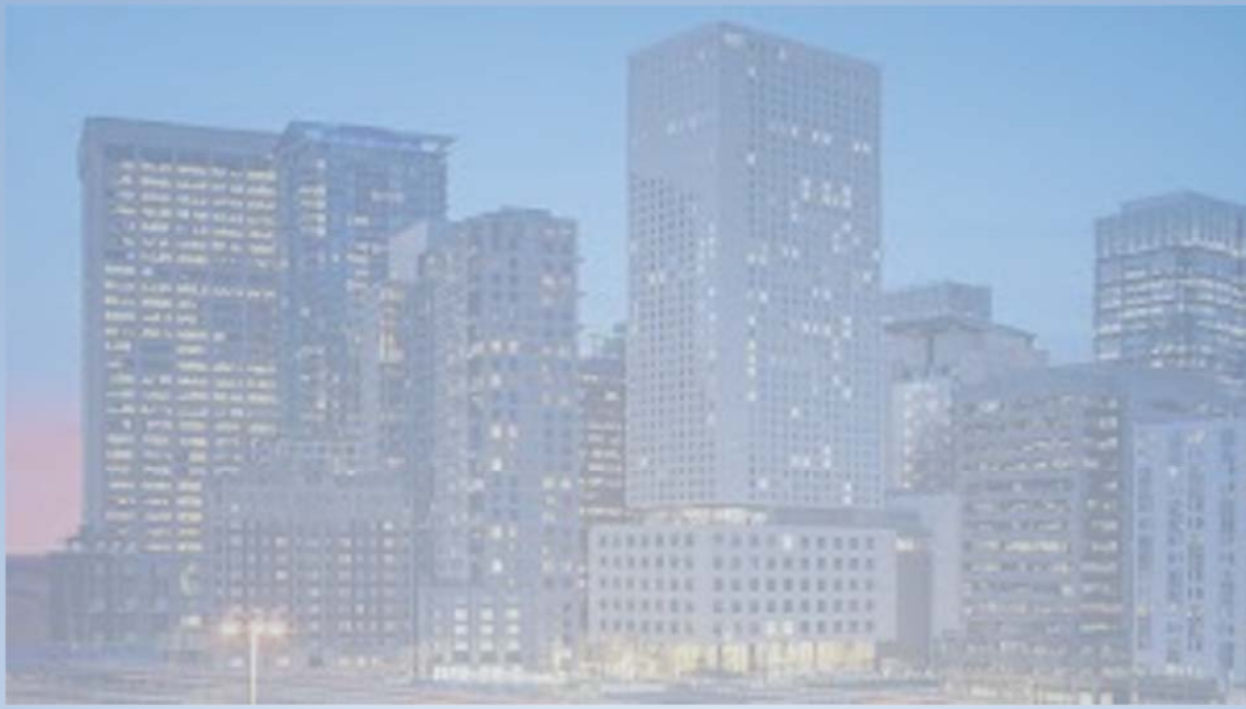


IST 687
Hyatt Hotel Data Analysis

TEAM 4

Elizabeth Schaub
Daniel Pearl
Travis Tran
Laura Lamoureux

Table of Contents

Introduction 2
Problem 2
Key Conclusions 2
Methodology 2
Initial Observations 2
Discussion 5
Unexecuted R Code 33

Introduction

Data science, at its core, can be described as extracting valuable information from data and transforming it into knowledge that can be used to make recommendations and drive decisions. This group project attempts to do just that.

Problem

Learn and explain something of value out of a data set. We chose to use the Hyatt Hotel data set provided to the class. Our group wanted to know if we could determine if any single variable or could have an effect on the Net Promoter Score. We chose many variables to analyze.

Key Conclusions

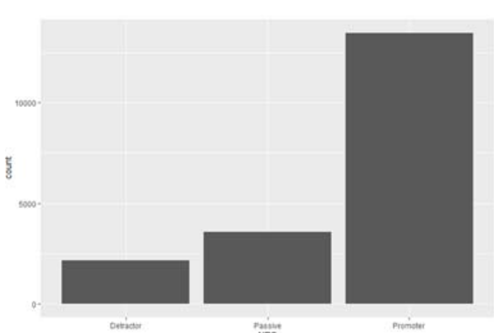
- A combination of Customer Service, Guest Room Satisfaction, Condition of Hotel, Staff Cared and Tranquility explains 65% of Likelihood to Recommend, which is directly related to Net Promoter Score.
- Being older and female leads to a great chance of being a Promoter.

Methodology

Our first step was to use the Variable Glossary document to help determine what variables we wanted to work with. We hypothesized 21 questions that could potentially help us to explain Net Promoter Score. We evenly divided the 21 questions among the team members so that each of us had 5-6 questions to answer. Eventually, we ended up adding one more question that we had not originally envisioned, for a total of 22 total attempts.

Initial Observations

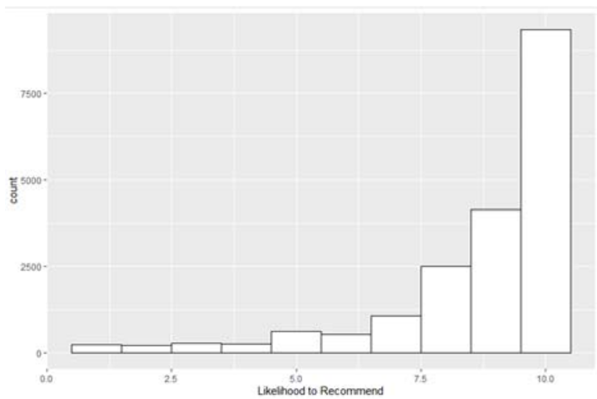
1. Looking at the distribution of Net Promoter Scores, we see that most hotel guests are classified as “Promoters”:



Net Promoter Score

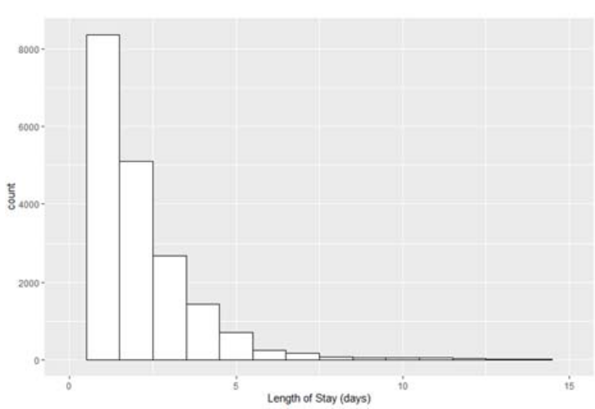
Detractor	Passive	Promoter
2153	3580	13486

2. Looking at the distribution of Likelihood to Recommend scores, we see, unsurprisingly, that most hotel guests would be highly likely to recommend the hotel:



Likelihood to Recommend
mean: 8.700193
median: 9
min: 1
max: 10
sd: 1.924673
quantile: 4 10
skewness: -2.02118

3. Looking at the distribution of Length of Stay, we see that most hotel guests stay for less than 5 days, with the majority staying for 1 day:



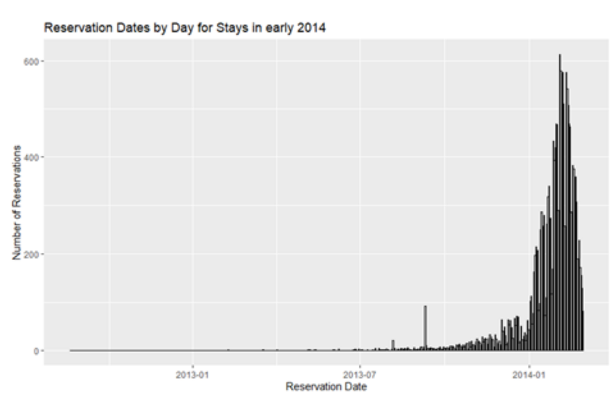
Length of Stay
mean: 2.404808
median: 2
min: 1
max: 97
sd: 2.992155
quantile: 1 5
skewness: 9.593808

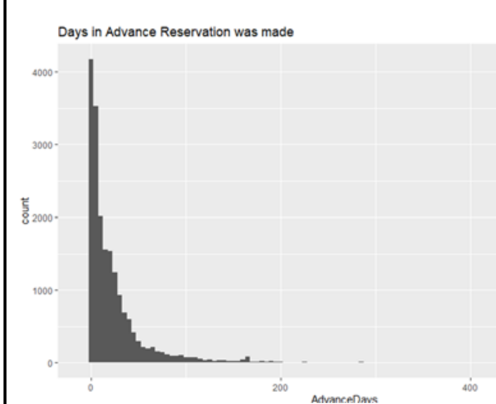
4. Looking at Reservation Dates, we learn that most reservations in this data set are from 2014, and most of those are from the month of February:

Year	Number of Stays
2013	1
2014	19218

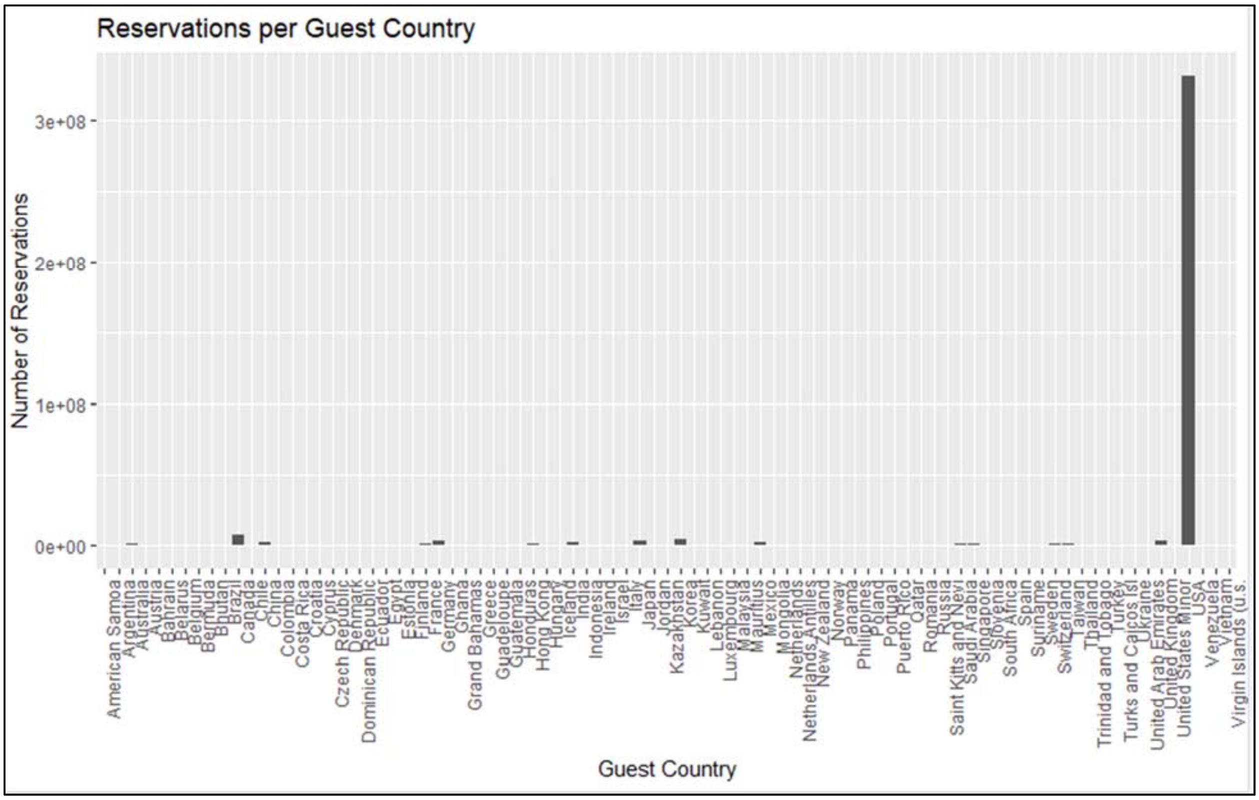
Month	Number of Stays
Jan	6
Feb	19192
Mar	21

5. A look at reservation dates and comparison to subsequent check in dates reveals that most reservations are made 1-3 months prior to check in.





6. A look at the distribution of Guest Country of Origin shows us that most hotel reservations in this data set were made by residents of the USA:



Discussion

Unfortunately, the Variable Glossary did not match the .csv file, so as we worked through data cleaning, we each realized that the actual data set did not contain data for one or more of the questions we had set out to answer.

Analysis Notes:

- *In many cases, Likelihood to Recommend (a numeric, continuous variable) was used as a proxy for NPS, since NPS is a discrete variable and couldn't be directly used in some of the models we wanted to create.*
- *It is noted here that while for some of the analyses in this report, all available data was used, for some specific questions where it made more sense (i.e. map visualization), we used only data from the US, where the bulk of the data was available from.*

Questions we posed for ourselves, and the results we were able to obtain:

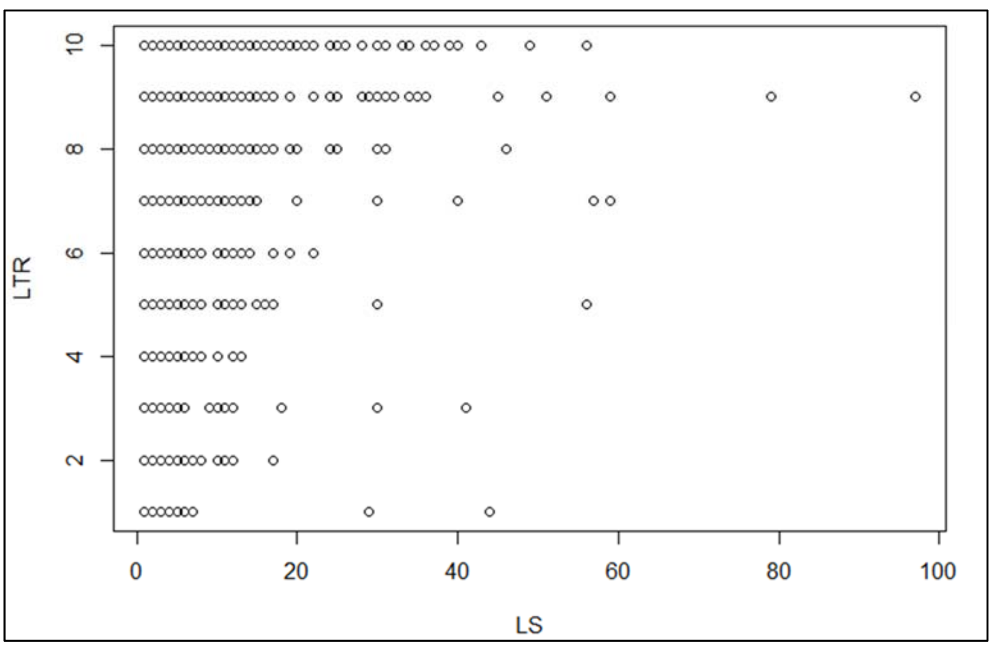
1. Does length of stay impact NPS?

We first tried fitting a line to actual Likelihood to Recommend (LTR) scores, which resulted in a scatter plot with groups of points clustered around each of the 10 scale options.

