Inteligencia de Negocios

Documentación de Clase Práctica 2

Marcos Ulises Sánchez

Escogimos el set de datos: Bike Sales in Europe (kaggle.com)

En este set de datos se encuentra información sobre las ventas de bicicletas, tenemos ingresos por transacción, ganancias, cantidad de productos vendidos, y otras variables como país, estado, categoría de productos o subcategoría. En los registros encontramos transacciones de varios años, con clientes de distintas edades y ubicaciones geográficas. Podríamos usar este dataset para **Tendencias de ventas**, **Segmentación de clientes**, o **Predicción de ventas**.

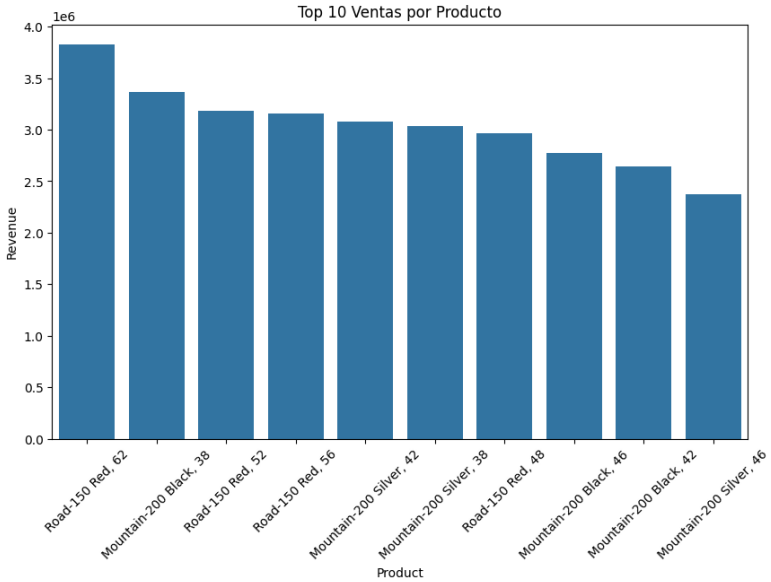
Pasos que seguimos:

