# Data analysis on Hyatt Chain of Hotels

In this Document, we describe analysis that we have carried on the customer data provided to us on the Hyatt Chain of Hotels. We describe our idea, the questions about the data, how we answer them, our analysis and finally the implementation. We concluded by determining the factors to help the Hyatt Brand of Hotels improve their Net Promoter Score.

# Table of Contents

# Introduction

Net promoter Score

The NPS Calculation - Calculate your NPS using the answer to a key question, using a 0-10 scale: How likely is it that you would recommend [brand] to a friend or colleague? Respondents are grouped as follows:

- Promoters (score 9-10) are loyal enthusiasts who will keep buying and refer others, fueling growth.
- Passives (score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.
- Detractors (score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.

# Business Questions

We tried to answer following business questions on the analysis of the dataset.

1) Which attributes influence NPS type the most?
2) Which lead to the biggest positive/negative impact on NPS?
3) Which country to focus on that will affect the NPS Score?
4) Which state should the client consider on improving the NPS score?
5) What Purpose of Visit for the customer has a great impact on the NPS Score?

# Data cleanse/munge/preparation

The dataset was divided into 13 csv files, separated by the month the data was collected. In total, the information was well over 10GB in size, and much of it was filled with empty fields or unfinished responses. Each spreadsheet contained 240 variables, many of which were placeholders for non-existent questions or improperly named. The first step we had to take to analyze this dataset was to transform it into a workable format to better understand and analyze what it contained.

To efficiently operate with the dataset, we selected a three-month quarter of the year to act as our sample (October, November, December), creating a 2GB collection of entries that we could

use with our computing resources quickly and effectively. While this limits our findings to a time of year, it allows us to have consistent findings not influencing each attribute by seasonal changes, as well as being a sample we could modify easily without having to condense all 13 CSV files. This selection contained over 1,300,000 individual entries, which was a large enough size to truly predictively analyze.

Finalizing the columns- Our focus of the project was on leisure, so we further focused on leisure and amenities.

We brainstormed and finalized the following column

| Likelihood_Recommend_H | Likelihood to recommend metric; value on a 1 to 10 scale |
|---|---|
| **Overall_Sat_H** | Overall satisfaction metric; value on a 1 to 10 scale |
| **Guest_Room_H** | Guest room satisfaction metric; value on a 1 to 10 scale |
| **Tranquility_H** | Tranquility metric; value on a 1 to 10 scale |
| **Condition_Hotel_H** | Condition of hotel metric; value on a 1 to 10 scale |
| **Customer_SVC_H** | Quality of customer service metric; value on a 1 to 10 scale |
| **Staff_Cared_H** | Staff cared metric; value on a 1 to 10 scale |
| **Internet_Sat_H** | Internet satisfaction metric; value on a 1 to 10 scale |
| **Check_In_H** | Quality of the check in process metric; value on a 1 to 10 scale |
| **F&B_Overall_Experience_H** | Overal F&B experience metric; value on a 1 to 10 scale |
| **All Suites_PL** | Flag indicating if the hotel is all suites |
| **Bell Staff_PL** | Flag indicating if the hotel has bell staff |
| **Boutique_PL** | Flag indicating if the hotel is boutique |
| **Business Center_PL** | Flag indicating if the hotel has a business center |
| **Casino_PL** | Flag indicating if the hotel has a casino |
| **Conference_PL** | Flag indicating if the hotel has a conference center nearby |
| **Convention_PL** | Flag indicating if the hotel has convention space |
| **Dry-Cleaning_PL** | Flag indicating if the hotel has dry-cleaning |
| **Elevators_PL** | Flag indicating if the hotel has elevators |
| **Fitness Center_PL** | Flag indicating if the hotel has a fitness center |
| **Fitness Trainer_PL** | Flag indicating if the hotel has fitness trainers |
| **Golf_PL** | Flag indicating if the hotel is near a golf space |
| **Indoor Corridors_PL** | Flag indicating if the hotel has indoor corridors |
| **Laundry_PL** | Flag indicating if the hotel has laundry space |

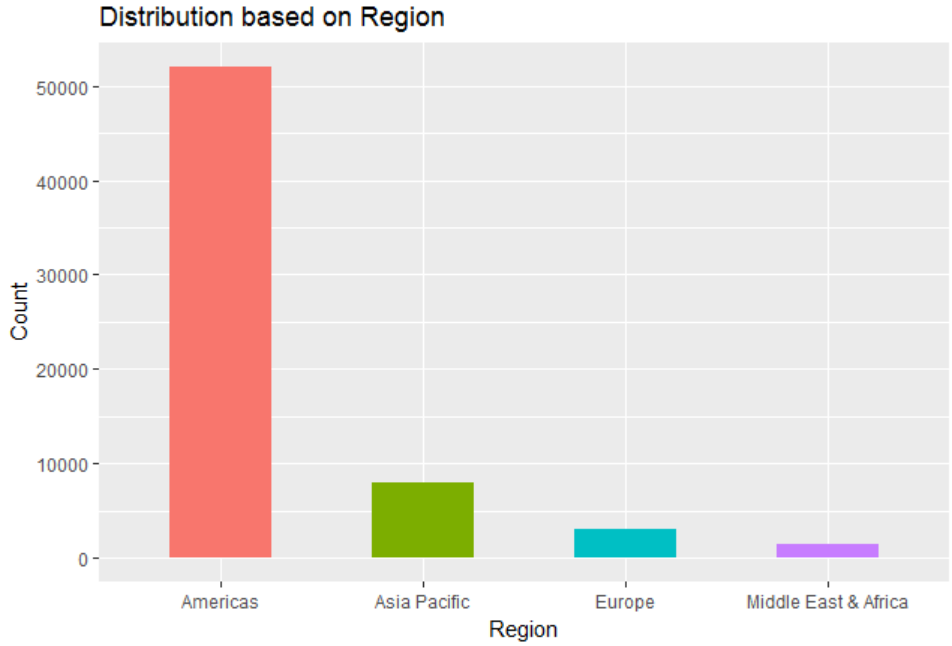
| | |
|---|---|
| **Limo Service_PL** | Flag indicating if the hotel has limo service |
| **Mini-Bar_PL** | Flag indicating if the hotel has mini-bar |
| **Pool-Indoor_PL** | Flag indicating if the hotel has an indoor poo |
| **Pool-Outdoor_PL** | Flag indicating if the hotel has an outdoor pool |
| **Regency Grand Club_PL** | Flag indicating if the hotel has a regency grand club |
| **Resort_PL** | Flag indicating if the hotel is a resort |
| **Restaurant_PL** | Flag indicating if the hotel has onsite restaurants |
| **Self-Parking_PL** | Flag indicating if the hotel has self-parking |
| **Shuttle Service_PL** | Flag indicating if the hotel has shuttle service |
| **Ski_PL** | Flag indicating if the hotel is near ski space |
| **Spa_PL** | Flag indicating if the hotel has a spa |
| **NPS_Type** | Indicates if the guest's HySat responses mark them as a promoter, a passive, or a detractor |
| **Gender_H** | Guest's gender |
| **State_PL** | State in which the hotel is located |
| **US Region_PL** | US region in which the hotel is located |
| **Postal Code_PL** | Zip code in which the hotel is located |
| **Country_PL** | Country in which the hotel is located |

As large set of data had '**NA'** in NPS type column. As NPS type is the most important column in this analysis. NA entries in this column would not be useful and hence we decided to delete NA records for NPS_type column.

