**Logotipo

Descripción generada automáticamente**

**Segmentación y análisis de patrones, para la identificación de productos rentables**

Ricardo Lastra Lopera

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
DAVID MANUEL VILLANUEVA VALDES, M.Sc.

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Lastra Lopera, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Lastra Lopera, R. (2023). *Segmentación y análisis de patrones, para la identificación de productos rentables*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, Cohorte V.

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Ricaro Lastra Lopera**

**Dedicatoria**

Quiero dedicar este trabajo a las dos personas más importantes en mi vida, quienes han sido mi fuente constante de inspiración y apoyo incondicional durante este viaje académico.

A mi esposa porque me amas y crees en mí, esta también es tu victoria, gracias por estar a mi lado, por tolerarme, por entenderme y por creer en mí todo el tiempo. Gracias, hijo, por la alegría y curiosidad dentro de ti que siempre han sido mi fuerza motivadora.

**Agradecimientos**

Quiero expresar mi más profundo agradecimiento a mi familia, cuyo apoyo ha sido fundamental en la consecución de este logro académico.

A mi esposa, por tu amor, paciencia y comprensión inagotables. Tu constante apoyo y tus palabras de aliento han sido mi mayor motivación en este arduo camino. Gracias por estar siempre a mi lado.

A mi hijo, por tu alegría y curiosidad que me han inspirado a seguir adelante. Tu presencia ha sido una fuente constante de energía.

A mis padres, por inculcarme desde pequeño el valor del esfuerzo y la perseverancia. Sin sus enseñanzas y su apoyo, este logro no habría sido posible.

A mis hermano, por su constante apoyo y por creer en mí en todo momento. Gracias por sus palabras de aliento y por estar siempre dispuestos a ayudarme.

**Tabla de contenido**

[Resumen 8](#_Toc167735477)

[Abstract 9](#_Toc167735478)

[1. Descripción del problema 10](#_Toc167735479)

[1.1. Problema de negocio 10](#_Toc167735480)

[1.2. Aproximación desde la analítica de datos 11](#_Toc167735481)

[1.3. Origen de los datos 12](#_Toc167735482)

[1.4. Métricas de desempeño 13](#_Toc167735483)

[2. Objetivos 14](#_Toc167735484)

[2.1. Objetivo general 14](#_Toc167735485)

[2.2. Objetivos específicos 15](#_Toc167735486)

[3. Datos 16](#_Toc167735487)

[3.1. Datos originales 16](#_Toc167735488)

[3.2. Datsets 17](#_Toc167735489)

[3.3. Analítica descriptiva 17](#_Toc167735490)

[Referencias 18](#_Toc167735491)

**Lista de figuras**

[Figure 1 resultado eliminación datos duplicado 19](#_Toc167737812)

[Figure 2 Normalización columnas 20](#_Toc167737813)

[Figure 3 Agrupar variables categóricas 22](#_Toc167737814)

[Figure 4 frecuencia variables categóricas 22](#_Toc167737815)

[Figure 5 Relación variables numéricas 23](#_Toc167737816)

[Figure 6 Correlacion de variables 24](#_Toc167737817)

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**Esp.** Especialista

**MP** Magistrado Ponente

**MSc** Magister Scientiae

**UdeA** Universidad de Antioquia

# Resumen

El proyecto busca mejorar la selección de productos para vendedores en plataformas de comercio electrónico, con el objetivo de aumentar las ventas y rentabilidad. Utilizando datos históricos de ventas, calificaciones y categorías de productos recolectados durante los últimos dos años, se desarrollará un modelo de clustering mediante la técnica de K-means.

La estrategia del proyecto incluye la recolección y preparación de datos, seguido por un análisis exploratorio para comprender la distribución y relaciones entre variables. Posteriormente, se implementará el modelo de K-means, determinando el número óptimo de clusters mediante el método del codo, y las métricas de evaluación de desempeño incluirán: Inertia (Within-Cluster Sum of Squares - WCSS), Silhouette Score, Elbow Method entre otras, adicional evaluaremos las ventas mensuales y el retorno de la inversión

Este proyecto no solo ayudara a mejorar la competitividad de los vendedores en el mercado de comercio electrónico, sino que también proporciona una base sólida para futuras investigaciones y optimizaciones en la selección de productos rentables.

