**INFORME**

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

**Grupo 6**

**Miembros**

Juan Pablo Arango Gaviria

**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**

Ejecución de procesos de extracción, limpieza, normalización, agregación, imputación, carga y visualización de datos para la empresa de refrescos.

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

Diagrama

El contenido generado por IA puede ser incorrecto.

**3.- Diccionario de Datos**

*Elabore un Diccionario de Datos. El Diccionario de Datos debe contener la información completa de cada tabla: nombre de la tabla, nombre del campo, tipo de dato, tamaño, valores que puede contener el campo, clave primaria, clave foránea, claves únicas, relación con otra tabla, entre otros. Se debe colocar aquí el diccionario (no fuera de este informe)*

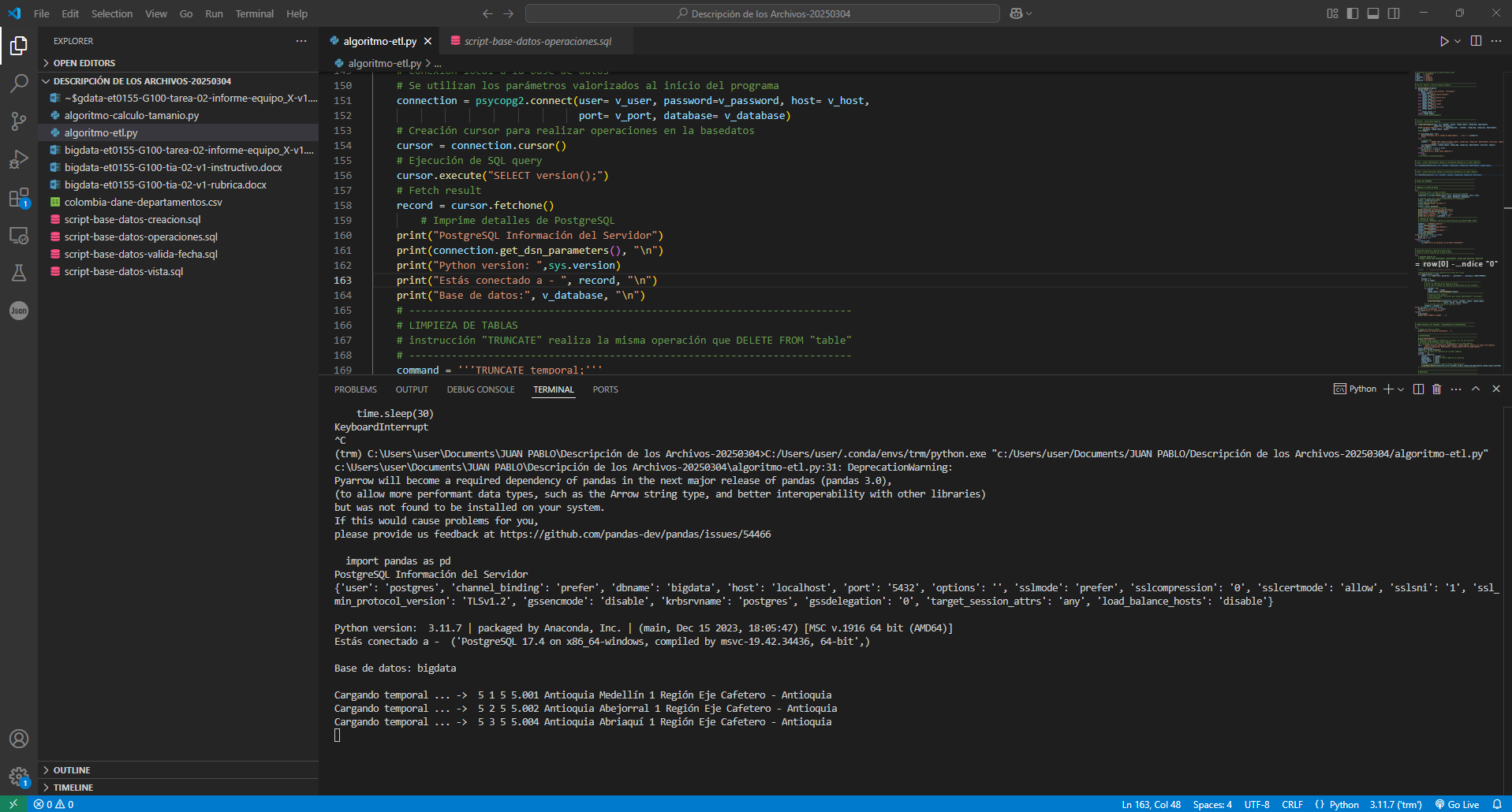
**4.- Corrida del algoritmo ETL**

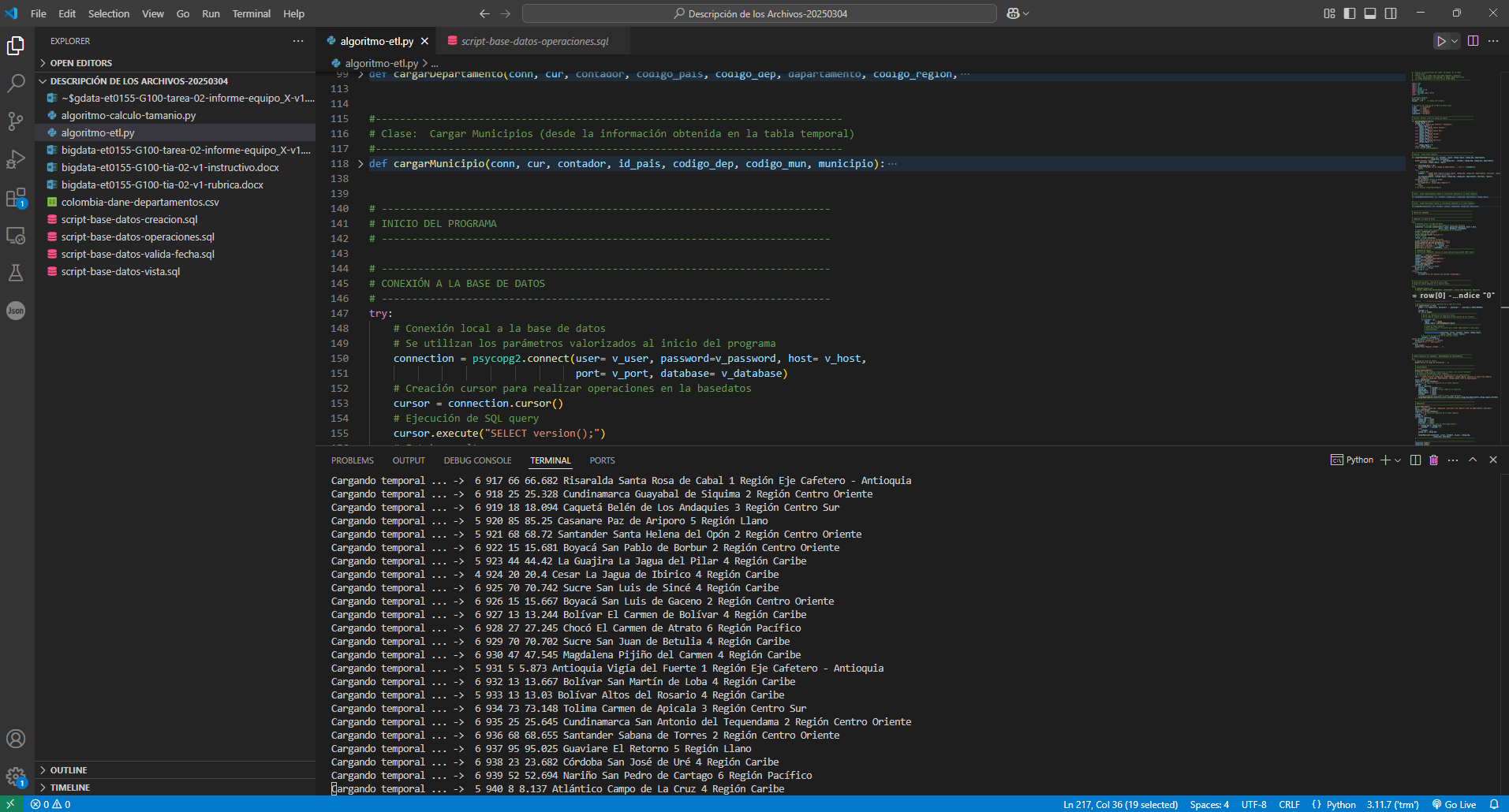
El algoritmo se encarga de extraer, transformar y cargar en una base de datos PostgreSQL toda la información de una fuente externa, en este caso un archivo Excel CSV que contiene los departamentos y municipios de Colombia.

