Airlines Satisfaction Analysis Report

IST 687 M002 Group 5
Qi Wang
Jiaming Guo
Yun Xiao
Mingyuan (Emma) Zhu
Kshitij Sankesara

Contents

Introduction

(Scope,context,background)

In order to help our client Southeast Airline to improve their quality, we designed this project to research factors that will highly influence their clients' satisfactions. Southeast collected raw data by doing survey via internet. This survey covers clients among 14 airline companies which are assumed to have the same level as Southeast Airplane. We are using data of Southeast airlines as our client data. This dataset contains total 28 variables. The most important - Satisfaction - will be divided into 5 levels from 1 to 5 ascendingly. All other variables are all research objects including age, gender, price sensitive and so on. The types of variables are numerical and factors. For some analysis, we divided Satisfaction scores to "Y" (Yes as satisfied) and "N" (No as not satisfied): Satisfaction Scores as 4 and 5 are "Y"s, and Satisfaction Scores as 1,2,3 are "N"s.

In this project, Association Rules and MCA will be performed after necessary data cleaning. Based on our Association Rules and MCA results, we will build linear regression model and ordinal regression model. Finally, we will use SVM to see the accuracy and determine whether and how to improve our modes. Results and recommended improvement methods will be discussed at the end of project based on our business questions. We hope this final optimal mode can help Southeast Airline to improve efficiently.

Business Questions Addressed

- Which factors tend to influence satisfaction, and which one is the most effective?
- Is there any relationship between these factors?
- If there is any relationship between explanatory factors and satisfaction variable, are they positively related or negatively related?
- Especially for Southeast Airline company, is there any solution could help them to stand out in achieving the high customer satisfaction?
- How does the current satisfaction variable diverse through different gender, airline status, airline class, age, price sensitivity, travel type, and flight cancellation status? Can we identify a trend or feature for people who are more likely to give higher satisfaction score as well as lower satisfaction score?
- What is the current average satisfaction rating? How does the Southeast company perform in customer’s satisfaction compared to the whole industry?
- What're the current problems that the airline industry faced in improving customer satisfaction?
- Is there any way we can promote to solve these problems to achieve higher customer satisfaction?

Data Acquisition, Cleansing, Transformation, and Munging

The Dataset for this Project consisted of 129,889 Observations and 28 Attributes. This dataset included 14 Airlines, their arrival and delay time, Origin and Destination place, Day and Date of flight, etc. It also had details about their Customer Age, Gender, Satisfaction, number of times which they travel by flying in a year, their Type of Travel, Loyalty cards which they have, their Shopping, Drinking and Eating expenses, Class by which they travel, etc. The dataset was for the first three months of 2014 i.e. from January 2014 to March 2014.

The main focus for us was to concentrate on Customer Satisfaction and to find out how customer satisfaction is affected and what are the ways of increasing customer satisfaction.

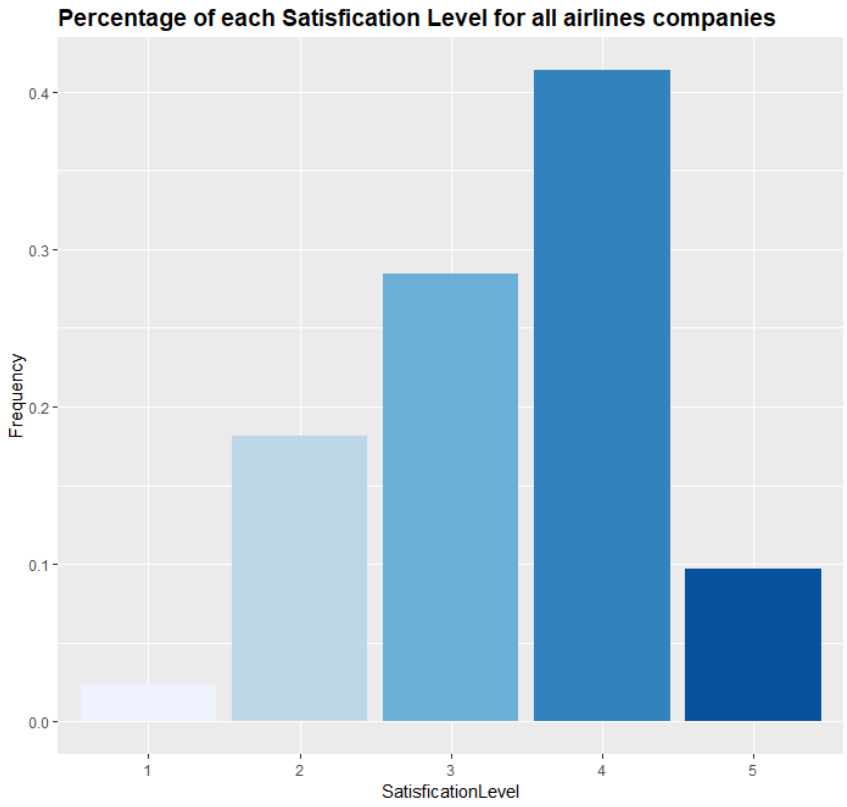
For our client Southeast Airlines, we considered only the data of Southeast Airlines, and we researched on all the factors which could improve the satisfaction of all the customers of Southeast Airlines.

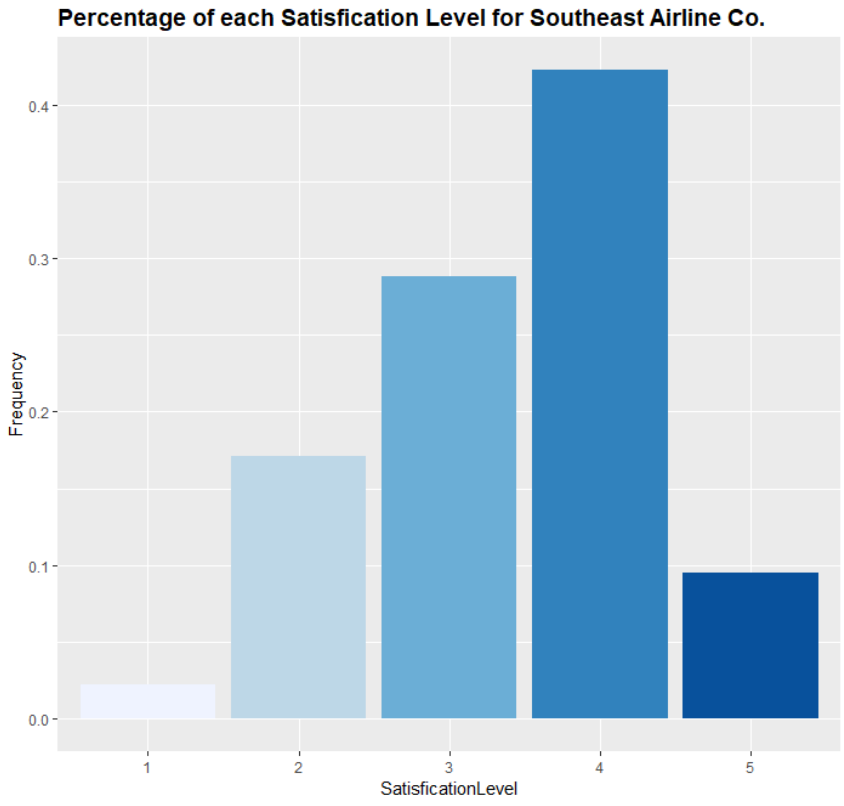
As we were going through the data for our analysis, we came across NA's in 3 columns (Flight time, Arrival and Departure time). We decided to consider 0 for all our NA's and ignore the 0 value later in our analysis as it would affect the quality of our study. We could not just omit NAs by removing the corresponding rows because our datasets will be modified. For example, the original data had 2,401 "Flight. Canceled" as "Yes"s and 127,488 as "No"s. When NAs were omitted, only 127,151 "Yes"s were kept, and all the "No"s were gone. We tried approaching other methods like replacing the mean, median values in place of NA which reduced the quality of the data and produced highly skewed results.

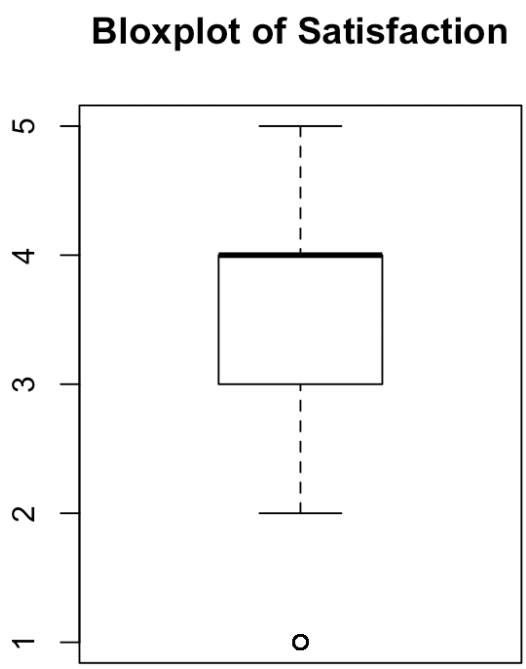
For our most important column Satisfaction, we cleared weird data for example “4.00.200”, and we decided to change the decimal values into the next whole number. For example, 2.5 to 3. As the decimal values were minimal in total (just 1 or 2 of each), we decided to replace it with the whole number. It was easier to analyze that way, and it didn't affect our data.

Descriptive Statistics & Visualization

1) Satisfaction







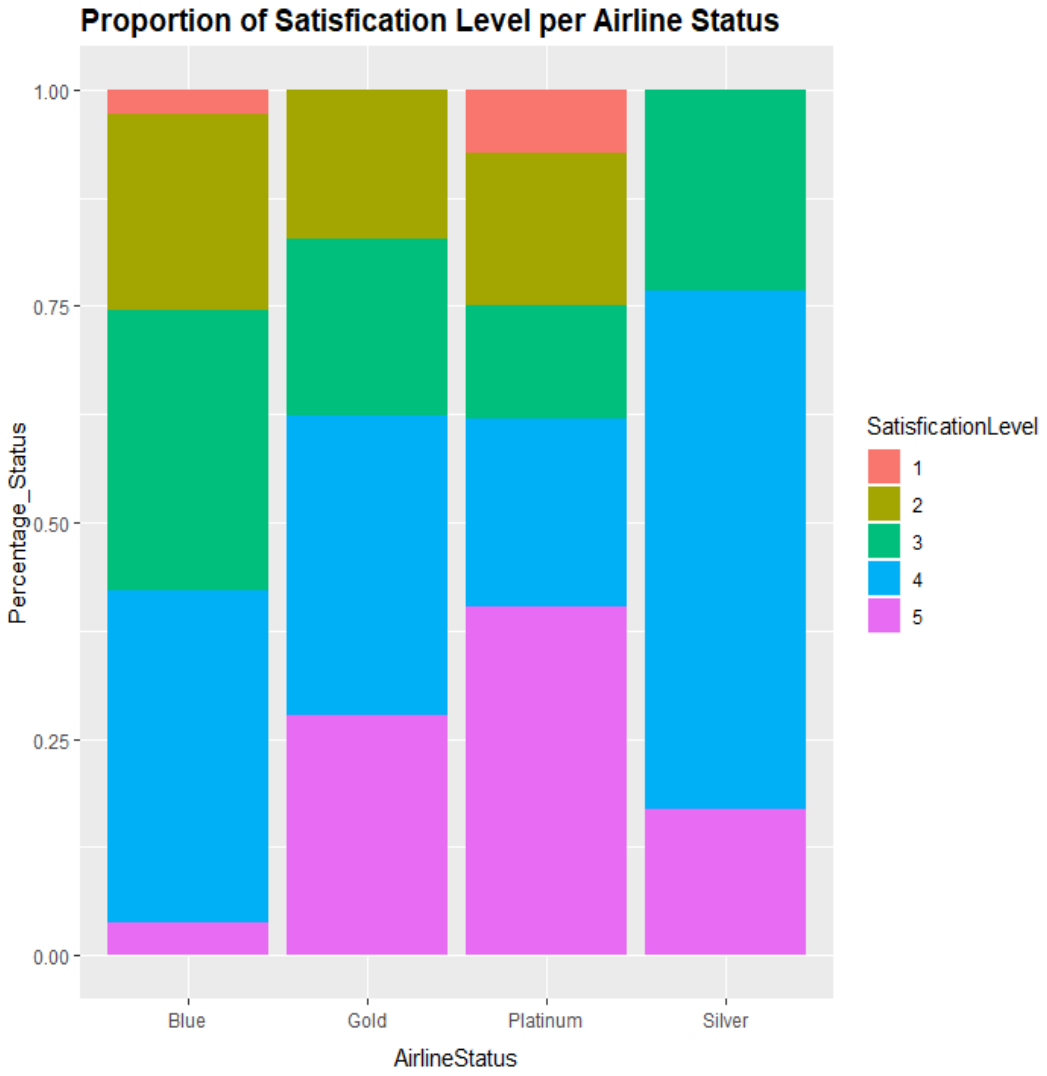
Statistics	
Mean	3.4
Variance	0.91
Std. Dev.	0.95
Minimum	1
Maximum	5
Range	4
1st Quartile	3
3rd Quartile	4
IQR	1
Median	4

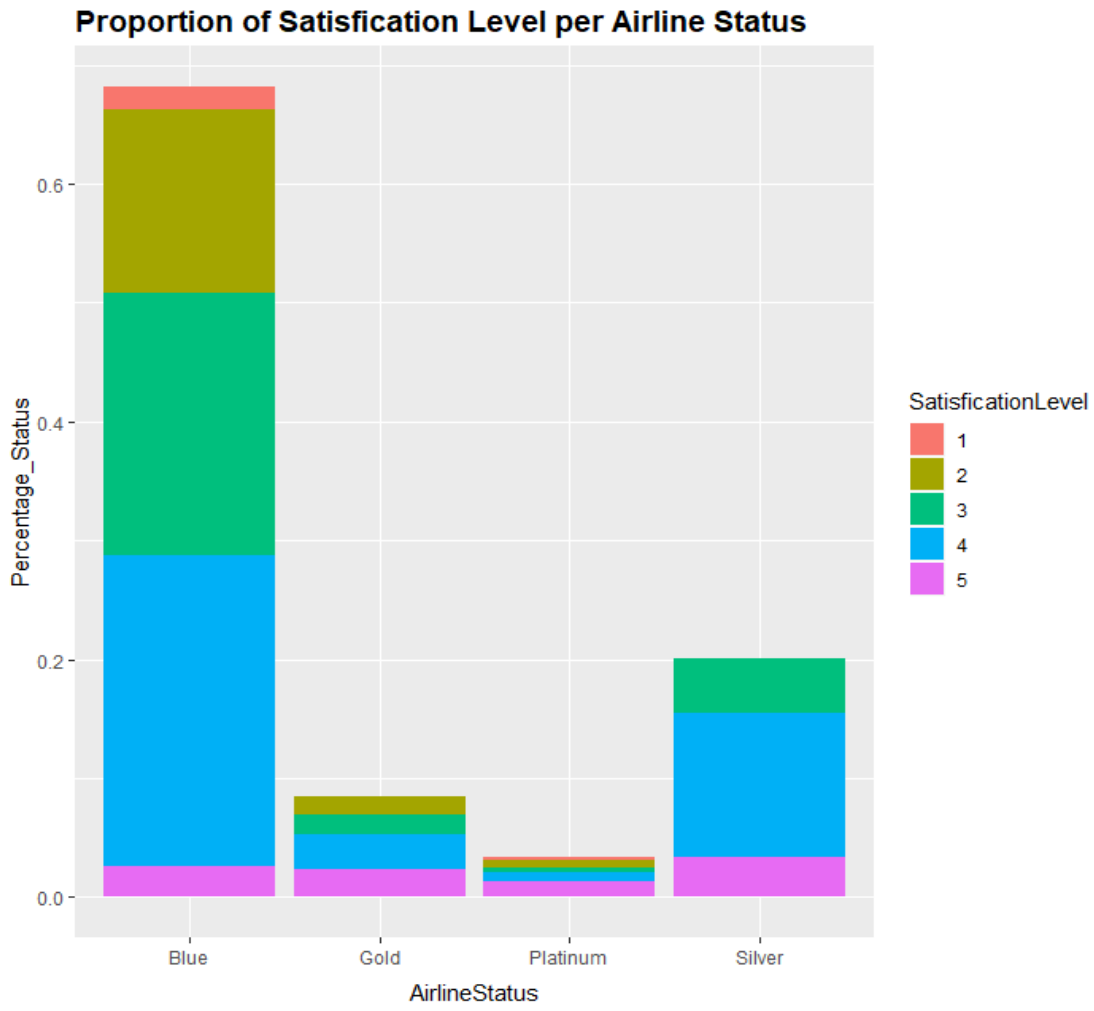
From the above comparison of Satisfaction bar chart in the whole industry and Satisfaction in Southeast Airline, it's obvious that the distribution of Satisfaction from them is almost alike.

Hence, in the whole industry (including Southeast Airline), Satisfaction in level 4 accounts for the highest proportion for all level; the second Satisfaction level is level 3. Hence, clients are more likely to give an above-average Satisfaction level.

The left two chart shown the Summary Statistics for Satisfaction in Southeast Company. We can tell that the mean is 3.4 and median is 4, and furtherly we conclude that Satisfaction above or equal to 4 is high score, and Satisfaction of 1,2&3 is low score.

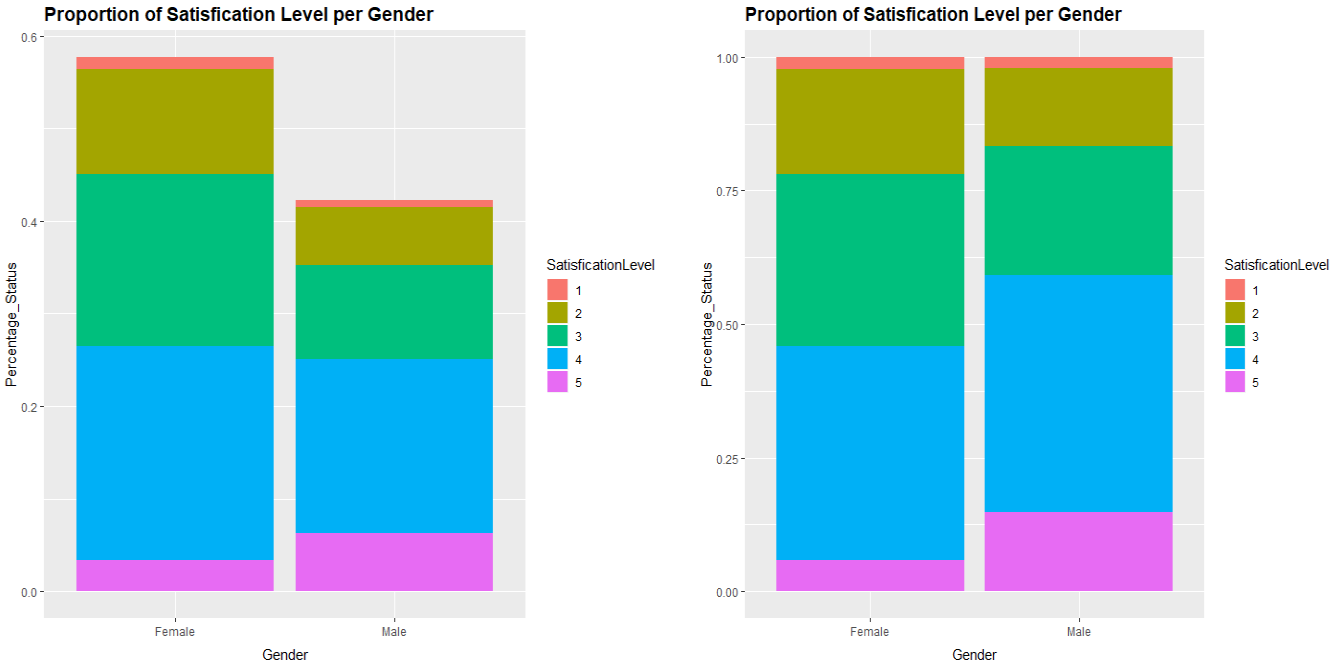
2) Satisfaction vs. Airline Status





Guests with Silver Airline status are more likely to give Satisfaction level above 3, with the highest proportion in level 4. Premium guests are the most likely group of guests to have the highest Satisfaction level. Besides, Gold guests are also very probable to give the highest Satisfaction evaluation. For Blue guests, Satisfaction level 3&4 are liable to happen.

3) Satisfaction vs. Gender

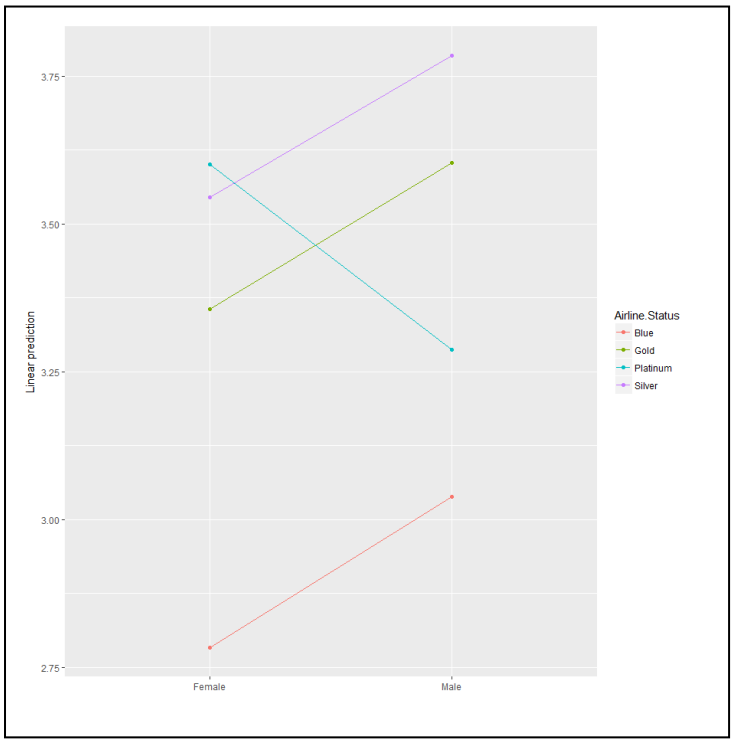


Females are more likely to give Satisfaction evaluation than males.

However, males are far more possible than females to give the highest Satisfaction Level.

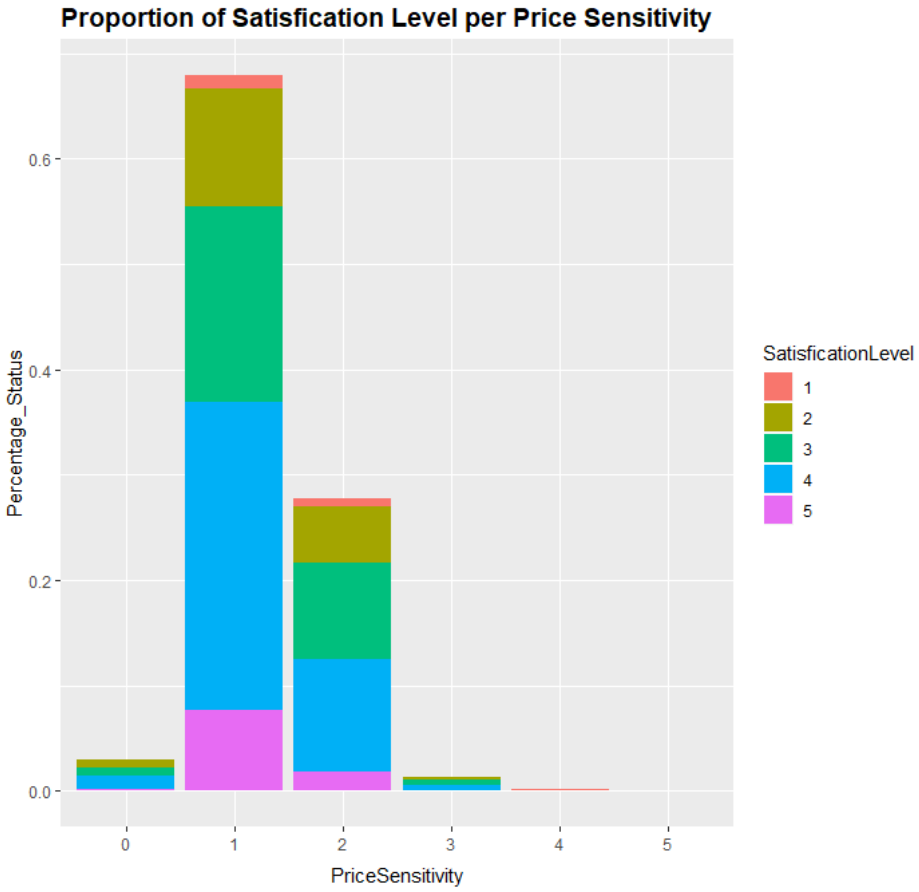
Besides that, females and males proportion in Satisfaction Levels, 3&4 are almost the same.

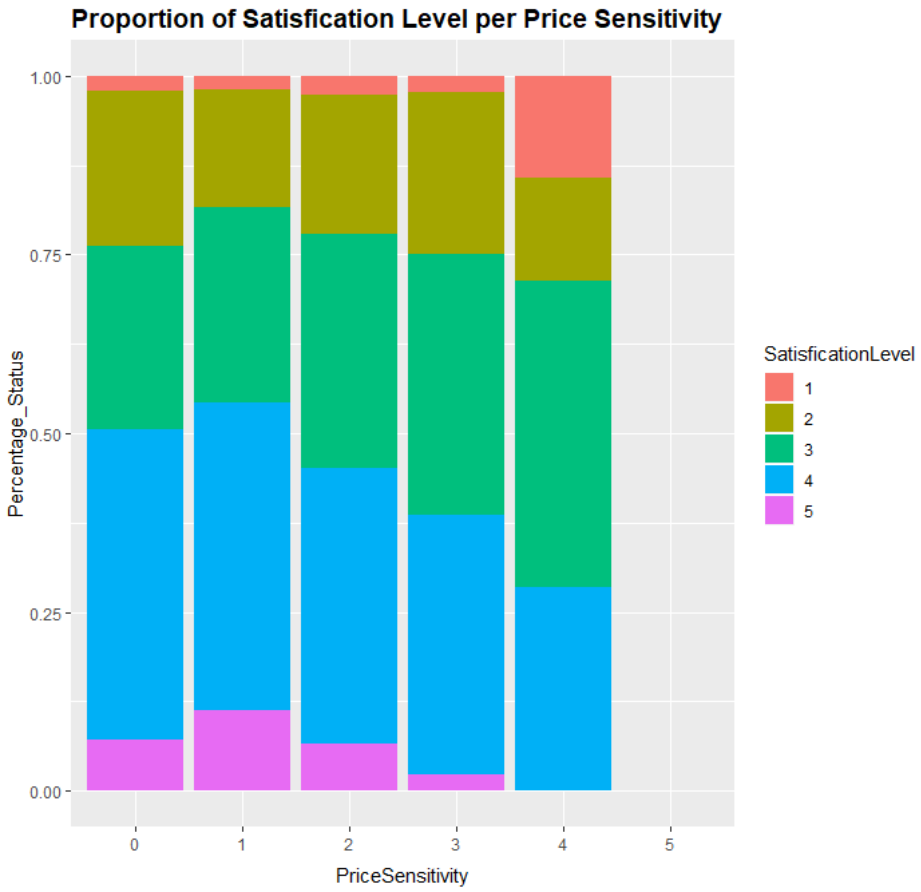
But, women are more alike to give Satisfaction Level 2 than males.



The left chart furtherly proves that males in three of total four Airline Status gave a higher Satisfaction score than females.