*Palabras clave*: Segmentación de productos, Análisis de patrones, Comercio electrónico, K-means, Clustering, Datos históricos de ventas, Optimización de producto, Estrategias de comercialización, Rentabilidad de productos, Retorno de la inversión (ROI).

Lastra Lopera, R. (2024). Segmentación y Análisis de Patrones (v1.0).GitHub. <https://github.com/rlastralop/seminario>

# Abstract

The project aims to improve product selection for sellers on e-commerce platforms to increase sales and profitability. Utilizing historical data on sales, ratings, and product categories collected over the past two years, a clustering model will be developed using the K-means technique.

The project strategy includes data collection and preparation, followed by exploratory analysis to understand the distribution and relationships between variables. Subsequently, the K-means model will be implemented, determining the optimal number of clusters using the elbow method, and performance evaluation metrics will include Inertia (Within-Cluster Sum of Squares - WCSS), Silhouette Score, Elbow Method among others. Additionally, we will evaluate monthly sales and return on investment (ROI).

This project will not only help improve the competitiveness of sellers in the e-commerce market but also provide a solid foundation for future research and optimizations in the selection of profitable products.

Keywords:

Product segmentation, Pattern analysis, E-commerce, K-means, Clustering, Historical sales data, Product optimization, Marketing strategies, Product profitability, Return on investment (ROI).

# Descripción del problema

Los vendedores en plataformas de comercio electrónico necesitan optimizar la selección de sus productos para incrementar sus ventas y rentabilidad. Para lograr esto, se analizarán datos históricos provenientes de la plataforma de líder en comercio Electrónico Amazon . Estos datos incluyen registros de ventas, calificaciones de clientes y categorización de productos recolectados durante los últimos dos años. El objetivo es desarrollar un modelo que utilice técnicas avanzadas de clustering, como K-means, para identificar patrones y segmentar los productos de manera efectiva.

Las métricas claves para evaluar el desempeño del modelo incluirán la precisión del modelo, el aumento de ventas mensuales, la Inertia (Within-Cluster Sum of Squares - WCSS) y el Silhouette Score. Además, se analizará el retorno de la inversión (ROI) para determinar la rentabilidad de los productos seleccionados.

Esta estrategia permitirá a los vendedores tomar decisiones informadas basadas en datos, mejorando su competitividad en el mercado de comercio electrónico y proporcionando una base sólida para futuras optimizaciones y estrategias de comercialización..

## Problema de negocio

Empresas o vendedores independientes en plataformas de comercio electrónico enfrentan el desafío de seleccionar y comercializar productos de manera eficaz para maximizar sus ventas y rentabilidad. Con la abundancia de datos disponibles, identificar cuáles productos tienen el mayor potencial de éxito se convierte en una tarea compleja. Sin una estrategia clara y basada en datos, los vendedores pueden tomar decisiones poco optimas, lo que resulta en inventarios mal gestionados, bajas ventas, y una disminución de la rentabilidad.

Actualmente, según el artículo escrito por Sophia (2023), muchos vendedores basan sus decisiones en intuiciones o análisis superficiales de los datos, lo que lleva a una falta de precisión en la selección de productos rentables . Esta ineficiencia se traduce en oportunidades perdidas y en una menor competitividad en el mercado.

El problema se agrava por la necesidad de entender múltiples variables, como las ventas históricas, las calificaciones de los productos, y las tendencias de las categorías. Sin una herramienta adecuada para analizar y segmentar estos datos, los vendedores no pueden identificar patrones clave que diferencian a los productos exitosos de los no exitosos.

Por lo tanto, es crucial desarrollar un modelo de análisis de datos que permita a los vendedores segmentar y categorizar productos de manera efectiva, basándose en datos históricos y patrones identificados. Este modelo no solo mejorará la precisión en la selección de productos, sino que también optimizará las estrategias de comercialización, incrementando así las ventas y la rentabilidad en el competitivo mundo del comercio electrónico.

