Instituto Tecnológico de Estudios Superiores de Monterrey

Proyecto Integrador

Avance1.#Equipo

Codigo de Etica:

"Confirmamos nuestro compromiso de acatar los principios y valores del compromiso de aprendizaje."

Equipo 13

César Alexis Nájera Mendoza	**A01376517**
Eduardo Martín Rico Sotomayor	**A01793775**
Ana Gabriela Fuentes Hernández	**A01383717**

Fecha de entrega:
04 de Mayo 2025

1. ¿Hay valores faltantes en el conjunto de datos? ¿Se pueden identificar patrones de ausencia?

   Los datos se obtienen de dos almacenes diferentes en la central de abastos pero las variables son las mismas, 4 categóricas (Producto, Etiqueta, Tamaño y Color) tambien tenemos 3 variables numéricas (Cajas, Kilos, Importe de Venta)

   Sí, hay valores faltantes o ausentes en múltiples columnas, como en las variables de Tamaño, Color, Cajas, Kilos, Importe de Venta: en muchos registros no se capturó información (especialmente Kilos y Color).

   En algunos productos como el Morrón, Tomate Cherry y Tomate Saladette, el campo Kilos aparece como 0.00, lo que probablemente indica dato faltante o no capturado (no una pérdida real de 0 kg).

2. ¿Cuáles son las estadísticas resumidas del conjunto de datos?

Estadísticas descriptivas de Almacen Bodega

```python
# Estadísticas resumidas del conjunto de datos
# Equipo 13 Mermas
import pandas as pd

# Cargar datos desde archivo CSV o Excel
df = pd.read_csv("/content/merma_bod_combinado_2022-2024.csv")  # datos bodega

# Ver estadísticas resumidas
print(df.describe(include='all'))
```

```
                Year Month Source         Producto  Etiqueta Tamaño    Color  \
count   2916.000000  2916   2916             2916      2916   2916      545
unique          NaN    12      1               24       235     53       10
top             NaN  Mayo    bod  Tomate Saladete  GENERICA     LG  COLOR 4
freq            NaN   303   2916             1671       292    415      199
mean    2023.022634   NaN    NaN              NaN       NaN    NaN      NaN
std        0.789404   NaN    NaN              NaN       NaN    NaN      NaN
min     2022.000000   NaN    NaN              NaN       NaN    NaN      NaN
25%     2022.000000   NaN    NaN              NaN       NaN    NaN      NaN
50%     2023.000000   NaN    NaN              NaN       NaN    NaN      NaN
75%     2024.000000   NaN    NaN              NaN       NaN    NaN      NaN
max     2024.000000   NaN    NaN              NaN       NaN    NaN      NaN

              Cajas        Kilos  Importe_Venta
count   2916.000000  2916.000000         2916.0
unique          NaN          NaN            NaN
top             NaN          NaN            NaN
freq            NaN          NaN            NaN
mean      67.495199    14.313563            0.0
std      138.369814   149.620900            0.0
min        0.000000     0.000000            0.0
25%        4.000000     0.000000            0.0
50%       21.000000     0.000000            0.0
75%       74.000000     0.000000            0.0
max     2024.000000  5559.000000            0.0
```

a. Valores Faltantes y Patrones de Ausencia

En el conjunto de datos se observan valores faltantes principalmente en la columna Color, ya que solo hay 545 registros con datos (de un total de 2916). Esto sugiere que en la mayoría de los casos no se registra esta variable. También se identifican muchos valores en cero en la

columna Kilos, lo que podría indicar ausencia de información o que efectivamente no hubo peso registrado para esas mermas.

b. Estadísticas Resumidas

El campo Year muestra que los registros abarcan de 2022 a 2024, con una media centrada en el año 2023. En cuanto a las variables numéricas:

- ➔ Cajas tiene un promedio de 67.5 cajas, pero una mediana de solo 21, lo que junto a una desviación estándar alta (138.3) indica que existen valores muy extremos. El valor máximo es de 2024 cajas, lo que probablemente representa un valor atípico.
- ➔ Kilos tiene una media de 14.31 kilos, pero los valores del 25%, 50% y 75% percentil son todos 0. Esto significa que más de la mitad de los registros no tienen peso registrado. Además, el valor máximo alcanza los 5559 kilos, lo que nuevamente apunta a la existencia de valores atípicos importantes.
- ➔ El importe de venta es igual a cero en todos los registros, por lo que esta variable no proporciona información útil para el análisis actual.

c. Valores Atípicos

Las columnas Cajas y Kilos contienen valores extremos que se alejan significativamente de la media y los percentiles. Esto indica la presencia de valores atípicos que pueden distorsionar los resultados si no se gestionan adecuadamente. En ambos datasets, la desviación estándar es considerablemente alta y los valores máximos registrados (ej. 5559 Kilos en BODEGA, 22418 Kilos en ISLA) están muy alejados de las medianas (0 Kilos en ambos casos). Esta gran diferencia es un claro indicador de que existen registros con valores inusualmente altos que deberían ser investigados más a fondo, o bien, que la mediana en 0 en kilos nos dice que hacen falta valores en la misma.

d. Cardinalidad de Variables Categóricas

El conjunto contiene varias variables categóricas con alta diversidad de valores:

Producto: 24 categorías, siendo Tomate Saladete el más frecuente con más del 57% del total.

