0.1 Question 0

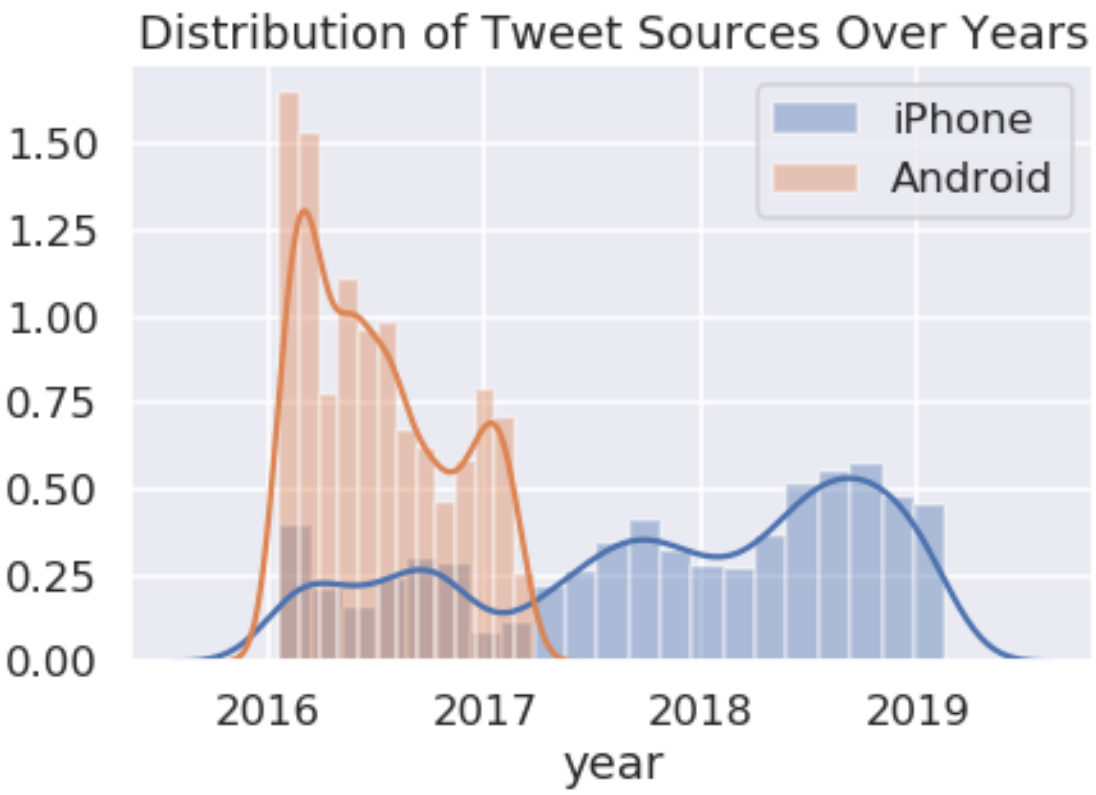
Why might someone be interested in doing data analysis on the President's tweets? Name one person or entity which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

News networks such as CNN and Fox might be interested in analyzing the president's tweets to track his social media activity during the election or other important times. Data analysis of the president's tweets could provide insight into the responses and activity his social media presence generates and also gives direct access to his thoughts and opinions on various national issues.

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 [82]: android_uses = trump[trump['source'] == 'Twitter for Android']['year']
         iphone_uses = trump[trump['source'] == 'Twitter for iPhone']['year']
         sns.distplot(iphone_uses)
         sns.distplot(android_uses)
         plt.legend(['iPhone','Android'])
         plt.title('Distribution of Tweet Sources Over Years')
```