## Aproximación desde la analítica de datos

Para abordar este problema, se desarrollará un modelo de clustering utilizando la técnica de K-means, que permitirá categorizar los productos de manera efectiva y precisa. Este modelo ayudará a identificar productos con alto potencial de éxito al agruparlos en categorías basadas en patrones y tendencias identificados en los datos históricos.

El proceso comenzará con la recolección y preparación de datos, que incluye información sobre ventas, calificaciones, categorías de productos y otros factores relevantes recolectados durante los últimos dos años. Luego, se realizará un análisis exploratorio de los datos para comprender mejor su distribución y las relaciones entre las variables.

El modelo K-means se implementará para segmentar los productos en clusters homogéneos. Utilizando métricas de evaluación como Inertia (Within-Cluster Sum of Squares - WCSS), Silhouette Score y el método del codo, se determinará el número óptimo de clusters. Estos clusters permitirán identificar patrones distintivos que diferencian a los productos exitosos de los no exitosos.

Además de la segmentación, se analizarán las ventas mensuales y el retorno de la inversión (ROI) para cada cluster, proporcionando una visión integral de la rentabilidad y el rendimiento de los productos. Este enfoque permitirá a los vendedores tomar decisiones informadas sobre la selección y comercialización de productos, optimizando así sus estrategias y mejorando la competitividad en el mercado de comercio electrónico.

## Origen de los datos

Los datos utilizados en este proyecto provienen de un conjunto de archivos CSV, los cuales son extraídos desde la plataforma de comercio electrónico mediante un aplicativo diseñado específicamente para este propósito. El dataset abarca un histórico de los últimos dos años y contiene información detallada y relevante

Estos datos proporcionan una base robusta para el análisis, permitiendo una visión detallada y completa del comportamiento de los productos en el mercado. Al utilizar esta información, el proyecto buscará identificar patrones y tendencias que permitan a los vendedores optimizar su selección de productos, mejorando así sus estrategias de comercialización y aumentando la rentabilidad.

El proceso de análisis incluirá la limpieza y normalización de los datos para asegurar su calidad y consistencia, seguido de un análisis exploratorio para comprender mejor las relaciones y distribuciones de las variables clave. Con esta base sólida, se implementará el modelo de clustering K-means para segmentar los productos y proporcionar recomendaciones estratégicas informadas.

## Métricas de desempeño

Las métricas de desempeño utilizadas para evaluar la efectividad del modelo de clustering incluirán:

* **Inertia (Within-Cluster Sum of Squares - WCSS):** Mide la suma de las distancias al cuadrado de cada punto a su centroide más cercano. Una menor inercia indica que los puntos dentro de un cluster están más cerca entre sí.
* **Silhouette Score:** Evalúa la calidad de los clusters midiendo qué tan similares son los puntos dentro del mismo cluster en comparación con otros clusters. Un valor más alto indica que los clusters están bien definidos.
* **Elbow Method:** Utilizado para determinar el número óptimo de clusters. Se grafica la inercia contra el número de clusters y se identifica el "codo" del gráfico, donde la inercia deja de disminuir significativamente.

Además de estas métricas, se evaluarán indicadores clave de rendimiento relacionados con el negocio, tales como:

* **Ventas mensuales:** Se analizarán las ventas mensuales para cada cluster de productos, permitiendo identificar tendencias y patrones de demanda.
* **Retorno de la inversión (ROI):** Se calculará el ROI para cada producto dentro de cada cluster, proporcionando una visión clara de la rentabilidad.
* **Ingresos generales:** Se monitorizarán los ingresos totales generados por los productos agrupados en los diferentes clusters, permitiendo evaluar el impacto global de las estrategias de comercialización.

