TWITTER ANALYSIS COURSE WORK

Tasks Part Part-1: Basic Stats

ANS:1.1

After removing the duplicates and anomalies the total number of tweets are 13857163. The Duplicates and anomalies are removed by the function nunique().

The total number of tweets in the data set before removing the duplicates or anomalies are 13861412.

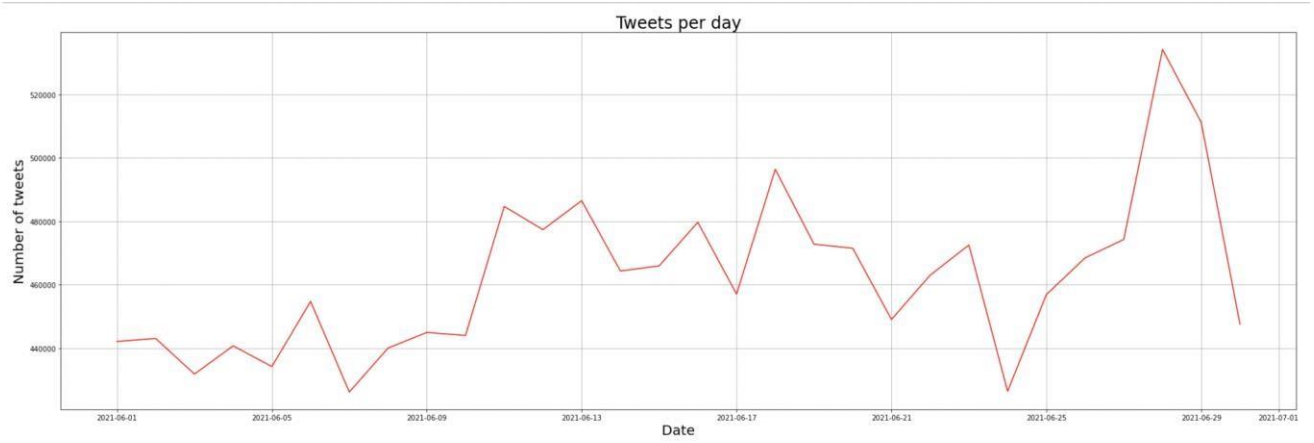
```
data['id_str'].nunique()
```

```
13857163
```

ANS:1.2

```
data['normalised_date'] = data['time'].dt.normalize()
bydate=data.groupby('normalised_date').count()
```

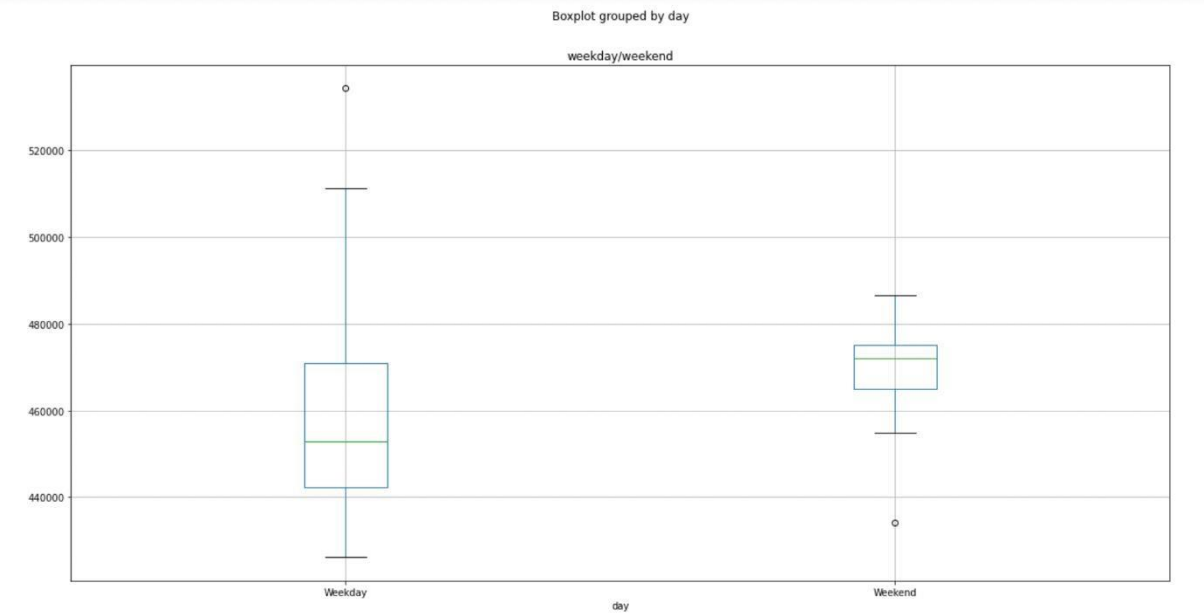
From the time-series plot below we can observe that average number of tweets in a day are above 40,000. The number of tweets is low at the start of month but after 11th June 2021 there is a sudden increase in number of tweets with the start of Football European Championship**.** We can observe decrease in number of tweets on 24th June 2021 and 31 June 2021 when there are no matches scheduled. We can observe highest number of tweets are made on 28th June 2021 when there are two matches held.



