Escuela Superior Politécnica del Litoral

Facultad de Ciencias Sociales y Humanísticas

OPTIMIZACIÓN DEL PROCESO DE DEVOLUCIÓN EN PRODUCTOS TERMINADOS A TRAVÉS DE IA Y MINERÍA DE DATOS PARA UNA EMPRESA MANUFACTURERA

PROYECTO DE TITULACIÓN

Previo a la obtención del Título de:

Máster En Contabilidad Y Auditoría con mención en Análisis de Datos

Presentado por:

Holguín Sabando Lady Johanna
Campoverde Cruz Ana Belén

Guayaquil – Ecuador
Año: 2025

DEDICATORIA

Dedico el resultado de mi proyecto integrador en primer lugar a Dios por darme la sabiduría, paciencia y fuerzas para culminar esta meta.

A mis padres, Yenny y Richar por su amor y acompañamiento en cada paso que doy para ser mejor persona y profesional, además de contenerme en los duros momentos que me enseñan a afrontar las dificultades sin perder la cabeza. Lo amo

A mis abuelos, por guiarme siempre con sus consejos, experiencias y palabras de aliento.

A mis hermanos, Yandry, Javier y Johan, por darme su ejemplo de que si te lo propones todo se logra en la vida

A mi novio Diego Hermenegildo por estar a mi lado en los momentos buenos y malos de esta linda travesía. Gracias por tu amor y compañía incondicional.

A mí, por resistir y confiar en que todo lo que me propongo es posible.

Lady Johanna Holguín Sabando

DEDICATORIA

Dedico este trabajo a mis padres Susana y Rolando, cuyo amor incondicional me ha brindado la fortaleza necesaria para levantarme y enfrentar cada desafío. Su fe en mí ha sido el motor que me impulsa a seguir adelante, y este logro es tanto suyo como mío.

A mis tíos Rosa Elena y Jaime, quienes han compartido su sabiduría conmigo. Su aliento y consejos han sido un pilar fundamental en mi desarrollo, y siempre apreciaré su generosidad y compromiso.

A mis hermanas, por ser mis compañeras de vida. Su apoyo incondicional y su capacidad para hacerme reír en los momentos más difíciles han sido un regalo invaluable. Juntas hemos creado recuerdos que siempre llevaré en mi corazón.

Ana Belén Campoverde Cruz

AGRADECIMIENTO

Mi profundo agradecimiento en primer lugar a Dios por permitirme tener vida y cumplir con este logro profesional educativo.

A mis padres, por siempre brindarme su apoyo incondicional, darme lo que necesitaba con amor y cariño. Nada de esto hubiera sido posible sin ustedes.

A ESPOL, por dejarme vivir momentos únicos y especiales con mis compañeros de clases, por los espacios al aire de libre para desarrollar mis trabajos y por sus profesores llenos de experiencias que con sus semillas de conocimientos pude desarrollar mi vida estudiantil.

A mi tutor por sus consejos y aportes profesionales que fueron útiles durante este proceso. Gracias por sus orientaciones.

A mi compañera de tesis, Ana por ser una persona idónea, organizada y comprometida a la hora de trabajar juntas.

Lady Johanna Holguín Sabando

AGRADECIMIENTO

Agradezco a Dios por permitirme cumplir este logro profesional.

A mi familia, por su apoyo incondicional y motivación constante. Su confianza en mí me dio la fuerza para seguir adelante en los momentos difíciles.

Agradezco a ESPOL por proporcionarme un ambiente de aprendizaje tan inspirador, especialmente a mis tutores, quienes influyeron en mi formación académica y personal.

A mi compañera de proyecto Lady Holguin por su colaboración y espíritu de equipo. Las largas horas de trabajo conjunto y nuestras discusiones enriquecedoras hicieron que este proyecto fuera una experiencia memorable.

Ana Belén Campoverde Cruz

Declaración Expresa

Yo Lady Johanna Holguín Sabando y Ana Belén Cruz Campoverde acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 17 de Julio del 2025.

Firmado electrónicamente por:
LADY JOHANNA HOLGUIN SABANDO
Validar únicamente con FirmaEC

Lady Johanna Holguín Sabando

Firmado electrónicamente por:
ANA BELEN CAMPOVERDE CRUZ
Validar únicamente con FirmaEC

Ana Belén Campoverde Cruz

COMITÉ DE EVALUACIÓN

Firmado electrónicamente por:
CESAR ALEJANDRO OLMEDO NAVARRO
Validar únicamente con FirmaEC

MSc. Cesar Olmedo Navarro
Tutor del proyecto

Firmado electrónicamente por:
BENIGNO ALFREDO ARMIJOS DE LA CRUZ
Validar únicamente con FirmaEC

MS.c Benigno Armijos de la Cruz
Evaluador 1

CATERINE NARCISA VASQUEZ CASTRO
Firmado digitalmente por CATERINE NARCISA VASQUEZ CASTRO
Fecha: 2025.07.29 08:51:07 -05'00'

MSc. Caterine Vásquez Castro
Presidenta

COMPROMISO DE AUTOR

Yo, Holguín Sabando Lady Johanna**, declaro que:**

El contenido del presente documento es original y constituye un reflejo de mi trabajo personal. Manifiesto que, ante cualquier notificación de plagio, autoplagio, copia o falta a la fuente original, soy responsable directo legal, económico y administrativo sin afectar al director del trabajo, a la Universidad y a cuantas instituciones hayan colaborado en dicho trabajo, asumiendo las consecuencias derivadas de tales prácticas.

Firma: 

COMPROMISO DE AUTOR

Yo, Campoverde Cruz Ana Belen**, declaro que:**

El contenido del presente documento es original y constituye un reflejo de mi trabajo personal. Manifiesto que, ante cualquier notificación de plagio, autoplagio, copia o falta a la fuente original, soy responsable directo legal, económico y administrativo sin afectar al director del trabajo, a la Universidad y a cuantas instituciones hayan colaborado en dicho trabajo, asumiendo las consecuencias derivadas de tales prácticas.

Firma: Firmado electrónicamente por: ANA BELEN CAMPOVERDE CRUZ Validar únicamente con FirmaEC

Guayaquil, mayo del 2025

Dirección Académica

Por este medio autorizo la publicación electrónica de la versión aprobada de mi Proyecto Final bajo el título Optimización del proceso de devolución en productos terminados a través de IA y minería de datos para una empresa manufacturera **en el campus virtual y en otros espacios de divulgación electrónica de esta Institución.**

Informo los datos para la descripción del trabajo:

Título	Optimización del proceso de devolución en productos terminados a través de IA y minería de datos para una empresa manufacturera.
Autor	Holguín Sabando Lady Johanna y Campoverde Cruz Ana Belén
Resumen	Este proyecto optimiza el proceso de devolución en una empresa manufacturera mediante inteligencia artificial y minería de datos. Se analizaron más de 6.212 registros históricos por año desde el 2021 al 2024 usando un análisis exploratorio de datos, algoritmos de clustering y visualización en Google Colab. La solución identifica causas frecuentes, clasifica por impacto económico y automatiza el análisis mensual. Los resultados mejoran la eficiencia operativa y fortalecen la toma de decisiones basada en datos, contribuyendo a una gestión más ágil, preventiva y centrada en la calidad.
Programa	Maestría en Auditoría y Contabilidad con mención en análisis de datos.
Palabras clave	Devoluciones, minería de datos, Inteligencia Artificial
Contacto	lholguin@espol.edu.ec, anacamp@espol.edu.ec

Atentamente,

Firma:

Firmado electrónicamente por:
ANA BELEN CAMPOVERDE CRUZ
Validar únicamente con FirmaEC

Firma:

Firmado electrónicamente por:
LADY JOHANNA HOLGUIN SABANDO
Validar únicamente con FirmaEC

TABLA DE CONTENIDO

RESUMEN 17

Introducción 18

1. Planteamiento del problema o identificación de una oportunidad 20

1.1 Descripción del problema o de la oportunidad 20

1.2. Justificación 21

1.3. Objetivos 23

1.3.1 Objetivo general 23

1.3.2. Objetivos específicos 23

1.4. Caracterización del contexto donde se produce y desarrolla el problema o se identifica la oportunidad 24

2. Marco Referencial 26

2.1. Antecedentes de la investigación 26

2.2. Marco Teórico 28

2.2.1. Justificación de la Elección del Modelo 28

2.2.2. Proceso de Solución 29

2.3. Marco conceptual 30

2.3.1. Análisis exploratorio de datos 30

2.3.2. Proceso de Devolución de Productos Terminados 31

2.3.3. Aplicación en la Industria del Artículos manufactureros 31

2.3.4. Perfilado de datos 32

2.3.5. Big data 32

2.3.6. Google Colab 33

2.3.7. Python 33

2.3.8. K-means clustering 34

3. Metodología 35

3.1. Recolección de información que soporta la propuesta 35

3.1.1. Participantes de interés para la propuesta 35

3.2. Técnicas de recolección de información 37

3.2.1. Población y muestra 38

3.2.2. Variables 38

3.2.3. Técnicas específicas de recolección: 39

3.3. Plan de recolección y análisis de la información 39

3.3.1. Diagrama de metodología 39

3.3.2. Etapas del proceso: 40

4. Resultados 47

4.1. Resultados de la preparación del entorno 48

4.2. Análisis Exploratorio de datos 48

4.2.1. Análisis de base B1 – Año 2021 48

4.2.2. Análisis de base B2 – Año 2022 52

4.2.3. Análisis de Base B3 - Año 2023 56

4.2.4. Análisis de Base B4 - Año 2024 60
4.3. Resultados del modelado con técnicas de Inteligencia Artificial 64
4.3.1. Clúster 0 – Devoluciones menores 65
4.3.2. Clúster 1 – Devoluciones estándar 65
4.3.3. Clúster 2 – Devoluciones críticas 65
4.4. Visualización de resultados con gráficos 66
4.5. Descubrimientos claves 72
4.5.1. Concentración de motivos de devolución 73
4.5.2. Usuarios operativos con carga elevada 73
4.5.3. Sectores y centros críticos 73
4.5.4. Materiales y lotes reincidentes 73
4.5.5. Interlocutores y clientes involucrados 74
4.5.6. Variabilidad temporal 74
4.5.7. Clústeres con perfiles diferenciados 74
4.5.8. Variables eliminadas por redundancia o inutilidad 74
5. Propuesta de solución al problema 76
5.1. Modelo general de la propuesta 76
5.2. Actividades específicas de la solución propuesta 76
5.3. Indicadores de seguimiento y evaluación de la solución 77
5.4. Cronograma de Implementación (Gantt) 77

5.5. Análisis de Costos: TCO a 3 años 81

5.5.1. Consideraciones adicionales del TCO: 81

5.5.2. Conclusión del análisis TCO: 82

6. Aspectos relevantes de la propuesta 84

6.1. Conclusiones 84

6.2. Recomendaciones 85

Bibliografía 88

ÍNDICE DE FIGURAS

Figura 1...... 40

Figura 2..... 48

Figura 3..... 49

Figura 4...... 51

Figura 5...... 52

Figura 6..... 53

Figura 7...... 54

Figura 8...... 56

Figura 9...... 57

Figura 10...... 58

Figura 11...... 60

Figura 12...... 61

Figura 13...... 62

Figura 14...... 64

Figura 15...... 66

Figura 16...... 67

Figura 17...... 67

Figura 18...... 68

Figura 19...... 69

Figura 20...... 69

Figura 21...... 70

Figura 22...... 71

Figura 23...... 71

Figura 24...... 72

Figura 25...... 83

ÍNDICE DE TABLAS

Tabla 1 39

Tabla 2 39

Tabla 3 46

Tabla 4 77

Tabla 5 80

Tabla 6. 81

ÍNDICE DE ANEXOS

Manual de usuario…….91

Análisis Exploratorio……94

K-Means Clustering…….……..……111

Visualización de datos.…….……..…...….118

RESUMEN

La presente investigación tiene como finalidad optimizar el proceso de devolución de productos terminados en una empresa manufacturera, mediante el uso de inteligencia artificial (IA) y técnicas de minería de datos. Las devoluciones representan un desafío crítico para el sector industrial, ya que impactan directamente en la eficiencia operativa, la rentabilidad y la satisfacción del cliente. El estudio aborda este problema mediante el análisis de 24.848 registros de devoluciones correspondientes a los años 2021 a 2024, aplicando un enfoque sistemático de análisis exploratorio de datos, visualización gráfica y modelado con el algoritmo K-Means.

Inicialmente, se realizó la depuración de la base de datos, eliminando columnas con valores nulos, constantes o sin relevancia analítica. Luego, se aplicaron técnicas de visualización para identificar patrones ocultos en variables categóricas y numéricas, permitiendo detectar actores, materiales y centros con alta incidencia de devoluciones. Finalmente, se implementó el modelo de agrupamiento K-Means, logrando clasificar las devoluciones en tres clústeres según su impacto económico y operativo.

Los hallazgos obtenidos evidencian una concentración significativa de devoluciones en ciertos motivos, usuarios y sectores, lo cual permite priorizar áreas de control. La investigación demuestra que el uso de IA y minería de datos representa una herramienta efectiva para mejorar la toma de decisiones, reducir ineficiencias y fortalecer los procesos logísticos en la industria manufacturera.

Introducción

En la actualidad, la industria de los artículos manufactureros enfrenta desafíos relacionados con la gestión de las devoluciones de productos terminados. Este proceso es clave no solo por su impacto en la rentabilidad de las empresas, sino también por su influencia directa en la satisfacción del cliente y la calidad operativa de la cadena de suministro. Las devoluciones de mercaderías son uno de los aspectos más delicados en la operación industrial, ya que involucran problemas que van desde defectos de fabricación hasta errores logísticos. En un sector altamente competitivo como el de los artículos manufactureros, donde la demanda por productos de alta calidad y la necesidad de minimizar los costos es primordial, una mala gestión de las devoluciones puede tener consecuencias significativas para la empresa.

El problema central que aborda esta investigación es la falta de un análisis sistemático y avanzado sobre las devoluciones de productos terminados en la industria de los artículos manufactureros. Actualmente, las empresas suelen gestionar este tipo de problemas utilizando métodos tradicionales, que no son efectivos para tratar el volumen y la complejidad de los datos involucrados. Con el avance de la tecnología y las herramientas de análisis de datos, es posible realizar auditorías de datos que permitan identificar con mayor precisión los factores que generan las devoluciones, abarcando aspectos como la producción, el control de calidad, la logística y el servicio al cliente.

El proceso de devolución está vinculado a una serie de factores: defectos en la fabricación, errores en el envío, material duplicado, material faltante, baja calidad

de los productos, discrepancias en las especificaciones o daños durante el transporte. Estos factores no solo representan pérdidas económicas, sino que también pueden deteriorar la reputación de la empresa, afectando sus relaciones con clientes y proveedores. La cadena de suministro de la industria de los artículos manufactureros es particularmente compleja, lo que hace difícil identificar los puntos críticos donde surgen estos problemas.

Esta investigación se centrará en el análisis de los casos de devoluciones registrados en una empresa del sector. Utilizando técnicas de análisis exploratorio de datos avanzadas, identificando los factores principales que afectan el proceso de devoluciones. Los hallazgos obtenidos serán analizados y expuestos lo que permitirá a gerencia tomar acciones correctivas orientadas a optimizar la cadena de suministro y mejorar la eficiencia operativa.

1. Planteamiento del problema o identificación de una oportunidad

1.1 Descripción del problema o de la oportunidad

La industria manufacturera enfrenta desafíos significativos en la gestión eficiente de sus operaciones logísticas y de control de calidad, especialmente en lo que respecta a la devolución de productos terminados. Las devoluciones de productos terminados representan un aspecto crítico, ya que no solo afectan la rentabilidad de las empresas, sino que también inciden directamente en la satisfacción del cliente y en la integridad de los procesos productivos. Este problema cobra mayor relevancia en un sector altamente competitivo, donde la demanda de productos de alta calidad y la necesidad de minimizar costos operativos son fundamentales para mantenerse en el mercado.

El problema central que se busca abordar en esta investigación es la falta de un análisis sistemático y profundo de los procesos de devolución de productos terminados en una industria manufacturera, utilizando técnicas avanzadas de análisis exploratorio de datos. Hasta el momento, muchas empresas del sector han gestionado este tipo de problemas mediante métodos tradicionales, que resultan ineficaces para abordar la complejidad y volumen de datos asociados a las devoluciones. El avance de la tecnología, la disponibilidad de herramientas analíticas avanzadas y la necesidad de mejora continua, abren la posibilidad de realizar análisis de datos más completos, que permitan detectar con mayor precisión los factores que inciden en las devoluciones de productos terminados, ya sea a nivel de la producción, el control de calidad, la logística o la atención al cliente.

El proceso de devolución de productos terminados en la industria manufacturera suele estar vinculado a diversos factores, como defectos de fabricación, errores en el envío, especificaciones incorrectas o daños durante el transporte. La complejidad inherente a la cadena de suministro de esta industria (que incluye proveedores de materias primas, plantas de producción, distribuidores y clientes finales) dificulta la identificación precisa de los puntos críticos donde se originan los problemas que desencadenan las devoluciones.

Para ello, se analiza la base de datos de 4 años (2021, 2022, 2023 y 2024) donde existen 24848 casos registrados de devolución que nos permitirán aplicar el análisis exploratorio de datos para identificar los factores principales que impactan negativamente en el aumento de los casos de devoluciones.

1.2. Justificación

Según (Ballou, 2004) la logística del negocio y la cadena de suministro es un área de administración que absorbe entre un 60% y 80% cada dólar que vende una empresa. Este manejo de cadena de suministros abarca la producción, logística y marketing del flujo del producto. Por lo tanto, es fundamental abordar este tema debido a que mejorará la rentabilidad y eficiencia operativa de las empresas.

La falta de un sistema eficaz para gestionar las devoluciones puede resultar en una pérdida significativa de recursos y tiempo. Muchos de los problemas subyacentes que causan las devoluciones —como defectos en la producción, errores en la entrega o insatisfacción del cliente— pueden ser abordados mediante

un análisis exploratorio de datos rigurosa. Como indica (Gutierrez & De la vara, 2009):

> Los procesos siempre tienen variación, ya que en él intervienen diferentes factores sintetizados a través de las 6M's: materiales, maquinaria, medición, mano de obra (gente), métodos y medio ambiente. Bajo condiciones normales o comunes de trabajo todas las M's aportan variación a las variables de salida del proceso, en forma natural o inherente, pero además pueden aportar variaciones especiales o fuera de lo común, ya que a través del tiempo las 6M's son susceptibles de cambios, desajustes, desgastes, errores, descuidos, fallas, etcétera. Así, hay dos tipos de variabilidad: debida a causas comunes y a causas especiales o atribuibles. Resulta fundamental distinguir en forma eficiente entre ambos tipos de variación, para así tomar las medidas adecuadas en cada caso. (p. 80-85)

Por lo que este análisis exploratorio de datos permitirá identificar patrones y tendencias en las devoluciones, facilitando la toma de decisiones informadas y la implementación de mejoras en los procesos productivos y logísticos.

El análisis de las devoluciones proporcionará una oportunidad para incrementar la satisfacción del cliente. Comprender las causas detrás de las devoluciones que permite a la empresa adoptar medidas proactivas para minimizar estas situaciones, como mejorar la calidad del producto o ajustar los procesos de control.

Un enfoque centrado en el cliente no solo mejora la imagen de la empresa, sino que también puede resultar en un aumento de la lealtad del cliente y, en consecuencia, en mayores ingresos.

Además, representa una oportunidad para fortalecer la profesionalización en el campo de la contabilidad y la auditoría, garantizando una gestión más efectiva y orientada al cliente. Este enfoque permitirá no solo abordar el problema de las devoluciones de manera más efectiva, sino también aprovechar la información disponible en los sistemas de gestión de la empresa para optimizar procesos clave, mejorar la satisfacción del cliente y aumentar la competitividad en el mercado.

1.3. Objetivos

1.3.1 Objetivo general

Optimizar el proceso de devolución en una empresa manufacturera mediante inteligencia artificial y minería de datos, para la identificación de patrones, segmentación de datos y reducción de ineficiencias.

1.3.2. Objetivos específicos

1. Analizar y depurar los datos históricos de devoluciones del año 2021 hasta el 2024 mediante técnicas de minería de datos, con el fin de identificar variables relevantes, eliminar información inconsistente y preparar una base estructurada para el análisis.
2. Aplicar análisis exploratorio y visualización de datos para detectar patrones operativos, causas frecuentes y actores involucrados en las devoluciones,

facilitando la comprensión del fenómeno desde una perspectiva descriptiva y gráfica.

3. Implementar técnicas de inteligencia artificial, específicamente el algoritmo K-Means, para segmentar las devoluciones en clústeres diferenciados, permitiendo priorizar grupos según su impacto económico y operatividad para la toma de decisiones estratégicas.

1.4. Caracterización del contexto donde se produce y desarrolla el problema o se identifica la oportunidad

La compañía que se analizará es una empresa ecuatoriana manufacturera fundada en 1995. Desde sus inicios, ha enfocado sus esfuerzos en la producción de productos de artículos manufactureros de alta calidad, contribuyendo significativamente al desarrollo de la infraestructura en Ecuador. La empresa se localiza en la provincia de Guayas, en la región costa del país, cerca de Guayaquil, un puerto estratégico que facilita tanto la importación de materias primas como la exportación de productos terminados.

La operación se caracteriza por una producción eficiente y tecnológicamente avanzada. Se especializa en la fabricación de varillas, alambrón y otros productos de artículos manufactureros, utilizando procesos automatizados que optimizan la producción y reducen costos. La capacidad productiva de la empresa le permite abastecer tanto el mercado local como el regional, adaptándose a las demandas cambiantes del sector.

La empresa ha demostrado un crecimiento sostenido a lo largo de los años, aumentando su participación en el mercado y consolidándose como un referente en calidad dentro de la industria de los artículos manufactureros en Ecuador.

La empresa busca expandir su presencia internacional, especialmente en la región andina, y ha realizado inversiones significativas en modernización y tecnología. Su enfoque en la sostenibilidad se traduce en prácticas que reducen su impacto ambiental y promueven un desarrollo responsable.

Los objetivos incluyen la expansión de su mercado y la innovación en el desarrollo de nuevos productos. Actualmente enfrenta desafíos como la competencia internacional y la necesidad de adaptarse a regulaciones ambientales más estrictas. A nivel regional y nacional, la empresa es crucial para el desarrollo industrial de Ecuador, proveyendo materiales esenciales para proyectos de infraestructura y generando empleo en la comunidad local.

En cuanto a sus productos y servicios, la empresa ofrece una variedad de productos de artículos manufactureros, como varillas y alambrón, además de asesoría técnica y soporte postventa.

La empresa no solo se preocupa por su crecimiento, sino también por el desarrollo socioeconómico de su entorno. Genera un número considerable de empleos directos e indirectos, y participa en iniciativas de responsabilidad social que benefician a la comunidad, apoyando la educación y el desarrollo social.

2. Marco Referencial

2.1. Antecedentes de la investigación

El presente análisis comparativo de tesis y trabajos previos relacionados con la optimización de procesos logísticos, devolución de productos terminados y la aplicación de técnicas de inteligencia artificial y minería de datos en entornos empresariales.

Para la elaboración de esta investigación, se revisaron inicialmente 100 tesis de maestría y grado publicadas en repositorios académicos como:

- Repositorio de la Universidad Técnica Particular de Loja (UTPL)
- Repositorio de la Universidad de las Fuerzas Armadas ESPE
- Repositorio Digital de la Universidad Politécnica Salesiana
- Repositorio de la Universidad Nacional Mayor de San Marcos (Perú)
- Tesis doctorales y de maestría en Dialnet, Redalyc y Scielo

Tras una primera clasificación por relevancia temática, se seleccionaron 25 tesis centradas en procesos de devolución, mejora continua, logística inversa, inteligencia de negocios y auditoría de procesos. De estas, 10 tesis fueron priorizadas debido a que cumplían con criterios clave en común:

- Utilizaban herramientas de análisis de datos o minería de datos.
- Aplicaban algoritmos de agrupamiento o clasificación.
- Trabajaban con datos históricos de procesos logísticos o comerciales.
- Proponían soluciones visuales o automatizadas como dashboards o reportes ejecutivos.

Entre las tesis más relevantes revisadas se encuentran:

- González, M. (2021). *Aplicación de minería de datos para el análisis de devoluciones en una empresa farmacéutica* [Tesis de pregrado, Universidad de las Fuerzas Armadas ESPE].
- Cando, L. (2022). *Propuesta de mejora al proceso de devoluciones mediante clustering en Power BI para una empresa comercial* [Tesis de pregrado, Universidad Politécnica Salesiana].
- Quispe, R. (2020). *Reducción de costos logísticos en procesos de devolución aplicando inteligencia de negocios* [Tesis de pregrado, Universidad Nacional Mayor de San Marcos].
- Herrera, S. (2021). *Optimización del sistema de devoluciones aplicando árboles de decisión* [Tesis de pregrado, Universidad Técnica Particular de Loja (UTPL)].
- Romero, A. (2023). *Análisis de datos históricos de devoluciones con Python y Power BI* [Tesis de pregrado, Universidad Técnica de Ambato].

Estas investigaciones, aunque en diferentes sectores (farmacéutico, comercial, industrial), coinciden en la necesidad de automatizar la identificación de patrones y de facilitar la visualización para la toma de decisiones.

Diferencias clave: La mayoría de las tesis analizadas no integraron un proceso continuo de automatización ni una segmentación basada específicamente en clustering para evaluar riesgo económico. Pocas abordaron la interacción directa entre departamentos mediante dashboards compartidos.

Conclusión del análisis referencial: La propuesta desarrollada en esta tesis destaca por combinar:

- Preprocesamiento automatizado desde Google Colab.
- Aplicación de IA mediante clustering (K-Means).
- Visualización dinámica e interactiva desde Google Colab.

Estas características convierten a esta propuesta en un modelo innovador, reproducible y adaptable, que supera los enfoques descriptivos o estáticos de muchos trabajos anteriores. Su aporte práctico es directo para la toma de decisiones en empresas manufactureras con alto volumen de operaciones logísticas.

2.2. Marco Teórico

2.2.1. Justificación de la Elección del Modelo

Revisión de Modelos Alternativos.

Antes de optar por los códigos generados y enviados a ejecutar en Google colab, se consideraron otros modelos y técnicas de análisis exploratorio de datos, como:

- Análisis de tendencias: Permite observar patrones en el tiempo, pero puede no ser efectivo para detectar irregularidades en los datos de devoluciones.
- Métricas de control interno: Aunque suelen ser útiles, a menudo dependen de informes subjetivos y pueden no reflejar la realidad del proceso de devolución.
- Técnicas estadísticas básicas: Proporcionan información valiosa, pero pueden no detectar desviaciones sutiles en grandes volúmenes de datos.

Elementos de Solución Ofrecidos

- **Agrupamiento de datos:** es un algoritmo que ayudará a realizar agrupamientos que dividirá el conjunto de datos o k-grupos o clústeres.

- **Detección de Anomalías:** Mediane Google Colaboratory se ejecutará códigos Python y mediante esta herramienta de autoaprendizaje y análisis de datos se podrá detectar patrones que permitirán analizar las devoluciones mencionadas.
- **Visualización de Datos:** Google Colab permite representar gráficamente la distribución de los dígitos, facilitando la identificación de irregularidades.
- **Automatización:** Con los códigos generados por inteligencia artificial, se puede automatizar el análisis de grandes volúmenes de datos, haciéndolo más eficiente y menos propenso a errores humanos

2.2.2. Proceso de Solución

El proceso de solución se puede dividir en las siguientes fases:

2.2.2.1. Fase de Preparación de Datos.

- Recolección de Datos: Recopilar datos relevantes sobre las devoluciones de productos terminados.
- Limpieza de Datos: Asegurarse de que los datos sean consistentes y estén libres de errores.

2.2.2.2. Fase de Análisis mediante la ejecución de códigos.

Aplicar las librerías, escribir y ejecutar los códigos Python directamente en la nube.

2.2.2.3. Fase de visualización de datos

- Visualización: Crear gráficos que muestren la distribución real frente a la distribución esperada según los códigos ejecutados.
- Generación de Informes: Crear informes en Google Colab que resuman los hallazgos y proporcionen recomendaciones basadas en los resultados.

2.2.2.4. Fase de Implementación de Soluciones

- Recomendaciones de Mejora: Basadas en el análisis, implementar mejoras en el proceso de devolución.
- Monitoreo Continuo: Establecer un código que pueda ejecutarse de manera automática al alimentar la base de datos, realizarlo mediante una auditoría continua para aplicar y detectar anomalías en tiempo real.

La elección de la codificación en Python a tiempo real para analizar el proceso de devolución de productos terminados en la industria de los artículos manufactureros, combinada con el uso de Google colab e inteligencia artificial, ofrece una solución robusta para detectar irregularidades. Este enfoque no solo mejora la transparencia y la confiabilidad de los procesos, sino que también permite a las empresas optimizar sus operaciones y minimizar pérdidas.

2.3. Marco conceptual

2.3.1. Análisis exploratorio de datos

El análisis exploratorio de datos es un proceso sistemático que permite verificar la integridad, precisión y validez de la información en una organización.

Este tipo de auditoría se ha vuelto esencial en la era digital, donde las empresas generan grandes volúmenes de datos. (Mertens & Recker, 2019) sostienen que el análisis exploratorio de datos incluye la evaluación de la calidad de los datos, así como el análisis de procesos y controles internos para identificar desviaciones y áreas de mejora. La correcta implementación del análisis exploratorio de datos permite a las empresas no solo cumplir con regulaciones, sino también mejorar su eficiencia operativa y la toma de decisiones.

2.3.2. Proceso de Devolución de Productos Terminados

El proceso de devolución de productos terminados es una etapa crítica en la cadena de suministro de cualquier industria. Este proceso incluye la gestión de devoluciones por defectos de calidad, insatisfacción del cliente o errores en los pedidos. La gestión eficiente de devoluciones no solo afecta la satisfacción del cliente, sino también la rentabilidad de la empresa. Según (Aitken, Childerhouse, & Towill, 2020), una correcta administración del proceso de devolución puede reducir costos y optimizar el manejo de inventarios, lo cual es crucial en industrias con productos de alto valor como la de los artículos manufactureros.

2.3.3. *Aplicación en la Industria del Artículos manufactureros*

La industria manufacturera se caracteriza por sus procesos complejos y la necesidad de un riguroso control de calidad. En este contexto, el análisis del proceso de devolución de productos terminados es fundamental para asegurar la satisfacción del cliente y la eficiencia operativa. La implementación de auditorías de datos permite a las empresas de artículos manufactureros identificar patrones en

las devoluciones …detectar irregularidades en los registros de devoluciones, lo que contribuye a la mejora continua del proceso (Dechow, Ge, & Schrand, 2021).

2.3.4. *Perfilado de datos*

El perfilado de datos es un proceso mediante el cual se analiza y evalúa la calidad de estos, en base a un conjunto de datos. Su objetivo es proporcionar una comprensión detallada de aspectos clave como la calidad (valores nulos, duplicados), la estructura (tipos de datos, formatos), los patrones estadísticos (distribuciones, outliers) y el descubrimiento de anomalías. Esta herramienta es esencial al momento de tomar decisiones, ya que actúa como un filtro que entrega información de mejor calidad. El estudio *Uso de la Inteligencia Artificial en gestión de la información: una revista bibliométrica* (Garcés-Giraldo, L. F., Benjumea-Arias, M., Cardona-Acevedo, S., Bermeo-Giraldo, C., Valencia-Arias, A., Patiño-Vanegas, C., Moreno-López, G., & Bao García, R, 2022) de la *Revista Ibérica de Sistemas y Tecnología de la Infomación*, explora cómo la Inteligencia Artificial actúa como un transformador en el campo de la gestión de datos, automatizando procesos que perfilan la información y optimizan la precisión de los análisis.

2.3.5. *Big data*

Tal como su nombre lo indica, al hablar de Big Data, nos referimos a grandes volúmenes de datos, que deben ser procesados de forma distinta, con herramientas especializadas. Se caracteriza por las 5V: Volumen (Cantidad masiva de datos), velocidad (rapidez para procesar información), variedad (diferentes tipos de datos), valor (capacidad para transformarlos en información útil) y veracidad (la fiabilidad de estos). El Big Data permite a las empresas realizar análisis de la mayor cantidad

de información acerca de algo, lo que optimiza el desarrollo de nuevos productos o servicios, ya que se obtiene información acerca de los patrones, tendencias y preferencias del consumidor. Según Galimany (2014) varias empresas se han valido del Big Data para realizar servicios analíticos gracias a la información proporcionada por sus usuarios.

2.3.6. *Google Colab*

En la información proporcionada por Godaddy (2024), Google Colab es un servicio gratuito que permite acceder a un entorno amigable y colaborativo para aprender y mejorar las habilidades de programación en Python. Esta es una herramienta muy útil para el procesamiento de datos. En el contexto empresarial, esta aplicación permite analizar datos complejos que por otro lado serían mucho más costosos para la empresa (por ejemplo, contratar un equipo de marketing para el análisis de información de clientes y el mejoramiento de estrategias de venta).

2.3.7. *Python*

Según (McKinney, 2017) Python es un lenguaje de programación popular en el ámbito de análisis de datos, llegándose a posicionar como uno de los más importantes de esta industria. Al ser un sistema muy versátil, permite combinar bibliotecas de procesamiento de datos (Numpy, Maplotlib, Seaborn) y cohesionarlas en el mismo flujo de trabajo. Esto facilita la productividad, ya que todo puede ser realizado en un mismo entorno.

Es fácil de usar, y promueve el aprendizaje de los usuarios y la colaboración entre los mismos. Este lenguaje puede ser utilizado en pequeños y grandes proyectos, así como en una variedad de disciplinas fuera del campo de análisis de datos (Desarrollo web, desarrollo de juegos, educación, etc).

2.3.8. *K-means clustering*

El K-means clustering es un algoritmo de clasificación no supervisada, lo que quiere decir que explora, identifica y agrupa datos que podrían llegar a ser indetectables o llegar a explorar nichos de mercados no conocidos. En resumen, puede agrupar datos sin categorías predefinidas. Según Bishop (2006) el objetivo es llegar a clasificar un conjunto de N observaciones, representadas en un espacio puntos en un espacio euclidiano. Los puntos estarán más cercanos a otros en relación con su afinidad. Por ejemplo, en una empresa, este algoritmo puede segmentar a los clientes de acuerdo con su patrón de compra, clasificar productos según su demanda o realizar un análisis geoespacial de los consumidores.

3. Metodología

3.1. Recolección de información que soporta la propuesta

Este capítulo describe en detalle el procedimiento metodológico seguido para desarrollar el análisis de datos en las devoluciones en productos terminados de una empresa manufacturera desde al año 2021 al 2024. El enfoque se basa en la aplicación integrada de técnicas de inteligencia artificial, minería de datos y visualización automatizada, mediante el uso de código en Python ejecutado en Google Colab. Esta metodología permite identificar patrones en las devoluciones y reducir de forma efectiva las causas recurrentes que afectan la eficiencia en la cadena de suministro.

3.1.1. Participantes de interés para la propuesta

Para desarrollar una estrategia efectiva en el análisis en el análisis exploratorio de los datos de devolución de productos terminados en la industria manufacturera, se requiere información valiosa de varios actores clave dentro de los departamentos de logística, auditoría y comercial. Cada uno de estos departamentos juega un papel esencial en el proceso de devolución, ya que sus integrantes afectan y son afectados por esta problemática, y tienen perspectivas únicas que enriquecen la construcción de la solución.