Estas métricas no solo permitirán evaluar la precisión y efectividad del modelo de clustering, sino que también ofrecerán insights valiosos sobre el rendimiento y la rentabilidad de los productos en el mercado de comercio electrónico. Con esta información, los vendedores podrán tomar decisiones más informadas y estratégicas, optimizando sus inventarios y estrategias de marketing para maximizar las ventas y la rentabilidad.

# Objetivos

## Objetivo general

El objetivo principal del proyecto es identificar patrones en los datos de ventas y realizar una categorización precisa de los productos para determinar cuáles tienen un alto potencial de ventas y rentabilidad. A través de un análisis detallado y un modelo de clustering utilizando la técnica de K-means, el proyecto buscará descubrir los factores que diferencian a los productos exitosos de los menos exitosos.

El análisis se centrará en identificar patrones y tendencias ocultas en los datos históricos que puedan proporcionar insights valiosos sobre el comportamiento del mercado y las preferencias de los consumidores. Al categorizar los productos en clusters basados en estos patrones, se pretende:

* Distinguir productos con alto potencial de ventas: Identificar aquellos productos que muestran una tendencia constante de buen rendimiento en ventas y alta rentabilidad.
* Optimizar la selección de productos: Ayudar a los vendedores a tomar decisiones más informadas sobre qué productos incluir en su inventario, basándose en datos y análisis sólidos.
* Aumentar la competitividad: Equipar a los vendedores con herramientas y conocimientos que les permitan ser más competitivos en el mercado de comercio electrónico.
* Fomentar el éxito sostenible: Crear una base de datos y un marco de análisis que pueda ser utilizado para futuras investigaciones y optimizaciones continuas.

## Objetivos específicos

* **Obtener datos**: Extraer datos históricos de ventas y costos de los productos de la plataforma de comercio electrónico utilizando un aplicativo especializado.
* **Limpiar los datos**: Realizar un proceso de limpieza de datos para eliminar valores nulos, corregir inconsistencias y normalizar las variables. Asegurar que los datos sean precisos y estén listos para el análisis.
* **Modelo de clasificación:** Implementar un modelo de clustering utilizando la técnica de K-means para segmentar los productos en clusters homogéneos.
* **Identificación de patrones:** Analizar los datos para identificar patrones y tendencias que diferencien a los productos exitosos de los no exitosos en términos de ventas y rentabilidad.
* **Análisis de resultados:** Interpretar los resultados del modelo de clustering para comprender las características comunes dentro de cada cluster.
* **Estrategias de comercialización:** Basado en los patrones identificados, desarrollar recomendaciones estratégicas para optimizar la selección de productos y las estrategias de comercialización. Sugerir tácticas específicas para promover los productos de alto potencial.
* **Desarrollar tecnica de analisis de datos**: Utilizar un modelo de clasificacion usando la tecnica de K-means para segmentar los productos identificando patrones en los datos.
* **Evaluación del modelo:** Probar el modelo de clustering con un conjunto de datos de prueba para validar su precisión y efectividad.
* **Optimización:** Ajustar los parámetros del modelo y mejorar los algoritmos utilizados para maximizar la precisión y utilidad del modelo. Realizar iteraciones para refinar los clusters y asegurar que las recomendaciones estratégicas sean lo más precisas y útiles posible.

# Datos

## Datos originales

Los datos utilizados en este proyecto provienen de múltiples archivos CSV extraídos mediante el aplicativo Helium10. Cada archivo contiene 200 filas y tiene la misma estructura de columnas. Esta consistencia facilita el análisis conjunto de los datos, aunque también presenta desafíos en términos de manejo y procesamiento debido al gran número de archivos.

