TECNICATURA SUPERIOR EN

Ciencia de Datos e Inteligencia Artificial

Trabajo Práctico

Exploración, Transformación y Limpieza de Datos utilizando Pandas

Proceso ETL para Análisis de Datos de Ventas y Clientes

Análisis de un archivo .csv

Objetivo del trabajo práctico:

El objetivo de este trabajo práctico es permitir a los estudiantes adquirir experiencia en el proceso ETL (Extracción, Transformación y Limpieza) de datos al trabajar con un Data Frame que combina información de ventas y datos de clientes.
A través de este trabajo, los estudiantes aprenderán a integrar y preparar datos de manera efectiva para su análisis.
Para ello, deberán trabajar con el archivo **"customer_data.csv" y "sales_data.csv"** del dataset "**Sales and Customer data".** Deberán buscarlo y bajarlo del repositorio de la plataforma **"Kaggle",** el cual pueden accederlo cómo está explicado en los videos subido a la plataforma.

El trabajo práctico deberá ser realizado, copiando el "script" de solución de los ítems consignados y una captura de su resolución del problema ejecutado en **Jupyter**.

Importante: Debe aclarar correctamente a que ítem está referenciando con cada solución de código y captura.

El formato del trabajo respetará las siguientes consignas:

- Letra: Arial Tamaño: 12
- Carátula con el nombre y apellido del alumno, materia y cohorte correspondiente.
- Orden y presentación. Se deberá respetar la estructura del trabajo práctico, como se establece en los apartados subsiguientes.

Estructura del trabajo práctico:

En este trabajo, se proporcionará a los estudiantes un archivo CSV que contiene dos conjuntos de datos: uno con información de **ventas** y otro con **datos de clientes**.
Los estudiantes deberán realizar las siguientes tareas:

Extracción de Datos (Extract):

1. Cargar los datos de ventas y clientes desde el archivo CSV en dos DataFrames distintos.

2. Describir el proceso de extracción y cómo se acceden a los datos en los DataFrames.

3. Concatenar los dos Data Frames anteriores, en uno final con información relevante. No borrar los Dataframes de ventas y clientes.

Transformación de Datos (Transform):

Realizar operaciones de limpieza y preparación de datos en ambos DataFrames. Esto incluirá la gestión de valores nulos, la estandarización de formatos de fecha, y la unión de datos de ventas y clientes utilizando el campo común **"id del cliente"**.

4. Realizar transformaciones adicionales, como determinar el modo de pago más frecuente de todos los clientes y a su vez categorizados por género. Realice una categorización de clientes de acuerdo a su forma de pago (por edad y género).

    Métodos de pagos realizados por el rango etario de 25 a 35 años

    Métodos de pago más utilizados por las mujeres.

    Precios por categoría de productos

5. Documentar las transformaciones realizadas en detalle y explicar su justificación.

Limpieza de Datos (Load):

Crear un nuevo Data Frame que contenga los datos limpios y transformados que serán utilizados para análisis posteriores.

6. Explicar cómo se carga este nuevo Data Frame y si se aplican restricciones de integridad en este paso.

Carga de datos y Consultas SQL:

Llevar los datos limpios del ETL a una base de datos relacional (MariaDB o similares) y ejecutar consultas SQL básicas que respondan a las mismas preguntas planteadas en el punto “Transformación de Datos (Transform)”

Análisis de Datos:

Realizar un análisis exploratorio de los datos para extraer información valiosa, como el comportamiento de compra por género o grupo de edad, precios más altos y bajos por categorías de productos, etc.

Realizar un resumen, evaluación y síntesis del estudio.

Entregables: Los estudiantes deberán entregar un informe que incluya los siguientes elementos:

- Una descripción detallada de las operaciones de extracción, transformación y limpieza realizadas en los datos de ventas y clientes.
- El código fuente utilizado para llevar a cabo el proceso ETL.
- Un DataFrame que contenga los datos limpios y transformados.
- El esquema de la base de datos y las sentencias SQL utilizadas para las consultas.
- En caso de realizar análisis, resumen final de la misma, cierre, resultados relevantes y síntesis del estudio.

Notas Importantes: Se recomienda a los estudiantes utilizar herramientas como Pandas en Python para llevar a cabo las operaciones de ETL. Además, se espera que sigan buenas prácticas de programación y documentación en su trabajo.

Actividades y evaluación:

- **Investigación bibliográfica:**

  Los estudiantes deberán investigar y analizar fuentes académicas para obtener una comprensión más profunda de los conceptos fundamentales.