# Notebook

February 13, 2021

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

Since Cristiano's top 5 most commonly used devices is more spread out across different devices, we can investigate each device's usage over the time of day using a line plot. Further, we can also do the same for AOC, but focus on investigating the relationship between AOC's rare usage of Twitter Media Studio device and the time of day. Similarly, for elonmusk, we can investigate the relationship between elonmusk's rare usage of Twitter Web App device and the time of day.

### 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?

Because we are comparing among different users, the proportions of tweets might be a better measure than the numbers of tweets because then the measure of tweets for each device would be relative to the total tweets, giving us a basis for comparison and removing any biases that arise from different number of total tweets. To elaborate, if the total tweets of user x is significantly more than that of user y and a certain count of device category for user x is also more than that of user y, we cannot conclude that user x uses this certain device more than user y because user x has a higher total count in the first place. Thus, the number of tweets is not very indicative when comparing across users since it does not provide the distribution of each device that they use relative to the total number of tweets.

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?

Elon Musk's number of tweets is generally evenly distributed around 80-100 tweets from hours 0 to 8 and hours 18 to 24, but we see a dramatic decrease in the number of tweets from hours 8 to 12.

AOC's number of tweets continues to decrease and stay around zero from hours 2 to 13, but we see a dramatic increase from hours 13 and onwards in comparison. The peak of AOC's plot is around 160 tweet counts at hour 18.

Cristiano's number of tweets is around zero from hours 0 to 6 but increased significantly from hours 6 to 17 and decreased from hours 17 to 24. The peak of Cristiano's plot is around 230 tweet counts at hour 17.

To compare Cristiano's distribution with those of AOC and Elon Musk, around hour 6, Cristiano's number of tweets increased dramatically whereas AOC's number decreased before hour 6 and continued to stay around 0 tweets after hour 6 while Elon Musk's number of tweets stayed around the same before and after hour 6. Possible causes of this difference in distribution are varying working schedules and timezones (time difference) of each user. Another observation from the distribution plot is that AOC's and Cristiano's peak number of tweets are both around the same hour (hour 17 to 18).

The data plotted above seems reasonable for AOC and Cristiano since there are times when they do not have any posts, which accounts for their sleep schedule and other down times. However, the plot does not seem reasonable for Elon Musk since he is constantly posting and has at least one tweet for every hour of the day.

### 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?

When grouping by mentions and aggregating the polarity of the tweets, the aggregation function that we should use is the median, which is not affected by outliers. Another method is to still use the mean aggregate function, but this time, use a standard deviation cut-off to first filter out any results +/-3 stds, then compute the mean.

Some drawbacks of using the mean is that it is heavily affected by outliers, which means that these extreme values would skew the polarity of the tweets and no longer be representative of the sentiment of the tweets. For example, if the polarity of the tweets are [99999, -12, -8, -3, -1], then the mean of the polarity would be skewed by the 99999. Another drawback is that the mean doesn't allow us to group the tweets into big bins without further analysis such as rouding polarities and complex bin classification, so we will have a lot of smaller bins that are difficult to interpret.

