IST 687 - Applied Data Science

USA Airline Customer Satisfaction Analysis

M007 Group 1

Logan Clark

Kripesh Jain

Simaant Patil

Neha Shah

Prof: Jeffrey Saltz

Lab Instructor: Stephen Wallace

TA: Ivan Shamshurin

Table of Contents

1.Introduction:

In this project, we have been provided airline data that will be analyzed to provide some actionable insights into what will help improve customer satisfaction for an airline. The data given is from different airlines in the United States, tracking the customer satisfaction by each flight. The satisfaction depends on multiple variables that a customer and airline experience during travel. The scope of the project is to determine the conditions and measures that an airline can provide to improve satisfaction for their customers. We will be focusing on variables that are statistically significant in the outcome of a high customer satisfaction that help us answer the key business questions. Our approaches will include descriptive visualizations, and various modeling techniques like, linear modeling, support vector machines and association rules to generate our insights.

2.Summary Of The Dataset:

The dataset contains about 129889 responses (rows) from airline customers survey over 3 months, and contains data from 14 airlines. It has 28 columns (attributes), which consists of data obtained from surveys submitted by its customers. The columns broadly focus on several categories, including customer's gender, age, number of flights, shopping amount at the airport, type of travel, class of travel, etc.

The main goal of this project is to understand how these factors affect the satisfaction of the customers, which eventually leads to them choosing a particular airline over other. There can different relationships between the customers and the attributes involved and not all would make sense.

Our main focus would be determining these significant factors affecting overall customer satisfaction considering all airlines and also acting as consultants for SouthEast Airlines Co. This, as a group, would help us determine the important business questions an airline company might need to consider and which could eventually help in increasing their profits.

3.Dataset:

The dataset has 29 features in all. They are:

1.**Satisfaction -** It is rated from 1 to 5, which shows customer satisfaction with the least being 1 and the highest being 5.

2.**Airline Status -** The status tells us about the airline company and also how well it is ranked, i.e, Blue, Platinum, Gold.

3.**Age**– The specific customer's age which ranges 15 to 85 years old.

4.**Gender** – Specifies whether a customer is male or female.

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

6.**Year of First Flight -** This attributes shows the first flight of each customer with range of year from 2003 until 2012.

7.**No of Flights p. a. -** This could be the number of flights that each customer has taken ranging from 0 to 100.

8.**Percentage of Flight with other Airlines -** It gives the total percent of flights taken by other airlines.

9.**Type of Travel -** It provides travelling purpose for each consumer, which are business travel, mileage tickets that based on loyalty cards, and personal travel.

10.**No. Of other Loyalty Cards -** Tt could be a kind of membership card of each customer, which would help them to gain benefits such as discounts.

11.**Shopping Amount at Airport -** Shows the amount of shopping done by each customer at the airport and ranges from 0 to 875.

12.**Eating and Drinking at Airport -** It is amount spent by each customer on eating and drinking at the airport and has a range from 0 to 895.

13.**Class -** Consists of three different types of classes such as economy, economy-plus and business class.

14.**Day of Month -** It gives the travelling day of each customer. In this attribute, it is shown as a period of 31 days.

15.**Flight Date -** All of the dates in this attribute are abbreviate the passenger's flight date travel, which were since 2014 and only in January, February, and March.

16.**Airline Code -** It is a unique code that is related to a specific airline.

17.**Airline Name -** There are several airlines company names such as West Airways, Southeast Airlines Co, etc. and they refer to the customers taking a particular airline.

18.**Origin City -** Refers to actual city that the customers have departed from.

19.**Origin State -** Refers exactly to the state the customer has departed from.

20.**Destination City** - The arrival city for a particular travelling customer.

21.**Destination State -** Specifies the arrival state where a flight lands.

22.**Scheduled Departure Hour -** Specifies the time at which passengers are scheduled to depart. Which is starting at 1 am until 23 pm.

23.**Departure Delay in Minutes -** Gives the delay of flight for each consumer and it ranges from 0 1128 minutes

24.**Arrival Delay in Minutes -** Gives the arrival delay for each flight that a consumer takes and ranges from 0 to 1115 minutes .

25.**Flight Cancelled -** Occurs when the airline does not operate a flight at all.

26.**Flight time in minutes** - Indicates the total flight time.

27.**Flight Distance -** The total distance between the two places that the flight connects.

28.**Arrival Delay greater than 5 Minutes** – It shows whether the arrival delay for a flight of a particular airline is more than 5 mins or not.

4.Data Analysis:(for the entire dataset)

The analysis of the data is divided into acquisition, cleaning of the, visualizations of the available data, developing models and predicting results.

5.Data Acquisition:

The first step we performed was to download the dataset provided to us and set up the environment for testing the code and conducting analysis on the data provided.

```
getwd()
setwd("E:/Syracuse University/Academics/Sem 1/IST 687/Projects/Final project")
project= read.csv("satisfactionSurvey.csv")
View(project)
str(project)
summary(project)
dim(project)
```

Fig 1: Data Acquisition

The dataset was downloaded on a local PC and R studio was used to run the different commands on it to perform the analysis.

6.Data Cleaning:

Data cleaning is the most important step as it is necessary to take out the garbage and irrelevant values and also edit the data in a specific manner so that it follows a specific appearance or format for each attribute.

```
project$Flight.date=gsub("-", "/", project$Flight.date, fixed=T )
project$Flight.date=as.Date(project$Flight.date, "%m/%d/%Y")
project$Flight.date= format(project$Flight.date, "%m-%d-%y")
project$Destination.City=gsub(",","", project$Destination.City)
project$Origin.City=gsub(",","", project$Origin.City)
project=project[order(project$Satisfaction),]

trimws(project$Type.of.Travel)
trimws(project$Class)
trimws(project$Airline.Code)
trimws(project$Airline.Name)
trimws(project$Origin.City)
trimws(project$Origin.State)
trimws(project$Destination.City)
trimws(project$Destination.State)

check<-which((is.na(project$Flight.time.in.minutes)&project$Flight.cancelled=="No"))
check
project<-project[-check,]
```

Fig 2: Data cleaning(1)

```
unique(project$Satisfaction)
project$Satisfaction[(project$Satisfaction == "2.5")]=3
project$Satisfaction[(project$Satisfaction == "4.00.2.00")|(project$Satisfaction == "4.00.5")
project$Satisfaction=factor(project$Satisfaction, levels = c("1", "2","3","4","5"))
project$Satisfaction[(project$Satisfaction == "4.5")]=5
unique(project$Satisfaction)
num=project$Satisfaction
num
project$Satisfaction=as.numeric(levels(num))[num]
project$Price.Sensitivity <- as.numeric(as.character(project$Price.Sensitivity))
project$Year.of.First.Flight <- as.numeric(as.character(project$Year.of.First.Flight))
project$No.of.Flights.p.a. <- as.numeric(as.character(project$No.of.Flights.p.a.))
project$Percentage.of.Flight.with.other.Airlines <- as.numeric(as.character(project$Percentag
project$No.of.other.Loyalty.Cards <- as.numeric(as.character(project$No.of.other.Loyalty.Card
project$Shopping.Amount.at.Airport <- as.numeric(as.character(project$Shopping.Amount.at.Airp
project$Eating.and.Drinking.at.Airport <- as.numeric(as.character(project$Eating.and.Drinking
project$Day.of.Month <- as.numeric(as.character(project$Day.of.Month))
project$Scheduled.Departure.Hour <- as.numeric(as.character(project$Scheduled.Departure.Hour)
project$Departure.Delay.in.Minutes <- as.numeric(as.character(project$Departure.Delay.in.Minu
project$Arrival.Delay.in.Minutes <- as.numeric(as.character(project$Arrival.Delay.in.Minutes)
project$Flight.time.in.minutes <- as.numeric(as.character(project$Flight.time.in.minutes))
project$Flight.Distance <- as.numeric(as.character(project$Flight.Distance))
```

Fig 3: Data Cleaning(2)

```
project=subset(project, Flight.cancelled=="No")


project$Arrival.Delay.in.Minutes[is.na(project$Arrival.Delay.in.Minutes)]=round(median(projec
project$Flight.time.in.minutes[is.na(project$Flight.time.in.minutes)]=round(median(project$Fl

dim(project)
sum(is.na(project))

na<-which(!(project$Satisfaction %in% c(1,2,3,4,5)))
project<-project[-na,]

View(project)
str(project)
summary(project)
```

Fig 4: Data Cleaning(3)

Fig 2, 3 and 4 show the code where we converted the date into one specific format and removed all the white spaces present before all the character values in the dataset. Moreover, ordering of the dataset was also done based on the satisfaction attribute from low(1) to high(5). The satisfaction levels were also brought in the range from 1 to 5 by getting rid to complicated values(4.00.2.00) and (4.00.5) and the factor levels which were numbers were converted to numeric values. The Flight.cancelled column was used to subset only that data in which the flights took of and finally, NA's in flight time in minutes and arrival delay in minutes were replaced by the median of that particular column.

7.Data Munging:

```
new=replicate(length(project$Satisfaction), "nil")
new[project$Satisfaction >= 4]="Happy"
new[project$Satisfaction < 4]="notHappy"
project$happycustomers=new
project$happycustomers

View(project)
str(project)
summary(project)
sum(is.na(project))
dim(project)
```

Fig 5: Adding a new column for happy and unhappy customers

Here, a new column is added based on the satisfaction values and this column specifies whether a customer is happy or unhappy based on the value range.

```
stddev_project=sapply(project, sd)
stddev_project
original=data.frame(stddev_project)
colnames(original)=c("original")
View(original)
dim(original)

set.seed(1)
index=sample(1:dim(project)[1], 60000 )
project_samp=project[index, ]
summary(project_samp)
str(project_samp)
View(project_samp)
dim(project_samp)
dim(project)

stddev_project_samp=sapply(project_samp, sd)
stddev_project_samp
sample=data.frame(stddev_project_samp)
colnames(sample)=c("sample")
View(sample)
dim(sample)

std=cbind(sample, original)
View(std)
```

Fig 6: Sampling the original dataset

The standard deviation of the original dataset was found out and the dataset was sampled to obtain 60,000 values. Finally, the standard deviation of the new sampled dataset was also calculated and the two deviations were compared. Because these two values were somewhat similar for each attribute in both, the new sampled and the original dataset, the new sampled dataset could be considered as a representative of the entire population.

8.Data Visualization:

1.Customer Satisfaction Based on Airline

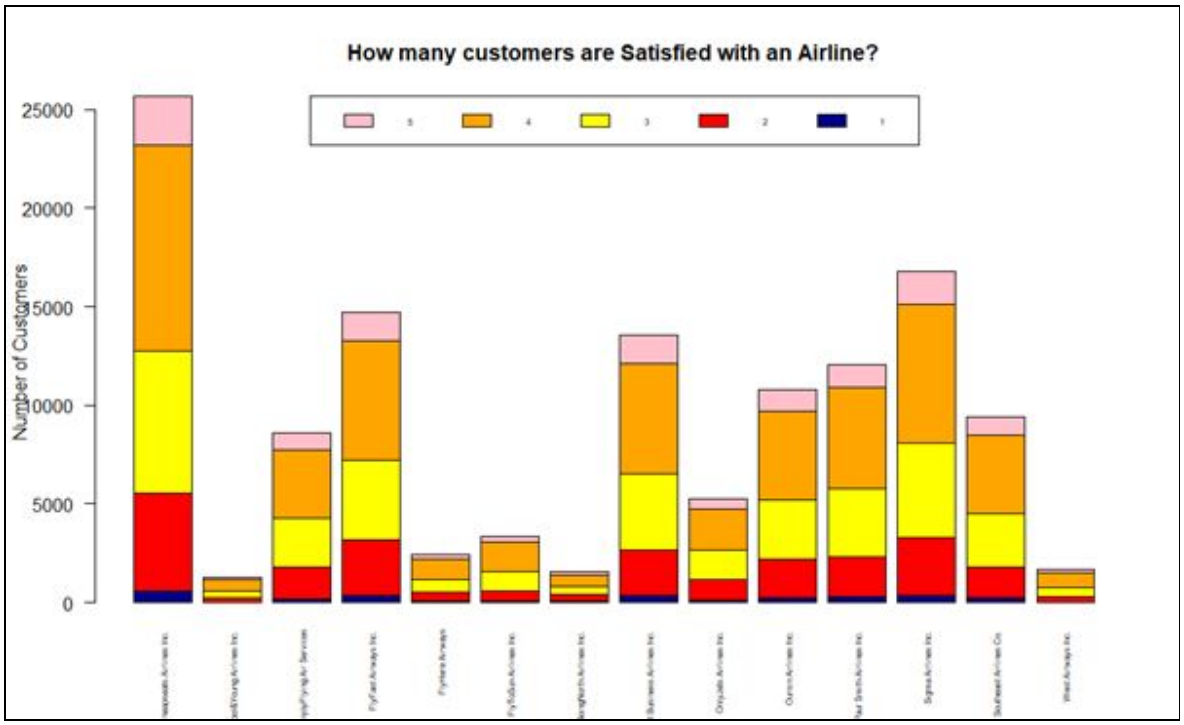


Fig 7

The above figure shows the amount of customers using a specific airline and also provides us with the level of satisfaction for each airline. From the graph, it can be seen that Cheapseats Airline Inc. has a larger base of customers and about a half of them have a high satisfaction(i.e from 4-5).

2. Price Sensitivity with respect to Airline

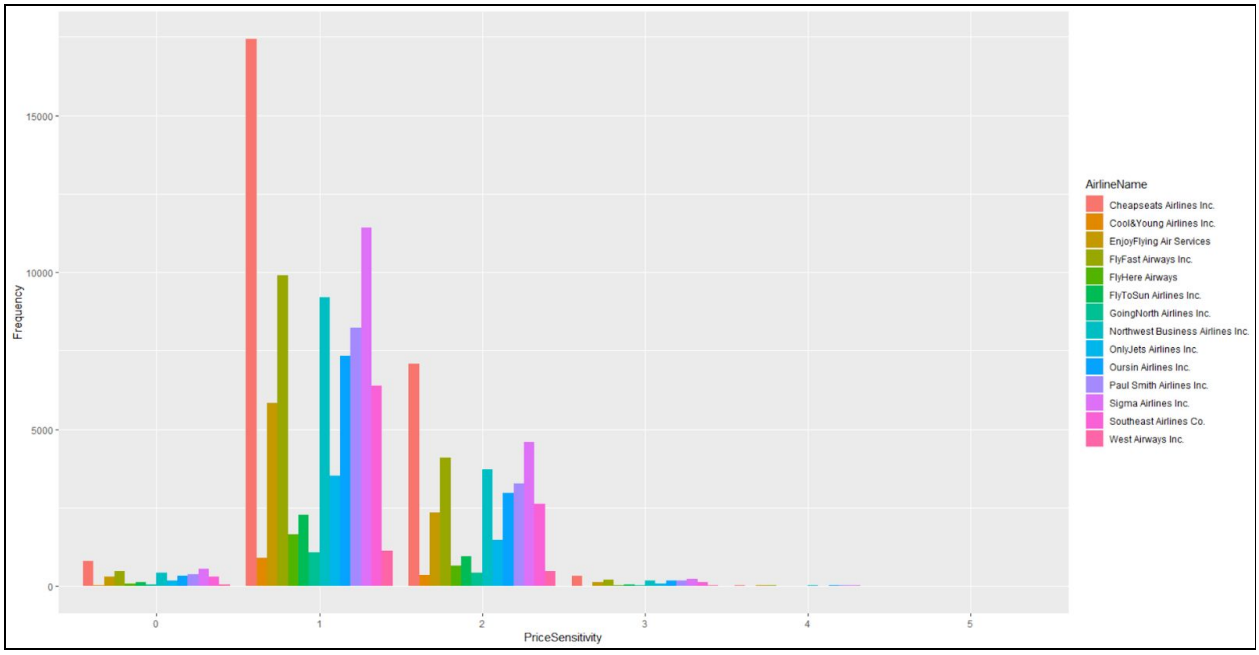


Fig 8

The above graph shows that most people prefer airlines that have a low sensitivity to pricing and here too, Cheapseats Airlines Inc. have a majority of people travelling through them.

