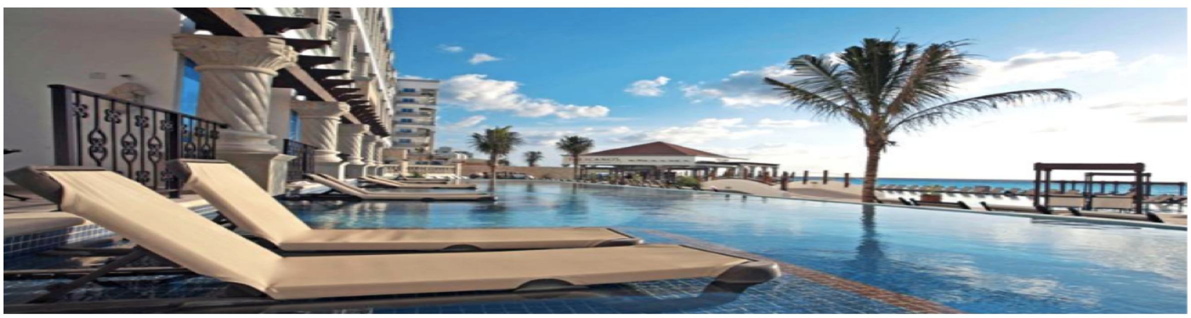
FINAL REPORT:

HYATT - HOTEL DATA ANALYSIS



TEAM 4 – YUNPENG LI

TUSHAR KARIA

TITUS ANDRADE

YAOJHUN ZHENG

ZHIDA ZHAO

AHINAV DEWAN

The Team was given Hyatt Hotel Data set and the project required data analysis on the huge data set and answer data questions. This Report consists of the following:

- **Data ETL**
- **Descriptive Analysis**
- **Regional Analysis: Middle East & Africa**
- **Predictive Analysis**
- **Linear Regression: Predict NPS Type**
- **SVM: Predict NPS Type**
- **Association Rules Mining**

Initially we started off by cleaning the data given to us. We used the December data that was provided to us. Firstly, we got rid of the columns which had NA values or no values in it. Our main focus was to look at the columns which had entries from the feedback form filled by the customers.

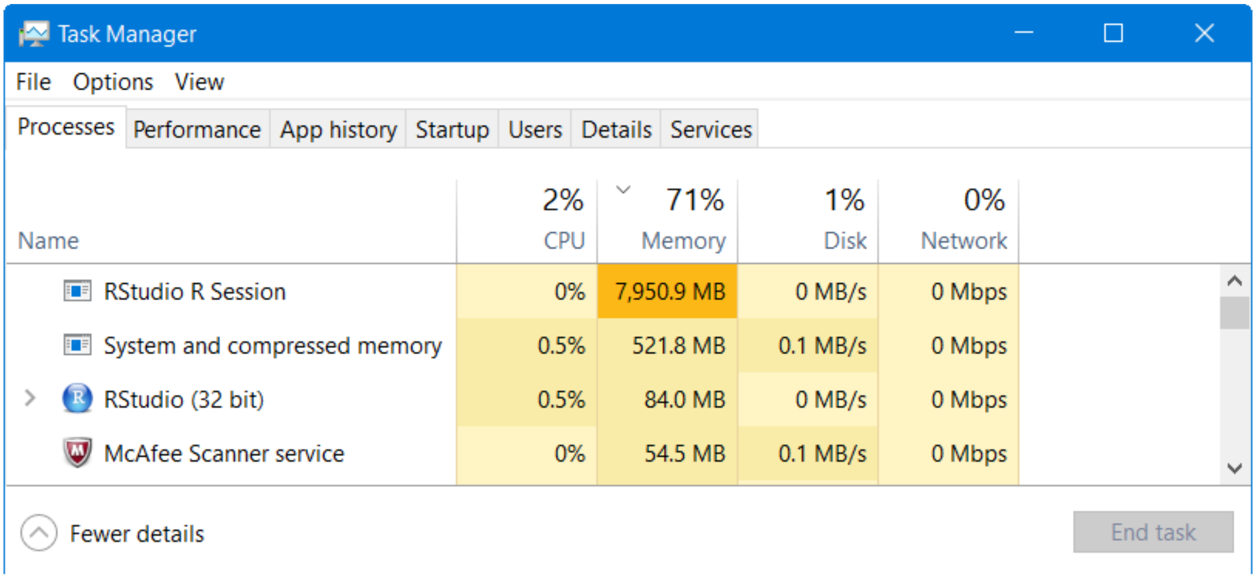
1. ***DATA ETL:***

R Code:

```
#######################################################
# IST 687 Project
# Group 4
#######################################################
# Set environment variables
Sys.setlocale(category="LC_ALL", locale="chs")

# Load libraries
library(ggplot2)
library(reshape2)
library(stringr)

# Read the raw dataset: December 2011-2014, and record the loading time.
inputFilePath <- "C:/TITUS/Assignments/SEM 2/IST 687/Project/out-hyatt_Dec_2011-2014_Clean.csv"
beginTime <- Sys.time()
hotelData <- read.csv(inputFilePath, header=TRUE, sep=",")
hotelData <- data.frame(hotelData)
endTime <- Sys.time()
loadTime <- endTime-beginTime
```

Dataset: out-hyatt_Dec_2011-2014.csv

- Size: 2,106MB
- 5,932,877 observations, 112 variables
- Load time: 4.9 minutes



Dataset: out-hyatt_Dec_2011-2014_Clean.csv

- 120,805 observations, 112 variables

	A	B	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU
1	ROW_ID	ROOM_TYPE_CO	Likelihood_Recommend_H	Overall_Sat_H	Guest_Room_H	Tranquility_H	Condition_Hotel_H	Customer_SVC_H	Staff_Cared_H	Internet_Sat_H	Check_In_H	F.B_FREQ_	F.B_Overa	average_c	Brand_PL
120782	5932565	KING	2	7	10	10	10	10	9	NA	10	4	8		Hyatt Rege
120783	5932568	ADKS	9	10	7	7	9	10	10	NA	10	2	10		Hyatt Rege
120784	5932583	ADKT	9	9	10	10	10	10	10	NA	10	NA	NA		Hyatt Rege
120785	5932601	KING	10	9	8	9	9	10	9	NA	9	3	7.666667		Hyatt Rege
120786	5932638	ST01	10	10	10	10	10	10	10	10	10	2	10		Hyatt Rege
120787	5932642	VW1K	8	9	9	10	9	9	9	8	9	2	8		Hyatt Rege
120788	5932683	VW2D	9	9	9	9	9	9	5	NA	9	2	9		Hyatt Rege
120789	5932688	PANO	7	8	8	7	8	8	9	10	8	NA	NA		Hyatt Rege
120790	5932692	PANO	9	9	10	9	10	10	8	9	10	1	9		Hyatt Rege
120791	5932695	VW2D	9	8	9	9	9	10	9	NA	9	3	9		Hyatt Rege
120792	5932707	HOSP	8	8	9	9	8	9	9	4	9	NA	NA		Hyatt Rege
120793	5932718	KING	1	1	2	5	4	8	8	5	8	2	8		Hyatt Rege
120794	5932722	PANO	10	10	10	10	10	10	10	10	10	1	10		Hyatt Rege
120795	5932726	PANO	10	10	9	10	10	10	10	10	10	3	9.666667		Hyatt Rege
120796	5932733	VW2D	10	10	10	10	10	10	10	NA	10	1	10		Hyatt Rege
120797	5932734	KING	10	9	9	5	9	10	10	10	10	1	10		Hyatt Rege
120798	5932743	KING	5	7	7	7	7	9	10	1	9	2	8		Hyatt Rege
120799	5932766	VW2D	10	10	10	10	10	10	10	10	10	1	8		Hyatt Rege

out-hyatt_Dec_2011-2014_Clean | Comment

	A	B	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA
1	ROW_ID	ROOM_TYPE_CO	All.Suites_	Bell.Staff_	Boutique_	Business.(	Casino_PL	Conferen(	Conventic	Dry.Clean	Elevators_	Fitness.Ce	Fitness.Tr	Golf_PL	Indoor.Co	Laundry_F	Limo.Serv	Mini.Bar_	Pool.Indo	Pool.Outd	Regency.(	Resort_PL	Restauran	Self.Parki	Shuttle.Se	Ski_PL	Spa_PI
120782	5932565	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120783	5932568	ADKS	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120784	5932583	ADKT	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120785	5932601	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120786	5932638	ST01	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120787	5932642	VW1K	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120788	5932683	VW2D	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120789	5932688	PANO	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120790	5932692	PANO	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120791	5932695	VW2D	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120792	5932707	HOSP	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120793	5932718	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120794	5932722	PANO	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120795	5932726	PANO	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120796	5932733	VW2D	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120797	5932734	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120798	5932743	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120799	5932766	VW2D	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120800	5932784	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120801	5932788	DDBL	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120802	5932795	VW1K	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120803	5932797	VW2D	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N
120804	5932798	KING	N	Y	N	Y	N	N	Y	Y	Y	Y	Y	N	Y	N	N	N	N	Y	N	N	Y	Y	N	N	N

out-hyatt_Dec_2011-2014_Clean | Comment

Summary of some data:

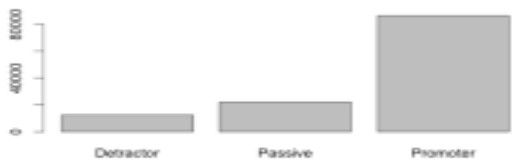
summary of some data

summary(Hnew$NPS_Type)

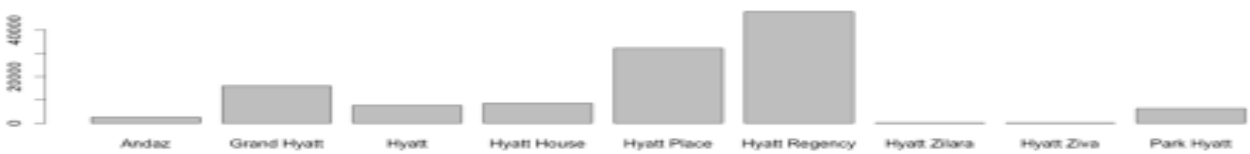
Detractor	Passive	Promoter
12379	22050	86376

summary(Hnew$Location_PL)

(blank)	Airport	Resort	Suburban	Urban
116	15671	11097	41451	52470



Andaz	GrandHyatt	Hyatt	Hyatt House	Hyatt Place	Hyatt Regency	Hyatt Zilara	Hyatt Ziva	Park Hyatt
2458	15867	7585	8464	32137	47799	112	99	6284



Most likelihood hotel brands

tapply(Hnew$Likelihood_Recommend_H,Hnew$Brand_PL,mean)

Andaz	Grand Hyatt	Hyatt	**Hyatt House**	**Hyatt Place**	Hyatt Regency	**Hyatt Zilara**	Hyatt Ziva	**Park Hyatt**
8.886086	8.688536	8.606196	**8.964319**	**8.923297**	8.610724	**9.258929**	7.404040	**9.014163**

Data questions asked:

- Regional Analysis to check which region is the best performing region based on NPS type?
- What are the things, the region is doing right in order to get high NPS Score?
- Do amenities play a significant role in increasing NPS Score?
- Interesting patterns or association among columns affecting NPS Score?

