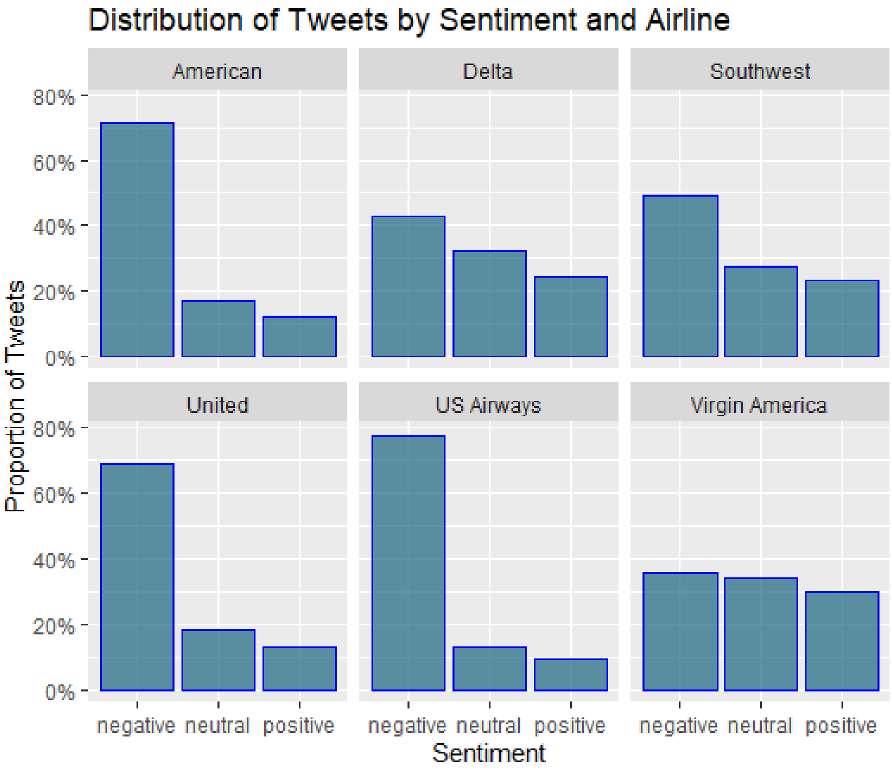
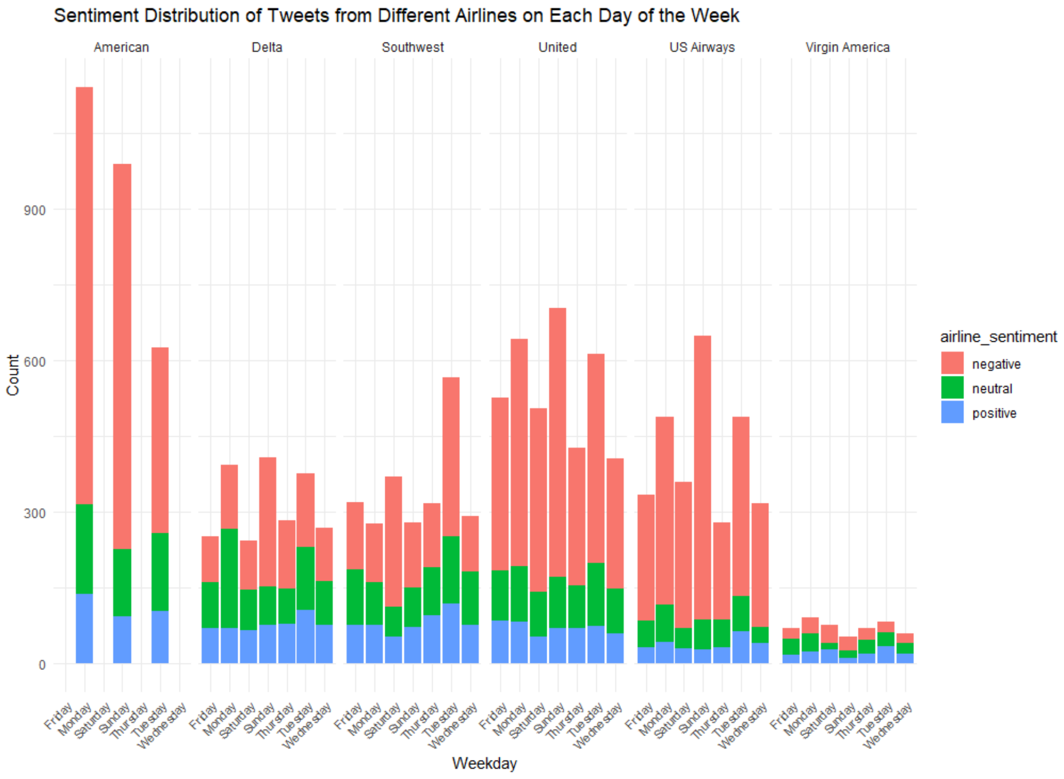
Insight Report

