0.1 Question 0

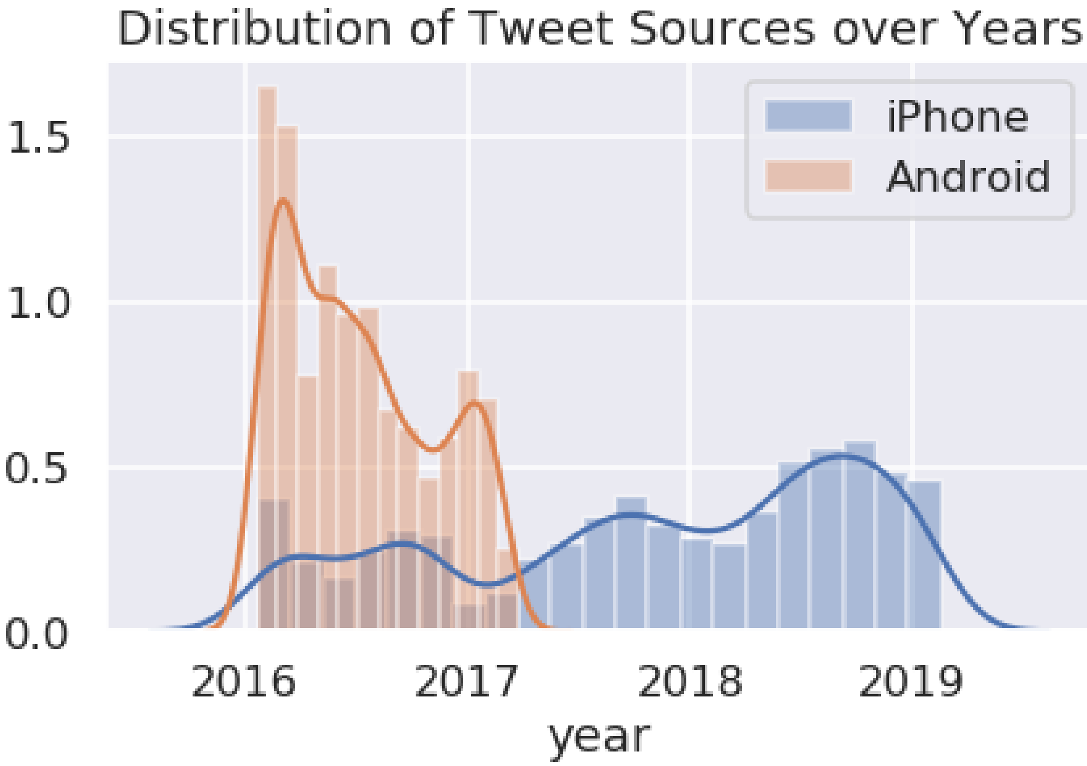
Why might someone be interested in doing data analysis on the President's tweets? Name one person or entity which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

Voters in America may be interested in this kind of analysis because an individual's tweet can expose a lot about a person's behavior, interest, intentions, and agenda. Voters may be also be interested in his tweet analysis to see if his values align with theirs so that they can determine whether or not they want to re-elect him or find reasons to dislike him.

Now, use `sns.distplot` to overlay the distributions of Trump's 2 most frequently used web technologies over the years. Your final plot should look similar to the plot below:

```
In [13]: sns.distplot(trump[trump['source'] == 'Twitter for iPhone']['year'])
         sns.distplot(trump[trump['source'] == 'Twitter for Android']['year'])

         plt.legend(['iPhone', 'Android'])
         plt.title('Distribution of Tweet Sources over Years');
```

