Syracuse University
School of Information Studies
Fall 2017

School of Information Studies
Syracuse University

IST687 APPLIED DATA SCIENCE

Final Project Report

Instructors: Jeffrey Saltz; Gary Krudys
TA: Ivan Shamshurin

Team members:
David Forteguerre
Bradley Wuon Seok Choi
Devin Shannon
Hem Adhikari

Academic Year 2017 – 2018

EXECUTIVE SUMMARY

This final report was prepared by a group of Syracuse University consultants, David Forteguerre, Bradley Wuon Seok Choi, Devin Shannon, and Hem Adhikari as part of IST 687: Applied Data Science (Syracuse University, Fall 17).

We were provided with a modified version of a large dataset from the Hyatt Hotel chain. The dataset contained customer information, survey results, and we were asked to serve as consultants to analyse the customer survey responses for our client. The goal was to identify and then answer interesting questions, to focus on NPS (Net Promoter Score) and likelihood to recommend, to identify the key drivers that could improve NPS, but no specific questions or goals were provided. The full analysis was determined by the team.

Primary *(for quality purposes)*

Here are final recommendations to the Hyatt Hotels Corporation management:

- Focus on all of the variables relating to customer service but put the most emphasis specifically on the top two variables: Customer_SVC_H and Guest_Room_H.
- Put less emphasis on Tranquility_H and F.B_Overall_Experience_H as they have less of an effect, but do not disregard them altogether as they do have a little bit of influence.
- Do not focus on variables such as Internet_Sat_H and Check_in_H as they don't have much if any influence on customer satisfaction.
- Based on SVM Classification analysis Hyatt hotel should pay close attention on Customer service.
- Provide more customer service training for management and staff.
- Cultivate and maintain a strong customer-oriented experience by way of providing exceptional personal (positive employee-to-customer interaction) service.

Secondary *(for marketing purposes)*

- NYC is an international tourist hub, as such it is disappointing that the majority of the customers are U.S.-based travellers. The Hyatt Hotel Corporation should try to attract more international travellers.
- Hyatt Hotel Corporation should consider a wider variety of demographics:
    - Female travelers
    - Leisure travelers
    - Younger and older travelers besides (40-60)

TABLE OF CONTENTS

I. INTRODUCTION

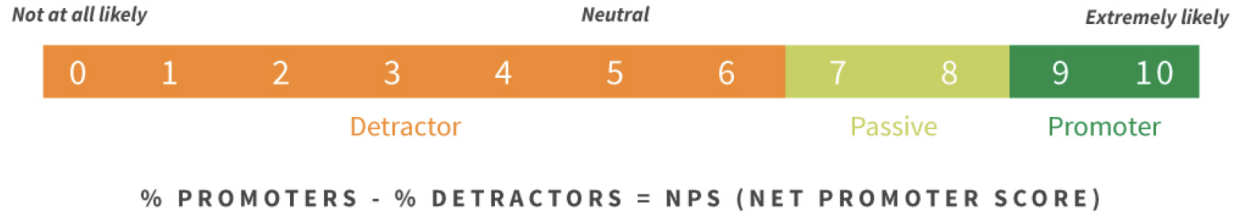
This final report was prepared by a group of Syracuse University consultants, David Forteguerre, Bradley Wuon Seok Choi, Devin Shannon, and Hem Adhikari. Our team was provided with a large dataset of 17.56 GB containing customer surveys from the Hyatt Hotels Corporation, an American-based multinational owner, operator, and franchiser of hotels, resorts, and vacation properties. The surveys were recorded over the span of a year (12 files, 1 month each) and included customer feedback from hotels all around the world. The dataset had 237 variables (columns) and millions of records (rows.)

Customer feedback is crucial for any company that wishes to improve their services, products, and overall customer experience. Thus, our key task was to clean and analyse the dataset in order to provide actionable insights to the hotel chain whose ultimate goal was to improve the Net Promoter Score of their hotels.

The concept of Net Promoter Score®, or NPS, was key in the analysis. NPS measures customer experience and predicts business growth. It is a proven metric that transformed the business world and now provides the core measurement for customer experience management programs around the world. Below is a quick explanation of how the metric works.[1]
Survey respondents are categorized as three different types thanks to their likelihood to recommend value.

- **Promoters** (score 9-10) are loyal enthusiasts who will keep buying and refer others, fueling growth.
- **Passives** (score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.
- **Detractors** (score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.

Then, subtracting the percentage of detractors from the percentage of promoters yields the NPS, which can range from a low of -100 (if every customer is a Detractor) to a high of 100 (if every customer is a Promoter).



[1] Source: https://www.netpromoter.com/know/

II. BUSINESS QUESTIONS

Our main goal was to investigate the contents of the full dataset and carefully analyse what data was available to us. We began by evaluating the given variables and determining which served to be most useful in our task of improving the customer experience at the Hyatt hotel chain. From this list of variables, along with looking at the data input, we were able to establish the following business questions which would guide us throughout our project:

1) Who are the customers? Who are the satisfied customers, and who are the unsatisfied customer ?
2) What factors directly affect a customer's satisfaction (more specifically, their "likelihood to recommend?")
3) Which modeling techniques can be used to predict the effect on likelihood to recommend?
4) How can Hyatt better target unsatisfied customers, and what steps can Hyatt management do to improve customer satisfaction?

III. METHODOLOGY

Our team used R, the programming language and software environment for statistical analysis, graphics representation and reporting, to conduct this analysis. We used RStudio to write and run all of the code for the project. Due to the large amount of code accumulated, we have decided to submit it as a separate .R file which can be opened directly in RStudio.[2] Please open the attached .R file (the **STEPS** in this report are denoted throughout the code).

1. Data importation, cleansing, munging, and preparation # STEP 1(a)

The first step consisted in importing the data into RStudio only keeping the variables we deemed relevant to this analysis. We used the read_csv command (library(readr) required) and stored each month into a variable. We then created one single dataset combining all 12 months, and saved the new full dataset to our local drive to be able to import it more easily (due to file size).

We decided to analyse the full dataset (all 12 months) instead of just picking 3 or 4 months as we had initially planned. The reason was that removing *NA's* from the data was required, which made us lose a considerable amount of data. Therefore, analyzing all 12 months was a good way to have enough data for the analysis and algorithms to be meaningful.

We also decided to keep **21** of the 237 in the original dataset for our first analysis. Below is a list of those variables and their definition.

[2] Professor Saltz's recommendation

	Variable Name	Meaning
1.	"Country_PL"	Country in which the hotel is located
2.	"City_PL"	City in which the hotel is located
3.	"Spirit_PL"	Unique hotel identifier (5-letter code)
4.	"Hotel Name-Long_PL"	Full hotel name
5.	"NPS_Type"	Indicates if the guest's HySat responses mark them as a promoter, a passive, or a detractor
6.	"Likelihood_Recommend_H"	Likelihood to recommend metric; value on a 1 to 10 scale
7.	"POV_CODE_C"	Purpose of visit
8.	"MEMBER_STATUS_R"	Tier of the GP program that the member belongs
9.	"LENGTH_OF_STAY_C"	Length of stay
10.	"GUEST_COUNTRY_R"	The country to which the actual guest belongs to. Different from Country_Code
11.	"Gender_H"	Guest's gender
12.	"Age_Range_H"	Guest's age range
13.	"Overall_Sat_H"	Overall satisfaction metric; value on a 1 to 10 scale
14.	"Guest_Room_H"	Guest room satisfaction metric; value on a 1 to 10 scale
15	"Tranquility_H"	Tranquility metric; value on a 1 to 10 scale
16.	"Condition_Hotel_H"	Condition of hotel metric; value on a 1 to 10 scale
17.	"Customer_SVC_H"	Quality of customer service metric; value on a 1 to 10 scale
18.	"Staff_Cared_H"	Staff cared metric; value on a 1 to 10 scale
19.	"Internet_Sat_H"	Internet satisfaction metric; value on a 1 to 10 scale
20.	"Check_In_H"	Quality of the check in process metric; value on a 1 to 10 scale
21.	"F&B_Overall_Experience_H"	Overall F&B experience metric; value on a 1 to 10 scale

Note that not all 21 variables were used throughout the full analysis. These variables were all chosen for specific reasons, and were removed from the dataset whenever they were not needed anymore.

