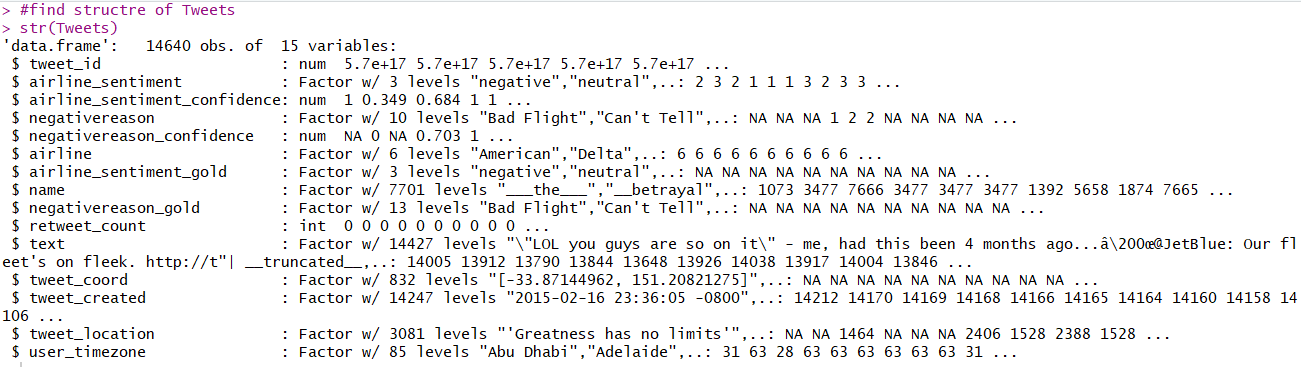
**Eric Vela**

**ABA Homework 1**

**Data Assessment and Findings**

Over the past few weeks, I was able to analyze the Twitter Airline Sentiments dataset found on Kaggle and found some quite interesting results. Using exclusively R, here is a brief outline of the variables measured in this dataset.



There were 14640 observations and 15 variables used in this dataset. However, for the sake of this analysis, only a few of the variables were used. After exploring this dataset, the most important variables used for analysis were:

***airline\_sentiment*** – shows negative, neutral, or positive depending on the user’s flight experience. Measured using a sentiment analysis model that determines the overall sentiment of the tweet given each individual word in the tweet.

***negativereason*** – the reason why the flight experience was not very good

***airline***– self-explanatory in this context

***name***– the username of the person who tweeted about the airline

***text***– the tweet/message posted on to the site

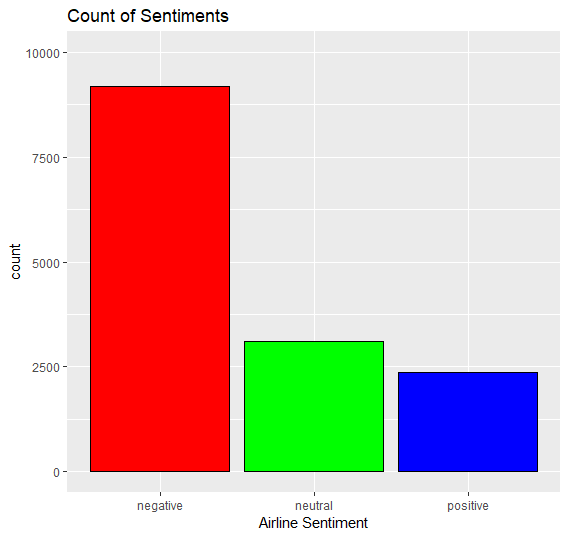
***user\_timezone*\*** - the relative location the tweet originated from (not the most useful measurement, not actually a time zone sometimes)

It is important to note that the exact algorithm used to classify the tweets was not outlined. However, confidence intervals were used to determine how likely the tweet was deemed negative, neutral, or positive, 0 being completely uncertain and 1 being completely certain. This same practice was outline for the ***negativereason\_confidence***.

