AIRLINE TWITTER SENTIMENT ANALYSIS

BY KAYLA CHO
MGSC 410

OVERVIEW OF DATA SET

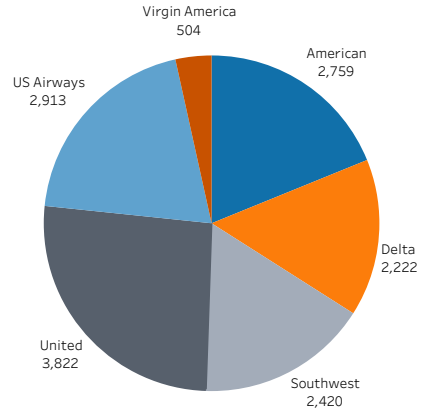
MOTIVATION

To understand a sentiment analysis done on tweets written about airlines by passengers in February 2015.

THE DATA SET

- Tweets from February 2015
- 14,640 Entries
- 15 Columns
- 6 Airlines
  - Majority about United
- Airline Sentiment Confidence
  - Range ≈ 0.34 – 1.0
  - Mean ≈ 0.9
- Negative Reason Confidence
  - Range ≈ 0.0-1.0
  - Mean ≈ 0.64

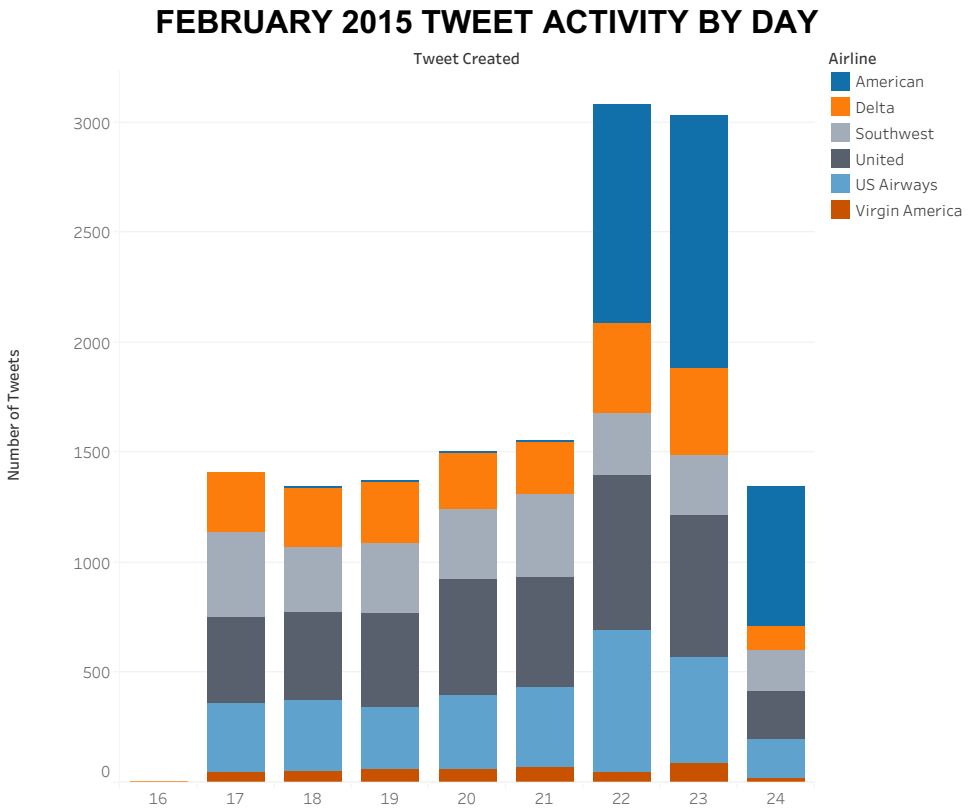
TWEET COUNT BY AIRLINE



SUMMARY STATISTICS

	tweet_id	airline_sentiment_confidence	negativereason_confidence	retweet_count
count	1.464000e+04	14640.000000	10522.000000	14640.000000
mean	5.692184e+17	0.900169	0.638298	0.082650
std	7.791112e+14	0.162830	0.330440	0.745778
min	5.675883e+17	0.335000	0.000000	0.000000
25%	5.685592e+17	0.692300	0.360600	0.000000
50%	5.694779e+17	1.000000	0.670600	0.000000
75%	5.698905e+17	1.000000	1.000000	0.000000
max	5.703106e+17	1.000000	1.000000	44.000000

TWEET ACTIVITY



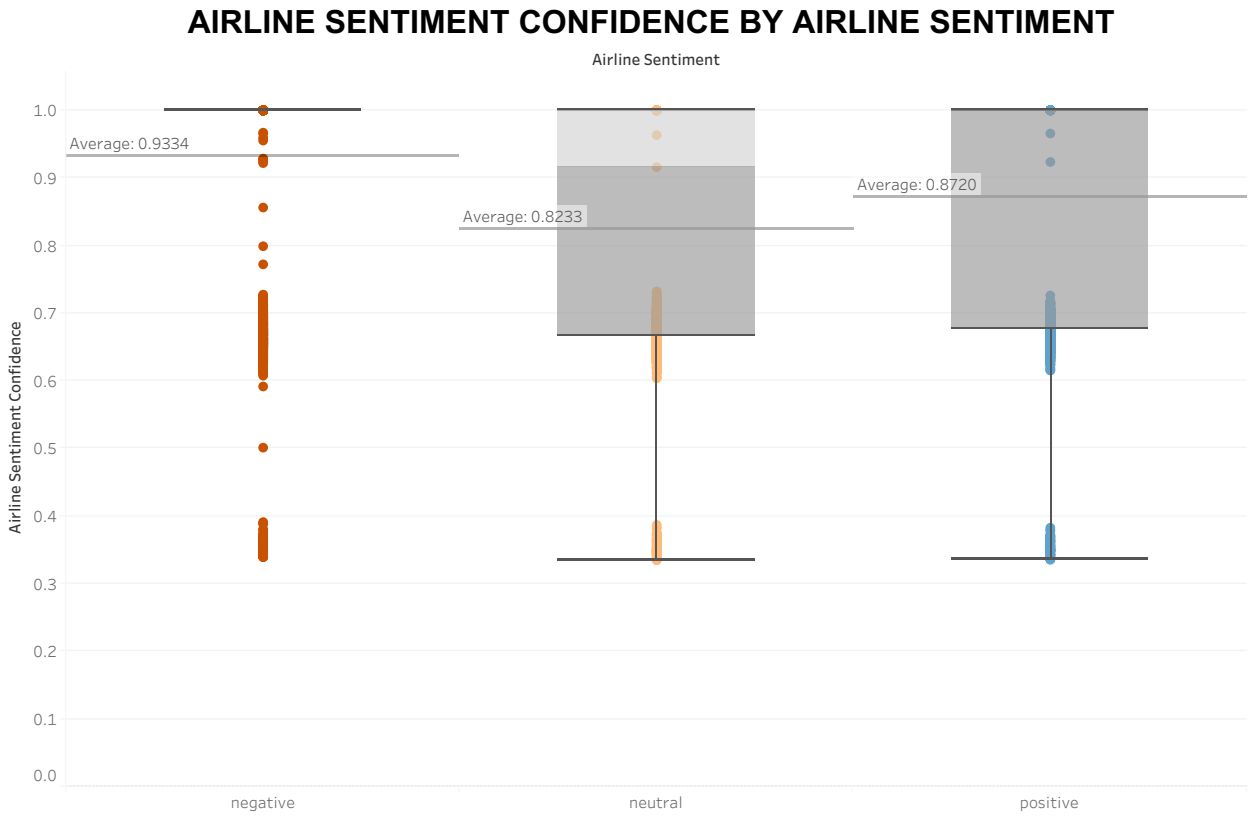
- Range: February 16th- 24th 2015
- Most Tweets Posted: February 22nd & 23rd

February 22nd – 24th

  - Number of tweets doubled compared to previous days
  - Tweets about American Airlines rose from one a day to over 1,000

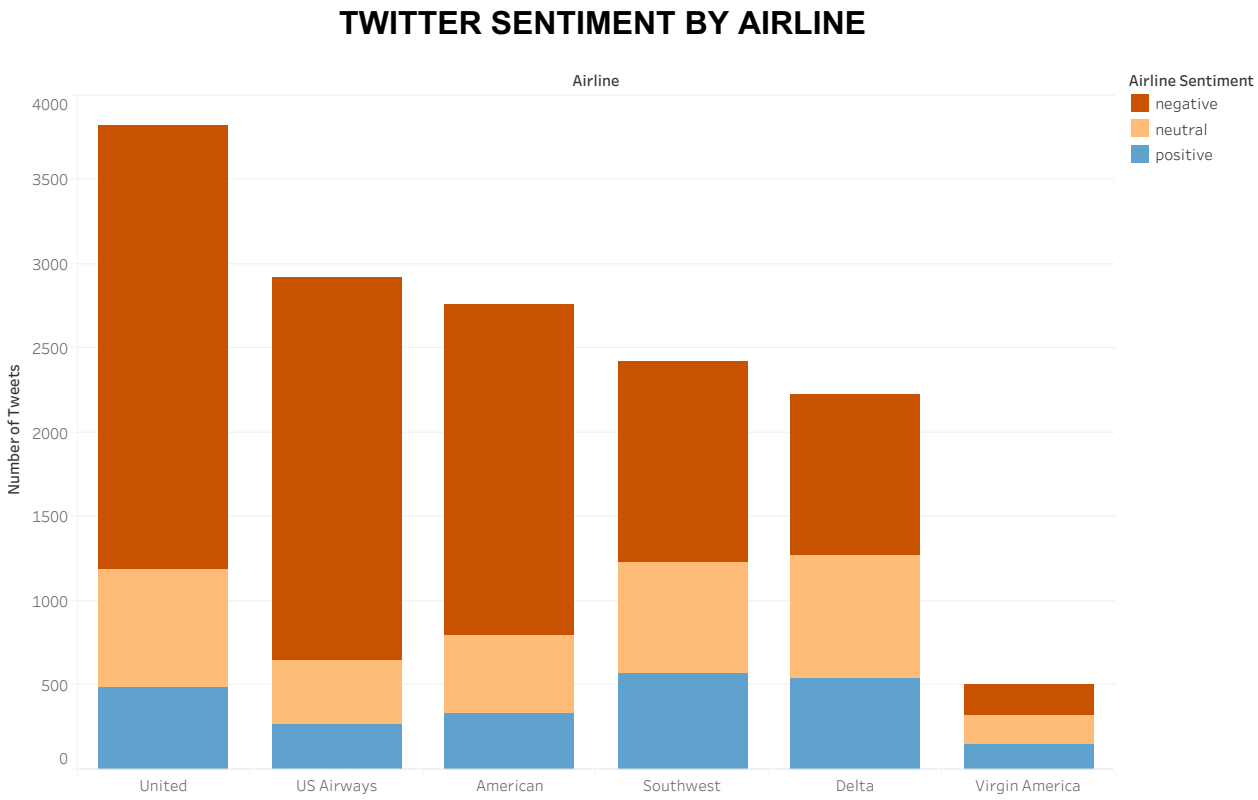
- Other airlines generally had steady daily tweet activity respective to their average number of tweets per day.

AIRLINE SENTIMENT CONFIDENCE



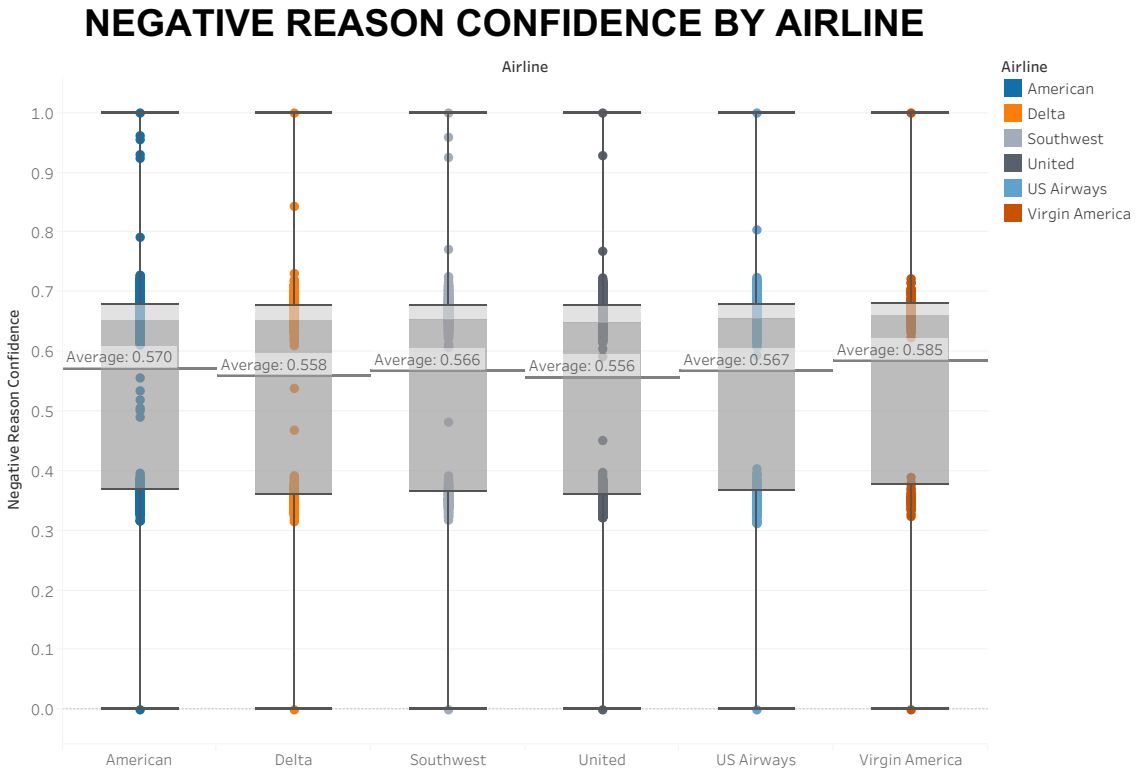
- Averages: ≈ 0.82-0.93
- Highest Average Level of Confidence: Negative
- Lowest Average Level of Confidence: Neutral
- Distribution of Twitter sentiment confidence is similar across the different sentiment classifications.
- Overall, approximately 90% confident that the tweet is put into the correct sentiment.

TWITTER SENTIMENTS



- Majority of the tweet sentiments were negative.
- Most tweets with negative sentiments: United Airlines
- Most tweets with positive sentiments: Southwest
- Airlines with more tweets appear have a higher negative sentiment ratio.

NEGATIVE REASON CONFIDENCE



Averages: ≈ 0.55-0.58

Highest Average Level of Confidence: Virgin America

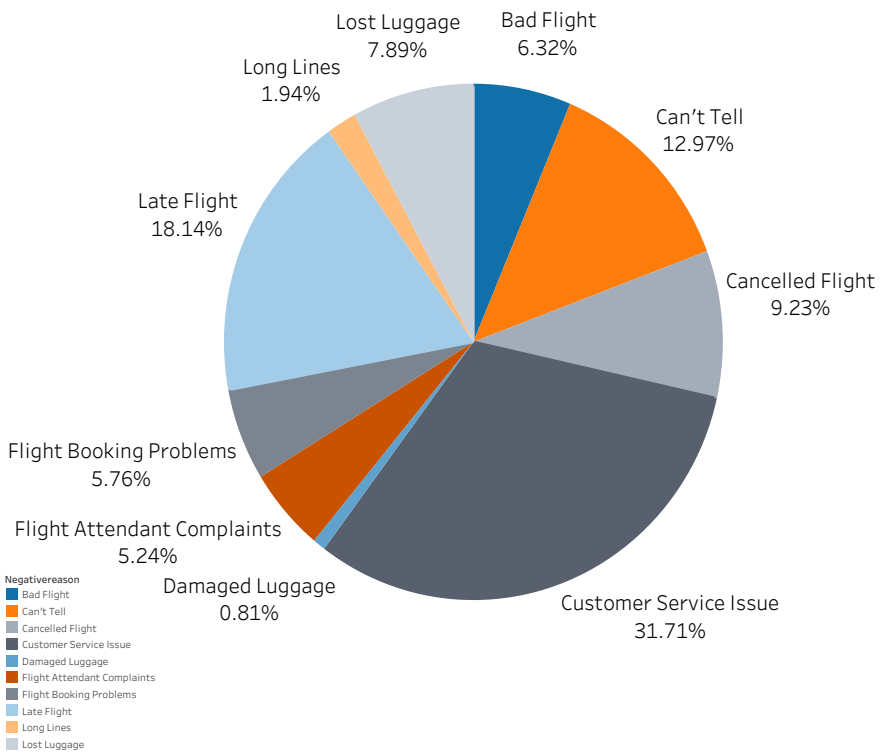
Lowest Average Level of Confidence: United

Distribution of Twitter sentiment confidence is similar across the different airlines.

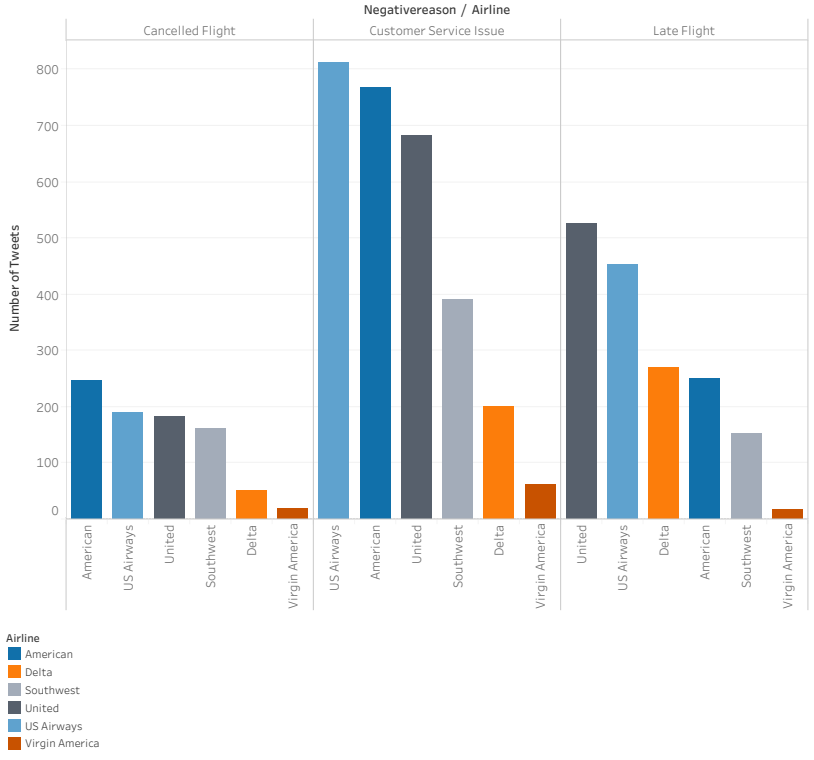
Overall, approximately 56.4% confident that the tweet's negative reason is accurately classified.

NEGATIVE REASONS FOR TWEETS

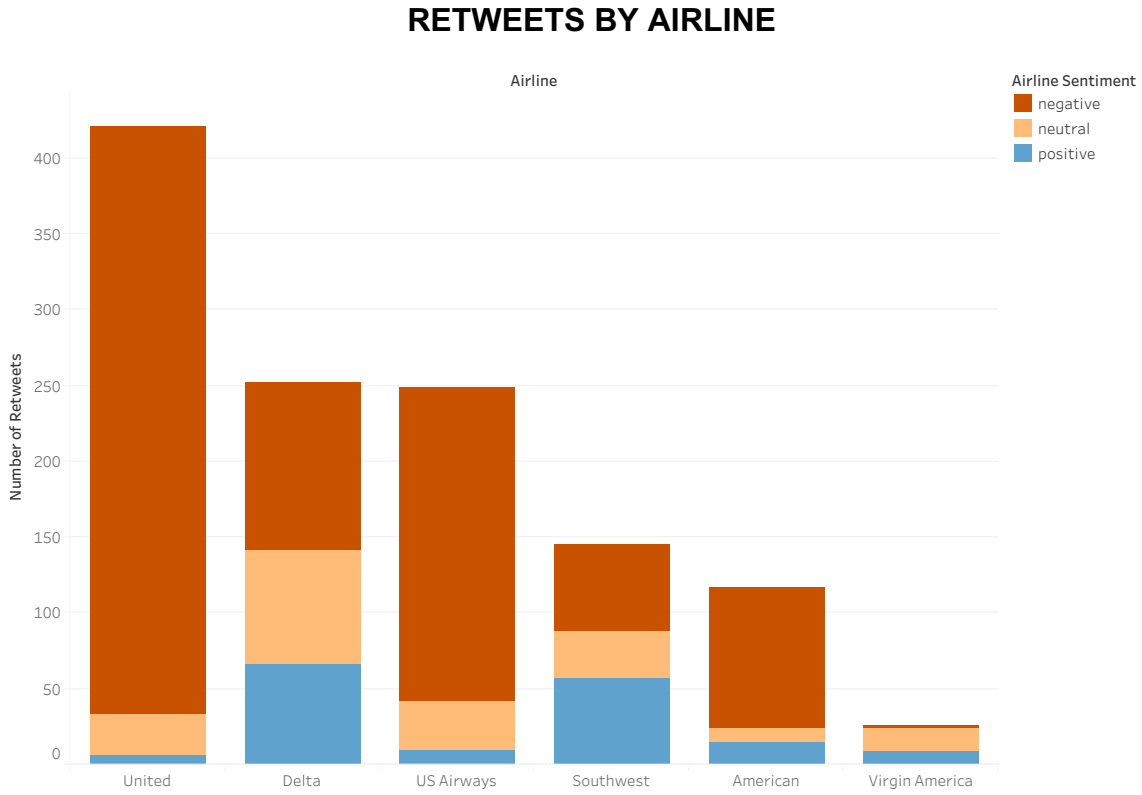
NEGATIVE REASON FOR TWEETS



TOP 3 NEGATIVE REASONS BY AIRLINE



AIRLINE RETWEETS



Most

- Total Retweets: United
- Negative Retweets: United
- Positive Retweets: Delta

Least

- Total Retweets: Virgin
- Negative Retweets: Virgin
- Positive Retweets: United

Delta had the 2nd lowest number of total tweets in 2015 but had the 2nd highest number of retweets & highest number of positive retweets

Overall, similar pattern to the total number of tweets and its corresponding sentiments.

INSIGHT HIGHLIGHTS

1

The sentiment analysis is on average 90% confident that it classified tweets correctly.

2

Airline tweets & retweets from February 2015 are largely classified as negative.

3

The top negative reason for airline tweets is customer service issues.

4

All airlines had the same top three negative reasons for tweets.

TAKEAWAYS

1

After looking through the tweets individually, I noticed some tweets were given a high sentiment and negative reason confidence score but were incorrect. Therefore, we might want to revaluate how the confidence score is curated and not use these confidence scores to make significant decisions.

2

While many tweets were classified as negative, approximately half of them were neutral or positive. Consider exploring reasons behind the positive and neutral tweets to understand what certain airlines do that is unique that passengers like.

3

Many tweets were created on February 22nd and 23rd, which should be researched further to understand what may have caused the increase. Also, keep in mind that the data set might be skewed because of those two days, so we might consider analyzing a period longer than nine days.

4

Since all airlines had the same top three negative reasons for their tweets, companies may want to understand these issues further to differentiate themselves from their competitors.

APPENDIX

APPENDIX A : DATA ASSESSMENT

Quality of data

The columns negativereason_gold, airline_sentiment_gold, and tweet_coord, were missing more than 99% of their rows so I chose to leave them out of the Executive Summary. I also didn't analyze tweet_id, name, or text because it was unique to the tweet posted and tweet_location and user_timezone because difficult to work with because it wasn't standardized.

Findings

When looking through the tweets, I found there were tweets that were misclassified when the sentiment confidence was extremely high. Therefore, we might not want to heavily rely on these sentiments to make any large decisions.

Issues

I spent a couple of hours researching how to manipulate the columns that contained locations but could not find a comprehensive way to do it. I tried to use Tableau, but since I am not super familiar with the platform, I wasn't sure why I only got a couple of results. I also couldn't find a way to manually change the location classification unless it was unknown to Tableau. I also struggled when working with the negative reasons column in Python and R because I could not exclude certain factors or null values. Another issues I had was attempting to create word clouds in Tableau because there were so many insignificant words like "you" which occurred the most and I couldn't filter through the all. Also when using Tableau I realized that Tableau defaults to using an aggregate measures that affected my results and I had to go back and change each graph so it accounted for all the data.

```
df.isnull().sum()
```

```
tweet_id                          0
airline_sentiment                 0
airline_sentiment_confidence      0
negativereason                 5462
negativereason_confidence      4118
airline                           0
airline_sentiment_gold        14600
name                              0
negativereason_gold           14608
retweet_count                     0
text                              0
tweet_coord                   13621
tweet_created                     0
tweet_location                 4733
user_timezone                  4820
dtype: int64
```

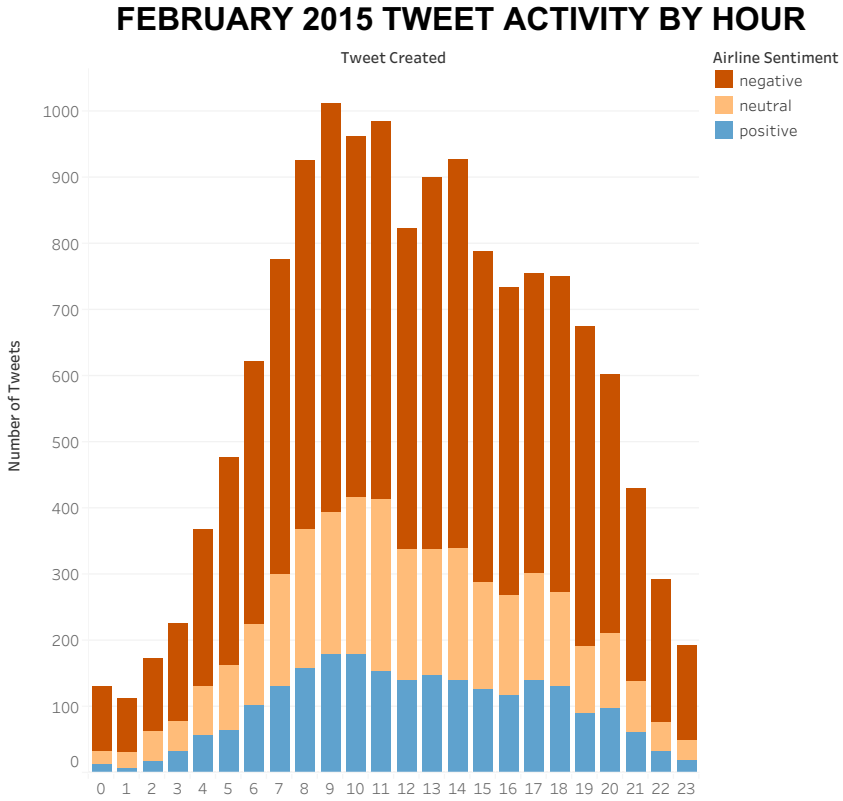
Misclassification Example

American Airlines
Negative Reason: Can't Tell
Sentiment Confidence: 1
Negative Sentiment Confidence: 0.701
@AmericanAir @Delta @JetBlue can you help? @SpiritAirlines is in the wrong, stuck in Florida 3days. They need to learn from real airlines.

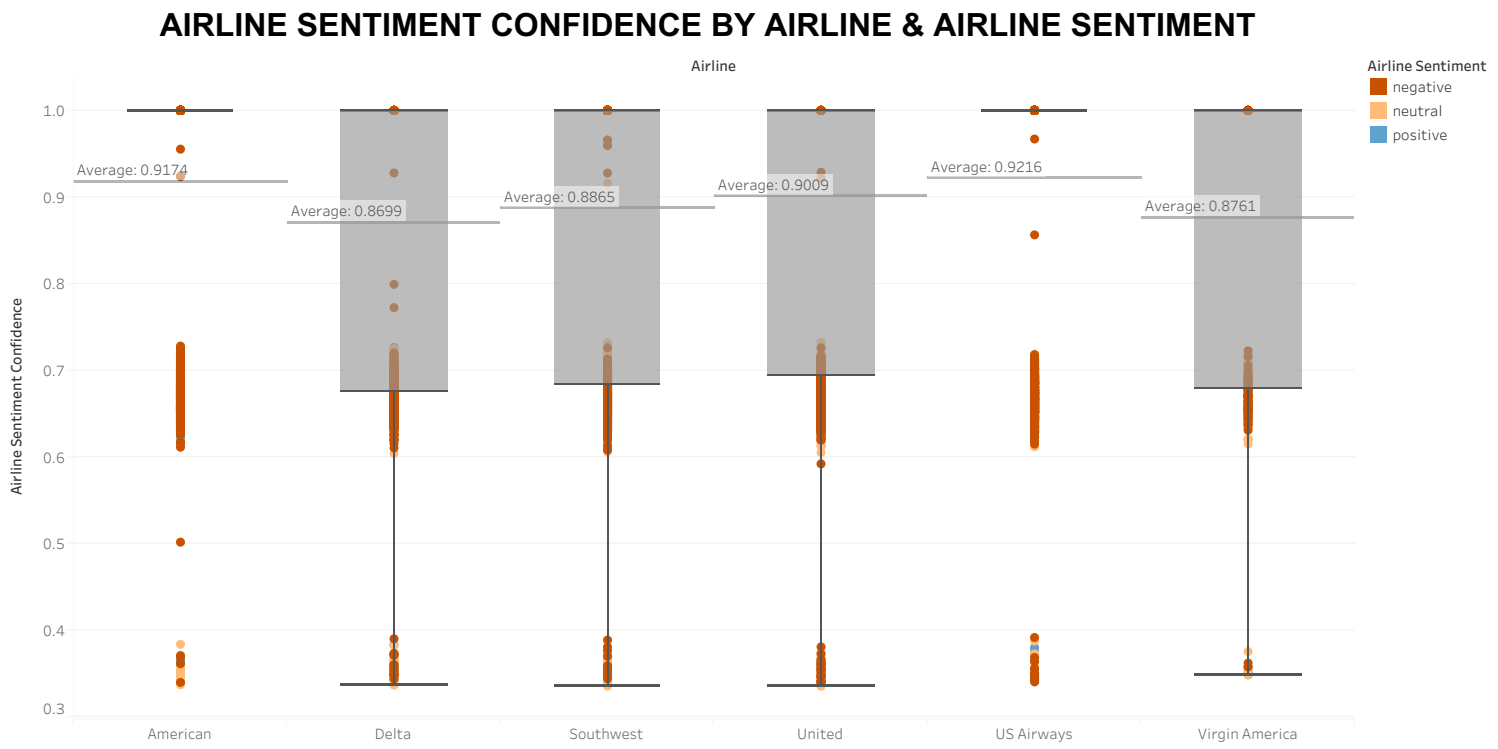
APPENDIX B : ACTION STEPS

1. **Explored the data set**
    a. Used Python and Excel to look at the shape of the data set (columns, rows, NULL values, summary statistics)
    b. Researched about the data set (learned what the variables meant, looked through Kaggle discussions and the original publisher's findings)
2. **Data Visualization Research**
    a. Researched how to clean the data to get uniform locations from Twitter data
    b. Looked at different possible graphs that I could create on different platforms
3. **Explored and analyzed data using graphs and charts**
    a. Started in Excel to look through the data and create basic bar charts and pie charts.
    b. Moved to using Python and then moved to R to be able to rename factor levels and mutate the data set.
    c. Converted data into appropriate data types
4. **Created PowerPoint Presentation**
    a. Created an outline that tells a story about the data
    b. Found a PowerPoint template online
    c. Learned how to use Tableau
    d. Took all graphs created in Excel, Python, and R and recreated it in Tableau to make it uniform.
    e. Made graph colors to be color blind friendly to ensure anyone who looks at it can understand it
    f. Presented presentation to family members to make sure it was understandable.

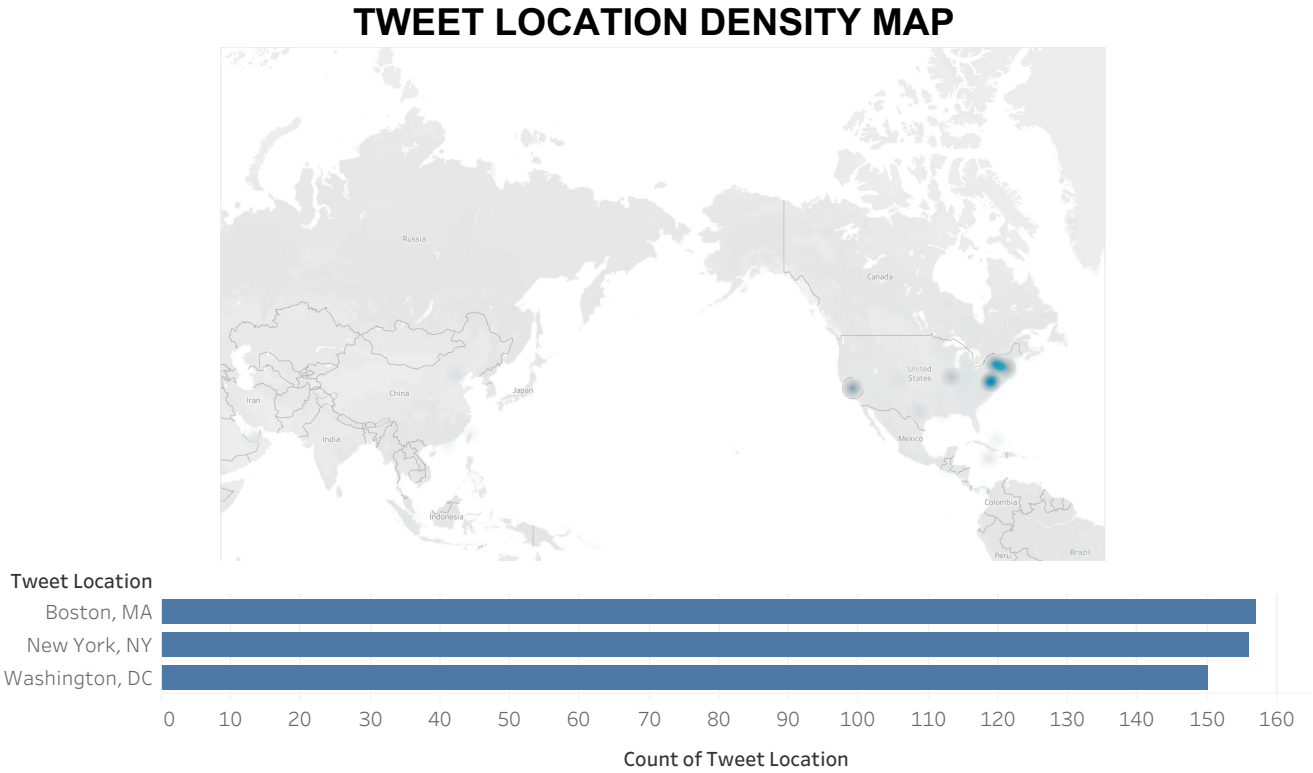
APPENDIX C : FEBRUARY TWEET ACTIVITY BY HOUR



APPENDIX D :
AIRLINE SENTIMENT CONFIDENCE BY AIRLINE



APPENDIX E : TWEET LOCATION DENSITY

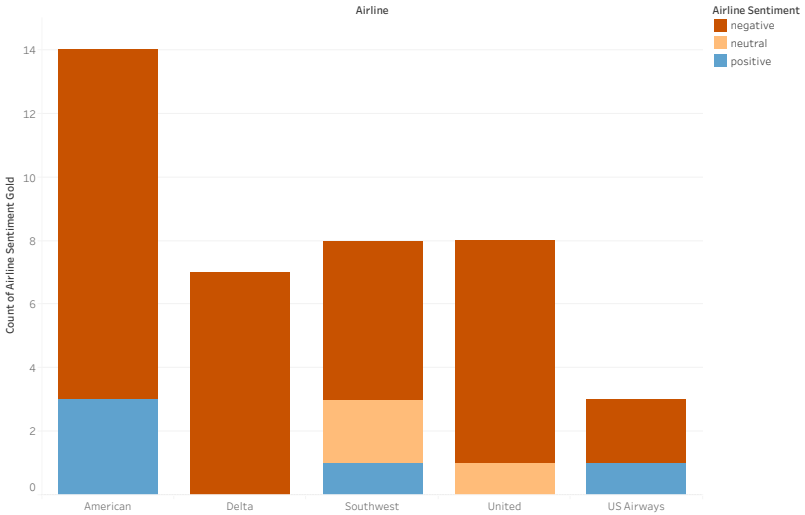


APPENDIX F : USER TIME ZONE DENSITY

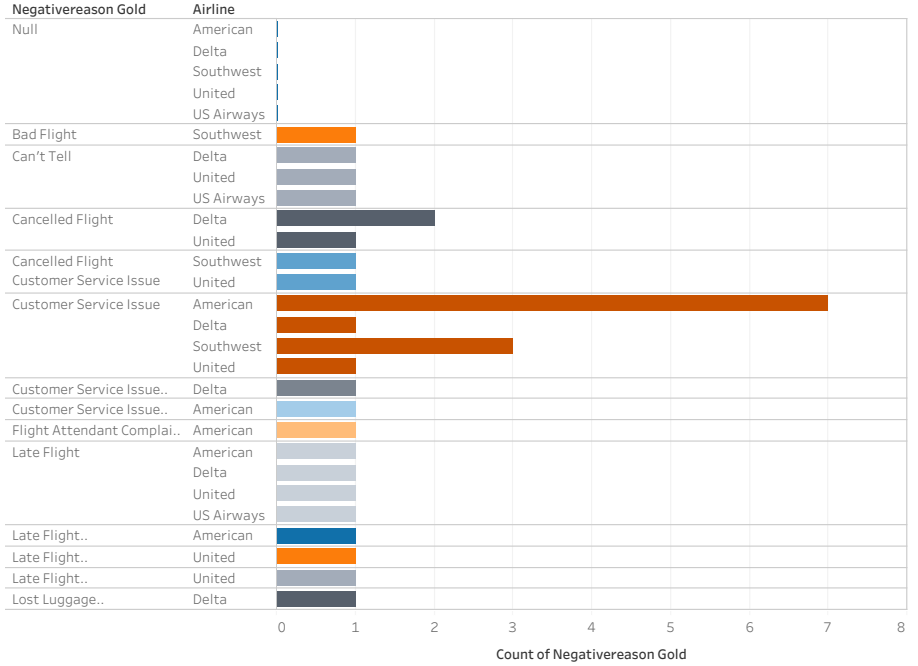
USER TIME ZONE DENSITY MAP

APPENDIX G : GOLD COLUMNS ANALYSIS

GOLD AIRLINE SENTIMENT BY AIRLINE



NEGATIVE REASONS GOLD



APPENDIX H : NUMBER OF TWEETS PER USER