INTRODUCTION TO DATA SCIENCE - COURSE WORK

Setting up Data

I've used MySQL database to store the twitter data. Using python code, I've parsed the json files inside the given zip files and extracted only those keys that I figured are useful for the analysis (by reading all the coursework questions before extracting data). Using csv writer, I've created 30 CSV files each containing 24 hours of tweet data. Using LOAD DATA LOCAL INFILE query(specified below), I've loaded those CSV files into a table in the MySQL Database.

```
mysql> LOAD DATA LOCAL  INFILE 'E:/Introduction to Data Science/new_twitter_data_01.csv' INTO TABLE
twitter_final_twitter_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n';
```

Some of the columns in my database are - id, user_id, user_name, user_created_at, place_country, place_bounding_box, place_name, created_at, text, coordinates, geo amongst a total of 39 fields.

Part 1. Basic Stats

1.1 Count the total number of tweets, describing how you deal with duplicates or other anomalies in the data set.

We've been provided with a 1% sample of tweets from 1st June 2021 to 30th June 2021. Some of the anomalies we've in the dataset are:

- Some tweets are just invalid/empty json objects.
- We have the data of ~7700 tweets from 31st May 2021. - We also have duplicate tweets in our dataset.

Each json has Tweet, User, Geo, Entity and Extended Entity objects.

There are some texts containing Unicode characters in the dataset which caused trouble(Unicode Decode Error) while loading the data into mysql db.

Unicode text contains a large number of characters some of which might prove to be troublesome while storing the data.

Such texts are handled separately where the word(not the sentence but just the word) causing error would be replaced with empty string.

To count unique tweets, I've considered the number of tweets with distinct tweet id, user id, tweet creation time, coordinates and the first 20 characters of every tweet using the below MySQL query.

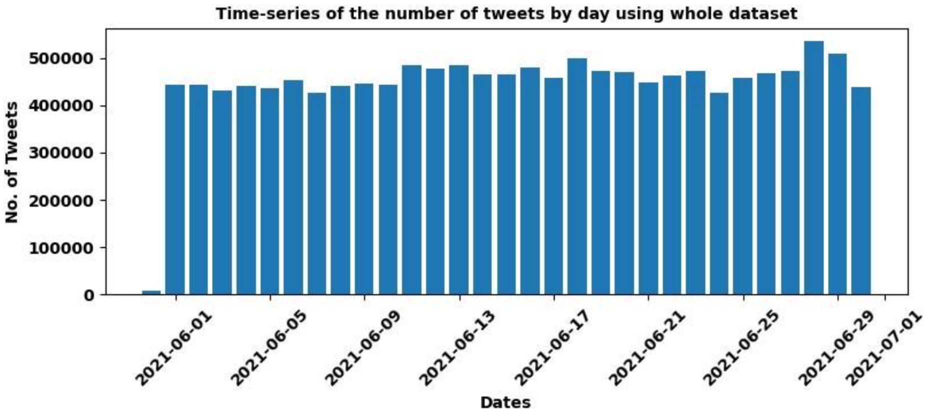
```
mysql> select count(*) from (select distinct id, tweet_created_at_parsed as date, left(text,20) from
twitter_db.final_twitter_2)T;
```

Total no. of Unique tweets is **138,57,161.**

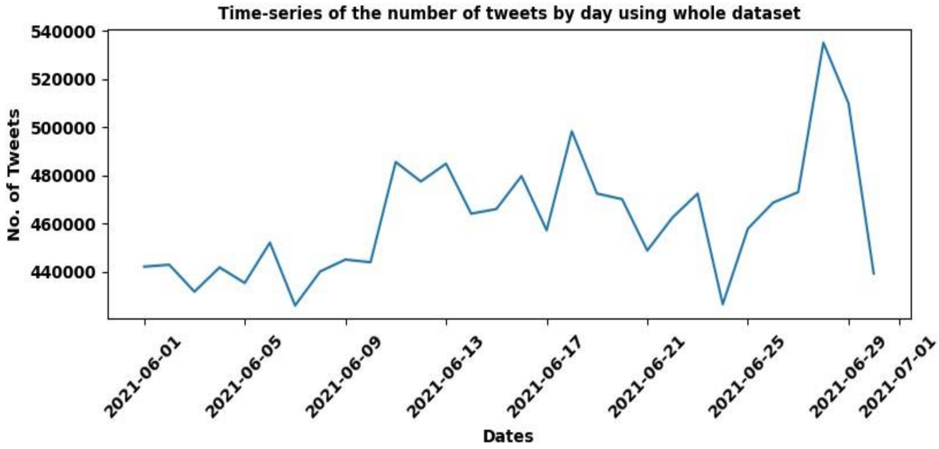
1.2 **Plot a time-series of the number of tweets by day using the whole dataset and comment on what you see.**

After establishing a connection between my python script and the MySQL db using MYSQLDB, I've used the following query in Python

```
mysql> select count(distinct id) as id, left(tweet_created_at_parsed,10) as date
from twitter_db.final_twitter_2;
```



Above is a bar chart depicting the time series of number of tweets by day using the whole dataset. We can see that there is a small bar showing data from *31st May 2021.*



