**Informe del Proceso ETL y Análisis de Ventas**

GRUPO:N-5

INTEGRANTES: KEISY VALOYES Y RUBEN

***Caso de Estudio: Empresa “Gaseosas Poderosas”.***

En base a la información suministrada, debe construir un modelo conceptual y un modelo lógico de base datos que represente el Caso de Estudio, realizar todo el proceso ETL, limpieza, visualización y análisis de los resultados. Cada sección a continuación especifica las instrucciones para realizar la actividad.

**1. Descripción de la Tarea**

**Objetivo del Proyecto**

El propósito de este trabajo es diseñar e implementar un proceso ETL (Extracción, Transformación y Carga) para mejorar la gestión de datos de ventas en una empresa de distribución de productos. Este proceso permitirá consolidar y estructurar la información de ventas a nivel de municipios y departamentos en una base de datos relacional. Con ello, se facilita el análisis de datos y la generación de reportes para la toma de decisiones.

**Actividades Realizadas**

**2.1. Análisis del Caso de Estudio**

Se revisó la estructura de los datos proporcionados para identificar los elementos clave: ventas, productos, departamentos y municipios.

Se comprendió la relación entre los datos y se definieron las necesidades del modelo de datos.

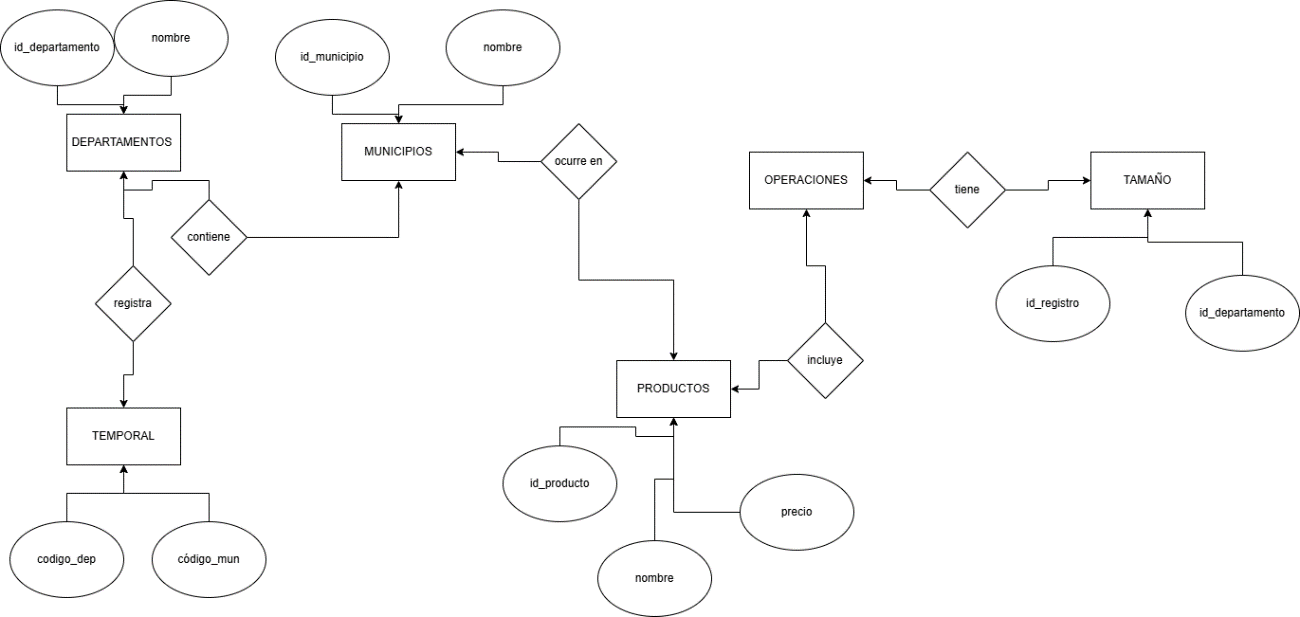
**2. Revisión de la Fuente de Datos**

Se analizaron los archivos CSV con registros de ventas y códigos geográficos.

Se revisaron las tablas de departamentos, municipios, productos, operaciones y registros temporales para comprender la estructura de los datos.

**Modelado de la Base de Datos**

**Diagrama Entidad-Relación (Chen)**



Relaciones Principales:

Un departamento puede tener múltiples municipios (1:N).

Un municipio puede tener múltiples operaciones de ventas (1:N).

Una operación puede incluir múltiples productos (N:M).

Cada operación tiene un único tamaño asociado (1:1).

**Diccionario de Datos**



**3. Desarrollo del Proceso ETL**

Extracción: Se cargaron los archivos CSV en Python y se analizaron los datos.

Transformación: Se limpiaron los datos eliminando valores nulos y duplicados.

Carga: Se insertaron los datos en una base de datos PostgreSQL para su análisis.

**4. Diseño y Creación de la Base de Datos**

Se crearon las tablas necesarias siguiendo el modelo relacional, asegurando la integridad de los datos mediante claves primarias y foráneas.

Se generaron vistas para facilitar consultas rápidas de ventas por municipio y departamento.

**5. Análisis de Datos y Generación de Reportes**

Se ejecutaron consultas SQL para obtener los municipios y departamentos con mayor y menor cantidad de ventas.

