TALLER SEGUNDO CORTE

Integrantes:

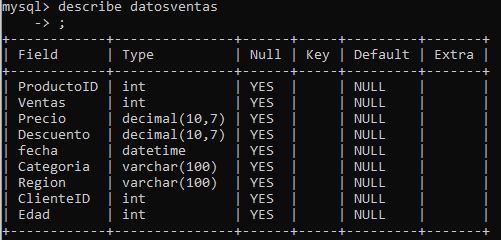
* Juan Esteban Vega Villamil. 1803894.
* Rafael Triana.

1. **Manejo de Datos con SQL:** - Crear una base de datos en MySQL que almacene los datos de ventas, productos, y clientes (tome como referencia el archivo DataSet.xlsx).

Tomando como referencia el archivo DataSet.xlsx, tenemos que realizado una limpieza con la base de datos para poder importarla a MySQL y a Python y poder realizar el taller, para la limpieza tuvimos en cuenta en los datos que no tenían sentido en los campos de la base de datos, y haciendo el esfuerzo de no eliminar registros para no perder información. Se realizo lo siguiente para la limpieza

En el campo de Ventas, se eliminaron caracteres especiales para asegurar que los datos fueran numéricos. Además, las ventas negativas fueron reemplazadas por 0, ya que no son válidas en la práctica, en el campo de Edad, Los valores de -1 fueron reemplazados por el promedio de las edades, calculado excluyendo dichos valores inválidos, en precio Los valores de -999 fueron reemplazados por 0, ya que representaban datos faltantes en esta columna, tanto en categoría y en Región Los valores faltantes fueron reemplazados por "NAN" para no perder registros importantes y poder recopilar datos faltantes en futuras etapas y por ultimo se eliminación de filas irrelevantes: Se eliminaron las filas sin valores disponibles en las columnas de Precio y Ventas, considerándolas irrelevantes para el análisis de datos.

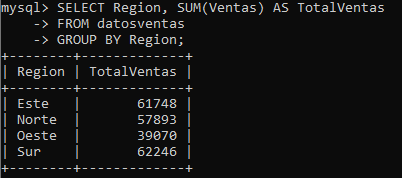
Este proceso de limpieza fue esencial para asegurar la calidad de los datos y obtener resultados confiables en el análisis posterior. La tabla resultado que se realiza para la importación de la base datos a MySQL quedo de la siguiente manera:



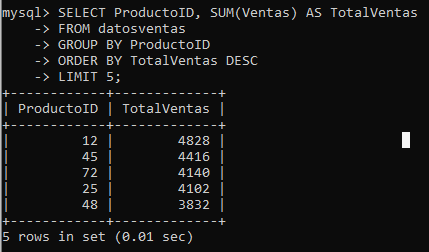
- Realizar consultas SQL para obtener resúmenes y filtrar la información según criterios específicos (e.g., ventas por región, productos más vendidos).

Después de poder insertar la base de datos a MySQL realizamos las siguientes pruebas, explorando la base de datos:

1. Consultas para saber el total de ventas de cada región:



1. Consulta para saber los mejores 5 productos en el total de ventas.



1. **Exploración de Datos (EDA) con Pandas:** - Importar los datos desde la base de datos a un DataFrame de Pandas.

- Realizar un análisis exploratorio para identificar tendencias, patrones y detectar valores atípicos o datos faltantes.

- Identificar las variables categóricas y numéricas dentro del conjunto de datos.

- Utilizar estadísticas descriptivas y visualizaciones (e.g., histogramas, boxplots) para resumir los datos.

1. **Visualización de Datos:** - Crear gráficos con Matplotlib y Seaborn para visualizar la distribución de las ventas, correlaciones entre variables, y tendencias temporales. - Generar gráficos interactivos utilizando Plotly para presentar los resultados de manera dinámica.
2. **Regresión y Predicción:** - Aplicar regresión lineal y logística para predecir las ventas futuras en función de variables independientes como precios, promociones, y características de los clientes. - Evaluar la calidad del modelo utilizando métricas como el R-cuadrado y el RMSE. - Presentar un análisis de residuos para verificar la adecuación del modelo.