4) Satisfaction vs. Price Sensitivity

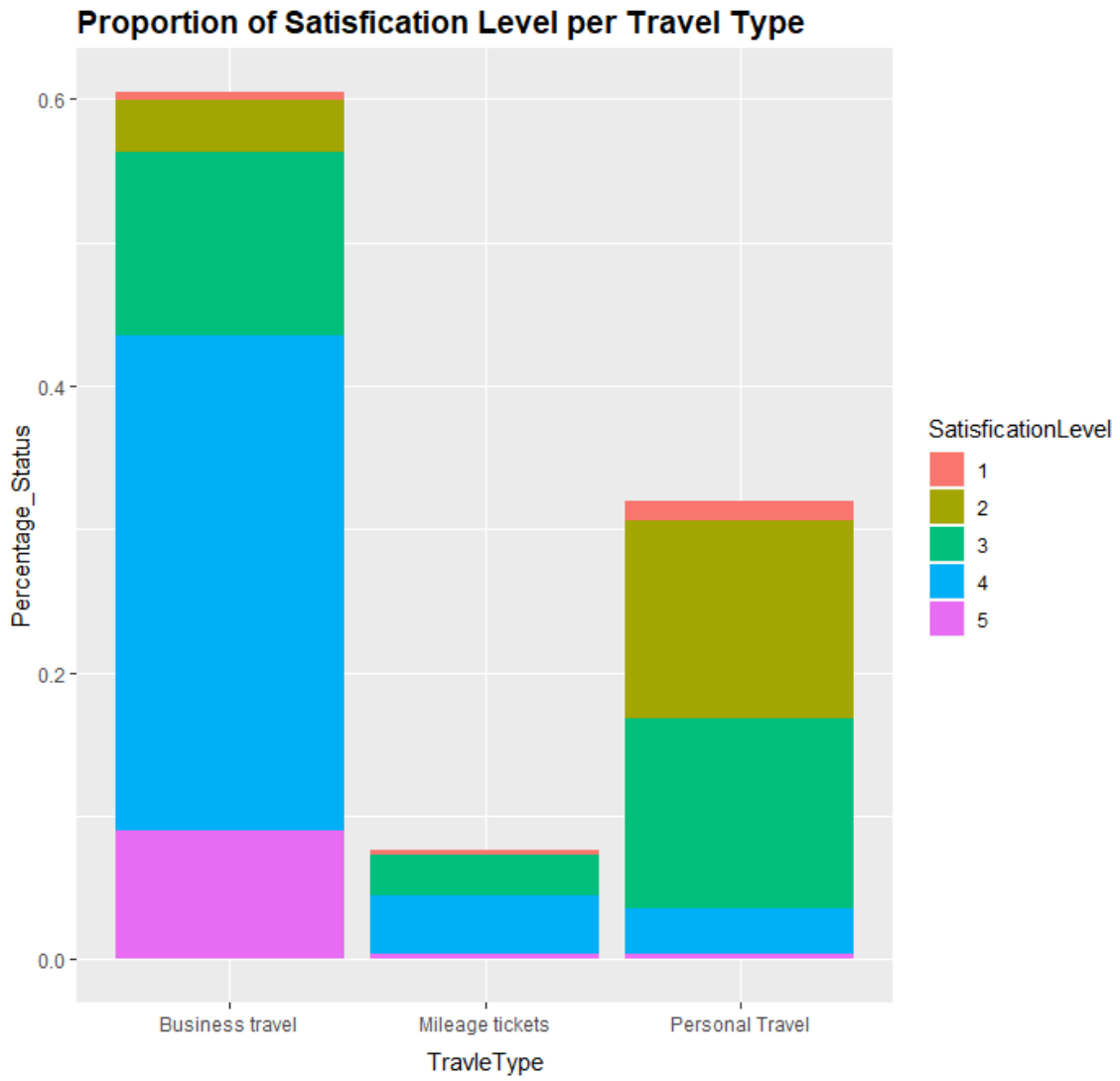


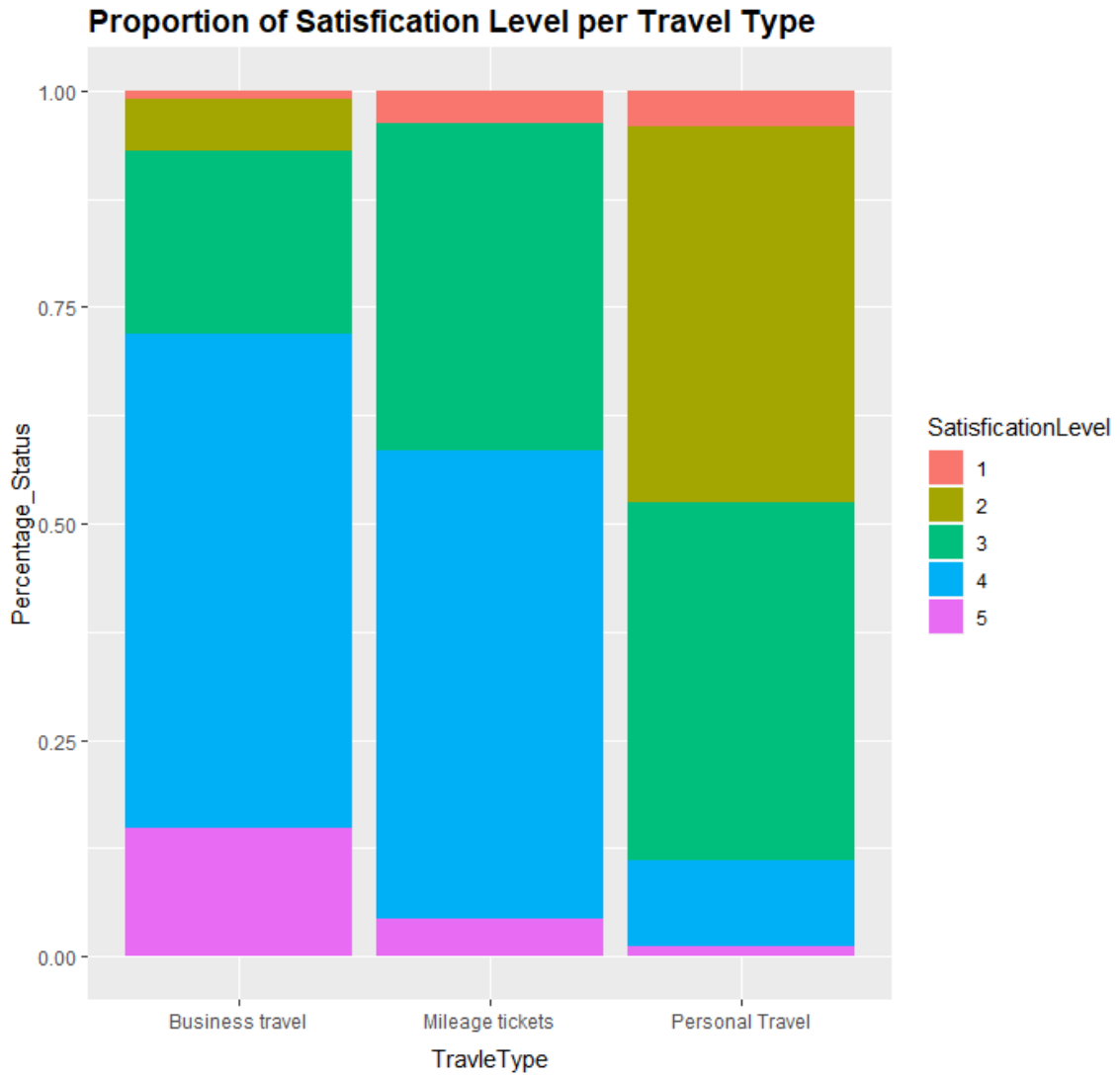


Among all guests and six different price sensitivity, they are most likely to have a price sensitivity in 1. And there is a dramatic decrease difference between the number of people in one-level price sensitivity group with the rest amount guests.

Guests with one price sensitivity are most likely to give highest satisfaction score, and guests with four price sensitivity are most possible to provide the lowest satisfaction score

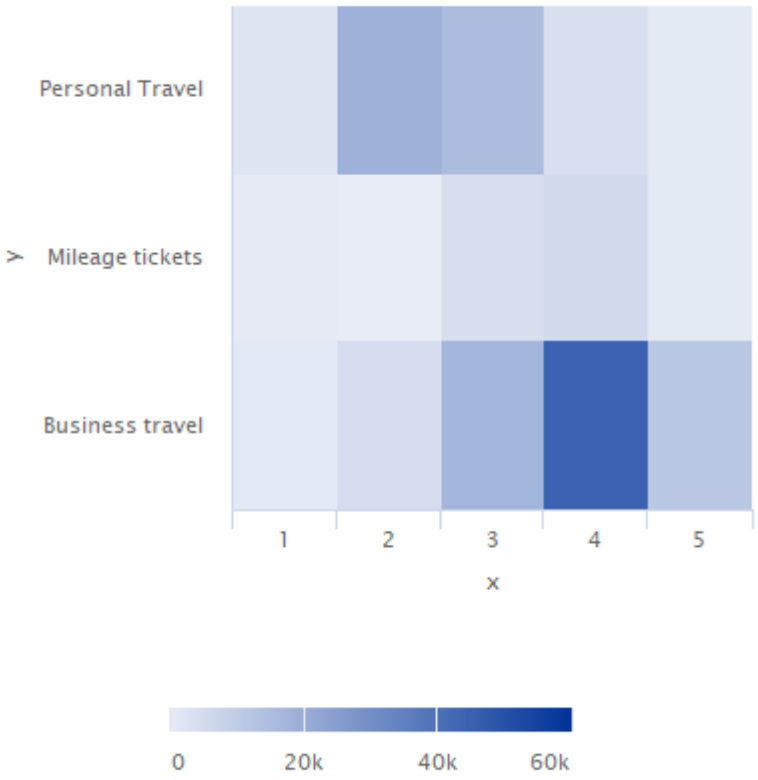
5) Satisfaction vs. Travel Type





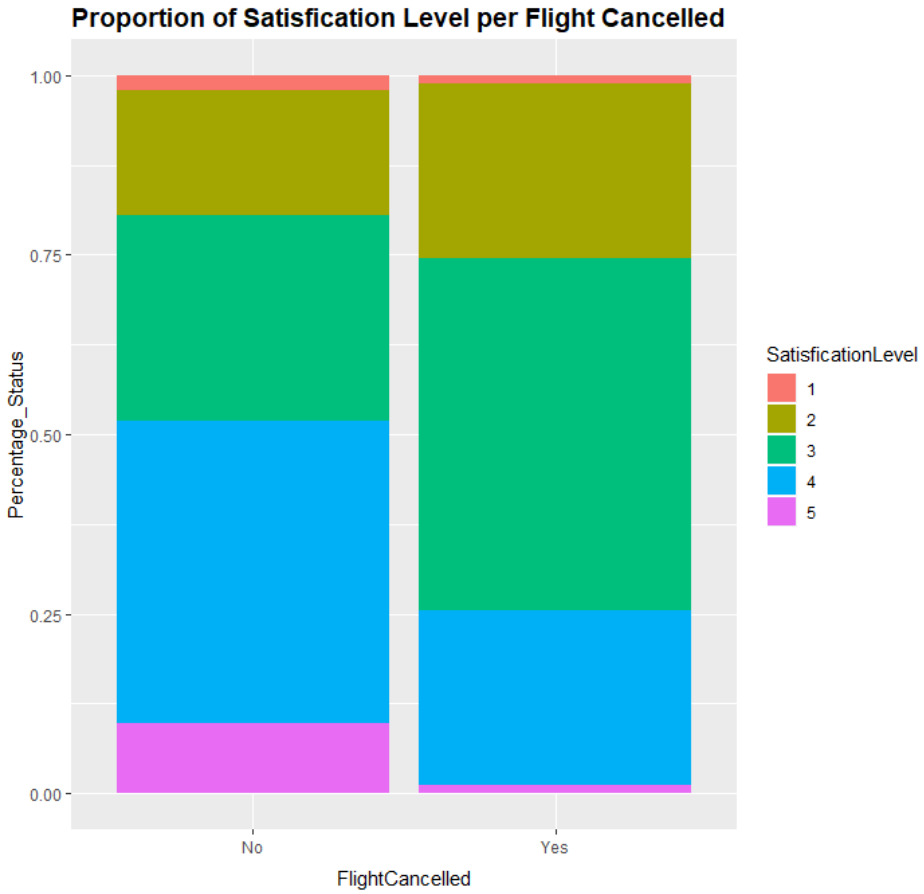
The highest travel type are Business travel, which has had a pretty high Satisfaction proportion in level 4. It also had the highest level-five satisfaction proportion. We could assume that Business travel group people are more likely to give a higher score; The second largest travel group is Personal travel, and in this group, guests are more likely to offer below average Satisfaction scores. And people in Personal travel group has the smallest proportion in giving level-5 Satisfaction. People in this group are probable to be critical and unsatisfied by airlines service. The smallest travel type is Mileage tickets. People in this group are more likely to give an approximately average score of Satisfaction.

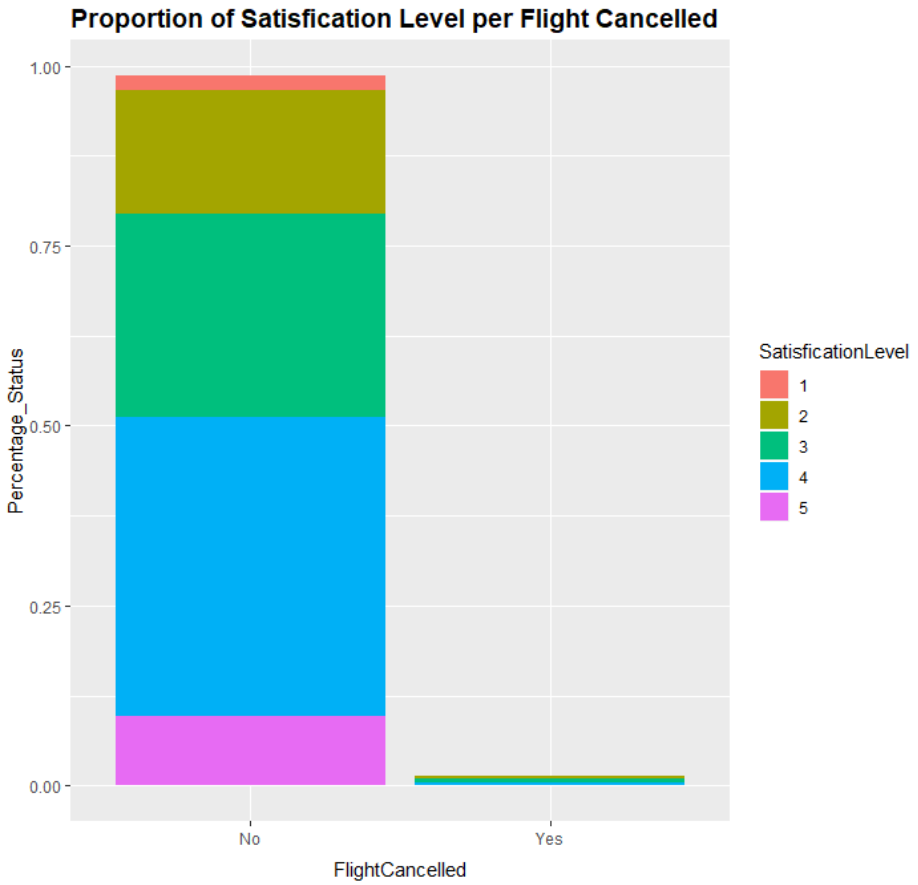
Heat Map of Satisfaction (x) vs. Type of Travel



Most of the flights are business travels, and Satisfaction score 4 occurs the most among the business travels. Most of the customers flying personal travels give Satisfaction 2 or 3. Customers fly personal travels appear to be less satisfied than customers fly business travels.