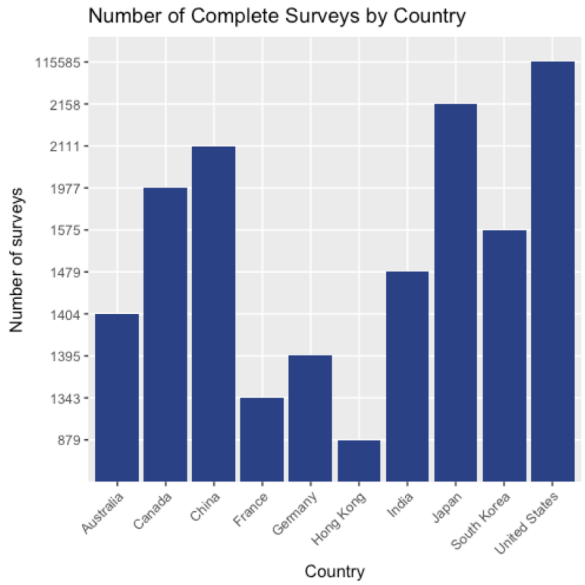
- **Variables 1 through 4** would help us choose which area to work on and narrow down the dataset to that area, and get information about the location/specific hotels as well.
- **Variables 5 and 6** were crucial in the analysis as they were our target variable. The goal of this project was to find ways to improve "Likelihood_Recommend_H". Note that NPS_Type was the actual category the customer belonged to (detractor, passive, or promoter.) We checked that this column had been created correctly according to "Likelihood_Recommend_H" and the definition we provided in the introduction.
- **Variables 7 to 12** would help us conduct a descriptive analysis of our customers. Customer profiles are always very helpful in understanding why they are satisfied or unsatisfied.
- **Variables 13 through 21** would help us discover the main factors why customers are either satisfied or unsatisfied through the use of modeling techniques.

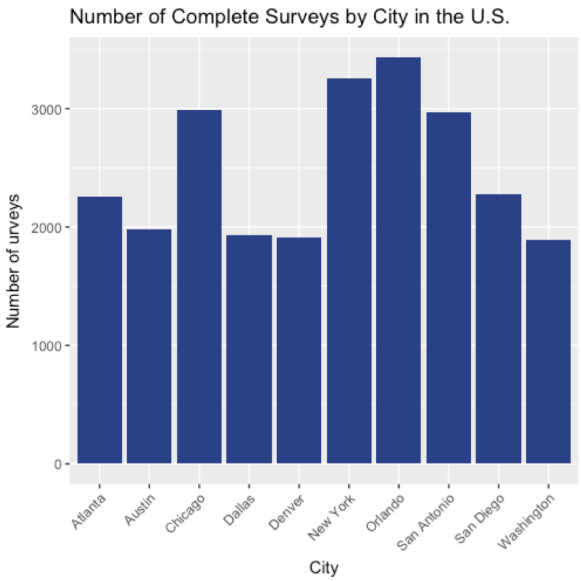
STEP 2

Then, we checked that the NPS_Type column had been created accordingly to the definition we gave in the introduction. We used the tapply and unique function to verify that, and everything was coherent.

STEP 3

After performing basic analysis on the dataset and plotting the results, we noticed that the United States was the country that had the largest amount of data, and that Orlando was the city that had the largest amount of surveys within the U.S.





STEP 4

However, we decided to focus on a city that had still enough data for the analysis to be meaningful, and therefore chose to work on New York City (NYC). Even though it was the second city with the largest number of surveys after the NAs had been removed, NYC still had a very complete set of data (3253 surveys in total).

STEP 1(b) (at the beginning)

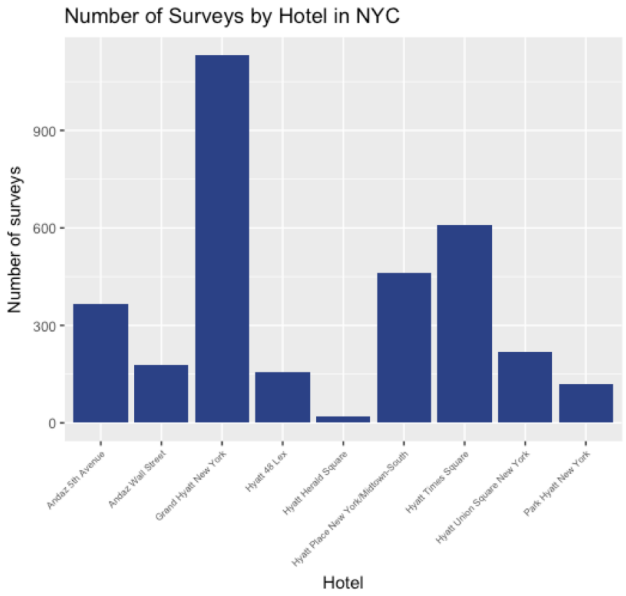
Before proceeding, we decided to do and extra step and reimport the full dataset at the very beginning of the code to compare the types of services that were offered by each NYC hotel on the last month of the survey. We only imported the services and created a table to be able to easily visualize how many NYC hotels offered the same services or not.

```
> HotelServicesSummary
  All.Suites_PL Bell.Staff_PL Boutique_PL Business.Center_PL Casino_PL Conference_PL Convention_PL Dry.Cleaning_PL Elevators_PL
N             6             1          7                  2         7             7             6               0            0
Y             1             3          0                  2         0             0             1               4            4
  Fitness.Center_PL Fitness.Trainer_PL Golf_PL Indoor.Corridors_PL Laundry_PL Limo.Service_PL Mini.Bar_PL Pool.Indoor_PL Pool.Outdoor_PL
N                 0                 3       7                   0          1              0          1              4              4
Y                 4                 1       0                   7          3              4          3              0              0
  Regency.Grand.Club_PL Resort_PL Restaurant_PL Self.Parking_PL Shuttle.Service_PL Ski_PL Spa_PL Spa.services.in.fitness.center_PL
N                     3         7             1               4                  4      7      6                                3
Y                     1         0             6               0                  0      0      1                                0
  Spa.online.booking_PL Spa.F.B.offering_PL Valet.Parking_PL
N                     1                   1                1
Y                     0                   0                3
```

We came to the conclusion that the services offered were quite consistent (i.e. all the NYC hotels had the same types of services), and thus decided not to include this data in our full analysis. Any complaints regarding the availability of hotel amenities would be reflected in the customer ratings anyway.

STEP 5

After keeping the NYC data only, we visualized how many hotels were in NYC and how many surveys each of them had. One of the hotels, the Grand Hyatt New York hotel, really stood out for having much more data than the others, which might emphasize its popularity.



