Obtención de estadísticas descriptivas

In [84]:

```
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
```

1. Lectura de CSV

In [85]:

```
def readData():
    data = pd.read_csv("covid19_tweets.csv")
    return data

df = readData()
df
```

Out[85]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_fa
0	Ꮙi☻⌊ℭ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	
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	
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	
4	DIPR-J&K	Jammu and Kashmir	🖊️ Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	
...	...	...	...	...	...	...	
74431	Laura Wolfrom	Lexington, KY	The only things I collect are memories.	2010-09-24 02:01:15	85	586	
74432	Professor Tonya M. Evans	😷 #stayathome	Law Prof @DickinsonLaw & Entrepreneur | Crypto...	2013-05-14 20:15:24	4289	1066	

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_fav
74433	People's Daily app	北京, 中华人民共和国	Our mission is to provide news and perspective...	2018-02-04 12:36:42	1413	102	
74434	M0ser	NaN	Reagan conservative and attorney raised in the...	2014-02-18 03:46:28	2554	1733	
74435	Your Friend & Sabre ⚔️	Chicago, IL	My spectral decomposition has a significant da...	2016-12-19 19:55:00	310	1748	

74436 rows × 13 columns

En este segmento se muestra los elementos y variables que contiene el documento a analizar

1. Información general (cantdad de datos, variables, vector de datos y tipos de variables)

In [86]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74436 entries, 0 to 74435
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   user_name         74436 non-null  object
 1   user_location     59218 non-null  object
 2   user_description  70079 non-null  object
 3   user_created      74436 non-null  object
 4   user_followers    74436 non-null  int64 
 5   user_friends      74436 non-null  int64 
 6   user_favourites   74436 non-null  int64 
 7   user_verified     74436 non-null  bool  
 8   date              74436 non-null  object
 9   text              74436 non-null  object
 10  hashtags          53002 non-null  object
 11  source            74424 non-null  object
 12  is_retweet        74436 non-null  bool  
dtypes: bool(2), int64(3), object(8)
memory usage: 6.4+ MB
```

En este segmento se muestran las variables (columnas) que contiene el csv, junto con el numero de elementos que contiene cada variable y su tipo de dato. Se puede observar que se manejan principalmente objetos, seguidos de integers de 64 bits y por último valores booleanos.

In [87]:
```
df.describe(include = object).transpose()
```

Out[87]:

	count	unique	top	freq
user_name	74436	44853	GlobalPandemic.NET	312

	count	unique	top	freq
user_location	59218	14622	India	1496
user_description	70079	42690	Breaking News & Critical Information to SURVIV...	312
user_created	74436	45554	2010-07-13 21:58:05	312
date	74436	56546	2020-07-29 16:30:00	26
text	74436	74312	Greenland has no active cases of the novel cor...	6
hashtags	53002	23445	['COVID19']	16004
source	74424	450	Twitter Web App	22974

En este segmento se puede observar un análisis más a fondo del csv. En esta tabla indica las variables junto con el número de elemento de cada una. Pero además, indica cuanto de esos elementos son únicos, los elementos que más se repiten en cada variable y la frecuencia con la que se repiten.

In [ ]:

En esta tabla se obtiene más información de las variables de tipo integer 64 bits. Muestra el número de valores, la media de los valores de los elementos de cada columna, la desviación estándar, el valor mínimo, máximo y percentiles.
En este caso, se puede observar que el rango de valores de las variables varia una de otra. Debido a esto, al hacer un análisis entre los rangos de esta variables, no obtendría un resultado congruente.

1. 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.

In [88]:
```
df.sort_values(['user_followers'], ascending = False).head(10)
```

Out[88]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favo
6959	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892841	69	
13450	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892839	69	
16194	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892837	69	
235	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892795	69	
2837	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892793	69	

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favo
5344	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892792	69	
20483	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	
20378	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	
24243	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	
23721	CGTN	Beijing, China	#SeeTheDifference with CGTN as we bring you st...	2013-01-24 03:18:59	13892212	69	

In [89]:
```
df.sort_values(['user_friends'], ascending = False).head(10)
```

Out[89]:

	user_name	user_location	user_description	user_created	user_followers
