IST 687

Applied Data Science

Hyatt Group Data analysis

Group 1

-Bhavani Shankar

-Prakhar Goyal

-Sahil Arora

-Sanjana Kataria

-Sapan Badjatiya

-Shikhar Agrawal

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. The entire dataset contains about 3 Million responses collected from the Hyatt Customer Survey from a time span of Feb 2014 to Jan 2015. There are about 240 columns for each observation, some columns give details about the person who responded to the survey (example: guest title, guest preferred language), some attributes about the hotel (example: Location, Spa, Type) and a column that indicates whether the person is promoter, passive or detractor. In addition, the survey was used to determine if a customer is likely to be a "promoter" or "detractor" for Hyatt Hotels, with complacent people being labelled as "passive".

Talking more about the Data

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 for the data analysis**
- **Visualize Data on the concluded basis**
- **Finding correlation**
- **Use modelling techniques**
- **Give recommendations based on analysis**

Business Aspect and Impact of the 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.

We have separated and strategized our data analysis into steps which are as follows :

Step 1: DATA CLEANING

We considered 12 months data i.e. from February 2014 to January 2015. We cleaned the different month's data thereby creating a new dataset which had 34 attributes of the original 237 and eliminated rows which had NA attribute from all the 12 months data.

While analysing the original datasets, we found that each dataset had about 13-14 lakh tuples out of which only 250,000-260,000 rows had a value. 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:

- State distribution of Customers
- Response of Customers towards Hotel facilities
- 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.

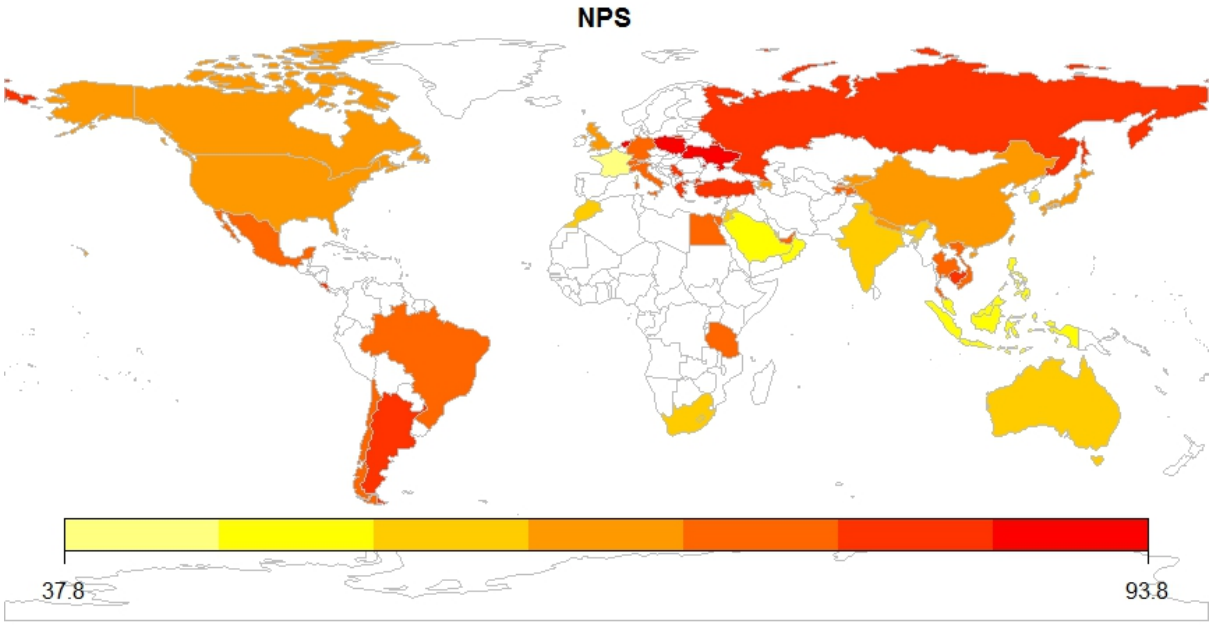
- *What is the region wise NPS of different countries?*
- *How important are hotel facilities in determining NPS?*
- *Which purpose of visit needs more focus: Business or Leisure?*
- *Do amenities play a significant role in increasing NPS?*

STEP 2: NPS CALCULATION

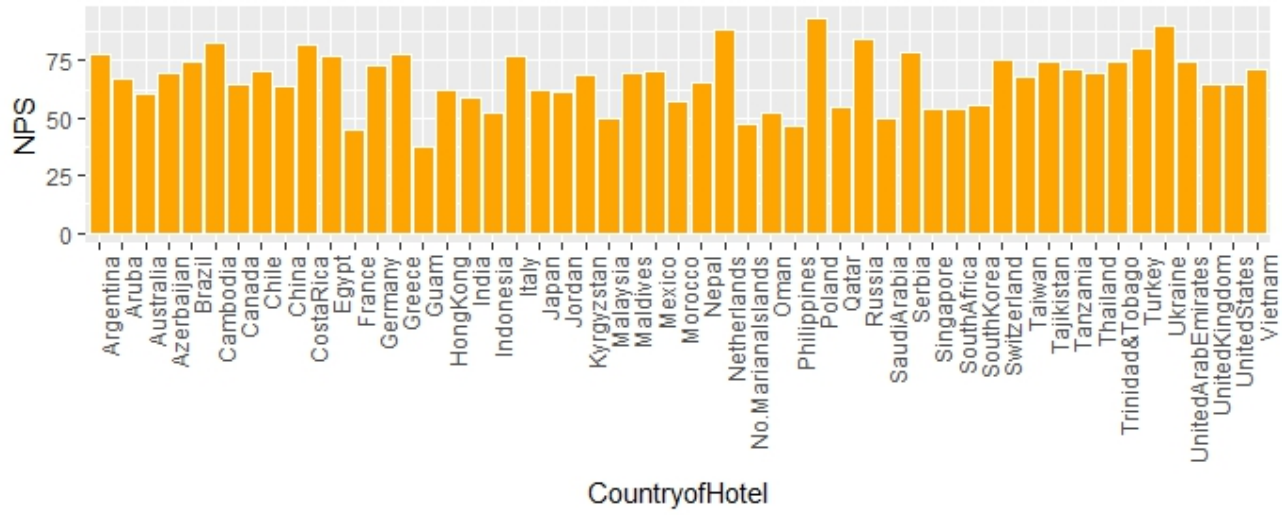
The **Net Promoter Score** is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company's products or services to others. It is used as a proxy for gauging the customer's overall satisfaction with a company's product or service and the customer's loyalty to the 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.