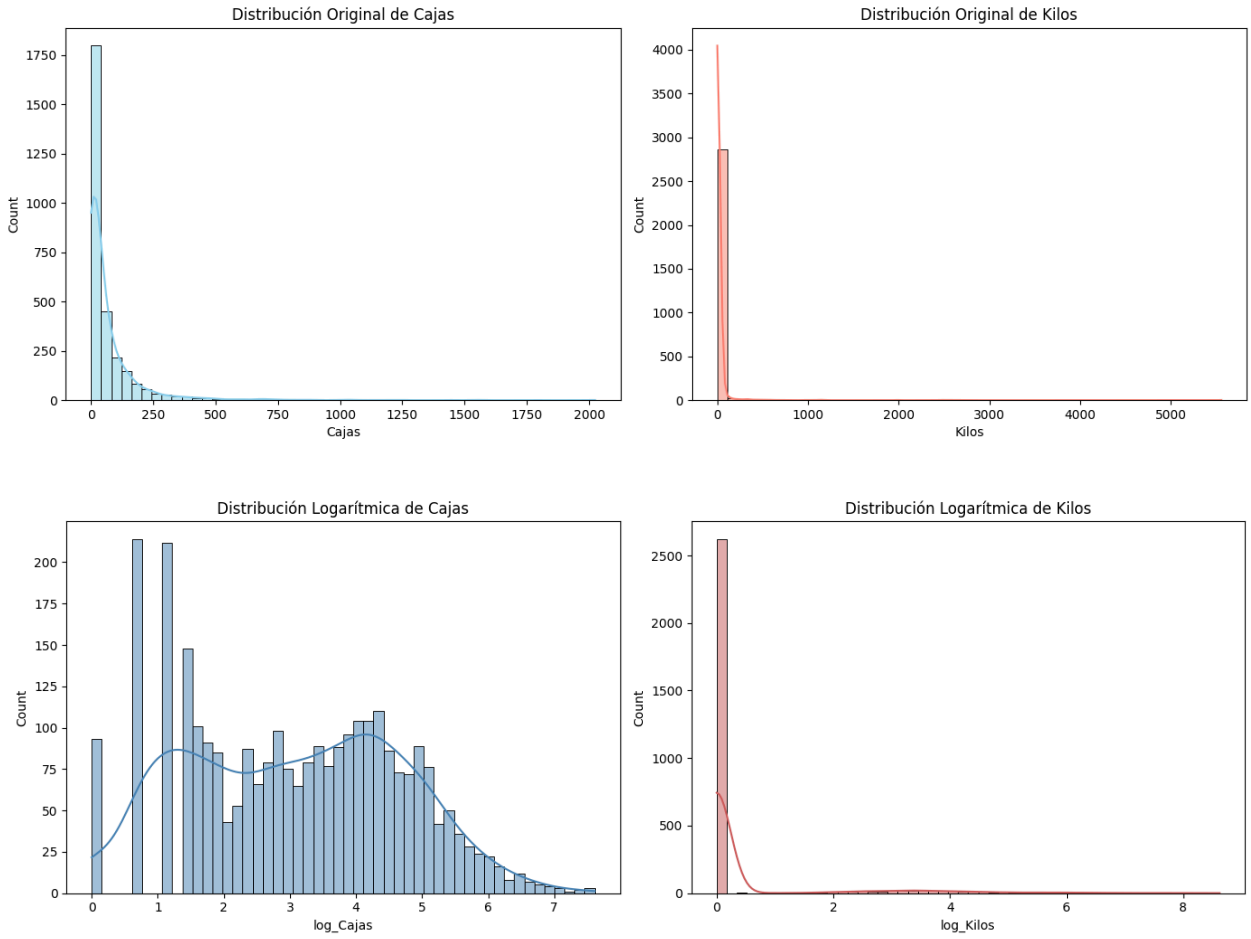
Etiqueta: 235 etiquetas distintas, lo que implica una alta cardinalidad. La más frecuente es GENÉRICA.

Tamaño: 53 valores únicos, con LG como el más común.

Color: 10 valores únicos, destacando COLOR 4 como el más frecuente. Sin embargo, la mayoría de los registros no tienen información en esta columna.

e. Distribuciones Sesgadas y Transformaciones

Tanto Cajas como Kilos presentan distribuciones altamente sesgadas hacia la derecha (muchos valores bajos y pocos valores muy altos). En estos casos, es recomendable aplicar transformaciones no lineales, como el logaritmo, para mejorar la representación y el análisis de los datos.

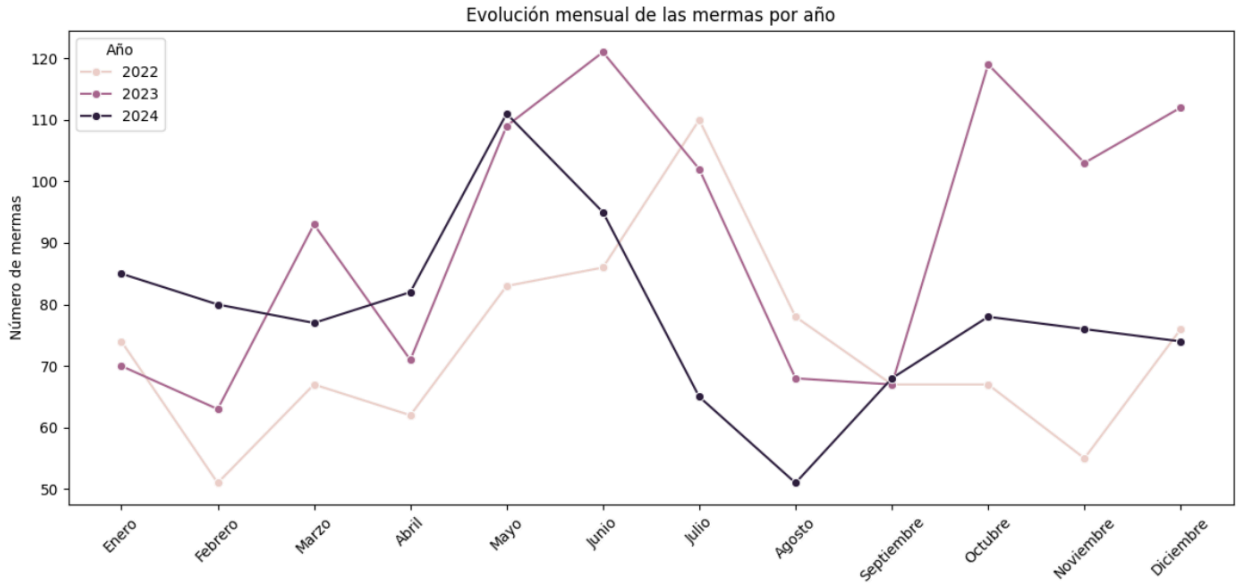


Esto significa que la mayoría de los datos se concentran en valores bajos, mientras que hay pocos casos con valores muy altos, lo que genera una cola larga en el extremo derecho de la distribución. Este tipo de sesgo puede impactar en la interpretación de los datos, ya que las medidas de tendencia central como la media pueden quedar distorsionadas y los modelos predictivos pueden ajustarse de forma ineficiente. Para corregir este problema y mejorar la representación de los datos, se aplicó una transformación logarítmica utilizando la función logarítmica natural ajustada (log1p), que permite incluir ceros sin generar errores. Como resultado, las distribuciones se vuelven más simétricas y se reducen los efectos de los valores extremos, facilitando así una visualización más clara, una mejor detección de patrones.

f. Tendencias Temporales

La variable Year permite analizar la evolución de las mermas a lo largo del tiempo, aunque sería necesario un desglose adicional por mes y año para detectar posibles tendencias.