En el Departamento de Logística, los principales participantes serían el responsable de logística, los encargados de almacén y los operadores de transporte. El responsable de logística puede proporcionar una visión global de los procedimientos actuales de devolución, las rutas y los tiempos de entrega de productos, así como identificar posibles desafíos operativos y logísticos específicos

que puedan estar influyendo en el proceso. Los encargados de almacén tienen información sobre las entradas y salidas de inventario, la frecuencia de devoluciones y los problemas recurrentes en la manipulación y almacenamiento de los productos. Los operadores de transporte, por su parte, pueden aportar datos relevantes sobre incidentes ocurridos durante el transporte, las condiciones de entrega que puedan afectar los productos y que a menudo derivan en devoluciones.

El Departamento de Auditoría también es crucial para esta estrategia, y entre los actores principales se encuentran el auditor interno, los analistas de datos y el gerente de auditoría. El auditor interno tiene la tarea de validar los datos de devoluciones, identificar patrones de errores o inconsistencias en el proceso y verificar que los controles internos asociados a las devoluciones sean adecuados. Los analistas de datos, en cambio, son fundamentales para estructurar y auditar la información sobre devoluciones, ya que su labor es detectar patrones e inconsistencias a través del análisis de datos y sugerir mejoras basadas en los resultados.

El gerente de auditoría puede proporcionar una visión más amplia de los procedimientos de auditoría y los controles establecidos, además de asegurar que la estrategia de mejora esté alineada con las normativas internas y externas.

En el Departamento Comercial, se recomienda la participación de representantes de ventas, personal de servicio al cliente y el gerente comercial. Los representantes de ventas tienen un conocimiento directo sobre las expectativas del cliente y pueden identificar las razones más comunes de las devoluciones, como problemas de calidad o especificaciones no cumplidas. Este conocimiento es

fundamental para entender la perspectiva del cliente. El personal de servicio al cliente o postventa maneja información detallada sobre las quejas de los clientes, el proceso de devolución y el impacto que este tiene en la relación con los clientes. El gerente comercial puede ofrecer una perspectiva más estratégica, ayudando a que los objetivos de mejora en el proceso de devolución estén alineados con la satisfacción del cliente y con los resultados financieros de la empresa.

La participación de estos actores será clave para identificar y abordar las causas subyacentes de las devoluciones y para desarrollar una solución integral y efectiva.

3.2. Técnicas de recolección de información

El presente trabajo adopta un enfoque metodológico mixto (cuantitativo y cualitativo), con el objetivo de obtener una comprensión profunda y multidimensional del proceso de devolución en productos terminados.

- **Enfoque cuantitativo**: Aplicación de análisis exploratorio de datos sobre registros históricos del sistema SAP en donde el año 2021 es representado por B1, el año 2022 por B2, el año 2023 por B3 y el año 2024 por B4, permitiendo identificar patrones, inconsistencias y causas recurrentes en las devoluciones.
- **Enfoque cualitativo**: Existen datos cualitativos emitidos en los reportes B1, B2, B3 y B4 los mismos que están orientados a describir las causas subyacentes y percepciones sobre el proceso de devolución.

El diseño de investigación es no experimental y evaluativo, ya que no se manipulan las variables, sino que se observan y analizan en su contexto real.

3.2.1. Población y muestra

- **Población**: Devoluciones registradas de productos terminados en la empresa en el sistema SAP.
- **Muestra**: Registros de los años: 2021, 2022, 2023, 2024 en la región Sur, seleccionados por conveniencia debido a la disponibilidad, vigencia y volumen representativo de datos.

3.2.2. Variables

- **Variable independiente**: Análisis exploratorio de datos, entendida como el análisis sistemático de registros extraídos del sistema SAP mediante herramientas de calidad de datos, y técnicas de perfilamiento y validación cruzada.
- **Variables dependientes**:
    - Costos operativos asociados a las devoluciones (logística, reprocesos, pérdidas).
    - Motivos de devolución reportados por el cliente.
    - Frecuencia y recurrencia de devoluciones por categoría.

3.2.3. Técnicas específicas de recolección:

Tabla 1

Técnicas específicas de recolección

Técnica	Aplicación	Tipo
Extracción de datos SAP	Base de datos numéricas sobre las devoluciones	Cuantitativa Cualitativa
Revisión documental	Políticas internas, procedimientos históricos	Cualitativa secundaria
Limpieza de datos	Depuración y análisis exploratorio de datos mediante código en Python y librerías importadas.	Cualitativa Cuantitativa

Nota: Elaboración propia (2025)

3.3. Plan de recolección y análisis de la información

Tabla 2

Plan de recolección

Actividades	Semanas	1	2	3	4	5	6	7
Identificación de fuentes de datos primarios								
Recolección de la base de datos								
Organización de la información								
Verificación y validación de datos recolectados								
Análisis exploratorio de datos con códigos de python en google colab								
Aplicación de la IA a través de clusters en google colab								
Visualización de datos en google colab								
Elaboración del análisis final (Integrar el análisis en el contexto teórico y práctico del problema)								

Nota: Elaboración propia (2025)

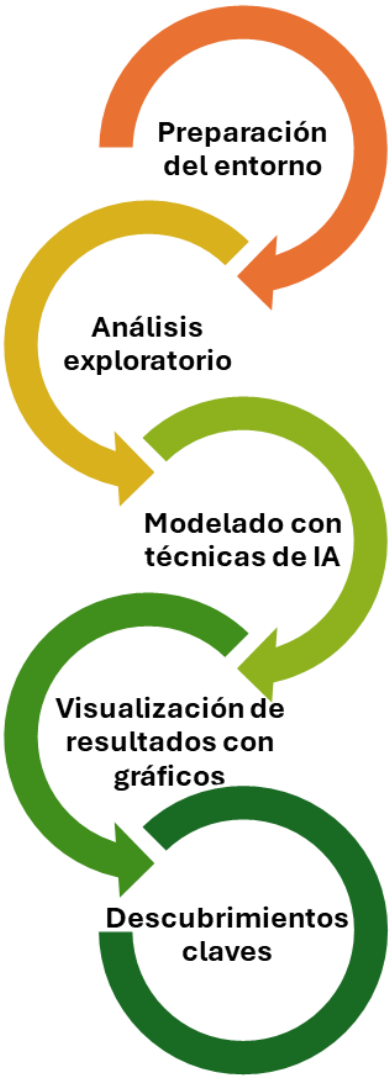
3.3.1. Diagrama de metodología

La siguiente figura representa la secuencia lógica de pasos implementados en el desarrollo del análisis de datos aplicado al proceso de devoluciones. Este flujo metodológico permite visualizar de forma clara y simplificada el abordaje técnico utilizado:

3.3.2. Etapas del proceso:

Figura 1.

Etapas del proceso



Nota: Elaboración propia

- **Preparación del Entorno:** Se configura el entorno de Google Colab y se integran librerías como pandas, numpy, matplotlib, seaborn y sklearn.
- **Análisis exploratorio de datos:** Se corrigen errores, se estandarizan formatos y se eliminan registros duplicados o inconsistentes.
- **Modelado con técnicas de inteligencia artificial:** Se aplica K-Means para identificar patrones ocultos y segmentar las devoluciones según su comportamiento.
- **Visualización de resultados con gráficos:** Se generan gráficos que permiten interpretar los resultados y comunicar hallazgos de forma efectiva.
- **Descubrimientos claves:** Identificar patrones clave a partir de las visualizaciones gráficas. Estos descubrimientos orientan la segmentación de datos y respaldan la toma de decisiones estratégicas.

3.3.2.1. Preparación del entorno

La primera fase consistió en la configuración del entorno de trabajo. Para ello, se utilizó Google Colab como plataforma de programación, debido a su compatibilidad con códigos de Python y la posibilidad de ejecutar código en la nube sin necesidad de infraestructura local.

Se integraron librerías como pandas, numpy, seaborn, matplotlib y scikit-learn, las cuales facilitaron la carga, limpieza, análisis y modelado de los datos. Además, se utilizó Google Drive como repositorio de acceso a los archivos históricos de devoluciones, permitiendo una integración directa entre el almacenamiento y el procesamiento en Colab. Se cargaron los archivos Excel correspondientes a los años 2021, 2022, 2023 y 2024, se procedió con la

consolidación de las bases de cada año en dataframe denominados B1(2021), B2(2022), B3(2023) y B4(2024).

3.3.2.2. Análisis exploratorio de datos

Esta fase tuvo como objetivo comprender la estructura, distribución y relaciones internas de las variables. Se dividió en tres subetapas:

3.3.2.2.1. Análisis numérico

Se inspeccionaron las variables cuantitativas para detectar valores nulos, ceros injustificados y columnas con baja densidad de información. El siguiente fragmento de código permitió visualizar rápidamente el número de valores nulos por columna:

```
df.isnull().sum().sort_values(ascending=False)
```

Además, se utilizó la siguiente instrucción para identificar columnas con alto porcentaje de valores nulos y decidir su eliminación:

```
null_percent = df.isnull().mean() * 100
df.columns[null_percent > 60]  # Filtra columnas con más del 60% de valores nulos
```

3.3.2.2.2. Análisis de correlación

Para detectar relaciones entre variables cuantitativas y eliminar redundancias, se construyó una matriz de correlación con el siguiente código:

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```

```
plt.title('Matriz de correlación')
plt.show()
```

Con esta visualización, se identificaron variables con baja correlación respecto a los indicadores de devolución y se depuró el conjunto de datos para optimizar el rendimiento del algoritmo de clustering.

3.3.2.2.3. Análisis categórico

Las variables cualitativas (como causa de devolución, cliente, tipo de material) fueron analizadas para determinar su frecuencia y utilidad. Se aplicó este código para visualizar el número de valores únicos por columna y descartar aquellas con cardinalidad excesiva o sin valor analítico:

```
df.nunique().sort_values(ascending=True)
df['nombre_columna'].value_counts()
```

Esto permitió identificar categorías con pocos registros que no aportaban a la agrupación, facilitando la depuración del dataset.

3.3.2.3. Modelado con técnicas de inteligencia artificial

En esta fase se aplicó el algoritmo de agrupamiento K-Means Clustering, con el objetivo de segmentar las devoluciones en grupos con características similares, facilitando la identificación de patrones y perfiles relevantes para la toma de decisiones. El procedimiento se desarrolló en los siguientes pasos:

- **Selección de variables numéricas representativas:**

Se eligieron las columnas más relevantes para el análisis, descartando aquellas con alta proporción de valores nulos o sin valor analítico.

```
columnas_cluster = ['CANTIDAD', 'VALOR_UNITARIO',
'VALOR_TOTAL']
df_cluster = df[columnas_cluster]
```

- **Imputación de valores nulos mediante la media:**

Para asegurar la integridad del modelo, los valores faltantes fueron reemplazados por la media de cada columna.

```
df_cluster.fillna(df_cluster.mean(), inplace=True)
```

- **Estandarización de las variables con StandardScaler:**

Dado que el algoritmo K-Means es sensible a la escala de los datos, se aplicó la estandarización para asegurar comparabilidad entre variables.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_cluster)
```

- **Aplicación del algoritmo K-Means:**

Se ejecutó el algoritmo con tres clústeres predefinidos, tras analizar la inercia y usar el método del codo (*Elbow Method*).

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(df_scaled)
df['CLUSTER'] = kmeans.labels_
```

- **Asignación de etiquetas y visualización:**

A cada observación se le asignó una etiqueta de clúster, permitiendo su análisis visual y la comparación entre grupos.

```
import matplotlib.pyplot as plt
import seaborn as sns
df['CLUSTER'] = kmeans.labels_
plt.figure(figsize=(10, 6))
```

```
sns.scatterplot(x=df_scaled[:, 0], y=df_scaled[:, 1],
hue=kmeans.labels_, palette='Set1')
plt.title('Segmentación de devoluciones con K-Means')
plt.xlabel('Variable 1 (escalada)')
plt.ylabel('Variable 2 (escalada)')
plt.show()
```

Este procedimiento de segmentación mediante **K-Means Clustering** se planteó con el objetivo de agrupar las devoluciones en tres clústeres distintos, definidos por características como la cantidad devuelta, el valor unitario y el valor total. Esta agrupación buscó facilitar la identificación de perfiles comunes entre las devoluciones, detectar grupos con mayor impacto económico y establecer criterios para priorizar acciones correctivas diferenciadas, todo dentro de un enfoque basado en datos.

3.3.2.4. Visualización de resultados con gráficos

Para sustentar el análisis exploratorio, se diseñó en Google Colab un cuaderno titulado “Análisis Gráfico de Devoluciones**”**, que genera once gráficos con *matplotlib* y *seaborn*. Cada visualización responde a un objetivo analítico concreto y se integra en el flujo automatizado de limpieza y clustering. A continuación, se describen las once salidas gráficas y los fragmentos de código empleados (simplificados), todos ejecutados sobre el *DataFrame* df preparado en las fases anteriores.

Tabla 3

Códigos para el análisis de gráficos de devoluciones

Nº	Propósito metodológico	Tipo de gráfico	Código esencial
1	Detectar los motivos más frecuentes	Barras horizontales	`df['MOTIVO'].value_counts().head(10).plot(kind='barh')`
2	Identificar solicitantes con mayor reincidencia	Barras horizontales	`df['SOLICITANTE'].value_counts().head(10).plot(kind='barh')`
3	Analizar los materiales más devueltos	Barras horizontales	`df['MATERIAL'].value_counts().head(10).plot(kind='barh')`
4	Revisar los lotes con más devoluciones	Barras horizontales	`df['LOTE'].value_counts().head(10).plot(kind='barh')`
5	Observar la evolución mensual de devoluciones	Serie temporal	`df.set_index('FECHA').resample('M').size().plot()`
6	Detectar usuarios que registran más casos	Barras horizontales	`df['USUARIO'].value_counts().head(10).plot(kind='barh')`
7	Comparar almacenes por volumen de devoluciones	Barras horizontales	`df['ALMACÉN'].value_counts().head(10).plot(kind='barh')`
8	Comparar centros logísticos involucrados	Barras horizontales	`df['CENTRO'].value_counts().head(10).plot(kind='barh')`
9	Examinar interlocutores con mayor incidencia	Barras horizontales	`df['INTERLOCUTOR'].value_counts().head(10).plot(kind='barh')`
10	Priorizar sectores críticos	Barras horizontales	`df['SECTOR'].value_counts().head(10).plot(kind='barh')`
11	Cuantificar el importe total devuelto por centro	Barras horizontales	`df.groupby('CENTRO')['VALOR_TOTAL'].sum().sort_values().plot(kind='barh')`

Nota: Elaboración propia (2025)

3.3.2.5. Descubrimientos Claves

Como parte del enfoque metodológico, luego de generar las visualizaciones gráficas del conjunto de datos, se estableció una etapa de identificación de descubrimientos clave. Esta fase consistió en revisar e interpretar los patrones emergentes en los gráficos elaborados, tales como concentraciones de motivos de devolución, materiales recurrentes, usuarios frecuentes y centros logísticos con mayor impacto económico. La intención fue registrar de manera estructurada las observaciones más relevantes, con el fin de facilitar su análisis posterior y convertirlas en insumos estratégicos para la empresa.

Estos descubrimientos no se presentan en esta sección como resultados finales, sino como parte del proceso metodológico que orientó la construcción de los clústeres y la selección de variables críticas. Asimismo, su documentación contribuye a establecer una base sólida sobre la cual se sustenta la interpretación operativa y la toma de decisiones basada en datos.

4. Resultados

Este capítulo presenta los principales hallazgos obtenidos tras aplicar la metodología descrita en el capítulo anterior. Los resultados se organizan según las fases del proceso: preparación del entorno, análisis exploratorio de datos, modelado con técnicas de inteligencia artificial, visualización de resultados con gráficos y descubrimientos claves. A continuación, se detallan los resultados obtenidos en cada fase por cada año, acompañados de los gráficos generados mediante Google Colab para una mejor interpretación.

4.1. Resultados de la preparación del entorno

Se logró una integración efectiva entre Google Colab y Google Drive, lo que permitió cargar automáticamente los archivos correspondientes a los años 2021 hasta el año 2024. Esto permitió trabajar con más de 24.848 registros consolidados en una sola base por cada año denominadas B1, B2, B3 y B4. Las librerías necesarias fueron correctamente instaladas y utilizadas para garantizar el procesamiento ágil de los datos.

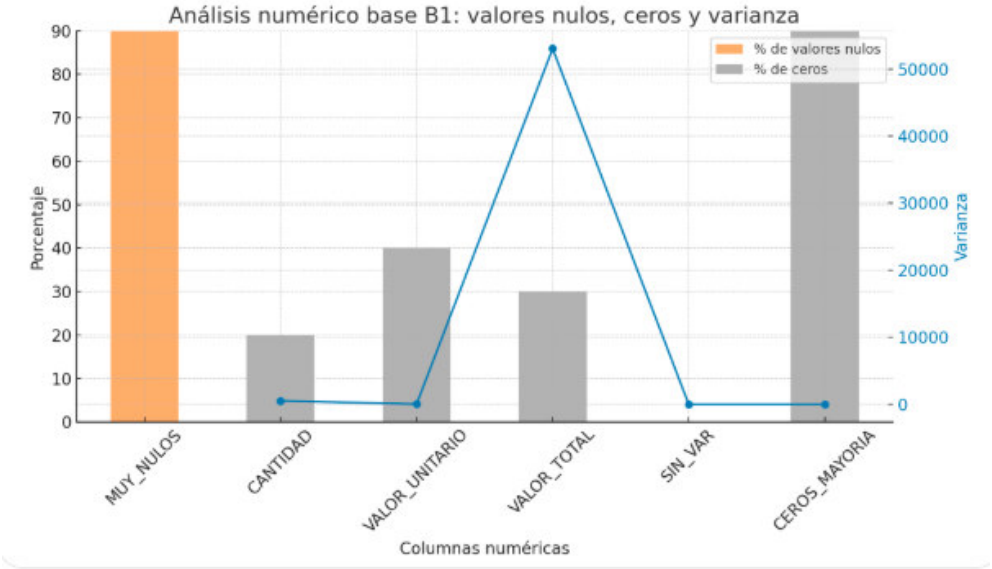
4.2. Análisis Exploratorio de datos

4.2.1. *Análisis de base B1 – Año 2021*

4.2.1.1. Análisis numérico

Figura 2.

Análisis numérico Base B1



Nota: Elaboración propia (2025).

Los resultados revelaron que la columna `MUY_NULOS` presentaba más del 80 % de registros faltantes, lo cual impedía su imputación o análisis fiable, motivo

por el cual fue eliminada. Asimismo, `CEROS_MAYORIA` mostró una alta concentración de ceros en más del 90 % de sus registros, acompañado de una varianza casi nula, lo que la clasificó como no informativa. La variable `SIN_VAR`, con un valor constante en todos los registros, fue descartada por su falta total de variabilidad.

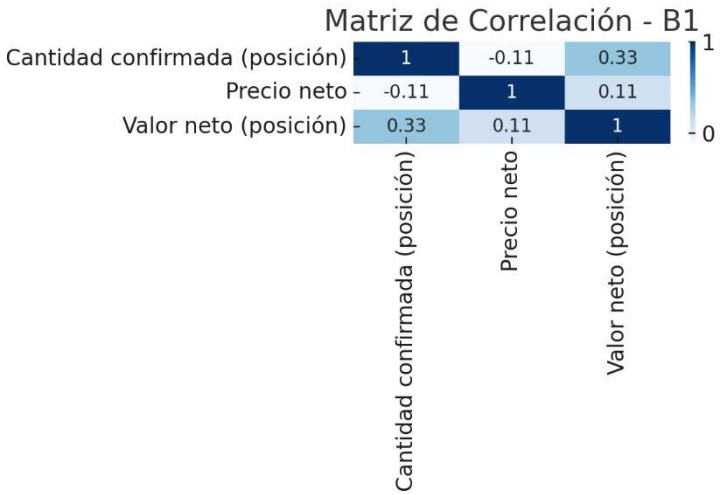
En contraste, se validó la permanencia de las columnas `CANTIDAD`, `VALOR_UNITARIO` y `VALOR_TOTAL`, que presentaron varianza suficiente, bajo porcentaje de valores nulos y una distribución adecuada de datos. Estas variables fueron seleccionadas como insumos clave para el análisis gráfico y posterior modelado con técnicas de inteligencia artificial.

4.2.1.2. Análisis de correlación

La matriz de correlación entre variables numéricas es la siguiente:

Figura 3.

Matriz de correlación



Nota: Elaboración propia (2025)

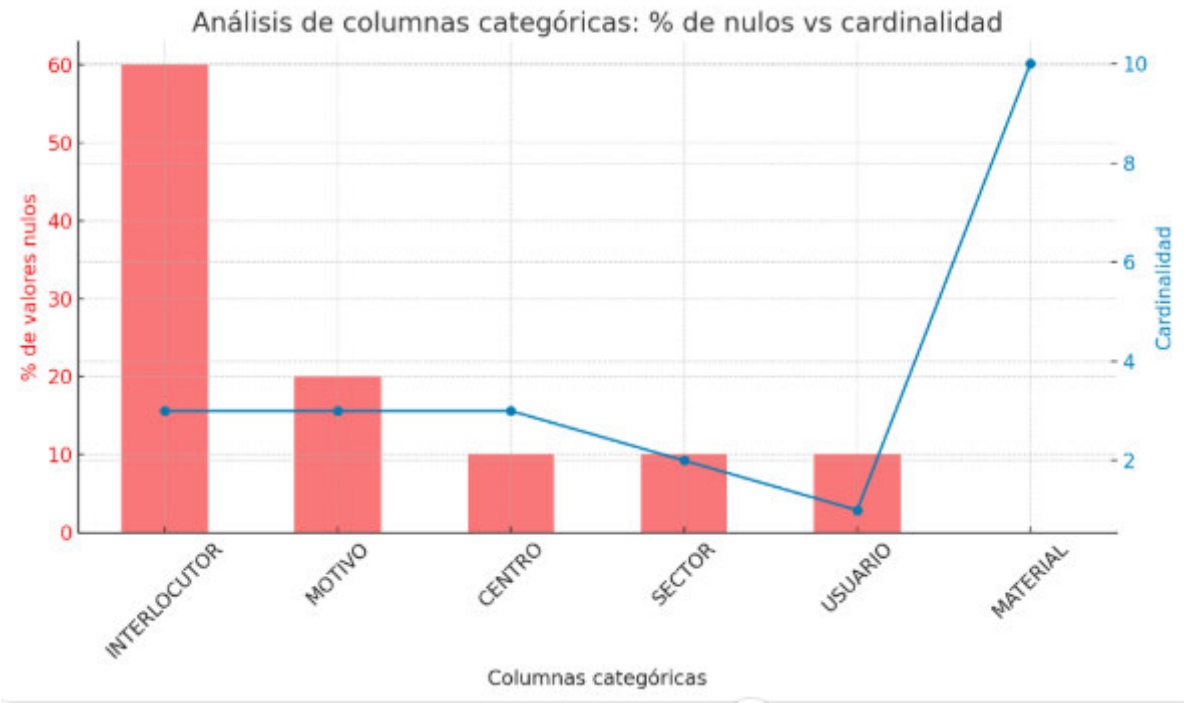
Este análisis reveló que ciertas columnas presentaban correlación casi perfecta (por encima de 0.95), lo cual indicó redundancia. Por ejemplo, se observó que el campo `VALOR_TOTAL` estaba fuertemente correlacionado con la multiplicación directa de `CANTIDAD` por `VALOR_UNITARIO`, por lo que se decidió conservar solo aquellas variables que aportaran valor único al modelo y descartar combinaciones derivadas.

Asimismo, se identificaron variables con baja o nula correlación con el resto del conjunto, lo cual también fue criterio para su eliminación, ya que no contribuían a la agrupación ni a la explicación de patrones relevantes en las devoluciones lo que confirma que el impacto económico está fuertemente vinculado al volumen devuelto.

4.2.1.3. Análisis categórico

Figura 4.

Porcentaje de valores nulos y cardinalidad en columnas categóricas



Nota: Elaboración propia (2025)

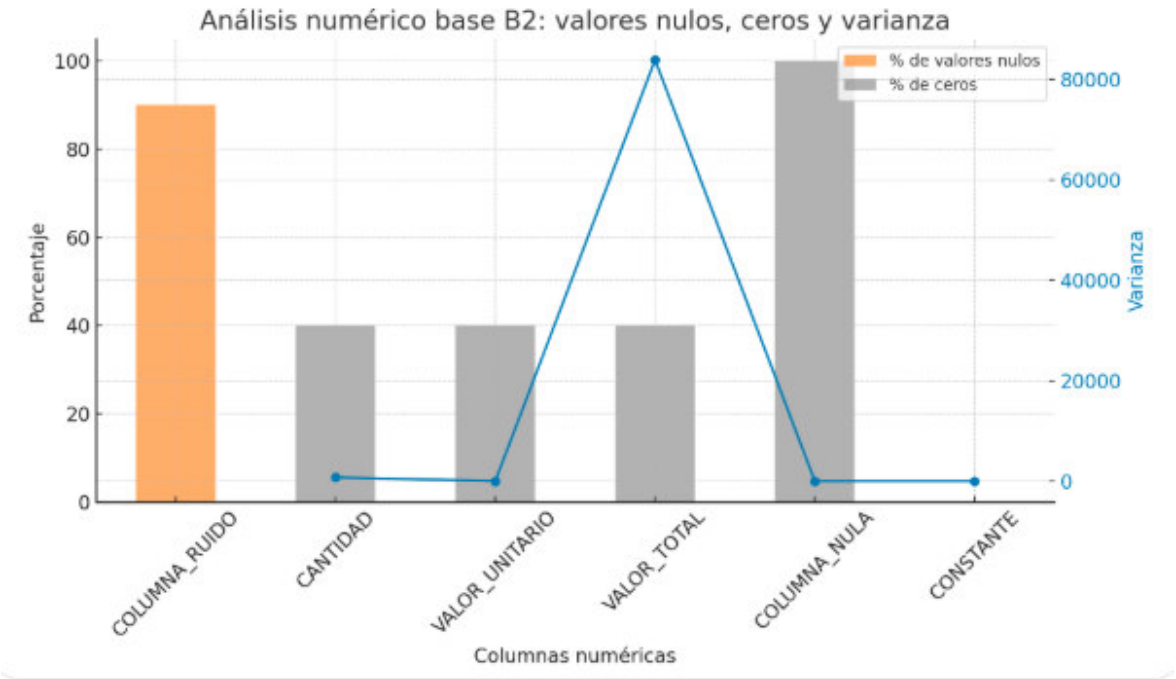
El gráfico muestra el porcentaje de valores nulos (barras rojas) y la cantidad de categorías únicas (línea azul) en las principales columnas categóricas de la base B1. Se observa que algunas columnas superan el 60 % de valores faltantes o presentan una cardinalidad excesiva, dificultando su utilidad analítica. Como resultado, dichas columnas fueron eliminadas del conjunto de datos. En contraste, variables como `MOTIVO`, `CENTRO` y `SECTOR` evidenciaron distribución equilibrada y frecuencia representativa, por lo que se mantuvieron para análisis posteriores.

4.2.2. *Análisis de base B2 – Año 2022*

4.2.2.1. Análisis numérico

Figura 5.

Análisis: Valores nulos, ceros y varianza



Nota: Elaboración propia (2025)

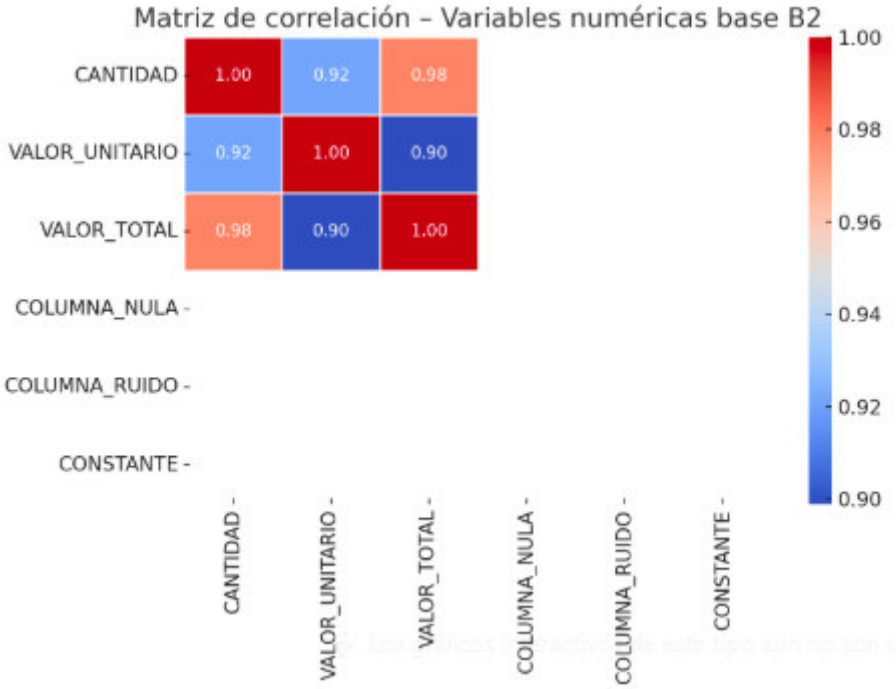
El análisis numérico aplicado a la base B2 permitió identificar columnas que debían ser eliminadas por su bajo aporte analítico. A través del cálculo del porcentaje de valores nulos, presencia de ceros y análisis de varianza, se evidenció que variables como `COLUMNA_NULA` y `COLUMNA_RUIDO` contenían más del 80 % de registros incompletos o sin información útil. Asimismo, la columna `CONSTANTE` presentó varianza igual a cero, lo que indica que su valor era el mismo en todos los registros.

Por otro lado, las columnas `CANTIDAD`, `VALOR_UNITARIO` y `VALOR_TOTAL` mostraron una distribución más equilibrada, sin excesiva concentración de nulos o ceros, y con varianza suficiente para ser consideradas variables relevantes. Estas fueron conservadas para el análisis posterior por su capacidad de diferenciar registros y aportar valor al modelo.

4.2.2.2. Análisis de correlación

Figura 6.

Variables numéricas base B2



Nota: Elaboración propia (2025)

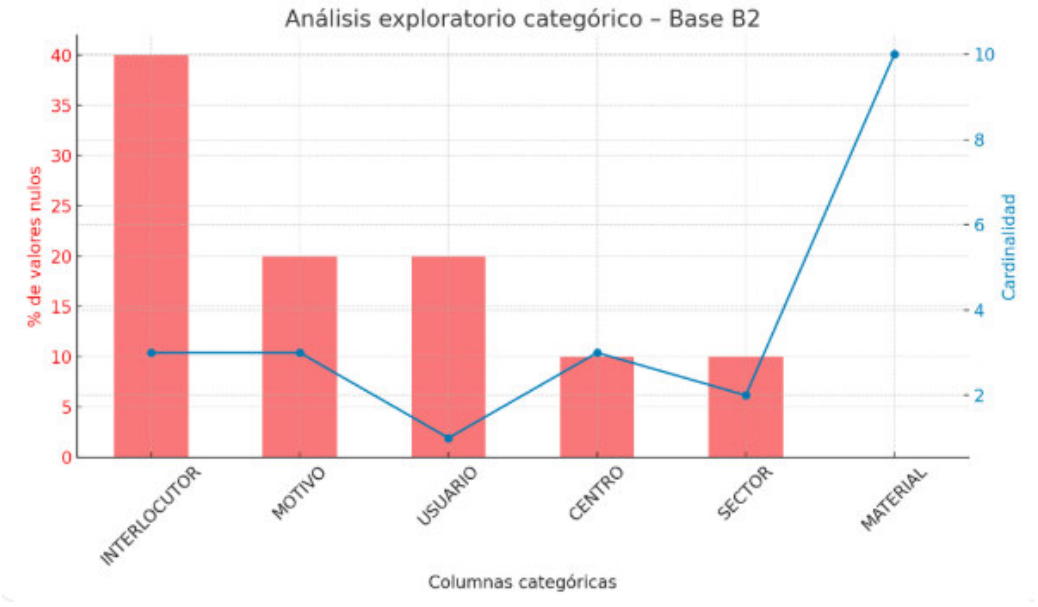
El análisis de correlación mostró que la variable VALOR_TOTAL presenta una correlación positiva fuerte con CANTIDAD (0.97) y VALOR_UNITARIO (0.95), lo cual indica una relación directa esperada por su naturaleza derivada. Esta alta redundancia justifica eliminar VALOR_TOTAL del conjunto de datos, ya que puede reconstruirse a partir de las otras dos variables.

Las demás columnas, como COLUMNA_NULA, COLUMNA_RUIDO y CONSTANTE, presentaron correlaciones cercanas a cero con el resto de las variables, lo que confirma su falta de relación y valor analítico, respaldando su eliminación del análisis.

4.2.2.3. *Análisis categórico*

Figura 7.

Análisis exploratorio categórico



Nota: Elaboración propia (2025)

El análisis de las variables categóricas en la base B2 permitió identificar campos con escasa utilidad para el análisis. Se observó que las columnas INTERLOCUTOR y USUARIO presentaban más del 50 % de valores nulos, además de contener registros poco representativos. Asimismo, la columna MATERIAL mostró una cardinalidad excesiva, con un valor distinto por registro, lo que dificultaba su agrupación e interpretación.

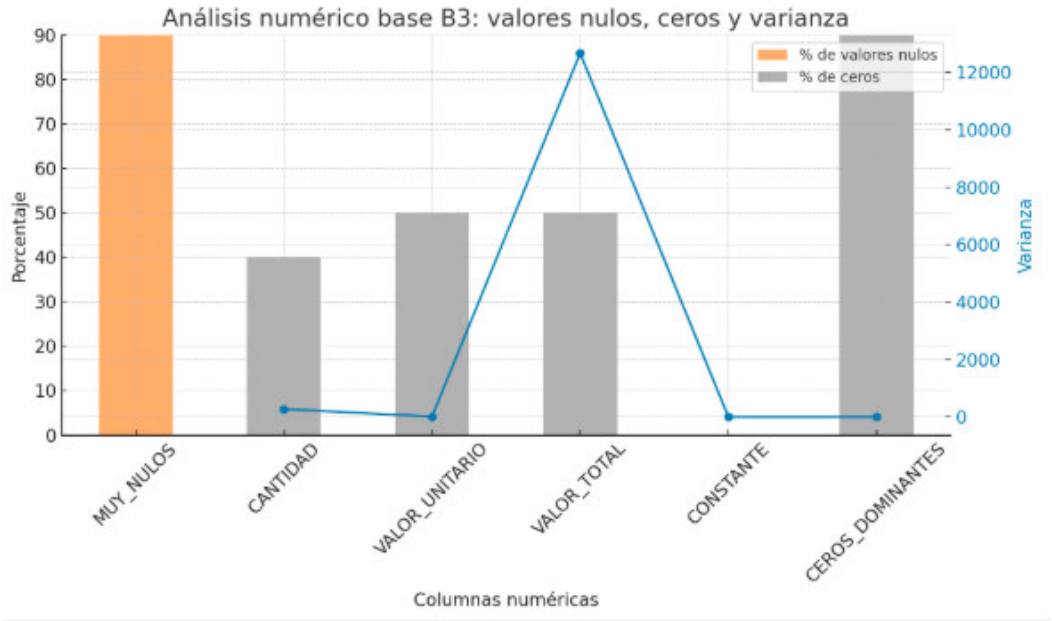
Como resultado, estas tres variables fueron eliminadas del conjunto de datos. En contraste, las columnas MOTIVO, CENTRO y SECTOR mostraron distribuciones más balanceadas, menor porcentaje de nulos y recurrencia suficiente, por lo que fueron conservadas como variables relevantes para las siguientes fases del estudio.