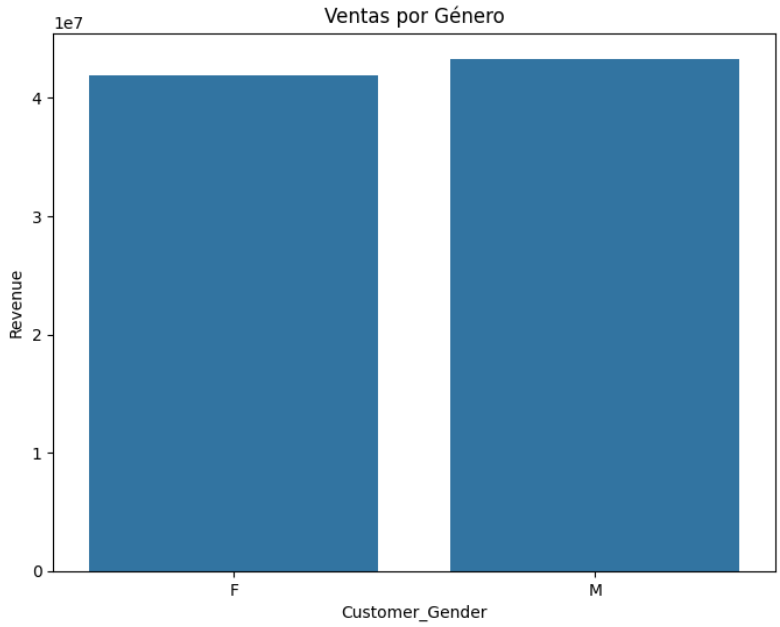
1. Cargar los datos

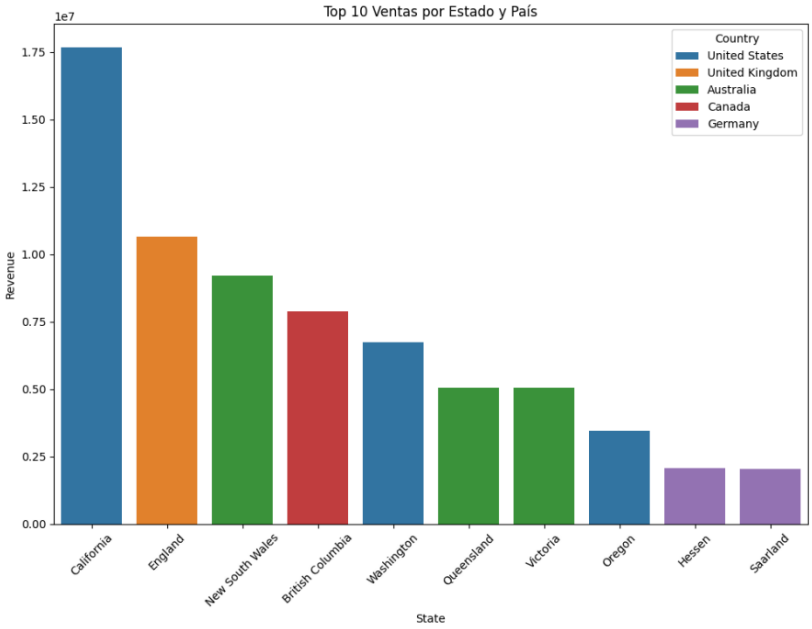
El primer paso fue cargar los datos utilizando **pandas**. Usamos una URL de un repositorio de GitHub. Imprimimos el head del dataframe para asegurarnos que los datos fueron cargados correctamente.

2. Exploración inicial

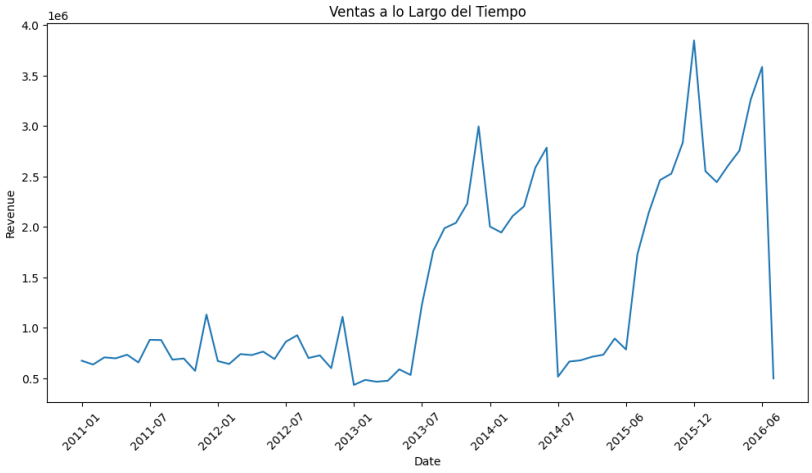
La exploración inicial permite conocer mejor los datos, identificando el número de registros, los tipos de variables y las estadísticas clave de las variables numéricas. Agrupamos las variables por género, país-estado, por producto para conocer mejor la naturaleza y distribución de nuestros datos.