Many subsequent variables were not useful for this analysis. Moreover, ***airline\_sentiment\_gold*** had 40 entries of the 14640 observations, meaning not even 1% of the data was populated for this variable. The ***negativereason\_gold*** variables was sparsely populated as well, with only 30 entries, and was not seen as useful for this analysis. Both of the prior variables were confusing to interpret and the purpose of them for this dataset was unclear. The variable ***retweet\_count*** had a maximum value of 44 retweets but 13873 of the observations (94.76%) had no retweets. The variable ***tweet\_coord*** had 832 unique geographical coordinates and was not very useful when analyzing where tweets came from. The variable ***tweet\_created*** could have been useful for this analysis if the tweets had been captured over a longer time period. The captured time frame started on February 16, 2015 around 11PM and ended on February 24, 2015 just before midnight. The variable ***tweet\_location*** offered some appropriate location information; however, it was littered with junk data such as “Lalaland,” “Punk is the preacher,” and “searching for coffee” to name a few entries. Moreover, it is inferred that this data was pulled from the user’s Twitter biographies where actual locations are not required. The ***user\_timezone*** variable was rarely an actual time zone and was only somewhat useful to this analysis.

**Dataset Analysis and Insight Report**

Sentiment analysis was the main approach (obviously) when looking at this dataset. Below is a bar chart plotting the sentiments by their frequency.



It is apparent that most of the sentiments toward flying is negative. This is likely caused by people’s natural tendencies to highlight negative experiences better than the positive ones, i.e. one is likely to remember a bad flight rather than a good one. The negative sentiment count is 9178 (62.69%), the neutral sentiment is 3099 (21.17%), and the positive sentiment is 2363 (16.14%). Moreover, below is a cross tabulation display of the Airline used when the customer reported their sentiment.

Tweets$airline\_sentiment

Tweets$airline negative neutral positive

American 1960 463 336

Delta 955 723 544

Southwest 1186 664 570

United 2633 697 492

US Airways 2263 381 269

Virgin America 181 171 152

As seen above, United has the most negative reviews with 2633, with US Airways and American not too far behind. Moreover, Southwest received the highest amount of positive tweets and United received most of the neutral tweets. Now we will analyze this same table, using percentages.

negative neutral positive

American 0.71040232 0.16781443 0.12178325

Delta 0.42979298 0.32538254 0.24482448

Southwest 0.49008264 0.27438017 0.23553719

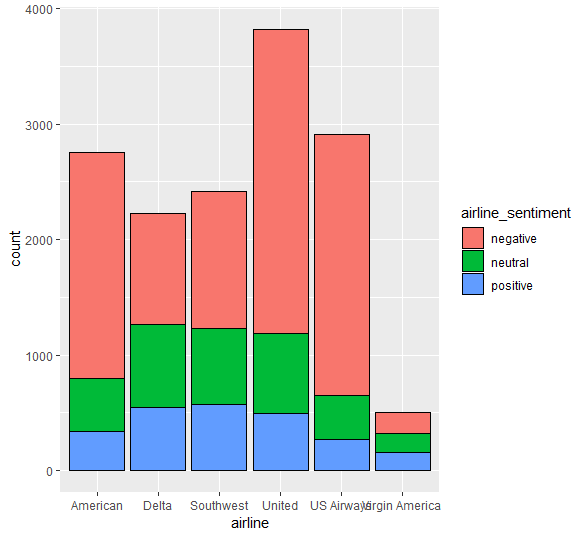
United 0.68890633 0.18236525 0.12872841

US Airways 0.77686234 0.13079300 0.09234466

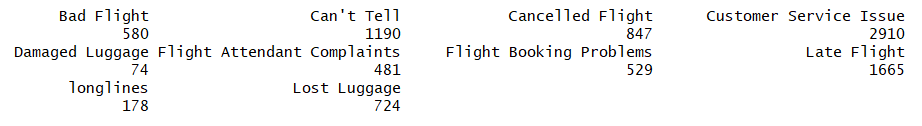
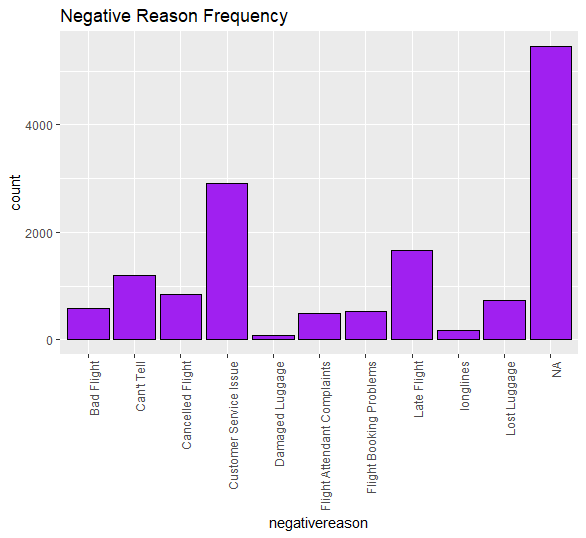
Virgin America 0.35912698 0.33928571 0.30158730

From this view, US Airways has the highest negative sentiment towards their flights by percentage with 77.68%. American has the second highest negative sentiment with 71.04%. Virgin America has the lowest amount of negative sentiment with 35.91%, highest neutral sentiment with 33.92%, and the highest positive sentiment with 30.16%. In the time frame of these tweets, Virgin America was the best airline to fly according to Twitter.

