**IST 687**

**Final Project**

Group3

Ishita Joshi, Leah Singer, Wei Mu, Xin Sun, Zhenlei Liu

# Introduction

In order to improve the satisfaction of passengers, this project investigates the data from 194,833 passenger surveys of 29 attributes. Our major findings are:

* The important demographic characteristics of the passengers
* Influential factors which contribute to customer satisfaction
* Impact of the arrival delay on airline customer satisfaction
* The difference in satisfaction between different types of travelers and airline status

# Business question

* Which demographics of people should we focus on to increase the overall satisfaction?
* Is there any difference between different partner companies in customer satisfaction?
* Is there a significant difference in customer satisfaction between different groups of age?
* How is the company’s service for different groups of Airline Status?
* How does the arrival delay affect customer satisfaction?

# Procedure

## Data Cleansing and preparation

We found out that the major reason for a record to contain missing value is that the flight is canceled. Table 1 compares whether flights are canceled or not between complete cases and incomplete cases:

|  |  |  |
| --- | --- | --- |
| **Table 1** | | |
| Missing value in a record? | Flight Cancellation | Number of records |
| No | Not canceled | 190720 |
| Yes | Not canceled | 510 |
| Yes | Canceled | 3603 |

If the flight is not canceled, in most cases the records are complete. If the flight is canceled, the record always contains missing values, because fields like “Arrival Delay” does not make sense in this situation. By comparing the average satisfaction score between canceled or not, we see that when a flight is canceled, the customer has a lower satisfaction score.

|  |  |
| --- | --- |
| **Table 2** | |
| Flight Cancelled | Average satisfaction score in the group |
| Canceled | 3.11 |
| Not canceled | 3.39 |

Only 1.85% of the customers experienced a flight cancelation problem. Because most of the missing values are caused by the cancellation of a flight, in all the following analysis we are going to remove all the 4113 records (out of 194833 total records) which contain missing values. The incomplete records are only 2.11% of total records, so they do not affect the following analysis.

The following attributes were dropped before doing the following analysis:

|  |
| --- |
| * Day.of.Month |
| * Year.of.First.Flight |
| * Flight.date |
| * Partner.Code |
| * Partner.Name |
| * Origin.City |
| * Origin.State |
| * Destination.City |
| * Destination.State |

## Get to know the data

After the cleaning of the data, we were ready to obtain the sample demographics of the raw dataset.

**Demographic characteristics**

Table 3 provides the frequency and percentage of the four characteristics which can be used to categories the passengers. The Blue and Silver airline status was dominated in all the customs which is 68.3% and 20.1%, respectively. The gender of the custom was uneven that more female customs have participated in the survey. 61.6% of the passengers are business travel, and 30.5% of the passenger are personal travel. However, personal travelers were more important to improve passenger satisfaction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 3** | | | |
| Characteristics | | Frequency | Percentage (%) |
| Airline Status | |  |  |
|  | Blue | 130525 | 68.3% |
|  | Silver | 38365 | 20.1% |
|  | Gold | 16124 | 8.4% |
|  | Platinum | 6208 | 3.2% |
|  |  |  |  |
| Gender | |  |  |
|  | Female | 107614 | 56.3% |
|  | Male | 83608 | 43.7% |
|  |  |  |  |
| Type of Travel | |  |  |
|  | Business travel | 117884 | 61.6% |
|  | Mileage tickets | 14934 | 7.8% |
|  | Personal Travel | 58404 | 30.5% |

**Satisfaction Comparison between Partner Companies**

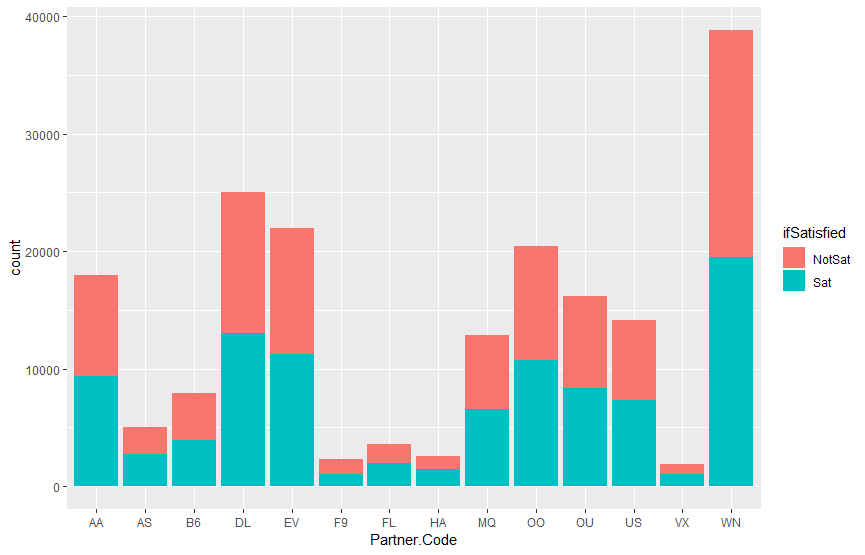


Figure Airline Partner Satisfaction Performance

Each airline segment in the dataset is partnered with one of the 14 different wholly- and partially-owned subsidiary companies. The bar plot shows that the responses from different partner companies are significantly different from each other. However, the average satisfaction scores for different partner companies are almost equal, so it is reasonable to use the entire dataset for the following analysis without separating into different companies.

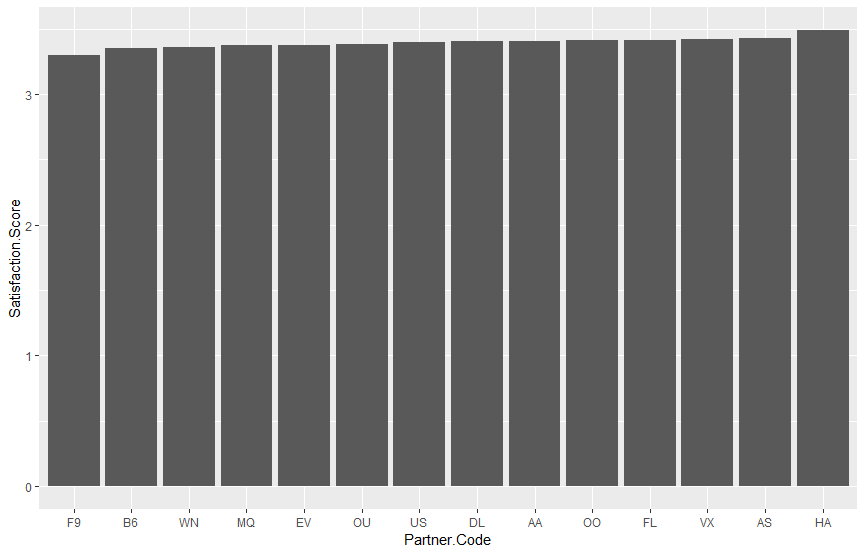


Figure Airline Partner Satisfaction Score

## Satisfaction Score

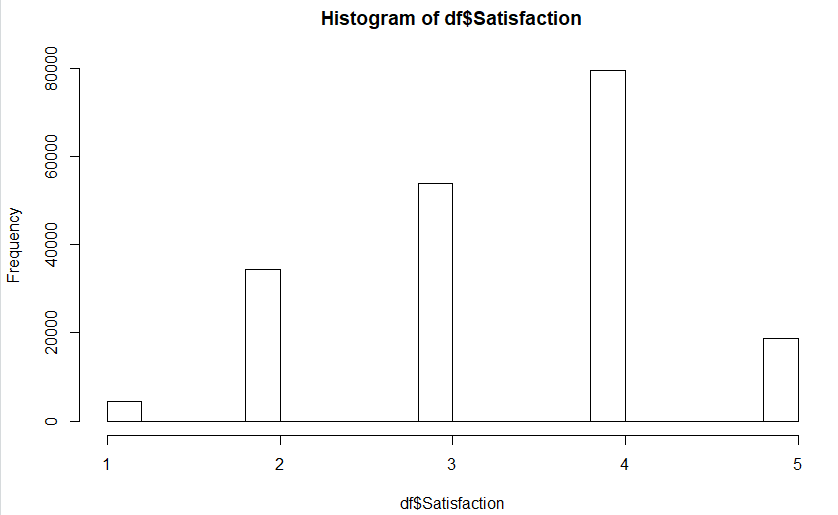


Figure Overall Satisfaction Distribution

The histogram of satisfaction score suggests the customer’s understanding of the satisfaction score: 4 is satisfied, 5 is very satisfied (which is much lesser than 4), and a score below 3 is not very satisfied. There are only 7 customers answered 2.5, 1 customer answered 3.5, and 3 customers answered 4.5, so these records can be ignored. By splitting the response variable into two classes: “satisfied” and “not satisfied”, some machine learning algorithm for classification problem can be applied to the dataset, such as decision tree and logistic regression.

