

IST 687: Introduction to Data Science:

Customer Intelligence on Airline Data for South-East Airlines

Akshita Chandiramani

Eashani Deorukhkar

Jay Kachhadia

Rohan Mahajan

Yesaswi Avula

**Professor: Jeffrey Saltz**

**TA: Ivan Shamshurin**

Contents

[Introduction 4](#_gjdgxs)

[Objective 4](#_30j0zll)

[Background 4](#_1fob9te)

[Context 4](#_3znysh7)

[Scope 4](#_2et92p0)

[Business Questions 5](#_tyjcwt)

[Initial business questions 5](#_3dy6vkm)

[Final Business questions 6](#_1t3h5sf)

[Data Analysis 6](#_4d34og8)

[Data Acquisition 6](#_2s8eyo1)

[Data Cleansing 7](#_17dp8vu)

[Data Transformation 8](#_3rdcrjn)

[Data Munging 8](#_26in1rg)

[Descriptive statistics & Visualizations 9](#_lnxbz9)

[Use of modeling techniques & Visualizations 16](#_35nkun2)

[Linear Models 16](#_1ksv4uv)

[Association rules 22](#_44sinio)

[Amenities data 22](#_2jxsxqh)

[Demographic Information 24](#_z337ya)

[For USA states (Florida, California and Texas) 26](#_3j2qqm3)

[Classification Models 26](#_1y810tw)

[Support vector machine 27](#_4i7ojhp)

[Naïve Bayes 31](#_2xcytpi)

[Actionable Insights / Overall interpretation of results 32](#_1ci93xb)

[BQ 1: How is the overall satisfaction of Southeast customers compared to other airlines? 32](#_3whwml4)

[BQ 2: Does flight distance, flight time, arrival or departure delay affect customer satisfaction? 33](#_2bn6wsx)

[BQ3: How does age affect customer satisfaction? Which age group gives lower ratings? 3](#_qsh70q)4

[BQ 4: Is there any relationship between gender and customer satisfaction?](#_3as4poj)  35

[Are females more likely to give](#_3as4poj) lower rating to the airline[?](#_3as4poj)

[BQ 6: Does Year of first flight and No. of flights p.a affect customer satisfaction?](#_49x2ik5)

[BQ 7: Is it possible to associate airline status, type of travel and class with customer rating? 34](#_2p2csry)

[BQ 8: How much does Shopping, eating and drinking at the airport affect customer satisfaction? 34](#_147n2zr)

B[Q 9: Does date of travel (day, week, month) affect customer satisfaction? 34](#_3o7alnk)

BQ 10: Do particular Origins and Destinations cause low customer satisfaction? 35

BQ 11: Do cancelled flights cause customers to give a low customer rating? 36

[Conclusion 34](#_23ckvvd)

[Appendix - Code 35](#_ihv636)

[Descriptive analysis 35](#_32hioqz)

[Linear models 41](#_1hmsyys)

[Arules 53](#_41mghml)

[SVM Models 59](#_2grqrue)

[Experiments (Descriptive analysis and linear models) 62](#_vx1227)

[Experiment 2 (arules) 69](#_3fwokq0)

[References 79](#_1v1yuxt)

# Introduction

## Objective

We are working as a consulting company for Southeast Airlines. The focus of our survey analysis project is to provide some notions and directions to Southeast Airlines so that they can improve its profit and consumer satisfaction. Concrete analysis of the airlines' data containing 14 major airline companies in the United States including Southeast Airlines helps us generate some useful insights into customers mindsets.

## Background

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

Our goal here is to understand what factors influence a customer’s satisfaction after taking a particular airline’s flight. There can be different relationships between customers and all these 28 variables. Some of them make sense whereas some of them won’t make sense. Also, Various recommendations could be provided to our client Southeast airlines based on these insights we get from data analysis and model creation.

There are many business questions to be answered for improving Southeast airlines’ service. So, they could better satisfy their customers and increase their profits. Also, It’s essential to understand the market in which you are playing your game. So, Understanding your competitor’s strategy gives you actionable insights on where you should stress on to improve your business.

## Context

The Airlines data is a dataset collected from information about the customers taking various airline flights and giving their satisfaction ratings about the overall experience they had with the flight. Getting actionable insights from this dataset helps them improve the services, win more customers and increase revenues. The dataset that we got was dirty and required much cleaning due to inappropriate values and missing values in multiple columns which needs to be handled thoughtfully.

Southeast Airlines for understanding their customers and their experiences. We think that answering these questions will help Southeast executives know more about the ways to improve Southeast airline Services.

Models were also created to make predictions on customer satisfaction so that we could better understand the critical features and whether or not all the models depict the same findings or not and why.

## 

## Scope

By playing with the features in the dataset doing some feature engineering, we can get insights on how we can improve Southeast services to improve customer satisfaction, attract new customers and encourage existing customers to avail Southeast airline services more frequently.

We have 29 features out of which some were not useful actually. here’s the detailed description of all the features.

1. **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer?
   1. 5 means higher satisfied, and 1 is lowest level of satisfaction.
2. **Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.
3. **Age** – the specific customer’s age. That is starting from 15 to 85 years old.
4. **Gender** – male or female.
5. **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.
6. **Year of First Flight** – this attributes shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.
7. **No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.
8. **Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many time that customer fly with other Airlines.
9. **Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation
10. **No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefits such as, discounts.
11. **Shopping Amount at Airport** – showing the costumer’s result of how many products have been purchased. The range of shopping amount is from 0 to 875.
12. **Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.
13. **Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.
14. **Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.
15. **Flight date** – all of these data are abbreviate the passenger’s flight date travel, which were since 2014 and only in January, February, and March.
16. **Airline Code** – basically, it is unique two or three digits that mean what is the specific type of airline. There are several codes that consumers have been going with. For example, AA, AS, B6, and DL.
17. **Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and FlyToSun Airlines Inc. This attribute provide what airline name that passenger have been used.
18. **Origin City** – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo HO.
19. **Origin State** – same thing as origin city such as, what state that customers have departed from? A good example, Texas, Ohio, Alaska, and Utah.
20. **Destination City** – the place to which passenger travels to. For example, Akron HO, Alpena MI, Austin TX, and Boston MA.
21. **Destination State** – also, it is the same thing as origin city, such as, to what state passenger travel to? Some example of destination states, Alaska, Kentucky, Iowa, and Florida.
22. **Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.
23. **Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage are starting from 0 until 1128 minutes.
24. **Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang of delayed minutes in this data are starting from 0 until 1115 minutes.
25. **Flight Cancelled** – occurs when the airline dose not operates the flight at all, and that is for a certain reason.
26. **Flight time in minutes** – indicate to period time to the destination.
27. **Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.
28. **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

# Business Questions

Questions that drive the success of any business are called business questions. There are some questions which are very essential to understand in order to improve the service, attract more customers and increase customer satisfaction. Our Client Southeast airlines is interested in the same type of questions which can help them increase their profits and gain more customers.

We here also try to compare all the features of the dataset for Southeast airline and other airlines inorder to figure out where they are lacking and can improve upon. Trends shown by each of the feature in association with the customer satisfaction is taken into account to create a list of initial business questions.

## Initial business questions

1. How much are the customers satisfied with Airline services in the USA?
2. In which areas Southeast airlines lag with respect to the other airlines?
3. What are the factors which really affect the customer satisfaction?
4. Are there any difference in how females rate airline services of flights than males?
5. Does Age have any relationship with customer satisfaction?
6. Can we model customer satisfaction using some of features in the dataset?
7. Does type of travel affects customer satisfaction?

After performing analysis on airlines data, We found that there are many other questions which can be answered for southeast airlines and in general for all airline services. Our team was also interested in knowing whether any actionable insights can be drawn from other features which are either customer centric or airline centric. So, Following are the business questions we came up with after doing the exploratory data analysis and modeling.

## Final Business questions

Our new business questions were:

1. [How is the overall satisfaction of Southeast customers compared to other airlines?](#_3whwml4)
2. [Does flight distance, flight time, arrival or departure delay affect customer satisfaction?](#_2bn6wsx)
3. [How does age affect customer satisfaction? Which age group gives lower ratings?](#_qsh70q)
4. [Is there any relationship between gender and customer satisfaction?](#_3as4poj)
5. [Are females more likely to give](#_3as4poj) lower rating to the airline[?](#_3as4poj)
6. [Does Year of first flight and No. of flights p.a affect customer satisfaction](#_49x2ik5)?
7. [Is it possible to associate airline status, type of travel and class with customer rating](#_2p2csry)?
8. [How much does Shopping, eating and drinking at the airport affect customer satisfaction?](#_147n2zr)
9. [Does date of travel (day, week, month) affect customer satisfaction?](#_3o7alnk)
10. Do particular Origins and Destinations cause low customer satisfaction?
11. Do cancelled flights cause customers to give a low customer rating?

# Data Analysis

## Data Acquisition

Our professor provided us with a dataset that contains customer feedback surveys obtained from the guests that have travelled through southeast airlines which is our client. This data set a csv file that contains surveys from the year 2014.

We downloaded the file using the following code:



We checked the information of the given dataset using the following code.



## Data Cleansing

For the cleaning process, our team decided to work with all the months but subset the dataset based on the columns that can help us to answer our business questions. We separated the southeast airlines as it is our client. We separated it from the other airlines for the sake of the results.



After selecting the columns, we decided to use the method na.omit to delete the missing values wherever needed.

We decided not to fulfil blanks with an average value because we considered that these assumptions can affect our final results. Also, the majority of the columns presented categorical data, so it was difficult to get an average or middle value.

## Data Transformation

As it is mentioned in the previous section, many of the columns that we were working with contain categorical data. In order to work with linear models, it was necessary to transform the categorical values into numerical values. We transformed these values into numeric values using the following code:



Our team was also interested on knowing if the amenities have any influence on the likelihood to recommend. We used linear models to answer this question but it was necessary to transform the data from categorical to numeric. The columns of the amenities were different from other columns because they only presented two different values: “Y” or “N”. We transformed these columns using the method previously presented, and the new values were one for Y and cero for N.

We performed this process for all the amenities.

For the customer opinion scores, it was not necessary to transform the data because these values were numerical.

## Data Munging

Data munging is the process of programmatically transforming data into a format that makes it easier to work with (Mode Analytics, 2018).

After doing some experiments with the dataset, we noticed that our results were not what our team was expecting. We had some talks with our Professor, and he recommended us to subset the data by state, specifically he mentioned that it was better to work with the states with more surveys completed. These states were California, Florida and Texas.

For our different experiments, we subset the dataset by state in order to provide better suggestions to specific locations. We worked with data frames and vectors.



# Descriptive statistics & Visualizations

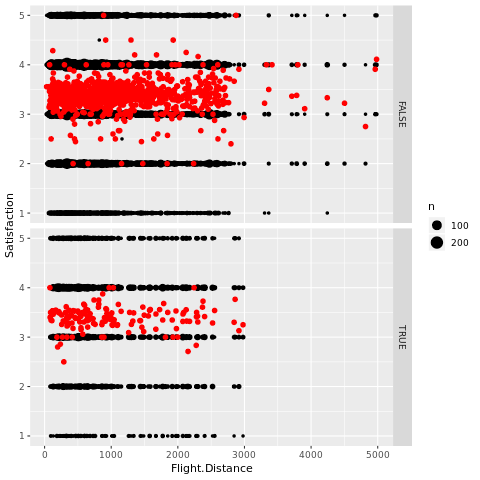
After cleaning the dataset, our dataset contains 129889 observations and 28 columns. We have customer survey data of our client Southeast as well as other airlines.

|  |  |  |  |
| --- | --- | --- | --- |
| **Airline Name** | **Total** | **Satisfied** | **Not Satisfied** |
| **Southeast** |  |  |  |
| **Others** |  |  |  |

[BQ 1: How is the overall satisfaction of Southeast customers compared to other airlines?](#_3whwml4)

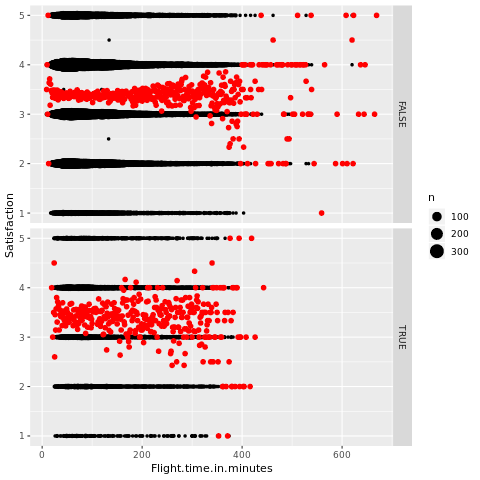
[BQ 2: Does flight distance, flight time, arrival or departure delay affect customer satisfaction?](#_2bn6wsx)

Satisfaction vs. Flight distance for Southeast and other airlines



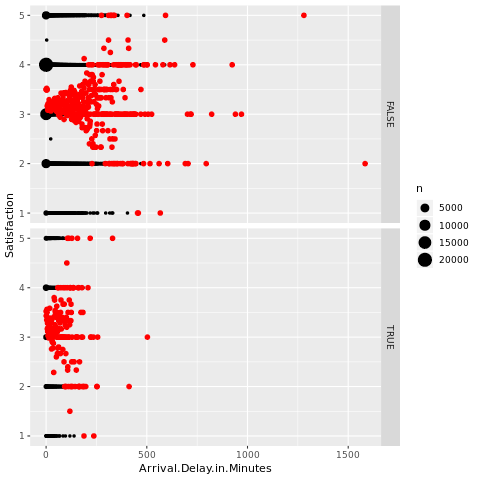
Satisfaction remains almost constant as flight distance increases

Satisfaction vs. Flight time for Southeast and other airlines



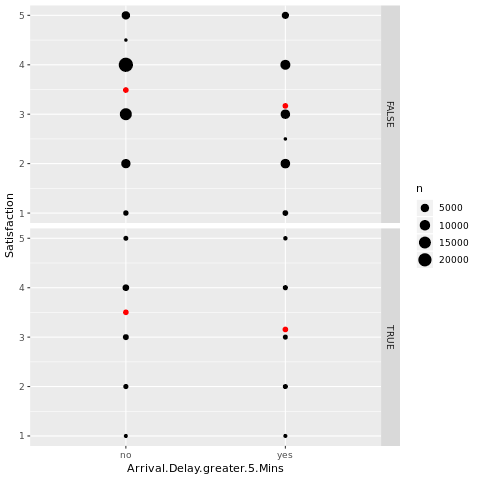
Satisfaction remains almost constant as Flight time increases

Satisfaction vs. Arrival Delay for Southeast and other airlines

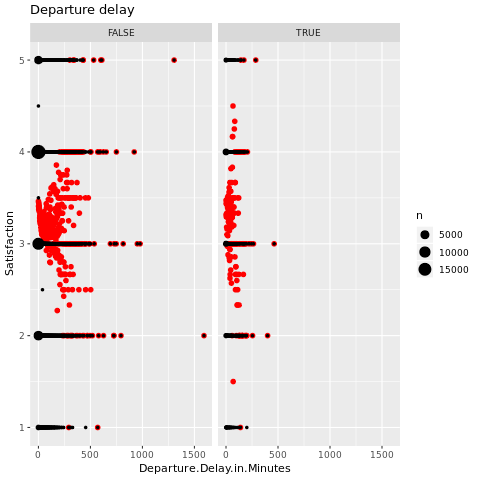


As arrival delay increases, the satisfaction ratings tend to reduce

Satisfaction vs. Arrival Delay greater than 5 minutes for Southeast and other airlines

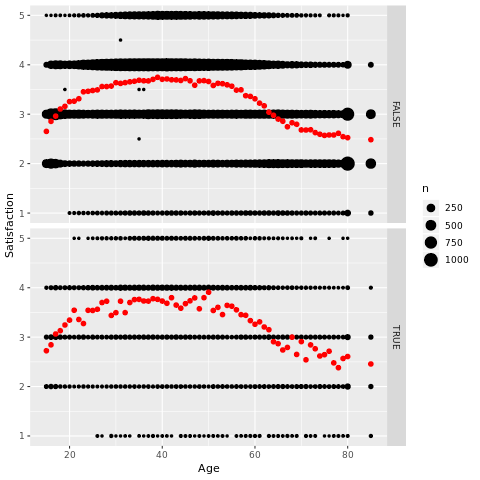


If the arrival delay is greater than 5 mins, the Satisfaction reduces by around 0.5 for South east and other airlines

Satisfaction vs. Departure Delay for Southeast and other airlines

Departure delay affects the Satisfaction rating by a small amount. There is evidence that as departure delay increases, the satisfaction reduces, but this is not true for all cases.

[BQ3: How does age affect customer satisfaction? Which age group gives lower ratings?](#_qsh70q)



Age affects satisfaction rating in both southeast and other airlines

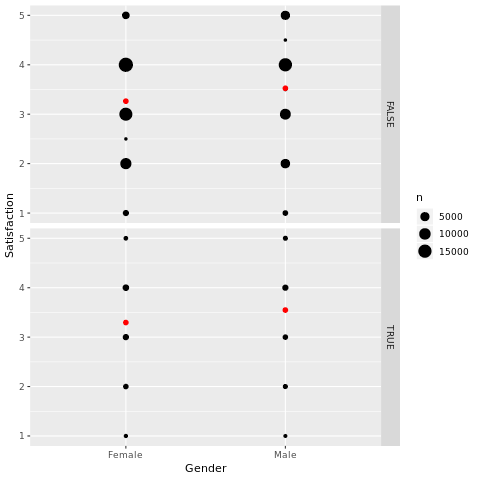
As customer’s age increases from 20 to 50, the satisfaction rating increases,

after which it starts decreasing till the age of 80

[BQ 4: Is there any relationship between gender and customer satisfaction?](#_3as4poj)

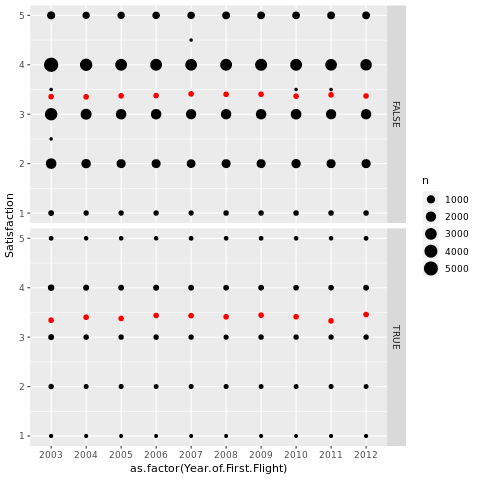
[Are females more likely to give](#_3as4poj) lower rating to the airline[?](#_3as4poj)

Satisfaction vs. Gender for Southeast and other airlines

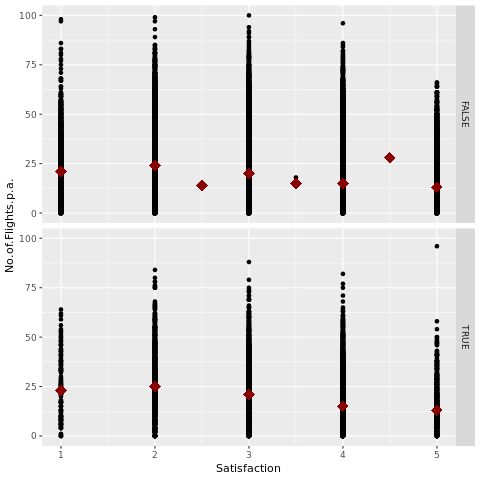


On an average, females give lesser satisfaction ratings compared to males

[BQ 6: Does Year of first flight and No. of flights p.a affect customer satisfaction?](#_49x2ik5)



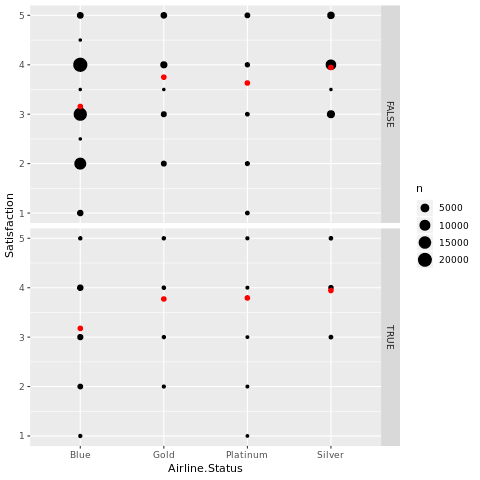
Year of flights does not affect the satisfaction ratings by a significant amount



For south east airlines, satisfaction tends to increase as the no. of flights p.a. decreases

[BQ 7: Is it possible to associate airline status, type of travel and class with customer rating?](#_2p2csry)

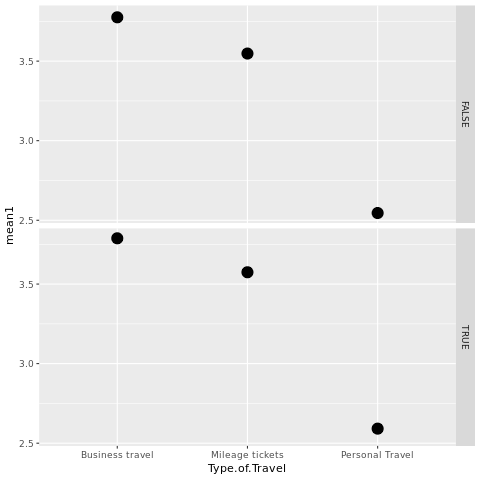
Satisfaction vs. Airline Status for Southeast and Other airlines



Customer’s travelling by blue status have the lowest satisfaction rating, followed by Platinum.

