Hyatt Customer Data Analysis

By:

Apurva Sharma

Table of Contents

Introduction

Background

Hyatt Hotels Corporation is an United States based multinational owner, operator, and franchiser of hotels, resorts, and vacation properties and caters to a range of customers like business, leisure and family travelers. By 2018, Hyatt has 777 properties in 54 countries.
As a leading international hotel chain Hyatt is passionate about maintaining best customer experience across all their hotels. To achieve this, they actively collect customer feedback along with detailed information about customers stay and hotel information.

Objective

The objective of this project was to use data science techniques and unlock actionable insights from the collected data to help Hyatt better understand levers for improving overall customer experience.

Data Provided

Hyatt accumulates an extensive range of data for each reservation at their hotels. A version of such data obtained was given to us as part of our course-" IST 687 Introduction to Data Science". This data was from 56 nations all across the globe and was for the period February 2014 - January 2015.

The data comprised the following key categories of information:

- **Customer Aspects**: Contained customer demographics and profile information such as Age, Gender
- **Booking Factors Information**: Contained data such as check-in date, length of stay and Type of room
- **Hotel Details**: Hotel characteristics like the Hyatt Corporation brand, location, and available amenities
- **Customer feedback**: detailed customer feedback on different aspects of the service, along with customer's eventual Likelihood to Recommend rating and NPS_Type

Problem Statement

The customer feedback component, especially *Likelihood to Recommend* rating and *NPS_Type*, was the most essential component of the feedback score it was the key performance index that hotel chains aim to ultimately impact to via better customer service.

With *NPS_Type* and *LTR* being the key performance index, the project's problem statement was defined as identifying a list of factors that impact overall rating and hence will be the levers to improve the customer experience.

Data Acquisition/Import and Cleaning and Pre-processing

Selecting Columns

The data provided had 12 csv files in all; one for every month, overall making it approx. 20 GB full dataset of 12 individual months. The project started off by analyzing a sample file of the full data and identifying fill rates of each column provided. The full list of columns was filtered on basis of the following:

1. Columns having more than 90% missing values were dropped
2. Columns which didn't have any relevance to LTR or NPS_Type were dropped

Column Name	Definition
ROOM_TYPE_DESCRIPTION_C	Room type description (specific to the property) of the guest's room upon checkout
CHECK_IN_DATE_C	Check in date; for WALK status adjusted to the first in-house stay date
CHECK_OUT_DATE_C	Check out date
ADULT_NUM_C	Number of adults on the last day of the stay
CHILDREN_NUM_C	Number of children on the last day of the stay
POV_CODE_C	Purpose of visit
ENTRY_TIME_R	Time the reservation was made
NUM_ROOMS_R	Total number of rooms booked under the reservation
ADULT_NUM_R	Number of adults for which the reservation was made
CHILDREN_NUM_R	Number of children associated with the reservation
LENGTH_OF_STAY_R	Total length of stay (calculated as Departure Date minus Arrival Date) represented in days
Gender_H	Guest's gender
Age_Range_H	Guest's age range
Likelihood_Recommend_H	Likelihood to recommend metric; value on a 1 to 10 scale
Overall_Sat_H	Overall satisfaction metric; value on a 1 to 10 scale
Guest_Room_H	Guest room satisfaction metric; value on a 1 to 10 scale
Tranquility_H	Tranquility metric; value on a 1 to 10 scale
Condition_Hotel_H	Condition of hotel metric; value on a 1 to 10 scale
Customer_SVC_H	Quality of customer service metric; value on a 1 to 10 scale
Staff_Cared_H	Staff cared metric; value on a 1 to 10 scale
Internet_Sat_H	Internet satisfaction metric; value on a 1 to 10 scale
Check_In_H	Quality of the check in process metric; value on a 1 to 10 scale
City_PL	City in which the hotel is located
State_PL	State in which the hotel is located
Postal Code_PL	Zip code in which the hotel is located
Country_PL	Country in which the hotel is located
Property Latitude_PL	Latitude of the hotel's location
Property Longitude_PL	Longitude of the hotel's location
Brand_PL	Hotel's brand
Business Center_PL	Flag indicating if the hotel has a business center
Convention_PL	Flag indicating if the hotel has convention space
Golf_PL	Flag indicating if the hotel is near a golf space
Mini-Bar_PL	Flag indicating if the hotel has mini-bar
Pool-Outdoor_PL	Flag indicating if the hotel has an outdoor pool
Resort_PL	Flag indicating if the hotel is a resort
Shuttle Service_PL	Flag indicating if the hotel has shuttle service
Spa_PL	Flag indicating if the hotel has a spa
Valet Parking_PL	Flag indicating if the hotel has valet parking
Booking_Channel	Defined booking channel as per the NPS analysis
NPS_Type	Indicates if the guest's HySat responses mark them as a promoter, a passive, or a detractor

Assimilated full year data of above columns ensured patterns from all months taken into account. This created a dataset which was further analyzed for missing values. Since, our available machine could handle the selected relevant columns/variables from the 12 months individual data sets which could further be read into a single full_Test_dataset, no initial sampling was applied on the observations combined and hence, the data from full_Test_dataset was further used in our analysis. This not only ensured higher confidence on the entire analysis, but also allowed capturing all seasonal variations.

GETTING ACCUSTOMED TO OUR COMBINED DATA SET:

```
> summary(full_cleanData)
  POV_CODE_C        Age_Range_H          Gender_H         Likelihood_Recommend_H Overall_Sat_H     Guest_Room_H
 Length:924378      Length:924378      Length:924378      Min.   : 1.000         Min.   : 1.000   Min.   : 1.000
 Class :character   Class :character   Class :character   1st Qu.: 8.000         1st Qu.: 8.000   1st Qu.: 8.000
 Mode  :character   Mode  :character   Mode  :character   Median : 9.000         Median : 9.000   Median : 9.000
                                                          Mean   : 8.646         Mean   : 8.626   Mean   : 8.701
                                                          3rd Qu.:10.000         3rd Qu.:10.000   3rd Qu.:10.000
                                                          Max.   :10.000         Max.   :10.000   Max.   :10.000
                                                                                 NA's   :1117     NA's   :9593
 Tranquility_H    Condition_Hotel_H Customer_SVC_H   Staff_Cared_H    Internet_Sat_H     Check_In_H       Golf_PL
 Min.   : 1.0     Min.   : 1.000    Min.   : 1.000   Min.   : 1.0     Min.   : 1.0     Min.   : 1.0     Length:924378
 1st Qu.: 8.0     1st Qu.: 8.000    1st Qu.: 9.000   1st Qu.: 8.0     1st Qu.: 8.0     1st Qu.: 9.0     Class :character
 Median : 9.0     Median : 9.000    Median :10.000   Median :10.0     Median : 9.0     Median :10.0     Mode  :character
 Mean   : 8.7     Mean   : 8.903    Mean   : 9.008   Mean   : 8.9     Mean   : 8.5     Mean   : 9.2
 3rd Qu.:10.0     3rd Qu.:10.000    3rd Qu.:10.000   3rd Qu.:10.0     3rd Qu.:10.0     3rd Qu.:10.0
 Max.   :10.0     Max.   :10.000    Max.   :10.000   Max.   :10.0     Max.   :10.0     Max.   :10.0
 NA's   :400188   NA's   :11473     NA's   :16598    NA's   :397956   NA's   :548330   NA's   :397909
   City_PL            State_PL         PostalCode_PL      Country_PL        PropertyLatitude_PL PropertyLongitude_PL
 Length:924378      Length:924378      Min.   :    10   Length:924378      Min.   :-37.81      Min.   :-159.44
 Class :character   Class :character   1st Qu.: 28273   Class :character   1st Qu.: 30.32      1st Qu.: -98.49
 Mode  :character   Mode  :character   Median : 55344   Mode  :character   Median : 36.16      Median : -84.38
                                       Mean   : 60640                      Mean   : 33.92      Mean   : -65.51
                                       3rd Qu.: 81620                      3rd Qu.: 40.58      3rd Qu.: -74.64
                                       Max.   :7630293                     Max.   : 56.84      Max.   : 151.21
                                       NA's   :237397
   Brand_PL         BusinessCenter_PL  Convention_PL      LimoService_PL     PoolIndoor_PL      MiniBar_PL
 Length:924378      Length:924378      Length:924378      Length:924378      Length:924378      Length:924378
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

 PoolOutdoor_PL      Resort_PL         ShuttleService_PL     Spa_PL          ValetParking_PL     NPS_Type
 Length:924378      Length:924378      Length:924378      Length:924378      Length:924378      Length:924378
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

 CHECK_IN_DATE_C    CHECK_OUT_DATE_C   ENTRY_TIME_R       NUM_ROOMS_R    LENGTH_OF_STAY_R   ADULT_NUM_R    CHILDREN_NUM_R
 Length:924378      Length:924378      Length:924378      Min.   :1.000   Min.   :  1.000   Min.   :1.000   Min.   :1.0
 Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:  1.000   1st Qu.:1.000   1st Qu.:1.0
 Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :  2.000   Median :1.000   Median :2.0
                                                          Mean   :1.009   Mean   :  2.558   Mean   :1.519   Mean   :1.6
                                                          3rd Qu.:1.000   3rd Qu.:  3.000   3rd Qu.:2.000   3rd Qu.:2.0
                                                          Max.   :9.000   Max.   :180.000   Max.   :8.000   Max.   :6.0
                                                                          NA's   :58        NA's   :882     NA's   :833095
 Booking_Channel    ROOM_TYPE_DESCRIPTION_C
 Length:924378      Length:924378
 Class :character   Class :character
 Mode  :character   Mode  :character

> str(full_cleanData)
Classes 'data.table' and 'data.frame':  924378 obs. of  40 variables:
 $ POV_CODE_C              : chr  "BUSINESS" "BUSINESS" "BUSINESS" "LEISURE" ...
 $ Age_Range_H             : chr  "66-75" "46-55" "66-75" "66-75" ...
 $ Gender_H                : chr  "Female" "Male" "Female" "Male" ...
 $ Likelihood_Recommend_H  : int  10 10 10 10 10 9 10 10 10 10 ...
 $ Overall_Sat_H           : int  9 9 10 10 9 9 8 10 10 10 ...
 $ Guest_Room_H            : int  10 10 10 10 10 9 9 10 8 10 ...
 $ Tranquility_H           : int  10 9 10 NA 10 9 NA NA 10 10 ...
 $ Condition_Hotel_H       : int  9 10 10 9 10 9 10 10 9 10 ...
 $ Customer_SVC_H          : int  10 10 10 10 10 10 10 10 9 10 ...
 $ Staff_Cared_H           : int  10 10 10 NA 10 10 NA NA 10 10 ...
 $ Internet_Sat_H          : int  10 7 NA NA NA 6 NA NA 9 9 ...
 $ Check_In_H              : int  10 9 10 NA 9 10 NA NA 8 10 ...
 $ Golf_PL                 : chr  "N" "N" "N" "N" ...
 $ City_PL                 : chr  "Sharm El Sheikh" "Sharm El Sheikh" "Sharm El Sheikh" "Sharm El Sheikh" ...
 $ State_PL                : chr  NA NA NA NA ...
 $ PostalCode_PL           : int  NA NA NA NA NA NA NA NA NA NA ...
 $ Country_PL              : chr  "Egypt" "Egypt" "Egypt" "Egypt" ...
 $ PropertyLatitude_PL     : num  28 28 28 28 28 ...
 $ PropertyLongitude_PL    : num  34.4 34.4 34.4 34.4 34.4 ...
 $ Brand_PL                : chr  "Hyatt Regency" "Hyatt Regency" "Hyatt Regency" "Hyatt Regency" ...
 $ BusinessCenter_PL       : chr  "Y" "Y" "Y" "Y" ...
 $ Convention_PL           : chr  "Y" "Y" "Y" "Y" ...
 $ LimoService_PL          : chr  "Y" "Y" "Y" "Y" ...
 $ PoolIndoor_PL           : chr  "N" "N" "N" "N" ...
 $ MiniBar_PL              : chr  "Y" "Y" "Y" "Y" ...
 $ PoolOutdoor_PL          : chr  "Y" "Y" "Y" "Y" ...
 $ Resort_PL               : chr  "N" "N" "N" "N" ...
 $ ShuttleService_PL       : chr  "Y" "Y" "Y" "Y" ...
 $ Spa_PL                  : chr  "Y" "Y" "Y" "Y" ...
 $ ValetParking_PL         : chr  "N" "N" "N" "N" ...
 $ NPS_Type                : chr  "Promoter" "Promoter" "Promoter" "Promoter" ...
 $ CHECK_IN_DATE_C         : chr  "2014-02-10" "2014-02-09" "2014-02-28" "2014-02-11" ...
 $ CHECK_OUT_DATE_C        : chr  "2014-02-20" "2014-02-16" "2014-03-14" "2014-03-05" ...
 $ ENTRY_TIME_R            : chr  "04:06:06" "09:04:08" "02:59:52" "06:27:56" ...
 $ NUM_ROOMS_R             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ LENGTH_OF_STAY_R        : int  10 7 14 22 7 7 7 7 10 7 ...
 $ ADULT_NUM_R             : int  2 2 2 2 2 2 2 2 2 2 ...
 $ CHILDREN_NUM_R          : int  NA NA NA NA NA NA NA NA NA NA ...
 $ Booking_Channel         : chr  "Hotel" "Hotel" "Hotel" "Global Contact Center" ...
 $ ROOM_TYPE_DESCRIPTION_C: chr  "Sea View King" "Sea View King" "Seafront View King" "Regency Suite Twin" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

Handling Missing Values – I

Missing values (such as blanks, NA's) and other noisy data was handled on several stages of our analysis. Any global changes to the full data were not applied as this might have caused loss of crucial data.

The data had many cells having blank values or missing values, including *Likelihood_to_recommend* and *NPS_Type*. Following steps were taken to handling missing values:

1. A comprehensive identification of missing values. Although R marks most missing values as NA, it doesn't mark blank strings as NA. This causes some missing values to have non-NA values. All blank cells were converted to the value NA to ensure all missing values are marked as NA
2. Filtering of rows having blank value for *NPS_Type.* This was done as NPS_Type is the key parameter being studied. It being unavailable makes the record irrelevant to the study. This was also checked for *Likelihood_to_recommend and Country_PL*
3. All rows having missing countries were removed

```
# Handling Missing Values
# -----------------------
# Removing Blanks - [A good idea is to set all of the "" (blank cells) to NA before any further analysis.]
chunk[chunk==""] <- NA

chunk<-chunk[!(is.na(chunk$NPS_Type)),]
row.names(chunk)<-NULL
chunk<-chunk[!(is.na(chunk$Likelihood_Recommend_H)),]
row.names(chunk)<-NULL
chunk<-chunk[!(is.na(chunk$Country_PL)),]
row.names(chunk)<-NULL
```

After initial data cleaning, our full clean data set looked like this:

Data Profile

- Initial total columns in the given 12 months of datasets: 237
- Columns retained: 40
- Rows read: 15,711,552 obs. of 40 variables
- **Final Clean Data: 924378 obs. of 40 variables**

Assumptions:

- Raising customer satisfaction on services and the quality of stay may help convert detractors into passives and eventually into promoters.
- Detractors take away Hyatt's future business, so the scope of the project would be decided by parts with less NPS /more number of detractors.

Relationship between Key Performance Indices(KPI's)

Likelihood_to_recommend vs. NPS_Type

Given we had two versions of customer feedback score: *NPS_Type* and *Likelihood_to_recommend,* the relationship between the two was first carefully examined Below table shows number of observations for each combination of the values:

		Likelihood to Recommend-----→									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
NPS_Type	Detractor	**11628**	**11379**	**13891**	**13147**	**30404**	**28341**	0	0	0	0
	Passive	0	0	0	0	0	0	**56956**	**121365**	0	0
	Promoter	0	0	0	0	0	0	0	0	**201018**	**436249**

Based on this analysis, it was established that Likelihood_to_recommend and NPS_Type are directly related.

Net Promoter Score

Hence, Net Promoter Score is a key performance metric which is defined as:

NPS = Percentage of Promoters – Percentage of Detractors

Based on above it can be said that all three components of the equation are directly related to each other. Focusing on any one of the component will allow us to study and impact all three.

Business Questions:

All Business questions(BQ) were formulated keeping Hyatt Corporations best interests in mind. The feedback provided by Hyatt's customer consisted of various parameters such as likelihood to recommend, overall satisfaction, tranquility, customer service, overall experience, guestroom, etc. Analyzing such data based on a visitor's feedback is essential to Hyatt to not only help improve their facilities but also recognize any fallacies before they become a major issue.

In this project the problem statement has been broken down into bite sized pieces and worked on each exclusively to get results. Use of various descriptive statistics, modelling techniques and visualization techniques has been done to address the following the business questions: These questions will lead to generation of trends and patterns which lead to tangible action Points for Hyatt CEO.

- **BQ 1.** How is Hyatt doing overall in general? In the Hyatt dataset provided, after cleaning, what is the count or spread of customers worldwide with respect to different Purpose of Visit and the three NPS types?
- **BQ 2.** What is the hotel details/ amenities effect on NPS score? Which of these factors related to service quality across different purpose of visits (Business and Leisure) are important in determining NPS score? How to improve NPS Score using these factors?
- **BQ 3.** What is the comparison between Hyatt Hotels sister brands and their effect on NPS?
- **BQ 4.** What is the relationship of various booking factors on NPS? How can it be improved?
- **BQ 5.** How are customer aspects/guest demographics analysis related to NPS & how do they impact on customer's likelihood to recommend? How can NPS Score be improved with these?

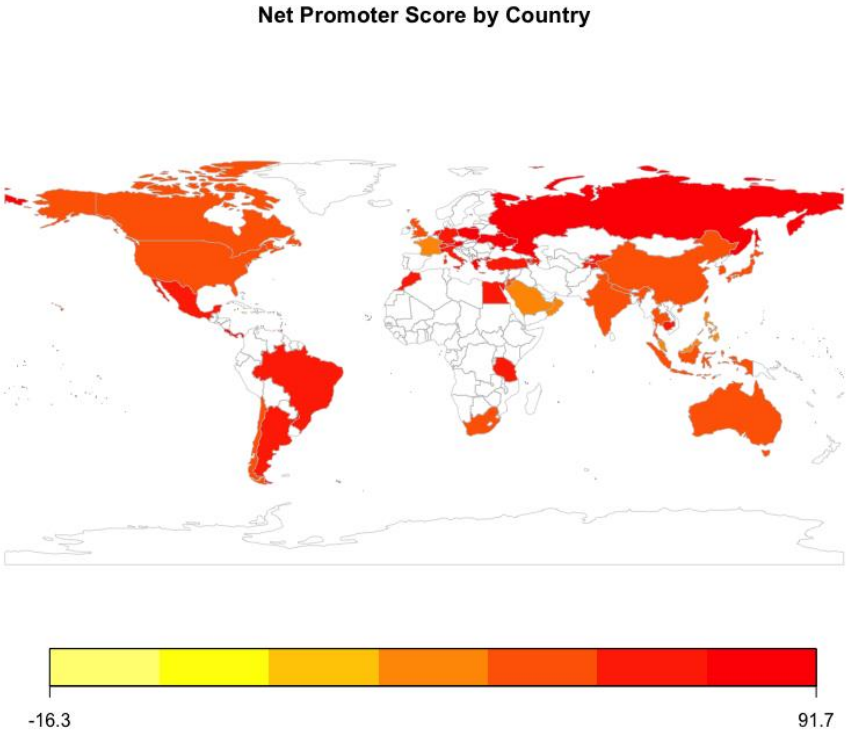
Data Exploration and Insights

Approach

Data across all geographies and brands was first analyzed in depth, and then it was iteratively sliced into subparts that had the highest data volume at each level. Hypothesis was that focusing on parts having most data will allow us to generate analysis based on wider customer behavior patterns. In United States, California had the largest dataset, because California also has the largest population. Similarly, Hyatt Regency was chosen as the next dataset as it was found to have the highest observations. Details of this analysis and visuals are presented further in the report.

The World

At start, analysis of NPS by country was performed in order to understand how NPS varies by country and geography.




```
#Join the promoters and detractors using country
PromDet_byCountry<-join(PromoterTotalbyCountry,DetractorTotalbyCountry,by="Country_name")
colnames(PromDet_byCountry)<- c("Country_name","Promoter_Freq","Detractor_Freq")

#Join TotalbyCountry and PromDet_byCountry
FullNPSData_byCountry<-join(TotalbyCountry,PromDet_byCountry,by="Country_name")
FullNPSData_byCountry[is.na(FullNPSData_byCountry)]<-0
FullNPSData_byCountry$NPS_Score <- (FullNPSData_byCountry$Promoter_Freq - FullNPSData_byCountry$Detractor_Freq)/FullNPSData_byCountry$TotalFreq *100

# Merge the two data frames(LTR and NPS dataframes)
TotalNPSdata<- merge(LTR_df,FullNPSData_byCountry,by.x = 'Country_Name',by.y  = 'Country_name')
TotalNPSdata$NPCformula<- npc(round(TotalNPSdata$Country_LTR), breaks = list(0:6, 7:8, 9:10)) #NPC formula
View(TotalNPSdata)
# - Total of 56 countries were found to have their NPS score more than their Goal Values(Goal NPS=53)
# - The average NPS throughout all countries is 57.7
```

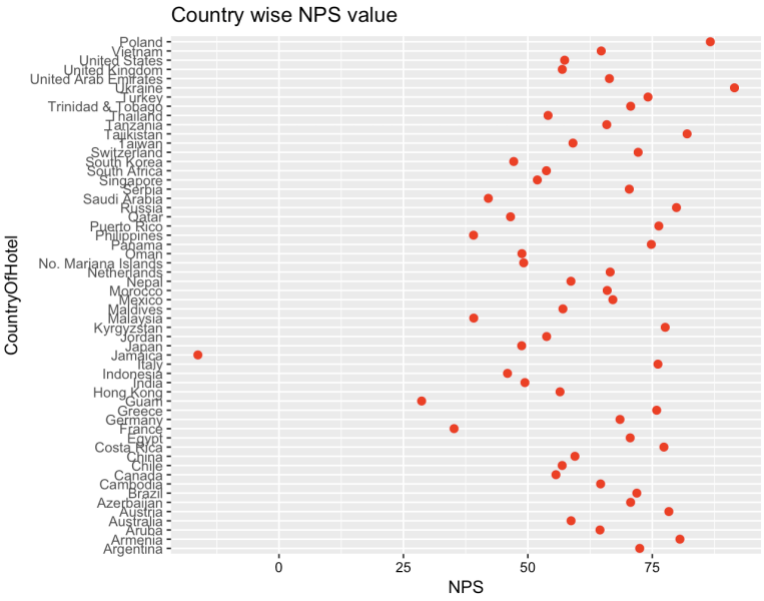
Key Observations and Insights:

- Average NPS across all countries has been found to be 60.69% (average LTR 8.76)
- Poland is having the highest NPS 91% (average LTR of 9.64)
- Jamaica having the lowest NPS -16% (average LTR of 6.22)

BQ 1. What is the count & spread of Hyatt hotel customers worldwide?

Detailed Data on Top and Bottom four countries in terms of NPS:

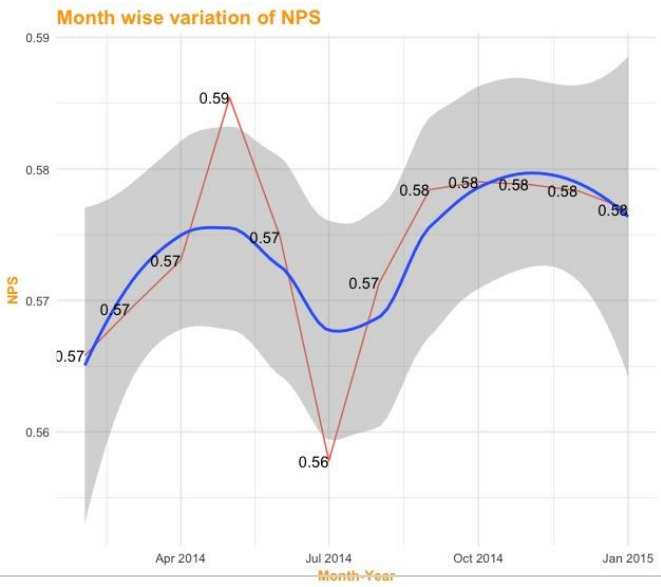
Country Name	LTR	Population	NPS
Poland	9.65	48	91.67
Ukraine	9.48	378	88.10
Russia	9.36	1887	80.71
Greece	9.29	369	79.13
…	…	…	…
United States	8.65	744959	57.39
…	…	…	…
Philippines	8.20	1239	38.01
France	7.99	10241	34.15
Guam	7.90	1503	27.28
Jamaica	6.23	92	-16.30

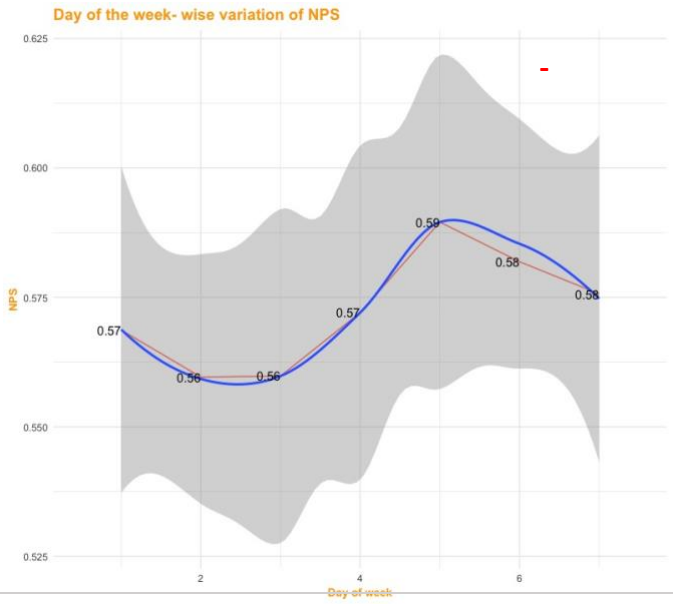


Action Points for Hyatt: for Hyatt:

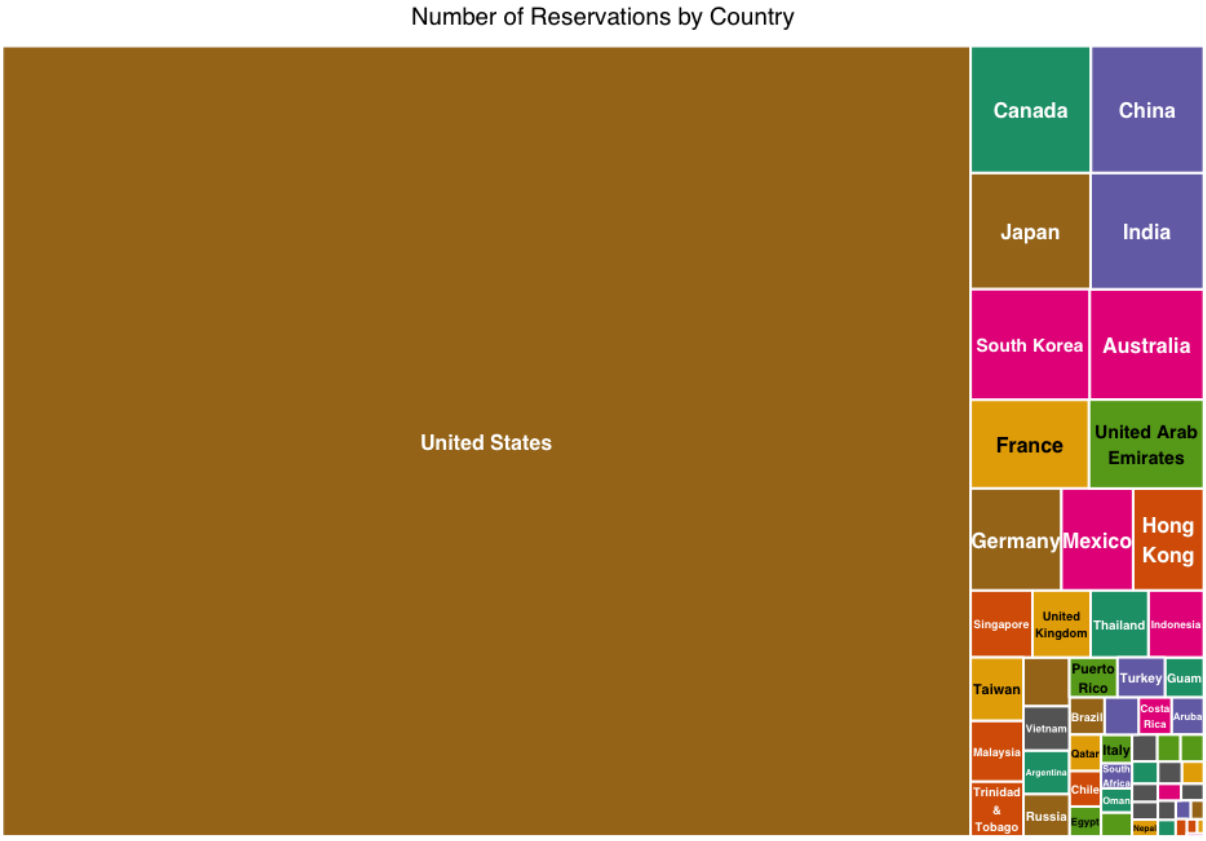
- There is scope for improvement in general for Hyatt Corporation
- In United States, Hyatt must focus on Jamaica to study what is causing the low NPS Score in Jamaica
- Jamaica is a popular tourist destination and improving NPS might lead to significant impact on the overall NPS

Month wise variation and Day of week wise variation in NPS Score was also performed Looks like responses provided in first quarter of the year had low NPS scores. Also, Fridays have a higher score as compared to other days of the week. However, doing this analysis did not generate much insights as the variation of scores was pretty less. Here are the visuals of analysis:





To study the factors that might be causing the difference in NPS values, the project focused on analyzing a country out of the total world data which had most amount of data. Having done this, provided us with enough data to play with and generate trends and patterns. The below chart shows the number of reservations available for each country.




```
# Population Data and plot
CountryTreeMap<- treemap(TotalNPSdata,index = c("Country_Name"),vSize="TotalFreq",type="index",
                         palette = "Dark2",title = "Number of Reservations by Country",fontsize.title = 14,
                         fontsize.labels = 12,border.col = "white")
```

Key Observations

- United States has the highest number of reservations which accounted for 81% percentage of total reservations. NPS score of United States is: 57.39
- Observations in many other countries are significantly less

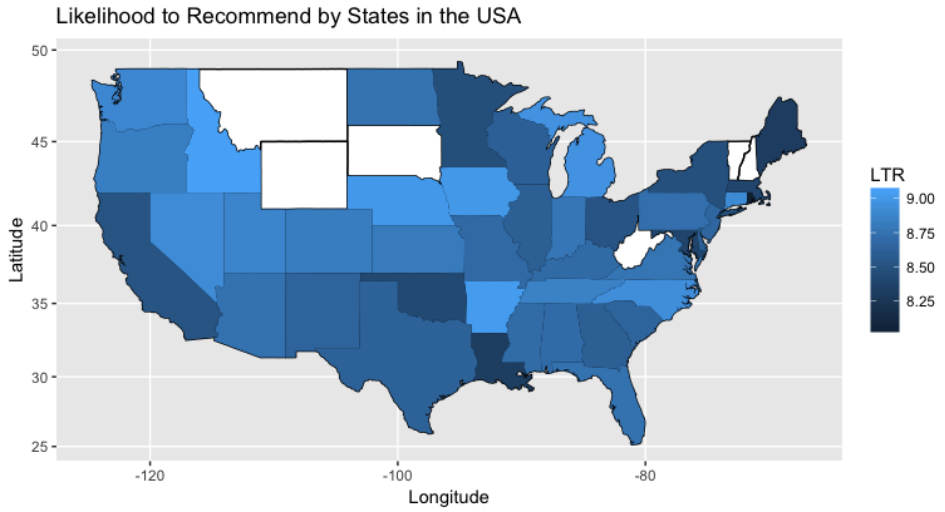
Action Points for Hyatt:

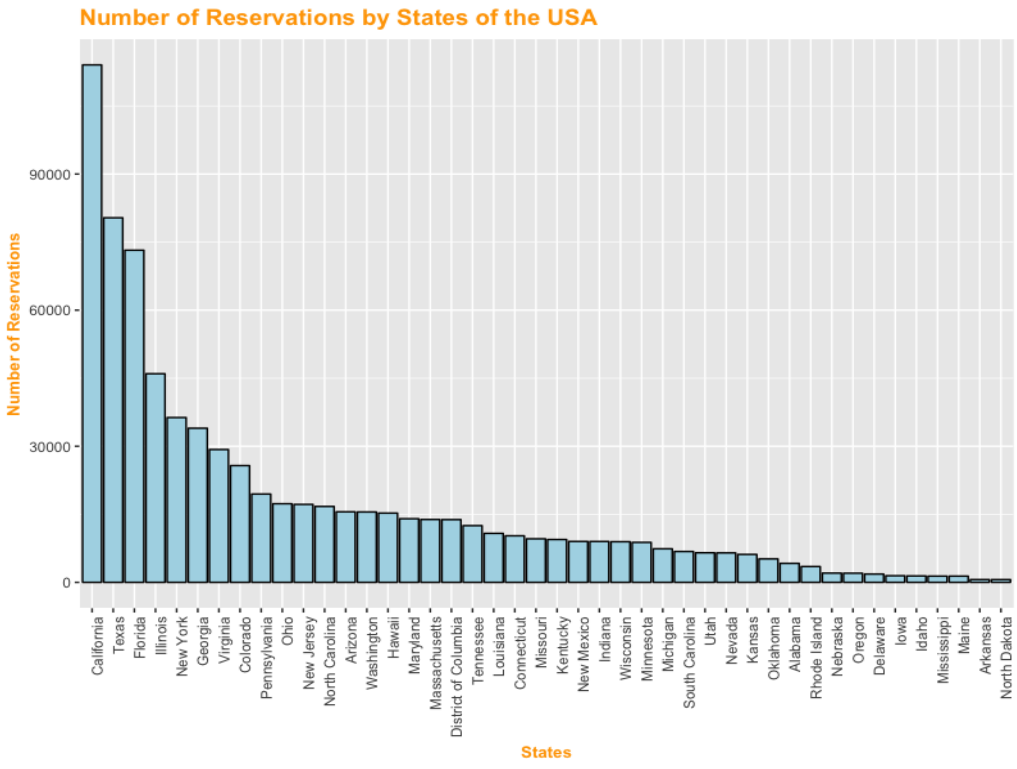
- United States brings in most business to Hyatt group and there is scope for improving performance as its NPS score of 57% is lower than average NPS score of 61%
- Further, Hyatt has a good opportunity to analyze expanding its presence in countries outside the United States, as currently it has very low presence outside United States including countries which are bigger and more populous than USA

United States had 744,959 reservations from the selected columns in the '12 months data' provided, making it a rich resource to analyze customer preferences. This could be due the fact that United States has the third highest population the whole world and Hyatt's customers are majorly from United States. Hence, this project has focused on United States data for most of the further analysis

Country: USA

After getting accustomed with our United States of America data, a further deep dive into NPS by State was taken to understand how NPS varies by state. Below plot shows average Likelihood_to_recommend, which is directly related to NPS, for each state in the USA. Further, analysis on the number of reservations by each state was also performed.





```
# plot LTR
us_map <- map_data("state")
simplemap <- ggplot()
simplemap <- simplemap +  geom_map(data=us_map,aes(map_id=region),map=us_map,fill="white",  color="black")
simplemap <- simplemap+ expand_limits(x= us_map$long, y=us_map$lat)+ coord_map()
States_LTR_df$States_LTR<- (as.numeric(as.character(States_LTR_df$LTR)))
State_NPSmap  <- simplemap +
  geom_map(data=States_LTR_df,map=us_map,aes( map_id=State_name_lower,fill=LTR),na.rm=TRUE,show.legend=TRUE)
State_NPSmap <- State_NPSmap + labs(x = "", y= "")+
  ggtitle("Likelihood to Recommend by States in USA")
State_NPSmap

# plot Population
ggplot(States_LTR_df,aes(x = reorder(State_name,-freq), y = freq )) +
  geom_bar(stat ="identity",col="black",fill="light blue") + theme(axis.text.x =element_text(angle = 90,hjust = 1))+
  ggtitle("Number of Reservations by States of the USA")+
  theme(plot.title = element_text(color="orange", face="bold", size=14, hjust=0)) +
  theme(axis.title = element_text(color="orange", face="bold", size=10))+
  labs(x="States",y="Number of Reservations")
```

Key Observations

1. Idaho state had the highest LTR of 9.05
2. Rhode Island state had the lower LTR of 8.01
3. Bigger states of California and Texas have highest reservation count of 114,063

Action Points for Hyatt:

1. For fastest gains, Hyatt must of focus on states of Colorado, Virginia and Florida as their average LTR is very close to 9 and they are amongst the top most contributors by reservation counts

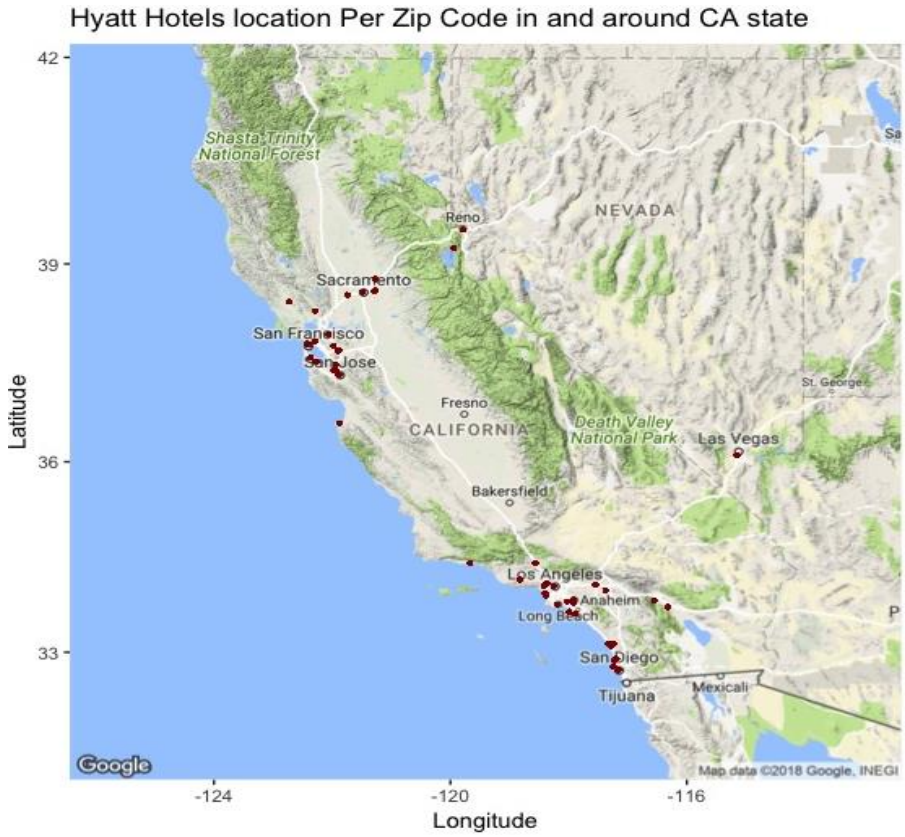
State_name	LTR	Reservation Count
Idaho	9.05	1412
Arkansas	9.03	611
Nebraska	9.00	2035
Iowa	9.00	1457
Michigan	8.97	7397
North Carolina	8.93	16734
Nevada	8.92	6491
Connecticut	8.91	10258
Washington	8.89	15512
Utah	8.88	6515
Tennessee	8.85	12502
Oregon	8.83	2014
Colorado	8.82	25735
Kansas	8.81	6157
Indiana	8.77	9011
Virginia	8.76	29277
North Dakota	8.75	606
Florida	8.74	73210
Arizona	8.74	15543
Pennsylvania	8.73	19464
Hawaii	8.71	15261
Mississippi	8.70	1368

State_name	LTR	Reservation Count
Alabama	8.70	4187
Kentucky	8.70	9446
Missouri	8.69	9602
New Jersey	8.66	17180
South Carolina	8.65	6802
New Mexico	8.64	9017
Texas	8.63	80361
Georgia	8.61	33974
Wisconsin	8.61	8956
Illinois	8.61	45980
Delaware	8.60	1830
District of Columbia	8.56	13829
Ohio	8.52	17329
California	8.52	114063
New York	8.49	36325
Minnesota	8.48	8808
Oklahoma	8.42	5169
Massachusetts	8.39	13859
Maryland	8.38	14021
Louisiana	8.32	10797
Maine	8.31	1366
Rhode Island	8.01	3488

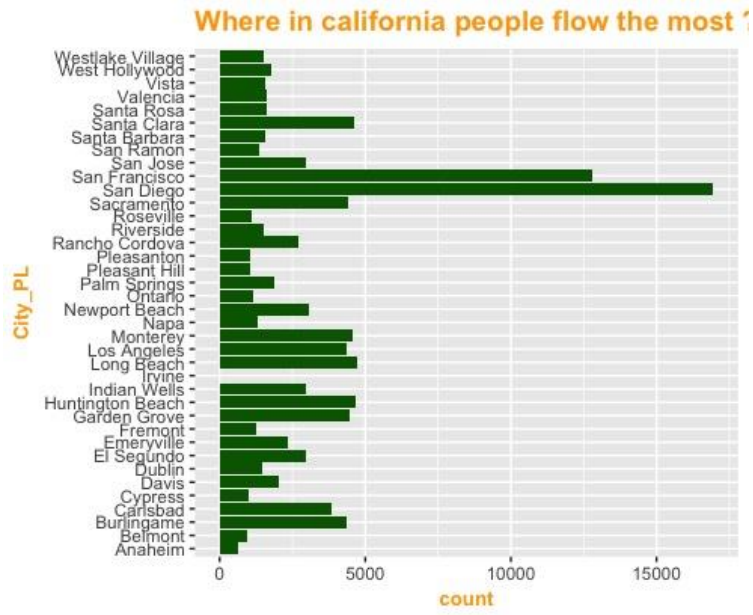
Moving forward this project has focused on the State of 'California' as it provided the most population for doing data analysis. Working on higher population data ensured a robust analysis. Further working on California allowed to focus on a small region and ensured capturing of regional phenomena well. The idea was to further try the patterns learnt in California on other states to analyze if other states required a different analysis.

