Airline Survey Data Analysis

Southeast Airline Co.



Prof Krudys  
  
  
Tamilselvan Tamilmani

Amanda Xiong

Kelly Spilman

IST 687 | December 14, 2020

Table of Contents

[Introduction 2](#_Toc500423205)

[Industry Analysis 2](#_Toc500423206)

Overview of Southeast Airline Co………………………………………………………………………………………. 4

Business Questions…………………………………………………………………………………………………………….. 5

Data Identification…………………………………………………………………………………………………………………… 5

[Descriptive Statistics](#_Toc500423211) 8

[Modeling Techniques](#_Toc500423212) 11

Association Rules ……………………………………………………………………………………………………………… 17

Visualizations …………………………………………………………………………………………………………………… 21

Business Questions Interpretations …………………………………………………………………………………. 32

Recommendations……………………………………………………………………………………………………………..38

# Introduction

Project Objective:

The data analysis project's objective is to analyze data collected through airline customers' satisfaction surveys. The focus is to identify patterns and provide concrete recommendations to Southeast Airline Co. to improve the satisfaction scores received from their customers.

Scope:

The scope of the project is limited to a flight booked within the United States. The survey data for January 2014, February 2014, and March 2014 have been analyzed, pre-processed, and analyzed again.

Deliverables:

* Identification of areas in which the airlines can make changes or educated decisions to improve the customer experience.
* Provide statistical data to Southeast Airline Co. to aid in future decisions to improve satisfaction scores for their company.
* Recommendations of different offers/accommodations airlines could make to customers to improve satisfaction scores.

# Industry Analysis

Airline industry:

The airline industry has four major strengths, High Income, Growing Tourism Industry, Continued Growth and Safety, and Speed. The tourism industry has been steadily increasing for the past decade, and airplanes are the only viable options when traveling long distances over short timeframes. Planes are safe and speedy; the staff is highly trained and well equipped.

However, the Airline industry also has several weaknesses, Slow Rate of Infrastructure, High Spoilage Rate, Huge Investments, Competitive Market. Spoilage rates in the airline industry equate to passengers missing their flights, and a refund is issued. Airline companies must invest large amounts of money early on, so long-term success is significant. Additionally, and most importantly, the airline industry is an extremely competitive market; these companies are not competitive in pricing and service; they can lose customers very quickly.

The Airline Industry Customer Experience Benchmarks have been identified, which are as follows:

* Airline Status
* Age
* Gender
* Price Sensitivity
* Year of First Flight
* The number of Flights P.A.
* Percentage of Flights taken with other airlines
* Type of Travel
* Number of other Loyalty Cards
* Shopping amount at Airport
* Eating and Drinking at the 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 Cancelled
* Flight time in minutes
* Flight Distance
* Arrival Delay greater than 5 minutes

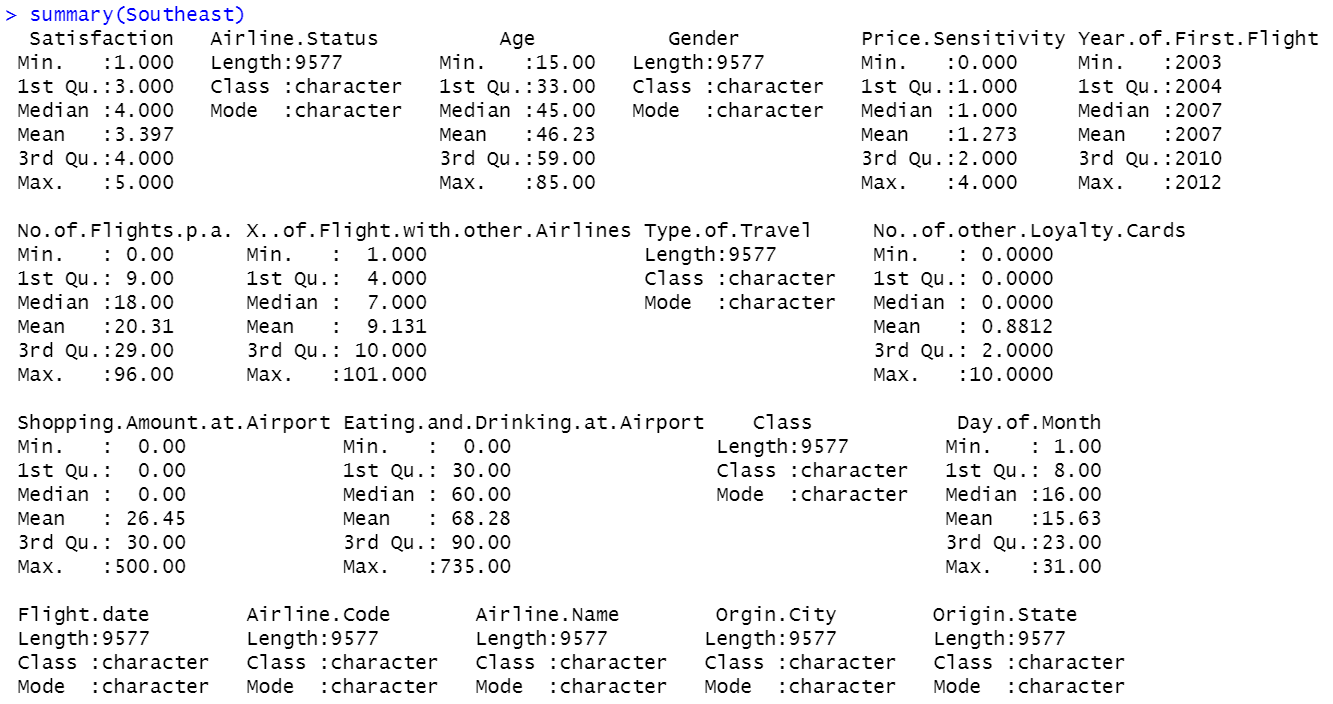
Satisfaction Score:

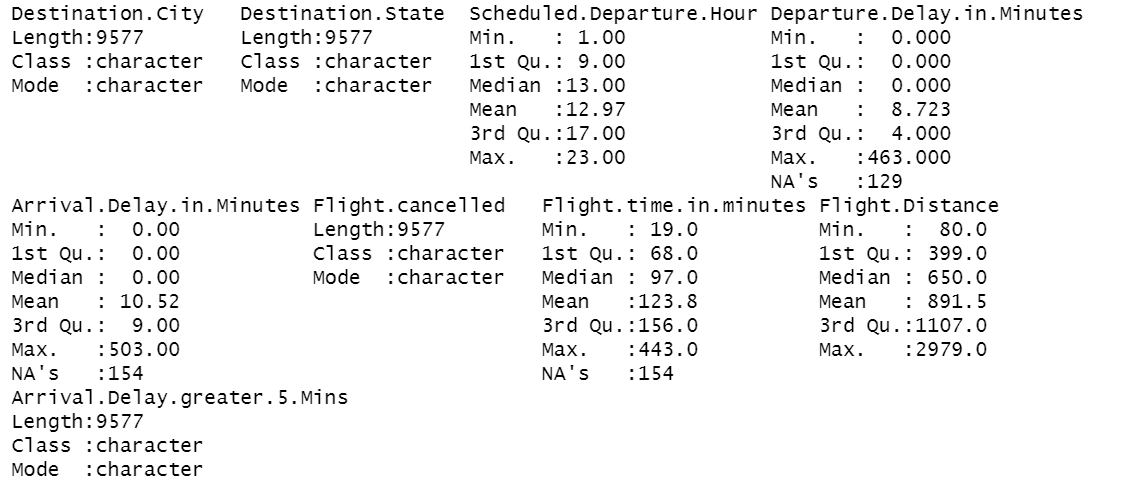
Satisfaction Score measures the Satisfaction of the customer's experience while traveling on one Fl; this score can be used to predict future business growth.

Satisfaction is rated for 1 to 5, 5 being the highest satisfied, and 1 is the lowest satisfaction level. Based on the above attributes, a customer ranks the Flight on the 1- 5 scale

# 

# Overview of Southeast Airline Co.





# Business Questions

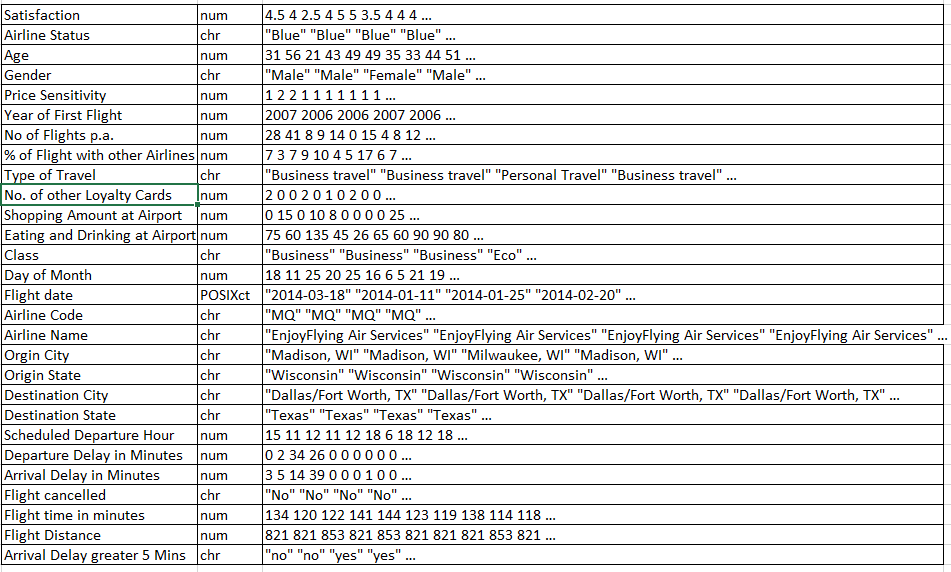
Original

1. Based on satisfaction, which airlines are performing well, and which are performing poorly?
2. How does Airline Class affect customer satisfaction based on age group?
3. Could eating and drinking at the airport before a flight affects satisfaction?
4. What are times of operation critical in affecting satisfaction score?
5. What elements contribute to the highest satisfaction Scores?

New from Data

1. How does Southeast Airlines Co. perform on satisfaction among customers that have never flown before?
2. How does Southeast Airline Co. perform based on the type of Travel?

# Data Identification

The Customer Satisfaction data from January, February, and March 2014 has 130,000customer data with around 28 relevant attributes of the Flight taken by the customer.

When viewing the data structure, we observed that 4 of the variables contained NA's.

1. Satisfaction had only three rows that contained NA's. Due to the low number, we dropped those rows.
   1. Flight Time and Arrival and Departure delay, we replaced the NA's with the mean.
   2. We tried the alternative option of replacing the NA's with 0. However, the mean gave us a better linear model.
2. Flight Time in Minutes we used the mean for that specific route (Source City to Destination City).
   1. There were 3 NA's after this process; we dropped those three rows.

as.data.frame(sapply(surveyDataRaw,function(X) sum(is.na(X))))

|  |  |
| --- | --- |
| **Column** | **Total of NA's** |
| Flight time in minutes | 2738 |
| Arrival Delay in Minutes | 2738 |
| Departure Delay in Minutes | 2345 |
| Satisfaction | 3 |
| Year of First Flight | 0 |
| Type of Travel | 0 |
| Shopping Amount at Airport | 0 |
| Scheduled Departure Hour | 0 |
| Price Sensitivity | 0 |
| Origin State | 0 |
| Origin City | 0 |
| No. of other Loyalty Cards | 0 |
| No of Flights p.a. | 0 |
| Gender | 0 |
| Flight Distance | 0 |
| Flight date | 0 |
| Flight canceled | 0 |
| Eating and Drinking at the Airport | 0 |
| Destination State | 0 |
| Destination City | 0 |
| Day of Month | 0 |
| Class | 0 |
| Arrival Delay greater 5 Mins | 0 |
| Airline Status | 0 |
| Airline Name | 0 |
| Airline Code | 0 |
| Age | 0 |
| % of Flight with other Airlines | 0 |

**Numerical & Categorical Columns**

We have 15 numerical columns and 12 categorical, and one date column. We have converted all categorical to factors

|  |  |
| --- | --- |
| Numerical | Categorical |
| "Satisfaction"  "Age"  "Price Sensitivity"  "Year of First Flight"  "No of Flights p.a."  "% of Flight with other Airlines"  "No. of other Loyalty Cards"  "Shopping Amount at Airport"  "Eating and Drinking at Airport"  "Day of Month"  "Flight date"  "Scheduled Departure Hour"  "Departure Delay in Minutes"  "Arrival Delay in Minutes"  "Flight time in minutes" | "Airline Status"  "Gender"  "Type of Travel"  "Class"  "Airline Code"  "Airline Name"  "Orgin City"  "Origin State"  "Destination City"  "Destination State"  "Flight cancelled"  "Arrival Delay greater 5 Mins"    Date: Flight Date |

We also tried breaking down the delays to multiple bins and found only (arrival/departure) delay more significant than 5 mins has some significance.

departure\_delay\_gt\_15

departure\_delay\_gt\_30

departure\_delay\_gt\_60

departure\_delay\_gt\_120

arrival\_delay\_gt\_5

arrival\_delay\_gt\_15

arrival\_delay\_gt\_30

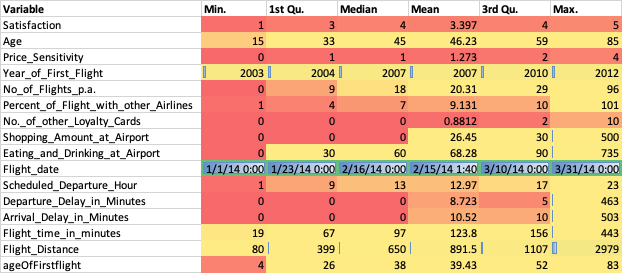
arrival\_delay\_gt\_60

arrival\_delay\_gt\_120

We created a new variable by combining Origin and destination city to create a path variable, but this was also not significant along with the origin and destination city

Descriptive Statistics

**Summary Statistics for Numerical Variables:**

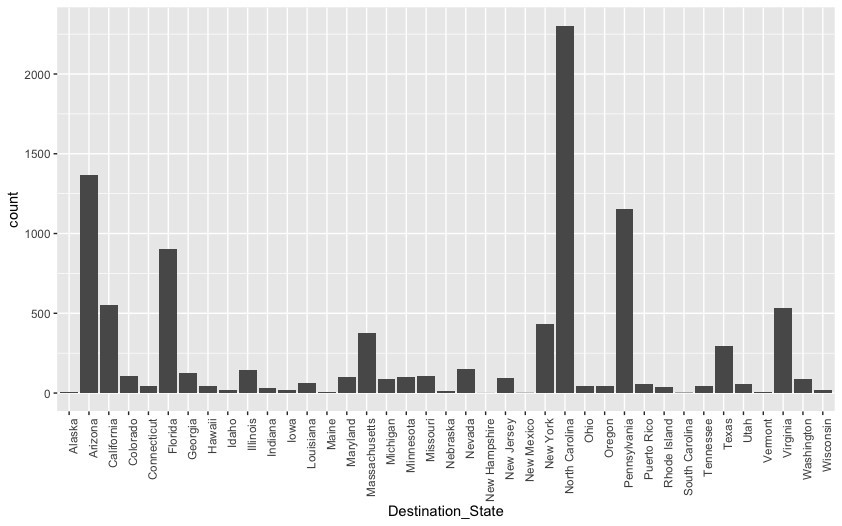


From the summary statistics of numerical variables, we can see the below observations

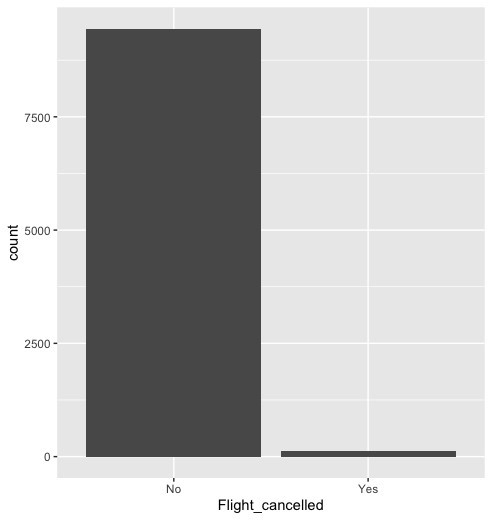
1. Average Satisfaction is 3.4
2. The Average flights taken by the passenger is 20
3. It is interesting to see why we have zero in here
4. The average percent of Flight with other airlines is around 9
5. Given 13 other airlines, we can say, on average, the passenger takes one other airline other than Southeast.
6. The flight date is the first quarter of 2014
7. We would not be able to explain the seasonality with this data set.
8. Nor can we predict how our airline is going to do for the remainder of the year.

**Summary Statistics for Categorical Variables:**

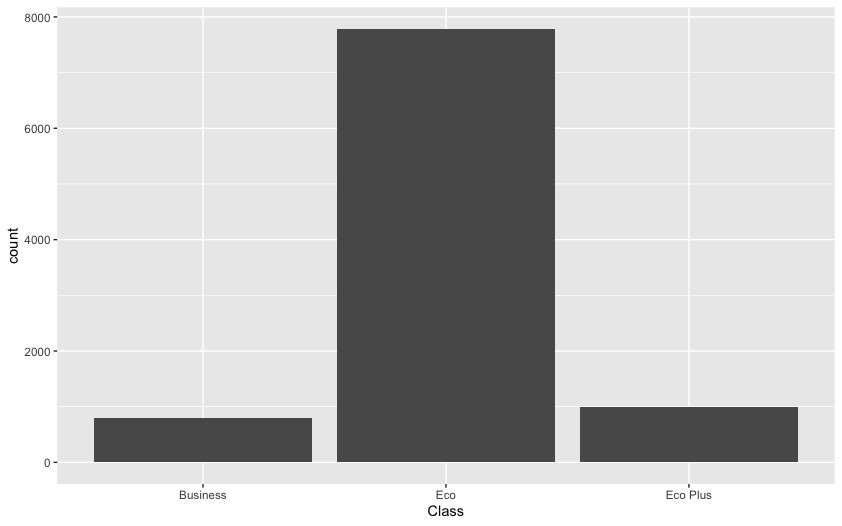
For the southeast, the most popular destination is North Carolina.



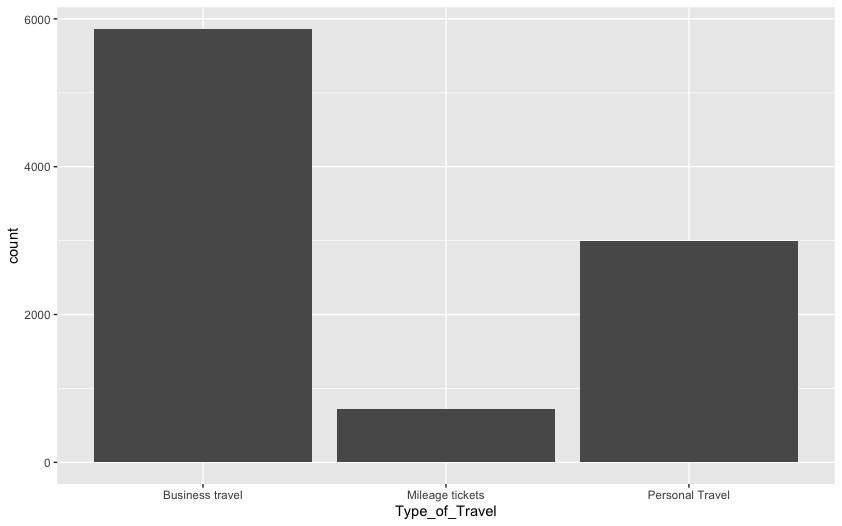
The number of flights canceled vs. not canceled is meager. This might lead to skewed distribution:



We have too many people traveling in Economy class, again, which will lead to skewed distribution:



Most southeast travelers are Business travelers:



**Modeling techniques**

We first ran our linear model with all variables to get the significant variables, which yielded a  43.5 R Squared value model and the below significant variables.

* + Airline\_Status
  + Gender
  + Age
  + No\_of\_Flights\_p.a.
  + Type\_of\_Travel
  + Eating\_and\_Drinking\_at\_Airport
  + Class
  + Scheduled\_Departure\_Hour
  + Flight\_time\_in\_minutes

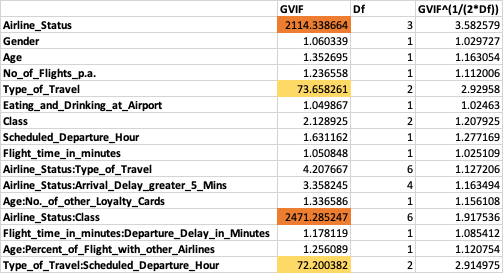
Then we ran our moderating effect analysis model, which yielded the below interactions:

* + Airline\_Status:Type\_of\_Travel
  + Airline\_Status:Arrival\_Delay\_greater\_5\_Mins
  + Age:No.\_of\_other\_Loyalty\_Cards
  + Airline\_Status:Class
  + Departure\_Delay\_in\_Minutes:Flight\_time\_in\_minutes
  + Age:Percent\_of\_Flight\_with\_other\_Airlines
  + Type\_of\_Travel:Scheduled\_Departure\_Hour

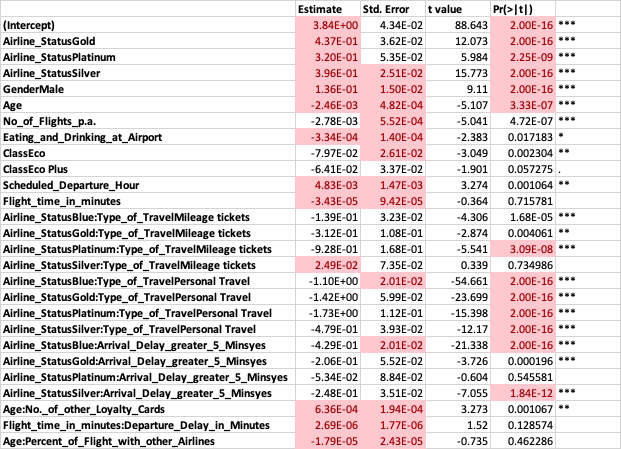
Finally, we used both the significant and interaction variables to run our final linear model, which gave an adjusted R Squared of 45.68, but it still had dependent columns.

Then we ran through the variable influence factor analysis and filtered the dependent variables

1. Airline Status and Airline Status: Class is dependent, we removed higher value "Airline Status: Class
2. Type\_of\_Travel:Scheduled\_Departure\_Hour & Type\_of\_Travel are dependent. We removed higher valued Travel



This is our Final Model Coefficients:



#Final LM Model

sig\_cols\_vif<-c("Airline\_Status","Gender","Age","No\_of\_Flights\_p.a.","Eating\_and\_Drinking\_at\_Airport","Class","Scheduled\_Departure\_Hour","Flight\_time\_in\_minutes")

mod\_cols\_vif<-c("Airline\_Status:Type\_of\_Travel","Airline\_Status:Arrival\_Delay\_greater\_5\_Mins","Age:No.\_of\_other\_Loyalty\_Cards","Departure\_Delay\_in\_Minutes:Flight\_time\_in\_minutes","Age:Percent\_of\_Flight\_with\_other\_Airlines")

formula\_sig\_mod\_vif<-paste("Satisfaction", paste(c(sig\_cols\_vif,mod\_cols\_vif), collapse=" + "), sep=" ~ ")

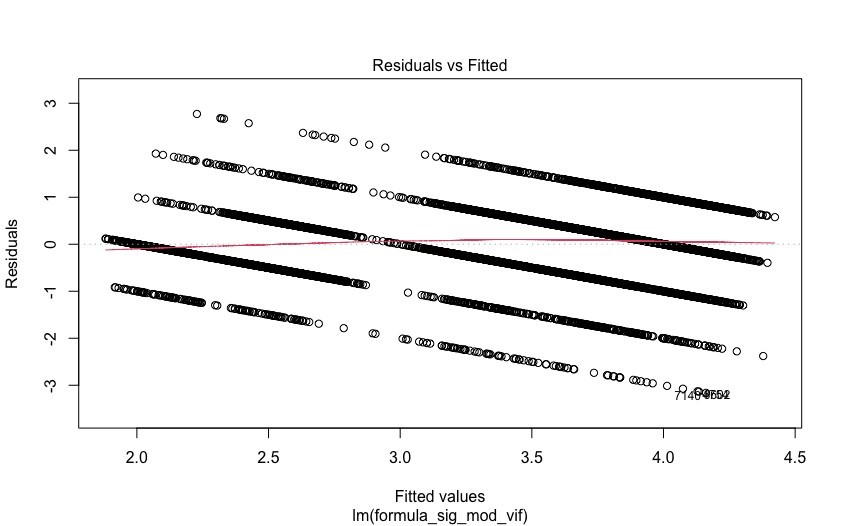
lm.sig\_mod\_vif\_model<- lm(formula = formula\_sig\_mod\_vif,lmData)

summary(lm.sig\_mod\_vif\_model)

vif(lm.sig\_mod\_vif\_model)

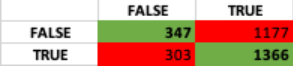
plot(lm.sig\_mod\_vif\_model)

Our model produced a residual plot with a pattern, which means our data is not linear, so the linear model may not be the right fit for this analysis.



Next, we ran a classification model:

We grouped Satisfaction above three as good and below three as bad and train our model to classify this binary option for our classification model.

* + We split our data 2:1 for train and test and
  + We got an accuracy of 53.64 percent with the below confusion matrix
  + 

#SVM

#Above 3 is good, and below is Satisfactiontion

randIndex <- sample(1:nrow(surveyDataRaw))

cutpoint<- floor(nrow(surveyDataRaw)\*2/3)

svm\_numerical\_columns <- numeric\_columns[!numeric\_columns %in% c("Flight\_date","ageOfFirstflight")]

svm\_categorical\_columns <- categorical\_columns[!categorical\_columns %in% c("path","Orgin\_City","Destination\_City","Origin\_State","Destination\_State","Day\_of\_Month","Month\_of\_Year","Flight\_Date")]

svmData<-surveyDataRaw[,svm\_numerical\_columns]

svmData$Satisfaction<-ifelse(svmData$Satisfaction>3,TRUE,FALSE)

#svmData$Satisfaction<-as.factor(ifelse(svmData$Satisfaction>3,1,0))

#svmData<-convertToFactors(svmData)

for(c in categorical\_columns){

  if(c %in% svm\_categorical\_columns){

    binary<-to.dummy(surveyDataRaw[[c]],c)

    svmData<- cbind(svmData,binary)

  }

}

svm\_trainData<- svmData[randIndex[1:cutpoint],]

svm\_testData <- svmData[randIndex[(cutpoint+1):nrow(svmDataMatrix)],]

svm\_trainDataMatrix<-as.matrix(svm\_trainData)

svm\_testDataMatrix<-as.matrix(svm\_testData)

#svm\_sig\_cols\_vif<-c("Airline\_Status","Gender","Age","No\_of\_Flights\_p.a.","Eating\_and\_Drinking\_at\_Airport","Class","Scheduled\_Departure\_Hour","Flight\_time\_in\_minutes")

#svm\_mod\_cols\_vif<-c("Airline\_Status\*Type\_of\_Travel","Airline\_Status\*Arrival\_Delay\_greater\_5\_Mins","Age\*No.\_of\_other\_Loyalty\_Cards","Departure\_Delay\_in\_Minutes\*Flight\_time\_in\_minutes","Age\*Percent\_of\_Flight\_with\_other\_Airlines")

svm\_formula\_sig\_mod\_vif<-paste("Satisfaction", paste(c(numeric\_columns,svm\_categorical\_columns), collapse=" + "), sep=" ~ ")

set.seed(2732)

svm.model<- ksvm(formula = Satisfaction ~.,svm\_trainDataMatrix,kernel = "rbfdot",kpar="automatic",C=0.1,cross=2,prob.model=TRUE)

#svm.sig\_mod\_vif\_model<- ksvm(formula = svm\_formula\_sig\_mod\_vif,svmData,kernel = "stringdot",kpar=list(length = 4, lambda = 0.5),C=0.1,cross=2,prob.model=TRUE)

summary(svm.model)

plot.svm(svm.model,svm\_testData)

svm\_testData$predicted<-predict(svm.model,svm\_testDataMatrix)

table(svm\_testData$Satisfaction,svm\_testData$predicted)

percent\_svm<-sum(ifelse(svm\_testData$predicted==svm\_testData$Satisfaction,1,0))/length(svm\_testData$predicted)

percent\_svm

After we ran that code, we went back and compared it against our linear model.

* That gave an accuracy of 57.69%
* We have a better prediction in the Linear Model.

We then ran a logit model, which gave a 34% accuracy model.

#Try a Logit

logit\_trainData<-lm\_trainData

logit\_testData<-lm\_testData

logit\_trainData$Satisfaction<-ifelse(logit\_trainData$Satisfaction>3,1,0)

logit\_testData$Satisfaction<-ifelse(logit\_testData$Satisfaction>3,1,0)

logit.model<-glm(formula =formula\_sig\_mod\_vif,data=logit\_trainData,family=binomial(logit))

summary(model.logit)

logit\_testData$predicted<-predict(logit.model,logit\_testData)

logit\_percent<-sum(ifelse(round(logit\_testData$predicted)==logit\_testData$Satisfaction,1,0))/length(logit\_testData$predicted)

logit\_percent

table(lm\_testData$Satisfaction,round(lm\_testData$predicted))

#Linear model for whole data set

#Whole data set

surveyDataFull<- readxl::read\_excel("~/Downloads/FinalProjectMaterial\_2\_2/Satisfaction Survey(2).xlsx",col\_types = "guess" )

surveyDataFull<- cleanData(surveyDataFull)

Overall LM Model

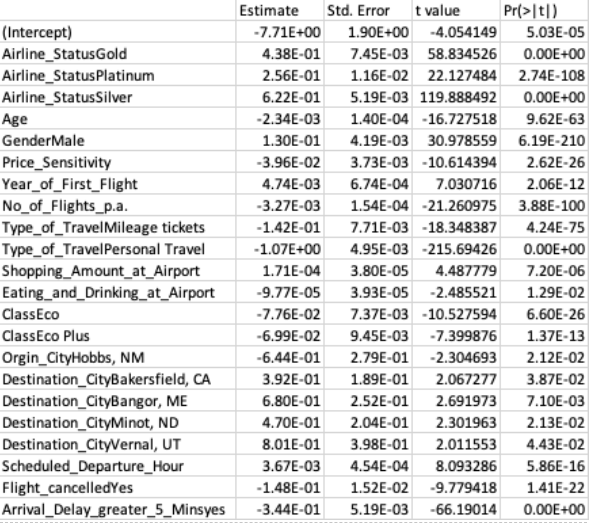
lm\_fulldata<- lm(formula = Satisfaction ~.,data=surveyDataFull)

lm\_coeff\_full<-summary(lm\_fulldata)$coeff

lm\_sig\_full<-lm\_coeff\_full[lm\_coeff\_full[,4] < 0.05,]

lm\_sig\_full

summary(lm\_fulldata)



Association Rules

We removed redundant columns from the data set and ran our apriori rules, and picked our only rules with a lift of above 2.5.

#Association Rules result

arules\_columns<-colnames(surveyDataRaw)

#remove redundant columns

arules\_columns<-arules\_columns[!arules\_columns %in% c("ageOfFirstflight","Flight\_Distance","departure\_delay\_gt\_5","arrival\_delay\_gt\_5","Arrival\_Delay\_greater\_5\_Mins","path","Origin\_State","Destination\_State")]

arulesData<-surveyDataRaw[,arules\_columns]

arulesTransactions<-as(arulesData, "transactions")

itemFrequencyPlot(arulesTransactions,support=0.5)

arules.model<-apriori(arulesTransactions,parameter = list(support=0.1,confidence=0.7))

summary(arules.model)

#inspect(arules.model)

#plot(arules.model)

goodrules <- arules.model[quality(arules.model)$lift > 2.5]

plot(goodrules)

inspect(goodrules)

     lhs                                   rhs                                support confidence  coverage     lift count

[1]  {Satisfaction=[1,3),

      No.\_of\_other\_Loyalty\_Cards=[0,1)} => {Type\_of\_Travel=Personal Travel} 0.1039992  0.8170632 0.1272841 2.610949   996

[2]  {Satisfaction=[1,3),

      Gender=Female}                    => {Type\_of\_Travel=Personal Travel} 0.1007622  0.8326143 0.1210191 2.660643   965

[3]  {Satisfaction=[1,3),

      Airline\_Status=Blue}              => {Type\_of\_Travel=Personal Travel} 0.1367860  0.7924985 0.1726010 2.532452  1310

[4]  {Satisfaction=[1,3),

      Price\_Sensitivity=[1,4],

      No.\_of\_other\_Loyalty\_Cards=[0,1)} => {Type\_of\_Travel=Personal Travel} 0.1006578  0.8155668 0.1234207 2.606167   964

[5]  {Satisfaction=[1,3),

      No.\_of\_other\_Loyalty\_Cards=[0,1),

      Flight\_cancelled=No}              => {Type\_of\_Travel=Personal Travel} 0.1017020  0.8157454 0.1246737 2.606738   974

[6]  {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Class=Eco}                        => {Type\_of\_Travel=Personal Travel} 0.1124569  0.7965976 0.1411716 2.545551  1077

[7]  {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Price\_Sensitivity=[1,4]}          => {Type\_of\_Travel=Personal Travel} 0.1317740  0.7927136 0.1662316 2.533139  1262

[8]  {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Flight\_cancelled=No}              => {Type\_of\_Travel=Personal Travel} 0.1339668  0.7914867 0.1692597 2.529219  1283

[9]  {No\_of\_Flights\_p.a.=[25,96],

      Type\_of\_Travel=Personal Travel,

      No.\_of\_other\_Loyalty\_Cards=[0,1)} => {Age=[53,85]}                    0.1135011  0.8917145 0.1272841 2.558403  1087

[10] {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Price\_Sensitivity=[1,4],

      Class=Eco}                        => {Type\_of\_Travel=Personal Travel} 0.1089068  0.7986217 0.1363684 2.552019  1043

[11] {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Class=Eco,

      Flight\_cancelled=No}              => {Type\_of\_Travel=Personal Travel} 0.1099509  0.7959184 0.1381435 2.543380  1053

[12] {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Price\_Sensitivity=[1,4],

      Flight\_cancelled=No}              => {Type\_of\_Travel=Personal Travel} 0.1290592  0.7918001 0.1629947 2.530220  1236

[13] {Price\_Sensitivity=[1,4],

      No\_of\_Flights\_p.a.=[25,96],

      Type\_of\_Travel=Personal Travel,

      No.\_of\_other\_Loyalty\_Cards=[0,1)} => {Age=[53,85]}                    0.1097421  0.8921902 0.1230030 2.559768  1051

[14] {No\_of\_Flights\_p.a.=[25,96],

      Type\_of\_Travel=Personal Travel,

      No.\_of\_other\_Loyalty\_Cards=[0,1),

      Flight\_cancelled=No}              => {Age=[53,85]}                    0.1112039  0.8912134 0.1247781 2.556965  1065

[15] {Satisfaction=[1,3),

      Airline\_Status=Blue,

      Price\_Sensitivity=[1,4],

      Class=Eco,

      Flight\_cancelled=No}              => {Type\_of\_Travel=Personal Travel} 0.1064008  0.7979640 0.1333403 2.549917  1019

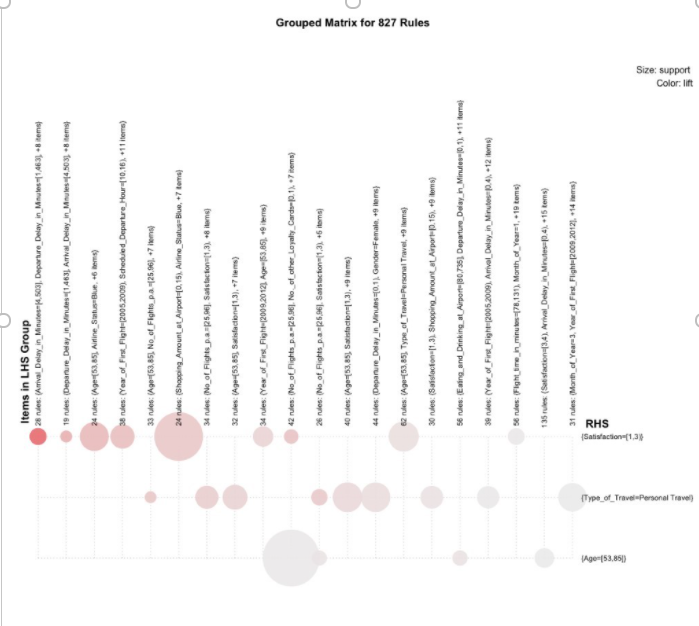
[16] {Price\_Sensitivity=[1,4],

      No\_of\_Flights\_p.a.=[25,96],

      Type\_of\_Travel=Personal Travel,

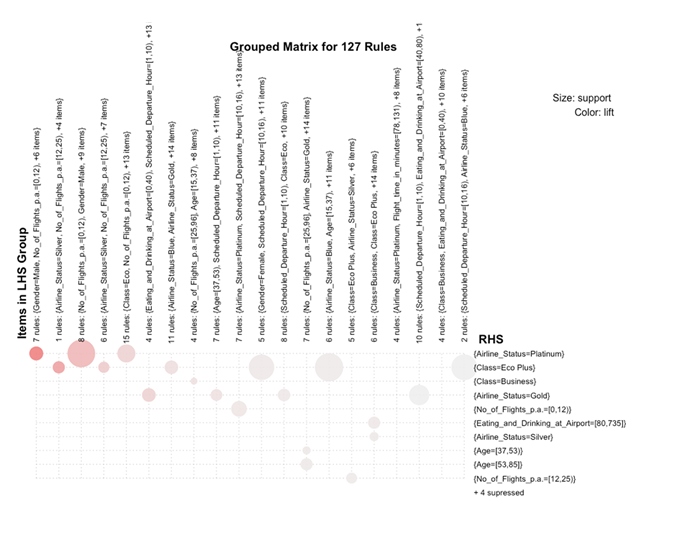
      No.\_of\_other\_Loyalty\_Cards=[0,1),

      Flight\_cancelled=No}              => {Age=[53,85]}                    0.1075493  0.8917749 0.1206014 2.558576  1030



We found by running these rules that:

Age group 53- 85 Have lower scores. The blue on the graph represents those that take a significant number of flights.



Logit Model

#Try a Logit

logit\_trainData<-lm\_trainData

logit\_testData<-lm\_testData

logit\_trainData$Satisfaction<-ifelse(logit\_trainData$Satisfaction>3,1,0)

logit\_testData$Satisfaction<-ifelse(logit\_testData$Satisfaction>3,1,0)

logit.model<-glm(formula =formula\_sig\_mod\_vif,data=logit\_trainData,family=binomial(logit))

summary(logit.model)

logit\_testData$predicted<-predict(logit.model,logit\_testData)

logit\_percent<-sum(ifelse(round(logit\_testData$predicted)==logit\_testData$Satisfaction,1,0))/length(logit\_testData$predicted)

logit\_percent

table(lm\_testData$Satisfaction,round(lm\_testData$predicted))

#Association Rules

Logit also didn't improve the accuracy much. It made it worse. We got only 33% accuracy

**Neural Net**

Finally, we tried neural net

#neuralnet

library(neuralnet)

nnData<-getSignificantDataAsSparse(surveyDataRaw)

nnData$Satisfaction<-ifelse(nnData$Satisfaction==TRUE,1,0)

nn\_trainData<- nnData[randIndex[1:cutpoint],]

nn\_testData <- nnData[randIndex[(cutpoint+1):nrow(nnData)],]

nn\_trainDataMatrix<-as.matrix(nn\_trainData)

nn\_testDataMatrix<-as.matrix(nn\_testData)

#nn\_formula <-formula\_sig\_mod\_vif<-paste("Satisfaction", paste(c(sig\_cols\_vif), collapse=" + "), sep=" ~ ")

nn\_model<- neuralnet(formula=Satisfaction ~.,data = nn\_trainDataMatrix,hidden = 12,lifesign = "minimal", linear.output = FALSE,threshold = 0.1)

plot(nn\_model)

nn\_predicted<-compute(nn\_model,nn\_testDataMatrix)

nn\_testData$predicted<-nn\_predicted$net.result

#table(nn\_testData$Satisfaction,nn\_testData$predicted)

percent\_nn<-sum(ifelse(round(nn\_testData$predicted)==nn\_testData$Satisfaction,1,0))/length(nn\_testData$predicted)

percent\_nn

The neural net gate the highest accuracy at 72%

# Visualization

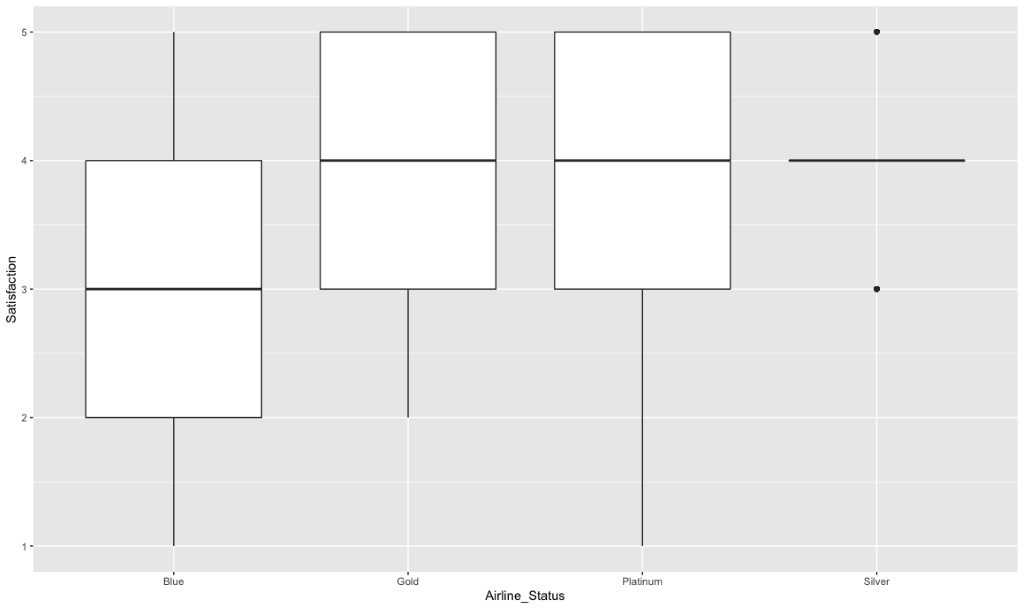
**Visualizations based on Significant LM vars**

**Airline Status**

#ggplot(surveyDataRaw,aes(Airline\_Status)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=Airline\_Status)) + geom\_boxplot()

# Except for Blue, the rest of the class have a higher average satisfaction score

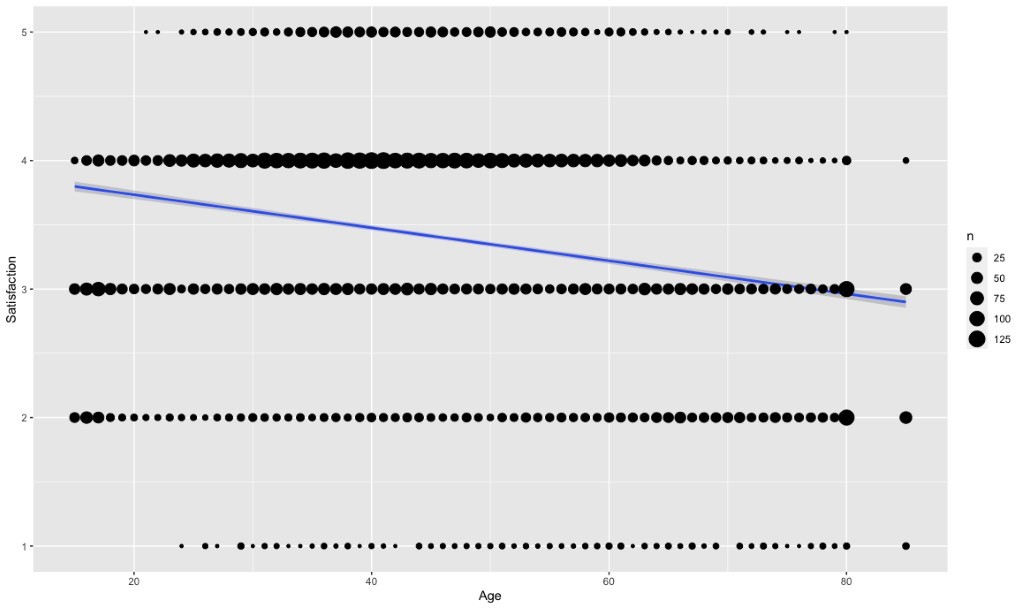


**Age**

#ggplot(surveyDataRaw,aes(Age)) + geom\_histogram(stat="count",line="blue")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=Age)) + geom\_smooth(method="glm") + geom\_count()

#We can see the satisfaction decreases as age increases

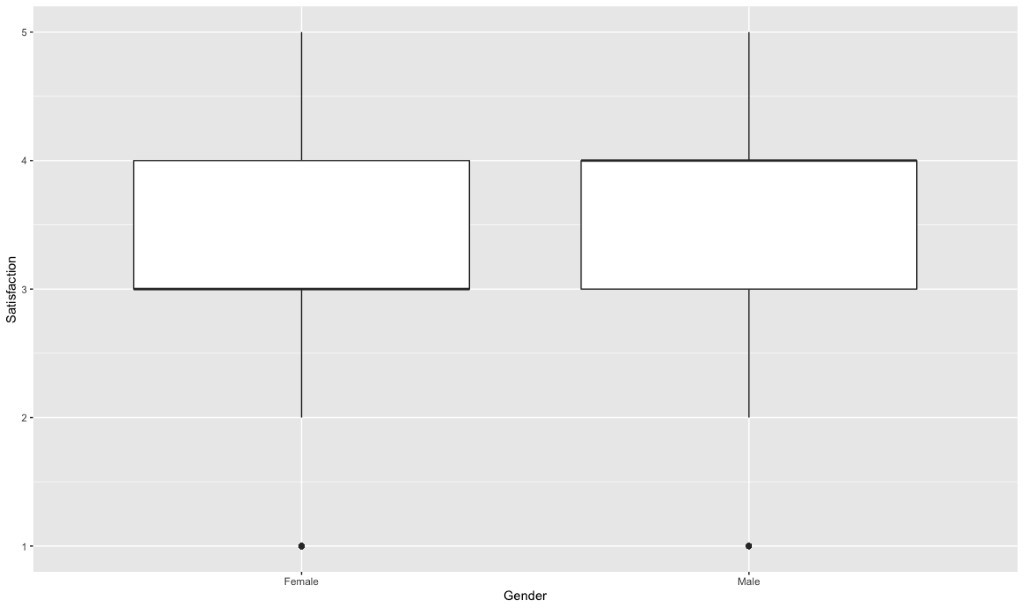


**Gender**

#ggplot(surveyDataRaw,aes(Gender)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=Gender)) + geom\_boxplot()

The mean and lower quatre is same for female, which means female tend to score as low

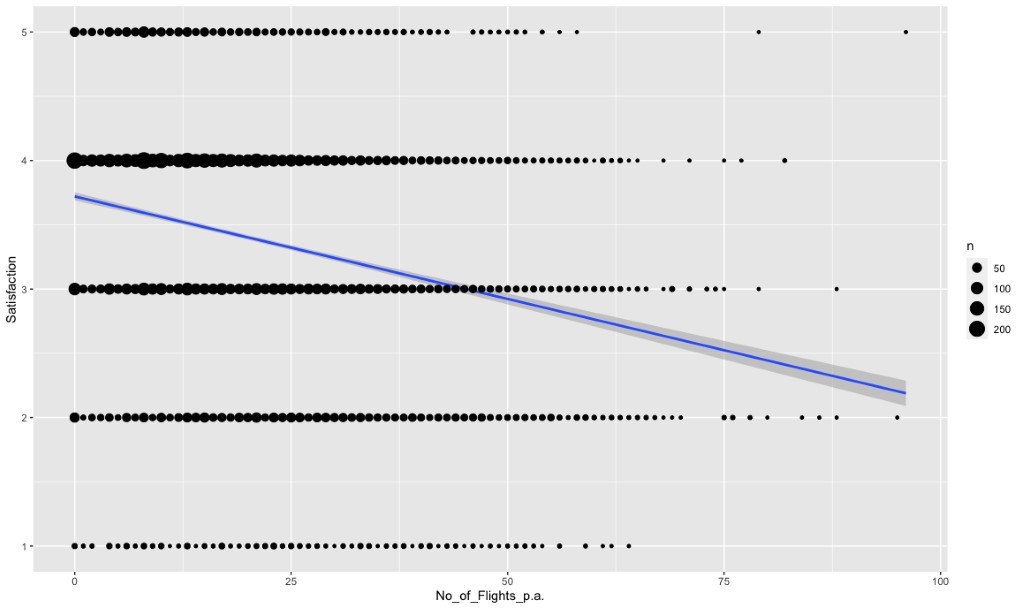


#No of flights

#ggplot(surveyDataRaw,aes(No\_of\_Flights\_p.a.)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=No\_of\_Flights\_p.a.)) + geom\_smooth(method="glm") + geom\_count()

# the higher the number of flight the lower the score is

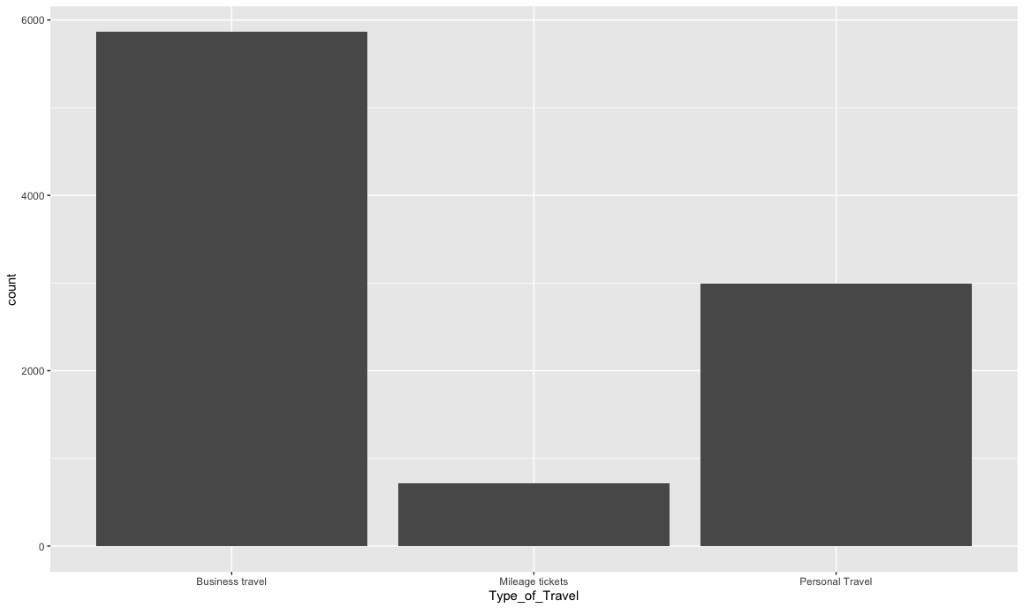


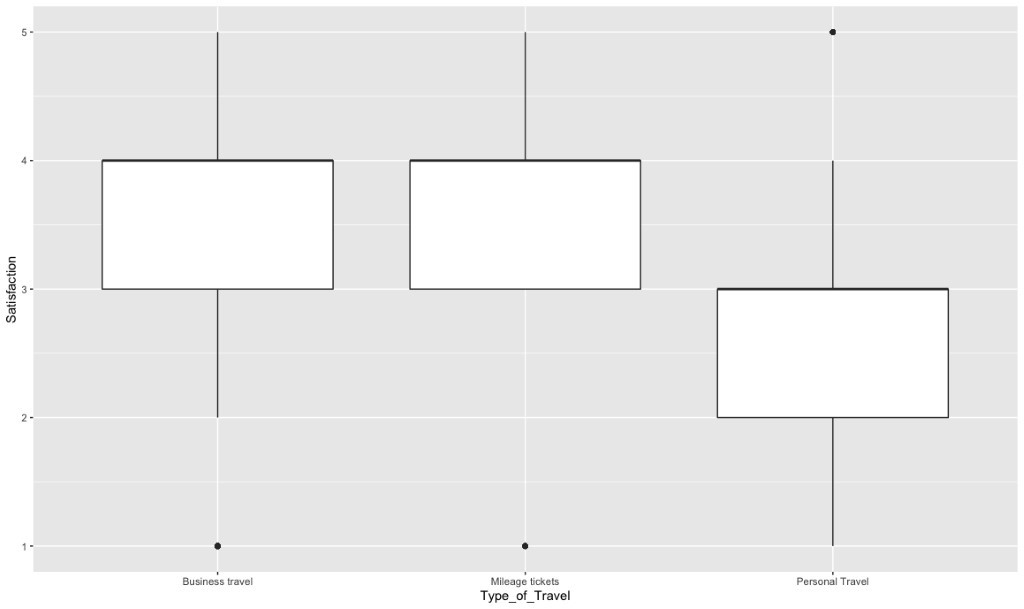
#Type\_of\_Travel

ggplot(surveyDataRaw,aes(Type\_of\_Travel)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=Type\_of\_Travel)) + geom\_boxplot()

Southeast has more business Travelers, and Personal travelers tend to score us less





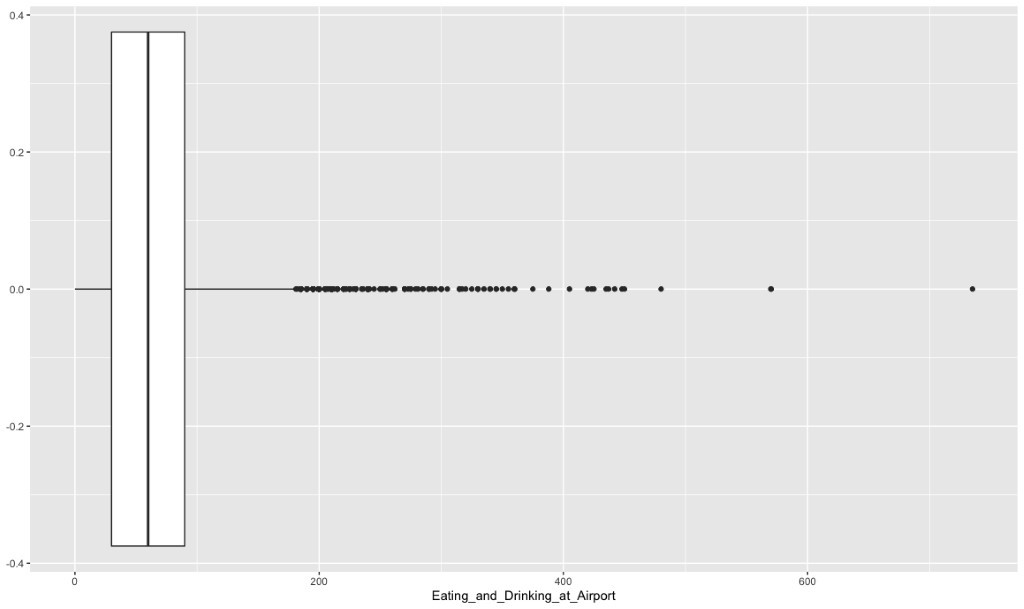
**Eating\_and\_Drinking\_at\_Airport**

#ggplot(surveyDataRaw,aes(Eating\_and\_Drinking\_at\_Airport)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(Eating\_and\_Drinking\_at\_Airport)) + geom\_boxplot()

#ggplot(surveyDataRaw,aes(y=Satisfaction,x=Eating\_and\_Drinking\_at\_Airport)) + geom\_smooth(method="glm") + geom\_count()

Concentrated on lower 10's and a lot of outliers

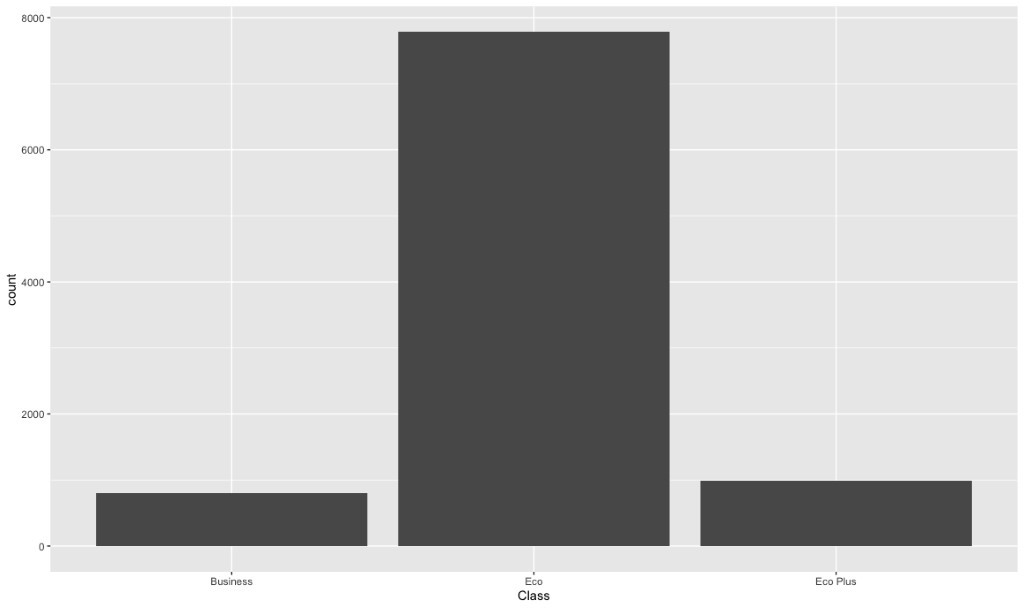


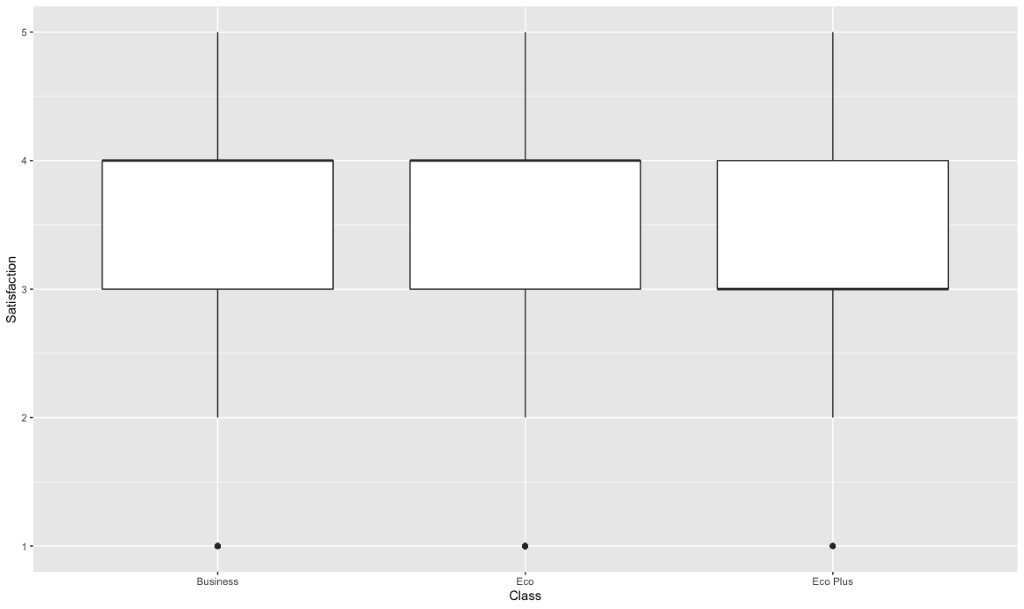
**Class**

ggplot(surveyDataRaw,aes(Class)) + geom\_histogram(stat="count")

ggplot(surveyDataRaw,aes(y=Satisfaction,x=Class)) + geom\_boxplot()

A lot of Economy Travelers & Economy Plus tends to score us lower

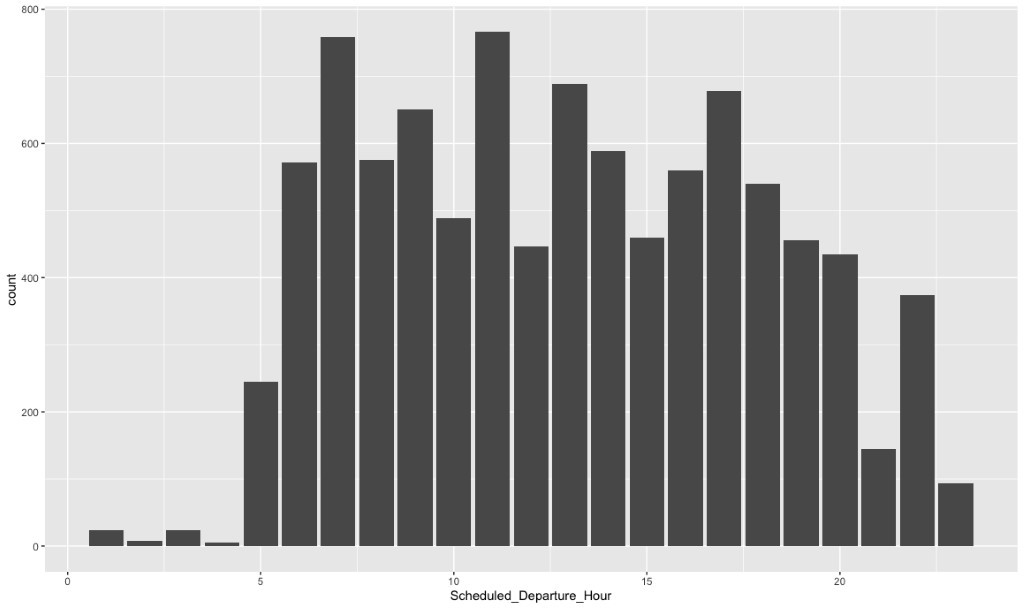




**Scheduled\_Departure\_Hour**

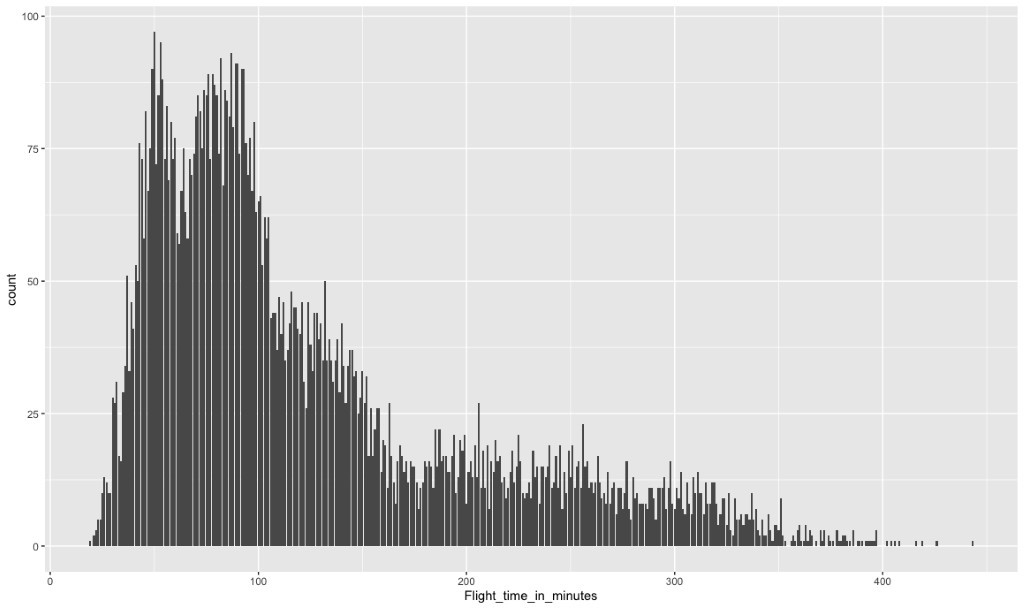
ggplot(surveyDataRaw,aes(Scheduled\_Departure\_Hour)) + geom\_histogram(stat="count")

