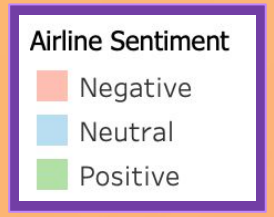
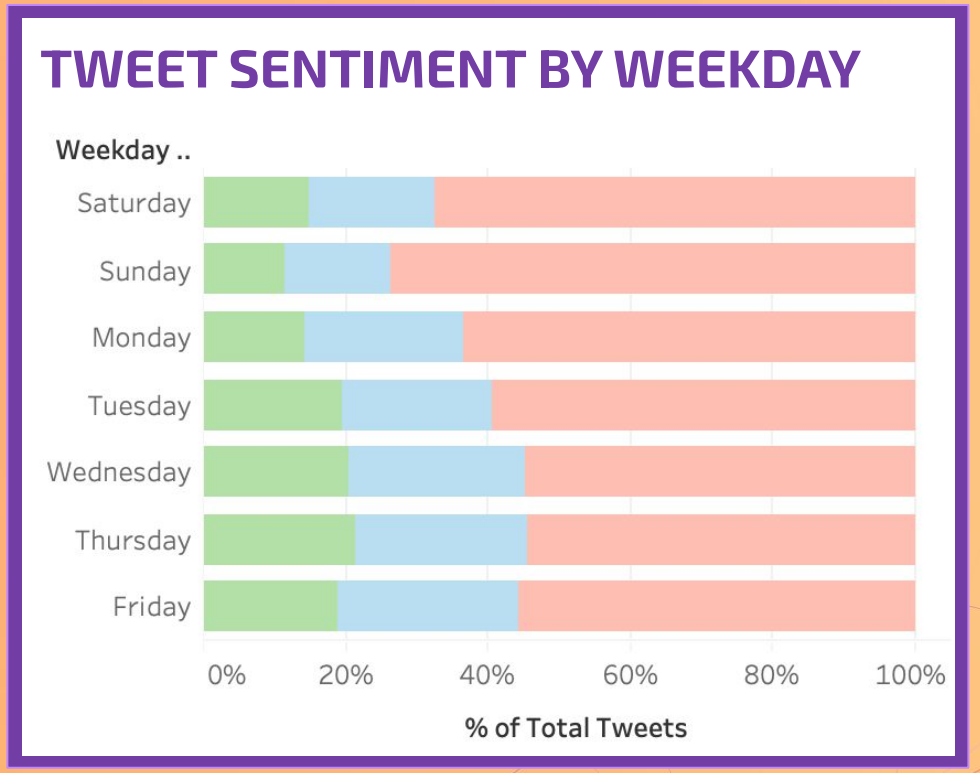
TWITTER AIRLINE SENTIMENT ANALYSIS

By Cady Stringer

UNDERSTANDING THE DATA

- A sentiment analysis was conducted on the text of 14,640 tweets to determine tone/attitude.
    - Each classification has **airline_sentiment** (positive, negative, or neutral) and an **airline_sentiment_confidence** level between 0 and 1 that expresses how confident the model was in that sentiment prediction.
- For tweets with negative sentiment, the analysis identified the **negativereason** for the tweet, like “Customer service” or a “Late flight,” and a corresponding **negativereason_confidence** level for that prediction.
- For some entries, **tweet_location**, **user_timezone**, and **tweet_coord** (coordinates) are included. These are user-entered and not helpful in analysis. For example, some locations entered don’t exist, like “Somewhere celebrating life.”
- Other variables include the **airline** tagged, **tweet_created** (date/time the tweet was posted), the **text** of the tweet, **retweet_count**, the user’s **name,** and a unique **tweet_id.**
- The variables **airline_sentiment_gold** and **negative_reason_gold** were mostly NAs.





- **Insight**: Saturdays, Sundays, and Mondays have the most negative sentiment tweets across all airlines
- **Analysis**: weekend travel likely causes customers to have short tempers, which could explain why their tweets are more negative
- **Solution**: to combat this, airlines should focus on excellent customer service on weekends and Mondays with the assumption that customers are more likely to air their grievances on twitter and create bad publicity