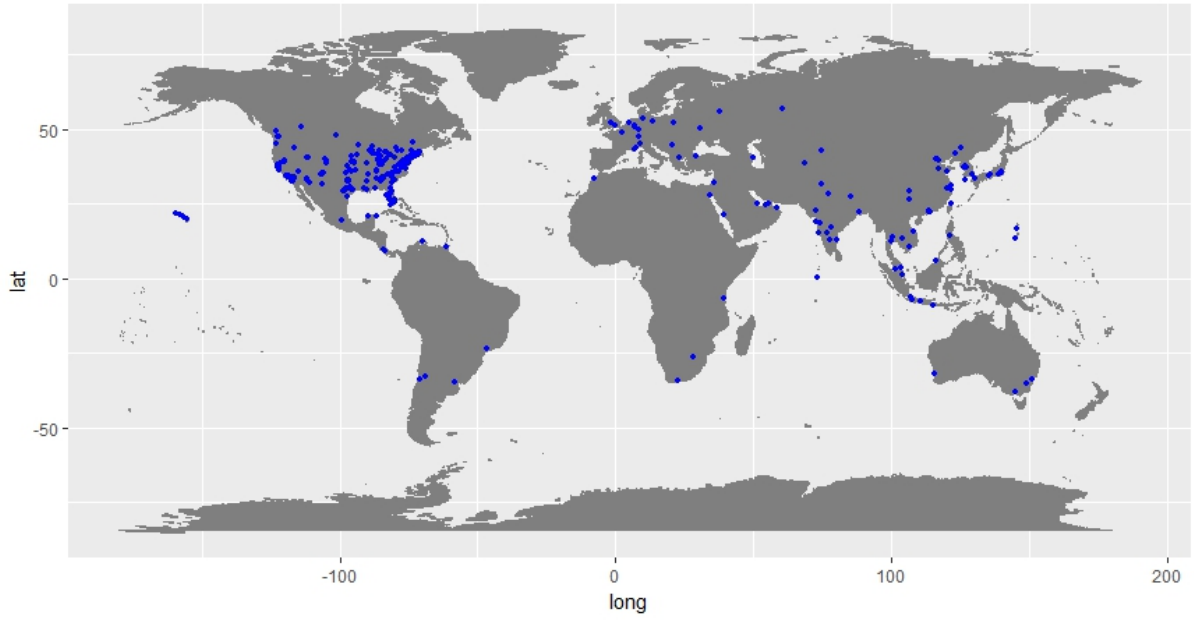


A few facts which can be observed from this plot have been stated below:

- 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

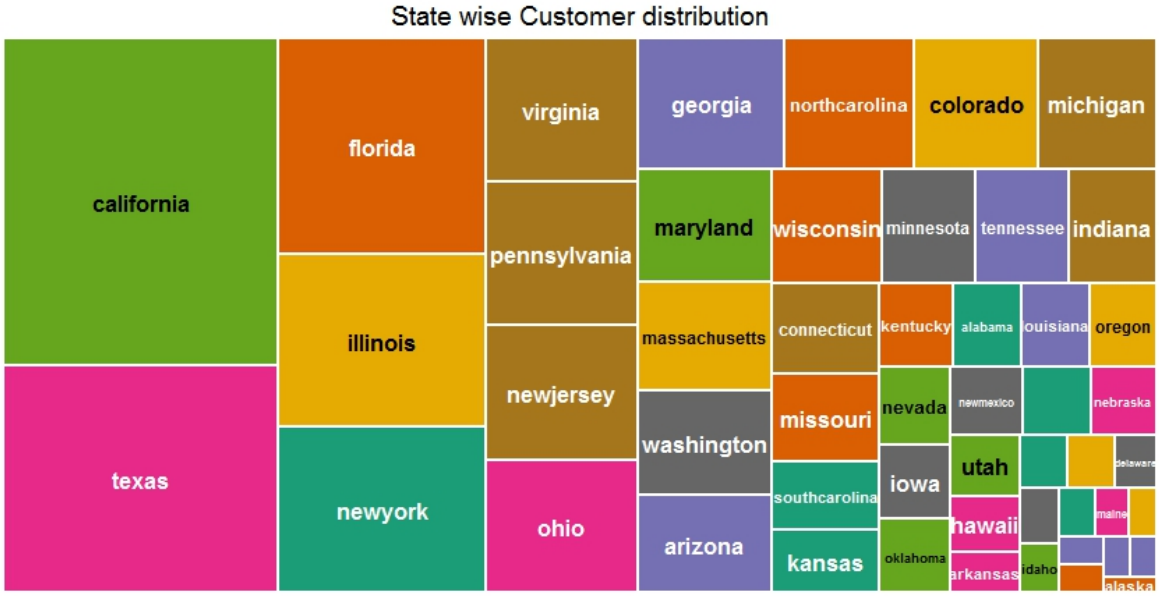
We plotted a world map which shows frequency of hotels visited around the world during Feb 2014 – Jan 2015.



- United States has highest number of hotels.
- Country with second largest number of hotels is China.

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 United States had the maximum number of hotels which were almost 10 times compared to other countries.

Since United States had highest number of customers and also NPS value is above average, we decided to further explore states in United States and their effect on overall NPS for United States



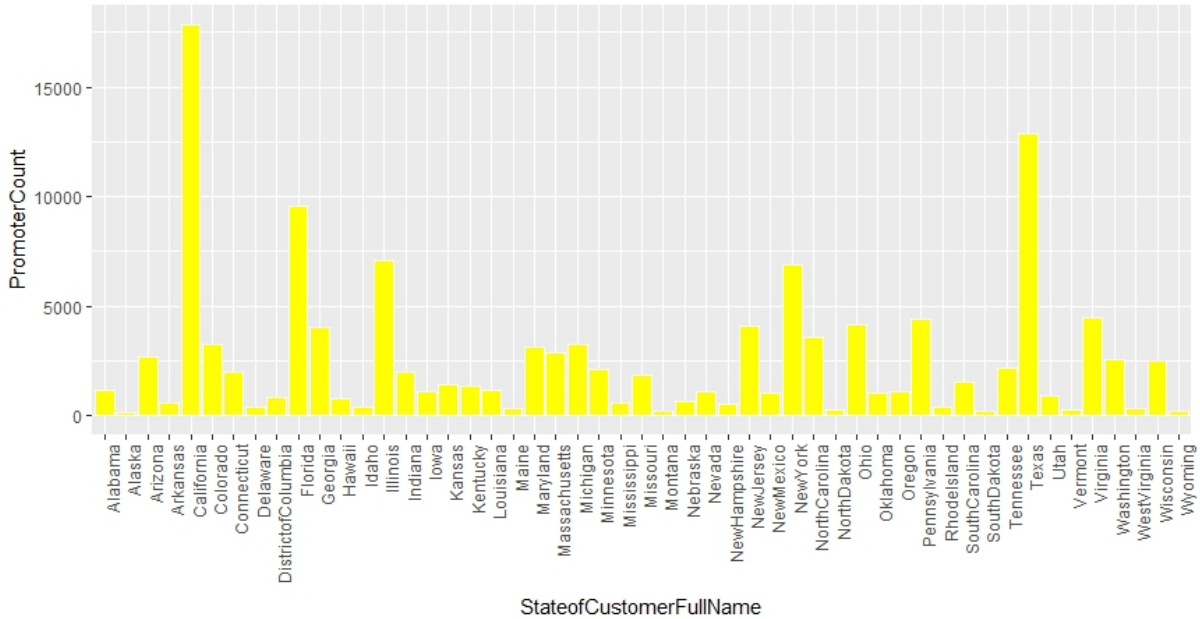
After plotting the tree map of states of United States based on their frequency of occurrence in the given data set, we found out that California is the state with highest number of customers followed by Texas.

