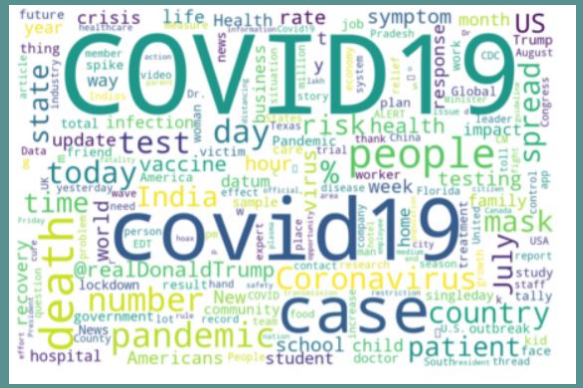
Reto

“Analisis a través de técnicas de analitica de datos”



Diógenes Grajales C - A01653251
Francisco Emiliano Rocha P - A01652082
Emilio Adrian Solis Velasco - A01659828

Diógenes Grajales

S Generar modelos gráficos que representen nuestros datos y variables

M Tener gráficas de tipo mapa de calor, histograma y de bigotes

A Utilizaremos los comandos y librerías vistos en clases

R Esto ayudará a poder analizar y comparar nuestras variables

T Su puede obtener estas gráficas en un corto periodo si ya se cuenta con datos y un ambiente apropiado

Francisco Emiliano Rocha

S Visualizar los hashtags más populares

M Obtener tablas y una visualización gráfica Además de un video y una presentación

A Haciendo uso de text mining y gráficos de barras

R Usar los elementos que aprendimos en esta Semana Tec

T El trabajo estará listo antes del 30 de octubre 2020 a las 10:00 am

Emilio Adrian Solis

S Clasificar las variables más importantes para hacer un análisis

M Generar diversos elementos como las gráficas para permitirnos sacar una conclusión

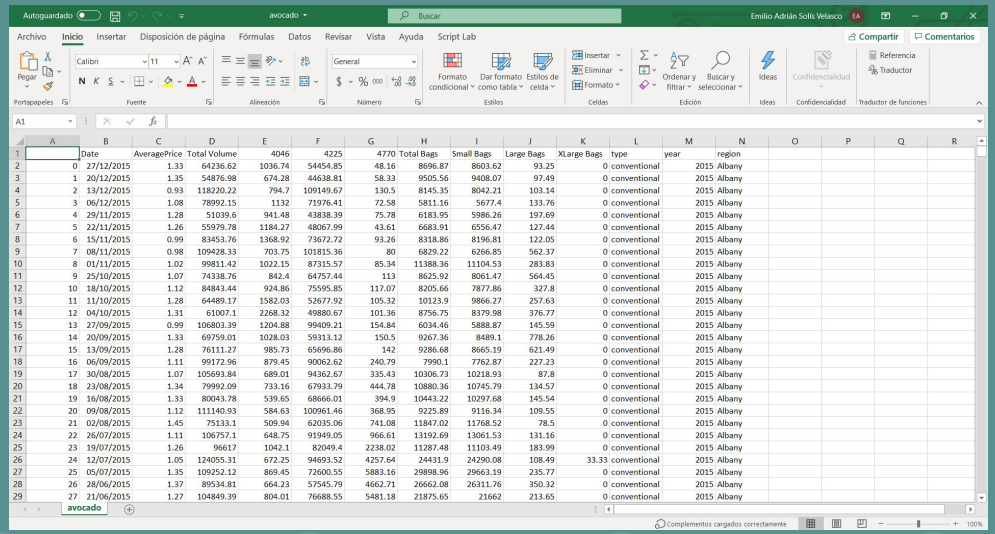
A Usando text mining, mapas de calor Usando textmining y los histogramas.

R Usando todas técnicas de analítica que estuvimos viendo a lo largo del curso

T Tenerlo listo antes del 30 de octubre 2020 en la clase

Datos Analizados

Para este reto se nos dio una base de datos en excel llamada “covid19_tweets_csv” la cual representaba los tweets que realizaban las personas y los datos de estos, los cuales cargamos en python para con ayuda de las librerías analizarlos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
2	0	27/12/2015	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0	conventional	2015	Albany
3	1	20/12/2015	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0	conventional	2015	Albany
4	2	13/12/2015	0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0	conventional	2015	Albany
5	3	06/12/2015	1.08	78992.15	1132	71976.41	72.58	5811.16	5677.4	133.76	0	conventional	2015	Albany
6	4	29/11/2015	1.28	51039.6	941.48	43838.39	75.78	6183.95	5986.26	197.69	0	conventional	2015	Albany
7	5	22/11/2015	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0	conventional	2015	Albany
8	6	15/11/2015	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0	conventional	2015	Albany
9	7	08/11/2015	0.98	109428.33	703.75	101815.36	80	6829.22	6266.85	562.37	0	conventional	2015	Albany
10	8	01/11/2015	1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0	conventional	2015	Albany
11	9	25/10/2015	1.07	74338.76	842.4	64757.44	113	8625.92	8061.47	564.45	0	conventional	2015	Albany
12	10	18/10/2015	1.12	84843.44	924.86	75595.85	117.07	8205.66	7877.86	327.8	0	conventional	2015	Albany
13	11	11/10/2015	1.28	64489.17	1582.03	52677.92	105.32	10123.9	9866.27	257.63	0	conventional	2015	Albany
14	12	04/10/2015	1.31	61007.1	2268.32	49880.67	101.36	8756.75	8379.98	376.77	0	conventional	2015	Albany
15	13	27/09/2015	0.99	106803.39	1204.88	99409.21	154.84	6034.46	5888.87	145.59	0	conventional	2015	Albany
16	14	20/09/2015	1.33	69759.01	1028.03	59313.12	150.5	9267.36	8489.1	778.26	0	conventional	2015	Albany
17	15	13/09/2015	1.28	76111.27	985.73	65696.86	142	9286.68	8665.19	621.49	0	conventional	2015	Albany
18	16	06/09/2015	1.11	99172.96	879.45	90062.62	240.79	7990.1	7762.87	227.23	0	conventional	2015	Albany
19	17	30/08/2015	1.07	105693.84	689.01	94362.67	335.43	10306.73	10218.93	87.8	0	conventional	2015	Albany
20	18	23/08/2015	1.34	79992.09	733.16	67933.79	444.78	10880.36	10745.79	134.57	0	conventional	2015	Albany
21	19	16/08/2015	1.33	80043.78	539.65	68666.01	394.9	10443.22	10297.68	145.54	0	conventional	2015	Albany
22	20	09/08/2015	1.12	111140.93	584.63	100961.46	368.95	9225.89	9116.34	109.55	0	conventional	2015	Albany
23	21	02/08/2015	1.45	75133.1	509.94	62035.06	741.08	11847.02	11768.52	78.5	0	conventional	2015	Albany
24	22	26/07/2015	1.11	106757.1	648.75	91949.05	966.61	13192.69	13061.53	131.16	0	conventional	2015	Albany
25	23	19/07/2015	1.26	96617	1042.1	82049.4	2238.02	11287.48	11103.49	183.99	0	conventional	2015	Albany
26	24	12/07/2015	1.05	124055.31	672.25	94693.52	4257.64	24431.9	24290.08	108.49	33.33	conventional	2015	Albany
27	25	05/07/2015	1.35	109252.12	869.45	72600.55	5883.16	29898.96	29663.19	235.77	0	conventional	2015	Albany
28	26	28/06/2015	1.37	89534.81	664.23	57545.79	4662.71	26662.08	26311.76	350.32	0	conventional	2015	Albany
29	27	21/06/2015	1.27	104849.39	804.01	76688.55	5481.18	21875.65	21662	213.65	0	conventional	2015	Albany

