

Análisis del comportamiento e identificación de patrones en el valor de las donaciones y la destrucción de alimentos, en una empresa del sector cárnico mediante técnicas de ciencia de datos

Luisa Fernanda Jiménez Ramírez

Maestría en Ingeniería – Universidad de Antioquia

luisa.jimenez1@udea.edu.co

Resumen—Para las compañías de alimentos, uno de los principales desafíos es el desperdicio de alimentos, que, por su triple impacto económico, social y ambiental, resulta pertinente abordarlo desde una perspectiva global. Este trabajo analiza la evolución del valor de las donaciones y la destrucción de alimentos en una empresa cárnica desde 2022, identificando patrones en las variables históricas que expliquen su comportamiento y apoyen la creación de modelos. Para cumplir este objetivo se utiliza la técnica de ciencia de datos CRISP-DM (Cross Industry Standard Process for Data Mining) [1], compuesta por 6 fases, sin embargo, para el alcance de este trabajo se realizará hasta la fase 3: entendimiento del negocio, entendimiento de los datos y preparación de los datos. La aplicación de CRISP-DM permitió identificar problemas ocultos en los datos, como destrucciones con valores negativos y productos con códigos “0”, lo que evidencia la necesidad de mejorar la calidad de la información. El uso de Cramer’s V redujo la dimensión de variables categóricas y Box-Cox mejoró la distribución de continuas. Aunque el análisis enfrentó limitaciones por el tamaño y naturaleza de la base, se obtuvieron transformaciones útiles y líneas futuras, como explorar una respuesta categórica y nuevas estrategias para comprender mejor la data de destrucciones.

Palabras Clave: Ciencia de Datos, Análisis de Datos, CRISP-DM (Cross-Industry Standard Process for Data Mining), Donación de Alimentos, Pérdida y Desperdicio de Alimentos.

I. INTRODUCCION

La problemática de las pérdidas y desperdicios de alimentos (PDA) es crítica a nivel global y nacional, con impactos económicos, sociales y ambientales significativos. En Colombia, se desperdician grandes cantidades de alimentos mientras más de la mitad de la población enfrenta inseguridad alimentaria, generando además emisiones contaminantes y pérdidas económicas del 1,3% del PIB [2]. Las causas son múltiples: desde ineficiencias en la producción, distribución y consumo, hasta vacíos normativos y logísticos. Aunque el país ha avanzado en legislación (Ley 1990 de 2019 [3] y Ley 2380 de 2024 [4]), los incentivos fiscales no se corresponden con la limitada capacidad logística, que solo permite recuperar el 12% del potencial alimentario disponible [5]. Esto evidencia la urgencia de transformar la cadena de recuperación alimentaria mediante modelos logísticos más eficientes, coordinados y adaptados al nuevo marco legal.

Actualmente, las empresas están apostando por políticas y medidas que contribuyen a reducir los riesgos medioambientales, causados directa o indirectamente por su actividad comercial. Para las compañías de alimentos, uno de los principales desafíos es el desperdicio de alimentos, que por su triple impacto; económico, social y ambiental, resulta pertinente abordarlo desde una perspectiva global. En dónde, las donaciones son un medio que permite alargar el ciclo de vida de los alimentos, asegurando que lleguen al público objetivo desde el punto de vista social.

Esta investigación está enmarcada en una empresa del sector de alimentos cárnicos, actualmente cuenta con una política de pérdida y desperdicio de alimentos (PDA), que internamente se denomina desguace, a lo largo de su cadena de abastecimiento. Esta política ha ayudado a entender el proceso y controlarlo por medio de indicadores clave. Se sabe que la PDA está clasificada en tres tipos, descritos a continuación:

Destrucción: se refiere a todos los alimentos que no son aptos para el consumo humano, que pierden todas sus cualidades organolépticas, por lo que se disponen por medio de la incineración. La decisión de disponer de estos alimentos es por medio de su fecha de vencimiento.

Reelaboración: aquí son clasificados todos los alimentos que, si bien no son aptos para la venta al público por diferentes motivos, por ejemplo, imperfecciones estéticas, abolladuras, etc. Cuentan con las características organolépticas para el consumo y se decide reincorporarlos al proceso productivo para evitar incinerarlos. Y hacen parte del indicador de PDA ya que no cumplen con el objetivo principal de ser vendidos como producto inicial.

Donación: son todos aquellos productos que, si bien no son aptos para la venta al público en general, por diferentes motivos, principalmente es por la fecha de caducidad. Aún cuentan con un tiempo disponible para su consumo, es decir, están en una ventana de tiempo en dónde la fecha de caducidad es muy corta para cumplir con todo el proceso de venta y consumo, sin embargo, no están vencidos y pueden ser donados a entidades encargadas de distribuirlos al público objetivo.

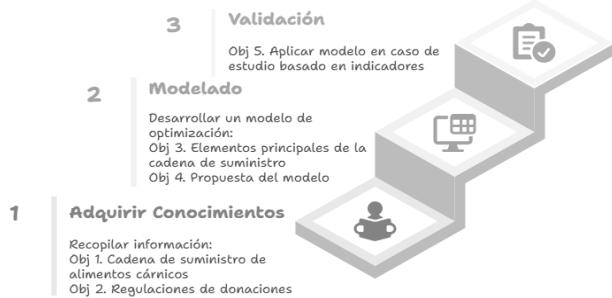
En la compañía, la política se enfocó en disminuir la destrucción de los alimentos, ya que es la clasificación que impacta de forma directa al rendimiento económico, por su parte, la reelaboración no pierde el 100% de su costo, ya que es reincorporada al proceso. En cuanto a las donaciones, actualmente son percibidas como una pérdida total, ya que no generan ningún valor monetario, más allá de la implicación

ambiental y social que motiva a la empresa. A la luz de las políticas de alivios fiscales para las donaciones postuladas por el gobierno, este proyecto se enfoca en esta clasificación. Lo trascendental es que la existencia de alivios fiscales de gran envergadura puede ayudar a modificar esta creencia, ya que el proceso de donación tendría un incentivo y no una pérdida del 100%. Adicionalmente contando con un plus, ya que las donaciones contribuyen directamente con propósitos sociales y ambientales.

El proyecto principal cuenta con tres etapas en la metodología, en primera medida se debe adquirir los conocimientos en cuanto a la cadena de abastecimiento y las regulaciones, en cuanto a la segunda etapa se desarrollará un modelo de optimización y finalmente se validará este modelo con simulaciones, como se puede evidenciar en la “Figura. 1”. Para la segunda etapa en el desarrollo del modelo de optimización, que es dónde se va a ejecutar y analizar toda la cadena de abastecimiento para la toma de decisiones, es necesario contar con la entrada de la data (instancias) previamente organizada y analizada. Y es en este punto dónde surge la necesidad de organizar, limpiar y analizar la data que ingresará al modelo. De esta forma, el objetivo de este estudio es: Analizar el comportamiento del valor de las donaciones y de la destrucción de alimentos en una empresa del sector cárnico, identificando patrones asociados a las variables que componen la data histórica desde el 2022, que permitan comprender su dinámica y servir como base para el desarrollo de modelos predictivos. Para lograr este objetivo general, se cuenta con cinco objetivos específicos:

- Recolectar y depurar la base de datos histórica de donaciones y del desperdicio de alimentos de la empresa del sector cárnico desde el año 2022, garantizando la calidad, consistencia y completitud de la información.
- Describir las variables relevantes asociadas al valor de las donaciones y del desperdicio de alimentos, diferenciando entre variables categóricas y numéricas, para comprender su estructura y posibles relaciones.
- Aplicar técnicas de análisis exploratorio de datos (EDA) para identificar tendencias, patrones y comportamientos significativos en el valor de las donaciones y del desperdicio de alimentos a lo largo del tiempo.
- Detectar y analizar valores atípicos en la base de datos para mejorar la calidad de la información, reducir sesgos y garantizar mayor confiabilidad en los resultados del procesamiento y modelado posterior.
- Estandarizar el conjunto de datos mediante la imputación de valores faltantes, el escalamiento de variables y la aplicación de transformaciones adecuadas, asegurando un preprocesamiento que mejore la interpretabilidad y desempeño de los modelos analíticos posteriores.