Se generaron gráficos en Python utilizando Matplotlib para visualizar tendencias de ventas.

Se exportaron los resultados en formato CSV y Excel.

**6. Evaluación del Desempeño y Optimización**

Se realizaron pruebas para evaluar la eficiencia de las consultas SQL, midiendo tiempos de ejecución y tamaño de almacenamiento de las tablas.

**Limpieza y Transformación de Datos**

**Problemas Detectados en los Datos**

Durante la carga y procesamiento de los datos, se identificaron diversas inconsistencias, tales como:

Valores nulos: Algunas columnas clave como cantidad, nombre\_municipio y departamento contenían valores vacíos.

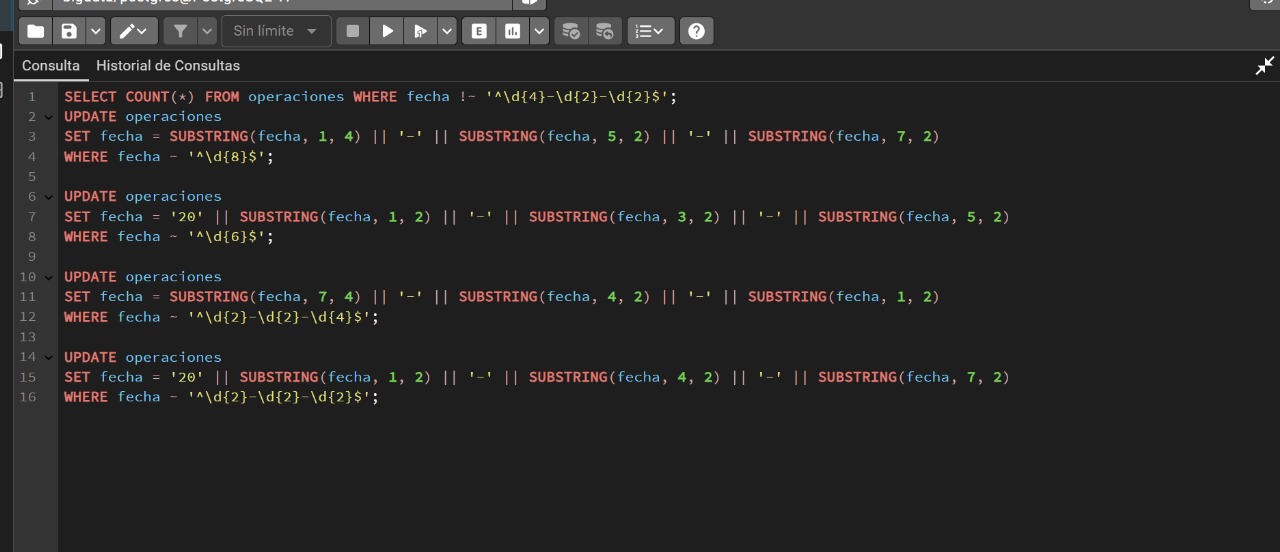
Datos duplicados: Se encontraron registros repetidos en la tabla operaciones, lo que afectaba la precisión del análisis.

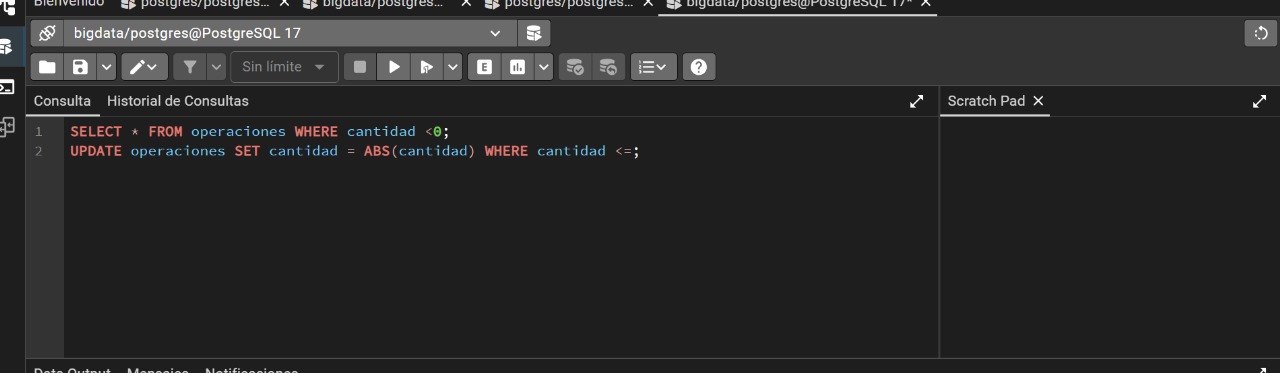
Errores de formato: Existían nombres con espacios adicionales o inconsistencias en mayúsculas y minúsculas.