Se analizaron los datos proporcionados y determinamos los datos más relevantes los cuales nos ayudarán a determinar un análisis acerca de las situación.

Análisis Ejecutados

¿Qué técnicas se usaron y cuáles no?

Librerías Utilizadas

- Pandas
- Matplotlib
- Numply
- Seaborn
- Spacey
- Kmeans

Analizar los datos separando los numéricos de los strings

Establecer los datos que eran strings para utilizarlos

Clasificación de los datos resultantes

Agrupación y relación entre categorías

Tablas de datos, Histogramas , Mapas de calor, Gráficas 2D y 3D, Boxplot (diagrama de bigotes y cajas)

Se definieron y cargaron los datos

```
In [1]: import pandas as pd
        import matplotlib
        import matplotlib.pyplot as plt
        import numpy as np; np.random.seed(0)
        import seaborn as sns
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sb
```

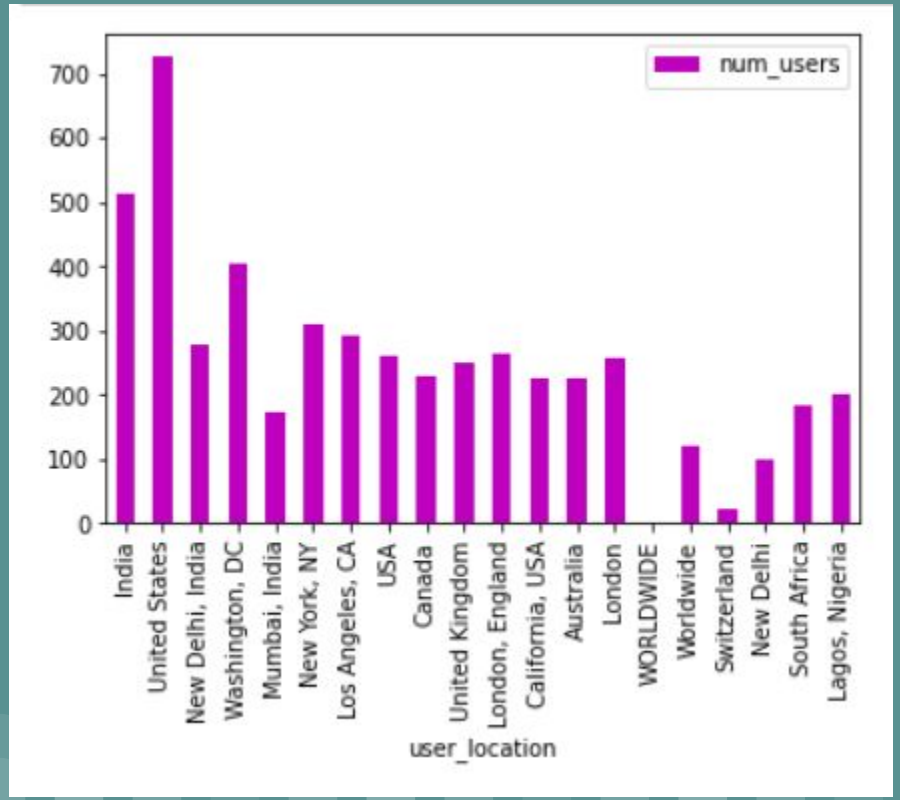
	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favouri
0	[illegible]	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	187
1	Tom Basile us	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	72
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	14
4	DIPR-J&K	Jammu and Kashmir	Official Twitter handle of Department of	2017-02-12 06:45:15	101009	168	1

Objetivo

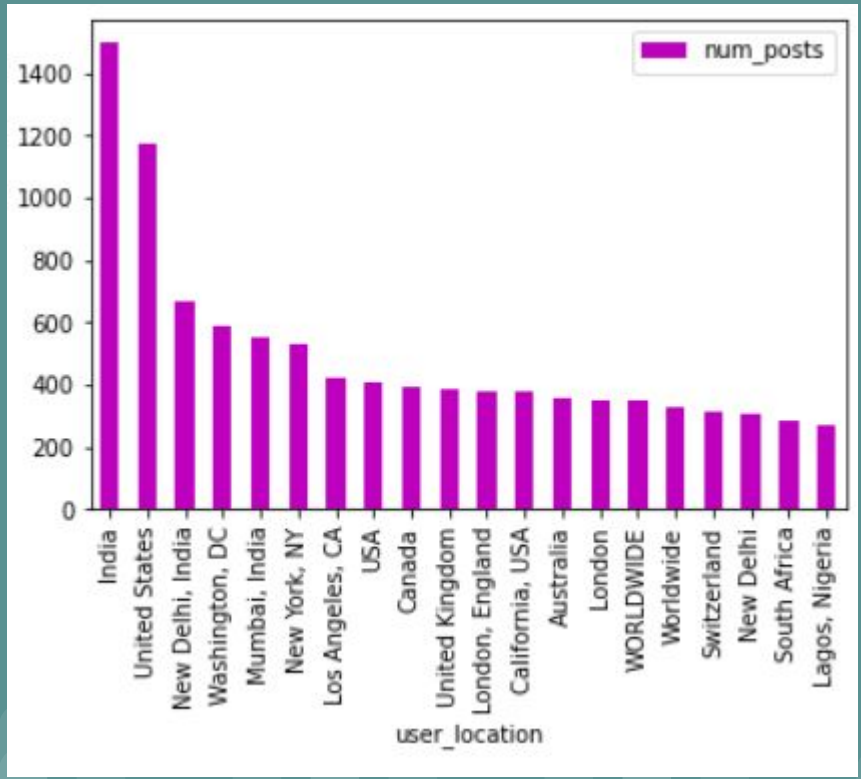
Text mining para conocer más estadísticas acerca de los tweets, los usuarios y los hashtags.

