Informe Técnico del Proyecto
Análisis Exploratorio de Datos (EDA)

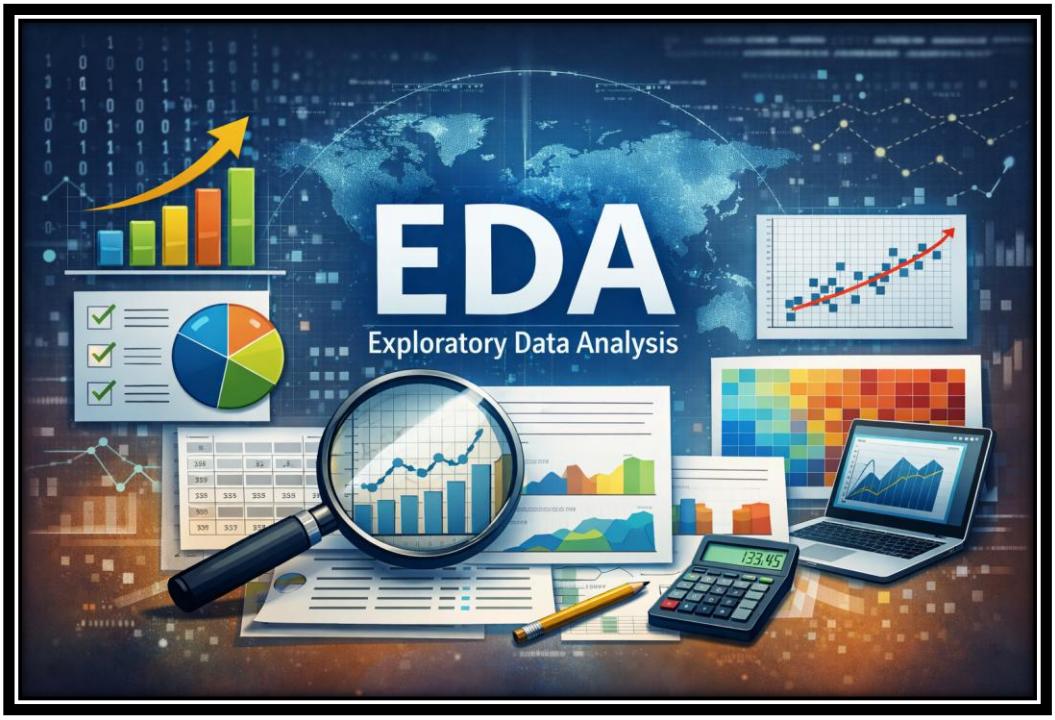
Proyecto Módulo 4 – Ciencia de Datos
Autora: Carolina Tapia Bahamonde
Fecha :08 de febrero 2026

1. Introducción

El presente informe técnico documenta el desarrollo del Proyecto del Módulo 4 de Análisis Exploratorio de Datos (EDA), cuyo propósito es aplicar técnicas estadísticas y de visualización para comprender un conjunto de datos reales y extraer información relevante que apoye la toma de decisiones comerciales.

El Análisis Exploratorio de Datos constituye una etapa fundamental dentro de cualquier proceso de análisis de datos, ya que permite conocer la estructura del dataset, identificar patrones, detectar valores atípicos y explorar relaciones entre variables antes de la construcción de modelos más complejos. En este contexto, el proyecto se orienta al análisis de datos de comercio electrónico, con especial énfasis en el comportamiento de los precios y descuentos de productos.

A través del uso de herramientas de ciencia de datos y un enfoque metodológico sistemático, este trabajo busca no solo describir los datos, sino también generar interpretaciones útiles desde una perspectiva comercial, demostrando la relevancia del EDA como apoyo a la gestión y a la toma de decisiones basada en datos.



2. Contexto del proyecto.

El análisis desarrollado en este proyecto se enmarca en el contexto de la empresa ficticia ComercioYA, una organización dedicada al comercio electrónico y a la comercialización de productos a través de plataformas digitales. En este tipo de entorno, la gestión eficiente de precios y descuentos constituye un elemento clave para la competitividad, la atracción de clientes y la optimización de ingresos.

Las empresas de comercio electrónico operan en mercados altamente dinámicos, donde las decisiones relacionadas con estrategias de precios, promociones y descuentos deben basarse en el análisis sistemático de datos históricos y actuales. En este escenario, el uso de herramientas de análisis de datos permite identificar patrones de comportamiento, evaluar el impacto de las políticas comerciales y apoyar la toma de decisiones informadas.

