Actividad Reto

Fernando Alfonso Arana Salas A01272933
Paola Fernández Gutiérrez Zamora A01658087
Sofia Donlucas Bañuelos A01655565
Isaac Jacinto Ruiz A01658578
Santiago Gabián A01658280

Objetivo Smart

Analizar la información de los tweets relacionados con covid 19 a través de la implementación de herramientas computacionales para conocer los países con mayor actividad y encontrar la relación entre seguiores, amigos y favoritos para ver si se relacionan con el hecho de que la cuenta sea verificada. Esto para determinar los países que más información presentaron sobre el tema y comprobar la legitimidad de la información de las cuentas verificadas. El análisis se completará en la semana del 9 al 13 de mayo.

In [2]:

```python
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import warnings
import numpy as np
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
from mpl_toolkits.mplot3d import Axes3D

import plotly.express as px

warnings.filterwarnings('ignore')
```

In [3]:

```python
data = pd.read_csv('covid19_tweets.csv')
data.head()
```

Out[3]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	u
0	Ꮙi☻ɭℭ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	
1	Tom Basile 🇺🇸	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	

In [6]:

```
data.columns
```

Out[6]:

```
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 [7]:

```
data.info
```

Out[7]:

```
<bound method DataFrame.info of                      user_name              u
ser_location  \
0                         ᏉᎥ☻ꞁℭ⌶                  astroworld
1                   Tom Basile 🇺🇸                New York, NY
2                Time4fisticuffs            Pewee Valley, KY
3                    ethel mertz        Stuck in the Middle 
4                       DIPR-J&K           Jammu and Kashmir
...                          ...                         ...
179103  AJIMATI AbdulRahman O.               Ilorin, Nigeria
179104                     Jason                     Ontario
179105               BEEHEMOTH ⏳                    🇨🇦 Canada
179106             Gary DelPonte               New York City
179107                   TUKY II  Aliwal North, South Africa

                                         user_description  \
0       wednesday addams as a disney princess keepin i...
1       Husband, Father, Columnist & Commentator. Auth...
2       #Christian #Catholic #Conservative #Reagan #Re...
3       #Browns #Indians #ClevelandProud #[]_[] #Cavs ...
4       🖊Official Twitter handle of Department of Inf...
...                                                   ...
179103    Animal Scientist|| Muslim|| Real Madrid/Chelsea
179104  When your cat has more baking soda than Ninja ...
179105  ⚒ The Architects of Free Trade ⚒ Really Did ...
179106  Global UX UI Visual Designer. StoryTeller, Mus...
179107  TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...

               user_created  user_followers  user_friends  user_favourites  \
0       2017-05-26 05:46:42             624           950            18775
1       2009-04-16 20:06:23            2253          1677               24
2       2009-02-28 18:57:41            9275          9525             7254
3       2019-03-07 01:45:06             197           987             1488
4       2017-02-12 06:45:15          101009           168              101
...                     ...             ...           ...              ...
179103  2013-12-30 18:59:19             412          1609             1062
179104  2011-12-21 04:41:30             150           182             7295
179105  2016-07-13 17:21:59            1623          2160            98000
179106  2009-10-27 17:43:13            1338          1111                0
179107  2018-04-14 17:30:07              97          1697              566

        user_verified                 date  \
0               False  2020-07-25 12:27:21
1                True  2020-07-25 12:27:17
2               False  2020-07-25 12:27:14
3               False  2020-07-25 12:27:10
4               False  2020-07-25 12:27:08
...               ...                  ...
179103          False  2020-08-29 19:44:21
179104          False  2020-08-29 19:44:16
179105          False  2020-08-29 19:44:15
179106          False  2020-08-29 19:44:14
179107          False  2020-08-29 19:44:08

                                                     text  \
0       If I smelled the scent of hand sanitizers toda...
1       Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2       @diane3443 @wdunlap @realDonaldTrump Trump nev...
3       @brookbanktv The one gift #COVID19 has give me...
4       25 July : Media Bulletin on Novel #CoronaVirus...
...                                                   ...
```

```
179103  Thanks @IamOhmai for nominating me for the @WH...
179104  2020! The year of insanity! Lol! #COVID19 http...
179105  @CTVNews A powerful painting by Juan Lucena. I...
179106  More than 1,200 students test positive for #CO...
179107  I stop when I see a Stop\n\n@SABCNews\n@Izinda...

                                 hashtags               source  is_retweet
0                                     NaN   Twitter for iPhone       False
1                                     NaN  Twitter for Android       False
2                             ['COVID19']  Twitter for Android       False
3                             ['COVID19']   Twitter for iPhone       False
4       ['CoronaVirusUpdates', 'COVID19']  Twitter for Android       False
...                                   ...                  ...         ...
179103                      ['WearAMask']  Twitter for Android       False
179104                        ['COVID19']  Twitter for Android       False
179105                                NaN      Twitter Web App       False
179106                        ['COVID19']   Twitter for iPhone       False
179107                                NaN  Twitter for Android       False

[179108 rows x 13 columns]>
```

In [8]:

```
data['source'].unique()
```

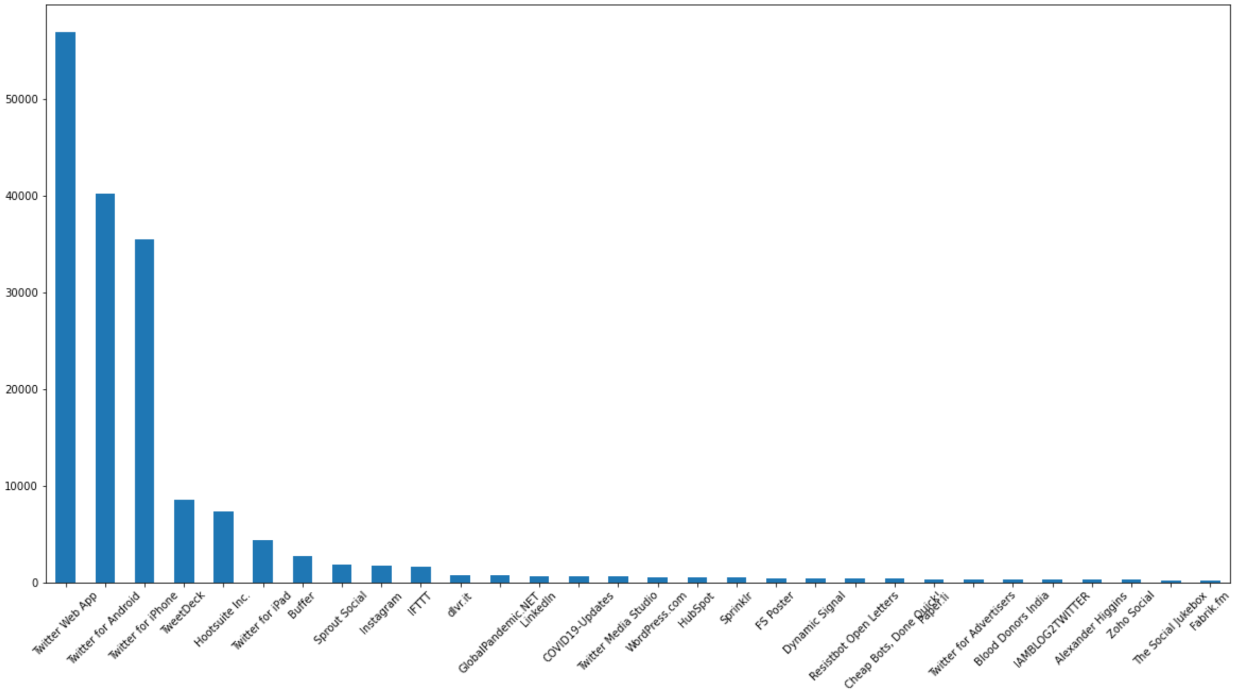
```
Out[8]: array(['Twitter for iPhone', 'Twitter for Android', 'Twitter Web App',
               'Buffer', 'TweetDeck', 'Twitter for iPad', 'Africa Newsroom',
               'Blood Donors India', 'TweetCaster for Android',
               'Alexander Higgins', 'IFTTT', 'Hootsuite Inc.', 'Sprout Social',
               'Sprinklr', 'assarofficial', 'IAMBLOG2TWITTER', 'CrowdControlHQ',
               'COVID19-Updates', 'EveryoneSocial', 'Dynamic Signal', 'Instagram',
               'TweetCaster for iOS', 'GlobalPandemic.NET', 'Venrap Radio',
               'HeyOrca', 'Twitter for Advertisers', 'Paper.li',
               'Twitter Media Studio', 'Twitter for Mac', 'dlvr.it',
               'Cheap Bots, Done Quick!', 'Prof. Shanku', 'LaterMedia',
               'SEMrush Social Media Tool', 'Twitterrific for iOS',
               "Sebastian's Twitter Bot", 'Threader_client', 'COVID19FactoidBot',
               'PwC UK SMART', 'tweet pro stiff', 'UK COVID-19 Alerts',
               'Resistbot Open Letters', 'preprint-alert', 'ContentStudio.io',
               'Peeping Moon', 'TweetAutomaticos', 'Orlo', 'AgoraPulse Manager',
               'Meltwater Social', 'Blog2Social APP',
               'Social Genie by Brighter Vision', 'Social Media Publisher App ',
               'VoiceToData', 'Hearsay Social', 'Metricool', 'SocialPilot.co',
               'Loomly', 'Owly', 'Facelift-Cloud', 'Khoros', 'Oktopost',
               'coronaData_Test', 'SocialOomph', 'SmarterQueue',
               'Salesforce - Social Studio', 'Twittimer', 'Dolar Değişti!',
               'COVID19 Update', 'LinkedIn', 'Socialbakers',
               'Bambu by Sprout Social', 'HubSpot', 'National Herald',
               'Twitter Ads', 'twootlk', 'WordPress.com',
               'Twitter Media Studio - LiveCut', 'Covid-19 Bot',
               'Tweetbot for iOS', 'Zoho Social', 'Mobile Web (M2)',
               'Global Citizen Mobile App', 'whatSaoCarlos', 'Tweetbot for Mac',
               'FS Poster', 'rate_twitte', 'corona-recoveries',
               'The Social Jukebox', 'ContentCal Studio', 'OneUp App',
               'Promo.com', 'TopHashtags', 'autotweet scheduler', 'NepalCorona',
               'Khoros Marketing', 'Hocalwire Social Share', 'DataScienceInfo',
               'crystalwind.ca', 'SNAP-Homeless Times',
               'COVID-19 Information Bot', 'eMartmarket', 'trackingbot2020',
               'MeetEdgar', 'Twidere for Android', 'econ b2b post', 'Aleph א ',
               'SocialBee.io v2', 'Spreaker', 'ExanteData Robots', 'Mention\xa0',
               'Sked Social', 'Periscope', 'CoronaVirus Bot by Sloth',
               'dailyindia', 'Tweet Suite', 'SurveyCircle Team', 'PTI_Tweets',
               'Swat.io', 'Restream.io', 'Powered by Sprinklr', 'GT_Backend',
               'HN_Comments', '@thedextazlab', 'China Xinhua News', 'ArmeniaITN',
               'Crowdfire App', 'Microsoft Power Platform', 'Bitly', 'Echofon',
               'Ripl App', 'WPwamnwebsitescript', 'corona_stats',
               'CoronaTrackerMY', 'The Tweeted Times', 'Qureet Leads',
               'Threat Intel Hub', 'My Khel', 'Falcon Social Media Management ',
               'Zapier.com', 'Echobox', 'Scoop.it', 'SocialRabbit Plugin',
               'covid19_tracker', 'Cubi.so', 'SocialChamp IO ', 'twittbot.net',
               'pixiv: Post to Twitter', 'drumup.io', 'mpsontwitter.co.uk',
               'Covid19daily', 'SocialFlow', 'Vattel Tracker Dev',
               'Radio.co now playing', 'POST.it - Edit,Share,Rediscover',
               'SocialNewsDesk', nan, 'Phone2Action', 'tweeter_biases',
               'Heropost', 'rhega.net', 'Nuzzel', 'DR Data', 'Seket Aanru LVLII',
               'Grabyo', 'sphere_ja_bot', 'Wildmoka', 'TrackerCV',
               "iContact's Social Tools", 'Igorotage', 'HWD', 'Mailchimp',
               'Sprinklr Publishing', 'Integromat', 'Fabrik.fm',
               'Clearview Social, Inc.', 'Friends Me', 'Stuffed Productions',
               'Revive Social App', 'Flamingo for Android',
               'Microsoft Azure Logic Apps', 'MavSocial App', 'Smarp.',
               'UberSocial for Android', 'Missinglettr', 'Canva', 'Talon Android',
               'Konnect Social', 'Siargao Guide', 'PinkVilla', 'Penname',
               'Kashmir Life', "Monty's Twitter Reposter", 'SnapStream TV Search',
               'WatchDog Uganda', 'True Anthem', 'MarketingSuite', 'Corona-Stats',
               'Twitter for  iPhone', 'Social Head', 'BookClubProTweet',
               'covid19_counter', 'PromoRepublic', 'PolitiHUB.us', 'Raven Tools',
               'COVID19-Tweet', 'Cancer Health Auto Tweet', 'remote.io',
               'Arena.im', 'Isrg', 'Constant Contact', 'Google', 'VegaLms',
               'DTR Auto Post', 'CoSchedule', 'COVID19 Daily Stats', 'CovidDay',
               'COVID-19MX', 'SmartNews | スマートニュース', 'Spire FM', 'eClincher',
               'Typepad', 'Publer ', 'Botbird tweets', 'shankar_live_bot',
               'Shauntv', "Bob's Python tweetbot", 'Sprinklr Publisher',
               'Commun.it Intelligence', 'Nintendo Switch Share',
               "Stalin's Twitter Reposter", 'WeVideo', 'recurpost.com',
               'El Cañonazo de las 9', 'Tweetings for Android', 'JoinDiaspora',
               'Post Planner Inc.', 'DopeyUncle2', 'COVIDRecoveredBot',
               'Qnary.io', 'Canada Covid-19 Stats', 'Twitter Test App P',
               'Nelio Content', 'PlayStation®Network', 'Covid19_MV',
               'Postcron App', 'TtwTimes Top News', 'Chicken Nugget',
               'Plume\xa0for\xa0Android', 'Fenix 2', 'ETRetail.com',
               'Campaign Share', 'Postfity.com', 'newsgovhk', 'Sociality.io',
               'Echofon  Android', "Rumpet's Twitter Poster", 'PostBeyond',
               'dnh twitter publisher', 'news_by_gatfil', 'Pardot',
               'App for IconicHipster.com', 'CovidUpdatesBot', 'KhuramKTS',
               'NIAAuto', 'fovle', 'Imminent News', 'PublishBestNews',
               'LFALSNAPAP', 'HW news english', 'StreamElements',
               'Mediaproxy LogJam', 'dancehallaudio',
               'Liberal Forum Political Chat', 'NippyTweet',
               'That Best Home Automation', 'Dear_Assistant', 'Bot Libre!',
               'NCoV-19 Tracker', 'tprzechlewski.app', 'DataBlogger',
               'POZ Auto Tweet', 'SocialDog for Twitter', 'AtlantaTechBlogs',
               'ReadyForSocial App', 'Traject Social', 'abrbuzz',
               'Auto-Post (IMW)', 'ME Construction News', 'NetNaija Twiit',
               'TrafficChiefNG', 'EUobserver', 'Flote.app', 'aa.com.tr',
               'Twitterrific for Mac', 'Bizcommunity.com', 'Covid Robot',
               'VoxPop Sync', 'stocktitan', 'StockTwits Web', 'Dr. Kill Pain',
               'SociabbleApp', 'Post to Social by SHIFT1', 'PRNewswire',
               'Hep Auto Tweet', 'Chorus publishing platform',
               'masterdebater.net', 'A Touch of Snark', 'Login CricketCountry',
               "Let's Talk Singapore", 'RiteKit', 'Etsy', 'Zlappo.com',
               'DirectorsTalk Auto Posting', 'GlobalVillageSpace',
               'TeamSight Publisher', 'naalikeram', 'Amplifr', 'News Medical',
               'Agenparl', 'Hacker Noon', 'Salsa Social Publishing', 'feedspora2',
