IST 687

Data Analysis on Hyatt Hotel Dataset

Farheen Safoora Najeeb

Harsh Takrani

Radhika Bhange

Ashik Liyakathali

Introduction:

This project was undertaken for the course IST 687: Applied Data Science at the iSchool, Syracuse University. The entire dataset contains about 3 Million responses collected from the Hyatt Customer Survey ranging from Feb 2014 to January 2015. This huge quantity of data consists of 237 attributes like details about the person who responded to the survey (example: guest title, guest preferred language), few more attributes about the hotel (for example: Location, Spa, Type) and a column that indicates whether the person is a promoter, passive or detractor. Considering the fact that Customer satisfaction is the key to the long-term success of any hotel chain. If you are not asking your customers for feedback, they may simply move on to your competitors. However, when you begin to use advanced online survey software to regularly administer customer satisfaction surveys, you can collect and analyze important feedback necessary to development important strategic business decisions. Owing to this, the major aim of the project was to carry out data analysis on this huge data set to gain insights for improving the Net Promoter Score of the Hyatt Chain of Hotels.

Insights on the Data:

For the most part, selecting the required attributes involves mining the data which generates information with a high value for an organization. Working with raw data makes the process even more complicated, energy-consuming and may give out incorrect results. Hence, it is necessary for us to select appropriate attributes for demonstrating the relationship with NPS. The following steps must be taken in order to increase the efficiency of our algorithm. Here are some steps which were devised to analyze the huge dataset:

- Data cleaning
- NPS Calculation
- Region Selection for the data analysis
- Visualization of the Data based on the analysis
- Find correlation
- Use modelling techniques
- Give recommendations based on the analysis

Net Promoter Score:

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.

Customers are surveyed on one single question. They are asked to rate on an 11-point scale the likelihood of recommending the company or brand to a friend or colleague. “On a scale of 0 to 10, how likely are you to recommend this company’s product or service to a friend or a colleague?” Based on their rating, customers are then classified in 3 categories: detractors, passives and promoters.

 Promoters (9 or 10)

Promoters are loyal and enthusiastic fans. They sing the company’s praises to their friends and colleagues. They are far more likely than others to remain customers and to increase their purchases over time.

 Passives (7 or 8)

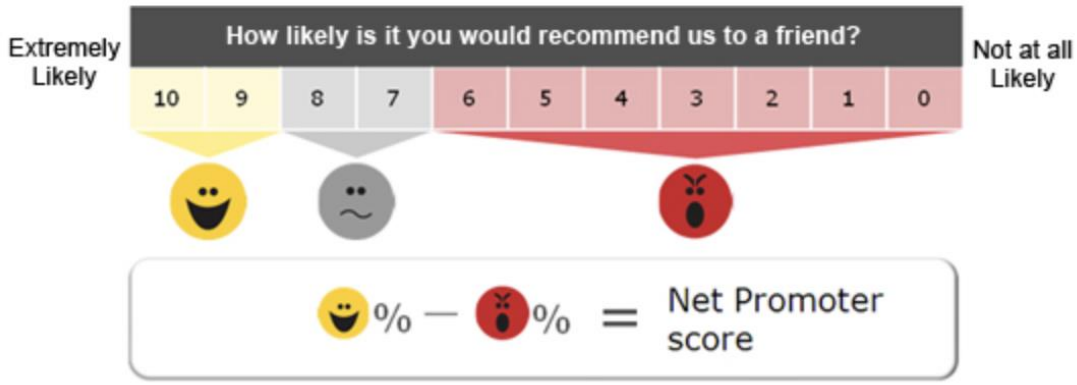
They are somewhat satisfied but could easily switch to a competitor’s offering if given the opportunity. They probably wouldn’t spread any negative word-of-mouth, but are not enthusiastic enough about your products or services to promote them.

 Detractors (0 to 6)

Detractors are unhappy customers. They, with all likelihood, won’t purchase again from the company, could potentially damage the company’s reputation through negative word of mouth.

NPS Calculation

The Net Promoter Score (NPS) is determined by subtracting the percentage of customers who are detractors from the percentage who are promoters. What is generated is a score between -100 and 100 called the Net Promoter Score. At one end of the spectrum, if when surveyed, all of the customers gave a score lower or equal to 6, this would lead to a NPS of -100. On the other end of the spectrum, if all the customers were answering the question with a 9 or 10, then the total Net Promoter Score would be 100.



Using NPS you can also track it for customer segments, geographic units or functional groups. It helps everyone focus on the twin goals of creating more promoters and fewer detractors. It is, quite simply, your customer balance sheet.

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 number 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 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.

Scope:

Which set of data did we deem the most valuable to our analysis? What is the scope of our data?

As the team started to work with the deluge of data, it soon became apparent that we would first need to focus on two important things:

Prioritize attributes:

We had 237 attributes to work with which contained varied data right from the demographic data of the Customer's, Hotel location and amenities data, Check-in information of the Customers, revenue data and so much more. We needed to prioritize the data which we considered important to our analysis and filter out the rest.

Clean the data:

The entire dataset includes the data for all the 12 months in a year with each file being around 1 GB. We have performed the analysis on the all the months in a year. The dataset is first loaded into R from CSV format file, here we have used data table library and fread function to load the data and then columns were filtered which we found relevant for our analysis.
Read.csv is slow because it reads everything in the memory as character and then converts it back into integer, numeric or character according to the data, while fread memory maps the file into memory and hence is faster. Next, we removed all the blank spaces in the Net Promotor Scale column as they were irrelevant.

Below is a list of the columns which we considered important in our data:

S No.	Column Name	Definition
1	ROOM_TYPE_CODE_C	Hyatt standard room type code of the guest's room upon checkout
2	ROOM_TYPE_DESCRIPTION_C	Room type description (specific to the property) of the guest's room upon checkout
3	POV_CODE_C	Purpose of visit
4	PMS_ROOM_REV_USD_C	Room revenue from the PMS in USD
5	PMS_TOTAL_REV_USD_C	Total from the PMS in USD
6	PMS_FOOD_BEVERAGE_REV_USD_C	F&B revenue from the PMS in USD
7	PMS_OTHER_REV_USD_C	Other revenue from the PMS in USD
8	GUEST_COUNTRY_R	The country to which the actual guest belongs to. Different from Country_Code
9	e_hy_gss_gender_I	Guest's gender
10	Likelihood_Recommend_H	Likelihood to recommend metric; value on a 1 to 10 scale
11	Overall_Sat_H	Overall satisfaction metric; value on a 1 to 10 scale
12	Guest_Room_H	Guest room satisfaction metric; value on a 1 to 10 scale
13	Tranquility_H	Tranquility metric; value on a 1 to 10 scale
14	Condition_Hotel_H	Condition of hotel metric; value on a 1 to 10 scale
15	Customer_SVC_H	Quality of customer service metric; value on a 1 to 10 scale
16	Staff_Cared_H	Staff cared metric; value on a 1 to 10 scale
17	Internet_Sat_H	Internet satisfaction metric; value on a 1 to 10 scale
18	Check_In_H	Quality of the check in process metric; value on a 1 to 10 scale
19	F&B_Overall_Experience_H	Overal F&B experience metric; value on a 1 to 10 scale
20	State_PL	State in which the hotel is located
21	US Region_PL	US region in which the hotel is located
22	Country_PL	Country in which the hotel is located
23	Guest NPS Goal_PL	Hotel's NPS goals
24	Business Center_PL	Flag indicating if the hotel has a business center
25	Casino_PL	Flag indicating if the hotel has a casino
26	Conference_PL	Flag indicating if the hotel has a conference center nearby
27	Convention_PL	Flag indicating if the hotel has convention space
28	Golf_PL	Flag indicating if the hotel is near a golf space
29	Limo Service_PL	Flag indicating if the hotel has limo service
30	Mini-Bar_PL	Flag indicating if the hotel has mini-bar
31	Pool-Indoor_PL	Flag indicating if the hotel has an indoor poo
32	Pool-Outdoor_PL	Flag indicating if the hotel has an outdoor pool
33	Resort_PL	Flag indicating if the hotel is a resort
34	Spa_PL	Flag indicating if the hotel has a spa
35	NPS_Type	Indicates if the guest's HySat responses mark them as a promoter, a passive, or a detractor

Steps taken to obtain our data:

1. Once the total records were reduced to around 35 columns we considered removing all the columns with N/As in it as that won't add any value to the model. However, while doing so we realized we were losing most of the important rows which would in turn make the model weak. A trade-off had to be made to include the rows with N/As for a stronger and more robust model. The next step was to consider mean/aggregate of the columns having NA in the dataset. However, we realized that wasn't a viable alternative either as the data might then become biased. We then made sure that we subset our data and take care of the NA's by removing them from the NPS Type and the Likelihood_Recommend_H column while creating our models, because these are the important columns that affect the business recommendations.

2. We decided to take an analyzed and mathematical approach for classifying this dataset
    - First, we separated all the data based on each country. This helped us look at the problem more closely for a smaller subset of data.
    - Then we decided to calculate the total number of detractors and NPS score per country in order to devise business solutions according to the country.
    - We realized population might play a crucial factor in determining the number of detractors.
    - Owing to this, we calculated a fraction of the number of detractors upon the total number of people in the given dataset per country.
    - This fraction and NPS score was used as a filter to determine the country we would consider for business solutions.
    - The country that stood out with highest detractor percentage and a low NPS score was France, Jamaica and Guam. We did not include Jamaica and Guam as they had a large number of outliers.
    - We also decided to use USA as one of our viable countries because of the large target audience, as USA has the largest number of Hyatt Hotels in the world.
    - After obtaining 37 columns, we removed the columns having more than 60% of NA values and then categorized the likelihood to recommend column to low(0-6), high(9,10) and medium(7,8).

3. Once all these calculations were factored in, we had a clean dataset to start building our model on and analyze business solutions.

Code cleaning snippet

```
FebData <- fread(file= "FebData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,65,71,89,126,127,137:147,168,169,171,179,198,202,203,204,205,208,210,211,21
3,214,215,216,217,218,222,223,224,226,232)]
```

```
FullData <-
rbind(FebData,MarchData,AprilData,MayData,JuneData,JulyData,AugustData,SeptemberData,O
ctoberData,NovemberData,DecemberData,JanuaryData)

#removing NA's based on NPS and likelihood to recommend
FullData <- FullData[!(FullData$NPS_Type == "" | is.na(FullData$Likelihood_Recommend_H)), ]
#checking
#removing columns with more than 60% NA values
FullData <-  FullData[, -which(colMeans(is.na(FullData)) > 0.4)]
# NPS function
percentage_DetractorsAndNPS <- function(countrydata){
  promoter <- length(which(countrydata$NPS_Type == "Promoter"))
  detractor <- length(which(countrydata$NPS_Type == "Detractor"))
  passives <- length(which(countrydata$NPS_Type == "Passive"))
  total <- promoter + detractor + passives
  percent_detractor <- (detractor/total)*100
  NPS <- ((promoter - detractor)/total)* 100
  return(paste( "percentage of detractors are", percent_detractor, "and NPS score is", NPS))
}
FranceData <- subset(FullData, FullData$Country_PL=="France")
percentage_Detractors(FranceData)
```

Business Questions:

1. To determine the relationship and effect of purpose of visit ie. Business and leisure on the likelihood to recommend

We used Association rules to find an answer to this question. Firstly, we used our France data for the analysis.
We found that out of 11,000 people, around 9000 are for business visit.
After that we found that the residents of France have high likelihood to recommend
Support =0.15, confidence=0.56, lift=0.9
Then we found that for their US guests, the likelihood to recommend is also quite high
Support =0.16, confidence= 0.62, lift = 1.12
As the number of people who visit hotels for leisure are comparatively less in number, no significant impact of the leisure customers is found on the likelihood to recommend.

2. What is the effect of the type of room on likelihood to recommend

We have used a linear model to find the effect of room type on the likelihood to recommend. Room type would be the independent variable and likelihood to recommend being the dependent variable.
We found that the value of France – no significant impact

For the US data, we found that king size room and DDBL rooms have a positive effect on the likelihood to recommend.

```
[1] {}                          => {Likelihood_Recommend_H_Char=high} 0.69408518
0.6940852  1.0000000 517065
[2] {ROOM_TYPE_CODE_C=DLXK} => {Likelihood_Recommend_H_Char=high} 0.01048246
0.6942568  1.0002472   7809
[3] {ROOM_TYPE_CODE_C=DLXN} => {Likelihood_Recommend_H_Char=high} 0.01258727
0.7498601  1.0803574   9377
[4] {ROOM_TYPE_CODE_C=1BKN} => {Likelihood_Recommend_H_Char=high} 0.02302811
0.7256768  1.0455155  17155
[5] {ROOM_TYPE_CODE_C=VW1K} => {Likelihood_Recommend_H_Char=high} 0.02416375
0.6934664  0.9991085  18001
[6] {ROOM_TYPE_CODE_C=QNQN} => {Likelihood_Recommend_H_Char=high} 0.05499095
0.7368118  1.0615582  40966
[7] {ROOM_TYPE_CODE_C=DDBL} => {Likelihood_Recommend_H_Char=high} 0.11612585
0.6748288  0.9722564  86509
[8] {ROOM_TYPE_CODE_C=KING} => {Likelihood_Recommend_H_Char=high} 0.23507871
0.6960552  1.0028383 175124
```

3. What factors influence likelihood to recommend?

We considered the 10 metric columns from the data for linear modelling.
Initially we checked likelihood to recommend with each individual column but that did not give any satisfactory results. We Excluded overall satisfaction and tranquility because it is not appropriate to consider these metrics for likelihood to recommend.
The best model that we got gave us a Multiple R-squared value of 0.7136 and the Adjusted R-squared value of 0.7132.
We found that the factors that affect likelihood to recommend the most are: guest room satisfaction, hotel condition, customer service satisfaction, staff cared and overall F&B experience.

Linear modelling

```
View(FranceData)
model1 <- lm(data = FranceData, formula = Likelihood_Recommend_H ~ Guest_Room_H +
Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H +`F&B_Overall_Experience_H`)
summary(model1)
```

We then put the same data for analysis using the Random forest algorithm for France and we got almost similar results like that from the linear model with linear model giving slightly better results.

4. What is the effect of different facilities that the hotel provides on the likelihood to recommend?

We Implemented association rules mining to know what kind of facilities do the business customers and leisure customers use.

For the France dataset, we found that no utilities or facilities are being used by the customers. Hence there is no effect of these different facilities that the hotels in France are providing to their customers. Keeping in mind majority of the customers come for business purposes.

For the US dataset, we found that
For Business customers:
Leisure facilities like spa, golf, casinos are not used, hence there is no effect on the likelihood to recommend.
Even the conference rooms are not used and still the likelihood to recommend is high.
We found out that business centers are important as they are being used and also have an effect on likelihood to recommend.
For leisure customers: The hotels not having utilities like spa, golf course, mini- bar, casinos also have a high likelihood to recommend from the leisure customers.
We observed that both the countries have a minority of leisure travelers and that might be due to the reason that a lot of Hyatt hotels do not have these leisure facilities.

5. How is total revenue distributed across different revenue categories and which revenue has the greatest impact?

To keep the data consistent, we converted our revenue to USD.
Then by using linear modelling, we kept total revenue as the dependent variable and other revenue categories as the independent variables and found that :
Total revenue is highly dependent on the room revenue : 0.9685
and the Hyatt hotel chain can improvise on others like the food and beverages revenues to get a higher revenue.

Our Recommendations:

- The hotel chain should invest more on the hotel condition, customer service satisfaction, staff cared and overall Food &Beverages experience, for France.
- The hotel chain should invest more on the hotel condition, customer service satisfaction, staff cared and overall Food &Beverages experience.
- Hyatt hotels get maximum of their revenue through their room revenue, but as they also have restaurants and other facilities involved, they can try to improve them to generate more revenue through them.
- They should device exclusive offers for their regular customers based on their feedback.
- The hotel should focus on their leisure customers and try to attract them through promotion and by including leisure facilities in their hotels as majority of their customers are business customers.
- As the king size room and double bed room have a greater impact on likelihood, they should maintain these rooms and try to improve the conditions of the other rooms.

Lessons Learnt:

- Learnt different methods of cleaning, sorting and classifying huge datasets according to the needs.
- Understanding the hospitality industry and the factors that drive this industry.
- In-depth understand of various modeling and regression techniques.
- To ask the right questions.
- Understand the customers and the factors that satisfy them.

Challenges faced:

- Understanding the huge dataset
- Devising methods to deal with cleaning and sub-setting the dataset to keep the important data.
- Technical difficulties
- Time management

Conclusion:

With the analysis carried out on the data, we can conclude that below are the factors which affects and influences the NPS:

- Customer Service
- Staff Care
- Condition of the hotel
- Guest Room
- Food and beverages

Future Scope:

- The profile of the Customers who are categorized as Passive can be carefully analyzed. There will be comparably fewer factors to influence their Likelihood to Recommend. If they can be converted into Promoters, ensuring Customer Loyalty, the revenue of the Hotel Chain can be positively influenced.
- Applying text mining and sentiment analysis to the data obtained for getting more accurate result.
- This time our analysis included two countries, namely US and France. We would like to stream line our focus on specific hotels by separating them into business and leisure hotels and study city-wise data to get more specific and detailed information related to them.

Reflection:

It has been a great learning experience while working on this project. We not just took in the fundamental ideas of R Programming and Data Analysis but also learnt about group coordination and work incorporation. In our task refreshes, we had a go at following kanban work process and ensured everyone is a spoke in the wheel. From fundamental learning like formation of vector, cleaning the data to machine learning, we tried to actualize our class learnings into the task. It was a continuous and incremental use of our insights. Working in a group turned out to be extremely fascinating as we figured out how to incorporate all the viewpoints and in addition absorbed learning from every one of our colleagues. The help we got from our professor and TA was admirable and that extremely persuaded us to be on the track and we figured out how to convey our outcome amid the last portrayal.

##############################CODE##################################

```
library(data.table)
library(bit)
library(bit64)
library(graphics)
library(gsubfn)
library(proto)
library(RSQLite)
library(sqldf)
library(arules)
library(grid)
library(arulesViz)
#detach("package:arules", unload=TRUE)


FebData <- fread(file= "FebData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
MarchData <- fread(file= "MarchData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
AprilData <- fread(file = "AprilData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
MayData <- fread(file = "MayData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
```

```
JuneData <- fread(file = "JuneData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
JulyData <- fread(file = "JulyData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
AugustData <- fread(file = "AugustData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
SeptemberData <- fread(file = "SeptData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
OctoberData <- fread(file = "OctData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
NovemberData <- fread(file = "NovData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
DecemberData <- fread(file = "DecData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]
JanuaryData <- fread(file = "JanData.csv", header = TRUE, sep = ",",stringsAsFactors = FALSE
)[,c(11,12,23,26,28,30,32,65,89,137:145,147,168,169,171,179,202,203,204,205,210,213,214,21
5,216,218,223,232)]

#merging whole data
FullData <-
rbind(FebData,MarchData,AprilData,MayData,JuneData,JulyData,AugustData,SeptemberData,O
ctoberData,NovemberData,DecemberData,JanuaryData)
str(FullData)
#removing row names
rownames(FullData) <- NULL
dim(FullData)
head(FullData)
#removing NA's based on NPS and likelihood to recommend
FullData <- FullData[!(FullData$NPS_Type == "" | is.na(FullData$Likelihood_Recommend_H)), ]
#checking
head(FullData$Likelihood_Recommend_H,20)
head(FullData$NPS_Type,20)
head(FullData)
#rm(FullData)
#removing columns with more than 60% NA values
View(colMeans(is.na(FullData)) > 0.6)
#adding a likelihood_recommend_h_char column in full data
```

```
FullData$Likelihood_Recommend_H_Char <- FullData$Likelihood_Recommend_H
FullData$Likelihood_Recommend_H_Char <-
as.factor(FullData$Likelihood_Recommend_H_Char)
FullData$Likelihood_Recommend_H_Char <- gsub(pattern = "9|10", replacement = "high"
,FullData$Likelihood_Recommend_H_Char)
FullData$Likelihood_Recommend_H_Char <- gsub(pattern = "7|8", replacement = "medium"
,FullData$Likelihood_Recommend_H_Char)
FullData$Likelihood_Recommend_H_Char <- gsub(pattern = "1|2|3|4|5|6", replacement =
"low" ,FullData$Likelihood_Recommend_H_Char)

str(FullData)
##############################################################################
############################
#adding a numeric type NPS_Type
FullData$NPS_Type_num <- FullData$NPS_Type
FullData$NPS_Type_num <- as.factor(FullData$NPS_Type_num)
FullData$NPS_Type_num <- gsub(pattern = "Promoter",replacement =
"10",FullData$NPS_Type_num)
FullData$NPS_Type_num <- gsub(pattern = "Detractor",replacement =
"6",FullData$NPS_Type_num)
FullData$NPS_Type_num <- gsub(pattern = "Passive",replacement =
"8",FullData$NPS_Type_num)
##############################################################################
##############################
#NPS for whole dataset
promoters <- length(which(FullData$NPS_Type == "Promoter"))
View(promoters)
detractors <- length(which(FullData$NPS_Type == "Detractor"))
View(detractors)
passives <- length(which(FullData$NPS_Type == "Passive"))
View(passives)

total <- passives + promoters + detractors
View(total)
NPS <- ((promoters-detractors)*100)/total
View(NPS)
#output
#57.17
##############################################################################
memory.limit(100000)

# %of detractors
percentage_Detractors <- function(countrydata){
  promoter <- length(which(countrydata$NPS_Type == "Promoter"))
```

```
  detractor <- length(which(countrydata$NPS_Type == "Detractor"))
  passives <- length(which(countrydata$NPS_Type == "Passive"))
  total <- promoter + detractor + passives
  percent_detractor <- (detractor/total)*100
  return(paste( "percentage of detractors are",percent_detractor))
}

# NPS function
percentage_DetractorsAndNPS <- function(countrydata){
  promoter <- length(which(countrydata$NPS_Type == "Promoter"))
  detractor <- length(which(countrydata$NPS_Type == "Detractor"))
  passives <- length(which(countrydata$NPS_Type == "Passive"))
  total <- promoter + detractor + passives
  percent_detractor <- (detractor/total)*100
  NPS <- ((promoter - detractor)/total)* 100
  return(paste( "percentage of detractors are", percent_detractor, "and NPS score is", NPS))
}
View(FullData$Guest_Room_H)
##############################################################################
####################
#adding character columns for each metric
FullData$Guest_Room_H_char <- FullData$Guest_Room_H
FullData$Guest_Room_H_char <- as.factor(FullData$Guest_Room_H_char)
FullData$Guest_Room_H_char <- gsub("10|9","high",FullData$Guest_Room_H_char)
FullData$Guest_Room_H_char <- gsub("8|7","medium",FullData$Guest_Room_H_char)
FullData$Guest_Room_H_char <- gsub("0|1|2|3|4|5|6","low",FullData$Guest_Room_H_char)

FullData$Condition_Hotel_H_char <- FullData$Condition_Hotel_H
FullData$Condition_Hotel_H_char <- as.factor(FullData$Condition_Hotel_H_char)
FullData$Condition_Hotel_H_char <- gsub("10|9","high",FullData$Condition_Hotel_H_char)
FullData$Condition_Hotel_H_char <- gsub("8|7","medium",FullData$Condition_Hotel_H_char)
FullData$Condition_Hotel_H_char <-
gsub("0|1|2|3|4|5|6","low",FullData$Condition_Hotel_H_char)

FullData$Customer_SVC_H_char <- FullData$Customer_SVC_H
FullData$Customer_SVC_H_char <- as.factor(FullData$Customer_SVC_H_char)
FullData$Customer_SVC_H_char <- gsub("10|9","high",FullData$Customer_SVC_H_char)
FullData$Customer_SVC_H_char <- gsub("8|7","medium",FullData$Customer_SVC_H_char)
FullData$Customer_SVC_H_char <-
gsub("0|1|2|3|4|5|6","low",FullData$Customer_SVC_H_char)

FullData$Staff_Cared_H_char <- FullData$Staff_Cared_H
FullData$Staff_Cared_H_char <- as.factor(FullData$Staff_Cared_H_char)
FullData$Staff_Cared_H_char <- gsub("10|9","high",FullData$Staff_Cared_H_char)
```

```
FullData$Staff_Cared_H_char <- gsub("8|7","medium",FullData$Staff_Cared_H_char)
FullData$Staff_Cared_H_char <- gsub("0|1|2|3|4|5|6","low",FullData$Staff_Cared_H_char
FullData$Internet_Sat_H_char <- FullData$Internet_Sat_H
FullData$Internet_Sat_H_char <- as.factor(FullData$Internet_Sat_H_char)
FullData$Internet_Sat_H_char <- gsub("10|9","high",FullData$Internet_Sat_H_char)
FullData$Internet_Sat_H_char <- gsub("8|7","medium",FullData$Internet_Sat_H_char)
FullData$Internet_Sat_H_char <- gsub("0|1|2|3|4|5|6","low",FullData$Internet_Sat_H_char)

FullData$Check_In_H_char <- FullData$Check_In_H
FullData$Check_In_H_char <- as.factor(FullData$Check_In_H_char)
FullData$Check_In_H_char <- gsub("10|9","high",FullData$Check_In_H_char)
FullData$Check_In_H_char <- gsub("8|7","medium",FullData$Check_In_H_char)
FullData$Check_In_H_char <- gsub("0|1|2|3|4|5|6","low",FullData$Check_In_H_char)

View(colMeans(is.na(FullData)) > 0.6)
FullData$Overall_Sat_H <- NULL
FullData$Tranquility_H <- NULL
str(FullData)

colnames(FullData)
##############################################################################
#######
FranceData <- subset(FullData, FullData$Country_PL=="France")
str(FranceData)

#FranceData
##############################################################################
#######
#linear modelling

FranceData <- na.aggregate(FranceData)
View(FranceData)
model1 <- lm(data = FranceData[1:100000,], formula = Likelihood_Recommend_H ~
Guest_Room_H + Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + Check_In_H)
summary(model1)
library(ggplot2)

model2 <- lm(data = FranceData, formula = PMS_TOTAL_REV_USD_C ~
PMS_ROOM_REV_USD_C)
summary(model2)
#Multiple R-squared:  0.9461,          Adjusted R-squared:  0.9461
#here we can see that total revenue is based on heavily relied on room revenue in hyatt hotels
in france
```

```
model3 <- lm(data = FranceData, formula = PMS_TOTAL_REV_USD_C ~
PMS_FOOD_BEVERAGE_REV_USD_C)
summary(model3)
#Multiple R-squared:  0.4276,          Adjusted R-squared:  0.4275
#can improve food and beverage services to improve revenue

model4 <- lm(data = FranceData, formula = PMS_TOTAL_REV_USD_C ~
PMS_OTHER_REV_USD_C + REVENUE_USD_R )
summary(model4)
#Multiple R-squared:  0.7828,          Adjusted R-squared:  0.7828

ggplot(FranceData, aes(x=FranceData$Guest_Room_H,
y=FranceData$Likelihood_Recommend_H  , color=FranceData$NPS_Type)) +
  geom_smooth(method = "lm") +  ylab("LTR Rating") +  xlab(" Guest Room satisfaction metric")
+
  ggtitle(" Effect of Guest Room Rating rating on Net Promoter Score")

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

ggplot(FranceData, aes(x=FranceData$Customer_SVC_H,
y=FranceData$Likelihood_Recommend_H  , color=FranceData$NPS_Type)) +
  geom_smooth(method = "lm") +  ylab("LTR Rating") +  xlab(" customer service satisfaction
metric") +
  ggtitle(" Effect of customer service rating on Net Promoter Score")

ggplot(FranceData, aes(x=FranceData$Check_In_H, y=FranceData$Likelihood_Recommend_H  ,
color=FranceData$NPS_Type)) +
  geom_smooth(method = "lm") +  ylab("LTR Rating") +  xlab(" check in satisfaction metric") +
  ggtitle(" Effect of check in rating on Net Promoter Score")

##############################################################################
#######
#KVSM and SVM
rm(data1)
colnames(FranceData)
data1 <- FranceData[1:100000,c(10:17)]
View(data1)
randIndx <- sample(1:dim(data1)[1])
#taking 2/3rd of random data for training dataset and remaining 1/3rd for testing dataset
point <- floor(2*dim(data1)[1]/3)
traindta <- data1[randIndx[1:point],]
```

```
View(traindta)
testdta <- data1[randIndx[(point+1):dim(data1)[1]],]
View(testdta)
str(traindta)
traindta <- na.omit(traindta)
testdta <- na.omit(testdta)
library(kernlab)
#step3
KSVM <- ksvm(Likelihood_Recommend_H ~.,data = traindta, kernel="rbfdot",kpar="automatic",
C=5, cross=3, prob.model=TRUE)
KSVM
#Objective Function Value : -1624.44
#Training error : 0.120246
#Cross validation error : 1.66121
#Laplace distr. width : 2.633128
#here we can see that the training and cross validation error is very low
library(e1071)
SVM <- svm(data= traindta,Likelihood_Recommend_H~.)
SVM
predict <- predict(SVM, testdta,type="votes")
View(predict)
compTable <- data.frame(predict,testdta[,1])
View(compTable)
#renaming columns
colnames(compTable) <- c("Pred","test")
#calculating RMSerror
error <- sqrt(mean(compTable$test - compTable$Pred)^2)
error
#0.07239554
#calculating error
compTable$error <- abs(compTable$test - compTable$Pred)
#creatting data.frame to plot
data <-
data.frame(testdta$Likelihood_Recommend_H,testdta$Guest_Room_H,compTable$error)
library(ggplot2)
#plotting
ScatPlot <- ggplot(data,aes(x=testdta.Likelihood_Recommend_H,y=testdta.Guest_Room_H)) +
  geom_point(aes(size=compTable.error,color=compTable.error))
ScatPlot
##########################################################################
######
#arules
colnames(FranceData)
str(FranceData)
```

```
rm(FranceData_Arules)
FranceData_Arules <- FranceData[,c (1,3,8,35)]
#FranceData_Arules <- gsub("","NA",FranceData_Arules)
str(FranceData_Arules)
View(FranceData_Arules)
FranceData_Arules <-  replace(FranceData_Arules,TRUE,lapply(FranceData_Arules, factor))
FranceData_Arules <- sapply(FranceData_Arules,as.factor)

str(FranceData_Arules)
length(FranceData$POV_CODE_C)
rules <- apriori(FranceData_Arules, parameter = list(support = 0.09, confidence = 0.5))
inspect(rules)
summary(rules)
plot(rules,method= "graph",measure="support",shading="lift",engine= "interactive")
goodrules <- rules[quality(rules)$lift>1]
inspect(goodrules)
summary(goodrules)
plot(goodrules,method= "graph",measure="support",shading="lift",engine= "interactive")

#3
x <- is.maximal(goodrules)
inspect(goodrules[x])
#lhs                          rhs                          support confidence     lift count
#[1] {}                        => {Likelihood_Recommend_H_Char=high} 0.5521922  0.5521922
1.000000  5655
#[2] {Likelihood_Recommend_H_Char=low}    => {POV_CODE_C=BUSINESS}           0.1915829
0.9091752 1.023284  1962
#[3] {Likelihood_Recommend_H_Char=medium} => {POV_CODE_C=BUSINESS}
0.2126745  0.8970346 1.009620  2178
#[4] {GUEST_COUNTRY_R=UNITED STATES}      => {Likelihood_Recommend_H_Char=high}
0.1794747  0.6313981 1.143439  1838
#[5] {POV_CODE_C=BUSINESS,
#GUEST_COUNTRY_R=UNITED STATES}      => {Likelihood_Recommend_H_Char=high}
0.1559418  0.6218847 1.126211  1597

#association rules 2
colnames(FranceData)
FranceData_Arules2 <- subset(FranceData[,c(3,24:35,36)])
FranceData_Arules2 <- na.omit(FranceData_Arules2)
str(FranceData_Arules2)
View(FranceData_Arules2)

FranceData_Arules2 <-  replace(FranceData_Arules2,TRUE,lapply(FranceData_Arules2, factor))
rules2 <- apriori(FranceData_Arules2,parameter = list(support=0.78, confidence= 0.6))
```

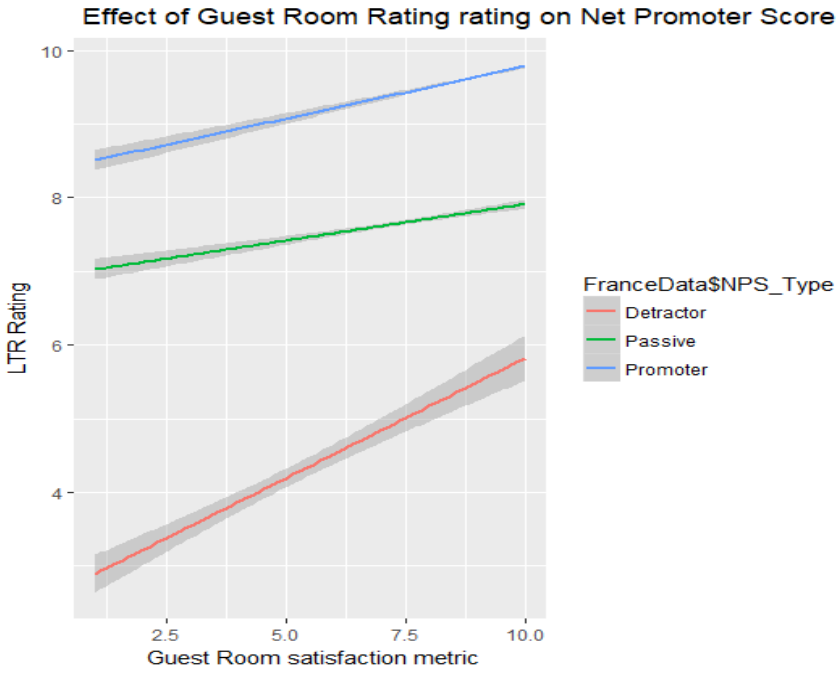
```
summary(rules2)
plot(rules2,method= "graph",measure="support",shading="lift",engine= "interactive")
inspect(rules2)

goodrules2 <- rules2[quality(rules2)$lift>1]
inspect(goodrules2)
summary(goodrules2)
plot(goodrules2,method= "graph",measure="support",shading="lift",engine= "interactive")

#3
x <- is.maximal(goodrules2)
inspect(goodrules2[x])

##########################################################################################
#####################################
##########################################################################################
#####################################
colnames(FranceData)
#random forest
install.packages("randomForest")
library(randomForest)
RandomforestData <- subset(FranceData[,c(10:17)])
str(RandomforestData)
View(RandomforestData)
RandomforestData <- na.omit(RandomforestData)

RandomforestData$Likelihood_Recommend_H <-
ifelse(RandomforestData$Likelihood_Recommend_H >= 6,1,0)
RandomforestData$`F&B_Overall_Experience_H` <-
as.integer(RandomforestData$`F&B_Overall_Experience_H`)

model4 <- randomForest(Likelihood_Recommend_H ~ Guest_Room_H + Condition_Hotel_H +
Staff_Cared_H + Internet_Sat_H  +  Customer_SVC_H + Staff_Cared_H , ntree=4 ,mtry=2,data =
RandomforestData)
model4
#% Var explained: 71.02
plot(model4)


##########################################################################################
#######
#USdata

USdata <- subset(FullData, FullData$Country_PL=="United States")
```

```
str(USdata)
#############################################################################
#######
#linear modelling
model5 <- lm(data= USdata, Likelihood_Recommend_H ~ Guest_Room_H + Condition_Hotel_H
+ Customer_SVC_H)
summary(model5)
#Multiple R-squared:  0.6829,          Adjusted R-squared:  0.6829
model8 <- lm(data = USdata, Likelihood_Recommend_H~ NPS_Type_num)
summary(model8)
#Multiple R-squared:  0.8656,          Adjusted R-squared:  0.8656
#############################################################################
#######
#A rules
colnames(USdata)
USdata_arules <- subset(USdata[,c(3,22:34)])
USdata_arules <- na.omit(USdata_arules)
str(USdata_arules)
View(USdata_arules)

USdata_arules <-  replace(USdata_arules,TRUE,lapply(USdata_arules, factor))
rules3 <- apriori(USdata_arules,parameter = list(support=0.9, confidence= 0.8))
summary(rules3)
plot(rules3)
plot(rules3,method= "graph",measure="support",shading="lift",engine= "interactive")
inspect(rules3)

goodrules3 <- rules3[quality(rules3)$lift>1]
inspect(goodrules3)
summary(goodrules3)
plot(goodrules3)

#3
x3 <- is.maximal(goodrules3)
inspect(goodrules3[x3])
#############################################################################
USleisureData <- subset(USdata, USdata$POV_CODE_C=="LEISURE")
View(USleisureData)
unique(USleisureData$POV_CODE_C)
colnames(USleisureData)

USleisureData_arules <- subset(USleisureData[,c(24:34,36)])
USleisureData_arules <-  replace(USleisureData_arules,TRUE,lapply(USleisureData_arules,
factor))
```

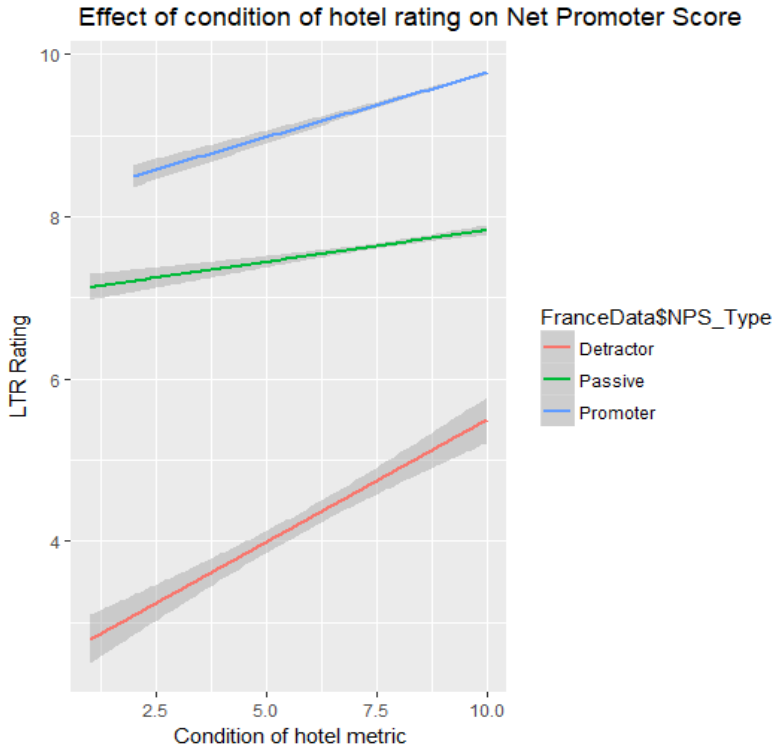
```
str(USleisureData_arules)

rules4 <- apriori(USleisureData_arules,parameter = list(support=0.6, confidence= 0.6))
summary(rules4)
plot(rules4)
inspect(rules4)

goodrules4 <- rules4[quality(rules4)$lift>1]
inspect(goodrules4)
summary(goodrules4)
plot(goodrules4)

#3
x4 <- is.maximal(goodrules4)
inspect(goodrules4[x4])
##############################################################################
# A rules for metrics
colnames(USdata)
data2 <- subset(USdata[,c(34:41)])
data2 <-  replace(data2,TRUE,lapply(data2, factor))
str(data2)

rules4 <- apriori(data2,parameter = list(support=0.6, confidence= 0.6))
summary(rules4)
plot(rules4)
inspect(rules4)

goodrules4 <- rules4[quality(rules4)$lift>1]
inspect(goodrules4)
summary(goodrules4)
plot(goodrules4,method= "graph",measure="support",shading="lift",engine= "interactive")

#3
x4 <- is.maximal(goodrules4)
inspect(goodrules4[x4])


##############################################################################
#plotting
library(sp)
install.packages("rworldmap")
library(rworldmap)

df <- data.frame(countryNames,Detractors,NPS)
```

```
df$countryNames <- gsub("([a-z])([A-Z])", "\\1 \\2",df$countryNames)
Worldmap <- joinCountryData2Map(df, joinCode="NAME",
                    nameJoinColumn="countryNames")
mapCountryData(Worldmap, nameColumnToPlot="NPS",colourPalette =
"topo",catMethod="fixedWidth",addLegend = TRUE,borderCol = "grey",mapTitle = "NPS
distribution across countries")
?mapCountryData
Worldmap2 <- joinCountryData2Map(df, joinCode="NAME",
                    nameJoinColumn="countryNames")
mapCountryData(Worldmap2, nameColumnToPlot="Detractors",colourPalette =
"heat",catMethod="fixedWidth",addLegend = TRUE,borderCol = "black",mapTitle = "Percentage
of Detractors across countries")
```

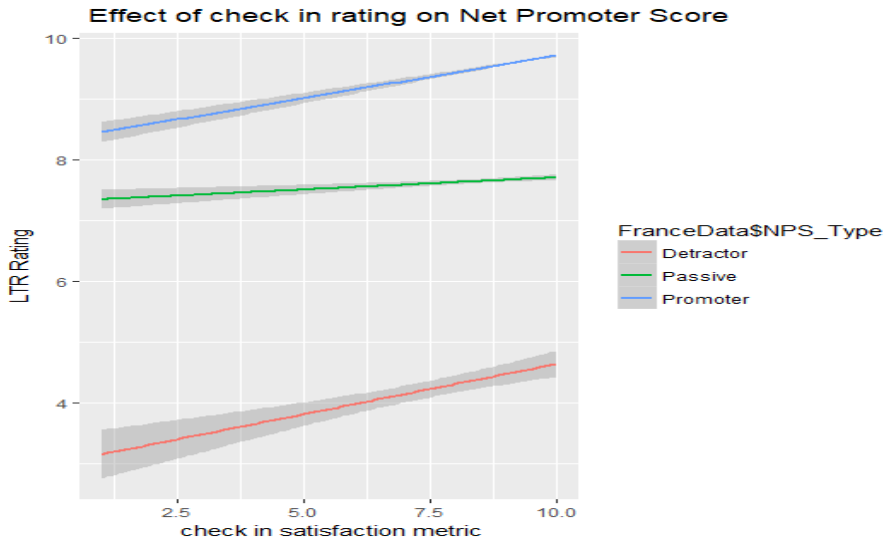
Visualization:

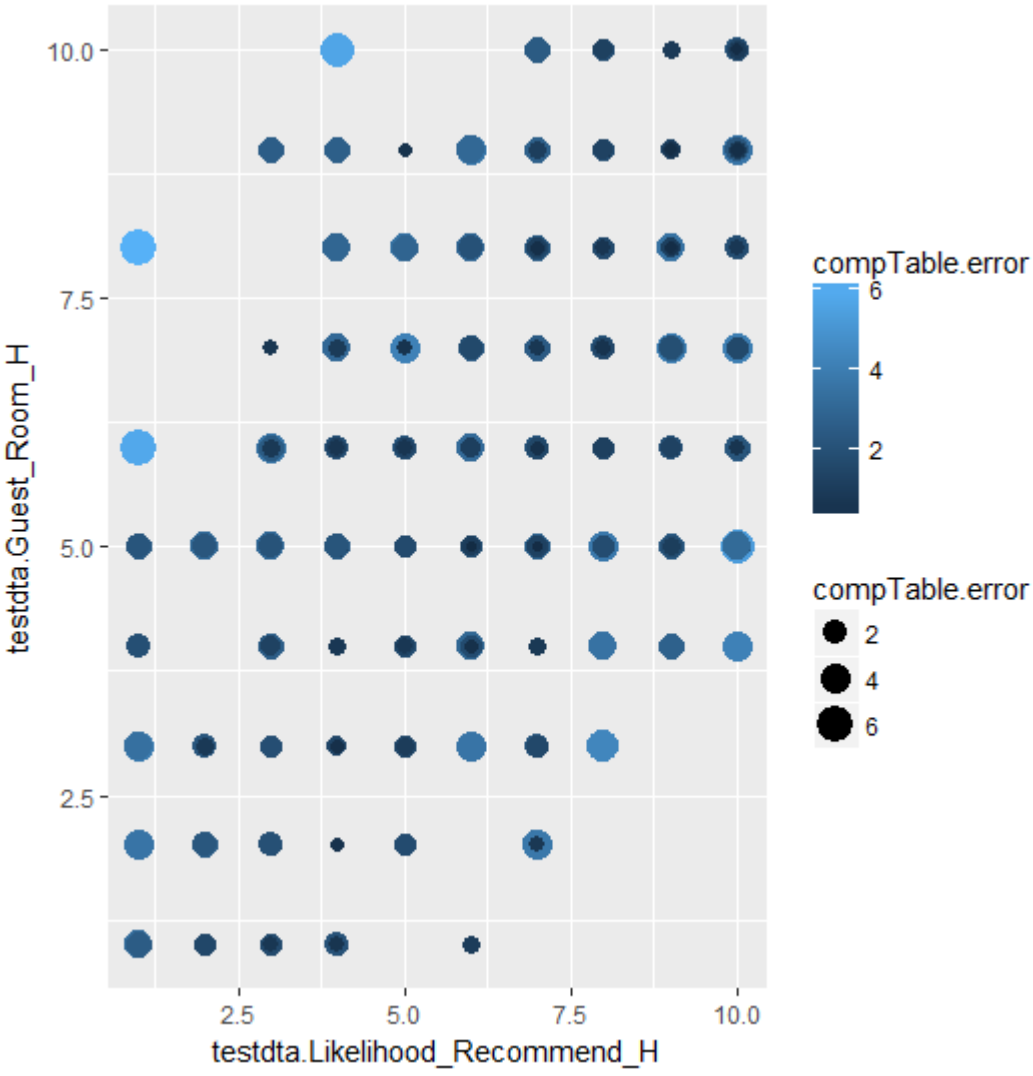


n:

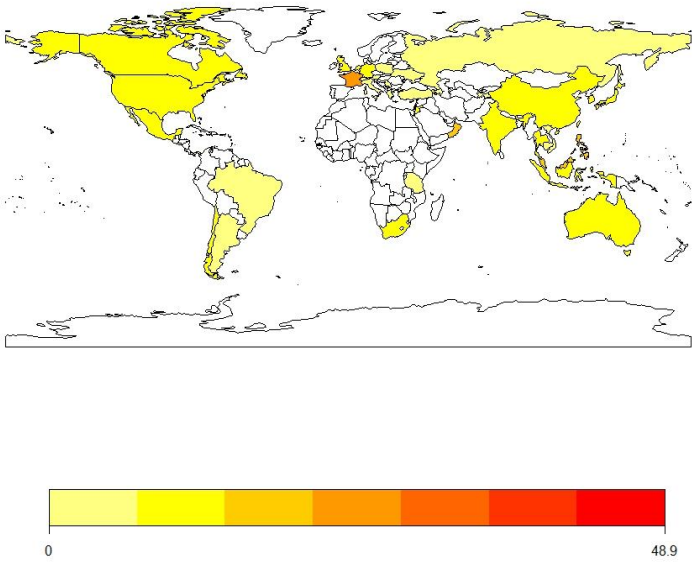

Effect of condition of hotel rating on Net Promoter Score
LTR Rating
10
8
6
4
2.5
5.0
7.5
10.0
Condition of hotel metric
FranceData$NPS_Type
Detractor
Passive
Promoter


Effect of customer service rating on Net Promoter Score
LTR Rating
10
8
6
4
2
2.5
5.0
7.5
10.0
customer service satisfaction metric
FranceData$NPS_Type
Detractor
Passive
Promoter

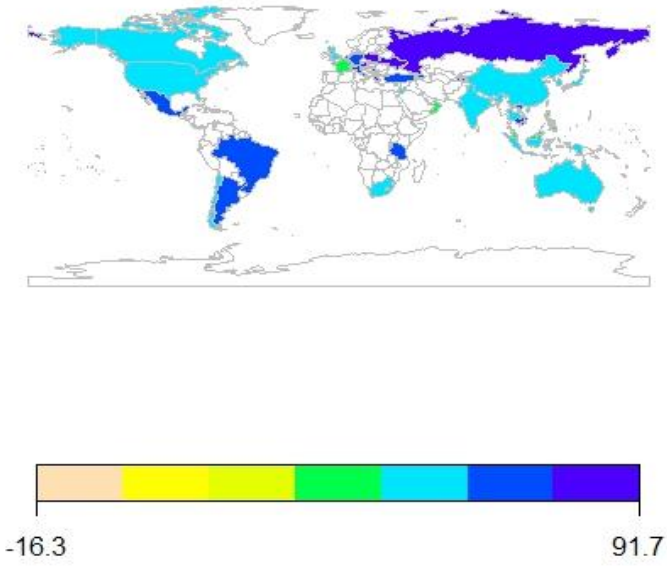

Effect of check in rating on Net Promoter Score
LTR Rating
check in satisfaction metric
FranceData$NPS_Type
Detractor
Passive
Promoter


testdta.Guest_Room_H
testdta.Likelihood_Recommend_H
compTable.error
compTable.error

Percentage of Detractors across countries



NPS



References:

1. Godwin, H. (2011) Merge all files in a directory using R into a single dataframe. Available at: https://www.r-bloggers.com/merge-all-files-in-a-directory-using-r-into-a-single-dataframe/

2. Byers, T. (2015) Ggplot 2.0.0. Available at: https://blog.rstudio.org/2015/12/21/ggplot2-2-0-0/

3. CheckMarket (2011) Net promoter score (NPS) - use, application and pitfalls. Available at: https://www.checkmarket.com/blog/net-promoter-score/

4. Geom_boxplot. Ggplot2 2.1.0 (no date) Available at: http://docs.ggplot2.org/current/geom_boxplot.html

5. Robk, R.K. - (2014) Quick-r: Pie charts. Available at: http://www.statmethods.net/graphs/pie.html

6. Systems, S. (2016) What is net promoter? Available at: https://www.netpromoter.com/know/.

7. Legends (ggplot2) Available at: http://www.cookbook-r.com/Graphs/Legends_(ggplot2).