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.

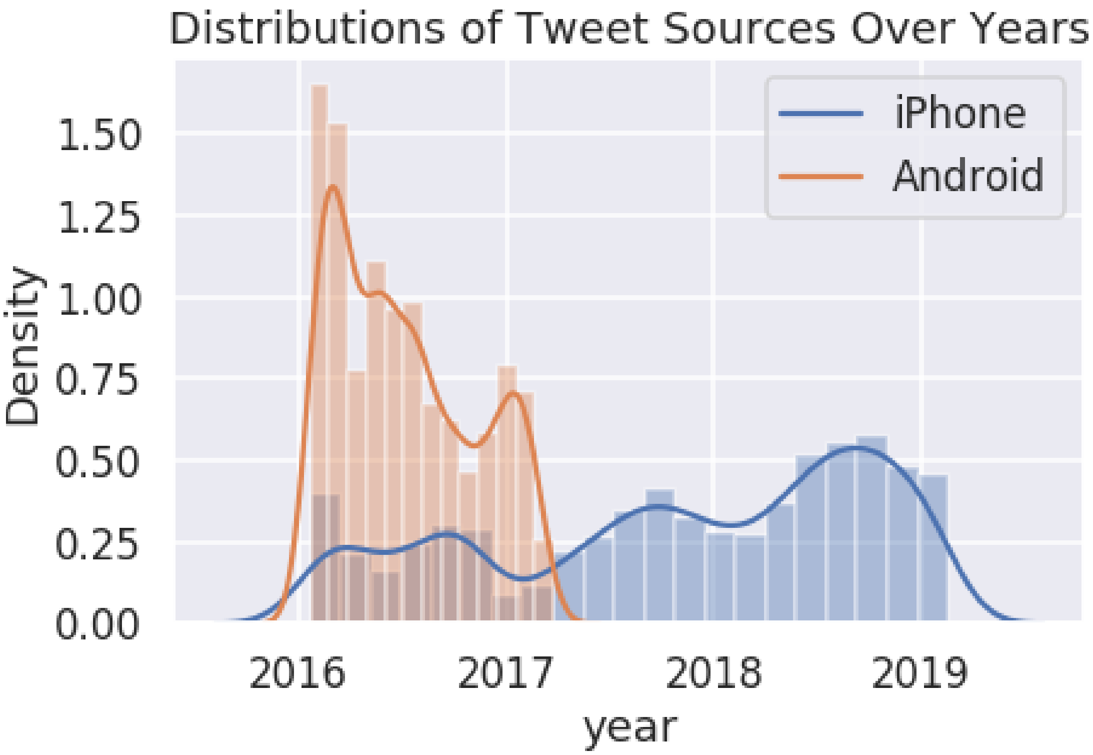
With election coming up it is important to analyze Trump's tweets in order to get an idea of Trumps tweeting behavior as well as the response toward his tweets from his followers. This information can be helpful to news compamnies such as CNN/Fox or FiveThirtyEight who want to report on updates about 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 [89]: 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.xlabel("year")
         plt.show()
```



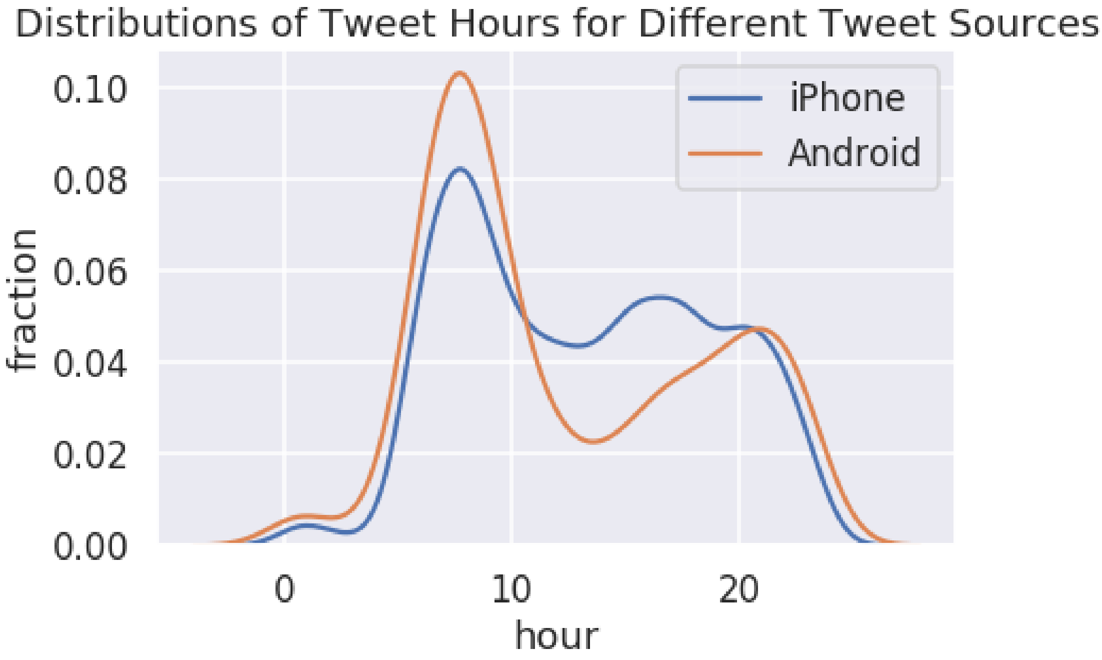
0.1.1 Question 4b

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 [94]: ### make your plot here
         trump_android = trump[trump['source'] == 'Twitter for Android']['hour']
         trump_iphone = trump[trump['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")
         plt.xlabel("hour")
         plt.ylabel("fraction")
         plt.show()
```