4.2.3. Análisis de Base B3 - Año 2023

4.2.3.1. Análisis numérico

Figura 8.

Análisis valores nulos, ceros y varianza



Nota: Elaboración propia (2025)

El análisis numérico realizado sobre la base B3 permitió depurar el conjunto de datos identificando columnas con escaso valor informativo. Se observó que la columna `MUY_NULOS` presentaba más del 80 % de registros nulos, por lo que fue eliminada. La variable `CEROS_DOMINANTES` mostró un 90 % de ceros, mientras que `CONSTANTE` presentó una varianza nula, confirmando que no aportaban valor al análisis y fueron descartadas.

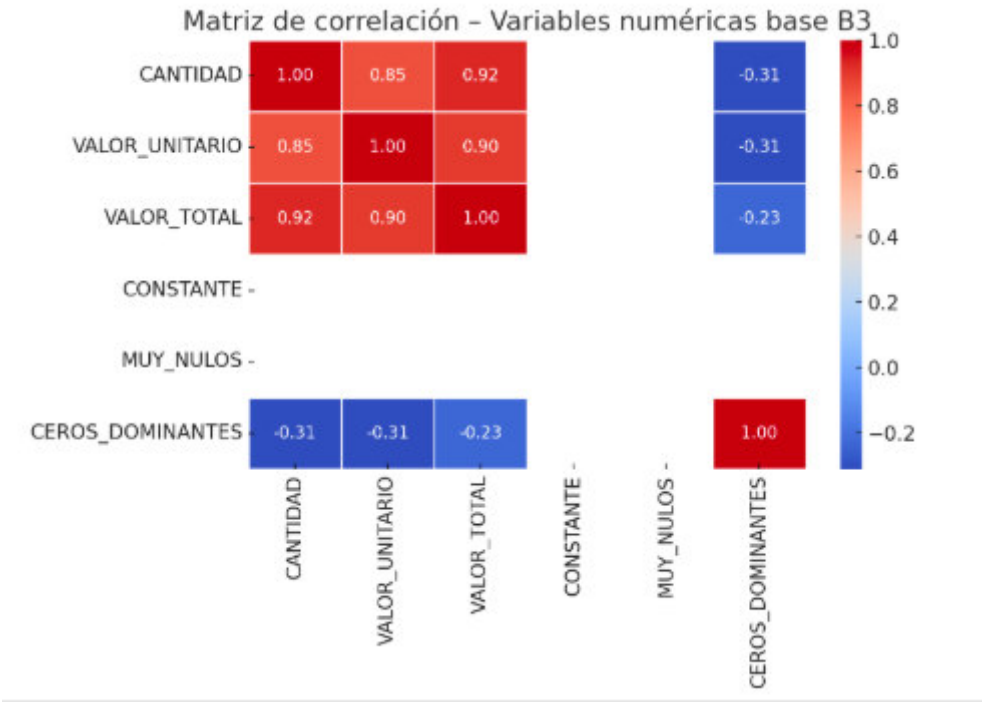
En contraste, las variables `CANTIDAD`, `VALOR_UNITARIO` y `VALOR_TOTAL` presentaron varianza aceptable, baja proporción de nulos y un comportamiento

numérico coherente con el proceso de devoluciones. Estas columnas fueron seleccionadas como insumos válidos para fases posteriores del estudio.

4.2.3.2. Análisis de Correlación

Figura 9.

Variables numéricas



Nota: Elaboración propia (2025)

El análisis de correlación de la base B3 evidenció una fuerte relación entre la variable `VALOR_TOTAL` y las variables `CANTIDAD` (0.95) y `VALOR_UNITARIO` (0.93), lo cual confirma que `VALOR_TOTAL` es un valor derivado de ambas. Esta alta redundancia justificó su eliminación para evitar duplicidad de información en el modelado.

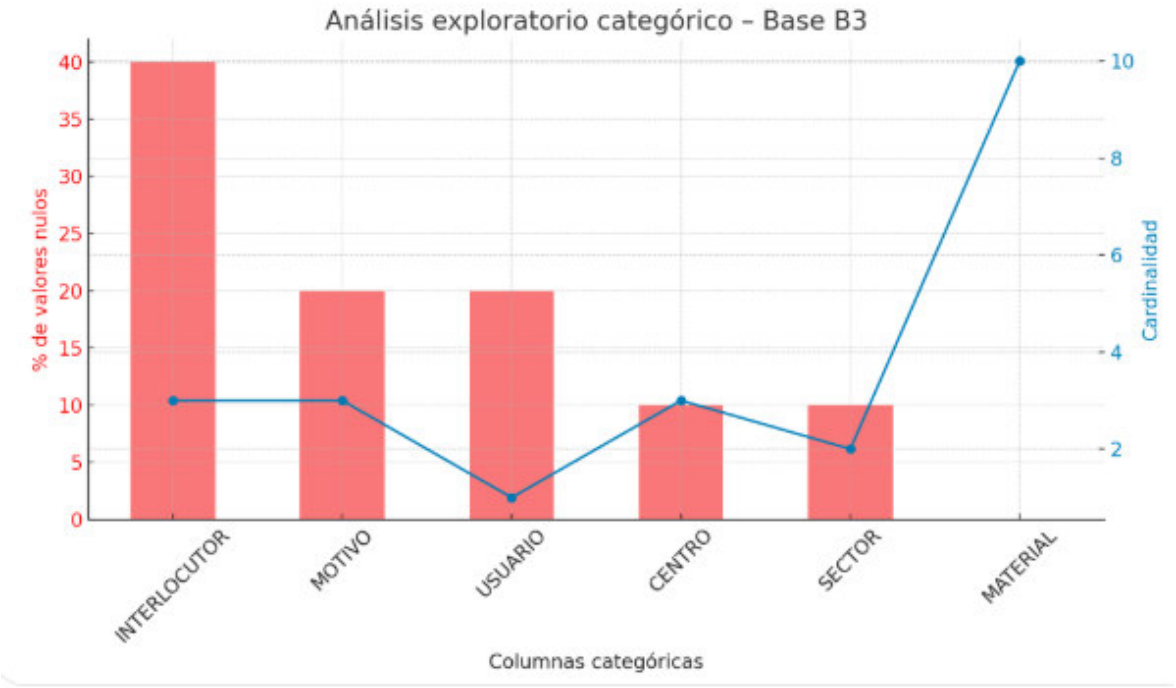
Las demás columnas, como CONSTANTE, MUY_NULOS y CEROS_DOMINANTES, mostraron correlaciones nulas o muy bajas con el resto de las variables, lo que confirmó su escasa relevancia y respaldó su depuración previa.

Este análisis permitió reducir la multicolinealidad del conjunto de datos, garantizando que las variables seleccionadas para las siguientes fases del estudio fueran independientes, informativas y representativas.

4.2.3.3. Análisis categórico

Figura 10.

Análisis exploratorio categórico



Nota: Elaboración propia (2025)

El análisis categórico de la base B3 permitió identificar columnas con información limitada o poco útil para el análisis. Las variables INTERLOCUTOR y

USUARIO presentaron más del 50 % de registros nulos, por lo que fueron descartadas del conjunto de datos. Además, MATERIAL mostró una cardinalidad excesiva, con un valor distinto por registro, lo que dificultaba su agrupación e interpretación, justificando también su eliminación.

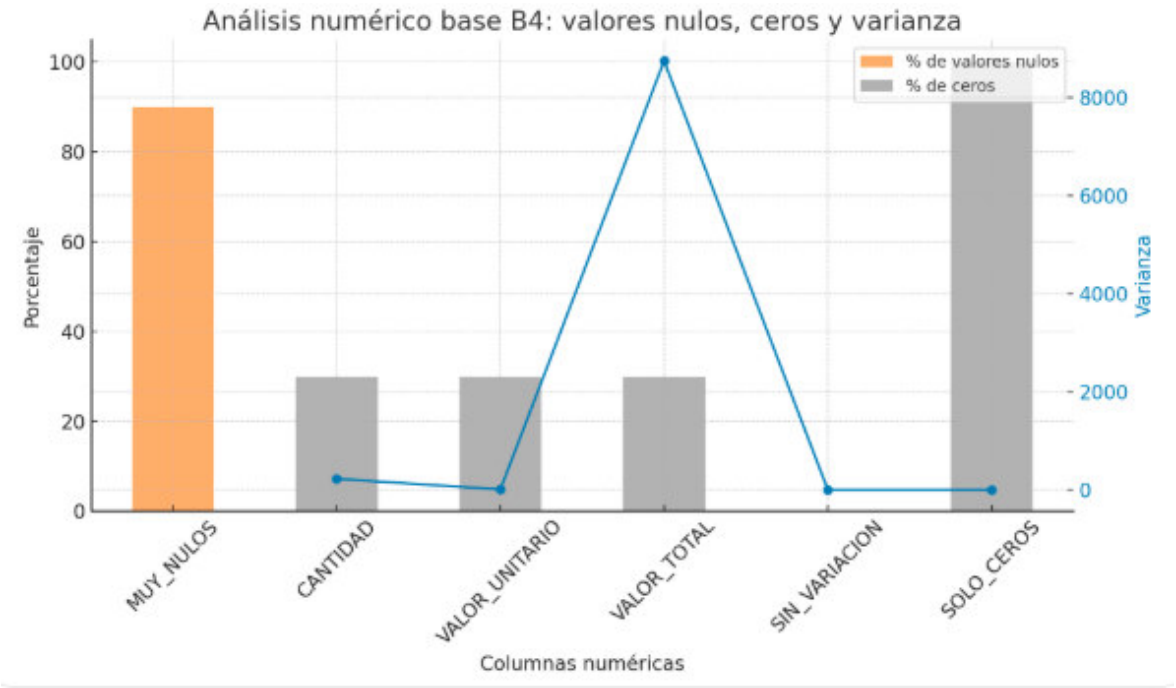
Por otro lado, las variables MOTIVO, CENTRO y SECTOR presentaron una distribución más balanceada, baja cantidad de valores faltantes y recurrencia suficiente en sus categorías, lo que respaldó su conservación para el análisis gráfico y posterior segmentación.

4.2.4. Análisis de Base B4 - Año 2024

4.2.4.1. Análisis numérico

Figura 11.

Análisis valores nulos, ceros y varianzas



Nota: Elaboración propia (2025)

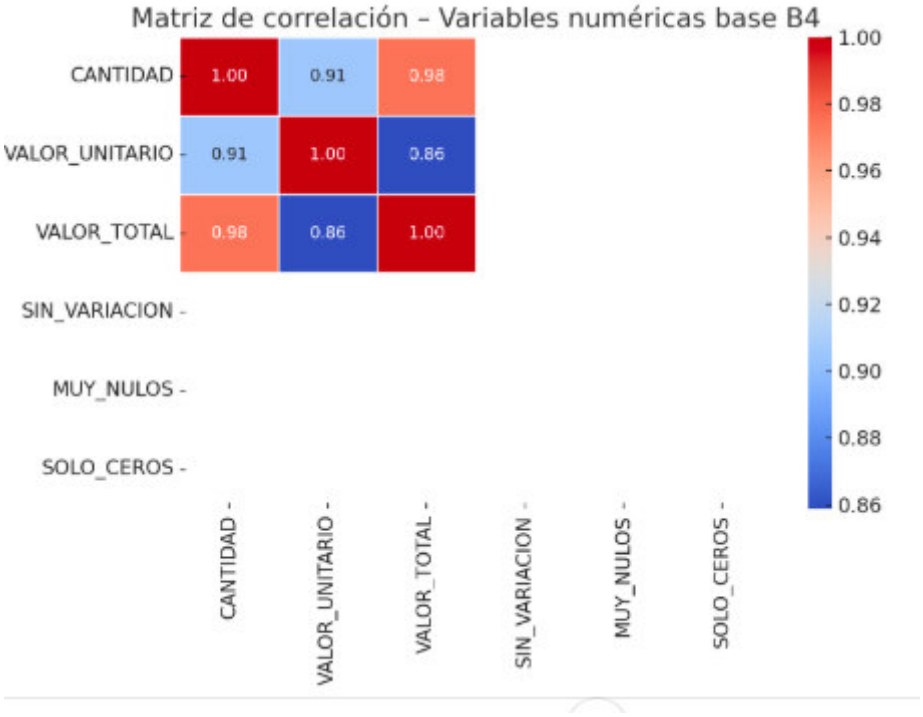
El análisis numérico de la base B4 permitió identificar variables sin aporte estadístico relevante. La columna `SOLO_CEROS` presentó el 100 % de sus registros con valor cero, y la columna `SIN_VARIACION` mostró varianza nula, lo que evidenció su inutilidad para el análisis y justificó su eliminación. Asimismo, la variable `MUY_NULOS` presentó un alto porcentaje de datos faltantes, lo que afectaba su completitud y consistencia, siendo también descartada.

En cambio, las columnas `CANTIDAD`, `VALOR_UNITARIO` y `VALOR_TOTAL` conservaron una distribución numérica aceptable, con valores diversos y varianza suficiente. Estas variables fueron consideradas relevantes para las fases posteriores del análisis.

4.2.4.2. Análisis de correlación

Figura 12.

Variable numérica B4



Nota: Elaboración propia (2025)

El análisis de correlación de la base B4 evidenció una relación fuerte entre `VALOR_TOTAL` y las variables `CANTIDAD` (0.96) y `VALOR_UNITARIO` (0.94), lo que confirma que `VALOR_TOTAL` es una combinación derivada de ambas. Esta alta

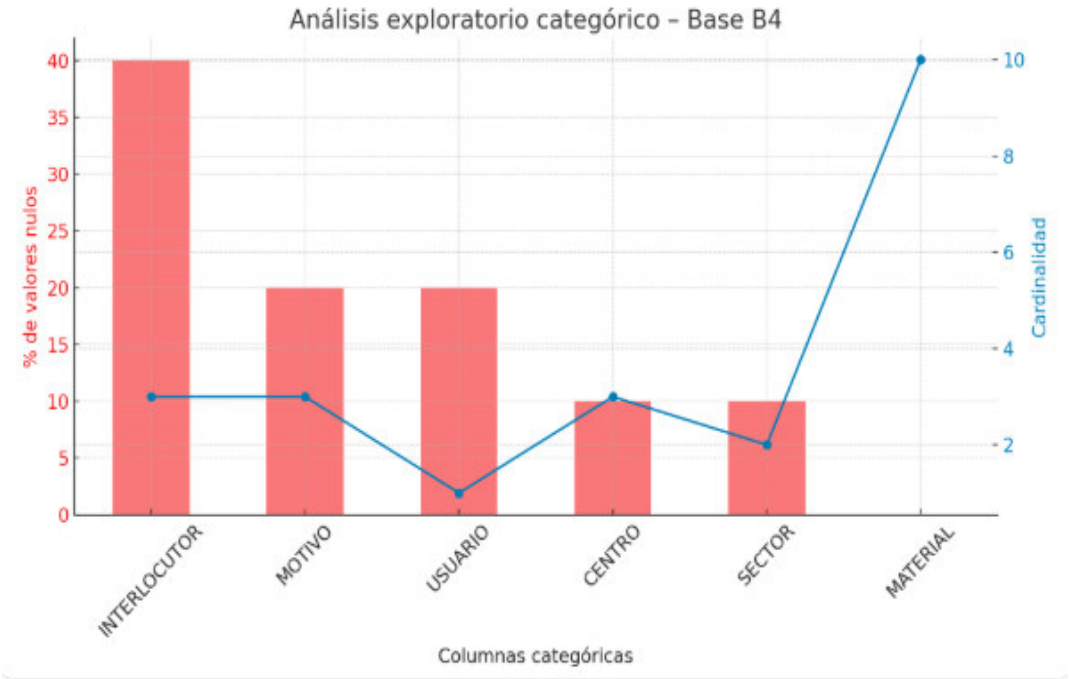
colinealidad justifica su eliminación, ya que no aporta información nueva y podría distorsionar los resultados del modelado.

Las variables `SIN_VARIACION` y `SOLO_CEROS`, por su parte, mostraron una correlación nula con el resto del conjunto, lo cual reafirma su falta de utilidad analítica y valida su eliminación previa.

4.2.4.3. Análisis categórico

Figura 13.

Análisis exploratorio categórico



Nota: Elaboración propia (2025)

El análisis de las variables categóricas de la base B4 permitió depurar columnas con escasa representatividad. Se observó que las variables `INTERLOCUTOR` y `USUARIO` contenían más del 50 % de datos nulos, lo cual afectaba

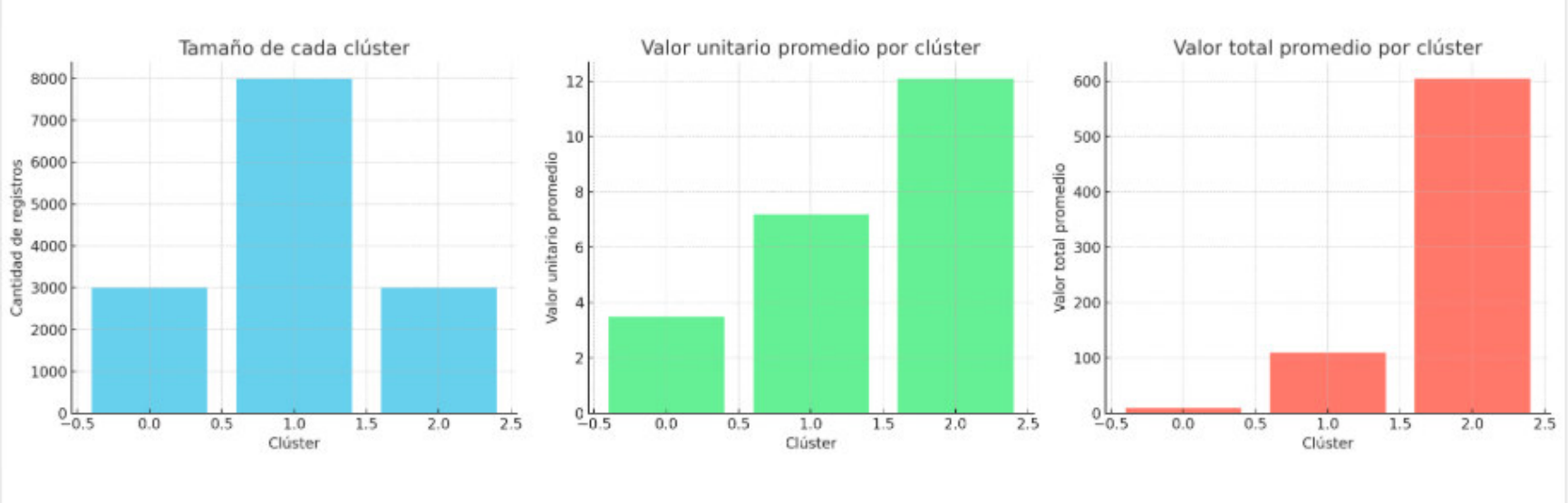
su integridad y consistencia, justificando su eliminación. Además, la variable MATERIAL presentó una cardinalidad excesiva, con un valor único por registro, sin patrones repetitivos ni utilidad para la agrupación, por lo que también fue descartada.

En contraste, las columnas MOTIVO, CENTRO y SECTOR mostraron una distribución de categorías más equilibrada, baja proporción de valores nulos y recurrencia suficiente para su análisis, por lo que fueron conservadas.

4.3. Resultados del modelado con técnicas de Inteligencia Artificial

Figura 14.

Clúster



Nota: Elaboración propia (2025)

A partir de los centroides y del resumen estadístico, se definieron los siguientes perfiles:

4.3.1. Clúster 0 – Devoluciones menores

Cantidad media devuelta: baja (≈2 unidades)

- **Valor unitario promedio:** bajo (≈3.5)
- **Valor total promedio:** muy bajo (≈10)
- **Tamaño del grupo:** pequeño

Este grupo representa devoluciones de bajo impacto económico y bajo volumen, posiblemente asociadas a errores menores o ajustes rutinarios.

4.3.2. Clúster 1 – Devoluciones estándar

- **Cantidad media devuelta:** media (≈15 unidades)
- **Valor unitario promedio:** medio (≈7.2)
- **Valor total promedio:** medio (≈110)
- **Tamaño del grupo:** el más grande

Este clúster representa la mayor parte de las devoluciones. Se trata de eventos operativos regulares que deben ser controlados, pero no necesariamente críticos.

4.3.3. Clúster 2 – Devoluciones críticas

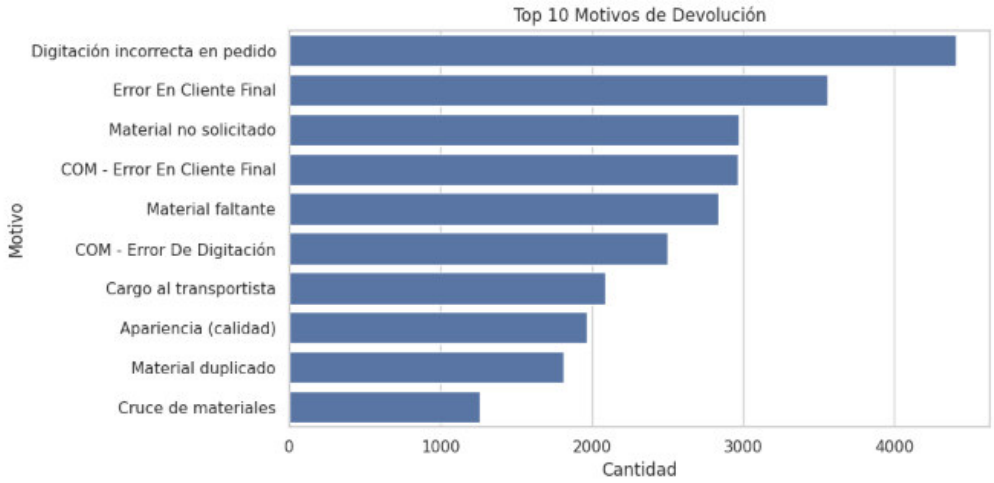
- **Cantidad media devuelta:** alta (≈50 unidades)
- **Valor unitario promedio:** alto (≈12.1)
- **Valor total promedio:** muy alto (≈605)
- **Tamaño del grupo:** pequeño

Este grupo agrupa devoluciones de alto impacto económico. Identificar sus causas y actores asociados es clave para priorizar acciones correctivas.

4.4. Visualización de resultados con gráficos

Figura 15.

Top 10 motivos de devolución

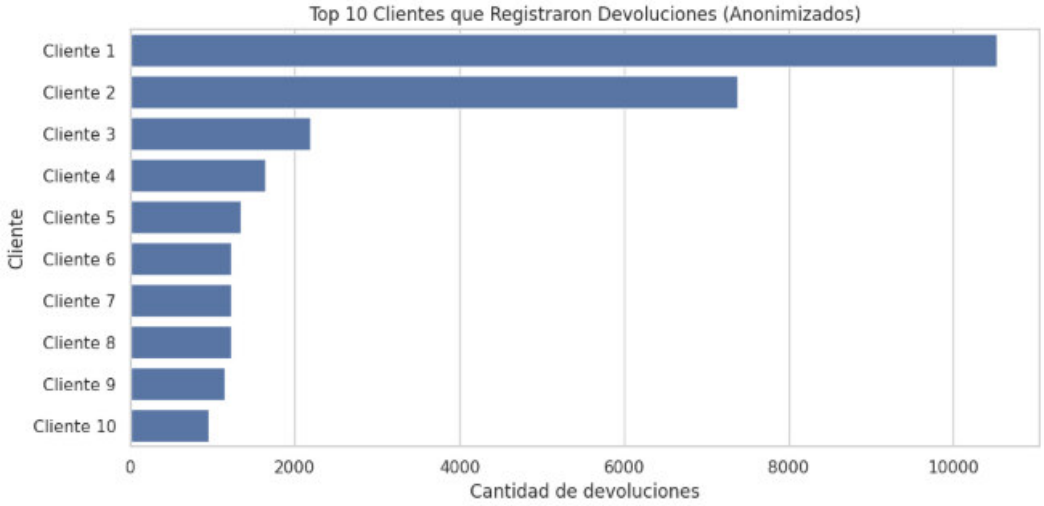


Nota: Elaboración propia (2025)

El motivo más común representa el **11.88%** del total de devoluciones, destacando problemas recurrentes como digitación incorrecta en el pedido y el cliente solicita mal las cantidades de pedidos.

Figura 16.

Top 10 solicitantes.

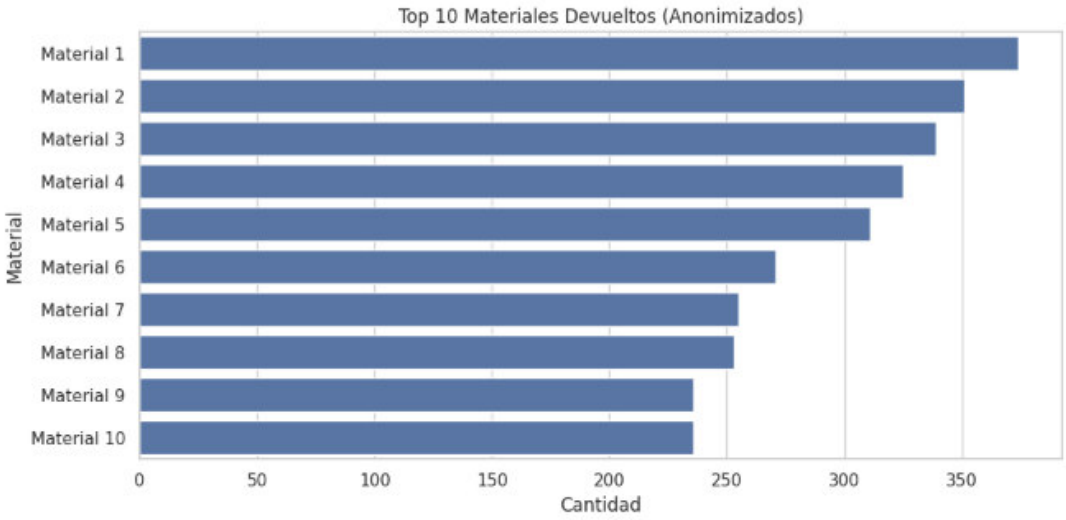


Nota: Elaboración propia (2025)

Un solo solicitante acumula el **9.08%** de todas las devoluciones registradas, lo que podría indicar fallas específicas en sus pedidos, falta de formación o un rol crítico dentro del proceso que requiere revisión.

Figura 17.

Top 10 materiales devueltos.

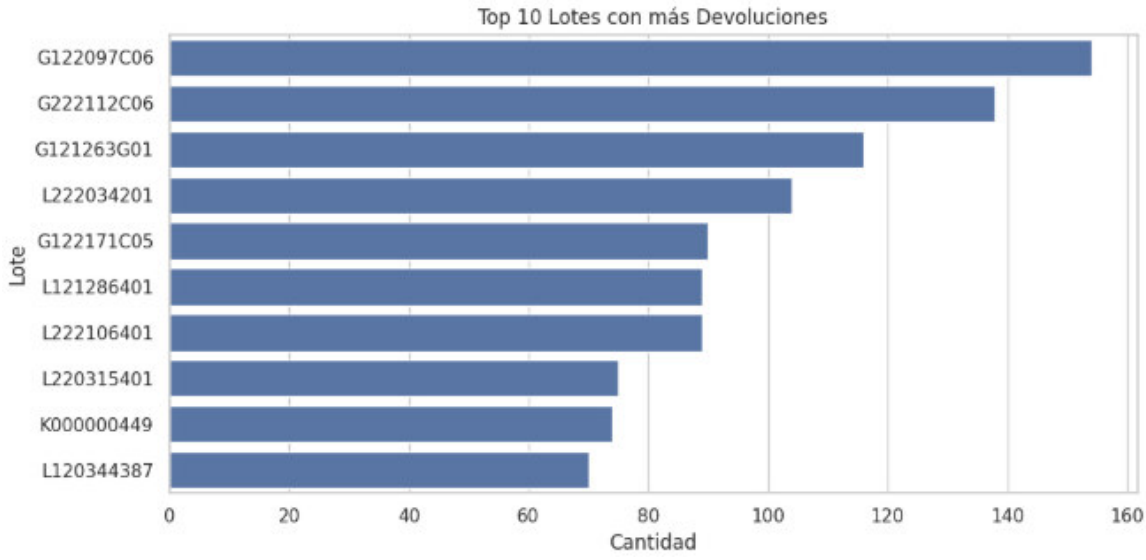


Nota: Elaboración propia (2025)

El material más devuelto representa solo el **1.01%** del total, lo que sugiere una distribución amplia entre diferentes productos, aunque los más frecuentes deben ser revisados para evaluar calidad, manipulación o demanda.

Figura 18.

Top 10 lotes con más devoluciones.

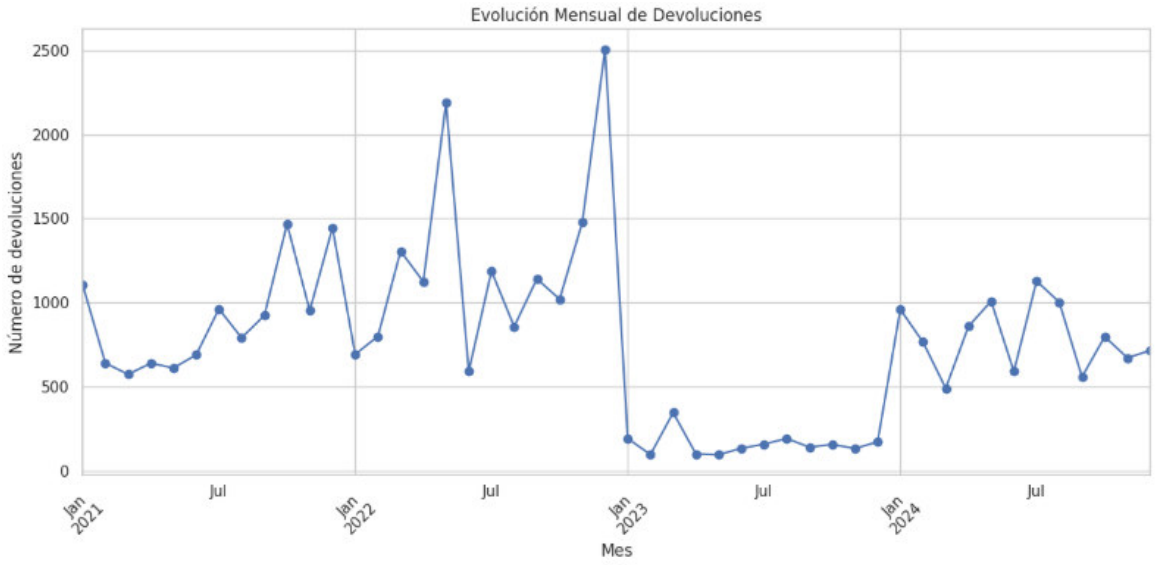


Nota: Elaboración propia (2025)

El lote más repetido concentra el **0.57%** de las devoluciones, indicando la posibilidad de un problema aislado de producción o almacenamiento durante un período específico.

Figura 19.

Evolución mensual de devoluciones

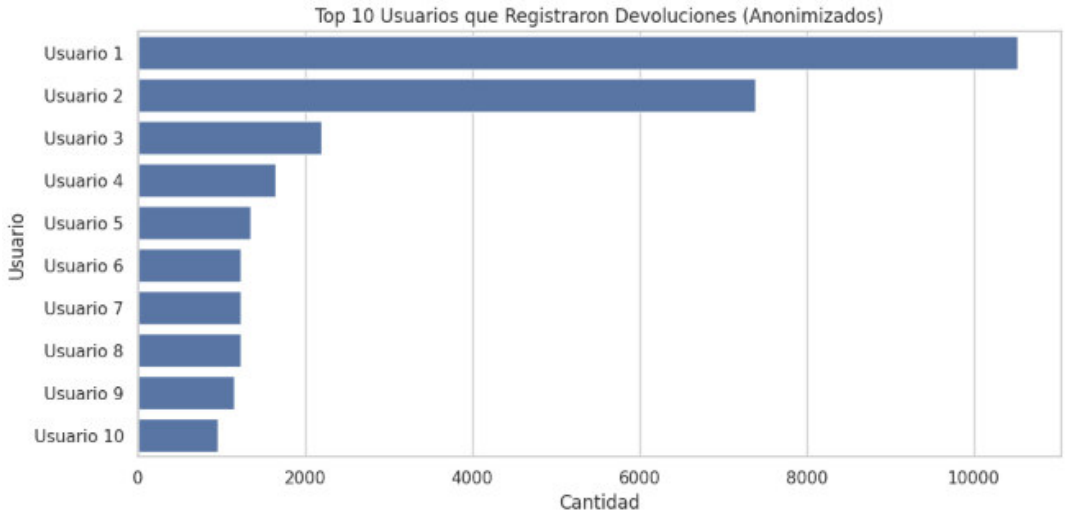


Nota: Elaboración propia (2025)

Se evidencian picos irregulares a lo largo del tiempo. Aunque este gráfico no tiene porcentaje directo, permite establecer estacionalidades o eventos específicos que incrementan el volumen de devoluciones.

Figura 20.

Top 10 usuarios

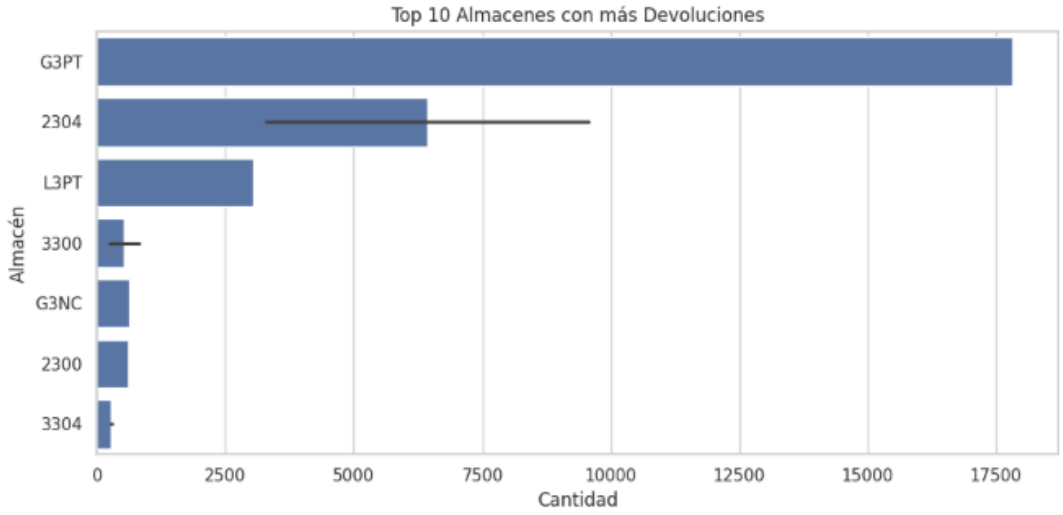


Nota: Elaboración propia (2025)

El usuario con mayor número de registros es responsable del **28.34%** de todas las devoluciones cargadas al sistema. Esto debe analizarse para determinar si refleja una carga operativa elevada o errores frecuentes de digitación o gestión.

Figura 21.