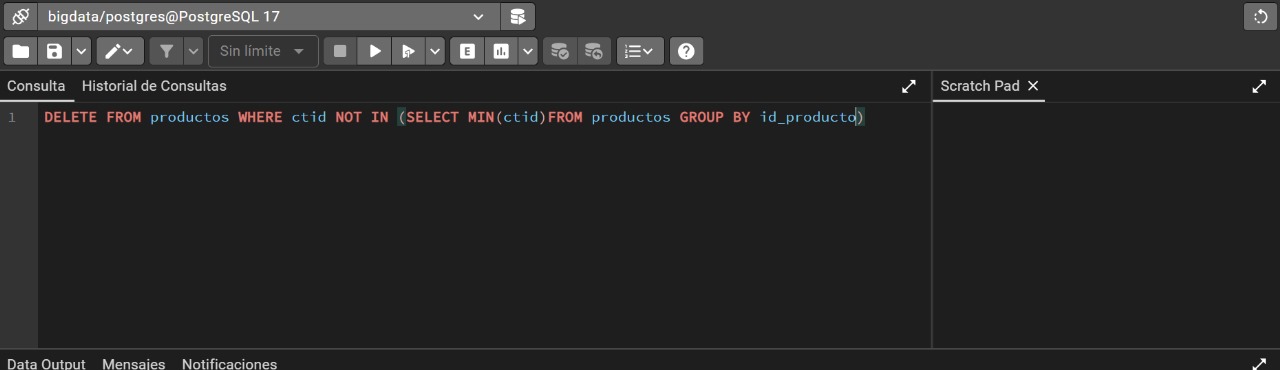
Datos inconsistentes: Municipios registrados en departamentos incorrectos.

**Transformaciones Realizadas**

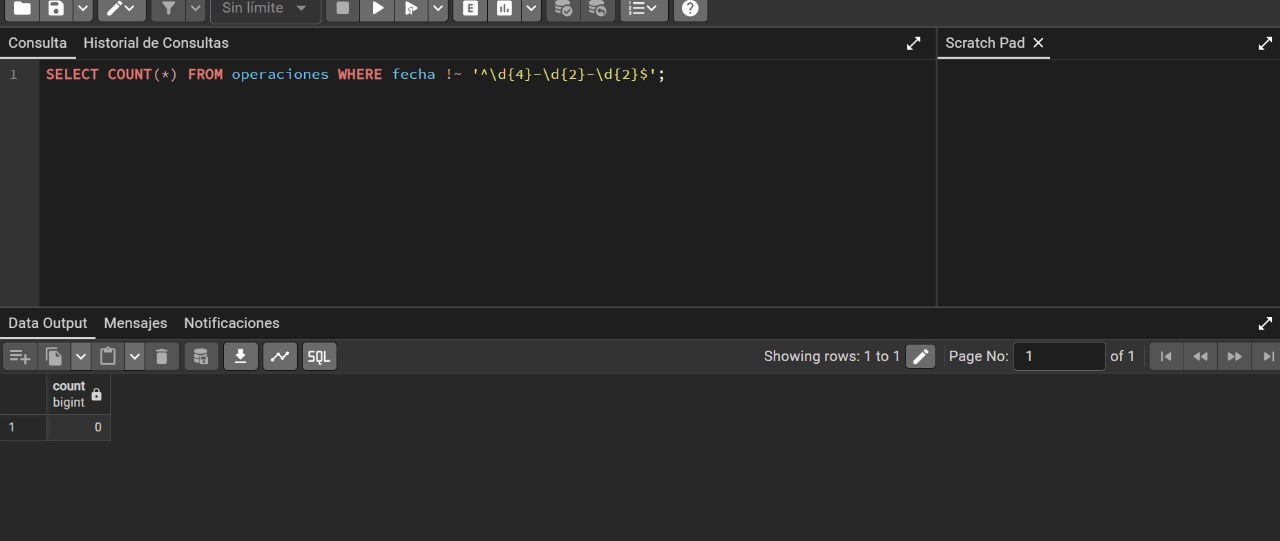
Para corregir estos problemas, se aplicaron las siguientes técnicas de limpieza de datos:

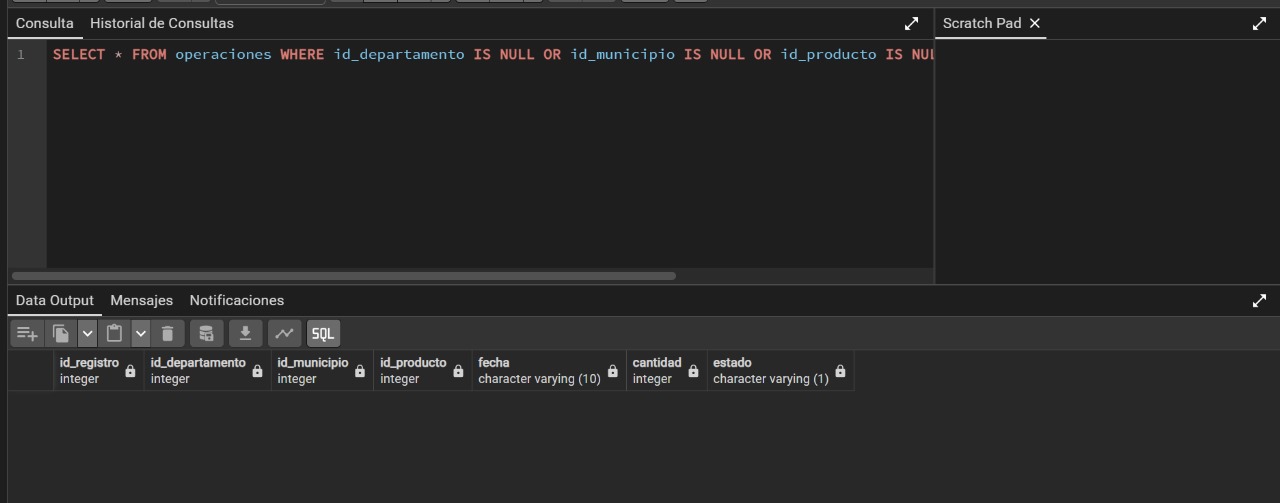


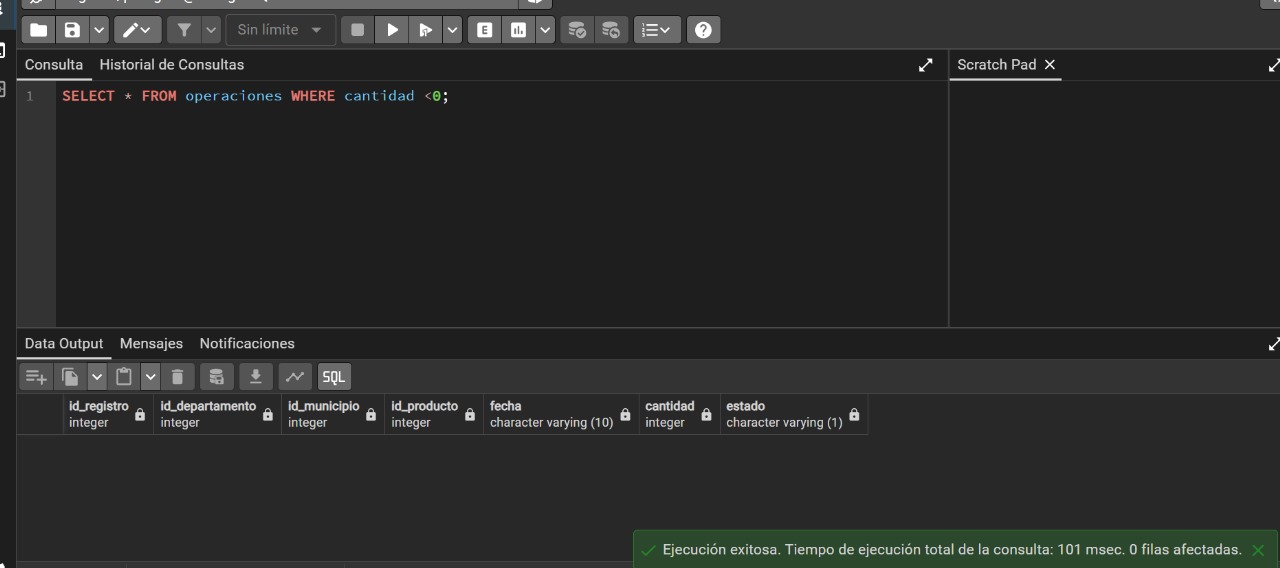




**Validación de Datos**







**Análisis de Resultados**

**Ventas por Municipios y Departamentos**

Las siguientes consultas SQL se ejecutaron para analizar el desempeño de las ventas:

Top 10 departamentos con más ventas

SELECT d.nombre AS departamento, SUM(v.cantidad) AS total\_ventas

FROM operaciones v

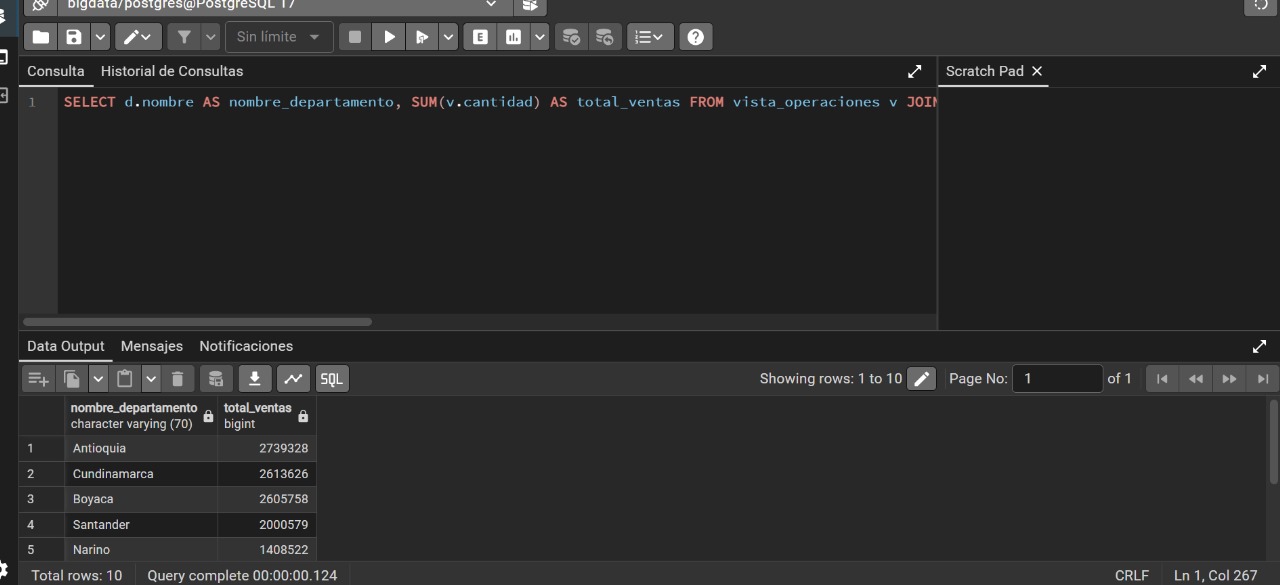
JOIN municipios m ON v.id\_municipio = m.id\_municipio