The above plot shows number of tweets tweeted per day by users from 01-06-2021 to 31-06-2021.

ANS:1.3

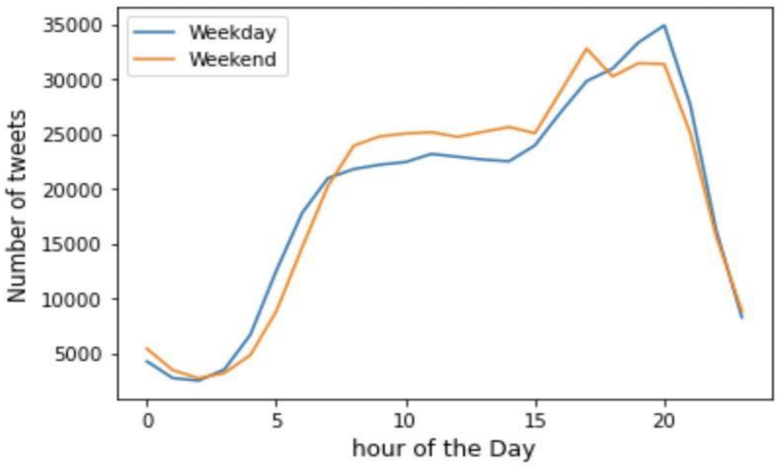
```
Boxplot=data.groupby(['normalised_date']).size()
Box=pd.DataFrame(Boxplot)
Box['day']=np.where(Box.index.dayofweek>4, "Weekend", "Weekday")
Box1=Box.rename(columns={0:'tweet'})
```




```
The Boxplot above shows average number of tweets made on weekdays in the dataset to the number in weekends.
```

The box plot shows average number of tweets on the weekdays in the dataset to the numbers for weekend days. From quartiles we can observe that minimum number of tweets are higher on weekend when compared to a weekday, but maximum number of tweets are higher on weekday. We can observe that on a particular day there are more tweets made on weekday which can be seen on weekday outlier. We can observe Average number of tweets on weekday is lesser compared to weekend. The data is right skewed on weekday and left skewed on weekend.