               'Tech_Tee_Shop', 'CensoredTodayPoster', 'Tumblr', 'Weebly App',
               'science faction twitter bot', 'SocialHub by maloon', 'PNSposter',
               'Social Reputation', 'RSS Masher - Dev', 'Planable', 'Trump OwO',
               "Bongani's Auto Poster", 'TTYtter', 'BAPSCharities',
               'thefivedaily', 'RegulatorWatch.com', 'BrentStafford.com',
               'Phil Ammann', 'flapol', 'Squarespace', 'VoiceFeed',
               'Libsyn On-Publish', 'Sharpspring', 'Twitseq', 'Social Press Kit',
               'Lightful', 'Gain Platform', 'OregonCovidApp',
               'Dynamics 365 for Marketing', 'yorkshire-times', 'Contento App',
               'XHNorthAmerica', 'ContentMX', 'opendorse', 'Greenfly',
               'Freshdesk', 'Triberr', 'EdgeTheory', 'Emma Social Connector',
               'POZ Global Auto Tweets', 'likefollowretweet', 'Healio Twitter',
               'Partnermarketing.com', 'EnrichrBot', 'Statusbrew',
               'Zift Platform', 'Constant Contact - Social Posts', 'Tweepsmap',
               'WayScript', 'SoCast Digital', 'allAfrica.com', 'paulcrypto',
               'nwindianabusiness.com', 'tc_covid_bot', 'Newry.ie',
               'SPTK: DVOrangeCounty', 'cmsoc', 'VoterVoice Tweet', 'Rallio',
               'FalconPro3', 'Cloud Campaign', 'RelojesESP', 'Socialchief',
               'SpaceRefTweet', 'PLAYLOADED.COM', 'TD Wealth Social Centre',
               'covid19oceanupdate', 'Janetter Pro for Android',
               'Fortinet Partner Social', 'Emphatic', 'BLOX CMS',
               'Onollo Software', 'GCCHD', 'floridapolitics.com', 'Web lider',
               'SOCi - Simplifying Social Media', 'Contentpilot', 'ayasetenchi',
               'Chimpify', "K's posts", 'Downtime Monkey', 'COVID19TwitterKenya',
               'Quotes as image', 'Lead Accel', 'Streamlabs Twitter',
               'Dynamic Tweets', 'LatelyAI', 'Foursquare', 'Maat_Interface',
               'coronaBot-Purvesh', 'OneCMS Social Connect', 'BTCNbot',
               'InterfaceCDM', 'QuickTake by Bloomberg', 'Fan Page Robot',
               'Hypefury', 'VotepledgeBot', 'covidestimates',
               'Phone2Action Advocacy', 'Enthusiast97', 'NewsAutocorrect',
               'Ohhtweet', 'TheLancsTimes', 'TW Blue', 'Apphi',
               'Canada COVID-19 Bot', 'standard3d', 'chirr.app',
               'GearGuide Twitter', 'SlackSocial', 'Pressenza IPA eng',
               'SiteAssist Social Posting', 'ZenRud', 'ThreadReaderApp',
               'Cognitive Website', 'UNIAN.info', 'UNIAN Info',
               'Tech Nation News updates', 'Fenix for Android', 'Percolate',
               'GroupTweet', 'Social Aider', 'CyberSecDN v.2',
               'Inclusibot_Revenge', 'Nolwa-Digital', 'OS X', "Janindu's Bot",
               'Influitive', 'Tweecha Lite', 'SocialShare PRD',
               'Woofy Social Media Scheduler', 'ISN Autotweet',
               'App for TonysTShirts', 'Market Beam', 'PLANOLY', 'COVIDTweetBot',
               'eupdater', 'Newswise Expert Pitch', 'Bigram Poetry',
               'Imbibe Website Publishing', 'PixelTweeter', 'Coosto', 'Nonli',
               'Blackbird Video Platform/Amplify', 'ALCOVID19',
               'Hub Central APP Agencies', 'Marketing Agency', '4strat-foresight',
               'audioBoom', 'Grapevine6', 'GaggleAMP', 'StockBot0001',
               'SocialPost App', 'Right Relevance', 'Social5',
               'COVID Articles on SP Journals', 'coronalivebot',
               'Ontario_Covid19', 'Zymplify', 'Whitley2020', 'ChimpReports_',
               'OnlyPultCom', 'Washington Square Parkerz', 'WashSquarePrkrz',
               'VA COVID-19 Updates', 'twmode', 'www.diolch.wales',
               'COVID-19 Alerts', 'TMZ iPhone Social Application',
               'Elgato Stream Deck', 'Bloglovin', 'Coronavirus tracker bot',
               "Joe's Politics", 'TwitPane for Android', 'EastMojo',
               'Co-Kinetic ', 'Moa Bridge', 'UhmazingPressRelease',
               'Ozclubbers Posts', '15 Minute Fun', 'plague-bot', 'Ryzely',
               'LegendLakeComApp', 'STL.News', 'Tweetlogix', 'WoopSocial',
               'Vaccines and Homeopathy News', 'Lately Social', 'SocialGest',
               'BS-SMAP', 'PressPage Manager', 'Snaps1', 'App Political Hispanic',
               '24liveblog', 'Choqok', 'GovernmentCyberWeb', 'PublishToTwitter',
               'xh_scitech', 'WP to Twitter EPR', 'Ground News',
               'Twitter Web Client', 'uwf.london', 'Posting Platform',
               'WP to Twitter IAR', 'Total Travel TAG', '_stranger_twitterbot_',
               'Paiger', 'CGS Tech Tweeter', ' Xinhua Sports', 'TanPaulus27',
               'Mastodon-Twitter Crossposter', 'TS Auto Tweet', 'RH Auto Tweet',
               'Twibble.io', 'Portal de CNMG', 'Revive Old NatCorn Posts',
               'The Recycler', 'Kontentino', 'PostPickr', 'HealthBuzz Tweets',
               'CIO Tech Asia Twitter', 'PolitiTweet Alerts', 'Africa News Hub',
               'Dram Scotland', 'oneindia entertainment', 'gifincric',
               'The Shirt List', 'Hail – Create Curate Communicate',
               "Hamish's Twitter Reposter", 'Sane Auto Tweet', 'forafriendbot',
               'Boma Marketing', 'Fox XRIO 2 News ', 'trumpDigest', 'CivilDialog',
               'Do It Later', 'ActivewearShopping', 'peetm', 'covid19statusbot',
               'Academia Meme Stack', 'The Unshackled',
               'Unshackled Content Poster', 'The Tweeted Times Mobile',
               'BT Production Site', 'Freebie-Depot', 'CareerIndia_Tweet_Counts',
               'LOOKBOOK.nu', 'FPTraffic', 'William Travis Hardman',
               'SPTK: PutnamDV', 'Daily Voice Suffolk County', 'Boost Old Posts',
               'TheWhiteHouse', 'Twitter VIT App for iOS', 'CletusDroo',
               'iHeartMedia Publishing', 'Metigy', 'WIZS 1450 AM',
               'Clarabridge Engage', 'NewsFromDonbass', 'Plurk',
               'feather for iOS  ', 'Partisan Issues?', 'POZ Army Auto Tweets',
               'InfoBlaze Southeast Asia', 'Conservative Daily News',
               'MomoRScrape', 'Sendible', 'News Users', 'twidered',
               ' autopo.st - @Radio_Lichfield', 'apherald', 'InfoBlaze India',
               'JustVent.LIVE', 'CyberSecDN v.7', 'SameraVinson', 'Hiplay',
               'doctroidbot', 'Poshmark', 'rtweet_token_sc', 'Paradym Social',
               'covid-19data', 'Covid19MT', 'Cawbird', 'Auto tweets',
               'Blaq for BlackBerry® 10', 'CoronaWatchUSA', 'oysttyer',
               'Radiology: AI app'], dtype=object)
```

</br> </br> Gráfica: Distribución de Orígenes </br> </br>

In [9]:
```
# Distribution of Sources
data['source'].value_counts()
```

Out[9]:
```
Twitter Web App             56891
Twitter for Android         40179
Twitter for iPhone          35472
TweetDeck                    8543
Hootsuite Inc.               7321
                            ...  
DataBlogger                     1
Dear_Assistant                  1
OnlyPultCom                     1
Washington Square Parkerz       1
Radiology: AI app               1
Name: source, Length: 610, dtype: int64
```

In [10]:
```
plt.figure(figsize=(20,10))
data['source'].value_counts().nlargest(30).plot(kind='bar')
plt.xticks(rotation=45)
plt.show()
```



In [11]:
```
data.drop(['user_description'],inplace = True, axis = 1)
```