Figure 1. Etapas metodología proyecto macro. (*Elaboración propia*)



El enfoque metodológico del proyecto se fundamenta en los principios de la ciencia de datos, orientados al análisis cuantitativo y al descubrimiento de patrones a partir de información estructurada. Se adopta un enfoque cuantitativo de tipo descriptivo y exploratorio, ya que el propósito principal es analizar el comportamiento de las variables contenidas en la base de datos e identificar relaciones o patrones significativos que expliquen el fenómeno de estudio. Este enfoque permite aplicar herramientas estadísticas, técnicas de limpieza y transformación de datos, así como métodos de visualización y modelado, que facilitan la comprensión objetiva y basada en evidencia del comportamiento de los datos.

Además, al tratarse de un proyecto en el marco de Fundamentos de Ciencia de Datos, este enfoque resulta pertinente porque favorece el desarrollo de competencias analíticas y metodológicas en la manipulación de datos, la formulación de hipótesis, la validación empírica de resultados y la interpretación de hallazgos para la toma de decisiones o la construcción de modelos predictivos.

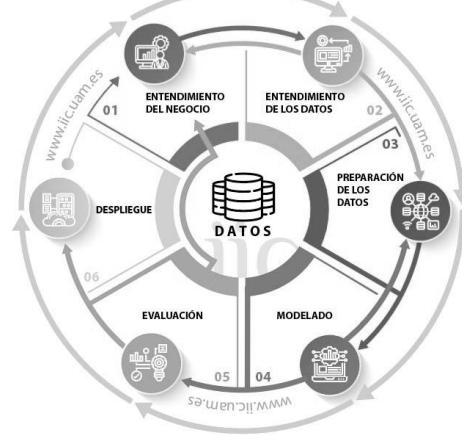
Para el respectivo análisis, la data obtenida tiene un histórico desde el año 2022, que constituye toda la trazabilidad de alimentos destruidos (incinerados) por diferentes causas en la compañía. Esta data madre tiene un campo llamado "Tipo de desguace" que tiene tres clasificaciones: "destrucción", "reelaboración" y "donaciones". En vista de que el trabajo está enfocado en las donaciones y las destrucciones, se omite toda la información relacionada con reelaboración. Para términos legales y de información confidencial, se generó un documento de confidencialidad por parte de la empresa, en donde no es posible utilizar su nombre propio para publicaciones o trabajos, sin embargo, para términos académicos el uso de sus datos es con total transparencia.

II. METODOLOGÍA

La herramienta utilizada que guía este proyecto, que contribuye al alcance del objetivo general, es basada en la metodología CRISP-DM, que cuenta con seis etapas detalladas en la “Figura. 2”, para el alcance de este proyecto, de desarrollar hasta la Fase tres “Preparación de los datos”, en donde los datos se encuentran completamente listos para el uso en modelos analíticos o de machine learning. El proyecto se encuentra implementado con la herramienta de Python y publicado en un [repositorio](#) de GitHub, llamado “Introducción_Ciencia_De_Datos” [6], este proyecto usó

como guía un repositorio de la clase “Introducción a la Ciencia de Datos” En este repositorio hay una carpeta nombrada “proyecto_aula”, que se encuentra organizada en cinco diferentes archivos, en los que su contenido y desarrollo se describe a continuación:

Figure 2. Fases de la metodología CRISP-DM. (Imagen tomada de ii.uam.es)



A. 01_Introduccion_Carga_luisa_jimenez

En este notebook se encuentra una introducción del problema, los objetivos que se quieren alcanzar, las librerías necesarias para correr el código, en este caso *pandas* y se cargan los datos. Los datos se obtienen de la empresa de alimentos cárnicos, por medio de una figura llamada “Proyecto U”, en dónde la empresa genera un vínculo con la academia para resolver problemas internos, este vínculo se crea por medio de un líder de la empresa que ayuda con la información o dudas alrededor del problema. En este caso, el líder es del área de analítica y su función es ser el puente ante dudas, avances e información requerida por parte del proyecto de maestría. Este notebook corresponde a la Fase 1, del entendimiento del negocio, dónde se alinean los objetivos del negocio con la ayuda que puede ofrecer la ciencia de datos.

B. 02_Descripcion_Limpieza_luisa_jimenez

La Fase 2 de la metodología “Entendimiento de los datos” está desarrollada en este notebook, en el que se utilizan las librerías *pandas*, *unicodedata* y *numpy*. En dónde se realiza un proceso de depuración de variables por medio de dos fases, en dónde la primera cumple el objetivo de eliminar variables redundantes, duplicadas, etc., y la segunda fase se basa en el entendimiento del problema y a partir de allí, se seleccionan las variables de interés.

Posterior a la identificación de las variables de interés, se realiza la limpieza de los data, por medio de la unificación y estandarización del dataframe, toda la data se encuentra en minúsculas, sin tildes y se modifican los nombres de las variables, adicionalmente se reasigna el tipo de variable según corresponda y se unifican los nombres de algunas variables con espacios u otro tipo de símbolo. Finalmente, todo este proceso se guarda en un archivo tipo csv nombrado “df_limpio”, que se utilizará como insumo para la siguiente fase.

C. 03_AnalisisEDA_luisa_jimenez

Con respecto a la Fase 3 de la metodología que corresponde a la “Preparación de los datos”, se desarrollan los tres notebooks siguientes con el actual inclusive, es decir: *03_AnalisisEDA*, *04_Atipicos* y *05_Imputacion_Escalamiento_Transformacion*. Para el correcto desarrollo de este notebook se utilizaron las siguientes librerías:

- import pandas as pd
- from scipy import stats
- import os
- import pandas as pd
- import numpy as np
- from scipy.stats import chi2_contingency
- import seaborn as sns
- import matplotlib.pyplot as plt
- import pingouin as pg
- from sklearn.ensemble import IsolationForest
- from sklearn.preprocessing import StandardScaler
- from sklearn.cluster import KMeans
- from sklearn.metrics import silhouette_score
- from tqdm import tqdm
- import matplotlib.dates as mdates
- from sklearn.feature_selection import mutual_info_regression
- from scipy.stats import pearsonr

En primera medida se realiza un análisis exploratorio general para saber los datos faltantes de todas las variables y los estadísticos descriptivos (promedio, mínimo, máximo, varianza, desviación estándar, moda) tanto de la data en general como de la data analizada individualmente entre donaciones y destrucciones.

Se procede al análisis univariado para las variables continuas y categóricas. Para las variables continuas, se realizan análisis de medidas de tendencia central graficando el comportamiento de las datas individuales de las donaciones y las destrucciones. En el caso de las variables categóricas primero se procede a realizar una estandarización asegurando que no haya categorías duplicadas o escritas diferente, y con la ayuda de los histogramas se analiza cada variable categórica de forma global e individualmente por el tipo de PDA, donaciones y destrucciones. Adicionalmente, se complementa el análisis, graficando en los histogramas el valor de las donaciones y las destrucciones por planta y centro de distribución, con las causales con mayor participación. Y finalmente, se realiza un análisis temporal, para analizar el comportamiento del PDA en general, las donaciones y las destrucciones a lo largo del tiempo.

Con respecto al análisis bivariado, por medio de una matriz de dispersión y valores de correlación, se analizan las variables continuas “desperdicio_kg”, “peso_kg” y “cantidad” y su relación con la variable respuesta “valor”, para los tipos de PDA, las donaciones y las destrucciones, complementado con un análisis de correlación de Pearson. Y para el caso de las variables categóricas, se realiza un estudio de correlación de Cramer’s V.

Finalmente, para el análisis multivariado, se implementa un clúster con K-Means, ya que para PCA solamente se cuenta con cuatro variables continuas y reducir su

dimensionalidad no contribuye en gran medida, y en el caso de Regresión lineal múltiple, más adelante en la sección de resultados se evidencia que los datos incumplen al menos uno de los supuestos, lo que impide su aplicación. En este sentido, el clúster se aplica para el tipo de PDA de donaciones y de destrucciones, guardando los cambios realizados en un data frame como archivo csv con el nombre de “df_nclean”, para utilizarlo en el siguiente notebook.

