**IST – 687 = Team# 3**

Mateo Arbelaez

Maggie Faiella

Teodor Fatu

Sonny Lizarraga

Daniel Tully

**Final Project Report**

Table of Contents

1. Project Background and Description

Business Question

2. Hotel data – Business questions

Data Acquisition, Cleaning, Transformation

3. Process description

4. Data Selected and why

5. What was your quality assessment

6. Which fields/Variables did you select and why

7. Cleaning process

8. Findings

Descriptive Statistics

9. Provide demographic statistics

10. Any early observations, nuggets of interest, interpretation, interesting findings

11. Graphs, Charts, tables, visuals, text

Use of Modeling Techniques

12. Linear modeling

13. Regression

14. KSVM

Overall Interpretation of Results/Actionable Insights

References

Appendix – Rstudio Code

**IST – 687 = Team# 3**

**Final Project Report**

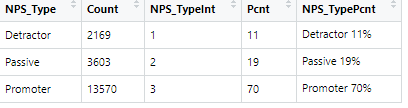
INTRODUCTION

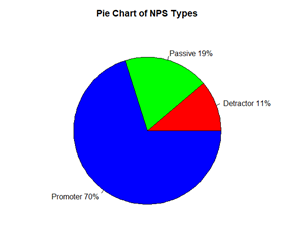
1. Company is trying to increase traffic to their chain of hotels. They would like us to see hwo would thay have to target and who could target dicounts to?

BUSINESS QUESTIONS

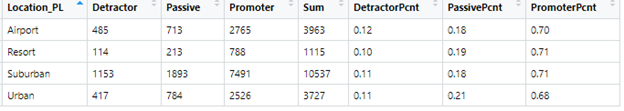
2. Hotel data – Business questions

1. How many stays were related to some discount? (See Figure 24)
2. What is the average duration of the guest stays? (See Figure 25)
3. Percentage of walk ins vs. reservations? (See Figure 26)
4. Do guests bring children? (See Figure 1)
5. Percentage of business vs leisure stays? (See Figure 2)
6. Which hotel is the most profitable? Relationship\_PL?
7. Is there a preferred booking method? (See figure 6)
8. What country has the most revenue? (See Figure 28)
9. Are Males vs. Females more likely to stay at which location? Average Ages? (see figures 3,4,5)
10. Which months are more popular among different age group? (See figure 7)
11. Which age group has the longest average stay?
12. Which age group is more like to book a stay multiple times in on month/year?
13. Likelihood of recommendation by Age group. (See figures 12 and 13)
14. What is the average of likelihood guest recommendation?
15. What is percentage of promoter and detractor? (See Figures 8, 9,10)
16. What are the different booking channels? What are the % of each of channels?
17. What is the percentage of each category in Relationship (owned, managed, franchise)? (See figure 11)
18. Which relationship category have the highest average of likelihood guest recommendation (See figure 28)
19. What are some of the drivers of NPS?
    1. We looked at how many of each NPS types we have in the data and concluded that we have a lot more Promoters in the data.

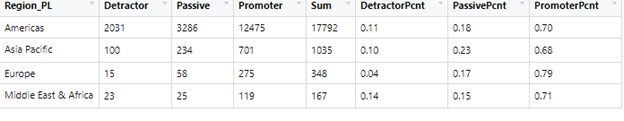




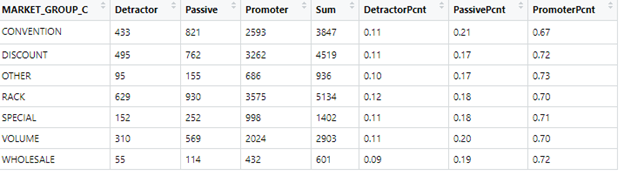
* 1. We looked at how many of each NPS Types we have for each Location and noted that the NPS Type follows the same patters across each Location as it did across all data.



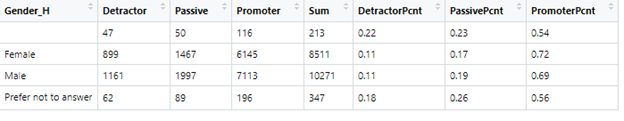
* 1. We looked at how many of each NPS Types we have for each Region and noted that we have a lot more NPS Types from Americas.



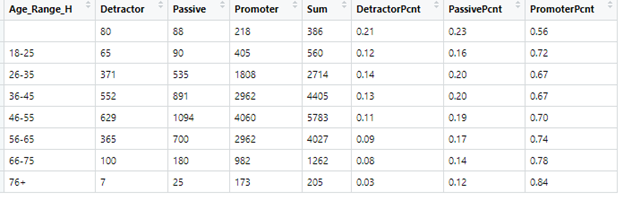
* 1. We looked at how many of each NPS Types we have for each Market Group and noted that the NPS Type follows the same patters across each Market Group as it did across all data.



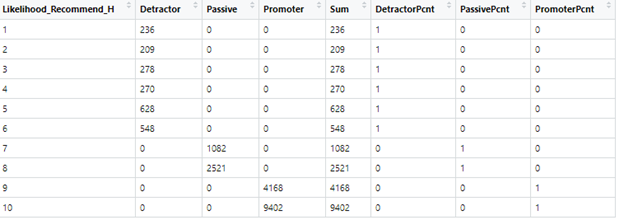
* 1. We looked at how many of each NPS Types we have for each:
     1. Gender



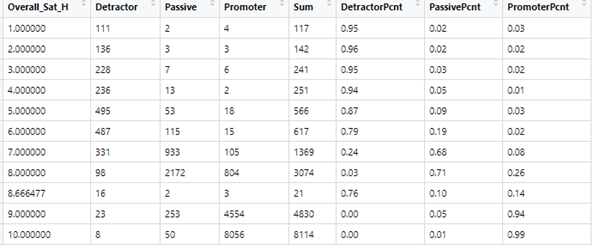
* + 1. Age Range



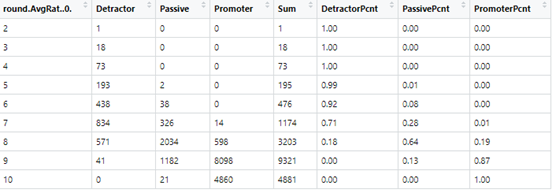
* 1. (We looked at how many of each NPS Types we have for each:
     1. Likelihood to Recommend the Hotel and noted that it’s a strong indicator of the NPS.



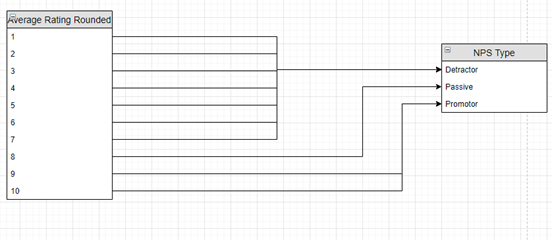
* + 1. Overall Satisfaction Hotel Score and noted that it’s a strong indicator of the NPS.



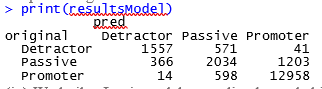
* + 1. Can we predict the NPS Type using and average of Likelihood to Recommend, Overall Satisfaction, Guest Room Rating, Tranquility Rating, Condition of Hotel Rating, Customer Service Rating, Staff Cared Rating, Internet Rating, Check In Rating, Overall Experience Rating?
       1. The rounded average across the columns gave us:



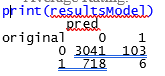
* + - 1. We built a predictive model using the following map:

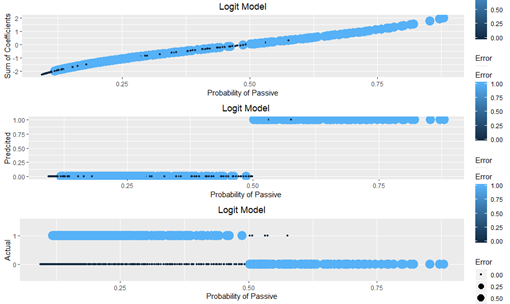


* + - 1. The result was that we did well to predict Promoters but not so well at predicting Detractors and Passives.

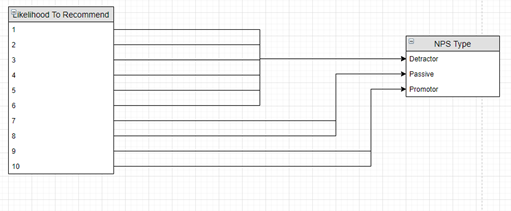


* + - 1. We built a Logit model to predict the probability of Passive based on the Average Rating.

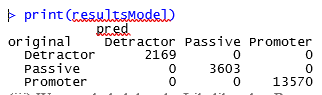




* + - 1. We found that the Logit model is not very good at predicting Passives from Average Rating.
      2. Can we predict the NPS Type using the Likelihood to Recommend?
         1. We built a predictive model using the following map:



* + - * 1. The result was:



* + - * 1. We concluded that the Likelihood to Recommend is a very strong predictor for the NPS Type.
    1. We found that the NPS type can be derived solely from Likelihood to Recommend. We found that if you use an average from all the “Rating” variables, then the Passive group becomes highly unpredictable. Our data has a lot more Promoters so it’s not surprising that we were better at predicting this NPS type. It’s also not surprising that any moderating of Likelihood to Recommend results in less predictive power.

Data Acquisition, Cleaning, Transformation

3. Process description

The data was provided via share drive in the company that has been recently updated in preparation of this project.

4. Data Selected and

The dataset provided had a significant amount of information that can be useful for this analysis. However, after reviewing the full dataset, we decided to mostly use the following:

Discount, duration of guest, purpose of stay, booking methods, Country, demographic and month of reservation.

5. What was your quality assessment

We felt that the data had a lot of useful information. There were some minimal claiming up to do but overall, we were able to use the data with satisfactory results.

6. Which fields/Variables did you select and why

We decided to use a wide variety of variables from the dataset in order to answer the Business question that the company wanted to us to answer. Find below the variables that we selected:

Discount, duration of guest, purpose of stay, booking methods, Country, demographic and month of reservation.

7. Cleaning process

One of the cleaning processes that was used, was the removal of NAs. The NAs were replaced by the mean of the column.

8. Findings.

Column of “Likelyhood of Recommendation” had a lot of NAs in there. A lot of the fields in the data were blank and or NA, making it dificult to create accurte regression models and predictive models.

Descriptive Statistics

9. Provide demographic statistics

Figures under the section of Graphs and charts provide explanation and demographics found from the collected data.

