Jorge Gerardo Iglesias Ortiz - A01653261

Alejandro Hernández De la Torre - A01651516

Santiago Orozco Quintero - A01658308

Carlos Andres Barredeaz Rios - A01653183

Jorge Yepez Frutos - A01652661

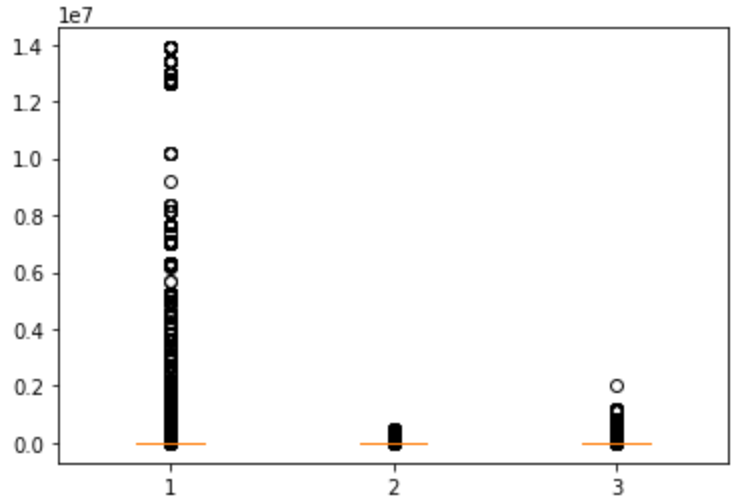
Actividad Evaluable: Mapas de calor y boxplots

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

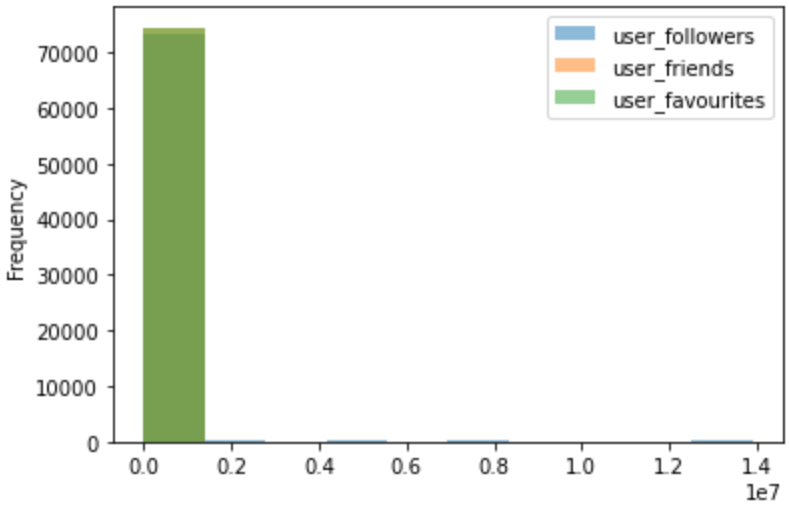
In [1]:
```
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
import numpy as np

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

Realiza el análisis de las variables usando diagramas de cajas y bigotes, histogramas y mapas de calor.

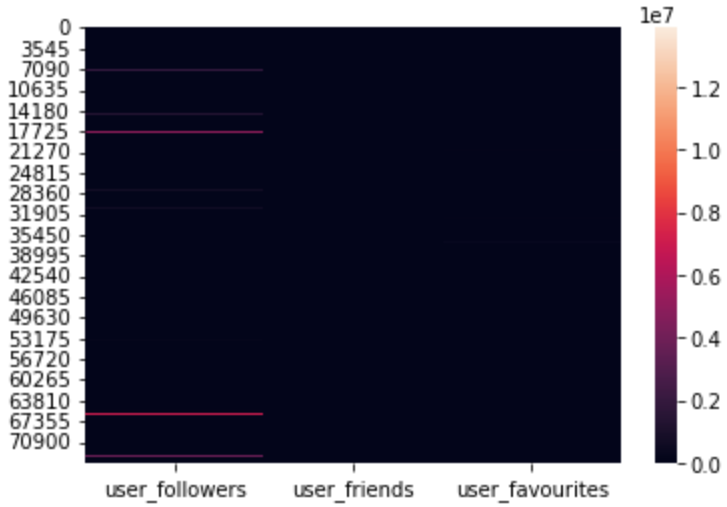
In [2]:
```
plt.boxplot(df[["user_followers","user_friends","user_favourites"]])

plt.show()
```



In [3]:
```
ax = df[["user_followers","user_friends","user_favourites"]].plot.hist(bins=10, alpha=0.
```



In [4]:
```
ax = sb.heatmap(df[["user_followers","user_friends","user_favourites"]])
```



In [5]:
```
details = pd.DataFrame(df[["user_followers","user_friends","user_favourites"]])

details = details.apply(lambda x : True if x['user_followers'] <1 and x['user_favourites

num_rows = len(details[details == True].index)

