Memoria

El presente documento describe el proceso y los hallazgos obtenidos durante el Análisis Exploratorio de Datos realizado sobre un dataset correspondiente a las operaciones de una tienda. Este dataset incluye información detallada sobre pedidos, compradores, productos y transacciones. El objetivo principal del EDA fue comprender la estructura de los datos, identificar patrones relevantes y posibles inconsistencias, y establecer una base para futuros análisis. Los datos fueron obtenidos desde Kaggle, una web que ofrece datos sin ningún coste para poder analizarlos y como nuestro caso practicar.

El dataset contiene las siguientes columnas:

Columna	Descripción
Order ID	Identificación de la Orden.
Order Date	Fecha de la Orden.
Product ID	Identificación del product adquirido.
Product Category	Categoria del producto adquirido.
Buyer Gender	Genero del comprador.
Buyer Age	Edad del comprador.
Order Location	Ubicación de donde se realizó la orden.
International Shipping	Si la orden requirió envío internacional.
Sales Price	Precio unico del producto.
Shipping Charges	Si hubo cargos de envío.
Sales Per Unit	Precio único si son 2 o más productos.
Quantity	Cantidad de productos por orden.
Total Sales	Pago total realizado.
Rating	Puntuación del comprador.

Durante la carga de datos realizamos algunas modificaciones, como cambiar los espacios por guiones bajos en los nombres de las variables, transformamos la columna “Order Date” a DateTime para facilitar el análisis de tiempos y también eliminamos una columna llamada “Review” eran comentarios de los clientes que no nos parecían significantes para el análisis.

Durante el análisis inicial realizamos un resumen general que incluye la cardinalidad, los tipos de datos y la clasificación de las categorías presentes en el conjunto de datos. En este análisis identificamos que contamos con dos variables categóricas nominales, que representan datos sin un orden específico, y dos variables binarias, las cuales solo pueden tomar dos valores posibles. Además, el conjunto de datos incluye siete variables numéricas discretas, una variable de tipo fecha que nos permite trabajar con información temporal, y varios índices que funcionan como identificadores únicos. Es importante destacar que no se encontraron valores faltantes, lo cual simplifica el análisis al no requerir procesos de imputación o eliminación de registros. En general, la mayoría de las variables presentan una baja cardinalidad, lo que significa que tienen pocos valores únicos, lo cual nos resulta beneficioso al momento de realizar interpretaciones y visuales.

Seguidamente realizamos un resumen de estadísticas de cada variable numéricas.

- **Order_ID**: Esta columna, aunque numérica, la excluimos del análisis estadístico al ser un identificador.
- **Buyer_Age**: La media fue de 26.45 años, con una distribución ligeramente sesgada hacia edades menores.
- **Sales_Price**: Media de 55.17$ con un rango amplio que refleja variabilidad significativa en los productos vendidos.
- **Shipping_Charges**: Valores predominantemente bajos con algunos extremos altos, generando una distribución sesgada.
- **Total_Sales**: Valor máximo de 1000 unidades monetarias, pero con una mediana de 90, lo que indica una concentración de pedidos en valores menores.
- **Rating**: Valoraciones concentradas en el rango 3-5, con una media de 3.5.

La curtosis y asimetría de las variables proporcionaron información adicional sobre la forma de las distribuciones. Por ejemplo, Total_Sales presentó una alta curtosis y asimetría positiva, indicando la presencia de outliers.

Luego de haber realizado el primer “tanteo” con los datos, guardamos el Dataframe actualizado con las modificaciones realizadas hasta el momento y entramos a realizar Test estadísticos como: Chi2 (entre la categoría de productos y genero), ANOVA (Entre la categoría de los productos y la venta total) y test de Grubbs (en las ventas totales para evaluar outliers).

Al continuar con el análisis, identificamos una cantidad significativa de valores atípicos en las ventas totales. Al examinar estos casos, comprobamos que se deben, principalmente, a órdenes con una mayor cantidad de productos y/o pedidos que incluyen productos más costosos. Tras analizar el contexto general del negocio y los datos, consideramos que estos valores atípicos son coherentes con la naturaleza de las transacciones y reflejan situaciones reales. Por este motivo, recomendamos no imputar ni eliminar estos valores, ya que su presencia es válida y representa a los patrones esperados en las operaciones comerciales de la tienda.

Visualizaciones

Se implementaron diversas visualizaciones para complementar el análisis:

- **Gráficos univariantes:** Histogramas y boxplots para identificar la distribución y outliers.
- **Gráficos bivariantes:** Mapas de calor y gráficos de barras apiladas para explorar relaciones entre variables categóricas.
- **Gráficos multivariantes:** Análisis de combinaciones de variables numéricas y categóricas para identificar tendencias más complejas.

Conclusiones

1. Las variables Sales_Price, Shipping_Charges y Total_Sales mostraron alta variabilidad, lo que podría estar relacionado con la diversidad de productos y destinos.
2. No se identificó una asociación significativa entre el género del comprador y la categoría del producto, aunque los patrones visuales sugieren preferencias aparentes.
3. Los valores atípicos en Total_Sales podrían indicar pedidos inusuales o errores en los datos.

En conclusión, el EDA permitió una comprensión profunda de los datos y estableció una base sólida para futuros análisis predictivos o de optimización de ventas y por suerte no tuvimos muchos inconvenientes a la hora de realizar el análisis, los datos estaban muy limpios.