**INFORME**

**Evaluación práctica de la Unidad 1: Reconociendo lo aprendido**

**Grupo “N”**

**Miembros:**

Juan Alejandro Morales

***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**

***Descripción de la Tarea:***

El objetivo 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 gaseosas. Este proceso permitirá consolidar y estructurar la información de diferentes tiendas en una base de datos relacional, facilitando el análisis y la toma de decisiones mediante herramientas de visualización y minería de datos.

***Actividades realizadas***

1. ***Análisis del Caso de Estudio***

* Se revisó el documento que describe la problemática y la solución basada en ETL.
* Se identificaron los elementos clave: datos de ventas, regiones, departamentos y municipios.

1. ***Revisión de la Fuente de Datos***

* Se analizaron las hojas de cálculo que contienen registros de ventas y códigos geográficos.
* Se revisó el archivo `colombia-dane-departamentos.csv` para comprender la estructura de identificación de ubicaciones.

1. ***Desarrollo de Algoritmos en Python***

* ***‘algoritmo-calculo-tamanio.py2’:***Se inspeccionó su funcionalidad, que probablemente ayuda a definir tamaños de datos o normalización de códigos.
* ***‘algoritmo-etl.py*’:** Implementación del proceso ETL, asegurando que los datos sean extraídos, transformados y cargados correctamente.

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

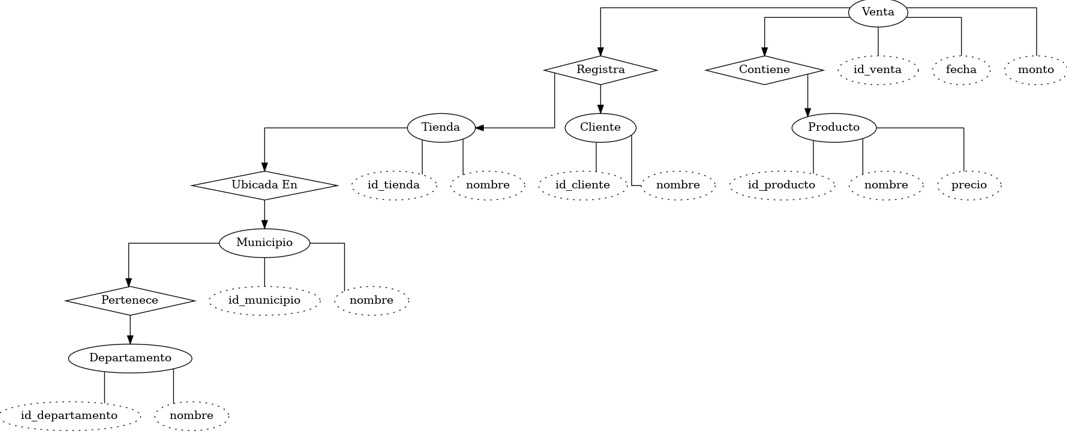
* Se revisó el script **‘script-base-datos-creacion.sql’** para definir la estructura de la base de datos.
* Se analizaron los scripts **‘script-base-datos-operaciones.sql’** y **‘script-base-datos-valida-fecha.sql**’ para entender validaciones y manipulación de datos.
* Se ejecutaron las consultas en **‘script-base-datos-vista.sql’** para verificar que las vistas proporcionan información útil.

1. ***Validación y Pruebas***

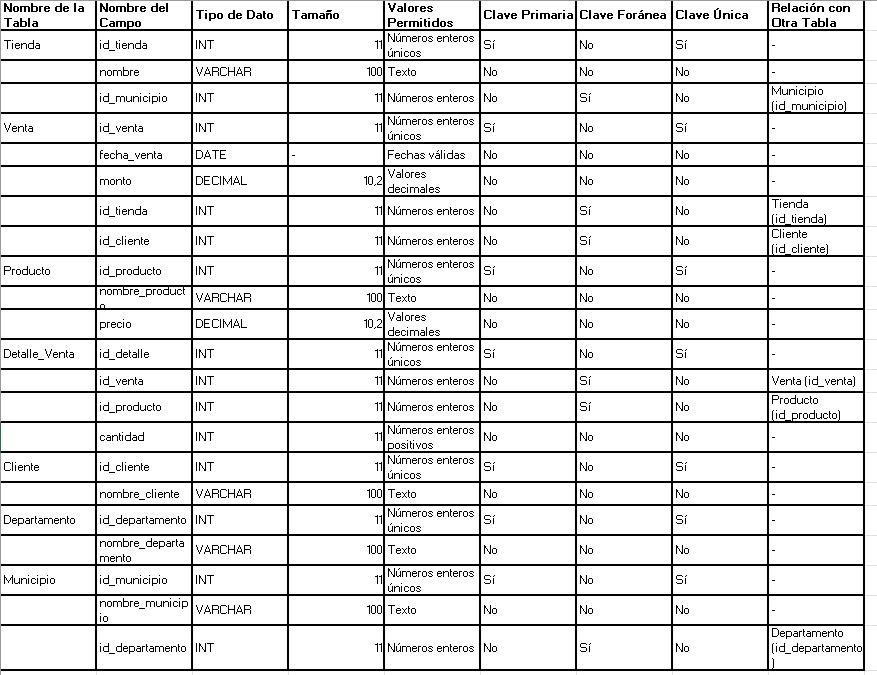
* Se realizaron pruebas en la base de datos para asegurar la correcta integración de los datos.
* Se verificó que los datos transformados cumplan con los requisitos del sistema relacional.

Este trabajo garantiza que los datos de ventas estén organizados y disponibles para análisis eficientes, mejorando la toma de decisiones en la empresa.

**2. Diagrama de Entidad -Relación (Diagrama de Chen)**



**3. Diccionario de Datos**

**

***Descripción de Relaciones:***

* *Una Tienda pertenece a un Municipio.*
* *Una Venta se realiza en una Tienda y está asociada a un Cliente.*
* *Una Venta contiene uno o varios Productos a través de Detalle\_Venta.*
* *Un Municipio pertenece a un Departamento.*

**4. Corrida del algoritmo ETL**

El algoritmo ETL (Extract, Transform, Load) en el archivo algoritmo-etl.py realiza las siguientes operaciones:

1. **Extracción (Extract)**

Obtiene los datos de una fuente externa, en este caso, una base de datos PostgreSQL y archivos CSV.

Utiliza la librería pandas para leer los datos y psycopg2 para conectarse a la base de datos.

