Análisis de datos - Trabajo práctico integrador. Parte 1

1. Introducción y motivación

Para este trabajo práctico deberán realizar el EDA completo de un set de datos y la idea es que trabajen en grupo.

1.1 Datasets disponibles

En la siguiente tabla tienen los datasets propuestos. Se espera que indaguen en las posibilidades de análisis y visualización según el dataset elegido.

Dataset
AirBnB Buenos Aires
Datos meteorológicos de Argentina
Encuesta mundial de salud escolar, Argentina 2018 (EMSE 2018)
Estadísticas sobre Ejecución de la Pena (SNEEP) (Elegir un año).
Precios Claros - Base SEPA: Elegir alguna de las cadenas grandes de supermercado (Carrefour, Disco, etc.)
Recorridos en Ecobicis (Elegir un año).
Crímenes reportados en Chicago (Elegir un año).
Denuncias a la policía de NY (para lo que va del 2025)
Uso de High-Volume For-Hire Services (HVFHS) en USA (Elegir un año).
Uso de taxis Yellow Cab en USA (Elegir un año).
Conflictos armados en ciudades (Cities and Armed Conflict Events, CACE)
Eventos de violencia organizada (UCDP Georeferenced Event Dataset, GED)
Full TMDB Movies Dataset 2024

1.2 Definición de grupos y elección de datasets

Sobre la conformación de grupos y elección de dataset:

- Los grupos pueden ser de 1, 2 o 3 personas.
- Un mismo dataset no puede ser elegido por más de dos grupos.

Completar esta planilla con los datos del grupo de trabajo y el dataset elegido antes de la clase 2: **3/7/2025**.

NOTA. Para el TP2 trabajarán en el mismo grupo y con el mismo dataset.

2. Consignas

El análisis debe abordar los siguientes aspectos:

- Exploración y comprensión de los datos:
  - Cargar el dataset proporcionado y realizar un análisis exploratorio de los datos.
  - Describir las características principales del dataset, incluyendo el número de observaciones, número de variables y tipos de datos.
  - Identificar patrones generales y distribuciones.
  - Identificar errores, outliers (anomalías), valores faltantes y su tipo (MCAR, MAR, MNAR).

- Aplicación de técnicas de visualización:
  - Utilizar técnicas de visualización adecuadas para ilustrar las principales características del dataset.
  - Asegurarse de que las visualizaciones sean claras, concisas y efectivas para comunicar la información.
  - Interpretar los resultados obtenidos a partir de las visualizaciones.

- Plantear un posible problema de ML supervisado a partir de los datos elegidos.
  - Describir el problema de clasificación o de regresión.
  - Definir la variable target.

3. Entrega

La entrega consistirá en entregar a través del campus una notebook con el análisis exploratorio.

Consideraciones:

- La entrega consiste en una única notebook, correctamente documentada y organizada.
- La notebook debe estar compartida en un repositorio GitHub público.
- Se realizará una única entrega por grupo, a través del campus virtual.
- En la entrega deberán incluir un archivo (txt, pdf) con el link al repositorio GitHub. No es necesario entregar material de apoyo, pdf, slides, zip, etc.; basta con el link. (**NOTA**: el campus obliga a entregar un adjunto. Si no adjuntan nada, la entrega puede fallar).
- La entrega estará habilitada desde el día jueves **10/07/2025 a las 19:00** hasta el día lunes **14/07/2025 a las 23:59** (hora Argentina). Pasada la ventana de tiempo, el campus cierra la posibilidad de entrega de forma automática. Ante cualquier eventualidad, contactar a las docentes.
- **IMPORTANTE!** Al finalizar, asegurarse que la actividad figure "**Entregada**".

4. Evaluación

Se evaluará:

- El entendimiento del dominio del dataset.
- El correcto planteo del problema de ML.
- La elección y aplicación de conceptos de visualización.
- La profundidad del EDA realizado.

ChatGPT y demás LLMs: usarlos con responsabilidad