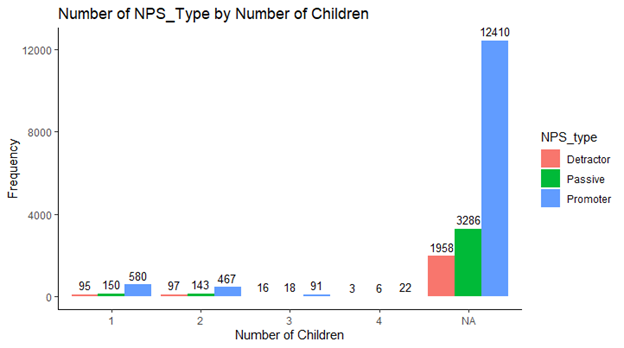
10. Any early observations, nuggets of interest, interpretation, interesting findings

One interesting finding was in Figure 7. The most popular month of reservations is February and followed by January. Based on the data, the age group of 45-55 is the most popular for this type of reservatios. That shows us that the current campaing is targeting this age group and with amazing results. Majority of the revenue is made from the hotels located in the USA.

11. Graphs, Charts, tables, visuals, text

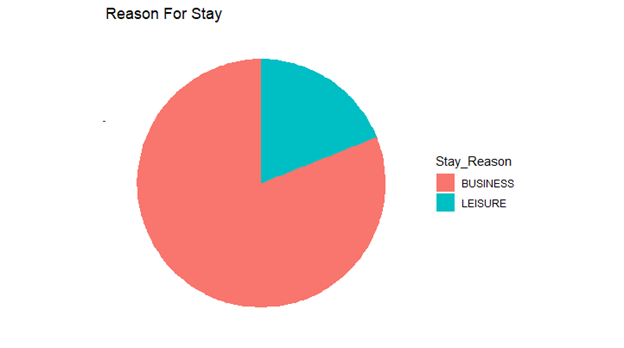
**Figure 1 (questions 4)**

Interpreting this chart let us see that majority of the respondents did not travel with children. Less than 2% of those that answered had children travel with them.



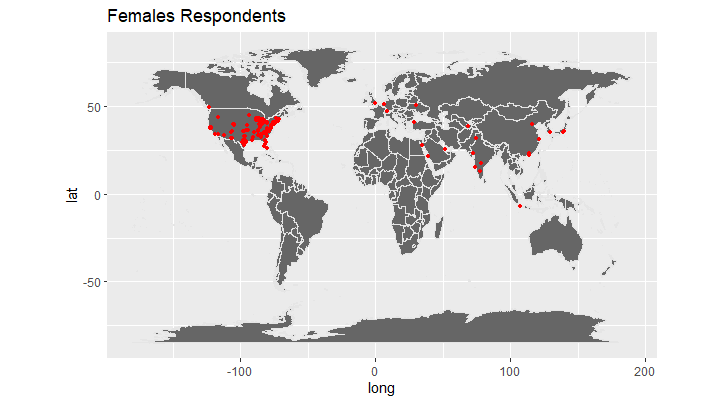
**Figure 2 (question 5)** - Percentage of business vs leisure stays

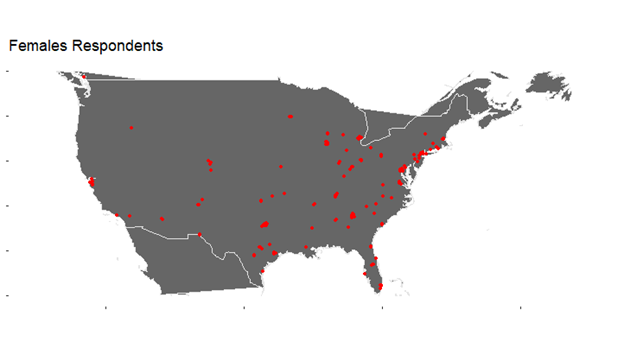
Over 2/3’s of those surveyed traveled for Business. Better to target groups or offer incentives to companies.



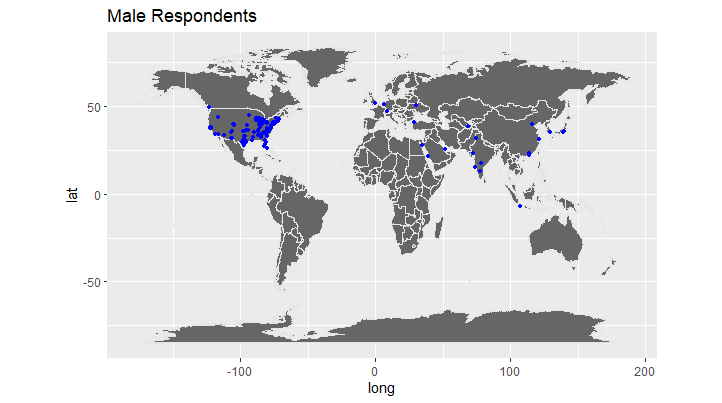
**Figure 3 (question 9)**

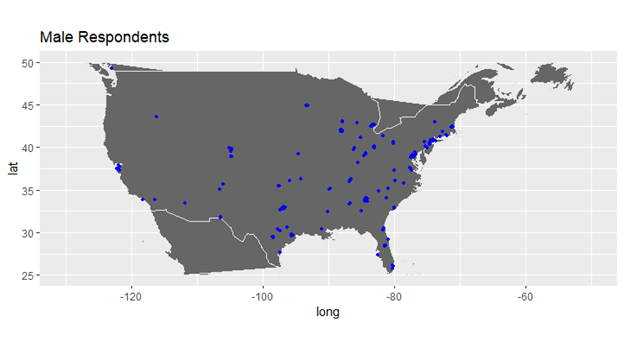
**Examining our respondents, majority of our population was staying in hotels in the USA.**





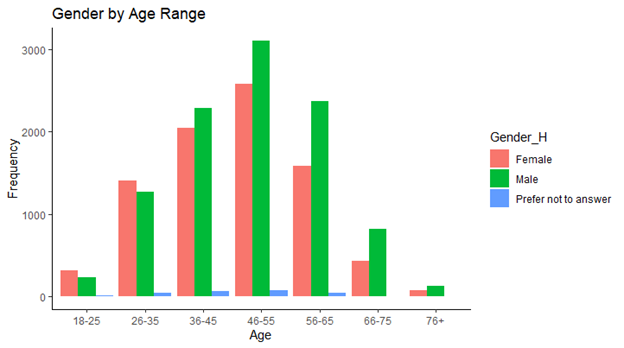
**Figure 4 (question 9)**

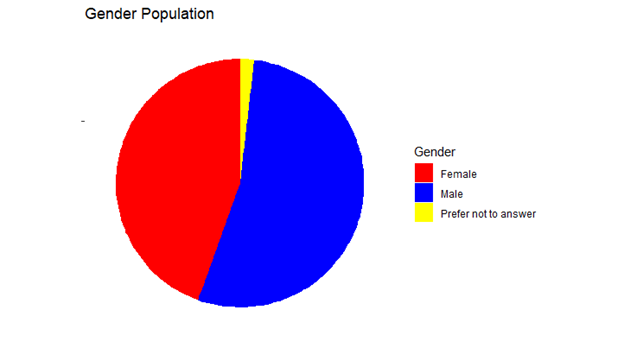




**Figure 5 (question 9)**

**When looking at our gender breakdown by age. We see that more females are staying with us at younger ages vs an increase in male occupants after the 36-45 age groups.**

Over all there is an even population between the genders



**Figure 6**

**As expected the majority of people are booking their reservation electronically. The least comon way to book the reservation based on this data, is by booking at the front desk of the Hotel**

A screenshot of a cell phone

Description automatically generated

**Figure 7**

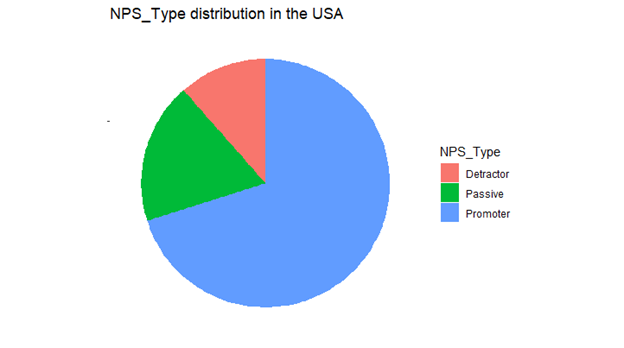
**It’s pretty clear by looking at the Figure 7 that the most popular month of reservations is February and followed by January and the age group of 45-55 . That shows us that the current campaing is targeting this age group and with amazing resutsl. In this case, I would suggest the campagin focus on age groups of 18-25, 26-35 and 36-45. I would also combine the age groups with the months of March, April and May since the reservatios seems to be extremely low.**