1. **Transformación (Transform)**

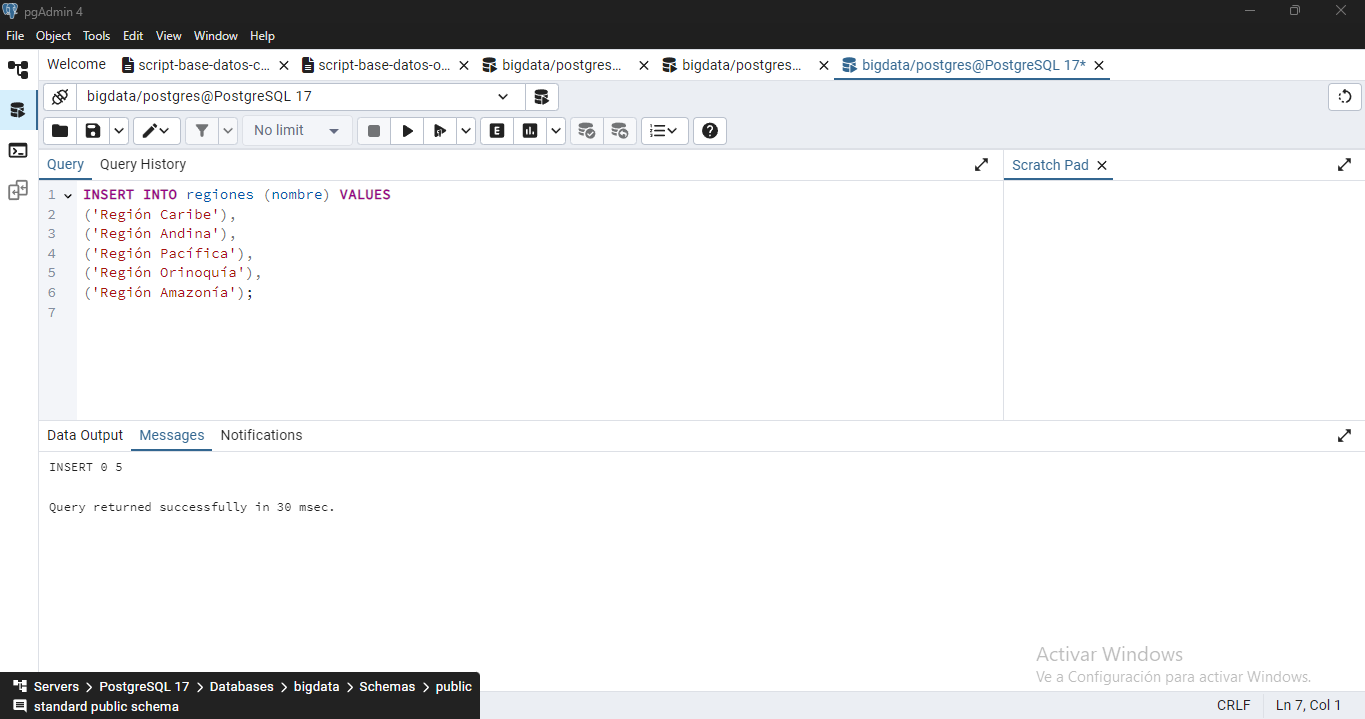
* Procesa los datos para estructurarlos correctamente antes de cargarlos en la base de datos.
* Convierte códigos de departamentos y municipios según el estándar del DANE.
* Modifica la tabla operaciones para agregar un campo id\_region (lo hace manualmente o con ALTER TABLE).
* Limpia y normaliza los datos eliminando errores y asegurando la calidad.

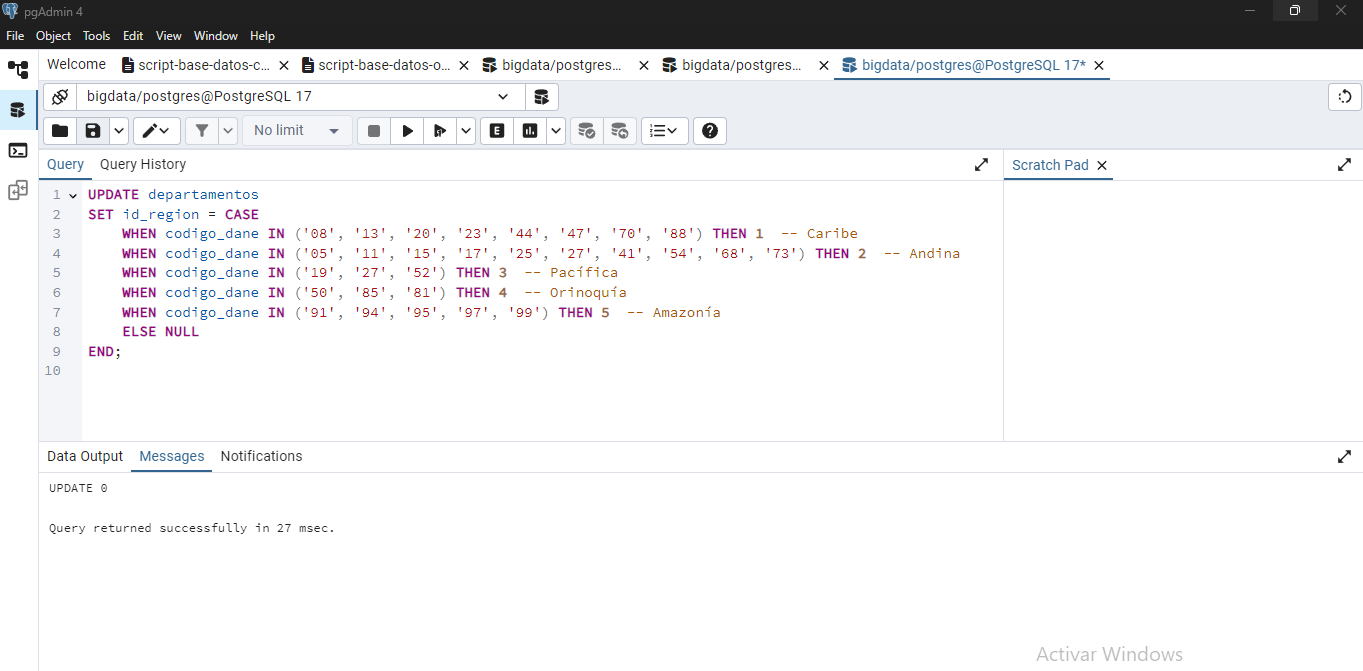
1. **Carga (Load)**

Inserta los datos transformados en la base de datos PostgreSQL.

