Syracuse University

School of Information Studies

Fall 2020

# **Flight Satisfaction Survey**

# IST687: Introduction to Data Science

A picture containing text, fireworks, outdoor object, dark

Description automatically generated

(Plot of all flight patterns in the Flight Survey dataset)

Instructor: Gary Krudys

Presenter: Alec Schneider afschnei@syr.edu

**Table of Contents**

**Introduction1**

Background2

Type chapter title (level 3)3

**Business Questions4**

**Data Cleaning & Feature Engineering4**

Data Cleaning2

Feature Engineering2

Categorization3

One Hot Encoding3

Additional Feature Engineering3

**Exploratory Data Analysis4**

**Modeling4**

Linear Regression5

Gradient Boosting Tree Classification5

Support Vector Machine Classification5

Naïve Bayes Classification5

**Conclusion4**

**Appendix4**

# Introduction

## Background

Airlines have been pioneers in the analytics community regarding pricing and route optimization, however they have long struggled with understanding customer satisfaction. It is possible they have struggled with this due to company bias, and lack of customer data from other airlines. To tackle this problem an industry funded customer survey gathered ~130,000 responses each with 25 attributes. I was tasked with analyzing this data thoroughly enough to provide actionable insights for the businesses to implement in their company strategies.

After cleaning and exploring the survey data, creating informative visualizations to assist in understanding the data, and applying machine learning algorithms to it, airlines are now closer to fully understanding what their customers value most, thus improving customer satisfaction. This will be a win for the airlines, as increased satisfaction is believed to increase demand of air travel, and the consumers who choose to travel on their aircrafts.

# Business Questions

1. Which airlines and origin cities have the highest rates of departure delays and cancellations?
2. Do vacation destinations have a higher customer satisfaction score?
3. Which demographics are most price sensitive?
4. Which airlines have the highest satisfaction scores?
5. Which factors affect the customer satisfaction score most?

# Data Cleaning & Feature Engineering

The analysis began with first understanding the tidiness of the dataset and understand what logic should be used to clean any naming conventions and missing values, before engineering any additional features. All rows and variables were used for the cleaning and feature engineering stages to ensure that the analysis would be non-biased and be strictly derived from the data itself. After all cleaning logic and feature engineering logic was decided upon, each step was coded as function that could be imported and reused, to save storage space by not having to create an additional file of the cleaned dataset. These functions were then combined in one function so the raw data could be transformed with one line of code inside an R script (see appendix XXX).

## Data Cleaning

Any column names that contained a space were substituted with an underscore to prevent any future difficulties in column extraction. In the raw data there were four columns that had missing data, Satisfaction, Departure\_Delay\_in\_Minutes, Arrival\_Delay\_in\_Minutes, and Flight\_time\_in\_minutes. A function, *fillNAs*, was created to apply the following logic to clean up the missing values (See appendix XXX).

1. Satisfaction column
   1. These observations with missing data were not associated with any other attributes
   2. It was decided to impute the missing values with the mean satisfaction score, excluding the missing values
2. 337 rows of data were removed from the data set as they were survey responses where the flight was not cancelled, but there was no arrival delay, flight time, or departure delay information. The decision was made to remove these rows since it as these data points could be reliably imputed with a mean value
3. Departure\_Delay\_in\_Minutes column
   1. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value
4. Arrival\_Delay\_in\_Minutes column
   1. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value
5. Flight\_time\_in\_minutes column
   1. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value

## Feature Engineering

In order to extract additional information from the existing variables, new variables were created to categorize discrete variables and one hot encode category variables so they can be used in a machine learning algorithm.

Categorization:

Four discrete variables were transformed into category variables for exploratory data analysis and possible one hot encoding for modeling purposes. The *category\_processing* function was created to handle this transformation (See appendix XXX).

1. No.\_of\_other\_Loyalty\_Cards
   1. Transformed into the cat\_loyalty\_cards variable with values of None (0), Low (0, 3], Medium (3, 6], and High (6, 12] to create an ordinal hierarchy of the number of loyalty cards a customer holds
2. %\_of\_Flight\_with\_other\_Airlines
   1. Transformed into the cat\_%\_of\_FLight\_with\_other variable with values of Low (0, 1/3], Medium (1/3, 2/3], and High (2/3, 1] to create an ordinal hierarchy for the percent of flights that customer takes with other airlines
3. No\_of\_Flights\_p.a.
   1. Transformed into the cat\_No\_of\_Flights variable with values of Low (0, 1/3], Medium (1/3, 2/3], and High (2/3, 1] to create an ordinal hierarchy of
4. Age
   1. Transformed into the cat\_Age variable with values of (-Inf, 20], (20, 30], (30, 40], (40, 50], (50, 60], (60, 70], and (70, Inf] to break up the age buckets to be of 10 years, excluding the small population of customers 20 and younger

One Hot Encoding:

One hot encoding is a form of data transformation that encodes categorical features as numeric features that take binary (0 or 1) values. One hot encoding allows these categorical columns to be used in machine learning models since their values become numeric. Using this methodology, I created n-1 columns for each categorical column below, where n is the number of unique categories. N-1 column methodology was used to remove any potential multi-collinearity in the dataset. The *dummy\_processing* function was created to one hot encode the following variables.

1. Airline\_Status
2. Type\_of\_Travel
3. Class
4. cat\_loyalty\_cards
5. cat\_No\_of\_Flights
6. cat\_Age

Additional Feature Engineering:

Since delay times that are of a higher proportion of the flight time may lead to decreased satisfaction, two additional variables were created to display these proportions.

1. DepDelayRatio
   1. Departure\_Delay\_in\_Minutes / Flight\_time\_in\_minutes
2. ArrDelayRatio
   1. Arrival\_Delay\_in\_Minutes / Flight\_time\_in\_minutes

# Exploratory Data Analysis

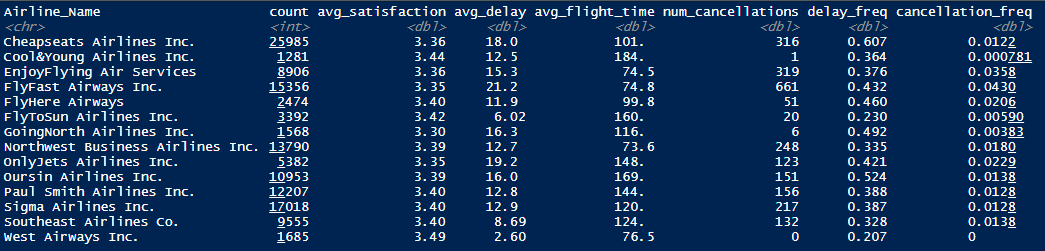
First, I wanted to explore our target variable, Satisfaction. The distribution of the satisfaction scores is left skewed, with most customers giving an above average score of four, with a mean score of 3.38:

Chart, bar chart

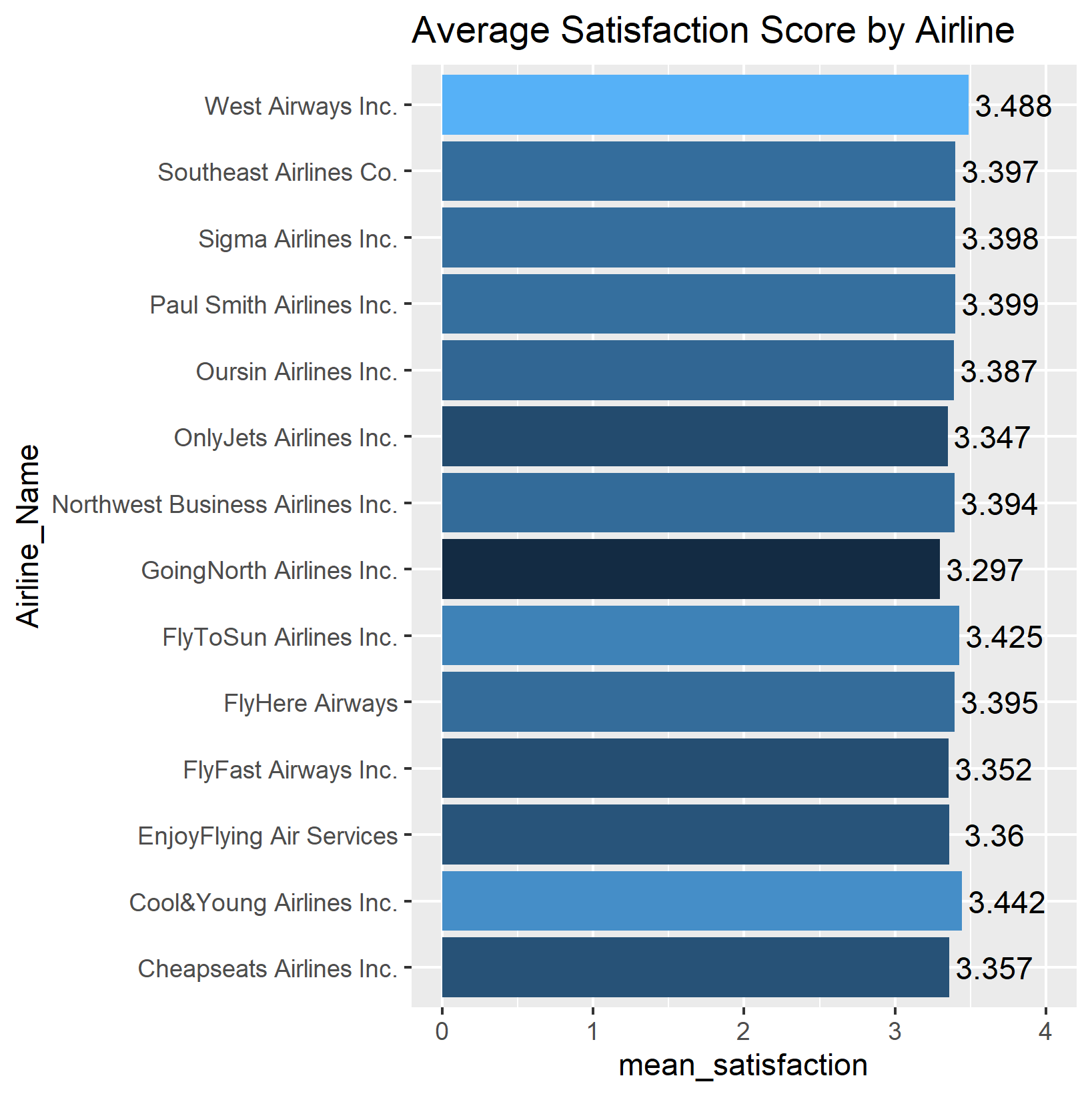
Description automatically generated

Second, it was important to analyze the individual airlines. When analyzing the mean satisfaction score across the whole population, large differences between the airlines would provoke further investigation to see which factors the higher performing airlines scored better in. The data was grouped by on the Airline\_Name column, and the following summary statistics were calculated:

* Average satisfaction score
* Average flight delay
* Average flight time
* Number of flight cancellations
* Departure delay frequency
* Cancellation frequency



All of the airlines had an average satisfaction score in a tight range, with no true outliers (See figure below). Looking further at the summary statistics, it is interesting to note that despite having them most amount of flights in the dataset, Cheapseats Airlines Inc. also had the highest frequency of delayed flights, with a margin of ~8% between them and the next closest airline Oursin Airlines Inc. The frequency of cancellations shows that FlyFast Airways Inc. scored the worst with a frequency of ~4%. It is very insightful to know that the airline with the best mean satisfaction score, West Airways Inc., had 0 cancellations on ~1700 flights, the lowest delay frequency at ~21%, and had the lowest average departure delay of only 2.6 minutes. Cool&Young Airlines Inc. had the second highest mean satisfaction score and score very well in the three aforementioned categories relative to peers.



Third, to answer our question regarding vacation destinations, there needed to be a mapping of vacation destinations. To accomplish this, all unique destination cities were written to a csv file using R, manually tagged as a vacation destination based on the criteria of being a known vacation spot (West Palm Beach/Palm Beach, FL) or a high entertainment area (Nashville, TN). This file was then read back into R and merged with the cleaned dataset. Before calculating the mean scores, business travel travel types were filtered out so they did not skew the data. The data was then grouped on whether or not the destination was a vacation destination, and calculated the count of records, mean satisfaction score, and standard deviation score for each group. The mean satisfaction score and standard deviations were almost exact, which meant that customers are not more likely to be satisfied with their travel experience if they are going to a vacation destination. This goes against the initial hypothesis.

> avg\_score\_vac\_dest <- df\_merged %>%

+ filter(Type\_of\_Travel != "Business travel") %>%

+ group\_by(Vacation) %>%

+ summarize(avg\_satisfaction = mean(Satisfaction),

+ count=n(),

+ std=sd(Satisfaction))

> avg\_score\_vac\_dest

# A tibble: 2 x 4

Vacation avg\_satisfaction count std

*<int>* *<dbl>* *<int>* *<dbl>*

1 0 2.75 22039 0.856

2 1 2.75 28088 0.859

ADD INTO APPENDIX??????

* Dallas/Fort Worth, TX
* Chicago, IL Nashville, TN
* New York, NY Miami, FL Los Angeles, CA
* Houston, TX Charleston, SC Lexington, KY
* Aspen, CO Washington, DC Tampa, FL
* Salt Lake City, UT San Antonio, TX Memphis, TN
* Denver, CO Key West, FL San Diego, CA
* Tucson, AZ Boston, MA New Orleans, LA
* Orlando, FL Dallas, TX Phoenix, AZ
* Fort Myers, FL West Palm Beach/Palm Beach, FL Myrtle Beach, SC
* Fort Lauderdale, FL Las Vegas, NV San Francisco, CA
* Honolulu, HI Kahului, HI Kona, HI
* Lihue, HI Palm Springs, CA Portland, OR
* Long Beach, CA Ponce, PR Aguadilla, PR
* Reno, NV Guam, TT Daytona Beach, FL
* Panama City, FL

Business travel is a large part of the dataset (>50%), so while this type of travel should be removed from any vacation analysis, it should be included when understanding satisfaction scores. When looking at the distribution of scores by travel type, it is interesting to see that business travel has a heavy left skew towards higher satisfaction scores, whereas personal travel has a right skew towards lower satisfaction. Mileage tickets’ distribution seems to mirror the overall distribution of scores, with a left skew as well. To improve satisfaction scores, it would be best to target business travelers, especially given how sparse they are due to the COVID-19 pandemic.

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Fourth, to provide concrete recommendations to the industry, it was absolutely necessary to analyze the demographic data, which are represented in the Age and Gender columns. The distribution of ages is right skewed with spikes at both edges on the distribution but shows that most customers surveyed were around the age of forty years old. There are also ~20,000 more females surveyed in the data versus males.

Chart, histogram

Description automatically generated

A screenshot of a computer

Description automatically generated

A density plot of the satisfactions scores by gender shows that males have a much higher likelihood to have a satisfaction score of 5 than females, which gives males a higher mean satisfaction score. The mean satisfaction scores for males and females are 3.53 and 3.27. The second plot below is also a density plot by age category, and displays that the three of the most popular age buckets (30, 40], (40, 50], and (50, 60] also have high mean satisfaction scores, 3.68, 3.68, and 3.51, that are well above the average of 3.38. The oldest age bucket (70, Inf] has a way below average satisfaction score of 2.58, perhaps because they do not feel properly accommodated given the health situations this age group has compared to the other age groups.

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Another thing to look deeper at, since airlines are concerned about pricing correctly, is the price sensitivities of the demographics. Seeing how different demographics react to price increases can provide insights on how to price seat tickets. The first chart below displays a density plot of the price sensitivities by gender, and the mean price sensitives of 1.28 for females and 1.26 for males. This shows that overall, the price sensitivities are not statistically different between the genders. The second chart plots the density of price sensitivities and mean price sensitivities by the age groups that were described in the feature engineering section of this analysis. As one would assume, the two youngest age buckets are the most price sensitive, along with the oldest age bucket. The middle age buckets are the least price sensitive, which is a good sign for airlines as they are the most common flyers. Airlines can extract additional pricing value on this age demographic.

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Lastly, before beginning to build a statistical model to predict customer satisfaction, it is important to understand the correlations in the dataset. This ensures that multicollinearity can be avoided, and the highest correlated features with the satisfaction score can be used. The figure below is a correlation matrix of all the numeric variables in the data set. At first glance it is easy to see that Arrival\_Delay\_in\_Mintues and Departure\_Delay\_in\_Minutes are almost perfectly correlated, so one of these variables and the ratio variables created from them should always be excluded. There are no strong correlations with Satisfaction, but No\_of\_Flights\_p.a. and Age have the strongest, with -0.24 and -0.22.

A picture containing graphical user interface

Description automatically generated

# Modeling

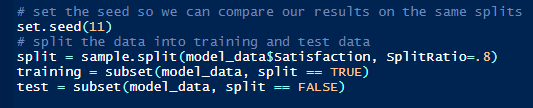
## Linear Regression

As a baseline, linear regression would be used to gauge if modeling results would be sufficient to continue to keep our target variable as a continuous variable, or if it should be transformed into a category variable to be used with classification algorithms. The distribution of the satisfaction scores displays the need for classification, but linear regression will put that to the test. Linear regression uses independent variables (the x variables), a coefficient variable for each independent variable, and an intercept to predict a continuous dependent variable (the y variable). These variables and coefficients are used to model a linear relationship in the data. Linear regression generally uses a least squares approach to attempt to fit the data.

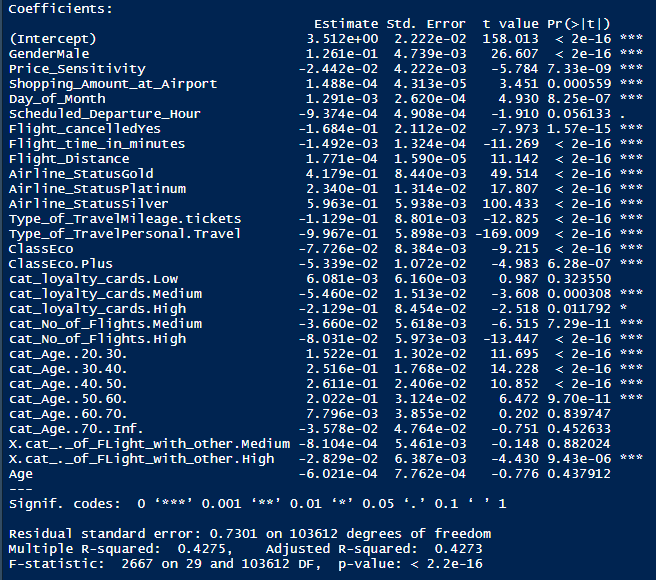
Linear regression models in R can only take numeric variables as input, which is what makes our prior data transformation steps important. Before building the model, we will need to create a list of the numeric variables, which includes our one hot encoded variables:

* Gender
* Price\_Sensitivity
* Shopping\_Amount\_at\_Airport
* Day\_of\_Month Scheduled\_Departure\_Hour
* Flight\_cancelled
* Flight\_time\_in\_minutes
* Flight\_Distance
* Airline\_StatusGold
* Airline\_StatusPlatinum
* Airline\_StatusSilver
* Type\_of\_TravelMileage.tickets Type\_of\_TravelPersonal.Travel
* ClassEco
* ClassEco.Plus
* cat\_loyalty\_cards.Low
* cat\_loyalty\_cards.Medium
* cat\_loyalty\_cards.High
* cat\_No\_of\_Flights.Medium
* cat\_No\_of\_Flights.High
* cat\_Age..20.30.
* cat\_Age..30.40.
* cat\_Age..40.50.
* cat\_Age..50.60.
* cat\_Age..60.70.
* cat\_Age..70..Inf.
* X.cat\_.\_of\_FLight\_with\_other.Medium
* X.cat\_.\_of\_FLight\_with\_other.High Age

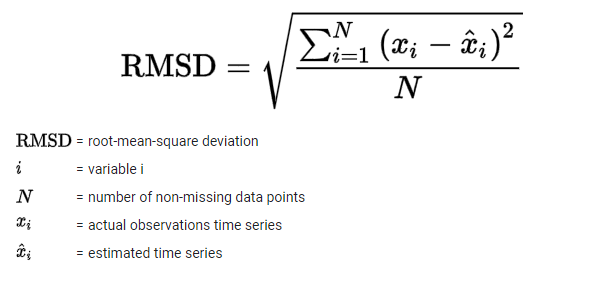
All of these variables will be passed to the *lm* function built into R to return a fitted model. To split the data between a training set used to fit the model, and a test set that the model can be evaluated on, the following code will be used to generate an 80/20 split:



The training data was then passed to the regression model. Summary of the trained linear regression model:



At first glance it is easy to see that there are a lot of statistically significant variables. These variables should be considered when building other models. The F-statistic of the equation is also statistically significant; however, the model seems to be a poor fit with an adjusted r-squared of only 42.7%. Now the model will be evaluated on test data, where we root mean square error will be used to evaluate the prediction results. Root mean square error is defined as the standard deviation of the prediction errors and can be calculated as:



After predicting the Satisfaction score using the linear regression model and comparing to the actual Satisfaction score in the test set, the root mean square comes out to 0.731. This can be interpreted to mean that the linear regression model’s errors have a standard deviation of 0.731, and given that our Satisfaction scores range from 1 to 5, this means about 68% of the errors are within a 15% difference of the actual result. This error rate is quite large, leading me to believe that a classification algorithm should be used to predict customer satisfaction scores.

Chart, histogram

Description automatically generated

## Gradient Boosting Tree Classification

Gradient boosting is an algorithm that uses weak learner trees (trees that do not have many nodes, or decisions, in them) to improve learning sequentially, as trees are built one after another. This varies from random forests that build decision tress in parallel and can be much more complex. Due to its learning algorithm, gradient boosting trees tend to have high bias, and low variance, but can even reduce variance by aggregating the output from many different models. Gradient boosting trees can be very powerful and fast to run, which are a big advantage over machine learning algorithms like support vector machines.

I

Chart, bar chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

## Support Vector Machine Classification

Chart, scatter chart

Description automatically generated

## Naive Bayes Classification

Chart, scatter chart

Description automatically generated

# Conclusion

# Appendix