JOYCE WOZNICA
AIRLINE SURVEY ANALYSIS PROJECT

Joyce Woznica
jlwoznic@syr.edu
March 19, 2019

Contents

List of Figures

Introduction

Presented with data from several travelers rating their overall satisfaction on select airline trips, I needed to study, analyze and review the data collected to look for any patterns or other indications that certain data collected for each trip has some bearing on the satisfaction rating of the individual traveler. When these pattern or indications are found, I must arrive at recommendations for ***Southeast Airlines*** (code: ***US***) so that they can use the studied information to improve their overall satisfaction with current and prospective travelers.

In addition, I wanted to provide Southeast Airlines with information on predicting which flyers that were likely to have low satisfaction. With this information, the airline can make decisions to market to their more satisfied base as well as target pulling up satisfaction in this lower satisfaction group of travelers or take other action.

This document provides details into the analysis done on the satisfaction survey provided, the conclusions and any possible recommendations or actions to be taken by Southeast Airlines to address poor customer satisfaction.

Business Questions

As I initially reviewed the dataset, I looked into what variables might be of interest in studying satisfaction. I gathered information on which variables to analyze to answer the business questions that I determined would be pertinent for Southeast Airlines (as well as any additional questions that might be appropriate as the analysis progresses). As I further analyzed the data, some alterations to the existing data as well as some additional questions were added. I selected the following as the business questions to research as well as some additional questions added as we reviewed the data and did analysis.

As part of reviewing the data, it is important to understand the following:

1. Which airlines (Airline Code/Airline Name) have the highest satisfaction rates?
    a. Why? What do they have in common?
    b. What is contributing to this rate?
    c. How does Southeast stack up against the other airlines?
1. How do cancellations and delays affect satisfaction?
2. What contributes to first time flyers satisfaction?
3. Does membership in a frequent flyer program affect satisfaction?
4. What variable(s) (or combination of variable(s)) contribute the most to satisfaction for flyers?
    1. Class of Service
    2. Origin/Destination City and State
    3. Travel Day of the month
    4. Type of plane (Airline Code)
    5. Gender and/or Age
    6. Airline Status
    7. Distance

To begin to analyze the data, I reviewed the variables collected associated with each Satisfaction value and came up with a list of variables that were pertinent to review. This does not mean that I did not evaluate other variables as part of this analysis, as will be seen in this document, but this was my initial list based on the selection of initial business questions.

Variable	Definition
Satisfaction	Rating from 1 (low) to 5 (high) of satisfaction
Airline Status	Status with airline (frequent flyer)
Age	Age of traveler
Gender	Gender of traveler – male or female
Number of Flights	Number of flights taken by this traveler
Class	Class of travel (first, business, etc.)
Day of Month	Day of the month of travel
Airline Code	Specific airline code name
Airline Name	Specific airline name
Origin City and Origin State	Origin City and State for flight
Destination City Destination State	Destination City and State for flight
Departure Delay in Minutes	How many minutes of departure delay
Arrival Delay in Minutes	How many minutes of arrival delay
Flight Cancelled	If flight was cancelled (yes/no)
Flight Distance	Distance between origin and destination

Figure 1: Table of Candidate Variables

Data Cleanse, Munge and Preparation

The dataset provided had approximately 130,000 survey responses with up to 25 fields in each survey. The amount of data per airline varied and further information on these details is provided later in this document. You will see later in the document that the sheer size of this data set led me to take subsets of the data as I performed analysis so that I could get results as on my personal computer, some specific functions around modeling took significant time to complete.

Data Load

To load the data, I took advantage of the "readxl" package and the *read_excel* R function to read the provided Excel file into R. I then converted this data into a data frame for usage and manipulation for analysis.

I renamed the columns to something simpler for manipulating and analyzing the data:

- Satisfaction = Satisfaction
- Airline Status = Status
- Age = Age
- Gender = Gender
- Price Sensitivity = PriceSens
- Year of First Flight = FFYear
- % of Flight with other Airlines = PercOther
- No of Flight p.a. = NumFlights
- Type of Travel = TravelType
- No. of other Loyalty Cards = NumCards

- Shopping Amount at Airport = ShopAmount
- Eating and Drinking at Airport = EatDrink
- Class = Class
- Day of Month = MonthDay
- Flight date = FlightDate
- Airline Code = AirlineCode
- Airline Name = Airline
- Origin City = OrigCity
- Origin State = OrigState
- Destination City = DestCity
- Destination State = DestState
- Scheduled Departure Hour = SchDeptHour
- Departure Delay in Minutes = DeptDelayMins
- Arrival Delay in Minutes = ArrDelayMins
- Flight cancelled = Cancelled
- Flight time in minutes = FlightMins
- Flight Distance = Distance
- Arrival Delay greater than 5 Mins = ArrDelayGT5

Data Cleanse

In some cases, the data was blank in a specific field. This was addressed in a way that would minimize the skewing or misrepresentation of the data. The sub-sections below describe how I determined the following would be used to handle missing data.

Cancelled Flights

When a flight was cancelled, no associated values for the following variables were in the survey results:

- Departure Delay in Minutes (DeptDelayMins)
- Arrival Delay in Minutes (ArrDelayMins)
- Flight time in Minutes (FlightMins)

In this situation, I determined that since there was no delay, arrival or flight time since the flight had not been taken at all, to set these values to 0 (zero).

Not Applicable

There were also some situations, other than a cancelled flight, that had with NA or no value for the following variables:

- Arrival Delay in Minutes (ArrDelayMins)
- Flight time in Minutes (FlightMins)

In this case, I also chose to set the arrival delay and flight time in minutes to a value of 0.

In a few more cases, I believe there was just bad data, so I removed any remaining rows (observations/surveys) that had NA values and did not use them in the data set analyzed.

> For the code used to do this data load, cleanse, munge and preparation, please see the section entitled Data Load and Clean-up beginning on page 28 in the

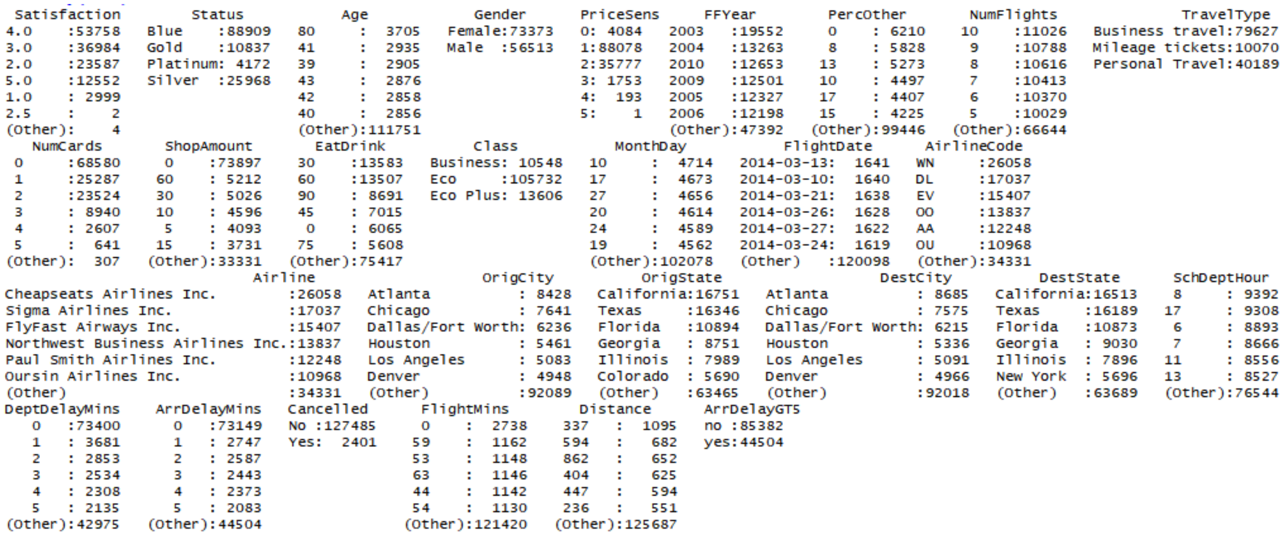
Appendix. Other operations to cleanse and manipulate the data for specific analysis will be found in the sections where that analysis took place.

Use of Descriptive Statistics

The initial analysis of the data included some general study of the data. To do this, I used some generic tools to find out a bit about the data before doing any extensive modeling and visualization.

General Information

First, I ran a general summary to see what type of data I had. For example, the frequency of various data in the satisfaction survey as shown in the following figure.

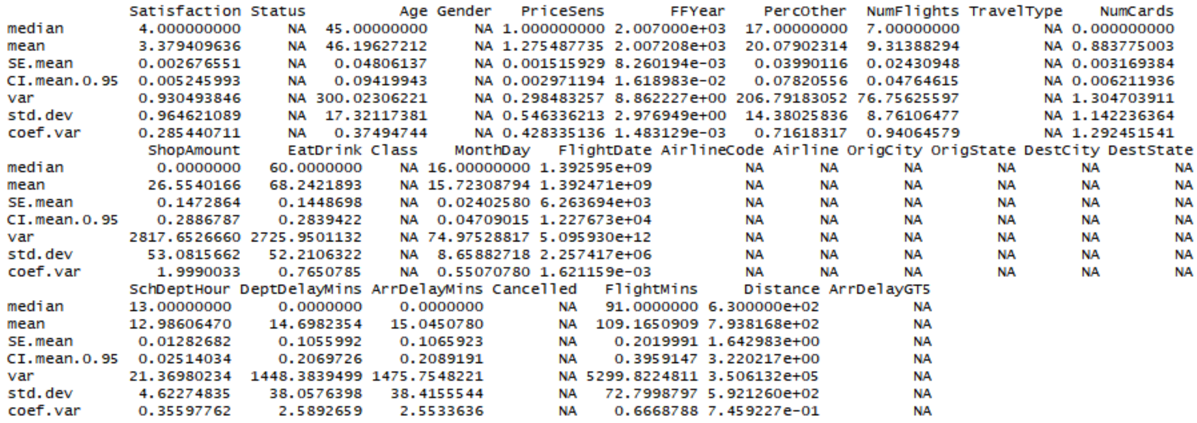


```
  Satisfaction        Status          Age             Gender         PriceSens    FFYear          PercOther        NumFlights               TravelType    
 4.0     :53758   Blue    :88909   80      :  3705   Female:73373   0: 4084   2003   :19552   0      : 6210   10     :11026   Business travel:79627  
 3.0     :36984   Gold    :10837   41      :  2935   Male  :56513   1:88078   2004   :13263   8      : 5828   9      :10788   Mileage tickets:10070  
 2.0     :23587   Platinum: 4172   39      :  2905                  2:35777   2010   :12653   13     : 5273   8      :10616   Personal Travel:40189  
 5.0     :12552   Silver  :25968   43      :  2876                  3: 1753   2009   :12501   10     : 4497   7      :10413                          
 1.0     : 2999                    42      :  2858                  4:  193   2005   :12327   17     : 4407   6      :10370                          
 2.5     :    2                    40      :  2856                  5:    1   2006   :12198   15     : 4225   5      :10029                          
 (Other):    4                    (Other):111751                            (Other):47392   (Other):99446   (Other):66644                          
    NumCards       ShopAmount       EatDrink            Class           MonthDay          FlightDate         AirlineCode   
 0      :68580   0      :73897   30      :13583   Business: 10548   10      :  4714   2014-03-13:  1641   WN      :26058  
 1      :25287   60     : 5212   60      :13507   Eco      :105732   17      :  4673   2014-03-10:  1640   DL      :17037  
 2      :23524   30     : 5026   90      : 8691   Eco Plus: 13606   27      :  4656   2014-03-21:  1638   EV      :15407  
 3      : 8940   10     : 4596   45      : 7015                     20      :  4614   2014-03-26:  1628   OO      :13837  
 4      : 2607   5      : 4093   0       : 6065                     24      :  4589   2014-03-27:  1622   AA      :12248  
 5      :  641   15     : 3731   75      : 5608                     19      :  4562   2014-03-24:  1619   OU      :10968  
 (Other):  307   (Other):33331   (Other):75417                      (Other):102078   (Other)   :120098   (Other):34331  
                              Airline                  OrigCity          OrigState                DestCity          DestState        SchDeptHour    
 Cheapseats Airlines Inc.          :26058   Atlanta          : 8428   California:16751   Atlanta          : 8685   California:16513   8      : 9392  
 Sigma Airlines Inc.               :17037   Chicago          : 7641   Texas     :16346   Chicago          : 7575   Texas     :16189   17     : 9308  
 FlyFast Airways Inc.              :15407   Dallas/Fort Worth: 6236   Florida   :10894   Dallas/Fort Worth: 6215   Florida   :10873   6      : 8893  
 Northwest Business Airlines Inc.:13837   Houston          : 5461   Georgia   : 8751   Houston          : 5336   Georgia   : 9030   7      : 8666  
 Paul Smith Airlines Inc.          :12248   Los Angeles      : 5083   Illinois  : 7989   Los Angeles      : 5091   Illinois  : 7896   11     : 8556  
 Oursin Airlines Inc.              :10968   Denver           : 4948   Colorado  : 5690   Denver           : 4966   New York  : 5696   13     : 8527  
 (Other)                           :34331   (Other)          :92089   (Other)   :63465   (Other)          :92018   (Other)   :63689   (Other):76544  
 DeptDelayMins     ArrDelayMins    Cancelled         FlightMins          Distance       ArrDelayGT5  
    0   :73400      0   :73149    No :127485      0      :  2738    337    :  1095   no :85382   
    1   : 3681      1   : 2747    Yes:  2401     59      :  1162    594    :   682   yes:44504   
    2   : 2853      2   : 2587                   53      :  1148    862    :   652               
    3   : 2534      3   : 2443                   63      :  1146    404    :   625               
    4   : 2308      4   : 2373                   44      :  1142    447    :   594               
    5   : 2135      5   : 2083                   54      :  1130    236    :   551               
 (Other):42975   (Other):44504                 (Other):121420   (Other):125687               
```

Figure 2: Summary Information from the Satisfaction Survey

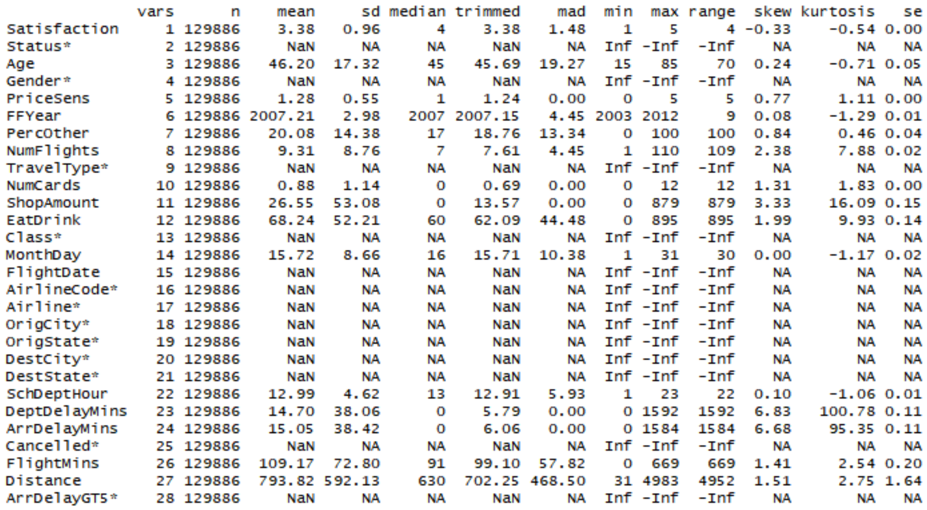
Descriptive Statistics

To begin to evaluate the variables and information, I reviewed standard descriptive statistics on the continuous variables in the dataset.



```
             Satisfaction Status          Age Gender   PriceSens       FFYear    PercOther NumFlights TravelType    NumCards
median        4.000000000     NA  45.00000000     NA 1.000000000 2.007000e+03  17.00000000  7.00000000         NA 0.000000000
mean          3.379409636     NA  46.19627212     NA 1.275487735 2.007208e+03  20.07902314  9.31388294         NA 0.883775003
SE.mean       0.002676551     NA   0.04806137     NA 0.001515929 8.260194e-03   0.03990116  0.02430948         NA 0.003169384
CI.mean.0.95  0.005245993     NA   0.09419943     NA 0.002971194 1.618983e-02   0.07820556  0.04764615         NA 0.006211936
var           0.930493846     NA 300.02306221     NA 0.298483257 8.862227e+00 206.79183052 76.75625597         NA 1.304703911
std.dev       0.964621089     NA  17.32117381     NA 0.546336213 2.976949e+00  14.38025836  8.76106477         NA 1.142236364
coef.var      0.285440711     NA   0.37494744     NA 0.428335136 1.483129e-03   0.71618317  0.94064579         NA 1.292451541
               ShopAmount      EatDrink Class    MonthDay   FlightDate AirlineCode Airline OrigCity OrigState DestCity DestState
median          0.0000000    60.0000000    NA 16.00000000 1.392595e+09          NA      NA       NA        NA       NA        NA
mean           26.5540166    68.2421893    NA 15.72308794 1.392471e+09          NA      NA       NA        NA       NA        NA
SE.mean         0.1472864     0.1448698    NA  0.02402580 6.263694e+03          NA      NA       NA        NA       NA        NA
CI.mean.0.95    0.2886787     0.2839422    NA  0.04709015 1.227673e+04          NA      NA       NA        NA       NA        NA
var          2817.6526660  2725.9501132    NA 74.97528817 5.095930e+12          NA      NA       NA        NA       NA        NA
std.dev        53.0815662    52.2106322    NA  8.65882718 2.257417e+06          NA      NA       NA        NA       NA        NA
coef.var        1.9990033     0.7650785    NA  0.55070780 1.621159e-03          NA      NA       NA        NA       NA        NA
             SchDeptHour DeptDelayMins ArrDelayMins Cancelled   FlightMins     Distance ArrDelayGT5
median       13.00000000     0.0000000    0.0000000        NA   91.0000000 6.300000e+02          NA
mean         12.98606470    14.6982354   15.0450780        NA  109.1650909 7.938168e+02          NA
SE.mean       0.01282682     0.1055992    0.1065923        NA    0.2019991 1.642983e+00          NA
CI.mean.0.95  0.02514034     0.2069726    0.2089191        NA    0.3959147 3.220217e+00          NA
var          21.36980234  1448.3839499 1475.7548221        NA 5299.8224811 3.506132e+05          NA
std.dev       4.62274835    38.0576398   38.4155544        NA   72.7998797 5.921260e+02          NA
coef.var      0.35597762     2.5892659    2.5533636        NA    0.6668788 7.459227e-01          NA
```

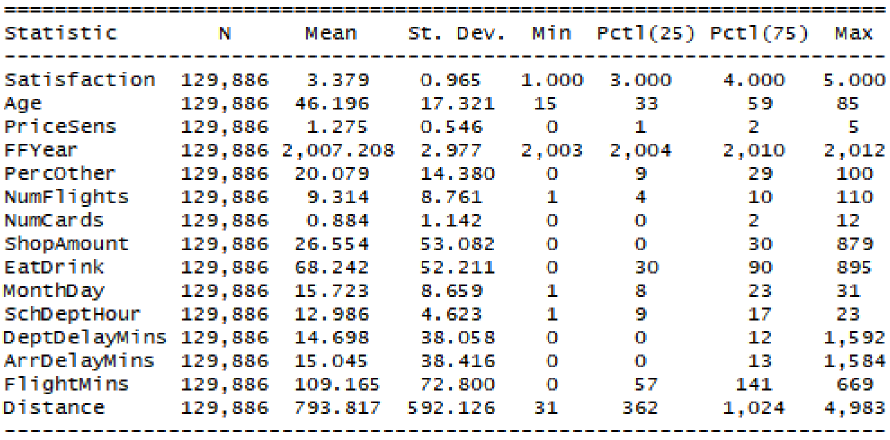
Figure 3: Summary Descriptive Statistics from the Satisfaction Survey



```
              vars      n    mean     sd median trimmed    mad  min  max range  skew kurtosis   se
Satisfaction     1 129886    3.38   0.96      4    3.38   1.48    1    5     4 -0.33    -0.54 0.00
Status*          2 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
Age              3 129886   46.20  17.32     45   45.69  19.27   15   85    70  0.24    -0.71 0.05
Gender*          4 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
PriceSens        5 129886    1.28   0.55      1    1.24   0.00    0    5     5  0.77     1.11 0.00
FFYear           6 129886 2007.21   2.98   2007 2007.15   4.45 2003 2012     9  0.08    -1.29 0.01
PercOther        7 129886   20.08  14.38     17   18.76  13.34    0  100   100  0.84     0.46 0.04
NumFlights       8 129886    9.31   8.76      7    7.61   4.45    1  110   109  2.38     7.88 0.02
TravelType*      9 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
NumCards        10 129886    0.88   1.14      0    0.69   0.00    0   12    12  1.31     1.83 0.00
ShopAmount      11 129886   26.55  53.08      0   13.57   0.00    0  879   879  3.33    16.09 0.15
EatDrink        12 129886   68.24  52.21     60   62.09  44.48    0  895   895  1.99     9.93 0.14
Class*          13 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
MonthDay        14 129886   15.72   8.66     16   15.71  10.38    1   31    30  0.00    -1.17 0.02
FlightDate      15 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
AirlineCode*    16 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
Airline*        17 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
OrigCity*       18 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
OrigState*      19 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
DestCity*       20 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
DestState*      21 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
SchDeptHour     22 129886   12.99   4.62     13   12.91   5.93    1   23    22  0.10    -1.06 0.01
DeptDelayMins   23 129886   14.70  38.06      0    5.79   0.00    0 1592  1592  6.83   100.78 0.11
ArrDelayMins    24 129886   15.05  38.42      0    6.06   0.00    0 1584  1584  6.68    95.35 0.11
Cancelled*      25 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
FlightMins      26 129886  109.17  72.80     91   99.10  57.82    0  669   669  1.41     2.54 0.20
Distance        27 129886  793.82 592.13    630  702.25 468.50   31 4983  4952  1.51     2.75 1.64
ArrDelayGT5*    28 129886     NaN     NA     NA     NaN     NA  Inf -Inf  -Inf    NA       NA   NA
```

Figure 4: Describing the Satisfaction Survey Data

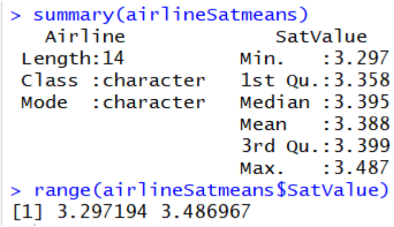
I then looked a little deeper at some of the variables of interest from my initial business questions. I reviewed the general descriptive statistics like mean, standard deviation, maximum and minimum for a few of the variables using a package called "stargazer".



```
=====================================================================
Statistic       N      Mean    St. Dev.  Min   Pctl(25) Pctl(75)  Max
---------------------------------------------------------------------
Satisfaction  129,886   3.379    0.965   1.000  3.000    4.000   5.000
Age           129,886  46.196   17.321    15     33       59      85
PriceSens     129,886   1.275    0.546     0      1        2       5
FFYear        129,886 2,007.208  2.977   2,003  2,004    2,010   2,012
PercOther     129,886  20.079   14.380     0      9       29      100
NumFlights    129,886   9.314    8.761     1      4       10      110
NumCards      129,886   0.884    1.142     0      0        2      12
ShopAmount    129,886  26.554   53.082     0      0       30      879
EatDrink      129,886  68.242   52.211     0     30       90      895
MonthDay      129,886  15.723    8.659     1      8       23      31
SchDeptHour   129,886  12.986    4.623     1      9       17      23
DeptDelayMins 129,886  14.698   38.058     0      0       12     1,592
ArrDelayMins  129,886  15.045   38.416     0      0       13     1,584
FlightMins    129,886  109.165  72.800     0     57       141     669
Distance      129,886  793.817 592.126    31     362     1,024   4,983
---------------------------------------------------------------------
```

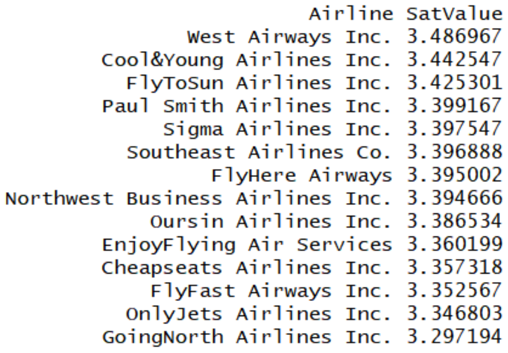
Figure 5: Summary Statistics of Some Satisfaction Survey Variables

To gain some further insight into the most successful airlines based on this satisfaction survey, I did more general analysis of airline information and satisfaction ratings.



```
> summary(airlineSatmeans)
   Airline             SatValue    
 Length:14          Min.   :3.297  
 Class :character   1st Qu.:3.358  
 Mode  :character   Median :3.395  
                    Mean   :3.388  
                    3rd Qu.:3.399  
                    Max.   :3.487  
> range(airlineSatmeans$SatValue)
[1] 3.297194 3.486967
```

Figure 6: Summary Satisfaction Information



```
                         Airline SatValue
              West Airways Inc. 3.486967
       Cool&Young Airlines Inc. 3.442547
         FlyToSun Airlines Inc. 3.425301
       Paul Smith Airlines Inc. 3.399167
            Sigma Airlines Inc. 3.397547
          Southeast Airlines Co. 3.396888
                  FlyHere Airways 3.395002
Northwest Business Airlines Inc. 3.394666
           Oursin Airlines Inc. 3.386534
       EnjoyFlying Air Services 3.360199
       Cheapseats Airlines Inc. 3.357318
            FlyFast Airways Inc. 3.352567
          OnlyJets Airlines Inc. 3.346803
       GoingNorth Airlines Inc. 3.297194
```

Figure 7: Airline Satisfaction Rating by Decreasing Mean

By reviewing this very general information, I can see that Southeast Airlines is in the top 6 out of the 14 airlines that were surveyed, but still has room for improvement. So, then I determine I work more extensively on visualization so I could determine where to focus my efforts for modeling and prediction.

To show this same information, I completed a simple bar graph.

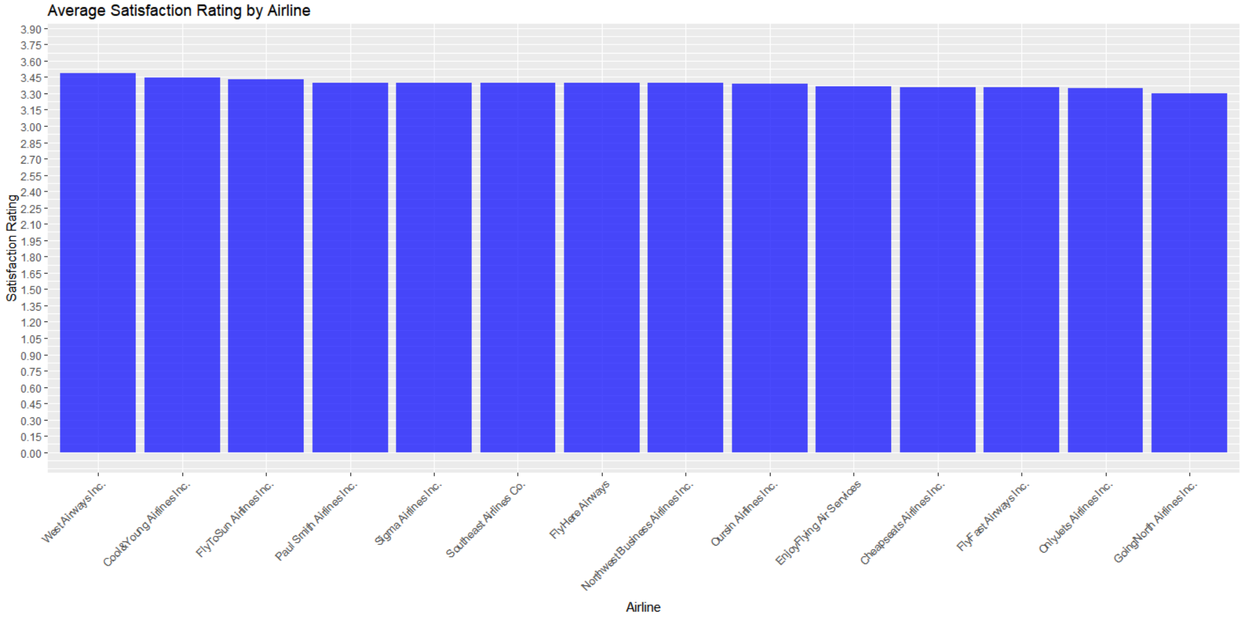


Figure 8: Mean Airline Satisfaction Rating Plotted in Decreasing Order

Modeling and Visualization

I found that looking at some plots and visualization of the data in different ways, I could see more patterns or dependencies than I could gather from just reviewing the raw statistics.

General Visualization

First, I did some very generic plots based on some hunches that I had around satisfaction results. Out of the over 73,000 females surveyed and the over 56,000 males surveyed, I thought the females would likely have higher satisfaction rates in general. However, my assumptions were not reproduced when I graphed this information (see following figure).

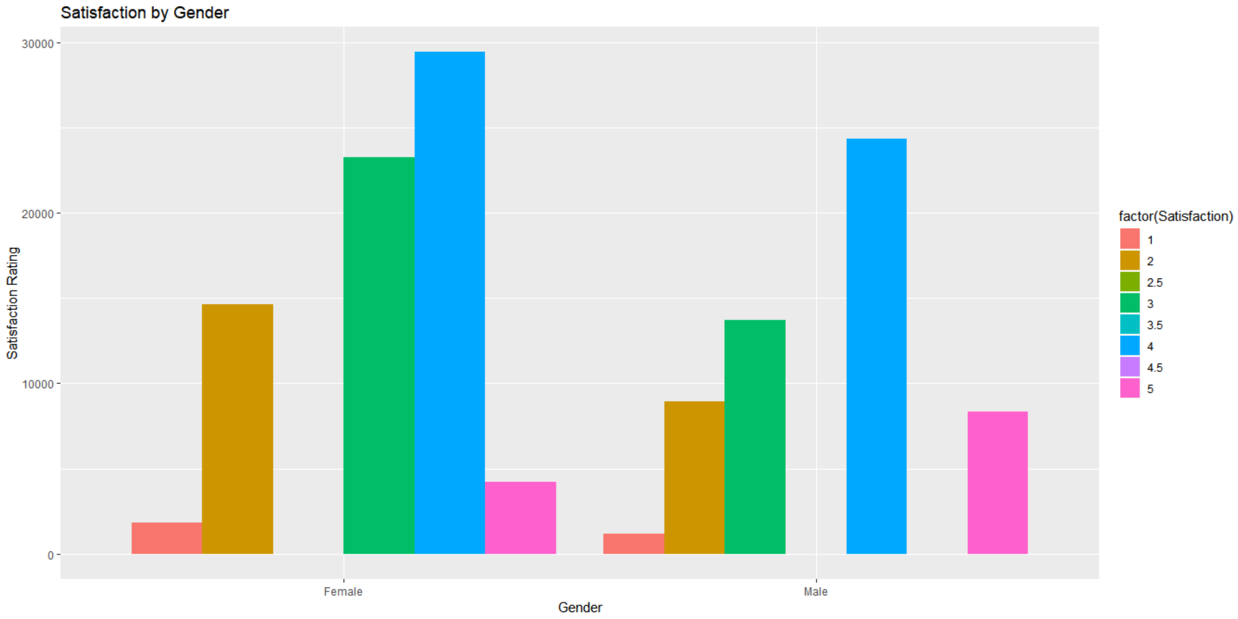


Figure 9: Airline Satisfaction Rating by Gender

I also wanted to see gather a little more data about the ages of those surveyed and elected to do this using a box plot. I used a function I wrote to come up with some general buckets for the ages.

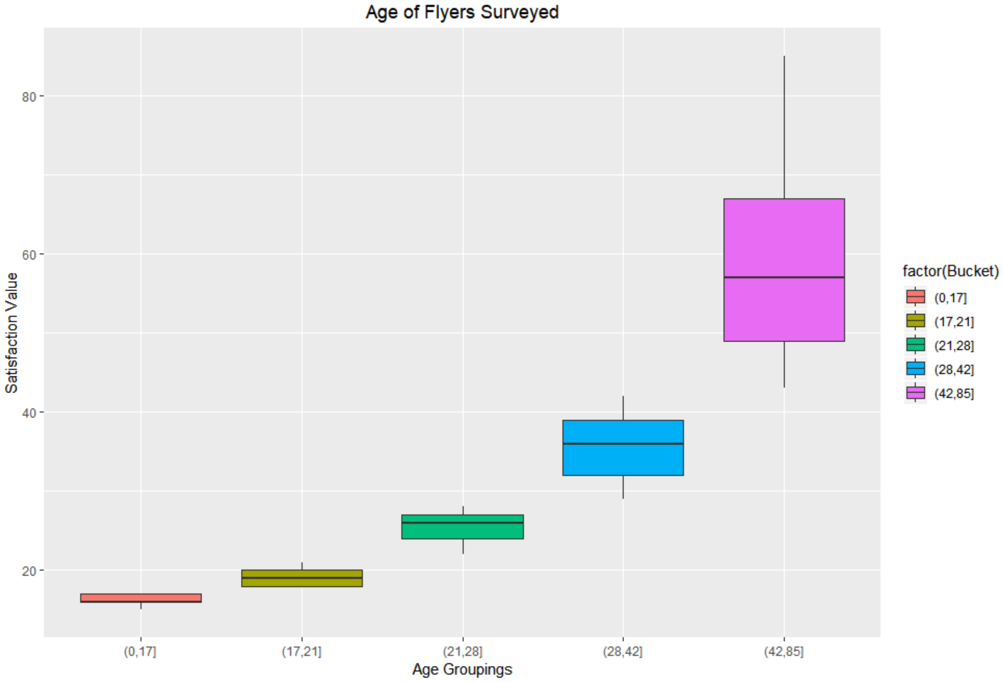


Figure 10: Ages of Survey Population

To further review the means by airlines and how the different airlines match up against each other, I plotted the satisfaction by airline as well.

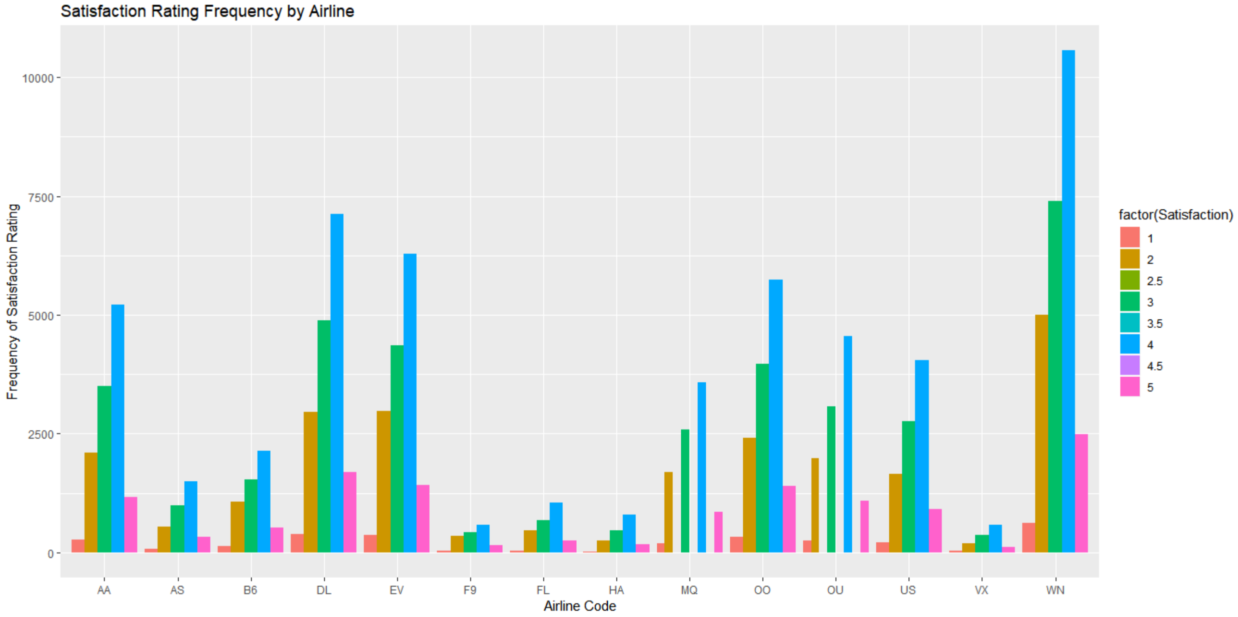


Figure 11: Frequency of Satisfaction Rating by Airline Code

You can see in the plot shown above that the WN (Cheapseats Airlines Inc.) has high satisfaction ratings; however, if we look at the overall mean satisfactions – it does not do well overall. This is probably due to the high 2 ratings as shown in the plot.

This is another representation of the same data showing the airlines and ratings in a different way.

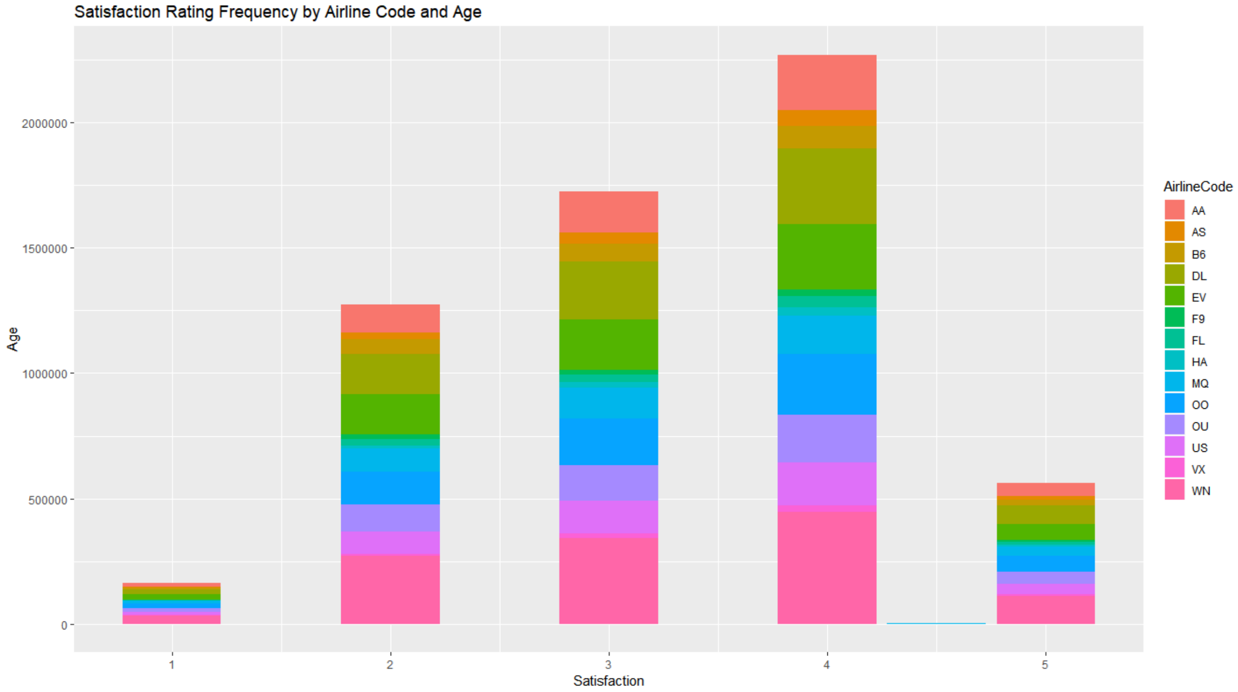


Figure 12: Satisfaction Rating by Airline Code Colored Bar Chart

As I looked in more detail at the visualizations and information, I started to fine tune the visualizations to try to gather more specific information around the Satisfaction Rating and only a few independent variables that have the most influence on the satisfaction rating.

I thought about myself and flying, as I have been an active business and personal flyer for over 30 years. On the airlines where I hold a Sliver, Gold or higher loyalty card, I tend to have a better experience. This is usually because I can board early, upgrade my seat for a lesser price, etc. So, I started to compare that to what I don't like about flying which is delays. So, I decided to plot that information by airline to see what I could determine based on satisfaction versus loyalty card and number of delayed minutes. The results of this plot are in the next figure.

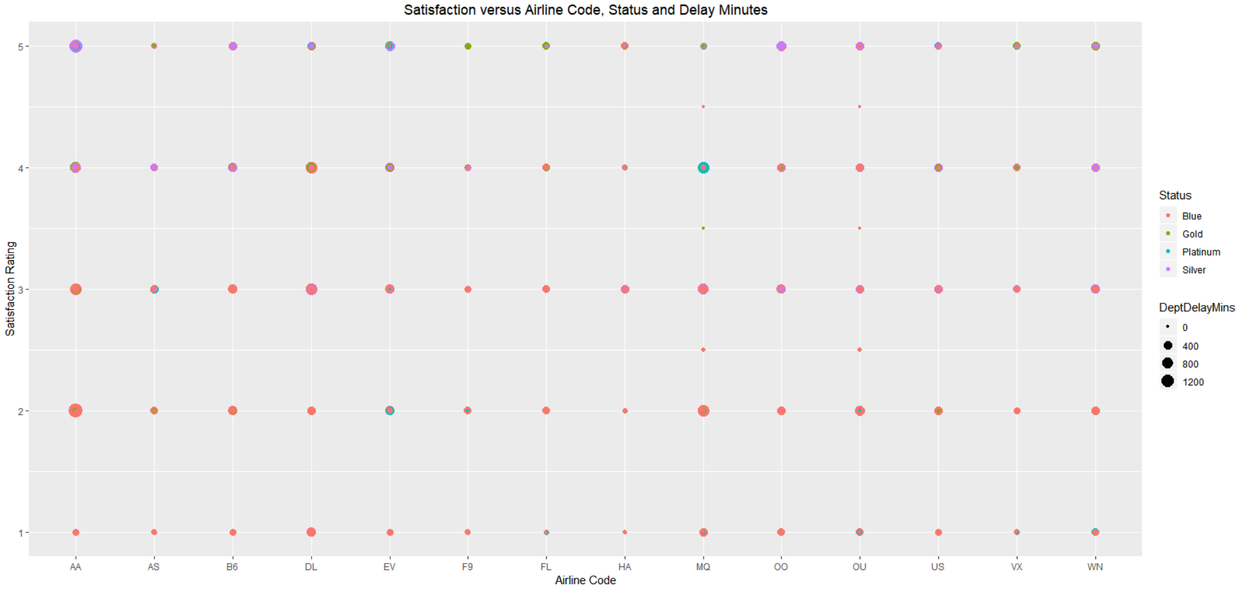


Figure 13: Satisfaction by Airline Code and Delay in Minutes

As shown in the figure above, the "2" satisfaction rate does have longer delays as well as more individuals with "Blue" (the lowest reward program status) loyalty cards. The majority of customers rating with the lowest satisfaction rate of "1" are "Blue" card members as well. As you can see by the color of the dots in the scatter plot above, the majority of flyers with Silver, Gold and Platinum cards rated with scores of 4 and 5. This is what I expected to see because of the perks associated with these higher levels of rewards programs.

I also wanted to know if there were issues with Satisfaction depending on the departure or arrival location, so I completed maps to show this information. In this case, I gathered information about the Origination and Destination Satisfaction Means by State as well as by City. For ease of viewing, I plotted this information and the Mean Satisfaction by Origination and Destination State is shown in the following figure.

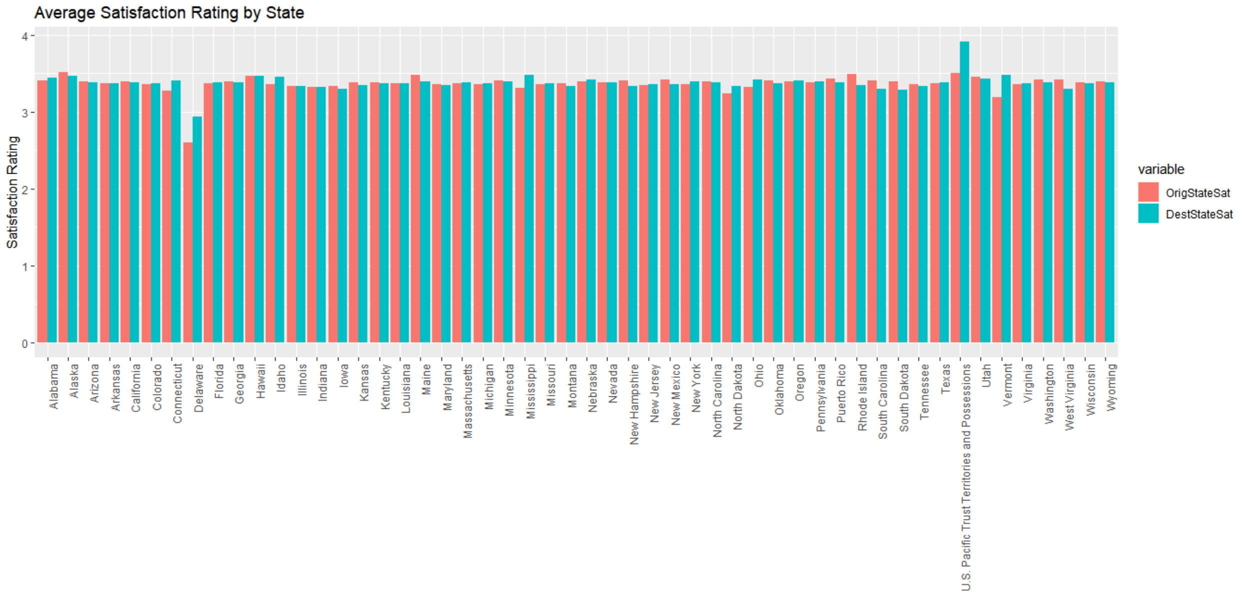


Figure 14: Average Airline Satisfaction Rating Plotted by Origination and Destination State

It becomes instantly obvious that flying in and out of Delaware is an issue and flying in and out of the U.S. Pacific Trust Territories and Possessions has the most satisfied flyers. This area is not within the United States. There does not seem to be another status that are specifically higher than others, but that might change when we look at individual airport cities.

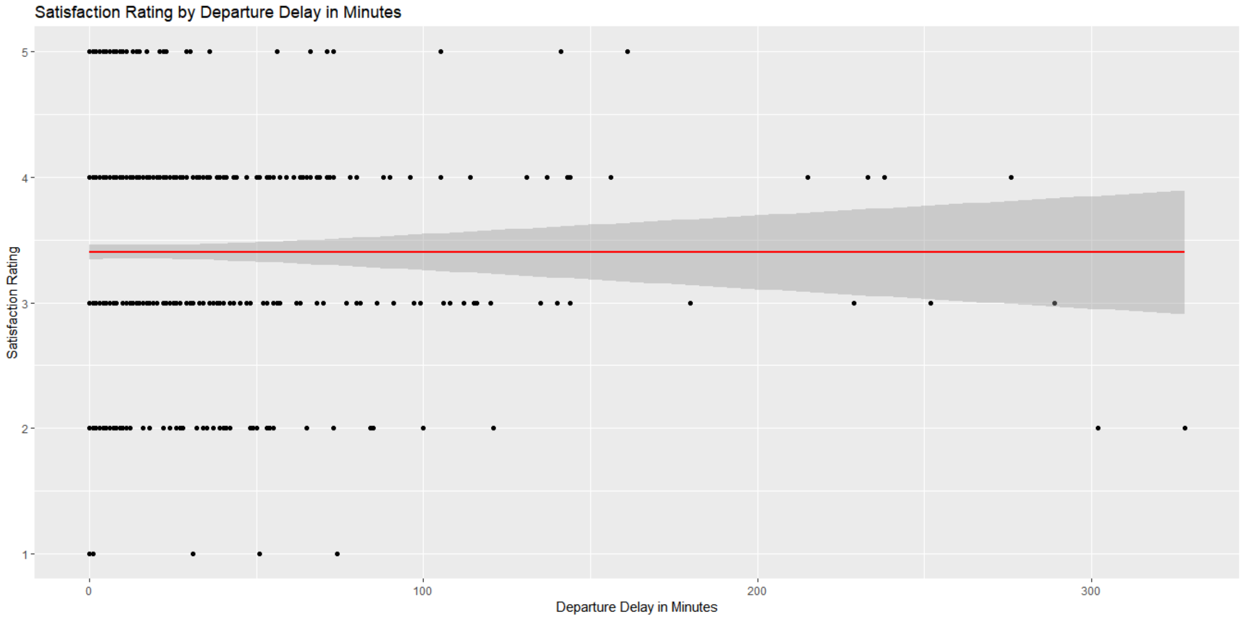


Figure 15: Satisfaction Rating by Departure Delay Minutes

Reviewing the plot above, we see that the average satisfaction rating is approximately 3.5 and as that that rating is lower with departure delays over 300 minutes.

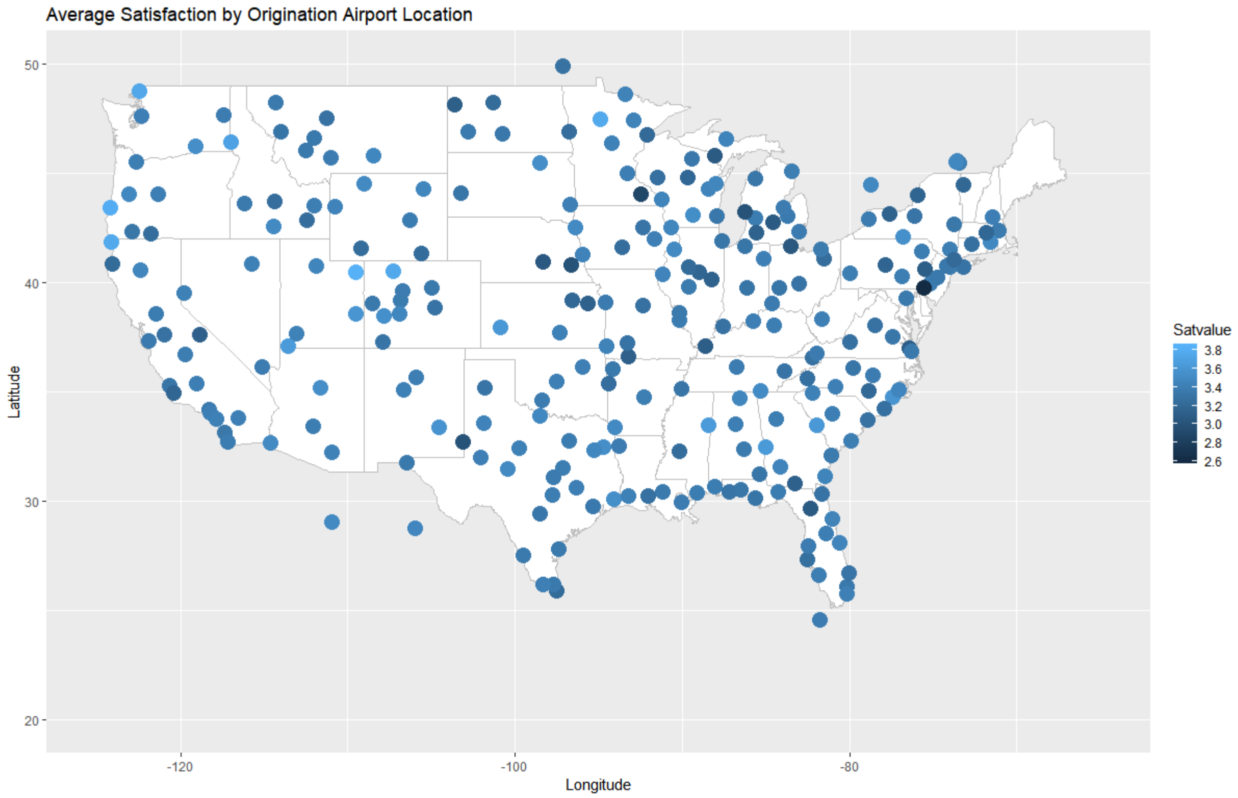


Figure 16: Airline Satisfaction Rating Origination Airport Location

The figure above indicates that there are some airport locations in Colorado, Washington, and California that seems to be higher mean satisfaction ratings than Delaware (darkest dot) on the plot above. This may have to do with the size, accessibility, etc. But clearly, satisfaction rates differ by departure airport location.

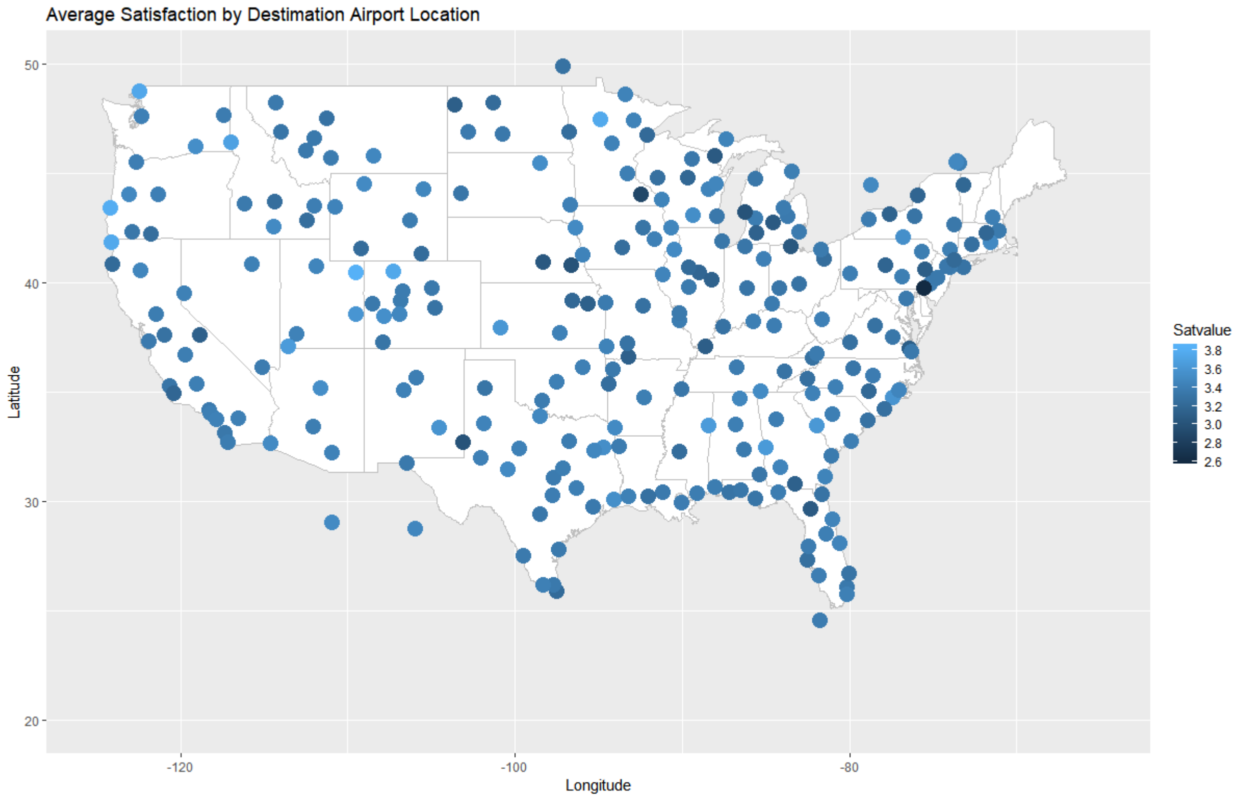


Figure 17: Airline Satisfaction Rating Destination Airport Location

Again, the figure above indicates that there are some airport locations in Colorado, Washington, and California that seems to be higher mean satisfaction ratings than Delaware (darkest dot) on the plot above.

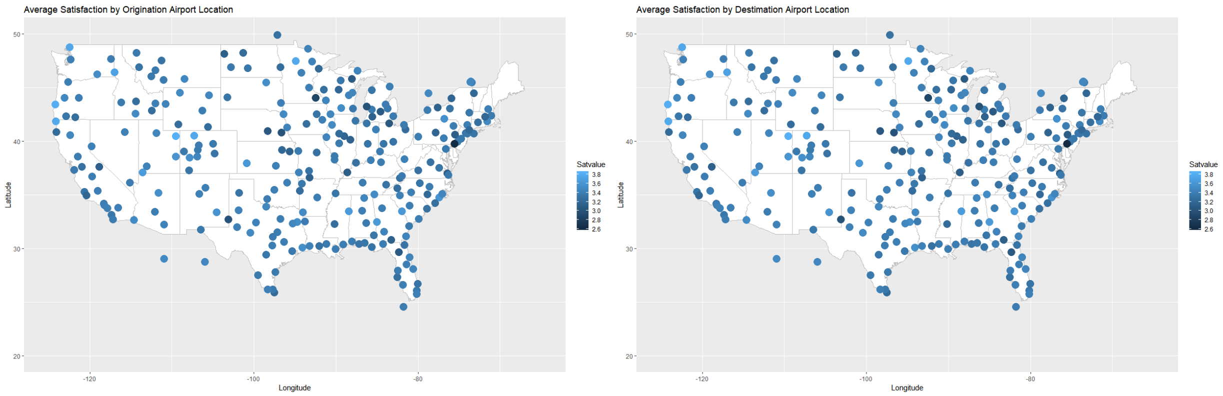

Average Satisfaction by Origination Airport Location
Average Satisfaction by Destination Airport Location
Latitude
50
40
30
20
-120
-100
-80
Longitude
Satvalue
3.8
3.6
3.4
3.2
3.0
2.8
2.6

Figure 18: Airline Satisfaction Rating Origination and Destination Airport Location Side by Side

As shown (hard to see in the small plots), the airport locations seem to be fairly consistent with satisfaction rates whether they are a departure airport or a destination airport. I do not feel the location has much influence on the experience with the exception a few airports (the ones mentioned with high and low satisfaction).

There were some additional variables that I felt might be influencing the Satisfaction and I did some plots of those variables alone as well. These are reflected in the plots in the following figures.

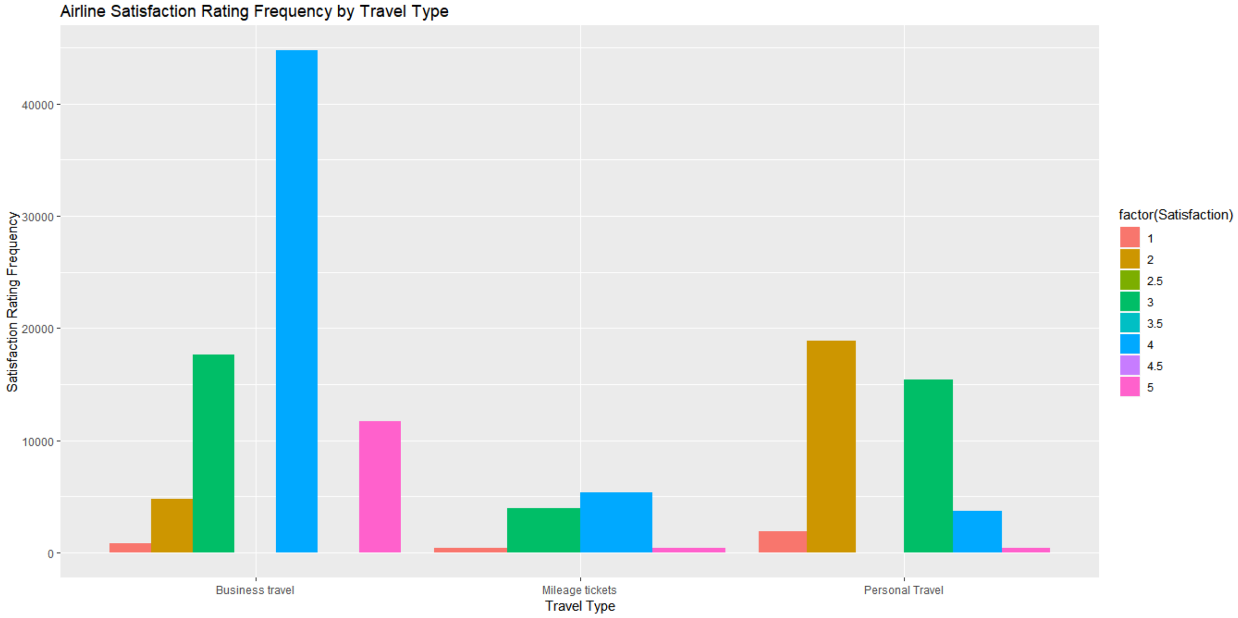


Figure 19: Airline Satisfaction Rating by Travel Type

As we see here, business travelers tend to be more satisfied overall than those traveling for personal reasons. Because less travelers were using mileage tickets, it is hard to tell, but they seem to be average in their satisfaction overall.

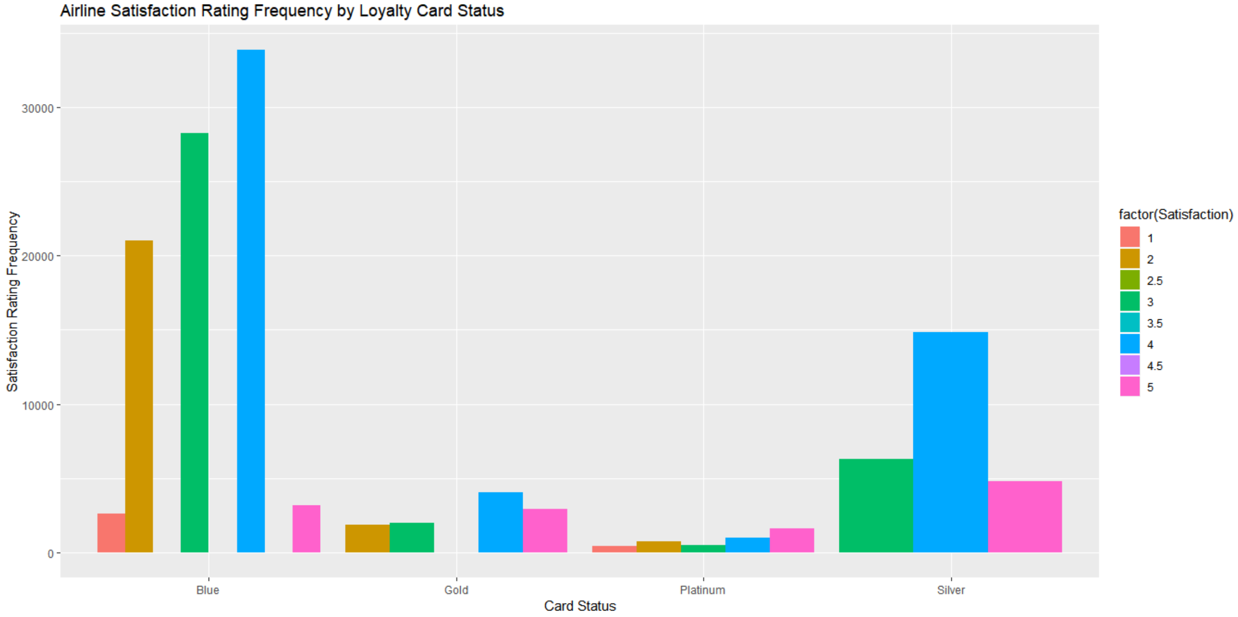


Figure 20: Airline Satisfaction Rating by Airline Card Status

Obviously, there are more flyers with *Blue* status than those of other status, but it does look like *Silver*, *Gold*, and *Platinum* customers are more satisfied overall (scores of 4 and 5 more popular) which is what we saw in our previous scatterplot showing satisfaction, delays and card status.

I also wanted to see how other travel affected the flyers. This included when they started flying, the number of flights they had take and the percentage on other airlines.

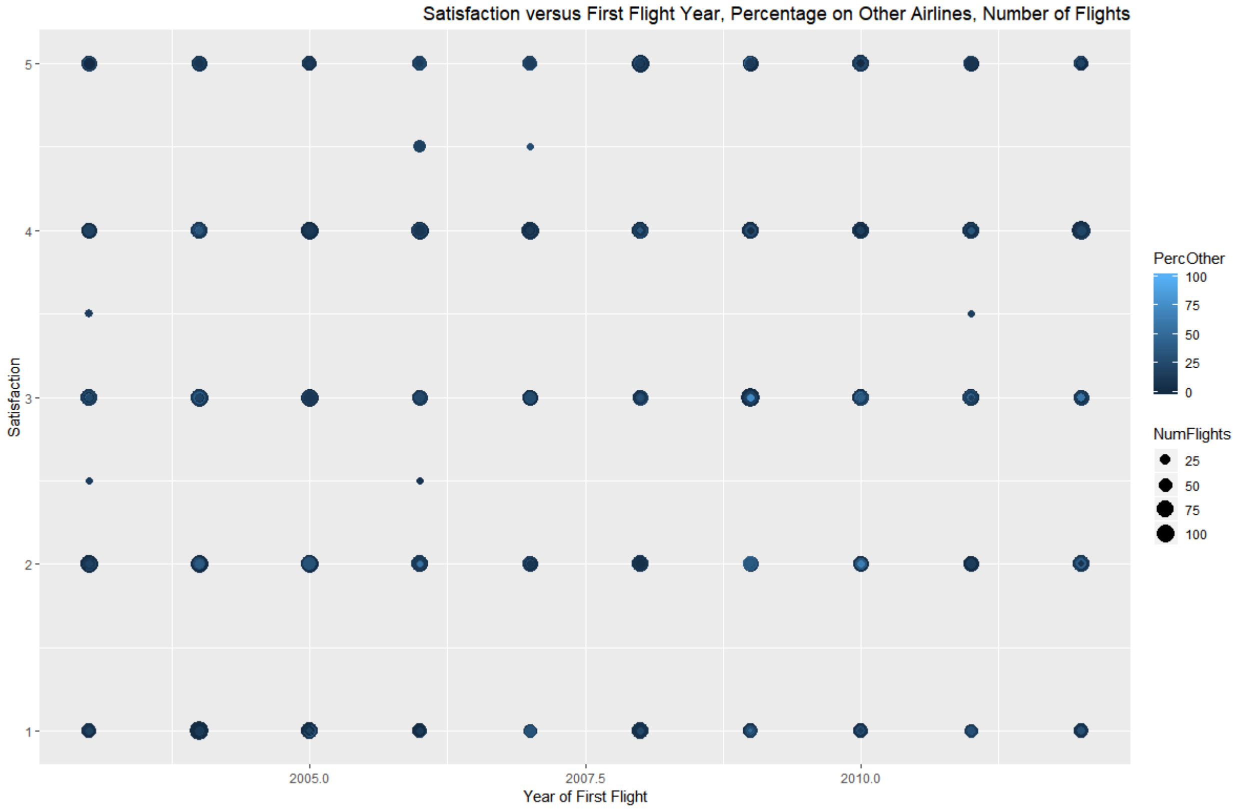


Figure 21: Satisfaction by % of Other Flights and Number of Flights by First Flight Year

It seems the more you fly the more satisfied you are with airlines. Maybe this is because you anticipate issues and know what to expect as opposed to expecting so much when you have not had many experiences to compare it against. This was also reflected in the bar graph with similar data shown below.

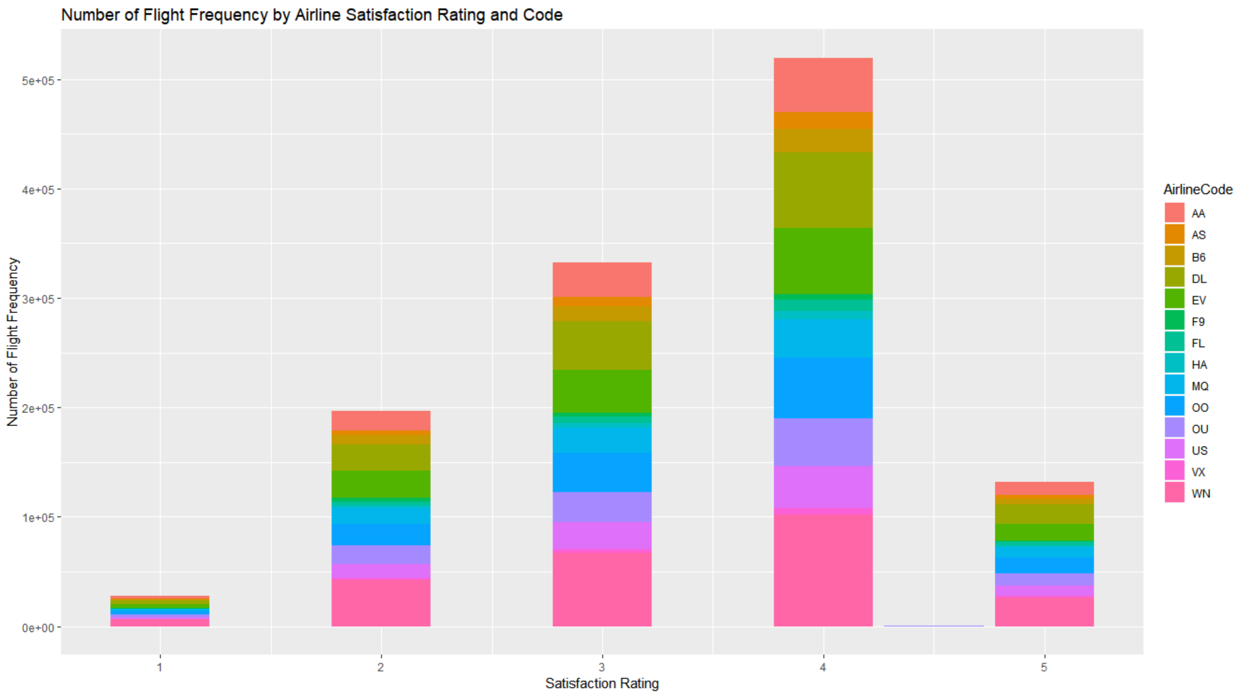


Figure 22: Airline Satisfaction as affected by Number of Flights by Airline

Modeling

Now that this initial investigation has been done, I can look into preparing models to determine what is really influencing the satisfaction for certain airlines. For modeling, I selected to use the following modeling techniques.

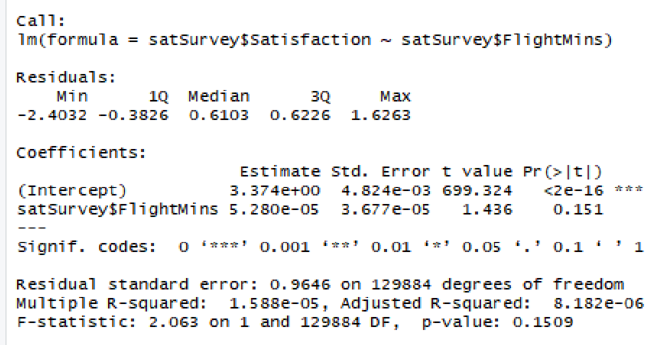
- Linear Regression (LM and utilized AIC to find the best model for the data)
- Support Vector Machines (SVM)
- Kernel Support Vector Machines (KSVM)
- Naïve Bayes (NB)

For building the SVM and NB models and testing the models, a subset of data was taken for the training dataset (15,000 surveys) and a subset was taken for the testing (predicting) data set (5,000 surveys). All plots were done with a testing data set of 5,000 surveys.

Linear Regression (LM)

The first model helped me to determine which, if any variables are influencing the Satisfaction Rating as well as determine how much influence they have. This is determined by looking at the Adjusted R Squared Value (to show the relevance/strength of the model) and the p-values which should be < 0.05 to show that the individual independent variable has significant influence on the Satisfaction Rating (the dependent variable).

Initially, I looked as some simple modeling and information to rule out certain variables. For example, I completed a quick linear regression model with just how long a flight took and that correlation to flight satisfaction as shown in the following figure.

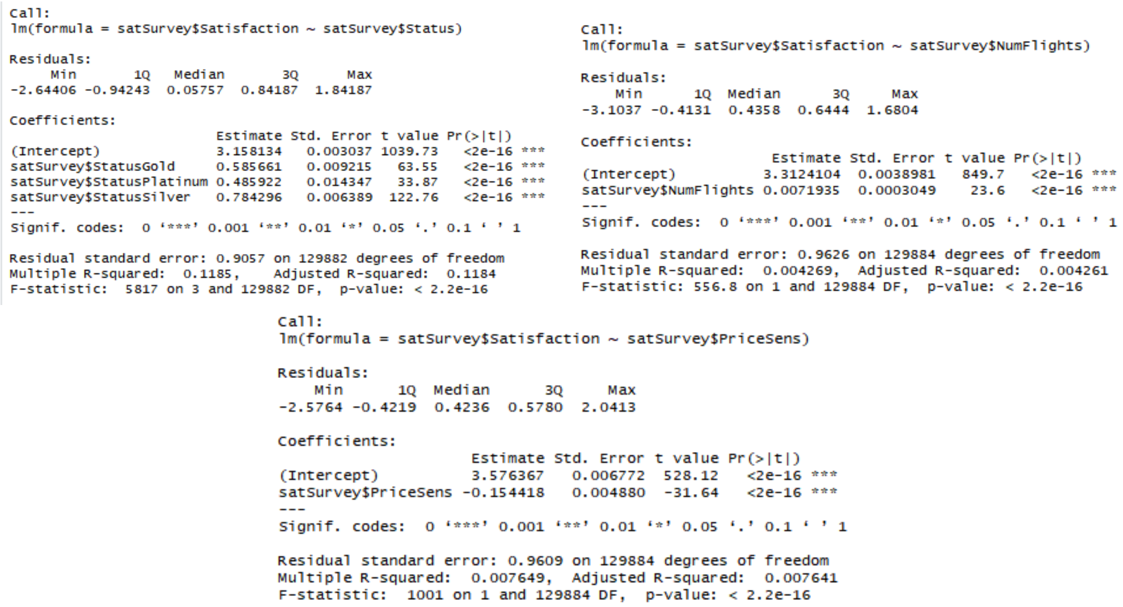


```
Call:
lm(formula = satSurvey$Satisfaction ~ satSurvey$FlightMins)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4032 -0.3826  0.6103  0.6226  1.6263 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.374e+00  4.824e-03 699.324   <2e-16 ***
satSurvey$FlightMins 5.280e-05  3.677e-05   1.436    0.151    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9646 on 129884 degrees of freedom
Multiple R-squared:  1.588e-05,	Adjusted R-squared:  8.182e-06 
F-statistic: 2.063 on 1 and 129884 DF,  p-value: 0.1509
```

Figure 23: Satisfaction based on Total Flight Minutes Linear Regression Model

As shown in the details from this linear regression model, the p-value value was quite a bit higher than 0.05 (0.151) and the Adjusted R-squared value is very low indicating no correlation between these two pieces of data.

I completed some more simple models and found that although the Adjusted R-squared value was low, the p-value was low for the Status (Silver, Gold, Platinum) with the airline. The same was true of the number of flights taken be the individual.



```
Call:
lm(formula = satSurvey$Satisfaction ~ satSurvey$Status)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64406 -0.94243  0.05757  0.84187  1.84187 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               3.158134   0.003037 1039.73   <2e-16 ***
satSurvey$StatusGold      0.585661   0.009215   63.55   <2e-16 ***
satSurvey$StatusPlatinum 0.485922   0.014347   33.87   <2e-16 ***
satSurvey$StatusSilver   0.784296   0.006389  122.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9057 on 129882 degrees of freedom
Multiple R-squared:  0.1185,	Adjusted R-squared:  0.1184 
F-statistic:  5817 on 3 and 129882 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = satSurvey$Satisfaction ~ satSurvey$NumFlights)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1037 -0.4131  0.4358  0.6444  1.6804 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.3124104  0.0038981   849.7   <2e-16 ***
satSurvey$NumFlights 0.0071935  0.0003049    23.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9626 on 129884 degrees of freedom
Multiple R-squared:  0.004269,	Adjusted R-squared:  0.004261 
F-statistic: 556.8 on 1 and 129884 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = satSurvey$Satisfaction ~ satSurvey$PriceSens)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5764 -0.4219  0.4236  0.5780  2.0413 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.576367   0.006772  528.12   <2e-16 ***
satSurvey$PriceSens -0.154418   0.004880  -31.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9609 on 129884 degrees of freedom
Multiple R-squared:  0.007649,	Adjusted R-squared:  0.007641 
F-statistic:  1001 on 1 and 129884 DF,  p-value: < 2.2e-16
```

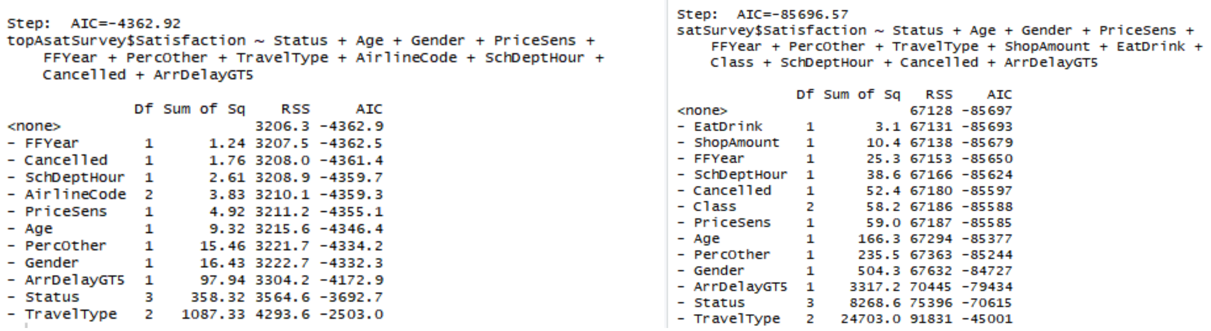
Figure 24: Satisfaction Model based on Status, Total Number of Flights, Price Sensitivity

To gather some additional data on the satisfaction, I decided to narrow down the survey and create a new data frame of just the top 3 airlines based on their mean satisfaction value:

- West Airways Inc (HA)
- Cool&Young Airlines Inc. (VX)
- FlytoSun Airlines Inc. (AS)

This can then be used to look at how we might be able to improve the satisfaction of Southeast Airlines Inc. (US). I obtained the Airline Codes for these by filtering the satSurvey data with RStudio.

After creating this new data, I created a linear regression on these airlines and found that the variables listed in the following figure were the ones that influenced satisfaction for these three airlines. I ran the AIC technique against all the ***all*** the airlines and got the following for the best model which appears to the right of the best airline model.



```
Step:  AIC=-4362.92
topAsatSurvey$Satisfaction ~ Status + Age + Gender + PriceSens + 
    FFYear + PercOther + TravelType + AirlineCode + SchDeptHour + 
    Cancelled + ArrDelayGT5

              Df Sum of Sq    RSS     AIC
<none>                     3206.3 -4362.9
- FFYear       1      1.24 3207.5 -4362.5
- Cancelled    1      1.76 3208.0 -4361.4
- SchDeptHour  1      2.61 3208.9 -4359.7
- AirlineCode  2      3.83 3210.1 -4359.3
- PriceSens    1      4.92 3211.2 -4355.1
- Age          1      9.32 3215.6 -4346.4
- PercOther    1     15.46 3221.7 -4334.2
- Gender       1     16.43 3222.7 -4332.3
- ArrDelayGT5  1     97.94 3304.2 -4172.9
- Status       3    358.32 3564.6 -3692.7
- TravelType   2   1087.33 4293.6 -2503.0
```

```
Step:  AIC=-85696.57
satSurvey$Satisfaction ~ Status + Age + Gender + PriceSens + 
    FFYear + PercOther + TravelType + ShopAmount + EatDrink + 
    Class + SchDeptHour + Cancelled + ArrDelayGT5

              Df Sum of Sq   RSS    AIC
<none>                     67128 -85697
- EatDrink     1       3.1 67131 -85693
- ShopAmount   1      10.4 67138 -85679
- FFYear       1      25.3 67153 -85650
- SchDeptHour  1      38.6 67166 -85624
- Cancelled    1      52.4 67180 -85597
- Class        2      58.2 67186 -85588
- PriceSens    1      59.0 67187 -85585
- Age          1     166.3 67294 -85377
- PercOther    1     235.5 67363 -85244
- Gender       1     504.3 67632 -84727
- ArrDelayGT5  1    3317.2 70445 -79434
- Status       3    8268.6 75396 -70615
- TravelType   2   24703.0 91831 -45001
```

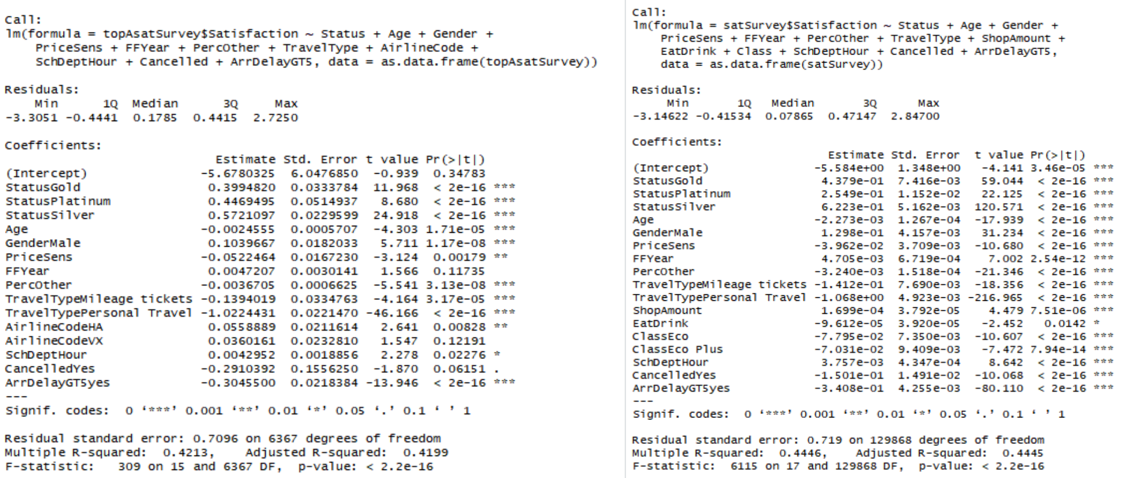
Figure 25: AIC Best Linear Regression Model for Top Three Airlines and All Airlines

You will see that these models have quite a few variables in common:

- Scheduled Departure Hour
- Flight Cancelled or Not
- Price Sensitivity
- Gender
- Arrival Delay Greater than 5 minutes
- Percent of travel on other airlines
- Type of Travel – business, personal, mileage award
- Age of traveler
- Year of First Flight

The variables that appear in only one model are as follows:

- Eating or Drinking in the airport
- Shopping Amount in the airport
- Class of Travel
- Airline Code – only in the top three airlines model

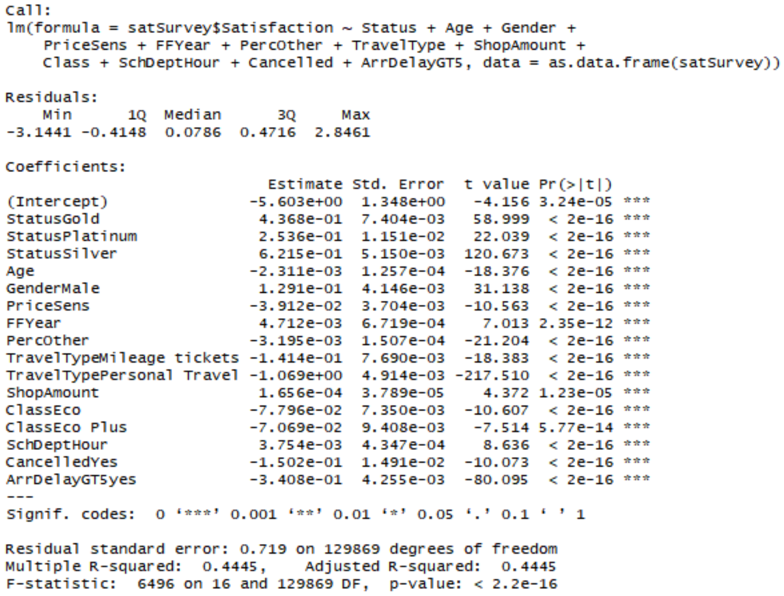


```
Call:
lm(formula = topAsatSurvey$Satisfaction ~ Status + Age + Gender +
    PriceSens + FFYear + PercOther + TravelType + AirlineCode +
    SchDeptHour + Cancelled + ArrDelayGT5, data = as.data.frame(topAsatSurvey))

Residuals:
    Min      1Q  Median      3Q     Max
-3.3051 -0.4441  0.1785  0.4415  2.7250

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -5.6780325  6.0476850  -0.939  0.34783
StatusGold                0.3994820  0.0333784  11.968  < 2e-16 ***
StatusPlatinum            0.4469495  0.0514937   8.680  < 2e-16 ***
StatusSilver              0.5721097  0.0229599  24.918  < 2e-16 ***
Age                      -0.0024555  0.0005707  -4.303 1.71e-05 ***
GenderMale                0.1039667  0.0182033   5.711 1.17e-08 ***
PriceSens                -0.0522464  0.0167230  -3.124  0.00179 **
FFYear                    0.0047207  0.0030141   1.566  0.11735
PercOther                -0.0036705  0.0006625  -5.541 3.13e-08 ***
TravelTypeMileage tickets -0.1394019  0.0334763  -4.164 3.17e-05 ***
TravelTypePersonal Travel -1.0224431  0.0221470 -46.166  < 2e-16 ***
AirlineCodeHA             0.0558889  0.0211614   2.641  0.00828 **
AirlineCodeVX             0.0360161  0.0232810   1.547  0.12191
SchDeptHour               0.0042952  0.0018856   2.278  0.02276 *
CancelledYes             -0.2910392  0.1556250  -1.870  0.06151 .
ArrDelayGT5yes           -0.3045500  0.0218384 -13.946  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7096 on 6367 degrees of freedom
Multiple R-squared:  0.4213,    Adjusted R-squared:  0.4199
F-statistic:   309 on 15 and 6367 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = satSurvey$Satisfaction ~ Status + Age + Gender +
    PriceSens + FFYear + PercOther + TravelType + ShopAmount +
    EatDrink + Class + SchDeptHour + Cancelled + ArrDelayGT5,
    data = as.data.frame(satSurvey))

Residuals:
     Min       1Q   Median       3Q      Max
-3.14622 -0.41534  0.07865  0.47147  2.84700

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)               -5.584e+00  1.348e+00   -4.141 3.46e-05 ***
StatusGold                 4.379e-01  7.416e-03   59.044  < 2e-16 ***
StatusPlatinum             2.549e-01  1.152e-02   22.125  < 2e-16 ***
StatusSilver               6.223e-01  5.162e-03  120.571  < 2e-16 ***
Age                       -2.273e-03  1.267e-04  -17.939  < 2e-16 ***
GenderMale                 1.298e-01  4.157e-03   31.234  < 2e-16 ***
PriceSens                 -3.962e-02  3.709e-03  -10.680  < 2e-16 ***
FFYear                     4.705e-03  6.719e-04    7.002 2.54e-12 ***
PercOther                 -3.240e-03  1.518e-04  -21.346  < 2e-16 ***
TravelTypeMileage tickets -1.412e-01  7.690e-03  -18.356  < 2e-16 ***
TravelTypePersonal Travel -1.068e+00  4.923e-03 -216.965  < 2e-16 ***
ShopAmount                 1.699e-04  3.792e-05    4.479 7.51e-06 ***
EatDrink                  -9.612e-05  3.920e-05   -2.452   0.0142 *
ClassEco                  -7.795e-02  7.350e-03  -10.607  < 2e-16 ***
ClassEco Plus             -7.031e-02  9.409e-03   -7.472 7.94e-14 ***
SchDeptHour                3.757e-03  4.347e-04    8.642  < 2e-16 ***
CancelledYes              -1.501e-01  1.491e-02  -10.068  < 2e-16 ***
ArrDelayGT5yes            -3.408e-01  4.255e-03  -80.110  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.719 on 129868 degrees of freedom
Multiple R-squared:  0.4446,    Adjusted R-squared:  0.4445
F-statistic:  6115 on 17 and 129868 DF,  p-value: < 2.2e-16
```

Figure 26: Summary of the Best Linear Regression Model for Top Three and All Airlines

The figure above provides a summary of the variables for each of the models (top 3 airlines and all airlines). As you can see in the model on the right, eating and drinking in the airport had the largest p-value, so I elected not to use that variable in the final linear regression model. In addition, the airline code had a larger p-value in the top three airline model and I elected not to allow this to interfere in the model. The final model chosen had a union of the two models shown above without EatDrink and AirlineCode.



```
Call:
lm(formula = satSurvey$Satisfaction ~ Status + Age + Gender +
    PriceSens + FFYear + PercOther + TravelType + ShopAmount +
    Class + SchDeptHour + Cancelled + ArrDelayGT5, data = as.data.frame(satSurvey))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1441 -0.4148  0.0786  0.4716  2.8461

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)               -5.603e+00  1.348e+00   -4.156 3.24e-05 ***
StatusGold                 4.368e-01  7.404e-03   58.999  < 2e-16 ***
StatusPlatinum             2.536e-01  1.151e-02   22.039  < 2e-16 ***
StatusSilver               6.215e-01  5.150e-03  120.673  < 2e-16 ***
Age                       -2.311e-03  1.257e-04  -18.376  < 2e-16 ***
GenderMale                 1.291e-01  4.146e-03   31.138  < 2e-16 ***
PriceSens                 -3.912e-02  3.704e-03  -10.563  < 2e-16 ***
FFYear                     4.712e-03  6.719e-04    7.013 2.35e-12 ***
PercOther                 -3.195e-03  1.507e-04  -21.204  < 2e-16 ***
TravelTypeMileage tickets -1.414e-01  7.690e-03  -18.383  < 2e-16 ***
TravelTypePersonal Travel -1.069e+00  4.914e-03 -217.510  < 2e-16 ***
ShopAmount                 1.656e-04  3.789e-05    4.372 1.23e-05 ***
ClassEco                  -7.796e-02  7.350e-03  -10.607  < 2e-16 ***
ClassEco Plus             -7.069e-02  9.408e-03   -7.514 5.77e-14 ***
SchDeptHour                3.754e-03  4.347e-04    8.636  < 2e-16 ***
CancelledYes              -1.502e-01  1.491e-02  -10.073  < 2e-16 ***
ArrDelayGT5yes            -3.408e-01  4.255e-03  -80.095  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.719 on 129869 degrees of freedom
Multiple R-squared:  0.4445,    Adjusted R-squared:  0.4445
F-statistic:  6496 on 16 and 129869 DF,  p-value: < 2.2e-16
```

Figure 27: Summary of the Final and Best Linear Regression Model

As I reviewed the coefficients associated with the variables of influence, you can see how much these individual variables increase or decrease satisfaction. For example, we can see that Price Sensitivity (PriceSens) have a negative coefficient, so it brings the overall satisfaction down if you are price sensitive. In addition, if the traveler has flown often on other airlines, is traveling Economy or Economy Plus status, the flight is cancelled, or the arrival delay greater than 5 minutes – satisfaction is likely to be decreased. However, predicted satisfaction is likely to increase if the traveler is male traveling that holds

Gold, Platinum or Silver status. In addition, if he/she has been shopping before his/her flight, satisfaction is higher.

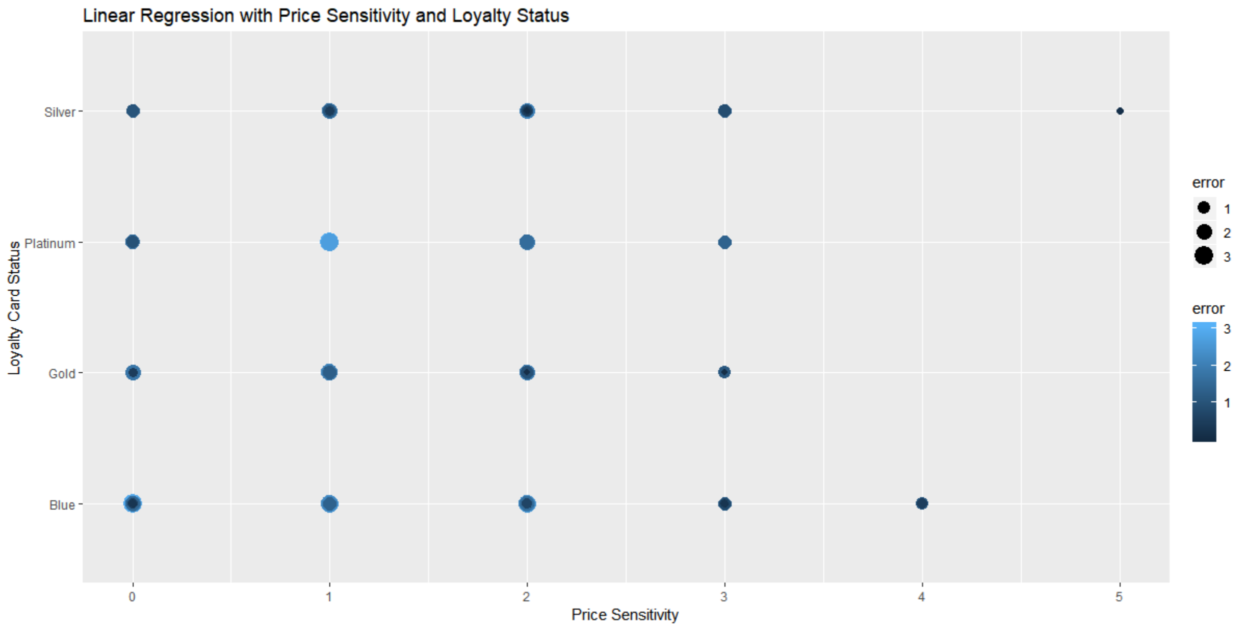


Figure 28: Best LM showing Price Sensitivity and Status with Prediction Error

As shown in the plot above, our linear regression model does a fairly good job predicting when the Status is Silver, Gold or Platinum with a higher level of Price Sensitivity.

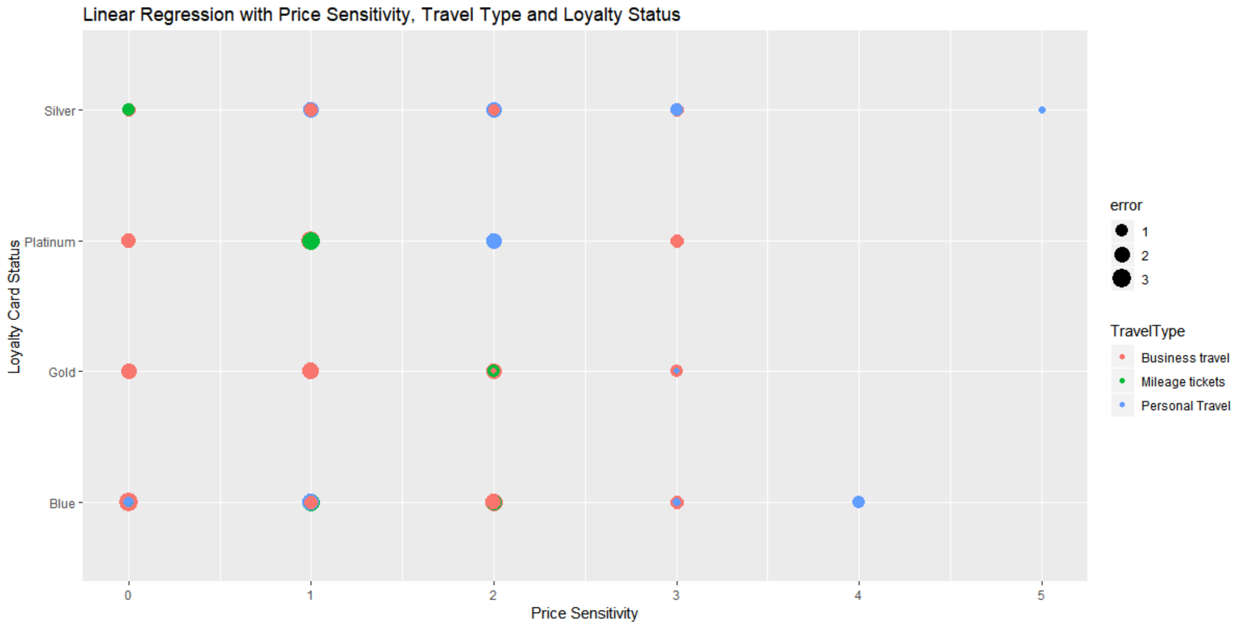


Figure 29: Best LM showing Price Sensitivity, Travel Type and Status with Prediction Error

As reflected in the figure above, again the model shows low errors between our predictions with the model and the actual values when we are looking at the Silver and Gold members with high price sensitivity.

With a linear regression model and our data, the plots do now show much of a pattern. This is due to the finite number of values like Price Sensitivity of 1 through 5 only, or Status of 4 different values, etc. There is very little continuous data except delays that can be found in our data, but that led me to the next plot.

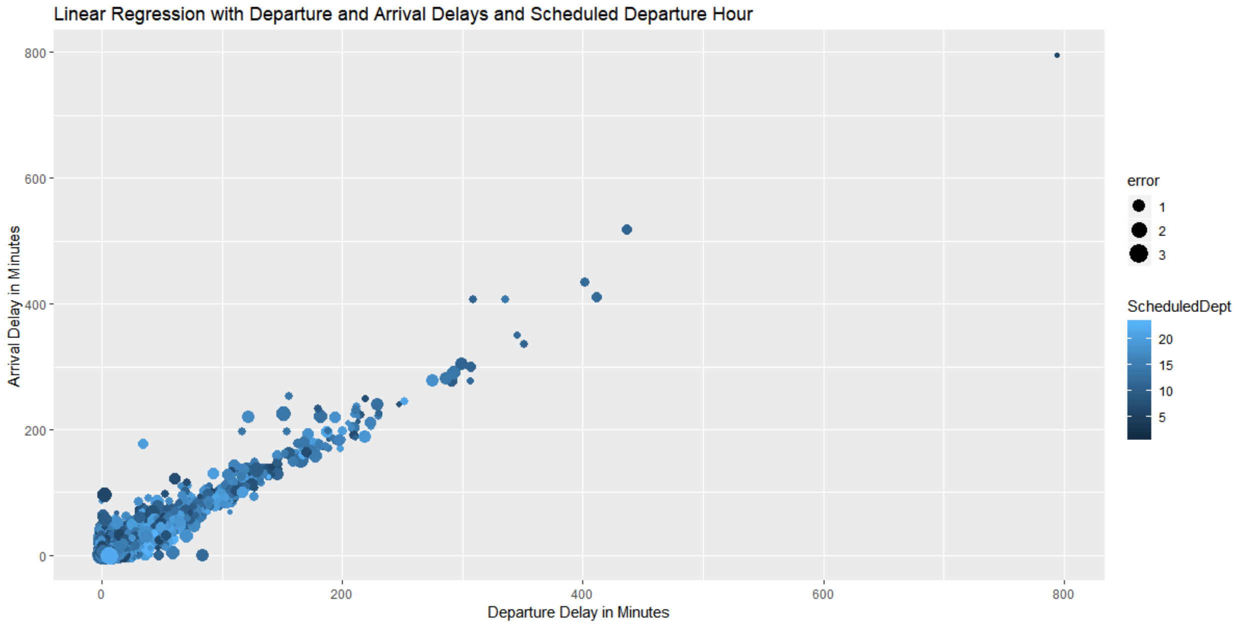


Figure 30: Best LM showing Departure and Arrival Delays with Prediction Error

As shown above, we can predict poor satisfaction when the delays are extensive. In fact, the errors get pretty obvious (very small) as we move to longer arrival delays and longer departure delays. However, it is harder to predict the satisfaction when the delays are more minimal.

In an attempt to see this a bit better, I changed the scale to blow up the plot. Although we missed 2 points taking this approach, this helps us to see the information at the shorter arrival and departure delays. It is clear that the predictions are better the longer the delays.

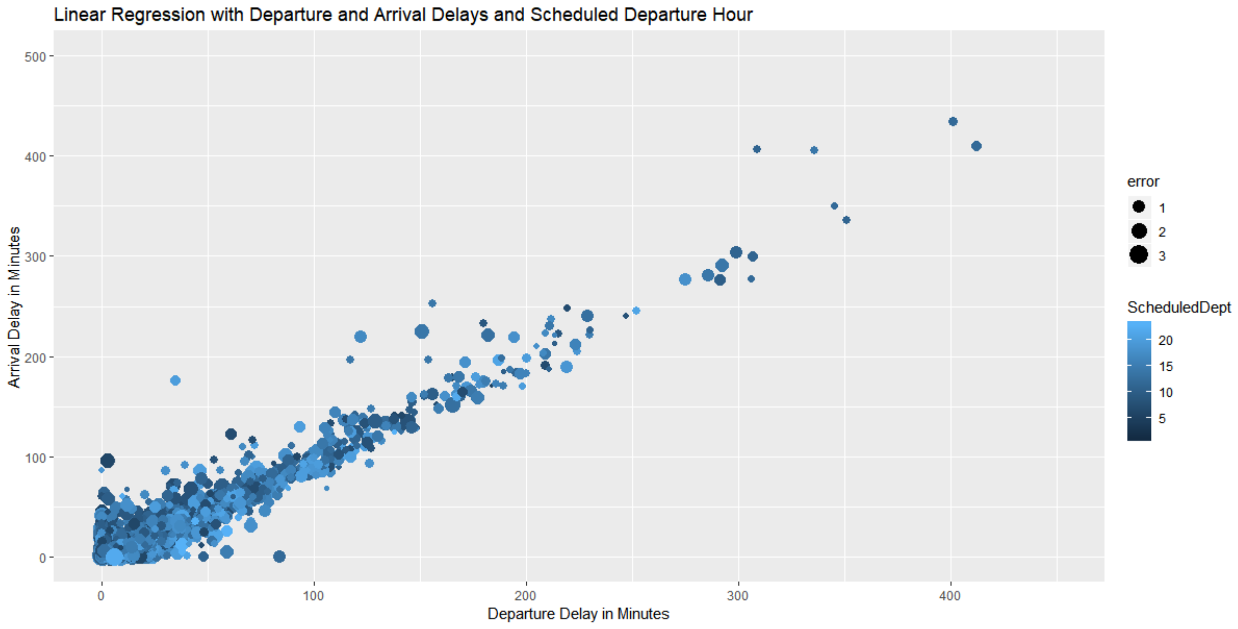


Figure 31: Zoomed Best LM showing Departure and Arrival Delays with Prediction Error

Zooming in as shown in the plots above and below show there is definitely a pattern, but as we look at the errors, the size of the dots are smaller as the delays get larger meaning our model is doing a better job at predicting in these circumstances. This next zoom reflects an area with a drop of 41 observations.

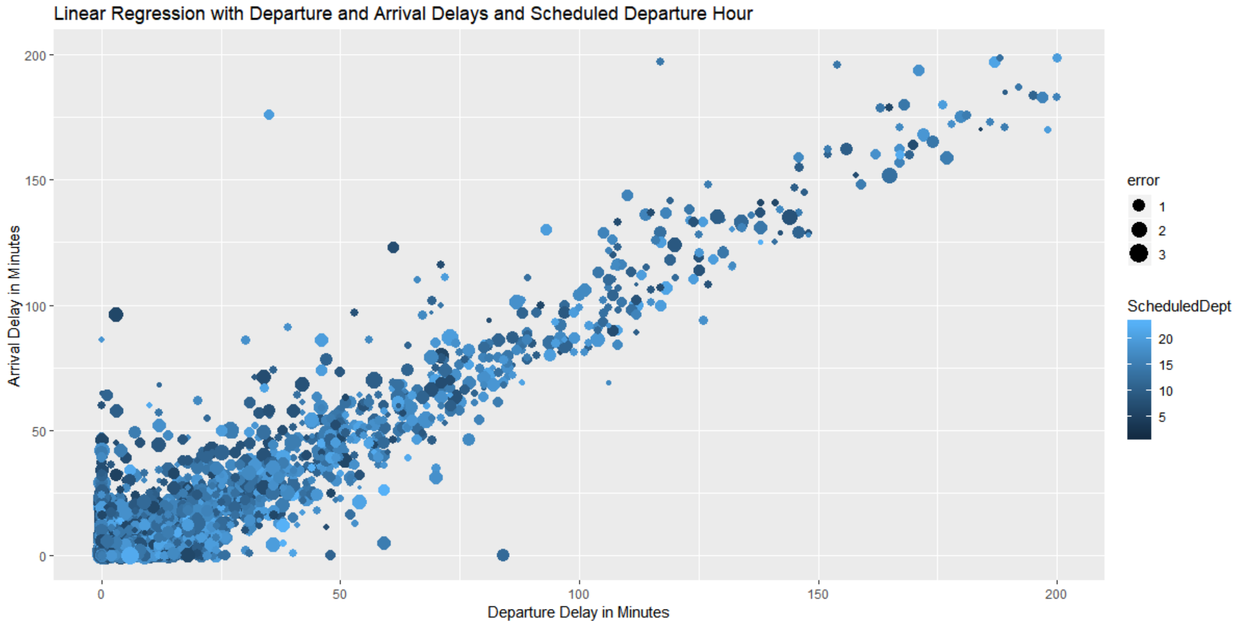


Figure 32: Zoomed Again Best LM showing Departure and Arrival Delays with Prediction Error

Support Vector Machine (SVM)

The next model that I used in reviewing the data was the Support Vector Machine (SVM) which is a "discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side."[1]

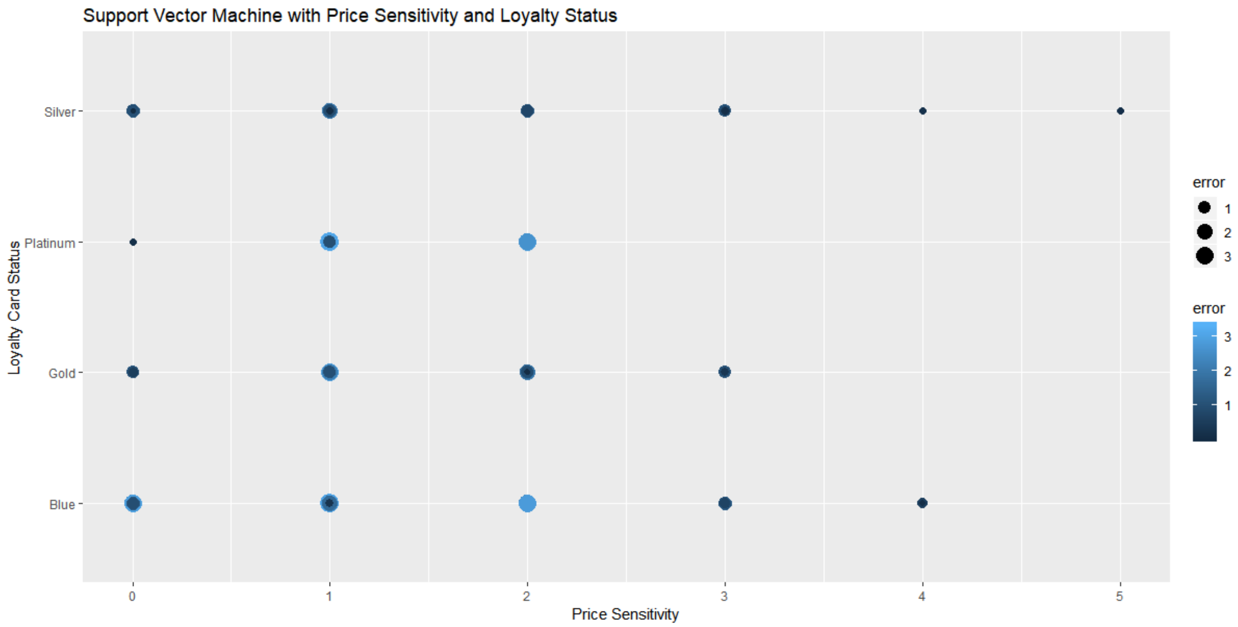


Figure 33: SVM showing Price Sensitivity and Status with Prediction Error

In reviewing the plot above, we see similar results to that in our LM model and that as our errors are smaller when both Price Sensitivity is higher and the loyalty card is Silver or Gold.

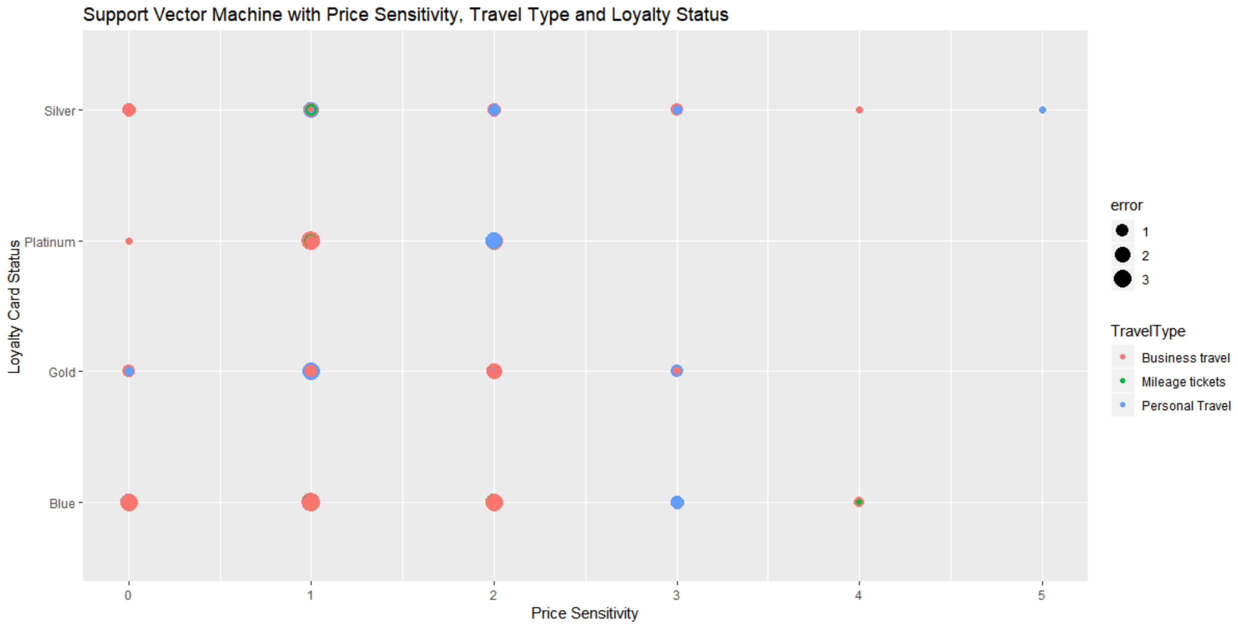


Figure 34: SVM showing Price Sensitivity, Travel Type and Status with Prediction Error

[1] Patel, Savan. Chapter 2: SVM (Support Vector Machine) – Theory. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

Similarly, our model does a better job prediction Satisfaction when Price Sensitivity is higher. This may mean that Price Sensitivity has a significant bearing on the overall satisfaction rating.

Zooming on a similar plot to that done of our Linear Regression Model This next zoom reflects an area with a drop of 33 observations from our test data. You will see that as the Arrival Delay increases and/or the Departure Delay increases, we do a better job prediction satisfaction (reflected by the smaller dot for a smaller error).

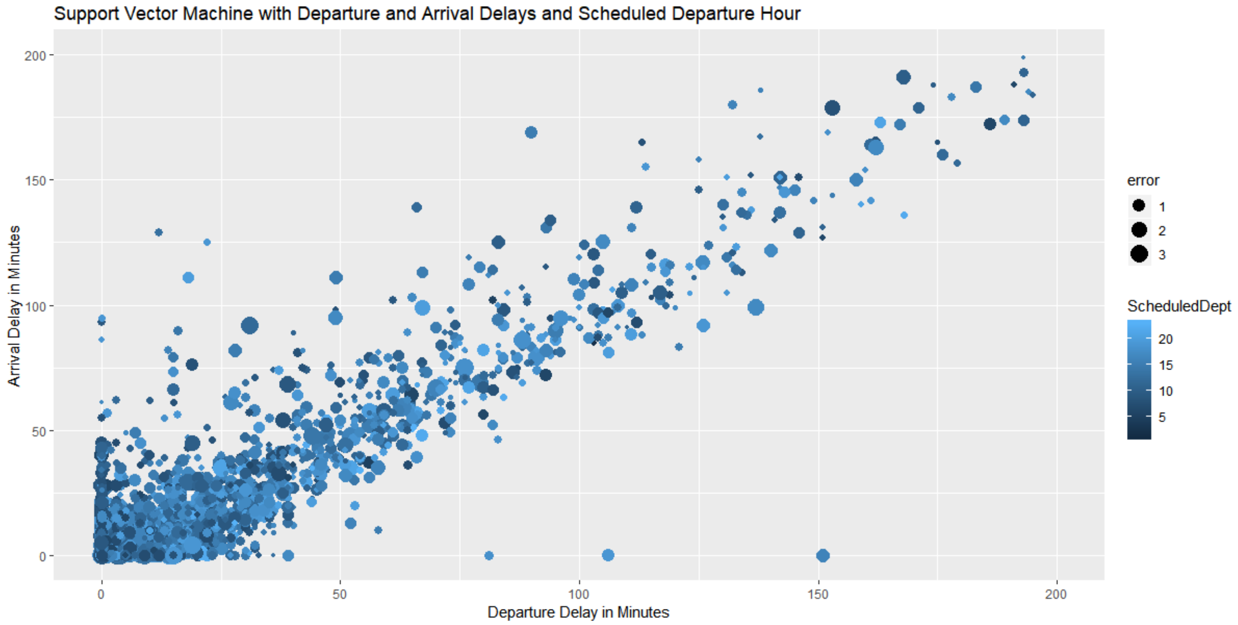


Figure 35: Zoomed SVM showing Departure and Arrival Delays with Prediction Error

Kernel Support Vector Machine (KSVM)

The next model selected was the Kernel Support Vector Machine (KSVM), which is like SVM, but supports the well-known C-svc, nu-svc, (classification) one-class svc (novelty), eps-svr, nu-svr (regression) formulations along with native multi-class classification formulations and the bound-constraint SVM formulations. KSVM also supports class-probabilities output and confidence intervals for regression.[2]

For KSVM, I just calculated the Root Squared Mean Error and did not produce additional plots for this model.

Naïve Bayes

The final model selected was the Naïve Bayes Classifier. "Naive Bayes classifier calculates the probabilities for every factor. Then it selects the outcome with highest probability."[3]

Unfortunately, I had some difficulties with this model and was unable to product results for this final project.

[2] KSVM. https://www.rdocumentation.org/packages/kernlab/versions/0.9-27/topics/ksvm

[3] Patel, Savan. Chapter 1: Supervised Learning and Naïve Bayes Classification – Part 1 (Theory). https://medium.com/machine-learning-101/chapter-1-supervised-learning-and-naive-bayes-classification-part-1-theory-8b9e361897d5

Summary of Modeling

As I review the models (Linear Regression, Support Vector Machines and Naive Bayes), I review the Root Mean Squared Error (RSME) for each of the models. These were as follows:

- For our best Linear Regression model, RSME = 0.7146178
- For our SVM model, RSME = 0.7500606
- For our KSVM model, RSME = 0.7250396
- For Naïve Bayes, RSME = *TBD*

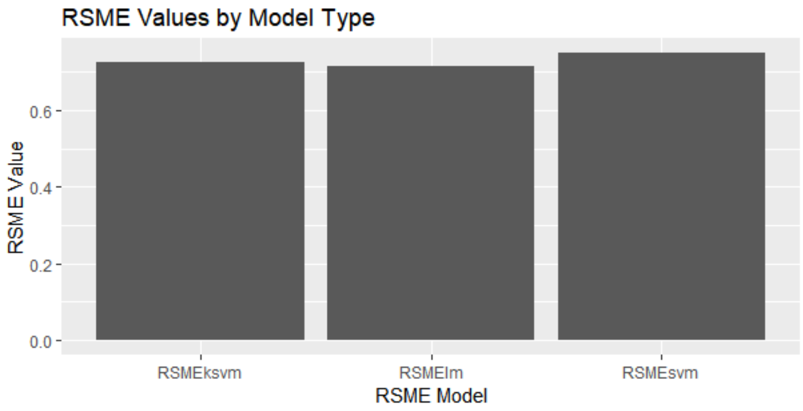


Figure 36: Root Mean Squared Error by Model Type

Using this as an evaluation, I would conclude that our Linear Regression model is the best predictor of Satisfaction based on the fact that it has the lowest Root Mean Squared Error.

Overall Interpretation of the Results and Actionable Insights

As shown in the previous section, there are quite a few variables that have a direct affect on how a traveler will rate his/her overall satisfaction with a specific trip. For example, the loyalty card or airline status (Blue, Silver, Gold, Platinum) has a strong effect on the satisfaction rating. One would assume the perks associated with those cards, has an impact on the satisfaction rating provided by most flyers. This is expected.

Ironically, on our modeling, we found that some unexpected items such as shopping amount in the airport before the flight as well as the eating or drinking in the airport before the flight had influence on the overall rating given to the airline by the flyer. The latter had much less influence, but still played a role in the overall satisfaction rating.

We can also see in our graphs in the visualization section that the destination and origination location have some effect on our overall satisfaction, but they are not of a great influence. We can also see that even though departure and arrival delays can help us predict satisfaction, they are not something that contributes (when reviewing our linear regression model) greatly to the overall satisfaction rating. That being said, that is a good thing since delays in most cases cannot be predicted or even minimized.

After all of our modeling and analysis, the final group of variables determined for our best model are found in the following table.

Variable	Definition
Satisfaction	**Rating from 1 (low) to 5 (high) of satisfaction**
Airline Status	Status with airline (frequent flyer)
Age	Age of traveler
Gender	Gender of traveler – male or female
Scheduled Departure Hour	Hour the Flight was scheduled to depart
Class	Class of travel (first, business, etc.)
Price Sensitivity	Sensitivity to ticket price
Arrival Delay Greater Than 5 Minutes	Arrival Delay greater than 5 minutes
Type of travel	Business, personal, mileage award
Percentage on other airlines	Percentage of travel on other airline carriers
Year of First Flight	The year the first airline flight was taken
Shopping Amount	Amount of money spent shopping in the airport
Flight Cancelled	If flight was cancelled (yes/no)

Figure 37: Final Variables Determined in Modeling

Recommendations

As a result of the analysis done on this data, I would recommend the following for Southeast Airlines:

- Target Male Flyers

  This can be done by sending specials deals to male travelers or creating a campaign to target getting more male travelers to fly with Southeast Airlines.

- Increase Loyalty Members

  Since it is clear that Silver, Gold, and Platinum flyers are more satisfied by our plots and the positive coefficients in our model, it would be good for Southeast Airlines to provide more ways to increase loyalty members. This could be done by finding ways to bump flyers up to the next loyalty level. For example, offering double miles for certain fights or decrease the requirements to obtain the next loyalty level.

- Pricing

  Since sensitivity to price affects satisfaction, offering reduced pricing for certain legs or flights or possibly discounted pricing for different types of travel would give Southeast Airlines the ability to increase their overall satisfaction rating for those customers that are more sensitive to price.

- Shopping

  I found this particular variable to be unexpected, but since it has some influence – it should be addressed. Southeast Airlines could provide coupons for shopping in the airports before a flight as a perk. For example, shop the XYZ Store for 15% off before your flight.

- Delays

  Obviously, minimizing delays would help the airline increase satisfaction. This might be as easy as extending the arrival time by 5 or 10 minutes so that it is a more attainable goal so that it seems that the flight is closer to the scheduled time.

Appendix

In the appendix we have provided the code used to make the package installation, data load and manipulation, analysis and to reproduce the charts and plots provided in this document. To help to compartmentalize the code associated with this project, I elected to separate the code into components or functional areas.

Package and Function Load Section

The code in this section provides the list of packages required and loading those into the R environment for execution of code in the other modules. I found that the function only worked sometimes and I had to often load a package manually while I was working within R-Studio.

```
#
# Course: IST687
# Name: Joyce Woznica
# Project Code: Package Loading
# Due Date: 03/19/2019
# Date Submitted:
#
# Package Section
# ------------------------------------------------------------------

install.packages("plyr", dependencies=TRUE)
library(plyr)

#specify the packages of interest
packages=c("readxl", "arules",  "arulesviz", "kernlab", "e1071", "gridExtra", "ggplot2", "caret", "CRAN", 
"zipcodes",
           "stargazer", "gmodels", "pastecs", "Hmisc", "reshape2", "plyr", "plotly", "psych", "maps", "ggmap",
           "dplyr")

#use this function to check if each package is on the local machine
#if a package is installed, it will be loaded
#if any are not, the missing package(s) will be installed and loaded
package.check <- lapply(packages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
})

#verify they are loaded
search()

# I find this does not always work, so added to install when required here
install.packages("plyr", dependencies = TRUE)
library(plyr)
```

Data Load and Clean-up

The code in this section provide the code used to load the data and clean it readying it for manipulation.

```
#
# Course: IST687
# Name: Joyce Woznica
# Project Code: Load and Manipulate Data
# Due Date: MM/DD/2019
# Date Submitted:
#
# read in a dataset so that it can be useful.
# use this package to read in a XLS file

#
airlineSatSurvey <- read_excel("C:/ProjectIST687/SatisfactionSurvey2.xlsx")

# this is excellent! from Hsmic
satSurvey <- as.data.frame(airlineSatSurvey)
```

```
# rename the columns to make easier:
# Satisfaction = Satisfaction
# Airline Status = Status
# Age = Age
# Gender = Gender
# Price Sensitivity = PriceSens
# Year of First Flight = FFYear
# % of Flight with other Airlines = PercOther
# No of Flight p.a. = NumFlights
# Type of Travel = TravelType
# No. of other Loyalty Cards = NumCards
# Shopping Amount at Airport = ShopAmount
# Eating and Drinking at Airport = EatDrink
# Class = Class
# Day of Month = MonthDay
# Flight date = FlightDate
# Airline Code = AirlineCode
# Airline Name = Airline
# Origin City = OrigCity
# Origin State = OrigState
# Destination City = DestCity
# Destination State = DestState
# Scheduled Departure Hour = SchDeptHour
# Departure Delay in Minutes = DeptDelayMins
# Arrival Delay in Minutes = ArrDelayMins
# Flight cancelled = Cancelled
# Flight time in minutes = FlightMins
# Flight Distance = Distance
# Arrival Delay greater than 5 Mins = ArrDelayGT5

newColNames <- c("Satisfaction","Status", "Age", "Gender", "PriceSens",
                 "FFYear", "PercOther", "NumFlights", "TravelType", "NumCards",
                 "ShopAmount", "EatDrink", "Class", "MonthDay", "FlightDate",
                 "AirlineCode", "Airline", "OrigCity", "OrigState",
                 "DestCity", "DestState", "SchDeptHour",
                 "DeptDelayMins", "ArrDelayMins", "Cancelled",
                 "FlightMins", "Distance", "ArrDelayGT5")

colnames(satSurvey)<-newColNames
View(satSurvey)

# Remove spaces (for beginning and end of lines)
satSurvey$Gender<-trim(satSurvey$Gender)
# remove ", STATE_ABBR" from Cities
satSurvey$OrigCity<-trimCity(satSurvey$OrigCity)
satSurvey$DestCity<-trimCity(satSurvey$DestCity)

# NA appears when plane doesn't take off
# if Cancelled = "Yes" - set all delay and arrival minutes to 0
satSurvey$DeptDelayMins[satSurvey$Cancelled=="Yes"]<-0
satSurvey$ArrDelayMins[satSurvey$Cancelled=="Yes"]<-0
satSurvey$FlightMins[satSurvey$Cancelled=="Yes"]<-0
satSurvey$FlightMins[is.na(satSurvey$FlightMins)]<-0
satSurvey$ArrDelayMins[is.na(satSurvey$ArrDelayMins)]<-0

# anymore NAs?
na_count<-sapply(satSurvey, function(x) sum(length(which(is.na(x)))))
na_count
# remove the lines that have no satistfaction value
satSurvey<-na.omit(satSurvey)
```

Project Functions

The code in this section has a set of functions that I created throughout the class term, of which some were used in this project around data cleansing, the addressing of blank data, and general functions to manipulate the data.

```
#
# Course: IST687
# Name: Joyce Woznica
# Project Code: Package Functions
# Due Date: 03/19/2019
# Date Submitted:
#

# function section
# ------------------------------------------------------------------
# helpful functions for removing spaces
# need to clean out extra spaces from ends of lines
trim.leading<-function(x) {sub("^\\s+","",x)}
trim.trailing<-function(x) {sub("\\s+$","",x)}
trim<-function(x) {sub("^\\s+|\\s+$","",x)}
trimCity<-function(x) {sub("\\,.*$","",x)}
trimSlash<-function(x) {sub("/.*$","",x)}

# Need to find "buckets" for the information
# maybe pick under 25%, 25% to 50%, 50% to 75%, 75% and up?
buildCutOffs<- function(mini, maxi, numcuts)
{
  index<-numcuts
  cutoffs<-c(0)
  while(index>=1)
  {
    cutoffs<- c(cutoffs, round(maxi/index))
    index<-index-1
  }
  return(cutoffs)
}

# convert exponential to decimal
toDecimal<-function(x) {format(x,scientific = FALSE)}

# State Name to Abbreviation
name2abbr <- function (st)
{
  statesDF <- data.frame(state.abb, state.name)
  colnames(statesDF)<-c("abbr","name")
  res<-statesDF$abb[match(st, statesDF$name, nomatch="NA")]
  substr(res,1,3)
}

## geocoding function using OSM Nominatim API
## details: http://wiki.openstreetmap.org/wiki/Nominatim
## made by: D.Kisler

install.packages("jsonlite")
install.packages("tidyverse")

library(jsonlite)
library(tidyverse)

nominatim_osm <- function(address = NULL)
{
  if(suppressWarnings(is.null(address)))
    return(data.frame())
  tryCatch(
    d <- jsonlite::fromJSON(
      gsub('\\@addr\\@', gsub('\\s+', '\\%20', address),
           'http://nominatim.openstreetmap.org/search/@addr@?format=json&addressdetails=0&limit=1')
    ), error = function(c) return(data.frame())
  )
  if(length(d) == 0) return(data.frame())
  return(data.frame(lon = as.numeric(d$lon), lat = as.numeric(d$lat)))
}

#dplyr will be used to stack lists together into a data.frame and to get the pipe operator '%>%'
suppressPackageStartupMessages(library(dplyr))

NewLatLon<-function(addresses){
  d <- suppressWarnings(lapply(addresses, function(address) {
    #set the elapsed time counter to 0
    t <- Sys.time()
    #calling the nominatim OSM API
    api_output <- nominatim_osm(address)
    #get the elapsed time
    t <- difftime(Sys.time(), t, 'secs')
    #return data.frame with the input address, output of the nominatim_osm function and elapsed time
    return(data.frame(address = address, api_output, elapsed_time = t))
  }) %>%
    #stack the list output into data.frame
    bind_rows() %>% data.frame())
  #output the data.frame content into console
  return(d)
}
```

