Herramientas compuatacionales: El arte de la analítica

Reto

Pablo Sánchez Aguirre A01662244
Carolina González Salinas A01662120

Obteniendo los datos.

Originalmente así se veían las columnas del archivo csv.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	user_name	user_location	user_descripti	user_created	user_follower	user_friends	user_favourit	user_verified	date	text	hashtags	source	is_retweet
2	á‰áŽ¥â˜»Õ¬	astroworld	wednesday ac	##########	624	950	18775	False	##########	If I smelled the scent of han		Twitter for iPh	False
3	Tom Basile ðŸ	New York, NY	Husband, Fath	##########	2253	1677	24	True	##########	Hey @Yankees @YankeesPR		Twitter for An	False
4	Time4fisticuff	Pewee Valley,	#Christian #Ca	##########	9275	9525	7254	False	##########	@diane3443 (	['COVID19']	Twitter for An	False
5	ethel mertz	Stuck in the M	#Browns #Ind	##########	197	987	1488	False	##########	@brookbankt	['COVID19']	Twitter for iPh	False
6	DIPR-J&K	Jammu and Ka	ðŸ–Šï¸Official	##########	101009	168	101	False	##########	25 July :	['CoronaVirus	Twitter for An	False
7	ðŸŽ¹ Franz Sch	ÐÐ¾Ð²Ð¾Ñ€Ð	ðŸŽ¼ #ÐÐ¾Ð	##########	1180	1071	1287	False	##########	#coronavirus	['coronavirus'	Twitter Web A	False
8	hr bartender	Gainesville, FL	Workplace tip	##########	79956	54810	3801	False	##########	How #COVID1	['COVID19', 'R	Buffer	False
9	Derbyshire LPC			##########	608	355	95	False	##########	You now have to wear face		TweetDeck	False
10	Prathamesh Bendre		A poet, reiki p	##########	25	29	18	False	##########	Praying for	['covid19', 'co	Twitter for An	False
11	Member of Cl	ðŸ‘‡ðŸ»locati	Just as the bo	##########	55201	34239	29802	False	##########	POPE AS	['HurricaneHa	Twitter for iPh	False
12	Voice Of CBSE Students			##########	8	10	7	False	##########	49K+		Twitter Web A	False
13	Creativegms	Dhaka,Bangla	I'm Motalib	##########	241	1694	8443	False	##########	Order here:	['logo', 'graph	Twitter Web A	False
14	SEXXYLYPPS	Hotel living - v	My ink "My	##########	0	8	32	False	##########	ðŸ‘‰ðŸ»@Patt	['COVID19']	Twitter Web A	False
15	Africa Youth A	Africa	Official accou	##########	830	254	3692	False	##########	Let's all	['COVID19']	Twitter Web A	False

Obteniendo los datos.

Lo primero que hicimos fue evaluar las variables y ajustar los datos.

```
[22]
datos['user_created'] = pd.to_datetime(datos['user_created'])
datos['user_created'] = datos['user_created'].dt.date

datos['date'] = pd.to_datetime(datos['date'])
datos['date'] = datos['date'].dt.date
```

Creamos tres columnas extras: dos para tener a las fechas como enteros y poder utilizarlas aún más y una para tener de forma numérica a la columna de 'user_verified'.

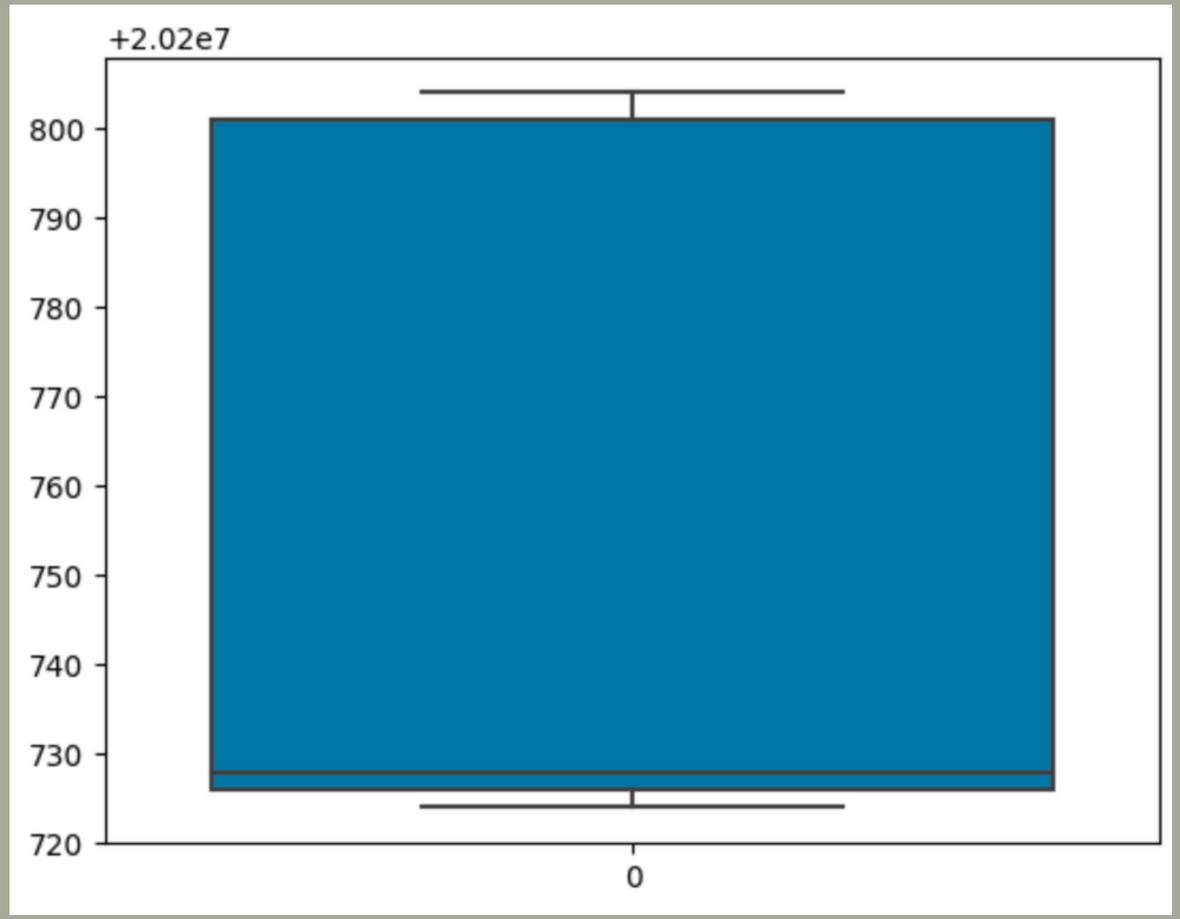
```
def fecha_a_numero(fecha):
    return int(fecha.strftime('%Y%m%d'))

def user_verified_num(booleano):
    if(booleano == True):
        return 1
    else:
        return 0

datos['user_created_entero'] = datos['user_created'].apply(fecha_a_numero)
datos['date_entero'] =  datos['date'].apply(fecha_a_numero)
datos['user_verified_entero'] = datos['user_verified'].apply(user_verified_num)
```

