HYATT HOTEL REVIEWS

DATA ANALYSIS

IST 687: Applied Data Science

Group Members:

Niti Saluja

Kathleen Flynn

Ananya Bhupathipalli

Project Index

Introduction

When a company considers its reputation, it must determine how that reputation is created. Since a company's image is largely due to word of mouth of customers, Fred Reichheld devised a system to calculate what he termed as a Net Promoter Score® (NPS®) (Bain & Company, 2017). This system divides customers into three groups: Promoters, Passives, and Detractors (Bain & Company, 2017). These groups are assigned following a survey asking how likely the customer is to recommend the company to someone else on a scale from zero to ten (Bain & Company, 2017). Those who give a score of zero to six are Detractors, a score of seven or eight are from Passives, and a nine or ten indicate Promoters (Bain & Company, 2017). The NPS® is calculated by subtracting the percentage of Detractors from the percentage of Promoters (Bain & Company, 2017). Therefore, the score can range from -100 to +100 with a score above zero demonstrating the company has more Promoters than Detractors. This score and the related survey questions can be beneficial in that they provide the company with possible explanations for why a customer likes or dislikes their business. With this information, the company can be prepared to continue satisfying their Promoters and win over the Passives or Detractors.

Originally based in the United States, Hyatt Hotels Corporation is the creator of several brands of hotels located worldwide (Hyatt Hotels Corporation, 2017). For this project we were provided with one year of data from Hyatt Hotels worldwide. Using R as a tool, our goal was to analyze this data until we could identify factors that affect Hyatt Hotel NPS® and suggest possible actions the company can take to improve it.

Data Selection

Prior to downloading the data, it became apparent that the full 12 months would consist of over 12 GB of data. Using this data set would result in slow processing time during the analysis. Therefore, we selected a subset of data for the analysis.

Initially, we chose a subset of hotels from four different countries and from four different months. These countries were: United Kingdom, Malaysia, Morocco, and Switzerland. The four months were February, May, August, and December. We chose those countries after sorting the data by country to identify the number of rows per country. This was due to our wish to have around 50,000 rows of data total, which we decided was a reasonable amount that would not cause slow processing speeds. We were also interested in countries with diverse climates and economies. Similarly, when choosing the four months we selected for months from different seasons as well as months where people are more likely to be on vacation from work or school. Unfortunately, after cleaning the data we were left with slightly more than 2,000 rows of data. After spending some time cleaning and analyzing the data using different modeling techniques, we concluded that we would achieve better results from a larger subset. Therefore, we began searching for a more substantial subset.

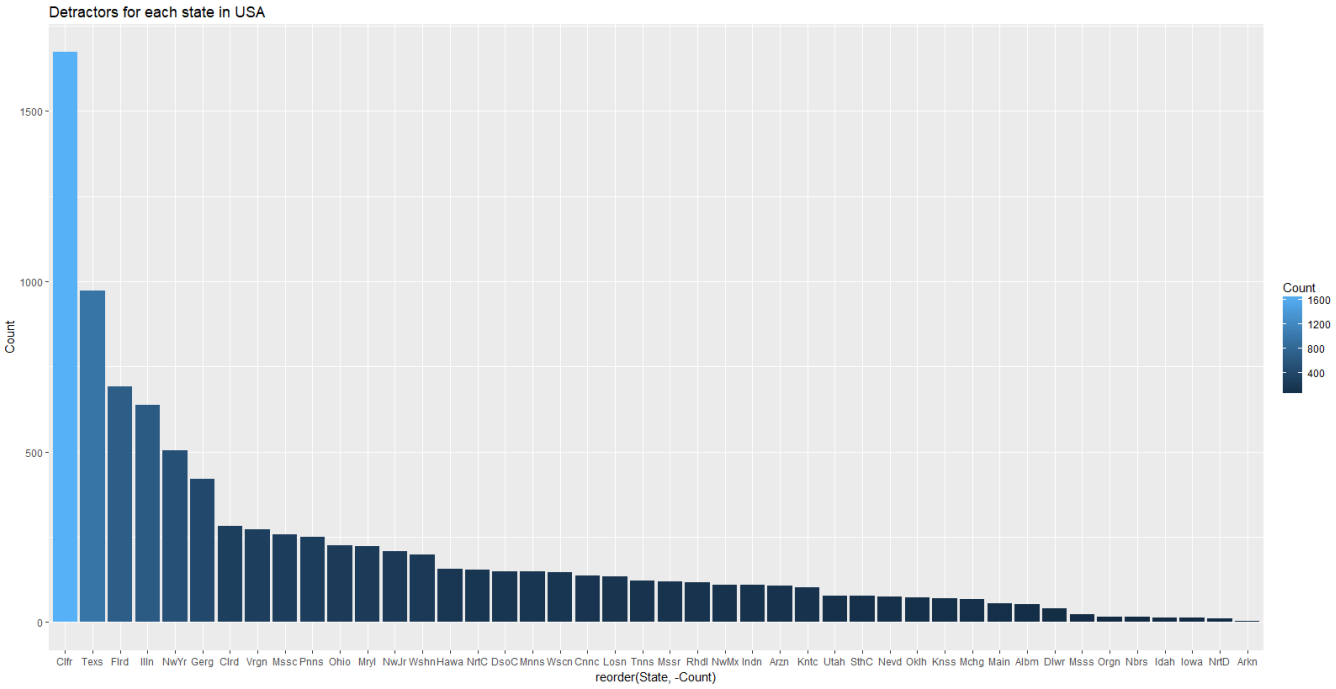
We settled on data from the United States. This was partly because more data was available compared to other countries. Another factor was since Hyatt Hotels originated in the United States, the hotels there were more likely to have all amenity features offered by Hyatt. Furthermore, more Hyatt hotel brands are likely to be represented in the United States.

Length of data in each of the countries:

United States	1058803
China	50464
India	33475
Japan	28072
Canada	25656
Australia	18963
South Korea	18890

We then decided to narrow our subset down to the state of California because it is a large state with not only a large population, but a sizable number of tourists and business travelers during the year. Also, we discovered that California was the country with the highest number of Detractors as shown below. Therefore, we could potentially have more significant observations from the analysis. Finally, we limited the data to the month of August because as a summer month was more likely to result in guests on vacation, especially families with children. Before data cleaning this subset consisted of 173,790 rows.

State Wise Detractors Count



Data Preparation

Using the metadata provided to us, we determined which columns we would include in our analysis and which could be removed as shown below. Once the data subset of the month of August was loaded as a csv file into R, we created a new data frame to store only the state of California's data. We then only included the columns were deemed useful. Next, we replaced any blank values with an 'NA'. And then we removed all rows which had an NA value for the column 'NPS_Type'. This was to ensure we were only left with rows from those customers who had completed the survey for NPS®. Following these actions, we were left with 11,526 rows of data.

Our next task was to observe the structure of the subset to identify the data types for each column. Certain column values were set as factors if they had character strings. These columns needed to contain factors to complete association rule mining and use support vector machine models. Thankfully, there were no issues with commas or odd spaces in the data that needed to be removed. Next, we noted the columns, notably the NPS® survey columns, that had NA values which required action. We experimented with removing the NAs and concluded that it would be better to replace them with the means of their respective columns. This was done partly because eliminating all rows with NA values sometimes resulted in a loss of all useable data. Occasionally, a guest would fail to assign a score to every column in the survey. Therefore, we could have been removing a row which had one NA value, but eight genuine values.

Columns selected to be included in the analysis:

1	CHECKOUT_HEADER_ID_C	24	Condition_Hotel_H	47	Dry-Cleaning_PL
2	CONS_GUEST_ID_C	25	Customer_SVC_H	48	Elevators_PL
3	ROOM_TYPE_CODE_C	26	Staff_Cared_H	49	Fitness Center_PL
4	ROOM_TYPE_DESCRIPTION_C	27	Internet_Sat_H	50	Fitness Trainer_PL
5	WALK_IN_FLG_C	28	Check_In_H	51	Golf_PL
6	CHECK_IN_DATE_C	29	F&B_FREQ_H	52	Indoor Corridors_PL
7	CHECK_OUT_DATE_C	30	F&B_Overall_Experience_H	53	Laundry_PL
8	LENGTH_OF_STAY_C	31	average_daily_rate_CC	54	Limo Service_PL
9	ADULT_NUM_C	32	Hotel Name-Long_PL	55	Mini-Bar_PL
10	CHILDREN_NUM_C	33	City_PL	56	Pool-Indoor_PL
11	PMS_ROOM_REV_USD_C	34	State_PL	57	Pool-Outdoor_PL
12	POV_CODE_C	35	Country_PL	58	Regency Grand Club_PL
13	ENTRY_HOTEL_CODE_R	36	Guest NPS Goal_PL	59	Resort_PL
14	REVENUE_USD_R	37	Brand_PL	60	Restaurant_PL
15	Survey_ID_H	38	Total Meeting Space_PL	61	Self-Parking_PL
16	Age_Range_H	39	Relationship_PL	62	Shuttle Service_PL
17	Clublounge_Used_H	40	All Suites_PL	63	Ski_PL
18	Spa_Used_H	41	Bell Staff_PL	64	Spa_PL
19	GP_Tier_H	42	Boutique_PL	65	Spa services in fitness center_PL
20	Likelihood_Recommend_H	43	Business Center_PL	66	Spa online booking_PL
21	Overall_Sat_H	44	Casino_PL	67	Spa F&B offering_PL
22	Guest_Room_H	45	Conference_PL	68	Valet Parking_PL
23	Tranquility_H	46	Convention_PL	69	NPS_Type

Business Questions

1. What factors should be considered while choosing the subset for analysis to achieve overall high NPS score?
2. What is the effect of NPS on the Hotel Brand and Purpose of Visit?
3. Which age group guests are most likely to recommend and who are least likely to recommend?
4. On what factors do the 'Likelihood to Recommend' and 'NPS Type' have the highest dependency?
5. What are the recommendations we can provide to improve the NPS score of Hyatt?

Business Question:

- What are the recommendations we can provide to improve the NPS score of Hyatt?

Actionable Insights

1. Focus on improving Hotel Condition, Guest Room Condition, and Customer Service.
    a. Hotel Condition, Guest Room Condition, Customer Service Satisfaction, Tranquility, and Age Range are the most influential factors affecting NPS. As shown later in the report, this was demonstrated in linear modeling, association rules, and descriptive statistics.

2. Improve the rooms 'Guest Room King' and 'Guest Room Double,' primarily in general Hyatt brand hotels.
    a. The general Hyatt Brand hotels have the lowest likelihood to recommend scores among the brands. Also, the room types 'Guest Room King' and 'Guest Room Double' are the most used rooms, but have lower likelihood to recommend and guest room condition scores. This is true in all brands and in the general Hyatt brand. This was shown using descriptive statistics, as seen later in the report. Therefore, actions should be taken to improve the room conditions of those two rooms in all brands, but especially in the general Hyatt brand as it is in more need of improvement.

3. Center the hotel promotional activities towards the age groups of 26-35, 36-45, and 46-55.
    a. Guests in the age ranges of 26-35, 36-45, and 46-55 are less likely to recommend, while those who are in the youngest and oldest age ranges are significantly more satisfied. These observations were found in association rules, linear modeling, and descriptive statistics found later in the report.

4. Increase the number of hotels which have Mini Bars.
    a. The majority of hotels in our subset lacked a Mini Bar. Since the presence of a Mini Bar results in a higher Hotel Condition score, and Hotel Condition has a significant impact on NPS, Hyatt should work on adding a Mini Bar to more hotels. The impact of mini bars was shown using descriptive statistics that appear later in the report.

Descriptive Statistics

- *NPS®*

We began by calculating the NPS® for the entire subset, which we created functions to accomplish. The result was a score of 51, which can be considered good since a score well above zero means there are significantly more Promoters than Detractors. For validation purposes we created more than one function to calculate the score. The results were the same.

Calculating NPS® for the entire subset.

```
> NPS <- function(vector){
+   n <- as.numeric(vector)
+   passives1 <- length(which(n==7))
+   passives2 <- length(which(n==8))
+   passives3 <- passives1 + passives2
+   p <- length(which(n>=9))
+   d <- length(which(n<=6))
+   total <- p + d + passives3
+   nps1 <- p-d
+   NPS1 <- nps1/total
+   perc <- NPS1 * 100
+   return(perc)
+ }
> #Use NPS function to calculate the NPS of the entire subset using the Likelihood_Recommend_H column.
> ab <- NPS(subset$Likelihood_Recommend_H)
> ab
[1] 51.04112
```

- ***Question***: Does the type of room affect the likelihood to recommend?

Method: Using tapply, we calculated the mean, min, max, and median likelihood to recommend scores for each room type from the ROOM_TYPE_DESCRIPTION_C column. We also calculated how many guests stayed in each room type. We combined all of those in a data frame to observe.


```
#Use tapply to see relationships between room type description and likelihood to recommend.
RoomTypeDMean <- tapply(subset$Likelihood_Recommend_H, subset$ROOM_TYPE_DESCRIPTION_C, mean)
RoomTypeDMin <- tapply(subset$Likelihood_Recommend_H, subset$ROOM_TYPE_DESCRIPTION_C, min)
RoomTypeDMax <- tapply(subset$Likelihood_Recommend_H, subset$ROOM_TYPE_DESCRIPTION_C, max)
RoomTypeDMedian <- tapply(subset$Likelihood_Recommend_H, subset$ROOM_TYPE_DESCRIPTION_C, median)
RTypeDLike <- data.frame(RoomTypeDMean, RoomTypeDMin, RoomTypeDMax, RoomTypeDMedian)
View(RTypeDLike)

#Use tapply to see how many people used each of the room types.
RoomTypeCount <- tapply(subset$ROOM_TYPE_DESCRIPTION_C, subset$ROOM_TYPE_DESCRIPTION_C, length)
RoomTypeCount

RoomTypeCountdf <- data.frame(ROOM_TYPE_DESCRIPTION_C=names(RoomTypeCount),length=RoomTypeCount)
View(RoomTypeCountdf)

#Combine the two to see ratings per room type and how many stayed in each.
RoomRatings <- data.frame(RoomTypeCountdf, RTypeDLike)
View(RoomRatings)
```

***Analysis*:** Most people stayed in the rooms called 'Guest Room King' and 'Guest Room Double'. However, they had lower means for likelihood to recommend than other room types as shown below.

Likelihood to recommend score by room type.

ROOM_TYPE_DESCRIPTION_C	length	RoomTypeDMean	RoomTypeDMin	RoomTypeDMax	RoomTypeDMedian
Guest Room King	1545	8.232362	1	10	9.0
Guest Room Double	1023	8.286413	1	10	9.0
Guest Room King Bed	714	8.745098	1	10	9.0
1 Bedroom King	655	8.897710	1	10	10.0
2 Queen Beds	310	8.906452	1	10	10.0
Guest Room 2 Double Bed	286	8.625874	1	10	9.0
Guest Room Queen/Queen	232	7.956897	1	10	9.0
Studio King	175	9.091429	1	10	10.0
Pool View King	167	8.574850	1	10	9.0
Deluxe King	154	8.532468	1	10	9.0
Club King	146	8.691781	1	10	9.5
View King	141	8.815603	1	10	9.0
High Floor King	135	8.288889	1	10	9.0
Bayview Balcony King	106	8.773585	4	10	9.0
ADA King Tub	103	8.495146	1	10	9.0
Guest Room Double/Double	100	7.570000	1	10	8.0
King Bed	98	8.530612	1	10	9.0
1 Bedroom King Bay View	97	9.072165	2	10	10.0
City View King	91	8.791209	1	10	9.0
Golf View King	86	8.232558	1	10	9.0
Deluxe View Queen/Queen	82	7.560976	1	10	8.0
Park King	79	9.050633	1	10	10.0
View Double	79	9.177215	3	10	10.0
Ocean View King	70	8.385714	2	10	9.0

- ***Question***: Does the type of room affect the guest room condition score?

***Method*:** Using tapply, we calculated the mean, min, max, and median guest room condition scores for each room type from the ROOM_TYPE_DESCRIPTION_C column. We also calculated how many guests stayed in each room type. We combined all of those in a data frame to observe.

```
#Look at Room condition rating and room type.
RoomConMean <- tapply(subset$Guest_Room_H, subset$ROOM_TYPE_DESCRIPTION_C, mean)
RoomConMin <- tapply(subset$Guest_Room_H, subset$ROOM_TYPE_DESCRIPTION_C, min)
RoomConMax <- tapply(subset$Guest_Room_H, subset$ROOM_TYPE_DESCRIPTION_C, max)
RoomConMedian <- tapply(subset$Guest_Room_H, subset$ROOM_TYPE_DESCRIPTION_C, median)

RTypeConLike <- data.frame(RoomConMean, RoomConMin, RoomConMax, RoomConMedian)

#Combine the two to see ratings per room type and how many stayed in each.
RoomConRatings <- data.frame(RoomTypeCountdf, RTypeConLike)
View(RoomConRatings)
```

***Analysis*:** Again, most people stayed in the rooms called ‘Guest Room King’ and ‘Guest Room Double’. However, they also had lower means for guest room condition than other room types, as shown below. Furthermore, the mean likelihood to recommend and guest room condition for those rooms were similar. This suggests that the condition of those rooms may be lowering the likelihood to recommend for guests who stay there.

Guest room condition by room type.

ROOM_TYPE_DESCRIPTION_C	length	RoomConMean	RoomConMin	RoomConMax	RoomConMedian
Guest Room King	1545	8.394732	1.00000	10	9.00000
Guest Room Double	1023	8.250042	1.00000	10	9.00000
Guest Room King Bed	714	8.899304	1.00000	10	9.00000
1 Bedroom King	655	8.854383	1.00000	10	9.00000
2 Queen Beds	310	8.998443	1.00000	10	10.00000
Guest Room 2 Double Bed	286	8.895346	2.00000	10	9.00000
Guest Room Queen/Queen	232	8.024078	1.00000	10	9.00000
Studio King	175	8.803152	1.00000	10	9.00000
Pool View King	167	8.760685	3.00000	10	9.00000
Deluxe King	154	8.626959	2.00000	10	9.00000
Club King	146	8.791214	1.00000	10	9.00000
View King	141	8.921986	1.00000	10	10.00000
High Floor King	135	8.185440	1.00000	10	9.00000
Bayview Balcony King	106	8.886792	2.00000	10	9.00000
ADA King Tub	103	8.349849	1.00000	10	9.00000
Guest Room Double/Double	100	8.065517	1.00000	10	9.00000
King Bed	98	8.847290	3.00000	10	9.00000
1 Bedroom King Bay View	97	9.293992	3.00000	10	10.00000
City View King	91	8.791209	2.00000	10	9.00000
Golf View King	86	8.064154	1.00000	10	9.00000
Deluxe View Queen/Queen	82	7.621951	1.00000	10	8.00000
Park King	79	9.075949	1.00000	10	10.00000
View Double	79	9.240942	5.00000	10	10.00000
Ocean View King	70	8.400000	2.00000	10	9.00000

- ***Question***: Does the type of room affect the hotel condition score?

Method: Using tapply, we calculated the mean, min, max, and median hotel condition scores for each room type from the ROOM_TYPE_DESCRIPTION_C column. We also calculated how many guests stayed in each room type. We combined all of those in a data frame to observe.

```
#Look at Hotel condition rating and room type.
HotelConMean <- tapply(subset$Condition_Hotel_H, subset$ROOM_TYPE_DESCRIPTION_C, mean)
HotelConMin <- tapply(subset$Condition_Hotel_H, subset$ROOM_TYPE_DESCRIPTION_C, min)
HotelConMax <- tapply(subset$Condition_Hotel_H, subset$ROOM_TYPE_DESCRIPTION_C, max)
HotelConMedian <- tapply(subset$Condition_Hotel_H, subset$ROOM_TYPE_DESCRIPTION_C, median)

RTypeHConLike <- data.frame(HotelConMean, HotelConMin, HotelConMax, HotelConMedian)

#Combine the two to see ratings per room type and how many stayed in each.
RoomHConRatings <- data.frame(RoomTypeCountdf, RTypeHConLike)
View(RoomHConRatings)
```

Analysis: Like the previous two tests, most people stayed in the rooms called ‘Guest Room King’ and ‘Guest Room Double’. However, they also had lower means for hotel condition than other room types, as shown below. Furthermore, the mean likelihood to recommend and guest room condition for those

rooms were similar. This suggests that the condition of those rooms may be lowering the likelihood to recommend for guests who stay there.

Room type effect on Hotel Condition

ROOM_TYPE_DESCRIPTION_C	length	HotelConMean	HotelConMin	HotelConMax	HotelConMedian
Guest Room King	1545	8.624081	1	10	9.000000
Guest Room Double	1023	8.439589	1	10	9.000000
Guest Room King Bed	714	9.021897	1	10	9.000000
1 Bedroom King	655	8.887893	1	10	9.000000
2 Queen Beds	310	9.210094	2	10	10.000000
Guest Room 2 Double Bed	286	8.952365	1	10	10.000000
Guest Room Queen/Queen	232	8.419726	1	10	9.000000
Studio King	175	9.023226	1	10	9.000000
Pool View King	167	8.870517	2	10	9.000000
Deluxe King	154	8.788578	1	10	10.000000
Club King	146	9.002578	1	10	10.000000
View King	141	9.113933	3	10	10.000000
High Floor King	135	8.462047	1	10	9.000000
Bayview Balcony King	106	8.965815	2	10	9.000000

Business Question

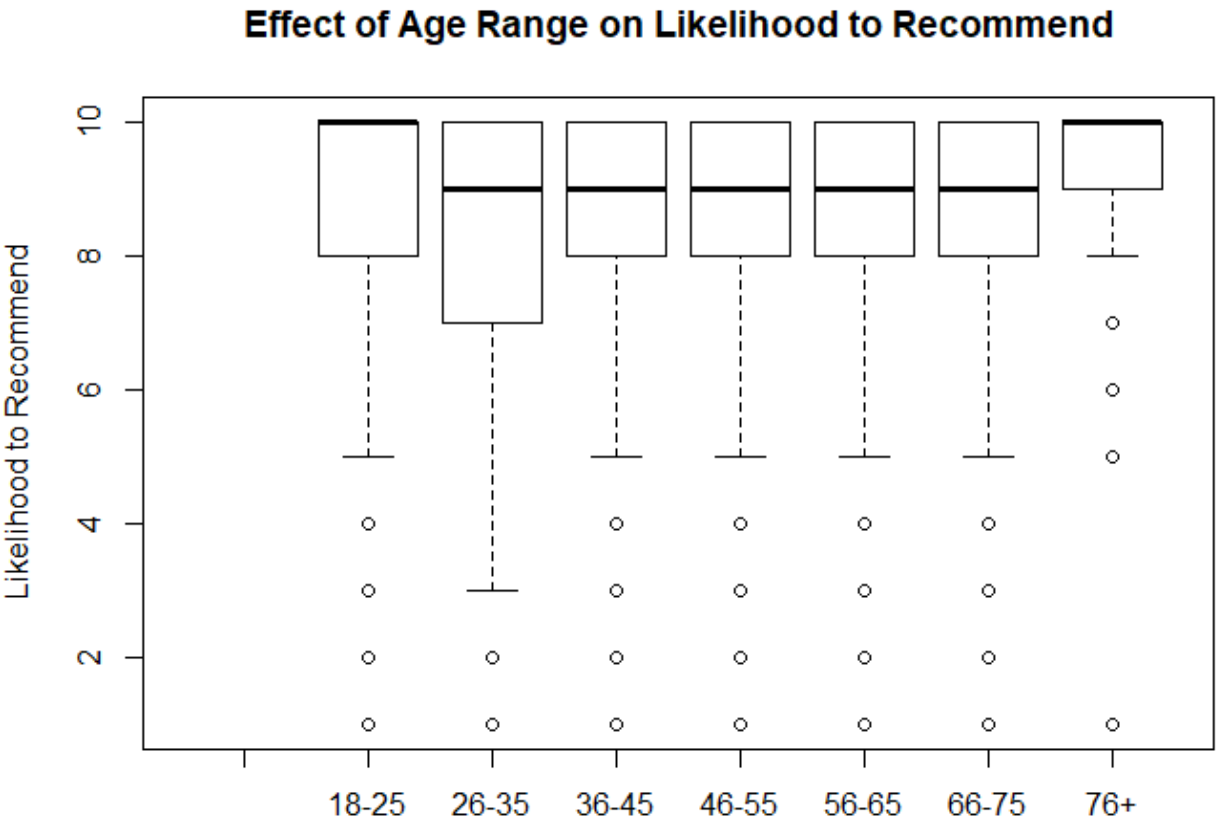
- Which age group guests are most likely to recommend and who are least likely to recommend?

***Method*:** Using tapply, we calculated the mean, min, max, and median likelihood to recommend scores for each age range. We also calculated how many guests were in each age range.

***Analysis*:** Guests in the youngest and oldest age ranges were considerably more likely to recommend than middle aged guests, as shown in below. The number of guests in the youngest and oldest age ranges were lower, but we could still observe that likelihood to recommend increases as age increases. This was possible because the amount of data for the middle age ranges was plentiful.

```
> tapply(subset$Likelihood_Recommend_H, subset$Age_Range_H, mean)
           18-25    26-35    36-45    46-55    56-65    66-75      76+
      NA 8.665138 8.277124 8.384921 8.417271 8.610707 8.710953 9.184713
```

Effect of age range on likelihood to recommend.



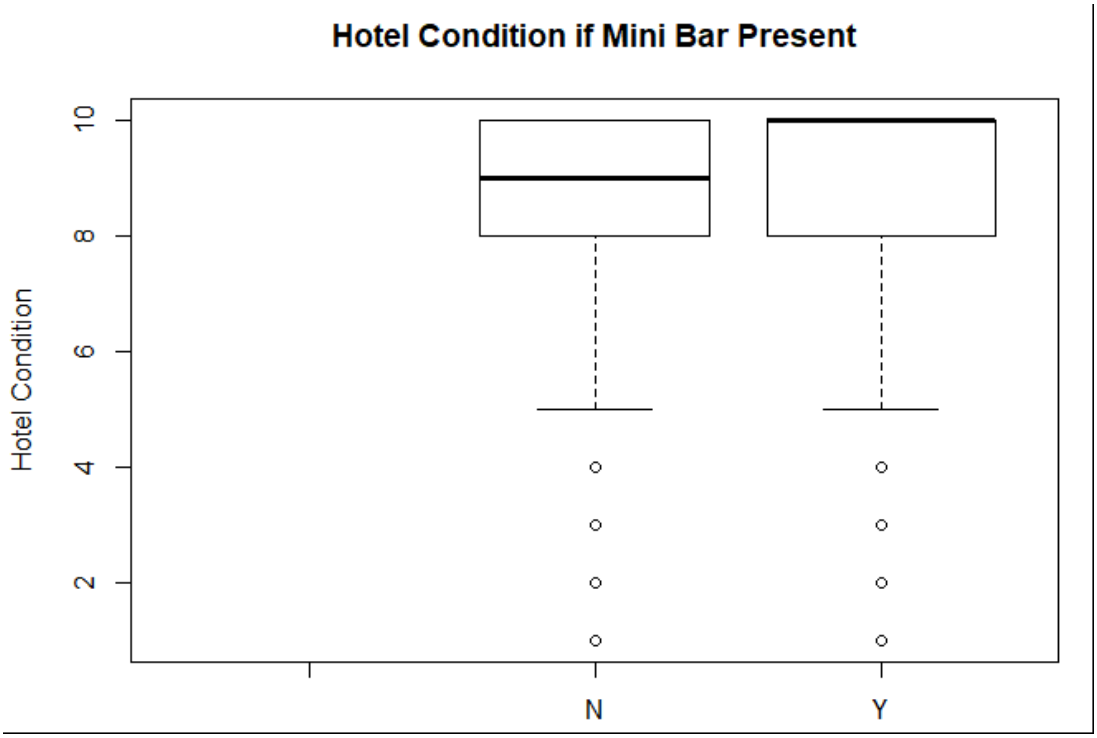
- ***Question***: Does the presence of a Mini Bar affect the hotel condition score?

Method: Using tapply, we calculated the mean hotel condition scores for guests whose hotels had mini bars and those who did not have them. We also calculated how many guests were in each situation.

Analysis: The mean hotel condition score was higher for those who had access to a mini bar than those who did not, as shown below.

```
> with(subset, plot(Mini.Bar_PL, Condition_Hotel_H))+ title(main="Hotel Condition if Mini Bar Present",
ylab="Hotel Condition")
```

Hotel condition based on if mini bar is present.



- ***Question***: Does using the Club lounge affect the likelihood to recommend score for those in the age ranges of 26-35 and 36-45?

***Method*:** We created a new data frame with only guests in those age ranges. Then, using tapply, we calculated the mean likelihood to recommend scores for guests who used the club lounge and those who did not. We also calculated how many guests were in each situation.

***Analysis*:** The mean likelihood to recommend score was higher for those who had used the Club lounge than those who did not.

```
> mid <- sqldf("select * from subset where Age_Range_H='26-35' and '36-45'")
> tapply(mid$Likelihood_Recommend_H, mid$Clublounge_Used_H, mean)
I don't know           No          Yes
    7.833333     8.098712     9.153846
```

- ***Question*:** Does using the Club lounge affect the likelihood to recommend score of all guests?

***Method*:** Using tapply, we calculated the mean likelihood to recommend scores for guests who used the club lounge and those who did not. We also calculated how many guests were in each situation.

***Analysis*:** The mean likelihood to recommend score was higher for those who had used the Club lounge than those who did not.

```
> l <- as.factor(subset$Clublounge_Used_H)
> tapply(subset$Likelihood_Recommend_H, l, mean)
        I don't know           No          Yes
          8.549451     8.407499     8.610973
```

- ***Question***: Is the guest room condition score affected by room type in Hotels of the general Hyatt brand.

***Method*:** Using tapply, we calculated the mean, min, max, and median guest room condition

scores for each room type from the ROOM_TYPE_DESCRIPTION_C column. We also calculated how many guests stayed in each room type. We combined all of those in a data frame to observe.

***Analysis*:** The guest room condition scores were lower for those in the rooms 'Guest Room King' and 'Guest Room Double'. These were also the most frequented room types for this brand. These results match those found in the whole subset for those room types.

Guest room condition score by room type for hotel brand: Hyatt.

ROOM_TYPE_DESCRIPTION_C	length	RoomConMean	RoomConMin	RoomConMax	RoomConMedian
Guest Room King	209	7.780234	1.00000	10	8.51722
Guest Room Double	87	7.649623	2.00000	10	8.00000
Premier Pool View 1 King	63	8.833607	3.00000	10	9.00000
Deluxe King	58	9.215814	4.00000	10	10.00000
Guest Room Queen/Queen	42	8.261905	2.00000	10	9.00000
Premier Pool View 2 Doubles	39	7.961980	1.00000	10	9.00000
Deluxe Room 1 King	32	8.562500	1.00000	10	9.00000
Guest Room 2 Queen Beds	32	7.531250	2.00000	10	8.00000
Premier Mountain View 1 King	28	8.964286	5.00000	10	9.50000
Balcony King	25	8.580689	6.00000	10	9.00000
Deluxe Room 2 Doubles	23	8.086957	3.00000	10	8.00000
Ocean View 2 Queen	23	8.173913	3.00000	10	9.00000
Balcony Double	21	7.285714	1.00000	10	8.00000

Modeling

Business Question:

- On what factors do the 'Likelihood to Recommend' and 'NPS Type' have the highest dependency?

Linear Modeling:

The aim of linear regression is to model a continuous variable *Y* as a mathematical function of one or more *X* variables, to predict the values of Y when only X value is known. Linear Regression is also a part of of supervised learning family and provides more accurate results for quantitative response.

Using the California subset, we have selected the NPS Survey and Amenities columns both quantitative and categorical data to perform linear regression. To accomplish this we considered Likelihood_Recommend_H as the dependent variable and different combinations of Amenities and survey columns as independent variables as they contribute towards the prediction of the dependent variable. We have tested the model using different combinations of dependent and independent variables and selected our best model as the one which has highest r squared value.