6) Satisfaction vs. Flight Cancelled

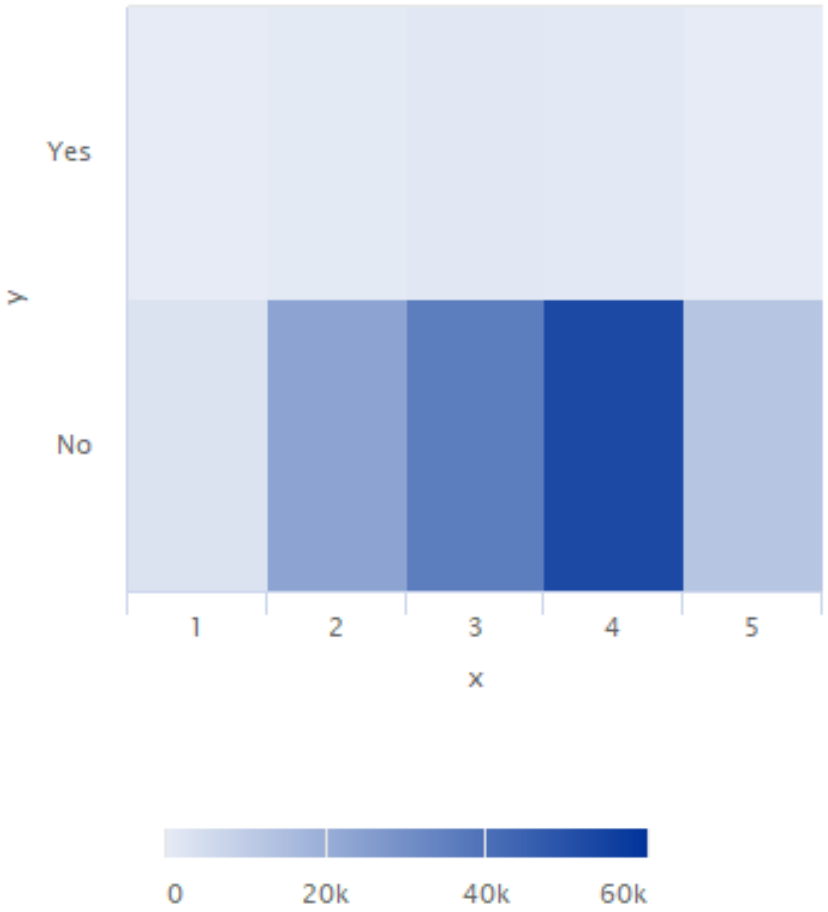




From the above pictures, there is a dramatic and exciting difference in the number of people for giving Satisfaction score between those who had a flight canceled and who hadn't. It is reasonable to imagine that people are much more likely to provide a review when they faced an issue of flight canceling. Moreover, people without flight cancellation are more likely to give higher Satisfaction score than people with flight cancellation. We believe this factor is an essential predictor for Satisfaction.

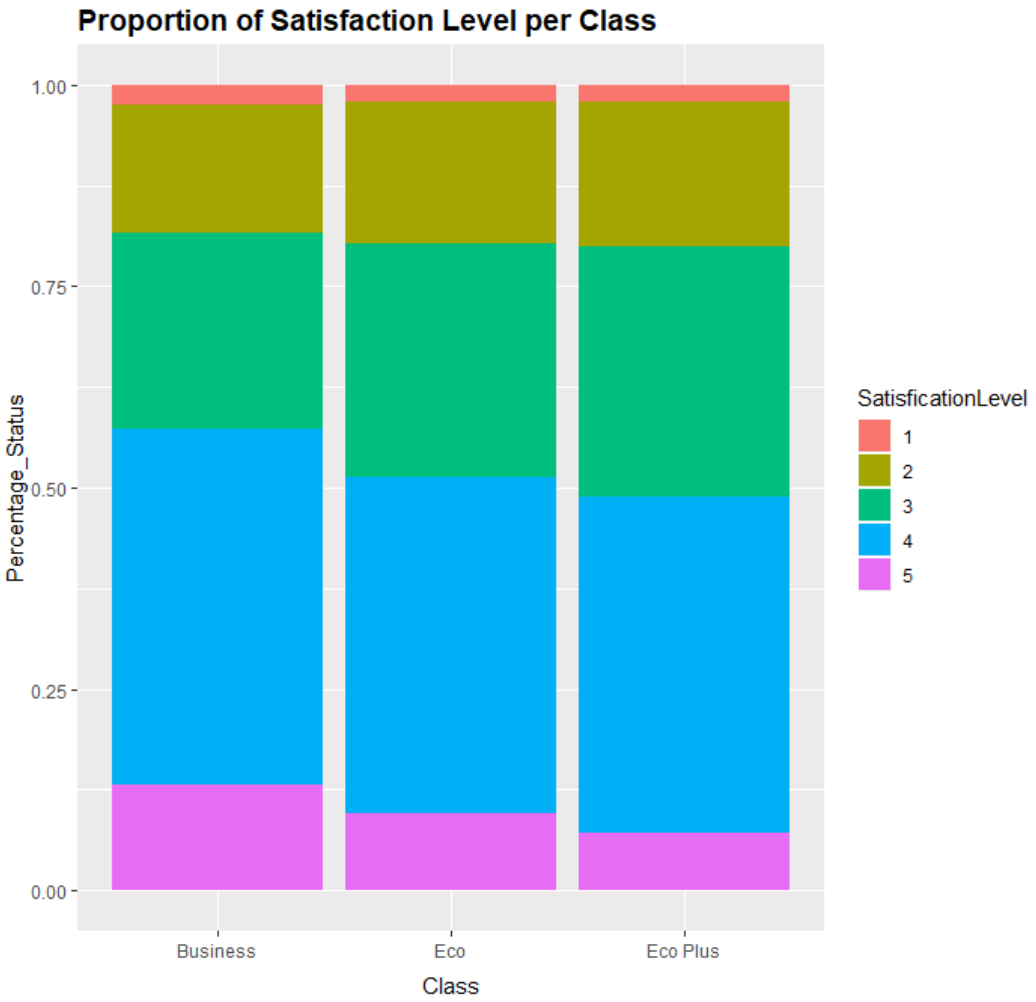
The above “Satisfaction vs Flight Cancelled” boxplot shows that Satisfaction scores as lower when there are cancelled flights than when no cancelled flights.

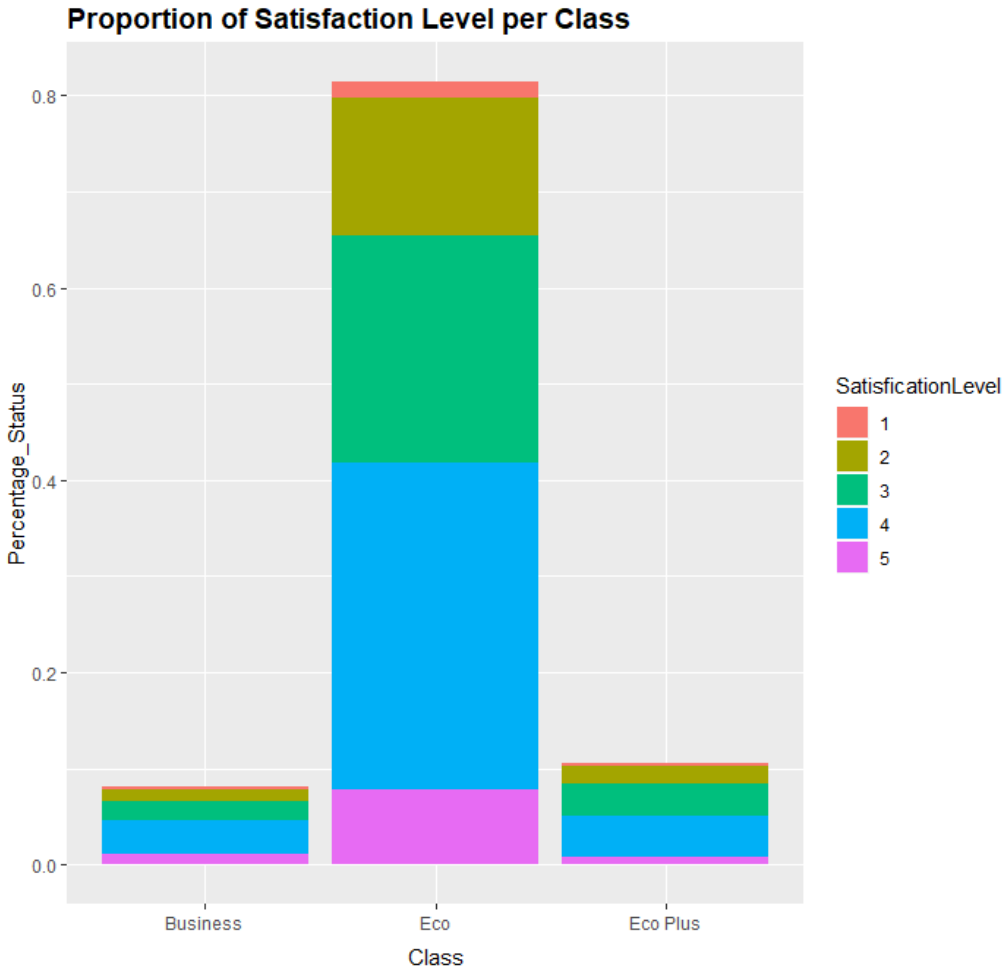
Heat Map Satisfaction (x) vs Flight Cancelled (y)



Most flights are not cancelled. Among the flights that are not cancelled, most customers give Satisfaction as 4.

7) Satisfaction vs. Class



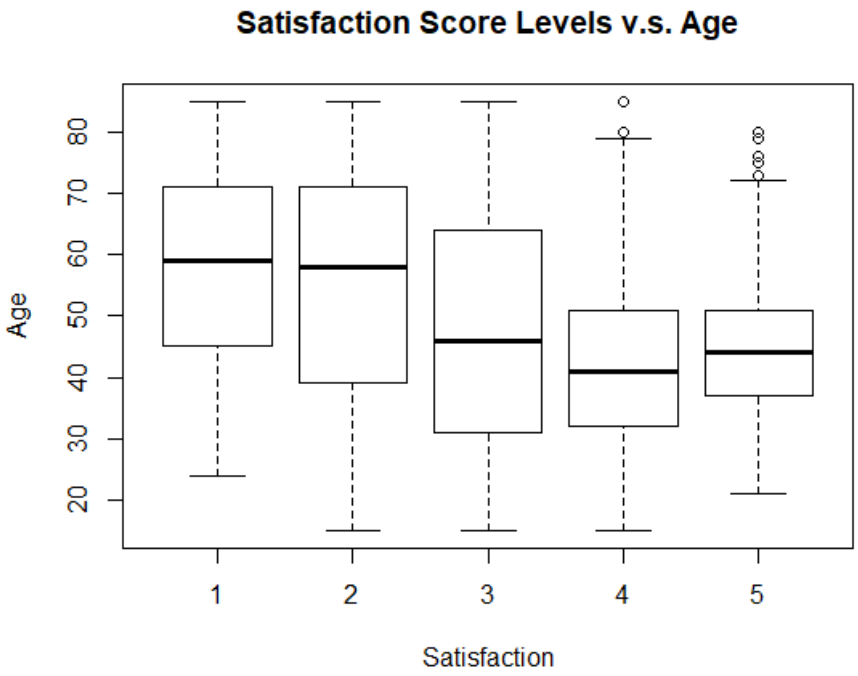


Among Business Class customers, more than half of them give satisfaction scores that are more than 3. Among Economy Class customers, about half rated more than 3. For all the Economy Plus Class customers, less than half gave more than 3.

Business class customers are more likely to give a higher score in Satisfaction. Eco class customers are less likely than Business class customers on giving a high score and more likely to give a score which is slightly surrounded the average rating. For Eco Plus class customers, they are almost the same as Eco class customers in providing the Satisfaction score, except that Eco Plus class customers are more less alike to give the 5 Satisfaction score.

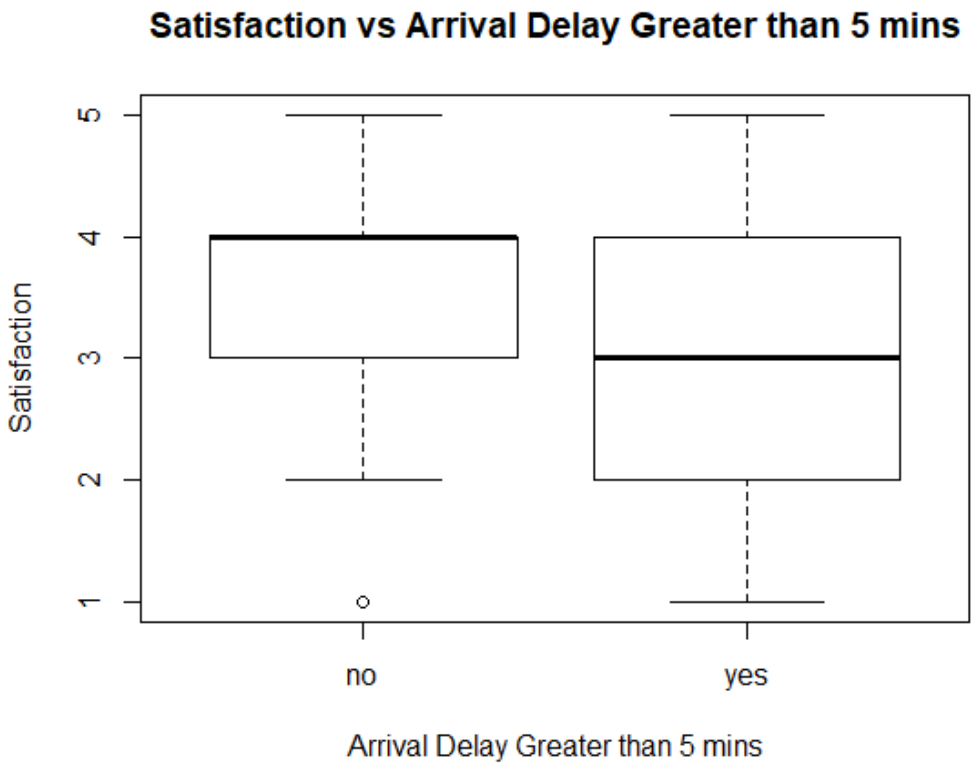
8) Satisfaction vs. Age

Age's statistics summary on different Satisfaction level					
	Satisfaction: 1	Satisfaction: 2	Satisfaction: 3	Satisfaction: 4	Satisfaction: 5
Mean	57	54	47	42	45
Variance	270	429	394	182	102
Std. Dev.	16	21	20	13	10
Median	59	58	46	41	44
Min	24	15	15	15	21
Max	85	85	85	85	80
Range	61	70	70	70	59
1st Quartile	45	39	31	32	37
3rd Quartile	71	71	64	51	51
IQR	26	32	33	19	14



Generally speaking, there is an apparent trend between the age and Satisfaction level. From the lowest Satisfaction score to the highest Satisfaction score, there is an approximately decreasing trend through Age. Except for the Age value for the Satisfaction level five, the mean value and median value of Age decrease from the lowest Satisfaction level to the lowest Satisfaction level. We assume that there is a negative relationship between Age and Satisfaction.

