0.1 Question 0

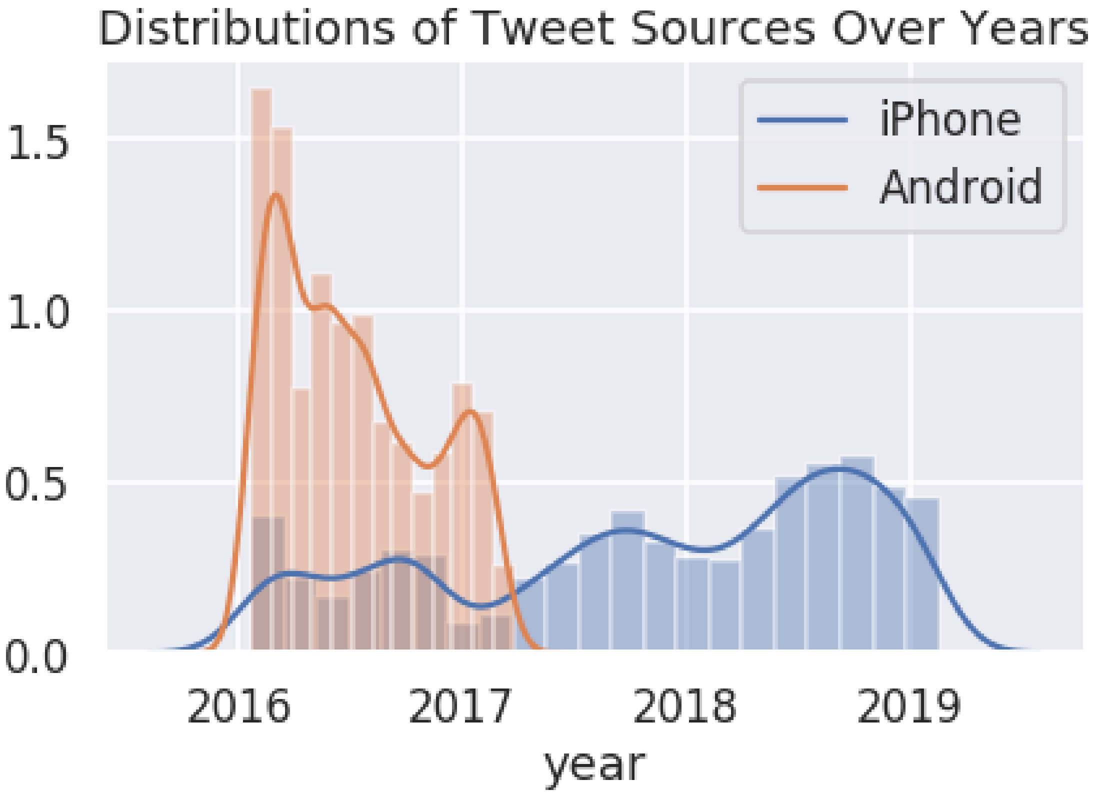
There are many ways we could choose to read the President's tweets. Why might someone be interested in doing data analysis on the President's tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

Someone who would be interested in this analysis may be a news anchor, Presidential Debate moderator, or someone working for the opposition's Presidential campaign, and institutions that may be interested would include news agencies e.g. CNN.

Presidential Debate moderators may be interested because they can then tailor their questions in the debate according to particular instances of the President's reaction to something, while someone working in the opposition's Presidential campaign may be interested because they can use the President's negligence/avoidance of a certain issue to further their agenda. Furthermore, seeing users' responses to his tweets would allow institutions such as news agencies and individuals such as news anchors to get an idea about how the people are feeling about the President and his policies, and this would allow them to make informed predictions about the upcoming elections.