==

```
# plot NPS vs Length of Stay
plot(LS, LTR)
```

==



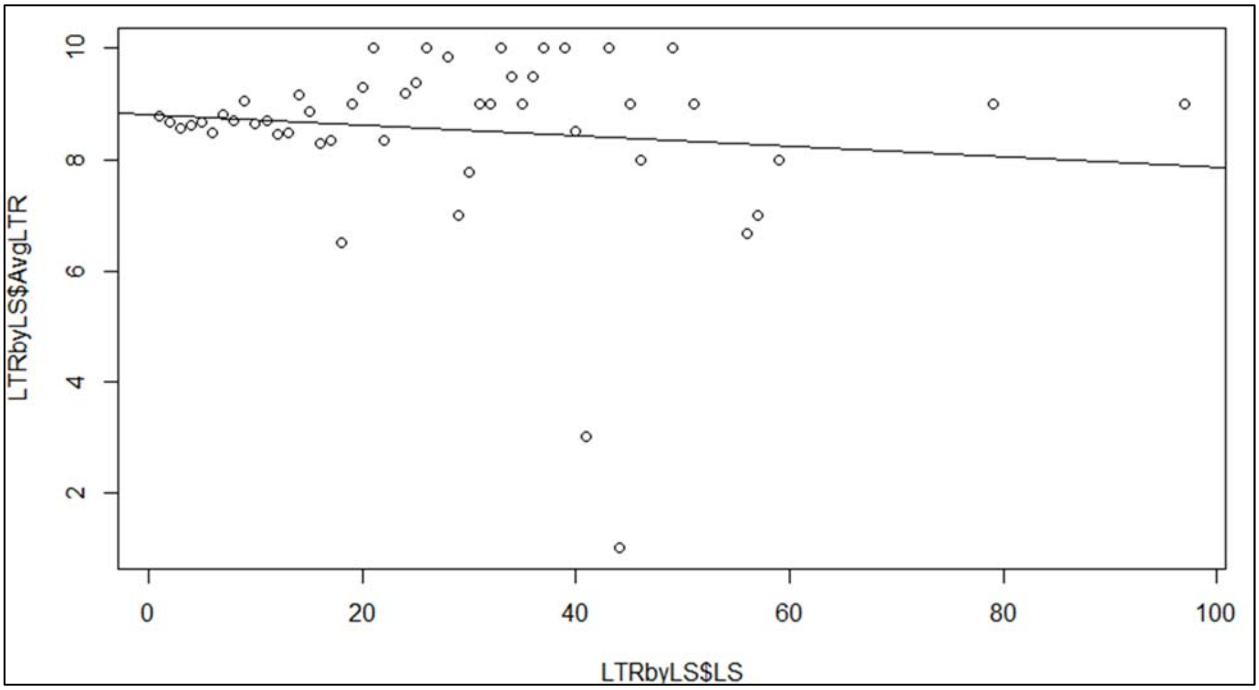
So instead, an Average LTR score for each Length of Stay (days) was calculated and then used in a linear regression model.

==

```
# plot AvgLTR by LS
plot(LTRbyLS$LS, LTRbyLS$AvgLTR)

# build a linear model
model <- lm(formula=AvgLTR ~ LS, LTRbyLS)
summary(model)
abline(model)
```

==

The line plotting AvgLikelihood to Recommend by Length of Stay(days) has an adjusted R^2 value of -0.007.



Conclusion: Due to the insignificant R^2 value, we conclude that Length of Stay does NOT significantly impact NPS.

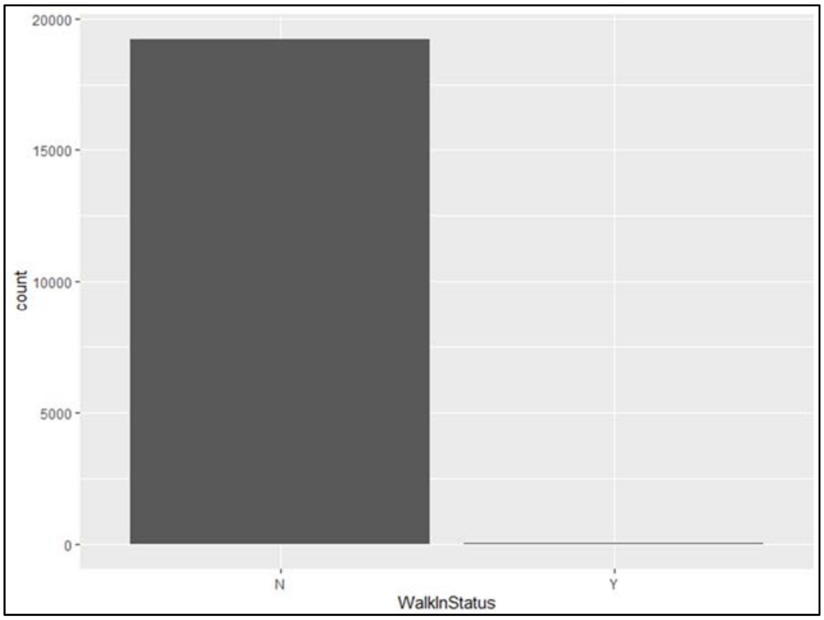
2(a). ***<u>Does whether guest stay was a walk-in or a reservation impact NPS?</u>***

==

```
WalkInStatus_Breakdown <- tapply(WalkInStatus, WalkInStatus, length)
WalkInStatus_Breakdown
bar_WalkInStatus <- ggplot(HotelData, aes(x=WalkInStatus)) + geom_bar()
bar_WalkInStatus
```

==

WalkIns	13
Reservations	19206



Conclusion: No conclusion could be drawn, because the base size of walk-ins is not big enough to draw conclusions about the impact of WalkInStatus on NPS**.**

2(b). <u>Is NPS affected by how far in advance the reservation was made?</u>

The variable Advance Days was created by subtracting Reservation Date from Check-In Date.

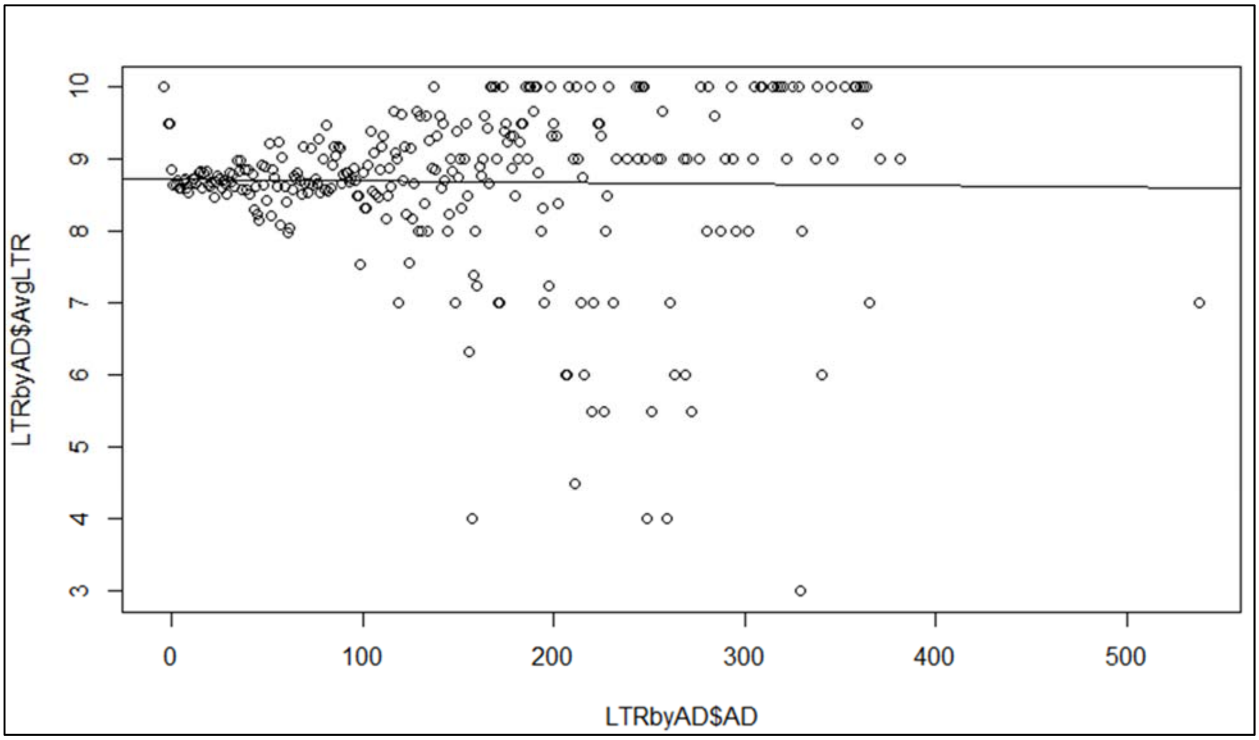
For the same reason given in #1 above, Average LTR was used to create the linear regression model to determine if Number of Days in Advance the reservation was made could help to explain LTR.

===

```
# how far in advance a reservation was made
AdvanceDays <- CheckInDate - ReserveDate
# make AdvanceDays numeric
AdvanceDays <- as.numeric(AdvanceDays)

# plot AvgLTR by AD
plot(LTRbyAD$AD, LTRbyAD$AvgLTR)

# build a linear model
model <- lm(formula=LTRbyAD$AvgLTR ~ LTRbyAD$AD, LTRbyAD)
summary(model)
abline(model)
```

===

The line plotting AvgLikelihood to Recommend by Advance Days has an adjusted R^2 value of -0.003:



Conclusion: Due to the insignificant R^2 value, we conclude that number of days reservation was made in advance of stay does NOT significantly impact NPS.

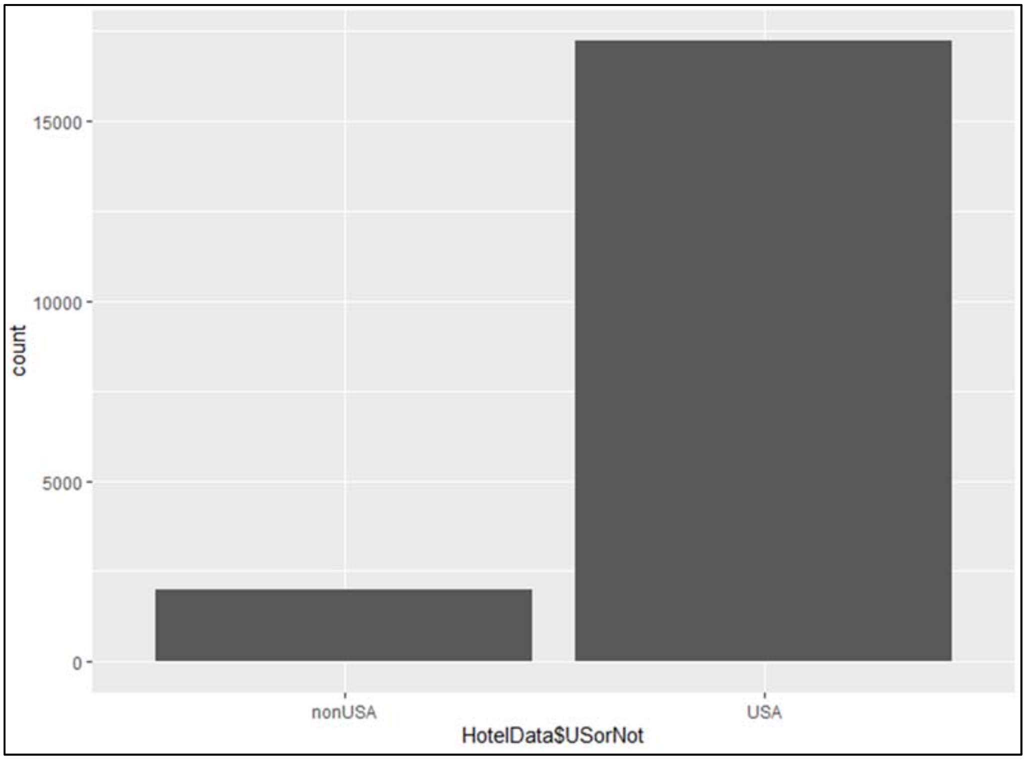
3(a). <u>Does guest country of origin impact NPS?</u>

Since most hotel guests were from the US, it was not possible to answer this question as originally envisioned.

We did, however, compare US vs. non-US guests' average ratings, and found them to be identical:

```
===========================================================================
USA_stays <- length(HotelData$GuestCountry[HotelData$GuestCountry=="USA"])
USA_stays
nonUSA_stays <- length(HotelData$GuestCountry[!(HotelData$GuestCountry=="USA")])
nonUSA_stays

HotelData$USorNot <- ifelse((HotelData$GuestCountry=="USA"), "USA", "nonUSA")
HotelData$USorNot
attach(HotelData)

# create a bar chart of Guest Country
gg_bar <- ggplot(HotelData, aes(x=HotelData$USorNot)) + geom_bar()
gg_bar
# calculate AvgLTR by whether country was US or Not
LTRbyUSorNot <- sqldf("SELECT AVG(LTR) AS AvgLTR,
    USorNot
```

```
    FROM HotelData
    GROUP BY USorNot")
LTRbyUSorNot <- data.frame(LTRbyUSorNot)
LTRbyUSorNot
```
==



Country of Origin		**avgLTR**
US	17220	8.7
nonUS	1998	8.7

Conclusion: Whether guest country of origin was US or not does NOT significantly impact NPS.

3(b). ***<u>How does country of origin impact NPS in countries other than the guest's country of origin?</u>***

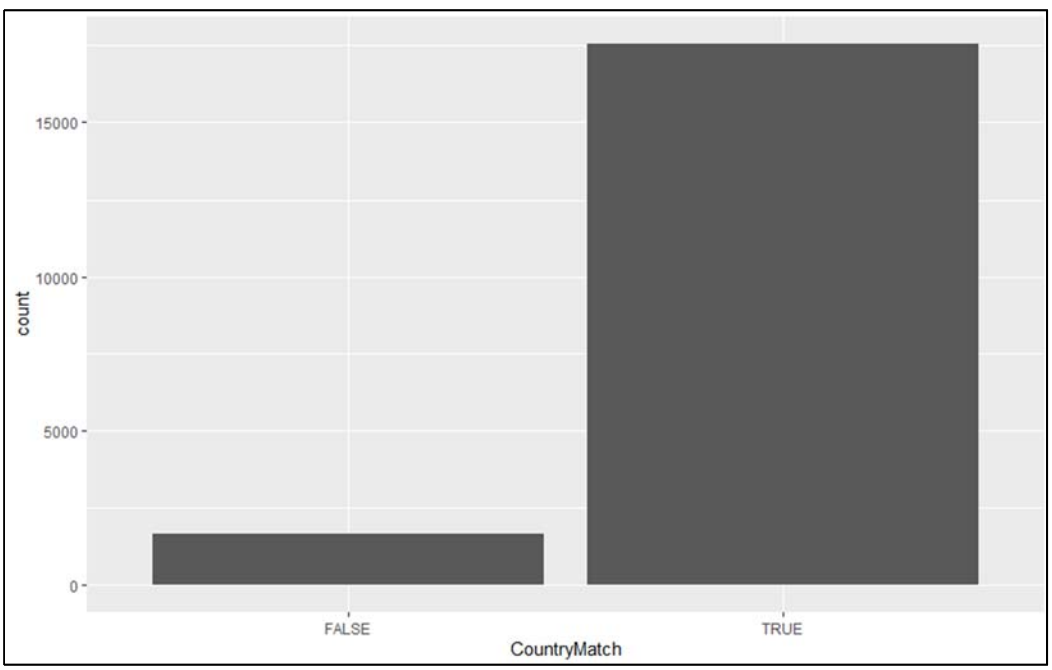
==

```
# CountryMatch
# create a new column called CountryMatch representing
# whether hotel guest stayed in is in their country
# of origin (Y) or not (N)

x <- c(1:length(HotelData[,1]))
head(x)

Compare <- function(x) {
```

```
  result <- GuestCountry[x]==HotelCountry[x]
  return(result)
  }

HotelData$CountryMatch <- Compare(x)
head(HotelData$CountryMatch)
length(HotelData$CountryMatch)

# create a bar chart of Country Match, depicting # of stays where hotel country was in the guest's
# country of origin or not

gg_bar <- ggplot(HotelData, aes(x=CountryMatch)) + geom_bar()
gg_bar

# calculate AvgLTR by whether guest stayed in home country or not
LTRbyCM <- sqldf("SELECT AVG(LTR) AS AvgLTR,
              CountryMatch AS CM
              FROM HotelData
              GROUP BY CountryMatch")
LTRbyCM <- data.frame(LTRbyCM)
LTRbyCM
```

