Hyatt Hotel Analysis: Recommendations to Improve Net Promoter Scores

By:

Hannah Greenberg,
Leslie Sanchez,
Minyang Wang,
Palaniappan Muthukkaruppan,
Shaojie Zhang

Table of Contents:

Background

A. Executive Summary

Hyatt Hotels Corporation is a leading global hospitality company with a portfolio of 14 premier brands including more than 700 properties in more than 50 countries across six continents. Hyatt Hotels Corporation's vision is a world of understanding and care in where they care for people so they can be their best. This leads to Hyatt's mission of delivering distinctive experiences for their guests. Hyatt senior management has requested our services to provide recommendations for improving their services that will specifically increase the Net Promoter Scores.[1]

The goal of this report is to demonstrate and exhibit a high level analysis based on the data gathered from Hyatt and their customers ratings which will inform our recommendations. Our team used an open source data analysis program, R, to help us perform our analysis process. Through this, we were able to identify patterns, gather information and ultimately predict results to make suggestions that could potentially increase the level of satisfaction of Hyatt customers. Several data analysis methods were used such as regression modeling, association rules and KSVM, Naive Bayes and visualizations to understand the data. Based on the given information and the analysis the team has performed, we are confident in the strength of the recommendations provided to improve NPS scores.

B. Introduction

The purpose of this project is to improve the overall Net Promoter Score, meaning that customers will have a higher likelihood of recommending their fellow friends and relatives to stay at a Hyatt Hotel. Net Promoter Score is calculated by subtracting the percentage of the detractors from the percentage of the promoters. In this dataset, the customers were categorized into three main groups:

Promoters: the class of people who are willing to recommend the hotel

Passives: the class of people who neither promote or demote the hotel

Detractors: the class of people who are not willing to recommend the hotel

For this dataset, our team has decided to focus on the urban business hotels located in the state of California, within a 12 month span. The urban business hotels located in California provided the most information needed to perform a feasible and detailed analysis in order to make actionable insights to increase the NPS score.

[1] https://www.hyatt.com/

C. Business Questions

Final Business Questions

1. Which state in the country and area within a state will provide the most useful information to improve NPS scores?
2. What type of hotel will provide the most useful information to improve NPS scores?
3. Which satisfaction metric including guest room satisfaction, tranquility satisfaction, hotel condition satisfaction, staff cared satisfaction, customer service satisfaction, internet satisfaction and check in satisfaction is most important in predicting a promoter NPS type?
4. Which satisfaction metric including guest room satisfaction, tranquility satisfaction, hotel condition satisfaction, staff cared satisfaction, customer service satisfaction, internet satisfaction and check in satisfaction is most important in predicting a detractor NPS type?
5. What room types should be increased or decreased to improve NPS?

Additional Business Questions Explored

1. What effect does Boutique hotel type, existence of a business center, having a convention center, having limo service, having a mini bar, having an outdoor pool, having a Regency Grand Club, having Self-Parking and having Spa Services in the Fitness Center have on NPS Type?
2. What effect does having all suites, a bell staff, dry cleaning, elevators, a fitness center, fitness trainers, a golf course, indoor corridors, an indoor pool, shuttle service, ski, spa and valet parking have on NPS type?

<u>Data Cleaning</u>

A. Data Acquisition

For this project, we used a dataset with survey data which includes 12 months of data on the Hyatt hotel chain. The data was broken down by month and included 237 variables and close to a million observations for each month. In order to answer our business questions, we downloaded the data in Excel and cleaned for the specific variables we were interested in investigating. We decided to focus our research on the NPS Type, Likelihood of customers to recommend the hotel, State, Location, Hotel Type, Room Type, Guest room Satisfaction, Tranquility, Hotel Condition, Staff Cared Satisfaction, Customer Service Satisfaction, Internet Satisfaction and Check in Satisfaction. For our first analysis, we used only winter months. In order to use the data, we narrowed down each month's excel sheet to only the specific columns needed for our analysis. Afterwards, we brought the data into R and combined all 3 months into one dataframe.

However, after narrowing down the data to urban California Business hotels during the winter months, we only had 3,000 observations. As this was not enough observations to be confident in our analysis, we decided to use all 12 months of data which produced 15,711,540 observations of 13 variables which we felt would provide a more accurate analysis.

In order to use the data for all 12 months, we brought the data directly into R by selecting only the specific columns as needed for our analysis and combining them into a single dataset.

B. Data Cleaning

Once all 12 months data was combined in R, we needed to clean the data to make it usable. We decided to omit all rows with "NA" data or when a participant did not answer one of the questions we were looking at. We decided this method would keep our data the most accurate because only observations with all questions answered would be included in the analysis. Additionally, we decided against using the mean to fill in for missing observations because we thought this would lead our data to no longer be representative of the true population and that it was not appropriate to do so for our variables. We used this cleaning method for all aspects of our data analysis.

Descriptive Statistics

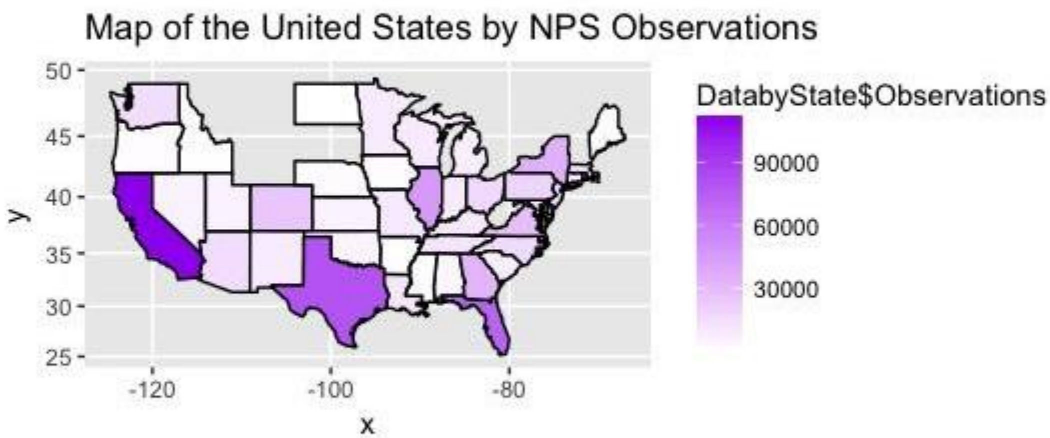
A. Using Descriptives Statistics to Choose our Focus

In order to understand the data, we then used descriptive statistics to inform the direction of our analysis.

1) **First we looked at the data broken down by state to determine which state to focus our analysis.**

```
DatabyState<-aggregate(DSData1$State_PL, by=list(DSData1$State_PL)   ,FUN=length)
View(DatabyState)
colnames(DatabyState) <- c("States", "Observations")
library(openintro)
library(maps)
library(ggplot2)
#install.packages("ggmap")
library(ggmap)
#make state names lowercase
e <- tolower(DatabyState$States)
DatabyState$stateName <- tolower(DatabyState$States)
#create data frame with state data for the US
us <-map_data("state")
```

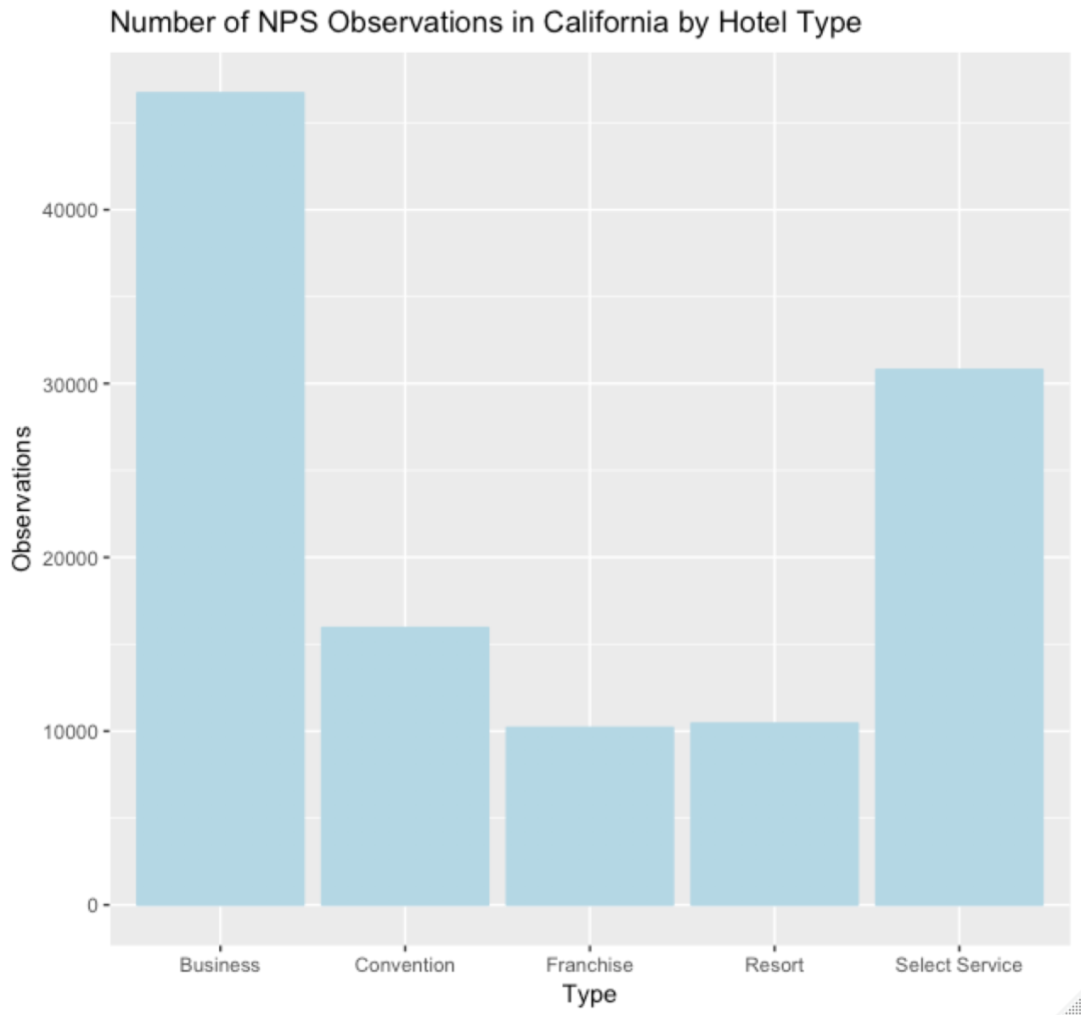
```
#Map of the United States with color representing Observations
map.simple <- ggplot(DatabyState, aes(map_id=stateName))
map.simple <-map.simple + geom_map(map=us, aes(fill=DatabyState$Observations), color=
"black")
map.simple <- map.simple + scale_fill_gradient(low="white", high ="purple")
map.simple <- map.simple + expand_limits(x= us$long, y=us$lat)
map.simple <- map.simple + coord_map()+ggtitle("Map of the United States by NPS
Observations")
map.simple
```



2) Second, we looked at the California hotels to see which hotel type had the most NPS observations. As evidenced below, California had the most business hotels so we decided to narrow our focus to California business hotels.

```
Cal<-which((DSData1$State_PL)=="California")
Cal<-DSData1[c(Cal),]
View(Cal)

DatabyType<-aggregate(Cal$Type_PL, by=list(Cal$Type_PL)   ,FUN=length)
colnames(DatabyType)<-c("Type", "Observations")
Typebar<- ggplot(DatabyType) +aes(x=Type, y=Observations) + geom_col(color=" light blue",
fill="light blue")+ ggtitle("Number of NPS Promoter Observations in California by Hotel Type")
View(DatabyType)
Typebar
```

	Type	Observations
1	Business	46710
2	Convention	15938
3	Franchise	10191
4	Resort	10437
5	Select Service	30787

