IST 687 Applied Data Science

Analysis of Hyatt Group of Hotels

M004 GROUP 1

Siddarth K Sampath
Vamsee Metlapalli
Sooraj Advani
Shreya Sawant
Aditya B Patnaik

Data Cleaning

The dataset for this project consisted of 15,711,552 observations and 236 attributes. This dataset was for an entire year from February 2014 to January 2015. The main focus for us was to concentrate on a short period to give out an accurate analysis.

For our analysis, we considered the data for six months (half yearly) – February to July. Each month consisted of over a million rows of data. As we are munging through the data we observed that there were lot of NA's and blanks in the dataset.

Since it is not possible to remove all NA's which would affect the quality of our dataset, we decided to focus on a particular set of variables such as the Likelihood to Recommend and NPS type. Since our Analysis is going to be dependent on these two variables, we decided to remove blank and NA values in these two columns.

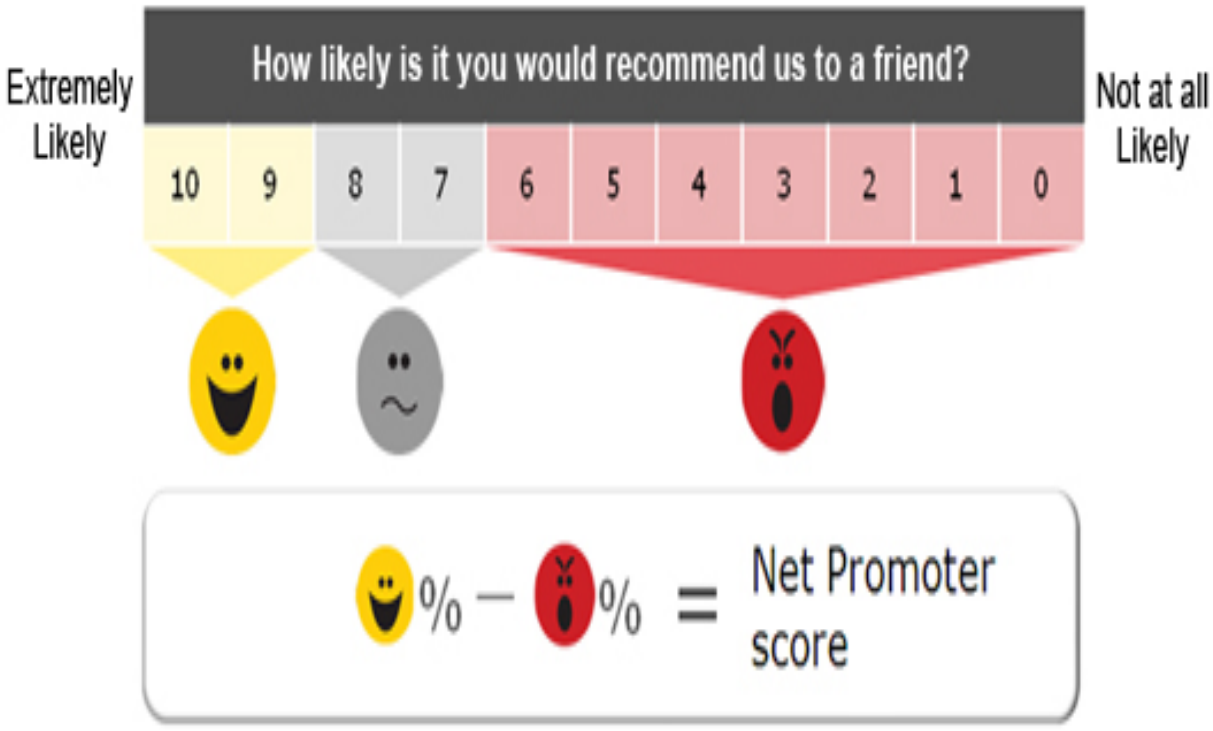
We tried approaching other methods like replacing the mean, median values in the place of NA which would reduce the quality of data and produce highly skewed results. To further narrow down, we considered the highest number of customers for a particular country and decided to analyze only the United States of America to generate better results.

Business Questions?

- How to improve the NPS score for hotels in USA?
- What are the services provided by a particular hotel that influence likelihood to recommend?
- Do most customers travel for business or leisure?
- Which state has the highest number of customers?
- Which regions in USA have highest number of promoters and detractors?

What is NPS score?

The Net Promoter Score is a loyalty metric developed by Fred Reichheld. It is used to measure the loyalty between the provider and a consumer. The Net Promoter Score is calculated based on responses to a single question: How likely is it that you would recommend our company/ product/ service to a friend or colleague? The scoring for this answer is most often based on a 0 to 10 scale. Those who respond with a score of 9 to 10 are called Promoters, and are considered likely to exhibit value-creating behaviors, such as buying more, remaining customers for longer, and making more positive referrals to other potential customers. Those who respond with a score of 0 to 6 are labeled Detractors, and they are believed to be less likely to exhibit the value-creating behaviors. Responses of 7 and 8 are customers who are "passively satisfied" and their behavior falls in the middle of Promoters and Detractors. The Net Promoter Score is calculated by subtracting the percentage of Detractors from the percentage of Promoters.



Net Promoter System practitioners ask customers the reasons for their ratings using open-ended questions. Their answers are then used by the organization to not only address customer issues but in turn also increase their NPS score by increasing the number of promoters and reducing the number of detractors.

Promoters (9 or 10)

These are the loyal customers who are satisfied with the services and enthusiastically recommend the hotel to their friends and family. They account for more than 80 percent referrals in most business.

Passive (7 or 8)

These types of customers are satisfied. However, they are not very enthusiastic about recommending the services to their friends or other prospective customers.

Detractors (0 to 6)

Detractors are unhappy customers. These people are primarily responsible for negative publicity of the hotel. These customers are bad for the hotel. However, they are important because they force the hotel to review its business practices and policies.

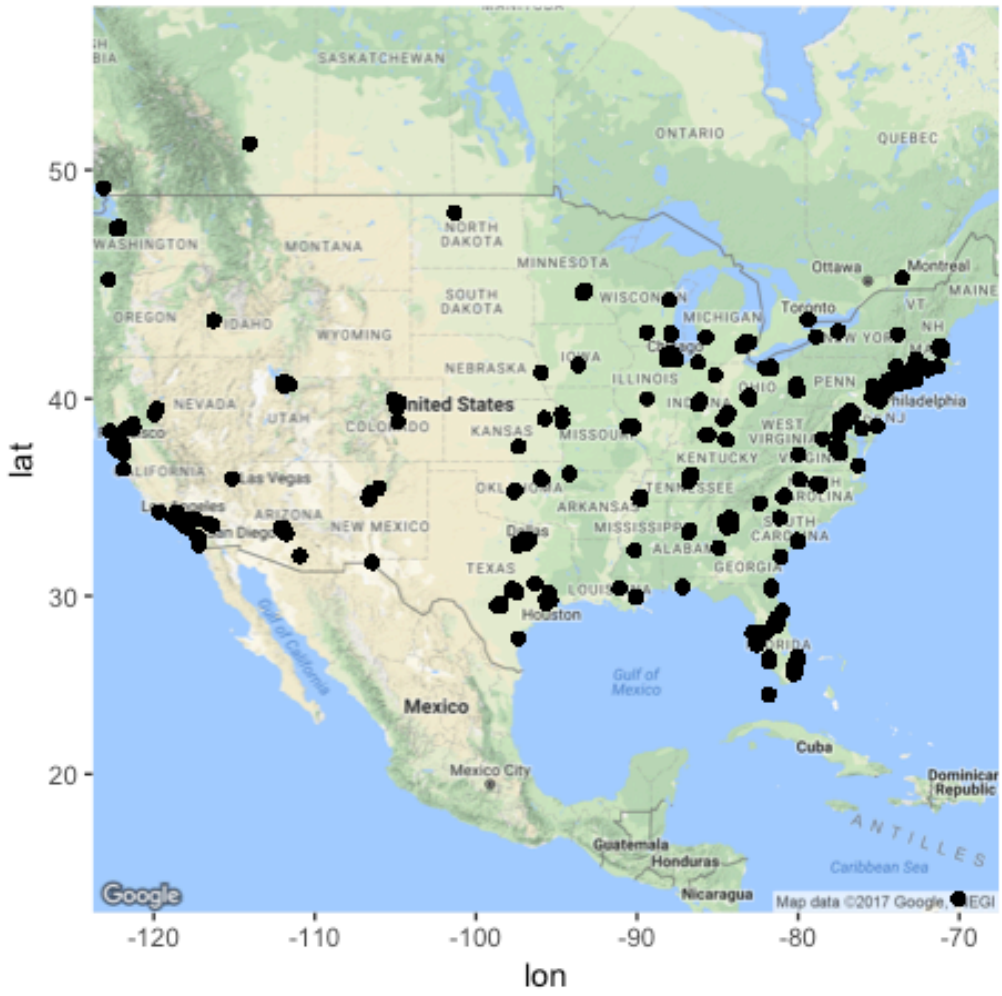
Descriptive Statistics

We generated a word cloud to see which country has the highest number of customers. We used the customer count as the frequency to generate the word cloud. We find that United States has the maximum number of customers. Hence, we conduct the analysis on USA only.

Code:

```
HotelCountryWordCloudFrame <- SpecifiedFreqMatrix(LocationDataSetYearly$CountryofHotel)
HotelCountryWordCloud <- wordcloud(HotelCountryWordCloudFrame$Name,HotelCountryWordCloudFrame$freq,colors = brewer.pal(3,"Dark2"),scale=c(2,1))
#
```

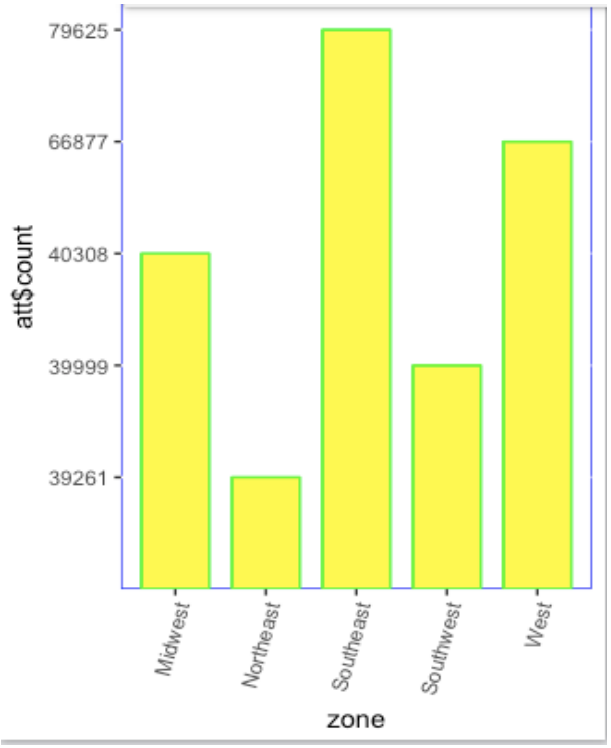
We first found out the cities in which the hotels are located. The black dots on the map represent the location of hotels in United States. We can observe that most of the hotels are concentrated on the east coast. Very few hotels are in the central region.

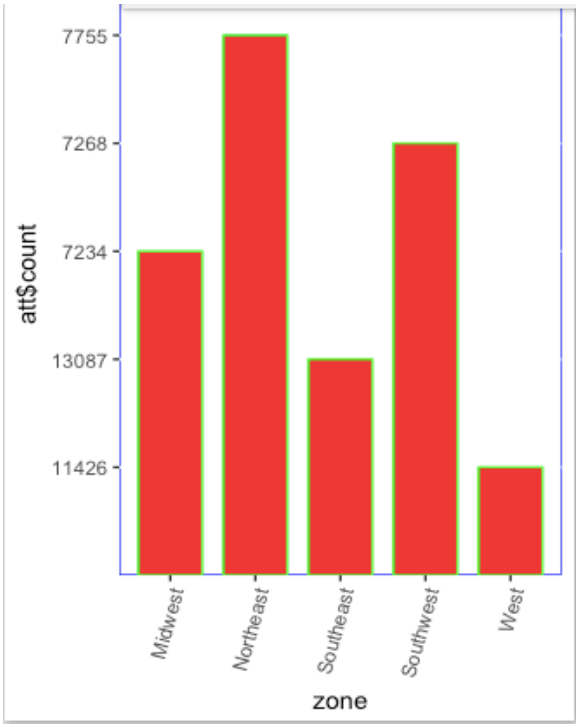


Code:

```
map <- get_map(location = "United States of America", zoom = 4)
mappoi <-ggmap(map)
mapPoi <- ggmap(map) + geom_point(data = america ,aes( x = america$`Property Longitude_PL` , y = america$`Property Latitude_PL`))
mapPoi

library(rworldmap)
newmap <- getMap(resolution = "low")
plot(newmap, xlim = c(-139.3, -58.8) , ylim = c(13.5, 55.7), asp = 1)

points(america$`Property Longitude_PL`, america$`Property Latitude_PL`, col = "red", cex = .6)
```

We have analyzed the regional distribution of promoters and detractors in USA. The southeast has the highest number of promoters and the northeast has the highest number of detractors.

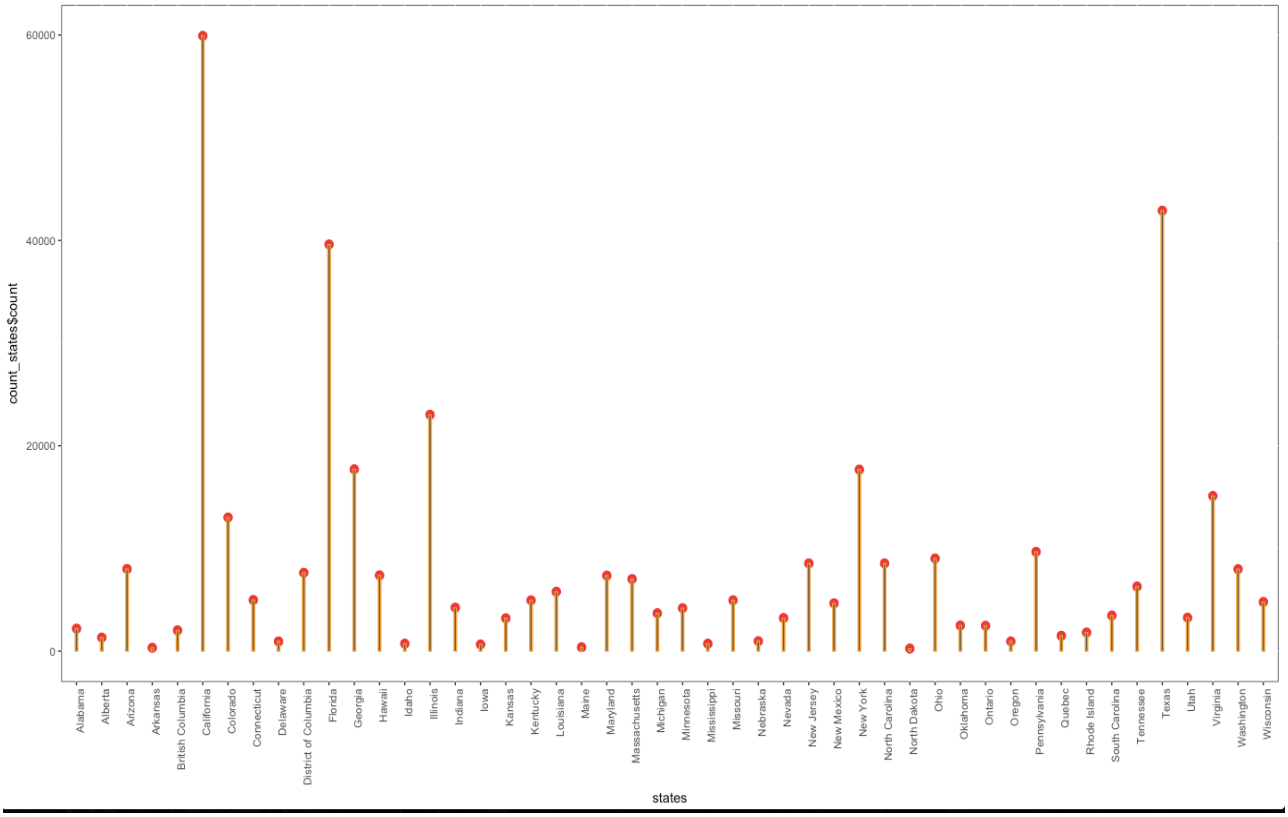




Code:

```
americaz$`US Region_PL` <- as.factor(americaz$`US Region_PL`)
americaz_detractor$`US Region_PL` <- as.factor(americaz_detractor$`US Region_PL`)
summary(americaz$`US Region_PL`)
summary(americaz$`US Region_PL`)

count_zone <- as.table(summary(americaz$`US Region_PL`))
count_det <- as.table(summary(americaz_detractor$`US Region_PL`))

names(count_zone) <- c("count")
names(count_det) <- c("detcount")

count_zone_det$zonal <- row.names(count_zone_det)

gg <- ggplot(count_zone,aes(x=zonal))
gg <- gg + geom_col(aes(y = count_zone$count), width = .1, color = "green")
gg <- gg + theme (axis.text.x = element_text(angle = 75,hjust = 1))
gg <- gg + theme (panel.background = element_rect(fill = 'white', colour = 'blue'))
gg
```

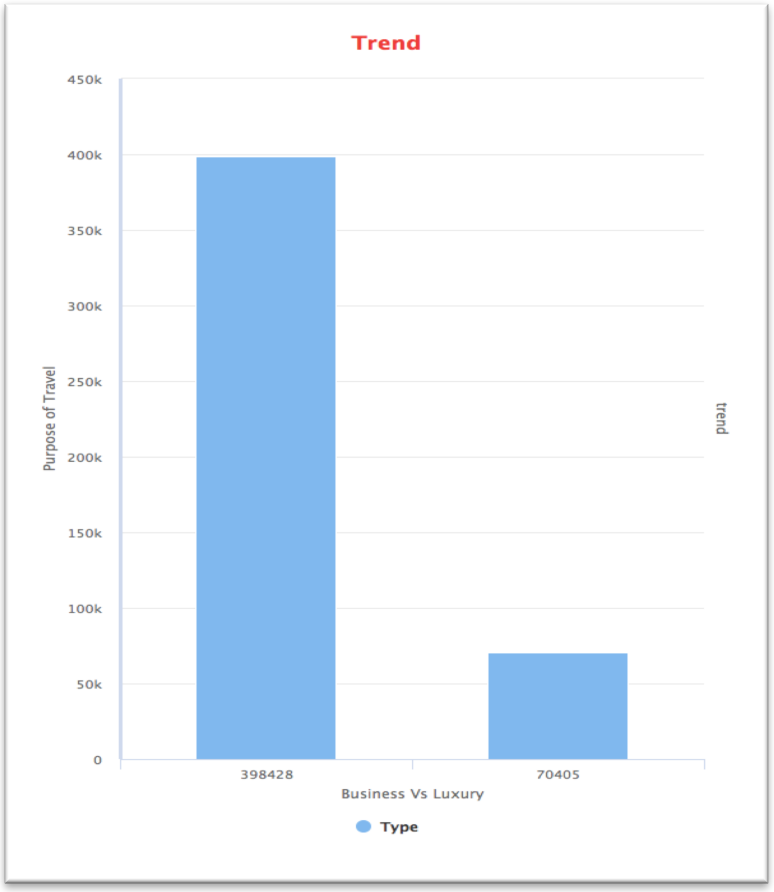
Next, we considered the number of customers per state. We found that California, Texas and Florida have the highest number of customers.