En este proyecto, ComercioYA se utiliza como un caso de estudio para aplicar técnicas de Análisis Exploratorio de Datos sobre un conjunto de datos de productos e-commerce, con el objetivo de analizar la relación entre precios originales, precios con descuento y porcentajes de descuento. Este enfoque permite simular un escenario real de análisis comercial, alineado con las necesidades habituales de gestión en organizaciones digitales.

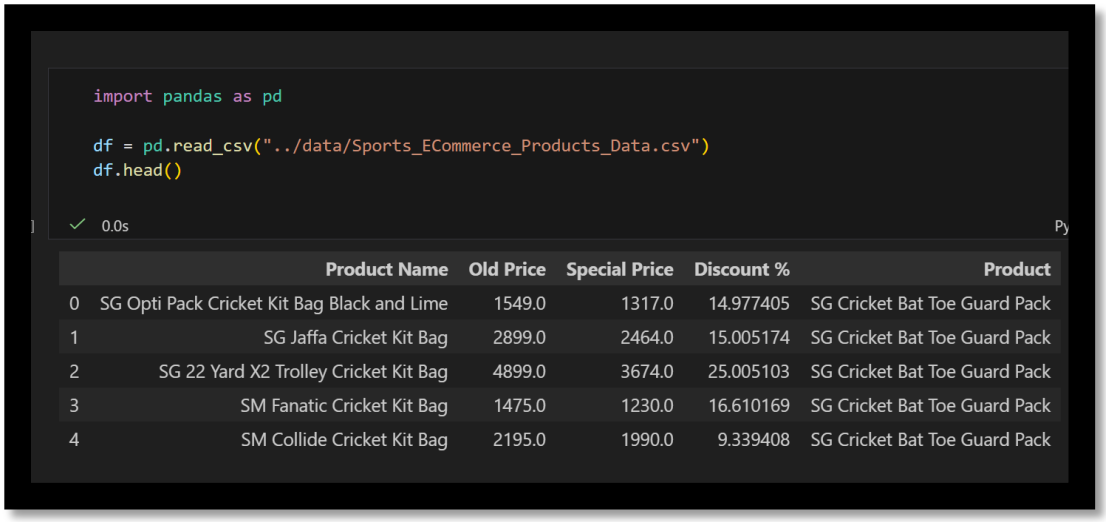
3. Dataset

Descripción del dataset

El dataset utilizado en este proyecto corresponde a un conjunto de datos de productos de comercio electrónico, obtenido desde la plataforma Kaggle, repositorio ampliamente utilizado para proyectos de análisis de datos y aprendizaje automático. El uso de este tipo de fuente permite trabajar con datos reales y representativos de escenarios comerciales actuales.

https://www.kaggle.com/datasets/shreypachauri123/ecommerce

El conjunto de datos contiene información asociada a productos e-commerce, incluyendo variables numéricas relacionadas con el precio original del producto, el precio final con descuento y el porcentaje de descuento aplicado. Adicionalmente, el dataset incorpora variables categóricas que permiten identificar y clasificar los productos, lo que resulta relevante para el análisis exploratorio y la segmentación básica.



```
import pandas as pd

df = pd.read_csv("../data/Sports_ECommerce_Products_Data.csv")
df.head()
```

✓ 0.0s

	Product Name	Old Price	Special Price	Discount %	Product
0	SG Opti Pack Cricket Kit Bag Black and Lime	1549.0	1317.0	14.977405	SG Cricket Bat Toe Guard Pack
1	SG Jaffa Cricket Kit Bag	2899.0	2464.0	15.005174	SG Cricket Bat Toe Guard Pack
2	SG 22 Yard X2 Trolley Cricket Kit Bag	4899.0	3674.0	25.005103	SG Cricket Bat Toe Guard Pack
3	SM Fanatic Cricket Kit Bag	1475.0	1230.0	16.610169	SG Cricket Bat Toe Guard Pack
4	SM Collide Cricket Kit Bag	2195.0	1990.0	9.339408	SG Cricket Bat Toe Guard Pack

Desde el punto de vista analítico, el dataset es especialmente adecuado para el estudio de estrategias de precios y descuentos, ya que permite explorar la relación entre el valor inicial de los productos y las políticas de descuento aplicadas. Este tipo de información es fundamental en entornos de comercio electrónico, donde la definición de precios influye directamente en la competitividad y en el comportamiento de compra de los clientes.

Durante la etapa inicial de análisis, se verificó que el dataset presenta una estructura consistente y una calidad adecuada para el desarrollo del Análisis Exploratorio de Datos, lo que permitió avanzar con las etapas posteriores sin la necesidad de realizar procesos complejos de limpieza o imputación de datos.

