# Trabajo práctico integrador.
# Análisis de datos

## 1. Introducción y motivación

Les proponemos para este trabajo final realizar el análisis completo para un set de datos, para ello les vamos a proponer varios de estos y la idea es que ustedes elijan uno.

### 1.1 Datasets disponibles

Se proponen 9 datasets destinos. Sin embargo, si les interesa trabajar en otro porque les motiva/entusiasma más el tema se puede considerar. En tal caso, enviar un mail al docente con el dataset deseado, una explicación de porque les interesaría analizarlo, que desafíos le ven y que preguntas se hacen. Es importante que el dataset no sea trivial y que sea necesario realizar tareas de limpieza de datos, conversión, imputación de missings, etc.

1. Datos de distintas estaciones meteorológicas de Australia.
    - Preguntas interesantes para considerar aquí: El objetivo es predecir si lloverá o no al día siguiente (variable *RainTomorrow*), en función datos meteorológicos del día actual.
2. Datos de distintas canciones en Spotify.
    - Preguntas interesantes para considerar aquí: El objetivo aquí es poder estimar si un tema nuevo será del gusto de la persona que tiene esta playlist activa. En este caso la variable *label* corresponde a nuestra variable de salida a analizar.
3. Uso de taxis *Yellow Cab* en USA en el año 2020.
    - Preguntas interesantes para considerar aquí: (elija una o dos)
        - ¿Existe una manera de caracterizar los lugares más recurrentes para inicio/fin de viaje?
        - ¿Cómo son los viajes típicamente en distancia y tiempo?
        - ¿Podremos segmentar los viajes de alguna manera? (clusterización)
4. Dataset de Casos de COVID para análisis de dependencia con vacunas: Vacunación y Covid Muertes
    - Preguntas interesantes para considerar aquí:
        - ¿ Qué relación hay entre decesos por covid y campaña de vacunación?
        - ¿Como varía en base a las distintas vacunas, países, dosis administradas, fechas, etc?
5. Dataset de piezas creadas por año de LEGO.
    - Preguntas interesantes para considerar aquí:
        - ¿Cómo evolucionaron los sets de lego en tamaño a través de los años?
        - ¿Existe alguna asociación entre los colores y las temáticas?
        - ¿Podría predecir a que temática pertenece un set basado en el

contenido de este? (recomendado)
- A través de los años, ¿Cuál o cuáles son los sets que tienen las piezas más raras?
- ¿Cómo evolucionaron los colores en los sets de lego a través de los años?

6. Dataset de comidas varias (Elijan uno).
   - Starbucks dataset
   - McDonald's dataset
   - Preguntas interesantes para considerar aquí:
     - ¿Podemos clasificar de alguna manera los ítems de los menues?
     - ¿Se puede diseñar una manera interesante de mostrar estos datos?
     - ¿Se puede estimar las calorías de un item basado en sus propiedades?
7. MNIST dataset
   - Preguntas interesantes:
     - ¿Se pueden encontrar heurísticas interesantes para clasificar los datos en función de sus valores?
     - ¿Es posible encontrar representaciones de baja dimensionalidad que nos permitan visualizar posibles grupos?
8. Airline Passenger Satisfaction
   - Preguntas interesantes
     - ¿Cómo influyen variables como la edad y el género en la satisfacción con el vuelo?
     - ¿Depende de en que clase viajó el pasajero?
     - ¿Influye que sea un pasajero frecuente?
9. AirBnB Buenos Aires
   - Preguntas interesantes
     - ¿Como es la distribución por barrio?
     - ¿Habra alguna relación entre el source y los valores nulos?
     - Relación entre las características del host y el score/precio, ¿se podrá crear un score para el host? ¿tiene sentido hacerlo?
     - Relación entre las características del lugar y el score/precio, ¿se podrá crear un score para el lugar? ¿tiene sentido hacerlo?

## 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, 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.

- **Transformación de Variables y Selección de Features:**

    - Aplicar técnicas de conversión de variables donde sea necesario.
    - Aplicar técnicas de escalamiento de ser necesario
    - Realizar una selección de variables basándose en la relevancia y significancia de las mismas.
    - Justificar la elección de las variables seleccionadas.