Code:

```
america_sample$State_PL <- as.factor(america_sample$State_PL)
count_states <- as.data.frame(summary(america_sample$State_PL))
count_states <- as.matrix(count_states)
count_states <- t(count_states)
View(count_states)
names(count_states) <- c("count")
count_states$state <- row.names(count_states)


gt <- ggplot(count_states,aes(x=state))
gt <- gt + geom_point(aes(y = count_states$count), size = 3, color = "red1")
gt <- gt + geom_col(aes(y = count_states$count), width = .1, color = "orange")
gt <- gt + theme (axis.text.x = element_text(angle = 90,hjust = 1))
gt <- gt + theme (panel.background = element_rect(fill = 'white', colour = 'black'))
gt
```

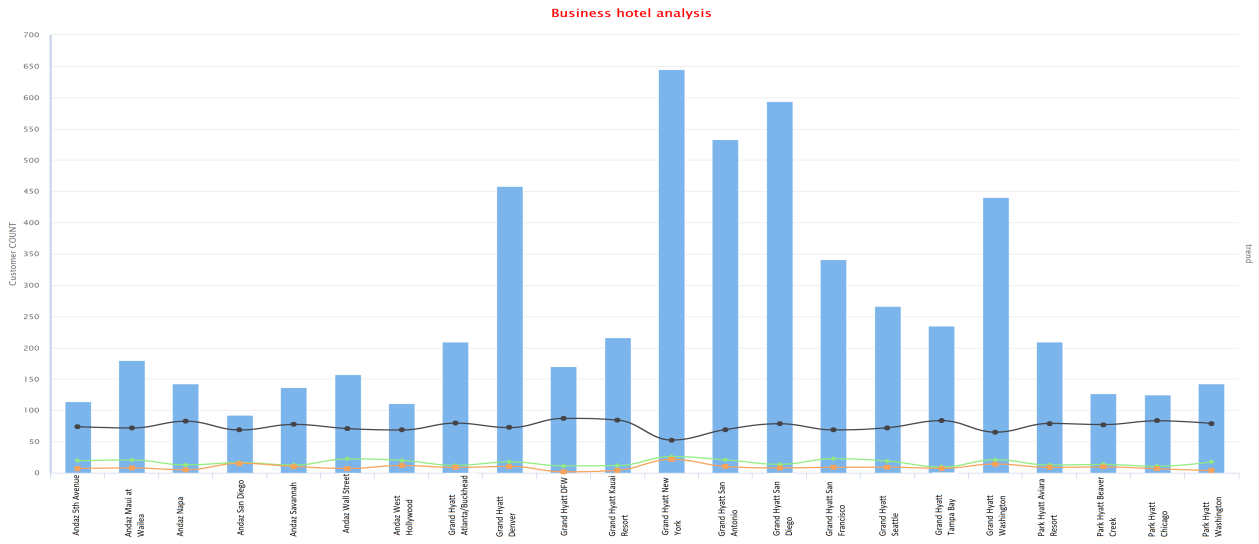
For further analysis, we considered the number of customers who visit the hotels for business or leisure. The plot below shows the total number of customers who visited for business and leisure. It is evident that most customers who are using the hotels are visiting for business purpose.



Code:

```
cal_business <- america_sample[america_sample$POV_CODE_C=="BUSINESS"]
cal_leisure <- america_sample[america_sample$POV_CODE_C=="LEISURE"]
count <- c(nrow(cal_business),nrow(cal_leisure))
names(count) <- c("Business","LEISURE")

highchart()
library(highcharter)
highchart() %>% hc_title(text = "<b>Purpose of Visit</b>",
                         margin = 20, align = "center",
                         style = list(color = "red", useHTML = TRUE)) %>%
  hc_yAxis_multiples(
    list(lineWidth = 3,title = list(text = "Customer COUNT"),min=0),
    list(showLastLabel = FALSE, opposite = TRUE,title = list(text = "trend"))
  ) %>%
  hc_xAxis(title=list(text ="Business   Leisure" ),categories = count) %>%
  hc_add_series(name = "customers", data = count ,type = "column") %>%
   hc_plotOptions(series = list(stacking = FALSE)) %>%
  hc_chart(type = "column")
```

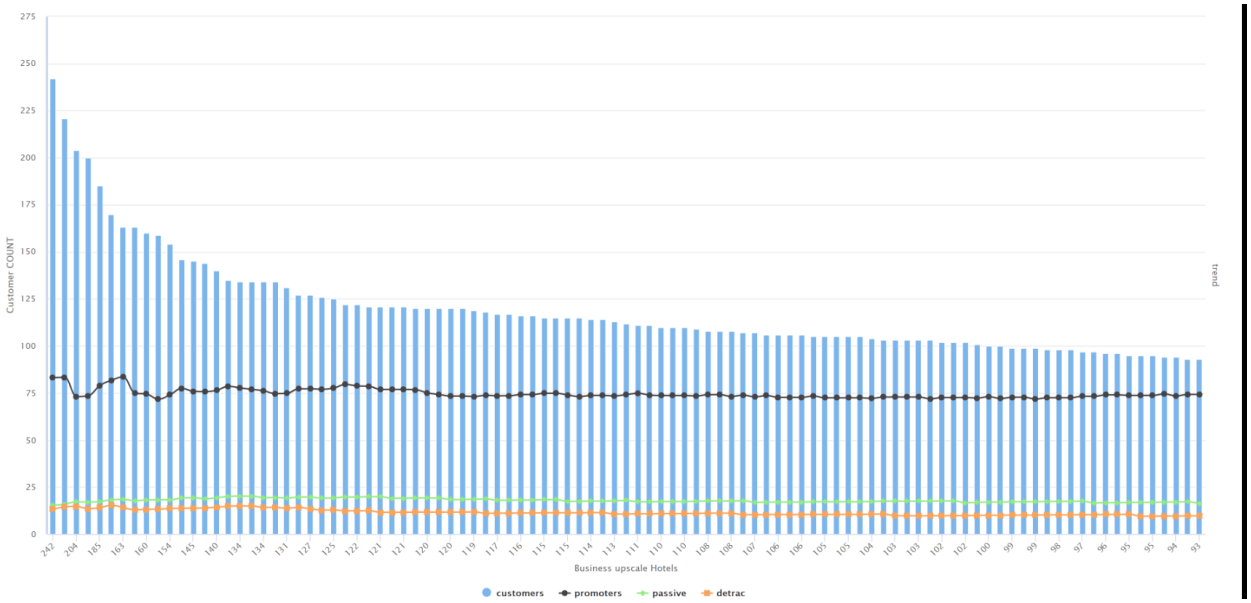
As most customers travel for business purpose, we have focused our analysis on the hotels which customers visit for business trips. Hotels which customers visit for business purpose can further be divided into two categories – luxury and upscale. The visualization below shows trends for luxury hotels. The black spline represents the trend of promoters for different hotels. Green spline represents the trend for passive customers and the yellow spline shows the trend for the detractors. The region in between black and yellow line represents the NPS score. From this analysis, we can conclude that maximum number of business travelers stay in New York. However, the detractor percent is significantly high in New York, when compared to other locations.



Code:

```
cust <- business_luxury$customer
prom <- business_luxury$promoter_percent
pass <- business_luxury$passive_perc
detrac <- business_luxury$detractor_perc
dag <- data.frame(cust,prom,pass,detrac)


library(ggplot2)

highchart() %>% hc_title(text = "<b>Business hotel analysis</b>",
                         margin = 20, align = "center",
                         style = list(color = "red", useHTML = TRUE)) %>%
  hc_yAxis_multiples(
    list(lineWidth = 3,title = list(text = "Customer COUNT"),min=0),
    list(showLastLabel = FALSE, opposite = TRUE,title = list(text = "trend"))
  ) %>%
  hc_xAxis(title=list(text ="Business luxury Hotels" ),categories = cust) %>%
  hc_add_series(name = "customers", data = cust ,type = "column") %>%
  hc_add_series(name = "promoters", data = prom  ,type = "spline") %>%
  hc_add_series(name = "passive", data = pass ,type = "spline") %>%
  hc_add_series(name = "detrac", data = detrac ,type = "spline") %>%
  hc_plotOptions(series = list(stacking = FALSE)) %>%
  hc_chart(type = "column")
```

However, for upscale hotels we found that the trends for promoters, detractors and passive remains the same. This has been shown in the visualization below.