Actualmente, Mayo es el mes con más registros (303), lo que podría indicar una estacionalidad, pero se requiere más análisis.

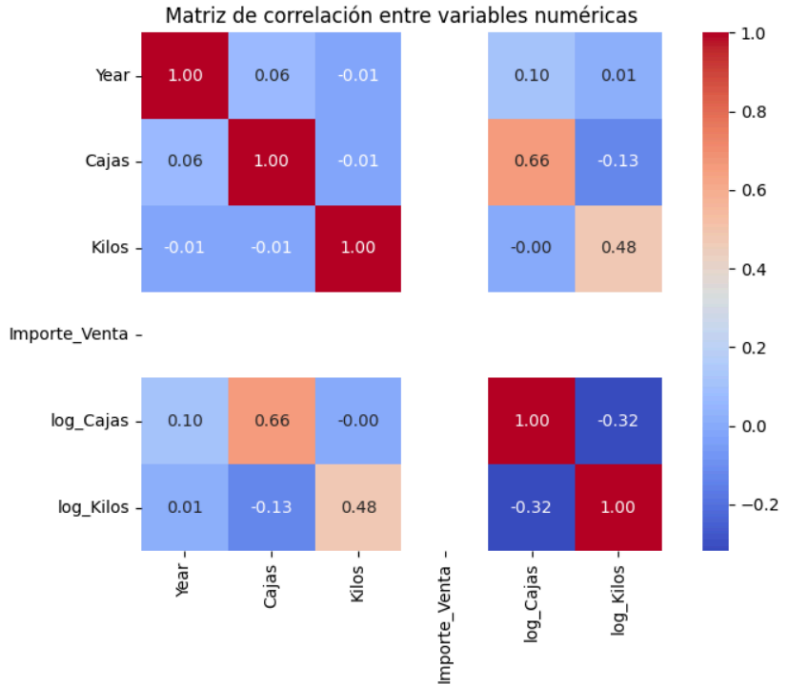


Este patrón sugiere que durante ciertos meses del año, particularmente en mayo, ocurren más pérdidas, lo cual podría estar relacionado con factores como la estacionalidad de la producción, condiciones climáticas o picos en la operación logística.

g. Correlaciones

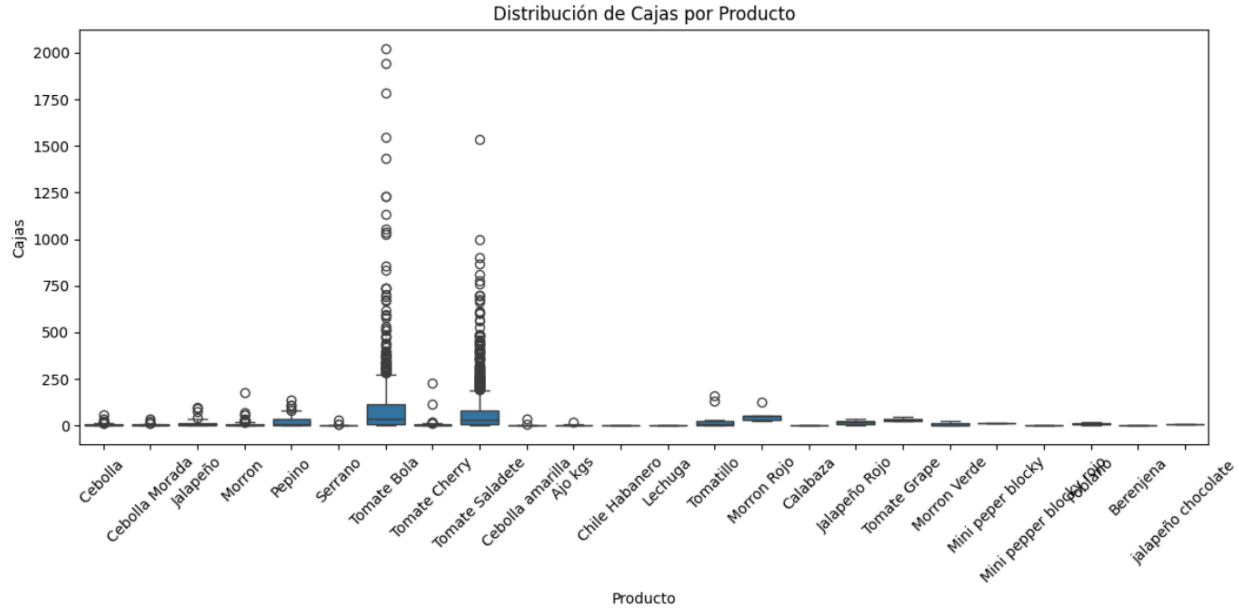
La matriz de correlación muestra que las variables Cajas y Kilos tienen una relación muy débil y negativa (correlación de -0.01). La variable Cajas tiene una correlación moderada con su versión logarítmica (log_Cajas, 0.66), lo que indica que la transformación logarítmica mejora la relación lineal. Kilos no muestra una fuerte correlación ni con Cajas ni con su transformación logarítmica (log_Kilos, -0.13), aunque hay una correlación moderada con log_Kilos (0.48), sugiriendo una relación más significativa cuando se transforman las variables. La variable Importe_Venta no muestra correlaciones con ninguna otra, lo que indica que no varía o no aporta información útil para el análisis