D. 04_Atipicos_luisa_jimenez

Para continuar con el proceso de preparación de los datos, en este notebook de detección y análisis de datos atípicos son necesarias las siguientes librerías:

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- from sklearn.cluster import DBSCAN
- from sklearn.preprocessing import StandardScaler

Útiles para graficar box-plots de cada una de las cuatro variables continuas inclusive la variable respuesta “valor”, distinguidas por el total de observaciones y de forma individual por el tipo: destrucción y donaciones. Para complementar este análisis, se grafica el método DBSCAN de la variable respuesta “valor” por cada una de las variables continuas, de forma global y detallado entre donaciones y destrucciones.

E.05_Imputacion_Escalamiento_Transformacion_luisa_jimenez

La Fase de preparación de los datos finaliza con este notebook de imputación, escalamiento y transformación de las variables que así lo requieran, en dónde es útil la importación de las siguientes librerías:

- import pandas as pd
- import seaborn as sns
- import matplotlib.pyplot as plt
- # Instalar librerías necesarias
- !pip install pingouin seaborn
- !pip install pandas numpy seaborn matplotlib scikit-learn openpyxl --quiet
- import numpy as np
- import pingouin as pg
- from sklearn.feature_selection import mutual_info_regression
- from scipy import stats
- from scipy.stats import skew, kurtosis
- from scipy.stats import boxcox
- from sklearn.ensemble import IsolationForest
- from sklearn.preprocessing import StandardScaler

El notebook inicia con el proceso de imputación de valores “0” como número y como texto, en dónde no es

coherente que exista una categoría de este tipo, se realiza por medio de la estrategia MAR — Missing at Random, ya que a partir de la relación directa con una tercera variable se pretende cambiar los valores ceros, que no deberían estar en ceros, según la naturaleza de la variable. Después de evaluar el porcentaje de cada uno con respecto al total, el restante se procede a imputar con la moda, ya que no es significativo. Posteriormente, se procede a imputar los valores con la misma lógica anterior, MAR por medio de la variable “material”, ya que las características de cada material son únicas y deberían ser iguales en todos los casos. Y en este caso, como son variables de materiales diferentes no es lógico imputarlo por la moda, así que se valida el porcentaje con respecto al total para proceder a eliminar.

Para realizar la verificación en dónde la imputación no modifique drásticamente la distribución de los datos originales, se realza una validación en dos sentidos, primero se grafica la distribución de los datos antes y después de la imputación, y en un segundo momento se analizan los estadísticos antes y después de la imputación, complementándolo con la prueba chi-cuadrado.

Una vez se complete el proceso de imputación, se procede a realizar la transformación para las variables continuas y las variables categóricas. Para las variables continuas, se utilizan las trasformaciones logarítmica y Box-Cox, a continuación, se justifica no utilizar las otras transformaciones:

- Yeo-Johnson incluye valores negativos o nulos y no aplica en este caso.
- Raíz cuadrada es útil para conteos y en este estudio solamente la variable “cantidad” es de conteo, lo que dificultaría la interpretación de resultado al no estandarizar el uso de las transformaciones.
- Recíproca, aumentaría la variabilidad, ya que comprime los valores grandes y expande los valores pequeños.
- Cuadrática y logarítmica reflejada, corrigen la asimetría negativa y en este caso, es positiva.
- Logit, es útil para datos como proporciones o probabilidades, y en este estudio ninguna variable es de este tipo.

Lo que se espera es comparar los resultados de ambas transformaciones y así aplicar la transformación que mejores ajustes refleje.

Para las variables categóricas, se decide dividir en dos grupos la transformación:

- Variables con menos de 20 categorías, se transforman con One Hot Encoding, ya que es una técnica muy útil para variables nominales sin orden y con pequeñas categorías.
- Variables con más de 20 categorías, son transformadas con Count Encoding, porque tiene en cuenta el número de veces que aparece cada categoría sin generar sesgo.

De este modo, las otras estrategias de transformación no aplican por los siguientes motivos:

- Ordinal Encoding, las variables no siguen un orden jerárquico.
- Label Encoding, al tener variables con más de 100 categorías los label serían muy grandes, y podría encontrarse sesgo en el análisis.
- Binary Encoding, al tener variables con más de 100 categorías el número binario sería muy grande.
- Embeddings Categóricos, al tener variables con más de 100 categorías, no es lógico tener un número tan grande de “embeddings”.

Finalmente, en el momento que se cuente con todas las imputaciones y transformaciones aplicadas, se unifica en un data frame “df_final” guardado y exportado como csv, para su posterior uso en la siguiente Fase 4 de modelado.

En este apartado está detallado el paso a paso utilizado metodológicamente para el desarrollo de este proyecto, puntualizando en la ubicación de los archivos, su contenido, las herramientas utilizadas y los criterios tenido en cuenta al momento de elegir o no una herramienta.

III. RESULTADOS Y ANÁLISIS

En el marco de la analítica y la ciencia de datos, el estudio y comprensión de los resultados que arrojan las diferentes herramientas es crucial para un óptimo desempeño de los datos en un modelo. Al momento de ejecutar cada uno de los procedimientos detallados anteriormente en la metodología, se producen salidas que ayudan a conocer la naturaleza de los datos, las limitaciones y oportunidades para analizarlos a partir de su contexto y como consecuencia tomar decisiones para el correcto procesamiento. Esta sección se desarrolla con base en los apartados mencionados en la metodología, según corresponda con la naturaleza de esta, como sigue.

A. Descripción y Limpieza

La base de datos inicial está compuesta por un dimensionamiento de 42 variables y 1.814.611 registros, en dónde es necesario hacer una limpieza, se realizan dos depuraciones, en la primera depuración se obvian variables sin objetivo propio, por ejemplo, variables redundantes, duplicadas, etc. La segunda depuración se basa en el contexto y el entendimiento del problema, para seleccionar las variables que funcionan como insumo para las instancias del modelo de optimización. Y de esta forma se obtiene una data con 17 variables necesarias para el análisis, que se detallan en la “Tabla. 1”; y una variable número 18 “nombre_material” que solamente es necesaria en la fase de imputación, pero que no es imprescindible para el análisis del problema, y además cuenta con 1.814.611 registros. El proceso de depuración se encuentra en el archivo de Excel “estadísticas_descriptivas”, en este documento hay tres pestañas, la primera pestaña “Depuración_Conocimiento” dónde se encuentran las 42 variables y subrayadas en rojo las variables que se depuran y cada una con un comentario explicativo. En la segunda pestaña “Depuración_Literatura” se encuentra la data depurada del proceso anterior con 23 variables y subrayado en amarillo las variables que se deciden depurar, para finalmente obtener 17 variables de interés para el estudio.

TABLE I. DESCRIPCIÓN DE LAS VARIABLES DEL DATAFRAME. ELABORACIÓN PROPIA

Variable	Tipo	Descripción
1. Cantidad	Discreto	Número de productos de alimentos donados o destruidos
2. Valor	Discreto	Costo de productos de alimentos donados o destruidos
3. Peso_kg	Discreto	Peso de cada producto de alimentos donados o destruidos
4. Desperdicio_kg	Discreto	Peso de productos de alimentos donados o destruidos
5. Material	Categórica	Código de identificación de los productos que donaron o destruyeron
6. Motivo	Categórica	Razón por la cual se dona o destruye los productos
7. FechaFact#	Date	Fecha en la que se hizo la donación o la destrucción
8. Planta	Categórica	Código de la planta en dónde se dona o destruye el producto
9. Centro	Categórica	CEDI en dónde se dona o destruye el producto
10. Ofc#Ventas	Categórica	Oficina en dónde se dona o destruye el producto
11. Categoría	Categórica	Tipo de material (alimentos larga vida, carnes frescas, etc.)
12. Subcategoria	Categórica	Tipo de material (larga vida cárnicos, tajados, etc.)
13. Línea	Categórica	Línea de producción de los productos
14. Marca	Categórica	Marca de los productos
15. Ord/Ext	Categórica	Tipo de demanda de los productos
16. Causales NUEVOS	Categórica	Causal de donación o destrucción de los productos
17. Tipo de Desguace	Categórica	Identifica si es donación, destrucción o reelaboración de alimentos

En vista de que el propósito del estudio es analizar las destrucciones y las donaciones, se procede a eliminar el tipo de PDA reelaboración, que comprende 1.063 registros, ya que no aporta información relevante en el estudio, de esta forma, finalmente se cuenta con “df_limpio” de 17 variables y 1.813.548 registros.