**Columnas del Dataset**

**URL**: Enlace al producto en la plataforma de comercio electrónico.

**ASIN**: Identificador único del producto en Amazon.

**Título**: Nombre descriptivo del producto.

**Marca**: Marca del producto.

**Cumplimiento**: Método de cumplimiento (FBA, FBM, Amazon).

**Categoría**: Categoría del producto.

**BSR**: Best Sellers Rank (Rango de los más vendidos).

**Subcategoría**: Subcategoría del producto.

**Precio**: Precio actual del producto.

**Tendencia de precios (90 días) (%):** Cambio porcentual en el precio en los últimos 90 días.

**Ventas mensuales:** Número de unidades vendidas mensualmente.

**Tendencia de ventas (90 días) (%):** Cambio porcentual en las ventas en los últimos 90 días.

**Ingresos mensuales:** Ingresos generados mensualmente por el producto.

**Cuenta de reseñas:** Número de reseñas del producto.

**Reseñas Calificación:** Calificación promedio de las reseñas.

**Vendedor:** Nombre del vendedor.

**Número de vendedores activos:** Número de vendedores que ofrecen el producto.

**Ventas del año pasado:** Ventas del producto en el año anterior.

**Ventas año tras año (%):** Cambio porcentual en las ventas año tras año.

**Rango de tamaño:** Clasificación del tamaño del producto.

**Longitud**: Longitud del producto.

**Ancho**: Ancho del producto.

**Alto**: Altura del producto.

**Peso**: Peso del producto.

**Precio almacenamiento (ene-sep):** Costo de almacenamiento del producto de enero a septiembre.

**Precio almacenamiento (oct-dic):** Costo de almacenamiento del producto de octubre a diciembre.

**Mejor período de ventas:** Fecha del mejor período de ventas del producto.

**Antigüedad (mes):** Antigüedad del producto en meses.

**Cantidad de imágenes:** Número de imágenes del producto.

**Recuento de variaciones:** Número de variaciones del producto.

**Ventas / Reseñas:** Proporción de ventas respecto a reseñas

**Tamaño de los Datos**

**Número de Registros:** Cada archivo contiene 200 filas y, dado que se pueden extraer multiples archivos, el numero puede ser variable. Al momento de realizar esta monografía se realizaron con 30 archivos para un total de seis mil (6.000) filas

**Tamaño Total:** El tamaño de un archivo en promedio es de aproximadamente 0.069 MB. el tamaño total es aproximadamente 2.07 MB.

## Datsets

Para construir los datasets de entrenamiento y validación a partir de los datos proporcionados en los archivos CSV, se seguirán los siguientes pasos:

**1. Recolección y Unificación de Datos**

* **Carga de Datos:** Se cargan todos los archivos CSV utilizando una biblioteca de análisis de datos como pandas en Python. Cada archivo contiene 200 registros con la misma estructura de columnas.
* **Concatenación de Archivos:** Todos los archivos CSV se concatenan en un único DataFrame para tener una vista completa de los datos.

**2. Limpieza y Preprocesamiento de Datos**

* **Manejo de Valores Nulos:** Se identifican y tratan los valores nulos. Esto puede incluir la eliminación de registros incompletos o la imputación de valores faltantes.
* **Normalización y Escalado:** Se normalizan o escalan las columnas numéricas para asegurar que los datos estén en una escala común, lo cual es importante para muchos algoritmos de aprendizaje automático.
* **Codificación de Variables Categóricas:** Las columnas categóricas, como "Categoría" y "Marca", se convierten en variables dummy o se codifican mediante la técnica que se conoce como "binning" o "discretización" basada en la frecuencia relativa de las categorías

**3. División del Dataset**

* **Definición de Variables Independientes y Dependientes**: Se seleccionan las variables independientes (features) y la variable dependiente (label) si se va a realizar un análisis supervisado. Por ejemplo, "Ventas mensuales" podría ser la variable dependiente.
* **División en Entrenamiento y Validación**: Se divide el dataset en conjuntos de entrenamiento y validación. Una práctica común es utilizar un 80% de los datos para entrenamiento y un 20% para validación. Esto se puede realizar utilizando la función train\_test\_split de la biblioteca sklearn.

**4. Creación del Dataset de Entrenamiento**

* **Selección de Features:** Se seleccionan las columnas relevantes que se utilizarán como features en el modelo.
* **Dataset de Entrenamiento:** El conjunto de entrenamiento se utiliza para entrenar el modelo. Esto incluye el ajuste del modelo a los datos y la optimización de sus parámetros.