The first two questions are targeted towards better understanding the data and to dig for statistics that were of interest. This graph shows the distribution of sentiment in percentages for each airline. We can see that for every single airline, negative sentiment is the majority. When a person tweets about an airline, this is usually to complain or highlight a negative experience rather than a positive one, which can explained by this observation. Another noticeable trend is that American, United, and US Airways have a significantly higher proportion of negative tweets when compared to the Delta, Southwest, and Virgin America who have a more even distribution.

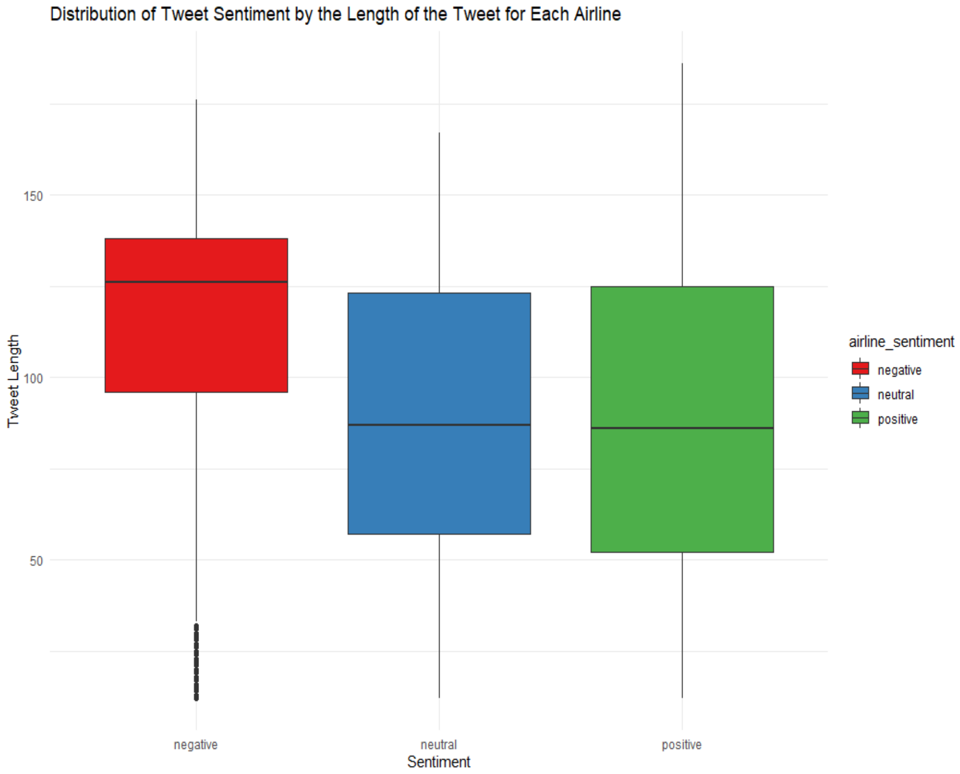


By separating the tweets into the days of the weeks, this graph tells us a few key important points. First, the most interesting point is that American only had tweets from Mon, Sun, and Tues. But, spreading out those tweets puts them about average for total number of tweets. The last point is that Virgin America had by far the least amount of tweets. This may explain the even distribution of sentiment in the last graph.

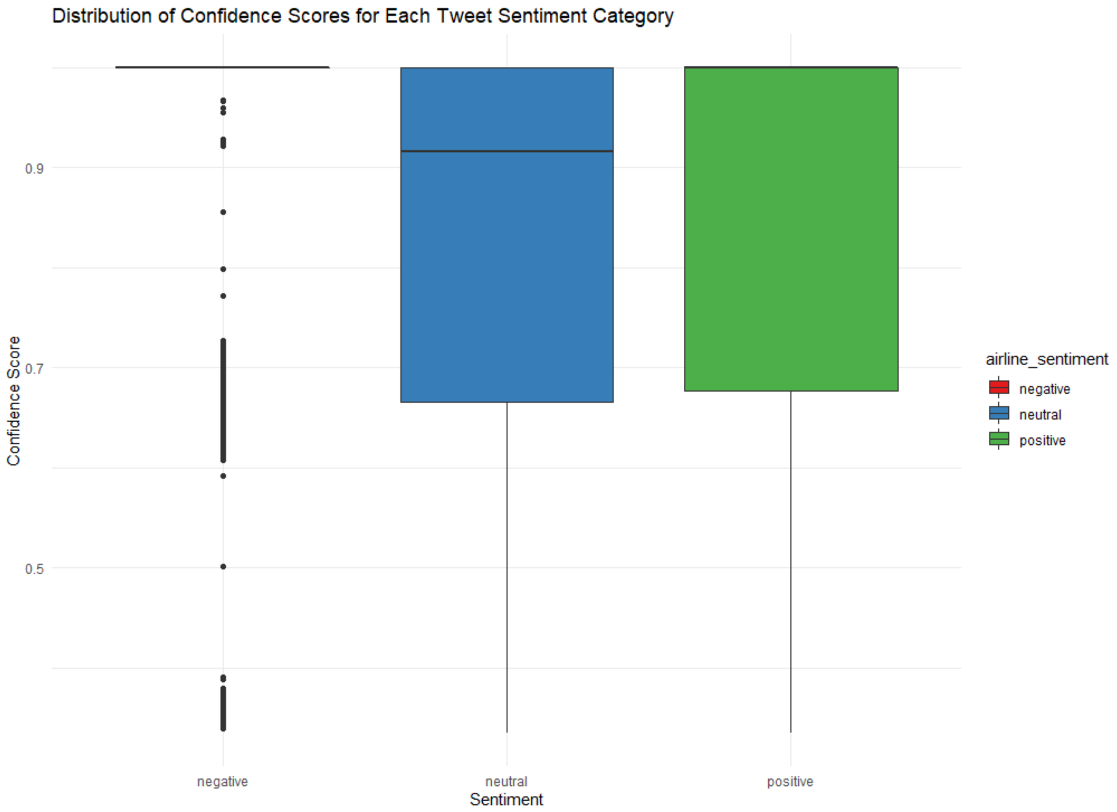


An interesting factor that could have played into the role of many tweets getting categorized as “negative” by the AI is the length of the tweet.

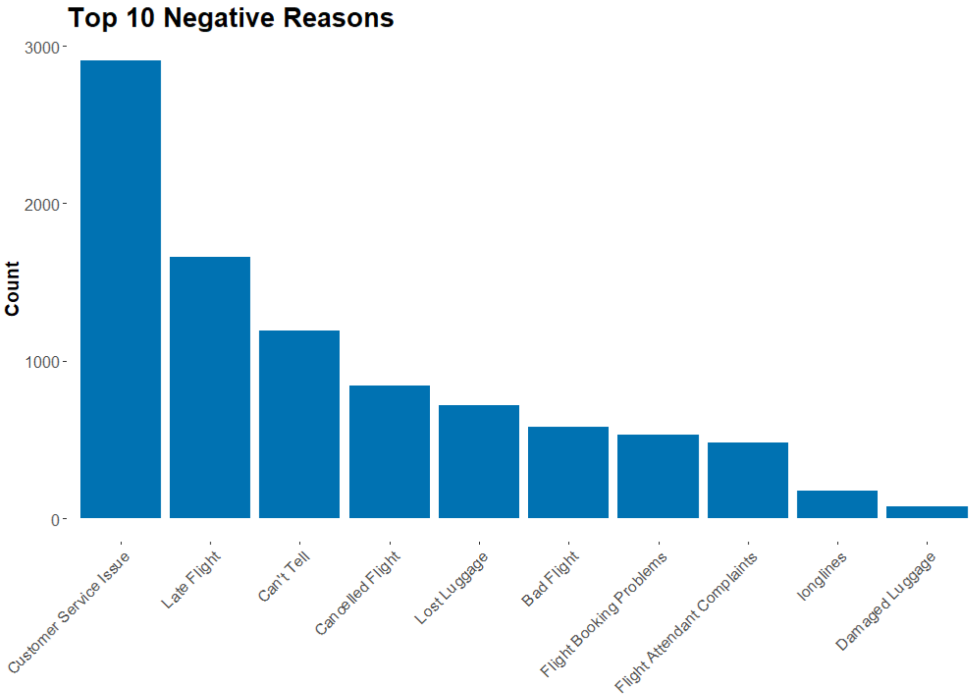
This box and whisker plot shows us that the average length of a negative tweet is about 25% larger than the neutral and positive tweets. From the way the AI categorizes tweets, this may be showing the bias towards categorizing longer tweets as negative.