0.1.2 Question 4c

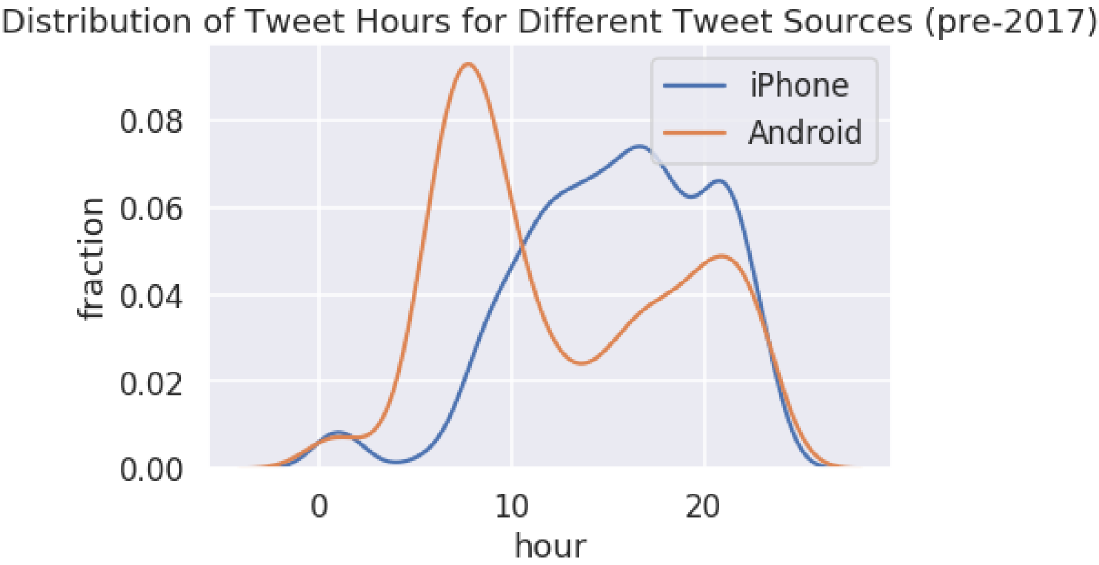
According to this Verge article, Donald Trump switched from an Android to an iPhone sometime in March 2017.

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 [95]: ### make your plot here
         trump_android = trump[trump['source'] == 'Twitter for Android'].query('year < 2017')['hour']
         trump_iphone = trump[trump['source'] == 'Twitter for iPhone'].query('year < 2017')['hour']

         sns.distplot(trump_iphone, hist = False)
         sns.distplot(trump_android, hist = False)

         plt.legend(["iPhone", "Android"])
         plt.title("Distribution of Tweet Hours for Different Tweet Sources (pre-2017)")
         plt.xlabel("hour")
         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?

There is slight evidence for this scenario. In the cases pre 2017, we see the android tweets pretty consistent with the overall distribtion shown in 4b, with a peak in the morning around 8 and slight peak at 22. The iphone tweets from pre 2017 have a different distribution, tweeting throughout the day making it seem like it could be from a different source of people. Potentially it could be his staff tweeting throughout the day.

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?

