**INFORME**

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

**Grupo “4”**

**Miembros**

***Sebastián Acevedo Molina***

***Juan Esteban Castaño Montoya***

***Maria Fernanda Gutiérrez Mira***

**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.

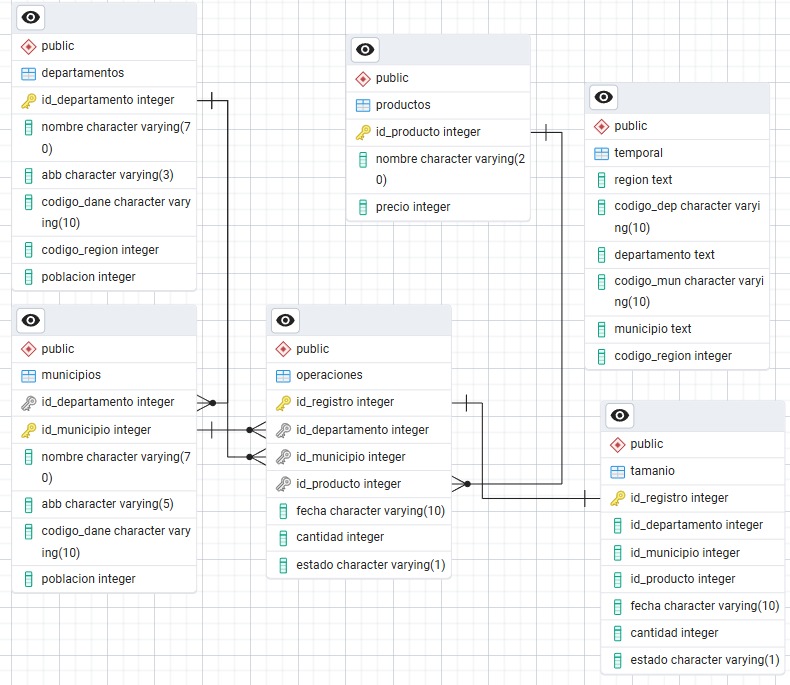
**1.- Descripción de la Tarea**

Este trabajo tiene como objetivo la correcta implementación de las herramientas ETL, para la empresa “gaseosas poderosas”. Su principal enfoque es migrar la información que se posee actualmente en hojas de cálculo, a bases de datos relacionales, usando PostgreSQL, esto con el afán de suplir la necesidad de tener la información organizada y centralizada, por regiones, departamento y municipios, que se dan a través de ventas; con el fin de evitar errores en la información.

A continuación, se proponen las acciones a realiza para cumplir con los requerimientos:

* Se analizará el caso de estudio, como también la fuente de datos con la que se cuenta (Hojas de cálculo), con el fin de encontrar donde se genera el déficit en la calidad de la información.
* Se diseñará el modelo entidad relación en notación de Chen.
* Diccionario de datos, donde se especifican los campos de cada tabla, tipos de datos, tamaño, restricciones y relaciones.
* Se extraen los datos por medio de la lectura de la información que se encuentre en las hojas de cálculo.
* Se hace limpieza de datos y corrección de estos.
* Se cargan los datos en tablas de BD.
* Se crearán consultas SQL que permitan analizar las ventas por cada región, municipio y periodo de tiempo.
* Se visualizan los resultados por medio de representaciones graficas de los principales indicadores de ventas, para apoyar la toma de decisiones.

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



**3.- Diccionario de Datos**













**4.- Corrida del algoritmo ETL**

A continuación, se describe la corrida del algoritmo ETL, implementado en Python, para el procesamiento de los datos de departamentos y municipios de Colombia. En este podemos integrar la información en una base de datos PostgreSQL garantizando la limpieza, consistencia y normalización de los registros.

