Actividad Evaluable: Obtención de estadísticas descriptivas
Nombre: Marco Uriel Pérez Gutiérrez
Matrícula: A01660337
Repositorio : https://github.com/MarcoPerez16/SemanaTecAnalitica.git
Archivo: Actividad Evaluable 2
Ubicada en carpeta “Dia 2”

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

In [15]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

data=pd.read_csv("semanaTec_Analitica/arte-de-analitica/covid19_tweets.csv")
data.head()
```

Out[15]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	
0	ᏉᎥ☻լꂅᏖ	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21	If I smelled the scent of hand sanitizers toda...	
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it...	
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14	@diane3443 @wdunlap @realDonaldTrump Trump nev...	
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10	@brookbanktv The one gift #COVID19 has give me...	
4	DIPR-J&K	Jammu and Kashmir	🖊Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08	25 July : Media Bulletin on Novel #CoronaVirus...	['CoronaVi

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

In [16]:
```
#Cantidad de usarios
print("Cantidad de usuarios:")
print(len(data.index))
#Variables
print("Variables:")
print(data.columns.values)
#Tipo de variables
print("Tipo de variables:")
data.dtypes
```

```
Cantidad de usuarios:
74436
Variables:
['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:
```

Out[16]:
```
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.

```
data_num.min()
```

```
user_followers     0
user_friends       0
user_favourites    0
dtype: int64
```

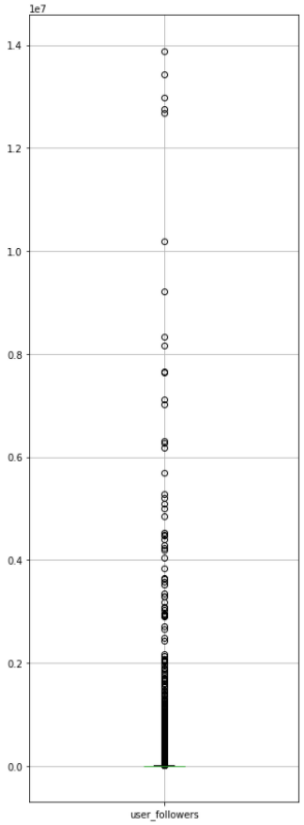
```
data_num.max()
```

```
user_followers     13892841
user_friends         497363
user_favourites     2047197
dtype: int64
```

In [76]:
```
d1=data.groupby(['user_name']).mean().sort_values(['user_followers'], ascending = False).groupby("user_name").head(10)
dU = d1[['user_followers']]
dU.head()
```

Out[76]:

user_name	user_followers
CGTN	1.389003e+07
NDTV	1.343905e+07
The Times Of India	1.298272e+07
United Nations	1.275416e+07
China Xinhua News	1.268052e+07



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

```
data_num= data[["user_followers","user_friends","user_favourites"]]
data_num.head()
```

	user_followers	user_friends	user_favourites
0	624	950	18775
1	2253	1677	24
2	9275	9525	7254
3	197	987	1488
4	101009	168	101

```
data_num.describe()
```

	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

Conclusiones:

Podemos observar que hay un total de 74,436 de observaciones en nuestro archivo .csv. Después, veamos la media aritmética o promedio, la cual es la suma de todos los valores dividida entre el número de valores, por lo cual, podemos ver que el promedio de los followers de los tweets es 105,951 , el promedio de amigos de los usuarios que publicaron tweets es 2154.72 y, por último, el promedio de likes de todos los tweets es 15,297. Al saber esto nos damos cuenta que en la mayoría de tweets existen bastantes reacciones que son los likes, y los usuarios de los tweets tienen gran cantidad de followers y amigos. Esto no es tan preciso de concluir, ya que puede ver un caso de que 100 tweets contengan una cantidad de followers, amigos y likes, y el resto de los tweets no tengan muchos, o hasta 0, como el mínimo encontrado.

Exactamente con la desviación estándar podemos observar y determinar qué fenómeno está ocurriendo, una desviación estándar baja nos diría que la mayor parte de los datos tienden a estar agrupados cerca de su media, o sea el promedio que ya vimos . Por el contrario, una desviación estándar con valor alto nos diría que los datos se extienden sobre un rango de valores más amplio, por lo tanto no se acercan tanto a la media, sino que varían, ya sea muy alejados o relativamente cercanos.

Las desviaciones estándar de los 3 valores numéricos de nuestros datos(followers, amigos y likes) son altas, por lo tanto podemos determinar que si existe variaciones en los valores y si están ocurriendo casos como tweets que tienen poca reacciones y followers, pero también hay casos de tweets con enormes reacciones e interacciones. Esto lo podemos observar claramente directo de los datos o en la gráfica que creamos en nuestro notebook,

donde vemos que existen tweets con demasiados followers, pero a la vez existen varios tweets con pequeñas cantidades de followers.