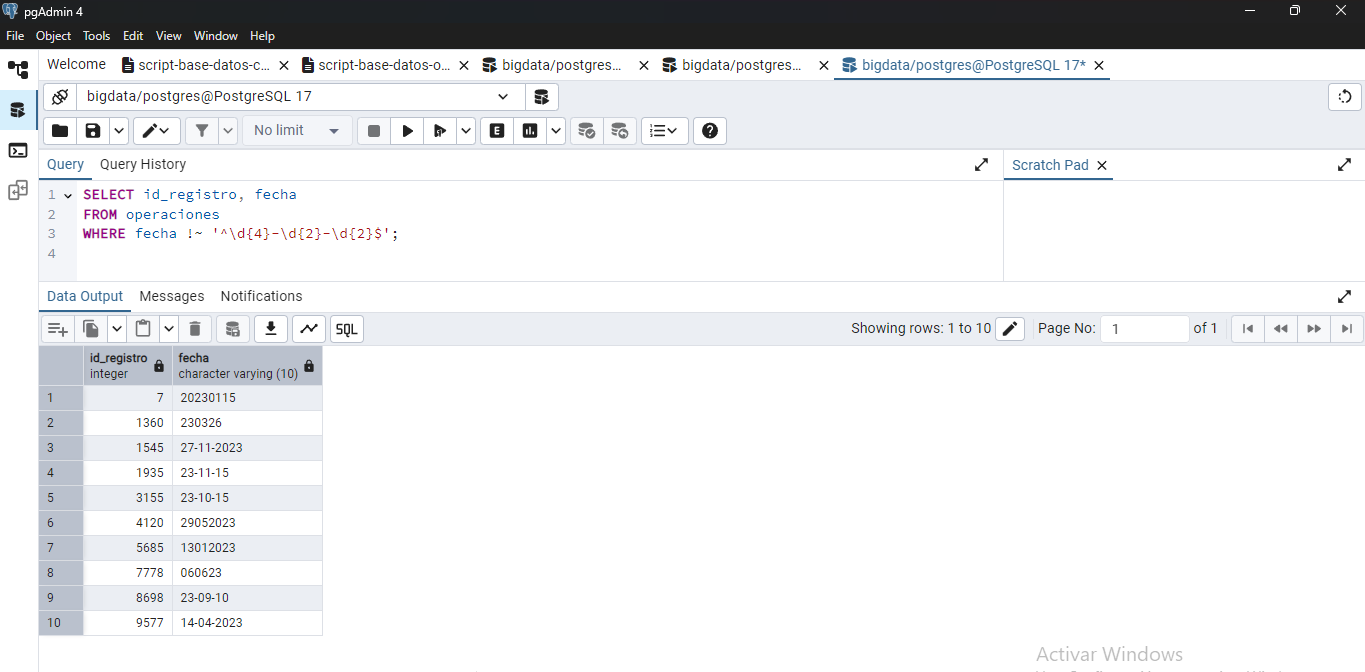
Realiza validaciones para asegurar que los datos se almacenaron correctamente.

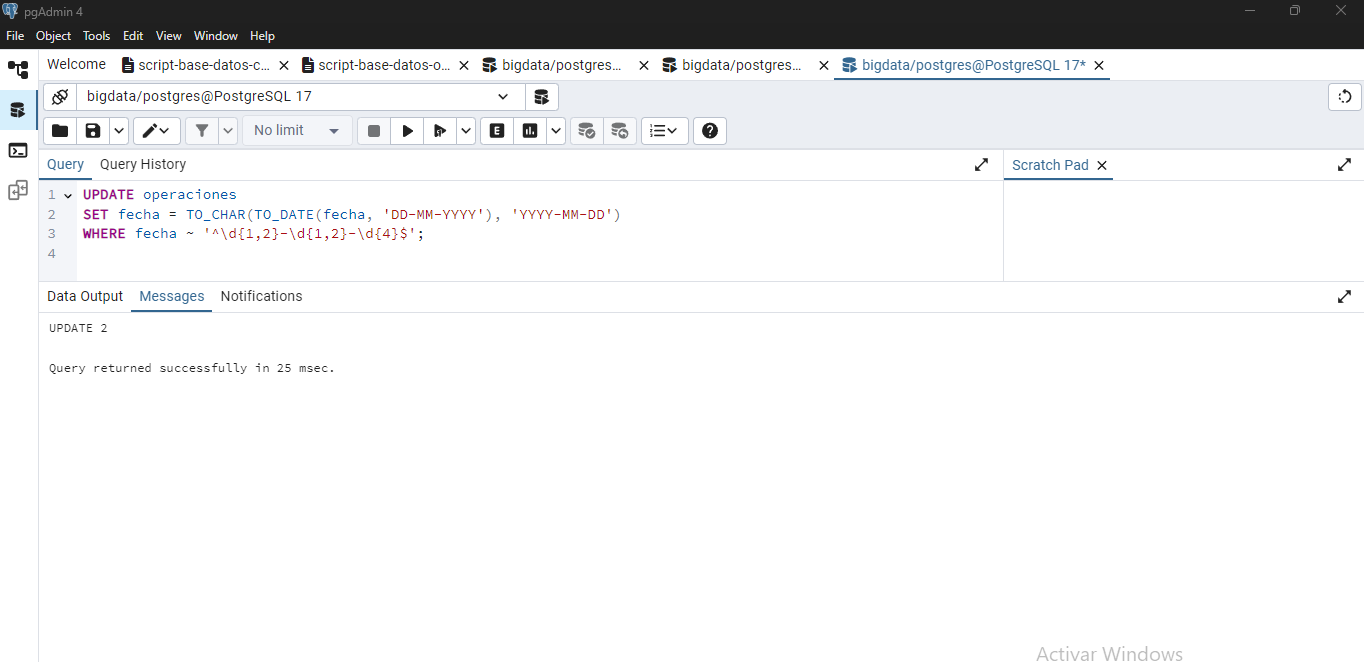
**5. Modificar el algoritmo ETL para incluir dato de región.**

