APPENDIX (DATA ASSESMENT)

-Initial Observations:

* 14,640 rows, 15 columns:
* 15 columns:
  + Tweet ID:
    - Data Type: Nominal Data/Unique Identifier
    - Action: Kept for purpose of having a unique identifier
  + Airline Sentiment:
    - Data Type: Ordinal Data because of levels related to one another
    - Action: Kept for data grouping
  + Airline Sentiment confidence:
    - Data Type: Finite value data ranging from 0 to 1 (range data)
    - Action: Kept for data grouping
  + Negative reason:
    - Data Type: Nominal Data/Labels
    - Action: kept for data grouping
  + Negative Reason confidence:
    - Data Type: Categorical data ranging from 0 to 1 (range data)
    - Action: kept for data grouping
  + Airline:
    - Data Type: Nominal Data/Labels
    - Action: kept for data grouping
  + Airline Sentiment Gold:
    - Data Type: Unknown/NA
    - Action: Dropped because all values were blank
  + Name:
    - Data Type: Nominal
    - Action: Dropped because irrelevant to intended analysis. Didn’t provide insight on gender
  + Negative Reason gold
    - Data Type: Unknown/NA
    - Action: Dropped because all values were blank
  + Retweet Count
    - Data Type: Discrete Data
    - Action Dropped: Retweet disparity isn’t wide enough, doesn’t reveal insight for intended analysis
  + Text
    - Data Type: String
    - Action: Originally kept but then substituted it for a text-cleaned version in order to observe how confidence levels were made more clearly
    - New column: fix\_text(cleaned version)
      * Removes special characters, emojis to strip to pure text
      * Used R code from https://towardsdatascience.com/text-mining-with-r-gathering-and-cleaning-data-8f8b0d65e67c
  + Tweet Coordinates
    - Data Type: Categorical/Nominal
    - Action: Dropped, too many N/As to be deemed useful. Close to 90% N/A
  + Tweet Created:
    - Data Type: Date/Nominal
    - Action: Dropped, only recorded for month Feb 2015, couldn’t observe trends
  + Tweet Location:
    - Data Type: Nominal/Categorical
    - Action: Dropped, data wasn’t reliant enough with formatting of location. Some location modified and not accurate
  + User Time zone:
    - Data Type: Nominal/Categorical
    - Action: Kept for data grouping, but used to create new variables: User Region. Used Vlookup function to group up time zone into the following regions: North America, Caribbean, South Central America, Africa, Oceania, East Asia, Western Europe, Eastern Europe/Central Asia, Middle East, South Asia. These groupings were based on inspiration from Homeland Security/United Nations region categorizations.
  + New columns:
    - Grade\_setiment\_confidence: new column that categorizes confidence level to : Perfect, Strong, Ok, Weak. Refer to page: for more detail
    - Grade\_negative\_confidence: new column that categorizes confidence level to : Perfect, Strong, Ok, Weak. Refer to page: for more detail
* Basic Data Interpretation (Refer to Excel Sheet (Pivot Tables for more info):
  + Airline Sentiment:
    - ~63% negative
    - ~21% neutral
    - ~16% positive
  + Airline Sentiment Confidence Average:
    - ~93. % Negative
    - ~82.3% Neutral
    - ~87.2
    - Confidence Level was slightly more accurate for negative tweets. This may be due to the easier interpretations of negative statements since negative words might be easier to sport
  + Negative Reason:
    - Top 4 Negative Reason based Tweets:
      * 1) Customer Service 31.71%
      * 2) Late Flight 18.14%
      * 3) Can’t tell 12.97%
      * 4) Cancelled Flight 9.23%
        + Customer service most likely highest % because of the broad use of the term, could relate to multiple things
  + Negative Reason Confidence:
    - Average of Confidence level: .63
    - Lowest Confidence Average Rating: Longlines (.194)
    - Highest Confidence Average Rating: Lost Luggage (.813)
    - Average negative reason confidence scored much lower than sentiment confidence. Sentiment is easier to discern between positive and negative because of buzzwords, ex: good and bad. Unlike negativereason.
  + Airline:
    - Highest percentage of negative tweets by Airline:
      * 1) United (17.98%)
      * 2) US Airways (15.46%
      * 3) American (13.39%)
    - Distribution of Airlines in Tweets:
      * United 26.11%
      * US Airways 19.90%
      * American 18.85%
      * Southwest 16.53%
      * Delta 15.18%
      * Virgin America 3.44%
    - User Region:
      * Majority of data is concentrated in North America:
        + 56.32%
      * N/A:
        + 32.92%
      * Central America:
        + 5.1%