Con este algoritmo resolvemos:

Implementa un proceso ETL sobre datos oficiales del DANE

El procedimiento se divide en tres fases:

1. Extracción: Lectura del archivo CSV que contiene regiones, departamentos y municipios.  
2. Transformación: Limpieza y estandarización de códigos, asignación de códigos de región, creación de identificadores únicos y normalización de nombres.  
3. Carga: Inserción de los datos en la base de datos PostgreSQL a través de una tabla temporal, para luego distribuirlos en las tablas finales (departamentos y municipios).

Los datos se obtienen de:

- Fuente primaria: Archivo CSV descargado de Datos Abiertos de Colombia:  
 https://www.datos.gov.co/Mapas-Nacionales/Departamentos-y-municipios-de-Colombia/xdk5-pm3f/data  
- Fuente destino: Base de datos PostgreSQL (data-2025) en un servidor local, gestionada con pgAdmin4.

Se realizan los siguientes procesos:

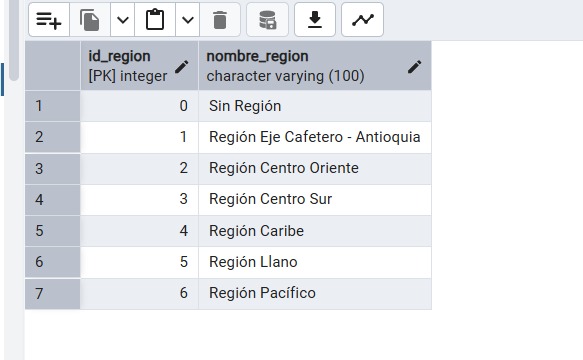
- Limpieza de tablas.  
- Inserción de registros en una tabla temporal.  
- Agrupamiento y conteo de municipios por departamento.  
- Construcción de claves primarias.  
- Inserción final en las tablas departamentos y municipios.

Los resultados obtenidos:

Como resultado, la base de datos queda organizada en tres tablas principales:

1. Temporal, que son los datos crudos provenientes del archivo CSV.  
2. departamentos, que tiene identificadores únicos y código de región.  
3. Municipios, que contiene identificadores únicos y relación con su departamento.  
  
Estos datos permiten generar consultas analíticas como:  
- Ventas realizadas pro municipio.  
- Departamentos con más ventas.  
- Municipios con ventas más bajas.

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

****

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











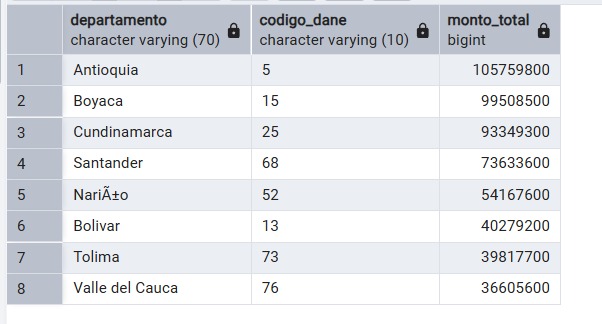
**7.- Tabla de consultas SQL**

|  |  |  |
| --- | --- | --- |
| **#** | **Descripción** | **Consulta SQL** |
| **7.1** | **Seleccionar los 8 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 | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT*** |
| **7.2** | **Seleccionar los 15 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 | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT*** |
| **7.3** | **Seleccionar los 5 departamentos con mayor cantidad de gaseosas vendidas del producto “MANZALOCA” 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. | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT*** |
| **7.4** | **Seleccione los 5 municipios con el menor monto de ventas de gaseosas ordenados de menor a mayor.** Datos solicitados: nombre del departamento al que pertenece, nombre municipio y monto total de ventas por municipio. Nota: Recuerde que tiene agrupar por municipio | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT*** |
| **7.5** | **Consultar la cantidad de gaseosas vendidas de cada producto por cada región ordenados de mayor a menor.** | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT. Requiere el campo id\_region.*** |
| **7.6** | **Consultar el total del monto de ventas de cada producto en Antioquia de mayor a menor.** | ***Entregue la consulta en un script. Ese script debe estar colocado en el repositorio GIT.*** |

