Introduction

In this paper, you will find step by step discovery of the Twitter dataset that is scraped from twitter. At the beginning of the report, I will give a brief explanation of the dataset, main challenges and my way to encounter these issues. After the introduction, you will find the results, graphs, evidence that answers the tasks questions one by one. Please note that I will keep my words succinct so as to keep the report short and crisp.

The dataset that I received contains 720 zip folders within the zip folder. Each zip folder contains a JSON file that is an hour of the Twitter data in June 2021. The size of data, outside of the zip folder, is almost 51.2 GB so it was impossible to handle this big data on my computer. Loading these files was the main challenge for me to solve. When I load one of the JSON files, I noticed that there are 35 features, columns in other words. The tricky part was that each column contains even more complicated data structures such as lists, dictionaries, tuples within each other. Additionally, I noticed that JSON files were not consistent since some of them had 37, 39 features. As a result, the dataset was big and complicated, additionally, in terms of data quality, it was poor and messy.

To solve the problem, first I go over the assignment to identify what information is needed to achieve the tasks. After my overview, I decided that the following ten features (*id_str, user.id_str, in_reply_to_user_id_str, timestamp_ms, place.country_code, entities.user_mentions, entities.hashtags, text, latitude, longitude*) will be helpful to answers all the questions in the tasks. In the next process, I collected all JSON files into one single folder to extract them in order. Secondly, I have created a pipeline that gets JSON files by a certain number of chunks, flattens the files, selects the features that I need and save them as pickle files. At the end of the process, I collected the entire data in six pickle files each with has a size of 0.85 GB. You can see the script of this process below.

```
for k in range(6): #collect entire data in 5 chuncks
    data = pd.DataFrame() # a data frame to append files one by one

    n=120*(k+1) # termination point for each chunks

    for i in files[n-120:n]: # loop over 120 (chunk size) JSON

        with open(i) as f:
            # normalize (flatten) a JSON file line by line and save in variable
            a = pd.DataFrame(json_normalize(
                [json.loads(line) for line in f.readlines()]),
                             columns=col_names)
            data = data.append(a) # append extracted JSON file in the data frame
            del a # delete variable to save memory and prevent posible problem.

    data.to_pickle(f'data{k+1}') # save the chunk of data in a pickle file.

    del data # delete the data frame to save memore and duplicated of data.
```

Figure 1. Pipeline to extract information from JSON files.

After collecting data into pickle files, each of them has gone through the data cleaning process. This time, I dealt with the quality of data. I have filled missing coordinates by getting data from the place column, have taken text data out from lists, converted data types of features into appropriate ones etc. As the final, step of this process merged and saved the entire twits into one single CSV file. Now, I have a 2.6 GB CSV file that contains entire twits but was not still convenient to use in a pandas data frame. The best way to handle such a problem is to create a database and loading data from there. Therefore I used SQL to create a database and work on it in my project.

To handle all the processes above, I used the following python libraries. zipfile to handle zip folders, json to open files, pandas.json_normalize to extract data within cells, swifter to speed up apply() function on data, sqlite3 and sqlalchemy to create a database and tweak data in the rest of my project.

This project aims to evaluate real-life data from different aspects, such as region of users, usage, a hot discussion topic between them, from a data scientist perspective and discuss some hidden insights of human reactions. Finally, to conclude with the effect of these publicly available data in human life with the pros and cons.

Tasks

The tasks are in a sequential order of a purpose to achieve the aim of the project. I will follow the order as stated in the description document and try to fulfil the expectations by my answers.

Basics Statistics of Dataset

There are *13,861,412* entries (rows) in the dataset, however, some of them are duplicated values. I decide to count the number of unique values of *id_str* and the result was 13,850,314. The difference, *11,098*, between these two values is not trivial, so it is worth digging into the detail. One of my findings is the NaN values in the id column which are *726* in total. On the other hand, I thought some of these data might be duplicated. I checked repeated entries and found *10,372* of them. As you can see the addition of these two numbers is exactly 11,098 which is the difference between the number of rows and the count of unique tweet ids. As a result, I found out some anomalies in the dataset that I need to take care of during my evaluations.