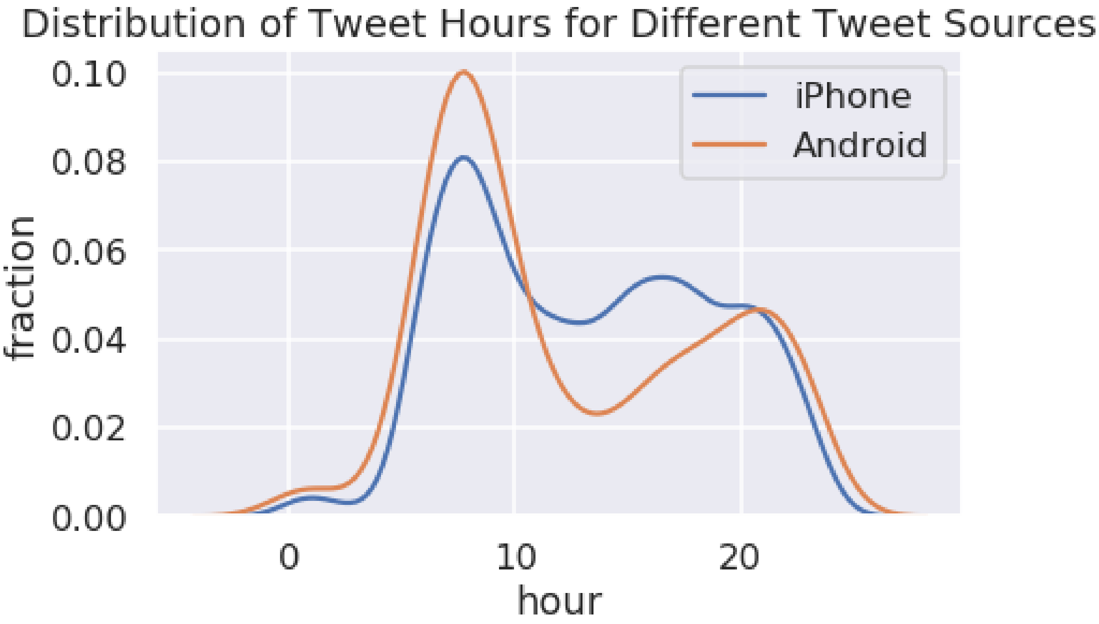


0.1.1 Question 4b

Use this data along with the seaborn `distplot` function to examine the distribution over hours of the day in eastern time that Trump tweets on each device for the 2 most commonly used devices. Your final plot should look similar to the following:

```
In [18]: ### make your plot here
         sns.distplot(trump[trump['source'] == 'Twitter for iPhone']['hour'], label = 'iPhone', hist =
         sns.distplot(trump[trump['source'] == 'Twitter for Android']['hour'], label = 'Android', hist

         plt.ylabel('fraction')
         plt.title('Distribution of Tweet Hours for Different Tweet Sources')
         plt.show()
```



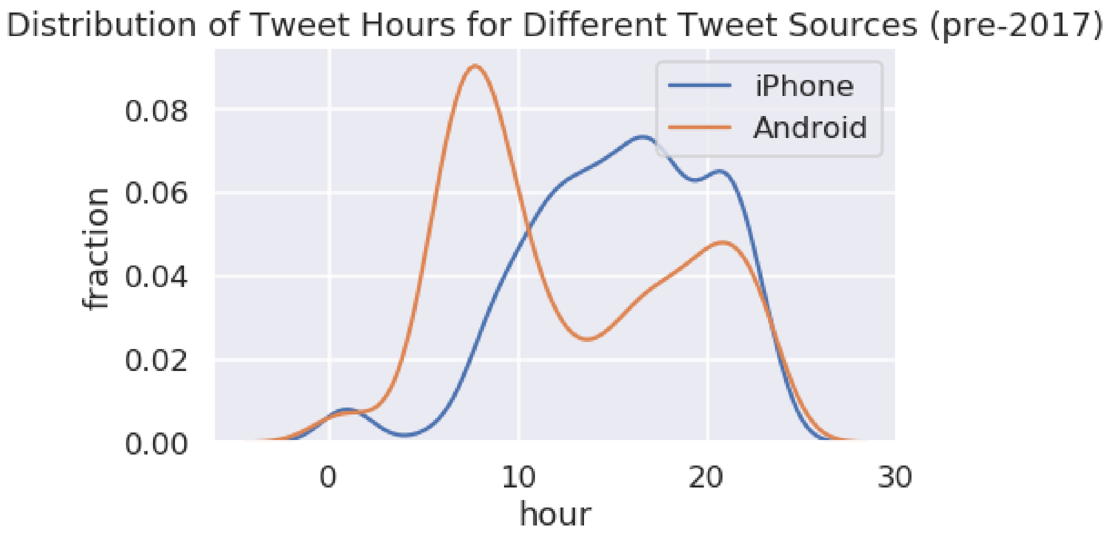
0.1.2 Question 4c

According to this Verge article, Donald Trump switched from an Android to an iPhone sometime in March 2017.