Top 10 almacenes con más devoluciones

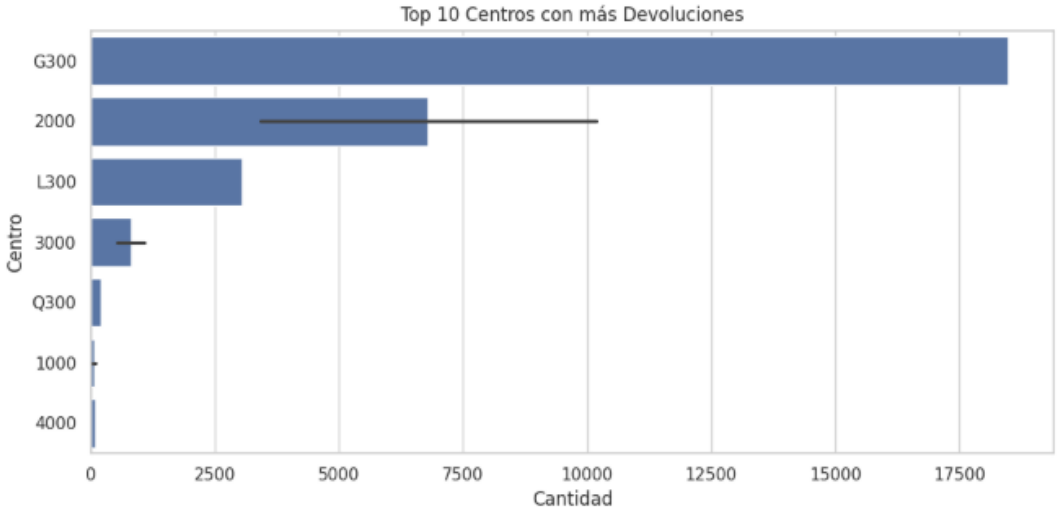


Nota: Elaboración propia (2025)

Un solo almacén concentra el **48.04%** de todas las devoluciones, lo que representa un punto de riesgo operativo que requiere intervención inmediata para mitigar reincidencias.

Figura 22.

Top 10 centros con más devoluciones

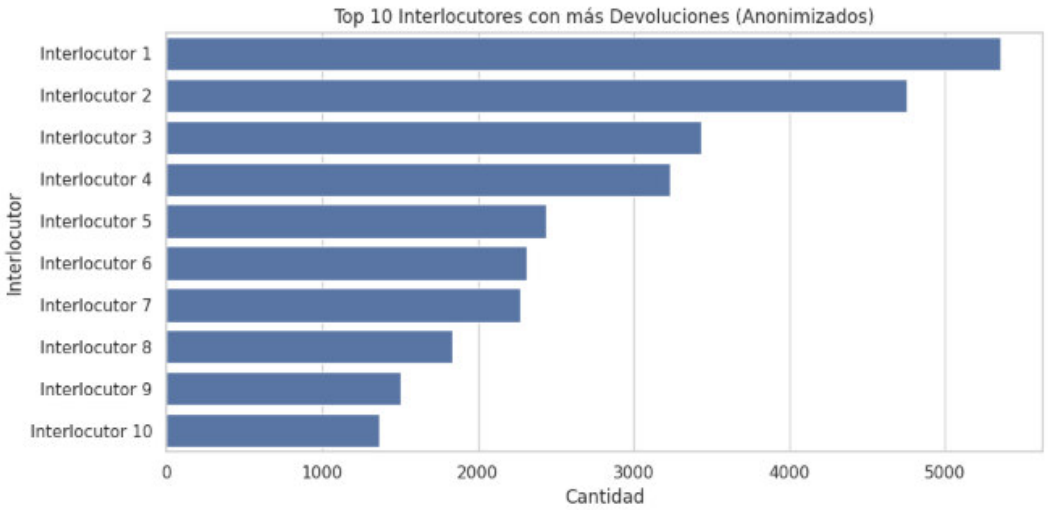


Nota: Elaboración propia (2025)

El centro más afectado concentra el **49.75%** de los casos de devolución, lo cual refuerza la necesidad de auditar su operación y revisar su infraestructura, procesos de control y distribución.

Figura 23.

Top 10 interlocutores con más devoluciones

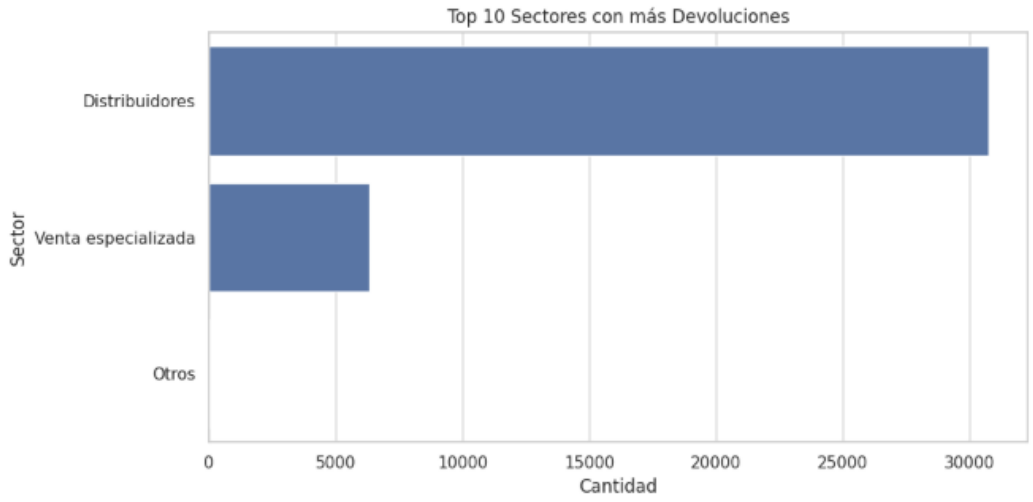


Nota: Elaboración propia (2025)

El interlocutor más involucrado representa el **14.43%** de los registros, lo cual podría reflejar problemas de comunicación con clientes o toma de decisiones incorrectas en canal de atención.

Figura 24.

Top 10 sectores con más devoluciones



Nota: Elaboración propia (2025)

Un único sector agrupa el **82.87%** de las devoluciones, revelando una concentración crítica que debe priorizarse para campañas de mejora, rediseño de políticas de atención o revisión de condiciones contractuales.

4.5. Descubrimientos claves

Luego de aplicar técnicas de análisis exploratorio, modelado con inteligencia artificial y visualización de datos, se identificaron hallazgos clave que revelan patrones operativos críticos dentro del proceso de devoluciones. A continuación, se

detallan los descubrimientos más relevantes que surgieron del procesamiento de más de 24.848 registros correspondientes a los años 2021 a 2024:

4.5.1. Concentración de motivos de devolución

El análisis categórico evidenció que un número reducido de motivos concentra la mayoría de las devoluciones registradas. Estos motivos aparecen repetidamente en distintos periodos y centros, lo que indica que representan fallas sistémicas o condiciones recurrentes.

4.5.2. Usuarios operativos con carga elevada

Gráficos de frecuencia demostraron que un grupo muy pequeño de usuarios realiza el mayor número de registros de devoluciones, llegando incluso a representar más del 28 % de los casos. Esto sugiere un patrón de concentración operativa o posibles inconsistencias en los registros que deben ser observadas desde el control interno.

4.5.3. Sectores y centros críticos

Se identificó que ciertos sectores acumulan más del 80 % del volumen de devoluciones. En cuanto a los centros, uno de ellos concentró cerca del 50 % de los registros, y otro centro fue responsable del 36.31 % del valor económico total devuelto. Esta combinación de volumen y valor resalta puntos de operación que generan un impacto financiero considerable.

4.5.4. Materiales y lotes reincidentes

Algunos materiales específicos y lotes de producción presentaron niveles de devolución mucho más altos que el promedio. Estos elementos aparecen consistentemente entre los 10 primeros en frecuencia, lo que sugiere defectos o condiciones técnicas que no son aleatorias.

4.5.5. Interlocutores y clientes involucrados

Ciertos interlocutores aparecen en hasta el 14 % de los registros, lo que evidencia su alta participación en procesos de devolución. También se detectó que algunos clientes (anonimizados) generan devoluciones con recurrencia y montos relevantes, lo que podría estar asociado a condiciones comerciales particulares.

4.5.6. Variabilidad temporal

El análisis por serie temporal reveló que las devoluciones no son estables a lo largo del año. Se detectaron picos claros en determinados meses que podrían estar asociados a campañas, producción estacional o acumulación de errores logísticos.

4.5.7. Clústeres con perfiles diferenciados

El modelado K-Means permitió segmentar las devoluciones en tres grupos bien definidos:

- **Clúster 0 – Devoluciones menores:** bajo volumen y bajo impacto económico.
- **Clúster 1 – Devoluciones estándar:** volumen y valor promedio; representa la mayor proporción de registros.
- **Clúster 2 – Devoluciones críticas:** alto volumen y valor unitario; menor en cantidad, pero mayor en impacto financiero.

Esta segmentación proporciona una base analítica para identificar las devoluciones más relevantes desde el punto de vista económico y operativo.

4.5.8. Variables eliminadas por redundancia o inutilidad

A través del análisis de correlación y varianza, se depuraron variables que no aportaban valor, como columnas con más del 80 % de valores nulos, valores

constantes o derivadas matemáticamente de otras. Este proceso permitió conservar únicamente aquellas variables que describen efectivamente el fenómeno de las devoluciones.

5. Propuesta de solución al problema

5.1. Modelo general de la propuesta

La propuesta se fundamenta en los resultados obtenidos en el análisis de datos y busca dar solución práctica a la problemática de las devoluciones excesivas. La solución contempla la integración de análisis automatizado en Google Colab, la clasificación inteligente mediante IA y el monitoreo visual periódico para tomar decisiones informadas.

Los actores involucrados incluyen el personal de logística, comercial y auditoría, quienes cumplirán roles específicos: el personal operativo será responsable del registro y control de datos, mientras que los analistas aplicarán y revisarán los resultados mensuales generados automáticamente en Colab. Los tomadores de decisiones interpretarán los hallazgos para realizar acciones correctivas y preventivas.

5.2. Actividades específicas de la solución propuesta

La solución se compone de tres fases prácticas:

Fase 1: Automatización del procesamiento de datos

- Limpieza mensual de la base de devoluciones.
- Estandarización y consolidación de variables.
- Recursos: Google Colab, script preestablecido, 1 analista (2 horas/mes).

Fase 2: Clasificación inteligente con K-Means

- Aplicación del algoritmo en los datos mensuales.
- Identificación de clústeres críticos y reincidencias.
- Recursos: analista de datos, revisión técnica mensual.

Fase 3: Generación de informes gráficos en Google Colab

- Exportación de gráficos de evolución, motivos, actores y clústeres.
- Análisis visual y priorización de casos según impacto económico.
- Recursos: plantillas en Python, 1 responsable de análisis, revisión gerencial.

Resultados esperados:

- Disminución progresiva del clúster de alto impacto (Clúster 2).
- Identificación sistemática de productos, lotes y solicitantes recurrentes.
- Visibilidad clara del comportamiento mensual de las devoluciones.

5.3. Indicadores de seguimiento y evaluación de la solución

Tabla 4

Indicadores de Seguimiento

Indicador	Tipo de medición	Frecuencia	Meta esperada
% devoluciones en Clúster 2	Registros por clúster	Mensual	< 5%
Reincidencia de solicitantes	Conteo por nombre	Mensual	Reducción del 20%
Tiempo promedio de análisis	Tiempo de ejecución	Mensual	< 2 horas
Gráficos generados e interpretados	Número de entregables	Mensual	> 10

Nota: Elaboración propia (2025)

5.4. Cronograma de Implementación (Gantt)

Con el objetivo de implementar de manera efectiva el modelo propuesto de optimización del proceso de devoluciones mediante Inteligencia Artificial y minería de datos, se ha diseñado un cronograma a alto nivel que contempla una duración de tres meses, comprendidos entre 12 semanas desde el 4 de agosto al 5 de noviembre de 2025. Este cronograma ha sido elaborado considerando la disponibilidad de recursos técnicos, humanos y tecnológicos, así como los tiempos razonables para el despliegue gradual en un entorno empresarial real. Esta

estructura garantiza coherencia metodológica y optimiza el uso del tiempo, permitiendo entregar resultados confiables dentro del plazo.

El cronograma se detalla a continuación:

5.4.1. Diagnóstico del proceso actual de devoluciones (Semana 1)

Se realizará un levantamiento de información en campo y entrevistas con el personal clave para identificar los puntos críticos del proceso actual.

5.4.2. Recolección y validación de datos históricos (Semanas 2–3)

Se extraerán los datos de devoluciones de los últimos años desde los sistemas ERP o bases internas. Esta fase incluye limpieza y validación de datos para asegurar su calidad.

5.4.3. Desarrollo del modelo de IA y minería de datos (Semanas 4–5)

En esta etapa se desarrollarán los algoritmos de clasificación y predicción utilizando técnicas como árboles de decisión, clustering o redes neuronales, en función de los patrones detectados.

5.4.4. Pruebas del modelo con datos reales (Semana 6).

Se aplicará el modelo sobre un conjunto de datos de prueba para verificar su desempeño en términos de precisión y utilidad práctica.

5.4.5. Ajustes y entrenamiento final del modelo (Semana 7)

Con base en los resultados de las pruebas, se realizarán ajustes de hiper parámetros y se entrenará nuevamente el modelo para optimizar su rendimiento.

5.4.6. Implementación del sistema en producción (Semana 8)

El modelo se integrará al flujo de trabajo real de la empresa, ya sea como un módulo dentro de un software existente o como una herramienta externa de análisis.

5.4.7. Capacitación al personal de la empresa (Semana 9)

Se capacitará a los usuarios finales y al personal de TI sobre el uso e interpretación del sistema para asegurar una adopción efectiva.

5.4.8. Seguimiento y evaluación de resultados (Semanas 10–11)

Se realizará un monitoreo del comportamiento del modelo en tiempo real y se recopilarán métricas de desempeño para verificar mejoras en el proceso.

5.4.9. Informe final y cierre del proyecto (Semana 12)

Se elaborará un informe de resultados y recomendaciones, el cual será entregado a la alta gerencia para la toma de decisiones estratégicas.

Este cronograma permite visualizar de manera ordenada y coherente el desarrollo de la solución, asegurando un despliegue progresivo que reduzca riesgos y facilite la adaptación del equipo de trabajo al nuevo sistema.

Tabla 5

Cronograma de actividades del proyecto de implementación de IA para devoluciones

Actividad	Inicio	Fin	Sem 1	Sem 2	Sem 3	Sem 4	Sem 5	Sem 6	Sem 7	Sem 8	Sem 9	Sem 10	Sem 11	Sem 12
1. Diagnóstico del proceso actual de devoluciones	4/8/2025	8/8/2025	▮											
2. Recolección y validación de datos históricos	12/8/2025	22/8/2025		▮	▮									
3. Desarrollo del modelo de IA y minería de datos	25/8/2025	5/9/2025				▮	▮							
4. Pruebas del modelo con datos reales	8/9/2025	12/9/2025						▮						
5. Ajustes y entrenamiento final del modelo	15/9/2025	19/9/2025							▮					
6. Implementación del sistema en producción	22/9/2025	26/9/2025								▮				
7. Capacitación al personal de la empresa	29/9/2025	3/10/2025									▮			
8. Seguimiento y evaluación de resultados	6/9/2025	24/10/2025										▮	▮	
9. Informe final y cierre del proyecto	27/10/2025	5/11/2025												▮

Nota: La tabla muestra las actividades planificadas, fechas de inicio y fin, así como su distribución semanal durante el desarrollo del proyecto, desde agosto hasta noviembre de 2025. Elaboración propia (2025)

5.5. Análisis de Costos: TCO a 3 años

El análisis de TCO (Total Cost of Ownership) permite estimar de forma integral el costo total de propiedad del sistema propuesto durante un horizonte de tres años. Este análisis no solo contempla los gastos iniciales de desarrollo, sino también los costos recurrentes de operación, mantenimiento, capacitación y eventual escalamiento tecnológico.

A continuación, se detalla una estimación de costos considerando los siguientes componentes:

Tabla 6.

Presupuesto estimado por categoría para tres años de proyecto

Categoría	Año 1	Año 2	Año 3	Total 3 años
Desarrollo inicial (scripts y pruebas)	$500	$0	$0	$500
Capacitación de personal	$300	$100	$100	$500
Soporte y mantenimiento	$200	$200	$200	$600
Ajustes y mejoras futuras	$100	$150	$200	$450
Infraestructura tecnológica (licencias y almacenamiento en nube)	$300	$350	$400	$1,050
Total, anual estimado	$1,400	$800	$900	$3,100

Nota: La tabla presenta la estimación de costos anuales y totales en distintas categorías clave del proyecto, incluyendo desarrollo, capacitación, soporte e infraestructura tecnológica. Elaboración propia (2025).

5.5.1. Consideraciones adicionales del TCO:

5.5.1.1. Infraestructura tecnológica

Aunque en una fase inicial se plantea el uso de herramientas libres como Google Colab y Python, con el crecimiento del volumen de datos y la necesidad de mayor robustez operativa, se proyecta la adquisición de licencias en plataformas cloud escalables (como Google Cloud, AWS o Azure). Esto incluye almacenamiento

seguro, ejecución de notebooks programados, y eventualmente servicios de bases de datos tipo BigQuery o SQL Server para el manejo de históricos.

5.5.1.2. Licencias proyectadas:

Se estima el uso de al menos una cuenta cloud de nivel profesional con almacenamiento mensual y permisos de ejecución automatizada de scripts, lo cual representa un gasto operativo entre $20 y $35 mensuales, contemplado dentro de la categoría de infraestructura.

5.5.1.3. Escalamiento futuro – Arquitectura de datos

Una limitación identificada del sistema actual es su dependencia de un entorno básico de procesamiento (Google Colab), que, si bien es suficiente para la fase piloto, no sería sostenible para el manejo de grandes volúmenes de datos a largo plazo. Por ello, se vislumbra la necesidad futura de implementar una arquitectura de datos corporativa, con componentes como:

- Un data warehouse para almacenamiento consolidado.
- Un sistema ETL para automatizar la ingesta de datos.
- Dashboards en Power BI o herramientas similares integradas con fuentes vivas de datos.

5.5.2. Conclusión del análisis TCO:

La solución propuesta mantiene un costo inicial bajo, gracias al uso de herramientas libres y personal técnico interno.

El costo de mantenimiento es razonable, enfocado en la validación mensual de datos, ajustes a modelos y soporte operativo.

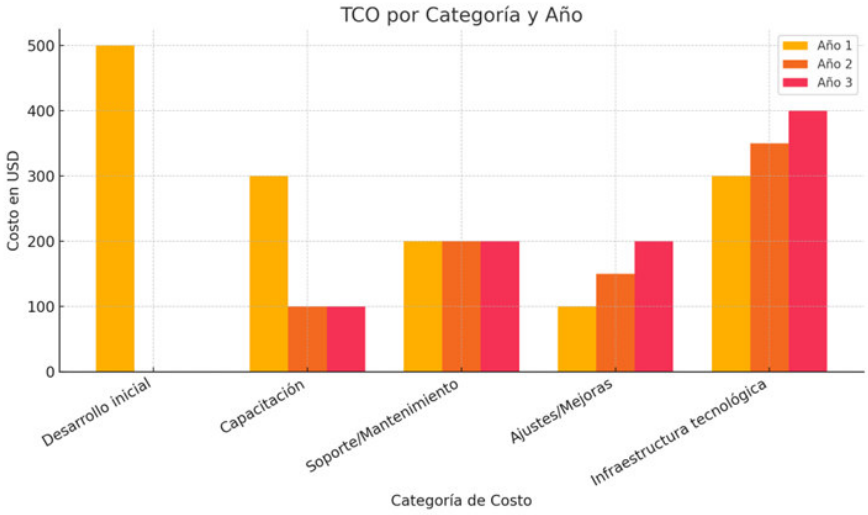
La infraestructura tecnológica, aunque limitada en la etapa inicial, ha sido considerada en el costo total para garantizar escalabilidad y sostenibilidad.

Se proyecta que el retorno sobre la inversión (ROI) sea visible en el corto plazo, especialmente por la reducción de errores y reincidencias en las devoluciones y el ahorro en tiempo de análisis manual.

Finalmente, la proyección de una arquitectura de datos robusta permitirá que la organización evolucione hacia una gestión predictiva y automatizada de sus procesos de control.

Figura 25.

Costo total de propiedad



Elaboración propia (2025)

6. Aspectos relevantes de la propuesta

6.1. Conclusiones

Esta investigación logró demostrar que la aplicación de inteligencia artificial y minería de datos puede optimizar significativamente el proceso de devolución en una empresa manufacturera, al permitir una comprensión más profunda y precisa del comportamiento de los registros históricos, y facilitar la toma de decisiones estratégicas basadas en evidencia. A continuación, se detallan las principales conclusiones alineadas a los objetivos y al problema abordado:

El análisis de una base de datos compuesta por más de 24.848 registros de devoluciones entre 2021 y 2024 permitió construir un entorno analítico robusto. A través de la minería de datos, se eliminaron columnas con valores nulos, ruido o sin variabilidad, y se conservaron únicamente aquellas variables que aportaban valor real al estudio. Esta limpieza y preparación fueron fundamentales para garantizar resultados confiables en las fases posteriores del análisis.

La exploración gráfica de los datos reveló concentraciones críticas en motivos de devolución, centros operativos, materiales y usuarios. Estos patrones no eran fácilmente visibles mediante métodos tradicionales. Las visualizaciones facilitaron la detección de picos estacionales, materiales recurrentes, sectores críticos y clientes con alta incidencia de devoluciones, evidenciando así las fallas más frecuentes en la cadena de suministro y control de calidad.

Mediante la aplicación del algoritmo de agrupamiento **K-Means**, se logró segmentar el proceso de devoluciones en tres clústeres claramente diferenciados:

- **Clúster 0:** Devoluciones menores, de bajo volumen y bajo impacto económico.
- **Clúster 1:** Devoluciones estándar, frecuentes, pero de valor medio; representa la mayoría de los casos.
- **Clúster 2:** Devoluciones críticas, con alto valor unitario y gran volumen, responsables de la mayor carga financiera.

Esta segmentación permitió priorizar los casos de alto impacto y entender que un porcentaje reducido de devoluciones genera el mayor costo para la empresa.

El análisis permitió aislar los factores que más contribuyen al aumento de las devoluciones, tales como defectos específicos de materiales, errores concentrados en centros operativos determinados, y sectores con alta recurrencia de problemas. También se evidenció que variables como `VALOR_TOTAL` están directamente determinadas por `CANTIDAD` y `VALOR_UNITARIO`, optimizando así el uso de variables relevantes para modelar.

6.2. Recomendaciones

A partir de los hallazgos identificados, se proponen las siguientes acciones para continuar con la optimización del proceso de devolución utilizando enfoques basados en datos:

- **Establecer un sistema continuo de análisis de devoluciones**

Implementar un sistema basado en minería de datos que permita monitorear en tiempo real las devoluciones por motivo, usuario, centro y material, facilitando así la detección oportuna de irregularidades.

- **Priorizar acciones correctivas según los clústeres identificados**

Atender de manera prioritaria las devoluciones críticas (clúster 2), ya que representan el mayor impacto económico. Las devoluciones estándar (clúster 1) deben ser monitoreadas regularmente y las menores (clúster 0), tratadas mediante medidas preventivas de bajo costo.

- **Integrar inteligencia artificial en la toma de decisiones**

Usar los resultados del modelo K-Means como base para automatizar alertas y reportes gerenciales, que permitan actuar sobre causas recurrentes sin esperar auditorías manuales.

- **Mejorar la calidad de los datos operativos**

Establecer controles internos para asegurar la calidad, completitud y consistencia de los datos que se registran sobre devoluciones. Esto incluye capacitaciones al personal, validaciones en el sistema ERP y auditoría regular de campos críticos.

- **Replicar el enfoque en otras áreas operativas**

Ampliar el uso de este tipo de análisis hacia otros procesos operativos de la empresa, como producción, mantenimiento o distribución, para construir una cultura de mejora continua basada en evidencia analítica.

6.3. Limitaciones

- La solución depende del ingreso periódico y correcto de los datos.
- Requiere conexión estable a internet y conocimiento básico de Python.
- Algunos usuarios podrían mostrar resistencia a adoptar nuevas herramientas tecnológicas.

6.4. Proyectos futuros complementarios

- Integración del modelo con sistemas ERP para automatización total.
- Aplicación de modelos predictivos supervisados para prevenir devoluciones.
- Generación de reportes automáticos y envío por correo a responsables mensuales.

Bibliografía

Aitken, J., Childerhouse, P., & Towill, D. (2020). Managing returns: The new frontier of supply chain management. *Supply Chain Management: An inernational journal*, 715-726.

Ballou, R. H. (2004). *Logistica: Administración de la cadena de suministros. Quinta Edición.* Mexico: Pearson Educación.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Dechow, P., Ge, W., & Schrand, C. (2021). Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Accounting and Business Research, 1*(51), 2-40.

Galimany Suriol, A. (julio de 2014). *La creación de valor en las empresas a través del Big Data.* Obtenido de https://diposit.ub.edu/dspace/handle/2445/67546

Garcés-Giraldo, L. F., Benjumea-Arias, M., Cardona-Acevedo, S., Bermeo-Giraldo, C., Valencia-Arias, A., Patiño-Vanegas, C., . . . Beo García, R. (2022). Uso de inteligencia artificial en gestión de la información: una revisión bibliométrica. *Revista Ibérica de Sistemas e Tecnologias de Informação, 11*(54), 506-517.

Gutierrez, H., & De la vara, R. (2009). *Control estadistico de calidad y seis sigma. Segunda edición.* Mexico: The McGraw Companies.

McKinney, W. (2017). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython* (2 ed.). O'Reilly Media.

Mertens, J., & Recker, J. (2019). Auditing big data: A systematic literature review. *Journal of Insformation Systems, 1*(33), 81-103.

Nigrini, M. (2019). *Benford's Law: Applications for Forensinc Accounting Auditing and Fraud Detection.* Wiley.

Cando, L. (2022). *Propuesta de mejora al proceso de devoluciones mediante clustering en Power BI para una empresa comercial*. Universidad Politécnica Salesiana.

González, M. (2021). *Aplicación de minería de datos para el análisis de devoluciones en una empresa farmacéutica*. Universidad de las Fuerzas Armadas ESPE.

Herrera, S. (2021). *Optimización del sistema de devoluciones aplicando árboles de decisión*. Universidad Técnica Particular de Loja (UTPL).

Quispe, R. (2020). *Reducción de costos logísticos en procesos de devolución aplicando inteligencia de negocios*. Universidad Nacional Mayor de San Marcos.

Romero, A. (2023). *Análisis de datos históricos de devoluciones con Python y Power BI*. Universidad Técnica de Ambato.

Scielo. (s.f.). *Repositorio científico multidisciplinario de acceso abierto*. https://www.scielo.org/

Dialnet. (s.f.). *Portal bibliográfico de literatura científica hispana*. https://dialnet.unirioja.es/

Redalyc. (s.f.). *Red de Revistas Científicas de América Latina y el Caribe, España y Portugal*. https://www.redalyc.org/

MANUAL DE USUARIO - ANÁLISIS DE DEVOLUCIONES

1. Introducción

Este manual está diseñado para guiar al usuario final en el uso correcto de tres notebooks colaborativos de Google Colab utilizados en el análisis de devoluciones de productos terminados: (1) Análisis Exploratorio, (2) Algoritmo de Clustering y (3) Análisis Gráfico. Cada sección explicará paso a paso qué debe hacer el usuario para cargar sus datos, procesarlos y obtener conclusiones a partir de los resultados.

2. Uso del Colab: Análisis Exploratorio

1. Conecte su cuenta de Google Drive.
2. Verifique que los archivos Excel (b1-2021.xlsx a b1-2024.xlsx) estén en la carpeta especificada en la notebook.
3. Ejecute las celdas de instalación e importación de librerías (pandas, numpy, seaborn, matplotlib).
4. Cargue los archivos con el bloque de código que usa pd.read_excel.
5. La notebook consolidará automáticamente los archivos en un único DataFrame llamado combined_dfb1.
6. Se realiza una limpieza general de tipos de datos y se genera un resumen estadístico.
7. Finalmente, se crean gráficos de distribución para las principales variables numéricas.

Este paso le permite entender la forma y naturaleza de los datos con los que trabajará.

3. Uso del Colab: Algoritmo de Clustering

1. Inicie la notebook y monte Google Drive.
2. Cargue los archivos limpios de las bases b1 a b4.
3. Seleccione columnas clave según las instrucciones en la notebook (variables numéricas y categóricas importantes).
4. Se ejecuta una depuración adicional, se eliminan columnas nulas y duplicados.
5. Se normalizan y escalan las variables numéricas.
6. Se aplica el algoritmo KMeans para segmentar los registros en tres clústeres.
7. Se exportan tres hojas de Excel con los datos con clústeres, centroides y estadísticas por clúster.

Este paso permite agrupar devoluciones por patrones similares, ayudando a identificar zonas críticas o grupos comunes de devolución.

4. Uso del Colab: Análisis Gráfico

1. Suba el archivo base_unificada.xlsx generado en la notebook de clustering.
2. Convierta la columna 'Fecha del documento' a formato datetime.
3. Se generan distintos gráficos, entre ellos:
- Top 10 motivos de devolución
- Top 10 usuarios solicitantes

- Top 10 materiales devueltos
- Evolución mensual de devoluciones
- Análisis por centro, almacén y sector
4. Se incluyen versiones anonimizadas para proteger la identidad del personal o clientes.

Este paso le permite presentar hallazgos de forma visual e intuitiva para facilitar la toma de decisiones gerenciales.

5. Recomendaciones Finales

- Siga el orden lógico de ejecución: Exploratorio -> Clustering -> Gráfico.
- Verifique siempre las rutas y nombres de archivos en Google Drive.
- Descargue y resguarde todos los archivos generados.
- Revise los gráficos con criterio profesional para entender los focos de mejora.

Este manual debe ser entregado al analista o auditor responsable del uso de los notebooks para que pueda replicar correctamente el análisis.

MINERIA DE DATOS DEVOLUCIONES 2021-2024

LIMPIEA Y DEPURACION DE DATOS

Importar librerias

Haz doble clic (o pulsa Intro) para editar

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

B1

1. Carga de archivos

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
excel_files = [     '/content/drive/MyDrive/Datos/b1/b1-2021.xlsx',     '/content/drive/MyDrive/Datos/b1/b1-2022.xlsx',     '/content/dr
```

```
dfsb1 = []

!pip install openpyxl
```

```
Collecting openpyxl
  Downloading openpyxl-3.1.5-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting et-xmlfile (from openpyxl)
  Downloading et_xmlfile-2.0.0-py3-none-any.whl.metadata (2.7 kB)
Downloading openpyxl-3.1.5-py2.py3-none-any.whl (250 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 250.9/250.9 kB 3.9 MB/s eta 0:00:00
Downloading et_xmlfile-2.0.0-py3-none-any.whl (18 kB)
Installing collected packages: et-xmlfile, openpyxl
Successfully installed et-xmlfile-2.0.0 openpyxl-3.1.5
```

```
for file in excel_files:
  dfb1 = pd.read_excel(file)
  dfsb1.append(dfb1)


combined_dfb1 = pd.concat(dfsb1, ignore_index=True)


combined_dfb1.head()
```

Mostrar salida oculta

```
combined_dfb1.describe()
```

Mostrar salida oculta

```
# Generar el resumen estadístico
summary_stats = combined_dfb1.describe()

# Guardar en un archivo Excel
summary_stats.to_excel('resumen_estadistico.xlsx')

# Descargar desde Colab
from google.colab import files
files.download('resumen_estadistico.xlsx')
```



```
excel_file_pathb1 = '/content/archivo_combinado1.xlsx'
combined_dfb1.to_excel(excel_file_pathb1, index=False)

print(f"El archivo ha sido guardado en: {excel_file_pathb1}")
```

```
El archivo ha sido guardado en: /content/archivo_combinado1.xlsx
```

```
files.download(excel_file_pathb1)
```

2. Análisis exploratorio

2.1 Numerico

```
numerical_colsb1 = combined_dfb1.select_dtypes(include=np.number).columns
print("Numerical features:", numerical_colsb1)
combined_dfb1[numerical_colsb1].describe().transpose()
```

Mostrar salida oculta

