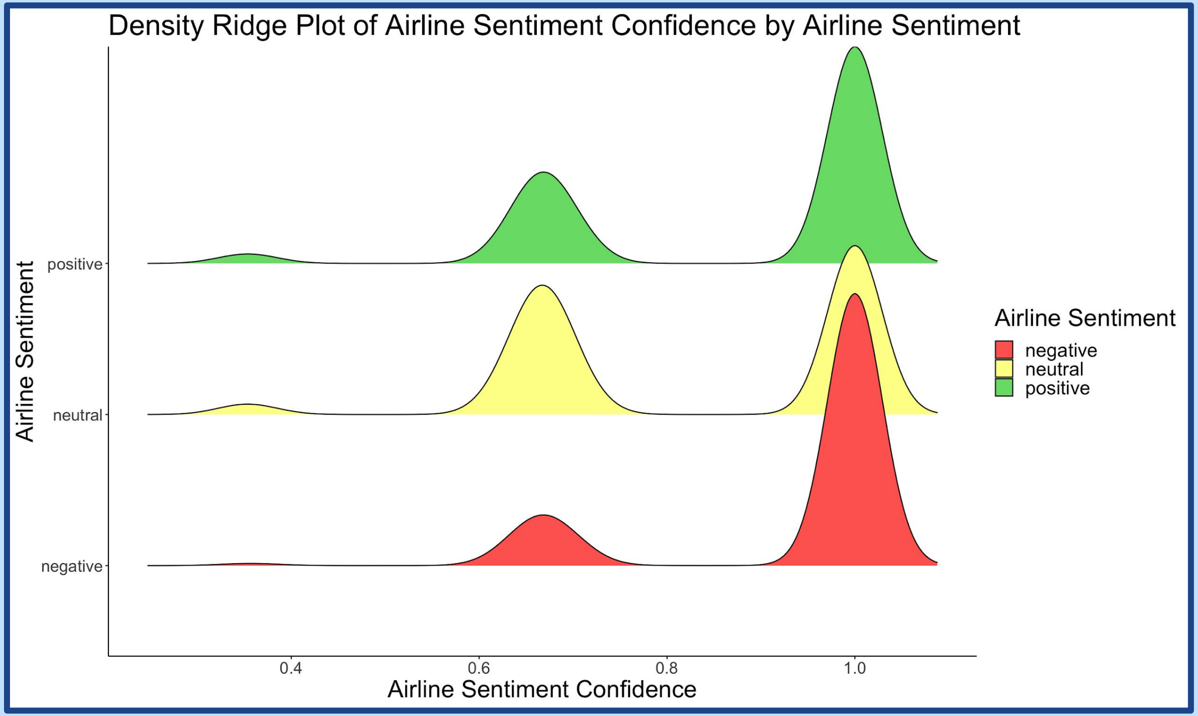
Airline Insights through Twitter

James Kistner

Understanding the Data

- The data in this presentation was taken over eight days.
- The information is for six different airlines: Delta, United, Southwest, US Airways, Virgin America, and American.
- The *airline_sentiment* attribute tells us which category a tweet is (negative, neutral, or positive) based on sentiment analysis of the text of the tweet. There is a confidence attribute associated with the sentiment as well.
- Of the cleaned data nearly 58% were negative tweets.
- For the negative tweets, there are two attributes, *negativereason* and *negativereason_confidence*, that explain the data further, giving the reason for the negative tweet and the confidence associated with that reason.

Caution About the Data



- Airline sentiment confidence has three distinct peaks instead of being continual or having one peak. This indicates a lack of continuity on the metric used to calculate confidence.
- Sentiment analysis uses a dictionary of words and has a score for each word. These scores don't take into account context or tone, therefore they can be misleading.
- 11% of negative reasons are "Can't Tell".

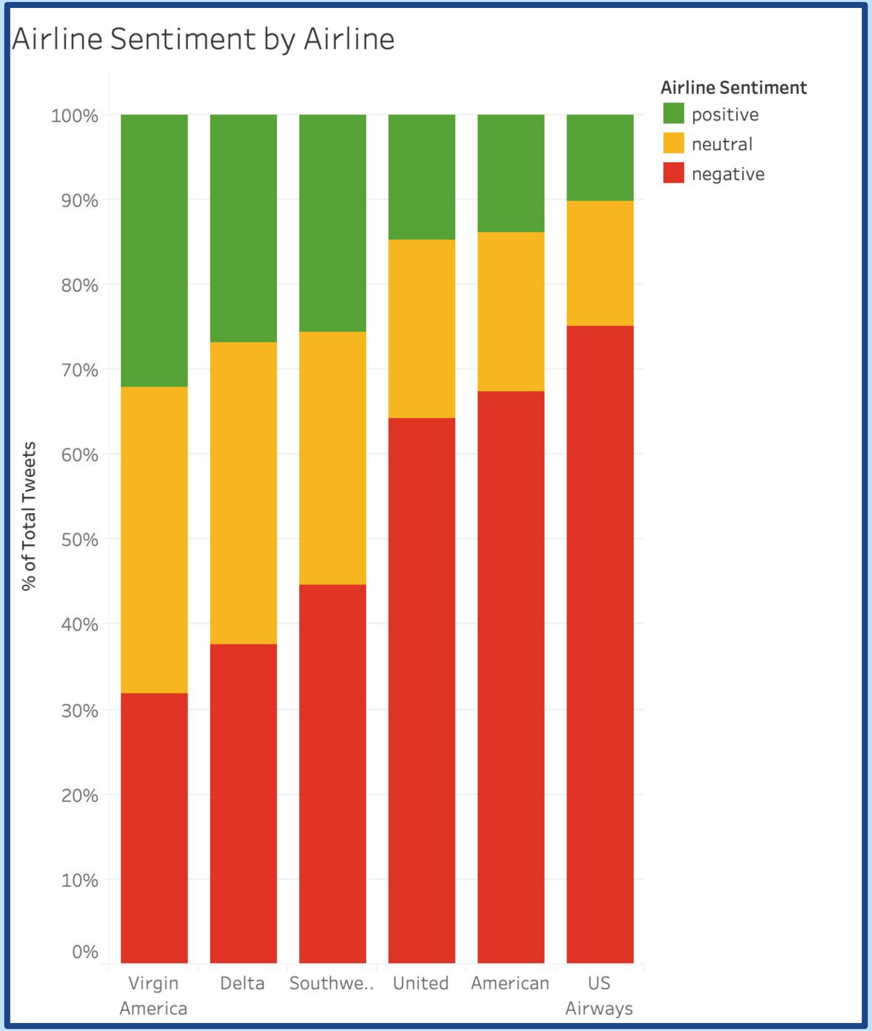
Best and Worst Airlines

Top 3 airlines based on the equation (positive tweets per airline / total tweets per airline):

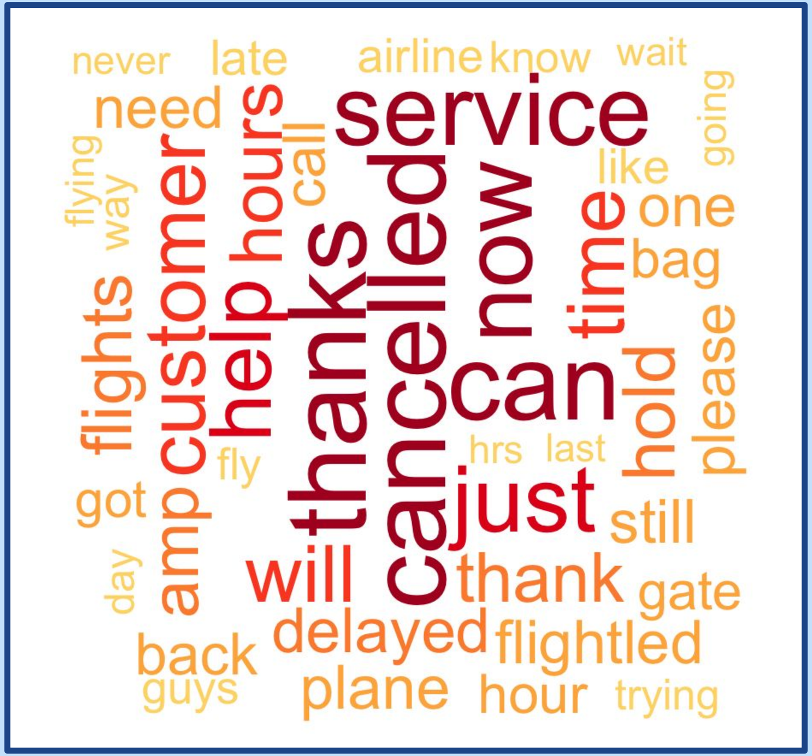
1. Virgin America = 0.32 = 32% positive
2. Delta = 0.27 = 27% positive
3. Southwest = 0.26 = 26% positive

Bottom 3 airlines based on the equation (negative tweets per airline / total tweets per airline):

1. US Airways = 0.75 = 75% negative
2. American = 0.67 = 67% negative
3. United = 0.64 = 64% negative



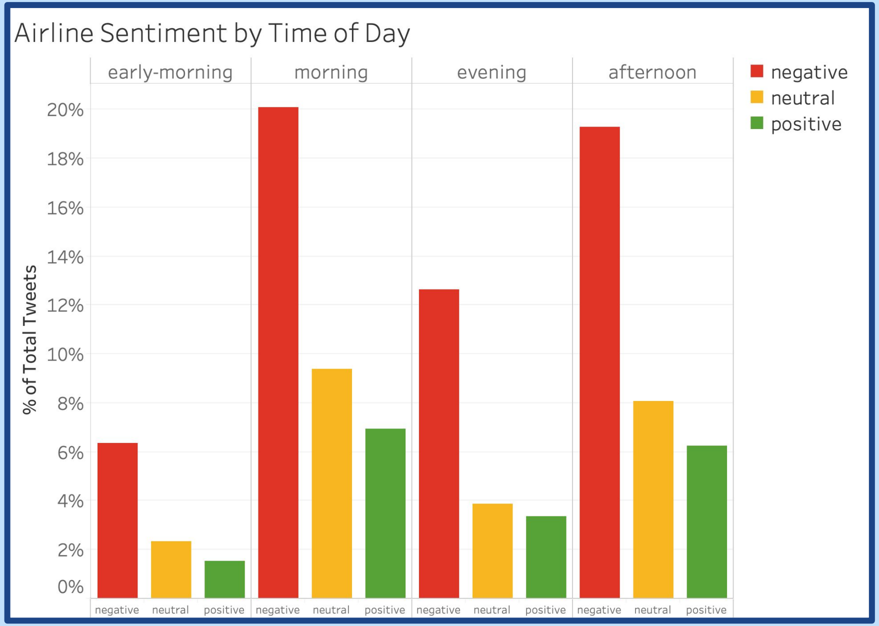
What are the Negative Tweets About?



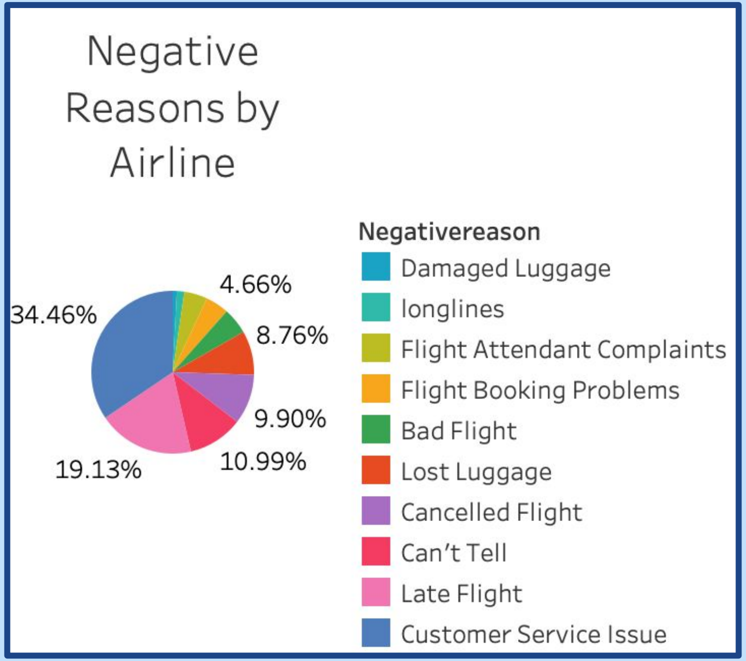
Key words found in negative tweets:

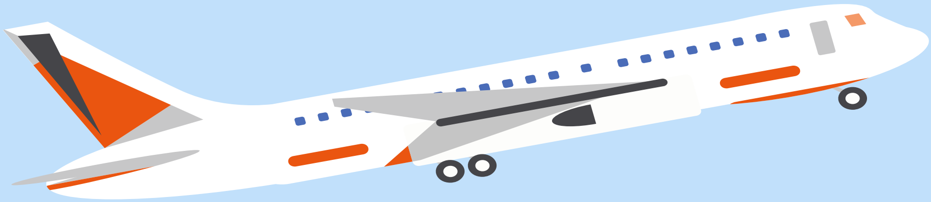
- Cancelled
- Thanks (sarcastic?)
- Service
- Help
- Now
- Delayed
- Customer

What can Airlines Improve upon?



See speaker notes.



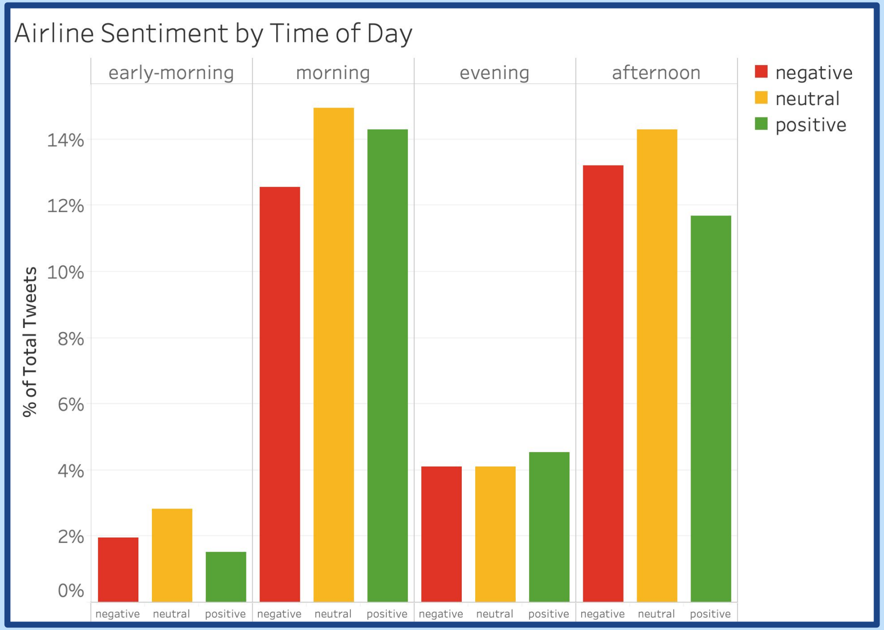


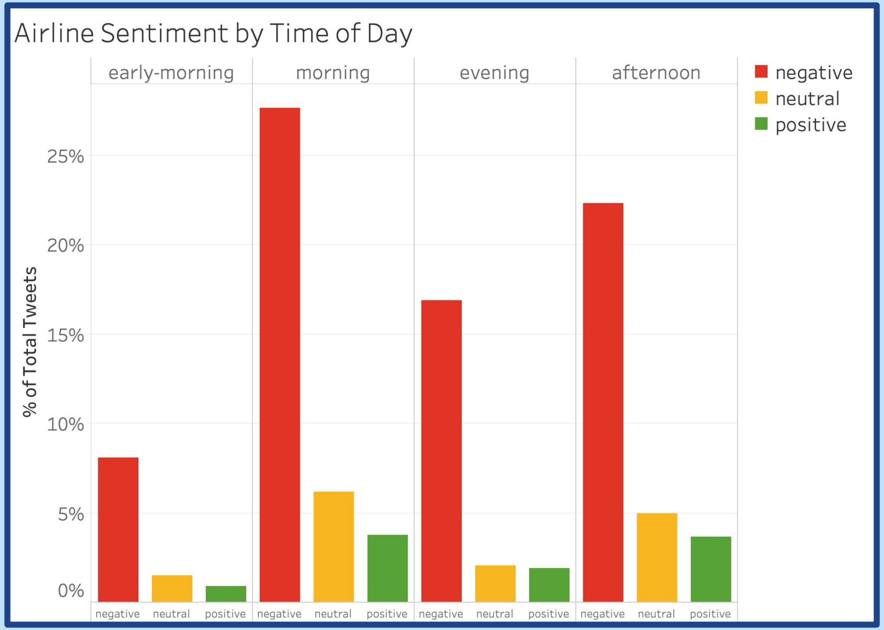
Looking at When Tweets Were Sent

Best: Virgin America

**See speaker notes.*

Worst: US Airways



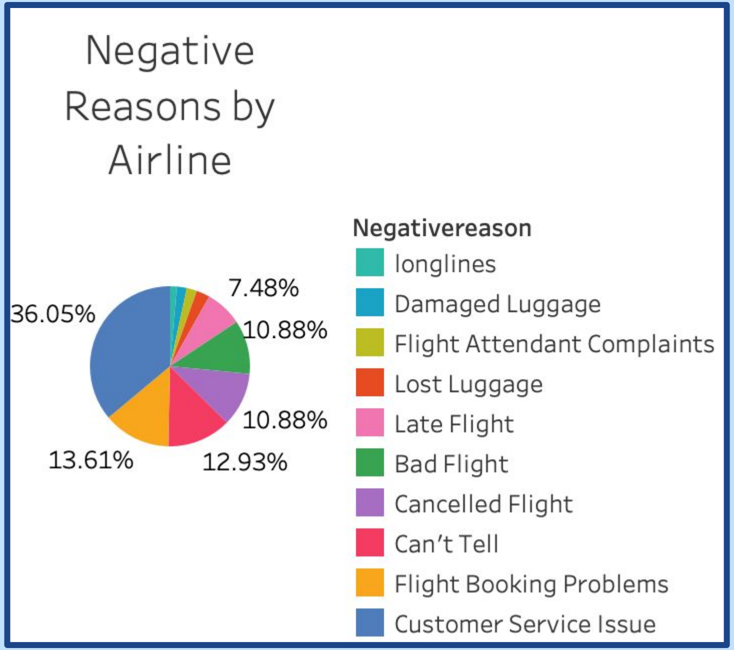


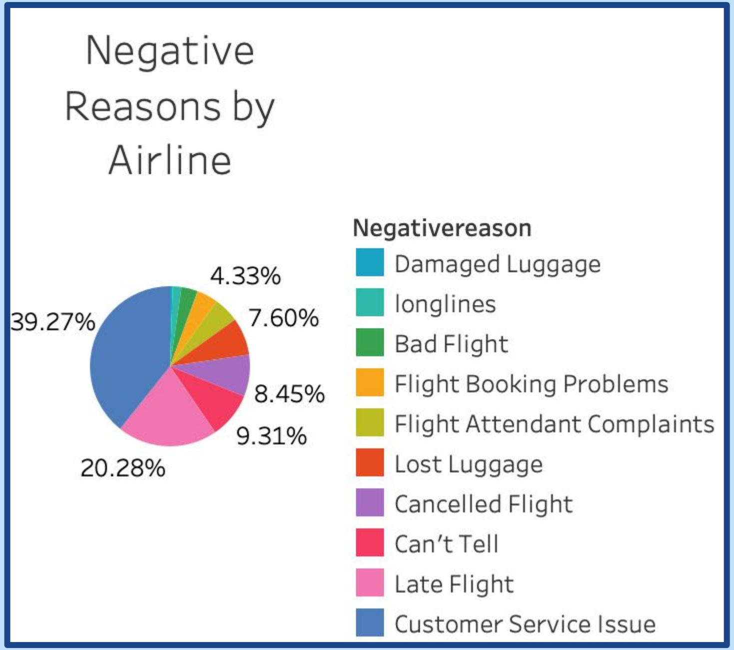
Reasons for Negative Tweets

Best: Virgin America

**See speaker notes.*

Worst: US Airways





Modelling

Confusion matrix:

	American	Delta	Southwest	United	US Airways	Virgin America	class.error
American	445	0	0	449	314	0	0.6316225
Delta	73	0	0	394	90	0	1.0000000
Southwest	169	0	0	313	241	0	1.0000000
United	286	0	0	920	341	0	0.4053006
US Airways	284	0	0	667	451	0	0.6783167
Virgin America	33	0	0	50	27	0	1.0000000

- I attempted a random forest model to predict the airline using the *Time of Day* and *Airline Sentiment* as predictors.
- This model is wildly inaccurate with a 100% error rate in three of the six airlines.
- This data had a clear use case that warranted a model.
- If there was more data, stock prices could be predicted using *Airline Sentiment* for each airline.

Appendix/Data Cleaning Expanded

- I removed airline_sentiment_gold and negativereason_gold because they were all NA.
- I removed tweet_coord, tweet_location, and user_timezone because, according to the Twitter docs, these features were user-generated which means they weren't necessarily accurate. For instance, you could add a location that wasn't where the tweet was sent from, or if the user was using a VPN, then the location would also not be accurate. From the samples, I looked at these three columns that didn't match up with each other. A side note on the user_timezone; this feature seemed fine with over 9,000 rows of seemingly okay data, but I didn't find this feature particularly useful without tweet_coord, and tweet_location.
- I feature engineered a time of day (ToD) column that translated the tweet_created column and changed it to be four dummy variables: early-morning (00:00:01-06:00:00), morning (06:00:01-12:00:00), afternoon (12:00:01-18:00:00), and evening (18:00:01-24:00:00)
- Next, I created three different datasets classified by their airline_sentiment
- For positive and neutral sentiments I removed negativereason and negativereason_confidence, as these columns don't relate to the data and are NA's.
- Lastly, to clean my data I got rid of airline_sentiment_confidence if the values were less than 0.5. If the airline_sentiment was negative, I got rid of the negativereason_confidence that were less than 0.5. I only wanted Tweets where the sentiment was at least 50% accurate.