Charts and Graphs:

* Figure 1: Was created to show how tweets mentioning different airlines were broken up by sentiment

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Count of airline\_sentiment** | **Column Labels** |  |  |  |  |  |  |
| **Row Labels** | **American** | **Delta** | **Southwest** | **United** | **US Airways** | **Virgin America** | **Grand Total** |
| negative | 13.39% | 6.52% | 8.10% | 17.98% | 15.46% | 1.24% | 62.69% |
| neutral | 3.16% | 4.94% | 4.54% | 4.76% | 2.60% | 1.17% | 21.17% |
| positive | 2.30% | 3.72% | 3.89% | 3.36% | 1.84% | 1.04% | 16.14% |
| **Grand Total** | **18.85%** | **15.18%** | **16.53%** | **26.11%** | **19.90%** | **3.44%** | **100.00%** |

* Figure 2: Was created to filter through each airline and examine percentage distribution tweet regarding negative reasons. This gives good insight if an airline is contributing more to a specific negative reason comparatively to other airlines.

|  |  |  |  |
| --- | --- | --- | --- |
| airline | (All) |  |  |
|  |  |  |  |
| **Row Labels** | **Count of negativereason** | **% of Total Negatitve Reason** | **Average of negativereason\_confidence** |
| Bad Flight | 580 | 6.32% | 0.631731379 |
| Can't Tell | 1190 | 12.97% | 0.629725798 |
| Cancelled Flight | 847 | 9.23% | 0.783095632 |
| Customer Service Issue | 2910 | 31.71% | 0.78005433 |
| Damaged Luggage | 74 | 0.81% | 0.733432432 |
| Flight Attendant Complaints | 481 | 5.24% | 0.659639293 |
| Flight Booking Problems | 529 | 5.76% | 0.606796597 |
| Late Flight | 1665 | 18.14% | 0.768906727 |
| longlines | 178 | 1.94% | 0.594075843 |
| Lost Luggage | 724 | 7.89% | 0.813018508 |
| (blank) |  | 0.00% | 0 |
| **Grand Total** | **9178** | **100.00%** | **0.63829828** |

* Figure 3: Was created to examine how accurate the sentiment analysis is on average for tweets related to major airline.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Average of airline\_sentiment\_confidence** | **Column Labels** |  |  |  |
| **Row Labels** | **negative** | **neutral** | **positive** | **Grand Total** |
| American | 0.944954694 | 0.826 | 0.8823 | 0.917351939 |
| Delta | 0.902202199 | 0.829 | 0.8671 | 0.869878263 |
| Southwest | 0.920532968 | 0.826 | 0.8861 | 0.88651595 |
| United | 0.933382719 | 0.81 | 0.856 | 0.900877682 |
| US Airways | 0.945713699 | 0.822 | 0.8597 | 0.921578441 |
| Virgin America | 0.901733149 | 0.838 | 0.888 | 0.876086111 |
| **Grand Total** | **0.933365319** | **0.823** | **0.872** | **0.900168852** |

* Figure 4: An insight to see how tweets in the dataset are being broken up by negative reason. As we can see, customer service issues are the most frequent/most important.

Table

Description automatically generated

Chart

Description automatically generatedChart, line chart

Description automatically generatedFigure 5:

|  |  |  |
| --- | --- | --- |
| Perfect Confidence | x = 1 | |
| Good Confidence | .7 < x < 1 | |
| Ok Confidence | .5 < x <.7 | |
| Weak Confidence | .3 < x < .5 | |
| Unacceptable Confidence | x < .3 |

|  |  |
| --- | --- |
| Perfect Confidence | x = 1 |
| Strong Confidence | .7 < x < 1 |
| Good Confidence | .5 < x <.7 |
| Weak Confidence | x < .5 |

I wanted to create a categorization for the confidence grades of both negative reason and sentiment. I did this by creating a visualization where I can see major cutoff points. I chose these cutoff points and created new columns: Grade\_negative\_confidence and Grade\_negative\_confidence. These columns used if statements. Here are the formulas:

Grade\_sentiment\_confidence:

=IF(C2=1,"Perfect",IF(AND(C2>0.5,C2<0.7),"Good",IF(AND(C2>0.7,C2<0.99),"Strong","Weak")))

Grade\_negative\_confidence:

=IF(E2=1,"Perfect",IF(AND(E2>0.7,E2<1),"Good",IF(AND(E2>0.5,E2<0.7),"Ok",IF(AND(E2>0.3,E2<0.5),"Weak","Unacceptable"))))

Confidence Grade Graphs

Figure 6:

I used the categorizations I made in Figure 5 to graph each airline based on what percent of respective airline tweets fall into each category of confidence grade. For both sentiment confidence and negative reason confidence.

