Trabajo Práctico 1 : Visualización de datos

Se espera que puedan hacer un análisis sobre algunas variables de los datasets y analizar relaciones entre entre ellas utilizando herramientas gráficas.
Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema. Por lo tanto, deben ser comprensibles por quien los vaya a leer. Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, etc.

El trabajo debe ser resuelto en una máquina virtual tipo júpiter en el sitio Kaggle, se debe entregar el enlace a la máquina correspondiente con los permisos de lectura y escritura necesarios para la evaluación.

Fecha máxima de entrega: Lunes 25 de abril.

Parte 1 - Entrada en calor

Dataset a utilizar

Seleccionar un set de datos de la fuente *FiveThirtyEight* (link), la cual brinda 80 datasets de distintos temas como seguridad aérea, consumo de alcohol, malos conductores, biografías, nacimientos, rock clásico, carreras universitarias, personajes-de-comics,etc.
El dataset elegido debe tener variables cualitativas y cuantitativas (tanto discretas como contínuas) y debe ser validado por el docente asignado. Recomendamos seleccionar más de uno ya que dos grupos no pueden utilizar el mismo.

Análisis exploratorio

Exploración Inicial

Realizar una descripción del dataset elegido detallando:

- Nombres y tipos de las columnas, y descripción de cada una.
- Resumen del dataset mostrando un conjunto reducido de filas (las primeras y las últimas)
- Cantidad de valores nulos por columna

Variables cualitativas

- Sobre las variables de este tipo mostrar los distintos valores existentes y la cantidad de filas correspondientes a cada uno.
    - Armar un nuevo *dataframe* con estos datos.
- Para cada variable compare en un gráfico de barras la cantidad de filas pertenecientes a cada valor.
    - ¿Es posible realizar este gráfico con un histograma? Explique las diferencias entre estos dos tipos de gráficos.

Variables cuantitativas

1) Sobre variables de este tipo calcular las siguientes medidas de resumen, y armar un nuevo dataframe con estos datos:
    - Media
    - Mediana
    - Moda
    - Primer y tercer cuartil
    - Rango

2) Correlación de atributos
    - Explorar las variables tomándolas de a pares utilizando un gráfico scatter_matrix
    - Calcular la correlación de Pearson y graficar las correlaciones obtenidas en un gráfico de tipo heatmap

3) Seleccionar un subconjunto de variables que resulten de interés y analizarlas utilizando:
    - Histogramas
    - Gráfico de violín (utilizar la librería matplotlib.pylot)
    - Boxplots
    - Density plot. Analizar si los datos presentan algún tipo de asimetría y explicar cuál.
    - Gráfico a elección. Seleccionar otro tipo de gráfico de una librería distinta a las propuestas en la materia y realizar un análisis a elección.

Conclusiones

Extraer conclusiones a partir de los análisis realizados en los puntos anteriores, y justificar cada conclusión.
Por ejemplo:

¿Se encontró algún tipo de comportamiento particular en alguna variable?
¿Existen variables correlacionadas y por qué? ¿Es esperable o no?

Parte 2 - Preguntas de investigación

Luego de la entrada en calor les proponemos trabajar con un nuevo set de datos (disponible para descargar en el siguiente link) y les planteamos los siguientes objetivos:

1.- Explorar el set de datos de forma libre aplicando las técnicas que consideren adecuadas (vistas en la materia)

2.- A partir de la exploración realizada plantear dos preguntas de investigación que les resulten interesantes para formular sobre los datos propuestos. Por ejemplo, ¿existe alguna relación entre el género musical y alguna de las otras variables?

3.- Elegir dos visualizaciones (como mínimo) que permitan abordar sus preguntas de investigación e interpretar los resultados.