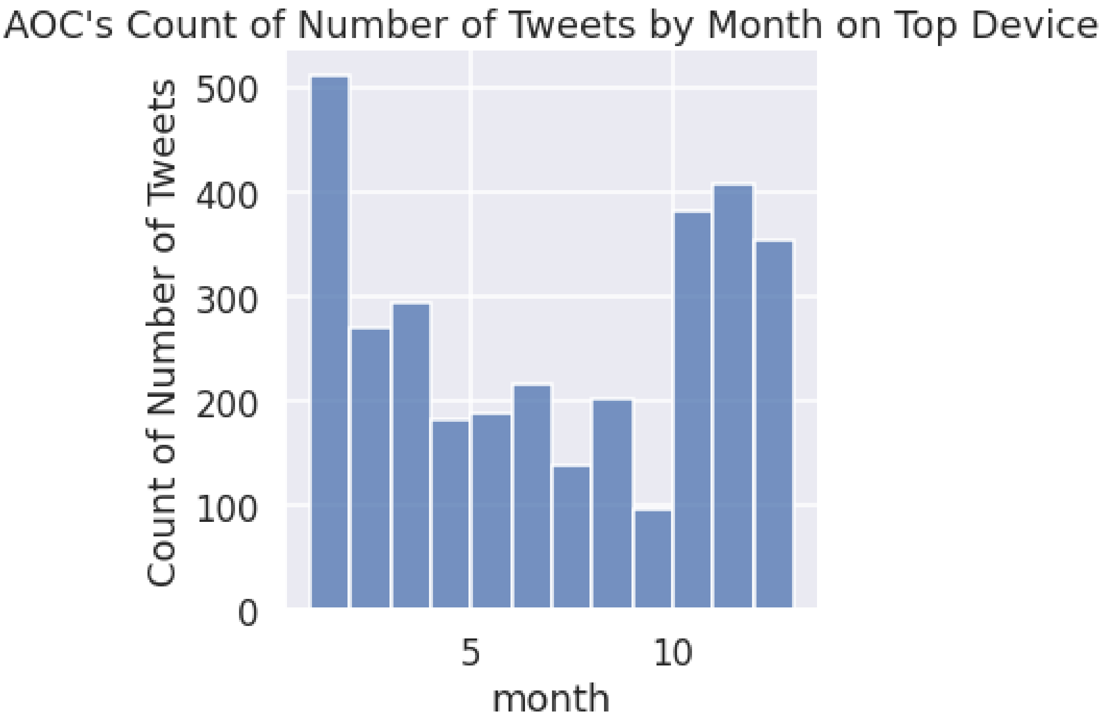
### 0.0.3 Question 5a

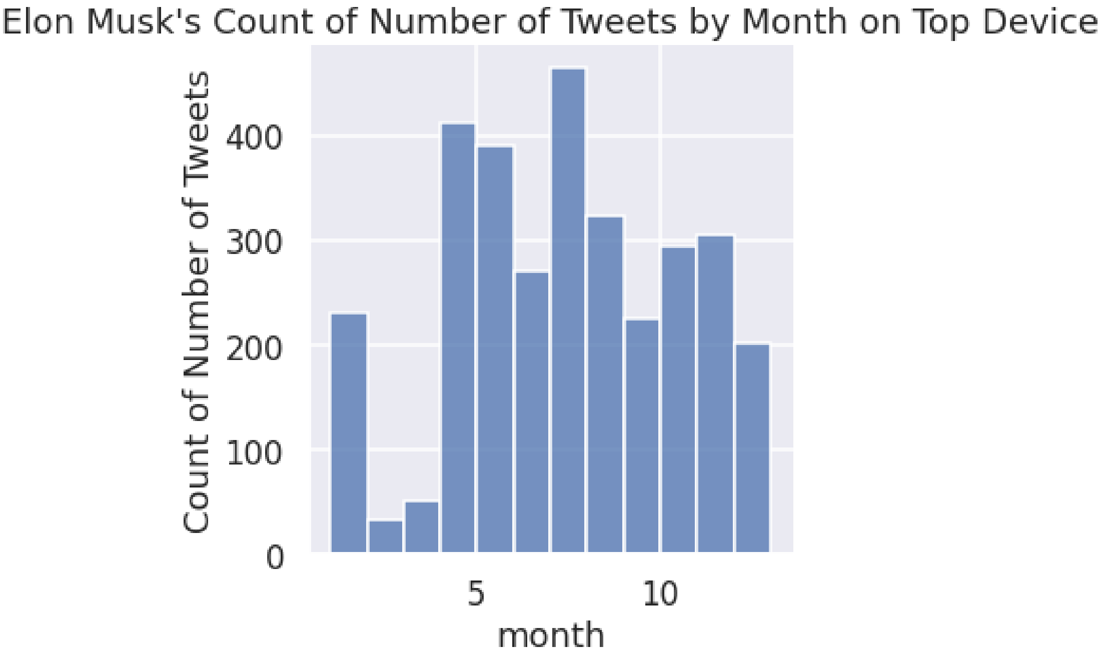
Use this space to put your EDA code.