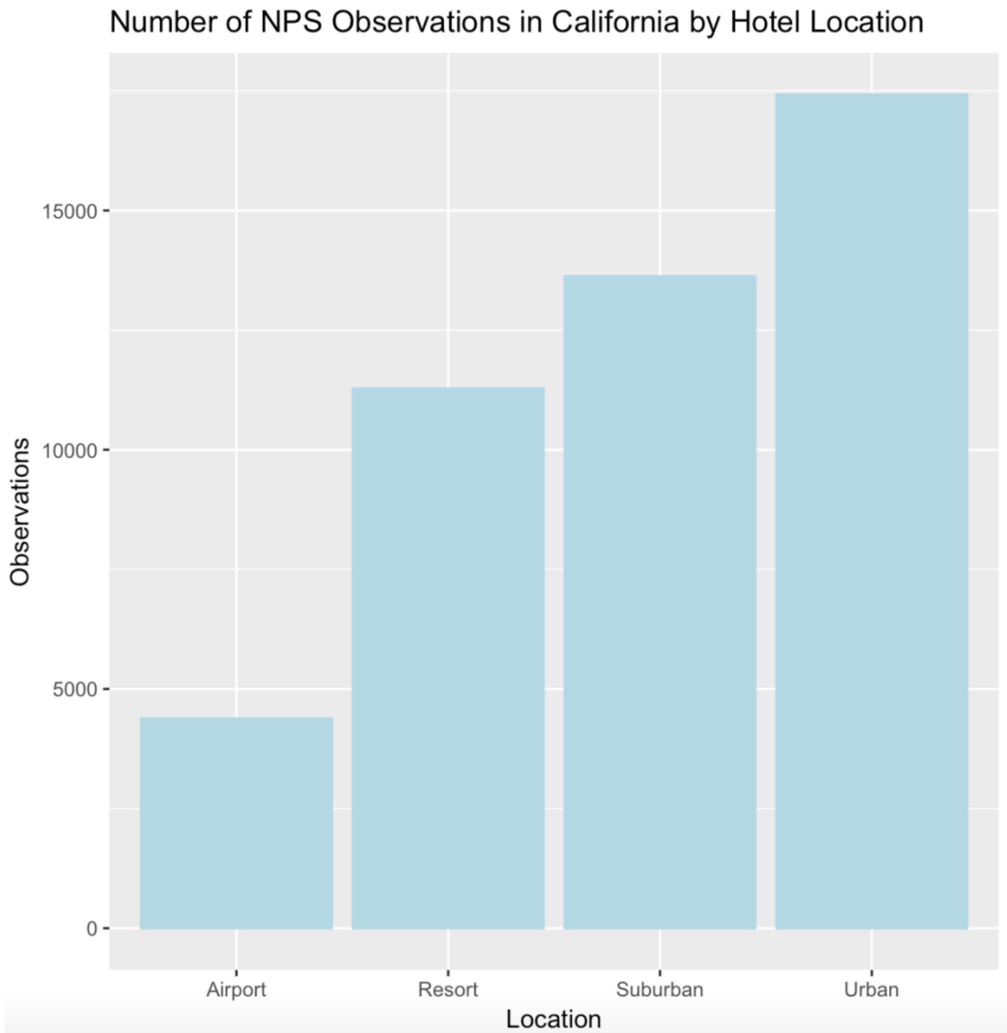


3) Third, we looked at the data within California business hotels broken down by location to determine which area to focus on. As evidenced below, the highest number of business hotels in California were in urban areas. Therefore, our analysis moving forward focuses on urban California business hotels.

```
DatabyLoc1<-aggregate(Cal1$Location_PL, by=list(Cal1$Location_PL)   ,FUN=length)
View(DatabyLoc1)
```

```
colnames(DatabyLoc1)<-c("Location", "Observations")
Locbar<- ggplot(DatabyLoc1) +aes(x=Location, y=Observations) + geom_col(color="light blue", fill="light blue ")+ ggtitle("Number of NPS Promoter Observations in California by Hotel Location")
LocbarBusrows<-which((Cal$Type_PL)=="Business")
CalBus<-Cal[c(Busrows),]
View(CalBus)

DatabyLoc1<-aggregate(CalBus$Location_PL, by=list(CalBus$Location_PL)   ,FUN=length)
View(DatabyLoc1)
colnames(DatabyLoc1)<-c("Location", "Observations")
Locbar<- ggplot(DatabyLoc1) +aes(x=Location, y=Observations) + geom_col(color="light blue", fill="light blue ")+ ggtitle("Number of NPS Promoter Observations in California by Hotel Location")
Locbar
```

	Group.1	x
1	Airport	4384
2	Resort	11278
3	Suburban	13625
4	Urban	17423



Modeling and Visualization

A. Linear Modeling

We performed linear modeling to determine which of the variables had the strongest relationship with likelihood to recommend and therefore would have the greatest influence on increasing NPS. As shown below all variables used aside from the Check in Satisfaction were significant and had a positive influence on LTR and NPS. The most influential were guest room and customer service satisfaction.

A one unit increase in guest room satisfaction leads to a 0.338 unit increase in likelihood to recommend, holding all else constant.

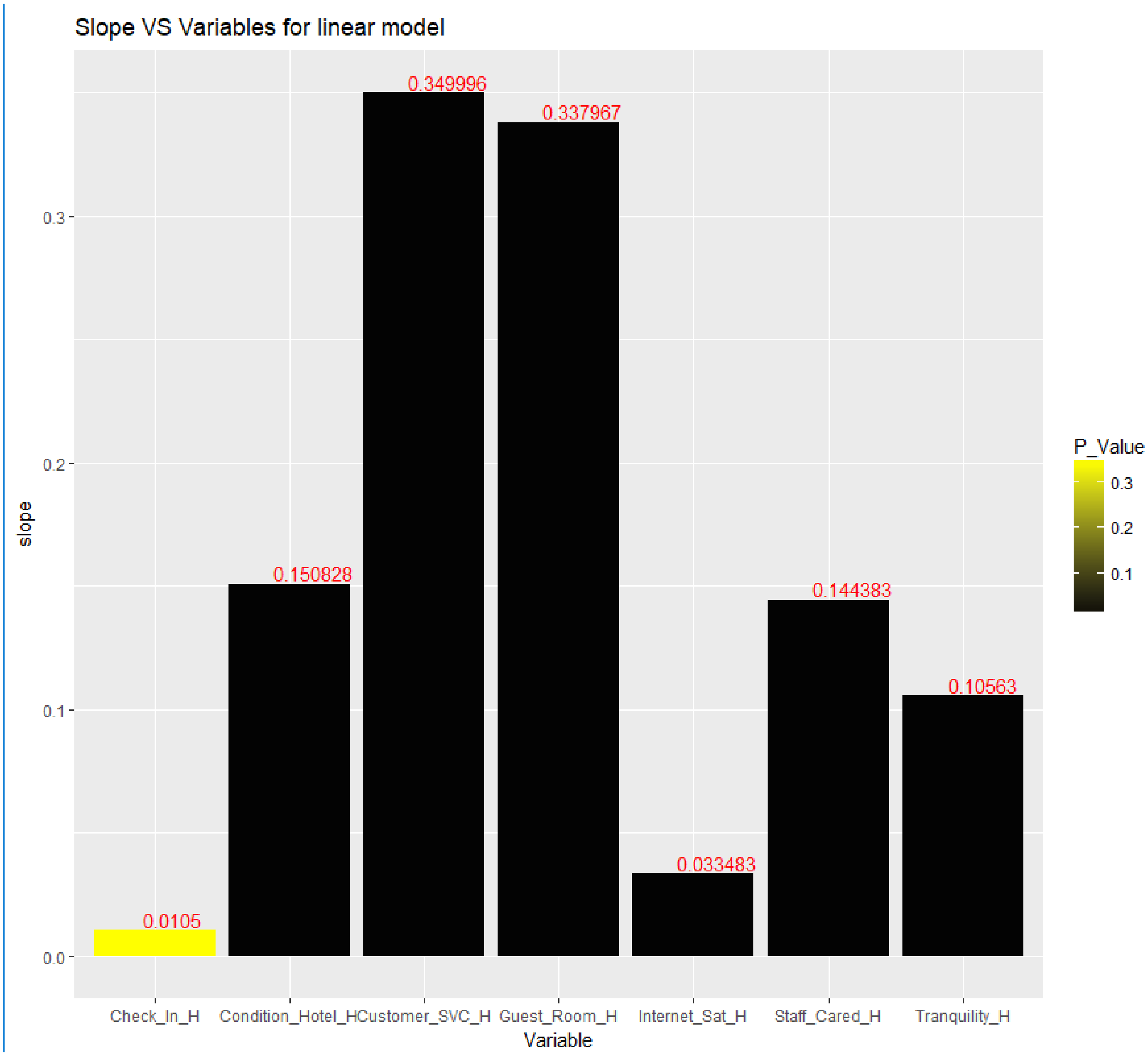
A one unit increase in customer service satisfaction leads to a 0.35 unit increase in likelihood to recommend, holding all else constant.

```
Regression_1<-fread(file="~/Desktop/Data Science/IST687-data/out-201501.csv",select=c(137,139:145,168,194,196,232))
Regression_2<-fread(file="~/Desktop/Data Science/IST687-data/out-201402.csv",select=c(137,139:145,168,194,196,232))
Regression_3<-fread(file="~/Desktop/Data Science/IST687-data/out-201403.csv",select=c(137,139:145,168,194,196,232))
Regression_4<-fread(file="~/Desktop/Data Science/IST687-data/out-201404.csv",select=c(137,139:145,168,194,196,232))
Regression_5<-fread(file="~/Desktop/Data Science/IST687-data/out-201405.csv",select=c(137,139:145,168,194,196,232))
Regression_6<-fread(file="~/Desktop/Data Science/IST687-data/out-201406.csv",select=c(137,139:145,168,194,196,232))
Regression_7<-fread(file="~/Desktop/Data Science/IST687-data/out-201407.csv",select=c(137,139:145,168,194,196,232))
Regression_8<-fread(file="~/Desktop/Data Science/IST687-data/out-201408.csv",select=c(137,139:145,168,194,196,232))
Regression_9<-fread(file="~/Desktop/Data Science/IST687-data/out-201409.csv",select=c(137,139:145,168,194,196,232))
Regression_10<-fread(file="~/Desktop/Data Science/IST687-data/out-201410.csv",select=c(137,139:145,168,194,196,232))
Regression_11<-fread(file="~/Desktop/Data Science/IST687-data/out-201411.csv",select=c(137,139:145,168,194,196,232))
Regression_12<-fread(file="~/Desktop/Data Science/IST687-data/out-201412.csv",select=c(137,139:145,168,194,196,232))

RegressionData<-rbind(Regression_1,Regression_2,Regression_3,Regression_4,Regression_5,Regression_6,Regression_7,Regression_8,Regression_9,Regression_10,Regression_11,Regression_12)

RegressionData[RegressionData==""]<-NA
RegressionData1<-na.omit(RegressionData)
View(RegressionData1)

RegressionRow<-which((RegressionData1$State_PL=="California") & (RegressionData1$Type_PL=="Business") & (RegressionData1$Location_PL=="Urban"))
RegressionData2<-RegressionData1[c(RegressionRow),]
View(RegressionData2)

RegressionData2$Promoter<-0
```

```
ProRow1<-which((RegressionData2$NPS_Type)=="Promoter")
RegressionData2$Promoter[c(ProRow1)]<-1

RegressionData2$Detractor<-0
DetRow1<-which((RegressionData2$NPS_Type)=="Detractor")
RegressionData2$Detractor[c(DetRow1)]<-1

install.packages("mfx")
library(mfx)

Lin1<-
lm(formula=RegressionData2$Likelihood_Recommend_H~RegressionData2$Guest_Room_H+
RegressionData2$Tranquility_H+RegressionData2$Condition_Hotel_H+RegressionData2$Cust
omer_SVC_H+RegressionData2$Staff_Cared_H+RegressionData2$Internet_Sat_H+Regression
Data2$Check_In_H, data=RegressionData2)
summary(Lin1)
```

```
Call:
lm(formula = RegressionData2$Likelihood_Recommend_H ~ RegressionData2$Guest_Room_H +
    RegressionData2$Tranquility_H + RegressionData2$Condition_Hotel_H +
    RegressionData2$Customer_SVC_H + RegressionData2$Staff_Cared_H +
    RegressionData2$Internet_Sat_H + RegressionData2$Check_In_H,
    data = RegressionData2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6832 -0.2556  0.0722  0.4457  8.8009

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.366575   0.108311 -12.617  < 2e-16 ***
RegressionData2$Guest_Room_H      0.337967   0.013823  24.450  < 2e-16 ***
RegressionData2$Tranquility_H     0.105630   0.008391  12.589  < 2e-16 ***
RegressionData2$Condition_Hotel_H 0.150828   0.016129   9.351  < 2e-16 ***
RegressionData2$Customer_SVC_H    0.349996   0.019052  18.371  < 2e-16 ***
RegressionData2$Staff_Cared_H     0.144383   0.015081   9.574  < 2e-16 ***
RegressionData2$Internet_Sat_H    0.033483   0.005780   5.793 7.28e-09 ***
RegressionData2$Check_In_H        0.010500   0.011170   0.940    0.347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.026 on 6122 degrees of freedom
Multiple R-squared:  0.6462,    Adjusted R-squared:  0.6458
F-statistic:  1597 on 7 and 6122 DF,  p-value: < 2.2e-16
```

```
slope<-c(0.337967,0.105630,0.150828,0.349996,0.144383,0.033483,0.010500)
Variable<-
c('Guest_Room_H','Tranquility_H','Condition_Hotel_H','Customer_SVC_H','Staff_Cared_H','Int
ernet_Sat_H','Check_In_H')
P_Value<-c(2e-16,2e-16,2e-16,2e-16,2e-16,7.28e-9,0.347)
dflm<-data.frame(Variable,slope,P_Value)

plotlm<-
ggplot(data=dflm,aes(x=Variable,y=slope))+geom_col(aes(fill=P_Value))+scale_fill_gradient(lo
w='grey1',high = 'yellow1')+ggtitle("Slope VS Variables for linear
model")+geom_text(aes(label=slope),colour="red",hjust=0.2, vjust=-0.2)
plotlm
```