By defining “satisfied” as >=4 and “not satisfied” as <4, the result could show a general picture of the response. The number of satisfied and not-satisfied responses are balanced in this situation:

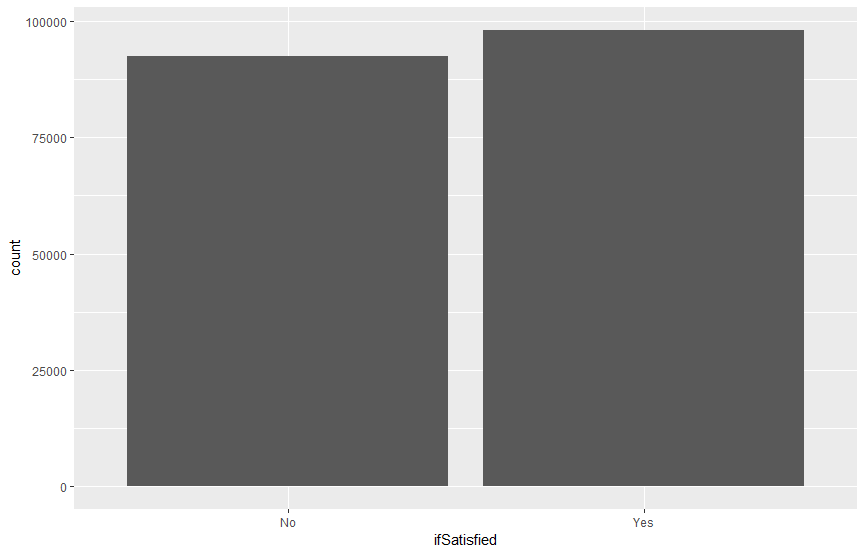


Figure Satisfaction Divided by Satisfaction > 3

By defining “satisfied” as =5 and “not satisfied” as <5, we could focus on the more interesting part of the data: why the customer is “very satisfied” or “angry” compared to other general responses.

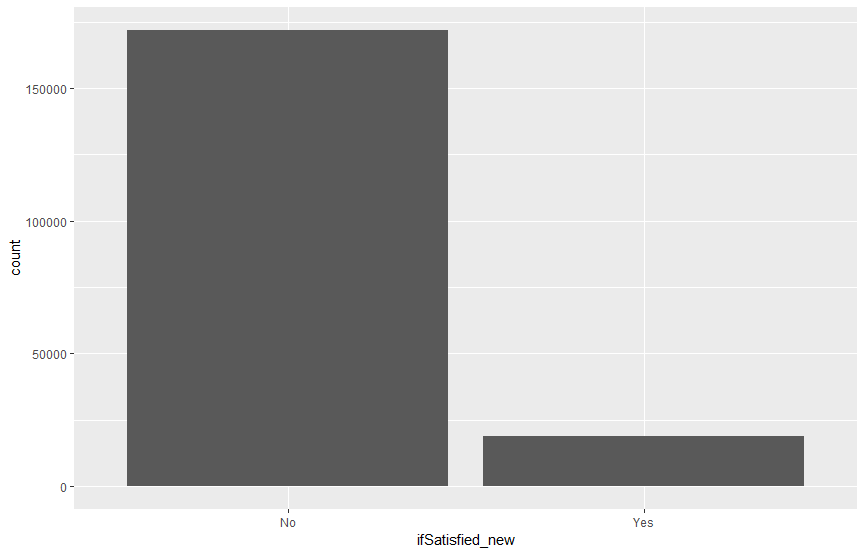


Figure Satisfaction Divided by Satisfaction >4

# Data Analysis and Discussion:

## Decision Tree Analysis using the whole dataset

Decision Tree is a supervised machine learning algorithm covering classification and regression problem. In a the classification problem, it automatically selects a set of variables that gives more information on whether the response variable belongs to a certain class. For this airline analysis problem, it chooses factors that are important on whether the customer is satisfied or not.

In this part of the analysis, we define “Satisfied” as a satisfaction score over or equal to 4, and “Not satisfied” as a satisfaction score lower than 4. In the data preparation part, we showed that the dataset is balanced between these two class, so it is safe to continue without concerns about overfitting to an unbalanced dataset.

First, we randomly split the dataset into 70 percent training data and 30 percent testing data. We first run decision tree algorithm on the training dataset, examine the tree we get, and then test in on the testing dataset. By using ‘Gini’ as the splitting index, and complexity parameter of 0.01, we get the following decision tree:

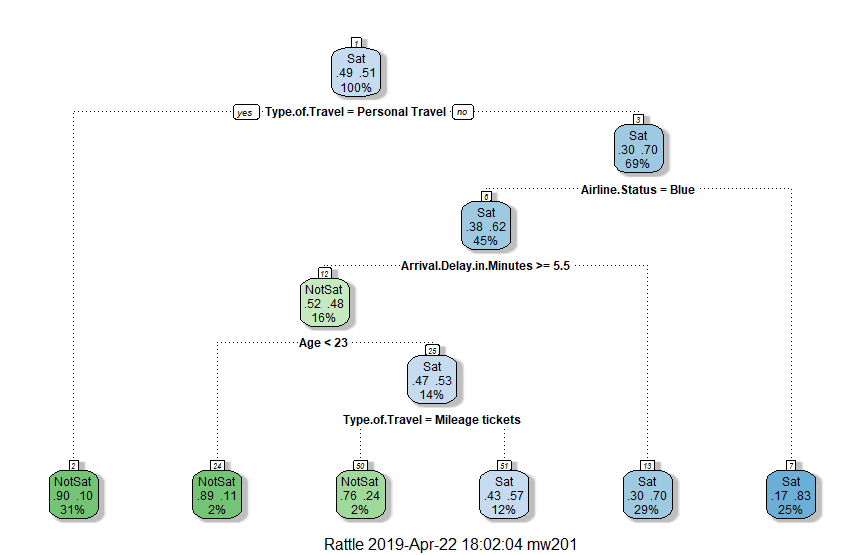


Figure Decision Tree

For each split, the left-hand leaf node means the splitting condition is true, while right leaf node means false. The decision tree shows that a personal traveler has a probability of 90% to be not satisfied. Non-personal young travelers (less than 23 years old) with Blue airline status are very likely to be not satisfied if the arrival delay in minutes in more than 5.5 minutes. All the variables in this tree are important compared to other irrelevant variables in the dataset.

By setting cp=0.005, we further expanded the tree like the following:

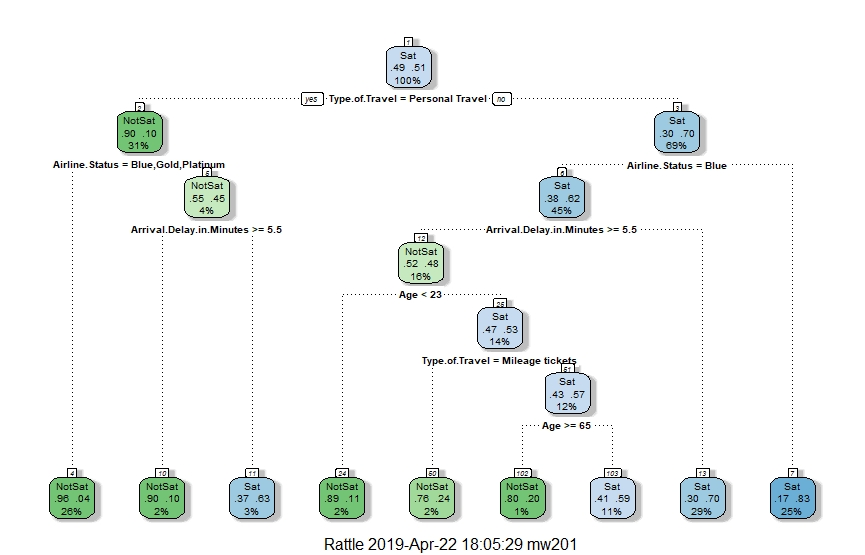


Figure Extended Decision Tree

It shows that whether arrival delay is more than 5.5 minutes is also important for the Silver type of customers who are going for personal travels. A business traveler who is more than 65 years old and has airline status of blue is very likely to be not satisfied when arrival delay is more than 5 minutes.