Typical Business Hours with some service during Night



**Flight\_time\_in\_minutes**

A lot of short flights indicates more of a regional airline



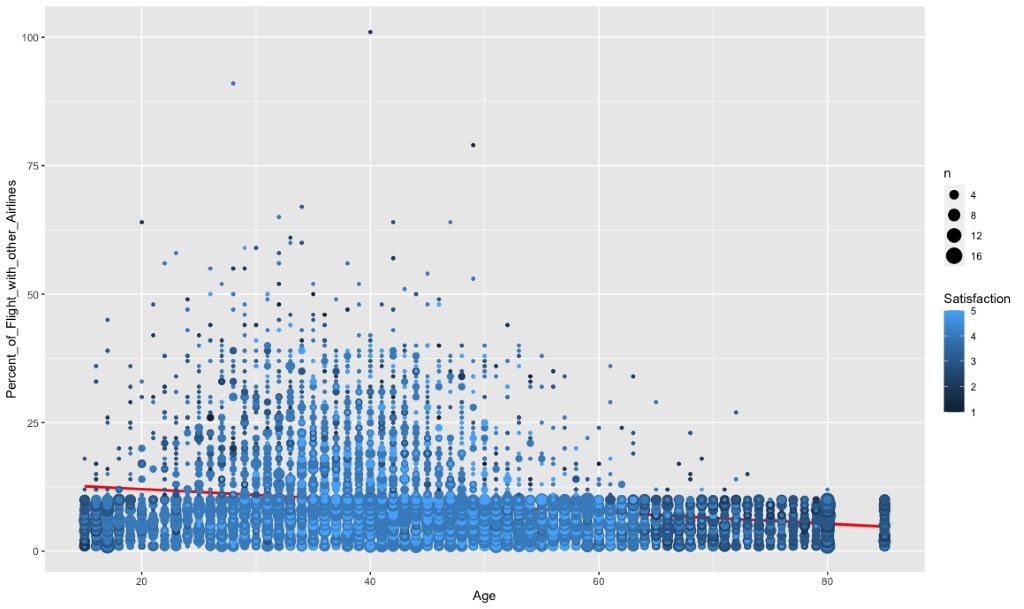
Moderating Variables

#**Age vs. Others**

#**Age vs Percent\_of\_Flight\_with\_other\_Airlines**

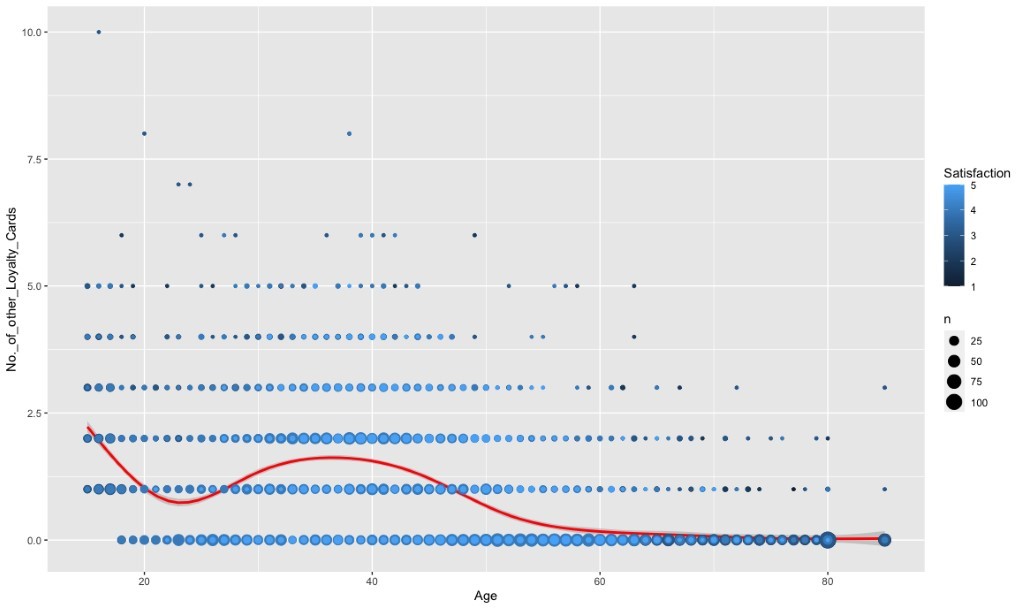
ggplot(surveyDataRaw,aes(x=Age,y=Percent\_of\_Flight\_with\_other\_Airlines,color=Satisfaction)) + geom\_smooth(color="red",method="glm") + geom\_count()

It appears all of the travelers took at least one other Flight



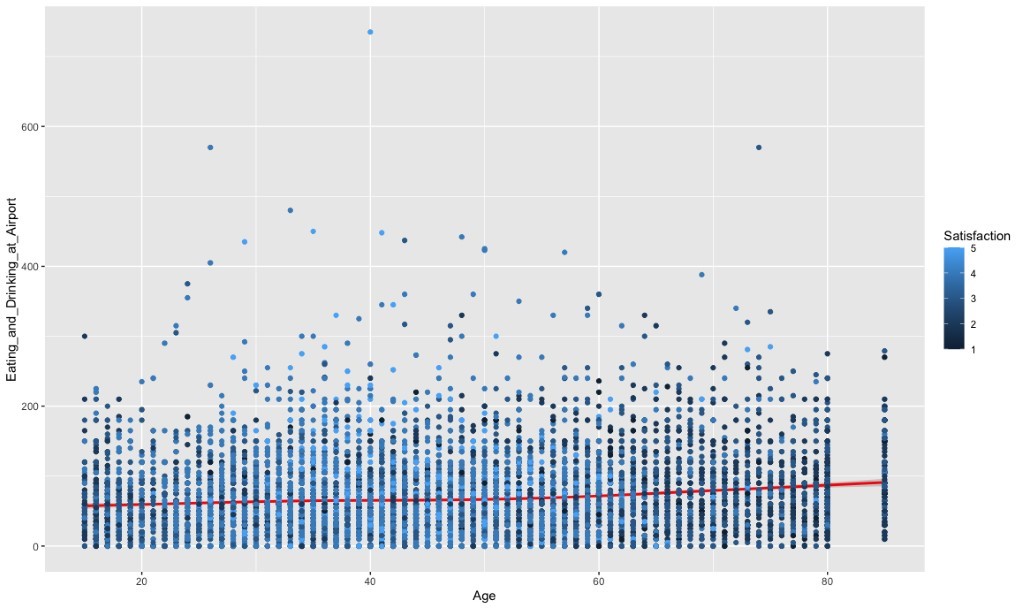
**Age vs No.\_of\_other\_Loyalty\_Cards**

ggplot(surveyDataRaw,aes(x=Age,y=No.\_of\_other\_Loyalty\_Cards,color=Satisfaction)) + geom\_smooth(color="red") + geom\_count()



#Age:Eating\_and\_Drinking\_at\_Airport

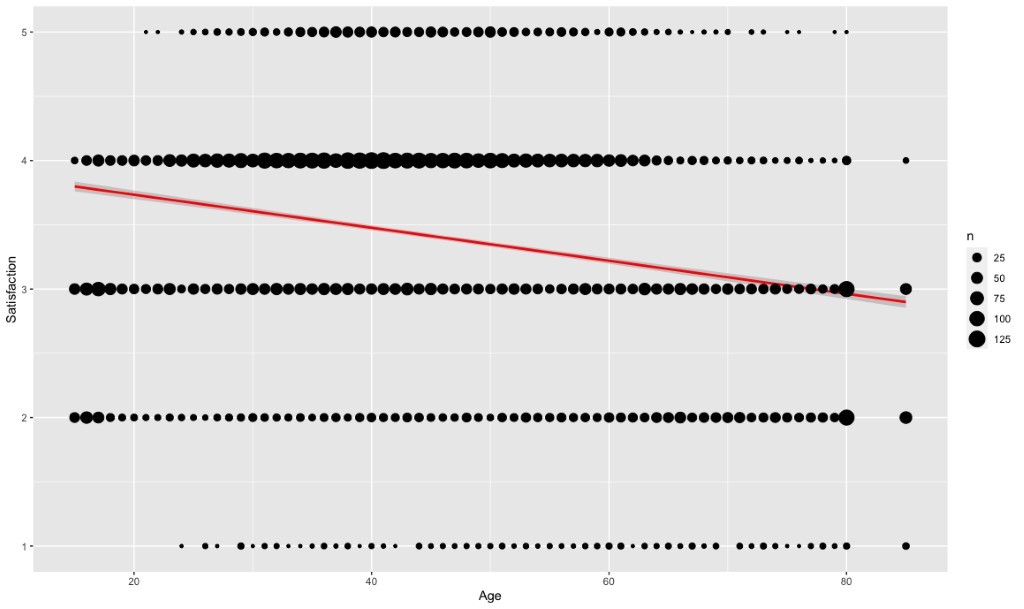
ggplot(surveyDataRaw,aes(x=Age,y=Eating\_and\_Drinking\_at\_Airport,color=Satisfaction)) + geom\_smooth(color="red") + geom\_point()



