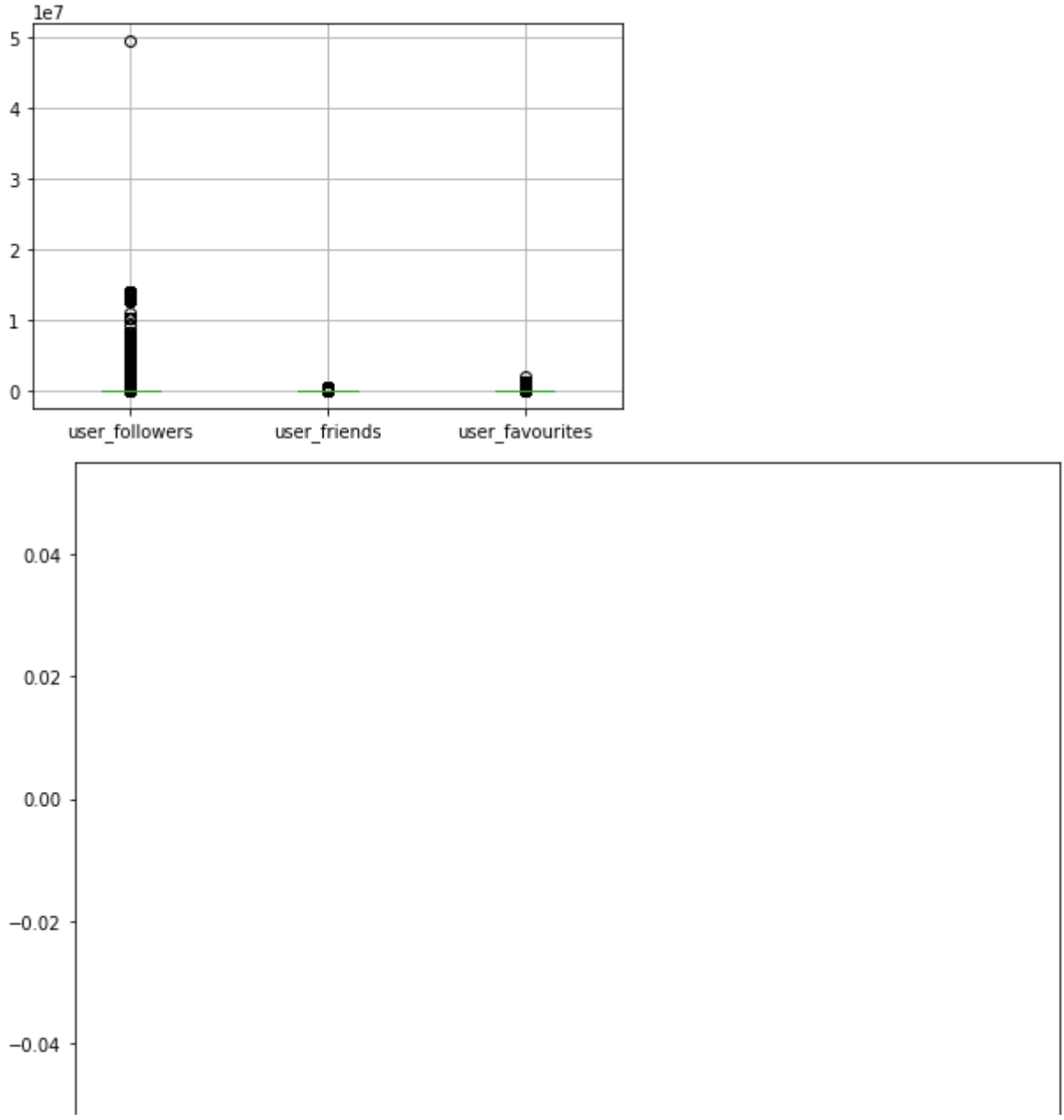
Equipo 4

Objetivo: Visualizar las diferencias entre followers que tienen los usuarios y los dispositivos con los que se hacen los Tweets (**sources**). Asímismo, identificar las relaciones entre si un usuario está verificado o no.

```
"""
Actividad Evaluable 3:  Mapas de calor y boxplots
"""
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
df = pd.read_csv('/content/sample_data/covid19_tweets.csv')
```