2. NPS Calculation

STEP 6

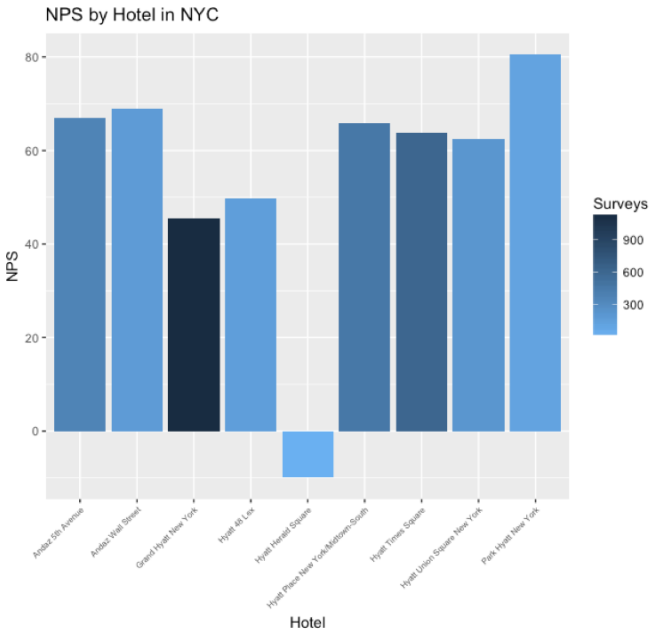
Only then the concept of Net Promoter Score became relevant. As the goal of this analysis was to *improve* NPS, we decided to calculate NPS for each hotel in NYC, to see if any of them would stand out for having (1) a very low NPS, (2) enough data for the low NPS to be meaningful. If that were the case, it would be relevant to only focus on those hotels to better analyse the data and be able to target the unsatisfied customers more efficiently.
We calculated NPS for each hotel using the following formula:

(Number of Promoters - Number of Detractors) / (Number of Respondents) x 100

We calculated it using a simple function which made the whole process much faster and easier. We then stored the NPS for each hotel into a data frame, and plotted it using ggplot2 again. Below are the results.

	HotelName	Surveys	NPS
1	Grand Hyatt New York	1130	45.48673
2	Hyatt Times Square	607	63.75618
3	Hyatt Place New York/Midtown-South	462	65.80087
4	Andaz 5th Avenue	365	66.84932
5	Hyatt Union Square New York	219	62.55708
6	Andaz Wall Street	177	68.92655
7	Hyatt 48 Lex	155	49.67742
8	Park Hyatt New York	118	80.50847
9	Hyatt Herald Square	20	-10.00000

The Net Promoter Scores across the NYC hotels were all quite similar (ranging from 45 to 80), and only one of them stood out for being very low: the Hyatt Herald Square. However, considering the amount of data that hotel had (20 surveys), we all agreed it would not be relevant to only focus on that specific hotel in particular (20 surveys are not enough!). Thus, we decided to remain focused on all the hotels in the NYC are



 # STEP 7

Before proceeding, we removed all the location information from the dataset which we did not need anymore (i.e. Country_PL, City_PL, Spirit_PL, Hotel.Name.Long_PL)

3. Data descriptive analysis

After getting a quick overview of the NYC hotels data, we started the analysis by creating customer profiles so we could better understand what type of guests were staying at the hotels.

The full descriptive analysis is developed in part IV. of this report.

4. Data modeling techniques

After the descriptive analysis of the customers, we decided to use three modeling techniques in order to see how the Hyatt hotel chain could improve. The three algorithms used for the predictions were **linear regression**, **association rules**, and **support vector machines**.

- **Linear regression**: linear modeling is one of the most frequently used techniques in statistics where we investigate the potential relationship between a variable of interest (the dependent variable) and a set of one of more variables (the independent variables).
- **Association rules**: association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness
- **Support vector machines:** in machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

The objective of such modeling techniques was to make sure that our final recommendations would *not* be solely based on descriptive analysis but also on predictive statistics and analysis, as the Hyatt Hotels Corporation needs to improve.

The full data modeling & predictive analysis is developed in part V. of this report.

5. Data generalization

After getting the final recommendations from Professor Saltz to make sure our analysis was relevant and our algorithms were outputting statistically relevant results, we decided to re-run the code solely on the United States data from the full dataset that was provided at the beginning of our this project.

The data generalization analysis is developed in part VI. of this report.

IV. DATA DESCRIPTIVE ANALYSIS

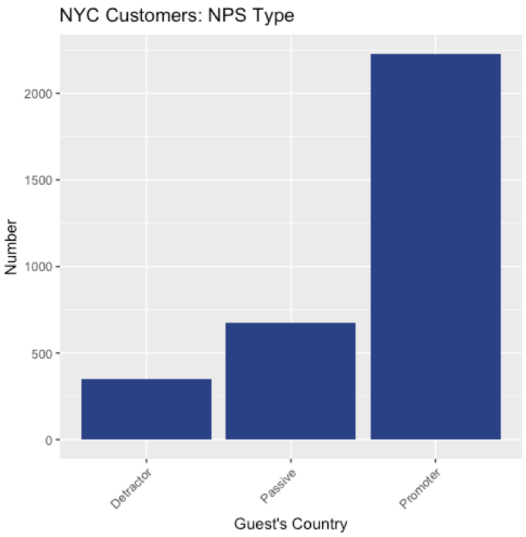
1. Basic descriptive analysis

 # STEP 8

Performing a descriptive analysis on the customer surveys data was a crucial step for us to understand who the customers were and better refine our analysis.

NPS Type:

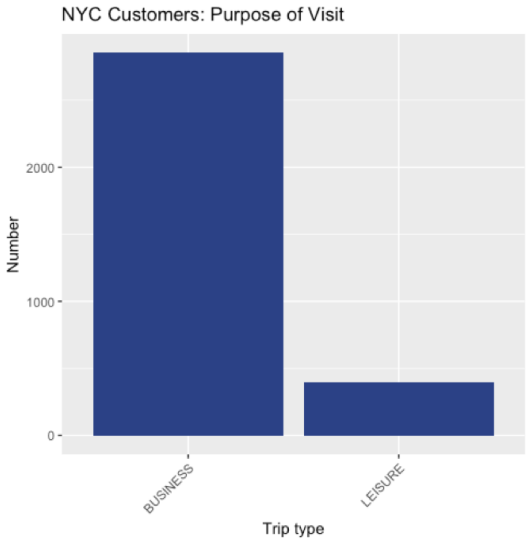
First of all, we found out that there were **2,228 promoters**, **675 passives**, and **350 detractors** among the NYC customers. That was a good sign, as the majority of the customers were actually very satisfied with the Hyatt Hotel chain. Only a fairly small portion of them were not satisfied, and understanding why was the purpose of our analysis.



	NPSType	Number
1	Promoter	2228
2	Passive	675
3	Detractor	350

Purpose of visit:

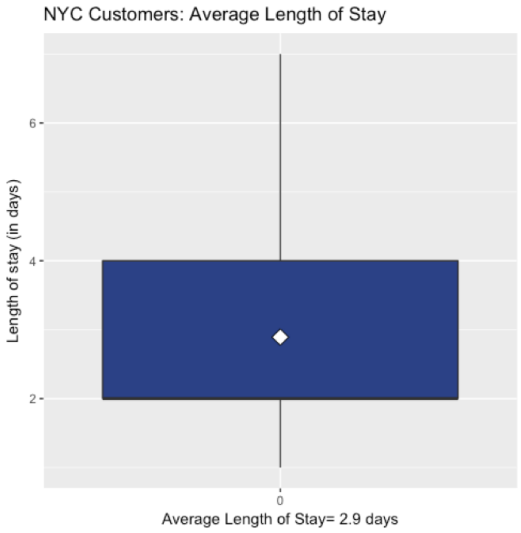
Then, we found out that **2,854** of the NYC customers were traveling for business, whereas **399** of them were traveling for leisure.



