Introduction to Data Science - Tweets Analysis – Student ID: 720065364

There is a total of 720 json files containing tweets from the 1st of June 2022 to the 30th of June 2022. Each json file contains 1 tweet per line, with multiple objects describing that specific tweet. Each line contains a tweet object, a user object, geo objects, an entities object and an extended entities object.

The tweet object has fundamental attributes about the tweet such as 'id', 'created_at', and 'text'. It acts as a parent for the other objects mentioned above. [1]

The coursework pdf detailing the instructions to follow has a link to the Twitter Standard v1.1 documentation although an 'extended_tweet' object only present on the Premium v1.1 documentation was found across the 720 files, so it was decided to make use of this documentation as well. [2]

Part 1. Basic Tasks

Question 1:

For counting the number of tweets present, the field 'id' of the tweet object was used. A total of 15.040.387 tweets were found. Not every line of the json files contains a tweet, some lines have something called 'missed_tweets', which collects how many tweets were not collected by the API and provide a timestamp and the count of missed tweets. There is a total of 7757 tweets missing.

To check and remove if any tweets were duplicated, we used the pandas library function 'drop_duplicates()'. We also checked for any null values. In total, there were 15.033.548 unique tweets present in the dataset. The following questions will be answered using these unique tweets.

Question 2:

For this analysis, we used the field 'timestamp_ms' and converted it to DateTime using 'datetime.datetime.fromtimestamp()', which converts the time of the tweets to UTC.

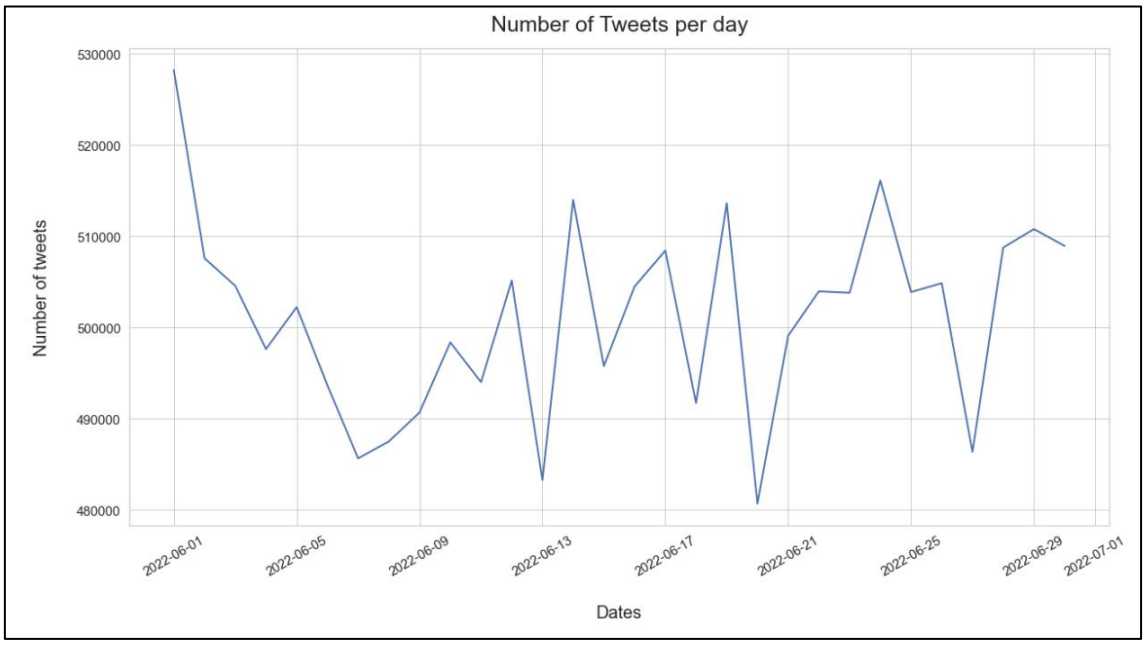


Figure 1. Time series of the number of tweets by day

Figure 1 shows a time series of the number of tweets by day in June 2022. By observing the graph, we can see that there is a pattern that repeats every week. Mondays are the day with the least number of tweets, and then the number of tweets per day starts to go up as the week continues, with a peak on Fridays, followed by a dip on Saturdays before getting to the maximum number of tweets on Sundays.

This pattern could be explained as Mondays are the start of the week, people go back to their jobs, and nothing has happened yet so they do not have a lot of things to talk about, as the week continues, more

things start to happen and people have more things to tweet about. On Fridays and weekends, people have finished work and users have more time to tweet about all the things that happened over the week.