**5. Creación del Dataset de Validación**

* **Evaluación del Modelo:** El conjunto de validación se utiliza para evaluar el desempeño del modelo. Esto permite medir la capacidad del modelo para generalizar a datos nuevos y no vistos durante el entrenamiento.

## Analítica descriptiva

* **Carga y unificación de archivos CSV**

#Carga y unifica todos los archivos CSV alojados en el data Storage. carga con spark y conversion a datagrame con pandas

files = dbutils.fs.ls(f"{MOUNTPOINT}/amazon\_data")

csv\_files = [file.path for file in files if file.name.startswith('US\_AMAZON\_blackBoxProducts\_') and file.name.endswith('.csv')]

df\_amazon\_spark = spark.read.csv(path=csv\_files, sep=',', header=True, quote='"', escape='"')

df\_amazon = df\_amazon\_spark.toPandas()

df\_amazon.head(15)

* **Datos duplicados**

# Identificación y eliminación de duplicados

duplicados\_en\_columna = df\_amz\_sales.duplicated()

df\_amz\_sales= df\_amz\_sales.drop\_duplicates()

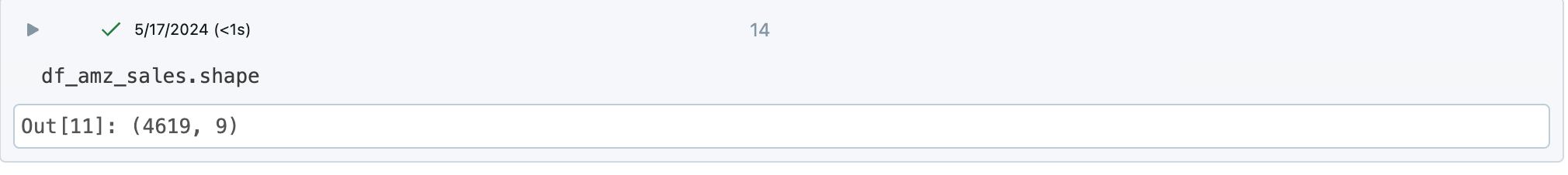
****

Figure 1 resultado eliminación datos duplicado

* **Casteo de datos**

#Se realiza la conversion de las columnas a su tipo especifico, en este caso ya que el dataset llega todo en formato String, se realiza su conversion a tipo numerico

df\_amz\_sales['ventas\_mensuales'] = pd.to\_numeric(df\_amz\_sales['ventas\_mensuales'], errors='coerce')

df\_amz\_sales['ingresos\_mensuales'] = pd.to\_numeric(df\_amz\_sales['ingresos\_mensuales'], errors='coerce')

df\_amz\_sales['resenas\_calificacion'] = pd.to\_numeric(df\_amz\_sales['resenas\_calificacion'], errors='coerce')

df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'] = pd.to\_numeric(df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'], errors='coerce')

df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'] = pd.to\_numeric(df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'], errors='coerce')

df\_amz\_sales['ventas\_del\_anio\_pasado'] = pd.to\_numeric(df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'], errors='coerce')

#normalizacion campos en porcentaje para dejarlos en formato a base de 100

df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'] = df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'] / 100

df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'] = df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'] / 100

# se castean las columnas a tipo float.

df\_amz\_sales['precio'] = df\_amz\_sales['precio'].astype(float)

df\_amz\_sales['ventas\_mensuales'] = df\_amz\_sales['ventas\_mensuales'].astype(float)

df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'] = df\_amz\_sales['tendencia\_de\_precios\_(90\_dias)\_(%)'].astype(float)

df\_amz\_sales['ingresos\_mensuales'] = df\_amz\_sales['ingresos\_mensuales'].astype(float)

df\_amz\_sales['resenas\_calificacion'] = df\_amz\_sales['resenas\_calificacion'].astype(float)

df\_amz\_sales['ventas\_del\_anio\_pasado'] = df\_amz\_sales['ventas\_del\_anio\_pasado'].astype(float)

df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'] = df\_amz\_sales['ventas\_anio\_tras\_anio\_(%)'].astype(float)