	PurposeOfVisit	Number
1	BUSINESS	2854
2	LEISURE	399

Length of stay:

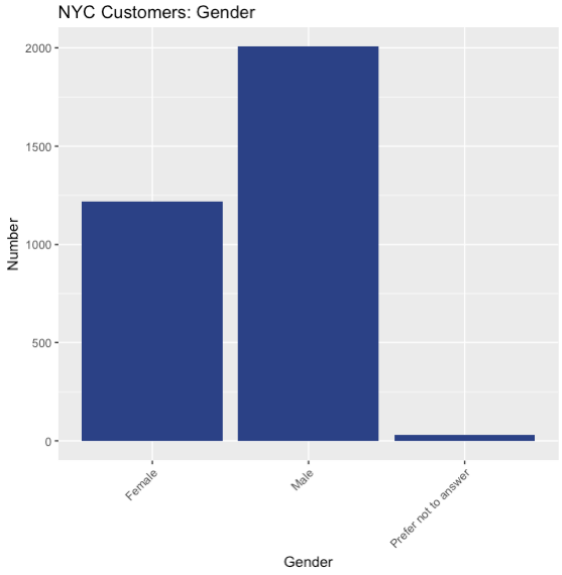
After computing the NYC customers' average length of stay, we found that that customers stayed **2.9** days on average. The average length of stay was coherent with our previous findings, as business trips tend to be short.



Gender:

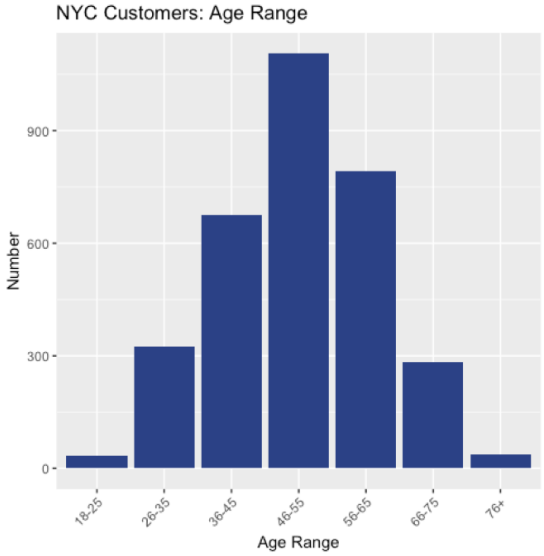
Then, we decided to look at gender. We found out that **2,005** of the NYC customers were men, **1,219** were women, and **29** preferred not to answer.

	Gender	Number
1	Male	2005
2	Female	1219
3	Prefer not to answer	29



Age:

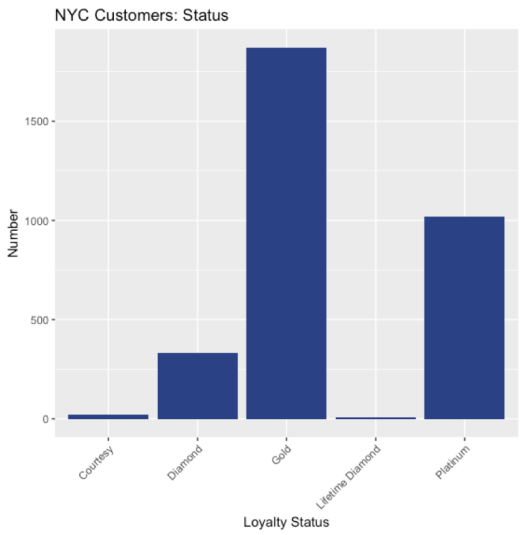
As for the NYC customers' age, the majority of them were in the **46-55 age range.** Not many customers were in the **18-25 age range**, which is important to consider as younger and older customers can have very different expectations when it comes to customer services and hotel amenities.



	AgeRange	Number
1	46-55	1105
2	56-65	791
3	36-45	676
4	26-35	326
5	66-75	284
6	76+	36
7	18-25	35

Member status:

After looking into the NYC customers' loyalty status, we found out that most of them were **Gold** and **Platinum** customers, which means that about **89%** of the customers (2,081 out of 3,225) were returning customers and already enrolled in the Hyatt Hotel loyalty program.

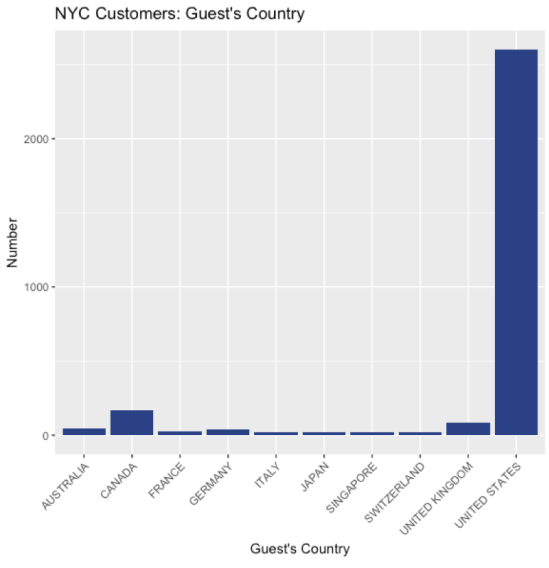


	Status	Number
1	Gold	1871
2	Platinum	1020
3	Diamond	331
4	Courtesy	23
5	Lifetime Diamond	8

Guest Country:

Finally, we decided to analyse the top-10 guests' countries to know more about where the customers came from. We found out that most of the customers were from the United States and Canada.

Culture is always a very important factor in customer service and it is crucial for any company to be sensitive to cultural differences. Every country tends to have different levels of expectations when it comes to quality and comfort, which is why deemed it relevant to also include Guest's country in the analysis. (Our four team members come from the U.S., the U.K., Nepal, and France.)



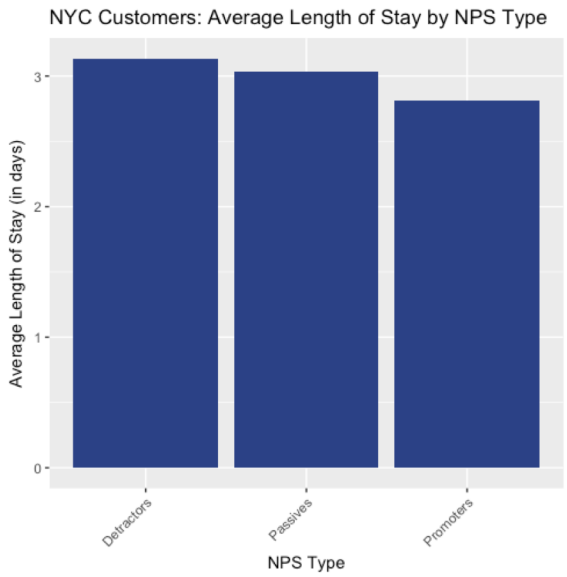
	GuestCountry	Number
1	UNITED STATES	2600
2	CANADA	168
3	UNITED KINGDOM	83
4	AUSTRALIA	42
5	GERMANY	37
6	FRANCE	23
7	SWITZERLAND	22
8	JAPAN	18
9	SINGAPORE	18
10	ITALY	17

2. Descriptive analysis according to NPS type

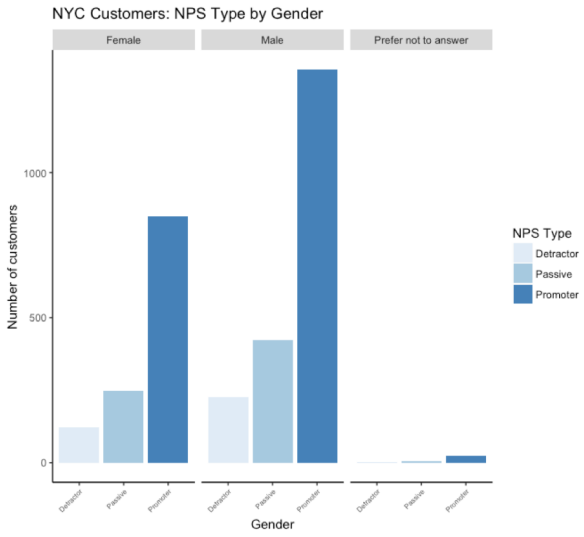
 # STEP 9

Finally, we performed most of the same descriptive analyses as above but this time combining the NPS type variable to really get a sense of who the unsatisfied customers were.[3]