Question 3:

To identify whether a day is a weekday or weekend, we used the 'datetime.date' and the '.weekend()' function which if it returns a value smaller than 5, means that it is a weekday, otherwise if it returns 5 or 6, it is a weekend.

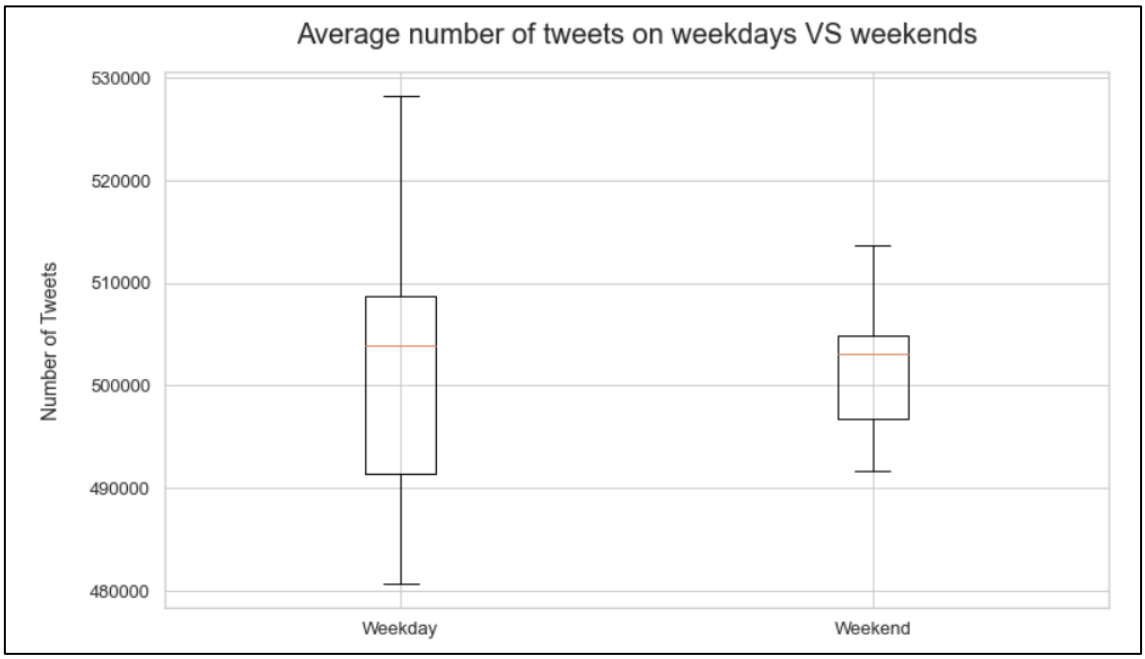


Figure 2. Average number of tweets on weekdays versus weekends

Figure 2 shows a box plot of the average number of tweets on weekdays versus weekends. By looking at the graph, we can see that the medians of each box plot are almost at the same level, meaning there is not a big difference between the tweets on weekdays and weekends. By looking at the interquartile ranges and whiskers of each box plot, on the one hand, we see that on weekends the number of tweets is more stable, meaning both Saturday and Sunday have a similar number of tweets. On the other hand, the number of tweets on weekdays varies more ranging from 480.000 to almost 530.000. Also, as shown in Figure 1, the number of tweets on Mondays is very low in comparison with the number of tweets on Fridays.

Question 4:

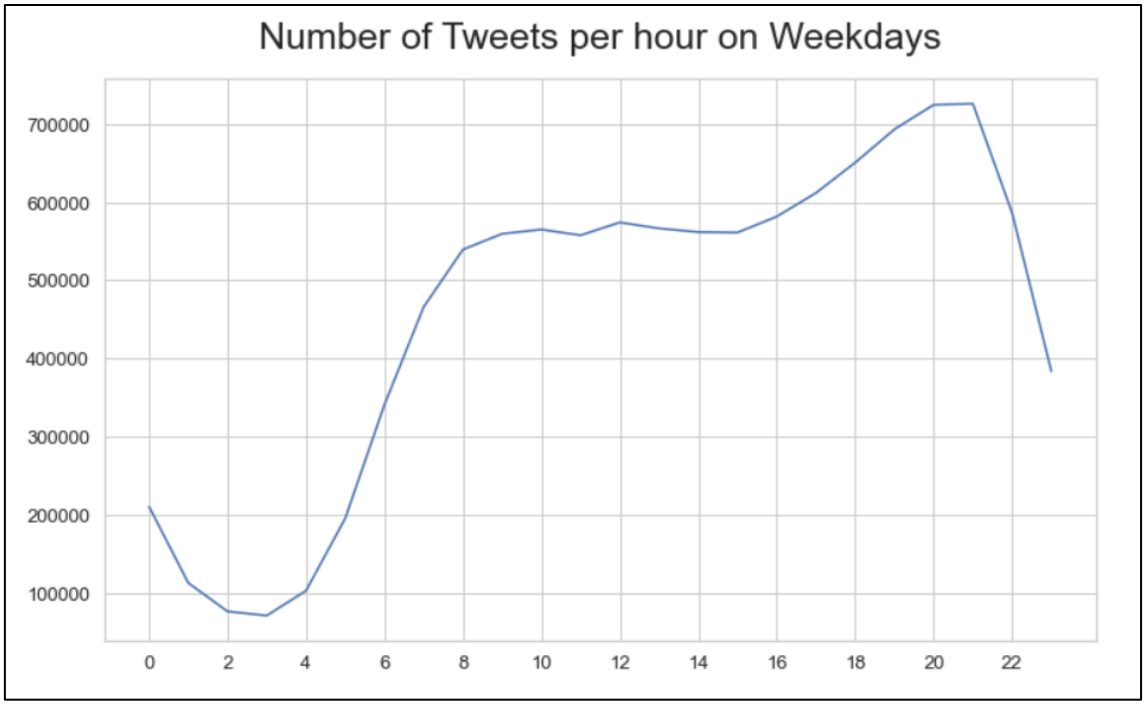


Figure 3. Number of tweets per hour averaged over all weekdays

Figure 3 shows the number of tweets per hour averaged over all weekdays. In this graph, we can see that the number of tweets starts dropping at 9pm having the lowest number of tweets between 1am and 4am. After this period, the number of tweets starts to increase and stabilise between 8am and 3pm, and right after that, they start increasing again to reach the maximum number of tweets of the day at around 8pm and 9pm.

This pattern is explained by the average timetable of a person. When people start to wake up between 5am and 9pm the number of tweets starts to increase, they then go to work where they don't have a lot of time to tweet and the number of tweets stabilises, and as they finish work at 4pm, they start to tweet more until they go to bed at 10pm.

Part 2. Users

Question 1:

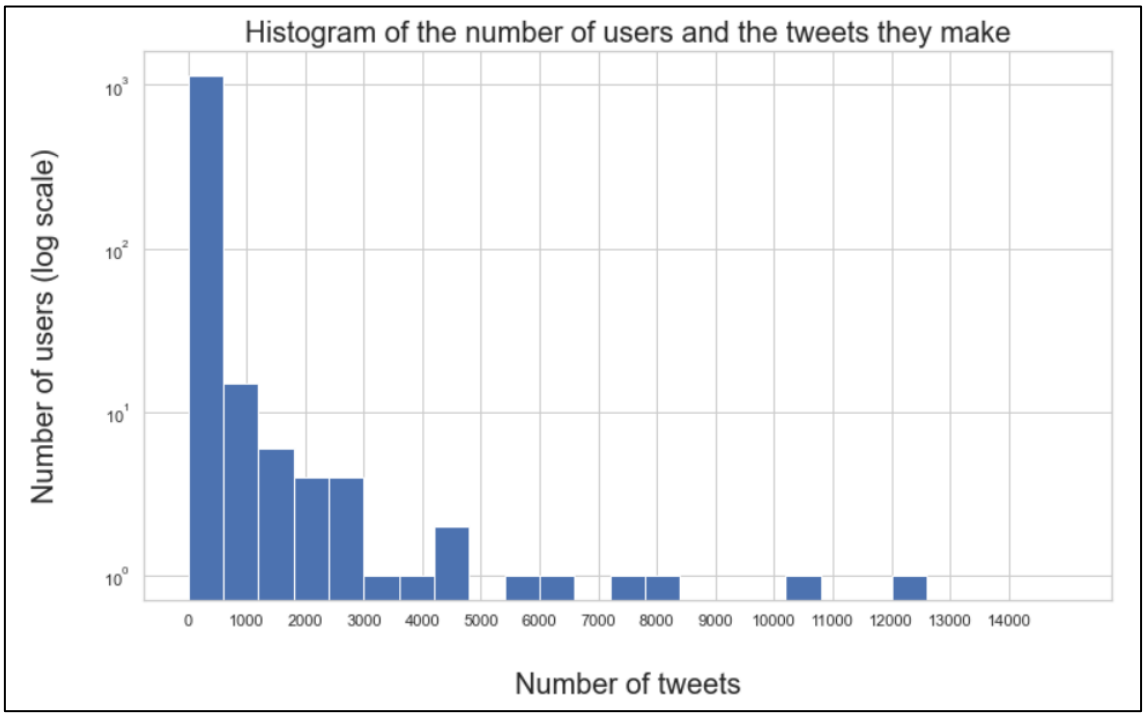


Figure 4. Histogram of the number of users and the number of tweets they make.