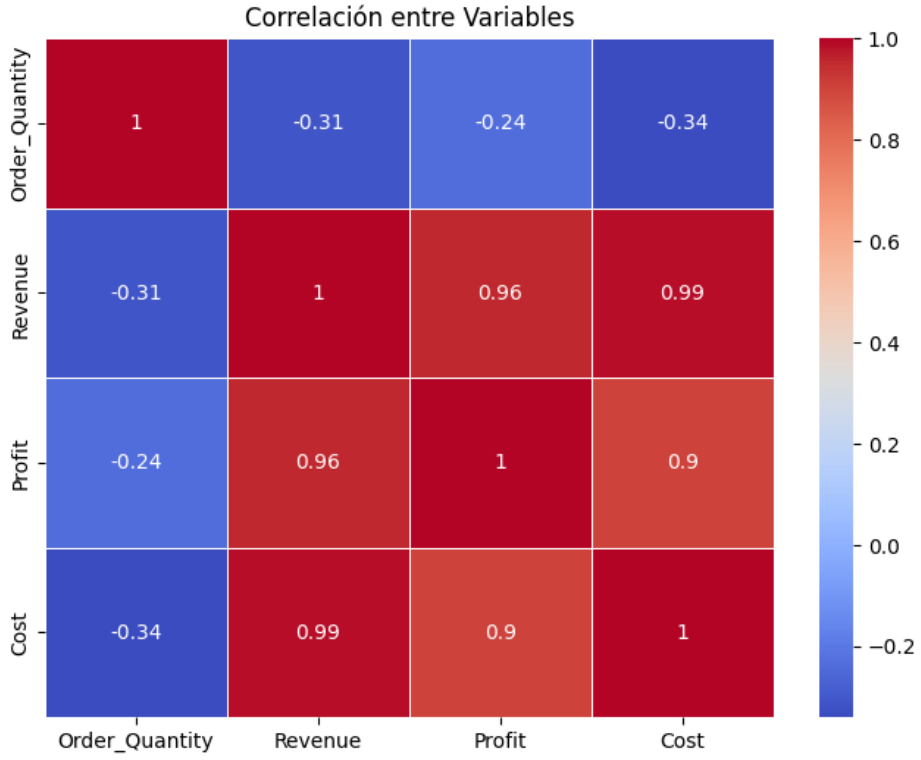




También podemos analizar el comportamiento de las ventas a lo largo del tiempo:

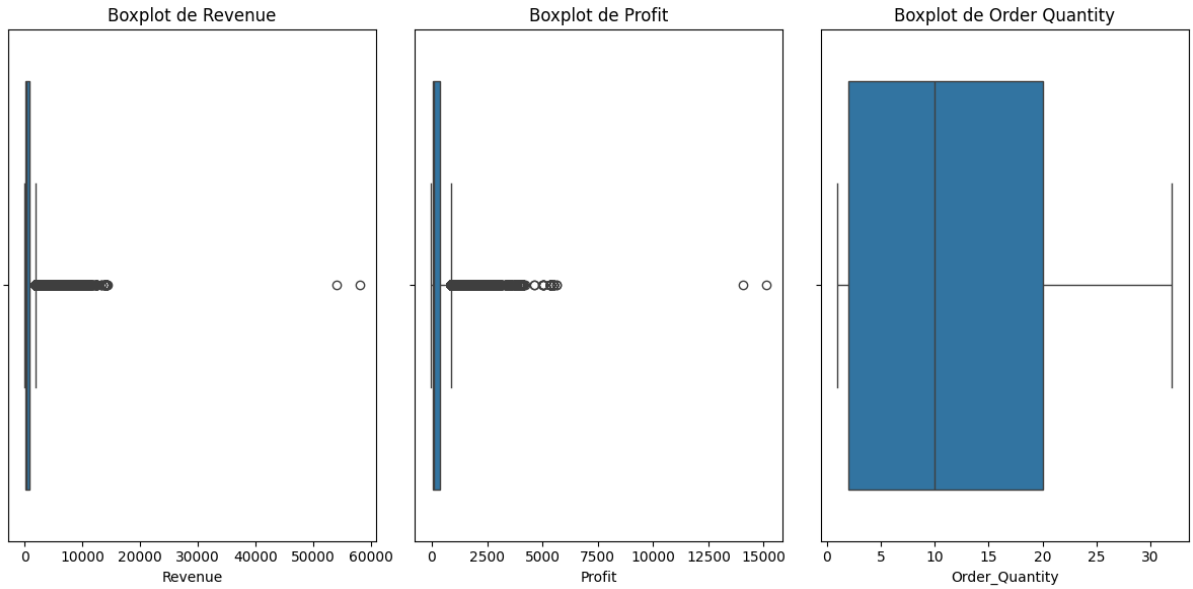


Con las variables cuantitativas más importantes, podemos hacer una matriz de correlación para determinar qué variables podríamos cruzar en nuestros análisis. Usamos costo, ganancia, ingreso y cantidad ordenada.