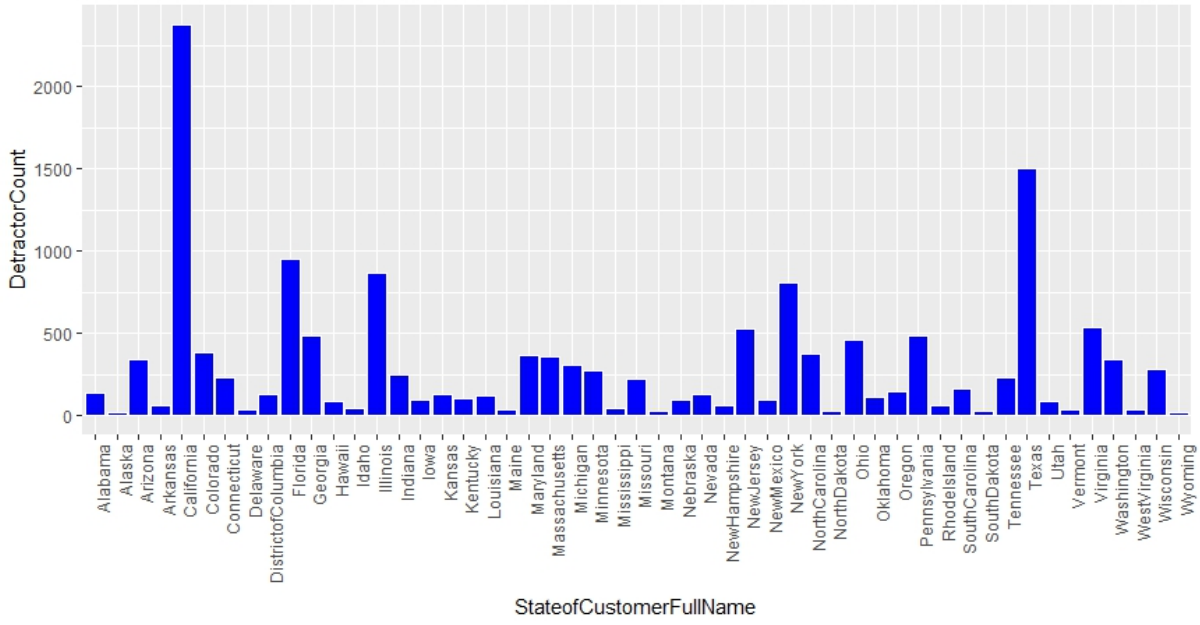
To further bolster our selection of United States as our region for NPS analysis, we plotted a word cloud depicting various countries. The size of the name of the country shows the country with maximum number of hotels.

Here it can be seen that United States has the highest number of Hotels.

STEP 4: VIZUALIZATION OF DATA

Visualizing data to understand its clarity and correlation is best way to answer various business questions.





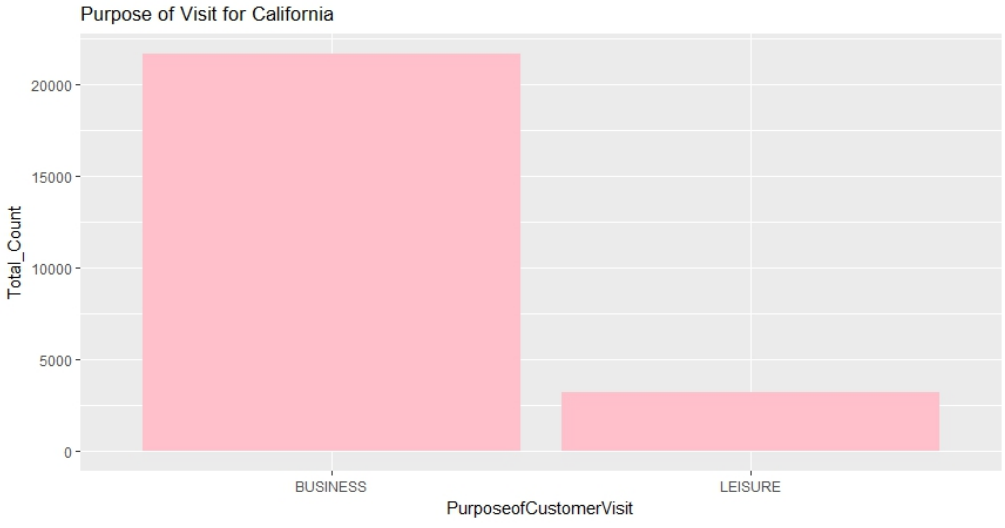
The above analysis shows number of promoters and detractors in United States in a span of 1 year.

From the above data following conclusions were made:

a. California has the number of promoter count, almost 17500.
b. California also has highest number of detractors, almost 2300.
c. Texas has second largest number of promoters, almost 12500.
d. Texas also has second largest number of detractor, almost 1500.

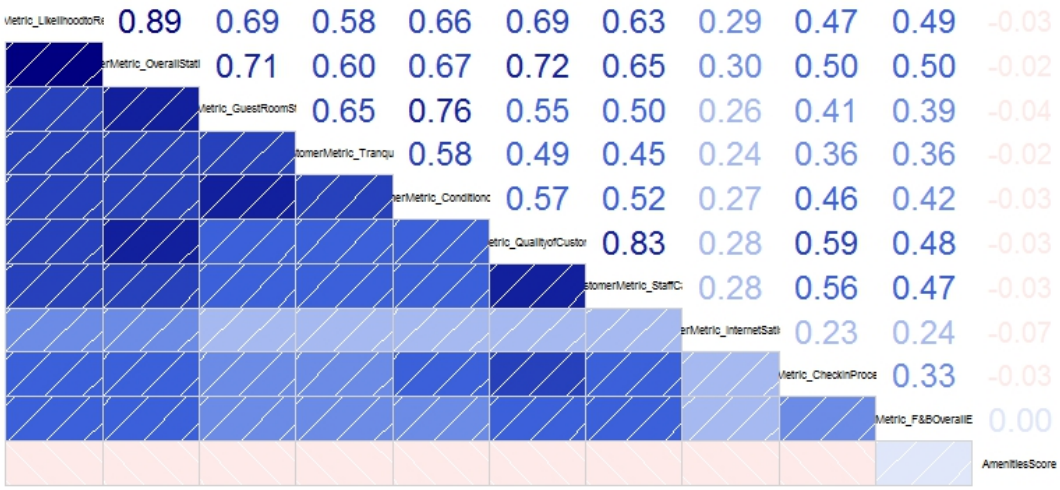
As we can see that, California and Texas being states with huge population, the number of promoters are way more than number of detractors.

Based on this analysis we tried to find out that from state of California, for what purposes people visited more.

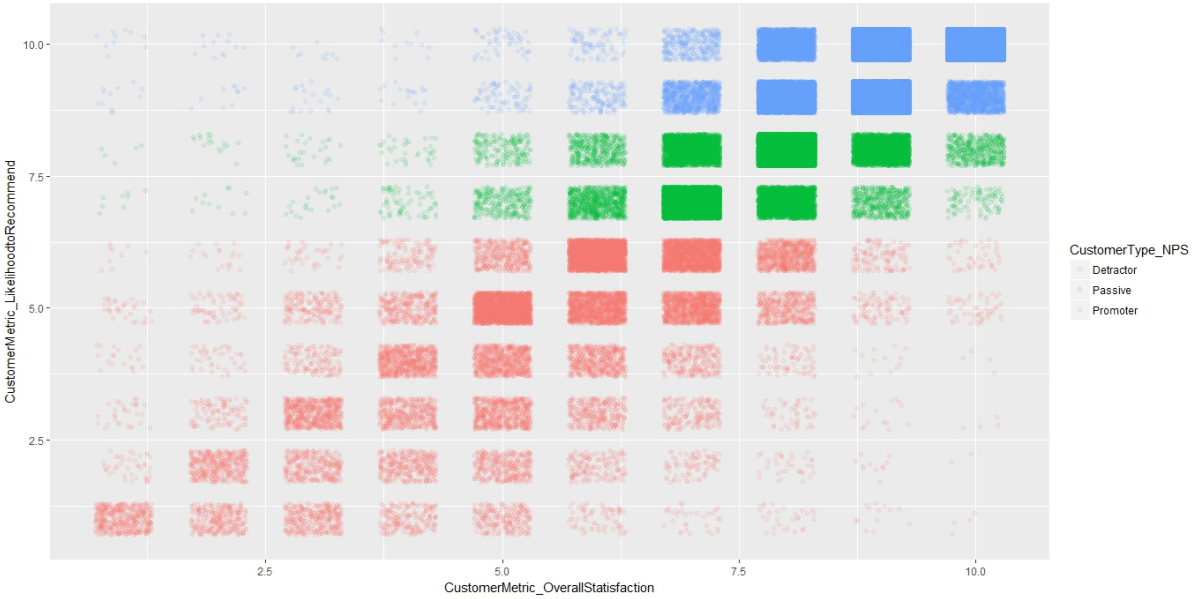


From above graph we can conclude that from state of California, maximum people visited Hyatt Hotel for business purposes. Therefore the people who affected the promoter score were business people.