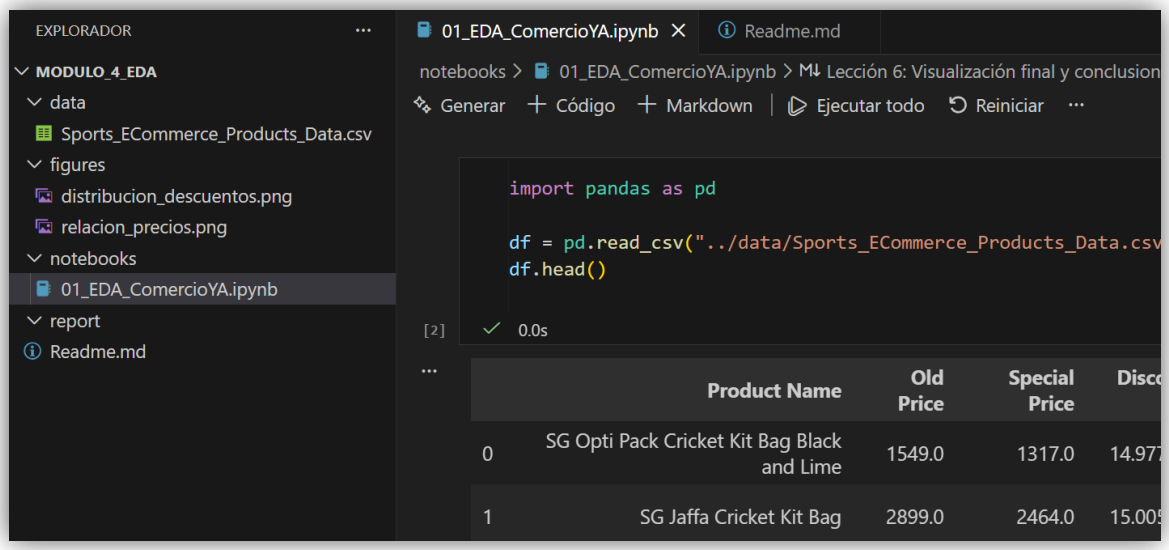
4. Entorno de desarrollo y configuración

El desarrollo del proyecto se realizó utilizando Visual Studio Code como entorno de desarrollo integrado, junto con Jupyter Notebook para la ejecución interactiva del análisis exploratorio de datos. Esta combinación permite trabajar de forma modular, documentada y reproducible, facilitando tanto el análisis como la presentación de resultados.

El lenguaje de programación utilizado fue Python, debido a su amplio uso en ciencia de datos y a la disponibilidad de librerías especializadas para análisis estadístico y visualización. Inicialmente, se evaluó el uso de versiones recientes de Python 3.13.x; sin embargo, durante la configuración del entorno se identificaron dificultades de compatibilidad y estabilidad del kernel de Jupyter en sistema operativo Windows, lo que afectaba la correcta ejecución de las celdas del notebook.

Como medida correctiva, se optó por utilizar Python versión 3.11.x, versión ampliamente estable y recomendada para proyectos de análisis de datos. Este ajuste permitió asegurar la correcta integración con Jupyter Notebook y el funcionamiento adecuado de las librerías requeridas para el desarrollo del proyecto.

La estructura del proyecto se organizó mediante carpetas diferenciadas para datos (`data`), notebooks (`notebooks`), visualizaciones (`figures`) e informe (`report`), siguiendo buenas prácticas de organización y facilitando la trazabilidad y reutilización del trabajo. Esta configuración permitió mantener un flujo de trabajo ordenado y alineado con estándares profesionales



5. Librerías utilizadas

Para el desarrollo del Análisis Exploratorio de Datos se utilizaron librerías estándar del ecosistema de ciencia de datos en Python, seleccionadas por su estabilidad, amplio uso y adecuación a los objetivos del proyecto.

La librería **pandas** fue empleada como herramienta principal para la carga, manipulación y análisis del dataset, permitiendo trabajar eficientemente con estructuras de datos tabulares y realizar operaciones estadísticas descriptivas. La librería **numpy** se utilizó como apoyo para operaciones numéricas y cálculos matemáticos, particularmente en el cálculo de métricas de error asociadas al modelo de regresión, tales como el error cuadrático medio (MSE) y el error absoluto medio (MAE).Para la visualización de datos, se emplearon las librerías **matplotlib** y **seaborn**. Matplotlib fue utilizada para la creación de gráficos personalizados y la exportación de visualizaciones finales, mientras que Seaborn permitió generar visualizaciones estadísticas avanzadas, como pairplots, violinplots y jointplots, facilitando la interpretación visual de las relaciones entre variables.