**A screenshot of a computer

Description automatically generated**

Figure 2 Normalización columnas

* **Segmentación de variables categoricas**

#Segmentacion de la columna subcategoria, para crear 3 grupos principales segun su frecuencia para asi facilitar el analisis de las variables categoricas. se dividen en tres grupos, "Alta Frecuencia","MediaFrecuencia","Baja Frecuencia" tanto para esta variable categorica como para la variable marca, se realiza con el metodo de cuartil para dividir los grupos, al final a ambas variables se le realiza un ajusto manual para tener un equilibrio en los grupos

freq\_counts = df\_amz\_sales['subcategoria'].value\_counts(normalize=True)

grupo\_alto\_frecuencia = freq\_counts[freq\_counts >= freq\_counts.quantile(0.85)].index.tolist()

grupo\_medio\_frecuencia = freq\_counts[(freq\_counts >= freq\_counts.quantile(0.33)) & (freq\_counts < freq\_counts.quantile(0.85))].index.tolist()

grupo\_bajo\_frecuencia = freq\_counts[freq\_counts < freq\_counts.quantile(0.33)].index.tolist()

grupos = {

'Alta Frecuencia:': grupo\_alto\_frecuencia,

'MediaFrecuencia:': grupo\_medio\_frecuencia,

'Baja Frecuencia:': grupo\_bajo\_frecuencia,

}

grupos

# Crear nueva columna con los datos de la segmentacion anterior para la columna categoria

def mapear\_grupo(subcategoria):

for grupo, subcategorias in grupos.items():

if subcategoria in subcategorias:

return grupo

#return 'Otro'

df\_amz\_sales['grupo\_categoria'] = df\_amz\_sales['subcategoria'].map(mapear\_grupo)



Figure 3 Agrupar variables categóricas

* **Visualizacion de datos**

A screenshot of a graph

Description automatically generated

Figure 4 frecuencia variables categóricas

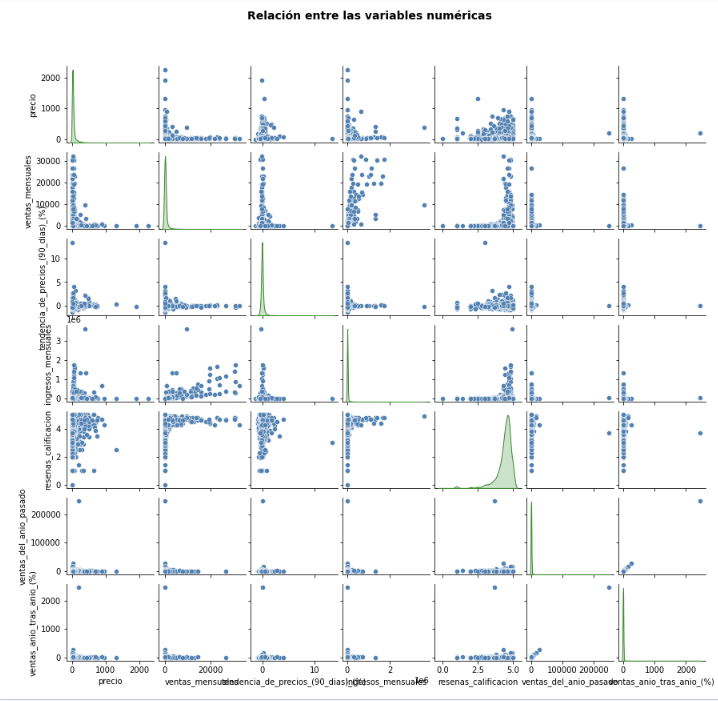


Figure 5 Relación variables numéricas

A screenshot of a computer screen

Description automatically generated

Figure 6 Correlacion de variables

## Referencias

Sohpia. (2023). Que porcentaje de vendedores de amazon fracasan - amzalgo.com. https://medium.com/@libradafigg42/que-porcentaje-de-vendedores-de-amazon-fracasan-amzalgo-com-2dccd18826fc