After this, using correlation we found out metrics which affected the NPS most by assuming that Likelihood to Recommend is directly proportional to NPS.

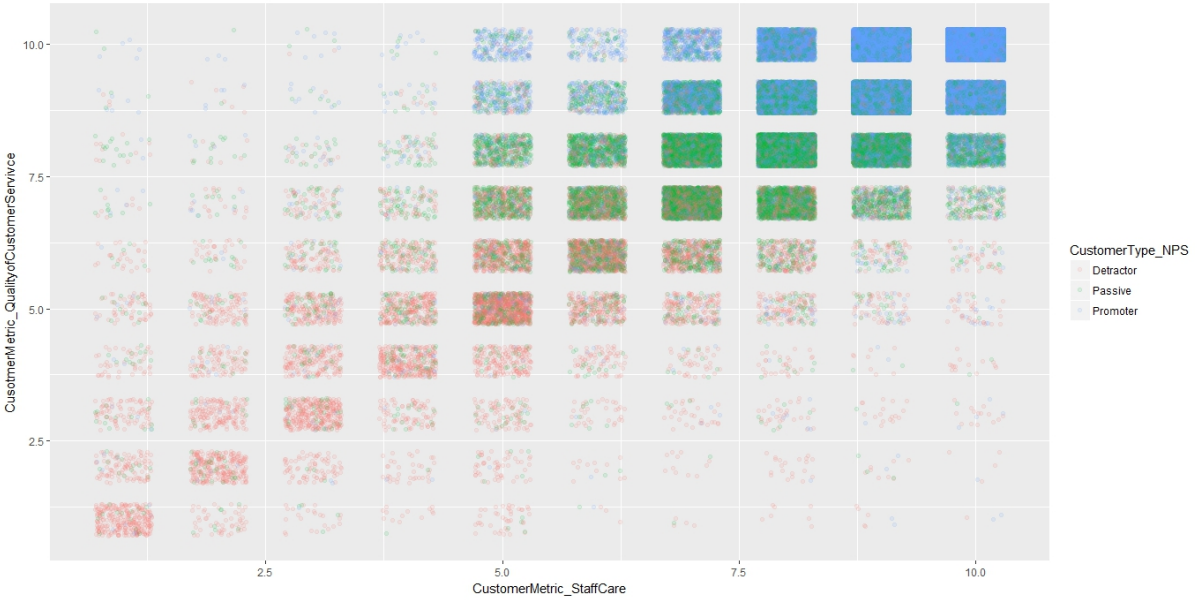


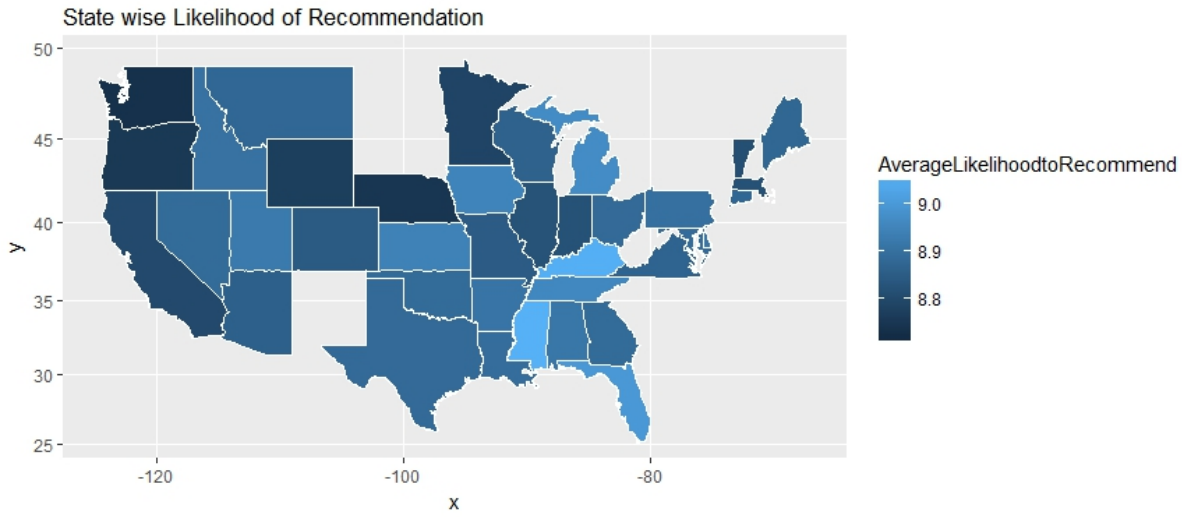
With this corrgram we found that, Likelihood to recommend is highly dependent on Customer Overall Satisfaction. We tried plotting a graph of Likelihood to Recommend vs Overall Satisfaction based on NPS_Type to understand NPS_Type varies with these two metrics.



We can see that as the values of Likelihood to Recommend and Overall Satisfaction increase proportionately, NPS_Type also changes from Detractors to Promoters.

From the corrgram above we also found out that two other metrics i.e. Quality of Customer Service and Staff Care were highly interdependent on each other. We again graphed the two metrics based on NPS_Type two understand how change in their value affects NPS_Type.





We mapped average likelihood to recommend on a state map of USA. This was done to further bolster our decision of selecting California as a state of our analysis as the value average likelihood to recommend for California is between 8.8-8.9. Even though promoter count is high in California compared to other states, there is still room for promoter base growth.

To further bolster our assumptions we did linear modelling on the metrics as described below.