B. Análisis EDA

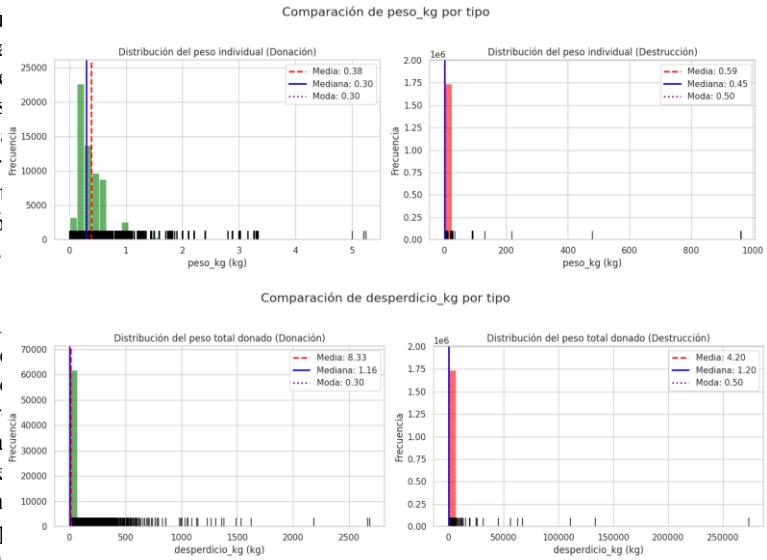
Análisis General: Con base en los resultados de los datos faltantes, la data de desperdicios contiene la gran mayoría, lo que podría significar falta de trazabilidad y de control de la calidad en este proceso. En el caso de los estadísticos descriptivos, se corrobora la hipótesis anterior con respecto a la data de destrucciones, ya que en todas las variables

continuas se evidencian datos negativos, lo que se encu-
ntra como extraño según la naturaleza de las varia-
bles. Adicionalmente, se evidencia que más del 96% de
los registros se encuentran tipificados como destrucciones.
Una varianza muy elevada en comparación con la tipifi-
cación de donaciones, dónde se cuenta con una varianza gr-
ande pero la información se encuentra más limpia y consis-
tente. De acuerdo con estos resultados, se toma la decisió-
n de eliminar los valores negativos, ya que son 741 registros
que aportan información al análisis.

Por último, en el análisis de la moda de todas las variables, se puede observar que para las variables “material” y “valor” la moda es “0”, lo que resulta extraño dado el significado de las variables, esto puede deberse a error en la digitación u otro tipo, ya que son digitaciones manuales subidas por el sistema SAP. Para este caso, se procede a analizar en detalle la variable “valor”, que por ningún motivo puede ser “0” y en cuanto a la variable “material”, se ajusta como pendiente para su respectivo análisis y tratamiento en la siguiente sección. Ya que los registros de “0” en “valor” es de 379 y representan menos del 1% se procede a eliminarlos, ya que podría interferir en análisis posteriores.

Análisis Univariado Continuas: Por medio de las gráficas de medidas de tendencia central que se encuentran ilustradas en la “Figura. 3”, se podría afirmar que, en las gráficas de cantidad, valor y donacion_kg, se confirma el sesgo que tienen y por ende la necesidad transformar las variables. Este comportamiento es lógico dentro del contexto, ya que se presenta una alta concentración de valores de donación muy pequeñas, es decir, que es poco probable que haya donaciones con valores muy altos, pero pueden existir, y por consecuencia, el peso debería tener este mismo comportamiento. En cuanto a las cantidades de las donaciones, se evidencia que es más común donar pocas cantidades que altas. Y por su parte en la gráfica de peso_kg se evidencia un comportamiento diferente, ya que se mide el peso de cada producto donado, en donde un peso muy bajo tiene la mayor cantidad de participación, entre 0 y 1 kg, es por esto por lo que la distribución se evidencia con una cola hacia la derecha, dónde es menos frecuente donar productos con pesos muy altos.

Figure 3. Medidas de tendencia central variables continuas. (*Elaboración propia*)



Ánalisis Univariado Categóricas: En primera medida por medio de los histogramas, se graficaron las frecuencias de cada variable y los resultados presentan una alta participación de donación por mala manipulación en el almacén o en el transporte, esto quiere decir que hay procesos en planta que pueden optimizarse y mejorar. Hay una congruencia en los resultados, ya que los alimentos de categoría y subcategoría “Larga vida” son elaborados en la planta de “La Ceja” y son los que cuentan con una participación mayor, así también como el tipo de demanda que en su mayoría son de “demanda larga vida”. Por otro lado, vemos que el Centro donde más se presentan donaciones es en “Comercial Nutresa Bogotá Op. Propia” y está muy relacionado con la Oficina de ventas con mayores devoluciones, que es “R Bogotá”. Es decir, que los alimentos que presentan una fecha de vencimiento más extensa son los que más se donan, elaborados en la Planta de “La Ceja” y distribuidos desde el CEDI de Comercial Nutresa en Bogotá. Adicionalmente, dentro del Top de productos hay uno con código “0”, que genera duda y se analizará en detalle más adelante.

Por su parte, las destrucciones tienen un comportamiento diferente, ya que en su mayoría son por el motivo de devoluciones, siendo las devoluciones otra base datos u otra información que impacta directamente, y el CEDI donde se generan más destrucciones es en el de Bogotá, que, si bien es diferente a las donaciones, es uno de los principales de la cadena de abastecimiento.

Como segundo análisis, la gráfica de temporalidad de las tres datas representada en la “Figura. 4”, se observa la diferencia en las proporciones entre donaciones y destrucciones, se podría afirmar que hay una oportunidad de disminuir esa brecha, aumentando el valor de las donaciones y disminuyendo el de las destrucciones. Los picos más altos se visualizan justamente en fechas de fin de año (2022 y 2023), que en el año 2024 se empezó a controlar. En los años 2023 y 2024 se puede observar un comportamiento embudo, de inicio de año valores altos y va disminuyendo a medida que avanza el tiempo. Adicionalmente se evidencia que desde el año 2024 las fluctuaciones han disminuido, generando un comportamiento más estable tanto en las donaciones, como en las destrucciones.

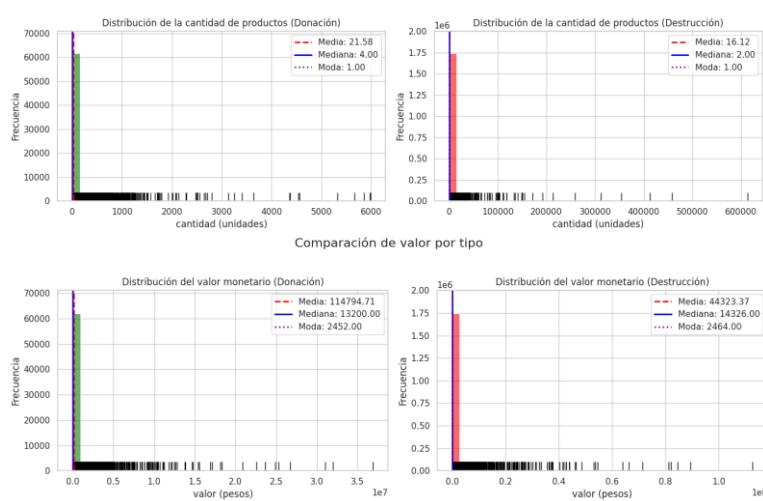


Figure 4. Comportamiento temporal del valor de PDA por tipo.
(Elaboración propia)



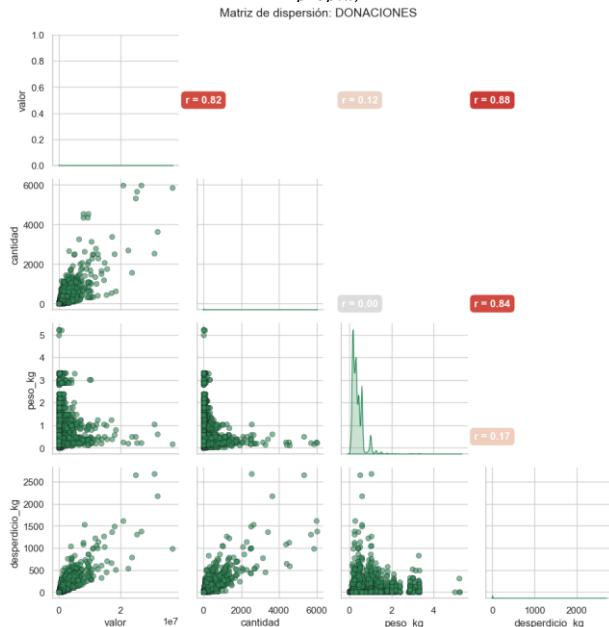
Análisis Bivariado Continuas: Por medio del gráfico de dispersión, se visualizan las correlaciones entre las variables, como se puede observar en la “Figura. 5”, allí se muestra una fuerte relación entre la variable respuesta que es "valor" con peso_kg y donacion_kg, ya que la cantidad de donaciones y el peso de las donaciones define el valor de la donación. Por el contrario, presenta baja relación con peso_kg, porque esta variable define el peso de cada unidad del producto, no el total del peso.

