Actividad Evaluable: Obtención de estadísticas descriptivas

Paola Fernández Gutiérrez Zamora - A01658087

TC1002S.222

Profesor:
Sergio Ruiz Loza

Fecha de entrega: 11 de Mayo de 2022

Actividad Evaluable: Obtención de estadísticas descriptivas

Para realizar esta actividad primero se cargaron los datos de los tweets de covid 19 utilizando la librería pandas.

```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np
        import seaborn as sns; sns.set_theme()

        # Leer el archivo utilizando read_csv
        data = pd.read_csv("covid19_tweets.csv")
```

Después se verificó la cantidad de datos que contenían las variables y se identificó el tipo de variable.

```
: data.count()

: user_name           74436
  user_location       59218
  user_description    70079
  user_created        74436
  user_followers      74436
  user_friends        74436
  user_favourites     74436
  user_verified       74436
  date                74436
  text                74436
  hashtags            53002
  source              74424
  is_retweet          74436
  dtype: int64
```

```
In [8]: data.columns

Out[8]: Index(['user_name', 'user_location', 'user_description', 'user_created',
               'user_followers', 'user_friends', 'user_favourites', 'user_verified',
               'date', 'text', 'hashtags', 'source', 'is_retweet'],
              dtype='object')
```

```
In [5]: data.dtypes

Out[5]: 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
```

Para analizar las variables y los rangos en las que se encontraban se utilizó la función describe con cada columna y se hizo una gráfica de cajas para las columnas con valores numéricos para poder ver el rango en el que se encontraban de manera visual.

```
In [12]: data.iloc[:, 0].describe()

Out[12]: count                 74436
         unique                44853
         top       GlobalPandemic.NET
         freq                    312
         Name: user_name, dtype: object
```

La variable user_name representa el nombre de usuario de la persona que realizó el tweet

```
In [13]: data.iloc[:, 1].describe()

Out[13]: count     59218
         unique    14622
         top       India
         freq       1496
         Name: user_location, dtype: object
```

La variable user_location representa el lugar desde el que se realizó ese tweet, es posible observar que el lugar más común fue India

```
In [14]: data.iloc[:, 2].describe()

Out[14]: count                                                 70079
         unique                                                42690
         top       Breaking News & Critical Information to SURVIV...
         freq                                                    312
         Name: user_description, dtype: object
```

La variable user_description representa la descripción del usuario que realizó el tweet. Podemos observar que el que más se repite tiene el mismo número de frecuencia que el user_name, esto es porque ese usuario fue el que más tweets sobre el covid 19 tiene, por lo tanto aparece que su username y su user_description es la más común.

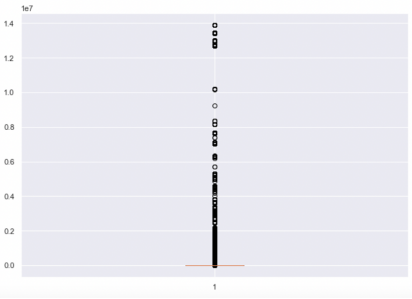
```
In [16]: data.iloc[:, 4].describe()

Out[16]: count    7.443600e+04
         mean     1.059513e+05
         std      8.222900e+05
         min      0.000000e+00
         25%      1.660000e+02
         50%      9.600000e+02
         75%      5.148000e+03
         max      1.389284e+07
         Name: user_followers, dtype: float64
```

```
In [3]: fig = plt.figure(figsize =(10, 7))

        # Creating plot
        plt.boxplot(data.iloc[:, 4])

        # show plot
        plt.show()
```



Finalmente se realizaron algunos análisis de los datos para complementar a lo encontrado con las funciones describe

```
In [25]: data.sort_values(by="user_followers", ascending=False).head(10)
```

Out[25]:

	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...

Podemos observar que la cuenta con más seguidores es CGTN, una cuenta verificada de Beijing, China que realizó diferentes tweets sobre el covid, incluso se puede ver la fluctuación de followers que tuvo a diferentes horas y en diferentes días

```
In [52]: data.corr(method ='pearson')
```

Out[52]:

	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

Así mismo se puede observar que hay una correlación relativamente alta entre el número de seguidores y si una cuenta es verificada. Una cuenta verificada podría indicar que es una fuente confiable, sin embargo si lo que determina la verificación de la cuenta es el número de seguidores, no necesariamente se tratan de datos verídicos y comprobables.

Con la información obtenida y añadiendo al pequeño análisis que se realizó en el desarrollo del análisis se pudo concluir lo siguiente:

Gracias a la información obtenida al describir cada variable es posible observar que se tiene un gran rango de datos, especialmente en el número de seguidores, esto puede ser bueno ya que nos permite observar una mayor muestra de datos y se puede hacer un análisis más completo al comparar los tweets de cuentas con más seguidores a las que tienen menos. Además podemos ver que en la muestra de datos se pueden observar tweets de las mismas cuentas por lo que es posible encontrar relación entre el número de tweets sobre el tema con el número de seguidores, si es una cuenta verificada o algún otro factor.
También es posible agrupar datos, al contar con un número de datos tan grande y tener un rango tan amplio se pueden agrupar los datos en distintos grupos para poder obtener información más significativa y poder llegar a conclusiones generales. Por ejemplo, se podría obtener el número de tweets que utilizaron un hashtag en específico o ver si las cuentas con más seguidores tienen más tweets sobre el covid-19.
Finalmente, podemos observar por las descripciones de las variables que la mayoría de los datos que se muestran son significativos y pueden ser de utilidad para realizar un análisis de datos, sin embargo, al analizar la relación que tienen podemos ver que hay algunas que no aportan mucho valor como la descripción de los usuarios o las fechas de creación de las cuentas.