==

We first checked to see how often the guest's country of origin was different from the country the hotel was in:



Since most guests in this data set were from the US and had stayed in a US hotel, we simply calculated means for each of these groups, and found them to be almost identical:

Guest Country = Hotel Country	**AvgLTR**
Yes	8.7
No	8.8

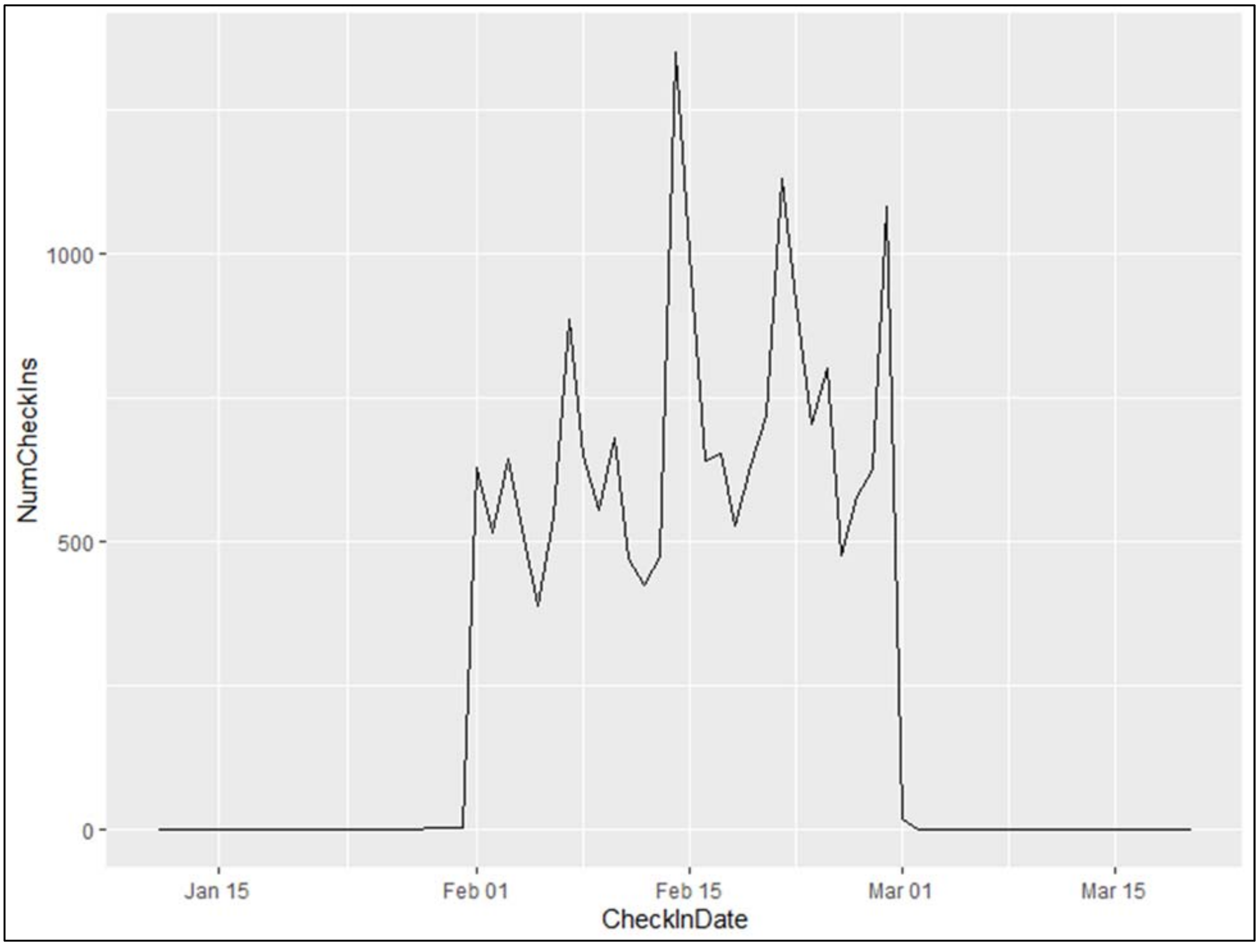
Conclusion: NPS is not impacted by whether a guest stays in a hotel in their own country or not.

4. <u>During which time is NPS the highest?</u>

Conclusion: Recalling the earlier data shown in our Initial Observations, since we found that the majority of stays were during February 2014, we could not answer this question based on seasonality or month of the year.

That said, we did look at Check-Ins Per day in JFM 2014:

```
==========================================================================
# determine the number of record(s) with 2013 & 2014 CheckInDates
# count rows
Year1 <- HotelData[HotelData$CIYear==2013,]
length(Year1[,1])
Year2<- HotelData[!(HotelData$CIYear==2013),]
length(Year2[,1])
# remove the 1 record with the 2013 CheckInDate from the dataset
HotelData <- Year2
# line plot of CheckIn dates
CheckIns <- sqldf("SELECT CheckInDate, COUNT(CheckInDate) AS NumCheckIns FROM
HotelData GROUP BY CheckInDate")
dfCheckIns <- data.frame(CheckIns)
dfCheckIns

line_CheckIn <- ggplot(dfCheckIns,aes(x=CheckInDate, y=NumCheckIns)) + geom_line()
line_CheckIn
==========================================================================
```



* We note a spike in Check-Ins on February 14 (Valentine's Day)

5. <u>How does whether travel is free independent travel vs. group travel impact NPS?</u>
Conclusion: No conclusion could be drawn, because the data did not exist for this variable in the data set.

Added Question:
<u>Which survey data questions might help us to understand and predict NPS?</u>
We created a multiple regression model that included each of the survey variables:

```
Call:
lm(formula = LTR ~ GuestRoom + Tranquility + Condition + CustServ +
    Staff + Internet + CIProcess, data = HotelData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3326 -0.1408  0.0508  0.4214  4.7935

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.926782   0.096290 -20.010   <2e-16 ***
GuestRoom    0.308893   0.012805  24.123   <2e-16 ***
Tranquility  0.130670   0.008720  14.985   <2e-16 ***
Condition    0.193097   0.013479  14.325   <2e-16 ***
CustServ     0.377980   0.016344  23.127   <2e-16 ***
Staff        0.149685   0.013660  10.958   <2e-16 ***
Internet     0.008682   0.006277   1.383   0.1667    
CIProcess    0.018587   0.010469   1.775   0.0759 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.012 on 8234 degrees of freedom
Multiple R-squared:  0.6535,  Adjusted R-squared:  0.6532 
F-statistic:  2219 on 7 and 8234 DF,  p-value: < 2.2e-16
==============================================================
```

The adjusted R^2 value was 0.65, and 5 of the inputs were statistically significant.

Then we created a model, stepwise, adding in each significant variable until we reached an equivalent adjusted R^2 value using only the variables that were significant in the original model, to achieve an equivalent R^2 value:

```
==============================================================
Call:
lm(formula = LTR ~ CustServ + GuestRoom + Condition + Staff + 
    Tranquility, data = HotelData)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3455 -0.1480  0.0601  0.4162  4.8189 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.840446   0.088688  -20.75   <2e-16 ***
CustServ     0.386171   0.015838   24.38   <2e-16 ***
GuestRoom    0.309030   0.012777   24.19   <2e-16 ***
Condition    0.197497   0.013324   14.82   <2e-16 ***
Staff        0.153901   0.013534   11.37   <2e-16 ***
Tranquility  0.131433   0.008713   15.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.012 on 8236 degrees of freedom
Multiple R-squared:  0.6533,  Adjusted R-squared:  0.6531 
F-statistic:  3104 on 5 and 8236 DF,  p-value: < 2.2e-16

==============================================================
```

Conclusion: We found that a combination of Customer Service, Guest Room Satisfaction, Condition of Hotel, Staff Cared and Tranquility explains 65% of Likelihood to Recommend, which is directly related to Net Promoter Score**.**

Likelihood to Recommend = -1.84 + 0.39(Customer Service) + 0.31(Guest Room)
+ 0.20(Hotel Condition) + 0.15(Staff Cared) + 0.13(Tranquility)

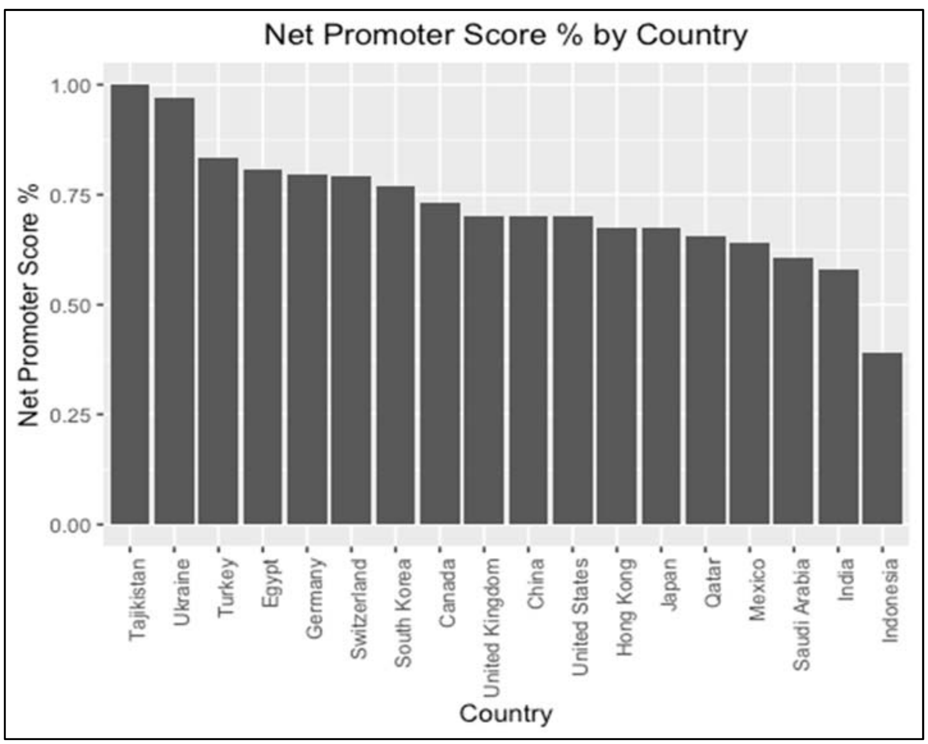
Note: Survey attributes, considered but not included in the final model, since they do not increase its predictive ability are: Internet Satisfaction and Quality of CheckIn Process.

6. Which regions have the highest NPS?

This question is important for determining if there are certain cultural or regional influences on the NPS. The regional data frame was queried and transformed to create a visual with the highest NPS. The percentages of guest's scores in these regions were used. If drastic differences can be seen, it may be necessary to explore cultural differences that contribute to their success. Perhaps there are certain contributors to NPS that can be replicated to other hotels. Perhaps it is exclusively the culture rather than the hotels that contribute to these high scores.

The following code was used to generate the region visualizations:

```
nps.region <- tapply(NPS_Type, list(Region_PL, NPS_Type), length) # Query Data
top_region <- get_percentages(nps.region) # Generate Percentage dataframe
generate_bar_graph(top_region, rownames(top_region), top_region$data, "Region", "Net
Promoter Score %" ) # Generate bar graph
```



Conclusion: European hotels have the greatest NPS percentage while Asia Pacific have the lowest. The difference is minimal but perhaps there are contributing factors. Further analysis would have to be conducted to determine what cultural factors are contributing to the differences in these scores.

7. Which hotels have the lowest NPS Percentage?

If the worst hotels are identified, necessary changes can be made to increase the overall NPS. A data frame was created to identify the hotels with the lowest NPS. The percentages of guests scores in these hotels were used. A histogram was also created to determine the distribution of NPS percentages

The following helper functions were created to assist in 2 tasks:

1) # Generate Bar Graph

```
generate_bar_graph <- function(df, x, y, x_label, y_label){
	# Create bar graph
	g <- ggplot(df, aes(x=reorder(x, -y), y=y)) + geom_bar(stat="identity")
	title <- paste(y_label, "by", x_label, sep = " ")
	g <- g + ggtitle(title) + theme(plot.title = element_text(hjust=0.5))
	g <- g + xlab(x_label) + ylab(y_label) +  theme(axis.text.x = element_text(angle = 90, 
hjust = 1))
	return(g)
}
```

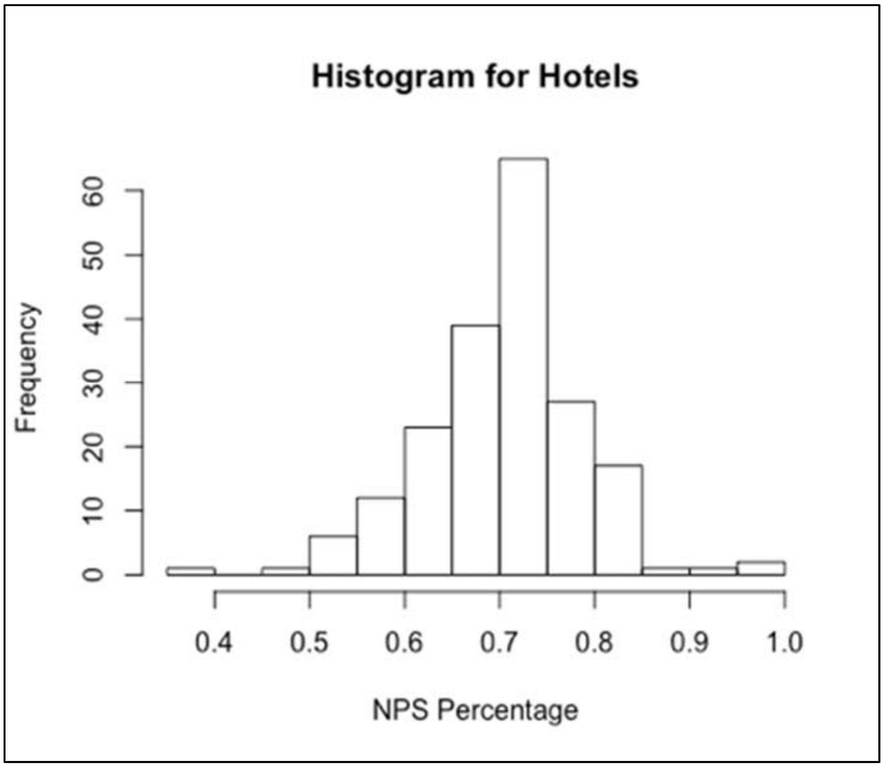
2) # Clean and return dataframe of percentages

```
get_percentages <- function(data){
	data[is.na(data)] <- 0
	data <- round(data[,"Promoter"] / (data[,"Promoter"] + data[,"Detractor"] + 
data[,"Passive"]),3)
	# Get Percentages
	data <- data[order(-data)] # Order data by descending
	data <- data.frame(data) # Sadatae as dataframe
	return(data)
}
```

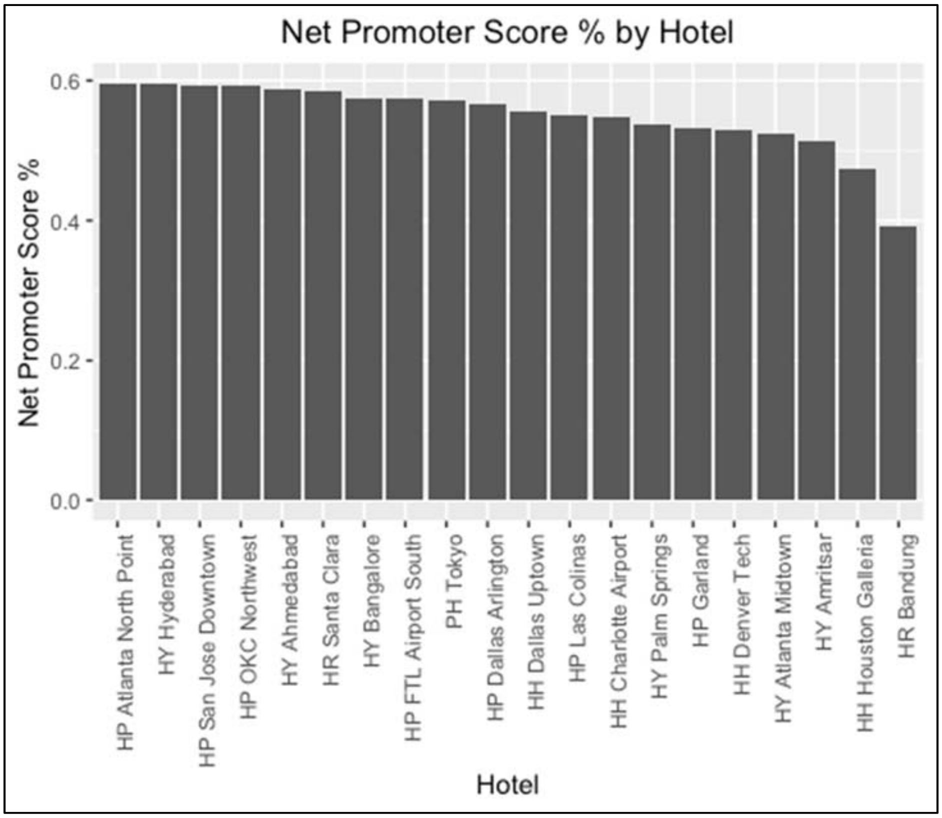
The following code was used to generate the hotel visualizations:

```
nps.hotel <- tapply(NPS_Type, list(Hotel.Name.Short_PL, NPS_Type), length) # Query Data
nps_hotel <- get_percentages(nps.hotel) # Generate Percentage dataframe
nps_hotel.hist <- hist(nps_hotel$data, main="Histogram for Hotels", xlab="NPS Percentage")
nps_hotel.hist
summary(nps_hotel)
worst_hotel <- tail(nps_hotel, 20)
best_hotel <- head(nps_hotel, 20)
```

generate_bar_graph(worst_hotel, rownames(worst_hotel), worst_hotel$data, "Hotel", "Net Promoter Score %") # Generate bar graph
generate_bar_graph(best_hotel, rownames(best_hotel), best_hotel$data, "Hotel", "Net Promoter Score %") # Generate bar graph



Min.:	0.3910
1st Qu.:	0.6610
Median:	0.7110
Mean:	0.7051
3rd Qu.:	0.7495
Max.:	1.0000



Conclusion: These low scoring hotels are compared to the top hotels which have an NPS that go as high as 100%. More successful hotels can be emulated and used as a model for these low scoring hotels. Further analysis will have to be conducted to determine what sets these hotels apart from the rest. Hopefully after identifying various contributors to NPS scores we will see the histogram distribution shift. This shift should also show a change in the measures of central tendency and distribution variation

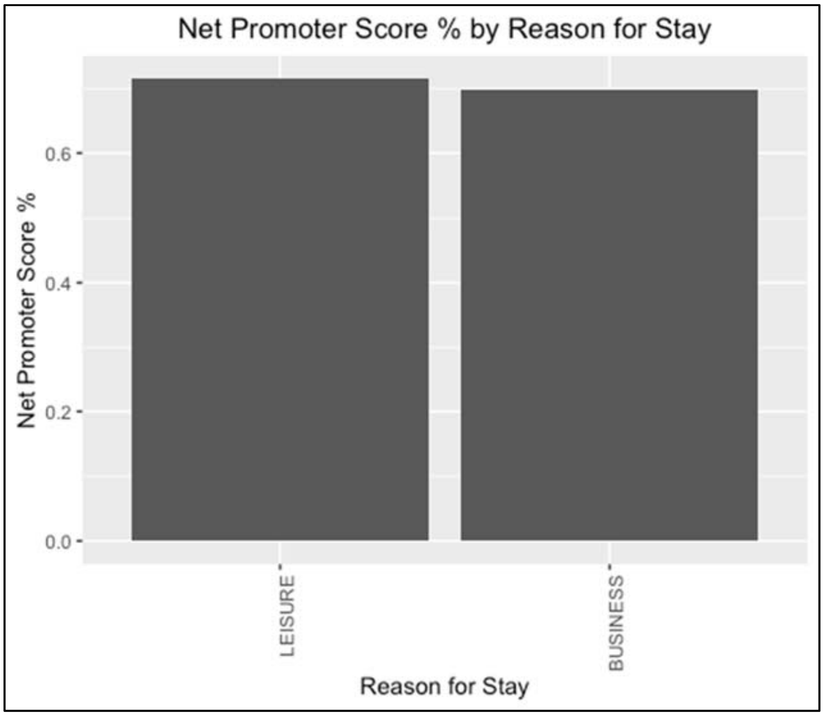
10. <u>How does reason for stay impact the NPS?</u>

The goal is to determine whether NPS is affected by guests staying for business or leisure. A data frame was created to separate the reason for stay. The percentages of guest's scores were used. A regression model was also created to determine if reason for stay impacted the likelihood to recommend score**.**

The following code was used to generate the reason for stay visualizations:

```
nps.stay <- tapply(NPS_Type, list(POV_CODE_C, NPS_Type), length) # Query data
top_stay <- get_percentages(nps.stay) # Generate Percentage dataframe
generate_bar_graph(top_stay, rownames(top_stay), top_stay$data, "Reason for Stay", "Net Promoter Score %" ) # Generate bar graph

rstay <- sqldf("Select POV_CODE_C as stay,  Likelihood_Recommend_H as score from hotel_data")
rmodel <- lm(score ~ stay, data = rstay) # Run linear model
summary(rmodel)
```

Reason for stay dataframe:
LEISURE 0.715
BUSINESS 0.698




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.69186    0.01537 565.686   <2e-16 ***
stayLEISURE  0.04149    0.03530   1.175     0.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.924 on 19340 degrees of freedom
Multiple R-squared:  7.141e-05,	Adjusted R-squared:  1.971e-05
F-statistic: 1.381 on 1 and 19340 DF,  p-value: 0.2399
```

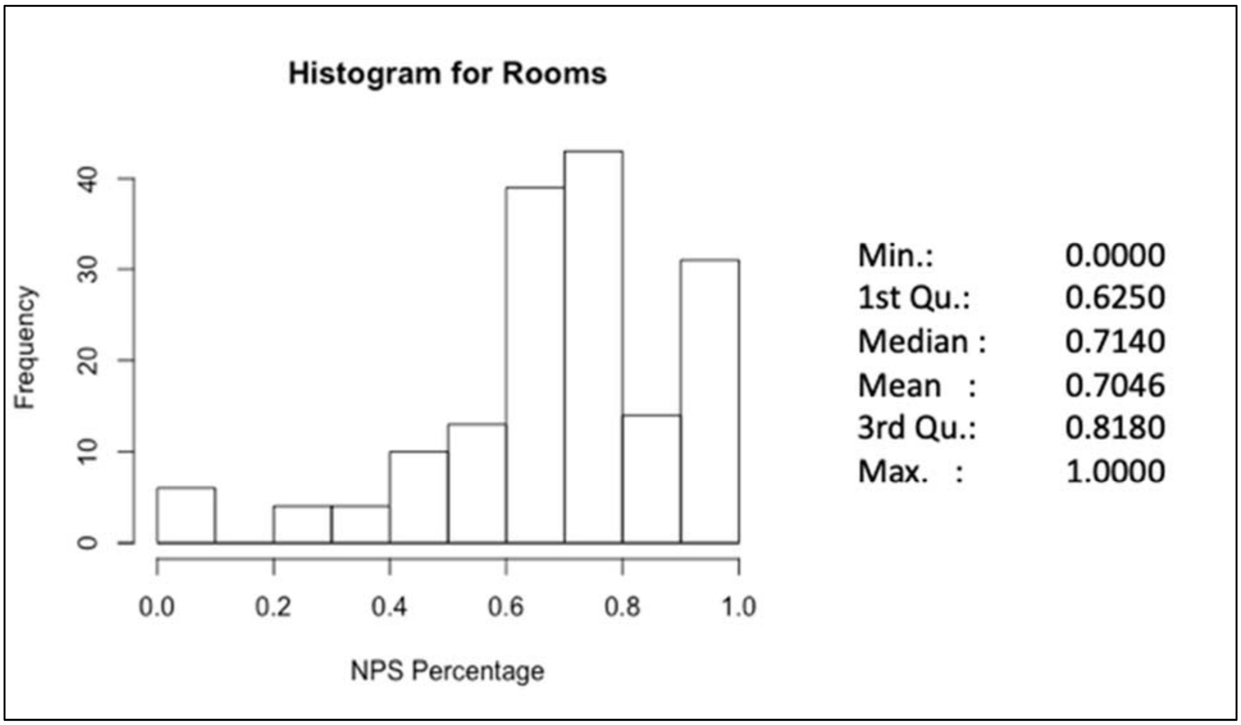
Conclusion: The difference in percentage between guest staying for business and leisure is 0.017. This is a very minimal difference between these two groups. The p-value for the regression is 0.2399 and the adjusted r-squared values is 1.971e-05. These values indicate that there is not a significant impact on NPS.

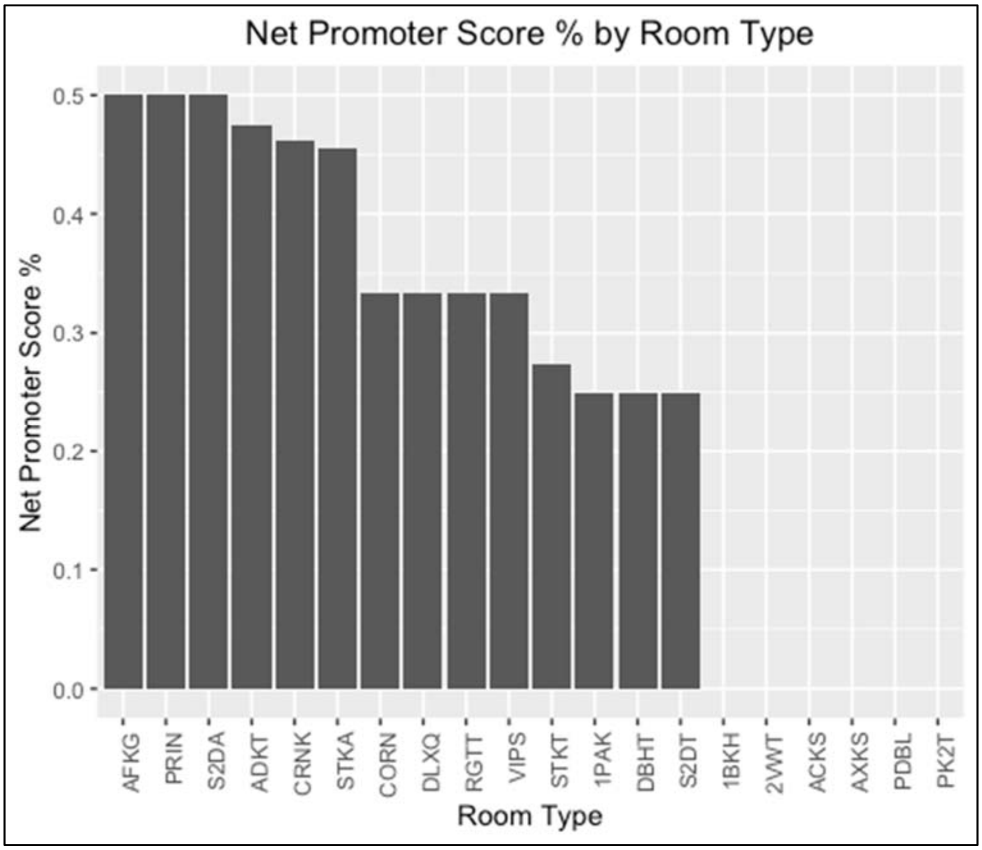
11. __Which rooms have the lowest net promoter score?__

Like the individual hotel analysis, we queried the dataset to find the lowest scoring hotel room categories. A data frame was created to identify the rooms with the lowest NPS. The percentages of guest's scores were used. Hopefully after identifying various contributors to NPS scores we will see the histogram distribution shift. This shift should also show a change in the measures of central tendency and distribution variation.

The following code was used to generate the reason for stay visualizations:

```
nps.room <- tapply(NPS_Type, list(ROOM_TYPE_CODE_C, NPS_Type), length) # Query Data
nps.room <- get_percentages(nps.room) # Generate Percentage dataframe
best_rooms <- head(nps.room, 20)
worst_room <- tail(nps.room, 20) # Get lowest pecentages from dataframe
generate_bar_graph(worst_room, rownames(worst_room), worst_room$data, "Room Type", "Net Promoter Score %" ) # Generate bar graph

nps.room.hist <- hist(nps.room$data, main="Histogram for Rooms", xlab="NPS Percentage")
summary(nps.room)
```





Conclusion: The top rooms have a 100% NPS while the worst rated rooms all score below 50%. This is a pretty notable difference. Perhaps the lowest ranking rooms can be altered and made nicer to raise the overall NPS. Further analysis will have to be conducted to determine the difference between the high and low rated rooms. Hopefully after identifying various contributors to NPS scores we will see the histogram distribution shift. This shift should also show a change in the measures of central tendency and distribution variation.

Added Question:
<u>How does Award impact NPS?</u>
A data frame was created using the award score and the corresponding NPS. A linear regression model was run to determine if these two variables are connected.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.641562   0.025342 340.991  < 2e-16 ***
df$award    0.026114   0.009535   2.739  0.00617 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.924 on 19340 degrees of freedom
Multiple R-squared:  0.0003877,	Adjusted R-squared:  0.000336
```

F-statistic: 7.501 on 1 and 19340 DF, p-value: 0.006172

Conclusion: Though the p-value is low, the adjusted r-squared value has determined that the award score does not explain the NPS.

Added Question:

Does the difference between expected and actual costs impact the likelihood to recommend?

A SQL query was run to return a dataframe containing 1) the difference between the actual and 2) expected cost and the NPS. A linear regression model was run to determine if these two variables are connected.

The following code was used to generate the difference in cost visualizations:

```
df <- sqldf("Select round(abs((REVENUE_USD_R - (QUOTED_RATE_C *
LENGTH_OF_STAY_C ))),2) as cost_diff,  Likelihood_Recommend_H as score from
hotel_data") # Get the difference between actual and expected cost
rmodel <- lm(df$score ~ df$cost_diff, data = df) # Run linear model
summary(rmodel) # Summarize model
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.700e+00  1.390e-02 626.071   <2e-16 ***
df$cost_diff -1.457e-08  1.079e-07  -0.135    0.893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.923 on 19243 degrees of freedom
  (97 observations deleted due to missingness)
Multiple R-squared:  9.486e-07,	Adjusted R-squared:  -5.102e-05
F-statistic: 0.01825 on 1 and 19243 DF,  p-value: 0.8925
```

Conclusion: Both the p-value and the adjusted r-squared value are too low to indicate a relationship between cost difference and NPS.

12. Does the room rate the guest paid stayed impact the NPS?

Conclusion: No conclusion could be drawn, because the data did not exist for this variable in the data set.

13. Does size of hotel (number of rooms &/or number of floors) impact the NPS?

Conclusion: No conclusion could be drawn, because the data did not exist for this variable in the data set.

