

IDS Coursework

Joshua Robert

December 5, 2022

Important

Code: Codes will not always be shown, but the logic will be written to increase readability and engagement.

Assumptions: The Twitter data does not have time offsets therefore this analysis did not consider local time zones. The calendar starts from 2022-06-01 — 2022-06-30 in this analysis.

Finally, the following sections present the solutions.

1 Question One

1.1 Solution

Code

Note: The codes below represent the core preprocessing utilised to solve other questions, with only the first function changing to suit a need.

```
# exclude tweet with no id, timestamp and user dictionary
# returns None for the excluded tweet
def question_1(tweet):
    if (t_id:=tweet.get('id_str')) and
        (timestamp_ms:=tweet.get('timestamp_ms')) and
        tweet.get('user'):
        return t_id, timestamp_ms

# returns only within the range of June 1st 2022 to June 30th 2022
def juneOnlyHour(panda):
    panda['time_utc'] = pd.to_datetime(panda["timestamp_ms"],
        utc=True, unit="ms")
    panda['time_utc'] = panda['time_utc'].dt.floor('H')
    june_only = (panda['time_utc'] >= junefirst) &
        (panda['time_utc'] < june_end)
    return panda[june_only]

# takes a pandas dataframe and a by(e.g tweet id) is a list
# also, drops the rows where all elements are missing
def unique(panda, by):
    return drop_duplicates(subset=by).dropna(how="all")
```

Method

The Zip files were read without unpacking them to save memory. Then the question_1 function was used to get the required values and inserted into a pandas data frame. Then juneOnlyHour function was used to exclude none June timestamps. Lastly, the resulting pandas data frame was passed into the unique function. To count the number of tweets, the pandas shape variable was utilised, getting only the row value.

Result

Amount with duplicates — 15,028,231 and without duplicates — 15,021,393

Comment

There was a low number of duplicated tweets, and it was assumed they were generally from Europe, so the count has some noise.

1.2 Solution

Code

```
#floors time_utc to days
def juneOnlyD(panda):
    panda['time_utc'] = panda['time_utc'].dt.floor('D')
    return panda

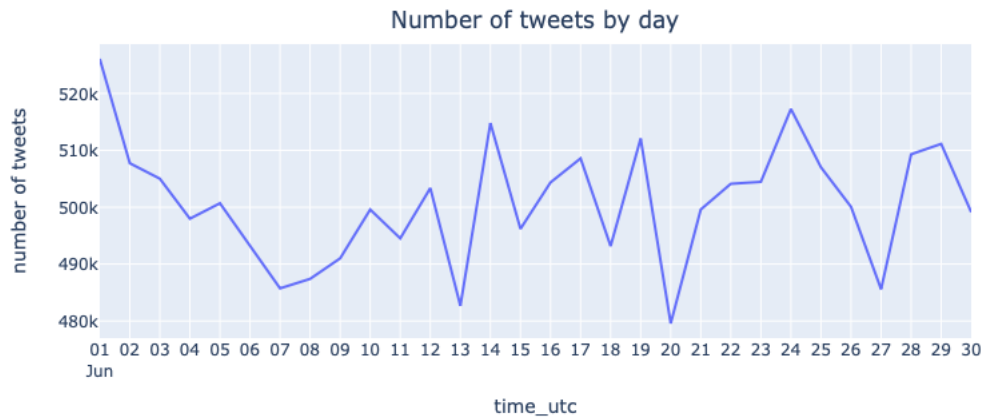
#takes the freq of tweets by day from 1 to 30 June
result12 = juneOnlyD(result11).groupby(by="time_utc").
    count()['tweet_id'].rename('number of tweets')

#plots the time series with plotly.express
plotly.express.line(df1_june12, x=df1_june12.index, y='number of
    tweets')
```

Method

Takes result11 which is a pandas data frame, floors the time to days with juneOnlyD function, groups the data frame by time, counts the freq for each day then insert it into result12 and finally plots the time series using plotly.express.line function

Result



Comment

The graph above indicates a high number of tweets in June. 1st June has the highest number of tweets which could mean that there was more than one event on that day and the evidence is shown in [1].

