Jorge Gerardo Iglesias Ortiz - A01653261

Actividad Evaluable: Obtención de estadísticas descriptivas

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

In [1]:

```python
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt

df = pd.read_csv('covid19_tweets.csv')
```

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

In [2]:

```python
print(df)
```

```
                     user_name        user_location  \
0                      ᏉᎥ☻լꂅꋊ           astroworld
1                Tom Basile 🇺🇸        New York, NY
2              Time4fisticuffs     Pewee Valley, KY
3                  ethel mertz  Stuck in the Middle
4                     DIPR-J&K    Jammu and Kashmir
...                        ...                  ...
74431           Laura Wolfrom        Lexington, KY
74432  Professor Tonya M. Evans       😷 #stayathome
74433        People's Daily app        北京, 中华人民共和国
74434                    M0ser                  NaN
74435       Your Friend & Sabre ⚔️        Chicago, IL

                                        user_description         user_created  \
0      wednesday addams as a disney princess keepin i...  2017-05-26 05:46:42
1      Husband, Father, Columnist & Commentator. Auth...  2009-04-16 20:06:23
2      #Christian #Catholic #Conservative #Reagan #Re...  2009-02-28 18:57:41
3      #Browns #Indians #ClevelandProud #[]_[] #Cavs ...  2019-03-07 01:45:06
4      🖊️Official Twitter handle of Department of Inf...  2017-02-12 06:45:15
...                                                  ...                  ...
74431              The only things I collect are memories.  2010-09-24 02:01:15
74432  Law Prof @DickinsonLaw & Entrepreneur | Crypto...  2013-05-14 20:15:24
74433  Our mission is to provide news and perspective...  2018-02-04 12:36:42
74434  Reagan conservative and attorney raised in the...  2014-02-18 03:46:28
74435  My spectral decomposition has a significant da...  2016-12-19 19:55:00

       user_followers  user_friends  user_favourites  user_verified  \
0                 624           950            18775          False
1                2253          1677               24           True
2                9275          9525             7254          False
3                 197           987             1488          False
4              101009           168              101          False
...               ...           ...              ...            ...
74431              85           586             1902          False
74432            4289          1066            53569          False
74433            1413           102               16          False
74434            2554          1733           129104          False
74435             310          1748            60133          False
```

```
                      date                                               text  \
0      2020-07-25 12:27:21  If I smelled the scent of hand sanitizers toda...
1      2020-07-25 12:27:17  Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2      2020-07-25 12:27:14  @diane3443 @wdunlap @realDonaldTrump Trump nev...
3      2020-07-25 12:27:10  @brookbanktv The one gift #COVID19 has give me...
4      2020-07-25 12:27:08  25 July : Media Bulletin on Novel #CoronaVirus...
...                    ...                                                ...
74431  2020-08-04 03:13:29  So far this summer I have filled up my lawn mo...
74432  2020-08-04 03:13:26  ICYMI: REPLAY: #TechIntersect #16: Isaiah "@B...
74433  2020-08-04 03:13:22  Community workers in Tianshan District of Urum...
74434  2020-08-04 03:13:19  If only we had a responsible media to warn us ...
74435  2020-08-04 03:13:15  MAGA: #COVID19 is just a cold &amp;amp; it'd be go...

                                     hashtags              source  \
0                                         NaN   Twitter for iPhone
1                                         NaN  Twitter for Android
2                                ['COVID19']  Twitter for Android
3                                ['COVID19']   Twitter for iPhone
4      ['CoronaVirusUpdates', 'COVID19']  Twitter for Android
...                                     ...                 ...
74431            ['COVID19', 'QuarantineLife']   Twitter for iPhone
74432  ['TechIntersect', 'Bitcoin', 'COVID19']      Twitter Web App
74433                   ['China', 'Xinjiang']      Twitter Web App
74434                            ['COVID19']   Twitter for iPhone
74435       ['COVID19', 'Hydroxycholoroquine']  Twitter for Android

       is_retweet
0           False
1           False
2           False
3           False
4           False
...           ...
74431       False
74432       False
74433       False
74434       False
74435       False

[74436 rows x 13 columns]
```

Se tienen 74436 objetos con 13 variables siendo cada una:

1. user_name: Nombre de usuario en string
2. user_location: Lugar donde vive el usuario en string
3. user_description: Descripción del usuario en string
4. user_created: Fecha de la creación de la cuenta en tipo de dato de fecha
5. user_followers: Cantidad de usuarios que lo siguen en entero
6. user_friends: Cantidad de usuarios que son amigos en entero
7. user_favourites: Cantidad de favoritos dados en entero
8. user_verified: Estado de verificación del usuario en booleano
9. date: Fecha cuando se publico el tweet en tipo de dato de fecha
10. text: Texto contenido en el tweet en string
11. hashtags: Hashtags usados en el tweet como lista de strings
12. source: En que dispositivo fue publicado el tweet en string
13. is_retweet: Si el tweet es de otro tweet en retweet en booleano

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.