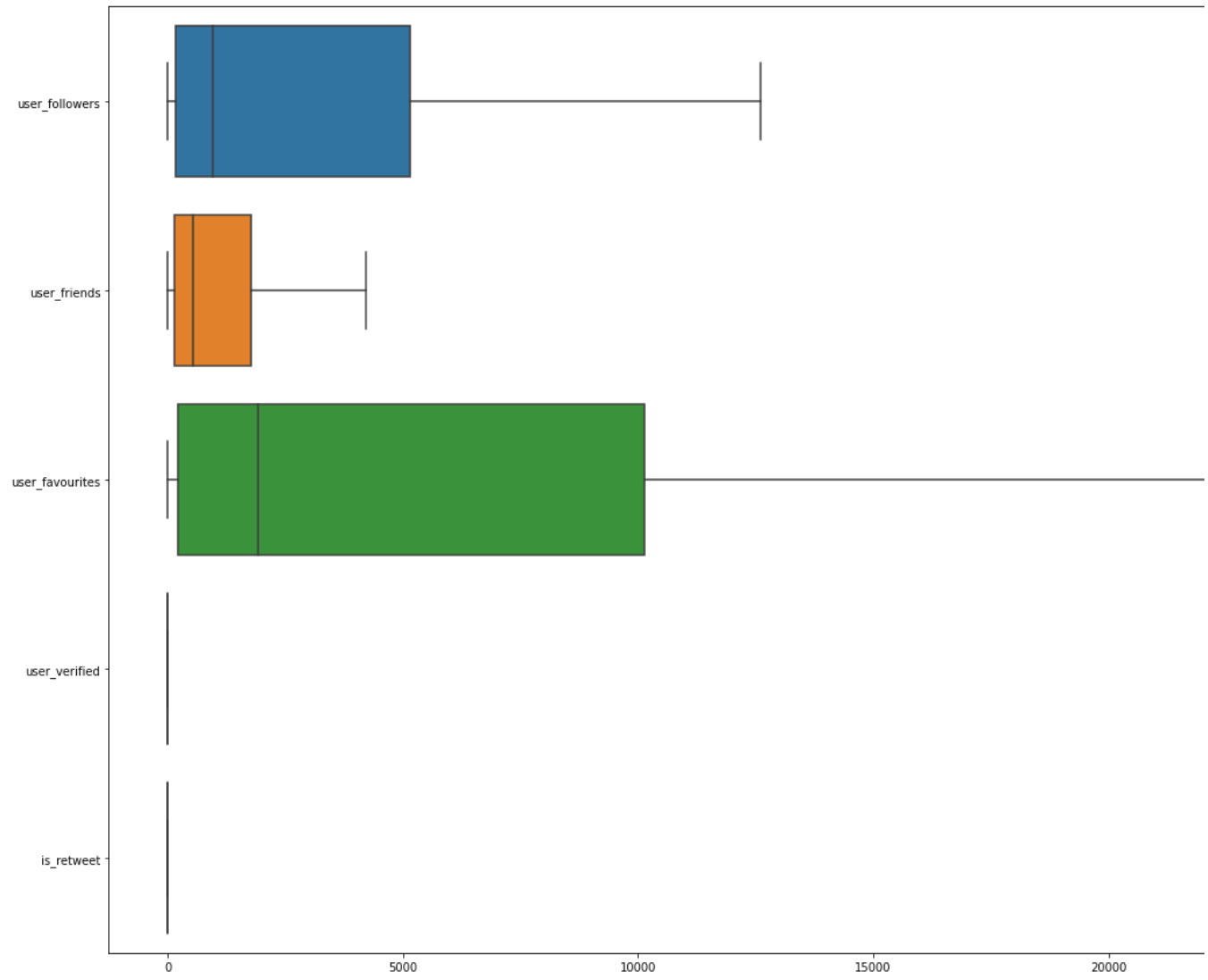
```
<ipython-input-22-bad2b88e6f6f>:8: DtypeWarning: Columns (7,12) have mixed types. Specif
  df = pd.read_csv('/content/sample_data/covid19_tweets.csv')
```

```
"""Diagrama de cajas y bigotes"""
df.boxplot()
#np.random.seed(10)
#data = np.random.normal(100, 20, 200)
data = df["user_followers"]
fig = plt.figure(figsize =(10, 7))
# Creating plot
plt.boxplot(data)
# show plot
plt.show()
sns.boxplot(data=df, orient="h")
```



