PROJECT REPORT

ANALYSIS ON HYATT REGENCY DATA SET

GUIDED BY:
PROF. JEFF SALTZ
PROF. ERIK ANDERSON
MR. IVAN SHAMSHURIN

PRESENTED BY:
ADITYA CHAUHAN
AKSHIT SALIAN
NISHITA SHETH
QIYI WU
ZHAOHENG LI

PRESENTED FOR:
IST 687-APPLIED DATA SCIENCE

Project Objective

The objective of this project is to analyze the data provided by the customers, who have visited Hyatt Regency hotels and provided feedback for their services and facilities, in order to improve customer satisfaction by providing better facilities and thereby increasing their revenue.

Project Scope

The scope of the project is limited to the chain of Hyatt Regency hotel which are located within the United States of America. For analysis we have produced a subset from existing dataset which consists of data from alternate months for the year 2014.

Deliverables

1. Identifying Key Performance Indicators (KPI) in Hyatt Regency hotels and distributing the relevance in R-programming to identify patterns.
2. Performing data munging to form clusters of data.
3. Performing data analytics on information calculate the Net Promoter Score (NPS) and represent the data in form of graphs using data visualization tools.

Data Requisition

Initially, we started with 55 attributes and 6 Lakh records for each alternate month, starting from February which sums up to 360,000 records. However, after munging the data set we designed a final data set whic h consisted of 24167 obs. with 25 variables:

Data Preprocessing

The initial data set has 360,000 instances of customer data and 44 columns describing various factors related to customers and hotels. However, there were many records where the “Likelihood to Recommend” column had null values. Also, there were many variables with bias values and several variables which did not have significant values for analysis. After eliminating these records, the data set then consisted of 47,300 records and 25 attributes.

Next, we checked the dataset for the month of February and as seen in the graph below, figured out that about 90 % of the data was of United States. As we proceeded further, we realized that the most optimal dataset should focus on the top states which contributed to the maximum portion of the data set i.e. Florida and California, which created our final data set consisted of 24167 records and 25 attributes.

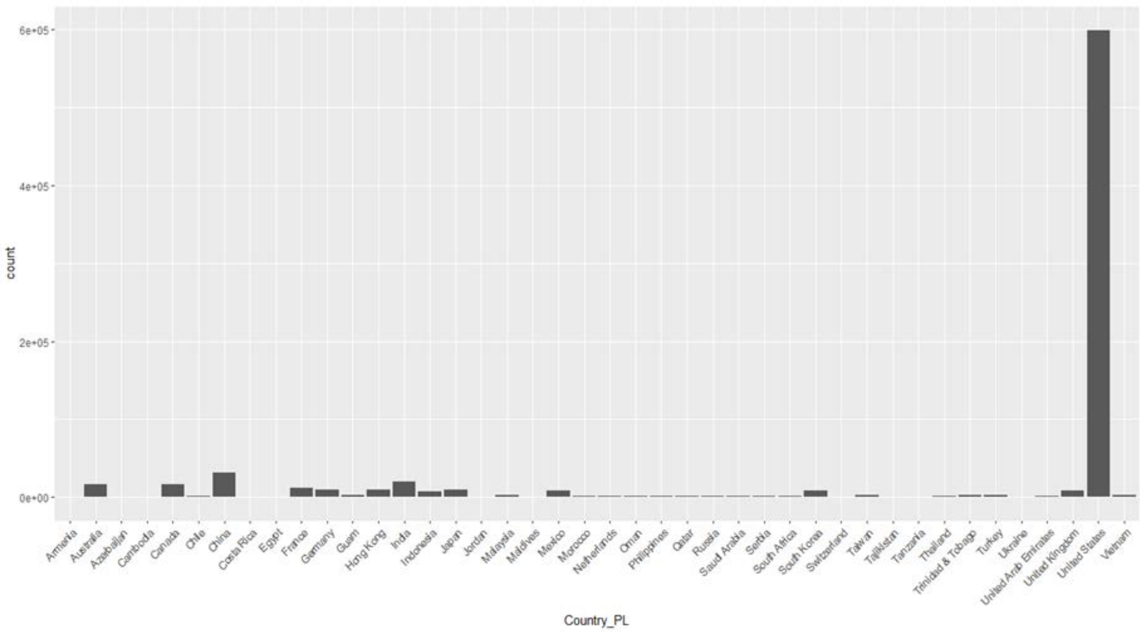


Figure : Distribution of Country Wise Data

In order to find out the count of customers, who visited Hyatt Regency hotels in the states of Florida and California we executed the below stated code. The graph below summarizes the count of customers for each individual Hyatt Regency hotels.

```
library(ggplot2)

HotelPlot <- ggplot(ModellingData, aes(x=Brand_PL)) + geom_bar(aes(fill=POV_CODE_C),
position="dodge") + facet_grid(State_PL ~ .) + scale_fill_manual(values=c("red", "green"))

HotelPlot
```

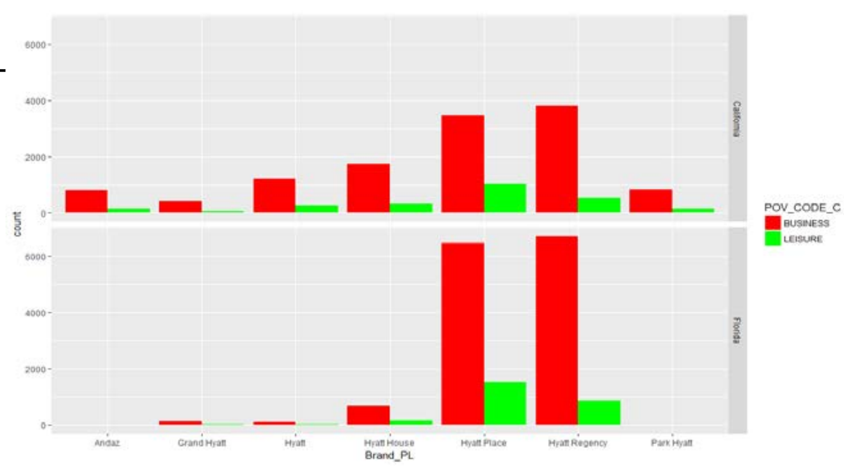


Figure: Count of Customers for Florida and California (Hotel Wise)

Business Questions

Feedback is provided by each and every customer who resides at one of the Hyatt Regency. This feedback consisted of various parameters such as likelihood to recommend, overall satisfaction, tranquility, customer service, overall experience, guestroom, etc. Analyzing this data based on their feedback is crucial to Hyatt Regency to improve their services before it becomes a major issue.

We have used various descriptive statistics, modelling techniques and visualization techniques to address the following the business questions:

1. What is the count of customers who visited hotels in Florida and California (hotel wise)?
2. State the hotel with maximum promoters and detractors.
3. What are the various services and parameters affect the promoters and detractors?
4. Which are the top two cities within the top two states with maximum number of customers?
5. Which are the top two cities within the top two states with maximum number of promoters and detractors?
6. What is the purpose of visit for maximum number of customers?
7. Calculate hotel wise Net Promoter Score (NPS) for each hotel within the city of California and Florida.
8. In which areas (overall satisfaction, tranquility, guestroom, customer service, etc.) other Hyatt hotels lag with respect to the best Hyatt hotels?

Data Cleaning

To clean the data and get rid of the NA values we initially checked for present NA values using the missingness map created with the help of "Amelia" Package in R studio.

In order to get rid of the NA values, we initially checked for percentage of NA values in the respective columns. If the percentage of NA value is greater than 55%, then the respective column was eliminated, else for the respective column the NA value was replaced with the mean value for that column.

```
# for Guest_Room_H
CountOfNA_GuestRoom <- sum(is.na(HotelData$Guest_Room_H))
CountOfNA_GuestRoom
NaGuestRoom_H <- (CountOfNA_GuestRoom/nrow(HotelData)) * 100
NaGuestRoom_H

# Percentage of NA Value : 18.36454
# for Tranquility_H
CountOfNA_Tranquility_H<- sum(is.na(HotelData$Tranquility_H))
CountOfNA_Tranquility_H
NaTranquility_H <- (CountOfNA_Tranquility_H/nrow(HotelData)) * 100
NaTranquility_H
# Percentage of NA Value : 54.61132

#Condition_Hotel_H
CountOfNA_Condition_Hotel_H<- sum(is.na(HotelData$Condition_Hotel_H))
CountOfNA_Condition_Hotel_H
NaCondition_Hotel_H <- (CountOfNA_Condition_Hotel_H/nrow(HotelData)) * 100
NaCondition_Hotel_H
#Percentage of NA Value : 18.6046

#Customer_SVC_H
CountOfNA_Customer_SVC_H<- sum(is.na(HotelData$Customer_SVC_H))
CountOfNA_Customer_SVC_H
NaCustomer_SVC_H <- (CountOfNA_Customer_SVC_H/nrow(HotelData)) * 100
NaCustomer_SVC_H
#Percentage of NA Value : 18.9516
```

```
#Staff_Cared_H
CountOfNA_Staff_Cared_H<- sum(is.na(HotelData$Staff_Cared_H))
CountOfNA_Staff_Cared_H
NaStaff_Cared_H <- (CountOfNA_Staff_Cared_H/nrow(HotelData)) * 100
NaStaff_Cared_H

##Percentage of NA Value : 54.45637

#Internet_Sat_H

CountOfNA_Internet_Sat_H<- sum(is.na(HotelData$Internet_Sat_H))
CountOfNA_Internet_Sat_H
NaInternet_Sat_H<- (CountOfNA_Internet_Sat_H/nrow(HotelData)) * 100
NaInternet_Sat_H

###Percentage of NA Value : 66.90018

#Check_In_H
CountOfNA_Check_In_H<- sum(is.na(HotelData$Check_In_H))
CountOfNA_Check_In_H
NaCheck_In_H<- (CountOfNA_Check_In_H/nrow(HotelData)) * 100
NaCheck_In_H

##Percentage of NA Value : 54.4782

#F.B_Overall_Experience_H
CountOfNA_F.B_Overall_Experience_H<- sum(is.na(HotelData$F.B_Overall_Experience_H))
CountOfNA_F.B_Overall_Experience_H
NaF.B_Overall_Experience_H<- (CountOfNA_F.B_Overall_Experience_H/nrow(HotelData)) * 100
NaF.B_Overall_Experience_H
```

```
##Percentage of NA Value : 63.7183

#Overall_Sat_H
CountOfNA_Overall_Sat_H<- sum(is.na(HotelData$Overall_Sat_H))
CountOfNA_Overall_Sat_H
NaOverall_Sat_H<- (CountOfNA_Overall_Sat_H/nrow(HotelData)) * 100
NaOverall_Sat_H

##Percentage of NA Value : 17.58107
```