B. Probit Modeling

We also performed probit modeling to look at the same variables' relationship with NPS type. We did probit modeling for promoters and detractors. Our findings confirmed what we found in our linear modeling, that all variables aside from Check in satisfaction were significant and positive. Again, Guest Room satisfaction and Customer Service satisfaction had the highest correlation.

A one unit increase in guest room satisfaction scores leads to a 11.39% higher percentage point probability of being a promoter, holding all else constant.

A one unit increase in customer service satisfaction leads to a 11.05% higher percentage point probability of being a promoter, holding all else constant.

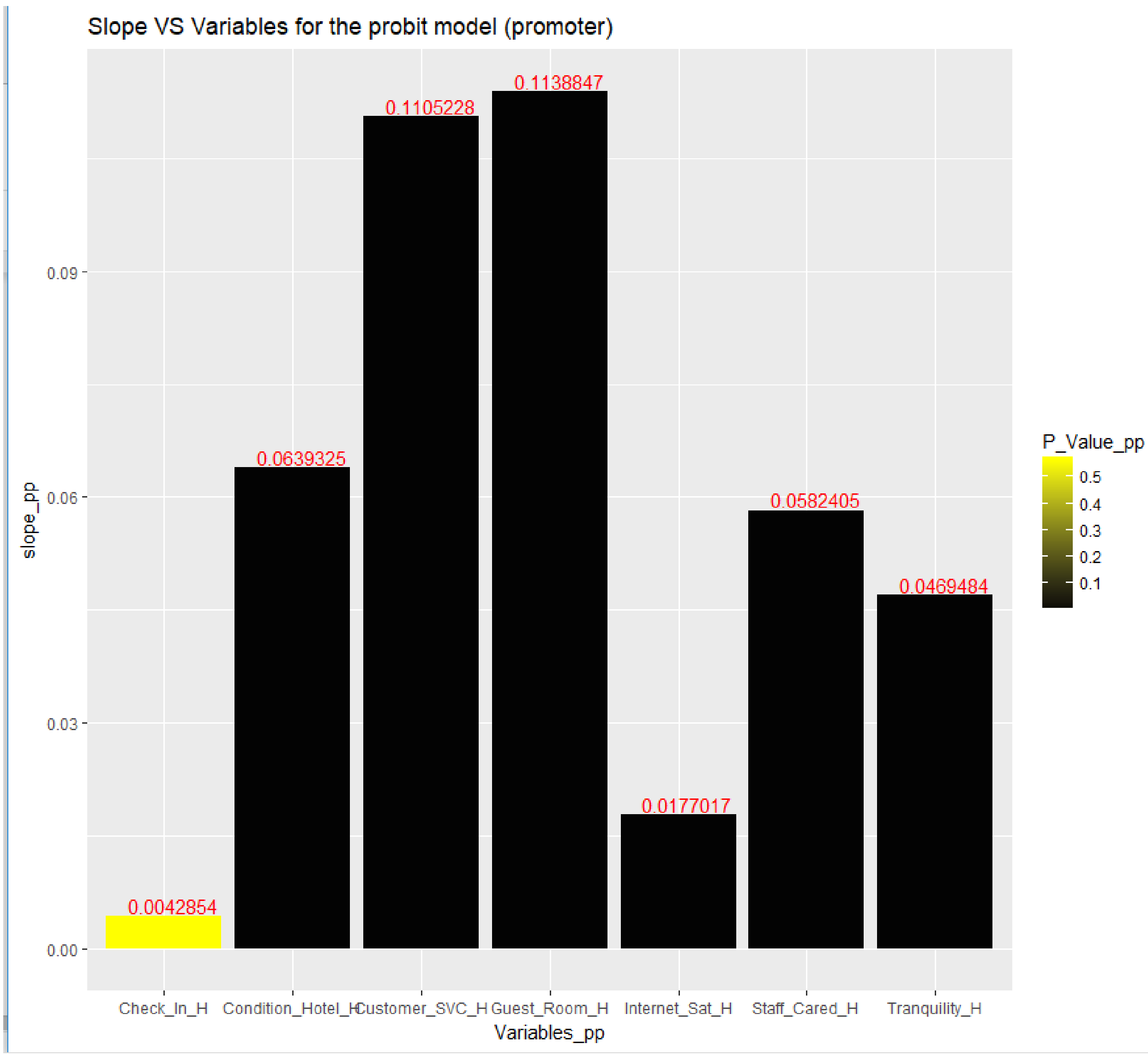
```
ProbPro<-
glm(RegressionData2$Promoter~RegressionData2$Guest_Room_H+RegressionData2$Tranquility_H+RegressionData2$Condition_Hotel_H+RegressionData2$Customer_SVC_H+RegressionData2$Staff_Cared_H+RegressionData2$Internet_Sat_H+RegressionData2$Check_In_H,
family=binomial(link="probit"),data=RegressionData2)
summary(ProbPro)
ProbProMfx<-
probitmfx(RegressionData2$Promoter~RegressionData2$Guest_Room_H+RegressionData2$Tranquility_H+RegressionData2$Condition_Hotel_H+RegressionData2$Customer_SVC_H+RegressionData2$Staff_Cared_H+RegressionData2$Internet_Sat_H+RegressionData2$Check_In_H,
data=RegressionData2,    atmean=TRUE, robust=TRUE)
ProbProMfx
```

```
Call:
probitmfx(formula = RegressionData2$Promoter ~ RegressionData2$Guest_Room_H +
    RegressionData2$Tranquility_H + RegressionData2$Condition_Hotel_H +
    RegressionData2$Customer_SVC_H + RegressionData2$Staff_Cared_H +
    RegressionData2$Internet_Sat_H + RegressionData2$Check_In_H,
    data = RegressionData2, atmean = TRUE, robust = TRUE)

Marginal Effects:
                                    dF/dx Std. Err.       z     P>|z|
RegressionData2$Guest_Room_H      0.1138847 0.0113408 10.0420 < 2.2e-16 ***
RegressionData2$Tranquility_H     0.0469484 0.0050536  9.2900 < 2.2e-16 ***
RegressionData2$Condition_Hotel_H 0.0639325 0.0114428  5.5871 2.309e-08 ***
RegressionData2$Customer_SVC_H    0.1105228 0.0143517  7.7010 1.350e-14 ***
RegressionData2$Staff_Cared_H     0.0582405 0.0088544  6.5776 4.782e-11 ***
RegressionData2$Internet_Sat_H    0.0177017 0.0031263  5.6623 1.494e-08 ***
RegressionData2$Check_In_H        0.0042854 0.0073517  0.5829      0.56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> PseudoR2(ProbPro)
     McFadden   Adj.McFadden      Cox.Snell     Nagelkerke McKelvey.Zavoina         Effron          Count
    0.4404376      0.4380989      0.4247740      0.5940219        0.6876500      0.5084450      0.8512235
    Adj.Count            AIC  Corrected.AIC
    0.5368207   4322.6912192   4322.7147448
```

```
slope_pp<-c(0.1138847,0.0469484,0.0639325,0.1105228,0.0582405,0.0177017,0.0042854)
```

```
Variables_pp<-
c("Guest_Room_H","Tranquility_H","Condition_Hotel_H","Customer_SVC_H","Staff_Cared_H","Internet_Sat_H","Check_In_H")
P_Value_pp<-c(2.2e-16,2.2e-16,2.309e-08,1.35e-14,4.782e-11,1.494e-08,0.56)

probitp<-data.frame(slope_pp,Variables_pp,P_Value_pp)

plotpp<-
ggplot(data=probitp,aes(x=Variables_pp,y=slope_pp))+geom_col(aes(fill=P_Value_pp))+scale_fill_gradient(low='grey1',high = 'yellow1')+ggtitle("Slope VS Variables for the probit model (promoter)")+geom_text(aes(label=slope_pp),colour="red",hjust=0.4, vjust=-0.2)
plotpp
```



ProbDet<-glm(RegressionData2$Detractor~RegressionData2$Guest_Room_H+RegressionData2$Tranquility_H+RegressionData2$Condition_Hotel_H+RegressionData2$Customer_SVC_H+RegressionData2$Staff_Cared_H+RegressionData2$Internet_Sat_H+RegressionData2$Check_In_H, family=binomial(link="probit"),data=RegressionData2)
ProbDetMfx<-probitmfx(RegressionData2$Detractor~RegressionData2$Guest_Room_H+RegressionData2$Tranquility_H+RegressionData2$Condition_Hotel_H+RegressionData2$Customer_SVC_H+RegressionData2$Staff_Cared_H+RegressionData2$Internet_Sat_H+RegressionData2$Check_In_H, data=RegressionData2,
atmean=TRUE, robust=TRUE)
ProbDetMfx

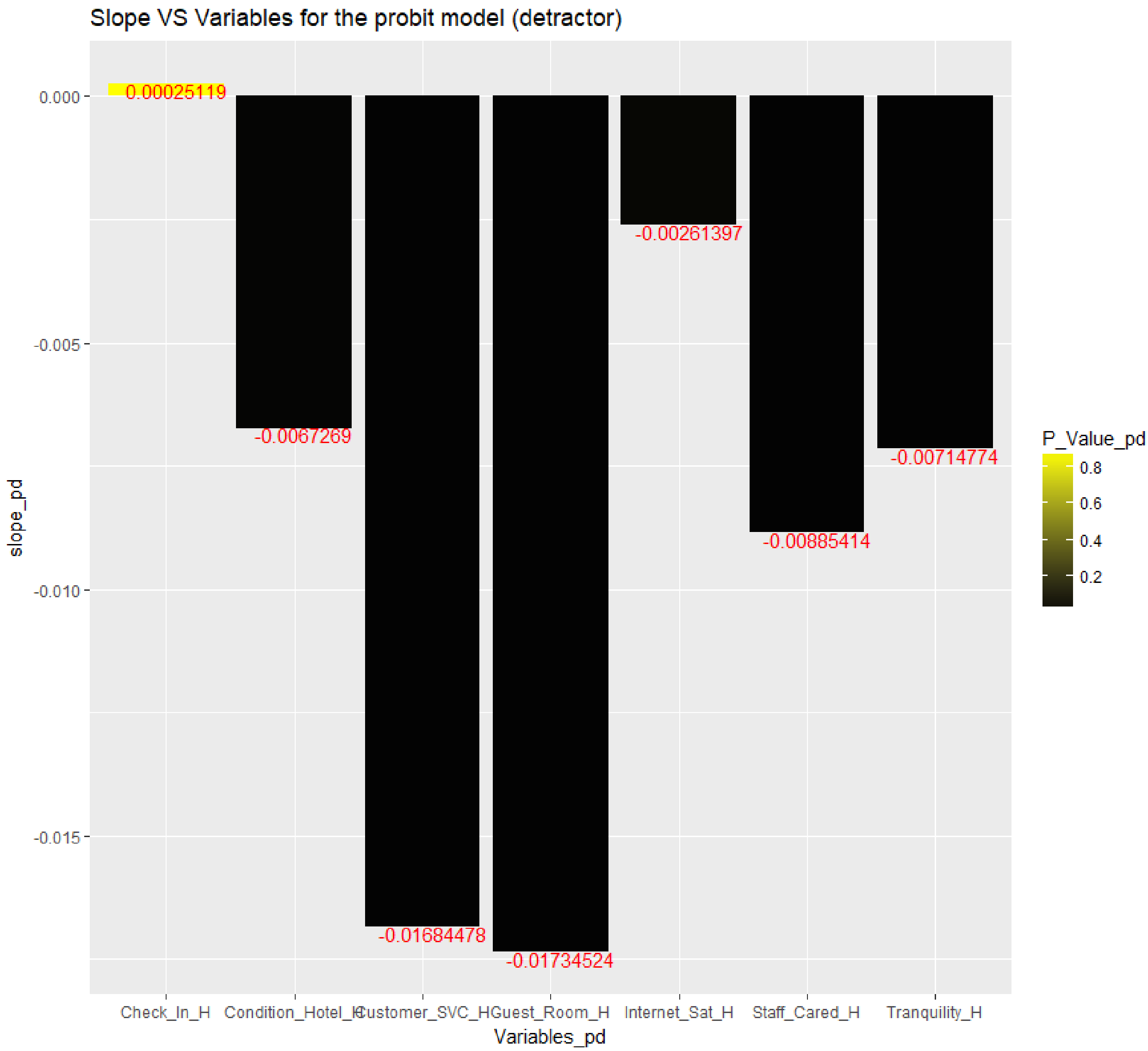
```
Call:
probitmfx(formula = RegressionData2$Detractor ~ RegressionData2$Guest_Room_H +
    RegressionData2$Tranquility_H + RegressionData2$Condition_Hotel_H +
    RegressionData2$Customer_SVC_H + RegressionData2$Staff_Cared_H +
    RegressionData2$Internet_Sat_H + RegressionData2$Check_In_H,
    data = RegressionData2, atmean = TRUE, robust = TRUE)
```

```
Marginal Effects:
                                      dF/dx   Std. Err.       z     P>|z|
RegressionData2$Guest_Room_H      -0.01734524 0.00205118 -8.4562 < 2.2e-16 ***
RegressionData2$Tranquility_H     -0.00714774 0.00142190 -5.0269 4.985e-07 ***
RegressionData2$Condition_Hotel_H -0.00672690 0.00233216 -2.8844  0.003922 **
RegressionData2$Customer_SVC_H    -0.01684478 0.00287850 -5.8519 4.859e-09 ***
RegressionData2$Staff_Cared_H     -0.00885414 0.00225946 -3.9187 8.903e-05 ***
RegressionData2$Internet_Sat_H    -0.00261397 0.00100487 -2.6013  0.009287 **
RegressionData2$Check_In_H         0.00025119 0.00187163  0.1342  0.893236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> PseudoR2(ProbDet)
       McFadden     Adj.McFadden        Cox.Snell       Nagelkerke McKelvey.Zavoina           Effron            Count
      0.4417886        0.4371320        0.2431486        0.5198555        0.4906723        0.4261897        0.9365416
      Adj.Count              AIC    Corrected.AIC
      0.3361775     2173.7818976     2173.8054232
```

slope_pd<-c(-0.01734524,-0.00714774,-0.00672690,-0.01684478,-0.00885414 ,-0.00261397,0.00025119)

```
Variables_pd<-
c("Guest_Room_H","Tranquility_H","Condition_Hotel_H","Customer_SVC_H","Staff_Cared_
H","Internet_Sat_H","Check_In_H")
P_Value_pd<-c(2.2e-16,4.985e-07,0.003922,4.859e-09,8.903e-05,0.009287, 0.893236)

probitd<-data.frame(slope_pd,Variables_pd,P_Value_pd)

plotpd<-
ggplot(data=probitd,aes(x=Variables_pd,y=slope_pd))+geom_col(aes(fill=P_Value_pd))+scale_
fill_gradient(low='grey1',high = 'yellow1')+ggtitle("Slope VS Variables for the probit model
(detractor)")+geom_text(aes(label=slope_pd),colour="red",hjust=0.4, vjust=0.9)
plotpd
```



C. Association Rules

Original we used 27 independent variables for our Arules analysis. However, after running analysis we found that many of the variables were either all yes or no and therefore offered no interesting information for the analysis. Thus, we eliminated these variables from the analysis and were left with 9 independent variables. We used Arules to look at these variables' effect on NPS type including Boutique hotel type, existence of a business center, having a convention center, having a limo service, having a mini bar, having an outdoor pool, having a Regency Grand Club, having Self Parking and having Spa Services in the Fitness Center. However, after inspecting our rules we found that several variables provided the same right hand side information. Therefore, we limited our analysis to only one of these variables so that it would not be redundant. For the next analysis, only the Boutique, Limo Service, Mini Bar, Self Parking and Spa Services in the fitness center variables were included. However, even with this more focused dataset, we did not find significant conclusions.

```
RULE_1<-fread(file="~/Desktop/intro to data/project/dataset/out-
201501.csv",select=c(168,194,196,199:227,232))
RULE_2<-fread(file="~/Desktop/intro to data/project/dataset/out-
201402.csv",select=c(168,194,196,199:227,232))
RULE_3<-fread(file="~/Desktop/intro to data/project/dataset/out-
201403.csv",select=c(168,194,196,199:227,232))
RULE_4<-fread(file="~/Desktop/intro to data/project/dataset/out-
201404.csv",select=c(168,194,196,199:227,232))
RULE_5<-fread(file="~/Desktop/intro to data/project/dataset/out-
201405.csv",select=c(168,194,196,199:227,232))
RULE_6<-fread(file="~/Desktop/intro to data/project/dataset/out-
201406.csv",select=c(168,194,196,199:227,232))
RULE_7<-fread(file="~/Desktop/intro to data/project/dataset/out-
201407.csv",select=c(168,194,196,199:227,232))
RULE_8<-fread(file="~/Desktop/intro to data/project/dataset/out-
201408.csv",select=c(168,194,196,199:227,232))
RULE_9<-fread(file="~/Desktop/intro to data/project/dataset/out-
201409.csv",select=c(168,194,196,199:227,232))
RULE_10<-fread(file="~/Desktop/intro to data/project/dataset/out-
201410.csv",select=c(168,194,196,199:227,232))
RULE_11<-fread(file="~/Desktop/intro to data/project/dataset/out-
201411.csv",select=c(168,194,196,199:227,232))
RULE_12<-fread(file="~/Desktop/intro to data/project/dataset/out-
201412.csv",select=c(168,194,196,199:227,232))
```

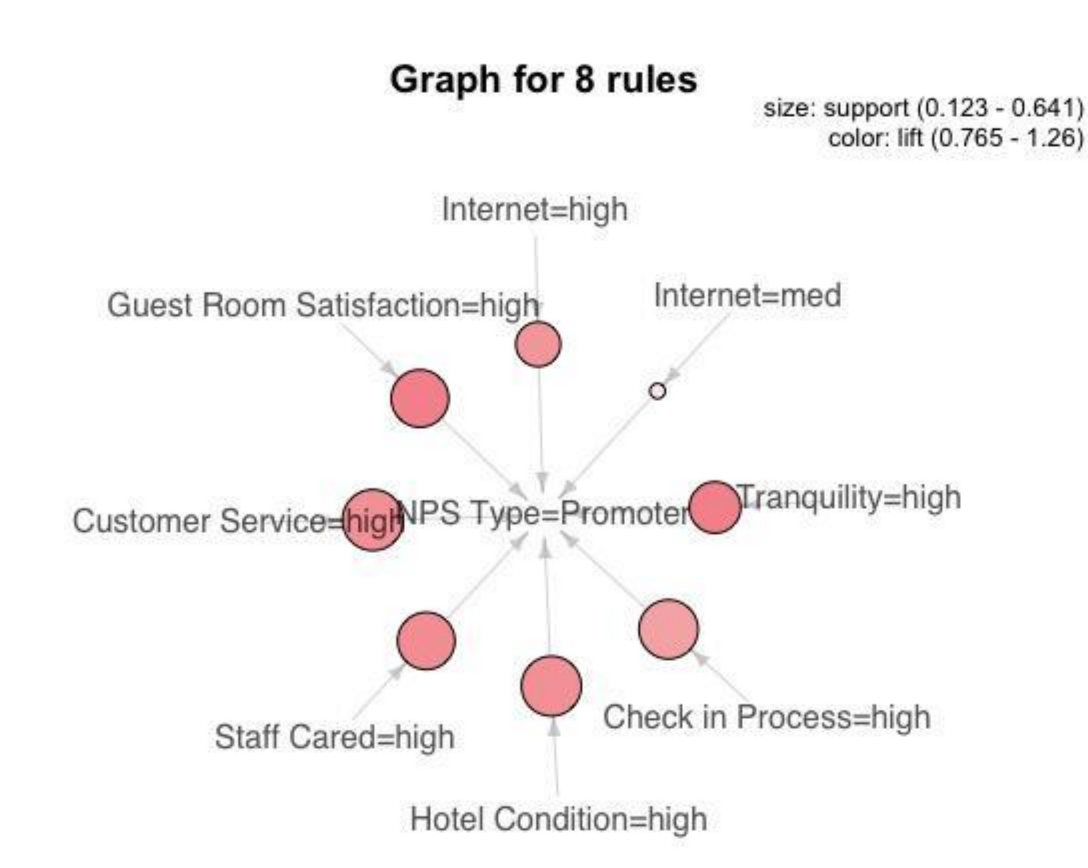
```
library(arulesViz)

RULEData<-
rbind(RULE_1,RULE_2,RULE_3,RULE_4,RULE_5,RULE_6,RULE_7,RULE_8,RULE_9,RU
LE_10,RULE_11,RULE_12)
summary(RULE_1)

summary(RULEData)

RULERow<-which((RULEData$State_PL=="California") &
(RULEData$Type_PL=="Business") & (RULEData$Location_PL=="Urban"))
RULEData1<-RULEData[c(RULERow),]
summary(RULEData1)

RULEData1[,4:33]<-lapply(RULEData1[,4:33],as.factor) # shift character to factor
summary(RULEData1)
RULEData2<-RULEData1[,-c(30,31)] # remove columns with all NAs
summary(RULEData2)
RULEData2[RULEData2==""]<-NA
RULEData3<-na.omit(RULEData2)
View(RULEData3)
summary(RULEData3)

RULEData4<-RULEData3[,c(6,7,10,18,19,21,22,25,29,31)] # select columns with both N and Y
responds
summary(RULEData4)
View(RULEData4)

install.packages("arules")
library(arules)
install.packages("arulesViz")
library(arulesViz)
RulesetPro <-apriori(RULEData4,parameter = list(support=0.1,
confidence=0.6,minlen=2,maxlen=2),appearance = list(rhs="NPS_Type=Promoter"))
summary(RulesetPro)
inspect(RulesetPro)
# We noticed that with in some rules there are the same support, confidence and lift with
different left hand sides

View(RULEData4[Boutique_PL=="N"])
```

```
summary(RULEData4[Boutique_PL=="N"])
summary(RULEData4[Boutique_PL=="Y"])

summary(RULEData4[`Spa services in fitness center_PL`=="Y"])

summary(RULEData4[`Mini-Bar_PL`=="N"])
# boutique=business center=convention=regency grand club
# mini bar=outdoor pool

RULEData5<-RULEData3[,c(18,19,25,29,31)]
summary(RULEData5)
View(RULEData5)
# boutique(Y)=limo(Y)  boutique(Y)=mini bar(Y)  boutique(Y)=self parking(N)
boutique(Y)=spa(N)
# limo(Y)=spa(N) mini bar(N)=spa(N)
RulesetPro1 <-apriori(RULEData5,parameter = list(support=0.1,
confidence=0.5,minlen=2,maxlen=2),appearance = list(rhs="NPS_Type=Promoter"))
summary(RulesetPro1)
inspect(RulesetPro1)
inspect(sort(RulesetPro1,decreasing=TRUE,by="confidence"))
```