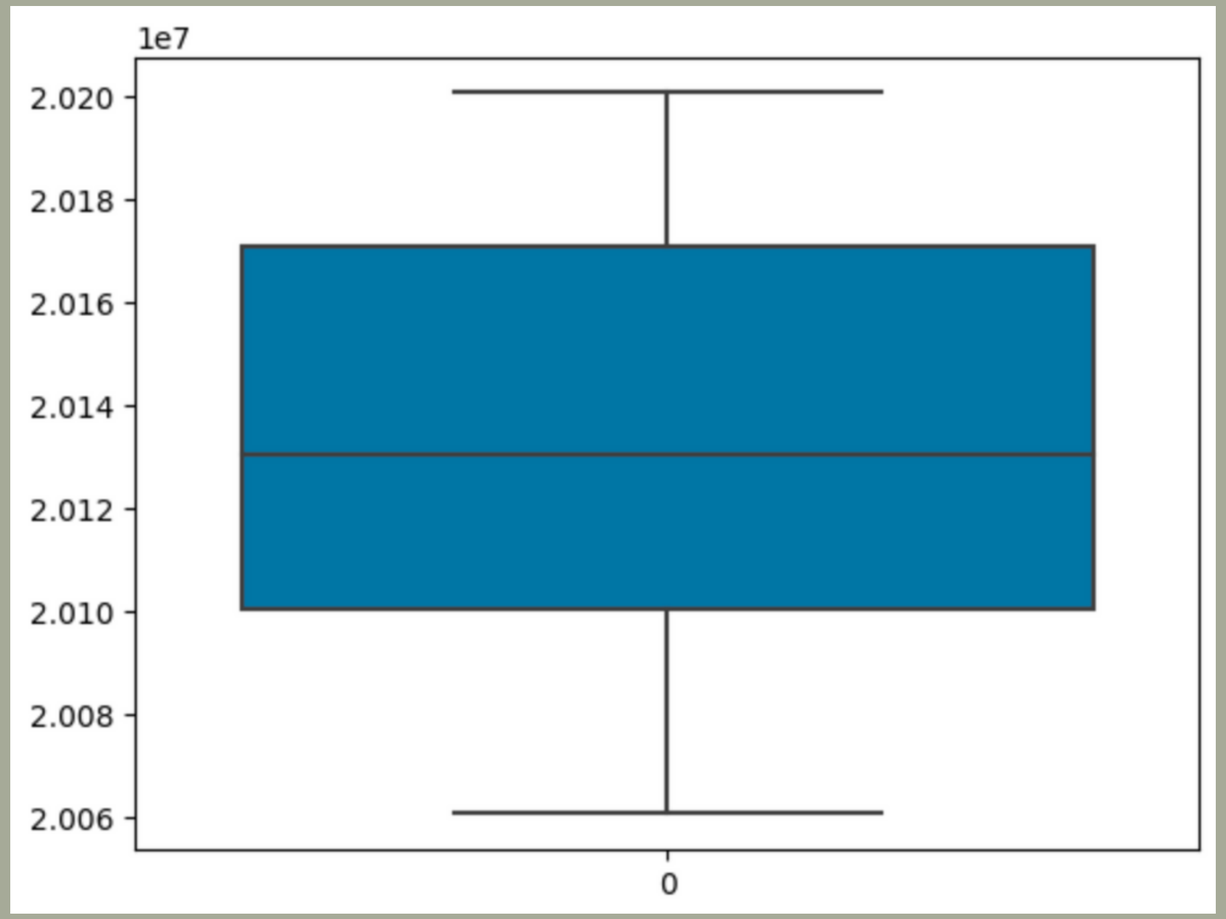
Analizando los datos

Fechas de subida de los tweets en números enteros.
Aquí podemos ver que la mayoría de las fechas se encuentran después de la mitad pero antes del tercer cuarto.



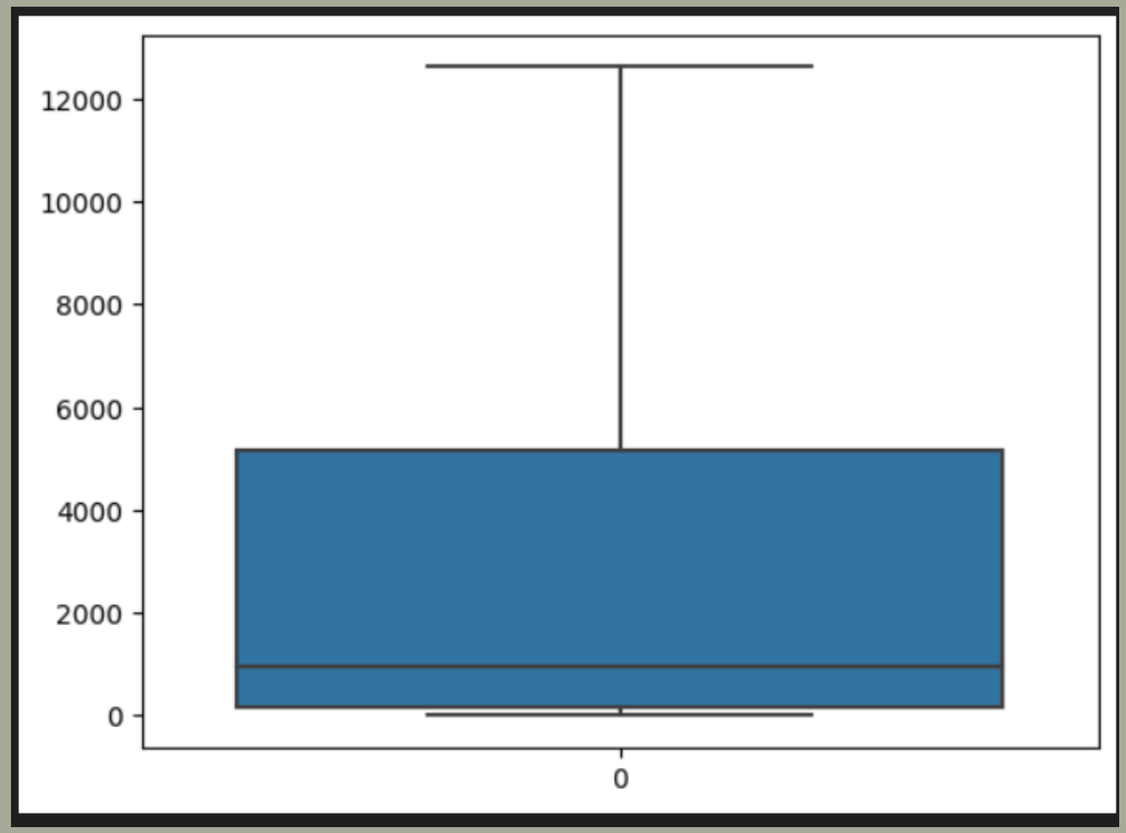
Analizando los datos

Fechas de creación de usuario con números enteros.
Aquí tenemos nuestro boxplot más balanceado. Las fechas se distribuyen casi de forma normal.