Based on the values seen above,columns F.B_Overall_Experience_H and Internet_Sat_H were eliminated. In order to find a solution for categorial columns with NULL Values where the answer is given as either Yes or No, we replaced the NULL values with NA values using the following steps:

1. Exported the dataset to the excel file
2. Called the exported file again in R and passed na.strings=c("","NA")) as the parameter while reading the file

Next, we checked the percentage of NA value in each column and later, followed the same steps as stated above. Image below represents the percentage of NA Values in each column.

```
> naCategorical
                Columns PercentageOfnaValues
1          All.Suites_PL            0.3382655
2          Bell.Staff_PL           29.8699315
3            Boutique_PL            0.3382655
4     Business.Center_PL            0.3382655
5              Casino_PL            0.3382655
6          Conference_PL            0.3382655
7          Convention_PL            0.3382655
8        Dry.Cleaning_PL           29.8699315
9           Elevators_PL           29.8699315
10     Fitness.Center_PL           29.8699315
11    Fitness.Trainer_PL           31.3823054
12               Golf_PL            0.3382655
13            Laundry_PL           29.8699315
14       Limo.Service_PL           29.8699315
15           Mini.Bar_PL           29.8699315
16        Pool.Indoor_PL           29.8699315
17       Pool.Outdoor_PL           29.8699315
18 Regency.Grand.Club_PL           29.8699315
19             Resort_PL            0.3382655
20         Restaurant_PL            0.3382655
21       Self.Parking_PL           29.8699315
22    Shuttle.Service_PL           29.8699315
23                Spa_PL            0.3382655
24       Valet.Parking_PL           29.8699315
```

Figure : Column Wise Count of NA values

Further, to optimize the data set we decided to first eliminate columns with 31 % of NA values and then omit the rest of the NA values. Next, we also eliminated columns with biased or less significant values. Also, there were some NA values in the column NPS_Type, whose Likelihood to recommend was 8.686550976. Since, such values made it difficult to categorize the customer as Promoter or Passive and added to the confusion we eliminated these values as well.

Initial Phase : The project represents "Likelihood to Recommend" column as the independent attribute, which contributes to improving the business of the hotel. Therefore, the analysis calculates the Net Promoter Score (NPS) for "Likelihood to Recommend". We have identified various parameters which could the possible reasons for customers being either detractors or promoters. These parameters along with their effect on the customer being categorized as Promoter, Passive or Detractor can be visualized based on the graphs below.

Parameter: Shuttle Service

```
Shuttle.Service_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 ) &   ((FinalDatasetFinal$Shuttle.Service_PL)=="N")])
Shuttle.Service_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 ) & ((FinalDatasetFinal$Shuttle.Service_PL)=="Y")])
piesShuttle.Service_PL <- c(Shuttle.Service_PL_n,Shuttle.Service_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctpiesShuttle.Service_PL <- round(piesShuttle.Service_PL/sum(piesShuttle.Service_PL)*100)
lblspiesShuttle.Service_PL <- paste(labels, pctpiesShuttle.Service_PL) # add percents to labels
lblspiesShuttle.Service_PL <- paste(lblspiesShuttle.Service_PL,"%",sep="") # ad % to labels
pie3D(piesShuttle.Service_PL, labels=lblspiesShuttle.Service_PL,explode=0.1, col =
rainbow(length(piesShuttle.Service_PL)))
```

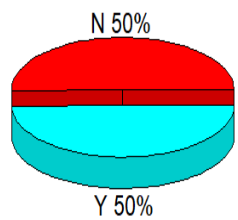


Figure: Effect of Shuttle Service on Detractors

As seen in the above pie chart, we can summarize that 50% of the detractors marked shuttle service as an important factor and probably the reason for them rating their experience low. While the other 50% of the detractors marked shuttle service as a not so important factor. To conclude, as a result of this even ratio we can say that shuttle service doesn’t have much effect on the customer being classified as promoter or detractor.

Parameter: Spa Service

```
Spa_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_
H) < 8 )  & ((FinalDatasetFinal$Spa_PL)=="N")])
```

```
Spa_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_
H) < 8 )  & ((FinalDatasetFinal$Spa_PL)=="Y")])
pies <- c(Spa_PL_n,Spa_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pct <- round(pies/sum(pies)*100)
lbls <- paste(labels, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

pie3D(pies, labels=lbls,explode=0.1, col = rainbow(length(pies)))
```

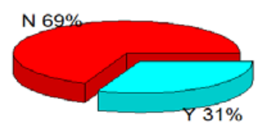


Figure: Effect of Spa Service on Detractors

From the above pie chart, we can summarize 69% of the detractors marked spa service as an important factor and probably a reason of them being a detractor. To conclude, we can say that absence of spa service may have an important effect on the customer low ratings for the hotel.

Parameter: Convention Place

```
Convention_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  &   ((FinalDatasetFinal$Convention_PL)=="N")])
Convention_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 ) & ((FinalDatasetFinal$Convention_PL)=="Y")])
piesConvention_PL <- c(Convention_PL_n,Convention_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctpiesConvention_PL <- round(piesConvention_PL/sum(piesConvention_PL)*100)
lblsConvention_PL <- paste(labels, pctpiesConvention_PL) # add percents to labels
lblsConvention_PL <- paste(lblsConvention_PL,"%",sep="") # ad % to labels
pie3D(piesConvention_PL, labels=lblsConvention_PL,explode=0.1, col =
rainbow(length(piesConvention_PL)))
```

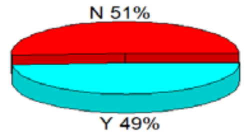


Figure: Effect of Convention Place on Detractors

From the above pie chart, we can summarize 51% of the detractors marked convention place as an important factor and probably a reason of them being a detractor. However, as the ratio between the two figures is relatively small, we can say that convention place may or may not have an important effect on the customer ratings on the hotel.

Parameter: Valet Parking

```
Valet.Parking_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  &   ((FinalDatasetFinal$Valet.Parking_PL)=="N")])
Valet.Parking_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$Valet.Parking_PL)=="Y")])
piesValet.Parking_PL <- c(Valet.Parking_PL_n,Valet.Parking_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctValet.Parking_PL <- round(piesValet.Parking_PL/sum(piesValet.Parking_PL)*100)
lblsValet.Parking_PL <- paste(labels, pctValet.Parking_PL) # add percents to labels
lblsValet.Parking_PL <- paste(lblsValet.Parking_PL,"%",sep="") # ad % to labels
pie3D(piesValet.Parking_PL, labels=lblsValet.Parking_PL,explode=0.1, col =
rainbow(length(piesValet.Parking_PL)))
```

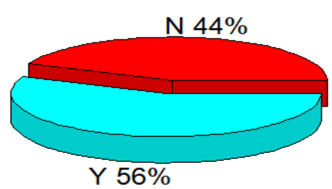


Figure: Effect of Valet Parking on Detractors

From the above pie chart, we can summarize 56% of the detractors marked available valet parking as an important factor and probably a reason of them being a detractor. To conclude, we can say that presence of valet parking may have an important negative effect on the customer ratings for the hotel.

Parameter: Suite Place

```
All.Suites_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$All.Suites_PL)=="N")])
All.Suites_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$All.Suites_PL)=="Y")])
piesAll.Suites_PL <- c(All.Suites_PL_n,All.Suites_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctAll.Suites_PLL <- round(piesAll.Suites_PL/sum(piesAll.Suites_PL)*100)
lblsAll.Suites_PL <- paste(labels, pctAll.Suites_PLL) # add percents to labels
lblsAll.Suites_PL <- paste(lblsAll.Suites_PL,"%",sep="") # ad % to labels
pie3D(piesAll.Suites_PL, labels=lblsAll.Suites_PL,explode=0.1, col =
rainbow(length(piesAll.Suites_PL)))
```

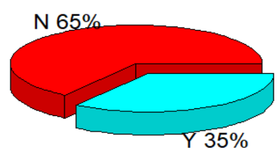


Figure: Effect of Suite Place on Detractors

From the above pie chart, we can summarize 65% of the detractors marked suite place as an important factor and probably a reason of them being a detractor. To conclude, we can say that suite may have an important effect on the low customer ratings for the hotel.

Parameter: Bell Staff

```
Bell.Staff_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$POV_CODE_C)=="BUSINESS")
&((FinalDatasetFinal$Bell.Staff_PL)=="N")])
Bell.Staff_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 ) & ((FinalDatasetFinal$POV_CODE_C)=="BUSINESS") &
((FinalDatasetFinal$Bell.Staff_PL)=="Y")])
piesBell.Staff_PL <- c(Bell.Staff_PL_n,Bell.Staff_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctBell.Staff_PL <- round(piesBell.Staff_PL/sum(piesBell.Staff_PL)*100)
lblsBell.Staff_PL <- paste(labels, pctBell.Staff_PL) # add percents to labels
lblsBell.Staff_PL <- paste(lblsBell.Staff_PL,"%",sep="") # ad % to labels
pie3D(piesBell.Staff_PL, labels=lblsBell.Staff_PL,explode=0.1, col =
rainbow(length(piesBell.Staff_PL)))
```

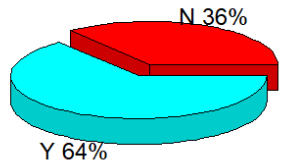


Figure: Effect of Bell Staff on Business Customers/ Detractors

From the above pie chart, we can summarize 64% of the detractors whose purpose of visit was business marked presence of bell staff as an important factor and probably a reason of them being a detractor. To conclude, we can say that presence of bell staff may have an important negative effect on the customer been classified as a detractor or a promoter.