ANS:1.4

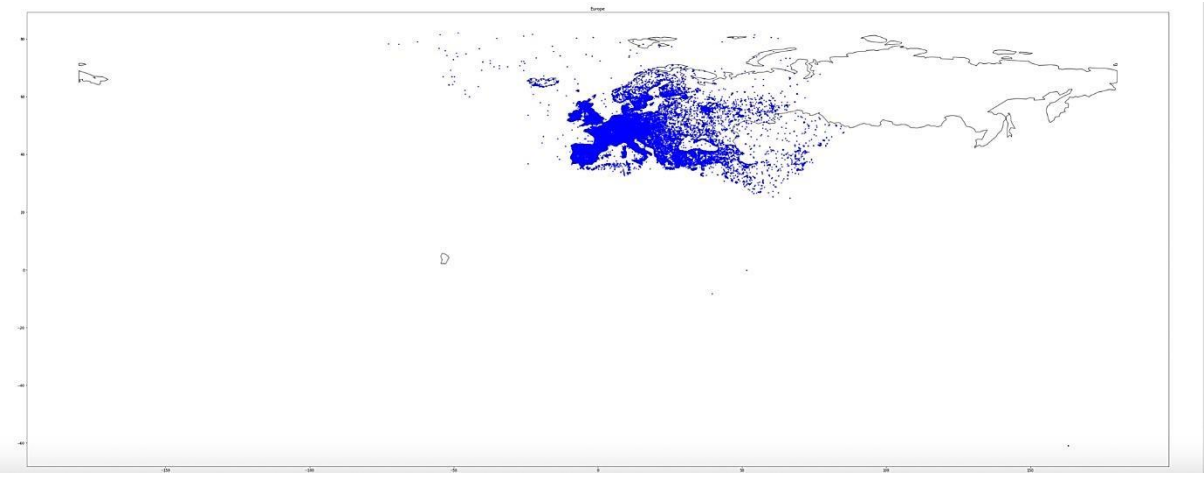


We can observe less number of tweets are made at the start of the day, after 5 hours there is a rapid increase in number of tweets made this can be because most of the users start their day after 5:00. We can also observe most tweets are made after 16:00 hours. It can be because most of the football European championship matches start after 16:00 and then we can see a decrease in tweets after 21:00. This can be because of most users end their day.

Part 2. Mapping

Ans:2.1

```
geo=gp.GeoDataFrame(data5,geometry=gp.points_from_xy(data5.long,data5.lat))
world = gp.read_file(gpd.datasets.get_path('naturalearth_lowres'))
axes=world[world.continent=='Europe'].plot(color='white', edgecolor='black',figsize=(60,60))
```



The above figure describes a map of Europe showing the location of the GPS-tagged tweets.

Ans:2.2

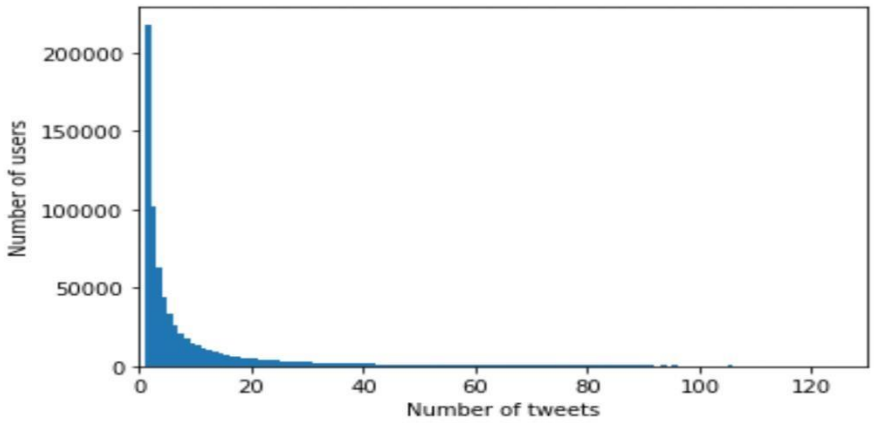
From the figure above we can observe that there are less or no tweets in some eastern parts of Europe which contains a part of Russia. This may be because of new laws on twitter in Russia in the recent times. We can also observe there are more no of GPS tagged tweets on western part of Europe. By this

pattern we can assume that most of the twitter users in western part of Europe are interested in freely sharing their GPS location. This helps in understanding the users interest in particular region. It also helps in understanding the problems, occurrence of events in a particular location.

Part 3. Users

Ans:3.1

```
data_user=data['user.id_str']
data_user.dropna()
user_count=data_user.groupby(data_user.values).size()
val=[]
for i in range(740650):
    val.append(user_count[i])
```

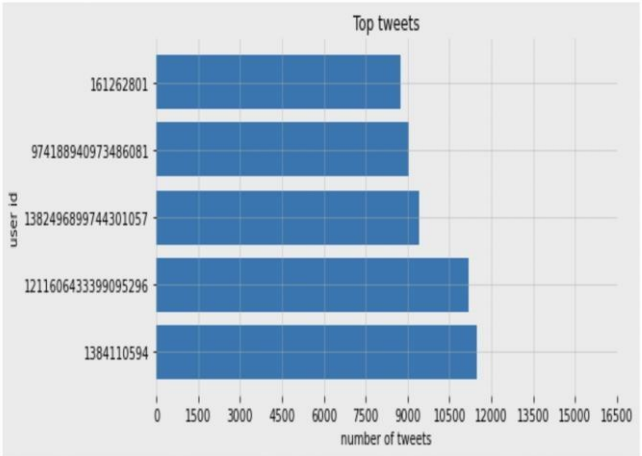


The above plot describes number of tweets made per user. We can observe most of the users tweeted less than 20 tweets for the given data. I have avoided the users with highly unusual number of tweets these are tweets made by bots mostly. We can observe inverse Gaussian Distribution from the above figure which is positively skewed. As the number of tweets increased the number of users who tweeted decreased rapidly.

