Twitter Data Analysis

CANDIDATE NUMBER: 087074

JANUARY 2021

1 Basic Statistics

1.1

In this research project, I will be analysing twitter data from Europe during the period of March 1st to March 31st 2020. The lower-left and upper-right coordinates of the bounding box used to generate the data has coordinates (-24.5, 34.8) and (69.1, 81.9). The data consisted of 31 compressed files, each storing 1 day of tweets, and was accessible via a SharePoint folder. I downloaded the compressed files and saved them locally to my computer.

1.2

The compressed files are around 450Mb in size and contain millions of tweets stored as JSON objects. A lot of the data in the JSON objects is redundant and would slow down analysis if it is not removed. Therefore, I accessed the files in their compressed format, extracted the required data points from each tweet and created a csv file containing the data that is required for analysis.

To extract the key information from each tweet, I created a function called ExtractInformation which iterates over each tweet of a file and stores the values that have a certain prefix. The code shows how the unique tweet ID of each tweet is located and appended to a list. I stored the date and time, language, country code and user ID of each tweet in a similar fashion.

```
if prefix == 'id_str':   # tweet ID
    tweet_IDs.append(value)
    count_tweet_ID += 1
```

Some of the JSON objects had missing data points. To ensure that this didn't corrupt the dataset, I added a NaN value to the appropriate list so that each index value in the lists contains data on the same tweet.

```
if count_tweet_ID == (num-1):  # tweet ID
    tweet_IDs.append(np.nan)
    count_tweet_ID += 1
```

I applied the ExtractInformation function to each of the 31 files and created a dataset which contained data on 24,803,671 tweets. I then removed the 16,220 duplicate entries from this dataset and formed a csv file called CleanedData which stored data on 24,787,451 unique tweets.

```
df.drop_duplicates(keep=False, inplace=True)
```

1.3

The date and time that each tweet was created is stored in Coordinated Universal Time (UTC). This needs to be converted to the local time zone so that the time that the tweets was created is in the perspective of the user who made the tweet. To convert the date and time from UTC to local time, I grouped each country in the dataset by time zone and amended the date and time column of each tweet in the CleanedData dataset depending on which time zone group the tweet came

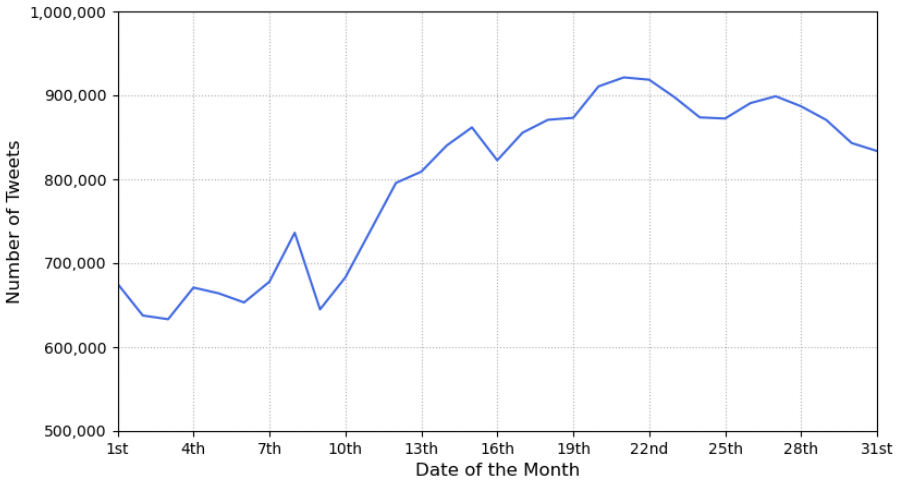
from. For example, I added three hours to the time of each tweet from Georgia and Turkey as the local time of these countries is three hours ahead of UTC. I stored the updated dataset in a csv file called CleanTime.

```
for c in GMT_plus_3:
    df.loc[df.Country_Code == c, 'Date_Time'] = df['Date_Time'] + timedelta(hours=3, minutes=0)
```

I created a dataset called Day_df that contained the number of tweets on each day by grouping the CleanTime dataset by day and counting the number of tweets.

```
Day_df = df.groupby(by=df['Date_Time'].dt.date).count()
```

I then plotted the Day_df dataset on a line graph.



The graph shows that the activity on twitter remained consistent until the 10th of March, fluctuating around 650,000 tweets per day. The number of tweets then started to steadily rise until it peaked on the 22nd of March at over 900,000. The increase in twitter activity is due to Covid-19 starting to spread across Europe and the national lockdowns that this caused. After the 22nd of March, the number of tweets start to decline, this is because the majority of European countries are now in national lockdown and the initial burst of communication that Covid-19 caused is starting to subside.