1. Average length of stay by NPS Type

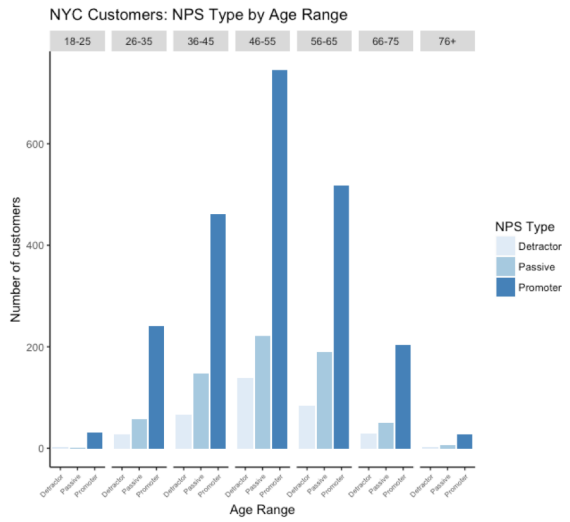


2. NPS type by Gender

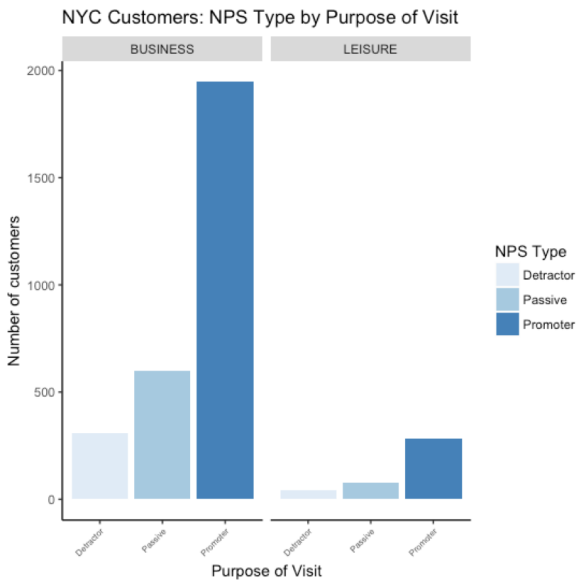


[3] The ggplot2 package came in handy to plot the results, even though we had to do a lot of online research to find ways to efficiently plot the results.

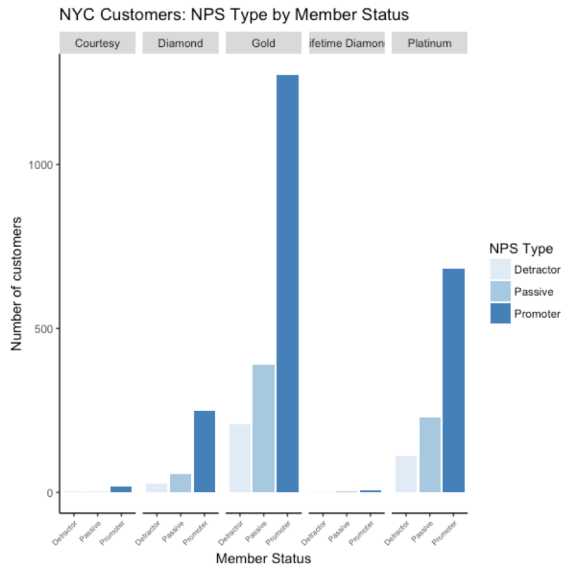
3. NPS type by Age Range



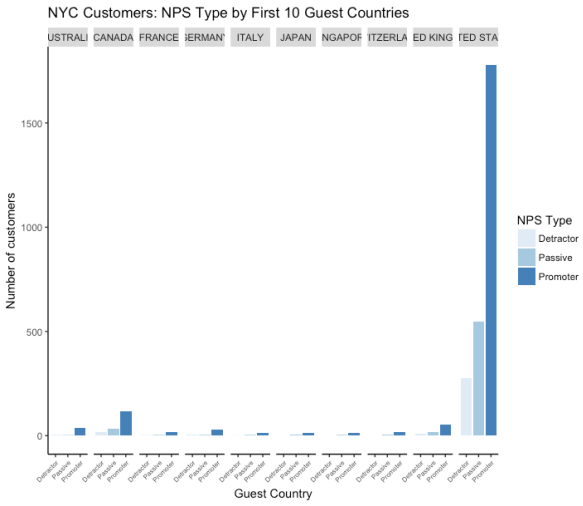
4. NPS type by Purpose of Visit



5. NPS type by Member Status



6. NPS type by (first 10) Guest Countries



Observations:

1. The NPS distribution seems to be proportional overall among the demographics.
2. NPS Type, seems not to be influenced by any demographics in particular, rather it reinforces our assumption that NPS Type is likely connected to the amenities and services provided by the establishments.
3. The observations above might be useful for marketing purposes to benefit the Hyatt Hotel Corporation (see below.)

First recommendations:

1. NYC is an international tourist hub, as such it is disappointing that the majority of the customers are U.S.-based travellers. The Hyatt Hotel Corporation should try to attract more international travellers.
2. Hyatt Hotel Corporation should consider a wider variety of demographics:
    a. Female travelers
    b. Leisure travelers
    c. Younger and older travelers besides (40-60)

V. DATA MODELING TECHNIQUES & PREDICTIVE ANALYSIS

1. Linear Regression # STEP 10

Using an analytical linear modeling (LM), (also known as linear regression) approach, we were able to analyze the customer service metrics relating to the survey scores—provided by users—that pertain to certain aspects of the customer experience. It was our goal to find if certain variables influence a customer's likelihood to recommend the Hyatt brand, which in turn would indicate their favorability toward the hotel chain. We also performed this analysis to see not just if certain factors influence a customer's decision to recommend the brand but also in what way, meaning to quantify, and show how certain factors of the experience have a provable influence on this decision, if at all.

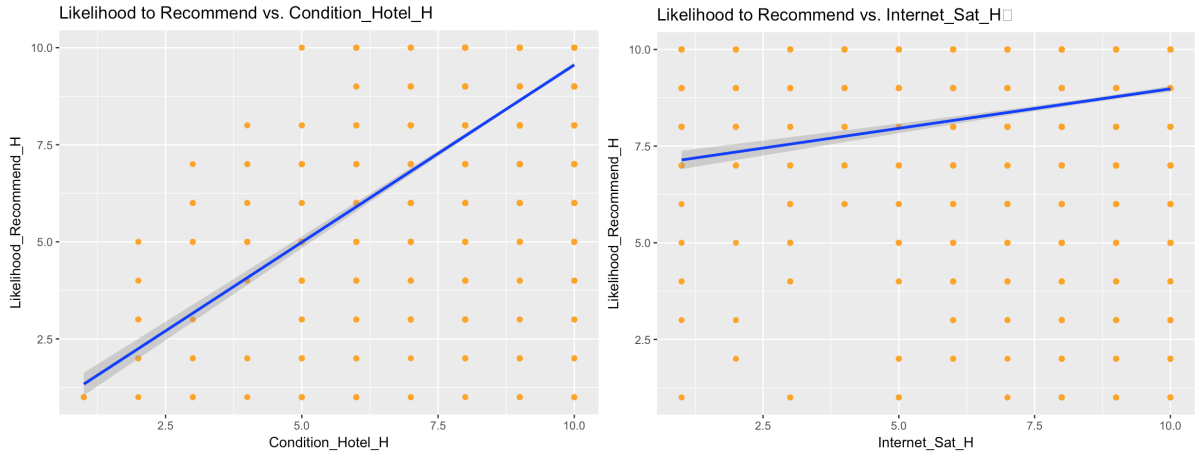
The steps we took in order to perform this analysis were to first determine which variables were associated with the customer experience, isolate the specific variables, and run a number of regression analyses to determine the most influential components of the customer experience in order to make recommendations/suggestions for the hotel management and staff, that will lead to further improvement and thus increase a customer's likelihood to recommend the Hyatt hotel brand. As data scientists, it was important to begin by seeking to disprove the null hypothesis, which is to assume that [x]variable(s) have no effect at all on the dependent variable (in this case, "Likelihood to Recommend").