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

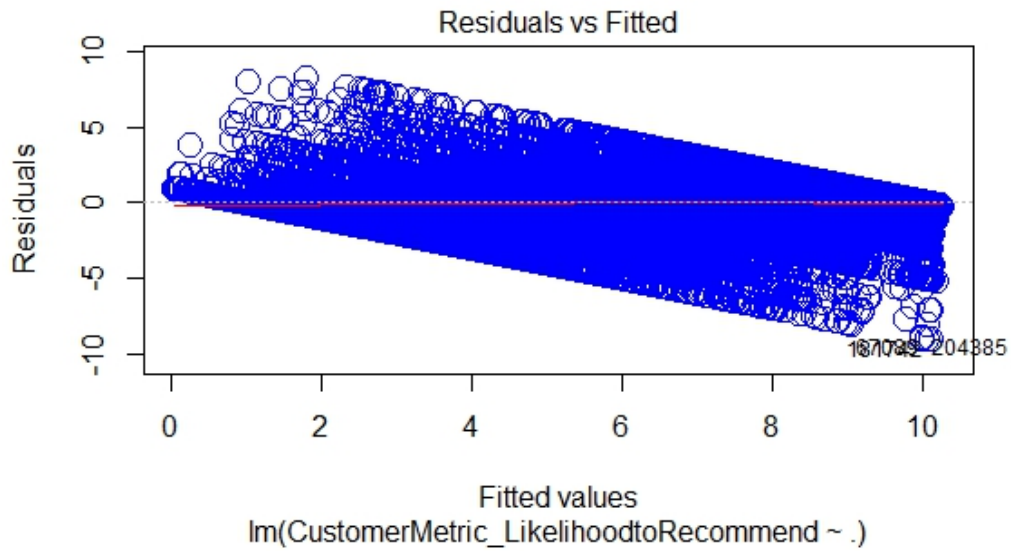
For hotels worldwide, we conducted Linear modelling on Hotel where people have quantified their experience with respect to various services provided by the hotel.

```
Call:
lm(formula = CustomerMetric_LikelihoodtoRecommend ~ ., data = correlationMatrixYearly)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1213 -0.1228 -0.0688  0.1197  8.1726

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -0.9978916  0.0130736 -76.329   <2e-16 ***
CustomerMetric_OverallStatisfaction  0.8097176  0.0017413 464.995   <2e-16 ***
CustomerMetric_GuestRoomStatisfaction  0.0657900  0.0017075  38.530   <2e-16 ***
CustomerMetric_Tranquility            0.0236159  0.0011642  20.285   <2e-16 ***
CustomerMetric_ConditionofHotel       0.0856671  0.0018211  47.042   <2e-16 ***
CusotmerMetric_QualityofCustomerServivce  0.0681731  0.0022749  29.968   <2e-16 ***
CustomerMetric_StaffCare              0.0453785  0.0018602  24.394   <2e-16 ***
CustomerMetric_InternetSatisfaction   0.0075030  0.0007698   9.747   <2e-16 ***
CustomerMetric_CheckInProcessQuality -0.0171345  0.0013751 -12.461   <2e-16 ***
`CustomerMetric_F&BOverallExperience`  0.0232102  0.0010857  21.378   <2e-16 ***
AmenitiesScore                        0.0004792  0.0009350   0.512    0.608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7333 on 256783 degrees of freedom
Multiple R-squared:  0.8098,    Adjusted R-squared:  0.8098
F-statistic: 1.093e+05 on 10 and 256783 DF,  p-value: < 2.2e-16
```

The value of R squared with amenities is 0.8094

From this model it is evident that all the services provided by the hotel play a significant role in determining the NPS except for Amenities used by the customer. If more amenities were accessed by the customer, the NPS would increase accordingly. This assumptions can be bolstered by the corrgram also, where correlation between likelihood to recommend and amenities score is negative.

To further advocate our linear model analysis, we plotted a Residual vs Fitted plot, to find out if our model was linear or not.



As we can see that the red line is straight, this concludes that the model was linear and significant.

STEP 6: PROVIDING RECOMMENDATIONS

- 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.
- 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.
- 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 to gain their loyalty towards the hotel could also lead to increase in the NPS score of the hotel
- They could implement some offers where in the customers are lured to use them on a trial basis and later if they like it, they can pay for it and hence promote better marketing strategies.
- More spending on renovation and maintenance of Hotel Facilities
- Training programs for staff like bell staff to improve customer service
- Plan a spa and a fitness centre to improve tranquillity
- Special offers for previous customers providing them with offers based on their feedback
- Customer service, Room Type and Room conditions should be of prime focus especially KY and MS states
- Investing on high end internet services should be of least priority because it has low impact on Net Promoter Score
- More family offers for winter season as customers prefer to stay with the family during the Christmas, Hanukkah and New Years
- Promotional offers in Hotels close to the beach to attract more customers during the Summer Seasons