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:

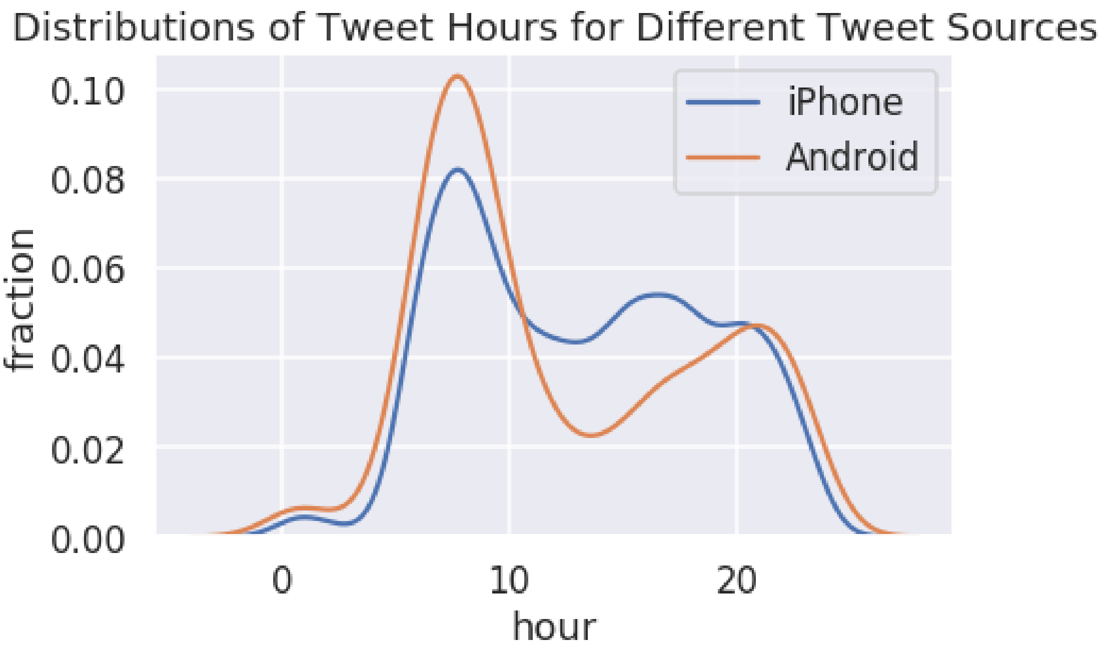
```
In [14]: plt.figure(figsize=(6, 3.9))
         trump_android = trump[trump['source'] == 'Twitter for Android']['year']
         trump_iphone = trump[trump['source'] ==  'Twitter for iPhone']['year']
         sns.distplot(trump_iphone);
         sns.distplot(trump_android);
         plt.legend(['iPhone', 'Android'])
         plt.title('Distributions of Tweet Sources Over Years')
         plt.ylabel('')
         plt.show()
```