Número de usuarios twitteando desde cada ubicación (Top)

user_location	num_users	num_posts
India	512	1496
United States	726	1172
New Delhi, India	279	669
Washington, DC	403	589
Mumbai, India	173	554
New York, NY	311	527
Los Angeles, CA	292	420
USA	261	405

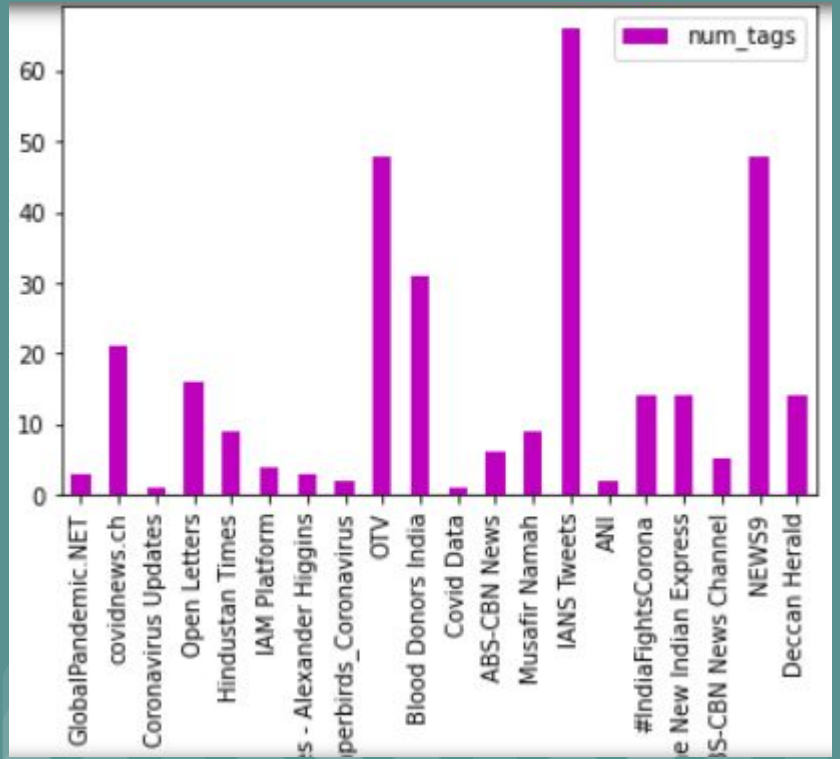


	user_name	num_posts
count	14622.000000	14622.000000
mean	2.478662	4.049925
std	13.068210	24.473345
min	1.000000	1.000000
25%	1.000000	1.000000
50%	1.000000	1.000000
75%	1.000000	2.000000
max	726.000000	1496.000000

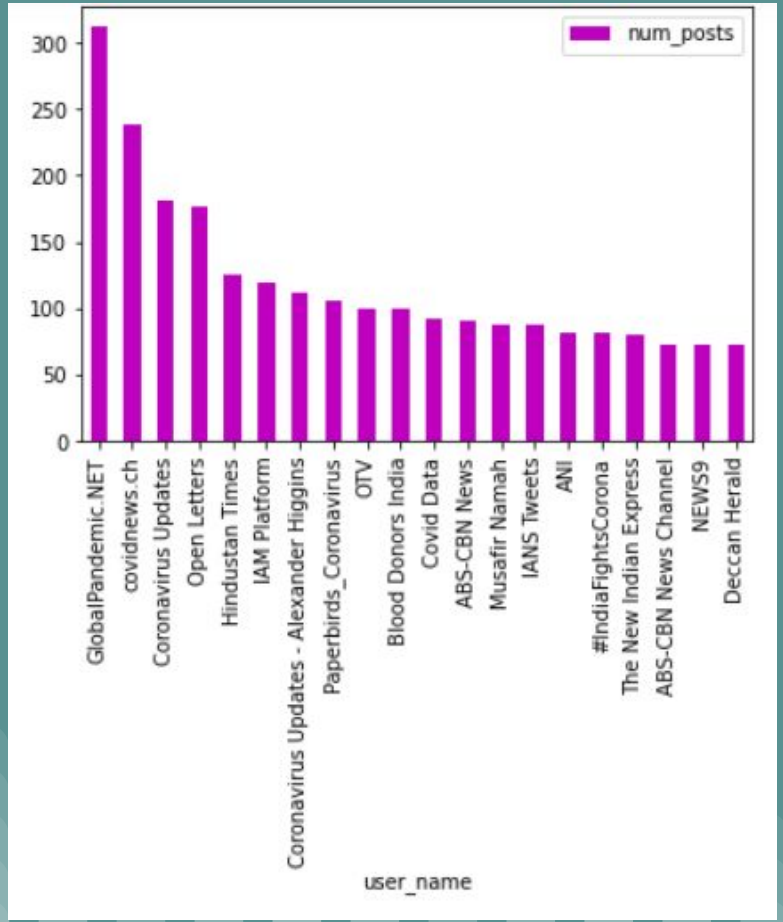


Número de hashtags usados por usuario (Top)

user_name	num_tags	num_posts
GlobalPandemic.NET	3	312
covidnews.ch	21	239
Coronavirus Updates	1	181
Open Letters	16	177
Hindustan Times	9	125
IAM Platform	4	120
Coronavirus Updates - Alexander Higgins	3	112
Paperbirds_Coronavirus	2	106