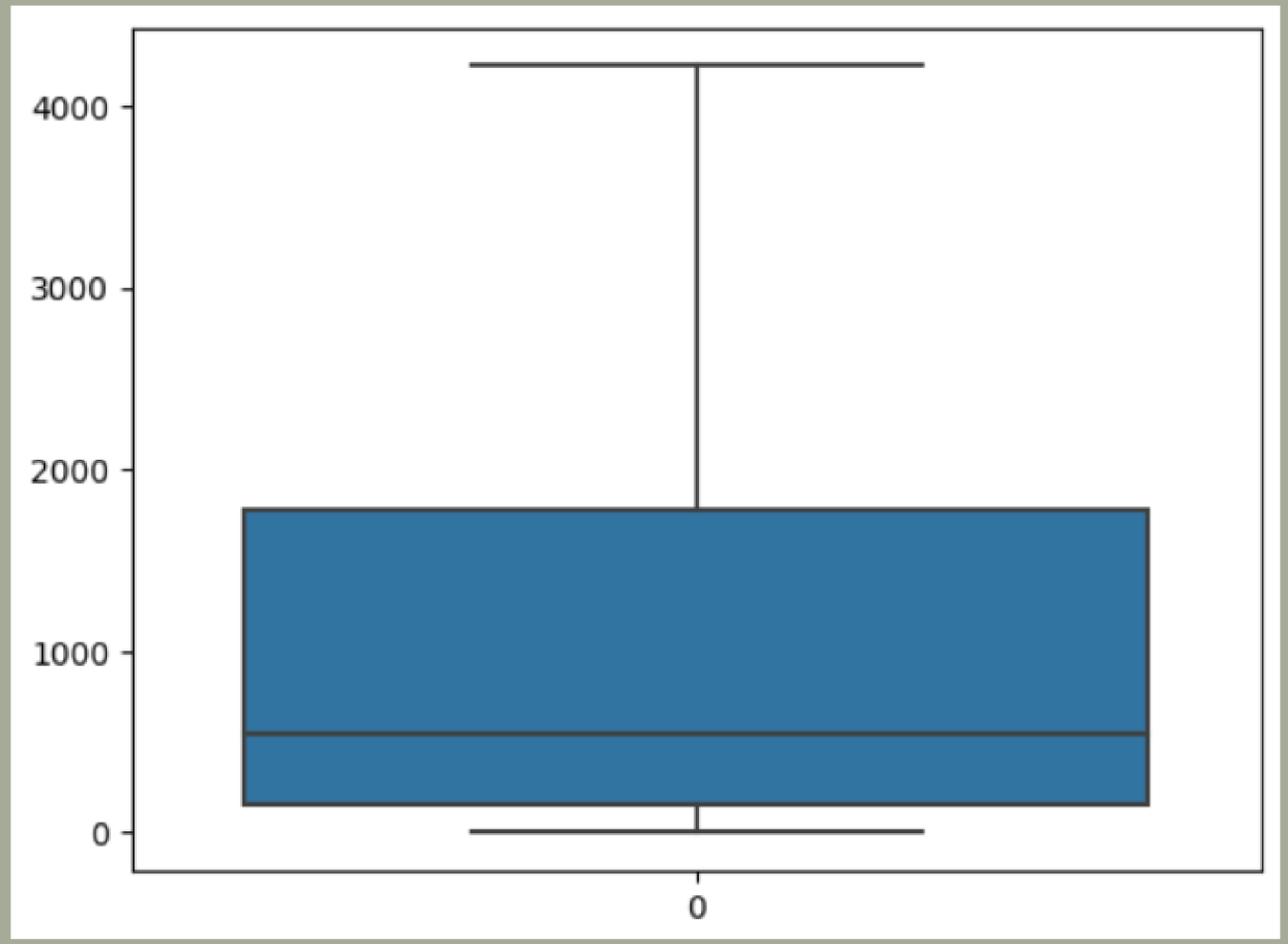
Analizando los datos

Cantidad de seguidores de cada usuario.
Se puede observar que la gran mayoría de los usuarios tiene menos seguidores que los pocos usuarios con más seguidores.



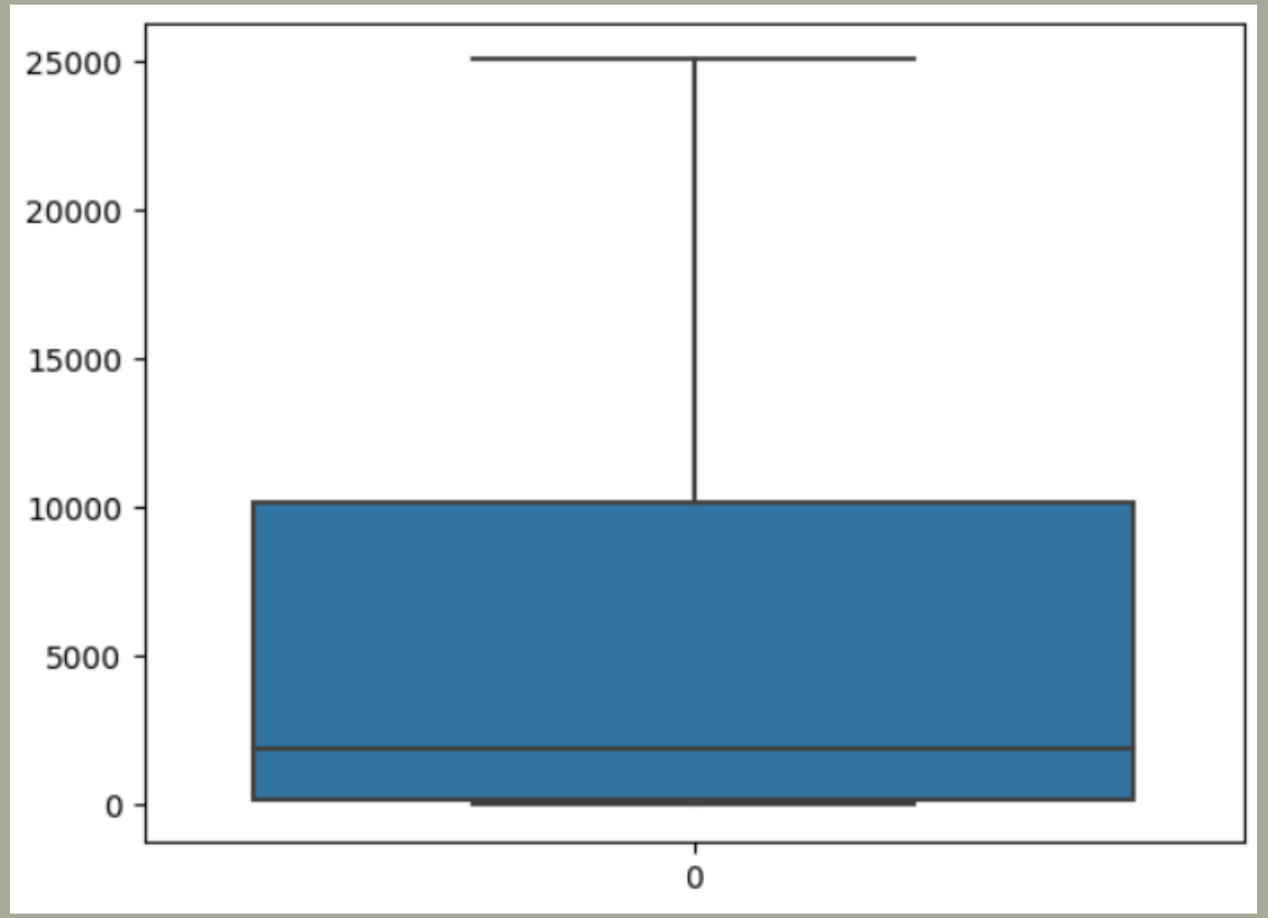
Analizando los datos

Cantidad de amigos de cada usuario.
Se observa que se comporta como la gráfica de la diapositiva anterior.



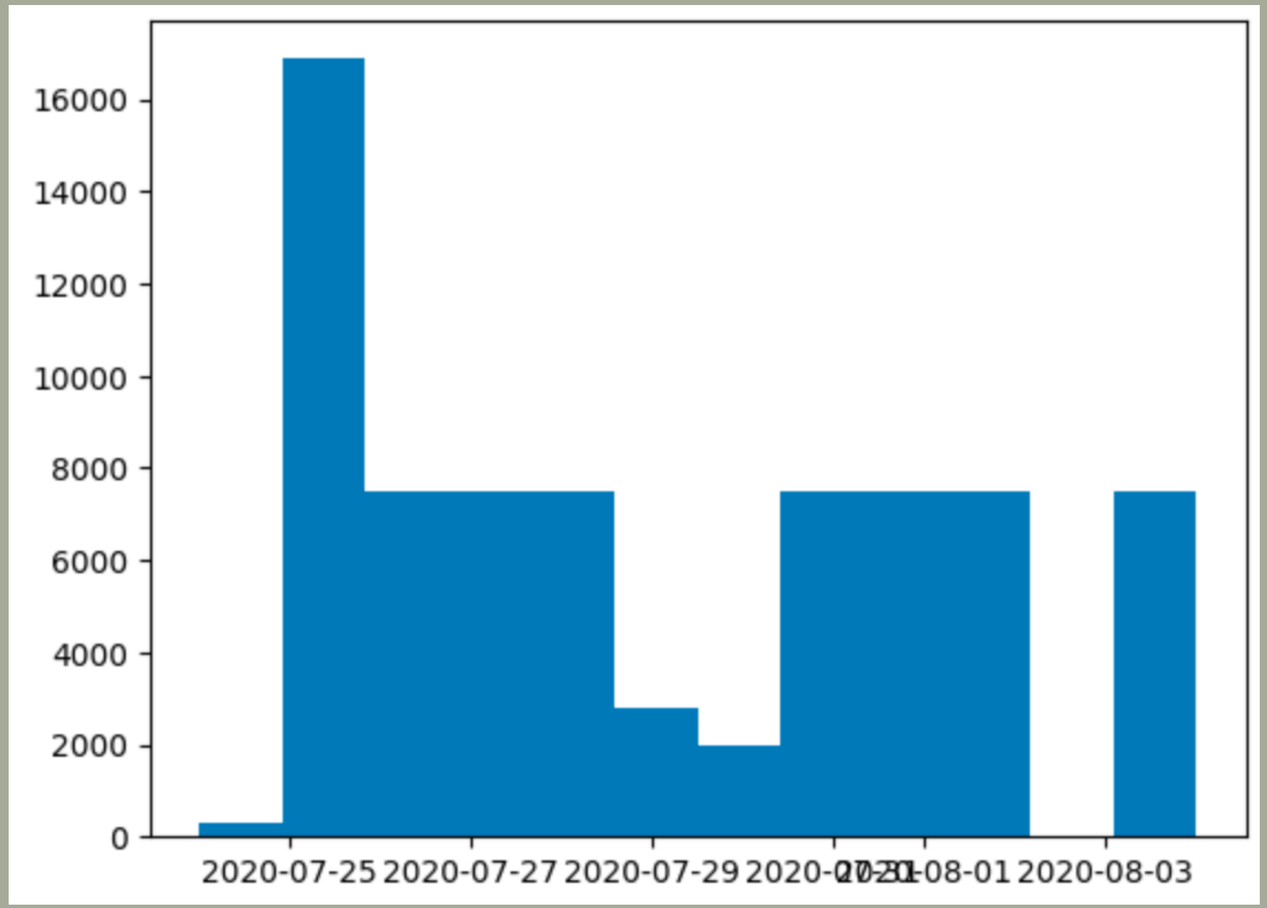
Analizando los datos

Cantidad de favoritos de un usuario. Se distribuyen de forma similar a las dos gráficas anteriores.



