0.1 Question 0

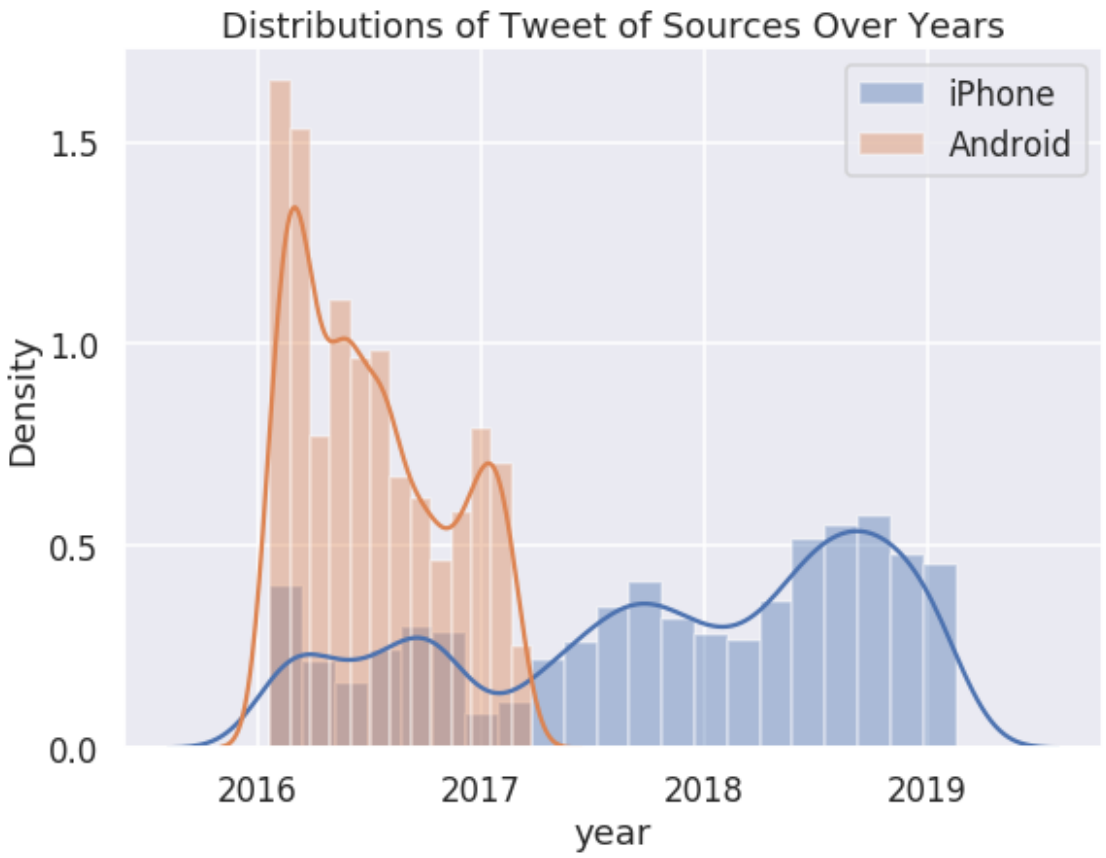
There are many ways we could choose to read the President's tweets. Why might someone be interested in doing data analysis on the President's tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's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

Any kind of news source like CNN might be interested in data analysis of th president's tweets. They might, for example, want to know how many times he has tweeted about something, like golfing or COVID. Or they might want to know how many tweets a day or week the presidents sends out to his followers.