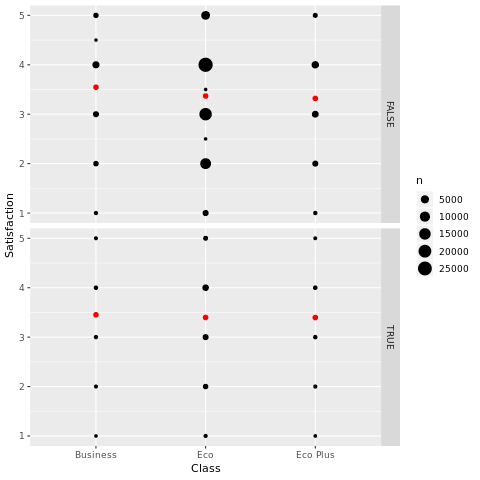
The customer ratings are higher for Gold and maximum for Silver. This holds true for both southeast and other airlines

Satisfaction vs. Type of travel for Southeast and other airlines



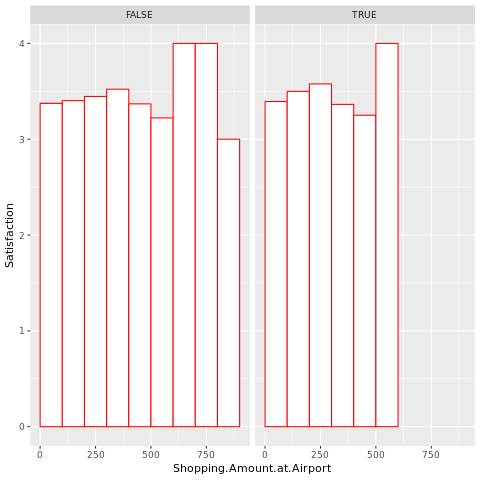
When customer’s travel for personal reasons, their satisfaction is the lowest (around 2.5)

Whereas, it is higher when the tickets are Mileage tickets and even higher for Business travel

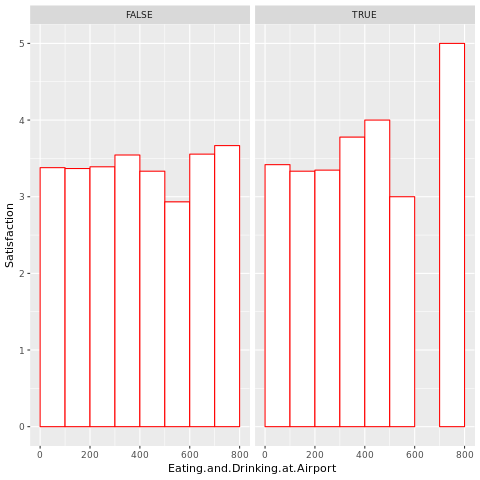


The satisfaction ratings are slightly higher for Business class compared to Eco and Eco plus

[BQ 8: How much does Shopping, eating and drinking at the airport affect customer satisfaction?](#_147n2zr)



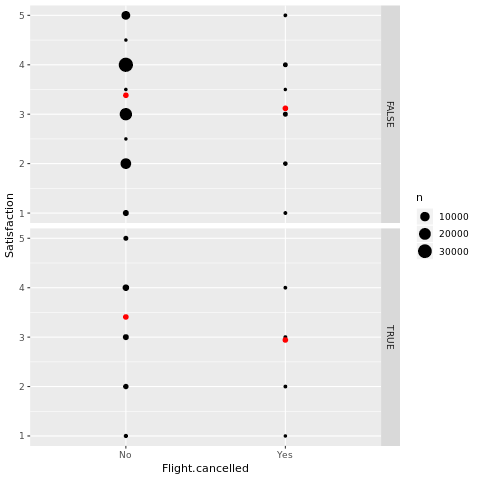
Customers who shop between 600-800 give a high customer satisfaction rating for other airlines whereas those who shop for 500-600 give a high customer rating for southeast airlines



B[Q 9: Does date of travel (day, week, month) affect customer satisfaction?](#_3o7alnk)

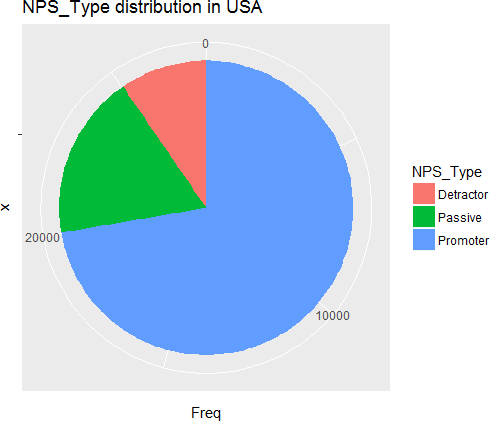
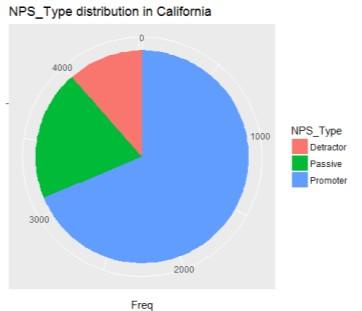
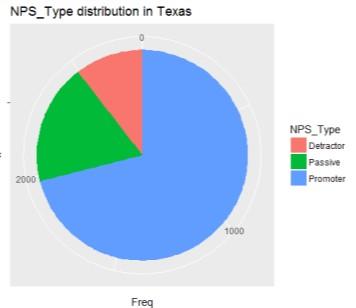
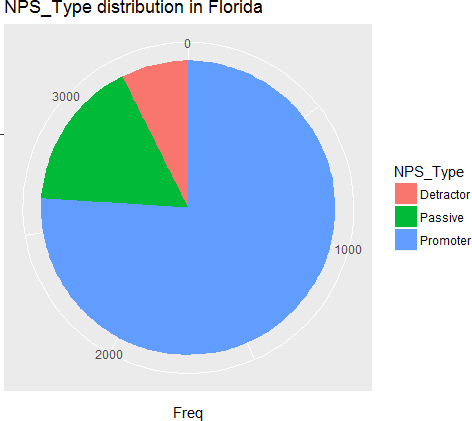
BQ 10: Do particular Origins and Destinations cause low customer satisfaction?

BQ 11: Do cancelled flights cause customers to give a low customer rating?



Cancelled flights cause customers to give a lower satisfaction rating for both southeast and other airlines

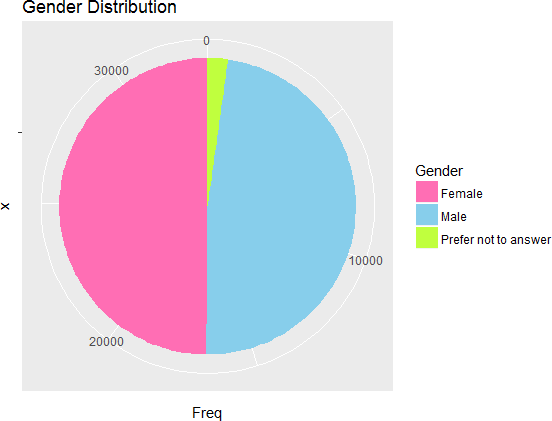
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **State** | **Detractors** | **Passives** | **Promoters** | **Total** | **NPS**  **(%Promoters - %Detractors)** |
| **California** | 519 | 895 | 3,107 | 4,521 | .57 |
| **Florida** | 255 | 573 | 2,627 | 3,455 | .68 |
| **Texas** | 290 | 516 | 1,977 | 2,783 | .60 |



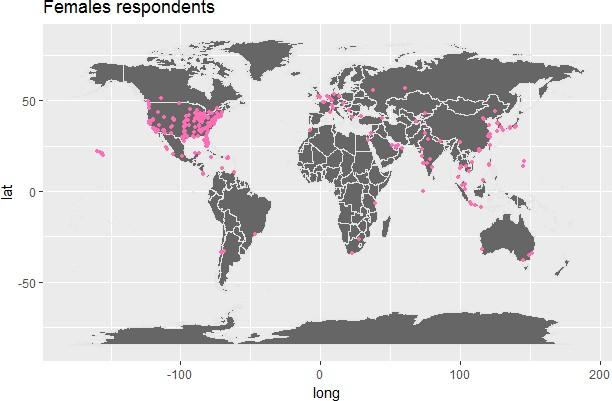
Based on the NPS score, our team concluded that Hotels in the USA have a good performance but it could be improved. In the following graphs, we can see the distribution of NPS in USA and in these three states.

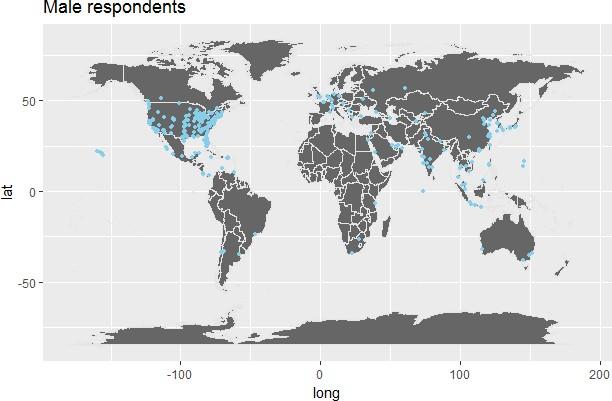
In the following sections, we explained how our team tried to use linear models to find the relationships between nationality and NPS\_Type.

Another important characteristic for our team was the gender of the survey respondents. In our dataset, 16566 of the survey respondents were females and 15916 were males.



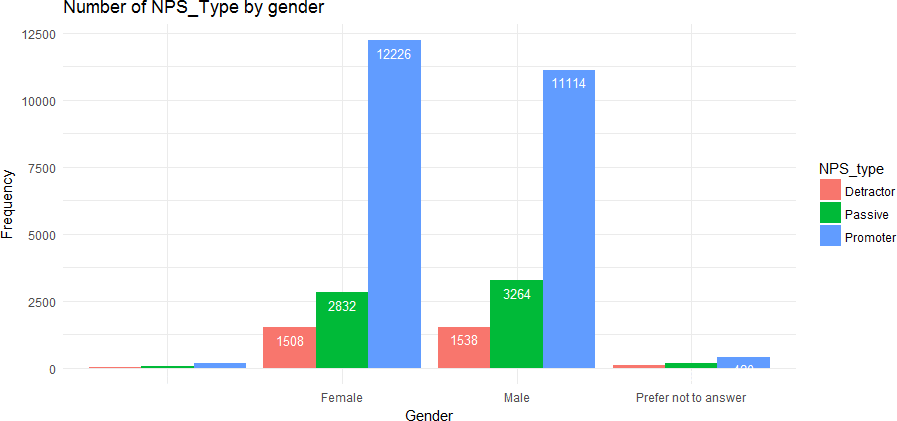
We used different graphs to locate the female and male respondents.



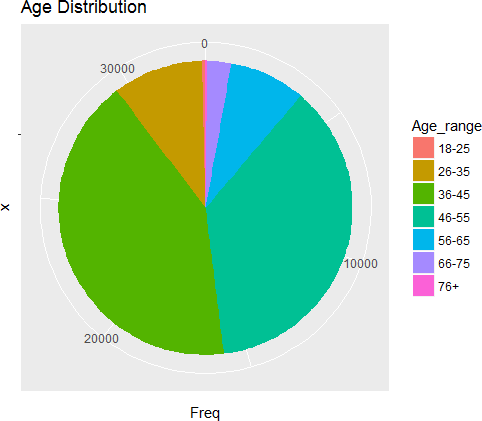


We were also interested on knowing the number of female detractors, and male detractors. From this descriptive analysis, we noticed that males tended to be more detractors than females.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender** | **Detractors** | **Passives** | **Promoters** | **Total** | **NPS**  **(%Promoters - %Detractors)** |
| **Female** | 1,508 | 2,832 | 12,226 | 16,566 | .64 |
| **Male** | 1,538 | 3,264 | 11,114 | 15,916 | .60 |



Another characteristic that our team considered important was the age range. We believed that the age has a direct influence in the NPS type, so created the following visualizations to know which age group gives the worst reviews to the hotel.

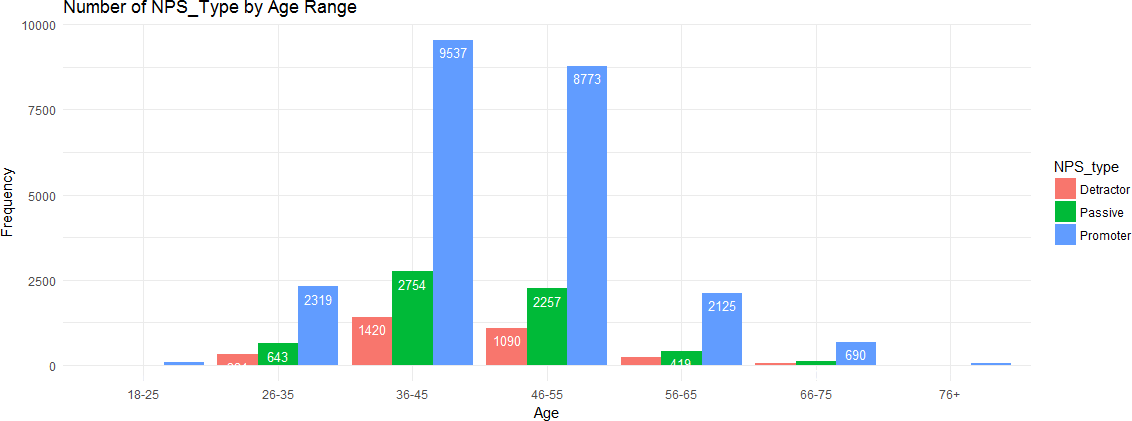


We observed that most of the respondents where between 36 to 55 years old. We were also interested on knowing which group has the best NPS score, and from our analysis we noticed that old people tend to be promoters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Age group** | **Detractors** | **Passives** | **Promoters** | **Total** | **NPS**  **(%Promoters - %Detractors)** |
| **18 – 25** | 11 | 10 | 97 | 118 | .72 |
| **26 – 35** | 321 | 643 | 2,319 | 3,283 | .60 |
| **36 – 45** | 1,420 | 2,754 | 9,537 | 13,711 | .59 |
| **46 – 55** | 1,090 | 2,257 | 8,773 | 12,120 | .63 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **56 – 65** | 224 | 419 | 2,125 | 2,768 | .68 |
| **66 – 75** | 51 | 109 | 690 | 850 | .75 |
| **76 +** | 1 | 13 | 57 | 71 | .78 |

Using this information, we created the following chart.

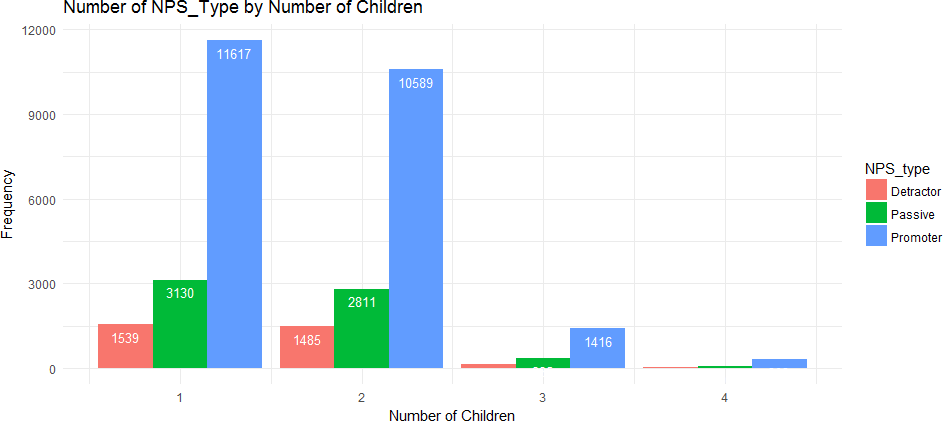


From this chart, we can observe that young people (18-25) and old people (+76) tend to be promoters but they completed less surveys than the other age groups.

Finally, we were interested on knowing if people with children give lowest scores to the hotel. Also, we were interested on knowing if the number of children relates to the NPS type. We noticed that people with children tend to give good scores to the hotel.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Number of children** | **Detractors** | **Passives** | **Promoters** | **Total** | **NPS**  **(%Promoters - %Detractors)** |
| **1** | 1,539 | 3,130 | 11,617 | 16,286 | .61 |
| **2** | 1,485 | 2,811 | 10,589 | 14,885 | .61 |
| **3** | 156 | 335 | 1,416 | 1,907 | .66 |
| **4** | 23 | 70 | 303 | 396 | .70 |

Our team noticed that the NPS score is not so low as we were expecting, but it could be improved



In the following sections, we explained with more detail the answer for our different business questions

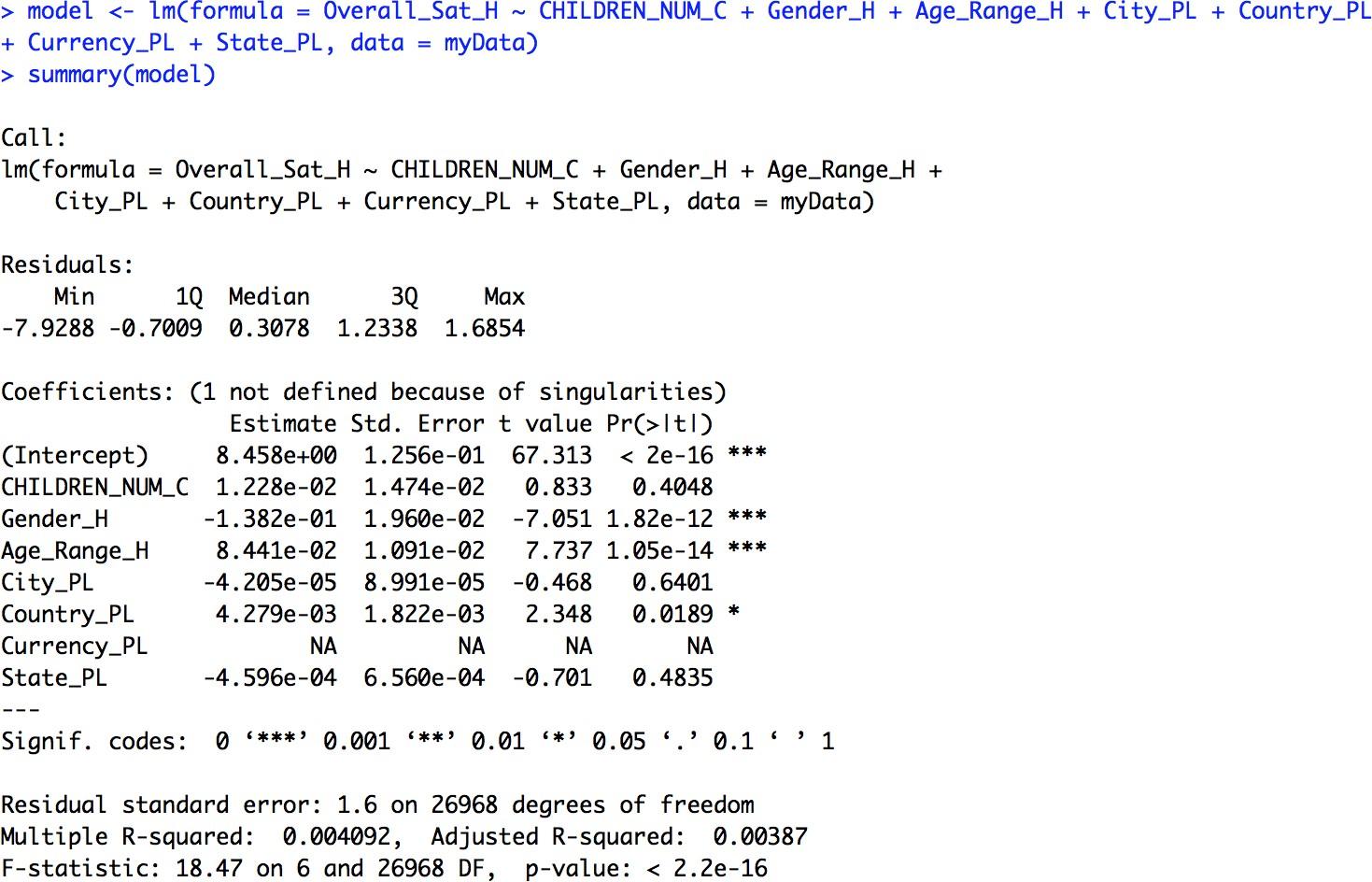
# Use of modeling techniques & Visualizations

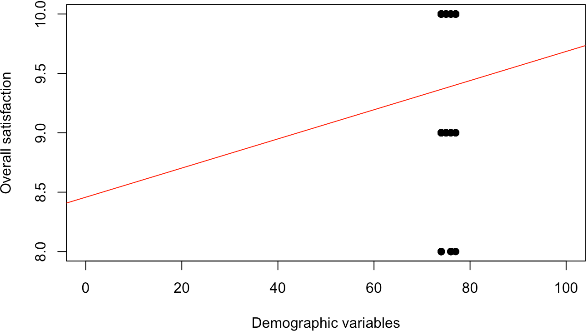
## Linear Models

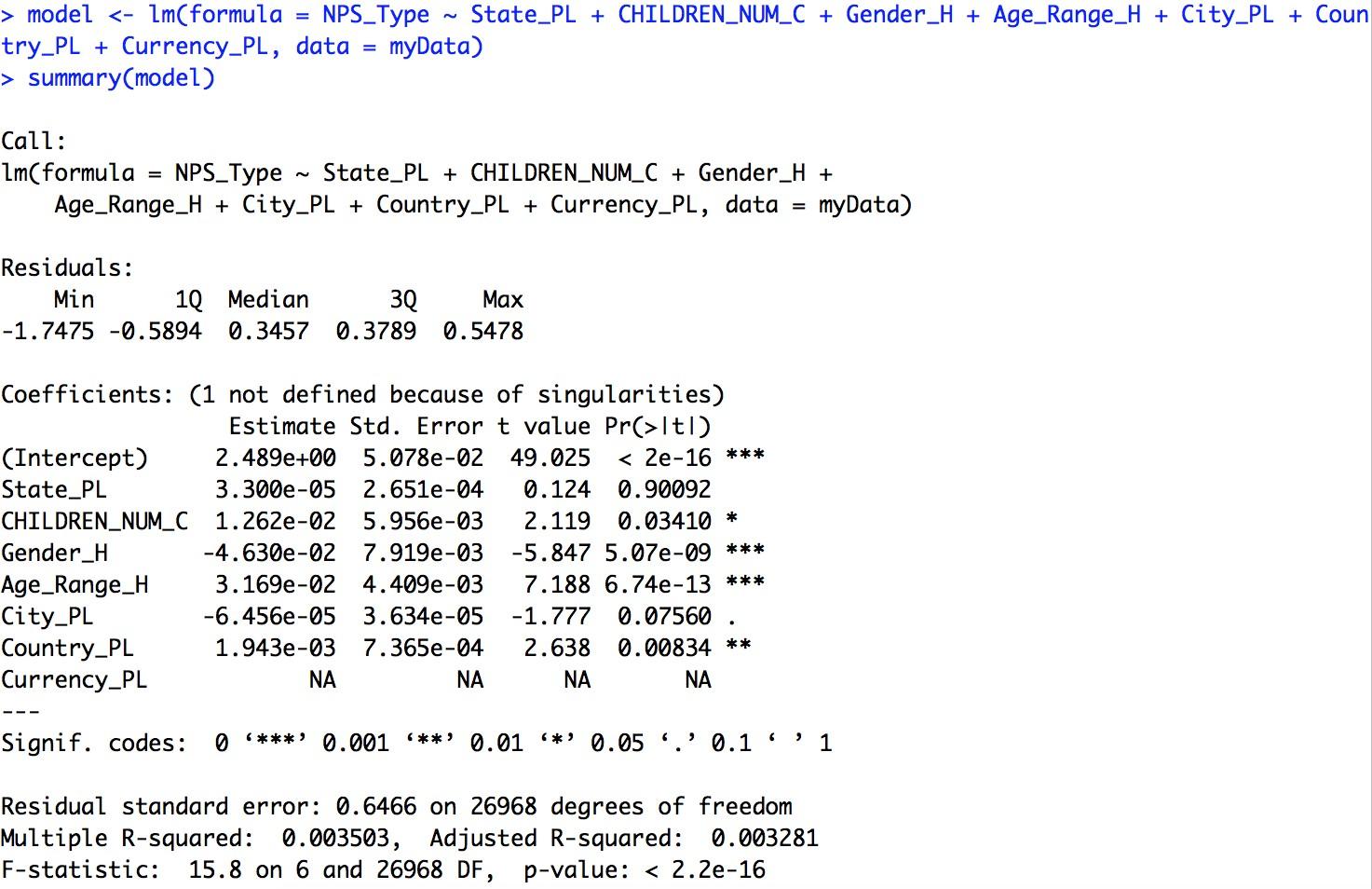
**Business Question 4. Is there any relationship between gender and hotel review? Are females more likely to give bad reviews to the hotel?**

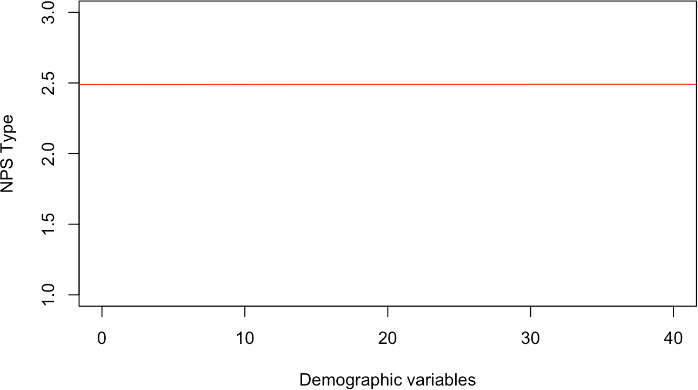
First, we are trying to find the relationship between the demographic and likelihood to recommend. We created a linear model to test their significance and using age, gender, the number of children, hotel city location, state location, country location and local currency for the hotel as independent variables against NPS type.

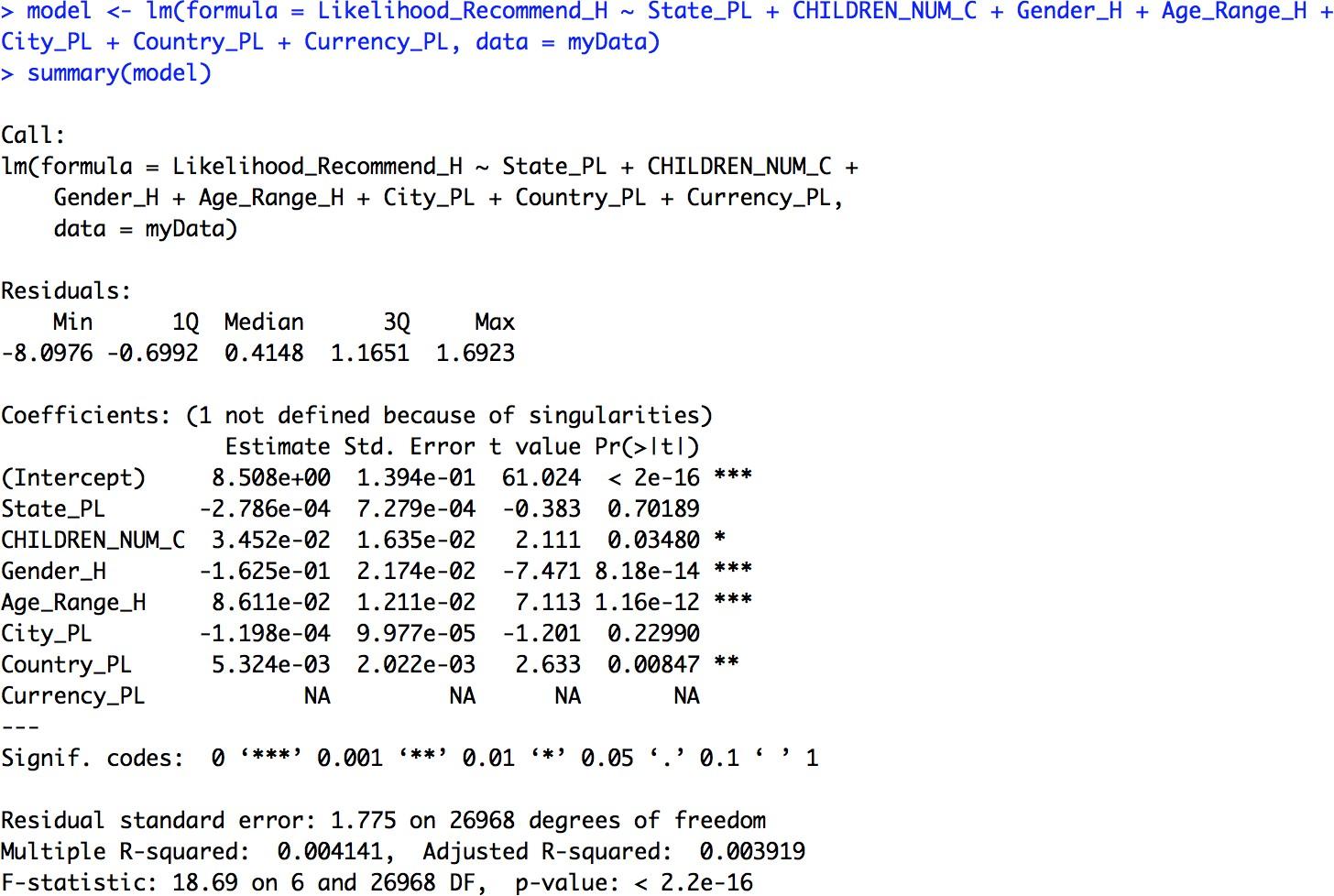
Before we begin the linear model process, we did some data cleaning. Here we ignored the white spaces in the columns of age, state and gender, and we also omit the rows of “Prefer not to answer” in gender column. Then, we converted the columns to numeric types for convenience. After that, we used lm() function to test the linear relationship between demographic characteristics with overall satisfaction, NPS type and likelihood to recommend.

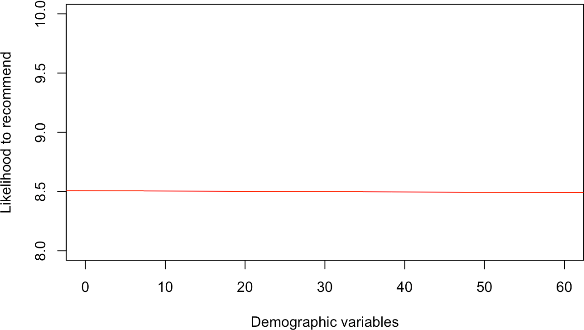




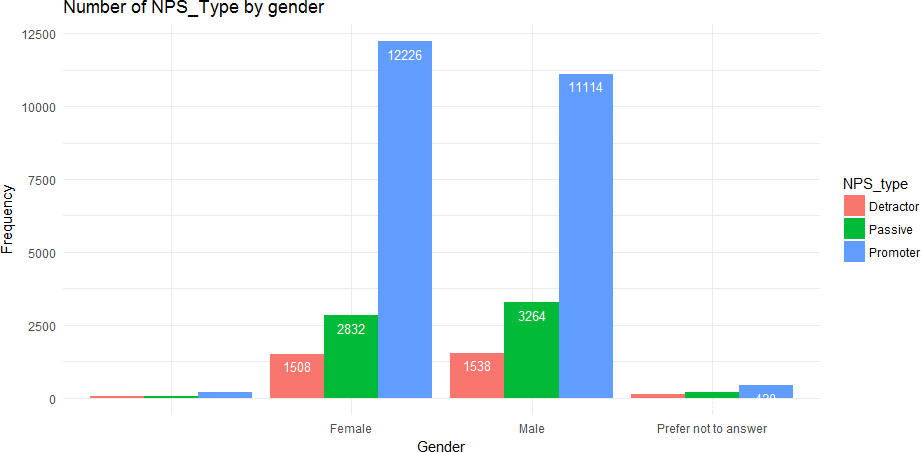






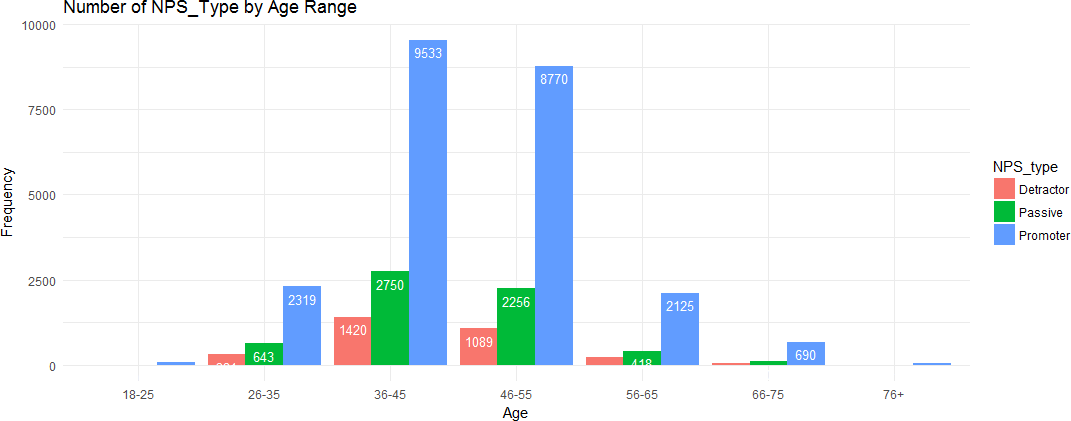


Although through the descriptive analysis, we found that female have more promoter and there are more male detractors and passives. As a result, we believe there is no significant linear relationship between demographic and general customers reviews since the R-square is too low to support the assumption.



**Business Question 5. Is there any relationship between age range and hotel reviews? Which age group gives the lowest rating?**

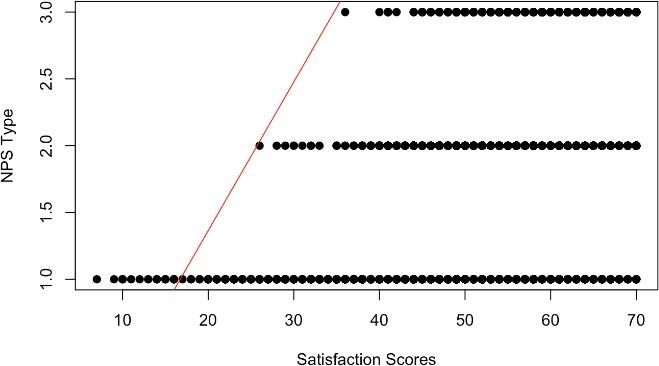
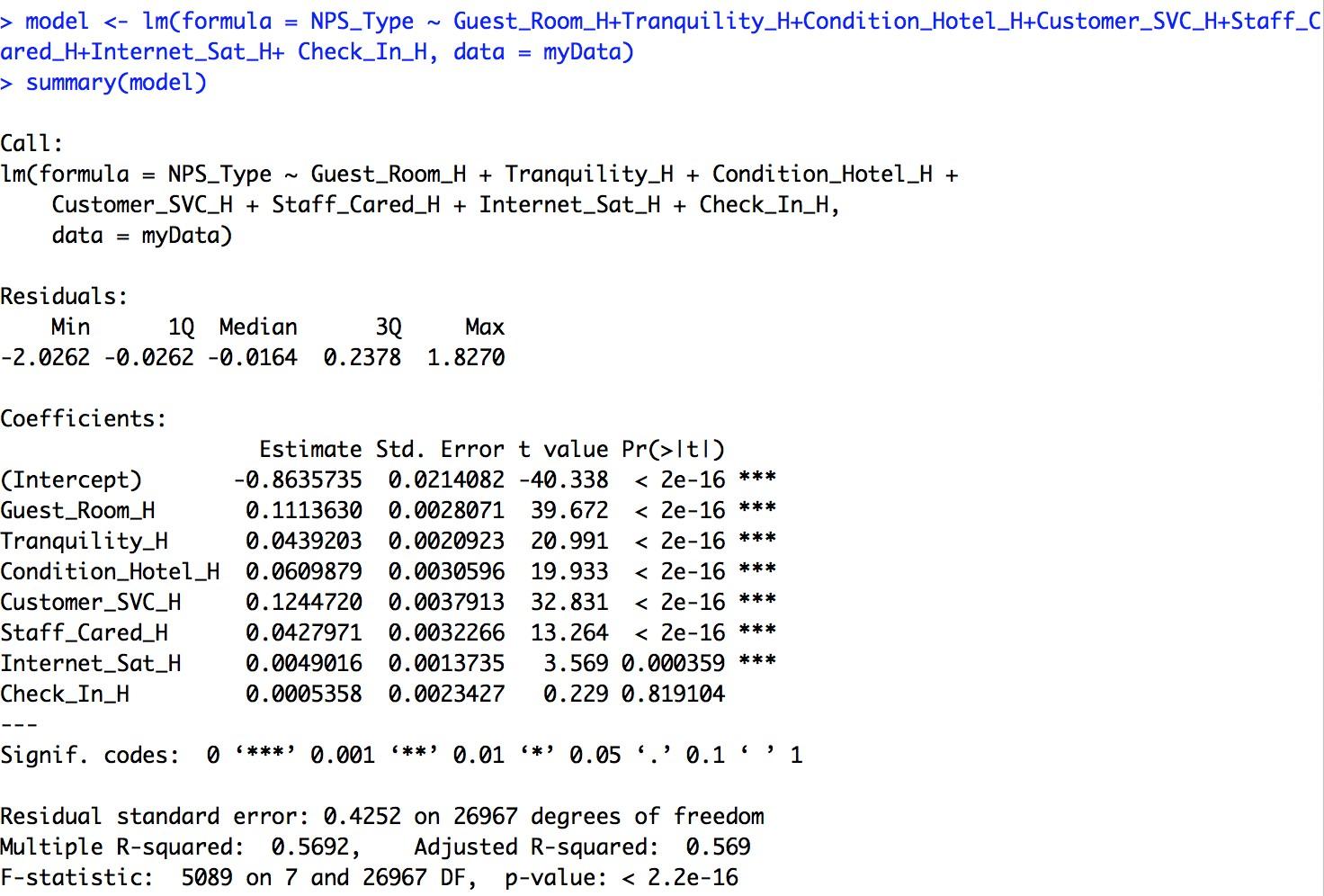
The same result with the last question, we did not find the significant linear relationship between age range and hotel reviews. From the descriptive analysis, we can see most of our survey results are from age range 36-45 and 46-55. Also, we can see the range of age 36-45 has the most promoters. And another interesting finding is most people over 75 give South-East a great rating.



**Business Question 6. Does the existence of certain hotel amenities relate to likelihood to recommend?**

Since the last linear model analysis did not bring us a good result, so we created another model using different satisfaction metrics (Guest\_Room\_H, Tranquility\_H, Condition\_Hotel\_H, Customer\_SVC\_H, Staff\_Cared\_H, Internet\_Sat\_H, Check\_In\_H) as independent variables to determine the relationship with overall satisfaction, NPS type and likelihood to recommend. And here we used lm() and varImp() function to get the importance of variables. The most interesting finding is the results show that the metrics of guest room satisfaction (Guest\_Room\_H), quality of customer service (Customer\_SVC\_H) and tranquility (Tranquility\_H) have the most significant influence on customer overall reviews.

In order to specific understand the customers’ requirements for the hotel's facilities, we created the third linear model using different amenities (Mini.Bar\_PL, Pool.Indoor\_PL, Pool.Outdoor\_PL, Regency.Grand.Club\_PL, Resort\_PL, Restaurant\_PL, Self.Parking\_PL, Shuttle.Service\_PL, Ski\_PL, Spa\_PL, Spa.services.in.fitness.center\_PL, Spa.online.booking\_PL) as independent variables. After we transform all the amenities values (Y,N) to numeric, we still used lm() and varImp() function to get importance of variables.

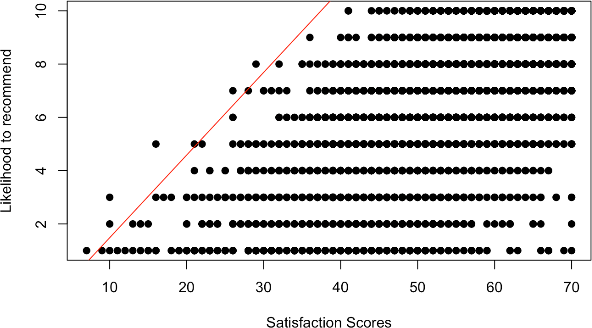


As a result, compared with the importance, we can see the hotel with the onsite restaurants (Restaurant\_PL), has a spa service (Spa\_PL) or near ski space (Ski\_PL) will have a significant impact on likelihood to recommend and NPS type, if the hotel has a mini-bar (Mini.Bar\_PL) will impact the overall satisfaction as well.

**Business question 2a. Does the guest opinion about the hotel services relates to likelihood to recommend?**

Based on our dataset, we focus on the three most representative states which had the most customer survey results and tried to find the relationship between hotel services and likelihood to recommend. After selecting the state, we used the same function to do the linear model analysis. And we found that in overall in these three states the metrics of guest room satisfaction (Guest\_Room\_H), quality of customer service (Customer\_SVC\_H) and the condition of hotel (Condition\_Hotel\_H) will have more influence on NPS type and likelihood to recommend. Also, overall satisfaction would be impacted by metrics of guest room satisfaction (Guest\_Room\_H), quality of customer service (Customer\_SVC\_H) and staff cared metrics (Staff\_Cared\_H).

Moreover, we made the fifth linear model analysis base on the dataset of three states and using amenities to be the independent variable. We found that whether the hotel has the resort (Resort\_PL), restaurants (Restaurant\_PL), has a spa service (Spa\_PL) or near ski space (Ski\_PL), will have an effect on NPS type and likelihood to recommend. Besides, if the hotel has shuttle service (Shuttle.Service\_PL) is another factor to influent overall satisfaction. And more interesting is that we found customers in Texas prefer to have a mini bar and outdoors pool.



## Association rules

Association rules are used to run the apriori algorithm. It makes it easier to find a pattern or the rules between the variables. If we talk in the retail terms the rule will specify what item is frequently bought with what item. It has a LHS and RHS part.

To answer our questions, we made subset of all the amenities, demographic information and the

states of the USA to see how they are associated with the “NPS\_Type” (Promoter, Detractor and Passive).

### Amenities data

**Business question 2b. Is it possible to associate NPS\_TYPE with certain Amenities?**

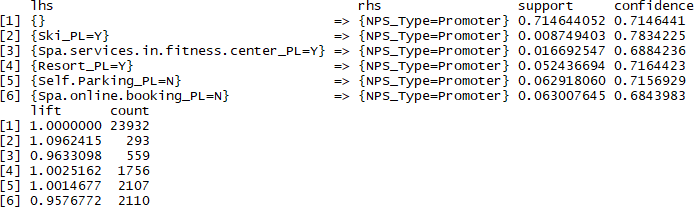
Starting with the amenities provided by the hotel like Resort\_PL, Regency.Grand.Club\_PL, Pool.Outdoor\_PL, Self.Parking\_PL, Shuttle.Service\_PL, Ski\_PL, "Spa\_PL", Spa.online.booking\_PL, Spa.services.in.fitness.center\_PL, Pool.Indoor\_PL, Mini.Bar\_PL, Restaurant\_PL

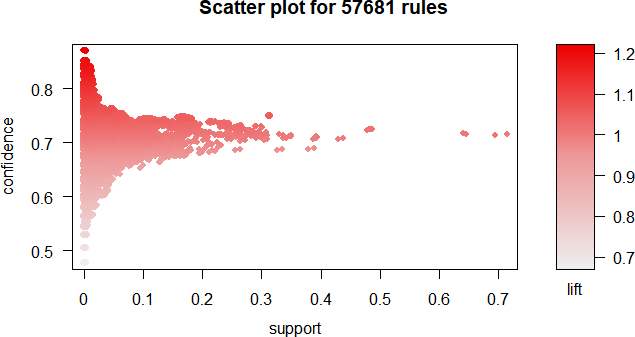
From this we can interpret which services affect the NPS type. Here It specifies for each variable whether it makes a promoter.

* **Promoters:**

We can see if there is a Ski service, it could be associated with promoters. The confidence of this rule is 0.78 which is very high compared to others.

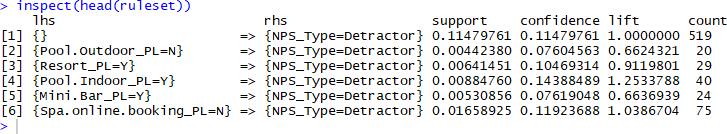
###### inspect(head(ruleset))

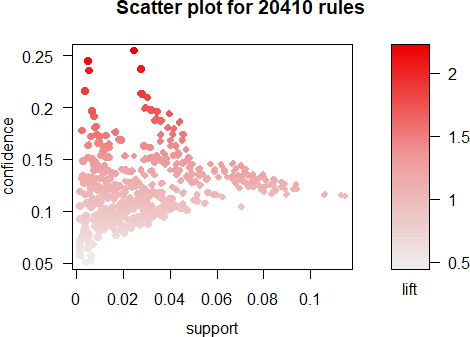




* **Detractors:**

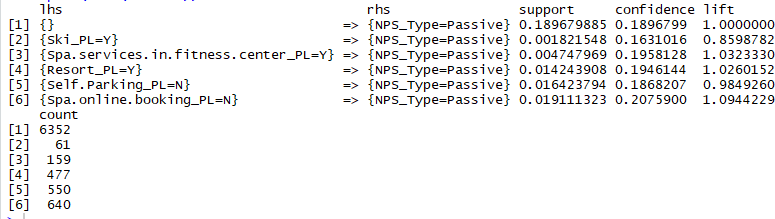
In these rule set, we see that when there is no pool out door, the guest tends to be a detractor (the confidence and support is really low, so we did not include this rule in our final presentation).

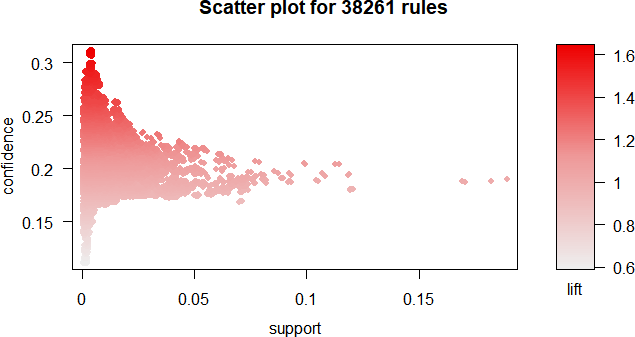




* **Passive:**

For passive NPS the support and confidence is low for almost all the parameters.





### Demographic Information

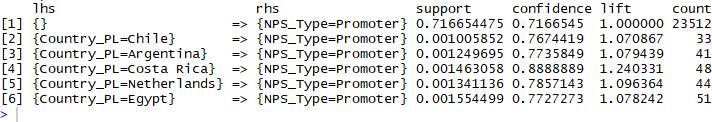
**How is NPS type associated with demographics?**

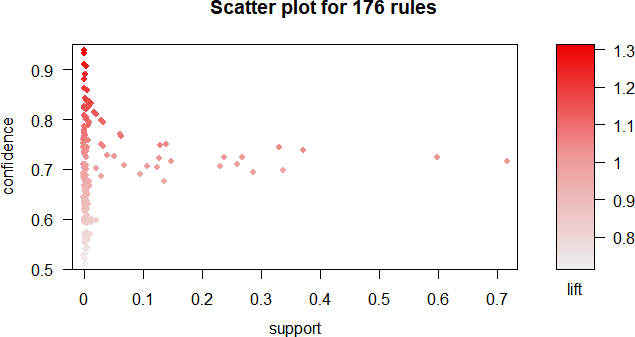
Next we made a subset of the demographic information of the customers such as Gender\_H, Age\_Range\_H ,Country\_PL and see how these variables are associated with NPS type.

Checking for each NPS\_Type individually:

* **Promoter:**

Here from the first result we can see people who don’t provide their country they are more promoters. Also, the other results where the confidence is very high for countries like Costa Rica and Netherlands, but the support is very low.

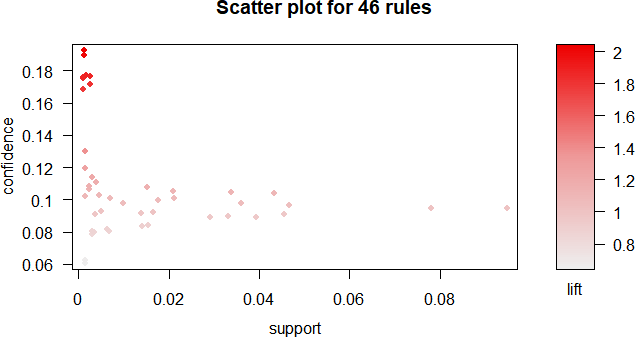
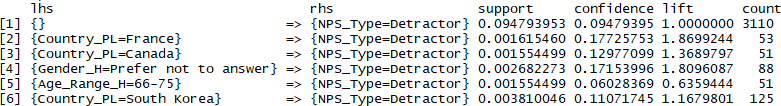




* **Detractor:**

For detractors the confidence and support both are low. So, we can say that the country

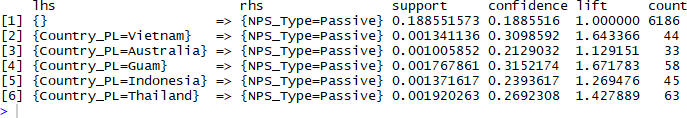
doesn’t associate the detractors.

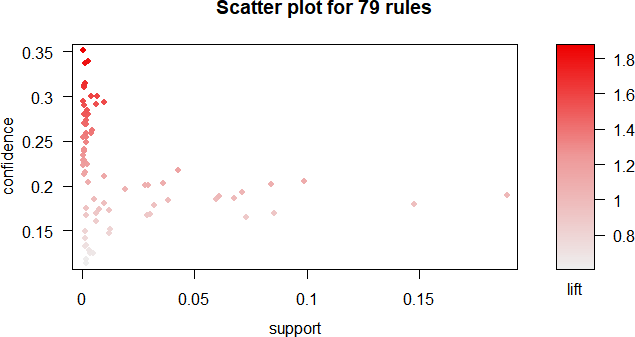


* **Passive:**

For country like Guam and Vietnam there is a confidence of 0.3 higher than the others but a very low

support. It doesn’t help in analysis for demographics with the NPS\_TYPE.





### For USA states (Florida, California and Texas)

We did all the above analysis for specific states. We did it for California, Florida and Texas. The reason we selected these three states is that because California is the state where there are more number of surveys completed followed by Florida and Texas respectively.

We are first checking how in California the amenities associate with the customer to be a promoter, detractor or a passive. If there are SPA services, in the hotels in California they tend to be promoters. The confidence was 0.743. Also, when there is a resort it has a good confidence as well as support which tells us that when there is a resort the customers are promoters. The Spa online booking Service has both confidence and support high as compared to other. So, when there is no spa online booking service, the customer is generally a detractor. Also, when there is a no pool outdoors, we have high support and confident associating with passive NPS TYPE.

The next state with more number of completed survey is Florida. The high support and confidence is for Spa services and fitness center. It shows that if there is a Spa and fitness services Customer tends to be a Promoter. Also, when there are mini bars in the room it tends to be a promoter (it has good confidence value). The next value which has a good support and confidence is when there Is spa online booking. For passive we can analyze that the support is very low for all the amenities. We can analyze from the result that when there is no Spa online booking Service it is a passive and it has better confidence and support as compared to other variables.

For the hotels in Texas, we analyzed that if there is a mini bar they tend to be promoters. Also, if there is a pool indoor there is a good confidence and support value, which shows that the customers are promoter if there is a Pool indoor.

Classification Models

**Business Question 3a. Is it possible to predict the NPS\_Type based on Hotel amenities / guest opinion scores?**

We used SVM model to classify the promoter and detractor and tried to predict the NPS types based on Hotel amenities and guest opinion scores. Here, we just omit the passive and just do the prediction for promoter and detractor and still chose the California, Florida and Texas.

First, we used amenities as input variables for our classifier, we can predict (NPS\_TYPE) if a guest is a promoter or detractor with an accuracy of 86.4% in California, 90.6% Florida and 87.3% in Texas.

Logistic regression

happyCust <- 1

med <- median(clean\_data$Satisfaction)

happyCust[clean\_data$Satisfaction>=med] <- 3

happyCust[clean\_data$Satisfaction<med] <- 2

happyCust <- as.factor(happyCust)

clean\_data1 <- cbind(clean\_data,happyCust)

clean\_data1 <- clean\_data1[,-c(1)]

clean\_data1$Departure.Delay.in.Minutes[is.na(clean\_data1$Departure.Delay.in.Minutes)] <- mean(clean\_data1$Departure.Delay.in.Minutes,na.rm=T)

clean\_data1$Arrival.Delay.in.Minutes[is.na(clean\_data1$Arrival.Delay.in.Minutes)] <- mean(clean\_data1$Arrival.Delay.in.Minutes,na.rm=T)

clean\_data1$Flight.time.in.minutes[is.na(clean\_data1$Flight.time.in.minutes)] <- mean(clean\_data1$Flight.time.in.minutes,na.rm=T)

# origin city, destination, airline code, flights with other airlines,arrival and departure delay and day of month don't affect happiness significantly(high p values)

model3 <- glm(happyCust~Airline.Status+Age+Gender+Price.Sensitivity+No.of.Flights.p.a.+ Type.of.Travel+Shopping.Amount.at.Airport+Class+Arrival.Delay.greater.5.Mins+Eating.and.Drinking.at.Airport+southeast,family="binomial",data=clean\_data1)

summary(model3)

# # evaluation of logistic regression model

logtraindata <- clean\_data1

logtestdata <- test

happyCust <- 1

med <- median(logtestdata$Satisfaction)

happyCust[logtestdata$Satisfaction>=med] <- 3

happyCust[logtestdata$Satisfaction<med] <- 2

happyCust <- as.factor(happyCust)

logtestdata$Departure.Delay.in.Minutes[is.na(logtestdata$Departure.Delay.in.Minutes)] <- mean(logtestdata$Departure.Delay.in.Minutes,na.rm=T)

logtestdata$Arrival.Delay.in.Minutes[is.na(logtestdata$Arrival.Delay.in.Minutes)] <- mean(logtestdata$Arrival.Delay.in.Minutes,na.rm=T)

logtestdata$Flight.time.in.minutes[is.na(logtestdata$Flight.time.in.minutes)] <- mean(logtestdata$Flight.time.in.minutes,na.rm=T)

logpred <- predict(model3,logtestdata,type="response")

pos\_or\_neg <- ifelse(logpred > 0.5, 3, 2)

happyCusttest <- factor(pos\_or\_neg)

x <- table(happyCust,happyCusttest)

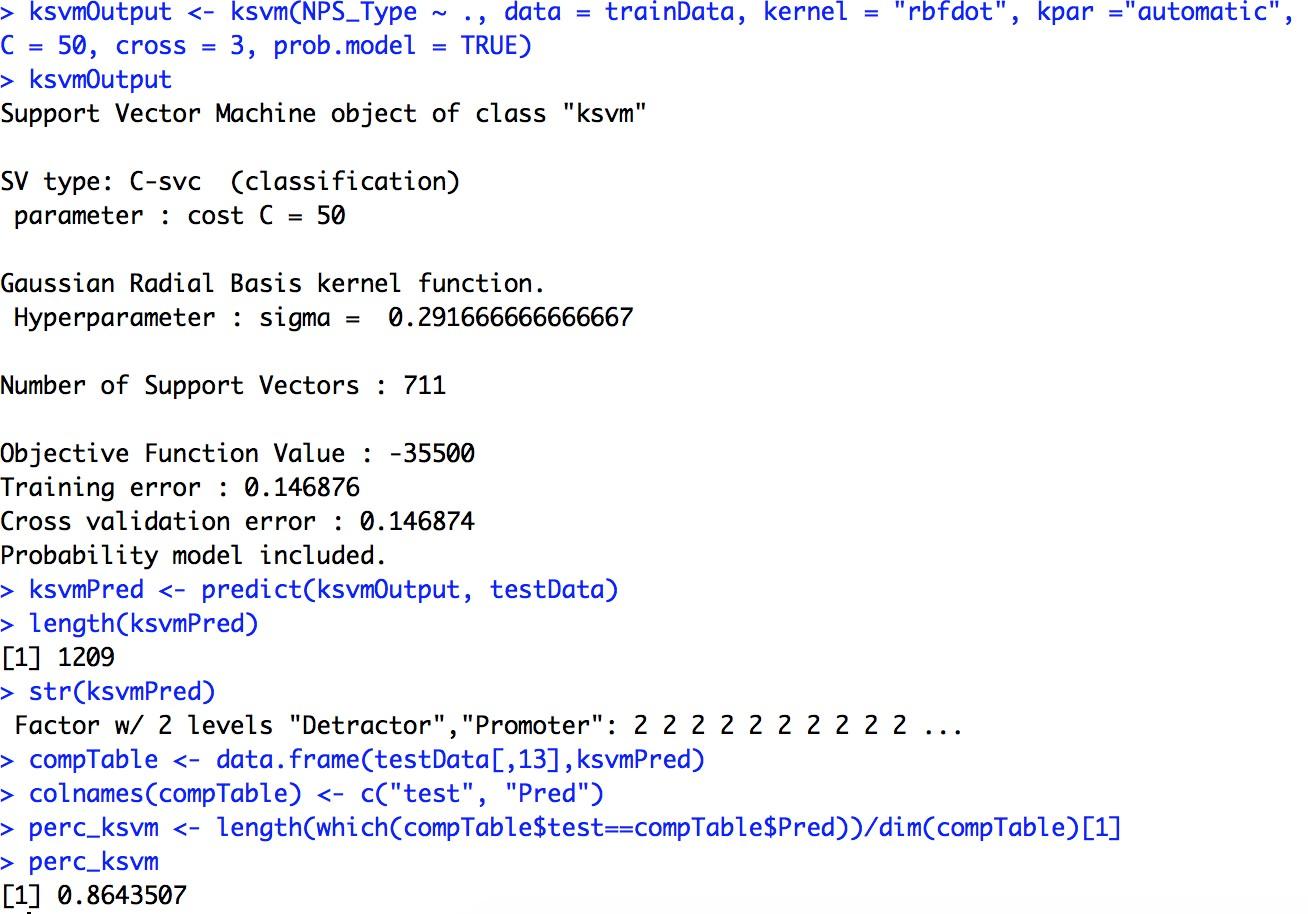
error <- (x[1,1]+x[2,2])/sum(x)