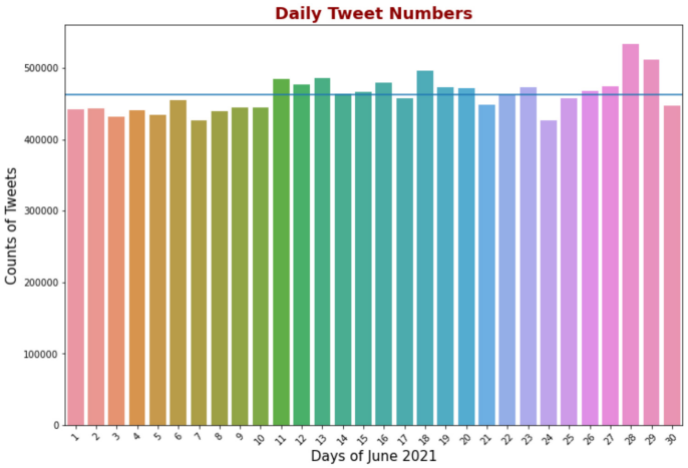


Figure 2. Daily Total Tweets in June 2021 in Europe

The graph above indicates the total number of daily

tweets in Europe in June 2021. The highest number occurred on the *28th of June* with 534,306 tweets. The daily mean value for the month is *462,047.1* tweets. As an interesting result, half of the days in the month are below the average tweet value and the other half is above the mean. Although the total tweet number is highest at weekends and least on Mondays, on the 28th of June which is the highest number of tweets shared by people is Monday. The second most tweeted day is Tuesday the 29th of June which is the following day of the highest value. That means there is an event on the 28th of June that people keep continuing to comment about it.

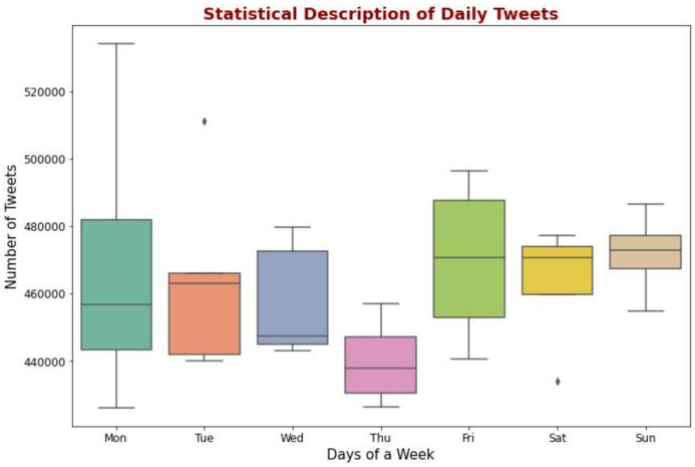


Figure 3. A boxplot describe statistical outcomes of dataset.

A boxplot provides detailed information such as median value, quantiles, outliers about a dataset in other words it is the descriptive statistic of a dataset. The boxplot above describes the daily number of tweets in June 2021. The median value of the number of tweets is the greatest on Sunday which means people, in general, react on Twitter during their leisure times such as Friday, Saturday, Sunday. However, on the other hand, the maximum value occurred on Monday as you noticed the top end of the whisker is the highest. Remember, we have identified that Monday, 28th of June was the highest twitter activity that took place. Thursdays in general people are not active on Twitter.

The size of the box in a boxplot indicates the intensity of the data. Since 50% of the data points are in the box. So, the smaller the box the intenser the data points. In our case, the size of the box indicates the sustainability of the Twitter activity of people. What I mean by that is people are active every other weekend on Twitter, since the box sizes are smaller on Saturday and Sunday.