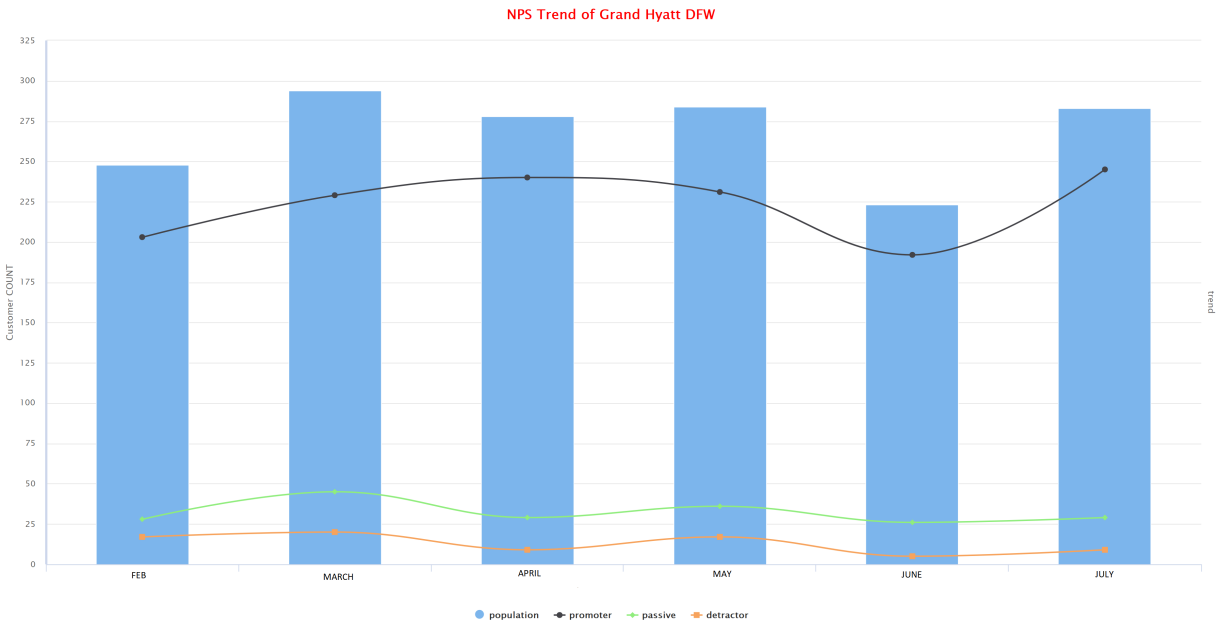
Code:

```
cust1 <- business_upscale$customer
prom1 <- business_upscale$promoter_percent
pass1 <- business_upscale$passive_perc
detrac1 <- business_upscale$detractor_perc
barplot(cust1)

highchart() %>% hc_title(text = "<b>Business hotel analysis</b>",
                         margin = 20, align = "center",
                         style = list(color = "red", useHTML = TRUE)) %>%
  hc_yAxis_multiples(
    list(lineWidth = 3,title = list(text = "Customer COUNT"),min=0),
    list(showLastLabel = FALSE, opposite = TRUE,title = list(text = "trend"))
  ) %>%
  hc_xAxis(title=list(text ="Business upscale Hotels" ),categories = cust1) %>%
  hc_add_series(name = "customers", data = cust1 ,type = "column") %>%
  hc_add_series(name = "promoters", data = prom1  ,type = "spline") %>%
  hc_add_series(name = "passive", data = pass1 ,type = "spline") %>%
  hc_add_series(name = "detrac", data = detrac1 ,type = "spline") %>%
  hc_plotOptions(series = list(stacking = FALSE)) %>%
  hc_chart(type = "column")
```

After conducting the analysis for luxury hotels, we decided to analyze the NPS scores of various hotels in the USA in order to select the one which currently has the highest NPS. So, we created a table to display the NPS scores of various hotels in the USA, so that we could choose the hotel with highest NPS as a benchmark and further analyze on why it is performing well.

X Show in new window	customer	promoter	detractor	passive	promoter_percent	detractor_perc	passive_perc	NPS_SCORE
Grand Hyatt DFW	170	148	3	19	87.05882	1.764706	11.176471	85.29412
Grand Hyatt Kauai Resort and Spa	216	182	9	25	84.25926	4.166667	11.574074	80.09259
Andaz Napa	143	118	7	18	82.51748	4.895105	12.587413	77.62238
Park Hyatt Chicago	125	104	8	13	83.20000	6.400000	10.400000	76.80000
Grand Hyatt Tampa Bay	235	196	17	22	83.40426	7.234043	9.361702	76.17021
Park Hyatt Washington	143	113	5	25	79.02098	3.496503	17.482517	75.52448
Grand Hyatt Atlanta in Buckhead	209	166	18	25	79.42584	8.612440	11.961722	70.81340
Grand Hyatt San Diego	594	466	47	81	78.45118	7.912458	13.636364	70.53872
Park Hyatt Aviara Resort	209	165	18	26	78.94737	8.612440	12.440191	70.33493
Park Hyatt Beaver Creek Resort and Spa	127	98	12	17	77.16535	9.448819	13.385827	67.71654
Andaz Savannah	137	106	14	17	77.37226	10.218978	12.408759	67.15328
Andaz 5th Avenue	114	84	8	22	73.68421	7.017544	19.298246	66.66667
Andaz Maui at Wailea	180	129	14	37	71.66667	7.777778	20.555556	63.88889
Andaz Wall Street	157	111	11	35	70.70064	7.006369	22.292994	63.69427
Grand Hyatt Seattle	266	192	24	50	72.18045	9.022556	18.796992	63.15789
Grand Hyatt Denver	458	331	47	80	72.27074	10.262009	17.467249	62.00873
Grand Hyatt San Francisco	341	234	30	77	68.62170	8.797654	22.580645	59.82405
Grand Hyatt San Antonio	533	369	54	110	69.23077	10.131332	20.637899	59.09944
Andaz West Hollywood	111	76	13	22	68.46847	11.711712	19.819820	56.75676
Andaz San Diego	92	63	14	15	68.47826	15.217391	16.304348	53.26087
Grand Hyatt Washington	440	286	62	92	65.00000	14.090909	20.909091	50.90909
Grand Hyatt New York	645	338	139	168	52.40310	21.550388	26.046512	30.85271

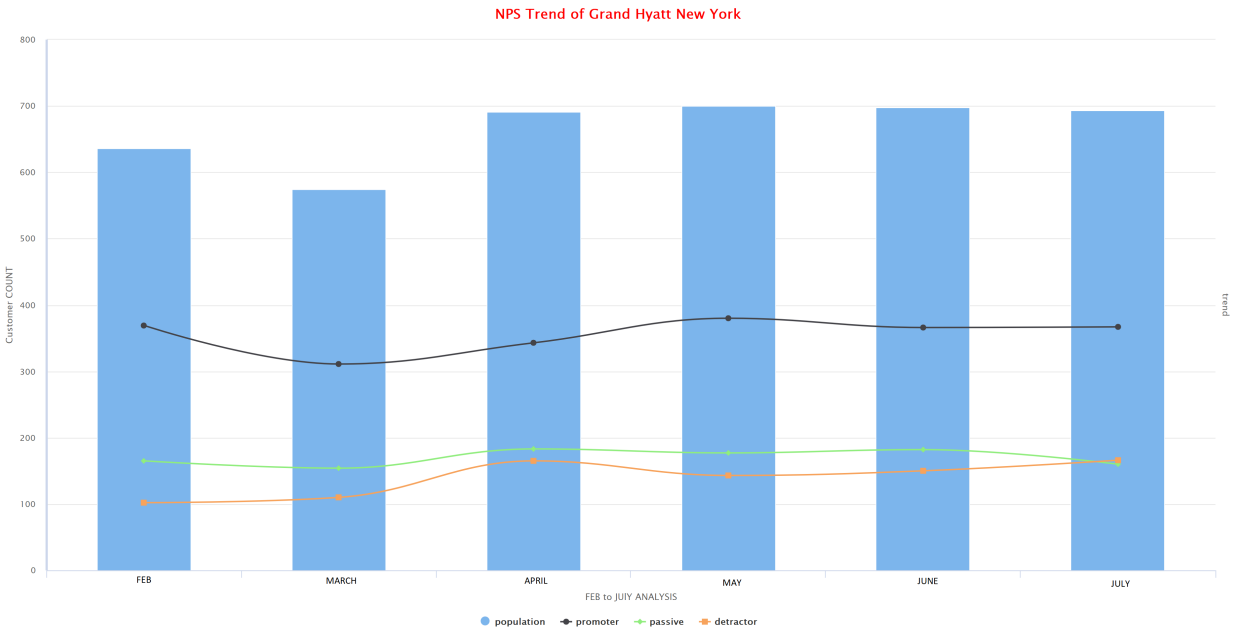
We found that Grand Hyatt DFW had the highest percent of promoters and the least percent of detractors. This resulted in a high NPS score. We selected this particular hotel to set a benchmark for other hotels in the same segment to reach a good NPS score. This analysis was conducted for a six-month period from February to July. Over a six-month period, the number of promoters had low variability.



Code:

```
customers <- c(ncust,ncust2,ncust3,ncust4,ncust5,ncust6)
prom <- c(nprom,nprom2,nprom3,nprom4,nprom5,nprom6)
pass <- c(npass,npass2,npass3,npass4,npass5,npass6)
detrac <- c(ndet,ndet2,ndet3,ndet4,ndet5,ndet6)
### highcharts

library(highcharter)
highchart() %>% hc_title(text = "<b>NPS Trend of Grand Hyatt DFW </b>",
                         margin = 20, align = "center",
                         style = list(color = "red", useHTML = TRUE)) %>%
  hc_yAxis_multiples(
    list(lineWidth = 3,title = list(text = "Customer COUNT"),min=0),
    list(showLastLabel = FALSE, opposite = TRUE,title = list(text = "trend"))
  ) %>%
  hc_xAxis(title=list(text = "FEB to JUNE ANALYSIS"),categories = customers) %>%
  hc_add_series(name = "population", data = customers ,type = "column") %>%
  hc_add_series(name = "promoter", data = prom ,type = "spline") %>%
  hc_add_series(name = "passive", data = pass ,type = "spline") %>%
  hc_add_series(name = "detractor", data = detrac ,type = "spline") %>%

  hc_plotOptions(series = list(stacking = FALSE)) %>%
  hc_chart(type = "column")
```

Along similar lines we found that Grand Hyatt New York had the least percentage of promoters and a highest percentage of detractors. This has been shown in the visualization below.



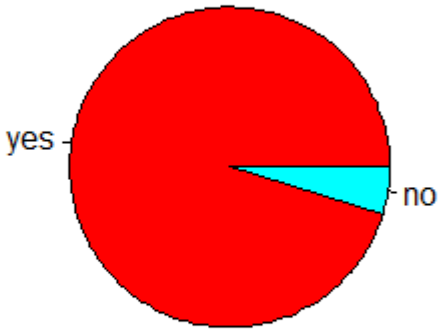
Code:

```
customers <- c(ncust,ncust2,ncust3,ncust4,ncust5,ncust6)
prom <- c(nprom,nprom2,nprom3,nprom4,nprom5,nprom6)
pass <- c(npass,npass2,npass3,npass4,npass5,npass6)
detrac <- c(ndet,ndet2,ndet3,ndet4,ndet5,ndet6)
### highcharts

library(highcharter)
highchart() %>% hc_title(text = "<b>NPS Trend of Grand Hyatt New York </b>",
                         margin = 20, align = "center",
                         style = list(color = "red", useHTML = TRUE)) %>%
  hc_yAxis_multiples(
    list(lineWidth = 3,title = list(text = "Customer COUNT"),min=0),
    list(showLastLabel = FALSE, opposite = TRUE,title = list(text = "trend"))
  ) %>%
  hc_xAxis(title=list(text = "FEB to JUlY ANALYSIS"),categories = customers) %>%
  hc_add_series(name = "population", data = customers ,type = "column") %>%
  hc_add_series(name = "promoter", data = prom ,type = "spline") %>%
  hc_add_series(name = "passive", data = pass ,type = "spline") %>%
  hc_add_series(name = "detractor", data = detrac ,type = "spline") %>%

  hc_plotOptions(series = list(stacking = FALSE)) %>%
  hc_chart(type = "column")
```

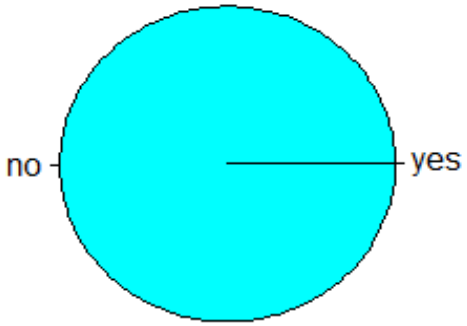
Amenities Used by Promoters

To increase the percentage of promoters we need to understand what amenities are mostly used by promoters. We have used pie charts to gain insight in this regard.

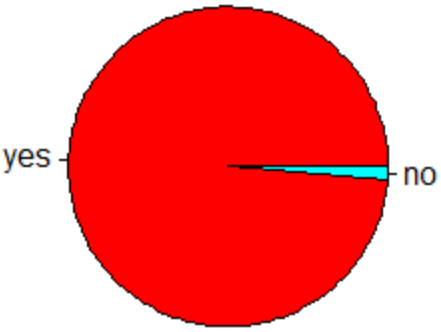
Business Centre Used



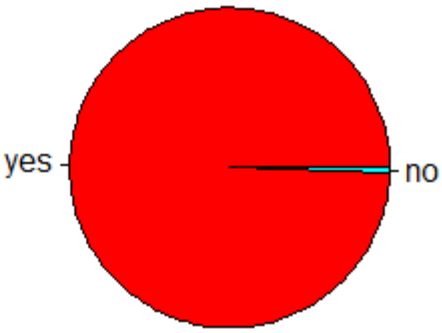
Conference Centre Used



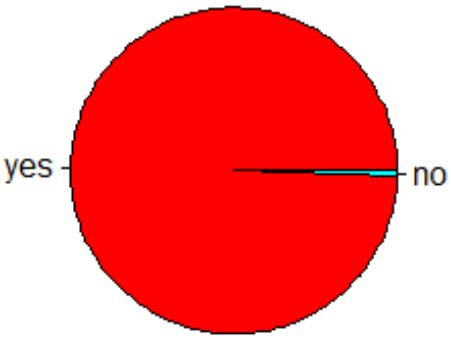
Dry Cleaning Used



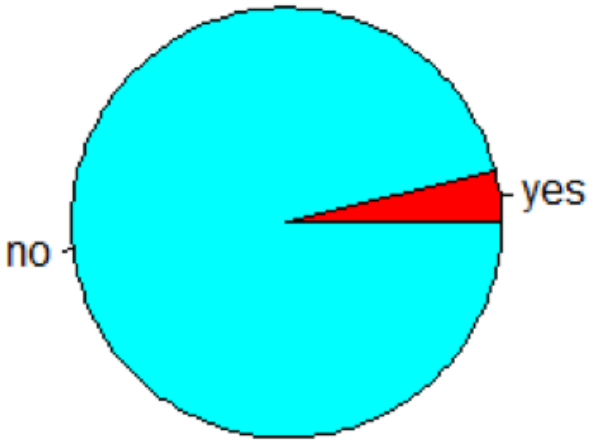
Elevators Used



Fitness Centres Used



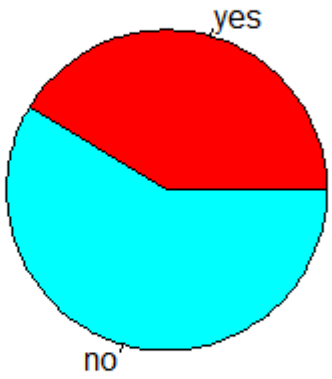
GOLF COURSES USED



Indoor Pool Used

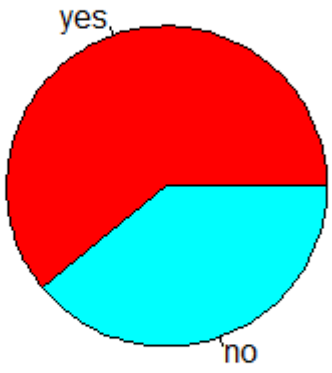
Laundry Used

Limo Services Used

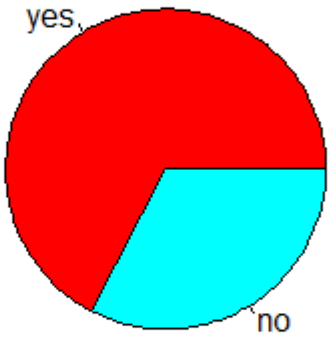


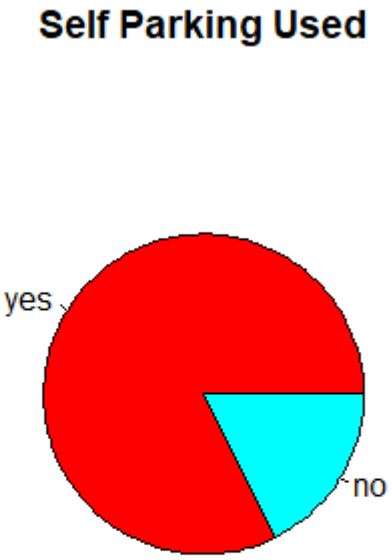
Mini Bar Used

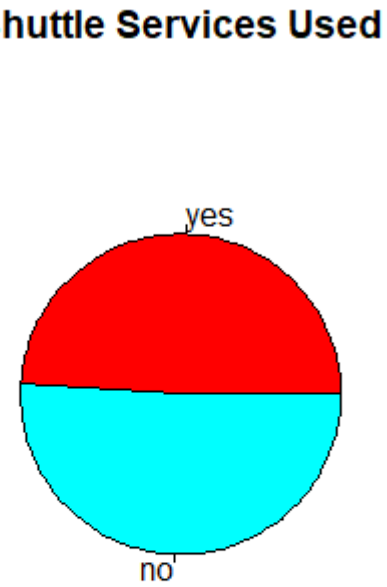
Outdoor Pool Used



Restaurant Used







From these pie charts, we can see that most promoters made use of the following amenities:

- Business center
- Dry cleaning
- Fitness center
- Laundry
- Restaurant
- Self-Parking

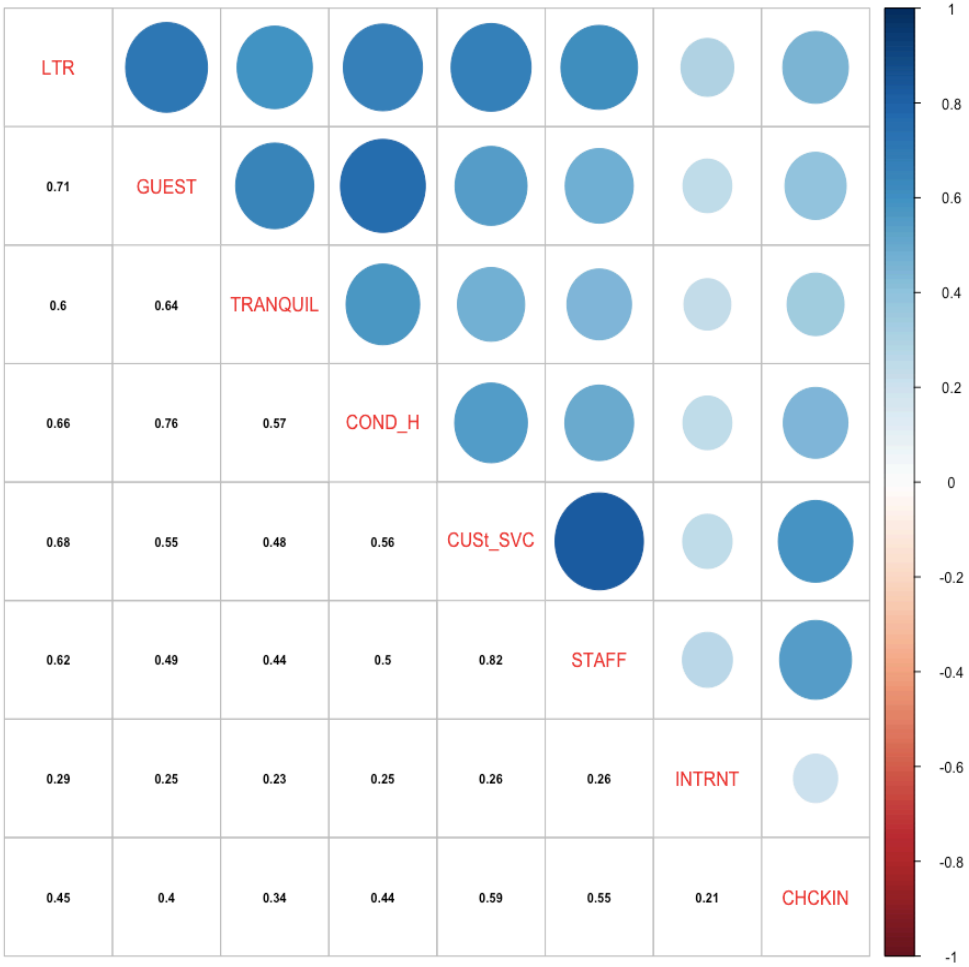
After generating the descriptive statistics, we got a clear idea about the data. Next, we moved on to building predictive models. We used multiple approaches to understand the importance of different variables and also their impact on the likelihood to recommend.