error

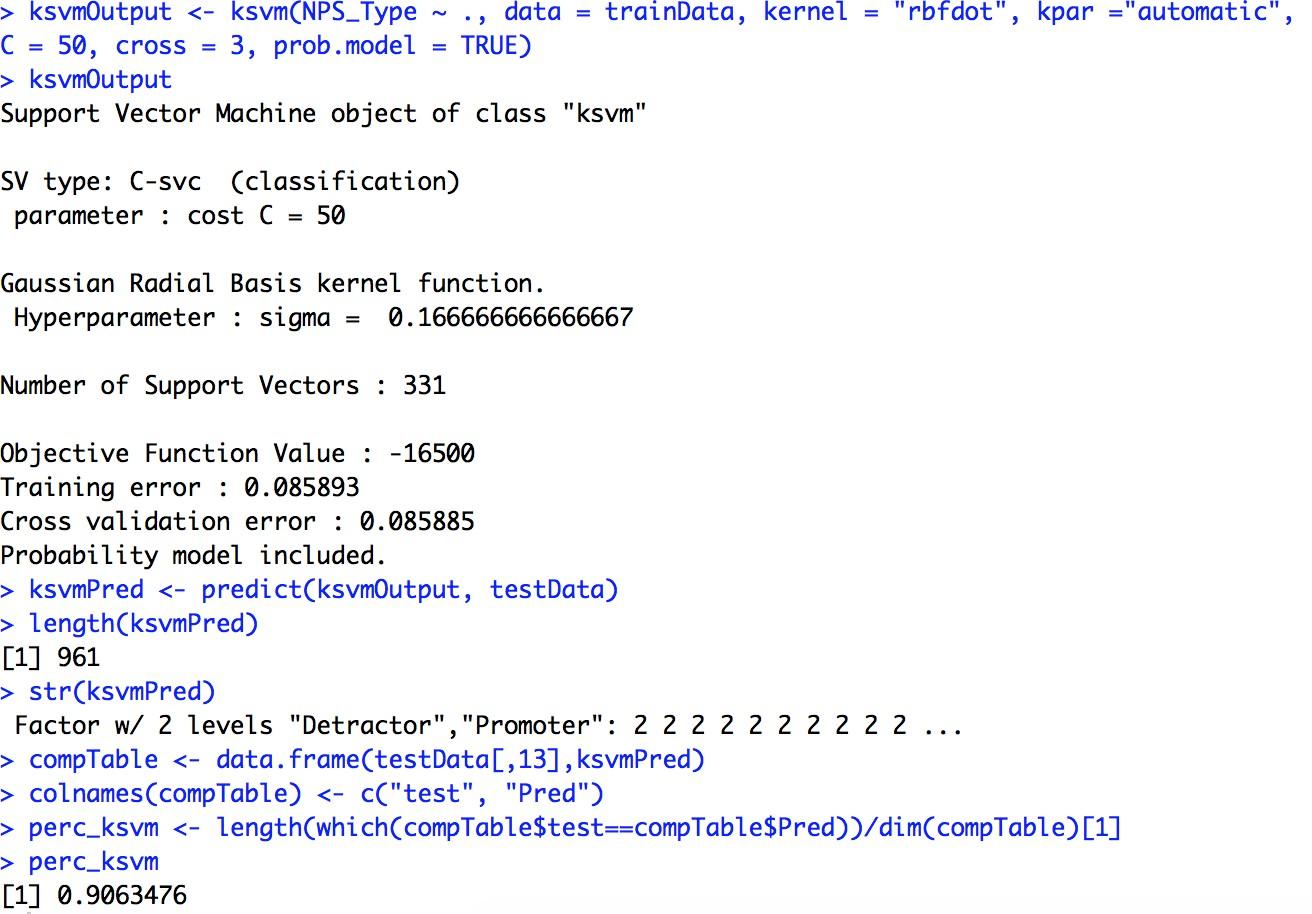
# 77% accuracy

## Support vector machine

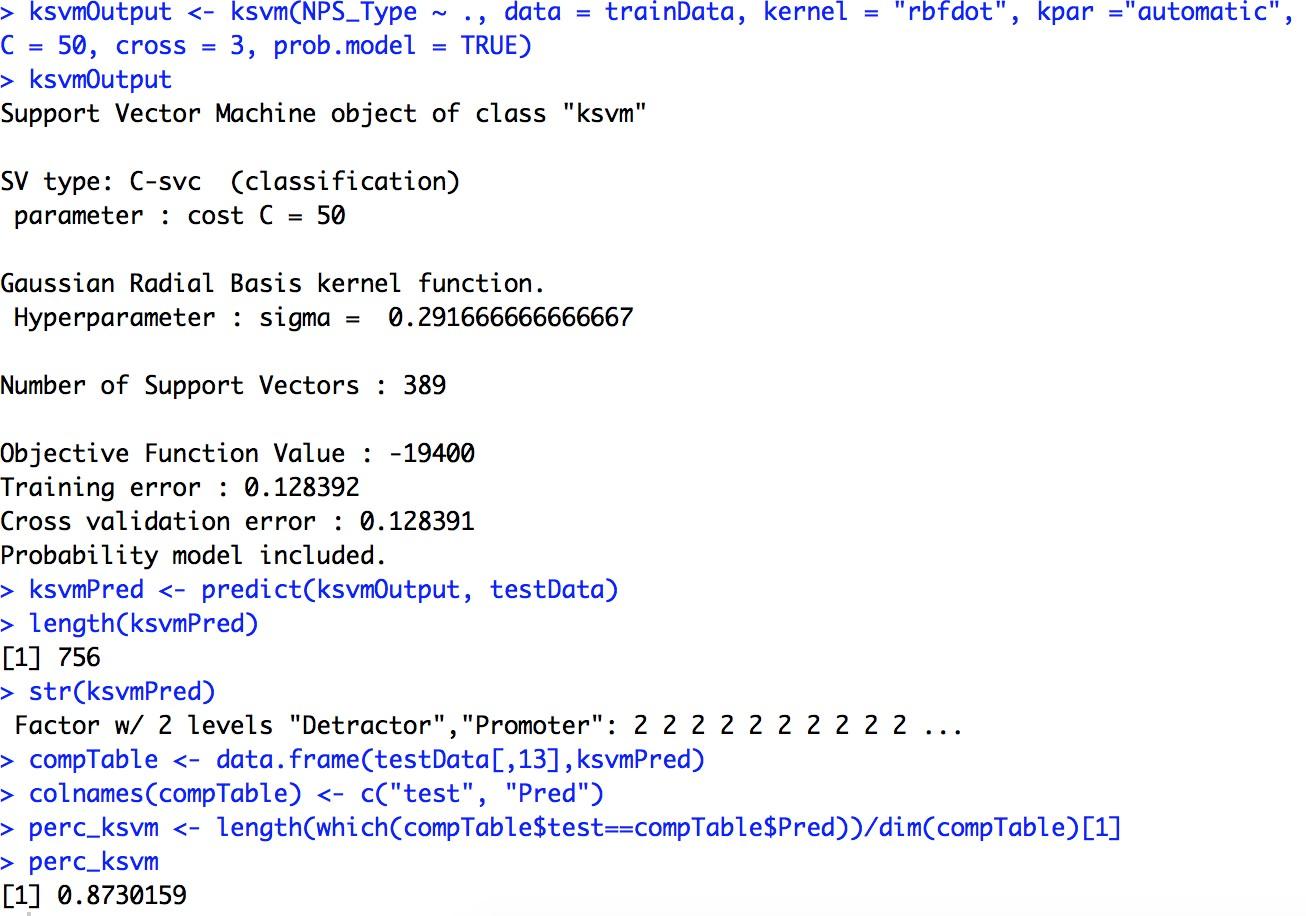
#California



#Florida

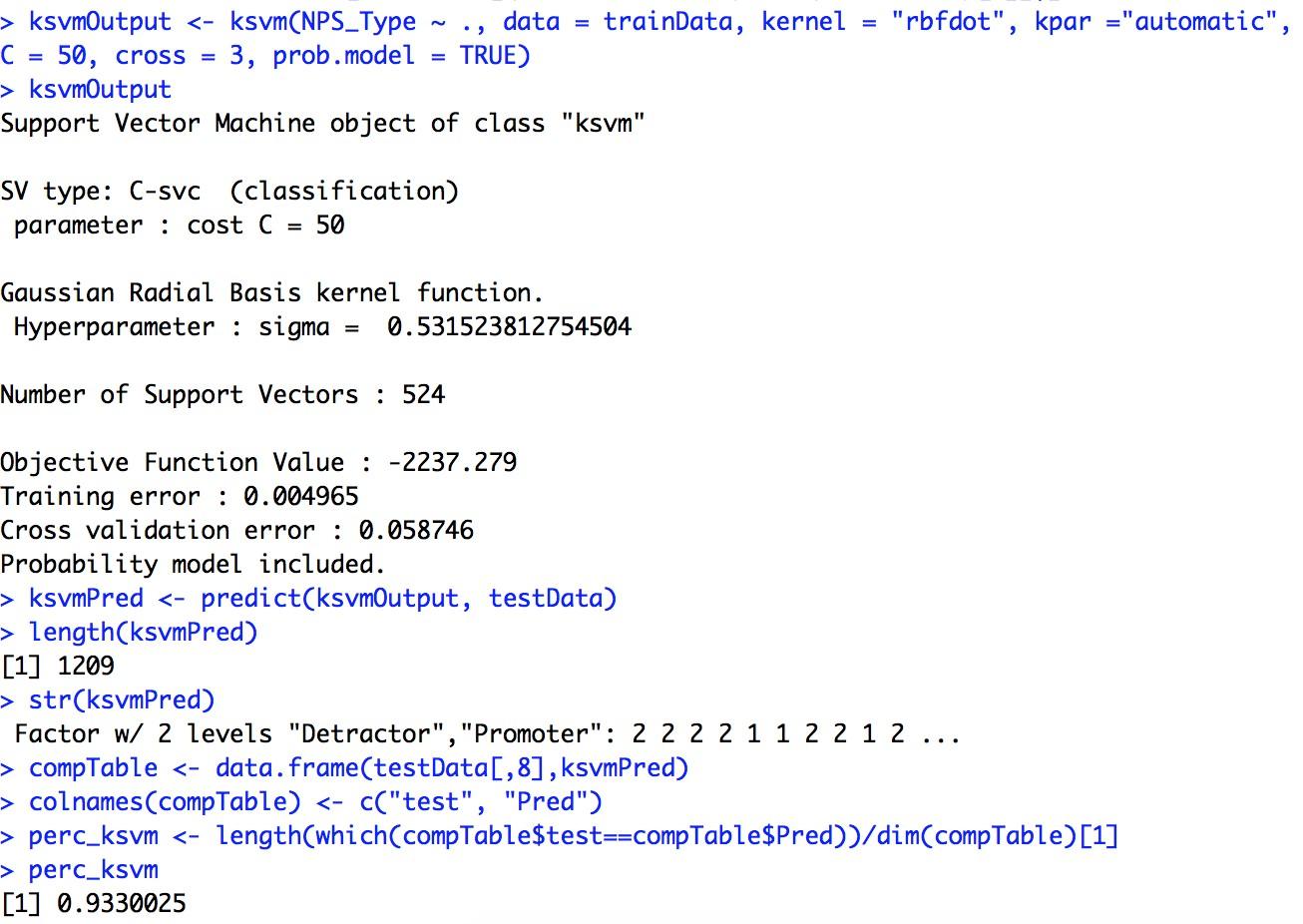


#Texas

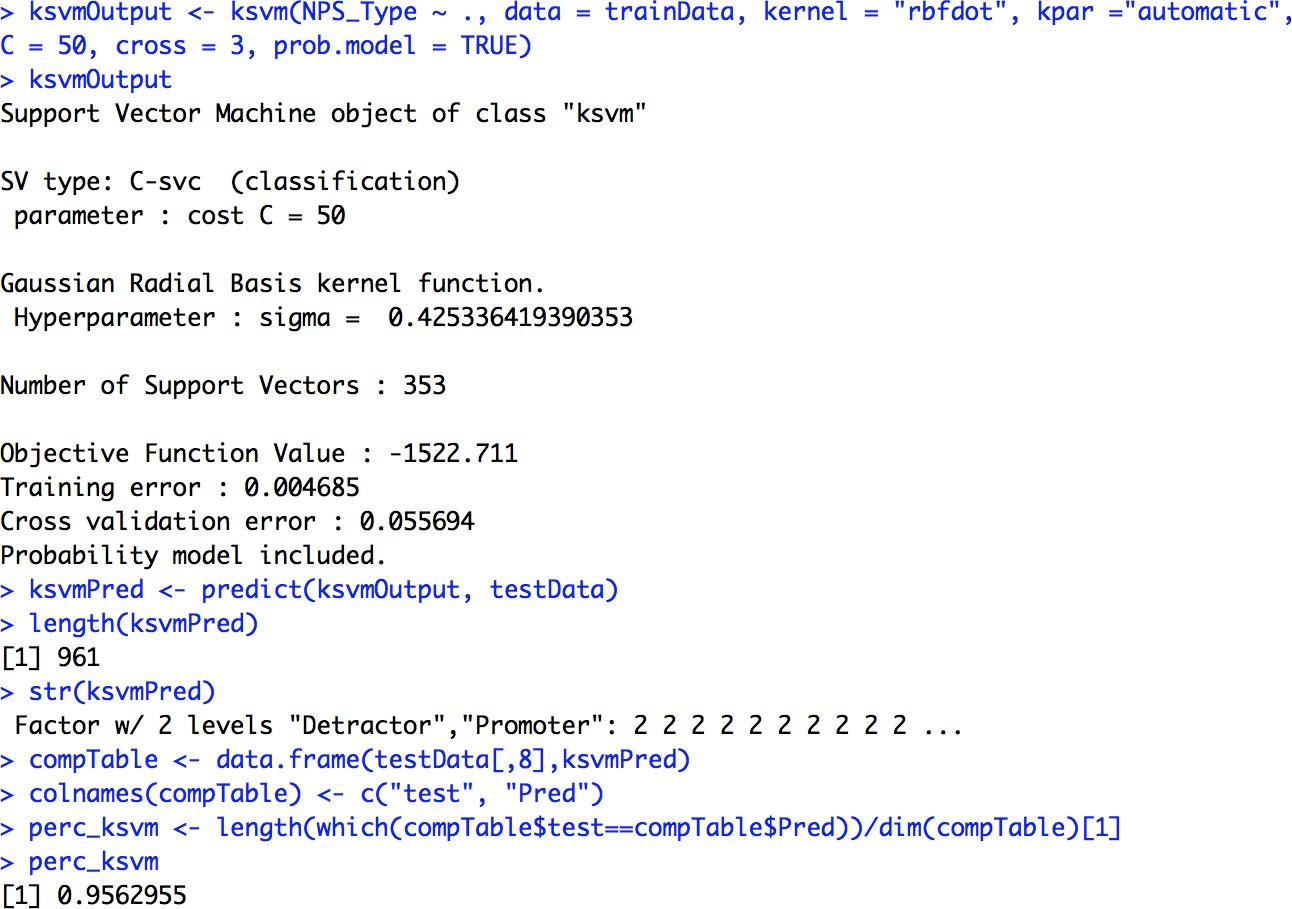


Second, we used services opinion as variables for our classifier. From the result, we know that with an accuracy of 93.3% a guest is a promoter or detractor in California, 95.6% in Florida and 94.3% in Texas.

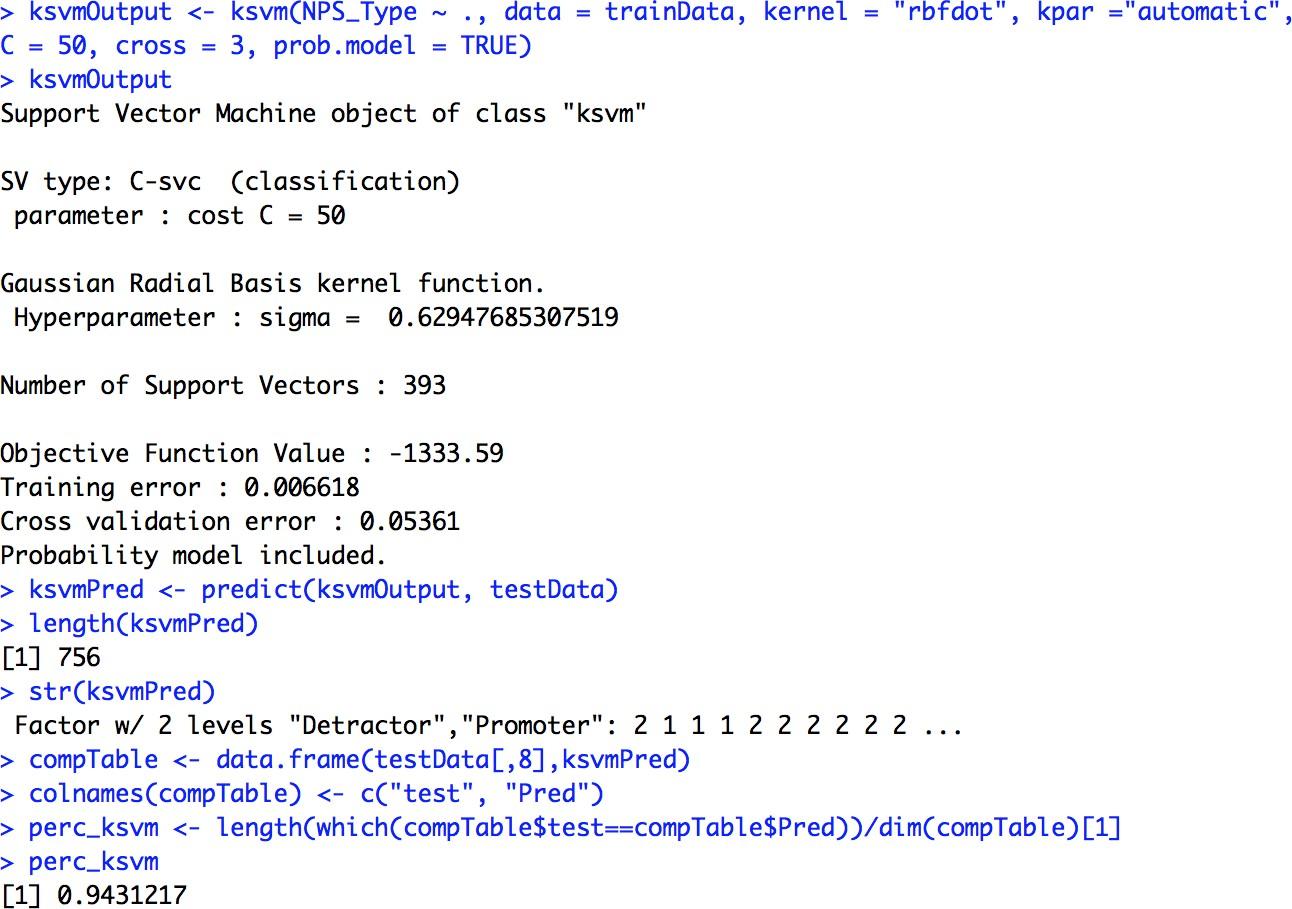
# California



# Florida



#Texas

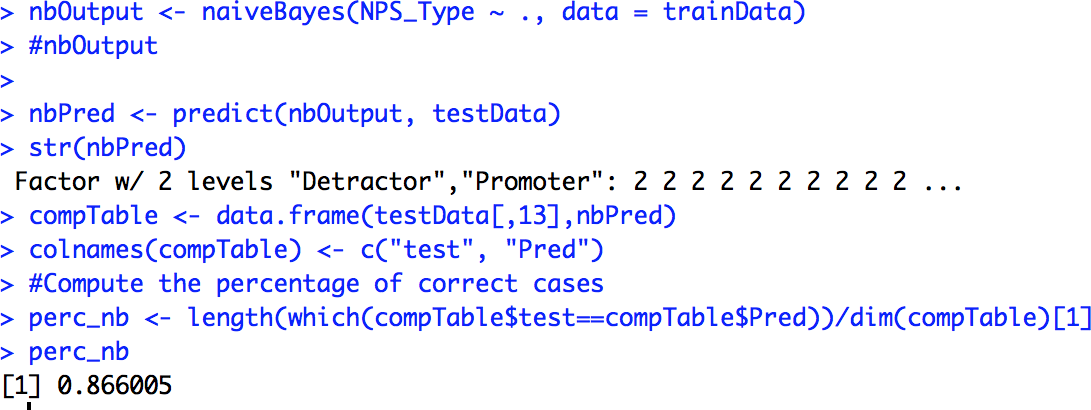


## Naïve Bayes

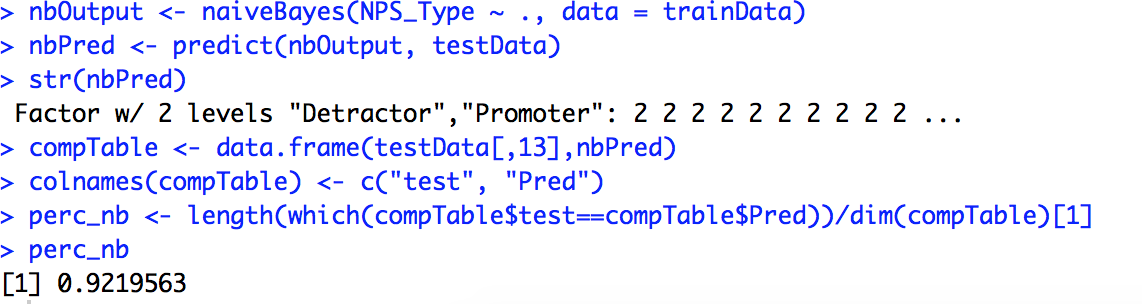
And then we used the 'naive bayes' function to do the prediction.