"Dog" would generally be a positive word with a score of 0.5 with the prevalence of dogs being pets and being accepted part of society. In the situation we are talking about police dogs regarding riots and BLM movement, this word can have a negative score. In this context is very important.

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?

You would properly not want to use VADER when working with large amounts of texts with a lot of context and correlated words, as it would be less accurate and increase inaccuracy. It would also be inneffective in detecting sarcasm or any situation where people dont mean what they are exactly saying.

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 the 5 most negative tweets are about negative topics such as ICE, drugs and illegal activity. The 5 most positive tweets are Trump thanking, congratulating and complimenting people.

0.4 Question 6

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

0.4.1 Question 6a

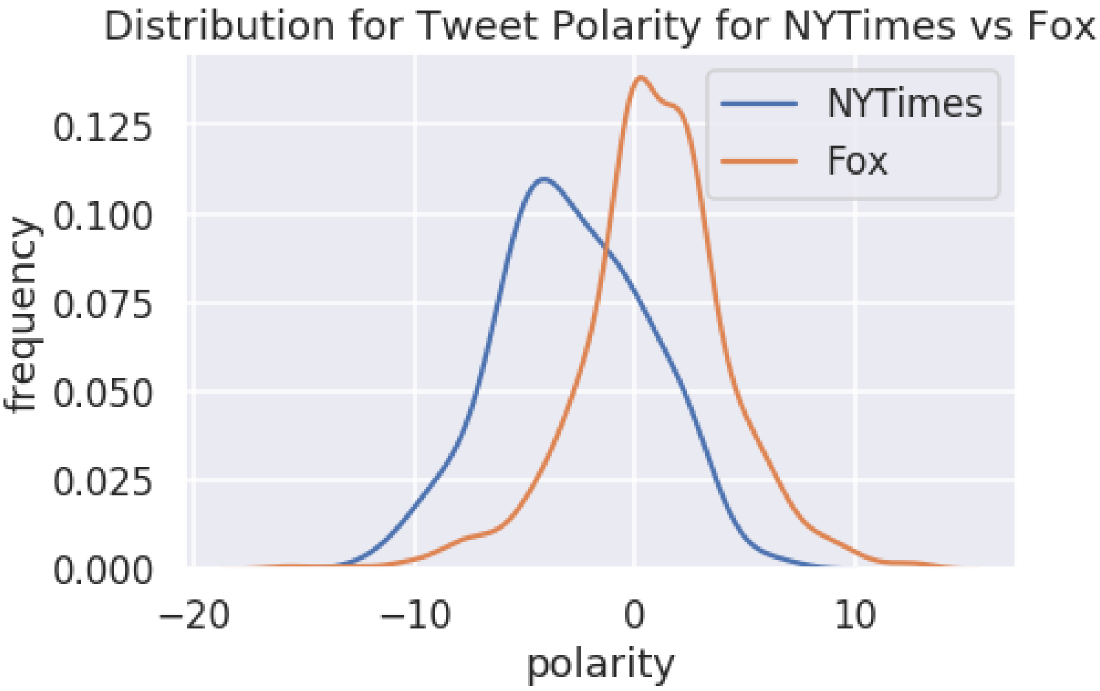
In the cell below, create a single plot showing both the distribution of tweet sentiments for tweets containing `nytimes`, as well as the distribution of tweet sentiments for tweets containing `fox`.

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

```
In [137]: nytimes = trump[trump["text"].str.contains("nytimes")]["polarity"]
          fox = trump[trump["text"].str.contains("fox")]["polarity"]

          sns.distplot(nytimes, hist = False)
          sns.distplot(fox, hist = False)

          plt.xlabel("polarity")
          plt.ylabel("frequency")
          plt.title("Distribution for Tweet Polarity for NYTimes vs Fox")
          plt.legend(["NYTimes", "Fox"])
          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 Fox distribution has a general higher polarity compared to NYTimes. The NyTimes and Fox distributions are unimodal. The NYTimes distribution has a mean around -5 and has a slight right skew while the Fox distribution has a center around 0 and is fairly symmetrical. Fox has a general polarity of which is higher to NYTimes.

We can see a difference between keywords ”mexican” and ”christian” with christians also having a higher mean, indicating that Trump had more positive tweet regarding christians.

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

The distribution of no hashtag or link has a mean around 0 and is fairly symmetrical. The distribution of hashtag or linked tweets has a mean slightly larger than 0 and has a right skew. Thus we can conlude that the hashtagged or linked data has a higher polarity.