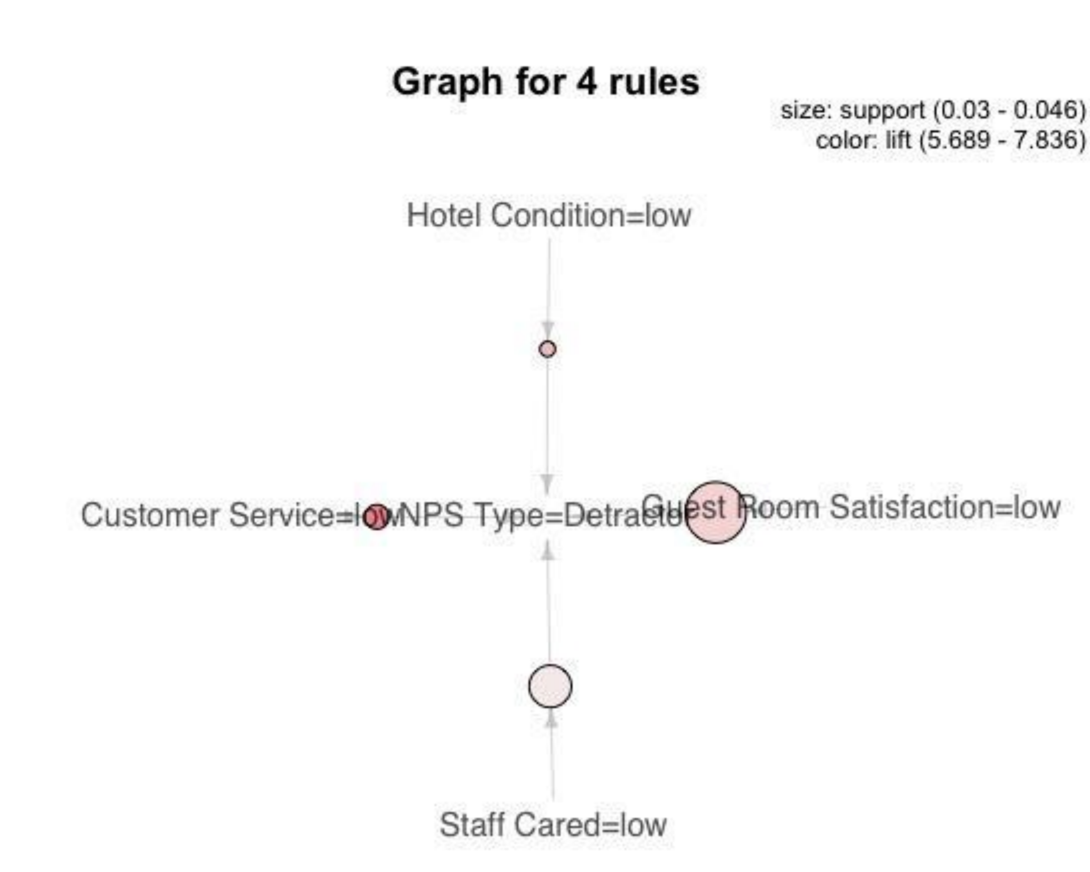
In an effort to find significant results, we changed the left hand side from the previous dummy variables to the variables used in the regression analysis. In order to make these variables usable, we had to categorize the numeric variables into categorical variables. We took satisfaction scores that were on a 1-10 scale and categorized them into “low”, “medium” and “high” groups. This allowed us to conduct arules modeling. We found that customer service satisfaction and hotel condition had the highest confidence and lift for low levels or detractors. Additionally, we found that customer satisfaction and guest room condition had the highest confidence and lift for medium levels or passive. Lastly, we found that guest room satisfaction had the highest confidence and lift for high levels or promoters. Customer satisfaction and hotel condition only had slightly lower confidence and should still be considered a significant result. Therefore, the most important takeaways from this modeling are to increase the customer service, hotel condition and guest room condition.

```
# Because of the fact that we are using the same columns as the regression dataframe,we use the
same dataset for the regression data.

RULEData_1<-RegressionData2

summary(RULEData_1)
```

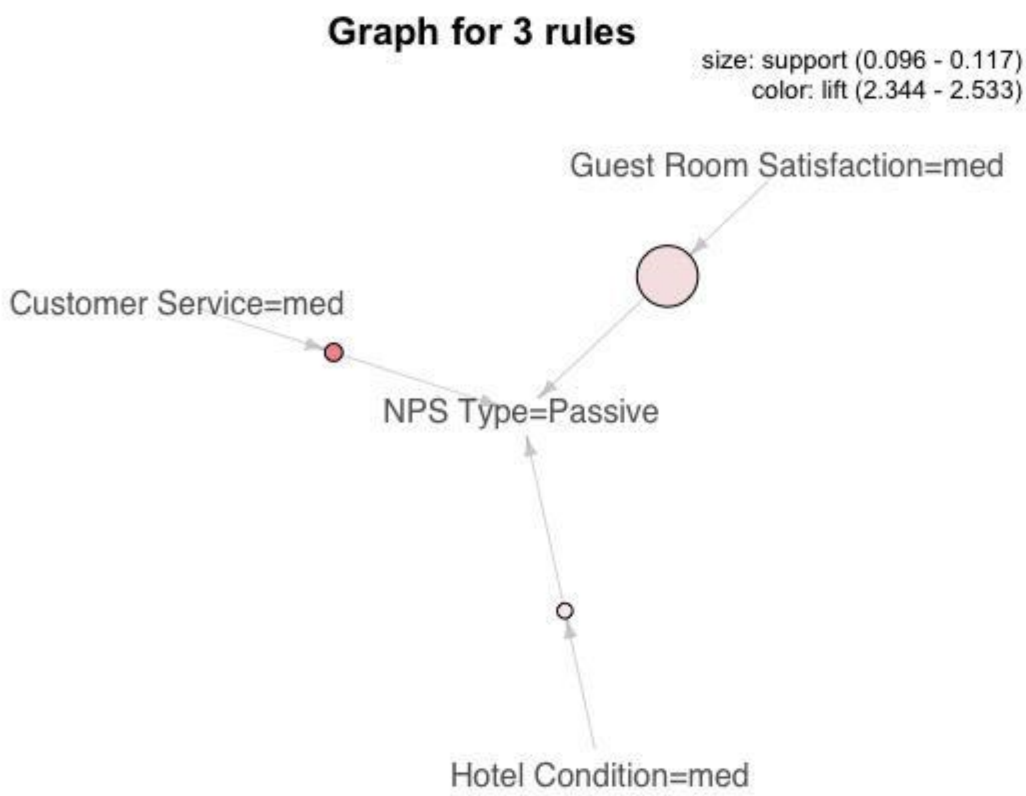
```
View(RULEData_1)
RULEData_1<-RULEData_1[,c(2:8,12)]
RULEData_1[,1:7]<-lapply(RULEData_1[,1:7],as.numeric)
str(RULEData_1)

RuleFunction<-function(c)
{
c[c>=1&c<=6]<-"low"
c[c>=7&c<=8]<-"med"
c[c==9|c==10]<-"high"
return(c)
}
RULEData_2<-
data.frame(RuleFunction(RULEData_1$Guest_Room_H),RuleFunction(RULEData_1$Tranquili
ty_H),RuleFunction(RULEData_1$Condition_Hotel_H),RuleFunction(RULEData_1$Customer
_SVC_H),RuleFunction(RULEData_1$Staff_Cared_H),RuleFunction(RULEData_1$Internet_S
at_H),RuleFunction(RULEData_1$Check_In_H),RULEData_1$NPS_Type)
View(RULEData_2)
colnames(RULEData_2)<-c("Guest Room Satisfaction","Tranquility","Hotel
Condition","Customer Service","Staff Cared","Internet","Check in Process","NPS Type")
RulesetPro_1 <-apriori(RULEData_2,parameter = list(support=0.1,
confidence=0.5,minlen=2,maxlen=2),appearance = list(rhs="NPS Type=Promoter"))

summary(RulesetPro_1)
inspect(RulesetPro_1)
topConfidencePro<-sort(RulesetPro_1,decreasing=TRUE,by="confidence")
inspect(topConfidencePro)
plot((topConfidencePro), method="graph", control=list(type="items"))
```




```
RulesetDet_1<-apriori(RULEData_2,parameter = list(support=0.005,
confidence=0.2,minlen=2,maxlen=2),appearance = list(rhs="NPS Type=Detractor"))
summary(RulesetDet_1)
inspect(RulesetDet_1)
topConfidenceDet<-sort(RulesetDet_1,decreasing=TRUE,by="confidence")
inspect(topConfidenceDet)
plot((topConfidenceDet), method="graph", control=list(type="items"))
```




```
RulesetPas_1 <-apriori(RULEData_2,parameter = list(support=0.01,
confidence=0.4,minlen=2,maxlen=2),appearance = list(rhs="NPS Type=Passive"))
summary(RulesetPas_1)
inspect(RulesetPas_1)
topConfidencePas<-sort(RulesetPas_1,decreasing=TRUE,by="confidence")
inspect(topConfidencePas)
plot((topConfidencePas), method="graph", control=list(type="items"))
```



Rules for Detractor

```
> inspect(topConfidenceDet)
    lhs                              rhs                    support    confidence lift     count
[1] {Customer Service=low}        => {NPS Type=Detractor} 0.03311582 0.7490775  7.835913 203
[2] {Hotel Condition=low}         => {NPS Type=Detractor} 0.02985318 0.6535714  6.836848 183
[3] {Guest Room Satisfaction=low} => {NPS Type=Detractor} 0.04616639 0.6021277  6.298707 283
[4] {Staff Cared=low}             => {NPS Type=Detractor} 0.03947798 0.5438202  5.688768 242
```

Rules for Passive

```
> inspect(topConfidencePas)
    lhs                              rhs                  support    confidence lift     count
[1] {Customer Service=med}        => {NPS Type=Passive} 0.09722675 0.5714286  2.532796 596
[2] {Guest Room Satisfaction=med} => {NPS Type=Passive} 0.11729201 0.5345725  2.369436 719
[3] {Hotel Condition=med}         => {NPS Type=Passive} 0.09608483 0.5287253  2.343519 589
```