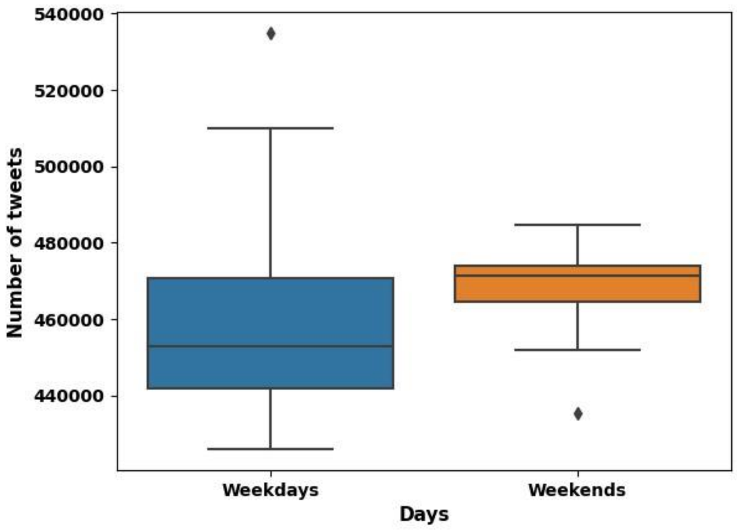
After removing the tweets dated May 31 from the database, I've plotted the above line chart , to see if there're any patterns. We can see that during 28th May, there is the highest spike in the number of tweets while we can see smaller spikes on 10th and 18th.

1.3 Using a box and whisker diagram compare the average number of tweets on the weekdays in the dataset to the numbers for weekend days. Are there statistically significant differences between the number of tweets on weekdays and weekends?

```
mysql> select distinct id, tweet_created_at_parsed as date, weekday from
twitter_db.final_twitter_2;
```

Using the above query, I’ve fetched the data and then grouped them based on weekday (which is the day of the week) and fetched the number of tweets on each day after which they’re categorised into weekdays and weekends.

- Below is a box plot showing the average number of tweets on weekdays and weekends.
- From the graph, we can see that the median no. of tweets posted on weekdays is around 450,000 while the same for weekends is around 470,000.
- To find out if there’re any significant statistical difference between those numbers, I’ve performed oneway ANOVA tests(and also independent t-test, separately).



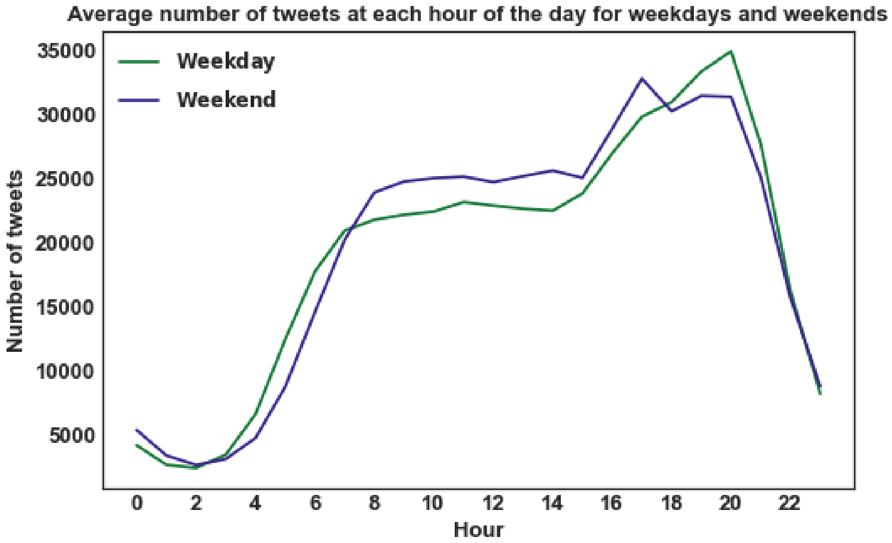
```
from scipy.stats import f_oneway
  dataframe = weekdays_df.groupby(['just_date','day','weekday'],\
                                  as_index=False)['id'].count()
                                  weekdays_list =
                                  dataframe[dataframe['day'] ==
                                  'Weekdays']['id'].tolist()
                                  weekends_list =
                                  dataframe[dataframe['day'] ==
                                  'Weekends']['id'].tolist()

t, p =  f_oneway(weekdays_list, weekends_list)
print(f' p-value: {p}')
```

In both the cases the p-value is around 0.515 which is not less than the predefined alpha – 0.05. Hence, we fail to reject the null hypothesis, implying that there is no significant difference between the number of tweets on weekdays vs weekends.

1.4 Plot the average number of tweets at each hour of the day for weekdays and weekends and comment. You should have two plots where the x-axis is time of day (from midnight to midnight) and the y-axis shows the number of tweets.

Below is the plot depicting the average no. of tweets at each hour of the day for weekdays vs weekends. We have grouped the data by the type of the day and hour of the day, where type of the day is either a weekday or a weekend.



- From the above plot, we can see that the average number of tweets start slowly growing up from 2:00 AM but reduce the pace from 4:00 AM where we can see early birds start being active on the platform.
- The highest number of tweets is recorded during 8:00 PM and 10:00 PM – most people engage themselves in social media during this time.
- We can also see that the average number of tweets between weekdays and weekends is almost the same and there could be no statistical significance, which has been confirmed in the previous analysis(1.3).

Part 2:

2.1 Draw a map of Europe showing the location of the GPS-tagged tweets - these are tweets which have a “coordinates” field in the metadata. The exact form of the map is up to you: marks will be given for accuracy, clarity, and presentation.

In all Tweet objects from the dataset, we’ve two fields that give information about the location –

1. Coordinates – Latitude, Longitude

2. Geo – Longitude, Latitude

Until recently, 'geo' is used to store the coordinates, which is currently deprecated, as per the twitter docs. But, in some of our tweets where Coordinates information is not provided, Geo information is available. Hence, I combined both of those fields using geo information only when coordinates field is null.

Using the data, I’ve used Folium to plot the following graphs:

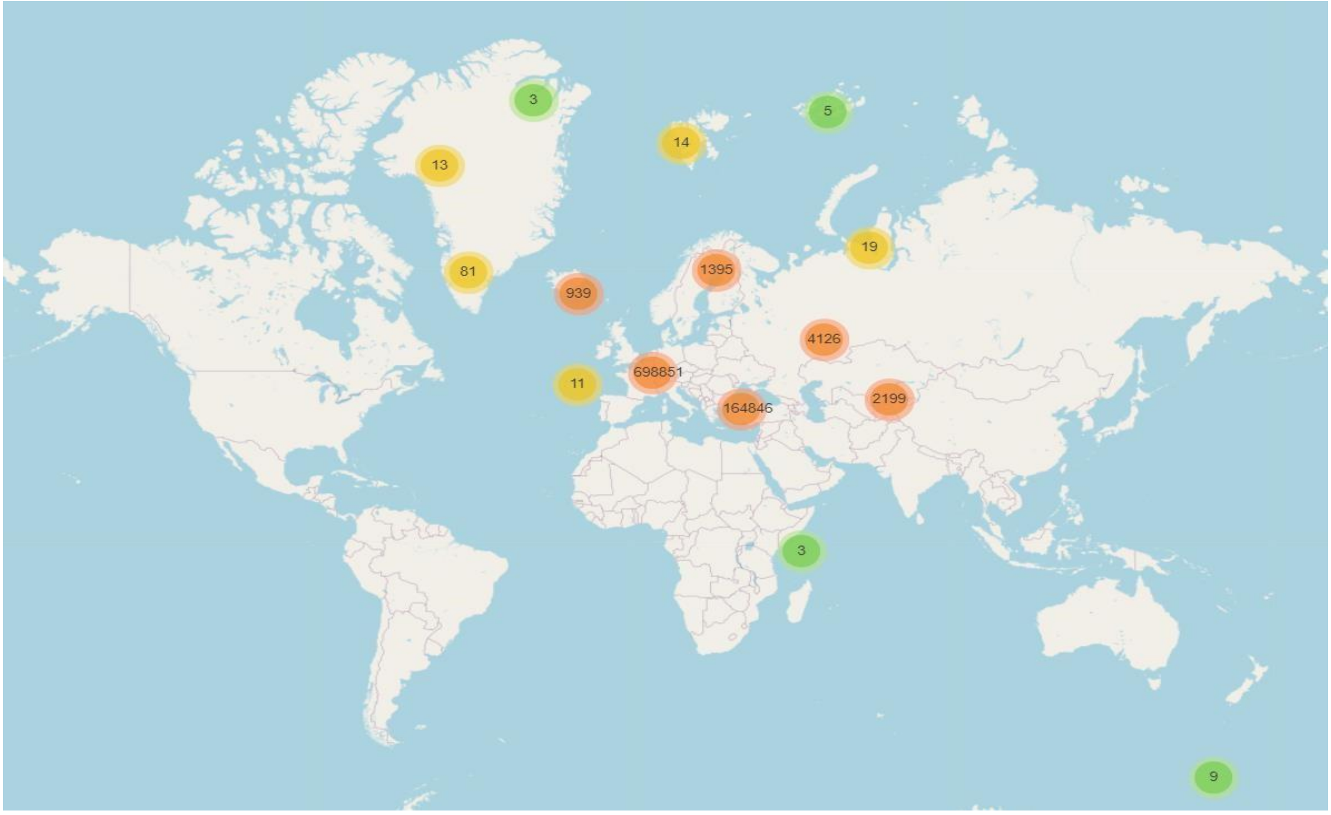


Figure 1

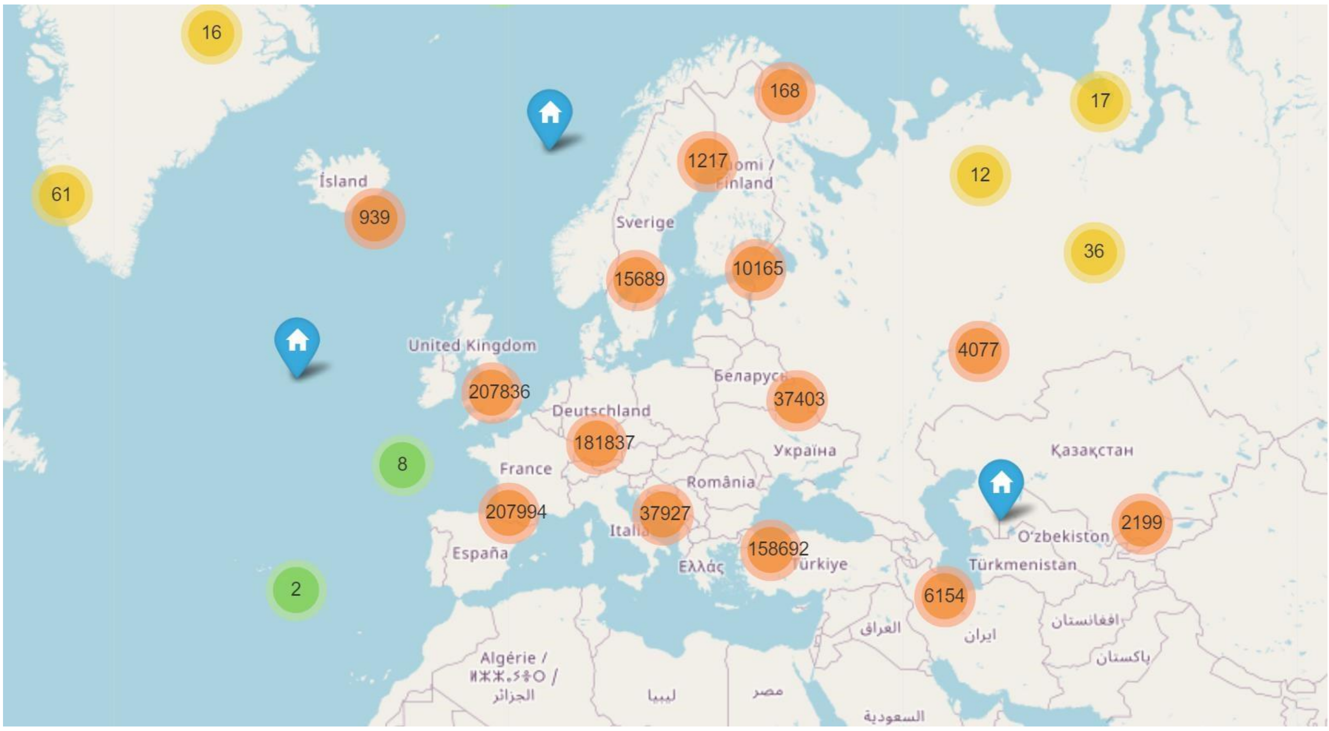


Figure 2

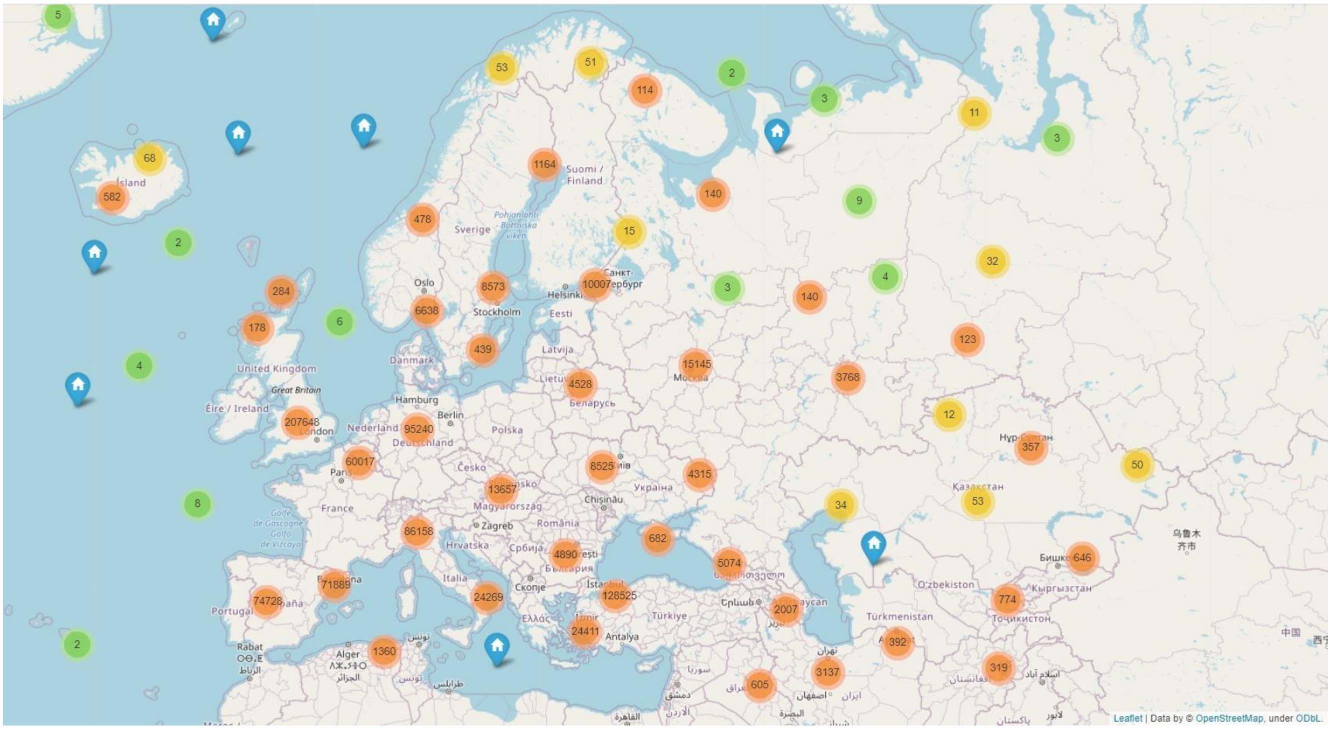


Figure 3

2.2 Explain any patterns you observe.

- We can see that most of the tweets are from the United Kingdom, and Turkey
- Some of the points are outside the bounding box mentioned in the coursework. This indicates that some users had given Twitter the access to their GPS. Hence, the locations of the users when the tweet was posted is depicted by the plot.