Parameter: Laundry Place

```
Laundry_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) >
8 )  & ((FinalDatasetFinal$Laundry_PL)=="N")])
Laundry_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) >
8 )  & ((FinalDatasetFinal$Laundry_PL)=="Y")])
piesLaundry_PL <- c(Laundry_PL_n,Laundry_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctLaundry_PL <- round(piesLaundry_PL/sum(piesLaundry_PL)*100)
lblsLaundry_PL <- paste(labels, pctLaundry_PL) # add percents to labels
lblsLaundry_PL <- paste(lblsLaundry_PL,"%",sep="") # ad % to labels
pie3D(piesLaundry_PL, labels=lblsLaundry_PL,explode=0.1, col = rainbow(length(piesLaundry_PL)))
```

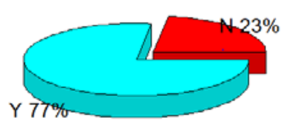


Figure: Effect of Laundry Place on Promoters

From the above pie chart, we can summarize only 77% of the promoters marked laundry place as an important factor and probably a reason of them being a promoter. To conclude, we can say that presence of laundry place may have an important effect on the customer rating the hotel high.

Parameter: Limo Service

```
Limo.Service_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$Limo.Service_PL)=="N")])
Limo.Service_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$Limo.Service_PL)=="Y")])
piesLimo.Service_PL <- c(Limo.Service_PL_n,Limo.Service_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctLimo.Service_PL <- round(piesLimo.Service_PL/sum(piesLimo.Service_PL)*100)
lblsLimo.Service_PL <- paste(labels, pctLimo.Service_PL) # add percents to labels
lblsLimo.Service_PL <- paste(lblsLimo.Service_PL,"%",sep="") # ad % to labels
pie3D(piesLimo.Service_PL, labels=lblsLimo.Service_PL,explode=0.1, col =
rainbow(length(piesLimo.Service_PL)))
```

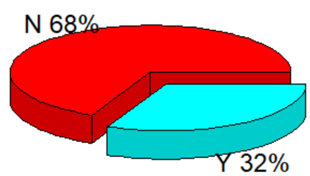


Figure: Effect of Limo Service on Detractors

From the above pie chart, we can summarize only 68% of the detractors marked absence of Limo service as an important factor and probably a reason of them being a detractor. To conclude, we can say that absence of Limo service may have an important negative effect on the customer ratings for the hotel.

Parameter: Regency Grand Place

```
Regency.Grand.Club_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) < 8 )  &
((FinalDatasetFinal$Regency.Grand.Club_PL)=="N")])
Regency.Grand.Club_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) < 8 )  &
((FinalDatasetFinal$Regency.Grand.Club_PL)=="Y")])
piesRegency.Grand.Club_PL <- c(Regency.Grand.Club_PL_n,Regency.Grand.Club_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctRegency.Grand.Club_PL <- round(piesRegency.Grand.Club_PL/sum(piesRegency.Grand.Club_PL)*100)
lblsRegency.Grand.Club_PL <- paste(labels, pctRegency.Grand.Club_PL) # add percents to labels
lblsRegency.Grand.Club_PL <- paste(lblsRegency.Grand.Club_PL,"%",sep="") # ad % to labels
pie3D(piesRegency.Grand.Club_PL, labels=lblsRegency.Grand.Club_PL,explode=0.1, col =
rainbow(length(piesRegency.Grand.Club_PL)))
```

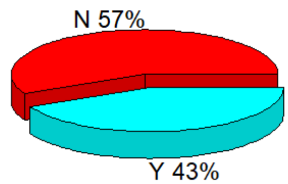


Figure: Effect of Regency Hyatt Place on Detractors

From the above pie chart, we can summarize only 57% of the detractors marked absence of Regency Grand Place as an important factor and probably a reason of them being a detractor. To conclude, we can say that absence of Regency Grand Place may have an important negative effect on the customer ratings for the hotel.

Parameter: Restaurant Place

```
Restaurant_PL_n <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$Restaurant_PL)=="N")])
Restaurant_PL_y <-
length(FinalDatasetFinal$Likelihood_Recommend_H[((FinalDatasetFinal$Likelihood_Recommend_H) <
8 )  & ((FinalDatasetFinal$Restaurant_PL)=="Y")])
piesRestaurant_PL <- c(Restaurant_PL_n,Restaurant_PL_y)
labels <- c("N","Y")
par(mar = rep(2, 4))
pctRestaurant_PL <- round(piesRestaurant_PL/sum(piesRestaurant_PL)*100)
lblsRestaurant_PL <- paste(labels, pctRestaurant_PL) # add percents to labels
lblsRestaurant_PL <- paste(lblsRestaurant_PL,"%",sep="") # ad % to labels
pie3D(piesRestaurant_PL, labels=lblsRestaurant_PL,explode=0.1, col =
rainbow(length(piesRestaurant_PL)))
```

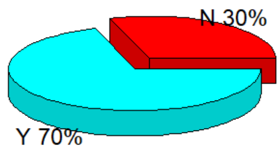


Figure: Effect of Restaurant Place on Detractors

From the above pie chart, we can summarize only 70% of the detractors marked presence of Restaurant Place as an important factor and probably a reason of them being a detractor. To conclude, we can say that presence of Restaurant Place may have an important negative effect on the customer ratings for the hotel.

Modelling

We decided to use a model based approach to carry out analysis of data. The various models used were:

1. Linear Regression
2. Support Vector Machine (SVM)
3. K Support Vector Machine (KSVM)
4. Associative Rule Mining

Points covered under this section are:

1. The modelling is based on various parameters such as overall satisfaction, guestroom, tranquility against the independent parameter, likelihood to recommend for determining the r-square value and the accuracy of these parameters.
2. Comparing the factors of the best hotel against the same parameters of other hotels.

Following code represents the points made above and recommends the various facilities that need to be improved in order to improve customer satisfaction and profitability.

Linear Regression

In order to determine linear relationship between our independent variable “Likelihood to recommend” and other variable we used Linear Regression as follows:

```
## Likelihood_Recommend_H vs Overall_Sat_H, Guest_Room_H,  Tranquility_H

LikeVsOverallGuestTranquility <- lm(Likelihood_Recommend_H ~ Overall_Sat_H + Guest_Room_H
+Tranquility_H,ModellingData)

summary(LikeVsOverallGuestTranquility)
```

```
> summary(LikeVsOverallGuestTranquility)

Call:
lm(formula = Likelihood_Recommend_H ~ Overall_Sat_H + Guest_Room_H +
    Tranquility_H, data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0552 -0.0944 -0.0552  0.0977  8.2031

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.594753   0.038704 -15.367  < 2e-16 ***
Overall_Sat_H  0.903575   0.004703 192.137  < 2e-16 ***
Guest_Room_H   0.132836   0.004777  27.806  < 2e-16 ***
Tranquility_H  0.032509   0.004763   6.825 8.99e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8616 on 24163 degrees of freedom
Multiple R-squared:  0.817,     Adjusted R-squared:  0.8169
F-statistic: 3.595e+04 on 3 and 24163 DF,  p-value: < 2.2e-16
```

Figure: Screenshot for Linear Regression

As seen above, values for linear regression using the variables Likelihood_Recommend_H vs Overall_Sat_H, Guest_Room_H, Tranquility_H are as follows:
Multiple R-squared: 0.817
Adjusted R-squared: 0.8169

```
## Removing Overall_Sat_H

LikeVsGuestTranquility <- lm(Likelihood_Recommend_H ~ Guest_Room_H
+Tranquility_H,ModellingData)

summary(LikeVsGuestTranquility)
```

```
> summary(LikeVsGuestTranquility)

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H,
    data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8448 -0.1227  0.1552  0.3746  8.5550

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.511664   0.060849   8.409   <2e-16 ***
Guest_Room_H  0.751355   0.005612 133.890   <2e-16 ***
Tranquility_H 0.181961   0.007471  24.355   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.37 on 24164 degrees of freedom
Multiple R-squared:  0.5373,    Adjusted R-squared:  0.5373
F-statistic: 1.403e+04 on 2 and 24164 DF,  p-value: < 2.2e-16
```

Figure 16: Screenshot for Linear Regression

As seen above, values for linear regression after removing Overall_Sat_H is as follows:
Multiple R-squared: 0.5373
Adjusted R-squared: 0.5373

```
#Likelihood_Recommend_H vs Guest_Room_H,  Tranquility_H, Condition_Hotel_H

LikeVsGuestTranquilityCondition <- lm(Likelihood_Recommend_H ~ Guest_Room_H +Tranquility_H
+ Condition_Hotel_H,ModellingData)

summary(LikeVsGuestTranquilityCondition)
```

```
> summary(LikeVsGuestTranquilityCondition)

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Condition_Hotel_H, data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9278  0.0137  0.1387  0.4646  5.9354

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.737452   0.060724  -12.14   <2e-16 ***
Guest_Room_H       0.450979   0.007317   61.63   <2e-16 ***
Tranquility_H      0.143855   0.007016   20.50   <2e-16 ***
Condition_Hotel_H  0.471695   0.008007   58.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.281 on 24163 degrees of freedom
Multiple R-squared:  0.5954,    Adjusted R-squared:  0.5954
F-statistic: 1.185e+04 on 3 and 24163 DF,  p-value: < 2.2e-16
```

Figure: Screenshot for Linear Regression

As seen above, values for linear regression using the variables Likelihood_Recommend_H vs Guest_Room_H, Tranquility_H, Condition_Hotel_H is as follows:
Multiple R-squared: 0.5954
Adjusted R-squared: 0.5954

```
#Guest_Room_H +Tranquility_H + Customer_SVC_H

LikeVsGuestTranquilityCustomer <- lm(Likelihood_Recommend_H ~ Guest_Room_H +Tranquility_H +
Customer_SVC_H + Condition_Hotel_H,ModellingData)

summary(LikeVsGuestTranquilityCustomer)
```

```
> summary(LikeVsGuestTranquilityCustomer)

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Customer_SVC_H + Condition_Hotel_H, data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9988 -0.0843  0.1294  0.4487  5.3332

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.104143   0.056925  -36.96   <2e-16 ***
Guest_Room_H       0.348460   0.006660   52.32   <2e-16 ***
Tranquility_H      0.106304   0.006281   16.93   <2e-16 ***
Customer_SVC_H     0.459788   0.005854   78.54   <2e-16 ***
Condition_Hotel_H  0.295745   0.007490   39.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.143 on 24162 degrees of freedom
Multiple R-squared:  0.6777,    Adjusted R-squared:  0.6777
F-statistic: 1.27e+04 on 4 and 24162 DF,  p-value: < 2.2e-16
```

Figure: Screenshot for Linear Regression

As seen above, values for linear regression using the variables Likelihood_Recommend_H vs Guest_Room_H, Tranquility_H, Customer_SVC_H is as follows:
Multiple R-squared: 0.6777
Adjusted R-squared: 0.6777

```
#Adding Staff_Cared_H

Model3AndStaff <- lm(Likelihood_Recommend_H ~ Guest_Room_H +Tranquility_H +
Customer_SVC_H + Staff_Cared_H + Condition_Hotel_H,ModellingData)

summary(Model3AndStaff)
```

```
> summary(Model3AndStaff)

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Customer_SVC_H + Staff_Cared_H + Condition_Hotel_H, data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0444 -0.0915  0.1607  0.4622  6.0250

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.550128   0.063817  -39.96   <2e-16 ***
Guest_Room_H       0.354361   0.006640   53.37   <2e-16 ***
Tranquility_H      0.068482   0.006729   10.18   <2e-16 ***
Customer_SVC_H     0.414691   0.006540   63.41   <2e-16 ***
Staff_Cared_H      0.123561   0.008138   15.18   <2e-16 ***
Condition_Hotel_H  0.298358   0.007456   40.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.138 on 24161 degrees of freedom
Multiple R-squared:  0.6808,    Adjusted R-squared:  0.6807
F-statistic: 1.03e+04 on 5 and 24161 DF,  p-value: < 2.2e-16
```

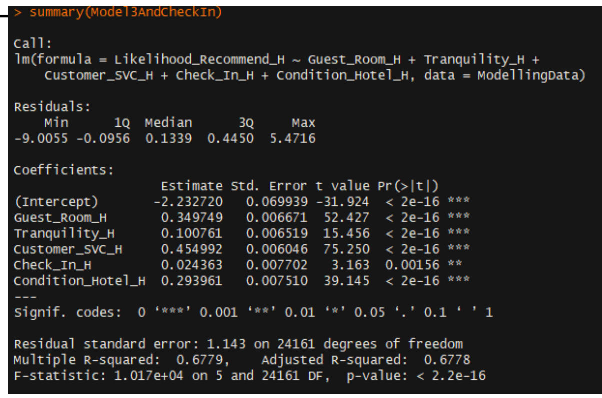
Figure: Screenshot for Linear Regression

As seen above, values for linear regression using the variables Likelihood_Recommend_H ~ Guest_Room_H +Tranquility_H + Customer_SVC_H + Staff_Cared_H + Condition_Hotel_H,ModellingData is as follows:
Multiple R-squared: 0.6808
Adjusted R-squared: 0.6807

```
# Adding Check_In_H

Model3AndCheckIn <- lm(Likelihood_Recommend_H ~ Guest_Room_H +Tranquility_H + Customer_SVC_H + Check_In_H,ModellingData)

summary(Model3AndCheckIn)
```



```
> summary(Model3AndCheckIn)

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Customer_SVC_H + Check_In_H + Condition_Hotel_H, data = ModellingData)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0055 -0.0956  0.1339  0.4450  5.4716

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.232720   0.069939 -31.924  < 2e-16 ***
Guest_Room_H       0.349749   0.006671  52.427  < 2e-16 ***
Tranquility_H      0.100761   0.006519  15.456  < 2e-16 ***
Customer_SVC_H     0.454992   0.006046  75.250  < 2e-16 ***
Check_In_H         0.024363   0.007702   3.163  0.00156 **
Condition_Hotel_H  0.293961   0.007510  39.145  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.143 on 24161 degrees of freedom
Multiple R-squared:  0.6779,    Adjusted R-squared:  0.6778
F-statistic: 1.017e+04 on 5 and 24161 DF,  p-value: < 2.2e-16
```

Figure: Screenshot for Linear Regression

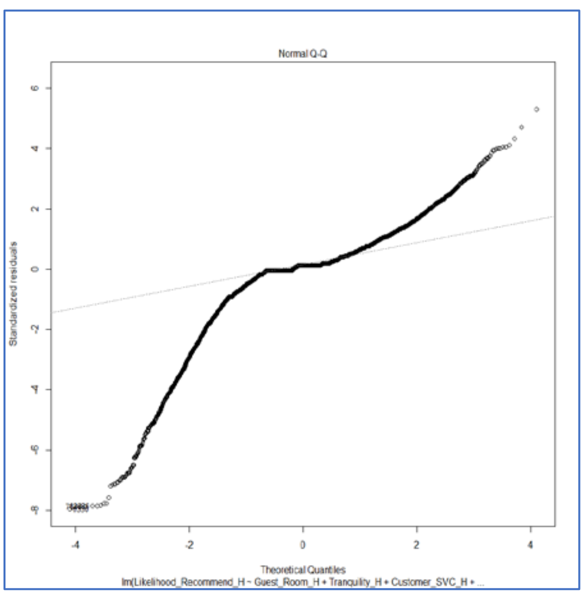


Figure: Linear Regression Model

From the Linear Model we shortlisted 5 attributes i.e Guest_Room_H ,Tranquility_H , Customer_SVC_H , Staff_Cared_H and Condition_Hotel_H. Received R-squared value for these variables is 0. 6808. We then used these variables to determine the Net Promoter Score (NPS).

Kernel- Support Vector Machine (KSVM)

To validate the accuracy or appropriateness of these shortlisted variables we first predicted NPS using KSVM (K Support Vector Machines). To run this model, we first installed and loaded "kernlab" package in R. Code for the same is as below.

We removed CheckerID, Room Number, Room Type Code, Check-in Date, Guest Country, Major Market Code, Country Code, Region, Hotel Inventory to have the data only of the Survey filed by the customer and whether a particular service is available in the hotel or not.

Also, we did summary of all the Flagged Values and kept only those attributes where we received a mixed proportion of Yes(Y) and No(N) values.

Only following attributes were kept:
All.Suites_PL,Bell.Staff_PL,Convention_PL,Laundry_PL,Limo.Service_PL,Regency.Grand.Club_PL,Restaurant_PL,Shuttle.Service_PL,Spa_PL,Valet.Parking_PL

```
write.csv(ModellingData2, "FinalDatasetFinal.csv")
# Creating Category " High", " Medium"," Low"

for(i in 1:nrow(ModellingData2)){
 if(ModellingData2[i,3] >= 9){
   ModellingData2[i,ncol(ModellingData2)] <- "High"
 } else if (ModellingData2[i,3]==7 | ModellingData2[i,3]==8){
   ModellingData2[i,ncol(ModellingData2)] <- "Medium"
 } else{
   ModellingData2[i,ncol(ModellingData2)] <- "Low"
 }
}
```

We created a new column named Likelihood Category and assigned the values on the basis of the below points which shall be used while creating a model with all the attributes (including flagged values)

- Likelihood Score of 9 and 10 is marked as High
- Likelihood Score of 7 and 8 is marked as Medium
- Likelihood score of less than 7 is marked as low

Below is the model created using the attributes selected from the Linear Model. We predicted Likelihood to recommend and created a confusion matrix after converting the predicted values to Promoter, Detractor and Passive.

```
svmOutput.Numeric <- ksvm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
Customer_SVC_H + Staff_Cared_H + Condition_Hotel_H ,  data=ModellingData.Train,
kernel="rbfdot", kpar = "automatic", C= 10, cross= 10, prob.model=TRUE)
svmOutput.Numeric
```

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

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 10

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.479017211750374

Number of Support Vectors : 8116

Objective Function Value : -28538.28
Training error : 0.225469
Cross validation error : 1.191575
Laplace distr. width : 1.64302
```

Figure: Screenshot for KSVM

```
> resultsNumeric.NPS
           PredictedNPS
TestNPS     Passive Promoter Detractor
   Passive      808      501       116
   Promoter     277     5357        33
  Detractor     181      101       682
```

Figure: KSVM Result

Accuracy for the above code was detected to be 84.99%. Next, in order to check the dependency of all variables together, we executed KSVM for all variable in our final data set. Code for the same is as seen below.

We predicted the likelihoodCategory (“ Medium”, “Low” and “High”) created above. Also, we converted all the flagged values to 1 (N) and (Y) and Likelihood Category values to 1(Promoter) , 2(Passive), 3(Detractor) and predicted them using the below model. Also, we removed Likelihood to recommend and Overall Satisfaction from the table to get better insights.

```
svmOutput.All <- ksvm(LikelihoodCategory ~.,  data=ModellingData.TrainAll, kernel="rbfdot", kpar = "automatic", C= 10, cross= 10, prob.model=TRUE)
summary(svmOutput.All)
svmOutput.All
```

We later created a confusion matrix to check the accuracy of the model and got the accuracy of 84.98%.

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

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 10

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0652954260735108

Number of Support Vectors : 7524

Objective Function Value : -41727.58
Training error : 0.413103
Cross validation error : 0.165734
Laplace distr. width : 1.183184
```

```
> results.All
    Predicted
Test    1    2    3
   1 5463  257    3
   2  518 1329   37
   3  116  279   54
```

Figure: Screenshot and confusion matrix for KSVM for all variables

We reached the conclusion that flagged values don't have much importance on the prediction as they both almost have same accuracy.

Support Vector Machine: In order to further validate the accuracy and appropriateness of the variables received from Linear Model, we executed Support Vector Machine (SVM) algorithm for the variables derived from Linear Model.

