0.1 Question 0

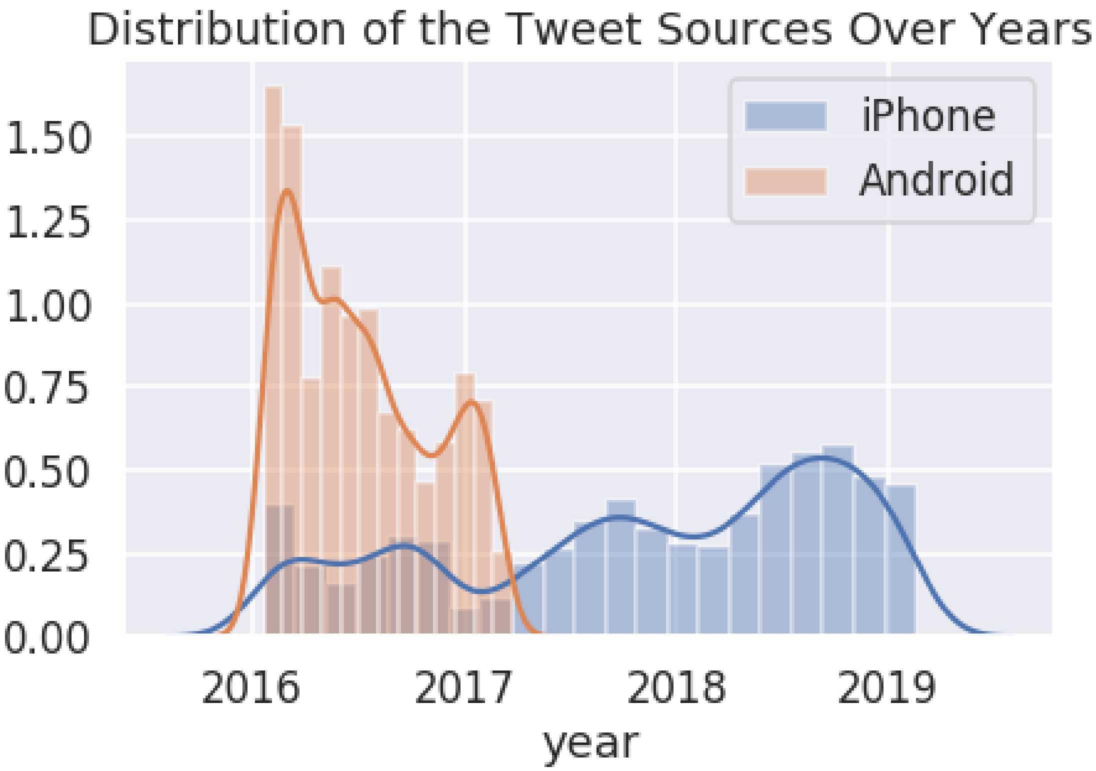
There are many ways we could choose to read the President's tweets. Why might someone be interested in doing data analysis on the President's tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's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

An institution like a newspaper or a journal might be interested in Trump's tweets. They could use these tweets to either fact check him, or try to analyze the impact of his tweets on the stock market.

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]:
```
and_col = trump.loc[trump["source"]=="Twitter for Android"]
iph_col = trump.loc[trump["source"]=="Twitter for iPhone"]

sns.set_style("darkgrid")
sns.distplot(iph_col[["year"]], label="iPhone")
sns.distplot(and_col[["year"]], label="Android")
plt.xlabel("year")
plt.ylabel("")
plt.legend(loc="upper right")
plt.title("Distribution of the Tweet Sources Over Years");
```