Part 3:

3.1 Make a histogram of tweets per user with number of users on the y-axis and number of tweets they make on the x-axis. Discuss the distribution that you see. All the users in the data set should be included! All the users in the data set should be included!

- In my MySQL table, each row contains the data of a distinct tweet. So, the frequency count of a user should be the number of tweets that he posted.

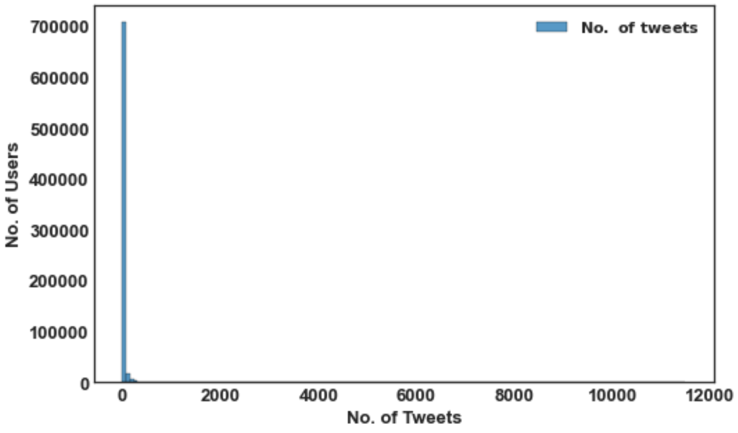
- Hence, I've selected all the users and executed value_counts() on the dataframe to get the number of tweets each user posted. (*I've used this method instead of grouping by user directly in MySQL because that was taking longer time than expected and hence, I used this logic to save time*.)

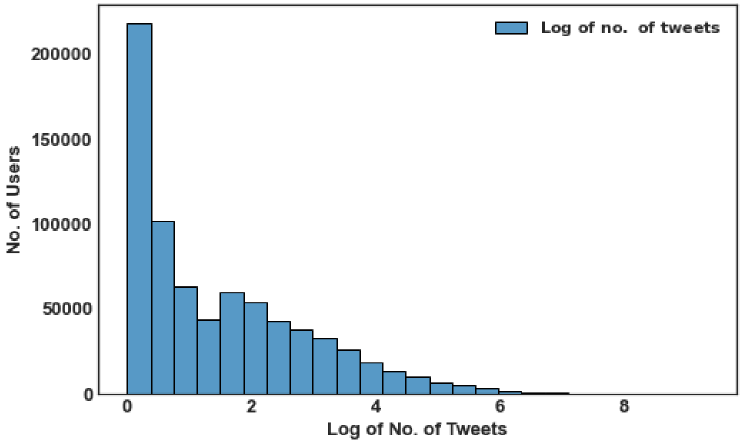
```
query = "select user_id from twitter_db.final_twitter_2;"
```

```
hist_df = pd.read_sql(query, conn)
tweet_count_per_user = hist_df.value_counts()
sns.histplot(tweet_count_per_user, bins = 150)
```

Using the dataframe, the below histogram was plotted.

On clear observation, we can see that most of the users recorded less than 10 tweets in the span of 30 days. Although we are unable to notice, there must be users who plotted more than 10,000 tweets in the span of 30 days*(which is almost 14 tweets an hour)* and hence the X ticks are spanning till 12000.





To get more details, I've plotted a log of the no. of tweets vs the no of users and this is what the plot looks like. We can see that approximately more than 80% of the people tweeted at most once every 2 days, on average (15 tweets or lesser in 30 days).

3.2 Find the top-5 users by total number of tweets. Do you think any are automated accounts (aka bots)? Justify your answer.

```
mysql> select distinct user_screen_name, count(distinct id)as no_of_tweets from
twitter_db.final_twitter_2 where id!= '###' group by 1 order by 2 desc limit 5;
```

user_screen_name	**no_of_tweets**	***Is Bot***
HoraCatalana	11483	*Yes*
Maria70221974	11176	*No*
AyferGl02976871	9415	*Yes*
casimiroperezc1	9059	*No*
RadioTeddyMusic	8733	*Yes*

Going by the sheer number of tweets, one could be tempted to declare all of them as bots –they should have tweeted 12-16 tweets an hour, on an average to score those numbers. But not all of them are bots – here's my justification: There can be more than one managing user managing an account as per the official website. So, it is possible that some accounts are being aggressively maintained and the reasons for such aggressive maintenance could for various purposes like advertisings, commercials, etc.

- **HoraCatalana**: This has only *348 distinct tweets*, teaching how to say time in Catalan language. This is maintained by a bot which tweets at certain intervals of time.
- **AyferGl02976871**: This account is a bot because most of the tweets are the same - retweets of some set of images.
- **Radio TEDDY Playlist**: This is the official twitter page of radioteddy.de. Tweets inform about all the songs played in their website
- **Maria70221974, casimiroperezc1**: They look like they're maintained by human(s). The tweets are not repetitive, and contain text, images, and videos. These users also retweeted and responded to some tweets.

