IST 687

Applied Data Science

Hyatt Group Data analysis

Ashish Chabbra

Amol Wani

Farees Patel

Sonali Karamchandani

Introduction

Hyatt group projects had stored around 13 GB of customer data which was given to various data analyst to give out meaningful results so as to increase the NPS of the organization. Generally, when an analyst would look at the data, it would look like a structured mess. Mining useful data from various attributes is one task. Hence it becomes important to select attributes which shows strong relationship with NPS i.e. the final factor to be increased. Hence following steps need to be taken in order to increase the efficiency of algorithm. Here are some steps which were devised so as to analyse huge amount of data:

- **Data cleaning**
- **NPS Calculation**
- **Region Selection**
- **Visualize Data**
- **Finding correlation**
- **Use modelling techniques**
- **Give recommendations based on analysis**

Before stepping towards technical aspect of the project it is important to understand the business requirement of analysis. Here, the company wants to increase sales through referring details provided by the customer. The feedback form has various columns of amenities which needed to be filled by the customer such as spa service, free parking, etc. We have made an attempt to analyse the effects of such amenities on the customers.

Hyatt group of hotels is an international brand, to make the analysis more specific we tried to find the NPS of hotels based on their regions. This region based analysis gave us a chance to formulate more questions and needs of the organization.

Step 1: DATA CLEANING

Since the size of the data was huge, we needed to trim it down so as to make the analysis stage much more time-efficient and also specific to the questions we wanted to focus on. We considered 8 months data i.e. from February 2011 to October 2011. We cleaned the different month's data thereby creating a new dataset which had 60 attributes of the original 237 and eliminated rows which had no value for NPS_Value attribute from all the 8 months data.

While analysing the original datasets, we found that each dataset had about 13-14 lakh tuples out of which only 70,000 - 80,000 rows had a value for the attribute NPS_Value. Since we were focussing on that attribute, we needed to make sure we had enough data which could be analysed using that attribute and also optimize the overall analysis. Hence, we cleaned the data this way thereby satisfying both our requirements.

In our project we have concentrated on 4 major factors:

- Region distribution
- Brand popularity
- Amenities
- Purpose of visit and their contribution towards being a detractor, promoter or a passive customer

Based on these factors, we came up with few major business questions which we think could be important to the client.

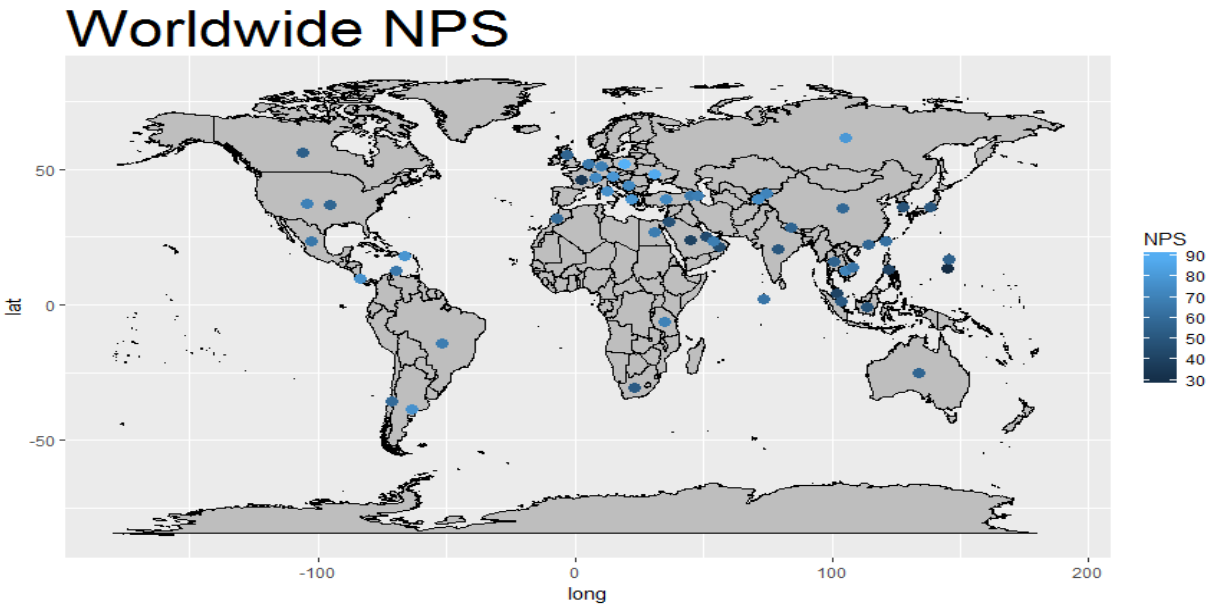
- *What is the region wise NPS of different countries?*
- *What are they doing to keep their NPS really high and meet their NPS goals?*
- *Which purpose of visit needs more focus: Business or Leisure?*
- *What are some interesting facts about various brands of Hyatt? Which one can give better analysis?*
- *Do amenities play a significant role in increasing NPS?*

STEP 2: NPS CALCULATION

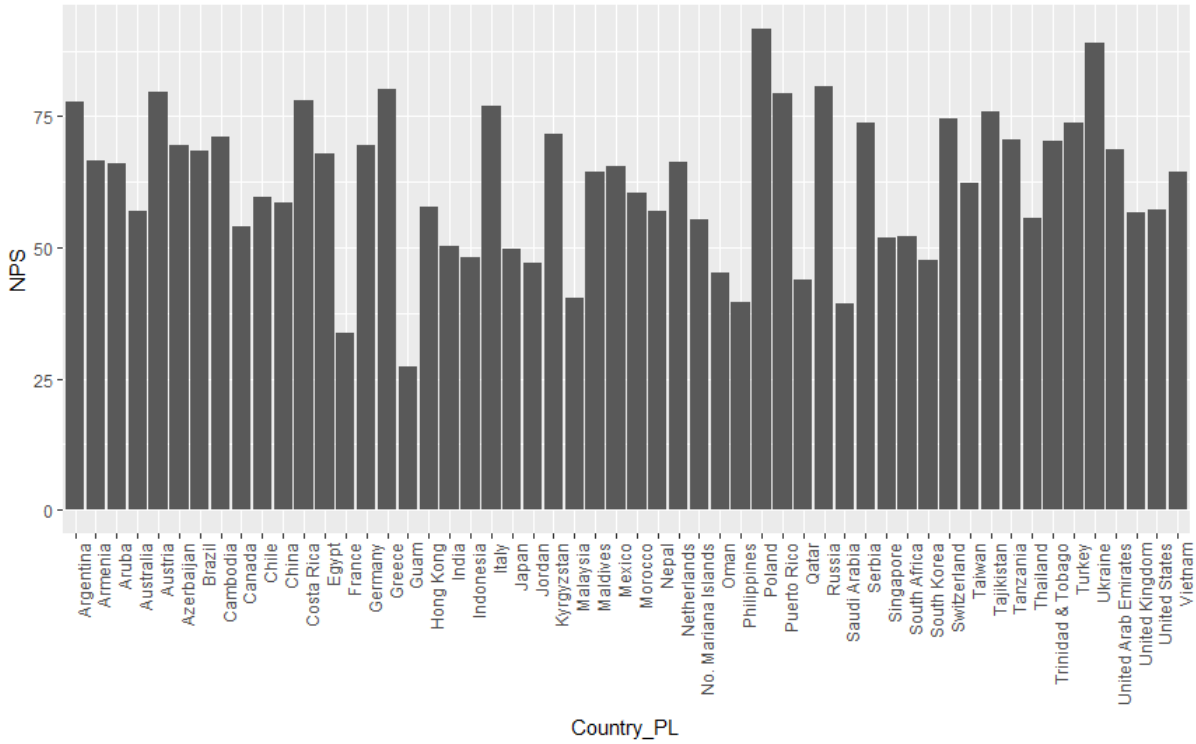
The **Net Promoter Score** is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company's products or services to others. It is used as a proxy for gauging the customer's overall satisfaction with a company's product or service and the customer's loyalty to the brand. In order to calculate the actual NPS of Hyatt Group of hotels we considered the column NPS_TYPE. This column helped us calculate the value for each country based on the no. of promoters, detractors and passives for the respective country. For NPS calculation, we used the formula:

Actual NPS = ((No. of Promoters – No. of Detractors) / (No. of respondents)) * 100

After obtaining the NPS values for each country, we plotted it on a world map so as to get a rough idea as to which countries need to be concentrated upon for improvements and which countries could be considered as an example of excellence and success.



We plotted a Bar plot in order to get an exact idea as to how the countries fared in comparison to each other on the basis of their NPS Value.

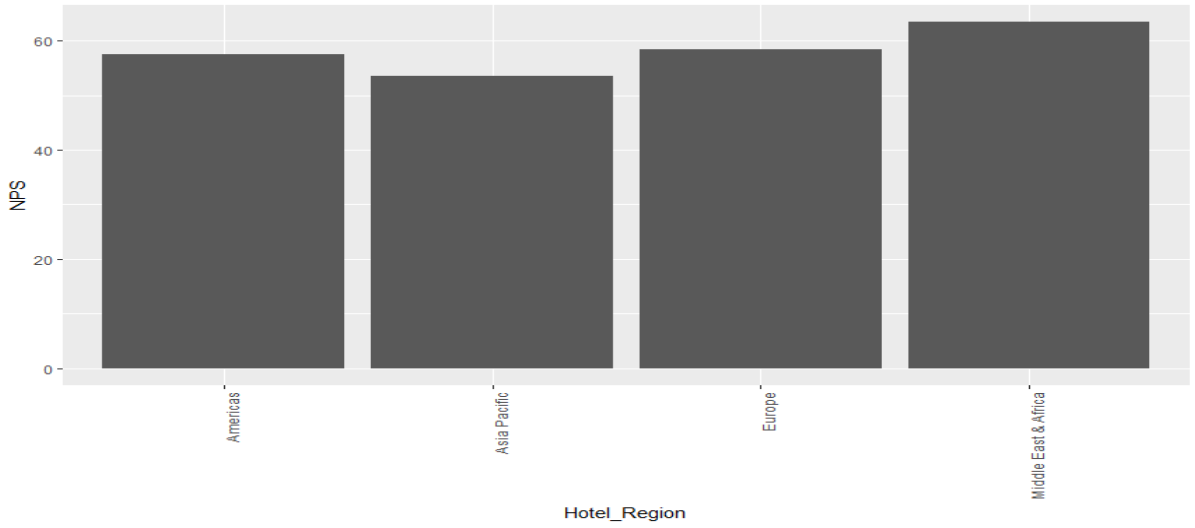


A few facts which can be observed from this plot have been stated below:

- Poland stands out from the list of the countries and has the highest NPS of 91.6
- Guam has the lowest NPS of 27.58
- Total of 28 countries were found to have their NPS score more than their Goal Values
- The average NPS throughout the countries is 57.7

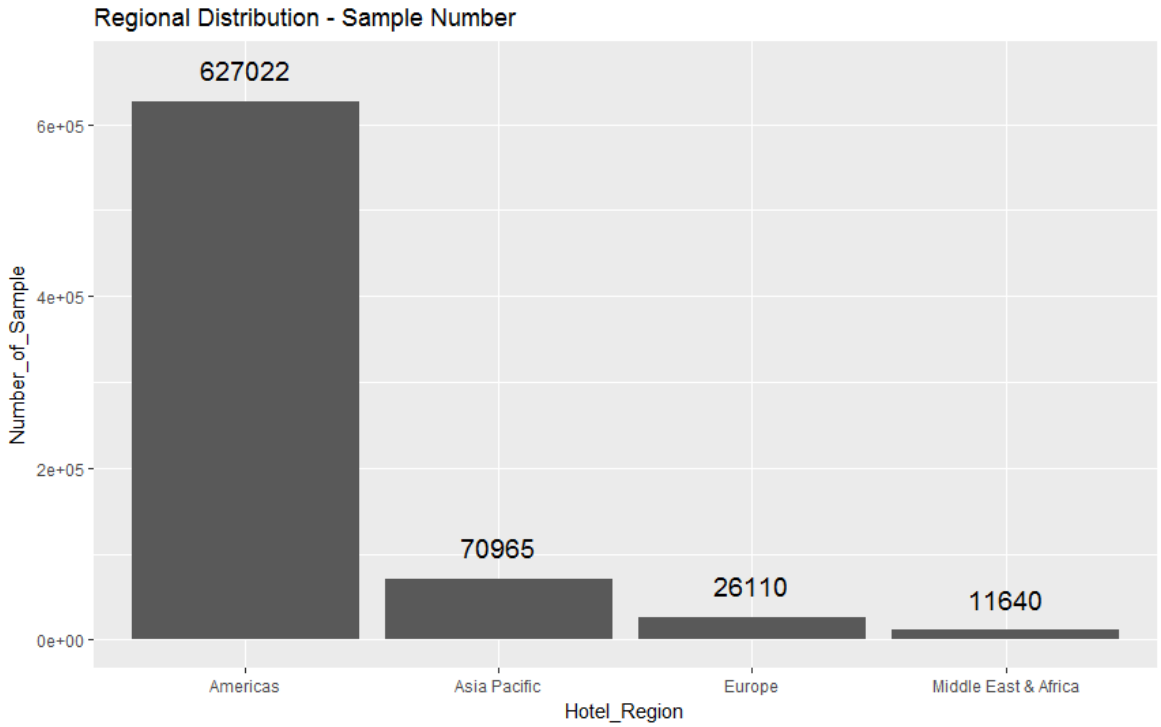
STEP 3: REGION SELECTION

We plotted a barplot which shows how each region fares in term of their overall NPS.



- Middle East and Africa regions have highest region wise NPS value
- Second largest NPS value in accordance to region is that of Europe.

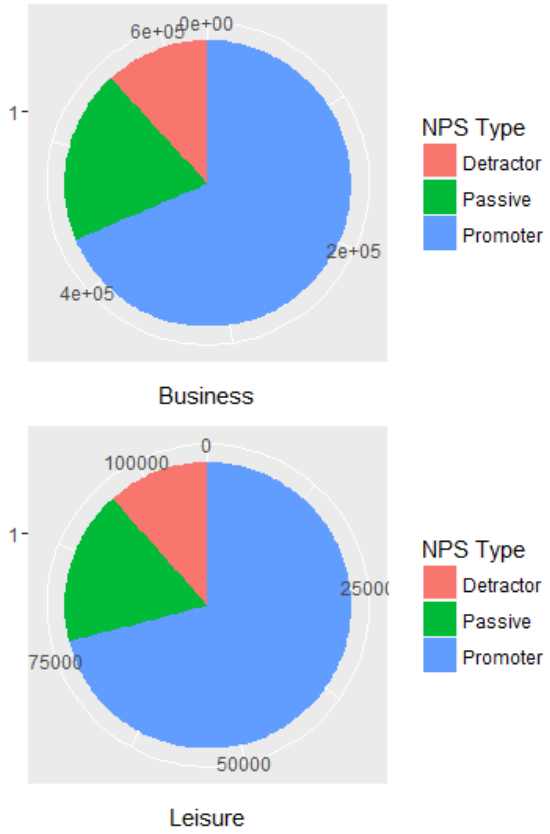
When we gave a closer look to the data, we wanted to consider the major contributors to the NPS calculation. By doing so, we made sure that our analysis reaches corners so that we can explore more and more data to find interesting correlations at further stages. We found that Americas region had the maximum number of feedbacks which were almost 10 times compared to other countries.



Since Middle East and Africa had the best NPS value compared to other regions, we selected it as one the regions which we would be focussing on in our analysis. Also, since Americas region had the highest number of feedbacks, it would be much more interesting to analyse it. Hence, we selected the Americas region as well to be focussed upon.

STEP 4: VIZUALIZATION OF DATA

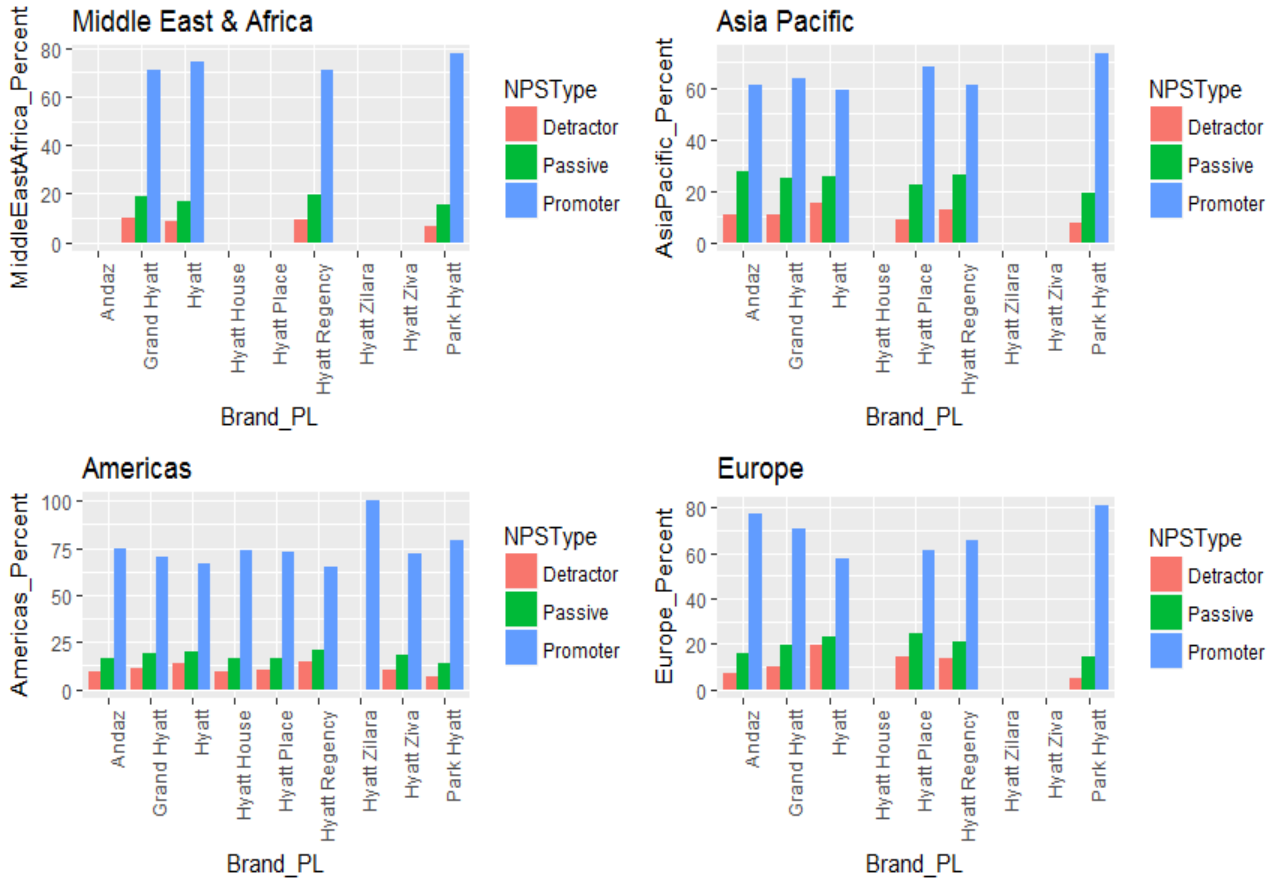
Visualizing data to understand its clarity and correlation is best way to answer various business questions. When we analysed the dataset, we found that the two major purposes of visit of customers were either Business or Leisure. We wanted to analyse this aspect a bit more in terms of how much did these two contribute towards the NPS Value for the countries. The pie chart below can be used to explain our findings.