1.3 Solution

Code

```
import plotly.graph_objects as go
fig = go.Figure()

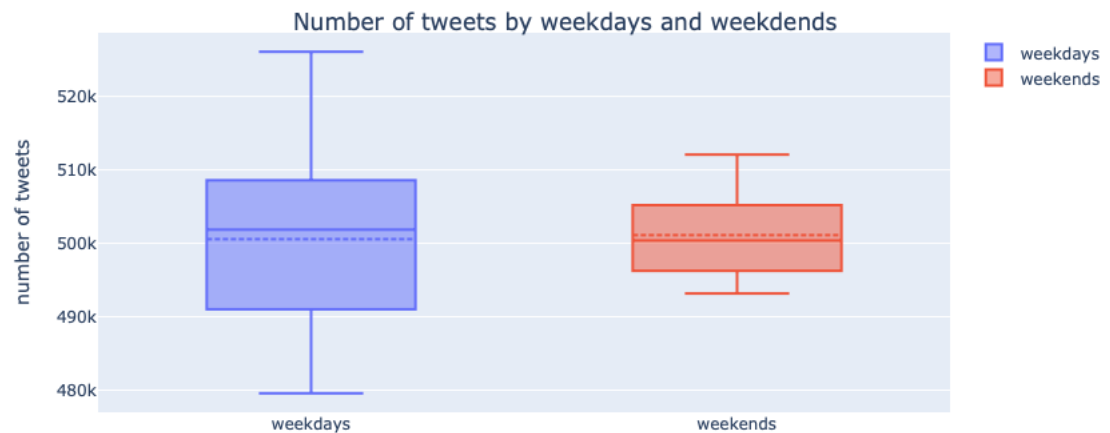
# adds the box plots to the figure object
fig.add_trace(go.Box(y=df1_june13[df1_june13.index.dayofweek<5],
                      name='weekdays', boxmean=True))
fig.add_trace(go.Box(y=df1_june13[df1_june13.index.dayofweek>=5],
                      name='weekends', boxmean=True))

import scipy.stats as stats
stats.mannwhitneyu(result13[result13.index.dayofweek<5],
                   result13[result13.index.dayofweek>=5])
```

Method

Selected both weekdays and weekends by filtering the index(time series) of the result12 series with the dayofweek attribute(0-4 represent weekdays and 5-6 represents weekends). The means are the dotted lines in the box plots. Mann Whitney U Test was used for statistical significance as both groups are not normally distributed.

Result



Avg in weekday $\approx 500.5673k$ and weekends $\approx 501.1141k$

MannwhitneyuResult(statistic=87.0, pvalue=0.49064607864644855)

Comment

The p-value indicates that no statistical significance exists even when considering the threshold to be 5%

1.4 Solution

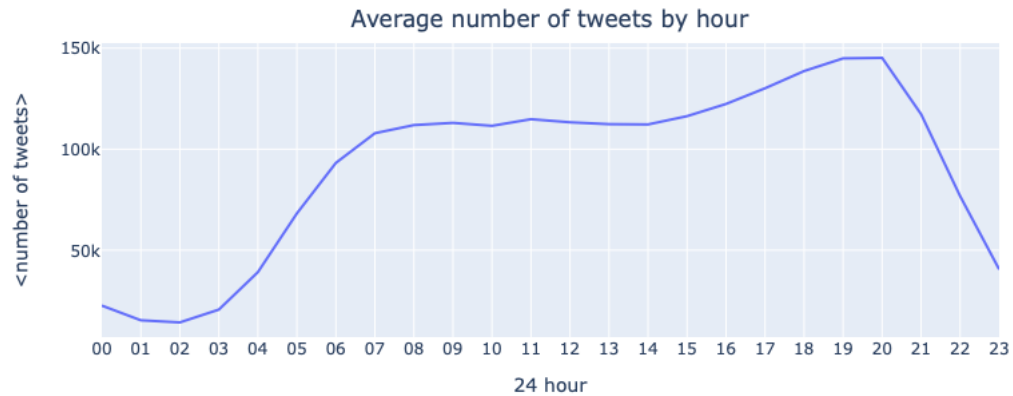
Code

```
result14 = (result1Hour.groupby(by="24  
hour").count()/5)['tweet_id'].rename('<number of tweets>')  
  
plotly.express.line(result14, x=result14.index, y='<number of  
tweets>')
```

