Notebook

October 17, 2019

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.

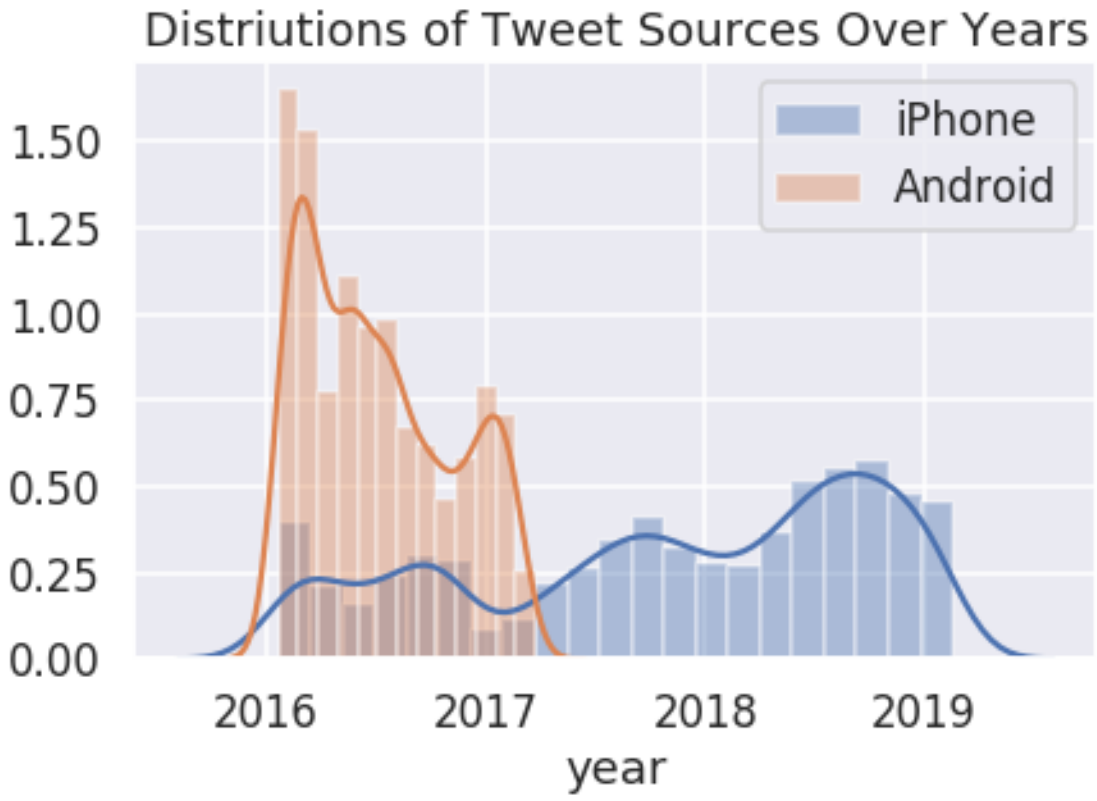
One entity that would be interested in analyzing President Trump's tweets is the financial industry and wall street. Because politics have an effect on the markets, the President's tweets could very well effect the movement of a stock. For example, JP Morgan created a Volfefe Index to track Trump's tweets.

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 [166]: iphone = trump['year'][trump['source']=='Twitter for iPhone']
          android = trump['year'][trump['source']=='Twitter for Android']

          sns.distplot(iphone, label='iPhone')
          sns.distplot(android,  label='Android')
          plt.legend()
          plt.title("Distriutions of Tweet Sources Over Years")
          plt.xlabel("year")

