**IST 687 M900:**

**APPLIED DATA SCIENCE**

Final Project Report

**Professor:**

Gary Krudys

**Team 1 - BJSR:**



Academic Year: Fall 2019

# EXECUTIVE OVERVIEW

Primary objective of this project was to review a provided large dataset from Airline Satisfaction Survey. The dataset contained various of information of satisfaction scoring and survey results. Our process was geared to analyze this satisfaction survey responses for the airlines and identify interesting business questions to identify the key drivers that could improve the satisfaction scoring.

The team took a two-step approach this process:

1. Qualify Data, looking to find quality data that would provide areas of focus and where to data mine for trends analysis
2. Marketing Data, looking to find data where it would help improve overall scoring and to allow more focus on trends or better return on investment.

**Focus 1: Qualify the Data.**

Here are our recommendations to the Airline Executive Management:

* Focus on all of the variables relating to customer service but put the most emphasis specifically on the top two variables: [Insert Variables Names].
* Put less emphasis on [Insert Variables Names] as they have less of an effect, but do not disregard them altogether as they do have a little bit of influence.
* Do not focus on variables such as [Insert Variables Names] as they don’t have much if any influence on customer satisfaction scoring.
* Based on SVM Classification analysis Airline Executives and Marketing Executives Key Strategy should pay close attention on [Insert Variables Names].
* Provide more programs and services that focus on servicing [Insert Variables Names].

**Focus 2: Expose/Explore the Market Data**

Here are our recommendations to the Airline Executive Management:

* Cultivate and maintain a strong customer-oriented experience by way of providing exceptional service and travel experience to Business Travelers who will likely continue regular frequent service and revenue to the airlines as well as high satisfaction rating.
* Provide more programs and services that focus on servicing [Insert Variables Names].
* The Airline Executive and Marketing should try to find more mechanisms and programs to ensure better overall travel experience to [Insert Variables Names].
* The Airline Executive and Marketing should consider a wider variety of demographics:
  + Type of Travel
  + Business Travelers
  + Better Service the Type of Class of Travel

**Table of Contents**

[EXECUTIVE OVERVIEW 2](#_Toc27221442)

[INTRODUCTION 6](#_Toc27221443)

[BUSINESS QUESTIONS AND APPROACH 7](#_Toc27221444)

[METHODOLOGY – Data importation, cleansing, munging, and preparation 9](#_Toc27221445)

[Methodology with Data Preparation and Selection 9](#_Toc27221446)

[Methodology with Data Cleaning 10](#_Toc27221447)

[Cleansing Focus on Variables of Value 11](#_Toc27221448)

[Cleansing Focus on NAs 12](#_Toc27221449)

[Data Analysis using Descriptive Statistics 13](#_Toc27221450)

[Statistical Analysis of NAs 13](#_Toc27221451)

[Basic descriptive analysis 14](#_Toc27221452)

[Statistical Analysis of Region in the United States 24](#_Toc27221453)

[Observations: 26](#_Toc27221454)

[First recommendations: 26](#_Toc27221455)

[DATA MODELING TECHNIQUES & PREDICTIVE ANALYSIS 27](#_Toc27221456)

[1. Linear Regression Model 27](#_Toc27221457)

[Comments on the Model 31](#_Toc27221458)

[2. The Ordered Logistic Model – Ordinal Regression 32](#_Toc27221459)

[Introduction 32](#_Toc27221460)

[Constructing the Model 33](#_Toc27221461)

[Interpreting the Model 34](#_Toc27221462)

[Further Analysis 37](#_Toc27221463)

[Comments on the Model 38](#_Toc27221464)

[3. Kernel-Based Support Vector Machines 39](#_Toc27221465)

[Comments on the Model 40](#_Toc27221466)

[DATA GENERALIZATION 41](#_Toc27221467)

[Data Generalization for Linear Regression 41](#_Toc27221468)

[Data Generalization for Ordinal Linear Model 42](#_Toc27221469)

[Data Generalization for Kernel-based Support Vector Machines 44](#_Toc27221470)

[DATA VALIDATION 45](#_Toc27221471)

[RESULTS 46](#_Toc27221472)

[Final Conclusions 46](#_Toc27221473)

[Final Recommendations 46](#_Toc27221474)

[Final Notes 46](#_Toc27221475)

[APPENDIX 49](#_Toc27221476)

[DATA DICTIONARY 49](#_Toc27221477)

[References 51](#_Toc27221478)

[Code Used 52](#_Toc27221479)

**Figure Table of Contents**

[Figure 1. Airline Satisfaction Survey Ratings/Scoring. 6](#_Toc27221395)

[Figure 2. Calculated Variable List. 11](#_Toc27221396)

[Figure 3. Evaluating NAs where exists. 12](#_Toc27221397)

[Figure 4. Initial findings of NAs within data set. 13](#_Toc27221398)

[Figure 5. Summary of NAs 13](#_Toc27221399)

[Figure 6. Graph representation of NAs compared to other variables 14](#_Toc27221400)

[Figure 7. Initial summary and findings of Satisfaction Survey data 14](#_Toc27221401)

[Figure 8. Graph representation for initial findings of Satisfaction Survey data 15](#_Toc27221402)

[Figure 9. Initial summary and findings of Type of Travel – Business Satisfaction Survey data 15](#_Toc27221403)

[Figure 10. Graph representation of initial findings of Type of Travel – Business Satisfaction Survey data 16](#_Toc27221404)

[Figure 11. Initial summary and findings of Type of Travel – Personal Satisfaction Survey data 16](#_Toc27221405)

[Figure 12. Graph representation of initial findings of Type of Travel – Personal Satisfaction Survey data 17](#_Toc27221406)

[Figure 13. Initial summary and findings of Type of Travel – Mileage Satisfaction Survey data 17](#_Toc27221407)

[Figure 14. Graph representation of initial findings of Type of Travel – Mileage Satisfaction Survey data 18](#_Toc27221408)

[Figure 15. Initial summary for standard deviation of Type of Travel – all Satisfaction Survey data 18](#_Toc27221409)

[Figure 16. Initial summary and findings of Tier Level of Airline Status Blue Satisfaction Survey data 18](#_Toc27221410)

[Figure 17. Graph of Blue Level of scores for Tier Level – Airline Status of Frequent Traveler 19](#_Toc27221411)

[Figure 18. Initial summary and findings of Tier Level of Airline Status Silver Satisfaction Survey data 19](#_Toc27221412)

[Figure 19. Graph of Silver Level of scores for Tier Level – Airline Status of Frequent Traveler 20](#_Toc27221413)

[Figure 20. Initial summary and findings of Tier Level of Airline Status Gold Satisfaction Survey data 20](#_Toc27221414)

[Figure 21. Graph of Gold Level of scores for Tier Level – Airline Status of Frequent Traveler 21](#_Toc27221415)

[Figure 22. Initial summary and findings of Tier Level of Airline Status Platinum Satisfaction Survey data 21](#_Toc27221416)

[Figure 23. Graph of Platinum Level of scores for Tier Level – Airline Status of Frequent travelers 22](#_Toc27221417)

[Figure 24. Standard Deviations of scores for Tier Status of Frequent travelers 22](#_Toc27221418)

[Figure 25. Standard Deviation, T-Value and P-Value of 22 of 28 variables 23](#_Toc27221419)

[Figure 26. Tapply calculation of Origin Satisfaction Scores and Ratings 24](#_Toc27221420)

[Figure 27. Tapply calculation of Destination Satisfaction Scores and Ratings 24](#_Toc27221421)

[Figure 28. Summary of Origin Satisfaction Scores and Ratings 24](#_Toc27221422)

[Figure 29. Summary of Destination Satisfaction Scores and Ratings 24](#_Toc27221423)

[Figure 30. By State Mean Satisfaction Score of Rating with Sum of Mean and Average 26](#_Toc27221424)

[Figure 31. Heat Map by State of Satisfaction based from data of travel from origin and to destination 26](#_Toc27221425)

[Figure 32. Summary of Linear Model 28](#_Toc27221426)

[Figure 33. Linear Model of Residuals vs Fitted 29](#_Toc27221427)

[Figure 34. Linear Model of Quantiles vs residuals 30](#_Toc27221428)

[Figure 35. Linear Model of Fitted Values vs Residuals based on Scale Location 30](#_Toc27221429)

[Figure 36. Linear Model of Residuals vs Leverage 31](#_Toc27221430)

[Figure 37. Bar Plot showing a positive skewed curve. 32](#_Toc27221431)

[Figure 38. Table produced with p-values calculated by comparing the t-values in the output to the normal distribution 33](#_Toc27221432)

[Figure 39. After running Model 2, the regression calculated the p-values table. 34](#_Toc27221433)

[Figure 40. Formula for Proportional Odds Model 34](#_Toc27221434)

[Figure 41. Regression Formula. 35](#_Toc27221435)

[Figure 42. Graph of high satisfaction scores and types of travel 38](#_Toc27221436)

[Figure 43. Prediction and Confusion Matrix Summary 39](#_Toc27221437)

[Figure 44. Linear Regression for Data Generalization 41](#_Toc27221438)

[Figure 45. t-values and p-values table matrix. 42](#_Toc27221439)

[Figure 46. Bar Graph of Satisfaction based on type of travel. 43](#_Toc27221440)

[Figure 47. Prediction and Confusion Matrix 44](#_Toc27221441)

# INTRODUCTION

With Airlines being a highly competitive industry, positive customer feedback is crucial for any company that wishes to improve or keep their market share, services, attract new ways to steal customers from competitors and increase profitability through overall and better customer experience. Our task and overall goal are to clean and analyze the dataset in order to provide actionable insights to the Airlines whose goal is to improve the Customer Satisfaction Scoring or Ratings.

**The Technical Aspects of the Dataset Provided**

Our team was provided with a 16,279 KB file dataset containing Airline Customer Satisfaction Survey. This survey data only spanned 3 months or a Calendar Quarter “Fiscal Type” (1 File, from 1/1/2014 to 3/31/2014). The dataset has 28 variables (columns) and 129,890 records (rows.)

The dataset provided did not come with a Scoring/Rating Measurement of Focus. Our team need to do an analysis of the scoring dataset to understand the Areas of Focus and Concerns. This Areas of Focus and Concerns measures customer experience and predicts business growth or business impacts. Our analysis is looking to provide a core measurement or process tool for measuring the likeliness for higher customer travel experience.

Survey respondents are categorized as three different types thanks to their likelihood to recommend value.

* + **Area of Focus** (score 4-5) are loyal enthusiasts who will keep buying and refer others, fueling growth.
  + **Area of Concern** (score 2.5-3.5) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.
  + **Area of Greater Concern** (score 0-2) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.

Then, subtracting the percentage of Areas of Concern (with focus on Greater Concern) from the percentage of Area of Focus yields a Scoring, which can range from a very low score (if customer is likely to give a bad rating) to a very high score (if customer is likely to give a great rating).



Figure . Airline Satisfaction Survey Ratings/Scoring.

# BUSINESS QUESTIONS AND APPROACH

Our main goal was to investigate the contents of the full dataset and carefully analyze what data was available to us. We began by evaluating the given variables and determining which served to be most useful in our task of improving the customer satisfaction experience for the airlines. From this list of variables, along with looking at the data input, we were able to establish the following business questions which would guide us throughout our project:

Business Rules and Assumptions

In order to analyze the data successfully our group decided to make a few assumptions that helped us in the report.

1. Assumption: Everyone who took the survey completed it with honesty and integrity.
2. Assumption: Airline Executive Management primary focus is focus on customers who will likely score higher. Secondary focus if there is time in our analysis is to find and enhance experience of low scores and passive scores
3. Assumption: If time allows to do analysis, it is possible to enhance the experience of low scores and passive scores to change them.
4. Assumption: Analyzing three months of data will give us enough information to provide solutions to satisfaction scoring practices.
5. Assumption: It is not possible to change the human variability of people wanting to travel to and from destination cities, it is only possible to change the experience or perception of one’s travel.
6. Assumption: The categorical variables can be objectively used by people traveling for either leisure or business or can be apparent in both.

Original Set of Questions

1. Which type of travelers have the highest satisfaction rate? What are the reasons why?
2. How do status and other factors impact satisfaction score?
3. What factors likely give a higher satisfaction score or rate? How does location factor in this scoring?
4. Are people more likely to recommend score higher satisfaction if they are travelling for business or leisure?
5. Are we able to quantify a lifetime customer and predict or produce a greater probability of this population?

Additional or Recently Added Set of Questions

1. Who are the customers? Who are the satisfied customers, and who are the unsatisfied customer?
2. What factors directly affect a customer’s satisfaction, more specifically, their likelihood to score higher?
3. Which modeling techniques can be used to predict the effect on likelihood to score a higher rating?
4. How can the airlines better target customers, and what steps can Airline Executive Management do to improve customer satisfaction?

# METHODOLOGY – Data importation, cleansing, munging, and preparation

Our team is using RStudio to write and run all the code for the project. RStudio is Code User Interface used for R programming. R the programming language and software environment for data scrubbing (data importation, Data Cleansing, Munging), doing statistical analysis, graphics representation and reporting, to conduct this analysis.

## Methodology with Data Preparation and Selection

As mentioned in the introduction, we were provided with a 16,279 KB file dataset containing Airline Customer Satisfaction Survey. This survey data only spanned 3 months or a Calendar Quarter “Fiscal Type” (1 File, from 1/1/2014 to 3/31/2014). The dataset has 28 variables (columns) and 129,890 records (rows.)

As a result of this our team decided to take a different approach to analyzing the data prior to cleansing and munging. This data preparation focused on evaluation and qualifying, based on:

* Understanding of the variables. Knowingly that not all variables show value based on type of analysis and how it relates to other dependent or independent variables
* Understanding NAs. Value of NAs and understanding where they exist
* Understanding of the information given. We took a high-level SWAG of our initial Business Questions or how we initially thought this data was going to help answer and find trends. We want to do quick analysis of the data with descriptive statistics and use of other data qualifying functions
* Take a high-level approach to the data to understand other aspects the data could answer (providing other business questions).

We wanted to take a simpler approach to determine and answer our initial business questions surrounding customer satisfaction rating information. Through use of initial graphs like histograms, we were able to quickly understand that using type of travel variable (business, personal, and miles) and class along with other variables would be good start and a good place to expand further.

With this type of initial data preparation, it compliments with our overall analysis conclusion, allowing of additional analysis and positioning more focused services that would improve the satisfaction rating or scoring, like offering drink coupons to its frequent traveler with delays for vacation mileage trips or “red-eye” flights to help ease already tense type of flights. And what factors contribute to long-term patronage and prediction if a person will become a lifetime customer. We have first flight information and percentage they use other airlines, so defining the probability of a lifetime customer may greater they fly with that specific airline to help figure out how much added services can affect ratings with customer threshold and their loyalty as a customer.