Method

After flooring the time to hours, the data(pandas data frame) was grouped by the hours and counted. Finally it was divided by 5 to give the average over the weekdays. The resulting series was plotted using plotly.

Result



Comment

The above plot shows that the number of tweets during working hours is relatively similar, and people tweet more after working hours, peaking at 8 PM, but tweet less after 8 PM.

2 Question Two

2.1 Solution

Code

```
result2 = unique(juneOnlyH(pd.DataFrame(result,
    columns=["tweet_id" ,"user_id","timestamp_ms"])),["tweet_id"])

result21 = result2["user_id"].value_counts().rename("tweet count")

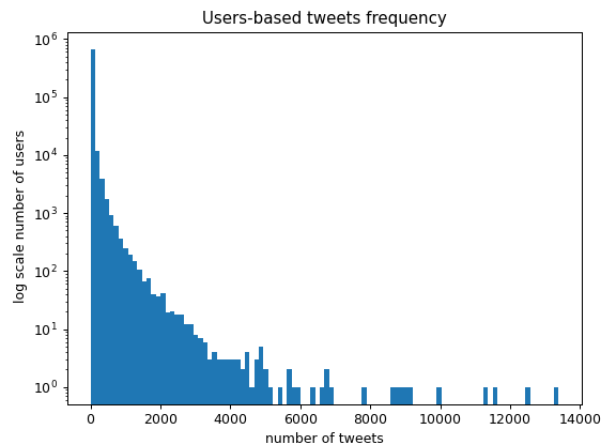
ans21 = result21.value_counts()

plt.hist(ans21.index, weights=ans21, bins=15)
```

Method

After removing duplicates by tweet id from the result, the user id column was selected, and the frequency of users(Also representing tweets) was counted using the pandas value_counts function. The resulting series was counted using the same pandas function. It means that the index of the pandas series is the number of tweets, and the values refer to the frequency of the number of tweets(Also represents users).

Result



2.1.1 Discussion

The plot above shows a positive skew distribution with a long tail. It means that most users tweeted approximately less than 130 times in June. The plot also signifies that the number of users is inversely proportional to the number of tweets made by those users(i.e. more users made fewer tweets).

2.2 Solution


Code

```
result 22 = result21[:5]
```

Method

Because result21 has been calculated, the top 5 users is the slicing from 0 to 5

Result

users	tweets count
uzlaşma af	13376
DailyNews 	12518
Christian Antolic	11619
L'hora catalana	11278
minijob-anzeigen.de	9944

2.2.1 Discussion

The top 5 users are bots. Using the 72 tweets per day guideline recommended by [2] for bot detection, 72×30 is 2160 tweets in June. This means that the top 5 users are clearly bots.

2.3 Solution

Code

```
""" to unpacks the list of str column in result23
data frame to be str"""
result23 = result23.explode("user_mention")

result23["user_mention"].value_counts()[:5]
```

Method

question_23 function expanded on question_1 function, the user mention data was added to the condition and returned. The data from the zip files returned as a lists of objects passed into a data frame, and the general functions were utilised. Because the user mention column was a lists of str, the pandas explode function was utilised on the column to turn it into str by repeating the rows with each value in the user mention list of str. The user mention column was selected, and the pandas function value_counts was utilised to count the frequency of user mentions.