CODE INTEGRATED

```
library(openintro)
library(ggplot2)
library(ggmap)
library(tm)
library(stringr)
library(wordcloud)
library(treemap)
library(Hmisc)
library(corrgram)
library(sqldf)
library(maps)
library(rworldmap)
```

```
library(data.table)
library(arules)
library(arulesViz)
#
#Reading the dataset
#
FebDataHyattHotel <- read.csv("out-
201402.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
MarchDataHyattHotel <-  read.csv("out-
201403.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
AprilDataHyattHotel <-  read.csv("out-
201404.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
MayDataHyattHotel <-  read.csv("out-
201405.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
JuneDataHyattHotel <-  read.csv("out-
201406.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
JulyDataHyattHotel <-  read.csv("out-
201407.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
AugDataHyattHotel <-  read.csv("out-
201408.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
SepDataHyattHotel <-  read.csv("out-
201409.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
OctDataHyattHotel <-  read.csv("out-
201410.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
NovDataHyattHotel <-  read.csv("out-
201411.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
DecDataHyattHotel <-  read.csv("out-
201412.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
JanDataHyattHotel <-  read.csv("out-
201501.csv")[,c(19,23,55,56,70,137:145,147,167,171,172,175:177,179,187,
191,202:205,214:216,218,221,232)]
#
#Combining all the data sets into a data set
#
YearlyDataSet <-
rbind(JanDataHyattHotel,FebDataHyattHotel,MarchDataHyattHotel,AprilData
HyattHotel,MayDataHyattHotel,JuneDataHyattHotel,JulyDataHyattHotel,AugD
ataHyattHotel,SepDataHyattHotel,OctDataHyattHotel,NovDataHyattHotel,Dec
DataHyattHotel)
#
#Removing NA's
#
YearlyDataSet <- na.omit(YearlyDataSet)
#
#Renaming row names
#
row.names(YearlyDataSet) <- NULL
```

```
#
#Renaming the columns
#
colnames(YearlyDataSet) <-
c("LengthofCustomerStay","PurposeofCustomerVisit","StateofCustomer","Co
untryofCustomer","FlagforOfferUsed","CustomerMetric_LikelihoodtoRecomme
nd","CustomerMetric_OverallStatisfaction","CustomerMetric_GuestRoomStat
isfaction","CustomerMetric_Tranquility","CustomerMetric_ConditionofHote
l","CusotmerMetric_QualityofCustomerServivce","CustomerMetric_StaffCare
","CustomerMetric_InternetSatisfaction","CustomerMetric_CheckInProcessQ
uality","CustomerMetric_F&BOverallExperience","CityofHotel","CountryofH
otel","OperationRegionofHotel","LatitudeofHotel","LongitudeofHotel","Ho
telLocalCurrency","Hotel'sNPSGoals","TotalMeetingSpaceinHotel","Regiono
fHotel","HotelFlag_BusinessCenter","HotelFlag_Casino","HotelFlag_Nearby
ConferenceCenter","HotelFlag_ConventionSpace","HotelFlag_MiniBar","Hote
lFlag_IndoorPool","HotelFlag_OutdoorPool","HotelFlag_Resort","HotelFlag
_ShuttleService","CustomerType_NPS")
#
#
#Converting CustomerType_NPS column to character vector
#
YearlyDataSet$CustomerType_NPS <-
as.character(YearlyDataSet$CustomerType_NPS)
#
#Segregating the dataset into different datasets
#
AmenitiesDataSetYearly <- subset(YearlyDataSet,select = c(25:33))
MetricDataSetYearly <- subset(YearlyDataSet,select = c(19,20,6:15))
LocationDataSetYearly <- subset(YearlyDataSet,select =c(16:20))
UserInfoDataSetYearly <- subset(YearlyDataSet,select = c(19,20,1:5,34))
#
#Changing row names
#
row.names(AmenitiesDataSetYearly) <- NULL
row.names(MetricDataSetYearly) <- NULL
row.names(LocationDataSetYearly) <- NULL
row.names(UserInfoDataSetYearly) <- NULL
#
#Converting state abbreviations to full names using library openintro
#
UserInfoDataSetYearly$StateofCustomerFullName <-
abbr2state(UserInfoDataSetYearly$StateofCustomer)
#
#In ameneties data set replacing all Y with 1 and N with 0
#
AmenitiesDataSetYearly <-
as.data.frame(ifelse(AmenitiesDataSetYearly=="Y",1,0))
#
#Converting user state and country name to character vector to find out
geocodes
#
UserInfoDataSetYearly$StateofCustomerFullName <-
as.character(UserInfoDataSetYearly$StateofCustomerFullName)
UserInfoDataSetYearly$CountryofCustomer <-
as.character(UserInfoDataSetYearly$CountryofCustomer)
UserInfoDataSetYearly$StateofCustomer <-
as.character(UserInfoDataSetYearly$StateofCustomer)
#
```

```
#Counting number of 1's in Amenities data set
#
AmenitiesDataSetYearly$Countof1 <- rowSums(AmenitiesDataSetYearly==1)
#
#Converting Customer State Full Name to charcater vector to convert it
into a word corpus
#
UserInfoDataSetYearly$StateofCustomer <-
as.character(UserInfoDataSetYearly$StateofCustomer)
#
#Clearing blankspaces in Country ans State Names
#
UserInfoDataSetYearly$CountryofCustomer <-
str_replace(UserInfoDataSetYearly$CountryofCustomer," ","")
UserInfoDataSetYearly$StateofCustomerFullName <-
str_replace_all(UserInfoDataSetYearly$StateofCustomerFullName," ","")
#
#Adding hotel latitude and longitude to amenities data set
#
AmenitiesDataSetYearly$LatitudeofHotel <-
LocationDataSetYearly$LatitudeofHotel
AmenitiesDataSetYearly$LongitudeofHotel <-
LocationDataSetYearly$LongitudeofHotel
#
#Adding statefullname column to each data set
#
AmenitiesDataSetYearly$CountryofHotel <-
LocationDataSetYearly$CountryofHotel
LocationDataSetYearly$StateofCustomerFullName <-
UserInfoDataSetYearly$StateofCustomerFullName
MetricDataSetYearly$StateofCustomerFullName <-
UserInfoDataSetYearly$StateofCustomerFullName
MetricDataSetYearly$CountryofHotel <-
LocationDataSetYearly$CountryofHotel
#
#create a map-shaped window
#
mapDevice('x11')
#
#Creating a function to generate a frequency data frame
#
SpecifiedFreqMatrix <- function(x){
  VectorSource <- VectorSource(x)
  WordCorpus <- Corpus(VectorSource)
  TDM <- TermDocumentMatrix(WordCorpus)
  WordMatrix <- as.matrix(TDM)
  WordCount <- rowSums(WordMatrix)
  WordCount <- sort(WordCount,decreasing = TRUE)
  CloudFrame <- data.frame(Name=names(WordCount),freq=WordCount)
  return(CloudFrame)
}
#
#Creating a data frame to find out frequency of most frequent state of
booking
#
StateCloudFrame <-
SpecifiedFreqMatrix(UserInfoDataSetYearly$StateofCustomerFullName)
#
```

```
#Creating a treemap to find out number of reservations made from each
state
#
UserStateTreeMap<- treemap(StateCloudFrame,index =
c("Name"),vSize="freq",type="index",palette = "Dark2",title = "State
wise Customer distribution",fontsize.title = 14,fontsize.labels =
12,border.col = "white")
#
#Creating a data frame to find out frequency of most frequent country
of booking
#
CountryCloudFrame <-
SpecifiedFreqMatrix(UserInfoDataSetYearly$CountryofCustomer)
#
#Creating a treemap to find out number of reservations made from each
country
#
UserCountryTreeMap<- treemap(CountryCloudFrame,index =
c("Name"),vSize="freq",type="index",palette = "Dark2",title = "Country
wise Customer distribution",fontsize.title = 14,fontsize.labels =
12,border.col = "white")
#
#Converting CityofHotel, CountryofHotel and OperationRegionofHotel in
LocationDataSet to character vector
#
LocationDataSetYearly$CityofHotel <-
as.character(LocationDataSetYearly$CityofHotel)
LocationDataSetYearly$CountryofHotel <-
as.character(LocationDataSetYearly$CountryofHotel)
LocationDataSetYearly$OperationRegionofHotel <-
as.character(LocationDataSetYearly$OperationRegionofHotel)
#
#Removing spaces from CityofHotel column in Location Data Set
#
LocationDataSetYearly$CityofHotel <-
str_replace_all(LocationDataSetYearly$CityofHotel," ","")
LocationDataSetYearly$CountryofHotel <-
str_replace_all(LocationDataSetYearly$CountryofHotel," ","")
#
#Creating a word cloud to find out most visted hotel citywise
#
HotelCityWordCloudFrame <-
SpecifiedFreqMatrix(LocationDataSetYearly$CityofHotel)
HotelCityWordCloud <-
wordcloud(HotelCityWordCloudFrame$Name,HotelCityWordCloudFrame$freq,rot
.per = 0.35,colors = brewer.pal(7,"Dark2"),scale = c(2,0.25))
#
#Creating a word cloud to find out most visted hotel country wise
#
HotelCountryWordCloudFrame <-
SpecifiedFreqMatrix(LocationDataSetYearly$CountryofHotel)
HotelCountryWordCloud <-
wordcloud(HotelCountryWordCloudFrame$Name,HotelCountryWordCloudFrame$fr
eq,colors = brewer.pal(3,"Dark2"),scale=c(2,1))
#
#Creating a word cloud to find out most visted hotel region wise
#
```

```
HotelRegionWordCloudFrame <-
SpecifiedFreqMatrix(LocationDataSetYearly$OperationRegionofHotel)
HotelRegionWordCloud <-
wordcloud(HotelRegionWordCloudFrame$Name,HotelRegionWordCloudFrame$freq
,colors = brewer.pal(7,"Accent"))
#
#Adding amenitites score in metric data
#
MetricDataSetYearly$AmenitiesScore <- AmenitiesDataSetYearly$Countof1
#
#subsetting metric data set to fnd out correlation
#
correlationMatrixYearly <-
subset(MetricDataSetYearly,select=c(3:12,15))
correlationDataYearly<-cor(correlationMatrixYearly)
correlationPlotYearly<-corrgram(correlationDataYearly,upper.panel =
panel.cor)
#
#Performing regression analysis to find out significant variables
#
MetricsLinearModelYearly<-
lm(CustomerMetric_LikelihoodtoRecommend~.,data =
correlationMatrixYearly)
summary(MetricsLinearModelYearly)
#
#Plotting regression model
#
plot(MetricsLinearModelYearly,col="blue",cex=2)
#
#Finding out NPS type countrywise
#
CountryWisePromoterDataYearly<- sqldf("select
count(CustomerType_NPS),CountryofCustomer,CustomerType_NPS from
UserInfoDataSetYearly where CustomerType_NPS = 'Promoter'group by
CountryofCustomer order by count(CustomerType_NPS) DESC")
#
#Creating subset to calculate NPS type state wise after selecting
United States as a state with maximum NPS type
#
StatePromoterSubsetYearly <-
subset(UserInfoDataSetYearly,CountryofCustomer=="UNITEDSTATES",select =
c(8:9))
#
#Grouping with respect to NPS type in United States
#
UnitedStatesPromoterCountYearly <- sqldf("select