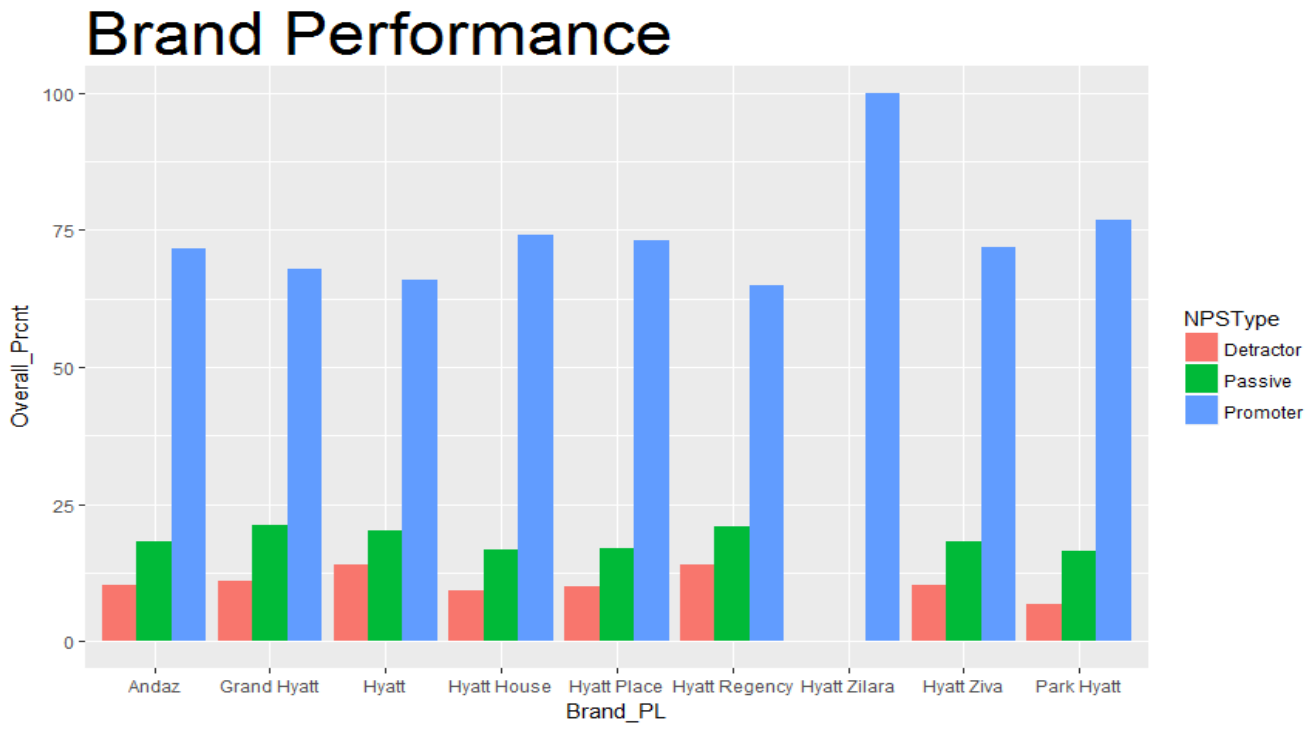


The pie chart is divided into three categories which are promoters, detractors and passives.

It can be seen that the percentage of promoters for Leisure is more compared to Business whereas the amount of passives for Business is more compared to Leisure. The percentage of detractors is almost same for both. The key finding in this analysis is that the customers who come for business trips are not bothered about the facilities provided by the hotel. Hence, most of the time, they give a neutral review i.e. are passives. If such customers are focussed upon and steps are taken to move them to the promoter category, the NPS value could get a considerable boost.

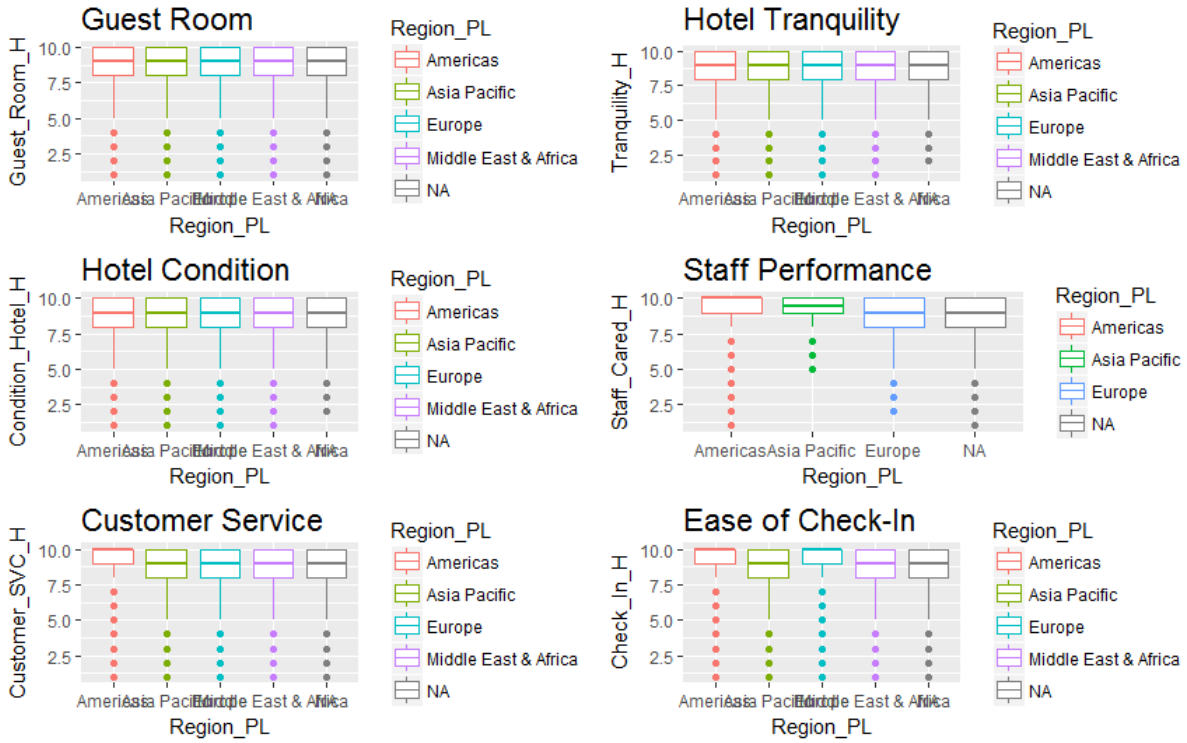
On completing NPS calculation for various aspects it is important to know what Hyatt is doing in each of its brands and how popular it is to various customers. Hyatt is parent to different genre of hotels like Andaz, Grand Hyatt, Hyatt, Hyatt House, Hyatt palace, Hyatt Regency, Hyatt Zilara, Hyatt Ziva and Park Hyatt. We used these brands to understand what percentage of promoters, detractors or passive customers are inclined towards which kind of brand. This analysis narrowed down to regions so as to keep the classification less complicated to regional level. Here we have used the column BRAND_PL so as to understand its relationship with the column REGION_PL.


Middle East & Africa
MiddleEastAfrica_Percent
0
20
40
60
80
Andaz
Grand Hyatt
Hyatt
Hyatt House
Hyatt Place
Hyatt Regency
Hyatt Zilara
Hyatt Ziva
Park Hyatt
Brand_PL
NPSType
Detractor
Passive
Promoter
Asia Pacific
AsiaPacific_Percent
0
20
40
60
Andaz
Grand Hyatt
Hyatt
Hyatt House
Hyatt Place
Hyatt Regency
Hyatt Zilara
Hyatt Ziva
Park Hyatt
Brand_PL
NPSType
Detractor
Passive
Promoter
Americas
Americas_Percent
0
25
50
75
100
Andaz
Grand Hyatt
Hyatt
Hyatt House
Hyatt Place
Hyatt Regency
Hyatt Zilara
Hyatt Ziva
Park Hyatt
Brand_PL
NPSType
Detractor
Passive
Promoter
Europe
Europe_Percent
0
20
40
60
80
Andaz
Grand Hyatt
Hyatt
Hyatt House
Hyatt Place
Hyatt Regency
Hyatt Zilara
Hyatt Ziva
Park Hyatt
Brand_PL
NPSType
Detractor
Passive
Promoter


Brand Performance
Overall_Prcnt
0
25
50
75
100
Andaz
Grand Hyatt
Hyatt
Hyatt House
Hyatt Place
Hyatt Regency
Hyatt Zilara
Hyatt Ziva
Park Hyatt
Brand_PL
NPSType
Detractor
Passive
Promoter

The above analysis shows how Hyatt Zilara has most number of promoters and Hyatt Regency has least. The facts became more interesting when we ran an analysis on Hyatt Regency since it has comparatively less number of promoters than others and slightly higher number of detractors than brands. With such pattern we noticed that Hyatt Regency has most number of data, and due to this the results can be much more accurate.

After selecting the brand that we need to focus on, we compared that with number of factors to find obvious, but verified findings.

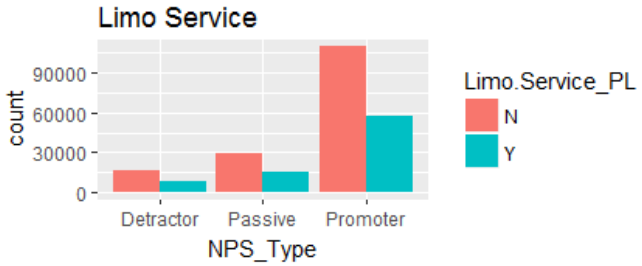


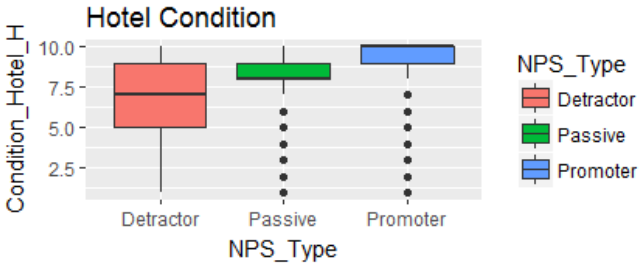
This analysis is solely for Hyatt Regency for different regions. From this analysis we can conclude that, Guest Room, Hotel Tranquillity and Hotel Conditions in each region if improved can have a better NPS and could bring about more promoters. Other factors seem to do fine, but there is always scope for improvement.