```
data
```

Out[11]:

	user_name	user_location	user_created	user_followers	user_friends	user_favourit
0	Ѵ̃i☻ₗℭ†	astroworld	2017-05-26 05:46:42	624	950	187
1	Tom Basile 🇺🇸	New York, NY	2009-04-16 20:06:23	2253	1677	2
2	Time4fisticuffs	Pewee Valley, KY	2009-02-28 18:57:41	9275	9525	725
3	ethel mertz	Stuck in the Middle	2019-03-07 01:45:06	197	987	148
4	DIPR-J&K	Jammu and Kashmir	2017-02-12 06:45:15	101009	168	1
...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	2013-12-30 18:59:19	412	1609	106
179104	Jason	Ontario	2011-12-21 04:41:30	150	182	729
179105	BEEHEMOTH ⌛	🇨🇦 Canada	2016-07-13 17:21:59	1623	2160	9800
179106	Gary DelPonte	New York City	2009-10-27 17:43:13	1338	1111	
179107	TUKY II	Aliwal North, South Africa	2018-04-14 17:30:07	97	1697	56

179108 rows × 12 columns

In [12]:

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
from nltk.corpus import subjectivity
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import re
from textblob import TextBlob

%matplotlib inline

# text cleaner
# turns \n, \t, \r into normal white spaces
# removes retweets, hyperlinks and non-alphanumeric values
# also, removes leading and trailing white spaces
def text_cleaner(text) :
    text = re.sub(r"\n"," ",text)
    text = re.sub(r"\t"," ",text)
    text = re.sub(r"\r"," ",text)
    text = re.sub(r"(@)|(#)|(RT[\s]+)|(https?:\/\/\S+)|([^a-zA-Z0-9 -])", "", t
    text = text.strip(" ")
    return text

# Covid_or_Coronavirus_remover
# removes words containing covid or corona
# also, removes leading and trailing white spaces
def Covid_or_Coronavirus_remover(text) :
    text = re.sub(r"((Covid)|(COVID)|(covid)|(Corona)|(corona)|(CORONA))+", "",
    text = re.sub(r"((Covid)|(COVID)|(covid)|(Corona)|(corona)|(CORONA))[A-Za-z
    text = text.strip(" ")
    return text

# hashtag collector
# finds all hashtags and puts them into a list
# removes # and . symbols to clean up list
# changes all hashtags to uppercase
def hashtag_collector(text) :
    list_of_hashtags = re.findall(r"#[A-Za-z0-9\-\.\_]+",text,re.DOTALL)
    if(list_of_hashtags != None) :
        list_of_hashtags = [word.replace('#', '') for word in list_of_hashtags]
        list_of_hashtags = [word.replace('.', '') for word in list_of_hashtags]
        list_of_hashtags = [word.upper() for word in list_of_hashtags]
    return list_of_hashtags

# at collector
# finds all mentions and puts them into a list
# removes @ symbols to clean up list
def at_collector(text) :
    list_of_ats = re.findall(r"@[A-Za-z0-9\-\.\_]+",text,re.DOTALL)
    if(list_of_ats != None) :
        list_of_ats = [name.replace('@', '') for name in list_of_ats]
    return list_of_ats

# compound sentiment score
# returns Vader sentiment polarity scores
def compound_sentiment_score(tweet):
    VaderSent = SIA()
    Overall_sentiment = VaderSent.polarity_scores(tweet)
    return Overall_sentiment

# getSubjectivity
# returns TextBlob subjectivity score
def getSubjectivity(tweet):
    return TextBlob.subjectivity(tweet)

# getPolarity
# returns TextBlob polarity score
def getPolarity (tweet):
    return TextBlob.polarity(tweet)

# doAnalysis
# returns simple connotation
def doAnalysis(score) :
    return 'Neutral' if (score == 0) else ('Negative' if (score < 0) else 'Posi

# wordClouder
# creates a word cloud
def wordClouder(string_column):
    Words = ''.join([words for words in string_column])
    wordcloud = WordCloud(width = 1000, height = 600, random_state = 10, max_fc
    wordcloud.generate(Words)
    plt.style.use('fivethirtyeight')
    plt.figure(figsize = (20,28))
    plt.imshow(wordcloud, interpolation = "bilinear")
    plt.axis('off')
    plt.show()
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/paofernandez/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

In [13]:
```
data["hashtags_in_tweets"] = data["text"].apply(hashtag_collector)
data
```

Out[13]:

	user_name	user_location	user_created	user_followers	user_friends	user_favourit
0	Ꮙ̃i☻ɭℭ†	astroworld	2017-05-26 05:46:42	624	950	187
1	Tom Basile 🇺🇸	New York, NY	2009-04-16 20:06:23	2253	1677	2
2	Time4fisticuffs	Pewee Valley, KY	2009-02-28 18:57:41	9275	9525	725
3	ethel mertz	Stuck in the Middle	2019-03-07 01:45:06	197	987	148
4	DIPR-J&K	Jammu and Kashmir	2017-02-12 06:45:15	101009	168	1
...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	2013-12-30 18:59:19	412	1609	106
179104	Jason	Ontario	2011-12-21 04:41:30	150	182	729
179105	BEEHEMOTH ⌛	🇨🇦 Canada	2016-07-13 17:21:59	1623	2160	980
179106	Gary DelPonte	New York City	2009-10-27 17:43:13	1338	1111	
179107	TUKY II	Aliwal North, South Africa	2018-04-14 17:30:07	97	1697	56

179108 rows × 13 columns

In [14]:
```
data["ats_in_tweets"] = data["text"].apply(at_collector)
data
```

Out[14]:

	user_name	user_location	user_created	user_followers	user_friends	user_favourit
0	Ꮙ'i☻ɭȻȶ	astroworld	2017-05-26 05:46:42	624	950	187
1	Tom Basile 🇺🇸	New York, NY	2009-04-16 20:06:23	2253	1677	2
2	Time4fisticuffs	Pewee Valley, KY	2009-02-28 18:57:41	9275	9525	725
3	ethel mertz	Stuck in the Middle	2019-03-07 01:45:06	197	987	148
4	DIPR-J&K	Jammu and Kashmir	2017-02-12 06:45:15	101009	168	1
...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	2013-12-30 18:59:19	412	1609	106
179104	Jason	Ontario	2011-12-21 04:41:30	150	182	729
179105	BEEHEMOTH ⌛	🇨🇦 Canada	2016-07-13 17:21:59	1623	2160	9800
179106	Gary DelPonte	New York City	2009-10-27 17:43:13	1338	1111	
179107	TUKY II	Aliwal North, South Africa	2018-04-14 17:30:07	97	1697	56

179108 rows × 14 columns

In [15]:
```
data["text"] = data["text"].apply(text_cleaner)
data
```

Out[15]:

	user_name	user_location	user_created	user_followers	user_friends	user_favourit
0	ᏉᎥ☻ɭℭϮ	astroworld	2017-05-26 05:46:42	624	950	187
1	Tom Basile 🇺🇸	New York, NY	2009-04-16 20:06:23	2253	1677	2
2	Time4fisticuffs	Pewee Valley, KY	2009-02-28 18:57:41	9275	9525	725
3	ethel mertz	Stuck in the Middle	2019-03-07 01:45:06	197	987	148
4	DIPR-J&K	Jammu and Kashmir	2017-02-12 06:45:15	101009	168	1
...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	2013-12-30 18:59:19	412	1609	106
179104	Jason	Ontario	2011-12-21 04:41:30	150	182	729
179105	BEEHEMOTH ⌛	🇨🇦 Canada	2016-07-13 17:21:59	1623	2160	9800
179106	Gary DelPonte	New York City	2009-10-27 17:43:13	1338	1111	
179107	TUKY II	Aliwal North, South Africa	2018-04-14 17:30:07	97	1697	56

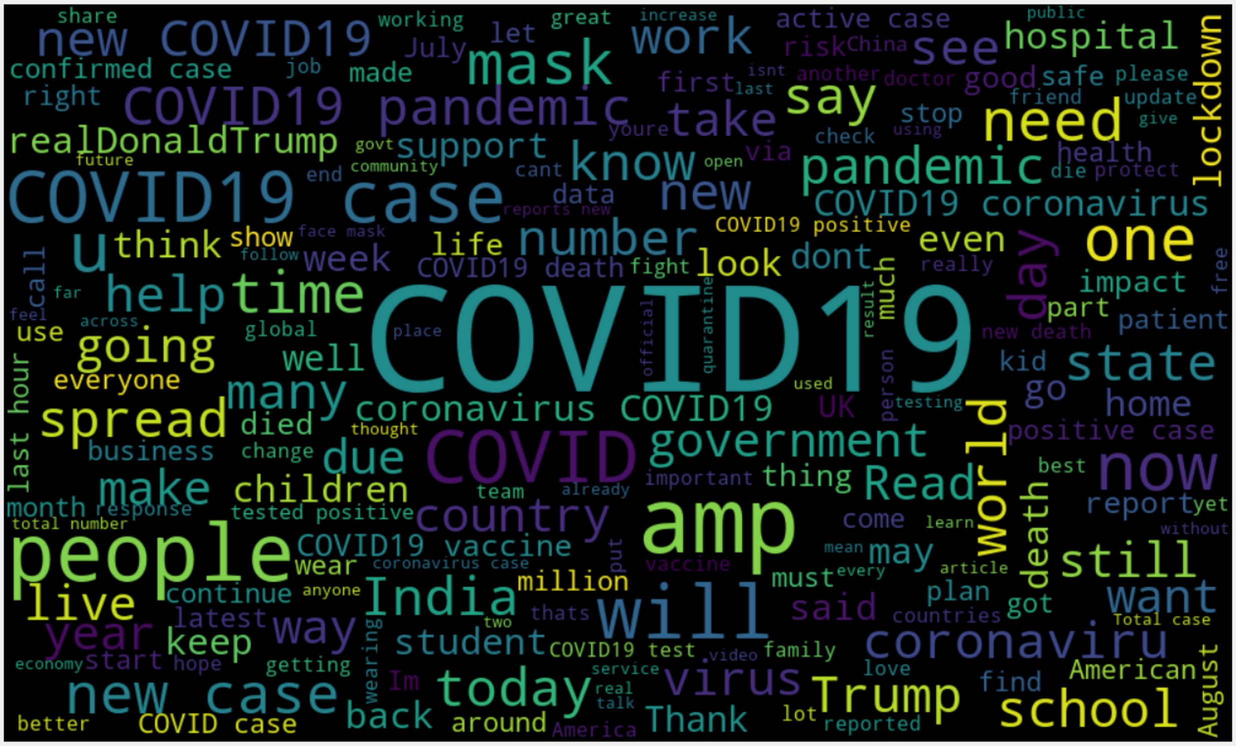
179108 rows × 14 columns

</br> </br> Gráfica: Nube de palabras </br>

Tweets

</br> </br>

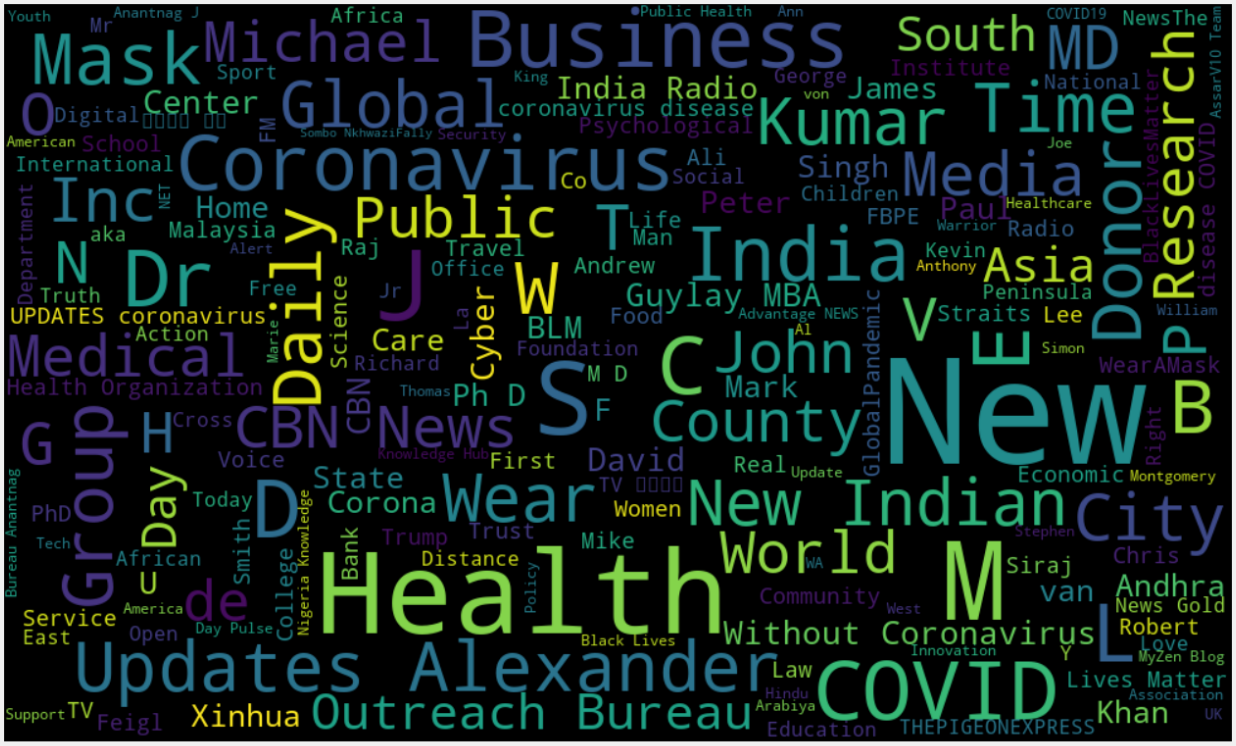
In [16]:
```
# Gráfica para identificar palabras en tweets
wordClouder(data['text'])
```



Hashtags

</br> </br>

In [17]:
```
# Gráfica para identificar palabras en Nombres de Usuarios
wordClouder(data['user_name'])
```



In [18]:
```
data_new = pd.read_csv('covid19_tweets.csv')
data_new.drop(data_new.columns[[0, 1, 2, 3, 6, 7, 8, 9, 10, 11]], axis='columns
data_new
```

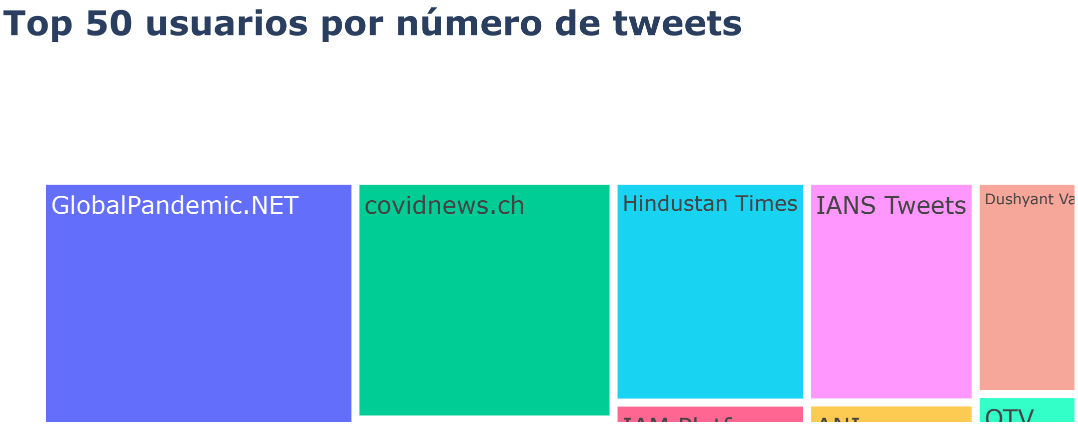
Out[18]:

	user_name	user_location	user_description	user_created	user_followers	user_frier
0	Ѵ̃i☻ɭℭ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	9
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	16
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	95
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	9
4	DIPR-J&K	Jammu and Kashmir	🖊️Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	1
...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	Animal Scientist|| Muslim|| Real Madrid/Chelsea	2013-12-30 18:59:19	412	16
179104	Jason	Ontario	When your cat has more baking soda than Ninja ...	2011-12-21 04:41:30	150	1
179105	BEEHEMOTH ⏳	🇨🇦 Canada	⚒️ The Architects of Free Trade ⚒️ Really Did ...	2016-07-13 17:21:59	1623	21
179106	Gary DelPonte	New York City	Global UX UI Visual Designer. StoryTeller, Mus...	2009-10-27 17:43:13	1338	1
179107	TUKY II	Aliwal North, South Africa	TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...	2018-04-14 17:30:07	97	16

179108 rows × 13 columns

</br> </br> Gráfica: Mapa de Arbol </br> </br>

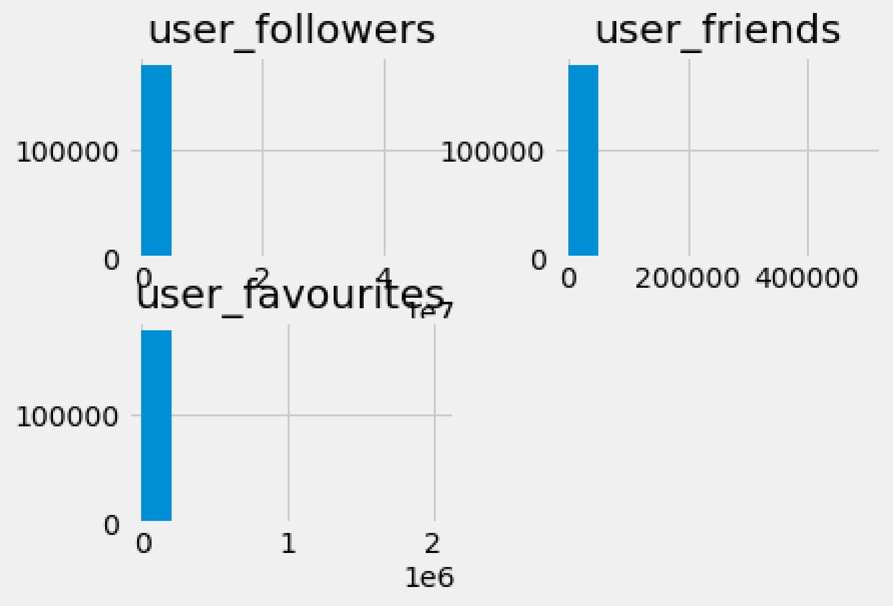
```
In [4]: usuarios_conteo = data['user_name'].value_counts().reset_index().rename(columns
             'user_name':'Numero_de_Tweets','index':'Usuario'})