Por otro parte, se evidencia un comportamiento extraño en las gráficas de estimación de densidad de la variable respuesta "valor", "cantidad" y "donación", seguramente debido a un problema de escala o distribución de los datos, ya que, como se mencionó en el apartado de Descripción y Limpieza:

- Escalas enormemente diferentes: valor tiene media 114,794 vs peso_kg con media 0.38
- Distribuciones muy sesgadas: Valores máximos mucho mayores que las medianas
- Presencia de outliers extremos: valor máximo 36.9M vs mediana 13,200
- Muchos ceros: Especialmente en cantidad (mínimo = 0)

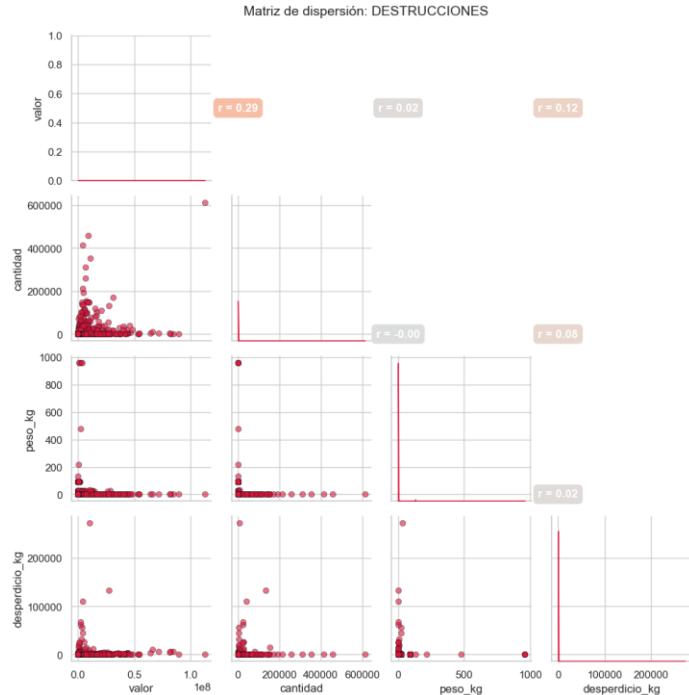
Es por esta razón que en el apartado de escalamiento y normalización es necesario ajustar estas variables.

Figure 5. Gráfico de dispersión por tipología donaciones. (Elaboración propia)



En el caso de las destrucciones, es crítico, ya que ninguna de las variables está relacionada, lo que indica que no hay relación lineal fuerte entre las variables. Las variables parecen independientes o solo ligeramente asociadas entre sí, ya que no hay pares de variables con correlación fuerte positiva o negativa. En cuanto al modelo, quiere decir que cada variable aporta información independiente, lo que puede ser útil si se quiere evitar multicolinealidad, pero limita predicciones basadas en relaciones lineales simples.

Figure 6. Gráfico de dispersión por tipología destrucciones. (Elaboración propia)



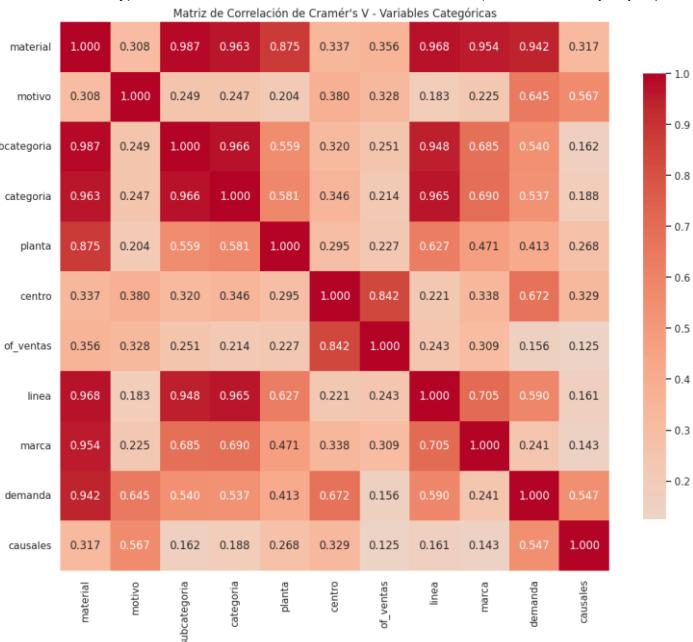
Para complementar este análisis, se utilizó la herramienta de correlación de Pearson en dónde, para el caso de las donaciones se identifican correlaciones muy fuertes y positivas entre tres variables clave: valor-cantidad ($r=0.82$), valor-donación_kg ($r=0.88$) y cantidad-donación_kg ($r=0.84$). Estas relaciones son estadísticamente significativas ($p<0.001$), indicando que el valor económico, el volumen de ítems y el peso de donaciones varían de forma coordinada. Ratificando lo evidenciado anteriormente. En caso contrario, el peso unitario (peso_kg) muestra escasa relación con las demás variables. Su correlación con cantidad es prácticamente nula ($r=0.002$) y estadísticamente no significativa ($p=0.58$), mientras que con valor y donación_kg presenta solo correlaciones débiles ($r=0.12$ y $r=0.17$ respectivamente), aunque significativas debido al gran tamaño muestral.

Este patrón sugiere que el valor de las donaciones se determina principalmente por el volumen y no por el peso individual de los ítems. La débil relación con peso_kg indica que los artículos donados probablemente incluyen mezclas de productos ligeros de alto valor y pesados de bajo valor, diluyendo cualquier patrón correlacional consistente. La consistencia de las correlaciones fuertes entre valor, cantidad y donación_kg sugiere que estas variables podrían servir como proxies entre sí para fines de estimación y planificación operativa.

En el caso de las destrucciones, la correlación más fuerte se observa entre valor y cantidad ($r \approx 0.28$), aunque sigue siendo débil, todas las demás correlaciones lineales son muy bajas ($r < 0.13$), lo que indica relaciones lineales casi nulas. Todos los valores p son prácticamente 0, lo que refleja alta significancia estadística debido al tamaño masivo de la muestra (~1.8 millones de registros). Es decir que las variables son casi independientes, y relaciones lineales simples no explican variaciones significativas entre ellas.