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 [19]: ### make your plot here
         plt.figure(figsize=(6, 3.9))
         trump_android_est = trump[trump['source'] == 'Twitter for Android']['hour']
         trump_iphone_est = trump[trump['source'] ==  'Twitter for iPhone']['hour']
         sns.distplot(trump_iphone_est, hist=False);
         sns.distplot(trump_android_est, hist=False);
         plt.legend(['iPhone', 'Android'])
         plt.title('Distributions of Tweet Hours for Different Tweet Sources')
         plt.ylabel('fraction')
         plt.show()
```

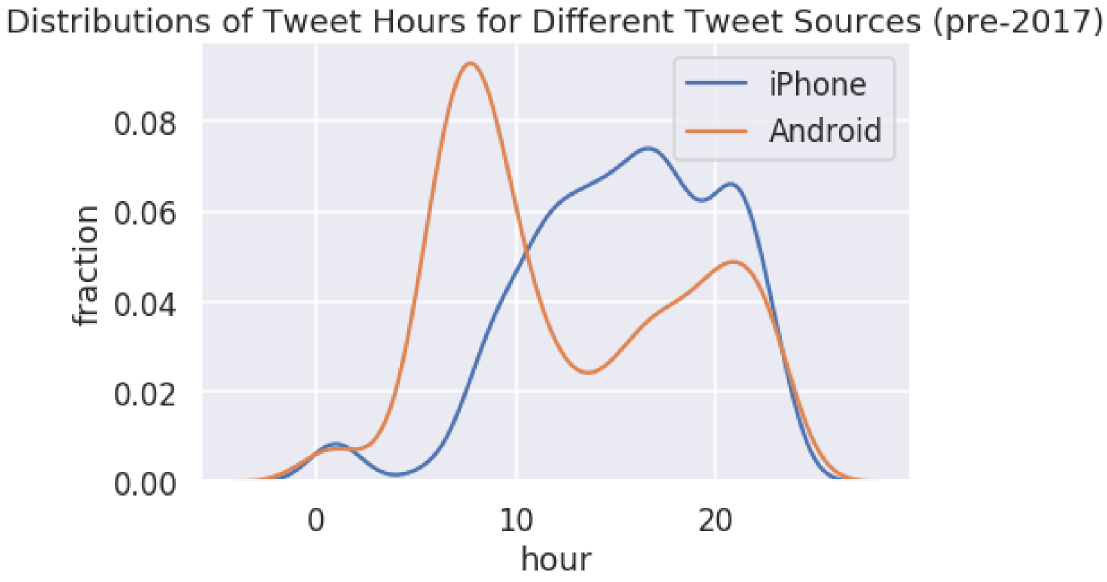


0.1.2 Question 4c

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 [20]: ### make your plot here
         trump_pre_2017 = trump[trump['year'] < 2017]
         plt.figure(figsize=(6.32, 4.2))
         trump_android = trump_pre_2017[trump_pre_2017['source'] == 'Twitter for Android']['hour']
         trump_iphone = trump_pre_2017[trump_pre_2017['source'] ==  'Twitter for iPhone']['hour']
         sns.distplot(trump_iphone, hist=False);
         sns.distplot(trump_android, hist=False);
         plt.legend(['iPhone', 'Android'])
         plt.title('Distributions of Tweet Hours for Different Tweet Sources (pre-2017)')
         plt.ylabel('fraction')
         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, it does. Firstly, the times at which peaks occur on the above graph prove our claim. The highest number of Android tweets are near 8-9am, and that is likely to be the time when he wakes up and hence uses Twitter to set the tone for the rest of the day. On the other hand, the the highest number of iPhone tweets occur at 4-5pm, and this is likely to be the time when Trump is most swamped with presidential responsibilities. This also makes sense because the maximum of the orange graph coincides with the minimum of the blue graph and vice versa. Whenever Trump is caught up with presidential responsibilities, his staff takes over and tweets for him.

Secondly, the "width" of each peak also supports our claim. The peak of the Android graph is narrower which makes sense because Trump is just one person, and cannot tweet consistently for so long. On the other hand, the peak of the iPhone graph is wider which aligns with our speculation because his staff consists of a group of individuals so they can all take turns tweeting.

Something that could support or reject this claim would be if we had information about the number of social media managers he had, and their working 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

Please score the sentiment of one of the following words: - 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?

I would give 'vaccine' a 0.6 score. The reason I would give it this score is because vaccine is mostly associated with positive sentiment i.e. the hope of curing COVID-19, which points towards the end of the pandemic that has cost people their lives, money and jobs. In more recent times, vaccines were making progress, with many even getting to the clinical trials stage.

