Obtención de estadísticas descriptivas

Jorge Iván Sánchez González A01761414
Ana Camila Jiménez Mendoza A01174422
Gustavo José Ortiz Zepeda A01637220

1. Carga los datos usando tu lector de csv o con pandas. Es recomendable hacerlo con pandas.

Actividad Evaluable: Obtención de estadisticas descriptivas

En esta actividad trabajaras con el conjunto de datos asignado para el reto.

```
[ ]  # imports
     from sklearn import datasets
     import pandas as pd
     import numpy as np
     from google.colab import files
     import matplotlib.pyplot as plt
     import seaborn as sns
```

Importamos las librerías necesarias para poder correr el programa y también añadimos el archivo csv con los datos a trabajar

```
[ ]  # importar tabla
     df = pd.read_csv('datos_2021.csv',na_values=' ')
```

Fig. 1.1

2. Verifica la cantidad de datos que tienen, las variables que contiene cada vector de datos e identifica el tipo de variables.

```
print(df.shape)
print(df.columns)
obj_columns = df.select_dtypes(include=np.object).columns.tolist()
df[obj_columns] = df[obj_columns].astype('string')
print(df.info())
```

```
(8760, 21)
Index(['Estación SIMAJ', 'Fecha', 'Hora', 'O3', 'NO', 'NO2', 'NOX', 'SO2',
       'CO', 'PM10', 'PM2.5', 'TMPI', 'TMP', 'RH', 'WS', 'WD', 'PP', 'RS',
       'PBA', 'UV', 'UVI'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Estación SIMAJ  8760 non-null   string
 1   Fecha           8760 non-null   string
 2   Hora            8760 non-null   int64
 3   O3              5568 non-null   float64
 4   NO              3216 non-null   float64
 5   NO2             3216 non-null   float64
 6   NOX             3216 non-null   float64
 7   SO2             2232 non-null   float64
 8   CO              4296 non-null   float64
 9   PM10            2088 non-null   float64
 10  PM2.5           561 non-null    float64
 11  TMPI            7794 non-null   float64
 12  TMP             7416 non-null   float64
 13  RH              7440 non-null   float64
 14  WS              7440 non-null   float64
 15  WD              7416 non-null   string
 16  PP              7896 non-null   float64
 17  RS              0 non-null      float64
 18  PBA             0 non-null      float64
 19  UV              0 non-null      float64
 20  UVI             0 non-null      float64
dtypes: float64(17), int64(1), string(3)
```

Fig. 1.2

Las variables que se muestran proporcionadas por la base de datos en la *figura 1.2* son de tipo flotante o decimal (float64), entero (int64) y texto (string).

En total contamos con 21 distintas variables con 8760 datos, en la mayoría de casos las variables presentan ausencia de datos y de información por lo que posteriormente las iremos descartando.

3. Analiza las variables para saber qué representa cada una y en qué rangos se encuentran. Si la descripción del problema no te lo indica, utiliza el máximo y el mínimo para encontrarlo.