A screenshot of a cell phone

Description automatically generated

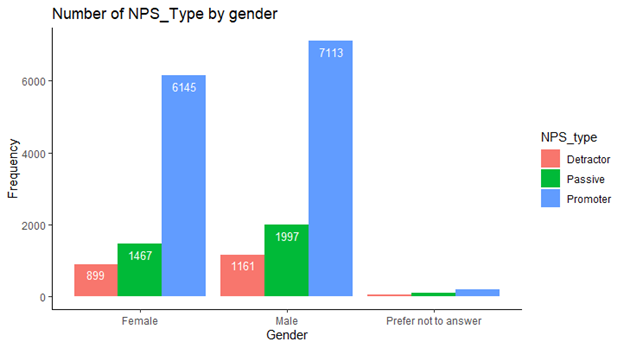
**Figure 8 (question 15)**

**Majority of our population from the data were Promoters**



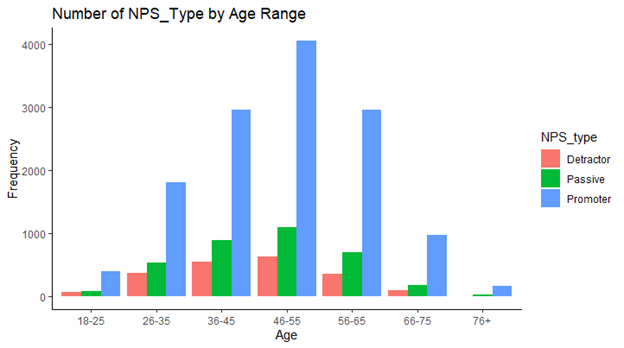
**Figure 9 (question 15)**

**Viewing a break down of our NPS\_type by gender, revealed that we had less than a 2% difference in our populations**

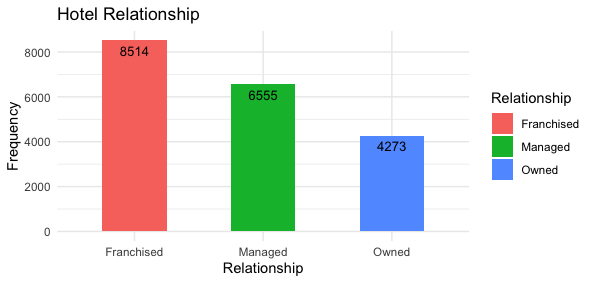


**Figure 10 (question 15)**

**Majority of our promoters were between the 36-65 age groups.**

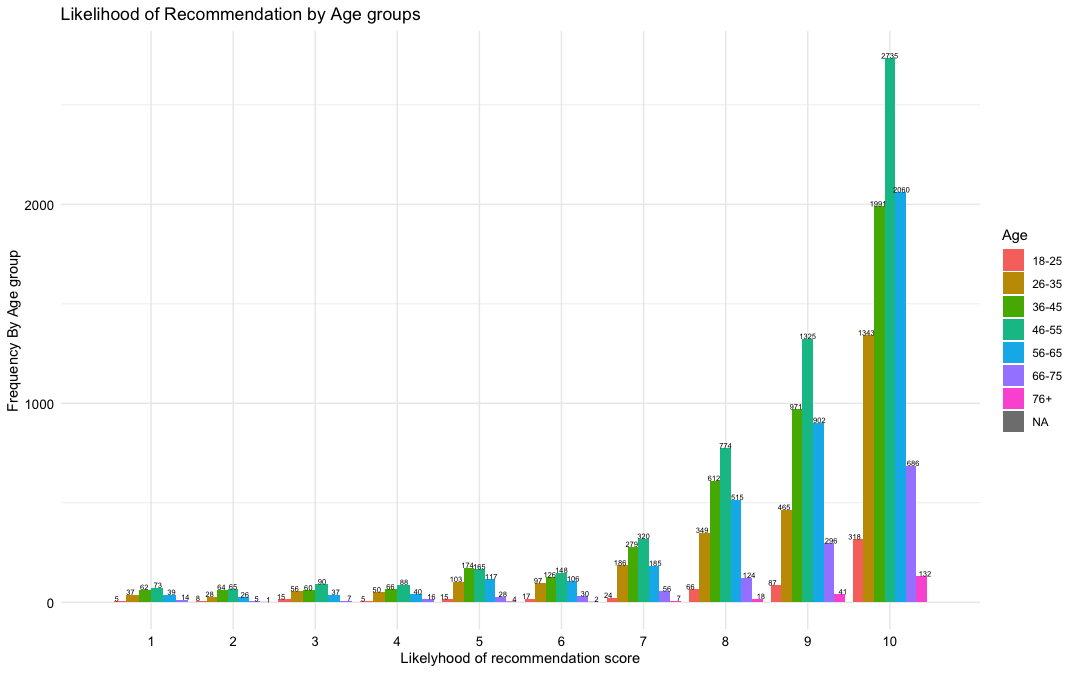


**Figure 11**



**Figure 12**

**Based on this data set, we can see that most people will likely recommend the hotel with a score of 10. In the next figure, we can see where the opportunities for improvement are.**



**Figure 13**

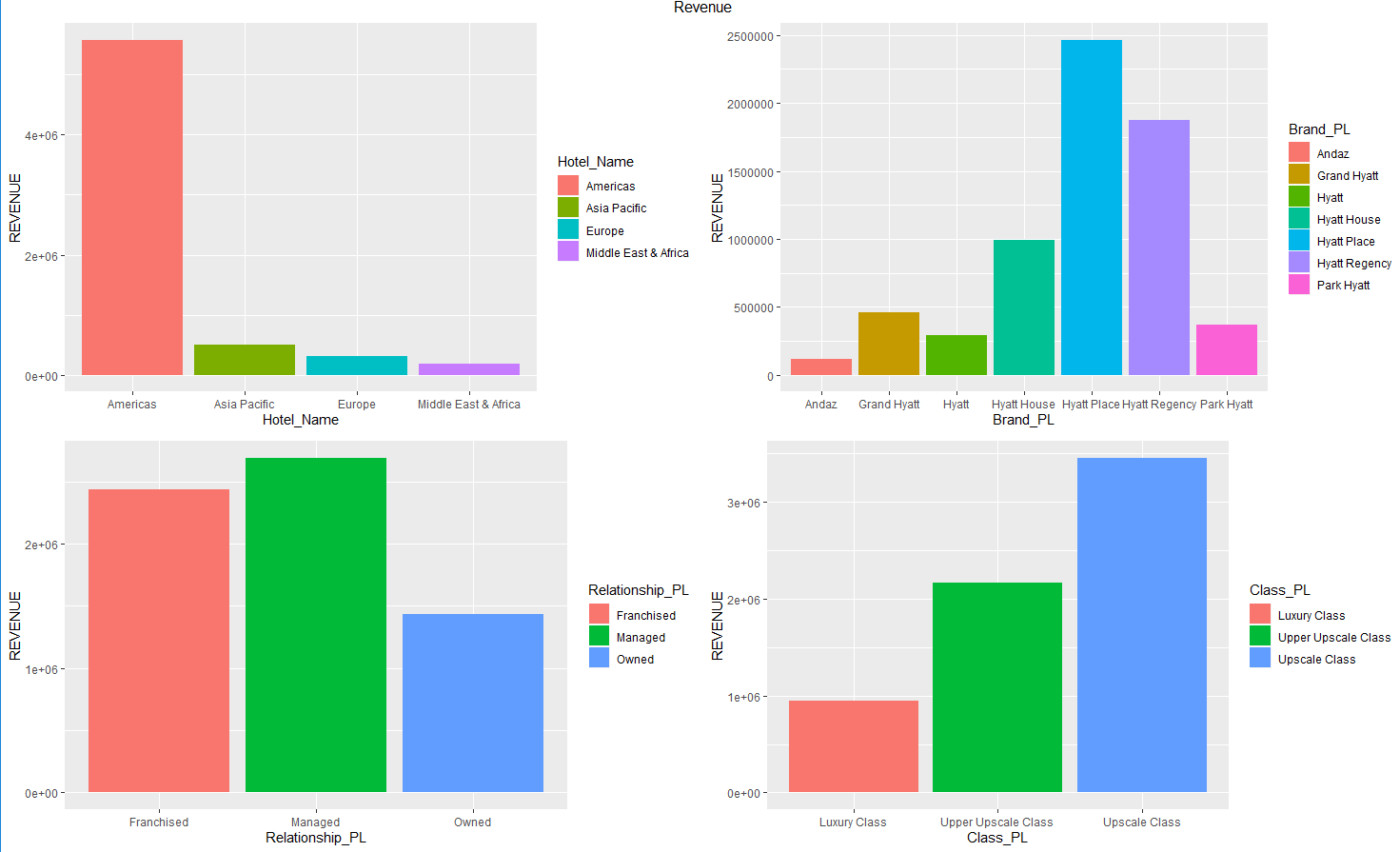
**In this graph, we have removed scores of 10 and 9s. This gives us a better picture of where we can focus to improve.**

A screenshot of a cell phone

Description automatically generated

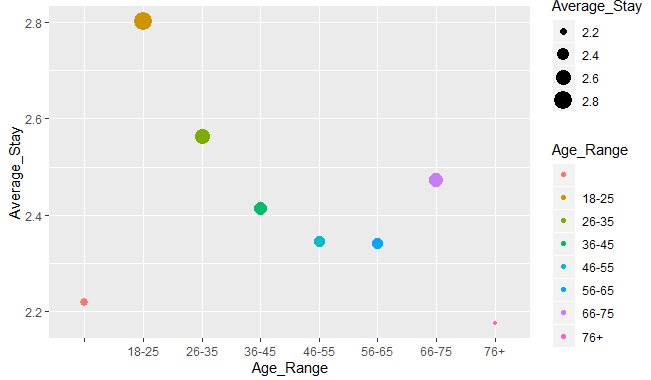
**Figure 14**

**Revenue by Region, Brand, Relationship and Class**

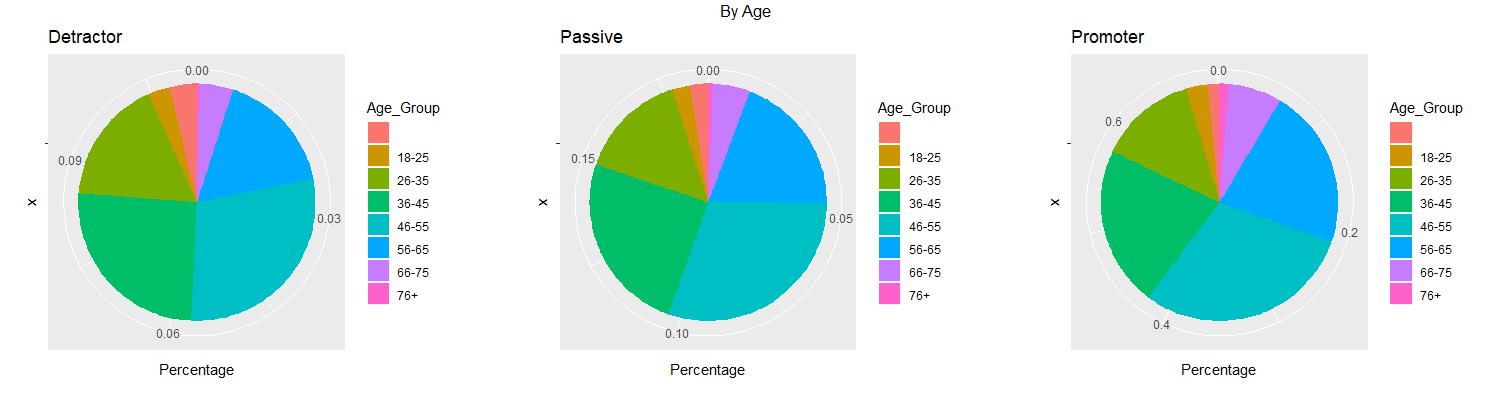


**Figure 15**

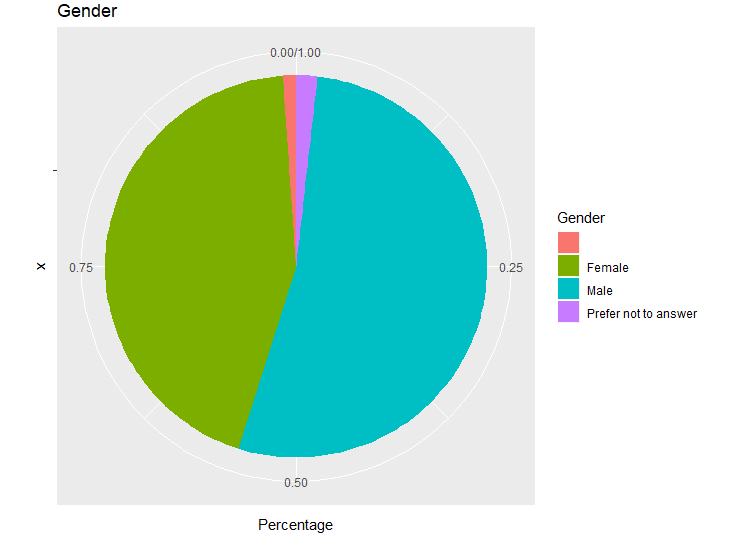
**Average Stay per group age: It appears group ages only stay around 2 nights, but the rounger age group tents to stay closer to 3 nights, which could mean that this age group spends more per capital**