First, we used amenities as variables for our classifier, we can predict (NPS\_TYPE) if a guest is a promoter or detractor with an accuracy of 86.6% in California, 90.6% Florida and 87.8% in Texas.

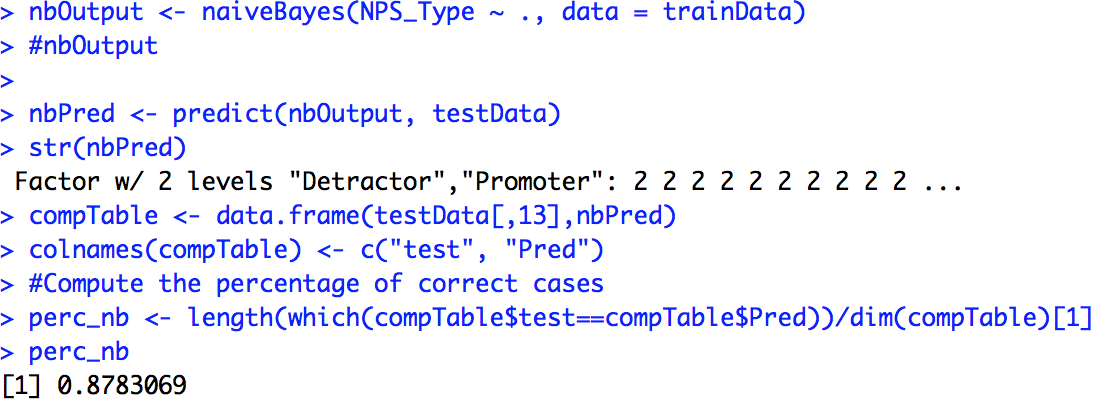
#California



#Florida

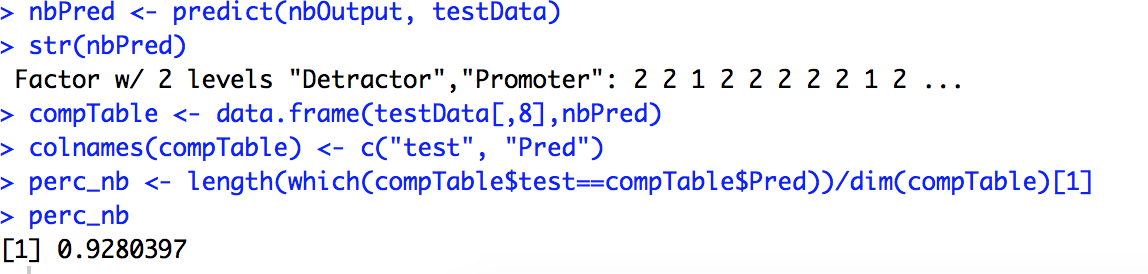


#Texas

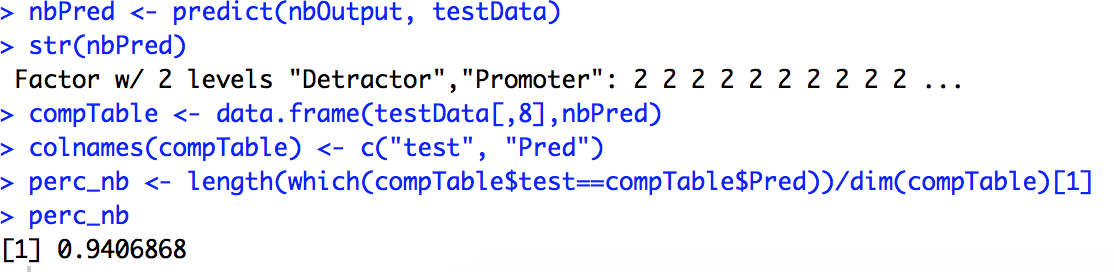


Second, we used services opinion as variables for our classifier. From the result, we know that with an accuracy of 92.8% a guest is a promoter or detractor in California, 94.1% in Florida and 94.3% in Texas.

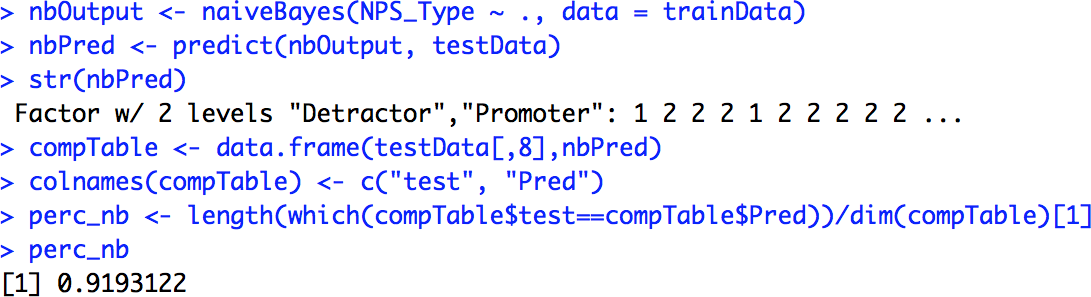
#California



#Florida



#Texas



Actionable Insights / Overall interpretation of results

Interpretation of Graph 1: The descriptive graphs on world map suggest that USA has a maximum number of promoters across the world and this is showcased using blue spots. The yellow spots show the passive guests across the world and the detractors are shown with red spots. This suggests that South-East Hotels has obtained more successfully completed surveys from the USA than any other country in the world.

(Please refer to the presentation for the number of the graphs)

### BQ 1: How is the Overall performance of Hotels in USA tomorrow?

Interpretation of Graph 2: Because of the concentration of survey participants in the USA, we are finding the performance of the hotels in the USA using pie chart. Overall, a little less than 75% of people in the USA are promoters, but the performance of hotels in Florida is greater than the national average. Texas and California are a little less than the national average. The reason why we drilled down to these 3 states is because there are more survey participants from these states.

### BQ 2: Is there any relationship between nationality and hotel reviews? Are US customers more likely to give bad reviews to the hotel?

Interpretation of Graph 3: The purpose of this graph is to answer our business question: Is there any relationship between nationality and hotel reviews? Are US customers more likely to give bad reviews to the hotel?

The graph suggests that there are more promoters, detractors and passive responders in USA, which basically is suggestive of more survey participants in the USA than anywhere else in the world. But the graph also suggests that even though the average NPS Score in USA is 0.62, there are countries where the NPS Average is worse.

### BQ3: 3. Customers with children tend to be detractors?

Interpretation of Graph 4: To answer our business question on customer behavior based on the number of children that they have, we found out that there is no significant difference between people with 1 child and people with 2 or 3 children with regards to their NPS behavior. The graph eloquently shows this insight. So, the answer to this question is also no.

### BQ 4: Is there any relationship between gender and hotel review? Are females more likely to give bad reviews to the hotel?

Interpretation of Graph 5: This graph clearly answers the 5th question. There is only a marginal difference between average NPS rates of females and males. So gender doesn’t play a role in predicting whether the person could be a promoter or a detractor.

Interpretation of Graph 6: To support the answer, we tried to distribute the female and male respondents on the world map and on a pie chart. The odds of a female being a detractor are as good as the odds of a male being a detractor. So again, the gender of a guest is insignificant.

### BQ 5: Is there any relationship between age range and hotel reviews? Which age group gives the lowest ratings?

Interpretation of Graph 7: The average NPS rating of the people between ages 26 and 55 revolves around 0.60 with a standard deviation of 0.2. Young people between ages 18 and 25 give 0.72 average NPS and older people (baby boomers?) are the most frequent promoters

### BQ 6: 6. Does the existence of certain hotel amenities relates to likelihood to recommend? (Linear Models)

Interpretation of Linear models: From the linear model, using amenities as independent variables and the likelihood to recommend as dependent variable, we were able to observe that restaurants have a high impact in the likelihood to recommend. We used the function variable importance (varImp) to support our conclusion.

Also, from the linear models, we can observe that:

* The existence of restaurants, spa service and resort place in the hotel relates to high scores of likelihood to recommend
* The existence of restaurants, Grand club and a pool indoor in the hotel relates to high overall satisfaction Regency

### BQ 7: Is it possible to associate NPS\_Type with certain amenities? (association rules)

By implementing association rules we are able to answer the above business question.

* If there is Resort then the guest could be a PROMOTER. (21,713 observations where the existence of Resort associates with Promoters)
* If there is not SPA nor SPA booking online service, then the guest could be a DETRACTOR
* The map graph shows hotels with no resorts in California, Florida and Texas. Florida’s overall NPS Type is better than the national average and in Central Florida respondents’ surveys show that this area has less hotels with resorts.

### BQ 8: Does the guest opinion about the hotel services relates to likelihood to recommend? How does it affect the overall satisfaction? (linear models)

* Guest room satisfaction score and Customer SVC satisfaction score directly relate to Overall satisfaction and likelihood to recommend. Customer opinion matters a lot.

Our analysis determines these as some of the most important amenities:

* 1. Guest room satisfaction score
  2. Tranquility satisfaction score
  3. Condition of the Hotel satisfaction score
  4. Customer SVC satisfaction score
  5. Staff Cared Satisfaction Internet satisfaction score
  6. Check\_In satisfaction score

### BQ9 Is it possible to predict the NPS\_Type based on Hotel amenities / guest opinion scores? (Support vector machines)

* Using support vector machine (svm), we can predict NPS type on the basis of Guest opinion Scores with an accuracy of 95%
* Using naïve bayes classifier, we can predict NPS type using amenities as input variables with 86% of accuracy

# Conclusion

From the detailed analysis of the data set from South-East Hotels, we have got a handful of actionable insights with regards to amenities, geography, important states of the business, ages range, amenities among other variables. After answering our business questions, we are able to come up with suggestions to South-East Hotels and we believe that considering these suggestions can help improve services of South-East Hotels and can improve revenues and profits.

Here are some of our suggestions:

* We suggest increasing the number of restaurants on South-East Properties. Availability of these often impacts likelihood to recommend
* Incorporating spas in the hotel properties may also help because for a majority of current hotels, the existence of spa directly impacts the likelihood to recommend
* South-East Hotel should improve customer service to increase overall satisfaction. Our analysis has

suggested that the customer’s overall satisfaction is impacted by customer services

* Providing extra training to staff members, maids, maintenance workers will help because these scores have a high effect on overall satisfaction
* Focus on improving amenities and services, demographic characteristics don’t necessarily have a

lot to do with overall experience

Appendix - Code

Understand Data

raw\_data<-data

#str(raw\_data)

#head(raw\_data)

# This data has 28 variables and 129889 observations

# Satisfaction is the dependent variable and all the other columns are independent variables

# The client is Southeast hence this project will have comparisons between Southeast

# and other providers

# The date column will be converted into month and day of week to determine if time of

# year and time of week have any significant effect on the Satisfaction

# find unique values in each column to check for abnormal values

sapply(raw\_data,function(y)unique(y))

## Descriptive analysis

library("ggplot2") library(ggmap) library(maps) library(mapdata)

#Descriptive Analysis tmp <- file.choose()

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

table(myData$Gender\_H, myData$NPS\_Type)

#Gender pie chart

gender <- as.data.frame(table(myData$Gender\_H)) colnames(gender) <- c("Gender", "Freq")

gender <- gender[gender$Gender != "",]

#gender <- gender[gender$Var1 != "Prefer not to answer",] bp<- ggplot(gender, aes(x="", y=Freq, fill=Gender))+

geom\_bar(width = 1, stat = "identity") bp

pie <- bp + coord\_polar("y", start=0) + ggtitle("Gender Distribution") +# for the main title

scale\_fill\_manual(values=c("hotpink1", "skyblue", "olivedrab1")) pie

#Gender map

Gender\_maps <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL","Gender\_H")]

#separating points by NPS\_TYPE unique(Gender\_maps$Gender\_H)

#Female

Female <- Gender\_maps[Gender\_maps$Gender\_H == "Female",] #Male

Male <- Gender\_maps[Gender\_maps$Gender\_H == "Male",]

world <- map\_data("world") South-East Map <- 0

South-East Map <- ggplot() + geom\_polygon(data = world, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

FemaleMap <-South-East Map + geom\_point(aes(x = Female$Property.Longitude\_PL, y = Female$Property.Latitude\_PL), color = "hotpink1", size = 1) + ggtitle("Females respondents")

MaleMap <-South-East Map + geom\_point(aes(x = Male$Property.Longitude\_PL, y = Male$Property.Latitude\_PL), color = "skyblue", size = 1) + ggtitle("Male respondents")

#NPS VS Gender myData$count <- 1

agg.data <- aggregate(myData$count, by = list(gender = myData$Gender\_H, NPS\_type = myData$NPS\_Type), FUN = sum)

p1 <- ggplot(agg.data, aes(x=gender, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by gender") + # for the main title xlab("Gender") + # for the x axis label

ylab("Frequency") # for the y axis label

#Another plot

p2 <- ggplot(agg.data, aes(x=gender, y=x, fill=NPS\_type)) + geom\_bar(stat="identity")+theme\_minimal()

#Number of children

myData <- transform(myData, similarityTag= ifelse(Similarity == 1, "Similar", ifelse(Similarity > 0.5, "some similar", "no similar") ))

unique(myData$CHILDREN\_NUM\_C)

myData <- myData[myData$CHILDREN\_NUM\_C != 5,] myData <- myData[myData$CHILDREN\_NUM\_C != 6,]

agg.data <- aggregate(myData$count, by = list(children = myData$CHILDREN\_NUM\_C, NPS\_type = myData$NPS\_Type),FUN = sum)

p1 <- ggplot(agg.data, aes(x=children, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by Number of Children") + # for the main title xlab("Number of Children") + # for the x axis label

ylab("Frequency") # for the y axis label

#Age

age <- as.data.frame(table(myData$Age\_Range\_H)) colnames(age) <- c("Age\_range", "Freq")

age <- age[age$Age\_range != "",]

bp<- ggplot(age, aes(x="", y=Freq, fill=Age\_range))+ geom\_bar(width = 1, stat = "identity")

bp

pie <- bp + coord\_polar("y", start=0) + ggtitle("Age Distribution") # for the main title

pie

agg.data <- aggregate(myData$count, by = list(age = myData$Age\_Range\_H, NPS\_type = myData$NPS\_Type),

FUN = sum)

agg.data <- agg.data[agg.data$age != "",]

p1 <- ggplot(agg.data, aes(x=age, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by Age Range") + # for the main title xlab("Age") + # for the x axis label

ylab("Frequency") # for the y axis label

#Country

#myData <- myData[myData$Country\_PL == "United States",] myData <- myData[myData$State\_PL == "Texas",]

USA <- as.data.frame(table(myData$NPS\_Type)) colnames(USA) <- c("NPS\_Type", "Freq")

bp<- ggplot(USA, aes(x="", y=Freq, fill=NPS\_Type))+ geom\_bar(width = 1, stat = "identity")

bp

pie <- bp + coord\_polar("y", start=0) +

ggtitle("NPS\_Type distribution in Texas") # for the main title pie

#Ammenities myData$Likelihood\_Recommend\_H

barplot(table(myData$Likelihood\_Recommend\_H))

ggplot(myData, aes(x=Spa.online.booking\_PL, y=Likelihood\_Recommend\_H)) + geom\_point() # Change the point size, and shape

ggplot(mtcars, aes(x=wt, y=mpg)) + geom\_point(size=2, shape=23)

library(ggplot2) library(maps)

tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

myData <- myData[myData$Country\_PL == "United States", ] myData <- myData[myData$Resort\_PL == "N",]

#load us map data

all\_states <- map\_data("state")

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = all\_states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

MissingAmmenities <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "blue", size = 1) + ggtitle("Hotels with Restaurants, Spa, and Resort")

#By state #California

states <- subset(all\_states, region %in% c( "california") ) myData <- myData[myData$State\_PL == "California",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

California <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in California")

#Florida

states <- subset(all\_states, region %in% c( "florida") ) myData <- myData[myData$State\_PL == "Florida",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

RestaurantMap <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in Florida")

#Texas

states <- subset(all\_states, region %in% c( "texas") ) myData <- myData[myData$State\_PL == "Texas",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

RestaurantMap <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in Texas")

#Hotels with low Guest room satisfaction score

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE) unique(myData$Guest\_Room\_H)

myData <- myData[myData$Country\_PL == "United States", ] myData <- myData[myData$Guest\_Room\_H < 6 ,]

#Texas

states <- subset(all\_states, region %in% c( "texas") ) myData <- myData[myData$State\_PL == "Texas",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

RestaurantMap <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in Texas")

#California

states <- subset(all\_states, region %in% c( "california") ) myData <- myData[myData$State\_PL == "California",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

California <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in California")

#Florida

states <- subset(all\_states, region %in% c( "florida") ) myData <- myData[myData$State\_PL == "Florida",]

ammenitiesMap <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL")]

South-East Map <- ggplot() + geom\_polygon(data = states, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

RestaurantMap <-South-East Map + geom\_point(aes(x = ammenitiesMap$Property.Longitude\_PL, y = ammenitiesMap$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Hotels with no Resorts in Florida")

## Linear models

library("caret") #Linear model code

#Reading the dataset tmp <- file.choose()

tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

#Deleting NA

myData <- na.omit(myData)

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

#Linear Models (General)

#Demographic information

#Using age, children and gener as independent variables against NPS\_TYPE #First we need to convert the columns to numeric types

#Age unique(myData$Age\_Range\_H) table(myData$Age\_Range\_H)

myData$Age\_Range\_H <- sapply(myData$Age\_Range\_H, as.numeric) #Ignoring white spaces

myData <- myData[myData$Age\_Range\_H != 1,] #"" -> 1

#18 - 25 -> 2

#26-35 -> 3

#36 - 45 -> 4

#46 - 55 -> 5

#56 - 65 -> 6

#66 - 75 -> 7

#76+ -> 8

#Gender unique(myData$Gender\_H) table(myData$Gender\_H)

myData$Gender\_H <- sapply(myData$Gender\_H, as.numeric) #Ignoring white spaces and "Prefer not to answer" myData <- myData[myData$Gender\_H != 1,]

myData <- myData[myData$Gender\_H != 4,] #"" -> 1

#Female -> 2

#Male -> 3

#Prefer not to answe -> 4

#Childs unique(myData$CHILDREN\_NUM\_C) table(myData$CHILDREN\_NUM\_C)

#City unique(myData$City\_PL) sort(table(myData$City\_PL))

myData$City\_PL <- sapply(myData$City\_PL, as.numeric)

#State unique(myData$State\_PL) sort(table(myData$State\_PL))

myData$State\_PL <- sapply(myData$State\_PL, as.numeric) #Ignoring white spaces

myData <- myData[myData$State\_PL != 1,]

#Country unique(myData$Country\_PL) sort(table(myData$Country\_PL))

myData$Country\_PL <- sapply(myData$Country\_PL, as.numeric)

unique(myData$Currency\_PL) table(myData$Currency\_PL)

myData$Currency\_PL<-sapply(myData$Currency\_PL, as.numeric)

#NPS\_TYPE

unique(myData$NPS\_Type)

table(myData$NPS\_Type)

myData$NPS\_Type <- sapply(myData$NPS\_Type, as.numeric) #Detractors -> 1

#Passive -> 2

#Promoter -> 3 unique(myData$Overall\_Sat\_H)

#Testing the linear Demographic carachteristics

model <- lm(formula = Overall\_Sat\_H ~ CHILDREN\_NUM\_C + Gender\_H + Age\_Range\_H + City\_PL + Country\_PL + Currency\_PL + State\_PL, data = myData)

summary(model)

plot(myData$City\_PL + myData$State\_PL +myData$Country\_PL + myData$Currency\_PL + myData$Gender\_H + myData$CHILDREN\_NUM\_C + myData$Gender\_H, myData$Overall\_Sat\_H, xlab = "Demographic variables", ylab = "Overall satisfaction", pch = 19, xlim = c(0,100), ylim = c(8,10))

abline(model, col = "red") varImp(model)

#Testing the linear Demographic carachteristics vs NPS\_Type

model <- lm(formula = NPS\_Type ~ State\_PL + CHILDREN\_NUM\_C + Gender\_H + Age\_Range\_H + City\_PL

+ Country\_PL + Currency\_PL, data = myData) summary(model)

plot(myData$City\_PL + myData$State\_PL+ myData$Country\_PL + myData$Currency\_PL + myData$Gender\_H + myData$CHILDREN\_NUM\_C + myData$Gender\_H, myData$NPS\_Type, xlab = "Demographic variables", ylab = "NPS Type", pch = 19, xlim = c(0,40), ylim = c(1,3))

abline(model, col = "red") varImp(model)