Result

YouTube	15529
Recep Tayyip Erdoğan	6610
SGC RGUKT BASAR	5413
Boris Johnson	5392
Elon Musk	5078

Comment

The top 5 user mentions are 4 celebrities and one prominent company, which is not surprising considering that any negative or positive events about them will most likely be tweeted.

2.4 Solution

Code

```
"""takes a data frame, tweet country code(ISO2) and
the user mention country code(ISO2), returns int"""
def mention_count(panda, twt_cc, userm_cc):
    panda = panda.copy()
    twt_incountry = panda[panda["country_code"]==f"{twt_cc}"]
    users_mentioned = twt_incountry["user_mentions"]
    merge_user_userm = pd.merge(panda, users_mentioned, how="left",
                                left_on="user_name",
                                right_on="user_mentions").dropna()

    only_userm_cc =
        merge_user_userm[merge_user_userm['country_code']==f"{userm_cc}"]
    return only_userm_cc.shape[0]
```

Method

Similar to the question_23 function with the addition of latitude and longitude from the coordinate field. It means that tweets with no coordinate field were not included but were returned as None. The data was passed into a data frame, and the general functions were used. Using reverse_geocoder module, the lats and longs were converted into countries ISO2 code and added to the data frame. The resulting data frame was passed into the mention_count function with two other arguments (e.g. ("UK", "UK"), ("UK", "FR")). The mention count function selects from the data frame based on the second argument and then passed into the twt_incountry variable, and the user mention column is selected from the twt_incountry and passed into users_mentioned. twt_incountry and users_mentioned were merged such that the user column in twt_incountry matched the users_mentioned series. In the resulting data frame, Nan values were removed, the last argument was used to select the required data, and the row value of the data frame was returned.

Result

	Great Britain	France	Italy	Netherlands
Great Britain	20582	154	36	29
France	711	566	26	6
Italy	46	34	9529	6
Netherlands	77	0	4	224

Comment

3 of the countries mentioned Great Britain more than they mentioned each other.

3 Question Three

3.1 Solution

Code

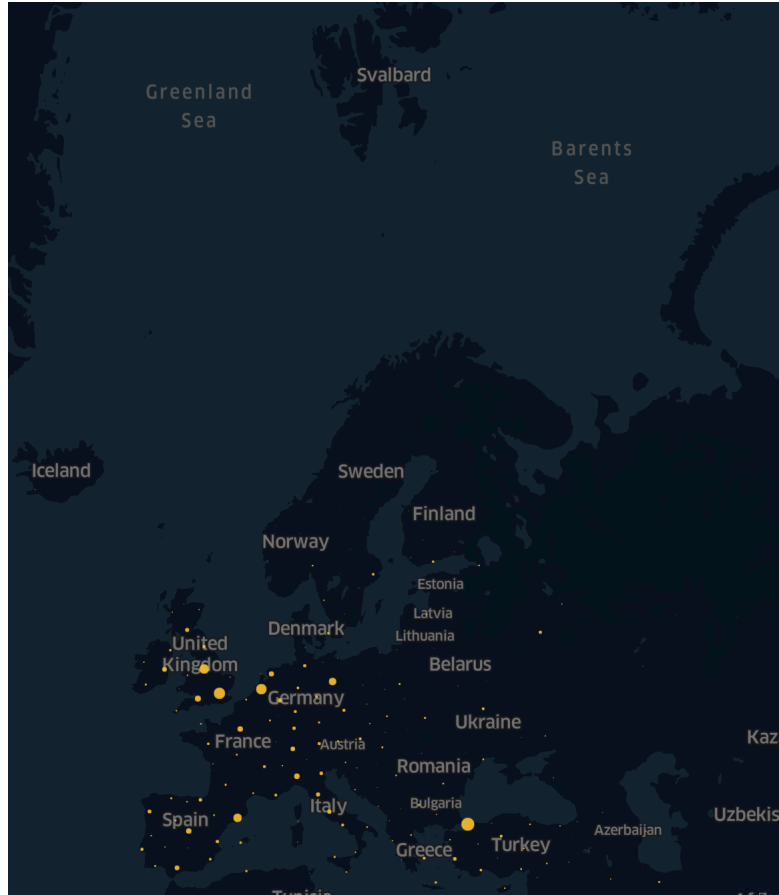
```
from keplergl import KeplerGl
map_1 = KeplerGl(height=600, config=config)
map_1.add_data(data= pd.DataFrame(coord, columns=["lat","lon"]),
               name='points')

map_1.save_to_html(data={'points': pd.DataFrame(coord,
          columns=["lat","lon"])},
                  file_name='privateers.html',config=config, read_only=True)
```