Out[166]: Text(0.5, 0, 'year')
```



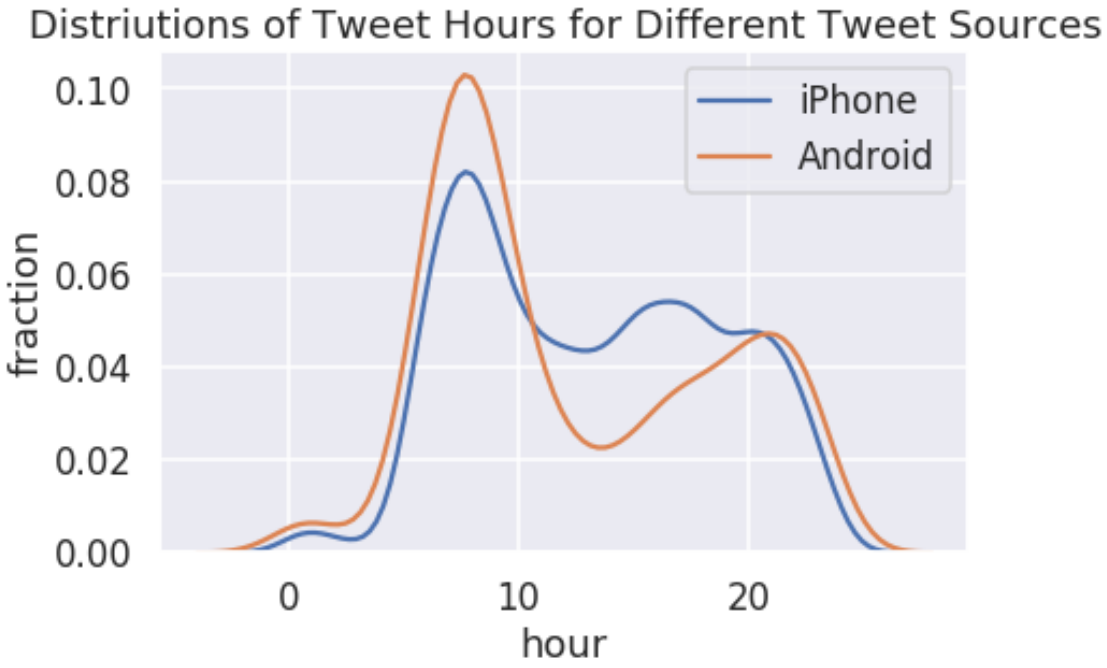
0.1.1 Question 4b

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 [171]: iphone_hour = trump['hour'][trump['source']=='Twitter for iPhone']
          android_hour = trump['hour'][trump['source']=='Twitter for Android']

          sns.distplot(iphone_hour, label='iPhone',hist=False)
          sns.distplot(android_hour,  label='Android',hist=False)
          plt.legend()
          plt.title("Distriutions of Tweet Hours for Different Tweet Sources")
          plt.xlabel("hour")
          plt.ylabel("fraction")

Out[171]: Text(0, 0.5, 'fraction')
```



0.1.2 Question 4c

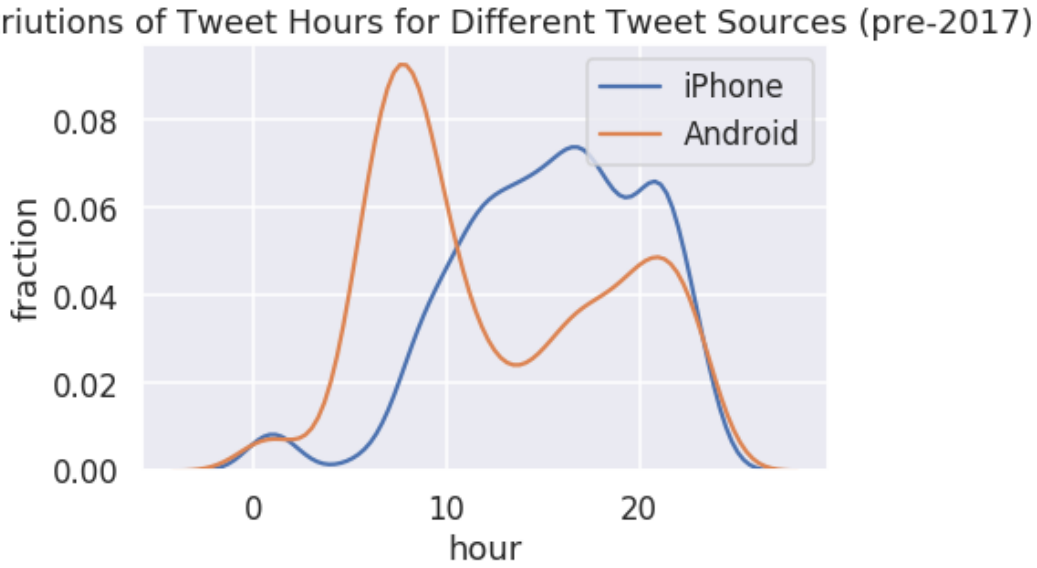
According to this Verge article, Donald Trump switched from an Android to an iPhone sometime in March 2017.

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 [172]: pre2017 = trump[trump['year']<2017]
          iphone_hour_pre2017 = pre2017['hour'][trump['source']=='Twitter for iPhone']
          android_hour_pre2017 = pre2017['hour'][trump['source']=='Twitter for Android']

          sns.distplot(iphone_hour_pre2017, label='iPhone',hist=False)
          sns.distplot(android_hour_pre2017,  label='Android',hist=False)
          plt.legend()
          plt.title("Distriutions of Tweet Hours for Different Tweet Sources (pre-2017)")
          plt.xlabel("hour")
          plt.ylabel("fraction")