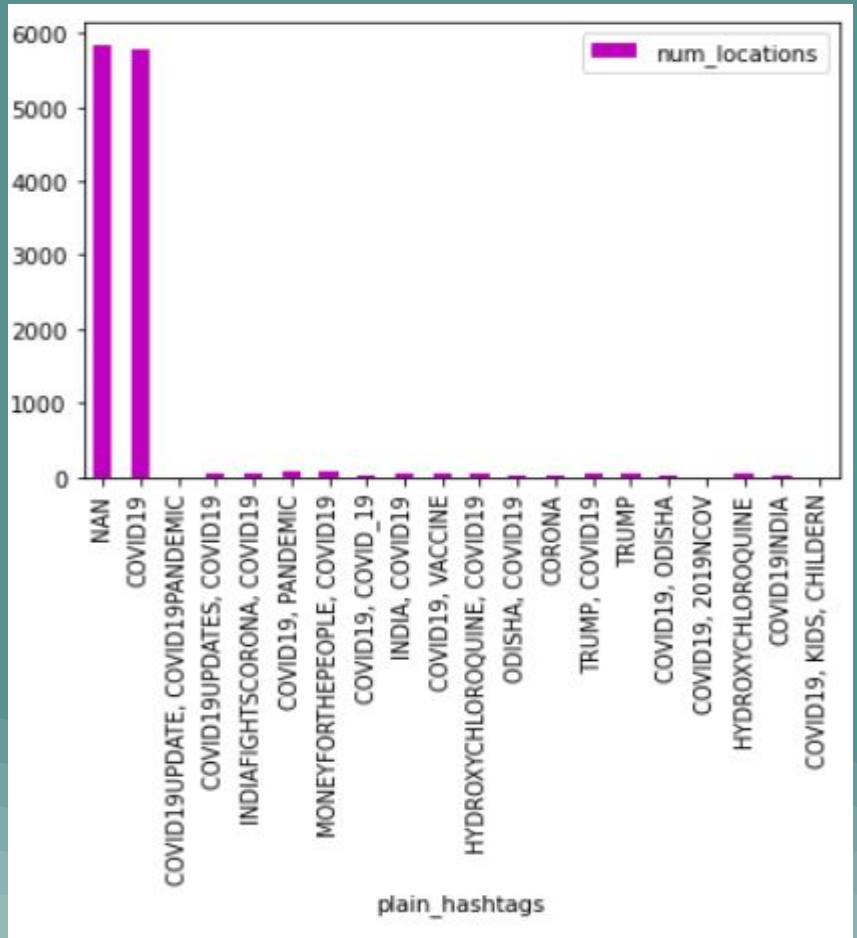


	num_tags	num_posts
count	44853.000000	44853.000000
mean	1.310838	1.659555
std	1.264445	3.684831
min	1.000000	1.000000
25%	1.000000	1.000000
50%	1.000000	1.000000
75%	1.000000	1.000000
max	66.000000	312.000000