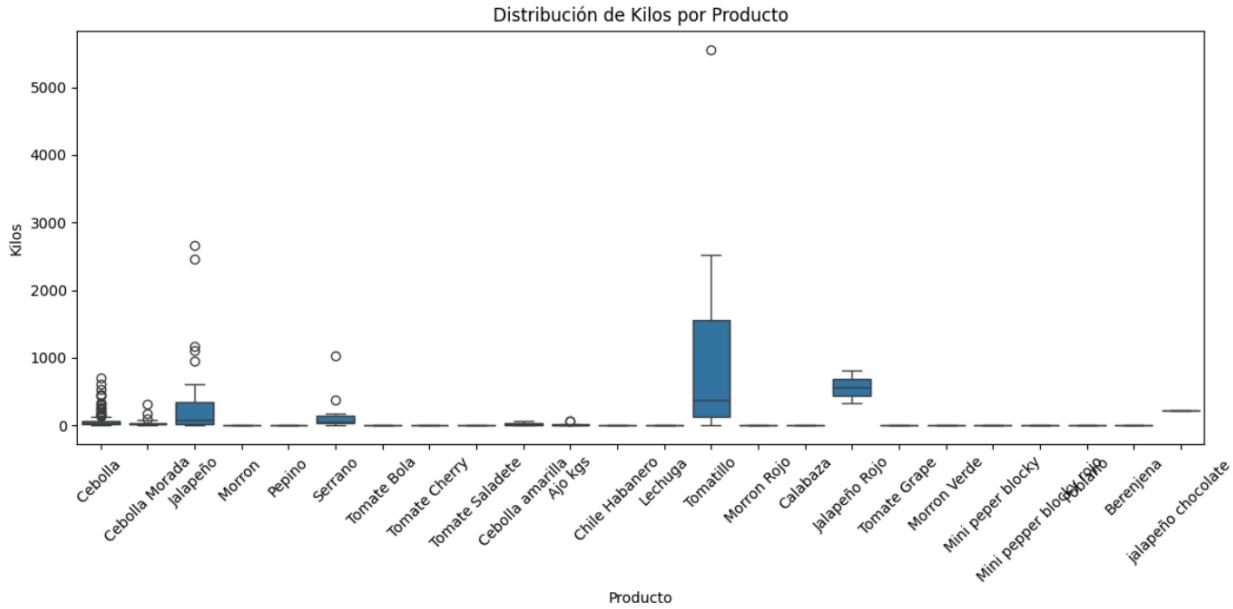
```
Matriz de correlación:
                    Year     Cajas     Kilos  Importe_Venta  log_Cajas  \
Year            1.000000  0.064460 -0.014462            NaN   0.101450
Cajas           0.064460  1.000000 -0.011521            NaN   0.656482
Kilos          -0.014462 -0.011521  1.000000            NaN  -0.002559
Importe_Venta        NaN       NaN       NaN            NaN        NaN
log_Cajas       0.101450  0.656482 -0.002559            NaN   1.000000
log_Kilos       0.009417 -0.131382  0.483865            NaN  -0.317585

                log_Kilos
Year             0.009417
Cajas           -0.131382
Kilos            0.483865
Importe_Venta         NaN
log_Cajas       -0.317585
log_Kilos        1.000000
```

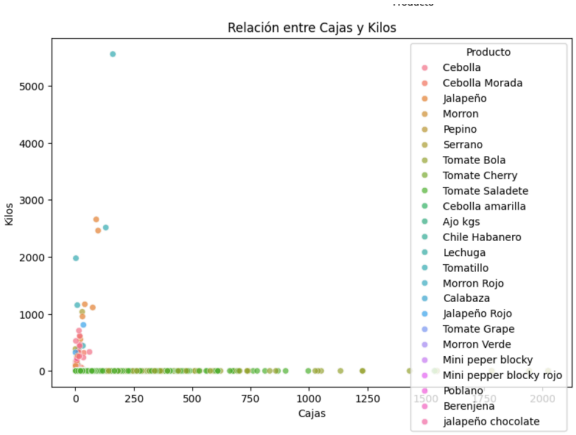


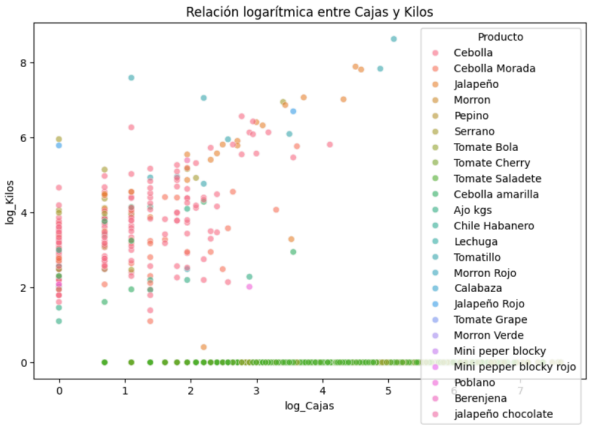
h. Distribución por Categorías (Análisis Bivariado)

Se identifica un fuerte sesgo hacia ciertos productos y etiquetas. Por ejemplo, Tomate Saladete aparece en más de la mitad de los registros. Esto indica un desequilibrio en las clases de la variable Producto, lo que debe considerarse en cualquier análisis comparativo o modelo predictivo.




Distribución de Kilos por Producto
Kilos
Producto
Cebolla
Cebolla Morada
Jalapeño
Morron
Pepino
Serrano
Tomate Bola
Tomate Cherry
Tomate Saladete
Cebolla amarilla
Ajo kgs
Chile Habanero
Lechuga
Tomatillo
Morron Rojo
Calabaza
Jalapeño Rojo
Tomate Grape
Morron Verde
Mini peper blocky
Mini pepper blocky rojo
Poblano
Berenjena
jalapeño chocolate


Relación entre Cajas y Kilos
Kilos
Cajas
Producto
Cebolla
Cebolla Morada
Jalapeño
Morron
Pepino
Serrano
Tomate Bola
Tomate Cherry
Tomate Saladete
Cebolla amarilla
Ajo kgs
Chile Habanero
Lechuga
Tomatillo
Morron Rojo
Calabaza
Jalapeño Rojo
Tomate Grape
Morron Verde
Mini peper blocky
Mini pepper blocky rojo
Poblano
Berenjena
jalapeño chocolate


Relación logarítmica entre Cajas y Kilos
log_Kilos
log_Cajas
Producto
Cebolla
Cebolla Morada
Jalapeño
Morron
Pepino
Serrano
Tomate Bola
Tomate Cherry
Tomate Saladete
Cebolla amarilla
Ajo kgs
Chile Habanero
Lechuga
Tomatillo
Morron Rojo
Calabaza
Jalapeño Rojo
Tomate Grape
Morron Verde
Mini peper blocky
Mini pepper blocky rojo
Poblano
Berenjena
jalapeño chocolate