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]:
```
trump_iphone = trump[trump['source'] == 'Twitter for iPhone']
trump_android = trump[trump['source'] == 'Twitter for Android']
fig = plt.figure(figsize = (7,5))
ax = fig.add_axes([0,0,1,1])
ax.set_yticks([0.0,0.5,1.0,1.5])
ax.set_xticks([2016, 2017, 2018, 2019])
sns.distplot(trump_iphone['year'], kde = True, label = 'iPhone')
sns.distplot(trump_android['year'], kde = True, label = 'Android')
plt.title('Distributions of Tweet of Sources Over Years')
plt.legend()
```

Out[14]:
```
<matplotlib.legend.Legend at 0x7f899d3423d0>
```

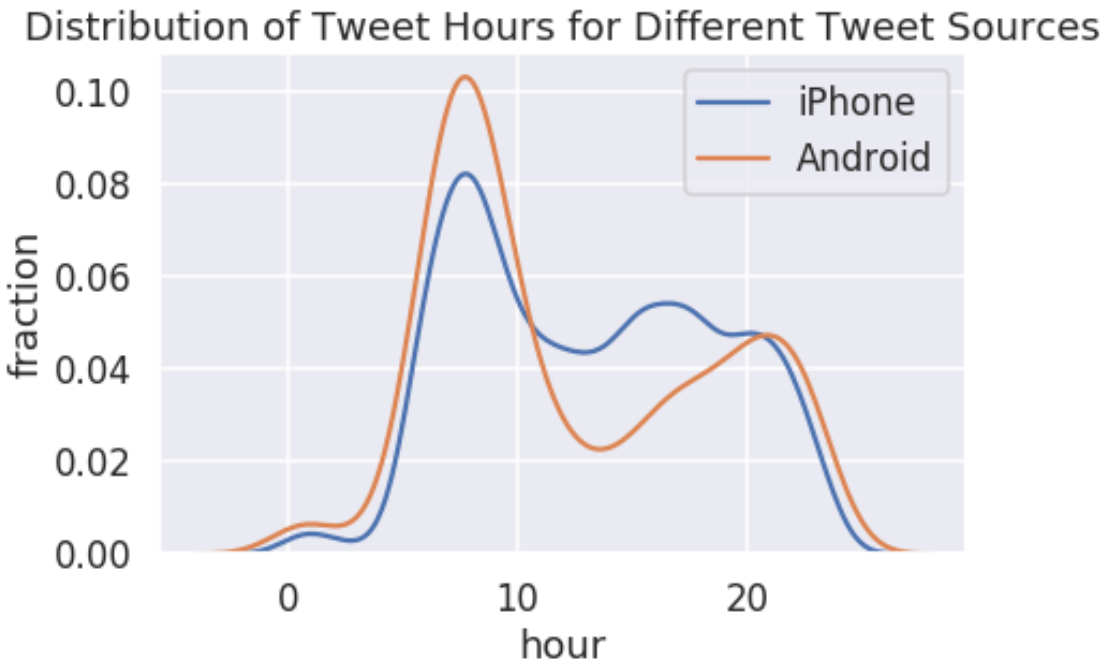


0.1.1 Question 4b

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 [28]:
```
### make your plot here
new_trump_iphone = trump[trump['source']=='Twitter for iPhone']
new_trump_android = trump[trump['source']=='Twitter for Android']
sns.distplot(new_trump_iphone['hour'], hist = False, kde = True, label='iPhone')
sns.distplot(new_trump_android['hour'], hist = False, kde = True, label='Android')
plt.title('Distribution of Tweet Hours for Different Tweet Sources')
plt.xlabel('hour')
plt.ylabel('fraction')
plt.legend()
```