```
# United States Analysis
# ----------------------
#Data preparation
usdata<-full_cleanData[full_cleanData$Country_PL=="United States",] #744959 obs. of  40 variables
row.names(usdata) <- NULL #to reindex
#percentage of US
dim(usdata)[1]/dim(full_cleanData)[1]*100
#[1] 80.5903
```

State: California



Extensive descriptive analytics was performed to gain in-depth understanding of US data and patters generation. This approach was best suited to analyze data and figure out hypothesis and business questions generation. Analysis of reservation counts for each brand within California showed that Hyatt Regency seems to have the most data and maximum no. of reservations of each type, making it the ideal brand to focus on

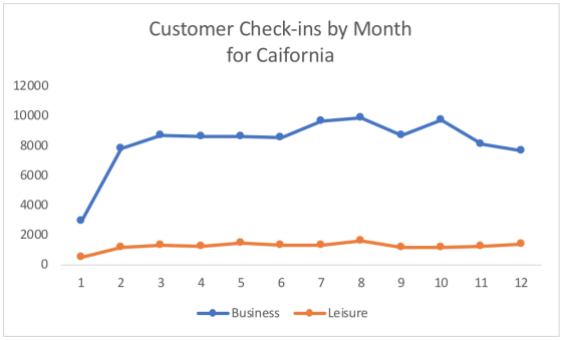


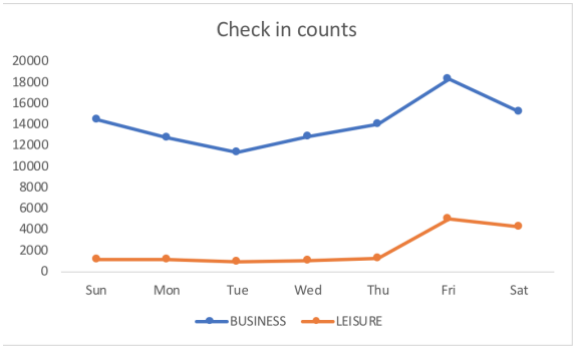
Insights:

- It was observed that weekenders and leisure seekers visit San Diego and San Francisco the most. It was also observed later through SVM modelling that having a Resort increases NPS. (Refer SVM modelling section)

Action Points for Hyatt:

- Having resorts in neighboring suburbs must be explored for catering better to requirements of these as having resort raises NPS for Leisure. This was proved further through association rules





Insight:

- Customers must be incentivized for January as this month of the year had the lowest NPS scores. (Please refer the graph below)
- Most Leisure people check in during the weekend

Action Points for Hyatt: for Hyatt:

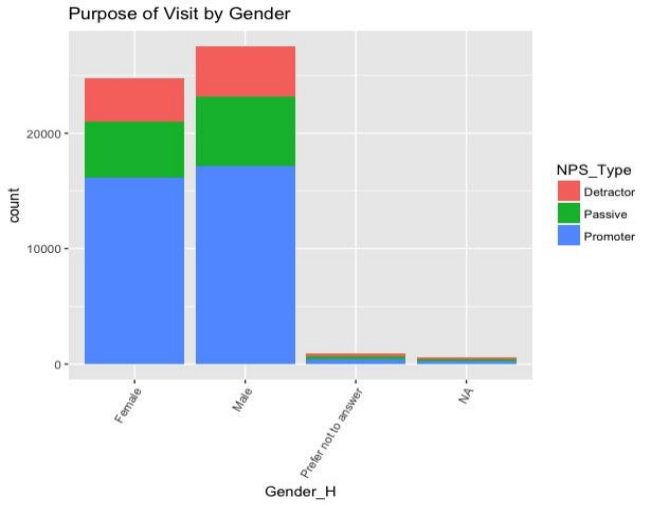
- Hyatt should reach out to such customers for more feedback
- Hyatt should also push them to higher Hyatt brands for more features they might be looking for.
- Weekend activities must be arranged as it turns out that Fridays have the maximum check-ins into the Hyatt Regency hotels of California.
- 50% of business check-ins on Friday have 2-day stay making them weekenders

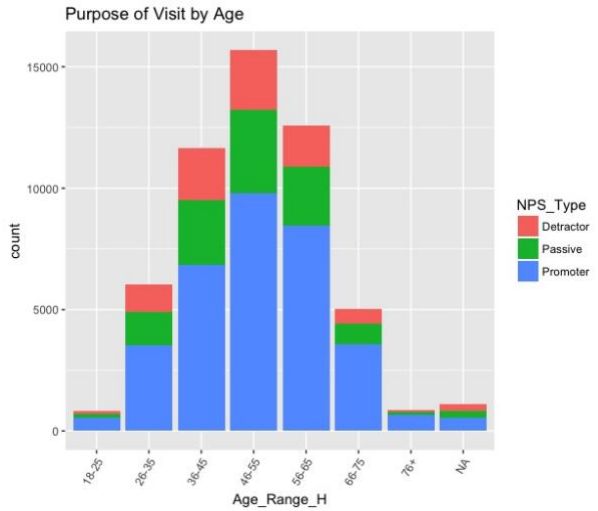
BQ 5. How are customer aspects/guest demographics analysis related to NPS & how do they impact on customer's likelihood to recommend?

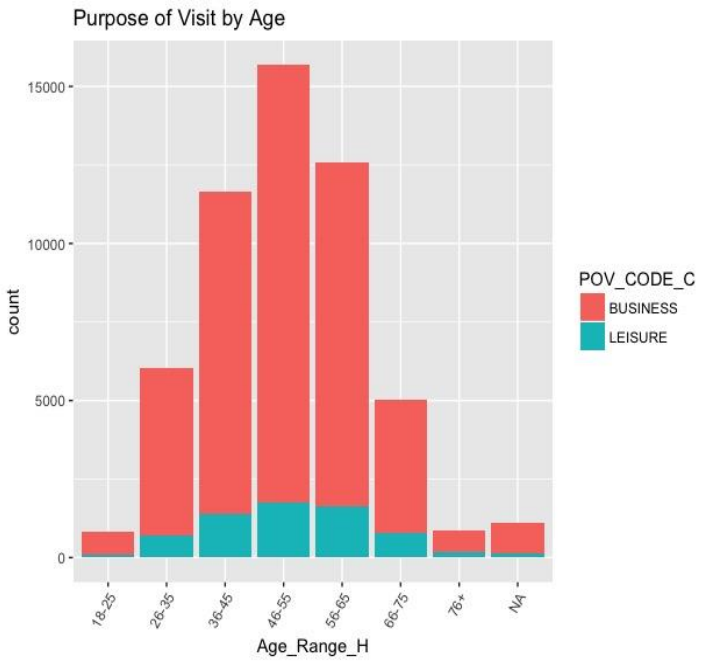
POV (Purpose of Visit) analysis vs Age, Gender and other customer demographics was achieved to figure out Hyatt attracts which section of age the most.

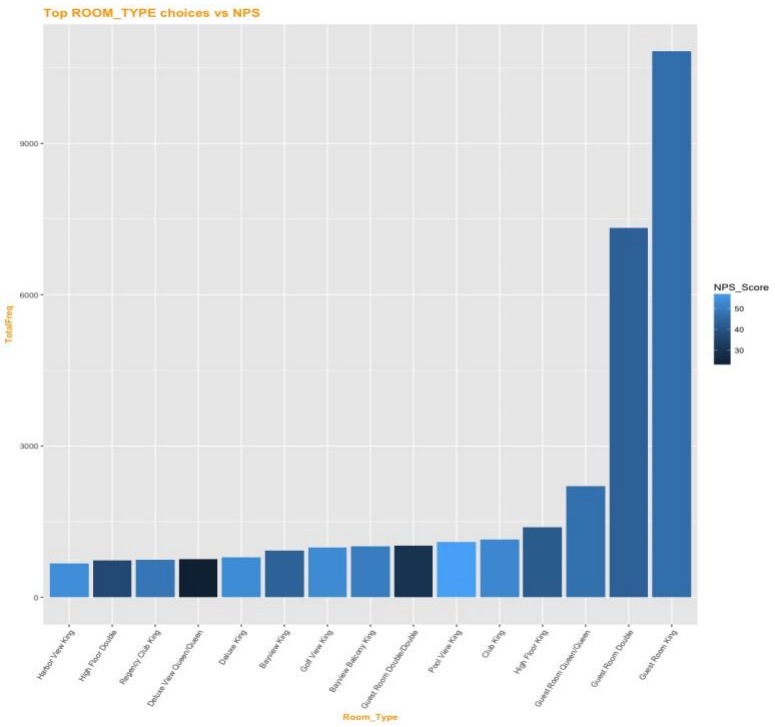
Insight:

- We found out that a typical customer for Hyatt is between the ages 26-75, with maximum customer in the age group of 46-75. This included both Business and Leisure visits but maximum travelers to Hyatt Regency brand in California have Business as Purpose of their visit.
- The NPS score varied accordingly but we couldn't find much insights as to why this was happening through descriptive analytics. Hence, we moved to various modelling techniques which are explained further in this report.
- Another trend that was noticed was the Room choice for maximum travelers is Guest Room King and Guest Room Double. The highest NPS contributing King room types were found to be- Pool View King and Harbor View King and Deluxe King. Hence, Hyatt must focus more on such rooms.



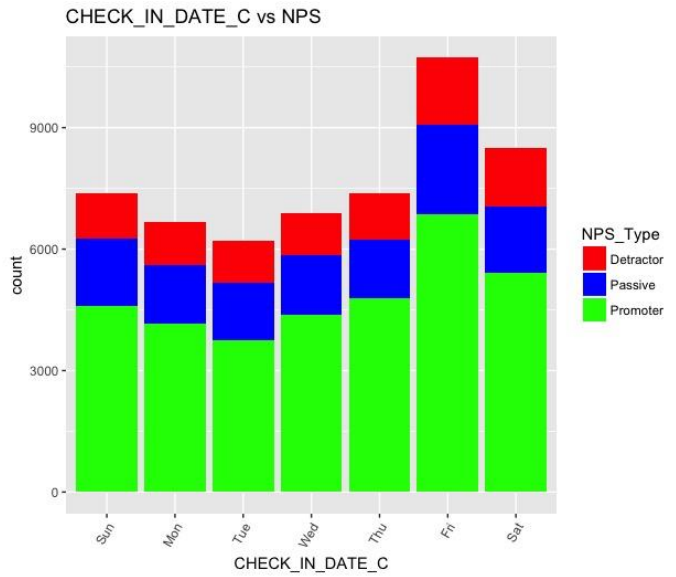


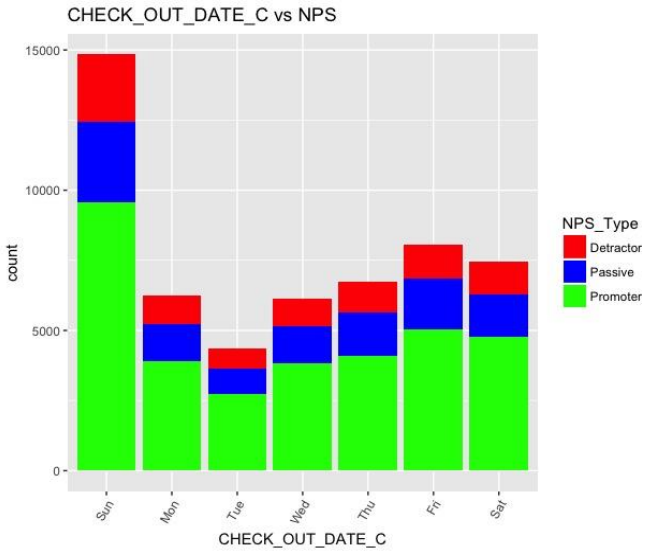


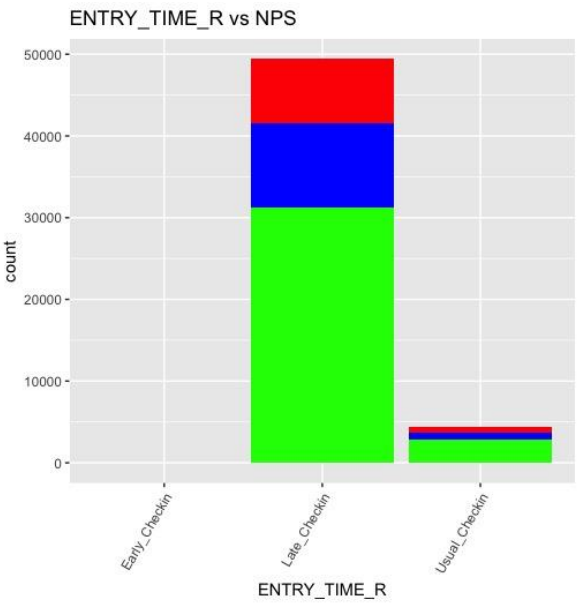


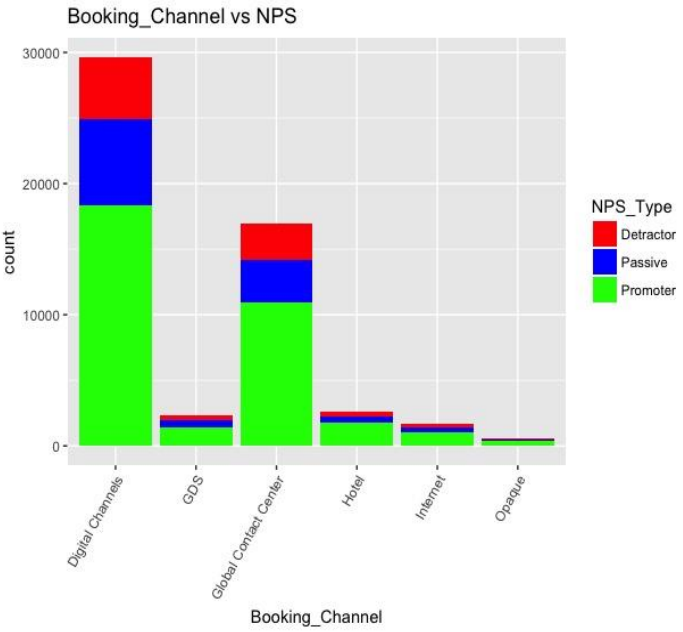
BQ 4. What is the relationship of various booking factors on NPS? What are the action Points ?

Booking factors such as Booking channel, Room Type selection preferences, Length of stay vs NPS Type, Check-in Day and Check-out Day and check-in entry timings, all against NPS types were plotted as below:







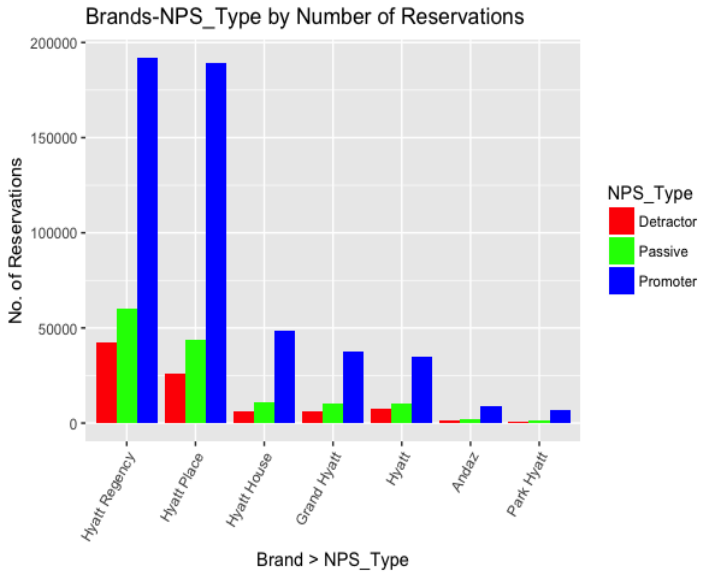


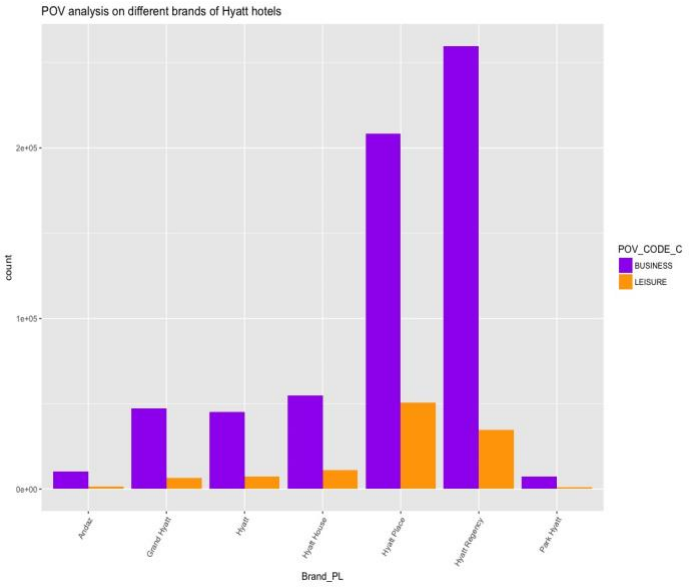
Insights and Action Points for Hyatt:

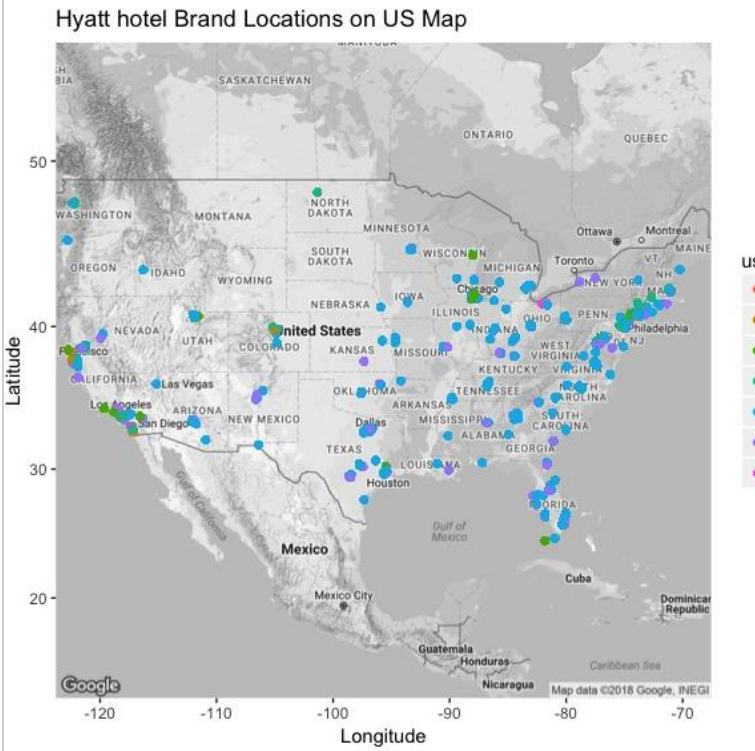
- Most check-ins happened on Fridays and most checkouts happened on Sunday.
- Most people checked in late at
- Digital Channels was the most preferred choice for making the bookings at Hyatt Regency. This was closely followed with Global Contact Center.

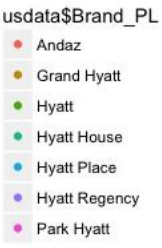
BQ 3. What is the comparison between Hyatt Hotels sister brands and their effect on NPS?

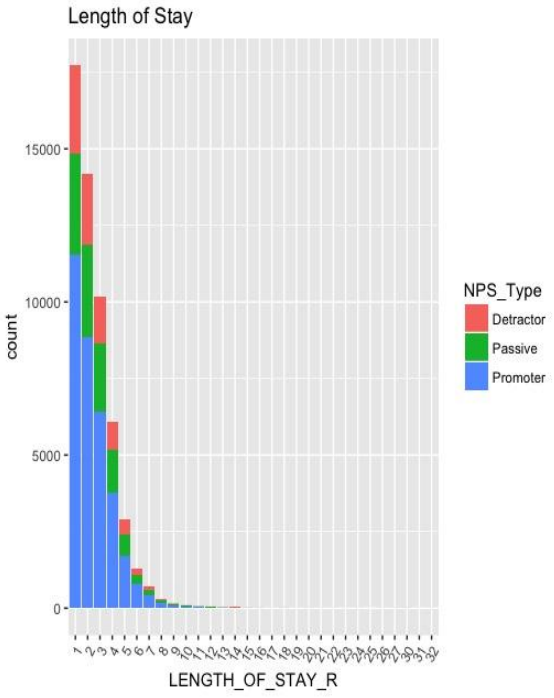
Insights: Brand Analysis showed that Regency which the most customers is performing the worst and Hyatt Park is performing the best. Hence Hyatt must consider understanding the difference between the two to improve performance. Maximum travelers to regency were on a business trip











A **US terrain map** was further plotted to analyze answers to questions such as “which Hyatt hotel brand do different customer prefers?” and plotted other descriptive analytics graphs to answer “What is the count of customers who visited hotels (hotel brand wise)?

As Hyatt Regency was found to have the highest number of detractors, further analysis is all based on Hyatt Regency.

Data Modelling for Advanced Insights and Validation

```
#NA HANDLING :(Removing unneeded columns and super biased columns)
#Determine if columns have any NA's
#summary(Cal_Regency_modelling)
#replacing NA's with mean values
Cal_Regency_modelling$Guest_Room_H[is.na(Cal_Regency_modelling$Guest_Room_H)] <- round(mean(Cal_Regency_modelling$Guest_Room_H, na.rm = TRUE))
Cal_Regency_modelling$Overall_Sat_H[is.na(Cal_Regency_modelling$Overall_Sat_H)] <- round(mean(Cal_Regency_modelling$Overall_Sat_H, na.rm = TRUE))
Cal_Regency_modelling$Tranquility_H[is.na(Cal_Regency_modelling$Tranquility_H)] <- round(mean(Cal_Regency_modelling$Tranquility_H, na.rm = TRUE))
Cal_Regency_modelling$Condition_Hotel_H[is.na(Cal_Regency_modelling$Condition_Hotel_H)] <- round(mean(Cal_Regency_modelling$Condition_Hotel_H, na.rm = TRUE))
Cal_Regency_modelling$Customer_SVC_H[is.na(Cal_Regency_modelling$Customer_SVC_H)] <- round(mean(Cal_Regency_modelling$Customer_SVC_H, na.rm = TRUE))
Cal_Regency_modelling$Staff_Cared_H[is.na(Cal_Regency_modelling$Staff_Cared_H)] <- round(mean(Cal_Regency_modelling$Staff_Cared_H, na.rm = TRUE))
Cal_Regency_modelling$Internet_Sat_H[is.na(Cal_Regency_modelling$Internet_Sat_H)] <- round(mean(Cal_Regency_modelling$Internet_Sat_H, na.rm = TRUE))
Cal_Regency_modelling$Check_In_H[is.na(Cal_Regency_modelling$Check_In_H)] <- round(mean(Cal_Regency_modelling$Check_In_H, na.rm = TRUE))
Cal_Regency_modelling$LENGTH_OF_STAY_R[is.na(Cal_Regency_modelling$LENGTH_OF_STAY_R)] <- round(mean(Cal_Regency_modelling$LENGTH_OF_STAY_R, na.rm = TRUE))
Cal_Regency_modelling$ADULT_NUM_R[is.na(Cal_Regency_modelling$ADULT_NUM_R)] <- round(mean(Cal_Regency_modelling$ADULT_NUM_R, na.rm = TRUE))
Cal_Regency_modelling$CHILDREN_NUM_R[is.na(Cal_Regency_modelling$CHILDREN_NUM_R)] <- round(mean(Cal_Regency_modelling$CHILDREN_NUM_R, na.rm = TRUE))

#Linear Modelling Data Preparation:#as it requires numeric variables . Saving this before converting to factor.
California_Regency_Cal_Regency_LMData<- Cal_Regency_modelling #Linear modelling main modelling data !
#summary(California_Regency_Cal_Regency_LMData)

# Converting all numerical columns into " High", " Medium"," Low"
str(Cal_Regency_modelling)
```

Data Preparation for Modelling

We applied a series of data processing steps to make our data ready for analysis using models

1. Removed redundant/biased columns: These were columns which had the same value in the full dataset. The columns included: Brand_PL, State_PL, Country_PL
2. All the rows with empty NPS_Type were removed in the process of analysis
3. Removed columns containing geographic details: These were columns latitude, longitude and zip codes as there were no clear state level signals that could be derived from these variables
4. Created categorical columns from numerical feedback columns
5. A bunch of variables were transformed so that better interpretation can be made from them
    a. Check-in time
    b. Check-in date
    c. Check out date
    d. Numeric to Factor

Handling Missing Values – II

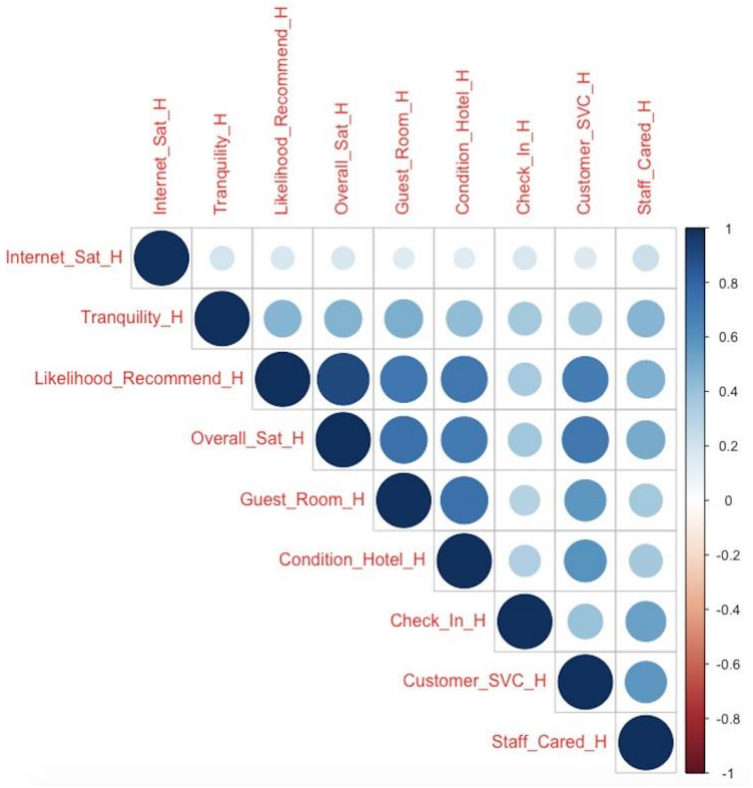
We again handled missing values and NA's at this step.

1. For numeric variables all missing values were replaced by their mean values
2. For factor variables NAs were added as a level of the factor

To further understand the relationship of these key drivers and others with individual Purpose of Visit and Booking factors, we performed Association Rule Mining individually for Business & Leisure customers

Identifying Correlated Variables

Common variables are key problem and often cause issues when models are implemented. Having studied correlation of all numeric variables upfront, we removed the variable Overall_Sat_H (Overall Satisfaction) from further analysis based on correlation analysis and Business intuition.



Correlation Table result:

	Likelihood_Recommend_H	Guest_Room_H	Overall_Sat_H	Tranquility_H	Condition_Hotel_H	Customer_SVC_H	Staff_Cared_H	Internet_Sat_H	Check_In_H
Likelihood_Recommend_H	1.00	0.73	0.90	0.45	0.71	0.70	0.48	0.18	0.36
Guest_Room_H	0.73	1.00	0.74	0.49	0.75	0.57	0.36	0.14	0.30
Overall_Sat_H	0.90	0.74	1.00	0.47	0.71	0.72	0.49	0.17	0.37
Tranquility_H	0.45	0.49	0.47	1.00	0.43	0.36	0.45	0.19	0.36
Condition_Hotel_H	0.71	0.75	0.71	0.43	1.00	0.59	0.36	0.14	0.33
Customer_SVC_H	0.70	0.57	0.72	0.36	0.59	1.00	0.58	0.15	0.41
Staff_Cared_H	0.48	0.36	0.49	0.45	0.36	0.58	1.00	0.22	0.54
Internet_Sat_H	0.18	0.14	0.17	0.19	0.14	0.15	0.22	1.00	0.18
Check_In_H	0.36	0.30	0.37	0.36	0.33	0.41	0.54	0.18	1.00

LINEAR REGRESSION MODELLING (LM)

Linear Modelling Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. (Source: https://www.statisticssolutions.com/what-is-linear-regression/)

Modelling parameters

- We used Likelihood_Recommend_H (LTR) for Linear Modelling analysis
- We based our predictions using R-sq. for Linear Modelling
- The idea of getting a list of key drivers was: multiple combinations of factors, wherein the one that provides highest R-sq. value is a list of key drivers; in LM we excluded factors with low coefficient (high p-value)
- We modelled our results against Likelihood_Recommend_H which was our dependent variable.

Individual R Squared Values(R^2)

```
model1 <- lm(formula=Likelihood_Recommend_H ~ Guest_Room_H, data=Cal_Regency_LM)
summary(model1) #R square= 0.5286 #important

model2 <- lm(formula=Likelihood_Recommend_H ~ Overall_Sat_H, data=Cal_Regency_LM)
summary(model2) #R square= 0.8169

model3 <- lm(formula=Likelihood_Recommend_H ~ Tranquility_H, data=Cal_Regency_LM)
summary(model3) #R square= 0.2068

model4 <- lm(formula=Likelihood_Recommend_H ~ Internet_Sat_H, data=Cal_Regency_LM)
summary(model4) #R square= 0.03152 #not important

model5 <- lm(formula=Likelihood_Recommend_H ~ Condition_Hotel_H, data=Cal_Regency_LM)
summary(model5) #R square= 0.5066 #important !

model6 <- lm(formula=Likelihood_Recommend_H ~ Customer_SVC_H, data=Cal_Regency_LM)
summary(model6) #R square= 0.4838 #important !

model7 <- lm(formula=Likelihood_Recommend_H ~ Staff_Cared_H, data=Cal_Regency_LM)
summary(model7) #R square= 0.2303

model8 <- lm(formula=Likelihood_Recommend_H ~ Check_In_H, data=Cal_Regency_LM)
summary(model8) #R square= 0.1274

model9 <- lm(formula=Likelihood_Recommend_H ~ LimoService_PL, data=Cal_Regency_LM)
summary(model9) #R square= 0.002634

model10 <- lm(formula=Likelihood_Recommend_H ~ MiniBar_PL, data=Cal_Regency_LM)
summary(model10) #R square= 0.002735

model11 <- lm(formula=Likelihood_Recommend_H ~ BusinessCenter_PL, data=Cal_Regency_LM)
summary(model11) #R square= 0.001348
```

Overall LM Result:

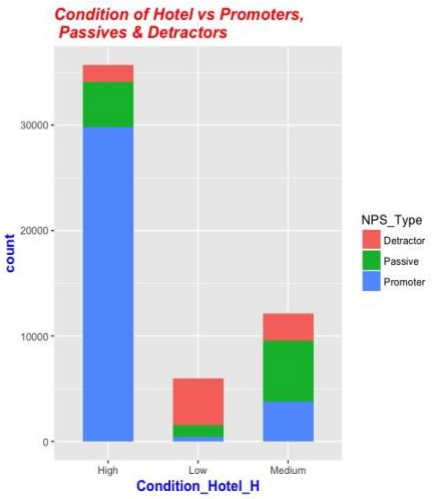
```
Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + Internet_Sat_H +
    Check_In_H + BusinessCenter_PL + Convention_PL + MiniBar_PL +
    PoolIndoor_PL + PoolOutdoor_PL + LimoService_PL + Resort_PL +
    ShuttleService_PL + Spa_PL + ValetParking_PL, data = Cal_Regency_LM)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9984 -0.2162  0.1903  0.5536  9.5124

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.632118   0.110016 -23.925  < 2e-16 ***
Guest_Room_H       0.343263   0.004392  78.161  < 2e-16 ***
Tranquility_H      0.067596   0.004265  15.848  < 2e-16 ***
Condition_Hotel_H  0.295488   0.004859  60.807  < 2e-16 ***
Customer_SVC_H     0.401817   0.004662  86.188  < 2e-16 ***
Staff_Cared_H      0.125172   0.005799  21.584  < 2e-16 ***
Internet_Sat_H     0.031921   0.003815   8.367  < 2e-16 ***
Check_In_H        -0.008240   0.005483  -1.503 0.132898
BusinessCenter_PLY 0.148342   0.024640   6.020 1.75e-09 ***
Convention_PLY    -0.068809   0.046565  -1.478 0.139498
MiniBar_PLY       -0.068560   0.038531  -1.779 0.075188 .
PoolIndoor_PLY    -0.022733   0.040595  -0.560 0.575486
PoolOutdoor_PLY    0.037811   0.038385   0.985 0.324600
LimoService_PLY   -0.027071   0.015195  -1.782 0.074834 .
Resort_PLY         0.061011   0.046413   1.315 0.188675
ShuttleService_PLY -0.056549  0.034399  -1.644 0.100202
Spa_PLY            0.051488   0.015362   3.352 0.000804 ***
ValetParking_PLY   0.075371   0.040442   1.864 0.062369 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.232 on 52290 degrees of freedom
  (1454 observations deleted due to missingness)
Multiple R-squared:  0.677,     Adjusted R-squared:  0.6769
F-statistic:  6448 on 17 and 52290 DF,  p-value: < 2.2e-16
```

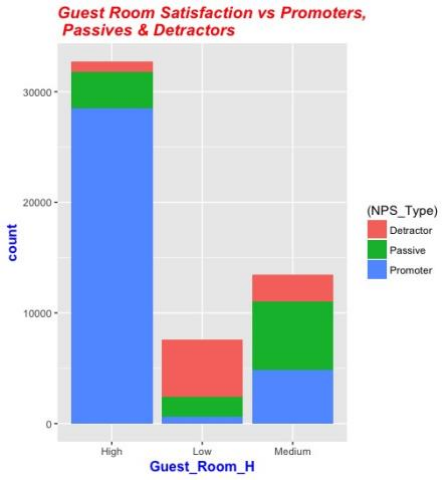
- Note that: All NA's were converted to a factor level called NA

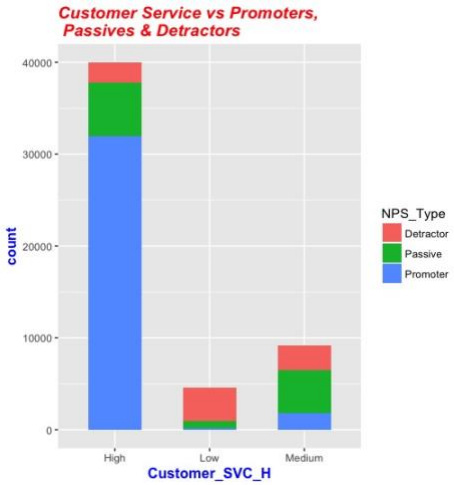
Insights from Linear Modelling:

BQ 2. What is the guest amenities relationship and effect on NPS score?

- Few key variables Guest_Room_H, Customer_SVC_H, Condition_Hotel_H, Staff_Cared_H, were most significant in contributing towards Likelihood_Recommend_H.
- Hyatt Should focus on keeping these well in order to improve its overall NPS Score.
- Hence, to do further Linear modelling and analyze significant values, we split our data into Business and Leisure subsets.
- Business and Leisure Linear models also showed 81% accuracy with these individual factors impacting the Likelihood_Recommend_H which further impacted NPS.
- Descriptive statistics plots are plotted here to visualize the impact of Detractors to the most NPS affecting Variables such as Guest_Room_H, Customer_SVC_H, Condition_Hotel_H.







- As seen in these graphs and through Linear Modelling, these three are top contributors towards Likelihood_to_recommend and NPS Scores for Hyatt Regency customers in California. (*Please note that all numerical variables were transformed from 1-10 scale to 'Low' (1-6), 'Medium' (7-8), 'High' (9,10) – similar to how NPS_Type is for SVM/ARules instead of LTR).*

Action Points for Hyatt:

- Extra attention must be paid to these variables as any improvement in the feedback or any amount of dissatisfaction pertaining to such variables will lead to large impacts on overall NPS Score.
- **A combination of above factors gave the highest R squared R^2 of 67%.**

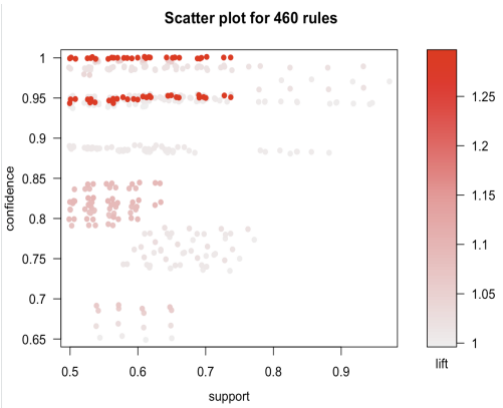
SUPPORT VECTOR MACHINES MODELLING (SVM)

The rationale

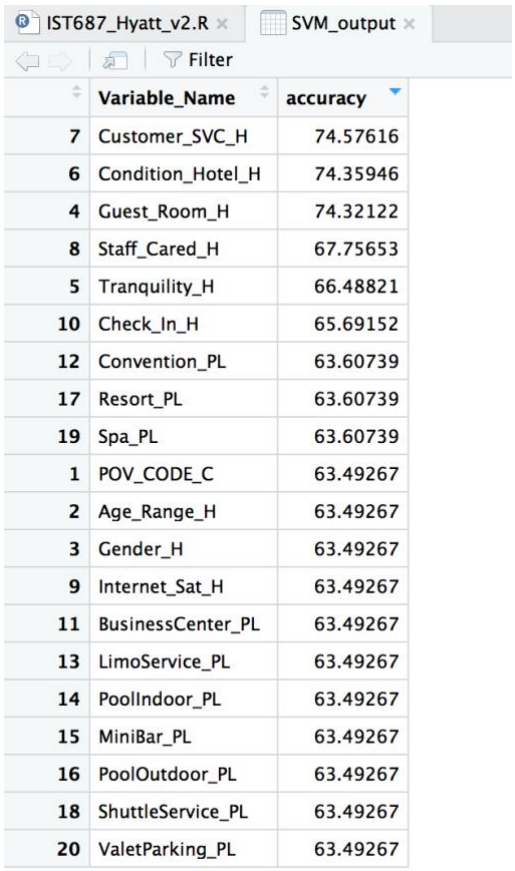
The basic idea behind SVMs can be illustrated by considering following diagram: a set of data points that belong to different classes, can be assumed to have a boundary separating the two classes is a straight line. As described on the referred website, SVM allow Class separation: basically, we are looking for the optimal separating hyperplane between the two classes by maximizing the margin between the classes' closest points. These points lying on the boundaries are called support vector
(Reference: https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf)

```
#Train and Test datasets

random_index<- sample(1:dim(Cal_Regency_modellingSVM_Leisure)[1])
cutPoint2_3 <- floor(2 * dim(Cal_Regency_modellingSVM_Leisure)[1]/3) #floor() function chops off any decimal part of the calculation.
cutPoint2_3 #31378

#Test and training sets:
trainData <- Cal_Regency_modellingSVM_Leisure[random_index[1:cutPoint2_3],] #102 obs
testData<- Cal_Regency_modellingSVM_Leisure[random_index[(cutPoint2_3+1):dim(Cal_Regency_modellingSVM_Leisure)[1]],] #51 observations
str(trainData)
```

```
>
>   svm_model<- svm(NPS_Type~Guest_Room_H,data=trainData,C=5)
>   svm_Predicted<- predict(svm_model,testData,type="responses")
>   comparison_Table_SVM <- data.frame(testData[["NPS_Type"]],svm_Predicted)
>   colnames(comparison_Table_SVM) <- c('Test_NPS','Predicted_NPS')
>   confusion_matrix_svm<- table(comparison_Table_SVM)
>   print(confusion_matrix_svm)
           Predicted_NPS
Test_NPS    Detractor Passive Promoter
  Detractor       235     104       39
  Passive          90     203      120
  Promoter         27     203     1211
>   Accuracy_svm <-((confusion_matrix_svm[1,1]+confusion_matrix_svm[2,2]+confusion_matrix_svm[3,3])/nrow(comparison_Table_SVM))*100
>   Accuracy_svm
[1] 73.87993
>
>   svm_model<- svm(NPS_Type~Customer_SVC_H,data=trainData,C=5)
>   svm_Predicted<- predict(svm_model,testData,type="responses")
>   comparison_Table_SVM <- data.frame(testData[["NPS_Type"]],svm_Predicted)
>   colnames(comparison_Table_SVM) <- c('Test_NPS','Predicted_NPS')
>   confusion_matrix_svm<- table(comparison_Table_SVM)
>   print(confusion_matrix_svm)
           Predicted_NPS
Test_NPS    Detractor Passive Promoter
  Detractor       181     107       90
  Passive          32     169      212
  Promoter         13      70     1358
>   Accuracy_svm <-((confusion_matrix_svm[1,1]+confusion_matrix_svm[2,2]+confusion_matrix_svm[3,3])/nrow(comparison_Table_SVM))*100
>   Accuracy_svm
[1] 76.5233
```



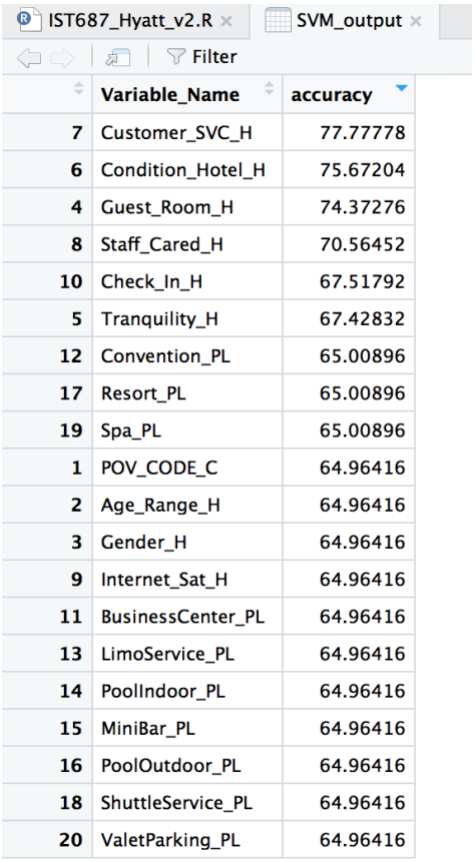
'Business' VS 'Leisure' SVM Variable Accuracy Outputs:



IST687_Hyatt_v2.R × SVM_output ×

Filter

	Variable_Name	accuracy
7	Customer_SVC_H	74.57616
6	Condition_Hotel_H	74.35946
4	Guest_Room_H	74.32122
8	Staff_Cared_H	67.75653
5	Tranquility_H	66.48821
10	Check_In_H	65.69152
12	Convention_PL	63.60739
17	Resort_PL	63.60739
19	Spa_PL	63.60739
1	POV_CODE_C	63.49267
2	Age_Range_H	63.49267
3	Gender_H	63.49267
9	Internet_Sat_H	63.49267
11	BusinessCenter_PL	63.49267
13	LimoService_PL	63.49267
14	PoolIndoor_PL	63.49267
15	MiniBar_PL	63.49267
16	PoolOutdoor_PL	63.49267
18	ShuttleService_PL	63.49267
20	ValetParking_PL	63.49267



IST687_Hyatt_v2.R × SVM_output ×

Filter

	Variable_Name	accuracy
7	Customer_SVC_H	77.77778
6	Condition_Hotel_H	75.67204
4	Guest_Room_H	74.37276
8	Staff_Cared_H	70.56452
10	Check_In_H	67.51792
5	Tranquility_H	67.42832
12	Convention_PL	65.00896
17	Resort_PL	65.00896
19	Spa_PL	65.00896
1	POV_CODE_C	64.96416
2	Age_Range_H	64.96416
3	Gender_H	64.96416
9	Internet_Sat_H	64.96416
11	BusinessCenter_PL	64.96416
13	LimoService_PL	64.96416
14	PoolIndoor_PL	64.96416
15	MiniBar_PL	64.96416
16	PoolOutdoor_PL	64.96416
18	ShuttleService_PL	64.96416
20	ValetParking_PL	64.96416

Model Outcomes and Insights Generated:

SVM Insights and Action Points for Hyatt:

The SVM Variable accuracy outputs generated the same result as our linear models and hence validated our assumption that

- Customer Service
- Hotel Condition
- Guest Room

are the most powerful variables affecting NPS Score and Likelihood_to_recommend. We further generated specific rules that govern each subset of Business Promoters vs Business Detractors, Leisure Promoters vs Leisure Detractors and similarly for Booking Factors. This has been done through ARules (Association Rules Mining)

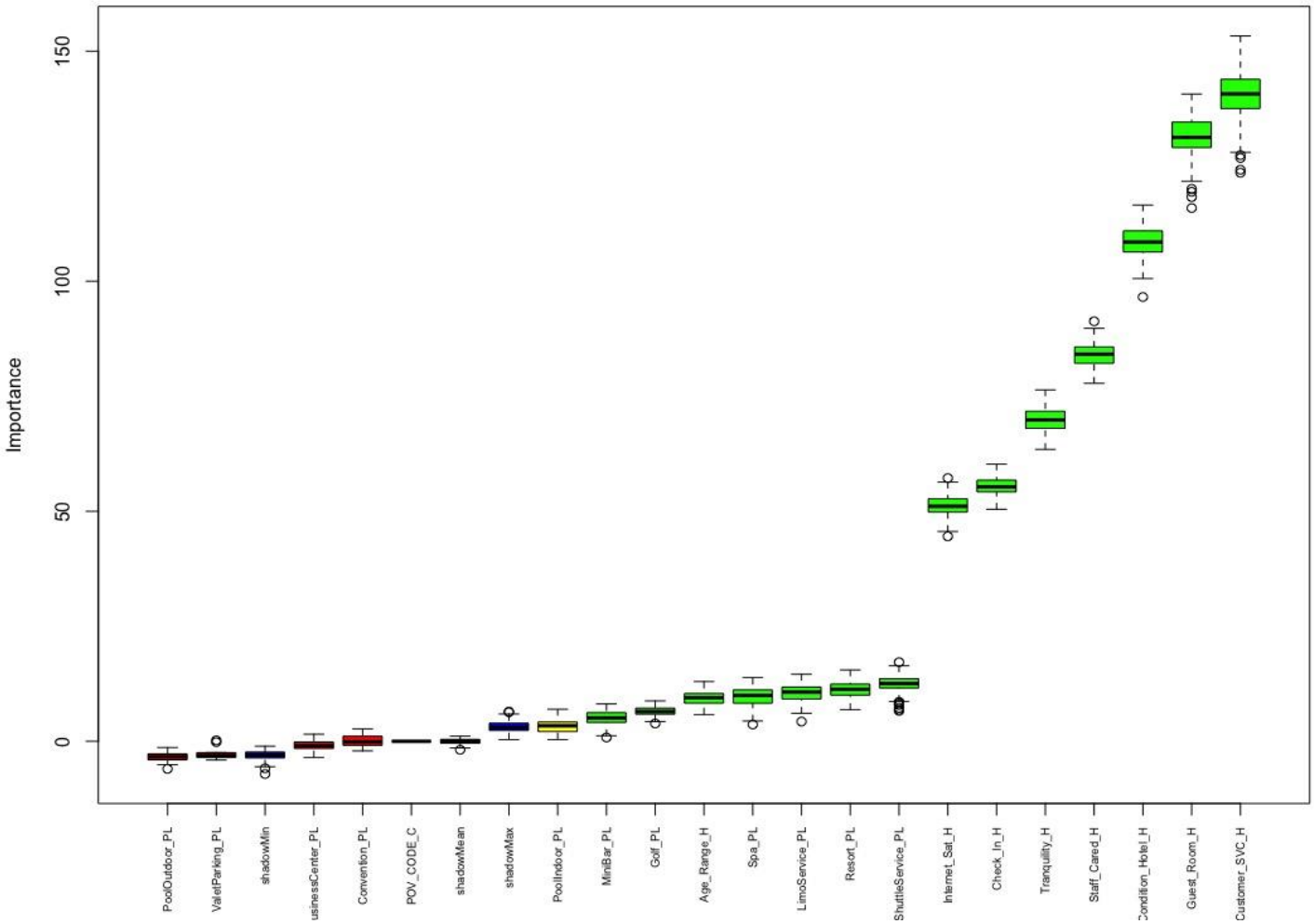
A combined version of these variables gave the highest accuracy of 81.%.

BORUTA

The machine learning Boruta algorithm, (a wrapper built around the Random Forest classification algorithm), and its implementation was also explored in the process as one last try before jumping to Association Rules. Boruta is useful to extract information from our high volume of data, hence it makes sense to use statistical techniques to reduce the noise or redundant data

```
plot(boruta.train, xlab = "", xaxt = "n")
lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)
  boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
names(lz) <- colnames(boruta.train$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
     at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.6)
```

Model Outcomes:



Boruta Insights Generated:

- Once again Boruta also proved our assumption and factors having the highest variable importance were: (in this order)
    - Customer_SVC_H
    - Guest_Room_H
    - Condition_Hotel_H
    - Staff_Cared_H
    - Tranquility_H
    - Check_In_H
    - Internet_Sat_H
    - ShuttleService_PL
    - LimoService_PL
    - Spa_PL

- Having figured out that in order to generate patterns and association between the above important factors, Association Rules mining was performed which will helped predicting rules that tell exactly what is the pattern between these factors and what Hyatt must do in order to help raise its overall NPS score.

ASSOCIATION RULES MINING (ARules) for factors Impacting NPS for Hyatt's Customers

Approach

- Done by creating ARules (Association Rules) Mining models and used apriori to study accuracy and relationship between key drivers.
- Done by studying key factors separately for each class of business and leisure
- Two models tried linear model and Apriori
- We used NPS_Type for ARules
- The idea of getting a list of key drivers was: multiple combinations of factors,
- For SVM/ARules: we transformed all numerical columns from 1-10 scale to low (1-6), _med (7-8), high (9,10) -- similarly to how we used NPS_Type for SVM/ARules instead of LTR
- RHS for ARules was specified as (NPS_Type)

Comparison of results:

Association rule learning is a rule-based machine learning method for discovering interesting relations between different variables in large databases like ours. This technique is intended to identify strong rules discovered in databases using some measures of interestingness. Scatterplot have also been used and these techniques allow an analyst effectively to explore large rule sets.

As association rule mining algorithms typically generate a large number of association rules, posing a major problem for understanding and analyzing rules, we used several visualization techniques implemented in arulesViz to explore and present sets of association rules. The metrics involved three main parameters:

- ✓ **Support**

Support is an indication of how frequently an item set appears in the dataset.

- ✓ **Confidence**

Confidence is an indication of how often the rule has been found to be true.

- ✓ **Lift**:

If the lift is > 1, it lets us know the degree to which occurrences are dependent on one another, and making those rules potentially useful for predicting the consequent in a future data sets

Data Preparation

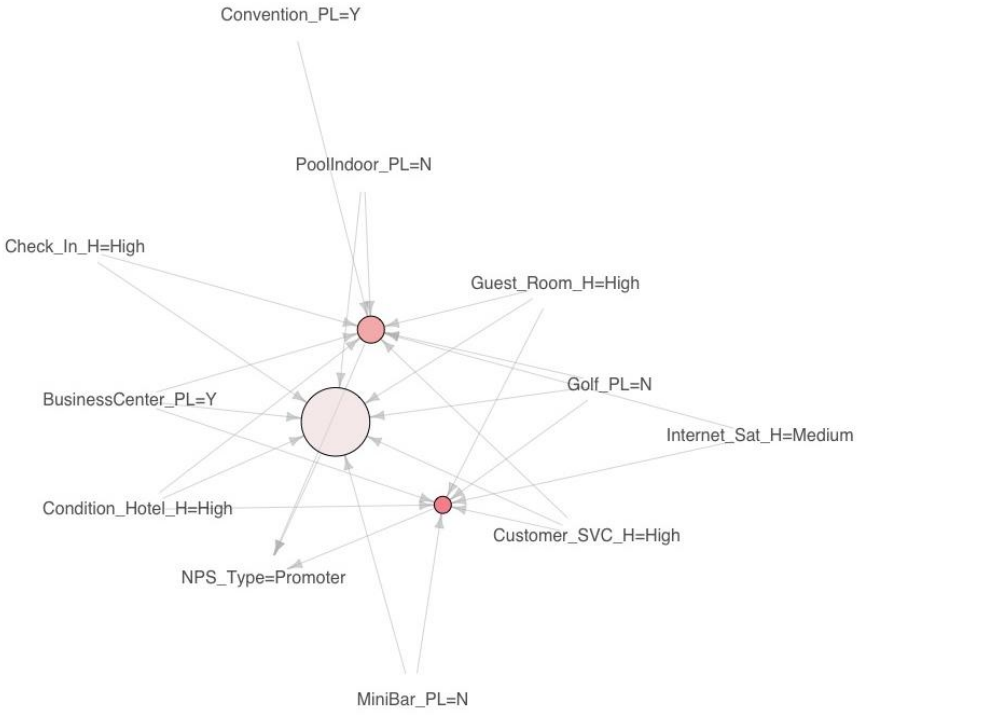
- Filtered rows based on POV_CODE_C=="BUSINESS" and POV_CODE_C =="LEISURE"
- For Booking factors, chose columns such as "Age_Range_H", "Gender_H", "ENTRY_TIME_R", "NUM_ROOMS_R", "LENGTH_OF_STAY_R", "ADULT_NUM_R", "CHILDREN_NUM_R", "Booking_Channel", "ROOM_TYPE_DESCRIPTION_C", "NPS_Type", "monthofcheckin", "checkInWeekDay", "checkOutWeekDay"

Analyzing Business Promoters

Graph for 3 rules

size: support (0.222 - 0.297)
color: lift (1.48 - 1.48)

Model Outcomes:



```
> inspect(rules_lift_Promoter)
    lhs                                rhs                    support confidence     lift count
[1] {Guest_Room_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Internet_Sat_H=Medium,
     BusinessCenter_PL=Y,
     MiniBar_PL=N}                 => {NPS_Type=Promoter} 0.2222529  0.9307768 1.480411 10461
[2] {Guest_Room_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Internet_Sat_H=Medium,
     Check_In_H=High,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     PoolIndoor_PL=N}              => {NPS_Type=Promoter} 0.2365301  0.9306194 1.480161 11133
[3] {Guest_Room_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Check_In_H=High,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     PoolIndoor_PL=N,
     MiniBar_PL=N}                 => {NPS_Type=Promoter} 0.2831648  0.9287158 1.477133 13328
[4] {Guest_Room_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Internet_Sat_H=Medium,
     Check_In_H=High,
     Convention_PL=Y,
     PoolIndoor_PL=N,
     MiniBar_PL=N}                 => {NPS_Type=Promoter} 0.2246112  0.9245300 1.470475 10572
[5] {Guest_Room_H=High,
     Tranquility_H=High,
     Customer_SVC_H=High,
     Check_In_H=High,
     BusinessCenter_PL=Y}          => {NPS_Type=Promoter} 0.2257797  0.9244085 1.470282 10627
[6] {Guest_Room_H=High,
     Tranquility_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Check_In_H=High,
     Convention_PL=Y,
     PoolIndoor_PL=N}              => {NPS_Type=Promoter} 0.2233577  0.9242198 1.469982 10513
[7] {Guest_Room_H=High,
     Tranquility_H=High,
     Condition_Hotel_H=High,
```

Insights:

- The above graph and the code showcases the key rules that have come up using association rule mining
- Beyond the three common categories of Guest Room, Condition of Hotel and Customer service which are universally true for making a customer be promoter, we also found that business customers also value tranquility, good internet connection and presence of business centers and convention centers

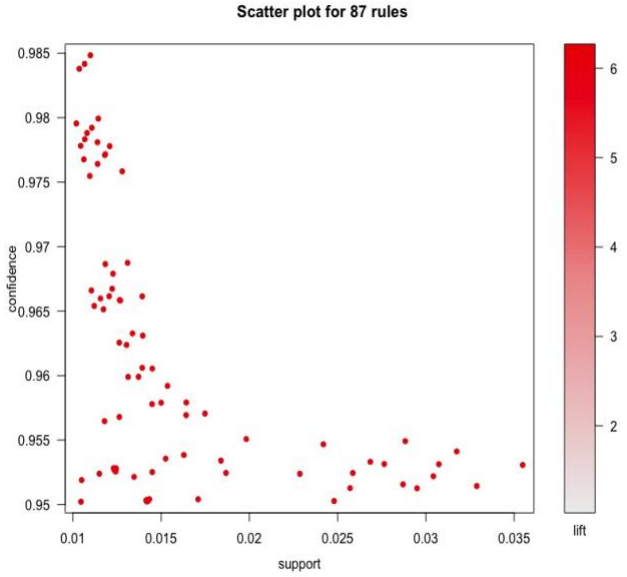
Action Points for Hyatt:

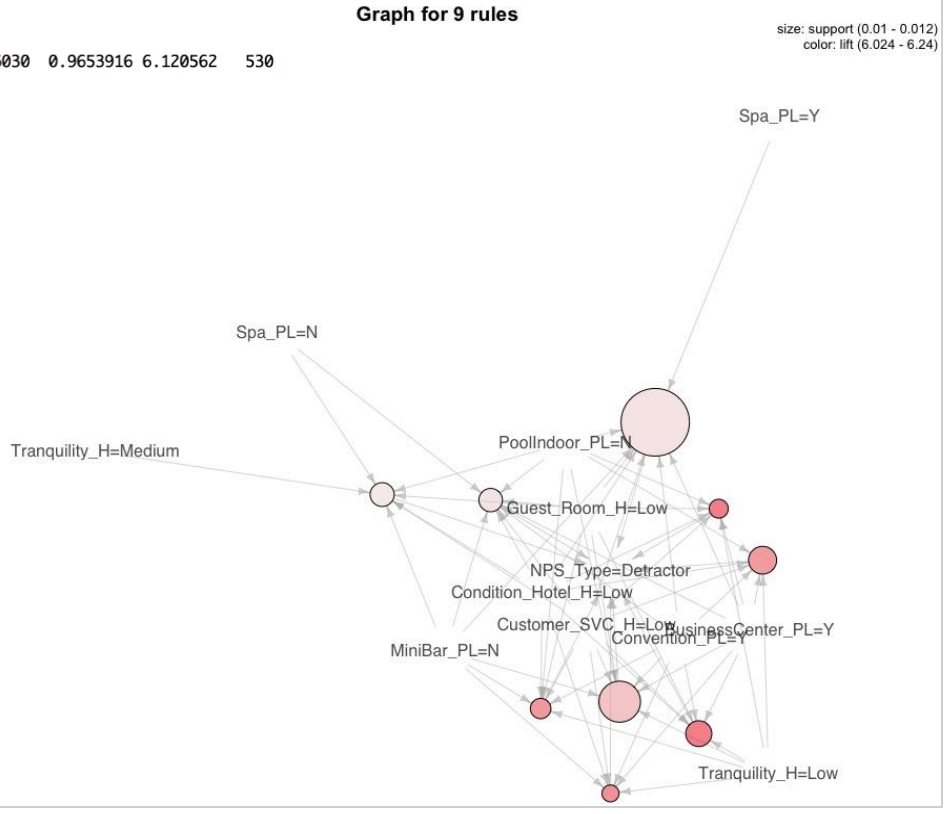
- To increase customer NPS, we recommend to Hyatt that they put high effort in providing guests with Tranquility and internet satisfaction in hotels along with business meeting facilities in the best way permissible. This makes sense because business meetings at hotels need to be peaceful and adequate infrastructure including high speed internet connection will boost business conversations and outcomes.
- This insight goes along with our linear models results and Boruta Analysis results and hence Hyatt must put extra care into improving Customer Service, Condition of the hotels and Customer Service along with better Check-in experience and high speed Internet service.

Analyzing Business Detractors

Model Outcomes:

```
> inspect(rules_lift_Detractor)
    lhs                                rhs                      support confidence     lift count
[1] {Guest_Room_H=Low,
     Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     Convention_PL=Y}           => {NPS_Type=Detractor} 0.01060168  0.9842209 6.239939   499
[2] {Guest_Room_H=Low,
     Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     PoolIndoor_PL=N}           => {NPS_Type=Detractor} 0.01030424  0.9837728 6.237098   485
[3] {Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     MiniBar_PL=N}              => {NPS_Type=Detractor} 0.01021926  0.9796334 6.210855   481
[4] {Guest_Room_H=Low,
     Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     Convention_PL=Y,
     PoolIndoor_PL=N,
     MiniBar_PL=N}              => {NPS_Type=Detractor} 0.01036798  0.9779559 6.200219   488
[5] {Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     PoolIndoor_PL=N}           => {NPS_Type=Detractor} 0.01068667  0.9766990 6.192251   503
[6] {Guest_Room_H=Low,
     Tranquility_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     PoolIndoor_PL=N,
     MiniBar_PL=N}              => {NPS_Type=Detractor} 0.01126030  0.9653916 6.120562   530
[7] {Guest_Room_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     BusinessCenter_PL=Y,
     Convention_PL=Y,
     PoolIndoor_PL=N,
```





Insights:

- The above graphs and rules show nine significant patterns observed within detractors who had travelled on business visit to Hyatt Regency in California
- The significant parameter is low tranquility amongst most of the patterns along with the standard parameters that lead to detractors, which are low condition of hotel and guest rooms along with poor customer service. The above increases the emphasis on ensuring tranquility within hotel for business customers

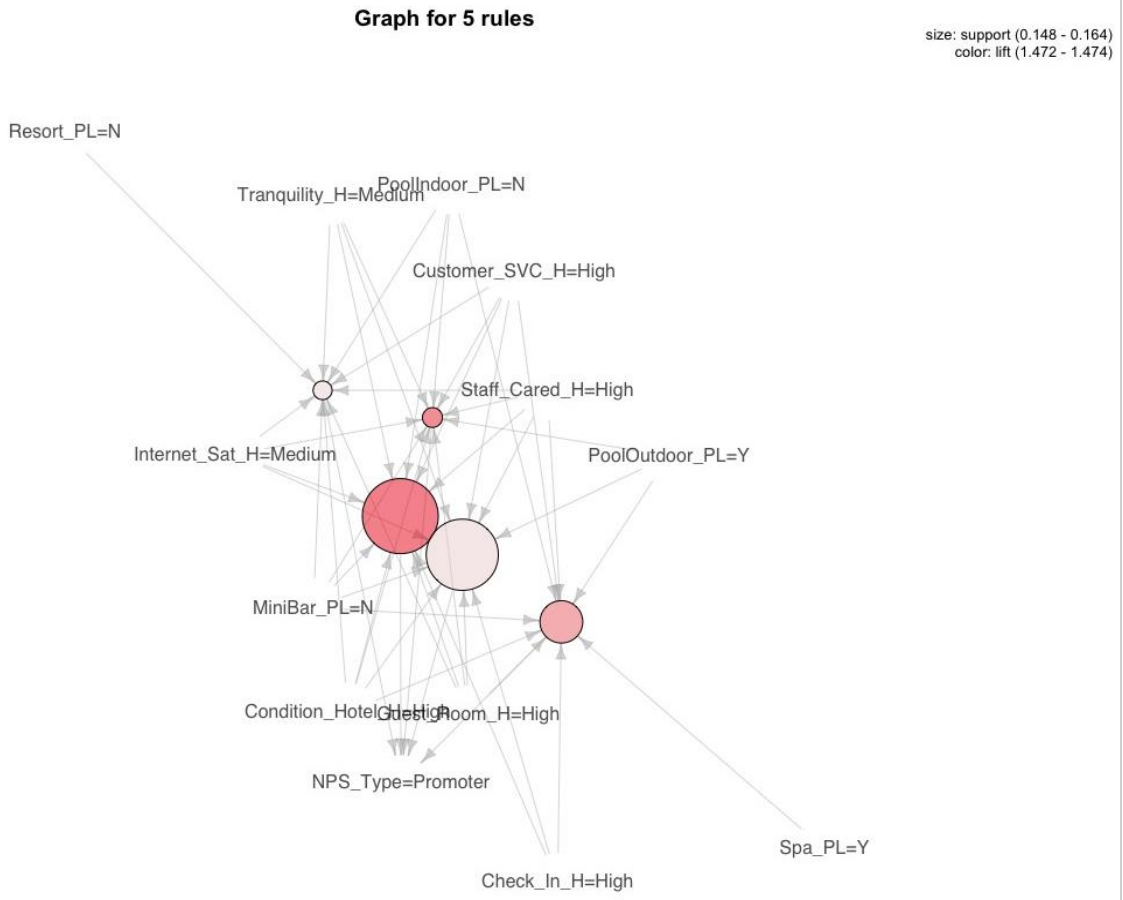
Action Points for Hyatt:

- To reduce number of detractors, in addition to the Guess Room condition, Condition of the hotel, and customer service, Hyatt must focus on ensuring tranquility within the hotels that have business centers and convention centers, lack of which pushes the customers to be detractors. Also, presence of indoor pool and minibar may affect the overall experience of these detractors.

Factors Impacting NPS for Leisure Travelers

Analyzing Leisure Promoters

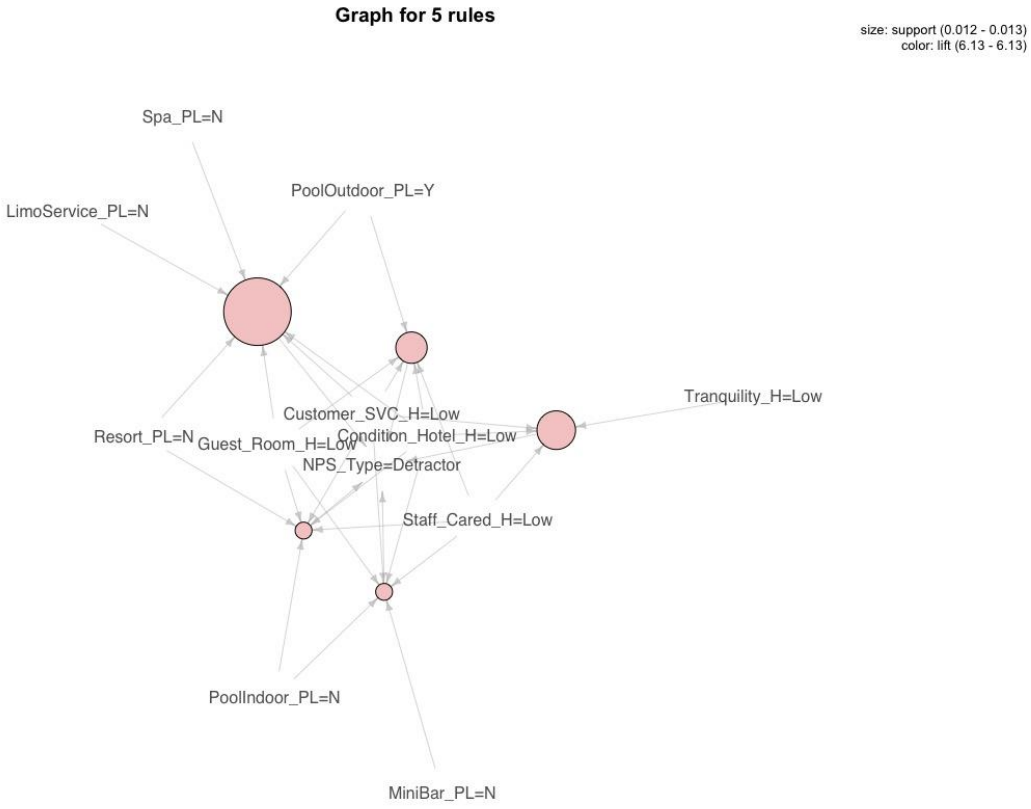
Model Outcomes:



```
[1] {Guest_Room_H=High,
     Tranquility_H=Medium,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Staff_Cared_H=High,
     Internet_Sat_H=Medium,
     Check_In_H=High,
     PoolIndoor_PL=N,
     MiniBar_PL=N}          => {NPS_Type=Promoter} 0.1640275  0.9614711 1.473801  1098
[2] {Guest_Room_H=High,
     Tranquility_H=Medium,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Staff_Cared_H=High,
     Internet_Sat_H=Medium,
     PoolIndoor_PL=N,
     MiniBar_PL=N,
     PoolOutdoor_PL=Y}      => {NPS_Type=Promoter} 0.1486406  0.9613527 1.473619   995
[3] {Guest_Room_H=High,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Staff_Cared_H=High,
     Check_In_H=High,
     PoolIndoor_PL=N,
     MiniBar_PL=N,
     PoolOutdoor_PL=Y,
     Spa_PL=Y}              => {NPS_Type=Promoter} 0.1549148  0.9610751 1.473194  1037
[4] {Guest_Room_H=High,
     Tranquility_H=Medium,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Staff_Cared_H=High,
     Internet_Sat_H=Medium,
     Check_In_H=High,
     MiniBar_PL=N,
     PoolOutdoor_PL=Y}      => {NPS_Type=Promoter} 0.1629818  0.9603873 1.472139  1091
[5] {Guest_Room_H=High,
     Tranquility_H=Medium,
     Condition_Hotel_H=High,
     Customer_SVC_H=High,
     Staff_Cared_H=High,
     Internet_Sat_H=Medium,
     PoolIndoor_PL=N,
     MiniBar_PL=N,
     Resort_PL=N}           => {NPS_Type=Promoter} 0.1483418  0.9603482 1.472079   993
>
```

- Above graphs and rules show 5 key patterns observed within Leisure travelers who were promoters. These were focus on presence of outdoor pool, spa, great check in experience and caring staff
- The parameters prove that leisure customers highly value personal experience with the hotel, like a good check in and caring staff, along and quality of amenities to relax and have fun, like outdoor pool and spas
- To increase promoter counts amongst leisure travelers we recommend Hyatt focuses on ensuring best customer service at check in and prompt and empathetic service staff

Analyzing Leisure Detractors




```
> inspect(L_rules_lift_Detractor)
    lhs                          rhs                     support confidence     lift count
[1] {Tranquility_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     Staff_Cared_H=Low}      => {NPS_Type=Detractor} 0.01239916          1 6.130037    83
[2] {Guest_Room_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     Staff_Cared_H=Low,
     PoolOutdoor_PL=Y}       => {NPS_Type=Detractor} 0.01224978          1 6.130037    82
[3] {Guest_Room_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     Staff_Cared_H=Low,
     PoolIndoor_PL=N,
     MiniBar_PL=N}           => {NPS_Type=Detractor} 0.01195100          1 6.130037    80
[4] {Guest_Room_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     Staff_Cared_H=Low,
     PoolIndoor_PL=N,
     Resort_PL=N}            => {NPS_Type=Detractor} 0.01195100          1 6.130037    80
[5] {Guest_Room_H=Low,
     Condition_Hotel_H=Low,
     Customer_SVC_H=Low,
     LimoService_PL=N,
     PoolOutdoor_PL=Y,
     Resort_PL=N,
     Spa_PL=N}               => {NPS_Type=Detractor} 0.01299671          1 6.130037    87
>
```

- The detractor patterns are highly correlated to patterns of promoters. The top parameters that lead to Leisure visit detractors are low tranquility, lack of caring staff and lack of amenities like Spa_PL and LimoService_PL
- We recommend that Hyatt considers our insights and works on improving these services

Comparing Business and Leisure

Insights and Action Points for Hyatt:

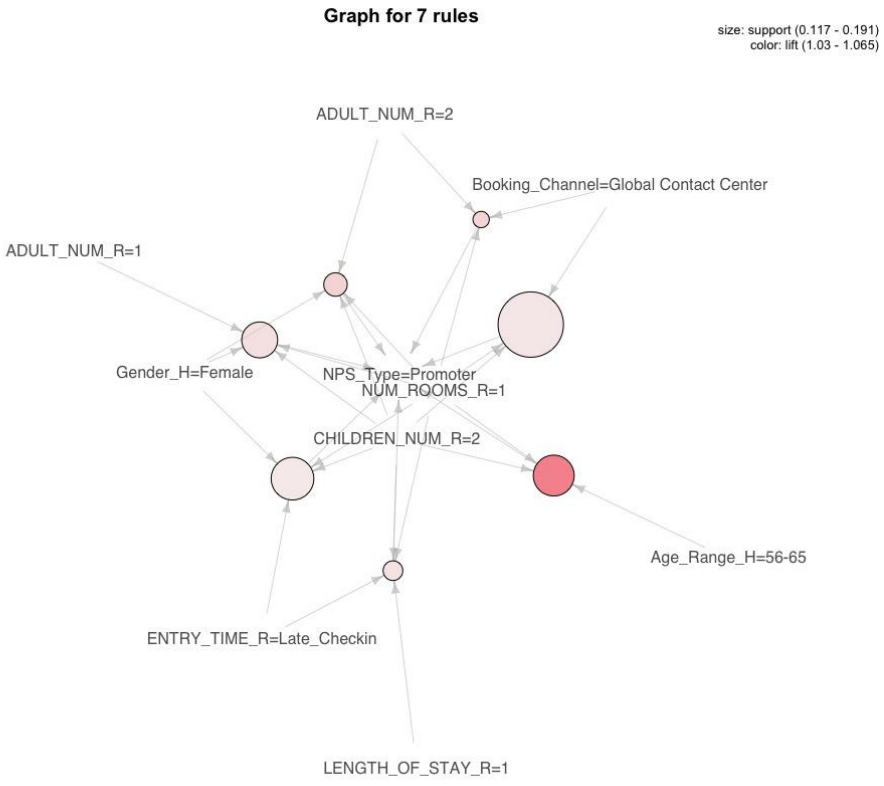
- Comparing business and leisure brings out a contradictory point on tranquility requirements. Business travelers prefer high tranquility, while leisure travelers are ok with medium tranquility which aligns with general understanding of customer requirements which is business require quiet for focusing on work, while leisure customers might look for some liveliness
- We recommend that Hyatt discusses tranquility requirement with customers during check in and allocates room accordingly to the customers, ensuring customers get the level of tranquility they are looking for more often
-

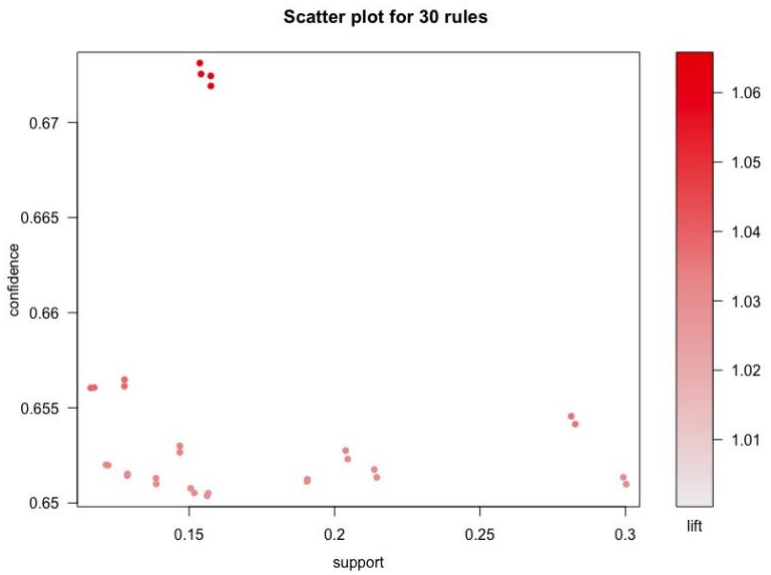
Impacting Net Promoter Score for Families

Approach

- Done by studying key factors separately for each class of business and leisure
- Two models tried linear model and Apriori
- Results of each model

Analyzing Booking Factors for Promoter Demographics






```
> inspect(BK_rules_lift_Promoter)
    lhs                                                rhs                    support confidence     lift count
[1] {Age_Range_H=56-65,
     NUM_ROOMS_R=1,
     CHILDREN_NUM_R=2}                          => {NPS_Type=Promoter} 0.1537331  0.6730456 1.065497  8265
[2] {Gender_H=Female,
     NUM_ROOMS_R=1,
     ADULT_NUM_R=2,
     CHILDREN_NUM_R=2}                          => {NPS_Type=Promoter} 0.1274506  0.6565105 1.039320  6852
[3] {NUM_ROOMS_R=1,
     ADULT_NUM_R=2,
     Booking_Channel=Global Contact Center}     => {NPS_Type=Promoter} 0.1166809  0.6560343 1.038566  6273
[4] {Gender_H=Female,
     NUM_ROOMS_R=1,
     ADULT_NUM_R=1,
     CHILDREN_NUM_R=2}                          => {NPS_Type=Promoter} 0.1462557  0.6529646 1.033707  7863
[5] {ENTRY_TIME_R=Late_Checkin,
     NUM_ROOMS_R=1,
     LENGTH_OF_STAY_R=1,
     CHILDREN_NUM_R=2}                          => {NPS_Type=Promoter} 0.1217217  0.6520526 1.032263  6544
[6] {NUM_ROOMS_R=1,
     CHILDREN_NUM_R=2,
     Booking_Channel=Global Contact Center}     => {NPS_Type=Promoter} 0.1908225  0.6511997 1.030913 10259
[7] {Gender_H=Female,
     ENTRY_TIME_R=Late_Checkin,
     NUM_ROOMS_R=1,
     CHILDREN_NUM_R=2}                          => {NPS_Type=Promoter} 0.1564488  0.6505530 1.029889  8411
>
```

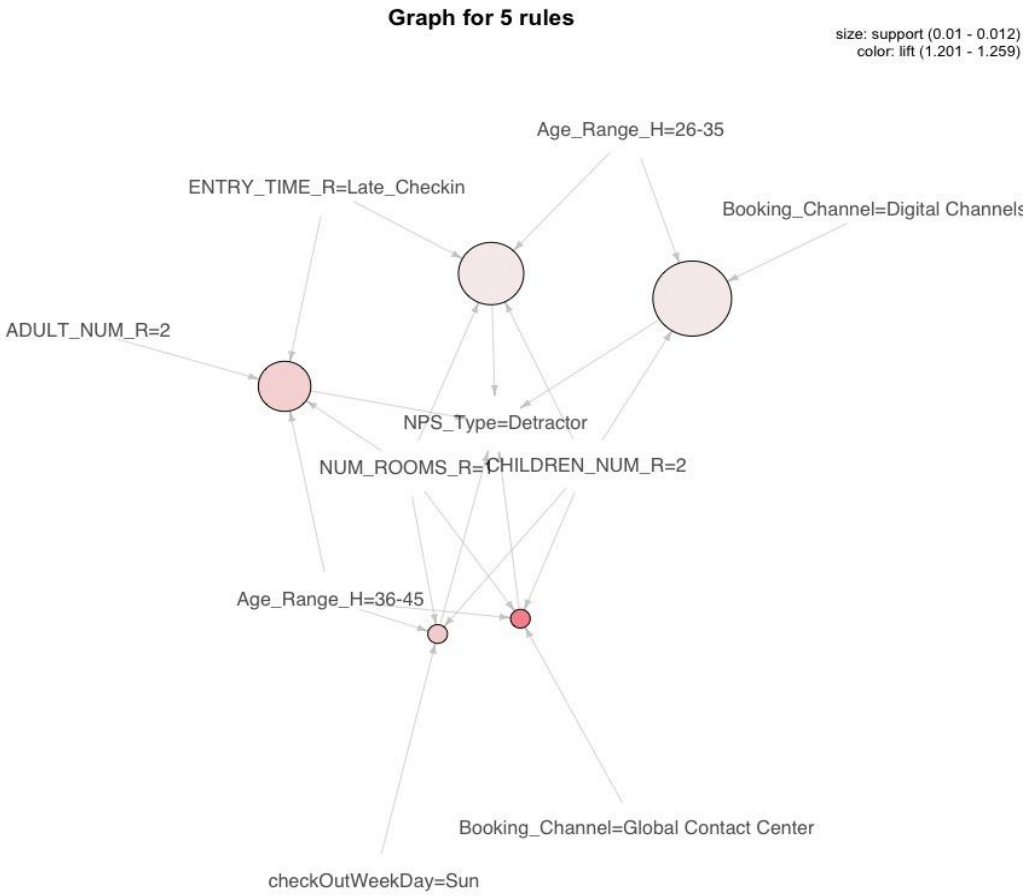
Insights:

- Younger parent(s), especially if single, seem to have a harder time at the hotel and give less ratings when with 2 children, especially when compared with older parents who are travelling together

Action Points for Hyatt: for Hyatt:

- We recommend making the hotel friendlier and helping to needs of single young parents, by providing services like play area of kids and care centers and staff for kids
- We also recommend providing discounts to parents visiting the hotel

Analyzing Booking Factors for Detractor Demographics



- Younger parent(s), especially if single, seem to have a harder time at the hotel and give less ratings when with 2 children, especially when compared with older parents who are travelling together
- We recommend making the hotel friendlier and helping to needs of single young parents, by providing services like play area of kids and care centers and staff for kids
- We also recommend providing discounts to parents visiting the hotel

Conclusion- Overall interpretation of results

In Total, this project used over 3 Data Science analytics models – Linear regression, Association Rules, Support Vector Machines and Boruta and generated following insights. Also, the data was viewed from multiple angles, and the business questions quoted before having led to generation of trends and patterns which were validated through modelling. Hence these tangible and actionable steps can be used by Hyatt Corporation CEO to improve NPS Score. This project accomplished following analysis:

- As, United States brings in most business to Hyatt group and there is scope for improving performance as its NPS score is 57% and is lower than average NPS score of 61%.
- Further, Hyatt has a good opportunity to analyze expanding its presence in countries outside the United States, as currently it has very low presence outside United States including countries which are bigger and more populous than USA (India and China).
- In United States, for fastest gains, Hyatt must of focus on states of Colorado, Virginia and Florida as their average LTR is very close to 9 and they are amongst the top most contributors by reservation counts after California.
- **In California, the project's scope was Hyatt Regency** as this brand had the highest number of records. This is due to the fact that Hyatt Regency's is Hyatt's flagship/ oldest brand in the company. On the other hand, Hyatt Regency as compared to its sister brands such as Grand Hyatt and Park Hyatt, has lesser resorts as many of these new additions to Hyatt Corporation have been styled as "resort" properties. Hyatt Regency is also a mid- to large-scale premium hotels chain intended for both leisure and business travelers, including those attending conventions, located in urban, suburban, airport, convention and resort destinations around the world. Some of these are and may have spas or other recreational facilities. Hence this project further focused on these individual visitors and took a deep dive into what factor affects these groups the most.
- Overall Key levers for NPS seem to be same across the board for all Hyatt Regency customer
  - ✓ **Customer_SVC_H:** Quality of customer service metric; value on a 1 to 10 scale
  - ✓ **Condition_Hotel_H:** Condition of hotel metric; value on a 1 to 10 scale
  - ✓ **Guest_Room_H:** Guest room satisfaction metric; value on a 1 to 10 scale

  These patterns in full data get sometimes mixed us and hence might look like noise , Hence in order to get finer local patterns we, divided the data into specific Business and Leisure population.

In general, customers can be incentivized for January as this month of the year had the lowest NPS scores. It was also noted that Weekend activities must be arranged as it turns out that Fridays have the maximum check-ins into the Hyatt Regency hotels of California. 50% of business check-ins on Friday have 2-day stay making them weekenders. Most Leisure people check in during the weekend. We further focused on Business and Leisure travelers from Hyatt regency brand of Hyatt. A combination of these factors gave the highest accuracy and R^2.

For Business Travelers:

- Beyond the three common categories of Guest Room, Condition of Hotel and Customer service which are universally true for making a customer to be promoter, we also found that business customers value Tranquility, Good internet connection and presence of Business centers and

Convention centers at Hyatt Regency in California. Hence, to increase customer NPS, we recommend to Hyatt that they put high effort in providing guests with Tranquility and internet satisfaction in hotels so that business meetings flow in the best way possible as meetings at hotels need to be peaceful and adequate infrastructure including high speed internet connection will boost business conversations and outcomes. This can be done by providing rooms to business travelers on higher floors and away from common lobbies and other noisy spaces. Apart from this Hyatt must also focus on constructing indoor pools and providing minibars as absence of these factors might lead to even worse overall satisfaction leading to low NPS.

For Leisure Travelers:

Although Hyatt Regency is most popular, but it also needs the biggest push. As Hyatt Regency Through our Association rule mining we found out that in order to increase Hyatt Regency's NPS, Hyatt must focus on more resort type products than hotels. As Hyatt Regency does not have any 'resorts' and all its properties are 'hotels', this absence was found to be linked to a low NPS. Hyatt Regency also wasn't found to have any Golf Resorts or Golf Facilities in United States, as it has internationally in Canada etc. Hence Hyatt must focus on getting these facilities as having Golf was found to have high variable importance through our machine learning algorithms. Younger parent(s), especially if single, seem to have a harder time at the hotel and give less ratings when with 2 children, especially when compared with older parents who are travelling together. We recommend making the hotel friendlier and helping to needs of single young parents, by providing services like play area of kids and care centers and staff for kids. We also recommend providing discounts to parents visiting the hotel. Leisure Promotor travelers also concentrate on presence of outdoor pool, spa, great check in experience and caring staff. The parameters prove that leisure customers highly value personal experience with the hotel, like a good check in and caring staff, along and quality of amenities to relax and have fun, like outdoor pool and spas. To increase promoter counts amongst leisure travelers we recommend Hyatt focuses on ensuring best customer service at check in and prompt and empathetic service staff. The detractor patterns are highly correlated to patterns of promoters. The top parameters that lead to Leisure visit detractors are low tranquility, lack of caring staff and lack of amenities like Spa_PL and LimoService_PL. Hence, we recommend that Hyatt considers our insights and works on improving these services. In California, most people flow to San Diego and San Francisco. Since Hyatt currently offers two types of products- hotels and resorts, having resorts in neighboring suburbs of such cities must be explored for catering better to requirements of these as having resort raises NPS for Leisure visits.

Appendix- Code

```
#Packages used:
# install.packages("data.table")
# install.packages("modeest")
# install.packages("ggplot2")
# install.packages("ggmap")
# install.packages("plyr")
# install.packages("rworldmap")
# install.packages("NPS")
# install.packages("wordcloud")
# install.packages("tm")
# install.packages("treemap")
# install.packages("countrycode")
# install.packages("lubridate")
# install.packages("arules")
# install.packages("arulesViz")
# install.packages("e1071")
# install.packages("kernlab")
# install.packages("openintro")
# install.packages("zipcode")
# install.packages("memisc")
# install.packages("openintro")
# install.packages("rminer")
# install.packages("ranger")
# install.packages("Boruta")
# install.packages("corrplot")


library('data.table')
library(modeest)
library(ggplot2)
library(ggmap)
library(plyr)
library(rworldmap)
library(NPS)
library(wordcloud)
library(tm)
library(treemap)
library(countrycode)
library(lubridate)
library(arules)
library(arulesViz)
```

```
library(e1071)
library(kernlab)
library(openintro)
library(zipcode)
library(memisc)
library(openintro)
library(rminer)
library(ranger)
library(Boruta)
library(corrplot)

#################################
# Data Import and Preprocessing
#################################

# Selecting Columns
#Choosing the columns from dataset(didnt choose columns with more than 90% blank columns and chose only 43 columns:)
columns<-c("POV_CODE_C", "Age_Range_H", "Gender_H", "Likelihood_Recommend_H", "Overall_Sat_H", "Guest_Room_H", "Tranquility_H",
        "Condition_Hotel_H", "Customer_SVC_H", "Staff_Cared_H", "Internet_Sat_H", "Check_In_H","Golf_PL",
        "City_PL", "State_PL", "Postal Code_PL","Country_PL", "Property Latitude_PL", "Property Longitude_PL",
        "Brand_PL", "Business Center_PL", "Convention_PL", "Limo Service_PL", "Pool-Indoor_PL", "Mini-Bar_PL",
        "Pool-Outdoor_PL", "Resort_PL", "Shuttle Service_PL", "Spa_PL", "Valet Parking_PL", "NPS_Type", "CHECK_IN_DATE_C",
        "CHECK_OUT_DATE_C", "ENTRY_TIME_R","NUM_ROOMS_R", "LENGTH_OF_STAY_R", "ADULT_NUM_R", "CHILDREN_NUM_R", "Booking_Channel","ROOM_TYPE_DESCRIPTION_C")

#--------------
# Data Import
# We used `fread` for every row to have the same number of columns.
Feb14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-201402.csv", header=TRUE, select=columns, verbose=TRUE)
Mar14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-201403.csv", header=TRUE, select=columns, verbose=TRUE)
Apr14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-201404.csv", header=TRUE, select=columns, verbose=TRUE)
May14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-201405.csv", header=TRUE, select=columns, verbose=TRUE)
Jun14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-201406.csv", header=TRUE, select=columns, verbose=TRUE)
```

```
Jul14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201407.csv", header=TRUE, select=columns, verbose=TRUE)
Aug14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201408.csv", header=TRUE, select=columns, verbose=TRUE)
Sep14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201409.csv", header=TRUE, select=columns, verbose=TRUE)
Oct14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201410.csv", header=TRUE, select=columns, verbose=TRUE)
Nov14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201411.csv", header=TRUE, select=columns, verbose=TRUE)
Dec14 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201412.csv", header=TRUE, select=columns, verbose=TRUE)
Jan15 <- fread(file="/Users/apurva_sharma/Downloads/IST687-data/Total_Hyatt/out-
201501.csv", header=TRUE, select=columns, verbose=TRUE)

#--------------
#Combining to one dataset
full_Test_dataset<-
rbind(Feb14,Mar14,Apr14,May14,Jun14,Jul14,Aug14,Sep14,Oct14,Nov14,Dec14,Jan15)

colnames(full_Test_dataset)<- c('POV_CODE_C', 'Age_Range_H', 'Gender_H',
'Likelihood_Recommend_H', 'Overall_Sat_H', 'Guest_Room_H', 'Tranquility_H',
                        'Condition_Hotel_H', 'Customer_SVC_H', 'Staff_Cared_H', 'Internet_Sat_H',
'Check_In_H','Golf_PL',
                        'City_PL', 'State_PL', 'PostalCode_PL','Country_PL', 'PropertyLatitude_PL',
'PropertyLongitude_PL',
                        'Brand_PL', 'BusinessCenter_PL', 'Convention_PL', 'LimoService_PL',
'PoolIndoor_PL', 'MiniBar_PL',
                        'PoolOutdoor_PL', 'Resort_PL', 'ShuttleService_PL', 'Spa_PL',
'ValetParking_PL', 'NPS_Type', 'CHECK_IN_DATE_C',
                        'CHECK_OUT_DATE_C', 'ENTRY_TIME_R','NUM_ROOMS_R',
'LENGTH_OF_STAY_R', 'ADULT_NUM_R', 'CHILDREN_NUM_R',
'Booking_Channel','ROOM_TYPE_DESCRIPTION_C')



# Handling Missing Values
# ----------------------
chunk<- full_Test_dataset
# Removing Blanks - [A good idea is to set all of the "" (blank cells) to NA before any further
analysis.]
chunk[chunk==""] <- NA

chunk<-chunk[!(is.na(chunk$NPS_Type)),]
```

```
row.names(chunk)<-NULL
chunk<-chunk[!(is.na(chunk$Likelihood_Recommend_H)),]
row.names(chunk)<-NULL
chunk<-chunk[!(is.na(chunk$Country_PL)),]
row.names(chunk)<-NULL

full_cleanData<- chunk
summary(full_cleanData)
str(full_cleanData)


###################################
# Project Scope
###################################

# World Level Analysis
# --------------------
full_cleanData1viz1<- full_cleanData

# LTR data for each Country
Country<- count(full_cleanData1viz1$Country_PL) #56 countries
Country_LTR<- tapply(full_cleanData1viz1$Likelihood_Recommend_H,
full_cleanData1viz1$Country_PL, mean, na.rm = TRUE)
LTR_df<- data.frame(Country_LTR,Country)
colnames(LTR_df)<- c("Country_LTR","Country_Name","No.ofobs")
LTR_df
#A few facts which can be observed from this plot have been stated below:
# - Poland stands out from the list of the countries and has the highest LTR of 9.64
# - Jamaica has the lowest LTR of 6.22

#NPS data for each country
#Get total number of observations in country
TotalbyCountry<-
tapply(factor(full_cleanData1viz1$NPS_Type),full_cleanData1viz1$Country_PL,length)
TotalbyCountry<- to.data.frame(TotalbyCountry,as.vars=0,name="Freq")
colnames(TotalbyCountry)<- c("Country_name","TotalFreq")

#Get number of promoters by country
full_cleanData_PromoterOnly<-
(full_cleanData1viz1[which(full_cleanData1viz1$NPS_Type=="Promoter"),])
PromoterTotalbyCountry<-
tapply(factor(full_cleanData_PromoterOnly$NPS_Type),(full_cleanData_PromoterOnly$Countr
y_PL),length)
PromoterTotalbyCountry<- to.data.frame(PromoterTotalbyCountry,as.vars=0,name="Freq")
```

```
colnames(PromoterTotalbyCountry)<- c("Country_name","PromFreq")

#Get number of detractors by country
full_cleanData_DetractorOnly<-
(full_cleanData1viz1[which(full_cleanData1viz1$NPS_Type=="Detractor"),])
DetractorTotalbyCountry<-
tapply(factor(full_cleanData_DetractorOnly$NPS_Type),full_cleanData_DetractorOnly$Country
_PL,length)
DetractorTotalbyCountry<- to.data.frame(DetractorTotalbyCountry,as.vars=0,name="Freq")
colnames(DetractorTotalbyCountry)<- c("Country_name","DetractorFreq")

#Join the promoters and detractors using country
PromDet_byCountry<-
join(PromoterTotalbyCountry,DetractorTotalbyCountry,by="Country_name")
colnames(PromDet_byCountry)<- c("Country_name","Promoter_Freq","Detractor_Freq")

#Join TotalbyCountry and PromDet_byCountry
FullNPSData_byCountry<-join(TotalbyCountry,PromDet_byCountry,by="Country_name")
FullNPSData_byCountry[is.na(FullNPSData_byCountry)]<-0
FullNPSData_byCountry$NPS_Score <- (FullNPSData_byCountry$Promoter_Freq -
FullNPSData_byCountry$Detractor_Freq)/FullNPSData_byCountry$TotalFreq *100

# Merge the two data frames(LTR and NPS dataframes)
TotalNPSdata<- merge(LTR_df,FullNPSData_byCountry,by.x = 'Country_Name',by.y  =
'Country_name')
TotalNPSdata$NPCformula<- npc(round(TotalNPSdata$Country_LTR), breaks = list(0:6, 7:8,
9:10)) #NPC formula
#View(TotalNPSdata)
# - Total of 56 countries were found to have their NPS score more than their Goal Values(Goal
NPS=53)
# - The average NPS throughout all countries is 57.7

#Plots
#loading ISO3 codes for countries from package 'countryExData'
data(countryExData)
Merged_Total_Dataframe <- merge(TotalNPSdata,countryExData,by.x = 'Country_Name',by.y  =
'Country')
sPDF <- joinCountryData2Map(Merged_Total_Dataframe, joinCode = "ISO3", nameJoinColumn
= "ISO3V10", verbose=TRUE)
mapCountryData(sPDF, nameColumnToPlot="NPS_Score",catMethod='fixedWidth',mapTitle =
"Net Promoter Score by Country") #NPS World map

# Population Data and plot
```

```
CountryTreeMap<- treemap(TotalNPSdata,index =
c("Country_Name"),vSize="TotalFreq",type="index",
                palette = "Dark2",title = "Number of Reservations by Country",fontsize.title = 14,
                fontsize.labels = 12,border.col = "white")




# United States Analysis
# ----------------------

#Data preparation
usdata<-full_cleanData[full_cleanData$Country_PL=="United States",] #744959 obs. of  40
variables
#percentage of US
dim(usdata)[1]/dim(full_cleanData)[1]*100
#[1] 80.5903


# Brand level Analysis
# Which Hyatt hotel brand do different customer prefer?
# What is the count of customers who visited hotels (hotel brand wise)?
# US hotel Hyatt brands count vs NPS_Type
tempBrandPLData<-data.frame(table(usdata$Brand_PL))
colnames(tempBrandPLData)<-c("Brand_PL","BrandStrength")
usdata2 <-merge(usdata,tempBrandPLData,by="Brand_PL")

BrandPlot <- ggplot(usdata2, aes(x = reorder(Brand_PL,-BrandStrength))) +
  geom_bar(aes(fill=NPS_Type), position="dodge") +
  scale_fill_manual(values=c("red", "green","blue")) +
  theme(axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Brand > NPS_Type",y="No. of Reservations")+
  ggtitle("Brands-NPS_Type by Number of Reservations")
BrandPlot


# California Analysis
# -------------------
CaliforniaData <- usdata[which(usdata$State_PL =="California"),]

# Data
CaliforniaData1<- CaliforniaData

#------------
```

```
#New Variables creation:

# Customer by month Variable
CaliforniaData1$monthofcheckin<-
month(as.Date(as.character.Date(CaliforniaData1$CHECK_IN_DATE_C)))
CAsummaryByMonth<-
data.frame(table(CaliforniaData1[,c("monthofcheckin","POV_CODE_C")]))

# Checkin and Checkout (Customers by day of week) Variables
CaliforniaData1$checkInWeekDay<-
wday(as.Date(as.character.Date(CaliforniaData1$CHECK_IN_DATE_C)),label = TRUE)
CaliforniaData1$checkOutWeekDay<-
wday(as.Date(as.character.Date(CaliforniaData1$CHECK_OUT_DATE_C)),label = TRUE)

# Weekender Custom Lesiure Variable
CaliforniaData1$weekender<-(CaliforniaData1$checkInWeekDay=="Fri" &
CaliforniaData1$checkOutWeekDay=="Sun")


rm(weekender, checkInWeekDay, checkOutWeekDay, monthofcheckin)


# Brand Data Preparation
# ----------------------
#splitting CalifoniaData into Hyatt Brands:
California_Regency<-CaliforniaData1[CaliforniaData1$Brand_PL=="Hyatt Regency",]

########################
# Model based analysis
########################


# Data Preparation
# ----------------
#(Removing unneeded columns and super biased columns with mostly NA's or Yes's or No's)
Cal_Regency_modelling <- subset(California_Regency,select = -
c(Brand_PL,State_PL,Country_PL,PostalCode_PL, #53762 obs. of  34 variables
                                        PropertyLatitude_PL,PropertyLongitude_PL))

# Early checkin age. Early check-in is defined as people checking in from 7AM-2PM (inclusive) as
usual chek-in is at 3PM
Cal_Regency_modelling$ENTRY_TIME_R <-
as.integer(substr(Cal_Regency_modelling$ENTRY_TIME_R, 1, 2))
```

```
tempColByAP<-
  ifelse((Cal_Regency_modelling$ENTRY_TIME_R>=7) &
(Cal_Regency_modelling$ENTRY_TIME_R<14) , "Early_Checkin",
       ifelse((Cal_Regency_modelling$ENTRY_TIME_R>=14) &
(Cal_Regency_modelling$ENTRY_TIME_R< 15 ), "Usual_Checkin",
              "Late_Checkin"
       ))
Cal_Regency_modelling$ENTRY_TIME_R<-tempColByAP


Cal_Regency_modelling$CHECK_IN_DATE_C <-
as.Date(Cal_Regency_modelling$CHECK_IN_DATE_C)
Cal_Regency_modelling$CHECK_IN_DATE_C<-
wday(Cal_Regency_modelling$CHECK_IN_DATE_C, label=TRUE)

Cal_Regency_modelling$CHECK_OUT_DATE_C <-
as.Date(Cal_Regency_modelling$CHECK_OUT_DATE_C)
Cal_Regency_modelling$CHECK_OUT_DATE_C<-
wday(Cal_Regency_modelling$CHECK_OUT_DATE_C, label=TRUE)


Cal_Regency_modelling$ENTRY_TIME_R<-as.factor(Cal_Regency_modelling$ENTRY_TIME_R)
Cal_Regency_modelling$CHECK_IN_DATE_C<-
as.factor(Cal_Regency_modelling$CHECK_IN_DATE_C)
Cal_Regency_modelling$CHECK_OUT_DATE_C<-
as.factor(Cal_Regency_modelling$CHECK_OUT_DATE_C)

str(Cal_Regency_modelling) #53762 obs. of  33 variables:
summary(Cal_Regency_modelling)

#NA HANDLING :(Removing unneeded columns and super biased columns)
#Determine if columns have any NA's

#replacing NA's with mean values
Cal_Regency_modelling$Guest_Room_H[is.na(Cal_Regency_modelling$Guest_Room_H)] <-
round(mean(Cal_Regency_modelling$Guest_Room_H, na.rm = TRUE))
Cal_Regency_modelling$Overall_Sat_H[is.na(Cal_Regency_modelling$Overall_Sat_H)] <-
round(mean(Cal_Regency_modelling$Overall_Sat_H, na.rm = TRUE))
Cal_Regency_modelling$Tranquility_H[is.na(Cal_Regency_modelling$Tranquility_H)] <-
round(mean(Cal_Regency_modelling$Tranquility_H, na.rm = TRUE))
Cal_Regency_modelling$Condition_Hotel_H[is.na(Cal_Regency_modelling$Condition_Hotel_H)
] <- round(mean(Cal_Regency_modelling$Condition_Hotel_H, na.rm = TRUE))
Cal_Regency_modelling$Customer_SVC_H[is.na(Cal_Regency_modelling$Customer_SVC_H)] <-
round(mean(Cal_Regency_modelling$Customer_SVC_H, na.rm = TRUE))
```

```
Cal_Regency_modelling$Staff_Cared_H[is.na(Cal_Regency_modelling$Staff_Cared_H)] <-
round(mean(Cal_Regency_modelling$Staff_Cared_H, na.rm = TRUE))
Cal_Regency_modelling$Internet_Sat_H[is.na(Cal_Regency_modelling$Internet_Sat_H)] <-
round(mean(Cal_Regency_modelling$Internet_Sat_H, na.rm = TRUE))
Cal_Regency_modelling$Check_In_H[is.na(Cal_Regency_modelling$Check_In_H)] <-
round(mean(Cal_Regency_modelling$Check_In_H, na.rm = TRUE))
Cal_Regency_modelling$LENGTH_OF_STAY_R[is.na(Cal_Regency_modelling$LENGTH_OF_STAY
_R)] <- round(mean(Cal_Regency_modelling$LENGTH_OF_STAY_R, na.rm = TRUE))
Cal_Regency_modelling$ADULT_NUM_R[is.na(Cal_Regency_modelling$ADULT_NUM_R)] <-
round(mean(Cal_Regency_modelling$ADULT_NUM_R, na.rm = TRUE))
Cal_Regency_modelling$CHILDREN_NUM_R[is.na(Cal_Regency_modelling$CHILDREN_NUM_R)
] <- round(mean(Cal_Regency_modelling$CHILDREN_NUM_R, na.rm = TRUE))

#Linear Modelling Data Preparation:#as it requires numeric variables . Saving this before
converting to factor.
California_Regency_Cal_Regency_LMData<- Cal_Regency_modelling #Linear modelling main
modelling data !
#summary(California_Regency_Cal_Regency_LMData)

# Converting all numerical columns into " High", " Medium"," Low"
str(Cal_Regency_modelling)

convertToCategorical<-function(givenDataFrame,numColName)
{
  givenDataFrame[[numColName]]<-as.numeric(givenDataFrame[[numColName]])
  tempCol<-ifelse(givenDataFrame[[numColName]]>=9, "High",
                  ifelse(givenDataFrame[[numColName]]>=7 , "Medium",
                         "Low" ))
  return (tempCol)
}

numericVariableList<-
c("Likelihood_Recommend_H","Overall_Sat_H","Guest_Room_H","Tranquility_H","Condition_
Hotel_H","Customer_SVC_H","Staff_Cared_H","Internet_Sat_H","Check_In_H")

for (colName in numericVariableList)
{
  Cal_Regency_modelling[[colName]]<-convertToCategorical(Cal_Regency_modelling,colName)
}
View(Cal_Regency_modelling)


#Converting to factor
Cal_Regency_modelling[ ] <- lapply(Cal_Regency_modelling, factor)
```

```
#Missing values in Categorical columns (we added a new level called NA in place of NA's by
using following command)
Cal_Regency_modelling$Age_Range_H<- addNA(Cal_Regency_modelling$Age_Range_H)
Cal_Regency_modelling$Gender_H<- addNA(Cal_Regency_modelling$Gender_H)
Cal_Regency_modelling$Age_Range_H<- addNA(Cal_Regency_modelling$Age_Range_H)
Cal_Regency_modelling$BusinessCenter_PL<-
addNA(Cal_Regency_modelling$BusinessCenter_PL)
Cal_Regency_modelling$Convention_PL<- addNA(Cal_Regency_modelling$Convention_PL)
Cal_Regency_modelling$LimoService_PL<- addNA(Cal_Regency_modelling$LimoService_PL)
Cal_Regency_modelling$MiniBar_PL<- addNA(Cal_Regency_modelling$MiniBar_PL)
Cal_Regency_modelling$PoolIndoor_PL<- addNA(Cal_Regency_modelling$PoolIndoor_PL)
Cal_Regency_modelling$PoolOutdoor_PL<- addNA(Cal_Regency_modelling$PoolOutdoor_PL)
Cal_Regency_modelling$ShuttleService_PL<-
addNA(Cal_Regency_modelling$ShuttleService_PL)
Cal_Regency_modelling$Spa_PL<- addNA(Cal_Regency_modelling$Spa_PL)
Cal_Regency_modelling$Resort_PL<- addNA(Cal_Regency_modelling$Resort_PL)
Cal_Regency_modelling$ValetParking_PL<- addNA(Cal_Regency_modelling$ValetParking_PL)
Cal_Regency_modelling$Golf_PL<- addNA(Cal_Regency_modelling$Golf_PL)

#summary(Cal_Regency_modelling)  # All NA's in numerical columns have been replaced .


Cal_Regency_modelling1<- Cal_Regency_modelling

############################
#POV level
############################
#POV analysis:
#Comparing within groups of hotels allows lower performing hotels to learn from higher
performing sister hotels.
HotelPlot <- ggplot(usdata, aes(x=Brand_PL)) +
  geom_bar(aes(fill=POV_CODE_C), position="dodge") +
  scale_fill_manual(values=c("Purple", "Orange")) +
  theme(axis.text.x = element_text(angle=60, hjust=1))+
  ggtitle("POV analysis on different brands of Hyatt hotels")
HotelPlot


###################
#Demographics Analysis:
###################
#Demographics and NPS: Purpose of Visit by Age
#pov, age bargraph
```

```
NPS_age<-ggplot(Cal_Regency_modelling, aes(Age_Range_H, fill = POV_CODE_C))+
  geom_bar()+
  ggtitle("Purpose of Visit by Age")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
NPS_age #maximum senior managers #age vs nps - max detractors: 46-55

#focus on males more:
Gender_Plot<-ggplot(Cal_Regency_modelling, aes(Gender_H, fill = POV_CODE_C))+
  geom_bar()+
  ggtitle("Purpose of Visit by Gender")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
Gender_Plot #maximum senior managers


#######################################
#Booking Factors
#-----------------------------------------------------------------
# focus on booking channel- digital
Booking_Channelplot<-ggplot(Cal_Regency_modelling, aes(Booking_Channel, fill = NPS_Type))+
  geom_bar()+ scale_fill_manual(values=c("red", "blue","green"))+
  ggtitle("Booking_Channel vs NPS")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
Booking_Channelplot


#which day of the week most imp for check in ?- friday !
CHECK_IN_DATE_Cplot<-ggplot(Cal_Regency_modelling, aes(CHECK_IN_DATE_C, fill =
NPS_Type))+
  geom_bar()+ scale_fill_manual(values=c("red", "blue","green"))+
  ggtitle("CHECK_IN_DATE_C vs NPS")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
CHECK_IN_DATE_Cplot

#sundays are critically imp for checkouts !
CHECK_OUT_DATE_Cplot<-ggplot(Cal_Regency_modelling, aes(CHECK_OUT_DATE_C, fill =
NPS_Type))+
  geom_bar()+ scale_fill_manual(values=c("red", "blue","green"))+
  ggtitle("CHECK_OUT_DATE_C vs NPS")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
CHECK_OUT_DATE_Cplot

#most detractors in late checkin !
ENTRY_TIME_Rplot<-ggplot(Cal_Regency_modelling, aes(ENTRY_TIME_R, fill = NPS_Type))+
  geom_bar()+ scale_fill_manual(values=c("red", "blue","green"))+
```

```
  ggtitle("ENTRY_TIME_R vs NPS")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
ENTRY_TIME_Rplot

#Max people stay for 1 day and most detractors are in this time frame
lengthplot<-ggplot(Cal_Regency_modelling, aes(LENGTH_OF_STAY_R, fill = NPS_Type))+
  geom_bar()+
  ggtitle("Length of Stay")+
  theme(axis.text.x = element_text(angle=60, hjust=1))
lengthplot
#Max people stay for 1 day and most detractors are in this time frame


####################
#Facility Analysis:
####################
# What are the various services and parameters affect the promoters and detractors?

#Parameter: Guest Room
HotelGuest_Room_H <-
  ggplot(Cal_Regency_modelling, aes(x=Guest_Room_H)) +
  geom_bar(aes(fill=(NPS_Type)),na.rm = TRUE) +
  ggtitle("Guest Room Satisfaction vs Promoters, \n Passives & Detractors")
HotelGuest_Room_H

#Parameter: Condition of the hotel
HotelCondition_Hotel_H <- ggplot(Cal_Regency_modelling, aes(x=Condition_Hotel_H)) +
  geom_bar(
    aes(fill=NPS_Type),na.rm = TRUE,width=.56) +
  ggtitle("Condition of Hotel vs Promoters, \n Passives & Detractors")
HotelCondition_Hotel_H

# Effect of Good Customer Service
HotelCustomerService <- ggplot(Cal_Regency_modelling, aes(x=Customer_SVC_H)) +
  geom_bar(aes(fill=NPS_Type),na.rm = TRUE,width=.56)  +
  ggtitle("Customer Service vs Promoters, \n Passives & Detractors")
HotelCustomerService


#-------------------------------Modelling :-------------------------------------------------------------
#################
#MODELLING
#################
```

```
#A)LINEAR MODELLING
#-----------------------------------------------------------------
#Linear Modelling:
colnames(California_Regency_Cal_Regency_LMData)
California_Regency_Cal_Regency_LMData1<-California_Regency_Cal_Regency_LMData
str(California_Regency_Cal_Regency_LMData1)
summary(as.factor(California_Regency_Cal_Regency_LMData1$Golf_PL))

#Removing Things that Hyatt cannot control:
Cal_Regency_LM <- subset(California_Regency_Cal_Regency_LMData1,select = -
c(City_PL,CHECK_IN_DATE_C,CHECK_OUT_DATE_C,ENTRY_TIME_R,

NUM_ROOMS_R,LENGTH_OF_STAY_R,ADULT_NUM_R,CHILDREN_NUM_R,
                                                NPS_Type,Booking_Channel))
summary(Cal_Regency_LM)

lm_Total <- lm(formula=Likelihood_Recommend_H ~ POV_CODE_C+Guest_Room_H+
Overall_Sat_H+Tranquility_H +Condition_Hotel_H+Customer_SVC_H+
            Staff_Cared_H+Internet_Sat_H+Check_In_H+BusinessCenter_PL+
Convention_PL+MiniBar_PL+PoolIndoor_PL+PoolOutdoor_PL+LimoService_PL+

Resort_PL+ShuttleService_PL+Spa_PL+ValetParking_PL+Age_Range_H+Gender_H,data=Cal_Reg
ency_LM)
summary(lm_Total) #0.8336



#Business:
California_Regency_Business_lm<-
Cal_Regency_LM[Cal_Regency_LM$POV_CODE_C=="BUSINESS",]
str(California_Regency_Business_lm)

lm_Business <- lm(formula=Likelihood_Recommend_H ~
Overall_Sat_H+Internet_Sat_H+BusinessCenter_PL+
Convention_PL+MiniBar_PL+PoolOutdoor_PL+

ShuttleService_PL+LimoService_PL+Spa_PL+ValetParking_PL,data=California_Regency_Business
_lm)
summary(lm_Business)
applySVMModel(lm_Business)

#Leisure:
```

```
California_Regency_Leisure_lm<-
Cal_Regency_LM[Cal_Regency_LM$POV_CODE_C=="LEISURE",]
str(California_Regency_Leisure_lm)

lm_Leisure <- lm(formula=Likelihood_Recommend_H ~
Overall_Sat_H+Internet_Sat_H+Spa_PL+ShuttleService_PL+ValetParking_PL+BusinessCenter_PL
+PoolOutdoor_PL+
              Convention_PL+MiniBar_PL+LimoService_PL,data=California_Regency_Leisure_lm)
summary(lm_Leisure)


#Booking factors
Cal_Regency_LM_BF<- subset(California_Regency_Cal_Regency_LMData1,select =
c(Gender_H,Age_Range_H, Likelihood_Recommend_H,Overall_Sat_H,City_PL,ENTRY_TIME_R,

NUM_ROOMS_R,LENGTH_OF_STAY_R,ADULT_NUM_R,CHILDREN_NUM_R,
                                                Booking_Channel))
summary(Cal_Regency_LM_BF)
str(Cal_Regency_LM_BF)

#Converting to factor
Cal_Regency_LM_BF$City_PL<- as.factor(Cal_Regency_LM_BF$City_PL)
Cal_Regency_LM_BF$Booking_Channel<- as.factor(Cal_Regency_LM_BF$Booking_Channel)
Cal_Regency_LM_BF$ENTRY_TIME_R<- as.factor(Cal_Regency_LM_BF$ENTRY_TIME_R)

lm_BookingF <- lm(formula=Likelihood_Recommend_H ~
Overall_Sat_H+ENTRY_TIME_R+NUM_ROOMS_R+LENGTH_OF_STAY_R+Booking_Channel+
               ADULT_NUM_R+CHILDREN_NUM_R,data=Cal_Regency_LM_BF)
summary(lm_BookingF)


Corrdf<- subset(Cal_Regency_LM,select =
c(Likelihood_Recommend_H,Guest_Room_H,Overall_Sat_H,Tranquility_H,Condition_Hotel_H,C
ustomer_SVC_H,Staff_Cared_H,
                           Internet_Sat_H,Check_In_H))