14. Does whether or not the guest was offered a promotion impact the NPS? Is either past or future offer more impactful?

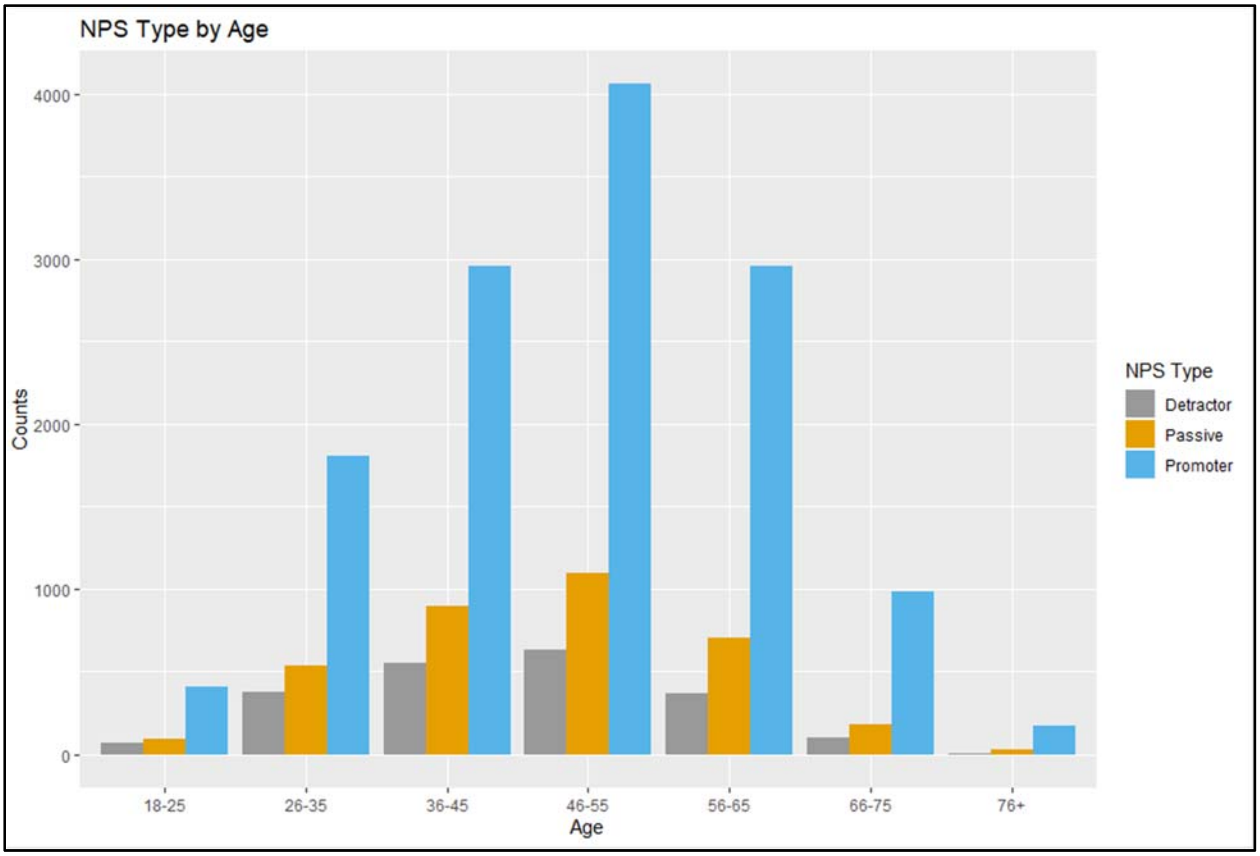
Conclusion: No conclusion could be drawn, because the data did not exist for this variable in the data set.

15. Which age groups give the highest NPS?

R was used to determine how the frequency for each age group by NPS type.

```
> npsAge <- npsAge[-c(1),]
> npsAge

        Detractor Passive Promoter
  18-25        65      90      405
  26-35       371     535     1808
  36-45       552     891     2962
  46-55       629    1094     4060
  56-65       365     700     2962
  66-75       100     180      982
  76+           7      25      173
```

	NPS TYPE						
AGE	**Detractor**	**% Detractor**	**Passive**	**% Passive**	**Promoter**	**% Promoter**	**Total by Gender**
18-25	65	12%	90	16%	405	72%	560
26-35	371	14%	535	20%	1808	67%	2714
36-45	552	13%	891	20%	2962	67%	4405
46-55	629	11%	1094	19%	4060	70%	5783
56-65	365	9%	700	17%	2962	74%	4027
66-75	100	8%	180	14%	982	78%	1262
76+	7	3%	25	12%	173	84%	205



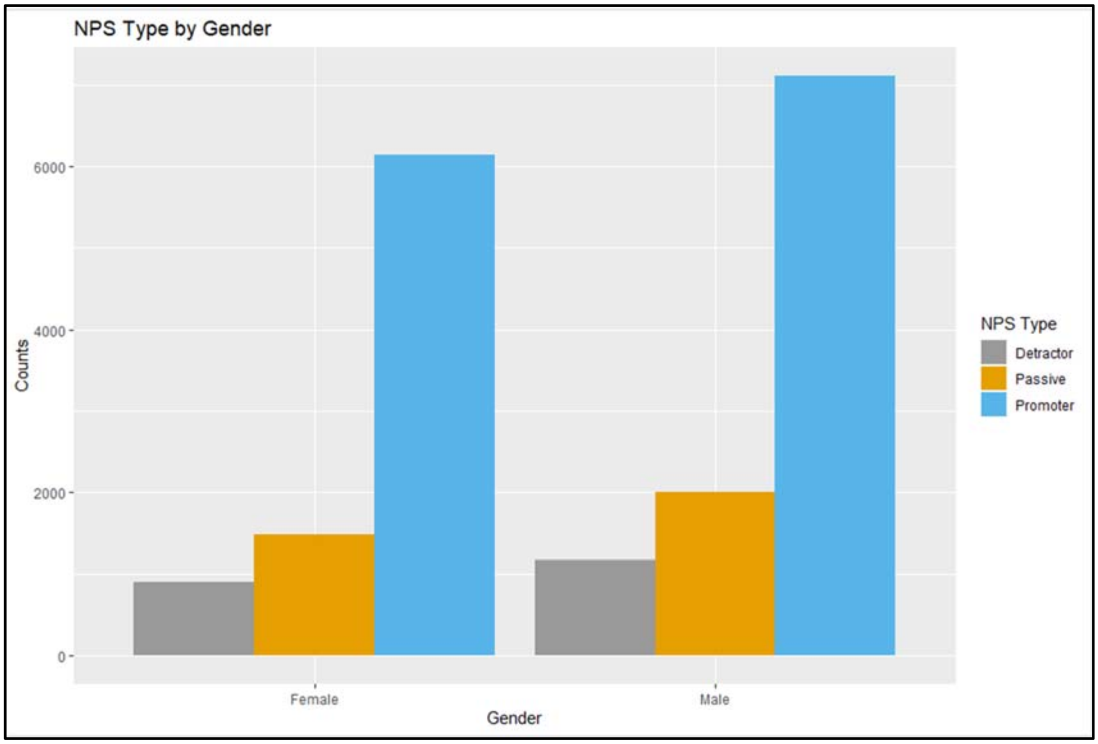
Conclusion: It was interesting that the percentages for "promoters" were higher the older the survey taker was. The highest percentage of "detractors" were 26-45 years old. Market research into amenities that those age-groups enjoy would provide opportunities to tailor hotel experiences to a specific age group.

16. Does gender of survey taker affect NPS?

We used R to determine the NPS by gender. Percentages by gender were also calculated. A frequency graph was created in R. More males than females took the survey. Which was generally an overall surprise. What might be interesting in the future to compare which gender took the survey by the reason for their stay at the hotel (business vs leisure).

```
> gnps
                gender   nps
1                        213
2               Female  8511
3                 Male 10271
4 Prefer not to answer   347
> #NPS Type by Gender
> npsGender <- table(hotels$Gender_H, hotels$NPS_Type)
> npsGender

                       Detractor Passive Promoter
                              47      50      116
  Female                     899    1467     6145
  Male                      1161    1997     7113
  Prefer not to answer        62      89      196
> c
```

GENDER	Detractor	% Detractor	Passive	% Passive	Promoter	% Promoter	Total by Gender
	NPS TYPE						
Female	899	11%	1467	17%	6145	72%	8511
Male	1161	11%	1997	19%	7113	69%	10271



Conclusion: There were generally more males than females taking the survey from our dataset. In general, females were more likely to be promoters than males. In today's world, marketing to a particular gender could prove to be difficult. With additional data around why the person was staying (business or leisure) some direct marketing campaigns could be developed.

19. <u>Does the class of the hotel impact the NPS?</u>

The goal was to determine whether the NPS is affected by what class the hotel was. There were three hotel classes in the data, Luxury, Upper Upscale, and Upscale. We compare the NPS scores as well as the likelihood rating per each hotel class. We also compared the average likelihood rating for each hotel class and compared it to the overall rating.

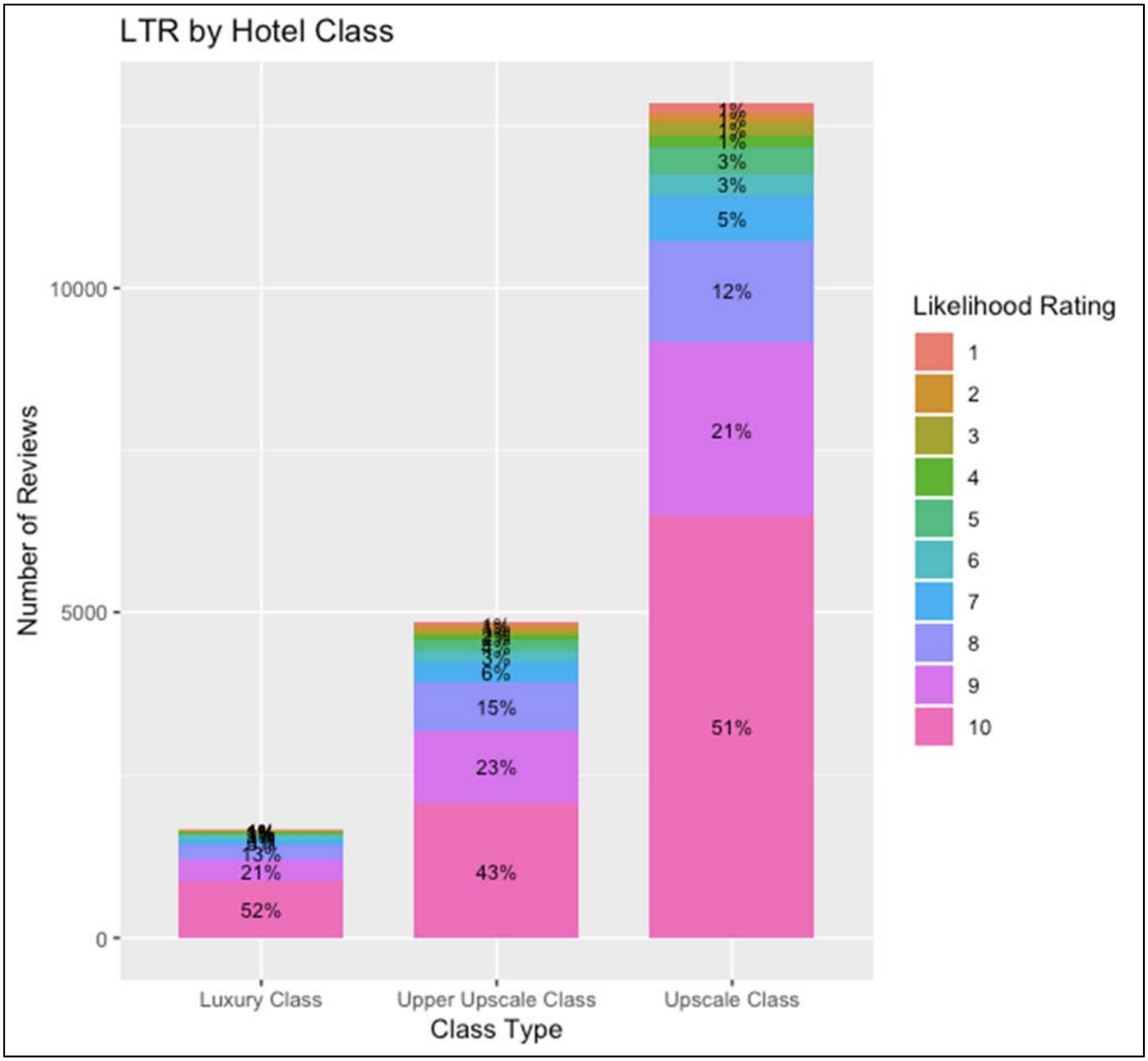
Here is the table:
new2data <- subset(hotel_data, select = c('Class_PL', 'NPS_Type'))

Cl.NPS <- as.data.frame(count(new2data))

```
           Class_PL  NPS_Type freq
1        Luxury Class Detractor  148
2        Luxury Class   Passive  297
3        Luxury Class  Promoter 1203
4 Upper Upscale Class Detractor  610
5 Upper Upscale Class   Passive 1039
6 Upper Upscale Class  Promoter 3187
7       Upscale Class Detractor 1410
8       Upscale Class   Passive 2266
9       Upscale Class  Promoter 9175
```

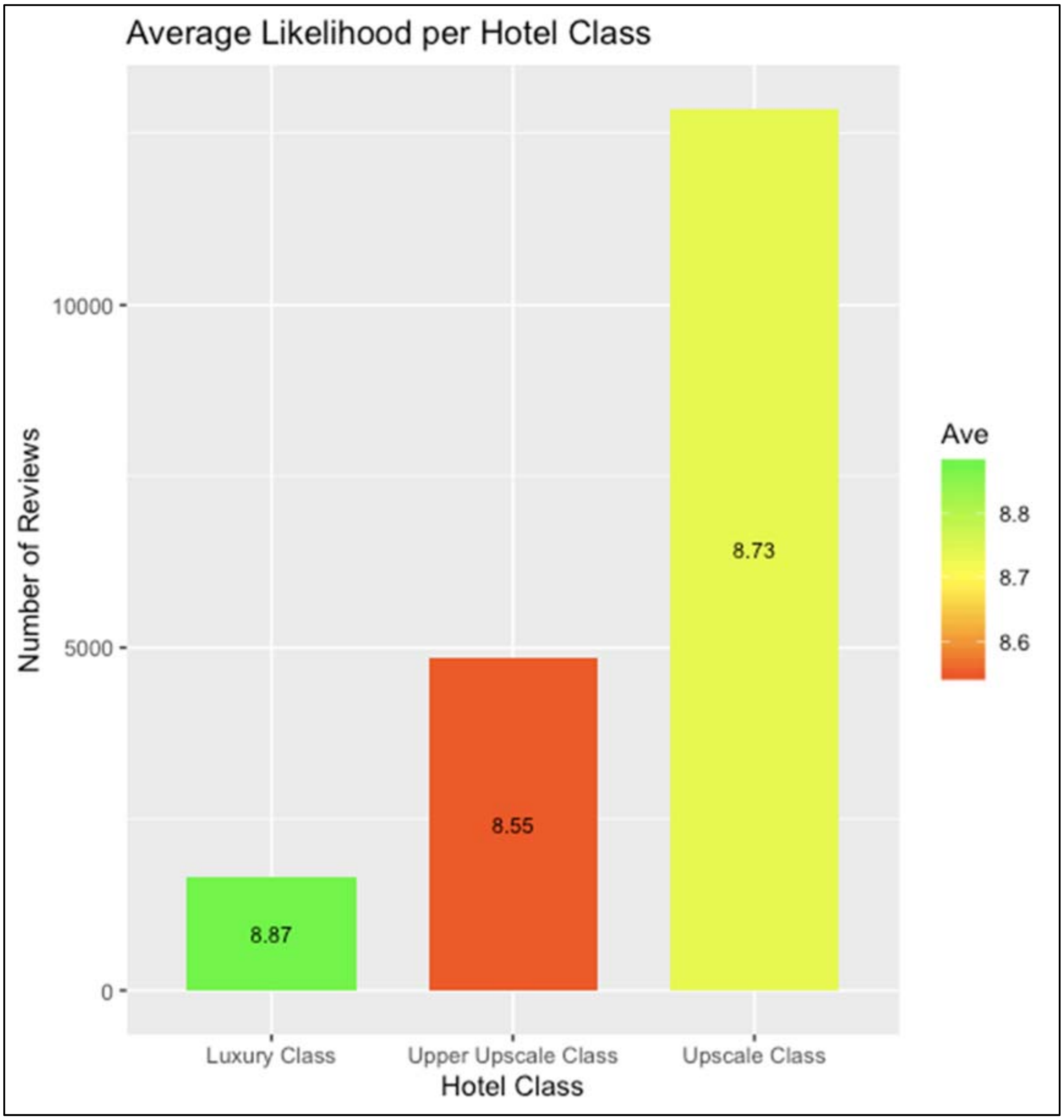
Here is the NPS Rating comparing each class:

For likelihood:
new2data2 <- subset(hotel_data, select = c('Class_PL', 'Likelihood_Recommend_H'))
CL.LIKE <- as.data.frame(count(new2data2))



Because most of the lower rating were very low and hard to read, we used another visual to see the overall rating comparing to the average rating for the whole data set:

```
CLL.DB <- count(hotel_data$Class_PL)
CLLiker <- tapply(hotel_data$Likelihood_Recommend_H, list(hotel_data$Class_PL), mean)
CLL.DB <- cbind(CLL.DB, CLLiker)
```



First the average rating for the entire set was calculated by:
likemean <- mean(hotel_data$Likelihood_Recommend_H)
[1] 8.699716

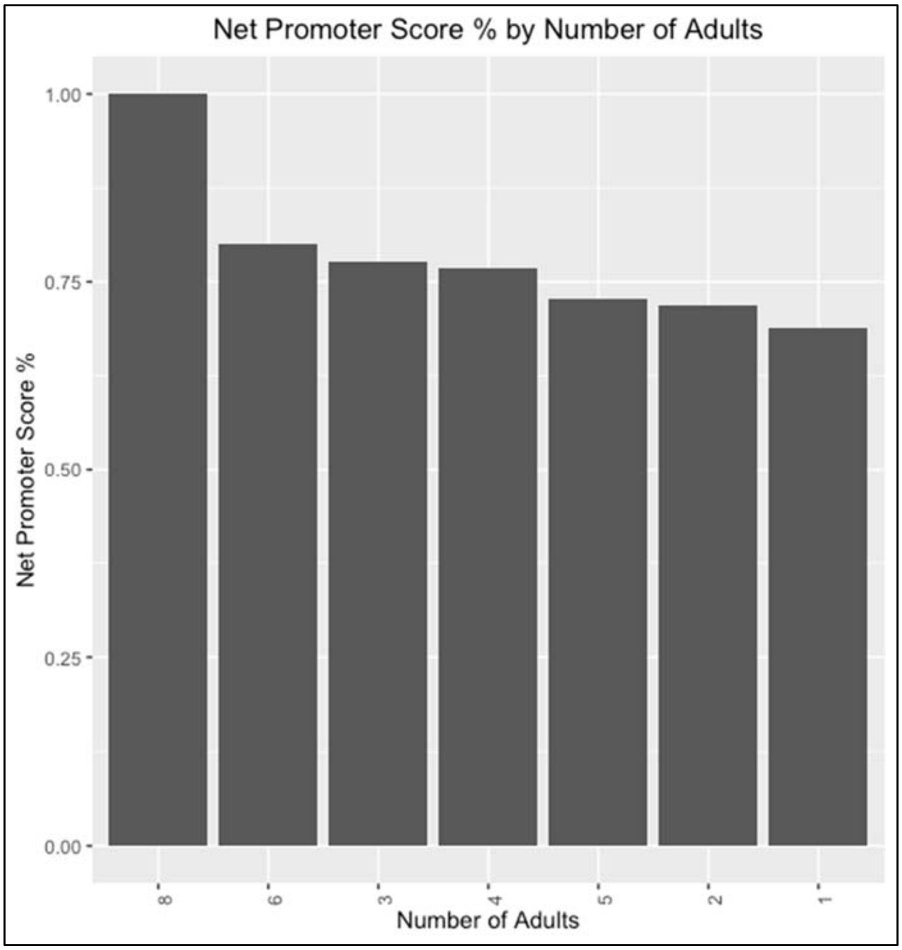
Comparing each hotel class to that would give different colors depending on its own average. More green would be better than the mean, yellow meaning closer to the mean, and red meaning lower than the mean.

Conclusion: Though the hotel class does affect the NPS score and the highest hotel class did have the highest average likelihood rating and the highest promoter score average, it still didn't translate to having much of an impact or the middle hotel class , upper upscale class, had worse averages and ratings compared to the lowest hotel class, upscale class.

20. <u>Does the amount of adults per reservation impact the NPS?</u>

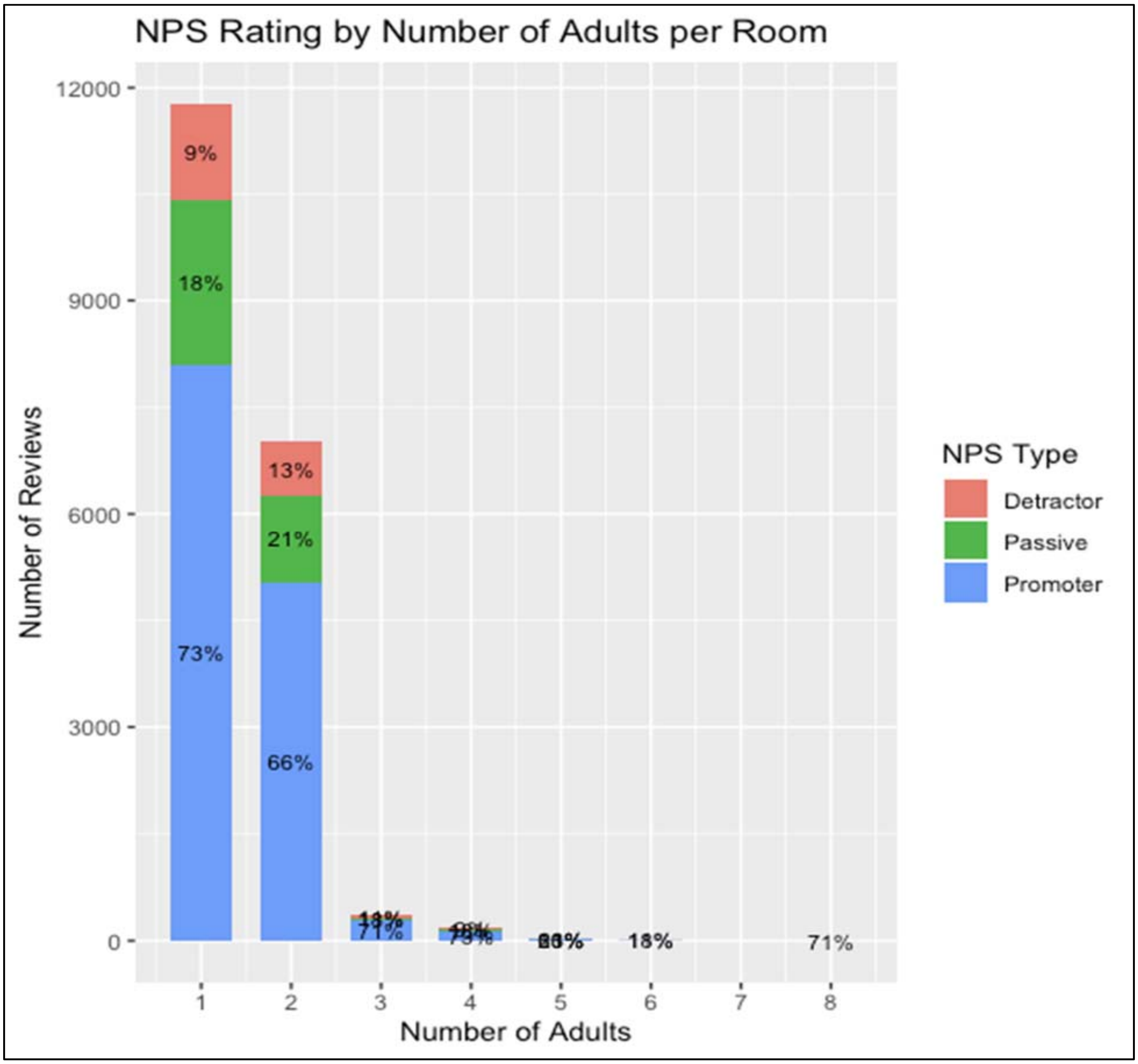
The goal was to determine if space or amount of adults affect the NPS rating. There was a total of 7 different amounts of adults per reservation, 1-6 and 8 adults. The NPS scores and likelihood rating per each amount was compared to each other. The average likelihood for each amount was also compared to the overall rating as well.

The first function we used was to see the percentage of the promoter scores first:



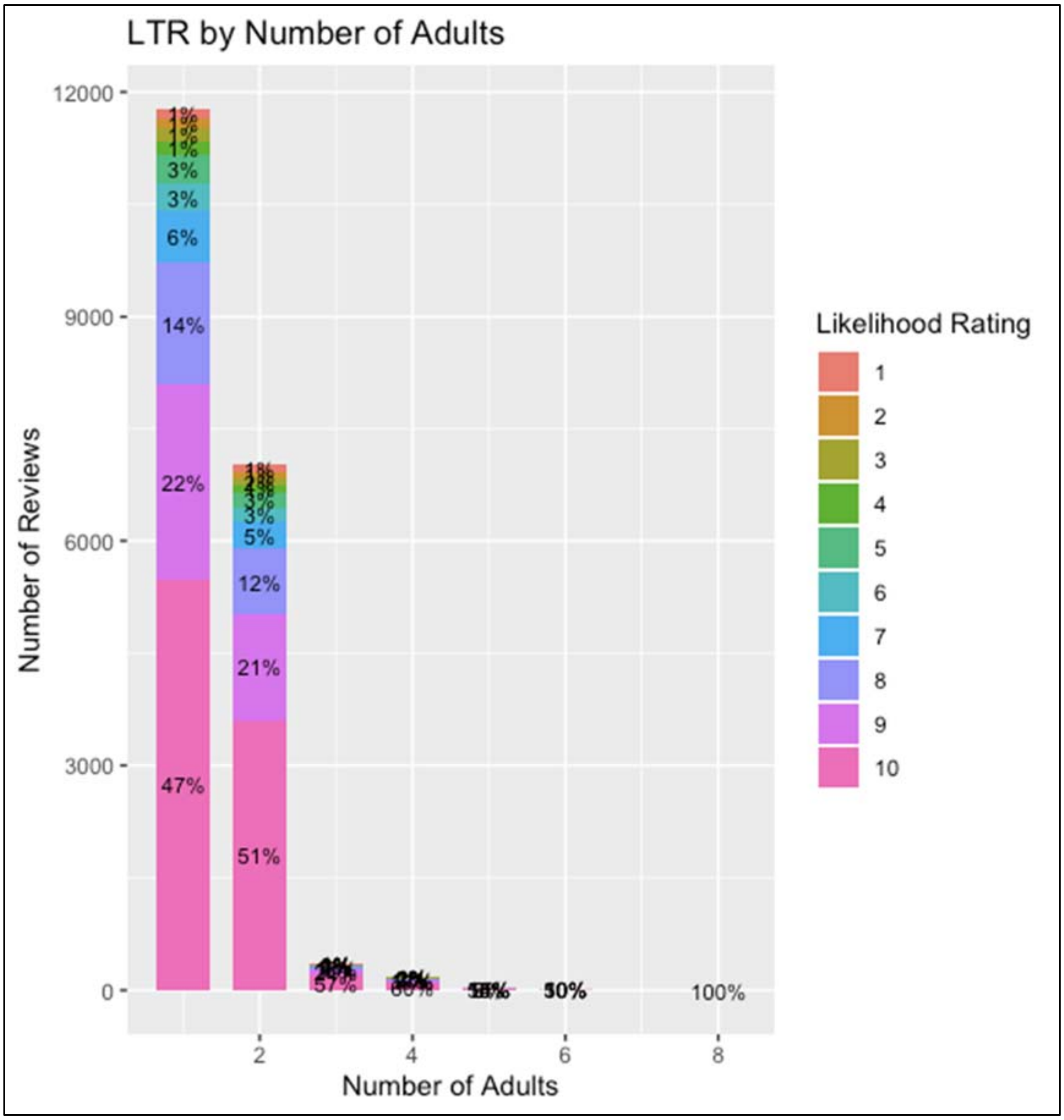
But the proportions were off, because there was only 1 reservation for 8 adults. So we then graphed vs the amount of reservations:

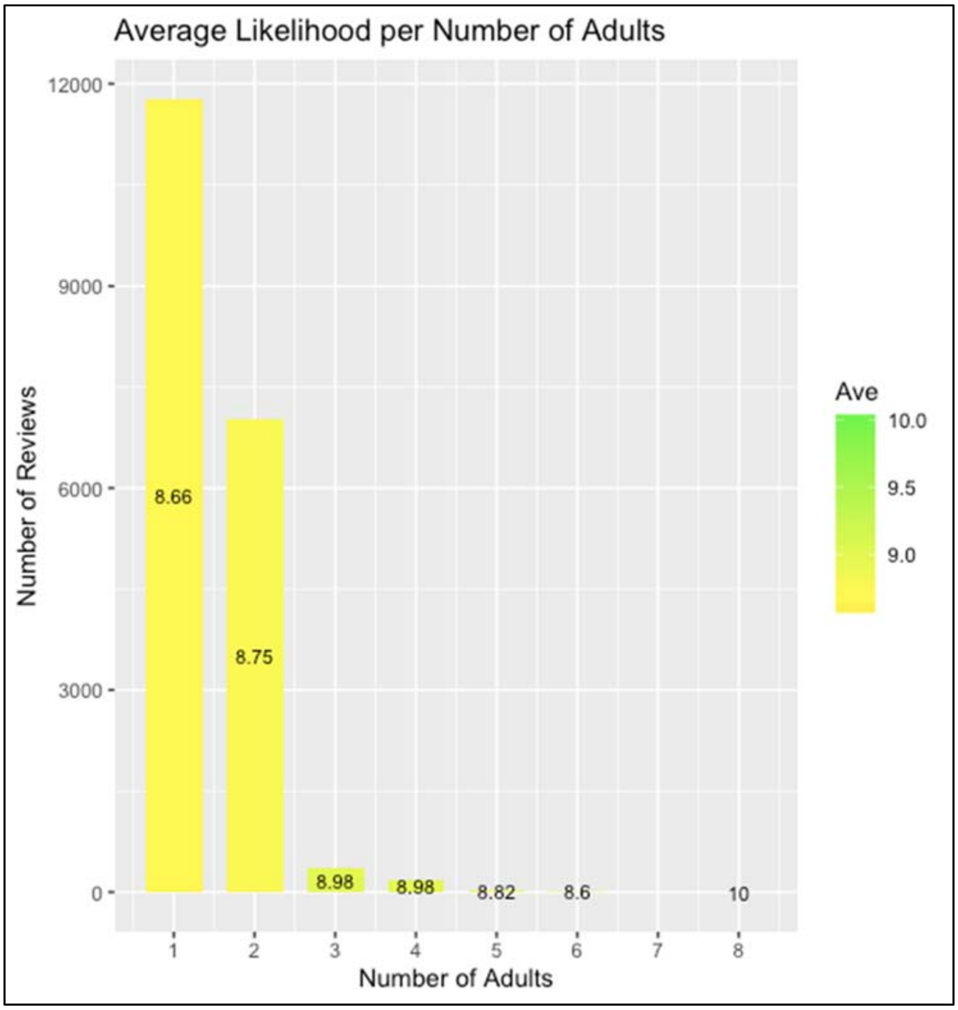
newdata <- subset(hotel_data, select= c("ADULT_NUM_C", "NPS_Type"))AD.NPS <- as.data.frame(count(newdata))



The number of adults were very disproportionate and unreadable, but this was to show the amounts of reservations per number of adults.