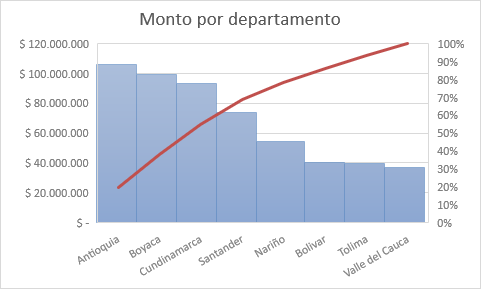
**8.- Gráficos**

**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.**

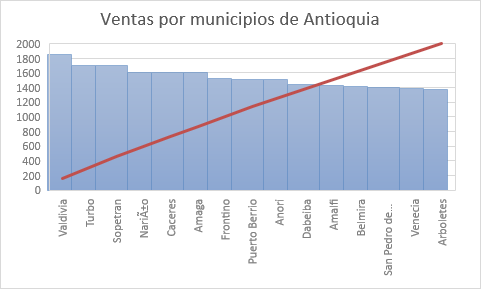


**8.2.- Gráfico de Pareto que muestra los resultados de la consulta #7.2.**

**8.2.1.- Resultados de la consulta.**

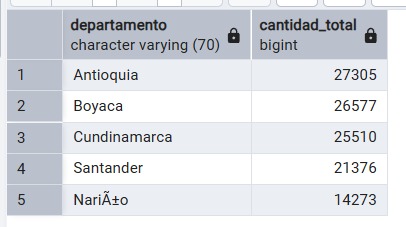
**

**8.2.2.- Gráfico.**

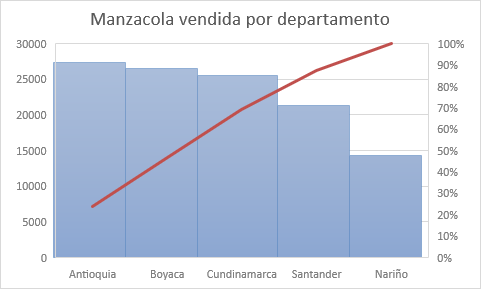


**8.3.- Gráfico de Pareto que muestra los resultados de la consulta #7.3.**

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

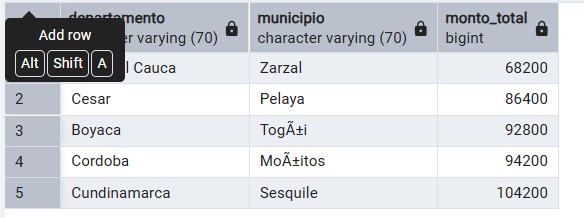
**

**8.3.2.- Gráfico.**

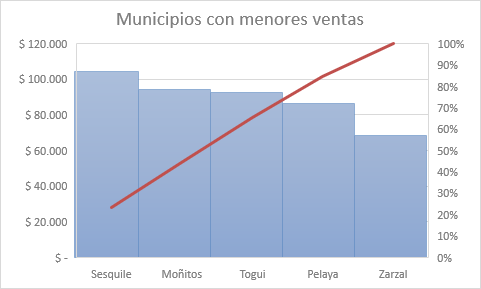


**8.4.- Gráfico de Pareto que muestra los resultados de la consulta #7.4.**

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

**

**8.4.2.- Gráfico.**

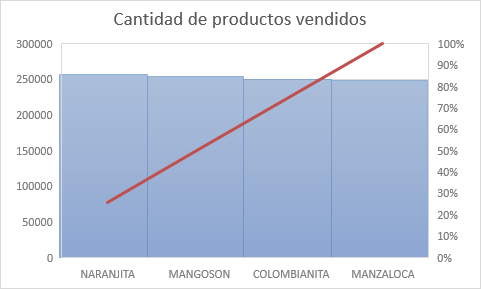


**8.5.- Gráfico de Tortas que muestra los resultados de la consulta #7.5.**

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

**

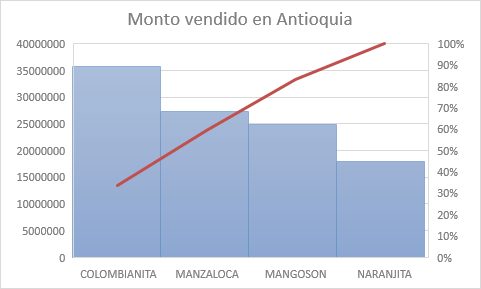
**8.5.2.- Gráfico.**



**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.**



**11.- Análisis de los resultados**

Una vez realizado el análisis correspondiente a los datos de ventas de “GASEOSAS PODEROSAS”, se da como resultado una serie de secuencias que nos pueden dar como resultado la toma de decisiones estratégicas. A continuación, se enumeran las más comunes:

* Existe un volumen de ventas mayor en pocos departamentos, como Antioquia y Valle del Cauca, lo que refleja que se encuentra consolidado en estas regiones.
* Hay departamentos que, a pesar de tener una alta población, no tiene una presencia sólida, como Cundinamarca y Antioquia.
* La manzacola es el producto mas vendido en varios departamentos, en cambio productos como naranjita han tenido menor desempeño en sus ventas.
* Se encontraron registros con ventas negativas, ceros y fechas mal formateadas, lo que afectaba los análisis iniciales.
* Tras la limpieza, se corrigieron estos valores, mejorando la confiabilidad de la información.

A raíz de los hallazgos encontrados, se realizan las siguientes recomendaciones:

* Reforzar estrategias de marketing en los departamentos que presentan bajo desempeño, como el Atlántico y Cundinamarca.