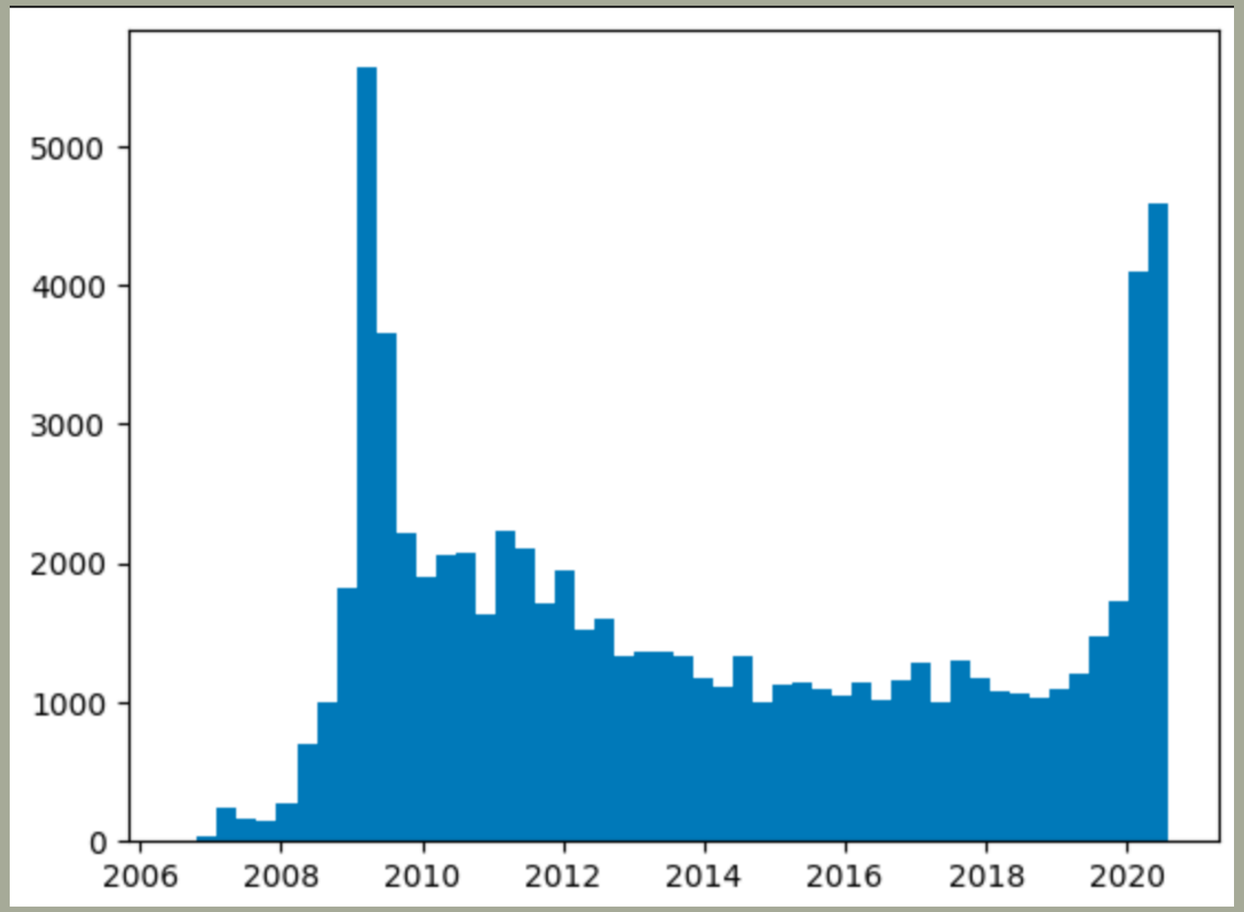
Analizando los datos

Histograma de fechas de subida de tweets.
Tenemos un pico en una fecha y 4 fechas en las que hubo menos tweets.



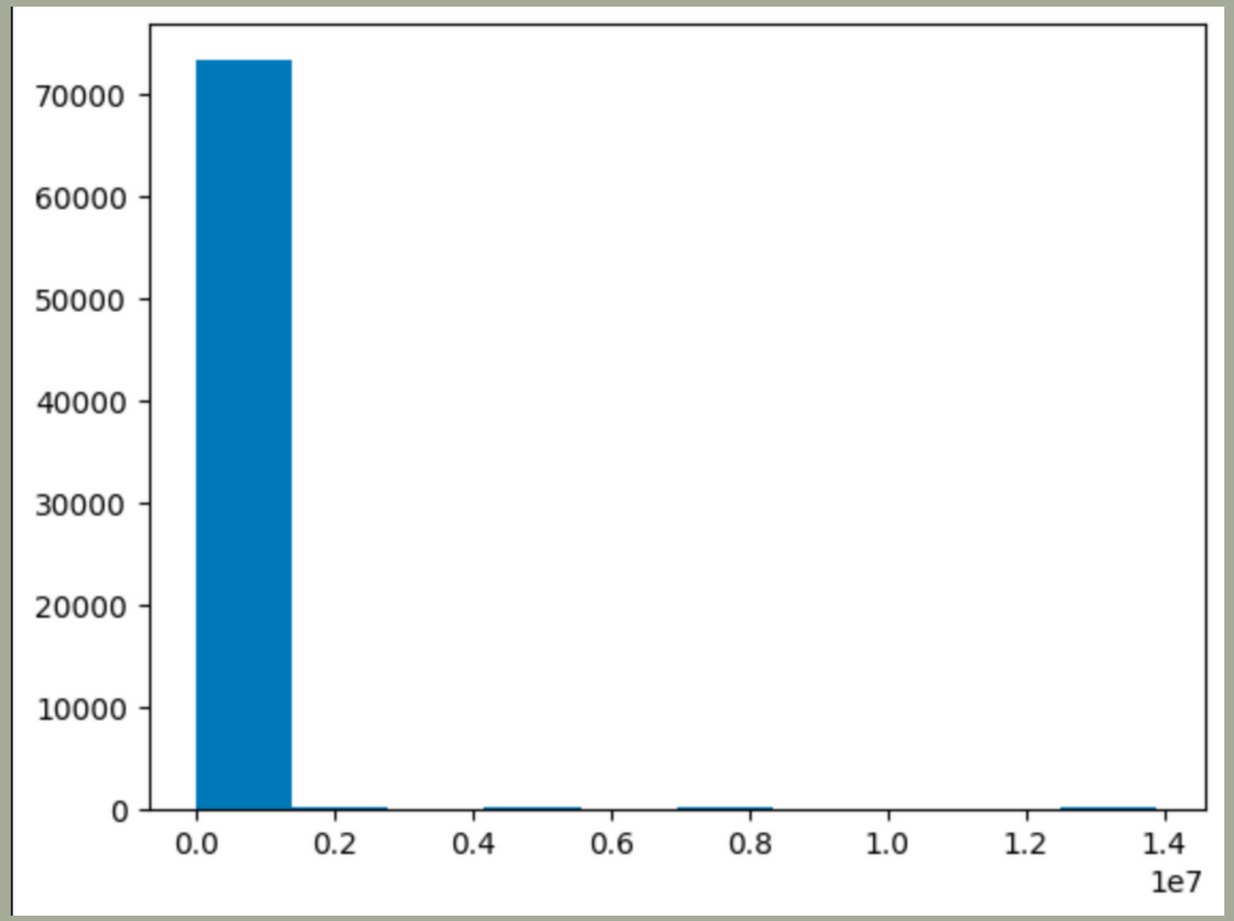
Analizando los datos

Histograma con las fechas de creación de usuario. Podemos ver que hay dos picos y en general la distribución es más o menos uniforme.