Project Code

The code in this section provides the code used to run descriptive statistics, visualizations, and modeling that was used for the project. The code used to create plots and outputs are noted in the comments.

```
#
# Course: IST687
# Name: Joyce Woznica
# Project Code: Descriptive Statistics
# Due Date: MM/DD/2019
# Date Submitted:
#

# ---------------------------------------------------------------------------------------------------------------
# DESCRIPTIVE STATISTICS
# from pastecs package
# only works on numeric variables in the set
# I did not find this useful
# Reflected in Figure 3
stat.desc(satSurvey, basic=F)

# overall description (lengthy)
describe(satSurvey)

test<-as.matrix(satSurvey)
# Reflected in Figure 2
summary(test)

# descriptive statistics from using stargazer package
# Reflected in Figure 5
stargazer(satSurvey, type="text")
# general Analysis
summary(satSurvey$Age)
summary(satSurvey$Satisfaction)
summary(satSurvey$DeptDelayMins)
summary(satSurvey$ArrDelayMins)
summary(satSurvey$Satisfaction)
mean(satSurvey$NumFlights)

# find the mean satisfaction by airline
airlineSatmeans<- ddply(satSurvey, .(Airline), summarize, SatValue = mean(Satisfaction))
# Reflected in Figure 7
newASM <-airlineSatmeans[order(-airlineSatmeans$SatValue),]
newASM
# Reflected in Figure 6
summary(airlineSatmeans)
range(airlineSatmeans$SatValue)
```

```
# from gmodels package - not right for this, but just playing around with statiistics
chisq.test(satSurvey$Gender,satSurvey$Satisfaction)

# ----------------------------------------------------------------------------------------------------------------
# INITIAL VISUALIZATION
# Plot the Mean Satisfaction Rate by Airline
# Reflected in Figure 8
g <- ggplot(newASM)
g <- g + geom_bar( aes(x = reorder(Airline, -SatValue), y=SatValue), stat="identity", fill="blue", alpha=0.7)
g <- g + theme(axis.text.x = element_text(angle = 45, hjust = 1))
g <- g + scale_y_continuous(name="Satisfaction Rating", limits=c(0, 3.75), breaks=seq(0,4,.15))
g <- g + ggtitle("Average Satisfaction Rating by Airline")
g <- g + xlab("Airline")
g

# satisfaction mean by location (state) - Origination
OrigState.means<- ddply(satSurvey, .(OrigState), summarize, SatValue = mean(Satisfaction))
colnames(OrigState.means)<- c("State", "OrigStateSat")
Ordered.OS.SatMeans <- OrigState.means[order(-OrigState.means$OrigStateSat),]

# satisfaction mean by location (state) - Destination
DestState.means<- ddply(satSurvey, .(DestState), summarize, SatValue = mean(Satisfaction))
colnames(DestState.means)<- c("State", "DestStateSat")
Ordered.DS.SatMeans <- DestState.means[order(-DestState.means$DestStateSat),]
StateSatdf <- merge(OrigState.means, DestState.means, by = "State")

# Melt for plotting
# Refected in Figure 14
meltedStatedf <- melt(StateSatdf, id.vars = "State")
g <- ggplot(meltedStatedf, aes(x=State, y=value, fill=variable))
g <- g + geom_bar(stat='identity', position='dodge')
g <- g + theme(axis.text.x = element_text(angle = 90, hjust = 1))
g <- g + ggtitle("Average Satisfaction Rating by State")
g <- g + xlab("State") + ylab("Satisfaction Rating")
g

# Some initial plotting of the data to see what we have
# maybe try sampling data from our set (maybe 1000 values) and using that
testS<-sample(satSurvey$Satisfaction,1000,replace=FALSE)
testDD<-sample(satSurvey$DeptDelayMins,1000, replace=FALSE)
testdf<-data.frame(testS,testDD)
plot(testS,testDD)

# Plotting a sample of 1000 items of Satisfaction and Departure Delays
# Reflected in Figure 15
g <- ggplot(testdf, aes(x=testDD, y=testS)) + geom_point()
g <- g + stat_smooth(method= "lm", col="red")
g <- g + ggtitle("Satisfaction Rating by Departure Delay in Minutes")
g <- g + xlab("Departure Delay in Minutes") + ylab("Satisfaction Rating")
g

# Are men or women more satisfied?
# Reflected in Figure 9
g<- ggplot(satSurvey, aes(x=Gender, fill=factor(Satisfaction)))
g<- g+ ggtitle("Satisfaction by Gender")
g<- g+ xlab("Gender") + ylab("Satisfaction Rating")
g<- g + geom_bar(position="dodge")
g

# boxplot male/female by satistfaction
# with boxplot
# do this again with buckets or something!
g<- ggplot(satSurvey, aes(group=Gender,x=Gender,y=Satisfaction))
g<- g + geom_boxplot(aes(fill=factor(Gender)))
g<- g + ggtitle("Satisfaction by Gender") + theme(plot.title=element_text(hjust=0.5))
g

# also did against month - just for my own benefit
g<- ggplot(satSurvey, aes(group=AirlineCode,x=AirlineCode,y=Satisfaction))
g<- g + geom_boxplot(aes(fill=factor(AirlineCode)))
g<- g + ggtitle("Satisfaction by Airline Code") + theme(plot.title=element_text(hjust=0.5))
g
```

```
hist(satSurvey$Satisfaction)
table(satSurvey$Gender)

# which airline has the best satisfaction
# Reflected in Figure 11
g<- ggplot(satSurvey, aes(x=AirlineCode, fill=factor(Satisfaction)))
g<- g + geom_bar(position="dodge")
g<- g+ ggtitle("Satisfaction Rating Frequency by Airline")
g<- g + xlab("Airline Code") + ylab("Frequency of Satisfaction Rating")
g

# Does Travel Type affect Satisfaction?
# Reflected in Figure 19
g<- ggplot(satSurvey, aes(x=TravelType, fill=factor(Satisfaction)))
g<- g + geom_bar(position="dodge")
g<- g+ ggtitle("Airline Satisfaction Rating Frequency by Travel Type")
g<- g+ xlab("Travel Type") + ylab("Satisfaction Rating Frequency")
g

# Does Status affect Satisfaction?
# Reflected in Figure 20
g<- ggplot(satSurvey, aes(x=Status, fill=factor(Satisfaction)))
g<- g + geom_bar(position="dodge")
g<- g+ ggtitle("Airline Satisfaction Rating Frequency by Loyalty Card Status")
g<- g+ xlab("Card Status") + ylab("Satisfaction Rating Frequency")
g

# Does Shopping and Eating in the Airport affect Satisfaction?
# Add scale for y axis
g <- ggplot(satSurvey, aes(x=EatDrink, y = Satisfaction,
                           fill = ShopAmount))
g <- g + geom_bar(stat = "identity")
g

# maybe plot by age buckets?
# create bins/buckets for box plot
# really want to check by Satisfaction though
# Reflected in Figure 10
plotBuckets<-buildCutOffs(min(satSurvey$Age),max(satSurvey$Age),5)
SatSurvWB <- satSurvey
SatSurvWB$Bucket<-cut(satSurvey$Age,plotBuckets)
g<- ggplot(SatSurvWB)
g<- g + geom_boxplot(aes(Bucket, Age, fill=factor(Bucket)))
g<- g + ggtitle("Age of Flyers Surveyed") + theme(plot.title=element_text(hjust=0.5))
g<- g + xlab("Age Groupings") + ylab("Satisfaction Value")
g

# now do a line plot against Date for satisfaction
# Does time of year matter?
g <- ggplot(satSurvey, aes(x=FlightDate, y=Satisfaction))
g <- g + geom_line(size=1, color="navy")
g <- g + ylab("Satisfaction")
g <- g + ggtitle("Satisfaction by Date")+theme(plot.title=element_text(hjust=0.5))
g

# Does time of year matter?
g <- ggplot(satSurvey, aes(x=FlightDate, y=Satisfaction))
g <- g + geom_point(size=1, color="orange")
g <- g + ylab("Satisfaction")
g <- g + ggtitle("Satisfaction by Date")+theme(plot.title=element_text(hjust=0.5))
g <- g + xlab("Date of Flight")
g


# data needs to be cleaned for missing values
# Add scale for y axis
g <- ggplot(satSurvey, aes(x=Age, y=Satisfaction, group=Gender ))
g <- g + geom_bar(stat="identity", position="dodge", color="steelblue",
                  aes(fill=factor(Gender)))
g
```

```
# Look at all the data via a scatter plot
#         Create a scatter chart (geom_point),
#         x-axis is AirlineCode
#         y-axis is Satisfaction
#         dot size represents Status
#         color represents DeptDelayMins
# Reflected in Figure 13
g <- ggplot(satSurvey, aes(x=AirlineCode, y=Satisfaction))
g <- g + geom_point(aes(color=Status, size=DeptDelayMins))
g <- g + ggtitle("Satisfaction versus Airline Code, Status and Delay Minutes")
g <- g + theme(plot.title=element_text(hjust=0.5))
g <- g + xlab("Airline Code") + ylab("Satisfaction Rating")
g

#         Create a scatter chart (geom_point),
#         x-axis is AirlineCode
#         y-axis is Satisfaction
#         dot size represents ArrDelayMins
#         color represents DeptDelayMins
g <- ggplot(satSurvey, aes(x=AirlineCode, y=Satisfaction))
g <- g + geom_point(aes(color=ArrDelayMins, size=DeptDelayMins))
g <- g + ggtitle("Satisfaction versus Airline Code, Arrival and Delay Minutes")
g <- g + theme(plot.title=element_text(hjust=0.5))
g <- g+ xlab("Airline Code") + ylab("Satisfaction Rating")
g

# Histogram of satisfaction rating vs number of flights
# Reflected in Figure 22
g <- ggplot(satSurvey, aes(x=Satisfaction,y=NumFlights, fill=AirlineCode))
g <- g + geom_bar(stat="identity")
g <- g + ggtitle("Number of Flight Frequency by Airline Satisfaction Rating and Code")
g <- g + xlab("Satisfaction Rating") + ylab("Number of Flight Frequency")
g

# Look at all the data via a scatter plot
#         Create a scatter chart (geom_point),
#         x-axis is First Flight Year
#         y-axis is Satisfaction
#         dot size represents Number of Flights
#         color represents Flights on other airlines
# Reflected in Figure 21
g <- ggplot(satSurvey, aes(x=FFYear, y=Satisfaction))
g <- g + geom_point(aes(color=PercOther, size=NumFlights))
g <- g + ggtitle("Satisfaction versus First Flight Year, Percentage on Other Airlines, Number of Flights")
g <- g + theme(plot.title=element_text(hjust=1))
g <- g + scale_x_continuous(name="Year of First Flight", limits=c(min(satSurvey$FFYear), max(satSurvey$FFYear)))
g

# play around with bar graphs
# Reflected in Figure 12
g <- ggplot(satSurvey, aes(x=Satisfaction, y = Age,
                           fill = AirlineCode))
g <- g + geom_bar(stat = "identity")
g <- g + ggtitle("Satisfaction Rating Frequency by Airline Code and Age")
g

# --------------------------------------------------------------------------------------------------------------
# MAP VISUALIZATIONS

us <- map_data("state")

# need to create a new frame - elected to create two frames (one for Orignation Airport, other for Destination
Airport)
# Satisfaction
# Origination - which is OrigCity, OrgState (but the state needs to be an abbrevation)
# Destation - which is DestCity, DestState (but the state needs to be an abbreviation)
# first remove all columns except the 3 we need
orig.mapDF <- data.frame(satSurvey$Satisfaction, satSurvey$OrigCity, satSurvey$OrigState)
dest.mapDF <- data.frame(satSurvey$Satisfaction, satSurvey$DestCity, satSurvey$DestState)
colnames(orig.mapDF) <- c("Satisfaction", "OrigCity", "OrigState")
colnames(dest.mapDF) <- c("Satisfaction", "DestCity", "DestState")
# trim the City to remove everything after the "/"
```

```r
orig.mapDF$OrigCity <- trimSlash(orig.mapDF$OrigCity)
dest.mapDF$DestCity <- trimSlash(dest.mapDF$DestCity)
# get state abbreviations
orig.mapDF$OrigStateAbbr <- name2abbr(orig.mapDF$OrigState)
dest.mapDF$DestStateAbbr <- name2abbr(dest.mapDF$DestState)

# get rid of pacific territories and other non-states
orig.mapDF<-na.omit(orig.mapDF)
dest.mapDF<-na.omit(dest.mapDF)

# remove Hawaii and Alaska (for easier plotting)
orig.mapDF <- subset(orig.mapDF, orig.mapDF$OrigStateAbbr !="HI")
dest.mapDF <- subset(dest.mapDF, dest.mapDF$DestStateAbbr !="HI")
orig.mapDF <- subset(orig.mapDF, orig.mapDF$OrigStateAbbr !="AK")
dest.mapDF <- subset(dest.mapDF, dest.mapDF$DestStateAbbr !="AK")

# create combination Origination and Destination with city, ST in lower case
orig.mapDF$Origination <- paste(orig.mapDF$OrigCity,orig.mapDF$OrigStateAbbr, sep=', ')
dest.mapDF$Destination <- paste(dest.mapDF$DestCity,dest.mapDF$DestStateAbbr, sep=', ')
# convert to all lower case
orig.mapDF$Origination<-tolower(orig.mapDF$Origination)
dest.mapDF$Destination<-tolower(dest.mapDF$Destination)

# remove the unnecessary columns now
orig.mapDF <- data.frame(orig.mapDF$Origination, orig.mapDF$Satisfaction, orig.mapDF$OrigState)
colnames(orig.mapDF)<-c("Origination", "Satisfaction", "State")
dest.mapDF <- data.frame(dest.mapDF$Destination,dest.mapDF$Satisfaction, dest.mapDF$DestState)
colnames(dest.mapDF)<-c("Destination", "Satisfaction", "State")

# need to summarize by mean satisfaction
# satisfaction mean by location (state) - Destination
Dest.means <- ddply(dest.mapDF, .(Destination), summarize, SatValue = mean(Satisfaction))
Orig.means <- ddply(orig.mapDF, .(Origination), summarize, Satvalue = mean(Satisfaction))

# add Latitudes and longitudes
#first.omeans <-head(Orig.means)
#first.omeans$geoCode <- NewLatLon(first.omeans$Origination)
Orig.means$geoCode<-NewLatLon(Orig.means$Origination)
Dest.means$geoCode<-NewLatLon(Dest.means$Destination)

# plot by Origination Airport Location
# Reflected in Figure 16 and 18
oplot <- ggplot(Orig.means,aes(geoCode$lon,geoCode$lat))
oplot <- oplot + geom_polygon(data=us,aes(x=long,y=lat,group=group),color='gray',fill='white')
oplot <- oplot + geom_point(aes(color = Satvalue),size=5)
oplot <- oplot +  xlim(-125,-65)+ylim(20,50)
oplot <- oplot + ylab("Latitude") + xlab("Longitude")
oplot <- oplot + ggtitle("Average Satisfaction by Origination Airport Location")
oplot

# plot by Destination Airport Location
# Reflected in Figure 17 and 18
dplot <- ggplot(Orig.means,aes(geoCode$lon,geoCode$lat))
dplot <- dplot + geom_polygon(data=us,aes(x=long,y=lat,group=group),color='gray',fill='white')
dplot <- dplot + geom_point(aes(color = Satvalue),size=5)
dplot <- dplot +  xlim(-125,-65)+ylim(20,50)
dplot <- dplot + ylab("Latitude") + xlab("Longitude")
dplot <- dplot + ggtitle("Average Satisfaction by Destimation Airport Location")
dplot

# random plotting
plot(satSurvey$FlightDate, satSurvey$Satisfaction)

# --------------------------------------------------------------------------------------------------------------
# REGRESSION ANALYSIS AND MODELING

# Just some test regressions on certain variables to see how they are correlated, if at all
# Reflected in Figure 23
testReg1 <- lm(satSurvey$Satisfaction ~ satSurvey$FlightMins)
summary(testReg1)
```

```r
# Reflected in Figure 24
testReg2 <- lm(satSurvey$Satisfaction ~ satSurvey$Status)
s.model <- summary(testReg2)
# Reflected in Figure 24
testReg3 <- lm(satSurvey$Satisfaction ~ satSurvey$NumFlights)
s.model <- summary(testReg3)
# Reflected in Figure 24
testReg4 <- lm(satSurvey$Satisfaction ~ satSurvey$PriceSens)
s.model <- summary(testReg4)


coefficients(testReg1) # model coefficients
confint(testReg1, level=0.95) # CIs for model parameters

# create a smaller dataframe of the top 3 airlines based on mean satisfaction
topAsatSurvey1 <- as.data.frame(subset(satSurvey, AirlineCode == "VX"))
topAsatSurvey2 <- as.data.frame(subset(satSurvey, AirlineCode == "HA"))
topAsatSurvey3 <- as.data.frame(subset(satSurvey, AirlineCode =="AS"))
topAsatSurvey <- rbind(topAsatSurvey1, topAsatSurvey2, topAsatSurvey3)

allTopReg <- lm(topAsatSurvey$Satisfaction ~., data = as.data.frame(topAsatSurvey))
top.model <- summary(allTopReg)
top.model$adj.r.squared
TopAICModel <- step(allTopReg, data=topAsatSurvey, direction="backward")
# AIC Results
#Step:  AIC=-4362.92
#topAsatSurvey$Satisfaction ~ Status + Age + Gender + PriceSens +
#  FFYear + PercOther + TravelType + AirlineCode + SchDeptHour +
#  Cancelled + ArrDelayGT5
# Reflected Figure 25
BestTopModel <- lm(formula=topAsatSurvey$Satisfaction ~ Status + Age + Gender + PriceSens +
                     FFYear + PercOther + TravelType + AirlineCode + SchDeptHour +
                     Cancelled + ArrDelayGT5, data = as.data.frame(topAsatSurvey))
summary(BestTopModel)

# output the data in a more readable format
#stargazer(BestlmModel, allReg, type="text", title="Linear Regression Output", align=TRUE)
SG <- stargazer(BestTopModel, type="text", title="Linear Regression Output", align=TRUE)

# look at linear regression on all variables
# this is useless - want a Multiple R-squared and p-value for those that matter!
# multiple linear regression to see if the multiple R-squared show correlation
# and review the p-values
allReg <- lm(satSurvey$Satisfaction ~., data=as.data.frame(satSurvey))
s.model <- summary(allReg)
s.model$coef[,4]
s.model$adj.r.squared

# step through to get the best model
AICModels<-step(allReg,data=satSurvey, direction="backward")
#AIC Results were as follows:
# Call:
#  lm(formula = satSurvey$Satisfaction ~ Status + Age + Gender +
#       PriceSens + FFYear + PercOther + TravelType + ShopAmount +
#       EatDrink + Class + SchDeptHour + Cancelled + ArrDelayGT5,
#     data = as.data.frame(satSurvey))
# AIC = -85696.57
BestlmModel <- lm(formula = satSurvey$Satisfaction ~ Status + Age + Gender +
                    PriceSens + FFYear + PercOther + TravelType + ShopAmount +
                    EatDrink + Class + SchDeptHour + Cancelled + ArrDelayGT5,
                  data = as.data.frame(satSurvey))
# Reflected Figure 25
summary(BestlmModel)

# Reflected in Figure 26
Final.Model <- lm(formula = satSurvey$Satisfaction ~ Status + Age + Gender +
                    PriceSens + FFYear + PercOther + TravelType + ShopAmount +
                    Class + SchDeptHour + Cancelled + ArrDelayGT5,
                  data = as.data.frame(satSurvey))
summary(Final.Model)

# output the data in a more readable format
```

```
#stargazer(BestlmModel, allReg, type="text", title="Linear Regression Output", align=TRUE)
stargazer(Final.Model, type="text", title="Linear Regression Output", align=TRUE)
summary.Model<- summary(Final.Model)

# SVM, KSVM, NB
# a new method for sampling the data for SVM and NB
# ** Due to the time it takes to run this output - I elected to take a random sample of 15,000 of the TrainData
# ** and 5,000 of the test data to test this model
svm.train <- satSurvey[sample(nrow(satSurvey), size=15000, replace=FALSE),]
table(svm.train$Satisfaction)/length(svm.train$Satisfaction)

svm.test <- satSurvey[sample(nrow(satSurvey), size=5000, replace=FALSE),]
nb.test <- satSurvey[sample(nrow(satSurvey), size=5000, replace=FALSE),]
ksvm.test <- satSurvey[sample(nrow(satSurvey), size=5000, replace=FALSE),]
lm.test <- satSurvey[sample(nrow(satSurvey), size=5000, replace=FALSE),]

table(svm.test$Satisfaction)/length(svm.test$Satisfaction)

# Linear Regression Prediction
lmPred <- predict(Final.Model, lm.test, type="response")
lmPred <- data.frame(lmPred)
compTable <- data.frame(lm.test[,1],lmPred[,1])
colnames(compTable) <- c("test", "Pred")
# this is the RMSE (how low)
RSMElm<-sqrt(mean((compTable$test-compTable$Pred)^2))
RSMElm

# plot some LM results
# compute absolute error for each case
compTable$error <- abs(compTable$test - compTable$Pred)
# create new dataframe for plotting
# Reflected in Figure 28
lmPlot <- data.frame(compTable$error, lm.test$PriceSens, lm.test$Status)
colnames(lmPlot) <- c("error", "PriceSens", "Status")
plotlm1 <- ggplot(lmPlot, aes(x=PriceSens, y=Status)) +
  geom_point(aes(size=error, color=error)) +
  ggtitle("Linear Regression with Price Sensitivity and Loyalty Status") +
  xlab("Price Sensitivity") + ylab("Loyalty Card Status")
plotlm1

# Reflected in Figure 29
lmPlot <- data.frame(compTable$error, lm.test$TravelType, lm.test$Status, lm.test$PriceSens)
colnames(lmPlot) <- c("error", "TravelType", "Status", "PriceSens")
plotlm2 <- ggplot(lmPlot, aes(x=PriceSens, y=Status)) +
  geom_point(aes(size=error, color=TravelType)) +
  ggtitle("Linear Regression with Price Sensitivity, Travel Type and Loyalty Status") +
  xlab("Price Sensitivity") + ylab("Loyalty Card Status")
plotlm2

# Reflected in Figure 30
lmPlot <- data.frame(compTable$error, lm.test$DeptDelayMins, lm.test$ArrDelayMins, lm.test$SchDeptHour)
colnames(lmPlot) <- c("error", "DeptDelay", "ArrivalDelay", "ScheduledDept")
plotlm3 <- ggplot(lmPlot, aes(x=DeptDelay, y=ArrivalDelay)) +
  geom_point(aes(size=error, color=ScheduledDept)) +
  ggtitle("Linear Regression with Departure and Arrival Delays and Scheduled Departure Hour") +
  xlab("Departure Delay in Minutes") + ylab("Arrival Delay in Minutes")
plotlm3

# change the scale
# Reflected in Figure 31
plotlm3 <- ggplot(lmPlot, aes(x=DeptDelay, y=ArrivalDelay)) +
  geom_point(aes(size=error, color=ScheduledDept)) +
  ggtitle("Linear Regression with Departure and Arrival Delays and Scheduled Departure Hour") +
  scale_x_continuous(limits=c(0,450)) + scale_y_continuous(limits=c(0,500)) +
  xlab("Departure Delay in Minutes") + ylab("Arrival Delay in Minutes")
plotlm3

# change the scale again
# Reflected in Figure 32
plotlm3 <- ggplot(lmPlot, aes(x=DeptDelay, y=ArrivalDelay)) +
  geom_point(aes(size=error, color=ScheduledDept)) +
  ggtitle("Linear Regression with Departure and Arrival Delays and Scheduled Departure Hour") +
  scale_x_continuous(limits=c(0,200)) + scale_y_continuous(limits=c(0,200)) +
  xlab("Departure Delay in Minutes") + ylab("Arrival Delay in Minutes")
plotlm3


# SUpport Vector Machine (SVM)
svm.model<-
svm(Satisfaction~Status+Age+PriceSens+PercOther+TravelType+Class+DeptDelayMins+ArrDelayMins,data=svm.train)
summary(svm.model)

svm.test$PredS<-predict(svm.model,svm.test, type="votes")
# Compare Observed and Predicted
table.svm <- table(pred = svm.validate$PredS,
                   true = svm.validate$Satisfaction)/length(svm.validate$Satisfaction)

# SVM
# create prediction
svmPred <- predict(svm.model, svm.test, type="votes")
svmPred <- (data.frame(svmPred))
compTable <- data.frame(svm.test[,1],svmPred[,1])
colnames(compTable) <- c("test", "Pred")
# this is the RMSE (how low)
RSMEsvm<-sqrt(mean((compTable$test-compTable$Pred)^2))
RSMEsvm

# plot some SVM results
# compute absolute error for each case
compTable$error <- abs(compTable$test - compTable$Pred)
# create new dataframe for plotting
# Reflected in Figure 33
svmPlot <- data.frame(compTable$error, svm.test$PriceSens, svm.test$Status)
colnames(svmPlot) <- c("error", "PriceSens", "Status")
plotsvm1 <- ggplot(svmPlot, aes(x=PriceSens, y=Status)) +
  geom_point(aes(size=error, color=error)) +
  ggtitle("Support Vector Machine with Price Sensitivity and Loyalty Status") +
  xlab("Price Sensitivity") + ylab("Loyalty Card Status")
plotsvm1

# Reflected in Figure 34
svmPlot <- data.frame(compTable$error, svm.test$TravelType, svm.test$Status, svm.test$PriceSens)
colnames(svmPlot) <- c("error", "TravelType", "Status", "PriceSens")
plotsvm2 <- ggplot(svmPlot, aes(x=PriceSens, y=Status)) +
  geom_point(aes(size=error, color=TravelType)) +
  ggtitle("Support Vector Machine with Price Sensitivity, Travel Type and Loyalty Status") +
  xlab("Price Sensitivity") + ylab("Loyalty Card Status")
plotsvm2

# Reflected in Figure 35
svmPlot <- data.frame(compTable$error, svm.test$DeptDelayMins, svm.test$ArrDelayMins, svm.test$SchDeptHour)
colnames(svmPlot) <- c("error", "DeptDelay", "ArrivalDelay", "ScheduledDept")
plotsvm3 <- ggplot(svmPlot, aes(x=DeptDelay, y=ArrivalDelay)) +
  geom_point(aes(size=error, color=ScheduledDept)) +
  ggtitle("Support Vector Machine with Departure and Arrival Delays and Scheduled Departure Hour") +
  scale_x_continuous(limits=c(0,200)) + scale_y_continuous(limits=c(0,200)) +
  xlab("Departure Delay in Minutes") + ylab("Arrival Delay in Minutes")
plotsvm3

# KSVM
ksvmOutput <- ksvm(Satisfaction~Status+Age+PriceSens+PercOther+TravelType+Class+DeptDelayMins+ArrDelayMins,
                   data=svm.train,kernel="rbfdot",kpar="automatic",C=10,cross=3,prob.model=TRUE)

# test the model
ksvmPred <- predict(ksvmOutput, ksvm.test, type="votes")
ksvmPred <- data.frame(ksvmPred)
str(ksvmPred)
compTable <- data.frame(ksvm.test[,1],ksvmPred[,1])
colnames(compTable) <- c("test", "Pred")
# this is the RMSE (how low)
RSMEksvm<-sqrt(mean((compTable$test-compTable$Pred)^2))
RSMEksvm
```

```
# Naive Bayes
# ** NOT USED **
nb.train <- satSurvey[sample(nrow(satSurvey), size=3000, replace=FALSE),]
nb.test <- satSurvey[sample(nrow(satSurvey), size=1000, replace=FALSE),]

nbOutput <- naiveBayes(Satisfaction~Age+PriceSens+PercOther,
                       data=nb.train)
str(nbOutput)
# create prediction
nbPred <- predict(nbOutput, nb.test)
nbPred <- as.data.frame(nbPred)
str(nbPred)
compTable <- data.frame(nb.test[,1],nbPred[,1])
colnames(compTable) <- c("test", "Pred")
# this is the RSME
RSMEnb <- sqrt(mean(compTable$test-compTable$Pred)^2)
RSMEnb

# Reflected in Figure 36
RSME.frame<-NULL
RSME.frame <- c(RSMEsvm, RSMElm, RSMEksvm)
RSME.frame<-melt(RSME.frame)
RSME.frame$model <- c("RSMEsvm", "RSMElm", "RSMEksvm")
g <- ggplot(data=RSME.frame, aes(x=model, y=value))
g <- g + geom_bar(stat="identity", position="dodge")
g <- g + ggtitle("RSME Values by Model Type")
g <- g + xlab("RSME Model") + ylab("RSME Value")
g
```