
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 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.

Data analysis can be useful for social media companies to gather data on how users can stay active on their applications. Data analysis of tweets might be useful to learn what kinds of things people tweet about to develop the timeline algorithm so that viewers who are interested can see more of those kinds of tweets. Another reason why data analysis might be useful is to learn about the times of day that people are most active and which tweets are interacted with the most to see what's most popular.

0.0.2 Question 2e

What might we want to investigate further based on the plot in 2d above? Write a few sentences below.

As shown in the plot, AOC and elonmusk have been tweeting mainly on Twitter for iPhone. However, Cristiano's devices have been pretty spread out. Similarly, he has been using Twitter for iPhone the must, but not nearly as much. He has also used Twitter Web Client and WhoSay almost as significantly. We might want to investigate further to see if the reason Cristiano uses different devices is because he is native to a different country and has different types of technology there.

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. Why might proportions of tweets be better measures than numbers of tweets?

Proportions of tweets are better measures than numbers of tweets because proportions demonstrate a comparison of the number of tweets on one device compared to other devices. Numbers of tweets is inaccurate because it assumes all users tweet the same number of times, which is clearly not the case.

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?

Cristiano's distribution is shaped more like a bell curve than AOC and Elon Musk. This means that he tweets more regularly within a specific time frame as compared to the other two. Some of the data plotted make sense. However, there are clearly several tweets made at hours where people are usually asleep, like between 3-5am. Otherwise, the data seem reasonable.

0.0.5 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?

police: -2 – When I first think of police I think of the unjust things they have done in the past, but there is a possibility they can be referred to when talking about how society is protected.

0.0.6 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?

The aggregation function we should use is mean. This is because we want the average sentiment of the tweets mentioning each user. A drawback of using the mean is there could be outliers that would skew the mean in the opposite direction – the mean is very sensitive to outliers. The median however is not that sensitive outliers, so that could be a better aggregate function to use. It would also capture the center of our spread more accurately, therefore giving a more accurate polarity for the users.

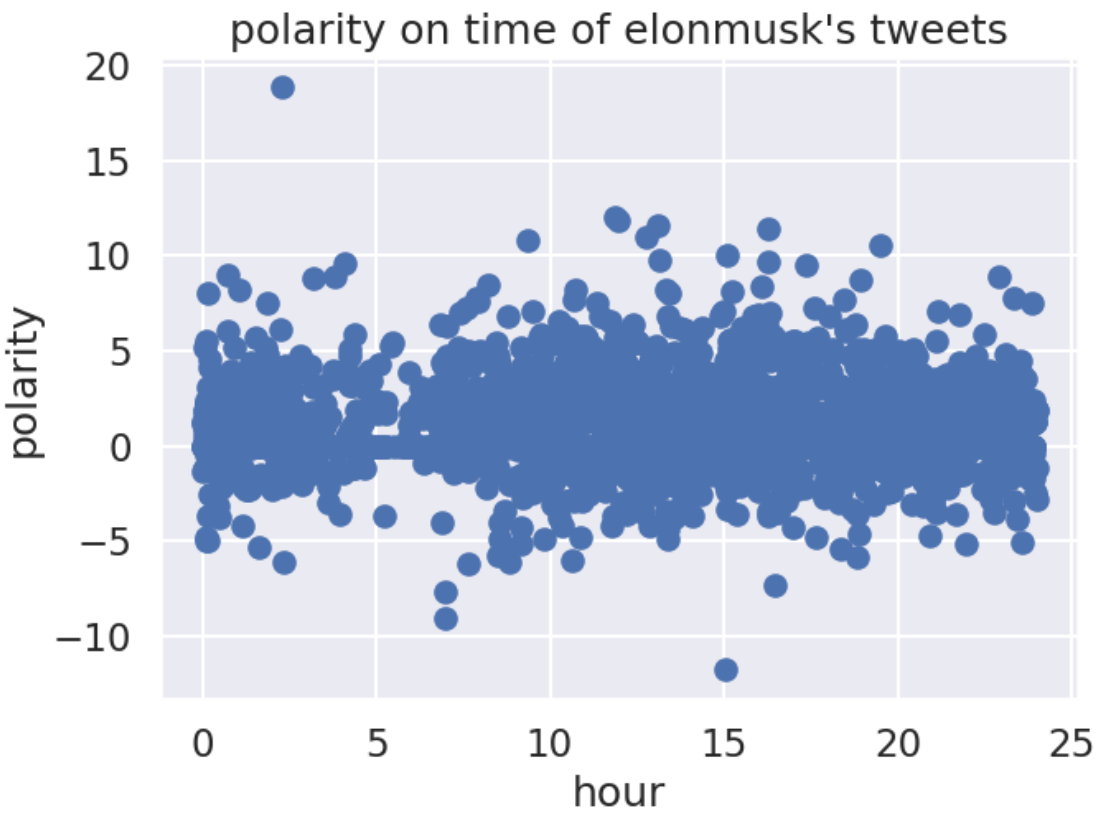
0.0.7 Question 5a

Use this space to put your EDA code.

```
In [44]: em = add_polarity(tweets["elonmusk"], to_tidy_format(tweets["elonmusk"]))
         em = convert_timezone(em, "America/Los_Angeles")
         em = add_hour(em, 'converted_time', 'hour').sort_values('hour').loc[:, ('hour', 'polarity')]
         pivot_em = em.pivot_table(index='hour', values='polarity', aggfunc='mean')

         plt.scatter(em['hour'], em['polarity'])
         plt.title('polarity on time of elonmusk\'s tweets')
         plt.xlabel('hour')
         plt.ylabel('polarity')
         ;
```

```
Out[44]: ''
```



0.0.8 Question 5b

Use this space to put your EDA description.

I used the tweets data to analyze the polarity of elonmusk's tweets at all times of the day. Based on the plot, we can see that the general polarity is roughly at 0, maybe slightly above 0. Roughly between hour 10 and 15, the range of polairty is much smaller, but towards the beginning and end of the day, the range widens. There are also a few outliers of tweets that an extreme polarity above 15 or below -5.