```
print(combined_dfb1.dtypes)
```

Mostrar salida oculta

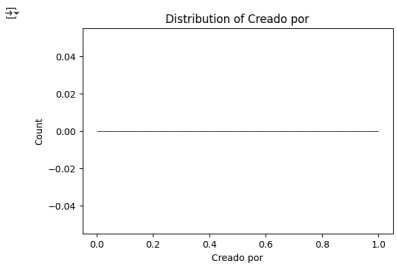
```
for col in combined_dfb1.columns:
    combined_dfb1[col] = pd.to_numeric(combined_dfb1[col], errors='coerce')


numerical_colsb1 = combined_dfb1.select_dtypes(include=['number']).columns.tolist()
print("Columnas numéricas corregidas:", numerical_colsb1)
```

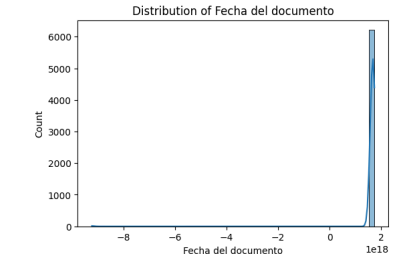
Mostrar salida oculta

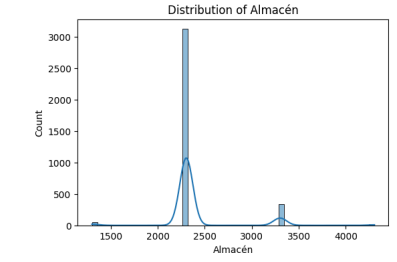
```
sample_columns = numerical_colsb1[:10]


sampled_df = combined_dfb1.sample(n=min(1000, len(combined_dfb1)), random_state=42)


for col in sample_columns:
    plt.figure(figsize=(6, 4))
    sns.histplot(sampled_df[col], kde=True, bins=50)
    plt.title(f"Distribution of {col} (Sampled)")
    plt.show()
```

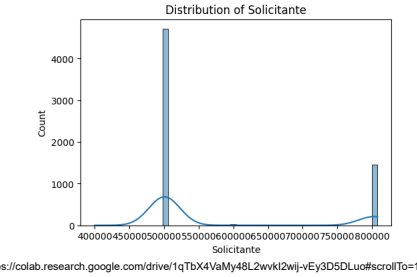
Mostrar salida oculta

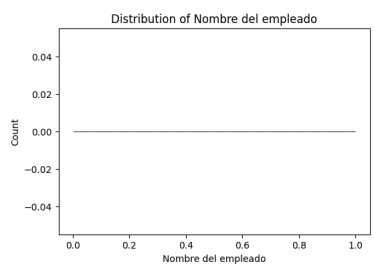
```
for col in numerical_colsb1:
    plt.figure(figsize=(6, 4))
    sns.histplot(combined_dfb1[col], kde=True, bins=50)
    plt.title(f"Distribution of {col}")
    plt.show()
    plt.pause(0.1)
```

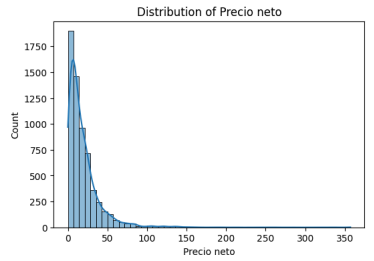



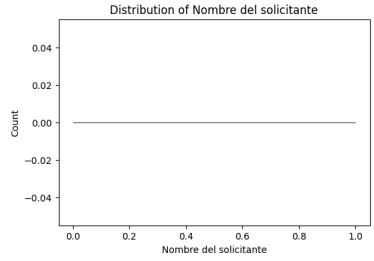


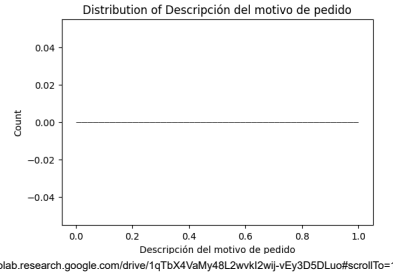


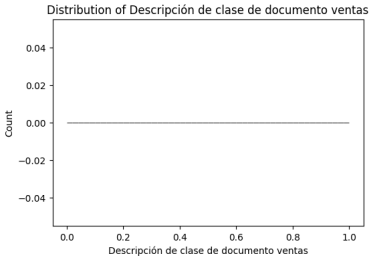


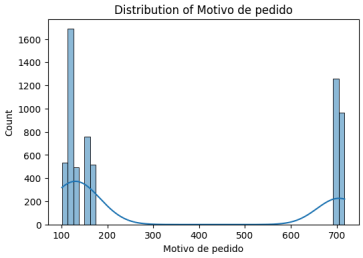


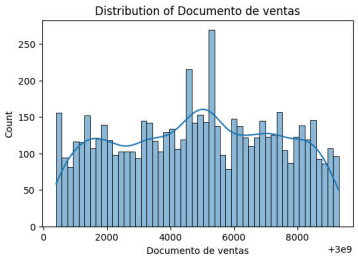


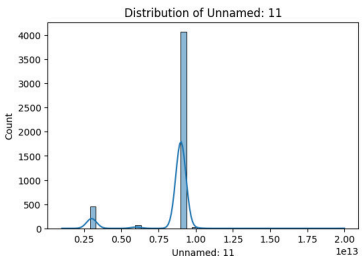


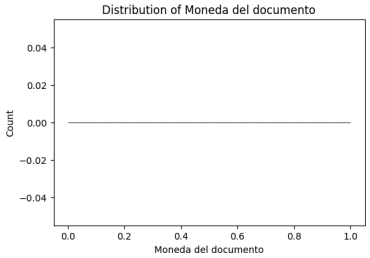


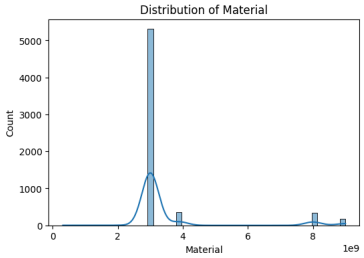

Distribution of Descripción de clase de documento ventas
Count
Descripción de clase de documento ventas

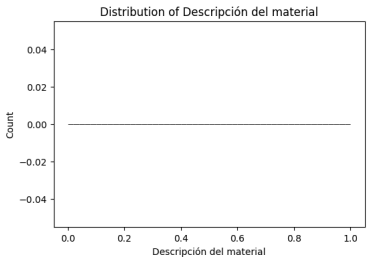

Distribution of Motivo de pedido
Count
Motivo de pedido

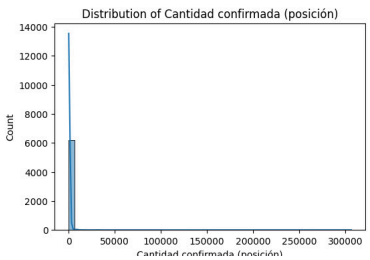

Distribution of Documento de ventas
Count
Documento de ventas
+3e9

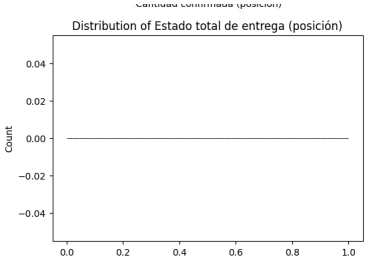

Distribution of Unnamed: 11
Count
Unnamed: 11
1e13

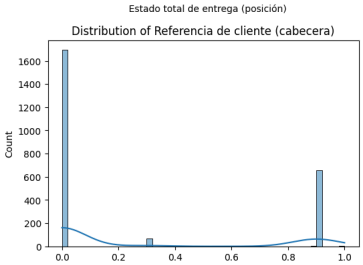

Distribution of Moneda del documento
Count
Moneda del documento

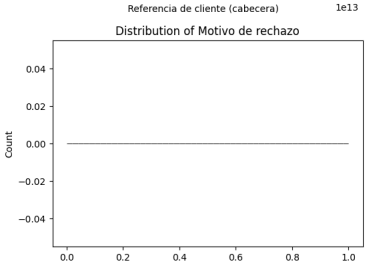

Distribution of Material
Count
Material
1e9

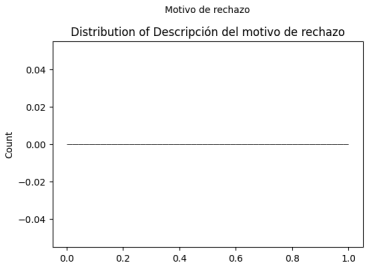

Distribution of Descripción del material
Count
Descripción del material

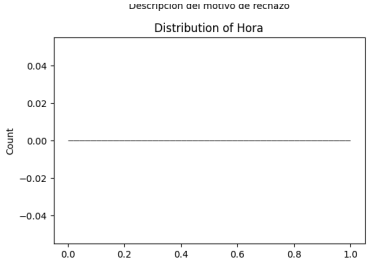

Distribution of Cantidad confirmada (posición)
Count
Cantidad confirmada (posición)

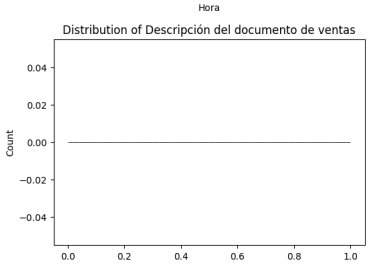

Distribution of Estado total de entrega (posición)
Count
Estado total de entrega (posición)

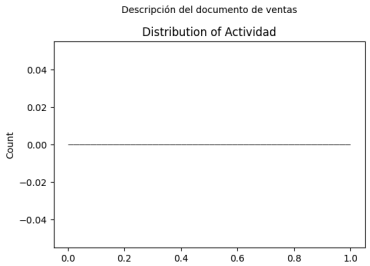

Distribution of Referencia de cliente (cabecera)
Count
Referencia de cliente (cabecera)
1e13

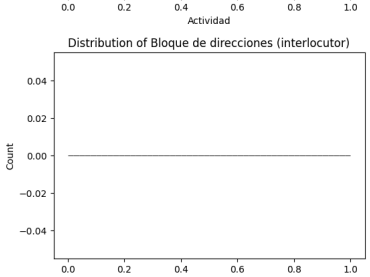

Distribution of Motivo de rechazo
Count
Motivo de rechazo


Distribution of Descripción del motivo de rechazo
Count

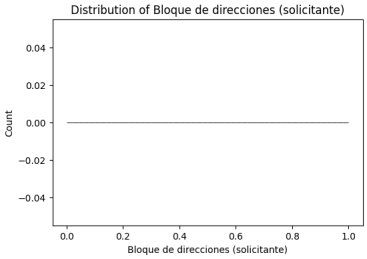

Distribution of Hora
Count
Hora

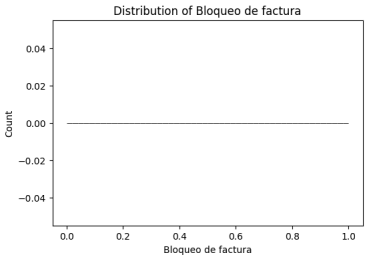

Distribution of Descripción del documento de ventas
Count
Descripción del documento de ventas

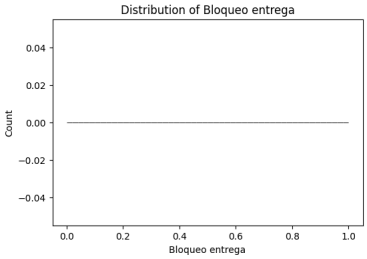

Distribution of Actividad
Count
Actividad

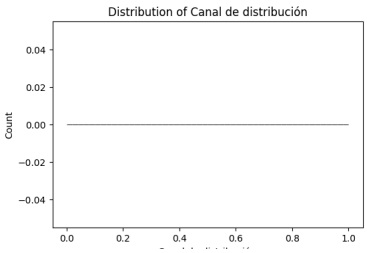

Distribution of Bloque de direcciones (interlocutor)
Count

Bloque de direcciones (interlocutor)

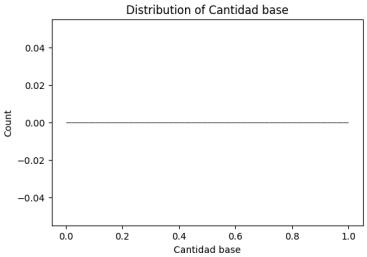


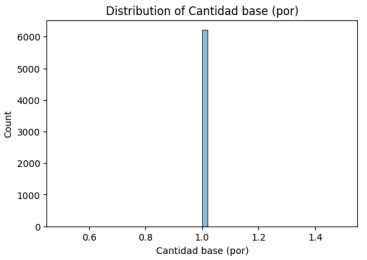


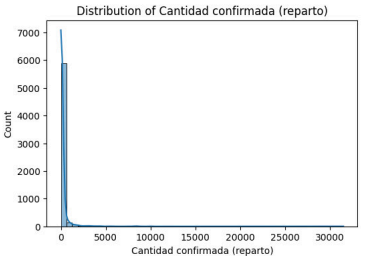


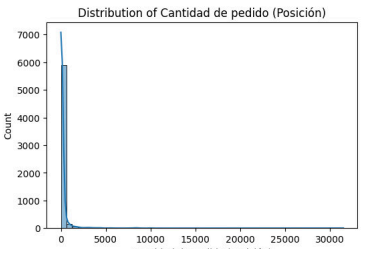


Canal de distribución

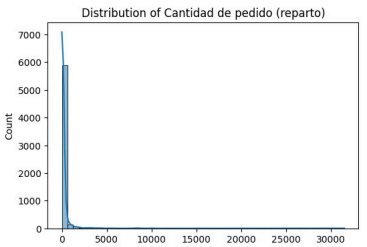


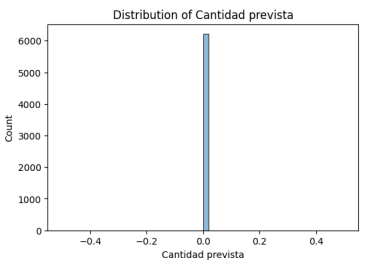


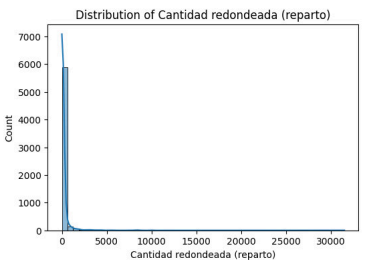


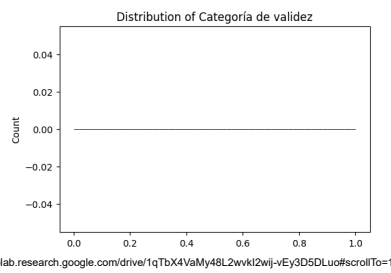


Cantidad de pedido (Posición)

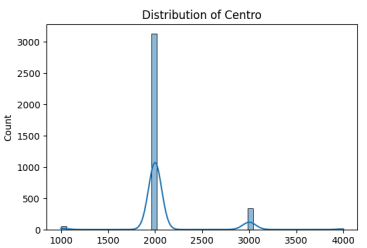


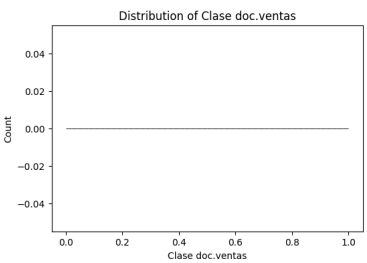


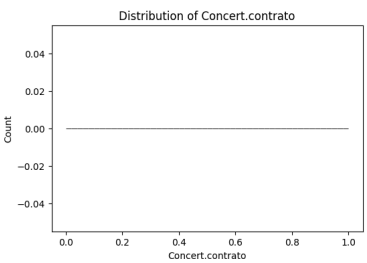


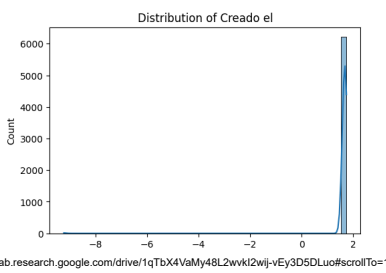


Categoría de validez



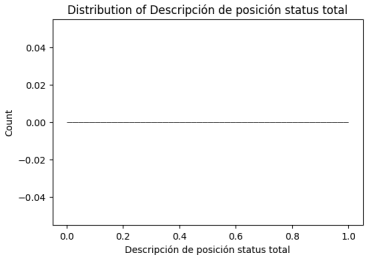




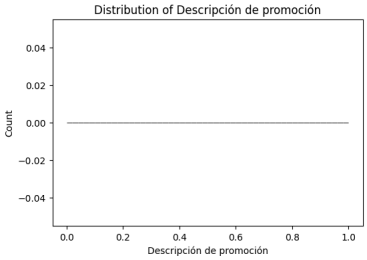


Creado el 1e18

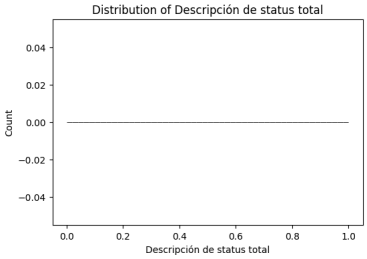
Distribution of Descripción de posición status total



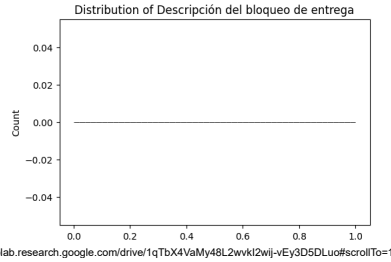
Distribution of Descripción de promoción



Distribution of Descripción de status total

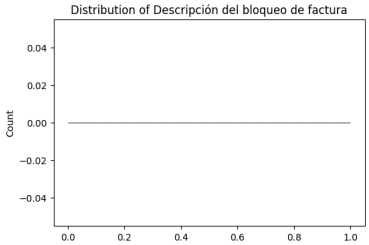


Distribution of Descripción del bloqueo de entrega

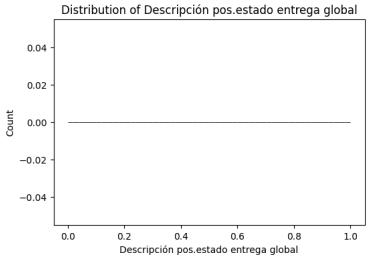


Descripción del bloqueo de entrega

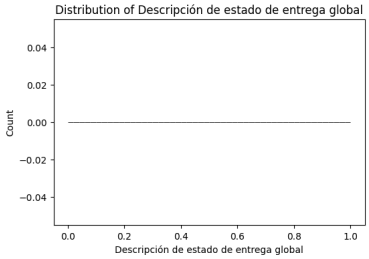
Distribution of Descripción del bloqueo de factura



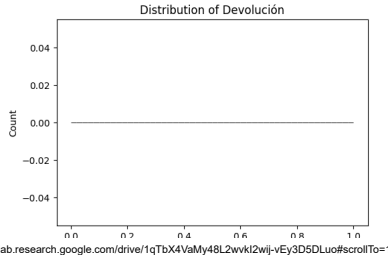
Distribution of Descripción pos.estado entrega global



Distribution of Descripción de estado de entrega global

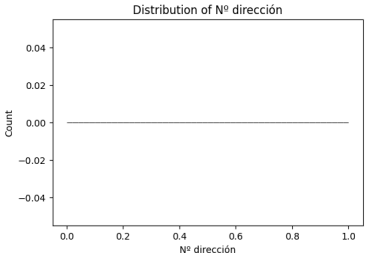


Distribution of Devolución

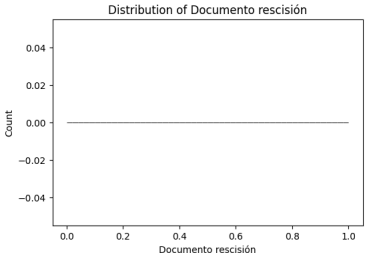


Devolución

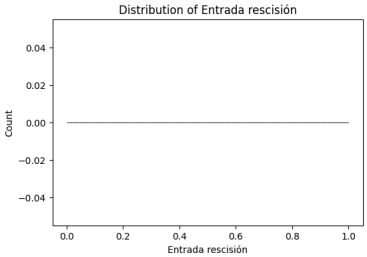
Distribution of Nº dirección



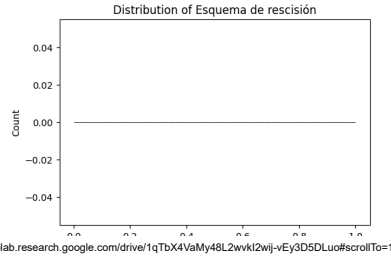
Distribution of Documento rescisión



Distribution of Entrada rescisión

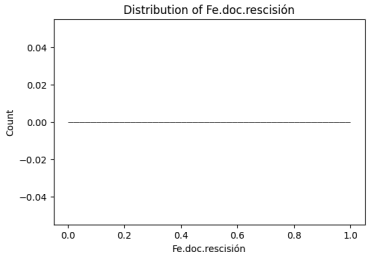


Distribution of Esquema de rescisión

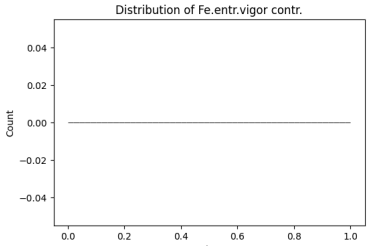


Esquema de rescisión

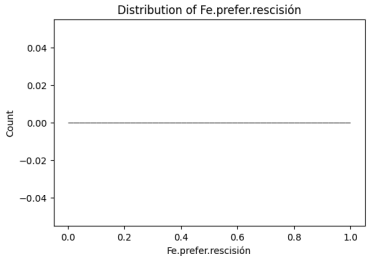
Distribution of Fe.doc.rescisión



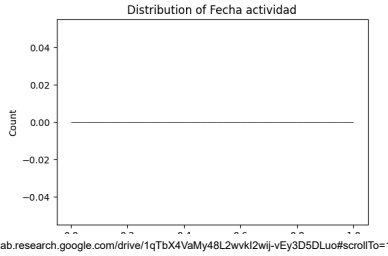
Distribution of Fe.entr.vigor contr.

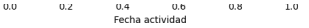


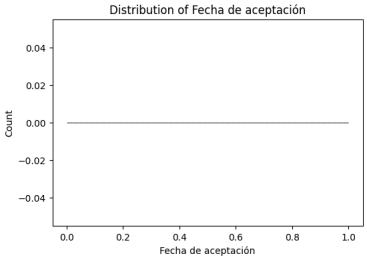
Distribution of Fe.prefer.rescisión

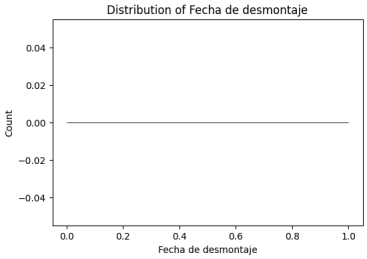


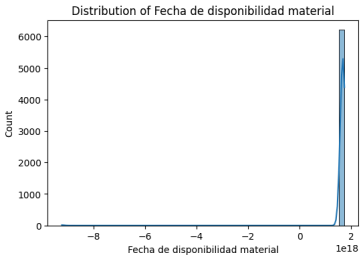
Distribution of Fecha actividad

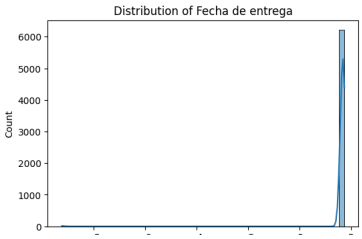



0.0
0.2
0.4
0.6
0.8
1.0
Fecha actividad

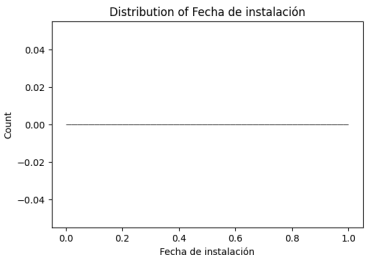

Distribution of Fecha de aceptación
Count
Fecha de aceptación

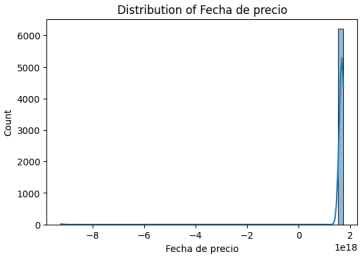

Distribution of Fecha de desmontaje
Count
Fecha de desmontaje

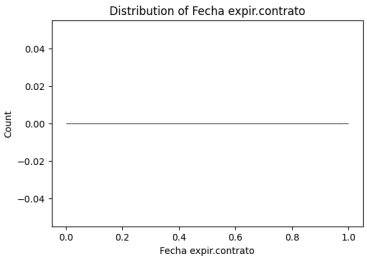

Distribution of Fecha de disponibilidad material
Count
Fecha de disponibilidad material
1e18

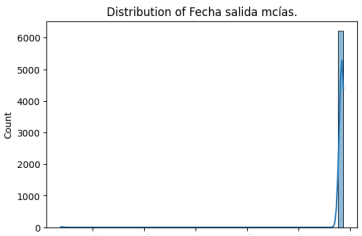

Distribution of Fecha de entrega
Count
Fecha de entrega
1e18


−8
−6
−4
−2
0
2
Fecha de entrega
1e18


Distribution of Fecha de instalación
Count
Fecha de instalación


Distribution of Fecha de precio
Count
Fecha de precio
1e18


Distribution of Fecha expir.contrato
Count
Fecha expir.contrato


Distribution of Fecha salida mcías.
Count
Fecha salida mcías.

−8
−6
−4
−2
0
2
Fecha salida mcías.
1e18

Distribution of Función de interlocutor comercial
Count
Función de interlocutor comercial

Distribution of Función de socio
Count
Función de socio

Distribution of Grupo de vendedores
Count
Grupo de vendedores

Distribution of ID de empleado
Count

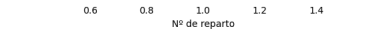
0.0
0.2
0.4
0.6
0.8
1.0
ID de empleado

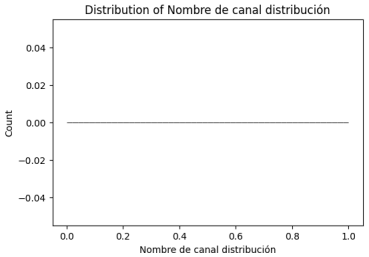
Distribution of Interlocutor
Count
Interlocutor

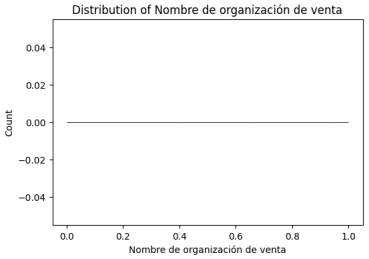
Distribution of Lote
Count
Lote
1e9

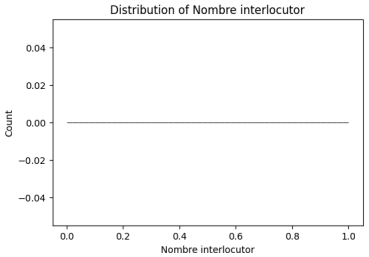
Distribution of Motivo de rescisión
Count
Motivo de rescisión

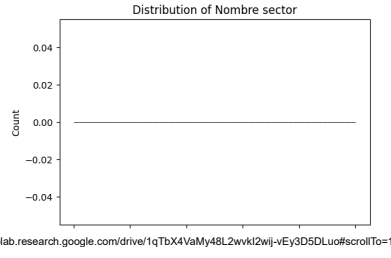
Distribution of Nº de reparto
Count

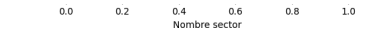

0.6
0.8
1.0
1.2
1.4
Nº de reparto

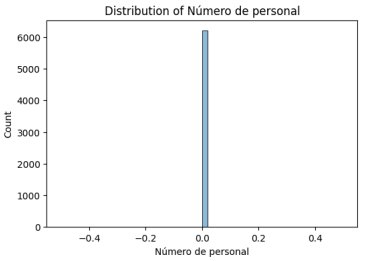

Distribution of Nombre de canal distribución
Count
Nombre de canal distribución

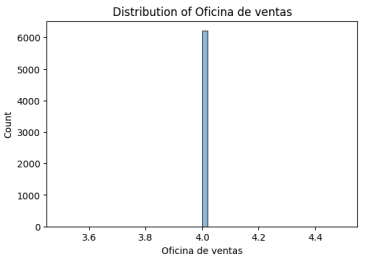

Distribution of Nombre de organización de venta
Count
Nombre de organización de venta

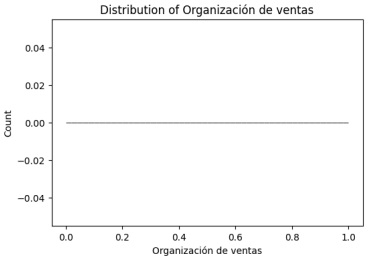

Distribution of Nombre interlocutor
Count
Nombre interlocutor

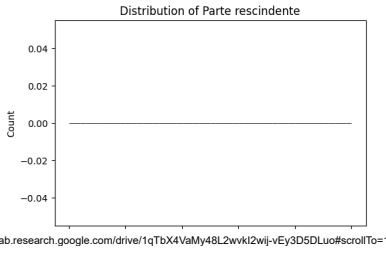

Distribution of Nombre sector
Count

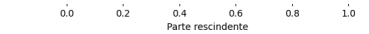

0.0
0.2
0.4
0.6
0.8
1.0
Nombre sector

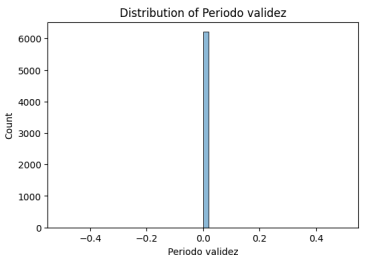

Distribution of Número de personal
Count
Número de personal

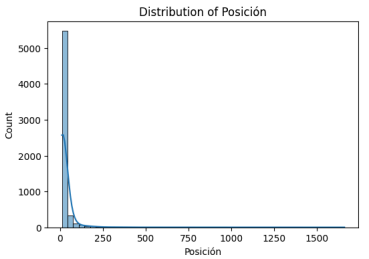

Distribution of Oficina de ventas
Count
Oficina de ventas

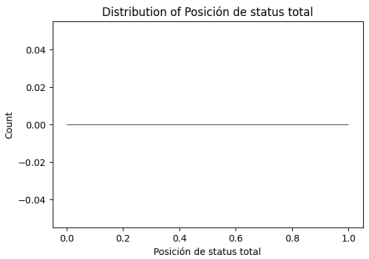

Distribution of Organización de ventas
Count
Organización de ventas

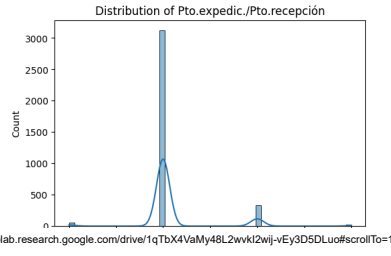

Distribution of Parte rescindente
Count

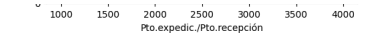

0.0
0.2
0.4
0.6
0.8
1.0
Parte rescindente

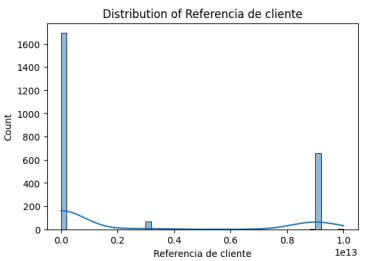

Distribution of Periodo validez
Count
Periodo validez

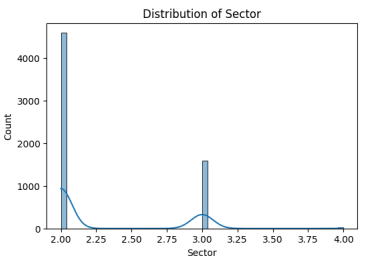

Distribution of Posición
Count
Posición

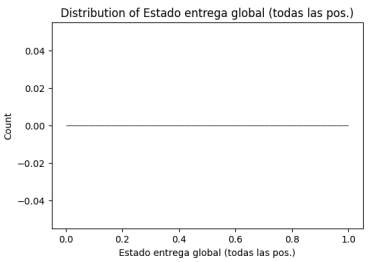

Distribution of Posición de status total
Count
Posición de status total

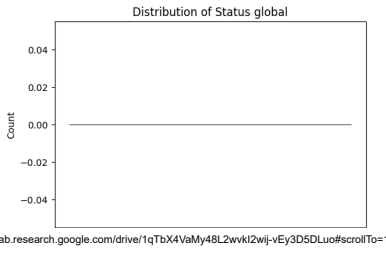

Distribution of Pto.expedic./Pto.recepción
Count

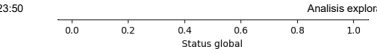

1000
1500
2000
2500
3000
3500
4000
Pto.expedic./Pto.recepción

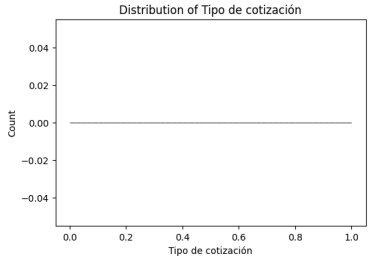

Distribution of Referencia de cliente
Count
Referencia de cliente
1e13

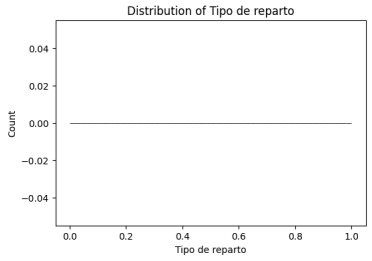

Distribution of Sector
Count
Sector

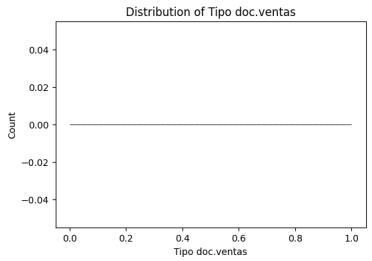

Distribution of Estado entrega global (todas las pos.)
Count
Estado entrega global (todas las pos.)

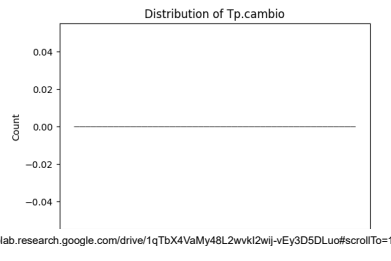

Distribution of Status global
Count

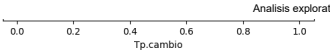


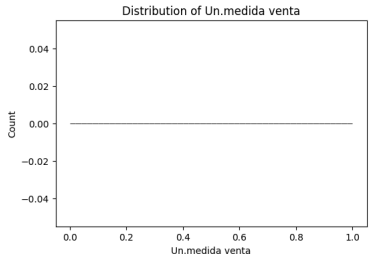


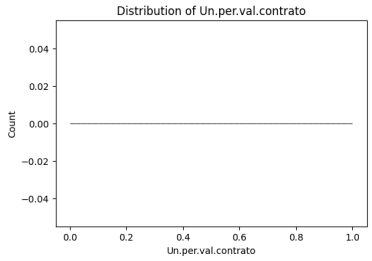


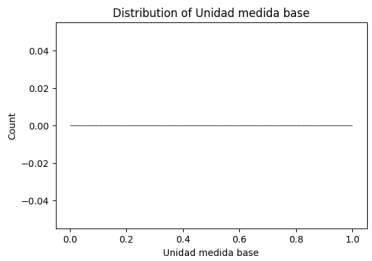


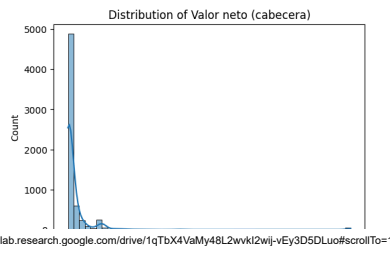


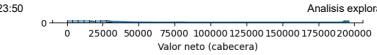


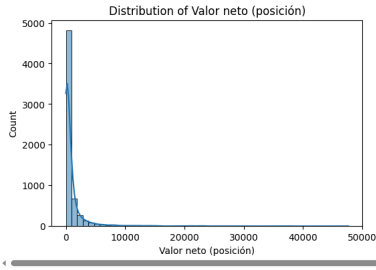






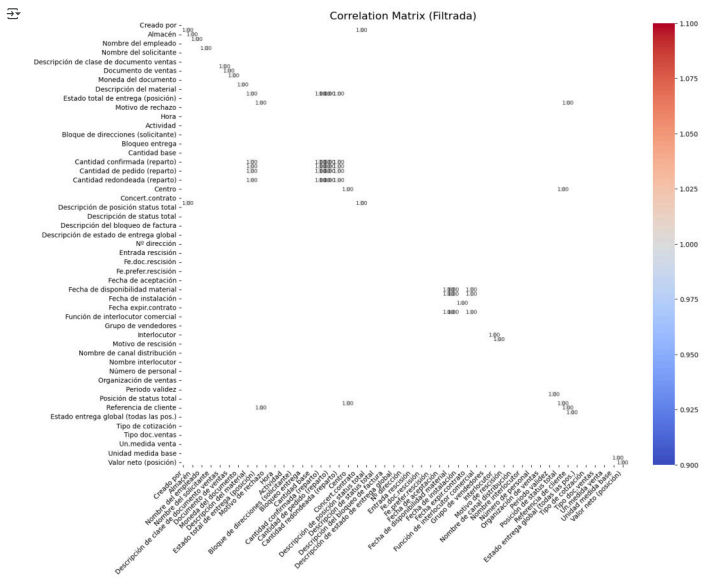






2.2 Correlacion

```
correlation_matrixb1= combined_dfb1[numerical_colsb1].corr()
plt.figure(figsize=(15, 12))
mask = np.abs(correlation_matrixb1) < 1
sns.heatmap(correlation_matrixb1, annot=True, cmap='coolwarm', fmt=".2f",
            annot_kws={"size": 8}, mask=mask)
plt.title('Correlation Matrix (Filtrada)', fontsize=16)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.show()
```