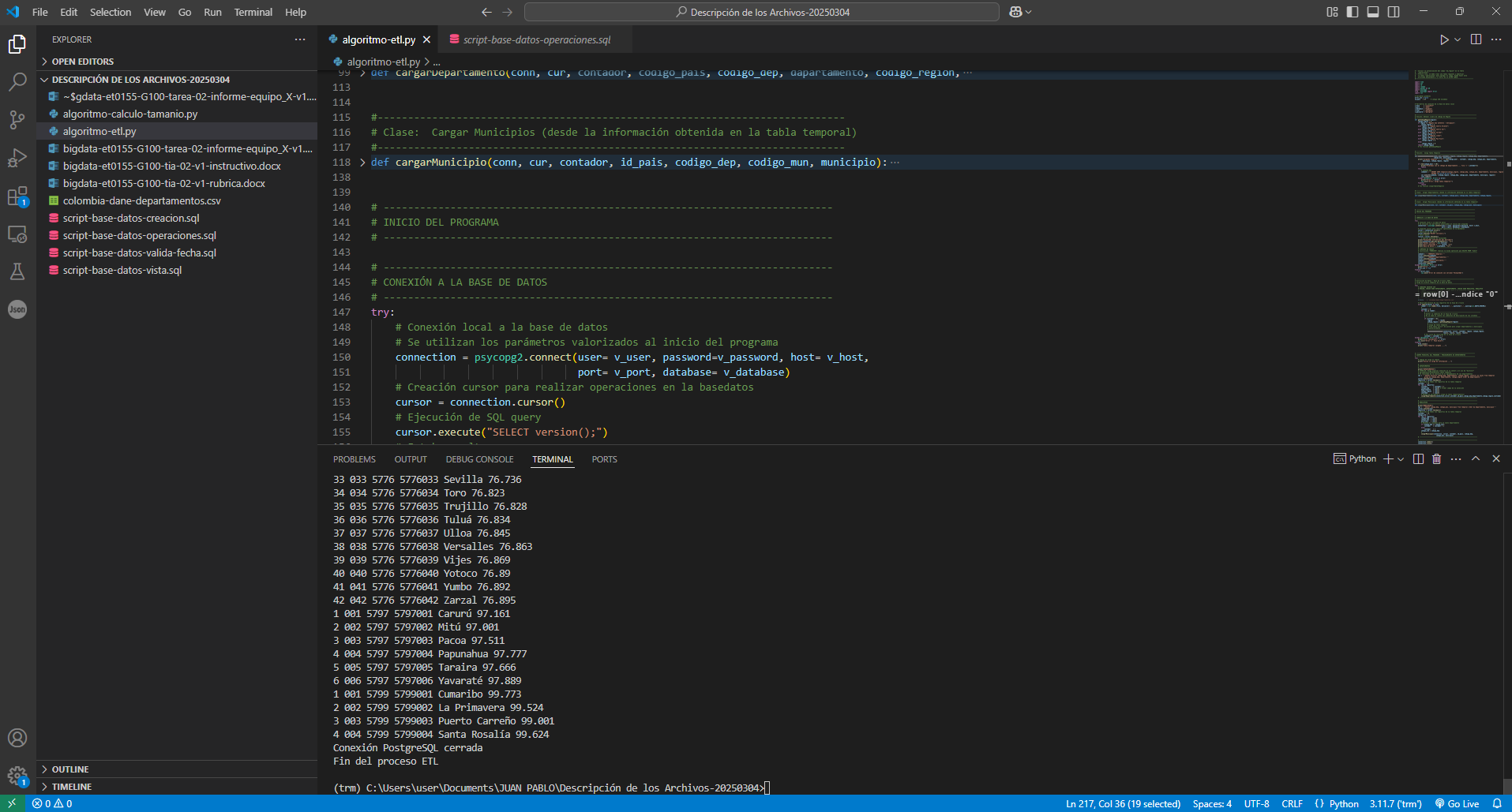
La ETL hace los siguientes procesos:

1. Conexión con la base de datos postgresql.
2. Realiza la extracción de departamentos y municipios desde el archivo CSV (colombia-dane-departamentos.csv).
3. Carga los datos a la tabla temporal de la base de datos bigdata.
4. Filtra e inserta los registros de los municipios que no se repiten a la tabla municipios.
5. Asigna un id único combinando id del país y el sufijo e inserta a la tabla departamentos.
6. Finaliza la conexión a la base de datos.