La librería **statsmodels** fue utilizada para el modelamiento estadístico, específicamente para la implementación del modelo de regresión lineal simple. Esta herramienta permitió obtener información detallada sobre los coeficientes del modelo, su significancia estadística y métricas de ajuste.

Durante el desarrollo del proyecto, se incorporó además la librería **scikit-learn (sklearn)** para el cálculo de métricas de evaluación del modelo de regresión, tales como el error cuadrático medio (MSE) y el error absoluto medio (MAE). Si bien posteriormente se evaluó una alternativa utilizando funciones de `numpy`, el uso de `sklearn` permitió contrastar resultados y validar el desempeño del modelo. Adicionalmente, se utilizó la librería estándar **os** para la gestión de rutas y verificación de archivos dentro del entorno de trabajo, contribuyendo a una correcta organización del proyecto y a la reproducibilidad del análisis.El uso conjunto de estas librerías permitió desarrollar un análisis completo, coherente y alineado con buenas prácticas en proyectos de análisis exploratorio de datos.

6. Metodología

La metodología aplicada en este proyecto corresponde al enfoque de Análisis Exploratorio de Datos (EDA), el cual tiene como objetivo comprender la estructura y características de un conjunto de datos antes de la aplicación de modelos más complejos o técnicas predictivas. Este enfoque permite identificar patrones, detectar valores atípicos, evaluar relaciones entre variables y generar hipótesis relevantes desde una perspectiva analítica y comercial.

El desarrollo del EDA se realizó de forma secuencial y sistemática, comenzando con un análisis inicial de los datos para evaluar su calidad y estructura, seguido de un análisis estadístico descriptivo que permitió comprender la distribución y variabilidad de las variables numéricas. Posteriormente, se exploraron las relaciones entre variables mediante análisis de correlación y visualizaciones, lo que facilitó la identificación de asociaciones relevantes.

Como parte del proceso metodológico, se incorporó un modelo de regresión lineal simple con el propósito de cuantificar la relación entre el precio original y el precio con descuento. Este modelo se utilizó como una herramienta exploratoria complementaria, permitiendo reforzar los hallazgos obtenidos en las etapas previas del análisis.

Finalmente, se emplearon técnicas de visualización avanzada para profundizar en la interpretación de los datos y comunicar de manera clara los resultados obtenidos. Este enfoque metodológico permitió desarrollar un análisis coherente, reproducible y alineado con los objetivos del proyecto, manteniendo siempre un vínculo directo con la toma de decisiones comerciales.

7. Desarrollo del análisis

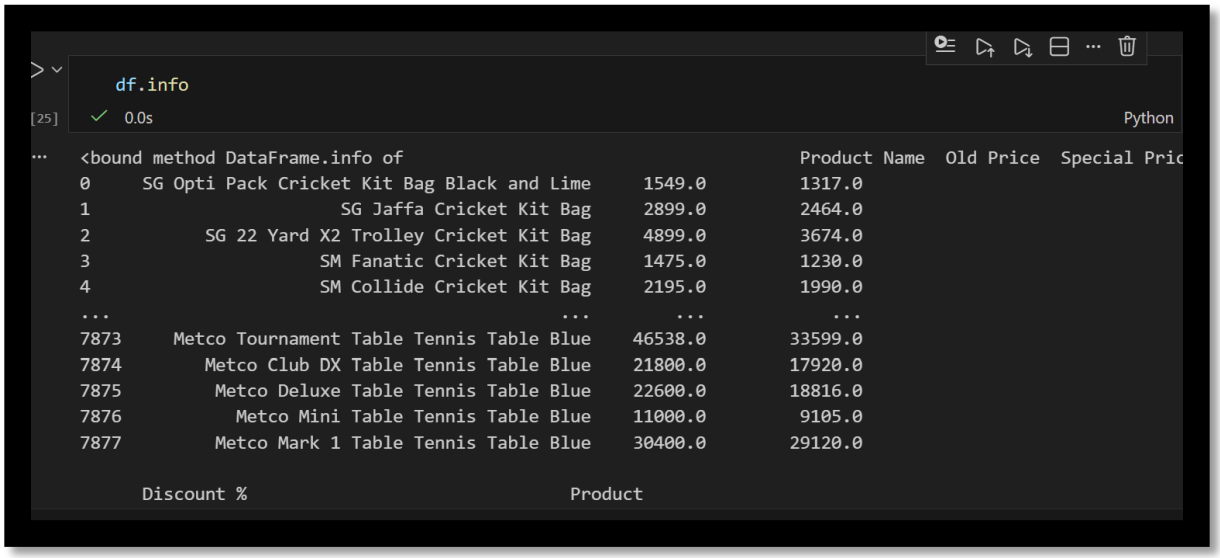
7.1 Análisis Inicial de Datos (IDA)