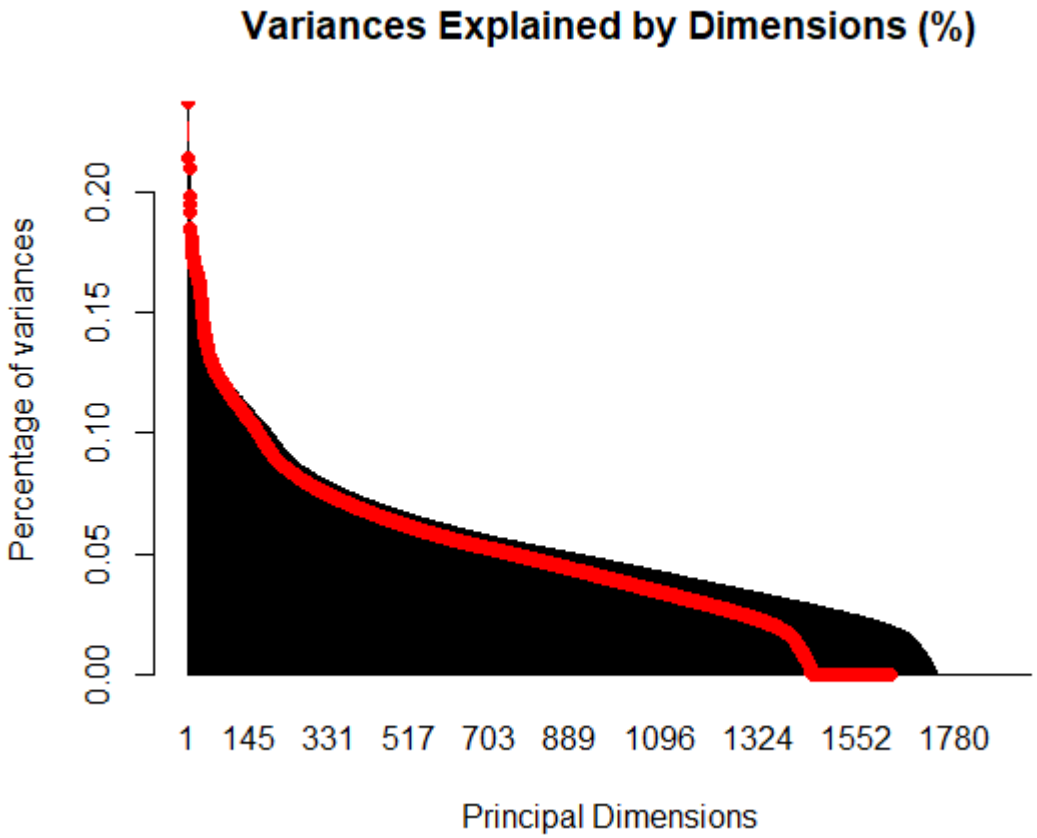
9) Satisfaction vs. Arrival Delay More Than 5 Minutes

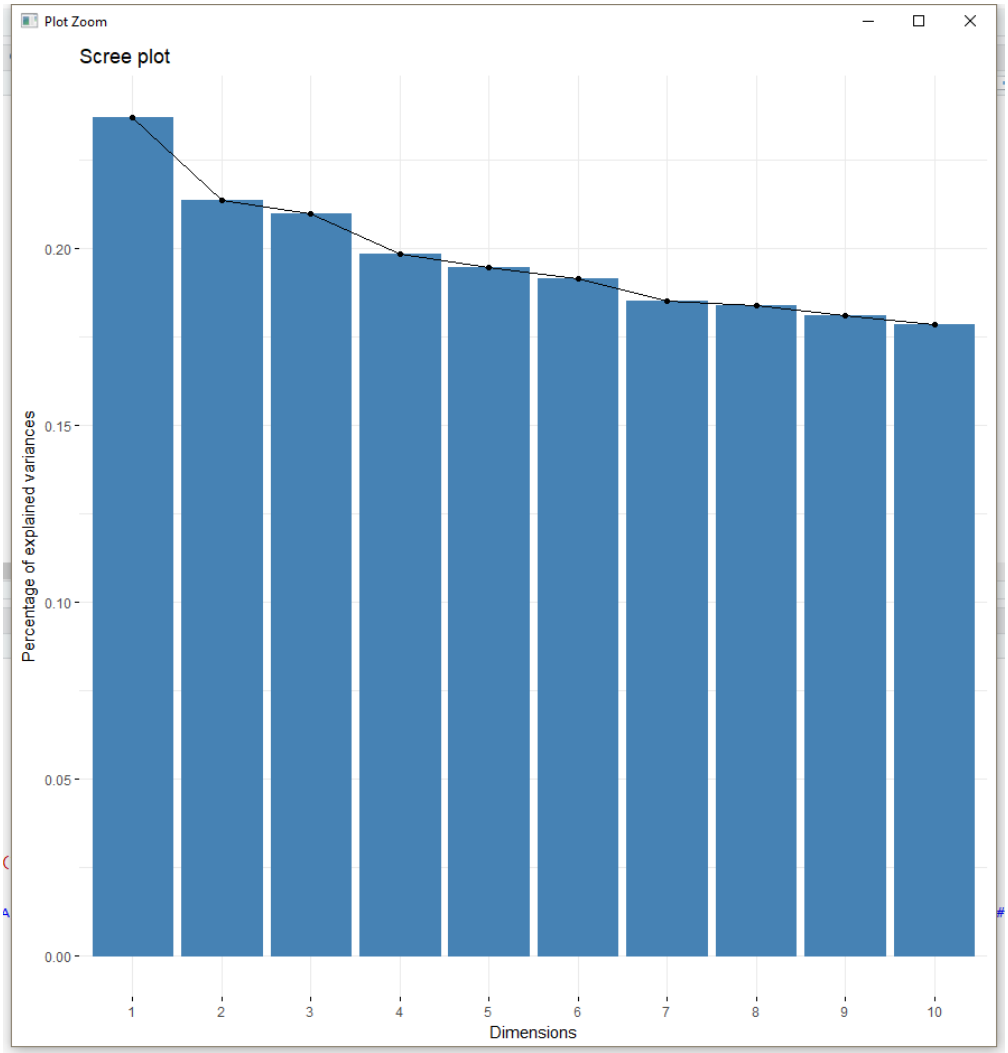


Satisfaction scores are more spread when arrival delay is more than 5 minutes and lower than when arrival delay is less than 5 minutes.

Use of Modeling Techniques & Visualizations

A. Multiple Correspondence Analysis





Due to the huge and complex dataset we have, we cannot simply put all the variables into our model, and then choose the significant variable. So, before we start working on a model we need to do a dimension reduce. Therefore, we turn to the MCA/PCA method and test if it can provide us some basic idea in feature selection.

After we run fit a mca model, as can be seen from the two tables above, the first 3 dimensions have the most explaining power contributing to the whole matrix. Although the percentage of variances explained is really tiny, we can still select learn some general information from this model.

From the significant table (appendix 2), we found some of the variables selected by the model (p value>0.05), and they provided the R squared between each variable and each dimension. So, from the table, we find that the 19 variables below are significant:

- Origin City
- Origin State
- Destination City
- Destination State
- Flight time in minutes
- Flight Distance
- Schedule Departure Hour
- Flight Date
- Day of Month
- Eating and Drinking at Airport
- Flight Cancelled
- Arrival Delay in Minutes
- Shopping Amount at Airport
- Number of Flights in the Past
- Age
- Departure Delay in Minutes
- Number of Flight with Other Airlines

Next, we selected more independent variables due to Association Rules.

B. Association Rules

We divide the dependent variable “Satisfaction” to 2 types. Satisfaction is “Y” (meaning the customers are satisfied) when scores are higher than 3, and Satisfaction is “N” (meaning the customers are not satisfied) when scores are equal to or lower than 3.

According to the top 5 association rules (appendix 3), Satisfaction being “Y” occurs when gender is “Male”, type of travel is “Business Travel”, arrival delay in minutes is “0”, departure delay in minutes is “0”, flight cancelled is “No”, arrival delay greater than 5 minutes is “No”, and class is “Eco”.

Satisfaction being “N” (appendix 4) occurs when gender is “Female”, type of travel is “Personal Travel”, airline status is “Blue”, number of other loyalty cards is “0”, flight cancelled is “yes”, shopping amount at airport is “0”, class is “Eco”.

Therefore, the significant variables according to Association Rules are:

- Gender
- Type of Travel
- Arrival Delay in Minutes
- Departure Delay in Minutes
- Flight Cancelled
- Arrival Delay is Greater than 5 minutes
- Class
- Airline Status
- Number of Other Loyalty Cards

- Shopping Amount at Airport

Let’s build linear regression models.

C. Linear Regression Model

Base on MCA and association rules, the significant independent variables are :

- Gender
- Type of Travel
- Arrival Delay in Minutes
- Departure Delay in Minutes
- Flight Cancelled
- Arrival Delay is Greater than 5 minutes
- Class
- Airline Status
- Number of Other Loyalty Cards
- Shopping Amount at Airport
- Origin City
- Origin State
- Destination City
- Destination State
- Flight time in minutes
- Flight Distance
- Schedule Departure Hour
- Flight Date
- Day of Month
- Eating and Drinking at Airport

- Number of Flights in the Past
- Age
- Number of Flight with Other Airlines

After building a multiple linear regression mode (appendix 5) with those independent variables using 70% of our client data (train), variables with P values greater than 0.01 are removed. A reduced model is built with the remaining variables :

- Airline Status
- Age
- Gender
- Number of Flights in the Past
- Type of Travel
- Flight Cancelled
- Arrival Delay is Greater than 5 minutes
- Class

This reduced model is not significantly different from the full model (with all independent variables) according to the “Anova” function (appendix 6) result as the P value is high.

To prove accuracy of the linear model, we take the rest 30% of our client data as the test data and predict the satisfaction scores by using the above reduced linear model with independent variables from the test data. After comparing the predicted satisfaction scores and the real scores from the test data, we had an accuracy rate of 55.71% (appendix 7). Therefore, it is an appropriate linear model to predict the

satisfaction scores. Airline status, age, gender, number of flights in the past, type of flights, class, whether the flights cancelled, and whether the arrival delay was greater than 5 minutes are significant predictors.

After proving that linear model with the 8 predictors are significant, we run a linear regression with all client data (appendix 8). Based on the coefficients of the variables, we have interpretations such as:

- Airline Status

  With other variables being the same, customers with Blue Status have the lowest satisfaction scores.

- Age:

  The larger the age, the lower the satisfaction scores.

- Gender:

  Males rate 0.127 higher than females.

- Type of flights:

  Customers with Business Travel have the highest satisfaction scores. If a customer takes a Personal Travel, the satisfaction score will be more than 1 point lower than as if the customer takes a Business Travel.

- Class:

  Customers fly Business Class are the most satisfied. Customers with Economy Plus Class are the least satisfied.

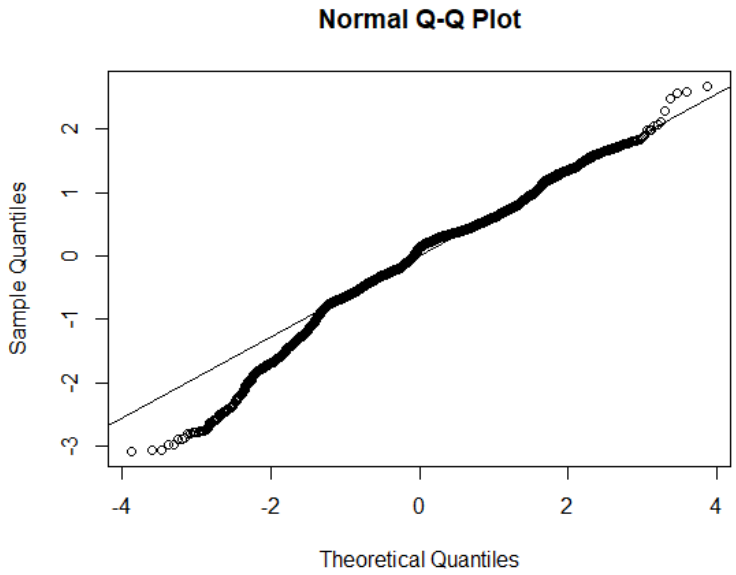
- Flight cancel:

  If the flight is cancelled, satisfaction score will be deducted by 0.248 point.

- Arrival delay:

If a flight's arrival time is delayed for more than 5 minutes, the satisfaction scores will be deducted by 0.346 point.

The 8 variables explain 43.38% variances of satisfaction scores (appendix 8). Then, we build another type of predictive model -- Ordinal Logistic Regression – to analyze the independent variables selected from MCA and Association Rules.



Based on the Q-Q plot, the overall fit of this linear regression model is fine. Only two tails a little apart. This is an appropriate model for our dataset.

D. Ordinal Logistic Regression

We build an ordinal logistic model with the variables from MCA and Association rules using 70% of our client data. To achieve the lowest AIC to have a better model, we remove variables which increase AIC. The remaining variables that kept the lowest AIC are: (model is at appendix 9)

- Gender
- Age

- Number of flight in the past
- Type of Travel
- Departure Delay in Minutes
- Arrival Delay in Minutes
- Flight Cancelled
- Arrival Delay is Greater than 5 minutes
- Class
- Airline Status
- Years of First Flight

To prove accuracy of this model, we use the rest 30% data as the test data and predict the satisfaction scores by using the above ordinal model with independent variables from the test data. After comparing the predicted satisfaction scores and the real scores from the test data, we concluded an accuracy rate of 59.46% (appendix 10). Thus, the ordinal model performs well. Next, we will further prove that the remaining variables influencing satisfaction scores the most with SVM.