Figure 4 shows a histogram of the number of users and the number of tweets they make using a logarithmic scale. This histogram is right highly skewed as there are a large number of users who make less than 3000 tweets per month, and a small number of users who make more than 5000 tweets per month. A logarithmic scale was used in this case as the distribution of the data is very large, having users that only tweeted once a month and other users that tweeted more than 10000 times in June.

Question 2:

The top-5 users by the number of tweets in June are shown in the table below:

Username	Number of tweets
@Kardeimcin1	13376
@DailyNews79	12518
@c_antolic	11628
@HoraCatalana	11294
@minijobanzeigen	10087

Figure 5. Top-5 users by the number of tweets

Figure 5 shows the top-5 users by the total number of tweets in June.

@Kardeimcin1: this account seems to be an automated account. It is a Turkish account that tweets about articles 158, 142, and 245. These articles reference laws about stealing and money laundering. [3]

@DailyNews79: is a suspended Twitter account that used to tweet news articles, probably similar to the Twitter account @minijobanzeigen mentioned below, which tweets every time a job appears on their website. [4]

@c_antolic: I couldn't find much information on this account. Looking at the activity it has, it has to be an automated account. It tweets a couple of times every few minutes throughout the day, and the tweets do not usually make sense (random words, videos, and photos) in German. [5]

@HoraCatalana: this is an automated account that posts the time in the Catalan language. It teaches how to say the time in traditional Catalan. This account is associated with another one which is also an automated account that tweets every quarter past, and every hour. [6]

@minijobanzeigen: this is an automated account that posts about new job offers in Germany being advertised on their website. [7]

Question 3:

```
how_many_users_mentions = []
for each in df['mentions_users_ids'].values:
        if len(each) > 1:
                for i in range(0, len(each)):
                        how_many_users_mentions.append(each[i])
        else:
                how_many_users_mentions.append(each[0])
how_many_users_mentions_count = dict(Counter(how_many_users_mentions))
```

The top-5 users who receive the most mentions can be seen in the table below:

Username	Number of mentions received
@YouTube	17851
@RTErdogan	13225
@BorisJohnson	10050
@BabyDogeCoin	9007
@elonmusk	8731

Figure 6. Top-5 users who receive the most mentions

Figure 6 shows the top-5 users who received the greatest number of mentions.

@YouTube received a total of 17851 mentions. As it is the biggest video platform, a lot of users tweet about videos they have seen where they include the link to the video and also a mention of the video platform.

@RTErdogan received 13225 mentions. Erdogan is the current president of Turkey, and in June he announced his intention to run for re-election in the next elections.

@BabyDogeCoin is a cryptocurrency which received 9007 mentions in June. A year after the launch, CoinMarketCap (a well-known website in the crypto world) listed the coin in June giving it more credibility, meaning a lot of people tweeted about that moment.

@elonmusk received a total of 8731 mentions. Elon is a very active user on Twitter and in June, he started talking about buying Twitter. Those two things are possibly the reasons why he got so many mentions in June.

Question 4:

```
users_mentions_list_GB = []
for each in df_GB['mentions_users_ids'].values:
        if len(each) > 1:
                for i in range(0, len(each)):
                        users_mentions_list_GB.append(int(each[i]))
        else:
                users_mentions_list_GB.append(int(each[0]))
df.loc[df['user_id'].isin(users_mentions_list_GB)].loc[df['country_code'].isin(country_codes)].
.groupby(by='country_code').count()
```

	… GB	… FR	… IE	… ES
GB mentions …	960.821	26.967	47.229	46.297
FR mentions …	71.886	223.765	9.701	30.959
IE mentions …	104.497	3.714	77.657	17.192
ES mentions …	103.791	16.378	10.476	495.377

Figure 7. Mentions between countries

Figure 7 shows the number of tweets a country mentions each other. I have chosen to focus on Great Britain, France, Ireland, and Spain.