DECISION TREE

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. We have created a decision tree to understand the variable importance (after removing the unnecessary factors such as hotel names, guest country name etc.) so that we can determine how the likelihood to recommend is influenced by various variables in the dataset. The decision tree below helps us to gain an understanding of why a customer gives a particular rating.

We can clearly observe that the main factors which influence a customer's decision in becoming a Promoter are the Guest Room Satisfaction, Condition of Hotel and the Quality of Customer Service.

CORRELATION PLOT



Code:

```
library(corrplot)
library(corrgram)
names(america_sample)
america_ltr <- america_sample[,c(46,48:54)]
names(america_ltr) <- c("LTR","GUEST","TRANQUIL","COND_H","CUSt_SVC","STAFF","INTRNT","CHCKIN")
america_ltr <-na.omit(america_ltr)
correlation_dat <- cor(america_ltr)
corplot1 <- corrplot.mixed(correlation_dat, lower.col = "black", number.cex = .7)
```

From the scatterplot, we see that guestroom satisfaction and customer service have strong positive correlation with the likelihood to recommend.

LINEAR MODELLING

Using data of 6 months for the United States region, we have conducted Linear Modelling on where the customers have rated Hyatt Hotels based on the below 7 factors.

- Guest Room Satisfaction Metric
- Tranquility Metric
- Condition of Hotel Metric
- Quality of Customer Service Metric
- Internet Satisfaction Metric
- Quality of Check in Process
- Staff Cared Metric

For linear modelling, likelihood to recommend is the dependent variable and all the above seven factors are considered as the independent variables. We have run linear modelling individually for each factor and then for a combination of all the factors. The outputs observed have been recorded below:

Independent Variable	Dependent Variable	R Squared Value
Likelihood to Recommend	Guest Room Satisfaction Metric	0.3240
Likelihood to Recommend	Tranquility Metric	0.2864
Likelihood to Recommend	Condition of Hotel Metric	0.3788
Likelihood to Recommend	Quality of Customer Service Metric	0.4936
Likelihood to Recommend	Internet Satisfaction Metric	0.0218
Likelihood to Recommend	Quality of Check-In Process	0.3078
Likelihood to Recommend	Staff Cared Metric	0.1876
Likelihood to Recommend	• Guest Room Satisfaction Metric • Tranquility Metric • Condition of Hotel Metric • Quality of Customer Service Metric • Internet Satisfaction Metric • Quality of Check-In Process • Staff Cared Metric	0.649

We observe that the R-Squared value is highest for a combination of the below factors:

- Guest Room Satisfaction Metric
- Tranquility Metric
- Condition of Hotel Metric
- Quality of Customer Service Metric
- Internet Satisfaction Metric
- Quality of Check-In Process
- Staff Cared Metric

Code:

```
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Customer_SVC_H +
    Condition_Hotel_H + Tranquility_H + Internet_Sat_H + +Check_In_H +
    Staff_Cared_H, data = america_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9260 -0.1810  0.0279  0.4422  5.5449

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.948059   0.056313 -34.593   <2e-16 ***
Guest_Room_H      0.340932   0.006977  48.865   <2e-16 ***
Customer_SVC_H    0.355725   0.009627  36.952   <2e-16 ***
Condition_Hotel_H 0.178231   0.007848  22.711   <2e-16 ***
Tranquility_H     0.132404   0.004713  28.093   <2e-16 ***
Internet_Sat_H    0.046053   0.003242  14.206   <2e-16 ***
Check_In_H        0.001309   0.006217   0.211    0.833
Staff_Cared_H     0.137358   0.007945  17.289   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.038 on 26138 degrees of freedom
  (37402 observations deleted due to missingness)
Multiple R-squared:  0.6491,    Adjusted R-squared:  0.649
F-statistic:  6908 on 7 and 26138 DF,  p-value: < 2.2e-16
```

We observe that the R-Squared value is 0.64 which means that a combination of all the 7 factors account for 64% of the rating for 'Likelihood to Recommend'. In other words, 64% variability in the likelihood to recommend can be explained by our explanatory variables. Also, except the Quality of Check-In Process, all other variables have a low p-value which means that they are statistically significant.

ASSOCIATION RULES

We have used Association Rules mining to determine the factors which would have a maximum impact on the NPS score of Hyatt Hotels. We set the RHS in the model as NPS type Promoters in order to analyze the factors which make a customer a promoter. Based on this, we generated some rules and picked up the top eight rules which had the highest value of 'Lift'. This enabled us to have an understanding of the combination of best factors which can be used to turn a customer into a promoter.

Code:

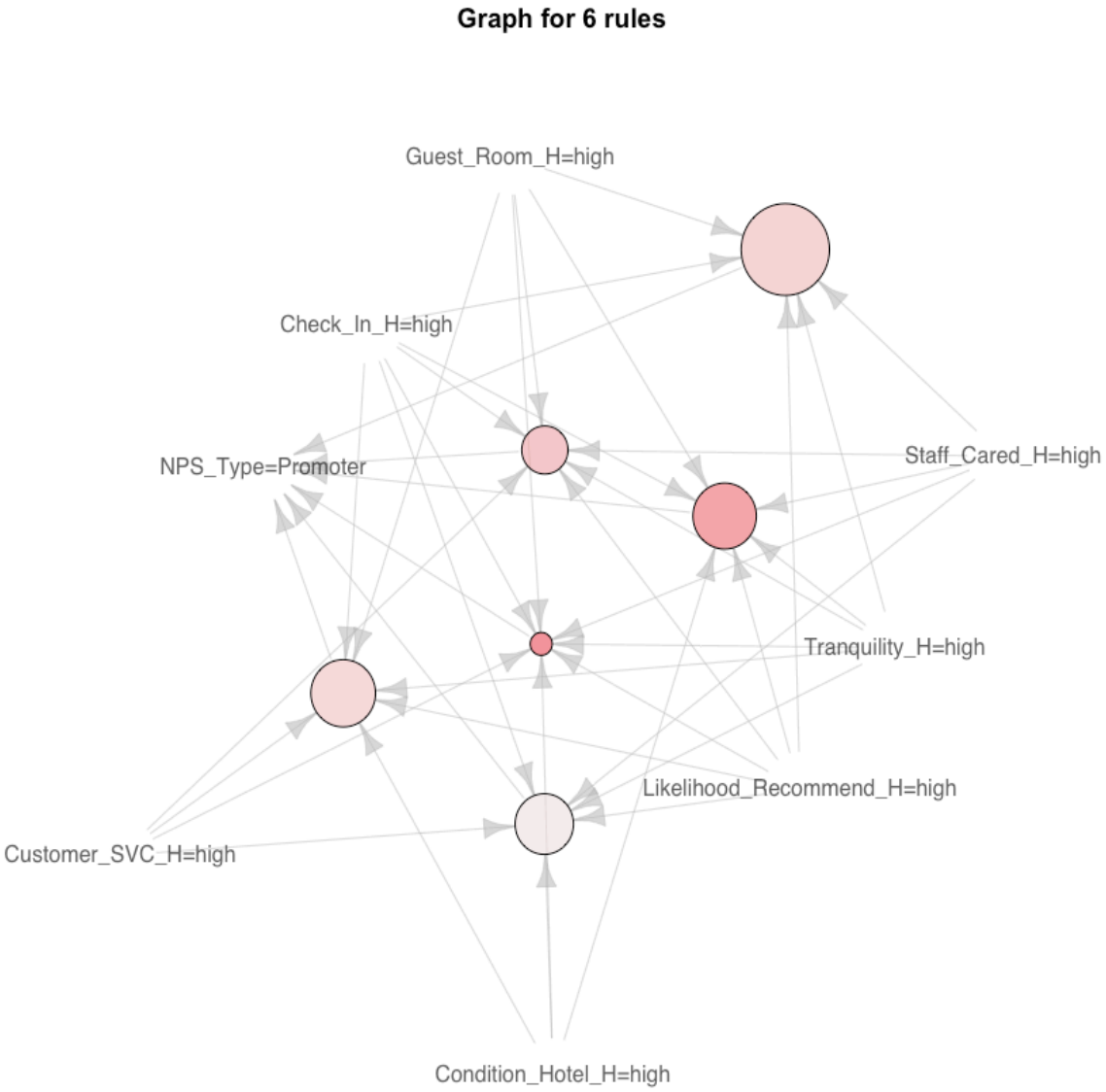
```
library(highcharter)
library(arules)
library(arulesViz)
america <- clean[clean$Country_PL=="United States",]
america_sample <- america[!(is.na(america$Likelihood_Recommend_H)),]
america_samplep <- america_sample[america_sample$`Hotel Name-Long_PL`=="Grand Hyatt DFW",]
america_samplep <- america_sample[,c(46:57,118)]
america_samplep <- as.data.frame(unclass(america_samplep))
america_samplep$Likelihood_Recommend_H <- as.factor(america_samplep$Likelihood_Recommend_H)
america_samplep$Guest_Room_H <- as.factor(america_samplep$Guest_Room_H)
america_samplep$Tranquility_H <- as.factor(america_samplep$Tranquility_H)
america_samplep$Condition_Hotel_H <- as.factor(america_samplep$Condition_Hotel_H)
america_samplep$Customer_SVC_H <- as.factor(america_samplep$Customer_SVC_H)
america_samplep$Staff_Cared_H <- as.factor(america_samplep$Staff_Cared_H)
america_samplep$Internet_Sat_H <- as.factor(america_samplep$Internet_Sat_H)
america_samplep$Check_In_H <- as.factor(america_samplep$Check_In_H)
america_samplep$NPS_Type <- as.factor(america_samplep$NPS_Type)

library(arules)
rules <- apriori(america_samplep[,-c(2,10,11)], parameter = list(minlen=3, supp=0.03, conf=0.5),
                 appearance = list(rhs="NPS_Type=Promoter",default="lhs"))
ruz <- sort(rules, decreasing=TRUE,by=c("confidence","lift"))
inspectDT(ruz)
sort_asso <- head(ruz)
library(arulesViz)
plot(sort_asso, method="graph", control=list(type="items"))
plot(ruz,col="purple")
```