#Age vs Satisfaction

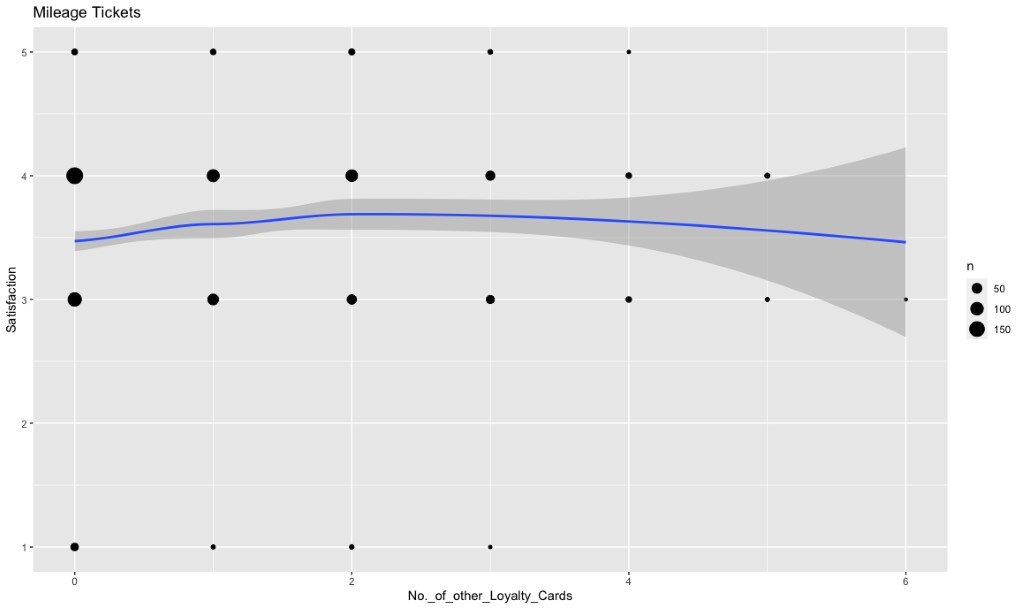
ggplot(surveyDataRaw,aes(x=Age,y=Satisfaction)) + geom\_smooth(color="red",method="glm") + geom\_count()

Satisfaction goes down with age



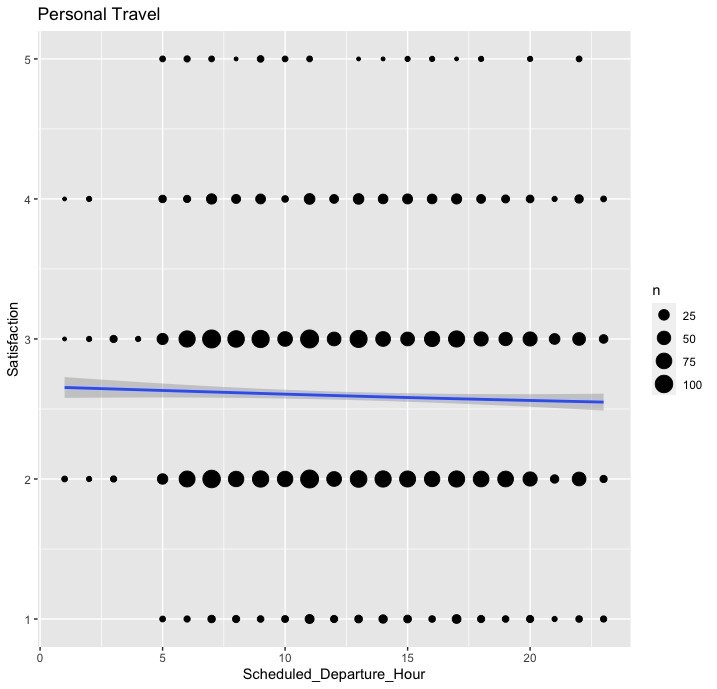
#Type\_of\_TravelMileage\_tickets:No.\_of\_other\_Loyalty\_Cards

ggplot(surveyDataRaw[surveyDataRaw$Type\_of\_Travel=="Mileage tickets",],aes(x=No.\_of\_other\_Loyalty\_Cards,y=Satisfaction))  + geom\_count() + ggtitle("Mileage Tickets") + geom\_smooth()



#Type\_of\_TravelPersonal\_Travel:Scheduled\_Departure\_Hour

ggplot(surveyDataRaw[surveyDataRaw$Type\_of\_Travel=="Personal Travel",],aes(x=Scheduled\_Departure\_Hour,y=Satisfaction))  + geom\_count() + ggtitle("Personal Travel")+ geom\_smooth()



# Business Questions Interpretation

The focus is to improve the Satisfaction scores of the airline customers for Southeast Airline Co. The customers must be heard, and the items viewed as negatives are adjusted or enhanced for future service provided by Southeast Airline Co. We have performed descriptive analysis and data mining to analyze the survey data set to answer the following questions:

1. Based on Satisfaction, which airlines are performing well, and which are performing poorly?
   * The data showed that all airline companies' mean satisfaction scores ranged between 3.30 and 3.53.
   * Southeast Airline Co's average score was 3.39
   * All airlines' average scores were within 0.24 of each other; therefore, no significant recommendations could be generated from this question.
   * Chart, scatter chart

     Description automatically generated
2. How does Airline Class affect customer satisfaction based on age group?
   * It was determined that Class and Age together did not significantly affect the satisfaction scores for Southeast Airlines. However, separately class groups economy and economy plus along with age did prove to be significant.
   * The data did show that the age group of 35-45 tend to score 3.5 and higher regardless of class affiliation.
   * Of Southeast's 9577 customers, close to 6000 of those travelers are business travelers. With 63% of their travelers flying business class, this question will not provide any useful airline recommendations.
   * Table

     Description automatically generated
   * Chart, scatter chart

     Description automatically generated
3. Could eating and drinking at the airport before a flight affect the satisfaction?
   * Based on the personal travel experience, it was our flight experience that flights were more enjoyable when there was no eating or drinking at the Airport.
   * The data showed that eating and drinking at the Airport was slightly significant (P-value- 0.013149) in determining Satisfaction.
   * There was a slight trend upward in satisfaction scores in the data for those customers who had a quantity of 200 or more.
   * More data is needed to understand the quantity count regarding the drinking and eating category.
   * 
   * Chart

     Description automatically generated
4. What times of operation are critical in affecting Satisfactiontion score?
   * Scheduled\_Departure\_Hour is significant with a P-Value of 0.001064.
   * We split the data into two groups—customers who rated 3.0 and higher and 3.0 and lower.
   * The data showed that Scheduled\_Depature\_Hour had lower satisfaction scores of 2.5 or lower at two critical times 3:00 A.M. and 2:00 P.M.
   * 
   * Chart, line chart

     Description automatically generated
   * Chart

     Description automatically generated
5. What elements contribute to the highest Satisfaction Scores?
   * The P-values for the following are significant. Age -2.46E-03, Airline\_StatusGold 4.37E.01, Airlinestatus\_statusPlatinum 3.20E.01, Airline\_StatusSilver 3.96E.01, GenderMale 1.36E.01, Age:No\_of\_other\_loyalty\_cards 6.36E.04, Scheduled\_depature\_Hour 4.83E.03.
   * Age was significant because you could see that these 53-85 take the most significant number of flights. However, they provide low satisfaction scores.
   * Arrival delay showed a significant negative correlation across all the loyalty programs.
   * The data for the elements of Age and Loyalty card show that 53-85 have the most loyalty cards and give higher satisfaction scores. However, 53-85 who take the most significant number of flights overall give lower satisfaction scores.
   * Table

     Description automatically generated
6. How does Southeast Airlines Co. perform on satisfaction among customers that have never flown before?
   * Southeast had 471 1st time flyers submit satisfaction ratings with a mean overall score of 3.45.
   * The mean of the satisfaction scores among 1st-time flyer goes up with age.
   * Chart, histogram

     Description automatically generated
   * Chart, line chart

     Description automatically generated
7. How does Southeast Airline Co. perform based on the type of Travel?
   * Of the 9577 travelers on Southeast Airlines during the 1st quarter of 2014, 61% were business travelers with a mean satisfaction score of 3.78. 7% were mileage ticket travelers with a mean satisfaction score of 3.56, and Personal travel tickets made up the final 32% with a mean satisfaction score of 2.59.
   * Chart

     Description automatically generated
   * Chart, scatter chart

     Description automatically generated

Recommendations:

After understanding the data provided to us by the Airlines, we as a group have come up with the following recommendations for Southeast Airlines Co. to use to help improve satisfaction scores in the future. If these recommendations are put into effect, the Southeast Airlines Co. will see a significant increase in their future satisfaction scores.

**Recommendation 1**: Create incentives for the age population of 53 and older.

The data showed that customers took a more significant number of flights at the age group of 53 and older compared to the age group of 53 and younger. Additionally, Southeast had better satisfaction scores from the same age group who also participated in multiple loyalty card programs. We recommend improving loyalty programs for older populations and granting more rewards as the traveler takes repeat flights on Southeast.

**Recommendation 2**: Critical Times to consider - 3 AM and 2 PM

The data provided insight into specific timeframes in which customers are more likely to give lower satisfaction scores at 3:00 A.M. and at 2:00 P.M. If there is a decision to be made regarding which Flight should be delayed the company should not choose a flight at either of those times. Southeast has a large population of business travelers, which correlate to those travelers trying to get to a meeting 1st thing in the morning and trying to get home before the end of the day. Additionally, due to the high demand from business flyers, we also recommend that the airline see if they have enough Flight available 1st thing in the morning and before the end of a typical business day.

**Recommendation 3**: Make subsequent flight experiences better for customers.

The Airline is doing a good job based on satisfaction scores for 1st time flyers with a mean satisfaction score of 3.45. The satisfaction scores for subsequent flights decreases after the 1st initial Flight. Implement a better loyalty program for frequent flyers, this will provide incentives for customers to come back for return flights. Take a survey of return customers on what features, they like best and which they like least. Make changes according to those results.

**Recommendation 4**: Find a way to identify personal airline travelers and provide additional benefits.

61% of the customers that are flying on your airline are business customers who have given a mean satisfaction score of 3.78. However, 32% of the population is personal travelers who have given the company a mean satisfaction score of 2.59. The personal ticket customers need to be provided some incentives, better customer service, or improved conditions while traveling on your airline. We recommend putting identify marks on personal travelers' tickets boarding passing so all airline employees can provide above and beyond customer service to those customers. Also, follow up with those personal travelers and asked them to list the top 3 things they would change about their experience with your airline.

**Recommendation 5**: Try to avoid Arrival Delays

From our significant variables, we can see that the arrival delay contributes more negatively to the Satisfaction, so even if we depart late, try to make up during the flight and arrive on time.