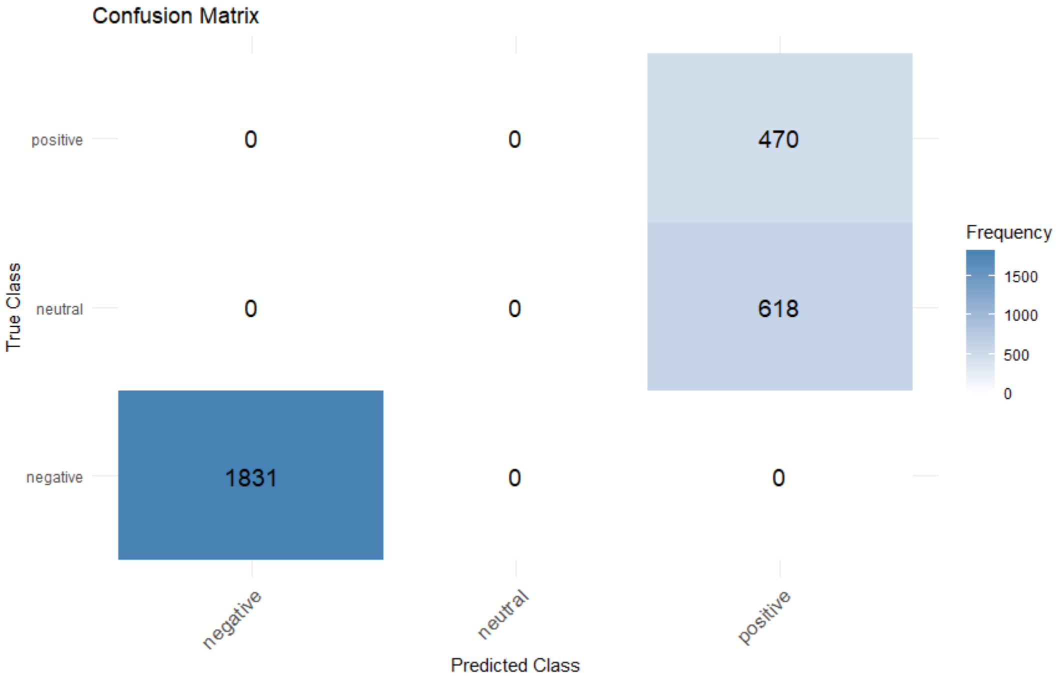
To confirm our suspicions, I looked into the distribution of confidence scores for each sentiment. Negative shows an almost perfect 1 average meaning that the AI is fully confident that it had recognized the negative tweets correctly. Pairing this with our last question, we can conclude that longer tweets are more likely to be negative with the AI worse at predicting neutral and positive tweets.


Distribution of Confidence Scores for Each Tweet Sentiment Category

Next, I looked into the reasons why airlines might be getting negative reviews. From the “negativereason” column, I created a bar chart to plot the frequencies of each reason. The top issue with the airlines was Customer Service. This shows how airlines may be able to improve their relationship with customers through investing into ensuring on-time flights and resolving customer service issues in a timely manner.



My last question was using a logistic regression model, I was using Airline, Reason, and Timezone to predict Airline Sentiment. I used an 80/20 training test split and overall the model performed well. With only a small portion of false positives in from neutral, the model was able to perfectly predict all of the negative reviews. This shows us that the logistic regression model is proficient in being able to predict Airline Sentiment given Reason and Timezone of the user. However, a fatal flaw to this model is it might have result from bias because Reason and Airline Sentiment are too closely related. In future iterations, this may have different predictors.


Confusion Matrix
True Class
positive
neutral
negative
negative
neutral
positive
Predicted Class
Frequency