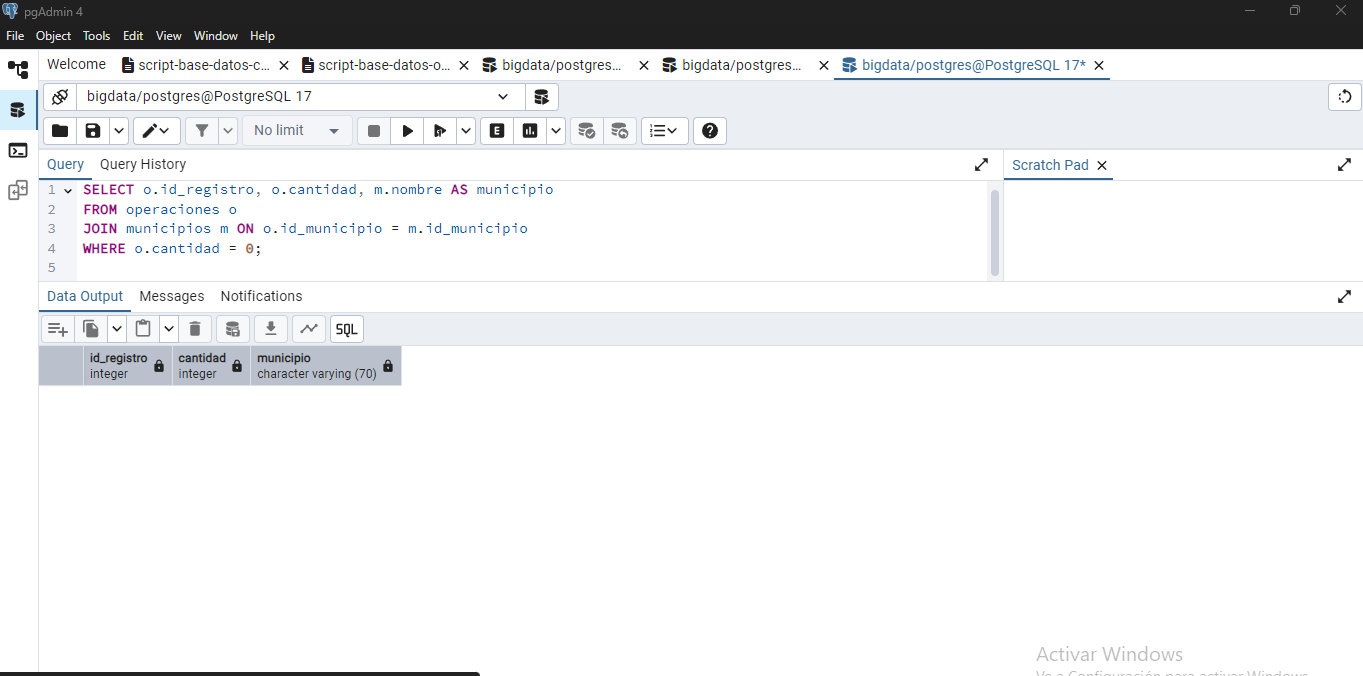
****

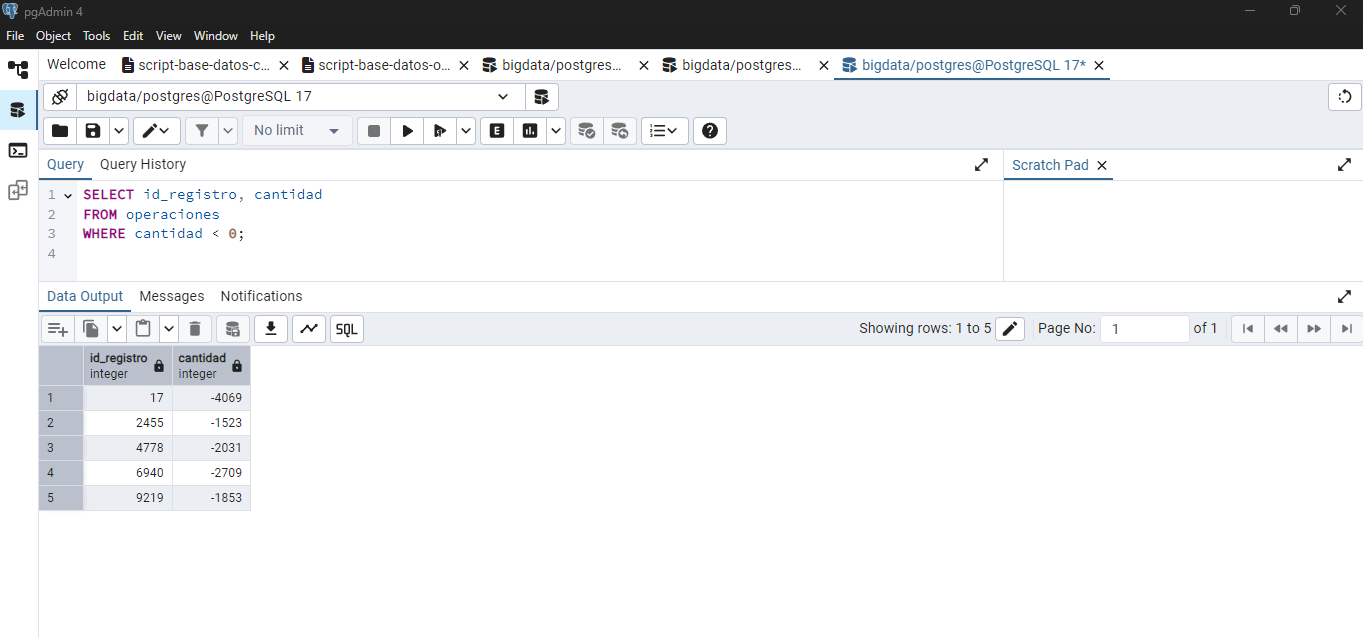
****

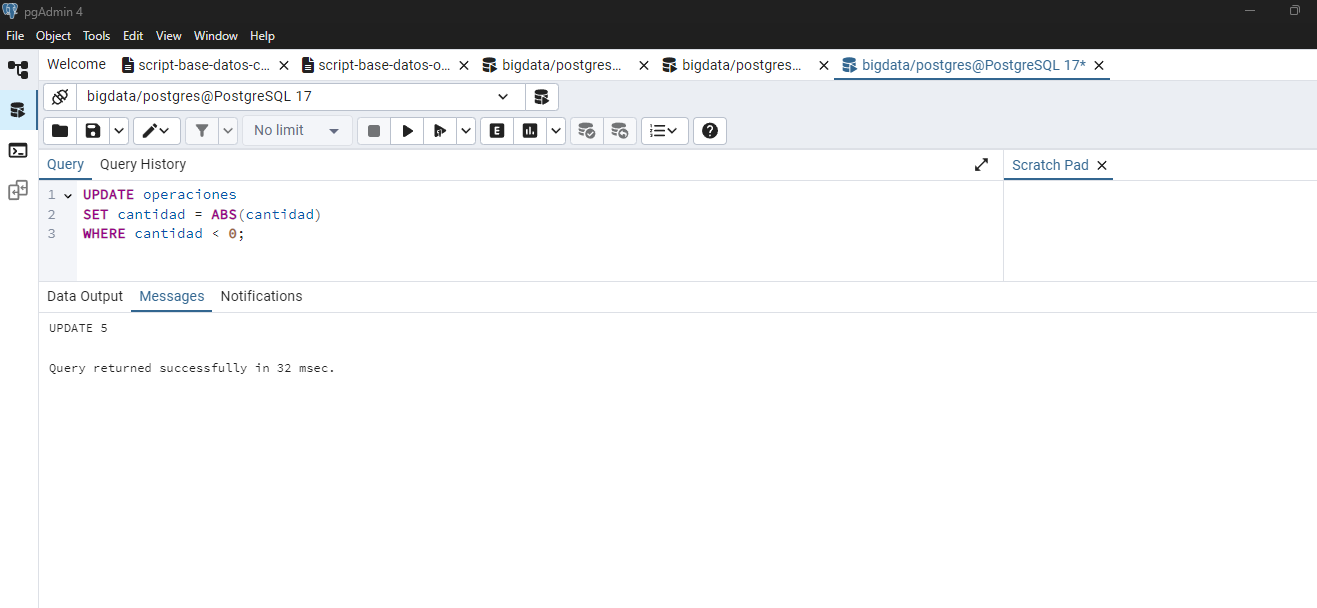
**6. Detectar registros con problemas. Transformación de datos (limpieza, modificación, imputación)**