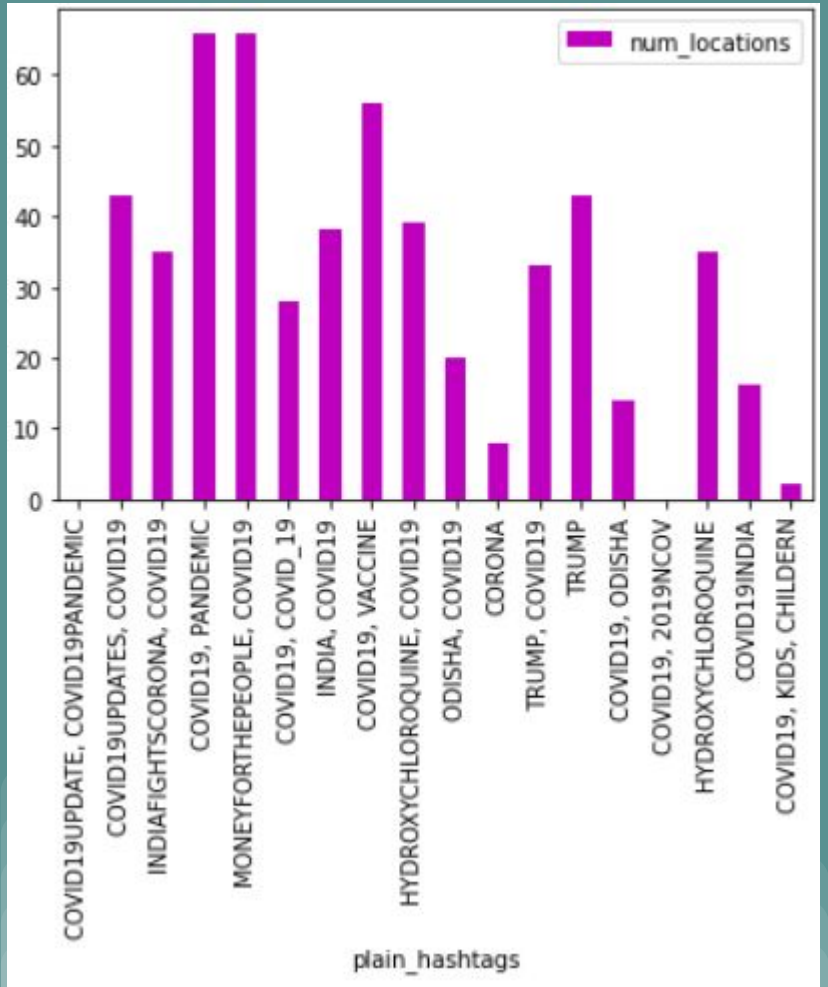


Número de ubicaciones que utilizaron un hashtag

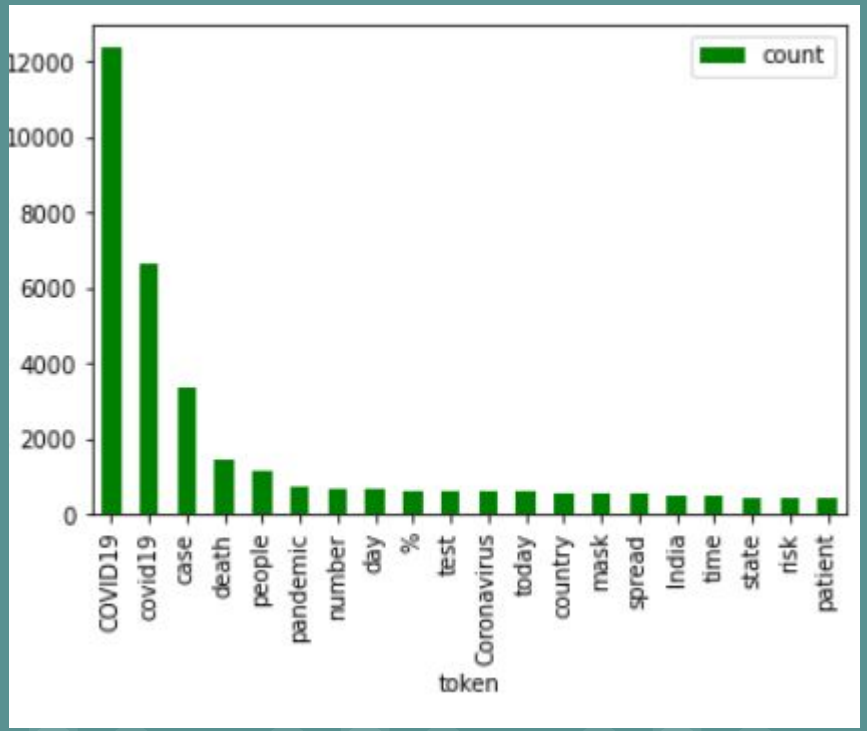
plain_hashtags	num_locations	num_posts
NAN	5847	21434
COVID19	5785	20747
COVID19UPDATE, COVID19PANDEMIC	0	181
COVID19UPDATES, COVID19	43	163
INDIAFIGHTSCORONA, COVID19	35	113
COVID19, PANDEMIC	66	113
MONEYFORTHEPEOPLE, COVID19	66	94
COVID19, COVID_19	28	82

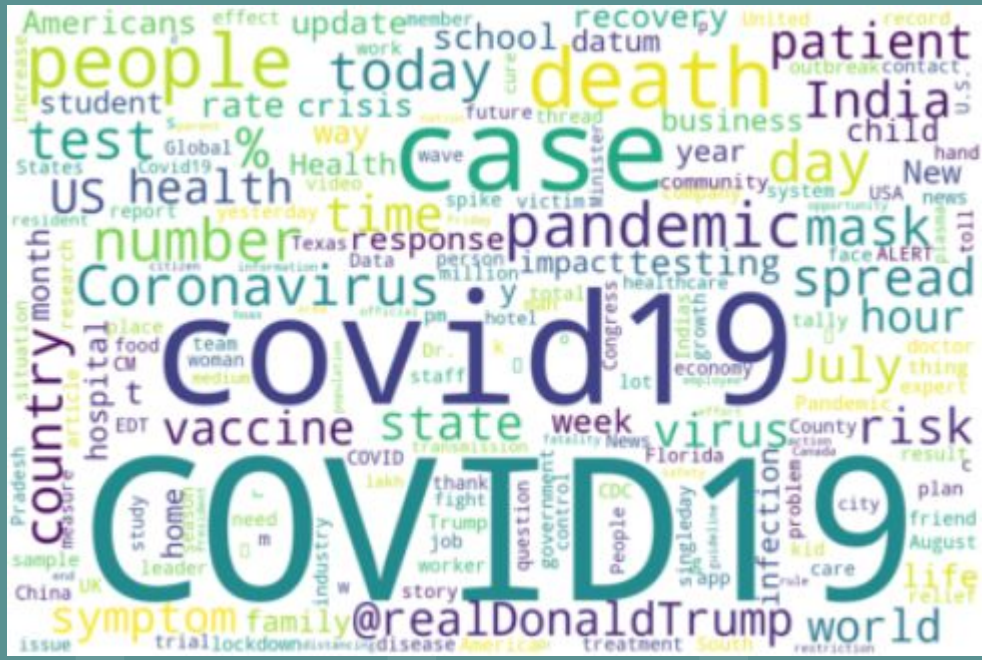


	user_locations	num_posts
count	21669.000000	21669.000000
mean	1.581891	3.435138
std	55.889327	202.654735
min	0.000000	1.000000
25%	1.000000	1.000000
50%	1.000000	1.000000
75%	1.000000	1.000000
max	5847.000000	21434.000000



