Action List of Steps

**Data Cleaning:**

I called the dataset and specific columns I wanted to use to create a new dataframe

*tweets\_cleaned <- tweets%>%select(c(airline, airline\_sentiment, airline\_sentiment\_confidence, negativereason, negativereason\_confidence, retweet\_count, text))*

**Mean of Airline Sentiment Confidence for each airline sentiment per airline:**

I dropped off the N/A’s from the airline\_sentiment\_confidence, then created three dataframe based on the airline sentiment whether it was positive, negative or neutral. Using the three dataframes, I found the mean positive, negative and neutral sentiment for each airline.

*pos\_ASC <- airlinesentiment\_confidence\_na %>% filter(airline\_sentiment == "positive")*

*mean\_pos\_airlinesentiment\_confidence\_na\_per\_airline <- pos\_ASC %>% group\_by(airline) %>% summarize\_at(vars(airline\_sentiment\_confidence),list(mean\_pos\_airlinesentiment\_confidence\_na\_per\_airline = mean))*

*mean\_pos\_airlinesentiment\_confidence\_na\_per\_airline %>% arrange(desc(mean\_pos\_airlinesentiment\_confidence\_na\_per\_airline))*

**Mean and Median Negative Reason Confidence for each airline:**

Since the airline sentiment equaled “negative” were the only rows that I had a negative reason, the tweets that had positive or neutral sentiment were removed (their Negative Reason = N/A). Before I removed those specific rows, there were 14640 observations and after removing them, there were 10522 observations.Then I created two dataframes, one for mean and the other for median, based on the negative reason confidence for each airline.

*negreason\_confidence\_na <- tweets\_cleaned %>% drop\_na(negativereason\_confidence)*

*mean\_negreason\_confidence\_per\_airline <- negreason\_confidence\_na %>% group\_by(airline) %>% summarize\_at(vars(negativereason\_confidence), list(mean\_negreason\_confidence\_per\_airline = mean))*

*mean\_negreason\_confidence\_per\_airline %>% arrange(desc(mean\_negreason\_confidence\_per\_airline))*

**Which airlines do consumers tweet most about?**

To find out how tweets were about each airline, I plotted a bar graph with airlines on the X-axis, and the Tweet Count on the y-axis.

*ggplot(tweets\_cleaned) + aes(airline,fill= airline) + geom\_bar() +geom\_text(stat='count', aes(label=..count..), vjust=1.6, color="black") + labs(title= 'Number of Tweets per Airline', fill="Airlines") + theme\_minimal()*

**What are the main reason’s negative tweets? Which airline has the worst rating for each negative reason?**

To understand the main reason for the negative tweets, I only used rows where the airline sentiment equaled “negative”. The tweets that had positive or neutral sentiment were removed (their Negative Reason = N/A). Then I changed the negative reason “longlines” to “Long Lines”, I did not like how that string looked before.

I graphed two different graphs. The first one was a bar graph that depicted the number of tweets per negative reason. And finally, I graphed a bar graph using the negative airline sentiment dataframe. Airline was on the x-axis and the count of tweets were on the y-axis. And to understand each of the airline's main negative reasons, I facet-wrapped negative reasons. The second graph was to understand each airline's main negative reasons.

*ggplot(tweets\_removed\_noreason) +*

*aes(x= airline, fill=airline) + facet\_wrap(~negativereason) + geom\_bar() + xlab("Airline")+ ylab("Count")+ggtitle("Count of Tweets per Negative Reason per Airline")+labs(fill="Negative Reason") + theme(axis.text.x = element\_text(angle = 0, size=0))*

**What is the overall positive, negative and neutral sentiment per airline? What are the top three negative reasons why for each airline?**

I plotted two graphs again. The first graph is to get an understanding of the number of tweets for each airline sentiment per airline graph. I used this graph as a baseline for my second graph. The second graph was to understand the negative reason for each airline based on airline sentiment. The positive and neutral airline sentiment did not have any negative reasons, however the negative tweets did have negative reasons. Based on the number of tweets about a specific negative reason for each airline, I was able to figure out the top three negative reasons for each airline. This showed how each airline could improve. I used a dataframe that had all three airline sentiments, not just the negative sentiments to the analysis for both these graphs.

*ggplot(tweets\_na\_cleaned, aes(x=airline\_sentiment, fill=negativereason))+ facet\_wrap(~airline)+*

*geom\_bar()+xlab("Airline Sentiment")+ ylab("Count")+ggtitle("Count of Tweet Sentiment per Airline and the Negative Reason")+ theme(axis.text.x = element\_text(angle = 90, size=10)) +labs(fill = "Negative Reason")*

**Was there a specific negative reason why the airline had a negative sentiment based on retweets?**

I wanted to know if multiple customers felt the same way as someone who tweeted, but decided to retweet and not tweet at the airline directly. I only used rows where the airline sentiment equaled “negative”. The tweets that had positive or neutral sentiment were removed (their Negative Reason = N/A). Based on the number of retweets, I could get a better understanding of the specific negative reasons across all airlines, as well as among each airline.

*ggplot(tweets\_removed\_noreason, aes(x=retweet\_count, fill=negativereason))+ facet\_wrap(~airline)+*

*geom\_bar()+xlab("Number of Retweets")+ ylab("Count")+ggtitle( "Count of Retweets per the Negative Reason from the original tweet and the Airline")+ theme(axis.text.x = element\_text(angle = 90, size=6)) +xlim(0,45)+ylim(0,200) +labs(fill="Negative Reason")*

**Does the length of the text show anything about the Sentiment?**

For the final graph, I wanted to understand if the text length correlated with the airline sentiment. First I needed to count the number of characters per tweet. I added a new to column to the dataframe (the original cleaned dataframe) called textlength. Then I graphed the the text length based on the airline sentiment using the geom\_density function.

*ggplot(tweets\_cleaned, aes(x = text\_length,*

*fill = airline\_sentiment)) +*

*geom\_density(alpha = 0.5) + xlab("Text Length") + ggtitle("Text Lengths with Sentiment Distribution") + theme(text = element\_text(size=10)) + theme\_minimal()*