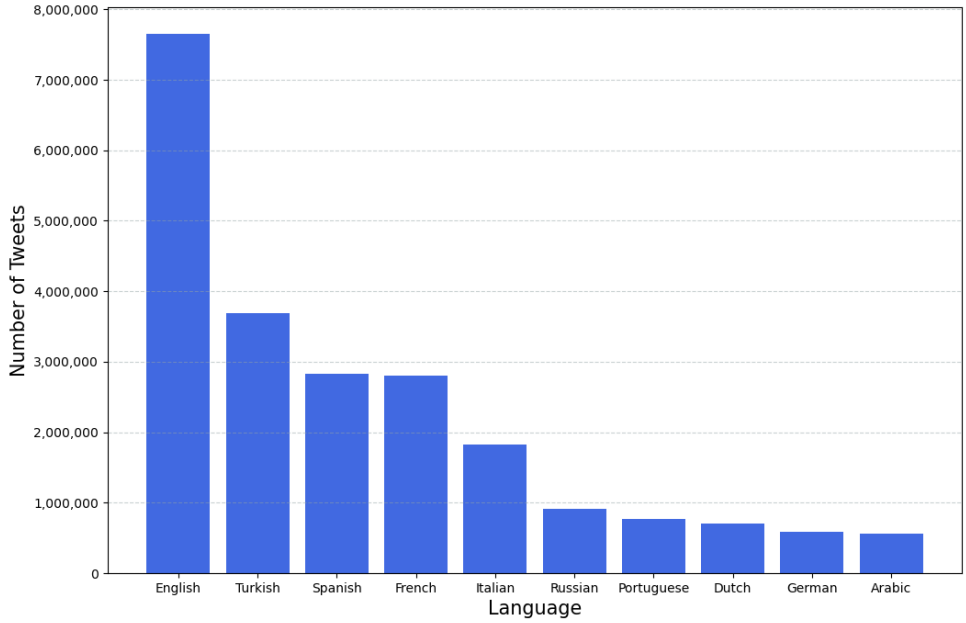
1.4

I grouped the CleanTime dataset by language and counted the number of tweets to create a dataset called Language_df.

```
Langauge_df = df.groupby(by=df['Language']).count()
```

I removed the group with language label ‘und’ as these tweets had an undetected language. 66 different languages were identified in the twitter data, but over 90% of the tweets were written

in only 10 languages. I created a histogram containing the information on the top ten most used languages so that the majority of the language data could be accurately analysed.



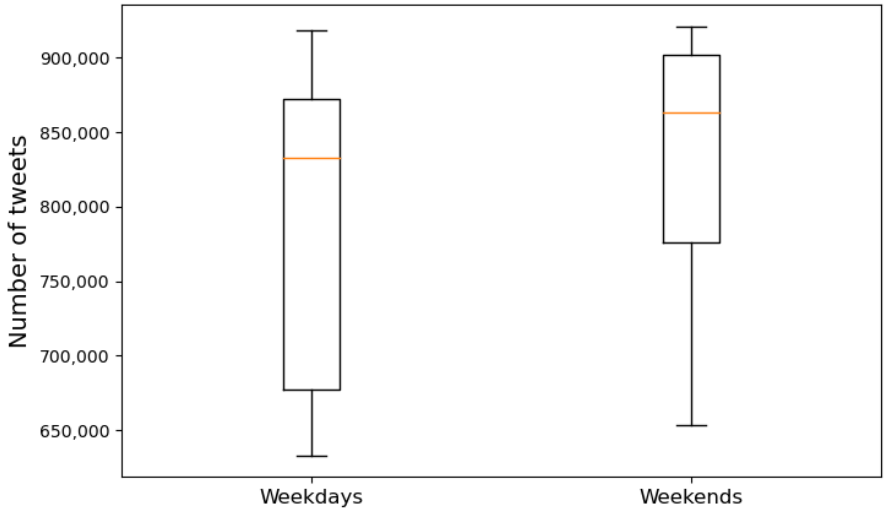
English is the most common language, forming around a quarter of all tweets (7,500,000) and is used almost twice as often as the second most common language, Turkish. Spanish and French form around 2,800,000 tweets each and Italy forms around 1,800,000. After these top five most used languages, the number of tweets per language starts to decline slowly, from Russian with 770,000 tweets to Tibetan with only 1 tweet.