Out[28]: <matplotlib.legend.Legend at 0x7f899d0ba7c0>



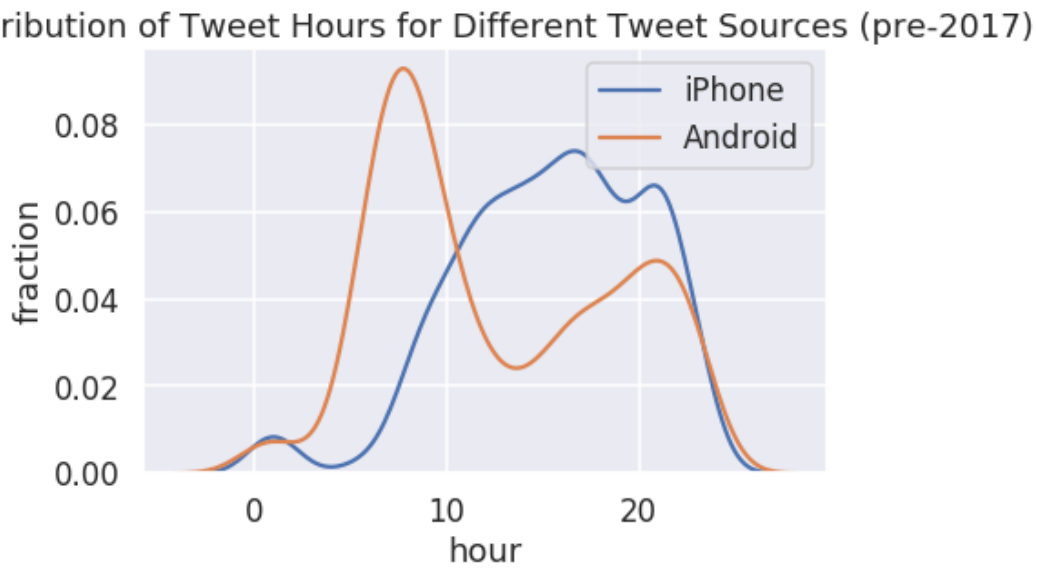
0.1.2 Question 4c

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 [29]:
```
### make your plot here
pre2017_trump = trump[trump['year'] < 2017]
pre2017_trump_iphone = pre2017_trump[pre2017_trump['source'] == 'Twitter for iPhone']
pre2017_trump_android = pre2017_trump[pre2017_trump['source'] == 'Twitter for Android']
sns.distplot(pre2017_trump_iphone['hour'], hist = False, kde = True, label = 'iPhone')
sns.distplot(pre2017_trump_android['hour'], hist = False, kde = True, label = 'Android')
plt.title('Distribution of Tweet Hours for Different Tweet Sources (pre-2017)')
plt.xlabel('hour')
plt.ylabel('fraction')
plt.legend()
```

Out[29]:
```
<matplotlib.legend.Legend at 0x7f899d028490>
```



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 figure does support this theory from the plot in 4b, we can see iPhone is being used more in the evening, which is usually when campaigns or speeches occur. From the plot in 4c, we can see the android graph remains very similar to 4b, while the iphone graph has changed. This plot shows an even bigger discrepancy between iphone and android, with android being used mostly around 8-9 am and 8-9 pm, which is often before and after important campaign accouncements/events. However, iphone usage peaks around 5-6 pm, which is right when speeches and such occur, supporting the idea that his staff is tweeeting out as trump is speaking. However, this is not concrete, and data about when trump is speaking would help prove or disprove this claim.

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?

vaccine: 0/9. I chose to score 'vaccine' 0.9. Given the current pandemic, I felt that mentioning a vaccine would be in a postitve context usually, and appeal to the American people who want to return to pre-COVID society, and a vaccine would help do that. However, there is a lot of distrust around vaccines so if could carry the opposite sentiment if someone is talking about how untrustworthy vaccines are.

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, we might not want to use VADER for a sentence with sarcasm. The VADER might give a positive value because it analyzes the words in it not the whole meaning of the sentence. An example, "I like your dress, just not with you in it." could have a positive value because the words that exist in the sentence does not contain a negative sentiment according to VADER.

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?

I think the tweets are accurately represented by their polarity scores. The negative tweets are about problems and crises such as drug epidemics, personal attacks, anti-semitism, etc. These are tweets with negative emotions attatched to them. The opposite is true for tweets with high positive polarities, which are congratulting or thanking tweets, which are positive in message.

