Análisis de datos - Trabajo práctico integrador. Parte 1

1. Introducción y motivación

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

1.1 Datasets disponibles

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

Dataset	**Preguntas sugeridas**
AirBnB Buenos Aires	¿Cuál es la relación entre el precio de las propiedades y su tipo de alojamiento? ¿Cómo varía la disponibilidad de las propiedades según el número de camas y habitaciones? ¿Existen patrones de precios según la ubicación, por ejemplo, en barrios específicos? ¿Cómo influye la política de cancelación en la calificación de las propiedades? ¿Qué características (número de camas, tipo de propiedad, etc.) están más correlacionadas con la puntuación de los huéspedes? ¿Cómo varía la distribución de precios según el tipo de propiedad y el número de reseñas recibidas?
Datos meteorológicos de Argentina	¿Cómo varía la temperatura a lo largo de las estaciones del año? ¿En qué meses se observan los valores más extremos de temperatura, humedad, viento y precipitación? ¿Las tendencias y patrones observados varían entre estaciones meteorológicas? ¿Cómo se compara el último año con la media de los últimos 30 años? ¿Las condiciones actuales y el pronóstico están dentro de lo esperado para esta época del año según los datos históricos?
Encuesta mundial de salud escolar, Argentina 2018 (EMSE 2018)	¿Cómo se distribuyen los comportamientos de riesgo (como consumo de alcohol, tabaco, actividad física insuficiente) entre estudiantes de diferentes géneros? ¿Cómo afecta el nivel de apoyo parental percibido a la probabilidad de involucrarse en comportamientos de riesgo? ¿Qué sector de la población adolescentes no adhiere a las recomendaciones de consumo de frutas y verduras? ¿Existe una asociación entre el tiempo dedicado a actividades sedentarias (como uso de pantallas) y la percepción de bienestar emocional?

Estadísticas sobre Ejecución de la Pena (SNEEP) Elegir un año.	¿Existe una asociación entre el nivel educativo de las personas detenidas y su estructura familiar? ¿Cómo se relaciona la frecuencia de las visitas familiares con el estado de salud mental o los incidentes de las personas detenidas? ¿Cómo influye la reincidencia en la duración de las condenas? ¿Cómo varía la participación en programas de rehabilitación o en programas educativos según la situación legal?
Precios Claros - Base SEPA Elegir alguna de las cadenas grandes de supermercado (Carrefour, Disco, etc.)	¿Cuáles son los productos que más cambian de precio según el día de la semana? ¿Existe un patrón de precios por región? (ejemplo: regiones o grupos de provincias del país con precios muy elevados) ¿Las marcas propias de los supermercados tienen precios más bajos que las marcas líderes? ¿Los precios de los productos esenciales (ejemplo: leche, pan, arroz) varían dependiendo del día de la semana? ¿Cómo influye el tipo de comercio (hipermercado, autoservicio, etc.) en los precios de los productos?
Recorridos en Ecobicis Elegir un año.	¿Existen patrones temporales en la cantidad de viajes (horas pico, días de la semana, meses)? ¿Qué estaciones tienen mayor flujo de entradas y salidas? ¿Cómo se distribuyen espacialmente los viajes? ¿Se pueden agrupar por zonas según la demanda? ¿La dirección de los viajes sigue alguna tendencia (por ejemplo, más viajes hacia el centro en la mañana y hacia los barrios en la tarde)?
Burnout en empleados corporativos	¿Existen diferencias significativas en los niveles de agotamiento laboral entre empleados de diferentes tipos de compañías? ¿En qué medida el apoyo percibido de la organización reduce los niveles de agotamiento laboral entre los empleados? ¿Cómo influye el equilibrio entre el trabajo y la vida personal, y la cantidad de horas de descanso en los niveles de agotamiento laboral? ¿Cómo varía el nivel de burnout según la antigüedad de los empleados, el género, el puesto y la disponibilidad de arreglos para trabajar en forma remota?
Crímenes reportados en Chicago Elegir un año.	¿Cómo varía la distribución de los crímenes a lo largo de las horas del día, los días de la semana y los meses del año?¿Se observan anomalías y/o patrones estacionales? ¿Hay diferencias significativas entre el número de crímenes en distintos distritos o comunas? ¿Están las fuerzas policiales bien distribuidas en relación a las características de cada zona? (ej: la mayor cantidad de actividad policial/arrestos se registra en las zonas críticas) ¿Cómo variaron los crímenes en la ciudad después de algún cambio o evento social importante? (ej: Covid-19, protestas, etc.)
Denuncias a la policía de NY (para lo que va del 2025)	¿Cómo varían las tasas de criminalidad según la ubicación geográfica? ¿Existen puntos críticos para tipos de crímenes específicos? ¿Qué delitos son los más reportados? ¿Existe alguna relación entre el tipo de crimen y factores demográficos y/o socioeconómicos en ciertos vecindarios? ¿Cómo varían los delitos según el tipo de localización?

