IST 687 INTRODUCTION TO DATA SCIENCE
SECTION M009

FINAL PROJECT REPORT

DECEMBER 6, 2018

GROUP 3

Xuehan Chen, Yanqi Yao, Advait Ramesh Iyer, Weicheng Wu

DATA ANALYSIS FOR PASSENGERS SATISFACTION INDEX

Objective:

To identify various trends in the customer data of flights in the US to formulate various business strategies for our client: **Southeast Airlines Co.**

Contents

1. EXECUTIVE SUMMARY

In this report, the team analyzed the performance of various airlines, including the client's, and addressed various business questions. The recommendations in the report span in the areas of marketing, finance, and operations. The analysis done in the report include four models:

- **Multiple Linear Regression Model**
- **Association Rules Mining Model**
- **Support Vector Machine Model**
- **Decision Tree Model**

The analysis in the report is done with the objective of identifying key attributes that affect the customer satisfaction index. Four different models were used to validate the claims made by each one, and to gain the ability to view data from different perspectives to gain more insight into the Key Performance Indicators (KPIs) of Satisfaction Index.

After all the models were developed, the key attributes that we could identify, which were consistently visible across all the models, are:

- **Age** has a negative impact as it increases. It has been one of the key decision nodes within the decision tree, which separated two different clusters across two clusters. Moreover, middle-aged individuals are seen as highly satisfied when travelling in premium classes
- **Arrival delay in minutes**, which is very strongly correlated with satisfaction index is the most important attribute
- **Airline status**, where Blue and Platinum were distinguishable generating low and moderate dissatisfaction, whereas Gold and Silver statuses generated moderate and high satisfaction
- **Gender** seems to play a big role in deciding how a customer would tend to react to the type of service. Women are clearly distinguishable sharing moderate or low satisfaction levels as compared to men
- **Type of travel,** i.e., Business travel, Mileage travel, and Personal travel drastically changes the customer's expectations, with personal travel frequently associated with low satisfaction levels while the latter two can be seen associated with high satisfaction levels

The team's recommendations are directed towards all age groups availing both economy commute and premium commute. The key recommendations include:

- Increasing operating and maintenance expenses within blue status airlines, and classes which can be associated with personal travel
- Increasing marketing expenses to increase market exposure of the services offered with premium class tickets
- Creating service bundles for customer segments such as business travelers and female travelers, which would lead to value addition for the customer

2. INTRODUCTION

In this report, we analyzed the satisfaction level of the passengers travelling through flights in the US through various attributes defining the quality of flight service, commute status: whether delayed or not as well as the economic status of the individuals. The team focused on identifying attributes that strongly affect the customer satisfaction ratings, and then validated their claim using various analysis methods.

Methodology

The model was initially analyzed using multiple linear regression analysis, which was done to check which attributes were strongly affecting the customer satisfaction index. Once they were identified, the team conducted association rules mining using apriori algorithm to validate the hypothesis suggested by the linear model.

After conducting association rules mining, the team conducted a machine learning exercise using support vector machine (SVM) method to classify passengers into clusters, and see how accurately the test dataset, i.e., one-third of the dataset performs when it learns the classification method from the training dataset, which is two-thirds of the data. The algorithm classifies the passengers into two categories: **Satisfied and Not Satisfied.** This method was used to conduct supervised machine learning, and the classification matrix gave us the error rate for the algorithm.

3. BUSINESS QUESTIONS

A. What are the key reasons for customer satisfaction or dissatisfaction?
B. Which clusters generally come off as happy passengers? Which of them come off as unhappy passengers?
C. What should be the pricing strategy of airline packages which might lead to a higher satisfaction level among the unhappy passengers?
D. What can be any strategic improvements to increase average customer satisfaction index?

4. DATA PREPROCESSING

The variables in the data were as follows:

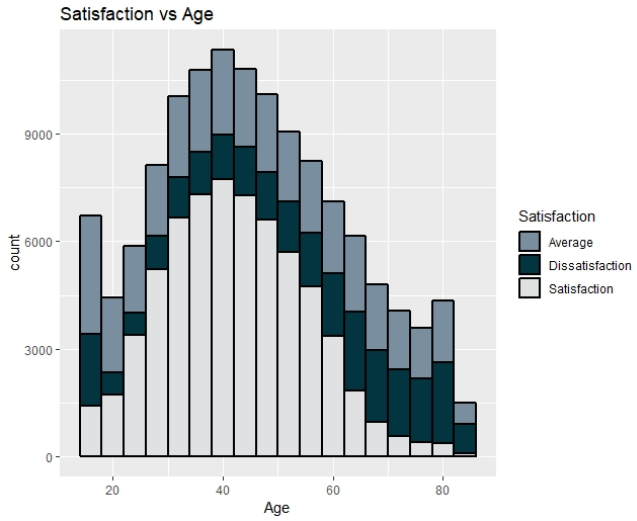
1. **Satisfaction (Range: 0-5)**
2. **Airline Status (Range: Blue, Gold, Silver, Platinum)**
3. **Age (Range: 15-85 years)**
4. **Gender**
5. **Price Sensitivity (Range: 0-5)**
6. **Year of First Flight (Range: 2003-2012)**

7. **Number of flights per annum (Range: 0-100)**
8. **Percentage of flights with other airlines (Range: 1-100)**
9. **Type of travel (Range: Business travel, Mileage travel, Personal travel)**
10. **Number of other loyalty cards (Range: 0-12)**
11. **Shopping amount at airport (Range: $0-$879)**
12. **Eating and drinking at airport (Range: $0-$895)**
13. **Class (Range: Economy, Economy Plus, Business)**
14. **Day of month (Range: 1-31)**
15. **Flight date (Range: 1/1/14 to 3/9/14)**
16. **Airline name: 14 different airlines**
17. **Origin city**
18. **Origin state**
19. **Destination city**
20. **Destination state**
21. **Scheduled departure hour**
22. **Departure delay in minutes (Range: 0-1592 min)**
23. **Arrival delay in minutes (Range: 0-1584 min)**
24. **Flight cancelled (Yes/No)**
25. **Flight time in minutes (Range: 8-669 min)**
26. **Flight distance (Range: 31-4983 miles)**
27. **Arrival delay greater than 5 minutes (Yes/No)**

3.1 Descriptive Statistics

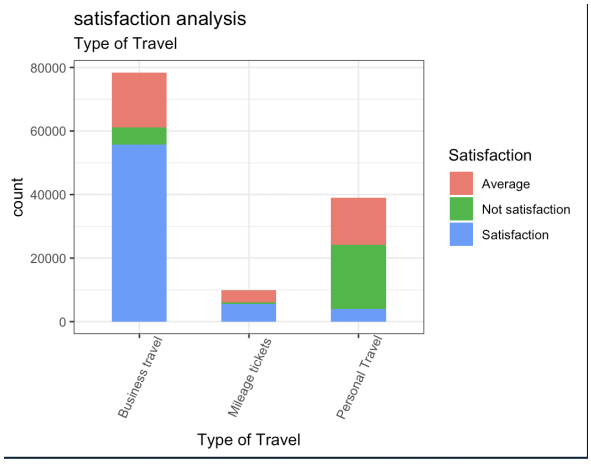
After importing the dataset into the R environment, the team conducted preliminary analysis of the contents, distribution and characteristics of the dataset. We focused upon visualizing if any factor has distinguishable characteristics when plotted against satisfaction index. We categorized the satisfaction index into three levels: **Satisfaction** if the index is **greater than or equal to 4**, **Average** if the index is **between 3 and 4**, and **Dissatisfaction** if the index is **less than 3**. Following are the results that the descriptive statistics exercise resulted in:

i. **Satisfaction Vs Age:**



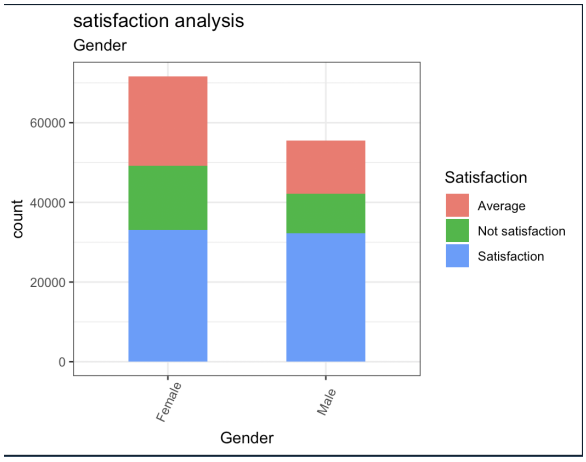
Judging by the distribution of the histogram, the dissatisfaction and average ratings increase with increase in age, and the age group 60-80 are the most dissatisfied.

ii. Satisfaction Vs Type of Travel:



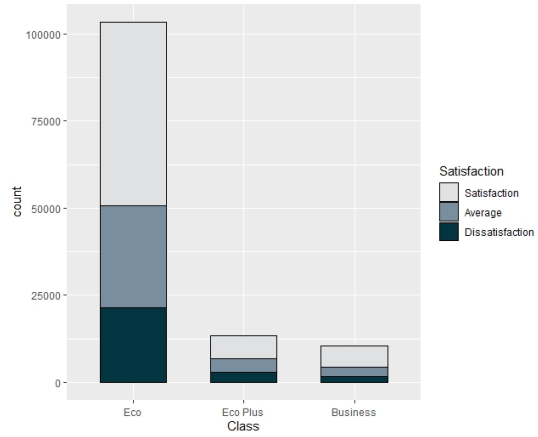
It is clearly visible that the business travel is a more satisfying experience than personal travel. This might be because of the allowances provided by the employer, as well as possibility of booking a business class ticket. However, personal travel is based on the individuals' own money, which might trigger dissatisfaction. Mileage tickets, although not availed that frequently, is seen to produce more satisfaction than personal travel.

iii. Satisfaction Vs Gender:



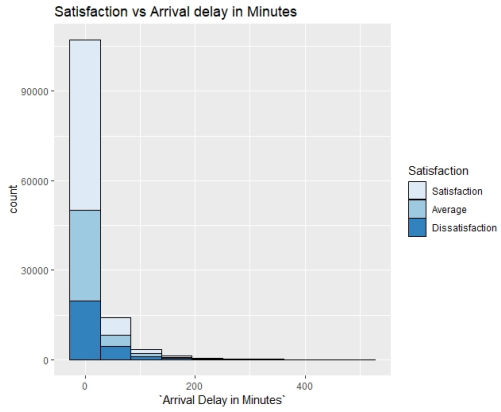
There is a higher satisfaction level among the male passengers as compared to female passengers. There are more female passengers who are either not satisfied or moderately satisfied.

iv. Satisfaction Vs Class:



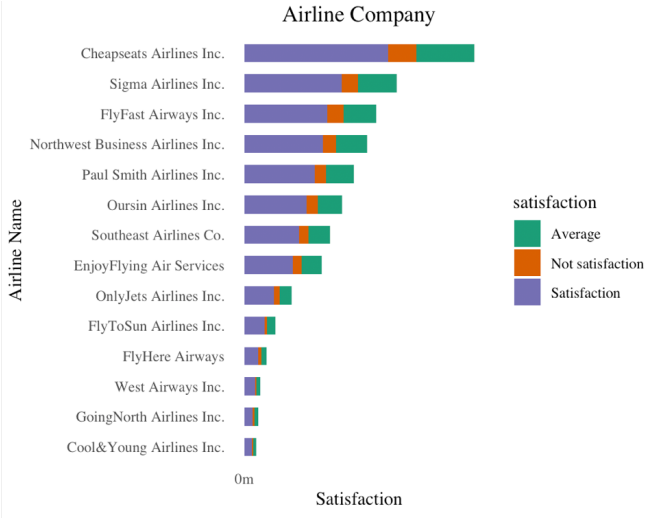
In the above bar plot, economy class has the highest proportion of moderately satisfied passengers, followed by economy plus.

v. **Satisfaction Vs Arrival delay in minutes:**



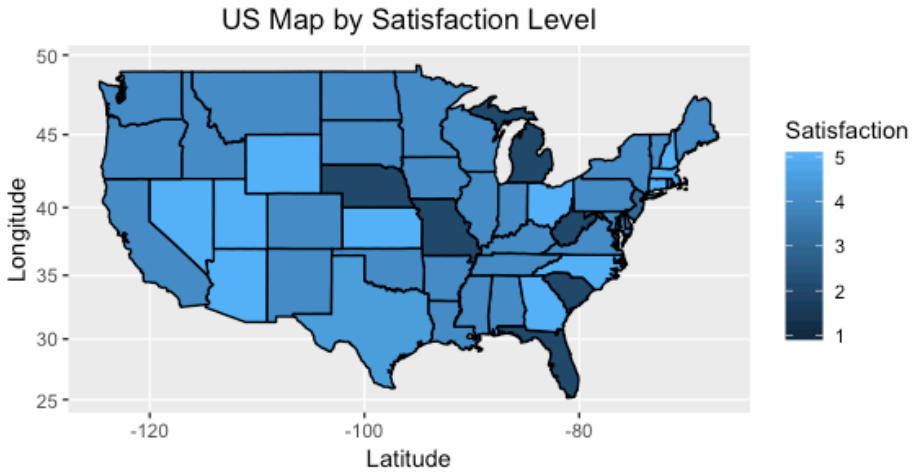
From the bar plot above, it can be concluded that a higher arrival delay would lead to a lesser customer satisfaction.