2. ***DESCRIPTIVE ANALYSIS:***

Brand Distribution:

Regional Analysis to check which region is the best performing region based on NPS type?

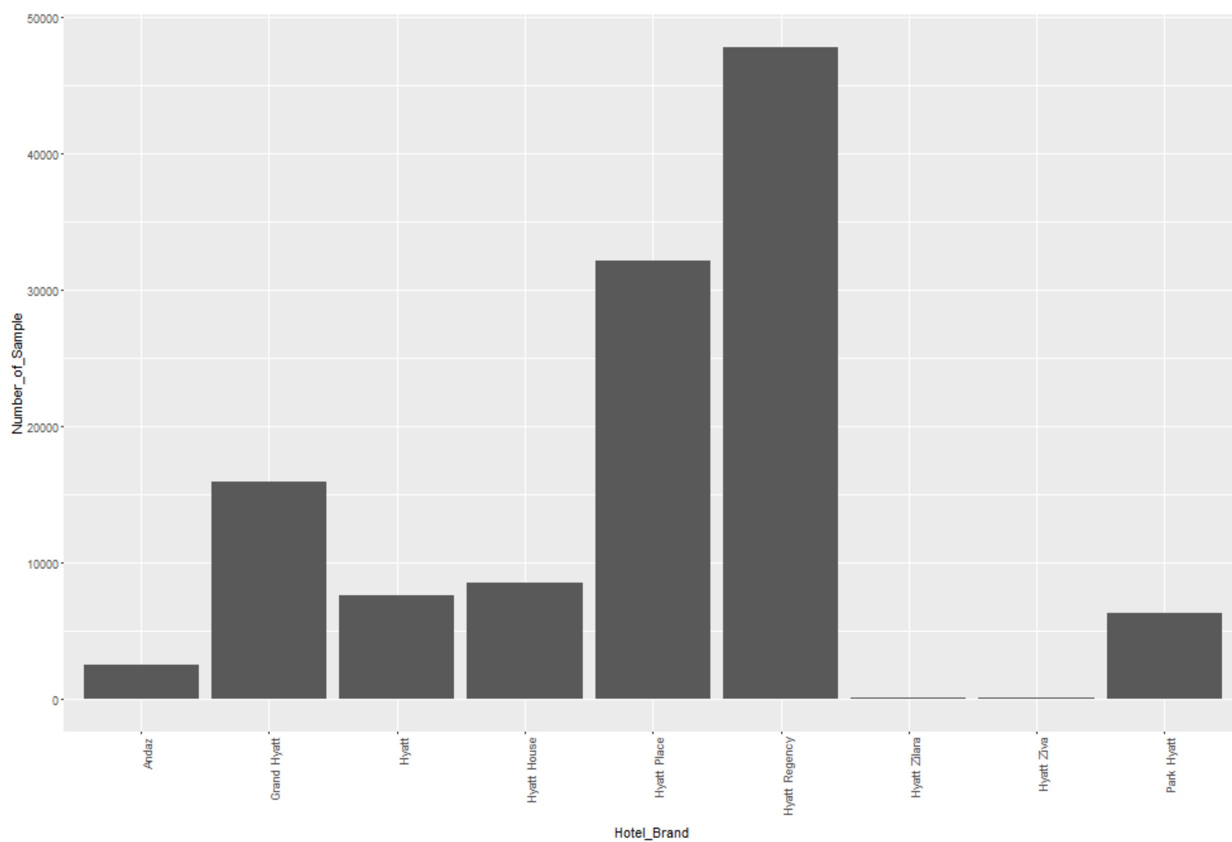
We used Descriptive analysis to find out the best performing region. Since we wanted to analyze the regions and find correlations between various columns, we thought descriptive analysis was the best method to go forward. We got the data samples of each region and then plotted the bar chart for NPS score of each region. Also we found out the overall satisfaction for each region. The most interesting fact we found was that there was high correlation between overall satisfaction and likelihood to recommend.

R code:

```
# Retrieve customer feedback and hotel amenity data
columns <- c("Brand_PL", "Region_PL", "G.Region_PL", "COUNTRY_CODE_R", "STATE_R",
        "Likelihood_Recommend_H", "Overall_Sat_H", "Guest_Room_H", "Tranquility_H",
        "Condition_Hotel_H", "Customer_SVC_H", "Staff_Cared_H", "Check_In_H", "NPS_Type",
        "All.Suites_PL", "Bell.Staff_PL", "Boutique_PL", "Business.Center_PL", "Casino_PL",
        "Conference_PL", "Convention_PL", "Dry.Cleaning_PL", "Elevators_PL", "Fitness.Center_PL",
        "Fitness.Trainer_PL", "Golf_PL", "Indoor.Corridors_PL", "Laundry_PL", "Limo.Service_PL",
        "Mini.Bar_PL", "Pool.Indoor_PL", "Pool.Outdoor_PL", "Regency.Grand.Club_PL", "Resort_PL",
        "Restaurant_PL", "Self.Parking_PL", "Shuttle.Service_PL", "Ski_PL", "Spa_PL")
hotelData <- hotelData[, columns]
# Remove missing data
hotelData <- na.omit(hotelData)
# Format columns: Region_PL
hotelData$Region_PL <- as.character(hotelData$Region_PL)
hotelData$Region_PL[str_trim(hotelData$Region_PL)==""] <- "Other"
```

```
#-----------------------------------------------------------------------------------------------------------#
# Descriptive Analysis
# 1) Brand distribution;
# 2) NPS distribution;
# 3) Correlation between Likelihood to Recommendation and Overall Satisfaction;
# 4) Overall Satisfaction distribution
#-----------------------------------------------------------------------------------------------------------#
```

```
# Brand distribution
hotelBrand <- data.frame(table(hotelData$Brand_PL))
colnames(hotelBrand) <- c("Hotel_Brand", "Number_of_Sample")
g <- ggplot(data=hotelBrand, aes(x=Hotel_Brand, y=Number_of_Sample))
g <- g + geom_bar(stat="identity")
g <- g + theme(axis.text.x=element_text(angle=90, hjust=1))
g
```

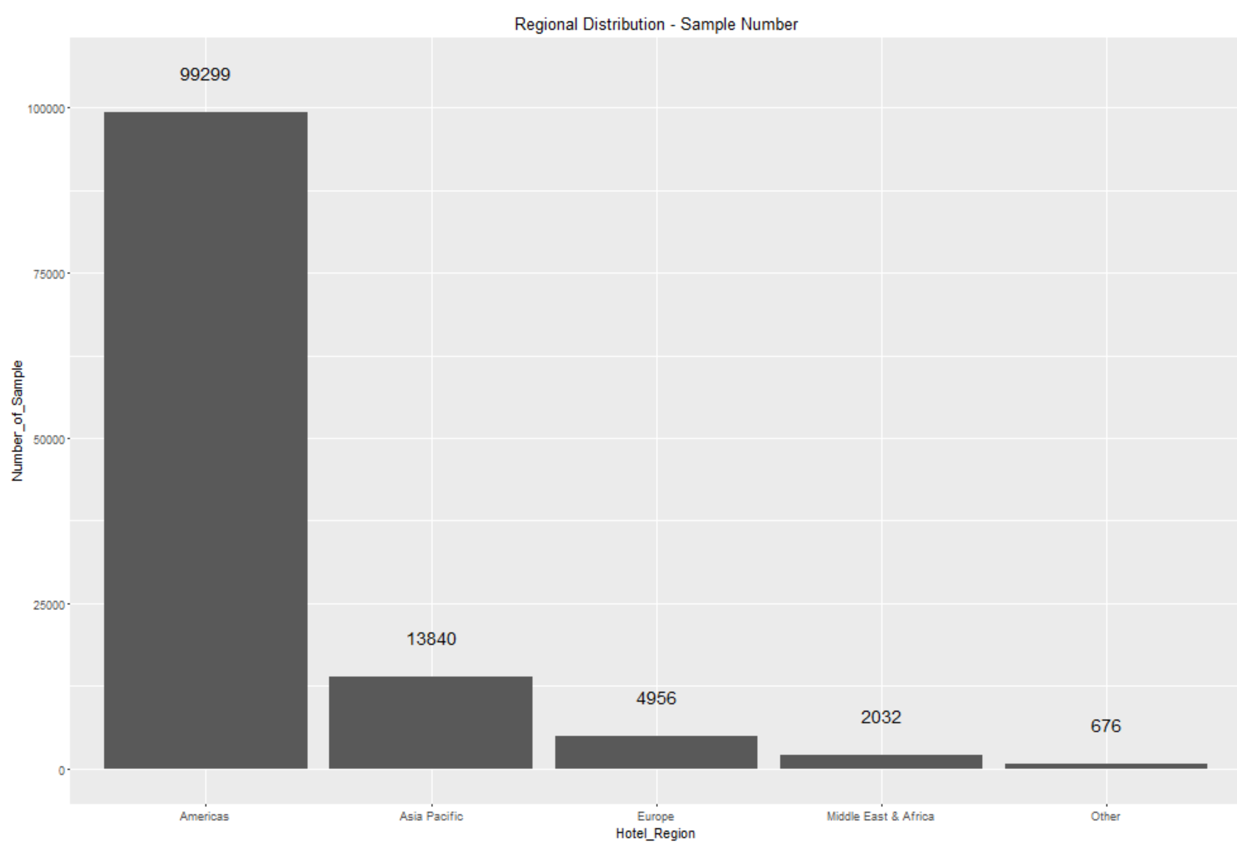


How did we go about?

- **Based on the data, we have first analyzed the region-wise data and retrieved the number of samples for each region.**
- **The overall satisfaction and NPS value for each region was calculated.**
- **The main factor to calculate NPS is Likelihood to Recommend. From the graphs below we can see, if it has high value – people are likely to recommend the hotels more.**

<u>Sample Distribution:</u>

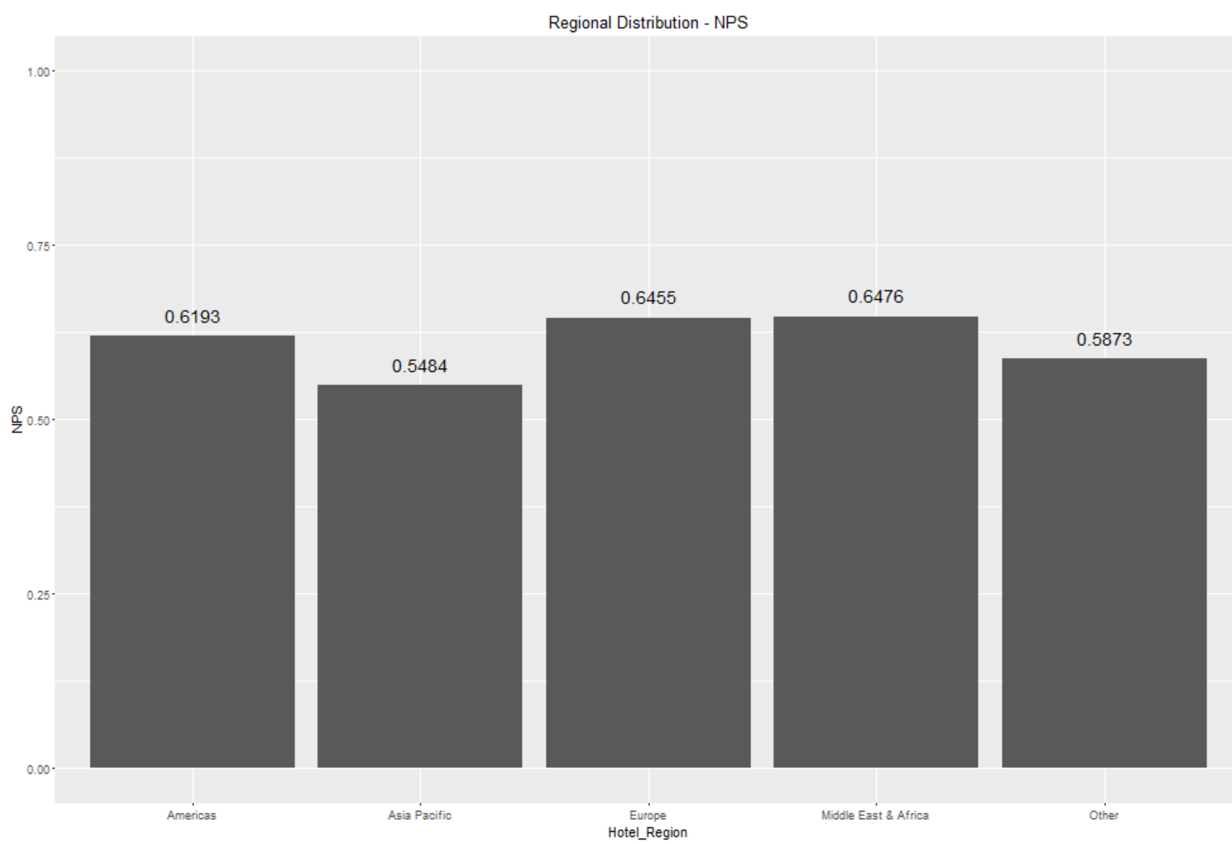
```
# Regional distribution: Sample number
hotelRegion <- data.frame(table(hotelData$Region_PL), stringsAsFactors=FALSE)
colnames(hotelRegion) <- c("Hotel_Region", "Number_of_Sample")
hotelRegion$Hotel_Region <- as.character(hotelRegion$Hotel_Region)
hotelRegion$Number_of_Sample <- as.numeric((hotelRegion$Number_of_Sample))
```



NPS Value: REGION WISE

R Code:

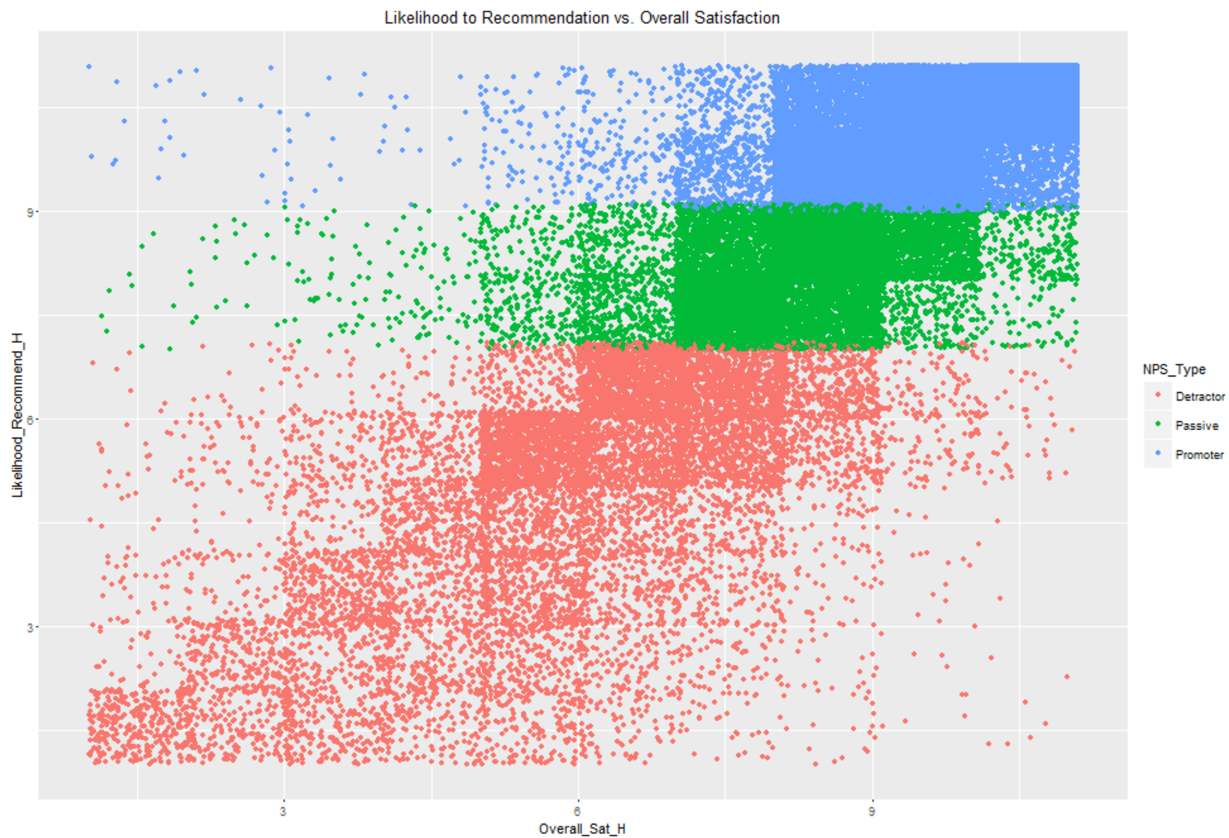
```
# Regional distribution: NPS Value
hotelRegionNPS <- data.frame(tapply(hotelData$NPS_Type, hotelData$Region_PL,
                         function(x){NPS = (sum(x=="Promoter")-sum(x=="Detractor"))/sum(x!="")}),
                    stringsAsFactors=FALSE)
hotelRegionNPS$Hotel_Region <- labels(hotelRegionNPS)[[1]]
colnames(hotelRegionNPS) <- c("NPS", "Hotel_Region")
hotelRegionNPS$Hotel_Region <- as.character(hotelRegionNPS$Hotel_Region)
hotelRegionNPS$NPS <- as.numeric((hotelRegionNPS$NPS))


# Bar chart
g <- ggplot(data=hotelRegionNPS, aes(x=Hotel_Region, y=NPS))
g <- g + geom_bar(stat="identity")
g <- g + geom_text(aes(x=Hotel_Region, y=NPS+0.05*mean(NPS), label=round(NPS,4)), size=5)
g <- g + ggtitle("Regional Distribution - NPS")
g <- g + ylim(0,1)
g
```



Finding the Correlation between Likelihood to Recommend and Overall Satisfaction:

R Code:

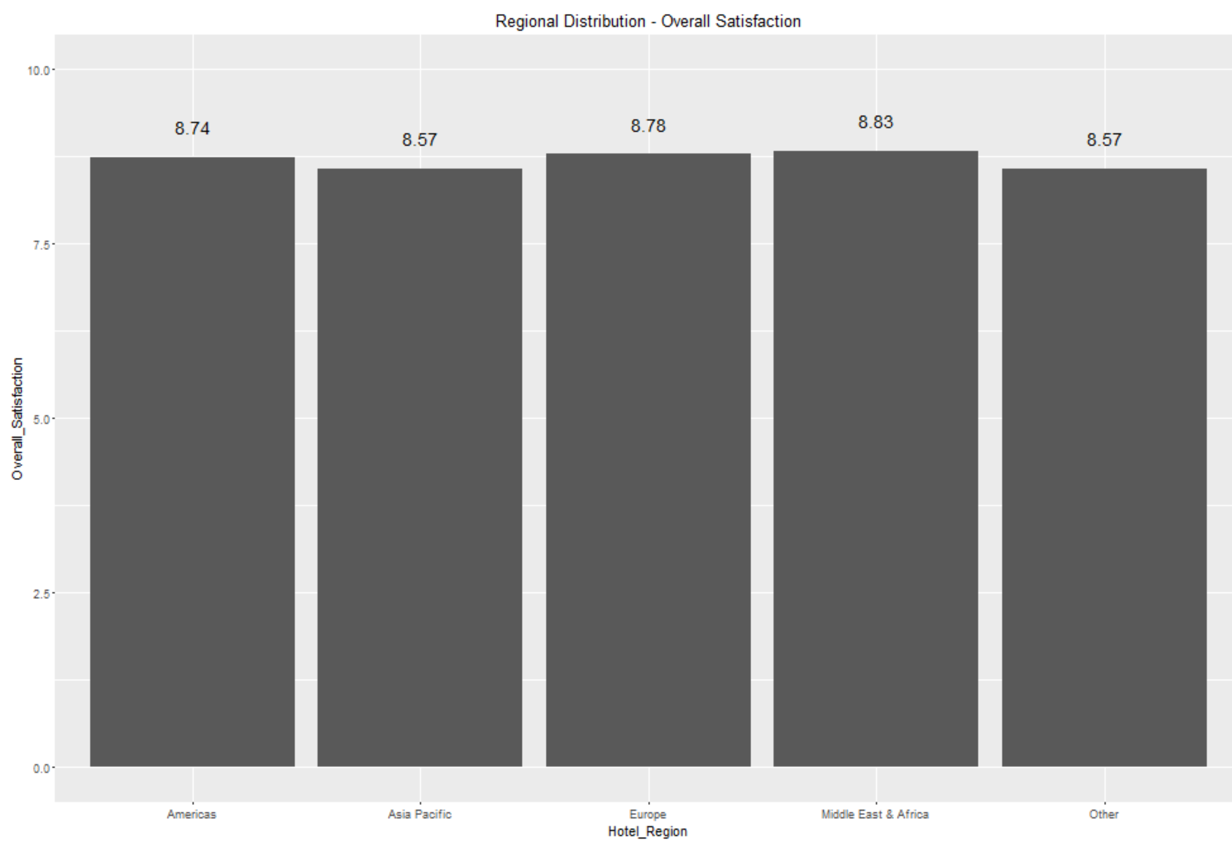
```
# Correlation between Likelihood to Recommendation and Overall Satisfaction

cor(hotelData$Likelihood_Recommend_H, hotelData$Overall_Sat_H)
# Scatter plot: x=overall satisfaction, y=likelihood to recommendation, color=NPS type
df <- hotelData[, c("Likelihood_Recommend_H","Overall_Sat_H","NPS_Type")]
df$Likelihood_Recommend_H <- df$Likelihood_Recommend_H + runif(nrow(df), min=0, max=1.1)
df$Overall_Sat_H <- df$Overall_Sat_H + runif(nrow(df), min=0, max=1.1)
g <- ggplot(data=df, aes(x=Overall_Sat_H))
g <- g + geom_point(aes(y=Likelihood_Recommend_H, color=NPS_Type))
g <- g + ggtitle("Likelihood to Recommendation vs. Overall Satisfaction")
g
```