As for the business suggestion for the airline company, the company should improve its services to personal travelers. Whether arrival delay is greater than 5 minutes matters for the Silver type of personal travelers and Blue type of non-personal travelers. In all the Blue type of customers, young (<23 years old) customers and old (>=65 years old) Business travelers are not satisfied, so the company should focus more on improving the experience of these two types of customers.

The plot of the complexity parameter is as follows:

A screenshot of a social media post

Description automatically generated

Figure Influence of cp value

It shows that it does not make much difference when cp is lower than 0.01. The ROC curve for cp=0.01 is as follows:

A screenshot of a cell phone

Description automatically generated

Figure

The AUROC score is 0.78 when the model is tested on the testing data set. It means that the model is generally a good model. Conclusions can also be verified by descriptive analysis, such as the following graph confirms that personal travelers are not satisfied.

A screenshot of a cell phone

Description automatically generated

Figure Satisfaction Distribution for Type of Travel

## Principal Component Analysis

Principal Component Analysis is an effective method to explore the numeric variables. If a numeric variable has higher (positive or negative) correlation score with the first few principal components, then this numeric variable is more likely to be important in modeling.

The relative importance of principal components are as follows:

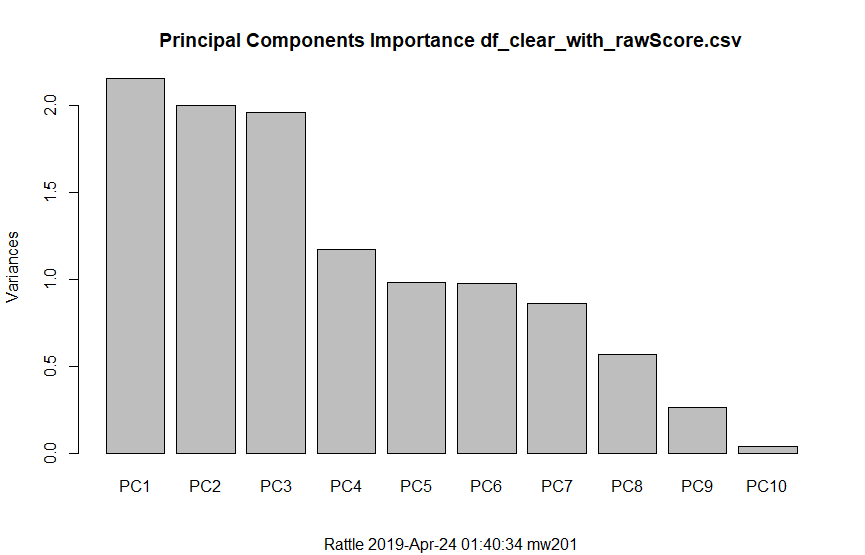
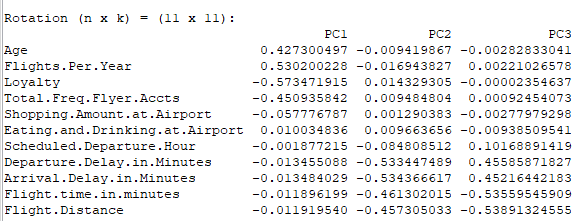


Figure Principle Importance Analysis

The R result of the first 3 principal components are as follows:



It can be seen that “shopping amount at airport”, “eating and drinking at the airport”, “scheduled departure hour” may not be very relevant in our analysis. Then we may focus on other more important numeric variables.

## Gender & Age Analysis

First, we looked at the difference of the number of responses between genders.

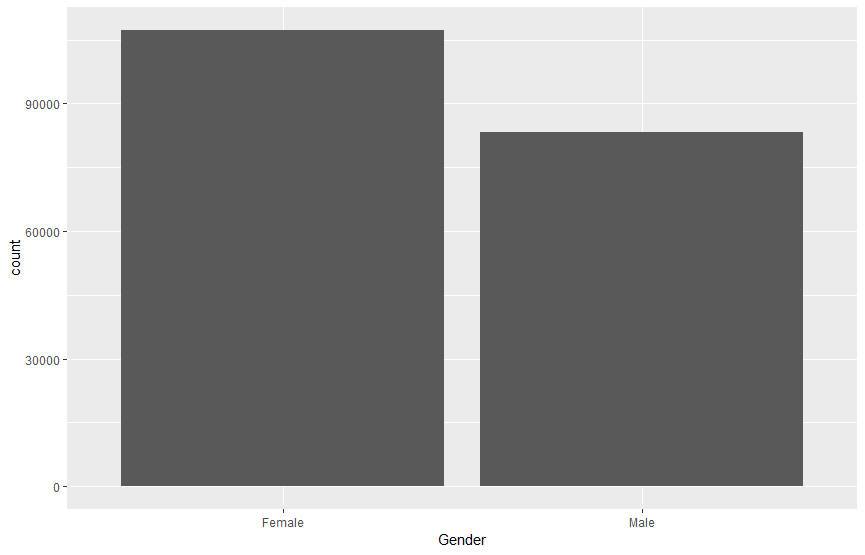


Figure Gender Distribution

Responses from female customer are significantly more than from male customer, so it is an unbalanced dataset.

By defining “satisfied” as >=4 and “not satisfied” as <4, we get the following visualization of stacked bar plot:

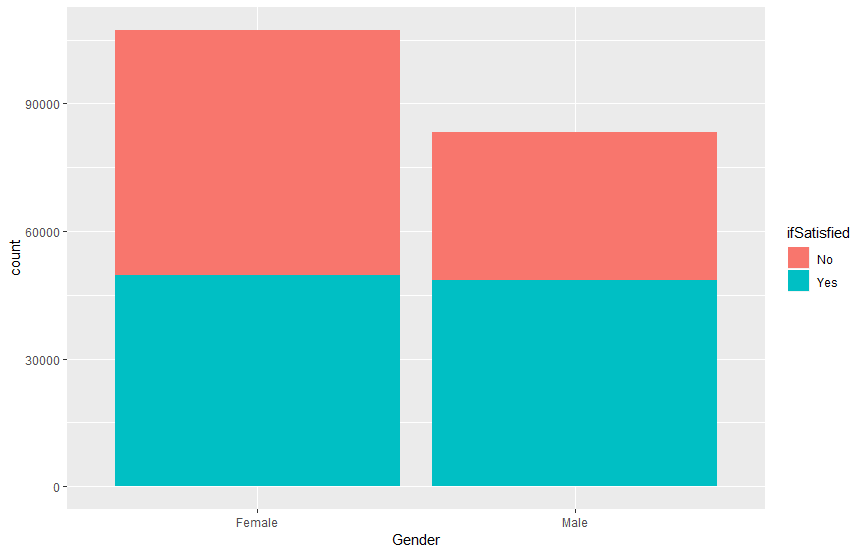


Figure Satisfaction Ratio of Gender

By observing the graph above, we know that male customers are more likely to be satisfied.

The following is the stacked histogram of age:

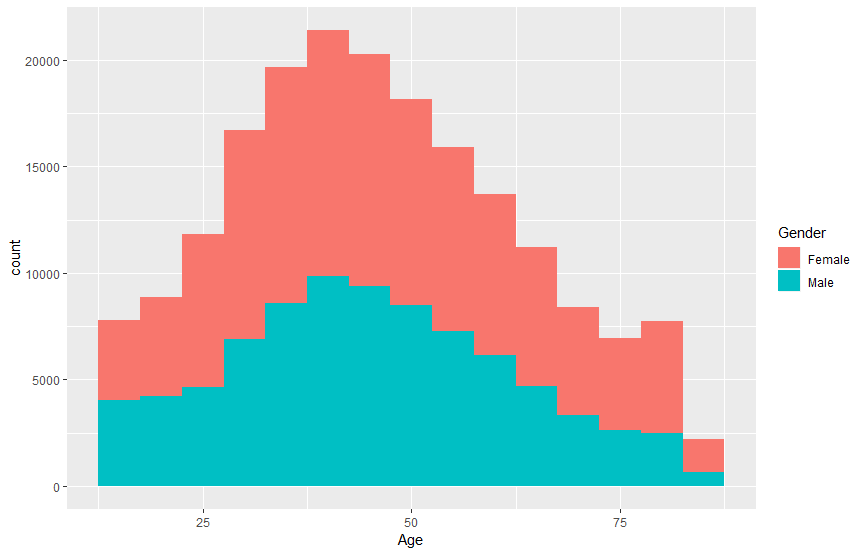


Figure Overall Age Distribution and Relation Between Satisfaction

The data is centered in age between 30 and 60. So if we define Age < 30 as "Young", 30 <= Age < 60 as "Middle" age, Age >= 60 as "old" age, then the following stacked bar plot show the age and gender distribution.

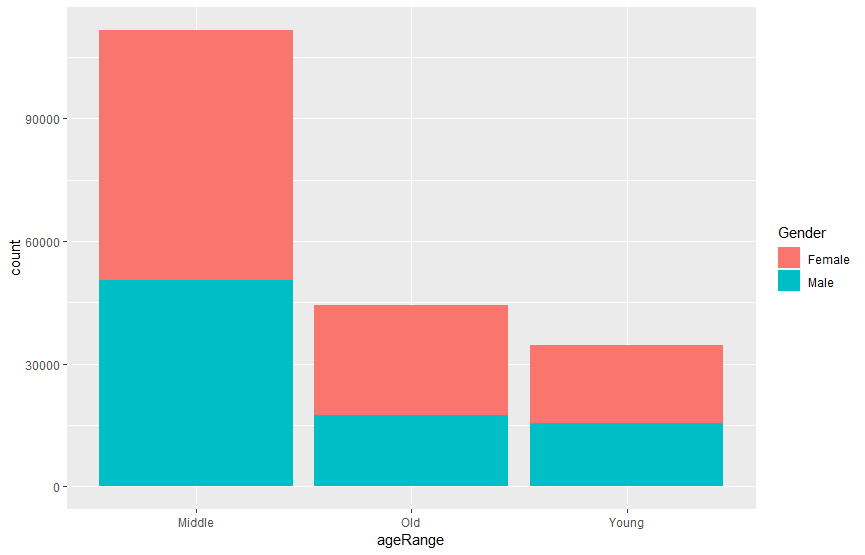


Figure Gender Distribution of Different Age Level

There are 34527 Young age responses, 111744 Middle age responses, 44449 Old age responses.

If we change the variable from the number of responses versus age-gender to the number of responses versus age-if Satisfied, the following graph shows the difference between different age range. Old customers (Age >=60) are likely to give lower satisfaction score, while Middle age customers are more likely to give higher satisfaction score.

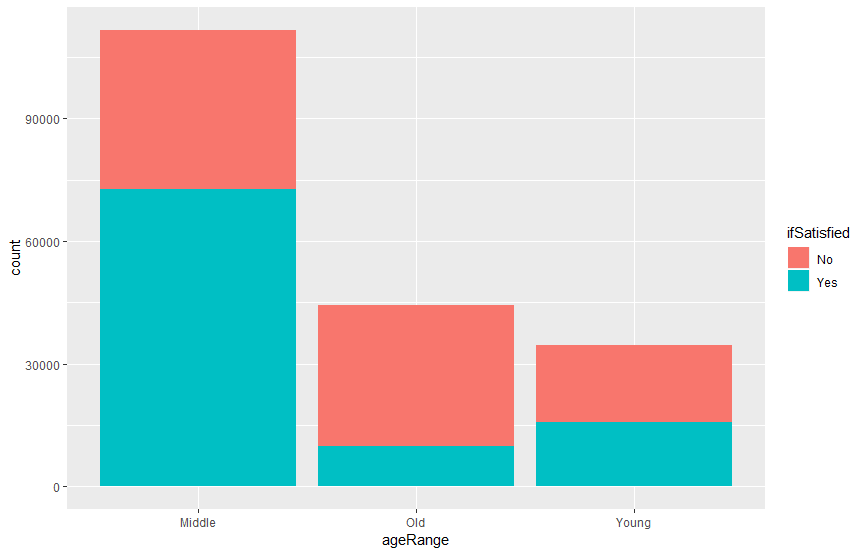


Figure Satisfaction Distribution of Different Age Level

A linear regression shows a similar result. The coefficient of Age is negative, and the coefficient of I(Male) is positive. The overall regression is significant with p-value < 2.2e-16. Both Age and I(Male) are significant with p<2e-16. The low R square is as expected because there are many other factors which determine the overall satisfaction. So the interpretation is:

With other factors constant, if the customer has a higher age, the customer gives a lower satisfaction score.

If everything else is the same, male customers give higher satisfaction score.

In general, the conclusion is the same as conclusions from the visualized graphs above.

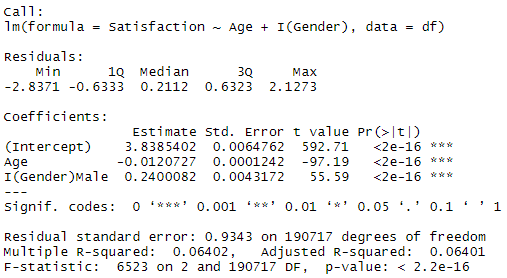
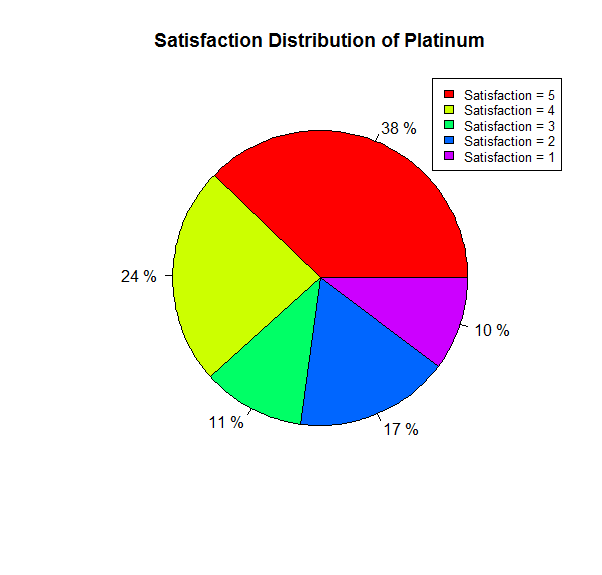
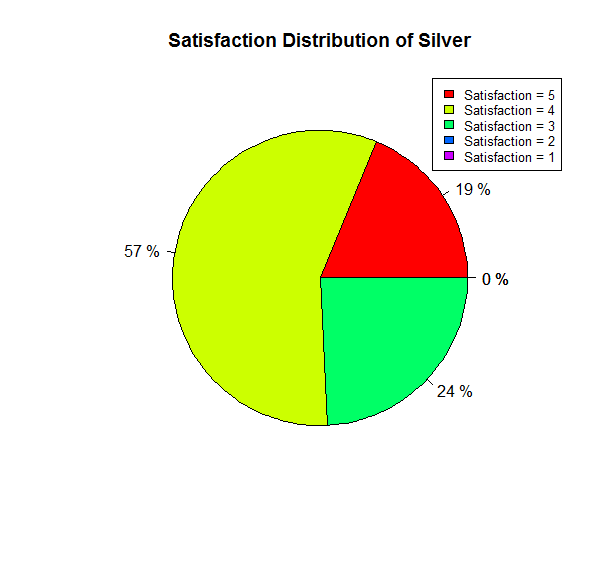
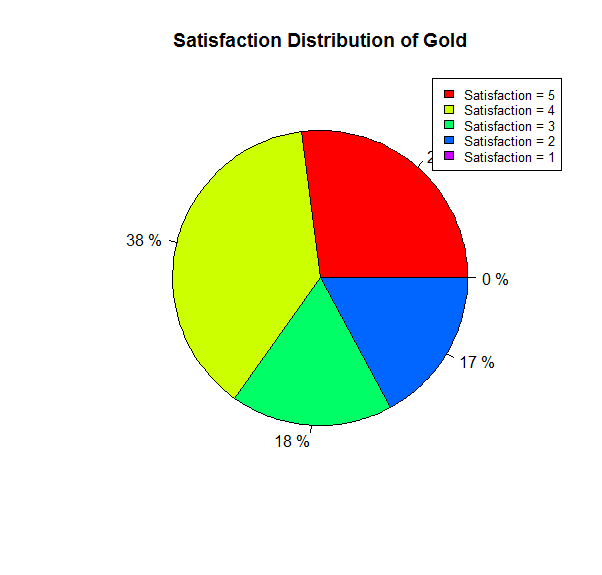
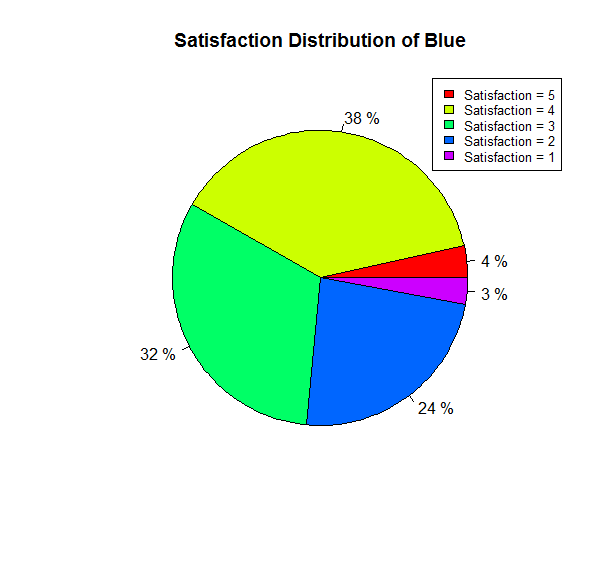


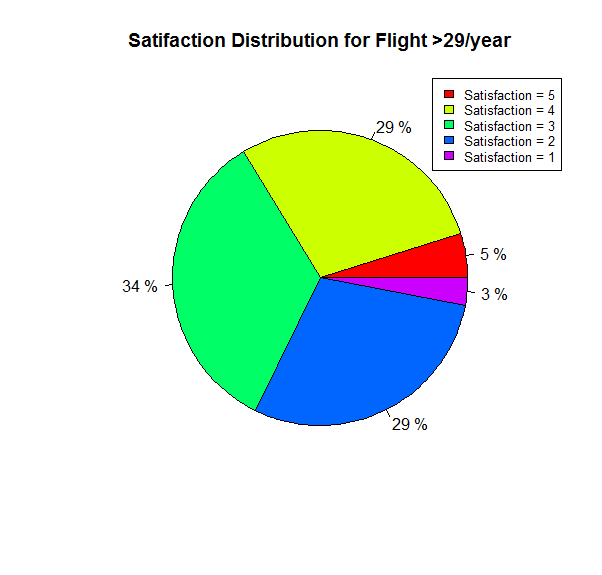
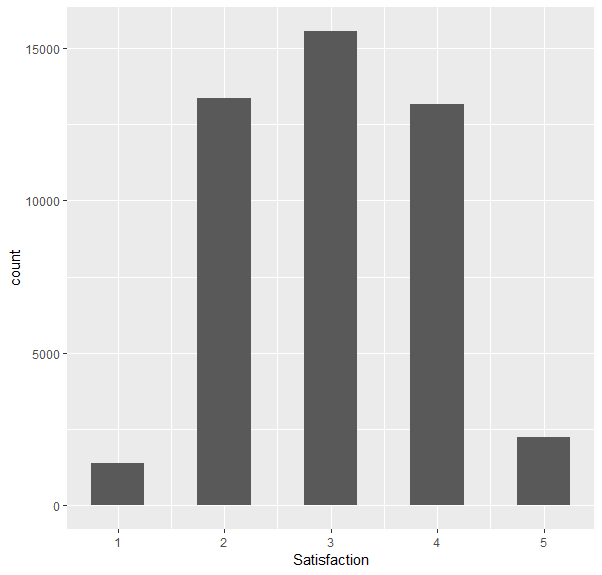
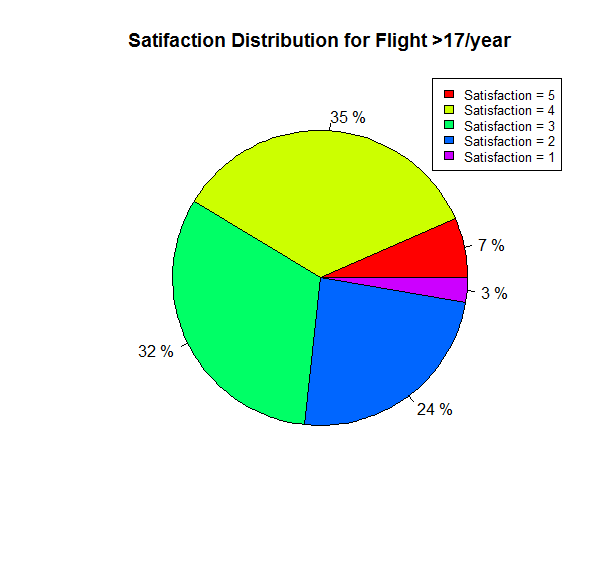
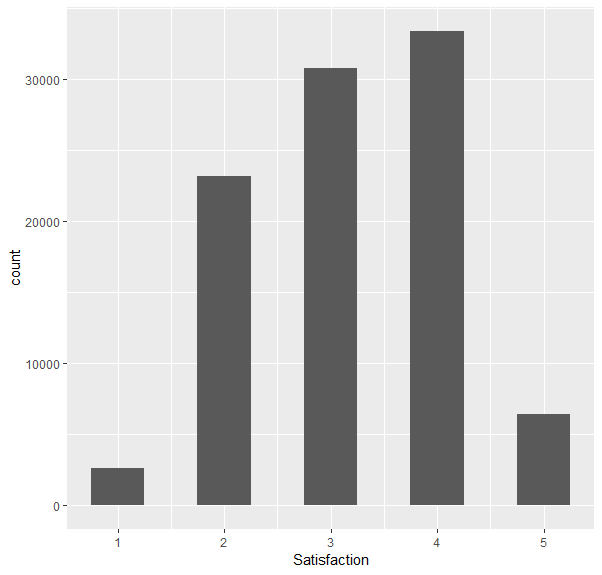
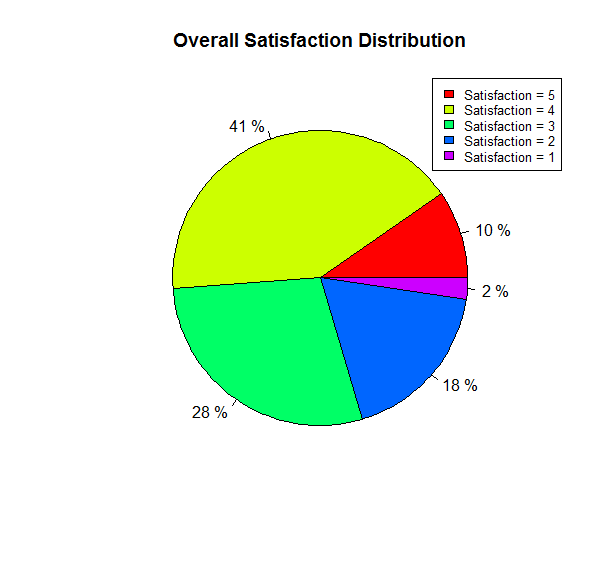
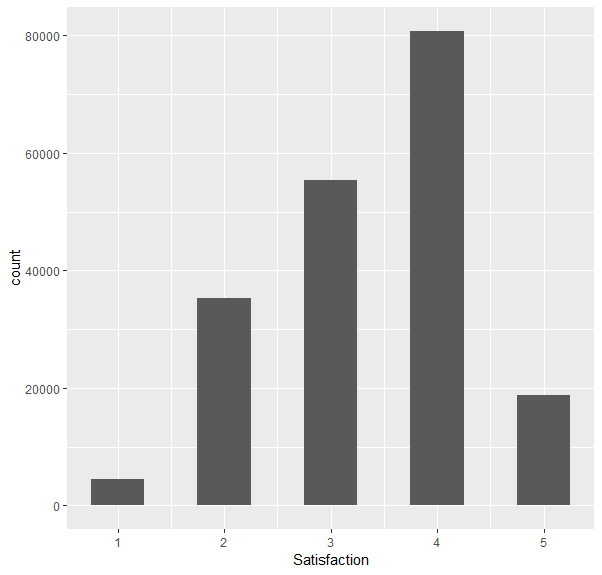
Figure 17 Satisfaction Ratio of Different Airline Status



## Airline Status Analysis

Figure 17 shows the satisfaction distribution of different airline status. The Blue customers give the lowest very satisfaction (satisfaction =5) which is 4% compared with other airline status customers, which are 19% - 38%. The Blue status also has the highest ratio (27%) of dissatisfaction (satisfaction < 3). The airline company should pay more attention to the Blue status since they have the highest frequency of all the passengers and the highest percentage of dissatisfaction. The Platinum customers are the most polar passengers who give both the highest percentage of satisfaction =5 and satisfaction = 1. However, the platinum status is only 3.2% of all the data, so in this project we focus on the Blue status which covered most type of customers.

Figure 18 Relation Between Flight Per Year and Satisfaction



## Flight per year Analysis

As shown in Figure 18, with increasing flight times per year, the customers are more sensitive to the service which has the trend to give dissatisfaction rating (<4). The frequent flight customers have more experience of travel and more opportunity to compare with the great service previously.

## Associate Rules Analysis

Based on the previous study, six features (Age, Arrival Delay, Travel Type, Airline Status, Price Sensitivity and Flight per Year) are the most important for the satisfaction. In order to find the linkage between the satisfaction and the six features, we applied the associate rules analysis to the data.

The first thing that we need is to convert the features with numeric data type to char. We convert satisfaction, price sensitivity and flight per year to three levels (high, average and low) by the criteria shown in Table 4.

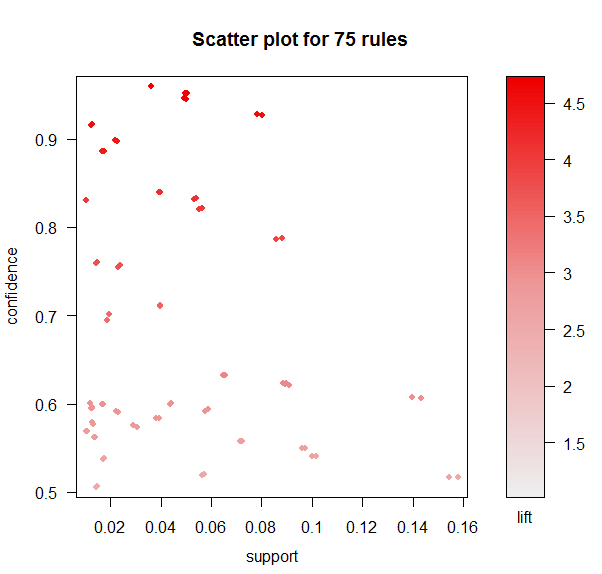
|  |  |  |
| --- | --- | --- |
| **Table 4** | | |
| Satisfaction and Price Sensitivity | Flight per year | Levels |
| 1,2 | <40% of median value | Low |
| 3 | 40% - 60% of median value | Average |
| 4,5 | >60% of median value | High |

The Age are separated based on the decision tree that young is less than 23 and senior is over 60. The arrival delay time is also separated based on the decision tree by 5 min.

Firstly, we tested the criteria that support = 0.1 and confidence = 0.5 of the associate rules analysis. There are only 5 rules which have been found which is listed below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 5** | | | | | | | |
|  | lhs |  | rhs | support | confidence | lift | count |
| [1] | {Typeoftravel=Personal Travel} | => | {Satisfaction=Low} | 0.1579578 | 0.5171735 | 2.541829 | 30205 |
| [2] | {Age=High,  Typeoftravel=Personal Travel} | => | {Satisfaction=Low} | 0.1017875 | 0.5400966 | 2.654493 | 19464 |
| [3] | {Airlinestatus=Blue,  Typeoftravel=Personal Travel} | => | {Satisfaction=Low} | 0.1433622 | 0.6064373 | 2.980547 | 27414 |
| [4] | {Typeoftravel=Personal  Travel,Price=Low} | => | {Satisfaction=Low} | 0.1543232 | 0.5167221 | 2.539611 | 29510 |
| [5] | {Airlinestatus=Blue,  Typeoftravel=Personal Travel,Price=Low} | => | {Satisfaction=Low} | 0.1399107 | 0.6072726 | 2.984652 | 26754 |

However, the arrival delay is not in the left-hand side ,which contradicts the decision tree analysis result due to the setting of the support value. So, we reduced the support value to 0.01. The results of 75 rules are shown in Figure xx. The best rules have confidence over 0.9 and lift value about 4.5, even the support value is about 4% -9%. The arrival delay time over 5 mins appears in all the best rules which indicates that arrival delay is an important feature of the dissatisfaction.



## Arrival Delay Time Analysis

In the previous paragraph, we found the importance of arrival delay time. In this part, we further analyze the relationship between satisfaction and average arrival delay time of each satisfaction level. The overall trend shows that average delay time decreases with the increasing of the satisfaction rate. We also performed a linear regression model to obtain the decay rate between satisfaction and delay time.

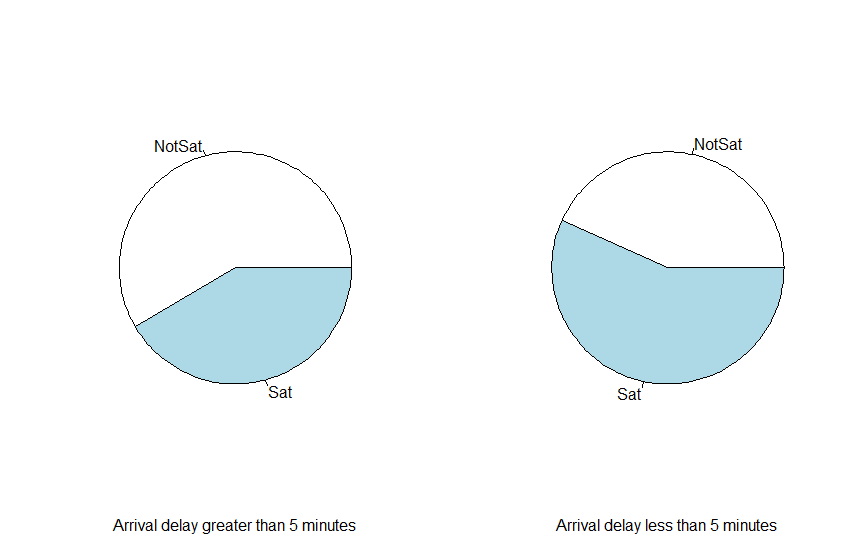


Figure Satisfaction Distribution of Different Delay Status

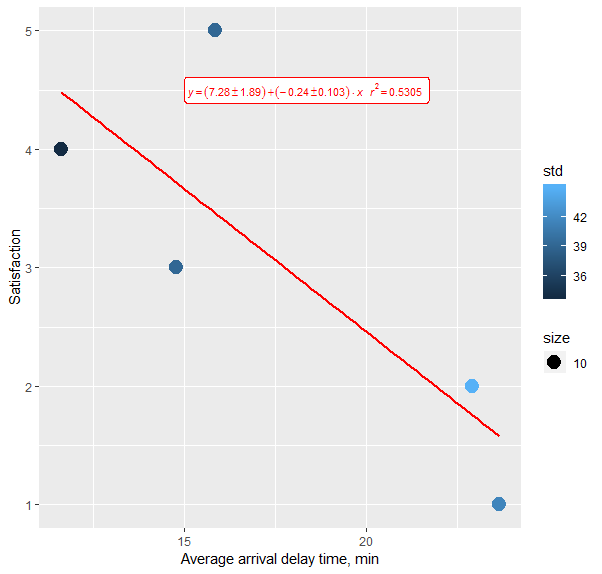


Figure 20 Linear Relation Between Satisfaction and Average Arrival Delay Time

# Conclusion and Recommendation

1. The company should improve its service for personal travelers. A large proportion of personal travelers are not satisfied with their experiences.
2. Whether the arrival time is late for more than 5 minutes is important for customers. Specifically, as an example, while most of the personal travelers are not satisfied, Silver customers who have arrived less than 5 minutes are more likely to be satisfied. Old (more than 65 years old) business travelers or young (less than 23 years old) non-personal travelers are likely to complain about the trip if the arrival is late for more than 5 minutes. If arrival delay cannot be avoided, the airline company should take some other measures to improve customer satisfaction.
3. While many of the middle-aged customers are satisfied with the airline’s service, old customers are more likely to report complaints about the service. The company may implement some marketing strategies for old customers to improve their satisfaction.

# Appendix

R code:

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

# IST687, Final Project

#

# Student name: Ishita Joshi, Leah Singer, Wei Mu, Xin Sun, Zhenlei Liu# Homework number: 7

# Date due: April 25

# Date submitted: April 25

# Attribution statement: (choose the one statement that is true)

# We did this homework with help from the book and the

# professor and these Internet sources: <provide the urls>

# https://ggplot2.tidyverse.org/reference/

# rattle()

# and other R documents

# Run these three functions to get a clean test of homework code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list = ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your homework data files

setwd("~/IntroDataScience/")

# Your homework specific code goes below here

if(!requireNamespace("ggplot2")) install.packages("ggplot2")

if(!requireNamespace("dplyr")) install.packages("dplyr")

if(!requireNamespace("rpart")) install.packages("rpart")

if(!requireNamespace("rattle")) install.packages("rattle")

if(!requireNamespace("ROSE")) install.packages("ROSE")

if(!requireNamespace("arules")) install.packages("arules")

if(!requireNamespace("arulesViz")) install.packages("arulesViz")

library(dplyr)

library(ggplot2)

library(rpart)

library(rattle)

library(ROSE)

library(arules)

library(arulesViz)

df <- read.table("spring19survey.csv", sep = ',', header = TRUE)

str(df)

# There are 190720 complete cases, 4133 cases contain "NA" answers.

sum(complete.cases(df))

# Average satisfaction between cancelled or not

df %>% group\_by(Flight.cancelled) %>% summarise(mean(Satisfaction, na.rm = T))

# A table which pivots into incomplete cases and cancelled cases

df\_whole\_addcomplete <- df

df\_whole\_addcomplete$ifNA <- as.factor(!complete.cases(df\_whole\_addcomplete))

df\_whole\_addcomplete %>% group\_by(ifNA) %>% count(Flight.cancelled==levels(Flight.cancelled)[2]) #Flight Cancelled

# drop incomplete cases

df <- df[complete.cases(df),]

# Examine satisfaction score

hist(df$Satisfaction)

table(df$Satisfaction)

# >=4 as Satisfied, <4 as not satisfied

df\_addif <- df

df\_addif$ifSatisfied <- as.factor(ifelse(df$Satisfaction>=4, "Sat", "NotSat"))

# balanced between two categories

ggplot(df\_addif, aes(ifSatisfied)) + geom\_bar()

# =5 as very satisfied, <5 as not very satisfied

df\_addif$ifSatisfied <- as.factor(ifelse(df$Satisfaction==5, "VerySat", "NotVerySat"))

# extremely unbalanced

ggplot(df\_addif, aes(ifSatisfied)) + geom\_bar()

# For now we define >=4 as Satisfied, <4 as not satisfied

df\_addif$ifSatisfied <- as.factor(ifelse(df$Satisfaction>=4, "Sat", "NotSat"))

# Examine Partner companies, 14 in total

summary(df\_addif$Partner.Code)

str(df\_addif)

# Very different number of responses

ggplot(df\_addif, aes(x=Partner.Code, fill=ifSatisfied)) + geom\_bar()

df\_addif$TFSatisfied <- as.factor(ifelse(df\_addif$Satisfaction>=4, TRUE, FALSE))

str(df\_addif)

# Satisfied percent for companies

# result: pretty much the same between different partner companies, we can combine them together

group\_by(df\_addif, Partner.Code) %>%

summarise(Satisfied.Percent = sum(TFSatisfied==TRUE)/n()) %>%

ggplot(aes(x=reorder(Partner.Code, Satisfied.Percent), y=Satisfied.Percent)) +

geom\_bar(stat = 'identity') +

xlab("Partner.Code")

# convert partner code to name

Partner.Coversion.Table <- group\_by(df\_addif, Partner.Code) %>%

summarise(Partner.Name[1])

Partner.Coversion.Table

# average raw satisfaction score of partner companies, almost the same

group\_by(df\_addif, Partner.Code) %>%

summarise(Satisfaction.Score = sum(Satisfaction)/n()) %>%

ggplot(aes(x=reorder(Partner.Code, Satisfaction.Score), y=Satisfaction.Score)) +

geom\_bar(stat = 'identity') +

xlab("Partner.Code")

# Decision Tree

# Split into training and test data

drop.cols <- c('Day.of.Month', 'Year.of.First.Flight', 'Flight.date', 'Partner.Code', 'Partner.Name', 'Orgin.City', 'Origin.State', 'Destination.City', 'Destination.State')

df\_cleared <- df %>% select(-drop.cols)

df\_cleared$ifSatisfied <- as.factor(ifelse(df\_cleared$Satisfaction>=4, "Sat", "NotSat"))

df\_cleared <- df\_cleared %>% select(-Satisfaction)

str(df\_cleared)

## Perform sampling on the dataset

set.seed(42)

sample.size <- round(0.70\*nrow(df\_cleared))

indexes <- sample(1:nrow(df\_cleared),size = sample.size,replace = F)

training.data <- df\_cleared[indexes,]

testing.data <- df\_cleared[-indexes,]

str(testing.data)

mytree <- rpart(ifSatisfied ~.,

data = df\_cleared,

parms = list(split= 'gini'),

method = "class", control=rpart.control(minsplit=2, minbucket = 1, cp = 0.01))

# Complexity parameter

plotcp(mytree)

# Plot the tree

fancyRpartPlot(mytree)

# Try expanding the tree

mytree2 <- rpart(ifSatisfied ~.,

data = df\_cleared,

parms = list(split= 'gini'),

method = "class", control=rpart.control(minsplit=2, minbucket = 1, cp = 0.005))

# Complexity parameter, not much difference when cp<=0.01

plotcp(mytree2)

# Plot the tree

fancyRpartPlot(mytree2)

# Test on testing dataset for cp=0.01 tree

predictions <- predict(mytree,testing.data, type = "class")

# ROC: 0.78

roc.curve(testing.data$ifSatisfied, predictions)

# Verify the conclusion that personal travelers are not satisfied using descriptive statistics

df\_temp <- df\_cleared %>%

group\_by(Type.of.Travel, ifSatisfied) %>%

count() %>%

mutate(flag = as.integer(ifSatisfied)\*2-3) %>%

mutate(Response = n\*flag)

ggplot(df\_temp, aes(x=Type.of.Travel, y=Response, fill=ifSatisfied)) +

geom\_bar(stat = 'identity') +

ylab("Number of Responses")

# PCA

# Note some of the following code is merged from rattle log

str(df)

drop.cols <- c('Day.of.Month', 'Year.of.First.Flight', 'Flight.date', 'Partner.Code', 'Partner.Name', 'Orgin.City', 'Origin.State', 'Destination.City', 'Destination.State')

df\_whole <- df %>% select(-drop.cols)

str(df\_whole)

numeric\_attr <- c("Age", "Price.Sensitivity", "Flights.Per.Year",

"Loyalty", "Total.Freq.Flyer.Accts",

"Shopping.Amount.at.Airport",

"Eating.and.Drinking.at.Airport",

"Scheduled.Departure.Hour",

"Departure.Delay.in.Minutes",

"Arrival.Delay.in.Minutes",

"Flight.time.in.minutes", "Flight.Distance")

set.seed(42)

nobs <- nrow(df\_whole)

train <- sample(nobs, 0.7\*nobs)

pc <- prcomp(na.omit(df\_whole[train, numeric\_attr]), scale=TRUE, center=TRUE, tol=0)

# Show the output of the analysis.

pc

# Summarise the importance of the components found.

summary(pc)

# Display a plot showing the relative importance of the components.

plot(pc, main="")

title(main="Principal Components Importance",

sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d %H:%M:%S"), Sys.info()["user"]))

axis(1, at=seq(0.7, ncol(pc$rotation)\*1.2, 1.2), labels=colnames(pc$rotation), lty=0)

# gender and age analysis

g <- ggplot(data = df, aes(Gender))

# Number of responses from different genders in the entire survey. Note that responses from female customers are more than male responses.

g + geom\_bar()

# If we define Satisfaction >= 4 as "satisfied", otherwise "not satisfied"

df$ifSatisfied <- as.factor(ifelse(df$Satisfaction>=4, "Yes", "No"))

# Stacked bar plot of customer satisfaction and gender

# Note the difference between genders. Male customer seems to be more likely to give higher satisfaction score.

g <- ggplot(data = df, aes(Gender))

g + geom\_bar(aes(fill=ifSatisfied))

# Show the stacked histogram of number of customer responses versus ages and gender

ggplot(df, aes(Age, fill = Gender)) + geom\_histogram(binwidth = 5)

# If we define Age < 30 as "Young", 30 <= Age < 60 as "Middle" age, Age >= 60 as "old" age

df$ageRange <- ifelse(df$Age < 30, "Young",

ifelse(df$Age < 60, "Middle", "Old"))

df$ageRange <- as.factor(df$ageRange)

# 34527 Young age responses, 111744 Middle age responses, 44449 Old age responses

summary(df$ageRange)

# Stacked bar plot of number of customer responses versus age and gender

ggplot(data = df, aes(ageRange)) + geom\_bar(aes(fill = Gender))

# Stacked bar plot of customer satisfaction and age

# Note the difference between different age range. Old customers(Age >=60) are likely to give lower satisfaction score,

# while Middle age customers are more likely to give higher satisfaction score.

ggplot(data = df, aes(ageRange)) + geom\_bar(aes(fill = ifSatisfied))

# Linear regression

checkage<- lm(Satisfaction ~ Age + I(Gender), data = df)

# Note to Ishita: if you write as Satisfaction ~ Age + Gender rather than I(rather), it gives the same result.

# It is a dummy variable.

summary(checkage)

# The coefficient of Age is negative and the coefficient of I(Male) is positive.

# The overall regression is significant with p value < 2.2e-16. Both Age and I(Male) are significant with p<2e-16.

# So the interpretation is:

# With other factors constant, if the customer has a higher age, the customer gives lower satisfaction score.

# If everything else is the same, male customers gives higher satisfaction score.

# In general, the conclusion is the same as conclusions from the visualized graphs above.

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

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

# Part 2 of the code from a different group member

# remove everything first but keep the library

rm(list=ls())

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

df <- read.csv("spring19survey.csv") # read from csv file

str(df)

summary(df)

# Preparing of the data for futher analysis

df <- df[-which(is.na(df$Satisfaction)),] # remove NA value in stisfaction

# Flight per year

df.flightlot <- df[df$Flights.Per.Year>17,]

df.flightlot2 <- df[df$Flights.Per.Year>29,]

# Price sensitivity

summary(df$Price.Sensitivity)

df.price <- df[df$Price.Sensitivity > 2,]

dim(df.price)

summary(df.price$Satisfaction)

# Departure Delay

summary(df$Departure.Delay.in.Minutes)

df.D.delay <- df[df$Departure.Delay.in.Minutes>60,]

dim(df.D.delay)

summary(df.D.delay$Departure.Delay.in.Minutes)

# Arrival Delay

summary(df$Arrival.Delay.in.Minutes)

df.A.delay <- df[df$Arrival.Delay.in.Minutes>480,]

dim(df.A.delay)

summary(df.A.delay$Arrival.Delay.in.Minutes)

# Airline Status

summary(df$Airline.Status)

Airline.Blue <- filter(df, Airline.Status == "Blue")

Airline.Silver <- filter(df, Airline.Status == "Silver")

Airline.Gold <- filter(df, Airline.Status == "Gold")

Airline.Platinum <- filter(df, Airline.Status == "Platinum")

# Data visualization

# function for ploting

plot\_pie <- function(df,title){

cust5 <- filter(df, Satisfaction == 5)

cust4 <- filter(df, Satisfaction == 4)

cust3 <- filter(df, Satisfaction == 3)

cust2 <- filter(df, Satisfaction == 2)

cust1 <- filter(df, Satisfaction == 1)

t1 <- c(dim(cust5)[1],dim(cust4)[1],dim(cust3)[1],dim(cust2)[1],dim(cust1)[1])

t1 <- t1/sum(t1)

print(t1)

p1 <- pie(t1,label = paste(round(t1\*100),"%"), main = title, col = rainbow(length(t1)))

legend("topright",c("Satisfaction = 5", "Satisfaction = 4","Satisfaction = 3",

"Satisfaction = 2","Satisfaction = 1"),cex = 0.8, fill = rainbow(length(t1)) )

return(p1)

}

plot\_hist <- function(df){

g1 <- ggplot(df)+ aes(x=Satisfaction)+geom\_histogram(binwidth=0.5)

# +ggtitle("Hist of Satisfaction")

return(g1)

}

p <- plot\_pie(Airline.Blue,"Satisfaction Distribution of Blue")

p <- plot\_pie(Airline.Silver,"Satisfaction Distribution of Sliver")

p <- plot\_pie(Airline.Gold,"Satisfaction Distribution of Gold")

p <- plot\_pie(Airline.Platinum,"Satisfaction Distribution of Plantinum")

# Associate rules

# Age, arrive delay, airline status, type of travel, price sensitivity, flight per year

summary(df)

rate1to5 <- function(vec){

level <- replicate(length(vec), "Average")

level[vec > 3] <- "High"

level[vec < 3] <- "Low"

return(level)

}

ratenum <- function(vec){

q <- quantile(vec, c(0.4, 0.6))

level <- replicate(length(vec), "Average") # num = (0.4,0.6) -> average

level[vec <= q[1]] <- "Low" # num = [0,0.4] -> low

level[vec > q[2]] <- "High" # num = [0.6,1] -> high

return(level)

}

ratearrive <- function(vec){

level <- replicate(length(vec), "Delay over 5 min")

level[vec < 5] <- "Delay less than 5 min"

return(level)

}

create\_df\_arule <- function(df){

Satisfaction <- df$Satisfaction

Age <- df$Age

Airlinestatus <- df$Airline.Status

Arrivedelay <- df$Arrival.Delay.in.Minutes

Typeoftravel <- df$Type.of.Travel

Price <- df$Price.Sensitivity

Flightsperyear <- df$Flights.Per.Year

df.arules <- data.frame(Satisfaction,Age,Airlinestatus,Arrivedelay,Typeoftravel,Price,Flightsperyear)

df.arules$Satisfaction <- rate1to5(df.arules$Satisfaction)

table(df.arules$Satisfaction)

df.arules$Price <- rate1to5(df.arules$Price)

table(df.arules$Price)

df.arules$Flightsperyear <- ratenum(df.arules$Flightsperyear)

table(df.arules$Flightsperyear)

df.arules$Age <- ratenum(df.arules$Age)

table(df.arules$Age)

df.arules$Arrivedelay <- ratearrive(df.arules$Arrivedelay)

return(df.arules)

}

df.arules <- create\_df\_arule(df)

summary(df.arules)

str(df.arules)

rulesets <- as(df.arules,"transactions")

itemFrequency(rulesets)

itemFrequencyPlot(rulesets)

r1 <- apriori(rulesets,

parameter = list(support=.1, confidence = .5),

appearance = list(default = "lhs",rhs =("Satisfaction=Low")))

summary(r1)

inspect(r1)

plot(r1)

#

# lhs rhs support confidence lift count

# [1] {Typeoftravel=Personal Travel} => {Satisfaction=Low} 0.1582189 0.5129635 2.515812 30825

# [2] {Age=High,

# Typeoftravel=Personal Travel} => {Satisfaction=Low} 0.1019479 0.5355371 2.626523 19862

# [3] {Airlinestatus=Blue,

# Typeoftravel=Personal Travel} => {Satisfaction=Low} 0.1435391 0.6013203 2.949155 27965

# [4] {Typeoftravel=Personal Travel,

# Price=Low} => {Satisfaction=Low} 0.1545336 0.5124421 2.513255 30107

# [5] {Age=High,

# Typeoftravel=Personal Travel,

# Price=Low} => {Satisfaction=Low} 0.1000693 0.5354573 2.626132 19496

# [6] {Airlinestatus=Blue,

# Typeoftravel=Personal Travel,

# Price=Low} => {Satisfaction=Low} 0.1400334 0.6020523 2.952745 27282

#

# Plot overall distribution

p1 <- plot\_pie(df,"Overall Satisfaction Distribution")

p2 <- plot\_hist(df)

p2

# Plot > mean distribution

p3 <- plot\_pie(df.flightlot,"Satifaction Distribution for Flight >17/year")

p4 <- plot\_hist(df.flightlot)

p4

# Plot >75% distribution

p5 <- plot\_pie(df.flightlot2,"Satifaction Distribution for Flight >29/year")

p6 <- plot\_hist(df.flightlot2)

p6

lm.model <- lm(df.A.delay$Satisfaction~df.A.delay$Arrival.Delay.in.Minutes, data = df.A.delay)

summary(lm.model)

plot(df.A.delay$Arrival.Delay.in.Minutes,df.A.delay$Satisfaction)

df.delayover5 <- filter(df, Arrival.Delay.greater.5.Mins == "yes")

df.arules <- create\_df\_arule(df.delayover5)

summary(df.arules)

str(df.arules)

rulesets <- as(df.arules,"transactions")

itemFrequency(rulesets)

itemFrequencyPlot(rulesets)

r1 <- apriori(rulesets,

parameter = list(support=.1, confidence = .5),

appearance = list(default = "lhs",rhs =("Satisfaction=Low")))

summary(r1)

inspect(r1)

plot(r1)