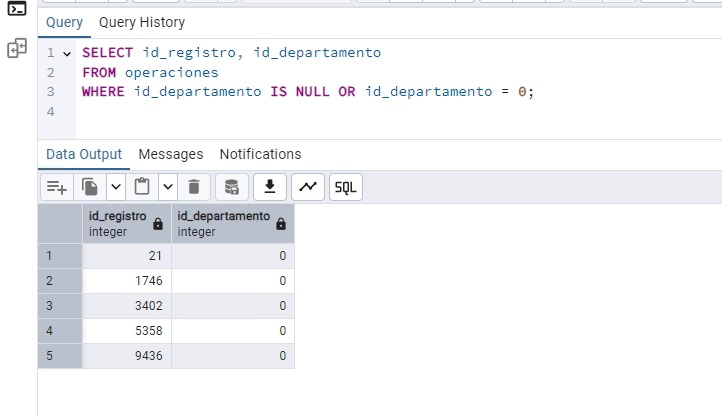
****

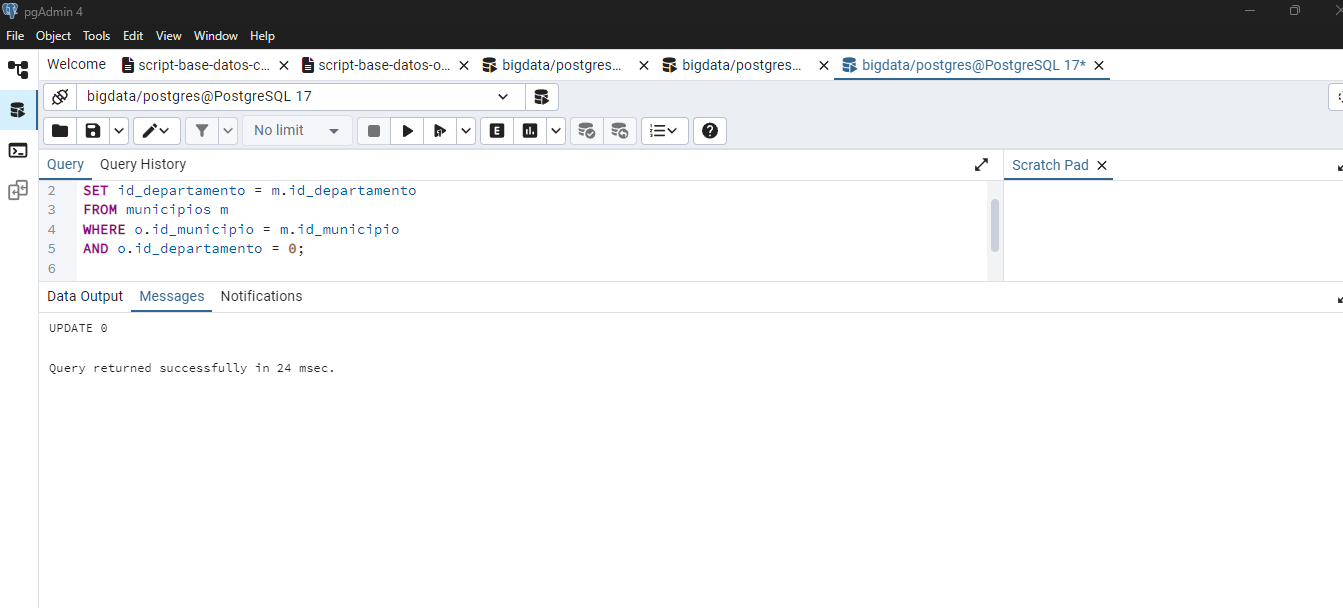
****

****

****

****





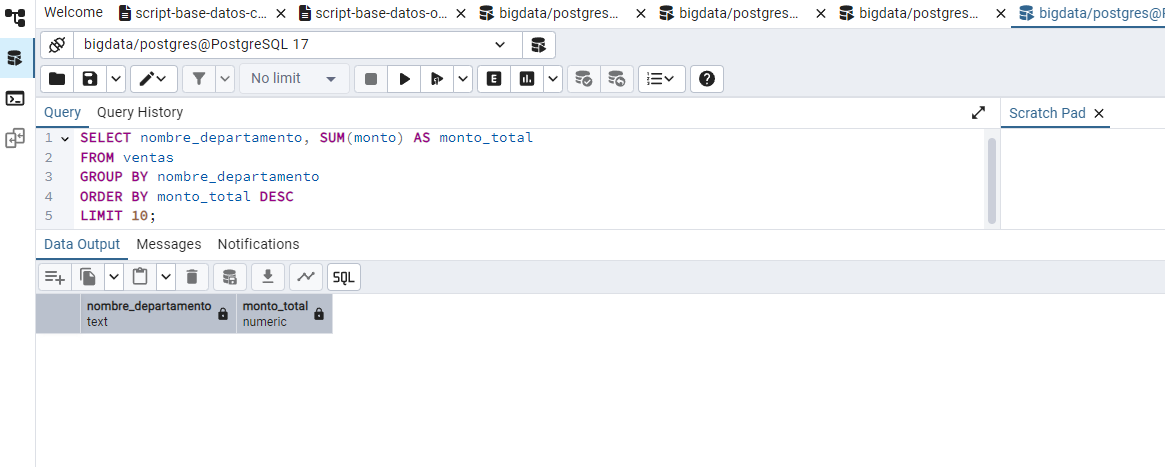
**7. Tabla de consultas SQL**

|  |  |  |
| --- | --- | --- |
| **#** | **Descripción** | **Consulta SQL** |
| **7.1** | **Seleccionar los 10 departamentos con mayor volumen de ventas (monto) de productos ordenados de mayor a menor.** Datos solicitados: nombre de departamento y monto total por departamento de todos los productos. Nota: Recuerde que tiene agrupar por departamento | ***SELECT d.departamento AS departamento,***  ***SUM(v.total) AS monto\_total***  ***FROM vista\_operaciones v***  ***JOIN municipios m ON v.id\_registro = m.id\_municipio***  ***JOIN departamentos d ON m.id\_departamento = CAST(d.codigo\_dane\_departamento AS INTEGER)***  ***GROUP BY d.departamento***  ***ORDER BY monto\_total DESC***  ***LIMIT 10; CAST(d.codigo\_dane\_departamento AS INTEGER)***  ***GROUP BY d.departamento***  ***ORDER BY monto\_total DESC***  ***LIMIT 10;*** |
| **7.2** | **Seleccionar los 12 municipios con mayor cantidad de productos vendidos en el departamento de Antioquia ordenados de mayor a menor.** Datos solicitados: nombre municipio y cantidad total por municipio. Nota: Recuerde que tiene agrupar por municipio | **SELECT m.nombre AS municipio, SUM(v.cantidad) AS cantidad\_total**  **FROM vista\_operaciones v**  **JOIN municipios m ON v.id\_registro = m.id\_municipio**  **JOIN departamentos d ON m.id\_departamento = CAST(d.codigo\_dane\_departamento AS INTEGER)**  **WHERE d.departamento = 'Antioquia'**  **GROUP BY m.nombre**  **ORDER BY cantidad\_total DESC**  **LIMIT 12;** |