Incidentes de tráfico en Los Ángeles desde 2010 a la actualidad	¿Cuáles son los días y horas con mayor cantidad de colisiones de tránsito en Los Ángeles? ¿Cómo han evolucionado las colisiones de tránsito a lo largo del tiempo (anualmente, mensualmente, etc.)? ¿Cuáles son las áreas o distritos con mayor cantidad de incidentes reportados?¿Existen ubicaciones recurrentes donde mejorar la señalización, iluminación o infraestructura vial podría reducir la cantidad de colisiones? ¿Se puede identificar algún patrón (edad, grupo demográfico, peatón/ciclista, etc.) asociado a las víctimas?
Préstamos a negocios en Los Ángeles (recuperación post COVID)	¿Cuántos préstamos fueron aprobados en función de la localización geográfica del negocio? ¿Existen zonas dentro de la ciudad con alta concentración de préstamos? ¿Hay alguna evidencia de posible sesgo en la aprobación de los préstamos? ¿Se puede inferir algo acerca de la efectividad/rapidez del proceso? ¿Cuáles parecen ser los factores que más influyen en la aprobación?
Uso de High-Volume For-Hire Services (HVFHS) en USA Elegir un año.	¿Existen diferencias en el número de viajes entre distintos proveedores (Uber, Lyft, etc.)? ¿Cómo varía el número de pasajeros según la hora del día y el tipo de servicio? ¿Hay una relación entre la tarifa y el tiempo de espera del pasajero? ¿Cuáles son los recorridos más rentables en términos de tarifa por milla? ¿Se pueden identificar patrones espaciales en los viajes? (Ejemplo: viajes dentro de Manhattan vs. viajes a los aeropuertos) ¿Existen diferencias en la tarifa o duración de los viajes según el día de la semana?
Uso de taxis Yellow Cab en USA Elegir un año.	¿Cuáles son los patrones de demanda a lo largo del tiempo? ¿Cómo varía la distancia de los viajes según el tipo de servicio? ¿Cuáles son las zonas con mayor actividad de pick-ups y drop-offs? ¿Cómo varían las tarifas según la distancia y el tipo de servicio? ¿Se pueden detectar patrones de fraude o comportamiento anómalo en los datos de viajes? (ejemplo viajes con distancias extremadamente cortas pero tarifas altas)
Conflictos armados en ciudades (Cities and Armed Conflict Events, CACE)	¿En qué ciudades se concentran los eventos más letales? ¿Cómo ha cambiado el tipo de violencia en contextos urbanos a lo largo del tiempo? ¿Hay ciertos años o décadas donde los eventos urbanos crecieron abruptamente en algunas regiones? ¿Qué países registraron más eventos armados en ciudades durante el período de la Segunda Guerra Mundial?
Eventos de violencia organizada (UCDP Georeferenced Event Dataset, GED)	¿Qué actores aparecen más frecuentemente involucrados en conflictos y cómo ha cambiado su presencia en las últimas décadas? ¿Cuáles son los patrones más comunes en la ocurrencia de eventos violentos por región del mundo? ¿Existe una relación entre la duración de los conflictos y el número de actores involucrados o el tipo de violencia? ¿Hay diferencias significativas en el número de muertes entre diferentes tipos de conflicto?

Full TMDB Movies Dataset 2024	¿Cómo evolucionó la popularidad promedio de las películas a lo largo de las décadas? ¿Hay relación entre el idioma original y la puntuación promedio? ¿Existen géneros con baja calificación pero alta popularidad? ¿Qué combinación de géneros aparece con mayor frecuencia? ¿Cómo ha cambiado esa combinación a lo largo del tiempo?

1.2 Definición de grupos y elección de datasets

La conformación de grupos y elección de dataset debe respetar los siguiente ítems:

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

Completar esta planilla con los datos del grupo de trabajo y el dataset elegido antes de la clase 2: **8/5/2025**.

2. Consignas

El análisis debe abordar los siguientes aspectos:

- Planteo de al menos tres preguntas a ser respondidas mediante análisis de datos.
    - Se pueden usar como ejemplo las preguntas sugeridas, o proponer otras.

- 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, distribuciones y cualquier anomalía inicial en los datos.
    - Visualizar las variables más importantes para entender sus relaciones y distribuciones.

- 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.

- Limpieza del dataset:
    - Identificar y tratar los valores faltantes en el dataset.
    - Detectar y manejar los outliers utilizando técnicas estadísticas o visuales apropiadas.

    - Realizar una limpieza general del dataset, eliminando o corrigiendo datos inconsistentes o irrelevantes.

3. Entrega

Consideraciones:

- Se deberá hacer una entrega por grupo a través del campus.
- Deberán informar el link al repositorio GitHub con la notebook correctamente documentada y organizada.
- La entrega estará habilitada desde el día jueves **29/05/2025 a las 00:00** hasta el día lunes **02/06/2025 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.
- Al finalizar, asegurarse que la actividad figure “Entregada”.

4. Evaluación

Se evaluará:

- El entendimiento del dominio del dataset.
- La medida en la que el análisis responde a las preguntas planteadas.
- La elección y aplicación de conceptos de visualización.
- La comprensión de los temas vistos hasta la fecha:
    - Limpieza general, EDA.
    - Manejo de valores faltantes
    - Manejo de datos atípicos (outliers)
- La calidad de las explicaciones, conclusiones y justificaciones de los pasos realizados.