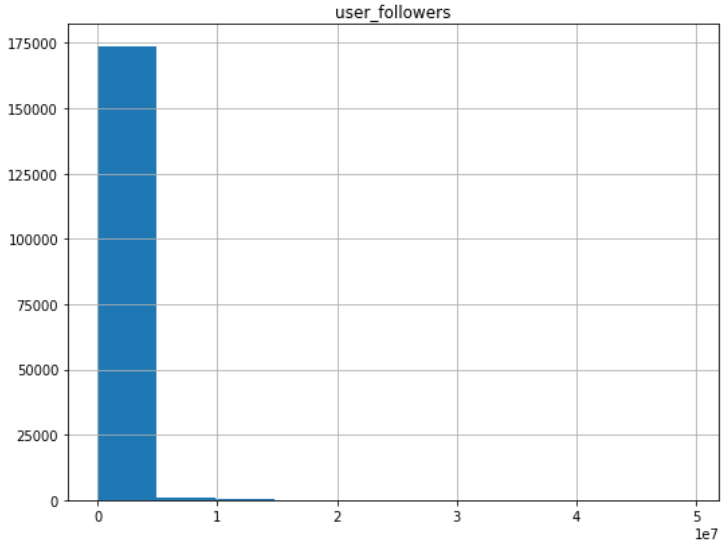

```
plt.figure(figsize =(20, 15))
sns.boxplot(data=df, orient="h",showfliers = False)
```

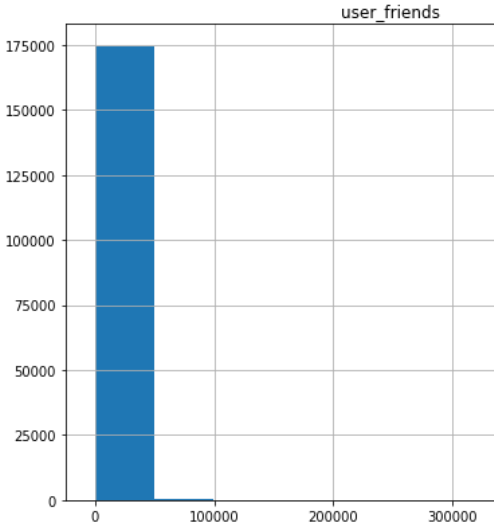
```
<Axes: >
```

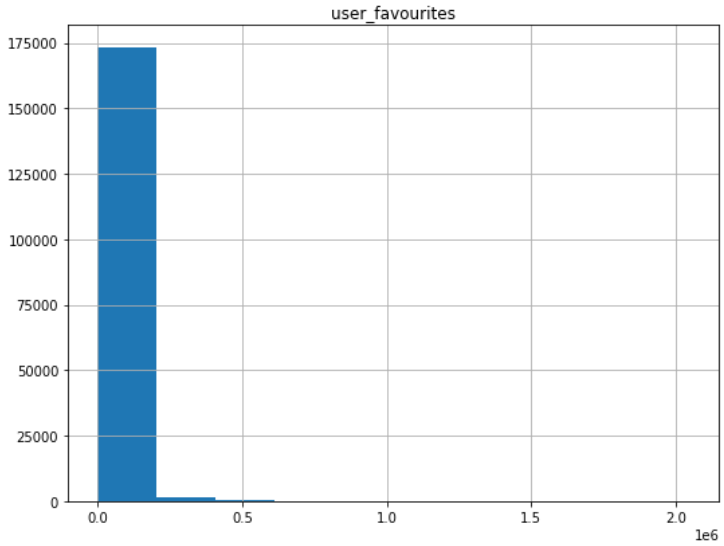



```
"""Histogramas"""

#df.hist(column=['user_followers', 'user_friends', 'user_favourites'], bins=20, figsize=(10,8
plt.show()
df.hist(figsize=(20,15))
plt.show()
```







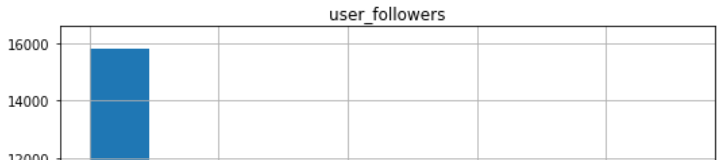

```
"""Mapas de calor"""
corrmat = df.corr()
plt.figure(figsize =(20, 15))
sns.heatmap(corrmat, annot=True)
plt.show()
```

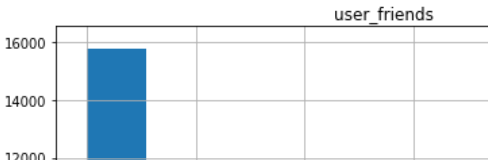

user_followers
user_friends
user_favourites
user_verified
is_retweet
1
-0.0027
-0.029
0.32
-0.0027
1
0.21
0.013
-0.029
0.21
1
-0.06
0.32
0.013
-0.06
1

```
df.describe()
```

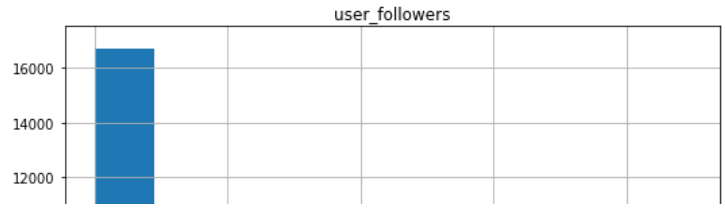
	user_followers	user_friends	user_favourites
count	7.443600e+04	74436.000000	7.443600e+04

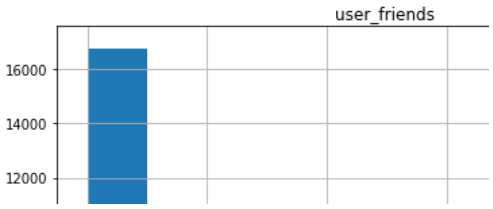
```
dfiphone=df[df["source"]=="Twitter for iPhone"]
dfiphone.hist(figsize=(20,15))
plt.show()
```



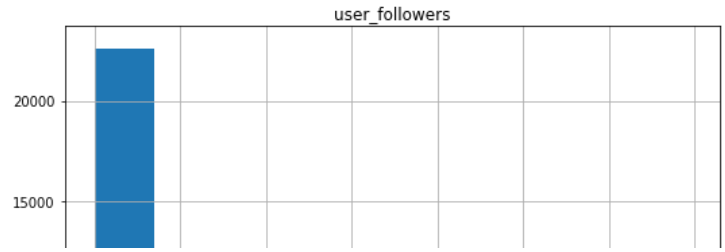


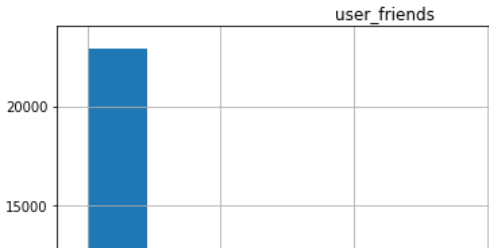

```
dfiphone=df[df["source"]=="Twitter for Android"]
dfiphone.hist(figsize=(20,15))
plt.show()
```






```
dfiphone=df[df["source"]=="Twitter Web App"]
dfiphone.hist(figsize=(20,15))
plt.show()
```