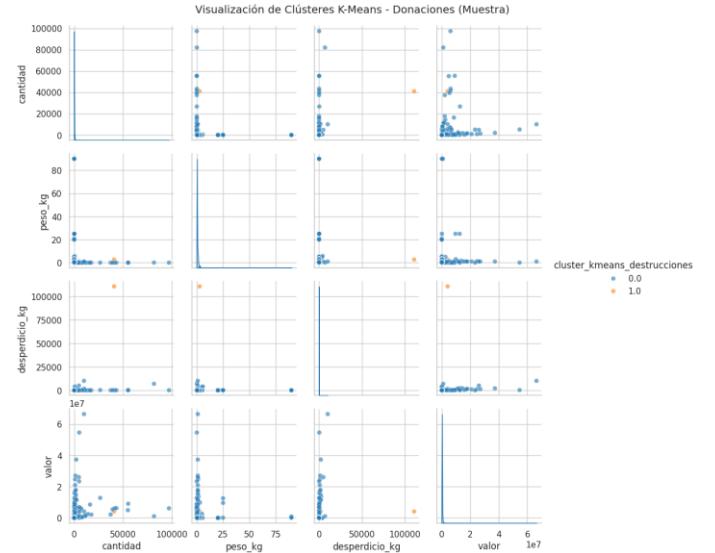
Análisis Bivariado Categóricas: Haciendo uso del análisis de correlación de Cramer's V y teniendo la necesidad de disminuir dimensionalidad, dado que la mayoría de las variables son categóricas. Según la literatura [7], se entiende como correlación alta los valores mayores al 30%, y de acuerdo con los resultados evidentes en la “Figura. 7”, se decide prescindir de las variables “causales”, “motivo”, “of_ventas” y “centro” por su baja correlación con las demás variables.

Figure 7. Gráfico correlación Cramer's V. (Elaboración propia)



Análisis Multivariado: De acuerdo con el análisis de clúster de K-Means, en la “Figura. 8” donde se grafica la tipología de destrucciones, se podría decir que no hay evidencia suficiente para visualizar una diferenciación en agrupamientos. Este comportamiento se puede responder, por los anteriores resultados de esta data, que muestra un comportamiento inusual, o también porque para este ejercicio de eligió una muestra muy pequeña en comparación con el tamaño total, debido al rendimiento del modelo. Pero en este caso puntual, los clúster no siguen ningún patrón evidente.

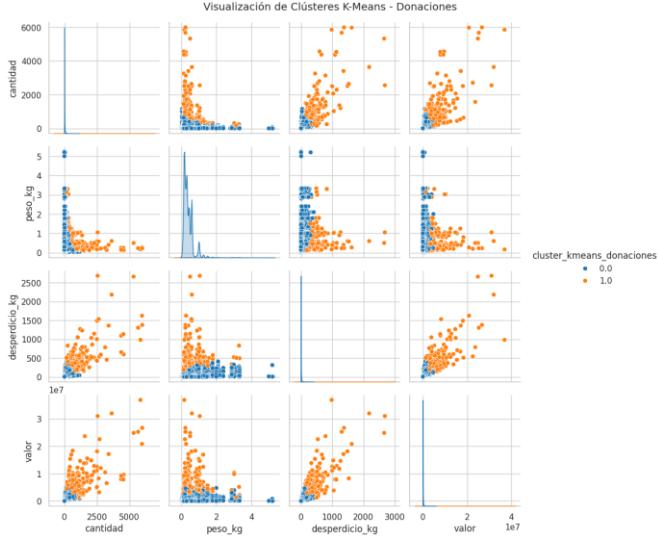
Figure 8. Clúster K-Means por tipología Destrucciones. (Elaboración propia)



Por parte de las donaciones que se grafican en la “Figura. 9” el algoritmo K-Means identificó dos grupos bien diferenciados dentro del conjunto de datos, caracterizados por distintos niveles de magnitud en producción, peso, donaciones y valor económico: Clúster 0: Representa la mayoría de las donaciones pequeñas, con cantidades promedio bajas (≈ 17 unidades), peso promedio bajo (0.38 kg), y desperdicio relativamente bajo (≈ 6.38 kg). Valor promedio de la donación también es bajo (≈ 88 mil). Clúster 1: Representa donaciones muy grandes o atípicas, con cantidades promedio mucho mayores ($\approx 1,117$ unidades), peso ligeramente mayor (0.57 kg), y desperdicio promedio extremadamente alto (≈ 447 kg). Valor promedio de las donaciones en este grupo es muy alto ($\approx 6,16$ millones), indicando que probablemente corresponden a donaciones excepcionales o institucionales.

En conjunto, el modelo evidencia una segmentación clara, útil para clasificar los registros en categorías de baja y alta escala. Podría ser útil transformar la variable respuesta “valor” que es de mi interés para crear un modelo que ayude a pronosticar la clasificación del valor de las donaciones en bajas o altas. Esta segmentación permite entender mejor la distribución de donaciones y puede ser útil para diseñar estrategias diferenciadas de logística, almacenamiento o análisis de impacto económico.

Figure 9. Clúster K-Means por tipología Donaciones. (*Elaboración propia*)



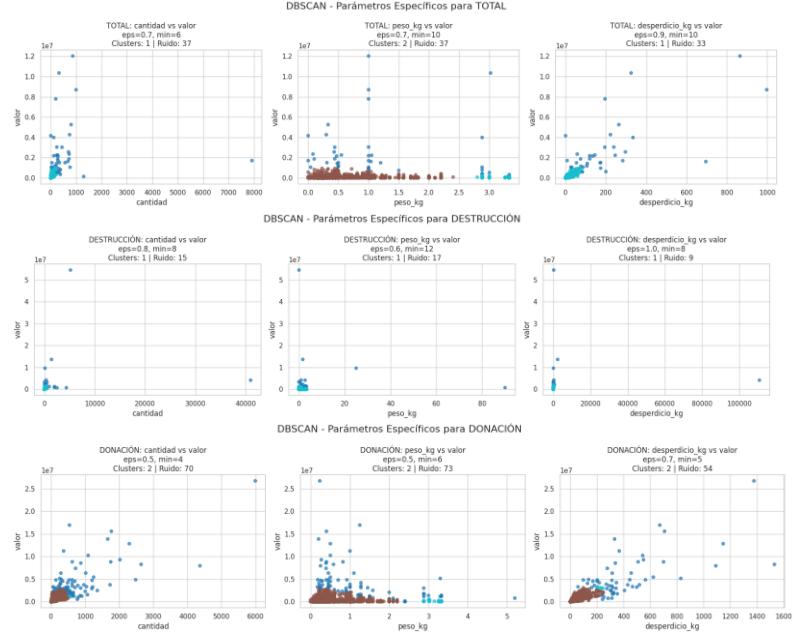
C. Atípicos

De acuerdo con los resultados del análisis por medio de Box Plot, se podría decir que las variables "cantidad", "valor" y "donación_kg", evidencian un gran número de outliers, sin embargo, con los análisis hechos anteriormente, sabemos que estas variables están altamente relacionadas y según sus estadísticos, era de esperarse este resultado. Por otro lado, la variable "peso_kg" presenta outliers, pero no son tan representativos en comparación con las demás variables, con una notoria cola hacia la derecha. Estos outliers, son lógicos en su esencia, ya que los valores que pueden tomar cada una de las variables, por ejemplo "valor", son en su mayoría pequeños y muy pocos con valores muy altos, y este comportamiento se ve reflejado en el Box-Plot, con muy pocos datos muy altos, que halan la gráfica hacia la derecha. En conclusión, estos datos no serían etiquetados como outliers, ya que en el contexto se pueden presentar estos casos. Por ejemplo, las donaciones en su mayoría son valores muy pequeños, sin embargo, hay situaciones, como en temporada alta (diciembre), donde las donaciones aumentan debido a la alta producción por la época del año.

Para complementar este análisis, se realizó un análisis de DBSCAN aplicado a la data total y por tipología de destrucciones y donaciones como se puede visualizar en la “Figura. 10”. Para el análisis del total de los datos y de la segmentación de donaciones, se logra diferenciar un agrupamiento en dos clases evidente, ya que se logran visualizar valores pequeños y altos, hay algunos datos que no logran ser diferenciados, pero la gran mayoría sí lo son. En cambio, para el caso de las destrucciones no es evidente una diferenciación ya que el grupo se encuentra muy sectorizado. Lo que lleva a puntualizar en que esta data de destrucciones tiene complicaciones en cuanto a análisis de correlación, de linealidad y de agrupamiento.

Este hallazgo reafirma lo expuesto anteriormente con los resultados obtenidos en K-Means y con Box-Plot, dónde los outliers son justificados con la naturaleza del contexto y es más visual en la data de donaciones.