Part 4

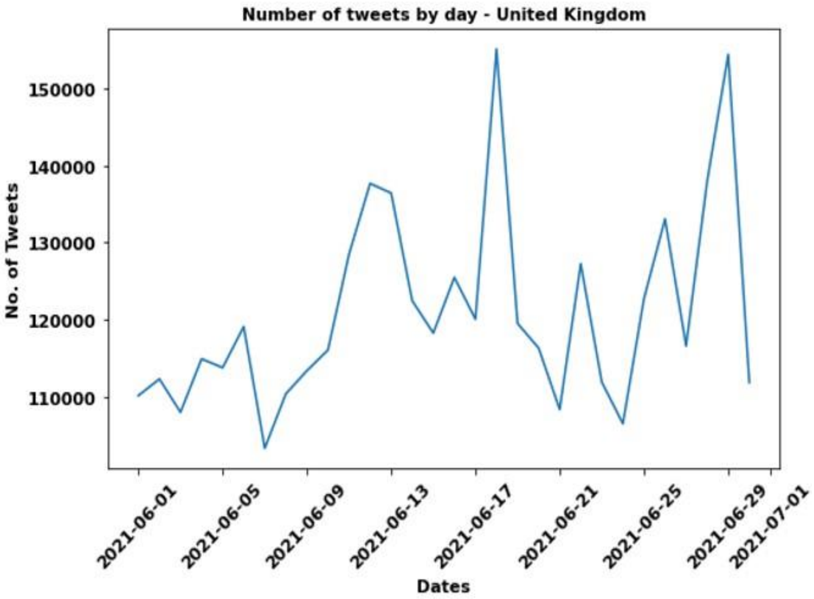
4.1 Identify 3 days with unusually high activity in 3 different countries of your choosing. Describe and justify how you identify 'unusual' days.

- To identify unusually high active days in various countries, I've calculated the monthly average tweet count per country based on the *country code* and calculated the average tweets for each of them.
- From those countries, I've selected the top 15 countries expecting to see high activity in at least some of those countries.
- For these 15 countries, I plotted a line graph with number of tweets per day on X-axis and no. of tweets on Y-axis and looked for those unusually high/sudden spikes in the graphs.

Country Code	Avg no. of Tweets
GB	117201.8
TR	72723.1
ES	62520.8
FR	33826.5
IT	22470.9
DE	22302.9
NL	15960.7
RU	11334.4
IE	9363.6
PL	7293.4
PT	7258.1
IR	6464.7
SE	5523.0
BE	4967.8
RS	3952.8

The above table shows the top 15 countries which have higher number of average tweets from the list of all countries.

From the graph, we can see that 18th June and 29th July have the highest number of tweets which indicates that they're the recorded unusually high activity in the UK. After plotting similar graphs for other countries as well, I've observed unusual spikes in Britain, Turkey and France on 18th June, 20th June, and 28th June respectively.



4.2 Characterise each of these three days. Exactly how you do this is up to you, but for example you could:

Display some indicative Tweets.
Make a word cloud from the tweet text. Plot tweets locations on a map.

Validate your conclusions with some other source of data e.g., government or news reports

For all the 3 days in the 3 countries, I've plotted word clouds to see what's causing the sudden spike in the number of tweets.

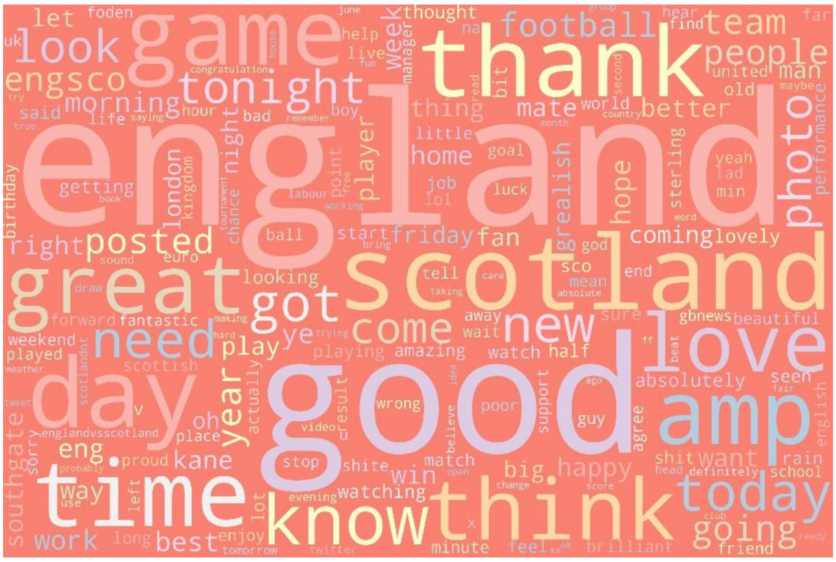
Text Preprocessing:

- Before plotting a word cloud, the following preprocessing steps have been done:
  - Using modules like NLTK and Spacy, I've updated the list of stop words that are to be removed from the text.
  - After removing stop words, all the text has been converted to lower case. Otherwise, we might end up getting the same word more than once in the cloud with it's original weight divided between those words.
  - After this, I've used the below code to plot a word cloud.

```
from wordcloud import WordCloud, STOPWORDS
text = tokenizeandstopwords(text)
wordcloud = WordCloud(width= 3000, height =
2000, random_state=1,
background_color='salmon',
colormap='Pastel1', collocations=False,
stopwords = STOPWORDS).generate(text)
plt.figure(figsize=(40, 30))
plt.imshow(wordcloud)
plt.axis("off")
wordcloud.to_file("wordcloud_UK.png")
```

United Kingdom (18th June 2021):

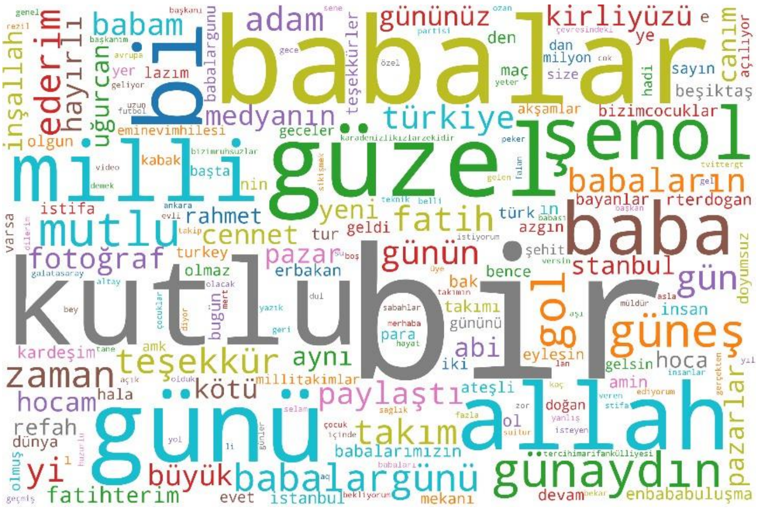
- From the word cloud, we can see that England, Scotland, game and football are some of the heavily weighted nouns in the cloud.



- We can also see words like good, great, thank and love expressing a positive emotion.
- From the word cloud, my guess would be that England had won a football match against Scotland.
- Upon doing some background research, I found that it was actually a goalless draw match at Wembley.
    - Source: **https://www.bbc.co.uk/sport/football/51197603**

Turkey (20th June 2021):

- All the words generated are in Turkish and so, there're no first impressions as to what might have caused the spike in number of tweets.
- We can see that the words babalar, gunu, babalar gunu have appeared multiple times in the word cloud.

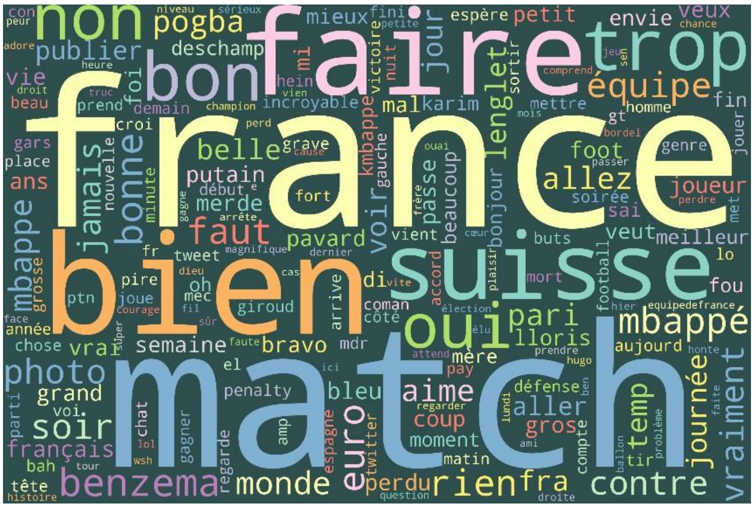


- A simple google search revealed that "Babalar Gunu" translates to Father's Day.
- While words like Kutlu and Guzel, translate to welfare and happiness, Tesekkur, Tesekkurler mean thank you.
- Based on these words, one can guess that people in Turkey were celebrating Father's Day on Twitter, exchanging greetings and thanks.

- As per Wikipedia, Google and other websites, Father's Day is celebrated on the third Sunday of June which happened to be 20th June in 2021.

France (28th June 2021):

- Although all the words are generated in French, there're a few words in English – France, Match which could mean that there was a match(sports event) of sort on this day.
- After using Google translate to find out more, some of the interesting words were equipe – team, bien – well, aujourd – today, bravo – well done, aime – love, Mbappé – French football player



- One can guess that there was a football event scheduled that day. After doing a bit of research, we come to know that our guess – supported by data - was true.
    - **https://www.eurosport.com/football/euro-2020/2021/euro-2020-france-v-switzerland-followlive_sto8392153/story.shtml**

When there's a Football Event, the number of tweets is bound to increase.

We can see similar spikes in the number of tweets of the following countries:

- Finland on 12th June
- Netherlands on 27th June
- Germany on 29th June
- Switzerland on 29th June

All the above are due to football matches that happened on those respective dates.