3. Satisfaction vs. Age

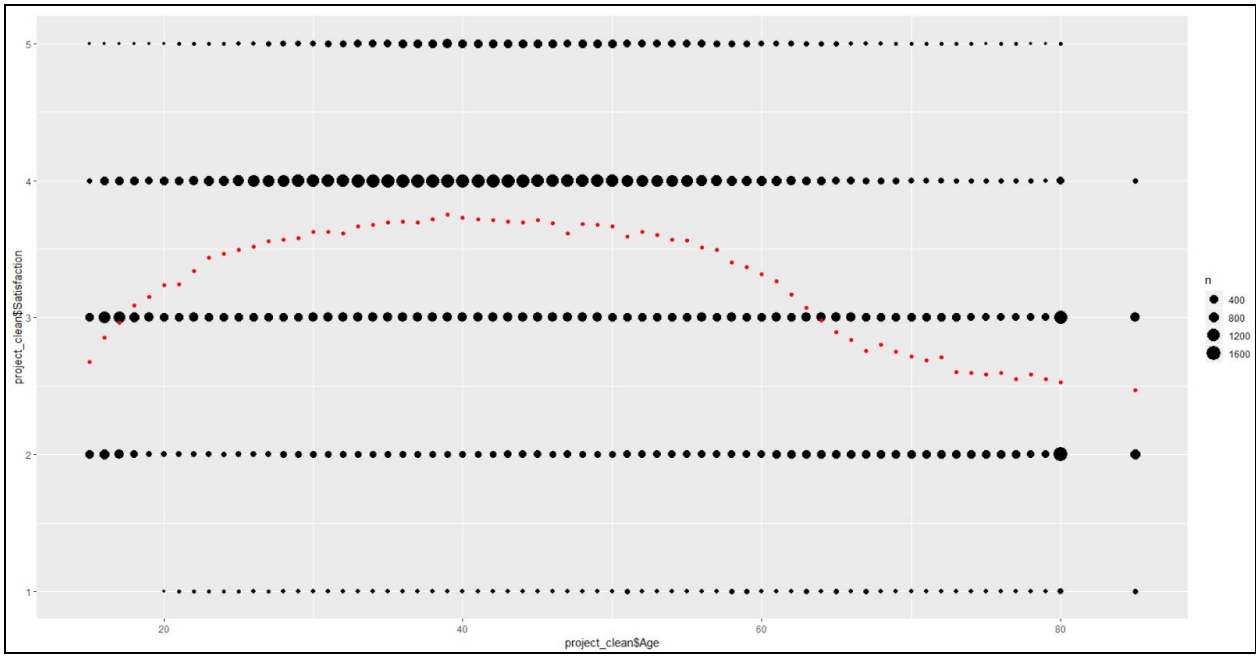


Fig 9

There are a majority of people having a satisfaction in the range of 3-4 and with the help of a trend line, it is quite clear that people in the age group of 20-60 provide a high satisfaction.

4.Class of Travel with respect to type of travel

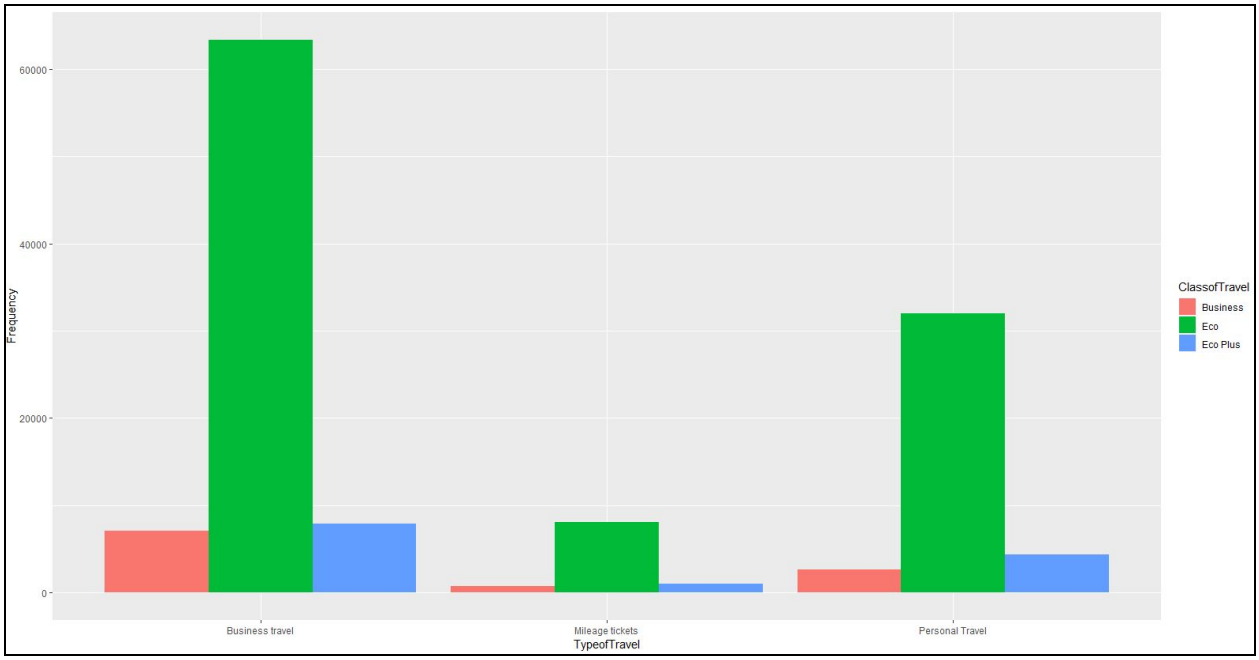


Fig 10

The above graph shows us that most of the people travelling for business purposes consider travelling through economy class.

5.Satisfaction vs. gender

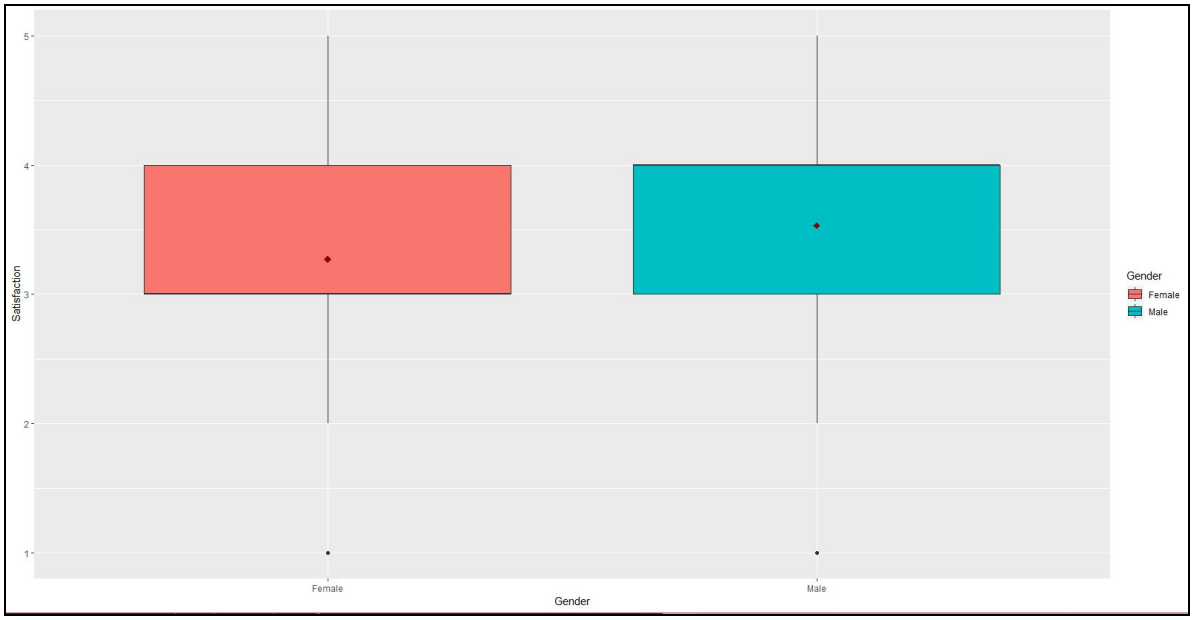


Fig 11

The above box plot helps in determining that male travellers have a high tendency to rate the travel well and thus have a high satisfaction.

6.Satisfaction vs. Type of travel

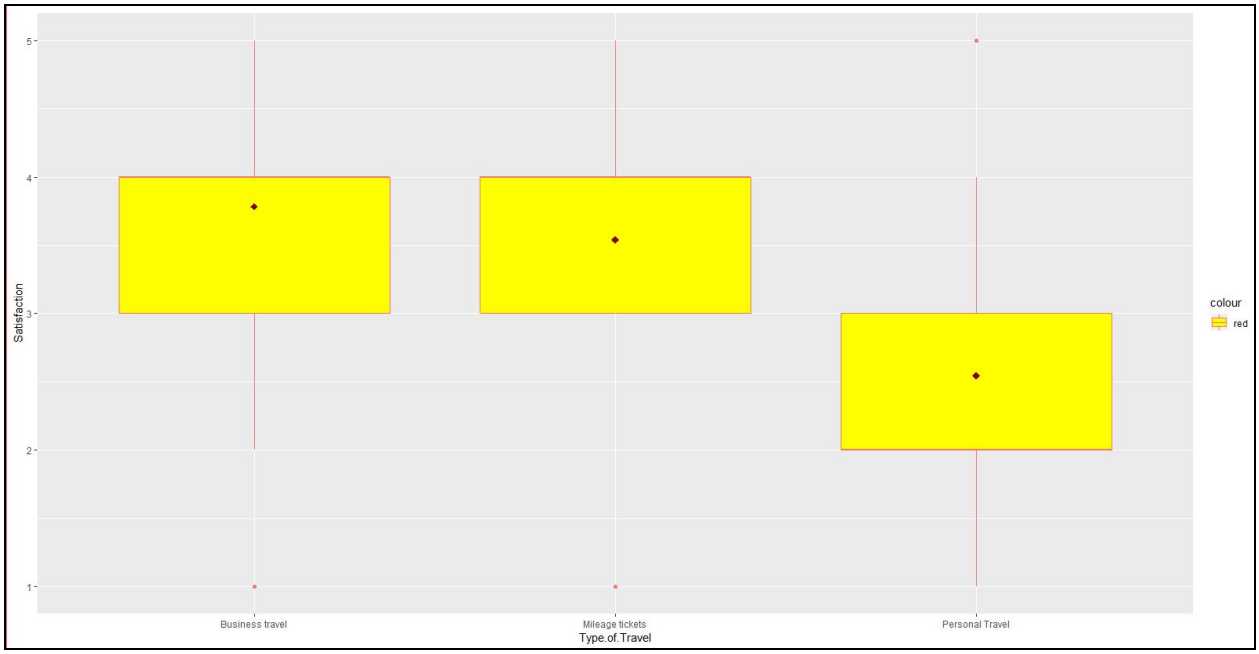


Fig 12

The type of travel also affects the satisfaction such that people travelling for business and using mileage tickets have a higher average satisfaction than people going for personal travel.

7.Satisfaction with respect to airline status

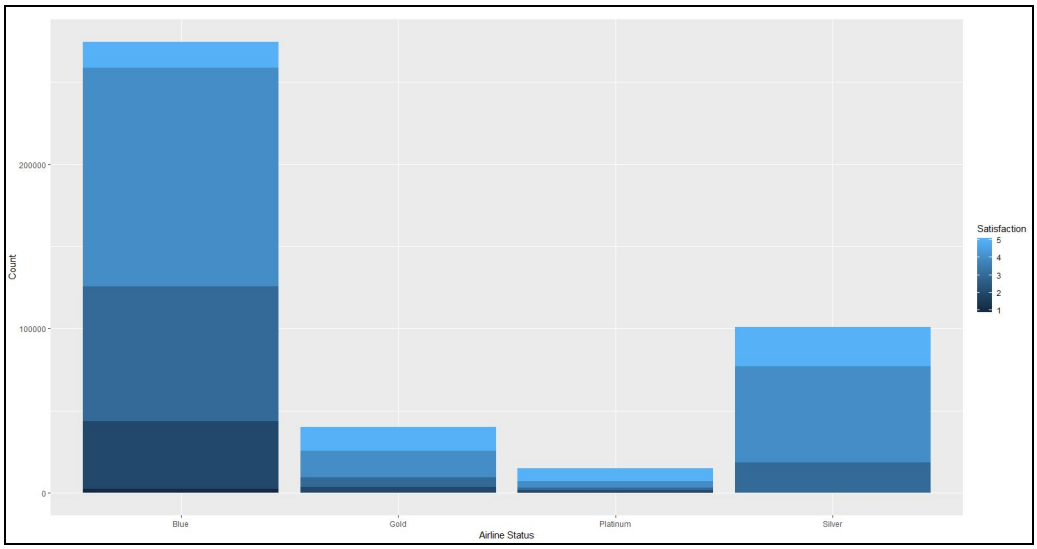


Fig 13

The above bar plot shows that most people consider airlines of blue status and the satisfaction also ranges from 4-5 from them. However, even though there are less customers towards other status of airlines, the satisfaction is relatively high for them as well.

8.Satisfaction with respect to price sensitivity

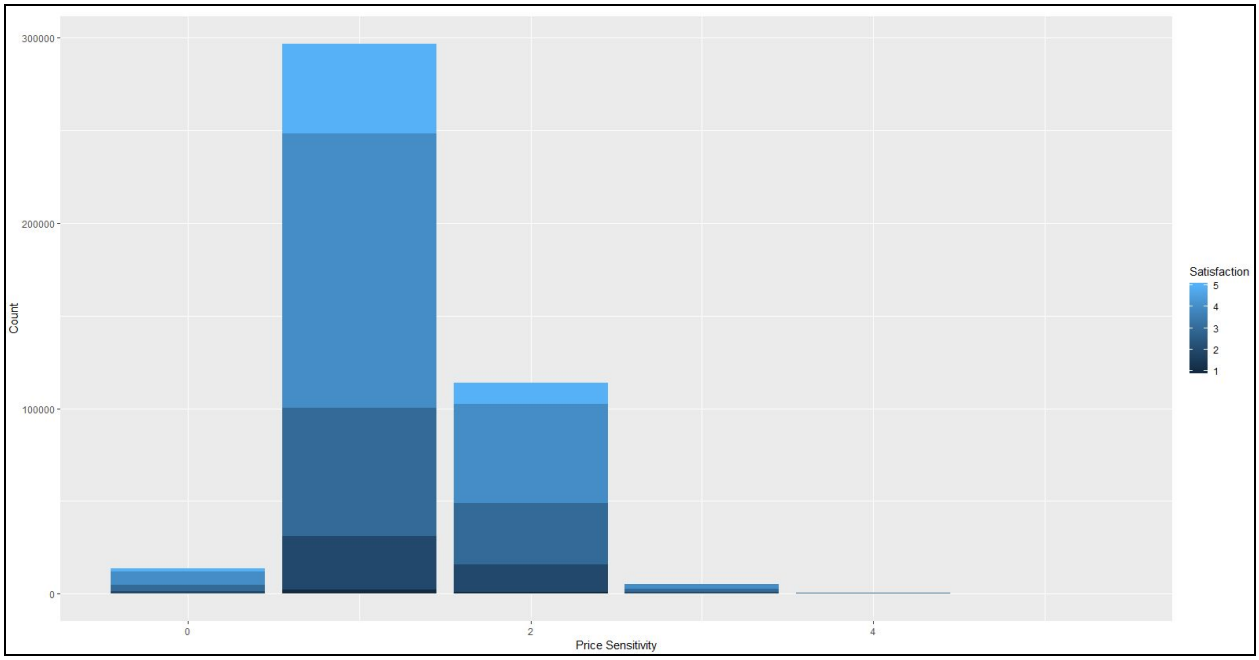


Fig 14

As mentioned above, there are a majority of people who consider a low price sensitivity to be apt for travel.

9.Data Modeling

This step is extremely important to learn about the variation in data and also how it behaves with respect to specific attributes.

1. Linear Model:

For our linear modeling our first model was to use all of the variables in efforts to narrow down what variables might be the most significant in predicting what influences a high customer satisfaction. We used every variable expect the flight cancelled variable. The reason that we did not use that variable is because it only had one level of factor. Linear modeling needs at least two levels for it to work properly. The satisfaction variable is derived from multiple variables, it is our dependant variable in our linear model analysis. The key measurements we will pay attention to in this analysis will be the Residual Standard Error, Multiple R-Squared value, and the p-value statistic based on the level of significance.

The code for linear model on the entire dataset was:

```
sat_orig=lm(formula = Satisfaction ~ Airline.Status + Gender + Age + Price.Sensitivity +
Year.of.First.Flight +No.of.Flights.p.a.+ Percentage.of.Flight.with.other.Airlines
        +Type.of.Travel
        +No.of.other.Loyalty.Cards
```

```
        +Shopping.Amount.at.Airport
        +Eating.and.Drinking.at.Airport
        +Class
        +Day.of.Month
        +Flight.date
        +Airline.Code
        +Airline.Name
        +Origin.City
        +Origin.State
        +Destination.City
        +Destination.State
        +Scheduled.Departure.Hour
        +Departure.Delay.in.Minutes
        +Arrival.Delay.in.Minutes
        +Flight.time.in.minutes
        +Flight.Distance
        +Arrival.Delay.greater.5.Mins
        , data= project)
summary(sat_orig)
```

```
Call:
lm(formula = Satisfaction ~ Airline.Status + Gender + Age + Price.Sensitivity +
    Year.of.First.Flight + No.of.Flights.p.a. + X..of.Flight.with.other.Airlines +
    Type.of.Travel + No..of.other.Loyalty.Cards + Shopping.Amount.at.Airport +
    Eating.and.Drinking.at.Airport + Class + Day.of.Month + Flight.date +
    Airline.Code + Airline.Name + Orgin.City + Origin.State +
    Scheduled.Departure.Hour + Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes +
    Flight.time.in.minutes + Flight.Distance + Arrival.Delay.greater.5.Mins,
    data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1740 -0.4134  0.0758  0.4709  3.0077

Coefficients: (65 not defined because of singularities)
                                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                     -5.807e+00  1.378e+00   -4.215 2.49e-05 ***
Airline.StatusGold                               4.420e-01  7.501e-03   58.923  < 2e-16 ***
Airline.StatusPlatinum                           2.656e-01  1.166e-02   22.776  < 2e-16 ***
Airline.StatusSilver                             6.200e-01  5.233e-03  118.472  < 2e-16 ***
GenderMale                                       1.321e-01  4.230e-03   31.241  < 2e-16 ***
Age                                             -2.329e-03  1.415e-04  -16.460  < 2e-16 ***
Price.Sensitivity                               -4.083e-02  3.768e-03  -10.835  < 2e-16 ***
Year.of.First.Flight                             4.897e-03  6.801e-04    7.201 6.02e-13 ***
No.of.Flights.p.a.                              -3.305e-03  1.555e-04  -21.248  < 2e-16 ***
X..of.Flight.with.other.Airlines                -5.661e-05  2.607e-04   -0.217 0.828082
Type.of.TravelMileage tickets                   -1.471e-01  7.799e-03  -18.864  < 2e-16 ***
Type.of.TravelPersonal Travel                   -1.077e+00  5.008e-03 -214.954  < 2e-16 ***
No..of.other.Loyalty.Cards                      -2.527e-03  2.148e-03   -1.177 0.239338
Shopping.Amount.at.Airport                       1.662e-04  3.839e-05    4.330 1.49e-05 ***
Eating.and.Drinking.at.Airport                  -8.935e-05  3.968e-05   -2.252 0.024339 *
ClassEco                                        -7.756e-02  7.396e-03  -10.486  < 2e-16 ***
ClassEco Plus                                   -7.059e-02  9.506e-03   -7.427 1.12e-13 ***
Day.of.Month                                     1.989e-04  8.966e-04    0.222 0.824421
```

Executing the summary function for this model gave us the above result and also meaningful insight as shown below :

Residual standard error: 0.719 on 126434 degrees of freedom
Multiple R-squared: 0.4498, Adjusted R-squared: 0.4467
F-statistic: 144.9 on 713 and 126434 DF, p-value: < 0.00000000000000022

From this we can see the most significant variables. The list of all variables the analysis was performed on is much too large to copy the whole things, so we selected the top few to show. From this we can analyze the most significant variables separately to see what variables might again be most significant on an individual scale. The variables that we select will have a p-value based on a 0.001 level of significance. Our model currently is very large therefore having a multiple R-Squared value of 0.4485, which is not great at predicting the customer satisfaction. It has all of the variables and that is why the R-squared value is so low. The model is still statistically significant at the 0.001 level of significance because the p-value is less than 0.001

Finding out the statistically significant variables helped us to determine relation between the satisfaction of customers and also how it varies with respect to the attributes.
Thus, considering each of those significant variables one at a time and plotting them against satisfaction, a measure of the most significant variable.

Airline Status vs. Satisfaction:

```
Call:
lm(formula = Satisfaction ~ Airline.Status, data = project)

Residuals:
     Min       1Q   Median       3Q      Max
-2.65999 -0.94502  0.05498  0.83861  1.83861

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.161388   0.003079 1026.74   <2e-16 ***
Airline.StatusGold    0.590088   0.009308   63.39   <2e-16 ***
Airline.StatusPlatinum 0.498598   0.014517   34.34   <2e-16 ***
Airline.StatusSilver  0.783631   0.006460  121.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9074 on 127147 degrees of freedom
Multiple R-squared:  0.1186,    Adjusted R-squared:  0.1186
F-statistic:  5705 on 3 and 127147 DF,  p-value: < 2.2e-16
```

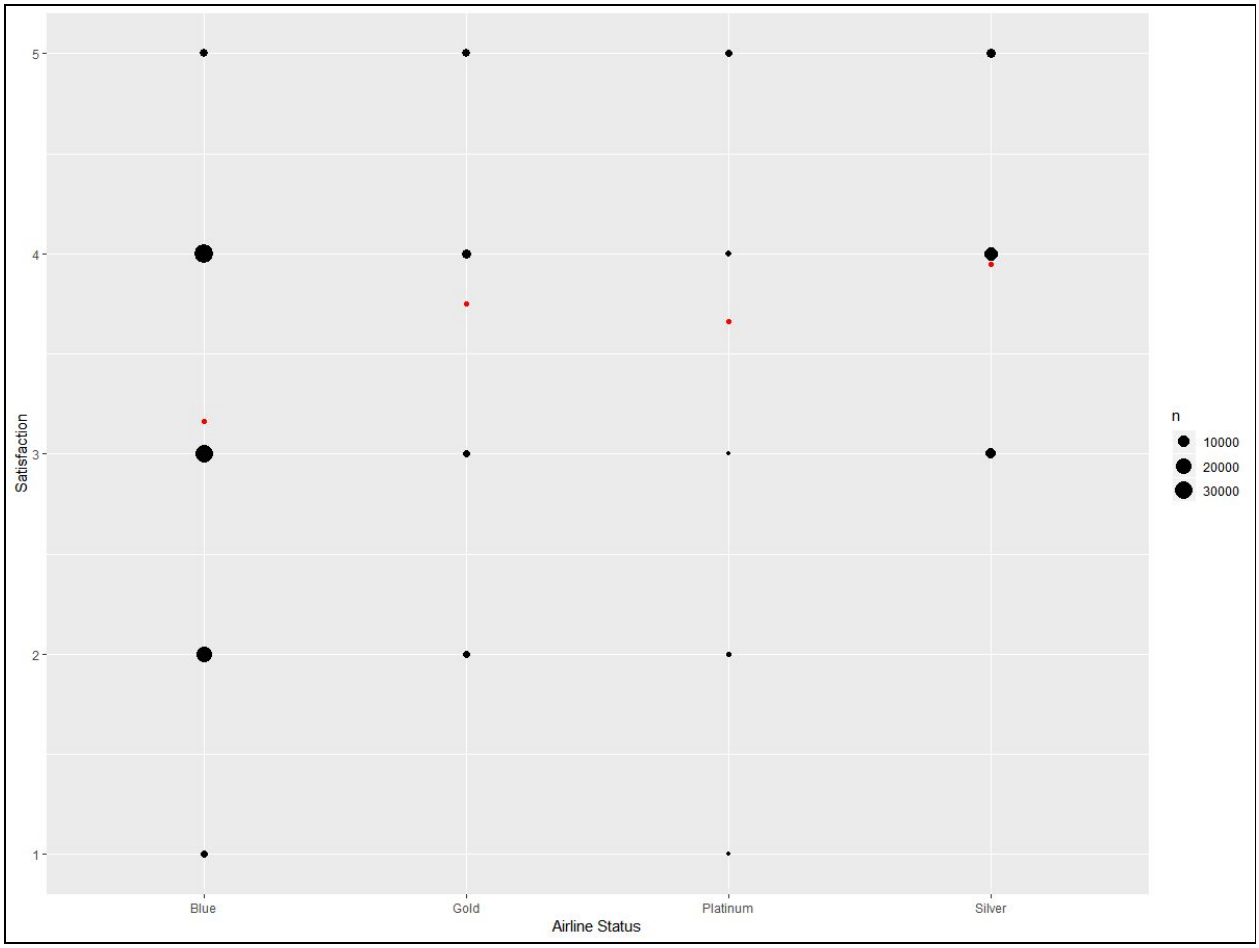


Fig 15

Gender vs. Satisfaction:

```
Call:
lm(formula = Satisfaction ~ Gender, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5316 -0.5316  0.4684  0.7296  1.7296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.270374   0.003577  914.16   <2e-16 ***
GenderMale  0.261241   0.005417   48.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9578 on 127149 degrees of freedom
Multiple R-squared:  0.01797,   Adjusted R-squared:  0.01796
F-statistic:  2326 on 1 and 127149 DF,  p-value: < 2.2e-16
```

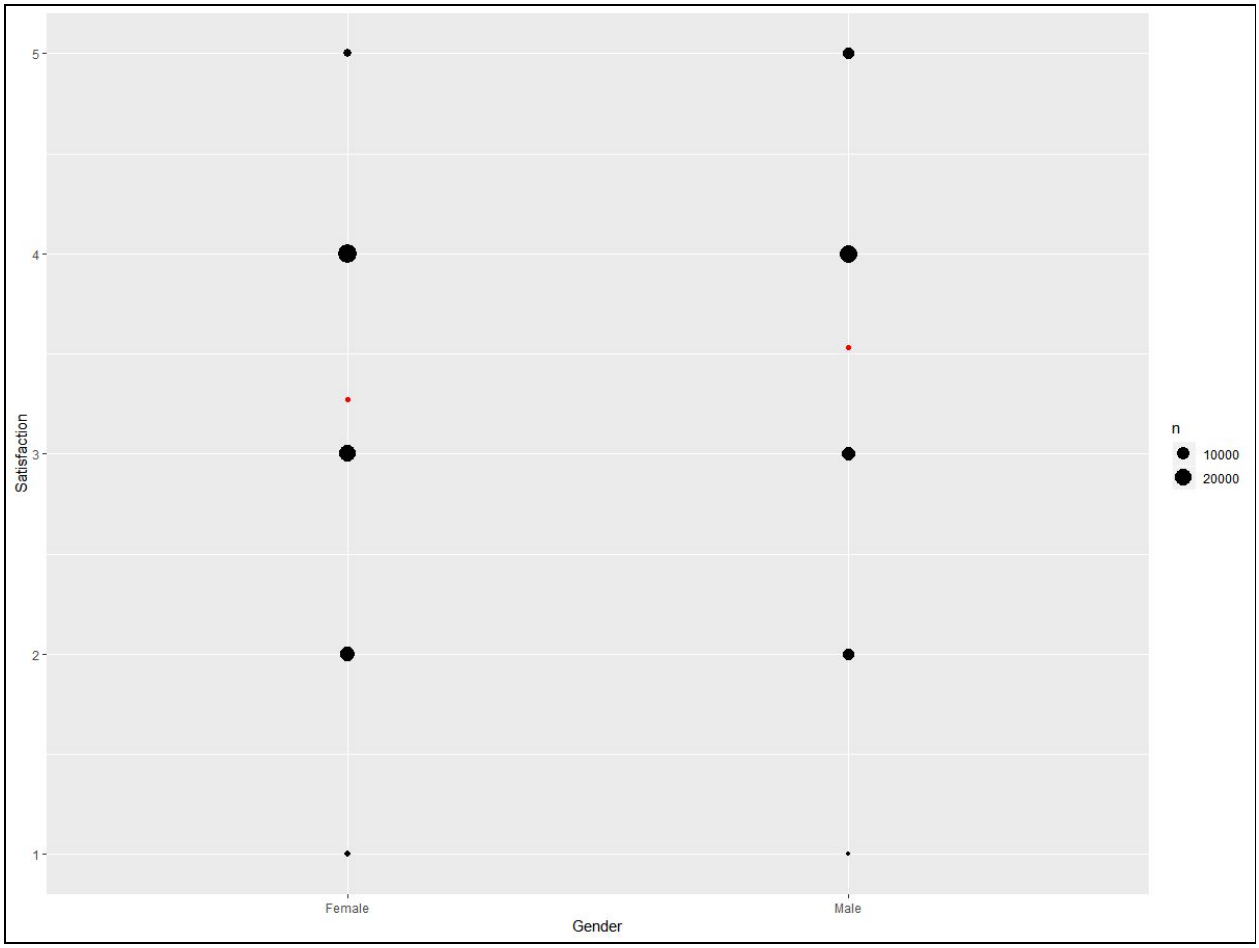


Fig 16

Age vs. Satisfaction:

```
Call:
lm(formula = Satisfaction ~ Age, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7095 -0.6722  0.2781  0.6261  2.0363

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.9581183  0.0075459  524.54   <2e-16 ***
Age         -0.0124299  0.0001531  -81.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9424 on 127149 degrees of freedom
Multiple R-squared:  0.04928,   Adjusted R-squared:  0.04927
F-statistic:  6591 on 1 and 127149 DF,  p-value: < 2.2e-16
```

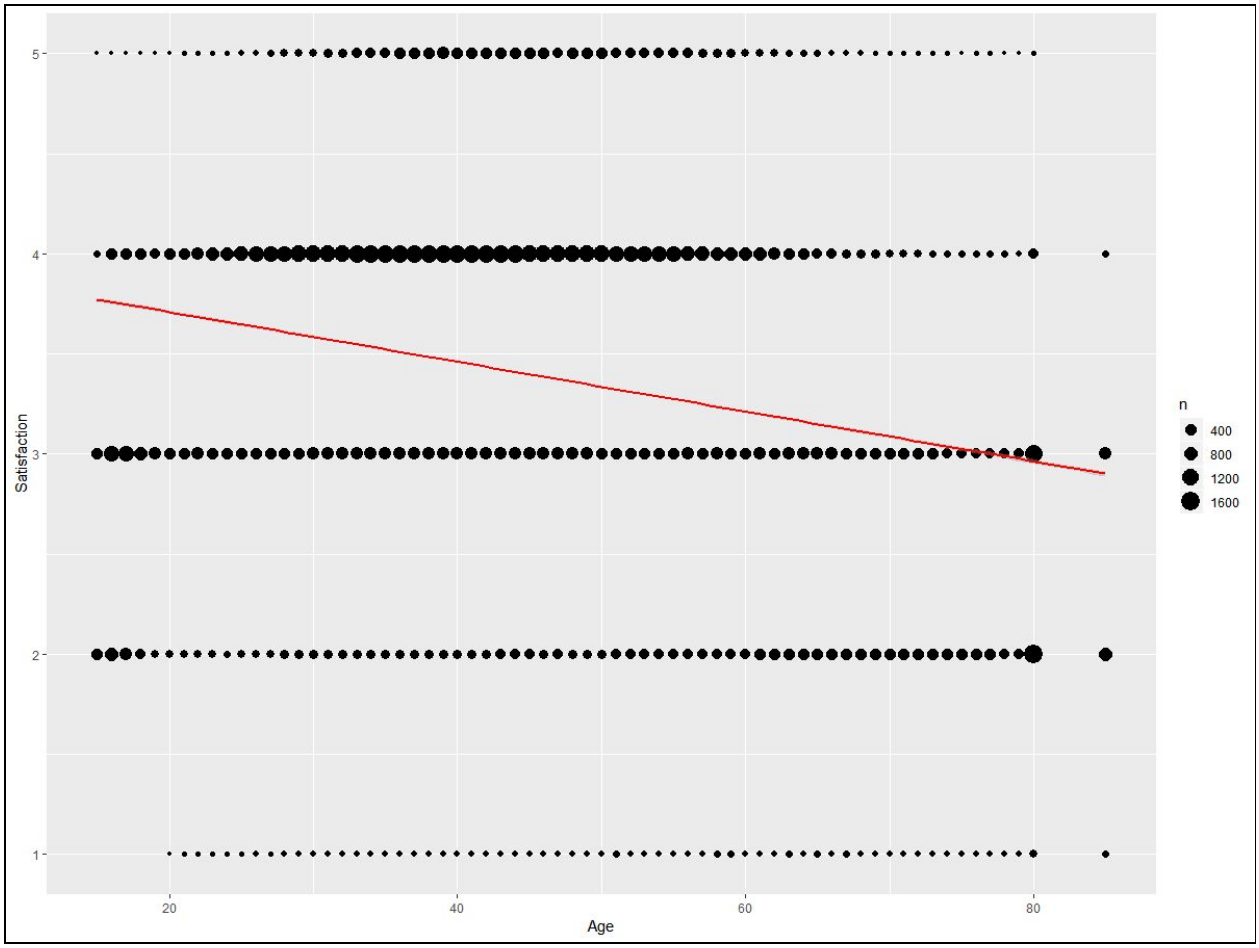


Fig 17

Price Sensitivity vs. Satisfaction:

```
Call:
lm(formula = Satisfaction ~ Price.Sensitivity, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5835 -0.4273  0.4165  0.5727  2.0412

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.583454   0.006857  522.61   <2e-16 ***
Price.Sensitivity -0.156177   0.004943  -31.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9628 on 127149 degrees of freedom
Multiple R-squared:  0.007789,  Adjusted R-squared:  0.007781
F-statistic: 998.1 on 1 and 127149 DF,  p-value: < 2.2e-16
```

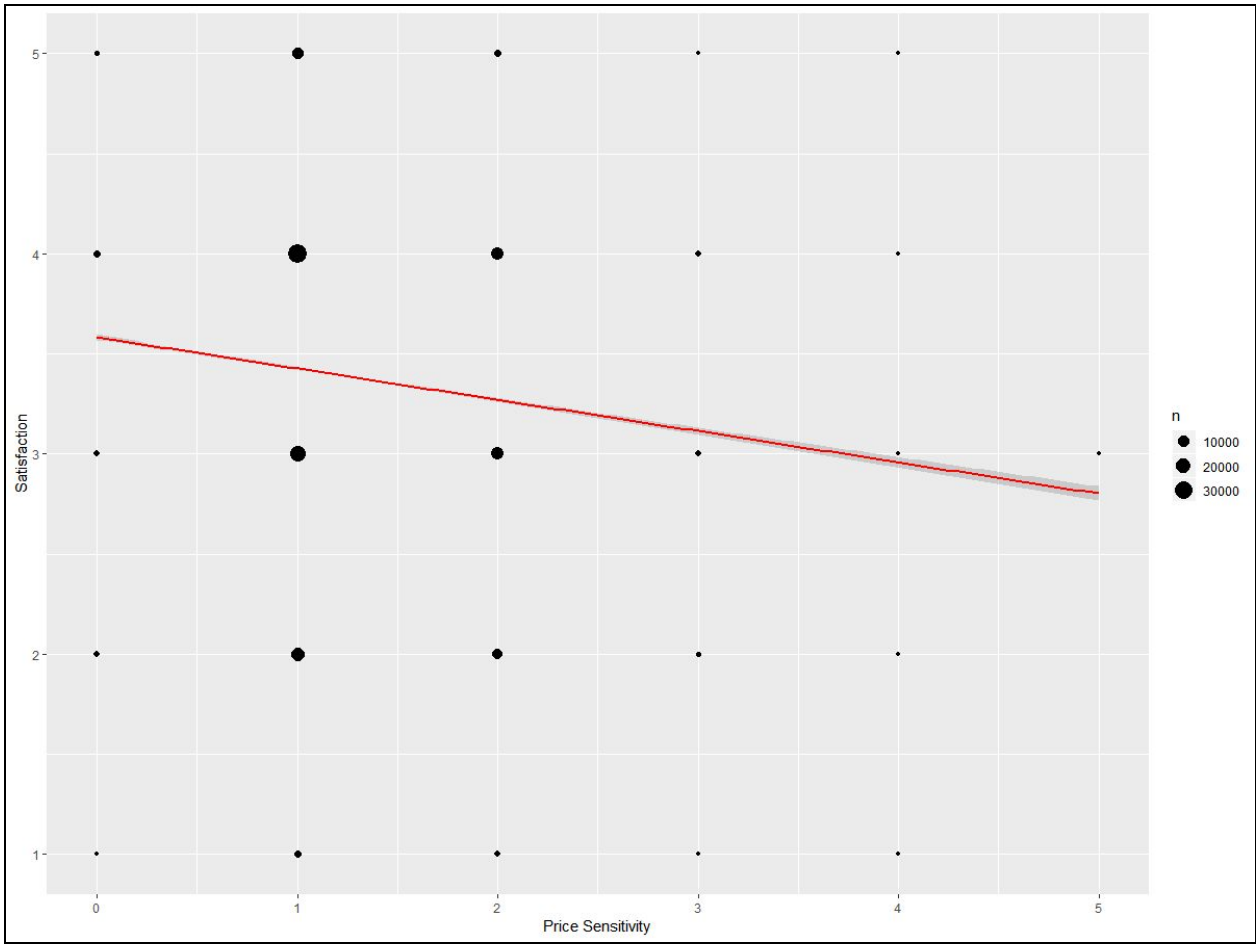


Fig 18

Year of First Flight vs. Satisfaction

```
Call:
lm(formula = Satisfaction ~ Year.of.First.Flight, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3965 -0.3940  0.6035  0.6188  1.6264

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.7358715  1.8273939  -0.950  0.34216
Year.of.First.Flight  0.0025509  0.0009104   2.802  0.00508 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9665 on 127149 degrees of freedom
Multiple R-squared:  6.174e-05, Adjusted R-squared:  5.388e-05
F-statistic: 7.851 on 1 and 127149 DF,  p-value: 0.005081
```

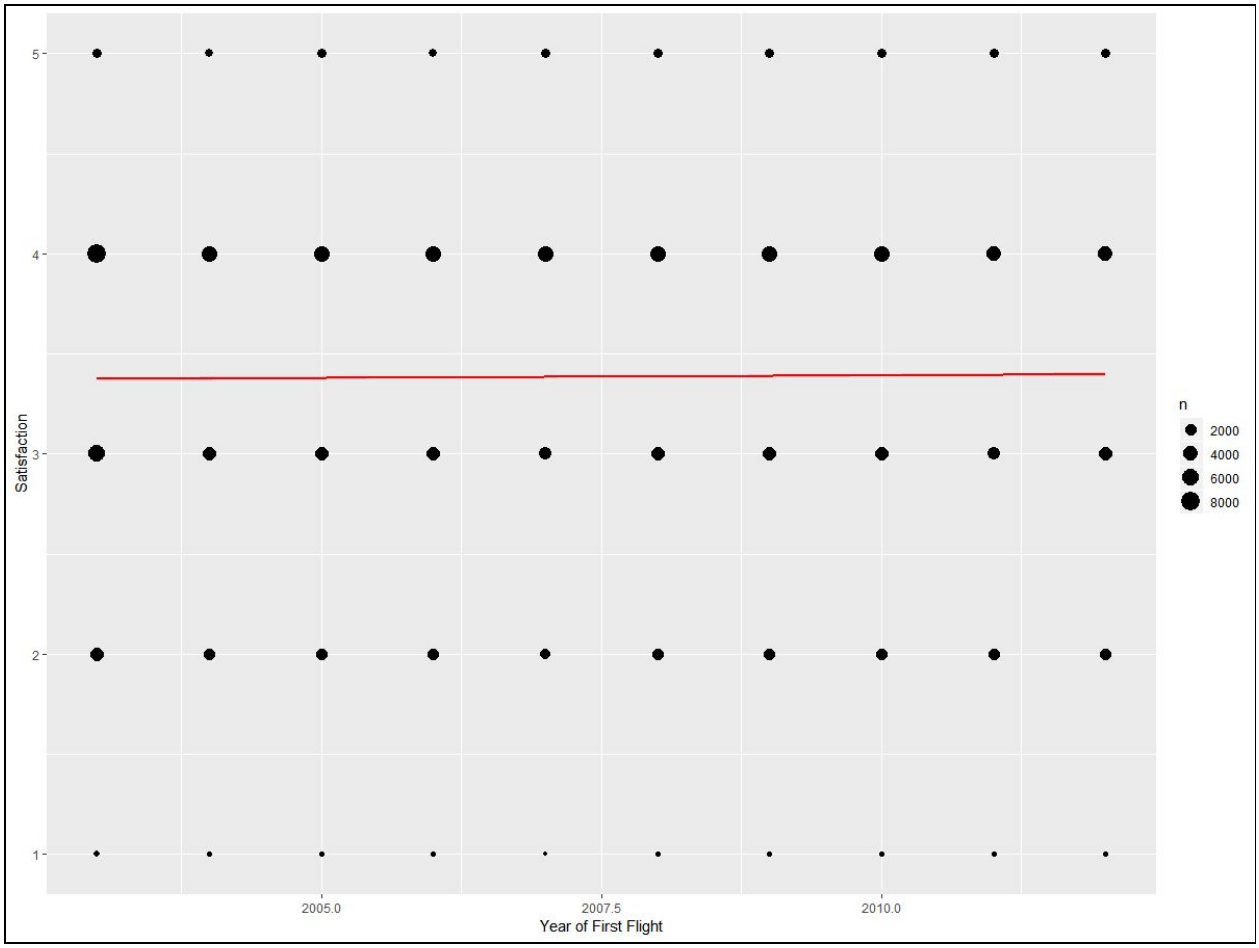


Fig 19

Number of Flights per Airline vs. Satisfaction

```
Call:
lm(formula = Satisfaction ~ No.of.Flights.p.a., data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7061 -0.5937  0.2939  0.5990  2.8356

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.7061121  0.0045170  820.49   <2e-16 ***
No.of.Flights.p.a. -0.0160595  0.0001832  -87.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9386 on 127149 degrees of freedom
Multiple R-squared:  0.05699,   Adjusted R-squared:  0.05699
F-statistic:  7685 on 1 and 127149 DF,  p-value: < 2.2e-16
```

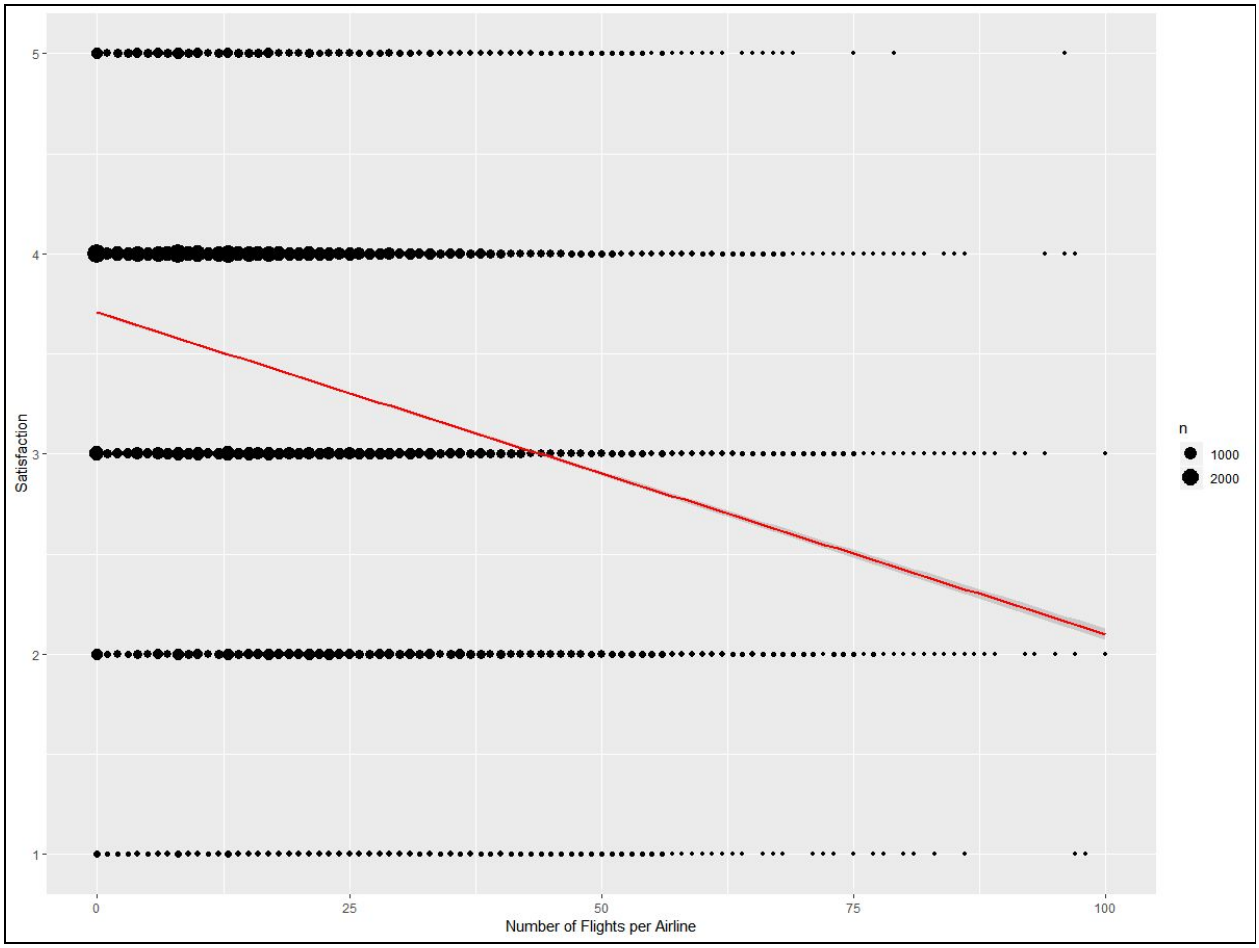


Fig 20

Type of Travel vs. Satisfaction

```
Call:
lm(formula = Satisfaction ~ Type.of.Travel, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7816 -0.5453  0.2184  0.4547  2.4548

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     3.781642   0.002811 1345.31   <2e-16 ***
Type.of.TravelMileage tickets -0.240642   0.008426  -28.56   <2e-16 ***
Type.of.TravelPersonal Travel -1.236390   0.004879 -253.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.787 on 127148 degrees of freedom
Multiple R-squared:  0.337,     Adjusted R-squared:  0.337
F-statistic: 3.232e+04 on 2 and 127148 DF,  p-value: < 2.2e-16
```

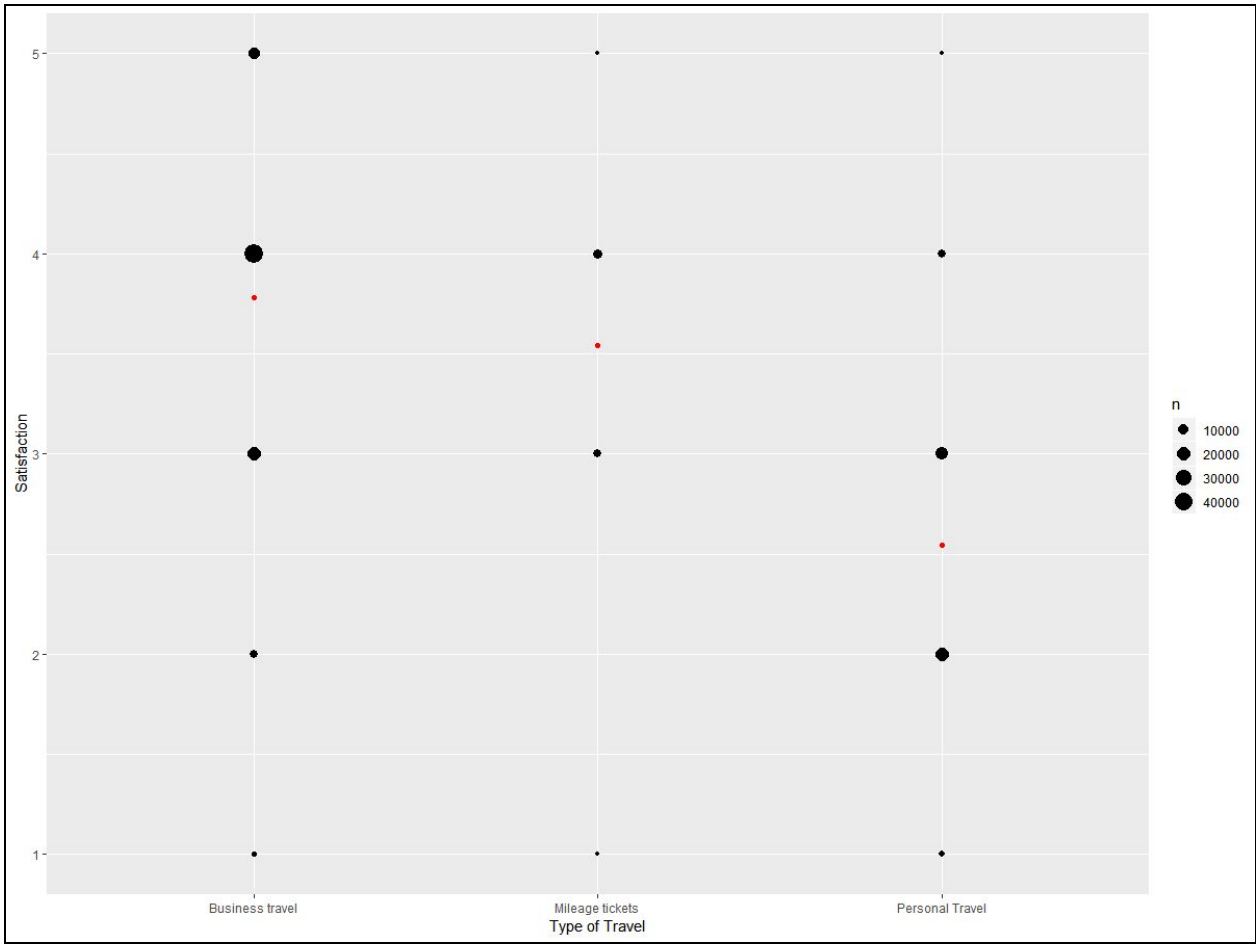


Fig 21

Shopping Amount at Airport vs. Satisfaction

```
Call:
lm(formula = Satisfaction ~ Shopping.Amount.at.Airport, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5047 -0.3950  0.5700  0.6241  1.6241

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               3.376e+00  3.031e-03 1113.908  < 2e-16 ***
Shopping.Amount.at.Airport 3.182e-04  5.106e-05    6.232 4.63e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9664 on 127149 degrees of freedom
Multiple R-squared:  0.0003053, Adjusted R-squared:  0.0002975
F-statistic: 38.84 on 1 and 127149 DF,  p-value: 4.625e-10
```

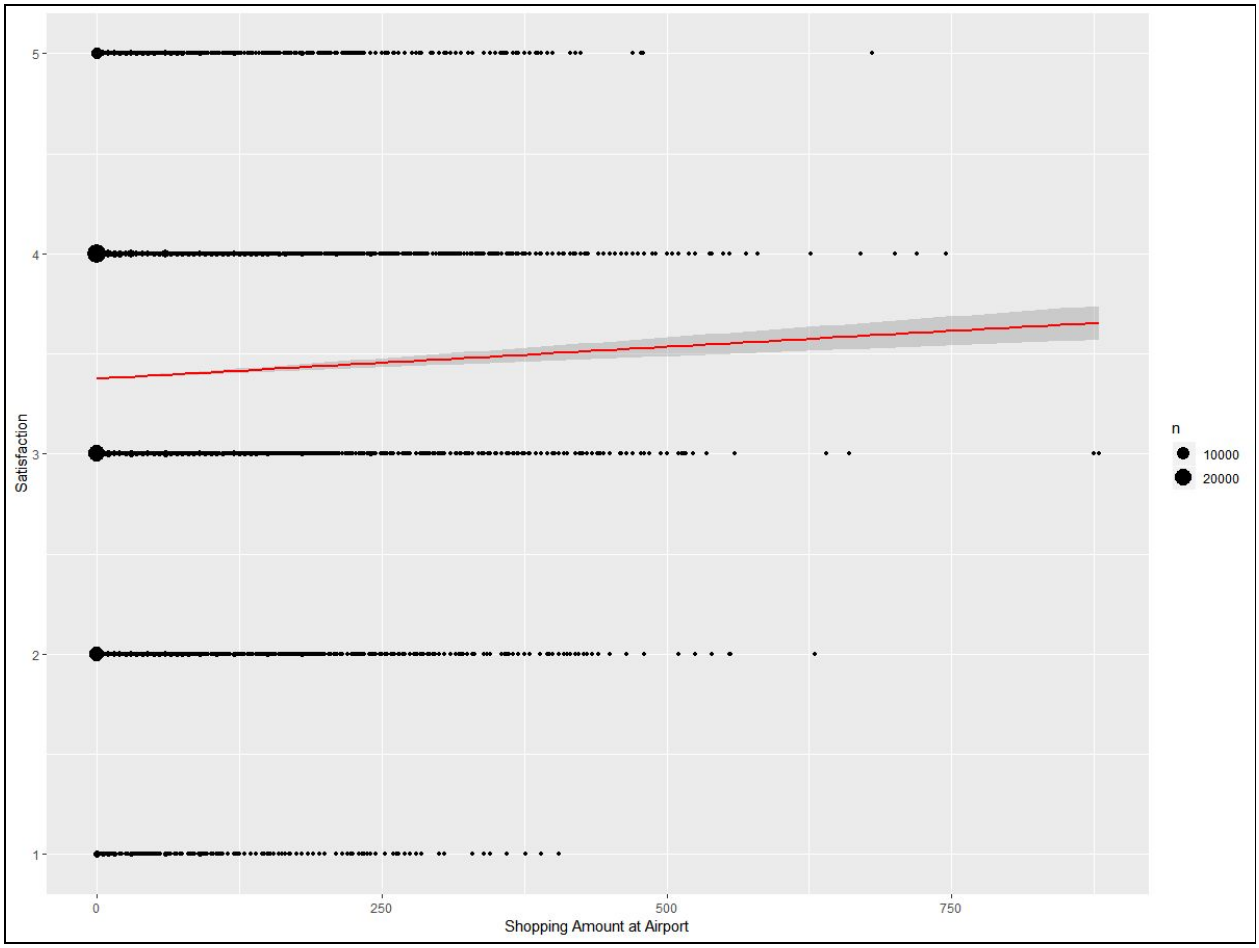

5
4
3
2
1
Satisfaction
0
250
500
750
Shopping Amount at Airport
n
10000
20000

Fig 22

Class vs. Satisfaction

```
Call:
lm(formula = Satisfaction ~ Class, data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5352 -0.3771  0.4648  0.6229  1.6781

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.535202   0.009442  374.42   <2e-16 ***
ClassEco      -0.158092   0.009908  -15.96   <2e-16 ***
ClassEco Plus -0.213258   0.012615  -16.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9654 on 127148 degrees of freedom
Multiple R-squared:  0.002485,  Adjusted R-squared:  0.002469
F-statistic: 158.4 on 2 and 127148 DF,  p-value: < 2.2e-16
```

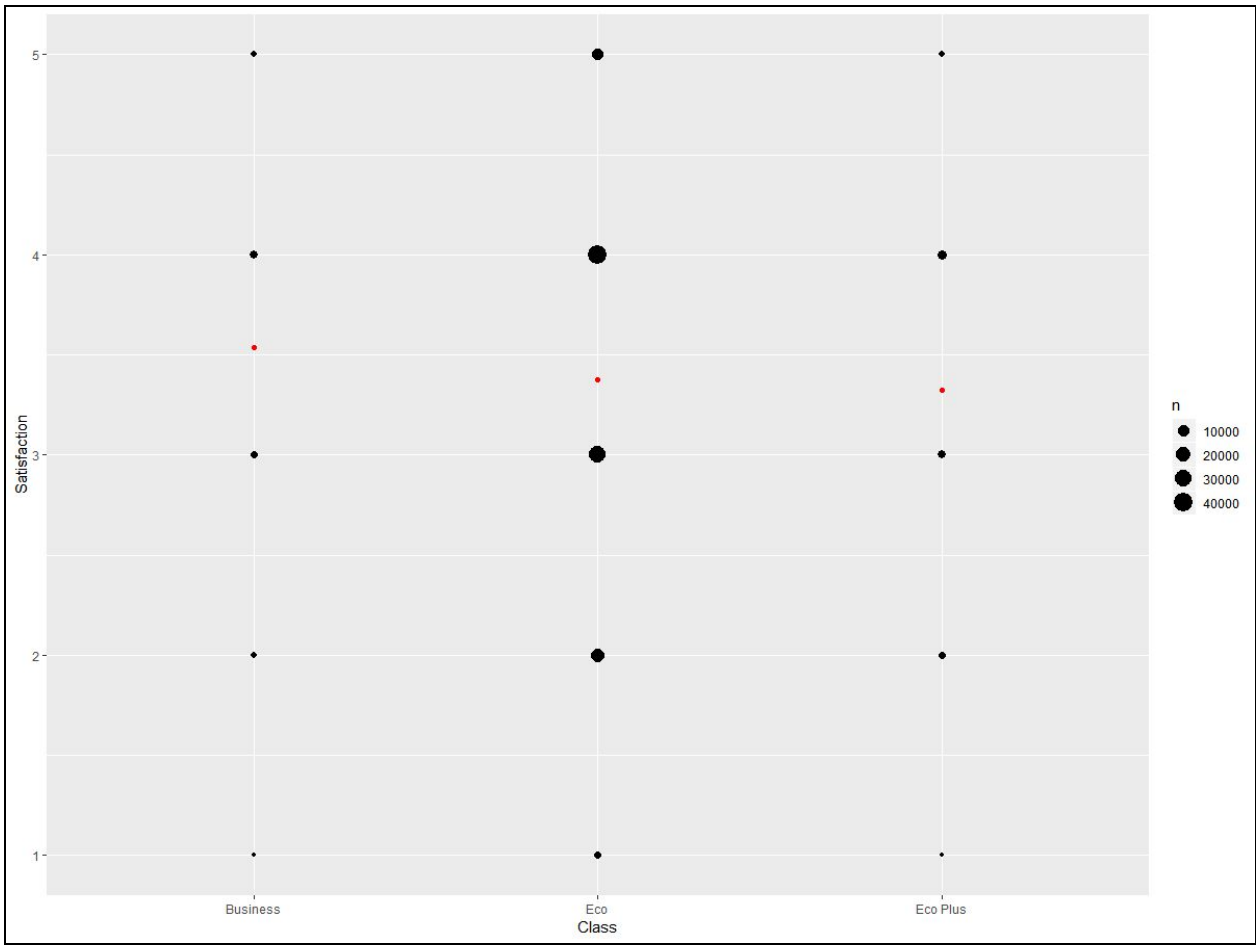


Fig 23

Based on the various linear models we are able to see which variables have the most statistical significance on customer satisfaction. This is important because the significant variables that are found through the first linear model that has all the variables might not be significant in

directly predicting customer satisfaction. As we can see the year of a customers' first flight is no longer significant in predicting customer satisfaction at a 0.001 level of significance. We can see through the analyses that Type of Travel and Airline Status are the key variables to focus on first. Type of travel has a Multiple R-Squared value of 0.337 and a Residual Standard Error of 0.787, Airline Status has a Multiple R-Squared value of 0.1186 and a Residual Standard Error of 0.9074. For an analysis on single variables against satisfaction these variables perform the best at predicting satisfaction. The visualizations help us see that people on business travel tend to have a higher satisfaction so we suggest that the airline focus on mileage travelers and personal travelers to help give a better experience for them. An airline status ranking is also very important. Blue status results in the lowest satisfaction rating, as a result we suggest either improving your ranking or improving the experience for people taking a flight with that airline ranking.

2.SVM

The support vector machine is used to find the accuracy of the prediction of a particular model. Here, we split the sampled dataset into 2 parts, training and testing and the train data was used on a model and the test data was used to determine it's accuracy of prediction.

Code:

```
library(kernlab)
randIndex=sample(1:dim(project_samp)[1])
randIndex
cutPoint2_3=floor(2 * dim(project_samp)[1]/3)
cutPoint2_3
trainData=project_samp[randIndex[1:cutPoint2_3],]
trainData
View(trainData)
testData=project_samp[randIndex[(cutPoint2_3+1):dim(project_samp)[1]],]
testData
str(testData)
View(testData)
dim(trainData)
dim(testData)
newtrainData=trainData[,c(2,3,4,5,6,7,9,11,13,28)]
newtestData=testData[,c(2,3,4,5,6,7,9,11,13,28)]
str(newtestData)
str(newtrainData)
svmop=ksvm(happycustomers ~., data=newtrainData, kernel = "vanilladot",kpar="automatic", C=5,cross=3, prob.model=TRUE)
svmop
```

Output:

```
Support Vector Machine object of class "ksvm"
SV type: C-svc  (classification)
 parameter : cost C = 5
Linear (vanilla) kernel function.
Number of Support Vectors : 20760
Objective Function Value : -97171.97
Training error : 0.242925
Cross validation error : 0.242925
Probability model included.
```

The code for predicting is as follows:

```
svmPred <- predict(svmop, newtestData, type = "votes")
svmPred
str(svmPred)
head(svmPred)
dim(svmPred)
svmPred[1,]
# Creating a confusion matrix
comTable=data.frame(newtestData[ ,10], svmPred[1, ])
comTable[comTable=="0"]="notHappy"
comTable[comTable=="1"]="Happy"
table(comTable)
#calculating the error rate
t<-table(comTable)
sum(t[1,2]+t[2,1])/sum(t)
```

The final result of prediction could be obtained with the help of a confusion matrix that is as follows:

```
                    svmPred.1...
newtestData...10. Happy notHappy
      Happy         9684     623
      notHappy      4200    5493
```

And the error rate was

0.24115

The above result shows that the model worked with an accuracy of about 76%

3.Associative Rule Mining:

Association rules are created using the apriori function. It is useful in finding variables that often occur together. It is created with a left hand side that are often together with one variable of the right hand side.

We converted all integer variables that we plan to use to categorical variables. We then created a data frame of the key variables to be able to perform an apriori analysis on the new data set. The variables we selected for the data frame are Satisfaction, Airline Status, Gender, Age, Price Sensitivity, Year of FirstFlight, No of Flights per airline, Type of Travel, Shopping Amount at Airport, Class.

Code:

```
change <- function(vec)
{
 vBuckets <- replicate(length(vec), "Average")
 vBuckets[vec > 4] <- "High"
  vBuckets[vec < 4] <- "Low"
  return(vBuckets)
}
Satisfaction <- change(project$Satisfaction)
Price.Sensitivity <- change(project$Price.Sensitivity)
change2=function(v)
{
  q <- quantile(v, c(0.4, 0.6))
  vBuckets <- replicate(length(v), "Average")
  vBuckets[v <= q[1]] <- "Low"
  vBuckets[v > q[2]] <- "High"
  return(vBuckets)
}
Age <- change2(project$Age)
No.of.Flights.p.a <- change2(project$No.of.Flights.p.a.)
Shopping.Amount.at.Airport <- change2(project$Shopping.Amount.at.Airport)
change3=function(v)
{
  q <- quantile(v, c(0.4, 0.6))
  vBuckets <- replicate(length(v), "Average")
  vBuckets[v <= q[1]] <- "Least Recent"
  vBuckets[v > q[2]] <- "Most Recent"
  return(vBuckets)
}
Year.of.First.Flight <- change3(project$Year.of.First.Flight)
ProjectSurvey <- data.frame(Satisfaction, project$Airline.Status, project$Gender, Age,
Price.Sensitivity, Year.of.First.Flight, No.of.Flights.p.a, project$Type.of.Travel,
Shopping.Amount.at.Airport, project$Class)
ProjectSurveyx <- as(ProjectSurvey, "transactions")
```

ruleset <- apriori(ProjectSurveyx, parameter = list(support=0.05, confidence =0.05), appearance = list(default="lhs",rhs=("Satisfaction=High")))

```
> inspect(ruleset)
     lhs                                                rhs                    support confidence      lift count
[1]  {}                                              => {Satisfaction=High} 0.09819821 0.09819821 1.0000000 12486
[2]  {No.of.Flights.p.a=Low}                         => {Satisfaction=High} 0.05622449 0.13573449 1.3822502  7149
[3]  {project.Gender=Male}                           => {Satisfaction=High} 0.06522953 0.14953574 1.5227951  8294
[4]  {Shopping.Amount.at.Airport=Low}                => {Satisfaction=High} 0.05258315 0.09249115 0.9418822  6686
[5]  {project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.09162335 0.14862537 1.5135243 11650
[6]  {project.Class=Eco}                             => {Satisfaction=High} 0.08031396 0.09878024 1.0059271 10212
[7]  {Price.Sensitivity=Low}                         => {Satisfaction=High} 0.09812742 0.09827582 1.0007904 12477
[8]  {No.of.Flights.p.a=Low,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.05239440 0.16798628 1.7106859  6662
[9]  {Price.Sensitivity=Low,
      No.of.Flights.p.a=Low}                         => {Satisfaction=High} 0.05615371 0.13602591 1.3852179  7140
[10] {project.Gender=Male,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.06091969 0.20531170 2.0907888  7746
[11] {project.Gender=Male,
      project.Class=Eco}                             => {Satisfaction=High} 0.05496614 0.14761231 1.5032078  6989
[12] {project.Gender=Male,
      Price.Sensitivity=Low}                         => {Satisfaction=High} 0.06519807 0.14968222 1.5242867  8290
[13] {Price.Sensitivity=Low,
      Shopping.Amount.at.Airport=Low}                => {Satisfaction=High} 0.05252810 0.09260954 0.9430879  6679
[14] {project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.07491880 0.15028792 1.5304549  9526
[15] {Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.09156043 0.14867885 1.5140689 11642
[16] {Price.Sensitivity=Low,
      project.Class=Eco}                             => {Satisfaction=High} 0.08025104 0.09886543 1.0067947 10204
[17] {Price.Sensitivity=Low,
      No.of.Flights.p.a=Low,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.05233148 0.16811096 1.7119555  6654
[18] {project.Gender=Male,
      project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.05137986 0.20411798 2.0786325  6533
[19] {project.Gender=Male,
      Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.06088824 0.20541803 2.0918715  7742
[20] {project.Gender=Male,
      Price.Sensitivity=Low,
      project.Class=Eco}                             => {Satisfaction=High} 0.05493468 0.14776189 1.5047311  6985
[21] {Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.07486374 0.15034827 1.5310694  9519
[22] {project.Gender=Male,
      Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.05134840 0.20421632 2.0796339  6529
```

We were able to generate 22 rules that we decided to take try and find some of the most relevant rules that can result in some actionable insight.

We have decided to the rules that have a lift of greater than 2 in efforts to find the variables that are help result in high satisfaction the most.

```
> inspect(ruleset_filter)
     lhs                                                rhs                    support confidence     lift count
[1]  {project.Gender=Male,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.06091969  0.2053117 2.090789  7746
[2]  {project.Gender=Male,
      Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel}        => {Satisfaction=High} 0.06088824  0.2054180 2.091871  7742
[3]  {project.Gender=Male,
      project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.05137986  0.2041180 2.078632  6533
[4]  {project.Gender=Male,
      Price.Sensitivity=Low,
      project.Type.of.Travel=Business travel,
      project.Class=Eco}                             => {Satisfaction=High} 0.05134840  0.2042163 2.079634  6529
```

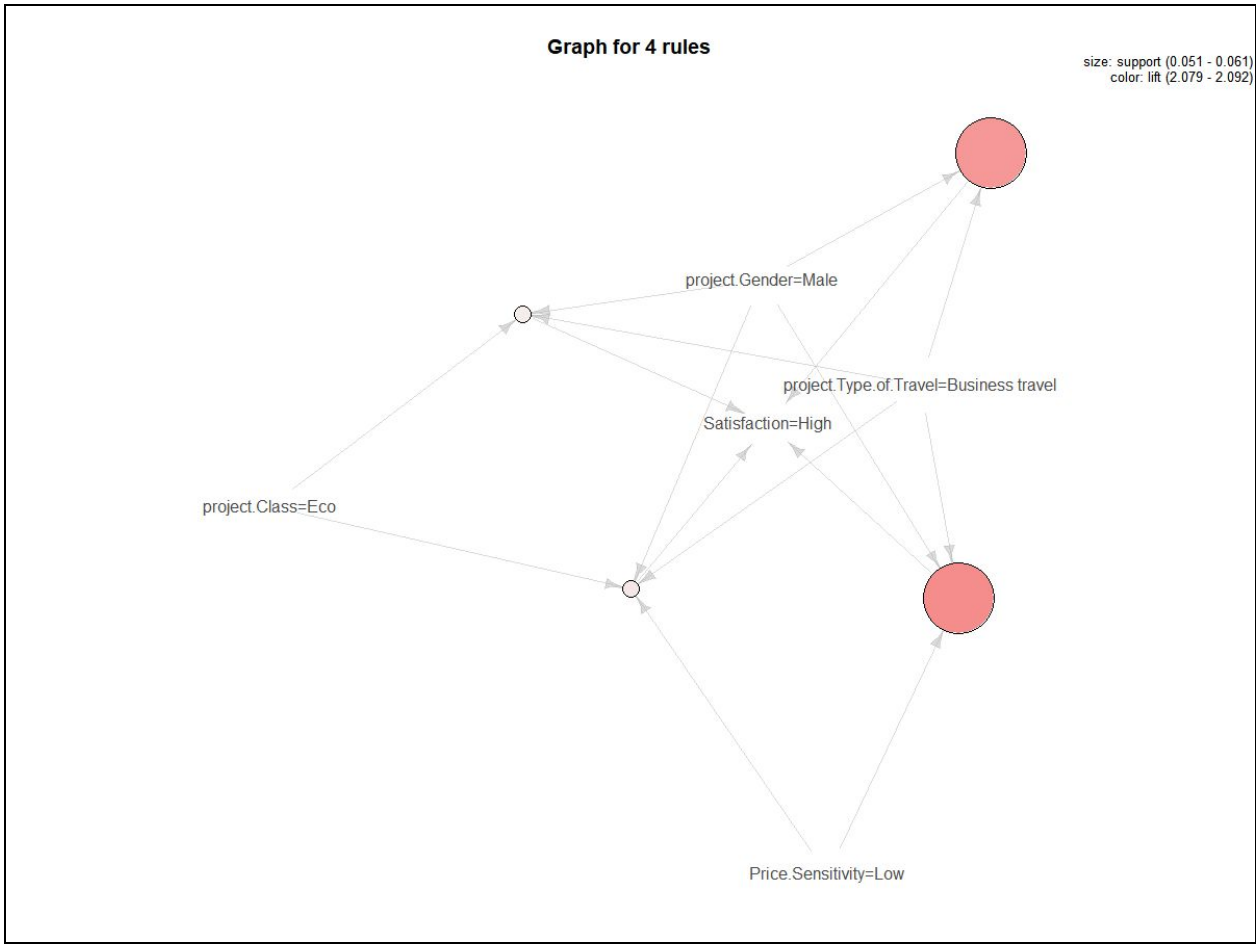


Fig 24

From our association rules analysis we have generated some useful insight. We can see that male passenger will oftentimes give a High satisfaction, from this we can say that we need to better market to female population to generate a higher satisfaction from them just as we get from the males. Other things that often occur together that generate a high customer satisfaction are having low prices, on business travel, travelling in the eco class. These attributes will be important for our actionable insights.

10.Business Questions:(Considering all Airlines)

- How to improve customer satisfaction rating for airlines?
    - Customer Satisfaction is based off of several variables given within the dataset. Through our analysis we have been able to narrow down those variables to a select few through our analysis. Within the report we have been able to find insight that variables such as Type of Travel and Airline Status are key variables. To be able to improve customer satisfaction the airline needs to formulate a business plan to better the experience on key variables that are statistically significant on having a high customer satisfaction.

- What are the attributes of an airline that influence customers' satisfaction rating?
    - Type of Travel, Class, Airline Status, Gender, Age, and Price Sensitivity are the key influences on a customers' satisfaction.
- Does airline status affect the consumer rating?
    - According to our association rules and linear modeling airline status is statistically significant in affecting customer satisfaction. We suggest an airline help improve the experience of the level that they are at or try to move up in airline status levels.
- Do most customers travel on business or for personal flights?
    - Most of the customers travel on Business flights, on average customers travelling for business give a higher customer satisfaction.
- How does the price affect a customer's decision on their rating?
    - Price is another key variable, according to our models having a lower price will result in a higher customer satisfaction. Most customers have a price sensitivity of 1, meaning they prefer lower prices. People at the price sensitivity of 1 most often give the highest satisfaction ratings.

11.Recommendations for SouthEast Airlines Co:

Having worked on the entire dataset, we even decided to take into consideration the satisfaction of customers when travelling through SouthEast Airlines. This helped us in determining the significant factors that people consider while travelling through their airlines and also how they affect their profits. In this case, our team would act as consultants to provide them solutions in the avenues they need to improve to make more profit.

This was done comparing the dataset which contained just SouthEast Airlines Co. and comparing its performance with other airlines.

To begin with, we made some visualizations:

1. **Satisfaction vs. arrival delay**

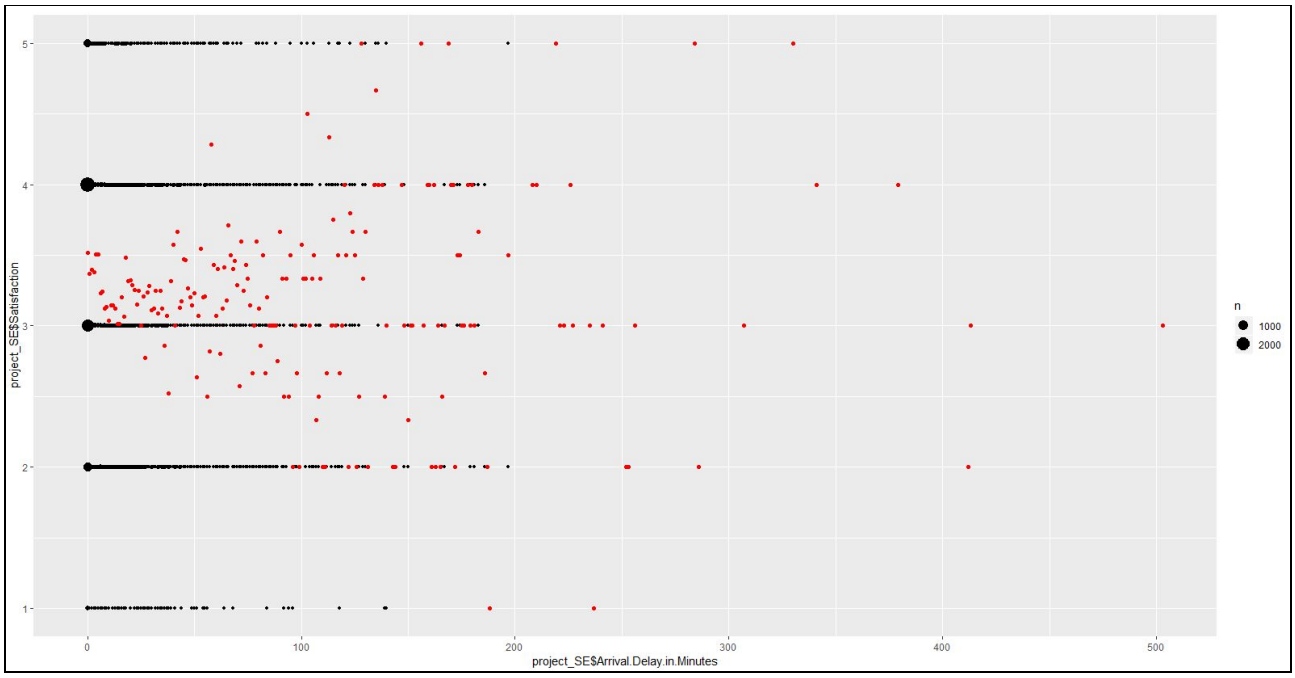


Fig 25

Here, as the arrival delay increases, the satisfaction rate tends to reduce. So, making arrangements for improving on-board experience of passengers is something SouthEast Airlines should consider because the arrival delay may be due to some external factors as well. So, keeping the passenger happy is something they could consider.

2. Satisfaction vs. departure delay

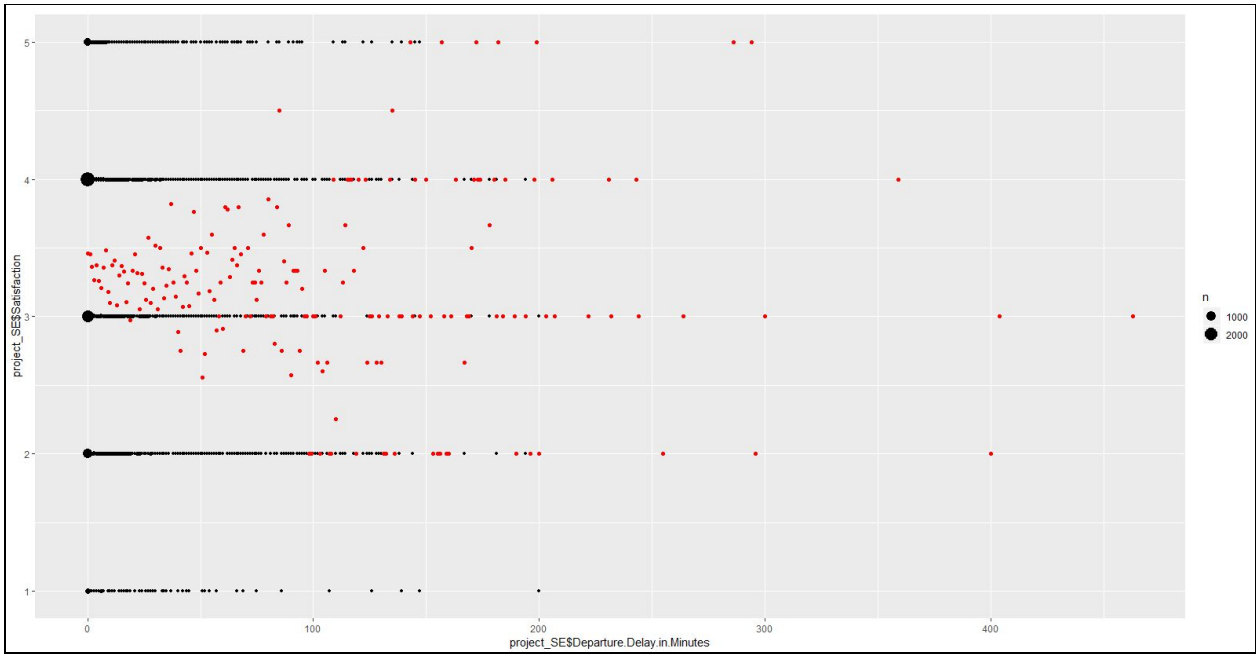


Fig 26

The departure delay doesn't affect the satisfaction rating for the airlines.

3. Satisfaction vs. Airline Status

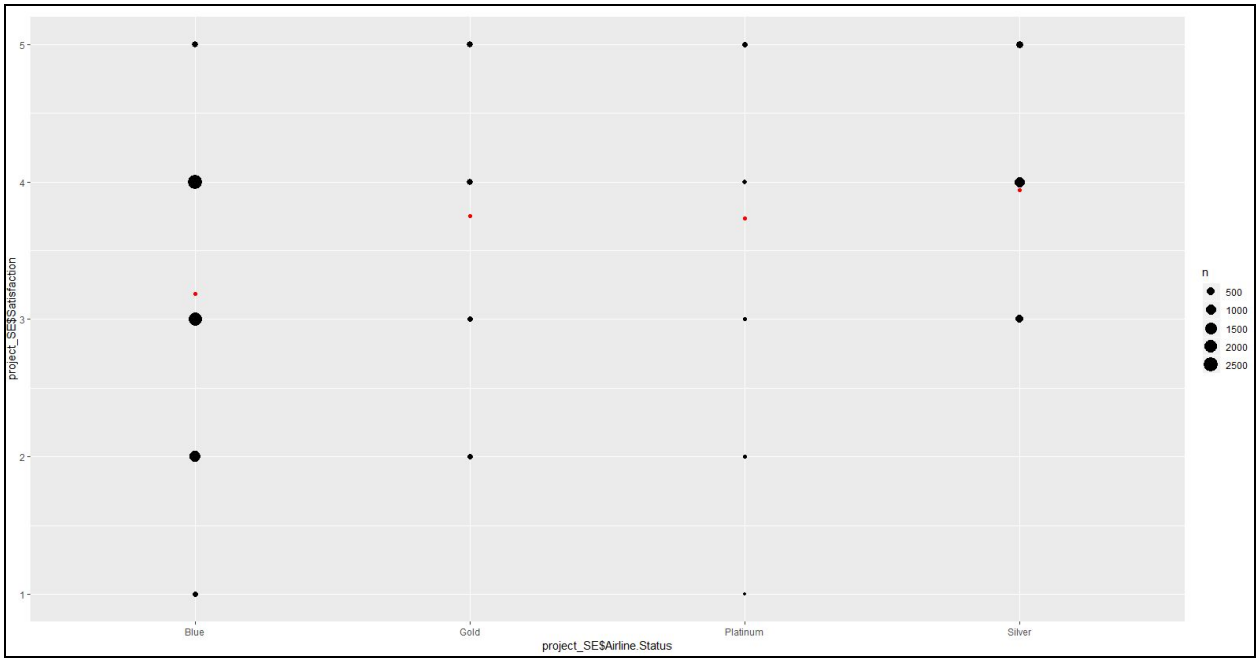


Fig 28

People travelling through blue status have a low satisfaction and those travelling through Silver and Platinum have a high satisfaction rating.

4. Satisfaction vs. no.of flights per airline

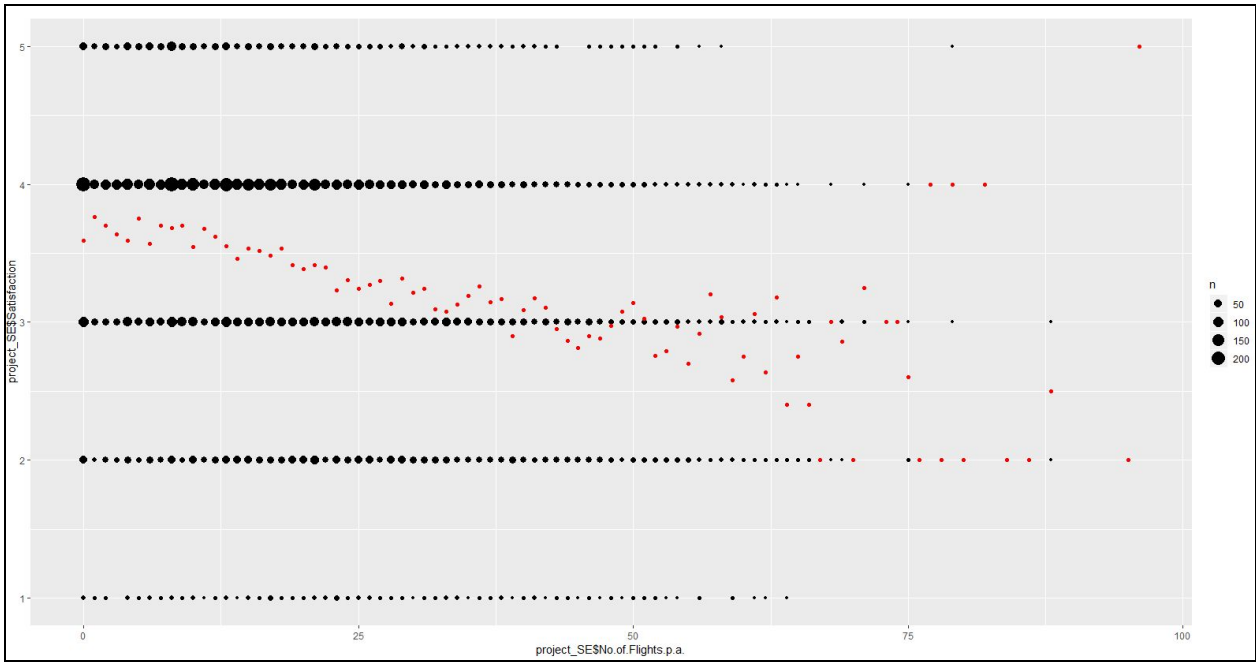


Fig 29

For SouthEast Airlines, the satisfaction tends to decrease as the number of airlines increases

5. Satisfaction vs. Class

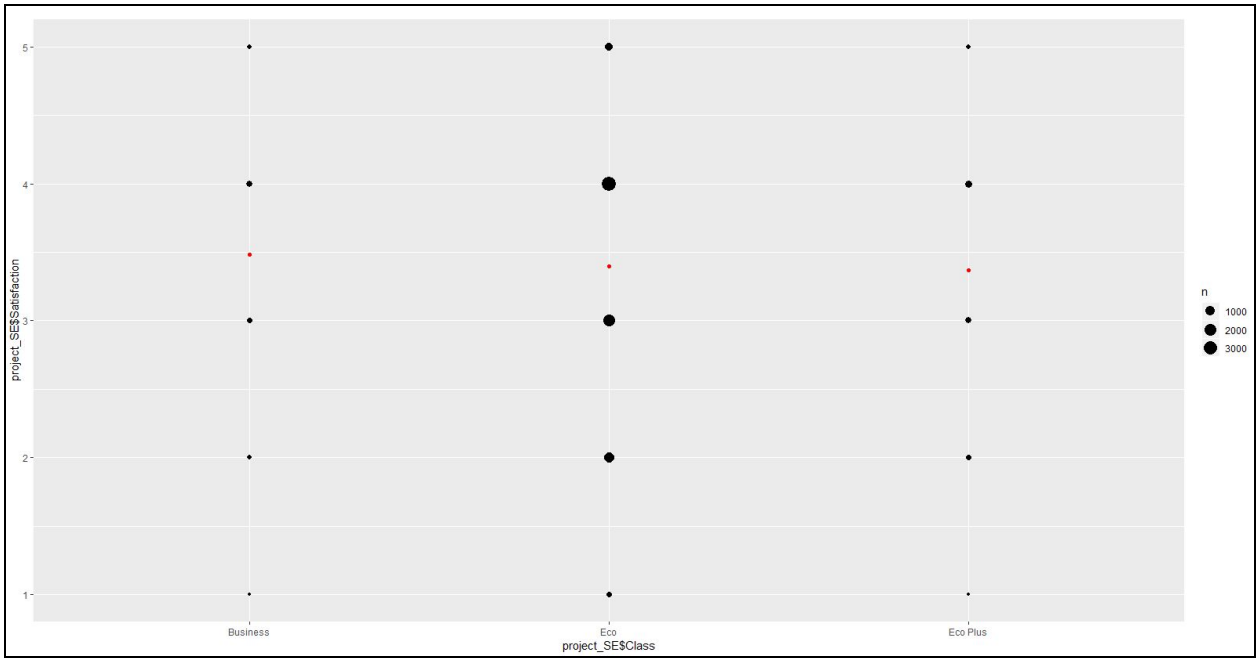


Fig 30

So, after Taking into account the above visualizations, and also the models we came up with for SouthEast Airlines, the following are our recommendations to them:

1. The satisfaction is less for people travelling with an airline having blue status. Thus, marketing the blue status of airline in such a way that it attracts more customers is one of the things that needs to be considered.
2. There are more people travelling through economy class for SouthEast Airlines. Thus, giving people rewards or improving their on-board experience will help them further.

12.Recommendation of Actionable Insights:

Our recommendations have been concluded as a result from data visualizations and modeling techniques. It is critical to focus on the the variables that are actionable and will generate the highest customer satisfactions. From the visualizations generated we have identified that airlines should concentrate more on the economy class passengers when majority travelling for business purposes. We suggest that they provide better deals and rates for corporate customers. Our descriptive statistics also suggest that having a good airline status would also help in improving the customer experience and hence, airline companies could focus on business plans and ideas to change their airline status to silver, gold or platinum. From linear modeling we have identified that customer type of travel and airline status have a high effect on the customer giving a high satisfaction rating. From this we can say that it the company should focus on why a customer might be travelling to tailor fit their experience to that. Also, they should focus on their airline ranking to either better the experience in their current rank or make an effort to move up their ranking. After applying association rules we find that class and price sensitivity are also important factors to focus on as well. The airline should also focus on a business plan to offer competitively low pricing for all class, especially the economy class in efforts to generate higher customer satisfaction ratings.