count(CustomerType_NPS) PromoterCount, StateofCustomerFullName from
StatePromoterSubsetYearly where CustomerType_NPS= 'Promoter'  group by
StateofCustomerFullName order by count(CustomerType_NPS) DESC")
UnitedStatesPassiveCountYearly <- sqldf("select count(CustomerType_NPS)
PassiveCount, StateofCustomerFullName from StatePromoterSubsetYearly
where CustomerType_NPS= 'Passive'  group by StateofCustomerFullName
order by count(CustomerType_NPS) DESC")
UnitedStatesDetractorCountYearly <- sqldf("select
count(CustomerType_NPS) DetractorCount, StateofCustomerFullName from
StatePromoterSubsetYearly where CustomerType_NPS= 'Detractor'  group by
StateofCustomerFullName order by count(CustomerType_NPS) DESC")
#
```

```
#Merging Data Frames
#
UnitedStatesNPSTypeYearly <-
merge(UnitedStatesPromoterCountYearly,UnitedStatesPassiveCountYearly,by
="StateofCustomerFullName")
UnitedStatesNPSTypeYearly <-
merge(UnitedStatesNPSTypeYearly,UnitedStatesDetractorCountYearly,by="St
ateofCustomerFullName")
UnitedStatesNPSTypeYearly <- na.omit(UnitedStatesNPSTypeYearly)
TotalCountYearly <-
aggregate(UnitedStatesNPSTypeYearly$PromoterCount+UnitedStatesNPSTypeYe
arly$PassiveCount+UnitedStatesNPSTypeYearly$DetractorCount,FUN=mean,dat
a=UnitedStatesNPSTypeYearly,by=list(UnitedStatesNPSTypeYearly$StateofCu
stomerFullName))
UnitedStatesNPSTypeYearly$TotalCount <- TotalCountYearly$x
#
#Plotting UnitedStates Promoter Type
#
PromoterUnitedStatesPlotYearly <-
ggplot(UnitedStatesNPSTypeYearly,aes(x=StateofCustomerFullName,y=Promot
erCount))+geom_bar(stat =
"identity",col="white",fill="yellow")+theme(axis.text.x =
element_text(angle=90,hjust=1))
PromoterUnitedStatesPlotYearly
#
#Plotting UnitedStates Detractor Type
#
DetractorUnitedStatesPlotYearly <-
ggplot(UnitedStatesNPSTypeYearly,aes(x=StateofCustomerFullName,y=Detrac
torCount),fill="red")+geom_bar(stat =
"identity",col="white",fill="blue")+theme(axis.text.x =
element_text(angle=90,hjust=1))
DetractorUnitedStatesPlotYearly
#
#Removing NA
#
MetricDataSetYearly <- na.omit(MetricDataSetYearly)
OverallSatisfactionDataSetYearly<-sqldf("select
avg(CustomerMetric_OverallStatisfaction)
AverageOverallSatisfaction,StateofCustomerFullName from
MetricDataSetYearly group by StateofCustomerFullName")
#
#PLotting Overall Customer Experience vs State
#
OverallCustomerExperienceUnitedStatesPlotYearly <-
ggplot(OverallSatisfactionDataSetYearly,aes(x=StateofCustomerFullName,y
=AverageOverallSatisfaction))+geom_bar(stat="identity",col="grey",fill=
"red")+theme(axis.text.x = element_text(angle = 90,hjust = 1))
OverallCustomerExperienceUnitedStatesPlotYearly
#
#Plotting Likelihood to Recommend vs State
#
LikelihoodtoRecommendDataSetYearly<-sqldf("select
avg(CustomerMetric_LikelihoodtoRecommend)
AverageLikelihoodtoRecommend,StateofCustomerFullName region from
MetricDataSetYearly group by StateofCustomerFullName")
#
#Loading US Map
```

```
#
USMap <- map_data("state")
#
#Finding latitude and longitude to plot a heat map
#
LikelihoodtoRecommendGeoData <-
geocode(LikelihoodtoRecommendDataSetYearly$region)
LikelihoodtoRecommendDataSetYearly$Latitude <-
LikelihoodtoRecommendGeoData$lat
LikelihoodtoRecommendDataSetYearly$Longitude <-
LikelihoodtoRecommendGeoData$lon
LikelihoodtoRecommendDataSetYearly$region <-
tolower(LikelihoodtoRecommendDataSetYearly$region)
#
#plotting heatmap using ggplot
#
LikelihoodtoRecommendHeatMapYearly <-
ggplot(LikelihoodtoRecommendDataSetYearly,aes(map_id=region))+geom_map(
map=USMap,aes(fill=AverageLikelihoodtoRecommend),col="white")+expand_li
mits(x=USMap$long,y=USMap$lat)+coord_map()+ggtitle("State wise
Likelihood of Recommendation")
LikelihoodtoRecommendHeatMapYearly
#
#Plotting world map with hotel locations
#
mapWorldHotelLocationsYearly <- borders("world", colour="gray50",
fill="gray50")
HotelLocationPlotYearly <- ggplot() +   mapWorldHotelLocationsYearly
HotelLocationPlotYearly <- HotelLocationPlotYearly+
geom_point(aes(x=LocationDataSetYearly$LongitudeofHotel,
y=LocationDataSetYearly$LatitudeofHotel) ,color="blue", size=1)
HotelLocationPlotYearly
#
#Code to Compute the NPS of each Country
#
UserInfoDataSetYearly$CountryofHotel <-
LocationDataSetYearly$CountryofHotel
NPS_Dummy <- sqldf('select count(*) as No_of_promoters ,CountryofHotel
from UserInfoDataSetYearly where CustomerType_NPS = "Promoter" Group By
CountryofHotel')
str(NPS_Dummy)
NPS_Dummy1 <- sqldf('select count(*) as No_of_detractors
,CountryofHotel from UserInfoDataSetYearly where CustomerType_NPS =
"Detractor" Group By CountryofHotel')
NPS_Dummy2 <- sqldf('Select
No_of_promoters,No_of_detractors,a.CountryofHotel from NPS_Dummy a ,
NPS_Dummy1 b where a.CountryofHotel = b.CountryofHotel')
sqldf('insert into NPS_Dummy2  select No_of_promoters, 0,
CountryofHotel from NPS_Dummy a where not exists (select * from
NPS_Dummy1 b where a.CountryofHotel = b.CountryofHotel)')
NPS_Dummy3 <- sqldf('select a.No_of_promoters, b.No_of_detractors,
a.CountryofHotel from NPS_Dummy a LEFT JOIN NPS_Dummy1 b ON
a.CountryofHotel = b.CountryofHotel')
NPS_Dummy3
NPS_Dummy4 <- sqldf('select count(*) as No_of_Passive ,CountryofHotel
from UserInfoDataSetYearly where CustomerType_NPS = "Passive" Group By
CountryofHotel')
NPS_Dummy4
```

```
NPS_Dummy5 <- sqldf('select a.No_of_promoters, a.No_of_detractors,
a.CountryofHotel,b.No_of_Passive from NPS_Dummy3 a LEFT JOIN NPS_Dummy4
b ON a.CountryofHotel = b.CountryofHotel')
NPS_Dummy5
NPS_Dummy5$No_of_promoters[is.na(NPS_Dummy5$No_of_promoters)] <- 0
NPS_Dummy5$No_of_detractors[is.na(NPS_Dummy5$No_of_detractors)] <- 0
NPS_Dummy5$No_of_Passive[is.na(NPS_Dummy5$No_of_Passive)] <- 0
sqldf('select * from NPS_Dummy1 a where not exists (select * from
NPS_Dummy b where a.CountryofHotel = b.CountryofHotel)')
NPS_Dummy5$Total_Voters <- NPS_Dummy5$No_of_promoters +
NPS_Dummy5$No_of_detractors + NPS_Dummy5$No_of_Passive
NPS_Dummy5$Promoter_Perc <- NPS_Dummy5$No_of_promoters/
NPS_Dummy5$Total_Voters * 100
NPS_Dummy5$Detractor_Perc <- NPS_Dummy5$No_of_detractors/
NPS_Dummy5$Total_Voters * 100
NPS_Dummy5$NPS <- NPS_Dummy5$Promoter_Perc - NPS_Dummy5$Detractor_Perc
NPS_Dummy5
sqldf('select * from  NPS_Dummy5 order by Total_Voters ')
#
#Scatter Plot of NPS for each country
#
ggplot(NPS_Dummy5, aes(x= CountryofHotel, y = NPS)) + geom_point() +
theme(axis.text.x = element_text(angle = 90,hjust = 1))
#
#Bar Plot of NPS for each Country
#
ggplot(NPS_Dummy5,aes(x = CountryofHotel, y = NPS )) + geom_bar(stat =
"identity",col="white",fill="orange") + theme(axis.text.x =
element_text(angle = 90,hjust = 1))
min(NPS_Dummy5$NPS)
sqldf('select min(NPS) from NPS_Dummy5 ')
#
#
#join to a coarse resolution map
#
NPS_Dummy5$CountryofHotel<- gsub("([a-z])([A-Z])", "\\1 \\2",
NPS_Dummy5$CountryofHotel)
spdf <- joinCountryData2Map(NPS_Dummy5, joinCode="NAME",
nameJoinColumn="CountryofHotel")
mapCountryData(spdf, nameColumnToPlot="NPS", catMethod="fixedWidth")
#
#PLotting a scatter of likelihood vs overall satisaction for different
NPS type
#
ScatterLikelihoodOverallSatData <- subset(YearlyDataSet,select
=c(6,7,34))
GGScatterPLot <-
ggplot(data=ScatterLikelihoodOverallSatData,aes(x=CustomerMetric_Overal
lStatisfaction,y=CustomerMetric_LikelihoodtoRecommend))
GGScatterPLot<-
GGScatterPLot+geom_point(aes(colour=CustomerType_NPS),shape=19,alpha=0.
1,position = position_jitter(w=0.3,h=0.3))
GGScatterPLot
#
#PLotting a scatter of likelihood vs amenities for different NPS type
#
YearlyDataSet$AmenitiesScore <- AmenitiesDataSetYearly$Countof1
```

```
GGScatterPLotAme <-
ggplot(data=YearlyDataSet,aes(x=CustomerMetric_StaffCare,y=CusotmerMetr
ic_QualityofCustomerServivce))
GGScatterPLotAme<-
GGScatterPLotAme+geom_point(aes(colour=CustomerType_NPS),shape=19,alpha
=0.1,position = position_jitter(w=0.3,h=0.3))
GGScatterPLotAme
#
#Performing Data Mining to find out association rules on metrics
#
MiningDataSet <- subset(MetricDataSetYearly, select=c(3:12,15))
MiningDataSet$CustomerMetric_LikelihoodtoRecommend <-
as.factor(MiningDataSet$CustomerMetric_LikelihoodtoRecommend)
MiningDataSet$CustomerMetric_OverallStatisfaction <-
as.factor(MiningDataSet$CustomerMetric_OverallStatisfaction)
MiningDataSet$CustomerMetric_GuestRoomStatisfaction <-
as.factor(MiningDataSet$CustomerMetric_GuestRoomStatisfaction)
MiningDataSet$CustomerMetric_Tranquility <-
as.factor(MiningDataSet$CustomerMetric_Tranquility)
MiningDataSet$CustomerMetric_ConditionofHotel <-
as.factor(MiningDataSet$CustomerMetric_ConditionofHotel)
MiningDataSet$CusotmerMetric_QualityofCustomerServivce <-
as.factor(MiningDataSet$CusotmerMetric_QualityofCustomerServivce)
MiningDataSet$CustomerMetric_StaffCare <-
as.factor(MiningDataSet$CustomerMetric_StaffCare)
MiningDataSet$CustomerMetric_InternetSatisfaction <-
as.factor(MiningDataSet$CustomerMetric_InternetSatisfaction)
MiningDataSet$CustomerMetric_CheckInProcessQuality <-
as.factor(MiningDataSet$CustomerMetric_CheckInProcessQuality)
MiningDataSet$`CustomerMetric_F&BOverallExperience` <-
as.factor(MiningDataSet$`CustomerMetric_F&BOverallExperience`)
MiningDataSet$AmenitiesScore <- as.factor(MiningDataSet$AmenitiesScore)
MiningRuleSet <- apriori(MiningDataSet,parameter =
list(support=0.005,confidence=0.5,maxlen=11))
inspect(MiningRuleSet)
MinigGoodRules <- MiningRuleSet[quality(MiningRuleSet)$lift >10.9]
MinigGoodRulesGraph<-
plot(MinigGoodRules,method="graph",measure="support",shading="lift",int
eractive=TRUE)
inspect(MinigGoodRules)
#
#Analyzing California Data Set
#
CalDataSet <- sqldf("select count(PurposeofCustomerVisit) Total_Count,
PurposeofCustomerVisit from YearlyDataSet where
StateofCustomer='CA'group by PurposeofCustomerVisit")
ggg <- ggplot(CalDataSet, aes(x =
PurposeofCustomerVisit))+geom_bar(aes(y=Total_Count),stat="identity",fi
ll="pink")
ggg <- ggg+ggtitle("Purpose of Visit for California")
ggg
```

Lessons Learnt

- ✓ Understanding the hospitality industry and the factors that drive the industry
- ✓ Detailed understanding of various modeling and regression techniques
- ✓ Understanding of the basic factors that influence a customer feedback.

Challenges faced

- ✓ Understanding the data
- ✓ Technical difficulties in setting and installing packages
- ✓ Project co-ordination and management
- ✓ Timeliness of deliverables and frequent team meeting schedules
- ✓ Limitation of geo code usage
- ✓ Time taken to load data sets and processor speeds

REFLECTION ON THE PROJECT AND WORKING IN A TEAM

The project and the data provided us a great platform to explore the possibilities and power of R programming. It helps us 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 learn about team management, team deliverables , module wise working and also helped us understand the importance of efficiency and importance of resources and deliverables. As a team, we did thorough analysis in the form of SWOT and determined root-cause of the problem which actually led us to correct our path of analysis. Each and every individual of the team made significant contributions at every point during the course of the project. We personally would like to make a point that even if someone misses the team meeting, it is really important at times that the team spirit and dedication should remain intact. Every time , we faced an issue with the moral , there was always someone or the other helping us out to uplift the mood and strive us to work harder for better data analysis of the project. Overall, it was a wonderful 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. If provided an opportunity or maybe if we cross our paths professionally, we definitely look forward to working together as a team once again. We would like to conclude with a team picture and probably the last deliverable for IST 687 this semester.