print('Cuentas sin seguidores ni usuarios seguidos : ',
      num_rows )
```

```
Cuentas sin seguidores ni usuarios seguidos :  254
```

In [6]:
```
df.describe(include = object).transpose()
```

Out[6]:

	count	unique	top	freq
user_name	74436	44853	GlobalPandemic.NET	312
user_location	59218	14622	India	1496
user_description	70079	42690	Breaking News & Critical Information to SURVIV...	312
user_created	74436	45554	2010-07-13 21:58:05	312
date	74436	56546	2020-07-29 16:30:00	26
text	74436	74312	Greenland has no active cases of the novel cor...	6
hashtags	53002	23445	['COVID19']	16004
source	74424	450	Twitter Web App	22974

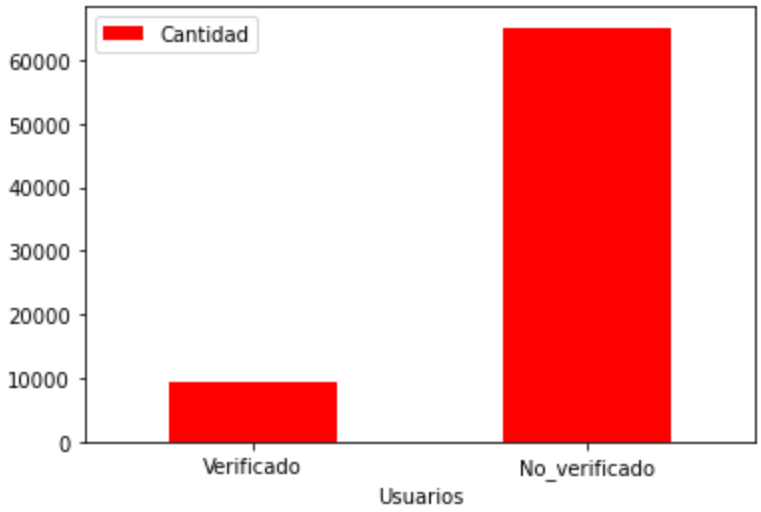
In [7]:
```
details = pd.DataFrame(df[["user_followers","user_friends","user_favourites","user_verif


details = details.apply(lambda x : True
            if x['user_verified']==True   else False, axis = 1)

# Count number of True in the series
num_rows = len(details[details == True].index)

print('Cuentas verificadas: ',
      num_rows )
```

```
Cuentas verificadas:  9354
```

In [8]:
```
details = pd.DataFrame(df[["user_followers","user_friends","user_favourites","user_verif


details = details.apply(lambda x : True
            if x['user_verified']==False   else False, axis = 1)

# Count number of True in the series
num_rows2 = len(details[details == True].index)

print('Cuentas no verificadas: ',
      num_rows2 )
```

```
Cuentas no verificadas:  65082
```

In [9]:
```
df1 = pd.DataFrame({'Usuarios':['Verificado', 'No_verificado'], 'Cantidad':[num_rows,  n
ax = df1.plot.bar(x='Usuarios', y='Cantidad', rot=0,color={"red"})
```



Responde las siguientes preguntas:

1. ¿Hay alguna variable que no aporte información?

No, todas las variables aportan información ya que representan fragmentos de datos relevantes de la cuenta de una persona, si bien por la naturaleza de los que se presentan existe cierta información que no se puede graficar de manera tan directa con un cierto análisis esta puede presentar datos útiles o interesantes, un ejemplo de esto podría ser la descripción donde se podría analizar la cantidad de personas que tienen una y que es lo más común que se suele escribir en éstas.

1. Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

User_description y Source

User_description: En nuestra opinión, este apartado es el que presenta los datos menos relevantes o más difíciles de manejar ya que si bien se pueden obtener correlaciones y tendencias con respecto a las personas y su preferencia al momento de escribir sus biografías no cual no es información indispensable al momento de analizar temas de covid.

Source: Es una variable que omitiremos ya que no representa una variable al tipo de análisis que queremos hacer. Si bien la proveniencia del tweet, sea de algún tipo de sistema puede significar algo, como proveniencia de un estrato social o para algún tipo de estudio de mercado, no está ligado al tipo de análisis que queremos hacer.

1. ¿Existen variables que tengan datos extraños?

Por supuesto siendo los nombres, descripción y algunos contenidos de tweets ya que utilizan caracteres que no pueden soportar dentro del código a menos que se pueda guardar el tipo de carácter en char además de tener en cuenta que los usuarios pueden escribir lo que quieran en estos apartados por lo que puede terminar poniendo información o caracteres sin sentido.

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

No, hay rango de valores en relación a la cantidad de seguidores, por ejemplo. Si bien los rangos van de 0 seguidores, a más 5000 hace que medir promedios generales sea algo innecesario ya que los rangos de valores son muy diferentes. Incluso el perfil de los trabajadores.

1. ¿Crees que esto afecte?

No, ya que si bien hace que los datos varían en gran medida esto no afecta el trabajo de analizarlos e interpretarlos, de hecho incluso esta distribución podría generar nueva información a contemplar observando cómo se distribuyen y acoplan los datos, aún más si tenemos en cuenta que se está revisando algo tan diverso como puede ser la actividad en una plataforma como twitter es de esperarse que los datos difieran mucho entre algunos usuarios.

1. ¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

Puede ser en la parte de seguidores, amigos y favoritos ya que en unas correlaciones hechas en una tarea anterior se encuentra que usuarios con seguidores tienen probabilidad de que también tengan un grupo de amigos al igual que si tienen amigos tienen más probabilidad de tener favoritos. De la misma manera en que se podrían contar la cantidad que hay por país y usuario para ver la influencia que tienen tanto nacional como internacional.

In [ ]: