Matthew Pascual-Mead

10/7/20

MGSC-410

Homework 1

**Project Steps**

To analyze this dataset, I chose to use Jupyter Notebook, as Python is the language I feel most comfortable using when working with data. Jupyter Notebook not only allowed me to analyze this dataset more efficiently, but it also offered libraries that make graphical representations both visually appealing and easy to understand.

Going into this assignment, I knew that I would be working with a messy dataset. As a result, I started off by getting a clear grasp of what the dataset consisted of, and what data needed to be cleaned. Based on the description on Kaggle, the dataset contained a collection of tweets from February 2015 about six major U.S. airlines. The dataset particularly focused on the sentiment of the tweets about the airlines, labeling each tweet with a distinct sentiment: negative, neutral, and positive. With a general idea of the data I was working with, I decided to load it into Jupyter Notebook to see how the data was structured. I first started by using the .info() function to see how the dataset was designed. The dataset consisted of a total of 14,640 observations (tweets), with a total of 15 variables. Out of the 15 variables, only four of them were quantitative(two floats and two integers). It was somewhat worrisome having such few numerical variables, as I had the suspicion that the categorical variables would require a lot of cleaning. But before I jumped to this conclusion, I wanted to see what the actual data looked like. From visualizing the first five rows, I was immediately greeted with numerous missing values. Now knowing that there were missing values in my dataset, I used the .isna().sum function to see the total count of missing values for each column. This showed me that there were missing values in 7 columns: “negativereason”, “negativereason\_confidence”, “airline\_sentiment\_gold”, “negativereason\_gold”, “tweet\_coord”, “tweet\_location”, and “user\_timezone”. I decided to remove the columns “airline\_sentiment\_gold”, “negativereason\_gold”, and “tweet\_coord” as the majority of their values were missing (over 90%). In addition, I did not feel that their non-missing values would offer too much useful information. I did choose to keep the remaining four variables with missing values as I felt that they had enough of their data intact for future usage.

Now that the dataset was a little more clean, I was ready to begin exploring and analyzing the data. Since the objective of this assignment was to analyze the sentiment of tweets about airlines, I directed most of my analysis toward two variables: “airline” and “airline\_sentiment”. To begin, I first wanted to visualize the number of tweets for each airline. To do this, I made a bar chart using the matplotlib.pylot library, where the x-axis represented the six airlines (United, US Airways, American, Southwest, Delta, and Virgin America), and the y-axis represented the number of tweets for each airline. From this visualization, it was clear that the number of tweets for each airline were not evenly distributed. The airlines with the leading number of tweets included United, US Airways, and American. On the other hand, Southwest, Delta, and Virgin America had the fewest number of tweets. To further exemplify the contrast in the number of tweets for each airline, I used the seaborn library to create a pie chart representing the proportion of tweets for each airline. The results were as follows: United – 26.11%, US Airways – 19.9%, American – 18.85%, Southwest – 16.53%, Delta – 15.18%, Virgin America – 3.44%. This representation was slightly better than the first bar chart, as it included the percentage of tweets for each airline. With prior knowledge about the negative reputation of some of the airlines (United in particular), I had a feeling that this imbalance might have had an underlying reason. However, I needed to do much more analysis before I could make any judgments.

As a result, I pivoted my attention toward the “airline\_sentiment” variable. I took the same approach as I did when looking at the airlines, and made a bar chart counting the number of tweets with each sentiment. The x-axis of the chart represented the sentiments, while the y-axis represented the number of tweets. In addition, I made a pie chart to show the proportion of sentiments for tweets. The results were as follows: Negative - 62.9%, Neutral - 21.17%, Positive - 16.4%. Both of these visualization showed that this dataset was dominated by tweets with a negative sentiment. From this, it seemed like more tweets for an airline might not be beneficial.

With knowledge of both the number of tweets for each airline and the proportion of sentiments for tweets, I now wanted to explore the frequency of sentiments for each airline. I decided to make bar charts for each airline, with the x-axis representing the sentiments of the tweet and the y-axis representing the frequency of sentiments for tweets. When comparing the bar charts of each airline, the three airlines with the highest number of tweets (United, US Airways, and American) also had highest number of tweets with a negative sentiment. The logical explanation for this might be that these airlines have more tweets with a negative sentiment simply because they have an overall higher number of tweets. So as a result, I needed to compare the proportions of tweet sentiment for each airline. By using a pie chart, I was able to make a representation showing the proportion of sentiments of tweets for each airline. The results showed that the three airlines with the most tweets also had the highest proportion of tweets with a negative sentiment. In contrast, the airlines with the fewest number of tweets also had the lowest proportion of tweets with a negative sentiment. This clarifies that airlines that are being tweeted about the most are not necessarily benefitting from it.

Now that I had learned that the three most popularly tweeted about airlines had the highest proportion of tweets with a negative sentiment, I wanted to see what the negative reasons were for the tweets. For this, I transitioned to using the “negativereason” variable, which stated the reason why a tweet had a negative sentiment. This variable did have missing values, but this was because some tweets were neutral or positive, meaning that they did not require a negative reason. As a result, I did not have to worry about fixing these missing values. I first used the .value\_counts() function to see the count of each negative reason across all negative tweets. The most popular was “Customer Service Issues”, being the negative reason for 2,910 tweets . At this point, I wanted to make a heatmap of the negative reason of tweets for each airline, but I did not know how to make it in Python. As a result, I transitioned to Tableau, which allowed me to make the heatmap with ease. The heatmap represented a count of each negative reason for a tweet for each airline, with the color darkness of each cell reflecting its frequency. The heatmap showed that for all but Delta, customer service is the number one area that needs improvement according to tweets.

Upon completion of my analysis, I learned that an airline having more tweets is not always advantageous. Based on my visualizations, the airlines with the highest number of tweets also had the highest proportion of tweets with a negative sentiment. I also learned that the leading reason for negative tweets about airlines is customer service issues. With this knowledge in mind, I think that the airlines should focus on improving their customer service. I think that if United, US Airways, and American were able to improve their customer service, it would significantly cut down on their frequency and proportion of negative tweets. In all, I feel that I uncovered some great insights that opened up potential for future analysis. For example, I would love to explore the “retweet\_count” variable to see if there is some sort of relationship with the number of retweets and sentiment. At some point, I would also like to try and implement a predictive model. I was not exactly sure how to make an accurate model with the variables that were provided, but I think that figuring this out down the line would be a great way to develop my machine learning skills.