1.5

I filtered the Day_df dataset to obtain two lists containing the number of tweets on weekend days and the number of tweets on weekdays.

```
weekends = df.loc[weekend_dates]['Number_of_tweets'].tolist()
weekdays = df.loc[weekday_dates]['Number_of_tweets'].tolist()
```

I created a graph containing two box plots to compare the distribution of the number of tweets on weekends and weekdays.



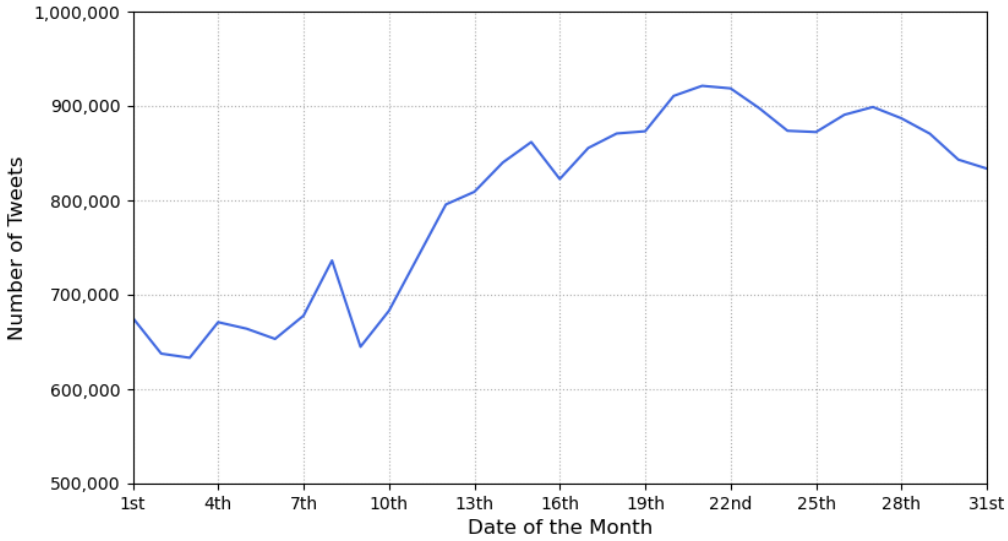
The boxplot shows that, on average there were more tweets on weekend days than there were on weekdays. The Inter Quartile range of tweets on weekend days is significantly smaller than weekdays, this suggests that twitter activity fluctuates more during the week than it does on weekends.

1.6

The CleanTime dataset stores each tweet's date and time in the time zone that it was created in. This ensures that the twitter activity at a certain hour is in the perspective of the user and is not affected by time zones. I grouped the CleanTime dataset by hour and counted the number of tweets to create a new csv file called Hour_df.

```
Hour_df = df.groupby(by=df['Date_Time'].dt.hour).count()
```

I plotted the Hour_df dataset as a line graph to show how the number of tweets varies throughout the day.



The graph shows that the number of tweets per hour is lowest at 5am at around 2,000,000. The level of twitter activity rises quickly throughout the morning and reaches a local maximum of around 12,500,000 tweets per hour at 2pm. Twitter activity then experiences a second increase in the late afternoon and reaches a daily maximum of 16,000,000 tweets per hour at 9pm.

2 Mapping

2.1

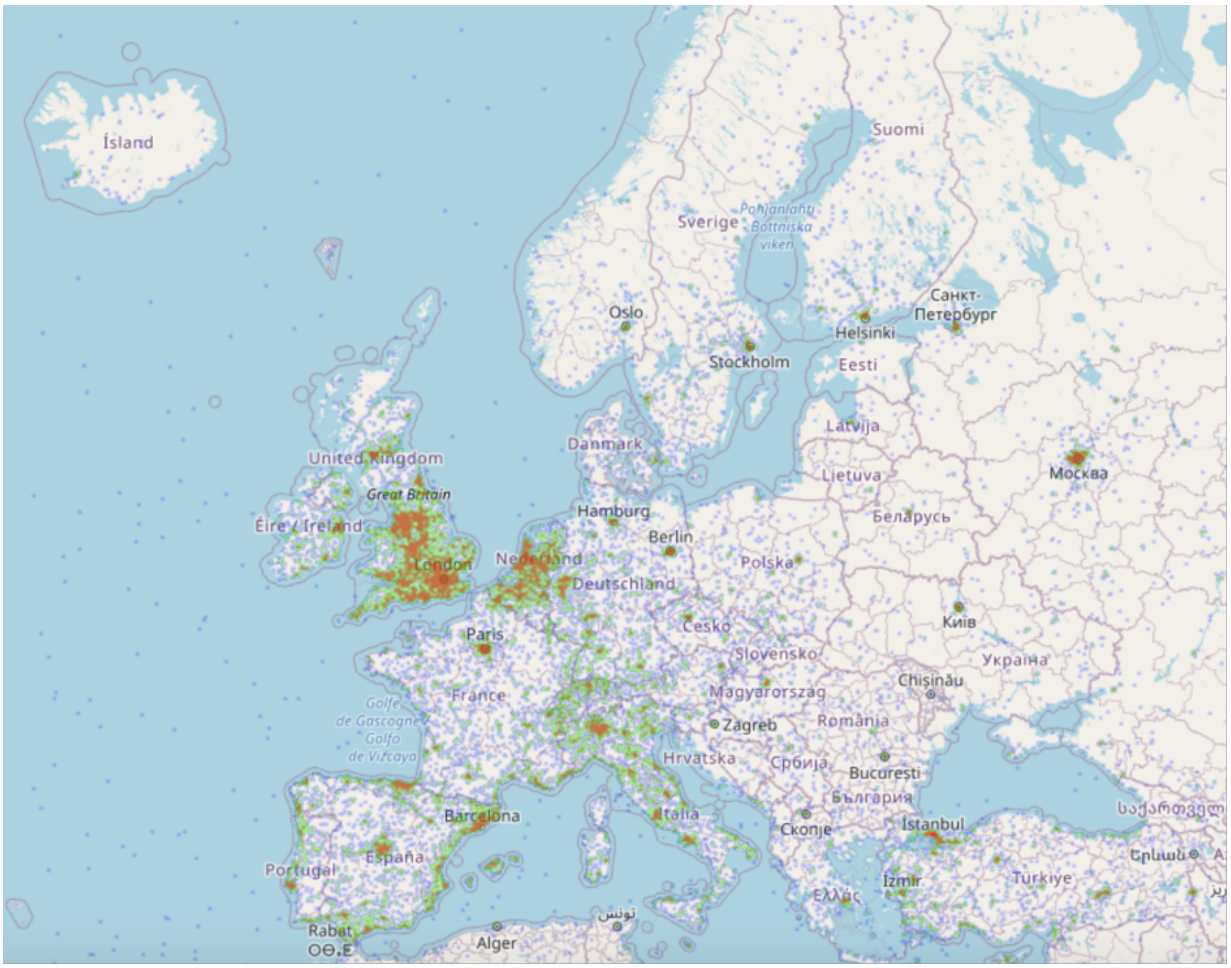
I amended the ExtractInformation function to obtain coordinate data from the tweets. To ensure that the longitude and latitude coordinates of each tweet were bound together, I saved each coordinate pair as a list within a list.

```
h = []
if len(h) == 2:
    h = h[::-1]
    both.append(h)
```

I extracted the longitude and latitude coordinates from each pair to form the columns of a csv file called Coordinate. I then used the Map function from the folium library to create a background map of Europe and inserted a list of zipped coordinates into folium's HeatMap function to project a heatmap layer onto the map.

```
HeatMap(list_coordinates, radius=1.4, min_opacity=0.3, blur=0.5,
        gradient={.3: 'blue', .4: 'lime', .9: 'red'}).add_to(World_Map)
```

The output was a html file showing a tweet density heatmap of Europe. The heatmap uses a colour scheme of blue for low density, green for medium density and red for high density.



2.2

The map shows that Eastern European countries have the lowest level of tweet density in Europe. There is a low to medium level of tweet density in Central European countries. Countries in Southern and Western Europe have the highest tweet density. This pattern of twitter activity strongly correlates with the population density across Europe.

Cities, particularly capitals, stand out in countries as the most tweet dense regions. For example, Moscow, Istanbul and Stockholm all have a dense number of tweets and can be identified as twitter hotspots. This is because cities have higher population densities. England, the Netherlands, Italy and the coast of Spain are the areas with the highest tweet density.