After the dataset was refined, we began by running an exploratory analysis (using LM) on each individual variable containing a customer score/metric—as the independent variable—against the customer variable for likelihood to recommend—as the dependent variable—and plotted the results to render a visualization of the supposed relationship. We did this to see if we could preliminarily make inferences about what the most influential variables might be.

VARIABLES WE KEPT FOR THIS ANALYSIS & DEFINITIONS FOR EACH

- Check_In_H - Quality of the check in process metric; value on a 1 to 10 scale
- Condition_Hotel_H - Condition of hotel metric; value on a 1 to 10 scale
- Customer_SVC_H - Quality of customer service metric; value on a 1 to 10 scale
- F&B_Overall_Experience_H - Overall F&B experience metric; value on a 1 to 10 scale
- Guest_Room_H - Guest room satisfaction metric; value on a 1 to 10 scale
- Internet_Sat_H - Internet satisfaction metric; value on a 1 to 10 scale
- Likelihood_Recommend_H - Likelihood to recommend metric; value on a 1 to 10 scale
- Overall_Sat_H - Overall satisfaction metric; value on a 1 to 10 scale
- Staff_Cared_H - Staff cared metric; value on a 1 to 10 scale
- Tranquility_H - Tranquility metric; value on a 1 to 10 scale

Below are a few examples of single-variable regression analysis



Observation: As you can see from the examples above, upon first inspection, Condition_Hotel_H appears to be a better predictor than Internet_Sat_H based on the slope of the line and its intercept with the Y axis.

In plotting the results and running a regression to see statistical data, we made a few observations. We noticed that one variable we initially included, Overall_Sat_H (i.e., "Overall Satisfaction"), was such a strong predictor compared to all the other variables, that something must be awry. It was eventually determined that we should exclude this variable because it had to have been likely derived from various other metrics and will as a result unfairly influence/skew our results and negatively impact our modeling. (For example, the R-squared value indicates 79% of the model can be explained using this variable, which is unlikely in comparison to all the other variables which hovered well below 50%.

We also noticed a few variables seemed to indicate more influence on the dependent variable than others, so we made note of these findings and proceeded to investigate further.

EXPLORATORY ANALYSIS *(Preliminary observations based on our single linear regression)*

The ***best*** independent predictors seem to be:

1. Condition_Hotel_H
2. Customer_SVC_H
3. Staff_Cared_H
4. Guest_Room_H
5. Tranquility_H

The ***worst*** independent predictors seem to be:

1. Internet_Sat_H
2. F.B_Overall_Experience_H

3. Check_In_H

The next step we took was to perform the same such regressions, but utilize a multiple linear regression to see if combinations of variables had a collective influence on the dependent variable. We performed a number of these regressions using the variables that we suspected would be the best predictors and then again incorporating the suspected worst predictors, and those that fell somewhere in the middle, to see if we would find anything interesting when variables are taken in consideration in combination with each other. We also proceeded to plot these results.

Tested Likelihood against:

- First: **Condition_Hotel_H+ Guest_Room_H+ Tranquility_H**
- Second: **Condition_Hotel_H+ Guest_Room_H+ Staff_Cared_H**
- Third: **Condition_Hotel_H+ Guest_Room_H+ Tranquility_H+ Customer_SVC_H**

Below are a few examples of multi-variable linear regressions analysis


```
Call:
lm(formula = Likelihood_Recommend_H ~ Condition_Hotel_H + Customer_SVC_H +
    Staff_Cared_H, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4775 -0.3308  0.1866  0.4358  4.2530

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.43256    0.14551  -9.845  < 2e-16 ***
Condition_Hotel_H  0.48260    0.01863  25.911  < 2e-16 ***
Customer_SVC_H     0.47406    0.02491  19.030  < 2e-16 ***
Staff_Cared_H      0.16792    0.02194   7.655 2.53e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112 on 3249 degrees of freedom
Multiple R-squared:  0.6171,    Adjusted R-squared:  0.6167
F-statistic:  1745 on 3 and 3249 DF,  p-value: < 2.2e-16

> summary(TestComboPred0)$r.squared
[1] 0.6170771
> summary(TestComboPred0)$adj.r.squared
[1] 0.6167235
```

The third step we took was to use a method referred to as backward elimination, where we would run an algorithm that tests all possible combinations of given variables together, to identify the most parsimonious model, meaning the one with the best predictive quality using the fewest amount of variables (i.e., predictors). In making this determination, we looked at the Akaike Information Criterion (AIC), (an estimator of the relative quality of statistical models for a given set of data), coefficients (t-value and PR(<t)), and adjusted R-squared to support our analysis and ultimately inform our suggestions. Upon conducting this very comprehensive analysis, we notice that all but two variables, 'Internet_Sat_H' and 'Check_In_H', were significant (good predictors), so we removed these non significant variables from the model and conducted the analysis once again to see if we would get an even better model.

After performing our second, and last, backward elimination, we were able to conclude that we could confidently reject the null hypothesis, showing each variable indeed has a significant effect on the dependent variable. The algorithm showed the model fits the data quite well, as the residuals are symmetrically distributed, with a median close to zero, and 67.38% of the data can be explained using the most parsimonious model. One interesting note is that the final model contained all variables, showing that each, when combined together, gives us the best predictability.

Although some variables proved to be better predictors than others—as we suspected—it is this holistic experience that enables us to predict if a customer is likely to recommend the Hyatt brand.

The estimates columns show the following predictors have the most effect on the dependent variable. The best predictors (all of which are highly significant "***") are:

1. Customer_SVC_H (0.37917)
2. Guest_Room_H (0.29947)
3. Condition_Hotel_H (0.17368)
4. Staff_Cared_H (0.13692)
5. Tranquility_H (0.09873)
6. F.B_Overall_Experience_H (0.04791)

Example of backward elimination:

```
Step:  AIC=170.11
Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H + Condition_Hotel_H +
    Customer_SVC_H + Staff_Cared_H + Internet_Sat_H + F.B_Overall_Experience_H

                          Df Sum of Sq    RSS    AIC
<none>                                  3410.8 170.11
- Internet_Sat_H           1      2.21 3413.0 170.22
- F.B_Overall_Experience_H 1     17.56 3428.4 184.81
- Staff_Cared_H            1     45.78 3456.6 211.48
- Tranquility_H            1     54.73 3465.5 219.89
- Condition_Hotel_H        1     68.28 3479.1 232.59
- Customer_SVC_H           1    277.44 3688.3 422.50
- Guest_Room_H             1    328.73 3739.5 467.42

Call:
lm(formula = Likelihood_Recommend_H ~ Guest_Room_H + Tranquility_H +
    Condition_Hotel_H + Customer_SVC_H + Staff_Cared_H + Internet_Sat_H +
    F.B_Overall_Experience_H, data = dataset)

Coefficients:
             (Intercept)              Guest_Room_H             Tranquility_H         Condition_Hotel_H
                -1.44008                   0.29809                   0.09787                   0.17374
          Customer_SVC_H             Staff_Cared_H            Internet_Sat_H  F.B_Overall_Experience_H
                 0.37921                   0.13529                   0.01370                   0.04593
```

Recommendations and suggestions

With all of this in mind, we have compiled the following suggestions for the hotel management and staff to incorporate, in order to increase a customer's likelihood to recommend:

1. The hotel should focus on all of the variables above as they each have an effect on the Likelihood to recommend, but more emphasis should be put specifically on the top two variables: Customer_SVC_H and Guest_Room_H, as they have the most impact. Condition_Hotel_H and Staff_Cared_H should also be paid attention to, although they have around half of the influence as the top two variables.
2. Cultivate and maintain a strong customer-oriented experience by way of providing exceptional personal (positive employee-to-customer interaction) service. For example, if staff were able to increase just the Customer_SV_H score of an individual by three points, the influence that variable alone has on the likelihood to recommend, would increase its score by one whole point! Hence, this little change would potentially take a passive customer (on the cusp) into the promoter category, or a detractor into the passive category.
3. Put less emphasis on Tranquility_H and F.B_Overall_Experience_H as they have less of an effect, but don't disregard them altogether as they do have a little bit of influence.