Como primera etapa del desarrollo del análisis, se realizó un Análisis Inicial de Datos (IDA) con el objetivo de comprender la estructura general del dataset y evaluar su calidad antes de aplicar técnicas estadísticas y de visualización más avanzadas.

En esta fase se verificó el número de registros y variables presentes en el conjunto de datos, así como los tipos de datos asociados a cada variable. El dataset contiene tanto variables numéricas, principalmente relacionadas con precios y descuentos, como variables categóricas que permiten identificar y clasificar los productos.

Adicionalmente, se evaluó la presencia de valores nulos y posibles inconsistencias en los datos. El análisis inicial evidenció que el dataset presenta una estructura consistente y un nivel adecuado de completitud, lo que permitió avanzar con las etapas posteriores del Análisis Exploratorio de Datos sin requerir procesos complejos de limpieza o imputación.

Este análisis preliminar fue fundamental para asegurar la fiabilidad de los resultados obtenidos en las siguientes etapas y para definir el enfoque metodológico del análisis, centrado en la exploración de precios, descuentos y sus relaciones.



```
df.info
[25]  ✓ 0.0s                                                    Python

<bound method DataFrame.info of                                        Product Name  Old Price  Special Pric
0     SG Opti Pack Cricket Kit Bag Black and Lime     1549.0        1317.0
1                         SG Jaffa Cricket Kit Bag     2899.0        2464.0
2           SG 22 Yard X2 Trolley Cricket Kit Bag     4899.0        3674.0
3                     SM Fanatic Cricket Kit Bag     1475.0        1230.0
4                     SM Collide Cricket Kit Bag     2195.0        1990.0
...                                          ...        ...           ...
7873     Metco Tournament Table Tennis Table Blue    46538.0       33599.0
7874        Metco Club DX Table Tennis Table Blue    21800.0       17920.0
7875         Metco Deluxe Table Tennis Table Blue    22600.0       18816.0
7876           Metco Mini Table Tennis Table Blue    11000.0        9105.0
7877         Metco Mark 1 Table Tennis Table Blue    30400.0       29120.0

      Discount %                              Product
```