**Figure 16**



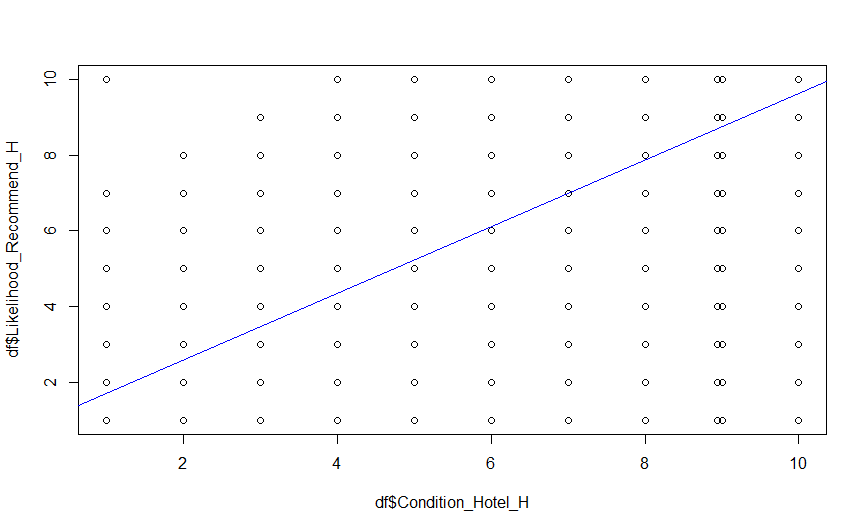
**Figure 17**



**Figure 18-21**

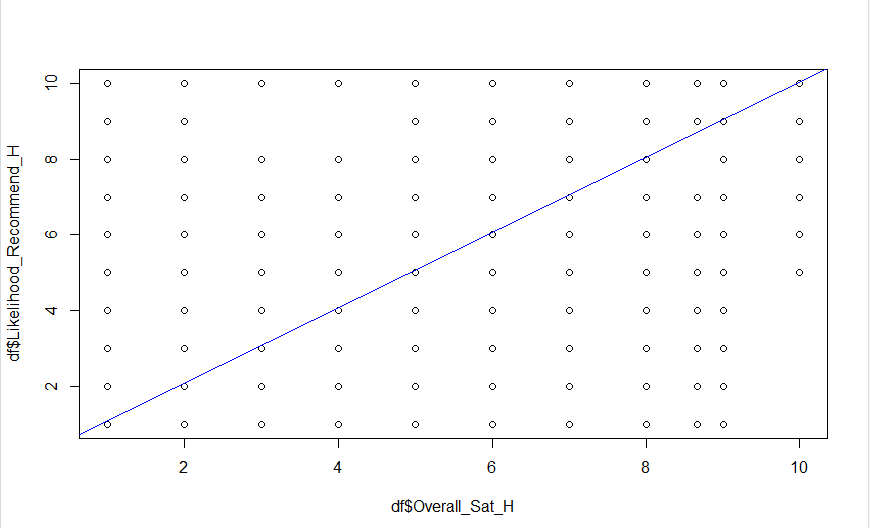
Regression model predicting Likelihood of Recommend based on Condition of Hotel

Adjusted R-squared: 0.4863



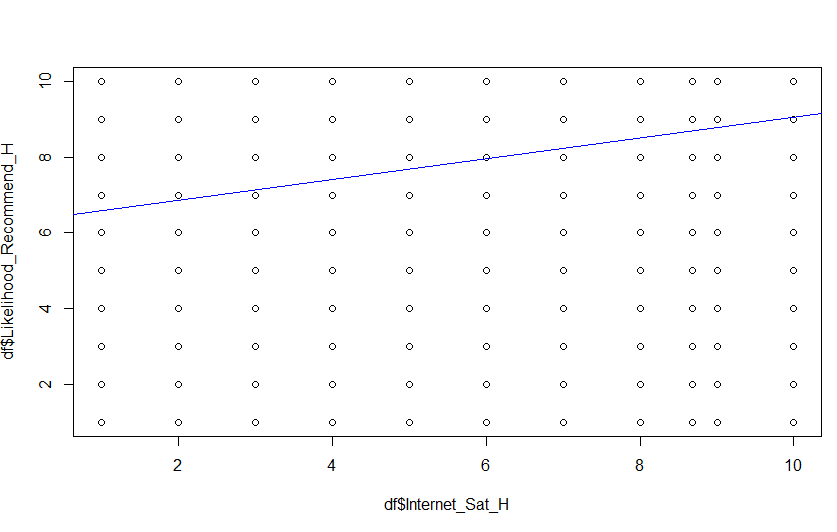
Regression model predicting Likelihood of Recommend based on Overall Satisfaction of Hotel

R-squared: 0.8047



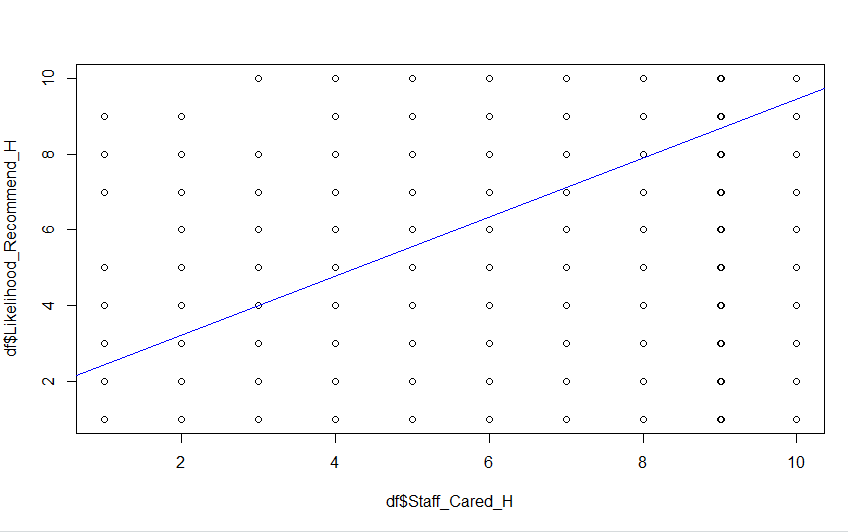
Regression model predicting Likelihood of Recommend based on internet Satisfaction

Adjusted R-squared: 0.03358



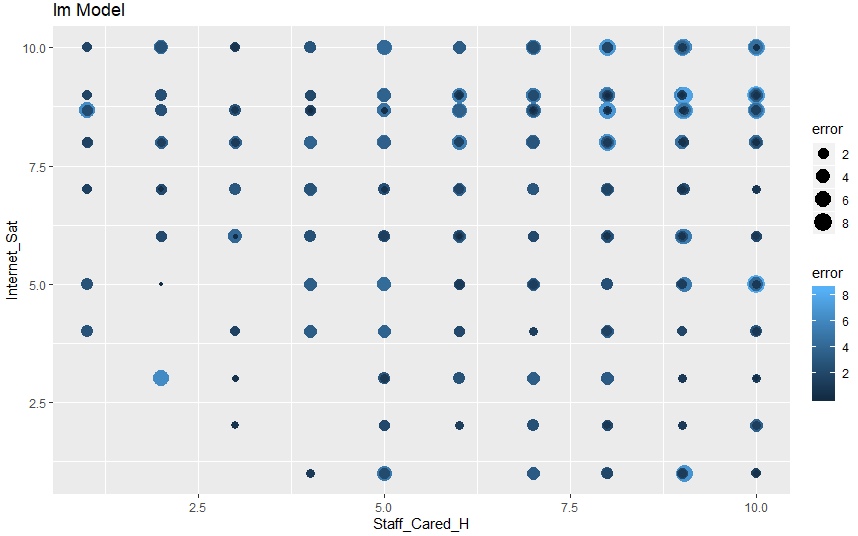
Regression model predicting Likelihood of Recommend based on Staff cared