The graph below displays the likelihood rating per each adult group:


LTR by Number of Adults
Number of Reviews
Number of Adults
Likelihood Rating
12000
9000
6000
3000
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
47%
22%
14%
6%
3%
3%
1%
51%
21%
12%
5%
3%
57%
50%
100%



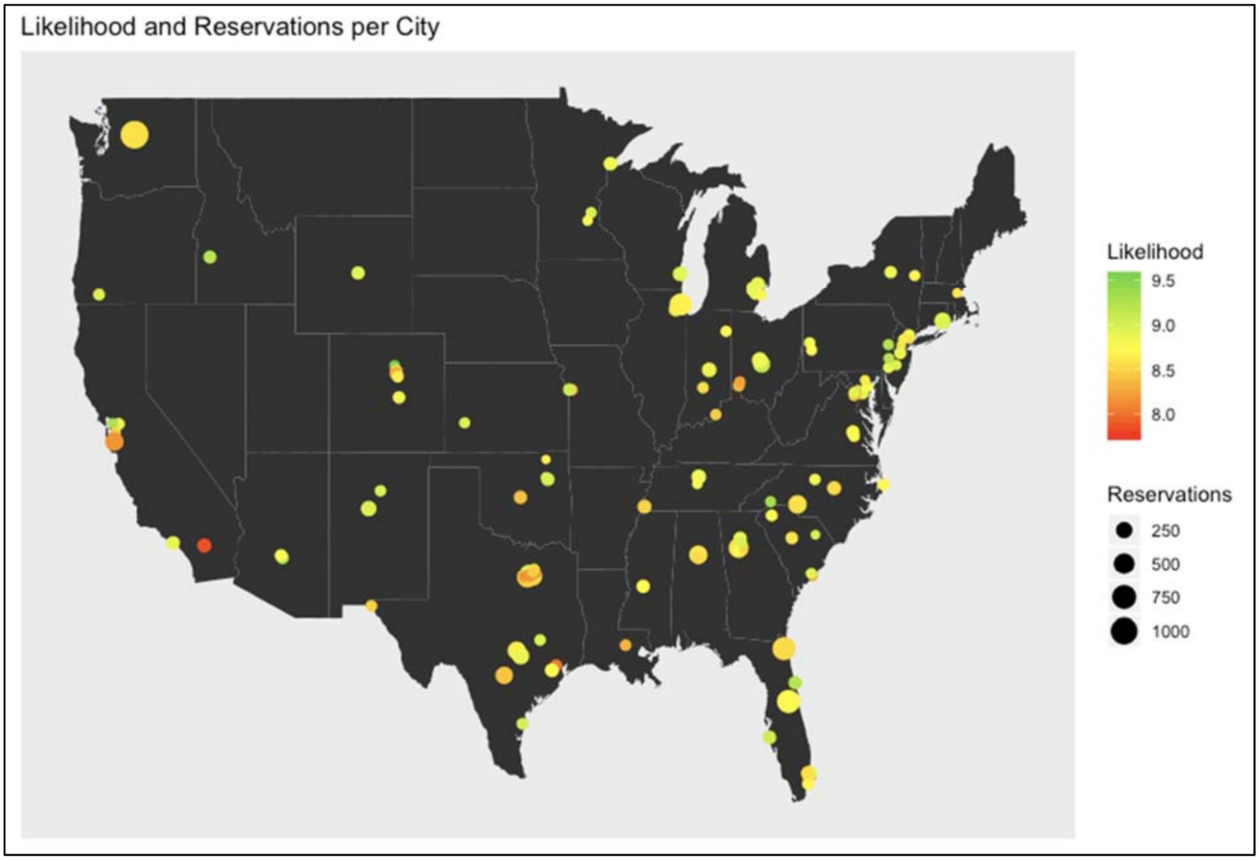
First the average rating for the entire set was calculated by:
likemean <- mean(hotel_data$Likelihood_Recommend_H)
[1] 8.699716

The most important for the larger sets of reservations would be the average likelihood. For the reservations for 1 and 2 adults seem to have low averages compared to the larger groups. That was due to the much larger set of data for those.

Conclusion: The trend seemed to be the more adults the higher the likelihood rating, but the increase was very slim and did not continue the entire trend. Also the proportions of review per each amount of adults staying were very lacking in the higher number of adults.

Likelihood and Reservations compared by city.

Using the google API we were able to plot the cities on the map with color compared to the average likelihood rating in the US only and the size of the dot in relation to the amount of reservations.