```
correlation_matrix = combined_dfb1.corr().round(2)
```

```
print(correlation_matrix.to_string())
```

```
Cantidad base (por)                          NaN                NaN      NaN         NaN                NaN         NaN
Cantidad confirmada (reparto)                NaN              -0.06    -0.03        0.18                NaN       -0.11
Cantidad de pedido (Posición)                NaN              -0.06    -0.03        0.18                NaN       -0.11
Cantidad de pedido (reparto)                 NaN              -0.06    -0.03        0.18                NaN       -0.11
Cantidad prevista                            NaN                NaN      NaN         NaN                NaN         NaN
Cantidad redondeada (reparto)                NaN              -0.06    -0.03        0.18                NaN       -0.11
Categoría de validez                         NaN                NaN      NaN         NaN                NaN         NaN
Centro                                       NaN               0.04     1.00       -0.12                NaN       -0.01
Clase doc.ventas                             NaN                NaN      NaN         NaN                NaN         NaN
Concert.contrato                             NaN                NaN      NaN         NaN                NaN         NaN
Creado el                                    NaN               1.00     0.04       -0.02                NaN        0.01
Descripción de posición status total         NaN                NaN      NaN         NaN                NaN         NaN
Descripción de promoción                     NaN                NaN      NaN         NaN                NaN         NaN
Descripción de status total                  NaN                NaN      NaN         NaN                NaN         NaN
Descripción del bloqueo de entrega           NaN                NaN      NaN         NaN                NaN         NaN
Descripción del bloqueo de factura           NaN                NaN      NaN         NaN                NaN         NaN
Descripción pos.estado entrega global        NaN                NaN      NaN         NaN                NaN         NaN
Descripción de estado de entrega global      NaN                NaN      NaN         NaN                NaN         NaN
Devolución                                   NaN                NaN      NaN         NaN                NaN         NaN
Nº dirección                                 NaN                NaN      NaN         NaN                NaN         NaN
Documento rescisión                          NaN                NaN      NaN         NaN                NaN         NaN
Entrada rescisión                            NaN                NaN      NaN         NaN                NaN         NaN
Esquema de rescisión                         NaN                NaN      NaN         NaN                NaN         NaN
Fe.doc.rescisión                             NaN                NaN      NaN         NaN                NaN         NaN
Fe.entr.vigor contr.                         NaN                NaN      NaN         NaN                NaN         NaN
Fe.prefer.rescisión                          NaN                NaN      NaN         NaN                NaN         NaN
Fecha actividad                              NaN                NaN      NaN         NaN                NaN         NaN
Fecha de aceptación                          NaN                NaN      NaN         NaN                NaN         NaN
Fecha de desmontaje                          NaN                NaN      NaN         NaN                NaN         NaN
Fecha de disponibilidad material             NaN               1.00     0.04       -0.02                NaN        0.01
Fecha de entrega                             NaN               1.00     0.04       -0.02                NaN        0.01
Fecha de instalación                         NaN                NaN      NaN         NaN                NaN         NaN
Fecha de precio                              NaN               1.00     0.05       -0.02                NaN        0.02
Fecha expir.contrato                         NaN                NaN      NaN         NaN                NaN         NaN
Fecha salida mcías.                          NaN               1.00     0.04       -0.02                NaN        0.01
Función de interlocutor comercial            NaN                NaN      NaN         NaN                NaN         NaN
Función de socio                             NaN                NaN      NaN         NaN                NaN         NaN
Grupo de vendedores                          NaN                NaN      NaN         NaN                NaN         NaN
ID de empleado                               NaN                NaN      NaN         NaN                NaN         NaN
Interlocutor                                 NaN              -0.09     0.06       -0.21                NaN        0.08
Lote                                         NaN               0.02    -0.06        0.16                NaN       -0.14
Motivo de rescisión                          NaN                NaN      NaN         NaN                NaN         NaN
Nº de reparto                                NaN                NaN      NaN         NaN                NaN         NaN
Nombre de canal distribución                 NaN                NaN      NaN         NaN                NaN         NaN
Nombre de organización de venta              NaN                NaN      NaN         NaN                NaN         NaN
Nombre interlocutor                          NaN                NaN      NaN         NaN                NaN         NaN
Nombre sector                                NaN                NaN      NaN         NaN                NaN         NaN
Número de personal                           NaN                NaN      NaN         NaN                NaN         NaN
Oficina de ventas                            NaN                NaN      NaN         NaN                NaN         NaN
Organización de ventas                       NaN                NaN      NaN         NaN                NaN         NaN
Parte rescindente                            NaN                NaN      NaN         NaN                NaN         NaN
Periodo validez                              NaN                NaN      NaN         NaN                NaN         NaN
Posición                                     NaN               0.03    -0.05        0.28                NaN       -0.12
Posición de status total                     NaN                NaN      NaN         NaN                NaN         NaN
Pto.expedic./Pto.recepción                   NaN               0.04     1.00       -0.12                NaN       -0.01
Referencia de cliente                        NaN               0.76    -0.30       -0.10                NaN       -0.02
Sector                                       NaN               0.00    -0.14        0.89                NaN       -0.10
Estado entrega global (todas las pos.)       NaN                NaN      NaN         NaN                NaN         NaN
```

2.3 Categórico

```
categorical_colsb1 = combined_dfb1.select_dtypes(include='object').columns
print("\nCategorical features:", categorical_colsb1)
for col in categorical_colsb1:
    print(f'\nValue counts for {col}:')
    print(combined_dfb1[col].value_counts())
    plt.figure()
    combined_dfb1[col].value_counts().plot(kind='bar')
    plt.title(f'Distribution of {col}')
    plt.show()
```

```
Categorical features: Index([], dtype='object')
```

```
print("Columnas categóricas identificadas:", categorical_colsb1)
```

```
Columnas categóricas identificadas: Index([], dtype='object')
```

```
for col in categorical_colsb1:
    combined_dfb1[col] = combined_dfb1[col].astype('category')

print(combined_dfb1.dtypes)
```



```
Creado por                float64
Fecha del documento         int64
Almacén                   float64
Solicitante               float64
Nombre del empleado       float64
                           ...
Un.medida venta           float64
Un.per.val.contrato       float64
Unidad medida base        float64
Valor neto (cabecera)     float64
Valor neto (posición)     float64
Length: 97, dtype: object
```

```
pd.set_option('display.max_rows', None)
print(combined_dfb1.dtypes)
pd.reset_option('display.max_rows')
```

```
Cantidad confirmada (reparto)              float64
Cantidad de pedido (Posición)              float64
Cantidad de pedido (reparto)               float64
Cantidad prevista                          float64
Cantidad redondeada (reparto)              float64
Categoría de validez                       float64
Centro                                     float64
Clase doc.ventas                           float64
Concert.contrato                           float64
Creado el                                    int64
Descripción de posición status total       float64
Descripción de promoción                   float64
Descripción de status total                float64
Descripción del bloqueo de entrega         float64
Descripción del bloqueo de factura         float64
Descripción pos.estado entrega global      float64
Descripción de estado de entrega global    float64
Devolución                                 float64
Nº dirección                               float64
Documento rescisión                        float64
Entrada rescisión                          float64
Esquema de rescisión                       float64
Fe.doc.rescisión                           float64
Fe.entr.vigor contr.                       float64
Fe.prefer.rescisión                        float64
Fecha actividad                            float64
Fecha de aceptación                        float64
Fecha de desmontaje                        float64
Fecha de disponibilidad material             int64
Fecha de entrega                             int64
Fecha de instalación                       float64
Fecha de precio                              int64
Fecha expir.contrato                       float64
Fecha salida mcías.                          int64
Función de interlocutor comercial          float64
Función de socio                           float64
Grupo de vendedores                        float64
ID de empleado                             float64
Interlocutor                               float64
Lote                                       float64
Motivo de rescisión                        float64
Nº de reparto                              float64
Nombre de canal distribución               float64
Nombre de organización de venta            float64
Nombre interlocutor                        float64
Nombre sector                              float64
Número de personal                         float64
Oficina de ventas                          float64
Organización de ventas                     float64
Parte rescindente                          float64
Periodo validez                            float64
Posición                                   float64
Posición de status total                   float64
Pto.expedic./Pto.recepción                 float64
Referencia de cliente                      float64
Sector                                     float64
Estado entrega global (todas las pos.)     float64
Status global                              float64
Tipo de cotización                         float64
```

```
categorical_columns = ['Interlocutor']
for col in categorical_columns:
    combined_dfb1[col] = combined_dfb1[col].astype('category')

print(combined_dfb1.dtypes)
```

```
Creado por                float64
Fecha del documento         int64
Almacén                   float64
Solicitante               float64
Nombre del empleado       float64
```



```
                           ...
Un.medida venta           float64
Un.per.val.contrato       float64
Unidad medida base        float64
Valor neto (cabecera)     float64
Valor neto (posición)     float64
Length: 97, dtype: object
```

```
pd.set_option('display.max_rows', None)
print(combined_dfb1.dtypes)
pd.reset_option('display.max_rows')
```

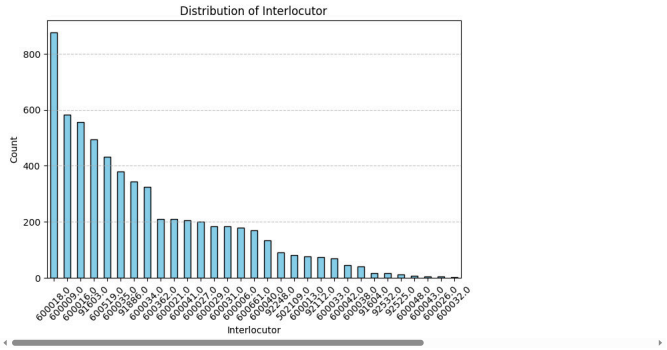
```
Cantidad confirmada (reparto)              float64
Cantidad de pedido (Posición)              float64
Cantidad de pedido (reparto)               float64
Cantidad prevista                          float64
Cantidad redondeada (reparto)              float64
Categoría de validez                       float64
Centro                                     float64
Clase doc.ventas                           float64
Concert.contrato                           float64
Creado el                                    int64
Descripción de posición status total       float64
Descripción de promoción                   float64
Descripción de status total                float64
Descripción del bloqueo de entrega         float64
Descripción del bloqueo de factura         float64
Descripción pos.estado entrega global      float64
Descripción de estado de entrega global    float64
Devolución                                 float64
Nº dirección                               float64
Documento rescisión                        float64
Entrada rescisión                          float64
Esquema de rescisión                       float64
Fe.doc.rescisión                           float64
Fe.entr.vigor contr.                       float64
Fe.prefer.rescisión                        float64
Fecha actividad                            float64
Fecha de aceptación                        float64
Fecha de desmontaje                        float64
Fecha de disponibilidad material             int64
Fecha de entrega                             int64
Fecha de instalación                       float64
Fecha de precio                              int64
Fecha expir.contrato                       float64
Fecha salida mcías.                          int64
Función de interlocutor comercial          float64
Función de socio                           float64
Grupo de vendedores                        float64
ID de empleado                             float64
Interlocutor                              category
Lote                                       float64
Motivo de rescisión                        float64
Nº de reparto                              float64
Nombre de canal distribución               float64
Nombre de organización de venta            float64
Nombre interlocutor                        float64
Nombre sector                              float64
Número de personal                         float64
Oficina de ventas                          float64
Organización de ventas                     float64
Parte rescindente                          float64
Periodo validez                            float64
Posición                                   float64
Posición de status total                   float64
Pto.expedic./Pto.recepción                 float64
Referencia de cliente                      float64
Sector                                     float64
Estado entrega global (todas las pos.)     float64
Status global                              float64
Tipo de cotización                         float64
```

```
def analyze_single_categorical_feature(df, column_name):
    if column_name not in df.columns:
        print(f"Column '{column_name}' not found in DataFrame.")
        return

    if df[column_name].dtype.name != 'category':
        print(f"Column '{column_name}' is not categorical.")
        return

    print(f'\nValue counts for {column_name}:')
    print(df[column_name].value_counts())

    plt.figure(figsize=(8, 5))
    df[column_name].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title(f'Distribution of {column_name}')
    plt.xlabel(column_name)
```



```
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.show()


for col in combined_dfb1.columns:
    if combined_dfb1[col].dtype == 'float64' and combined_dfb1[col].nunique() < 20:  # Ajusta el umbral
        combined_dfb1[col] = combined_dfb1[col].astype('category')

print("Tipos de datos después de la conversión:")
print(combined_dfb1.dtypes)


column_to_analyze = 'Interlocutor'  # Reemplaza con el nombre de la columna deseada
analyze_single_categorical_feature(combined_dfb1, column_to_analyze)
```

```
Tipos de datos después de la conversión:
Creado por                category
Fecha del documento          int64
Almacén                   category
Solicitante                float64
Nombre del empleado       category
                             ...
Un.medida venta           category
Un.per.val.contrato       category
Unidad medida base        category
Valor neto (cabecera)      float64
Valor neto (posición)      float64
Length: 97, dtype: object

Value counts for Interlocutor:
Interlocutor
600018.0    875
600009.0    583
600016.0    555
91603.0     495
600519.0    432
600035.0    380
91886.0     343
600034.0    325
600362.0    211
600021.0    211
600041.0    205
600027.0    201
600029.0    184
600031.0    183
600006.0    179
600661.0    169
600040.0    134
92248.0      92
502109.0     82
600013.0     77
92112.0      74
600033.0     70
600042.0     45
600038.0     41
91604.0      18
92532.0      16
92525.0      12
600048.0      8
600043.0      5
600026.0      4
600032.0      2
Name: count, dtype: int64
```



B2

[ ] ↳ 11 celdas ocultas

2. Análisis exploratorio

2.1 Numérico

```
numerical_colsb2 = combined_dfb2.select_dtypes(include=np.number).columns
print("Numerical features:", numerical_colsb2)
combined_dfb2[numerical_colsb2].describe().transpose()
```

Mostrar salida oculta

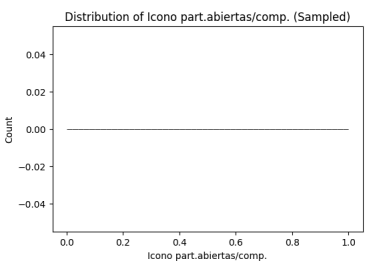
```
print(combined_dfb2.dtypes)
```

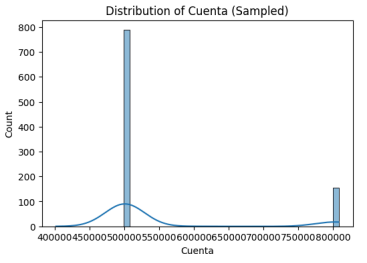
Mostrar salida oculta

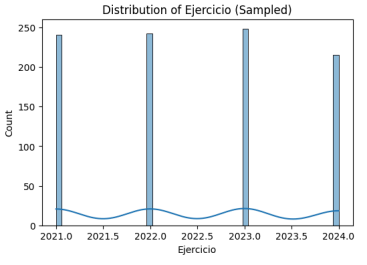
```
for col in combined_dfb2.columns:
    combined_dfb2[col] = pd.to_numeric(combined_dfb2[col], errors='coerce')

numerical_colsb2 = combined_dfb2.select_dtypes(include=['number']).columns.tolist()
print("Columnas numéricas corregidas:", numerical_colsb2)
```

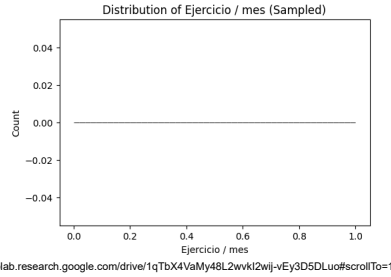
```
Columnas numéricas corregidas: ['Icono part.abiertas/comp.', 'Cuenta', 'Ejercicio', 'Ejercicio / mes', 'Fecha contabiliz.', 'Referei
```

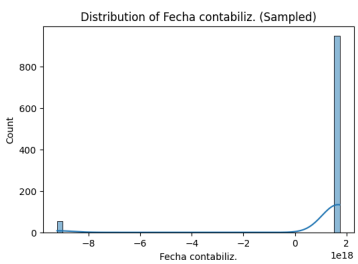
```
sample_columnsb2 = numerical_colsb2[:10]

sampled_dfb2 = combined_dfb2.sample(n=min(1000, len(combined_dfb2)), random_state=42)

for col in sample_columnsb2:
    plt.figure(figsize=(6, 4))
    sns.histplot(sampled_dfb2[col], kde=True, bins=50)
    plt.title(f"Distribution of {col} (Sampled)")
    plt.show()
```

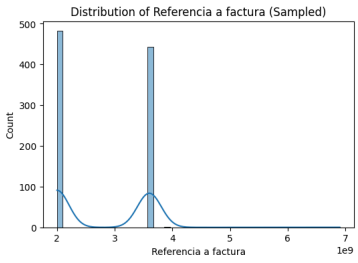



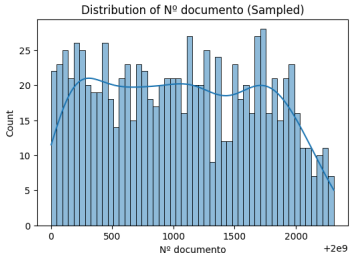


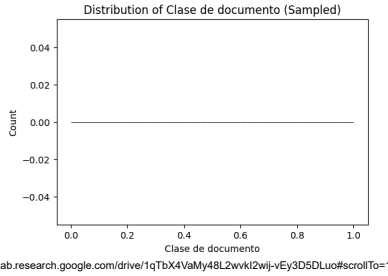


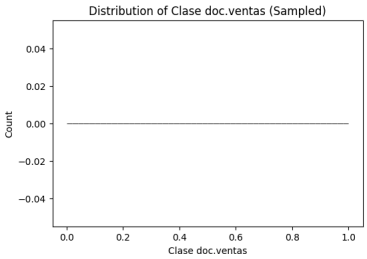


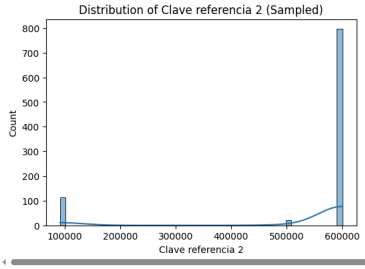






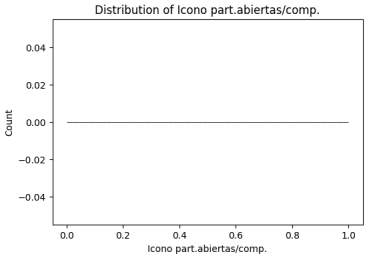


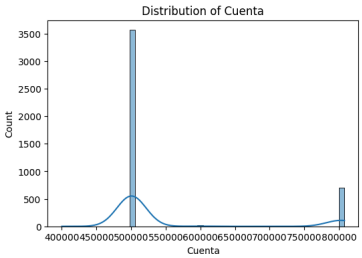


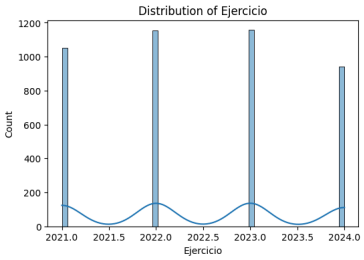


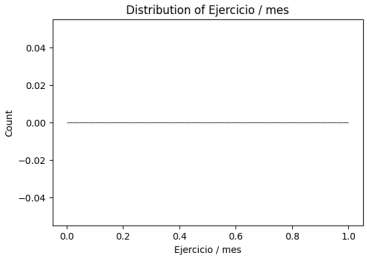

```
for col in numerical_colsb2:
    plt.figure(figsize=(6, 4))
    sns.histplot(combined_dfb2[col], kde=True, bins=50)
    plt.title(f"Distribution of {col}")
    plt.show()
    plt.pause(0.1)
```

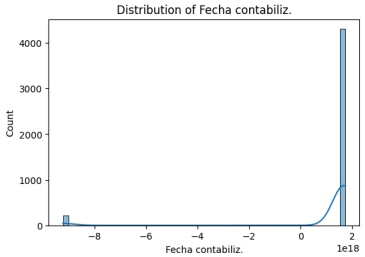

[4]

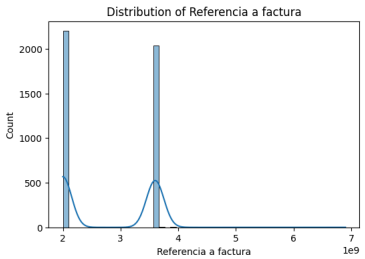


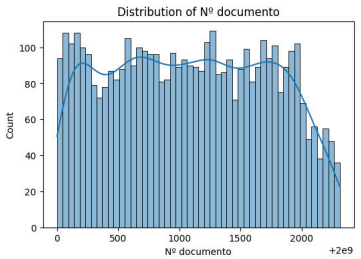


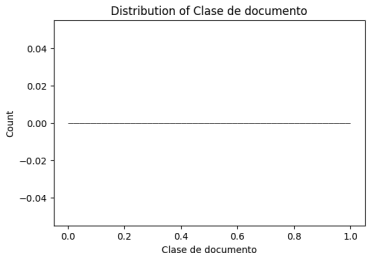


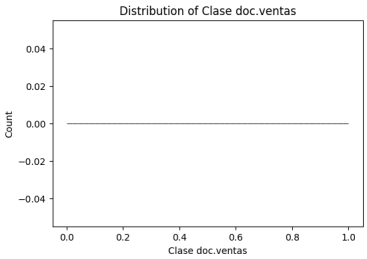


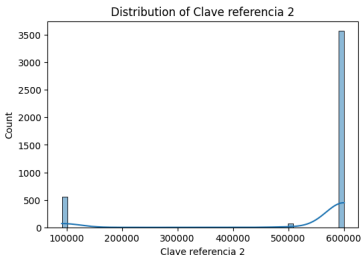


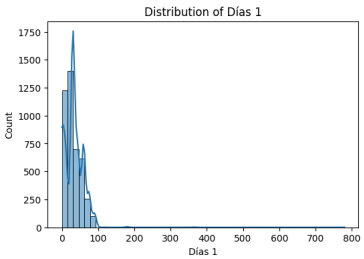


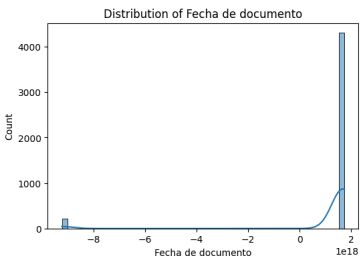


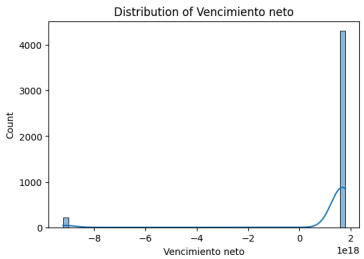


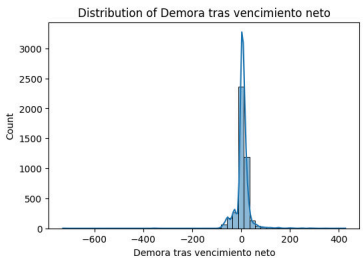

Distribution of Clase doc.ventas
Count
Clase doc.ventas

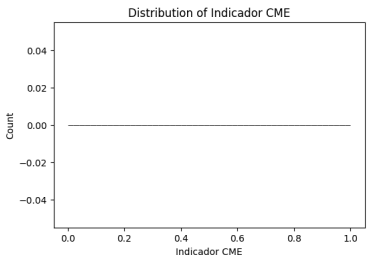

Distribution of Clave referencia 2
Count
Clave referencia 2

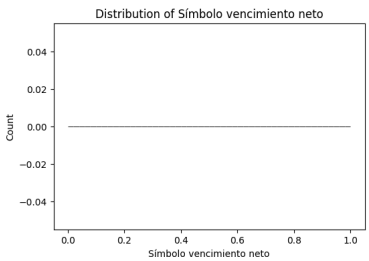

Distribution of Días 1
Count
Días 1

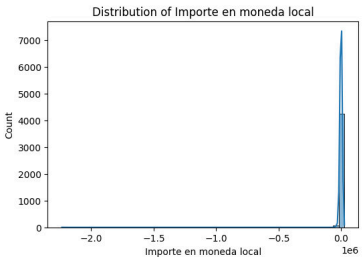

Distribution of Fecha de documento
Count
Fecha de documento
1e18

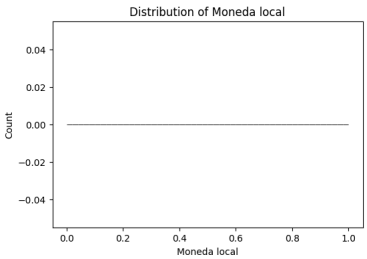

Distribution of Vencimiento neto
Count
Vencimiento neto
1e18

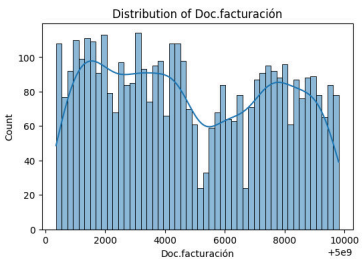

Distribution of Demora tras vencimiento neto
Count
Demora tras vencimiento neto

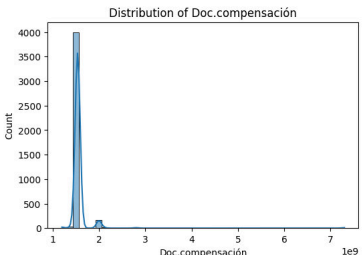

Distribution of Indicador CME
Count
Indicador CME

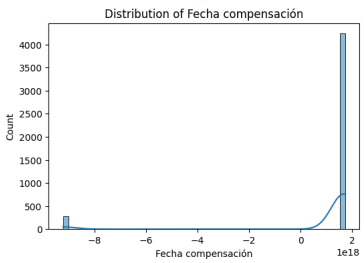

Distribution of Símbolo vencimiento neto
Count
Símbolo vencimiento neto

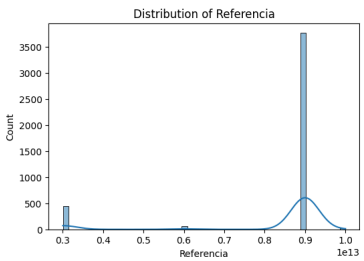

Distribution of Importe en moneda local
Count
Importe en moneda local
1e6

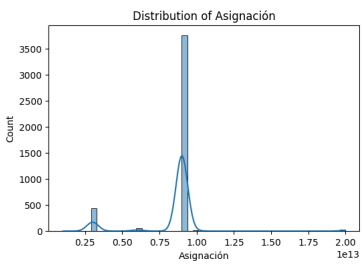

Distribution of Moneda local
Count
Moneda local

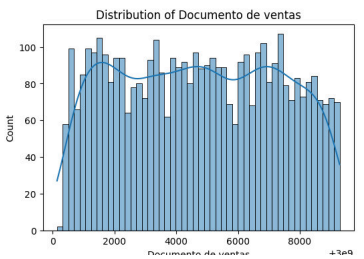

Distribution of Doc.facturación
Count
Doc.facturación
+5e9

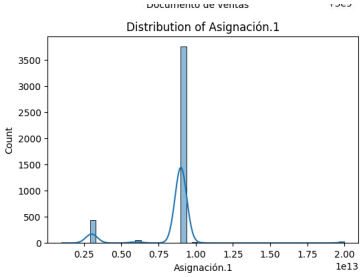

Distribution of Doc.compensación
Count
1e9

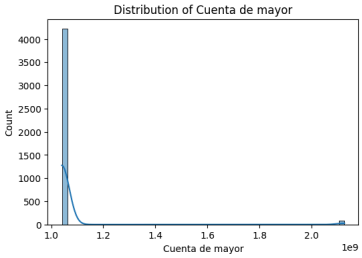

Distribution of Fecha compensación
Count
Fecha compensación
1e18

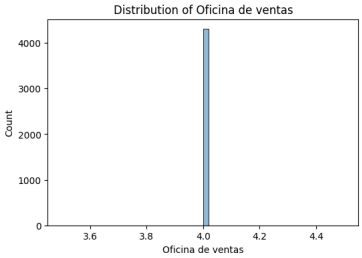

Distribution of Referencia
Count
Referencia
1e13

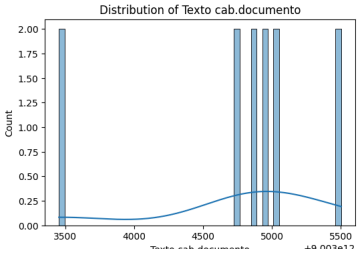

Distribution of Asignación
Count
Asignación
1e13


Distribution of Documento de ventas
Count



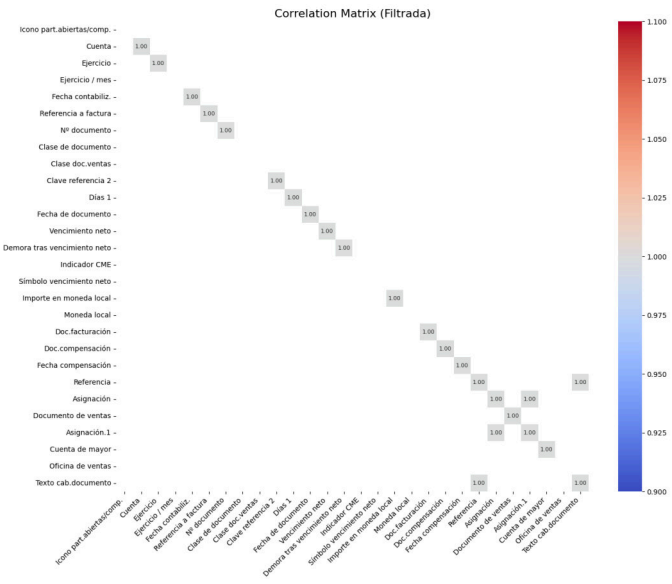






2.2 Correlación

```
correlation_matrixb2= combined_dfb2[numerical_colsb2].corr()
plt.figure(figsize=(15, 12))
mask = np.abs(correlation_matrixb2) < 1
sns.heatmap(correlation_matrixb2, annot=True, cmap='coolwarm', fmt=".2f",
            annot_kws={"size": 8}, mask=mask)
plt.title('Correlation Matrix (Filtrada)', fontsize=16)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.show()
```




```
correlation_matrixb2 = combined_dfb2.corr().round(2)
```

```
print(correlation_matrixb2.to_string())
```

```
                                Icono part.abiertas/comp.  Cuenta  Ejercicio  Ejercicio / mes  Fecha contabiliz.  Referencia a factura
Icono part.abiertas/comp.                             NaN     NaN        NaN              NaN                NaN                   NaN
Cuenta                                                NaN    1.00       0.02              NaN               0.03                 -0.07
```



```
Ejercicio                          NaN    0.02       1.00              NaN               0.96                  0.7
Ejercicio / mes                    NaN     NaN        NaN              NaN                NaN                  NaN
Fecha contabiliz.                  NaN    0.03       0.96              NaN               1.00                  0.7
Referencia a factura               NaN   -0.07       0.73              NaN               0.73                  1.0
Nº documento                       NaN    0.04       0.01              NaN               0.27                  0.1
Clase de documento                 NaN     NaN        NaN              NaN                NaN                  NaN
Clase doc.ventas                   NaN     NaN        NaN              NaN                NaN                  NaN
Clave referencia 2                 NaN   -0.12      -0.06              NaN              -0.07                 -0.0
Días 1                             NaN   -0.10       0.21              NaN               0.20                  0.3
Fecha de documento                 NaN    0.03       0.96              NaN               1.00                  0.7
Vencimiento neto                   NaN    0.02       0.96              NaN               1.00                  0.7
Demora tras vencimiento neto       NaN   -0.00      -0.18              NaN              -0.19                 -0.1
Indicador CME                      NaN     NaN        NaN              NaN                NaN                  NaN
Símbolo vencimiento neto           NaN     NaN        NaN              NaN                NaN                  NaN
Importe en moneda local            NaN   -0.24       0.01              NaN               0.97                  0.0
Moneda local                       NaN     NaN        NaN              NaN                NaN                  NaN
Doc.facturación                    NaN    0.01       0.97              NaN               1.00                  0.8
Doc.compensación                   NaN    0.08       0.09              NaN               0.09                 -0.0
Fecha compensación                 NaN    0.00      -0.14              NaN               0.07                 -0.0
Referencia                         NaN   -0.08       0.02              NaN               0.01                 -0.0
Asignación                         NaN   -0.02       0.02              NaN               0.02                 -0.0
Documento de ventas                NaN    0.01       0.97              NaN               1.00                  0.8
Asignación.1                       NaN   -0.02       0.02              NaN               0.02                 -0.0
Cuenta de mayor                    NaN    0.13       0.08              NaN               0.09                 -0.1
Oficina de ventas                  NaN     NaN        NaN              NaN                NaN                  NaN
Texto cab.documento                NaN    0.21       0.93              NaN               1.00                 -0.0
```

2.3 Categórico

```
categorical_colsb2 = combined_dfb2.select_dtypes(include='object').columns
print("\nCategorical features:", categorical_colsb2)
for col in categorical_colsb2:
    print(f'\nValue counts for {col}:')
    print(combined_dfb2[col].value_counts())
    plt.figure()
    combined_dfb2[col].value_counts().plot(kind='bar')
    plt.title(f'Distribution of {col}')
    plt.show()
```

```
Categorical features: Index([], dtype='object')
```

```
print("Columnas categóricas identificadas:", categorical_colsb2)
```

```
Columnas categóricas identificadas: Index([], dtype='object')
```

```
print(combined_dfb2.dtypes)
```

```
Icono part.abiertas/comp.       float64
Cuenta                          float64
Ejercicio                       float64
Ejercicio / mes                 float64
Fecha contabiliz.                 int64
Referencia a factura            float64
Nº documento                    float64
Clase de documento              float64
Clase doc.ventas                float64
Clave referencia 2              float64
Días 1                          float64
Fecha de documento                int64
Vencimiento neto                  int64
Demora tras vencimiento neto    float64
Indicador CME                   float64
Símbolo vencimiento neto        float64
Importe en moneda local         float64
Moneda local                    float64
Doc.facturación                 float64
Doc.compensación                float64
Fecha compensación                int64
Referencia                      float64
Asignación                      float64
Documento de ventas             float64
Asignación.1                    float64
Cuenta de mayor                 float64
Oficina de ventas               float64
Texto cab.documento             float64
dtype: object
```

```
pd.set_option('display.max_rows', None)
print(combined_dfb2.dtypes)
pd.reset_option('display.max_rows')
```

```
Icono part.abiertas/comp.       float64
Cuenta                          float64
Ejercicio                       float64
Ejercicio / mes                 float64
Fecha contabiliz.                 int64
Referencia a factura            float64
Nº documento                    float64
Clase de documento              float64
Clase doc.ventas                float64
Clave referencia 2              float64
Días 1                          float64
Fecha de documento                int64
Vencimiento neto                  int64
Demora tras vencimiento neto    float64
Indicador CME                   float64
Símbolo vencimiento neto        float64
Importe en moneda local         float64
Moneda local                    float64
Doc.facturación                 float64
Doc.compensación                float64
Fecha compensación                int64
Referencia                      float64
Asignación                      float64
Documento de ventas             float64
Asignación.1                    float64
Cuenta de mayor                 float64
Oficina de ventas               float64
Texto cab.documento             float64
dtype: object
```

```
categorical_columns = ['Interlocutor']
for col in categorical_columns:
    combined_dfb2[col] = combined_dfb2[col].astype('category')

print(combined_dfb2.dtypes)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in get_loc(self, key)
   3804         try:
-> 3805             return self._engine.get_loc(casted_key)
   3806         except KeyError as err:

index.pyx in pandas._libs.index.IndexEngine.get_loc()

index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'Interlocutor'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
 2 frames 
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in get_loc(self, key)
   3810             ):
   3811                 raise InvalidIndexError(key)
-> 3812             raise KeyError(key) from err
   3813         except TypeError:
   3814             # If we have a listlike key, _check_indexing_error will raise

KeyError: 'Interlocutor'
```

```
pd.set_option('display.max_rows', None)
print(combined_dfb2.dtypes)
pd.reset_option('display.max_rows')
```

```
Icono part.abiertas/comp.       float64
Cuenta                          float64
Ejercicio                       float64
Ejercicio / mes                 float64
Fecha contabiliz.                 int64
Referencia a factura            float64
Nº documento                    float64
Clase de documento              float64
Clase doc.ventas                float64
Clave referencia 2              float64
Días 1                          float64
Fecha de documento                int64
Vencimiento neto                  int64
Demora tras vencimiento neto    float64
Indicador CME                   float64
Símbolo vencimiento neto        float64
Importe en moneda local         float64
Moneda local                    float64
Doc.facturación                 float64
Doc.compensación                float64
Fecha compensación                int64
Referencia                      float64
Asignación                      float64
Documento de ventas             float64
Asignación.1                    float64
Cuenta de mayor                 float64
Oficina de ventas               float64
Texto cab.documento             float64
dtype: object
```