3. Limpieza de datos

No había datos faltantes, entonces no hubo necesidad de imputación de estos. Sin embargo, el siguiente paso de la limpieza es el tratamiento de outliers. Los outliers pueden tener un impacto significativo en el análisis estadístico y en el rendimiento de los modelos de machine



learning. Utilizando el Z-score con un umbral de 3 desviaciones estándar, se detectaron y eliminaron outliers en las variables clave como Revenue, Profit, y Order_Quantity.

4. Transformación de variables

Estandarización: Asegurando que tuvieran una media de 0 y una desviación estándar de 1. Esto nos permitió trabajar con datos en una escala uniforme, lo que es particularmente útil cuando se comparan variables que tienen magnitudes muy diferentes.

Normalización: En paralelo, se aplicó la normalización, que escala las variables en un rango entre 0 y 1. Esta transformación es útil cuando se trabaja con modelos sensibles a la magnitud de las características. También, mitiga el impacto de magnitudes demasiado grandes, haciendo menos sesgados nuestros modelos.

5. Ingeniería de características

La ingeniería de características es una etapa crucial en el análisis de datos, donde se crean nuevas variables a partir de las existentes para obtener información adicional. Las variables creadas fueron:

Profit_margin: Mide el margen de ganancia dividiendo Profit entre Revenue. Esta métrica es vital para evaluar qué tan rentable es cada venta, ya que el Revenue por sí solo no siempre refleja el rendimiento financiero si los costos son altos.

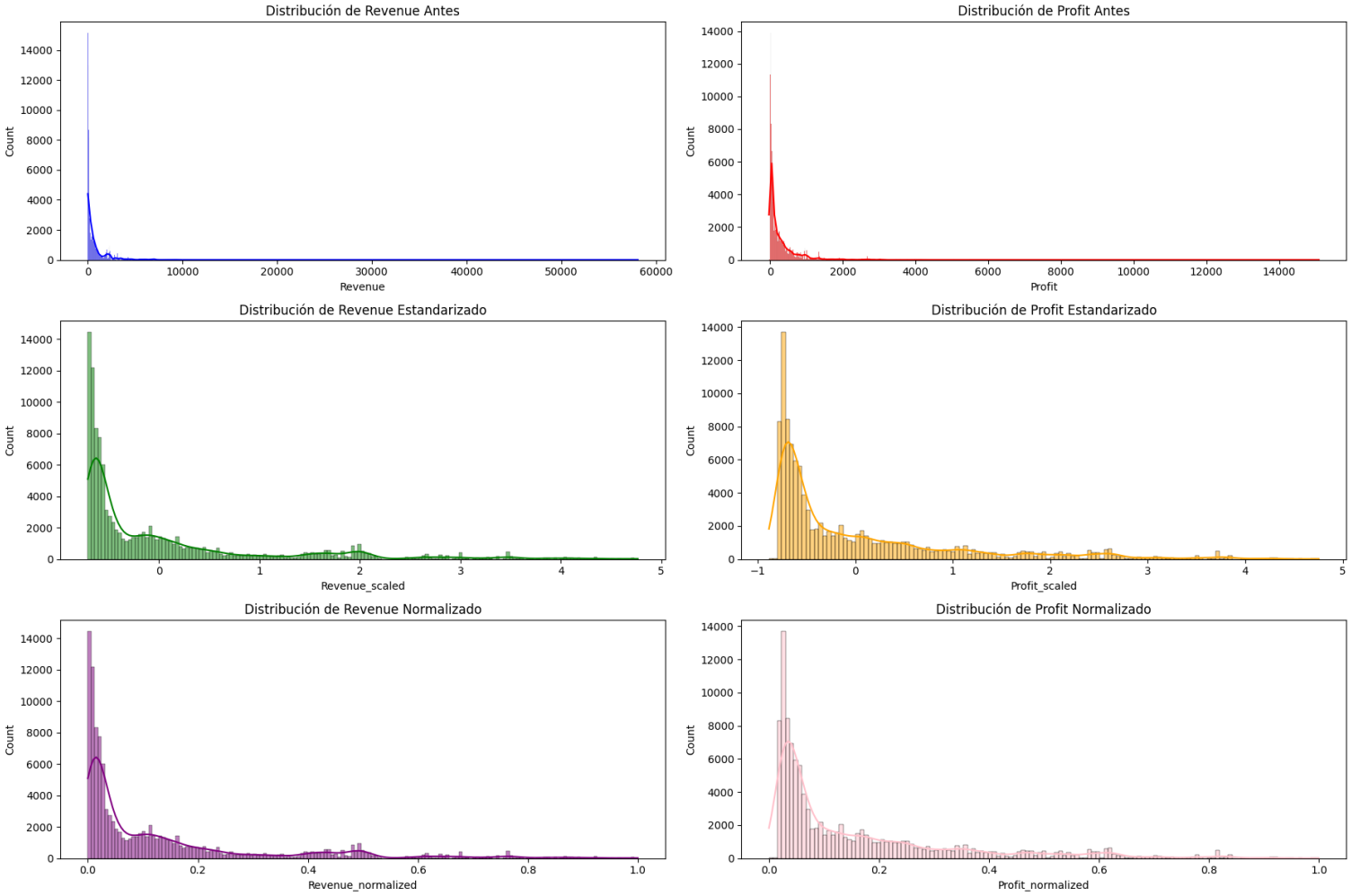
Revenue_Diff: Permitirá detectar cualquier discrepancia entre el ingreso real y el ingreso esperado en función de la cantidad y el precio unitario.