Estadísticas descriptivas de Almacén La Isla

```python
# Estadísticas resumidas del conjunto de datos
# Equipo 13 Mermas
import pandas as pd

# Cargar datos desde archivo CSV o Excel
df = pd.read_csv("/content/merma_isla_combinado_2022-2024.csv")  # datos bodega

# Ver estadísticas resumidas
print(df.describe(include='all'))
```

```
                Year      Month Source         Producto  Etiqueta Tamaño  \
count    525.000000        525    525              525       525    525
unique          NaN         12      1                9        89     22
top             NaN  Diciembre   isla  Tomate Saladete  GENERICA     XL
freq            NaN         70    525              453       138    126
mean    2023.337143        NaN    NaN              NaN       NaN    NaN
std        0.790359        NaN    NaN              NaN       NaN    NaN
min     2022.000000        NaN    NaN              NaN       NaN    NaN
25%     2023.000000        NaN    NaN              NaN       NaN    NaN
50%     2024.000000        NaN    NaN              NaN       NaN    NaN
75%     2024.000000        NaN    NaN              NaN       NaN    NaN
max     2024.000000        NaN    NaN              NaN       NaN    NaN

        Color        Cajas         Kilos  Importe_Venta
count     1.0   525.000000    525.000000          525.0
unique    NaN          NaN           NaN            NaN
top       NaN          NaN           NaN            NaN
freq      NaN          NaN           NaN            NaN
mean      2.0   414.262857     50.380952            0.0
std       NaN   828.875224    983.110576            0.0
min       2.0     0.000000      0.000000            0.0
25%       2.0    31.000000      0.000000            0.0
50%       2.0    87.000000      0.000000            0.0
75%       2.0   360.000000      0.000000            0.0
max       2.0  7505.000000  22418.000000            0.0
```

a. Valores Faltantes y Patrones de Ausencia

En el conjunto de datos, se observan valores faltantes principalmente en la columna Color, ya que solo hay un registro con datos en esta columna, lo que sugiere que no se ha registrado información en la mayoría de los casos. Además, la variable Importe_Venta presenta un valor constante de 0 en todos los registros, lo que indica que no se ha generado información sobre las ventas, lo cual podría hacerla irrelevante para el análisis.

b. Estadísticas Resumidas

La variable Year abarca los años 2022 a 2024, con una media cercana al 2023. En términos de las variables numéricas:

- ➔ Cajas tiene una media de 414.26, pero una mediana de 87, lo que, junto a una alta desviación estándar (828.88), sugiere la existencia de valores extremos en esta variable. El valor máximo es 7505, lo que podría ser un valor atípico.
- ➔ Kilos tiene una media de 50.38, pero una mediana de 0. Esto implica que una gran parte de los registros no contienen información sobre el peso, y algunos registros presentan valores extremos de hasta 22418 kilos.
- ➔ Importe_Venta es constante en 0 para todos los registros, por lo que no proporciona información útil en su forma actual.

c. Valores Atípicos

Se identifican valores atípicos significativos en las columnas Cajas y Kilos, debido a su alta dispersión y la existencia de registros con valores extremadamente altos. Estos valores extremos pueden afectar el análisis y la interpretación de los datos, por lo que se recomienda gestionarlos adecuadamente.

d. Cardinalidad de Variables Categóricas

El conjunto de datos contiene varias variables categóricas con alta diversidad de valores:

Producto: 9 categorías, siendo el Tomate Saladete el más frecuente con 453 registros, lo que indica que este producto domina en la muestra.

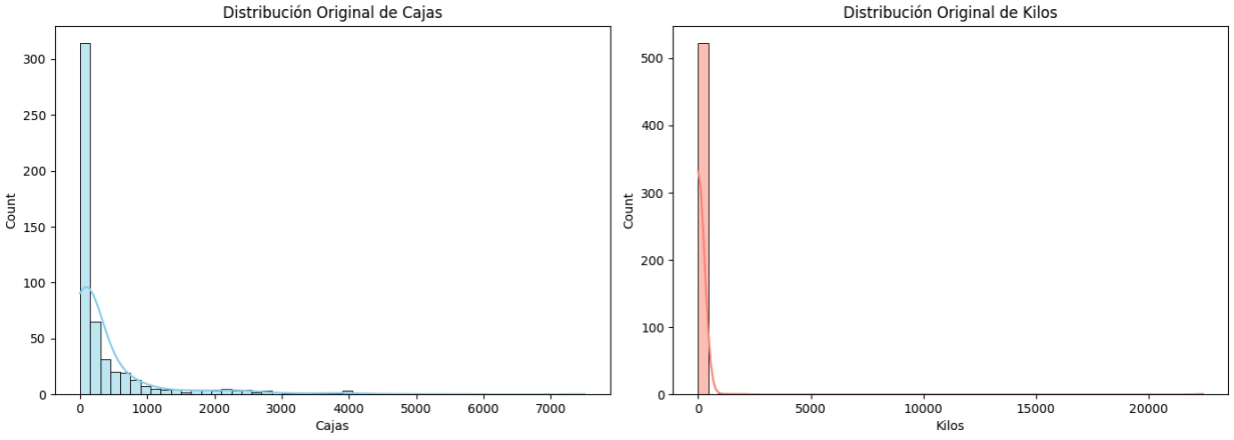
Etiqueta: 89 etiquetas diferentes, con GENERICA siendo la más común.

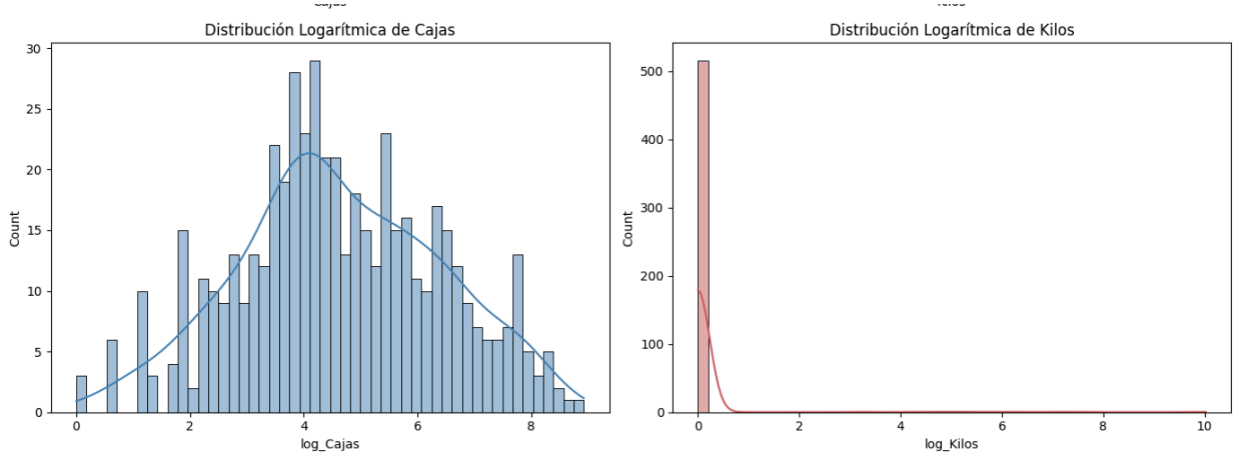
Tamaño: 22 tamaños diferentes, con XL como el más frecuente.