Method

After getting the data and using the general functions, keplergl python was utilised to plot the points because of the volume of points. Due to the points being above 100,000, clustering in keplergl python was used to join data near each other. The cluster geospatial map still shows the general pattern or trends even though countries with low points will have no clusters.

Result



3.2 Solution

Result

Cities, regions, or counties with large populations have a high number of tweets. Looking at England based on the map above, it can be seen that there are more tweets in London, and according to [3] London has the highest population in the UK.

3.3 Solution

Code

```
from shapely.geometry import box
europe_box=box(-24.5, 34.8, 69.1, 81.9)

# see if there are bounding boxes that does not intersect
(question33.geometry.intersects(europe_box)==False).sum()

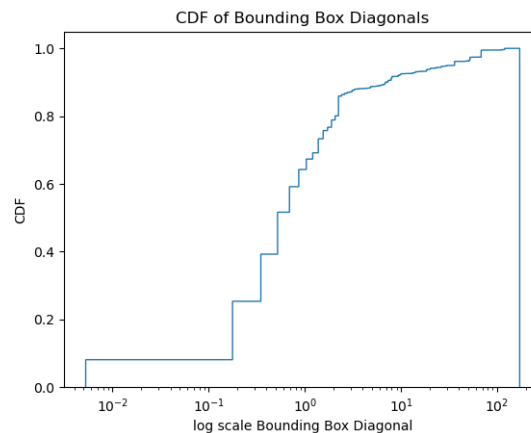
diag = ((question33.area**2 / question33.length**2) +
        question33.length**2).apply(np.sqrt)

plt.hist(diag, bins=1000, cumulative=1, density=1,
         histtype='step')
```

Method

question33 (Geopandas Dataframe) takes the data needed after processing the zip files, and check if there are bounding boxes that does not intersect (Answer is zero). The diagonal is calculated using the area, length, and width. The resulting pandas series is plotted, and the x-axis is log scaled to show the extremely low values.

Result



Comment

Most of the bounding box diagonals have low magnitudes, implying that the areas are also small. It also indicates that the bounding boxes could likely be used to identify an individual's location.