We initially wanted to stick with one confidence level (in terms of regression) to help with initial formatting and to give initial direction on how the data would be presenting with our business questions.

We later decided to organize and prepare a more graphical representation of the data. We felt that doing a heatmap graph/diagram, showing data organized by customer satisfaction origin and destination cities, was a more preferred method to show how this data looked in respects with a map. We took these steps to prepare the data:

* Create a vector with the lengths of each city both origin and destination.
* Get average customer satisfaction for each city.
* Use a US map and have the cities highlighted with a light to dark scale.
* This will be an average of the city, not origin city or destination city they will be average out to get a city average.
* Analyze through descriptive statistics if there is any difference between a city being and origin and destination.
* Looking and verifying if they are around the same.

## Methodology with Data Cleaning

The information was given to us in the form of MS Excel based file which we read into R studio. From there we pulled the necessary columns we decided to look at and pulled them into R studio. We initial kept all 28 columns from the dataset until running a variable evaluation designed to scrub and find variables of value. Through the variable evaluation process, we kept columns pertinent to our analysis.

Part of our process for data scrubbing, we took a different approach. We conducted a thorough analysis of our results to ensure a more meaningful finding.

1. The first step consisted in importing the data into RStudio only keeping the variables we deemed relevant to this analysis.
2. We used either the read\_csv or readxl command (library(readr) and library(readxl) required) and stored each month into a variable.
3. We decided to analyze the full dataset (all 3 months) because we need to make sure we had as much data as possible to do our analysis.
4. We used the gsub function to help with additional cleansing of spaces, commas, and other edit type fixes
5. We wanted to understand the quality of the information given. using more advanced techniques and algorithms to first give insight to influential candidate variables (evaluation of variable p-values) and running a multitude of summary functions, descriptive statistic functions, histograms, and predictive models against the dataset.
6. Once reading the data set into R, prior to removing the NAs, we used evaluation process to help better understand mechanisms to remove, understand if the NAs had any value to remove, and understand reason why it existed.

### Cleansing Focus on Variables of Value

We also decided to keep **19** of the 28 variables in the original dataset for our first analysis.

Below is a list of those variables and their definition.



Figure . Calculated Variable List.

Note that not all 28 variables were used throughout the full analysis. These variables were all chosen for specific reasons and removed variables from the dataset whenever they were not needed anymore or had no significant value.

### Cleansing Focus on NAs

After reading the 3-months of data into R we needed to evaluate NAs within the dataset to better understand both mechanisms to remove and understand if the NAs had any value to remove. Our analysis reviewed the NAs by p-value and showed that the variable with NAs (“Departure Delay in Minutes”, “Arrival Delay in Minutes”, and “Flight Time in Minutes”) did not have strong enough value to keep rows with NAs and that the total number was not enough impact if removed. Even removing the NAs, we felt that there was enough data for the analysis and algorithms to be meaningful. We used the na.omit function to remove the NAs.



Figure . Evaluating NAs where exists.

# Data Analysis using Descriptive Statistics

This section of the general use of Descriptive Statistics can best be describes as:

* Help give guideline and initial understanding of the data
* How it initially helps to answer the business questions
* How it would help determine what Models we would like use to answer our business questions
* Providing better answers on the quality of information and data given within the data set

We took a different approach to using and analyzing the descriptive statistics. We had early preconceptions of data set, wanted to understand the initial variable drivers and wanted to understand how NAs impacted the data set.

## Statistical Analysis of NAs

We looked to tackle the understanding the of the NAs within the data set. We ran scripts to find where the NAs existed.

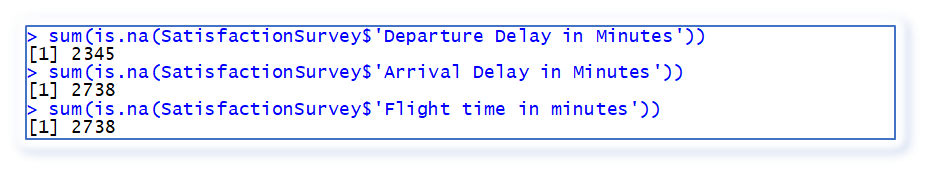


Figure . Initial findings of NAs within data set.

Then we wanted to understand how NAs impacted the data set. If there was a great potential to impact, we initial thought that we could do analysis and run various models to provide the various insights. As we dived more into this NAs understanding, we say that the NAs, overall, had little to no impact to the larger scope of the data. Figure 5 show such detail. We decided to use na.omit to clean NAs from the data set.

Departure Delay in Minutes: 2345

Arrival Delay in Minutes: 2738

Flight time in minutes: 2738

Total % of NA’s in the dataset is .22%

Figure . Summary of NAs

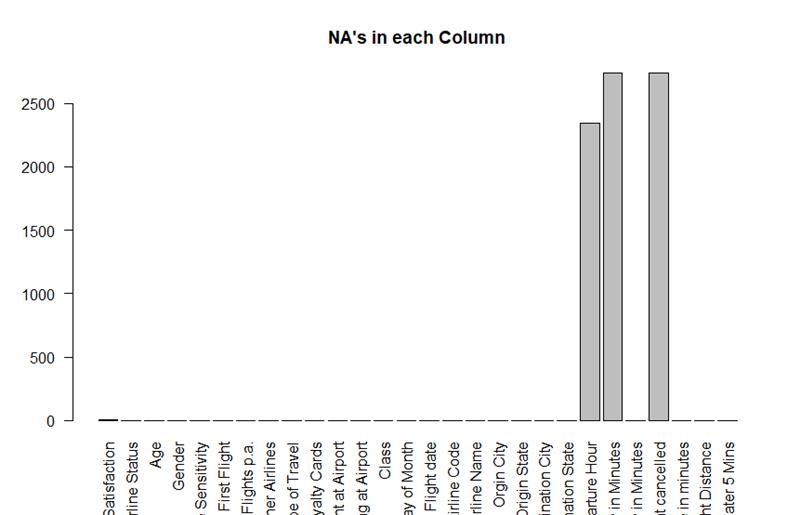


Figure . Graph representation of NAs compared to other variables

As part of the overall exercise at looking at NAs, we built a graphic representation to show where the NAs existed and how it compared to the other variables.

Basic descriptive analysis

Performing a descriptive analysis on the customer surveys data was a crucial step for us to understand who the customers were and better refine our analysis. As mentioned earlier, we have some preconceptions of the data and wanted to initially focus on where we identify some initial variable drivers.

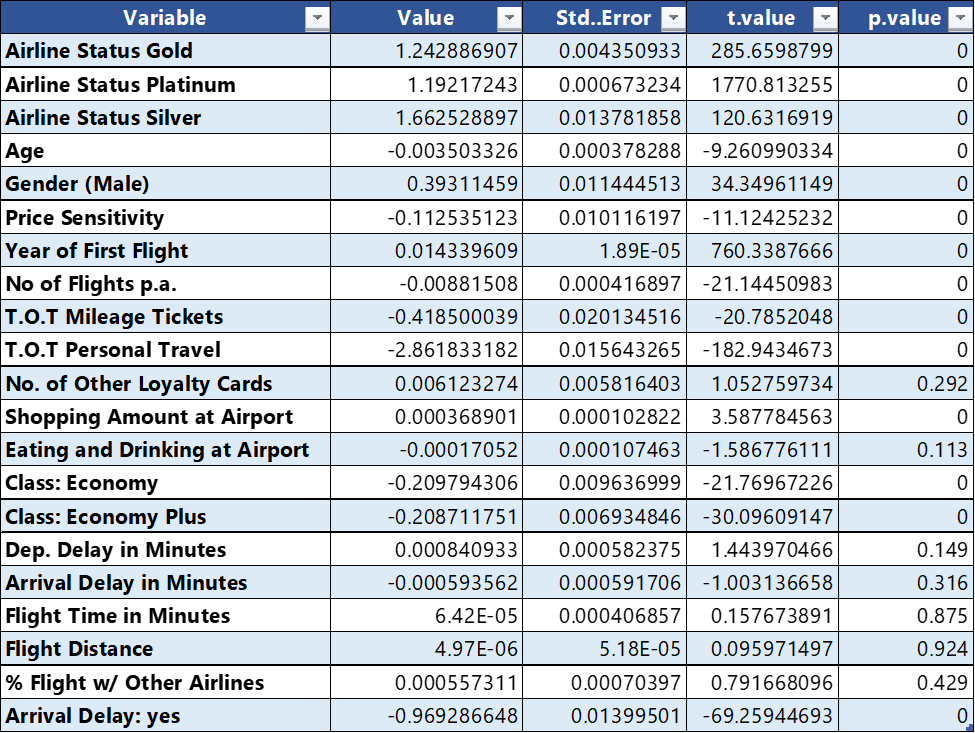


Figure . Standard Deviation, T-Value and P-Value of 22 of 28 variables

Although we initially felt that Airline Status and Type of Travel were good variable drivers. This table, Figure 7, helped our focus and validated our areas of focus (variables to use).

In Figure 8, we were looking to do a quick glance at the initial range of data, quantiles, median and mean of survey satisfaction.



Figure . Initial summary and findings of Satisfaction Survey data

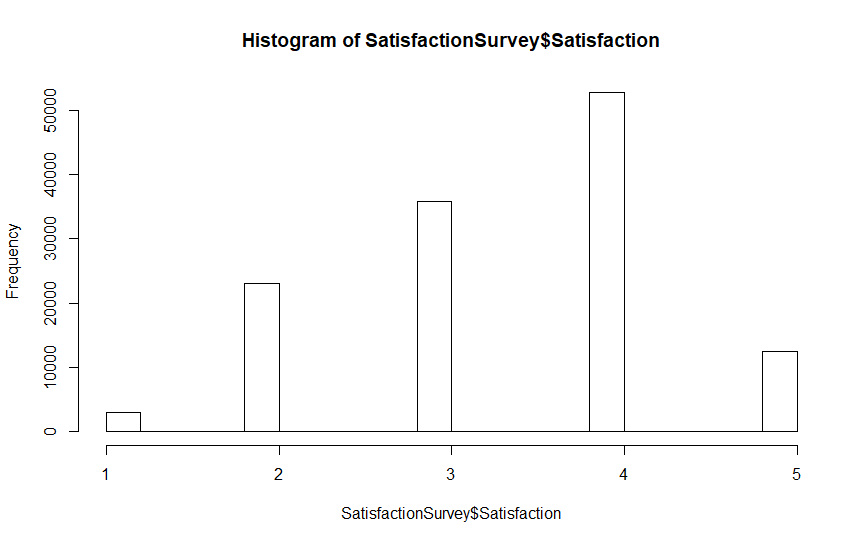


Figure . Graph representation for initial findings of Satisfaction Survey data

In Figure 9, we ran histograms to initial understand the baseline of satisfaction scoring or rating. We wanted to understand where satisfaction was generally at to allow for us to better understand where we needed improve or how we needed to use the models to find recommendations.

We wanted to focus on two areas, where we thought specific variables drivers would help provide better insights and validate our findings. We wanted to review the data of the variables to help understand initial impact to models and help drive out what models we needed to use for answering the business questions.

These two variable areas were:

* Type of Travel. Classification of the Traveler (Business, Personal, Mileage)
* Airline Status. What tiered level is the traveler (Blue, Silver, Gold, Platinum)

We wanted a quick glance at the initial range of data, quantiles, median and mean of the variables; and have a quick graphical representation, like a histogram.

For the type of traveler, we looked at the Business Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this type of traveler, shown in Figure 10.

Using a histogram, allowed to see at a high-level of the initial satisfaction for business travelers, see Figure 11. Looking at this graph, it shows that we need to get more insights and understand why and how we can either retain or improve on this area.



Figure . Initial summary and findings of Type of Travel – Business Satisfaction Survey data

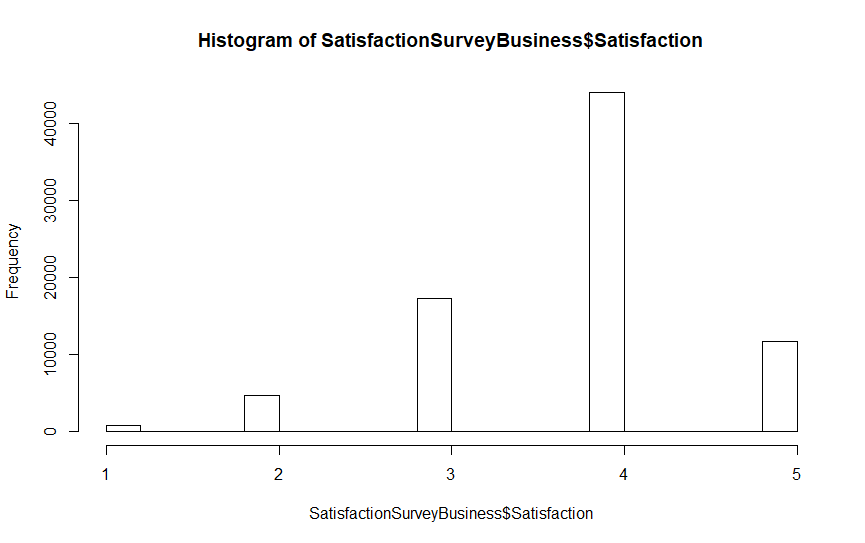


Figure . Graph representation of initial findings of Type of Travel – Business Satisfaction Survey data

For the type of traveler, we looked at the Personal Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this type of traveler, shown in Figure 12. We wanted to see if there was a difference if a traveler scored differently for personal travel as opposed to business or with miles to purchase the ticket.

Using a histogram, allowed to see at a high-level of the initial satisfaction for personal travelers, see Figure 13. Looking at this graph, it shows that we need to get more insights and understand why this rating is bad (or reflect poorly) and how we can improve on this area.



Figure . Initial summary and findings of Type of Travel – Personal Satisfaction Survey data

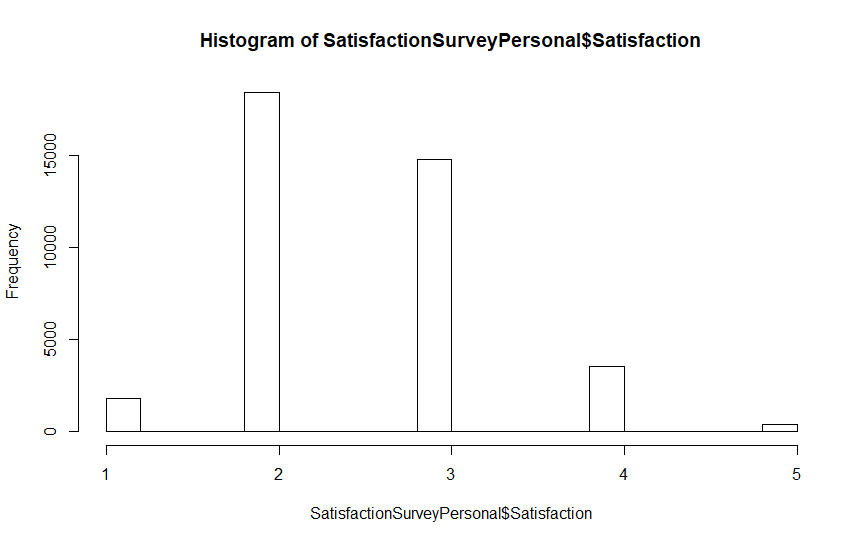


Figure . Graph representation of initial findings of Type of Travel – Personal Satisfaction Survey data

For the type of traveler, we looked at the Mileage Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this type of traveler, shown in Figure 14. We wanted to see if there was a difference if a traveler used miles, instead of paying for business or personal travel.

Using a histogram, allowed to see at a high-level of the initial satisfaction for mileage usage for travelling, see Figure 15. Looking at this graph, it shows that we need to get more insights and understand why and how we can either retain or improve on this area.



Figure . Initial summary and findings of Type of Travel – Mileage Satisfaction Survey data

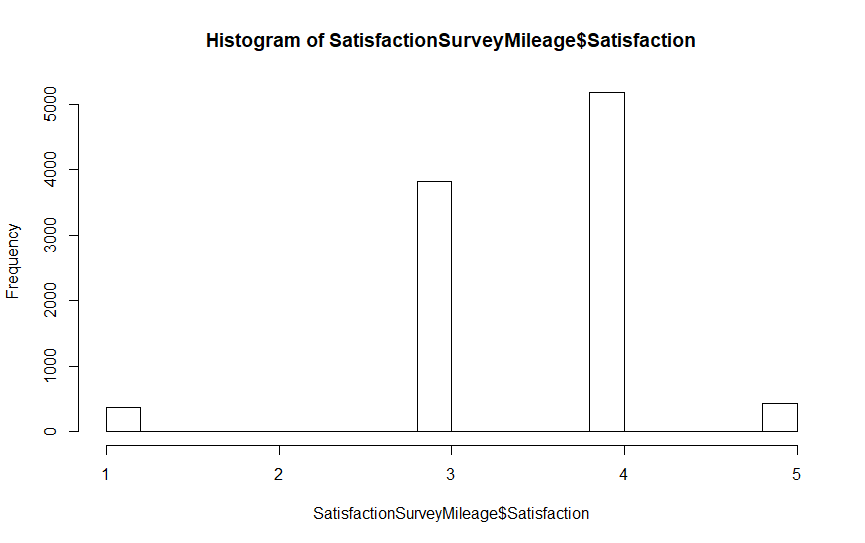


Figure . Graph representation of initial findings of Type of Travel – Mileage Satisfaction Survey data

We ran the standard deviation on each type of travel to best understand the range of satisfaction scoring. This along with the histogram and summary details gives us an initial understanding of where travelers are likely to score and who should we initial focus to help improve with scoring higher satisfaction ratings.



Figure . Initial summary for standard deviation of Type of Travel – all Satisfaction Survey data

For the Airline Status, we looked at the Blue Tiered-Level Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this status tiered level, shown in Figure 17.

Using a histogram, allowed to see at a high-level of the initial satisfaction for Blue Tiered-Status Level travelers, see Figure 18. Looking at this graph, it shows that we need to get more insights and understand why and how we can either retain or improve on this area.



Figure . Initial summary and findings of Tier Level of Airline Status Blue Satisfaction Survey data

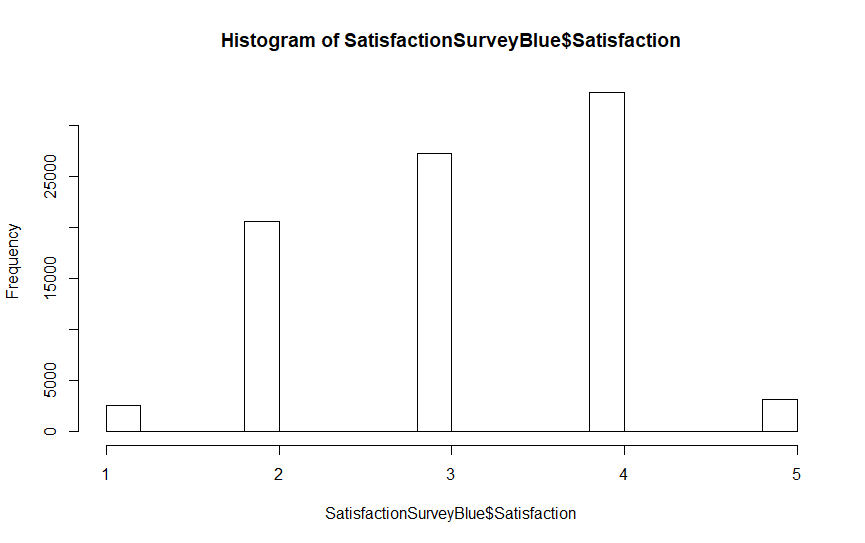


Figure . Graph of Blue Level of scores for Tier Level – Airline Status of Frequent Traveler

For the Airline Status, we looked at the Silver Tiered-Level Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this status tiered level, shown in Figure 19. We wanted to see if there was a difference on the satisfaction scoring based the status level, especially if higher than Blue Level.

Using a histogram, allowed to see at a high-level of the initial satisfaction for Silver Tiered-Status Level travelers, see Figure 20. Looking at this graph, it shows that we need to get more insights and understand why and how we can either retain or improve on this area.



Figure . Initial summary and findings of Tier Level of Airline Status Silver Satisfaction Survey data

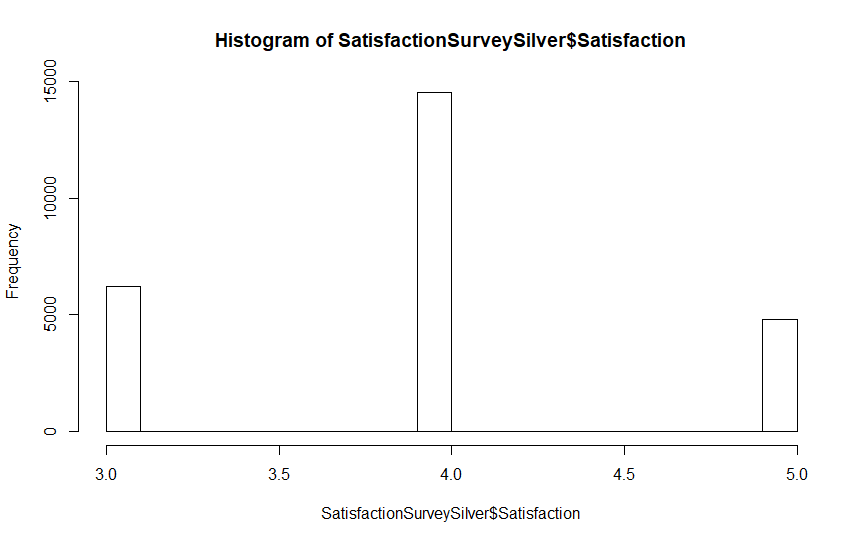


Figure . Graph of Silver Level of scores for Tier Level – Airline Status of Frequent Traveler

For the Airline Status, we looked at the Gold Tiered-Level Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this status tiered level, shown in Figure 21. We wanted to see if there was a difference on the satisfaction scoring based the second highest status level, especially if higher than Silver Level.

Using a histogram, allowed to see at a high-level of the initial satisfaction for Gold Tiered-Status Level travelers, see Figure 22. Looking at this graph, it shows that we need to get more insights and understand why and how we can make sure these travelers stay happy or find ways to improve on this area.



Figure . Initial summary and findings of Tier Level of Airline Status Gold Satisfaction Survey data

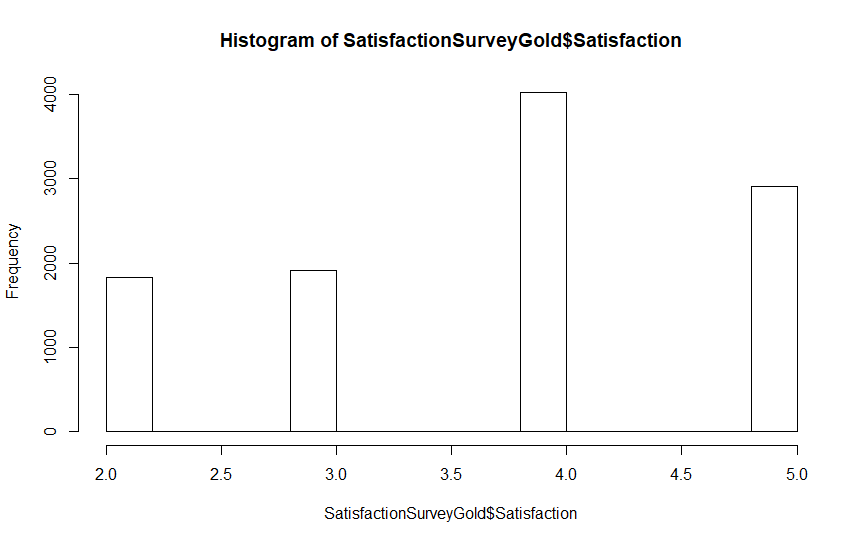


Figure . Graph of Gold Level of scores for Tier Level – Airline Status of Frequent Traveler

For the Airline Status, we looked at the Platinum Tiered-Level Satisfaction scoring. We wanted to see the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this status tiered level, shown in Figure 21. We wanted to see if there was a difference on the satisfaction scoring based the highest status level, especially if higher than Gold Level.

Using a histogram, allowed to see at a high-level of the initial satisfaction for Platinum Tiered-Status Level travelers, see Figure 22. Looking at this graph, it shows that we need to get more insights and understand why and how we can make sure these travelers stay happy or find ways to improve on this area.



Figure . Initial summary and findings of Tier Level of Airline Status Platinum Satisfaction Survey data

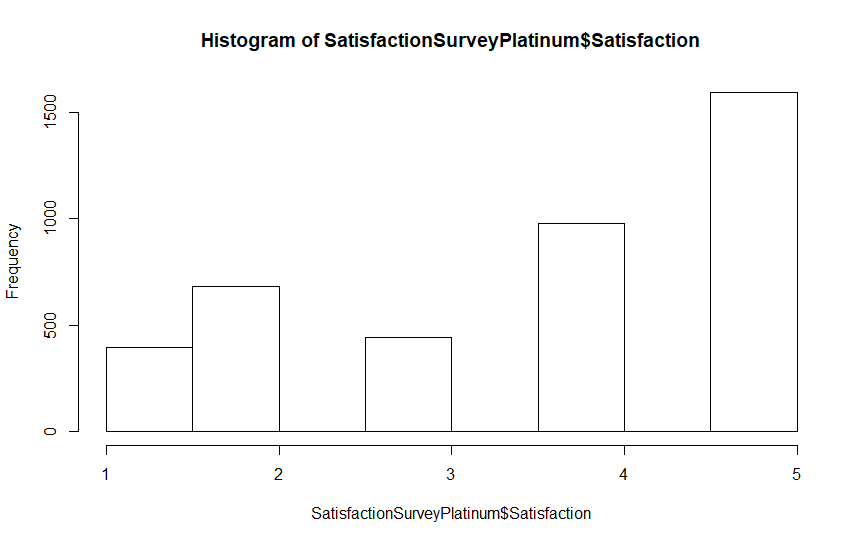


Figure . Graph of Platinum Level of scores for Tier Level – Airline Status of Frequent travelers

We ran the standard deviation on each tiered level within the Airline Status to best understand the range of satisfaction scoring. This along with the histogram and summary details gives us an initial understanding on how a traveler’s status will likely be scoring and who should we initial focus to help improve with scoring higher satisfaction ratings.



Figure . Standard Deviations of scores for Tier Status of Frequent travelers

## 

## Statistical Analysis by State of Travel

As a group, we decided that we need a better visualization for looking at the initial data set. We all felt that a heat map of the United States would provide better visualization (other than a bar or histogram) and help with better initial insights from a descriptive statistics standpoint.

Like with the basic descriptive statistics, we wanted a quick glance at the initial range of data, quantiles, median and mean of the variables; and have a quick graphical representation with focus of a heat map.

One of the first steps prior to displaying a heat map, we need to focus on what variables with satisfaction rating would work with this type of graph. We chose data that would display both city and state. Origin and Destination of travel were good variables to use in this exercise.

Part of this exercise was to use tapply function for calculating the mean. Figure 26 and 27, both show this function calculating mean for Origin and Destination.



Figure . Tapply calculation of Origin Satisfaction Scores and Ratings



Figure . Tapply calculation of Destination Satisfaction Scores and Ratings

For the both Origin and Destination of Travel, we looked at the initial range of data, quantiles, median and mean of satisfaction scoring or rating for this status tiered level, shown in Figure 28 and 29.

Figure 30, we created a view of a table by state of origin and destination mean, sum value, and average mean value. This table gave us a clean perspective and areas of focus. It was quite interesting to see that only Delaware was prone to providing low satisfaction scoring compared to all other states (which were all close in range to each other).



Figure . Summary of Origin Satisfaction Scores and Ratings



Figure . Summary of Destination Satisfaction Scores and Ratings

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **State** | **states** | **Origin Rate** | **Destination Rate** | **Sum** | **Average** |
| **Alabama** | alabama | 3.412281 | 3.458626 | 6.870906 | 3.435453 |
| **Alaska** | alaska | 3.515906 | 3.46875 | 6.984656 | 3.492328 |
| **Arizona** | arizona | 3.402906 | 3.386332 | 6.789238 | 3.394619 |
| **Arkansas** | arkansas | 3.380623 | 3.385737 | 6.76636 | 3.38318 |
| **California** | california | 3.397549 | 3.383105 | 6.780654 | 3.390327 |
| **Colorado** | colorado | 3.361554 | 3.378359 | 6.739913 | 3.369956 |
| **Connecticut** | connecticut | 3.276151 | 3.415054 | 6.691204 | 3.345602 |
| **Delaware** | delaware | 2.6 | 2.9375 | 5.5375 | 2.76875 |
| **Florida** | florida | 3.37824 | 3.383596 | 6.761836 | 3.380918 |
| **Georgia** | georgia | 3.403934 | 3.386015 | 6.789949 | 3.394974 |
| **Hawaii** | hawaii | 3.474623 | 3.467919 | 6.942542 | 3.471271 |
| **Idaho** | idaho | 3.372751 | 3.462338 | 6.835088 | 3.417544 |
| **Illinois** | illinois | 3.34123 | 3.350204 | 6.691435 | 3.345717 |
| **Indiana** | indiana | 3.333721 | 3.330969 | 6.66469 | 3.332345 |
| **Iowa** | iowa | 3.326203 | 3.301546 | 6.62775 | 3.313875 |
| **Kansas** | kansas | 3.3861 | 3.354244 | 6.740344 | 3.370172 |
| **Kentucky** | kentucky | 3.391304 | 3.371591 | 6.762895 | 3.381448 |
| **Louisiana** | louisiana | 3.378665 | 3.38 | 6.758665 | 3.379333 |
| **Maine** | maine | 3.43956 | 3.394737 | 6.834297 | 3.417149 |
| **Maryland** | maryland | 3.359188 | 3.354805 | 6.713993 | 3.356996 |
| **Massachusetts** | massachusetts | 3.367139 | 3.388138 | 6.755277 | 3.377638 |
| **Michigan** | michigan | 3.3713 | 3.375954 | 6.747254 | 3.373627 |
| **Minnesota** | minnesota | 3.40837 | 3.399314 | 6.807685 | 3.403842 |
| **Mississippi** | mississippi | 3.315436 | 3.480263 | 6.795699 | 3.39785 |
| **Missouri** | missouri | 3.358525 | 3.378849 | 6.737373 | 3.368687 |
| **Montana** | montana | 3.374648 | 3.349162 | 6.72381 | 3.361905 |
| **Nebraska** | nebraska | 3.395948 | 3.436735 | 6.832683 | 3.416342 |
| **Nevada** | nevada | 3.385444 | 3.381434 | 6.766878 | 3.383439 |
| **New Hampshire** | new hampshire | 3.401515 | 3.342857 | 6.744372 | 3.372186 |
| **New Jersey** | new jersey | 3.352744 | 3.371257 | 6.724002 | 3.362001 |
| **New Mexico** | new mexico | 3.420382 | 3.355878 | 6.77626 | 3.38813 |
| **New York** | new york | 3.362199 | 3.401464 | 6.763663 | 3.381831 |
| **North Carolina** | north carolina | 3.404216 | 3.38349 | 6.787706 | 3.393853 |
| **North Dakota** | north dakota | 3.246862 | 3.342975 | 6.589837 | 3.294919 |
| **Ohio** | ohio | 3.336819 | 3.422172 | 6.758991 | 3.379496 |
| **Oklahoma** | oklahoma | 3.416953 | 3.372632 | 6.789585 | 3.394792 |
| **Oregon** | oregon | 3.403924 | 3.408624 | 6.812548 | 3.406274 |
| **Pennsylvania** | pennsylvania | 3.387974 | 3.389654 | 6.777627 | 3.388814 |
| **Rhode Island** | rhode island | 3.488189 | 3.360902 | 6.849091 | 3.424546 |
| **South Carolina** | south carolina | 3.418118 | 3.307692 | 6.725811 | 3.362905 |
| **South Dakota** | south dakota | 3.413636 | 3.272727 | 6.686364 | 3.343182 |
| **Tennessee** | tennessee | 3.369637 | 3.340347 | 6.709984 | 3.354992 |
| **Texas** | texas | 3.375437 | 3.385474 | 6.760911 | 3.380455 |
| **Utah** | utah | 3.463512 | 3.431116 | 6.894629 | 3.447314 |
| **Vermont** | vermont | 3.240741 | 3.507042 | 6.747783 | 3.373891 |
| **Virginia** | virginia | 3.368573 | 3.378578 | 6.747151 | 3.373575 |
| **Washington** | washington | 3.42349 | 3.383396 | 6.806886 | 3.403443 |
| **West Virginia** | west virginia | 3.436364 | 3.25 | 6.686364 | 3.343182 |
| **Wisconsin** | wisconsin | 3.39345 | 3.368014 | 6.761464 | 3.380732 |
| **Wyoming** | wyoming | 3.38189 | 3.375527 | 6.757417 | 3.378709 |

Figure . By State Mean Satisfaction Score of Rating with Sum of Mean and Average

Using heat map graph, allowed to see at a high-level of the initial satisfaction for satisfaction by state travelers, see Figure 31. Looking at this graph, did not give us an avenue to focus as much as we got from the other descriptive statistics. This was due to the graph didn’t seem to have as much range or differentiation as other graphs used.

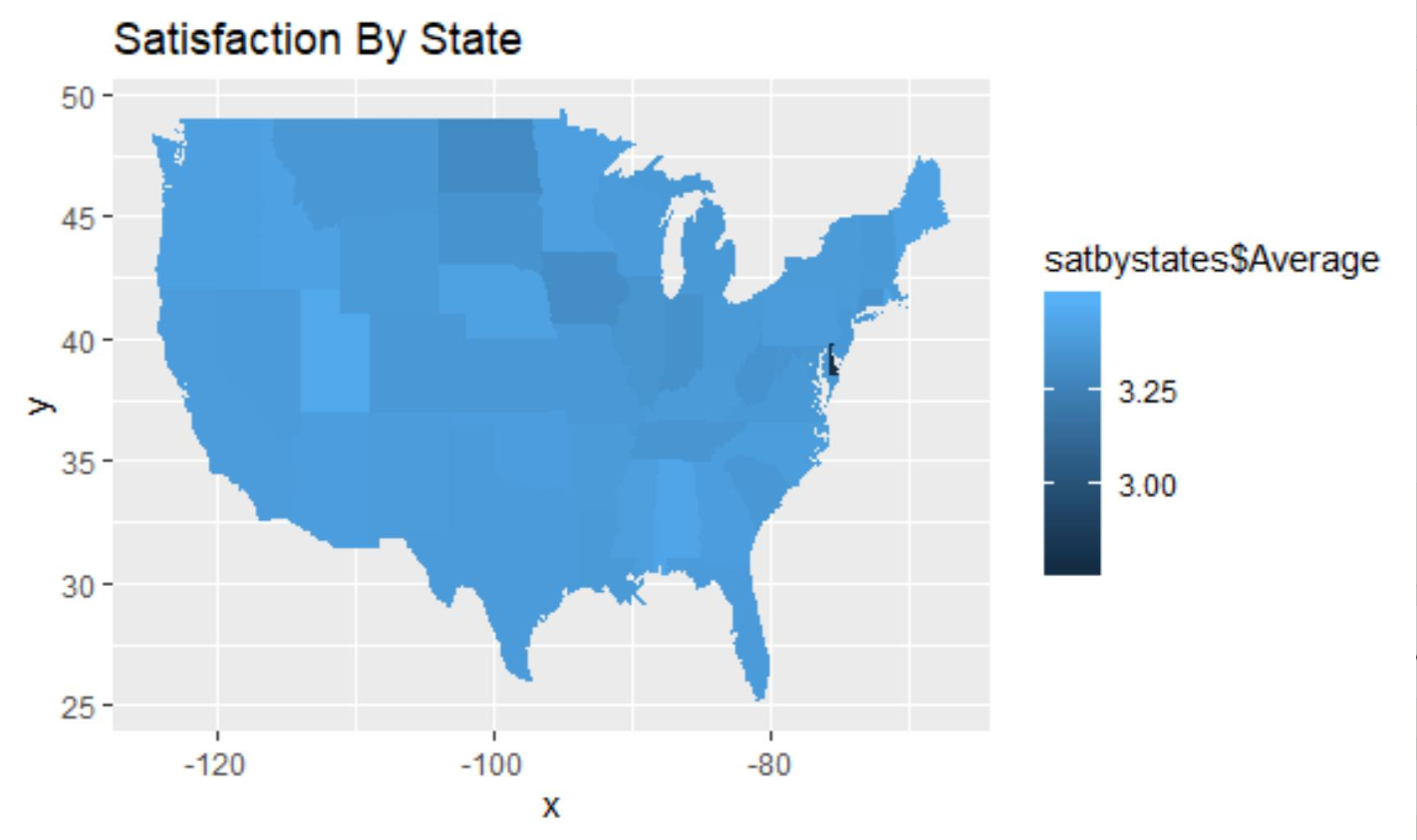


Figure . Heat Map by State of Satisfaction based from data of travel from origin and to destination

# DATA MODELING TECHNIQUES & PREDICTIVE ANALYSIS

We decided to use three modeling techniques in order to see how the airline satisfaction ratings/scoring could improve. The three algorithms used for the predictions were **linear regression**, **ordinal linear regression**, and **kernel-based support vector machines**.

* **Linear regression**: linear modeling is one of the most frequently used techniques in statistics where we investigate the potential relationship between a variable of interest (the dependent variable) and a set of one of more variables (the independent variables).
* **Ordinal linear regression**: can be performed using a coefficient vector and a set of thresholds to a dataset. It is a type of regression analysis used for predicting value that exists on an arbitrary scale where only the relative ordering between different values is significant. Ordinal regression turns up often in modeling of levels of preference scales and rating (i.e. 1 to 5 scoring), also known as ranking learning
* **Kernel-based support vector machines:** in machine learning, support vector machines (KSVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A kernel classifier typically computes a weighted sum of similarities. It is a type of regression analysis used for predicting a variable, where only the relative ordering between different values are significant or between regression and classification

The objective of such modeling techniques was to make sure that our final recommendations would *not* be solely based on descriptive analysis but also on predictive statistics and analysis, as the airline ratings or scoring needs to improve.

## 1. Linear Regression Model

A linear model was constructed after the first initial analysis of the data. The data was cleaned of NA’s. We opted not to replace NA’s with another value because this method still provided over 120,000 observations to use in the model. The variable ‘satisfaction’ was regressed against the following variables:

* Airline Status
* Age
* Gender
* Price Sensitivity
* Year of First Flight
* No of Flights p.a.
* Type of travel
* No. of other Loyalty Cards
* Shopping Amount at Airport
* Eating and Drinking at Airport
* Class
* Day of Month
* Flight Date
* Scheduled Departure Hour
* Departure Delay in Minutes
* Arrival Delay in Minutes
* Flight Time in Minutes
* Flight Distance
* Arrival Delay Greater Than 5 Mins
* % of Flight with other Airlines

Out of the 20 variables used in the model, 12 were statistically significant at the 95% level. The geographic data was omitted because it created a confused output. There were far too many coefficients and variables to analyze by incorporating the geographic data. The variable ‘Flight Cancelled’ was omitted because it only contained one factor. None of the flights in the data set were cancelled. The summary of the model was difficult to interpret.

> summary(lm1)

Residuals:

Min 1Q Median 3Q Max

-3.15893 -0.41299 0.07812 0.47086 2.85087

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.665e+00 1.898e+00 -4.040 5.36e-05 \*\*\*

data$’Airline Status’Gold 4.418e-01 7.489e-03 58.986 < 2e-16 \*\*\*

data$’Airline Status’Platinum 2.661e-01 1.164e-02 22.860 < 2e-16 \*\*\*

data$’Airline Status’Silver 6.199e-01 5.226e-03 118.615 < 2e-16 \*\*\*

data$Age -2.332e-03 1.413e-04 -16.513 < 2e-16 \*\*\*

data$GenderMale 1.319e-01 4.223e-03 31.239 < 2e-16 \*\*\*

data$’Price Sensitivity’ -4.078e-02 3.763e-03 -10.838 < 2e-16 \*\*\*

data$’Year of First Flight’ 4.857e-03 6.790e-04 7.153 8.55e-13 \*\*\*

data$’No of Flights p.a.’ -3.309e-03 1.553e-04 -21.307 < 2e-16 \*\*\*

data$’Type of Travel’Mileage tickets -1.469e-01 7.784e-03 -18.868 < 2e-16 \*\*\*

data$’Type of Travel’Personal Travel -1.076e+00 5.000e-03 -215.277 < 2e-16 \*\*\*

data$’No. of other Loyalty Cards’ -2.542e-03 2.144e-03 -1.186 0.2358

data$’Shopping Amount at Airport’ 1.635e-04 3.833e-05 4.267 1.99e-05 \*\*\*

data$’Eating and Drinking at Airport’ -8.640e-05 3.961e-05 -2.181 0.0292 \*

data$ClassEco -7.725e-02 7.384e-03 -10.461 < 2e-16 \*\*\*

data$ClassEco Plus -7.066e-02 9.490e-03 -7.445 9.73e-14 \*\*\*

data$’Day of Month’ -2.072e-04 2.488e-04 -0.833 0.4049

data$’Flight date’ 1.282e-09 9.521e-10 1.346 0.1782

data$’Scheduled Departure Hour’ 3.802e-03 4.434e-04 8.575 < 2e-16 \*\*\*

data$’Departure Delay in Minutes’ 6.088e-05 2.136e-04 0.285 0.7756

data$’Arrival Delay in Minutes’ 3.088e-05 2.168e-04 0.142 0.8867

data$’Flight time in minutes’ -1.117e-06 1.389e-04 -0.008 0.9936

data$’Flight Distance’ 4.555e-06 1.678e-05 0.271 0.7861

data$’Arrival Delay greater 5 Mins’yes -3.445e-01 5.097e-03 -67.590 < 2e-16 \*\*\*

data$’% of Flight with other Airlines’ -7.293e-05 2.603e-04 -0.280 0.7794

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7189 on 127123 degrees of freedom

Multiple R-squared: 0.4469, Adjusted R-squared: 0.4468

F-statistic: 4279 on 24 and 127123 DF, p-value: < 2.2e-16

Figure . Summary of Linear Model

Although the summary was able to identify p-value of variables that showed value. With the 4 plots shown below it was difficult to make sense of plot graphs.

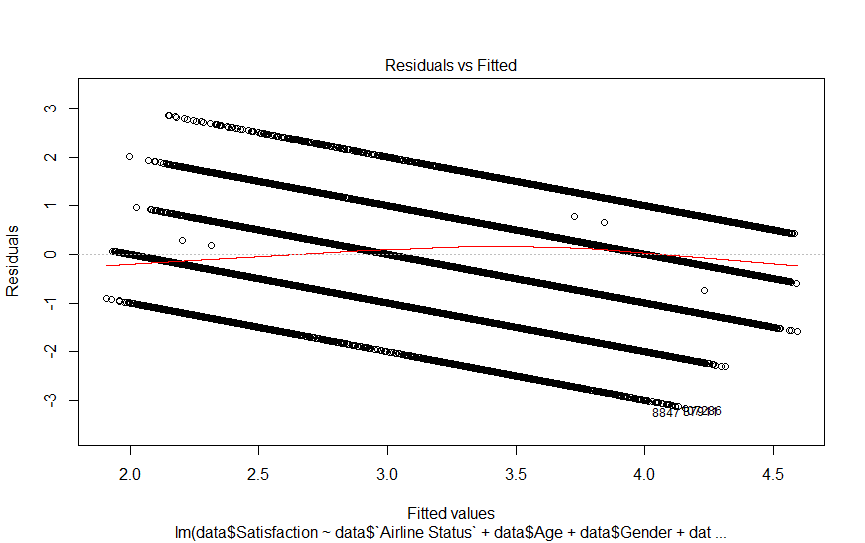


Figure . Linear Model of Residuals vs Fitted

For Figure 5, the plot shows how the affects of the variables gives a less propensity or likeliness (negative) not to achieve a higher score than lower score. It is difficult to understand for this plot of what negatively impacts the score.

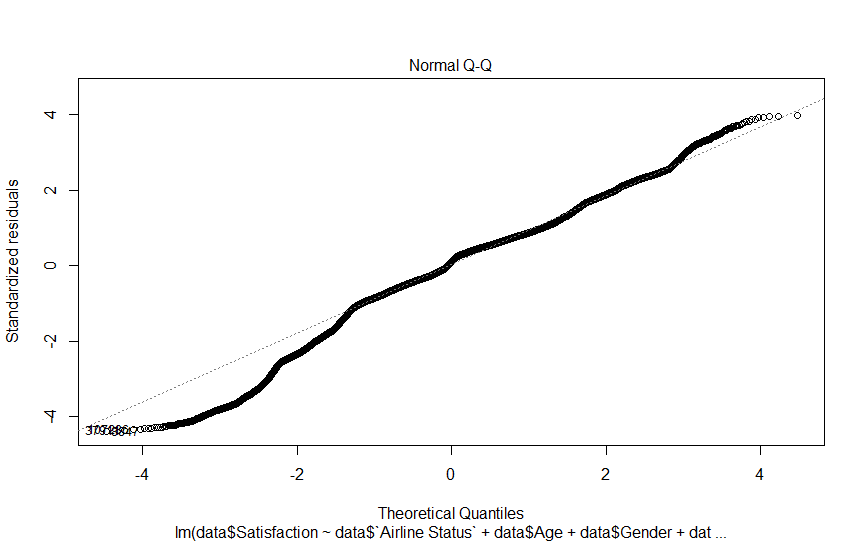


Figure . Linear Model of Quantiles vs residuals

For Figure 6, it shown a near mirror image between the Quantiles and Residuals, but it is difficult to understand the reasons why or what variables impact the reasons why square root of error exists.

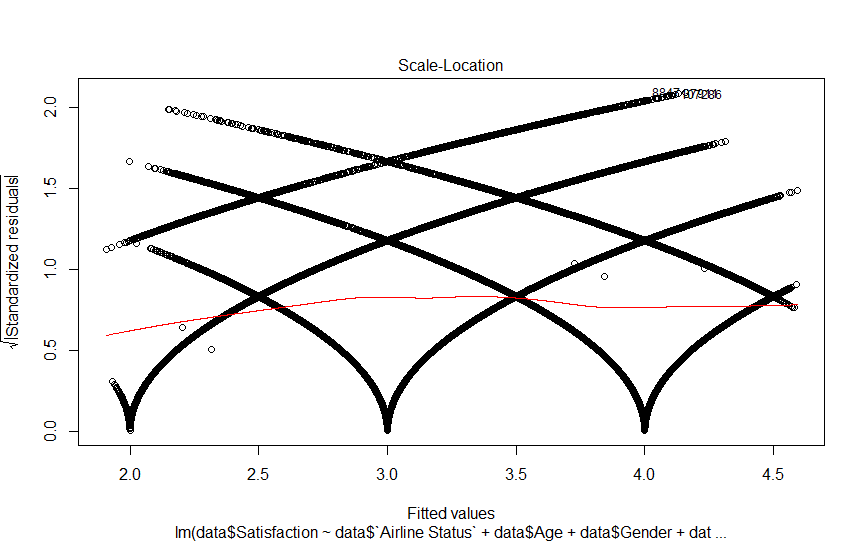


Figure . Linear Model of Fitted Values vs Residuals based on Scale Location

For Figure 7, it is a cool looking graph, but it seems misleading because satisfaction scoring can have .5 value. This graph interprets that there cannot have any value other than a whole value. This graph explains that coefficients has continuous values, which is better suited for an Ordinal Linear Model than Simple Linear.

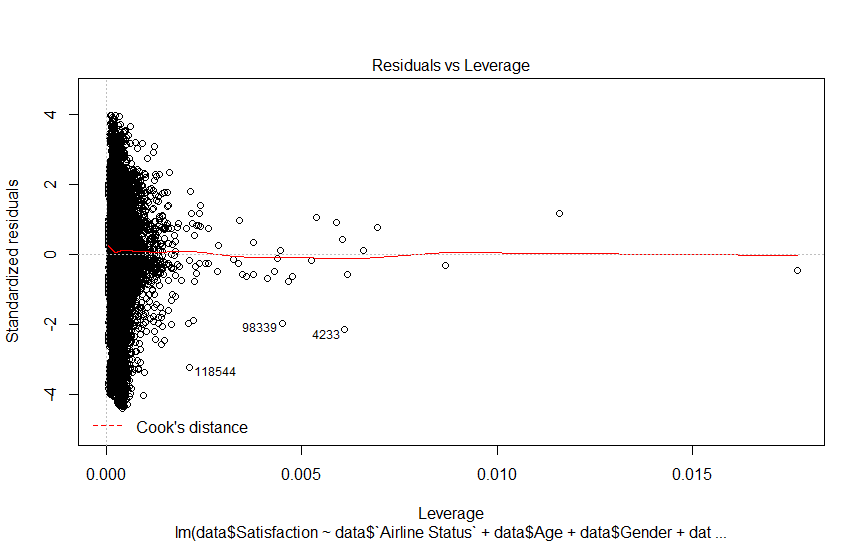


Figure . Linear Model of Residuals vs Leverage

Figure 8, shows the p-value of the observations of the dataset. Although this does show the value of both observations and variables, it is difficult to understand what variables demonstrate the better p-value than others.

### Comments on the Model

The coefficients were continuous values. However, our output consists of ordinal response values. After some research, we determined the linear model was not the correct method to determine the dependence of our ordinal response output. That is when the proportional-odds model was introduced and constructed to conduct a more thorough analysis of the data.

## 2. The Ordered Logistic Model – Ordinal Regression

### Introduction

The survey data we are analyzing do not contain a simple continuous output. The satisfaction level represents an ordered and categorical variable. The range of satisfaction in our data is from 1 to 5. This can be thought of as from low to high (ordered). Simple linear regression does not suffice in our case. We will use the proportional odds logistic model. This model will allow us to properly identify the variables which explain the variation in the satisfaction scores more precisely.

The first step was to omit any NA’s from our data. The data set is very robust. The decision was opted to eliminate any rows of data with NA’s versus replacing them with an average or another variable. There are still well over 120,000 observations in the data when using this methodology. The next step is to define our dependent variable and order it. We use the factor () function to accomplish this in R. We also define the levels within the factor function and set the variable to ordered (see code). We use the levels () function to confirm our variable is now ordered 1 through 5. Passing the variable through the max () function gives an output of 5 and also shows our variable is ordered properly (the min () function will do the same).

We can view the distribution of our dependent variable using a bar plot. A histogram doesn’t suffice because we are not dealing with continuous data. Building the bar plot gives us the following output:

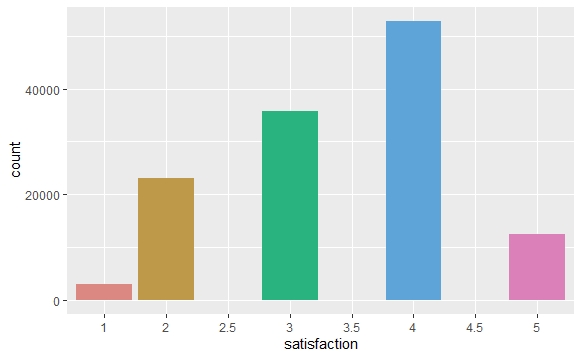


Figure . Bar Plot showing a positive skewed curve.

The data is a bit right skewed according to the plot. The most common score is 4. The average score is 3.38 (the average is taken from passing the satisfaction variable through the summary() function).

### Constructing the Model

Now that we have a baseline, we’ll start to build our model. There are a set of variables which were omitted from the model because they are not continuous, and they don’t work as categorical variables in the model when passed through the factor () function. We’ll use other methods of analysis to determine if they’re significant indicators of satisfaction scores. These variables are:

* Flight Date
* Airline Code
* Airline Name
* Origin City
* Origin State
* Destination City
* Destination State

The first model consists of 19 variables. The continuous variables were all included and some of the categorical variables were passed through the factor () function. See the code. The polr () function doesn’t contain a p-value in the output. The p-values are calculated by comparing the t-values in the output to the normal distribution. See to the code. The following table is produced after calculating the p-values:



Figure . Table produced with p-values calculated by comparing the t-values in the output to the normal distribution

The p-values for the variables are in the right-hand column of the above table. The variables which weren’t determined to be statistically significant are labeled in red. We will remove those variables and construct a new model. See the code for Model 2. After running the regression and calculating the p-values we obtain the following the table:



Figure . After running Model 2, the regression calculated the p-values table.

All the variables included in the model are now statistically significant. Now we will focus on interpretation of the model.

### Interpreting the Model

The ordinal regression model is not interpreted the same as a normal linear regression model. Let J be the total number of categories and M be the number of independent variables. Our model has 8 categories and 15 independent variables (J = 8 and M = 15). The formula of the Proportional Odds Model is given as:

Figure . Formula for Proportional Odds Model

Where j = 1, …, J -1

(see note below for explanation of J-1)

Our regression formula is as follows:

Figure . Regression Formula.

Where j = 1, …, J-1 and i = 1, …, M

Note: We use J-1 in the above formulas because . For example, in our model, the probability that a satisfaction score is 5 or lower is 100%. Therefore, we use J-1 to calculate our predictions.

For our model:

* j = 1 refers to a satisfaction score of 1
* j = 2 refers to a satisfaction score of 2
* j = 3 refers to a satisfaction score of 2.5
* j = 4 refers to a satisfaction score of 3
* j = 5 refers to a satisfaction score of 3.5
* j = 6 refers to a satisfaction score of 4
* j = 7 refers to a satisfaction score of 4.5
* j = 8 refers to a satisfaction score of 5
* i = 1 refers to airline status: Gold
* i = 2 refers to airline status: Platinum
* …
* i = 9 refers to type of travel: Mileage tickets
* …
* i = 15 refers to arrival delay: yes

#### Coefficients:

The “value” column in the table above represents the coefficient values for the respective variables. The continuous and categorical variables have slightly different interpretations. For our categorical variable “gender” the coefficient is interpreted as:

* A person with gender type “male” is associated with a higher probability of giving a higher satisfaction rating than a person with gender type “female.”
* For our categorical variable “type of travel: Personal Travel” the coefficient is interpreted as:
* A person with travel type “personal” has a lower probability of giving a higher satisfaction rating than a person of travel type “Mileage” or travel type “Business” (because of the coefficient is negative).
* For a continuous variable like shopping amount, the coefficient is interpreted as:
* A one unit increase in shopping amount increases the probability that an individual will give a higher satisfaction rating, holding all else constant.

#### Intercepts:

The last seven values in Column 1 are the intercepts of the model. This is a departure from the normal linear regression which only has one intercept.

* The intercept 3 | 3.5 corresponds to . It is the log of odds of an individual giving a satisfaction score of 1, 2, 2.5 or 3 versus giving a score of 3.5, 4, 4.5 or 5.
* The intercept 4 | 4.5 corresponds to . It is the log of odds of an individual giving a satisfaction score of 1, 2, 2.5, 3, 3.5, or 4 versus giving a score of 4.5 or 5.

#### Predictions:

We can use the intercept and coefficients to predict the probability of a satisfaction score given individual attributes. For example, lets calculate for an individual who has the following attributes:

* Airline Status: Silver
* Age: 35
* Gender: Female
* Price Sensitivity: 1
* Year of First Flight: 2005
* No. of Flights pa: 10
* Type of Travel: Personal
* Shopping Amount: 60
* Class: Eco
* Departure Hour: 18
* Arrival Delay: yes

The calculation is as follows:

= 23.5546295609 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

(0.0003610423\*60) +

(-0.2098222068\*1) +

(0.0113269368\*18) +

(-0.9756755701\*1)) = -3.80456

Remember, this value represents the log-odds this specific individual will give a score of 1. We need to use exponentiation to determine the probability. We calculate as follows:

= = 0.02178389

Therefore, we can determine that the probability of a person with the defined attributes above will give a satisfaction score of 1 with a probability of approximately 2.2%. What about the probability this same person gives a score of 4?

= 0.1864952

Therefore, we expect the probability of a person with the defined attributes will give a satisfaction score of 4 with a probability of approximately 18.65%.

### Further Analysis

The above calculations are useful if we’re trying to predict how an individual with specific attributes might rate their experience on the given scale of 1 to 5. However, the data set is robust, and this specificity is probably not as useful on a macro scale. It might be more useful to identify those variables which contribute the most to satisfaction and dissatisfaction within the model. The three variables which seem to have the biggest impact on satisfaction scores are:

* Airline Status
* Arrival Delay
* Type of Travel

The model indicates that customers who belong to a higher airline status than blue are far more probable to rate their experience highly. The statuses above blue are silver, gold and platinum. Their respective coefficients are 1.66, 1.24 and 1.93. Remember, these coefficients represent an increase to the log-odds of an individual giving a high score. Therefore, these positive values being over 1 is significant. Platinum being nearly 2 is very significant.

Arrival delay is one of our many categorical variables. A customer experiencing an arrival delay is less likely to give a high satisfaction score by a value of -0.976. This means when a customer experiences an arrival delay, all else equal, they are more likely to give a lower satisfaction score by nearly a full point on the log-odds scale.

The most prominent variable in the model corresponds to passengers conducting personal travel. The coefficient of -2.86 is striking. This means that a passenger conducting personal travel is far more probably to give a lower satisfaction score. Let’s use a visualization to investigate this further.

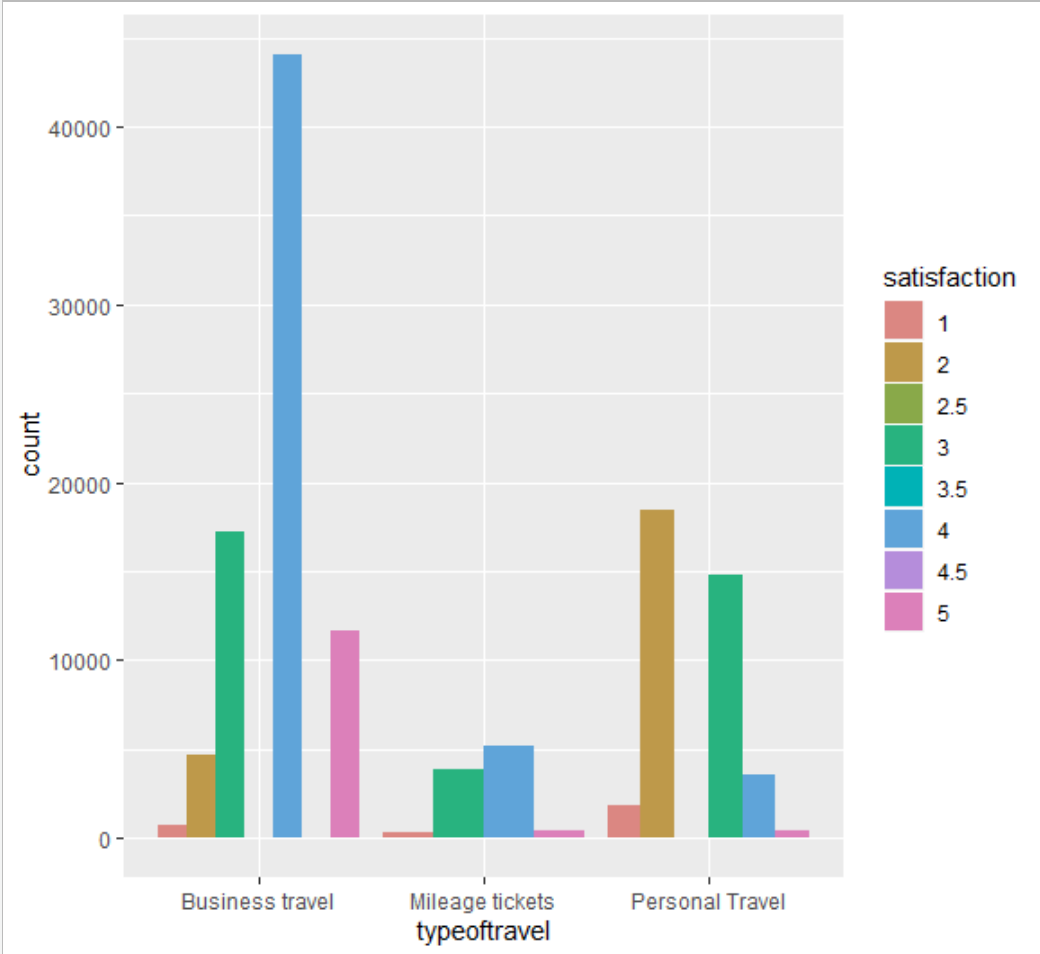


Figure . Graph of high satisfaction scores and types of travel

The graph above shows a huge disparity between high satisfaction scores and types of travel. Most of the scores 4 and above were given by people who conducted personal travel. Most of the scores below 3 were given by people conducting personal travel. While there are a multitude of significant factors contributing to how a person will rate their satisfaction, it seems type of travel is the biggest determinant.

### Comments on the Model

This model does not take geographic considerations into account. This is a weakness of the model. Other means of analysis have been used to determine the effect geography has on satisfaction scores. Also, not all the covariates in the model follow the parallel slopes assumption. This is not a surprise given the robustness of the data we are analyzing. we created a table within R testing the “goodness of fit” of the model that shows this. Obtaining a perfect fit in an ordinal regression with a data set of over 120,000 observations is unlikely. Therefore, we don’t believe the analysis gleaned from model has lost much magnitude.

## 3. Kernel-Based Support Vector Machines

For the Support vector machine, we want to increase our overall average. Although we can shoot for only 5/5 and make that into a binary variable, a more realistic approach would be to just increase our average.  So, we wanted to look at what factors can determine whether or not the satisfaction will be above average. After testing all the variables in different combinations, the model that continuously showed the highest accuracy was the final one.

Using the full dataset, it would not finish and would crash R. To get a working model, we used a subset of the data.

We took the smallest airline, by number of flights. The airline was Cool Air. This airline had 1280 flights. We started by using 1 variable and working a way to use all variables. There is 1 variable we could not use, Flight Cancelled. After testing the variables, the ones that gave the highest accuracy was Age + Class + Price Sensitivity + Airline Status + Type of Travel. After using the confusion matrix, the accuracy was consistently over 70%. It seems the model will guess more times that the review will leave an above average review but will give a below average review.

We had a similar problem with crashing R with using neural network.



Figure . Prediction and Confusion Matrix Summary

### Comments on the Model

For the KSVM when using the full Data Set the model would not finish and would crash R. This is a weakness of the model in order to get a working model, we used a subset of the data. We took the smallest airline, by number of flights. Because of our concern with crashing the system and R, we started by using a single variable and worked to use all variables. Through this research, we found that this variable, Flight Cancelled, we could not use because it caused system issues. Our findings through this exercise, we found the highest accuracy was with this set of variables Age + Class + Price Sensitivity + Airline Status + Type of Travel. The confusion matrix returned an accuracy over 70%. These findings found that the model will guess more times and will indicate an above average review but will show an overall below average review.

# DATA GENERALIZATION

Data evaluations. Why did go in the direction with the data and why did we use the Models picked.

After talking with Professor Krudys, we decided to go outside the realms of class teachings and look into over Linear Regressions like Ordinal Linear Model – Proportional Odds. Our objective for go this direction was a result from our findings during the data preparation and variable validation process. We determined the linear model was not the correct method to determine the dependence of our ordinal response output. That is when the proportional-odds model was introduced and constructed to conduct a more thorough analysis of the data.

## Data Generalization for Linear Regression



Figure . Linear Regression for Data Generalization

**Observations:** We saw the coefficients were continuous values, which consists of ordinal response values. Because it was more of an ordinal response values, we determined the simple linear model was not the correct method to determine the dependence of our ordinal response output. The proportional-odds model made more sense to conduct a more thorough analysis of the data.

## Data Generalization for Ordinal Linear Model



Figure . t-values and p-values table matrix.

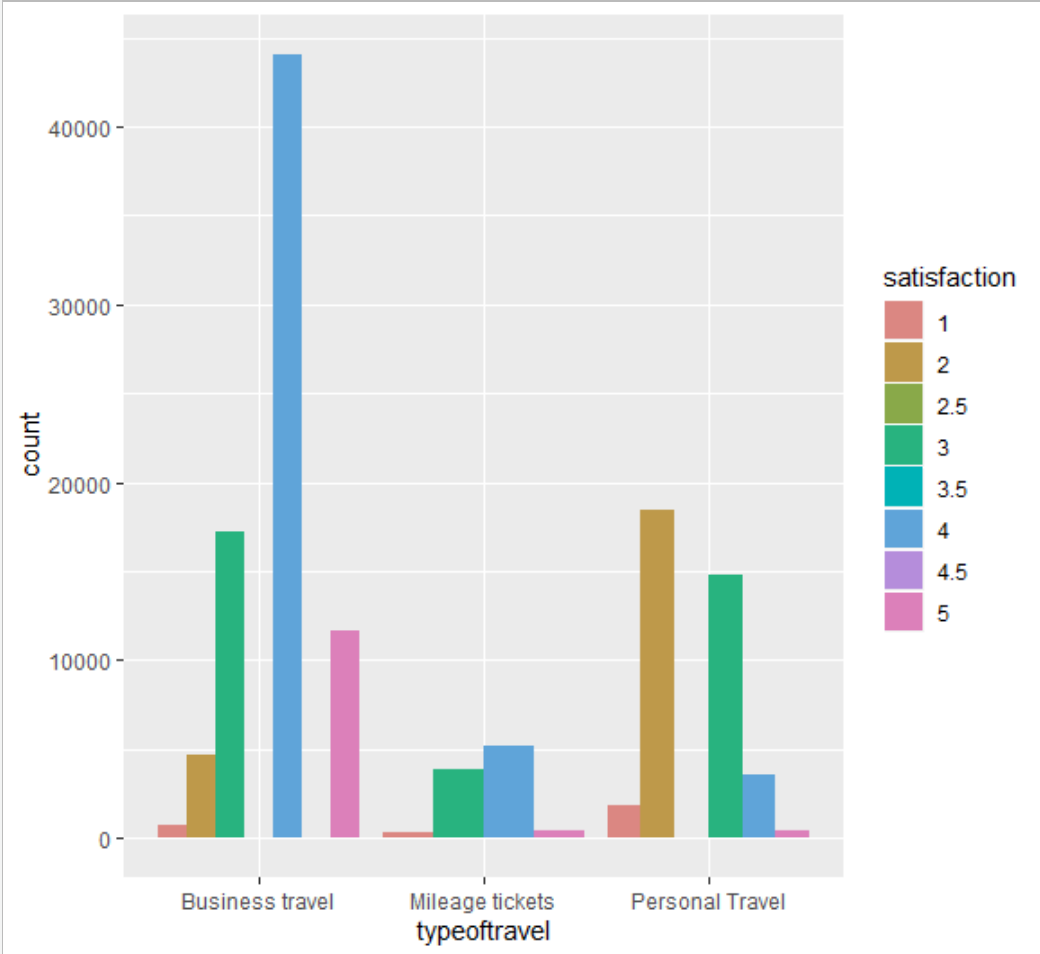


Figure . Bar Graph of Satisfaction based on type of travel.

**Observations**: The model allowed to narrow the variables based on p-value. The t-value show the propensity of likeliness of the variable impact to a rating or score. A higher t-value showed the propensity and likeliness to score favorable, while a negative score showed propensity and likeliness to score lower.

On other notes, this model does not consider geographical considerations. Not all the covariates in the model follow the parallel slopes assumption, due to the data we are analyzing. The ordinal regression seems like it would perform better if data sets were over 120,000 observations, even though our analysis seemed enough.

## Data Generalization for Kernel-based Support Vector Machines



Figure . Prediction and Confusion Matrix

**Observations:**  We had a lot of various issues with using this model. We first encounter problems trying to get the code to work. We had issues with a too broad of focus and impacts of variables that could work with KVSM model.

KVSM works with on binary classification type of data, so not all the data can work with this model. Variable we could not use was `Flight cancelled`, because it caused system issues. Variables that lowered accuracy was `Departure Delay in Minutes`, `Arrival Delay in Minutes`, `Flight Distance`, `No. of other Loyalty Cards`.

Next issue, we ran into was performance issues. Running too many variables on our laptops would either crash R or would not finish. We need to use a subset of the data in order to have a successful run of the model. For example, we took the smallest airline, by number of flights in order to ensure a successful run and execution of the model.

Our findings through this exercise, we found the highest accuracy was with this set of variables Age + Class + Price Sensitivity + Airline Status + Type of Travel, returning the best result from confusion matrix with an accuracy over 70%.

Trying to continue with other functions like NeuralNet failed with only one variable, Age, which prevented our further efforts of our analysis beyond simple execution and use of the confusion matrix.

# DATA VALIDATION

Data validation was an important step throughout our project. We conducted a thorough analysis of our results to ensure a more meaningful finding. We took a different approach to our information, using more advanced techniques and algorithms to first give insight to influential candidate variables (evaluation of variable p-values) and running a multitude of summary functions, descriptive statistic functions, histograms, and predictive models against the dataset. We constantly needed to make sure that our results were meaningful and accurate. As a result, we used several techniques.

1. Since our team is diverse with different skill sets, we all use different and common validation tools to make sure our calculations are sound and correct. We utilize tools like Microsoft Excel, calculators, RStudio, and MS SQL Server.
2. Within RStudio we used ggplot2 and histogram visualizations to help verify and compare the results. This way we can visually see the output results and see if there are any differences to our initial expectations.
3. We involved advice from Professor Krudys for guidance throughout the entire project, and always took his recommendations into consideration.
4. We also take advice from various R websites like cran.r-project.org, w3schools.in, stackoverflow.com, manuals.bioinformatics.ucr.edu, rdocumentation.org, rstudio.com
5. We ran combinations of descriptive statistics using a variety of different variables, ensuring that we were looking at the right variables and that our findings were supported effectively. In addition to this, we accompanied several of our findings with visualizations (charts and histograms) to back up our results.
6. Finally, we checked one another's work and do code reviews, especially when there are issues and concerns. We use a shared folder process with MS Teams. This way, it gives us the ability to to download one's work, run the code, do a code review, and test (where applicable).

# RESULTS

## Final Conclusions

We conducted a thorough analysis of our results to ensure a more meaningful finding. We took a different approach to our information, using more advanced techniques and algorithms to first give insight to influential candidate variables (evaluation of variable p-values) and running a multitude of summary functions, descriptive statistic functions, histograms, and predictive models against the dataset.

## Final Recommendations

**Focus 1: Qualify the Data.**

Here are our recommendations to the Airline Executive Management:

* Focus on all of the variables relating to customer service but put the most emphasis specifically on the top two variables: [Insert Variables Names].
* Put less emphasis on [Insert Variables Names] as they have less of an effect, but do not disregard them altogether as they do have a little bit of influence.
* Do not focus on variables such as [Insert Variables Names] as they don’t have much if any influence on customer satisfaction scoring.
* Based on SVM Classification analysis Airline Executives and Marketing Executives Key Strategy should pay close attention on [Insert Variables Names].
* Provide more programs and services that focus on servicing [Insert Variables Names].

**Focus 2: Expose/Explore the Market Data**

Here are our recommendations to the Airline Executive Management:

* Cultivate and maintain a strong customer-oriented experience by way of providing exceptional service and travel experience to Business Travelers who will likely continue regular frequent service and revenue to the airlines as well as high satisfaction rating.
* Provide more programs and services that focus on servicing [Insert Variables Names].
* The Airline Executive and Marketing should try to find more mechanisms andprograms to ensure better overall travel experience to [Insert Variables Names]..
* The Airline Executive and Marketing should consider a wider variety of demographics:
  + Type of Travel
  + Business Travelers
  + Better Service the Type of Class of Travel

## Final Notes

Finally, we will add that the Hyatt Hotels Corporation should:

* + …incentivise or invite the customers to be more responsive to its previous hotel stay surveys to overtly improve the hotel metrics.
  + … refine their surveys and make sure to get more input in the specific areas where customers are unsatisfied (i.e. that make customers unsatisfied), as these areas are negatively impact the NPS. This will help them better target the unsatisfied guests and understand the reasons of their dissatisfaction.
  + … always make sure to better organize their surveys and avoid having duplicates for the same question categories.
  + ... keep in mind that collecting high quality data is the most important step in the analysis.

Original Set of Questions

* Which type of travelers have the highest satisfaction rate? What are the reasons why?
* How do status and other factors impact satisfaction score?
* What factors likely give a higher satisfaction score or rate? How does location factor in this scoring?
* Are people more likely to recommend score higher satisfaction if they are travelling for business or leisure?
* Are we able to quantify a lifetime customer and predict or produce a greater probability of this population?

Additional or Recently Added Set of Questions

* Who are the customers? Who are the satisfied customers, and who are the unsatisfied customer?
* What factors directly affect a customer’s satisfaction, more specifically, their likelihood to score higher?
* Which modeling techniques can be used to predict the effect on likelihood to score a higher rating?
* How can the airlines better target customers, and what steps can Airline Executive Management do to improve customer satisfaction?

Recommendations

After looking at the above data supplied to us by Hyatt Hotels we as a group have decided to make the following recommendations to Hyatt Hotels.

Recommendation #1: Improve Hotel Customer Service

Using the data we saw that there was a very high correlation between hotel customer service and NPS type. Our recommendation would be to continue hotel customer service operations that are going well while potentially adding more representatives to ensure quality and speedy customer service.

Recommendation #2: Improve the condition of hotels

We saw a high correlation between hotel condition and NPS type. Our recommendation would be to ensure the condition of hotels meet or exceed those of the expectations of guests. With this Hyatt could run inspections every quarter to ensure all condition standards are met at a given location.

Recommendation #3: Ensure Mini-bars are available

Given the data we were able to see that certain hotel amenities affected the NPS scores of customers in the Hyatt Regency Hotels. Mini-Bars had a good effect on likelihood to recommend. We would recommend that Hyatt continues to keep mini-bars stocked and potentially even increase the supply of drinks in each mini-bar.

Recommendation #4: Make sure business centers are available

Our data showed that business centers had a great effect on likelihood to recommend in the hotels. We would recommend that business centers are available in Hyatt Hotels. We would also recommend that the technology and furniture within these centers are up to par as well.

# APPENDIX

## DATA DICTIONARY

**Satisfaction** – it is rated from 1 to 5, 5 means higher satisfied, and 1 is lowest level of satisfaction.

**Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.

**Age** – the specific customer’s age. That is starting from 15 to 85 years old.

**Gender** – male or female.

**Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.

**Year of First Flight** – this attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.

**No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.

**Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many times that customer fly with other Airlines.

**Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation

**No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefit such as, discounts.

**Shopping Amount at Airport** – showing the costumer’s result of how many products have been purchased. The range of shopping amount is from 0 to 875.

**Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.

**Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.

**Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.

**Flight date** – all of these data are abbreviating the passenger’s flight date travel, which were since 2014 and only in January, February, and March.

**Airline Code** – basically, it is unique two or three digits that mean what is the specific type of airline. There are several codes that consumers have been going with. For example, AA, AS, B6, and DL.

**Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and Fly ToSun Airlines Inc. This attribute provides what airline name that passenger have been used.

**Origin City** – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo HO.

**Origin State** – same thing as origin city such as, what state that customers have departed from? A good example, Texas, Ohio, Alaska, and Utah.

**Destination City** – the place to which passenger travels to. For example, Akron HO, Alpena MI, Austin TX, and Boston MA.

**Destination State** – also, it is the same thing as origin city, such as, to what state passenger travel to? Some example of destination states, Alaska, Kentucky, Iowa, and Florida.

**Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.

**Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage is starting from 0 until 1128 minutes.

**Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang of delayed minutes in this data are starting from 0 until 1115 minutes.

**Flight Cancelled** – occurs when the airline dose not operates the flight at all, and that is for a certain reason.

**Flight time in minutes** – indicate to period time to the destination.

**Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.

**Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

## References

* <https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/>
* <https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5>
* <https://stats.idre.ucla.edu/r/faq/ologit-coefficients/>

## Code Used

############################################################

# ------------------------------------------------------------------------------

# Descriptive Statistics

# Variables excluded from the model:

# ------------------------------------------------------------------------------

############################################################

# ------------------------------------------------------------------------------

# Airline Name # Produced only NA's in the output

# Airline code # It made the model output to difficult to interpret because it assigned coefficients to every airline code

# Origin City, Origin State, Destination City # same problem as above

# Flight Cancelled # None of the flights were canceled so the factor is only one level

# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------

# NA’s

# Descriptive Statistics of NAs

# Finding and analyzing where NAs exist and its value

# ------------------------------------------------------------------------------

library(readxl)

library(graphics)

SatisfactionSurvey <- read\_excel("C:/Users/16023/Desktop/IST 687/SatisfactionSurvey2\_2\_2\_2.xlsx")

sum(is.na(SatisfactionSurvey$Satisfaction))

sum(is.na(SatisfactionSurvey$’Airline Status’))

sum(is.na(SatisfactionSurvey$Age))

sum(is.na(SatisfactionSurvey$Gender))

sum(is.na(SatisfactionSurvey$’Price Sensitivity’))

sum(is.na(SatisfactionSurvey$’Year of First Flight’))

sum(is.na(SatisfactionSurvey$’No of Flights p.a.’))

sum(is.na(SatisfactionSurvey$’Type of Travel’))

sum(is.na(SatisfactionSurvey$’No. of other Loyalty Cards’))

sum(is.na(SatisfactionSurvey$’Shopping Amount at Airport’))

sum(is.na(SatisfactionSurvey$’Eating and Drinking at Airport’))

sum(is.na(SatisfactionSurvey$Class))

sum(is.na(SatisfactionSurvey$’Day of Month’))

sum(is.na(SatisfactionSurvey$’Flight date’))

sum(is.na(SatisfactionSurvey$’Airline Code’))

sum(is.na(SatisfactionSurvey$’Airline Name’))

sum(is.na(SatisfactionSurvey$’Orgin City’))

sum(is.na(SatisfactionSurvey$’Origin State’))

sum(is.na(SatisfactionSurvey$’Destination City’))

sum(is.na(SatisfactionSurvey$’Destination State’))

sum(is.na(SatisfactionSurvey$’Scheduled Departure Hour’))

sum(is.na(SatisfactionSurvey$’Departure Delay in Minutes’))

sum(is.na(SatisfactionSurvey$’Arrival Delay in Minutes’))

sum(is.na(SatisfactionSurvey$’Flight cancelled’))

sum(is.na(SatisfactionSurvey$’Flight time in minutes’))

sum(is.na(SatisfactionSurvey$’Flight Distance’))

sum(is.na(SatisfactionSurvey$’Arrival Delay greater 5 Mins’))

sum(is.na(SatisfactionSurvey$’% of Flight with other Airlines’))

View(SatisfactionSurvey)

Nasincol <- c(sum(is.na(SatisfactionSurvey$Satisfaction)),sum(is.na(SatisfactionSurvey$’Airline Status’)), sum(is.na(SatisfactionSurvey$Age)), sum(is.na(SatisfactionSurvey$Gender)), sum(is.na(SatisfactionSurvey$’Price Sensitivity’)), sum(is.na(SatisfactionSurvey$’Year of First Flight’)), sum(is.na(SatisfactionSurvey$’No of Flights p.a.’)), sum(is.na(SatisfactionSurvey$’Type of Travel’)), sum(is.na(SatisfactionSurvey$’No. of other Loyalty Cards’)), sum(is.na(SatisfactionSurvey$’Shopping Amount at Airport’)), sum(is.na(SatisfactionSurvey$’Eating and Drinking at Airport’)), sum(is.na(SatisfactionSurvey$Class)), sum(is.na(SatisfactionSurvey$’Day of Month’)), sum(is.na(SatisfactionSurvey$’Flight date’)), sum(is.na(SatisfactionSurvey$’Airline Code’)), sum(is.na(SatisfactionSurvey$’Airline Name’)), sum(is.na(SatisfactionSurvey$’Orgin City’)), sum(is.na(SatisfactionSurvey$’Origin State’)), sum(is.na(SatisfactionSurvey$’Destination City’)), sum(is.na(SatisfactionSurvey$’Destination State’)), sum(is.na(SatisfactionSurvey$’Scheduled Departure Hour’)), sum(is.na(SatisfactionSurvey$’Departure Delay in Minutes’)), sum(is.na(SatisfactionSurvey$’Arrival Delay in Minutes’)), sum(is.na(SatisfactionSurvey$’Flight cancelled’)), sum(is.na(SatisfactionSurvey$’Flight time in minutes’)), sum(is.na(SatisfactionSurvey$’Flight Distance’)), sum(is.na(SatisfactionSurvey$’Arrival Delay greater 5 Mins’)), sum(is.na(SatisfactionSurvey$’% of Flight with other Airlines’)))

ColumnNames <- "Column Names"

barNA <- barplot(Nasincol, main="NA's in each Column", names.arg=colnames(SatisfactionSurvey), las=2)

# ------------------------------------------------------------------------------

# Descriptive statistics by type of travel

# Data Cleansing

# ------------------------------------------------------------------------------

SatisfactionSurvey <- na.omit(SatisfactionSurvey)

SatisfactionSurvey$Satisfaction <- as.numeric(gsub(",","", SatisfactionSurvey$Satisfaction))

SatisfactionSurvey$’No of Flights p.a.’ <- as.numeric(gsub(",","", SatisfactionSurvey$’No of Flights p.a.’))

SatisfactionSurvey$’Departure Delay in Minutes’ <- as.numeric(gsub(",","", SatisfactionSurvey$’Departure Delay in Minutes’))

SatisfactionSurvey$’Arrival Delay in Minutes’ <- as.numeric(gsub(",","", SatisfactionSurvey$’Arrival Delay in Minutes’))

SatisfactionSurvey$’Flight time in minutes’ <- as.numeric(gsub(",","", SatisfactionSurvey$’Flight time in minutes’))

SatisfactionSurvey$’Flight Distance’ <- as.numeric(gsub(",","", SatisfactionSurvey$’Flight Distance’))

SatisfactionSurvey$’Type of Travel’ <- as.character(gsub(" ","", SatisfactionSurvey$’Type of Travel’))

# ------------------------------------------------------------------------------

# Summary

# Sections: will contain all the Graphing/ histograms of initial Satisfaction findings

# ------------------------------------------------------------------------------

SatisfactionSurveyBusiness <- SatisfactionSurvey[SatisfactionSurvey$’Type of Travel’== 'Businesstravel' , ]

SatisfactionSurveyPersonal <- SatisfactionSurvey[SatisfactionSurvey$’Type of Travel’=='PersonalTravel',]

SatisfactionSurveyMileage <- SatisfactionSurvey[SatisfactionSurvey$’Type of Travel’=='Mileagetickets',]

colnames(SatisfactionSurvey)

length(SatisfactionSurveyBusiness)

# ------------------------------------------------------------------------------

# Summary and graph/histogram of all the data

# ------------------------------------------------------------------------------

hist(SatisfactionSurvey$Satisfaction)

summary(SatisfactionSurvey$Satisfaction)

sd(SatisfactionSurvey$Satisfaction)

# ------------------------------------------------------------------------------

# This is the data Subcategorized by Type of Travel

# ------------------------------------------------------------------------------

SatisfactionSurveyBusiness <- SatisfactionSurvey[SatisfactionSurvey$'Type of Travel'== 'Businesstravel' , ]

SatisfactionSurveyPersonal <- SatisfactionSurvey[SatisfactionSurvey$'Type of Travel'=='PersonalTravel',]

SatisfactionSurveyMileage <- SatisfactionSurvey[SatisfactionSurvey$'Type of Travel'=='Mileagetickets',]

summary(SatisfactionSurveyBusiness$Satisfaction)

summary(SatisfactionSurveyPersonal$Satisfaction)

summary(SatisfactionSurveyMileage$Satisfaction)

hist(SatisfactionSurveyBusiness$Satisfaction)

hist(SatisfactionSurveyPersonal$Satisfaction)

hist(SatisfactionSurveyMileage$Satisfaction)

sd(SatisfactionSurveyBusiness$Satisfaction)

sd(SatisfactionSurveyPersonal$Satisfaction)

sd(SatisfactionSurveyMileage$Satisfaction)

# ------------------------------------------------------------------------------

# This is the data subcategorized by Airline Status

# ------------------------------------------------------------------------------

SatisfactionSurveyGold <- SatisfactionSurvey[SatisfactionSurvey$'Airline Status' == 'Gold',]

SatisfactionSurveyBlue <- SatisfactionSurvey[SatisfactionSurvey$'Airline Status' == 'Blue',]

SatisfactionSurveySilver <- SatisfactionSurvey[SatisfactionSurvey$'Airline Status' == 'Silver',]

SatisfactionSurveyPlatinum <- SatisfactionSurvey[SatisfactionSurvey$'Airline Status' == 'Platinum',]

hist(SatisfactionSurveyGold$Satisfaction)

hist(SatisfactionSurveyBlue$Satisfaction)

hist(SatisfactionSurveySilver$Satisfaction)

hist(SatisfactionSurveyPlatinum$Satisfaction)

summary(SatisfactionSurveyGold$Satisfaction)

summary(SatisfactionSurveyBlue$Satisfaction)

summary(SatisfactionSurveySilver$Satisfaction)

summary(SatisfactionSurveyPlatinum$Satisfaction)

sd(SatisfactionSurveyGold$Satisfaction)

sd(SatisfactionSurveyBlue$Satisfaction)

sd(SatisfactionSurveySilver$Satisfaction)

sd(SatisfactionSurveyPlatinum$Satisfaction)

# ------------------------------------------------------------------------------

# Satisfaction by State

# Creating an initial Heat Map of initial travel for data preparation

# and Descriptive Statistical analysis

# ------------------------------------------------------------------------------

install.packages("reshape2")

library(reshape2)

satairport <- SatisfactionSurvey[ ,c(1,17)]

airportdf <- tapply(satairport$Satisfaction, satairport$’Airline Name’, mean)

View(airportdf)

satcity <- SatisfactionSurvey[ , c(1,18,20)]

Ocitydf <- tapply(satcity$Satisfaction, satcity$’Orgin City’ , mean)

Dcitydf <- tapply(satcity$Satisfaction, satcity$’Destination City’ , mean)

Ocitydf

Dcitydf

satmap <- SatisfactionSurvey[ ,c(1,19,21)]

str(satmap)

satmap$’Origin State’ <- as.factor(satmap$’Origin State’)

satmap$’Destination State’ <- as.factor(satmap$’Destination State’)

origin <- tapply(satmap$Satisfaction, satmap$’Origin State’, mean)

origin

destination <- tapply(satmap$Satisfaction, satmap$’Destination State’, mean)

satbystates <- data.frame(origin, destination, origin+destination, (origin+destination)/2)

colnames(satbystates) <- c("Origin Rate", "Destination Rate", "Sum", "Average")

library(ggplot2)

install.packages("maps")

str(state.name)

head(state.name)

state.name[3]

#

dummyDF <- data.frame(state.name, stringsAsFactors=FALSE)

dummyDF$state <- tolower(dummyDF$state.name)

#

us <- map\_data("state") ## map\_data is a function in maps package

us

#

map.simple <- ggplot(dummyDF, aes(map\_id = state))

map.simple <- map.simple +

geom\_map(map = us, fill="white", color="black")

map.simple

map.simple <- map.simple +

expand\_limits(x = us$long, y = us$lat)

map.simple

###########################################################

# ------------------------------------------------------------------------------

# Final Project - Simple Linear Regression Model

# ------------------------------------------------------------------------------

###########################################################

# Read in the data

library(pacman)

p\_load(readxl)

data <- read\_excel("SatisfactionSurvey2\_2\_2\_2.xlsx")

# Now search for NA's in the data

any(is.na(data))

# Remove the NA's

data <- na.omit(data)

# Now create the model

lm1 <- lm(data = data, data$Satisfaction ~ data$’Airline Status’ +

data$Age +

data$Gender +

data$’Price Sensitivity’ +

data$’Year of First Flight’ +

data$’No of Flights p.a.’ +

data$’Type of Travel’ +

data$’No. of other Loyalty Cards’ +

data$’Shopping Amount at Airport’ +

data$’Eating and Drinking at Airport’ +

data$Class +

data$’Day of Month’ +

data$’Flight date’ +

data$’Scheduled Departure Hour’ +

data$’Departure Delay in Minutes’ +

data$’Arrival Delay in Minutes’ +

data$’Flight time in minutes’ +

data$’Flight Distance’ +

data$’Arrival Delay greater 5 Mins’ +

data$’% of Flight with other Airlines’)

summary(lm1)

plot(lm1)

############################################################

# ------------------------------------------------------------------------------

# Final Project - Ordinal Logistic Regression Model

# ------------------------------------------------------------------------------

############################################################

# here is the website I used for reference:

# https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5

# Install some packages (MASS is the most important)

install.packages("pacman")

library(pacman)

p\_load(broom, MASS, tidyverse, ggplot2, readxl)

# ------------------------------------------------------------------------------

# Ordinal Logistic Regression

# ------------------------------------------------------------------------------

# 1. Read in the data and omit the NA's:

survey\_data <- na.omit(SatisfactionSurvey2\_2\_2\_2)

# Now lets check out some summary statistics of our model

summary(survey\_data)

# the average satisfaction score turns out to be 3.384

hist(typeoftravel, satisfaction)

# check the number of observations we have

NROW(survey\_data)

# we need to make the satisfaction score a factor and order it

# This basically orders the scores as 5 > 4 > 3 > 2 > 1

satisfaction <- survey\_data$Satisfaction

satisfaction <- factor(satisfaction, levels = c("1","2","2.5",

"3", "3.5", "4","4.5",

"5"), ordered = TRUE)

# Check the levels

levels(satisfaction)

max(satisfaction)

# lets look at a bar plot of the satisfaction levels

ggplot(survey\_data, aes(satisfaction, fill = satisfaction)) +

geom\_bar() + scale\_fill\_hue(c = 60) +

theme(legend.position = 'none')

#============================================================================

# We can use categorical and continuous data in the model

# we'll define some variables before we build the model.

# flight\_cancelled will be run through the factor function

flight\_cancelled <- factor(survey\_data$’Flight cancelled’, ordered = FALSE)

# check the levels

levels(flight\_cancelled)

# none of the flights were cancelled. We'll omit this variable from the model

# now we'll do arrival delay > 5 mins

arrival\_delay <- factor(survey\_data$’Arrival Delay greater 5 Mins’, ordered = FALSE)

levels(arrival\_delay)

# price sensitivity

# note: the continuous variables DO NOT need to be ran through the factor function

price\_sensitivity <- survey\_data$’Price Sensitivity’

# airline status

# originally I ran this through the factor function but it was unecessary

airline\_status <- survey\_data$’Airline Status’

# Gender

gender <- factor(survey\_data$Gender, ordered = FALSE)

levels(gender)

# Type of Travel

typeoftravel <- survey\_data$’Type of Travel’

typeoftravel <- factor(typeoftravel, ordered = FALSE)

levels(typeoftravel)

# Model 1

# we'll use the plor function from the MASS package

# plor = proportional linear odds regression

#====================================================================================

model1 <- polr(satisfaction ~ airline\_status +

survey\_data$Age +

gender +

price\_sensitivity +

survey\_data$’Year of First Flight’ +

survey\_data$’No of Flights p.a.’ +

survey\_data$’% of Flight with other Airlines’ +

typeoftravel +

survey\_data$’No. of other Loyalty Cards’ +

survey\_data$’Shopping Amount at Airport’ +

survey\_data$’Eating and Drinking at Airport’ +

survey\_data$Class +

survey\_data$’Day of Month’ +

survey\_data$’Scheduled Departure Hour’ +

survey\_data$’Departure Delay in Minutes’ +

survey\_data$’Arrival Delay in Minutes’ +

survey\_data$’Flight time in minutes’ +

survey\_data$’Flight Distance’ +

arrival\_delay,

data = survey\_data, method = "logistic", Hess = TRUE)

summary(model1)

# We'll use the t-value and compare it to the normal distribution to calculate

# p-values and test for statistical significance

summary\_table <- coef(summary(model1))

pval <- pnorm(abs(summary\_table[,"t value"]), lower.tail = F)\*2

summary\_table <- cbind(summary\_table, "p value" = round(pval, 3))

summary\_table

# Put the summary into a data frame

model1\_df <- data.frame(summary\_table)

model1\_df

# Export the table to excel

write.csv(model1\_df, 'model1\_df.csv')

# Model 2

#========================================================================

# Ok, now I'll omit the variables which weren't statistically significant

# They are:

# % of flight w/ other airlines

# no. of other loyalty cards

# eating and drinking at the airport

# Day of month

# Dep. Delay in mins

# Arrival delay in mins

# Flight time in mins

# Flight Distance

# Before I begin, let me re-name variables:

age <- survey\_data$Age

yearfirstflight <- survey\_data$’Year of First Flight’

flightspa <- survey\_data$’No of Flights p.a.’

dephour <- survey\_data$’Scheduled Departure Hour’

shoppingamount <- survey\_data$’Shopping Amount at Airport’

class <- factor(survey\_data$Class)

# Now I'll rebuild the model

model2 <- polr(satisfaction ~

airline\_status + age +

gender + price\_sensitivity +

yearfirstflight +

flightspa +

typeoftravel +

shoppingamount +

class +

dephour + arrival\_delay,

data = survey,

method = "logistic",

Hess = TRUE)

# Now lets see the summary

summary(model2)

# We'll use the t-value to calculate the p-value from the model

summary\_table2 <- coef(summary(model2))

pval <- pnorm(abs(summary\_table2[,"t value"]), lower.tail = F)\*2

summary\_table2 <- cbind(summary\_table2, "p value" = round(pval, 3))

summary\_table2

# put it into a data frame

model2\_df <- data.frame(summary\_table2)

model2\_df

# Now lets export the table.

write.csv(model2\_df, 'Model2.csv')

# calculating a prediction

#============================================================

# use the intercepts and coefficients to make predictions

logit\_1 <- 23.5546295609 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_2 <- 26.5647146709 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_3 <- 26.5650099453 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_4 <- 28.7597098017 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_5 <- 28.7597841666 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_6 <- 31.8373790535 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

logit\_7 <- 31.8373897345 - ((1.6624047883\*1) +

(35\*-0.0038116973) +

(-0.1125294852\*1) +

(0.0148895899\*2005) +

(-0.0088588513\*10) +

(-2.8623654934\*1) +

( 0.0003610423\*60) +

( -0.2098222068\*1) +

( 0.0113269368\*18) +

(-0.9756755701\*1))

# now use the logits and exponentiate to determine probabilities

# these are ONLY probabilities for the individual attributes

# listed in the word document write up for this model

#============================================================

prob\_1 <- exp(logit\_1)/(1 + exp(logit\_1))

prob\_2 <- exp(logit\_2)/(1 + exp(logit\_2))

prob\_2 <- prob\_2 - prob\_1

prob\_3 <- exp(logit\_3)/(1 + exp(logit\_3))

prob\_3 <- prob\_3 - prob\_2 - prob\_1

prob\_4 <- exp(logit\_4)/(1 + exp(logit\_4))

prob\_4 <- prob\_4 - prob\_3 - prob\_2 - prob\_1

prob\_5 <- exp(logit\_5)/(1 + exp(logit\_5))

prob\_5 <- prob\_5 - prob\_4 - prob\_3 - prob\_2 - prob\_1

prob\_6 <- exp(logit\_6)/(1 + exp(logit\_6))

prob\_6 <- prob\_6 - prob\_5 - prob\_4 - prob\_3 - prob\_2 - prob\_1

prob\_7 <- exp(logit\_7)/(1 + exp(logit\_7))

prob\_7 <- prob\_7 - prob\_6 - prob\_5 - prob\_4 - prob\_3 - prob\_2 - prob\_1

prob\_8 <- 1 - (prob\_1 + prob\_2 + prob\_3 + prob\_4 + prob\_5 + prob\_6 + prob\_7)

# here is my barplot for type of travel

#===========================================================

ggplot(survey\_data, aes(typeoftravel, fill = satisfaction)) +

geom\_bar(position = position\_dodge()) + scale\_fill\_hue(c = 60)

# now we're gonna determine the goodness of fit of model2.

# this is where i reference paralell slopes assumption in word

#============================================================

p\_load(Hmisc)

summary(model2)

levels(satisfaction)

sf <- function(y) {

c('Y>=1' = qlogis(mean(y >= 1)),

'Y>=2' = qlogis(mean(y >= 2)),

'Y>=2.5' = qlogis(mean(y >= 2.5)),

'Y>=3' = qlogis(mean(y >= 3)),

'Y>=3.5' = qlogis(mean(y >= 3.5)),

'Y>=4' = qlogis(mean(y >= 4)),

'Y>=4.5' = qlogis(mean(y >= 4.5)),

'Y>=5' = qlogis(mean(y >= 5)))

}

(s <- with(model2\_df, summary(as.numeric(satisfaction) ~ airline\_status +

age +

gender +

price\_sensitivity +

yearfirstflight +

flightspa +

typeoftravel +

shoppingamount +

class +

dephour +

arrival\_delay,

fun = sf)))

#=============================================================

#https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

#https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5

#https://stats.idre.ucla.edu/r/faq/ologit-coefficients/

############################################################

# ------------------------------------------------------------------------------

# Final Project – Kernel-Base Support Vector Model (KSVM)

# ------------------------------------------------------------------------------

###########################################################

SatisfactionSurvey <- na.omit(SatisfactionSurvey\_CoolAir)

avg <- mean(SatisfactionSurvey$Satisfaction)

install.packages("kernlab")

library("kernlab")

library(e1071)

install.packages("neuralnet")

library(neuralnet)

install.packages("caret")

library(caret)

summary(SatisfactionSurvey)

randIndex <- sample(1:dim(SatisfactionSurvey)[1])

head(randIndex)

length(randIndex)

dim(SatisfactionSurvey)

#

# # In order to split data, create a 2/3 cutpoint and round the number

cutpoint2\_3 <- floor(2\*dim(SatisfactionSurvey)[1]/3)

# check the 2/3 cutpoint

cutpoint2\_3

#

# create train data set, which contains the first 2/3 of overall data

#

trainData <- SatisfactionSurvey[randIndex[1:cutpoint2\_3],]

dim(trainData)

head(trainData)

#

# create test data, which contains the left 1/3 of the overall data

#

testData <- SatisfactionSurvey[randIndex[(cutpoint2\_3+1):dim(SatisfactionSurvey)[1]],]

dim(testData) # check test data set

head(trainData)

trainData$goodsat <- ifelse(trainData$Satisfaction<avg, 0, 1)

testData$goodsat <- ifelse(testData$Satisfaction<avg, 0, 1)

trainData <- trainData[,-1]

testData <- testData[,-1]

str(trainData)

testData$goodsat <- as.factor(testData$goodsat)

modelcs <- ksvm(goodsat ~ Age + Class + ‘Price Sensitivity’ + ‘Airline Status’ + ‘Type of Travel’,

data = trainData,

kernel="rbfdot",

kpar = "automatic",

C = 10,

cross = 10,

prob.model = TRUE)

prediction <- predict(modelcs, testData)

results <- data.frame(testData$goodsat,round(prediction))

colnames(results) <- c("Actual", "Prediction")

str(results)

results$Prediction <- as.factor(results$Prediction)

confusionMatrix(results$Actual,results$Prediction)

#Variables I couldnt Use ‘Flight cancelled’

#Variables that Lowered the accuracy ‘Departure Delay in Minutes’, ‘Arrival Delay in Minutes’, ‘Flight Distance’, ‘No. of other Loyalty Cards’

# Failed NeuralNet Code the only variable that was used was Age

satSurvey <- SatisfactionSurvey2\_2

satSurvey <- na.omit(satSurvey)

randIndex <- sample(1:dim(satSurvey)[1])

head(randIndex)

length(randIndex)

dim(satSurvey)

#

# # In order to split data, create a 2/3 cutpoint and round the number

cutpoint2\_3 <- floor(2\*dim(satSurvey)[1]/3)

# check the 2/3 cutpoint

cutpoint2\_3

#

trainData2 <- satSurvey[randIndex[1:cutpoint2\_3],]

dim(trainData2)

head(trainData2)

testData2 <- satSurvey[randIndex[(cutpoint2\_3+1):dim(satSurvey)[1]],]

dim(testData2) # check test data set

head(trainData2)

trainData2$goodsat <- ifelse(trainData2$Satisfaction<avg, 0, 1)

testData2$goodsat <- ifelse(testData2$Satisfaction<avg, 0, 1)

trainData2 <- trainData2[,-1]

testData2 <- testData2[,-1]

testData2$goodsat <- as.factor(testData2$goodsat)

trainData2$goodsat <- as.factor(trainData2$goodsat)

trainData2$’Flight Distance’ <- as.numeric(trainData2$’Flight Distance’)

str(trainData2)

x <- neuralnet(goodsat ~ Age, trainData2, hidden = 2, lifesign="minimal", linear.output = FALSE, threshold = .1)

plot(x)