However, 'vaccine' can also carry a negative sentiment in some cases. If the COVID-19 vaccine were delayed or if it would take time to be distributed/manufactured on a mass scale, negative sentiments around the word 'vaccine' would rise. Examples include: "Trial of Moderna Covid-19 vaccine delayed" or "Most people likely won't get a coronavirus vaccine until the middle of 2021".

0.2.2 Question 5b

VADER aggregates the sentiment of words in order to determine the overall sentiment of a sentence, and further aggregates sentences to assign just one aggregated score to a whole tweet or collection of tweets. This is a complex process and if you'd like to learn more about how VADER aggregates sentiment, here is the info at this link.

Are there circumstances (e.g. certain kinds of language or data) when you might not want to use VADER? What features of human speech might VADER misrepresent or fail to capture?

Yes, there are. These circumstances include memes and sarcastic comments on the internet, where some words are often used in a different context than they are presumed to be used. Furthermore, VADER may not be the best for long paragraphs of human speech because in those cases the word usage depends heavily on the context. Another case where we may not want to use VADER is for ancient/historical texts because word meanings (and hence, sentiments) change over time. An example is the word ”awful” which used to mean awesome and inspiring (”awe-ful”) but means disgusting and terrible today.

Some features that VADER fails to take into account are the pitch, loudness and tone of human speech. Some words used by humans totally depend on the way they are said and also on the context in which they are used. E.g. ”damn” could be used angrily to express discontent towards a particular situation but it can also be used positively to express admiration of something/someone (”I hate this damn pandemic!” versus ”Damn, did you see Rihanna perform last night?”). The two vary in the context in which the individual uses the word as well as how loud/in what pitch they say it.

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 talk about killings, anti-semitism, hate, and his critisism of people (James Comey). On the other hand, the positive tweets express gratitude towards his supporters and congratulatory remarks to people/teams.

0.4 Question 6

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

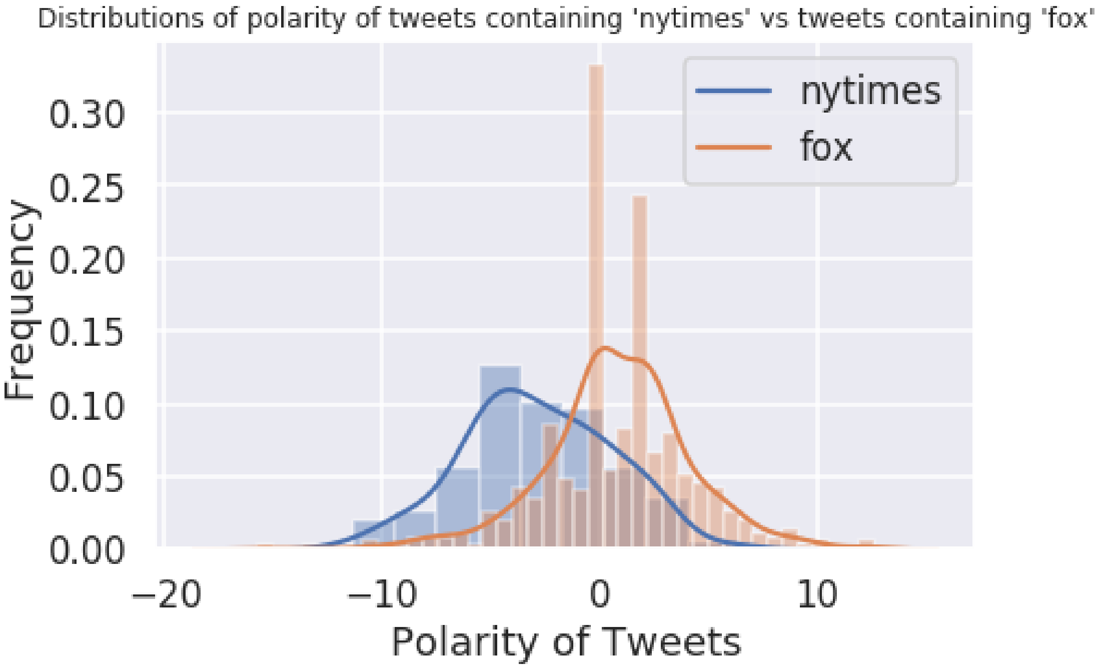
0.4.1 Question 6a

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 [34]: contains_nytimes = trump[trump['text'].str.contains('nytimes')]['polarity']
         contains_fox = trump[trump['text'].str.contains('fox')]['polarity']
         plt.figure(figsize=(6, 4))
         sns.distplot(contains_nytimes)
         sns.distplot(contains_fox)
         plt.legend(['nytimes', 'fox'])
         plt.title("Distributions of polarity of tweets containing 'nytimes' vs tweets containing 'fox'
         plt.xlabel('Polarity of Tweets')
         plt.ylabel('Frequency')
```

```
Out[34]: Text(0, 0.5, 'Frequency')
```