Con este proceso finalmente se obtienen los registros en la tabla municipios y departamentos y adicionalmente una tabla temporal con datos que aún se deben limpiar y normalizar.







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

*Debe agregar el campo “id\_region” en la tabla operaciones y valorizar el campo con el código región. Utilice cualquier estrategia que considere oportuna para resolver el problema. Este campo es necesario para resolver la consulta relacionada con Regiones. Debe crear una nueva tabla “regiones” para obtener la descripción eventualmente.*

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

*Una vez que haya cargado la tabla “operaciones”, debe realizar un análisis exploratorio de los datos. Según sea el caso, debe realizar: limpieza de datos (eliminación en caso que falte mucha información), corrección de algún dato que considere erróneo y/o imputación de datos faltantes. Puede utilizar la función de fechas válidas y no válidas para verificar cuáles fechas son correctas y cuáles no.*

*Para resolver estos problemas o algunos de ellos, puede hacerlo a través de análisis exploratorio con comandos SQL o código Python (utilice librería Pandas) o a través de programación Python. El docente le suministró código SQL para resolver parcialmente algunos problemas de fechas. Debe ingeniarse para resolver el resto.*

*Utilice el cuadro para colocar los registros que detectó con problemas en la tabla “operaciones”. Se requieren los siguientes datos: número de registro, dato (campo) con problemas, tipo de problema, campo de la bases de datos afectado por el problema, descripción del problema y la solución para “salvar” el dato y el registro.*

***NOTA:*** *Esta información será muy útil para implementar la tarea de recuperación del registro para hacerlo válido. Debe agregar un campo en la tabla operaciones que muestre que el dato fue “modificado” y otro campo que diga la causa. El resto de los registros deben ser valorados cómo válido en este nuevo campo.*

***Estos son los problemas que puede encontrar en los registros:***

***1.- Fechas con un formato diferente a “AAAA-MM-DD”.*** *En el caso particular de las fechas que no se encuentren en formato AAAA-MM-DD, se debe transformar la fecha a dicho formato.*

***2.- Cantidades en 0.*** *Se puede utilizar el PROMEDIO DE VENTAS del municipio para imputar el dato (obviamente, considerando solamente los registros con cantidades).**Los registros con cantidades en “0”, se les debe imputar un dato sacando el promedio de cantidades del municipio al que pertenece*

***3- Cantidades negativas*** *(error de ingreso del dato por parte del operador colocando el signo negativo, el dato está bien) ¿Qué de hacer para resolver?**En el caso de números negativos, debe convertirlos en positivos porque fue un error de introducción de datos.*

***4.- Falta el código de departamento*** *¿Cómo puedo lograr determinar ese código? Utilice datos que ya conozca dentro la tabla “operaciones”. Imputar el código de departamento dónde se encuentre en “0”.*

***5.- Falta el código de producto en el Municipio “Tame” (solamente se vende “COLOMBIANITA” en ese municipio****)* ***del departamento*** *“****Arauca****”. Asignar el código de producto dónde se encuentran en “0”*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Número Registro** | **Tipo de problema** | **Campo (BD)** | **Problema detectado** | **Solución** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**7.- Tabla de consultas SQL**

*Antes de proceder a realizar las siguientes consultas, debe haber realizado el proceso de “limpieza” de la tabla operaciones. No debe proceder a realizar estas consultas sin haberlo hecho porque la información no será correcta. La gerencia estará muy pendiente de este tipo de omisiones.*

*La forma de verificar que hubo una limpieza total es que la VIEW “view\_opeaciones” arroje el mismo número de registros que la tabla operaciones.*