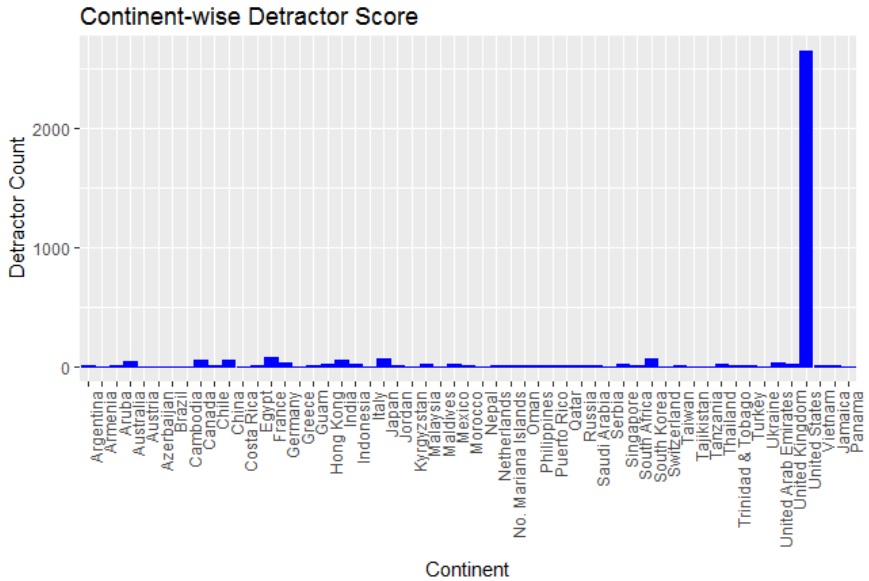
## Use of Descriptive statistics

After cleaning the data and shortening the data, we further wanted to identify what part should we focus on. So, we did following analysis and concluded to further focus on the data restricted to USA

1) Total number of Hotels according to countries. We can clearly see that tha America has the maximum number of hotels

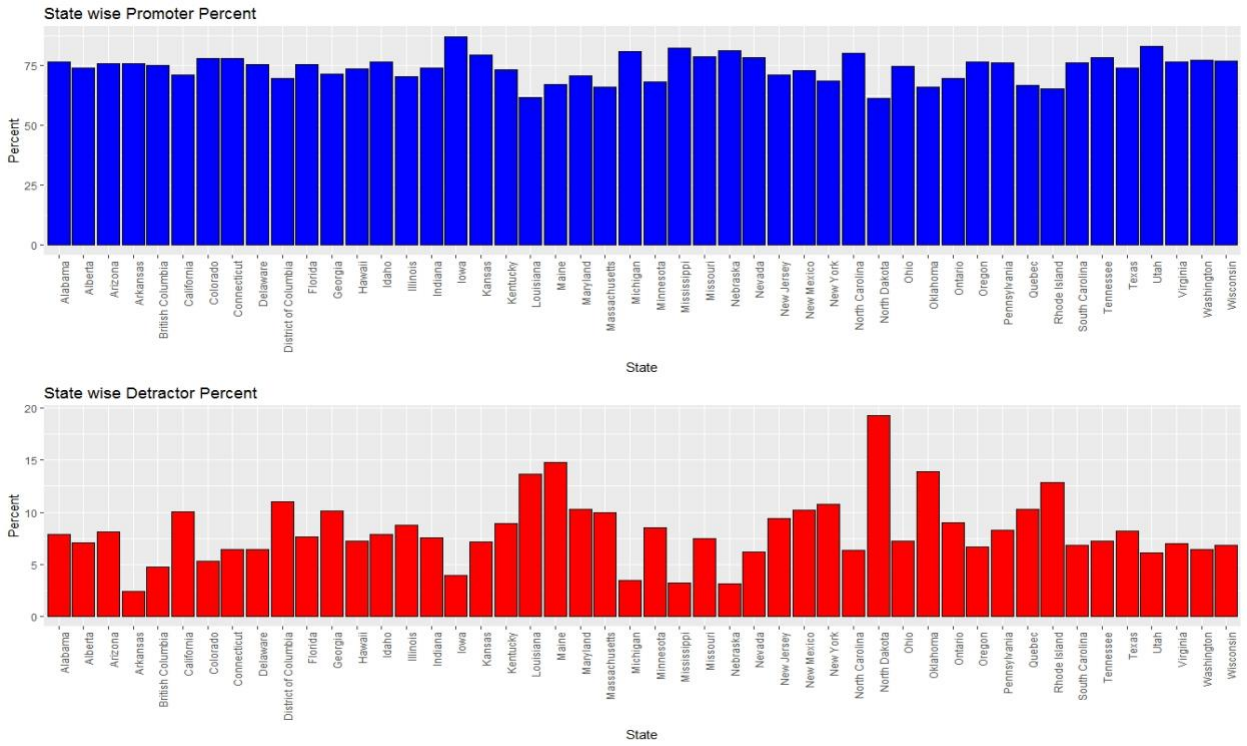



2) Finding number of detractors country wise, yet again we conclude that USA has maximum number of detractors as seen from the below graph.

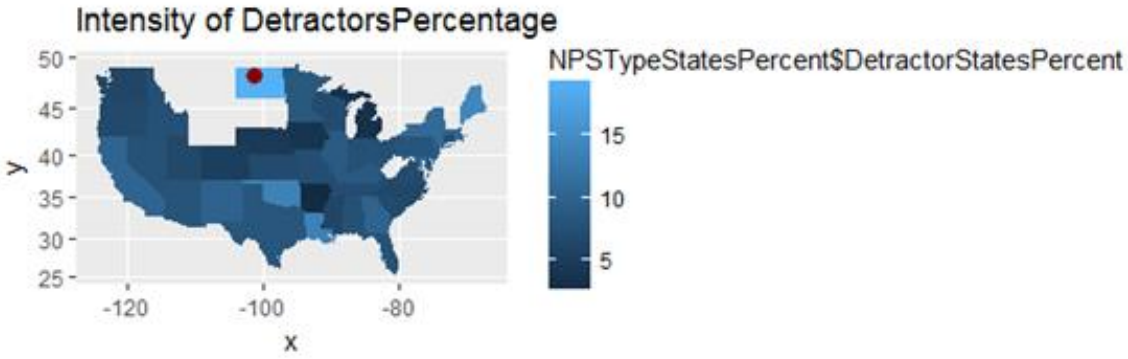


After finalizing the country, we further analyzed the percentage of the detractors in each state of the USA. We discussed and found that state with maximum number of detractors wont yield us appropriate findings so we decided to focus on percentage detractors and not maximum detractors.

3) Following plot shows the percentage detractors for each state and we can conclude that North Dakota has highest percentage detractors.




4) Further plotting the percentage of detractors on US map, we can see that North Dakota, which is marked with a red dot, has highest detractor percentage.




Hence, we will further do our analysis on North Dakota Data.

# Use of modeling techniques

We used following methods to establish relationship between the variables

1) Linear Modelling
2) Random Forest
3) Data Association rules using Apriori
4) Support Vector Machine

The whole purpose of using these modelling techniques was to draw conclusion on the important factors affecting the NPS

# 1.Linear Modelling

Linear models describe a continuous response variable as a function of one or more predictor variables. It helps to understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.

The model describes the relationship between a dependent variable yy (also called the response) as a function of one or more independent variables XiXi (called the predictors). The general equation for a linear model is:

$y=\beta 0+\sum \beta iXi+\epsilon iy=\beta 0+\sum \beta iXi+\epsilon i$

where ββ represents linear parameter estimates to be computed and ϵϵ represents the error terms.

Linear modelling technique has proven useful to identify important factors influencing the likelihood to recommend.

We targeted on performing Linear Modelling on the ratings columns.

