Proyecto 1– Introducción a la ciencia de los datos
Héctor Fabio Ocampo Arbeláez
R.A. 1

El objetivo es aplicar los conocimientos adquiridos en bases de datos transaccionales y analíticas, bodega de datos, ETL y análisis descriptivo.

Los estudiantes deben trabajar en grupos de 3 personas.

Data Set:

La base de datos recomendada para este proyecto es:

https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset/data?select=customer_shopping_data.csv

Este dataset contiene información sobre compras de clientes en una tienda minorista, incluyendo productos, categorías, precios, cantidad comprada y métodos de pago.

Nota: Si desean utilizar otro conjunto de datos, deben solicitar aprobación previa antes de proceder con su proyecto.

Desarrollo del proyecto:

1. **Diseño del Modelo de la Bodega de Datos (20%)**
    a. Analizar la estructura del dataset y comprender sus atributos.
    b. Decidir qué modelo de bodega de datos utilizar (**Estrella o Copo de Nieve**) y justificar la decisión.
    c. Diseñar el **diagrama de tablas** para el modelo elegido.
    d. Crear la bodega de datos en **PostgreSQL** según el diseño elegido.

    Entregables:
        i. Documento con el diagrama de la bodega de datos y justificación del modelo seleccionado.
        ii. Script SQL con la creación de las tablas en PostgreSQL.

2. **Extracción, Transformación y Carga de Datos (30%)**

    Diseñar y desarrollar un proceso ETL que:

    a. Extraiga los datos desde el dataset de Kaggle usando **Pandas**.
    b. Transforme los datos al modelo definido anteriormente.
    c. Cargue los datos en la bodega de datos de PostgreSQL.

d. Usar **SQLAlchemy** para la carga de datos.

Entregables:

i. Código del proceso ETL en Python.
ii. Documento con la explicación del proceso y las transformaciones realizadas.
iii. Comprobación de que los datos han sido correctamente insertados en la base de datos.

3. Consultas Analíticas en SQL (20%)

a. Diseñar consultas SQL que respondan preguntas de negocio:
   i. Total de ventas por categoría de producto.
   ii. Clientes con mayor volumen de compras.
   iii. Métodos de pago más utilizados.
   iv. Comparación de ventas por mes.
b. Optimizar las consultas utilizando índices y agregaciones.

Entregables:

i. Documento con las consultas SQL y su explicación.
ii. Capturas de pantalla o resultados obtenidos de PostgreSQL.

4. Análisis Descriptivo y Visualización de Datos (20%)

a. Elegir las gráficas más adecuadas para representar los datos obtenidos (barras, líneas, tortas, histogramas, etc.).
b. Implementar visualizaciones usando Python (Matplotlib, Seaborn, Power BI o Tableau).
c. Realizar un análisis descriptivo sobre:
   i. Tendencias en los datos.
   ii. Insights obtenidos de las consultas.
   iii. Posibles mejoras para el negocio.

Entregables:

iv. Código Python de las visualizaciones.
v. Documento con capturas de pantalla y explicaciones de los análisis realizados.

5. Conclusiones y Presentación Final (10%)

a. Elaborar un informe final con:
   i. Diseño de la bodega de datos.

   ii. Explicación del proceso ETL.
   iii. Consultas analíticas desarrolladas.
   iv. Visualizaciones y análisis de datos.
b. Escribir las conclusiones relacionadas con el proyecto

Evaluación del Proyecto

Fase	Ponderación (%)
Diseño del modelo de bodega de datos	20%
Implementación de ETL	30%
Consultas Analíticas en SQL	20%
Análisis Descriptivo y Visualización	20%
Conclusiones y Presentación	10%
Total	**100%**