E. SVM

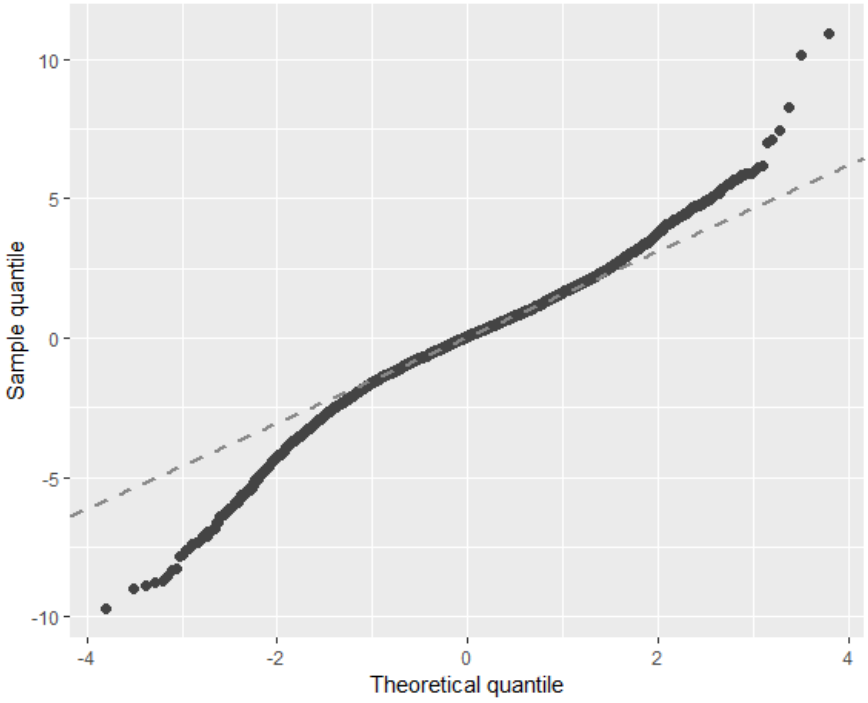
According to linear regression models and ordinal regression models, the independent variables they have in common are:

- Airline Status
- Age
- Gender
- Number of Flights in the Past
- Type of Travel

- Flight Cancelled
- Arrival Delay is Greater than 5 minutes
- Class

We build an SVM model with the remaining variables using 70% data of our client data and predicted satisfaction based on the SVM model by utilizing default settings (cost of constraints is 5 and the training error is 0.099; cross refers to cross-validation model) (appendix 11).

The error rate is 19.97% after comparing the real data and the predicted results (appendix 12). The error rate is low, so the predictors are proved to be appropriate.



This is the Q-Q plot for ordinal regression model. The overall fit is a little bit worse than the linear model, but there are no obvious extreme outliers. Only two tails are not perfectly fit. Combine adjusted R square and this Q-Q plot, we can still use this ordinal regression model.

F. Validation

To validate our models:

- Linear Regression Model

  After building a model with the train data, predicting with the test data, and comparing real and predicted Satisfaction, the accuracy rate of the model is 55.71%. Then we run a multiple linear regression with all client data, the adjusted R squared is 43.48% which is good. This model performs fine.

- Ordinal Logistic Model Rate

  Here we pre-process the data with same techniques and then use in an ordinal logistic regression model, the accuracy rate is 59.46% which is acceptable. This model works better than Linear Model.

- SVM error rate

  After utilizing the same method in data selection, we then fulfill an SVM model with the parameters setting mentioned above. After doing a prediction on test data, the error rate is 19.97% which is low compared to our ground truth level (a 50% to 50%), so our SVM works fine.

Actionable Insights

- Personal Travel

  In the linear model, the coefficient of personal travel is very low. In order to improve the user experience of personal travel customers, **we recommend to do a survey just based on these customers**. Further analysis on the survey results can help Southeast to come up with a specific plan to increase the satisfaction level of personal travel customers.

- Female

  In the associational rule, all female related components appear to be dissatisfaction. In linear and ordinal regression model, female has lower coefficient than male. In the stacked barplot, women are more active in survey, but they prefer to give low satisfaction. **We advice Southeast can provide more feminine care to improve women's satisfaction level. Southeast can also provide more help to mothers who carry babies. For example, we can set some area for baby-carrying mothers. By doing this, these babies won't influence other passengers and mothers won't feel uncomfortable when they take care of babies.**

- Blue status(most people, least satisfaction)

  From association rule and linear regression model, the clients of blue status are tend to give low satisfaction. The reason might be as clients become members of Southeast, they do not receive what they expect. The differences of expectations and realistics disappoint clients. **We think Southeast might need to change some policies for blue status or improve their service.**

- Economy Plus Class

  Based on the linear model and ordinal logistic model, customers flew Economy Plus were the least satisfied (worse than Economy), even though Economy Plus is more

expensive and meant to be better than Economy. **Southeast can investigate their Economy Plus services and management to find out what is making Economy Plus customers unhappy**. One of the reasons might be that customers paid more for Economy Plus but did not feel that the additional services worthed the extra cost.

- The Older people

  Based on boxplot, linear model and ordinal model, older people intend to give lower satisfaction. The old need more care than other clients. **Southeast could provide wheelchairs and accompanies at the airports and emergency aid for olds at the planes. So that the old won't feel lonely. The more reasons should be discussed by employees of Southeast.**

Appendix (code)

2.

```
> res$`Dim 1`$quali
                                          R2       p.value
Orgin.City                         0.726029859  0.000000e+00
Origin.State                       0.712358233  0.000000e+00
Destination.City                   0.732139508  0.000000e+00
Destination.State                  0.722927950  0.000000e+00
Flight.time.in.minutes             0.453850391  0.000000e+00
Flight.Distance                    0.941109379  0.000000e+00
Scheduled.Departure.Hour           0.102516948 9.103016e-205
Flight.date                        0.034014085  2.306194e-29
Day.of.Month                       0.017819526  7.749392e-22
Eating.and.Drinking.at.Airport     0.042683243  3.434529e-14
Flight.cancelled                   0.004850966  8.891953e-12
Arrival.Delay.in.Minutes           0.033871766  8.088502e-09
Shopping.Amount.at.Airport         0.035799806  9.034627e-09
No.of.Flights.p.a.                 0.017948871  6.321524e-08
Age                                0.014179102  8.385025e-07
Departure.Delay.in.Minutes         0.029704892  6.705693e-06
X..of.Flight.with.other.Airlines   0.012224033  1.029683e-04
No..of.other.Loyalty.Cards         0.002729297  1.925961e-03
```

3.

```
> inspect(ordered_rules[1:5])
    lhs                                                rhs                 support confidence     lift count
[1] {Gender=Male,
     Type.of.Travel=Business travel,
     Arrival.Delay.in.Minutes=0,
     Flight.cancelled=No}                          => {Satisfaction=Y} 0.1414292  0.8000000 1.555220   948
[2] {Gender=Male,
     Type.of.Travel=Business travel,
     Arrival.Delay.in.Minutes=0,
     Flight.cancelled=No,
     Arrival.Delay.greater.5.Mins=no}              => {Satisfaction=Y} 0.1414292  0.8000000 1.555220   948
[3] {Gender=Male,
     Type.of.Travel=Business travel,
     Class=Eco,
     Departure.Delay.in.Minutes=0,
     Flight.cancelled=No,
     Arrival.Delay.greater.5.Mins=no}              => {Satisfaction=Y} 0.1127853  0.8008475 1.556868   756
[4] {Gender=Male,
     Type.of.Travel=Business travel,
     Class=Eco,
     Arrival.Delay.in.Minutes=0,
     Flight.cancelled=No}                          => {Satisfaction=Y} 0.1202447  0.8011928 1.557539   806
[5] {Gender=Male,
     Type.of.Travel=Business travel,
     Class=Eco,
     Arrival.Delay.in.Minutes=0,
     Flight.cancelled=No,
     Arrival.Delay.greater.5.Mins=no}              => {Satisfaction=Y} 0.1202447  0.8011928 1.557539   806
```

4.

```
> inspect(ordered_rules[1:5])
    lhs                                                rhs                 support confidence     lift count
[1] {Airline.Status=Blue,
     Gender=Female,
     Type.of.Travel=Personal Travel,
     No..of.other.Loyalty.Cards=0}                 => {Satisfaction=N} 0.1044309  0.9803922 2.018915   700
[2] {Airline.Status=Blue,
     Gender=Female,
     Type.of.Travel=Personal Travel,
     No..of.other.Loyalty.Cards=0,
     Flight.cancelled=No}                          => {Satisfaction=N} 0.1011487  0.9797688 2.017631   678
[3] {Airline.Status=Blue,
     Type.of.Travel=Personal Travel,
     No..of.other.Loyalty.Cards=0,
     Shopping.Amount.at.Airport=0}                 => {Satisfaction=N} 0.1038341  0.9734266 2.004571   696
[4] {Airline.Status=Blue,
     Type.of.Travel=Personal Travel,
     No..of.other.Loyalty.Cards=0,
     Shopping.Amount.at.Airport=0,
     Flight.cancelled=No}                          => {Satisfaction=N} 0.1007012  0.9726225 2.002915   675
[5] {Airline.Status=Blue,
     Gender=Female,
     Type.of.Travel=Personal Travel,
     Class=Eco}                                    => {Satisfaction=N} 0.1262121  0.9724138 2.002485   846
```

5.

```
Call:
lm(formula = as.numeric(Satisfaction) ~ Airline.Status + Age +
    Gender + No.of.Flights.p.a. + Type.of.Travel + Class + Flight.cancelled +
    Arrival.Delay.greater.5.Mins, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0861 -0.4258  0.1057  0.4501  2.6627

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.8913524  0.0408835  95.182  < 2e-16 ***
Airline.StatusGold             0.4431065  0.0316805  13.987  < 2e-16 ***
Airline.StatusPlatinum         0.3284433  0.0518050   6.340 2.45e-10 ***
Airline.StatusSilver           0.6216132  0.0224996  27.628  < 2e-16 ***
Age                           -0.0032522  0.0005465  -5.951 2.80e-09 ***
GenderMale                     0.1383541  0.0180417   7.669 1.99e-14 ***
No.of.Flights.p.a.            -0.0031874  0.0006548  -4.868 1.15e-06 ***
Type.of.TravelMileage tickets -0.1091439  0.0347512  -3.141  0.00169 **
Type.of.TravelPersonal Travel -1.0213255  0.0212991 -47.951  < 2e-16 ***
ClassEco                      -0.0923984  0.0314467  -2.938  0.00331 **
ClassEco Plus                 -0.0948901  0.0407364  -2.329  0.01987 *
Flight.cancelledYes           -0.2308569  0.0761166  -3.033  0.00243 **
Arrival.Delay.greater.5.Minsyes -0.3374664  0.0191903 -17.585  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7135 on 6690 degrees of freedom
Multiple R-squared:  0.4395,    Adjusted R-squared:  0.4385
F-statistic: 437.1 on 12 and 6690 DF,  p-value: < 2.2e-16
```

6.

```
> anova(ModelFull,ModelReduced)
Analysis of Variance Table

Model 1: as.numeric(Satisfaction) ~ Airline.Status + Age + Gender + Price.Sensitivity +
    Year.of.First.Flight + No.of.Flights.p.a. + X..of.Flight.with.other.Airlines +
    Type.of.Travel + No..of.other.Loyalty.Cards + Shopping.Amount.at.Airport +
    Eating.and.Drinking.at.Airport + Class + Day.of.Month + Flight.date +
    Orgin.City + Origin.State + Destination.City + Destination.State +
    Scheduled.Departure.Hour + Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes +
    Flight.cancelled + Flight.time.in.minutes + Flight.Distance +
    Arrival.Delay.greater.5.Mins
Model 2: as.numeric(Satisfaction) ~ Airline.Status + Age + Gender + No.of.Flights.p.a. +
    Type.of.Travel + Class + Flight.cancelled + Arrival.Delay.greater.5.Mins
  Res.Df    RSS   Df Sum of Sq      F Pr(>F)
1   6437 3241.8
2   6690 3379.8 -253   -138.03 1.0833 0.1784
>
```

7.

```
> Predict<-predict(ModelReducedFinal, newdata = test, type = "response")
> test$SatisfactionPred <-predict(ModelReducedFinal, newdata = test, type = "response")
> # accuracy rate
> id <- which(test$Satisfaction!=round(test$SatisfactionPred), arr.ind = TRUE)
> Rate <- 1-length(id)/nrow(test)
> print(Rate)
[1] 0.5570633
```

8.

```
Call:
lm(formula = as.numeric(Satisfaction) ~ Airline.Status + Age +
    Gender + No.of.Flights.p.a. + Type.of.Travel + Class + Flight.cancelled +
    Arrival.Delay.greater.5.Mins, data = ClientData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0697 -0.4213  0.1268  0.4398  2.6682

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.8898651  0.0344703 112.847  < 2e-16 ***
Airline.StatusGold             0.4253867  0.0270334  15.736  < 2e-16 ***
Airline.StatusPlatinum         0.2974239  0.0419783   7.085 1.49e-12 ***
Airline.StatusSilver           0.6044553  0.0188256  32.108  < 2e-16 ***
Age                           -0.0030523  0.0004600  -6.636 3.41e-11 ***
GenderMale                     0.1273885  0.0151751   8.395  < 2e-16 ***
No.of.Flights.p.a.            -0.0030852  0.0005519  -5.590 2.33e-08 ***
Type.of.TravelMileage tickets -0.1320373  0.0286085  -4.615 3.98e-06 ***
Type.of.TravelPersonal Travel -1.0290593  0.0179223 -57.418  < 2e-16 ***
ClassEco                      -0.0799212  0.0266173  -3.003  0.00268 **
ClassEco Plus                 -0.0818114  0.0342686  -2.387  0.01699 *
Flight.cancelledYes           -0.2476917  0.0631302  -3.924 8.79e-05 ***
Arrival.Delay.greater.5.Minsyes -0.3461423  0.0161264 -21.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7165 on 9564 degrees of freedom
Multiple R-squared:  0.4345,    Adjusted R-squared:  0.4338
F-statistic: 612.4 on 12 and 9564 DF,  p-value: < 2.2e-16
```

9.

```
Call:
polr(formula = factor(Satisfaction) ~ Airline.Status + Age +
    Gender + No.of.Flights.p.a. + Type.of.Travel + Class + Departure.Delay.in.Minutes +
    Flight.cancelled + Arrival.Delay.greater.5.Mins + Year.of.First.Flight +
    Arrival.Delay.in.Minutes, data = train, Hess = TRUE)

Coefficients:
                                    Value Std. Error  t value
Airline.StatusGold                1.104676  9.549e-02   11.568
Airline.StatusPlatinum            1.574715  1.017e-02  154.877
Airline.StatusSilver              1.594335  6.364e-02   25.052
Age                              -0.005919  1.464e-03   -4.042
GenderMale                        0.365227  4.970e-02    7.349
No.of.Flights.p.a.               -0.008475  1.779e-03   -4.764
Type.of.TravelMileage tickets    -0.386031  8.916e-02   -4.330
Type.of.TravelPersonal Travel    -2.765194  6.700e-02  -41.269
ClassEco                         -0.174921  4.155e-02   -4.210
ClassEco Plus                    -0.294281  3.031e-02   -9.709
Departure.Delay.in.Minutes        0.006046  2.404e-03    2.515
Flight.cancelledYes              -0.577979  2.192e-03 -263.648
Arrival.Delay.greater.5.Minsyes  -0.949881  6.826e-02  -13.915
Year.of.First.Flight              0.019158  6.303e-05  303.946
Arrival.Delay.in.Minutes         -0.005278  2.528e-03   -2.088

Intercepts:
    Value      Std. Error t value
1|2    32.0931     0.0019 16608.6417
2|3    34.9951     0.0816   428.6501
3|4    37.2346     0.0911   408.8569
4|5    40.3782     0.1106   365.1370

Residual Deviance: 14040.12
AIC: 14078.12
```