   Min - Max

   - Fecha: Representa el día
   - Hora: Representa la hora, 0 - 23
   - O3: Indica la cantidad de ozono, 0 - 0.139
   - NO: Indica la cantidad de óxido nítrico, 0 - 0.268
   - NO2: Indica la cantidad de dióxido de nitrógeno, 0 - 0.135
   - NOX: Indica la cantidad de óxidos de nitrógeno, 0 - 0.349
   - SO2: Indica la cantidad de dióxido de azufre, 0 - 0.0101
   - CO: Indica la cantidad de monóxido de carbono, 0 - 3.896
   - PM10: Indica la cantidad de partículas de 10µm de diámetro, 0 - 454.6
   - PM2.5: Indica la cantidad de partículas de 2.5µm de diámetro, 11.5 - 180.9
   - TMPI: temperatura media diaria (por sus siglas en inglés, Daily Mean Temperature). 0 - 44.9
   - RH: humedad relativa (por sus siglas en inglés, Relative Humidity). 0 - 95.2
   - WS: velocidad del viento (por sus siglas en inglés, Wind Speed). 0 - 13.58
   - WD: dirección del viento (por sus siglas en inglés, Wind Direction).
   - PP: precipitación (por sus siglas en inglés, Precipitation). 0 - 2.84
   - RS: radiación solar (por sus siglas en inglés, Solar Radiation).
   - PBA: presión barométrica (por sus siglas en inglés, Barometric Pressure).
   - UV: radiación ultravioleta (por sus siglas en inglés, Ultraviolet Radiation).
   - UVI: índice de radiación ultravioleta (por sus siglas en inglés, Ultraviolet Index).

```
[ ]  print(df.describe())
```

```
              Hora           O3           NO          NO2         NOX  \
count  8760.000000  5568.000000  3216.000000  3216.000000  3216.00000
mean     11.500000     0.028869     0.013967     0.021505     0.03547
std       6.922582     0.021738     0.023529     0.016476     0.03568
min       0.000000     0.000000     0.000000     0.000000     0.00000
25%       5.750000     0.011000     0.002000     0.011000     0.01400
50%      11.500000     0.026000     0.006000     0.019000     0.02600
75%      17.250000     0.042000     0.015000     0.030000     0.04600
max      23.000000     0.139000     0.268000     0.135000     0.34900

               SO2           CO         PM10        PM2.5         TMPI  \
count  2232.000000  4296.000000  2088.000000  561.000000  7794.000000
mean      0.001104     0.685496    45.375383   46.036542    23.199666
std       0.000984     0.432117    35.664791   21.425526     4.278118
min       0.000000     0.000000     0.000000   11.500000     0.000000
25%       0.000500     0.432000    22.500000   30.200000    21.600000
50%       0.000800     0.585000    38.400000   42.500000    22.200000
75%       0.001400     0.830250    59.525000   56.500000    22.800000
max       0.010100     3.896000   454.600000  180.900000    44.900000

               TMP           RH           WS           PP   RS  PBA   UV  UVI
count  7416.000000  7440.000000  7440.000000  7896.000000  0.0  0.0  0.0  0.0
mean     23.795577    47.760914     3.817481     0.010992  NaN  NaN  NaN  NaN
std       5.395954    23.517940     2.311717     0.126035  NaN  NaN  NaN  NaN
min       0.000000     0.000000     0.000000     0.000000  NaN  NaN  NaN  NaN
25%      20.300000    29.600000     2.090000     0.000000  NaN  NaN  NaN  NaN
50%      23.200000    45.100000     3.370000     0.000000  NaN  NaN  NaN  NaN
75%      27.725000    67.000000     5.110000     0.000000  NaN  NaN  NaN  NaN
max      37.600000    95.200000    13.580000     2.845000  NaN  NaN  NaN  NaN
```

Fig 1.3

4. Basándose en la media, mediana y desviación estándar de cada variable, qué conclusiones puedes entregar de los datos.

Basándonos en los rangos de la *figura 1.3*: podemos observar que hay datos como que en promedio, los niveles de contaminantes del aire como el ozono y el dióxido de nitrógeno son relativamente bajos. También los datos muestran poca variación de estos niveles promedio de contaminantes a lo largo del tiempo. Hay pico ocasionales en los niveles de contaminantes. Los niveles de contaminación a menudo cumplen (o están cerca) los estándares ambientales, es decir, una calidad de aire medianamente buena. Algunos contaminantes tienen consistentemente niveles promedio más altos que los demás lo que podría ser grave para el aire. La temperatura (TMP) y la humedad (RH) parece que influyen en los niveles de contaminantes, a las altas temperaturas y la baja humedad. Esto a su vez contribuye a la formación de contaminantes como el ozono. La velocidad del viento (WS) es importante para determinar el esparcimiento de los contaminantes.