Rules for Promoters

```
> inspect(topConfidencePro)
    lhs                              rhs                   support   confidence lift      count
[1] {Guest Room Satisfaction=high} => {NPS Type=Promoter} 0.6019576 0.8551564  1.2598195 3690
[2] {Tranquility=high}             => {NPS Type=Promoter} 0.5331158 0.8514851  1.2544109 3268
[3] {Staff Cared=high}             => {NPS Type=Promoter} 0.5990212 0.8249831  1.2153681 3672
[4] {Customer Service=high}        => {NPS Type=Promoter} 0.6411093 0.8160299  1.2021782 3930
[5] {Hotel Condition=high}         => {NPS Type=Promoter} 0.6282219 0.8131334  1.1979111 3851
[6] {Internet=high}                => {NPS Type=Promoter} 0.4566069 0.7990294  1.1771330 2799
[7] {Check in Process=high}        => {NPS Type=Promoter} 0.6199021 0.7659746  1.1284365 3800
[8] {Internet=med}                 => {NPS Type=Promoter} 0.1234910 0.5192044  0.7648938  757
```

Room Type Analysis

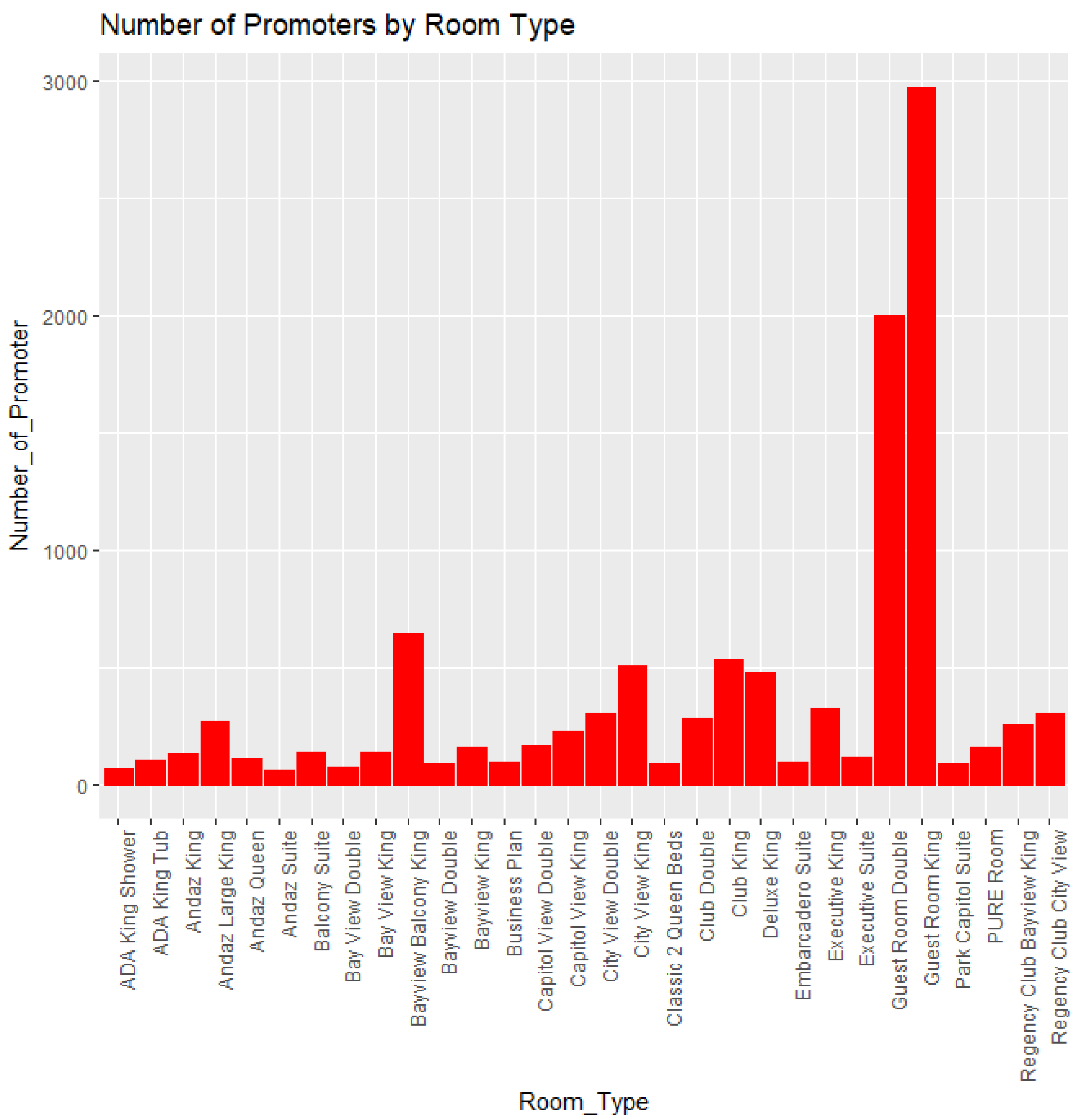
Due to finding guest room condition as a consistent significant result in our modeling, we decided to conduct further analysis on room type. We looked at the number of promoters, detractors and passive observations by room type.

```
l1<-aggregate(CalUrbanBus1$ROOM_TYPE_DESCRIPTION_C,
by=list(CalUrbanBus1$ROOM_TYPE_DESCRIPTION_C)   ,FUN=length)

pr1<-
tapply(CalUrbanBus1$NPS_Type=="Promoter",CalUrbanBus1$ROOM_TYPE_DESCRIPTIO
N_C, sum)
pr1<-data.frame(pr1)
pr1<-na.omit(pr1)

de1<-
tapply(CalUrbanBus1$NPS_Type=="Detractor",CalUrbanBus1$ROOM_TYPE_DESCRIPTIO
N_C, sum)
de1<-data.frame(de1)
de1<-na.omit(de1)

pa1<-
tapply(CalUrbanBus1$NPS_Type=="Passive",CalUrbanBus1$ROOM_TYPE_DESCRIPTION_
C, sum)
pa1<-data.frame(pa1)
pa1<-na.omit(pa1)

score<-aggregate(CalUrbanBus1$Likelihood_Recommend_H,
by=list(CalUrbanBus1$ROOM_TYPE_DESCRIPTION_C)   ,FUN=mean)
NPS1<-data.frame(l1,pr1,de1,pa1,score$x)
View(NPS1)
NPS1$prr1<-NPS1$pr1/NPS1$x
```

```
NPS1$der1<-NPS1$de1/NPS1$x
NPS1$par1<-NPS1$pa1/NPS1$x

colnames(NPS1)<-
c("Room_Type","Total_Number","Number_of_Promoter","Number_of_Detractor","Number_of
_Passive","LTR_Score","Rate_of_Promoter","Rate_of_Detractor","Rate_of_Passive")

View(NPS1)

NPS1<-NPS1[NPS1$Total_Number>100,]
View(NPS1)
```

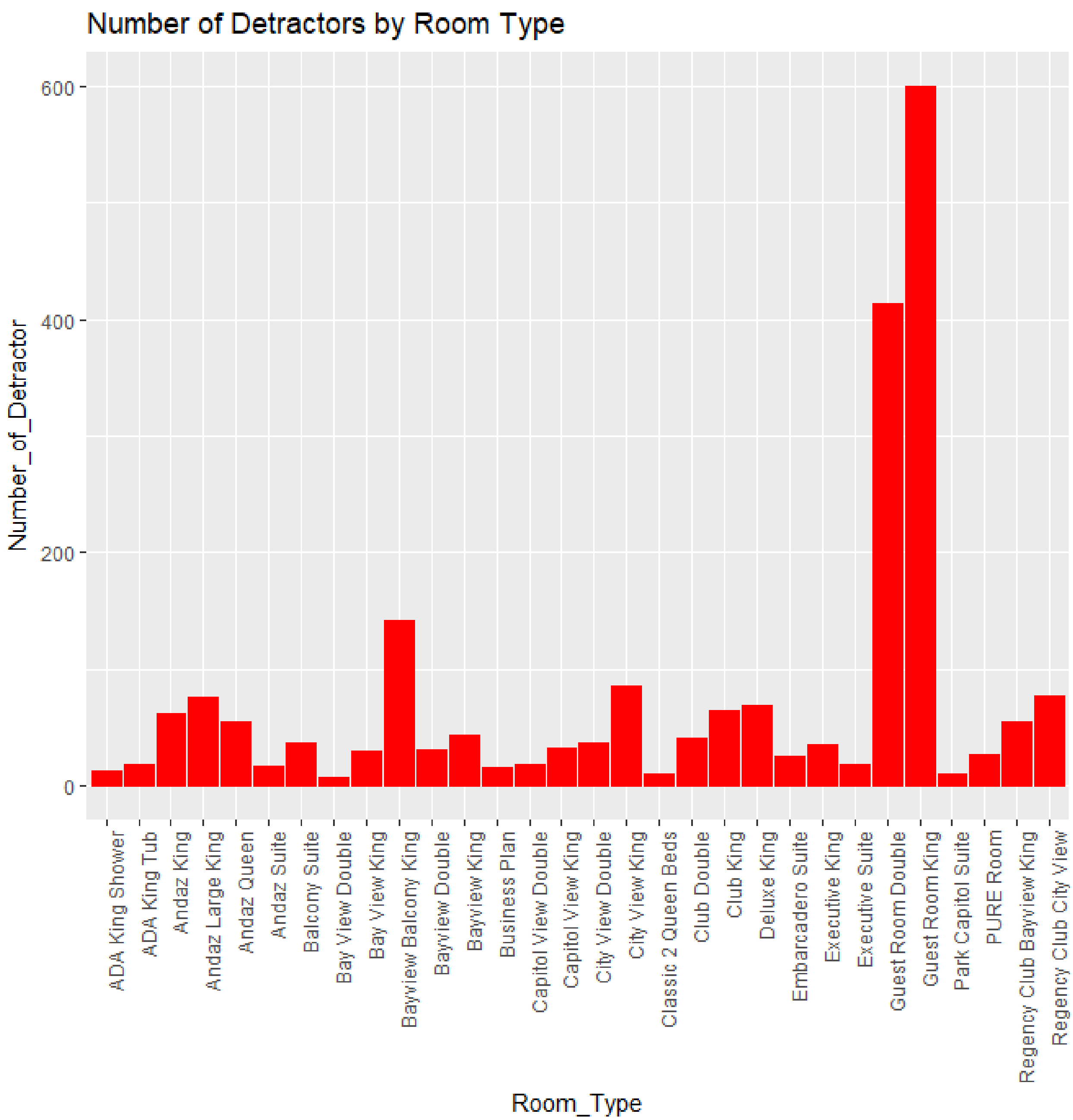
Number of Promoter Observations by Room Type

```
Data.1 <- ggplot(NPS1) + aes(x=Room_Type, y=Number_of_Promoter) +
geom_col(color="red", fill="red") + ggtitle("Number of Promoters by Room Type")
Data.1 <- Data.1 + theme(axis.text.x = element_text(angle=90, hjust = 1))
Data.1
```