In case numbers are too inundating, here is a graph that reemphasizes my previous findings.

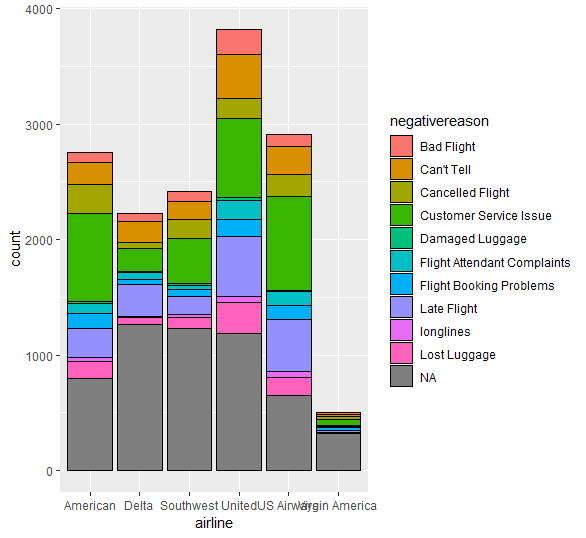


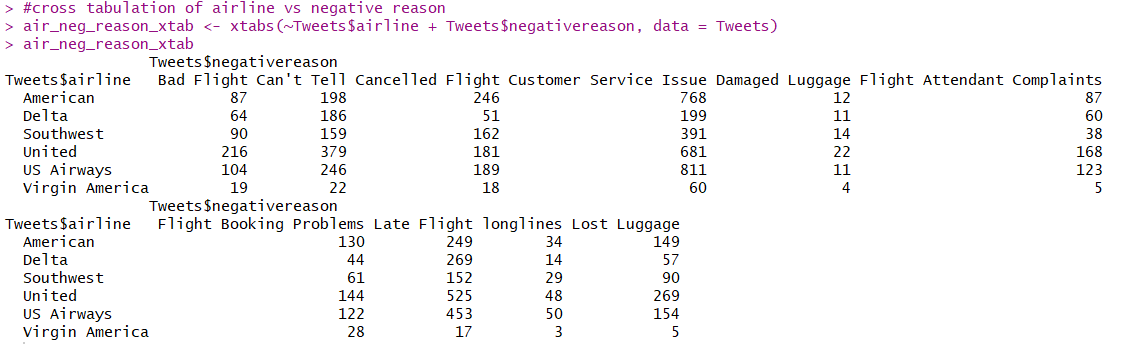
Furthermore, it is important to realize what problems Twitter users encountered that caused them to have a negative experience. Below is a bar chart describing this issue.



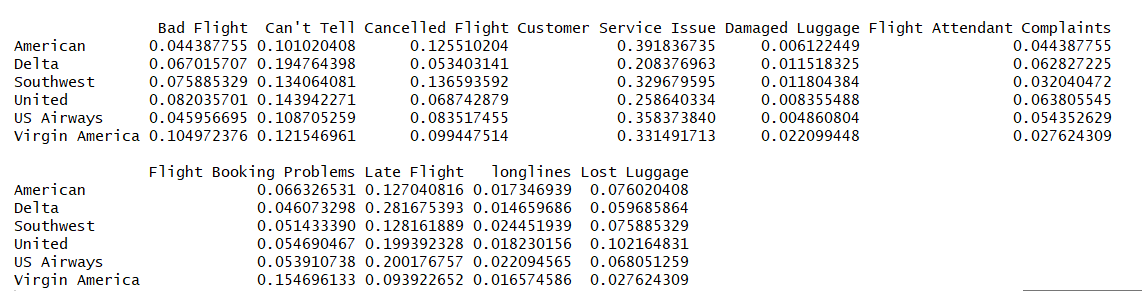
Most of the issues with flights came from a Customer Service Issue with 2910 of the 16460 observations or with a Late Flight (1665 issues). It is important to note that the fast majority of the issues could not be identified from the sentiment analysis algorithm (62.69% or 9178 observations).

