Tecnológico de Monterrey

Herramientas computacionales: el arte de la analítica

Entregable

Actividad Evaluable 2: Obtención de estadísticas descriptivas

Presenta

Javier Valente Rodríguez A01662693

Docente

Sergio Ruiz Loza

Ciudad de México Mayo, 11 de 2022

1. Cargamos los datos con ayuda de *pandas*.

```python
import pandas as pd

data = pd.read_csv("covid19_tweets.csv")
data.head(n = 4)
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashta
0	ᏉᎥ☻լꂅ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21	If I smelled the scent of hand sanitizers toda...	N
1	Tom Basile us	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it...	N
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14	@diane3443 @wdunlap @realDonaldTrump Trump nev...	['COVID'
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10	@brookbanktv The one gift #COVID19 has give me...	['COVID'

2. Verificamos la cantidad de datos con los que contamos e identificamos el tipo de dato de las variables.

```python
#cantidad de objetos
print("Cantidad de datos: ")
print(len(data.index)*(len(data.columns)))

print("\nTipo de variables: ")
# tipo de variables
print(data.dtypes)
```

```
Cantidad de datos:
967668

Tipo de variables:
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
```

3. Analiza las variables para saber lo que representa cada una y en qué rangos se encuentran.

   Tres de las variables (en las columnas) son analizados de manera numérica:

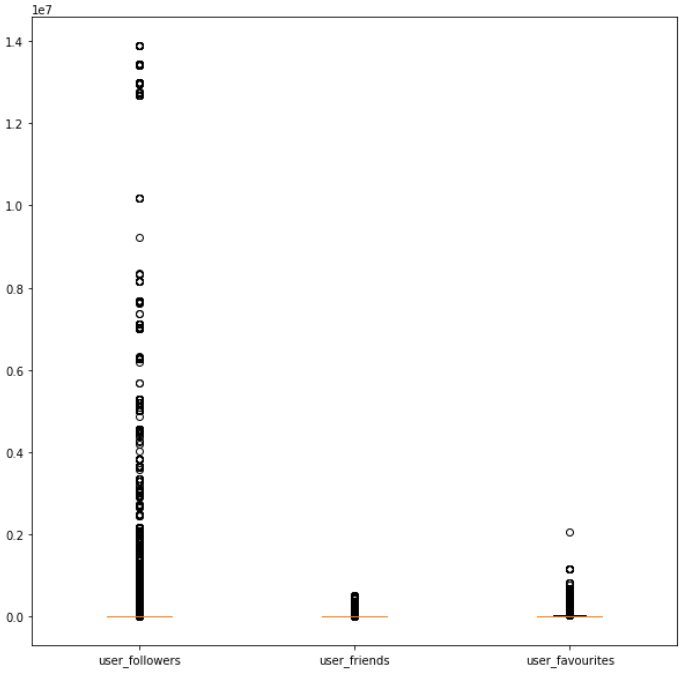
   - User_followers (número de seguidores del usuario)
   - User_friends (número de amigos del usuario)
   - User_favourites (cantidad de favoritos añadidos por el usuario)

   El resto de variables son de tipo *string*, es decir, de tipo texto; para el análisis de las variables numéricas se usaron las funciones:

   - df.describe() → Para obtener el número total de datos, la media, desviación estándar, valores mínimos y máximos.

   - df.median() → Para obtener la mediana de los mismos datos.

   Adicional a lo anterior, se realizó un diagrama de Boxplot:



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

   - User_followers:

     Cuenta con la dispersión más grande de los datos, esto comprobándose en el diagrama de Boxplot, su mediana y su media nos hacen inferir que esta dispersión de datos pudo verse afectada por tweets publicados por cuentas de gran número de seguidores.

   - User_friends:

     Se puede notar que su desviación estándar es muy pequeña, esto nos hace sentido debido a que, la mayor parte de los usuarios no siguen a un gran número de cuentas.

   - User_favourites:

     Cuenta con una desviación estándar en el medio, la mayor parte de los usuarios tiene aproximadamente unos 1927 tweets, esto lo sabemos gracias a la mediana, sin embargo, hay usuarios que la media es mayor a la cantidad anterior, por lo que podemos inferir que podrían pasar más tiempo dentro de la aplicación.