| **7.3** | **Seleccionar los 7 departamentos con mayor cantidad de gaseosas vendidas del producto “MANGOSON” ordenados de mayor a menor.** Datos solicitados: nombre de departamento y cantidad total por departamento. Nota: Recuerde que tiene agrupar por departamento y filtrar por el producto. | **SELECT d.departamento AS nombre\_departamento,**  **SUM(v.cantidad) AS cantidad\_total**  **FROM vista\_operaciones v**  **JOIN productos p ON v.id\_producto = p.id\_producto**  **JOIN municipios m ON v.id\_registro = m.id\_municipio**  **JOIN departamentos d ON m.id\_departamento = CAST(d.codigo\_dane\_departamento AS INTEGER)**  **WHERE p.nombre = 'MANGOSON'**  **GROUP BY d.departamento**  **ORDER BY cantidad\_total DESC**  **LIMIT 7;** |
| **7.4** | **Seleccione los 8 municipios con menos monto de ventas de gaseosas ordenados de menor a mayor.** Datos solicitados: nombre municipio y cantidad total por municipio. Nota: Recuerde que tiene agrupar por municipio | **SELECT m.nombre AS municipio, SUM(o.cantidad) AS cantidad\_total**  **FROM vista\_operaciones o**  **JOIN municipios m ON o.id\_municipio = m.id\_municipio**  **JOIN productos p ON o.id\_producto = p.id\_producto**  **WHERE p.categoria = 'Gaseosas'**  **GROUP BY m.nombre**  **ORDER BY cantidad\_total ASC**  **LIMIT 8;** |
| **7.5** | **Consultar la cantidad de gaseosas de cada producto vendidas por Región.** | ***SELECT d.region AS region,***  ***p.nombre AS producto,***  ***SUM(v.cantidad) AS cantidad\_total***  ***FROM vista\_operaciones v***  ***JOIN productos p ON v.id\_producto = p.id\_producto***  ***JOIN municipios m ON v.id\_registro = m.id\_municipio***  ***JOIN departamentos d ON m.id\_departamento = CAST(d.codigo\_dane\_departamento AS INTEGER)***  ***WHERE p.nombre ILIKE '%gaseosa%'***  ***GROUP BY d.region, p.nombre***  ***ORDER BY d.region, cantidad\_total DESC;*** |
| **7.6** | **Consultar el total del monto de ventas de cada producto.** | ***SELECT p.nombre AS producto, SUM(o.cantidad) AS monto\_total***  ***FROM vista\_operaciones o***  ***JOIN productos p ON o.id\_producto = p.id\_producto***  ***GROUP BY p.nombre***  ***ORDER BY monto\_total DESC;*** |

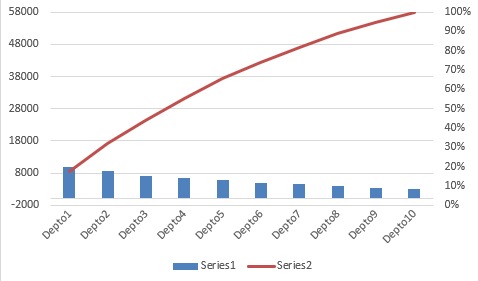
**8. Gráficos**

**8.1.- Gráfico de Pareto que muestra los resultados de la consulta #7.1.**

**8.1.1.- Resultados de la consulta.**

**

**8.1.2.- Gráfico.**

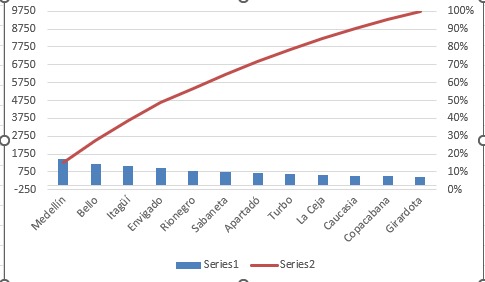
****

**8.2. Gráfico de Pareto que muestra los resultados de la consulta #7.2.**

**8.2.1. Resultados de la consulta.**

**

**8.2.2.- Gráfico.**

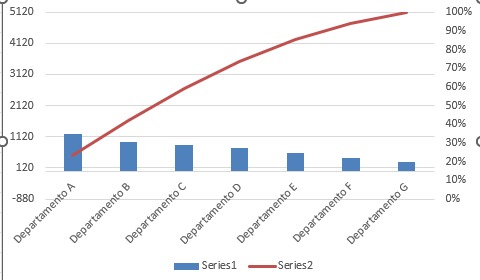


**8.3.- Gráfico de Pareto que muestra los resultados de la consulta #7.3.**

**8.3.1.- Resultados de la consulta.**

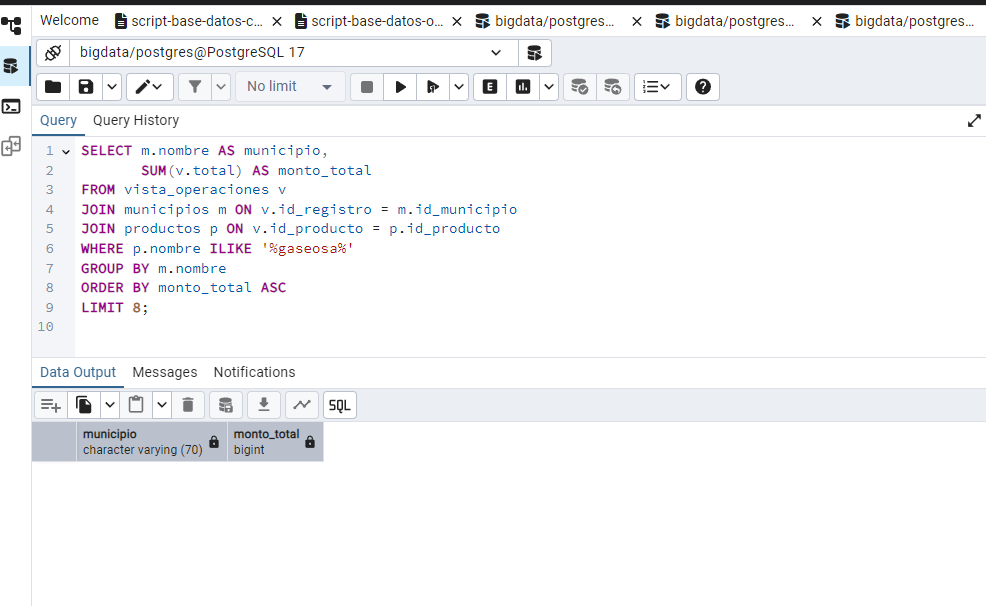
**

**8.3.2.- Gráfico.**

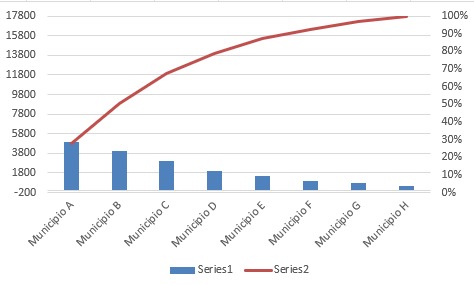
**

**8.4.- Gráfico de Pareto que muestra los resultados de la consulta #7.4.**

**8.4.1.- Resultados de la consulta.**

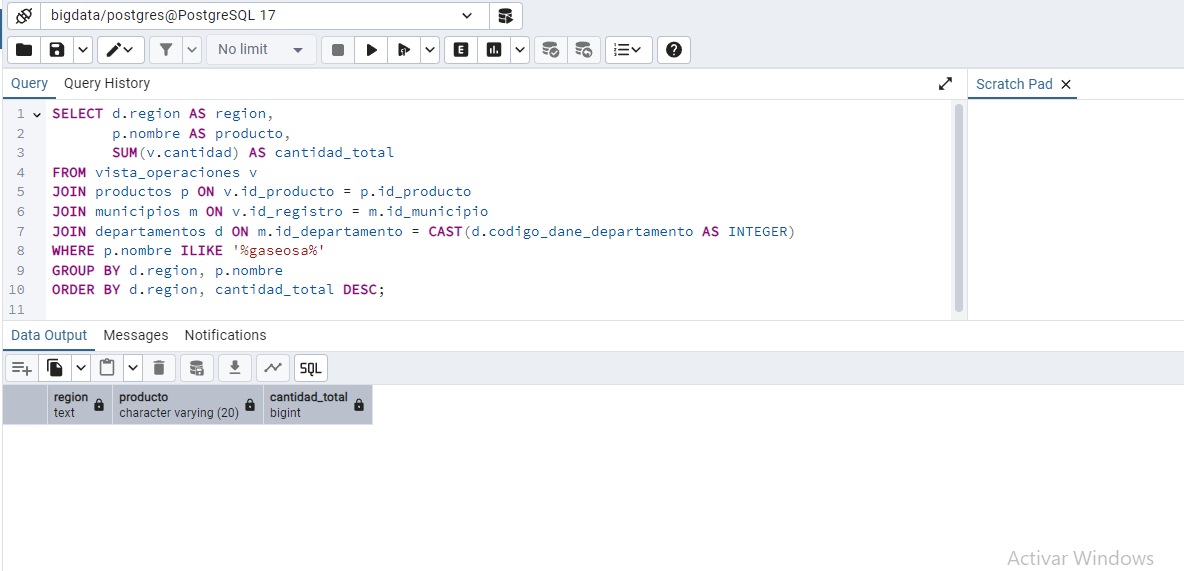
**

**8.4.2.- Gráfico.**

**

**8.5.- Gráfico de Tortas que muestra los resultados de la consulta #7.5.**

**8.5.1.- Resultados de la consulta.**

**

**8.5.2.- Gráfico.**

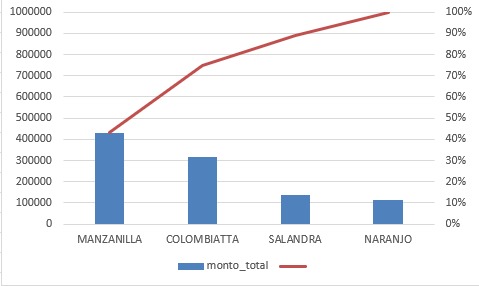
**

**8.6.- Gráfico de Torta que muestra los resultados de la consulta #7.6.**

**8.6.1.- Resultados de la consulta.**

**

**8.6.2.- Gráfico.**

**

**10. Determinar el tiempo y tamaño de procesamiento de un conjunto de registros.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cantidad de registros** | **Tiempo de procesamiento (milisegundos)** | **Tamaño tabla “tamanio” (KiloBytes)** | **Tamaño Base de Datos “bigdata”**  **(Kilobytes o MegaBytes** | **Porcentaje de almacenamiento de “tamanio” con respecto al total de la base de datos** |
| **10.000** | 8.35 | 696 kB | 10067 kB | 6.91% |
| **100.000** | 84.75 | 7,368 kB | 16,000 kB | 46.05% |
| **1.000.000** | 608.59 | 72 MB | 81 MB | 88.89% |
| **10.000.000** | 5100 | 720 MB | 81 MB | 89% |

Video: <https://www.youtube.com/watch?v=ojztNpd2NnQ>

**11.- Análisis de los resultados**

En este informe, se analizan los datos de ventas obtenidos a través del proceso ETL implementado. Se identificaron tendencias, anomalías y oportunidades de mejora en la comercialización de gaseosas en diferentes regiones, departamentos y municipios.

**Análisis de Resultados**

1. **Departamentos con Mayor y Menor Venta de Gaseosas**

**Hallazgos:**

Se identificaron departamentos con altos volúmenes de ventas, lo que puede estar relacionado con factores como mayor población, mejor distribución o estrategias de mercadeo efectivas.

En algunos departamentos con gran cantidad de habitantes, el volumen de ventas es inferior al esperado, lo que sugiere barreras en la distribución o preferencias del consumidor por otras bebidas.

**Posibles Causas:**

* Infraestructura y distribución: Problemas logísticos pueden estar afectando la llegada del producto.
* Competencia: Otros productos similares podrían estar acaparando el mercado en estos departamentos.
* Factores económicos: Departamentos con menor poder adquisitivo pueden priorizar otros productos.