Figure 10. Resultados DBSCAN totalizado y por tipología. (*Elaboración propia*)



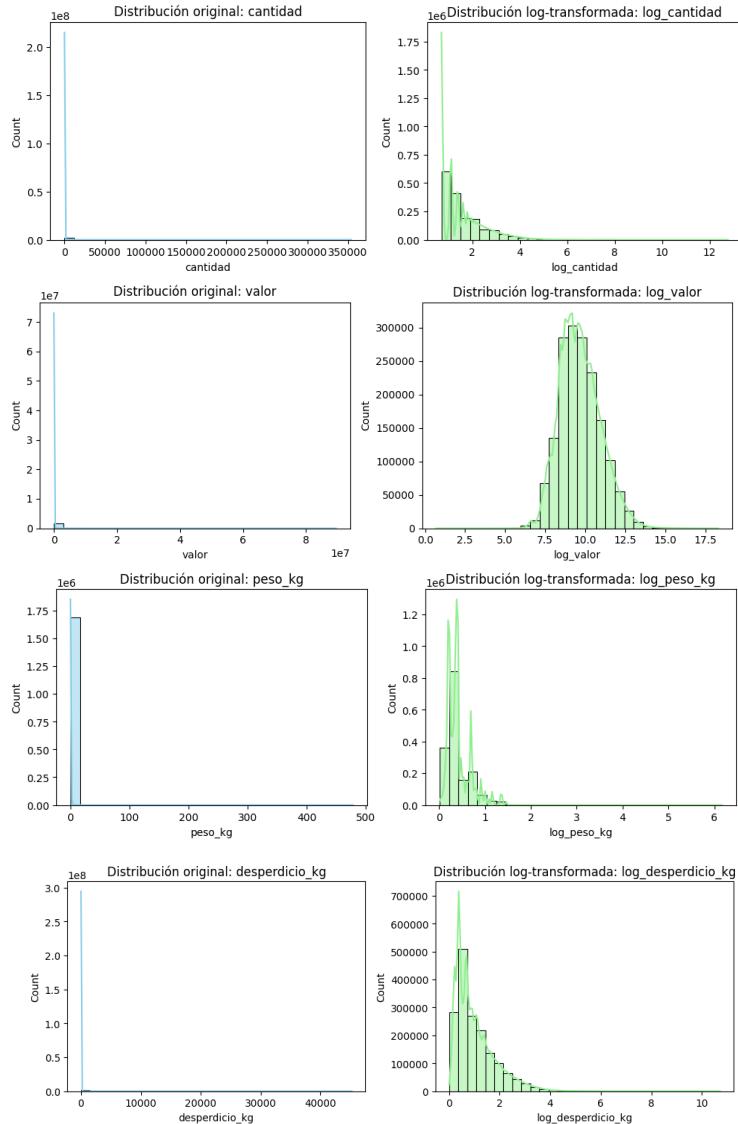
D. Imputación, Escalamiento y Transformación

Al realizar la imputación se confirma que la data ya no tiene datos en cero (que no sean lógicos), NaN o vacíos y la validación estadística de este primer proceso muestra un aumento en el count, ya que el número de datos pasó de $1,684,739 \rightarrow 1,686,017$. Esto significa que se imputaron 1,278 valores faltantes (se agregaron valores donde había NaN). La media y mediana casi no cambiaron, ya que la media pasó de $0.5732544 \rightarrow 0.5732574$, un cambio mínimo, y la mediana (50%), cuartiles y el mínimo/máximo se mantuvieron iguales, esto indica que la imputación no deformó la distribución original y el método conservó la estructura estadística del dato. Lo mismo aplica para desperdicio_kg, ya que solo se aumentó el count al imputar los NaN, y los valores estadísticos esenciales (mediana, cuartiles, min, max) se mantuvieron casi iguales. En cuanto a la validación de Chi-cuadrado = 0.0000, significa que no hay diferencia entre las distribuciones de las categorías “Antes” y “Después”. Los valores observados coinciden exactamente con los esperados bajo la hipótesis nula. Y el valor p-valor = 1.0000 es muy alto, mucho mayor que cualquier umbral típico (0.05, 0.01). Esto indica que no se rechaza la hipótesis nula, es decir, estadísticamente no hay diferencia entre las dos tablas.

Posteriormente se realiza la transformación de las variables continuas utilizando la función logarítmica y Box-Cox. Para el caso de la distribución logarítmica como se puede observar en la “Figura. 11”, redujo sustancialmente la asimetría de todas las variables, especialmente cantidad y valor, haciendo que las distribuciones sean más cercanas a la simetría y más adecuadas para análisis estadísticos basados en supuestos de normalidad. Sin embargo, aunque mejoró la forma, los datos todavía no son completamente normales (como mostró el test KS). Esto es normal en datos económicos, productivos o de demanda (con colas largas), pero ahora los datos son más tratables: puedes usar estos log-

transformados para modelos lineales, ANOVA, o clustering sin violar tanto los supuestos.

Figure 11. Resultados transformación logarítmica. (*Elaboración propia*)



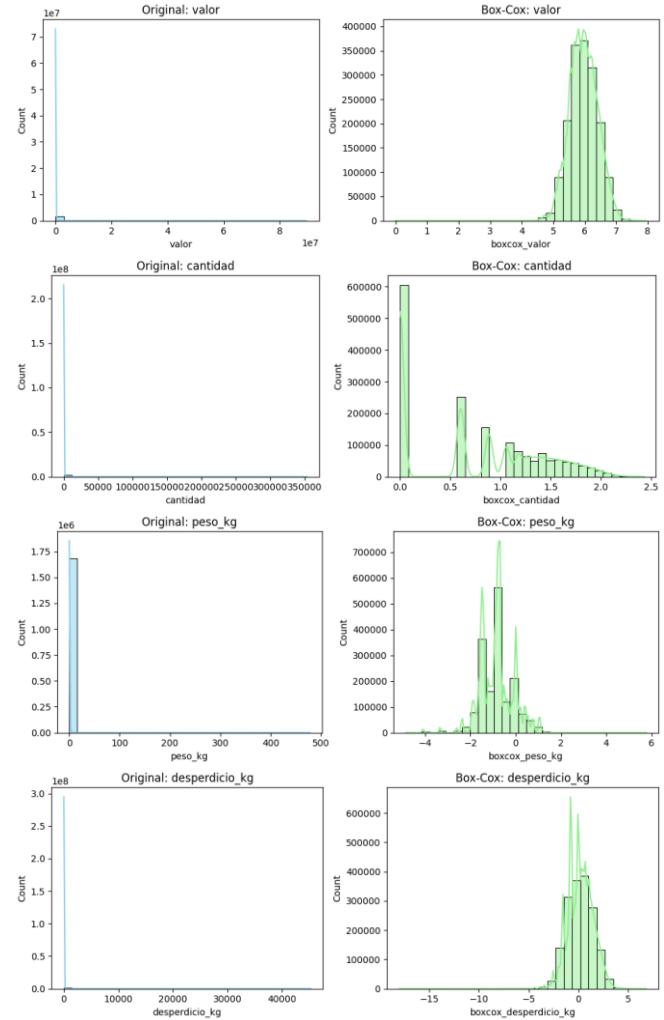
Para el caso de la transformación Box-Cox graficada en la “Figura. 12” muestra que los valores de λ son cercanos a cero o negativos pequeños, lo que indica que las distribuciones presentan asimetría positiva moderada y que se benefician de una ligera transformación logarítmica para acercarse a la normalidad. Se logra visualizar en el caso de cantidad, peso_kg y desperdicio_kg una distribución alrededor de cero y para el caso de valor, presenta una mejor distribución de campana. Adicionalmente, la transformación de Box-Cox logra reducir parcialmente la asimetría en algunas variables (por ejemplo, cantidad pasa de 334.77 a 0.21), pero los valores de λ cercanos a cero o negativos indican que los efectos son modestos en otras variables. La prueba de normalidad de Kolmogorov-Smirnov confirma que, incluso después de la transformación, ninguna variable sigue una distribución normal (p -valor = 0.0000), por lo que se mantiene la no normalidad de los datos.