Below is a chart and a cross-tabulation of the bad flight experiences per airline.



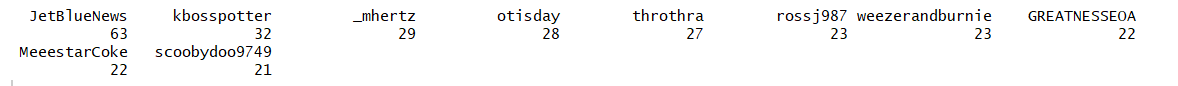


As seen above, US airways has the most Customer Service Issue with 811 which was the highest specific complaint across all airlines. The stacked chart offers a unique perspective on the data. As mentioned previously, Virgin America had the least amount of issues but also had substantially less tweets about their services. United Airways had the most Late Flights and the most amount of ambiguous issues. Below is the percentage view of this data similar to the previous issue.

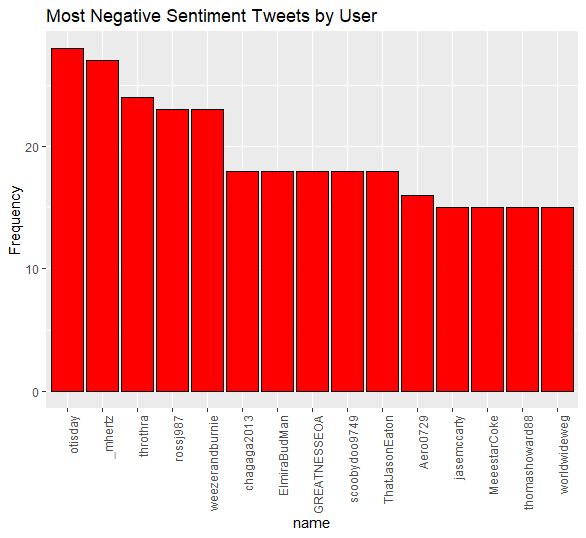


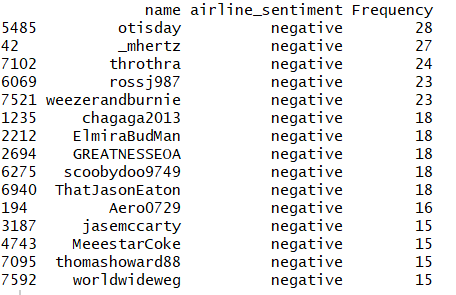
Most of every airline’s issues stem from a Customer Service Issue. For the sake of this analysis, I will continue moving forward.

In this time period, here are the most frequent Twitter users. If an airline company would like to receive more feedback on the issues encountered, these users appeared to be the most responsive.

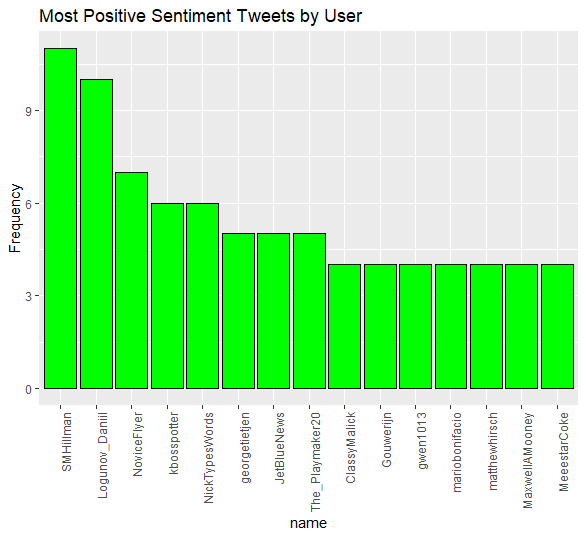


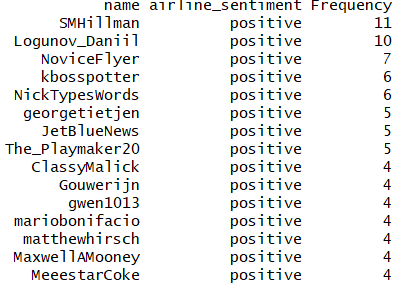
Moreover, here are the 15 biggest “haters,” or Twitter users with the most negative sentimental tweets.