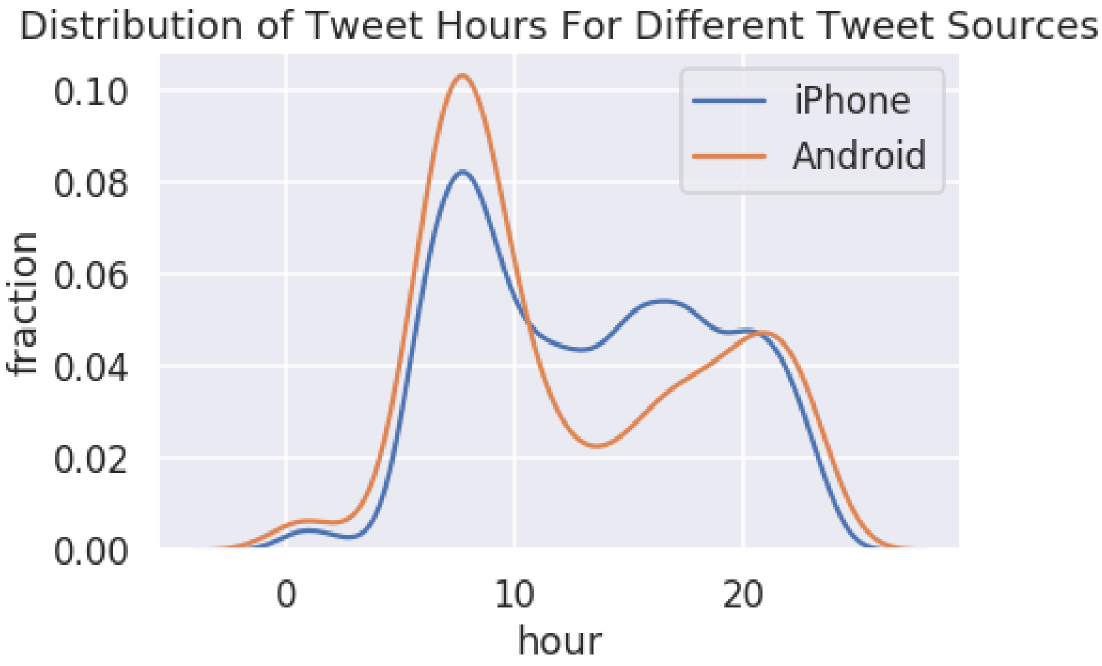


0.1.1 Question 4b

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 [19]: ### make your plot here
         and_col = trump.loc[trump["source"]=="Twitter for Android"]
         iph_col = trump.loc[trump["source"]=="Twitter for iPhone"]