vi. **Satisfaction Vs Airline Company:**

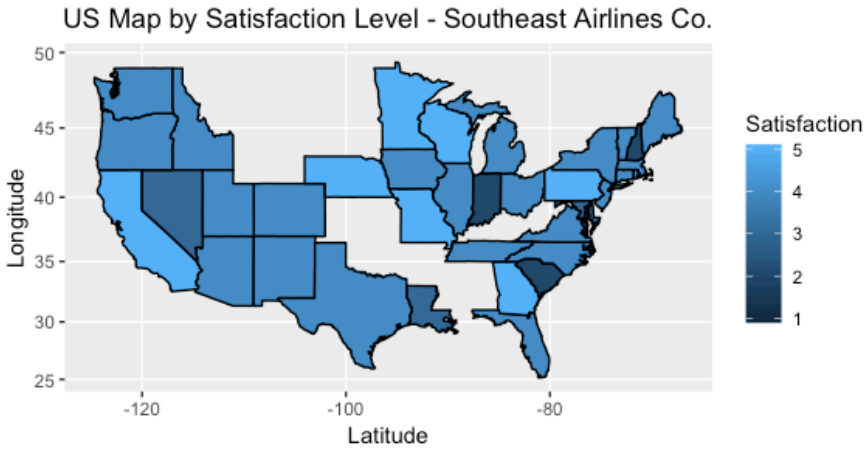


From the above data, we can conclude that **West Airways Inc.** and **Cool&Young Airlines Inc.** have lesser dissatisfaction than **Southeast Airlines Co.** Overall, Southeast Airlines performs better than most of its competitors.

vii. **Comparing satisfaction levels by geography:**



The map shown above denotes **satisfaction level of destination states**, wherein they are color-coded according to the level of satisfaction.



The map shown here denotes the satisfaction level of destination states, considering **only Southeast Airlines Co.** When compared with the above map, we can see that there is scope for improvement of service in the following states: **Nevada, Utah, Arizona, New Hampshire, Virginia and Ohio.**

3.2 Data Cleaning

After viewing the descriptive statistics of various attributes within the data, the team checked for consistency in the data by checking the type of data in each column (whether it is a factor, number, or character) as well as the number of NA (null) cells in each column.

Following are the observations:

```
> str(Satisfaction_Survey)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       129889 obs. of  28 variables:
 $ Satisfaction                   : num  4.5 4 2.5 4 5 5 3.5 4 4 4 ...
 $ Airline Status                 : chr  "Blue" "Blue" "Blue" "Blue" ...
 $ Age                            : int  31 56 21 43 49 49 35 33 44 51 ...
 $ Gender                         : chr  "Male" "Male" "Female" "Male" ...
 $ Price Sensitivity              : int  1 2 2 1 1 1 1 1 1 1 ...
 $ Year of First Flight           : int  2007 2006 2006 2007 2006 2010 2011 2010 2003 2005 ...
 $ No of Flights p.a.             : int  28 41 8 9 14 0 15 4 8 12 ...
 $ % of Flight with other Airlines: int  7 3 7 9 10 4 5 17 6 7 ...
 $ Type of Travel                 : chr  "Business travel" "Business travel" "Personal Travel" "Business travel" ...
 $ No. of other Loyalty Cards     : int  2 0 0 2 0 1 0 2 0 0 ...
 $ Shopping Amount at Airport     : int  0 15 0 10 8 0 0 0 0 25 ...
 $ Eating and Drinking at Airport : int  75 60 135 45 26 65 60 90 90 80 ...
 $ Class                          : chr  "Business" "Business" "Business" "Eco" ...
 $ Day of Month                   : int  18 11 25 20 25 16 6 5 21 19 ...
 $ Flight date                    : chr  "3/18/14" "1/11/14" "1/25/14" "2/20/14" ...
 $ Airline Code                   : chr  "MQ" "MQ" "MQ" "MQ" ...
 $ Airline Name                   : chr  "EnjoyFlying Air Services" "EnjoyFlying Air Services" "EnjoyFlying Air Services" "EnjoyFlying Air Services" ...
 $ Orgin City                     : chr  "Madison, WI" "Madison, WI" "Milwaukee, WI" "Madison, WI" ...
 $ Origin State                   : chr  "Wisconsin" "Wisconsin" "Wisconsin" "Wisconsin" ...
 $ Destination City               : chr  "Dallas/Fort Worth, TX" "Dallas/Fort Worth, TX" "Dallas/Fort Worth, TX" "Dallas/Fort Worth, TX" ...
 $ Destination State              : chr  "Texas" "Texas" "Texas" "Texas" ...
 $ Scheduled Departure Hour       : int  15 11 12 11 12 18 6 18 12 18 ...
 $ Departure Delay in Minutes     : int  0 2 34 26 0 0 0 0 0 0 ...
 $ Arrival Delay in Minutes       : int  3 5 14 39 0 0 0 1 0 0 ...
 $ Flight cancelled               : chr  "No" "No" "No" "No" ...
 $ Flight time in minutes         : int  134 120 122 141 144 123 119 138 114 118 ...
 $ Flight Distance                : int  821 821 853 821 853 821 821 821 853 821 ...
 $ Arrival Delay greater 5 Mins   : chr  "no" "no" "yes" "yes" ...
```

In the result of the str() command, we observe that there are **129889 observations of 28 variables**, with one column “Satisfaction” in number format, 1**3 columns** namely “Airline Status”, “Gender”, “Type of Travel”, “Class”, “Flight date”, “Airline Code”, “Airline Name”, “Origin State”, “Origin City”, “Destination State”, “Destination City”, “Flight cancelled” and “Arrival Delay greater 5 Mins” are **in character format**, and **14 columns** namely “Age”, “Price Sensitivity”, “Year of First Flight”, “No. of Flights p.a.”, “% of Flights with other Airlines”, “No. of other Loyalty cards”, “Shopping amount at Airport”, “Eating and Drinking at Airport”, “Day of Month”, “Scheduled Departure Hour”, “Departure Delay in Minutes”, “Arrival Delay in Minutes”, “Flight time in minutes”, and “Flight Distance” **in integer format**.

```
> summary(Satisfaction_Survey)
  Satisfaction   Airline Status          Age          Gender          Price Sensitivity Year of First Flight No of Flights p.a. % of Flight with other Airlines
Min.   :1.000   Length:129889      Min.   :15.0   Length:129889      Min.   :0.000     Min.   :2003        Min.   :  0.00     Min.   :  1.000
1st Qu.:3.000   Class :character   1st Qu.:33.0   Class :character   1st Qu.:1.000     1st Qu.:2004        1st Qu.:  9.00     1st Qu.:  4.000
Median :4.000   Mode  :character   Median :45.0   Mode  :character   Median :1.000     Median :2007        Median : 17.00     Median :  7.000
Mean   :3.379                      Mean   :46.2                      Mean   :1.276     Mean   :2007        Mean   : 20.08     Mean   :  9.314
3rd Qu.:4.000                      3rd Qu.:59.0                      3rd Qu.:2.000     3rd Qu.:2010        3rd Qu.: 29.00     3rd Qu.: 10.000
Max.   :5.000                      Max.   :85.0                      Max.   :5.000     Max.   :2012        Max.   :100.00     Max.   :110.000
NA's   :3
Type of Travel     No. of other Loyalty Cards Shopping Amount at Airport Eating and Drinking at Airport    Class           Day of Month   Flight date
Length:129889      Min.   : 0.0000           Min.   :  0.00              Min.   :  0.00                Length:129889      Min.   : 1.00   Length:129889
Class :character   1st Qu.: 0.0000           1st Qu.:  0.00              1st Qu.: 30.00                Class :character   1st Qu.: 8.00   Class :character
Mode  :character   Median : 0.0000           Median :  0.00              Median : 60.00                Mode  :character   Median :16.00   Mode  :character
                   Mean   : 0.8838           Mean   : 26.55              Mean   : 68.24                                   Mean   :15.72
                   3rd Qu.: 2.0000           3rd Qu.: 30.00              3rd Qu.: 90.00                                   3rd Qu.:23.00
                   Max.   :12.0000           Max.   :879.00              Max.   :895.00                                   Max.   :31.00

Airline Code       Airline Name       Orgin City         Origin State       Destination City   Destination State  Scheduled Departure Hour
Length:129889      Length:129889      Length:129889      Length:129889      Length:129889      Length:129889      Min.   : 1.00
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 9.00
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :13.00
                                                                                                                 Mean   :12.99
                                                                                                                 3rd Qu.:17.00
                                                                                                                 Max.   :23.00

Departure Delay in Minutes Arrival Delay in Minutes Flight cancelled   Flight time in minutes Flight Distance  Arrival Delay greater 5 Mins
Min.   :  0.00             Min.   :  0.00           Length:129889      Min.   :  8.0          Min.   :  31.0   Length:129889
1st Qu.:  0.00             1st Qu.:  0.00           Class :character   1st Qu.: 59.0          1st Qu.: 362.0   Class :character
Median :  0.00             Median :  0.00           Mode  :character   Median : 92.0          Median : 630.0   Mode  :character
Mean   : 14.98             Mean   : 15.37                              Mean   :111.5          Mean   : 793.8
3rd Qu.: 13.00             3rd Qu.: 13.00                              3rd Qu.:142.0          3rd Qu.:1024.0
Max.   :1592.00            Max.   :1584.00                             Max.   :669.0          Max.   :4983.0
NA's   :2345               NA's   :2738                                NA's   :2738
```

Using the summary() command it is observed that there are **3 NA values** in the **"Satisfaction"** column, **2345 NA values** in the **"Departure Delay in Minutes"** column, and **2738 NA values** in the **"Arrival Delay in Minutes**" column. The NA's in Arrival delay and Departure delay can be explained by the fact that there are a lot of flights that are cancelled. As seen below, there are **2401 flights** that are cancelled in the dataset.

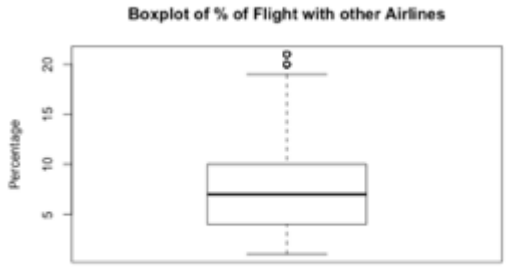
```
> summary(Satisfaction_Survey$`Flight cancelled` == "Yes")
   Mode   FALSE    TRUE
logical  127488    2401
```

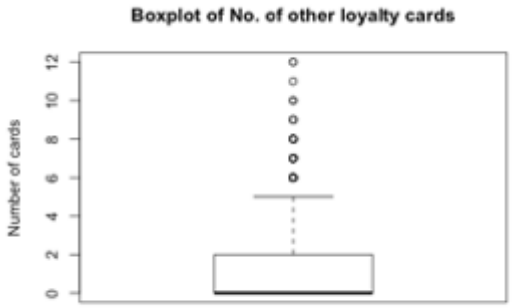
To analyze the flights that successfully reached the destination, the team cleaned out all the flights that were cancelled from total flights. Following was the clean dataset that the team created:

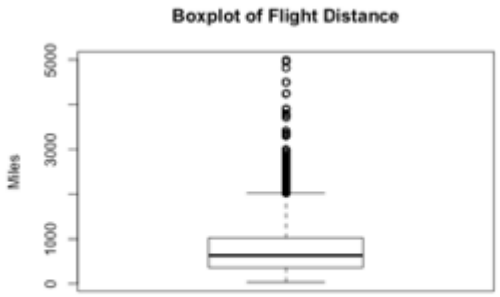
```
> summary(s_not_cancelled)
  Satisfaction   Airline Status          Age          Gender          Price Sensitivity No of Flights p.a. % of Flight with other Airlines Type of Travel
Min.   :1.000   Length:127148      Min.   :15.00   Length:127148      Min.   :0.000     Min.   :  0.00     Min.   : 1.000               Length:127148
1st Qu.:3.000   Class :character   1st Qu.:33.00   Class :character   1st Qu.:1.000     1st Qu.:  9.00     1st Qu.: 4.000               Class :character
Median :4.000   Mode  :character   Median :45.00   Mode  :character   Median :1.000     Median : 17.00     Median : 7.000               Mode  :character
Mean   :3.384                      Mean   :46.16                      Mean   :1.275     Mean   : 20.04     Mean   : 8.324
3rd Qu.:4.000                      3rd Qu.:58.00                      3rd Qu.:2.000     3rd Qu.: 29.00     3rd Qu.:10.000
Max.   :5.000                      Max.   :85.00                      Max.   :5.000     Max.   :100.00     Max.   :21.000
No. of other Loyalty Cards Shopping Amount at Airport Eating and Drinking at Airport    Class           Day of Month   Airline Name       Departure Delay in Minutes
Min.   : 0.0000            Min.   :0.0000            Min.   :  0.00                Length:127148      Min.   : 1.00   Length:127148      Min.   :   0.00
1st Qu.: 0.0000            1st Qu.:0.0000            1st Qu.: 30.00                Class :character   1st Qu.: 8.00   Class :character   1st Qu.:   0.00
Median : 0.0000            Median :0.0000            Median : 60.00                Mode  :character   Median :16.00   Mode  :character   Median :   0.00
Mean   : 0.8838            Mean   :0.4969            Mean   : 66.37                                   Mean   :15.78                      Mean   :  14.91
3rd Qu.: 2.0000            3rd Qu.:1.1526            3rd Qu.: 90.00                                   3rd Qu.:23.00                      3rd Qu.:  13.00
Max.   :12.0000            Max.   :1.1526            Max.   :180.00                                   Max.   :31.00                      Max.   :1592.00
Arrival Delay in Minutes Flight cancelled   Flight time in minutes Flight Distance  Number_of_years
Min.   :  0.00           Length:127148      Min.   :  8.0         Min.   :  31.0   Min.   :0.000
1st Qu.:  0.00           Class :character   1st Qu.: 59.0         1st Qu.: 363.0   1st Qu.:1.000
Median :  0.00           Mode  :character   Median : 92.0         Median : 631.0   Median :4.000
Mean   : 15.29                              Mean   :111.5         Mean   : 795.8   Mean   :4.211
3rd Qu.: 13.00                              3rd Qu.:142.0         3rd Qu.:1027.0   3rd Qu.:7.000
Max.   :500.00                              Max.   :669.0         Max.   :4983.0   Max.   :9.000
>
```

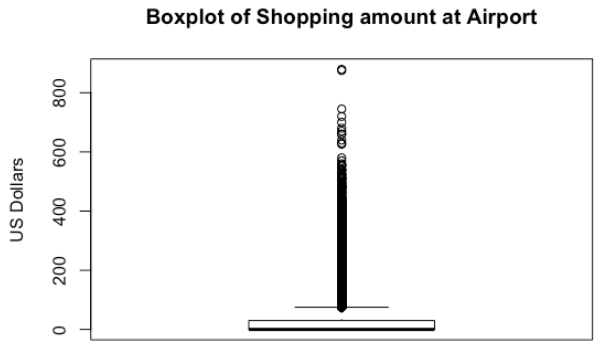
Once the preliminary data cleaning was done, we searched for outliers which we might need to remove from the dataset to improve the accuracy of the linear regression model.

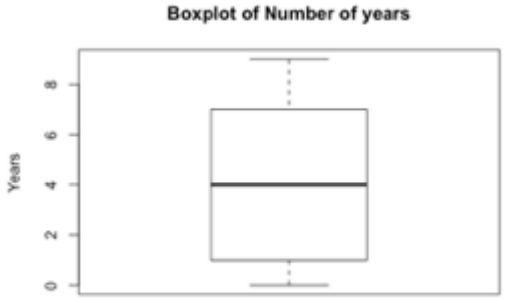
We used Boxplots to view the outliers:











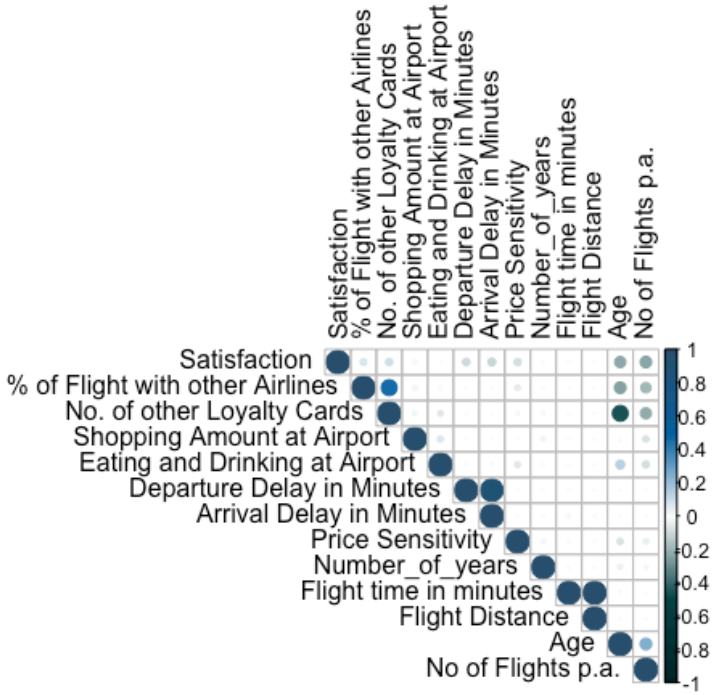
Before even removing outliers, one of the major concerns that the team had was difference in scaling of different variables. On one hand, there was **"satisfaction index**" which **varies from 1-5**, and then there were variables such as **"arrival delay in minutes" varying from 0-500**, and **"flight distance" varying from 0-5000**. This difference in scaling would result in discrepancies within the linear model.

To solve this problem, the team decided to transform all the variables using **Z-transformation.** Following was the method of Z-transformation:

```
#### Z-score Normalization ####
# select numeric variables
SZ <- s_not_cancelled
SZ$Age <- as.numeric(paste(SZ$Age))
SZ$`Price Sensitivity` <- as.numeric(paste(SZ$`Price Sensitivity`))
SZ$`No of Flights p.a.` <- as.numeric(paste(SZ$`No of Flights p.a.`))
SZ$`No. of other Loyalty Cards` <- as.numeric(paste(SZ$`No. of other Loyalty Cards`))
SZ$`Departure Delay in Minutes` <- as.numeric(paste(SZ$`Departure Delay in Minutes`))
SZ$`Flight time in minutes` <- as.numeric(paste(SZ$`Flight time in minutes`))
SZ$`Flight Distance` <- as.numeric(paste(SZ$`Flight Distance`))
SZ$`% of Flight with other Airlines` <- as.numeric(paste(SZ$`% of Flight with other Airlines`))
SZ$`Arrival Delay in Minutes` <- as.numeric(paste(SZ$`Arrival Delay in Minutes`))
SZ$`Shopping Amount at Airport` <- as.numeric(paste(SZ$`Shopping Amount at Airport`))
SZ$`Eating and Drinking at Airport` <- as.numeric(paste(SZ$`Eating and Drinking at Airport`))

library("dplyr")
Z <- select_if(SZ, is.numeric)
Z$`Day of Month` <- NULL
Z <- scale(Z)
Z <- cbind(Z,select_if(SZ, is.character),select_if(SZ, is.integer))
# The data frame "Z" is the transformed dataset which is used in the linear model
Final_Dataset <- Z
```

Once the Z-transformation is done, the outliers were removed. All the values **above 2 Standard Deviations (more than 95%)** were considered as outliers, hence we removed them to improve the predictability of the model. Moreover, while removing outliers, one thing that we observed was that there were outliers only on the upper side across almost all variables. Hence the outliers were removed.

Once the outliers were removed, our next concern was **Multicollinearity.** Keeping highly correlated variables together in the explanatory side of the linear model results in inaccuracies, because the coefficients generated in such a linear model deviates from the real characteristic of that variable in explaining the dependent variable. Following was the correlation matrix that we observed:



The variables **"% of flights with other airlines" and "Number of Loyalty Cards", "Departure delay in Minutes" and "Arrival delay in Minutes", "Flight time in minutes" and "Flight distance"** are highly correlated with each other. So, the variables that we **dropped** from the linear model are:

1. **Number of loyalty cards** - since age is a much more important factor to analyze
2. **Departure delay in minutes** - since arrival delay is a much more important factors that passengers care about
3. **Flight distance** - Because flight time is a much more important predictor

The final correlation matrix is shown below:

4. DATA ANALYSIS

4.1 Multiple Linear Regression Analysis

To understand the relationship between satisfaction index and other attributes in the data, the team decided to build a multiple linear regression model. Once the relationship between satisfaction index and other variables was explained, we were able to approach certain business questions to build the foundation of the core strategy that we built for Southeast Airlines Co.

After cleaning the dataset, we plotted the final correlation matrix as seen in the figure above. Also considering the categorical variables that we thought were important for the model, for the first iteration, following were the explanatory variables that were finalized:

1. Age
2. Price sensitivity
3. Shopping amount at Airport
4. Eating and drinking at Airport
5. Day of month
6. Arrival delay in minutes
7. Flight time in minutes
8. Number of flights per annum
9. Airline status
10. Gender
11. Type of travel
12. Class
13. Airline name
14. Number of years

First Iteration:

The results from the linear model were as follows:

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1567779	0.0099372	15.777	2E-16
Z$Age	-0.0406698	0.0024196	-16.809	2E-16
Z$`Price Sensitivity`	-0.0222363	0.0023285	-9.55	2E-16
Z$`Shopping Amount at Airport`	0.0125943	0.0025647	4.911	9.0855E-07
Z$`Eating and Drinking at Airport`	-0.004559	0.0024555	-1.857	0.063368
Z$`Day of Month`	0.0003424	0.0002462	1.391	0.164298
Z$`Arrival Delay in Minutes`	-0.1523555	0.0030797	-49.471	2E-16
Z$`Flight time in minutes`	-0.0032836	0.0024972	-1.315	0.188549
Z$`No of Flights p.a.`	-0.049981	0.0024197	-20.656	2E-16
Z$`Airline Status`Gold	0.4558827	0.0078639	57.971	2E-16
Z$`Airline Status`Platinum	0.2750426	0.0122245	22.499	2E-16
Z$`Airline Status`Silver	0.643022	0.0054875	117.18	2E-16
Z$GenderMale	0.136064	0.0044387	30.654	2E-16
Z$`Type of Travel`Mileage tickets	-0.1505805	0.0081696	-18.432	2E-16
Z$`Type of Travel`Personal Travel	-1.114512	0.0052361	-212.85	2E-16
Z$ClassEco	-0.079147	0.0077512	-10.211	2E-16
Z$ClassEco Plus	-0.0723589	0.0099579	-7.266	3.71E-13
Z$`Airline Name`Cool&Young Airlines Inc.	0.0675172	0.0217734	3.101	0.00193
Z$`Airline Name`EnjoyFlying Air Services	0.0105663	0.0094531	1.118	0.263672
Z$`Airline Name`FlyFast Airways Inc.	0.0186058	0.0078664	2.365	0.018021
Z$`Airline Name`FlyHere Airways	0.0108679	0.0160422	0.677	0.498116
Z$`Airline Name`FlyToSun Airlines Inc.	0.0301018	0.0139763	2.154	0.031259
Z$`Airline Name`GoingNorth Airlines Inc.	-0.0519042	0.0196754	-2.638	0.00834
Z$`Airline Name`Northwest Business Airlines	0.025106	0.0080724	3.11	0.001871
Z$`Airline Name`OnlyJets Airlines Inc.	0.0094252	0.0115288	0.818	0.41362
Z$`Airline Name`Oursin Airlines Inc.	0.02694	0.0089405	3.013	0.002585
Z$`Airline Name`Paul Smith Airlines Inc.	0.0267386	0.008462	3.16	0.001579
Z$`Airline Name`Sigma Airlines Inc.	0.0272093	0.0075243	3.616	0.000299
Z$`Airline	0.0288294	0.0091325	3.157	0.001595

Name`Southeast Airlines Co.				
Z$`Airline Name`West Airways Inc.	0.0819638	0.0190202	4.309	1.639E-05
Z$Number_of_years	0.0146102	0.002122	6.885	5.804E-12
Z$`% of Flight with other Airlines`	-0.0011197	0.0024603	-0.455	0.649043

Residual standard error: 0.7546 on 127116 degrees of freedom
Multiple R-squared: 0.4307, Adjusted R-squared: 0.4305
F-statistic: 3102 on 31 and 127116 DF, p-value: < 0.00000000000000022

There are **7 variables whose P-values are higher than alpha = 0.05 (Level of confidence = 95%)**.