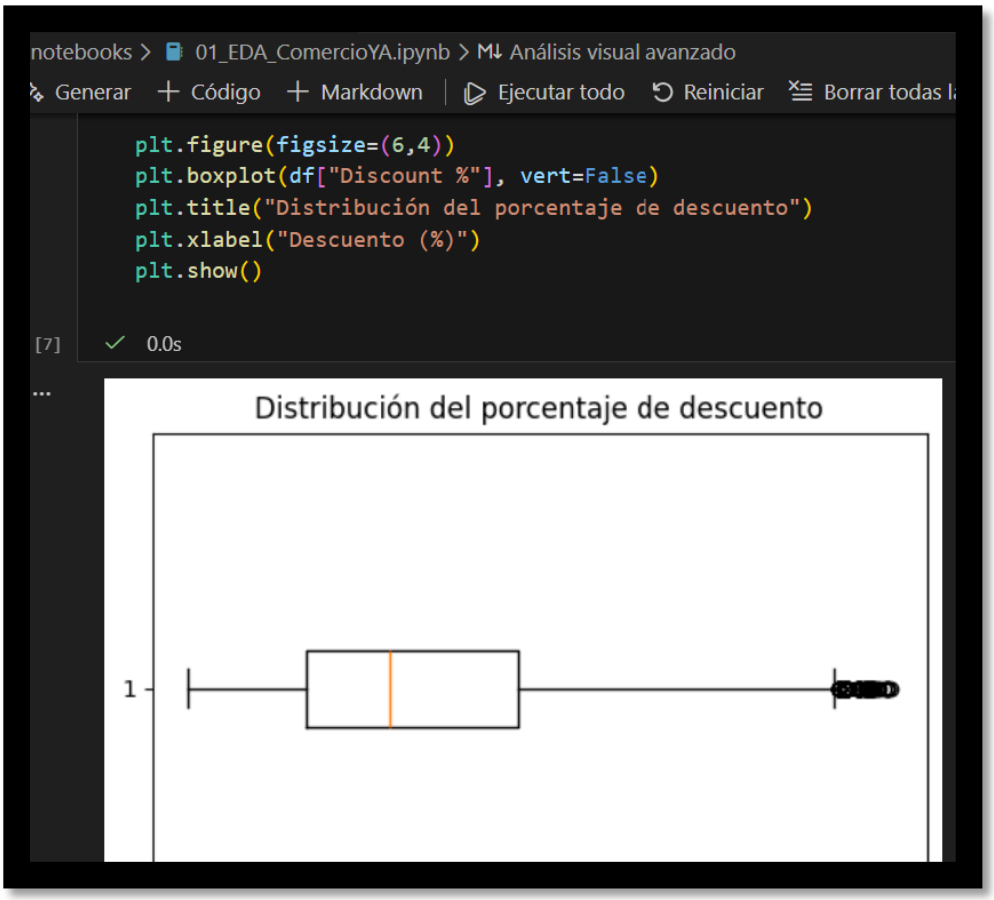
7.2 Estadística descriptiva

Una vez validada la estructura y calidad del dataset, se procedió al análisis de estadística descriptiva con el objetivo de comprender el comportamiento general de las variables numéricas asociadas a precios y descuentos.

En esta etapa se calcularon medidas de tendencia central, tales como la media y la mediana, así como medidas de dispersión como la desviación estándar y el rango. Estos indicadores permitieron identificar diferencias significativas en los valores de precios originales y precios con descuento, evidenciando la variabilidad presente en el conjunto de datos.

Complementariamente, se utilizaron visualizaciones como histogramas y diagramas de caja (boxplots) para analizar la distribución de los datos y detectar la presencia de valores atípicos. Los resultados muestran que, si bien la mayoría de los productos se concentra en determinados rangos de precios, existen valores extremos que reflejan una amplia diversidad de productos y estrategias de precio dentro del comercio electrónico analizado.

Este análisis descriptivo proporcionó una base sólida para comprender la estructura de los datos y sirvió como insumo fundamental para las etapas posteriores de análisis de correlación y modelamiento.



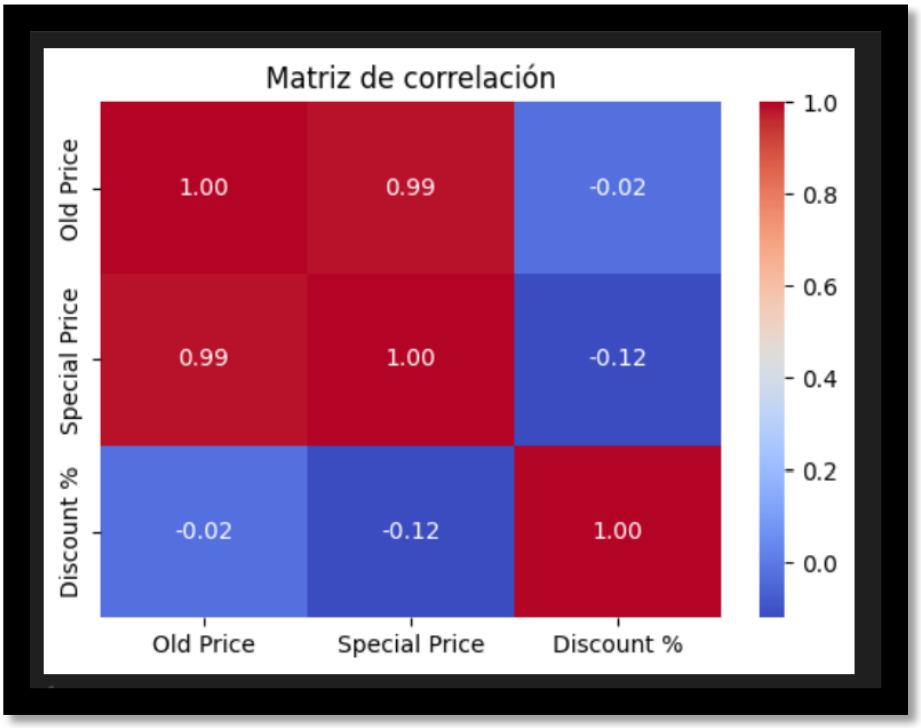
7.3 Análisis de correlación

Con el fin de explorar las relaciones existentes entre las variables numéricas del dataset, se realizó un análisis de correlación enfocado principalmente en el precio original, el precio con descuento y el porcentaje de descuento aplicado a los productos.