         sns.set_style("darkgrid")
         sns.distplot(iph_col[["hour"]], label="iPhone", hist=False)
         sns.distplot(and_col[["hour"]], label="Android", hist=False)
         plt.xlabel("hour")
         plt.ylabel("fraction")
         plt.legend(loc="upper right")
         plt.title("Distribution of Tweet Hours For Different Tweet Sources");
```



0.1.2 Question 4c

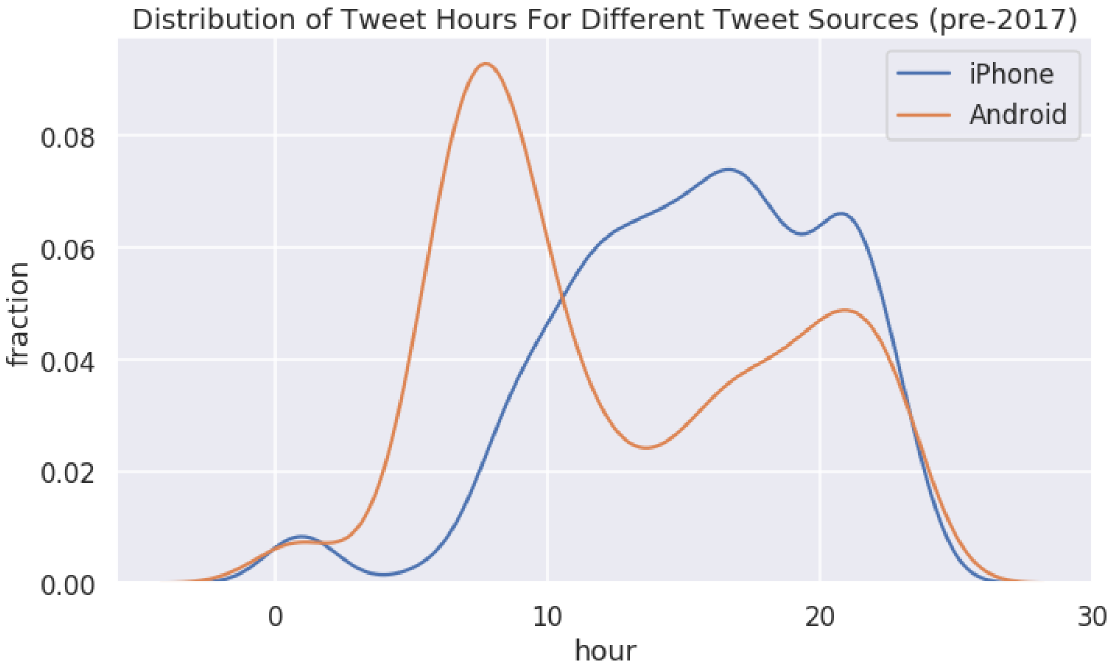
According to this Verge article, Donald Trump switched from an Android to an iPhone sometime in March 2017.

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 [100]: ### make your plot here
          n_trump = trump[trump["year"]<2017]

          and_col = n_trump.loc[n_trump["source"]=="Twitter for Android"]
          iph_col = n_trump.loc[n_trump["source"]=="Twitter for iPhone"]

          sns.set_style("darkgrid")
          plt.figure(figsize=(10,6))
          sns.distplot(iph_col[["hour"]], label="iPhone", hist=False)
          sns.distplot(and_col[["hour"]], label="Android", hist=False)
          plt.xlabel("hour")
          plt.ylabel("fraction")
          plt.xticks([0, 10,20,30])
          plt.legend(loc="upper right")
          plt.title("Distribution of Tweet Hours For Different Tweet Sources (pre-2017)");
```



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 peak hours for android and iPhone are different and so are the distributions. The peak in distribution for Android right before 10 am which could suggest a routine of tweeting right before starting 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?

I would give it a score of -2 due to the use of words such as police, order, gun which could suggest violence. On the other hand, if we focus on words such as dog, TikTok could give it a more neutral score of 1.

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?

Type your answer here, replacing this text.

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 they have been accurately classified.

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`.