3.4 Solution

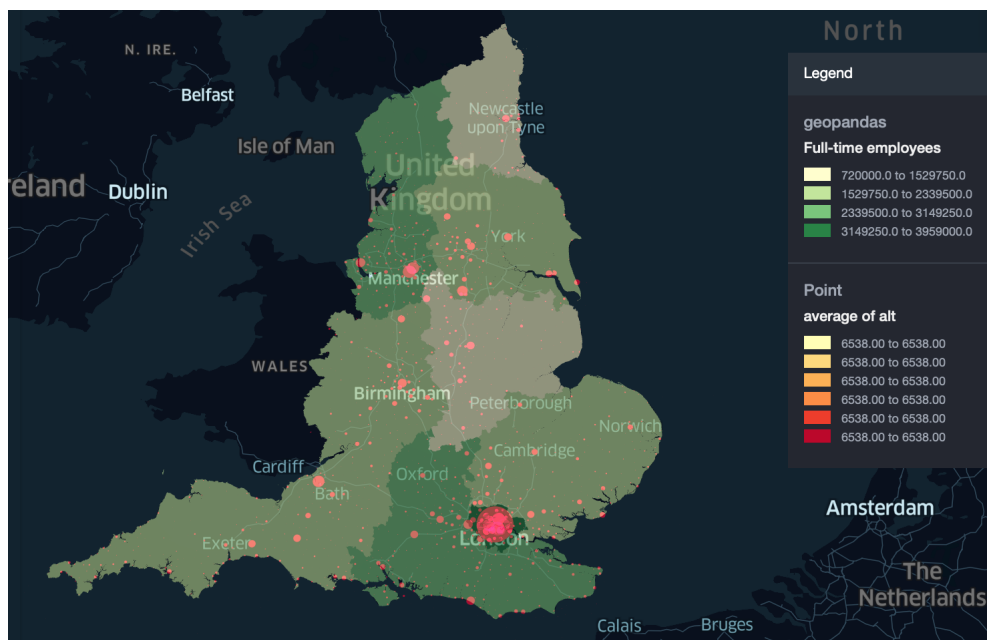
Code

```
from keplergl import KeplerGl
config={}
map_1 = KeplerGl(height=800, config=config)
map_1.add_data(data=employ_reg_gpd, name='geopandas')
map_1.add_data(data=ques34_uk, name='twitter')
```

Method

Because of the difficulty in finding a python package for free reverse geocoding of large data points, KeplerGl was utilised. The tweets were clustered, while a choropleth was utilised for the full employment data.

Result



Comment

The plot above shows that the number of full-time employees correlates with the number of tweets. The above plot also indicates that tweets are very high in London and London has a higher income than the other cities.

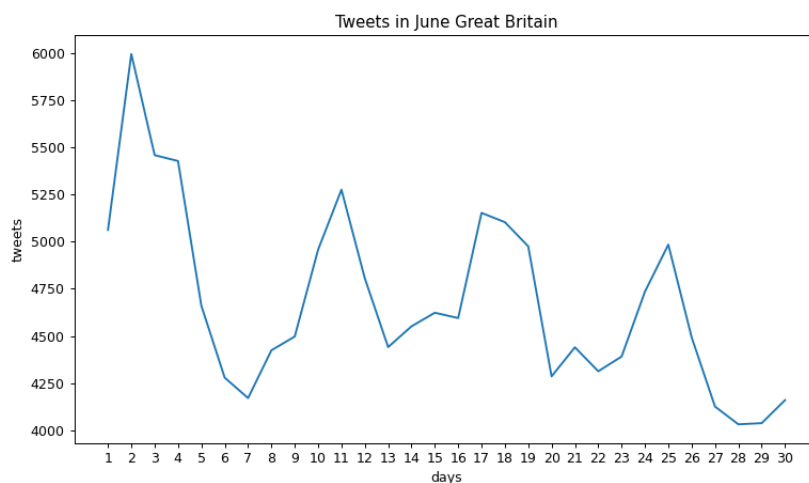
4 Question Four

4.1 Solution

Method

After preprocessing the data, the tweets are plotted similarly to question one, part two. The steepness and depth of the valley are the two criteria used to identify an unusual day. Steepness refers to the day that spikes the most, and deep valley refers to the relative decrease on both sides of the steepest day. It is important to note that coordinates were used and not place meta data.

Result



Above is an example showcasing the steepness and depth valley criteria.

The day is 2nd June in Great Britain, 10th June in Ireland, and 19th June in Spain.