2. Association Rules # STEP 11

For the association rules analysis, the given dataset was cleaned considerably. The most important variables were left in the dataset, while all other columns that were considered unnecessary were removed. The following are the columns that we utilised for association rules analysis:

- NPS_Type
- Guest_Room_H
- Tranquility_H
- Condiiton_Hotel_H
- Customer_SVC_H
- Staff_Cared_H
- Internet_Sat_H
- Check_In_H
- F.B_Overall_Experience_H

We cleaned the full dataset of Hyatt Hotels Corporations, we focused on the NYC hotels data for which 3,253 entries and nine columns remained. As a result, we focused on creating rules with a higher confidence level (we used 0.7 as the higher benchmark) in order to get statistically relevant rules.

	NPS_Type	Guest_Room_H	Tranquility_H	Condition_Hotel_H	Customer_SVC_H	Staff_Cared_H	Internet_Sat_H	Check_In_H	F.B_Overall_Experience_H
1	Promoter	10	10	10	10	10	10	10	10
2	Promoter	10	10	10	10	10	10	10	9
3	Detractor	10	10	10	9	9	3	10	9
4	Promoter	9	7	9	8	10	10	9	9
5	Detractor	6	4	9	7	9	9	9	9
6	Promoter	9	10	10	10	9	10	10	9
7	Promoter	10	7	10	10	9	10	10	10
8	Promoter	10	10	10	10	10	9	10	10
9	Promoter	10	10	10	10	10	10	10	10
10	Promoter	6	10	10	10	10	10	9	10
11	Promoter	10	10	10	9	9	10	10	8
12	Promoter	10	10	10	10	10	10	10	10
13	Promoter	9	4	9	10	10	7	9	9
14	Promoter	10	10	10	10	10	10	10	10
15	Promoter	10	10	10	10	10	10	10	10
16	Promoter	10	9	10	10	10	10	10	10
17	Promoter	10	10	10	10	10	10	10	10
18	Detractor	9	8	9	2	2	9	8	8
19	Promoter	10	10	10	10	10	10	10	3
20	Promoter	8	10	8	8	5	6	7	8
21	Passive	9	9	9	7	7	9	9	4
22	Promoter	10	8	10	10	10	10	10	9
23	Promoter	10	10	10	7	9	10	10	10

The screenshot above of the is the cleaned dataset (making sure Overall_Sat_H had been taken out, consistent with what we did in linear modeling.)

	NPS_Type	Guest_Room_H	Tranquility_H	Condition_Hotel_H	Customer_SVC_H	Staff_Cared_H	Internet_Sat_H	Check_In_H	F.B_Overall_Experience_H
1	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
2	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
3	Detractor	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_LOW	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
4	Promoter	Guest_Room_H_HIGH	Tranquility_H_MID	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
5	Detractor	Guest_Room_H_MID	Tranquility_H_LOW	Condition_Hotel_H_HIGH	Customer_SVC_H_MID	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
6	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
7	Promoter	Guest_Room_H_HIGH	Tranquility_H_MID	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
8	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
9	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
10	Promoter	Guest_Room_H_MID	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
11	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
12	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
13	Promoter	Guest_Room_H_HIGH	Tranquility_H_LOW	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_MID	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
14	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
15	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
16	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
17	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
18	Detractor	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_LOW	Staff_Cared_H_LOW	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
19	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_LOW
20	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_MID	Internet_Sat_H_MID	Check_In_H_MID	F.B_Overall_Experience_H_HIGH
21	Passive	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_MID	Staff_Cared_H_MID	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_LOW
22	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_HIGH	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH
23	Promoter	Guest_Room_H_HIGH	Tranquility_H_HIGH	Condition_Hotel_H_HIGH	Customer_SVC_H_MID	Staff_Cared_H_HIGH	Internet_Sat_H_HIGH	Check_In_H_HIGH	F.B_Overall_Experience_H_HIGH

The screenshot above shows the dataset after we transformed numerics into categorical data (i.e. new data type = factors) by creating three bins for each metric. Here is how the bins were created: 0-4 as LOW, 5-7 as MID, 8-10 as HIGH.

We used the apriori function to mine rules with the minimum support of 0.01.

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} = P(A \cap B)$$

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)}$$

$$ExpectedConfidence = \frac{Number\ of\ transactions\ with\ B}{Total\ number\ of\ transactions} = P(B)$$

$$Lift = \frac{Confidence}{Expected\ Confidence} = \frac{P(A \cap B)}{P(A).P(B)}$$

Below is a quick explanation of the aRules parameters we used:

- **Support** - is an indication of how frequent the item appears in the database
- **Confidence** - is to identify the most important relationships
    - Indicates the number of times if/then statement has been found true
- **Lift** - is the factor by which, the co-occurence of A and B exceeds the expected probability of A and B co-occurring. Higher the lift, higher the chance of A and B occuring together.
- **Maxlen** - the maximal length of itemsets/rules (which we restricted to 3 to make sure the algorithm would only output a small *and* relevant combination of itemsets under the LHS).

Here is a sample line of code that shows the parameter we picked to generate our rules.

- RulesPromoters = apriori(dataset, parameter=list(support=**0.01**, confidence=**0.8**, maxlen=**3**), appearance=list(rhs='**NPS_Type=Promoter**'))

After noticing that the rules generated for passives were not conclusive, we decided to follow Professor Krudys' recommendation and keep the analysis binary (i.e. focus on promoters and detractors *only*.) Indeed, we realized that in order to find rules for passive customers, we had to lower the confidence under 0.2. As a reminder, any rule that has a <0.5 confidence is not statistically relevant. Lastly, Professor Saltz, recommended that despite having a low confidence or support, the rules are still relevant because it tells us something about the quality of the dataset.

The following screenshots are the console output that illustrates the rules.