* Expandir los puntos de venta en los municipios que presentan mejor acogida del producto.
* Potenciar los productos mas vendidos, en este caso Manzacola, usándola como un gancho comercial.

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

Enlace de acceso:  
<https://miro.com/app/board/uXjVJJ3OKBY=/>

**13.- Conclusiones INDIVIDUALES**

* El análisis revela que, aunque la empresa tiene un mercado sólido en regiones clave, existen oportunidades de crecimiento en departamentos y municipios con bajo rendimiento. Además, la limpieza de datos fue fundamental para obtener información confiable.
* La implementación de este proceso, dio la posibilidad de centralizar y limpiar la información de ventas, por departamentos y municipios en una base de datos relacional confiable. Eliminando de esta manera la redundancia y los errores presentes en las hojas de cálculo, garantizando la calidad en la información, para mejorar la toma de decisiones.
* De acuerdo al análisis de los resultados se evidenciaron patrones de consumo y oportunidades de mejora en la estrategia comercial. Se identificaron departamentos y municipios con alto desempeño y otros con ventas por debajo de lo esperado, lo que abre la posibilidad de diseñar campañas focalizadas, fortalecer la distribución y mejorar la gestión de productos con bajo rendimiento.

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

*Se debe elaborar un video dónde aparezcan todos los miembros del equipo de trabajo. Se debe explicar el desarrollo de la tarea y mostrar en pantalla la ejecución de los algoritmos y la construcción de las consultas SQL. Mostrar cómo se realiza la actividad en el pgAdmin4 y cómo construyeron los gráficos en base a las consultas SQL realizadas. Coloque el LINK del video en esta sección. El vídeo debe tener una duración mínima de 5 minutos y máximo de 15 minutos. De tal modo, que deben practicar la presentación para que sean eficientes en el uso del tiempo y aborden de manera concreta todas las actividades. Cada miembro del equipo debe presentarse rápidamente (se debe ver su rostro o al menos el avatar) y explicar su aporte al trabajo.* ***NOTA: No se trata de leer el informe, se trata de mostrar código en ejecución***