Every country except Ireland mentions users from their country the most. It is possible that users in Ireland also mention users living in Northern Ireland, but those mentions are counted for Great Britain. Therefore, that would explain why Irish users mention users from Great Britain the most.

Users from Great Britain mention French users the least, making it a total of 26.967 times in June. This could be because of a language barrier, or also because of the history of rivalry between France and Great Britain, although French users mention Great Britain users almost three times more than Great Britain users do.

Users from Spain and France mention Irish users the least, it could also be because of a language barrier and because not many things are shared between these countries in terms of politics and sports.

Part 3. Mapping

Question 1:

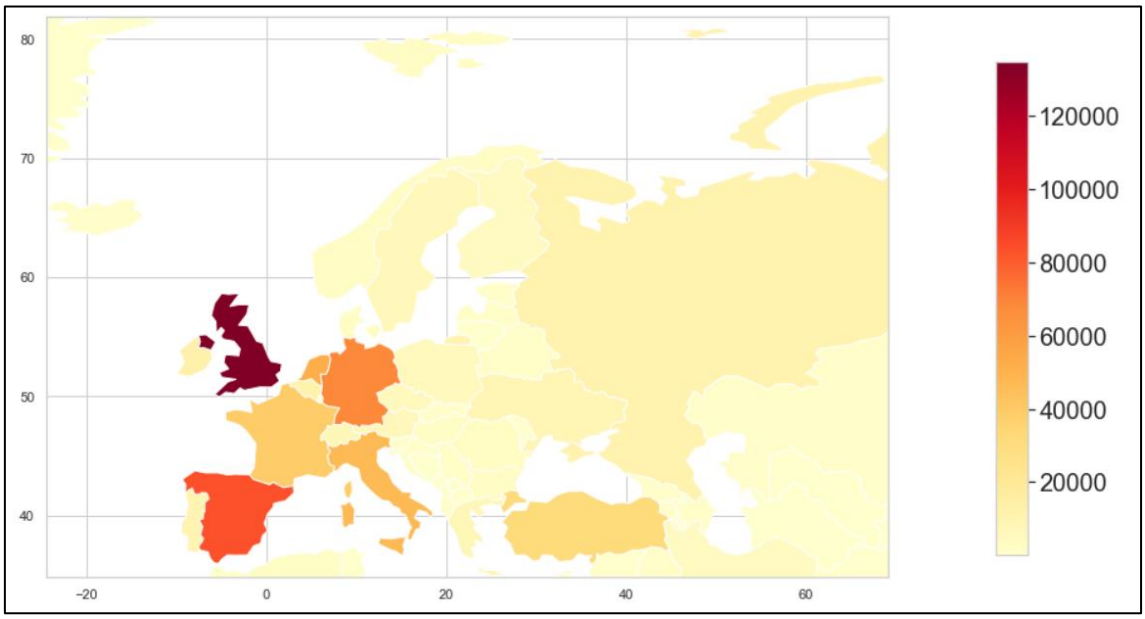


Figure 8. Choropleth map displaying the use of Twitter by country.

Question 2:

The map in Figure 8 clearly shows that Great Britain, Spain, Germany, the Netherlands, Italy, France, and Turkey are the countries where Twitter is used the most. A report from S. Dixon found that Great Britain, Turkey, France, and Spain are within the top-20 countries with the greatest number of users. [8]

Turkey, Germany, Great Britain, France, Italy, and Spain have a population between 45.000.000 and 85.000.000 according to 2022 statistics. [9] In the map, we observe that these countries with larger populations are the ones which tweet the most. This can be explained because the higher the population, the higher users of Twitter. Although this is only true for the countries mentioned above, for example, Russia, Ukraine, Poland, and Romania only have less than 40.000 tweets recorded.

Russia has a population of around 145.000.000 habitants but the number of tweets recorded in June is only around 40.000. This is because the Russian government banned the use of Twitter early in the year and has an equivalent social media called V Kontakte or VK.