Figure 7: **Refer to Pivot Tables on Excel Sheet. Image is easier to read**

This image represents the sheer dominance of North American based tweets in data set, with next majority laying in N/A based tweets. If these airlines were international, there may be better diversity. Motivation was to illustrate the global market each US based airline has.

Table

Description automatically generated

**Airline Strength Assessment:**

* Purpose: To create an assessment that compares the relative strength of airlines through social media reactions on Twitter.
* Metrics:
  + Negative Reason Complaints: To examine whether airlines are contributing proportionately more to negative reason-based tweets comparatively to the % total of all airlines
    - For example, if the % Total of Negative Reason for customer service-related issues is 31% of all negative reason tweets, Delta Airline would be rewarded if their OWN proportion of customer service-related issues contributed LESS. (ex:20.84%).
  + Market Hold: To examine twitter engagement globally. This rewards airlines with more tweets in the dataset.
    - For example, from a scoring metric, United Airlines would score the highest in Market hold in the United States as their mentioned tweets make up the highest percentage in the dataset for US (14.35% / 56.32%).
  + Sentiment Strength: To examine and reward airlines that have proportionately fewer negative tweets and proportionately higher positive tweets. Treat neutral tweets as positive but multiply outcome by .5 to lessen magnitude.
* How exactly were strength ratings made for each Airline and factor?
  + Negative Reason Rating:
    - Step 1: Compare Pivot tables of negative reason distribution between

Graphical user interface, application, table, Excel

Description automatically generated“all” airlines through filter function vs negative reason distribution of an individual airline. Then take the difference between two (‘all’ airlines – individual airline to see if the individual airline has proportionally more or less negative reason tweets than the holistic view of all airlines.

* + - Step 2: Take the difference and multiply it by importance number. I assigned an importance to each category in negativereason. Customer Service was the most frequent issue appearing in the airline industry, so it got the highest importance number of 5. “Can’t tell” and “Late flight” followed suit with the next highest frequencies, so both received importance numbers of 3. The rest of the reasons were around single digit percentage frequency, and all earned the same importance number of 1.5
    - Step 3: Adjust “Difference \* Importance” by adding the negative reason confidence score respective to the following category in “negativereason” for each airline. I divided each negative reason confidence score by 10, as the values of “Difference\*Importance” were much smaller than the confidence score values.
    - Step 4: Sum all of the totals calculated for each negative reason and repeat process for each airline.
    - Step 5: Take average of all sum of totals for airlines and compare to each respective sum of individual airline by dividng airline sum by total average sum. (Ex: Virgin America Sum / Avg Total Sum for All Airlines).
    - *How strength ratings were made:*
      * If (Airline Sum / Total Average Sum) = ~ 1(around), Score: 3
      * If (Airline Sum / Total Average Sum) is the highest among all other airlines, Score: 5. Does not apply if highest airline grade is barely higher, Ex: .01%
      * If (Airline Sum / Total Average Sum) is significantly over 1.0, Score: 4
      * If (Airline Sum / Total Average Sum) is significantly below 1.0, Score: 2
      * If (Airline Sum / Total Average Sum) is the lowest among all other airlines, Score:1. Does not apply if highest airline grade is barely lower, Ex: .01%
      * DO PROCESS FOR EACH AIRLINE TO GET STRENGTH SCORE
    - Photos of Virgin America Calculation: Table

      Description automatically generated
    - Photo of Strength Rating Designation:Table

      Description automatically generated
* Market Hold Rating:
  + Step 1: Examine Pivot Table of all “user\_region” and inspect what percentage of airlines contribute to the total % of Tweets based on each region. Then for each row pertaining to each airline, take the percentage of a region and divide it by the total percentage of that same region. Ex: Delta(North America) = 9.81%, Grand Total(North America) = 56.32% -> 9.81% / 56.32%. Repeat process for each airline and region to get % of Total Tweets.
    - Table

      Description automatically generatedPivot Table:
  + Step 2: Create Importance Score for each region. North America was given the highest importance score of 5 because of sheer volume of North American based tweets, this volume makes sense since these are American Airlines. Central America was given an importance score of 3 because it proceeded as the 2nd most populous twitter user region. Additionally, all European countries were given a score of 3 because of the favorable travel laws and international relations between the USA. The rest of the countries were given ratings of 2 because of their small % in dataset and N/A values were deemed a rating of 1.
  + Step 3: Multiply ‘% of Total Tweets’ by Importance Score for each region. Repeat for every airline.
  + Step 4: Sum up all ‘% of Total Tweets \* Importance Score for each to get total sum for each airline.
  + Step 5: Strength Assessment. Identical to Step 5 from the Negative Reason Rating portion of the assessment 9

-Virgin America Example:

Table

Description automatically generated

Strength Rating Distribution:

Table

Description automatically generated

* Sentiment Strength Rating:
  + Process is exactly same as the Negative Reason strength portion of assessment. Compare pivot tables of a filtered sentiment information for a specific airline vs the total percentage.
  + Side Notes:
    - Importance Score was -1 for negative differences in order to punish positive values for having more negative tweets than the aggregate %. Neutral was given .5 importance, still wanted to reward engagement but keep in mind uncertainty of neutral tweets.
    - Added average of sentiment confidence and divided by 10 to provide reliability to sum
* Photos:

Graphical user interface, application, table, Excel

Description automatically generated

Table

Description automatically generated

COMPLETED STRENGTH ASSESMENT TABLE

A picture containing application, table

Description automatically generated

* Importance Weight Designations:
  + Negative Reason Complaint and Sentiment (.4 and .4):
    - Although the sentiment importance gives a broad measure of the interactions twitter users had with the airline, I did not rank it higher than the negative reason measure. It’s normal instinct to write publicly about a service if it goes wrong rather than right. Airlines doing a good job is expected of the service more than rewarded. Both measurements ranked the same as I felt the negative reason strength score gives good insight on airline performance in tangible/specific customer service issues where airlines can improve.
  + Market Hold (.2):
    - This had the lowest weight because these Airlines are American. The discrepancy between North American related tweets are much higher than any other regions comparatively. However, it is a good insight to which Airline dominates the North American market with twitter engagement. This market hold is not considered 100% good publicity, since most engagement is negative, but still shows level of social media reach.
* Rankings:
  + 1) Virgin America - 4
  + 2) United – 3.8
  + 3) Delta – 3.6
  + 4) Southwest - 3
  + 5) American – 2.2
  + 6) US Airways -1.8
* Results:
  + Virgin America performed the best because:
    - Solid performance throughout all metrics
    - In Negative Reasons, the aggregate contribution of all airlines regarding Late Flights accounted for 18% of negative reason tweets. Proportionately, Virgin America had only half as many Late Flight complaints in their respective tweets. They also proportionately contributed 60% less Lost Luggage tweets than the aggregate total
    - Performed well with the sentiment score. Proportionately, Virgin America was made up of nearly 2x more positive tweets than the total proportion and nearly 2x less negative tweets than the total proportion. It was surprising to see users to mention positive things about the airline comparatively to other airlines.
  + Concern about Virgin America’s victory:
    - Airline only makes up 3.44% of dataset
    - Sample size might not be big enough to show true colors of airline
    - In future, would up-sample Virgin Airline but would need to build model that would predict likely tweets according to existing 3.44% typical behavior
  + United performed second best because:
    - Dominated in terms of market hold. No surprise because United makes up 26.11% of the mentioned dataset tweets. Additionally, United made up the most North American user mentioned tweets by a significant amount, which helped their score tremendously in market hold
    - United suffered, however, when it came to sentiment proportionally, they had higher % of negative tweets than the aggregate percentage and lower % of positive tweets than the aggregate percentage
    - If the weight was higher for market hold, United would have won, however I stand by the reasoning for deeming this a .2 weight for the strength assessment
  + US Airways performed the worst because:
    - US Airways performed poorly through all categories
    - They managed to be average in market hold but that was mostly due to having the second most mentioned tweets in the dataset (20%)
* Takeaways:
  + Airlines should seek the top 2 scoring candidates in terms of how to improve customer service to allow for more positive tweets and less complaints on twitter. This might be through rewards programs, internal system technology to allow for minimal late flights, or social media marketing to appeal to more consumers globally.
  + Consulting firms can use these grades to investigate and research practices of Virgin America and United to offer best practices to other airlines
  + Must be cautious with Virgin’s #1 rank, as they account for a small slice of the data.
    - Look for a method to up-sample Virgin America through predictions.
  + Major Limitation affecting Virgin America: If there were only a couple of data points in, for example Africa, and Virgin America had one of those data points. It could significantly help its market hold score even though it had one or two more tweets in that region than competitors. The model awards ‘% of’ rather than Count, which can greatly increase score. This would work better if more data points were in the other regions.