Color: Aunque esta variable tiene 2 valores distintos, solo se ha registrado un único valor, lo que sugiere que la variable Color no está siendo utilizada adecuadamente.

e. Distribuciones Sesgadas y Transformaciones.

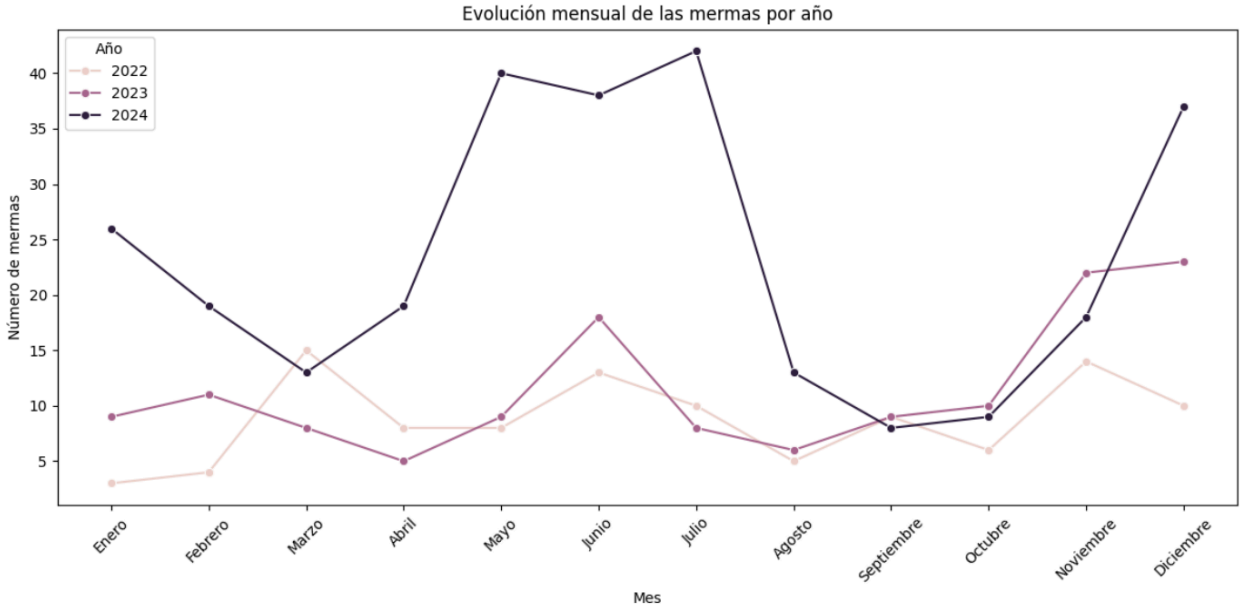
Las variables Cajas y Kilos presentan distribuciones altamente sesgadas hacia la derecha, con muchos valores bajos y pocos valores extremadamente altos. En estos casos, aplicar transformaciones no lineales como el logaritmo puede ser útil para mejorar la representación de los datos y facilitar el análisis.





f. Tendencias Temporales

La variable Year muestra una evolución de los datos a lo largo del tiempo, aunque un análisis más detallado por Mes sería necesario para identificar tendencias estacionales. Diciembre es el mes con más registros, lo que podría indicar una mayor actividad en ese período. Sin embargo, se requieren más análisis para confirmar si hay una estacionalidad clara.



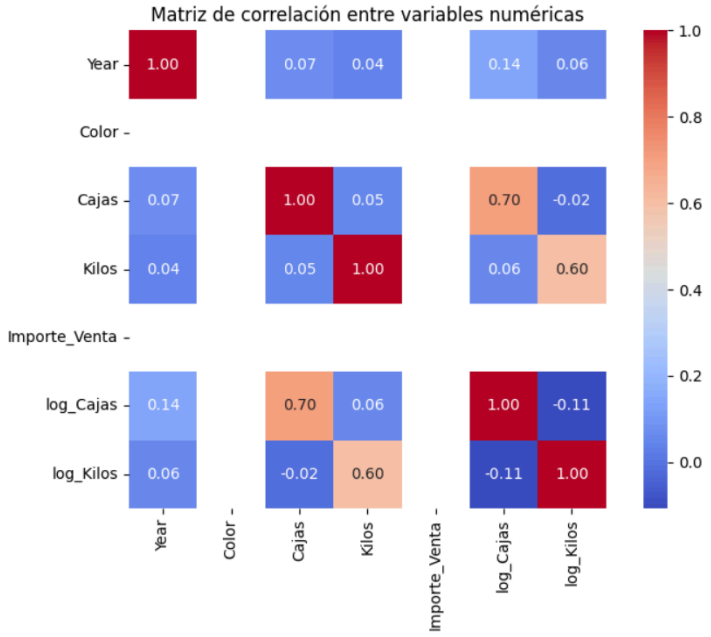
g. Correlaciones.

La matriz de correlación revela relaciones débiles entre la mayoría de las variables numéricas del conjunto de datos. La correlación entre el número de cajas y su transformación logarítmica (0.70), así como entre los kilos y su logaritmo (0.60), indica que las transformaciones logarítmicas son útiles para mejorar la linealidad y reducir el sesgo en estas variables. Sin embargo, la relación entre Cajas y Kilos es prácticamente nula (0.048), lo que sugiere que el

número de cajas no está fuertemente vinculado al peso registrado, posiblemente por inconsistencias o particularidades del producto. Del mismo modo, la variable Year muestra una correlación muy débil con Cajas y Kilos, lo que indica que no hay una tendencia temporal clara en los registros. Además, variables como Color e Importe_Venta no participan en la matriz debido a falta de datos o ausencia de variabilidad, por lo que no aportan valor al análisis de correlación actual.