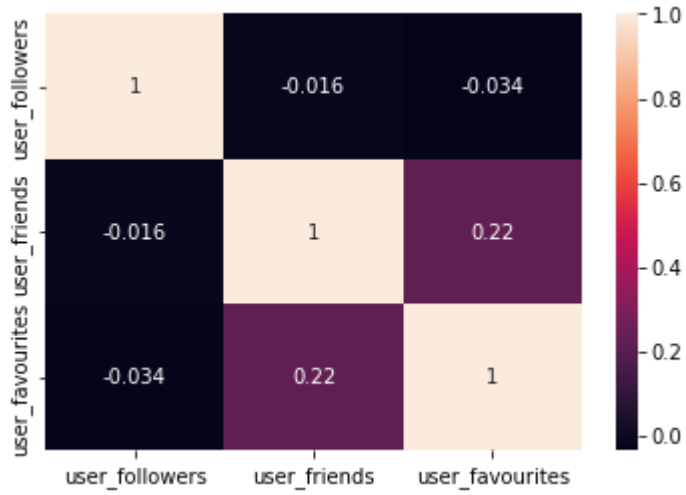



```
df['source'].value_counts().nlargest(5)
```

```
Twitter Web App        55651
Twitter for Android    39441
Twitter for iPhone     34539
TweetDeck               8399
Hootsuite Inc.          7134
Name: source, dtype: int64
```

```
plt.figure(figsize=(20,10))
df['source'].value_counts().nlargest(10).plot(kind='bar')
plt.xticks(rotation=45)
plt.show()
```


```
df2 = df[df["source"] == "Twitter Web App"]
corrmat = df2.corr()
sns.heatmap(corrmat, annot=True)
plt.show()
```



PREGUNTAS Y RESPUESTAS

Luis Gerardo Magaña Yáñez

¿Hay alguna variable que no aporta información?

So, la variable de is_retweet

Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué?

Yo eliminaría la variable de is_retweet, porque la verdad no es necesaria para todos nuestros resultados.

¿Existen variables que tengan datos extraños?

- En relación con nuestros resultados no existen variables con datos extraños, aunque como anteriormente comenté la variable is_retweet no contiene algo que sea de utilidad para nosotros.

Si comparas las variables, ¿todas están en rangos similares?

- No necesariamente todas, pero las que si son las de usuarios de android y usuarios de iphone..

¿Crees que esto afecte el análisis de los datos?

- No, la verdad no creo que afecte, ya que son indispensables para nuestro planteamiento.

Marco Iván Pacheco Martínez

Hay alguna variable que no aporta información? En este caso, las variables que no parecen aportar mucha información son user_name y user_description,user_location, user_created y is_retweet

Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué? user_name ,user_description y user_location se eliminarían ya que no son relevantes para el análisis, a menos que fuera para analizar la actividad de ciertos usuarios pero este no es el caso. Las variables user_created, y is_retweet se eliminarian Tambien ya que no son importantes para el análisis de datos relacionado al covid19, y no se encontraría ninguna correlación.

¿Existen variables que tengan datos extraños? Aparentemente no se ve una variable que contenga datos extraños

Si comparas las variables, ¿todas están en rangos similares? No necesariamente están dentro de los rangos, por ejemplo, la variable user_followers puede variar desde unos pocos hasta millones, mientras que user_favourites puede variar desde cero hasta varios miles. También algo similar puede ocurrir con los user_friends

¿Crees que esto afecte el análisis de los datos? Si afecta ya que al tener algunos datos fuera de la escala de las variables al realizar el análisis, es necesario considerar estandarizar o normalizar los datos para poder asi compararlos adecuadamente.

¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos? Si es posible hallar grupos que se parezcan, y que supone que hay tweets que compartan características similares. Identificando grupos usando clustering con k-menas , y definiendo las características y como se medirán las similitudes de los tweets en base a la verificación y el source de donde saldrían los datos, Personas con source Android son los que tienen altos valores en user_friends, mientras que las personas con source Web son los que tienen valores mas altos en user_followers.

Iñaki Vigil Arrechea A01662274 Generamos mapas de calor y gráficas de barras para poder comparar los datos de cada tipo de usuario. Existen muchos sources, pero los tres dispositivos de usuario son: 1) Twitter Web App, 2) Twitter for Android y 3) Twitter for iPhone A partir de estos tres es que vimos las relaciones entre la verificación, la cantidad de favoritos, amistades y seguidores de las cuentas, y si el source siendo diferente mostraba algo interesante.

Notamos que las personas de Android son las que mayor cantidad de amistades tienen, dado que las cifras significativas se dejan de mostrar alrededor de los 100,000. Después van los de Web App acabando poco antes de 100,000 sus cifras significativas. Al final van los usuarios de iPhone, con

sus cifras significativas acabando poco después de los 50,000. Este truco de fijarnos en cómo se distribuían los ejes ayudó bastante para el análisis de datos.