Part 5. Reflection (20 marks) Using social media to study the real world is very common in academia, media, and industry. Now that you have some experience analysing Twitter data discuss:

5.1 The strengths and weaknesses of Twitter as a data source from a technical/statistical perspective.

Strengths:

- Twitter has become one of the foremost studied platforms for computational social science research. It is one of the most reliable data sources for use in emergencies, disasters, and other extreme situations.
- It's one of the most popular, if not the best, social media platforms for academic research because of its more open and accessible API, tweet search feature, and hashtag culture.
- Allows for near-real-time notifications of an incident's occurrence, as well as easy access to first-hand reports of an incident's impact and community responses to emergency warnings.
- Information is freely shared on this platform which can be retrieved by using numerous tools and APIs available.

Weaknesses:

- Using the API, we can only access 1% of Twitter data, causing data sparsity.
- Each tweet is limited to 140 characters. So, in many cases, the entire meaning of a tweet is hidden behind a link to an image, a website, or a video. This also means that there would be so many acronyms – existing and invented, to suit the need of the day.
- Retweets could be considered as an endorsement of a position or a criticism of it. Figuring out how to handle retweets is one of the difficulties that twitter data presents us with.
- Reply tweets when taken out of context have no clear meaning. To associate the reply tweet to its parent and figure out the meaning is difficult.

5.2 Biases in Twitter data and how they might be mitigated

Bias

- **Participation Bias:** ○ Twitter users do not represent the entire population of the world or any region. The majority of users are younger and have higher incomes than the general population, coming from an affluent background. Eighty percent of tweets come from the top ten percent of most active users. This could imply that we only see the opinions of a small subset of users rather than the opinions of all users. Demographics also cause a type of bias in any statistical analysis. Ex: Twitter's global audience skews male (62%); People on Twitter have above average education credentials, etc - **Bias from Malicious Actors:** ○ Astroturfing: This is a classic case of online astroturfing, in which bot accounts are used to create fake posts and connections. These malicious actors collaborate to make their voices heard at disproportionate levels, resulting in a bias.

○ Crowdturfing: This is the process where people maliciously crowdsource systems in order to spread fake reviews, and spreading disinformation about competitive brands - **Dimensions of Social Data Bias:** ○ This causes the data to be collected in such a way that inferences drawn upon it are skewed.

Mitigation:

There are a variety of techniques available to mitigate each of the above mentioned and many other biases.

- Bot detention approaches can be used to identify bots based on the content of posts, activity, profile, connections, and other factors to mitigate bias from bots.

- To detect workers on crowdtrufing operations, various machine learning frameworks are proposed.
- To avoid Participation Bias, we can use the Firehose API, which allows us to access all of the data rather than just a sample.
- To reduce participation/demographic bias, we could reweight the data using various reweighting techniques based on different factors and attempt to adjust the bias.

5.3 Ethical and legal concerns about using Twitter data

There are legal and ethical implications to using Twitter data posted by people. Someone tweeting during an emergency, for example, may not realise that their tweet will be recorded and analysed — either to aid in the coordination of relief efforts or to undertake an analysis.

Twitter could be considered more of a public space where the profiles and tweets are set to public visibility. "What you share on Twitter may be viewed all around the world instantly. You are what you Tweet!" However, it is disputed whether individual Twitter users are aware of this or adjust their behaviour to accommodate for it.

When social media data is used, traditional notions of informed consent may be challenged. Researchers who might be working on large datasets would find it impossible to gain informed consent from all users included in the dataset simply because some of the accounts may no longer be maintained and some other users choose not to respond.

Although Twitter's terms and conditions note that "Twitter includes the right to make content … available to other companies", when we download tweets using the Twitter Advanced Search, any protection policies offered by Twitter would be void.

It's also necessary to highlight that, tweets posted by users on this platform are not owned by the users, as they're not considered to be original messages. Hence, the right of ownership and copyright law are not applicable.

5.4 The use of Twitter to study the effectiveness of lockdown policies

There has research already been made using Twitter to find out the effectiveness of lockdown policies, analysis of lockdown perception in the Society, understand the effect of health benefits on the society etc. However, those studies are not without limitations. Nevertheless, Twitter is one of the best resources currently available to gauge any real-time policy changes at a huge scale.

- Twitter allows near-real-time communications at a societal scale. The sentiments of the public can be monitored on a real-time basis.
- A Study suggest that "As the pandemic progressed, the portion of positive tweets remained levelled and that of the negative Tweets raised, suggesting that pandemic fatigue, stress, and isolation started taking a toll in how people felt about lockdowns."
- Many similar studies are made to understand various public sentiments followed after major announcements and lockdown implementations.
- But, as has already been pointed out, not everybody uses Twitter and Twitter is not an unbiased representation of the population.
- So, stand-alone sentiments from Twitter do not prove themselves useful in administrative decision making but they can be used as a supplement to other sources of data where they'll provide immense value.

If one is careful enough to avoid misinformation and over-represented sentiments, Twitter could be a very powerful tool to evaluate the effectiveness of lockdown policies or for that matter, any policy which affects people at large scale.

References: https://core.ac.uk/download/pdf/97835144.pdf

https://www.isi.edu/~fredmors/paperpdfs/osnem_preprint.pdf

https://journals.sagepub.com/doi/pdf/10.1177/20539517211013869

https://link.springer.com/content/pdf/10.1140/epjs/s11734-021-00265-z.pdf