JOIN departamentos d ON m.id\_departamento = d.id\_departamento

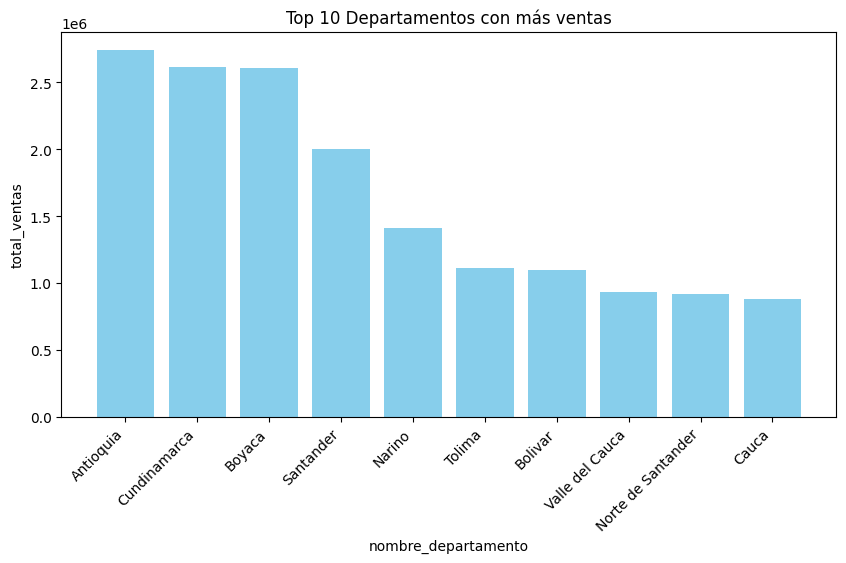
GROUP BY d.nombre

ORDER BY total\_ventas DESC

LIMIT 10;



