
0.0.1 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 that 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.

X/Twitter/Big Corporations are in their interest to analyze the tweets as they get the most benefit for it.

It allows them to personalize the 'for you' page more efficiently for its consumers, the 'for you' page is something that has completely taken over social media and is what divides the successful social media from the unsuccessful.

Furthermore, the data that they collect can be sold off to other companies in order to turn some revenue.

0.0.2 Question 2e

Given the plot above, what might we want to investigate during EDA? Name some possible questions you may have about the dataset in light of the information shown in the plot.

Why does Cristiano use so many different devices? Why do the twitter counts change so much? Is it useful to have these extra sources if only one person uses them?

0.0.3 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 (i.e., what percentage of all tweets for a user were published from each device). Why might the proportions of tweets be better measures than the number of tweets?

Because the amount of tweets a twitter user tweets varies from person to person. Some people may tweet a few hundred, some a few thousand. If we are trying to investigate devices and only devices, it would be more important to figure out their time spread out between each device not the count from each device per user.

0.0.4 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?

Hint: If you are not familiar with who Cristiano, AOC, and Elon Musk are, it may be helpful to Google information about these people, their occupations, and where they live.

AOC and Elon Musk are signficantly more active before hour 6 compared to Cristiano. This might be due to multiple differences.

AOC and Elon Musk are people that would benefit significantly more from tweeting. Garnering attention or support through social media is more impactful in their lives whereas Cristiano might be practicing for Soccer or doing things that are more important. Another potential explanation would be the fact that they live in different timezones so while Cristiano would be sleeping between hour 24 and hour 6, Elon Musk and AOC sleep between hour 8 and hour 14 roughly.

0.0.5 Question 4a

Using your own personal interpretation, please score the sentiment of one of the following words using the VADER scale (-4 means the word is extremely negative. +4 means the word is extremely positive). No code is required for this question!

- order
- dog
- cat
- technology
- TikTok
- security
- science
- climate change

What score did you give it and why? Can you describe a situation where this word would carry the opposite sentiment to the one you've just assigned? If not, explain why.

'TikTok' = -2.

I gave it a -2 because even the people that use the app are aware that the app is not a healthy thing to use. It has a lot of recognition, and a large amount of the population uses it on a regular basis. It has changed the global landscape in terms of trends and other such things. But it is also noted that it is not something that is healthy. I personally only use it since that is how my girlfriend sends me funny videos and such, and maybe she would carry the opposite sentiment as to her it is what she uses to get through some parts of her day. Her daily trip to work, or maybe even sometimes on break, she would use it to pass the time and not be bored.

0.0.6 Question 4g

In `q4f` above, we aggregated the polarity of the tweets by computing the mean sentiment score of tweets mentioning each user. What are some drawbacks of the decision to use the mean as an aggregation function? What other aggregation function(s) might be more appropriate than the mean?

Using mean as the average function could cause the discovered average to be skewed by outliers. Using median instead of mean would get rid of this worry especially if we knew that we are working with outlier data.

0.0.7 Question 5a

Use this space to put your EDA code.

In [83]:
```
df = pd.concat([tweets['AOC'], tweets['Cristiano'], tweets['elonmusk']],axis=0)
sentiment = df[['device', 'polarity']]
EDA = sentiment.groupby('device').agg(['min','mean','max','count'])
EDA
```

Out[83]:

	polarity			
	min	mean	max	count
device				
10 - Sport Through Your Lens	0.0	3.553846	16.3	52
Crowdfire Inc.	0.0	0.250000	0.5	2
Facebook	0.0	0.000000	0.0	2
Google	-1.2	-0.300000	0.0	4
Instagram	0.0	2.820000	9.0	40
MobioINsider.com	0.0	1.667361	12.1	144
Players' Post	-1.2	2.236364	8.6	11
Twitter Media Studio	-3.6	2.198765	15.3	81
Twitter Web App	-3.2	1.709231	13.1	65
Twitter Web Client	-4.4	2.301773	17.0	959
Twitter for Android	-6.7	2.840741	12.9	108
Twitter for BlackBerry®	3.9	3.900000	3.9	1
Twitter for Websites	5.2	5.550000	5.9	2
Twitter for iPad	-4.0	2.956522	10.2	92
Twitter for iPhone	-15.7	0.928440	25.7	7630
Viva Ronaldo	-1.2	-0.704000	2.8	25
WhoSay	-2.9	3.730905	13.4	453
iOS	0.0	0.000000	0.0	1
swonkie	-1.2	1.433333	6.7	12

0.0.8 Question 5b

Use this space to put your EDA description.

I grouped the tweets by devices, and I found a few interesting things. I found that the people at Viva Ronaldo on average, send negative tweets, and Twitter, for BlackBerry® on average, send the most positive tweets. However, Twitter for iPhone has sent the most mean tweet by far with a minimum of '-15.7'. Conversely, they have also sent the most positive tweets at '25.7'. Overall, it is mostly as expected since the lower count platforms tend to have farther from 0 means since there are less counts to center round 0 in the long run.