The line chart below illustrates the average number of tweets per hour. To evaluate the usages on weekends and weekdays, it will be a good idea to use one canvas for both lines to compare the characteristics. In terms of the average number of tweets, we can say that people have a similar tendency to use Twitter on weekends and weekdays because the line graph pattern almost matches each other. However, the peak hour on weekdays is 22.00 whereas it is 19.00 at weekends. In addition to this, the average number of tweets in primetime is more on weekdays.

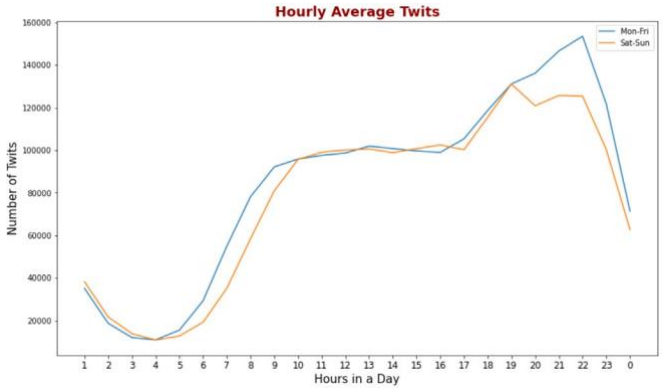


Figure 4. The line graph shows the trends of the Twitter usage.

Mapping Data

There are a couple of attributes such as coordinates, place.coordinates, country code etc within the dataset that allow us to identify the position of the users. However, some of these features were missing in the dataset. In my data evaluation process, I collected all location information into one single column to reach as much as location information to be able to respond the task 2. I stored this region information under latitude and longitude columns and they only have 726 missing data points out of 13.8 M entries.

Instead of python libraries, I used another tool, tableau, to visualise the location of users. The figure below indicates the number of tweets shared from each country within our dataset. I used a colour code to indicate the activities in these countries. The darker the country colour is more active on Twitter.

As you can see the most active country is the UK with 3,633,262 twits during June 2021. Turkey (2,254,437 twits), Spain (1,938,145 twits) and France (1,048,625 twits) are the followings on the leaderboard. Although these numbers are meant as an extended response, it will be meaningful to find the ratio of these values over the population of the countries. This ratio will provide a clear understanding of the usage of Twitter in the countries. I have calculated the result and you can see the percentage of the Twitter user on the table here. As you can see the order for activity of people is different than the total number of twits.

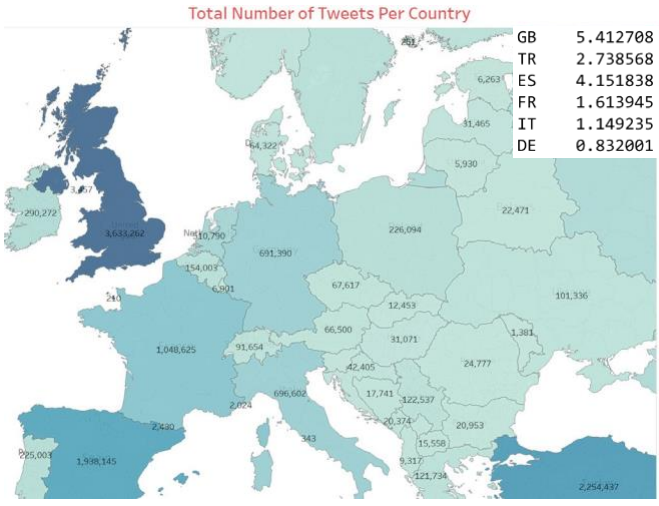


Figure 5. The map show total number of tweets from countries.