Question 3:

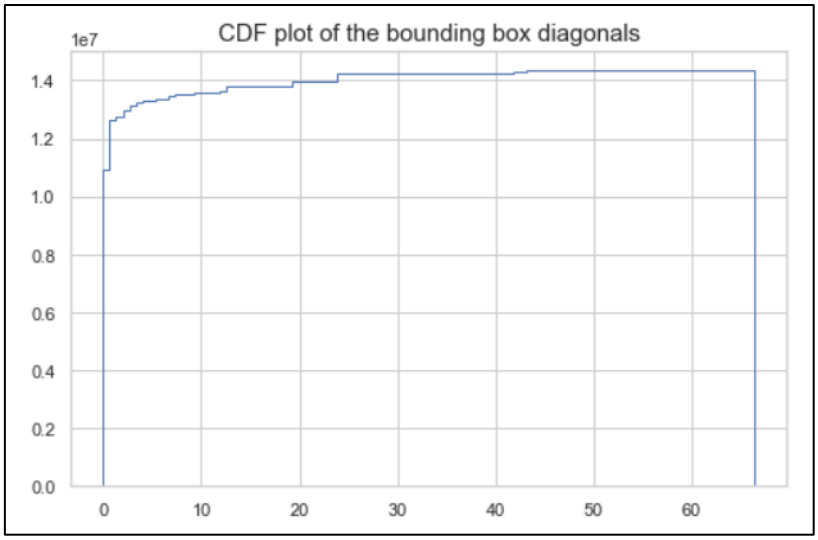


Figure 9. CDF of the bounding box diagonals.

Figure 9 shows the CDF of the distances of the bounding box diagonal of each tweet. The steeper parts of the graph are between 0 and 3, meaning that most of the distances are between those values. The following values of the graph show small increments in the distance, these are values between 4 and 23. After those, the line flattens showing that not many distances are between the values 24 and 67. It then drops to the y-value 0 at the x-value 68 showing the absence of values with a greater distance than 68.

Question 4:

We plotted a dataset containing the average electricity use per capita in kW/h. [11] This dataset is represented below in Figure 12.

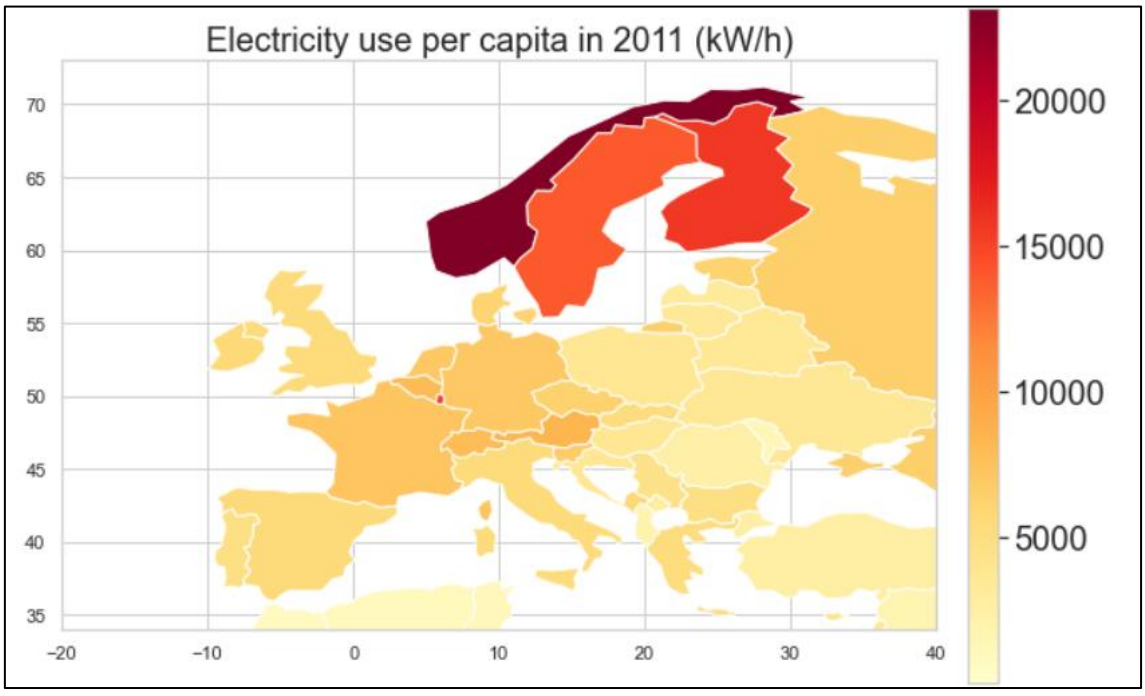


Figure 12. Choropleth map displaying the use of electricity per capita in 2011.

After analysing this dataset and comparing it to the initial one, we could not find any similarities or patterns which could indicate that they are related in any way.

Part 4. Events

Question 1:

We are going to analyse Ireland, France, and Great Britain. An unusual day can be a day where the number of tweets is clearly below or above the average number of tweets for that country. We are going to define unusual days as the day with the greatest number of tweets for each country.