We predicted Likelihood to recommend and created a confusion matrix after converting the predicted values to Promoter, Detractor and Passive.

```
svmOutput2.Numeric <- svm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
Customer_SVC_H + Staff_Cared_H + Condition_Hotel_H ,  data=ModellingData.Train)
svmOutput2.Numeric
summary(svmOutput2.Numeric)
svmPred.Numeric <- predict(svmOutput2.Numeric,ModellingData.Test, type="votes")
svmPred.Numeric
```

After predicting the values, we created converted the predicted Likelihood to recommend values to Promoter, Detractor and Passive and created the confusion matrix to check the accuracy.

```
for(i in 1:nrow(CompTable2Numeric.ex)){
 if(CompTable2Numeric.ex[i,1] >= 9){
  CompTable2Numeric.ex[i,3] <- " Promoter"
 } else if (CompTable2Numeric.ex[i,1]==7 | CompTable2Numeric.ex[i,1]==8){
  CompTable2Numeric.ex[i,3] <- " Passive"
 } else{
  CompTable2Numeric.ex[i,3] <- "Detractor"
 }
}

View(CompTable2Numeric.ex)

# Putting NPS Type for Predicted Column on the basis of Value

for(i in 1:nrow(CompTable2Numeric.ex)){
 if(CompTable2Numeric.ex[i,2] >= 9){
  CompTable2Numeric.ex[i,4] <- " Promoter"
 } else if (CompTable2Numeric.ex[i,2]==7 | CompTable2Numeric.ex[i,2]==8){
  CompTable2Numeric.ex[i,4] <- " Passive"
 } else{
  CompTable2Numeric.ex[i,4] <- "Detractor"
 }
```

```
> resultsNumeric2.NPS
           PredictedNPS
TestNPS      Passive Promoter Detractor
  Passive        801      487       137
  Promoter       353     5292        22
  Detractor      240       85       639
>
```

Figure: SVM Result

Accuracy for the above code was detected to be 83.56%. Next, to check the dependency of all variables together, we executed SVM code for all variables in our final data set. Code for the same is as below.

We predicted Likelihood Category using the same approach for KSVM (for all variables i.e including flagged variables and numeric variables and converted them to 0 and 1 form). Later, in order to check the accuracy we created the confusion matrix of the predicted and the original values.

```
svmOutput2.All <- svm(LikelihoodCategory ~.,  data=ModellingData.TrainAll)
summary(svmOutput2.All)
```

```
> results2.All
    Predicted
Test    1    2    3
   1 5446  277    0
   2  550 1333    1
   3  121  320    8
```

Figure: SVM Results (all variables)

Accuracy for the above model was detected as 84.27%. The accuracy was less as compared to KSVM and hence we decided to discard this model and take KSVM into consideration.

In order to check the validation and rules of the concluded factors, we proceeded to check the rules through Association Rule Mining.

Association Rule Mining

In order to implement Associative Rule Mining, we shall be requiring installation and loading of two important packages named “arules” and “arulesViz” in R.

We have created a separate column named “LikelihoodCategory” whose value will be “high”, “medium” or “low” on the basis of Likelihood to Recommend score.

1. Score of 9 or more will be considered as High
2. Score of 7 or 8 will be considered as Medium
3. Score of less than 7 will be considered as Low

In association Rule Mining, we removed Overall Satisfaction, Likelihood to recommend and NPS Type to get better insights.

The code to get the rules for Likelihood Category “High” is as under:

```
ruleset <-
apriori(ModellingData2.factor,parameter=list(support=0.3,confidence=0.8,maxlen=10),appearance=lis
t(default="lhs",rhs=("LikelihoodCategory=High")))
summary(ruleset)
```

```
> summary(ruleset)
set of 31 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5
 3 18  9  1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.258   4.000   5.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.3003   Min.   :0.9040   Min.   :1.290   Min.   : 7258
 1st Qu.:0.3173   1st Qu.:0.9145   1st Qu.:1.305   1st Qu.: 7667
 Median :0.3334   Median :0.9399   Median :1.341   Median : 8058
 Mean   :0.3561   Mean   :0.9397   Mean   :1.341   Mean   : 8607
 3rd Qu.:0.3862   3rd Qu.:0.9579   3rd Qu.:1.367   3rd Qu.: 9333
 Max.   :0.5121   Max.   :0.9774   Max.   :1.395   Max.   :12375

mining info:
                 data ntransactions support confidence
 ModellingData2.factor         24167     0.3        0.8
>
```

Figure: Ruleset- Association Rule Mining

```
goodrules <- ruleset[quality(ruleset)$lift > 1.3]
goodrules
inspect(goodrules)
goodrules <- sort(goodrules,by='lift',decreasing=T)
summary(goodrules)
```

```
[1]  {Guest_Room_H=10,
      Condition_Hotel_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.3637191  0.9774269 1.394996  8790
[2]  {POV_CODE_C=BUSINESS,
      Guest_Room_H=10,
      Condition_Hotel_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.3012786  0.9761362 1.393154  7281
[3]  {Guest_Room_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.3893326  0.9741174 1.390273  9409
[4]  {POV_CODE_C=BUSINESS,
      Guest_Room_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.3227128  0.9722014 1.387539  7799
[5]  {Condition_Hotel_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.4099392  0.9610982 1.371692  9907
[6]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=10,
      Customer_SVC_H=10}        => {LikelihoodCategory=High} 0.3412918  0.9601863 1.370390  8248
[7]  {Guest_Room_H=10,
      Condition_Hotel_H=10}     => {LikelihoodCategory=High} 0.3994290  0.9596381 1.369608  9653
[8]  {POV_CODE_C=BUSINESS,
      Guest_Room_H=10,
      Condition_Hotel_H=10}     => {LikelihoodCategory=High} 0.3311127  0.9583234 1.367732  8002
[9]  {Condition_Hotel_H=10,
      Customer_SVC_H=10,
      Laundry_PL=Y}             => {LikelihoodCategory=High} 0.3123681  0.9573874 1.366396  7549
[10] {Guest_Room_H=10,
      Condition_Hotel_H=10,
      Laundry_PL=Y}             => {LikelihoodCategory=High} 0.3055820  0.9569781 1.365812  7385
[11] {Guest_Room_H=10,
      Spa_PL=N}                 => {LikelihoodCategory=High} 0.3081061  0.9538816 1.361392  7446
[12] {Guest_Room_H=10,
      Limo.Service_PL=N}        => {LikelihoodCategory=High} 0.3130715  0.9537376 1.361187  7566
[13] {Guest_Room_H=10}          => {LikelihoodCategory=High} 0.4351388  0.9492688 1.354809 10516
[14] {POV_CODE_C=BUSINESS,
      Guest_Room_H=10}          => {LikelihoodCategory=High} 0.3614433  0.9471915 1.351844  8735
[15] {Guest_Room_H=10,
      Laundry_PL=Y}             => {LikelihoodCategory=High} 0.3320644  0.9455638 1.349521  8025
[16] {Condition_Hotel_H=10,
      Regency.Grand.Club_PL=N} => {LikelihoodCategory=High} 0.3003269  0.9399119 1.341455  7258
[17] {Condition_Hotel_H=10,
      Limo.Service_PL=N}        => {LikelihoodCategory=High} 0.3270576  0.9398335 1.341343  7904
[18] {Condition_Hotel_H=10,
      Spa_PL=N}                 => {LikelihoodCategory=High} 0.3214300  0.9383909 1.339284  7768
[19] {Condition_Hotel_H=10}     => {LikelihoodCategory=High} 0.4593868  0.9328628 1.331394 11102
[20] {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=10}     => {LikelihoodCategory=High} 0.3832085  0.9326284 1.331060  9261
>
```

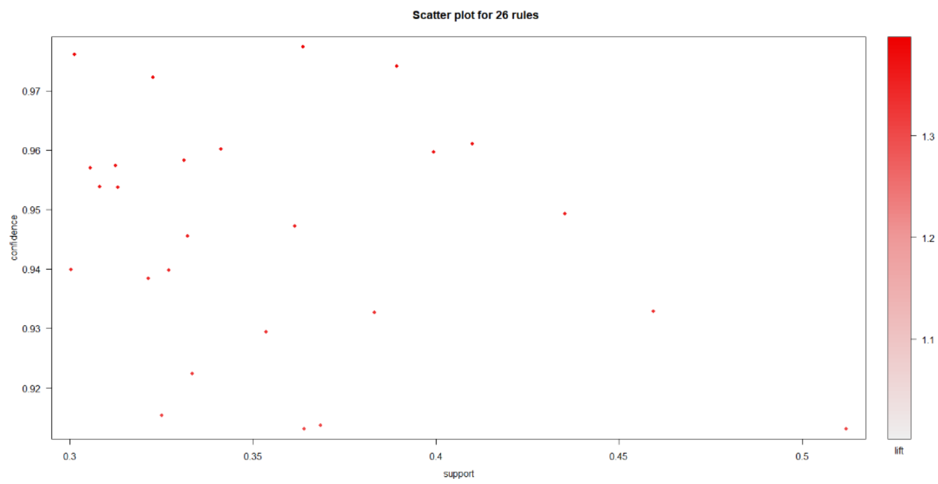


Figure: Scatter Plot for Ruleset

Association Rules for LikelihoodCategory “Low”

```
rulesetlow<-
apriori(ModellingData2.factor,parameter=list(support=0.01,confidence=0.8,maxlen=5),appearance=list(default="lhs",rhs=("LikelihoodCategory=Low")))

summary(rulesetlow)
```

```
> summary(rulesetlow)
set of 67 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5
 4 21 27 15

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   3.791   4.000   5.000

summary of quality measures:
    support          confidence          lift           count
 Min.   :0.01001   Min.   :0.8066   Min.   :3.336   Min.   :242.0
 1st Qu.:0.01043   1st Qu.:0.8203   1st Qu.:3.393   1st Qu.:252.0
 Median :0.01088   Median :0.8265   Median :3.418   Median :263.0
 Mean   :0.01142   Mean   :0.8283   Mean   :3.426   Mean   :276.1
 3rd Qu.:0.01173   3rd Qu.:0.8333   3rd Qu.:3.447   3rd Qu.:283.5
 Max.   :0.01870   Max.   :0.9244   Max.   :3.823   Max.   :452.0

mining info:
                  data ntransactions support confidence
 ModellingData2.factor         24167    0.01        0.8
```

Figure: Ruleset- Association Rule Mining

```
> inspect(goodrulesLow[1:20])
     lhs                            rhs                          support confidence     lift count
[1]  {Guest_Room_H=3}            => {LikelihoodCategory=Low} 0.01113088  0.9243986 3.823368   269
[2]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      Bell.Staff_PL=Y}           => {LikelihoodCategory=Low} 0.01071709  0.8464052 3.500783   259
[3]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      All.Suites_PL=N,
      Bell.Staff_PL=Y}           => {LikelihoodCategory=Low} 0.01022055  0.8430034 3.486713   247
[4]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      Bell.Staff_PL=Y,
      Restaurant_PL=Y}           => {LikelihoodCategory=Low} 0.01022055  0.8430034 3.486713   247
[5]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      Restaurant_PL=Y}           => {LikelihoodCategory=Low} 0.01088261  0.8429487 3.486487   263
[6]  {Condition_Hotel_H=5,
      Bell.Staff_PL=Y}           => {LikelihoodCategory=Low} 0.01253776  0.8416667 3.481184   303
[7]  {Condition_Hotel_H=5,
      Regency.Grand.Club_PL=N} => {LikelihoodCategory=Low} 0.01046882  0.8405316 3.476489   253
[8]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      All.Suites_PL=N}           => {LikelihoodCategory=Low} 0.01042744  0.8400000 3.474291   252
[9]  {POV_CODE_C=BUSINESS,
      Condition_Hotel_H=5,
      All.Suites_PL=N,
      Restaurant_PL=Y}           => {LikelihoodCategory=Low} 0.01042744  0.8400000 3.474291   252
[10] {Condition_Hotel_H=5,
      Restaurant_PL=Y}           => {LikelihoodCategory=Low} 0.01253776  0.8393352 3.471541   303
[11] {Condition_Hotel_H=5,
      All.Suites_PL=N,
      Bell.Staff_PL=Y}           => {LikelihoodCategory=Low} 0.01187570  0.8391813 3.470904   287
[12] {Condition_Hotel_H=5,
      Bell.Staff_PL=Y,
```

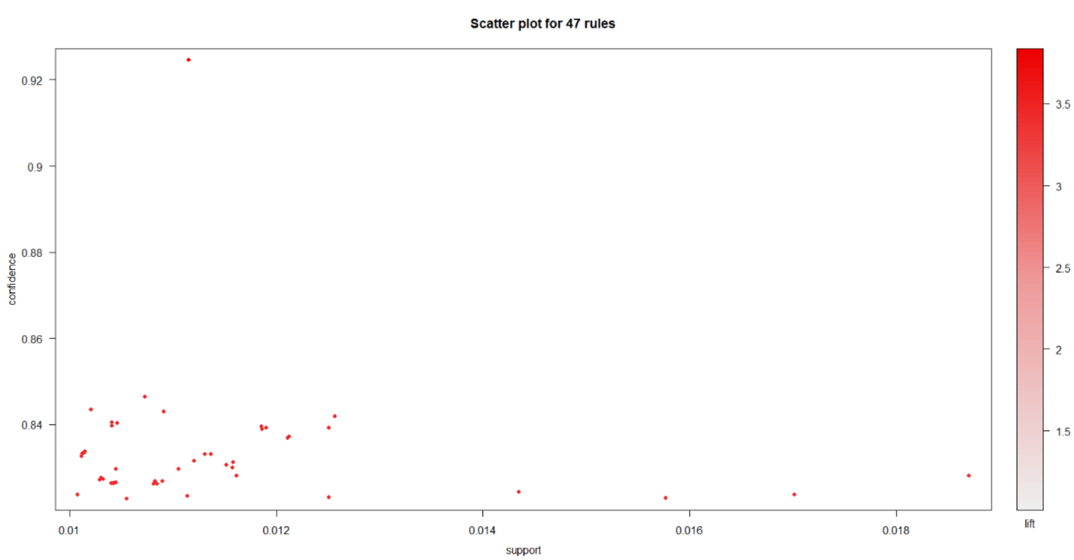


Figure: Scatter Plot for Ruleset

The rules derived from the Association Rules confirmed our insights derived from Linear Model and KSVM Model. In order to get high NPS Score / High Likelihood to Recommend, the hotel need to focus on Customer Service, Guest Room and Hotel Condition. Also, hotels need to focus more on Business clients as they determine towards a hotel being a promoter or a detractor.

Also, we concluded that a hotel can have a Laundry Service, if not available as they too play an important role.

Business Analysis

Based on the above study and modelling results, we can recommend solutions for each of the above stated business question as follows:

1. What is the count of customers who visited hotels in Florida and California (hotel wise)?

```
HotelPlot <- ggplot(FinalDatasetFinal, aes(x=Brand_PL)) + geom_bar(aes(fill=POV_CODE_C), position="dodge") + facet_grid(State_PL ~ .) + scale_fill_manual(values=c("red", "green"))
HotelPlot
```

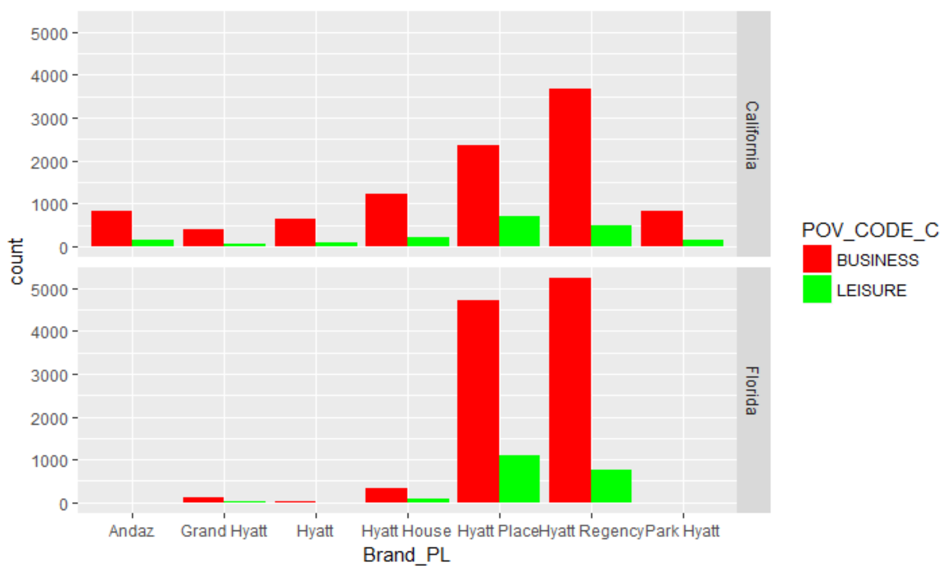


Figure: Customer Count for Florida and California

2. State the hotel with maximum promoters and detractors.

```
HotelPromoter <- ggplot(FinalDatasetFinal, aes(x=Brand_PL)) + geom_bar(aes(fill=NPS_Type), position="dodge") + facet_grid(State_PL ~ .) + scale_fill_manual(values=c("red", "green","blue"))
HotelPromoter
```

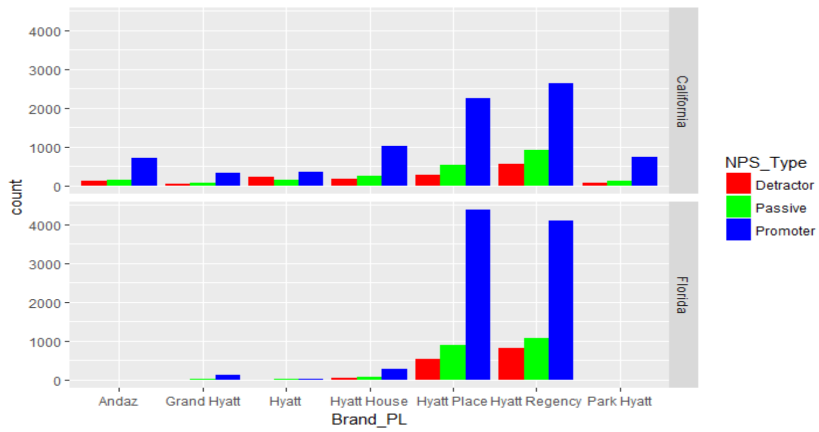


Figure: Count for Promoter, Passive Customers and Detractors (City Wise)

3. What are the various services and parameters affect the promoters and detractors?

<u>Parameter: Guest Room</u>

```
HotelGuest_Room_H <- ggplot(FinalDatasetFinal, aes(x=Guest_Room_H)) +
geom_bar(aes(fill=(NPS_Type),width=0.3)) + scale_fill_manual(values=c("red", "green","blue"))
+scale_x_continuous(breaks=seq(0.0, 10.0, 1.0))
HotelGuest_Room_H
```

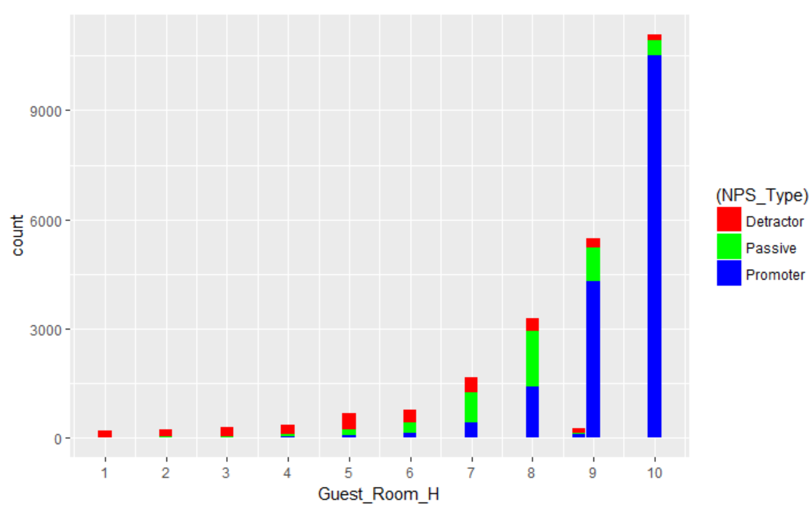


Figure: Effect of Guest Room on Customers

<u>Parameter: Condition of the hotel</u>

```
HotelCondition_Hotel_H <- ggplot(FinalDatasetFinal, aes(x=Condition_Hotel_H)) +
geom_bar(aes(fill=NPS_Type),width=0.3) +  scale_fill_manual(values=c("red",
"green","blue"))+scale_x_continuous(breaks=seq(0.0, 10.0, 1.0))

HotelCondition_Hotel_H
```