Let's see if this information significantly changes our plot. Create a figure similar to your figure from question 4b, but this time, only use tweets that were tweeted before 2017. Your plot should look similar to the following:

```
In [19]: ### make your plot here
         sns.distplot(trump[(trump['source'] == 'Twitter for iPhone') & (trump['year'] < 2017)]['hour']
         sns.distplot(trump[(trump['source'] == 'Twitter for Android') & (trump['year'] < 2017)]['hour']

         plt.ylabel('fraction')
         plt.title('Distribution of Tweet Hours for Different Tweet Sources (pre-2017)')
         plt.show()
```



0.1.3 Question 4d

During the campaign, it was theorized that Donald Trump's tweets from Android devices were written by him personally, and the tweets from iPhones were from his staff. Does your figure give support to this theory? What kinds of additional analysis could help support or reject this claim?

Yes the figure does support this theory, looking at the two graphs, the curve for android stays the same for both graphs while we see a shift and increase in the usage of iphones in the pre-2017 graph. This may indicate that, prior to 2017, his staff were tweeting more (around 10am-8pm which are working hours) with an iphone during his campaign. The fraction of his iphone usage then decreased and shifted to almost match his android usage (as shown in the first graph) when he made the switch to an iphone after he got elected; he would have less tweets because he no longer is campaigning and the shift would mean he's tweeting at the same hours that he would with his android. Additional analysis may include comparing his iphone usage exclusively after 2017 (when he made the switch) to that of his android usage prior to 2017 to see if the frequencies match at the right hours.

0.2 Question 5

The creators of VADER describe the tool's assessment of polarity, or "compound score," in the following way:

"The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate."

As you can see, VADER doesn't "read" sentences, but works by parsing sentences into words assigning a preset generalized score from their testing sets to each word separately.

VADER relies on humans to stabilize its scoring. The creators use Amazon Mechanical Turk, a crowdsourcing survey platform, to train its model. Its training set of data consists of a small corpus of tweets, New York Times editorials and news articles, Rotten Tomatoes reviews, and Amazon product reviews, tokenized using the natural language toolkit (NLTK). Each word in each dataset was reviewed and rated by at least 20 trained individuals who had signed up to work on these tasks through Mechanical Turk.

0.2.1 Question 5a

Given the above information about how VADER works, name one advantage and one disadvantage of using VADER in our analysis.

An advantage is that it allows us to assess the polarity of a tweet and we can assign a positive or negative sentiment without having to read through all of the tweets individually. A disadvantage is that it looks it breaks up the tweets into words and then assign values on that word. This might be an issue when words might carry different connotations or are linked with other words to gather meaning, it's very difficult to interpret different contexts in this case and thus might incorrectly assign a sentiment.

0.2.2 Question 5b

Are there circumstances (e.g. certain kinds of language or data) when you might not want to use VADER? Please answer “Yes,” or “No,” and provide 1 reason for your answer.

Yes, because VADER is trained on things like reviews, tweets, etc., it’s very catered towards language used in that context. For texts that are highly specialized towards a field like healthcare (which contain it’s own jargon), VADER might not be able to catch and accurately score every word.

0.3 Question 5h

Read the 5 most positive and 5 most negative tweets. Do you think these tweets are accurately represented by their polarity scores?

Yes I think these tweets are accurately represented by their polarity scores. The negative tweets mention a lot of dark things that I would classify as negative topics while the positive tweets are more encouraging & uplifting (includes congradulations).

0.4 Question 6

Now, let's try looking at the distributions of sentiments for tweets containing certain keywords.

