Juan Francisco García Rodríguez
A01660981
Equipo 1

# Actividad Evaluable 2: Obtención de estadísticas descriptivas

Enlace al repositorio del GitHub:
https://github.com/JFranciscoGR03/arte_de_analitica_equipo1.git

1- Verifica la cantidad de datos que tienes, las variables que contiene cada vector de datos e identifica el tipo de variables.

Hay 74,436 objetos (filas), las cuales contienen un tweet relacionado con el Covid-19, y cada uno de esos tweets contiene 13 variables (columnas), que nos dan información adicional; por lo que contabilizando toda la base de datos, se tienen un total de 967,668 datos.

Las variables contenidas en cada vector de datos junto con el tipo de variables se muestran a continuación:

- 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

Las variables que contienen "object" significa que contienen una cadena de caracteres alfanuméricos o incluso caracteres especiales; las que aparecen como "int64" contienen solo números enteros; y las que aparecen con la leyenda "bool" significa que solo son verdadero o falso.

2- Analiza las variables para saber qué representa cada una y en qué rangos se encuentran. Si la descripción del problema no te lo indica, utiliza el máximo y el mínimo para encontrarlo.

Todas las variables que aparecen tienen importancia porque revelan información que permite tener un mayor contexto de cada tweet. Analizándolas más a fondo y de acuerdo con el tipo de dato que contienen, daremos una descripción más amplia acerca de cada una de ellas.

1. user_name: El nombre de usuario asignado a la cuenta de twitter.
2. user_location: El lugar geográfico donde se encontró la persona al momento de publicar el tweet.
3. user_description: Alguna información adicional que el usuario agregó a su perfil.
4. user_created: La fecha y la hora en la que fue creada esa cuenta de twitter.
5. user_followers: La cantidad de seguidores que tiene cada usuario en su cuenta.
6. user_friends: La cantidad de amigos que la persona tiene en su cuenta.
7. user_favourites: El número de personas que cada usuario tiene marcado como favoritos.
8. user_verified: Sólo aparece verdadero si el usuario tiene su cuenta verificada.
9. date: La fecha y la hora en la que cada usuario publicó su respectivo tweet.
10. text: Muestra el contenido del tweet de cada usuario.
11. hashtags: Aparece si el usuario utilizó algún hashtag al realizar su publicación.
12. source: Revela la fuente digital donde se utilizó twitter al realizar el tweet.
13. is_retweet: Sólo es verdadero si el tweet de cada usuario fue compartido por más personas.

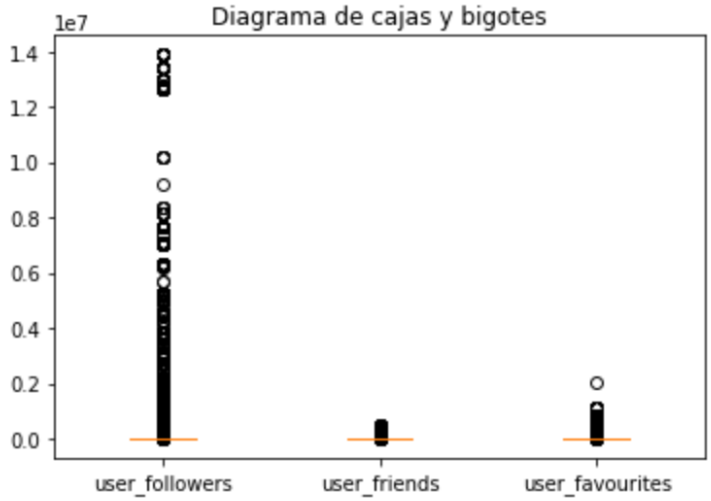
Por obvias razones, solo obtuvimos la media, mediana y desviación estándar de las variables con datos numéricos enteros, las cuáles son “user_followers”, “user_friends”, y “user_favourites”. En la tabla que se adjunta acá se obtienen más datos de los solicitados, que igual tienen cierta importancia, pero los más importantes son “mean”, “std”, “min”, “max”, y “50%”; los cuales representan la media, la desviación estándar, el valor mínimo, el valor máximo y la mediana respectivamente.

| | user_followers | user_friends | user_favourites |
|---|---|---|---|
| **count** | 7.443600e+04 | 74436.000000 | 7.443600e+04 |
| **mean** | 1.059513e+05 | 2154.721170 | 1.529747e+04 |
| **std** | 8.222900e+05 | 9365.587474 | 4.668971e+04 |
| **min** | 0.000000e+00 | 0.000000 | 0.000000e+00 |
| **25%** | 1.660000e+02 | 153.000000 | 2.200000e+02 |
| **50%** | 9.600000e+02 | 552.000000 | 1.927000e+03 |
| **75%** | 5.148000e+03 | 1780.250000 | 1.014800e+04 |
| **max** | 1.389284e+07 | 497363.000000 | 2.047197e+06 |

Añado que también se obtuvieron tablas similares para las variables tipo “object” y tipo “bool”, que de igual manera nos arrojan información relevante para el análisis de los tweets; sin embargo, dichas tablas solo se plasmaron en el libro de jupyter porque, al no contener datos numéricos, no se pudieron obtener resultados estadísticos similares a los de las variables numéricas.

3- Basándose en la media, mediana y desviación estándar de cada variable, ¿Qué conclusiones puedes entregar de los datos?

Se puede concluir que están muy dispersos los datos porque hay usuarios que tienen muchos seguidores, muchos amigos y favoritos, pero hay otros que apenas y tienen algunos.




Por último, en la gráfica de arriba observamos que user_followers es la variable más dispersa, lo cual conlleva a que tenga una desviación estándar más grande que las otras dos variables; que si bien están algo dispersas, su rango no es tan variado.