Palabras más utilizadas



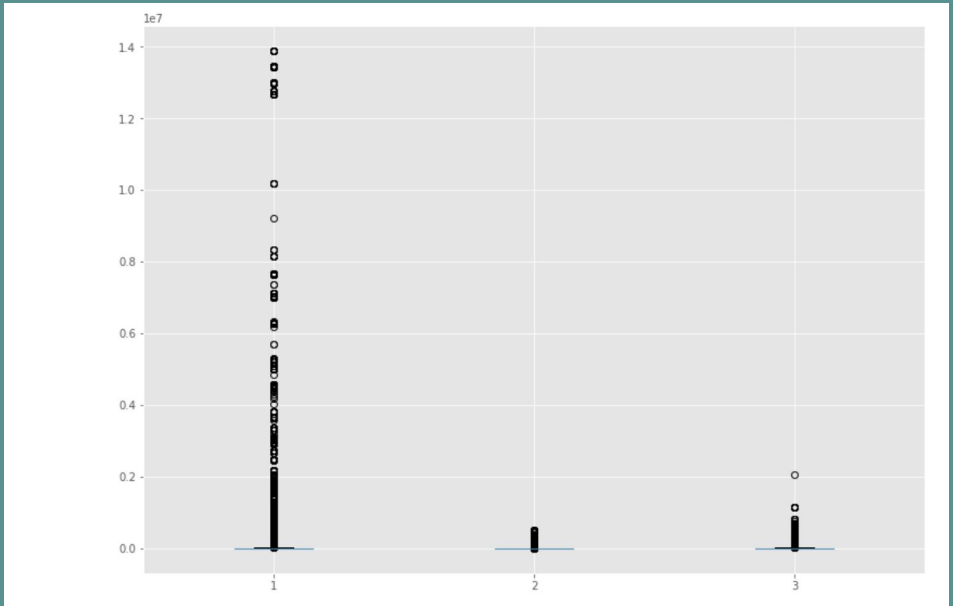


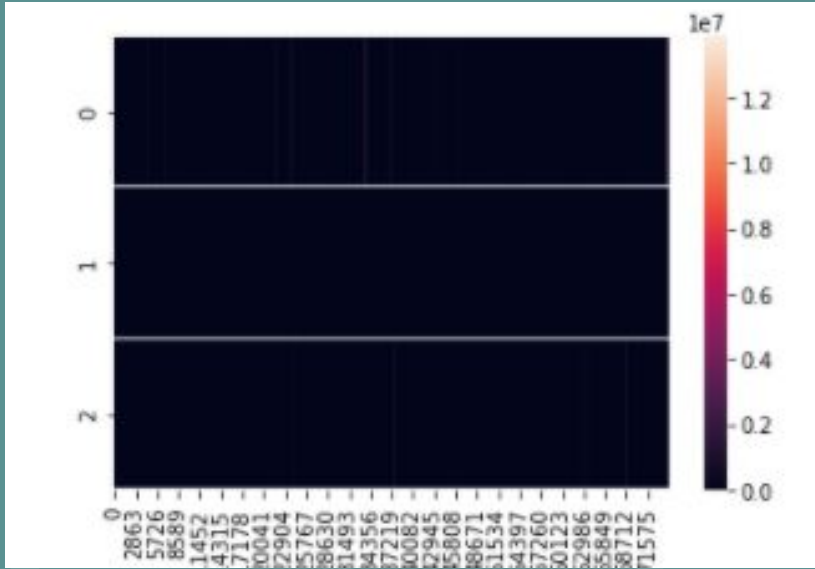
Text Clustering with K-Means

Clustering national anthems with unsupervised learning

Lucas de Sá Follow

Dec 17, 2019 · 8 min read

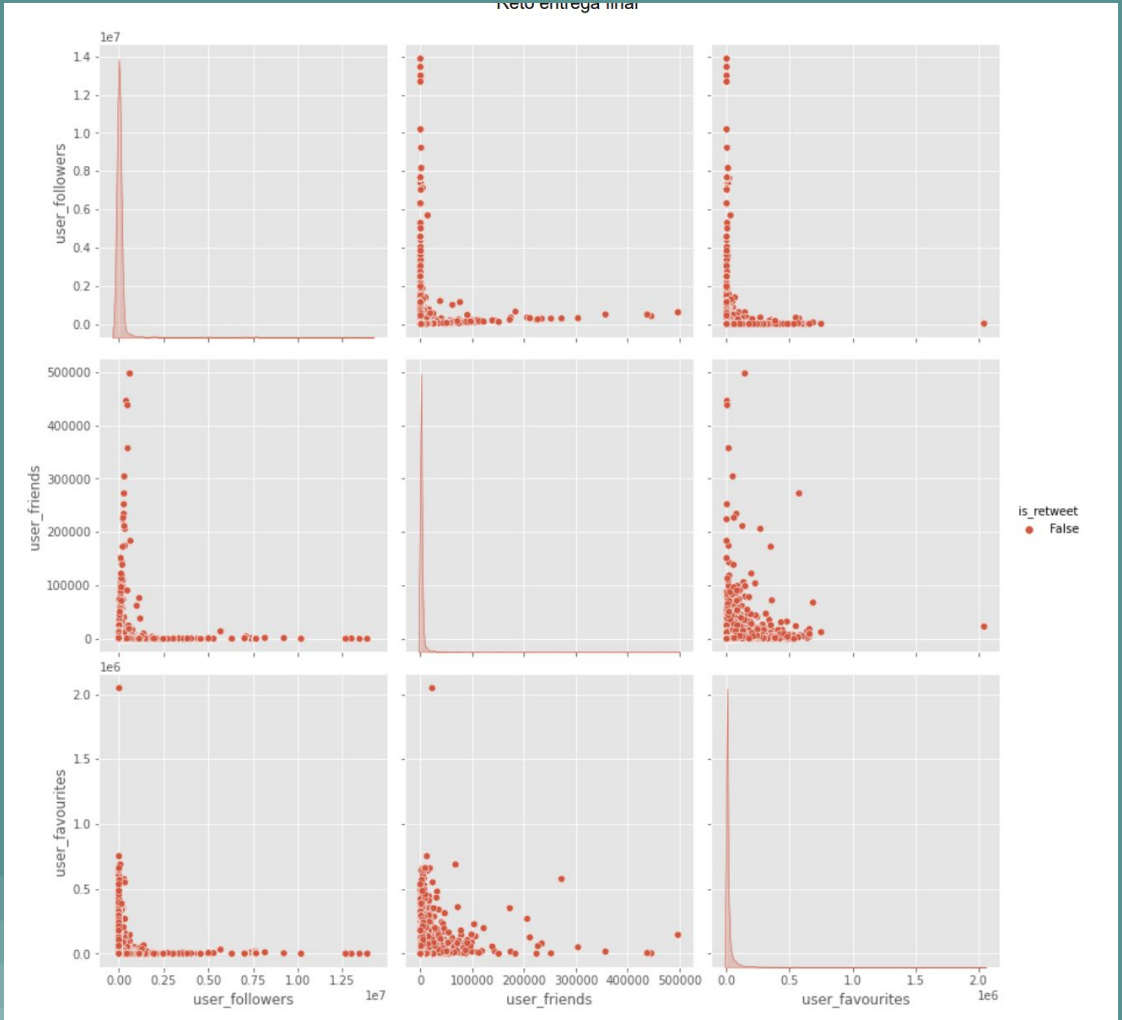

1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3


1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2

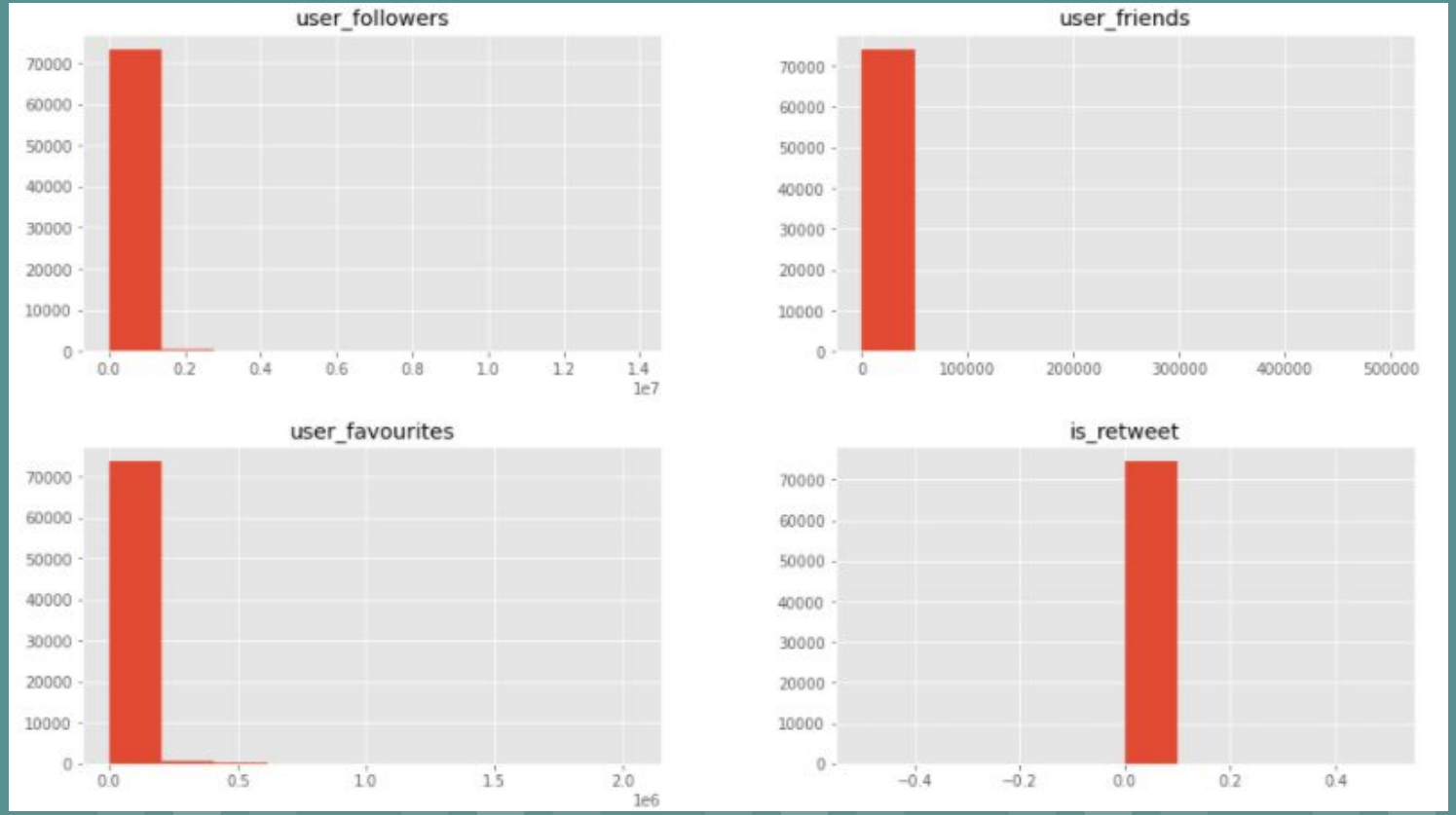