        fig = px.treemap(usuarios_conteo.head(50), path=['Usuario'], values='Numero_de_
                         title="<b>Top 50 usuarios por número de tweets</b>")

        fig.show()
```



</br> </br> Análisis con K-means </br> </br>

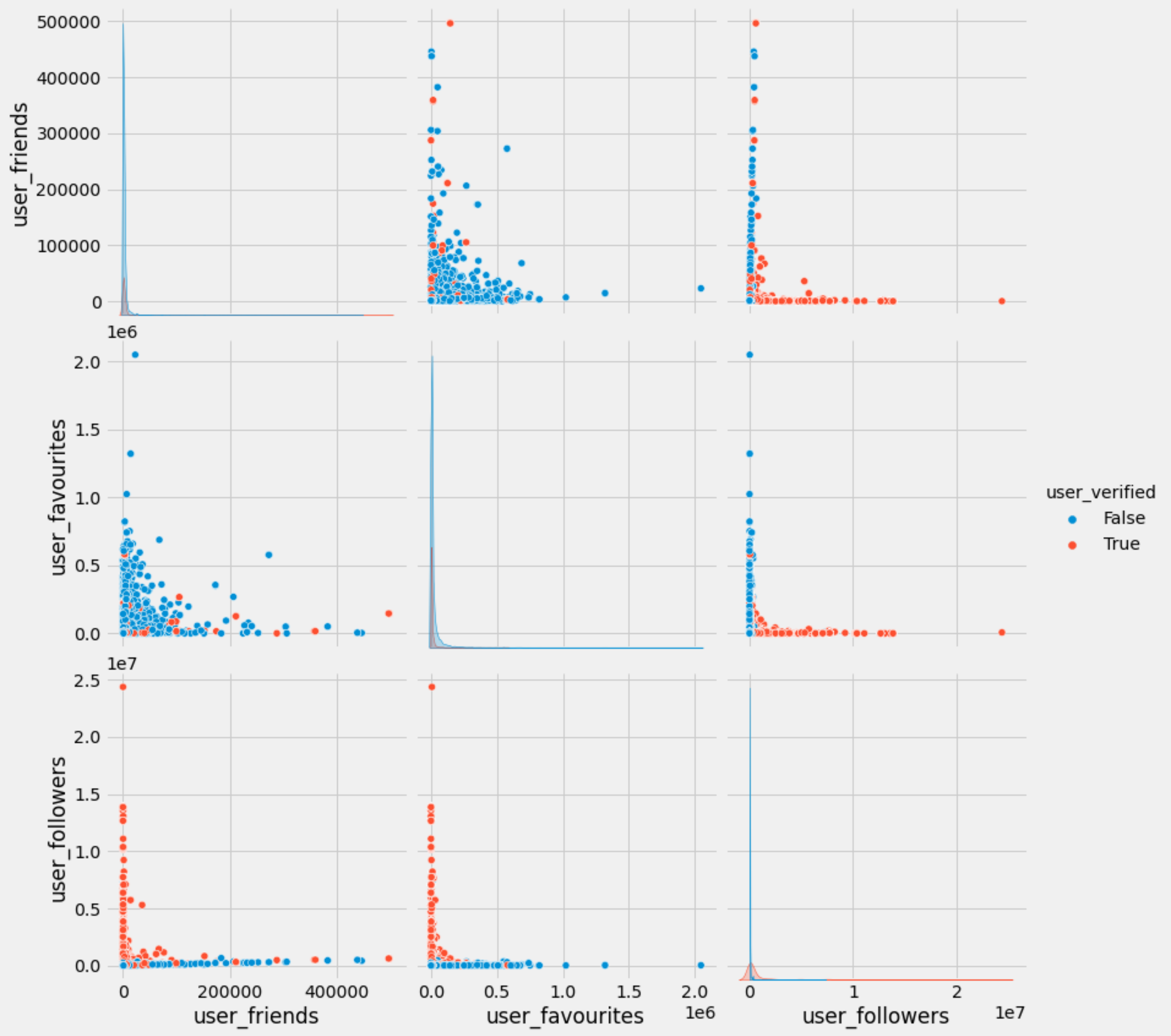
¿Es posible incluir el uso del algoritmo Kmeans para apoyar con la descripción de los datos? Basado en los datos que se tienen, consideramos que si es posible utilizar un diagrama K-means para apoyar con la descripción de datos. Observamos que es posible comparar los seguidores que tiene un usuario, sus amigos y sus favoritos y encontrar si hay alguna relación con que su cuenta esté verificada. Esto puede servir para determinar si la verificación de una cuenta se ve afectada por estos datos y así determinar si esa verificación en realidad significa que sus tweets son información verídica y significativa o no si no lo asegura. Aunque en este caso se hizo este ejemplo se podrían analizar otros valores para encontrar diferentes grupos significativos por la información obtenida ya que el método k-means ayuda a estudiar comportamientos en páginas web entonces podría ser útil en este caso.

In [20]:
```
# Se realizará analizando los usuarios verificados
data.drop(['user_verified'],1).hist()
plt.show()
```



In [21]:
```
sns.pairplot(data.dropna(), hue='user_verified',size=4,vars=["user_friends","us
```

Out[21]:
```
<seaborn.axisgrid.PairGrid at 0x7f8b9331e610>
```



In [22]:

```
# Se utilizarán los valores numéricos y se hará de 3 dimensiones
X = np.array(data[["user_friends","user_favourites","user_followers"]])
y = np.array(data['user_verified'])
X.shape
```

Out[22]:

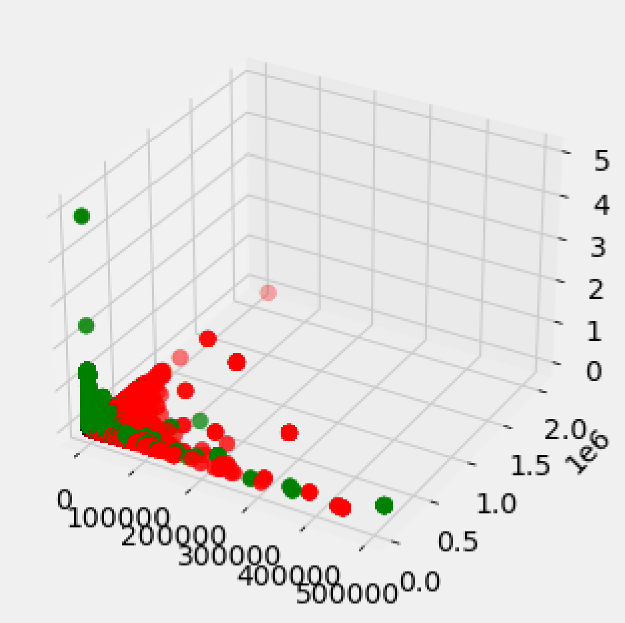
```
(179108, 3)
```

In [23]:

```
fig = plt.figure()
ax = Axes3D(fig)
colores=['red', 'green']
asignar=[]
for row in y:
    asignar.append(colores[row])
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar,s=60)
```

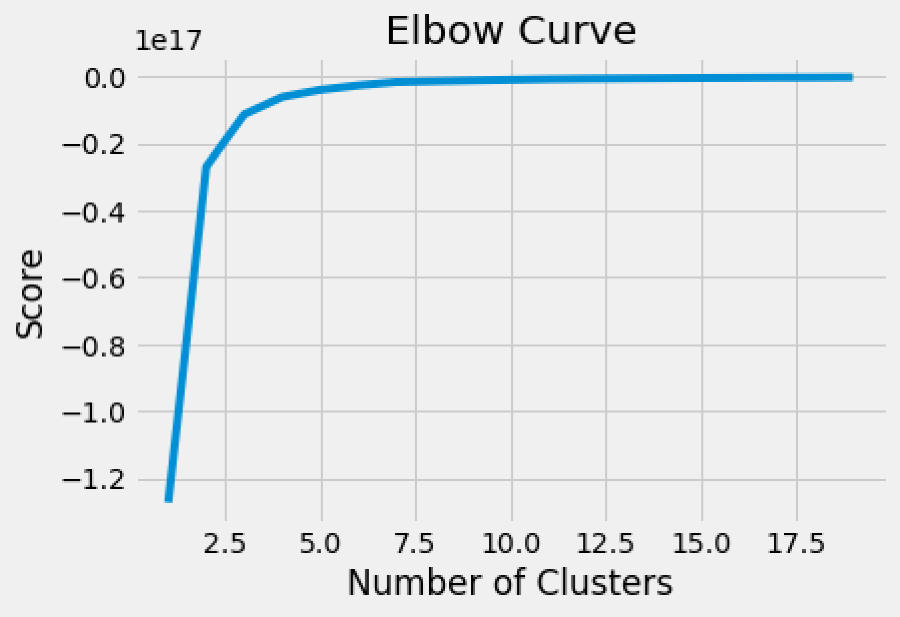
Out[23]:

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7f8b313e3910>
```



In [24]:

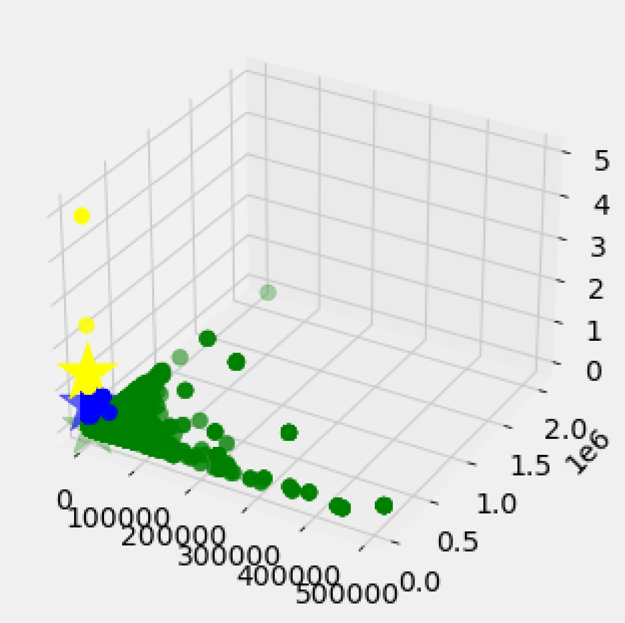
```
# Obteniendo el valor K
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
score
plt.plot(Nc,score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()

# Ejecutando K-Means para 3 clusters
kmeans = KMeans(n_clusters=3).fit(X)

# Obteniendo etiquetas y centroids
centroids = kmeans.cluster_centers_
print(centroids)

# Grafica 3D - estrellas marcan el centro
# Prediccion de clusters
labels = kmeans.predict(X)
# Obteniendo los centros de los clusters
C = kmeans.cluster_centers_
colores=['green','blue','yellow']
asignar=[]
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar,s=60)
ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c=colores, s=1000)
```



```
[[2.13836740e+03 1.45774127e+04 3.20424443e+04]
 [6.64066007e+02 2.59558086e+03 5.71070570e+06]
 [2.40845960e+02 1.23171717e+02 1.31393268e+07]]
```