Our descriptive statistics and our three models all pointed toward similar results. For the airline to be able to generate a higher satisfaction we suggest that they focus their business plan on the type of travel customers are using, their airline status, price sensitivity, and the class that customers are flying on. We suggest that they build a model that can cater to not only business travel but mileage and personal travel that is offered at a competitively low price with a focus on selling economic class tickets while improving the experience of flying in other classes.

13.MIDST

As a part of our regular homework submissions, we used midst, a platform where group members could collaborate with each other and work on the tasks assigned to them remotely. For this project, we too, created a separate space where we could combine different nodes related to different parts the project was divided in.

Kanban Board is shown below:

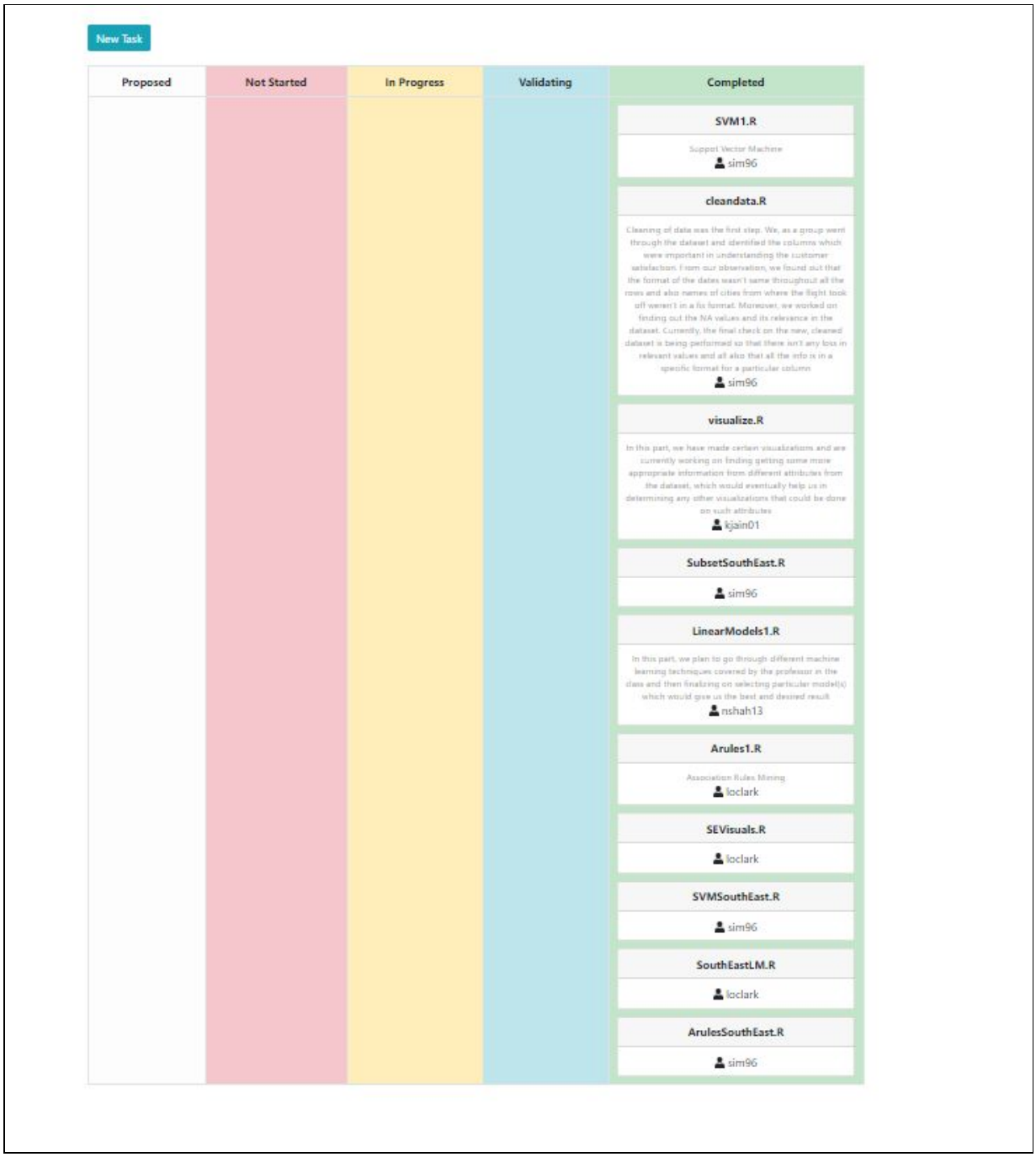


Fig 31

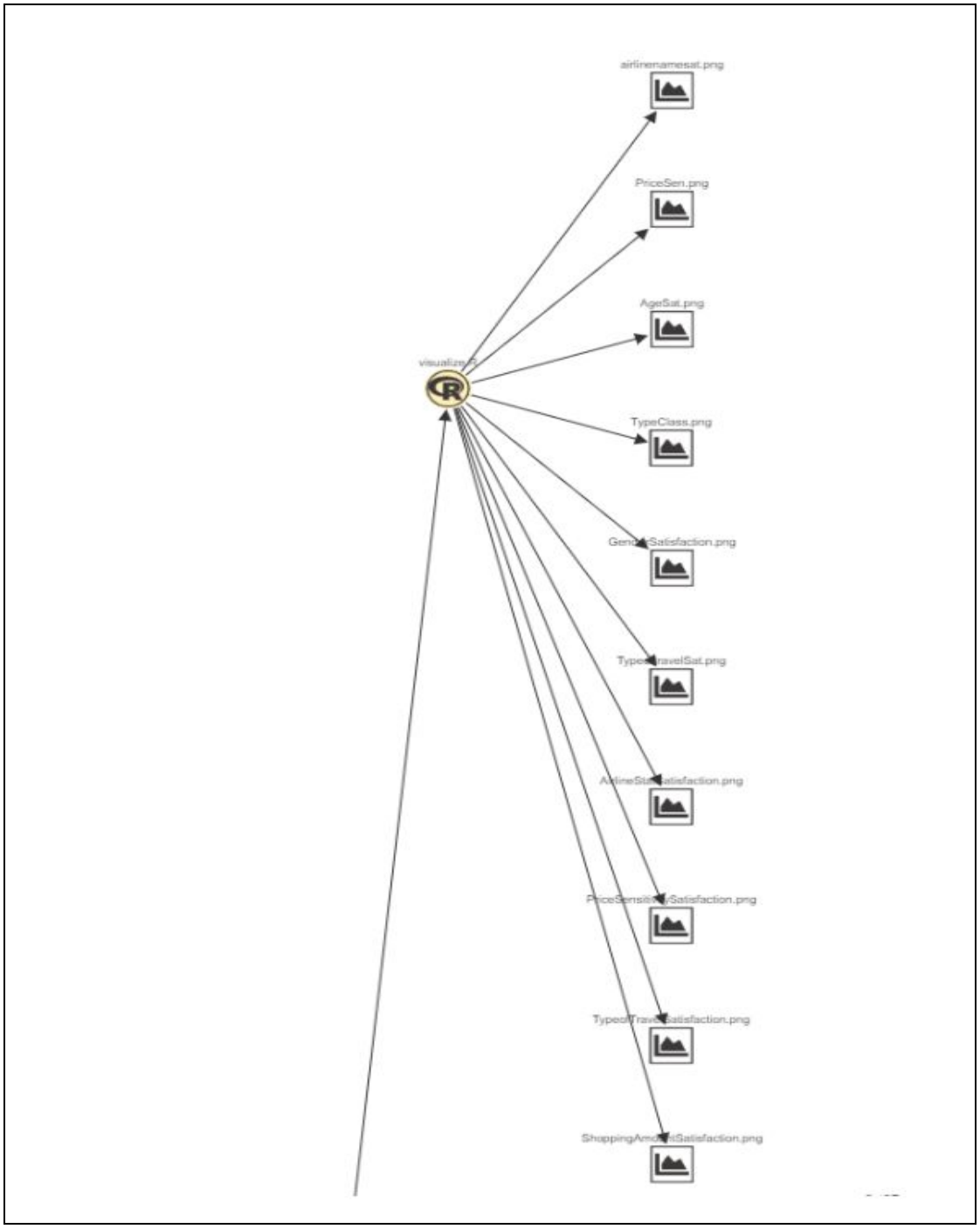
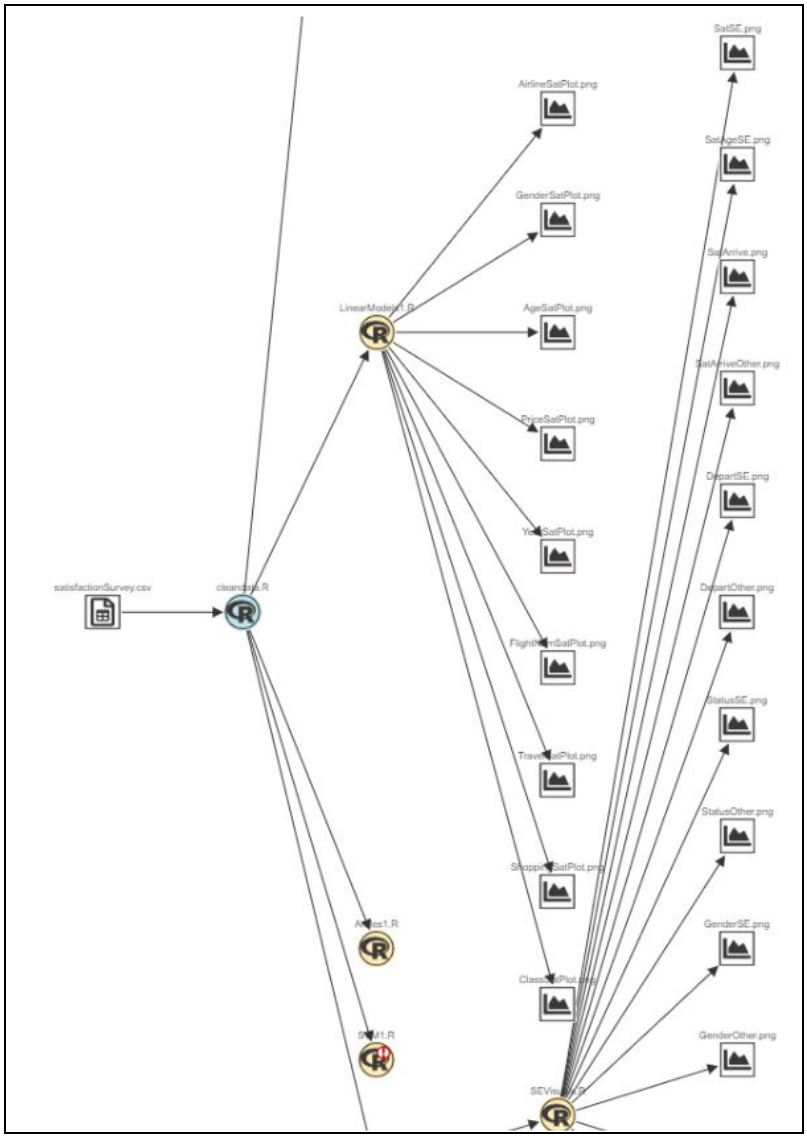
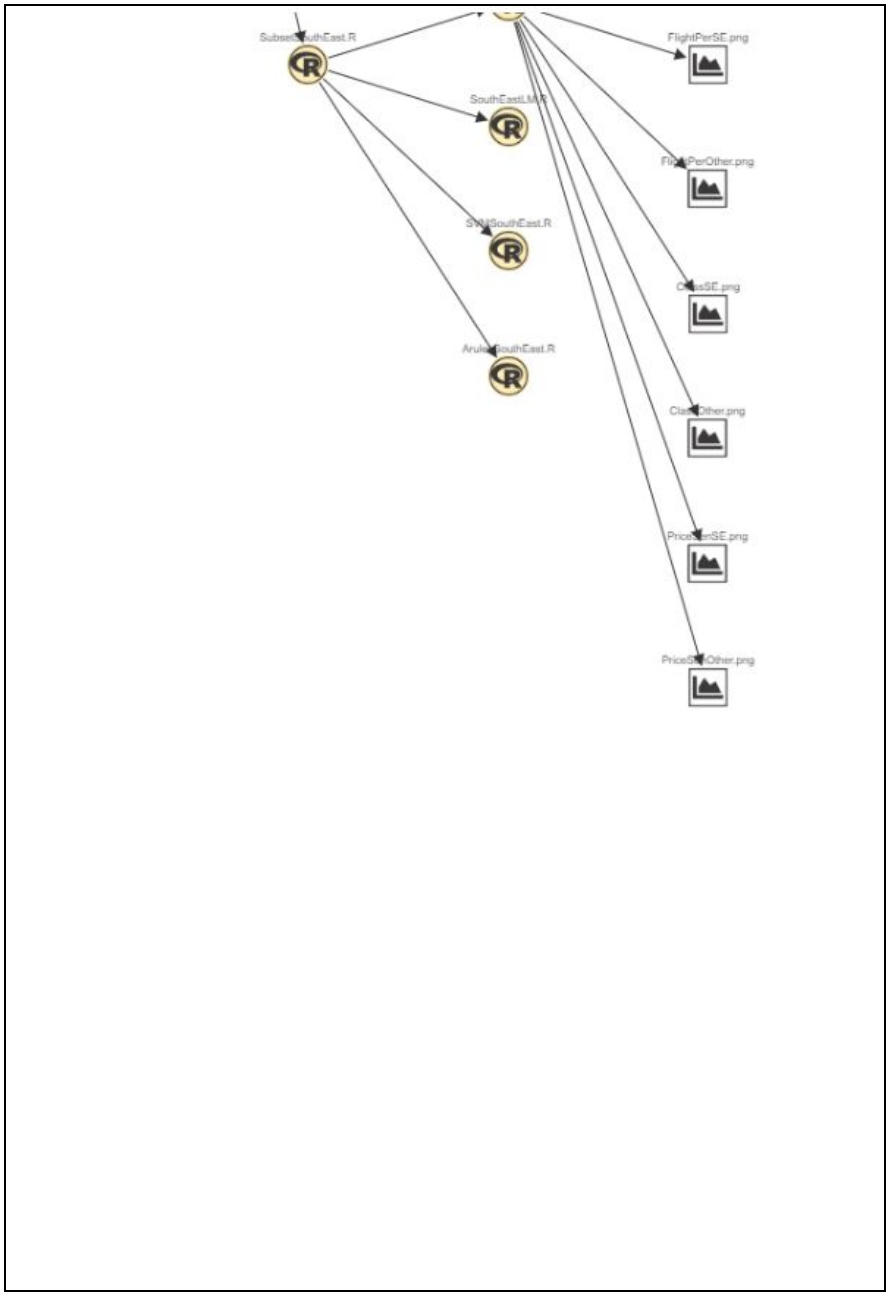


Fig 32: Screenshot of midst


SatSE.png
AirlineSatPlot.png
SatAgeSE.png
GenderSatPlot.png
SatArrive.png
LinearModel1.R
AgeSatPlot.png
SatArriveOther.png
PriceSatPlot.png
DepartSE.png
satisfactionSurvey.csv
DepartOther.png
FlightDistSatPlot.png
StatusSE.png
StatusOther.png
GenderSE.png
GenderOther.png



14. Appendix:

Link to the Project code: https://notepad.pw/simaant