As per our model and combination of our support and confidence values, around 3570 were generated. Based on the highest values of both Lift and Confidence, we have taken the top six rules to be displayed below:

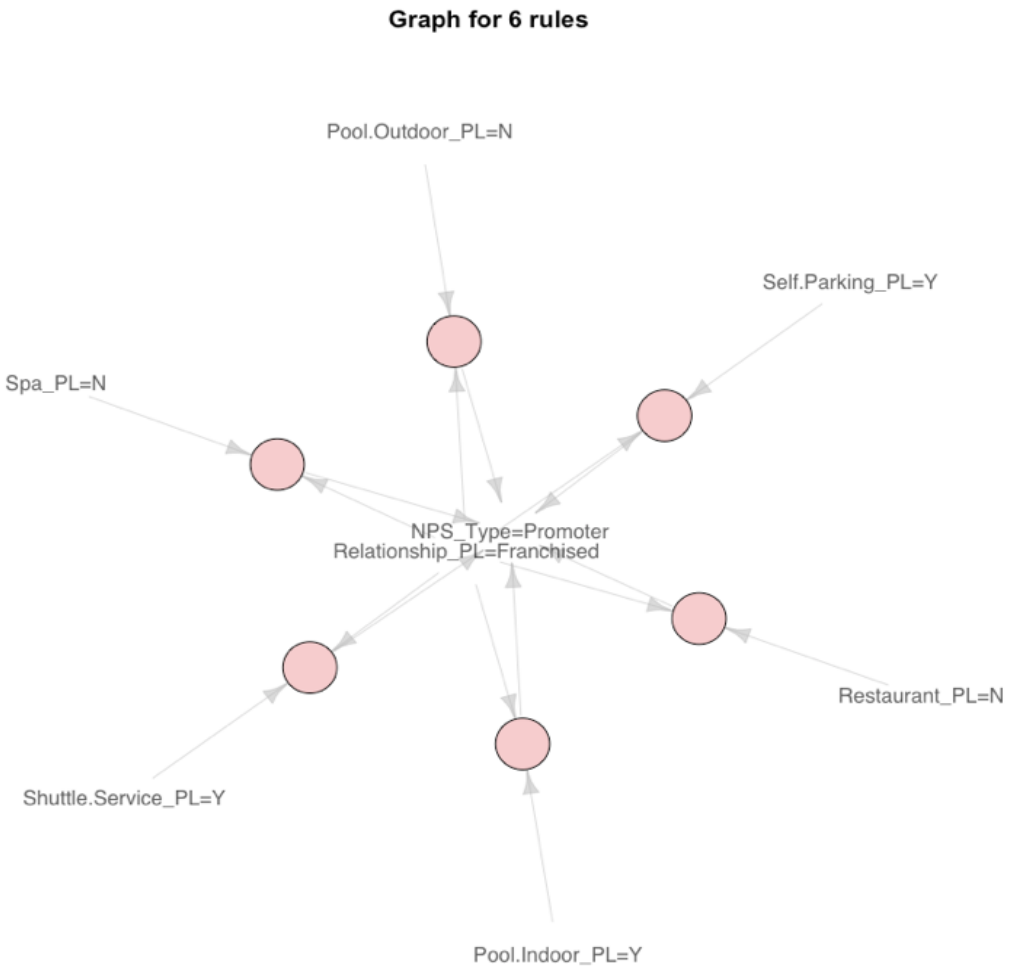
LHS	RHS	SUPPORT	CONFIDENCE	LIFT	COUNT
{Likelihood_Recommend_H=high, Guest_Room_H=high, Tranquility_H=high, Condition_Hotel_H=high, Customer_SVC_H=high, Staff_Cared_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.3350852	0.9216288	1.335198	157099
{Likelihood_Recommend_H=high, Guest_Room_H=high, Tranquility_H=high, Condition_Hotel_H=high, Staff_Cared_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.3420429	0.920652	1.333782	160361
{Likelihood_Recommend_H=high, Guest_Room_H=high, Tranquility_H=high, Customer_SVC_H=high, Staff_Cared_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.339157	0.9187337	1.331003	159008
{Likelihood_Recommend_H=high, Guest_Room_H=high, Tranquility_H=high, Staff_Cared_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.3462192	0.917689	1.32949	162319
{Likelihood_Recommend_H=high, Guest_Room_H=high, Tranquility_H=high, Condition_Hotel_H=high, Customer_SVC_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.3422541	0.9174857	1.329195	160460
{Likelihood_Recommend_H=high, Tranquility_H=high, Condition_Hotel_H=high, Customer_SVC_H=high, Staff_Cared_H=high, Check_In_H=high}	{NPS_Type=Promoter}	0.3411855	0.9155163	1.326342	159959

On running the association rules mining for the Grand Hyatt Hotel at DFW, we have taken the top 6 rules and plotted the below graph to determine the factors which influence promoters in the hotel with highest NPS:

PLOT



We have also used the Association rules mining to determine the amenities which would have a great impact on the NPS score of the Hyatt Hotel at DFW. Even in this scenario, we have set the RHS as NPS type promoters in order to determine the amenities which appeal the most to a promoter. The corresponding plot for the six rules having the highest Lift can be found below:

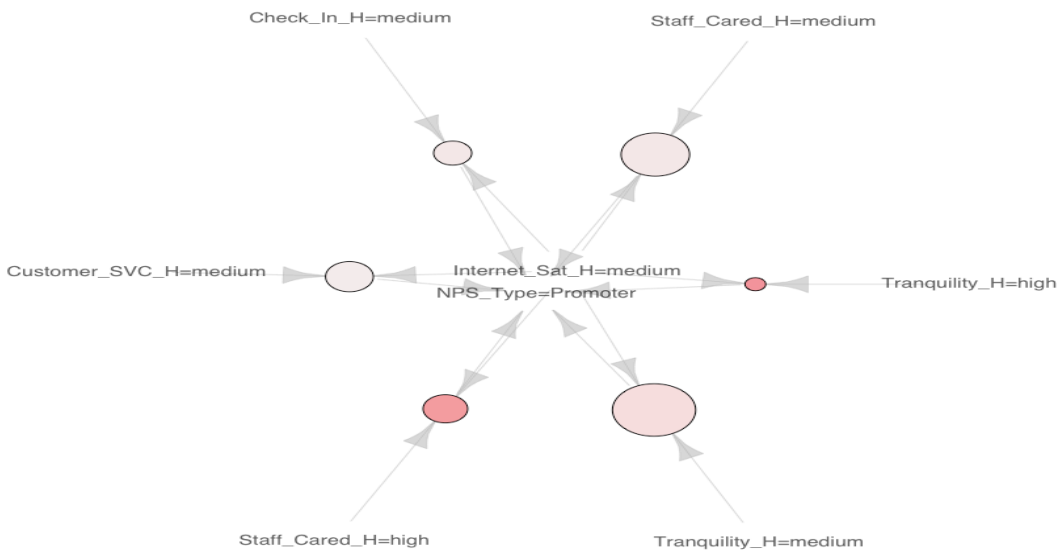


From the above plot, we can determine the most important amenities which influence a customer to become a promoter in the case of the Hyatt Hotel at DFW are:

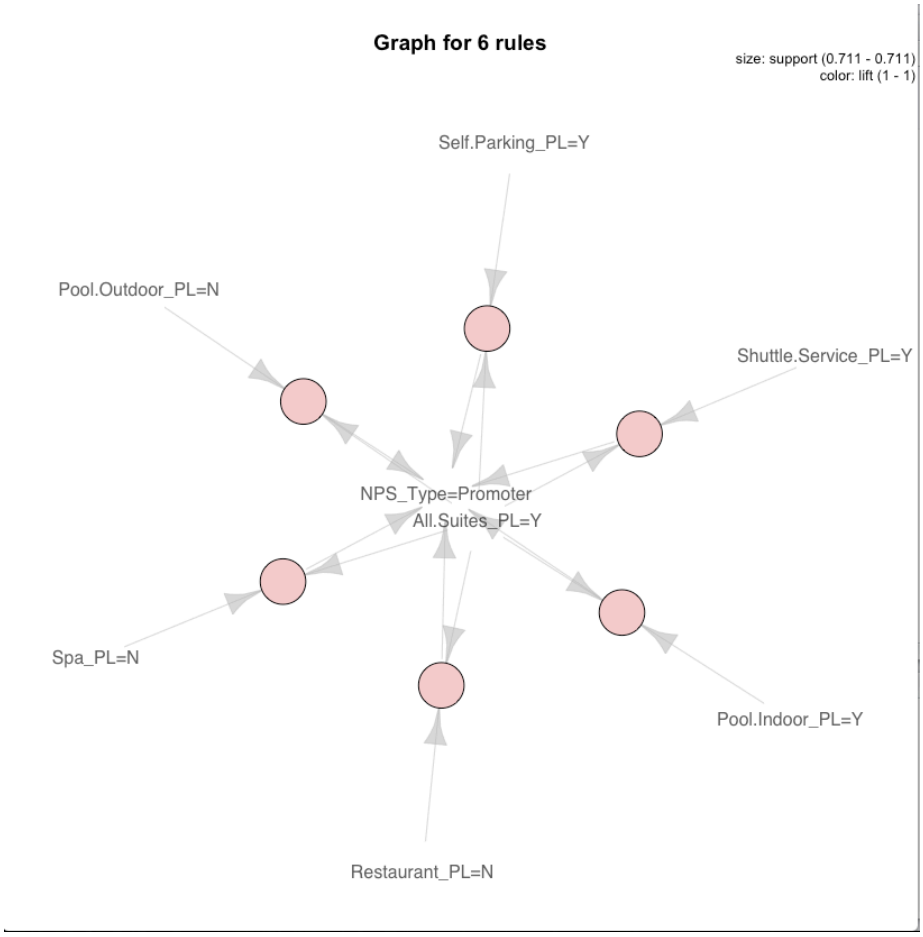
- Outdoor or indoor pool
- Self-parking
- Restaurant
- Shuttle service
- Spa

By improving these amenities in a hotel, we can influence the customer to become a promoter thereby increasing the NPS.

