Data 100/200 Homework 4 Written

Jackie Hu

TOTAL POINTS

8.5 / 9

QUESTION 1

1 Question 2e 2 / 2

✓ **+ 2 pts** **Proposes an interesting problem based on the data generated (e.g. why Christiano uses so many devices, exploring the few AOC tweets from Twitter media studio)**

+ 1 pts Proposes a problem not based on the data

+ 0 pts Blank/incorrect

QUESTION 2

2 Question 2f 1 / 1

✓ **+ 1 pts** ****Correct****

Some users might tend to tweet more often than the others; need to have a consistent scale.

+ 0 pts **Incorrect/Blank**

QUESTION 3

3 Question 3b 1 / 1

✓ **+ 1 pts** **Identification of difference, cause, and whether or not the data plotted seem reasonable**

+ 0.5 pts One or more of difference, cause, or identification of whether or not the data seem reasonable missing

+ 0 pts Incorrect/Blank

QUESTION 4

4 Question 4f 0.5 / 1

+ 1 pts Median; explains how outliers affect mean

+ 0.5 pts Median; no explanation of outliers

✓ **+ 0.5 pts** **Mean, sum, mode, min/max, or some other statistic**

+ 0 pts Blank or completely incorrect

QUESTION 5

5 Question 5a 2 / 2

✓ **+ 2 pts** **Produces a mostly informative plot or output that addresses the question posed in the student's description and uses at least one of the following methods: groupby, agg, merge, pivot_table, str, apply**

+ 1 pts Attempts to produce a plot or manipulate data but the output is unrelated to the proposed question, doesn't utilize at least one of the listed methods, or is difficult to interpret due to the way it is displayed (eg overplotting)

+ 0 pts No attempt

QUESTION 6

6 Question 5b 2 / 2

✓ **+ 2 pts** **Describes the analysis question and procedure comprehensively and summarizes results correctly**

+ 1 pts Attempts to describe analysis and results but description of results is incorrect or analysis of results is disconnected from the student’s original question

+ 0 pts No attempt

What might we want to investigate further? Write a few sentences below and be prepared to discuss in next week's small group meeting.

- The distribution seems polarize, AOC and Elon Musk has a large number of Iphone. Maybe there's information not show in the graph because the unproportional count, since the barplot only counts of number not based on individual counts.

0.0.1 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?

The base value for each user's posts count is different, so only look at the number of tweets can be arbitrary to their tweating frequenct. While using proportion is more solid to compare device uses based on individual's tweeting frequency.

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?

- The general posting trend is different between Cristiano and AOC, Elon Musk; while Christiano's numer of tweets starting to increase around 6, but AOC and Elon Musk's number of tweets does not increase untill 11.
- It might because they are from different timezones. So the data plotted is reasonable.

0.0.2 Question 4f

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

- we can use mean(), or sum() or len.
- it's talking into account the number of retweets, so if the number of repost is high, the mean polarity score will be smoothed out despite the individual score.