In [3]:
```
df.describe()
```

Out[3]:

	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

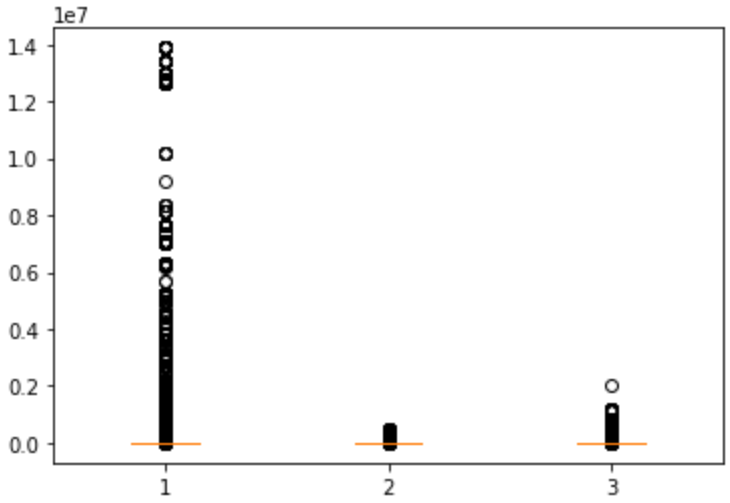
In [4]:
```
df.sort_values(['user_followers'], ascending = False).head(10)
```

Out[4]:

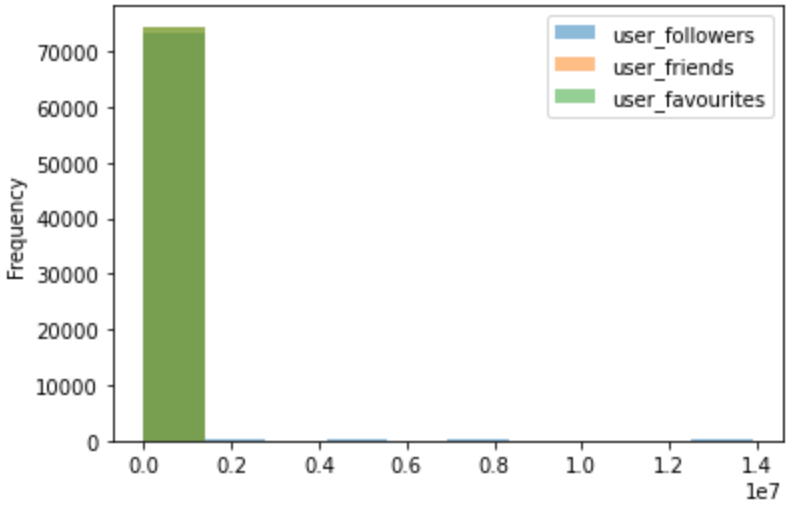
	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_
6959	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892841	69	104	
13450	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892839	69	104	
16194	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892837	69	104	
235	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892795	69	104	
2837	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892793	69	104	
5344	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892792	69	104	
20483	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	
20378	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	
24243	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	
23721	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	

In [5]:
```
plt.boxplot(df[["user_followers","user_friends","user_favourites"]])

plt.show()
```



In [6]:
```
ax = df[["user_followers","user_friends","user_favourites"]].plot.hist(bins=10, alpha=0.5)
```



In [7]:
```
df[["user_followers","user_friends","user_favourites","user_verified"]].corr(method ='pear
```

Out[7]:

	user_followers	user_friends	user_favourites	user_verified
user_followers	1.000000	-0.002722	-0.028724	0.322896
user_friends	-0.002722	1.000000	0.207825	0.013099
user_favourites	-0.028724	0.207825	1.000000	-0.060316
user_verified	0.322896	0.013099	-0.060316	1.000000

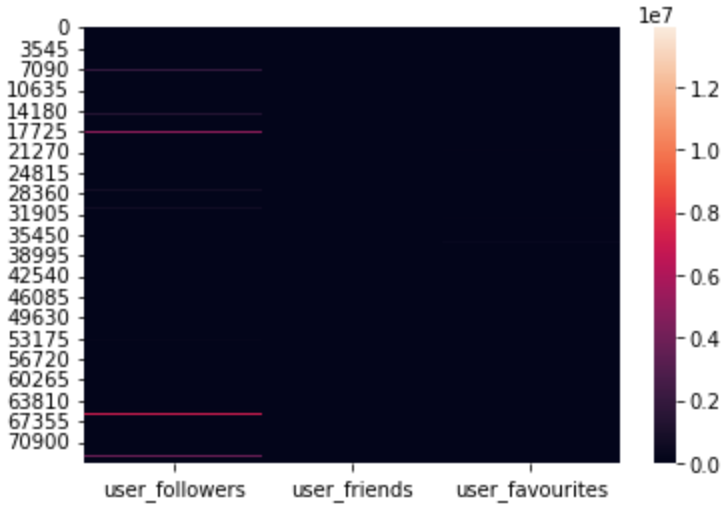
In [8]:
```
df[["user_followers","user_friends","user_favourites","user_verified"]].corr(method ='kend
```

Out[8]:

	user_followers	user_friends	user_favourites	user_verified
user_followers	1.000000	0.410663	0.242534	0.415364
user_friends	0.410663	1.000000	0.439099	-0.003114
user_favourites	0.242534	0.439099	1.000000	-0.032220

	user_followers	user_friends	user_favourites	user_verified
user_verified	0.415364	-0.003114	-0.032220	1.000000

In [9]:
```
ax = sb.heatmap(df[["user_followers","user_friends","user_favourites"]])
```



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

A partir del análisis en los valores cuantificables, se encuentra que, para la cantidad de seguidores por usuario, se observa que la media es de 105951 seguidores, pero su desviación es de 822289 seguidores, significando que algunos usuarios pueden tener desde 0 seguidores (siendo que no es posible tener una cantidad de seguidores negativo) hasta casi un millón de seguidores. De esta misma medida se sabe que la mediana es de 960 seguidores, siendo que la distribución de los usuarios en esta parte es que haya más usuarios con muy pocos o ningún seguidor mientras que muy pocos usuarios tengan varios seguidores de hasta a casi un millón.

Para la parte de la cantidad de amigos en Twitter, la media es de 2154 amigos, pero su desviación es de 9366 amigos, significando como en la anterior que hay usuarios pueden tener desde 0 amigos hasta más de diez mil amigos. De esta misma medida tenemos que la mediana es de 552, siendo que una mitad tiene una frecuencia de tener menos de 500 amigos, mientras que la otra mitad tiende a tener más de 500 amigos.

Finalmente, para la cantidad de favoritos por usuario hay una media de 15297 favoritos, pero su desviación es 46690 favoritos, siendo que algunos usuarios no tienen ningún favorito mientras que otros tienen más de 60000 favoritos. Para la mediana tenemos 1927 favoritos, significando que una mitad de los usuarios han tenido 2000 menos de favoritos y la otra tiene más de 2000 favoritos.

En otros análisis tenemos que aunque en un análisis de correlación no muestra altos grados de confiabilidad que estén correlacionadas las variables de la cantidad de seguidores, amigos, favoritos y verificación, se tiene otro que si muestra un grado un poco mayor alrededor del 42%, donde la cantidad de usuarios esta más o menos relacionado con la cantidad de amigos y la verificación, siendo que usuarios con mayor cantidad de seguidores tenga también una cantidad de amigos mayor y estén verificados. Para la cantidad de favoritos si está relacionado con la cantidad de amigos, siendo que un usuario con una mayor cantidad de amigos también tenga una mayor cantidad de favoritos. De cualquier manera, no se puede hacer fiables estas conclusiones porque el porcentaje de correlación es muy bajo como para poder aprobarlos bajo un nivel de confiabilidad del 95%.

Por último, se intentó sacar los diagramas de cajas y el mapa de calor, sin embargo, como los datos están muy apartados e inestables en general para las secciones de seguidores, amigos y favoritos, realmente no se puede observar una forma de análisis concluyente que pueda tener sentido para este caso y por lo tanto no se puede hacer alguna correlación de interés.