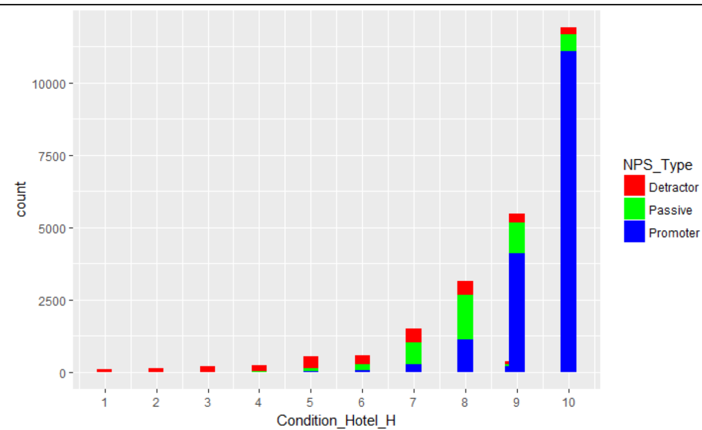


Figure: Effect of Condition of the hotel on Customers

Parameter: Customer Service

```
HotelCustomer_Service_H <- ggplot(FinalDatasetFinal, aes(x=Customer_SVC_H)) +
geom_bar(aes(fill=NPS_Type),width = 0.3)  + scale_fill_manual(values=c("red", "green","blue"))+
scale_x_continuous(breaks=seq(0.0, 10.0, 1.0))
HotelCustomer_Service_H
```

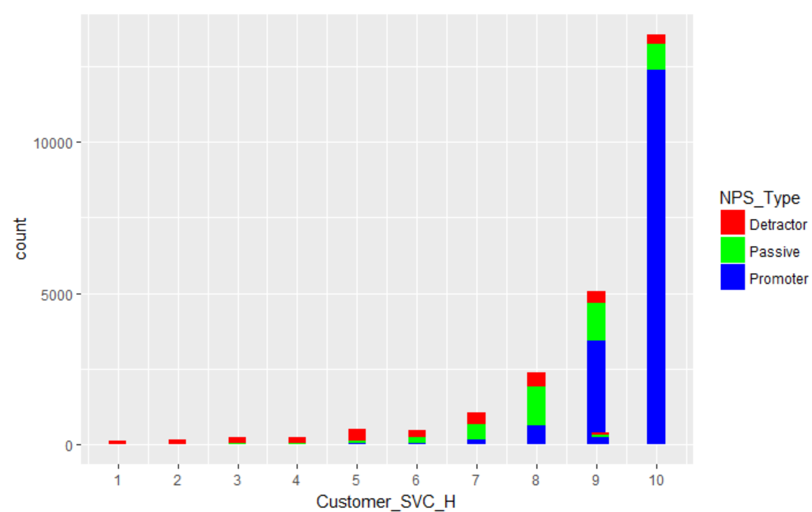


Figure: Effect of Customer Service on Customers

4. Which are the top two cities within the top two states with maximum number of customers?

```
CityCustomers <-ggplot(ModellingData, aes(x=City_PL)) + geom_bar(aes(fill=X), position =”dodge”) +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
CityCustomers
```

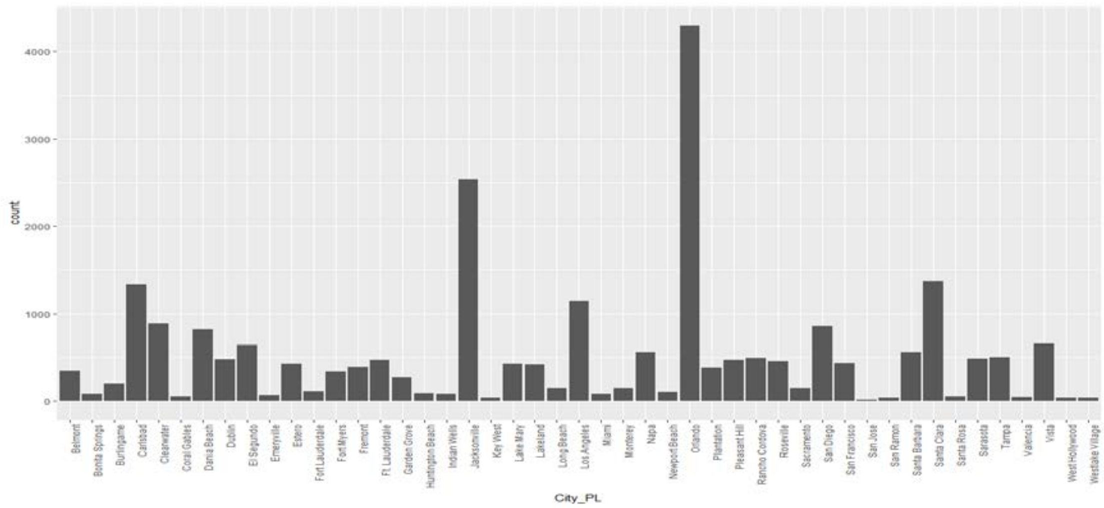


Figure: Count of Customers in Top Cities

5. Which are the top two cities within the top two states with maximum number of promoters?

```
Promoters <- ModellingData[which(ModellingData$NPS_Type=="Promoter"),]
View(Promoters)
CityPromoters <-ggplot(Promoters, aes(x=City_PL)) + geom_bar(aes(fill=X),
position ="dodge") + theme(axis.text.x = element_text(angle = 90, hjust = 1))
CityPromoters
```

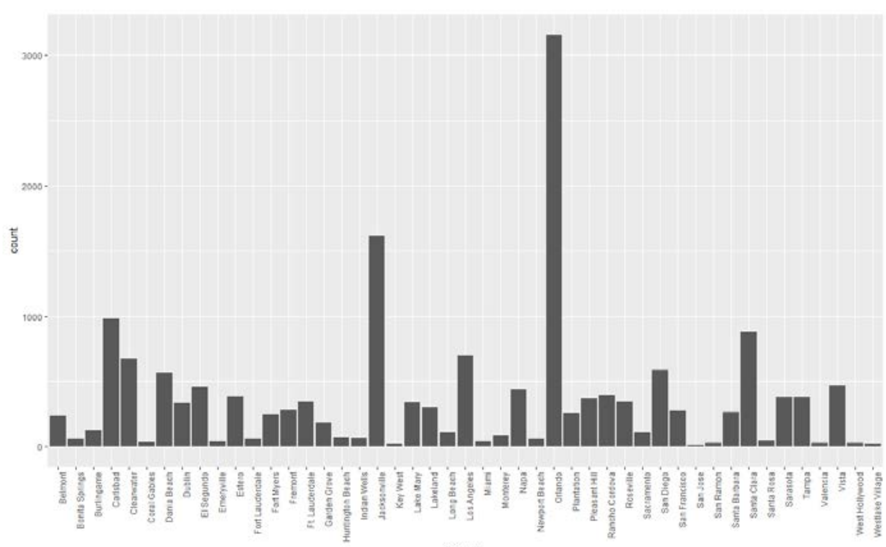


Figure: Cities with Maximum number of Promoters

6. What is the purpose of visit for maximum number of customers?

```
library(ggplot2)
HotelPlot <- ggplot(ModellingData, aes(x=Brand_PL)) + geom_bar(aes(fill=POV_CODE_C), position="dodge") + facet_grid(State_PL ~ .) + scale_fill_manual(values=c("red", "green"))
HotelPlot
```

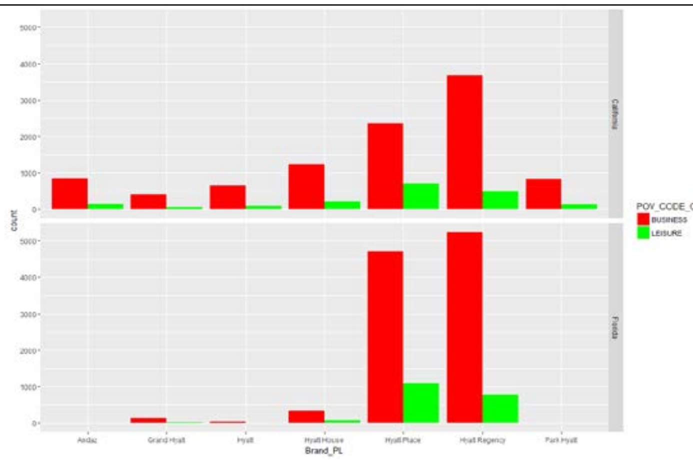


Figure: Purpose of Visit Statistics

7. Calculate hotel wise Net Promoter Score (NPS) for each hotel.

```
library(sqldf)
install.packages("sqldf")
# For Hyatt Regency
HyattRegencyPromoter <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Regency" AND NPS_Type="Promoter"')
HyattRegencyDetractor <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Regency" AND NPS_Type="Detractor"')
HyattRegencyPassive <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Regency" AND NPS_Type="Passive"')

NPSHyattRegency <- ((HyattRegencyPromoter-HyattRegencyDetractor)/(HyattRegencyPromoter+HyattRegencyDetractor+HyattRegencyPassive)) * 100
NPSHyattRegency

# For Hyatt Place
HyattPlacePromoter <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Place" AND NPS_Type="Promoter"')
HyattPlaceDetractor <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Place" AND NPS_Type="Detractor"')
HyattPlacePassive <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt Place" AND NPS_Type="Passive"')


NPSAndaz<- ((AndazPromoter-AndazDetractor)/(AndazPromoter+AndazDetractor+AndazPassive)) * 100
```

```
HyattDetractor <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt" AND
NPS_Type="Detractor"')
HyattPassive <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Hyatt" AND
NPS_Type="Passive"')

NPSHyatt<- ((HyattPromoter-HyattDetractor)/(HyattPromoter+HyattDetractor+HyattPassive)) * 100
NPSHyatt

#GrandHyatt
```

Creating Table

```
GrandHyattPromoter <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Grand
Hyatt" AND NPS_Type="Promoter"')
GrandHyattDetractor <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Grand
Hyatt" AND NPS_Type="Detractor"')
GrandHyattPassive <- sqldf('Select Count(X) FROM ModellingData WHERE Brand_PL="Grand
Hyatt" AND NPS_Type="Passive"')

NPSGrandHyatt<- ((GrandHyattPromoter-
```

```
NPSHotel <- data.frame(NPSHyattRegency,NPSHyattPlace, NPSHyattHouse, NPSAndaz,
NPSParkHyatt, NPSHyatt, NPSGrandHyatt)
NPSHotel
colnames(NPSHotel) <- c("NPSHyattRegency","NPSHyattPlace", "NPSHyattHouse", "NPSAndaz",
"NPSParkHyatt", "NPSHyatt", "NPSGrandHyatt")
View(NPSHotel)

NPSHotel <- t(NPSHotel)
NPSHotel <- as.data.frame(NPSHotel)
colnames(NPSHotel) <- c("NPSPercentage")
```

```
barplot(NPSHotel$NPSPercentage,xlab="Comparision of NPS Values",ylab="Percentage of Values",
ylim=c(0,100),main="Comparison",names.arg = c("NPSHyattRegency","NPSHyattPlace",
"NPSHyattHouse", "NPSAndaz", "NPSParkHyatt", "NPSHyatt", "NPSGrandHyatt"),col="red")
```