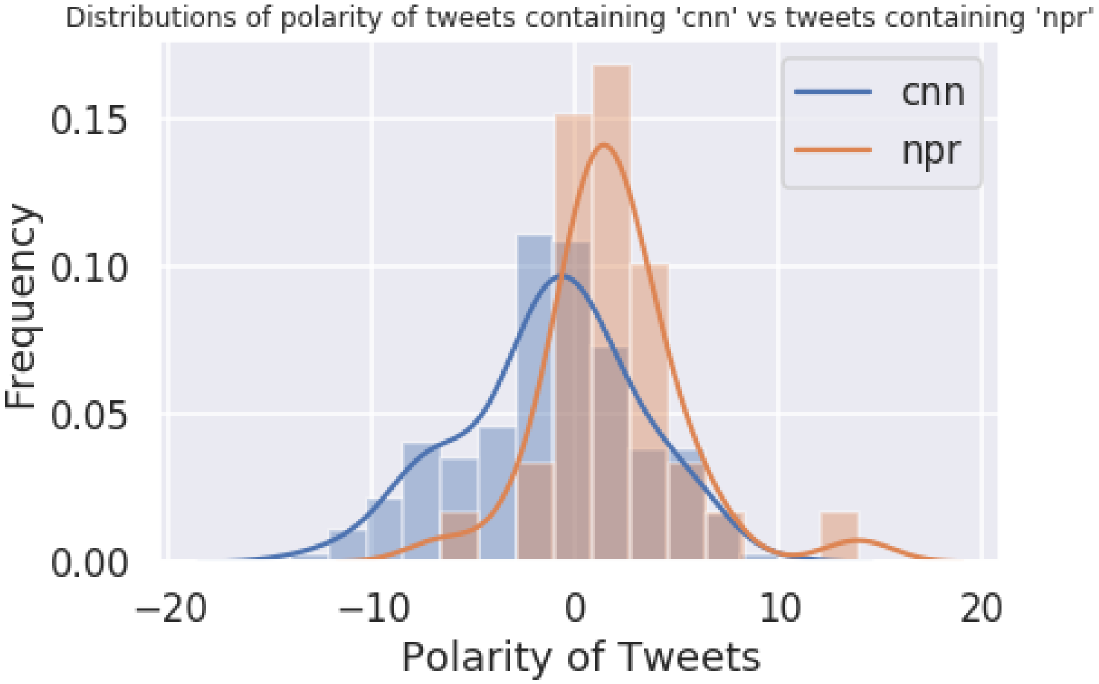


0.4.2 Question 6b

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).

```
In [35]: contains_cnn = trump[trump['text'].str.contains('cnn')]['polarity']
         contains_npr = trump[trump['text'].str.contains('npr')]['polarity']
         plt.figure(figsize=(6, 4))
         sns.distplot(contains_cnn)
         sns.distplot(contains_npr)
         plt.legend(['cnn', 'npr'])
         plt.title("Distributions of polarity of tweets containing 'cnn' vs tweets containing 'npr'", f
         plt.xlabel('Polarity of Tweets')
         plt.ylabel('Frequency')
```

```
Out[35]: Text(0, 0.5, 'Frequency')
```



In general, tweets containing the word ”fox” had a higher polarity than tweets containing the word ”nytimes”. This means that there was a more positive sentiment in tweets containing ”fox” than ”nytimes”. If we observe the words ”cnn” and ”npr”, a similar relationship is observed. What is interesting is that CNN seems to have the highest frequency near a polarity of 0, i.e. the sentiment surrounding CNN is roughly neutral. It may be interesting to wonder if this is why Georgia (the home of CNN) is regarded as a swing state.

It is also interesting to note that the word with the higher polarity has a higher maximum frequency e.g. ’fox’ has a higher maximum frequency than ’nytimes’ and ’npr’ has a higher maximum frequency than ’cnn’. Perhaps this is because a greater proportion of Trump’s tweets mention positive news associated to him, and Fox News and NPR were more pro-Trump than the NY Times and CNN.

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

In general, the distributions show that tweets with hashtags or links seem to have a more neutral polarity than tweets without hashtags or links because the frequency for the blue graph is higher than the frequency for the orange graph at polarity = 0. Furthermore, the distribution for tweets without hashtag or link is more of an 'even' distribution i.e. it has less gaps and looks more like a normal distribution than the distribution for tweets with hashtag or link.