0.4 Question 6

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

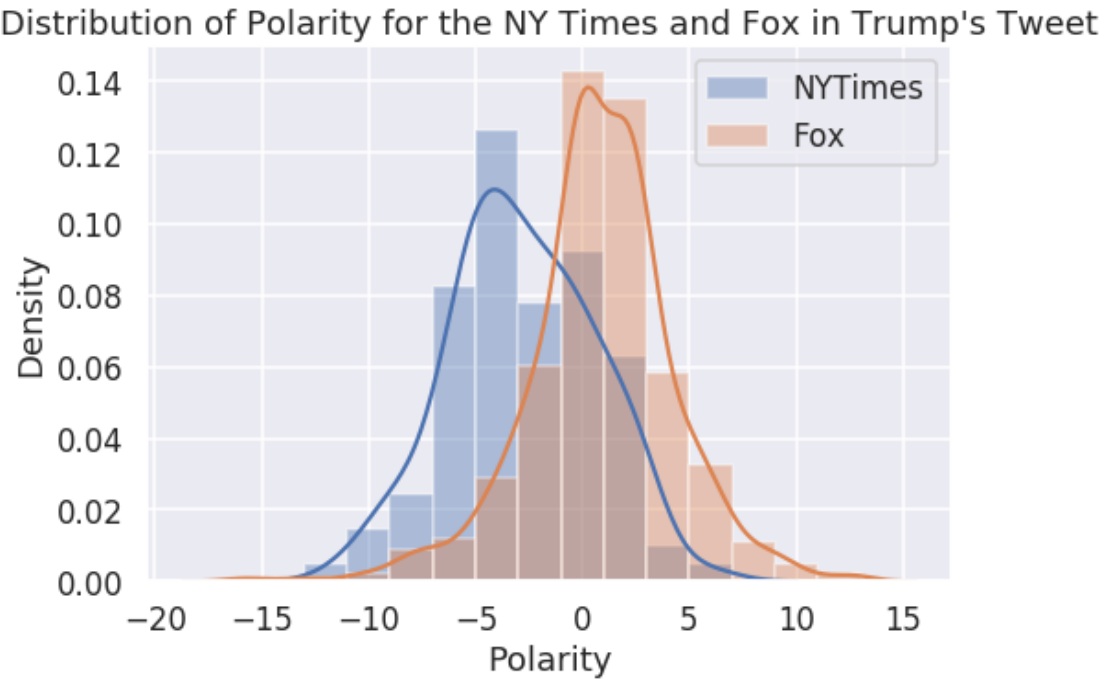
0.4.1 Question 6a

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 [108]: plt.figure(figsize = (7, 5))
          sns.distplot(trump[trump['text'].str.contains('nytimes')]['polarity'], label = 'NYTimes', bins
          sns.distplot(trump[trump['text'].str.contains('fox')]['polarity'], label = 'Fox', bins = np.a
          plt.title("Distribution of Polarity for the NY Times and Fox in Trump's Tweet")
          plt.xlabel('Polarity')
          plt.legend()
```

```
Out[108]: <matplotlib.legend.Legend at 0x7f8991332790>
```



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 distribution of polarity between the NY times and Fox are different. The polarity distribution of the NY times is centered around -5, while the polarity distribution of 'fox' is centered around 2. It appears trump's tweet about the NY times contains more negative polarity words, contrating with his tweet about Fox which contains more positive polarity words.

I looked for 'bernie' and 'ivanka' polarity distributions. The center of polarity distribution for 'bernie' is around -1, while for 'ivanka' is around 3. The words that trump uses on his twitter for 'bernie' are more likely to contain negative polarity words since the polarity distribution graph are skewed to the negative side, while the words used for 'ivanka' are more likely to have positive polarity words since the distribution is largely positive.

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

The polarity distribution of tweets without hastag or link appears approximately normal, centered around 0 with a larger variance of polarities than the distribution with hashtag or link. The polarity distribution of tweets with hastag or link looks like a right-skewed distribution, centered around 0 as well and have a smaller variance of polarities.