4.2 Solution

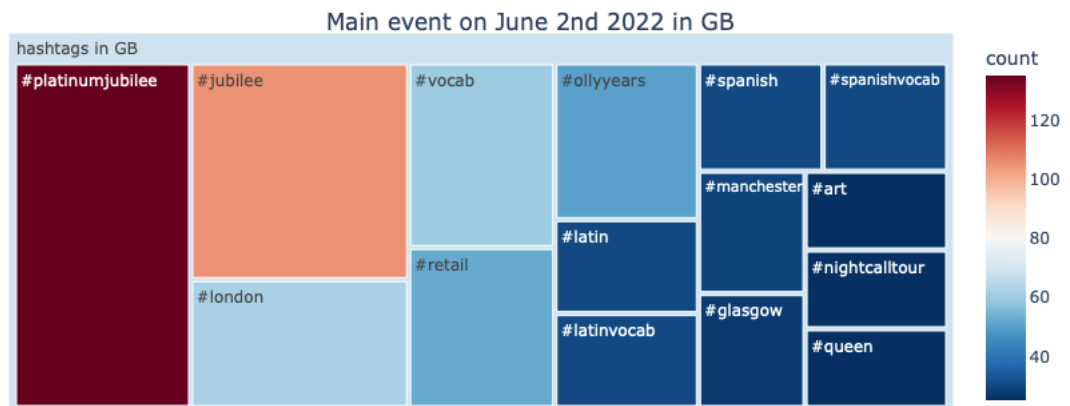
Method

A tree map was plotted using hashtags for part B. Regarding hashtags, it is important to remember that they can represent the topics covered by multiple tweets.

Result



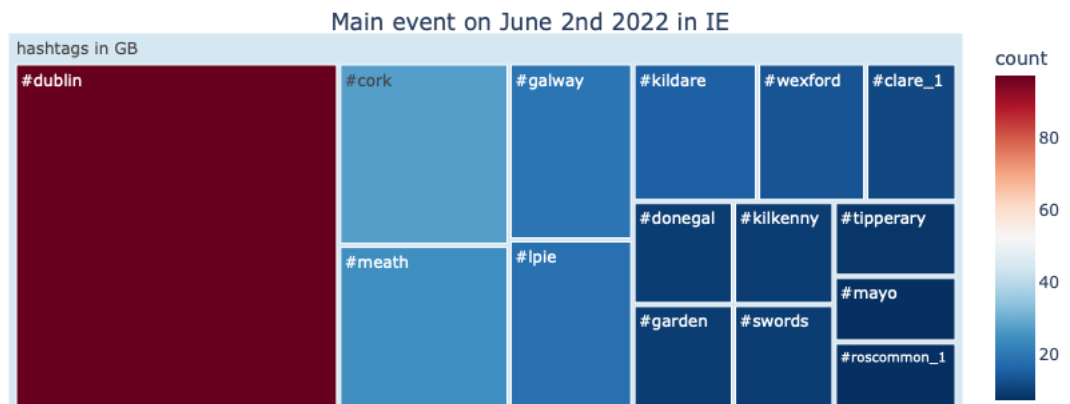
Word cloud of 2nd June in United Kingdom