Users' Statistics

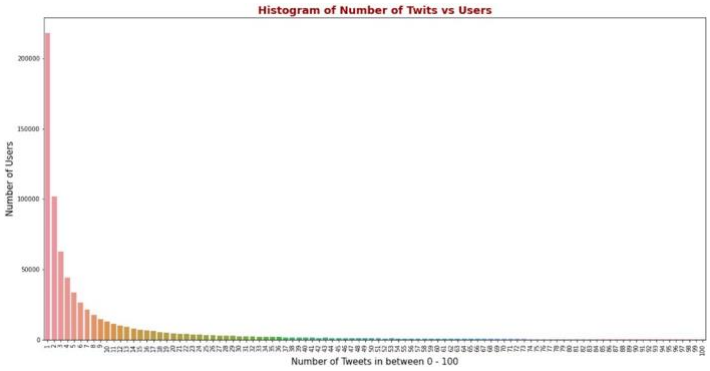
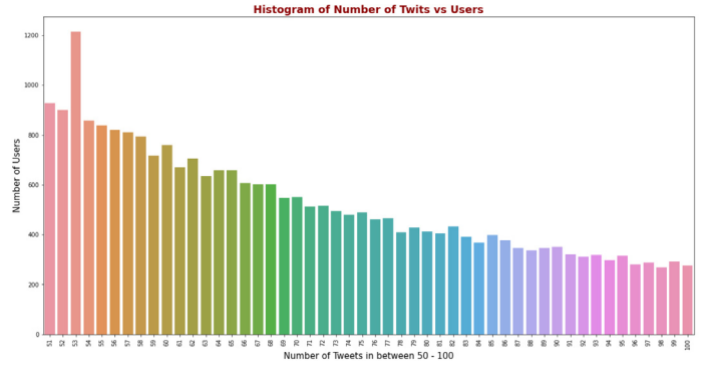


Figure 6. Histogram of number of tweets versus number of users

Figure 6 gives a general overview of or the pattern of the relation of the numbers of tweets with the numbers of users. For visualisation purposes, I subset the first 100 tweets but the layout is identical to the layout of the whole dataset. However below you see the histogram of some other parts as well.

Now I would like to give insights into this graph. As you notice the number of users, *218,091*, who share 1 tweet is dramatically greater than the others. The total number of users in our dataset is *740,659* so 29,4% of our users only shared 1 single tweet. What these statistics tell us is that the vast majority of people follow Twitter although they don't comment about the topics. Or these users might not find the topics in June 2021 important to react to.



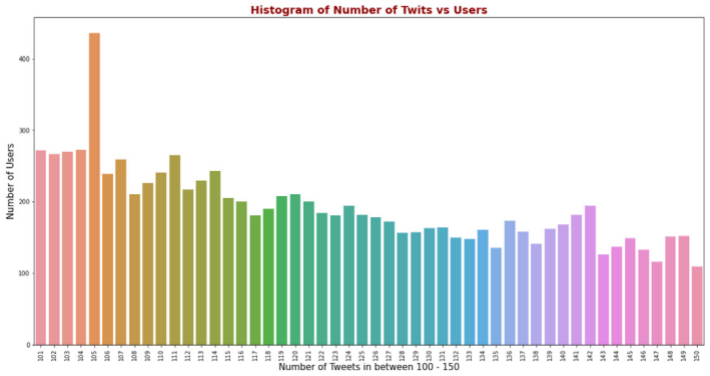


Figure 7. Histograms of some other chunks of data.

Above you see histograms for 50-100 tweets and 100-150 tweets. Although there are few spikes, in general, the number of users is inversely proportional to the number of tweets.

user ID	num_of_tweets
1.384111e+09	11485
1.211606e+18	11178
1.382497e+18	9416
9.741889e+17	9059
1.612628e+08	8736
1.608746e+08	8308
1.352620e+18	7962
6.122073e+07	7877
9.935837e+17	7827
7.225066e+07	6626
4.578781e+08	6350

Figure 8. The leaderboard of the most twits.

In this chart, you can see the leader-board of users in terms of the total number of tweets. The leader has shared 11,485 tweets in total. I assume a person sleeps 8 hours and let say spends 8 hours for work and other human activities finally in the leisure time spend 8 hours non-stop on Twitter so with a little math the leader shares 48 twits in an hour per day. To be honest this is not a real case. Therefore, in my opinion, most of the users within the leader-board are bot accounts.

user ID	total mention & comment
1.315706e+18	18891.0
1.382497e+18	17945.0
1.248631e+18	15730.0
9.286116e+17	15660.0
1.312367e+18	13568.0
1.211606e+18	12967.0
1.321011e+18	12540.0
3.391965e+09	11989.0
3.121726e+08	11730.0
3.531280e+08	11547.0

Figure 9. The leaderboard of the most mentions.