For the second iteration:

1. **Removed "Eating and drinking at Airport"** due to high P-value
2. **Removed "Day of Month"**
3. **Removed "Flight time in minutes"**
4. **Removed "% of Flights with other Airlines"**
5. **Kept `Airline Name`EnjoyFlying Air Services, `Airline Name`FlyHere Airways, `Airline Name`OnlyJets Airlines Inc.** as they are levels of the variable "Airline Name"

Second Iteration:

Coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.163645	0.009116	17.952	2E-16
Z$Age	-0.040994	0.002336	-17.549	2E-16
Z$`Price Sensitivity`	-0.021992	0.002322	-9.471	2E-16
Z$`Shopping Amount at Airport`	0.012298	0.00256	4.803	1.5617E-06
Z$`Arrival Delay in Minutes`	-0.152884	0.003064	-49.901	2E-16
Z$`No of Flights p.a.`	-0.049351	0.002387	-20.674	2E-16
Z$`Airline Status`Gold	0.454715	0.007834	58.042	2E-16
Z$`Airline Status`Platinum	0.27363	0.012197	22.434	2E-16
Z$`Airline Status`Silver	0.642075	0.005454	117.716	2E-16
Z$GenderMale	0.135663	0.004401	30.826	2E-16
Z$`Type of Travel`Mileage tickets	-0.150765	0.008169	-18.457	2E-16
Z$`Type of Travel`Personal Travel	-1.115242	0.005218	-213.73	2E-16
Z$ClassEco	-0.079178	0.007751	-10.215	2E-16
Z$ClassEco Plus	-0.07247	0.009945	-7.287	3.18E-13
Z$`Airline Name`Cool&Young Airlines Inc.	0.063831	0.021617	2.953	0.00315
Z$`Airline Name`EnjoyFlying Air Services	0.011758	0.009408	1.25	0.21142
Z$`Airline Name`FlyFast Airways Inc.	0.019886	0.007812	2.546	0.01091
Z$`Airline Name`FlyHere	0.010605	0.016041	0.661	0.50854

Airways				
Z$`Airline Name`FlyToSun Airlines Inc.	0.027316	0.01384	1.974	0.04842
Z$`Airline Name`GoingNorth Airlines Inc.	-0.052528	0.019668	-2.671	0.00757
Z$`Airline Name`Northwest Business Airlines	0.026189	0.008018	3.266	0.00109
Z$`Airline Name`OnlyJets Airlines Inc.	0.00753	0.011424	0.659	0.50983
Z$`Airline Name`Oursin Airlines Inc.	0.023893	0.008656	2.76	0.00578
Z$`Airline Name`Paul Smith Airlines Inc.	0.024742	0.008338	2.967	0.003
Z$`Airline Name`Sigma Airlines Inc.	0.026212	0.007495	3.497	0.00047
Z$`Airline Name`Southeast Airlines Co.	0.02769	0.009099	3.043	0.00234
Z$`Airline Name`West Airways Inc.	0.082946	0.018998	4.366	1.2669E-05
Z$Number_of_years	0.014636	0.002122	6.898	5.296E-12

Residual standard error: 0.7546 on 127120 degrees of freedom
Multiple R-squared: 0.4306, Adjusted R-squared: 0.4305
F-statistic: 3561 on 27 and 127120 DF, p-value: < 0.00000000000000022

There are **3 variables whose P-values are higher than alpha = 0.05 (Level of confidence = 95%)**. But since they are levels of the same variable, we need to either keep all of them, or remove the variable "Airline Name" from the model.

We tried removing "Airline name" from the linear model, however **the adjusted R-squared value due to removal of "Airline Name" from the model dropped down to 0.4303. Hence we do not remove it.**

Interpretation of the Linear Model:

From the model, the team interpreted the following:

1. Following are the respective base levels of the following variables:
    a) Airline Status - Blue
    b) Gender - Female
    c) Type of travel - Business
    d) Class - Business
    e) Airline Name - Cheapseats Airlines Inc.
2. Following are the key interpretations:

Variable	**Change in variable**	**Effect on Satisfaction Level**
Age	Increase	Decrease
Price Sensitivity	Increase	Decrease
Shopping Amount at Airport	Increase	Increase

Number of flights per annum	Increase	Decrease
Airline Status	Change in Status	Silver > Gold > Platinum > Blue
Gender	Change in gender	Male > Female
Type of travel	Change in type of travel	Business > Mileage > Personal
Class	Change in class	Business > Eco Plus > Eco
Number of years	Increase	Increase

Following is the order of satisfaction among different airlines (High to Low):

1. West Airways Inc.
2. Cool&Young Airlines Inc.
3. Southeast Airlines Inc.
4. FlyToSun Airlines Inc.
5. Sigma Airlines Inc.
6. Northwest Business Airlines Inc.
7. Paul Smith Airlines Inc.
8. Oursin Airlines Inc.
9. FlyFast Airlines Inc.
10. EnjoyFlying Air Services
11. FlyHere Airways
12. OnlyJets Airlines Inc.
13. Cheapseats Airlines Inc.

Recommendation:

- Satisfaction level for the client is high among their passengers, so they should now focus on the segment of passengers who are price sensitive, and implement flexibility of pricing
- There must be more capital invested into improving the service in economy class and blue status carriers, which is targeted towards the personal travel segment
- There must be more marketing done for the silver and gold status carriers, as well as the business class seats; the target segment for this marketing must be mileage travel and business travel passengers

4.2 Association Rules

After analyzing the data using multiple linear regression analysis, we can conclude that Age, Price Sensitivity, Number of flights per annum negatively impact the satisfaction index. However, we needed to validate the claims made by the linear model. Due to the type of data collected in the survey, there were a lot of categorical variables. This might have caused a variation in the results of the linear regression from the reality. Hence, we mined association rules to understand the effect on satisfaction index through various combination of factors.

Association rules method requires that the data type of variables should be categorical variables, so we decided to classify other numeric variables to categorical variables by bucketing them into categories in adherence to their distribution. This is how they were classified:

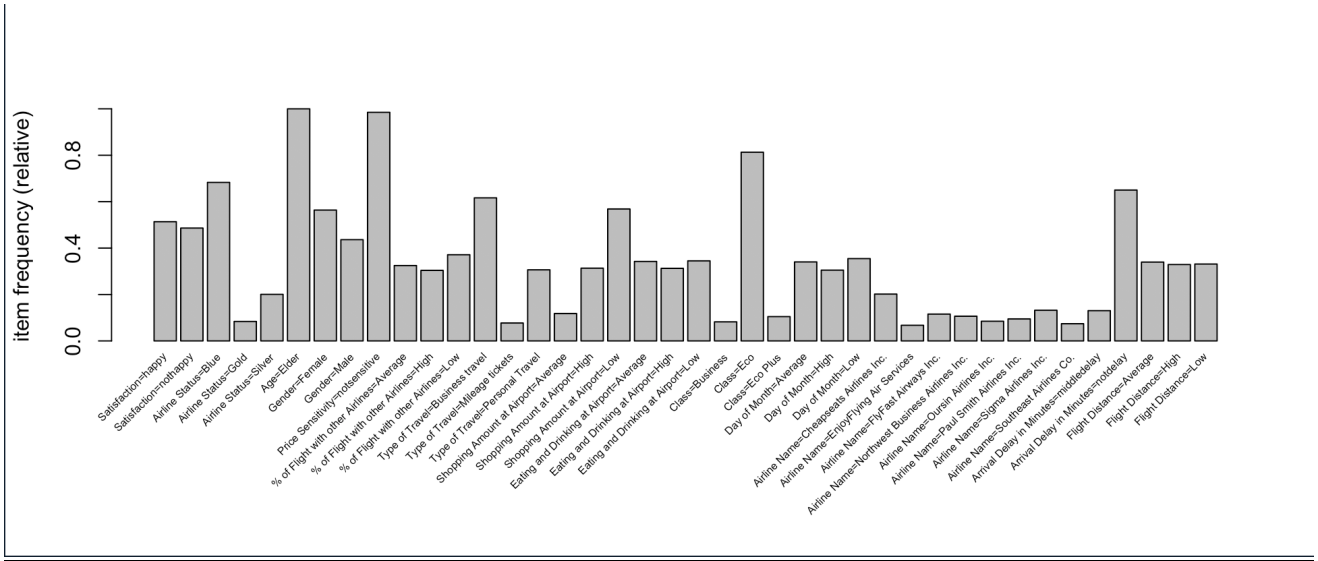
```
C$Satisfaction[C$Satisfaction >= 4] <- 'Satisfaction'
C$Satisfaction[C$Satisfaction < 4 & C$Satisfaction >=3] <- 'Average'
C$Satisfaction[C$Satisfaction < 3] <- 'Dissatisfaction'
C$Satisfaction = as.factor(C$Satisfaction)

C$Age[C$Age >= 55] = 'Elder'
C$Age[C$Age > 30 & C$Age<=55 ] = 'MiddleAgedpeople'
C$Age[C$Age<= 30] = 'Youngpeople'

C$`Arrival Delay in Minutes`[C$`Arrival Delay in Minutes`>5]  ='delay'
C$`Arrival Delay in Minutes`[C$`Arrival Delay in Minutes`<=5] ='notdelay'

category <- function(vec){
  q <- quantile(vec, c(0.33, 0.67))
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}
```

We defined the satisfaction level which is equal and greater than 4 as “Satisfaction”, between 3 to 4 as “Average”, and less than 3 as “Dissatisfaction”. For Age variable, we used 30 and 55 as cutoffs to distinguish passengers to Young people, Middle aged people and Elder. For price sensitivity, an index greater than 3 were defined as sensitive, and less than 3 were defined as not sensitive. For arrival delay in minutes, less than 5 minutes were classified as “Not Delayed”, and greater than 5 minutes were classified as “Delayed”. For other variables, we used quartiles to classify the variables.

Moreover, we removed the columns that were irrelevant to the association rules analysis:

1. **Day of month:** Service providers cannot change their service offerings across different days of the month
2. **Airline name:** We do not focus on the carrier or company in association rules mining; rather we analyze all carriers
3. **Departure delay in minutes:** Passengers tend to focus more upon arrival delay than departure delay
4. **Flight cancelled:** We mined the association rules for both cancelled as well as successful flights

After classifying the variables, we drew the item frequency plot to see all the categorical variables:

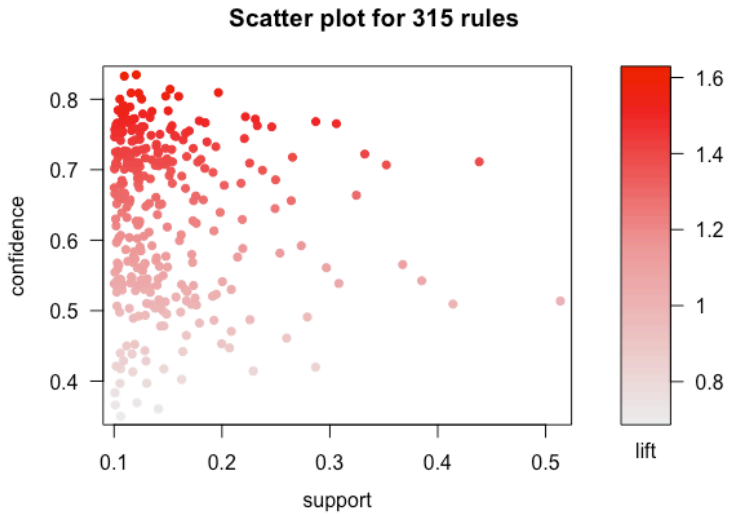


In order to find the association between those variables, we used the apriori algorithm to identify the association between these variables and the degree of satisfaction of the customer. To find out the association rules, we declare the level of satisfaction we wish to analyze on the left hand side (LHS), and all the variables which we want to check the association with on the right hand side (RHS).

4.1.1 High-level Satisfaction

First, we set “Satisfaction=Satisfaction” on the right hand side. Then we set the minimum support as 0.1 and minimum confidence as 0.3. Then we draw a plot to see the lift of rules.

```
#predict happy customers (as defined by their overall satisfaction >=4).
ruleset <- apriori(CX,parameter=list(support=0.1, confidence=0.3),appearance = list(rhs="Satisfaction=Satisfaction",default="lhs"))
plot(ruleset,jitter=0)
goodrules <- ruleset[quality(ruleset)$lift > 1.5]
inspect(goodrules)
top.lift <- sort(goodrules, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift, 10))
```



We sort the ruleset by the value of lift and choose the top rules which have higher lift.

The ruleset of “Satisfaction=High”:

```
> inspect(head(top.lift, 10))
     lhs                                    rhs                           support confidence     lift count
[1]  {Airline Status=Silver,
      Type of Travel=Business travel}    => {Satisfaction=Satisfaction} 0.1205996  0.8346851 1.625047 15334
[2]  {Airline Status=Silver,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1095023  0.8327153 1.621212 13923
[3]  {Age=MiddleAgedpeople,
      Price Sensitivity=Low,
      Type of Travel=Business travel,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1519332  0.8141436 1.585054 19318
[4]  {Age=MiddleAgedpeople,
      Type of Travel=Business travel,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1966763  0.8094190 1.575856 25007
[5]  {Age=MiddleAgedpeople,
      Gender=Male,
      Price Sensitivity=Low,
      Type of Travel=Business travel}    => {Satisfaction=Satisfaction} 0.1157234  0.8088615 1.574771 14714
[6]  {Age=MiddleAgedpeople,
      Price Sensitivity=Low,
      Type of Travel=Business travel,
      Class=Eco,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1228254  0.8085426 1.574150 15617
[7]  {Age=MiddleAgedpeople,
      Gender=Male,
      Type of Travel=Business travel}    => {Satisfaction=Satisfaction} 0.1478985  0.8043887 1.566063 18805
[8]  {Age=MiddleAgedpeople,
      Type of Travel=Business travel,
      Class=Eco,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1598295  0.8040356 1.565375 20322
[9]  {Age=MiddleAgedpeople,
      Type of Travel=Business travel,
      Shopping Amount at Airport=Low,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Satisfaction} 0.1053339  0.8002031 1.557914 13393
[10] {Age=MiddleAgedpeople,
      Gender=Male,
      Type of Travel=Business travel,
      Class=Eco}                         => {Satisfaction=Satisfaction} 0.1255545  0.8002005 1.557909 15964
```

Insights:

1. Passengers who are travelling through business class are highly satisfied
2. Traits such as low price sensitivity, economical, someone who values reaching on time seem to give a good satisfaction score
3. Highest lifts are given to the rules: have a silver airline status, is a middle-aged person, preferably a male, and someone travelling through business class. The support indicates that they have given highest satisfaction scores

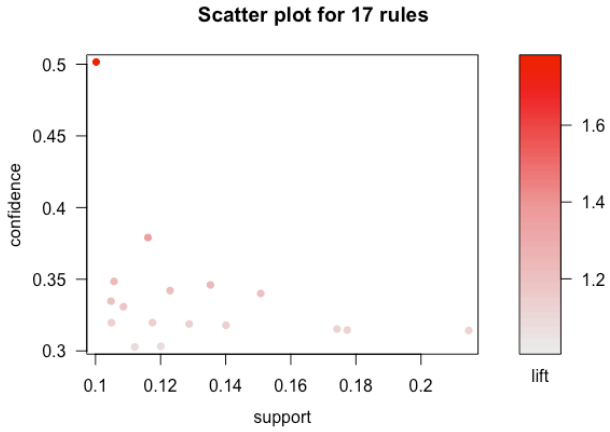
Recommendation:

- The client could focus on marketing the silver status package more, to improve sales. Even if they evolve the pricing strategy into a dynamic pricing, the cluster of passengers who are on business travel, or are not price sensitive, would be attracted due to the services that are available to them
- The airline routes which have lesser traffic should nest a silver or gold status airline, as the target customer demographics would comprise of older people, who highly value arrival-on-time and good hospitality
- The client should focus on creating partnership with any Multinational Companies, and developing bundle-packages for them in the business class segment. The target segment comprises of many middle-aged business class travelers, who would be highly satisfied

4.2.1 Average satisfaction

After analyzing the association rules of passengers who have higher satisfaction degree, we want to explore which passengers are more likely to give a moderate satisfaction degree. we set “Satisfaction=average” on the right hand side. Then we set the minimum support as 0.1 and minimum confidence as 0.1. Then we draw a plot to see the lift of rules.

```
ruleset <- apriori(CX,parameter=list(support=0.1, confidence=0.1),appearance = list(rhs="Satisfaction=Average",default="lhs"))
plot(ruleset,jitter=0)
goodrules <- ruleset[quality(ruleset)$lift > 1]
inspect(goodrules)
top.lift <- sort(goodrules, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift, 10))
```



We sort the rulesets by the value of lift and choose the top rules which have higher lift.

The ruleset of “Satisfaction=Average”:

```
     lhs                                    rhs                      support confidence     lift count
[1]  {Type of Travel=Personal Travel,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Average} 0.1001982  0.5015748 1.779266 12740
[2]  {Type of Travel=Personal Travel}    => {Satisfaction=Average} 0.1161245  0.3790855 1.344752 14765
[3]  {Airline Status=Blue,
      Gender=Female,
      Class=Eco}                         => {Satisfaction=Average} 0.1056800  0.3485061 1.236276 13437
[4]  {Airline Status=Blue,
      Gender=Female}                     => {Satisfaction=Average} 0.1353148  0.3460309 1.227496 17205
[5]  {Airline Status=Blue,
      Class=Eco,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Average} 0.1229119  0.3420966 1.213540 15628
[6]  {Airline Status=Blue,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Average} 0.1507535  0.3400873 1.206412 19168
[7]  {Age=Elder}                         => {Satisfaction=Average} 0.1047598  0.3346818 1.187237 13320
[8]  {No of Flights p.a.=High}           => {Satisfaction=Average} 0.1086057  0.3308258 1.173558 13809
[9]  {Gender=Female,
      Arrival Delay in Minutes=notdelay} => {Satisfaction=Average} 0.1175087  0.3197372 1.134223 14941
[10] {Airline Status=Blue,
      Shopping Amount at Airport=Low,
      Class=Eco}                         => {Satisfaction=Average} 0.1048935  0.3197171 1.134152 13337
```

Insights:

1. Females who travel through blue airline status are predominant in this customer segment

2. Passengers who travel through blue airline status do not shop much from the airport
3. Blue airline carriers can trigger a relatively higher satisfaction level by focusing on arrival on time

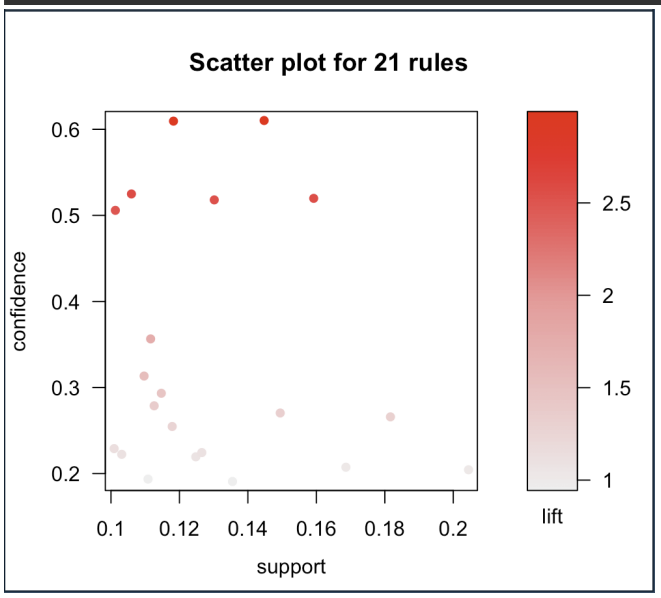
Recommendation:

- Low-priced business class plans must be created for this segment. This would result in a higher satisfaction level
- The client must focus on improving service in the blue status airlines, which should improve the satisfaction level in the passengers in this segment
- Females traveling through economy give a strong priority on the overall value associated with the airline's service. The client should consider creating value-oriented bundle plans, or lowering the ticket prices, specifically for women

4.3.1 Low-level Satisfaction

Now, we focus on the passengers who have low satisfaction level. We set "Satisfaction=Dissatisfaction" on the right hand side.

```
ruleset <- apriori(CX,parameter=list(support=0.1, confidence=0.1),appearance = list(rhs="Satisfaction=Dissatisfaction",default="lhs"))
plot(ruleset,jitter=0)
goodrules <- ruleset[quality(ruleset)$lift > 1.5]
inspect(goodrules)
top.lift <- sort(goodrules, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift, 10))
```



We sort the rulesets by the value of lift and choose the top rules which have higher lift.

```
> inspect(head(top.lift, 10))
    lhs                                  rhs                                 support confidence     lift count
[1] {Airline Status=Blue,
     Type of Travel=Personal Travel} => {Satisfaction=Dissatisfaction} 0.1447447  0.6102122 2.984470 18404
[2] {Airline Status=Blue,
     Type of Travel=Personal Travel,
     Class=Eco}                      => {Satisfaction=Dissatisfaction} 0.1182638  0.6095010 2.980992 15037
[3] {Gender=Female,
     Type of Travel=Personal Travel} => {Satisfaction=Dissatisfaction} 0.1059789  0.5248909 2.567174 13475
[4] {Type of Travel=Personal Travel} => {Satisfaction=Dissatisfaction} 0.1592239  0.5197823 2.542189 20245
[5] {Type of Travel=Personal Travel,
     Class=Eco}                      => {Satisfaction=Dissatisfaction} 0.1301790  0.5179946 2.533445 16552
[6] {Price Sensitivity=Low,
     Type of Travel=Personal Travel} => {Satisfaction=Dissatisfaction} 0.1012835  0.5058329 2.473964 12878
[7] {Age=Elder}                      => {Satisfaction=Dissatisfaction} 0.1115865  0.3564914 1.743554 14188
[8] {Arrival Delay in Minutes=delay} => {Satisfaction=Dissatisfaction} 0.1096753  0.3133426 1.532519 13945
```

Insights:

1. Passengers who travel through Blue airline status, and are on personal travel tend to be dissatisfied
2. Passengers availing economy seats always tend to be dissatisfied

Recommendation:

- The client must focus upon improving services for the blue airline status, as well as the economy class seats
- Lower price setting for business class seats in blue airline status might trigger a higher satisfaction level

4.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

We base the model on previous outcomes from linear model and association rules to help us make decisions which variable should be added into SVM model. The key variables over which satisfaction is dependent are Class, Type of Travel, Airline Status, Age, Arrival Delay in Minutes, Price Sensitivity, Shopping Amount at Airport and Gender. For developing a predictive model through SVM algorithm, we classify satisfaction in 2 levels, unlike the previous analysis methods. The first level is “Satisfaction”, when the Satisfaction Index is equal or higher than 4. The second level is “Dissatisfaction” when the grade is less than 4. We wanted to use SVM model to validate the precision of our previous models and hope to use SVM to predict passengers’ Satisfaction Index.

Firstly, we divided the dataset into a training set and a test set. We used **two-thirds of the data set to train and the remainder to test**. We generated a randomized index that let us choose cases for our training and test sets.Then, we created a new vector variable that randomly sampled portions of the dataset “S”, including the complete range of the data to the original dataset. Next, we calculated the cut point that would divide ⅔ of data i.e., the train set and ⅓ of data i.e., the test set based on the number of rows in the original dataset.
After dividing “S” into two separate data sets, we trained our support vector model.

```
##############SVM#################
###### not use z score  ########
library("dplyr")
S <- s_not_cancelled[which(s_not_cancelled$`Airline Name`=='Cheapseats Airlines Inc.'),]
S$Satisfaction[S$Satisfaction >= 4] <- 'Satisfaction'
S$Satisfaction[S$Satisfaction < 4] <- 'Dissatisfaction'
S$Satisfaction = as.factor(S$Satisfaction)

randomindex = sample(1:nrow(S))
cutpoint = floor(2*nrow(S)/3)
traindata = S[randomindex[1:cutpoint],]
testdata = S[randomindex[(cutpoint+1):nrow(S)],]
dim(traindata)
str(S)
#install.packages("kernlab")
library(kernlab)
svmOutput <- ksvm(Satisfaction ~Class+`Type of Travel`+`Airline Status`+Age
                  +`Arrival Delay in Minutes`+`Price Sensitivity`
                  +`Shopping Amount at Airport`+Gender,data=traindata,
                  kernel= "rbfdot", kpar = "automatic", C = 50, cross = 3, prob.model = TRUE)
svmOutput
svmPred <- predict(svmOutput, testdata, type = "votes")
compTable <- data.frame(data.frame(testdata$Satisfaction=='Satisfaction'),svmPred[1,])
table(compTable)
res <- table(compTable)
# Calculate an error rate
errorRate <- (res[1,1]+res[2,2])/(sum(res))
errorRate
```

```
> svmOutput
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 50

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.101800930134435

Number of Support Vectors : 8033

Objective Function Value : -361313.6
Training error : 0.199626
Cross validation error : 0.225573
Probability model included.
```

Validation:

```
> table(compTable)
                                    svmPred.1...
testdata.Satisfaction.....Satisfaction.    0    1
                                  FALSE 1547 2767
                                  TRUE  3859  384
```

	0	1
FALSE	1547	2767
TRUE	3859	384

To validate the precision of the SVM model, we calculated the error rate of this model.
The error rate is 22.56%, which indicates how many predictions deviated from the real-life result.

```
> errorRate <- (res[1,1]+res[2,2])/(sum(res))
> errorRate
[1] 0.2256632
```