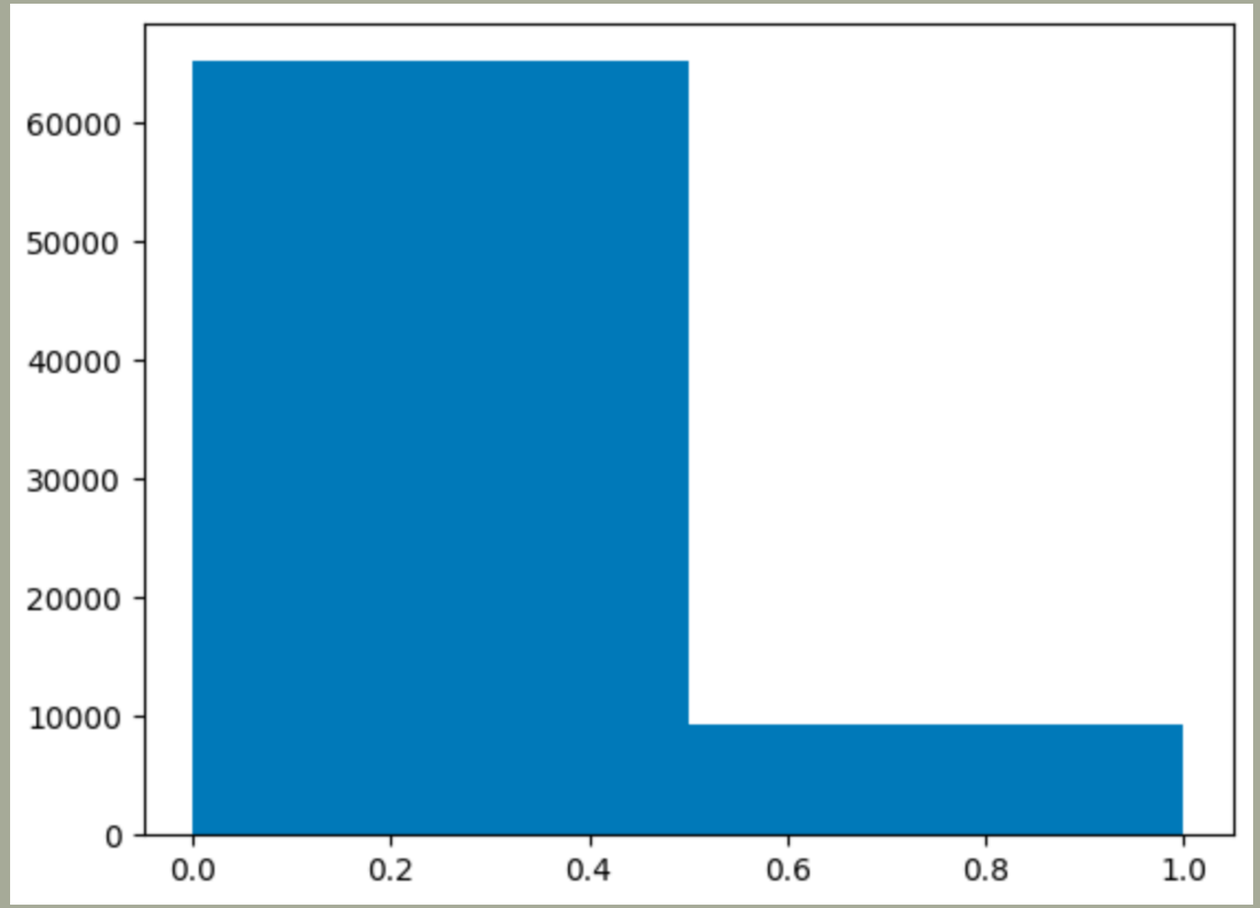
Analizando los datos

Histograma de la función:
“user_followers”
No se puede obtener mucha información de esta gráfica.



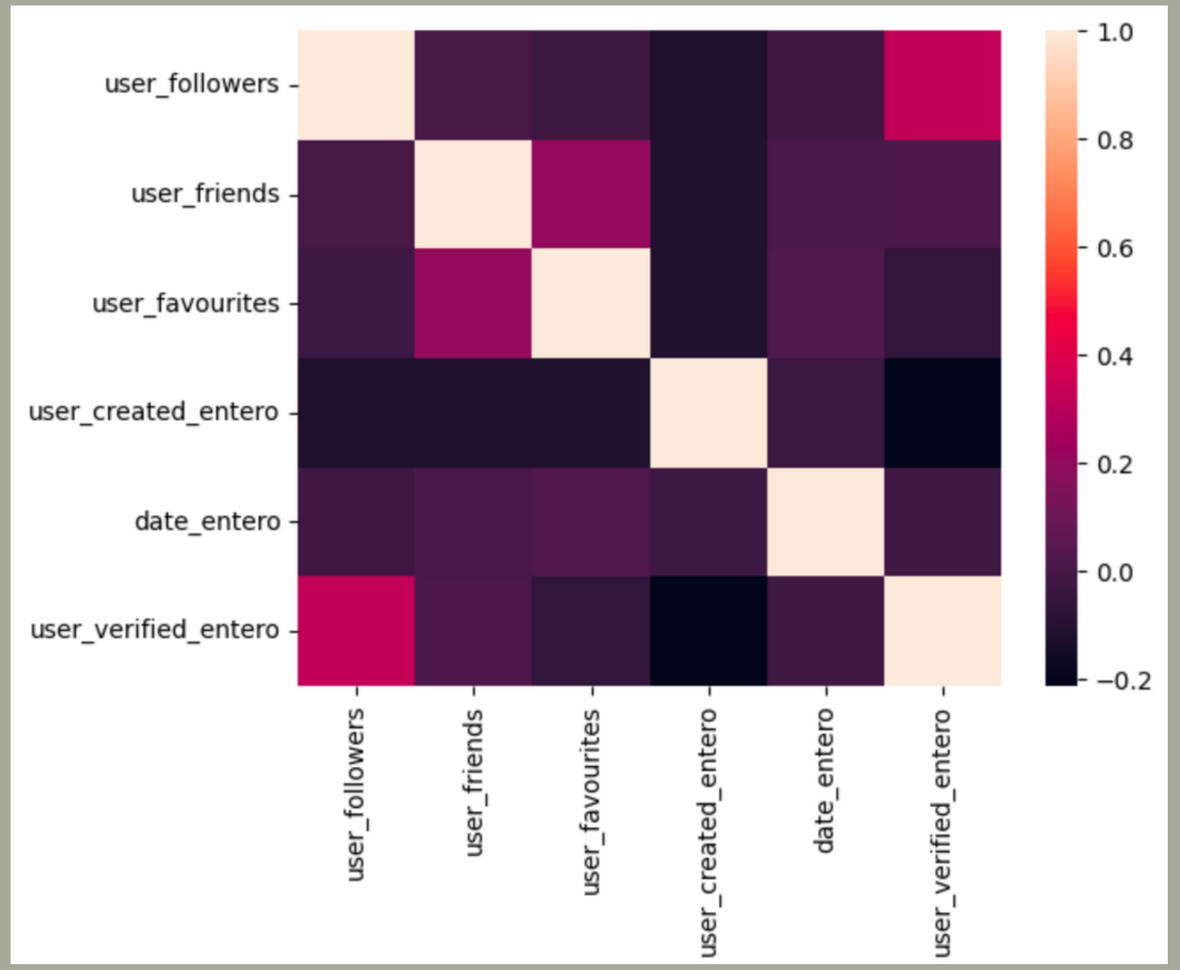
Analizando los datos

Histograma de la función “users_verified”.
Notamos que la mayoría de los usuarios no están verificados.