Para este propósito, se utilizaron gráficos de dispersión (scatterplots) que permitieron visualizar de manera clara la relación entre pares de variables, así como el cálculo del coeficiente de correlación de Pearson y su representación mediante un mapa de calor (heatmap). Estas herramientas facilitaron la identificación de asociaciones relevantes entre las variables analizadas.

Los resultados del análisis evidencian una relación positiva fuerte entre el precio original y el precio con descuento, lo que resulta coherente desde una perspectiva comercial, ya que los productos de mayor valor tienden a mantener precios finales más elevados incluso después de aplicar descuentos. En contraste, el porcentaje de descuento presenta un comportamiento más variable, sin una relación lineal directa claramente definida con los precios.

Es importante destacar que, si bien se identifican correlaciones entre determinadas variables, estas relaciones no implican necesariamente una relación de causalidad. El análisis de correlación se utilizó como una herramienta exploratoria para orientar la comprensión de los datos y apoyar las etapas posteriores del análisis.



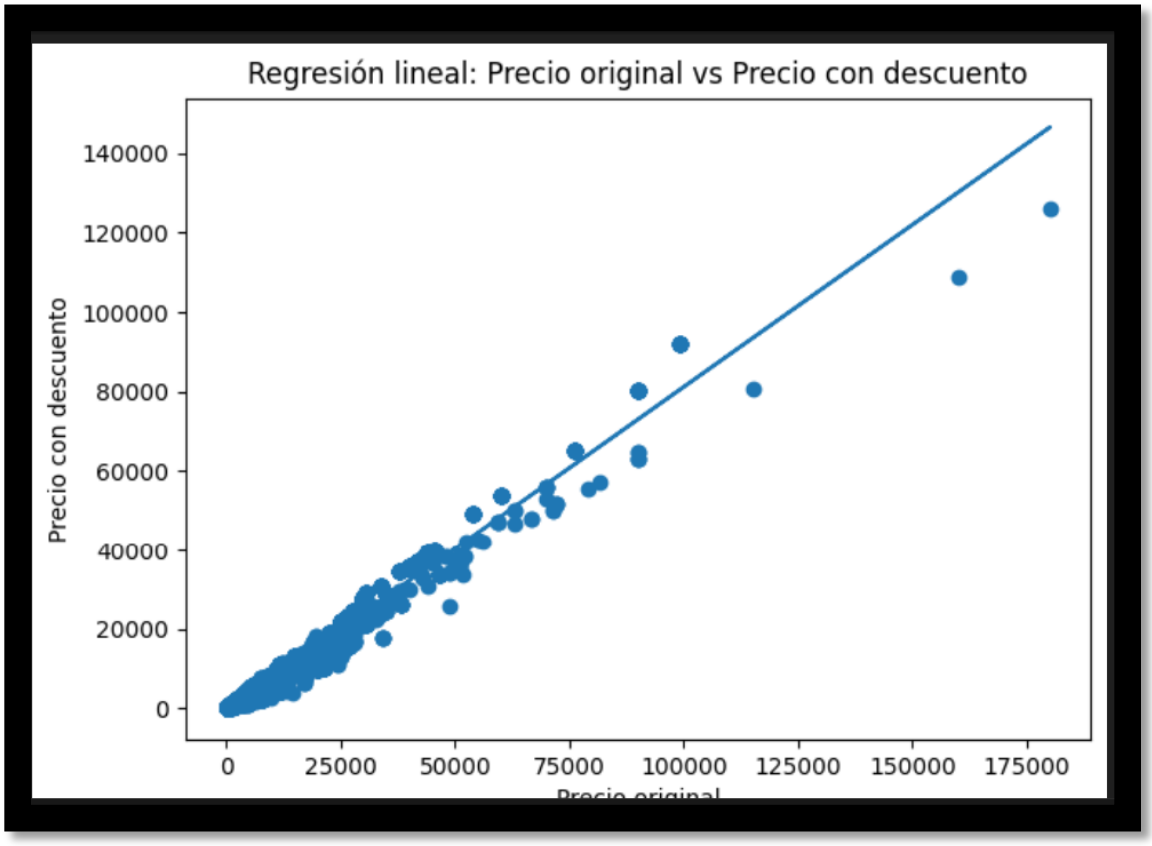
7.4 Regresión lineal

Como complemento al análisis de correlación, se implementó un modelo de regresión lineal simple con el objetivo de cuantificar la relación existente entre el precio original del producto y su precio con descuento. En este modelo, el precio original fue considerado como variable independiente, mientras que el precio con descuento se definió como variable dependiente.