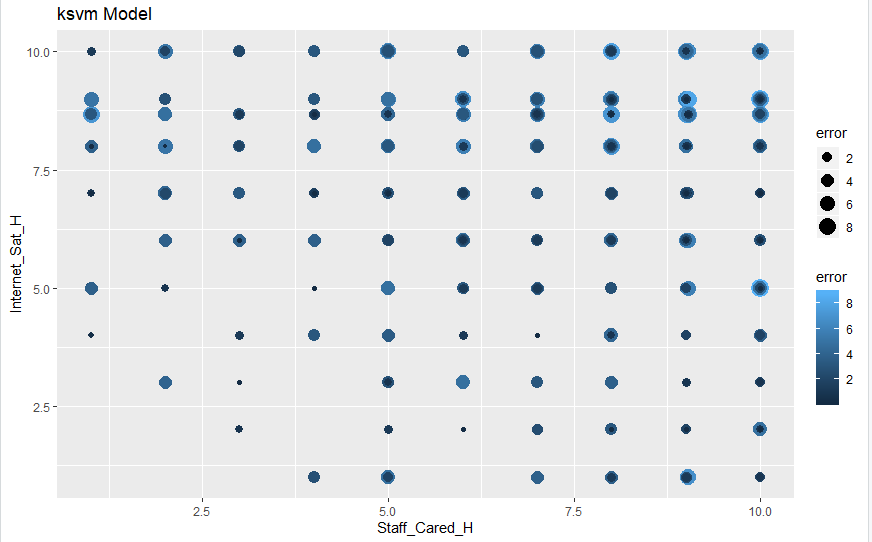
Adjusted R-squared: 0.2408



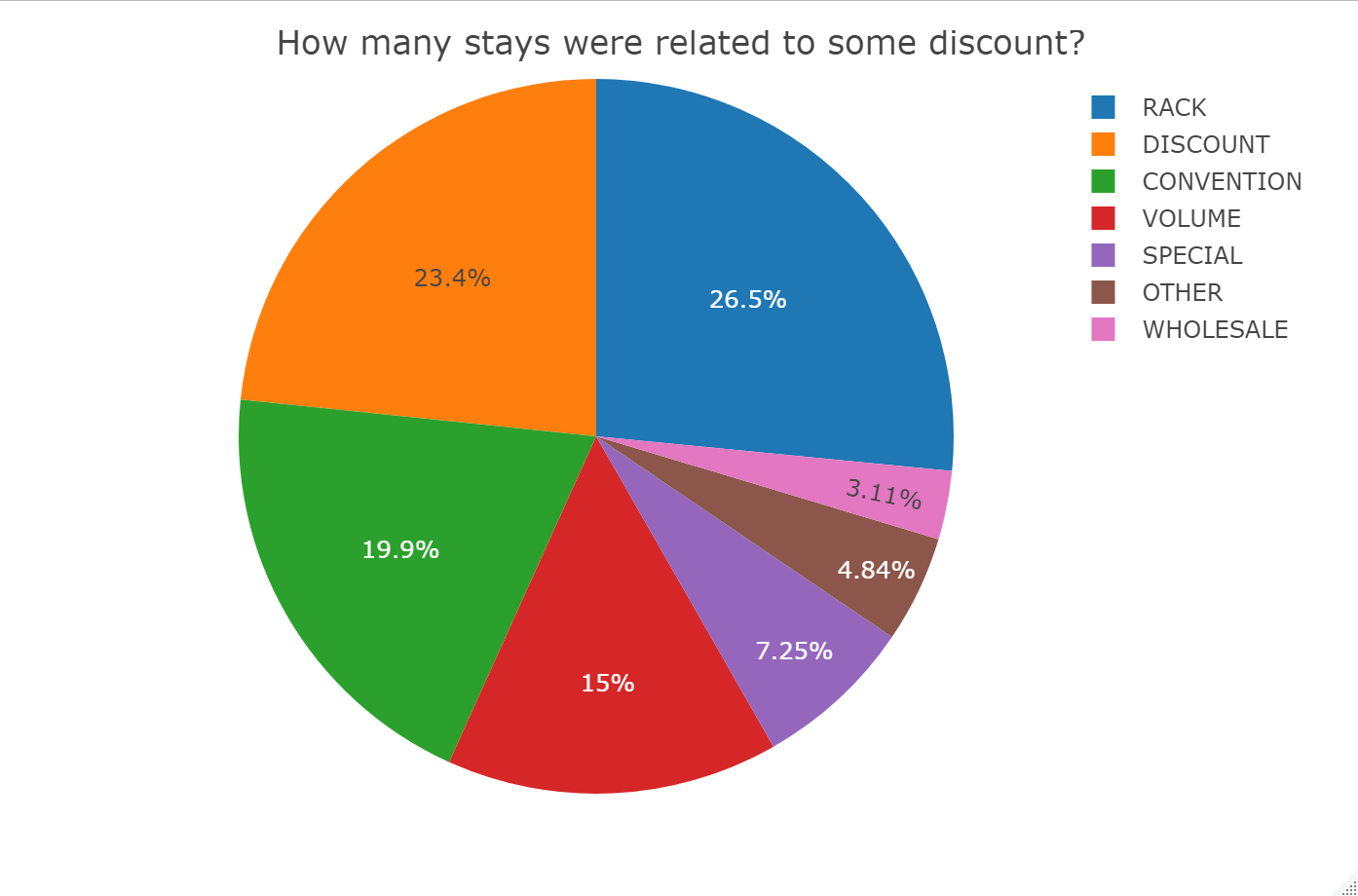
**Figure 22- 23**

**Predictive models using Internet Satisfaction and Staff Cared**

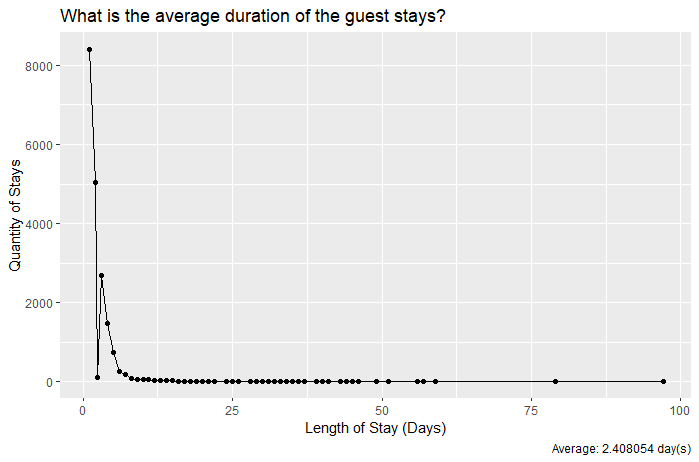




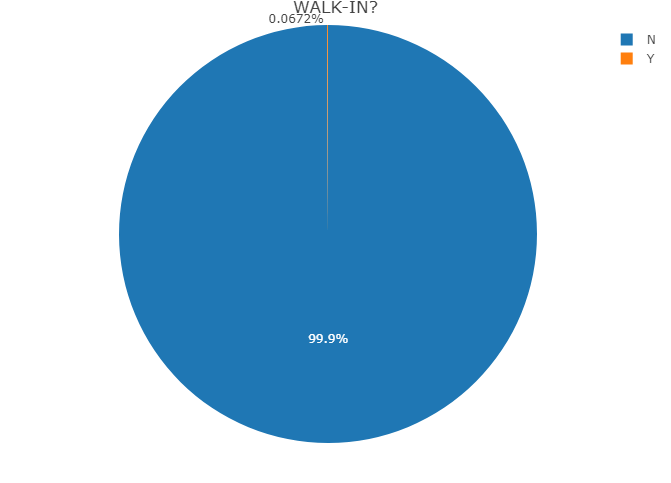
**Figure 24 (question 1)**



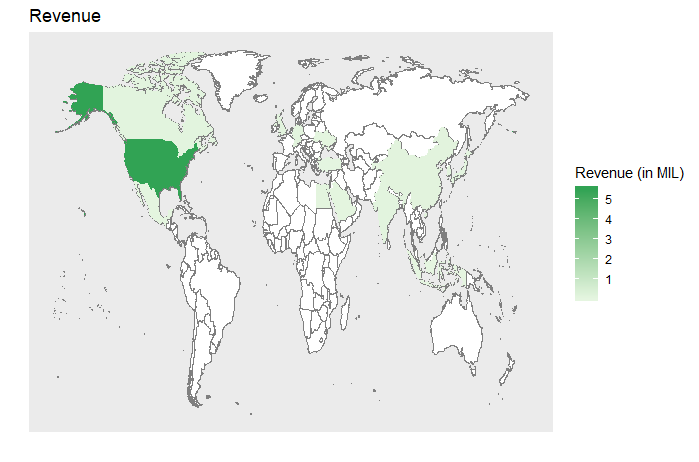
**Figure 25 (question 2)**



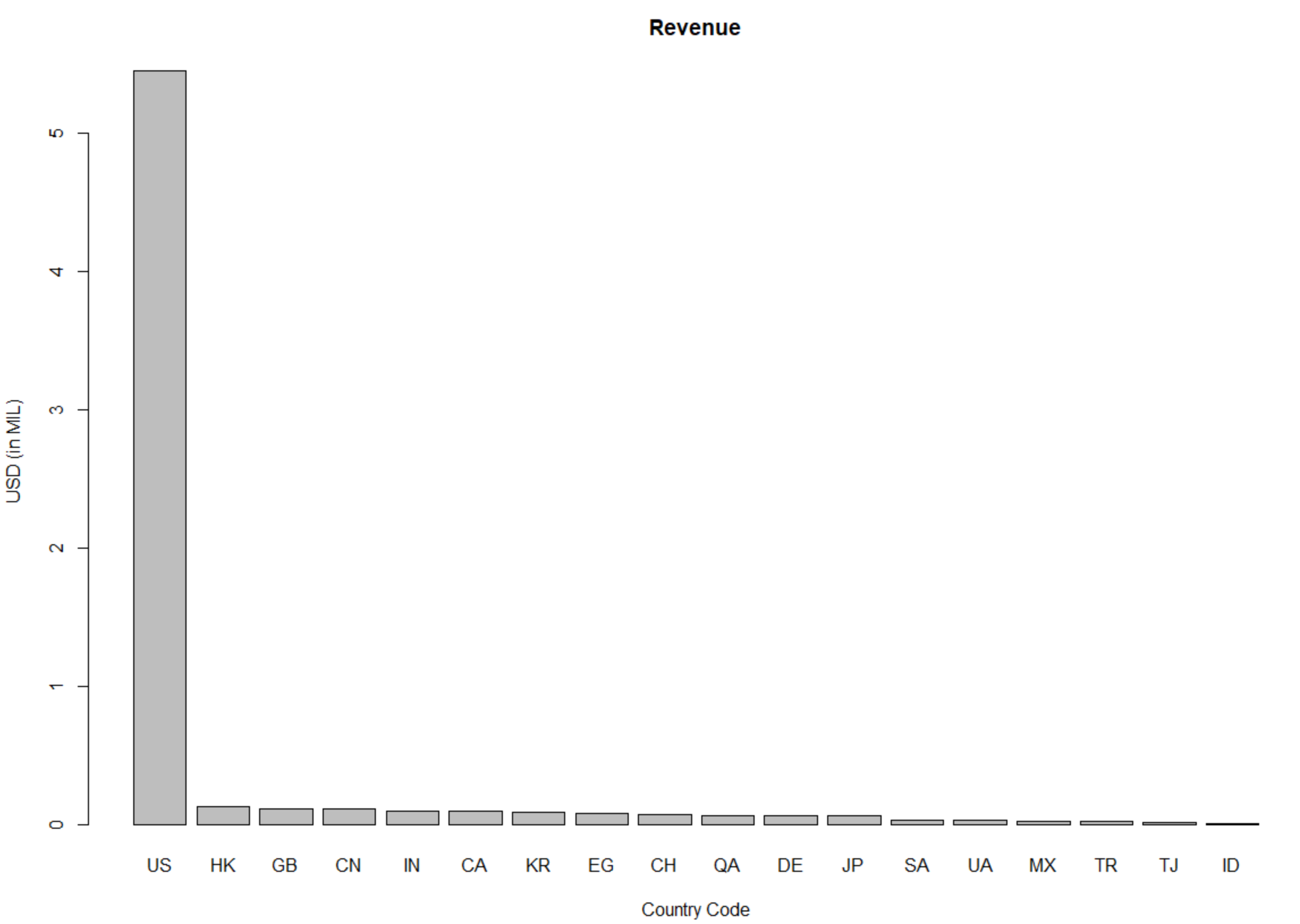
**Figure 26 (question 3)**



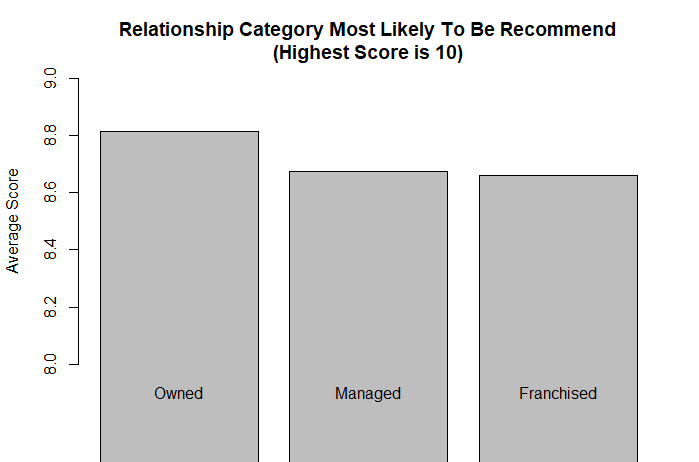
**Figure 27 (question 8)**



**Figure 27 (question 8) cont.**



**Figure 28 (question 18)**



Use of Modeling Techniques

12. Linear modeling

13. Regression

14. KSVM

Overall Interpretation of Results/Actionable Insights

References

Appendix – Rstudio Code

Codes for Q4 Figure 1

Library (ggplot)