To find the users who got the most reactions on Twitter, I grouped the user IDs and count all the mentions and comments to that user. As a result, you see the users' list with the total interaction number and user ID here.

If you compare the user IDs in the charts above, you will find out that most mentioned people are not the ones who share more tweets. This supports my claim of bot accounts.

Interesting Events

Finding an unusually high activity day is a tricky task. The unusualness may vary by your perspective. I mean, I could say the unusual day is the one in which users shared the most number of tweets. However, from another view angle, you can point out the unusual date is the one in

which a topic is highly mentioned. Although I can handle both approaches by the dataset, I will go with the number of tweets and identify the unusually high activity day by the most number of tweets in the day.

I picked the UK, Turkey and Spain to evaluate the days. The high activity day for the UK is *Friday, 18th of June* with 154,524 tweets, for Turkey it is *Sunday, 20th June* with 88,977 tweets and for Spain, it is *Monday, 28th of June* with 85,304 tweets. As I mentioned above, I could choose these days by the trend topics which would be meaningful too but I preferred this way to ease my job. Below you will find word clouds of tweets for these days to find out what people shared at most these days.

Figure 10. Word Cloud of tweets on 18th of June in the UK

It is clearly seen that the topic is football on Twitter on the 18th of June in the UK. I am not a big fan of football so even I didn't know that it is the time of the Euro cup. I checked the date and game schedules on the web and I noticed that **on the 18th of June England played with Scotland** and won the game. To be honest, it is a shame on me to learn that England and Scotland matched in the same group by my data project.

Figure 11. Word Cloud of tweets on 20th of June in Turkey

I searched these words on google and it is clear to me that people are celebrating father's day. In addition to this, there is a special name in the content who is the coach of the Turk- ish national team. I read the news about him at the time, he has been fired after the team`s bad performance of theteam in the Euro Cup

And finally, the last one in Spain. The main topic again is football and a little bit of politics. Spain played against Croatia in Euro Cup on the 28th of June and won 5-3. In addition to this people also talk about a political party in Spain but unfortunately, I could not find anything about it on the web.

Figure 12. Word Cloud of tweets on 28th of June in Spain

Final Words

Technically and statistically speaking, Twitter API provide valuable and huge data that a data scientist can discover a lot of characteristics, tendency and will of the users. In this project, I received 52 GB of data from Europe for a month of interaction. The data is being provided via JSON format which is a widely used unstructured data format and able to store tons of information. However, on the other hand, handling this data is not easy. To get some reliable results and to generalize them, you need to handle multiple features and maybe a year-long data at the same time and this is impossible to do with a personal laptop. Therefore, the pros of Twitter is the amount of data they provide, but the downside is handling this data is not a trivial task.

Secondly, there were some problems in the quality of Twitter data such as duplicated values, inconsistent number of columns, and missing values. Although we are able to fix some parts of these values, it is important to conduct a reliable data collection process to prevent these issues. Twitter should be aware of the problem and set the data collection process accordingly.

In terms of privacy, the data that I worked on was seriously scared me. As an aspiring data scientist, I can identify users' political, religious, or any other opinion, I can find their location and address and target them for harm so I don't even want to think about what a malicious professional person can do. Therefore, this data is totally sensitive and vulnerable to attacks. That's why the GDPR guideline strictly exists here in Europe.

Social media data is an enormous source to evaluate the social reactions of people. It is essential for the policymakers to know the feeling of people about certain regulations and adjust them in accordance with the public's' expectations. For instance, announcing lock-down without getting input from communities may trigger other harmful activities so it will be better to evaluate social behaviours in a timely manner to align the lock-down procedures accordingly to prevent any side effects on the public's' social well-being.