#Testing the linear Demographic carachteristics vs Likelihood to recommend

model <- lm(formula = Likelihood\_Recommend\_H ~ State\_PL + CHILDREN\_NUM\_C + Gender\_H + Age\_Range\_H + City\_PL + Country\_PL + Currency\_PL, data = myData)

summary(model)

plot(myData$City\_PL + myData$State\_PL + myData$Country\_PL + myData$Currency\_PL + myData$Gender\_H + myData$CHILDREN\_NUM\_C + myData$Gender\_H, myData$Likelihood\_Recommend\_H ,

xlab = "Demographic variables", ylab = "Likelihood to recommend", pch = 19, xlim = c(0,60), ylim = c(8,10))

abline(model, col = "red") varImp(model)

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

#Using satisfaction of services for model #Trying to create another using recomendations

model <- lm(formula = NPS\_Type ~ Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$NPS\_Type, xlab = "Satisfaction Scores", ylab = "NPS Type", pch = 19)

abline(model, col = "red") varImp(model)

#Likelihood to recommend

model <- lm(formula = Likelihood\_Recommend\_H ~ Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$Likelihood\_Recommend\_H, xlab = "Satisfaction Scores", ylab = "Likelihood to recommend", pch = 19)

abline(model, col = "red") varImp(model)

#Overall satisfaction

model <- lm(formula = Overall\_Sat\_H ~ Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$Overall\_Sat\_H, xlab = "Satisfaction Scores", ylab = "Overall satisfaction", pch = 19, xlim = c(5,20))

abline(model, col = "red") varImp(model)

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

#Using ammenities for lineal models

#Transform all the ammenities values (Y,N) to numeric unique(myData$Mini.Bar\_PL)

myData$Mini.Bar\_PL <- sapply(myData$Mini.Bar\_PL, as.numeric)

unique(myData$Pool.Indoor\_PL)

myData$Pool.Indoor\_PL <- sapply(myData$Pool.Indoor\_PL, as.numeric)

unique(myData$Pool.Outdoor\_PL)

myData$Pool.Outdoor\_PL <- sapply(myData$Pool.Outdoor\_PL, as.numeric)

unique(myData$Regency.Grand.Club\_PL)

myData$Regency.Grand.Club\_PL <- sapply(myData$Regency.Grand.Club\_PL, as.numeric)

unique(myData$Resort\_PL)

myData$Resort\_PL <- sapply(myData$Resort\_PL, as.numeric)

unique(myData$Restaurant\_PL)

myData$Restaurant\_PL <- sapply(myData$Restaurant\_PL, as.numeric)

unique(myData$Self.Parking\_PL)

myData$Self.Parking\_PL <- sapply(myData$Self.Parking\_PL, as.numeric)

unique(myData$Shuttle.Service\_PL)

myData$Shuttle.Service\_PL <- sapply(myData$Shuttle.Service\_PL, as.numeric)

unique(myData$Ski\_PL)

myData$Ski\_PL <- sapply(myData$Ski\_PL, as.numeric)

unique(myData$Spa\_PL)

myData$Spa\_PL <- sapply(myData$Spa\_PL, as.numeric)

unique(myData$Spa.services.in.fitness.center\_PL)

myData$Spa.services.in.fitness.center\_PL <- sapply(myData$Spa.services.in.fitness.center\_PL, as.numeric)

unique(myData$Spa.online.booking\_PL)

myData$Spa.online.booking\_PL <- sapply(myData$Spa.online.booking\_PL, as.numeric)

#Likelihood to recommend as dependent variable

model <- lm(formula = Likelihood\_Recommend\_H ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL

+ Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$Likelihood\_Recommend\_H, xlab = "Ammenities count", ylab = "Likelihood to recommend", pch = 19)

abline(model, col = "red") varImp(model)

#Overallsatisfaction as dependent variable

model <- lm(formula = Overall\_Sat\_H ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL + Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$Overall\_Sat\_H , xlab = "Ammenities", ylab = "Overall satisfaction", pch = 19, ylim = c(9,13), xlim = c(0,10))

abline(model, col = "red") varImp(model)

#NPS\_Type as dependent variable

model <- lm(formula = NPS\_Type ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL + Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$NPS\_Type , xlab

= "Amenities", ylab = "NPS\_TYPE", pch = 19, xlim = c(0,100), ylim = c(1, 3)) abline(model, col = "red")

varImp(model)

##########################################################################3

#Linear Models for Different states

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE) sort(table(myData$State\_PL))

#Converting to numeric NPS\_TYPE unique(myData$NPS\_Type) table(myData$NPS\_Type)

myData$NPS\_Type <- sapply(myData$NPS\_Type, as.numeric)

#Selecting the state

myData <- myData[myData$State\_PL == "Florida",]

#Using satisfaction of services for model

#Trying to create another using recomendations model <- lm(formula = NPS\_Type ~

Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$NPS\_Type, xlab = "Satisfaction Scores", ylab = "NPS Type", pch = 19)

abline(model, col = "red") varImp(model)

#Likelihood to recommend

model <- lm(formula = Likelihood\_Recommend\_H ~ Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$Likelihood\_Recommend\_H, xlab = "Satisfaction Scores", ylab = "Likelihood to recommend", pch = 19)

abline(model, col = "red") varImp(model)

#Overall satisfaction

model <- lm(formula = Overall\_Sat\_H ~ Guest\_Room\_H+Tranquility\_H+Condition\_Hotel\_H+Customer\_SVC\_H+Staff\_Cared\_H+Internet\_Sat\_H+ Check\_In\_H, data = myData)

summary(model)

plot(myData$Guest\_Room\_H+myData$Tranquility\_H+myData$Condition\_Hotel\_H+myData$Customer\_SVC\_H+m yData$Staff\_Cared\_H+myData$Internet\_Sat\_H+ myData$Check\_In\_H, myData$Overall\_Sat\_H, xlab = "Satisfaction Scores", ylab = "Overall satisfaction", pch = 19, xlim = c(5,20))

abline(model, col = "red") varImp(model)

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

#Using ammenities for lineal models

#Transform all the ammenities values (Y,N) to numeric myData$Mini.Bar\_PL <- sapply(myData$Mini.Bar\_PL, as.numeric) myData$Pool.Indoor\_PL <- sapply(myData$Pool.Indoor\_PL, as.numeric) myData$Pool.Outdoor\_PL <- sapply(myData$Pool.Outdoor\_PL, as.numeric)

myData$Regency.Grand.Club\_PL <- sapply(myData$Regency.Grand.Club\_PL, as.numeric) myData$Resort\_PL <- sapply(myData$Resort\_PL, as.numeric)

myData$Restaurant\_PL <- sapply(myData$Restaurant\_PL, as.numeric) myData$Self.Parking\_PL <- sapply(myData$Self.Parking\_PL, as.numeric) myData$Shuttle.Service\_PL <- sapply(myData$Shuttle.Service\_PL, as.numeric) myData$Ski\_PL <- sapply(myData$Ski\_PL, as.numeric)

myData$Spa\_PL <- sapply(myData$Spa\_PL, as.numeric)

myData$Spa.services.in.fitness.center\_PL <- sapply(myData$Spa.services.in.fitness.center\_PL, as.numeric)

myData$Spa.online.booking\_PL <- sapply(myData$Spa.online.booking\_PL, as.numeric)

#Likelihood to recommend as dependent variable

model <- lm(formula = Likelihood\_Recommend\_H ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL

+ Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$Likelihood\_Recommend\_H, xlab = "Ammenities count", ylab = "Likelihood to recommend", pch = 19)

abline(model, col = "red") varImp(model)

#Overallsatisfaction as dependent variable

model <- lm(formula = Overall\_Sat\_H ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL + Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$Overall\_Sat\_H , xlab = "Ammenities", ylab = "Overall satisfaction", pch = 19)

abline(model, col = "red") varImp(model)

#NPS\_Type as dependent variable

model <- lm(formula = NPS\_Type ~ Mini.Bar\_PL + Pool.Indoor\_PL + Pool.Outdoor\_PL + Regency.Grand.Club\_PL + Resort\_PL + Restaurant\_PL + Self.Parking\_PL + Shuttle.Service\_PL + Ski\_PL + Spa\_PL + Spa.services.in.fitness.center\_PL + Spa.online.booking\_PL, data = myData)

summary(model)

plot(myData$Mini.Bar\_PL + myData$Pool.Indoor\_PL + myData$Pool.Outdoor\_PL + myData$Regency.Grand.Club\_PL + myData$Resort\_PL + myData$Restaurant\_PL + myData$Self.Parking\_PL + myData$Shuttle.Service\_PL + myData$Ski\_PL + myData$Spa\_PL + myData$Spa.services.in.fitness.center\_PL + myData$Spa.online.booking\_PL,myData$NPS\_Type , xlab

= "Amenities", ylab = "NPS\_TYPE", pch = 19, xlim = c(0,100), ylim = c(1, 3)) abline(model, col = "red")

varImp(model)

Maps:

#install.packages("mapdata") library(ggplot2) library(ggmap)

library(maps) library(mapdata)

tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE) colnames(myData)

str(myData)

#Subsetting the columns for maps of NPS\_TYPE

NPS\_subset <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL","NPS\_Type")]

#separating points by NPS\_TYPE unique(NPS\_subset$NPS\_Type) #Detractors

Detractors <- NPS\_subset[NPS\_subset$NPS\_Type == "Detractor",] #Passive

Passives <- NPS\_subset[NPS\_subset$NPS\_Type == "Passive",] #Promoters

Promoters <- NPS\_subset[NPS\_subset$NPS\_Type == "Promoter",]

world <- map\_data("world") South-East Map <- 0

South-East Map <- ggplot() + geom\_polygon(data = world, aes(x=long, y = lat, group = group), fill="grey40", colour="grey90", alpha=1) +

coord\_fixed(1.3)

detractorsMap <-South-East Map + geom\_point(aes(x = Detractors$Property.Longitude\_PL, y = Detractors$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Detractors")

passivesMap <-South-East Map + geom\_point(aes(x = Passives$Property.Longitude\_PL, y = Passives$Property.Latitude\_PL), color = "yellow", size = 1) + ggtitle("Passives")

promotersMap <-South-East Map + geom\_point(aes(x = Promoters$Property.Longitude\_PL, y = Promoters$Property.Latitude\_PL), color = "blue", size = 1) + ggtitle("Promoters")

#Density maps for passives

passivesHeatMap <- ggplot(Passives, aes(Property.Longitude\_PL, Property.Latitude\_PL))

passivesHeatMap <- passivesHeatMap + geom\_polygon(data=world,aes(x=long,y=lat,group=group),color='gray',fill="black",alpha=.35)

passivesHeatMap <- passivesHeatMap + geom\_density\_2d()

passivesHeatMap <- passivesHeatMap + stat\_density\_2d(aes(fill = ..level..), geom = "polygon")

+ ggtitle("Passives") passivesHeatMap #Zooming US

passivesHeatMap <- passivesHeatMap + xlim(-130,-50)+ylim(25,50) #getting the northeast U.S. passivesHeatMap <- passivesHeatMap + coord\_fixed(1.3) #fixing coordenates

passivesHeatMap

#Density maps for detractors

detractorsHeatMap <- ggplot(Detractors, aes(Property.Longitude\_PL, Property.Latitude\_PL))

detractorsHeatMap <- detractorsHeatMap + geom\_polygon(data=world,aes(x=long,y=lat,group=group),color='gray',fill="black",alpha=.35)

detractorsHeatMap <- detractorsHeatMap + geom\_density\_2d()

detractorsHeatMap <- detractorsHeatMap + stat\_density\_2d(aes(fill = ..level..), geom = "polygon") + ggtitle("Detractors")

detractorsHeatMap #Zooming US

detractorsHeatMap <- detractorsHeatMap + xlim(-130,-50)+ylim(25,50) #getting the northeast U.S.

detractorsHeatMap <- detractorsHeatMap + coord\_fixed(1.3) #fixing coordenates detractorsHeatMap

#Density maps for promoters

promotersHeatMap <- ggplot(Promoters, aes(Property.Longitude\_PL, Property.Latitude\_PL))

promotersHeatMap <- promotersHeatMap + geom\_polygon(data=world,aes(x=long,y=lat,group=group),color='gray',fill="black",alpha=.35)

promotersHeatMap <- promotersHeatMap + geom\_density\_2d()

promotersHeatMap <- promotersHeatMap + stat\_density\_2d(aes(fill = ..level..), geom = "polygon") + ggtitle("Promoters")

promotersHeatMap #Zooming US

promotersHeatMap <- promotersHeatMap + xlim(-130,-50)+ylim(25,50) #getting the northeast U.S. promotersHeatMap <- promotersHeatMap + coord\_fixed(1.3) #fixing coordenates

promotersHeatMap

#Most popular cities unique(myData$City\_PL)

sort(table(myData$City\_PL))

## Arules

#Creating arules #loading library library("arules") library("arulesViz") tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

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

#Creating subset for ammenities

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "Self.Parking\_PL", "Shuttle.Service\_PL", "Ski\_PL", "Spa\_PL", "Spa.online.booking\_PL" , "Spa.services.in.fitness.center\_PL", "Pool.Indoor\_PL", "Mini.Bar\_PL", "Restaurant\_PL"

,"NPS\_Type")] myDataAmmenities[myDataAmmenities==""] <- NA

#Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset))

itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 2.5] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.64] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

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

#Using demographic information

myDataDemographics <- myData[, c("Gender\_H","Age\_Range\_H" ,"Country\_PL","NPS\_Type")] myDataDemographics[myDataDemographics==""] <- NA

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

#Promoters

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.5] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.6] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#

#Subsetting by state

myData <- myData[myData$State\_PL == "California",] ####################################################################################

#Creating subset for ammenities

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "Self.Parking\_PL", "Shuttle.Service\_PL", "Ski\_PL", "Spa\_PL", "Spa.online.booking\_PL" , "Spa.services.in.fitness.center\_PL", "Pool.Indoor\_PL", "Mini.Bar\_PL", "Restaurant\_PL"

,"NPS\_Type")] myDataAmmenities[myDataAmmenities==""] <- NA

#Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.4] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(head(goodrules))

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(head(goodrules))

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.64] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

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

#Using demographic information

myDataDemographics <- myData[, c("Gender\_H","Age\_Range\_H" ,"NPS\_Type")] myDataDemographics[myDataDemographics==""] <- NA

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

#Promoters

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.5] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.6] goodrules <- sort(goodrules, by="lift", decreasing = TRUE)

summary(goodrules) inspect(goodrules)

## SVM Models

#install.packages("rminer") library(rminer) library("caret") library(kernlab) library("ggplot2") library("e1071") library(gridExtra)

tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE) colnames(myData)

#Classify supporters and detractors unique(myData$NPS\_Type)

NPS\_subset <- myData[myData$NPS\_Type != "Passive",] NPS\_subset$NPS\_Type <- factor(NPS\_subset$NPS\_Type) unique(NPS\_subset$NPS\_Type)

#Subsetting by state

NPS\_subset <- NPS\_subset[NPS\_subset$State\_PL == "California",]

#Subsetting the columns for ammenities #NPS\_subset <-

NPS\_subset[,c("Mini.Bar\_PL","Pool.Indoor\_PL","Pool.Outdoor\_PL","Regency.Grand.Club\_PL","Resort

\_PL","Restaurant\_PL","Self.Parking\_PL","Shuttle.Service\_PL","Ski\_PL","Spa.online.booking\_PL"," Spa\_PL","Spa.services.in.fitness.center\_PL","NPS\_Type")]

#Subsetting the columns for demographics #NPS\_subset <-

NPS\_subset[,c("CHILDREN\_NUM\_C","Guest\_Country\_H","Gender\_H","Age\_Range\_H","NPS\_Type")]

#Subsetting for services opinion

#NPS\_subset <- NPS\_subset[,c("Guest\_Room\_H","Tranquility\_H","Condition\_Hotel\_H","Customer\_SVC\_H","Staff\_Cared

\_H","Internet\_Sat\_H","Check\_In\_H","NPS\_Type")]

#Converting to NA blank spaces NPS\_subset[NPS\_subset==""] <- "N"

#creating training and testing #Shuffle dataset

randIndex <- sample(1:dim(NPS\_subset)[1])

#calculate "cut point"

cutPoint2\_3 <- floor(2 \* dim(NPS\_subset)[1]/3) cutPoint2\_3

#Training data

trainData <- NPS\_subset[randIndex[1:cutPoint2\_3], ]

#Testing data

testData <- NPS\_subset[randIndex[(cutPoint2\_3 + 1):dim(NPS\_subset)[1]],]

#1) Build a model (using the 'ksvm' function)

ksvmOutput <- ksvm(NPS\_Type ~ ., data = trainData, kernel = "rbfdot", kpar ="automatic", C = 50, cross = 3, prob.model = TRUE)

ksvmOutput

ksvmPred <- predict(ksvmOutput, testData) length(ksvmPred)

str(ksvmPred)

compTable <- data.frame(testData[,8],ksvmPred) colnames(compTable) <- c("test", "Pred") #Compute the percentage of correct cases

perc\_ksvm <- length(which(compTable$test==compTable$Pred))/dim(compTable)[1] perc\_ksvm

#Create confusion Matrix

ksvmResults <- table(test = compTable$test, pred = compTable$Pred) print(ksvmResults)

#2) Build a model (using the 'svm' function) svmOutput <- svm(NPS\_Type ~ ., data = trainData) svmOutput

svmPred <- predict(svmOutput, testData) str(svmPred)

compTable <- data.frame(testData[,13],svmPred) colnames(compTable) <- c("test", "Pred") #Compute the percentage of correct cases

perc\_svm <- length(which(compTable$test==compTable$Pred))/dim(compTable)[1] perc\_svm

#Create confusion Matrix

svmResults <- table(test = compTable$test, pred = compTable$Pred) print(ksvmResults)

#3) Build a model (using the 'naive bayes' function) nbOutput <- naiveBayes(NPS\_Type ~ ., data = trainData) nbOutput

nbPred <- predict(nbOutput, testData) str(nbPred)

compTable <- data.frame(testData[,13],nbPred) colnames(compTable) <- c("test", "Pred") #Compute the percentage of correct cases

perc\_nb <- length(which(compTable$test==compTable$Pred))/dim(compTable)[1] perc\_nb

#Create confusion Matrix

nbResults <- table(test = compTable$test, pred = compTable$Pred) print(nbResults)

#4) Trying other algorithms

modelTree <- fit(NPS\_Type ~ ., data = trainData,model="ctree") plot(modelTree@object) # show model

## Experiments (Descriptive analysis and linear models)

#install.packages("sqldf") library("caret") library("sqldf")

tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

myData <- na.omit(myData)

myData <- myData[complete.cases(myData),] str(myData$NPS\_Type)

str(myData$Gender\_H) unique(myData$NPS\_Type)

#Relationship between gender and NPS Type barplot(table(myData$Gender\_H)) barplot(table(myData$NPS\_Type))

agg.data <- aggregate(myData$Overall\_Sat\_H, by = list(gender = myData$Gender\_H, NPS\_type = myData$NPS\_Type),

FUN = sum) bar.colors <- rep("blue", nrow(agg.data))

bar.colors[agg.data$gender == "Female"] <- "pink" bar.colors[agg.data$gender == "Male"] <- "blue" bar.colors[agg.data$gender == "Prefer not to answer"] <- "yellow" bar.colors[agg.data$gender == ""] <- "white"

barplot(agg.data$x, names.arg = agg.data$NPS\_type

, las = 2

, horiz = T

, col = bar.colors) #Linear Model

#Changing to numeric values to make linear model myData$NPS\_Type\_numeric <- sapply(myData$NPS\_Type, as.numeric) myData$Gender\_H\_numeric <- sapply(myData$Gender\_H, as.numeric) unique(myData$Gender\_H\_numeric) unique(myData$NPS\_Type\_numeric)

model1 <- lm(formula = NPS\_Type\_numeric ~ Gender\_H\_numeric, data = myData) summary(model1)

plot(myData$Gender\_H\_numeric, myData$NPS\_Type\_numeric, xlab = "Gender", ylab = "NPS\_Type", pch

= 19)

abline(model1, col = "red")

#Age range barplot(table(myData$Age\_Range\_H))

agg.data <- aggregate(myData$Overall\_Sat\_H, by = list(age = myData$Age\_Range\_H, NPS\_type = myData$NPS\_Type),