Ans:3.2

```
user_count=data.groupby('user.id_str').size().sort(ascending=False)
```

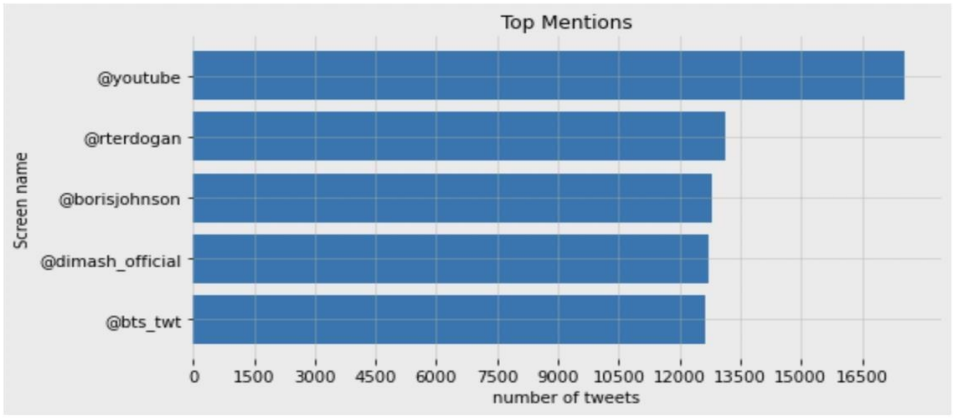
	number of tweets
1384110594	11485
1211606433399095296	11178
1382496899744301057	9416
974188940973486081	9059
161262801	8736



The figures above show the top five users with highest number of tweets in the given data. It is highly impossible for a human to make this high number of tweets in the given period of data. So, we can conclude these users are bots.

Ans:3.3

```
import advertools as adv
data.text=data.text.astype('str')
mention_summary = adv.extract_mentions(data['text'])
mention_summary.keys()
mention_summary['top_mentions'][:5]
```



The figure above describes the top mentions of users in the given data. We can observe top accounts mentioned are very popular verified accounts with most number of followers. The top account is social media app account and second mentioned is prime minister of United Kingdom. Third mentioned is president of Turkey. Fourth and fifth is accounts of popular singers. We can observe most of mentions are in political, social and entertainment category.

Ans 3.4

In the user entities we don't have any columns related to mentioned user country. With the limitation of data (1st June to 30th June) it may not be possible to compute mentioned user location. Part 4. Events

Ans 4.1

```
highest=data.groupby(['normalised_date','place.country_code']).size().sort_values(ascending=False)
highest.head(50)
```

I have observed unusually high activity of tweets in the following countries:

a) Great Britain on 18th June 2021
b) Turkey on 20th June 2021
c) Spain on 28th June 2021

Three of the countries played European Championship football on the respective days with competitive opponents. It could be because of fans tweeting to cheer their respective teams on those days lead to highly usual number of tweets.

Ans 4.2

```
data1=data[data['normalised_date']=='2021-06-28']
data_gb=data1[data1['place.country_code']=='ES']

from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
import pandas as pd

comment_words = ''
stopwords = set(STOPWORDS)
new_words =open('ignorewords.txt').read()
new_stopwords=stopwords.union(new_words)
for val in data_gb.text:
    val = str(val)
    tokens = val.split()
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()
    comment_words += " ".join(tokens)+" "
wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = stopwords,
                colormap='RdBu',
                normalize_plurals= True,
                collocations=False,
                min_font_size = 10).generate(comment_words)
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)
plt.show()
```

The Above Word Cloud figure describes the tweets in the Great Britain on 18th June 2021. From the figure we can understand that a football match was held between England and Scotland as part of European championship. We can observe many fans expressing their emotions and thoughts about the match. We can observe the match was held during the late hours of the day.