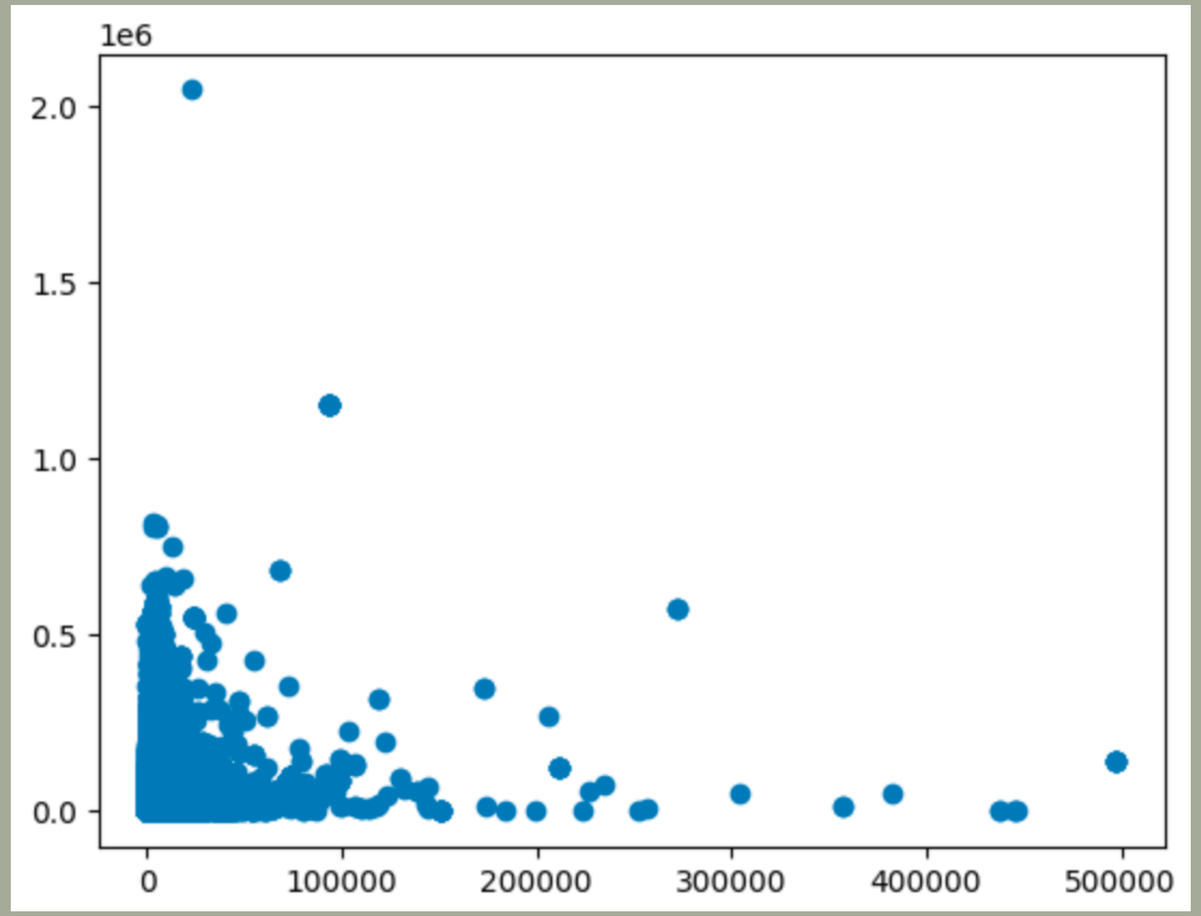
Analizando los datos

Mapa de calor que muestra que la correlación más grande es entre la cantidad de seguidores y si un usuario está o no verificado.



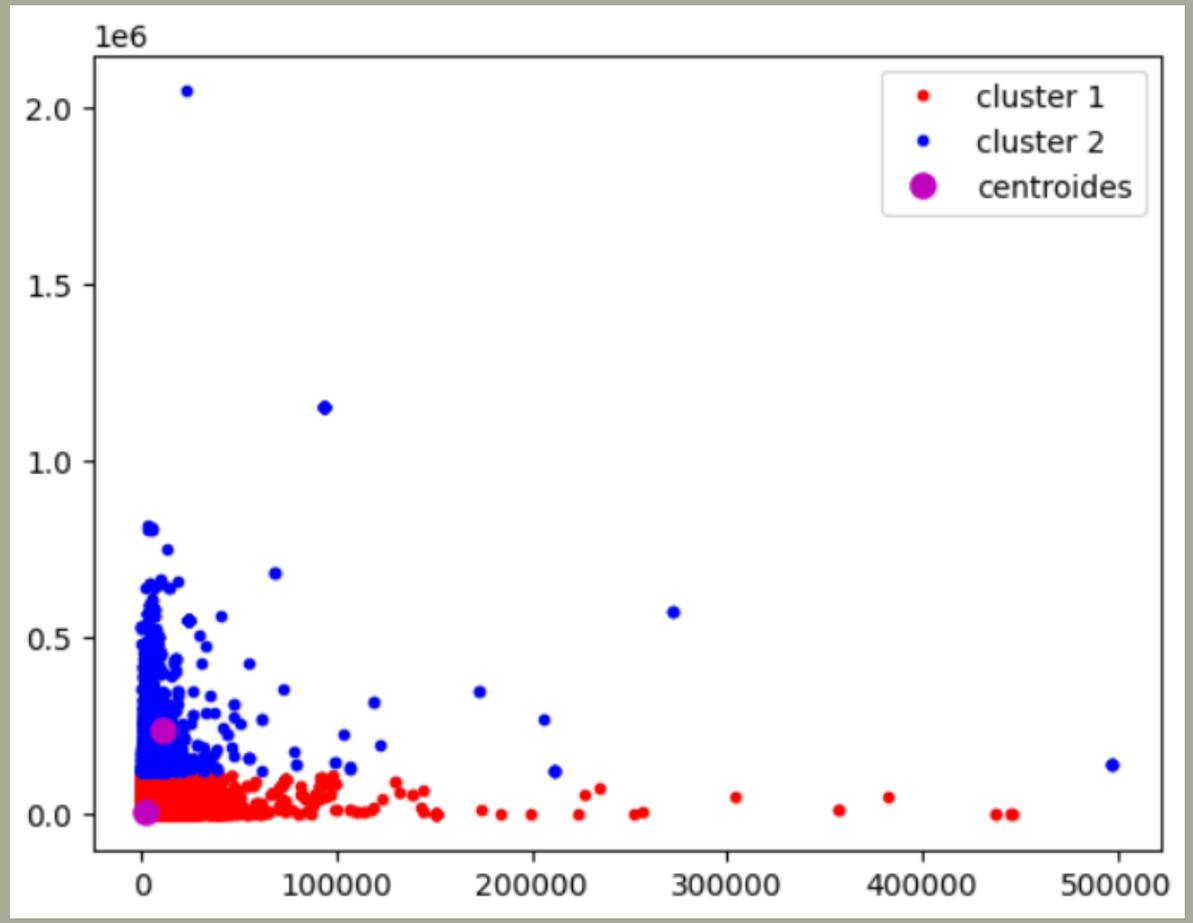
Analizando los datos

Gráfica de user_friends contra user_favourites. Son las dos variables que más correlación tuvieron, aunque fue una correlación baja.