Tree map of 2nd June in United Kingdom



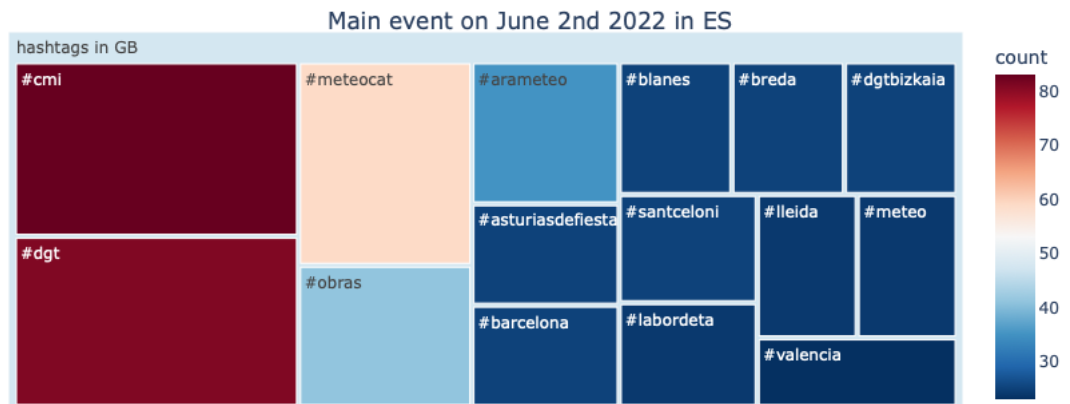
Word cloud of 10th June in Ireland



Tree map of 10th June in Ireland



Wordcloud of 19th June in Spain



Treemap of 19th June in Spain

4.3 question

Result

Platinum jubilee was the trending event on Twitter on the 2nd of June in UK. The increase in tweets on the 2nd could likely be to the extra holiday announced on June[4]. In Ireland majority of the hashtags were about cities or counties. June is a festive season in Spain. In the word cloud of Spain, there was foto(photo) and publicar (publish), mostly about festivals.

5 Question Five

Since the advent of the internet, social media platforms and applications have gained popularity among the general public and are frequently used for entertainment, socialization, and information dissemination. Twitter, a microblogging platform, is one of the prominent social media platforms and also among the few platforms that allow almost free access to its data for public research. As a result, researchers can easily collect data and carry out some tests with the help of this tool. Twitter is becoming increasingly useful in politics for mobilizing voters during elections[5] and in education for communicating with students[6]. Although Twitter data may be a good source of data for researchers, it does have some disadvantages. The data from Twitter are skewed and do not represent the entire population. [7] states that Twitter retweets do not indicate valuable tweets and are typically used as endorsements, so they cannot measure the quality of tweets. Because of the word limit, it makes it difficult to have contextual tweet text for natural language processing. Apart from the drawbacks of Twitter data, there are ethical concerns about using Twitter data, namely informed consent and privacy[8] and anonymity [9]. It may be necessary to inform most Twitter users that their data is used for research and what kind of research it is used for. For research that might require data which a user would consider private, the user does not have the option to opt out. Regarding anonymity, there is a high possibility of a unique user being reidentified since Twitter does not prevent unique user tweets from being fetched.

References

- [1] S. S. , “1 June 2022: History, News, Top Tweets, Social Media & Day Info - UK,” <https://www.wincalendar.com/Calendar-UK/date/1-June-2022>.
- [2] B. Nimmo, “BotSpot: Twelve ways to spot a bot,” *Atlantic Council’s Digital Forensic Research Lab*, vol. 28, 2017.
- [3] “London Population 2022,” <https://worldpopulationreview.com/world-cities/london-population>.
- [4] “Extra Bank Holiday to mark The Queen’s Platinum Jubilee in 2022,” <https://www.gov.uk/government/news/extra-bank-holiday-to-mark-the-queens-platinum-jubilee-in-2022>.
- [5] A. Jungherr, “Twitter use in election campaigns: A systematic literature review,” *Journal of Information Technology & Politics*, vol. 13, no. 1, pp. 72–91, Jan. 2016.
- [6] “Using Twitter for education: Beneficial or simply a waste of time?” *Computers & Education*, vol. 106, pp. 97–118, Mar. 2017.
- [7] “Limitations of Twitter Data. Due to Twitter’s accessible and... — by Kourosh Alizadeh — Towards Data Science,” <https://towardsdatascience.com/limitations-of-twitter-data-94954850cacf>.
- [8] C. M. Rivers and B. L. Lewis, “Ethical research standards in a world of big data,” Aug. 2014.
- [9] W. Ahmed, P. A. Bath, and G. Demartini, “Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges,” in *The Ethics of Online Research*, ser. Advances in Research Ethics and Integrity, K. Woodfield, Ed. Emerald Publishing Limited, Jan. 2017, vol. 2, pp. 79–107.