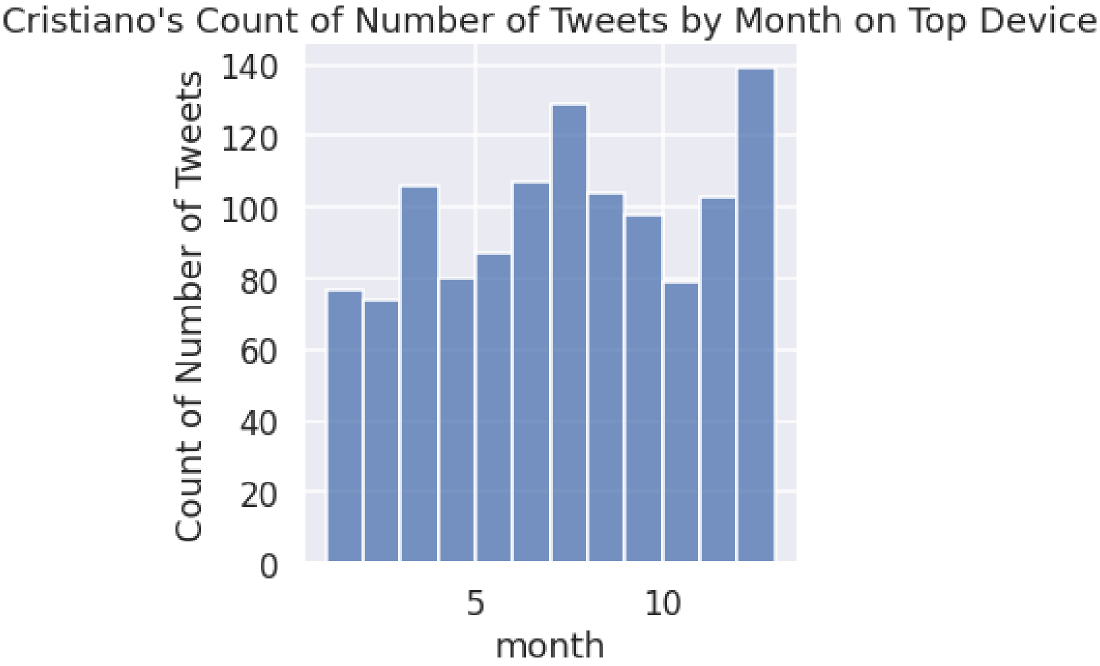
```
[99]: # perform your text analysis here
tweets["AOC"]["month"] = (tweets["AOC"]["converted_time"].dt.month)
topAOC = tweets["AOC"].groupby("device").agg("count").sort_values("created_at",␣
 ↪ascending=False).index[0]
ax_1 = sns.displot(tweets["AOC"][tweets["AOC"]["device"] == topAOC]["month"],␣
 ↪bins = range(1, 14));
ax_1.set(xlabel='month', ylabel='Count of Number of Tweets', title = "AOC's␣
 ↪Count of Number of Tweets by Month on Top Device");

tweets["elonmusk"]["month"] = tweets["elonmusk"]["converted_time"].dt.month
topelonmusk = tweets["elonmusk"].groupby("device").agg("count").
 ↪sort_values("created_at", ascending=False).index[0]
ax_2 = sns.displot(tweets["elonmusk"][tweets["elonmusk"]["device"] ==␣
 ↪topelonmusk]["month"], bins = range(1, 14));
ax_2.set(xlabel='month', ylabel='Count of Number of Tweets', title = "Elon␣
 ↪Musk's Count of Number of Tweets by Month on Top Device");
```

```
tweets["Cristiano"]["month"] = tweets["Cristiano"]["converted_time"].dt.month
topCristiano = tweets["Cristiano"].groupby("device").agg("count").
 ↪sort_values("created_at", ascending=False).index[0]
ax_3 = sns.displot(tweets["Cristiano"][tweets["Cristiano"]["device"] ==␣
 ↪topCristiano]["month"], bins = range(1, 14));
ax_3.set(xlabel='month', ylabel='Count of Number of Tweets', title =␣
 ↪"Cristiano's Count of Number of Tweets by Month on Top Device");
```







### 0.0.4 Question 5b

Use this space to pur your EDA description.

Description:

For this EDA, I digged deeper into the relationship between the number of tweet counts and the month that the tweet was created for the top device of the user. To elaborate, I tried to investigate and identify trends in how Elon Musk's, AOC's, and Cristiano's usage of their top device changed based on the month of year. To do this, I first extracted the month from the converted_time column and added the month as a new column to the original table. Then, I identified the top device of the user by grouping the devices and aggregating the counts. Afterwards, I created a mask to filter only the top device of the user and made a histogram plot of the count of number of tweets based on month. Lastly, I set the appropriate x-axis, y-axis, and title of the histogram. I repeat this process for all three users that I want to investigate.

Analysis of Result:

AOC's number of tweets on her top device (Twitter for iPhone) is around 500 twitter counts in January, and this continues to decrease through September to around 100 twitter counts. Then there is a significant increase in tweet counts for both November and December, to around 350 counts in December. Elon Musk's number of tweets on his top device (Twitter for iPhone) is around 220 twitter counts in January, which saw a dramatic drop in the following two months to under 50 counts for both months. However, this number increased to around 400 in April and May. The month with the most number of tweets is July, and the counts decreases through December.Cristianos's number of tweets on his top device (Twitter for iPhone) is more evenly distributed throughout the year: around 78 twitter counts in January and stayed around this number for the months February, April, and October, with a few spikes in March, June, July, and August. The month with the most number of tweets is December. One similarity between the AOC and Cristiano's usage is that they both use their top devices to tweet the most during January/December, which is likely due to the holiday seasons. Another observation is that the top device for all three users is Twitter for iPhone, possibly because this device is the most accessible and convenient.