12021	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611718
22924	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611780
20391	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611780
29229	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611801
37508	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611800
59568	Tim Fargo 😁	2x Inc. 500 winner	CEO of Social Jukebox #digitalnomad #socialtoo...	2010-10-09 17:40:24	611965
14520	#ismyhairmessedup	Long Beach CA	Von Wolf Clothing\nhttps://t.co/Fkb3oC5fOU \nh...	2010-12-02 18:16:42	410132
45590	#ismyhairmessedup	Long Beach CA	Von Wolf Clothing\nhttps://t.co/Fkb3oC5fOU \nh...	2010-12-02 18:16:42	409673

	user_name	user_location	user_description	user_created	user_followers
50218	#ismyhairmessedup	Long Beach CA	Von Wolf Clothing\nhttps://t.co/Fkb3oC5fOU \nh...	2010-12-02 18:16:42	409673
62866	Anthony Casas	United States	Work matters!	2008-10-30 18:27:32	490705

In [90]:
```
df.sort_values(['user_favourites'], ascending = False).head(10)
```

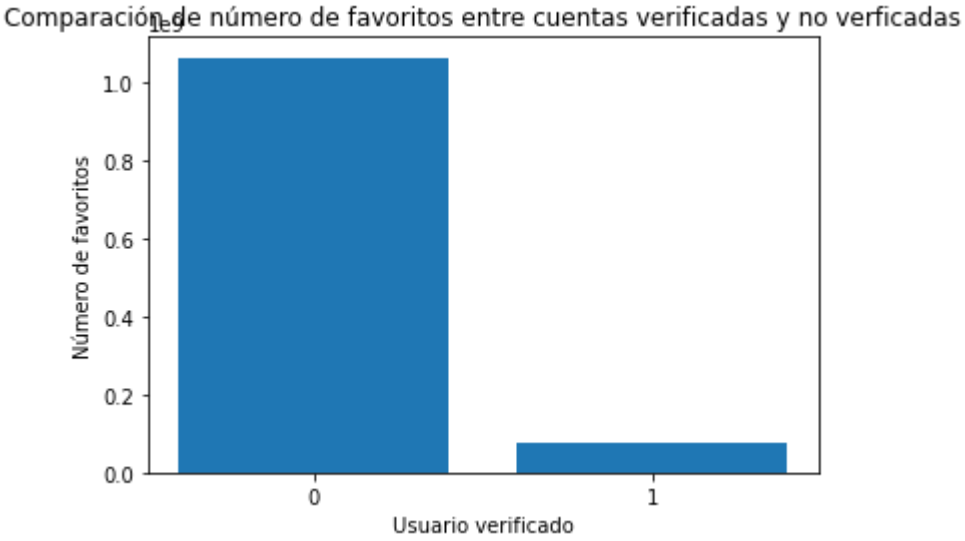
Out[90]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favou
14504	Chelsea Anderson.♥	South Dakota, USA	Activist, photography, reading, drawing, and s...	2013-03-06 04:11:45	22864	22900	204
71764	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87559	93252	115
69943	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87559	93252	115
70610	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87559	93252	115
62200	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87568	93275	115
62446	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87568	93275	115
62126	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87568	93275	115
53316	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87575	93309	115
53438	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87575	93309	115

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favou
44330	paolo ignazio marong	NaN	libero professionista, analista sereno navigat...	2013-04-10 09:29:25	87572	93383	115

En estos tres segmentos se pueden obtener las cuentas con mayor número de followers, friends, así como el tweet con mayor número de favourites.

In [91]:
```
df2 = df.iloc[1:, [7,6]]
verified_favourites = df2.groupby(['user_verified'], as_index = False).agg('sum')
#verified_favourites
plt.bar(verified_favourites.user_verified, verified_favourites.user_favourites)
plt.title('Comparación de número de favoritos entre cuentas verificadas y no verficadas
plt.xlabel('Usuario verificado')
plt.ylabel('Número de favoritos')
plt.xticks(verified_favourites.user_verified)
plt.show()
```



In [92]:
```
df3 = df.iloc[1:, [7, 4]]
verified_followers = df3.groupby(['user_verified'], as_index = False).sum()
#verified_followers
plt.bar(verified_followers.user_verified, verified_followers.user_followers)
plt.title('Comparación de número de followers entre cuentas verificadas y no verficadas
plt.xlabel('Usuario verificado')
plt.ylabel('Número de followers')
plt.xticks(verified_followers.user_verified)
plt.show()
```

En estos dos segmentos se puede ver una comparación entre cuentas verificadas y no verificadas, específicamente en el número de favoritos y número de followers.
En el caso de los favoritos, se puede ver que los tweets de cuentas no verificadas obtuvieron un mayor número de favoritos.
Por otro lado, el número de followers en cuentas verificadas sobrepasa al de las cuentas no verificadas.
Con estos datos se puede concluir que aunque las cuentas verificadas tienen un mayor número de seguidores, no asegura que los tweets que provengan de estas cuentas tendrán un mayor número de favoritos o mayor alcance en la red social.

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

In [93]:
```
df.describe()
```

Out[93]:

	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

Usando la función "describe()" se puede obtener valores como la media, desviación estandar, mínimo, máximo y percentiles, pero únicamente para las variables numéricas del archivo csv, es decir para "user_followers", "user_friends" y "user_favourites".
Con esta función se puede observar que el promedio de númerto de followers y amigos por cuenta

es de 105,951 y 2,154 respectivamente. Por otro lado, el promedio de número de favoritos por tweet fue de 15,297.

Al analizar los otros atributos de la tabla, se puede ver que el valor mínimo fue el mismo para cada una de las variables, es decir 0, mientras que el valor máximo cambio significativamente en cada una de las variables. Al observar el valor máximo se puede ver que la cuenta con mayor número followers (13,892,840) sobrepaso al tweet con mayor número de favoritos (1,047,197). Debido a la diferencias de rangos entres estas 3 variables, se puede decir que si se quisiera realizar un análisis de boxplots entre estas 3, no se obtendría un análisis congruente.