R Code:

```
# Regional distribution: Overall Satisfaction
hotelRegionOS <- data.frame(tapply(hotelData$Overall_Sat_H, hotelData$Region_PL, mean), stringsAsFactors=FALSE)
hotelRegionOS$Hotel_Region <- labels(hotelRegionOS)[[1]]
colnames(hotelRegionOS) <- c("Overall_Satisfaction", "Hotel_Region")
```

```
hotelRegionOS$Hotel_Region <- as.character(hotelRegionOS$Hotel_Region)
hotelRegionOS$Overall_Satisfaction <- as.numeric((hotelRegionOS$Overall_Satisfaction))
# Bar chart
g <- ggplot(data=hotelRegionOS, aes(x=Hotel_Region, y=Overall_Satisfaction))
g <- g + geom_bar(stat="identity")
g <- g + geom_text(aes(x=Hotel_Region, y=Overall_Satisfaction+0.05*mean(Overall_Satisfaction), label=round(Overall_Satisfaction,2)), size=5)
g <- g + ggtitle("Regional Distribution - Overall Satisfaction")
g <- g + ylim(0,10)
g
```



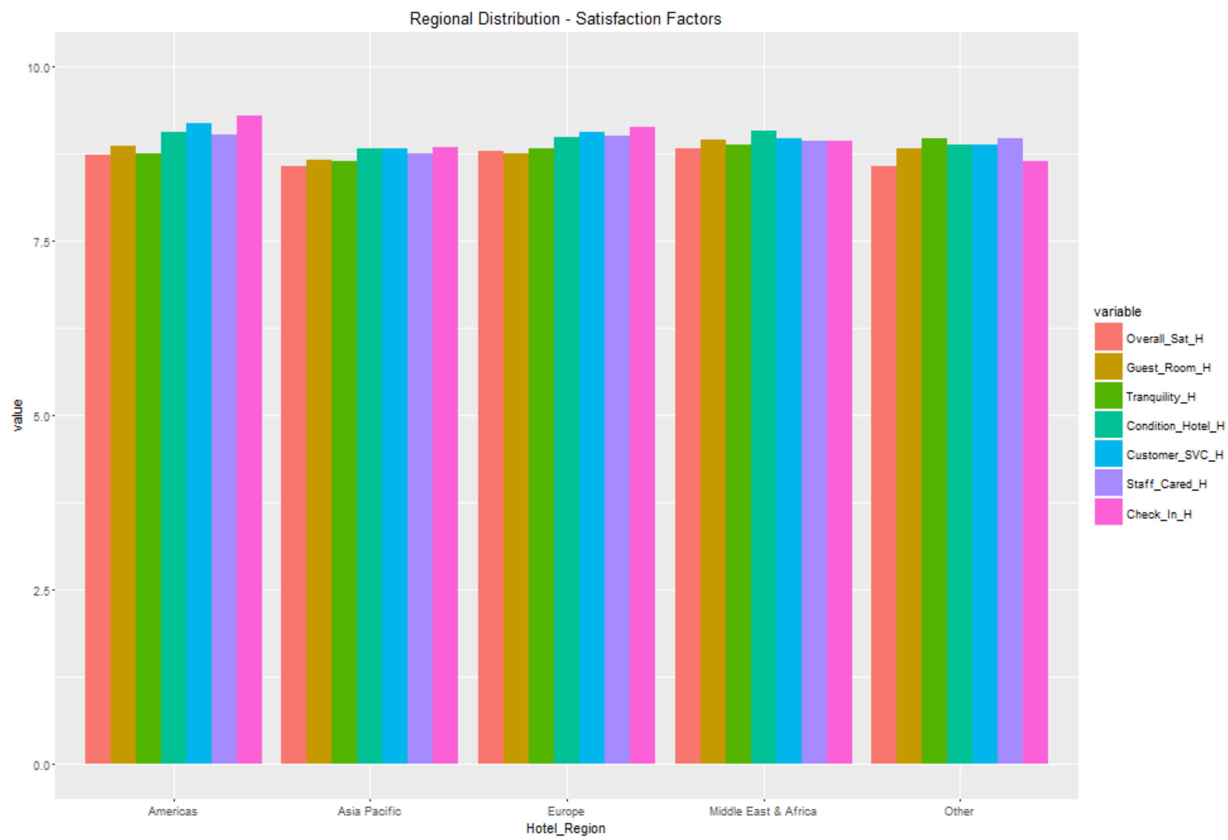
R Code:

```
# Regional distribution: Overall Satisfaction and its factors
df <- hotelData[,c("Region_PL", "Overall_Sat_H", "Guest_Room_H", "Tranquility_H",
        "Condition_Hotel_H", "Customer_SVC_H", "Staff_Cared_H", "Check_In_H")]
hotelRegionSFactors <- aggregate(df[,-1], list(df$Region_PL), mean)
colnames(hotelRegionSFactors)[1] <- "Hotel_Region"
hotelRegionSFactors.m <- melt(hotelRegionSFactors, id.vars="Hotel_Region")

# Bar chart
g <- ggplot(data=hotelRegionSFactors.m, aes(x=Hotel_Region, y=value))
g <- g + geom_bar(aes(fill=variable), position="dodge", stat="identity")
g <- g + ggtitle("Regional Distribution - Satisfaction Factors")
```

```
g <- g + ylim(0,10)
g
```

<u>What are the things, the region is doing right in order to get high NPS Score?</u>

There are various factors which the hotels are doing in the right way. We found out how each region is performing depending on the amenities in those hotels.

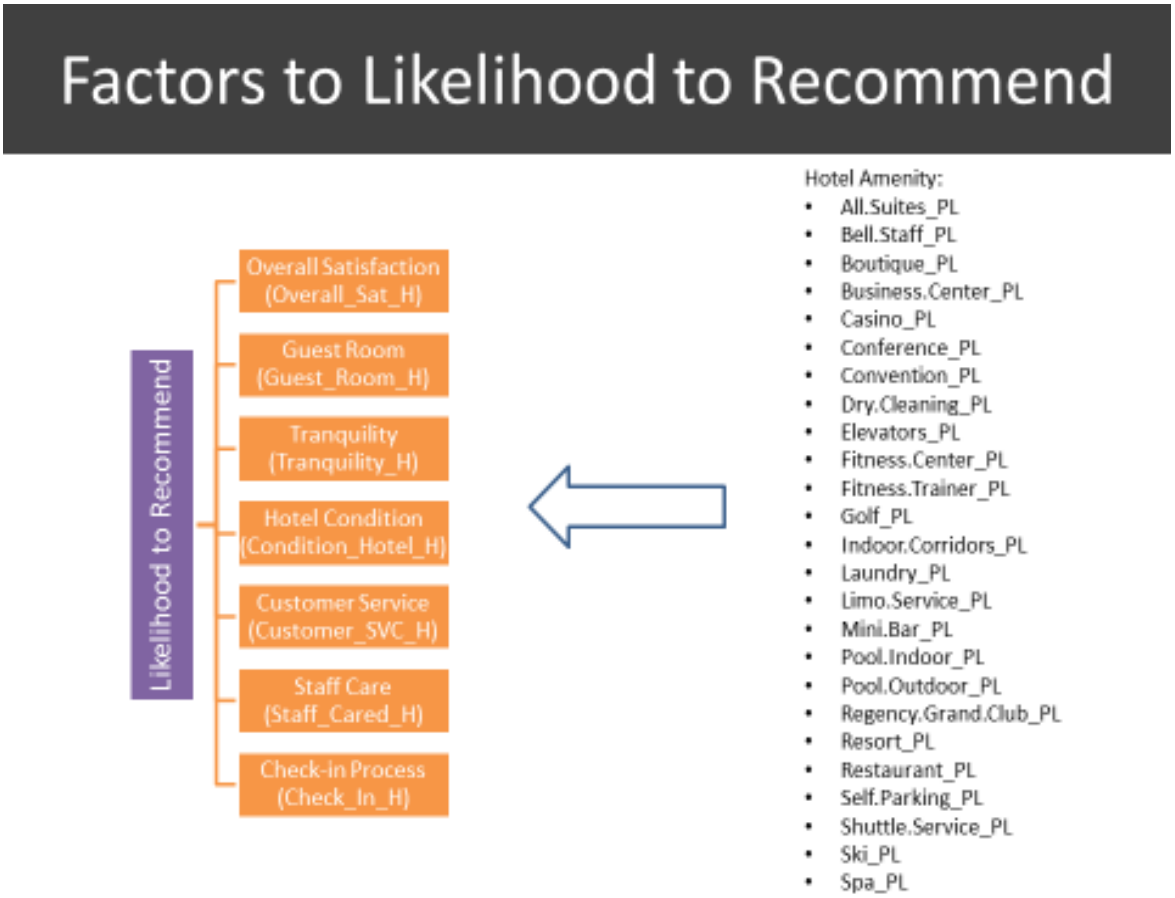


What we found out was: Likelihood to recommend is affected by many factors such as:

- Overall Satisfaction
- Check in Process
- Customer Service
- Tranquility of the room
- Guest Room
- Staff Care
- Hotel Condition

And these factors are dependent on the amenities of the hotel such as: Casino, Conference Room, Golf, Laundry, Fitness Center, Etc.

We noticed that for the Middle East and Africa region, almost all the seven factors are of the same importance.



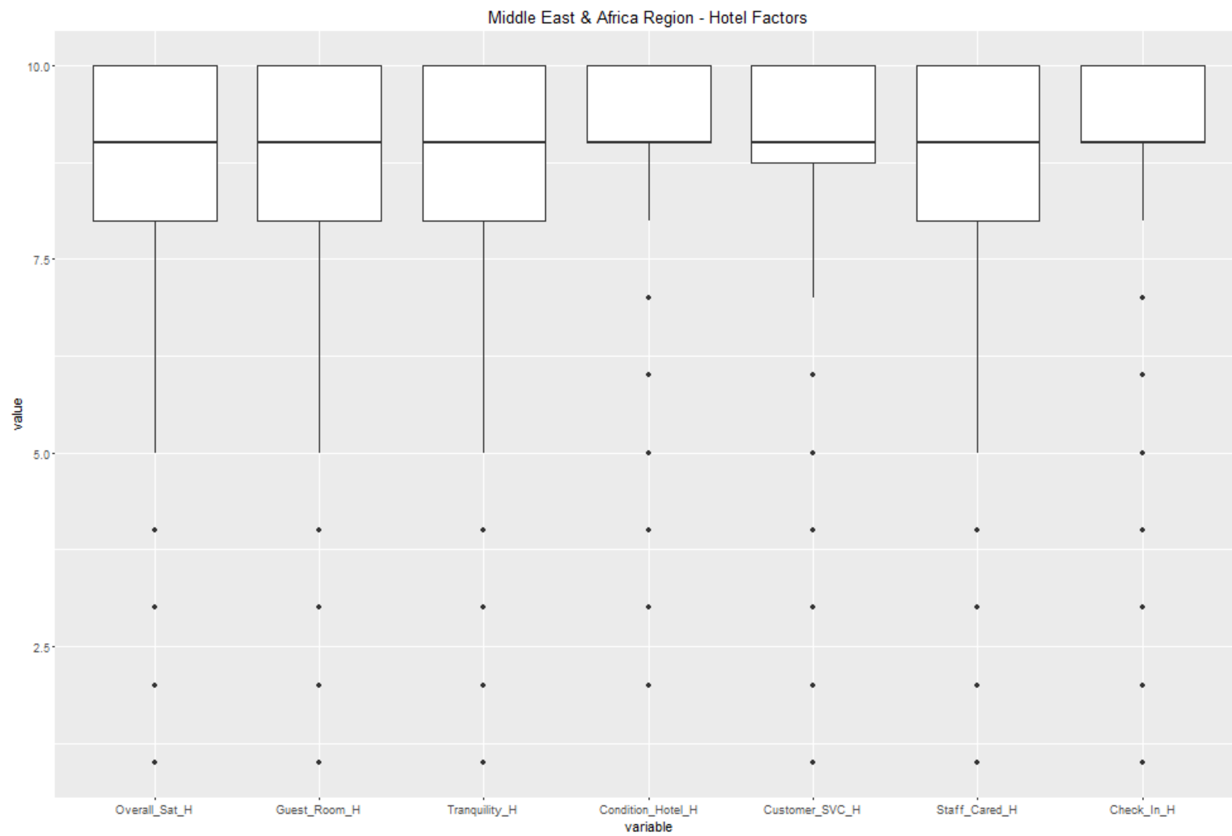
We also noticed that The Middle East & Africa region has the highest NPS value (64.76%) and overall satisfaction rate (8.83/10). Hence we selected the Middle and Africa Region for further Analysis

3. ***Regional Analysis: MIDDLE EAST AND AFRICA***

```
# Sub dataset for Middle East & Africa region
hotelMEA <- hotelData[hotelData$Region_PL=="Middle East & Africa", ]
# Keep columns: "Likelihood_Recommend_H", "Overall_Sat_H", "Guest_Room_H", "Tranquility_H",
# "Condition_Hotel_H", "Customer_SVC_H", "Staff_Cared_H", "Check_In_H", "NPS_Type"
df <- hotelMEA[,c("Likelihood_Recommend_H", "Overall_Sat_H", "Guest_Room_H", "Tranquility_H",
            "Condition_Hotel_H", "Customer_SVC_H", "Staff_Cared_H", "Check_In_H", "NPS_Type")]
hotelMeaSat <- df[,2:8]
df <- df[,2:8]
df <- na.omit(df)
df.m <- melt(df)
```

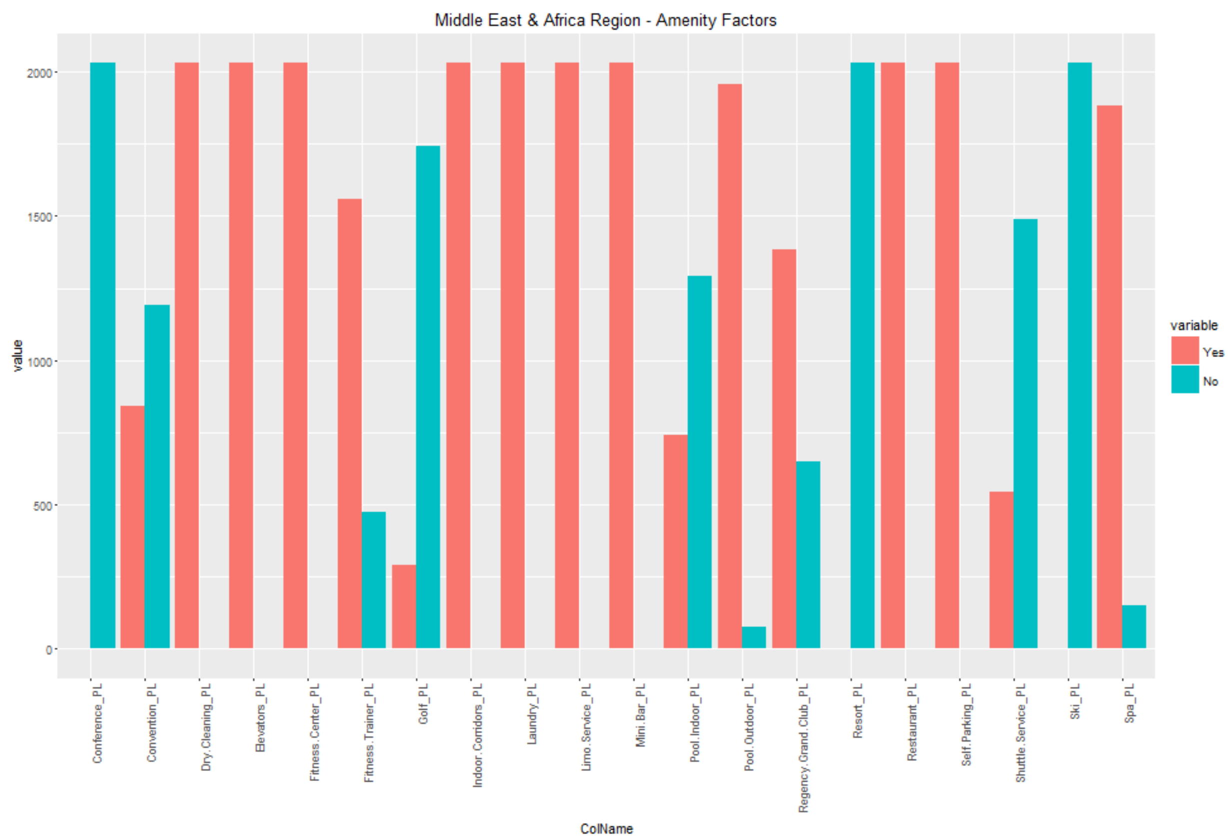
```
# Box-plot for each satisfaction factor
```

```
g <- ggplot(data=df.m)
g <- g + geom_boxplot(aes(x=variable, y=value, group=variable))
g <- g + ggtitle("Middle East & Africa Region - Hotel Factors")
g
```




```
# Bar-chart for amenity factors
amenity <- hotelMEA[,c("Conference_PL", "Convention_PL", "Dry.Cleaning_PL", "Elevators_PL", "Fitness.Center_PL",
             "Fitness.Trainer_PL", "Golf_PL", "Indoor.Corridors_PL", "Laundry_PL", "Limo.Service_PL",
             "Mini.Bar_PL", "Pool.Indoor_PL", "Pool.Outdoor_PL", "Regency.Grand.Club_PL", "Resort_PL",
             "Restaurant_PL", "Self.Parking_PL", "Shuttle.Service_PL", "Ski_PL", "Spa_PL")]
amenity <- na.omit(amenity)
amenity$Spa_PL[amenity$Spa_PL=="Yes"] <- "Y"
df <- data.frame(ColName=character(), Yes=character(), No=character())
for (column in colnames(amenity)) {
  contingency <- data.frame(table(amenity[,column]))
  newrow <- data.frame(ColName=column, Yes=contingency[,2][contingency[,1]=="Y"], No=contingency[,2][contingency[,1]=="N"])
  df <- rbind(df, newrow)
}
df.m <- melt(df, id.vars="ColName")
# Bar chart
```

g <- ggplot(data=df.m, aes(x=ColName, y=value))
g <- g + geom_bar(aes(fill=variable), position="dodge", stat="identity")
g <- g + theme(axis.text.x=element_text(angle=90, hjust=1))
g <- g + ggtitle("Middle East & Africa Region - Amenity Factors")
g



4. **<u>Linear Regression: Predict the NPS Type:</u>**

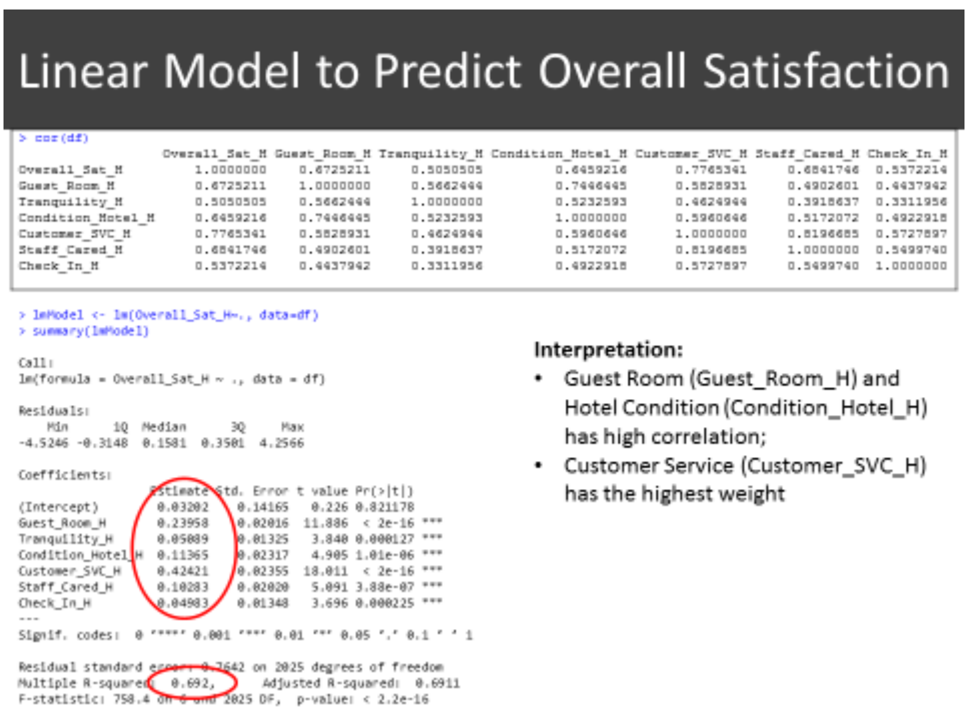
<u>Do amenities play a significant role in increasing NPS Score?</u>

We used the Linear modelling method as well as Support Vector Machine technique to find out whether amenities help the hotels having a high NPS value. We wanted to check how the various amenities affect each other and likelihood to recommend. Hence, linear model was used to predict the correlation between the various factors

```
# Correlation analysis and linear regression model
cor(hotelMeaSat)
lmModel <-
lm(Overall_Sat_H~Guest_Room_H+Tranquility_H+Condition_Hotel_H+Customer_SVC_H+Staff_Cared_H
+Check_In_H, data=hotelMeaSat)
summary(lmModel)
```

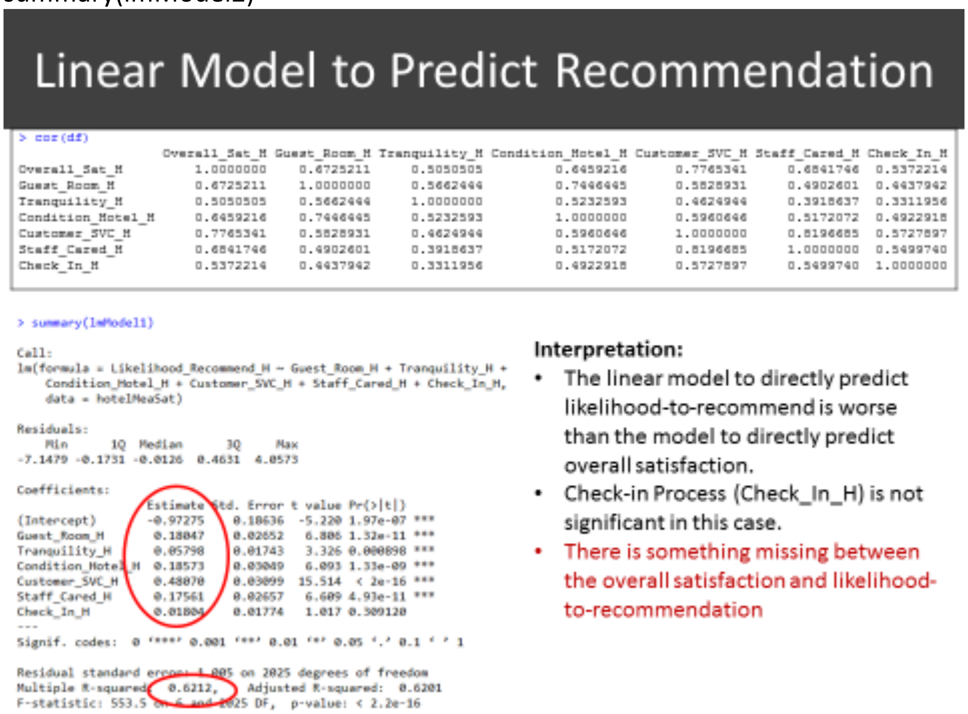
```
Call:
```

```
lm(formula = Overall_Sat_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H
+
    Customer_SVC_H + Staff_Cared_H + Check_In_H, data = hotelMeaSat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5246 -0.3148  0.1581  0.3501  4.2566

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.03202    0.14165   0.226 0.821178
Guest_Room_H       0.23958    0.02016  11.886  < 2e-16 ***
Tranquility_H      0.05089    0.01325   3.840 0.000127 ***
Condition_Hotel_H  0.11365    0.02317   4.905 1.01e-06 ***
Customer_SVC_H     0.42421    0.02355  18.011  < 2e-16 ***
Staff_Cared_H      0.10283    0.02020   5.091 3.88e-07 ***
Check_In_H         0.04983    0.01348   3.696 0.000225 ***
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.7642 on 2025 degrees of freedom
Multiple R-squared:  0.692,	Adjusted R-squared:  0.6911
F-statistic: 758.4 on 6 and 2025 DF,  p-value: < 2.2e-16
```




```
# Create dataset for model training
hotelMeaSat <- hotelMEA[,c("Likelihood_Recommend_H", "Guest_Room_H", "Tranquility_H",
                  "Condition_Hotel_H",  "Customer_SVC_H",  "Staff_Cared_H",  "Check_In_H",  "Amenity",
"NPS_Type")]
df <- hotelMeaSat[,-9]
```

# Create linear regression model
cor(df)
# Linear model without Amenity
lmModel1 <- lm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + Check_In_H, data=hotelMeaSat)
summary(lmModel1)
# Linear model with Amenity
lmModel2 <- lm(Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + Check_In_H + Amenity, data=hotelMeaSat)
summary(lmModel2)



Linear Model to Predict Recommendation

```
> cor(df)
                   Overall_Sat_H Guest_Room_H Tranquility_H Condition_Hotel_H Customer_SVC_H Staff_Cared_H Check_In_H
Overall_Sat_H          1.0000000    0.6725211     0.5050505         0.6459216      0.7765341     0.6841746  0.5372214
Guest_Room_H           0.6725211    1.0000000     0.5662444         0.7446445      0.5828931     0.4902601  0.4437942
Tranquility_H          0.5050505    0.5662444     1.0000000         0.5232593      0.4624944     0.3918637  0.3311956
Condition_Hotel_H      0.6459216    0.7446445     0.5232593         1.0000000      0.5960646     0.5172072  0.4922918
Customer_SVC_H         0.7765341    0.5828931     0.4624944         0.5960646      1.0000000     0.8196685  0.5727897
Staff_Cared_H          0.6841746    0.4902601     0.3918637         0.5172072      0.8196685     1.0000000  0.5499740
Check_In_H             0.5372214    0.4437942     0.3311956         0.4922918      0.5727897     0.5499740  1.0000000
```

```
> summary(lmModel1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.1479 -0.1731 -0.0126  0.4631  4.0573

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.97275    0.18636  -5.220 1.97e-07 ***
Guest_Room_H       0.18047    0.02652   6.806 1.32e-11 ***
Tranquility_H      0.05798    0.01743   3.326 0.000898 ***
Condition_Hotel_H  0.18573    0.03049   6.093 1.33e-09 ***
Customer_SVC_H     0.48070    0.03099  15.514  < 2e-16 ***
Staff_Cared_H      0.17561    0.02657   6.609 4.93e-11 ***
Check_In_H         0.01804    0.01774   1.017 0.309120
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.005 on 2025 degrees of freedom
Multiple R-squared:  0.6212,	Adjusted R-squared:  0.6201
F-statistic: 553.5 on 6 and 2025 DF,  p-value: < 2.2e-16
```

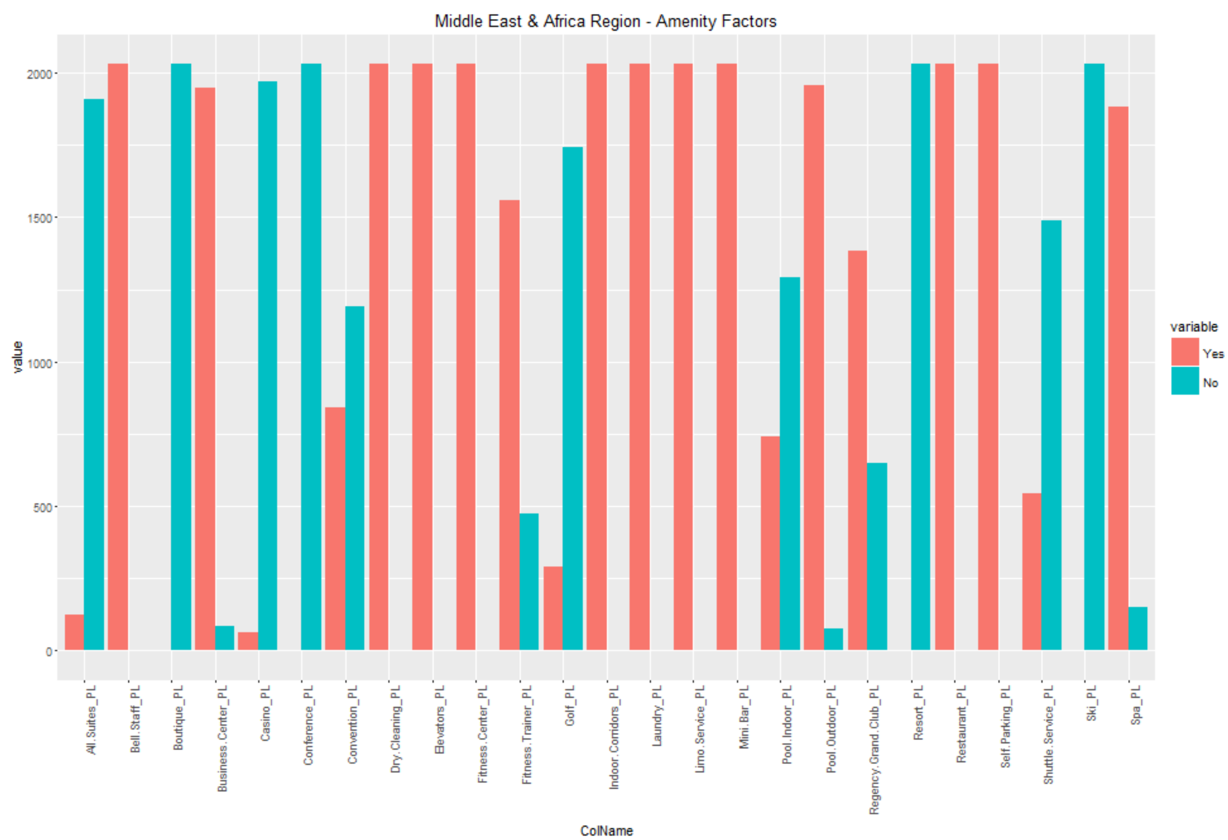
Interpretation:
- The linear model to directly predict likelihood-to-recommend is worse than the model to directly predict overall satisfaction.
- Check-in Process (Check_In_H) is not significant in this case.
- There is something missing between the overall satisfaction and likelihood-to-recommendation

We have also taken the Amenity Factors in the Middle East and Africa Region.

<u>R Code:</u>

```
# Bar-chart for amenity factors
amenity <- hotelMEA[,c("All.Suites_PL", "Bell.Staff_PL", "Boutique_PL", "Business.Center_PL", "Casino_PL",
            "Conference_PL", "Convention_PL", "Dry.Cleaning_PL", "Elevators_PL", "Fitness.Center_PL",
            "Fitness.Trainer_PL", "Golf_PL", "Indoor.Corridors_PL", "Laundry_PL", "Limo.Service_PL",
            "Mini.Bar_PL", "Pool.Indoor_PL", "Pool.Outdoor_PL", "Regency.Grand.Club_PL", "Resort_PL",
            "Restaurant_PL", "Self.Parking_PL", "Shuttle.Service_PL", "Ski_PL", "Spa_PL")]
amenity <- na.omit(amenity)
amenity$Spa_PL[amenity$Spa_PL==1] <- "Y"
df <- data.frame(ColName=character(), Yes=character(), No=character())
for (column in colnames(amenity)) {
 contingency <- data.frame(table(amenity[,column]))
```

```
  newrow <- data.frame(ColName=column, Yes=contingency[,2][contingency[,1]=="Y"], No=contingency[,2][contingency[,1]=="N"])
  df <- rbind(df, newrow)
}
df.m <- melt(df, id.vars="ColName")

# Bar chart
g <- ggplot(data=df.m, aes(x=ColName, y=value))
g <- g + geom_bar(aes(fill=variable), position="dodge", stat="identity")
g <- g + theme(axis.text.x=element_text(angle=90, hjust=1))
g <- g + ggtitle("Middle East & Africa Region - Amenity Factors")
g
```



We see that the amenities play an important role for a hotel. If there are amenities such as guest room, Conference room, fitness center, pool, etc, there are more number of promoters than without amenities in that hotel.

5. **SVM: Prediction of the NPS Type:**

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

```
testData <- hotelMeaSat[randIndex[(cutPoint2_3+1):dim(hotelMeaSat)[1]],]

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

# for dataset only containing "Promoter" and "Passive"
svmModel1 <- ksvm(NPS_Type~Guest_Room_H + Tranquility_H + Condition_Hotel_H +
            Customer_SVC_H + Staff_Cared_H, data=trainData, kernel="rbfdot", kpar="automatic", C=20,
cross=3)

# Test the SVM model
svmPred <- round(predict(svmModel1, testData, type="votes"))
compTable <- data.frame(testData[,"NPS_Type"], svmPred[3,])

# Create a confusion matrix based on the prediction result
table(compTable)

# for dataset only containing "Promoter" and "Passive"
svmModel2 <- ksvm(NPS_Type~Guest_Room_H + Tranquility_H + Condition_Hotel_H +
            Customer_SVC_H + Staff_Cared_H + Amenity, data=trainData, kernel="rbfdot",
kpar="automatic", C=20, cross=3)

# Test the SVM model
svmPred <- round(predict(svmModel2, testData, type="votes"))
compTable <- data.frame(testData[,"NPS_Type"], svmPred[3,])

# Create a confusion matrix based on the prediction result
table(compTable)
```

SVM to Predict NPS Type: With/Without Amenity Factors

NPS_Type ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + **Amenity**

Model 1:

```
> table(compTable)
                          svmPred.3...
testData....NPS_Type..    0   1   2
             Detractor   40  16  10
             Passive     22  33  65
             Promoter     7  29 456
```

Overall Accuracy =
(40+33+456)/(66+120+492)
= 78.02%

Model 2:

```
> table(compTable)
                          svmPred.3...
testData....NPS_Type..    0   1   2
             Detractor   43   8  15
             Passive     22  38  60
             Promoter     5  23 464
```

Overall Accuracy =
(43+38+464)/(66+120+492)
= 80.38%

Interpretation:

- The overall accuracy is around 80%
- The accuracy for "Promoter" is pretty high (92.7% and 94.3%), because the training dataset is skewed to promoter

6. Association Rule Mining:

Interesting patterns or association among columns affecting NPS Score?

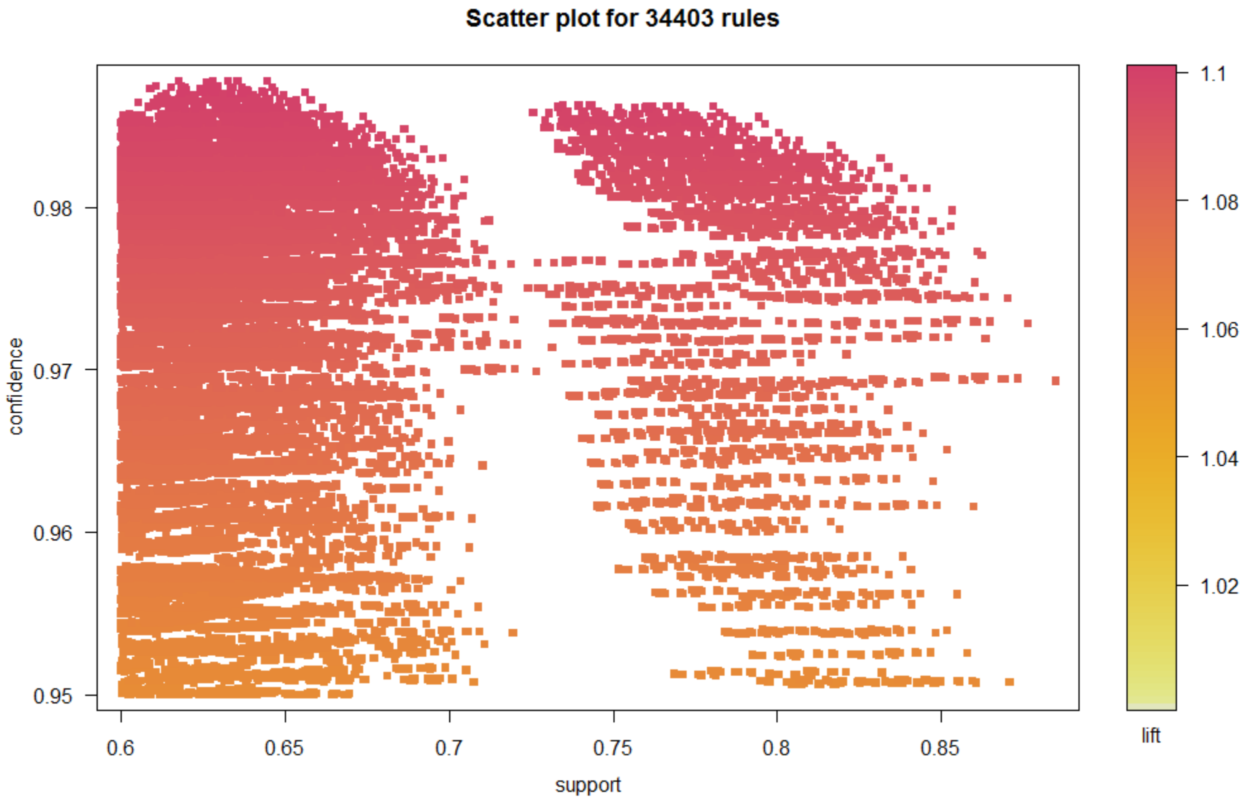
We resorted to association rule mining so that we can discover if the factors affecting the Middle East and Africa are the same across the world. We found out some interesting rules to support our data findings.

```
#Create a new data frame with only the columns which play a significant role in predicting nps type
#these columns were selected from linear and svm modelling
#Also including each individual amenities to get interesting rules

testData <- data.frame(hotelData[,3])
testData$LENGTH_OF_STAY_C <- hotelData$LENGTH_OF_STAY_C
testData$NUMBER_OF_ROOMS_C <- hotelData$NUMBER_OF_ROOMS_C
testData$POV_CODE_C <- hotelData$POV_CODE_C
testData$GROUPS_VS_FIT_R <- hotelData$GROUPS_VS_FIT_R
testData <- testData[,-6]
testData$Age_Range_H <- hotelData$Age_Range_H
testData$Gender_H <- hotelData$Gender_H
```

```
testData$POV_H <- hotelData$POV_H
testData$Clublounge_Used_H <- hotelData$Clublounge_Used_H
testData$Spa_Used_H <- hotelData$Spa_Used_H
testData$Likelihood_Recommend_H <- hotelData$Likelihood_Recommend_H
testData$Overall_Sat_H <- hotelData$Overall_Sat_H
testData$Guest_Room_H <- hotelData$Guest_Room_H
testData$Tranquility_H <- hotelData$Tranquility_H
testData$Condition_Hotel_H <- hotelData$Condition_Hotel_H
testData$Customer_SVC_H <- hotelData$Customer_SVC_H
testData$Staff_Cared_H <- hotelData$Staff_Cared_H
testData$Check_In_H <- hotelData$Check_In_H
testData$F.B_Overall_Experience_H <- hotelData$F.B_Overall_Experience_H
testData$Brand_PL <- hotelData$Brand_PL
testData$Region_PL <- hotelData$Region_PL
testData$Pool.Indoor_PL <- hotelData$Pool.Indoor_PL
testData$Pool.Outdoor_PL <- hotelData$Pool.Outdoor_PL
testData$Ski_PL <- hotelData$Ski_PL
testData$Spa_PL <- hotelData$Spa_PL
testData$Spa.online.booking_PL <- hotelData$Spa.online.booking_PL
testData$Golf_PL <- hotelData$Golf_PL
testData$Casino_PL <- hotelData$Casino_PL
testData$Laundry_PL <- hotelData$Laundry_PL
testData$Boutique_PL <- hotelData$Boutique_PL
testData$Mini.Bar_PL <- hotelData$Mini.Bar_PL
testData$Elevators_PL <- hotelData$Elevators_PL
testData$Bell.Staff_PL <- hotelData$Bell.Staff_PL
testData$Conference_PL <- hotelData$Conference_PL
testData$Convention_PL <- hotelData$Convention_PL
testData$Restaurant_PL <- hotelData$Restaurant_PL
testData$Dry.Cleaning_PL <- hotelData$Dry.Cleaning_PL
testData$Limo.Service_PL <- hotelData$Limo.Service_PL
testData$Self.Parking_PL <- hotelData$Self.Parking_PL
testData$Valet.Parking_PL <- hotelData$Valet.Parking_PL
testData$Fitness.Center_PL <- hotelData$Fitness.Center_PL
testData$Business.Center_PL <- hotelData$Business.Center_PL
testData$Shuttle.Service_PL <- hotelData$Shuttle.Service_PL
testData$Spa.F.B.offering_PL <- hotelData$Spa.F.B.offering_PL

#getting rid of the first three columns as well as POV_CODE_C
testData <- testData[,-1:-3]
testData <- testData[,-1]

#Checking the structure of the data again
str(testData)


#Changing likelihood to recommend into yes or no type i.e. >= 7 is Yes or else No
testData$Likelihood_Recommend_H[testData$Likelihood_Recommend_H >= 7] <- "Yes"
```

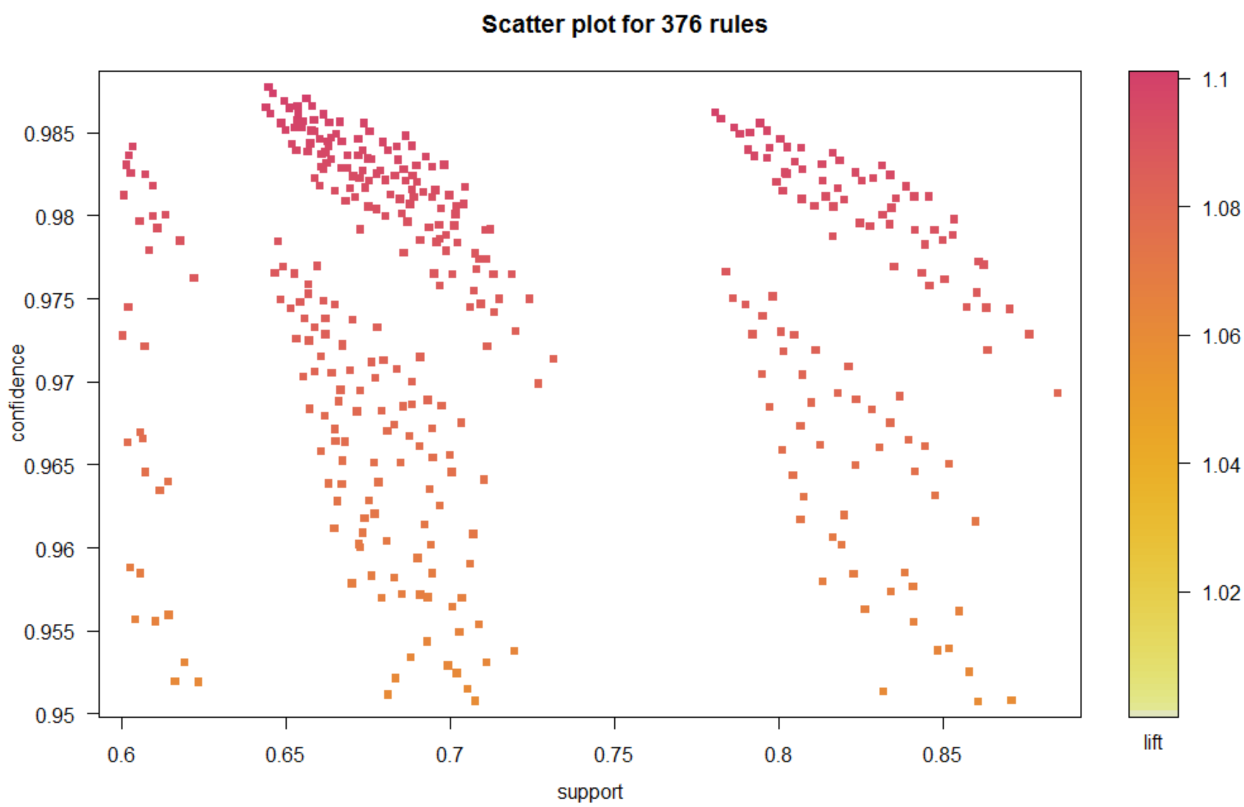
```
testData$Likelihood_Recommend_H[testData$Likelihood_Recommend_H < 7] <- "No"
testData$Likelihood_Recommend_H <- as.factor(testData$Likelihood_Recommend_H)

#Converting all the columns into categorical type by assigning Satisfied or not satisified values
#Also converting them into factors
#greater or equal to 7 is satisfied and less than 7 is not satisfied
testData$Overall_Sat_H[testData$Overall_Sat_H >= 7] <- "Satisfied"
testData$Overall_Sat_H[testData$Overall_Sat_H < 7] <- "Not Satisfied"
testData$Overall_Sat_H <- as.factor(testData$Overall_Sat_H)

testData$Guest_Room_H[testData$Guest_Room_H >= 7] <- "Satisfied"
testData$Guest_Room_H[testData$Guest_Room_H < 7] <- "Not Satisfied"
testData$Guest_Room_H <- as.factor(testData$Guest_Room_H)

testData$Tranquility_H[testData$Tranquility_H >= 7] <- "Satisfied"
testData$Tranquility_H[testData$Tranquility_H < 7] <- "Not Satisfied"
testData$Tranquility_H <- as.factor(testData$Tranquility_H)

testData$Condition_Hotel_H [testData$Condition_Hotel_H  >= 7] <- "Satisfied"
testData$Condition_Hotel_H [testData$Condition_Hotel_H  < 7] <- "Not Satisfied"
testData$Condition_Hotel_H  <- as.factor(testData$Condition_Hotel_H )

testData$Customer_SVC_H[testData$Customer_SVC_H >= 7] <- "Satisfied"
testData$Customer_SVC_H[testData$Customer_SVC_H < 7] <- "Not Satisfied"
testData$Customer_SVC_H <- as.factor(testData$Customer_SVC_H)

testData$Staff_Cared_H[testData$Staff_Cared_H >= 7] <- "Satisfied"
testData$Staff_Cared_H[testData$Staff_Cared_H < 7] <- "Not Satisfied"
testData$Staff_Cared_H <- as.factor(testData$Staff_Cared_H)

testData$Check_In_H[testData$Check_In_H >= 7] <- "Satisfied"
testData$Check_In_H[testData$Check_In_H < 7] <- "Not Satisfied"
testData$Check_In_H <- as.factor(testData$Check_In_H)

testData$F.B_Overall_Experience_H[testData$F.B_Overall_Experience_H >= 7] <- "Satisfied"
testData$F.B_Overall_Experience_H[testData$F.B_Overall_Experience_H < 7] <- "Not Satisfied"
testData$F.B_Overall_Experience_H <- as.factor(testData$F.B_Overall_Experience_H)

#installing and loading arules and arulesViz packages
library(arules)
library(arulesViz)

#using apriori command and setting parameters
rules <- apriori(testData,parameter=list(support=0.60,confidence=0.95))
summary(rules)

#rules.1  is a subset of rules to get rhs as likelihood to recommend
rules.1 <-  subset( rules, subset = rhs %pin% "Likelihood_Recommend_H=" )
```

```
summary(rules.1)
inspect(rules.1)

#creating a plot for rules 1
plot(rules.1)
```




```
#Changing the lift values constantly to get interesting rules
goodrules <- rules.1[quality(rules.1)$lift > 1.059 & quality(rules.1)$lift < 1.06]

#summarizing good rules and inspecting rules
summary(goodrules)
inspect(goodrules)



#Now trying to create rules without the amenities columns
#creating a new data frame without amenities column
testData.1 <- testData[,-5:-6]
testData.1 <- testData.1[,-16:-38]

#checking the structure of testData.1
str(testData.1)

#creating rules for it using apriori
```

```
rules.2 <- apriori(testData.1,parameter=list(support=0.60,confidence=0.95))
summary(rules.2)

#creating a subset of it to have rhs as likelihood to recommend
rules.3 <-  subset( rules.2, subset = rhs %pin% "Likelihood_Recommend_H=" )
#summarizing and inspecting the rules
summary(rules.3)
inspect(rules.3)

#using the plot command of arulesViz package to create a plot
#Plot helps comapring support, confidence and lift at the same time
plot(rules.3)
```




```
#Playing with values of lift to get interesting rules
goodrules <- rules.3[quality(rules.3)$lift > 1.07 & quality(rules.1)$lift < 1.08]

#summarizing and inspecting the rules
summary(goodrules)
inspect(goodrules)
```

Association Rules Mining

LHS	RHS	Support	Confidence	Lift
Customer_SVC_H=Satisfied, F.B_Overall_Experience_H=Satisfied	Likelihood_Recommend_H=Yes	57.27%	98.75%	1.10
Check_In_H=Satisfied, Region_PL=Americas	Likelihood_Recommend_H=Yes	71.99%	91.51%	1.02
POV_H=Leisure	Likelihood_Recommend_H=Yes	50.32%	90.05%	1.01
GROUPS_VS_FIT_R=FIT, Customer_SVC_H=Satisfied	Likelihood_Recommend_H=Yes	73.27%	93.67%	1.04
GROUPS_VS_FIT_R=FIT, Overall_Sat_H=Satisfied, Guest_Room_H=Satisfied, Tranquility_H=Satisfied, Customer_SVC_H=Satisfied, Staff_Cared_H=Satisfied, Check_In_H=Satisfied, Region_PL=Americas	Likelihood_Recommend_H=Yes	52.15%	98.73%	1.10

SUGGESTIONS TO HYATT:

- ***Guest room has high correlation with hotel condition***
- ***Amenities does play a significant factor when it comes to likelihood to recommend***
- ***Check in process experience is important in the Americas region***
- ***Customers travelling as FIT's compared to groups get influenced majorly by Customer service***
- ***Customers travelling for Business purposes get highly influenced by the tranquility of the room***
- ***Customers travelling for leisure purposes are more likely to recommend than those travelling for business***
- ***A good experience with the food and beverage services in the hotels also has a significant impact to the likelihood to recommend for customers***

REFLECTION ON THE PROJECT AND WORKING IN A TEAM:

It was a very good experience working on this project. The fact a that the data was provided to us by the Professor allowed us to focus and give more time to the project than to finding suitable data. The few things that we would definitely take from this project would be the techniques introduced by the professor to keep a track of the progress. The methods used to keep a good note of what's in progress

already done and what needs to be done further helped a lot when it came to organization and time management.

Working in a team was really pleasant experience, as it brought a lot of different perspective and skills to the table. It eventually helped us achieve more than we expected and aimed. To conclude it was a really insightful experience and sharpened not only our R-programming skills but also our communication skills.