10.

```
> Predict<-predict(OLM_ModelReducedF, newdata = test, type = "class")
> test$SatisfactionPred <-predict(OLM_ModelReducedF, newdata = test, type = "class")
> ## accuracy rate
> id <- which(test$Satisfaction!=test$SatisfactionPred, arr.ind = TRUE)
> Rate <- 1-length(id)/nrow(test)
> print(Rate)
[1] 0.5946416
```

11.

```
> library(kernlab)
> library(e1071)
> ### set up data
> vBuckets <- replicate(length(ClientData$Satisfaction), "N")
> vBuckets[ClientData$Satisfaction > 3] <- "Y"
> ClientDataSVM <- ClientData
> ClientDataSVM$Satisfaction <- vBuckets
> dim(ClientDataSVM)
[1] 9577   26
> trainSVM <- SamF(ClientDataSVM,0.7)[[1]]
> testSVM <- SamF(ClientDataSVM, 0.7)[[2]]
> svm_Reduced <-
+   ksvm(Satisfaction ~ Airline.Status + Age + Gender  + No.of.Flights.p.a.+Type.of.Travel+Class +Flight.cancelled+Arrival
.Delay.greater.5.Mins ,data = trainSVM,kernel = "rbfdot",kpar = "automatic", C = 5,cross = 3,prob.model = TRUE)
> svm_Reduced
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.242975337810414

Number of Support Vectors : 2939

Objective Function Value : -13225.7
Training error : 0.187677
Cross validation error : 0.201851
Probability model included.
```

12.

```
> svmPred <- predict(svm_Reduced,testSVM, type = "votes")
> compTable <- data.frame(testSVM$Satisfaction,svmPred[2,])
> table(compTable)
                    svmPred.2...
testSVM.Satisfaction    0    1
                   N  901  483
                   Y   91 1399
> errorRate <- (table(compTable)[2,1]+table(compTable)[1,2])/
+   (table(compTable)[1,1]+table(compTable)[1,2]+table(compTable)[2,1]+table(compTable)[2,2])
> errorRate
[1] 0.1997216
```

13.

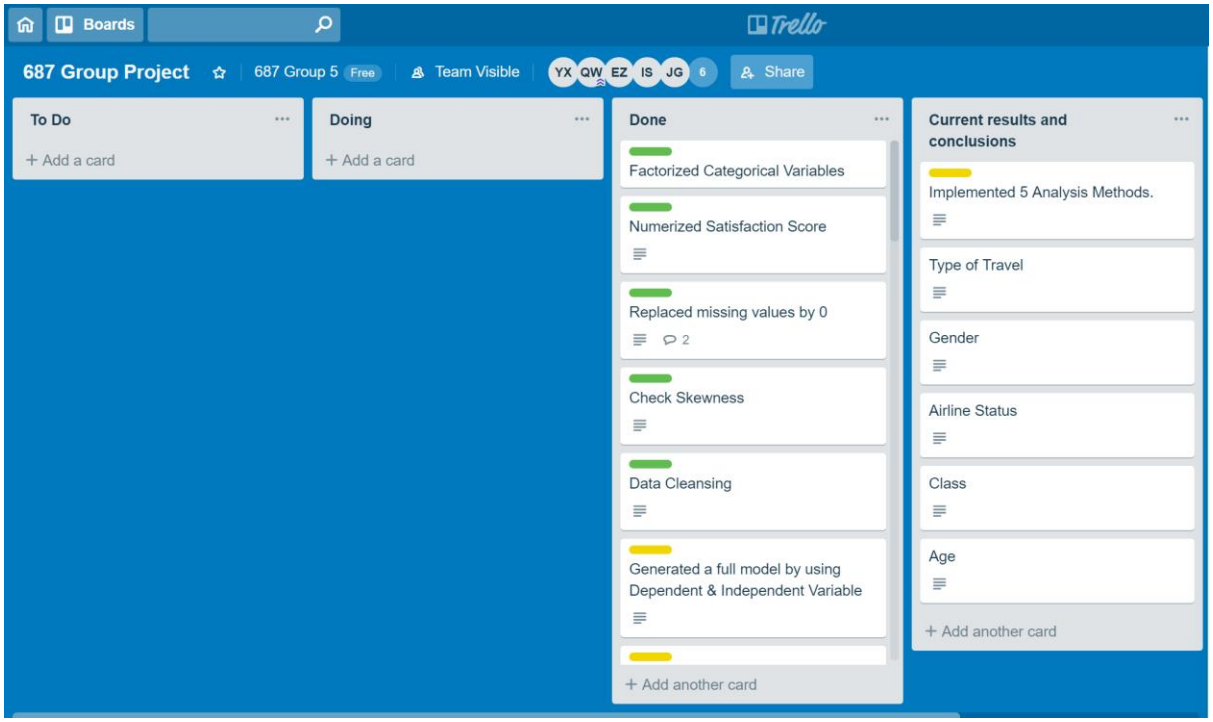
```
> ############################ Satisfication Level Bar Chart ############################
> propSat<- prop.table(table(CleanData$Satisfaction))
> smallData = as.data.frame(propSat)
> colnames(smallData)=c('SatisficationLevel', 'Frequency')
> str(smallData)
'data.frame':   5 obs. of  2 variables:
 $ SatisficationLevel: Factor w/ 5 levels "1","2","3","4",..: 1 2 3 4 5
 $ Frequency         : num  0.0231 0.1816 0.2848 0.4139 0.0967
>
> gbar=ggplot(smallData, aes(x=SatisficationLevel , y=Frequency, fill=SatisficationLevel ))
>
> gbar + geom_bar(stat = 'identity') + scale_fill_brewer() + ggtitle('Percentage of each Satisfication Level for all airlines companies') +
+   guides(fill = FALSE) + theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1))
>
> propSat<- prop.table(table(ClientData$Satisfaction))
> smallData_ = as.data.frame(propSat)
> colnames(smallData_)=c('SatisficationLevel', 'Frequency')
>
> gbar=ggplot(smallData_, aes(x=SatisficationLevel , y=Frequency, fill=SatisficationLevel ))
>
> gbar + geom_bar(stat = 'identity') + scale_fill_brewer() + ggtitle('Percentage of each Satisfication Level for Southeast Airline Co.') +
+   guides(fill = FALSE) + theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1))
>
>
> ######################## Proportion of Satisfication Levle per Airline Status ####################
> smallData1<- as.data.frame(prop.table(table(train$Airline.Status, train$Satisfaction)))
> colnames(smallData1)=c('AirlineStatus','SatisficationLevel' ,'Percentage_Status')
> View(smallData1)
>
> gbar = ggplot(smallData1, aes(x = AirlineStatus, y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfication Level per Airlin
e Status') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
>
> ######################## Proportion of Satisfication Levle per gender ####################
> smallData2<- as.data.frame(prop.table(table(train$Gender, train$Satisfaction)))
> colnames(smallData2)=c('Gender','SatisficationLevel' ,'Percentage_Status')
> View(smallData2)
>
> gbar = ggplot(smallData2, aes(x = Gender, y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfication Level per Gender') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
```

```
>
> ######################## Proportion of Satisfication Levle per Price Sencitivity ###################
> smallData3<- as.data.frame(prop.table(table(train$Price.Sensitivity, train$Satisfaction)))
> colnames(smallData3)=c('PriceSensitivity','SatisficationLevel' ,'Percentage_Status')
> View(smallData3)
>
> gbar = ggplot(smallData3, aes(x = PriceSensitivity, y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfication Level per Pri
ce Sensitivity') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
Warning message:
Removed 5 rows containing missing values (geom_bar).
> ######################## Proportion of Satisfication Levle per type of travel  ###################
> smallData4<- as.data.frame(prop.table(table(train$Type.of.Travel, train$Satisfaction)))
> colnames(smallData4)=c('TravleType','SatisficationLevel' ,'Percentage_Status')
> View(smallData4)
>
> gbar = ggplot(smallData4, aes(x = TravleType, y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfication Level per Travel Ty
pe') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
>
> ######################## Proportion of Satisfication Levle per Flight Cancelled  ###################
> smallData5<- as.data.frame(prop.table(table(train$Flight.cancelled, train$Satisfaction)))
> colnames(smallData5)=c('FlightCancelled','SatisficationLevel' ,'Percentage_Status')
> View(smallData5)
>
> gbar = ggplot(smallData5, aes(x = FlightCancelled , y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfication Level per Fli
ght Cancelled') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
> ######################## Proportion of Satisfication Levle per Class  ###################
> smallData6<- as.data.frame(prop.table(table(train$Class, train$Satisfaction)))
> colnames(smallData6)=c('Class','SatisficationLevel' ,'Percentage_Status')
> View(smallData6)
>
> gbar = ggplot(smallData6, aes(x = Class , y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfaction Level per Class') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> ######################## Proportion of Satisfication Levle per Class  ###################
> smallData6<- as.data.frame(prop.table(table(train$Class, train$Satisfaction)))
> colnames(smallData6)=c('Class','SatisficationLevel' ,'Percentage_Status')
> View(smallData6)
>
> gbar = ggplot(smallData6, aes(x = Class , y = Percentage_Status, fill = SatisficationLevel )) + ggtitle('Proportion of Satisfaction Level per Class') +
+   theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_text(vjust = -1))
>
> plot1 = gbar + geom_bar(stat = 'identity')
> plot1
> plot2 = gbar + geom_bar(stat = 'identity', position = 'fill')
> plot2
>
```

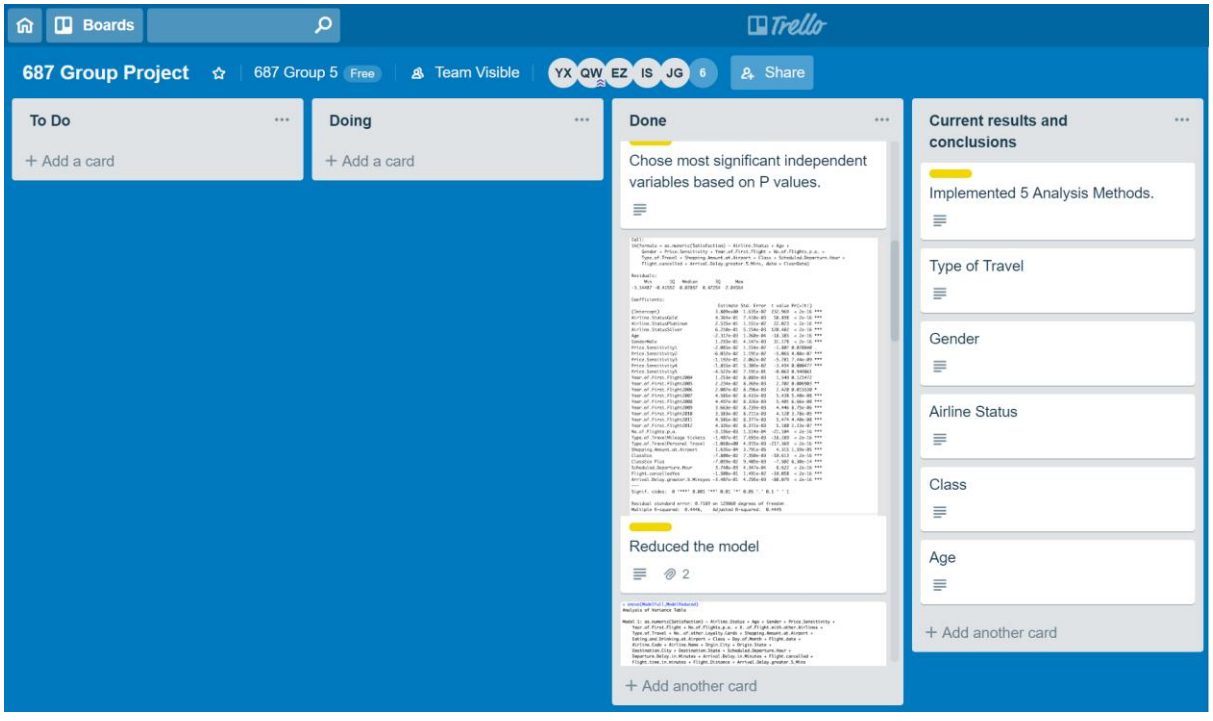
14.

```
> ##############################################################################################################
> # Read Data
> ProjectData <-read.csv( "Satisfaction Survey.CSV")
> ################################################ Data Cleansing ##############################################
>
> # Show NAs in each colum
> colSums(is.na(ProjectData))
                  Satisfaction                 Airline.Status                            Age                         Gender                Price.Sensi
tivity            Year.of.First.Flight              No.of.Flights.p.a.
                             0                              0                              0                              0
     0                        0                              0
X..of.Flight.with.other.Airlines                 Type.of.Travel       No..of.other.Loyalty.Cards       Shopping.Amount.at.Airport   Eating.and.Drinking.at.A
irport                    Class                      Day.of.Month
                             0                              0                              0                              0
     0                        0                              0
                  Flight.date                   Airline.Code                   Airline.Name                    Orgin.City                    Origin
.State                Destination.City               Destination.State
                             0                              0                              0                              0
     0                        0                              0
      Scheduled.Departure.Hour       Departure.Delay.in.Minutes        Arrival.Delay.in.Minutes              Flight.cancelled          Flight.time.in.m
inutes                Flight.Distance     Arrival.Delay.greater.5.Mins
                             0                           2345                           2738                              0
  2738                       0                              0
>
> # Treat NA values
> CleanData <- ProjectData
> CleanData$Departure.Delay.in.Minutes[is.na(ProjectData$Departure.Delay.in.Minutes)] <-0
> CleanData$Arrival.Delay.in.Minutes[is.na(ProjectData$Arrival.Delay.in.Minutes)] <-0
> CleanData$Flight.time.in.minutes[is.na(ProjectData$Flight.time.in.minutes)] <-0
> CleanData <- na.omit(CleanData)
> colSums(is.na(CleanData))
                  Satisfaction                 Airline.Status                            Age                         Gender                Price.Sensi
tivity            Year.of.First.Flight              No.of.Flights.p.a.
                             0                              0                              0                              0
     0                        0                              0
X..of.Flight.with.other.Airlines                 Type.of.Travel       No..of.other.Loyalty.Cards       Shopping.Amount.at.Airport   Eating.and.Drinking.at.A
irport                    Class                      Day.of.Month
                             0                              0                              0                              0
     0                        0                              0
                  Flight.date                   Airline.Code                   Airline.Name                    Orgin.City                    Origin
.State                Destination.City               Destination.State
                             0                              0                              0                              0
     0                        0                              0
      Scheduled.Departure.Hour       Departure.Delay.in.Minutes        Arrival.Delay.in.Minutes              Flight.cancelled          Flight.time.in.m
inutes                Flight.Distance     Arrival.Delay.greater.5.Mins
                             0                              0                              0                              0
     0                        0                              0
```

