Isaac Jacinto Ruiz A01658578

Profesor: Sergio Ruiz Loza

Fecha de entrega: Mayo 2022

Campus: CCM

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

```
import seaborn, matplotlib
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import math

# Cargando datos
data = pd.read_csv("covid19_tweets.csv")
```

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

```
# Cantidad de datos que se tienen
datosVariable = data.count()
datosTotales = sum(data.count())

print("\n ---> Cantidad de datos por variable")
print(datosVariable)
print("\n ---> Cantidad de datos que se tienen = ",datosTotales)
```

```
 ---> Cantidad de datos por variable
user_name           74436
user_location       59218
user_description    70079
user_created        74436
user_followers      74436
user_friends        74436
user_favourites     74436
user_verified       74436
date                74436
text                74436
hashtags            53002
source              74424
is_retweet          74436
dtype: int64

 ---> Cantidad de datos que se tienen =  926647
```

```
# Variables que contiene cada vector
def variablesVector(matriz):
    variables = data.columns.values
    for i in range(0, len(variables)):
        print(variables[i])

print(" \n ---> Variables que contiene cada vector")
variablesVector(data)
```

```
 ---> Variables que contiene cada vector
user_name
user_location
user_description
user_created
user_followers
user_friends
user_favourites
user_verified
date
text
hashtags
source
is_retweet
```

```
# Tipo de variables
tiposVariables = data.dtypes

print("\n --->Tipos de variables")
print(tiposVariables)
```

```
 --->Tipos de variables
user_name           object
user_location       object
user_description    object
user_created        object
user_followers       int64
user_friends         int64
user_favourites      int64
user_verified         bool
date                object
text                object
hashtags            object
source              object
is_retweet            bool
dtype: object
```

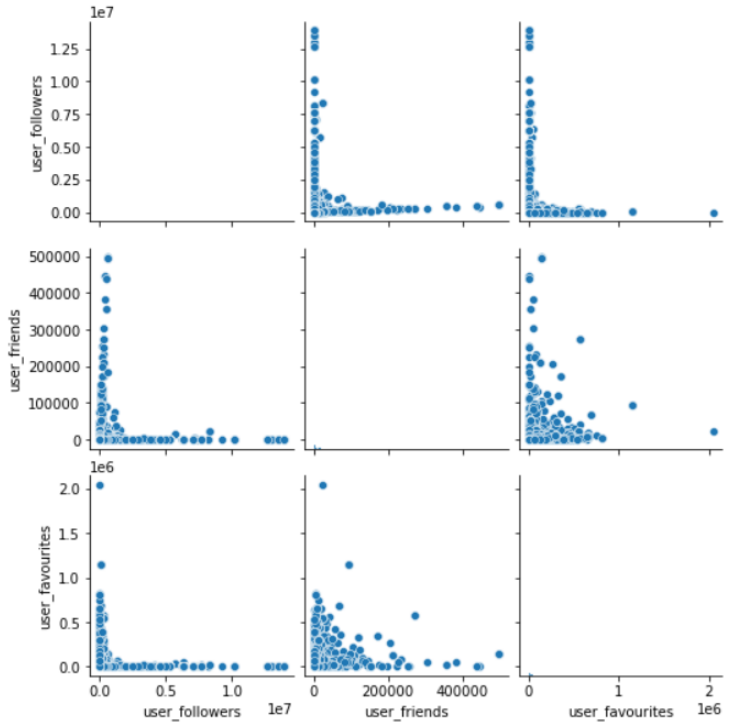
3. Analiza las variables para saber qué representa cada una y en qué rangos se encuentran. Si la descripción del problema no te lo indica, utiliza el máximo y el mínimo para encontrarlo.

En el análisis de las variables se observan 13 de ellas, en donde cada una maneja un tipo de dato diferente. Para el rango de las variables, solo trabajaremos con las únicas variables de tipo de dato numérico que son user_followers, user_friends y user_favourites.

user_name representa el nombre de usuario; user_location representa la locación de donde esta la cuenta; user_description muestra la información de presentación del perfil; user_created hace referencia a la fecha de creación de la cuenta; user_followers referencia a los seguidores que tiene la cuenta; user_friends: representa la cantidad de amigos que tiene la cuenta; user_favourites muestra la cantidad de Tweets marcados como favoritos por parte del usuario; user_verified muestra si el usuario está verificado o no; date es la fecha del tweet; text es el tweet; hashtag es la etiqueta que usó el tweet para un bloque de tweets o etiqueta; source sería el dispositivo donde fue publicado el tweet y por ultimo is_retweet si el tweet es propio o se ha compartido de otra cuenta.

```
#Correlacion
new_data = data[['user_followers','user_friends','user_favourites']]
sns.pairplot(new_data)
plt.show()
```

```
# Rangos en los que se encuentran las variables
print("\n ---> Rango de user_followers = de ",data["user_followers"].min(skipna=0), "hasta",data["user_followers"].max())
print("---> Rango de user_friends = de ",data["user_friends"].min(), "hasta",data["user_friends"].max())
print("---> Rango de user_favourites = de ",data["user_favourites"].min(), "hasta",data["user_favourites"].max())
```



En la gráfica se observa la comparación de las variables, en donde se observa que hay usuarios con actividad rara, ya que se observa una dispersión en ciertos sectores. Analizando, nos damos cuenta que la información y los datos son inusuales, ya que por ejemplo, en user_favourites vs user_friends, la cantidad de "me gustas" es demasiada para la cantidad de personas que sigue.

```
 ---> Rango de user_followers = de  0 hasta 13892841
---> Rango de user_friends = de  0 hasta 497363
---> Rango de user_favourites = de  0 hasta 2047197
```

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

```
# Obtención de media, mediana y desviación estándar
estadistica = data.describe()
print("\n --->Media, Mediana y Desviación estándar")
print(estadistica)
```

```
 --->Media, Mediana y Desviación estándar
       user_followers   user_friends  user_favourites
count    7.443600e+04   74436.000000     7.443600e+04
mean     1.059513e+05    2154.721170     1.529747e+04
std      8.222900e+05    9365.587474     4.668971e+04
min      0.000000e+00       0.000000     0.000000e+00
25%      1.660000e+02     153.000000     2.200000e+02
50%      9.600000e+02     552.000000     1.927000e+03
75%      5.148000e+03    1780.250000     1.014800e+04
max      1.389284e+07  497363.000000     2.047197e+06
```

Media → `mean  1.059513e+05  2154.721170  1.529747e+04`

Mediana → `50%  9.600000e+02  552.000000  1.927000e+03`

Desviación estándar → `std  8.222900e+05  9365.587474  4.668971e+04`

Analizando los datos se puede concluir que durante el brote de covid19, se tuvo un promedio de 105,951 seguidores en las cuentas dedicadas a enviar información. que a su vez varía en un aproximado de 822,290 seguidores por cuenta.
Cada cuenta tiene aproximadamente 2,154 amigos con una variación de 552 amigos por cuenta; también se tiene un aproximado de 15,297 tweets como favoritos con una variación de 1927 favoritos.
En cuanto a la mediana, se puede determinar que la mitad de los usuarios de Twitter tenía una cantidad de seguidores durante este periodo de pandemia igual o por debajo de los 960 y la otra mitad igual o por encima de esta cantidad. A su vez, una parte de la misma división de usuarios contaba con una cantidad de amigos agregados menor o igual a 522 y la otra parte mayor o igual a dicha cantidad.
Por otro lado, en cuanto a Tweets favoritos se puede determinar que la mitad de los usuarios destacó una cantidad igual o menor a 1927 y la otra mitad una cantidad igual o mayor a la mencionada.