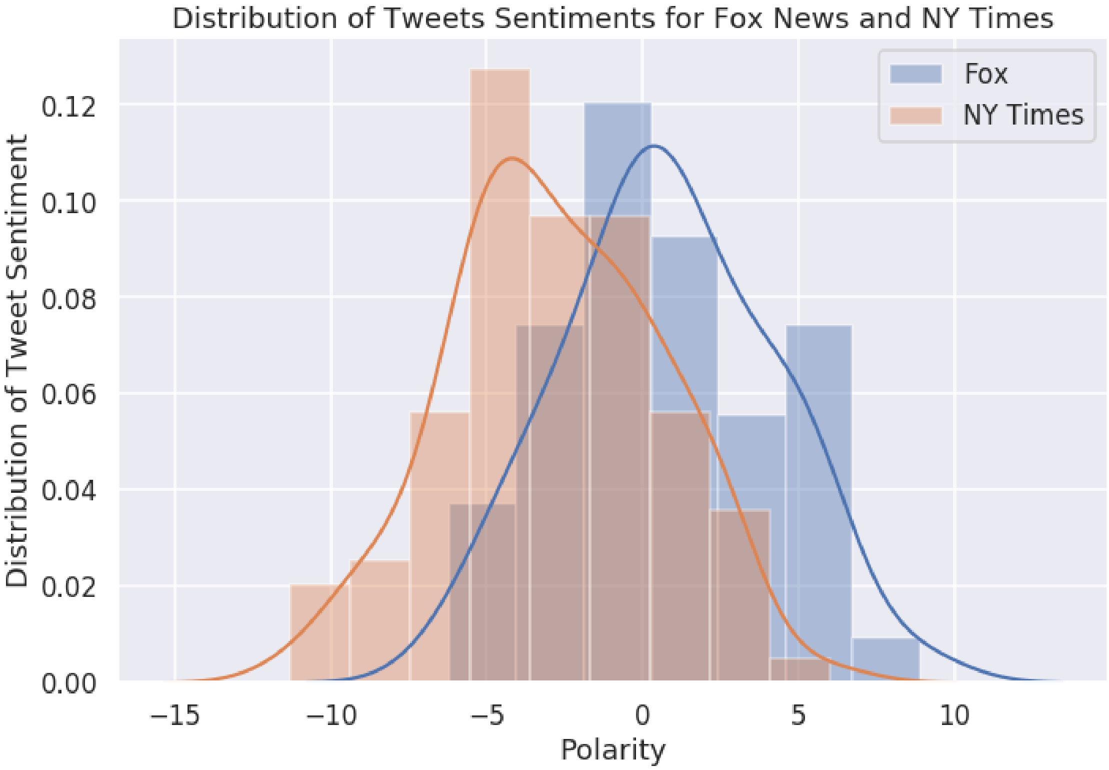
```
In [ ]: n_trump = trump[trump["year"]<2017]

        and_col = n_trump.loc[n_trump["source"]=="Twitter for Android"]
        iph_col = n_trump.loc[n_trump["source"]=="Twitter for iPhone"]

        sns.set_style("darkgrid")
        plt.figure(figsize=(10,6))
        sns.distplot(iph_col[["hour"]], label="iPhone", hist=False)
        sns.distplot(and_col[["hour"]], label="Android", hist=False)
        plt.xlabel("hour")
        plt.ylabel("fraction")
        plt.xticks([0, 10,20,30])
        plt.legend(loc="upper right")
        plt.title("Distribution of Tweet Hours For Different Tweet Sources (pre-2017)");
```

```
In [46]: ny_ind = tidy_format[tidy_format["word"]=="nytimes"]
         fox_ind = tidy_format[tidy_format["word"]=="fox"]
         trump_nytime = trump[trump.index.isin(ny_ind.index)]
         trump_fox = trump[trump.index.isin(fox_ind.index)]

         sns.set_style("darkgrid")
         plt.figure(figsize=(10,7))
         sns.distplot(trump_fox[["polarity"]], label="Fox")
         sns.distplot(trump_nytime[["polarity"]], label="NY Times")
         plt.xlabel("Polarity")
         plt.ylabel("Distribution of Tweet Sentiment")
         plt.title("Distribution of Tweets Sentiments containing Fox News and NY Times")
         plt.legend(loc="upper right");
```


Distribution of Tweets Sentiments for Fox News and NY Times
Fox
NY Times
Distribution of Tweet Sentiment
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−15
−10
−5
0
5
10
Polarity

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 nytimes sentiments distribution is somewhat symmetric and bellshaped, with a center around -5 and the distribution is shifted to the left compared to the tweets containing fox.