En cuanto a la cantidad de seguidores, los usuarios de la Web App son los que usan las personas más famosas. Tanto le ganó a los demás que su eje requirió un exponente mayor. Se notan datos significantes hasta 1.4e7, o 14e6, lo que deja muy atrás a los otros usuarios.Los de iPhone quedan en 2do con sus cifras acabando poco antes del 2 1e6, y al fina Android acabando sus cifras pcoo antes del 1 1e6.

Por último, en cuanto a los usuarios favoritos, tienen cantidades similares, pero notamos que a los usuarios de iPhone es a los que más les gusta guardar favoritos, ya que, aunque el máximo es muy similar en los tres, hay más usuarios de iPhone guardando favoritos. Esto lo vemos por el número de columnas desplegadas. En Android se despliegan 3 columnas, en web 5, y en iPhone se muestran 7 columnas.

Antes nos servía la variable de ser usuario verificado, ya que solo gente con una audiencia expansa podía serlo; sin embargo, dado que ahora cualquiera puede pagar para estar con una cuenta verificada, esta variable, en mi opinión, se ha vuelto inútil para nuestro análisis de los datos. Teníamos en mente asociar la verificación con la fama, y checar qué dispositivo era el más común entre famosos, pero la nueva verificación no permite este análisis con esta variable.

Es interesante ver cómo las variables son bastante similares entre sí aún entre los dispositivos. No existe un dispositivo que sobresalga demasiado en amistades, por ejemplo. Obvio hay diferencias, pero ninguna como una obvia ganadora. La única cercana es la de followers, donde se tienen más columnas en la Web App, mostrando que hay más gente en los rangos altos de seguidores.

Antonio Machorro A01782114

1.La variable **user_created**, **user_location** y **text** no aportan ningún dato de interés para resolver el objetivo. 2.Quitaría **is_retweet** así como las variables que no nos interesan en el punto anterior. 3.No hay variables con datos extraños pero es curioso observar la relación entre la **source** del Tweet así como los números de followers y favourites que tienen. Debido a que en los hotmaps podemos observar que en iPhone hay una relación más débil entre los usuarios verificados y el número de seguidores que cada usuario tiene, entonces podemos afirmar que hay menos personas verificadas que utilizan iPhone o que hay más personas no verificadas con un gran número de seguidores en iPhone que en las otras plataformas mayores.

4. Los rangos varían dependiendo de la variable. Para número de followers, los números serán mayores que número de favoritos. Asímismo, hay variables que son caracteres como el usuario, hashtags, el texto del tweet, etc. Además tenemos **user_verified** que tiene un valor booleano.

5. Eso afecta ciertos análisis de datos pues no se puede comparar ciertos datos entre ellos. Por ejemplo, no es posible hacer un histograma de variables booleanas debido a que no se puede contar la cantidad de valores cuando sólamente hay dos posibles opciones. Asímismo, no se pueden hacer gráficas significativas con texto.
6. Los grupos que más se parecen son **user_followers**, **user_favourites** y **user_friends**, pues todos son números enteros y, además, muchas veces se encuentran en un rango similar.

Erick Trinidad Limón Ace A01735902 ¿Hay alguna variable que no aporta información? R=Para nuestro análisis las variables user_create, is_retweet no aporta información Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué? R=Debido a que no aportan información eliminaria la variable is_retweet, IFTT, Instagram y Sprout Social ¿Existen variables que tengan datos extraños? R= No encuentro variables con datos extraños, pero si es un poco bizarro el numero de usuarios activos ya que es un número muy grande Si comparas las variables, ¿todas están en rangos similares? R=No necesariamente, algunas variables pueden tener rangos parecidos como user_followers vs user_favourites vs user_friends. Sin embargo hay unos con rangos muy diferentes, como por ejemplo la variable "user_followers_count" en comparación con la variable "retweet_count". ¿Crees que esto afecte el análisis de los datos? R=Depende del análisis que se quiera realizar, pero en nuestro caso es bueno que las variables estén en rangos similares y así evitar que una variable con valores muy grandes o muy pequeños tenga un peso excesivo en el análisis a pesar de que si existe una diferencia entre el tipo de usuario no representa diferencia muy grande(por lo menos en el top 3; Web, Android y Iphone) ¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos? R=Si hay grupos con rangos parecidos como user_followers vs user_favourites vs user_friends.