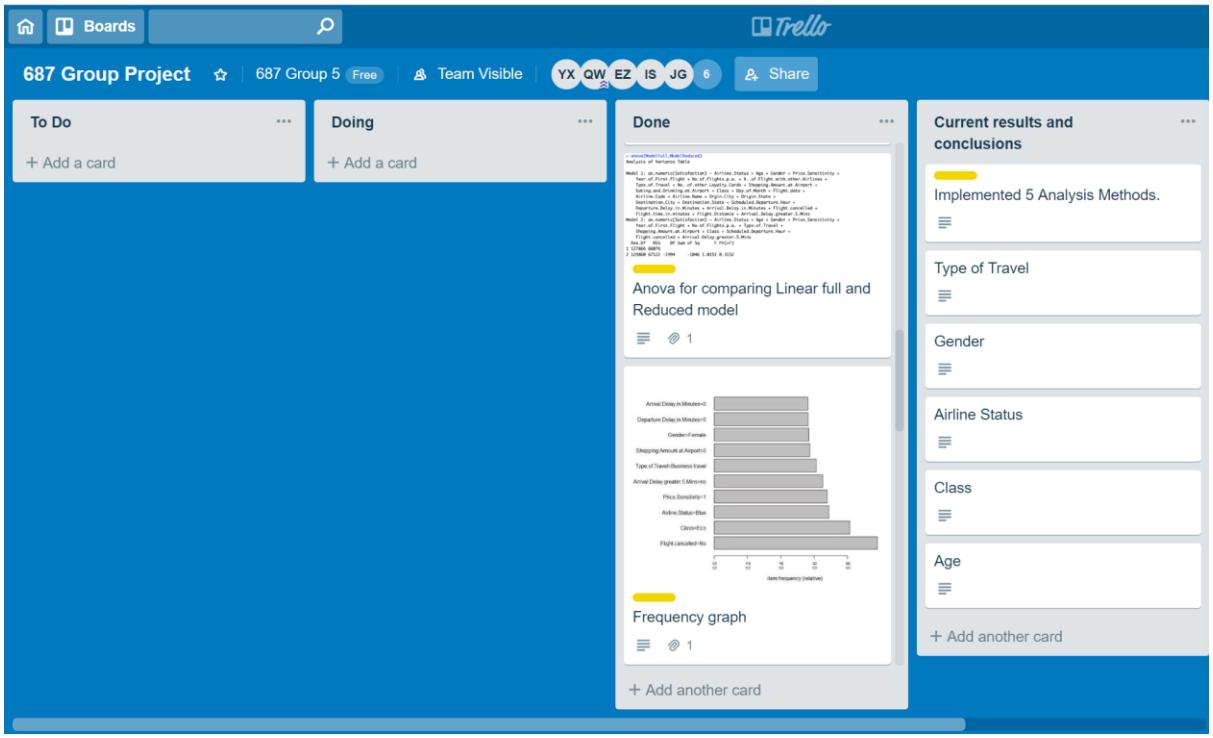
```
    0                             0                             0
> table(CleanData$Satisfaction)

       1        2      2.5        3      3.5        4 4.00.2.00   4.00.5      4.5        5
    2999    23587        2    36984        2    53758        2        1        2    12552
> # Formating the data
> ind1 <- which(CleanData$Satisfaction == '4.00.2.00')
> CleanData$Satisfaction[ind1[1]] <- 4
> CleanData$Satisfaction[ind1[2]] <- 4
> ind2 <- which(CleanData$Satisfaction == '4.00.5')
> CleanData$Satisfaction[ind2[1]] <- 5
> ind3 <- which(CleanData$Satisfaction == '2.5')
> CleanData$Satisfaction[ind3] <- 3
> ind4 <- which(CleanData$Satisfaction == '3.5')
> CleanData$Satisfaction[ind4] <- 4
> ind5 <- which(CleanData$Satisfaction == '4.5')
> CleanData$Satisfaction[ind5] <- 5
> CleanData$Satisfaction <- factor(CleanData$Satisfaction)
> table(factor(CleanData$Satisfaction))

    1     2     3     4     5
 2999 23587 36986 53762 12555
> CleanData$Satisfaction <- as.numeric(CleanData$Satisfaction)
> CleanData$Year.of.First.Flight <-factor(CleanData$Year.of.First.Flight)
> CleanData$Price.Sensitivity <- factor(CleanData$Price.Sensitivity)
> CleanData$Flight.Distance <- factor(CleanData$Flight.Distance)
> CleanData$X..of.Flight.with.other.Airlines <-factor(CleanData$X..of.Flight.with.other.Airlines)
> CleanData$Flight.cancelled <- factor(CleanData$Flight.cancelled)
>
```



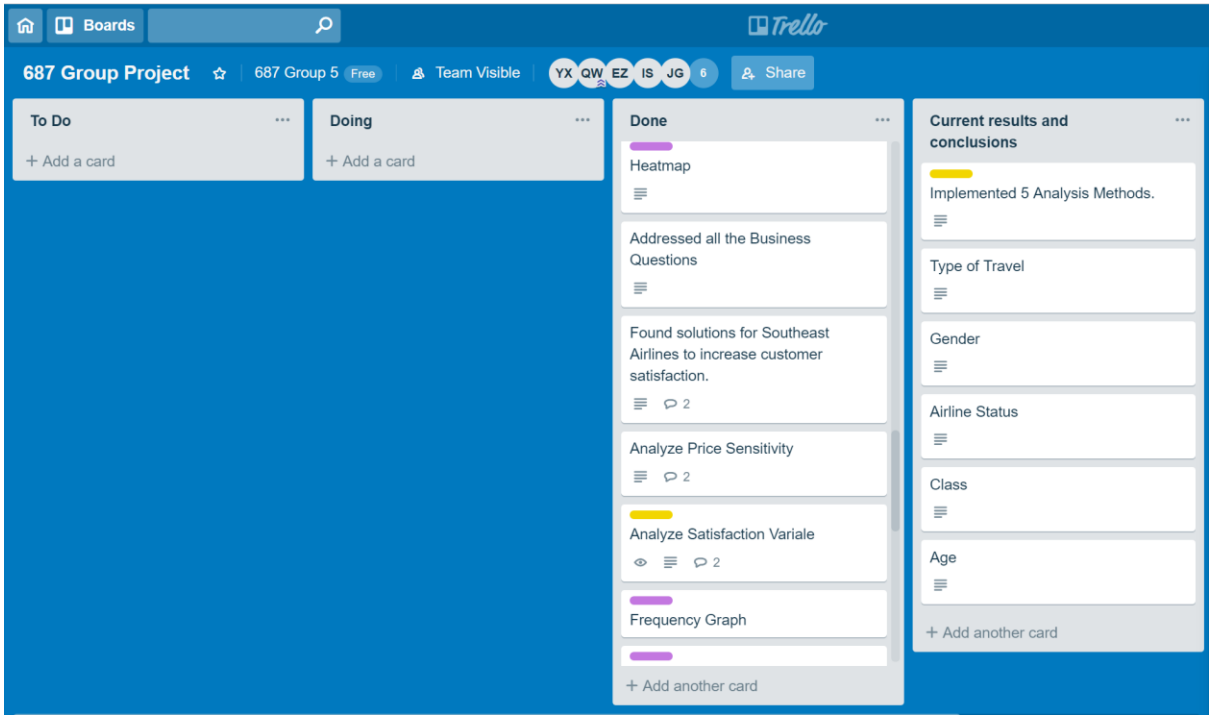
We cleanse our data by removing weird Satisfaction, rounding numbers, and replacing missing values with 0.



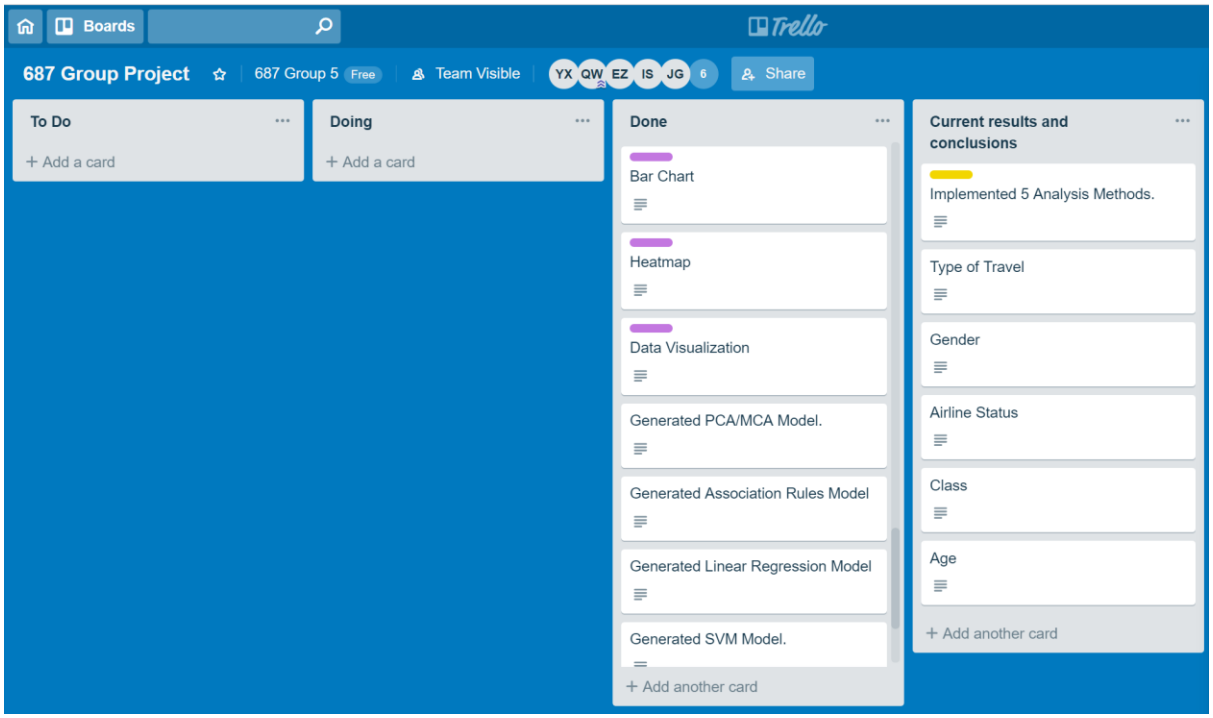
We generate full linear models and removed variables with large P values and build a reduced linear model.



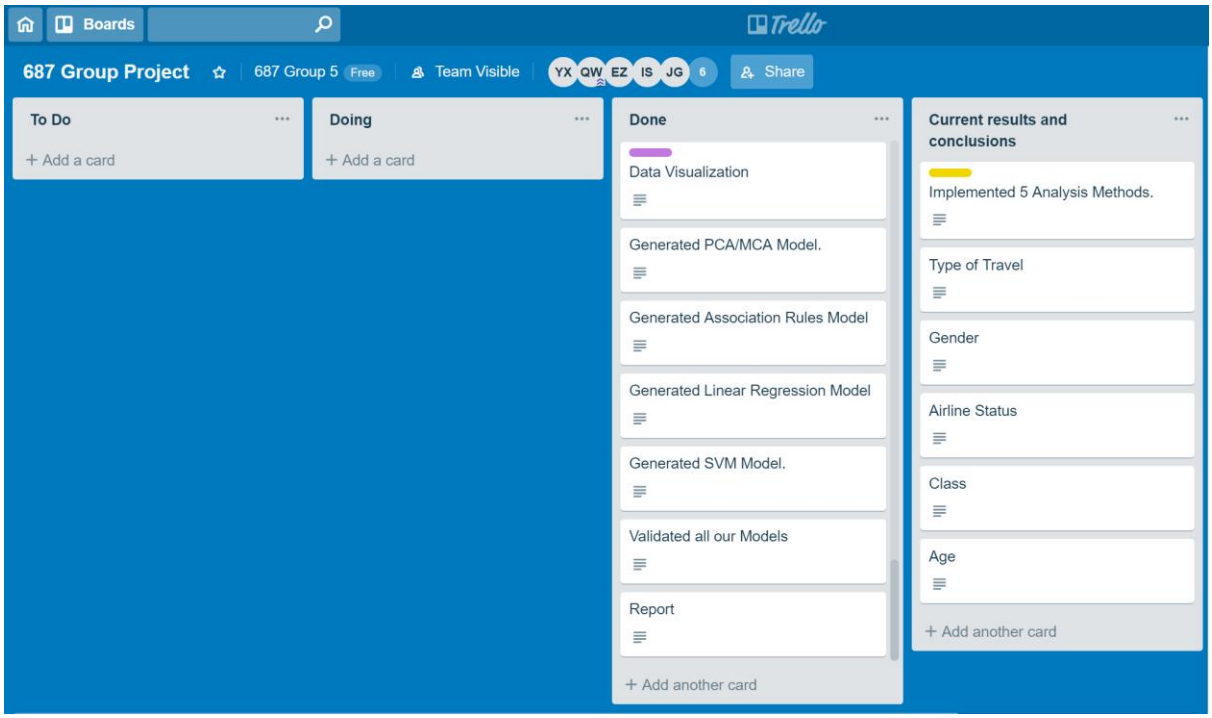
We compare full linear model and reduced linear model, and they are not very different. The reduced model is fine.



We make heatmaps to visualize Flight cancelled, Flight delay, and etc with Satisfaction. We Addressed business questions that can be found in the report. We analyzed dependent variable Satisfaction in the aspects of median, mean, quartiles, range, and etc.



We finalized our analyzing models to be MCA, Association Rules, Linear Regression, Ordinal Logistic, and SVM.



Among all the variables, Type of travel, Gender, Airline Status, Class, and Age appeared the most interesting. We come up conclusions based on the 5 variables.