Codificación de Variables Categóricas:

Las variables categóricas como Product_Category, Sub_Category, y Country no pueden ser procesadas directamente en los modelos de machine learning porque estos requieren variables numéricas. Para resolver este problema, se aplicó One-Hot Encoding, que convierte las categorías en una serie de variables binarias. Este proceso convierte cada valor único de una categoría en una columna separada, asignando un valor de 1 o 0 dependiendo de si esa categoría está presente.

Una parte fundamental del preprocesamiento es verificar cómo cambian las distribuciones de las variables antes y después del tratamiento de outliers y las transformaciones. En este caso, se generaron gráficos de **Revenue** y **Profit** en tres etapas:

- **Antes del preprocesamiento**: Las distribuciones originales mostraban sesgos y valores extremos.
- **Estandarizado**: Las variables fueron transformadas para tener una media de 0 y una desviación estándar de 1, lo que facilitó su comparación y análisis.
- **Normalizado**: Se escaló cada variable en un rango de 0 a 1, lo que permitió compararlas independientemente de sus magnitudes originales.



Conclusiones

La eliminación de outliers fue clave para mejorar la robustez del análisis, asegurando que los valores extremos no distorsionaran los resultados.

La aplicación de transformaciones (estandarización y normalización) preparó los datos para ser utilizados en futuros modelos, garantizando que las características estuvieran alineadas correctamente en términos de escala.

La creación de nuevas características ayudó a obtener insights adicionales sobre la eficiencia de las ventas y la rentabilidad, proporcionando una visión más completa del rendimiento del negocio.

La codificación de variables categóricas permitió que las variables no numéricas fueran procesadas de manera adecuada por los modelos de machine learning, eliminando cualquier barrera para el análisis predictivo. Como mencionamos antes, usamos one-hot encoding porque las variables categóricas no son ordinales. Label-encoding sería apropiado si tuviésemos ratings de calidad de servicio como bueno, normal o malo.