```
Out[82]: Text(0.5, 1.0, 'Distribution of Tweet Sources Over Years')
```

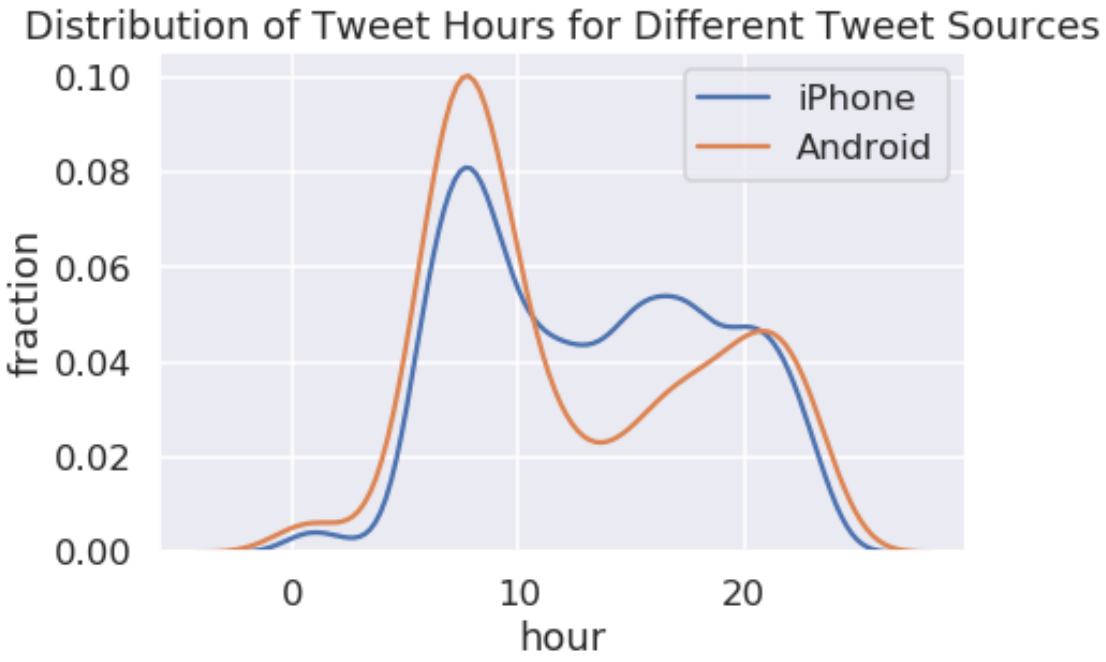


0.1.1 Question 4b

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 [87]: ### make your plot here
         sns.distplot(trump[trump['source'] == 'Twitter for iPhone']['hour'], hist=False, label='iPhone
         sns.distplot(trump[trump['source'] == 'Twitter for Android']['hour'], hist=False, label='Andro
         plt.ylabel('fraction')
         plt.title('Distribution of Tweet Hours for Different Tweet Sources')
```

```
Out[87]: Text(0.5, 1.0, 'Distribution of Tweet Hours for Different Tweet Sources')
```



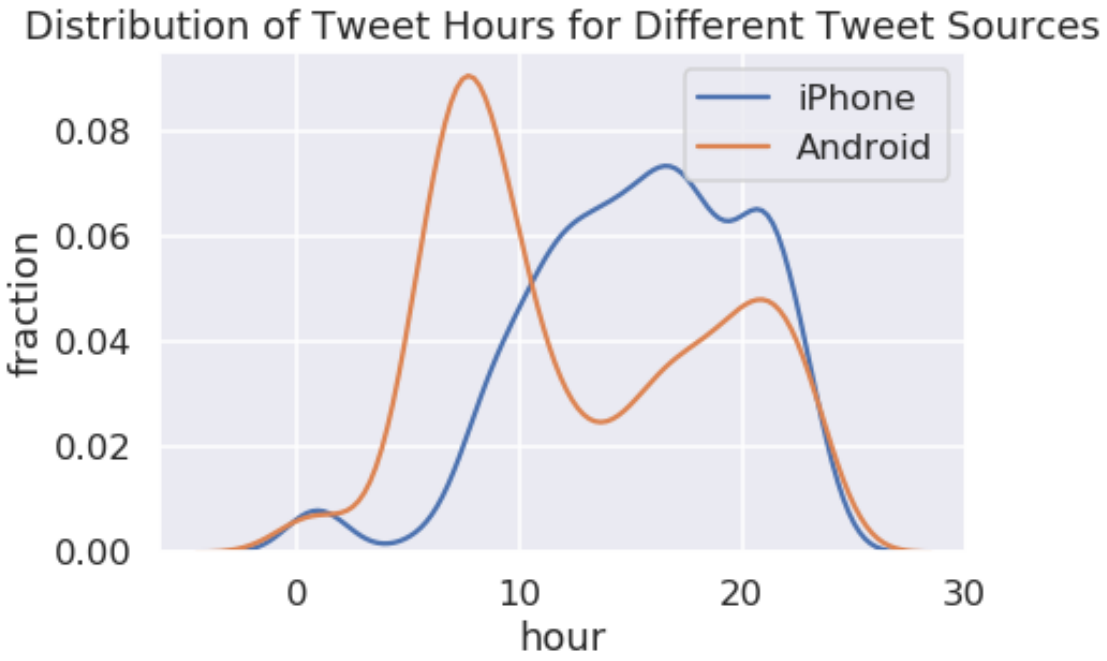
0.1.2 Question 4c

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 [88]:
```
### 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')
```

Out[88]:
```
Text(0.5, 1.0, 'Distribution of Tweet Hours for Different Tweet Sources')
```



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?