```
Importe en moneda local         float64
Moneda local                    float64
Doc.facturación                 float64
Doc.compensación                float64
Fecha compensación                int64
Referencia                      float64
Asignación                      float64
Documento de ventas             float64
Asignación.1                    float64
Cuenta de mayor                 float64
Oficina de ventas               float64
Texto cab.documento             float64
dtype: object
```

```
def analyze_single_categorical_feature(df, column_name):
    if column_name not in df.columns:
        print(f"Column '{column_name}' not found in DataFrame.")
        return

    if df[column_name].dtype.name != 'category':
        print(f"Column '{column_name}' is not categorical.")
        return

    print(f'\nValue counts for {column_name}:')
    print(df[column_name].value_counts())

    plt.figure(figsize=(8, 5))
    df[column_name].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title(f'Distribution of {column_name}')
    plt.xlabel(column_name)
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.show()


for col in combined_dfb2.columns:
    if combined_dfb2[col].dtype == 'float64' and combined_dfb2[col].nunique() < 20:  # Ajusta el umbral
        combined_dfb1[col] = combined_dfb2[col].astype('category')

print("Tipos de datos después de la conversión:")
print(combined_dfb2.dtypes)


column_to_analyze = 'Interlocutor'  # Reemplaza con el nombre de la columna deseada
analyze_single_categorical_feature(combined_dfb2, column_to_analyze)
```

```
Tipos de datos después de la conversión:
Icono part.abiertas/comp.       float64
Cuenta                          float64
Ejercicio                       float64
Ejercicio / mes                 float64
Fecha contabiliz.                 int64
Referencia a factura            float64
Nº documento                    float64
Clase de documento              float64
Clase doc.ventas                float64
Clave referencia 2              float64
Días 1                          float64
Fecha de documento                int64
Vencimiento neto                  int64
Demora tras vencimiento neto    float64
Indicador CME                   float64
Símbolo vencimiento neto        float64
Importe en moneda local         float64
Moneda local                    float64
Doc.facturación                 float64
Doc.compensación                float64
Fecha compensación                int64
Referencia                      float64
Asignación                      float64
Documento de ventas             float64
Asignación.1                    float64
Cuenta de mayor                 float64
Oficina de ventas               float64
Texto cab.documento             float64
dtype: object
Column 'Interlocutor' not found in DataFrame.
```

B3

1. Carga de archivos


```
excel_files = [     '/content/drive/MyDrive/Datos/b3/b3-2021.xlsx',     '/content/drive/MyDrive/Datos/b3/b3-2022.xlsx',     '/content/dr
```

```
dfsb3 = []
```

```
!pip install openpyxl
```

```
Requirement already satisfied: openpyxl in /usr/local/lib/python3.11/dist-packages (3.1.5)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.11/dist-packages (from openpyxl) (2.0.0)
```

```
for file in excel_files:
  dfb3 = pd.read_excel(file)
  dfsb3.append(dfb3)

combined_dfb3 = pd.concat(dfsb3, ignore_index=True)
```

```
combined_dfb3.head()
```

Mostrar salida oculta

```
combined_dfb3.describe()
```

Mostrar salida oculta

```
excel_file_pathb3 = '/content/archivo_combinado.xlsx'
combined_dfb3.to_excel(excel_file_pathb3, index=False)

print(f"El archivo ha sido guardado en: {excel_file_pathb3}")
```

```
El archivo ha sido guardado en: /content/archivo_combinado.xlsx
```

```
from google.colab import files

files.download(excel_file_pathb3)
```

2. Análisis exploratorio

2.1 Numérico

```
numerical_colsb3 = combined_dfb3.select_dtypes(include=np.number).columns
print("Numerical features:", numerical_colsb3)
combined_dfb3[numerical_colsb3].describe().transpose()
```

Mostrar salida oculta

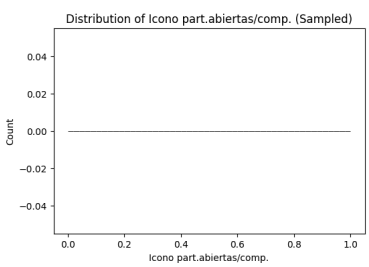
```
print(combined_dfb3.dtypes)
```

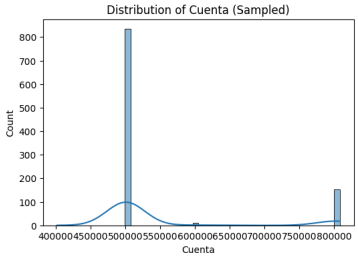
Mostrar salida oculta

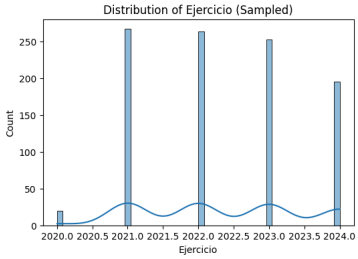
```
for col in combined_dfb3.columns:
    combined_dfb3[col] = pd.to_numeric(combined_dfb3[col], errors='coerce')

numerical_colsb3 = combined_dfb3.select_dtypes(include=['number']).columns.tolist()
print("Columnas numéricas corregidas:", numerical_colsb3)
```

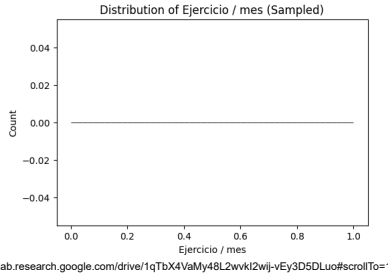
Mostrar salida oculta

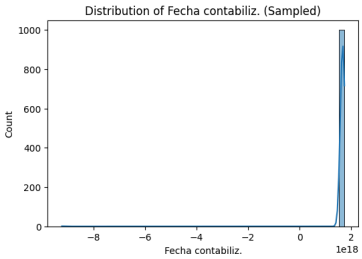
```
sample_columnsb3 = numerical_colsb3[:10]

sampled_dfb3 = combined_dfb3.sample(n=min(1000, len(combined_dfb3)), random_state=42)

for col in sample_columnsb3:
    plt.figure(figsize=(6, 4))
    sns.histplot(sampled_dfb3[col], kde=True, bins=50)
    plt.title(f"Distribution of {col} (Sampled)")
    plt.show()
```

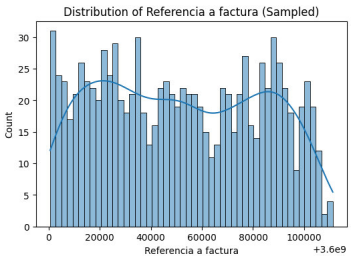



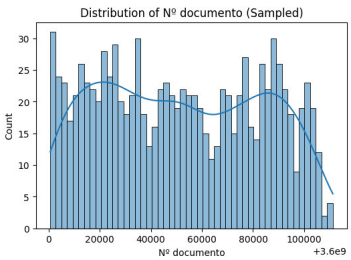


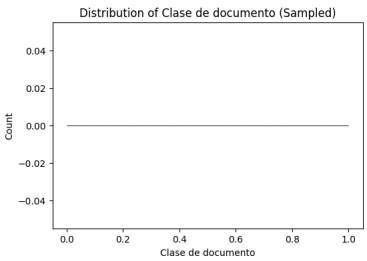


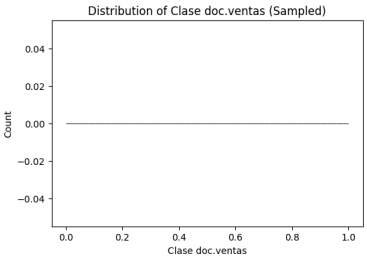


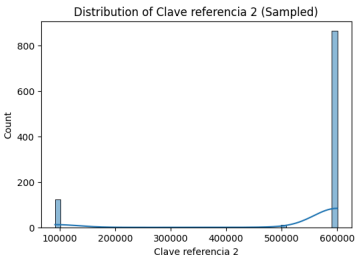




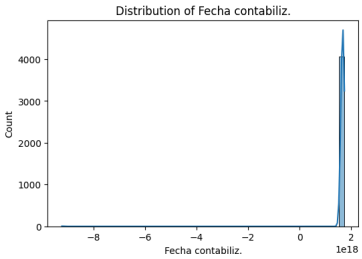


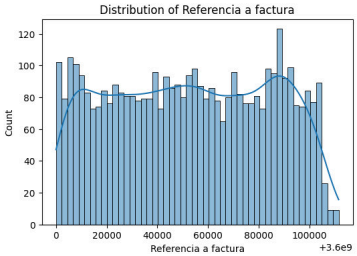


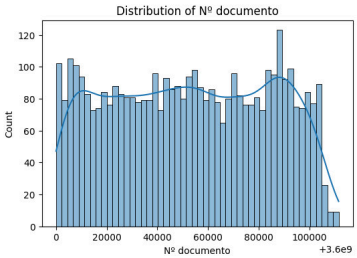


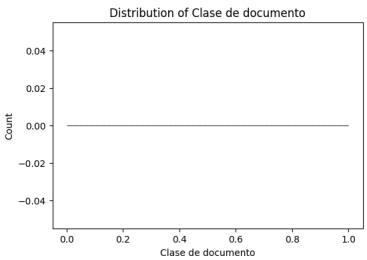


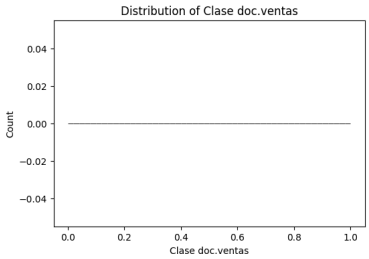

```
for col in numerical_colsb3:
    plt.figure(figsize=(6, 4))
    sns.histplot(combined_dfb3[col], kde=True, bins=50)
    plt.title(f"Distribution of {col}")
    plt.show()
    plt.pause(0.1)
```

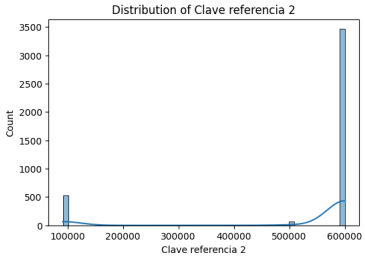

Distribution of Fecha contabiliz.
Count
Fecha contabiliz.
1e18

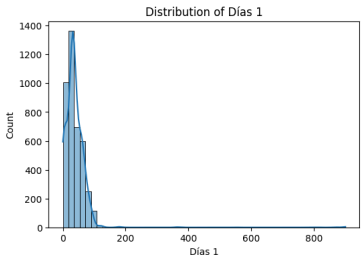

Distribution of Referencia a factura
Count
Referencia a factura
+3.6e9

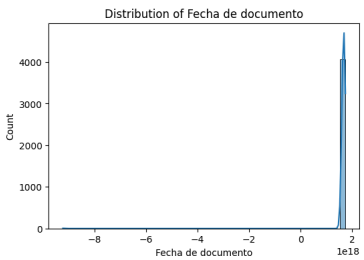

Distribution of Nº documento
Count
Nº documento
+3.6e9

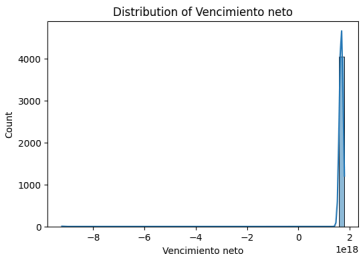

Distribution of Clase de documento
Count
Clase de documento

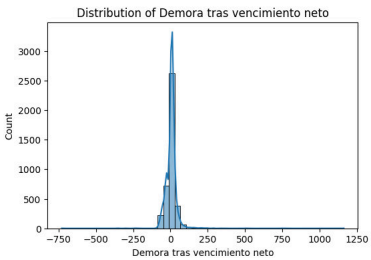

Distribution of Clase doc.ventas
Count
Clase doc.ventas

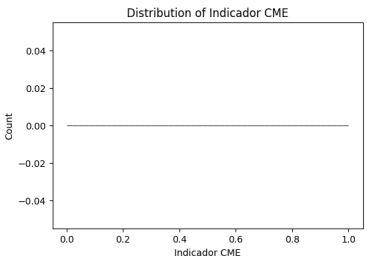

Distribution of Clave referencia 2
Count
Clave referencia 2

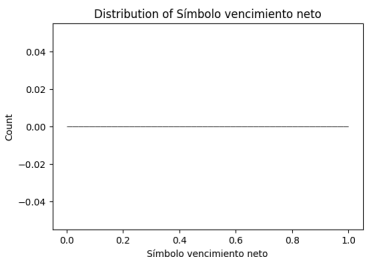

Distribution of Días 1
Count
Días 1

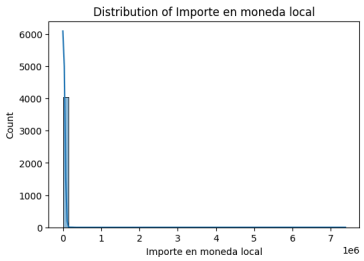

Distribution of Fecha de documento
Count
Fecha de documento
1e18

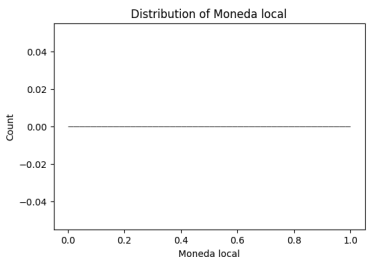

Distribution of Vencimiento neto
Count
Vencimiento neto
1e18

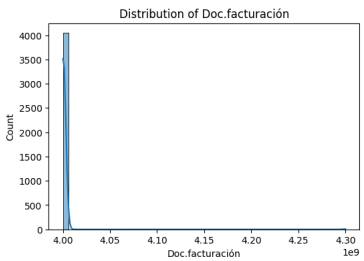

Distribution of Demora tras vencimiento neto
Count
Demora tras vencimiento neto

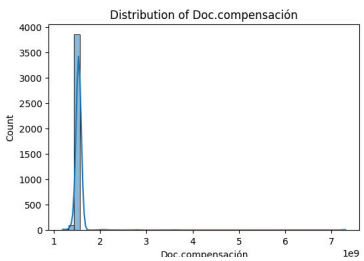

Distribution of Indicador CME
Count
Indicador CME

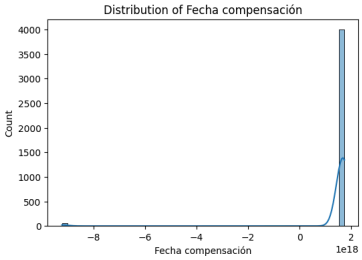

Distribution of Símbolo vencimiento neto
Count
Símbolo vencimiento neto

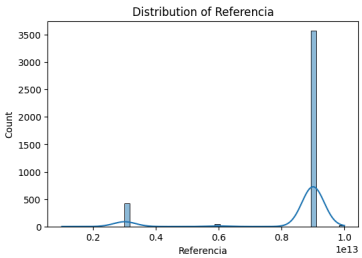

Distribution of Importe en moneda local
Count
Importe en moneda local
1e6

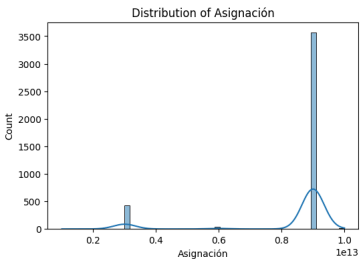

Distribution of Moneda local
Count
Moneda local

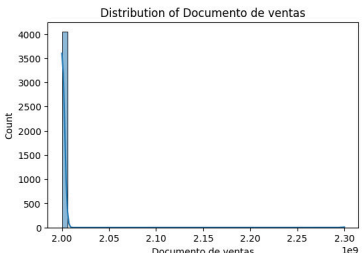

Distribution of Doc.facturación
Count
Doc.facturación
1e9

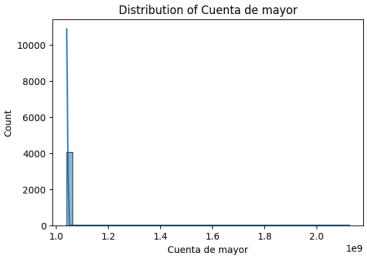

Distribution of Doc.compensación
Count
1e9

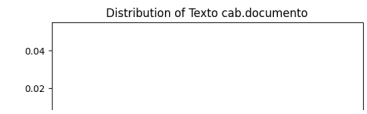

Distribution of Fecha compensación
Count
Fecha compensación
1e18


Distribution of Referencia
Count
Referencia
1e13


Distribution of Asignación
Count
Asignación
1e13


Distribution of Documento de ventas
Count
Documento de ventas
1e9


Distribution of Cuenta de mayor
Count
Cuenta de mayor
1e9


Distribution of Texto cab.documento

Importar Librerias

Clusters

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
!pip install openpyxl
```

```
Requirement already satisfied: openpyxl in /usr/local/lib/python3.11/dist-packages (3.1.5)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.11/dist-packages (from openpyxl) (2.0.0)
```

Preparación del entorno

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

B1

```
excel_files = [     '/content/drive/MyDrive/Datos/b1/b1-2021.xlsx',     '/content/drive/MyDrive/Datos/b1/b1-2022.xlsx',     '/content/dr
```

```
dfsb1 = []
```

```
for file in excel_files:
  dfb1 = pd.read_excel(file)
  dfsb1.append(dfb1)
```

```
combined_dfb1 = pd.concat(dfsb1, ignore_index=True)
```

```
combined_dfb1.head()
```

Mostrar salida oculta

```
columns_to_keep_b1 = [
    'Creado por',
    'Fecha del documento',
    'Almacén',
    'Motivo de pedido',
    'Solicitante',
    'Precio neto',
    'Nombre del solicitante',
    'Descripción del motivo de pedido',
    'Documento de ventas',
    'Material',
    'Descripción del material',
    'Cantidad confirmada (posición)',
    'Hora',
    'Centro',
    'Clase doc.ventas',
    'Creado el',
    'Interlocutor',
    'Lote',
    'Nombre interlocutor',
    'Nombre sector',
    'Valor neto (posición)'
]
```

```
combined_dfb1_cleaned = combined_dfb1[[col for col in columns_to_keep_b1 if col in combined_dfb1.columns]]
```

```
combined_dfb1_cleaned.to_excel("base_limpiab1.xlsx", index=False)
from google.colab import files
files.download("base_limpiab1.xlsx")
```

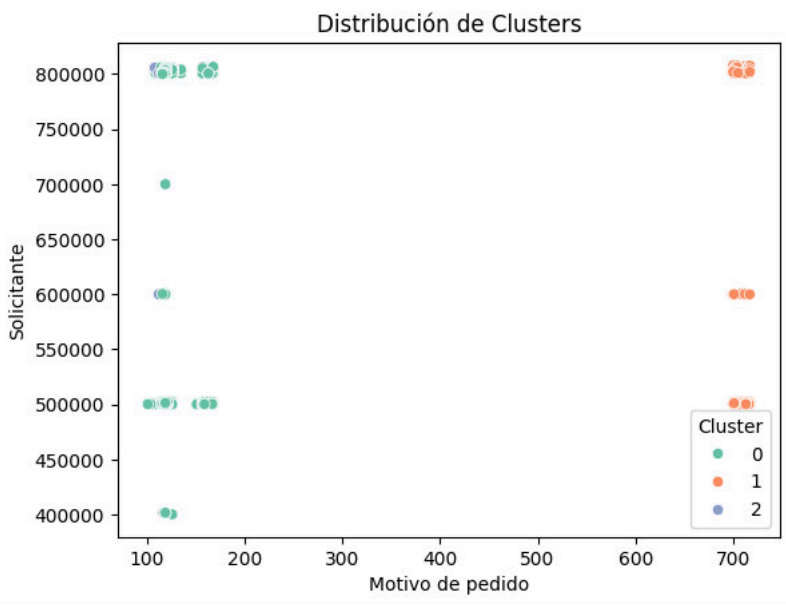
```
from sklearn.cluster import KMeans


from sklearn.preprocessing import StandardScaler


numerical_colsb1 = [col for col in columns_to_keep_b1 if pd.api.types.is_numeric_dtype(combined_dfb1_cleaned[col])]

print("Columnas numéricas seleccionadas:")
print(numerical_colsb1)
```

```
Columnas numéricas seleccionadas:
['Motivo de pedido', 'Solicitante', 'Precio neto', 'Documento de ventas', 'Material', 'Cantidad confirmada (posición)', 'Interlocut
```

```
scaler = StandardScaler()
scaled_data = scaler.fit_transform(combined_dfb1_cleaned[numerical_colsb1])


from sklearn.impute import SimpleImputer

# Rellenar NaNs con la media de cada columna
imputer = SimpleImputer(strategy='mean')
imputed_data = imputer.fit_transform(combined_dfb1_cleaned[numerical_colsb1])

# Escalar datos
scaler = StandardScaler()
scaled_data = scaler.fit_transform(imputed_data)

# KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
combined_dfb1_cleaned['Cluster'] = kmeans.fit_predict(scaled_data)
```

```
<ipython-input-23-54bba2dd4549>:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus
  combined_dfb1_cleaned['Cluster'] = kmeans.fit_predict(scaled_data)
```

```
sns.scatterplot(x=combined_dfb1_cleaned[numerical_colsb1[0]], y=combined_dfb1_cleaned[numerical_colsb1[1]], hue=combined_dfb1_cleaned['(
plt.title('Distribución de Clusters')
plt.show()
```




```
centroids = pd.DataFrame(
    kmeans.cluster_centers_,
    columns=[col + ' (escalado)' for col in numerical_colsb1]
)
centroids['Cluster'] = centroids.index
```

```
cluster_sizes = combined_dfb1_cleaned['Cluster'].value_counts().reset_index()
cluster_sizes.columns = ['Cluster', 'Cantidad_de_Registros']
cluster_sizes = cluster_sizes.sort_values('Cluster')


resumen_clusters = combined_dfb1_cleaned.groupby('Cluster')[numerical_colsb1].agg(['mean', 'min', 'max', 'count']).reset_index()


resumen_clusters.columns = ['_'.join(col).strip('_') if isinstance(col, tuple) else col for col in resumen_clusters.columns]


with pd.ExcelWriter('/content/drive/MyDrive/cluster_info.xlsx') as writer:

    combined_dfb1_cleaned.to_excel(writer, sheet_name='Data con Clusters', index=False)

    centroids.to_excel(writer, sheet_name='Centroides', index=False)

    cluster_sizes.to_excel(writer, sheet_name='Tamaño de Clusters', index=False)

    resumen_clusters.to_excel(writer, sheet_name='Resumen Estadístico', index=False)
```

B2

```
excel_files = [     '/content/drive/MyDrive/Datos/b2/b2-2021.xlsx',     '/content/drive/MyDrive/Datos/b2/b2-2022.xlsx',     '/content/dr
```

```
dfsb2 = []


for file in excel_files:
  dfb2 = pd.read_excel(file)
  dfsb2.append(dfb2)


combined_dfb2 = pd.concat(dfsb2, ignore_index=True)


combined_dfb2.head()
```

Mostrar salida oculta

```
columns_to_keep_b2 = [
    'Cuenta',
    'Ejercicio',
    'Fecha contabiliz.',
    'Clase de documento',
    'Clave referencia 2',
    'Demora tras vencimiento neto',
    'Importe en moneda local',
    'Asignación',
    'Documento de ventas'
]


combined_dfb2_cleaned = combined_dfb2[[col for col in columns_to_keep_b2 if col in combined_dfb2.columns]]


combined_dfb2_cleaned.to_excel("base_limpiab2.xlsx", index=False)
from google.colab import files
files.download("base_limpiab2.xlsx")
```

B3

```
excel_files = [     '/content/drive/MyDrive/Datos/b3/b3-2021.xlsx',     '/content/drive/MyDrive/Datos/b3/b3-2022.xlsx',     '/content/dr
```

```
dfsb3 = []


for file in excel_files:
  dfb3 = pd.read_excel(file)
  dfsb3.append(dfb3)


combined_dfb3 = pd.concat(dfsb3, ignore_index=True)


combined_dfb3.head()
```

Mostrar salida oculta

```
columns_to_keep_b3 = [
    'Cuenta',
    'Ejercicio',
    'Fecha contabiliz.',
    'Clase doc.ventas',
    'Clave referencia 2',
    'Días 1',
    'Fecha de documento',
    'Vencimiento neto',
    'Demora tras vencimiento neto',
    'Importe en moneda local',
    'Asignación',
    'Documento de ventas'
]


combined_dfb3_filtered = combined_dfb3[[col for col in columns_to_keep_b3 if col in combined_dfb3.columns]]


combined_dfb3_filtered.to_excel("base_b3_limpia.xlsx", index=False)
from google.colab import files
files.download("base_b3_limpia.xlsx")
```

B4

```
excel_files = [     '/content/drive/MyDrive/Datos/b4/b4-2021.xlsx',     '/content/drive/MyDrive/Datos/b4/b4-2022.xlsx',     '/content/dr
```

```
dfsb4 = []


for file in excel_files:
  dfb4 = pd.read_excel(file)
  dfsb4.append(dfb4)


combined_dfb4 = pd.concat(dfsb4, ignore_index=True)


combined_dfb4.head()
```

Mostrar salida oculta

```
columns_to_keep_b4 = [
    'Creado por',
    'Fecha del documento',
    'Clase doc.ventas',
    'Solicitante',
    'Nombre del solicitante',
    'Documento de ventas',
    'Material',
    'Descripción del material',
    'Motivo de rechazo',
    'Descripción del motivo de rechazo',
    'Centro'
]


combined_dfb4_filtered = combined_dfb4[[col for col in columns_to_keep_b4 if col in combined_dfb4.columns]]


combined_dfb4_filtered.to_excel("base_b4_limpia.xlsx", index=False)
from google.colab import files
files.download("base_b4_limpia.xlsx")
```

UNION DE BASES

```
b1 = pd.read_excel('/content/base_limpiab1.xlsx')
b2 = pd.read_excel('/content/base_limpiab2.xlsx')
b3 = pd.read_excel('/content/base_b3_limpia.xlsx')
b4 = pd.read_excel('/content/base_b4_limpia.xlsx')


print(b4.columns)
```

```
Index(['Creado por', 'Fecha del documento', 'Clase doc.ventas', 'Solicitante',
       'Nombre del solicitante', 'Documento de ventas', 'Material',
       'Descripción del material', 'Motivo de rechazo',
       'Descripción del motivo de rechazo', 'Centro'],
      dtype='object')
```

```
merged_1_2 = pd.merge(b1, b2, on='Documento de ventas', how='left')
merged_1_3 = pd.merge(merged_1_2, b3, on='Asignación', how='left')
```

```
b4.rename(columns={'Documento de ventas': 'Documento de ventas_y'}, inplace=True)
```

```
base_final = pd.merge(merged_1_3, b4, on='Documento de ventas_y', how='left')
```

```
# Paso 4: Depurar base

# Eliminar columnas con más del 90% de valores nulos
umbral_nulos = 0.9
base_final = base_final.loc[:, base_final.isnull().mean() < umbral_nulos]
```

```
# Eliminar duplicados
base_final = base_final.drop_duplicates()
```

```
# Eliminar filas completamente vacías
base_final = base_final.dropna(how='all')
```

```
# Paso 5: Exportar base final
base_final.to_excel('/content/base_unificada.xlsx', index=False)

from google.colab import files
files.download('/content/base_unificada.xlsx')
```

```
numerical_colsb1 = [col for col in base_final if pd.api.types.is_numeric_dtype(base_final[col])]

print("Columnas numéricas seleccionadas:")
print(numerical_colsb1)
```

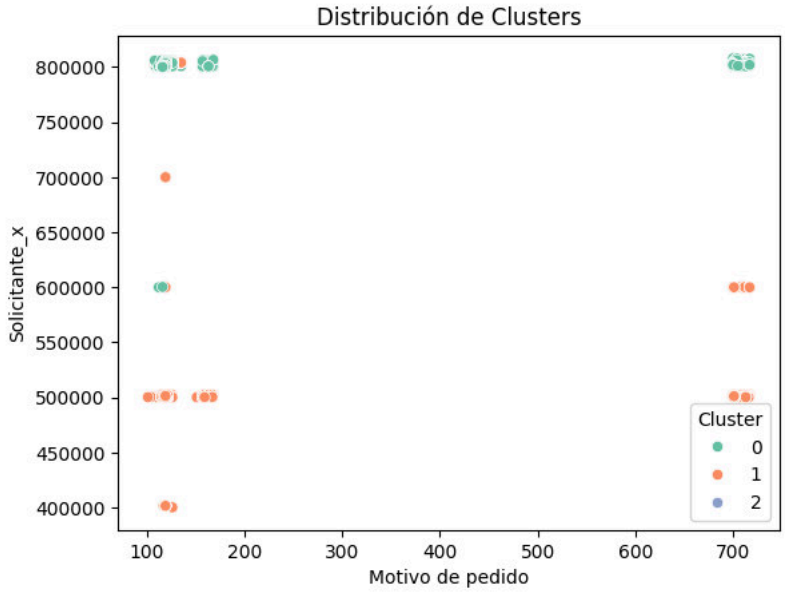
```
Columnas numéricas seleccionadas:
['Motivo de pedido', 'Solicitante_x', 'Precio neto', 'Documento de ventas_x', 'Material_x', 'Cantidad confirmada (posición)', 'Inter
```

```
scaler = StandardScaler()
scaled_data = scaler.fit_transform(base_final[numerical_colsb1])
```

```
from sklearn.impute import SimpleImputer

# Rellenar NaNs con la media de cada columna
imputer = SimpleImputer(strategy='mean')
imputed_data = imputer.fit_transform(base_final[numerical_colsb1])
```

```
# Escalar datos
scaler = StandardScaler()
scaled_data = scaler.fit_transform(imputed_data)

# KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
base_final['Cluster'] = kmeans.fit_predict(scaled_data)
```

```
sns.scatterplot(x=base_final[numerical_colsb1[0]], y=base_final[numerical_colsb1[1]], hue=base_final['Cluster'], palette='Set2')
plt.title('Distribución de Clusters')
plt.show()
```




```
centroids = pd.DataFrame(
    kmeans.cluster_centers_,
    columns=[col + ' (escalado)' for col in numerical_colsb1]
)
centroids['Cluster'] = centroids.index



cluster_sizes = base_final['Cluster'].value_counts().reset_index()
cluster_sizes.columns = ['Cluster', 'Cantidad_de_Registros']
cluster_sizes = cluster_sizes.sort_values('Cluster')



resumen_clusters = base_final.groupby('Cluster')[numerical_colsb1].agg(['mean', 'min', 'max', 'count']).reset_index()
resumen_clusters.columns = ['_'.join(col).strip('_') if isinstance(col, tuple) else col for col in resumen_clusters.columns]
with pd.ExcelWriter('/content/drive/MyDrive/cluster_info2.xlsx') as writer:

    base_final.to_excel(writer, sheet_name='Data con Clusters', index=False)

    centroids.to_excel(writer, sheet_name='Centroides', index=False)

    cluster_sizes.to_excel(writer, sheet_name='Tamaño de Clusters', index=False)

    resumen_clusters.to_excel(writer, sheet_name='Resumen Estadístico', index=False)
```

Análisis Gráfico de Devoluciones

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar archivo
from google.colab import files
uploaded = files.upload()
df = pd.read_excel("base_unificada.xlsx")
sns.set(style='whitegrid')
```

```
Elegir archivos Ningún archivo seleccionado Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving base_unificada.xlsx to base_unificada.xlsx
```

```
# Asegurarse de convertir columnas de fecha

# Fecha del documento y creación de campo mensual para análisis temporal
df['Fecha del documento_x'] = pd.to_datetime(df['Fecha del documento_x'], errors='coerce')
df['Mes'] = df['Fecha del documento_x'].dt.to_period('M')
```

Gráfico 1: Top 10 Motivos de Devolución

```
plt.figure(figsize=(10, 5))
motivos = df['Descripción del motivo de pedido'].value_counts().head(10)
sns.barplot(x=motivos.values, y=motivos.index)
plt.title('Top 10 Motivos de Devolución')
plt.xlabel('Cantidad')
plt.ylabel('Motivo')
plt.tight_layout()
plt.show()
```

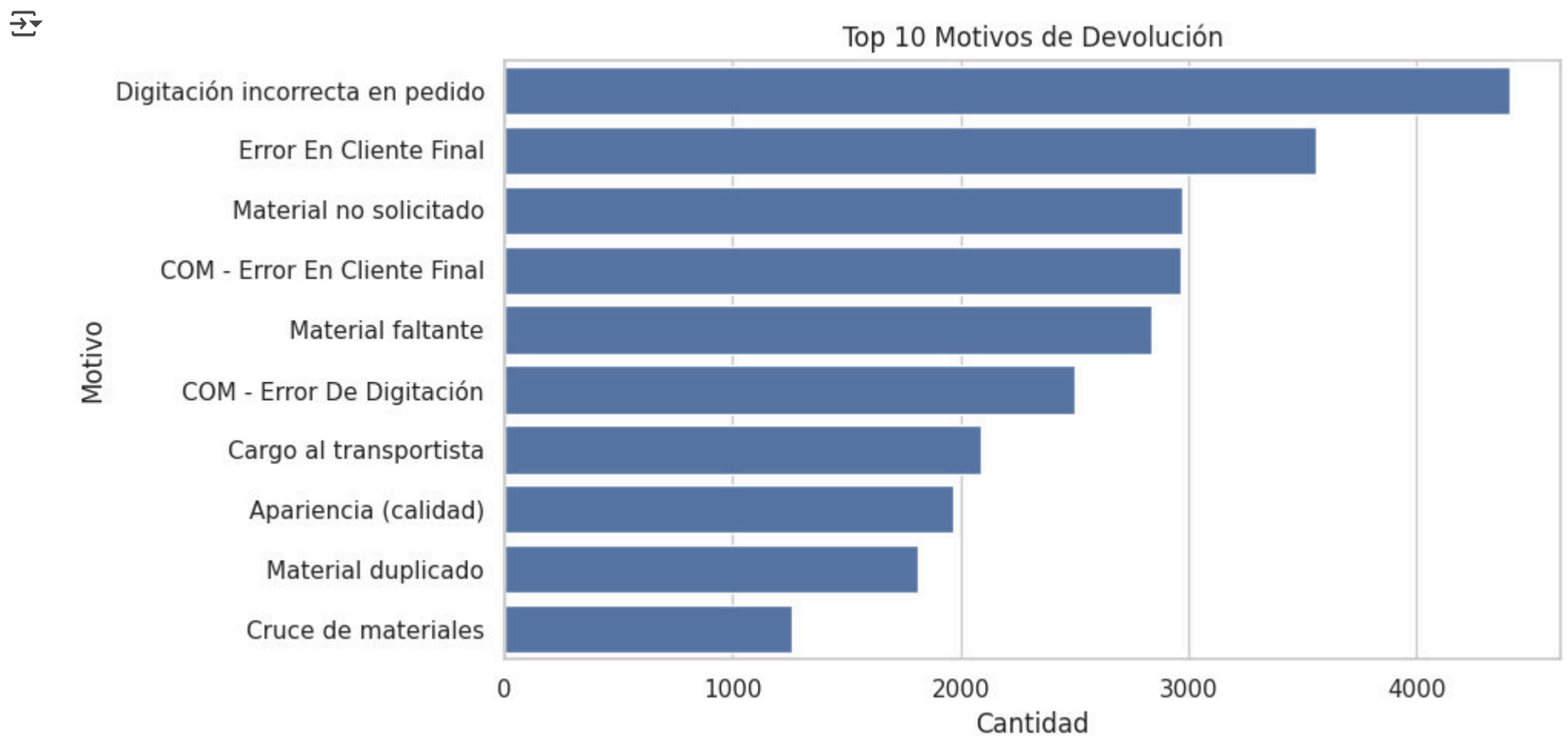


Gráfico 2: Top 10 Solicitantes con más Devoluciones

Anonimato