4.4 Decision Tree Model

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. We used the decision tree model to help us analyze what elements are decisive and we would like to explain to our clients with easy-understanding and visualization ways. The following pictures are the codes used to create this model and visualize its outcomes.

```
##########decision treeeeeeeeeeeee
dt = s_not_cancelled
dt$Satisfaction[dt$Satisfaction >= 4] <- 'Satisfaction'
dt$Satisfaction[dt$Satisfaction < 4 & dt$Satisfaction >=3 ] <- 'Average'
dt$Satisfaction[dt$Satisfaction <= 2] <- 'Dissatisfaction'
set.seed(1888)
install.packages("rpart")
library(rpart)
randomindex1 = sample(1:nrow(dt), size = 10000)
cutpoint1 = floor(2*10000/3)
traindata1 = dt[randomindex1[1:cutpoint1],]
testdata1 = dt[cutpoint1+1:10000,]
dtree<-rpart(Satisfaction ~ Class +`Type of Travel`+`Airline Status`+ Age+ `Arrival Delay in Minut
,data=traindata1, method="class", parms=list(split="information"))
#Class+`Type of Travel`+`Airline Status`+Age+`Arrival Delay in Minutes`+`Price Sensitivity`+`Shopp
printcp(dtree)
tree<-prune(dtree,cp=0.0125)
tree<-prune(dtree,cp=dtree$cptable[which.min(dtree$cptable[,"xerror"]),"CP"])
opar<-par(no.readonly = T)
par(mfrow=c(1,2))
install.packages("rpart.plot")
library(rpart.plot)
rpart.plot(dtree,branch=1 ,type=4, fallen.leaves=T,cex=0.7, sub="Decision Tree for Airline Satisfa
rpart.plot(tree, branch=1, type=4,fallen.leaves=T,cex=0.7, sub="Decision Tree for Airline Satisfac
par(opar)
dev.off()
predtree<-predict(tree,newdata=testdata1,type="class")
table(testdata1$Satisfaction,predtree,dnn=c("real one","predict one"))
```

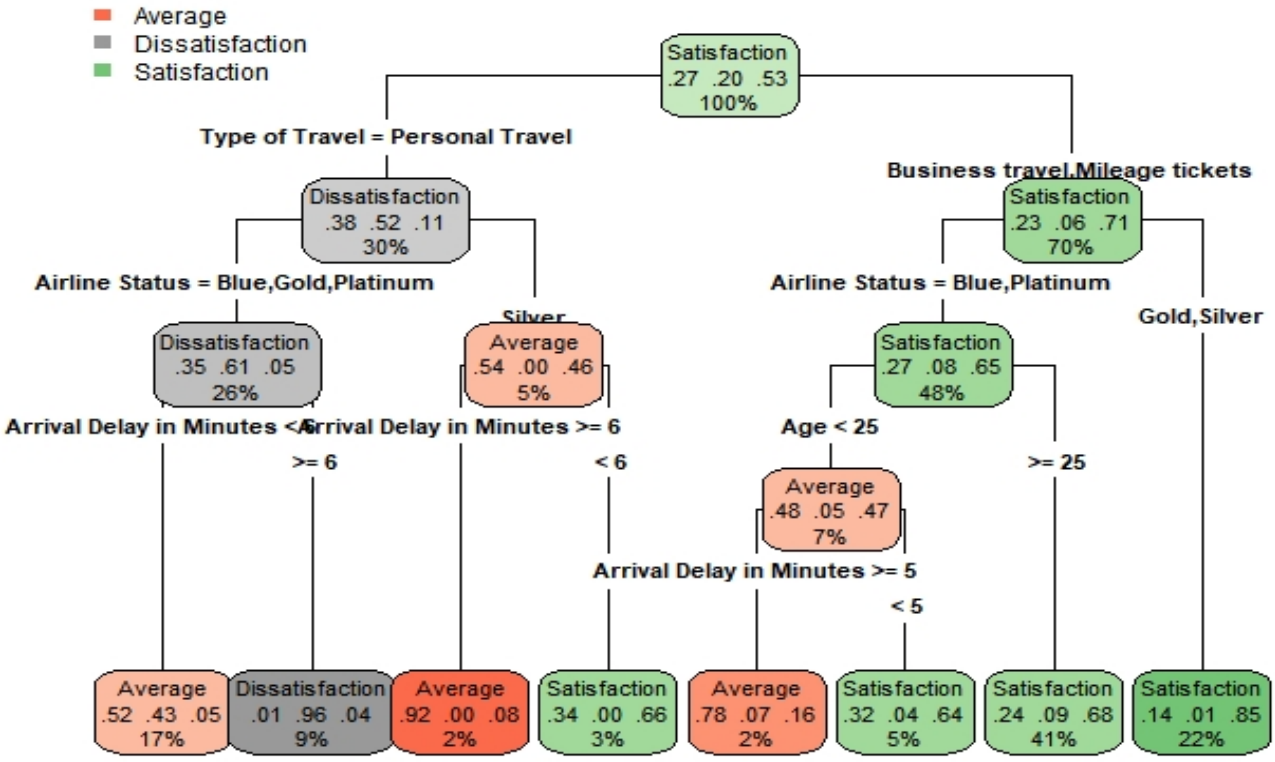
We selected 10000 random samples to develop this model. Two thirds of sample was used as training data and one third of data was used as test data. Based on our previous findings, we chose the variables: Class, Type of Travel, Airline Status, Age, Arrival Delay in Minutes, Price Sensitivity, Shopping Amount at Airport and Gender.

```
D:/2018 fall/687/project/
> testdata1 = dt[cutpoint1:10000,]
> predtree<-predict(tree,newdata=testdata1,type="class")
> table(testdata1$Satisfaction,predtree,dnn=c("real one","predict one"))
                predict one
real one         Average Dissatisfaction Satisfaction
  Average            355              3          517
  Dissatisfaction    208            303          148
  Satisfaction        48              7         1746
```

Real one	Predicted one		
	Average	Dissatisfaction	Satisfaction
Average	355	3	517
Dissatisfaction	208	303	148

Satisfaction	48	7	1746

The result of this model shows in the table above. We calculated the precision rate of this model as 72.07%. It is very close to SVM's precision value. So we not only validated the efficiency of the SVM model, but we could positively use this model to help us make decisions.



Decision Tree for Airline Satisfaction

We interpreted the decision tree as follows:

1. The dependent variable of this decision tree is Satisfaction index, which has 3 levels: Dissatisfaction, Average and Satisfaction
2. The most influential attributes which influence the satisfaction index are type of travel, airline status, age and arrival delay
3. If a customer is on personal travel:
    a) Under blue gold or platinum airline status, if there is a small delay (less than 6 min), most of the population would most likely be moderately satisfied or dissatisfied. However, if the delay exceeds 6 minutes, the passengers are most likely dissatisfied
    b) Under silver airline status, if there is a delay of more than 6 min, the passengers are mostly moderately satisfied. However, if there is a delay of less than 6 min, they are mostly satisfied
4. If a customer is on business travel:
    a) Under blue or platinum airline status, if the customer is on a business or a mileage travel, is less than 25 years of age and faces a delay of more than 5 min, they would most likely be moderately satisfied. However, in case of a delay of less than 5 min, they would most

likely be satisfied. If a customer is more than 25 years of age, they would most likely be satisfied

b) Under gold and silver airline status, passengers in business travel or mileage are most likely satisfied even if they face delays

5. Recommendations

a) There is scope for improvement for client in terms of overall service in the following states: Nevada, Utah, Arizona, New Hampshire, Virginia and Ohio.
b) Satisfaction level for the client is high among their passengers, so they should now focus on the segment of passengers who are price sensitive, and implement flexibility of pricing
c) There must be more capital invested in improving the service in economy class and blue status carriers, which is targeted towards the personal travel segment
d) There must be more marketing done for the silver and gold status carriers, as well as the business class seats; the target segment for this marketing must be mileage travel and business travel passengers
e) Low-priced business class plans must be created for the moderately satisfied segment. This would result in a higher satisfaction level in the middle aged segment
f) Females traveling through economy class give a strong priority on the overall value associated with the airline's service. The client should consider creating value-oriented bundle plans, or lowering the ticket prices, specifically for women
g) Lower price setting for business class seats in blue airline status might trigger a higher satisfaction level

6. Conclusion

The team validated all the models: Multiple linear regression model, Association Rules, Support Vector Machine model and Decision Tree model. All of the models provided us similar insights and pointed us to look at Age, Airline Status, Type of travel, Gender, and Arrival delay. The client must focus on cater to its different target customer segments by keeping these factors in mind, and devise strategies which cater to different segments, created by variations in these factors.

7. Trello Board

For the IST 687 project, we divided the tasks to four main parts (Data preparation, Data analysis, Data visualization, Report & presentation preparation)

A .Data preparation:

All team members participated in understanding, cleaning and munging the dataset.

B .Data analysis: We used four models to analysis the dataset.

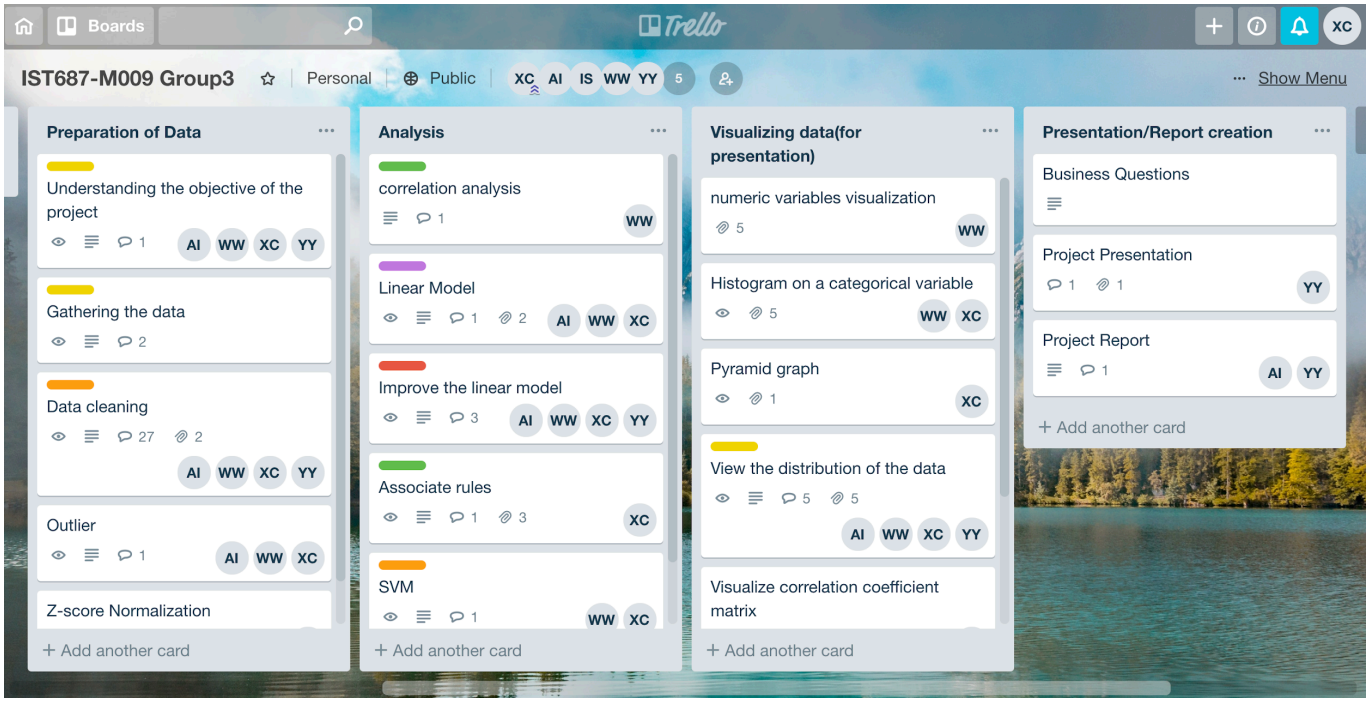
1. Yanqi Yao was responsible for Linear modelling;
2. Xuehan Chen was responsible for association rules;
3. Weicheng Wu was responsible for SVM and Decision tree.
4. Advait Ramesh Iyer was responsible for helping to analyze those models together, and generate business insights

C .Data visualization:

Weicheng Wu and Xuehan Chen were responsible to create descriptive statistics graphs

D .Report & presentation preparation:

1. Advait Ramesh Iyer was responsible for recording the process of the project
2. Yanqi Yao was responsible for preparing the presentation



