Airline Customer Satisfaction

Benjamin Cowan

Professor: Dr. Landowski

IST 652 - Final Project

Due: 12/6/2020

Data Source and Description, including preprocessing:

Structured data: an excel spreadsheet for airline customer satisfaction, supplied by IST 687. The spreadsheet includes 129890 rows and 28 columns; each row represents a customer at the airline while columns represent general demographic information – customer’s age and gender - and airline flight data – a customer’s airline satisfaction score, airline status, flight date and delays - to name a few. Through exploration of the airline data, I decided to analyze potential fields that predict an airline customer’s satisfaction score.

Semi-structured data: I used twitter data, in the form of JSON, to analyze tweets for large passenger airline manufacturers – Boeing and Airbus – as well as Delta, United, and Southwest airlines. Twitter data included text data, as well as favorites, friends, and followers count.

I chose to analyze those twitter fields because they are possible predictors of airline customer satisfaction.

Preprocessing: Structured data had two errors, specifically in the satisfaction variable, that needed to be removed before importing into python. Through MongoDB, I checked to make sure my hashtags, that I used to retrieve tweets, were returning relevant data. If I used a hashtag that was not specific, such as “#Delta,” it would return irrelevant tweets. To avoid irrelevant tweets, I used specific hashtags. I collected 300 tweets, for each manufacturer and airline, because the data for specific hashtags was scarce.

Methods of analysis, including analysis questions, data fields and expected results:

Structured data question:

* + Using data summaries, what are some potential predictors of airline customer satisfaction?

Structured data fields:

* + Satisfaction score – customer's satisfaction with the airline, ranging from 1 to 5. 5 being highly satisfied
  + Airline status – statuses rank from Blue, Silver, Gold, and Platinum. Platinum being the highest status, Blue being the lowest status.
  + Gender – customer's gender.
  + Age – customer's age.
  + Flight date – date of the customer’s most recent flight with the airline.

Structured data comparison:

* + Compare mean satisfaction scores for each airline status.
  + Compare mean satisfaction scores, filtered by airline status and gender.
  + Compare mean satisfaction scores, filtered by airline status and age.
  + Compare mean satisfaction scores, filtered by day of the week.

Semi-structured data question:

* + Using data summaries, what twitter data could an airline potential use to predict customer satisfaction?

Semi-structured data fields:

* + Text data– the text in each airline tweet.
  + Followers – for a given tweet, how many followers the user has on twitter.
  + Favorites– how many favorites a tweet received.
  + Friends – for a given tweet, how many friends the user has on twitter.

Semi-structured data comparison:

* + Compare mean followers, favorites, and friends for each airline manufacturer.
  + Compare sentiment analysis, using airline tweet texts, for each airline – Delta, United, and Southwest.

Structured and Semi-structured expected results:

* + Expected to create summary statistics for each comparison so I could identify differences and potential predictors between filtered customer demographics as well as airline tweet data.

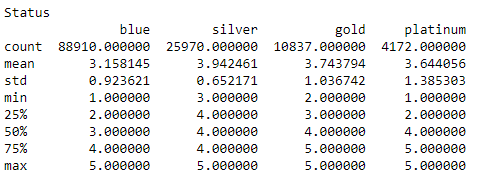
Overall Program Descriptions:

The structured section of my program begins by reading in the airline satisfaction spreadsheet, identifying the appropriate lines and replacing any missing data. Afterwards, each line is parsed into a dictionary and appended to a single list. The new list, containing dictionaries, is converted into a pandas data frame for analysis. Using the data frame, the program filters satisfaction scores for each airline status into new variables. Variables are combined into a new pandas data frame and the “describe.()” function is used to generate summary statistics to analyze. A similar filter is used to create three more summary statistics for satisfaction scores filtered by airline status and gender, weekdays, and airline status and age. Age was filtered into older and younger customer variables – older customers were above the mean age, while younger customers were below the mean age. Weekdays were generated by using the datetime function - dt.dayofweek - on the variable “Flight Date.” Once my four summary statistics were generated, they were sent to an output file to analyze. Also, three histograms were generated to give distributions of the data based on airline status, age, and gender.

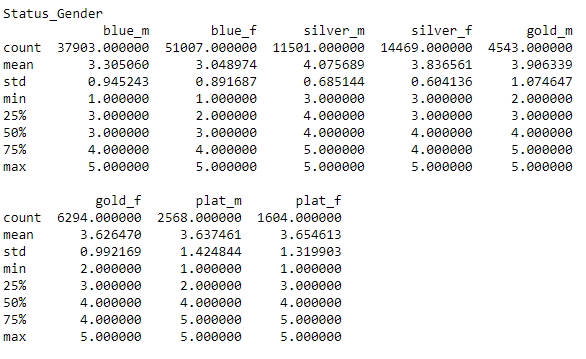
The semi-structured section of my program begins by gaining access to tweepy\_api, utilizing one of the course’s program “twitter\_login\_fn,” to access tweets from twitter. Another course program was used – run\_twitter\_simple\_search\_save.py - to import twitter login access, run queries on tweets, specifying tweet quantity and specific tweet hashtags, and save tweets into specific database collections. The program was run five times in an anaconda prompt to generate each MongoDB airline manufacturer and airlines database collections. Once the Delta, United, and Southwest collections were created, tweet texts were pulled into separate empty lists. A loop was run to calculate each text’s polarity, using textblob for sentiment analysis. Resulting polarities were put into pandas data frames and the mean polarity of each airline was calculated and stored. Similar to scraping tweet texts, the followers, favorites, and friends count for each manufacturer was appended to an empty list and brought into a pandas data frame to generate summary statistics. Both airline manufacturer’s summary statistics, as well as the three mean sentiment scores for the airlines, were exported to a csv file for analysis.

Results:

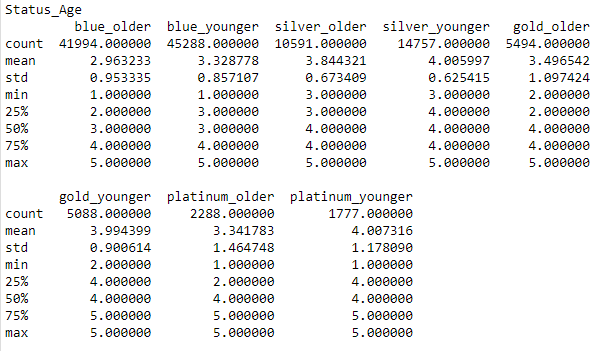
Structured data output:



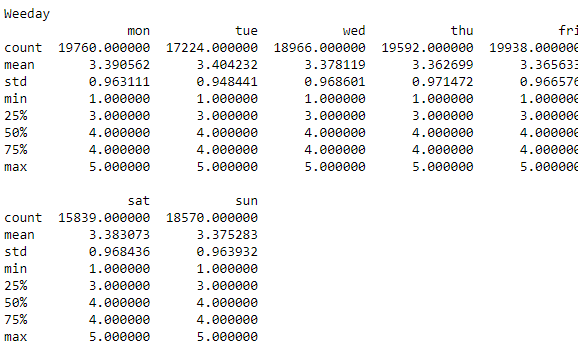
* Summary statistics for satisfaction scores between airline statuses: customers in blue status have the lowest mean satisfaction score and highest count while customers in silver status have the highest mean satisfaction score. This may suggest that silver status customers enjoy their benefits more than other statuses.



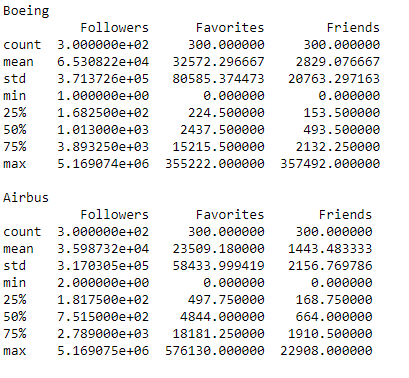
* Summary statistics for satisfaction scores between airline status and gender: there are more females in each status than males, except for platinum status. Females makeup the majority of the airline’s customers, however, males are, on average, more satisfied. There may be services males enjoy more than females in each airline status, excluding platinum.



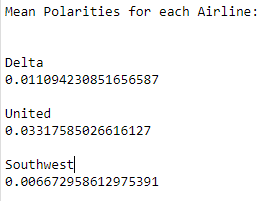
* Summary statistics for satisfaction scores between airline status and age: younger members - below 46 years old - are, on average, are more satisfied than older member - above 46 years old.



* Summary statistics for satisfaction scores between days of the week: satisfaction scores stay around 3.4 throughout the week.



* Summary statistics for airline manufacturer tweet data: tweets referring to Boeing have a much higher mean and spread of followers, favorites and friends than Airbus, this may suggest that Boeing has a bigger presence on twitter than Airbus.



* Mean output from tweet text sentiment analysis: Southwest had the lowest mean sentiment while United had the highest sentiment. This could indicate people are more satisfied with United airline’s services than Southwest airlines.

Conclusions:

Summary statistics, on the structured data, indicate airline status, gender, and age as potential predictors of customer satisfaction. A multinomial logistic regression could quantify each field’s significance for predicting customer satisfaction.

Summary statistics, on the semi-structured data, indicate differences in airline manufacturer’s popularity on twitter. If twitter popularity mirrors customer preference, customers may prefer to fly in a vehicle made by Boeing than Airbus. This preference could be a potential predictor of customer satisfaction. Differences in tweet sentiment analysis, between airlines, could be a predictor of customer satisfaction, as well; due to differences in sentiment analysis, individuals who have flown with United may enjoy their services more than someone who has flown with Southwest.