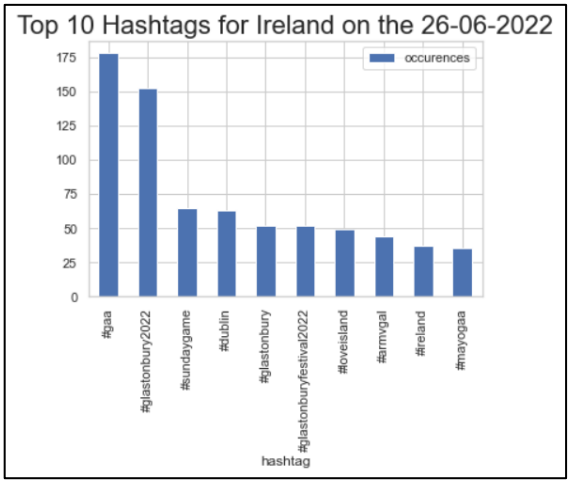
Ireland had the greatest number of tweets on 26-06-2022. It had a total of 11796 tweets.

France achieved a maximum for June on 19-06-2022, with a total of 41537 tweets.

On 24-06-2022, Great Britain registered a total of 133194 tweets.

Question 2 & Question 3:

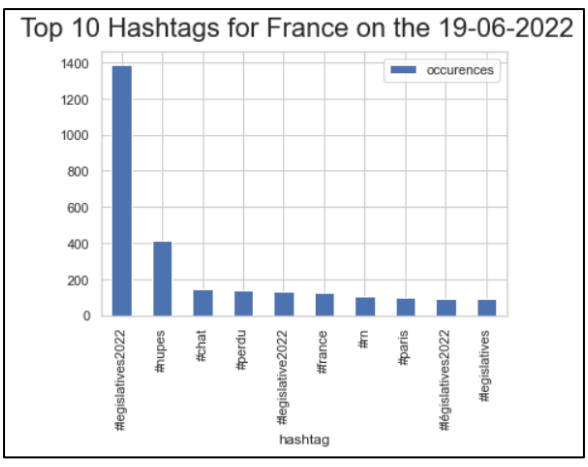
Figure 13 shows a word cloud of the tweets in Ireland on 26-06-2022, and Figure 14 shows the top-10 most used hashtags on that day.



Figures 13 & 14. Word cloud and hashtag analysis for Ireland on 26-06-2022.

As we can observe from both figures above, some of the most repeated words joined with the most common hashtags tell us that there was a GAA football game, that there was a festival called Glastonbury Festival, and users were probably tweeting about the 'time' at which the game started.

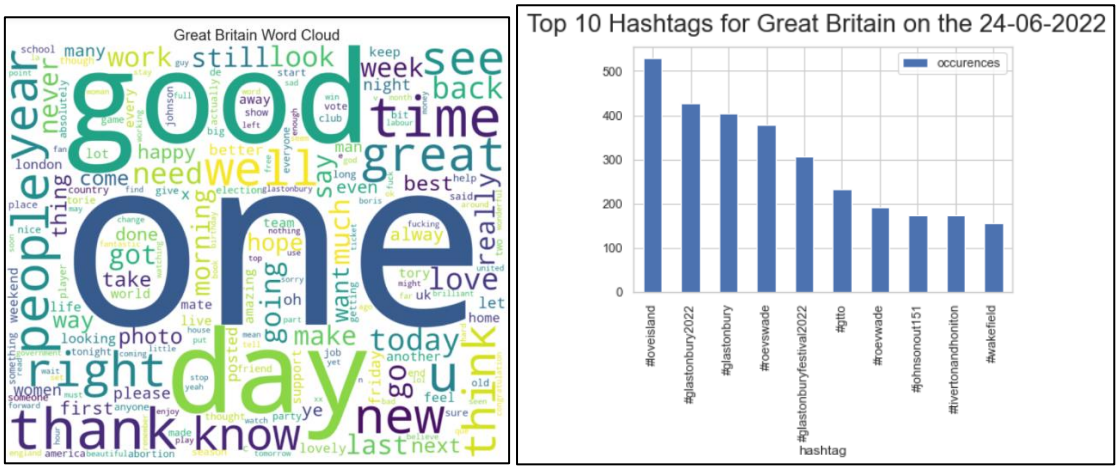
Figure 15 shows a word cloud of the tweets in France on 19-06-2022, and Figure 16 shows the top-10 most used hashtags on that day.