FUN = sum)

unique(agg.data$age)

bar.colors <- rep("gray", nrow(agg.data)) bar.colors[agg.data$age == "18-25"] <- "pink" bar.colors[agg.data$age == "26-35"] <- "blue" bar.colors[agg.data$age == "36-45"] <- "yellow" bar.colors[agg.data$age == "46-55"] <- "white" bar.colors[agg.data$age == "56-65"] <- "red" bar.colors[agg.data$age == "66-75"] <- "green" bar.colors[agg.data$age == "76+"] <- "orange"

barplot(agg.data$x, names.arg = agg.data$NPS\_type

, las = 2

, horiz = T

, col = bar.colors)

#Linear Model

#Changing to numeric values to make linear model myData$Age\_Range\_H\_numeric <- sapply(myData$Age\_Range\_H, as.numeric) unique(myData$Age\_Range\_H\_numeric)

unique(myData$NPS\_Type\_numeric)

model1 <- lm(formula = NPS\_Type\_numeric ~ myData$Age\_Range\_H\_numeric, data = myData) summary(model1)

plot(myData$Age\_Range\_H\_numeric, myData$NPS\_Type\_numeric, xlab = "Age", ylab = "NPS\_Type", pch

= 19)

abline(model1, col = "red")

#Linear models with likelihood to recommend as the dependent variable #Pool outdoor

barplot(table(myData$Pool.Outdoor\_PL))

myData$Pool.Outdoor\_PL\_numeric <- sapply(myData$Pool.Outdoor\_PL, as.numeric) myData <- myData[myData$Pool.Outdoor\_PL\_numeric != 1,]

model2 <- lm(formula = Likelihood\_Recommend\_H ~ Pool.Outdoor\_PL\_numeric, data = myData) summary(model2)

plot(myData$Pool.Outdoor\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Pool Outdoor", ylab = "Likelihood to recommend", pch = 19)

abline(model2, col = "red")

#Age range

#Regency Grand club\_PL

myData$Regency.Grand.Club\_PL\_numeric <- sapply(myData$Regency.Grand.Club\_PL, as.numeric) myData <- myData[myData$Regency.Grand.Club\_PL\_numeric != 1,]

model3 <- lm(formula = Likelihood\_Recommend\_H ~ Regency.Grand.Club\_PL\_numeric, data = myData) summary(model3)

plot(myData$Regency.Grand.Club\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Regency Grand Club", ylab = "Likelihood to recommend", pch = 19)

abline(model3, col = "red")

#Resort\_PL

myData$Resort\_PL\_numeric <- sapply(myData$Resort\_PL, as.numeric) myData <- myData[myData$Resort\_PL\_numeric != 1,]

model4 <- lm(formula = Likelihood\_Recommend\_H ~ myData$Resort\_PL\_numeric, data = myData) summary(model4)

plot(myData$Resort\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Resort", ylab = "Likelihood to recommend", pch = 19)

abline(model4, col = "red")

#Combining all the variables

model5 <- lm(formula = Likelihood\_Recommend\_H ~ Resort\_PL\_numeric + Regency.Grand.Club\_PL\_numeric + Pool.Outdoor\_PL\_numeric, data = myData)

summary(model5)

step(model5,direction = c("both", "backward", "forward")) varImp(model5)

#Checking gender and NPS variable table(myData$Gender\_H, myData$NPS\_Type) #Checking countries reviews unique(myData$Country\_PL) table(myData$Country\_PL, myData$NPS\_Type) sort(table(myData$Country\_PL))

#Code to check number of promoters, detractors and passive users in hotel place

myDataCountries <- myData[myData$Country\_PL == "United States" | myData$Country\_PL == "South Korea" | myData$Country\_PL == "United Arab Emirates" | myData$Country\_PL == "Canada" | myData$Country\_PL == "China" | myData$Country\_PL == "France" | myData$Country\_PL == "India" | myData$Country\_PL == "Japan" | myData$Country\_PL == "Mexico" | myData$Country\_PL == "Thailand",]

myDataCountries$Country\_PL <- factor(myDataCountries$Country\_PL) table(myDataCountries$Country\_PL)

barplot(table(myDataCountries$NPS\_Type,myDataCountries$Country\_PL), col = c("red", "yellow","green"))

#Checking without gender table(myDataCountries$Country\_PL,myDataCountries$NPS\_Type) #Adding gender

table(myDataCountries$Country\_PL,myDataCountries$NPS\_Type, myDataCountries$Gender\_H)

#Code to check how many detractors, passives and promoters we have based on their nationality #Checking unique nationalities

unique(myData$Guest\_Country\_H) #selecting the top 10 sort(table(myData$Guest\_Country\_H))

myNationalities <- myData[myData$Guest\_Country\_H == "USA" | myData$Guest\_Country\_H == "Canada"

| myData$Guest\_Country\_H == "Korea" | myData$Guest\_Country\_H == "Australia" | myData$Guest\_Country\_H == "Japan" | myData$Guest\_Country\_H == "United Kingdom" | myData$Guest\_Country\_H == "Mexico" | myData$Guest\_Country\_H == "Germany" | myData$Guest\_Country\_H == "China" | myData$Guest\_Country\_H == "India", ]

myNationalities$Guest\_Country\_H <- factor(myNationalities$Guest\_Country\_H) table(myNationalities$Guest\_Country\_H, myNationalities$NPS\_Type)

barplot(table(myNationalities$NPS\_Type,myNationalities$Guest\_Country\_H), col = c("red", "yellow", "green"))

#Creating arules #loading library library("arules") #Creating subset

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "NPS\_Type")]

myDataAmmenities[myDataAmmenities==""] <- NA

myDataAmmenities <- myDataAmmenities[complete.cases(myDataAmmenities),] #Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset))

itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Combining all the ammenities #Creating arules

#Creating subset

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "Self.Parking\_PL", "Shuttle.Service\_PL", "Ski\_PL", "Spa\_PL", "Spa.online.booking\_PL" , "Spa.services.in.fitness.center\_PL", "Pool.Indoor\_PL", "Mini.Bar\_PL", "Restaurant\_PL"

,"NPS\_Type")] myDataAmmenities[myDataAmmenities==""] <- NA

#Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 2.5] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.64] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

## Experiment 2 (arules)

#install.packages("sqldf") library("caret") library("sqldf") library("ggplot2")

tmp <- file.choose() tmp

myData <- read.csv(tmp, header = TRUE, stringsAsFactors = TRUE)

myData <- na.omit(myData)

myData <- myData[complete.cases(myData),] str(myData$NPS\_Type)

str(myData$Gender\_H) unique(myData$NPS\_Type)

#Relationship between gender and NPS Type barplot(table(myData$Gender\_H)) barplot(table(myData$NPS\_Type)) myData$count <- 1

agg.data <- aggregate(myData$count, by = list(gender = myData$Gender\_H, NPS\_type = myData$NPS\_Type),

FUN = sum) bar.colors <- rep("blue", nrow(agg.data))

bar.colors[agg.data$gender == "Female"] <- "pink" bar.colors[agg.data$gender == "Male"] <- "blue" bar.colors[agg.data$gender == "Prefer not to answer"] <- "yellow" bar.colors[agg.data$gender == ""] <- "white"

barplot(agg.data$x, names.arg = agg.data$NPS\_type

, las = 2

, horiz = T

, col = bar.colors)

p1 <- ggplot(agg.data, aes(x=gender, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by gender") + # for the main title xlab("Gender") + # for the x axis label

ylab("Frequency") # for the y axis label

p2 <- ggplot(agg.data, aes(x=gender, y=x, fill=NPS\_type)) + geom\_bar(stat="identity")+theme\_minimal()

#Linear Model

#Changing to numeric values to make linear model myData$NPS\_Type\_numeric <- sapply(myData$NPS\_Type, as.numeric) myData$Gender\_H\_numeric <- sapply(myData$Gender\_H, as.numeric) unique(myData$Gender\_H\_numeric) unique(myData$NPS\_Type\_numeric)

model1 <- lm(formula = NPS\_Type\_numeric ~ Gender\_H\_numeric, data = myData)

summary(model1)

plot(myData$Gender\_H\_numeric, myData$NPS\_Type\_numeric, xlab = "Gender", ylab = "NPS\_Type", pch

= 19)

abline(model1, col = "red")

#Age range barplot(table(myData$Age\_Range\_H))

agg.data <- aggregate(myData$Overall\_Sat\_H, by = list(age = myData$Age\_Range\_H, NPS\_type = myData$NPS\_Type),

FUN = sum)

unique(agg.data$age)

bar.colors <- rep("gray", nrow(agg.data)) bar.colors[agg.data$age == "18-25"] <- "pink" bar.colors[agg.data$age == "26-35"] <- "blue" bar.colors[agg.data$age == "36-45"] <- "yellow" bar.colors[agg.data$age == "46-55"] <- "white" bar.colors[agg.data$age == "56-65"] <- "red" bar.colors[agg.data$age == "66-75"] <- "green" bar.colors[agg.data$age == "76+"] <- "orange"

barplot(agg.data$x, names.arg = agg.data$NPS\_type

, las = 2

, horiz = T

, col = bar.colors)

agg.data <- agg.data[agg.data$age != "",]

p1 <- ggplot(agg.data, aes(x=age, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by Age Range") + # for the main title xlab("Age") + # for the x axis label

ylab("Frequency") # for the y axis label

#Linear Model

#Changing to numeric values to make linear model myData$Age\_Range\_H\_numeric <- sapply(myData$Age\_Range\_H, as.numeric) unique(myData$Age\_Range\_H\_numeric)

unique(myData$NPS\_Type\_numeric)

model1 <- lm(formula = NPS\_Type\_numeric ~ myData$Age\_Range\_H\_numeric, data = myData) summary(model1)

plot(myData$Age\_Range\_H\_numeric, myData$NPS\_Type\_numeric, xlab = "Age", ylab = "NPS\_Type", pch

= 19)

abline(model1, col = "red")

#Linear models with likelihood to recommend as the dependent variable #Pool outdoor

barplot(table(myData$Pool.Outdoor\_PL))

myData$Pool.Outdoor\_PL\_numeric <- sapply(myData$Pool.Outdoor\_PL, as.numeric) myData <- myData[myData$Pool.Outdoor\_PL\_numeric != 1,]

model2 <- lm(formula = Likelihood\_Recommend\_H ~ Pool.Outdoor\_PL\_numeric, data = myData) summary(model2)

plot(myData$Pool.Outdoor\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Pool Outdoor", ylab = "Likelihood to recommend", pch = 19)

abline(model2, col = "red")

#Age range

#Regency Grand club\_PL

myData$Regency.Grand.Club\_PL\_numeric <- sapply(myData$Regency.Grand.Club\_PL, as.numeric) myData <- myData[myData$Regency.Grand.Club\_PL\_numeric != 1,]

model3 <- lm(formula = Likelihood\_Recommend\_H ~ Regency.Grand.Club\_PL\_numeric, data = myData) summary(model3)

plot(myData$Regency.Grand.Club\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Regency Grand Club", ylab = "Likelihood to recommend", pch = 19)

abline(model3, col = "red")

#Resort\_PL

myData$Resort\_PL\_numeric <- sapply(myData$Resort\_PL, as.numeric) myData <- myData[myData$Resort\_PL\_numeric != 1,]

model4 <- lm(formula = Likelihood\_Recommend\_H ~ myData$Resort\_PL\_numeric, data = myData) summary(model4)

plot(myData$Resort\_PL\_numeric, myData$Likelihood\_Recommend\_H, xlab = "Resort", ylab = "Likelihood to recommend", pch = 19)

abline(model4, col = "red")

#Combining all the variables

model5 <- lm(formula = Likelihood\_Recommend\_H ~ Resort\_PL\_numeric + Regency.Grand.Club\_PL\_numeric + Pool.Outdoor\_PL\_numeric, data = myData)

summary(model5)

step(model5,direction = c("both", "backward", "forward")) varImp(model5)

#Checking gender and NPS variable table(myData$Gender\_H, myData$NPS\_Type) #Checking countries reviews unique(myData$Country\_PL) table(myData$Country\_PL, myData$NPS\_Type) sort(table(myData$Country\_PL))

#Code to check number of promoters, detractors and passive users in hotel place

myDataCountries <- myData[myData$Country\_PL == "United States" | myData$Country\_PL == "South Korea" | myData$Country\_PL == "United Arab Emirates" | myData$Country\_PL == "Canada" | myData$Country\_PL == "China" | myData$Country\_PL == "France" | myData$Country\_PL == "India" | myData$Country\_PL == "Japan" | myData$Country\_PL == "Mexico" | myData$Country\_PL == "Thailand",]

myDataCountries$Country\_PL <- factor(myDataCountries$Country\_PL) table(myDataCountries$Country\_PL)

barplot(table(myDataCountries$NPS\_Type,myDataCountries$Country\_PL), col = c("red", "yellow","green"))

#Checking without gender table(myDataCountries$Country\_PL,myDataCountries$NPS\_Type) #Adding gender

table(myDataCountries$Country\_PL,myDataCountries$NPS\_Type, myDataCountries$Gender\_H)

#Plotting results

myDataCountries <- myData[myData$Country\_PL == "United States" | myData$Country\_PL == "South Korea" | myData$Country\_PL == "United Arab Emirates" | myData$Country\_PL == "Canada" | myData$Country\_PL == "China",]

myDataCountries$Country\_PL <- factor(myDataCountries$Country\_PL)

agg.data <- aggregate(myDataCountries$count, by = list(country = myDataCountries$Country\_PL, NPS\_type = myDataCountries$NPS\_Type),

FUN = sum) unique(agg.data$country)

bar.colors <- rep("gray", nrow(agg.data)) bar.colors[agg.data$country == "Canada"] <- "red" bar.colors[agg.data$country == "China"] <- "blue" bar.colors[agg.data$country == "France"] <- "yellow" bar.colors[agg.data$country == "India"] <- "white" bar.colors[agg.data$country == "Japan"] <- "purple" bar.colors[agg.data$country == "Mexico"] <- "green" bar.colors[agg.data$country == "Thailand"] <- "orange" bar.colors[agg.data$country == "United Arab Emirates"] <- "pink" bar.colors[agg.data$country == "United States"] <- "blue"

p1 <- ggplot(agg.data, aes(x=country, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by Country") + # for the main title xlab("Country") + # for the x axis label

ylab("Frequency") # for the y axis label

#Code to check how many detractors, passives and promoters we have based on their nationality #Checking unique nationalities

unique(myData$Guest\_Country\_H) #selecting the top 10 sort(table(myData$Guest\_Country\_H))

myNationalities <- myData[myData$Guest\_Country\_H == "USA" | myData$Guest\_Country\_H == "Canada"

| myData$Guest\_Country\_H == "Korea" | myData$Guest\_Country\_H == "Australia" | myData$Guest\_Country\_H == "Japan" | myData$Guest\_Country\_H == "United Kingdom" | myData$Guest\_Country\_H == "Mexico" | myData$Guest\_Country\_H == "Germany" | myData$Guest\_Country\_H == "China" | myData$Guest\_Country\_H == "India", ]

myNationalities$Guest\_Country\_H <- factor(myNationalities$Guest\_Country\_H) table(myNationalities$Guest\_Country\_H, myNationalities$NPS\_Type)

barplot(table(myNationalities$NPS\_Type,myNationalities$Guest\_Country\_H), col = c("red", "yellow", "green"))

#Creating arules #loading library library("arules") library("arulesViz") #Creating subset

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "NPS\_Type")]

myDataAmmenities[myDataAmmenities==""] <- NA

myDataAmmenities <- myDataAmmenities[complete.cases(myDataAmmenities),] #Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE)

summary(goodrules) inspect(goodrules)

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.0] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Combining all the ammenities #Creating arules

#Creating subset

myDataAmmenities <- myData[, c("Resort\_PL", "Regency.Grand.Club\_PL", "Pool.Outdoor\_PL", "Self.Parking\_PL", "Shuttle.Service\_PL", "Ski\_PL", "Spa\_PL", "Spa.online.booking\_PL" ,

"Spa.services.in.fitness.center\_PL", "Pool.Indoor\_PL", "Mini.Bar\_PL", "Restaurant\_PL"

,"NPS\_Type")] myDataAmmenities[myDataAmmenities==""] <- NA

#Promoters

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Detractors

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Detractor"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 2.5] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Passive

ruleset <- apriori( myDataAmmenities, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Passive"), control = list (verbose=F))

summary(ruleset)

inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.64] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Demographic rules #Creating subset

#myDataDemographics <- myData[, c("Gender\_H","Age\_Range\_H" ,"Country\_PL","NPS\_Type")] myDataDemographics <- myData[, c("Gender\_H","NPS\_Type")] myDataDemographics[myDataDemographics==""] <- NA

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

ruleset <- apriori( myDataDemographics, parameter = list(supp=0.001,conf = 0.05), appearance = list(default="lhs",rhs="NPS\_Type=Promoter"), control = list (verbose=F))

summary(ruleset) inspect(head(ruleset)) itemFrequencyPlot(items(ruleset)) plot(ruleset)

#Best Rules

goodrules <- ruleset[quality(ruleset)$lift > 1.2] goodrules <- sort(goodrules, by="lift", decreasing = TRUE) summary(goodrules)

inspect(goodrules)

#Plot Frequency children barplot(table(myData$CHILDREN\_NUM\_C))

agg.data <- aggregate(myData$Overall\_Sat\_H, by = list(children = myData$CHILDREN\_NUM\_C, NPS\_type = myData$NPS\_Type),

FUN = sum) unique(agg.data$children)

barplot(agg.data$x, names.arg = agg.data$NPS\_type

, las = 2

, horiz = T

, col = bar.colors)

agg.data <- agg.data[agg.data$age != "",]

p1 <- ggplot(agg.data, aes(x=children, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)

+

theme\_minimal() +

ggtitle("Number of NPS\_Type by Number of Children") + # for the main title xlab("Number of Children") + # for the x axis label

ylab("Frequency") # for the y axis label

# References

#### Business Dictionary. (2017, March 25). *Market Segmentation Definition*. Retrieved from Business Dictionary: <http://www.businessdictionary.com/definition/market-segmentation.html>

Mode Analytics. (2018, April 29). *Mode Community*. Retrieved from Data wrangling with sql: <https://community.modeanalytics.com/sql/tutorial/data-wrangling-with-sql/>

Logistic regression

**Output**

Call:  
glm(formula = happyCust ~ Airline.Status + Age + Gender + Price.Sensitivity +   
 No.of.Flights.p.a. + Type.of.Travel + Shopping.Amount.at.Airport +   
 Class + Arrival.Delay.greater.5.Mins + Eating.and.Drinking.at.Airport +   
 southeast, family = "binomial", data = clean\_data1)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-2.6091 -0.5274 0.3286 0.8190 2.8538   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.6106194 0.0491168 32.792 < 2e-16 \*\*\*  
Airline.StatusGold 0.9183398 0.0324448 28.305 < 2e-16 \*\*\*  
Airline.StatusPlatinum 0.4675932 0.0473527 9.875 < 2e-16 \*\*\*  
Airline.StatusSilver 1.7059911 0.0256114 66.611 < 2e-16 \*\*\*  
Age -0.0055458 0.0005726 -9.685 < 2e-16 \*\*\*  
GenderMale 0.3572619 0.0179032 19.955 < 2e-16 \*\*\*  
Price.Sensitivity -0.1627882 0.0160193 -10.162 < 2e-16 \*\*\*  
No.of.Flights.p.a. -0.0128533 0.0006598 -19.480 < 2e-16 \*\*\*  
Type.of.TravelMileage tickets -0.4169442 0.0283718 -14.696 < 2e-16 \*\*\*  
Type.of.TravelPersonal Travel -3.0739273 0.0254484 -120.791 < 2e-16 \*\*\*  
Shopping.Amount.at.Airport 0.0007472 0.0001654 4.518 6.25e-06 \*\*\*  
ClassEco -0.2880530 0.0323666 -8.900 < 2e-16 \*\*\*  
ClassEco Plus -0.2720050 0.0410649 -6.624 3.50e-11 \*\*\*  
Arrival.Delay.greater.5.Minsyes -0.9276402 0.0182323 -50.879 < 2e-16 \*\*\*  
Eating.and.Drinking.at.Airport -0.0003434 0.0001702 -2.017 0.0437 \*   
southeastTRUE 0.0316929 0.0331638 0.956 0.3393   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 120595 on 87022 degrees of freedom  
Residual deviance: 81797 on 87007 degrees of freedom  
AIC: 81829  
  
Number of Fisher Scoring iterations: 5

The columns Airline status, Gender male, shopping amount and southeast have a positive impact on the rating being higher

whereas the columns age, price sensitivity, no of flights p.a., type of travel(mileage tickets), type of travel(personal travel), class(economy), class(eco plus),and eating and drinking at airport have a negative impact

Out of these,