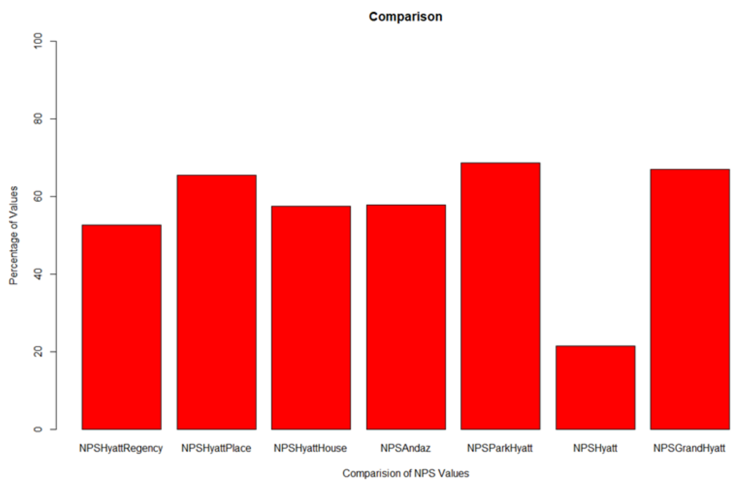


Figure: Hotel Wise NPS Score

NPS Park Hyatt have the highest NPS followed by NPS Grand Hyatt and then NPS Hyatt Place.

8. In which areas (overall satisfaction, tranquility, guestroom, customer service, etc.) other Hyatt hotels lag with respect to the best Hyatt hotels?

   To perform comparison between the Hyatt hotel with maximum NPS score and the one with low NPS Score, we selected the best performing hotel from Orlando, city with maximum number of Promoters in Florida against the hotels in Dania Beach, Florida. Results of the comparison can be seen below.

```
#Place with highest Promoters is Orlando and the best hotel is Hyatt Place, we shall compare it with
Hyatt House
#Analyzing the factors

ModellingData.Orlando <- ModellingData[which(ModellingData$City_PL=="Orlando" &
ModellingData$Brand_PL=='Hyatt Place'),]
View(ModellingData.Orlando)
ModellingData.DaniaBeach <- ModellingData[which(ModellingData$City_PL=="Dania Beach" ),]
ModellingData.compare <- rbind(ModellingData.Orlando,ModellingData.DaniaBeach)
View(ModellingData.compare)
#Checking the Factors

# Overall Satisfaction:
BestFactors1 =
mean(ModellingData.compare$Overall_Sat_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Overall Satisfaction: ",BestFactors1))

#Guest_Room_H:

BestFactors2 =
mean(ModellingData.compare$Guest_Room_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Guest Room Satisfaction: ",BestFactors2))
```

```
#Tranquility_H

BestFactors3 =
mean(ModellingData.compare$Tranquility_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Tranquility Satisfaction: ",BestFactors3))

#Condition_Hotel_H
BestFactors4 =
mean(ModellingData.compare$Condition_Hotel_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Condition Satisfaction: ",BestFactors4))

#Customer_SVC_H

BestFactors5 =
mean(ModellingData.compare$Customer_SVC_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Customer Service Satisfaction: ",BestFactors5))

#Staff_Cared_H

BestFactors6 =
mean(ModellingData.compare$Staff_Cared_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Staff Service Satisfaction: ",BestFactors6))


#Check_In_H
BestFactors8 =
mean(ModellingData.compare$Check_In_H[(ModellingData.compare$City_PL=="Orlando")],
na.rm=T)
print(paste("Desired Check In Satisfaction: ",BestFactors8))



#########Hotels in Dania Beach#####


DBFactors1 =
mean(ModellingData.compare$Overall_Sat_H[(ModellingData.compare$City_PL=="Dania Beach")],
na.rm=T)
print(paste("Current Overall Satisfaction: ",DBFactors1))

#Guest_Room_H:

DBFactors2 =
mean(ModellingData.compare$Guest_Room_H[(ModellingData.compare$City_PL=="Dania
Beach")], na.rm=T)
print(paste("Current Guest Room Satisfaction: ",DBFactors2))
```

```
#Tranquility_H

DBFactors3 =
mean(ModellingData.compare$Tranquility_H[(ModellingData.compare$City_PL=="Dania Beach")],
na.rm=T)
print(paste("Current Tranquility Satisfaction: ",DBFactors3))

#Condition_Hotel_H
DBFactors4 =
mean(ModellingData.compare$Condition_Hotel_H[(ModellingData.compare$City_PL=="Dania
Beach")], na.rm=T)
print(paste("Current Condition Satisfaction: ",DBFactors4))

#Customer_SVC_H

DBFactors5 =
mean(ModellingData.compare$Customer_SVC_H[(ModellingData.compare$City_PL=="Dania
Beach")], na.rm=T)
print(paste("Current Customer Service Satisfaction: ",DBFactors5))

#Staff_Cared_H

DBFactors6 =
mean(ModellingData.compare$Staff_Cared_H[(ModellingData.compare$City_PL=="Dania Beach")],
na.rm=T)
print(paste("Current Staff Service Satisfaction: ",DBFactors6))


#Check_In_H
DBFactors8 =
mean(ModellingData.compare$Check_In_H[(ModellingData.compare$City_PL=="Dania Beach")],
na.rm=T)
print(paste("Current Check In Satisfaction: ",DBFactors8))



##########Plotting Graph###
Factors <-
c("Overall_Sat_H","Guest_Room_H","Tranquility_H","Condition_Hotel_H","Customer_SVC_H","St
aff_Cared_H","Check_In_H")
Desired <-
c(BestFactors1,BestFactors2,BestFactors3,BestFactors4,BestFactors5,BestFactors6,BestFactors8)
Current <- c(DBFactors1,DBFactors2,DBFactors3,DBFactors4,DBFactors5,DBFactors6,DBFactors8)
ComparisonDF <- data.frame(Factors,Desired,Current)

library(ggplot2)
library(reshape2)
meltedComparisonDF = melt(ComparisonDF, id = "Factors")
ggplot(meltedComparisonDF, aes(Factors, value)) +geom_bar(aes(fill = variable), position = "dodge",
stat="identity")+ggtitle("Orlando vs Dania Beach") + theme(axis.text.x = element_text(angle = 90,
hjust = 1)) + scale_fill_manual(values=c("#aec6d7","#ffc0cb")) + theme_dark()
```

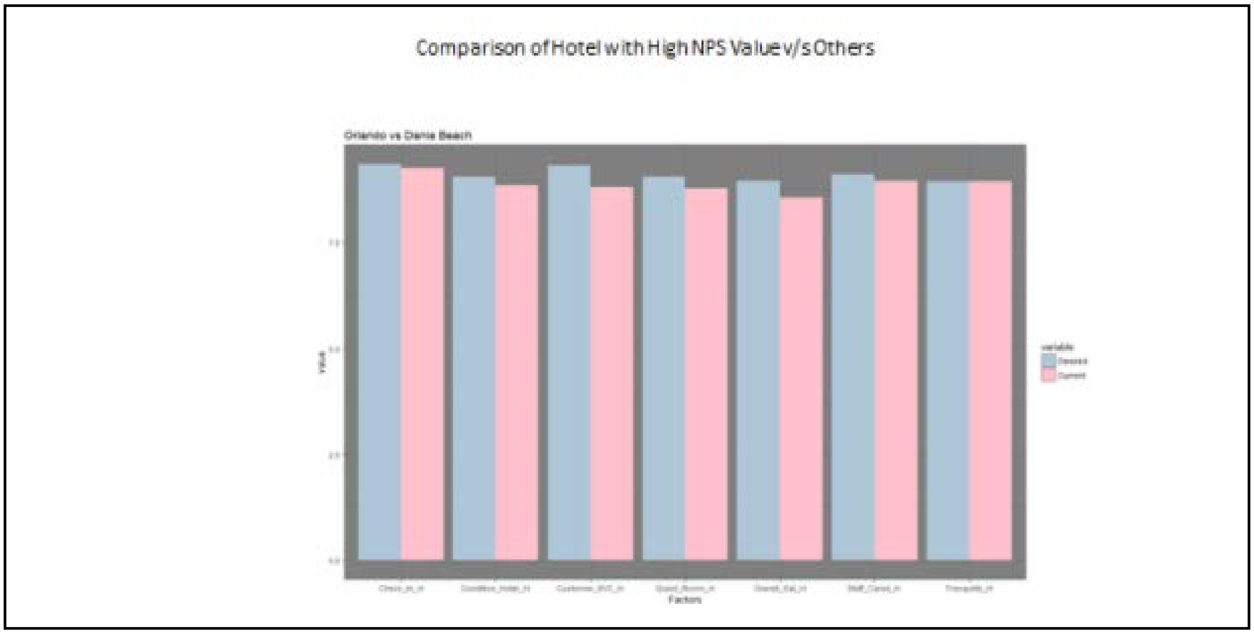


Figure: Comparison between Hotels

Conclusion

Based on the insights provided above and the Net Promoter Score calculated, Hyatt Regency can work towards implementing necessary changes, improving customer satisfaction and thus their profitability on the basis of the suggestions given below :

1) Focus on Guest Room Condition
2) Focus on Condition Hotel
3) Focus on Customer Service
4) Focus on Tranquility
5) Improvement in Check In process
6) Focus on Business Class people and have all the suites
7) Have Laundry Services in the hotel, if not.

The model which suits the best is the KSVM model, however, the hotels can also use the comparison depicted above (i.e comparing with the ratings of the best hotel) to determine the services they can improve in to increase the Likelihood To recommend and NPS Score.