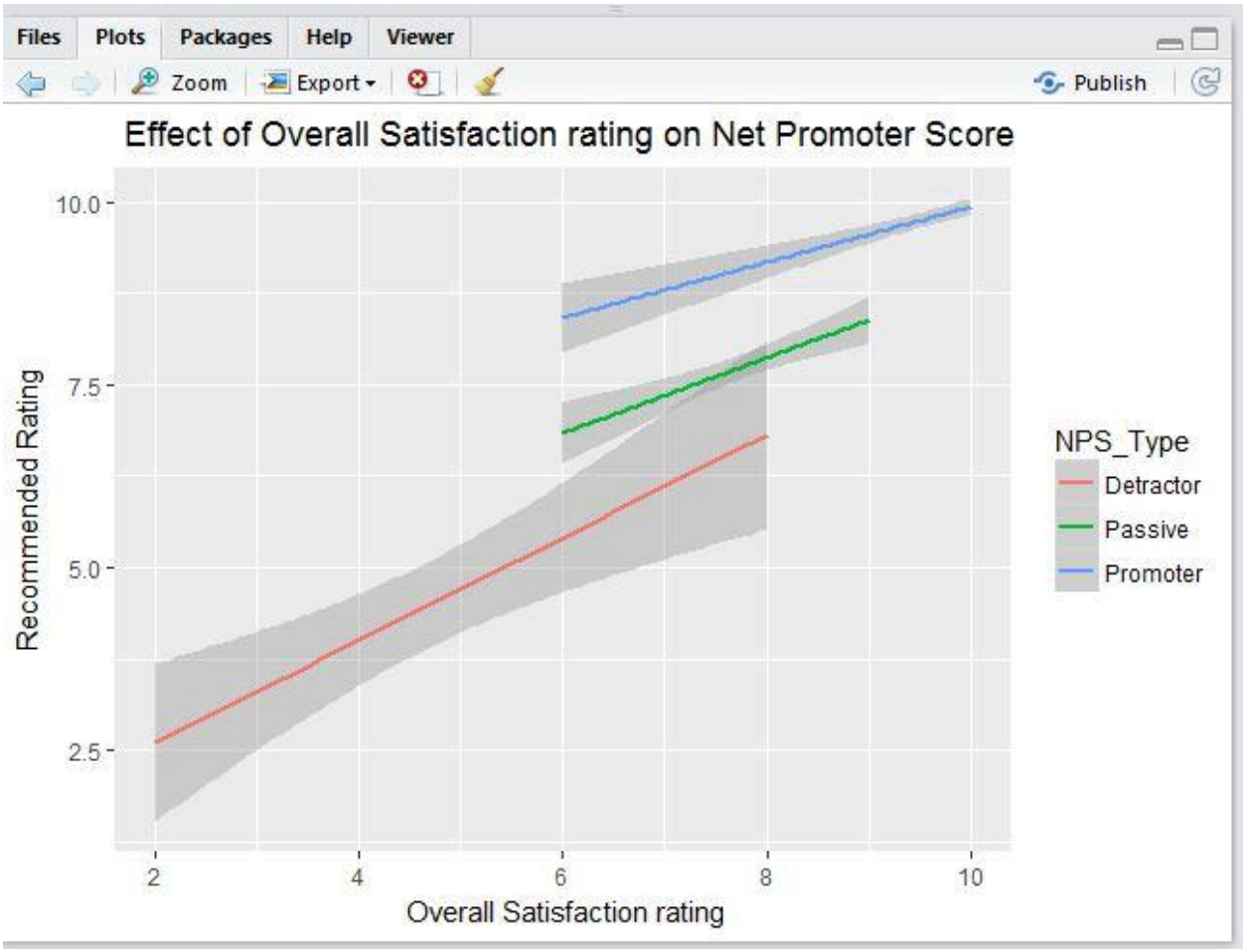
We calculated Independent variable which is "likelihood to recommend" on all of the ratings column and further calculated the Adjusted R-Squared value for each dependent variables.

Adjusted R-Squared value after performing linear modeling for the following factors are listed-

| **User Ratings** | **R-Squared value** |
|---|---|
| 1. Overall satisfaction metric | 0.9115 |
| 2. Guest room satisfaction metric | 0.4294 |
| 3. Tranquility metric | 0.4046 |
| 4. Condition of hotel metric | 0.4737 |
| 5. Quality of customer service metric | 0.4723 |
| 6. Staff cared metric | 0.3274 |
| 7. Internet satisfaction metric | 0.2171 |
| 8. Quality of the check in process metric | 0.08948 |
| 9. Overall F&B experience metric | 0.1386 |

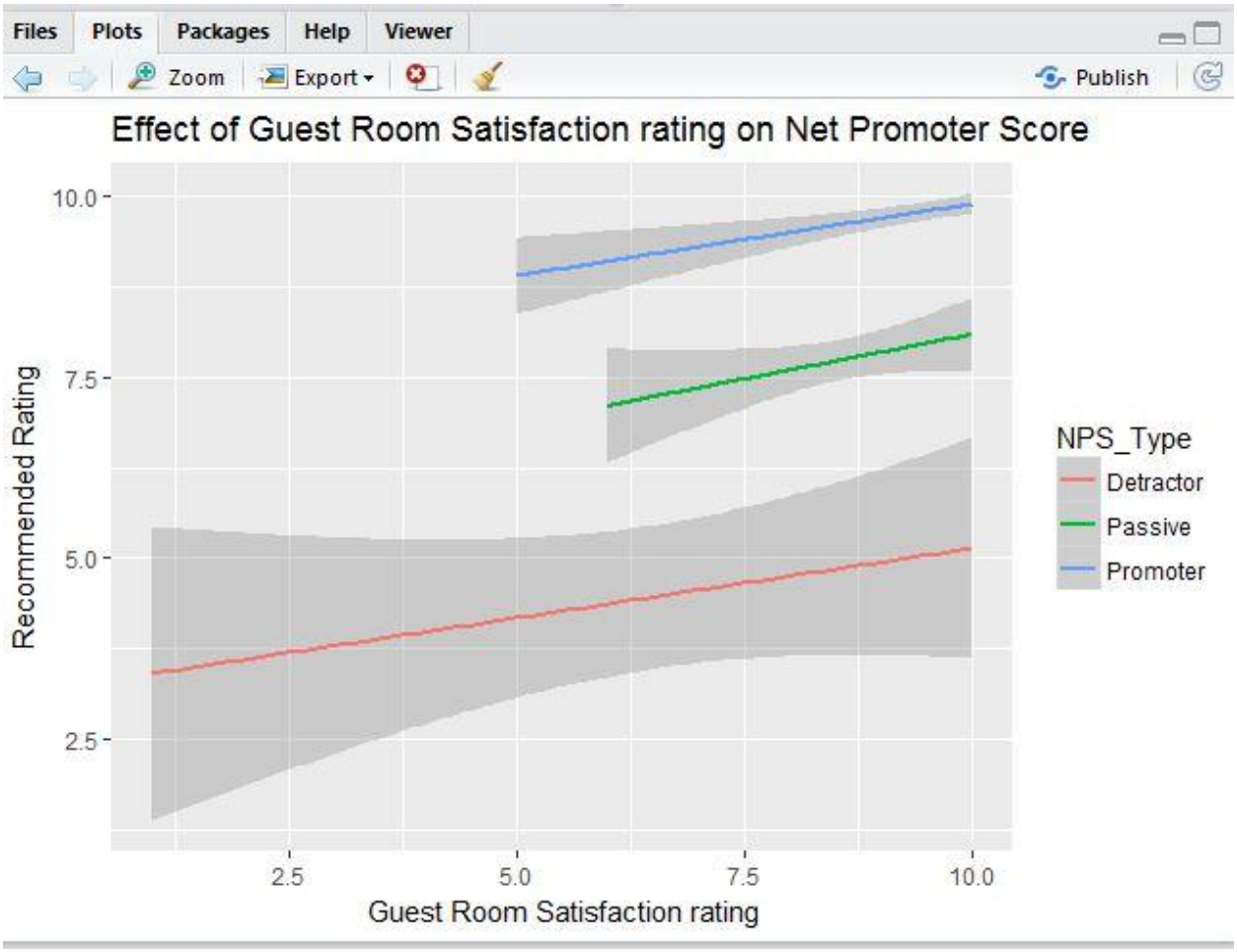
**#1 Overall_Sat_H**

lmoverallsat <- lm(formula = Likelihood_Recommend_H ~ Overall_Sat_H, data = USNorthDakotadata)
summary(lmoverallsat)
#0.9115
ggplot(USNorthDakotadata, aes(x=Overall_Sat_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Overall Satisfaction rating") + ggtitle(" Effect of Overall Satisfaction rating on Net Promoter Score")