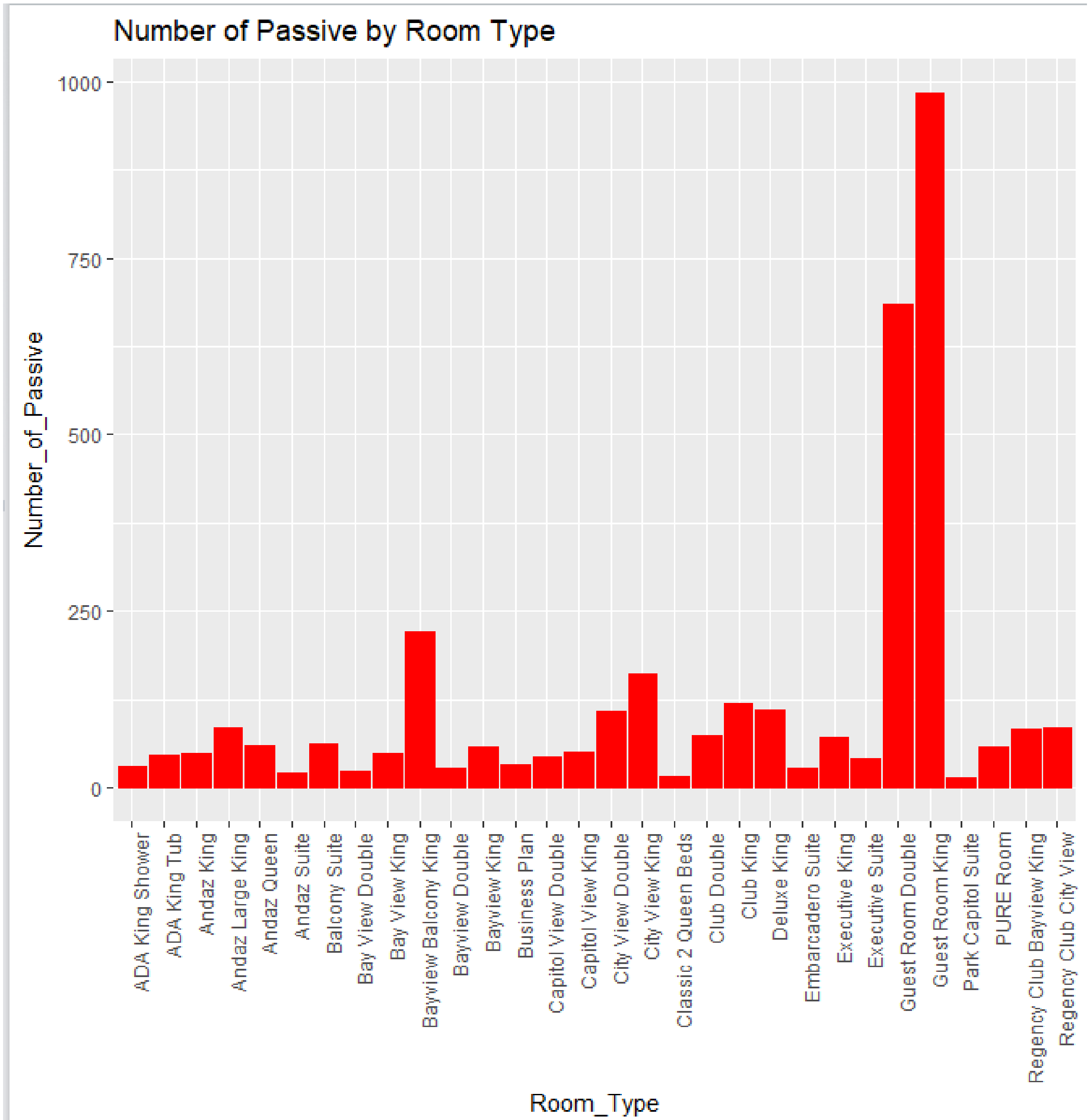
Number of Detractor Observations by Room Type

```
Data.2 <- ggplot(NPS1) + aes(x=Room_Type, y=Number_of_Detractor) +
geom_col(color="red", fill="red") + ggtitle("Number of Detracters by Room Type")
Data.2 <- Data.2 + theme(axis.text.x = element_text(angle=90, hjust = 1))
Data.2
```



Number of Passive Observations by Room Type

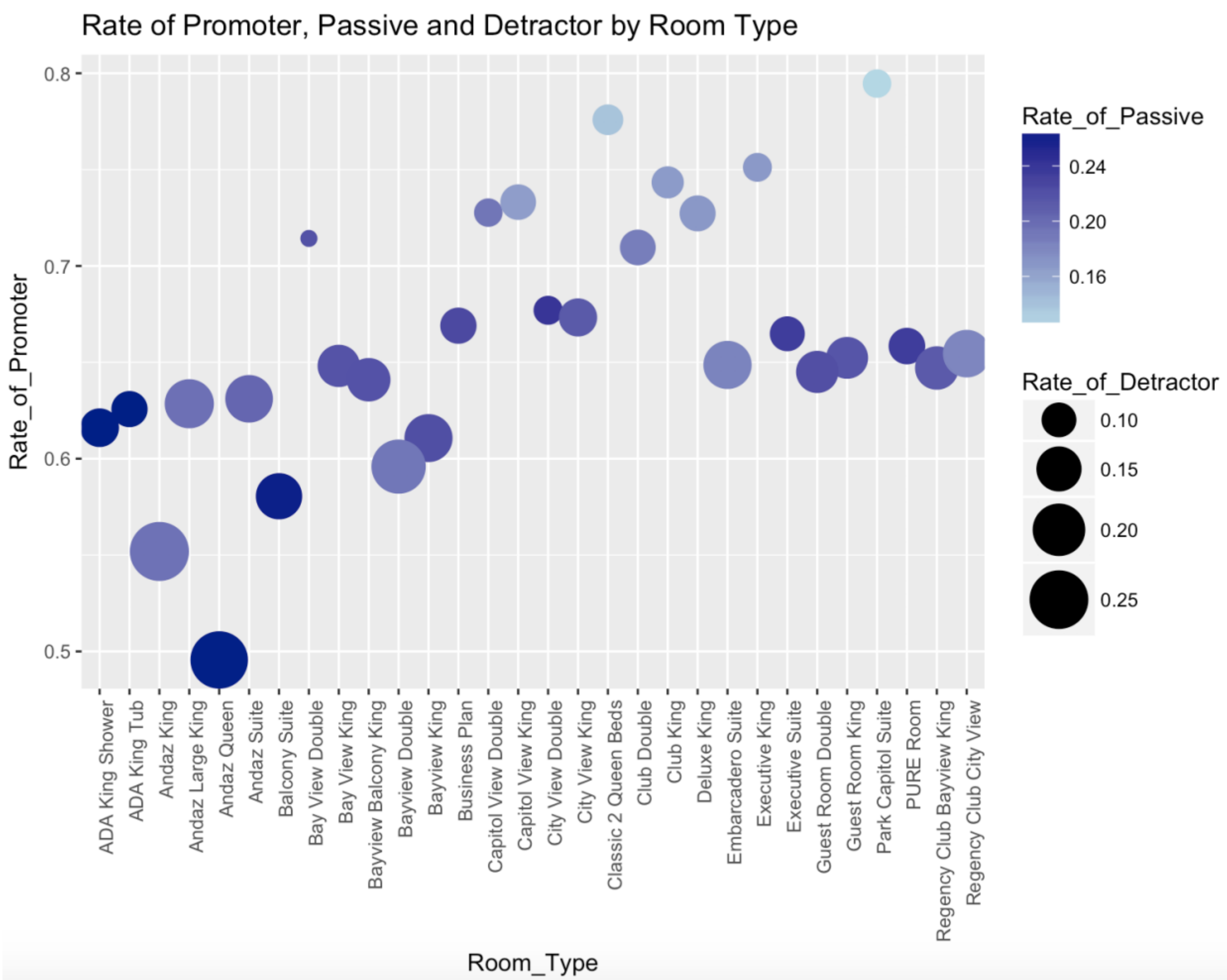
```
Data.3 <- ggplot(NPS1) + aes(x=Room_Type, y=Number_of_Passive) + geom_col(color="red",
fill="red") + ggtitle("Number of Passive by Room Type")
Data.3 <- Data.3 + theme(axis.text.x = element_text(angle=90, hjust = 1))
Data.3
```

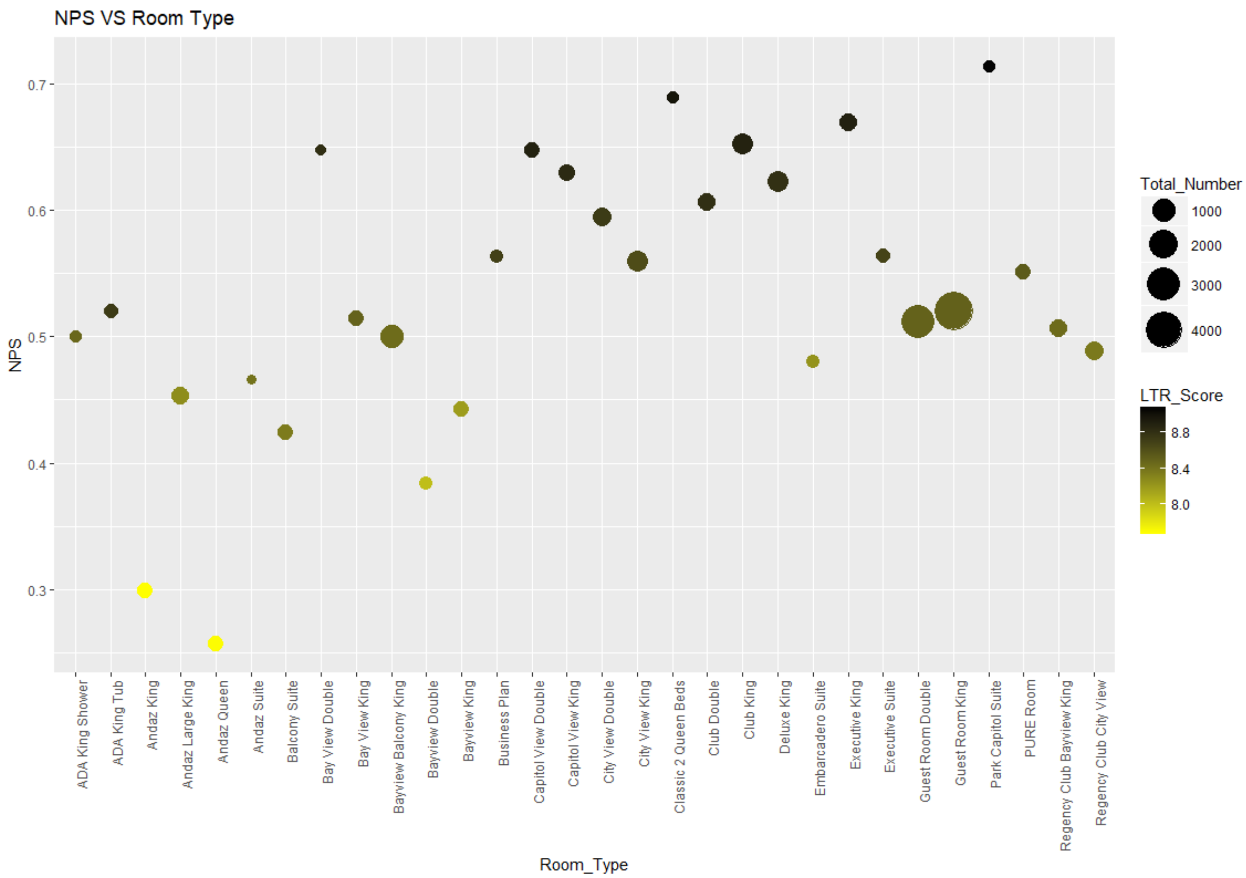


The number of promoter, passive and detractor observations did not provide the information we ultimately needed because it did not account for the frequency of observations to number of room types. Therefore, we needed to calculate the rate of promoters, passives and detractor observations by room type.

```
Rate<- ggplot(data=NPSsortD) +  geom_point(aes(x=Room_Type,
y=Rate_of_Promoter,col=Rate_of_Passive,size=Rate_of_Detractor))
```

```
+scale_color_gradient(low="light blue",high='dark blue')+scale_size(range = c(3,12))
+ggtitle("Rate of Promoter, Passive and Detractor by Room Type")
Rate<-Rate+theme(axis.text.x = element_text(angle=90, hjust = 1))
Rate
```






```
LHR<- ggplot(data=NPSsortL) +  geom_point(aes(x=Room_Type,
y=NPS,col=LTR_Score,size=Total_Number))
+scale_color_gradient(low="Yellow1",high='grey1')+scale_size(range = c(3,12)) +ggtitle("NPS
VS Room Type")
LHR<-LHR+theme(axis.text.x = element_text(angle=90, hjust = 1))
LHR
```

<u>Summary of Room Type Statistics Analysis</u>

1) Andaz Queen and Andaz King have the 2 lowest LTR and promoter rates and the highest detractor rates.
2) Park Capitol Suite, Classic 2 Queen Beds have the 2 highest LTR and promoter rates. They have relatively lowest passive and detractor rates.
3) Executive King and Club King also have very high LTR with a relative high demand. We also recommend increasing the number of these types of rooms.