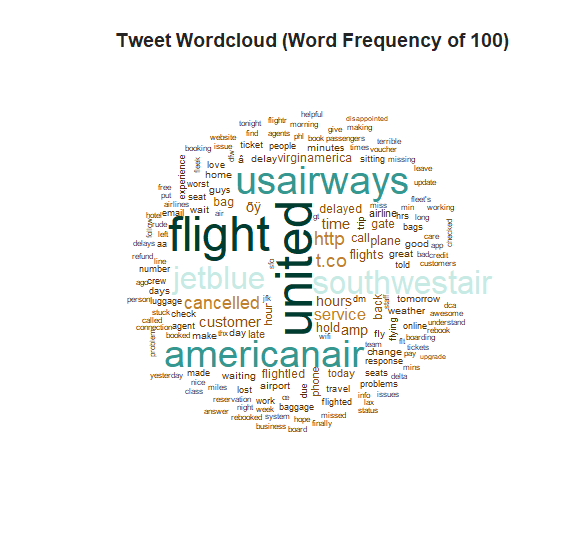


@otisday is the biggest hater with 28 negative tweets. This information would be useful for airlines to see the twitter users that are most likely to complain or experience an issue when flying.

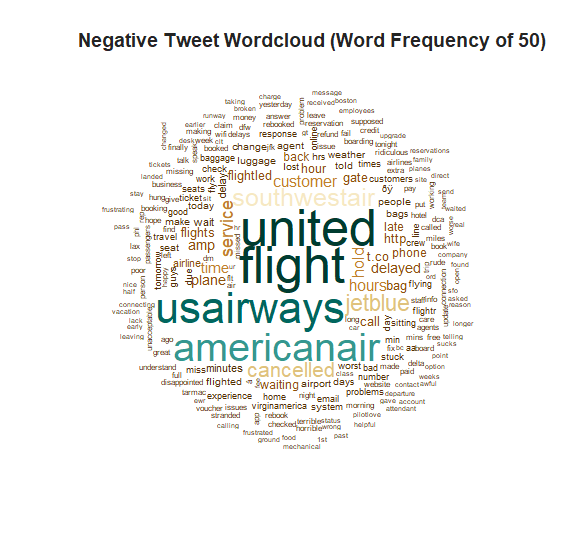




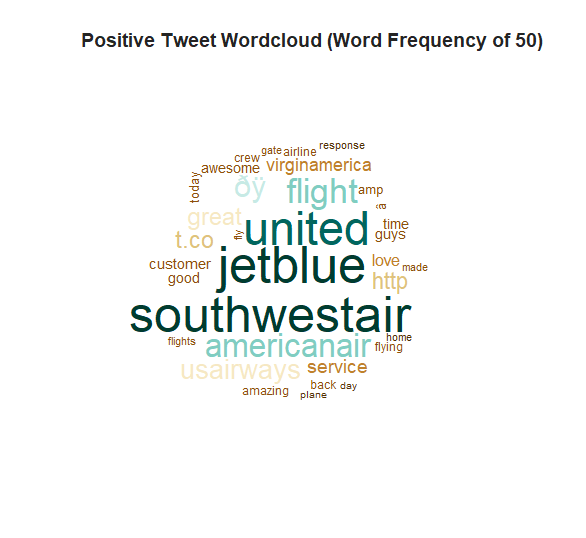
Now looking directly at frequency of words that occur in tweets, here are a few word clouds regarding all tweets, negative tweets, and positive tweets.



The most frequently occurring words are the airline companies because users tweet at their accounts using the “@” symbol. This algorithm has removed symbols which is why they appear more frequently. Words like cancelled, service, customer, and time also occur quite frequently.



The negative word cloud is quite similar to the previous version with all tweets included. This occurs because an overwhelming majority of the tweets were negative. However, words like delayed, worst, wait, lost, hold, etc. occur more frequently.



The positive tweet word cloud was slightly underwhelming considering most tweets were negative. However, words like love, good, amazing, etc. occur here more often.

Overall, this data was useful to some extent and was an interesting application to research and better understand how sentiment analysis works. The graphs and exploratory data analysis provided a data profile to better understand what data is being provided from Twitter.

**Action Steps Taken**

1. Downloaded data set and loaded into R
2. Explored structure and utilized summary function to understand data set
3. Identified usable variables from the list of variables
4. Created graphs and charts relevant for data profiling
5. Created wordclouds to better understand the frequency of words used in each airline sentiment
6. Brainstormed building some sort of predictive model on this data
7. Attempted to create a Support Vector Machine and my own Sentiment Analysis algorithm to no avail (7 hours later)