El modelo de regresión se utilizó con un enfoque exploratorio, permitiendo estimar cómo varía el precio con descuento en función del precio original. La implementación se realizó utilizando herramientas de modelamiento estadístico, lo que permitió obtener información detallada sobre el ajuste del modelo y la significancia de los coeficientes estimados.

Para evaluar el desempeño del modelo, se consideraron métricas de error como el error cuadrático medio (MSE) y el error absoluto medio (MAE), además del coeficiente de determinación (R^2). Estas métricas permitieron evaluar el grado en que el modelo logra representar la relación observada entre las variables, sin pretender construir un modelo predictivo final.

Los resultados del modelo refuerzan los hallazgos obtenidos en las etapas previas del análisis, confirmando la existencia de una relación lineal consistente entre el precio original y el precio con descuento, lo que resulta coherente con el comportamiento esperado en un entorno de comercio electrónico.



7.5 Visualización avanzada

Con el propósito de profundizar en la comprensión de los datos y enriquecer la interpretación de los resultados obtenidos, se incorporaron técnicas de visualización avanzada. Este tipo de visualizaciones permite analizar simultáneamente múltiples variables y facilita la identificación de patrones que no siempre son evidentes mediante análisis estadísticos básicos.

En esta etapa se utilizaron gráficos multivariados, tales como pairplots, violinplots y jointplots, los cuales permitieron observar de manera integrada la distribución de las variables, sus relaciones y la densidad de los datos. Estas visualizaciones complementan el análisis previo, aportando una visión más completa del comportamiento de los precios y de los porcentajes de descuento.

La visualización avanzada cumplió un rol clave en la comunicación de los resultados del análisis, ya que permitió representar de forma clara y comprensible relaciones complejas entre variables. De este modo, los gráficos se constituyen como una herramienta fundamental para apoyar la interpretación analítica y facilitar la toma de decisiones en contextos comerciales.

8. Resultados y hallazgos

El Análisis Exploratorio de Datos realizado permitió identificar una serie de hallazgos relevantes respecto al comportamiento de los precios y descuentos en el comercio electrónico analizado. A partir del análisis estadístico y visual, se evidenció una estructura de precios coherente, con una relación consistente entre el precio original de los productos y su precio con descuento.

Los resultados del análisis descriptivo muestran una alta variabilidad en los valores de precios y porcentajes de descuento, lo que refleja la diversidad de productos y la aplicación de distintas estrategias comerciales dentro del conjunto de datos. La presencia de valores atípicos sugiere la existencia de productos de alto valor o promociones específicas, elementos habituales en entornos de comercio electrónico.

El análisis de correlación y el modelo de regresión lineal confirmaron la existencia de una relación positiva entre el precio original y el precio con descuento, reforzando la idea de que el precio inicial es un factor determinante en la definición del precio final. Sin embargo, el porcentaje de descuento mostró un comportamiento más heterogéneo, lo que indica que las políticas de descuento no siguen un patrón único y pueden variar según el tipo o valor del producto.

En conjunto, estos resultados aportan una visión integral del comportamiento de los precios y descuentos, proporcionando información relevante que puede ser utilizada como base para

la evaluación y optimización de estrategias comerciales en plataformas de comercio electrónico.

9. Conclusiones

El desarrollo del presente proyecto permitió aplicar de manera práctica y sistemática las técnicas de Análisis Exploratorio de Datos sobre un conjunto de datos reales de comercio electrónico, evidenciando la utilidad del EDA como una herramienta fundamental para la comprensión y análisis de información comercial.

A lo largo del análisis se logró identificar patrones relevantes en el comportamiento de los precios y descuentos, así como relaciones consistentes entre variables clave, tales como el precio original y el precio con descuento. Estos hallazgos fueron respaldados mediante análisis estadístico, visualizaciones y un modelo de regresión lineal con enfoque exploratorio, lo que permitió reforzar la interpretación de los resultados desde distintas perspectivas.

El proyecto también puso de manifiesto la importancia de contar con un entorno de desarrollo correctamente configurado y con el uso adecuado de librerías especializadas, aspectos clave para garantizar la reproducibilidad y confiabilidad del análisis. Asimismo, la integración de visualizaciones avanzadas facilitó la comunicación de resultados y la comprensión de relaciones complejas entre variables.

En conclusión, el Análisis Exploratorio de Datos realizado aporta información valiosa que puede servir como base para la toma de decisiones comerciales orientadas a la definición de estrategias de precios y promociones en entornos de comercio electrónico, reafirmando el valor del análisis de datos como apoyo a la gestión empresarial.