Validation

A. Regression Modeling

In our linear model, the adjusted r squared is 0.6458. This means that 64.58% of the variance in the model is explained by Guest Room Satisfaction, Tranquility Satisfaction, Staff Care Satisfaction, Customer Service Satisfaction, Hotel Condition Satisfaction, Internet Satisfaction and Check in Satisfaction.

In our probit model, the Pseudo R squared is 0.68765. This means that 68.77% of the variance in our model is explained by Guest Room Satisfaction, Tranquility Satisfaction, Staff Care Satisfaction, Customer Service Satisfaction, Hotel Condition Satisfaction, Internet Satisfaction and Check in Satisfaction.

B. 1. KSVM Modeling for regression models

We used KSVM modeling to validate probit and linear regression modeling. As our probit and linear modeling predicted that guest room and customer satisfaction would have the biggest influence on NPS score, we used data for guest room and customer service satisfaction scores to predict the NPS type. We then compared the predictions to the actual data to validate our conclusions.

First, we divided the data into two groups, a train and test group for data. The cut point between the groups was ⅔ in the training group and ⅓ is in the test group. By using the KSVM model, we can see there are 4,875 promoters. 3,727 observations out of the 4,875 are predicted as promoters. There are 1,148(256+892) cases which were promoters but were not classified as promoters by KSVM model. Similarly, 162 observations out of 304 were classified as passive and 380 cases out of 513 were predicted as detractors. Therefore, by using the guest room and customer service satisfaction metric, we built a KSVM model with 75% prediction accuracy, while predicting the NPS. In terms of NPS score, the predicted NPS score is 59.56%, while the actual NPS score is 53.93%.

```
> table(RegCompTable)
           RegKsvmPred.1...
NPS_Type       0    1    2
  Detractor  256   94  380
  Passive    892  162  108
  Promoter  3727   48   25
```

RegSvm_1<-fread(file="~/Desktop/intro to data/project/dataset/out-201501.csv",select=c(168,194,196,139,141,142,232))

```
RegSvm_2<-fread(file="~/Desktop/intro to data/project/dataset/out-201402.csv",select=c(168,194,196,139,141,142,232))
RegSvm_3<-fread(file="~/Desktop/intro to data/project/dataset/out-201403.csv",select=c(168,194,196,139,141,142,232))
RegSvm_4<-fread(file="~/Desktop/intro to data/project/dataset/out-201404.csv",select=c(168,194,196,139,141,142,232))
RegSvm_5<-fread(file="~/Desktop/intro to data/project/dataset/out-201405.csv",select=c(168,194,196,139,141,142,232))
RegSvm_6<-fread(file="~/Desktop/intro to data/project/dataset/out-201406.csv",select=c(168,194,196,139,141,142,232))
RegSvm_7<-fread(file="~/Desktop/intro to data/project/dataset/out-201407.csv",select=c(168,194,196,139,141,142,232))
RegSvm_8<-fread(file="~/Desktop/intro to data/project/dataset/out-201408.csv",select=c(168,194,196,139,141,142,232))
RegSvm_9<-fread(file="~/Desktop/intro to data/project/dataset/out-201409.csv",select=c(168,194,196,139,141,142,232))
RegSvm_10<-fread(file="~/Desktop/intro to data/project/dataset/out-201410.csv",select=c(168,194,196,139,141,142,232))
RegSvm_11<-fread(file="~/Desktop/intro to data/project/dataset/out-201411.csv",select=c(168,194,196,139,141,142,232))
RegSvm_12<-fread(file="~/Desktop/intro to data/project/dataset/out-201412.csv",select=c(168,194,196,139,141,142,232))

RegSvmData<-RegressionData[,c(2,4,9:12)]

RegSvmData[RegSvmData==""]<-NA
RegSvmData1<-na.omit(RegSvmData)
View(RegSvmData1)

RegSvmRow<-which((RegSvmData1$State_PL=="California") & (RegSvmData1$Type_PL=="Business") & (RegSvmData1$Location_PL=="Urban"))
RegSvmData2<-RegSvmData1[c(RegSvmRow),]
View(RegSvmData2)
summary(RegSvmData2)
RegSvmData2[,3:6]<-lapply(RegSvmData2[,3:6],factor)

randIndex1<-sample(1:dim(RegSvmData2)[1])
summary(randIndex1)

head(randIndex1)
```

```
CutPoint2_3<-floor(2*dim(RegSvmData2)[1]/3)
CutPoint2_3

RegTrainData<-RegSvmData2[randIndex1[1:CutPoint2_3],]
RegTrainData<-RegTrainData[,-3:-5]
RegTestData<-RegSvmData2[randIndex1[(CutPoint2_3+1):dim(RegSvmData2)[1]],]
RegTestData<-RegTestData[,-3:-5]

install.packages("kernlab")
library(kernlab)

RegKsvmOutput<-ksvm(NPS_Type~., data=RegTrainData, kernel="rbfdot",
kpar="automatic",C=100,cross=10,prob.model=TRUE)
RegKsvmPred<-predict(RegKsvmOutput, RegTestData, type="votes")
RegCompTable<-data.frame(RegTestData[,3],RegKsvmPred[1,])
table(RegCompTable) # 75.0%
```

B.2. Naive Bayes modeling for regression models

We also used the Naive Bayes model to predict the NPS type and NPS score,while using the two main factor from the regression model. we divided the data into two groups, a train and test group for data. The cut point between the groups was ⅔ in the training group and ⅓ is in the test group. Then, we train the Naive Bayes using the training group, and use the model to predict the test group. Out of the 5692 total observations, the Naive bayes model prediction got 4445 correct,which give us a accuracy of 78%.

```
install.packages("e1071")
library(e1071)
RegNbModel<-naiveBayes(NPS_Type~.,data=RegTrainData)
RegNbPred<-predict(RegNbModel,RegTestData)
RegTestData$NB_Prediction<-RegNbPred
RegTestData$nb_YesorNot<-
ifelse(RegTestData$NPS_Type==RegTestData$NB_Prediction,"Correct","Wrong")
View(RegTestData)
Regaccuaracy<-
length(which(RegTestData$nb_YesorNot=="Correct"))/length(RegTestData$nb_YesorNot)
Regaccuaracy # 78.1%
```

B.3 KSVM modeling for Arule models

	AruleKsvmPred.1...		
AruleTestData...4.	0	1	2
Detractor	140	83	477
Passive	675	310	154
Promoter	3658	54	47

Again, we divided the data into two groups, a train and test group for data. The cut point between the groups was ⅔ in the training group and ⅓ is in the test group. By using the KSVM model, we can see there are 4,473 promoters. 3658 observations out of the 4,473are predicted as promoters. There are 815(140+675) cases which were promoters but were not classified as promoters by KSVM model. Similarly, 310 observations out of 447 were classified as passive and 477 cases out of 678 were predicted as detractors. Therefore, by using the ,hotel condition,guest room and customer service satisfaction metric, we built a KSVM model with 79.4% prediction accuracy, while predicting the NPS . In terms of NPS score,we use the table to calculate the predicted NPS score which is 56.82%, while the actual NPS score is 53.93%. The error for the KSVM NPS score prediction is 5.4%.

```
AruleData<-RegressionData[,c(2,4,5,9:12)]
summary(AruleData)
AruleData[AruleData==""]<-NA
AruleData1<-na.omit(AruleData)
View(AruleData1)

AruleSvmRow<-which((AruleData1$State_PL=="California") &
(AruleData1$Type_PL=="Business") & (AruleData1$Location_PL=="Urban"))
AruleData2<-AruleData1[c(AruleSvmRow),]
View(AruleData2)
summary(AruleData2)

RuleFunction<-function(c)
{
  c[c>=1&c<=6]<-"low"
  c[c>=7&c<=8]<-"med"
  c[c==9|c==10]<-"high"
  return(c)
}
AruleData3<-
data.frame(RuleFunction(AruleData2$Guest_Room_H),RuleFunction(AruleData2$Condition_H
otel_H),RuleFunction(AruleData2$Customer_SVC_H),AruleData2$NPS_Type)
View(AruleData3)
```

```
colnames(AruleData3)<-c("Guest Room Satisfaction","Hotel Condition","Customer
Service","NPS_Type")

str(AruleData3)
randIndex2<-sample(1:dim(AruleData2)[1])
summary(randIndex2)

head(randIndex2)
CutPoint22_3<-floor(2*dim(AruleData2)[1]/3)
CutPoint22_3

AruleTrainData<-AruleData3[randIndex2[1:CutPoint22_3],]
AruleTestData<-AruleData3[randIndex2[(CutPoint22_3+1):dim(AruleData2)[1]],]

AruleKsvmOutput<-ksvm(NPS_Type~., data=AruleTrainData, kernel="rbfdot",
kpar="automatic",C=5,cross=5,prob.model=TRUE)
AruleKsvmPred<-predict(AruleKsvmOutput, AruleTestData, type="votes")
AruleCompTable<-data.frame(AruleTestData[,4],AruleKsvmPred[1,])
table(AruleCompTable) # 79.4%
```

B.4 Naive Bayes for Arule models

We also used the Naive Bayes model to predict the NPS type and NPS score,while using the two main factor from the regression model. we divided the data into two groups, a train and test group for data. The cut point between the groups was ⅔ in the training group and ⅓ is in the test group. Then, we train the Naive Bayes using the training group, and use the model to predict the test group. Out of the 5598 total observations, the Naive bayes model prediction got 4477 correct,which give us a accuracy of 80%. In terms of NPS score, the predicted NPS score is 61.65%, while the actual NPS score is 53.93%. The error for the Naive Bayes NPS score prediction is 14.31%.

```
AruleNbModel<-naiveBayes(NPS_Type~.,data=AruleTrainData)
AruleNbPred<-predict(AruleNbModel,AruleTestData)
AruleTestData$NB_Prediction<-AruleNbPred
AruleTestData$nb_YesorNot<-
ifelse(AruleTestData$NPS_Type==AruleTestData$NB_Prediction,"Correct","Wrong")
View(AruleTestData)
Aruleaccuaracy<-
length(which(AruleTestData$nb_YesorNot=="Correct"))/length(AruleTestData$nb_YesorNot)
Aruleaccuaracy # 80.0%
```

```
### calculating the predicted NPS.
NBPro<-length(which(AruleTestData$NB_Prediction=="Promoter"))
NBDet<-length(which(AruleTestData$NB_Prediction=="Detractor"))
NBNPSPre<-(NBPro-NBDet)/length(AruleTestData$NB_Prediction)
NBNPSPre
```

Actionable Insights/Recommendations

For urban California business hotels, we recommend:

- Hyatt should focus more on guest room satisfaction, customer service quality and hotel condition in order to increase NPS score.
    - Increasing the number of Park Capitol Suite, Classic 2 Queen Beds, Executive King and Club King room types available
    - Decreasing Andaz Queen and Andaz King room type
    - Investing in better or more frequent staff customer service training and upgrade expectations of hotel cleanliness and train staff on new policies
    - Updating furniture, technology and facilities to ensure hotel is in its best condition

These policy changes would help move detractors to the passive level and passive to the promoter level which would improve overall NPS scores.

Further Research

A. Based on our results, we also recommend further research and analysis of the different brands of Hotels within the overall Hyatt brand.
B. Additionally, we recommend doing more detailed research on the room types most important for promoters and detractors to understand what specific characteristics of those room types are causing the difference in NPS type. This will give Hyatt more actionable insights on how to best improve guest rooms.