8. Appendix

```
str(Satisfaction_Survey)
summary(Satisfaction_Survey)
# use summary() to view the distribution of the dataset.
# Then I found that there are four variables
# (Satisfaction, Departure Delay in Minutes,Arrival Delay in Minutes, Flight time in mintues) have NA value.
# So we need to analyse these NA value to decide whether we need to remove them.
s <- Satisfaction_Survey
s$Number_of_years <- s$`Year of First Flight`-2003
s$`Flight date` <- NULL
s$`Airline Code` <- NULL
s$`Orgin City` <- NULL
s$`Origin State` <- NULL
s$`Destination City` <- NULL
s$`Destination State` <- NULL
s$`Scheduled Departure Hour` <- NULL
s$`Year of First Flight` <- NULL
# Removing NA values from the Satisfaction column
s <- s[-which(is.na(s$Satisfaction)),] # remove NA row

# separate data to flight canceled and not cancelled.
```

```
s_cancelled <- subset(s, s$`Flight cancelled` == "Yes")
s_cancelled$`Departure Delay in Minutes` <- NULL
s_cancelled$`Flight time in minutes` <- NULL
s_cancelled$`Arrival Delay in Minutes` <- NULL
summary(s_cancelled)
#remove 3 variables in flight cancelled dataset:Departure Delay in Minutes, arrival Delay in
Minutes,Flight time in minutes
s_not_cancelled <- subset(s, s$`Flight cancelled` == "No")
S_not_cancelled1 <- s_not_cancelled
s_not_cancelled <-s_not_cancelled[-which(is.na(s_not_cancelled$`Flight time in minutes`)),]
s_not_cancelled <- s_not_cancelled[,-ncol(s_not_cancelled)+1]
summary(s_not_cancelled)

# Creating boxplots to view outliers
#% of Flight with other Airlines
summary(s_not_cancelled$`% of Flight with other Airlines`)
boxplot(s_not_cancelled$`% of Flight with other Airlines`, main="Boxplot of % of Flight with other
Airlines", ylab="Percentage")

#Shopping Amount at Airport
summary(s_not_cancelled$`Shopping Amount at Airport`)
boxplot(s_not_cancelled$`Shopping Amount at Airport`, ylim=c(0,2), main="Boxplot of Shopping amount
at Airport", ylab="US Dollars")

#Eating and Drinking at Airport
summary(s_not_cancelled$`Eating and Drinking at Airport`)
boxplot(s_not_cancelled$`Eating and Drinking at Airport`, main="Boxplot of eating and drinking at
Airport", ylab="US Dollars")

#Arrival Delay in Minutes
summary(s_not_cancelled$`Arrival Delay in Minutes`)
boxplot(s_not_cancelled$`Arrival Delay in Minutes`, main="Boxplot of Arrival delay in minutes",
ylab="Minutes")

#Number of other Loyalty cards
summary(s_not_cancelled$`No. of other Loyalty Cards`)
boxplot(s_not_cancelled$`No. of other Loyalty Cards`, main="Boxplot of No. of other loyalty cards",
ylab="Number of cards")

#Number of Flight distance
summary(s_not_cancelled$`Flight Distance`)
boxplot(s_not_cancelled$`Flight Distance`, main="Boxplot of Flight Distance", ylab="Miles")

#Number of years
summary(s_not_cancelled$Number_of_years)
boxplot(s_not_cancelled$Number_of_years, main="Boxplot of Number of years", ylab="Years")

#### Z-score Normalization ####
# select numeric variables
```

```
SZ <- s_not_cancelled
SZ$Age <- as.numeric(paste(SZ$Age))
SZ$`Price Sensitivity` <- as.numeric(paste(SZ$`Price Sensitivity`))
SZ$`No of Flights p.a.` <- as.numeric(paste(SZ$`No of Flights p.a.`))
SZ$`No. of other Loyalty Cards` <- as.numeric(paste(SZ$`No. of other Loyalty Cards`))
SZ$`Departure Delay in Minutes` <- as.numeric(paste(SZ$`Departure Delay in Minutes`))
SZ$`Flight time in minutes` <- as.numeric(paste(SZ$`Flight time in minutes`))
SZ$`Flight Distance` <- as.numeric(paste(SZ$`Flight Distance`))
SZ$`% of Flight with other Airlines` <- as.numeric(paste(SZ$`% of Flight with other Airlines`))
SZ$`Arrival Delay in Minutes` <- as.numeric(paste(SZ$`Arrival Delay in Minutes`))
SZ$`Shopping Amount at Airport` <- as.numeric(paste(SZ$`Shopping Amount at Airport`))
SZ$`Eating and Drinking at Airport` <- as.numeric(paste(SZ$`Eating and Drinking at Airport`))

library("dplyr")
Z <- select_if(SZ, is.numeric)
Z$`Day of Month` <- NULL
Z <- scale(Z)
Z <- cbind(Z,select_if(SZ, is.character),select_if(SZ, is.integer))
# Z is the normalization result and use this dataset to build linear model

########## Removing Outliers #################
Final_Dataset <- Z

ul1 <- quantile(Final_Dataset$Age,0.975)
ll1 <- quantile(Final_Dataset$Age,0.025)
Z[Z$Age<ll1,]$Age <- ll1
Z[Z$Age>ul1,]$Age <- ul1

ul2 <- quantile(Final_Dataset$`Price Sensitivity`,0.975)
ll2 <- quantile(Final_Dataset$`Price Sensitivity`,0.025)
Z[Z$`Price Sensitivity`<ll2,]$`Price Sensitivity` <- ll2
Z[Z$`Price Sensitivity`>ul2,]$`Price Sensitivity` <- ul2

ul3 <- quantile(Final_Dataset$`No of Flights p.a.`,0.975)
ll3 <- quantile(Final_Dataset$`No of Flights p.a.`,0.025)
Z[Z$`No of Flights p.a.`<ll3,]$`No of Flights p.a.` <- ll3
Z[Z$`No of Flights p.a.`>ul3,]$`No of Flights p.a.` <- ul3

ul4 <- quantile(Final_Dataset$`No. of other Loyalty Cards`,0.975)
ll4 <- quantile(Final_Dataset$`No. of other Loyalty Cards`,0.025)
Z[Z$`No. of other Loyalty Cards`<ll4,]$`No. of other Loyalty Cards` <- ll4
Z[Z$`No. of other Loyalty Cards`>ul4,]$`No. of other Loyalty Cards` <- ul4

ul5 <- quantile(Final_Dataset$`Shopping Amount at Airport`,0.975)
ll5 <- quantile(Final_Dataset$`Shopping Amount at Airport`,0.025)
Z[Z$`Shopping Amount at Airport`<ll5,]$`Shopping Amount at Airport` <- ll5
Z[Z$`Shopping Amount at Airport`>ul5,]$`Shopping Amount at Airport` <- ul5

ul6 <- quantile(Final_Dataset$`Eating and Drinking at Airport`,0.975)
ll6 <- quantile(Final_Dataset$`Eating and Drinking at Airport`,0.025)
```

```
Z[Z$`Eating and Drinking at Airport`<ll6,]$`Eating and Drinking at Airport` <- ll6
Z[Z$`Eating and Drinking at Airport`>ul6,]$`Eating and Drinking at Airport` <- ul6

ul7 <- quantile(Final_Dataset$`Departure Delay in Minutes`,0.975)
ll7 <- quantile(Final_Dataset$`Departure Delay in Minutes`,0.025)
Z[Z$`Departure Delay in Minutes`<ll7,]$`Departure Delay in Minutes` <- ll7
Z[Z$`Departure Delay in Minutes`>ul7,]$`Departure Delay in Minutes` <- ul7

ul8 <- quantile(Final_Dataset$`Arrival Delay in Minutes`,0.975)
ll8 <- quantile(Final_Dataset$`Arrival Delay in Minutes`,0.025)
Z[Z$`Arrival Delay in Minutes`<ll8,]$`Arrival Delay in Minutes` <- ll8
Z[Z$`Arrival Delay in Minutes`>ul8,]$`Arrival Delay in Minutes` <- ul8

ul9 <- quantile(Final_Dataset$`Flight time in minutes`,0.975)
ll9 <- quantile(Final_Dataset$`Flight time in minutes`,0.025)
Z[Z$`Flight time in minutes`<ll9,]$`Flight time in minutes` <- ll9
Z[Z$`Flight time in minutes`>ul9,]$`Flight time in minutes` <- ul9

ul10 <- quantile(Final_Dataset$`Flight Distance`,0.975)
ll10 <- quantile(Final_Dataset$`Flight Distance`,0.025)
Z[Z$`Flight Distance`<ll10,]$`Flight Distance` <- ll10
Z[Z$`Flight Distance`>ul10,]$`Flight Distance` <- ul10

ul11 <- quantile(Final_Dataset$Number_of_years,0.975)
ll11 <- quantile(Final_Dataset$Number_of_years,0.025)
Z[Z$Number_of_years<ll11,]$Number_of_years <- ll11
Z[Z$Number_of_years>ul11,]$Number_of_years <- ul11

ul12 <- quantile(Final_Dataset$`% of Flight with other Airlines`,0.975)
ll12 <- quantile(Final_Dataset$`% of Flight with other Airlines`,0.025)
Z[Z$`% of Flight with other Airlines`<ll12,]$`% of Flight with other Airlines` <- ll12
Z[Z$`% of Flight with other Airlines`>ul12,]$`% of Flight with other Airlines` <- ul12

#### correlation chart####
str(s_not_cancelled)
install.packages("corrplot")
library(corrplot)
newdfs <- select_if(Z, is.numeric)
newdfs
newdfs$`Day of Month` <- NULL
str(newdfs)
col1 <- colorRampPalette(c("#002627","#084447","white","#0762a2","#29516d"))
res = cor(newdfs)
corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 90, col=col1(200))

####### linear regression model  ##########
# We remove % of flight with other airlines, No. of other loyalty cards, Flight distance, departure delay in minutes due to multicollinearity issues as they are highly correlated with other explanatory variables

Z$`No. of other Loyalty Cards` <- NULL
```

```
Z$`Flight Distance` <- NULL
Z$`Departure Delay in Minutes` <- NULL
attach(Z)
jitter(Z$Age)
jitter(Z$`Price Sensitivity`)
jitter(Z$`Shopping Amount at Airport`)
jitter(Z$`Eating and Drinking at Airport`)
jitter(Z$`No of Flights p.a.`)
jitter(Z$`Flight time in minutes`)
jitter(Z$Number_of_years)
jitter(Z$`Day of Month`)
jitter(Z$`Arrival Delay in Minutes`)

l1 = lm(formula = Z$Satisfaction ~ Z$Age+Z$`Price Sensitivity`+Z$`Shopping Amount at
Airport`+Z$`Eating and Drinking at Airport`+Z$`Day of Month`+Z$`Arrival Delay in Minutes`+Z$`Flight
time in minutes`+Z$`No of Flights p.a.`+Z$`Airline Status`+Z$Gender+Z$`Type of
Travel`+Z$Class+Z$`Airline Name`+Z$Number_of_years+Z$`% of Flight with other Airlines, data = Z)
summary(l1)

###### Optimizing the model by removing bad predictors ###############
# Removed eating and drinking at airport
# Removed flight time in minutes
# Removed day of month
# Removed % of flight with other airlines

l2 = lm(formula = Z$Satisfaction ~ Z$Age+Z$`Price Sensitivity`+Z$`Shopping Amount at
Airport`+Z$`Arrival Delay in Minutes`+Z$`No of Flights p.a.`+Z$`Airline Status`+Z$Gender+Z$`Type of
Travel`+Z$Class+Z$`Airline Name`+Z$Number_of_years, data = Z)
summary(l2)

# Removed airline names. However Adjusted R-squared goes down. So, keeping Airline names
l3 = lm(formula = Z$Satisfaction ~ Z$Age+Z$`Price Sensitivity`+Z$`Shopping Amount at
Airport`+Z$`Arrival Delay in Minutes`+Z$`No of Flights p.a.`+Z$`Airline Status`+Z$Gender+Z$`Type of
Travel`+Z$Class+Z$Number_of_years, data = Z)
summary(l3)

####Association rules####
C <- s_not_cancelled1
C$`Departure Delay in Minutes` <- NULL
C$`No. of other Loyalty Cards`<- NULL
C$`Flight Distance` <- NULL
C$`Day of Month` <- NULL
C$`Eating and Drinking at Airport` <- NULL
C$`Airline Name` <- NULL
C$`Flight cancelled` <- NULL
C$`Arrival Delay greater 5 Mins` <- NULL

C$Satisfaction[C$Satisfaction >= 4] <- 'Satisfaction'
C$Satisfaction[C$Satisfaction < 4 & C$Satisfaction >=3] <- 'Average'
C$Satisfaction[C$Satisfaction < 3] <- 'Dissatisfaction'
```

```
C$Satisfaction = as.factor(C$Satisfaction)

C$Age[C$Age >= 55] = 'Elder'
C$Age[C$Age > 30 & C$Age<=55 ] = 'MiddleAgedpeople'
C$Age[C$Age<= 30] = 'Youngpeople'

C$`Arrival Delay in Minutes`[C$`Arrival Delay in Minutes`>5]  ='delay'
C$`Arrival Delay in Minutes`[C$`Arrival Delay in Minutes`<=5] ='notdelay'

category <- function(vec){
  q <- quantile(vec, c(0.33, 0.67))
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}