Cal_Regency_LM2<-Cal_Regency_LM

#
# my_data<- Corrdf
# res <- cor(my_data)
# View(round(res, 2))
# cor(my_data, use = "complete.obs")
# res2 <- rcorr(as.matrix(my_data))
```

```
#
# flattenCorrMatrix <- function(cormat, pmat)
# {
#   ut <- upper.tri(cormat)
#   data.frame(
#     row = rownames(cormat)[row(cormat)[ut]],
#     column = rownames(cormat)[col(cormat)[ut]],
#     cor  =(cormat)[ut],
#     p = pmat[ut]
#   )
# }
#
# library(Hmisc)
# res2<-rcorr(as.matrix(my_data))
# flattenCorrMatrix(res2$r, res2$P)
# symnum(res, abbr.colnames = FALSE)

# #Plot:
# corrplot(res, type = "upper", order = "hclust",
#         tl.col = "black", tl.srt = 45)
#
# # Insignificant correlation are crossed
# corrplot(res2$r, type="upper", order="hclust",
#         p.mat = res2$P, sig.level = 0.01, insig = "blank")
# # Insignificant correlations are leaved blank
# corrplot(res2$r, type="upper", order="hclust",
#         p.mat = res2$P, sig.level = 0.01, insig = "blank")
#

#---------------------------------
model1 <- lm(formula=Likelihood_Recommend_H ~ Guest_Room_H, data=Cal_Regency_LM)
summary(model1) #R square= 0.5286 #important

model2 <- lm(formula=Likelihood_Recommend_H ~ Overall_Sat_H, data=Cal_Regency_LM)
summary(model2) #R square= 0.8169

model3 <- lm(formula=Likelihood_Recommend_H ~ Tranquility_H, data=Cal_Regency_LM)
summary(model3) #R square= 0.2068

model4 <- lm(formula=Likelihood_Recommend_H ~ Internet_Sat_H, data=Cal_Regency_LM)
summary(model4) #R square= 0.03152 #not important
```

```
model5 <- lm(formula=Likelihood_Recommend_H ~ Condition_Hotel_H,
data=Cal_Regency_LM)
summary(model5) #R square= 0.5066 #important !

model6 <- lm(formula=Likelihood_Recommend_H ~ Customer_SVC_H, data=Cal_Regency_LM)
summary(model6) #R square= 0.4838 #important !

model7 <- lm(formula=Likelihood_Recommend_H ~ Staff_Cared_H, data=Cal_Regency_LM)
summary(model7) #R square= 0.2303

model8 <- lm(formula=Likelihood_Recommend_H ~ Check_In_H, data=Cal_Regency_LM)
summary(model8) #R square= 0.1274

model9 <- lm(formula=Likelihood_Recommend_H ~ LimoService_PL, data=Cal_Regency_LM)
summary(model9) #R square= 0.002634

model10 <- lm(formula=Likelihood_Recommend_H ~ MiniBar_PL, data=Cal_Regency_LM)
summary(model10) #R square= 0.002735

model11 <- lm(formula=Likelihood_Recommend_H ~ BusinessCenter_PL,
data=Cal_Regency_LM)
summary(model11) #R square= 0.001348

model12 <- lm(formula=Likelihood_Recommend_H ~ POV_CODE_C, data=Cal_Regency_LM)
summary(model12) #R square= 1.512e-05

model13 <- lm(formula=Likelihood_Recommend_H ~ PoolIndoor_PL, data=Cal_Regency_LM)
summary(model13) #R square= 0.0004427

model14 <- lm(formula=Likelihood_Recommend_H ~ PoolOutdoor_PL, data=Cal_Regency_LM)
summary(model14) #R square= 0.0001833

model15 <- lm(formula=Likelihood_Recommend_H ~ Resort_PL, data=Cal_Regency_LM)
summary(model15) #R square= 0.002776

model16 <- lm(formula=Likelihood_Recommend_H ~ ShuttleService_PL,
data=Cal_Regency_LM)
summary(model16) #R square= 0.003479

model17 <- lm(formula=Likelihood_Recommend_H ~ ValetParking_PL, data=Cal_Regency_LM)
summary(model17) #R square=  3.79e-06

model19 <- lm(formula=Likelihood_Recommend_H ~ Age_Range_H, data=Cal_Regency_LM)
summary(model19) #R square=  0.008222
```

```
model20 <- lm(formula=Likelihood_Recommend_H ~ Gender_H, data=Cal_Regency_LM)
summary(model20) #R square= 0.002638

# #According to the linear model, Guest_Room_H +Condition_Hotel_H+Customer_SVC_H are
all powerful columns
# Customer Service contributes the most to LTR followerd by Guest_Room_H,
Condition_Hotel_H and Staff_Cared_H

#-------------------------------------------------------------------------------------------------------

#B)Association Rule Mining to further prove the relationship between amanities that affect
Business and Leisure people individually.
#-------------------------

#Splitting into business, leisure and Facility !

#Business:
California_Regency_Business<-
Cal_Regency_modelling[Cal_Regency_modelling$POV_CODE_C=="BUSINESS",]
California_Regency_Business <- subset(California_Regency_Business,select = -c(POV_CODE_C))
str(California_Regency_Business)

ModellingData_Business<- subset(California_Regency_Business,select = -
c(Age_Range_H,Gender_H,Overall_Sat_H,Resort_PL,ShuttleService_PL,CHILDREN_NUM_R,

Staff_Cared_H,ValetParking_PL,LimoService_PL,PoolOutdoor_PL,CHECK_IN_DATE_C,CHECK_OU
T_DATE_C,

ENTRY_TIME_R,NUM_ROOMS_R,LENGTH_OF_STAY_R,ADULT_NUM_R,CHILDREN_NUM_R,Boo
king_Channel,
                                                ROOM_TYPE_DESCRIPTION_C,Golf_PL))

str(ModellingData_Business)

#Leisure:
California_Regency_Leisure<-
Cal_Regency_modelling[Cal_Regency_modelling$POV_CODE_C=="LEISURE",]
California_Regency_Leisure <- subset(California_Regency_Leisure,select = -c(POV_CODE_C))
str(California_Regency_Leisure)

ModellingData_Leisure <- subset(California_Regency_Leisure,select = -
c(Age_Range_H,Gender_H,Overall_Sat_H,Convention_PL,BusinessCenter_PL,ValetParking_PL,
```

```
CHECK_IN_DATE_C,CHECK_OUT_DATE_C,ENTRY_TIME_R,NUM_ROOMS_R,LENGTH_OF_STAY_
R,ADULT_NUM_R,CHILDREN_NUM_R,Booking_Channel,

ROOM_TYPE_DESCRIPTION_C,Likelihood_Recommend_H,Golf_PL))

str(ModellingData_Leisure)


#Booking Factors Analysis:
#Business:
Modelling_Booking_Factors <- subset(Cal_Regency_modelling,select =
c("Age_Range_H","Gender_H", "ENTRY_TIME_R","NUM_ROOMS_R","LENGTH_OF_STAY_R",

"ADULT_NUM_R","CHILDREN_NUM_R","Booking_Channel","ROOM_TYPE_DESCRIPTION_C","N
PS_Type"
                                                ))
str(Modelling_Booking_Factors)



#A)Business
#####################################

#Business_Promoters:
#---------------------------------

rules_Promoter <-
apriori(ModellingData_Business,parameter=list(support=.53,confidence=0.92,maxlen=5),appea
rance =
                        list(rhs=c("NPS_Type=Promoter"),default="lhs"), control=list(verbose=F))
summary(rules_Promoter)

#good rules_Promoter
goodrules_Promoter <- rules_Promoter[quality(rules_Promoter)$lift > 1.58] #just 1 rule--- play
around with it to get atleast 5-6 rules
#inspect(goodrules_Promoter)
summary(goodrules_Promoter)
plot(goodrules_Promoter)

#Pick the  most interesting & useful rules.
max_Promoter<- is.maximal(goodrules_Promoter) #Find Maximal Itemsets
#inspect(goodrules[max_Promoter])
MostInteresting_Promoter<-(goodrules_Promoter[max_Promoter])
```

```
summary(MostInteresting_Promoter)

 detach(package:tm, unload=TRUE)
library(arules)

#the NPS recommendation rules:
rules_conf_Promoter <- sort(MostInteresting_Promoter, decreasing = TRUE,by="confidence")
#high-confidence rules
inspect(rules_conf_Promoter) #98 % confidence
rules_lift_Promoter <- sort(MostInteresting_Promoter, decreasing = TRUE,by="lift") # high-lift
rules
inspect(rules_lift_Promoter)

plot(rules_lift_Promoter, method="graph")



#Business_Detractors:
#--------------------------------
rules_Detractor <-
apriori(ModellingData_Business,parameter=list(support=.086,confidence=0.95,maxlen=6),appe
arance =
                        list(rhs=c("NPS_Type=Detractor"),default="lhs"), control=list(verbose=F))
summary(rules_Detractor) #87 rules
#plot(rules_Detractor)
#inspect(rules_Detractor)

#good rules_Detractor
goodrules_Detractor <- rules_Detractor[quality(rules_Detractor)$lift > 6.339]
#inspect(goodrules_Promoter) # 87
summary(goodrules_Detractor)
plot(goodrules_Detractor)

#Pick the 3 most interesting & useful rules.
max_Detractor<- is.maximal(goodrules_Detractor) #Find Maximal Itemsets
#inspect(goodrules[max_Detractor])
MostInteresting_Detractor<-(goodrules_Detractor[max_Detractor])
summary(MostInteresting_Detractor)#9 rules

#the NPS recommendation rules:
rules_conf_Detractor <- sort(MostInteresting_Detractor, decreasing = TRUE,by="confidence")
#high-confidence rules
inspect(rules_conf_Detractor)
```

```
rules_lift_Detractor <- sort(MostInteresting_Detractor, decreasing = TRUE,by="lift") # high-lift
rules
inspect(rules_lift_Detractor)

plot(rules_lift_Detractor, method="graph")

#Business_Passive:
#-------------------------------
rules_Passive <-
apriori(ModellingData_Business,parameter=list(support=.1,confidence=0.75),appearance =
                    list(rhs=c("NPS_Type=Passive"),default="lhs"), control=list(verbose=F))
summary(rules_Passive)

#good rules_Passive
goodrules_Passive <- rules_Passive[quality(rules_Passive)$lift > 4]
summary(goodrules_Passive)
plot(goodrules_Passive)

#Pick the 3 most interesting & useful rules.
max_Passive<- is.maximal(goodrules_Passive) #Find Maximal Itemsets
MostInteresting_Passive<-(goodrules_Passive[max_Passive])
summary(MostInteresting_Passive)

#the NPS recommendation rules:
rules_conf_Passive <- sort(MostInteresting_Passive, decreasing = TRUE,by="confidence") #high-
confidence rules
inspect(rules_conf_Passive)
rules_lift_Passive <- sort(MostInteresting_Passive, decreasing = TRUE,by="lift") # high-lift rules
inspect(rules_lift_Passive)

plot(rules_lift_Passive, method="graph")


####################################
#B)Leisure
####################################
str(ModellingData_Leisure)
#----------------------------------
#Leisure_Promoters:
#---------------------------------

#Relation between purpose of visit, city, age range, NPS type and Gender
```

```
L_ruleSetH <-
apriori(ModellingData_Leisure,parameter=list(support=0.14,confidence=0.96),appearance =
                  list(rhs=c("NPS_Type=Promoter"),default="lhs"), control=list(verbose=F))
summary(L_ruleSetH)

#good rules_Promoter
L_goodrules_Promoter <- L_ruleSetH[quality(L_ruleSetH)$lift > 1.4]
summary(L_goodrules_Promoter)
plot(L_goodrules_Promoter)

#Pick the  most interesting & useful rules.
L_max_Promoter<- is.maximal(L_goodrules_Promoter)
L_MostInteresting_Promoter<-(L_goodrules_Promoter[L_max_Promoter])
summary(L_MostInteresting_Promoter)
L_rules_conf_Promoter <- sort(L_MostInteresting_Promoter, decreasing =
TRUE,by="confidence") #high-confidence rules
inspect(L_rules_conf_Promoter)
L_rules_lift_Promoter <- sort(L_MostInteresting_Promoter, decreasing = TRUE,by="lift") # high-
lift rules
inspect(L_rules_lift_Promoter)
plot(L_rules_lift_Promoter, method="graph")



#Leisure_Detractors:
#---------------------------------
L_ruleSetL <-
apriori(ModellingData_Leisure,parameter=list(support=0.0119,confidence=1),appearance =
                  list(rhs=c("NPS_Type=Detractor"),default="lhs"), control=list(verbose=F))
summary(L_ruleSetL)

#good rules_Detractor
L_goodrules_Detractor <- L_ruleSetL[quality(L_ruleSetL)$lift > 6]
summary(L_goodrules_Detractor)
plot(L_goodrules_Detractor)

L_max_Detractor<- is.maximal(L_goodrules_Detractor)
L_MostInteresting_Detractor<-(L_goodrules_Detractor[L_max_Detractor])
summary(L_MostInteresting_Detractor)


L_rules_conf_Detractor <- sort(L_MostInteresting_Detractor, decreasing =
TRUE,by="confidence") #high-confidence rules
inspect(L_rules_conf_Detractor)
```

```
L_rules_lift_Detractor <- sort(L_MostInteresting_Detractor, decreasing = TRUE,by="lift") # high-lift rules
inspect(L_rules_lift_Detractor)

plot(L_rules_lift_Detractor, method="graph")



#####################################
#C)Booking Factors
#####################################

#Business:


str(Modelling_Booking_Factors)
#A)Business_Booking Factors:
#-------------------------------------------------------

Business_Booking_Prom <-
apriori(Modelling_Booking_Factors,parameter=list(support=0.11,confidence=0.65),appearance
=
                        list(rhs=c("NPS_Type=Promoter"),default="lhs"), control=list(verbose=F))
summary(Business_Booking_Prom)

#good rules_Promoter
BK_Bus_goodrules_Promoter <- Business_Booking_Prom[quality(Business_Booking_Prom)$lift
> 1]
summary(BK_Bus_goodrules_Promoter)
plot(BK_Bus_goodrules_Promoter)

#Pick the  most interesting & useful rules.
BK_Bus_max_Promoter<- is.maximal(BK_Bus_goodrules_Promoter)
BK_Bus_MostInteresting_Promoter<-(BK_Bus_goodrules_Promoter[BK_Bus_max_Promoter])
summary(BK_Bus_MostInteresting_Promoter)

BK_rules_conf_Promoter <- sort(BK_Bus_MostInteresting_Promoter, decreasing =
TRUE,by="confidence") #high-confidence rules
inspect(BK_rules_conf_Promoter) #99 # accuracy
BK_rules_lift_Promoter <- sort(BK_Bus_MostInteresting_Promoter, decreasing =
TRUE,by="lift") # high-lift rules
inspect(BK_rules_lift_Promoter)

plot(BK_rules_lift_Promoter, method="graph")
```

```
#-------------------------------------------------------------------------------

Business_Booking_Detr <-
apriori(Modelling_Booking_Factors,parameter=list(support=0.0117,confidence=0.19),appearan
ce =
                        list(rhs=c("NPS_Type=Detractor"),default="lhs"), control=list(verbose=F))
summary(Business_Booking_Detr)

#good rules_Promoter
BK_Bus_goodrules_Detractor <- Business_Booking_Detr[quality(Business_Booking_Detr)$lift >
1]
summary(BK_Bus_goodrules_Detractor)
plot(BK_Bus_goodrules_Detractor)

#Pick the  most interesting & useful rules.
BK_Bus_max_Detractor<- is.maximal(BK_Bus_goodrules_Detractor)
BK_Bus_MostInteresting_detractor<-(BK_Bus_goodrules_Detractor[BK_Bus_max_Detractor])
summary(BK_Bus_MostInteresting_detractor)

#, Insights: Booking_Channel=Digital Channels,Gender_H=Male ;NUM_ROOMS_R=1,and
ADULT_NUM_R=1,
#the NPS recommendation rules:

BK_rules_conf_Detractor <- sort(BK_Bus_MostInteresting_detractor, decreasing =
TRUE,by="confidence") #high-confidence rules
inspect(BK_rules_conf_Detractor) #99 # accuracy
BK_rules_lift_Detractor <- sort(BK_Bus_MostInteresting_detractor, decreasing = TRUE,by="lift")
# high-lift rules
inspect(BK_rules_lift_Detractor)

plot(BK_rules_lift_Detractor, method="graph")

#Booking_Channel=Digital Channels, is most contributing towards detractors.

#, Insights: NUM_ROOMS_R=1,
#CHILDREN_NUM_R=2,
#Likelihood_Recommend_H=2
#Booking_Channel=Global Contact Center,
#Give discounts to family children =2


#-------------------------------------------------------------------------------------------------------
```

```
################################################################################
#C)SVM Modelling for Booking factors:


#A) Splitting into business, leisure

Cal_Regency_modellingSVM<-  subset(Cal_Regency_modelling1,select = -
c(City_PL,CHECK_IN_DATE_C,CHECK_OUT_DATE_C,ENTRY_TIME_R,
                                          NUM_ROOMS_R,LENGTH_OF_STAY_R,Golf_PL,

Overall_Sat_H,Likelihood_Recommend_H,ROOM_TYPE_DESCRIPTION_C))
str(Cal_Regency_modellingSVM)
#Business:
Cal_Regency_modellingSVM_Business<-
Cal_Regency_modellingSVM[Cal_Regency_modellingSVM$POV_CODE_C=="BUSINESS",]
str(Cal_Regency_modellingSVM_Business)
summary(Cal_Regency_modellingSVM_Business)


#A)Business
###############

#Train and Test datasets
random_index<- sample(1:dim(Cal_Regency_modellingSVM_Business)[1])
cutPoint2_3 <- floor(2 * dim(Cal_Regency_modellingSVM_Business)[1]/3) #floor() function
chops off any decimal part of the calculation. #We want to get rid of any decimal because an
index variable needs to be an integer.
cutPoint2_3 #31378

#Test and training sets:
trainData <- Cal_Regency_modellingSVM_Business[random_index[1:cutPoint2_3],] #102 obs
testData<-
Cal_Regency_modellingSVM_Business[random_index[(cutPoint2_3+1):dim(Cal_Regency_model
lingSVM_Business)[1]],] #51 observations
str(trainData)

#Leisure:
Cal_Regency_modellingSVM_Leisure<-
Cal_Regency_modellingSVM[Cal_Regency_modellingSVM$POV_CODE_C=="LEISURE",]
str(Cal_Regency_modellingSVM_Leisure)


#Train and Test datasets
random_index<- sample(1:dim(Cal_Regency_modellingSVM_Leisure)[1])
```

```
cutPoint2_3 <- floor(2 * dim(Cal_Regency_modellingSVM_Leisure)[1]/3) #floor() function chops off any decimal part of the calculation. #We want to get rid of any decimal because an index variable needs to be an integer.
cutPoint2_3 #31378

#Test and training sets:
trainData <- Cal_Regency_modellingSVM_Leisure[random_index[1:cutPoint2_3],]
testData<-
Cal_Regency_modellingSVM_Leisure[random_index[(cutPoint2_3+1):dim(Cal_Regency_modellingSVM_Leisure)[1]],]
str(trainData)


#SVM Modelling  to prove LM results
#--------

  svm_model<- svm(NPS_Type~Condition_Hotel_H,data=trainData,C=5)
  svm_Predicted<- predict(svm_model,testData,type="responses")
  comparison_Table_SVM <- data.frame(testData[["NPS_Type"]],svm_Predicted)
  colnames(comparison_Table_SVM) <- c('Test_NPS','Predicted_NPS')
  confusion_matrix_svm<- table(comparison_Table_SVM)
  print(confusion_matrix_svm)
  Accuracy_svm <-
((confusion_matrix_svm[1,1]+confusion_matrix_svm[2,2]+confusion_matrix_svm[3,3])/nrow(comparison_Table_SVM))*10
  Accuracy_svm

  svm_model<- svm(NPS_Type~Guest_Room_H,data=trainData,C=5)
  svm_Predicted<- predict(svm_model,testData,type="responses")
  comparison_Table_SVM <- data.frame(testData[["NPS_Type"]],svm_Predicted)
  colnames(comparison_Table_SVM) <- c('Test_NPS','Predicted_NPS')
  confusion_matrix_svm<- table(comparison_Table_SVM)
  print(confusion_matrix_svm)
  Accuracy_svm <-
((confusion_matrix_svm[1,1]+confusion_matrix_svm[2,2]+confusion_matrix_svm[3,3])/nrow(comparison_Table_SVM))*100
  Accuracy_svm

  svm_model<- svm(NPS_Type~Customer_SVC_H,data=trainData,C=5)
  svm_Predicted<- predict(svm_model,testData,type="responses")
  comparison_Table_SVM <- data.frame(testData[["NPS_Type"]],svm_Predicted)
  colnames(comparison_Table_SVM) <- c('Test_NPS','Predicted_NPS')
  confusion_matrix_svm<- table(comparison_Table_SVM)
  print(confusion_matrix_svm)
```

```
  Accuracy_svm <-
((confusion_matrix_svm[1,1]+confusion_matrix_svm[2,2]+confusion_matrix_svm[3,3])/nrow(c
omparison_Table_SVM))*100
  Accuracy_svm

# ##BORUTA !----------------------Commented as it takes time. Please uncomment to run this-------
-----------------------------------
# dim(Cal_Regency_modellingSVM_Business)
# random_index<- sample(1:dim(Cal_Regency_modellingSVM_Business)[1])
# cutPoint2_3 <- floor(2 * dim(Cal_Regency_modellingSVM_Business)[1]/3)
# cutPoint2_3
#
# #Test and training sets:
# trainData <- Cal_Regency_modellingSVM_Business[random_index[1:cutPoint2_3],] #102 obs
# testData<-
Cal_Regency_modellingSVM_Business[random_index[(cutPoint2_3+1):dim(Cal_Regency_model
lingSVM_Business)[1]],] #51 observations
#
# boruta.train <- Boruta(NPS_Type~., data = trainData, doTrace = 2)
# print(boruta.train)
#
# plot(boruta.train, xlab = "", xaxt = "n")
# lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)
#   boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
# names(lz) <- colnames(boruta.train$ImpHistory)
# Labels <- sort(sapply(lz,median))
# axis(side = 1,las=2,labels = names(Labels),
#      at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.6)
#
# final.boruta <- TentativeRoughFix(boruta.train)
#
# getSelectedAttributes(final.boruta, withTentative = F)
# boruta.df <- attStats(final.boruta)


#-----------------------------------------------------end-------------------------------------
```