Figures 14 & 15. Word cloud and hashtag analysis for France on 19-06-2022.

On the 19th of June 2022, France had national elections as can be seen from the hashtags and word cloud above. It seems like people were happy with the results of the elections ('plu' in the word cloud meaning 'like') and it was one of the most mentioned topics that day.

Figures 16 and 17 show a word cloud and hashtags used in Great Britain on the 24th of June 2022.



Figures 16 & 17. Word cloud and hashtag analysis for Great Britain on 24-06-2022.

The figures above show a word cloud and the 10 most used hashtags in Great Britain on the 24th of June. The most used hashtag is one referencing love island, which is a reality show program that is broadcasted every Sunday. Some of the hashtags also refer to the Glastonbury Festival. Another hashtag references the Roe VS Wade case which was overturned and made famous across the world on the 24th of June. Some of the words from Figure 16 also show the word 'abortion', 'right', and 'women' which are likely to be linked to the Roe VS Wade case.

Part 5. Reflection

Using Twitter data for research projects in academia, the media, and the industry has some known drawbacks. First, let us dive into the Twitter demographics. The Omnicore agency published an article where they explained the demographics present on Twitter. They found that 70.4% of Twitter users are male, and only 29.6% are female. In the United States, only 23% of adults use Twitter and most Twitters users are between 25 and 34 years old. Only 18% of users come from rural areas, and more than 30% of users have at least a college degree. Twitter is also banned in places like Russia, China, Iran, and North Korea.[12]All of these things mentioned and related to demographics already suppose a big problem as we do not have a good representative sample of the population. Also, as we saw in the histogram of the number of users and the number of tweets they make (Figure 4), the distribution is highly skewed to the right as some users are very active in comparison to others, meaning that we have more data from those users who tweet more often.

Additionally, as we saw in a previous question, the top-5 users in terms of the number of tweets are all bots, meaning that they can affect the analysis of tweets if they, for example, post about a topic a lot, making it seem like people are talking about a certain topic whilst it is the bots making those posts and not humans.

Tweets are also short in length, which sometimes makes the meaning difficult to understand without context. Also, a lot of the tweets are replies to a conversation and without the original tweet, it is very difficult to know what they are trying to discuss.

If twitter data is used in research projects and studies, it should be made clear that they do not represent a good sample of the population as this can affect the outcome and validity of the studies made. Some ethical problems also arise as to if the researchers should inform the tweet users that their information is being used for a particular study. Although contacting users for their agreement is considered a breach of

the GDPR. All of these points mentioned above are things that need to be considered and taken into account when using Twitter data for a research project.

Bibliography

[1] "Introduction | Docs | Twitter Developer Platform." https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview (accessed Nov. 28, 2022).

[2] "Twitter API: Premium data dictionary | Docs | Twitter Developer Platform." https://developer.twitter.com/en/docs/twitter-api/premium/data-dictionary/overview (accessed Nov. 28, 2022).

[3] "(3) Büşra (@Kardeimcin1) / Twitter." https://twitter.com/Kardeimcin1 (accessed Nov. 29, 2022).

[4] "(3) Perfil / Twitter." https://twitter.com/DailyNews79 (accessed Nov. 29, 2022).

[5] "(3) Christian Antolic (@c_antolic) / Twitter." https://twitter.com/c_antolic (accessed Nov. 29, 2022).

[6] "(3) L'hora catalana (@HoraCatalana) / Twitter." https://twitter.com/HoraCatalana (accessed Nov. 29, 2022).

[7] "(3) minijob-anzeigen.de (@minijobanzeigen) / Twitter." https://twitter.com/minijobanzeigen (accessed Nov. 29, 2022).

[8] "Countries with most Twitter users 2022 | Statista." https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed Nov. 29, 2022).

[9] "European Countries by Population (2022) - Worldometer." https://www.worldometers.info/population/countries-in-europe-by-population/ (accessed Nov. 29, 2022).

[10] "2022/W46: Cocaine & Heroin Prices (1990-2020) - dataset by makeovermonday | data.world." https://data.world/makeovermonday/2022w46 (accessed Nov. 29, 2022).

[11] "gapminder Electricity use per - dataset by brianray | data.world." https://data.world/brianray/gapminder-electricity-use-per (accessed Nov. 29, 2022).

[12] "Twitter by the Numbers (2022): Stats, Demographics & Fun Facts." https://www.omnicoreagency.com/twitter-statistics/ (accessed Nov. 29, 2022).