Preprocesamiento de datos

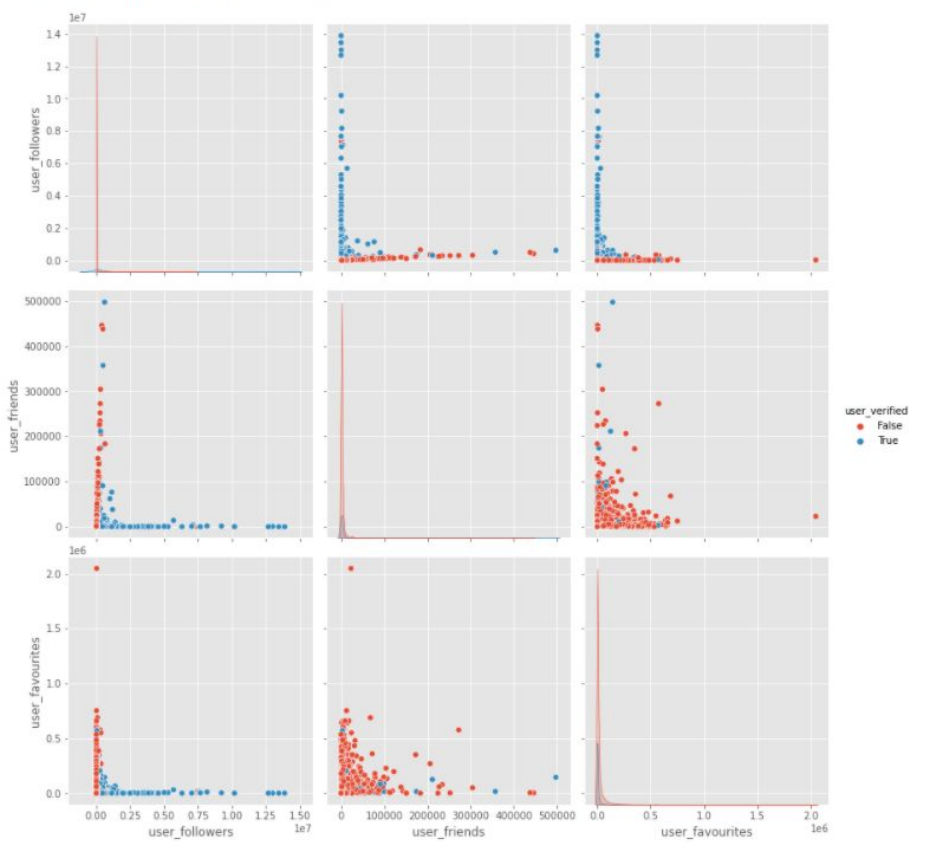
	user_followers	user_friends	user_favourites
count	7.443600e+04	74436.000000	7.443600e+04
mean	1.059513e+05	2154.721170	1.529747e+04
std	8.222900e+05	9365.587474	4.668971e+04
min	0.000000e+00	0.000000	0.000000e+00
25%	1.660000e+02	153.000000	2.200000e+02
50%	9.600000e+02	552.000000	1.927000e+03
75%	5.148000e+03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