#Number of children

Path <- "C:/Users/mesom/iCloudDrive/Syracuse/IST 687/Project Info/ProjectSurveyData\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2.xlsx"

myData<-read\_xlsx(Path)

unique(myData$CHILDREN\_NUM\_C)

myData <- myData[myData$CHILDREN\_NUM\_C != 5,]

myData <- myData[myData$CHILDREN\_NUM\_C != 6,]

agg.data <- aggregate(myData$count, by = list(children = myData$CHILDREN\_NUM\_C, NPS\_type = myData$NPS\_Type),FUN = sum)

p3 <- ggplot(agg.data, aes(x=children, y=x, fill=NPS\_type)) +

geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=-.5, color="black", position = position\_dodge(.9), size=3.5) +

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.line = element\_line(colour = "black")) +

ggtitle("Number of NPS\_Type by Number of Children") + # for the main title

xlab("Number of Children") + # for the x axis label

ylab("Frequency") # for the y axis label

Code for Q5 and Figure 2

Library (ggplot)

#business vs leisure

Path <- "C:/Users/mesom/iCloudDrive/Syracuse/IST 687/Project Info/ProjectSurveyData\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2.xlsx"

myData<-read\_xlsx(Path)

reason<-as.data.frame(table(myData$POV\_CODE\_C))

colnames(reason)<-c("Stay\_Reason", "Freq")

rc<- ggplot(reason, aes(x="", y=Freq, fill=Stay\_Reason))+ geom\_bar(width = 1, stat = "identity")+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

rc

pie <- rc + coord\_polar("y", start=0) + ggtitle("Reason For Stay") # for the main title

pie

Codes for Q9 and Figures 3, 4, 5

Library (ggplot)

Path <- "C:/Users/mesom/iCloudDrive/Syracuse/IST 687/Project Info/ProjectSurveyData\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2.xlsx"

myData<-read\_xlsx(Path)

Gender\_maps <- myData[,c("Property.Latitude\_PL","Property.Longitude\_PL","Gender\_H")]

#separating points by NPS\_TYPE

unique(Gender\_maps$Gender\_H)

#Female

Female <- Gender\_maps[Gender\_maps$Gender\_H == "Female",]

#Male

Male <- Gender\_maps[Gender\_maps$Gender\_H == "Male",]

world <- map\_data("world")

hotelMap <- 0

hotelMap <- ggplot() + geom\_polygon(data = world, aes(x=long, y = lat, group = group),

fill="grey40", colour="grey90", alpha=1) + coord\_fixed(1.3)

FemaleMap <-hotelMap + geom\_point(aes(x = Female$Property.Longitude\_PL, y =

Female$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Females Respondents")+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

FemaleMap <-hotelMap + geom\_point(aes(x = Female$Property.Longitude\_PL, y =

Female$Property.Latitude\_PL), color = "red", size = 1) + ggtitle("Females Respondents")+xlim(-130,-50)+ylim(25,50)+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

MaleMap <-hotelMap + geom\_point(aes(x = Male$Property.Longitude\_PL, y =

Male$Property.Latitude\_PL), color = "blue", size = 1) + ggtitle("Male Respondents")+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

MaleMap

MaleMap <-hotelMap + geom\_point(aes(x = Male$Property.Longitude\_PL, y =

Male$Property.Latitude\_PL), color = "blue", size = 1) + ggtitle("Male Respondents")+xlim(-130,-50)+ylim(25,50)+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

#Gender by age

agg.data <- aggregate(myData$count, by = list(age = myData$Age\_Range\_H, Gender\_H = myData$Gender\_H), FUN = sum)

agg.data <- agg.data[agg.data$age != "",]

p5 <- ggplot(agg.data, aes(x=age, y=x, fill=Gender\_H)) +

geom\_bar(stat="identity", position=position\_dodge()) +

#geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.line = element\_line(colour = "black")) +

ggtitle("Gender by Age Range") + # for the main title

xlab("Age") + # for the x axis label

ylab("Frequency") # for the y axis label

p5

#Gender pie chart

gender <- as.data.frame(table(myData$Gender\_H))

colnames(gender) <- c("Gender", "Freq")

gender <- gender[gender$Gender != "",]

#gender <- gender[gender$Var1 != "Prefer not to answer",]

bp<- ggplot(gender, aes(x="", y=Freq, fill=Gender))+

geom\_bar(width = 1, stat = "identity")+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

bp

pie <- bp + coord\_polar("y", start=0) +

ggtitle("Gender Population") +# for the main title

scale\_fill\_manual(values=c("red", "blue", "yellow"))

pie

Codes for Q7 and Figure 6 (MF)

install.packages("readxl")

install.packages("base")

library(ggplot2)

library(base)

library(readxl)

newdata <- read\_excel("Desktop/School/IST/Project1data.xlsx")

colnames(newdata)

n#<- ggplot(newdata, aes(newdata$Booking\_Channel)) + geom\_histogram(binwidth = 2, color="yellow")

#n

n<-as.data.frame(table(newdata$Booking\_Channel))

colnames(n)<-c("Booking\_Channel", "Frequency")

bar1<- ggplot(n, aes(x=Booking\_Channel, y=Frequency, fill=Booking\_Channel)) + geom\_bar(stat = "identity")+

ggtitle("Preferred booking methods") +

xlab("Booking Channels") +

ylab("Frequency")

bar1

Code to Q10 and figure 7 (MF)

library(readxl)

mydata <- read\_excel("Desktop/School/IST/Project1data.xlsx")

#11. Which day of the month is more popular among different age group for reservations? (MF)

str(mydata)

mydata$RESERVATION\_DATE\_R <- as.Date(mydata$RESERVATION\_DATE\_R)

mydata$RESERVATION\_DATE\_R = as.numeric(format(mydata$RESERVATION\_DATE\_R, "%m"))

mydata$count <- 1

data2 <- aggregate(mydata$count, by = list(age = mydata$Age\_Range\_H, month= mydata$RESERVATION\_DATE\_R), FUN=sum)

data2 <- data2[data2 != "",]

plot4 <- ggplot(data2, aes(x=month, y=x, fill=age)) +

geom\_bar(stat="identity", position=position\_dodge()) +

#geom\_text(aes(label=x), vjust=0, color="white", position = position\_dodge(1), size=2)+

theme\_minimal() +

ggtitle("Popular month of reservation by age group") + # for the main title

xlab("Month of reservation") + # for the x axis label

ylab("Frequency") +

theme(axis.text.x = element\_text(face="plain", color="black",

size=10, angle=0),

axis.text.y = element\_text(face="plain", color="black",

size=10, angle=0)) +

scale\_x\_discrete(limits=c("1", "2", "3","4", "5", "6", "7", "8", "9", "10", "11", "12"))

plot4

Codes Q13 and figures 12 and 13) (MF)

library(readxl)

mydata <- read\_excel("Desktop/School/IST/Project1data.xlsx")

str(mydata)

mydata$count <- 1

data3 <- aggregate(mydata$count, by = list(Age = mydata$Age\_Range\_H, score= mydata$Likelihood\_Recommend\_H), FUN=length)

######

data3 <- data3[data3 != "",]

plot8 <- ggplot(data3, aes(x=score, y=x, fill=Age)) +

geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=0, color="black", position = position\_dodge(1), size=2)+

theme\_minimal() +

ggtitle("Likelihood of Recommendation by Age groups") +

xlab("Likelyhood of recommendation score") +

ylab("Frequency By Age group") +

theme(axis.text.x = element\_text(face="plain", color="black",

size=10, angle=0),

axis.text.y = element\_text(face="plain", color="black",

size=10, angle=0)) +

scale\_x\_discrete(limits=c("1", "2", "3","4", "5", "6", "7", "8", "9", "10"))

plot8

#plot without 10 and 9s scores

mydata$count <- 1

ndata <- mydata

ndata <- ndata[ndata$Likelihood\_Recommend\_H != "10", ]

ndata <- ndata[ndata$Likelihood\_Recommend\_H!= "9", ]

ndata

ndata <- na.omit(ndata)

ndata

ndata <- aggregate(ndata$count, by = list(Age = ndata$Age\_Range\_H, score= ndata$Likelihood\_Recommend\_H), FUN=length)

ndata <- ndata[ndata != "",]

plot9 <- ggplot(ndata, aes(x=score, y=x, fill=Age)) +

geom\_bar(stat="identity", position=position\_dodge()) +

geom\_text(aes(label=x), vjust=0, color="black", position = position\_dodge(1), size=2)+

theme\_minimal() +

labs(title= "Likelihood of Recommendation by Age groups",

subtitle= "Without 9 and 10 scores") +

xlab("Likelihood of Recommendation scores") +

ylab("Frequency by Age group") +

theme(axis.text.x = element\_text(face="plain", color="black",

size=10, angle=0),

axis.text.y = element\_text(face="plain", color="black",

size=10, angle=0)) +

scale\_x\_discrete(limits=c("1", "2", "3","4", "5", "6", "7", "8"))

plot9

Codes to Q7 and figure 11 (MF)

library(readxl)

data <- read\_excel("Desktop/School/IST/Project1data.xlsx")

str(data)

any(is.na(tdata))

n1<-as.data.frame(table(data$Relationship\_PL))

colnames(n1)<-c("Relationship", "Frequency")

bar2<- ggplot(n1, aes(x=Relationship, y=Frequency, fill= Relationship)) + geom\_bar(stat = "identity", width = 0.5)+

ggtitle("Hotel Relationship") +

xlab("Relationship") +

ylab("Frequency") +

geom\_text(aes(label=Frequency), vjust=1.6, color="black", position = position\_dodge(0.9), size=3.5)+

theme\_minimal()

bar2

Code for Q15 and Figures 8, 9 and 10

Path <- "C:/Users/mesom/iCloudDrive/Syracuse/IST 687/Project Info/ProjectSurveyData\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2.xlsx"

myData<-read\_xlsx(Path)

#Country

myData <- myData[myData$Country\_PL == "United States",]