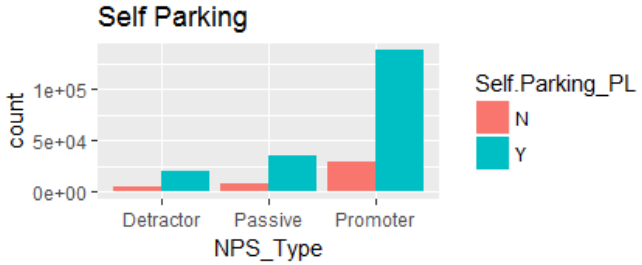
We mentioned only those visualizations which gave us some interesting finding during our presentation. Below are the ones which we plotted but didn't use it since it wasn't enough for us to give conclusive evidence for a recommendation to be made.

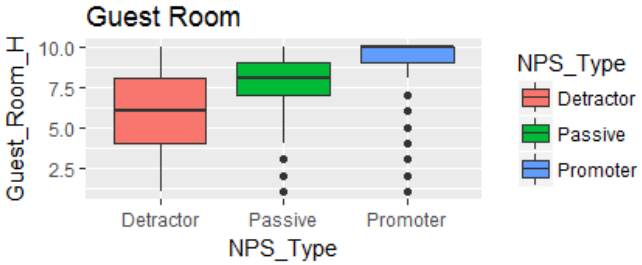
The plots shown below give us an idea as to how much do the secondary amenities contribute to the the NPS value of the region. Different graphs and plots were used to understand more about the secondary amenities. These plots revealed the fact that the customers which used these facilities were more likely to turn out as promoters and those who didn't would turn out to be detractors.

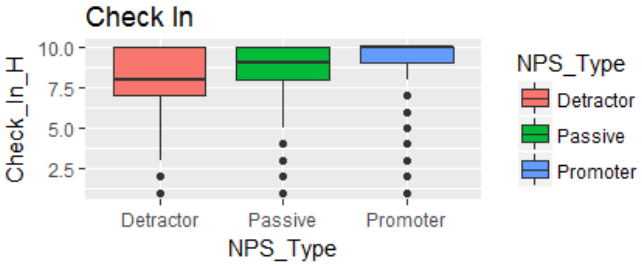
1

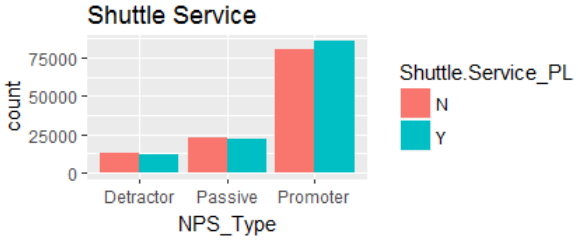

Limo Service
count
90000
60000
30000
0
Detractor
Passive
Promoter
NPS_Type
Limo.Service_PL
N
Y

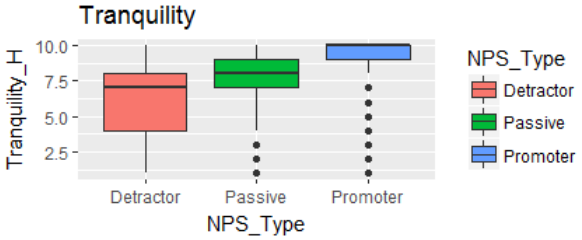

Hotel Condition
Condition_Hotel_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter

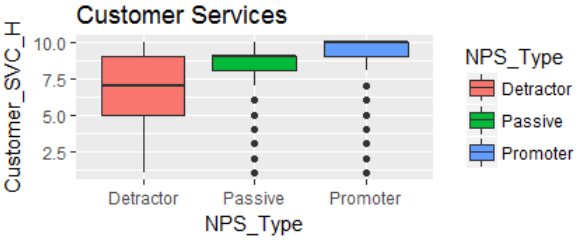

Self Parking
count
1e+05
5e+04
0e+00
Detractor
Passive
Promoter
NPS_Type
Self.Parking_PL
N
Y

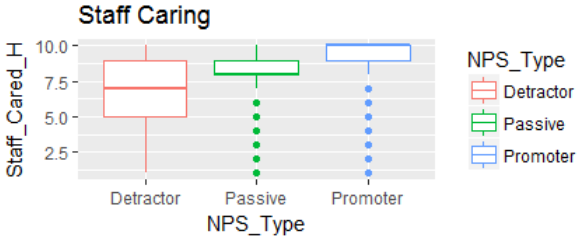

Guest Room
Guest_Room_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter

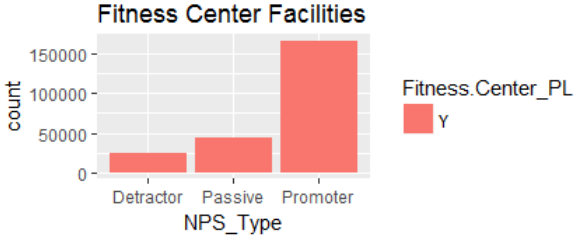

Check In
Check_In_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter

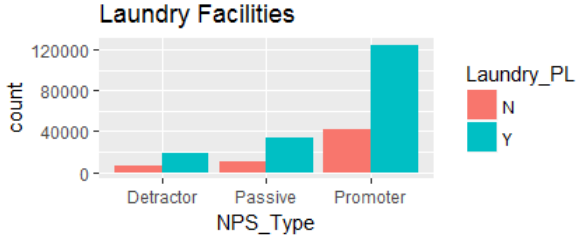

Shuttle Service
count
75000
50000
25000
0
Detractor
Passive
Promoter
NPS_Type
Shuttle.Service_PL
N
Y

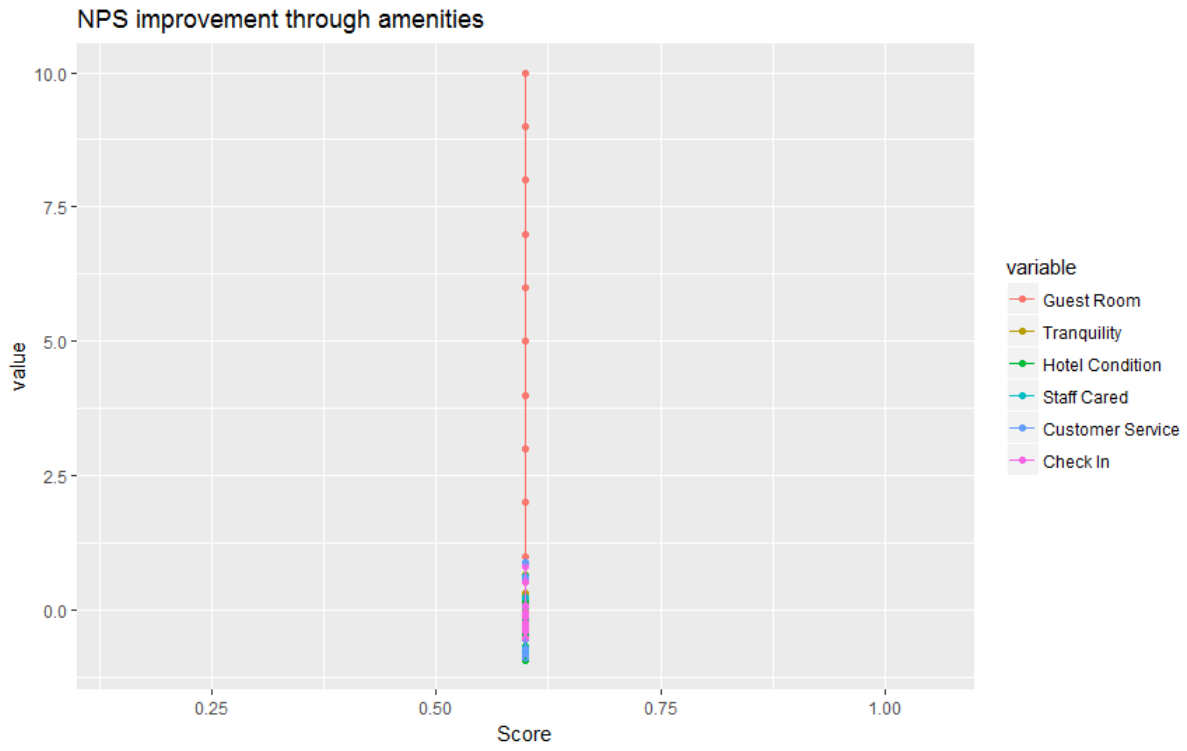

Tranquility
Tranquility_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter


Customer Services
Customer_SVC_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter

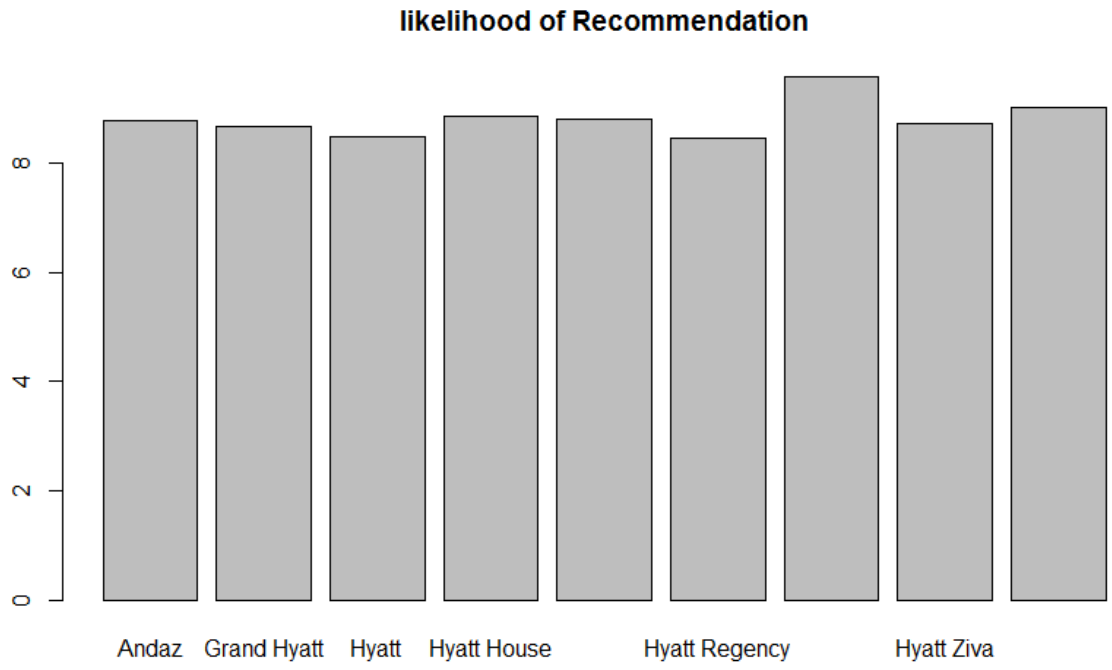

Staff Caring
Staff_Cared_H
10.0
7.5
5.0
2.5
Detractor
Passive
Promoter
NPS_Type
NPS_Type
Detractor
Passive
Promoter


Fitness Center Facilities
count
150000
100000
50000
0
Detractor
Passive
Promoter
NPS_Type
Fitness.Center_PL
Y


Laundry Facilities
count
120000
80000
40000
0
Detractor
Passive
Promoter
NPS_Type
Laundry_PL
N
Y

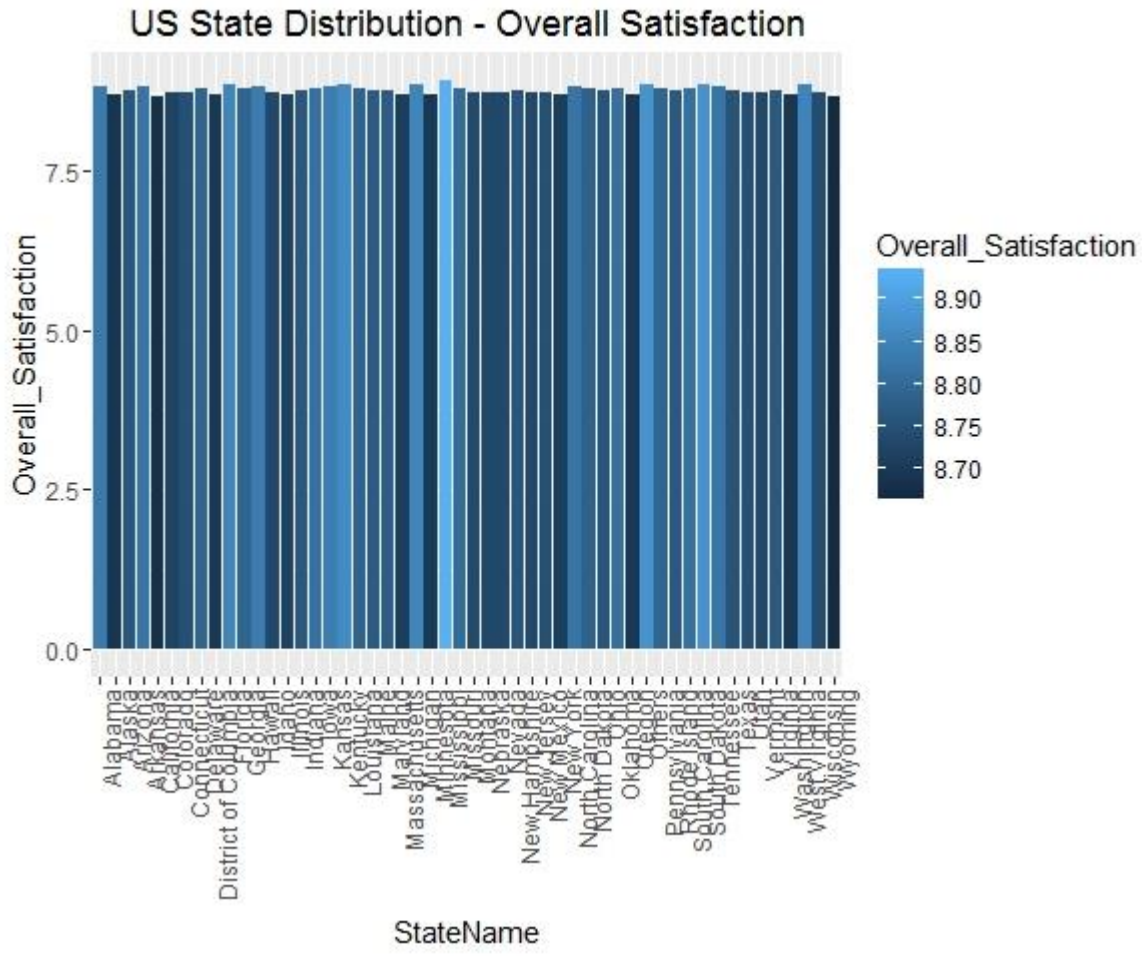


The plot below gives us an idea which brand of Hyatt group of hotels would most likely be recommended. This analysis was carried throughout all regions data and it was found the Hyatt Regency had the highest likelihood to be recommended.



In the barplot below, we tried to analyse the overall satisfaction factor for all the states throughout United States. We had tried this in order to understand whether we could have

focussed on specific cities who's NPS could be improved. We dropped the idea later on and focussed only on the regions.



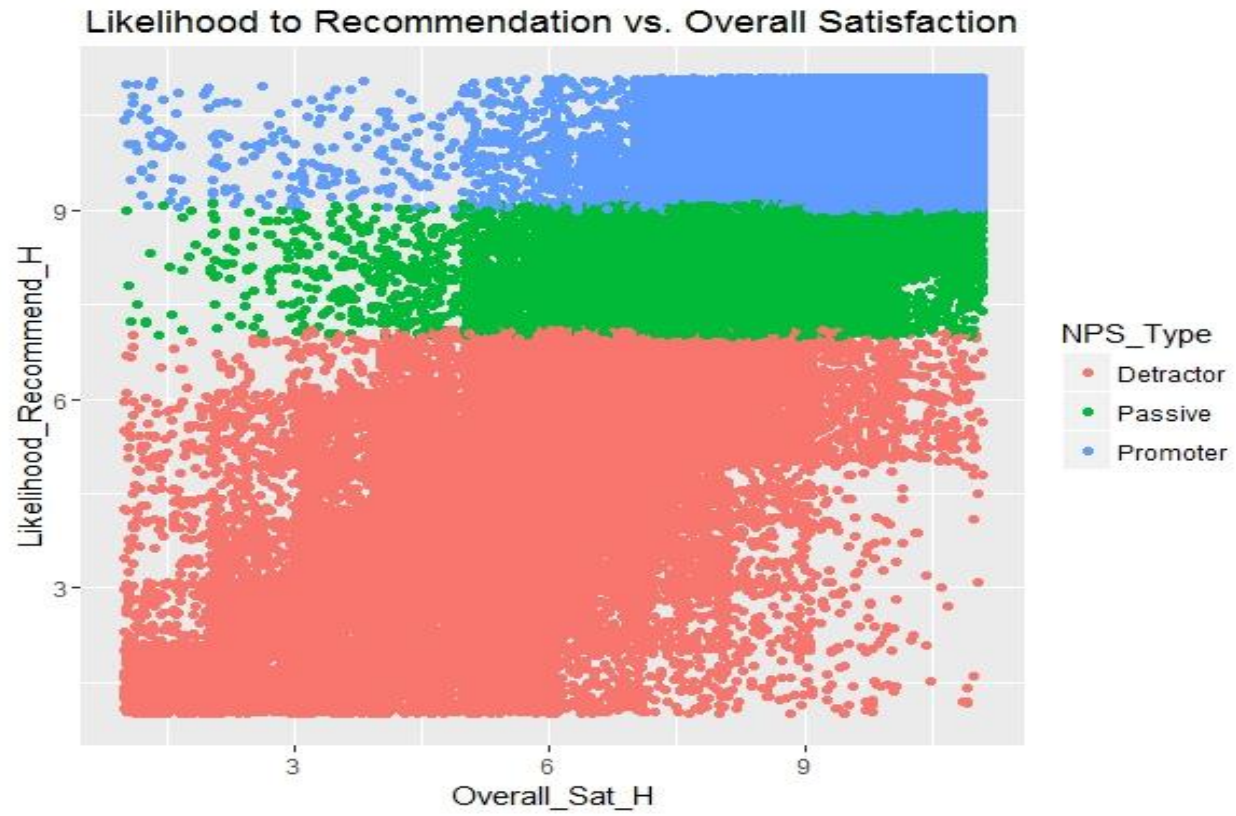
STEP 5: FIND CORRELATION BETWEEN VARIOUS COLUMNS

Correlations means to have a relationship between one or more items. If the correlation is strong then analysis can be carried on by using dependencies.

How did we go about this analysis?

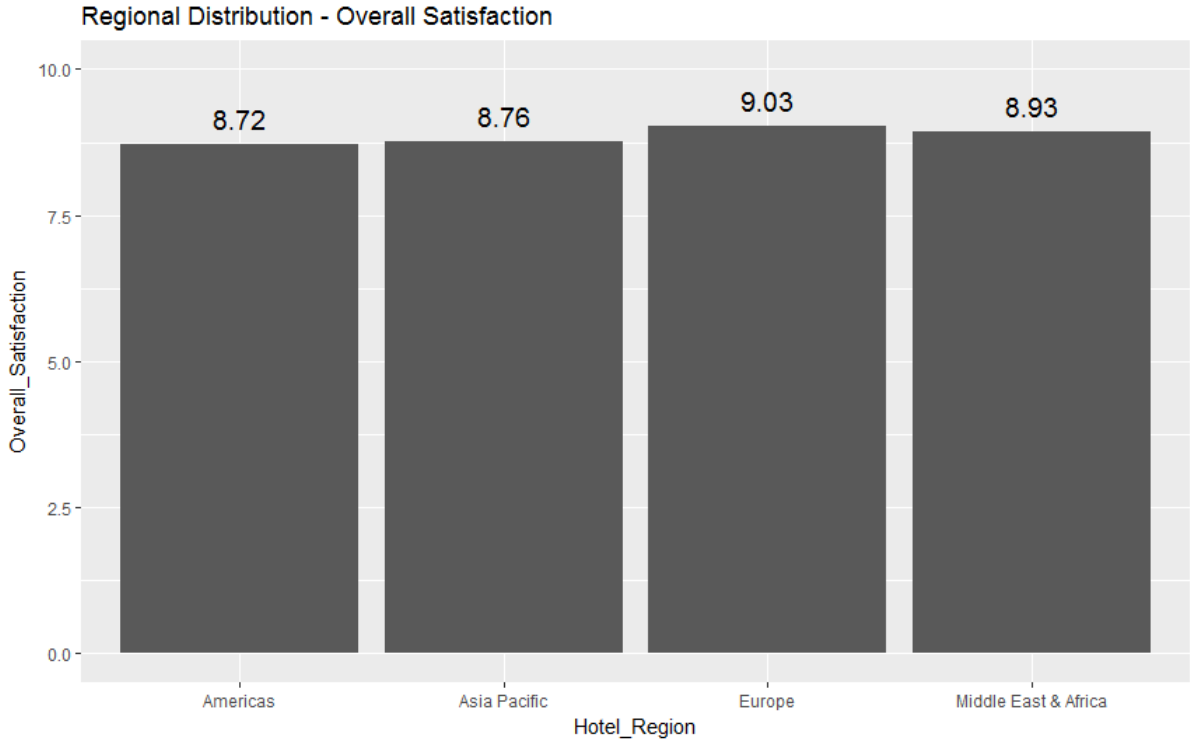
Now why perform correlation between Likelihood to Recommendation VS. overall satisfaction?

Based on the given data, for region wise analysis we figured that Middle East and Africa have highest NPS. The main factor to calculate NPS is Likelihood to Recommendation. From the graph we can see that it is high then people are likely to recommend the hotel. Here overall satisfaction is an independent variable.



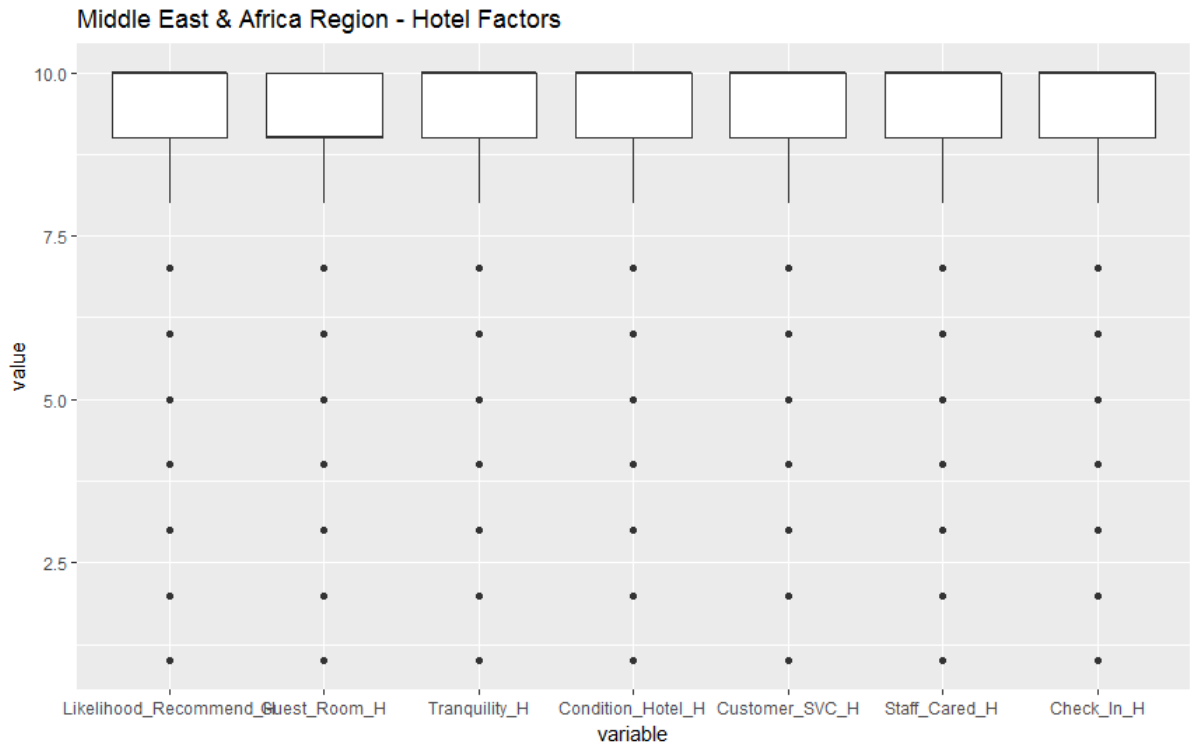
Regional distribution of for Overall Satisfaction

Why do this?
Overall satisfaction of various regions has been plotted here.

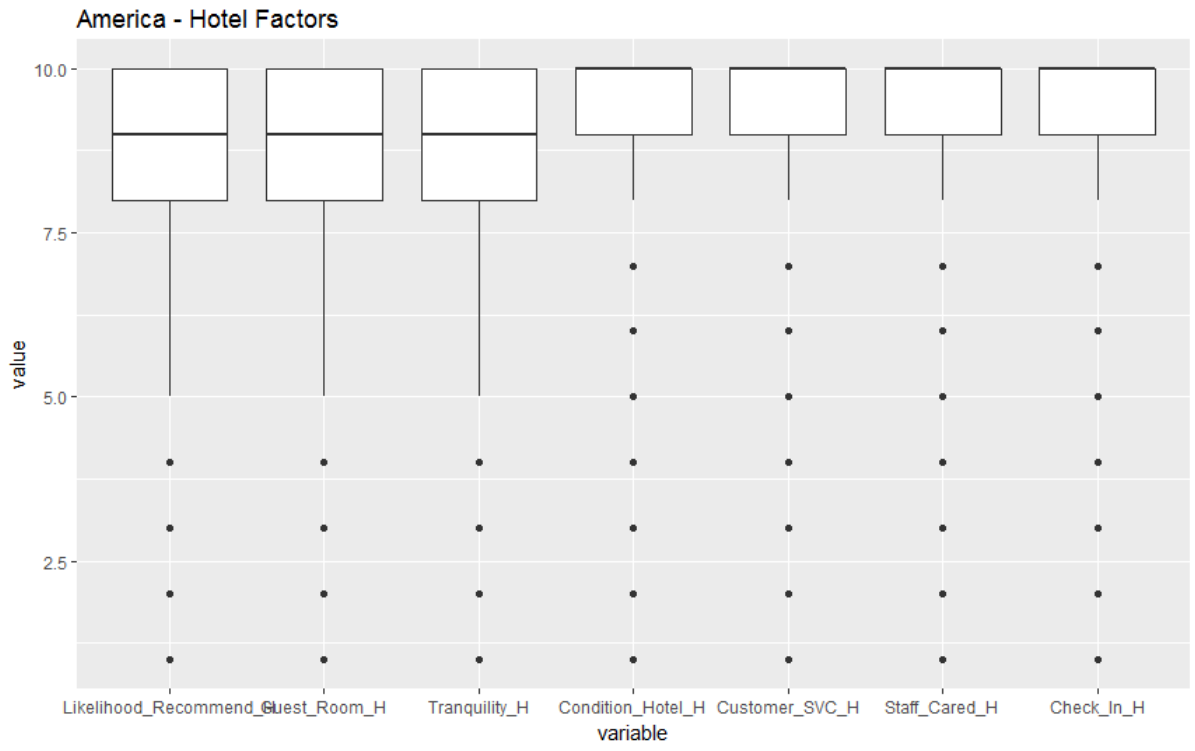


For our further analysis, we have used Middle East & Africa since they have highest NPS and America since they have lowest overall satisfaction. This made it easier for us to target regions which we need to pay attention to.

MIDDLE EAST & AFRICA



AMERICA



From comparing the two plots it can be seen that Americas low overall satisfaction can be result of low Likelihood to Recommend, Guest Rooms and Tranquillity which is lower than that of Middle East and Africa region.

Final correlation between various columns.

```
                      Likelihood_Recommend_H Guest_Room_H Tranquility_H Condition_Hotel_H Customer_SVC_H Staff_Cared_H
Likelihood_Recommend_H             1.00000000   0.68877570    0.57028031        0.62157261     0.75293462   0.653893534
Guest_Room_H                       0.68877570   1.00000000    0.64467899        0.78425986     0.65272207   0.537611497
Tranquility_H                      0.57028031   0.64467899    1.00000000        0.57106924     0.50654440   0.420317387
Condition_Hotel_H                  0.62157261   0.78425986    0.57106924        1.00000000     0.58342376   0.479957015
Customer_SVC_H                     0.75293462   0.65272207    0.50654440        0.58342376     1.00000000   0.809052334
Staff_Cared_H                      0.65389353   0.53761150    0.42031739        0.47995701     0.80905233   1.000000000
Check_In_H                         0.48695081   0.41670082    0.31719331        0.41765451     0.56915500   0.527210011
Amenity                            0.02330576  -0.03793637   -0.03143061       -0.02436138     0.03065717   0.004682554
                       Check_In_H      Amenity
Likelihood_Recommend_H 0.486950812  0.023305763
Guest_Room_H           0.416700824 -0.037936375
Tranquility_H          0.317193312 -0.031430612
Condition_Hotel_H      0.417654510 -0.024361381
Customer_SVC_H         0.569154999  0.030657170
Staff_Cared_H          0.527210011  0.004682554
Check_In_H             1.000000000  0.006471117
Amenity                0.006471117  1.000000000
```

Here we pick top 3 strongly correlated items in the data frame:

- **Staff cared and Customer service with the value of 0.80905233**
- **Condition of the hotel and Guest room with the value of 0.78425986**
- **Customer service and likelihood to recommend with the value of 0.75293462**

Now next step would be to perform meaningful modelling on factors with high correlation

STEP 5: USE MODELLING TECHNIQUES (Linear and SVM Modelling)

For hotels of Middle East & Africa region, we conducted Linear modelling on Hotel where people have used amenities and where customers have not used any kind of amenities.

```
> lmModel2 <- lm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H +
+                  Customer_SVC_H + Staff_Cared_H + Check_In_H + Amenity, data=hotelMeaSat)
> summary(lmModel2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.9671 -0.0564  0.0350  0.3646  3.4408

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.73642    0.26626  -2.766  0.00580 **
Guest_Room_H       0.21719    0.03940   5.513 4.70e-08 ***
Tranquility_H      0.13242    0.02688   4.926 1.01e-06 ***
Condition_Hotel_H  0.11190    0.04142   2.701  0.00704 **
Customer_SVC_H     0.44663    0.04610   9.688  < 2e-16 ***
Staff_Cared_H      0.11755    0.03687   3.189  0.00148 **
Check_In_H         0.04217    0.02331   1.809  0.07085 .
Amenity            0.01143    0.00930   1.229  0.21938
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8902 on 846 degrees of freedom
Multiple R-squared:  0.6564,    Adjusted R-squared:  0.6535
F-statistic: 230.9 on 7 and 846 DF,  p-value: < 2.2e-16
```

The value of R squared with amenities is 0.6564

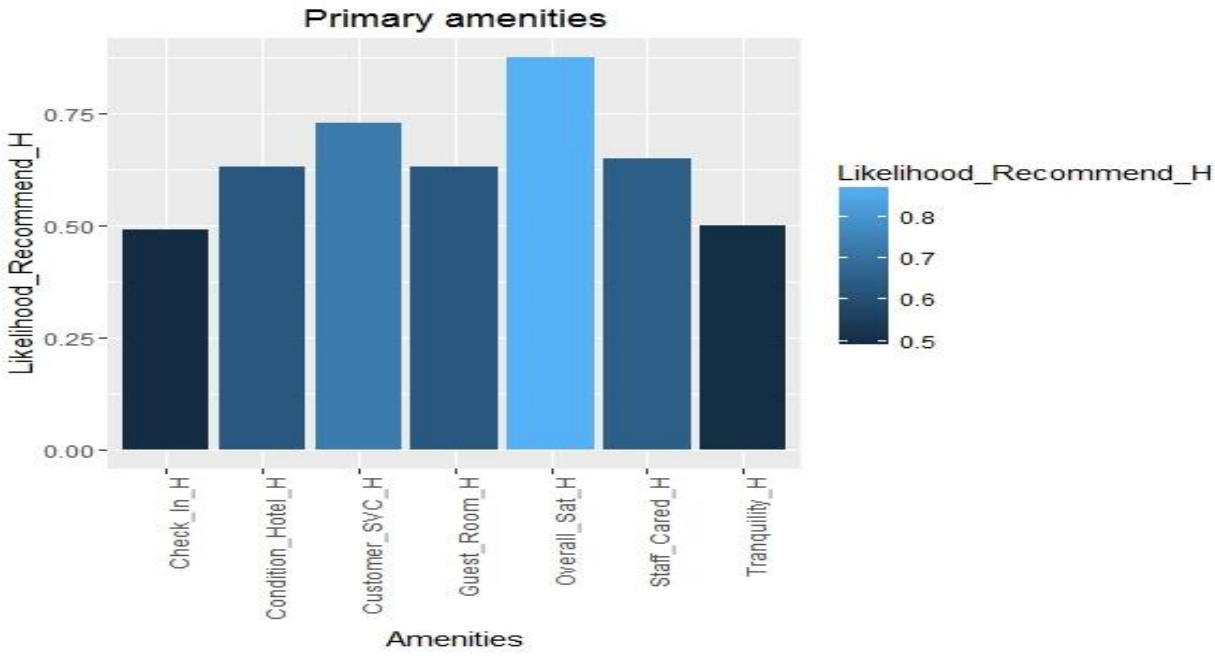
```
> lmModel1 <- lm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H +
+                  Customer_SVC_H + Staff_Cared_H + Check_In_H, data=hotelMeaSat)
> summary(lmModel1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.9175 -0.0472  0.0012  0.3623  3.4620

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.67774    0.26202  -2.587  0.00986 **
Guest_Room_H       0.21493    0.03937   5.460 6.27e-08 ***
Tranquility_H      0.13176    0.02689   4.901 1.14e-06 ***
Condition_Hotel_H  0.11192    0.04144   2.701  0.00705 **
Customer_SVC_H     0.45085    0.04599   9.804  < 2e-16 ***
Staff_Cared_H      0.11611    0.03686   3.150  0.00169 **
Check_In_H         0.04207    0.02332   1.804  0.07157 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8905 on 847 degrees of freedom
Multiple R-squared:  0.6558,    Adjusted R-squared:  0.6533
F-statistic: 268.9 on 6 and 847 DF,  p-value: < 2.2e-16
```

The value of Adjusted R squared for linear modelling without amenities is 0.6558

From this comparison is it clear that higher the value of R squared, higher is the probability of data fitting the regression line. In Layman terms, NPS recorded is better when a customer who has used to amenities is recommending. Also, NPS is encountered to be lower if the customer has never used any of the amenities.

SVM modelling

We first decided to analyse on what factors does the Likelihood to recommendation is dependent on. Using the correlation function, we were able to find some of the amenities which are highly correlated with the Likelihood to recommendation. These amenities were categorized into primary activities because irrespective of the customer’s purpose of travel, whether it be for business or leisure the primary amenities are going to have the similar correlation. Hence, the amenities that are highly correlated are categorized to primary amenities and the rest of the amenities are categorized into secondary amenities.



Further, while analysing data using modelling techniques, we were able to determine that secondary amenities too play an important role in deciding whether the customer is either a promoter or a detractor.

Output:

```
                      svmPred.3...
testData....NPS_Type..    0    1    2
             Detractor   88   23   30
             Passive     80  136  169
             Promoter    25   76 1561
> |
```

Prediction Efficiency= [88+136+1561] / [88+23+30+80+136+169+25+76+1561] = 81.5%~ 82%

So, here we can see from analysing Model 1 for which all we have considered for analysis was just the primary amenities and we were able to draw a conclusion based on our findings that if a customer just utilizes the primary amenities and not the secondary amenities. The Hotel is able to categorize him either as a promoter or a detractor correctly 82% of the times.

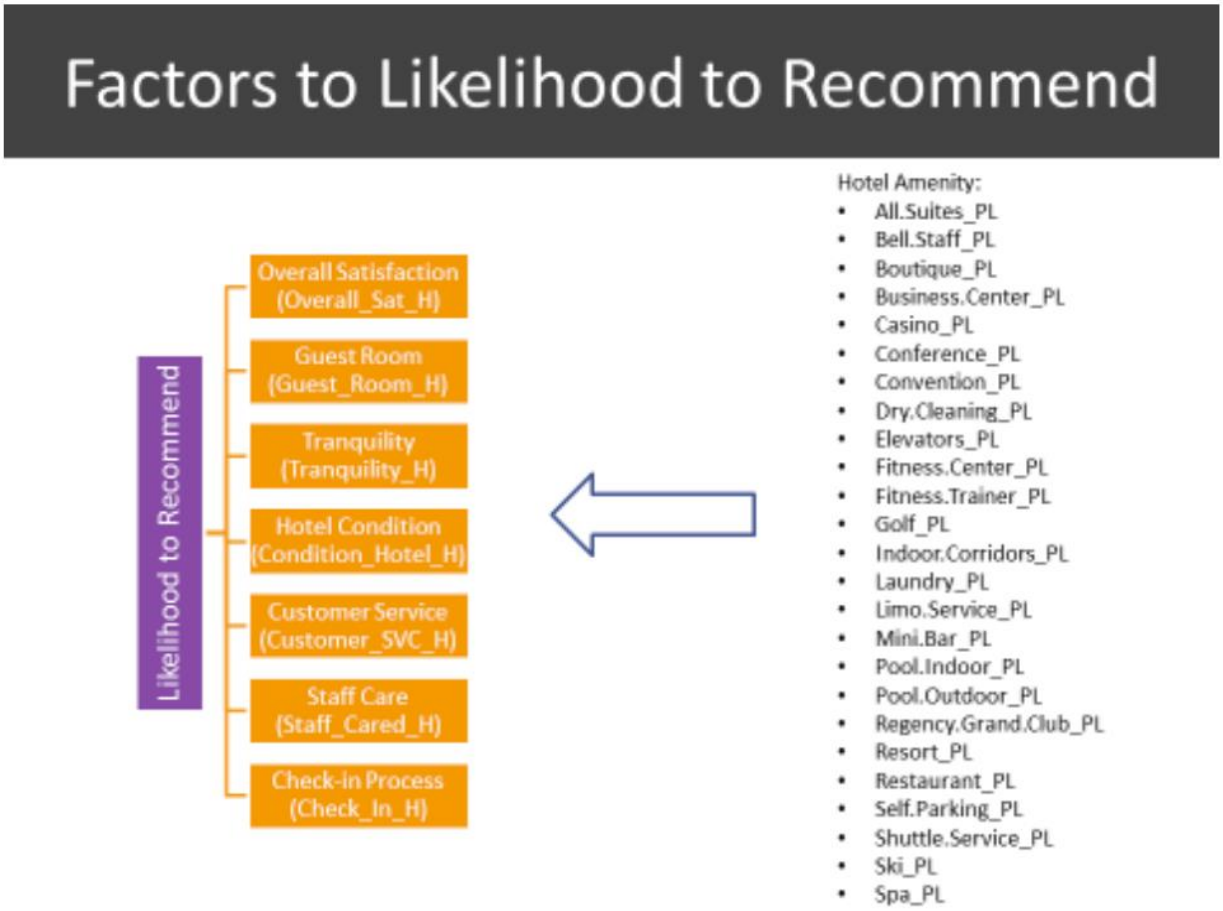
Output:

```
                      svmPred.3...
testData....NPS_Type..    0    1    2
             Detractor  141    0    0
             Passive    133  252    0
             Promoter     1    0 1661
> |
```

Prediction efficiency= [141+252+1661] / [141+0+0+133+252+0+1+0+1661] = 94%

Whereas, while analysing Model 2 for which all we have considered for analysis is the primary and the secondary amenities and we were able to draw a conclusion based on our findings that if a customer just utilizes the primary amenities and not the secondary amenities. The Hotel is able to categorize him either as a promoter or a detractor correctly 81.5% of the times.

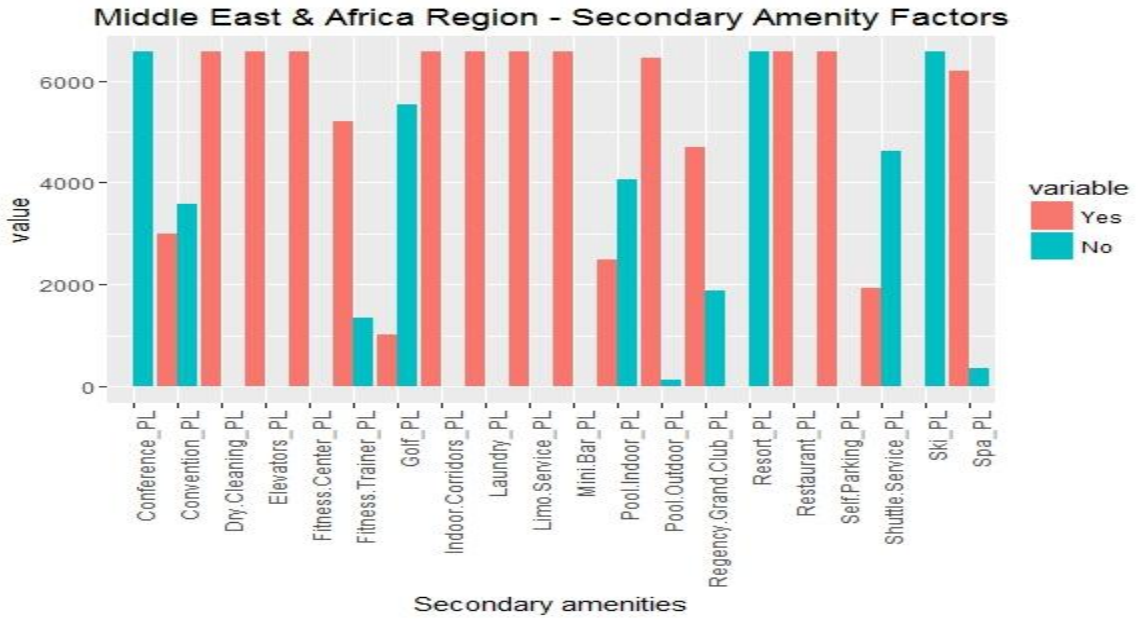
Hence we can summarize that that Likelihood to recommendation is dependent on primary as well as secondary amenities. The below image will give a broad picture of how primary, secondary and likelihood to recommendation are correlated.

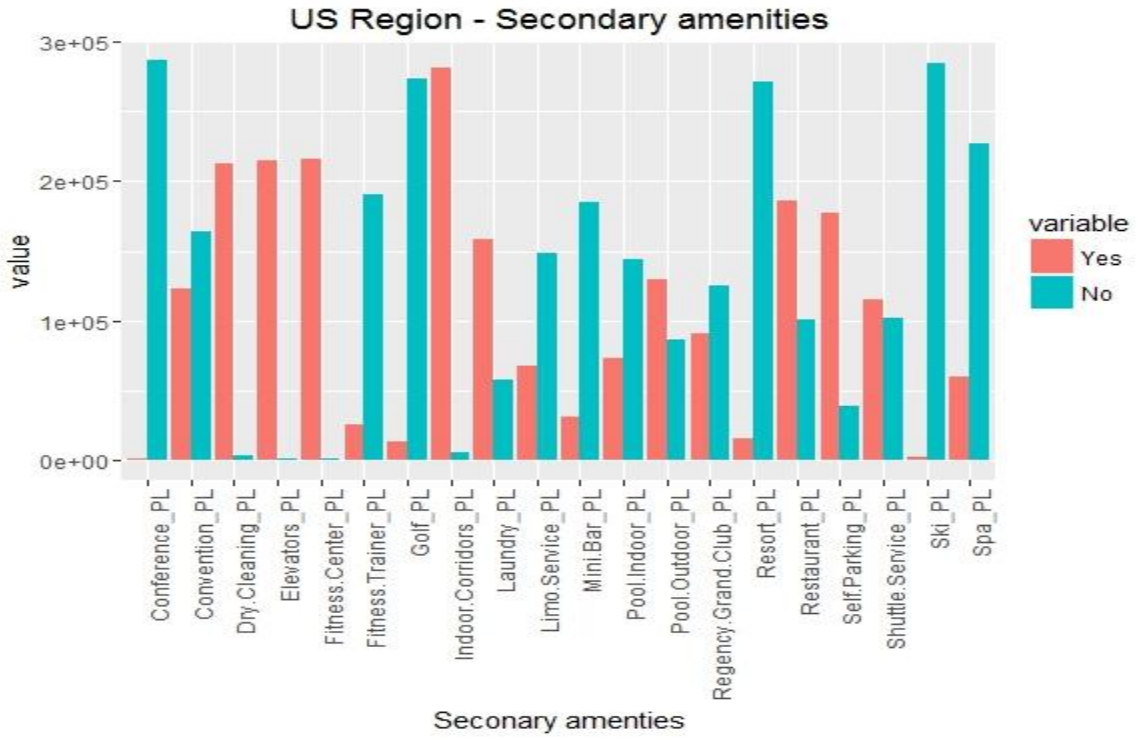


Recommendations resulting from SVM modelling:

Since, you can analyse Middle east & Africa is doing good because of the high quality of primary amenities that they offer to their customers. So America should implement the same quality of service. As we can see one of the differences that curbs America from achieving its

NPS goal is because of the condition of the hotel room and the Tranquillity. If America can somehow try to hire more staff and monitor room conditions on a daily basis, they can improve on this vertical. Similarly, America can reinforce building sound proof rooms so that customers' can have peace in the hotel room. By focusing on improving the primary amenities, Hyatt America can well be on their way to reach their goal NPS.





Similarly, for secondary amenities, we can analyse that customer's accommodating in the Middle East have utilized secondary amenities more as compared to customers' from Hyatt

America. Using the above SVM modelling we found that customers' using secondary amenities are more likely to recommend the hotel. On the other hand, if we compare the bar plot of Hyatt America, most of the customers' have not utilized the secondary amenities and hence are unaware about Hyatt offerings. So if Hyatt America is able to offer some perks or complimentary secondary amenities for their customers to increase their overall satisfaction and likelihood to recommendation eventually increasing the NPS.

STEP 6: PROVIDE RECOMMENDATIONS

For recommendations to the hotel

- **They can include referral offers, the person who has referred the other person can either get some special offers like discount coupons or complimentary secondary amenities.**
- **Customers traveling for business purpose should be focused upon in spite of their relatively shorter stays. Focusing on customers on a business trip by providing them with perks and allowances in order to gain their loyalty towards the hotel could also lead to increase in the NPS score of the hotel**
- **America needs to improve on low Likelihood to Recommend, Guest Rooms and Tranquillity.**
- **The Hotels who fail to reach their goal NPS score should focus on improving their amenities provided to the customers since our analysis has led us to believe that the amenities provided to the customer play a significant role in the customer's feedback.**
- **They could implement some offers where in the customers are lured to use them on a trial basis and later on if they like it, they can pay for it and hence promote better marketing strategies.**
- **This way the customer will be more involved and obliged to give a positive feedback and recommend the hotel.**

CODE INTEGRATED


Data_Cleaning.R


NPS Calculation, Vizualization and M


Actual NPS vs Goal NPS.R


SVM Modelling on amenities.R

REFLECTION ON THE PROJECT AND WORKING IN A TEAM

The project was a great platform for us to explore the possibilities R programming gives us at analysing huge datasets and helped us applying almost all the concepts we learned during the lectures as well the lab sessions. Working in a team helped us to complete huge, time-consuming tasks efficiently since we had divided the work between us. The one thing which we did well as a group was to sit down and discuss every time we faced a problem at any stage. We tried to understand the root-cause of the problem which actually led us to correcting our path of analysis. Each individual made useful contributions at every point during the course of the project. There were times when all of us planned to meet but someone couldn't join at the last moment due to a prior commitment. In such times, what helped was each member's dedication towards the project. Even though they had missed the team meeting, they made sure they were informed about what had been discussed and what the person needed to do for the next meet. Overall, it was a great experience working as a team on this project since it not only helped us augment our data analysing skill-set but also helped in understanding how working on a real-life project would be like. It was a great experience for us and we look forward to working together as a team in the future if given a chance.