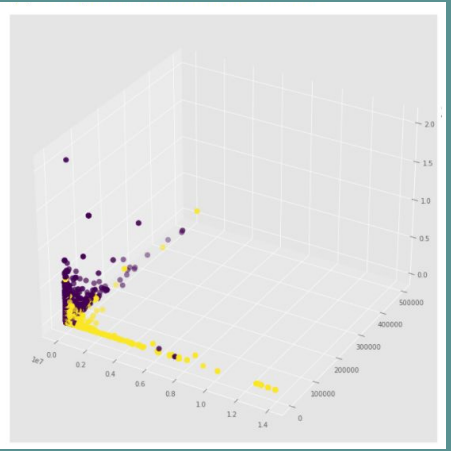
```
Out[4]: Index(['user_name', 'user_location', 'user_description', 'user_created',
               'user_followers', 'user_friends', 'user_favourites', 'user_verified',
               'date', 'text', 'hashtags', 'source', 'is_retweet'],
              dtype='object')
```

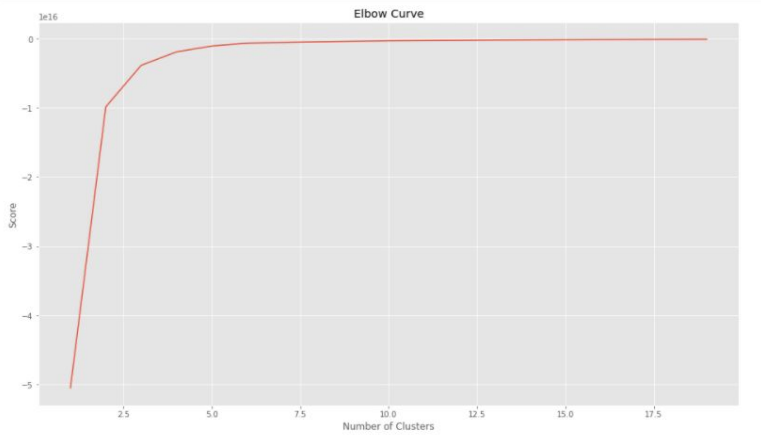


Hipótesis: is_retweet

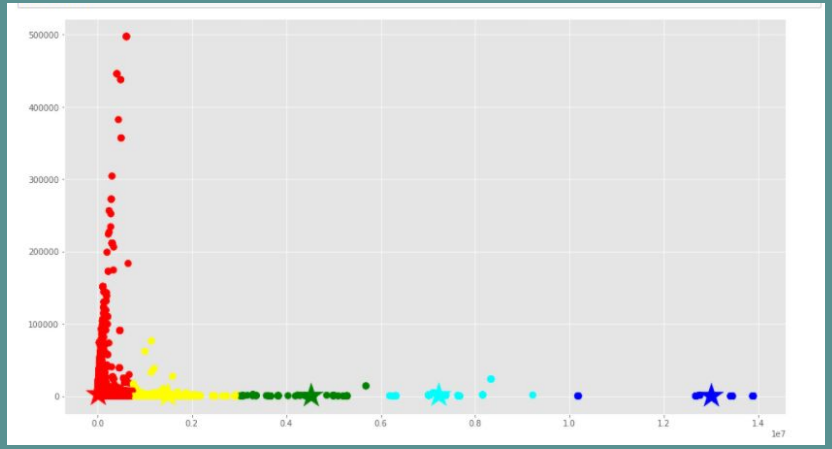

user_followers
70000
60000
50000
40000
30000
20000
10000
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e7
user_friends
70000
60000
50000
40000
30000
20000
10000
0
0
100000
200000
300000
400000
500000
user_favourites
70000
60000
50000
40000
30000
20000
10000
0
0.0
0.5
1.0
1.5
2.0
1e6
is_retweet
70000
60000
50000
40000
30000
20000
10000
0
−0.4
−0.2
0.0
0.2
0.4

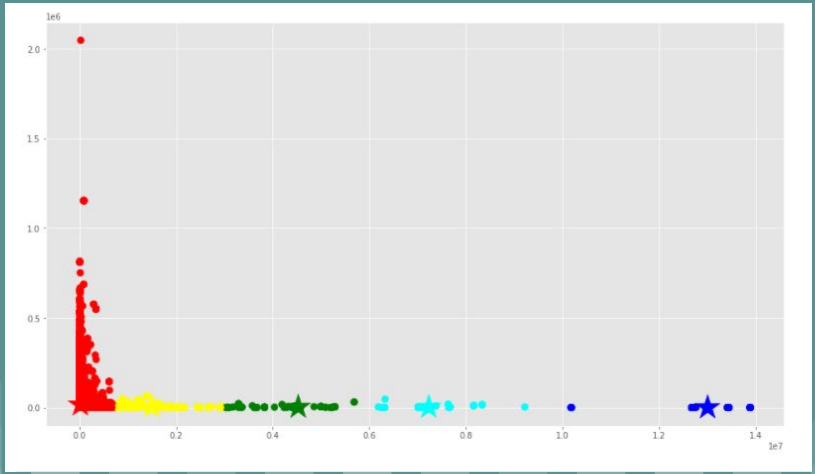

user_followers
user_friends
user_favourites
user_verified
False
True
1e7
1e6

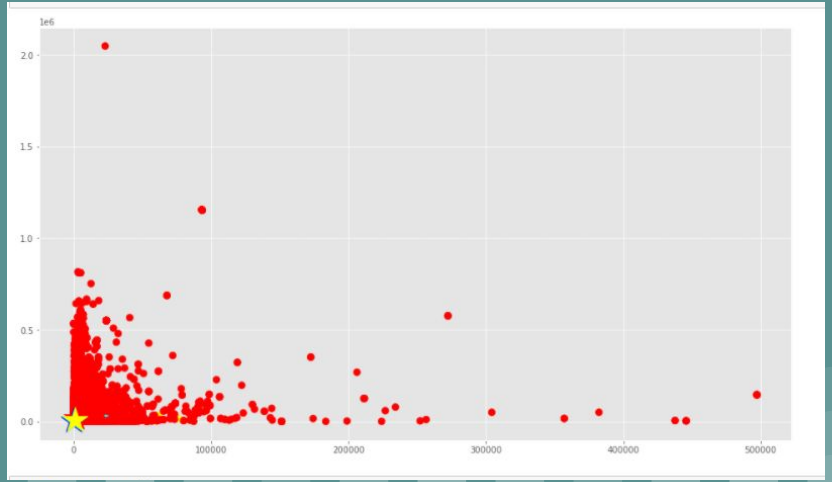



Elbow Curve
Score
Number of Clusters
1e16

user_verified	cantidad
False	65069
True	7828







Resultados obtenidos y Conclusiones