Exploración y visualización de datos en Python

Ejercicio de Aplicación Autónomo 2

Objetivo y alcance del trabajo

Desarrollar habilidades prácticas en la adquisición, limpieza, transformación y análisis inicial de datos utilizando Python y librerías como Pandas, integrando datos de fuentes externas (APIs o datasets públicos). Para esto, se recomienda utilizar Google Colab que permite escribir y ejecutar código Python en la nube, disponible mediante el navegador web, e integrarlo con bloques de texto en formato Markdown para documentación complementaria. Se debe presentar un solo documento tipo notebook en formato "ipynb" y adjuntar los archivos que se soliciten en el desarrollo. El nombre del archivo a entregar debe tener el siguiente formato: "AMGD_CP_W2_G#", en donde '#' es el número de grupo.

Contexto del Problema:

Los estudiantes deberán seleccionar un dataset público o una API de su interés (preferiblemente relacionada con su área de estudio o trabajo). El objetivo es obtener datos de esta fuente, realizar un proceso de limpieza y exploración inicial, y luego aplicar transformaciones para obtener insights significativos

Fases I: Selección de la Fuente de Datos (Investigación Inicial):

Tarea: Identificar una fuente de datos accesible y de interés. Las opciones pueden incluir:

1. **APIs Públicas:** (Ejemplos: APIs de clima, datos gubernamentales abiertos, APIs de redes sociales, APIs de películas/series, APIs de finanzas, etc.). Se recomienda buscar en sitios como data.gov, kaggle.com/datasets, github.com/public-apis/public-apis.
2. **Datasets Públicos:** Archivos CSV, JSON, Excel disponibles en repositorios como Kaggle, UCI Machine Learning Repository, data.world, etc.

Consideraciones:

1. La fuente debe ser accesible públicamente (sin necesidad de claves API complejas o autenticación avanzada inicialmente, a menos que el estudiante se sienta cómodo explorándolo).
2. Preferiblemente, los datos deben tener cierta complejidad (más de un fichero, o una API con estructura anidada, o datos que requieran limpieza).
3. El estudiante debe poder justificar la elección de la fuente de datos y qué tipo de preguntas podría responder con ella.

Entregable de esta fase: Un párrafo breve describiendo la fuente de datos elegida, su URL, el tipo de datos (CSV, JSON, API) y qué se espera investigar.

Fase II: Adquisición de Datos

Tarea: Escribir el código Python (usando pandas, requests, o sqlite3 si es una base de datos SQL accesible) para descargar o acceder a los datos.

Requisitos:

1. Mostrar el código de carga/acceso.
2. Si es una API, indicar la URL y los parámetros utilizados.
3. Si es un archivo, indicar la URL o cómo acceder a él.

Fase III: Limpieza y Exploración Inicial (Análisis Exploratorio de Datos - EDA)

Tarea: Aplicar técnicas de limpieza y exploración para entender los datos.

Requisitos:

1. Mostrar las primeras filas (.head()).
2. Utilizar .info() para ver tipos de datos y valores no nulos.
3. Usar .describe() para estadísticas descriptivas.
4. Identificar y cuantificar valores nulos (.isnull().sum()).

Decisión de Limpieza: Los estudiantes debe tomar una decisión sobre cómo manejar los valores nulos (eliminar filas/columnas, imputar valores). **Justificar esta decisión**. Aplicar al menos **una técnica de limpieza** (ej. renombrar columnas, cambiar tipo de dato, imputar/eliminar nulos).

Fase IV: Transformación y Procesamiento de Datos

Tarea: Aplicar transformaciones para preparar los datos para un análisis más profundo o para obtener nuevos insights.

Requisitos: Demostrar el uso de al menos **dos** de las siguientes técnicas:

1. **Filtrado y Ordenamiento:** Seleccionar subconjuntos de datos basados en condiciones.
2. **Agrupamiento (groupby)**: Agrupar datos por una o más columnas y aplicar funciones de agregación (mean, sum, count, median, etc.).
3. **Creación de Nuevas Columnas:** Usar apply(), funciones lambda, o operaciones vectorizadas para crear nuevas variables (ej. ratios, categorías, fechas procesadas).
4. **Reorganización de Datos:** Utilizar pivot_table() o melt() para cambiar la estructura del DataFrame.
5. **Combinación de Datos (Opcional/Avanzado):** Si se usan múltiples fuentes o se extraen datos de una API anidada, intentar combinar o aplanar la información.

Fase V: Análisis y Presentación de Insights Preliminares

Tarea: Basándose en las transformaciones realizadas, extraer al menos un insight inicial.

Requisitos:

1. Responder a una pregunta simple sobre los datos, apoyándose en los resultados de las transformaciones.

   Ejemplo: "¿Cuál es el día de la semana con el gasto promedio más alto?", "¿Qué ciudad tiene la mayor cantidad de usuarios en la API?", "¿Cuál es la temperatura promedio máxima en la semana para Quito?".

2. Mostrar el resultado de manera clara (ej. una tabla, un texto explicativo).
3. Mostrar dos gráficos y la interpretación de estos.

Fase VI: Conclusiones del proceso y análisis de datos

Documentación y Presentación:

- **Entregable Final:** Un notebook (Jupyter o Google Colab) que contenga todo el proceso:
  - Título claro y nombres de los estudiantes.
  - Introducción breve sobre la fuente de datos y los objetivos del análisis.
  - Código ejecutable y comentado.
  - Salidas de las operaciones clave (tablas, información).
  - Explicaciones concisas de las decisiones tomadas (especialmente en la limpieza y transformación).
  - Respuesta a la pregunta del insight preliminar.
  - Conclusiones sobre lo aprendido y los hallazgos
- **Formato de Entrega:** Archivo .ipynb.