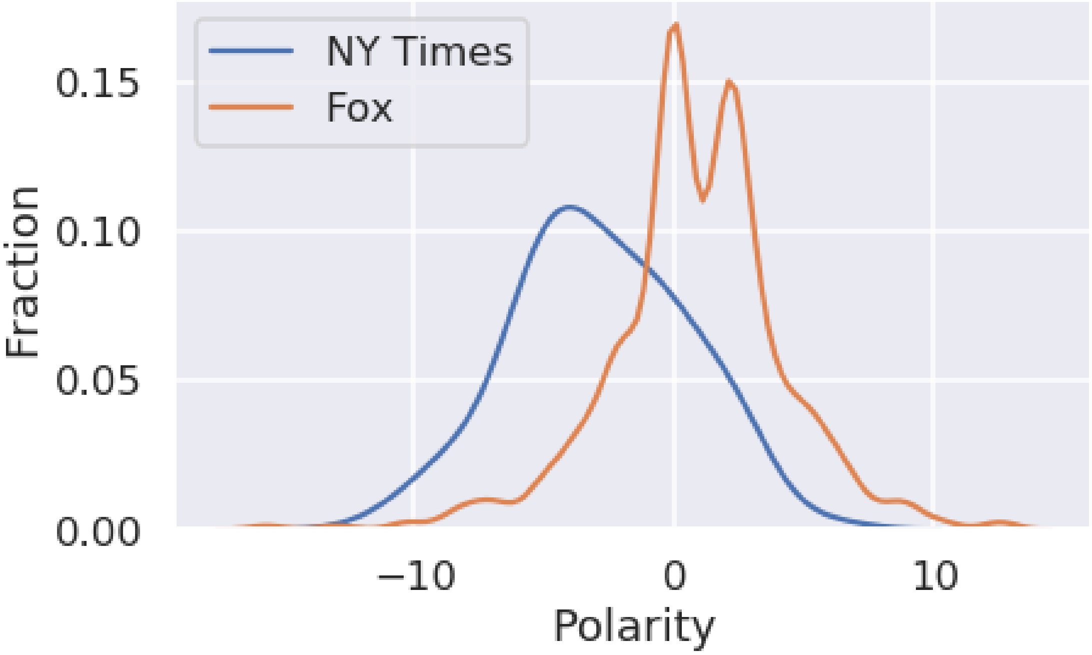
0.4.1 Question 6a

In the cell below, create a single plot showing both the distribution of tweet sentiments for tweets containing `nytimes`, as well as the distribution of tweet sentiments for tweets containing `fox`.

Be sure to label your axes and provide a title and legend. Be sure to use different colors for `fox` and `nytimes`.

```
In [33]: sns.distplot(trump[trump['text'].str.contains("nytimes")][['polarity']], label = 'NY Times', hi
         sns.distplot(trump[trump['text'].str.contains("fox")][['polarity']], label = 'Fox', hist = Fals


         plt.xlabel('Polarity')
         plt.ylabel('Fraction')
         plt.legend()
         plt.show()
```



0.4.2 Question 6b

Comment on what you observe in the plot above. Can you find another pair of keywords that lead to interesting plots? Describe what makes the plots interesting. (If you modify your code in 6a, remember to change the words back to `nytimes` and `fox` before submitting for grading).

The plot above shows that tweets with 'Fox' have generally have a higher polarity than tweets with 'nytimes'. This is interesting because Fox is known to be conservative leaning so the fact that these tweets generally have a higher polarity is not surprising.

What do you notice about the distributions? Answer in 1-2 sentences.

Tweets without hashtags/links have a greater spread (larger standard deviation) than tweets with hashtags/links and tweets with hashtags/links are more centered around 0 (comparatively). There is a higher peak around the neutral point of 0 for tweets with hashtags while the peak for tweets without hastags is much lower. Additionally, the distribution for tweets without hashtags/links is more evenly spread out across -10 to 10 polarity scores (half above 0 and half below) while the distribution for tweets with hashtags/links is shifted a little bit to the right such that most of the scores are above 0.