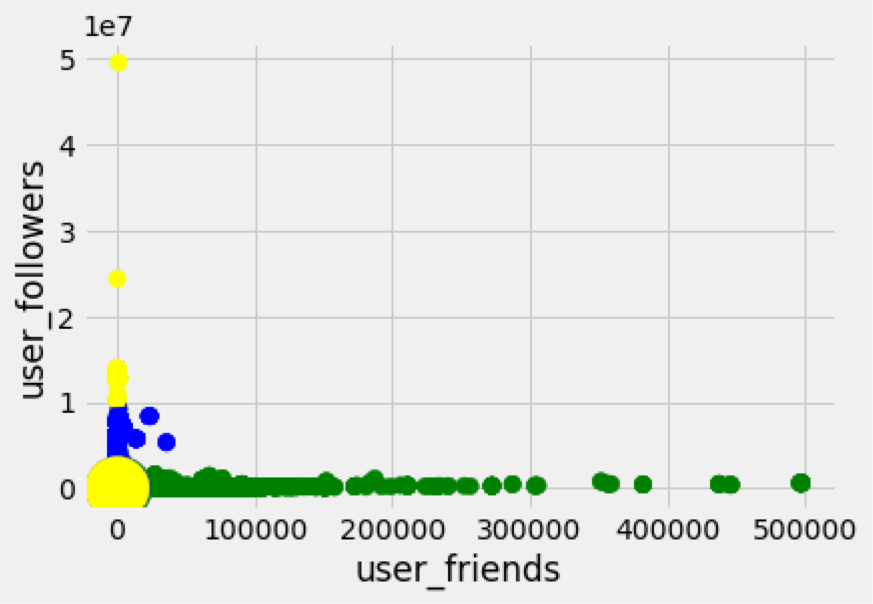
Out[24]:

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7f8b964780d0>
```



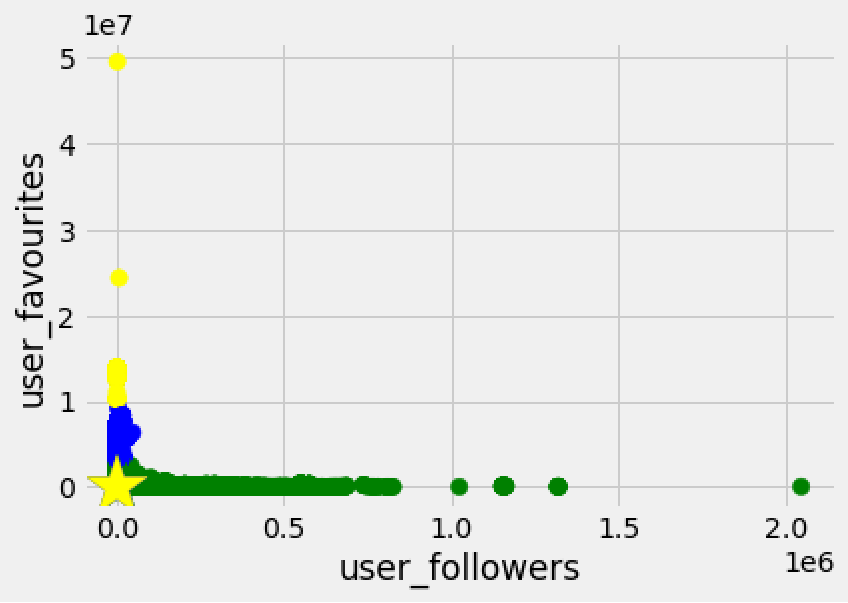
In [25]:

```
friends = data['user_friends'].values
followers = data['user_followers'].values

plt.figure()
plt.scatter(friends, followers, c=asignar, s=70)
plt.scatter(C[:,0], C[:,1], c=colores, s=1000)
plt.xlabel('user_friends')
plt.ylabel('user_followers')
plt.show()
print(C)
```

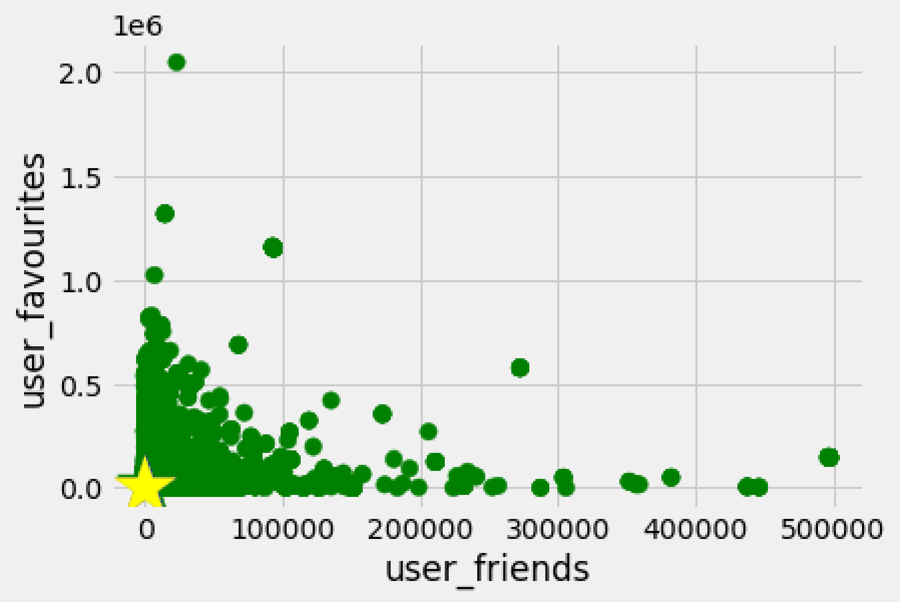


```
[[2.13836740e+03 1.45774127e+04 3.20424443e+04]
 [6.64066007e+02 2.59558086e+03 5.71070570e+06]
 [2.40845960e+02 1.23171717e+02 1.31393268e+07]]
```

In [26]:
```
import random
favourites = data['user_favourites'].values
followers = data['user_followers'].values

plt.xlabel('user_followers')
plt.ylabel('user_favourites')
plt.scatter(favourites, followers, c=asignar, s=70)
plt.scatter(C[:,0], C[:,1], marker="*", c=colores, s=1000)
plt.show()
```



In [27]:
```
friends = data['user_friends'].values
favourites = data['user_favourites'].values

plt.scatter(friends, favourites, c=asignar, s=70)
plt.scatter(C[:,0], C[:,1], marker="*", c=colores, s=1000)
plt.xlabel('user_friends')
plt.ylabel('user_favourites')
plt.show()
```



In [28]:

```
copy =  pd.DataFrame()

copy['user_verified']=data['user_verified'].values
copy['label'] = labels
cantGrp =  pd.DataFrame()
cantGrp['color']=colores
cantGrp['cantidad']=copy.groupby('label').size()
cantGrp
```

Out[28]:

	color	cantidad
0	green	177197
1	blue	1515
2	yellow	396

In [ ]: