

Actividad Evaluable: Obtención de estadísticas descriptivas

Santiago Orozco Quintero A01658308

La primera observación importante que se hizo sobre los datos proporcionados fue que el apartado de `user_friends` se refiere muy probablemente a los mutual followers que tiene una cuenta, es decir a toda cuenta seguida por ellos que a la vez los siguen a ellos, esto se vio por la forma en cómo estaba distribuido este dato entre los distintos usuarios y prestando atención a los otros datos presentes en las tablas.

Los siguientes datos nos dejan ver algunas realidades con respecto de los tweets, en primera como el método describe nos muestra que algunos usuarios que twitearon no tienen ningún amigo o ni ningún dato de favoritos, esto nos puede indicar que esos usuarios son nuevos y no han tenido ninguna actividad o, lo más probable, que son bots utilizados para aumentar las interacciones en el tópico o para empujar otra narrativa.

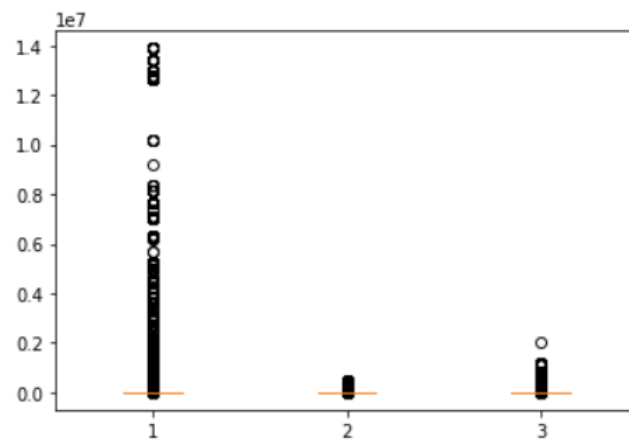
Otro dato muy interesante que se puede apreciar es que existe una relación directa entre estar verificado y tener un mayor número de seguidores. Esto puede deberse a un par de factores como por ejemplo el hecho de que muchos usuarios verificados son personas famosas o de la vida pública por lo que tienen mayor notoriedad y por ende atraer más suscriptores, del mismo modo puede ser que debido a lo tardado que puede ser el proceso para verificarse estos usuarios simplemente lleven más tiempo en la plataforma.

Las tablas de correlaciones también nos dejaron ver un par de cuestiones muy interesantes, entre ellas el hecho de que la mayoría de los usuarios verificados tienen una mayoría de amigos que no están verificados, esto tiene sentido si pensamos que no toda persona puede, o le interesa, verificarse, por lo mismo los usuarios registrados son una minoría en la plataforma. Observando otro dato de la correlación podemos apreciar que la mayoría de los usuarios siguen cuentas que no son sus amigos, o dicho de otra forma la mayoría de los seguidores de un usuario no son sus amigos.

Con las gráficas podemos apreciar primeramente que los datos es un poco complicado relacionarlos en bruto pero aun así podemos obtener información de ellos si se hace la relación de la manera correcta, por ejemplo con la gráfica de tipo box podemos notar que los followers varían mucho entre algunos usuarios donde se aprecia una gran disparidad en la distribución de estos mismos en la plataforma pero que así mismo una gran cantidad de usuarios coinciden en la cantidad de followers que tienen, es decir si bien hay usuarios con muchos seguidores y otros con muy pocos, en general se pueden apreciar tendencias estándar entre la distribución de seguidores. En esta misma gráfica podemos ver que los amigos son más escasos que los seguidores y que a su vez el rango entre estos datos varía menos, esto tiene sentido con la forma en cómo opera la plataforma donde la mayoría de las personas siguen más personas de las que reciben follower mutuo, otro dato que podemos apreciar observando este mismo campo es que los amigos deben siempre ser igual o menores que el número de user followers debió a que no tendría sentido que una persona tenga más followers mutuos que las personas que siguen.

Finalmente cuando vemos las cuentas más prominentes que hablan sobre el tema son de China lo cual hace sentido por ser el país donde se originó la pandemia, hablando de países podemos apreciar que el país que más twiteo fue India al ser de acuerdo con la moda el nombre de ubicación que más se repite.

```
In [7]: plt.boxplot(df[["user_followers","user_friends","user_favourites"]])
plt.show()
```



```
In [8]: df.corr(method = 'kendall')
```

Out[8]:

	user_followers	user_friends	user_favourites	user_verified	is_retweet
user_followers	1.000000	0.410663	0.242534	0.415364	NaN
user_friends	0.410663	1.000000	0.439099	-0.003114	NaN
user_favourites	0.242534	0.439099	1.000000	-0.032220	NaN
user_verified	0.415364	-0.003114	-0.032220	1.000000	NaN
is_retweet	NaN	NaN	NaN	NaN	1.0

```
In [9]: df.corr(method='pearson')
```

```
Out[9]:
```

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

```
In [11]: import seaborn as sns; sns.set_theme()
```

```
ax = sns.heatmap(df[["user_followers","user_friends","user_favourites"]])
```



```
In [15]: df.sort_values(['user_followers'],ascending=False).head(10)
```

```
Out[15]:
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text
6959	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892841	69	104	True	2020-07-25 08:00:00	#VoicesfromBeltandRoad: #COVID19 rap song aler...
13450	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892839	69	104	True	2020-07-25 02:43:57	#China's civil aviation recovers as daily flig...
16194	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892837	69	104	True	2020-07-25 00:27:38	On Friday, the #Chinese mainland reported:\n \...
235	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892795	69	104	True	2020-07-25 12:20:00	#APEC reaffirms #COVID19 economic recovery pri...
2837	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892793	69	104	True	2020-07-25 10:46:35	#COVID19 recovery can take weeks even for youn...
5344	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892792	69	104	True	2020-07-25 09:03:28	#HongKong reports 133 new confirmed #COVID19 c...
20483	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	True	2020-07-26 06:40:00	#COVID19 #HongKong SAR government introduces n...
20378	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	True	2020-07-26 06:46:36	Global #COVID19 cases have surpassed 16 millio...
24243	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	True	2020-07-26 02:31:57	Live: Students and recent grads from China and...
23721	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	104	True	2020-07-26 03:04:23	#DPRK sees 1st suspected #COVID19 case, adopts...

```
In [20]: df["user_location"].mode()[0]
```

```
Out[20]: 'India'
```