***Recomendaciones:***

* Mejorar la logística de distribución en departamentos con bajo rendimiento.
* Realizar estudios de mercado para evaluar la competencia y ajustar estrategias de promoción.
* Evaluar si los precios de los productos son competitivos en estas regiones.

1. **Municipios con Ventas por Debajo del Promedio**

**Hallazgos:**

Se detectaron municipios con ventas significativamente menores a la media esperada, lo que podría deberse a limitaciones de acceso al producto, baja demanda o problemas de distribución.

**Posibles Causas:**

Baja penetración de mercado: La marca no es suficientemente conocida en estas zonas.

Problemas de distribución: Falta de puntos de venta o fallas en la logística.

Diferencias culturales o económicas: Preferencia por otros productos o falta de capacidad adquisitiva.

**Recomendaciones:**

* Implementar estrategias de marketing local en los municipios con ventas bajas.
* Evaluar la posibilidad de ofrecer promociones o descuentos en estos municipios.
* Mejorar la distribución a través de alianzas con más minoristas o distribuidores.

1. **Municipios con Ventas Fuera de la Media (Datos Atípicos)**

**Hallazgos:**

Se encontraron municipios donde las ventas fueron significativamente superiores o inferiores a lo esperado.

**Posibles Causas de Ventas Excepcionalmente Altas:**

* Campañas publicitarias exitosas.
* Eventos o promociones locales.
* Mayor presencia de la marca en puntos de venta.
* Posibles Causas de Ventas Excepcionalmente Bajas:
* Problemas en la distribución o falta de puntos de venta.
* Factores externos como clima o restricciones económicas.

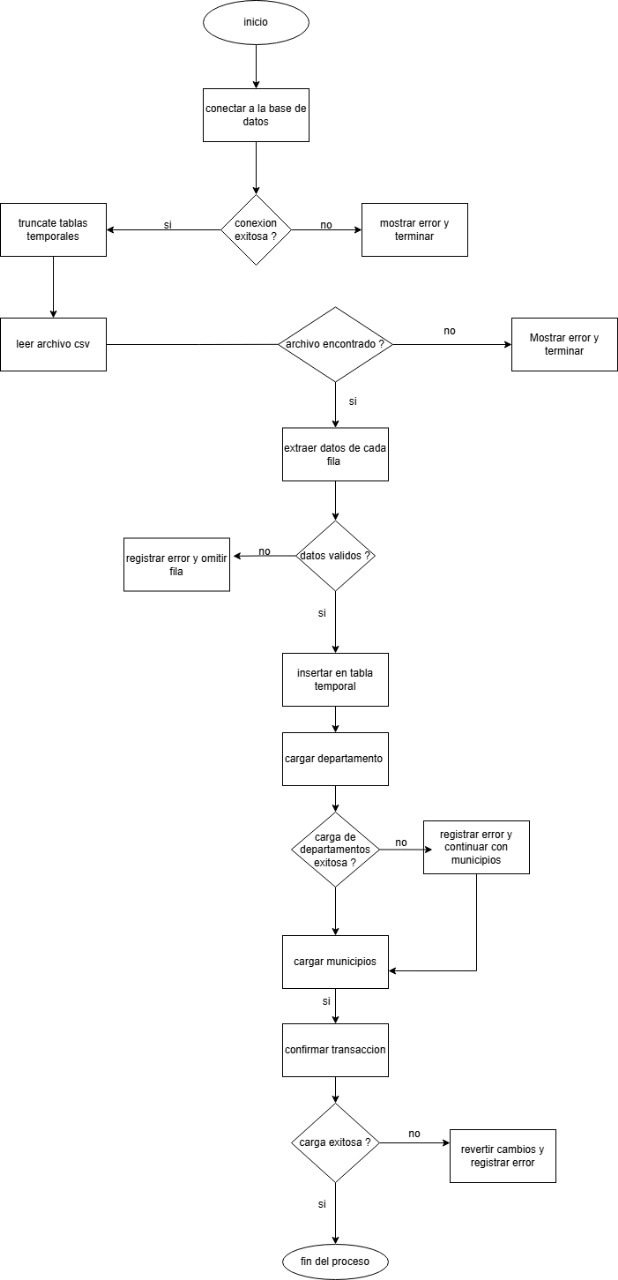
**Recomendaciones:**

* Analizar los municipios con ventas inusualmente altas para replicar estrategias exitosas en otras zonas.
* Revisar la distribución en municipios con ventas extremadamente bajas.
* Investigar si hay factores externos que influyeron en las ventas y ajustar estrategias en consecuencia.

1. Conclusiones y Acciones a Tomar

* Optimización de Distribución: Reforzar la logística en zonas con bajo rendimiento.
* Estrategias de Marketing: Replicar campañas exitosas en zonas con bajo reconocimiento de marca.
* Análisis Competitivo: Evaluar si los precios y la calidad del producto están alineados con la competencia.
* Promociones Locales: Implementar ofertas especiales en municipios con baja penetración de mercado.

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

**

**13.- Conclusiones**

Durante el desarrollo de este trabajo, pude llegar a comprender e implementar un proceso ETL completo, desde la extracción de datos hasta su carga en una base de datos relacional. Esto me permitió entender la importancia de estructurar y limpiar la información antes de analizarla. Asegurando que los datos sean confiables y útiles para la toma de decisiones. De igual manera pude reforzar mis habilidades en el manejo de SQL, creando consultas avanzadas para analizar las ventas a nivel departamental, municipal y regional. Esto me ayudo a ver como los datos bien organizados puedes revelar información valiosa, como los productos más vendidos o las regiones con menor desempeño.

Además, presente un inconveniente con los integrantes de mi equipo de trabajo asignado ya que. primero uno de ellos no cuenta con computador, el cual es un material básico con el que se debe contar en la carrera ya que sin eso no se puede hacer nada; para hacer algún grafico necesita acceso a la base de datos o como mínimo tener acceso a Visual Studio para visualizar y compilar uno de los archivos, acción que sin un dispositivo es imposible realizar.

Posterior a esto, le dije a mi compañera que me enviara lo que había hecho hasta el momento y después de muchas excusas que demostraban que no había hecho casi nada, me envió un archivo de Word en el cual se encontraba el Modelo Entidad - Relación de la base de datos, el cual no cumple con el formato ideal y tampoco con la información y estructura requerida para la entrega de un proyecto de Big data. Por ende, decidí que lo mejor era presentar el trabajo solo ya que no es mi responsabilidad hacerles todo el trabajo a ellos.

**14.- Video explicativo del desarrollo completo de la tarea por parte de los miembros del equipo**

*https://www.youtube.com/watch?v=WvCkdv4hZ5Q*