Date: 14 August 2016

Title: Data Story write-up

Author: Aakant Taurani

1) Pick a dataset - ideally the dataset for your Capstone. If for some reason you want to do this on a different data set, you can find one on Mode Analytics or Google's public data sets directory, or pick another one you like from elsewhere.

I mined the dataset I need for my capstone from Twitter via the Tweepy API. I obtained the most recent 3200 Tweets by 19 well-known publications including the Finanical Times and Forbes, resulting in approximately 60,000 tweets.

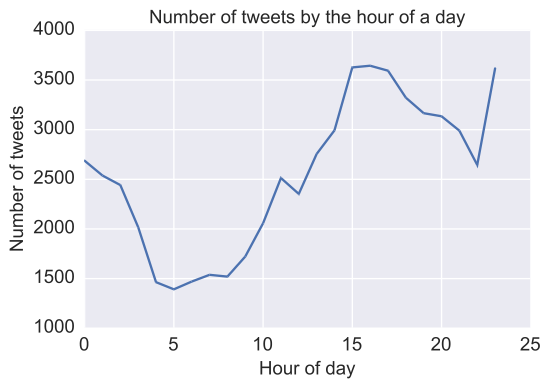
2) Get going by asking the following questions and looking for the answers with some code and plots:

a) Can you count something interesting?

I counted the number of tweets obtained for each publication and saw something interesting. The max limit of 3200 tweets for each user imposed by twitter is not a hard limit. There range of tweets received per publication ranged from 3184 tweets to 3248 tweets.

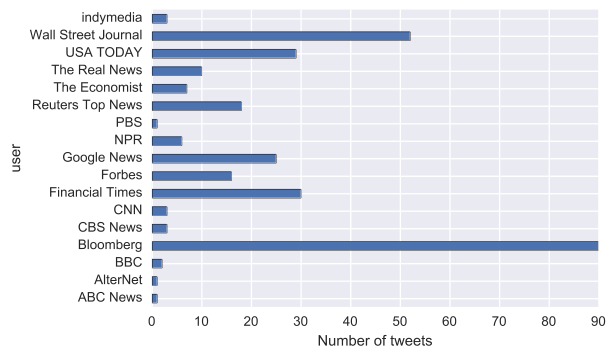
b) Can you find some trends (high, low, increase, decrease, anomalies)?

An interesting trend is that the number of tweets published is the lowest at around 5AM and increases until around 3-5PM.



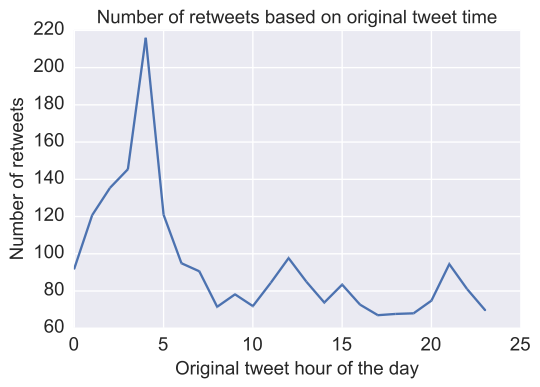
c) Can you make a bar plot or a histogram?

I created a horizontal bar plot showing the number of tweets that contained the word “stock” across each publication.



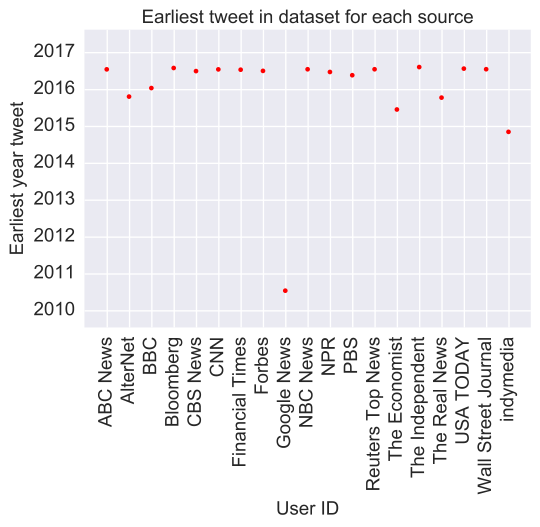
d) Can you compare two related quantities?

I compared the number of retweets vs. the time the original tweet was posted.

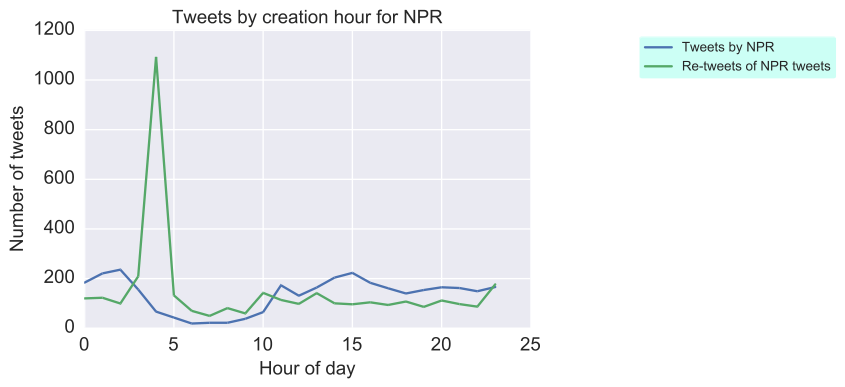


e) Can you make a scatterplot?

I created a scatterplot to see the earliest tweet in the dataset by publication.



f) Can you make a time-series plot?

I created a time-series plot for the number of tweets posted throughout the day

3) Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?

Some of the insights that I see are:

1) For NPR, CNN and CBS, there is a massive increase in the number of tweets between 4-6AM. Why is its so? Are the followers of these news sources awake by this time?

2) For some publications such as Bloomberg and Google News, the number of retweets is always less than the number of tweets. Why is this so? Why is it different from publications such as CNN where the retweet count is higher than the number of tweets?

4) By now you’ve asked a bunch of questions, and found some neat insights. Is there an interesting narrative, a way of presenting the insights using text and plots from the above, that tells a compelling story? As you work out this story, what are some other trends/relationships you think will make it more complete?

It is clear that different news sources engage differently with Twitter. Whereas some publications such as CNN write material that are highly re-tweeted, others such as Google News publish material that are barely re-tweeted. However, just because publications are highly re-tweeted, it doesn't apply that they deliver a lot of stock market information. For example, Bloomberg ranks 1st by a huge margin over other publications for having the most number of times the word 'stock' is in their tweets. However, their retweet counts is far less than the number of tweets posted. Perhaps, this suggests that not a lot of people are looking for financial information on Twitter.