#install.packages("arules")
#install.packages("arulesViz")
#install.packages("grid")
#install.packages("digest")
library(arules)
library(arulesViz)
library(grid)

# Coerce the Survey data frame into a sparse transactions matrix :
C$Satisfaction <- as.factor(C$Satisfaction)
C$Age <- as.factor(C$Age)
C$`Price Sensitivity` <- as.factor(category(C$`Price Sensitivity`))
C$`No of Flights p.a.` <- as.factor(category(C$`No of Flights p.a.`))
C$`% of Flight with other Airlines` <- as.factor(category(C$`% of Flight with other Airlines`))
C$`Shopping Amount at Airport` <- as.factor(category(C$`Shopping Amount at Airport`))
C$`Arrival Delay in Minutes` <- as.factor(C$`Arrival Delay in Minutes`)
C$`Flight time in minutes`<- as.factor(category(C$`Flight time in minutes`))
C$`Airline Status`<- as.factor(C$`Airline Status`)
C$Gender <- as.factor(C$Gender)
C$`Type of Travel` <- as.factor(C$`Type of Travel`)
C$Class <- as.factor(C$Class)
C$Number_of_years <- as.factor(category(C$Number_of_years))
View(C)

CX <- as(C,"transactions")

itemFrequencyPlot(CX,support=0.05,cex.names=0.5)

#predict happy passengers (as defined by their overall satisfaction >=4).
ruleset <- apriori(CX,parameter=list(support=0.1, confidence=0.3),appearance =
list(rhs="Satisfaction=Satisfaction",default="lhs"))
plot(ruleset,jitter=0)
goodrules <- ruleset[quality(ruleset)$lift > 1.5]
inspect(goodrules)
```

```
top.lift <- sort(goodrules, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift, 10))

#predict Average passengers (as defined by their overall satisfaction >=4).
ruleset1 <- apriori(CX,parameter=list(support=0.1, confidence=0.3),appearance =
list(rhs="Satisfaction=Average",default="lhs"))
plot(ruleset1,jitter=0)
goodrules1 <- ruleset1[quality(ruleset1)$lift > 1]
inspect(goodrules1)
top.lift1 <- sort(goodrules1, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift1, 10))

#predict Dissatisfied passengers (as defined by their overall satisfaction >=4).
ruleset2 <- apriori(CX,parameter=list(support=0.1, confidence=0.3),appearance =
list(rhs="Satisfaction=Dissatisfaction",default="lhs"))
plot(ruleset2,jitter=0)
goodrules2 <- ruleset2[quality(ruleset2)$lift > 1]
inspect(goodrules2)
top.lift2 <- sort(goodrules2, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(top.lift2, 10))

################SVM#################
###### not use z score  ########
library("dplyr")
S <- s_not_cancelled
dim(S)
S$Satisfaction[S$Satisfaction >= 4] <- 'Satisfaction'
S$Satisfaction[S$Satisfaction < 4] <- 'Dissatisfaction'
randomindex = sample(1:nrow(S))
cutpoint = floor(2*nrow(S)/3)
traindata = S[randomindex[1:cutpoint],]
testdata = S[randomindex[(cutpoint+1):nrow(S)],]
dim(traindata)
str(S)
#install.packages("kernlab")
library(kernlab)
svmOutput <- ksvm(Satisfaction ~Class+`Type of Travel`+`Airline Status`+Age+`Arrival Delay in
Minutes`+`Price Sensitivity`+`Shopping Amount at Airport`+Gender,data=traindata, kernel= "rbfdot", kpar
= "automatic", C = 50, cross = 3, prob.model = TRUE)
svmOutput
svmPred <- predict(svmOutput, testdata, type = "votes")
dim(svmPred)
dim(testdata)
compTable <- data.frame(data.frame(testdata$Satisfaction=="happy",svmPred[1,]))
table(compTable)
res <- table(compTable)
#Calculate an error rate
errorRate <- (res[1,1]+res[2,2])/(sum(res))
errorRate
```

```
###################ggplot
#install.packages("ggplot2")
library(ggplot2)
as <- s_not_cancelled
######3 level for satisfaction
as$Satisfaction[as$Satisfaction >= 4] <- 'Satisfaction'
as$Satisfaction[as$Satisfaction < 4 & as$Satisfaction >=3 ] <- 'Average'
as$Satisfaction[as$Satisfaction < 3] <- 'Dissatisfaction'
as$Satisfaction = as.factor(as$Satisfaction)

########GGplot Tasks

# Airline Status
g <- ggplot(as, aes(as$`Airline Status`))
g + geom_bar(aes(fill=as$Satisfaction), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="satisfaction analysis",
       subtitle="Airline Status")

# Type of Travel
g <- ggplot(as, aes(as$`Type of Travel`))
g + geom_bar(aes(fill=as$Satisfaction), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="satisfaction analysis",
       subtitle="Type of Travel")

# Gender
g <- ggplot(as, aes(Gender))
g + geom_bar(aes(fill=Satisfaction), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="satisfaction analysis",
       subtitle="Gender")

# Airline Status
g <- ggplot(as, aes(`Airline Status`))
g + geom_bar(aes(fill=Satisfaction), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="satisfaction analysis",
       subtitle="Airline Status")

# Price Sensitivity
g <- ggplot(as, aes(`Price Sensitivity`))
g + geom_bar(aes(fill=Satisfaction), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="satisfaction analysis",
       subtitle="Price Sensitivity")

# Airline Name
#install.packages('ggthemes')
library(ggthemes)
```

```
options(scipen = 999)  # turns of scientific notations like 1e+40
# X Axis Breaks and Labels
brks <- seq(-15000000, 15000000, 5000000)
lbls = paste0(as.character(c(seq(15, 0, -5), seq(5, 15, 5))), "m")
# Plot
freqtable <- table(s_not_cancelled$`Airline Name`)
freq <-as.data.frame(sort(freqtable))
s_not_cancelled$`Airline Name` <- factor(s_not_cancelled$`Airline Name`, levels = freq$Var1)
satisfaction <- as$Satisfaction
ggplot(s_not_cancelled, aes(x = `Airline Name`, y = Satisfaction, fill = satisfaction)) +   # Fill column
  geom_bar(stat = "identity", width = .6) +   # draw the bars
  scale_y_continuous(breaks = brks,   # Breaks
                     labels = lbls) + # Labels
  coord_flip() +  # Flip axes
  labs(title="Airline Company") +
  theme_tufte() +  # Tufte theme from ggfortify
  theme(plot.title = element_text(hjust = .5),
        axis.ticks = element_blank()) +   # Centre plot title
  scale_fill_brewer(palette = "Dark2")  # Color palette


####Age and Satisfaction
ageS <- ggplot(as, aes(Age)) + scale_fill_manual(values=c("Satisfaction"="#DFE1E3",
"Dissatisfaction"="#798E9E", "Average"="#02353F"))
ageS + geom_histogram(aes(fill=Satisfaction), binwidth = 4, col="black", size=1) +ggtitle("Satisfaction vs
Age")
ageS + geom_histogram(aes(fill=Satisfaction), bins=5, col="black", size=.1) +ggtitle("Satisfaction vs Age")

#####arrival delay vs Satisfaction###############
adS <- ggplot(as, aes(`Arrival Delay in Minutes`)) + scale_fill_brewer()
adS + geom_histogram(aes(fill=Satisfaction), binwidth = 20, col="black", size=1) +ggtitle("Satisfaction vs
Arrival delay in Minutes")
adS + geom_histogram(aes(fill=Satisfaction), bins=10, col="black", size=.1) +ggtitle("Satisfaction vs
Arrival delay in Minutes")

Ardy96 <- quantile(as$`Arrival Delay in Minutes`,0.96)
Sati_Ardy <- ggplot(as[as$`Arrival Delay in Minutes`<Ardy96,],aes(x=`Arrival Delay in Minutes`))
Sati_Ardy <- Sati_Ardy + geom_histogram(aes(fill=Satisfaction),position = "dodge", bins = 30)
Sati_Ardy <- Sati_Ardy + ggtitle("Satisfaction versus Arrival Delay")
Sati_Ardy


############class and satisfaction
as$Class <- factor(as$Class, levels = c("Eco","Eco Plus","Business"))
cs <- ggplot(as, aes(Class))
cs + geom_bar(aes(fill=Satisfaction), col="black", width = 0.6) + theme(axis.text.x =
element_text(angle=0, vjust=0.6)) + scale_fill_manual(values=c("Average"="#DFE1E3",
"Satisfaction"="#798E9E", "Dissatisfaction"="#02353F"))

##############Day of month######
```

```
as1 = s_not_cancelled
as1$Satisfaction = as.factor(as1$Satisfaction)
dss <- ggplot(as1, aes(`Day of Month`)) +  scale_fill_brewer(palette = "Spectral")
dss + geom_histogram(aes(fill=Satisfaction), binwidth = 1, col="black", size=1) +ggtitle("Satisfaction vs Day of Month")

#############shopping
ss <- ggplot(as, aes(`Shopping Amount at Airport`)) +  scale_fill_brewer(palette = "Spectral")
ss + geom_histogram(aes(fill=Satisfaction), binwidth = 4, col="black", size=1) +ggtitle("Satisfaction vs Day of Shopping Amount")

########Maps#########
Satisfaction_Survey$`Destination State` <- tolower(Satisfaction_Survey$`Destination State`)
summary(Satisfaction_Survey$`Destination State`)
str(x17)

map1.us <- ggplot(Satisfaction_Survey, aes(map_id = `Destination State`))
map1.us <- map1.us + geom_map(map = us, aes(fill=Satisfaction), color="black")
map1.us <- map1.us + expand_limits(x = us$long, y = us$lat)
map1.us <- map1.us + coord_map() + labs(x="Latitude", y="Longitude", title="US Map by Satisfaction Level") + theme(plot.title = element_text(hjust = 0.5))
map1.us

Satisfaction_Survey1 <- filter(Satisfaction_Survey,`Airline Name`=="Southeast Airlines Co.",`Flight cancelled`=="No")
Satisfaction_Survey1$`Destination State` <- tolower(Satisfaction_Survey1$`Destination State`)
map2.us <- ggplot(Satisfaction_Survey1, aes(map_id = `Destination State`))
map2.us <- map2.us + geom_map(map = us, aes(fill=Satisfaction), color="black")
map2.us <- map2.us + expand_limits(x = us$long, y = us$lat)
map2.us <- map2.us + coord_map() + labs(x="Latitude", y="Longitude", title="US Map by Satisfaction Level - Southeast Airlines Co.") + theme(plot.title = element_text(hjust = 0.5))
map2.us

##########Decision Tree######
dt = s_not_cancelled
dt$Satisfaction[dt$Satisfaction >= 4] <- 'Satisfaction'
dt$Satisfaction[dt$Satisfaction < 4 & dt$Satisfaction >=3 ] <- 'Average'
dt$Satisfaction[dt$Satisfaction <= 2] <- 'Dissatisfaction'
set.seed(2000)
install.packages("rpart")
library(rpart)
randomindex1 = sample(1:nrow(dt), size = 10000)
cutpoint1 = floor(2*10000/3)
traindata1 = dt[randomindex1[1:cutpoint1],]
testdata1 = dt[cutpoint1+1:10000,]
dtree<-rpart(Satisfaction ~ Class +`Type of Travel`+`Airline Status`+ Age+ `Arrival Delay in Minutes`
+`Price Sensitivity`+`Shopping Amount at Airport`+Gender
,data=traindata1, method="class", parms=list(split="information"))
#Class+`Type of Travel`+`Airline Status`+Age+`Arrival Delay in Minutes`+`Price Sensitivity`+`Shopping Amount at Airport`+Gender
```

```
printcp(dtree)
tree<-prune(dtree,cp=0.0125)
opar<-par(no.readonly = T)
par(mfrow=c(1,2))
install.packages("rpart.plot")
library(rpart.plot)
rpart.plot(dtree,branch=1,type=4, fallen.leaves=T,cex=0.7, sub="Decision Tree for Airline Satisfaction")
rpart.plot(tree,branch=1, type=4,fallen.leaves=T,cex=0.7, sub="Decision Tree for Airline Satisfaction")
par(opar)
dev.off()
predtree<-predict(tree,newdata=testdata1,type="class")
table(testdata1$Satisfaction,predtree,dnn=c("real one","predict one"))
```