The Word Cloud figure describes the tweets in the Turkey on 20th June 2021. From figure we can observe people mostly tweeted related to Father's Day celebration and football match help between Turkey and Switzerland. We can observe fans tweeting their emotions and praying god to help their team secure a win. We can observe people showing disappoint in tweets from which we can say their team lost the match.

The Word Cloud figure describes the tweets in the Spain on 28th June 2021. From figure we can observe people mostly tweeted on football matches held between a) Spain and Croatia b)France and

Switzerland. We can observe many fans expressing their emotions and thoughts about the match. We can observe users expressing happiness in their tweets which say Spain won the match.

I have validated the data by looking for the events of specific dates for mentioned countries and observed their occurrence in the headlines on google search and news websites.

Part-5: Reflection

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

Answer:

Twitter as a data source is well structured and openly accessible, which has valuable metadata including geographical information, can be used for a wide range of research such as social science research. As the popularity of social media has grown in recent years, so has the number of Twitter users. Therefore, the Twitter data source has a large amount of data on various topics that can be useful in research. It can be used to analyse the user's interest, trending issues/topics of past or current time. Statistical results on data can also be used to observe political events, pandemic situations, disasters and take necessary measures in real-time. But handling a large amount of data is quite a difficult task technically and needs a lot of time to refine data to get desired results. Hashtags derived from Twitter data can be used to determine user's interest in a topic. Even though statistical results obtained from research on a topic for a particular location are of some use, they might not be accurate to make conclusions as many people might not use Twitter or the internet. Also, most of the tweets are made by highly active users, so that data collected from Twitter may not necessarily contain opinions of different users. Nowadays we see a lot of Bot tweets on any particular topic, which skews the statistical research results. Users may use a hashtag of trending topics to their posts to grab interest, which may skew the analytic results on that topic. Although Twitter data is very useful in research, a lot of data should be refined to avoid noises and get valuable results.

5.2 Biases in Twitter data and how they might be mitigated.

Answer:

Twitter data has a lot of tweets made by my bots which creates a lot of noise in data. These tweets affect the results of research, particularly when taken in a specific geographic location and limited time frame. For example, taking tweets of the Europe region for a one-month time frame. This can be mitigated by avoiding the tweets from user ids whose number of tweets is unusual for the given time frame.

Retweets can also affect the results of the data, this can be mitigated by taking the original unique tweets and removing the duplicates and anomalies. Twitter data is provided in UTC time which may affect the results when we compare the day-to-day activity of data for a different time zone. To mitigate this we have to convert the time stamp to the time zone of the location of data.

The Twitter data source has a large amount of data related to different topics, regions, time periods. It is a difficult task to load that huge amount of data and work on it. This can be avoided by avoiding the columns of data that are not required for analysis and refining the selected columns.

5.3 Ethical and legal concerns about using Twitter data

Answer:

Many Twitter users are not aware that their tweets are available for analysis. Data from Twitter is used by many academic researchers to understand the user's behaviour in different circumstances. When working on large datasets, it is not possible for the researcher to approach every user for informed consent as it is very time-consuming and needs a lot of physical effort.

“By agreeing to Twitter terms and service agreement, users will consent for their information to be collected and used by third parties (Twitter, 2016A).”

Reuse of twitter data is well inside the Twitter privacy policy without the consent of the user, but it is important to understand that scrapping of tweets is against the policies

“Scrapping Services without the prior consent of Twitter is expressly prohibited” (Twitter, 2016B)

Also, removing the user id and reproducing /altering the tweet is against the policy and it is significant to understand and work according to the policies.

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

Answer:

Twitter data was very useful in understanding the outbreak and behaviour of its users during the COVID-19 pandemic. Different plans were implemented by the government with Twitter data using the geographical location of the user. The movement of users was tracked, lockdown policies were implemented where there is an increase in cases and large movement in people. Users used Twitter as a source to get information on the availability of beds in the local hospital, availability of medicine in local pharmacies, this information helped the government in understanding the rate of infection of virus in that location, hence helping in taking precautions like increasing the medical facilities and decreasing mobility of users. Users posted their thoughts, emotions, beliefs in the vaccine, this helped the government in understanding the interest of people in the vaccine. Twitter data is also used post vaccination in identifying the locations where there is more public gatherings hence mandating rules of wearing masks.