Analysis:

Among the 40 linear models created for linear regression, the combination of MEMBER_STATUS_R + Fitness.Center_PL + Staff_Cared_H + Condition_Hotel_H + Customer_SVC_H + Guest_Room_H + Tranquility_H as independent variable and Likelihood_Recommend_H as dependent variable proved to be the best one fetching the r-squared value of 71.02%


```
Call:
lm(formula = Likelihood_Recommend_H ~ MEMBER_STATUS_R + Fitness.Center_PL +
    Staff_Cared_H + Condition_Hotel_H + Customer_SVC_H + Guest_Room_H +
    Tranquility_H, data = Cali_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9278 -0.1796  0.1215  0.4847  5.1994

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.56520    0.35231  -7.281 3.74e-13 ***
MEMBER_STATUS_RDiamond           0.61345    0.30893   1.986   0.0471 *
MEMBER_STATUS_RGold              0.47790    0.30424   1.571   0.1163
MEMBER_STATUS_RLifetime Diamond  0.85296    0.59212   1.441   0.1498
MEMBER_STATUS_RPlatinum          0.55747    0.30529   1.826   0.0679 .
Fitness.Center_PLY               0.08604    0.12882   0.668   0.5043
Staff_Cared_H                    0.09493    0.01600   5.934 3.13e-09 ***
Condition_Hotel_H                0.31237    0.01333  23.438  < 2e-16 ***
Customer_SVC_H                   0.38975    0.01235  31.546  < 2e-16 ***
Guest_Room_H                     0.33299    0.01215  27.410  < 2e-16 ***
Tranquility_H                    0.07157    0.01254   5.706 1.21e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.136 on 5922 degrees of freedom
  (5593 observations deleted due to missingness)
Multiple R-squared:  0.7107,    Adjusted R-squared:  0.7102
F-statistic:  1455 on 10 and 5922 DF,  p-value: < 2.2e-16
```

Some of the other combinations of variables which fetched best results for linear regressions are below:

1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.8759 -0.1762  0.1241  0.4636  4.9750

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.928197   0.077261  -24.96   <2e-16 ***
Condition_Hotel_H  0.307807   0.009974   30.86   <2e-16 ***
Customer_SVC_H     0.428235   0.008422   50.84   <2e-16 ***
Guest_Room_H       0.339516   0.009257   36.68   <2e-16 ***
Tranquility_H      0.116507   0.008365   13.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.156 on 11345 degrees of freedom
  (176 observations deleted due to missingness)
Multiple R-squared:  0.6983,    Adjusted R-squared:  0.6982
F-statistic:  6564 on 4 and 11345 DF,  p-value: < 2.2e-16

>
```

2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.8759 -0.1762  0.1241  0.4636  4.9750

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.928197   0.077261  -24.96   <2e-16 ***
```

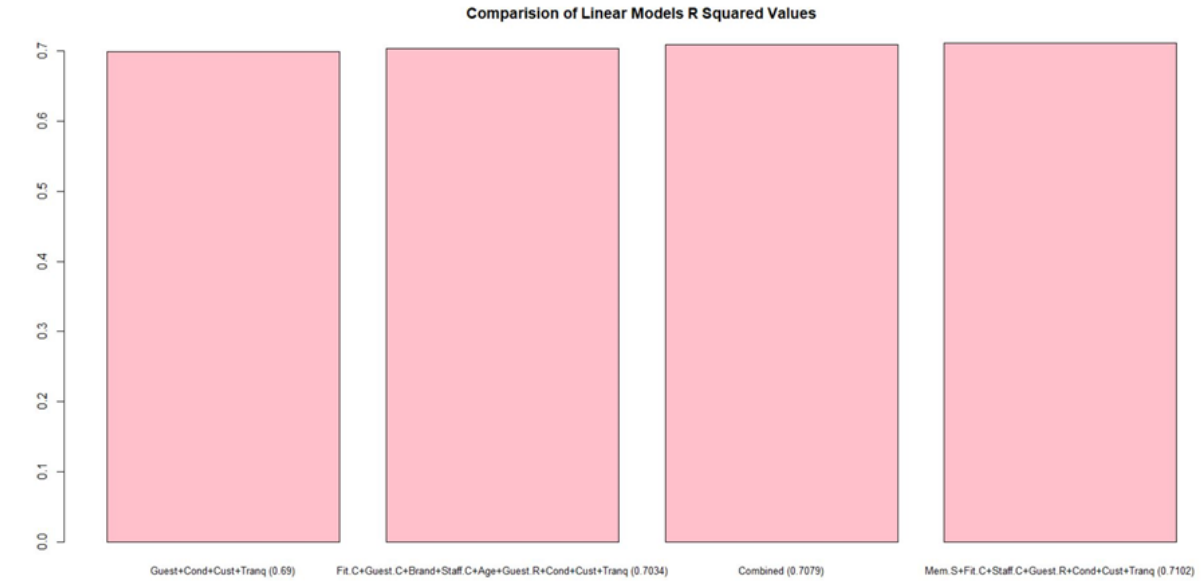
```
Residual standard error: 1.142 on 8803 degrees of freedom
  (2648 observations deleted due to missingness)
Multiple R-squared:  0.7074,    Adjusted R-squared:  0.7049
F-statistic: 287.6 on 74 and 8803 DF,  p-value: < 2.2e-16
```

3)

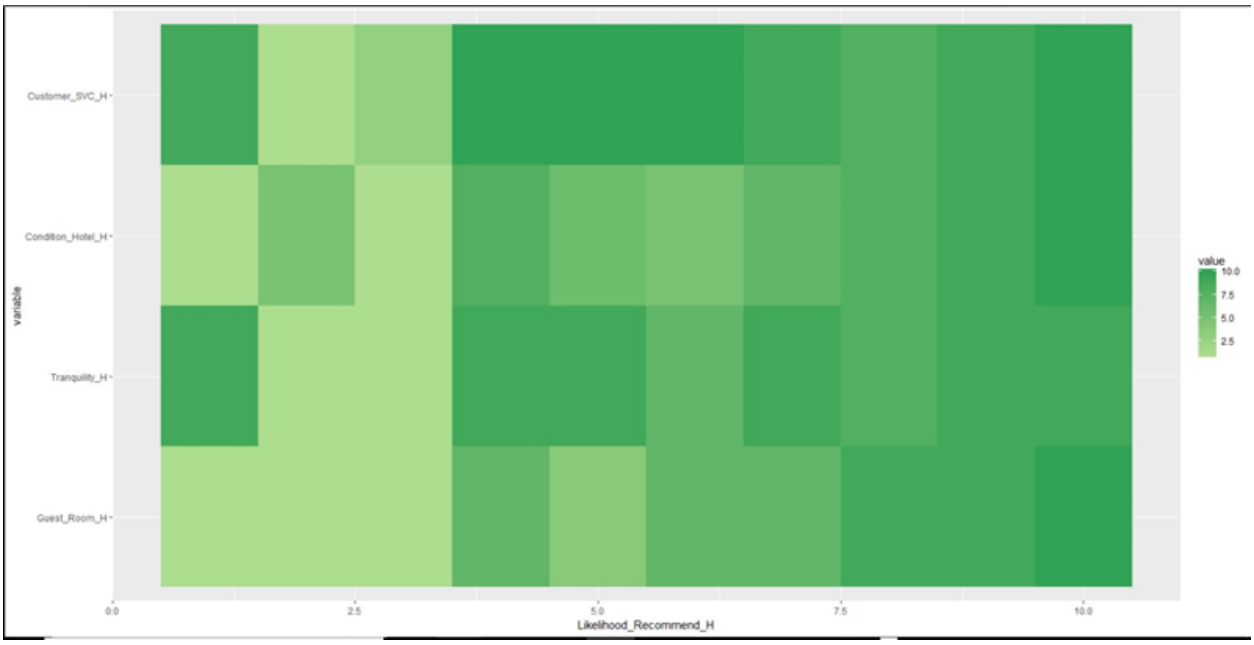
```
Call:
lm(formula = Likelihood_Recommend_H ~ MEMBER_STATUS_R + Fitness.Center_PL +
    GUEST_COUNTRY_R + Brand_PL + Staff_Cared_H + Age_Range_H +
    Condition_Hotel_H + Customer_SVC_H + Guest_Room_H + Tranquility_H,
    data = Cali_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7446 -0.1998  0.0989  0.4884  5.0905
```

```
Residual standard error: 1.134 on 5742 degrees of freedom
  (5717 observations deleted due to missingness)
Multiple R-squared:  0.7116,    Adjusted R-squared:  0.7083
F-statistic: 214.6 on 66 and 5742 DF,  p-value: < 2.2e-16
```

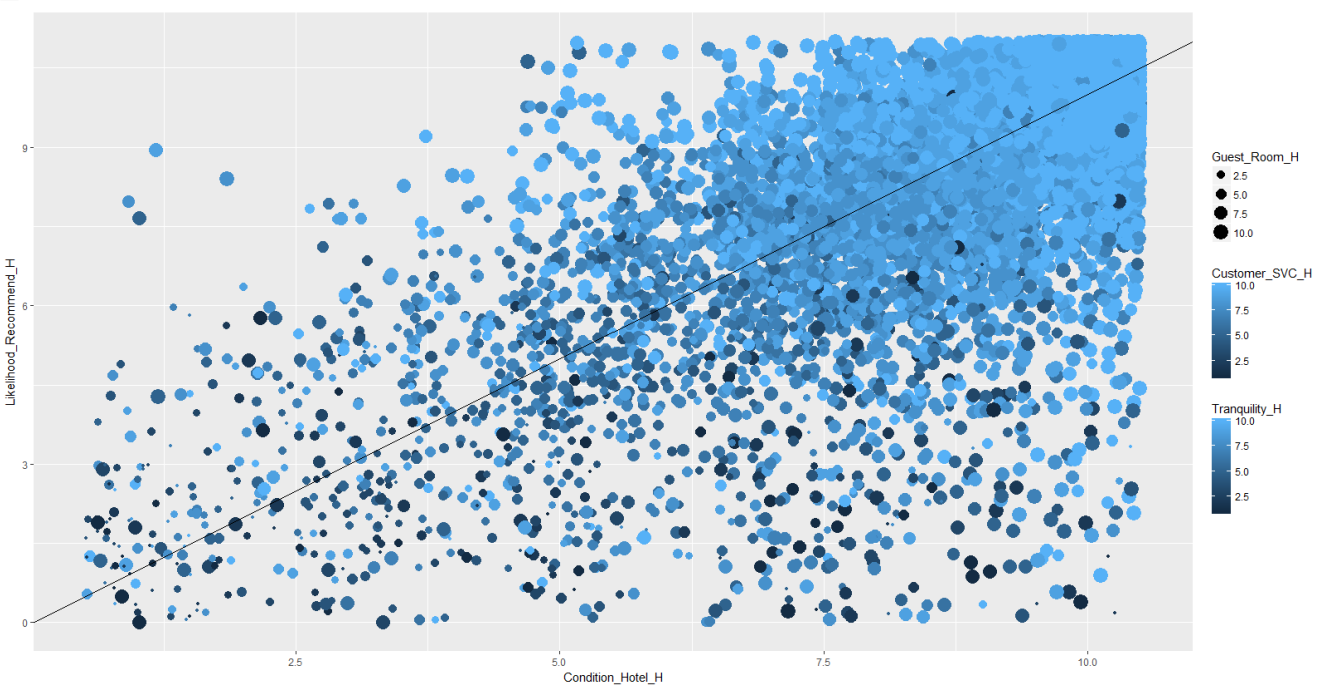
Comparison of R-square values of different models:



Results of Best Linear model plotted with a heat map



Scatter plot against Likelihood_Recommend for the factors obtained from linear model:



Association Rules:

- We performed association mining to determine if there were correlations between columns. To accomplish this, we created a new data frame as a test and changed any data types to factors. Next, for the survey columns shown below, we converted the factor levels to “low,” “mid,” and “high” based on the number ranges for NPS type. We created a data frame only containing the columns to be tested in each test. First, we performed the association mining with NPS_Type as the dependent variable, and the survey columns as the independent variables. We completed this with each NPS_Type: Promoter, Passive, and Detractor. Any rules that contain factors with numbers occurred because we replaced the NA values with the means. Therefore, those columns can be ignored.

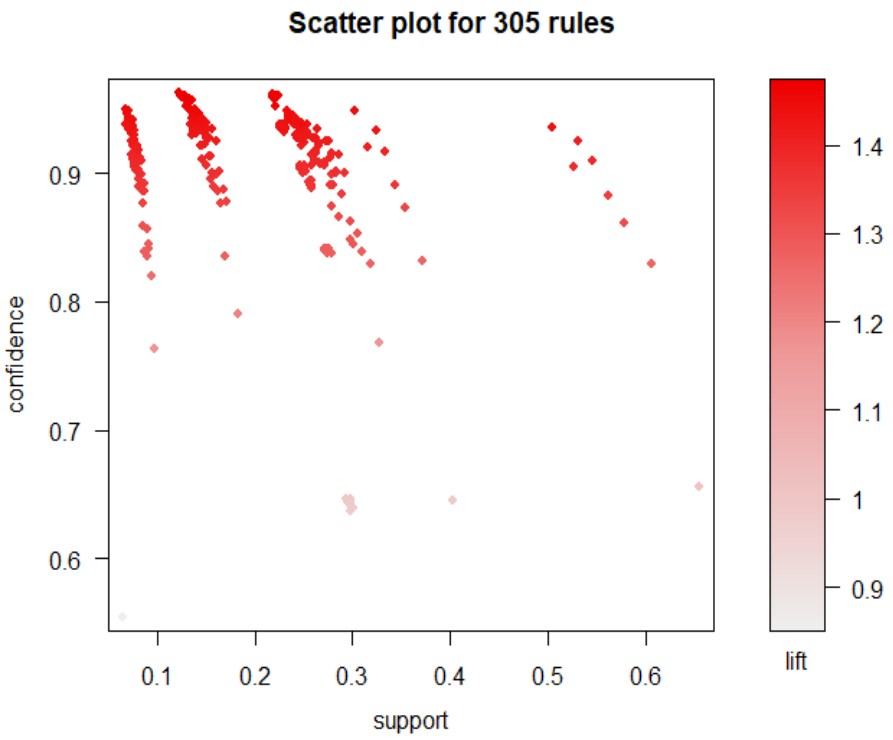
Guest_Room_H
Tranquility_H
Condition_Hotel_H
Customer_SVC_H
Staff_Cared_H
Internet_Sat_H
Check_In_H
F&B_FREQ_H
NPS_Type

- ***Results:***
  - *Promoters*: The top rules demonstrated that the columns for Hotel Condition, Guest Room Condition, and Customer Service Satisfaction were most likely to have high values when the NPS type was Promoter. The lifts for these rules were 1.46, which is significant as they are above 1. Furthermore, the support values were around 26% and the confidence values were around 96%. The results were plotted. The visualization shows the strong association between Promoters and the columns above by visualizing the rules with larger sizes and darker colors for the circles.
  - *Detractors:* The rules provided when NPS type was Detractor showed similar results as those for Promoters. When NPS type is Detractor, low values are likely to be present for Hotel Condition, Guest Room Condition, and Customer Service Satisfaction. The lifts for these rules ranged from 4.8 to 5.2, which is even better than the results for Promoters. The support and confidence values were lower. However, since the NPS was very good for this subset, there were not many detractors. Therefore, it makes sense that the support values, which were between 5 and 9 percent, were lower. The support is lower because they don't appear as often in the dataset. The confidence values were still decent at 70-86%.
  - *Passives:* The results were the same as those for Promoters and Detractors. A Passive NPS type was likely to show up with a mid-range score for Hotel Condition, Guest Room Condition, and Customer Service Satisfaction. The support for this was around 6%, the confidence was around 65%, and the lift was around 3.2.

Association mining for survey columns where NPS-Type=Promoter:

```
> ruleset <- apriori(test,
+                    parameter=list(support=0.05,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Promoter")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5    0.05      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 576

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[31 item(s), 11526 transaction(s)] done [0.00s].
sorting and recoding items ... [26 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.00s].
writing ... [305 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

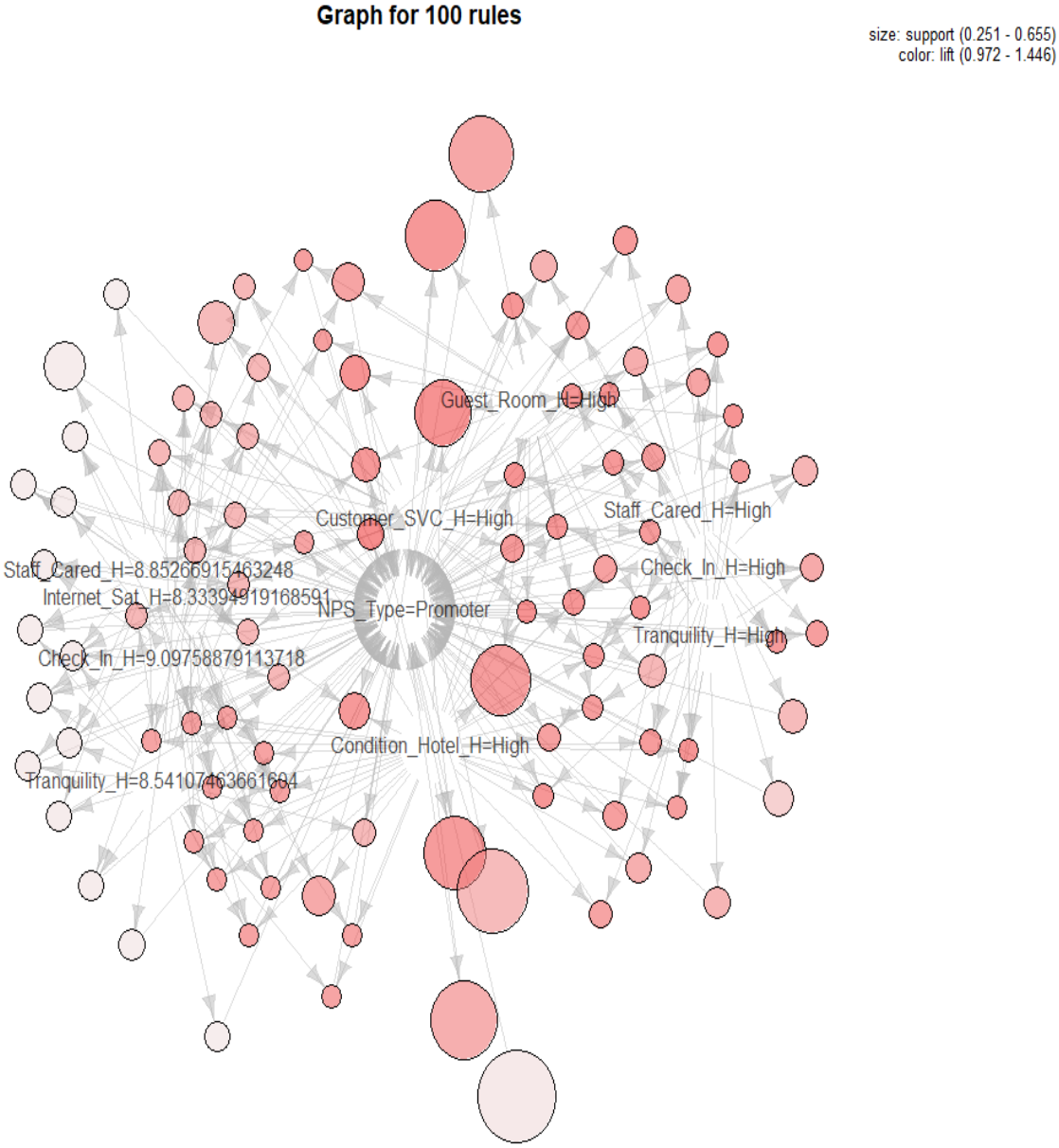
Plot of association mining for survey columns where NPS-Type=Promoter.



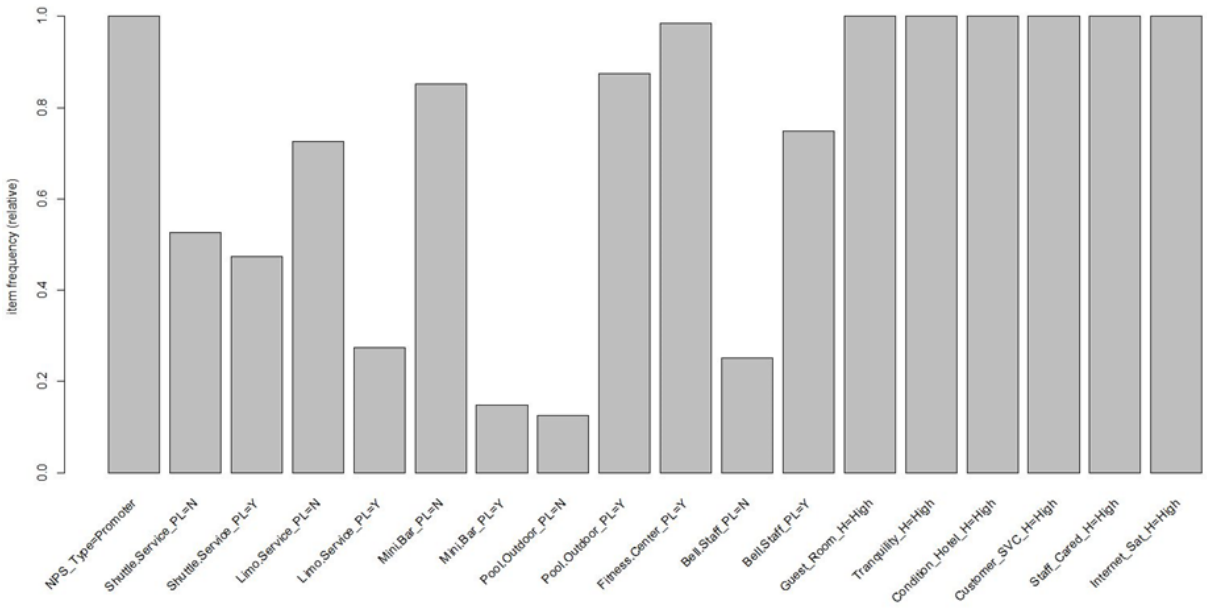
Rules for association mining for survey columns where NPS-Type=Promoter.

```
> interesting <- ruleset[quality(ruleset)$lift > 1.464]
> inspect(interesting)
     lhs                                  rhs                    support confidence     lift count
[1]  {Guest_Room_H=High,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Staff_Cared_H=8.85266915463248}  => {NPS_Type=Promoter} 0.2211522  0.9597139 1.464151  2549
[2]  {Guest_Room_H=High,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Staff_Cared_H=8.85266915463248,
      Check_In_H=9.09758879113718}     => {NPS_Type=Promoter} 0.2195037  0.9601518 1.464819  2530
[3]  {Guest_Room_H=High,
      Tranquility_H=8.54107463661604,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Staff_Cared_H=8.85266915463248}  => {NPS_Type=Promoter} 0.2195037  0.9601518 1.464819  2530
[4]  {Guest_Room_H=High,
      Tranquility_H=8.54107463661604,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Check_In_H=9.09758879113718}     => {NPS_Type=Promoter} 0.2195037  0.9601518 1.464819  2530
[5]  {Guest_Room_H=High,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Internet_Sat_H=8.33394919168591,
      Check_In_H=9.09758879113718}     => {NPS_Type=Promoter} 0.2206316  0.9599849 1.464565  2543
[6]  {Guest_Room_H=High,
      Tranquility_H=8.54107463661604,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Internet_Sat_H=8.33394919168591} => {NPS_Type=Promoter} 0.2200243  0.9602423 1.464957  2536
[7]  {Guest_Room_H=High,
      Tranquility_H=High,
      Condition_Hotel_H=High,
      Staff_Cared_H=High,
      Internet_Sat_H=High,
      Check_In_H=High}                 => {NPS_Type=Promoter} 0.1251085  0.9600533 1.464669  1442
[8]  {Guest_Room_H=High,
      Tranquility_H=High,
      Customer_SVC_H=High,
      Staff_Cared_H=High,
      Internet_Sat_H=High,
      Check_In_H=High}                 => {NPS_Type=Promoter} 0.1263231  0.9616909 1.467167  1456
[9]  {Tranquility_H=High,
      Condition_Hotel_H=High,
      Customer_SVC_H=High,
      Staff_Cared_H=High,
      Internet_Sat_H=High,
      Check_In_H=High}                 => {NPS_Type=Promoter} 0.1259761  0.9603175 1.465072  1452
```

Association rules for survey columns with NPS_Type=Promoter.



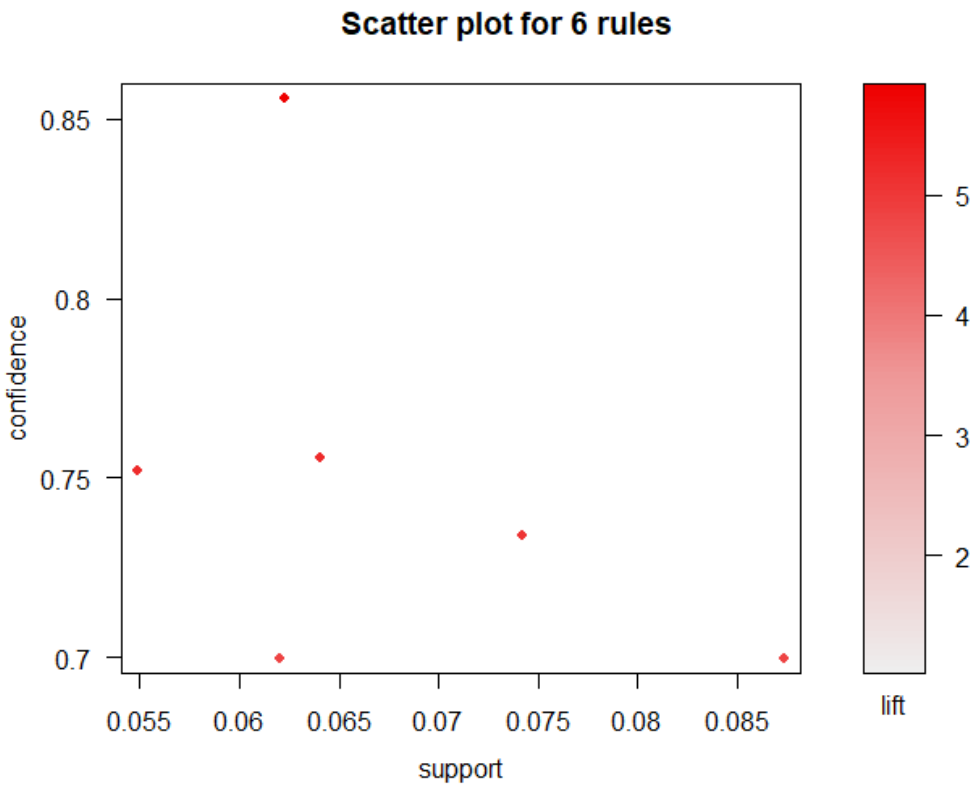
Factors contributing to high promoter score:

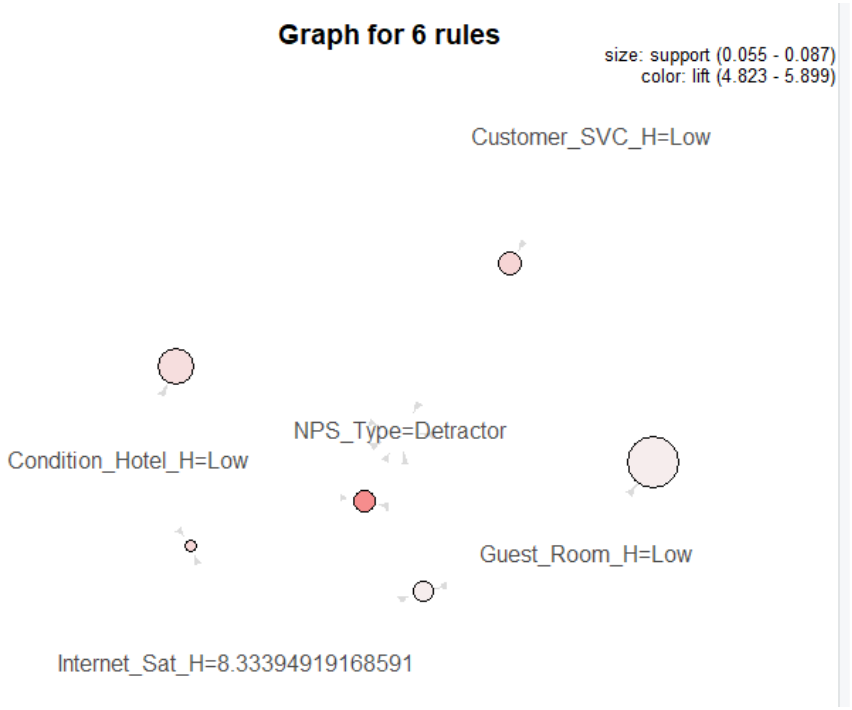


Association mining for survey columns where NPS-Type=Detractor.

```
> ruleset <- apriori(test,
+                    parameter=list(support=0.05,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Detractor")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5    0.05      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 576

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[31 item(s), 11526 transaction(s)] done [0.00s].
sorting and recoding items ... [26 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.00s].
writing ... [6 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

Plot of association mining for survey columns where NPS-Type=Detractor.





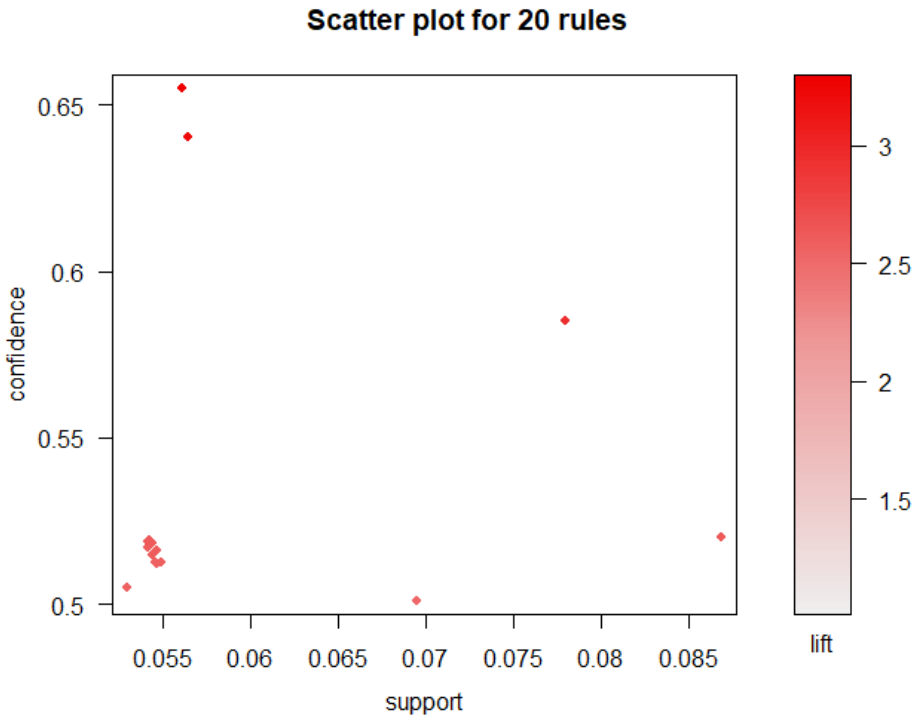
Rules for association mining for survey columns where NPS-Type=Detractor.

```
> inspect(ruleset)
    lhs                                                    rhs                    support    confidence lift     count
[1] {Customer_SVC_H=Low}                                => {NPS_Type=Detractor} 0.06402915 0.7553736  5.207199  738
[2] {Condition_Hotel_H=Low}                             => {NPS_Type=Detractor} 0.07418011 0.7339056  5.059208  855
[3] {Guest_Room_H=Low}                                  => {NPS_Type=Detractor} 0.08736769 0.6997915  4.824041 1007
[4] {Guest_Room_H=Low,Condition_Hotel_H=Low}            => {NPS_Type=Detractor} 0.06229394 0.8557807  5.899359  718
[5] {Condition_Hotel_H=Low,Internet_Sat_H=8.33394919168591} => {NPS_Type=Detractor} 0.05491931 0.7517815  5.182436  633
[6] {Guest_Room_H=Low,Internet_Sat_H=8.33394919168591}  => {NPS_Type=Detractor} 0.06203366 0.6996086  4.822780  715
>
```

Association mining for survey columns where NPS-Type=Passive.

```
> ruleset <- apriori(test,
+                    parameter=list(support=0.05,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Passive")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5    0.05      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 576

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[31 item(s), 11526 transaction(s)] done [0.00s].
sorting and recoding items ... [26 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.00s].
writing ... [20 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Plot of association mining for survey columns where NPS-Type=Passive.



Rules for association mining for survey columns where NPS-Type=Passive.

```
> inspect(interesting)
    lhs                                                rhs                    support    confidence lift     count
[1] {Condition_Hotel_H=Mid,Customer_SVC_H=Mid} => {NPS_Type=Passive} 0.05639424 0.6403941  3.210606 650
[2] {Guest_Room_H=Mid,Customer_SVC_H=Mid}      => {NPS_Type=Passive} 0.05596044 0.6554878  3.286278 645
```

Guest Related Columns Not From Survey.

- We also performed association mining on columns related to guest information as shown below. These columns were either already set as factors, or this was corrected. We left their factor levels as they were and investigated if certain column values are more likely to appear with each of the NPS types.

- **Results:**
    - *Promoter:* The results showed that the use of the Club lounge was associated with Promoters. The support was around 6%, the confidence was around 83%, and the lifts were around 1.4. There were also rules which contained values, such as length of stay and number of children, that we did not consider significant. We evaluated those columns using descriptive statistics, as shown later in the report, and noted that the amount of data for each was too low to be significant.
    - There were no rules for Detractors or Passives.

```
> str(Test1)
'data.frame':   243 obs. of  10 variables:
 $ ROOM_TYPE_DESCRIPTION_C: Factor w/ 1862 levels "","1 BD KING",..: 753 948 981 981 731 753 538 981 1365 731 ...
 $ LENGTH_OF_STAY_C       : Factor w/ 44 levels "1","2","3","4",..: 2 2 4 7 3 3 2 3 3 4 ...
 $ CHILDREN_NUM_C         : Factor w/ 4 levels "1","2","3","4": 2 1 2 1 2 2 2 2 1 1 ...
 $ POV_CODE_C             : Factor w/ 2 levels "BUSINESS","LEISURE": 1 1 1 1 1 1 1 1 2 1 ...
 $ Gender_H               : Factor w/ 4 levels "","Female","Male",..: 3 2 2 3 3 3 3 2 2 2 ...
 $ Age_Range_H            : Factor w/ 8 levels "","18-25","26-35",..: 4 4 4 6 4 6 4 5 5 4 ...
 $ Clublounge_Used_H      : Factor w/ 4 levels "","I don't know",..: 3 3 3 3 3 3 4 3 3 3 ...
 $ Spa_Used_H             : Factor w/ 3 levels "","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ GP_Tier_H              : Factor w/ 6 levels "","CARD","DIAM",..: 4 4 4 4 4 4 3 4 4 4 ...
 $ NPS_Type               : Factor w/ 4 levels "","Detractor",..: 3 2 4 2 4 4 3 4 3 3 ...
```

Association mining for non-survey columns where NPS_Type=Promoter.

```
> ruleset <- apriori(Test1,
+                    parameter=list(support=0.05,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Promoter")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5    0.05      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 12

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[91 item(s), 243 transaction(s)] done [0.00s].
sorting and recoding items ... [28 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 done [0.00s].
writing ... [499 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

Plot of association mining rules for non-survey columns where NPS_Type=Promoter.

Rules for non-survey columns where NPS_Type=Promoter.

```
> inspect(interesting)
     lhs                                rhs                     support confidence     lift count
[1]  {Clublounge_Used_H=Yes,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.06172840  0.8333333 1.377551    15
[2]  {LENGTH_OF_STAY_C=4,
      CHILDREN_NUM_C=2}              => {NPS_Type=Promoter} 0.06584362  0.8421053 1.392052    16
[3]  {POV_CODE_C=BUSINESS,
      Clublounge_Used_H=Yes,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.05349794  0.8125000 1.343112    13
[4]  {Clublounge_Used_H=Yes,
      Spa_Used_H=No,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.05349794  0.8125000 1.343112    13
[5]  {LENGTH_OF_STAY_C=4,
      Gender_H=Male,
      Clublounge_Used_H=No}          => {NPS_Type=Promoter} 0.05761317  0.8235294 1.361345    14
[6]  {LENGTH_OF_STAY_C=4,
      CHILDREN_NUM_C=2,
      POV_CODE_C=BUSINESS}           => {NPS_Type=Promoter} 0.05349794  0.8125000 1.343112    13
[7]  {LENGTH_OF_STAY_C=4,
      CHILDREN_NUM_C=2,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.05349794  0.8125000 1.343112    13
[8]  {LENGTH_OF_STAY_C=4,
      CHILDREN_NUM_C=2,
      Spa_Used_H=No}                 => {NPS_Type=Promoter} 0.05761317  0.8235294 1.361345    14
[9]  {LENGTH_OF_STAY_C=3,
      Age_Range_H=36-45,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.05761317  0.8235294 1.361345    14
[10] {LENGTH_OF_STAY_C=3,
      Age_Range_H=36-45,
      Spa_Used_H=No,
      GP_Tier_H=GOLD}                => {NPS_Type=Promoter} 0.05761317  0.8235294 1.361345    14
```

Amenities Parts 1 and 2.

- We completed association mining for the columns related to the hotel amenities using NPS_Type as the dependent variable. We were especially interested in observing this since we had noted that Hotel Condition and Customer Service Satisfaction were important. We prepared the data in a similar way as the other association mining tests. We also divided the columns into two tests because we were getting an abundance of rules when using a large number of columns.

- **Results for both tests**:
    - *Promoters:* The lifts for these rules were barely above 1.0, which shows that they were barely significant. Furthermore, we noted from our descriptive statistics that most of the hotels would have or not have an amenity. For instance, if almost every hotel that did not have an NA value had laundry services, we could expect to see it in a rule.

However, we would not find that significant. We did not find any significant rules from these tests.

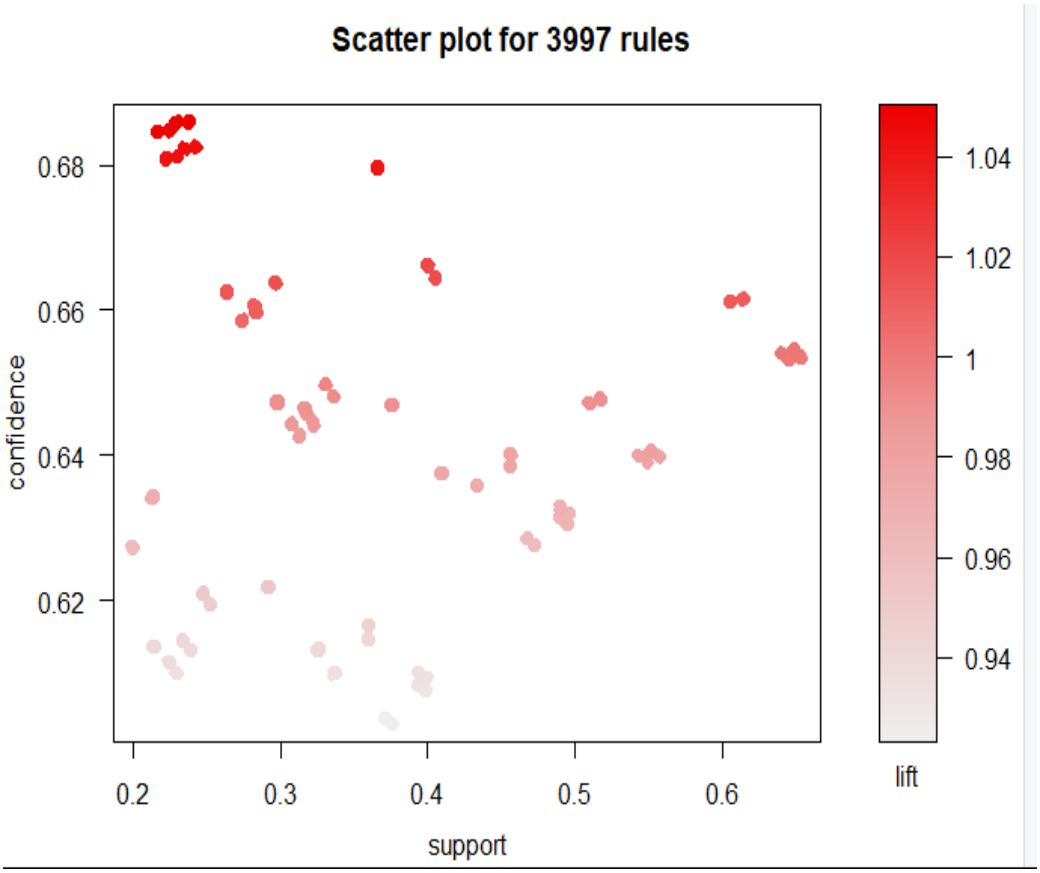
- There were no rules for Detractors or Passives in either test.

```
> test3 <- subset[,c(49:60, 79)]
> str(test3)
'data.frame':   11526 obs. of  13 variables:
 $ Relationship_PL   : Factor w/ 4 levels "Franchised","Leased",..: 1 1 1 1 1 1 1 1 1 ...
 $ All.Suites_PL     : Factor w/ 3 levels "","N","Y": 3 3 3 3 3 3 3 3 3 ...
 $ Bell.Staff_PL     : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ Boutique_PL       : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 ...
 $ Business.Center_PL: Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ Casino_PL         : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 ...
 $ Conference_PL     : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 ...
 $ Convention_PL     : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 ...
 $ Dry.Cleaning_PL   : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ Elevators_PL      : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ Fitness.Center_PL : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ Fitness.Trainer_PL: Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA ...
 $ NPS_Type          : Factor w/ 4 levels "","Detractor",..: 4 4 4 3 3 4 3 3 3 ...
```

Association mining for hotel amenities columns Part 1 where NPS_Type=Promoter.

```
> ruleset <- apriori(test3,
+                    parameter=list(support=0.2,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Promoter")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.2      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1989

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[25 item(s), 9947 transaction(s)] done [0.00s].
sorting and recoding items ... [17 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.01s].
writing ... [3997 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Plot for hotel amenities columns Part 1 where NPS_Type=Promoter.



Rules for hotel amenities columns Part 1 where NPS_Type=Promoter.

```
> inspect(interesting)
     lhs                                  rhs                    support confidence     lift count
[1]  {Business.Center_PL=Y,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[2]  {Boutique_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[3]  {Boutique_PL=N,
      Business.Center_PL=Y,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[4]  {Business.Center_PL=Y,
      Convention_PL=N,
      Fitness.Center_PL=Y}             => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[5]  {Business.Center_PL=Y,
      Convention_PL=N,
      Dry.Cleaning_PL=Y}               => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[6]  {Business.Center_PL=Y,
      Conference_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[7]  {Business.Center_PL=Y,
      Casino_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[8]  {Boutique_PL=N,
      Convention_PL=N,
      Fitness.Center_PL=Y}             => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[9]  {Boutique_PL=N,
      Convention_PL=N,
      Dry.Cleaning_PL=Y}               => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[10] {Boutique_PL=N,
      Conference_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[11] {Boutique_PL=N,
      Casino_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[12] {Boutique_PL=N,
      Business.Center_PL=Y,
      Convention_PL=N,
      Fitness.Center_PL=Y}             => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[13] {Boutique_PL=N,
      Business.Center_PL=Y,
      Convention_PL=N,
      Dry.Cleaning_PL=Y}               => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
[14] {Boutique_PL=N,
      Business.Center_PL=Y,
      Conference_PL=N,
      Convention_PL=N}                 => {NPS_Type=Promoter} 0.2383633  0.6858548 1.049569  2371
```

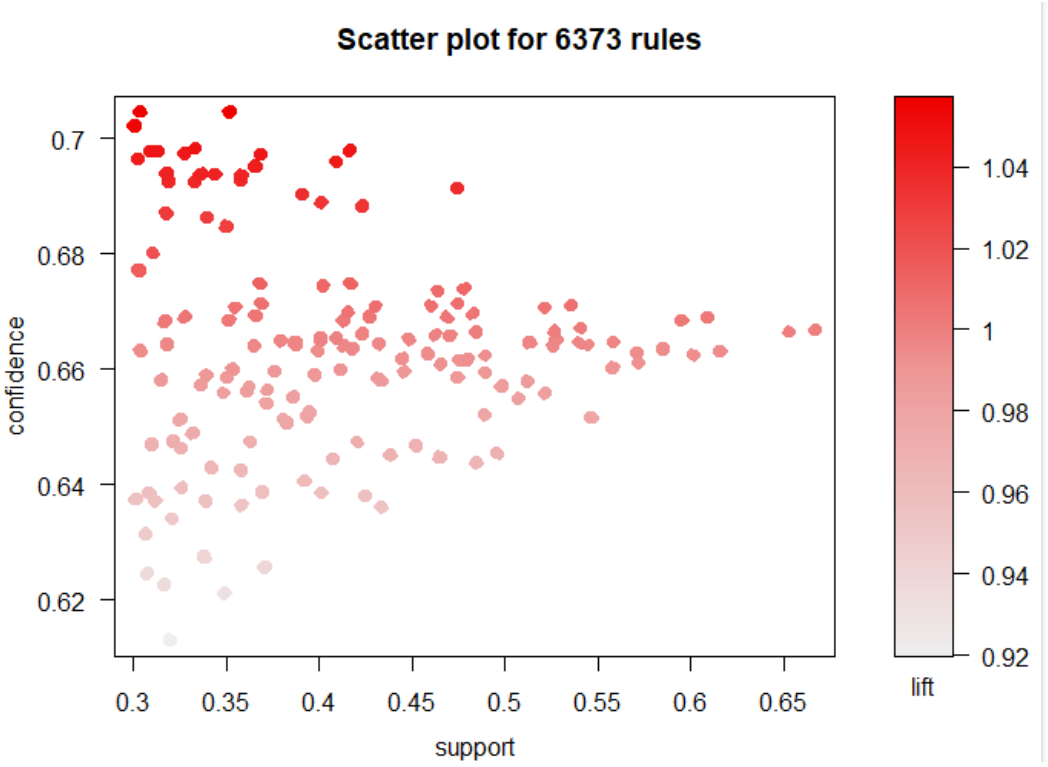
List of columns used before NA values removed.

```
> test4 <- subset[,c(61:75,78:79)]
> str(test4)
'data.frame':   11526 obs. of  17 variables:
 $ Golf_PL                          : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Indoor.Corridors_PL              : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Laundry_PL                       : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Limo.Service_PL                  : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Mini.Bar_PL                      : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Pool.Indoor_PL                   : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Pool.Outdoor_PL                  : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Regency.Grand.Club_PL            : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Resort_PL                        : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Restaurant_PL                    : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Self.Parking_PL                  : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Shuttle.Service_PL               : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Ski_PL                           : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Spa_PL                           : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Spa.services.in.fitness.center_PL: Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ Valet.Parking_PL                 : Factor w/ 3 levels "","N","Y": NA NA NA NA NA NA NA NA NA NA ...
 $ NPS_Type                         : Factor w/ 4 levels "","Detractor",..: 4 4 4 3 3 4 3 3 3 3 ...
```

Association mining for hotel amenities columns Part 2 where NPS_Type=Promoter.

```
> ruleset <- apriori(test4,
+                    parameter=list(support=0.3,confidence=0.5),
+                    appearance=list(default="lhs",rhs=("NPS_Type=Promoter")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.3      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1743

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[31 item(s), 5810 transaction(s)] done [0.00s].
sorting and recoding items ... [21 item(s)] done [0.01s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.02s].
writing ... [6373 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Plot for hotel amenities columns Part 2 where NPS_Type=Promoter.



Rules for hotel amenities columns Part 2 where NPS_Type=Promoter.

```
> inspect(interesting)
     lhs                                rhs                    support confidence     lift count
[1]  {Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y}             => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[2]  {Indoor.Corridors_PL=Y,
      Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y}             => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[3]  {Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Spa_PL=N}                      => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[4]  {Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Ski_PL=N}                      => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[5]  {Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Resort_PL=N}                   => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[6]  {Golf_PL=N,
      Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y}             => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[7]  {Indoor.Corridors_PL=Y,
      Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Spa_PL=N}                      => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[8]  {Indoor.Corridors_PL=Y,
      Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Ski_PL=N}                      => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
[9]  {Indoor.Corridors_PL=Y,
      Laundry_PL=Y,
      Limo.Service_PL=N,
      Pool.Indoor_PL=N,
      Pool.Outdoor_PL=Y,
      Resort_PL=N}                   => {NPS_Type=Promoter} 0.3037866  0.7045908 1.056432  1765
```

Support Vector Machine:

1) KSVM Model

We used the KSVM classification model for predicting the effect of the Amenities and the Survey columns on the NPS. To achieve this, firstly we created a data frame with a combination of quantitative and categorical data. We then cleaned the data set by omitting the NA values,converting to factors and excluding the columns which had less that 2 levels in the factor.

Thereafter we divided the subset into 2 parts i.e TrainData and Test Data. The TrainData was selected by dividing two third of the subset and the remaining one third of the subset was taken as the Test Data.


```
#Creating a random index
randIndex <- sample(1:dim(svm_new)[1])
randIndex
summary(randIndex)

#Making a training cutpoint
train_cutpoint2_3 <- floor((2*dim(svm_new)[1])/3)
train_cutpoint2_3
trainData <- svm_new[randIndex[1:train_cutpoint2_3],]
dim(trainData)[1]
View(trainData)
str(trainData)

#making testing cutpoint
testCutpoint <- dim(svm_new)[1]-(train_cutpoint2_3+1)
testCutpoint
testData <- svm_new[randIndex[train_cutpoint2_3+1:testCutpoint],]
View(testData)
dim(testData)
```

The TrainData was then applied to create the KSVM model. The model was then tested on the testData by predicting the NPS value. We then created a comparison table of predicted NPS value and the actual NPS value which was plotted using a histogram

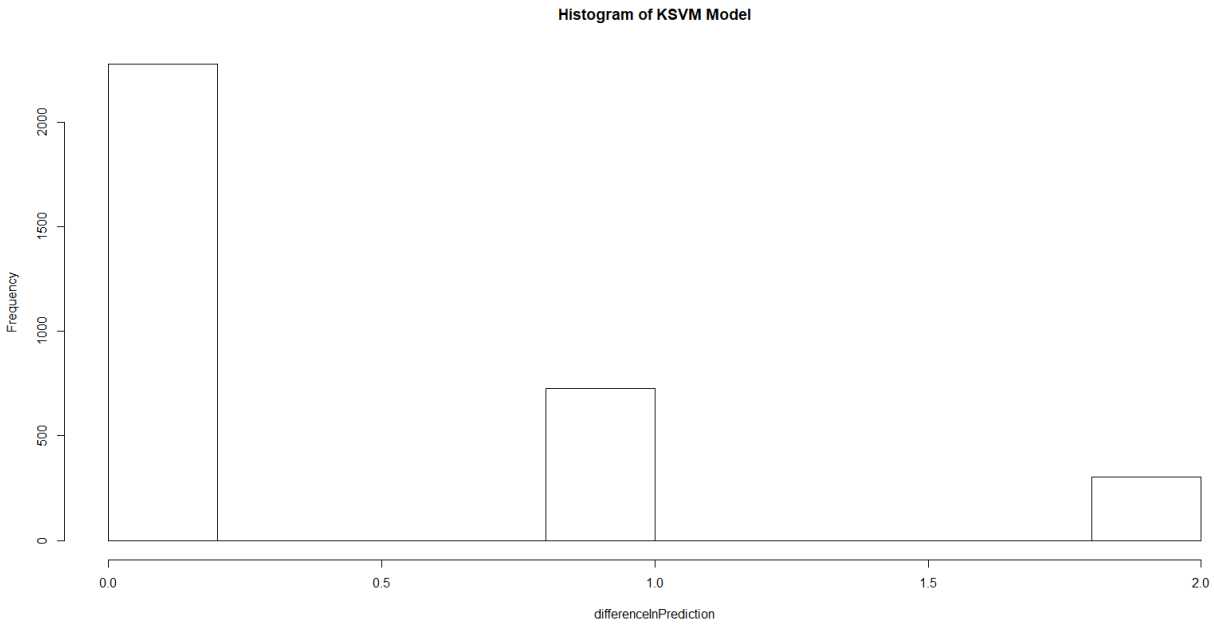
Cross Validation Error of the KSVM Model:

```
> ksvmOutput_NPS1
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 10

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.216666666666667

Number of Support Vectors : 3446

Objective Function Value : -11672.43 -12264.5 -18491.8
Training error : 0.22407
Cross validation error : 0.302387
Probability model included.
```

Comparison table showing the actual values in the test data vs the values predicted by the model:

```
> table(compTable1)
                ksvmpredNPS
testData...3.     0     1     2
            0   108    99   263
            1    88   170   429
            2    42   109 1998
```

Histogram showing the difference between actual and values predicted using KSVM:



2) SVM Model:

We used the SVM classification model for predicting the effect of the Amenities and the Survey columns on the NPS. To achieve this, we used the same subsets of TrainData and Test Data, as the

KSVM. The TrainData was selected by dividing two third of the original subset and the remaining one third was taken as the Test Data.

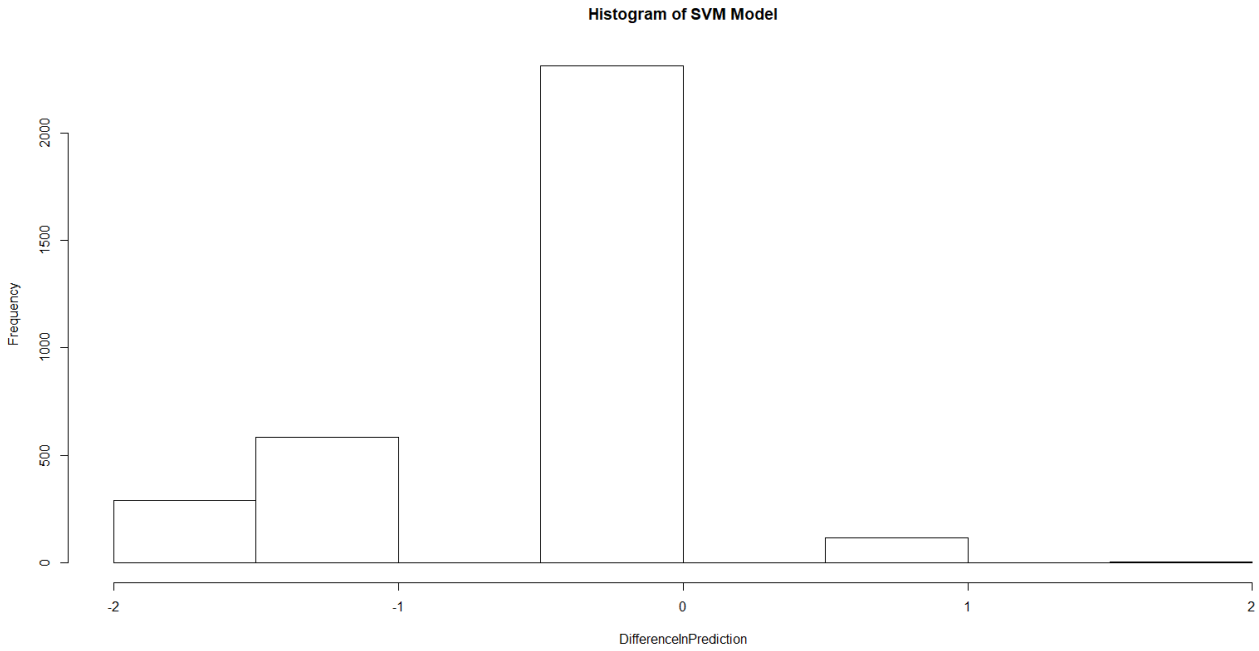
SVM Output:

```
> svmOutput

Call:
svm(formula = NPS_Type ~ Condition_Hotel_H + Tranquility_H + Staff_Cared_H + Check_In_H + Internet_Sat_H +
    Customer_SVC_H + Convention_PL + Business.Center_PL + Spa_PL + Mini.Bar_PL + Valet.Parking_PL + Self.Parking_PL +
    Shuttle.Service_PL + Limo.Service_PL + Pool.Outdoor_PL + Pool.Indoor_PL + Golf_PL + Fitness.Center_PL,
    data = trainData)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.04

Number of Support Vectors:  3529
```

Comparison table showing the actual values in the test data vs the values predicted by the model:

```
> table(compTable)
              svmpred
testData...22.    0    1    2
             0   15   24   38
             1    8  120  200
             2   59  275 2567
```

Histogram showing the difference between actual and values predicted using SVM:



Non-Significant Analysis

- ***Question*:** Does the length of stay have an impact on the likelihood to recommend?

Method: Using tapply, we calculated the mean, min, max, and median likelihood to recommend scores for each length of stay. We also calculated how many unique values there were for length of stay. We combined all of those in a data frame to observe.


```
#Use tapply to see relationships between length of stay and likelihood to recommend.
StayLengthMean <- tapply(subset$Likelihood_Recommend_H, subset$LENGTH_OF_STAY_C, mean)
StayLengthMin <- tapply(subset$Likelihood_Recommend_H, subset$LENGTH_OF_STAY_C, min)
StayLengthMax <- tapply(subset$Likelihood_Recommend_H, subset$LENGTH_OF_STAY_C, max)
StayLength <- data.frame(StayLengthMean, StayLengthMin, StayLengthMax)
View(StayLength)

#Use tapply to see how many values there are for each length of stay.
StayCount <- tapply(subset$LENGTH_OF_STAY_C, subset$LENGTH_OF_STAY_C, length)
StayCount

StayLength1 <- tapply(subset$LENGTH_OF_STAY_C, subset$LENGTH_OF_STAY_C, unique)
View(StayLength1)
StayLength1 <- data.frame(StayLength1)
names(StayLength1) <- c("Length of Stay")
StayLength2 <- data.frame(StayLength, StayLength1, StayCount)
View(StayLength2)
```

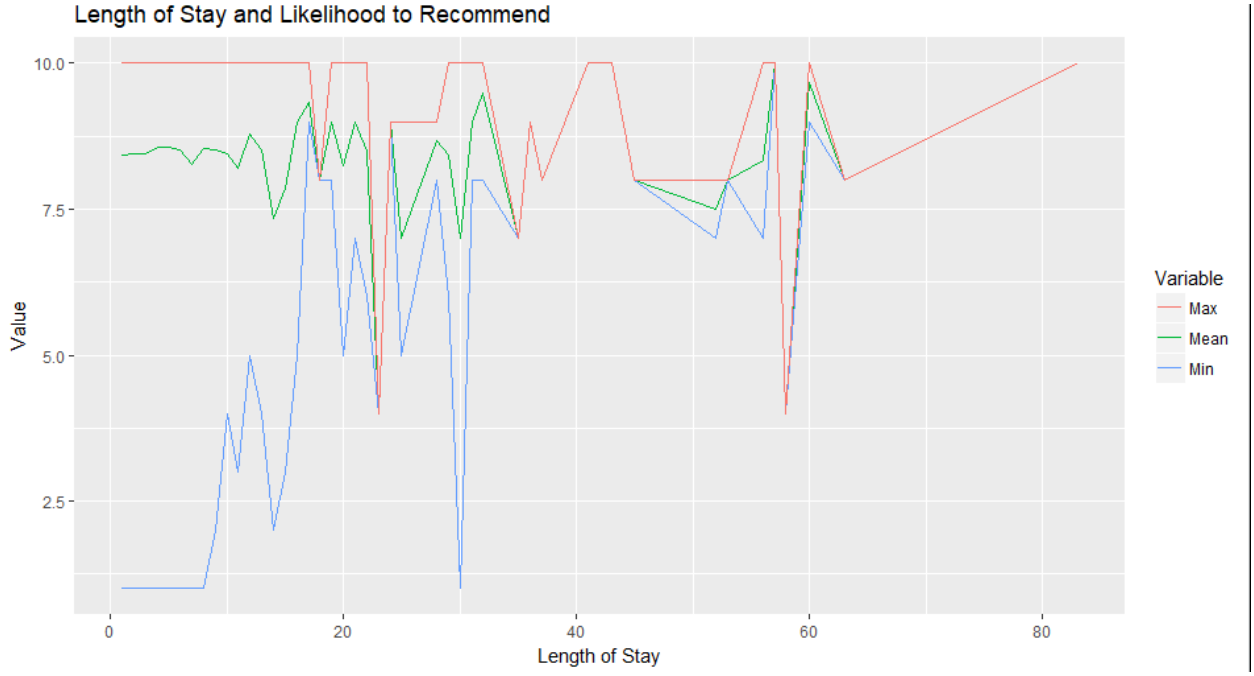
***Analysis*:** Most people stayed for fewer than seven days, but there were no significant differences among those first seven days. The mean likelihood to recommend did increase as guests stayed longer than seven days, but there were not enough guests staying for that long for it to be meaningful.

Effect of length of stay on likelihood to recommend score.

StayLengthMean	StayLengthMin	StayLengthMax	Length.of.Stay	StayCount
8.434054	1	10	1	3700
8.447885	1	10	2	3262
8.449513	1	10	3	1951
8.564428	1	10	4	1102
8.573746	1	10	5	678
8.513678	1	10	6	329
8.266667	1	10	7	195
8.541667	1	10	8	72
8.511111	2	10	9	45
8.444444	4	10	10	27
8.208333	3	10	11	24
8.789474	5	10	12	19
8.526316	4	10	13	19
7.333333	2	10	14	15
7.857143	3	10	15	7
9.000000	5	10	16	7
9.333333	9	10	17	3


```
LenStayLine <- ggplot(StayLength2, aes(x=StayLength2$Length.of.Stay, y=StayLength2$StayLengthMean, color="Mean")) + geom_line()
LenStayLine <- LenStayLine + geom_line(aes(x=StayLength2$Length.of.Stay, y=StayLength2$StayLengthMin, color="Min"))
LenStayLine <- LenStayLine + geom_line(aes(x=StayLength2$Length.of.Stay, y=StayLength2$StayLengthMax, color="Max"))
LenStayLine <- LenStayLine + labs(x="Length of Stay", y="Value", color="Variable", title="Length of Stay and Likelihood to Recommend")
LenStayLine
```

Effect of length of stay on likelihood to recommend.

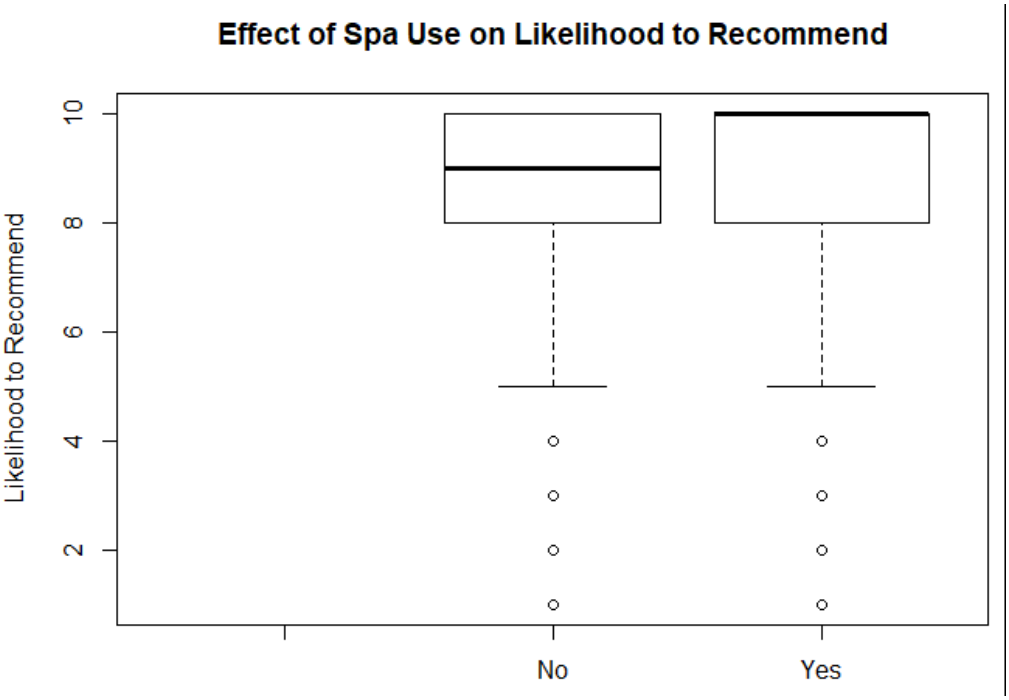


- ***Question*:** Does the use of the Spa affect the likelihood to recommend or room condition score?

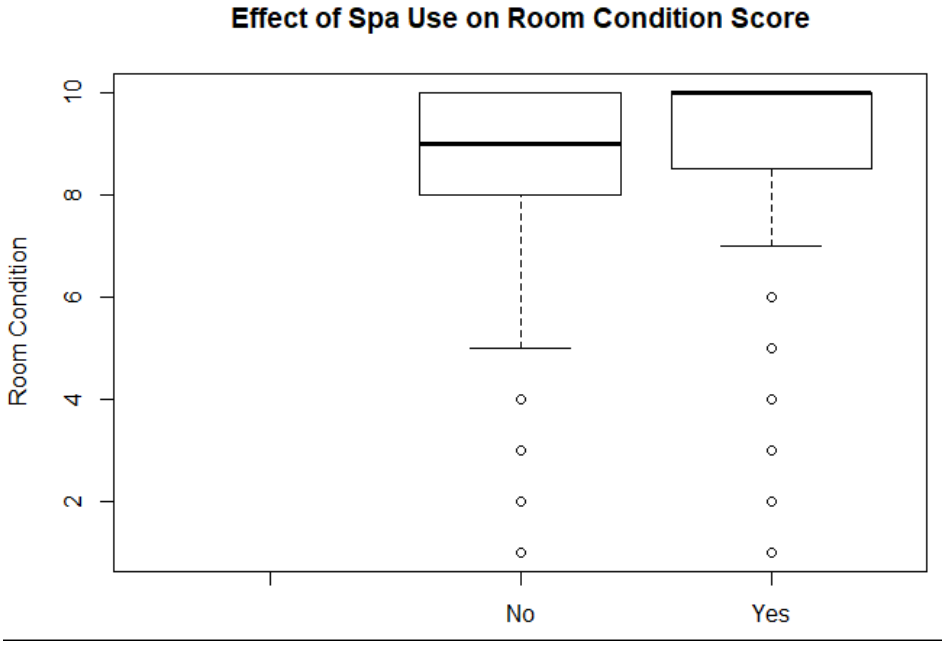
***Method*:** We converted the data type in the column ''Spa_Used_H' to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for those who used and did not use the Spa. We also calculated the number of yes and no values for each. The effect of Spa Usage on the Room Condition score was also evaluated by using the same method.

***Analysis*:** There were significantly higher likelihood to recommend and room condition means for those who used the Spa as shown in figures below. However, the number of people who used it was very low; and although it did appear in some association rules, it was most likely just appearing because almost all guests had a 'No' value for that column. Therefore, we decided it was not a significant factor for likelihood to recommend.

Effect of spa use on likelihood to recommend score.



Effect of spa use on room condition score.



- ***Question*:** Can the use of certain amenities in a hotel, such as Valet Parking or an Indoor Pool, affect the likelihood to recommend?

***Method*:** We converted the data types in the columns chosen to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for the presence or absence of each amenity. We also calculated the number of yes and no values for each, whether it was number of people who used an amenity or if it was present in the hotel.

***Analysis*:** There was a lower likelihood to recommend mean for the hotels who offered Valet Parking, but the median values were both 9. Furthermore, significantly more guests were in hotels that offered it. It is possible that the increased amount of data negatively affected the mean. The likelihood to recommend was slightly lower for those who were in hotels with an Indoor Pool, but very few guests did not have access to an Indoor Pool. Therefore, it is possible the mean was higher because there were less data.

Effect of valet parking on likelihood to recommend mean.

```
> tapply(subset$Likelihood_Recommend_H, t, mean)
       N        Y
NA 8.795980 8.330653
```

```
> tapply(subset$Valet.Parking_PL, subset$Valet.Parking_PL, length)
     N    Y
NA 2985 7089
```

- ***Question*:** Does a guest's gender have an impact on their likelihood to recommend?

***Method*:** We converted the data types in the columns chosen to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for each gender. We also calculated the number of values for each.

***Analysis*:** Females are very slightly more likely to recommend. The total amounts of males and females were almost equal, so there were enough data values to make a conclusion. The difference was so small that it was deemed unimportant.


```
> tapply(subset$Likelihood_Recommend_H, subset$Gender_H, mean)
                                Female                   Male Prefer not to answer
        NA                    8.542857               8.427276                 7.796813
> tapply(subset$Gender_H, subset$Gender_H, length)
                                Female                   Male Prefer not to answer
        NA                      5215                   5954                      251
```

- ***Question*:** Does a guest's GP_Tier, a Loyalty Program tier, have an impact on their likelihood to recommend?

***Method*:** We converted the data types in the columns used to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for each tier of the Loyalty Program. We also calculated the number of values for each.

***Analysis*:** The rank from most likely to recommend to least likely was: 'DIAM', 'CARD', 'GOLD', 'PLAT.' However, very few people were in the 'CARD' or 'DIAM' tiers. The vast majority were in 'GOLD.' Therefore, we decided not to consider the Gold Passport tier a significant factor for likelihood to recommend.

***Business Question**:*

- What is the effect of NPS_Type on the Hotel Brand and Purpose of Visit?

- ***Question*:** Does a guest's POV, their purpose of visit, have an impact on their likelihood to recommend?

***Method*:** We converted the data types in the columns used to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for each factor, 'Business' and 'Leisure'. We also calculated the number of values for each.

***Analysis*:** Almost all guests were staying for business reasons and there was almost no difference in the likelihood to recommend scores for them compared to those staying for leisure. Although business travelers appeared in some association rules, it was most likely because almost all guests were traveling

for business. Furthermore, since there was almost no difference in likelihood to recommend, we elected to consider it not important.

```
> tapply(subset$Likelihood_Recommend_H, v, mean)
BUSINESS  LEISURE
8.453894 8.500000
```

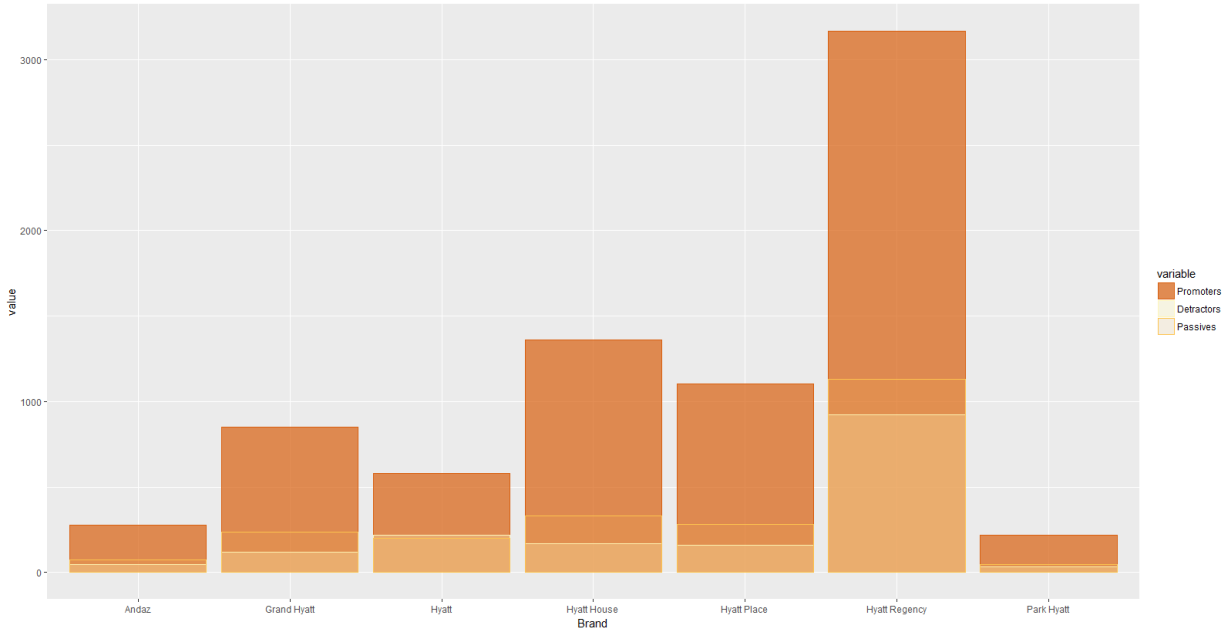
```
> tapply(subset$POV_CODE_C, subset$POV_CODE_C, length)
BUSINESS  LEISURE
    9912     1614
```

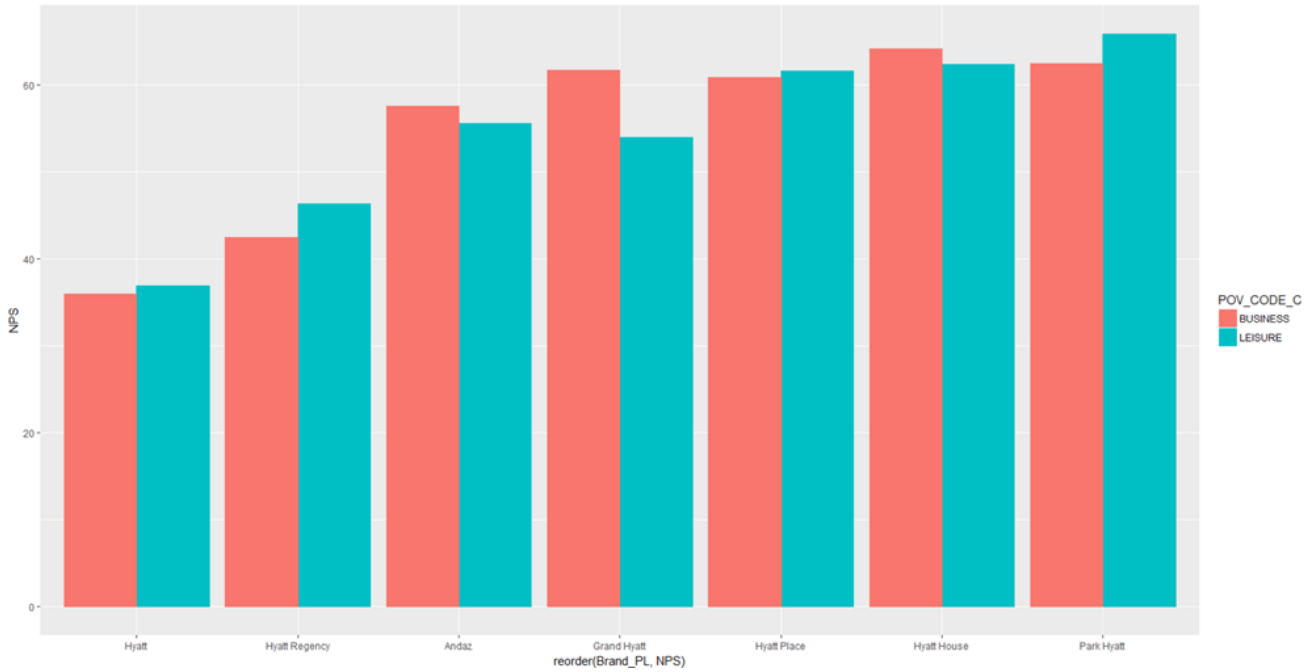
- ***Question***: Does the hotel brand have an impact on their likelihood to recommend? How does it vary by purpose of visit?

Method: We calculated the NPS for each type of guest, business or leisure, based on hotel brand. We also plotted them to see trends.

Analysis: The Park Hyatt hotel brand had the highest NPS, but it also had a lower number of guests. The general Hyatt brand had the lowest NPS.

Impact on likelihood to recommend based on hotel brand type:





- ***Question***: Does the number of children staying with a guest affect the likelihood to recommend?

Method: We converted the data types in the columns used to factor and used tapply to calculate the mean, min, max, and median likelihood to recommend for each factor. We also calculated the number of values for each.

Analysis: Although the likelihood to recommend means increased as the number of children increased, very few guests had the number of children which had a significantly higher mean. Furthermore, although number of children appeared in some association rules, the support and lift values were low. Therefore, we decided there were not enough data to consider it significant.

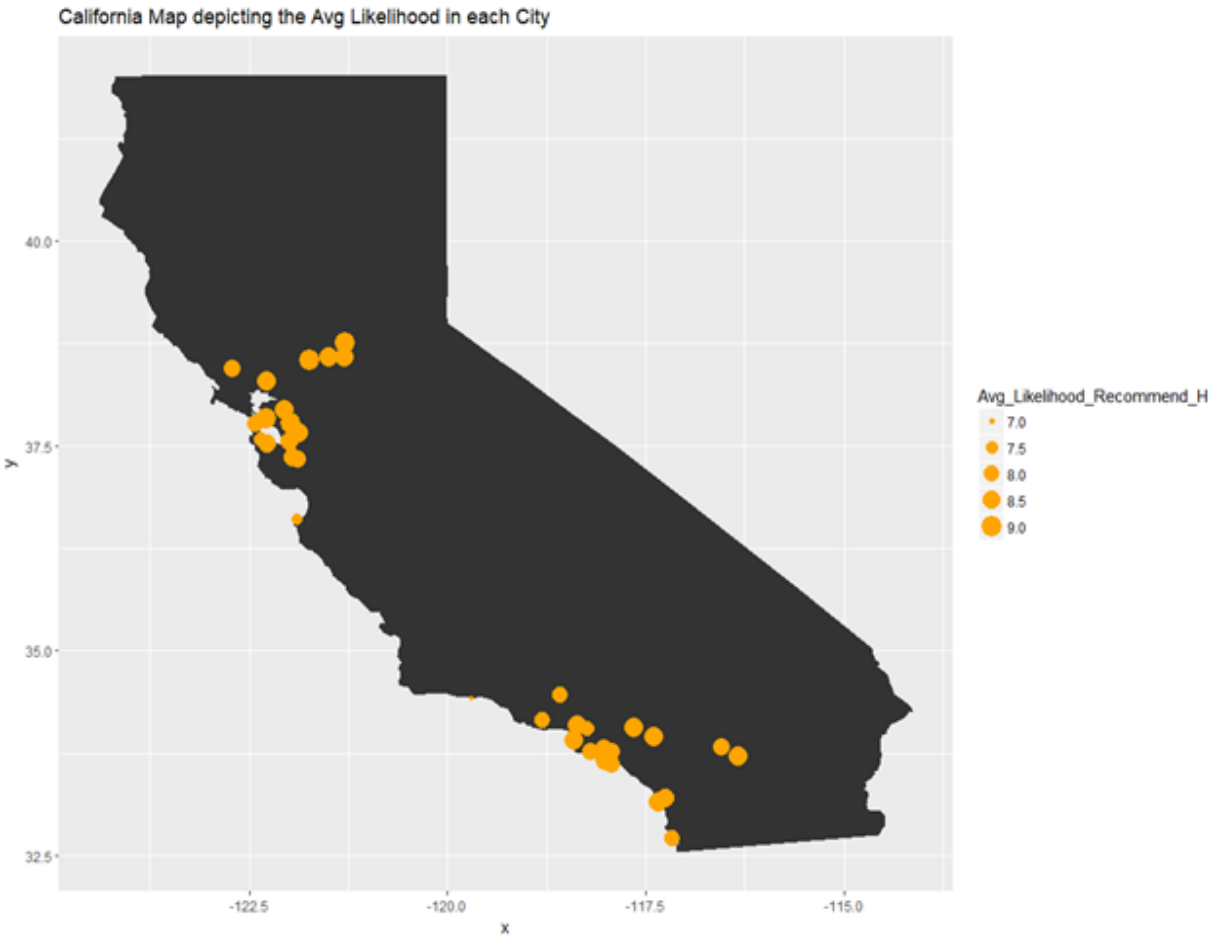
- ***Question***: Does Likelihood to Recommend vary significantly by city?

***Method*:** We used tapply and calculated the mean likelihood to recommend for each city along with the number of guests staying in each city. We also created a map to help visualize the differences geographically.

***Analysis*:** The Monterey had the lowest average likelihood to recommend. The cities of San Francisco and San Diego had the most guests, but slightly lower average likelihood to recommend as shown below.

Mean likelihood to recommend by city, with number of rows for each city.

	city	cityl
San Diego	8.205336	1724
San Francisco	8.443338	1306
Los Angeles	7.977551	490
Santa Clara	8.678351	485
Huntington Beach	9.075000	480
Carlsbad	8.775744	437
Long Beach	8.315534	412
Sacramento	8.788835	412
Monterey	7.452500	400
Garden Grove	8.372340	376
San Jose	8.367470	332
Indian Wells	8.833333	324
Burlingame	7.956250	320
Emeryville	9.016077	311
Rancho Cordova	8.807818	307



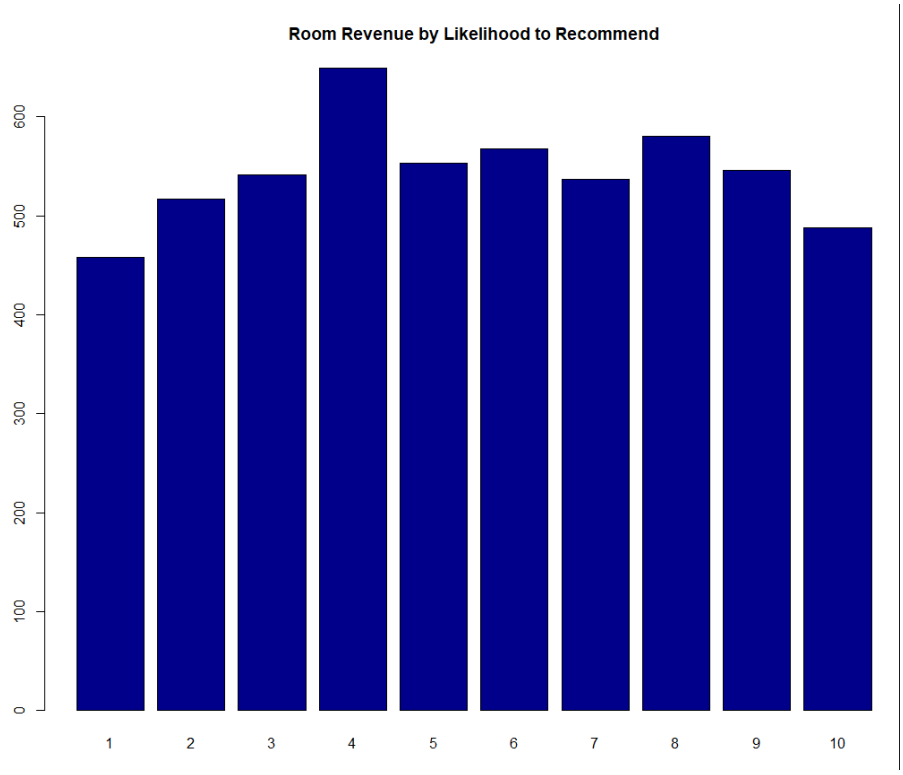
- ***Question***: Does the room revenue influence Likelihood to recommend?

Method: We used tapply to find the mean likelihood to recommend values for each revenue range for the columns of Room Revenue and Average Daily Rate. Then, we plotted them in bar charts to observe any interesting results. We also did the same for room revenue with Guest Room Condition as the dependent variable.

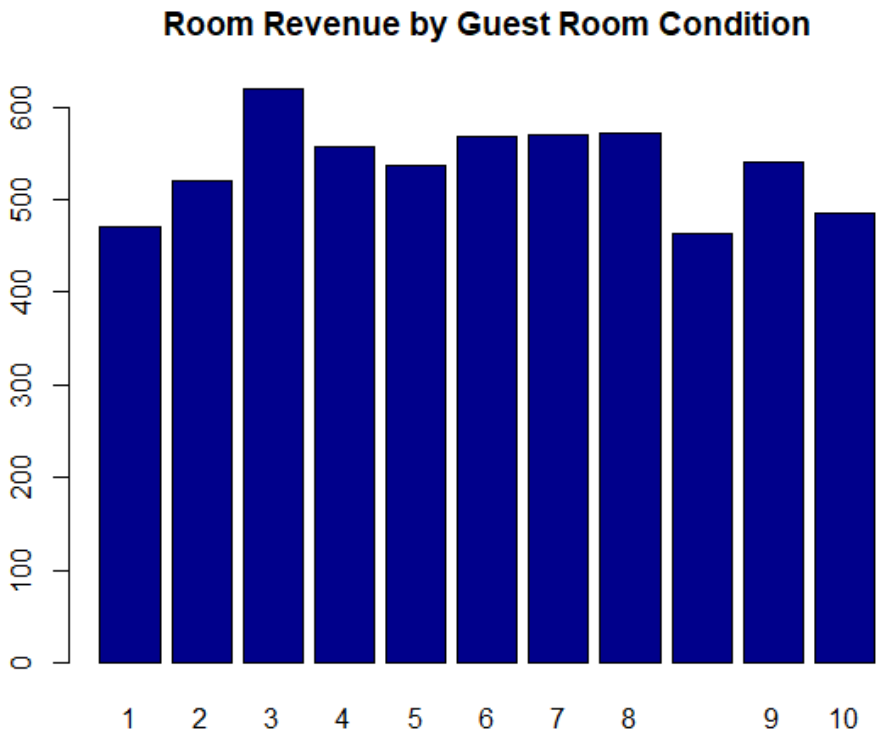
Analysis: The average daily rates were very similar for each value of likelihood to recommend. However, the total revenue was highest for those who had a likelihood to recommend of 4. This trend of higher revenue equaling a lower survey score was also noted for Guest Room Condition. However, room revenue did not appear in any models, and our subset's food and beverage column was excluded

from modeling because it contained over 60% NA values. Therefore, we could not create an actionable insight from these results.

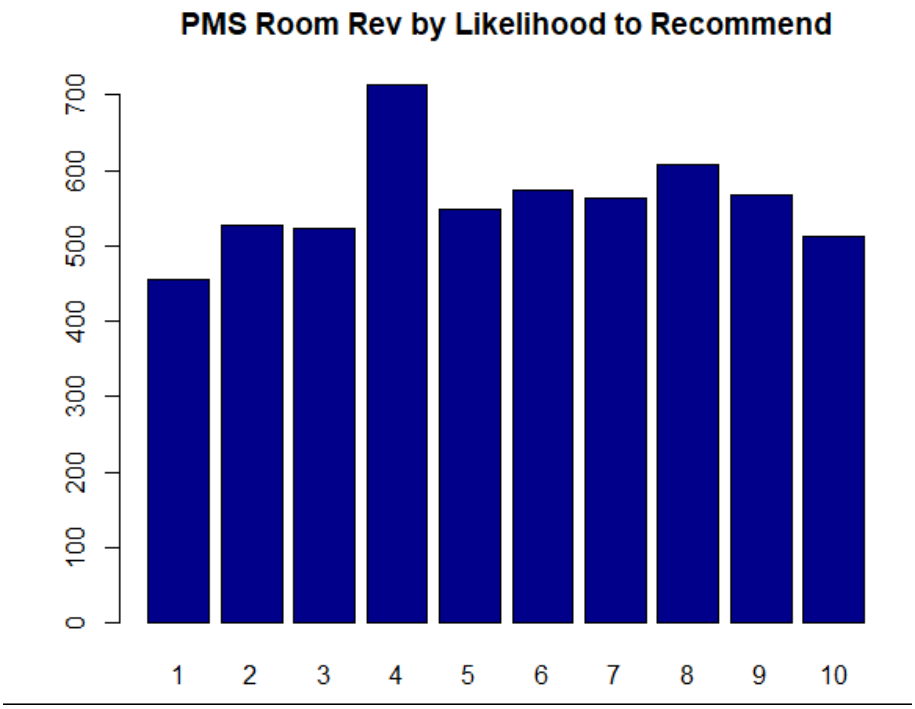
Room revenue by likelihood to recommend.



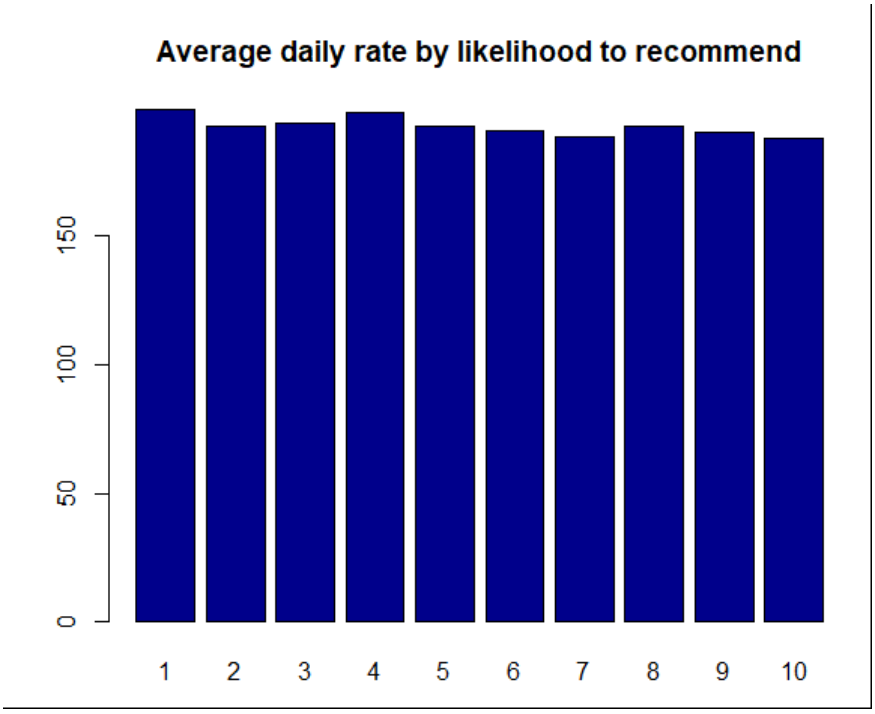
Room revenue by guest room condition.



PMS room revenue by likelihood to recommend.



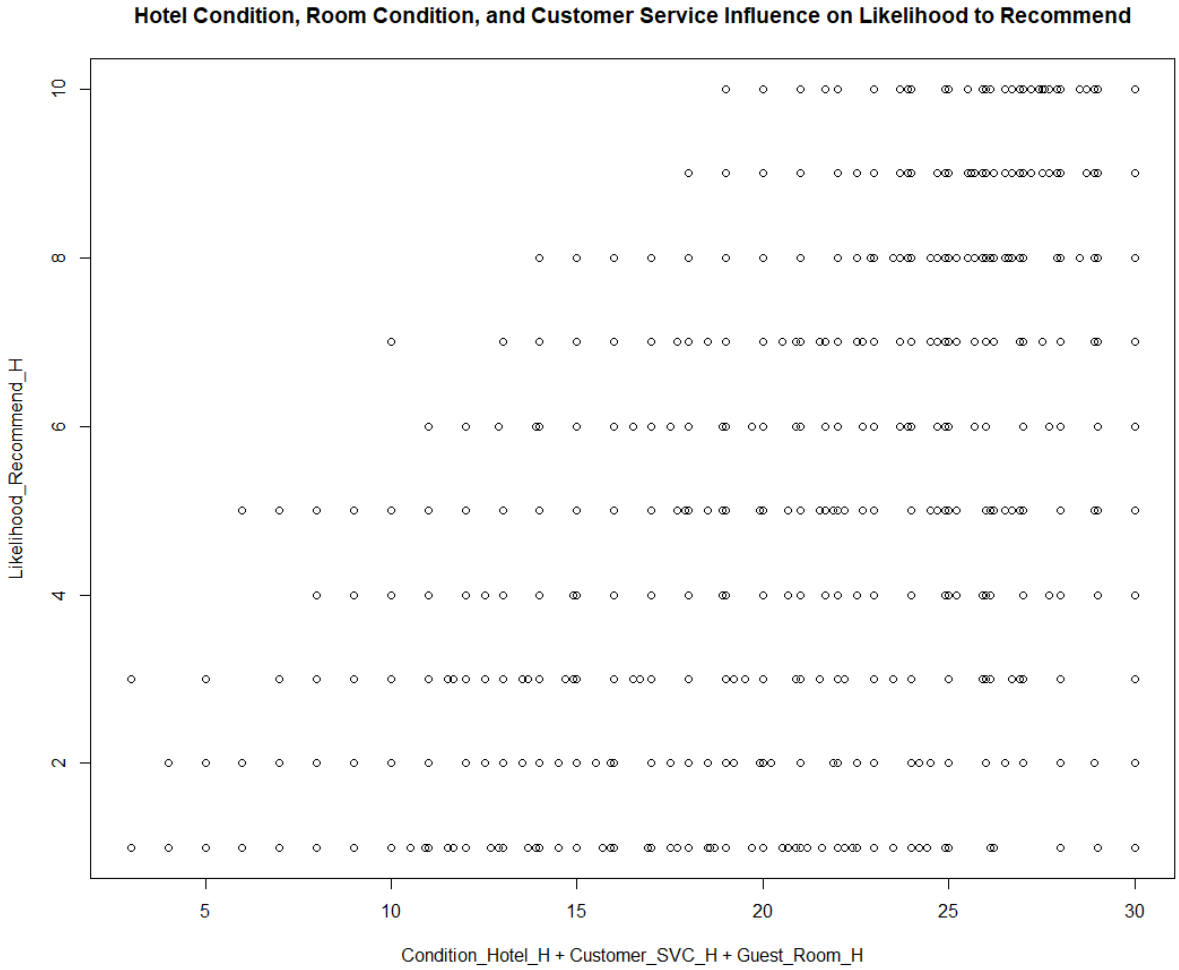
Average daily rate by likelihood to recommend.



Validation

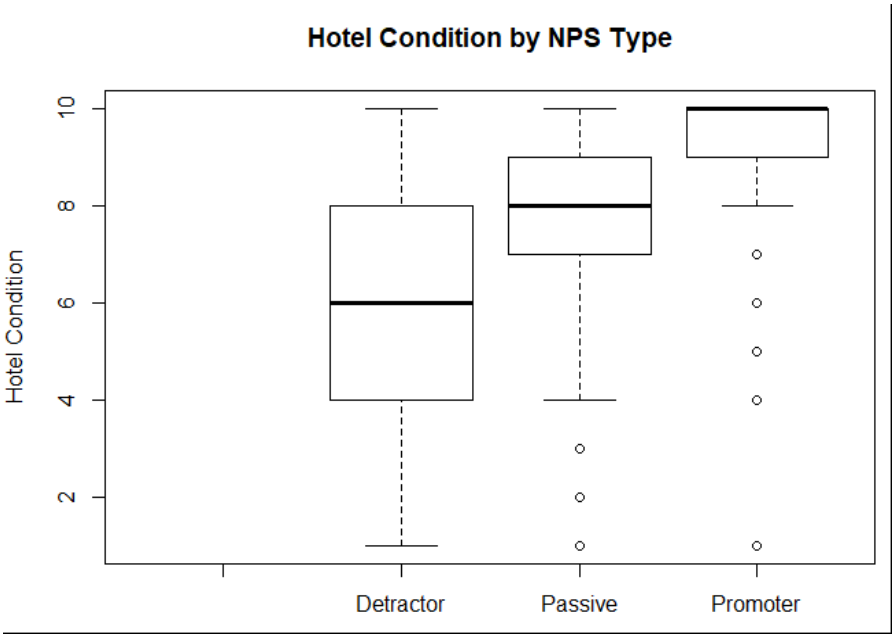
- A plot of the linear modeling results shows that as the scores for Hotel Condition, Guest Room Condition, and Customer Service Satisfaction increase, the Likelihood to Recommend scores increase as shown below.

Influence of hotel condition, room condition, and customer service on likelihood to recommend.



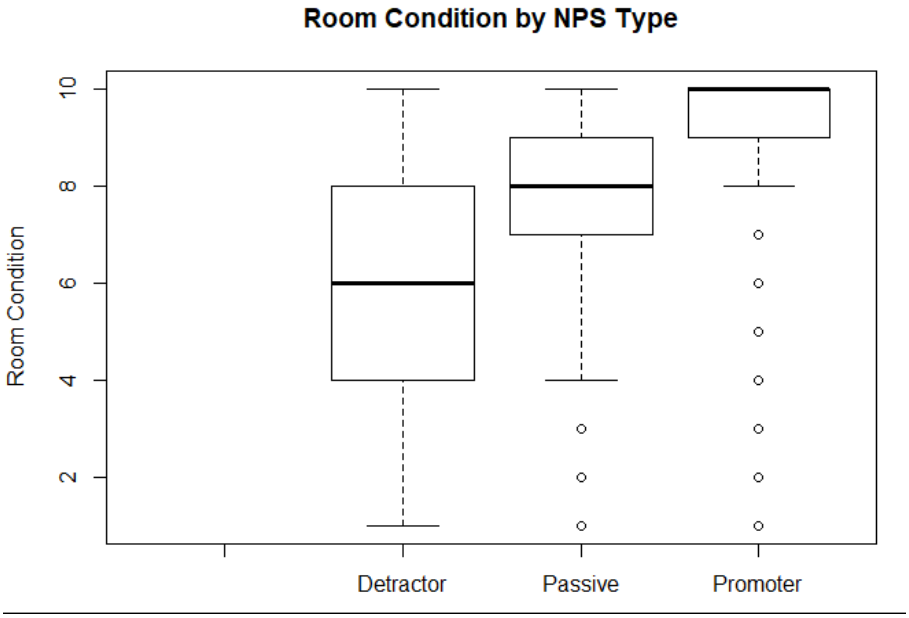
- Using descriptive statistics, we plotted the Hotel Condition scores for each NPS type and noted that Promoters are likely to have much higher scores than Passives or Detractors, as shown below. Also, the outlier values for Promoters were higher than the mean of the Detractors in some cases.

Hotel condition score by NPS type.



- Using descriptive statistics, we plotted the Guest Room Condition scores for each NPS_Type and noted that Promoters are likely to have much higher scores than Passives or Detractors, as shown below. Also, like the plot for Hotel Condition, the outlier values for Promoters were higher than the mean of the Detractors in some cases.

Room condition score by NPS type.



References

Bain & Company. (2017). Measuring your net promoter score. Retrieved from http://www.netpromotersystem.com/about/measuring-your-net-promoter-score.aspx

Hyatt Hotels Corporation. (2017). Hyatt history. Retrieved from https://about.hyatt.com/en/hyatthistory.html