0.1 Question 1d

There are many ways we could choose to read tweets. Why might someone be interested in doing data analysis on tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of tweets might be interesting or useful for them. Answer in 2-3 sentences.

We might be interested in doing data analysis on tweets to understand the person tweeting them better. Someone or some institution might be interested in doing data analysis on tweets probably for marketing purposes and showing relevant targeted ads toward them. By performing the data analysis, the institutions can better understand what their user likes or doesn't like (such as by finding the polarity of certain things they say) and showing things that are relevant and likely to be clicked on by the user.

0.1.1 Question 2e

What might we want to investigate further? Write a few sentences below.

We might want to investigate when the tweets were posted in comparison to the type of source used. For example, I see twitter for BlackBerry so the tweets sent by that must have been much older than those with the iPhone. We might also want to look at how long Cristiano has been on twitter as he has much more variation in his sources compared to the others.

0.1.2 Question 2f

We just looked at the top 5 most commonly used devices for each user. However, we used the number of tweets as a measure, when it might be better to compare these distributions by comparing *proportions* of tweets. Why might proportions of tweets be better measures than numbers of tweets?

Proportions might be better because they might account for differences that happen due to either using the same device but also because one person just might on average not post as much as someone else. For example, celebrities probably have social media people who will make more posts for them whereas smaller producers may not post as much.

0.1.3 Question 3b

Compare Cristiano's distribution with those of AOC and Elon Musk. In particular, compare the distributions before and after Hour 6. What differences did you notice? What might be a possible cause of that? Do the data plotted above seem reasonable?

Cristiano tends to tweet a lot more during the hours of 10 to 15 and in this time frame tweets more than either AOC or elonmusk. This is probably due to the fact that Cristiano is likely located in a different hemisphere than Elon and AOC and then the data does seem reasonable.

0.1.4 Question 4a

Please score the sentiment of one of the following words, using your own personal interpretation. No code is required for this question!

- police
- order
- Democrat
- Republican
- gun
- dog
- technology
- TikTok
- security
- face-mask
- science
- climate change
- vaccine

What score did you give it and why? Can you think of a situation in which this word would carry the opposite sentiment to the one you've just assigned?

vaccine 0, I adamantly agree with vaccine policies but this word feels like it has a lot of unwanted discourse around it that I am simple too unbothered to be involved in right now. I would usually feel pretty positively about this word because vaccines usually bring hope.

0.1.5 Question 4g

When grouping by mentions and aggregating the polarity of the tweets, what aggregation function should we use? What might be one drawback of using the mean?

We might use the mean but then depending on how much data there is for each mention, outliers could really skew the mean and one word that doesn't really represent the polarity of the mentioned person makes it inaccurate.

0.1.6 Question 5a

Use this space to put your EDA code.

```
In [131]: for col in tweets['Cristiano'].columns:
              print(col)
```