Unexecuted R Code

```
# save project survey data to a dataframe called df
df <- data.frame(ProjectSurveyData)
# structure of dataframe
str(df)

# Data Preparation Steps

# convert Reservation Date to date format
df$RESERVATION_DATE_R <- as.Date(df$RESERVATION_DATE_R)
# look at first few rows
head(df$RESERVATION_DATE_R)

# convert Check in Date to date format
df$CHECK_IN_DATE_C <- as.Date(df$CHECK_IN_DATE_C)
# look at first few rows
head(df$CHECK_IN_DATE_C)

# check for NAs in the variables needed for my analysis
any(is.na(NPS))
any(is.na(df$Likelihood_Recommend_H))
any(is.na(df$LENGTH_OF_STAY_C))
any(is.na(WALK_IN_FLG_C))
any(is.na(CHECK_IN_DATE_C))
any(is.na(RESERVATION_DATE_R))
any(is.na(Guest_Country_H))
any(is.na(Country_PL))

# calculate average length of stay
meanLS <- mean(df$LENGTH_OF_STAY_C,na.rm=TRUE)
roundedMeanLS <- round(meanLS)
roundedMeanLS

# replace NAs from LengthofStay with mean length of stay
df$LENGTH_OF_STAY_C[is.na(df$LENGTH_OF_STAY_C)]<- roundedMeanLS

# remove rows with no GuestCountry
df <- df[!(is.na(df$Guest_Country_H)),]
length(df[,1])

# assign each cleaned variable to a renamed vector
# NPS
NPS <- df$NPS
head(NPS)

tapply(NPS, NPS, length)
```

```
# Likelihood to Recommend
LTR <- df$Likelihood_Recommend_H
head(LTR)

# Length of Stay
LS <- df$LENGTH_OF_STAY_C
head(LS)

# WalkIn Flag
WalkInStatus <- df$WALK_IN_FLG_C
head(WalkInStatus)

# CheckIn Date
CheckInDate <- df$CHECK_IN_DATE_C
head(CheckInDate)

# Reservation Date
ReserveDate <- df$RESERVATION_DATE_R
head(ReserveDate)

# GuestCountry
GuestCountry <- df$Guest_Country_H
head(GuestCountry)

# HotelCountry
HotelCountry <- df$Country_PL
head(HotelCountry)

# change United States to USA in HotelCountry Column
HotelCountry[HotelCountry=="United States"] <- "USA"
HotelCountry[HotelCountry=="United States"]
HotelCountry[HotelCountry=="USA"]
HotelCountry

# AdvanceDays
# calculate a new column called AdvanceDays representing
# how far in advance a reservation was made
AdvanceDays <- CheckInDate - ReserveDate
# make AdvanceDays numeric
AdvanceDays <- as.numeric(AdvanceDays)
# look at first few rows of AdvanceDays
head(AdvanceDays)

# Free Independent vs. Group Travel
FITvGroup <- df$GROUPS_VS_FIT_R
head(FITvGroup)
```

```
# data does not exist for this variable

# create a dataframe called HotelData to hold the variables needed for analysis
HotelData <-
data.frame(NPS,LTR,LS,WalkInStatus,CheckInDate,ReserveDate,GuestCountry,AdvanceDays,
HotelCountry)
# look at first few rows of HotelData
head(HotelData)
str(HotelData)

# use printVecInfo function for all continuous variables
printVecInfo(LTR)
printVecInfo(LS)

# create a bar chart for each discrete variable in HotelData

# NPS
bar_NPS <- ggplot(HotelData, aes(x=NPS)) + geom_bar()
bar_NPS

# WalkIn Flag
bar_WalkInStatus <- ggplot(HotelData, aes(x=WalkInStatus)) + geom_bar()
bar_WalkInStatus

# create a histogram for each continous variable in HotelData

# LTR
hist_LTR <- ggplot(HotelData, aes(x=LTR)) + geom_histogram(binwidth=1,color="black",
fill="white")
hist_LTR <- hist_LTR + scale_x_continuous(name="Likelihood to Recommend")
hist_LTR

# LS
hist_LS <- ggplot(HotelData, aes(x=LS)) + geom_histogram(binwidth=1,color="black",
fill="white")
hist_LS <- hist_LS + scale_x_continuous(name="Length of Stay (days)", limits=c(0, 15))
hist_LS

# figure out how many outlier dates exist
# Breakdown CheckInDate onto its components: Year, Month, Day
# then add these columns to HotelData

# create Year from CheckInDate
CIYear <- year(HotelData$CheckInDate)
unique(CIYear)
HotelData$CIYear <- CIYear
```

```
# create Month from CheckInDate
CIMonth <- month(HotelData$CheckInDate)
unique(CIMonth)
HotelData$CIMonth <- CIMonth
# create Day from CheckInDate
CIDay <- day(HotelData$CheckInDate)
unique(CIDay)
HotelData$CIDay <- CIDay

# create a variable called "season"
# determine which months are in this variable
tapply(CIMonth,CIMonth,length)
# since all months are in Winter, no need to create a variable called season

head(HotelData)
str(HotelData)

sqldf("SELECT CIDay AS Day FROM HotelData GROUP BY CIDay ORDER BY CIDay")

# determine the number of record(s) with 2013 & 2014 CheckInDates
# count rows
Year1 <- HotelData[HotelData$CIYear==2013,]
length(Year1[,1])
Year2<- HotelData[!(HotelData$CIYear==2013),]
length(Year2[,1])
# remove the 1 record with the 2013 CheckInDate from the dataset
HotelData <- Year2

attach(HotelData)
head(HotelData)
str(HotelData)

# CountryMatch
# create a new column called CountryMatch representing
# whether hotel guest stayed in is in their country
# of origin (Y) or not (N)

x <- c(1:length(HotelData[,1]))
head(x)

Compare <- function(x) {
  result <- GuestCountry[x]==HotelCountry[x]
  return(result)
  }

HotelData$CountryMatch <- Compare(x)
head(HotelData$CountryMatch)
```

```
length(HotelData$CountryMatch)

# histogram of CheckIn dates
hist_CheckIn <- ggplot(HotelData, aes(x=CheckInDate)) +
geom_histogram(binwidth=1,color="black", fill="white")
hist_CheckIn <- hist_CheckIn + xlab("CheckIn Date") + ylab("Number of CheckIns") +
ggtitle("CheckIns per Day in early 2014")
hist_CheckIn <- hist_CheckIn + scale_x_date(limits=c("2014-01-01","2014-04-01"))
hist_CheckIn

count <- tapply(LTR, CIMonth, length)
count

# line plot of CheckIn dates
CheckIns <- sqldf("SELECT CheckInDate, COUNT(CheckInDate) AS NumCheckIns FROM
HotelData GROUP BY CheckInDate")
dfCheckIns <- data.frame(CheckIns)
dfCheckIns

line_CheckIn <- ggplot(dfCheckIns,aes(x=CheckInDate, y=NumCheckIns)) + geom_line()
line_CheckIn


# ReserveDate
hist_ReserveDate <- ggplot(HotelData, aes(x=ReserveDate)) +
geom_histogram(binwidth=1,color="black", fill="white")
hist_ReserveDate <- hist_ReserveDate + xlab("Reservation Date") + ylab("Number of
Reservations") + ggtitle("Reservation Dates by Day for Stays in early 2014")
hist_ReserveDate <- hist_ReserveDate + scale_x_date(limits=c("2013-07-01","2014-04-01"))
hist_ReserveDate

# determine where the spike is
# count the number of reservations by date, then order them by the number of reservations
sqldf("SELECT ReserveDate, COUNT(ReserveDate) AS NumberReservations FROM
HotelData
    GROUP BY ReserveDate ORDER BY NumberReservations DESC")

# AdvanceDays
hist_AD <- ggplot(HotelData, aes(x=AdvanceDays)) + geom_histogram(binwidth=5)
hist_AD <- hist_AD + ggtitle("Days in Advance Reservation was made")
hist_AD

# GuestCountry
bar_GuestCountry <- ggplot(HotelData, aes(x=GuestCountry, y=length(GuestCountry))) +
geom_bar(stat="identity")
bar_GuestCountry <- bar_GuestCountry + xlab("Guest Country") + ylab("Number of
Reservations") + ggtitle("Reservations per Guest Country")
```

```
bar_GuestCountry <- bar_GuestCountry + theme(axis.text.x = element_text(angle = 90, hjust =
1))
bar_GuestCountry

##################################################################

# Q1 Does length of stay impact NPS?

# look at Likelihood to Recommend by Length of Stay
tapply(LTR, list(LS==1,NPS), length)

# What relationship exists between NPS and length of stay?
# plot NPS vs Length of Stay
plot(LS, LTR)

LTRbyLS <- sqldf("SELECT AVG(LTR) AS AvgLTR,
    LS
    FROM HotelData
    GROUP BY LS")
LTRbyLS <- data.frame(LTRbyLS)
head(LTRbyLS)

# plot AvgLTR by LS
plot(LTRbyLS$LS, LTRbyLS$AvgLTR)

# build a linear model
model <- lm(formula=AvgLTR ~ LS, LTRbyLS)
summary(model)
abline(model)

# Conclusion: Length of Stay does NOT significantly impact NPS
# (Adjusted R^2 value is negative and very small)

##################################################################

# Q2(a): Does whether guest stay was a walk-in or a reservation impact NPS?
# determine how many were WalkIns vs. not
WalkInStatus_Breakdown <- tapply(WalkInStatus, WalkInStatus, length)
WalkInStatus_Breakdown

# Conslusion: Base size of walk-ins is not big enough to draw conclusions about
# the impact of WalkInStatus on NPS

###################################################################
# Q2(b): Is NPS affected by how far in advance the reservation was made?

# plot Likelihood to Recommend by Advance Days
```

```
plot(AdvanceDays, LTR)

# build a linear model
model <- lm(LTR ~ AdvanceDays)
summary(model)
abline(model)

# create a data frame containing average likelihood to recommend by Advance Days
LTRbyAD <- sqldf("SELECT AVG(LTR) AS AvgLTR,
            AdvanceDays AS AD
            FROM HotelData
            GROUP BY AD")
LTRbyAD <- data.frame(LTRbyAD)
head(LTRbyAD)

# plot AvgLTR by AD
plot(LTRbyAD$AD, LTRbyAD$AvgLTR)

# build a linear model
model <- lm(formula=LTRbyAD$AvgLTR ~ LTRbyAD$AD, LTRbyAD)
summary(model)
abline(model)

# Conclusion: AdvanceDays does NOT significantly impact NPS
# (Adjusted R^2 value is negative and very small)

################################################################################

# Q3: Does guest country of origin impact NPS?

USA_stays <- length(HotelData$GuestCountry[HotelData$GuestCountry=="USA"])
USA_stays
nonUSA_stays <- length(HotelData$GuestCountry[!(HotelData$GuestCountry=="USA")])
nonUSA_stays

HotelData$USorNot <- ifelse((HotelData$GuestCountry=="USA"), "USA", "nonUSA")
HotelData$USorNot
attach(HotelData)

# create a bar chart of Guest Country
gg_bar <- ggplot(HotelData, aes(x=HotelData$USorNot)) + geom_bar()
gg_bar

result <- sqldf("SELECT GuestCountry, AVG(LTR) AS LTR FROM HotelData GROUP BY
GuestCountry ORDER BY LTR DESC")
LTRbyCountry <- data.frame(result)
LTRbyCountry
```

```
plot(LTRbyCountry$GuestCountry, LTRbyCountry$LTR)

# inspect the range of average LTRs by country
printVecInfo(LTRbyCountry$LTR)

# create a linear model of LTR by Country
CountryModel <- lm(LTR ~ GuestCountry, HotelData)
summary(CountryModel)
abline(CountryModel)

# calculate AvgLTR by whether country was US or Not
LTRbyUSorNot <- sqldf("SELECT AVG(LTR) AS AvgLTR,
    USorNot
    FROM HotelData
    GROUP BY USorNot")
LTRbyUSorNot <- data.frame(LTRbyUSorNot)
LTRbyUSorNot

##################################################################################

# Q3(b): How does country of origin impact NPS in countries other than the guest's country of
origin?

# create a bar chart of Country Match, depicting # of stays where hotel country was in the guest's
# country of origin or not

gg_bar <- ggplot(HotelData, aes(x=CountryMatch)) + geom_bar()
gg_bar

# calculate AvgLTR by whether guest stayed in home country or not
LTRbyCM <- sqldf("SELECT AVG(LTR) AS AvgLTR,
            CountryMatch AS CM
            FROM HotelData
            GROUP BY CountryMatch")
LTRbyCM <- data.frame(LTRbyCM)
LTRbyCM

##################################################################################

# Q4: During which time is NPS the highest? (season/month/day of week/weekday vs. weekend)

# since all months are in Winter, no need to create a variable called season
# most data was collected in February; base size is too small to compare to January or March

# I'm not sure how to determine which weekday each stay was on
```

```
################################################################################

# Q5: How does whether travel is free independent travel vs. group travel impact NPS?

# data does not exist for this variable

################################################################################


# convert survey data to numeric format
df$Guest_Room_H <- as.numeric(df$Guest_Room_H)
df$Tranquility_H <- as.numeric(df$Tranquility_H)
df$Condition_Hotel_H <- as.numeric(df$Condition_Hotel_H)
df$Customer_SVC_H <- as.numeric(df$Customer_SVC_H)
df$Staff_Cared_H <- as.numeric(df$Staff_Cared_H)
df$Internet_Sat_H <- as.numeric(df$Internet_Sat_H)
df$Check_In_H <- as.numeric(df$Check_In_H)

# look at first few rows of each
head(df$Guest_Room_H)
head(df$Tranquility_H)
head(df$Condition_Hotel_H)
head(df$Customer_SVC_H)
head(df$Staff_Cared_H)
head(df$Internet_Sat_H)
head(df$Check_In_H)

# Which portions of survey data might lend some insight into Likelihood to Recommend Scores?

# Data Preparation Steps

# convert survey data to numeric format
df$Guest_Room_H <- as.numeric(df$Guest_Room_H)
df$Tranquility_H <- as.numeric(df$Tranquility_H)
df$Condition_Hotel_H <- as.numeric(df$Condition_Hotel_H)
df$Customer_SVC_H <- as.numeric(df$Customer_SVC_H)
df$Staff_Cared_H <- as.numeric(df$Staff_Cared_H)
df$Internet_Sat_H <- as.numeric(df$Internet_Sat_H)
df$Check_In_H <- as.numeric(df$Check_In_H)

# look at first few rows of each
head(df$Guest_Room_H)
head(df$Tranquility_H)
head(df$Condition_Hotel_H)
head(df$Customer_SVC_H)
head(df$Staff_Cared_H)
head(df$Internet_Sat_H)
```

```
head(df$Check_In_H)

# check for NAs in the variables needed for my analysis
any(is.na(df$NPS))
any(is.na(df$Likelihood_Recommend_H))
any(is.na(df$Guest_Room_H))
any(is.na(df$Tranquility_H))
any(is.na(df$Condition_Hotel_H))
any(is.na(df$Customer_SVC_H))
any(is.na(df$Staff_Cared_H))
any(is.na(df$Internet_Sat_H))
any(is.na(df$Check_In_H))

# Likelihood to Recommend
LTR <- df$Likelihood_Recommend_H
head(LTR)

# Guest Room Satisfaction
GuestRoom <- df$Guest_Room_H
head(GuestRoom)

# Tranquility
Tranquility <- df$Tranquility_H
head(Tranquility)

# Condition
Condition <- df$Condition_Hotel_H
head(Condition)

# Customer Service
CustServ <- df$Customer_SVC_H
head(CustServ)

# Staff Cared
Staff <- df$Staff_Cared_H
head(Staff)

# Internet
Internet <- df$Internet_Sat_H
head(Internet)

# CIProcess
CIProcess <- df$Check_In_H
head(CIProc)

# create a dataframe called HotelData to hold the variables needed for analysis
```

```
HotelData <-
data.frame(LTR,GuestRoom,Tranquility,Condition,CustServ,Staff,Internet,CIProcess)
# look at first few rows of HotelData
head(HotelData)
str(HotelData)

# remove rows containing NAs
HotelData <- na.omit(HotelData)
str(HotelData)

###################################################################################

# Which portions of survey data might lend some insight into Likelihood to Recommend Scores?

SurveyDataModel_1 <- lm(LTR ~ GuestRoom + Tranquility + Condition + CustServ + Staff +
Internet + CIProcess, HotelData)
summary(SurveyDataModel_1)

SurveyDataModel_2 <- lm(LTR ~ CustServ, HotelData)
summary(SurveyDataModel_2)

SurveyDataModel_3 <- lm(LTR ~ CustServ + GuestRoom, HotelData)
summary(SurveyDataModel_3)

SurveyDataModel_4 <- lm(LTR ~ CustServ + GuestRoom + Condition, HotelData)
summary(SurveyDataModel_4)

SurveyDataModel_5 <- lm(LTR ~ CustServ + GuestRoom + Condition + Staff + Tranquility,
HotelData)
summary(SurveyDataModel_5)

# recommend using combination of Customer Service, GuestRoom & Condition of Hotel to
predict LTR
```

```
------------------------------------------------------------------------------------------------------------------------
----------
# Load file
file_path <- "~/Syracuse/IST687_Intro_DS/GroupProject/ProjectSurveyData.csv"
hotel_data <- read.csv(file=file_path, header=TRUE, sep=",", stringsAsFactors = FALSE)

attach(hotel_data)

# ---------
# Functions
# ---------

# Generate Bar Graph
generate_bar_graph <- function(df, x, y, x_label, y_label){
  # Create bar graph
  g <- ggplot(df, aes(x=reorder(x, -y), y=y)) + geom_bar(stat="identity")
  title <- paste(y_label, "by", x_label, sep = " ")
  g <- g + ggtitle(title) + theme(plot.title = element_text(hjust=0.5))
  g <- g + xlab(x_label) + ylab(y_label) +  theme(axis.text.x = element_text(angle = 90, hjust =
1))
  return(g)
}

# Clean and return dataframe of percentages
get_percentages <- function(data){
  data[is.na(data)] <- 0
  data <- round(data[,"Promoter"] / (data[,"Promoter"] + data[,"Detractor"] + data[,"Passive"]),3)
# Get Percentages
  data <- data[order(-data)] # Order data by descending
  data <- data.frame(data) # Sadatae as dataframe
  return(data)
}

# -------------------------------------------------
# 6) Which regions have the highest NPS Percentage?
# -------------------------------------------------

nps.region <- tapply(NPS_Type, list(Region_PL, NPS_Type), length) # Query Data
top_region <- get_percentages(nps.region) # Generate Percentage dataframe
generate_bar_graph(top_region, rownames(top_region), top_region$data, "Region", "Net
Promoter Score %" ) # Generate bar graph

# -------------------------------------------------
# 6) Which hotels have the worst NPS Percentage?
# -------------------------------------------------

nps.hotel <- tapply(NPS_Type, list(Hotel.Name.Short_PL, NPS_Type), length) # Query Data
```

```
nps_hotel <- get_percentages(nps.hotel) # Generate Percentage dataframe
nps_hotel.hist <- hist(nps_hotel$data, main="Histogram for Hotels", xlab="NPS Percentage by promoter")
nps_hotel.hist
summary(nps_hotel)
worst_hotel <- tail(nps_hotel, 20)
best_hotel <- head(nps_hotel, 20)
generate_bar_graph(worst_hotel, rownames(worst_hotel), worst_hotel$data, "Hotel", "Net Promoter Score %" ) # Generate bar graph
generate_bar_graph(best_hotel, rownames(best_hotel), best_hotel$data, "Hotel", "Net Promoter Score %" ) # Generate bar graph

# ---------------------------------------
# 7) Which countries have the highest NPS?
# ---------------------------------------

nps.country <- tapply(NPS_Type, list(Country_PL, NPS_Type), length) # Query Data
top_countries <- get_percentages(nps.country) # Generate Percentage dataframe
top_countries <- head(top_countries, 20) # Get highest percentages from dataframes
generate_bar_graph(top_countries, rownames(top_countries), top_countries$data, "Country", "Net Promoter Score %" ) # Generate bar graph

# ---------------------------------------------
# 10) How does the reason for stay impact the NPS
# ---------------------------------------------

nps.stay <- tapply(NPS_Type, list(POV_CODE_C, NPS_Type), length) # Query data
top_stay <- get_percentages(nps.stay) # Generate Percentage dataframe
generate_bar_graph(top_stay, rownames(top_stay), top_stay$data, "Reason for Stay", "Net Promoter Score %" ) # Generate bar graph

rstay <- sqldf("Select POV_CODE_C as stay,  Likelihood_Recommend_H as score from hotel_data")
rmodel <- lm(score ~ stay, data = rstay) # Run linear model
summary(rmodel)

# -------------------------------------------------
# 11) Which rooms have the lowest net promoter score?
# -------------------------------------------------

nps.room <- tapply(NPS_Type, list(ROOM_TYPE_CODE_C, NPS_Type), length) # Query Data
nps.room <- get_percentages(nps.room) # Generate Percentage dataframe
best_rooms <- head(nps.room, 20)
worst_room <- tail(nps.room, 20) # Get lowest pecentages from dataframe
generate_bar_graph(worst_room, rownames(worst_room), worst_room$data, "Room Type", "Net Promoter Score %" ) # Generate bar graph
```

```
nps.room.hist <- hist(nps.room$data, main="Histogram for Rooms", xlab="NPS Percentage")
summary(nps.room)

# ------------------------------------------
# How does Award Impact likelihood to recommend
# ------------------------------------------

df <- sqldf("Select hotel_data.'Award.Category_PL' as award,  Likelihood_Recommend_H as
score from hotel_data")
rmodel <- lm(df$score ~ df$award, data = df) # Run linear model
summary(rmodel) # Summarize Model

# ---------------------------------------------------------------------------------
# Does the difference between expected and actual costs impact the likelihood to recommend
# ---------------------------------------------------------------------------------

df <- sqldf("Select round(abs((REVENUE_USD_R - (QUOTED_RATE_C *
LENGTH_OF_STAY_C ))),2) as cost_diff,  Likelihood_Recommend_H as score from
hotel_data") # Get the difference between actual and expected cost
rmodel <- lm(df$score ~ df$cost_diff, data = df) # Run linear model
summary(rmodel) # Summarize model

# ------------------------------------------------------------------------------
# Which of the other hotel features contributes to the likelihood to recommend score?
# ------------------------------------------------------------------------------

s <- sqldf("Select Likelihood_Recommend_H, Guest_Room_H, Tranquility_H,
Condition_Hotel_H, Customer_SVC_H, Staff_Cared_H, Internet_Sat_H, Check_In_H from
hotel_data")
rmodel <- lm(Likelihood_Recommend_H ~ ., data = s) # Run linear model
summary(rmodel) # Summarize Model

s_new <- sqldf("Select Likelihood_Recommend_H, Guest_Room_H, Tranquility_H,
Condition_Hotel_H, Customer_SVC_H, Staff_Cared_H from hotel_data")
rmodel <- lm(Likelihood_Recommend_H ~ ., data = s_new) # Run linear model
summary(rmodel) # Summarize Model
------------------------------------------------------------------------------------------------------------
----------########################################################################
# Q12
# Does the room rate the guest paid stayed impact the NPS?
# Unable to answer due to the lack of data

########################################################################
# Q13
# Does size of hotel (number of rooms &/or number of floors) impact the NPS?
# Unable to answer due to the lack of data
```

```
########################################################################
# Q14
# Does whether or not the guest was offered a promotion impact the NPS?
# - Is either past or future offer more impactful?
# Unable to answer due to the lack of data

########################################################################
# Q15
# Which age groups give the highest NPS?
npsAge <- table(hotels$Age_Range_H, hotels$NPS_Type)
npsAge

# Removing NAs
npsAge <- na.omit(npsAge)
npsAge
View(npsAge)
npsAge <- npsAge[-c(1),]
npsAge
dfnpsAge <- as.data.frame(npsAge)
dfnpsAge

# Plot 1: stacked histogram
hotels$count <- 1
npsageCounts <- aggregate(hotels$count, by = list(age=hotels$Age_Range_H,
                                    NPS_Type=hotels$NPS_Type), FUN=sum)
npsagePlot1 <- ggplot(npsageCounts, aes(x=age, y=x, fill=NPS_Type)) +
  geom_bar(stat = "identity")
npsagePlot1

# determine color pallette - color-blind friendly
cbfPalette <- c("#999999", "#E69F00", "#56B4E9")

# Plot 2: plot seperated by age group and promoter score
npsagePlot2 <- ggplot(dfnpsAge, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_bar(stat = "identity", position=position_dodge()) +
  scale_fill_manual(values=cbfPalette, name="NPS Type") + labs(x="Age", y="Counts") +
  ggtitle("NPS Type by Age")
npsagePlot2

########################################################################
# Q16
# Does gender of survey taker affect NPS?


# Removing NAs
npsGender <- table(hotels$Gender_H, hotels$NPS_Type)
```

```
npsGender
npsGender <- na.omit(npsGender)
npsGender
View(npsGender)
npsGender <- npsGender[-c(1),]
npsGender <- npsGender[-c(3),]
npsGender
View(npsGender)
dfnpsGender

# Plot 1: stacked histogram
hotels$count <- 1
npsGenderCounts <- aggregate(hotels$count, by = list(gender=hotels$Gender_H,
                                 NPS_Type=hotels$NPS_Type), FUN=sum)
npsGenderPlot1 <- ggplot(npsGenderCounts, aes(x=gender, y=x, fill=NPS_Type)) +
  geom_bar(stat = "identity")
npsGenderPlot1

# determine color pallette - color-blind friendly
cbfPalette <- c("#999999", "#E69F00", "#56B4E9")

# Plot 2: plot seperated by gender and promoter score
npsGenderPlot2 <- ggplot(dfnpsGender, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_bar(stat = "identity", position=position_dodge()) +
  scale_fill_manual(values=cbfPalette, name="NPS Type") + labs(x="Gender", y="Counts") +
  ggtitle("NPS Type by Gender")
npsGenderPlot2
```