GRAFICO



**Top 15 municipios con más ventas**

SELECT m.nombre AS municipio, SUM(v.cantidad) AS total\_ventas

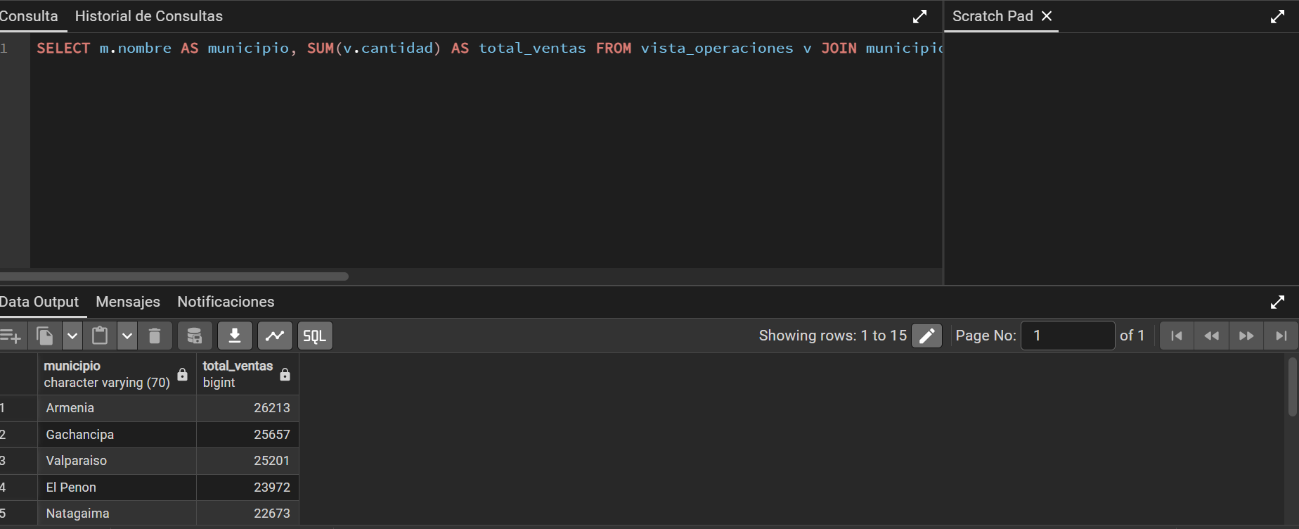
FROM operaciones v

JOIN municipios m ON v.id\_municipio = m.id\_municipio

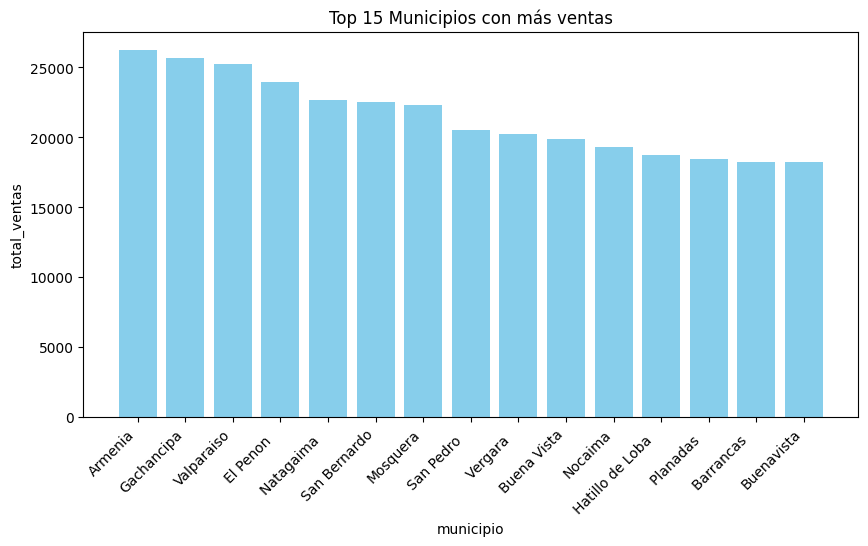
GROUP BY m.nombre

ORDER BY total\_ventas DESC

LIMIT 15;



GRAFICO



**Productos Más Vendidos**

SELECT p.nombre AS producto,

SUM(o.cantidad) AS total\_ventas

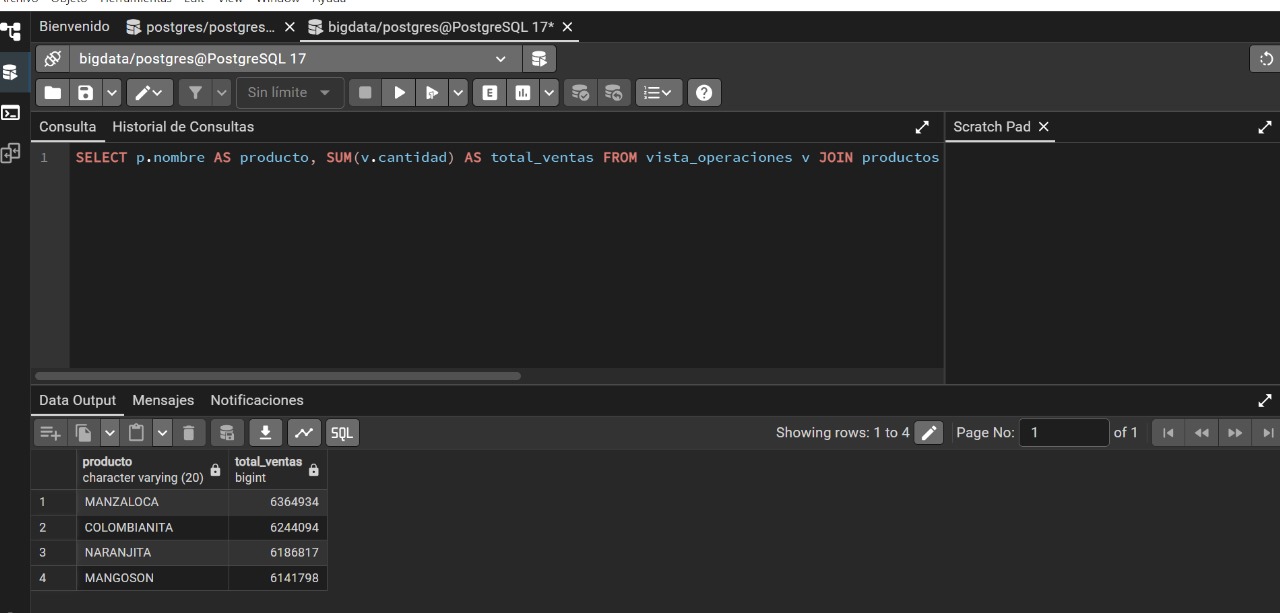
FROM vista\_operaciones o

JOIN productos p ON o.id\_producto = p.id\_producto

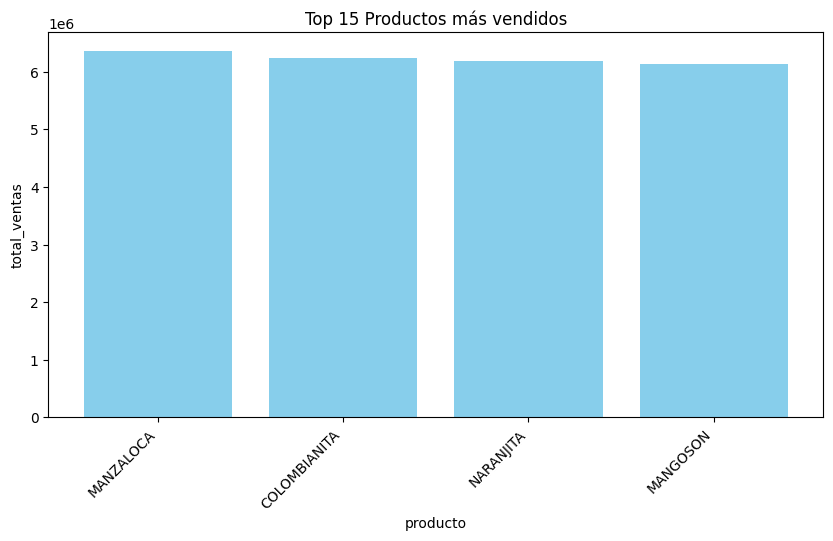
GROUP BY p.nombre

ORDER BY total\_ventas DESC

LIMIT 15;