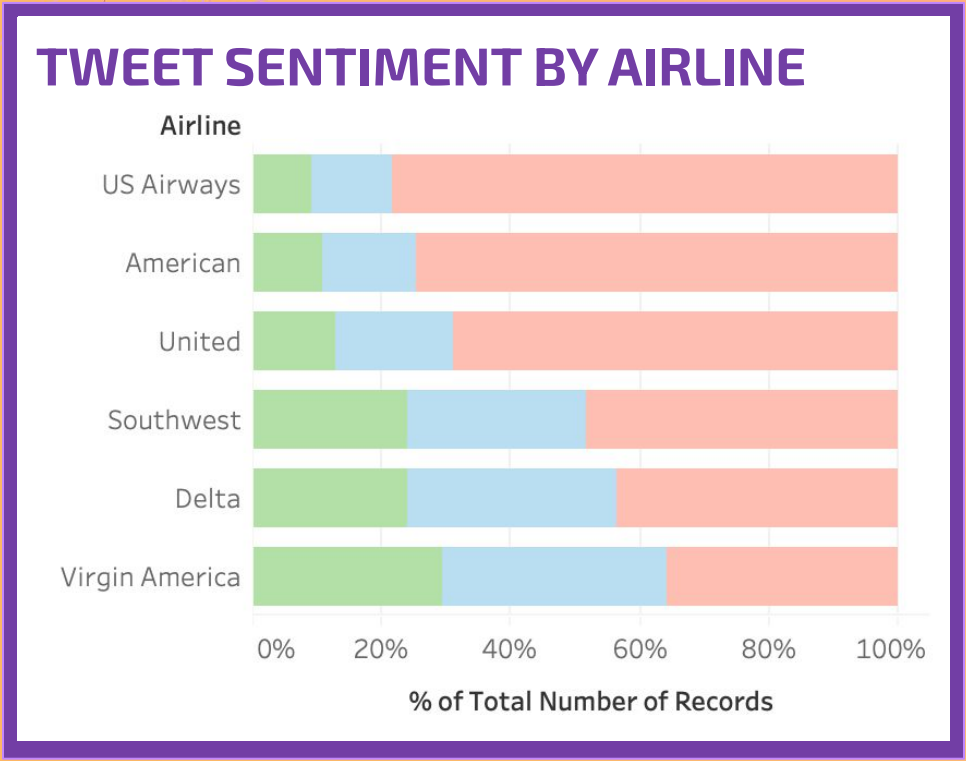
Most Negative Sentiment:

- US Airways
- American
- United

Least Negative Sentiment:

- Southwest
- Delta
- Virgin America

Solution: all airlines should identify the main reasons for tweets with negative sentiment and focus on improving these areas, it's especially important for the bottom 3 airlines.

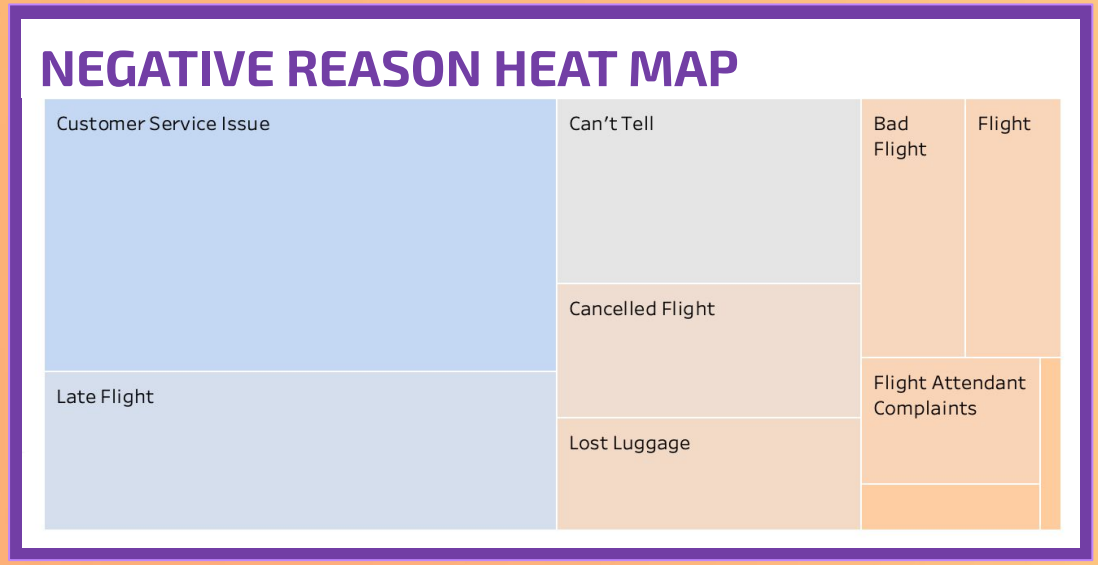


Most Common Negative Reasons Across All Airlines:

- Customer Service
- Late Flight

"Can't Tell" is the fault of the sentiment analysis. Airlines should treat it as an "Other" category and disregard it until a more effective sentiment analysis is conducted, and the category is eliminated or reduced.

Solution: to improve sentiment, airlines should identify the areas that most commonly cause negative sentiment tweets, and funnel resources into improvement in those categories.



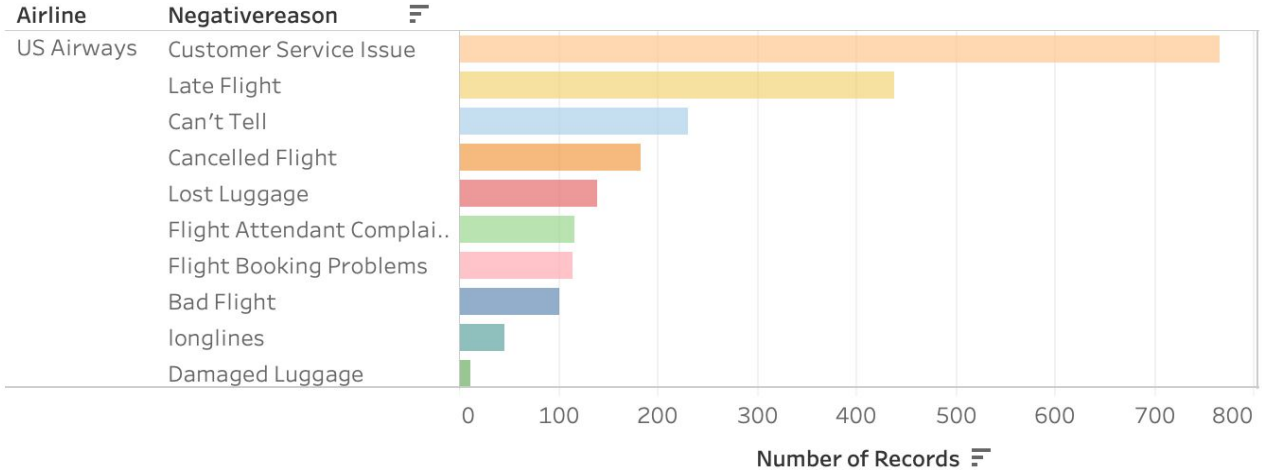
Airlines should also **compare** their most common negative reasons to the most common ones across all airlines to see ways they can stand out and improve their service compared to their competitors.

Most Common Negative Reasons:

- Customer Service
- Late Flight

Solution: to improve sentiment, US Airways should focus time and resources on training their customer service team, and preventing late flights or keeping customers informed about causes of late flights (especially if they're beyond the airline's control) and easily accessible solutions and options for other flights

NEGATIVE REASON: US AIRWAYS



**See attached packaged Tableau workbook for interactive plots and to filter visualizations by airline.*

MODELING DATA PREPARATION

- **Class imbalance problem:** there are about 4000 negative sentiment tweets, 1100 positive tweets, and 1500 neutral tweets. This class imbalance would lead a model to predict all or mostly negative tweets, instead of identifying patterns or finding relationships between the predictors and outcome.
- **Solution:** upsample the minority classes, neutral and positive.
    - This entails a randomly resampling (with replacement) the underrepresented classes to create additional rows to train the model with.
    - This choice is better than downsampling because it allows us to mitigate imbalance while still maintaining as much data as possible.

RANDOM FOREST CLASSIFICATION: VARIABLE IMPORTANCE

Purpose: use the variable importance feature to identify the most important variables for predicting tweet sentiment, to see whether sentiment is airline-dependent or if there's a different factor that influences sentiment.

MODEL

Number of trees: selected 300 decision trees to minimize error and maximize computational efficiency.
Predictors used: airline, airline_sentiment_confidence, retweet_count, tweet_created, & user_timezone.
Predicting: multinomial prediction of negative, positive, or neutral sentiment.

RESULTS

Model accuracy: ~44% error rate, so insights should guide further analysis but not taken at face value.
Insights: plotting variable importance reveals that airline_sentiment_confidence and airline were the most important.
So, airline does impact sentiment, and airlines should pay close attention to their negative tweet reasons and improve those areas.

LASSO CLASSIFICATION: VARIABLE IMPORTANCE

Purpose: a Lasso model penalizes coefficients (which symbolize the relationship between our predictors and outcome) in a way that sets unimportant variables to 0, and these variables likely have little to no impact on the outcome, tweet sentiment. A Lasso model will identify which predictors are the most important, which will help us identify which have a relationship with sentiment and thus customer satisfaction.

MODEL

Predictors used: airline, airline_sentiment_confidence, retweet_count, & tweet_created
Prediction: binary prediction of negative or positive sentiment

RESULTS

Model accuracy: ~60% accurate predictions.
Insights: Lasso selected airline_sentiment_confidence, airline, and tweet_created as the most important variables for predicting sentiment.
The **retweet_count** coefficient was set to 0, which means that retweet_count doesn't have enough of an impact on prediction accuracy to use in the model.
Airlines should be wary of making decisions based on insights from a model that predicts only slightly better than a coin flip.

https://linkedin.com/in/cadence-stringer/

https://github.com/cadystringer/

THANKS!

Please see attached Tableau Packaged Workbook for interactive plots and to filter visualizations by airline, and the attached RMD file for detailed data cleaning, preparation, and analysis.

APPENDIX: DATA CLEANING AND PREPARATION

- Removed **airline_sentiment_gold** and **negativereason_gold** columns because they're almost entirely NAs
- Removed **tweet_cord** because it's over 93% NAs
- Removed **tweet_location** column because although it's only 33% NAs, it includes many locations that aren't real locations, and the same locations are written different ways, like "New York, New York," "NYC," and "new york, new york." This would create too many factor levels for meaningful modelling
- Changed date/time column **tweet_created** to be only date, to limit the number of factor levels for modelling
- Proceed with caution for **negativereason** and **negativereason_confidence** columns: if the tweet sentiment is positive or neutral, those columns are filled with NAs. Dropping NAs from the whole dataset would remove all positive and neutral entries.