

Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad Evaluable 2: Obtención de estadísticas descriptivas

Semana TEC

TC1002S

Fernando Alfonso Arana Salas - A01272933

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

1.

```
In [4]: # Carga los datos usando tu lector de csv o con pandas.
import pandas as pd
data = pd.read_csv('covid19_tweets.csv')
```

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

2.

```
In [36]: # Verifica la cantidad de datos que tienes,
len(data)
Out[36]: 74436
```

Cantidad de datos: 74,436

```
In [27]: # Verifca las variables que contiene cada vector de datos
         data.columns
Out[27]: 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 [26]: # Identifica el tipo de variables
                              data.dtypes
                    Out[26]: user name
                                                    object
                              user location
                                                    object
                              user_description
                                                    object
                              user_created
                                                    object
                              user_followers
                                                     int64
                              user_friends
                                                     int64
                                                    int64
                              user_favourites
                              user verified
                                                      bool
                              date
                                                     object
                              text
                                                    object
                              hashtags
                                                    object
                              source
                                                    object
                              is_retweet
                                                     bool
                              dtype: object
```

3. 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.

Tenemos 13 variables (columnas) que describen los tweets:

- 1. Nombres de usuario
- 2. Ubicación de usuario
- 3. Descripción de usuario
- 4. Fecha de creación de usuario
- 5. Numero de seguidores
- 6. Numero de amigos del usuario
- 7. Número de favoritos del usuario
- 8. El usuario está verificado
- 9. Fecha del tweet
- 10. Contenido del tweet
- 11. Hashtags del tweet
- 12. Fuente
- 13. Es o no es re tweet

De estas, solo 3 son numéricas de tipo entero:

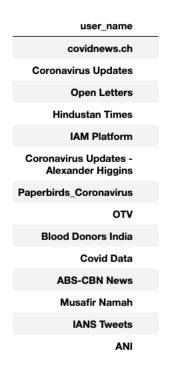
- 1. Numero de seguidores
- 2. Numero de amigos del usuario
- 3. Número de favoritos del usuario

El total de tweets (74,436) provienen de 44,853 cuentas.

```
len(data.groupby('user_name'))
Out[46]: 44853
```

Las 15 cuentas con el mayor número de tweets son:

```
selected = data.groupby('user_name').count().sort_values(by='user_followers', ascending=False)
selected[1:15]
```



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



Podemos observar que las 3 columnas cuyo tipo de datos es numérica es la que se nos presenta. Estas 3 columnas nos dan información sobre el usuario que publicó un tweet. Debido a esto podemos observar como hay tweets que corresponden a usuarios con cero seguidores, y tweets que corresponden a usuarios con 1.389e07 seguidores. El promedio de seguidores en los más de 74 mil usuarios que publicaron tweets es de 1.05e05.