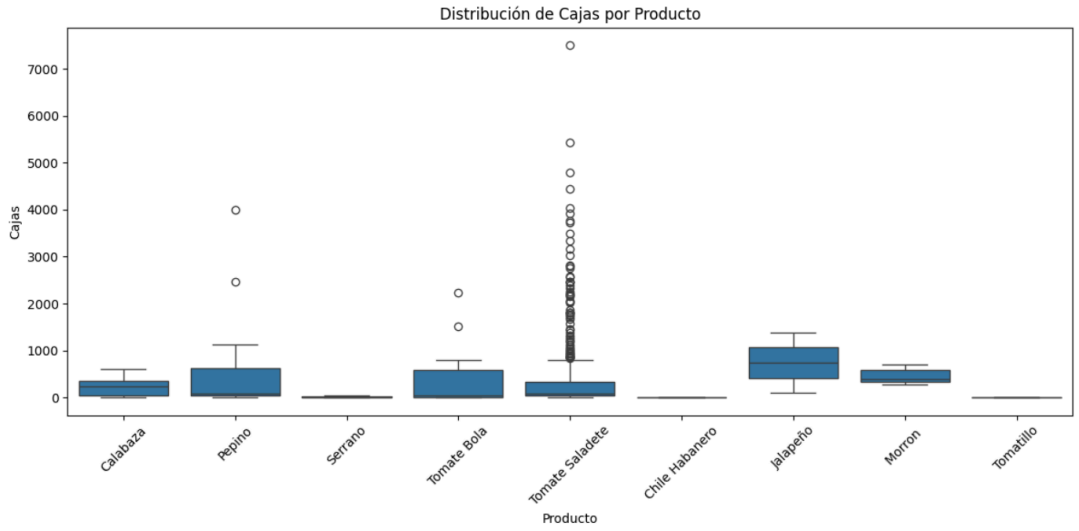
Matriz de correlación:

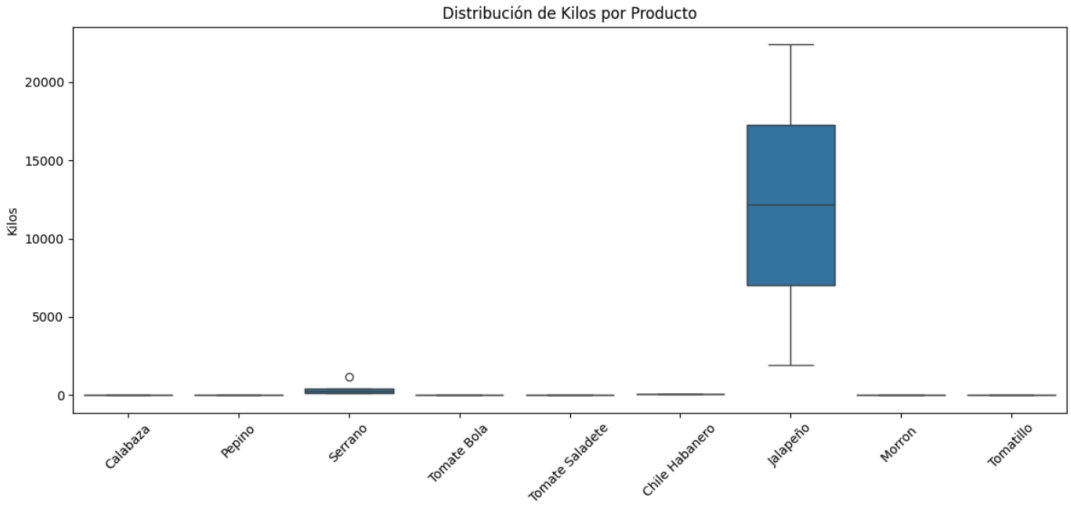
	Year	Color	Cajas	Kilos	Importe_Venta	log_Cajas	log_Kilos
Year	1.000000	NaN	0.074198	0.041922	NaN	0.135805	0.060151
Color	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Cajas	0.074198	NaN	1.000000	0.047622	NaN	0.702930	-0.019526
Kilos	0.041922	NaN	0.047622	1.000000	NaN	0.059266	0.601190
Importe_Venta	NaN	NaN	NaN	NaN	NaN	NaN	NaN
log_Cajas	0.135805	NaN	0.702930	0.059266	NaN	1.000000	-0.106741
log_Kilos	0.060151	NaN	-0.019526	0.601190	NaN	-0.106741	1.000000

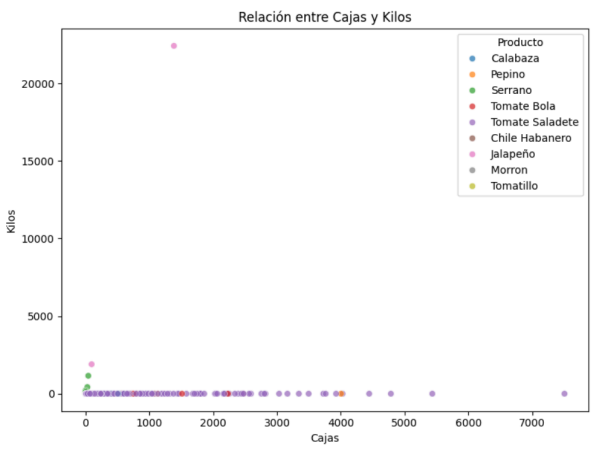


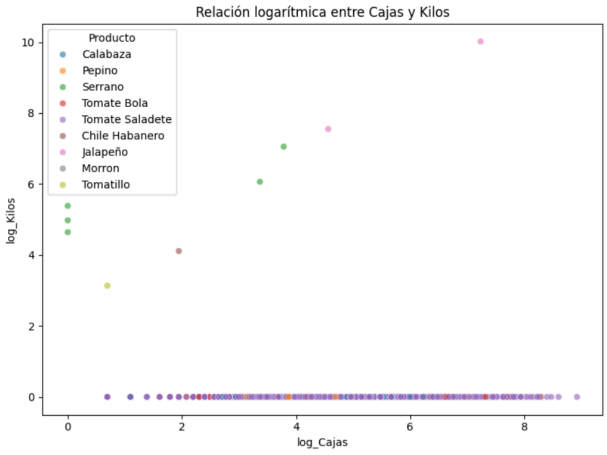
h. Distribución por Categorías (Análisis Bivariado)

El análisis muestra un fuerte sesgo hacia ciertos productos, especialmente **Tomate Saladete**, que aparece en más de la mitad de los registros. Esto indica un desequilibrio en las clases de la variable **Producto**, lo que puede influir en los resultados si se utiliza esta variable en modelos predictivos o análisis comparativos.


Distribución de Cajas por Producto
Cajas
Producto
Calabaza
Pepino
Serrano
Tomate Bola
Tomate Saladete
Chile Habanero
Jalapeño
Morron
Tomatillo

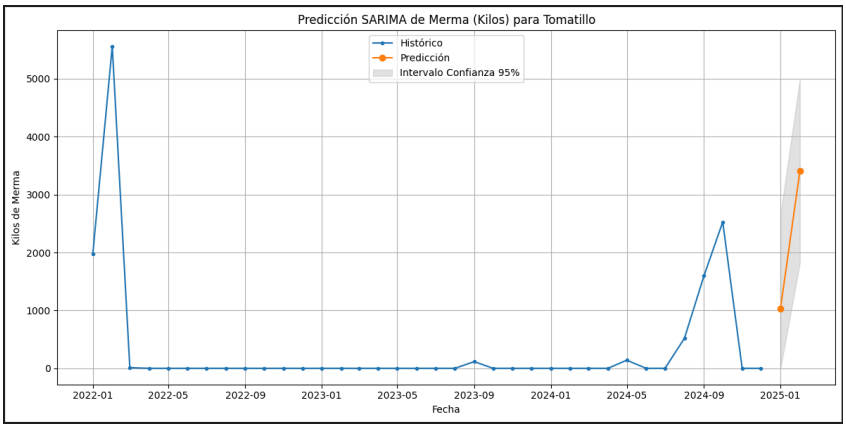

Distribución de Kilos por Producto
Kilos
Calabaza
Pepino
Serrano
Tomate Bola
Tomate Saladete
Chile Habanero
Jalapeño
Morron
Tomatillo


Relación entre Cajas y Kilos
Producto
Calabaza
Pepino
Serrano
Tomate Bola
Tomate Saladete
Chile Habanero
Jalapeño
Morron
Tomatillo
Kilos
Cajas


Relación logarítmica entre Cajas y Kilos
Producto
Calabaza
Pepino
Serrano
Tomate Bola
Tomate Saladete
Chile Habanero
Jalapeño
Morron
Tomatillo
log_Kilos
log_Cajas

3. Análisis Sarima.

Se modelaron los siguientes 10 productos principales: Tomatillo, Jalapeño, Cebolla , Serrano, Cebolla Morada, Jalapeño Rojo, Cebolla amarilla, jalapeño chocolate , Ajo kgs, Poblano



4. ¿Se deberían normalizar las imágenes para visualizarlas mejor? No aplica
5. ¿Hay desequilibrio en las clases de la variable objetivo? No aplica

Conclusión General del Análisis Exploratorio de los Datos de Mermas

En este análisis exploratorio de datos para los almacenes Bodega e Isla de la central de abastos podemos observar importantes desafíos de calidad, distribución y representación que deben ser abordados para un análisis y construcción de modelos predictivos más adelante. En ambos conjuntos se identificaron valores faltantes significativos, particularmente en las variables Color y Kilos. En el caso de Kilos, la abundancia de ceros sugiere datos no capturados más que una ausencia real de peso, mientras que en Color, la falta de registros en más del 80% de los casos limita su utilidad. Además, la variable Importe de Venta resulta no significativa en su forma actual, ya que es constante (0) en todos los registros.

Las estadísticas resumidas revelan una fuerte presencia de valores atípicos en Cajas y Kilos, reflejados en sus medias muy alejadas de las medianas y desviaciones estándar elevadas. Estos valores extremos podrían estar distorsionando las métricas tradicionales y deben gestionarse cuidadosamente. Las distribuciones de estas variables también presentan un sesgo positivo pronunciado, por lo que se aplicaron transformaciones logarítmicas (log1p) para normalizar y mejorar su análisis.

En cuanto a las variables categóricas, se observó una alta cardinalidad, especialmente en Etiqueta (235 y 89 etiquetas en Bodega e Isla respectivamente), lo que sugiere una posible sobre segmentación o falta de estandarización. Además, existe un fuerte desequilibrio en la variable Producto, donde el Tomate Saladete domina más del 50% de los registros, lo que

puede influir negativamente en modelos predictivos o análisis comparativos si no se corrige el sesgo.

Las correlaciones entre variables numéricas fueron generalmente débiles. La única relación moderada se dio entre las variables originales y sus transformaciones logarítmicas, lo que valida su aplicación. Sin embargo, la baja correlación entre Cajas y Kilos indica que no existe una relación lineal clara entre cantidad y peso, probablemente por diferencias entre productos o errores de captura.

Finalmente, se identificaron posibles patrones temporales, como una mayor frecuencia de registros en ciertos meses (mayo en Bodega y diciembre en Isla), lo que sugiere una posible estacionalidad en las mermas.

En conjunto, estos hallazgos resaltan la necesidad de un preprocesamiento riguroso, incluyendo limpieza, transformación de variables, manejo de valores atípicos y equilibrio de clases, antes de proceder con modelos de análisis o predicción más avanzados.

Dado que el objetivo del proyecto es la predicción de mermas a lo largo del tiempo, teniendo en cuenta el análisis anterior, podemos observar que los datos dados no son los suficientes para poder una predicción de este tipo, necesitamos más información, por ejemplo, de la cantidad de producto comprado o condiciones del almacenamiento.

Bibliografía

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., y Plöd, M. (2023). CRISP-ML(Q). The ML Lifecycle Process. MLOps. INNOQ. https://ml-ops.org/content/crisp-ml

Kumar Mukhiya, S., y Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python. Packt Publishing. https://learning.oreilly.com/library/view/hands-on-exploratory-data/9781789537253/0957090f-fa4d-4145-95dd-6d3782e5c04d.xhtml