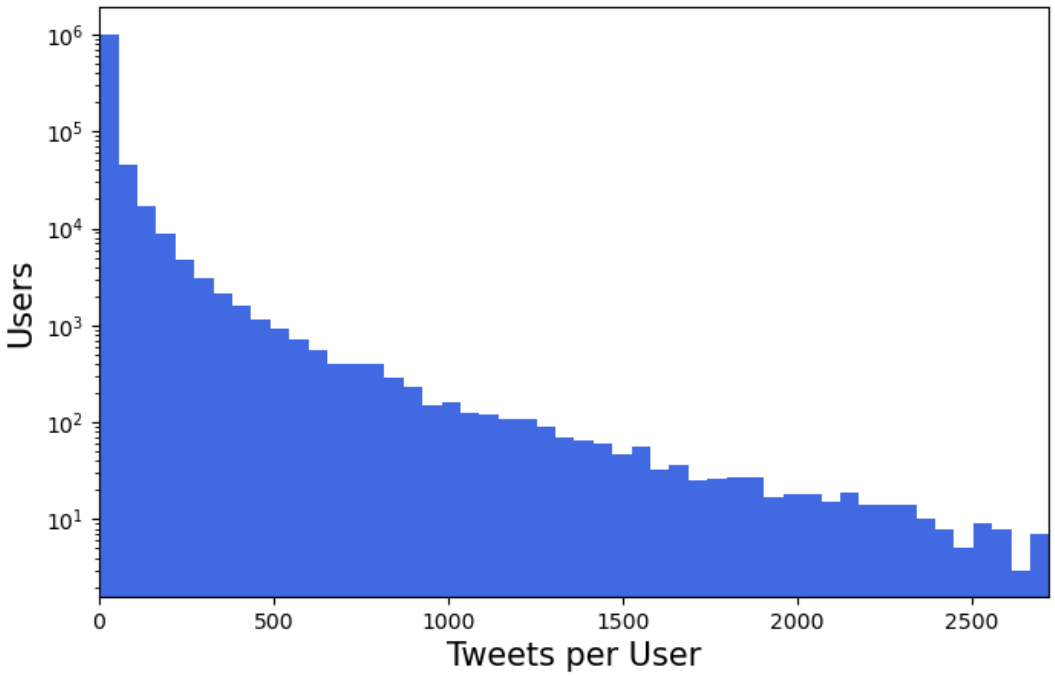
3 Users

3.1

I grouped the CleanTime dataset by user ID and counted the number of tweets to create a dataset with the number of tweets per user. I then sorted this dataset in descending order and created a csv file containing the data called User_df.

```
User_df = df.groupby(by=df['User_ID']).count().sort_values(by=['Tweet_ID'], ascending=False)
```

To plot a histogram of the data, I turned the number of tweets column of the User_df dataset into a list. Preliminary data visualisation showed that outliers were distorting the axis of the histogram. Therefore, I removed the 0.01% of accounts with the highest number of tweets and plotted a histogram of the data with the outliers removed. The y axis is scaled logarithmically because the number of users decays exponentially as the number of tweets increase.



The average number of tweets per user is 16 and 99% of users have under 300 tweets. This is shown in the histogram as the distribution has a very large tail.

3.2

I printed the top ten rows of the User_df dataset to obtain the user IDs of the accounts that had the greatest number of tweets. I then used a website called https://tweeterid.com to convert the user

IDs into usernames to find out information about the accounts. The website could not identify two of the User_IDs, therefore I'm assuming that these accounts no longer exist.

Username	Number of Tweets
@WhatsOnOLIO	37350
Not Found	21420
@AnimalsHolbox	13630
@korayd999	12015
@HoraCalalana	11885
@_BB_RADIO_MUSIC	8565
@RadioTeddyMusic	8546
@haykakan_top	8522
@MathieuRonsard	7604
Not Found	7447

The account with the highest number of tweets is the twitter account of OLIO, which is an app that seeks to connect people to reduce food waste. The twitter account automatically tweets each time a new food item is available on the app. It is evident that the tweets are automated because they have an identical format and occur at a very high rate.

Two of the accounts are created to raise awareness for a specific cause. @AnimalsHolbox is an account that promotes animal welfare and @haykakan_top promotes awareness of the military aggression of Azerbaijan in Artsakh. The tweets of the accounts show minimal variation and occur at a rate that could not be consistently produced by a human, strongly suggesting that they are automated.

@_BB_RADIO_MUSIC and @RadioTeddyMusic are the twitter accounts of European radio stations. The accounts automatically tweet the information of the song that each radio station is playing. The layout of the tweets of both accounts are identical, therefore the accounts must share the same automation software.

The other accounts vary in their specific nature, however they all share similar characteristics which indicate that they employ automated tweeting software. The accounts all have a small number of followers, unhuman like consistency of tweets and the tweets of each account are in the same format. Therefore, I think that all of the ten accounts use a high level of automation in their tweeting.