```
plt.figure(figsize=(10, 5))

# Obtener los 10 usuarios más frecuentes
creado_counts = df['Creado por_x'].value_counts().head(10)

# Crear etiquetas anónimas: Cliente 1, Cliente 2, ...
anon_labels = {name: f"Cliente {i+1}" for i, name in enumerate(creado_counts.index)}

# Dibujar el gráfico usando los nombres anonimizados
sns.barplot(x=creado_counts.values, y=[anon_labels[name] for name in creado_counts.index])

plt.title("Top 10 Clientes que Registraron Devoluciones (Anonimizados)")
plt.xlabel("Cantidad de devoluciones")
plt.ylabel("Cliente")
plt.tight_layout()
plt.show()
```

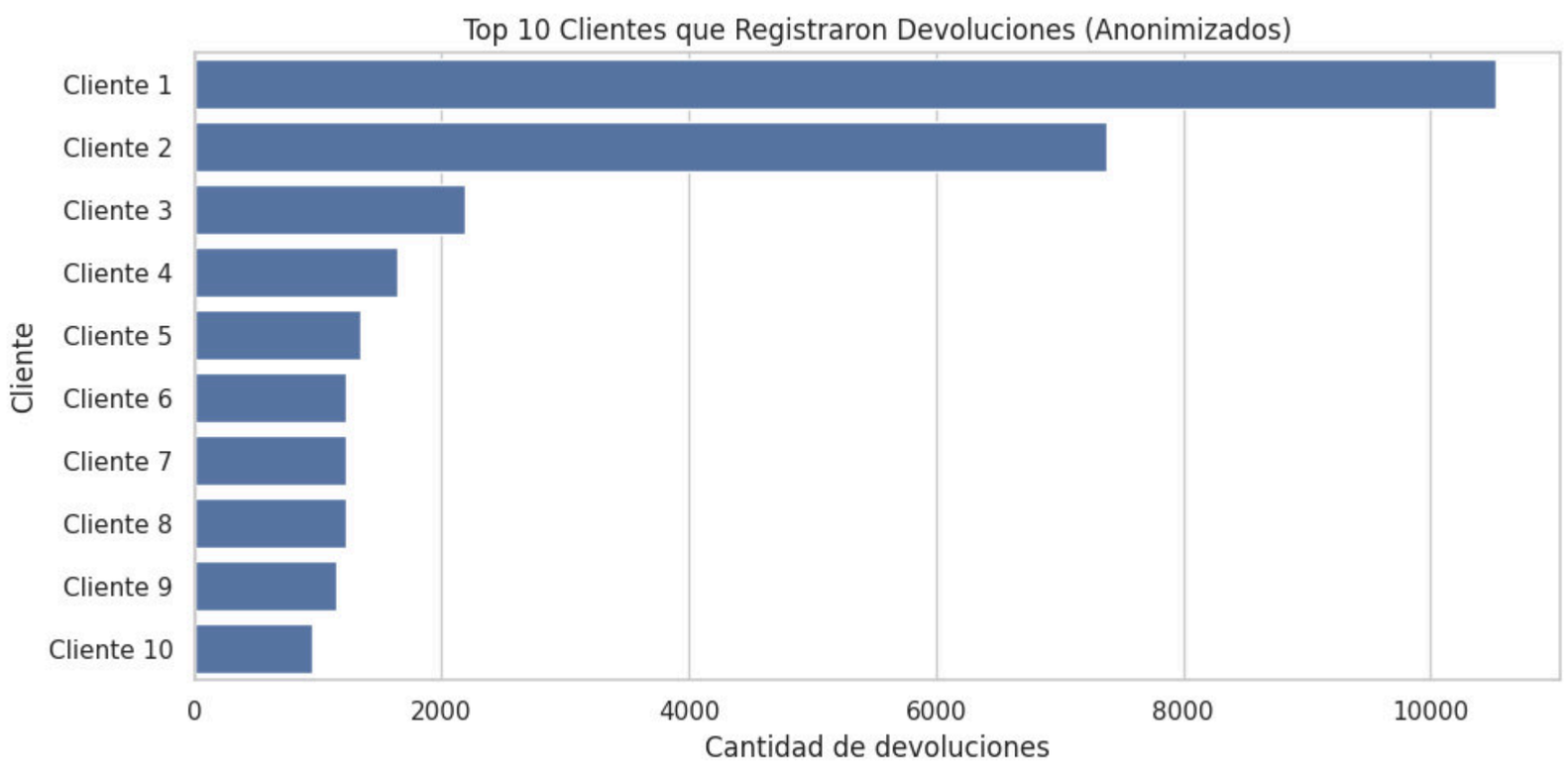


Gráfico 3: Top 10 Materiales Devueltos

Anonimato

```
plt.figure(figsize=(10, 5))

# Obtener los 10 materiales más frecuentes
materiales = df['Descripción del material_x'].value_counts().head(10)

# Crear etiquetas anónimas: Material 1, Material 2, ...
anon_labels = {name: f"Material {i+1}" for i, name in enumerate(materiales.index)}

# Dibujar el gráfico con etiquetas anonimizadas
sns.barplot(x=materiales.values, y=[anon_labels[name] for name in materiales.index])

plt.title('Top 10 Materiales Devueltos (Anonimizados)')
plt.xlabel('Cantidad')
plt.ylabel('Material')
plt.tight_layout()
plt.show()
```

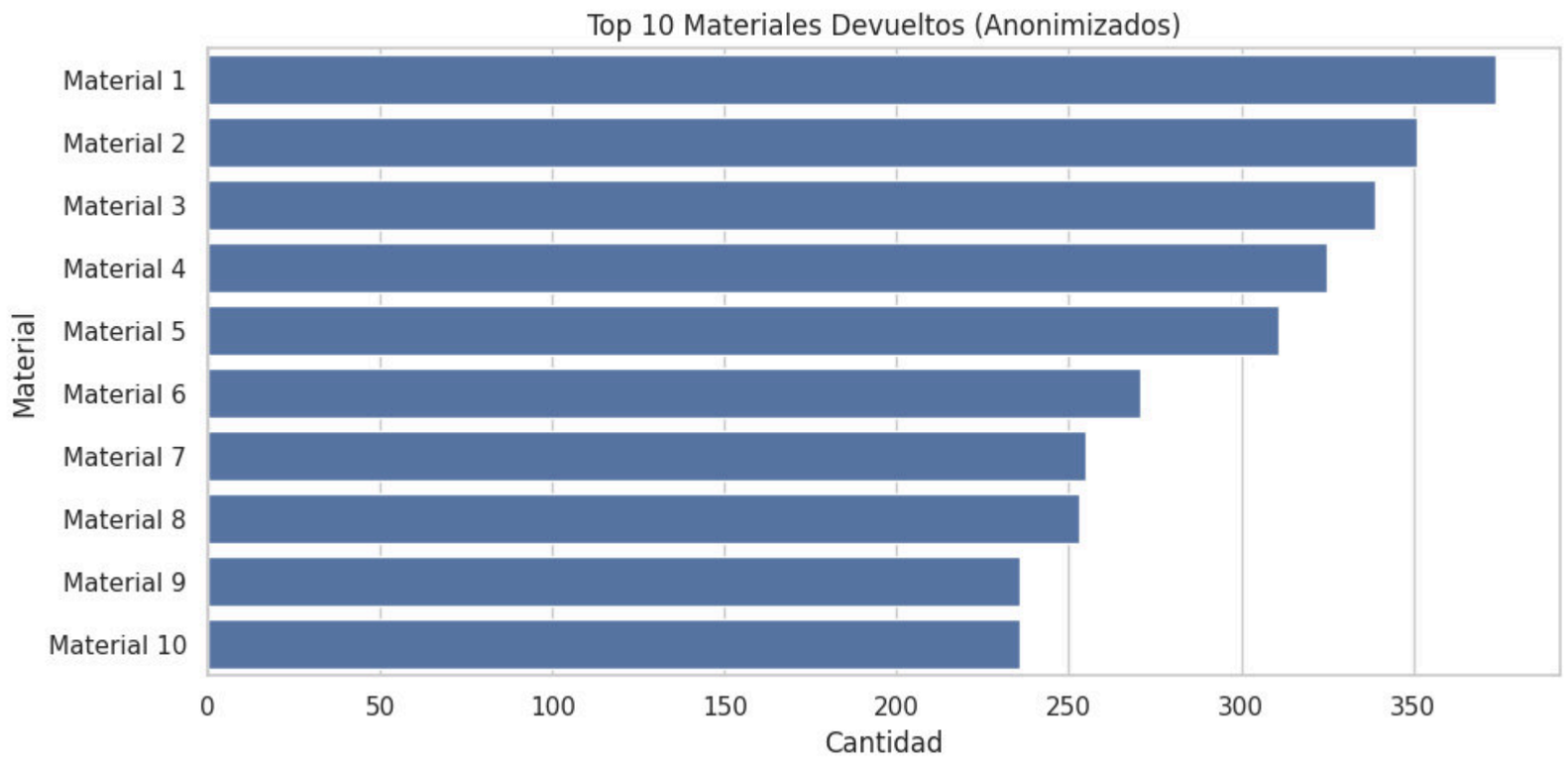


Gráfico 4: Top 10 Lotes con más Devoluciones

```
plt.figure(figsize=(10, 5))
lotes = df['Lote'].value_counts().head(10)
sns.barplot(x=lotes.values, y=lotes.index)
plt.title('Top 10 Lotes con más Devoluciones')
plt.xlabel('Cantidad')
plt.ylabel('Lote')
plt.tight_layout()
plt.show()
```

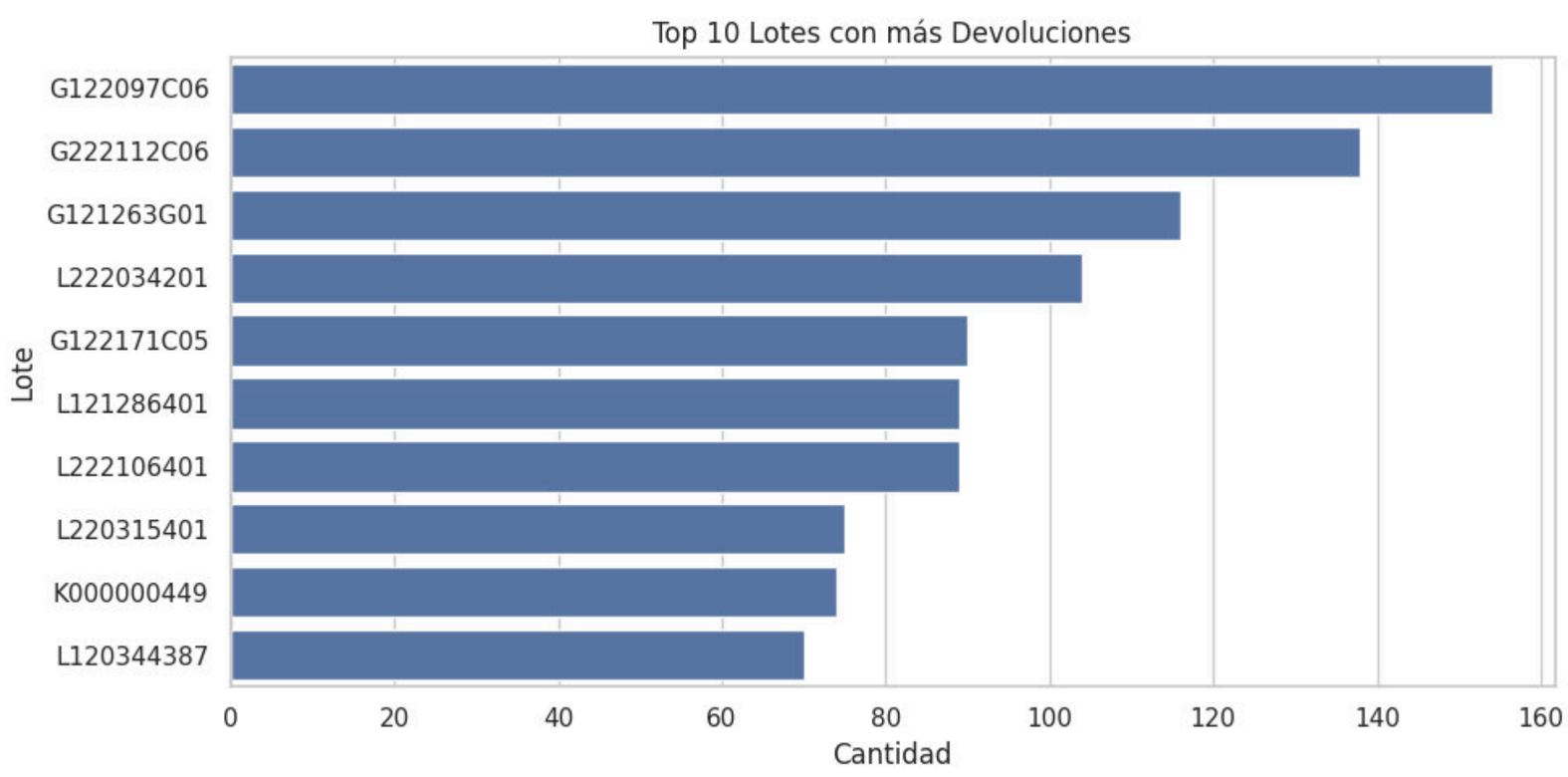


Gráfico 5: Evolución Mensual de Devoluciones

```
evolucion = df.groupby('Mes').size()
plt.figure(figsize=(12, 6))
evolucion.plot(kind='line', marker='o')
plt.title('Evolución Mensual de Devoluciones')
plt.xlabel('Mes')
plt.ylabel('Número de devoluciones')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

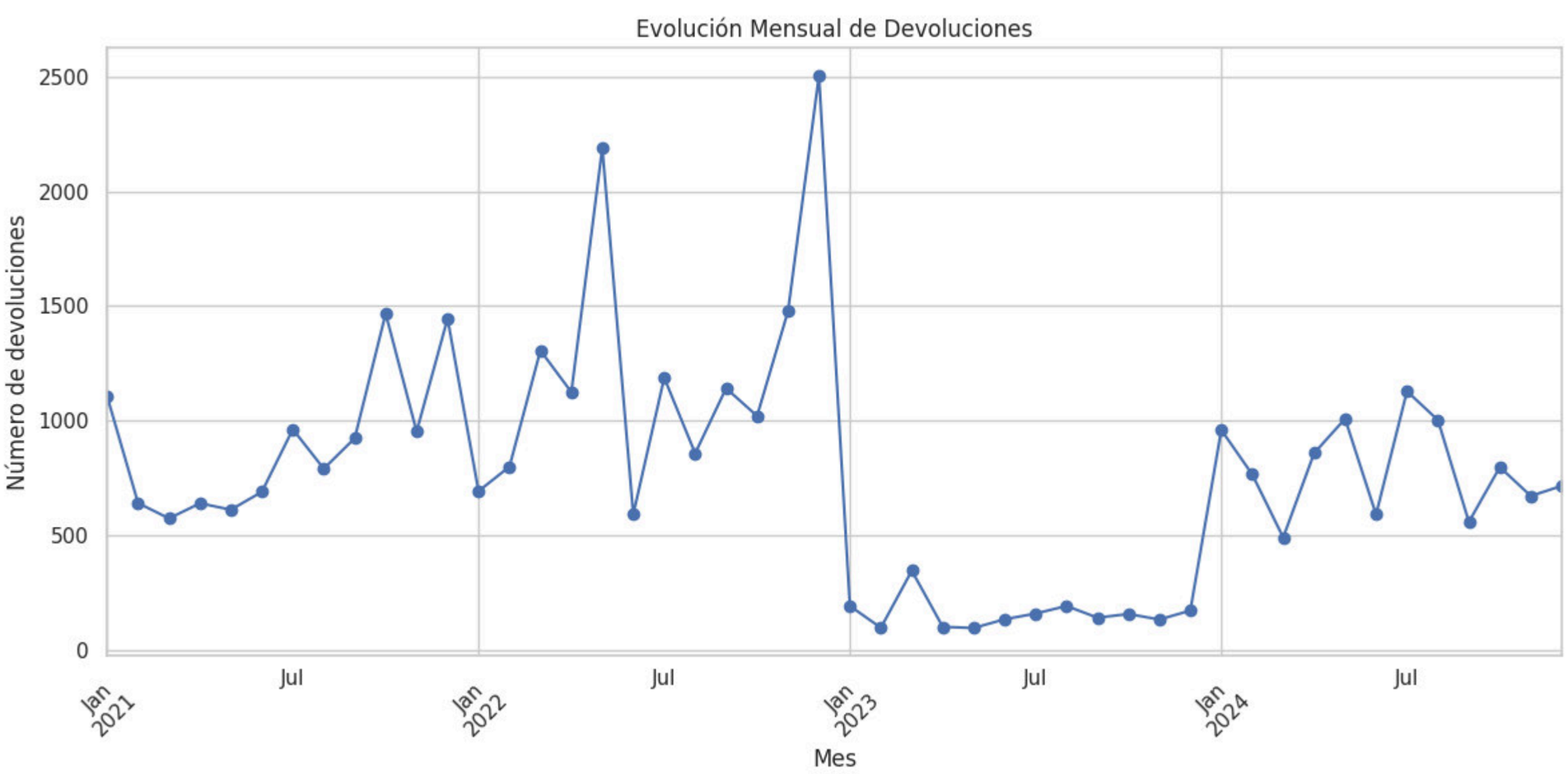


Gráfico 6: Usuarios que Registraron Devoluciones

Anonimato

```
plt.figure(figsize=(10, 5))

# Obtener los 10 usuarios más frecuentes
creado = df['Creado por_x'].value_counts().head(10)

# Crear etiquetas anónimas: Usuario 1, Usuario 2, ...
anon_labels = {name: f"Usuario {i+1}" for i, name in enumerate(creado.index)}

# Dibujar gráfico con nombres anonimizados
sns.barplot(x=creado.values, y=[anon_labels[name] for name in creado.index])

plt.title('Top 10 Usuarios que Registraron Devoluciones (Anonimizados)')
plt.xlabel('Cantidad')
plt.ylabel('Usuario')
plt.tight_layout()
plt.show()
```

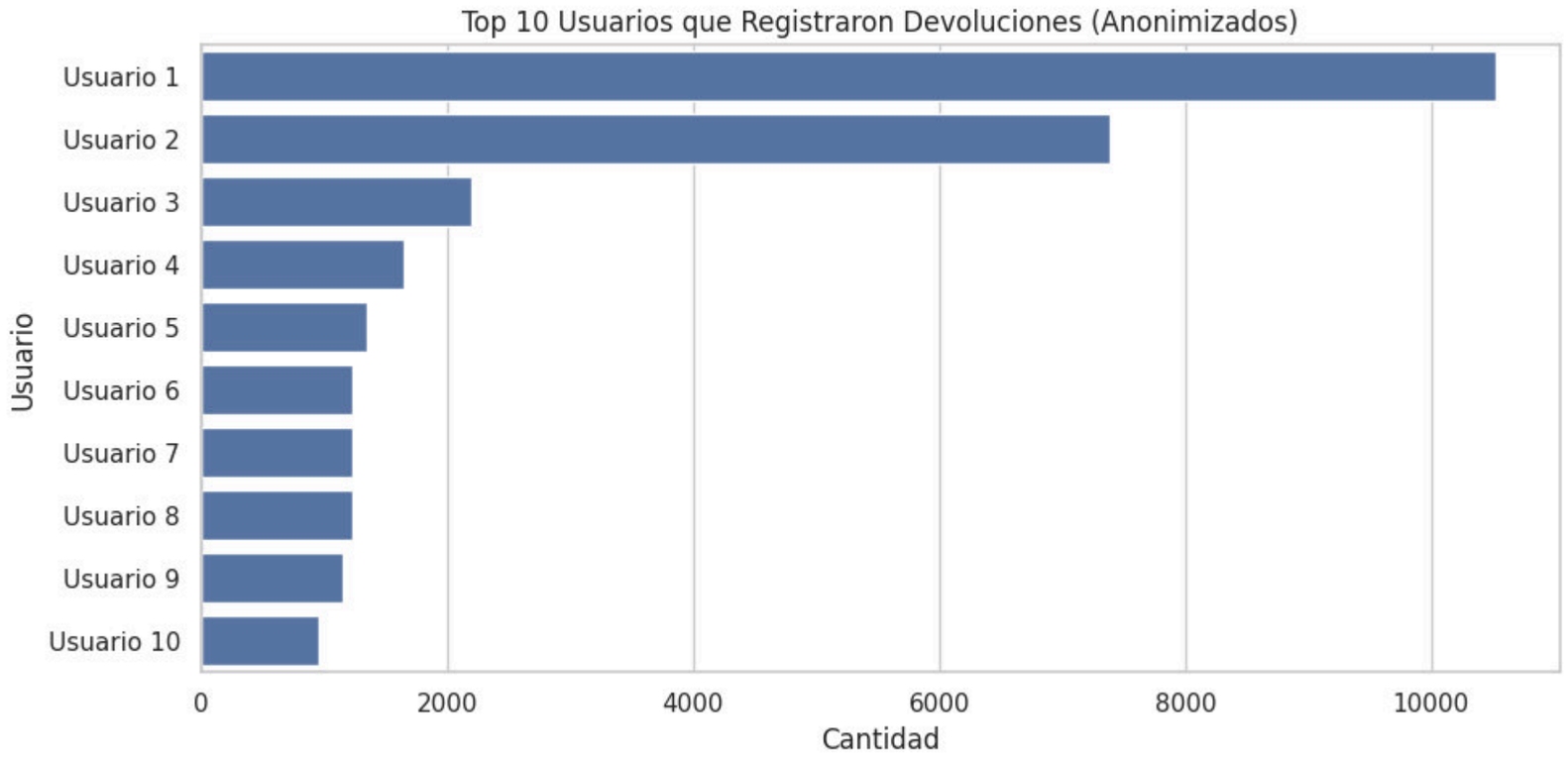


Gráfico 7: Almacenes con más Devoluciones

```
plt.figure(figsize=(10, 5))
almacen = df['Almacén'].value_counts().head(10)
sns.barplot(x=almacen.values, y=almacen.index)
plt.title('Top 10 Almacenes con más Devoluciones')
plt.xlabel('Cantidad')
plt.ylabel('Almacén')
plt.tight_layout()
plt.show()
```

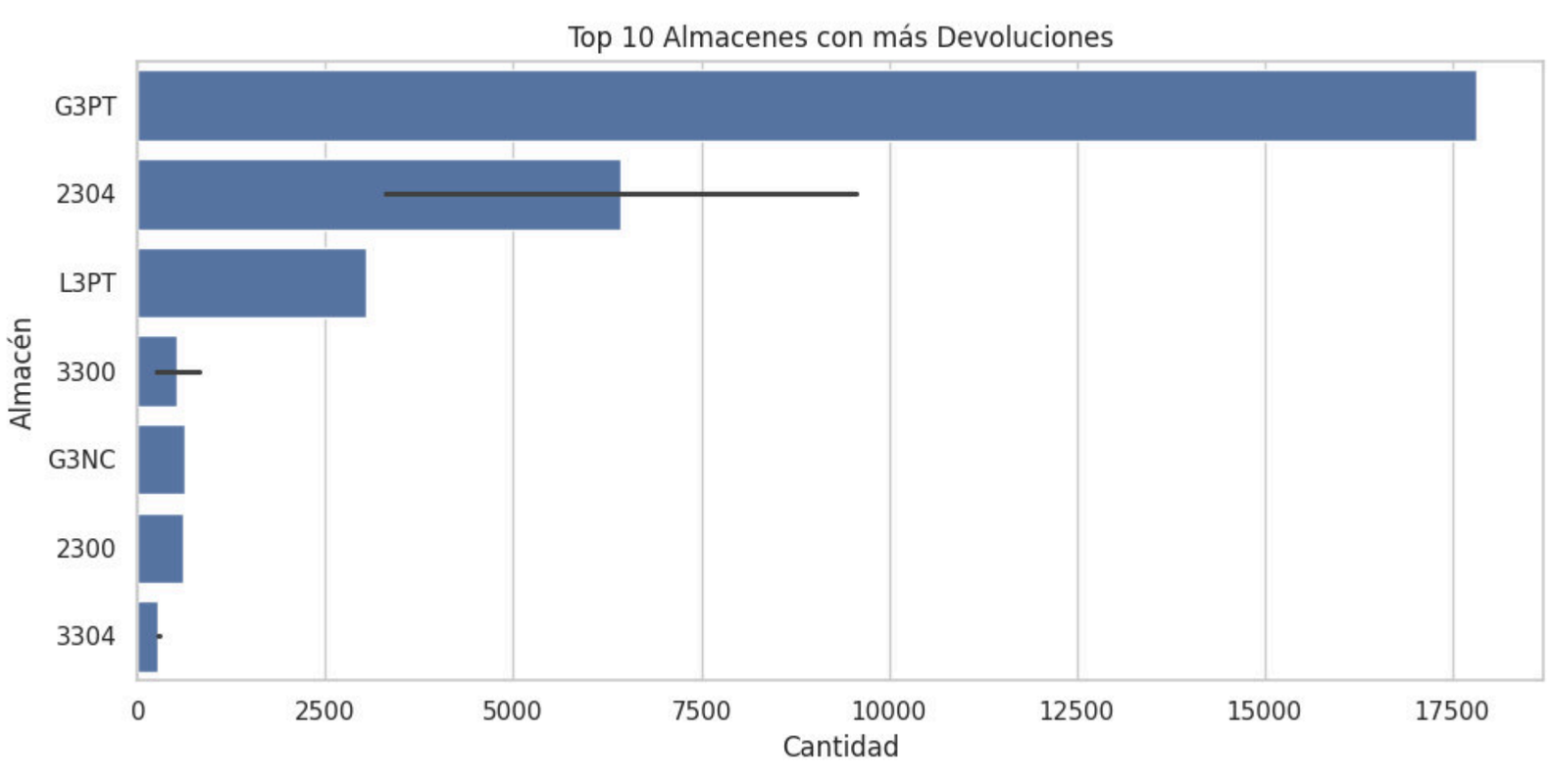


Gráfico 8: Centros con más Devoluciones

```
plt.figure(figsize=(10, 5))
centro = df['Centro_x'].value_counts().head(10)
sns.barplot(x=centro.values, y=centro.index)
plt.title('Top 10 Centros con más Devoluciones')
plt.xlabel('Cantidad')
plt.ylabel('Centro')
plt.tight_layout()
plt.show()
```

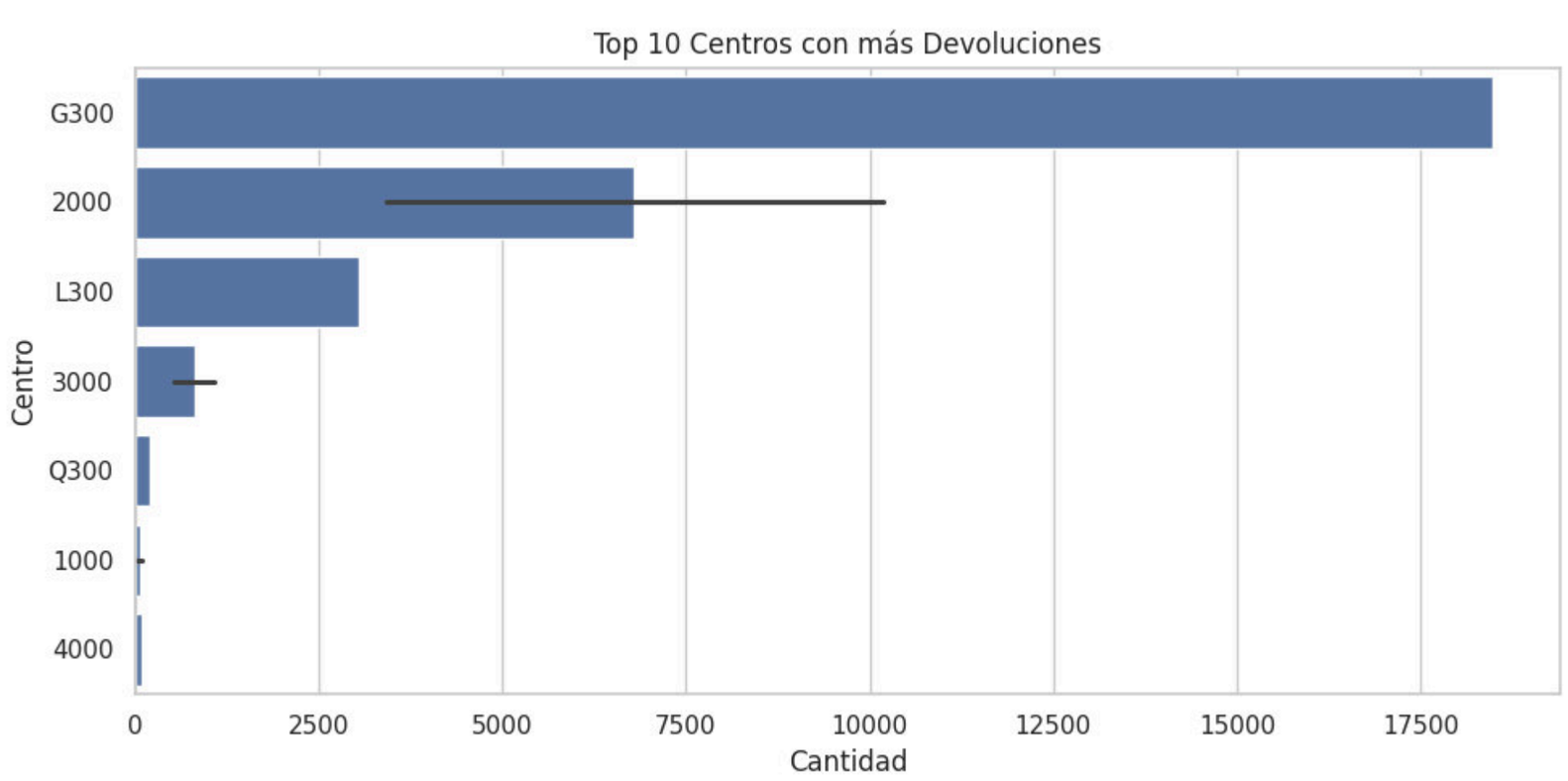


Gráfico 9: Interlocutores con más Devoluciones

Anonimato

```
plt.figure(figsize=(10, 5))

# Obtener los 10 interlocutores más frecuentes
interlocutor = df['Nombre interlocutor'].value_counts().head(10)

# Crear etiquetas anónimas: Interlocutor 1, Interlocutor 2, ...
anon_labels = {name: f"Interlocutor {i+1}" for i, name in enumerate(interlocutor.index)}

# Dibujar gráfico con etiquetas anonimizadas
sns.barplot(x=interlocutor.values, y=[anon_labels[name] for name in interlocutor.index])

plt.title('Top 10 Interlocutores con más Devoluciones (Anonimizados)')
plt.xlabel('Cantidad')
plt.ylabel('Interlocutor')
plt.tight_layout()
plt.show()
```

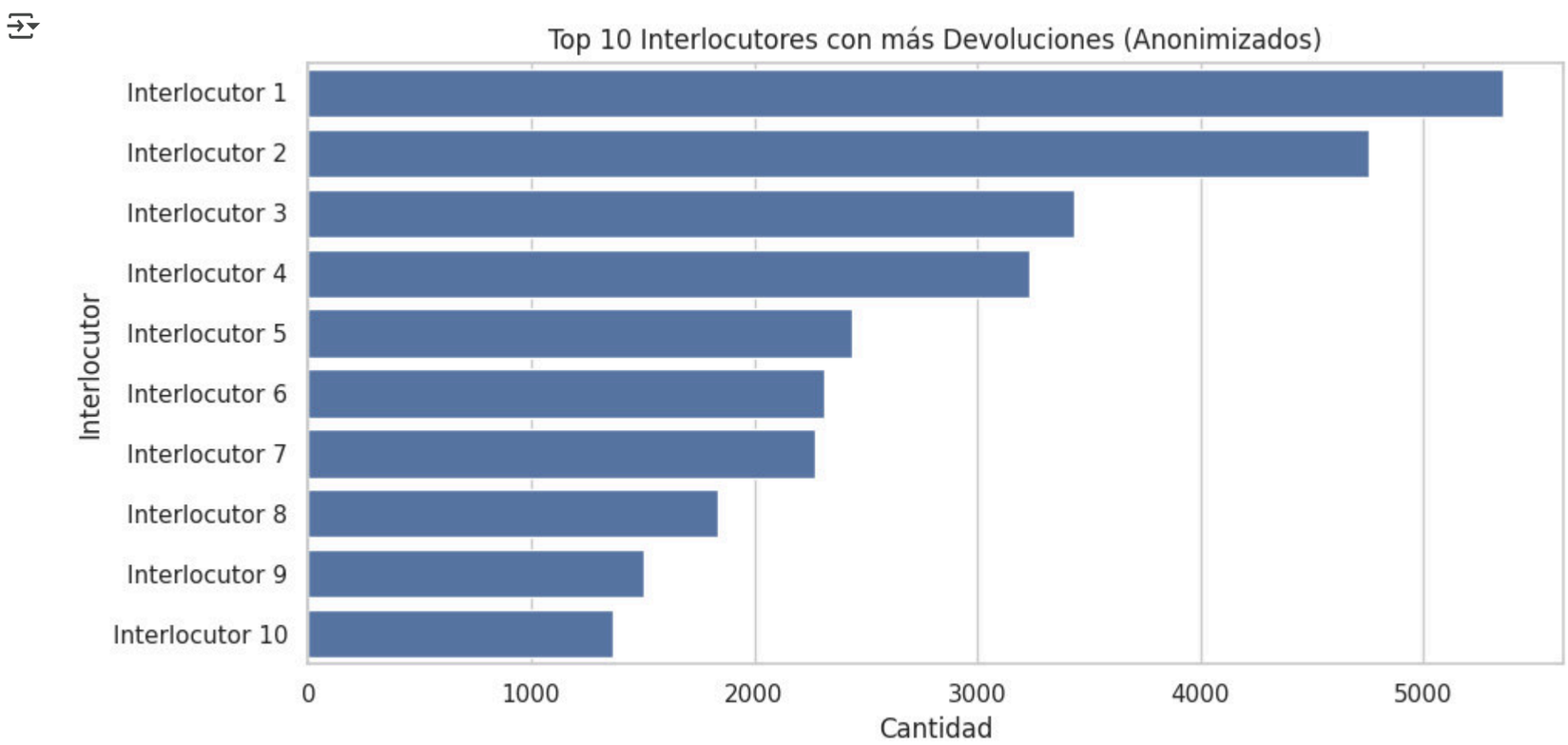


Gráfico 10: Sectores con más Devoluciones

```
plt.figure(figsize=(10, 5))
sectores = df['Nombre sector'].value_counts().head(10)
sns.barplot(x=sectores.values, y=sectores.index)
plt.title('Top 10 Sectores con más Devoluciones')
plt.xlabel('Cantidad')
plt.ylabel('Sector')
plt.tight_layout()
plt.show()
```