Out[172]: Text(0, 0.5, 'fraction')
```



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?

Because Trump switched to an iPhone in March 2017, the figure above shows Trump's usage of the iPhone and Android pre March 2017. If we look at the 20th hour, it shows that Trump more frequently tweets on his iPhone than on his Android. Because Trump did not use an iPhone pre March 2017, it could only make sense that his staff is tweeting for him. Some addtional analysis that we could look at to help support this claim is looking at the usage of other devices pre 2017, and the fraction they make up. If the total fraction is relatively high for all the devices, then it is highly probable that his staff is tweeting 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.

One advantage of using VADER in the analysis is that it is well established and relatively accurate given that each word is reviewed and rated by at least 20 trained individuals, and that it utilizes a normalized weighted composite score. One disadvantge of using VADER is that it doesn't recognize sentences but instead words, and so it might misinterpret the meaning of the entire sentence. For example, it would attribute a score to sentences that don't even make sense, or might even give a positive score to phrases like "not good", which is negative.

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, for example it would attribute a score to sentences that don't even make sense, or give a positie score to phrases like "not good", which is negative.

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 believe these tweets are accurately represented by their polarity scores. I say this because all the 5 negative tweets are actually negative in nature, and all the 5 positive tweets are positive in nature.

0.4 Question 6

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

0.4.1 Question 6a

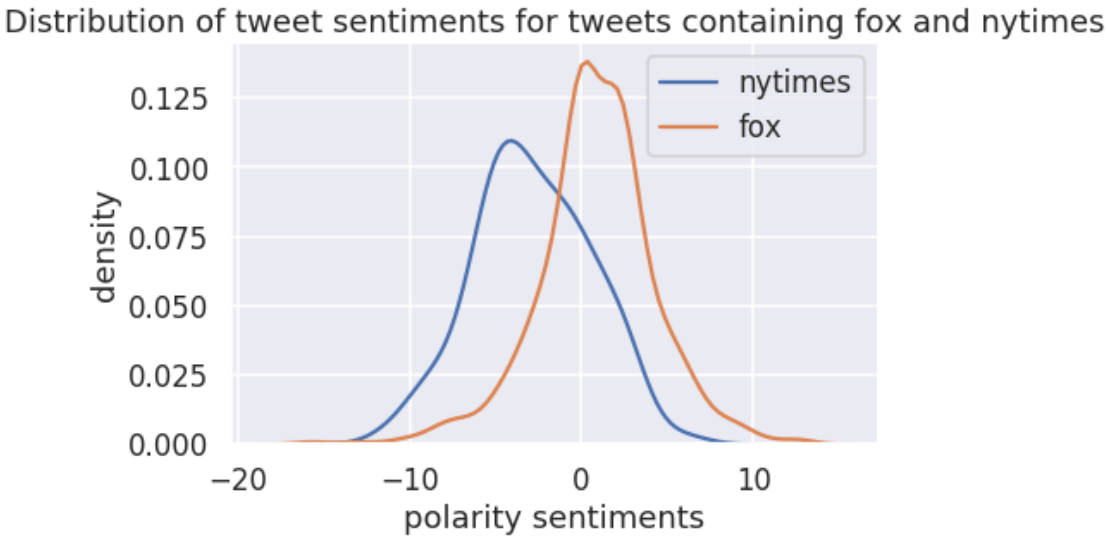
In the cell below, create a single plot showing both the distribution of tweet sentiments for tweets containing `nytimes`, as well as the distribution of tweet sentiments for tweets containing `fox`.

Be sure to label your axes and provide a title and legend. Your colors don't have to match ours, but you should use different colors for `fox` and `nytimes`.

In [299]:
```
nytimes = trump[trump['no_punc'].str.contains('nytimes')]
nytimes_polarity = nytimes['polarity']
fox = trump[trump['no_punc'].str.contains('fox')]
fox_polarity = fox['polarity']

sns.distplot(nytimes_polarity, label='nytimes',hist=False)
sns.distplot(fox_polarity,  label='fox',hist=False)
plt.legend()
plt.title("Distribution of tweet sentiments for tweets containing fox and nytimes")
plt.xlabel("polarity sentiments")
plt.ylabel("density")
```

Out[299]:
```
Text(0, 0.5, 'density')
```



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).

Looking at the plot above, we see that tweets that reference fox has a higher density of positive polarity sentiments compared to tweets that reference nytimes. This makes sense, because knowing trump, he lauds fox news and detests new york times. Similarly, when we change the keywords and compare the keywords democrats and republicans, tweets containing republics have a higher density of positive polarity sentiments compared to tweets containing nytimes.

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

Lookinga at the distributions, we see that tweets with a hashtag or link has the highest density when the polarity is 0. This means that tweets with hashtags or links are probably very neutral in nature. Whereas tweets without hashtags or links range from being very negative to very positive. I'd say this makes sense because tweets with hashtags and links tend to be less subjective and emotional, and more factual.