3.3

I amended the ExtractInformation function to add an account's user ID to a list each time the account was mentioned in a tweet.

```
parser = ijson.parse(new_string, multiple_values=True)
for prefix, type_of_object, value in parser:
    if prefix == 'entities.user_mentions.item.id_str':
        Mention_IDs.append(value)
        count_mention_ID += 1
```

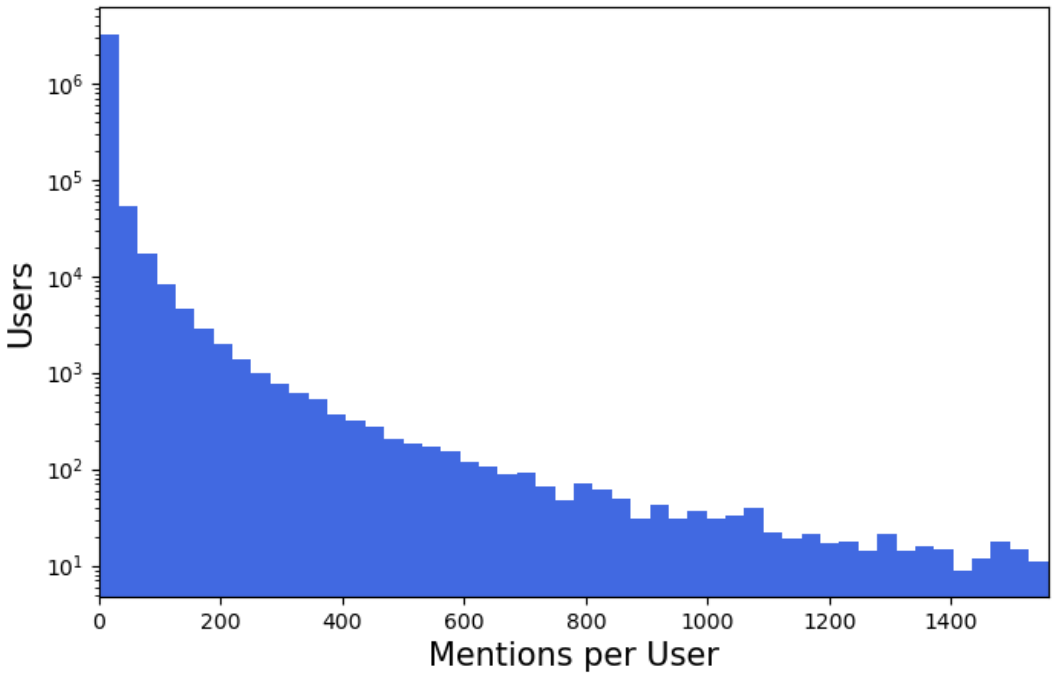
I used this list to establish the unique ID's and the number of times that each ID was mentioned. I sorted this data into descending order and created a dataset called Mention_df.

```
Tweet_ID, Number_of_mentions = np.unique(np.array(total_mention_IDs), return_counts=True)
```

To plot a histogram of the data, I turned the number of mentions column into a list. Preliminary data visualisation showed that outliers were distorting the axis of the histogram. Therefore, I removed the 0.01% of accounts with the highest number of mentions.

```
df_out = df[df.Number_of_mentions < df.Number_of_mentions.quantile(.9999)]
```

I then plotted a histogram of the data with the outliers removed. The y axis is scaled logarithmically because the number of mentions per user decays rapidly as the number of mentions increase.



The average number of mentions per user is 4 and 99% of users have under 75 tweets. Therefore, the histogram has a high density for low number of mentions per user and a long tail.

3.4

I printed the top twenty rows of the Mention_df dataset to obtain a selection of the most mentioned accounts.

Account	Number of Mentions
Boris Johnson	41149
YouTube	37254
Piers Morgan	36493
Dr. Fahrettin Koca	29981
BTS	27382
Donald Trump	23339
Pedro Sanchez	19644
Recep Tayyip Erdoğan	18123
Animal Defence BZ	17995
Matteo Salvini	13013
Sky News	10942
BBC News	9394
Matt Handcock	9005
Emmanuel Macron	8382
Ziya Selçuk	8242
Tesco	7951
Rishi Sunak	7905
Pablo Iglesias Turrión	7657
Good Morning Britain	7478
Süleyman Soylu	7452

The infection and death rates of Covid-19 were comparatively low in Europe during March 2020, however the rate of infection was high causing the virus to quickly spread across Europe. Political leaders implemented national lockdowns in an attempt to stem the spread of the virus, sparking debate on news channels, in parliament and at dinner tables. 12 out of the 20 most mentioned twitter accounts belong to European political leaders such as Boris Johnston, Pedro Sanchez (Prime Minister of Spain) and Recep Tayyip Erdoğan (President of Turkey). This shows that citizens were discussing the national lockdowns and sharing their political views via social media.