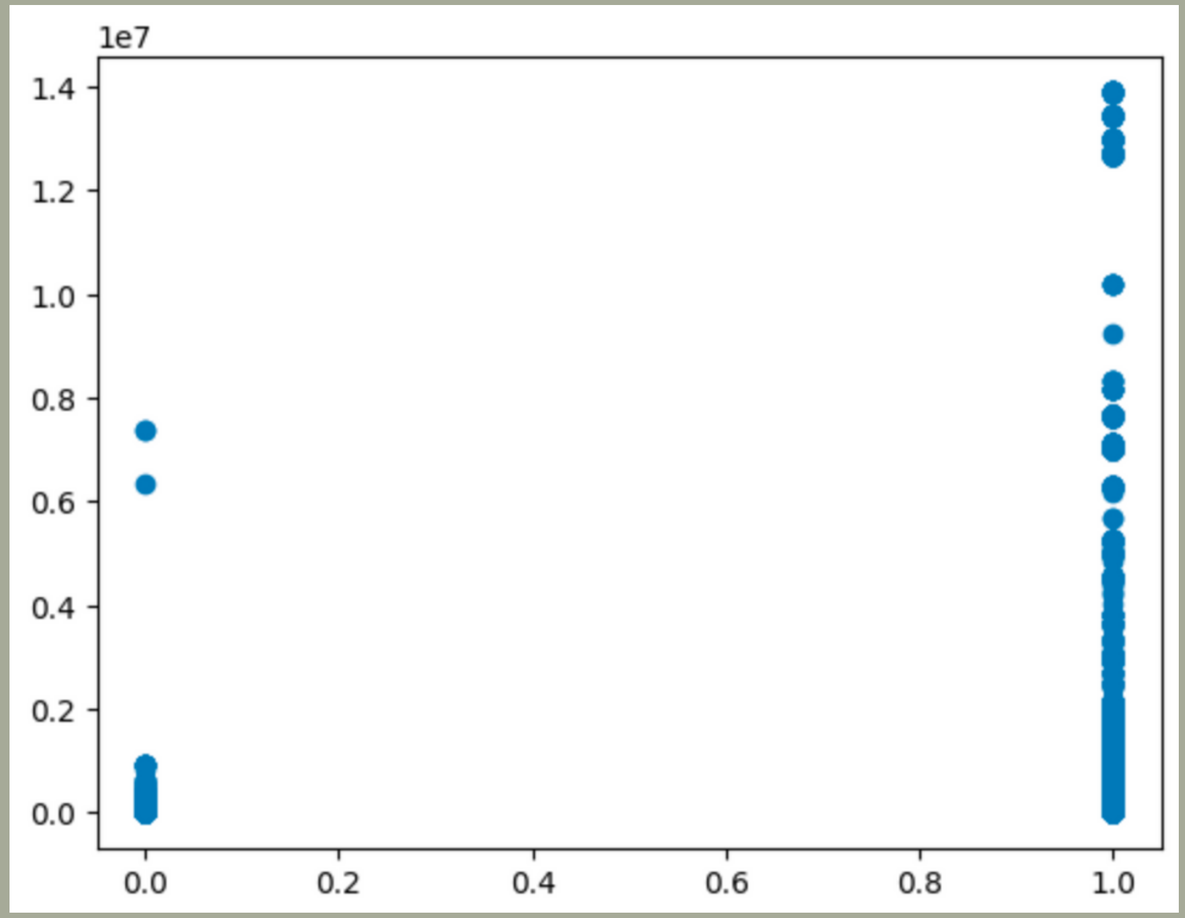
Analizando los datos

Utilizamos k = 2 pero en realidad se nota que únicamente hay un cluster claro.



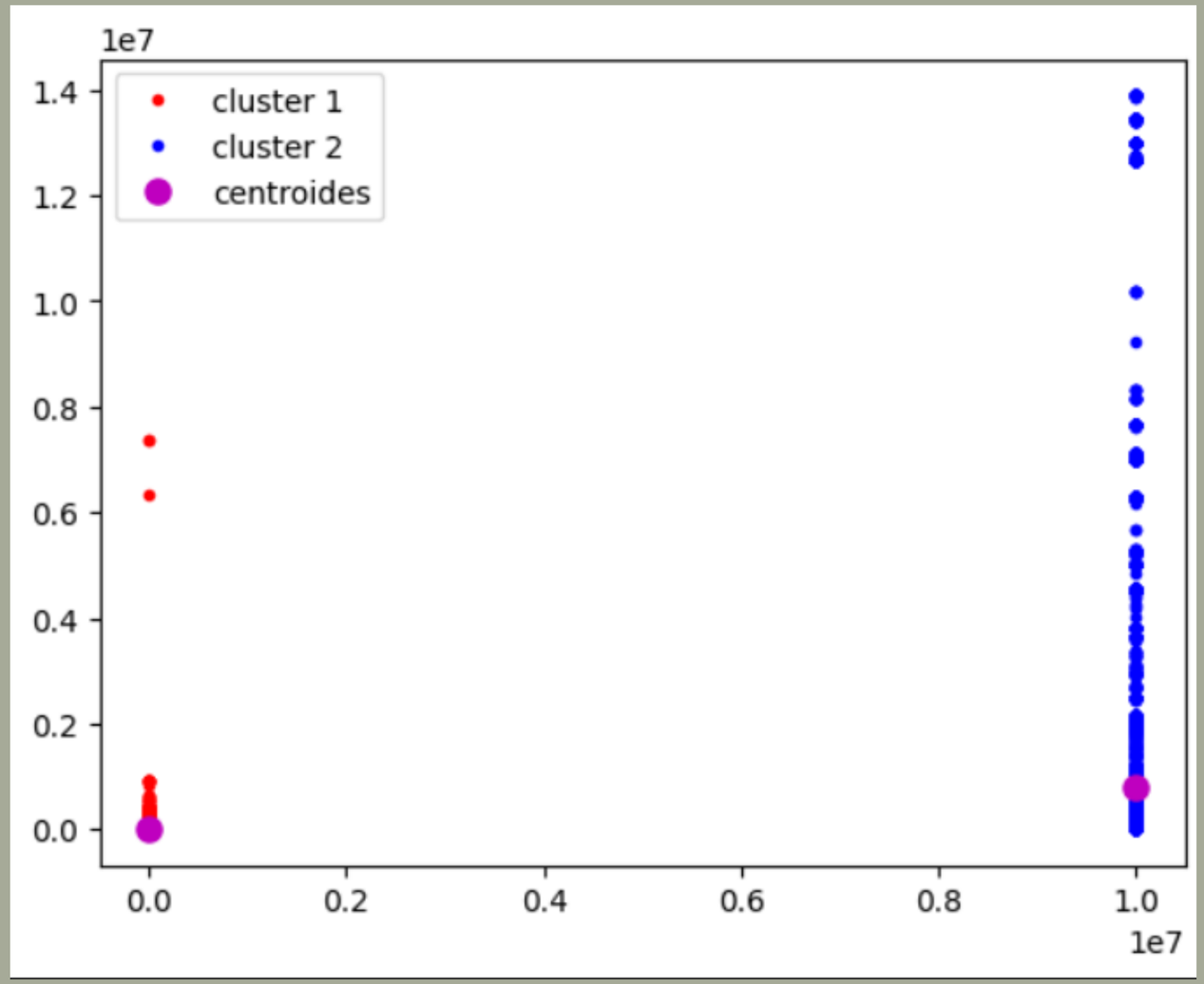
Analizando los datos

Gráfica de user_verified contra user_followers.



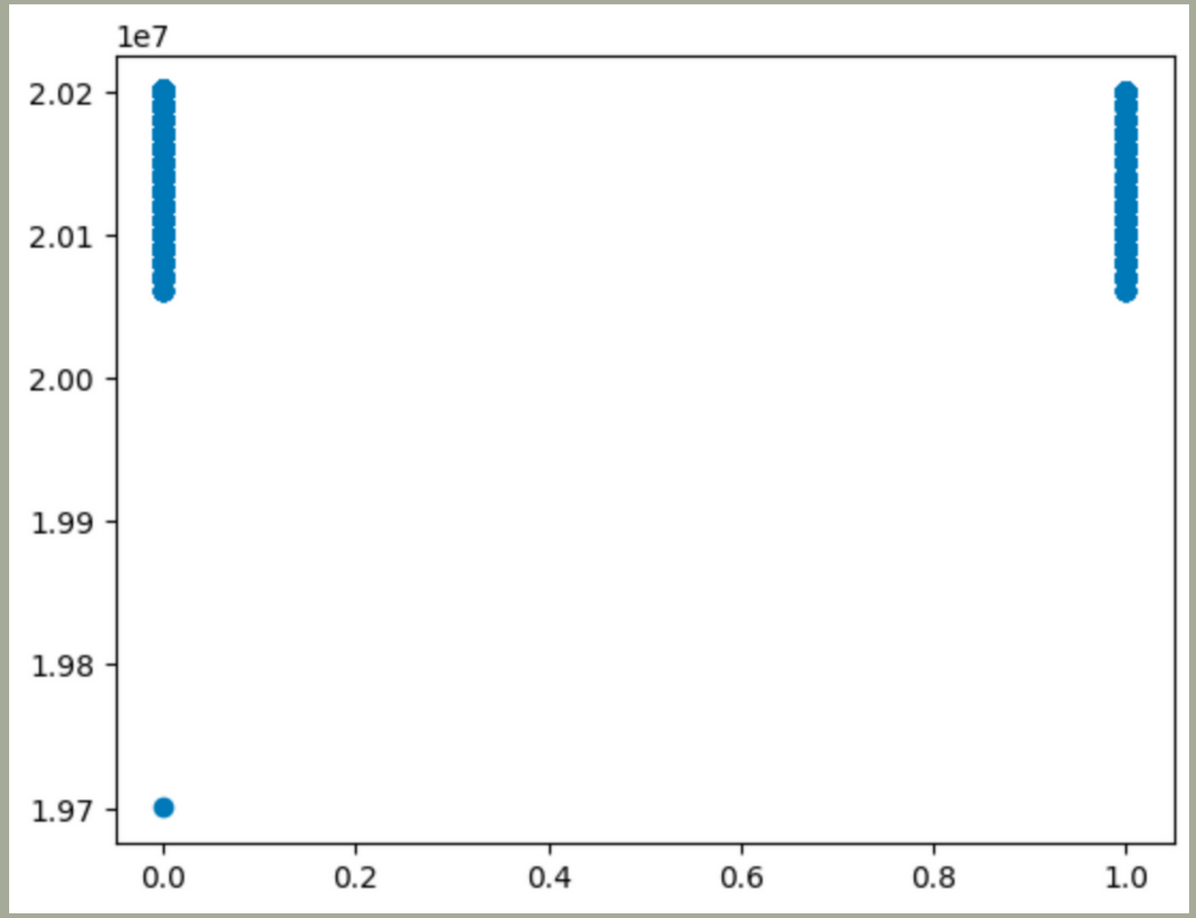
Analizando los datos

Gráfica de user_verified contra user_followers. Tuvimos dos centros claramente distinguibles.



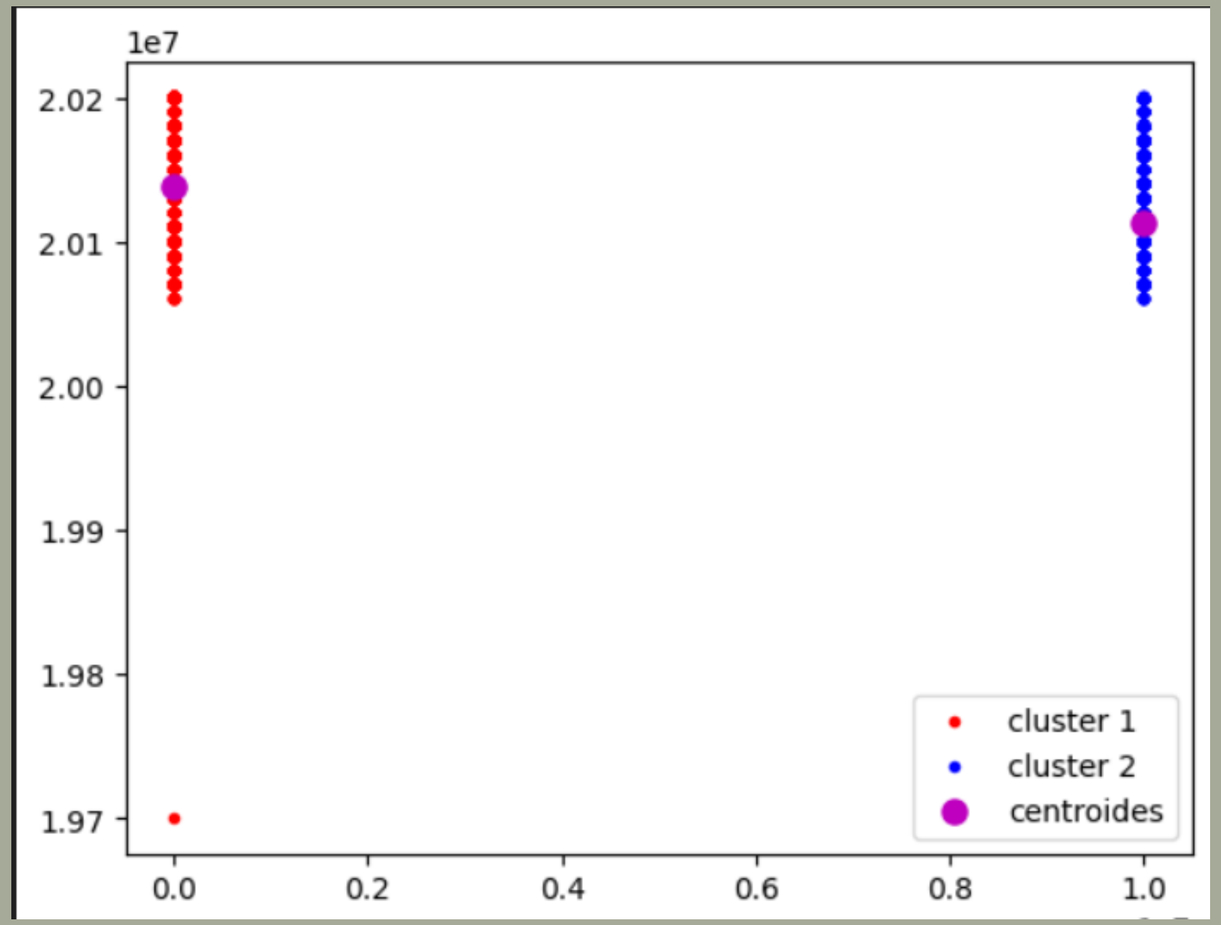
Analizando los datos

Gráfica de user_verified contra user_created.



Analizando los datos

Gráfica de user_verified contra user_followers. Tuvimos dos centros claramente distinguibles.



Preguntas a responder:

¿Hay alguna variable que no aporta información?
Sí, las variables user_name, user_location y user_description. Son personalizables y cada usuario pone lo que quiera en ellas. También la variable is_retweet, dado que siempre es False.

Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?
Quitaríamos user_name, user_location, user_description y is_retweet por lo mencionado en la pregunta anterior. También quitaríamos text porque tendríamos que procesar muchísimos textos distintos y hashtags por lo mismo. De igual forma quitaríamos source debido a que no aporta información que vayamos a utilizar.

¿Existen variables que tengan datos extraños?
Sí, las variables user_name, user_location y user_description. Son personalizables y cada usuario pone lo que quiera en ellas, incluyendo símbolos extraños y caracteres que no reconoce el software que estamos utilizando.

Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?
Hay distintos rangos. Los rangos de fechas son exclusivos para las fechas. Los rangos de las variables user_followers, user_friends y user_favourites son similares porque miden datos similares, dados por interacciones de personas. Los valores que puede tomar user_verified_entero son {0,1}, lo cual lo hace único. Sí va a afectar al análisis de los datos, cuando tomamos en cuenta los centros usando kmeans. Si una variable consta de 0 y 1 y otra variable tiene datos que se encuentran en los millones, la variable menor va a tener muy poca influencia al momento de hallar los centros. Es por esto que en la actividad siguiente, de los kmeans, haremos que el orden de magnitud de ambas variables sea el mismo, para que ambas influyan de igual forma al encontrar los centros.

¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?
Sí, los grupos de fechas son los que mejor se vieron reflejados en histogramas, ya que los demás grupos tenían muchos datos similares y pocos datos muy grandes o únicamente dos posibles valores. Las fechas tienen diagramas de bigotes centrados, mientras que user_followers, user_friends y user_favourites tienne diagramas de bigotes más bien inclinados a estar hacia abajo del centro.

Preguntas a responder:

¿Crees que estos centros pueden ser representativos de los datos? ¿Por qué?
Los de la primera gráfica no tanto, pues se nota una acumulación de puntos en una única zona, es decir, un único cluster de datos.
Los de las otras dos gráficas sí, se ve claramente que hay una acumulación de datos en dos lugares exactamente.

¿Cómo obtuviste el valor de k a usar?
Vimos las gráficas que generamos y notamos dónde había clusters de datos.

¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?
Es muy situacional y depende de los datos. En estas gráficas en particular, menos centros fue más representativo. En gráficas en la que se noten varios clusters, serán más representativos más centros.

¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?
En la primera gráfica están bastante cercanos, mientras que en las otras dos gráficas si se encuentran más alejados.

¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?
Los centros estarían más alejados de los datos, pues los datos estarían más dispersos.

¿Qué puedes decir de los datos basándote en los centros?
En la primera gráfica podemos ver que los centros están muy juntos, entonces la acumulación de datos es en realidad en un sólo cluster. Los datos que están lejos de los centros están dispersos y no forman ningún cluster.
En las otras dos gráficas, podemos notar una clara acumulación de datos en dos centros, debido a que tenemos una variable que sólo puede tener dos valores y una acumulación de datos en cada uno de sus valores.

Muchas Gracias