USA <- as.data.frame(table(myData$NPS\_Type))

colnames(USA) <- c("NPS\_Type", "Freq")

bp<- ggplot(USA, aes(x="", y=Freq, fill=NPS\_Type))+ geom\_bar(width = 1, stat = "identity")+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.title = element\_blank(), axis.text = element\_blank())

bp

pie <- bp + coord\_polar("y", start=0) +

ggtitle("NPS\_Type distribution in the USA") # for the main title

pie

#NPS VS Gender

myData$count <- 1

agg.data <- aggregate(myData$count, by = list(gender = myData$Gender\_H, NPS\_type =

myData$NPS\_Type), FUN = sum)

p1 <- ggplot(agg.data, aes(x=gender, y=x, fill=NPS\_type)) + geom\_bar(stat="identity", position=position\_dodge()) + geom\_text(aes(label=x), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)+ theme\_minimal() + ggtitle("Number of NPS\_Type by gender") + # for the main title

xlab("Gender") + # for the x axis label

ylab("Frequency") # for the y axis label

p1

#Number of NPS\_Type by age range

agg.data <- aggregate(myData$count, by = list(age = myData$Age\_Range\_H, NPS\_type = myData$NPS\_Type), FUN = sum)

agg.data <- agg.data[agg.data$age != "",]

p4 <- ggplot(agg.data, aes(x=age, y=x, fill=NPS\_type)) +

geom\_bar(stat="identity", position=position\_dodge()) +

#geom\_text(aes(label=x), vjust=-.5, color="black", position = position\_dodge(0.9), size=3.5)+

theme\_bw() + theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.line = element\_line(colour = "black")) +

ggtitle("Number of NPS\_Type by Age Range") + # for the main title

xlab("Age") + # for the x axis label

ylab("Frequency") # for the y axis label

p4

**Code to Figure 14 – 22**

library("RCurl")

library("RJSONIO")

library("sqldf")

library("jsonlite")

library(readxl)

library(moments)

library(ggplot2)

library(arulesViz)

library(e1071)

library(arules)

library(caret)

library(gridExtra)

library(kernlab)

setwd("H:\\School\\IST 687")

proj\_data <- read.csv("data.csv",stringsAsFactors = FALSE)

df<- proj\_data

str(df)

View(df)

length(df[df=='NA'])

colnames(df)[colSums(is.na(df))>0]

df$LENGTH\_OF\_STAY\_C[is.na(df$LENGTH\_OF\_STAY\_C)] <- mean(df$LENGTH\_OF\_STAY\_C,na.rm=TRUE)

df$NUMBER\_OF\_ROOMS\_C[is.na(df$NUMBER\_OF\_ROOMS\_C)] <- mean(df$NUMBER\_OF\_ROOMS\_C,na.rm=TRUE)

df$ADULT\_NUM\_C[is.na(df$ADULT\_NUM\_C)] <- mean(df$ADULT\_NUM\_C,na.rm=TRUE)

df$CHILDREN\_NUM\_C[is.na(df$CHILDREN\_NUM\_C)] <- mean(df$CHILDREN\_NUM\_C,na.rm=TRUE)

df$Overall\_Sat\_H[is.na(df$Overall\_Sat\_H)] <- mean(df$Overall\_Sat\_H,na.rm=TRUE)

df$Guest\_Room\_H[is.na(df$Guest\_Room\_H)] <- mean(df$Guest\_Room\_H,na.rm=TRUE)

df$Tranquility\_H[is.na(df$Tranquility\_H)] <- mean(df$Tranquility\_H,na.rm=TRUE)

df$Condition\_Hotel\_H[is.na(df$Condition\_Hotel\_H)] <- mean(df$Condition\_Hotel\_H,na.rm=TRUE)

df$Customer\_SVC\_H[is.na(df$Customer\_SVC\_H)] <- mean(df$Customer\_SVC\_H,na.rm=TRUE)

df$Staff\_Cared\_H[is.na(df$Staff\_Cared\_H)] <- mean(df$Staff\_Cared\_H,na.rm=TRUE)

df$Internet\_Sat\_H[is.na(df$Internet\_Sat\_H)] <- mean(df$Internet\_Sat\_H,na.rm=TRUE)

df$Check\_In\_H[is.na(df$Check\_In\_H)] <- mean(df$Check\_In\_H,na.rm=TRUE)

df$F.B\_FREQ\_H[is.na(df$F.B\_FREQ\_H)] <- mean(df$F.B\_FREQ\_H,na.rm=TRUE)

df$F.B\_Overall\_Experience\_H[is.na(df$F.B\_Overall\_Experience\_H)] <- mean(df$F.B\_Overall\_Experience\_H,na.rm=TRUE)

length(df[df=='na'])

Q6<- sqldf("select Region\_PL, sum(REVENUE\_USD\_R) FROM df group by Region\_PL")

Q6

df3 <-data.frame(Q6)

colnames(df3) <- c("Hotel\_Name", "REVENUE")

plotQ6<- ggplot(df3, aes(x=Hotel\_Name, y=REVENUE, fill=Hotel\_Name)) + geom\_bar(stat = "identity")

plotQ6

#

Q6b<- sqldf("select Brand\_PL, sum(REVENUE\_USD\_R) FROM df group by Brand\_PL")

Q6b

df3b <-data.frame(Q6b)

colnames(df3b) <- c("Brand\_PL", "REVENUE")

plotQ6b<- ggplot(df3b, aes(x=Brand\_PL, y=REVENUE, fill=Brand\_PL)) + geom\_bar(stat = "identity")

plotQ6b

#

Q6c<- sqldf("select Relationship\_PL, sum(REVENUE\_USD\_R) FROM df group by Relationship\_PL")

Q6c

df3c <-data.frame(Q6c)

colnames(df3c) <- c("Relationship\_PL", "REVENUE")

plotQ6c<- ggplot(df3c, aes(x=Relationship\_PL, y=REVENUE, fill=Relationship\_PL)) + geom\_bar(stat = "identity")

plotQ6c

#

Q6d<- sqldf("select Class\_PL, sum(REVENUE\_USD\_R) FROM df group by Class\_PL")

Q6d

df3d <-data.frame(Q6d)

colnames(df3d) <- c("Class\_PL", "REVENUE")

plotQ6d<- ggplot(df3d, aes(x=Class\_PL, y=REVENUE, fill=Class\_PL)) + geom\_bar(stat = "identity")

plotQ6d

#

grid.arrange(plotQ6,plotQ6b, plotQ6c, plotQ6d, ncol=2, nrow=2, top="Revenue")

#----------------------------

Q72 <- sqldf("select Age\_Range\_H, avg(LENGTH\_OF\_STAY\_C) FROM df group by Age\_Range\_H")

df2 <-data.frame(Q72)

colnames(df2) <- c("Age\_Range", "Average\_Stay")

plotage2 <- ggplot(df2, aes(x=Age\_Range, y=Average\_Stay)) + geom\_point(aes(size=Average\_Stay, color=Age\_Range))

plotage2

#----------------------------

q13<- sqldf("select NPS\_Type, Age\_Range\_H, count(\*) FROM df where NPS\_Type like '%Detra%' group by NPS\_Type, Age\_Range\_H")

q13

dfq13 <-data.frame(q13)

colnames(dfq13) <- c("NPS\_Type", "Age\_Range\_H","Count")

dfq13\_perc <- (dfq13$Count/dim(df)[1])

plotq13<- data.frame(dfq13$NPS\_Type,dfq13$Age\_Range\_H,dfq13\_perc )

colnames(plotq13) <- c("NPS\_Type","Age\_Group", "Percentage")

pie13<- ggplot(plotq13, aes(x="", y=Percentage, fill=Age\_Group))+ geom\_bar(width = 1, stat = "identity")

pie13

pie2 <- pie13 + coord\_polar("y", start=0) + ggtitle("Detractor")

pie2

q13a<- sqldf("select NPS\_Type, Age\_Range\_H, count(\*) FROM df where NPS\_Type like '%Passi%' group by NPS\_Type, Age\_Range\_H")

q13a

dfq13a <-data.frame(q13a)

colnames(dfq13a) <- c("NPS\_Type", "Age\_Range\_H","Count")

dfq13a\_perc <- (dfq13a$Count/dim(df)[1])

plotq13a<- data.frame(dfq13a$NPS\_Type,dfq13a$Age\_Range\_H,dfq13a\_perc )

colnames(plotq13a) <- c("NPS\_Type","Age\_Group", "Percentage")

pie13a<- ggplot(plotq13a, aes(x="", y=Percentage, fill=Age\_Group))+ geom\_bar(width = 1, stat = "identity")

pie13a

pie2a <- pie13a + coord\_polar("y", start=0) + ggtitle("Passive")

pie2a

q13b<- sqldf("select NPS\_Type, Age\_Range\_H, count(\*) FROM df where NPS\_Type like '%Prom%' group by NPS\_Type, Age\_Range\_H")

q13b

dfq13b <-data.frame(q13b)

colnames(dfq13b) <- c("NPS\_Type", "Age\_Range\_H","Count")

dfq13b\_perc <- (dfq13b$Count/dim(df)[1])

plotq13b<- data.frame(dfq13b$NPS\_Type,dfq13b$Age\_Range\_H,dfq13b\_perc )

colnames(plotq13b) <- c("NPS\_Type","Age\_Group", "Percentage")

pie13b<- ggplot(plotq13b, aes(x="", y=Percentage, fill=Age\_Group))+ geom\_bar(width = 1, stat = "identity")

pie13b

pie2b <- pie13b + coord\_polar("y", start=0) + ggtitle("Promoter")

pie2b

NPS\_BY\_AGE<-grid.arrange(pie2,pie2a, pie2b, ncol=3, nrow=2, top="By Age")

q9<- sqldf("select Gender\_H, count(\*) FROM df group by Gender\_H")

q9

dfq9 <-data.frame(q9)

colnames(dfq9) <- c("Gender","Count")

dfq9\_perc <- (dfq9$Count/dim(df)[1])

plotq9<- data.frame(dfq9$Gender,dfq9\_perc )

colnames(plotq9) <- c("Gender", "Percentage")

pie9<- ggplot(plotq9, aes(x="", y=Percentage, fill=Gender))+ geom\_bar(width = 1, stat = "identity")

pie9

pie9 <- pie9 + coord\_polar("y", start=0) + ggtitle("Gender")

pie9

#----------------------------

randIndex <- sample(1:dim(df)[1])

length(randIndex)

str(df)

cutpoint2\_3 <- floor(2\*dim(df)[1]/3)

cutpoint1\_3 <- floor(1\*dim(df)[1]/3)

trainData <- df[randIndex[1:cutpoint2\_3],]

dim(trainData)

testData <- df[randIndex[(cutpoint2\_3+1):dim(df)[1]],]

dim(testData)

#regression Models

model1 <- lm(formula=Likelihood\_Recommend\_H ~ Condition\_Hotel\_H, data=df)

summary(model1)

plot(df$Condition\_Hotel\_H, df$Likelihood\_Recommend\_H)

abline(model1,col="blue")

model2 <- lm(formula=Likelihood\_Recommend\_H ~ Overall\_Sat\_H, data=df)

plot(df$Overall\_Sat\_H, df$Likelihood\_Recommend\_H)

abline(model2,col="blue")

summary(model2)

model3 <- lm(formula=Likelihood\_Recommend\_H ~ Internet\_Sat\_H, data=df)

plot(df$Internet\_Sat\_H, df$Likelihood\_Recommend\_H)

abline(model3,col="blue")

summary(model3)

model4 <- lm(formula=Likelihood\_Recommend\_H ~ Staff\_Cared\_H, data=df)

plot(df$Staff\_Cared\_H, df$Likelihood\_Recommend\_H)

abline(model4,col="blue")

summary(model4)

#----------------------------

#LM Model

model <- lm(Likelihood\_Recommend\_H ~Internet\_Sat\_H +Staff\_Cared\_H,data=trainData)

lmPred <- predict(model,testData)

str(lmPred)

compTable3 <- data.frame(testData[,26],lmPred)

colnames(compTable3) <- c("test","Pred")

sqrt(mean((compTable3$test-compTable3$Pred)^2))

#lm plot

compTable3$error <- abs(compTable3$test - compTable3$Pred)

svmPlot3 <- data.frame(compTable3$error, testData$Staff\_Cared\_H, testData$Internet\_Sat\_H)

colnames(svmPlot3) <- c("error","Staff\_Cared\_H","Internet\_Sat")

ggplot(svmPlot3, aes(x=Staff\_Cared\_H,y=Internet\_Sat)) + geom\_point(aes(size=error, color=error)) + ggtitle("lm Model")

lm.plot <- ggplot(svmPlot3, aes(x=Staff\_Cared\_H,y=Internet\_Sat)) + geom\_point(aes(size=error, color=error)) + ggtitle("lm Model")

#--------------------------------------------------

#svm

svmOutput <- ksvm(Likelihood\_Recommend\_H~Internet\_Sat\_H +Staff\_Cared\_H,

data = trainData,

kernel = "rbfdot",

kpar = "automatic",

C = 10,

cross = 10,

prob.model = TRUE

)

svmOutput

svmPred <- predict(svmOutput,

testData,

type = "votes"

)

str(svmPred)

compTable <- data.frame(testData[,26], svmPred[,1])

colnames(compTable) <- c("test","Pred")

sqrt(mean((compTable$test-compTable$Pred)^2))

compTable$error <- abs(compTable$test - compTable$Pred)

svmPlot <- data.frame(compTable$error, testData$Internet\_Sat\_H, testData$Staff\_Cared\_H, testData$Likelihood\_Recommend\_H)

colnames(svmPlot) <- c("error","Internet\_Sat\_H","Staff\_Cared\_H", "Likelihood\_Recommend\_H")

plot.ksvm <- ggplot(svmPlot, aes(y=Internet\_Sat\_H,x=Staff\_Cared\_H)) +

geom\_point(aes(size=error, color=error))+

ggtitle("ksvm Model")

plot.ksvm

# Library

library(sqldf)

library(ggplot2)

library(plotly)

library(maps)

library(formattable)

#----Step 1: read in the data file

urlToRead <- "C:/Users/copla/OneDrive/Documents/my\_school/IST687/19299 Hotel ProjectSurveyData\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2\_2-1.csv"

hs <- read.csv(urlToRead,stringsAsFactors = FALSE)

str(hs)

colnames(hs)[colSums(is.na(hs))>0]

#Replace NA's with mean column

hs$LENGTH\_OF\_STAY\_C[is.na(hs$LENGTH\_OF\_STAY\_C)] <- mean(hs$LENGTH\_OF\_STAY\_C,na.rm=TRUE)

hs$NUMBER\_OF\_ROOMS\_C[is.na(hs$NUMBER\_OF\_ROOMS\_C)] <- mean(hs$NUMBER\_OF\_ROOMS\_C,na.rm=TRUE)

hs$ADULT\_NUM\_C[is.na(hs$ADULT\_NUM\_C)] <- mean(hs$ADULT\_NUM\_C,na.rm=TRUE)

hs$CHILDREN\_NUM\_C[is.na(hs$CHILDREN\_NUM\_C)] <- mean(hs$CHILDREN\_NUM\_C,na.rm=TRUE)

hs$Overall\_Sat\_H[is.na(hs$Overall\_Sat\_H)] <- mean(hs$Overall\_Sat\_H,na.rm=TRUE)

hs$Guest\_Room\_H[is.na(hs$Guest\_Room\_H)] <- mean(hs$Guest\_Room\_H,na.rm=TRUE)

hs$Tranquility\_H[is.na(hs$Tranquility\_H)] <- mean(hs$Tranquility\_H,na.rm=TRUE)

hs$Condition\_Hotel\_H[is.na(hs$Condition\_Hotel\_H)] <- mean(hs$Condition\_Hotel\_H,na.rm=TRUE)

hs$Customer\_SVC\_H[is.na(hs$Customer\_SVC\_H)] <- mean(hs$Customer\_SVC\_H,na.rm=TRUE)

hs$Staff\_Cared\_H[is.na(hs$Staff\_Cared\_H)] <- mean(hs$Staff\_Cared\_H,na.rm=TRUE)

hs$Internet\_Sat\_H[is.na(hs$Internet\_Sat\_H)] <- mean(hs$Internet\_Sat\_H,na.rm=TRUE)

hs$Check\_In\_H[is.na(hs$Check\_In\_H)] <- mean(hs$Check\_In\_H,na.rm=TRUE)

hs$F.B\_FREQ\_H[is.na(hs$F.B\_FREQ\_H)] <- mean(hs$F.B\_FREQ\_H,na.rm=TRUE)

hs$F.B\_Overall\_Experience\_H[is.na(hs$F.B\_Overall\_Experience\_H)] <- mean(hs$F.B\_Overall\_Experience\_H,na.rm=TRUE)

length(hs[hs=='na'])

#----Question 1. How many stays were related to some discount?

#

q1 <- sqldf("Select MARKET\_GROUP\_C [Discounts], count(\*) [QTY] from hs group by MARKET\_GROUP\_C ORDER BY [QTY] desc")

q1

p1 <- plot\_ly(q1, labels = ~Discounts, values = ~QTY, type = 'pie') %>%

layout(title = 'How many stays were related to some discount?',

xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),

yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))

p1

#

#----Question 3. What is the average duration of the guest stays?

#

los\_mean <- mean(hs$LENGTH\_OF\_STAY\_C)

los\_min <- min(hs$LENGTH\_OF\_STAY\_C)

los\_max <- max(hs$LENGTH\_OF\_STAY\_C)

c <- cat("LENGTH OF STAY\nmin:",los\_min, "day(s) | average:", los\_mean, "day(s) | max:", los\_max,"day(s)")

q3 <- sqldf("SELECT LENGTH\_OF\_STAY\_C [LOS], Count(LENGTH\_OF\_STAY\_C) [QTY] FROM hs GROUP BY LENGTH\_OF\_STAY\_C ORDER BY [QTY] desc")

gp <- ggplot(q3, aes(x=LOS, y=QTY)) + geom\_line()

gp <- gp + geom\_point()

gp <- gp + ggtitle("What is the average duration of the guest stays?") +

ylab("Quantity of Stays") + xlab("Length of Stay (Days)")

gp <- gp + labs(caption = "Average: 2.408054 day(s)")

gp

sdf <- sqldf("SELECT NPS\_Type, AVG(LENGTH\_OF\_STAY\_C) [avgLOS] from hs group by NPS\_Type Order by [avgLOS] desc")

#

#----Question 4. Percentage of walk ins vs. reservations?

#

q4 <- sqldf("SELECT WALK\_IN\_FLG\_C [wi\_res], count(\*) [qty] from hs group by WALK\_IN\_FLG\_C")

p4 <- plot\_ly(q4, labels = ~wi\_res, values = ~qty, type = 'pie') %>%

layout(title = 'WALK-IN?',

xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),

yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))

p4

#

#----Question 7. What country has the most revenue?

#

q7 <- sqldf("SELECT COUNTRY\_PL [Country], FLOOR(sum(REVENUE\_USD\_R)) [Rev\_USD] from hs group by COUNTRY\_PL order by [Rev\_USD] desc")

q7

cc <- c('US','HK','GB','CN','IN','CA','KR','EG','CH','QA','DE','JP','SA','UA','MX','TR','TJ','ID')

q7$Code <- cc

world\_map<-map\_data("world")

world\_map <- subset(world\_map, region!="Antarctica")

head(world\_map)

str(q7)

q7\_mod <- gsub("United States","USA",q7$Country)

q7\_mod <- gsub("United Kingdom","UK",q7\_mod)

q7$mod\_country <- q7\_mod

q7$Rev\_USD <- as.numeric(q7$Rev\_USD)

#http://colorbrewer2.org

#return currency in millions

#q7$Rev\_USD <- currency(q7$Rev\_USD, digits = 0L)

q7$mod\_rev <- round(q7$Rev\_USD)/1000000

#create the map

gg <- ggplot(q7)

gg <- gg + geom\_map(dat=world\_map, map = world\_map, aes(map\_id=region),

fill="white", color="#7f7f7f", size=0.25)

gg <- gg + geom\_map(map = world\_map, aes(map\_id = mod\_country, fill = mod\_rev), size=0.25)

gg <- gg + scale\_fill\_gradient(low="#e5f5e0", high="#31a354", name="Revenue (in MIL)")

gg <- gg + expand\_limits(x = world\_map$long, y = world\_map$lat)

gg <- gg + labs(x="", y="", title="Revenue")

gg <- gg + theme(panel.grid=element\_blank(), panel.border=element\_blank())

gg <- gg + theme(axis.ticks=element\_blank(), axis.text=element\_blank())

gg <- gg + theme(legend.position="right")

gg

barplot(q7$mod\_rev, main="Revenue", horiz=FALSE,

names.arg=q7$Code, ylab="USD (in MIL)", xlab="Country Code")

#

#----Question 19. Which relationship category have the highest average of likelihood guest recommendation?

#

q19 <- sqldf("SELECT RELATIONSHIP\_PL [relationship], avg(LIKELIHOOD\_RECOMMEND\_H) [avg\_lr] from hs group by RELATIONSHIP\_PL order by [avg\_lr] desc")

q19

barplot(q19$avg\_lr, main="Relationship Category Most Likely To Be Recommend\n(Highest Score is 10)", horiz=FALSE,

names.arg=q19$relationship, ylab="Average Score", xlab="",ylim=c(8,9))

#