BBC News and Sky News are the two main news networks in Britain. During March 2020, these networks provided reliable sources of information for the British public regarding the Covid-19 outbreak. Their twitter accounts were mentioned heavily, indicating that viewers were discussing content that was being broadcasted on their shows. Piers Morgan is a British political commentator and was heavily critical of the government's reaction to the pandemic on his talk show Good Morning Britain. This caused the twitter accounts of Piers Morgan and Good Morning Britain to be frequently mentioned in tweets.

The account Animal Defence BZ is an account that promotes the awareness of animal welfare and was the ninth most mentioned account. I previously identified @AnimalsHolbox as an automated account, this account mentioned Animal Defence BZ in almost every tweet. Therefore, the mentions of Animal Defence BZ are artificial and have most likely been done to promote the account.

4 Events

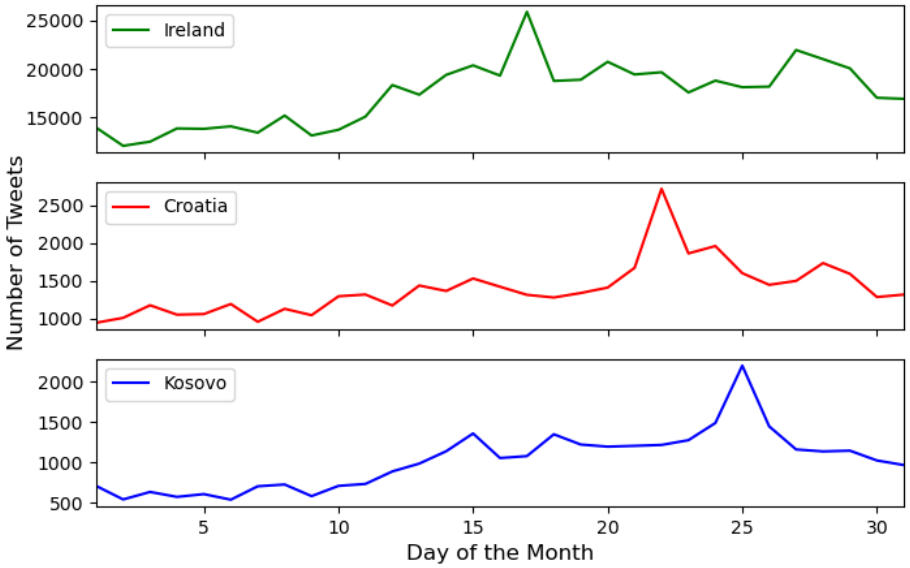
4.1

I grouped the UnionDataset by country and time, then counted the number of instances to obtain the number of tweets in each country on each day. I created a dataset called Country_date_df to contain the data.

```
Country_Date_df = df.groupby(['Country_Code', pd.Grouper(key='Date_Time', freq="D")]).size()
```

The number of tweets that each of the 72 countries in the dataset produces daily varies in magnitude. Therefore, I plotted line graphs containing countries with a similar order of magnitude of tweets per day. By analysing these line graphs, I was able to identify days where there was an unusual number of tweets in a country. A day is classed as unusual if the number of tweets on that day is significantly different from the trend of that country.

I identified three instances of countries experiencing a sharp increase of the number of tweets for only one day and plotted this data on line graphs. The graphs share x axes but have different y axes because the number of tweets in the countries are in a different order of magnitude.



The unusual day in each of the three countries is characterised by a spike in the number of tweets. The 17th of March is St. Patrick's Day which is a national holiday in Ireland, the capital of Croatia was hit by its largest earthquake in 140 years on the 22nd of March [10] and the Kosovo government lost a vote of no confidence on the 25th of March.

4.2

March 17th - Ireland

St. Patrick's Day is a cultural and religious celebration held on the 17th March to commemorate

the death of the patron saint of Ireland, Saint Patrick. The day is a celebration of the heritage and culture of Ireland, characterised by parades, music, dancing and drinking. However, the Irish Prime Minister announced that all St. Patrick's Day parades and festivals in the Republic of Ireland would not go ahead due to Covid-19.

Twitter became an outlet for people to express their St. Patrick's Day cheer. This tweet by an Irish farmer shows his cows marching through a field like people would have been doing through the streets of Dublin.