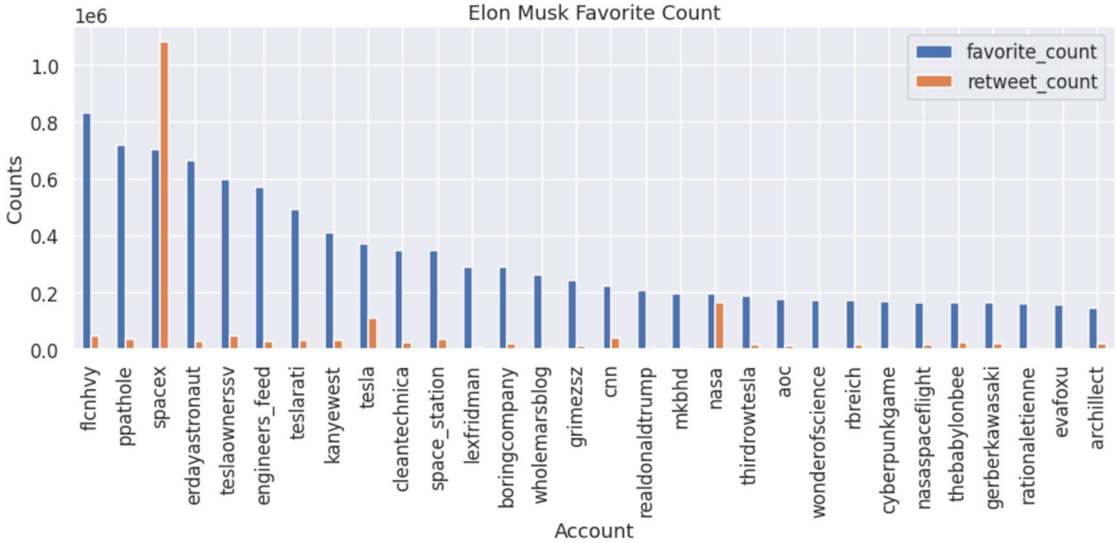
0.0.3 Question 5a

Use this space to put your EDA code.

```
In [53]: # perform your text analysis here
         em = tweets['elonmusk']
         em.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3239 entries, 1357991946082418690 to 1242881125049085956
Data columns (total 36 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   created_at                  3239 non-null   datetime64[ns, UTC]
 1   id_str                      3239 non-null   int64
 2   full_text                   3239 non-null   object
 3   truncated                   3239 non-null   bool
 4   display_text_range          3239 non-null   object
 5   entities                    3239 non-null   object
 6   extended_entities           248 non-null    object
 7   source                      3239 non-null   object
 8   in_reply_to_status_id       2643 non-null   float64
 9   in_reply_to_status_id_str   2643 non-null   float64
 10  in_reply_to_user_id         2643 non-null   float64
 11  in_reply_to_user_id_str     2643 non-null   float64
 12  in_reply_to_screen_name     2643 non-null   object
 13  user                        3239 non-null   object
 14  geo                         0 non-null      float64
 15  coordinates                 0 non-null      float64
 16  place                       0 non-null      float64
 17  contributors                0 non-null      float64
 18  is_quote_status             3239 non-null   bool
 19  retweet_count               3239 non-null   int64
 20  favorite_count              3239 non-null   int64
 21  favorited                   3239 non-null   bool
 22  retweeted                   3239 non-null   bool
 23  possibly_sensitive          458 non-null    float64
 24  lang                        3239 non-null   object
 25  retweeted_status            213 non-null    object
 26  quoted_status_id            74 non-null     float64
 27  quoted_status_id_str        74 non-null     float64
 28  quoted_status_permalink     74 non-null     object
 29  quoted_status               67 non-null     object
 30  device                      3239 non-null   object
 31  hour                        3239 non-null   float64
 32  converted_time              3239 non-null   datetime64[ns, America/Los_Angeles]
 33  converted_hour              3239 non-null   float64
 34  clean_text                  3239 non-null   object
 35  polarity                    3239 non-null   float64
dtypes: bool(4), datetime64[ns, America/Los_Angeles](1), datetime64[ns, UTC](1), float64(14), int64(3),
memory usage: 847.7+ KB
```

```
In [55]: df_ori = em.merge(mentions['elonmusk'], how= 'left', on= 'id')
         df1 = df_ori.groupby('mentions').sum()[['retweet_count','favorite_count']]
         fav_count = df1[['favorite_count','retweet_count']].sort_values(by= 'favorite_count',ascending
         make_bar_plot(fav_count, title='Elon Musk Favorite Count', xlabel= 'Account', ylabel= 'Counts'
```



```
In [62]: df2 = df_ori.groupby('mentions').mean()[['retweet_count','favorite_count','polarity']]
         polarity_top = df2['polarity'].sort_values(ascending=False).head(10).to_frame()
         polarity_top
```

```
Out[62]:                  polarity
         mentions
         viktaur27            11.9
         picot_john           11.4
         vm_one1               9.8
         arvnp                 9.5
         suvitruf              7.3
         businessinsider       7.3
         tegmark               7.1
         adlanbogatyryov       7.1
         hamoon__              7.0
         isaaclatterell        7.0
```

```
In [64]: polarity_buttom = df2['polarity'].sort_values(ascending=True).head(10).to_frame()
         polarity_buttom
```

```
Out[64]:                  polarity
         mentions
         naval                -6.1
         robotbeat            -5.9
         l_vaux               -4.9
         sjvtesla             -4.9
         timothybuffett       -4.8
         mygrindelwald        -4.3
         adamdraper           -3.8
         tomdestella          -3.7
         john_gardi           -3.6
         modernnotoriety      -3.6
```

0.0.4 Question 5b

Use this space to pur your EDA description.

- what were you looking for?
  - what are some of the accounts Elon Musk tends to interact the most; are there any trends in the account he likes and reposts?
- What did you find?
  - That Elon Musk reposts lots of tweets from his company, while the tweets he likes are mostly from personal account.
- How did you go about answering your question?
  - I start by creating the corresponding dataframes and plot bar chart to see the trends. Moreover, I create 2 table to see the polarity within the people Elon Musk interact.