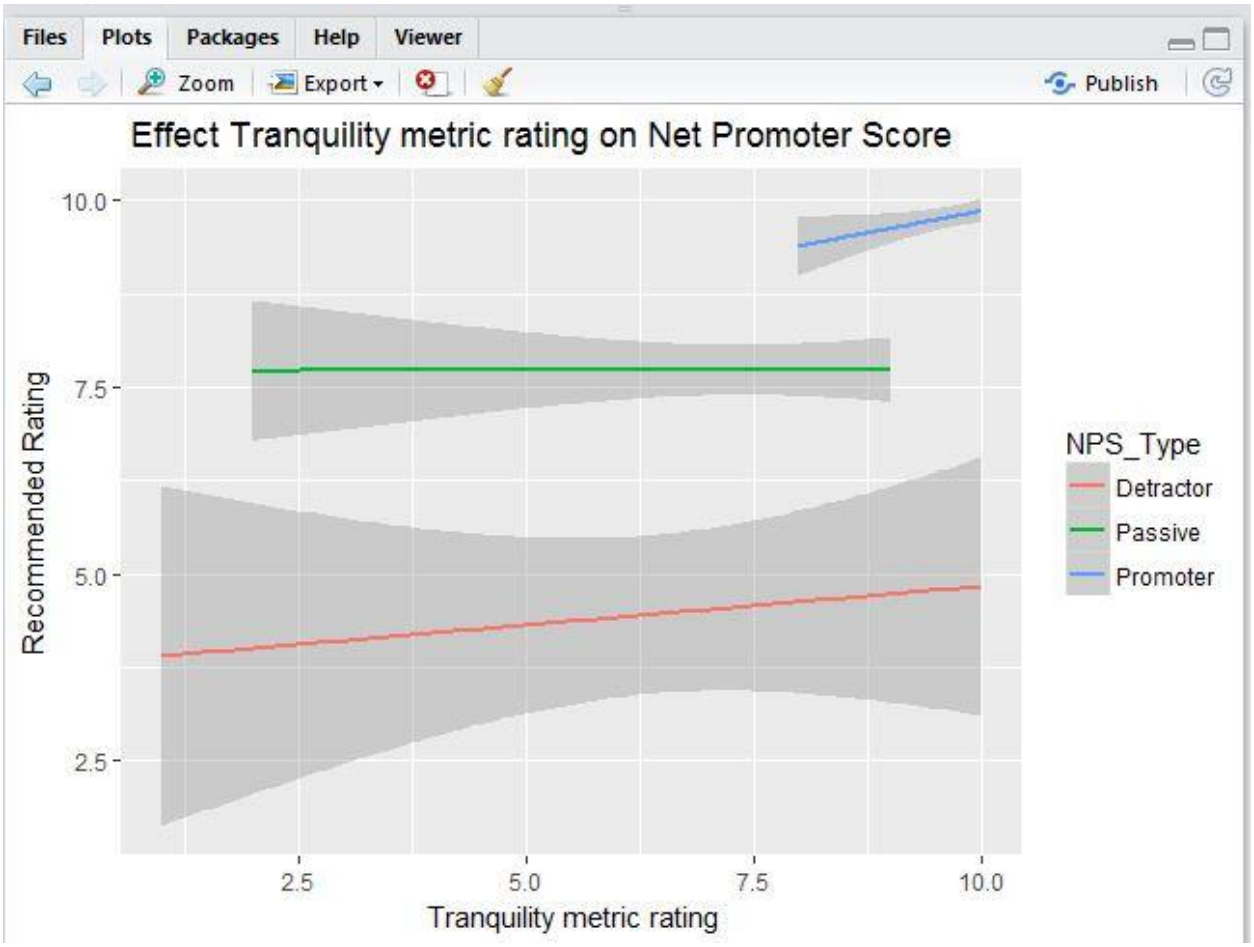

**#2 Guest_Room_H**
lmguestroom1 <- lm(formula = Likelihood_Recommend_H ~ Guest_Room_H, data = USNorthDakotadata)
summary(lmguestroom1)
#0.4294

ggplot(USNorthDakotadata, aes(x=Guest_Room_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Guest Room Satisfaction rating") + ggtitle("Effect of Guest Room Satisfaction rating on Net Promoter Score")




**#3 Tranquility_H**

lmTranquility1 <- lm(formula = Likelihood_Recommend_H ~ Tranquility_H, data = USNorthDakotadata)
summary(lmTranquility1)
# 0.4046
ggplot(USNorthDakotadata, aes(x=Tranquility_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Tranquility metric rating") + ggtitle(" Effect Tranquility metric rating on Net Promoter Score")



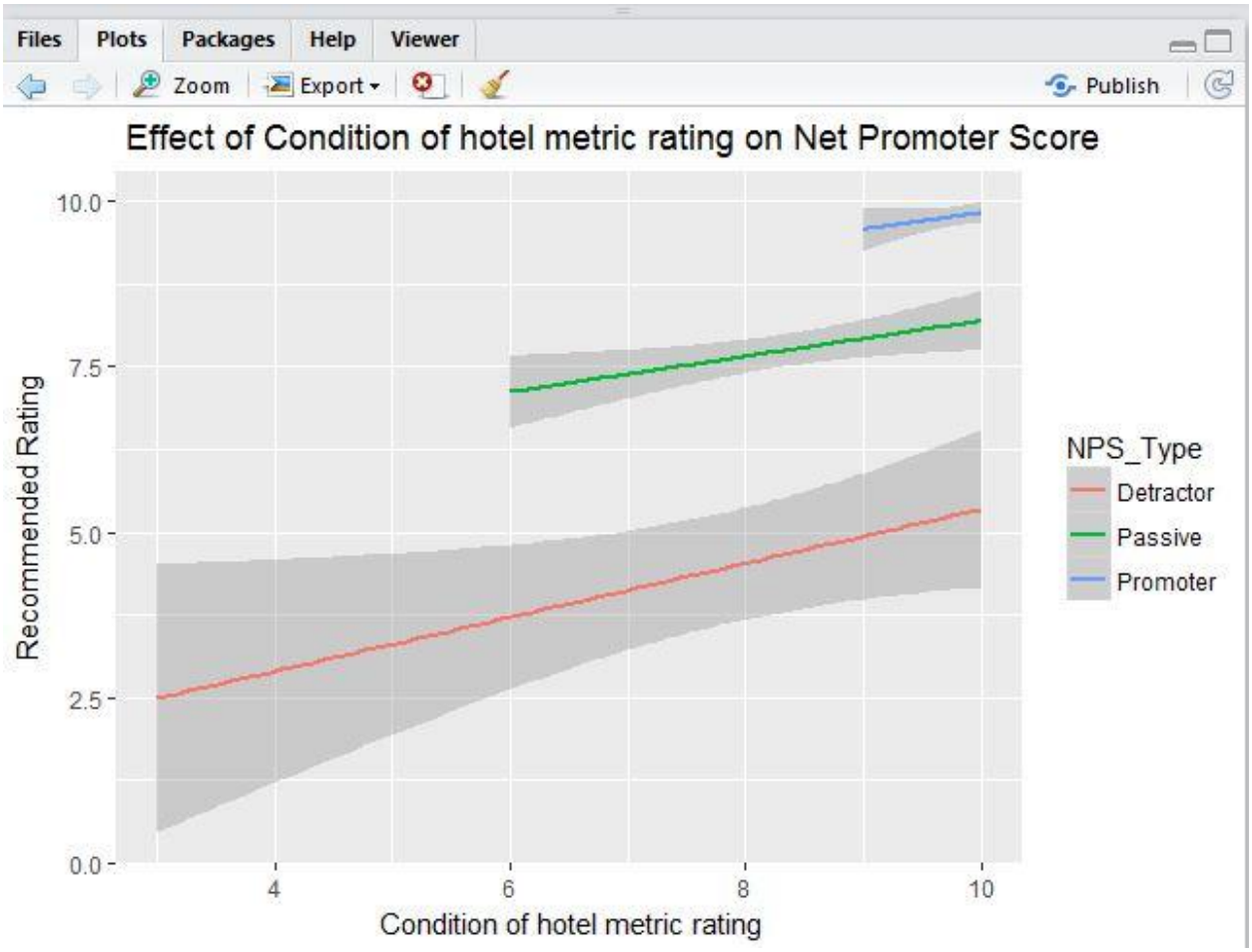

**#4 Condition_Hotel_H**

lmCondition_Hotel1 <- lm(formula = Likelihood_Recommend_H ~ Condition_Hotel_H, data = USNorthDakotadata)

summary(lmCondition_Hotel1)

# 0.4737

ggplot(USNorthDakotadata, aes(x=Condition_Hotel_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Condition of hotel metric rating") + ggtitle(" Effect of Condition of hotel metric rating on Net Promoter Score")



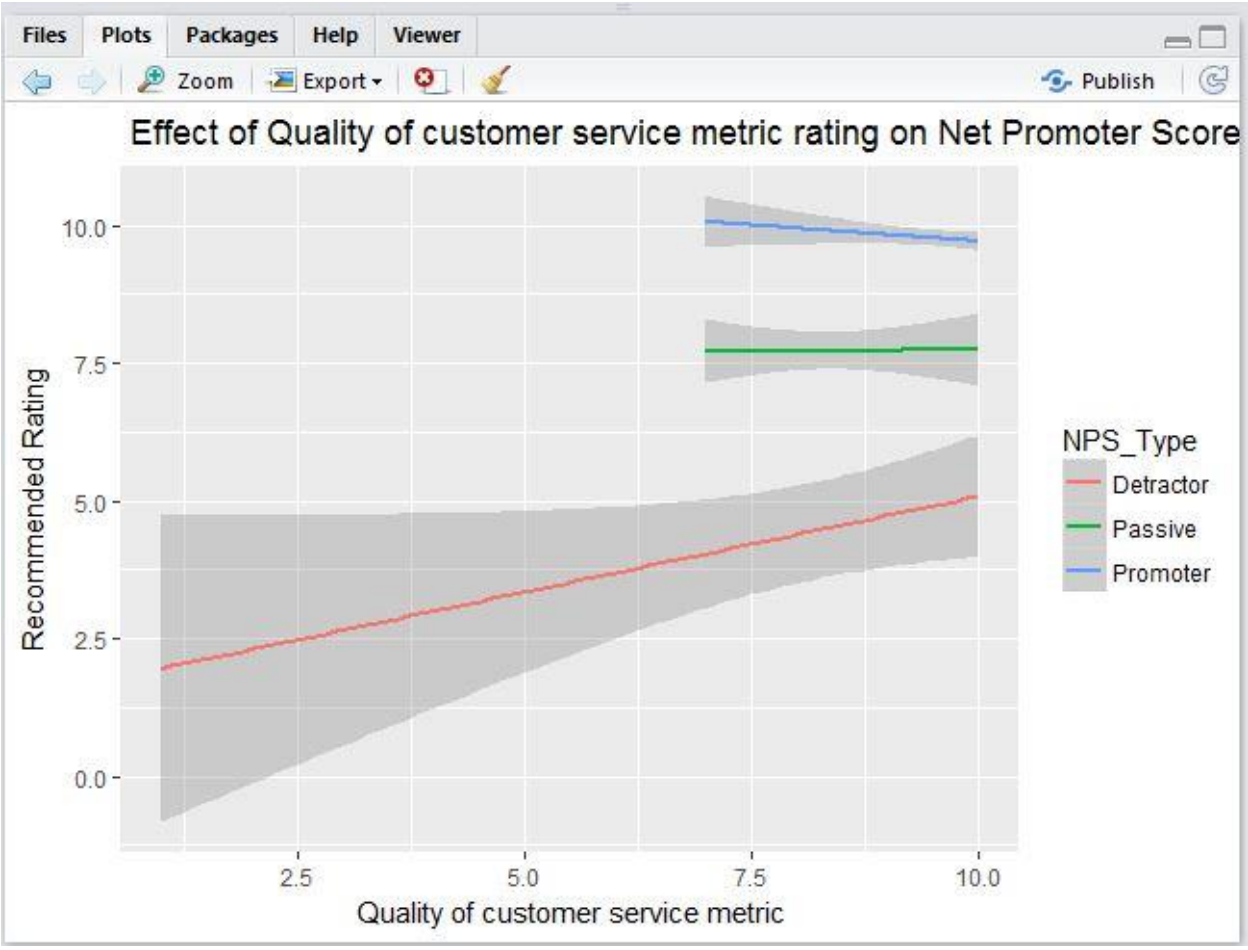

**#5 Internet_Sat_H**

lmInternet_Sat_1 <- lm(formula = Likelihood_Recommend_H ~ Internet_Sat_H, data = USNorthDakotadata)

summary(lmInternet_Sat_1)

#0.2171

ggplot(USNorthDakotadata, aes(x=Internet_Sat_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Quality of customer service metric") + ggtitle(" Effect of Quality of customer service metric rating on Net Promoter Score")



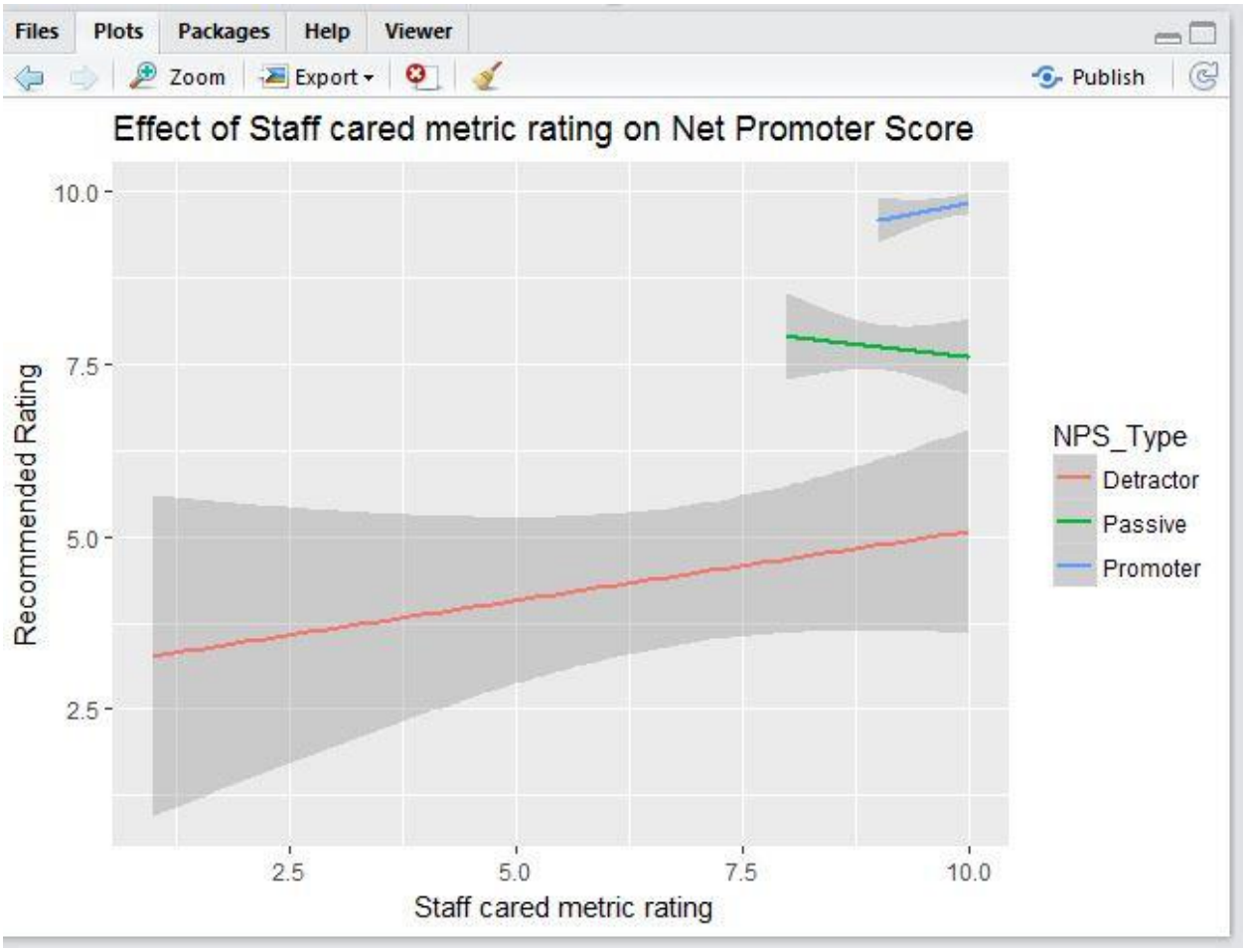

**#6 Customer_SVC_H**

lmCustomer_SVC1 <- lm(formula = Likelihood_Recommend_H ~ Customer_SVC_H, data = USNorthDakotadata)

summary(lmCustomer_SVC1)

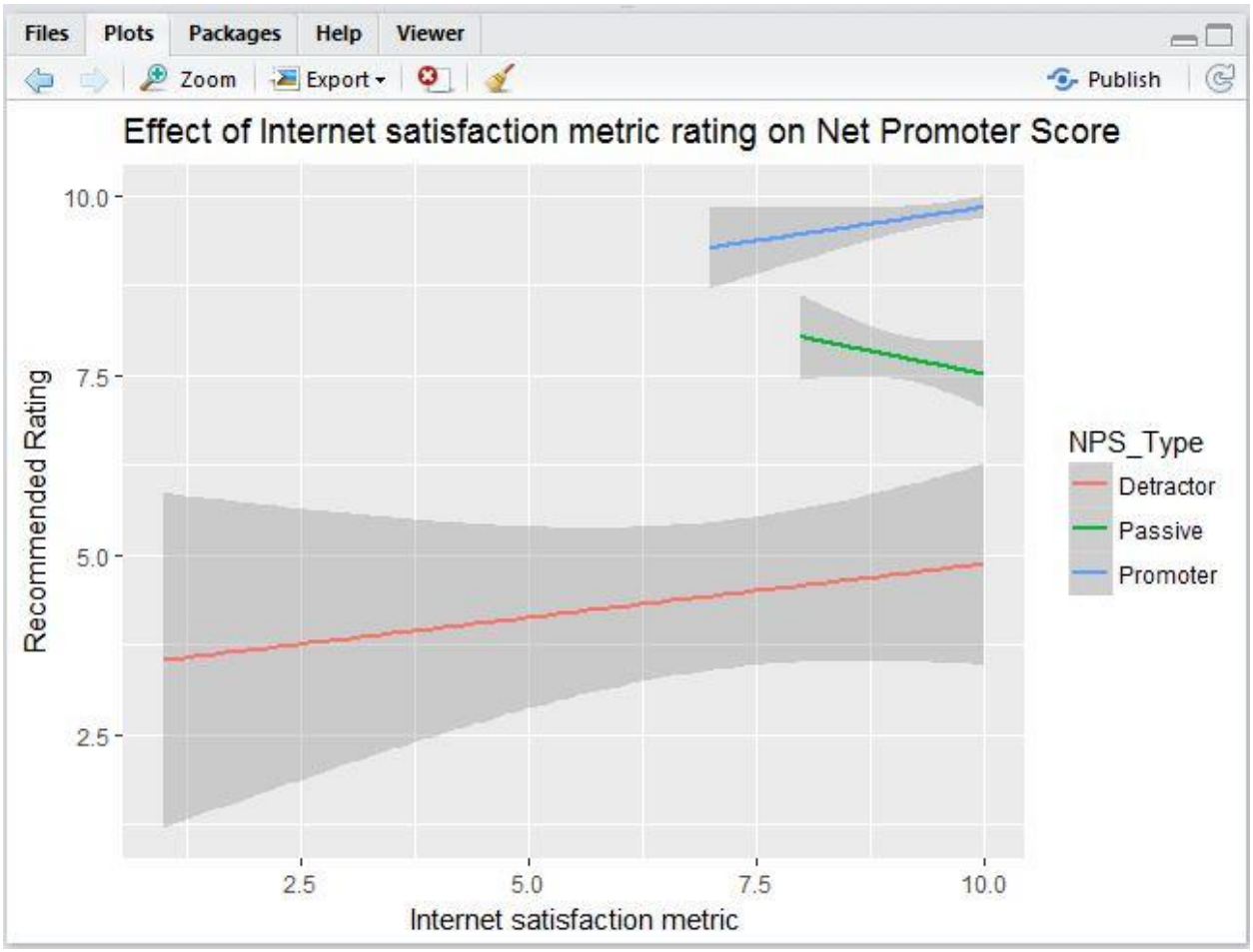
#0.4723

ggplot(USNorthDakotadata, aes(x=Customer_SVC_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Staff cared metric rating") + ggtitle("Effect of Staff cared metric rating on Net Promoter Score")




**#7 Staff_Cared_H**

lmStaff_Cared1 <- lm(formula = Likelihood_Recommend_H ~ Staff_Cared_H, data = USNorthDakotadata)
summary(lmStaff_Cared1)
#0.3274
ggplot(USNorthDakotadata, aes(x=Staff_Cared_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Internet satisfaction metric") + ggtitle("Effect of Internet satisfaction metric rating on Net Promoter Score")




**#8 Check_In_H**

lmCheck_In1 <- lm(formula = Likelihood_Recommend_H ~ Check_In_H, data = USNorthDakotadata)
summary(lmCheck_In1)
#0.08948
ggplot(USNorthDakotadata, aes(x=Check_In_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Quality of the check in process metric rating") + ggtitle(" Effect of Quality of the check in process metric rating on Net Promoter Score")

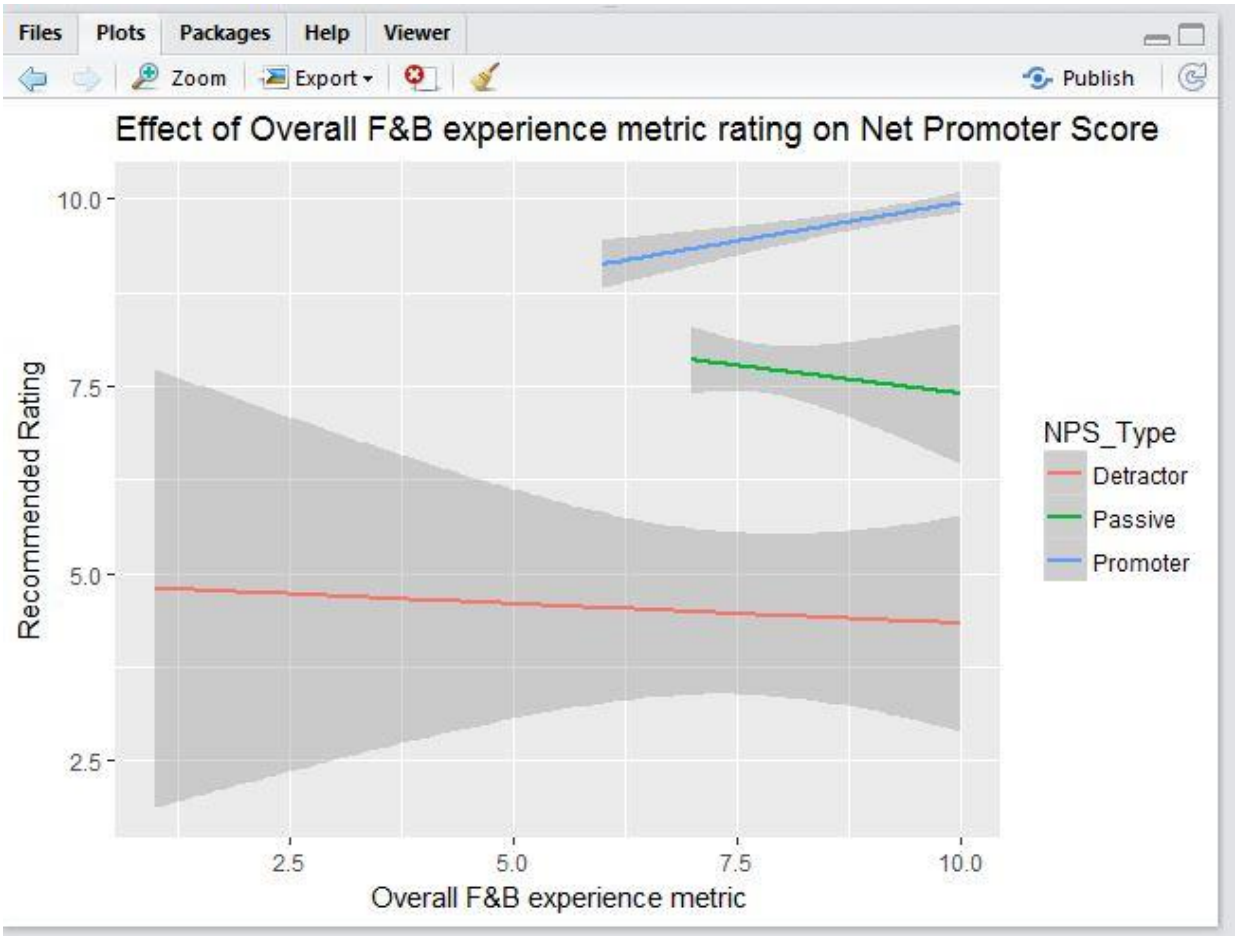

**#9 F.B_Overall_Experience_H**

lmF.B_Overall1 <- lm(formula = Likelihood_Recommend_H ~ F.B_Overall_Experience_H, data = USNorthDakotadata)

summary(lmF.B_Overall1)

#0.1386

ggplot(USNorthDakotadata, aes(x=F.B_Overall_Experience_H, y=Likelihood_Recommend_H, color=NPS_Type)) + geom_smooth(method = "lm") + ylab("Recommended Rating") + xlab(" Overall F&B experience metric") + ggtitle("Effect of Overall F&B experience metric rating on Net Promoter Score")




**# Overall**

lmOverall <- lm(formula = Likelihood_Recommend_H ~ Overall_Sat_H + Guest_Room_H + Tranquility_H + Condition_Hotel_H + Customer_SVC_H, data = USNorthDakotadata)

summary(lmOverall)

#0.9152

We are getting an Adjusted R-Squared value of '0.9152'. It depicts that the independent variable has a high influence on the dependent variable.

On the basis of the model, we found out that Customer_SVC_H, Guest_Room_H, Condition_Hotel_H and Tranquility_H are important ones.

# 2) Random Forest

**Concept**

Random forests is a learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. We have used Random forest algorithm as a backing up algorithm to linear model. The variables that we have identified in Linear model are used in Random Forest.

**Code-**

#Random Forest

install.packages("randomForest")

library("randomForest")

```
USNorthDakotadata_RF <- USNorthDakotadata
USNorthDakotadata_RF$Likelihood_Recommend_H <- ifelse(USNorthDakotadata_RF$Likelihood_Recommend_H >= "6",1,0)
USNorthDakotadata_RF$Likelihood_Recommend_H <- as.character(USNorthDakotadata_RF$Likelihood_Recommend_H)
USNorthDakotadata_RF$Likelihood_Recommend_H <- as.factor(USNorthDakotadata_RF$Likelihood_Recommend_H)
View(USNorthDakotadata_RF)
output.forest1 <- randomForest(Likelihood_Recommend_H ~ Overall_Sat_H + Customer_SVC_H + Condition_Hotel_H + Tranquility_H + Guest_Room_H, data = USNorthDakotadata_RF)
print(output.forest1)
#OOB estimate of  error rate: 14.04%
#      0     1     class.error
# 0   30   5      0.1428571
# 1   3     19   0.1363636
```

As Random Forest is a classification algorithm, we have made Likelihood_Recommend_H a finite variable. We have used Likelihood. Recommend_H as a dependent variable and the variables that we have found out by linear model as independent variables.
We have got an accuracy of 86.96% which is very good. Also, we can interpret that 30 + 19 (49) values are correctly predicted while just 5 + 3 (8) are incorrectly predicted.

# 3. Association Mining Rules

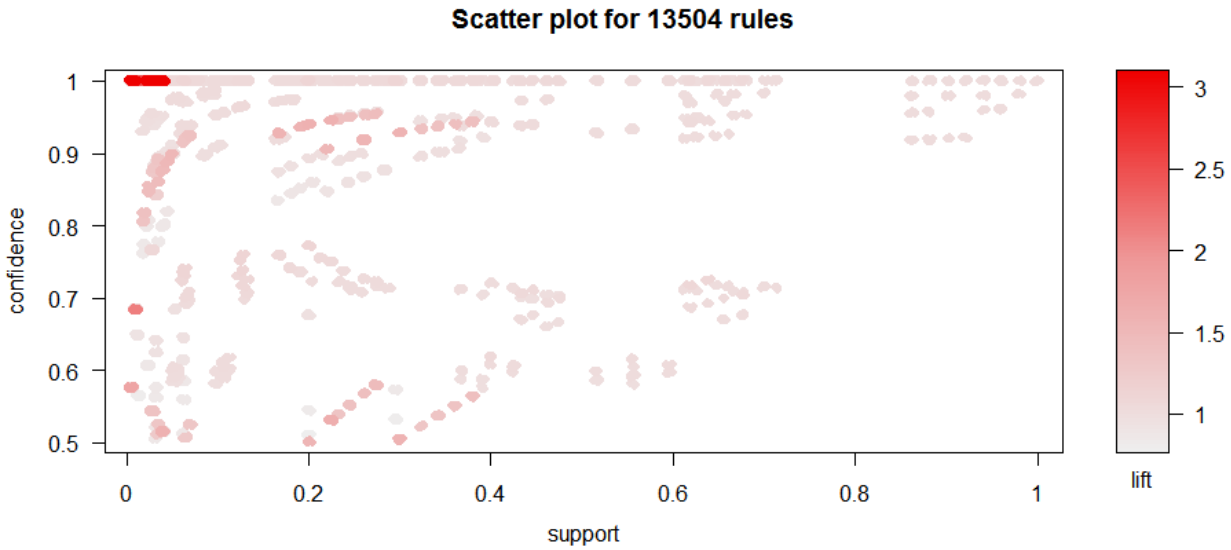
**Concept**

When we were analyzing the reasons for North Dakota having the most number of detractors, we realized that North Dakota doesn't have the amenities needed to increase the promoters.
We thought of analyzing the state that had highest number of promoters, so that North Dakota can learn from their data and try to include those amenities in their hotels. This way the detractors can be converted to promoters.
As we had seen earlier, California was the state with highest number of promoters. It would be very impractical and incorrect if we expect North Dakota to include all the amenities at once to improve the NPS Value. Hence, we analyzed and recommend including only those amenities which affect the promoters a lot. This association of amenities can be justified using the association rules.
We took only the columns of amenities of California.

**Code-**

```
columns1 <- c("Likelihood_Recommend_H","Boutique_PL", "Casino_PL" ,"Conference_PL", "Convention_PL", "Golf_PL", "Indoor.Corridors_PL", "Ski_PL", "Spa_PL", "NPS_Type")
```

```
c<-Calidata[ ,columns1]
c[c==""] <- NA
c <- na.omit(c)
str(c)
#Convert the data type into factors
c$Convention_PL <- as.factor(c$Convention_PL)
c$Conference_PL <- as.factor(c$Conference_PL)
c$Boutique_PL <- as.factor(c$Boutique_PL)
c$Casino_PL <- as.factor(c$Casino_PL)
c$Golf_PL <- as.factor(c$Golf_PL)
c$Indoor.Corridors_PL <- as.factor(c$Indoor.Corridors_PL)
c$Ski_PL <- as.factor(c$Ski_PL)
c$Spa_PL <- as.factor(c$Spa_PL)
c$NPS_Type <- as.factor(c$NPS_Type)

#Importing the packages needed for association mining
import.package("arules")
imort.package("arulesViz")
library(arules)
library(arulesViz)
#Applying the rule
rules_c<-apriori(c,parameter=list(support=0.005,confidence= 0.5))
#Inspect the rules obtained
inspect(rules_c)
#plot the rules
```




```
#Make a dataframe which contains the LHS, RHS, Support, #Confidence and Lift of all the rules
rules_info <-data.frame(LHS = labels(lhs(rules_c)), RHS = labels(rhs(rules_c)),quality(rules_c))
#Narrowed our dataframe such that our RHS contains only #NPS_Type as Promoters, so that we get only
the rules which affect #the NPS_Type=Promoters
```

x <- rules_info[rules_info$RHS=='{NPS_Type=Promoter}',]

As we learnt in the class and in our homework, the larger the value of lift, the more interesting the rule may be. Hence, we sorted our data in the dataframe according to the decreasing value of the lift. We decided to recommend only the top 5 amenities, according to the lift values.

| LHS | RHS | support | confidence | lift |
|---|---|---|---|---|
| {Golf_PL=Y} | {NPS_Type=Promoter} | 0.03347846 | 0.8415094 | 1.179814 |
| {Golf_PL=Y,Spa_PL=Y} | {NPS_Type=Promoter} | 0.03347846 | 0.8415094 | 1.179814 |
| {Convention_PL=Y,Golf_PL=Y} | {NPS_Type=Promoter} | 0.03347846 | 0.8415094 | 1.179814 |
| {Golf_PL=Y,Indoor.Corridors_PL=Y} | {NPS_Type=Promoter} | 0.03347846 | 0.8415094 | 1.179814 |
| {Boutique_PL=N,Golf_PL=Y} | {NPS_Type=Promoter} | 0.03347846 | 0.8415094 | 1.179814 |

## 4.SVM:

We performed SVM to ensure that the list of amenities that we obtained from the Association Rules was accurate or not.

```
# Create the Support Vector Machine (SVM) model
library(kernlab)

# Convert the column of Likelihood to Recommend to factors
c$Likelihood_Recommend_H <-ifelse(c$Likelihood_Recommend_H>=6,1,0)
c$Likelihood_Recommend_H <- as.factor(c$Likelihood_Recommend_H)

# Randomly sample 2/3 data as a training dataset and the rest data # as a test dataset
set.seed(10)
randIndex <- sample(1:dim(c)[1])
cutPoint2_3 <- floor(2*dim(c)[1]/3)
trainData <- c[randIndex[1:cutPoint2_3],]
testData <- c[randIndex[(cutPoint2_3+1):dim(c)[1]],]

# Apply the function of SVM
```

svmModel1 <- ksvm(Likelihood_Recommend_H ~ Spa_PL + Indoor.Corridors_PL + Golf_PL + Convention_PL, data = trainData, kernel="rbfdot", kpar="automatic", C=20, cross=3)

svmModel1

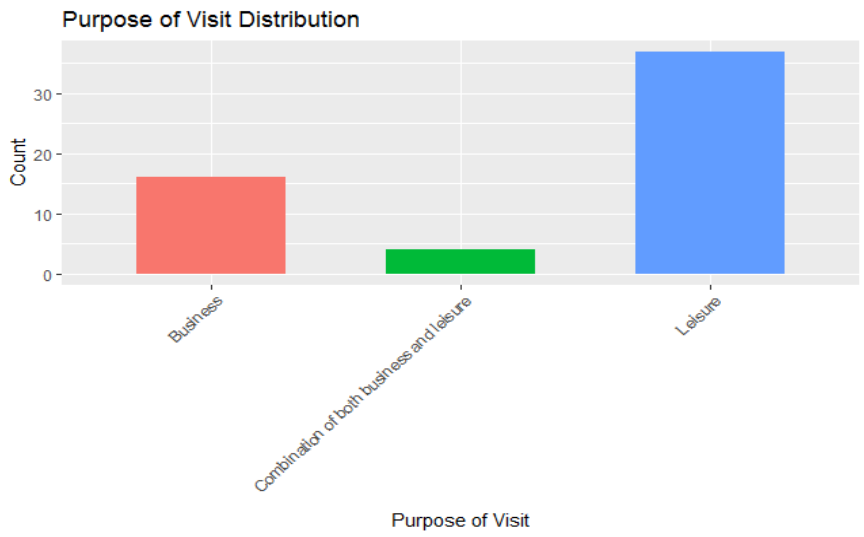
```
> svmModel1
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 20

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.625

Number of Support Vectors : 595

Objective Function Value : -11880
Training error : 0.066892
Cross validation error : 0.066892
```

From the output, we infer that the training error was around 6%, which tells that the amenities from the association rules rightly affect the promoters.

From the data of North Dakota, we see that it has the indoor.corridors facility already. We recommend that along with this amenity, if the amenities of Spa, Golf and convention space are incorporated, the number of promoters could be increased.

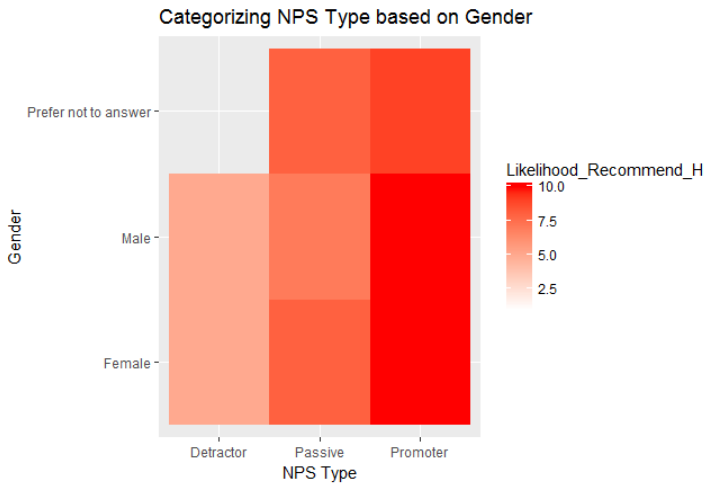
# Visualization

1) Our first major step was to divide the data based on purpose of visit.



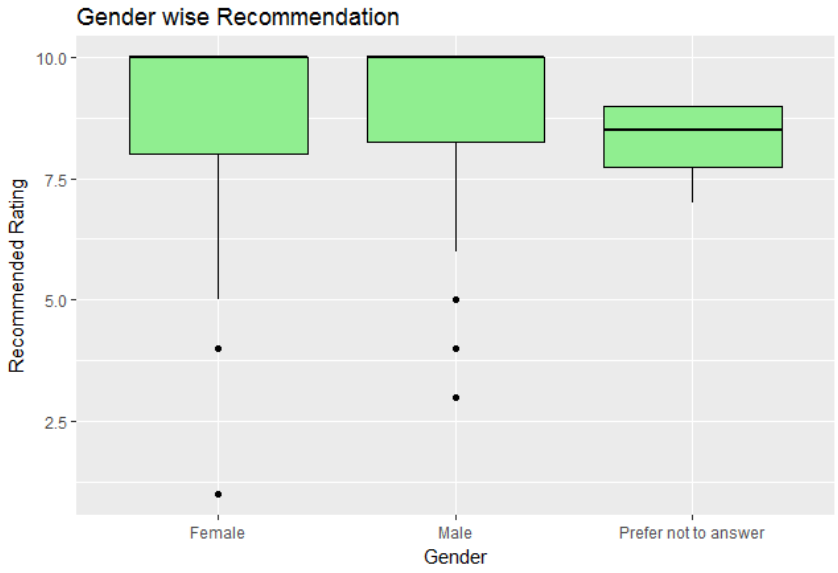

Upon plotting the bar graph, we can see that the count of people and their reason as to why they visit North Dakota's hotel and we see that the highest count is for leisure.
In addition, from the above Associative Rules we can see that Spa and Golf were the recommendations that we received, which fits in well with the guest who comes for leisure

2) We plotted a heat map to understand number of detractors, promoters and passive based on gender.



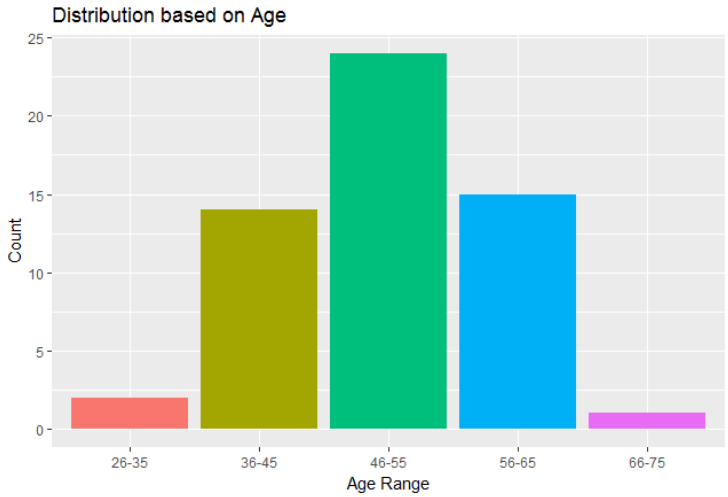
From this Visualization, we can easily decipher that the majority of the male customers in the Passive or Detractor have a less likelihood to recommend the hotel as compared to the females. This is one of the main reasons we are planning to focus on males and for that, we have done further analysis.

3) To further support our aurgument of focussing on males we plotted box plots as seen below.

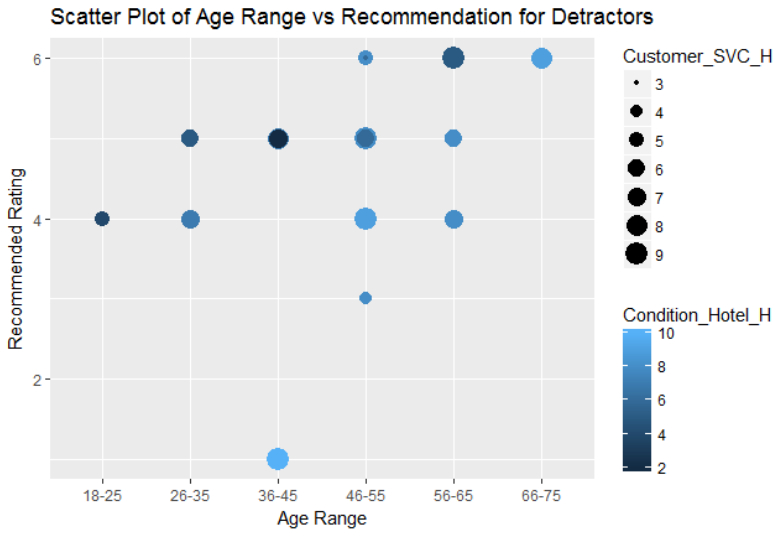



In addition, we can see that the Men are less likely to recommend as compared to the males and our main goal is to reduce the percentage of detractors so focusing on male customers should be a priority for the Hyatt Hotel chain in North Dakota.

4) To get a better understanding of the guests in North Dakota we plotted a bar graph to visualize the count of the age demographics of the North Dakota customers, in which we can see that the highest number of customers fall in the age range of 46-55.




5) To get a better insight of the data with respect to the age range and rating of the Detractors data we plotted the scatter plot as seen below.




From our linear modelling we got that, the most important factors are Customer Service and the Condition of Hotel. From the scatter plot we can infer that the people in the age range of 46-55 have a higher chance of recommending the hotel if these factors have a high rating.

. This along with age were the only two major demographic characteristics that had predictable and noticeable impact on overall NPS scores.

This along with age were the only two major demographic characteristics that had predictable and

noticeable impact on overall NPS scores.

**5)**

## Interpretation of the results/Actionable Insights

## Validation

We have used four models – Linear Modelling, Random Forest, SVM and Association Mining. When we applied Linear Modelling on influential facilities, adjusted R2 Value comes around 92%. This is highly accurate. We also performed Random Forest with the 4 most important variables identified by Linear Modelling. Since the accuracy came out to be 86.96%, we concluded through validation that these factors should be consider at an immediate basis.

When we applied Association Mining on all amenities, we found 3 important amenities. We also performed SVM on the 3 most important variables identified by Association Mining. Since the accuracy came out to be 94% we concluded through validation that these factors need to be taken into consideration at an immediate basis.