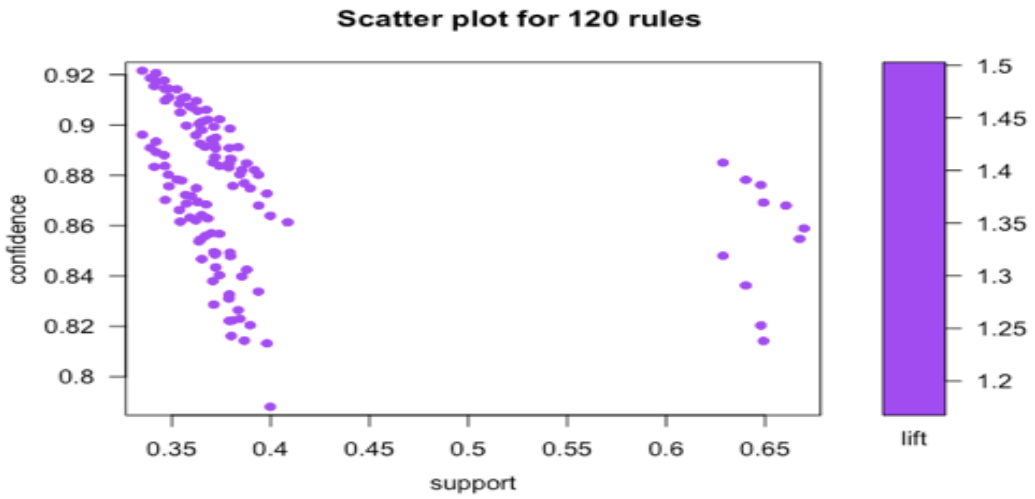
Following this, we have run the association rules mining for the Grand Hyatt Hotel having the least NPS i.e. the hotel at New York. As before, we have considered the RHS as NPS type promoters and plotted the below graph for the first six rules in order to determine the factors which influence promotors:



We have also used the Association rules mining to determine the amenities which would have a great impact on the NPS score of the Hyatt Hotel at New York. Even in this scenario, we have set the RHS as NPS type promoters in order to determine the amenities which appeal the most to a promoter. The corresponding plot for the six rules having the highest Lift can be found below:



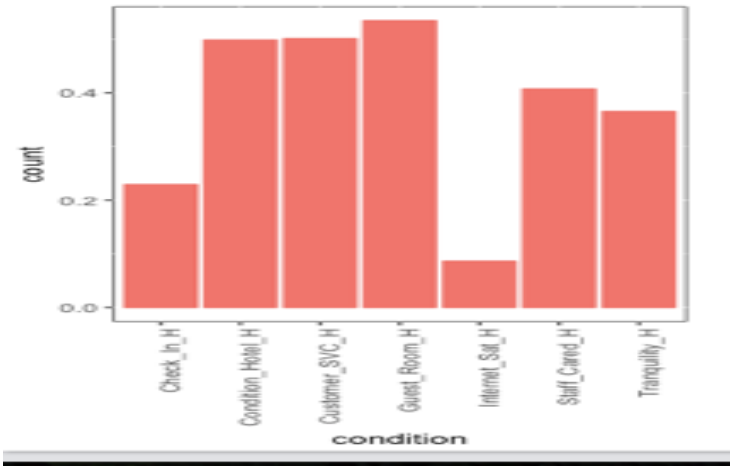
A scatter plot which we have generated in order to show the support and confidence has been given below:



SVM

To understand how the NPS TYPE is dependent on other parameters, it is necessary to determine the relationship between the NPS TYPE and several other rated metrics namely Quality of Check-In Process, Condition of Hotel, Quality of Customer Service, Guest Room Satisfaction, Internet Satisfaction, Staff Cared and Tranquility.

The below graph helps us to represent how the NPS TYPE is dependent on other parameters.



We have used the SVM model to predict how the above factors influence the NPS TYPE. The corresponding snippets of the code used for the model and the output observed have been given below:

Code:

```
america <- clean[clean$Country_PL=="United States",]
america_sample <- america[!(is.na(america$Likelihood_Recommend_H)),]
america_sampler <- america_sample[,c(46:55,117)]
america_sampler <- america_sampler[,-c(2,10)]


############ Code to convert parameters to low medium and high

america_sampler$Guest_Room_H <- ifelse(america_sampler$Guest_Room_H>=8,"high",ifelse(america_sampler$Guest_Room_H<=6,"low","medium"))
america_sampler$Tranquility_H <- ifelse(america_sampler$Tranquility_H>=8,"high",ifelse(america_sampler$Tranquility_H<=6,"low","medium"))
america_sampler$Condition_Hotel_H <- ifelse(america_sampler$Condition_Hotel_H>=8,"high",ifelse(america_sampler$Condition_Hotel_H<=6,"low","medium"))
america_sampler$Customer_SVC_H <- ifelse(america_sampler$Customer_SVC_H>=8,"high",ifelse(america_sampler$Customer_SVC_H<=6,"low","medium"))
############# Running Model#########################
set.seed(123)
split <- sample(seq_len(nrow(america_sampler)), size = floor(0.75 * nrow(america_sampler)))
train1 <- america_sampler[split, ]
test1 <- america_sampler[-split, ]

### BUILDING SVM
library(e1071)
svmModel1 <- svm(NPS_Type ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H
                 + Customer_SVC_H ,data=train1, kpar="automatic", C=5,cross=3, prob.model=TRUE)

summary(svmModel1)
```

Output:

```
Call:
svm(formula = NPS_Type ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H + Customer_SVC_H, data = train1, kpar = "automatic", C = 5,
    cross = 3, prob.model = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1111111

Number of Support Vectors:  11273

 ( 4373 1818 5082 )


Number of Classes:  3

Levels:
 Detractor Passive Promoter

3-fold cross-validation on training data:

Total Accuracy: 78.64594
Single Accuracies:
 78.19835 78.82216 78.91732
```

We find that a combination of Condition of Hotel, Quality of Customer Service, Guest Room Satisfaction and Tranquility help to maintain an accuracy of 78.64%

Validation:

To check the accuracy of our models we split the dataset into two parts- training set and testing set. 75% of the data was the training set and 25% was the test set. We got an accuracy of 60% for our linear model and an accuracy of 78.56 % for our SVM model.

Code for SVM:

```
library(Metrics)
test1 <- na.omit(test1)
test1 <- test1[!(is.na(test1$NPS_Type)),]
test1$predictions <- predict(svmModel1,test1)
View(test1)
acc1 <- accuracy(test1$predictions,test1$NPS_Type)
acc1
```

Output for SVM (accuracy)

```
> test1$predictions <- predict(svmModel1,test1)
> View(test1)
> acc1 <- accuracy(test1$predictions,test1$NPS_Type)
> acc1
[1] 0.7856273
```

Code for Linear Modelling:

```
library(Metrics)
clean <- read.csv("halfyearly.csv")
america <- clean[clean$Country_PL=="United States",]
america_sample <- america[!(is.na(america$Likelihood_Recommend_H)),]
america_sampler <- america_sample[,c(47:57,118)]
america_sampler <- america_sampler[,-c(2,10,11,12)]

set.seed(123)
split <- sample(seq_len(nrow(america_sampler)), size = floor(0.75 * nrow(america_sampler)))
train <- america_sampler[split, ]
test <- america_sampler[-split, ]

fit_5 <- lm(formula = Likelihood_Recommend_H~Guest_Room_H+Customer_SVC_H+Condition_Hotel_H+Staff_Cared_H+Internet_Sat_H ,data=(train))
summary(fit_5)

test$predictions <- predict.lm(fit_5,test)
test <- test[!(is.na(test$predictions)),]
test$predictions <- round(test$predictions)
accu <- accuracy(test$predictions,test$Likelihood_Recommend_H)
accu
```

Output for Linear Model (accuracy)

```
> test$predictions <- predict.lm(fit_5,test)
> test <- test[!(is.na(test$predictions)),]
> test$predictions <- round(test$predictions)
> accu <- accuracy(test$predictions,test$Likelihood_Recommend_H)
> accu
[1] 0.6036424
```

Recommendations to promote NPS:

The below recommendations can be made for the Business Luxury range of Hyatt hotels:

- Using descriptive statistics such as bar graphs, we found out that Hyatt should concentrate on the Business hotels as they have the maximum number of customers. Thus, Hyatt hotel chains should provide corporate firms with better deals for conference rooms and subsidized room prices.
- Amenities which a promoter uses the most are an outdoor pool, self-parking, restaurant, shuttle service and spa. These were found out using the pie charts. By making sure that these amenities are available and well maintained, the NPS can be improved. These can be improved by providing free parking, discount vouchers for restaurant and spa, having a clean and well-maintained swimming pool and free pickup and drop services.
- By using Pie Charts, we found that amenities such as Conference Rooms, Golf Courses, Limo Services, Mini bars and Indoor swimming pools are least used by the promoters and hence they should not be given more importance at the cost of other amenities. Thus, resources allocated to these amenities must be reduced and concentrated on other important ones.
- Decision tree showed us that the condition of the hotel must be maintained to a very high standard to obtain a better NPS score. The hotel should thus ensure that it looks elegant on the inside as well as the outside, provide a peaceful atmosphere, great ambience and must maintain the highest possible standard of cleanliness.
- Guest satisfaction and Customer services are the amenities that came forth through the Correlation matrix which have maximum impact with the Likelihood to Recommend. Thus, customers must be provided with clean and maintained rooms, laundry services, fresh toiletries, well trained and polite hotel staff and basic amenities such as drinking water, coffee, tea, etc.
- Using SVM we found that Quality of Internet Service can be given the minimum importance compared to other factors which influence the Likelihood to recommend and it should not be given more preference at the cost of other factors. Most of the customers may require just a basic internet connection which would be able to power their phones for messages and mails.