- **Reducción de la Dimensionalidad:**

    - Aplicar técnicas de reducción de la dimensionalidad.
    - Comparar el dataset original con el dataset reducido, evaluando las ventajas y desventajas de la reducción.
    - Evaluar cómo la reducción de la dimensionalidad afecta la interpretación y el rendimiento de futuros modelos predictivos.

Deberá enviarse a través del campus la notebook correctamente documentada y organizada. La fecha límite es la noche anterior al día de la defensa.

## 3. Evaluación

- El TP se defenderá de forma oral la última clase. La presentación podrá ser mediante la notebook o con slides. Este último caso suma más puntos.
- Cada grupo tendrá 20 minutos para su exposición, pasarse mucho de este tiempo implica reducción de la nota final
- Es importante que todos los miembros del grupo expongan de igual manera en términos de tiempo
- Todos los miembros del grupo deberán tener la cámara prendida durante la presentación
- La nota final de la materia será en base al desempeño individual y grupal de este trabajo

    - Desempeño individual: Durante la presentación es importante que se note que el alumno trabajó y comprende en profundidad lo hecho por el grupo.

    - Desempeño grupal: El trabajo debe reflejar una buena comprensión de los temas vistos en la materia a la vez que muestre un análisis en profundidad del dataset seleccionado.

**Rubrica**

| Criterios | 1 - Necesita Mejora | 2 - Satisfactorio | 3 - Excelente |
|---|---|---|---|
| **Entendimiento del Dominio del Dataset** | Comprensión limitada del dataset; dificultades para explicar su contexto y relevancia. | Comprensión básica; explica el contexto y relevancia con cierta claridad. | Comprensión profunda; articula el contexto y la relevancia del dataset de manera clara y confiada. |
| **Limpieza de Datos** | Esfuerzo mínimo en la limpieza de datos; permanecen errores significativos. | Limpieza adecuada; se abordan algunos errores. | Limpieza exhaustiva; los datos están bien preparados y sin errores. |
| **Conversión de Variables** | Poca o ninguna conversión aplicada; métodos poco claros. | Se aplica cierta conversión; se demuestran métodos básicos. | Aplica eficazmente técnicas de conversión; los métodos están bien articulados. |
| **Manejo de Valores Faltantes** | Sin estrategias para valores faltantes; se compromete la integridad de los datos. | Se utilizan algunas estrategias; comprensión básica de la importancia. | Estrategias completas empleadas; se mantiene la integridad de los datos. |

| | | | |
|---|---|---|---|
| **Selección de Features** | Pobre selección de características; se incluyen características irrelevantes. | Proceso de selección básico; se identifican algunas características relevantes. | Selección de características reflexiva; todas las características son relevantes y justificadas. |
| **Manejo de Atípicos** | Sin consideración por los atípicos; se afecta la integridad de los datos. | Conciencia básica de los atípicos; algunos son abordados. | Identifica y maneja proactivamente los atípicos; mantiene la integridad de los datos. |
| **Visualización de Datos** | Visualizaciones ineficaces o poco claras; no apoyan el análisis. Pocos gráficos. | Se utilizan visualizaciones básicas pero correctamente; Apoyan el análisis. | Visualizaciones claras y relevantes que muestran o aportan patrones relevantes en los datos. Buena variedad de gráficos y correcta utilización de los colores. |
| **Reducción de Dimensionalidad** | Sin aplicación de reducción de dimensionalidad; impacto en el análisis poco claro. | Aplicación básica; cierta comprensión de su relevancia. | Aplicación de más de una reducción; Clara comprensión de sus resultados . |
| **Calidad de la Presentación** | Presentación desorganizada; Repetición de análisis; Falta de comentarios | Notebook bien organizada y documentada, sin redundancias. | Presentación bien estructurada; uso efectivo de slides. |
| **Presentación oral individual** | No comprende correctamente lo que se hizo; Intervención muy corta | Comprende lo hecho en el trabajo; Intervención adecuada | |