The distribution of tweet hours suggests that his staff wrote tweets for him. Trump seems to tweet the most around 8-9 AM and around 9-10PM using his android, which seems like a reasonable time for one person to tweet. Trump's tweets from his iphone constantly from around 11AM - 8PM, which seems like his staff may be doing it for him.

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 of using VADER is that we are able to differentiate the intensity between words of similar meaning, such as love and like. A disadvantage of using VADER is that sentences are scored word by word instead of the entire sentence, which could result in a sentiment score that does not correctly represent the overall positivity/negativity of the sentence.

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. Certain types of language, such as sarcasm and indirect insults, could have hidden implied meaning that are opposite to the actual meaning.

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?

The polarity scores seem to represent most of the tweets well. Most of the negative tweets did have some sort of negative message while the positive ones definitely had positive words and messages. However, one of the tweets discussing about fighting anti-semitism was labeled negative due to the presence of words such as hate. Therefore, not all the tweets are accurately represented.

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`.

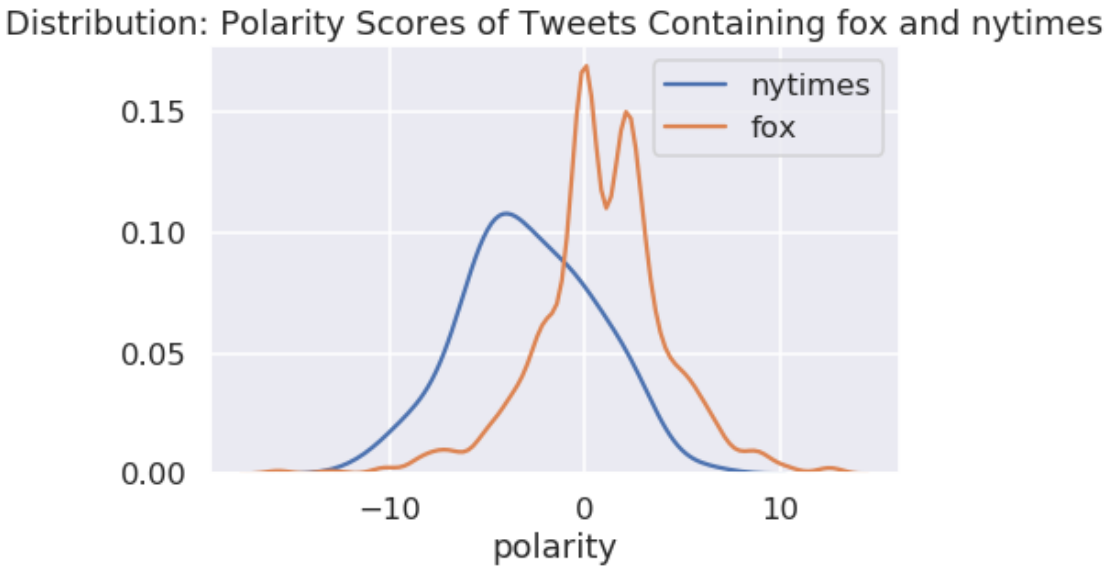
```
In [115]: ny = trump.loc[trump['text'].str.contains('nytimes')]['polarity']
          fox = trump.loc[trump['text'].str.contains('fox')]['polarity']
          #mx = trump.loc[trump['text'].str.contains('mexico')]['polarity']
          #border = trump.loc[trump['text'].str.contains('border')]['polarity']

          sns.distplot(ny, hist=False, label = 'nytimes')
          sns.distplot(fox, hist=False, label='fox')
          #sns.distplot(mx, hist=False, label = 'mexico')
          #sns.distplot(border, hist=False, label='border')



          plt.title('Distribution: Polarity Scores of Tweets Containing fox and nytimes')
```

```
Out[115]: Text(0.5, 1.0, 'Distribution: Polarity Scores of Tweets Containing fox and nytimes')
```



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).

From the plot, the average sentiment score for Trump's tweets containing nytimes were mostly negative while those containing fox were mostly positive. Another pair of keywords that lead to interesting plots were 'mexico' and 'border'. These tweets appear to have a positive average score, even though Trump mainly tweeted about mexico and the border with negative sentiments during his campaign, which is the opposite of what I expected.

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

The spread of tweets not containing a hashtag is much wider than tweets containing hashtags or links. The no hashtag or link distribution appears very close to a normal distribution with a bell-shaped curve while the distribution with a hashtag or link has a left tail and is not symmetric.