***NOTA- ATENCIÓN****: SE DEBE USAR EXCLUSIVAMENTE la VIEW vista\_operaciones en TODAS LAS CONSULTAS*

|  |  |  |
| --- | --- | --- |
| **#** | **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 | ***Escriba aquí la consulta SQL que cumple con la información solicitada*** |
| **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 | ***Escriba aquí la consulta SQL que cumple con la información solicitada*** |
| **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. | ***Escriba aquí la consulta SQL que cumple con la información solicitada*** |
| **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 | ***Escriba aquí la consulta SQL que cumple con la información solicitada*** |
| **7.5** | **Consultar la cantidad de gaseosas de cada producto vendidas por Región.** | ***Escriba aquí la consulta SQL que cumple con la información solicitada. Requiere el campo id\_region.*** |
| **7.6** | **Consultar el total del monto de ventas de cada producto.** | ***Escriba aquí la consulta SQL que cumple con la información solicitada*** |

**8.- Gráficos**

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

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

*Debe utilizar el “Query Editor” de pgAdmin para correr la consulta. Capture y coloque aquí un “pantallazo” del resultado de registros obtenidos en pgAdmin.*

**8.1.2.- Gráfico.**

El gráfico debe mostrar en el “eje horizontal X” los departamentos y en el “eje vertical Y” las cantidades vendidas por departamento. Realizar el gráfico en Excel. Debe utilizar en la Hoja de cálculo la información obtenida en pgAdmin. Inserte aquí el gráfico obtenido en la hoja de cálculo o la captura de pantalla del mismo. Nota: recuerde que es un Diagrama de Pareto.

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

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

*Ídem al punto 8.1.1*

**8.2.2.- Gráfico.**

*Ídem al punto 8.1.2 (solo que se trata de municipios)*

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

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

*Ídem al punto 8.1.1*

**8.3.2.- Gráfico.**

*Ídem al punto 8.1.2*

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

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

*Ídem al punto 8.1.1*

**8.4.2.- Gráfico.**

*Ídem al punto 8.1.2*

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

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

*Ídem al punto 8.1.1*

**8.5.2.- Gráfico.**

*Ídem al punto 8.1.2*

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

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

*Ídem al punto 8.1.1*

**8.6.2.- Gráfico.**

*Ídem al punto 8.1.2*

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

*Debe utilizar el algoritmo de cálculo de tiempo y almacenamiento. Este algoritmo realiza una generación aleatoria de operaciones y almacena los resultados en la tabla “****tamanio****”. Debe correr el algoritmo para las diferentes cantidades de registros solicitados y registrar los resultados obtenidos. Por otro lado, se debe modificar el algoritmo para implementar un control de tiempo de procesamiento desde el momento que arranca la generación de registros hasta que termina. En el video debe mostrar el algoritmo en ejecución para la cantidad de 10.000 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** |  |  |  |  |
| **100.000** |  |  |  |  |
| **1.000.000** |  |  |  |  |
| **10.000.000** |  |  |  |  |

**11.- Análisis de los resultados**

*Debe presentar ante la Junta Directiva el análisis de los resultados de esta situación de ventas hipotética. Debe pensar y razonar como un “científico de datos” que ha descubierto información valiosa y la quiere trasmitir a sus superiores para que estos tomen decisiones. Por lo tanto, además de analizar los resultados obtenidos, debe proponer o sugerir acciones a tomar. Recuerde, es una situación hipotética.*

*Posibles preguntas que se puede hacer un Científico de Datos:*

* *¿Por qué en este departamento se vendieron más gaseosas de todas los tipos?*
* *¿Por qué en este departamento tan importante se vendieron menos gaseosas que en otros departamentos con menor población?*
* *¿Por qué estos municipios tuvieron ventas tan bajas?*
* *¿Qué sucedió en estos municipios que arrojaron ventas fuera de la media esperada? (Datos atípicos)*

*Así como las preguntas planteadas anteriormente, pueden surgir muchas más. La Ciencia de Datos es un Arte … todo depende de la mano del artista que de sus mejores pinceladas para obtener información valiosa para las instituciones.*

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

*Debe presentar un diagrama de flujo detallado del programa que realiza el procesamiento ETL.*

**13.- Conclusiones**

*Presente sus conclusiones en relación al trabajo realizado y cómo impacta su vida académica y profesional*

**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 8 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.*