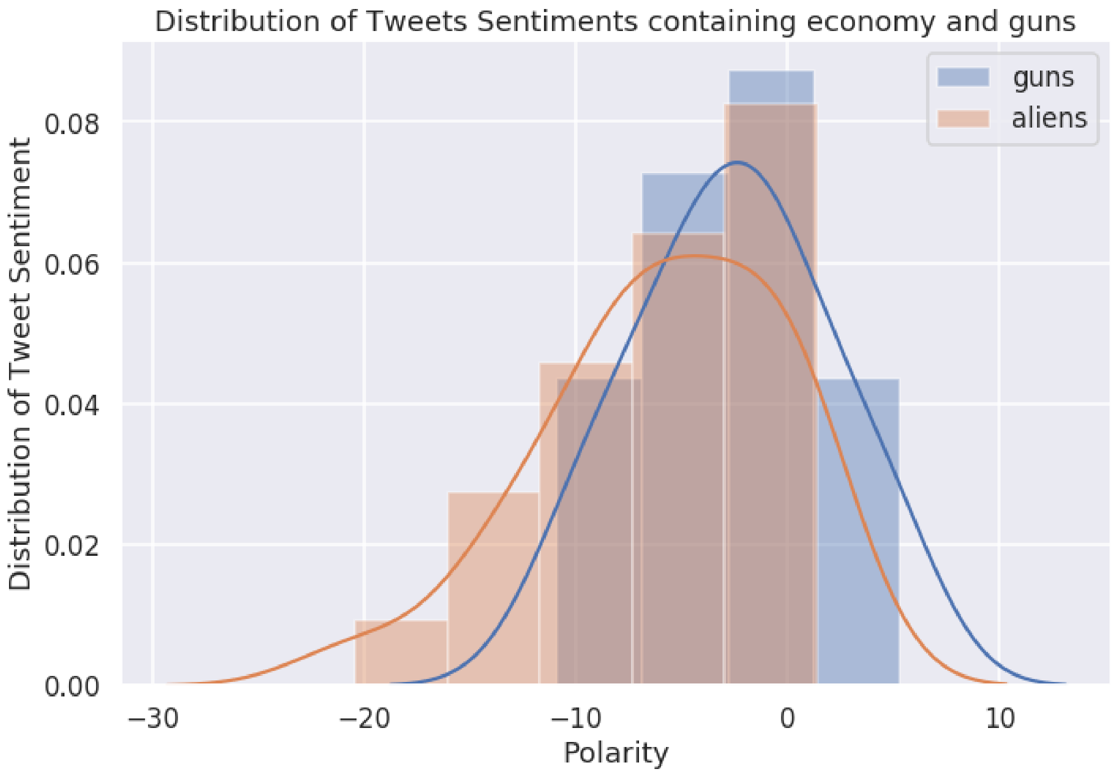
The fox sentiments distribution is also fairly symmetric and bell-shaped with a center around 1. The tweets have a greater number of positive polarity value.

The two key words sentiments' distributions overlapped between polarity of -6 to 6.

The two words that can also be compared are guns and aliens. It was interesting to see that the distribution of the polarity of the the tweets containing ”alien” is more negative compared to the word ”guns” which is a word with strong negative sentiment. The distribution for aliens is centered around -3 and the one for guns is centered at -1. There is also a lot of overlap.

```
In [82]: gun_ind = tidy_format[tidy_format["word"]=="gun"]
         economy_ind = tidy_format[tidy_format["word"]=="aliens"]
         trump_gun = trump[trump.index.isin(gun_ind.index)]
         trump_economy = trump[trump.index.isin(economy_ind.index)]

         sns.set_style("darkgrid")
         plt.figure(figsize=(10,7))
         sns.distplot(trump_gun[["polarity"]], label="guns")
         sns.distplot(trump_economy[["polarity"]], label="aliens")
         plt.xlabel("Polarity")
         plt.ylabel("Distribution of Tweet Sentiment")
         plt.title("Distribution of Tweets Sentiments containing economy and guns")
         plt.legend(loc="upper right");
```


Distribution of Tweets Sentiments containing economy and guns
Distribution of Tweet Sentiment
0.08
0.06
0.04
0.02
0.00
−30
−20
−10
0
10
Polarity
guns
aliens

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

1. For the tweets with a hashtag or link,the distribtution of polarity of the tweets is roughly bell-shaped.The peak is higher than that for tweets without a hasktag or link. It is centered around 0.
2. For the tweets with a hashtag or link,the distribtution of polarity of the tweets is roughly bell-shaped and symmetric. It is centered at 0.