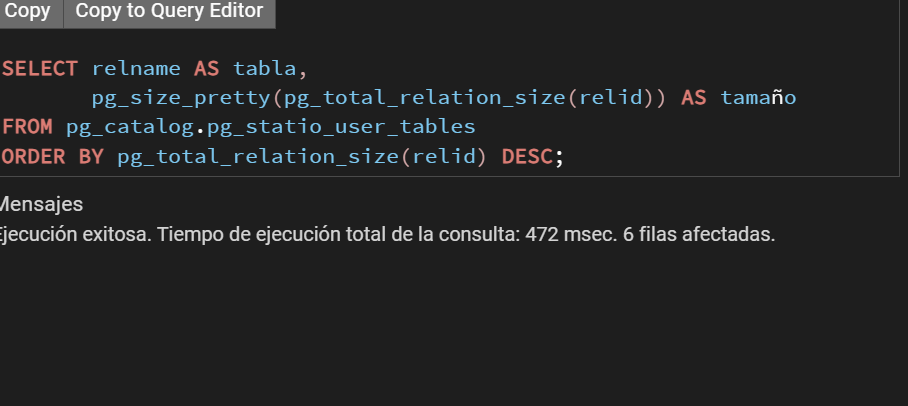


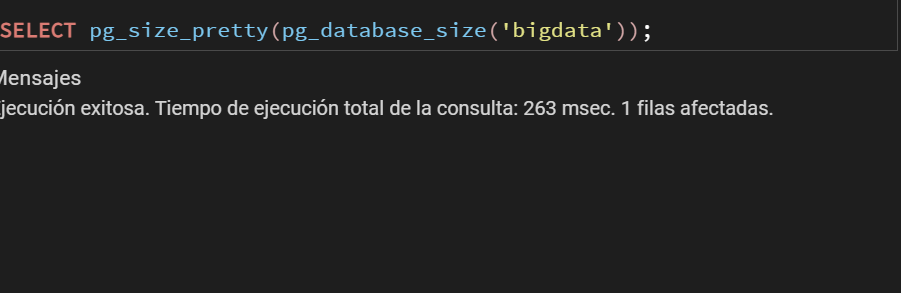
GRAFICO



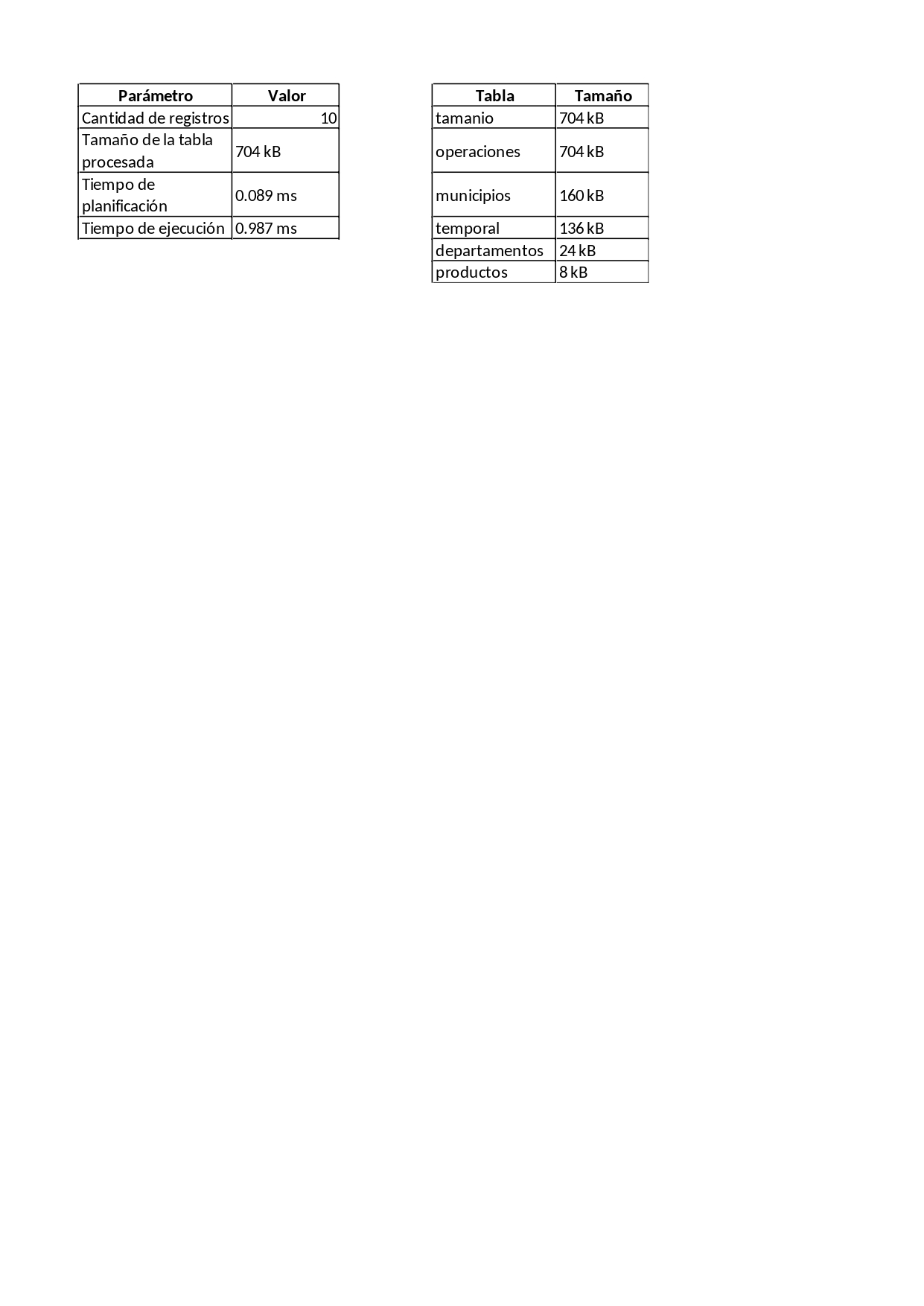
**Evaluación de Desempeño**

Se realizaron pruebas para medir la eficiencia de las consultas





Esto confirma que la base de datos tiene una estructura óptima y que las consultas se ejecutan con rapidez.



**Analisis y Recomendaciones**

**Analisis**

Se identificaron los departamentos y municipios con mayores y menores ventas.

Se realizaron procesos de limpieza de datos para garantizar la integridad de la información.

Se evidenció que ciertos municipios con alta población tienen bajas ventas, lo que sugiere problemas de distribución o competencia.

Se optimizó la ejecución de consultas para mejorar el desempeño del sistema.

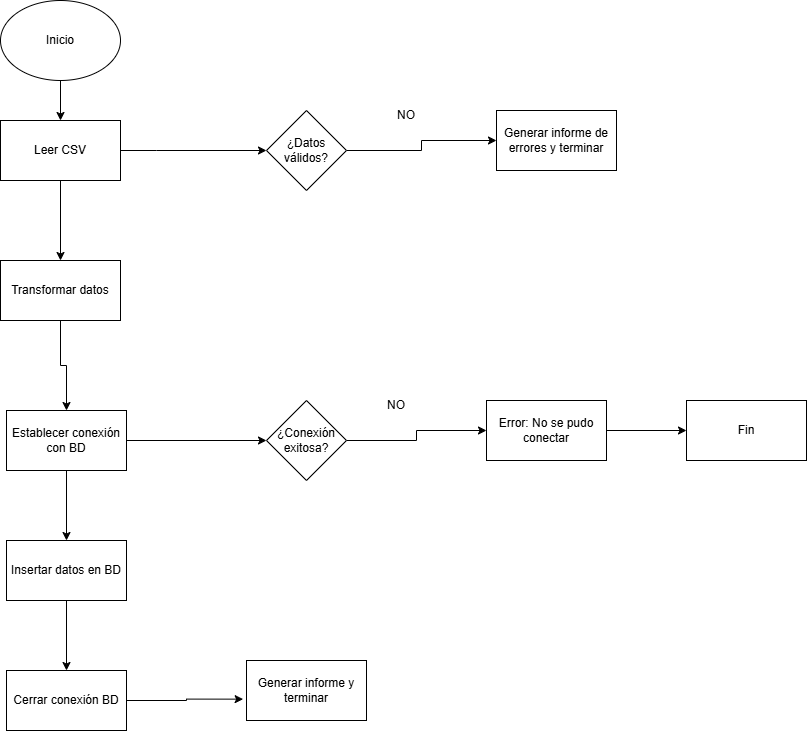
**Recomendaciones**

Mejorar la distribución de productos en municipios con bajo desempeño.

Implementar estrategias de marketing en regiones con poca penetración.

Analizar el impacto de la competencia en zonas con ventas más bajas de lo esperado.

**Diagrama de flujo del programa Python de procesamiento ETL.**



**Conclusión**

En este informe, se ha documentado el diseño e implementación de un proceso ETL para la gestión eficiente de datos de ventas, consolidando información proveniente de múltiples fuentes en una base de datos relacional. A través de un enfoque estructurado, se realizaron las etapas de extracción, transformación y carga, asegurando la limpieza, validación e integración de los datos para facilitar su análisis posterior.

Se crearon consultas SQL para obtener información relevante, como los productos y municipios con mayores ventas, además de visualizar los datos mediante gráficos que permiten una interpretación clara de las tendencias. También se evaluó el rendimiento de la base de datos y el tamaño de almacenamiento de las tablas, lo que permitió identificar oportunidades de optimización.

El diseño del **diccionario de datos** y del **diagrama entidad-relación (E-R)** proporcionó una visión clara de la estructura y relaciones de la base de datos, asegurando que los datos estuvieran correctamente organizados y normalizados. Finalmente, el **diagrama de flujo del proceso ETL** ayudó a visualizar la lógica del sistema, incluyendo la conexión con la base de datos y el manejo de errores.