I made a string containing the tweets in Ireland on the 17th of March and created a word cloud of the string using the wordcloud library.

```
wordcloud = WordCloud(width=3000, height=2000, random_state=1, background_color='white',
                      colormap='Set2', collocations=False, stopwords=STOPWORDS).generate(text)
```

To ensure that the word cloud was not cluttered with common words, I removed 18 words such as 'and', 'it' and 'that'.

The word cloud shows that people were wishing each other a happy St. Patrick's Day as the words 'Day', 'Patrick' and 'happy' were very common. There is a strong sense of community in the tweets, shown by words such as 'Ireland', 'family' and 'all'.

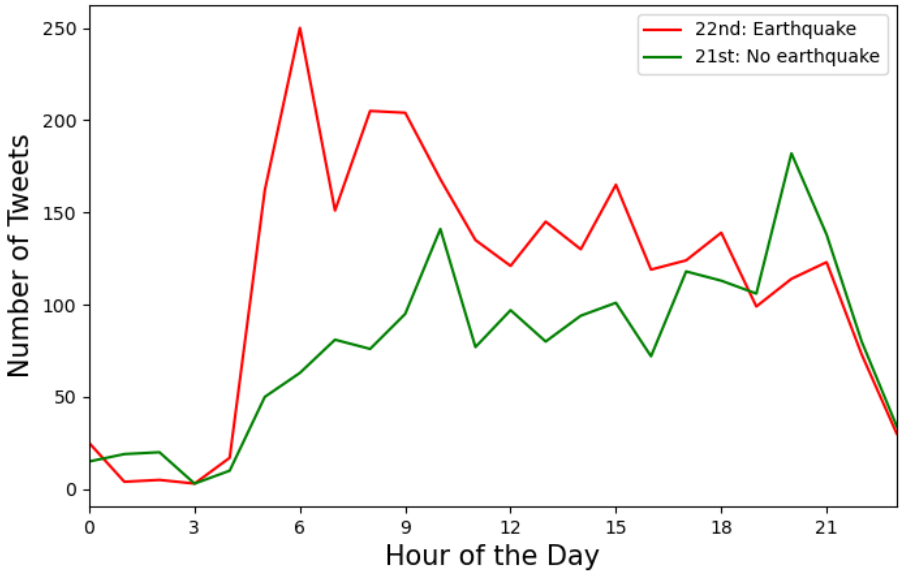
March 22nd - Croatia

At 5:23 am on Sunday the 22nd of March 2020, an earthquake hit the Croatian capital of Zagreb causing substantial damage to the historical city centre. The 5.3-magnitude earthquake was the largest to affect the city in 140 years. The timing of the Earthquake came at a particularly bad time as the country had to "face two serious crises, the earthquake and the epidemic".

I created a dataset containing the tweets that came from Croatia on the 21st and 22nd of March. I did so by applying the ExtractInformation function to obtain lists of tweet data and then removing the data of tweets that did not have the Croatian country code, HR.

```
if country_codes[num-1] != 'HR':
    del tweet_IDs[-1]
    del tweet_texts[-1]
    del datetimes[-1]
    del languages[-1]
    del country_codes[-1]
    del user_IDs[-1]
```

I plotted this data on a graph to compare twitter activity between the two days. There was a sudden spike in twitter activity as the earthquake hit the capital and a slow decline of activity until the number of tweets reached normal levels by the end of the day.



This tweet shows the damage that the earthquake caused to Zagreb.

March 25nd - Kosovo

Albin Kurti was sworn in to be Kosovo's Prime Minister on the 3rd February 2020 after he formed a coalition government with rival parties. Disputes over how to tackle coronavirus caused a

domestic power struggle which culminated in a vote of no confidence being presented by the president of the country, Hashim Thaci. On the 25th of March 2020, Albin Kurti's government lost the vote of no confidence, plunging the country into constitutional chaos when it should have been devising a response to the Covid-19 pandemic. The headline of an article in the Guardian conveys a disappointing tone while the Prime Minister looks on helplessly.

Kosovans look on aghast as government falls while coronavirus bites

Row over sacking of minister leads to coalition partner pulling out

The timing of the no confidence vote caused particular distress to the people of Kosovo. Residents took to banging pots and pans on their balconies to express their disapproval of the political upheaval. The editor-in-chief of Kosovo's biggest daily newspaper wrote "People are very disappointed and angry about what is going on; they feel it's very selfish from the parties and leaders who decided to tackle the government at this particular moment, when we all need to address the coronavirus issue." Citizens of Kosovo expressed their reactions to the loss of the no confidence vote on twitter which led to the sharp increase in tweets.