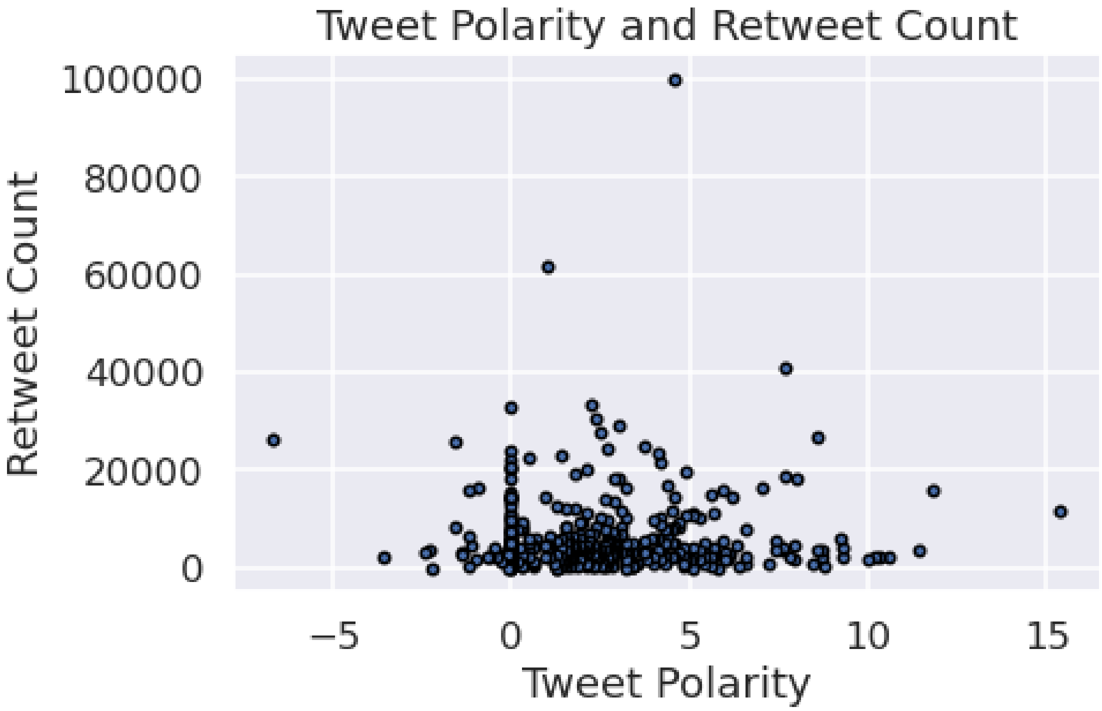
```
created_at
id_str
full_text
truncated
display_text_range
entities
extended_entities
source
in_reply_to_status_id
in_reply_to_status_id_str
in_reply_to_user_id
in_reply_to_user_id_str
in_reply_to_screen_name
user
geo
coordinates
place
contributors
is_quote_status
retweet_count
favorite_count
favorited
retweeted
possibly_sensitive
lang
quoted_status_id
quoted_status_id_str
quoted_status_permalink
quoted_status
retweeted_status
device
hour
converted_time
converted_hour
clean_text
polarity
```

```
In [132]: mentions_re = r'#(\w*)'
```

```
#re.findall(punt,tweets ["Cristiano"].iloc[2]["full_text"])

def extract_hashtags(full_texts):
    hashtags = pd.DataFrame({"hashtags" :  full_texts.str.lower().str.findall(mentions_re)})
    hashtags = hashtags.set_index(full_texts.index)
    hashtags['hashtags'] = hashtags['hashtags'].apply(lambda x: ','.join(map(str, x)))
    return hashtags[["hashtags"]]

mentions = {handle: extract_hashtags(df["full_text"]) for handle, df in tweets.items()}

def hashtag_polarity(df, hashtag_df):
    series = df.join(hashtag_df)[['hashtags', 'polarity', 'retweet_count']]
    series = series.set_index('hashtags')
    series = series.groupby(series.index).agg(np.mean)
    return series


cristiano_mention_polarity = hashtag_polarity(tweets["Cristiano"],mentions["Cristiano"]).sort_
display(cristiano_mention_polarity)

plt.scatter(cristiano_mention_polarity['polarity'], cristiano_mention_polarity['retweet_count
plt.title("Tweet Polarity and Retweet Count");
plt.ylabel("Retweet Count");
plt.xlabel("Tweet Polarity");
```

```
                                                        polarity  \
hashtags
goldenfoot2020                                              15.4
globesoccer                                                 11.8
happybirthdaycr7                                            11.4
als                                                         10.6
bethe1donor,abbott                                          10.3
…                                                          …
ocnn                                                        -2.2
cr7,mercurial,nikefootball                                  -2.3
cr7crunchfitness,cr7,gym                                    -2.4
keepaclearhead,clearmen,nodoubts,nodandruff,menshampoo      -3.6
prayers4paris                                               -6.7

                                                        retweet_count
hashtags
goldenfoot2020                                                11769.0
globesoccer                                                   15808.0
happybirthdaycr7                                               3865.0
als                                                            2328.0
bethe1donor,abbott                                             2023.0
…                                                               …
ocnn                                                             51.0
cr7,mercurial,nikefootball                                     3731.0
cr7crunchfitness,cr7,gym                                       3348.0
keepaclearhead,clearmen,nodoubts,nodandruff,menshampoo         2243.0
prayers4paris                                                 26490.0
```

```
[426 rows x 2 columns]
```



0.1.7 Question 5b

Use this space to put your EDA description.

I was interested in learning about the polarity of each hashtag and how that compared to the count of retweets for Cristiano in the tweets dataframe. Are certain hashtags relevant to a higher or lower polarity and do tweets with more or less polarity have more retweets? First I used regex to pull the hashtags used in the full_text column and put this into its own column. Then I compared it to the polarity of the entire full text it was located in. This would provide a polarity for the hashtag used in context. As we can see with the results of Cristiano's hashtags, it looks like goldenfoot2020 has the highest polarity which makes sense because it is the highest football award. We can then see that the lowest polarity is "prayers4paris" which is also relevant because that was a traumatic worldly event. Then, I used matplotlib to plot the polarity compared to the retweet count and saw that the majority of the tweets with hashtags were actually within thte 0-10 range and only a few were outside of this range. I also found that there was not a significantly higher retweet count for extreme polarity counts although, there was one outlier where the tweet polarity was about 5 and the retweet count was an astounding 100k!