US Airlines Twitter Sentiment Analysis

Kayelin Santa Elena
MGSC 410-01

The dataset

Twitter US Airline Sentiment – Kaggle

14 **Columns**

The number of columns in the dataset

14,640 **Rows**

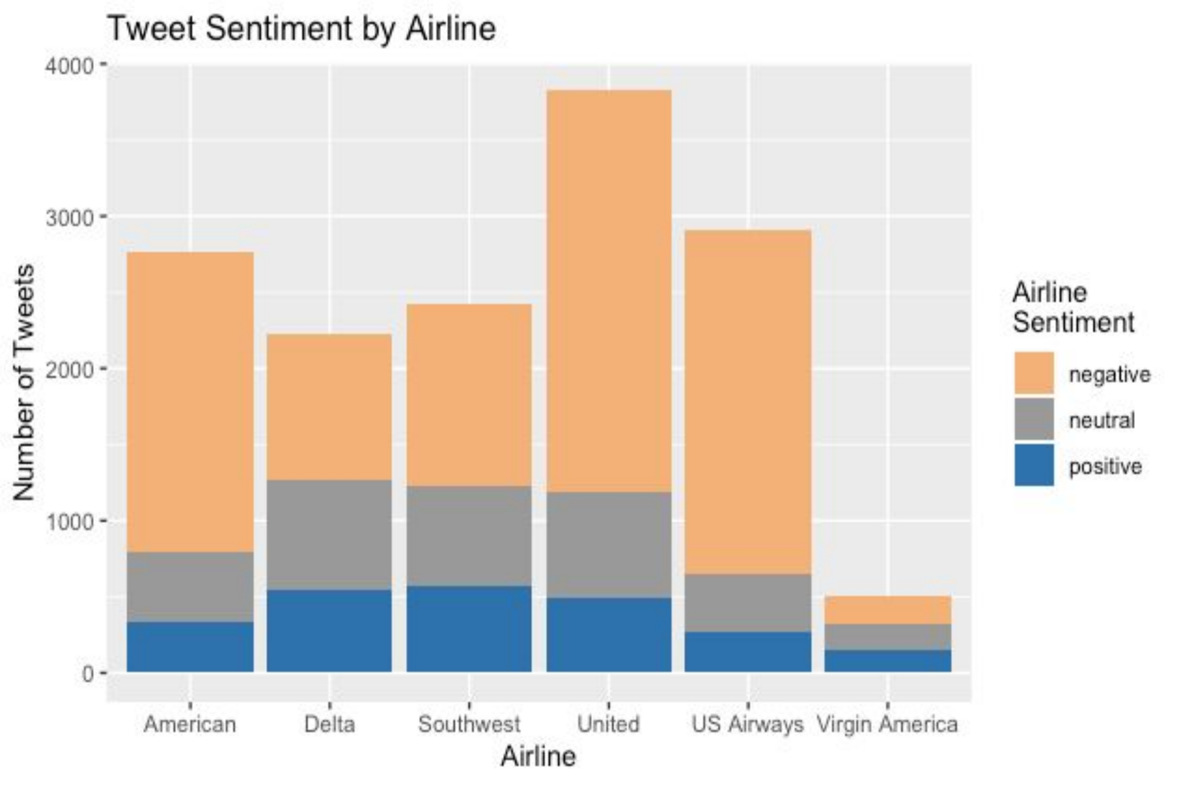
The number of rows in the dataset

6 **Airlines**

The number of airlines in the dataset

7,701 **Usernames**

The number of unique usernames in the dataset



Who contained the most of each tweet sentiment?

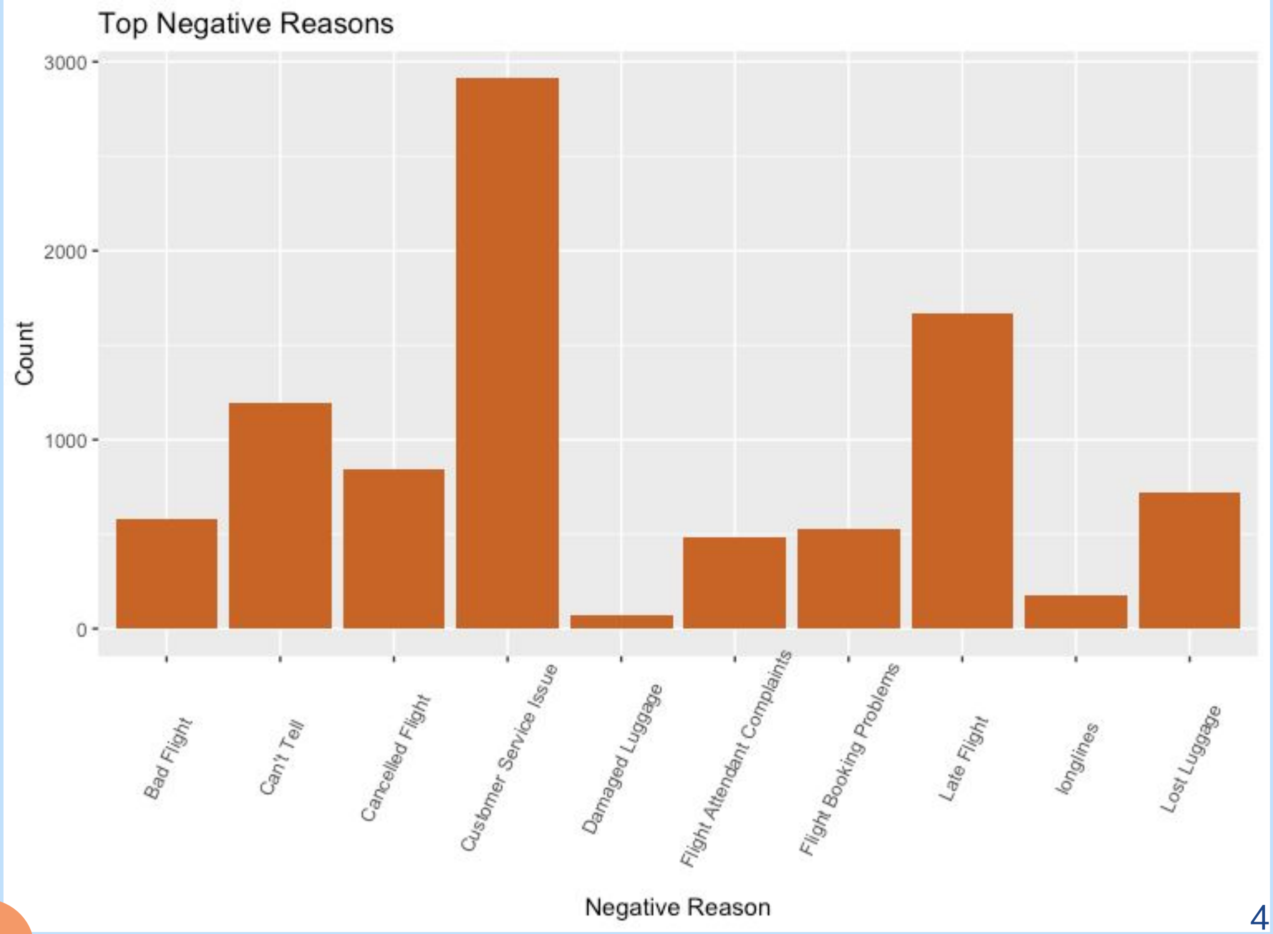
Negative: United

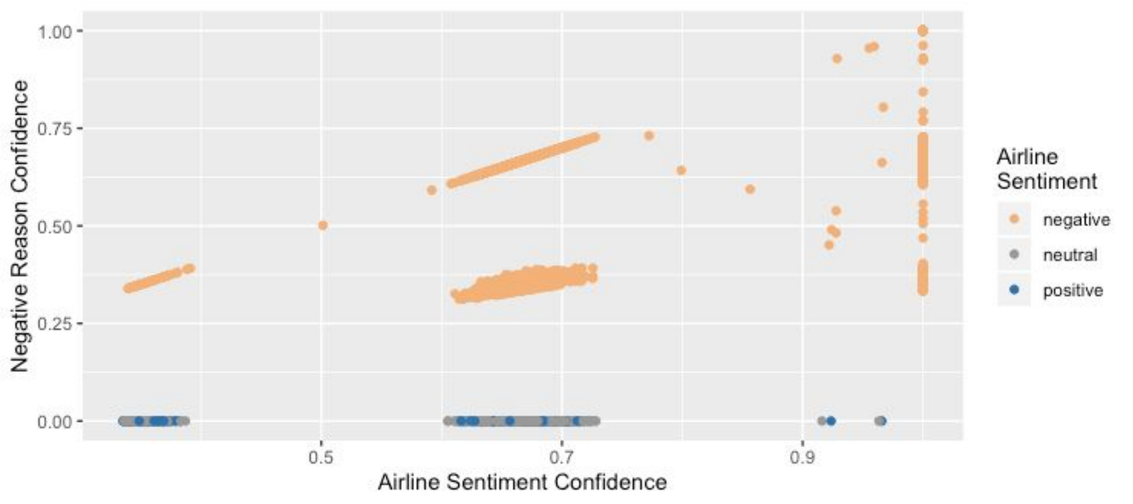
Neutral: Delta

Positive: Southwest

Top negative reasons:

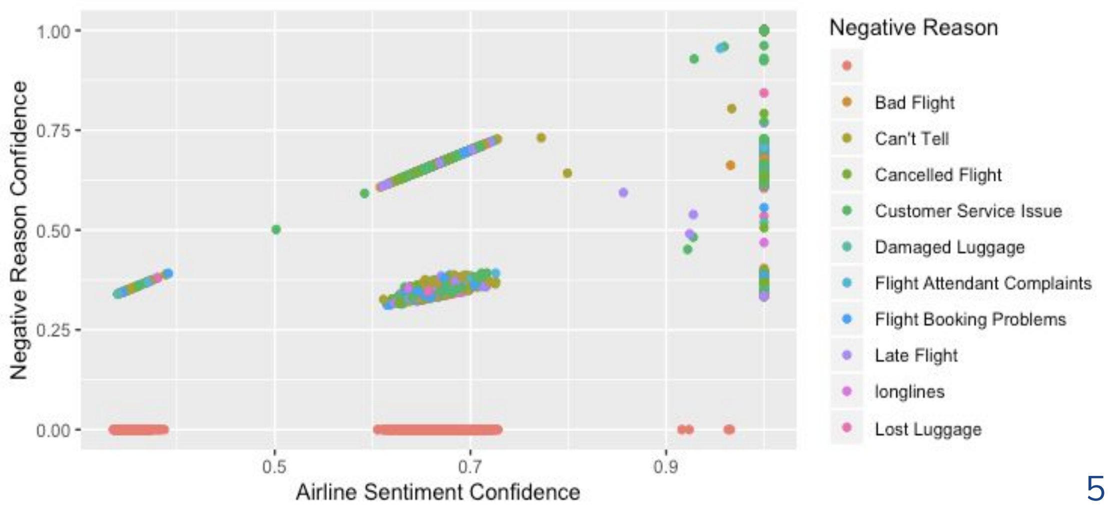
- Customer service issue
- Late flight
- Can’t tell





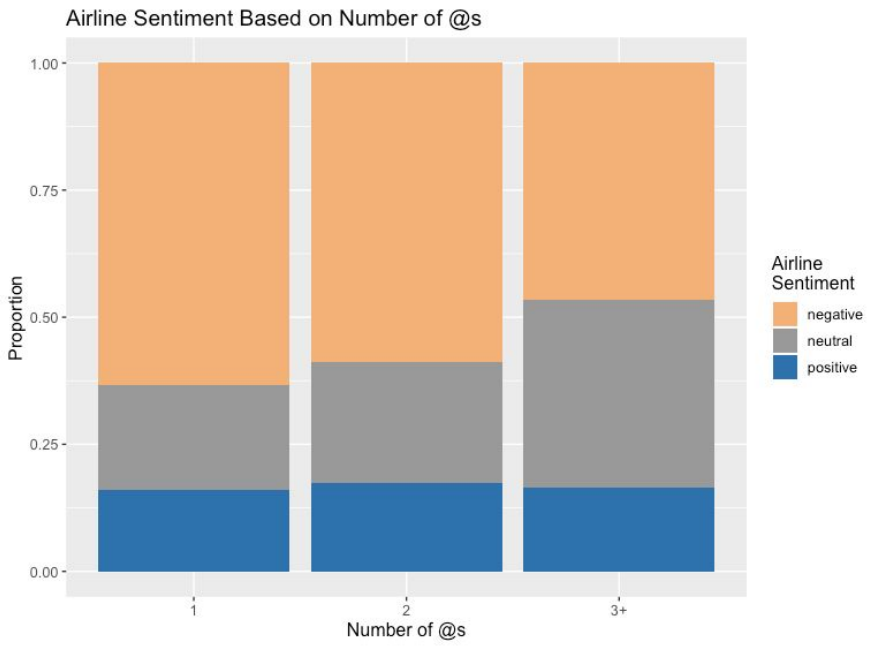
- The closer to 1 for Airline Sentiment Confidence the more likely a tweet is to be negative
- The closer to 0 for Airline Sentiment Confidence the more likely a tweet is to be neutral or positive

- The tweets with a high negative reason and airline sentiment confidence are most likely to be about a canceled flight or customer service issue

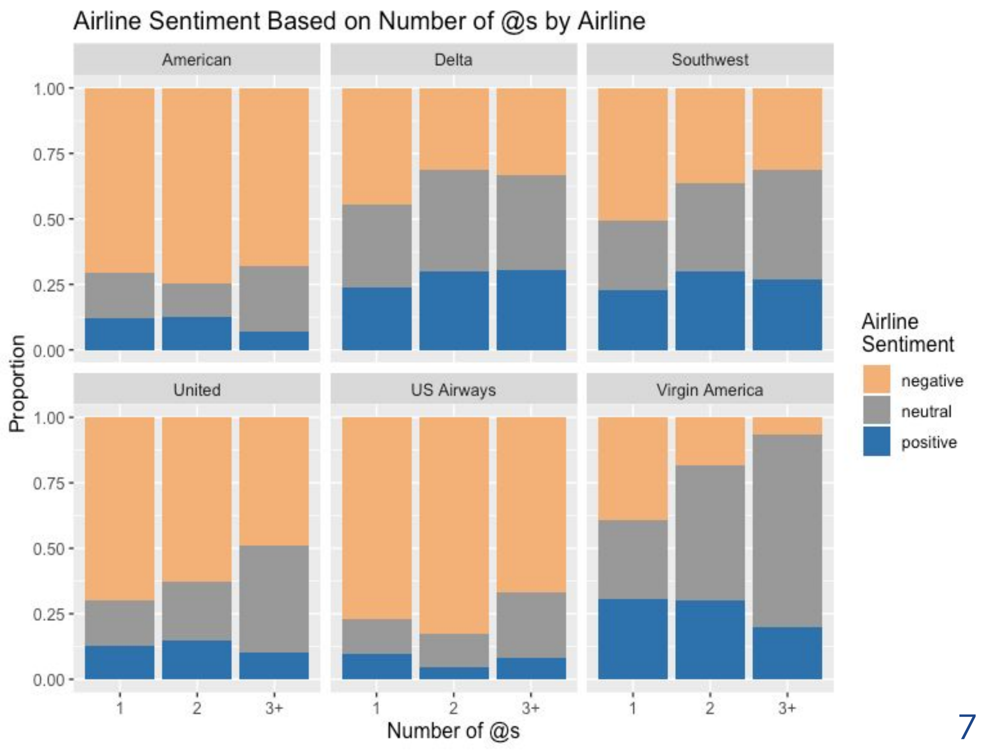


Is a sentiment more likely to be negative if a tweet has multiple @s and a long text length?

Negative tweets tend to have less @s and are considerably longer than positive or neutral ones.

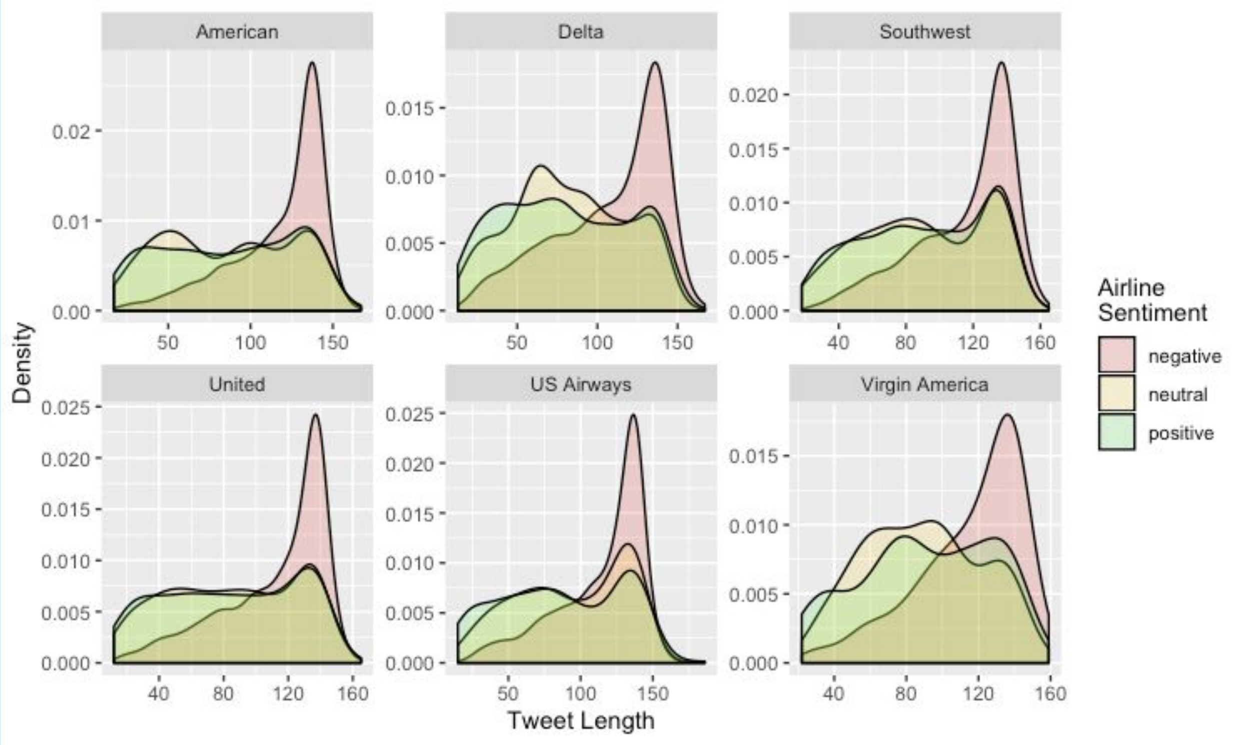


- As the number of @s in a tweet increase the sentiment is more likely to be neutral



- US Airways has the most negative tweets with 2 @s
- Virgin America has the most neutral tweets with 3+ @s

The tweets that reach the 170 character limit are mostly directed at Virgin America



What 3 factors are most important when calculating an airline's sentiment confidence?

Linear regression can tell us that the most important factors are negative reason confidence, the number of @s, and retweet count.

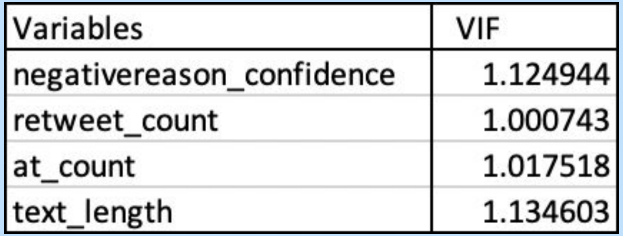
Linear Regression

Model Scores

Train RMSE: 0.1493641
Test RMSE: 0.1536577
Train R2: 0.143136

Issues

Heteroskedasticity
Multicollinearity

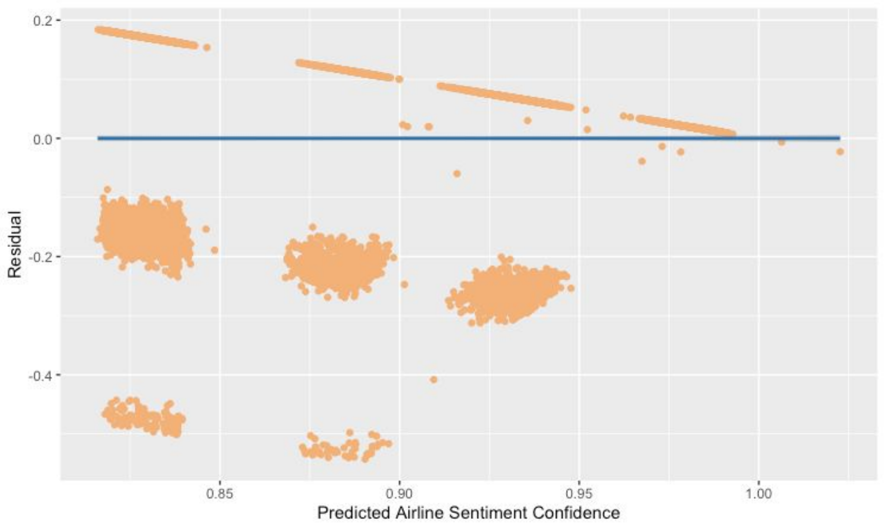


Variables	VIF
negativereason_confidence	1.124944
retweet_count	1.000743
at_count	1.017518
text_length	1.134603

VIF > 10 indicates problematic level of multicollinearity

There is not much of an overfitting issue since there is no big difference between the RSMEs

R2 is very low, the model did not do well



There are signs of heteroskedasticity which may contribute to the low R2.

Appendix

I originally planned to do a logistic regression model with airline_sentiment as my dummy variables. I changed “Negative” to 0 and “Neutral” and “Positive” to 1. When I ran the model I would always get a warning message “algorithm did not converge” and “fitted probabilities numerically 0 or 1 occurred.” This led me to believe I was not making the model correctly so I decided to make a linear regression model instead.

The *Twitter US Airline Sentiment* dataset from Kaggle was hard to work with. There was no background info to explain the columns so I had to learn what they were with further analysis. There were 4118 missing variables, and only 3 columns with usable continuous number data.There were not enough continuous variables as I would have liked, so I had to make some out of the other columns.

The coefficients for the linear regression:
negativereason_confidence: 0.1493116124
at_count: -0.0030643271
retweet_count: 0.0007922204
text_length: 0.0001663162