Al comparar las dos transformaciones aplicadas a las variables, se observa que tanto el logaritmo como Box-Cox buscan reducir la fuerte asimetría presente originalmente,

pero con diferentes grados de efectividad. Para cantidad, la asimetría original era extremadamente alta (334.77); el logaritmo la reduce drásticamente a 1.42, mientras que Box-Cox la lleva a 0.21, mostrando que Box-Cox logra una distribución más cercana a la simetría. Para valor, peso_kg y desperdicio_kg, ambas transformaciones reducen la asimetría considerablemente respecto al original, pero Box-Cox produce valores de asimetría ligeramente menores o más cercanos a cero, indicando un mejor ajuste para aproximarse a simetría.

Sin embargo, las pruebas de normalidad de Kolmogorov-Smirnov muestran que ninguna transformación logra normalidad perfecta (p -valor = 0.0000 en todos los casos), por lo que, aunque las transformaciones mejoran la distribución, los datos siguen siendo no normales, y los métodos estadísticos que asumen normalidad deberán aplicarse con precaución o considerar alternativas robustas. En resumen: logaritmo es simple y efectivo, mientras que Box-Cox logra un ajuste más cercano a la simetría, aunque ninguna garantiza normalidad completa.

Figure 12. Resultados transformación Box-Cox. (*Elaboración propia*)



En cuanto a las variables categóricas, según la cantidad de valores únicos que tuviera cada una, como se puede observar en la “Tabla. II” se determinó la herramienta de transformación, de esta forma, las variables: planta, categoría, marca y demanda fueron transformadas con one hot encoding

y las variables con más valores únicos como material, linea y subcategoria con count encoding.

TABLE II. CANTIDAD DE VALORES ÚNICOS POR CADA VARIABLE CATEGÓRICA. ELABORACIÓN PROPIA

Variable	Cantidad de valores únicos
material	1256
linea	57
subcategory	30
planta	20
categoria	14
marca	13
demand	5

De esta forma, se unificó en un dataframe nombrado “df_final” las transformaciones de box-cox de las variables continuas y las transformaciones de one hot encoding y count encoding de las variables categóricas, con un tamaño final de 68 columnas y 1.686.017 registros. Así se finaliza todo el proceso metodológico para cumplir con el objetivo principal de analizar la data de donaciones y destrucciones de una empresa de alimentos cárnicos, reestructurándola, por medio de limpieza, análisis estadísticos y transformaciones para tener un producto óptimo que funcione como entrada al siguiente paso que es la modelación.

A modo de conclusión se crea la “Tabla. III” con el objetivo de resumir las principales modificaciones a las que fueron intervenidas cada una de las variables y poder contar con esa trazabilidad.

TABLE III. RESÚMEN MODIFICACIONES IMPORTANTES DE CADA VARIABLE. ELABORACIÓN PROPIA

Variable	Trazabilidad
1. Cantidad	- Eliminación de valores negativos - Transformación Box-Cox
2. Valor	- Eliminación de valores negativos - Eliminación de valores “0” - Transformación Box-Cox
3. Peso_kg	- Transformación Box-Cox
4. Desperdicio_kg	- Eliminación de valores negativos - Transformación Box-Cox
5. Material	- Imputación campos en “0” - Transformación Count Encoding
6. Motivo	- Se elimina por baja correlación
7. FechaFact#	- Clasificación como Date - Se utiliza para graficar los datos de acuerdo con la línea de tiempo.
8. Planta	- Imputación campos en “0” - Transformación One Hot Encoding
9. Centro	- Se elimina por baja correlación
10. Ofc#Ventas	- Se elimina por baja correlación
11. Categoría	- Imputación campos en “0” - Transformación One Hot Encoding
12. Subcategoria	- Imputación campos en “0” - Transformación Count Encoding
13. Línea	- Imputación campos en “0” - Transformación Count Encoding
14. Marca	- Imputación campos en “0” - Transformación One Hot Encoding
15. Ord/Ext	- Transformación One Hot Encoding
16. Causales NUEVOS	- Se elimina por baja correlación
17. Tipo de	- Se utilizó para separar la data

Desguace	principal y analizar por destrucción y donación.
----------	--

IV. CONCLUSIONES

La herramienta CRISPDPM es una guía que cubre la mayoría de los ítems para tener en cuenta para el tratamiento, limpieza y transformación de datos, sin embargo, es crucial para el análisis de datos, contar con un conocimiento pleno del contexto de los datos, ya que algunas decisiones necesitan una carga crítica. Esta etapa de análisis y preparación de los datos es fundamental a la hora de ejecutar modelos, ya que previamente se evidencian sesgos y comportamientos naturales de los datos, que al momento de ejecutar el modelo no son evidentes estas causas. Esta metodología permitió conocer detalles del sistema que no eran reconocidos a simple vista ni con el contexto, por ejemplo, la oportunidad de mejora a la hora de reportar las destrucciones en los CEDI, ya que se permiten valores negativos, y esto no es lógico, adicionalmente en la data completa hay valores “0” en los códigos de los productos, y saber que en estos momentos con esa data se están reportando indicadores con este tipo de fallas, sería interesante validar si con un proceso de limpieza y transformación, se podrían mejorar este tipo de reportería.

Para esta data en particular, el uso de la herramienta Cramer’s V para analizar la correlación de las variables categóricas, fue de una importancia relevante, ya que ayudó a disminuir la dimensión y consecuentemente simplificar el análisis posterior de toda la data. Por otra parte, la transformación de las variables continuas con Box-Cox, representó una mejora en su distribución y una unicidad de todas las variables, ya que interpretar los resultados será más sencillo. Y en cuanto a las variables categóricas, elegir estrategia dos herramientas de transformación: One Hot Encoding y Count Encoding, permitió una unificación de la data final, sin aumentar en gran medida su tamaño.

Durante el desarrollo hubo varias limitaciones, pero hubo dos que realmente significaron un reto, en primero lugar el tamaño y las características de la base de datos, no solamente con más de un millón de registros, sino con 42 variables. Dentro de esa cantidad de variables con las 17 que representaban valor en cuanto a la información y el interés del proyecto, había solamente 4 continuas, ya que la gran mayoría eran categóricas y las herramientas para su análisis son limitadas. Por otro lado, al realizar el análisis de “dos bases de datos”, tanto para destrucciones como para donaciones, significó duplicar el trabajo y el análisis, sin contar que la data de destrucciones tenía características difícilmente de analizar y entender.

Si bien se obtuvo como resultado una data que corresponde a una transformación de Box-Cox de la variable respuesta “valor”, sería interesante validar la opción de transformar la variable respuesta de forma categórica, como se evidenció en los resultados obtenidos del clúster de K-Means, en este caso para las donaciones. Por otro lado, es importante hacer la validación de las causas, del porqué la data de destrucciones tiene características complicadas de abordar, por ejemplo, ningún tipo de correlación, ningún tipo de agrupamiento, con las estrategias utilizadas en este estudio, puede ser que con otro tipo de estrategias su comportamiento se pueda explicar mejor.

REFERENCIAS

- [1] IBM, *IBM SPSS Modeler CRISP-DM Guide*, version 18.4.0. International Business Machines Corporation, 2025. Disponible: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPD_M.pdf
- [2] Departamento Nacional de Planeación (DNP), *Lineamientos para la formulación de la política pública para la reducción de pérdidas y desperdicios de alimentos en Colombia*, 2016. Disponible: https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Social/Lineamientos_Perde_Desperd_Alimentos.pdf
- [3] Congreso de Colombia, «Ley 1990 de 2019: Política pública para la reducción de pérdidas y desperdicio de alimentos en Colombia», Diario Oficial de la República de Colombia, 2019.
- [4] Congreso de Colombia, «Ley 2380 de 2024: Por la cual se promueve la donación de alimentos, la seguridad alimentaria y se aporta al objetivo de “Hambre Cero” en Colombia y se dictan otras disposiciones», Diario Oficial No. 52.818, 15 jul. 2024.
- [5] J. A. C. Pesca, A. Niño, and G. E. Niño, “Análisis de las tendencias especulativas de los precios de alimentos en Colombia”, Panorama Económico, vol. 31, no. 4, pp. 294–310, Jul. 2024, doi: 10.32997/pe-2023-4771.
- [6] mariabda2, “intro_data_2025,” GitHub. [Online]. Available: https://github.com/mariabda2/intro_data_2025. Accessed: Nov. 27, 2025.
- [7] J. Cohen, Statistical Power Analysis for the Behavioral Sciences. 2013. doi: 10.4324/9780203771587.