Rules for Promoters:

```
> inspect(RulesPromoters)
     lhs                                                         rhs                  support   confidence lift     count
[1]  {Tranquility_H=Tranquility_H_HIGH,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.5548724  0.8306489 1.212792  1805
[2]  {Guest_Room_H=Guest_Room_H_HIGH,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.5742392  0.8422002 1.229658  1868
[3]  {Staff_Cared_H=Staff_Cared_H_HIGH,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.5788503  0.8162115 1.191713  1883
[4]  {Customer_SVC_H=Customer_SVC_H_HIGH,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.5871503  0.8005029 1.168777  1910
[5]  {Guest_Room_H=Guest_Room_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.5616354  0.8091231 1.181363  1827
[6]  {Guest_Room_H=Guest_Room_H_HIGH,
      Tranquility_H=Tranquility_H_HIGH}                       => {NPS_Type=Promoter} 0.6151245  0.8197460 1.196873  2001
[7]  {Tranquility_H=Tranquility_H_HIGH,
      Staff_Cared_H=Staff_Cared_H_HIGH}                       => {NPS_Type=Promoter} 0.6126652  0.8321503 1.214984  1993
[8]  {Tranquility_H=Tranquility_H_HIGH,
      Customer_SVC_H=Customer_SVC_H_HIGH}                     => {NPS_Type=Promoter} 0.6228097  0.8185859 1.195179  2026
[9]  {Tranquility_H=Tranquility_H_HIGH,
      Check_In_H=Check_In_H_HIGH}                             => {NPS_Type=Promoter} 0.6108208  0.8080521 1.179800  1987
[10] {Guest_Room_H=Guest_Room_H_HIGH,
      Staff_Cared_H=Staff_Cared_H_HIGH}                       => {NPS_Type=Promoter} 0.6369505  0.8412505 1.228271  2072
[11] {Guest_Room_H=Guest_Room_H_HIGH,
      Customer_SVC_H=Customer_SVC_H_HIGH}                     => {NPS_Type=Promoter} 0.6474024  0.8252351 1.204888  2106
[12] {Guest_Room_H=Guest_Room_H_HIGH,
      Check_In_H=Check_In_H_HIGH}                             => {NPS_Type=Promoter} 0.6344912  0.8174257 1.193486  2064
```

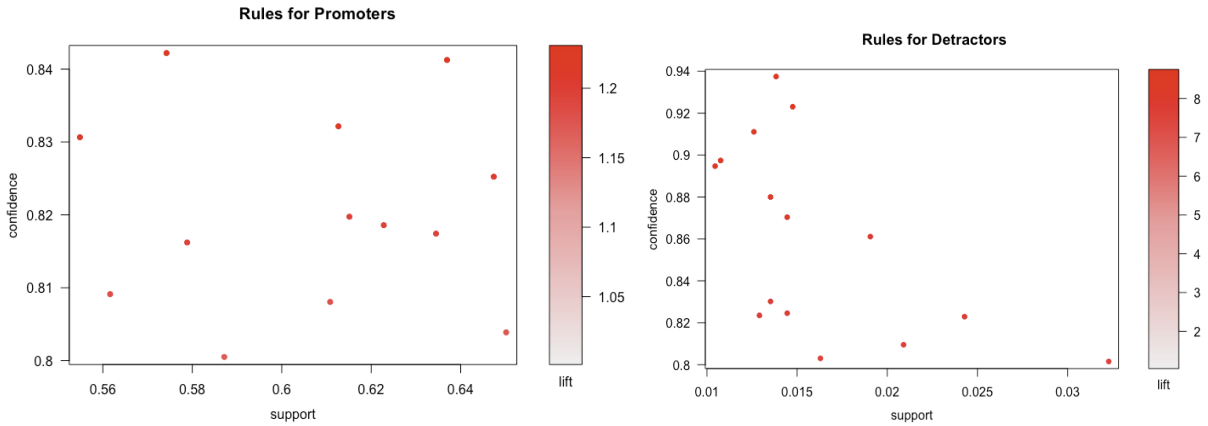
Rules for Detractors:

```
> inspect(RulesDetractors)
     lhs                                                                rhs                          support confidence     lift count
[1]  {Condition_Hotel_H=Condition_Hotel_H_LOW}                       => {NPS_Type=Detractor} 0.01045189  0.8947368 8.315940    34
[2]  {Customer_SVC_H=Customer_SVC_H_LOW}                            => {NPS_Type=Detractor} 0.01905933  0.8611111 8.003413    62
[3]  {Staff_Cared_H=Staff_Cared_H_LOW}                              => {NPS_Type=Detractor} 0.02428528  0.8229167 7.648423    79
[4]  {Guest_Room_H=Guest_Room_H_LOW}                                => {NPS_Type=Detractor} 0.03227790  0.8015267 7.449618   105
[5]  {Customer_SVC_H=Customer_SVC_H_LOW,
      Staff_Cared_H=Staff_Cared_H_LOW}                              => {NPS_Type=Detractor} 0.01475561  0.9230769 8.579341    48
[6]  {Customer_SVC_H=Customer_SVC_H_LOW,
      Internet_Sat_H=Internet_Sat_H_HIGH}                           => {NPS_Type=Detractor} 0.01444820  0.8703704 8.089471    47
[7]  {Condition_Hotel_H=Condition_Hotel_H_MID,
      Staff_Cared_H=Staff_Cared_H_LOW}                              => {NPS_Type=Detractor} 0.01260375  0.9111111 8.468127    41
[8]  {Guest_Room_H=Guest_Room_H_MID,
      Staff_Cared_H=Staff_Cared_H_LOW}                              => {NPS_Type=Detractor} 0.01075930  0.8974359 8.341026    35
[9]  {Guest_Room_H=Guest_Room_H_LOW,
      Tranquility_H=Tranquility_H_LOW}                              => {NPS_Type=Detractor} 0.01444820  0.8245614 7.663709    47
[10] {Guest_Room_H=Guest_Room_H_LOW,
      Condition_Hotel_H=Condition_Hotel_H_MID}                      => {NPS_Type=Detractor} 0.01629265  0.8030303 7.463593    53
[11] {Guest_Room_H=Guest_Room_H_LOW,
      Customer_SVC_H=Customer_SVC_H_MID}                            => {NPS_Type=Detractor} 0.01383338  0.9375000 8.713393    45
[12] {Guest_Room_H=Guest_Room_H_LOW,
      Staff_Cared_H=Staff_Cared_H_MID}                              => {NPS_Type=Detractor} 0.01352598  0.8301887 7.716011    44
[13] {Guest_Room_H=Guest_Room_H_LOW,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_MID}        => {NPS_Type=Detractor} 0.01352598  0.8800000 8.178971    44
```

Below is a screenshot of why we decided not to generate rules for passives. As you can see, the rules below are not relevant as the confidence is below 0.5.

```
> inspect(RulesPassives)
     lhs                                                                rhs                         support confidence     lift count
[1]  {Internet_Sat_H=Internet_Sat_H_LOW}                             => {NPS_Type=Passive} 0.01660006  0.3033708 1.462022    54
[2]  {Check_In_H=Check_In_H_MID}                                     => {NPS_Type=Passive} 0.02797418  0.3760331 1.812201    91
[3]  {Condition_Hotel_H=Condition_Hotel_H_MID}                       => {NPS_Type=Passive} 0.02920381  0.3429603 1.652815    95
[4]  {Customer_SVC_H=Customer_SVC_H_MID}                            => {NPS_Type=Passive} 0.03381494  0.3728814 1.797012   110
[5]  {Staff_Cared_H=Staff_Cared_H_MID}                              => {NPS_Type=Passive} 0.05256686  0.4373402 2.107656   171
[6]  {Guest_Room_H=Guest_Room_H_MID}                                => {NPS_Type=Passive} 0.05441131  0.4317073 2.080509   177
[7]  {Tranquility_H=Tranquility_H_MID}                              => {NPS_Type=Passive} 0.05564095  0.3867521 1.863859   181
[8]  {F.B_Overall_Experience_H=F.B_Overall_Experience_H_MID}        => {NPS_Type=Passive} 0.06394098  0.3573883 1.722347   208
[9]  {Internet_Sat_H=Internet_Sat_H_LOW,
      Check_In_H=Check_In_H_HIGH}                                   => {NPS_Type=Passive} 0.01444820  0.3012821 1.451956    47
[10] {Condition_Hotel_H=Condition_Hotel_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_LOW}                            => {NPS_Type=Passive} 0.01475561  0.3037975 1.464079    48
```

Below are the visualizations of the rules:



Observations:

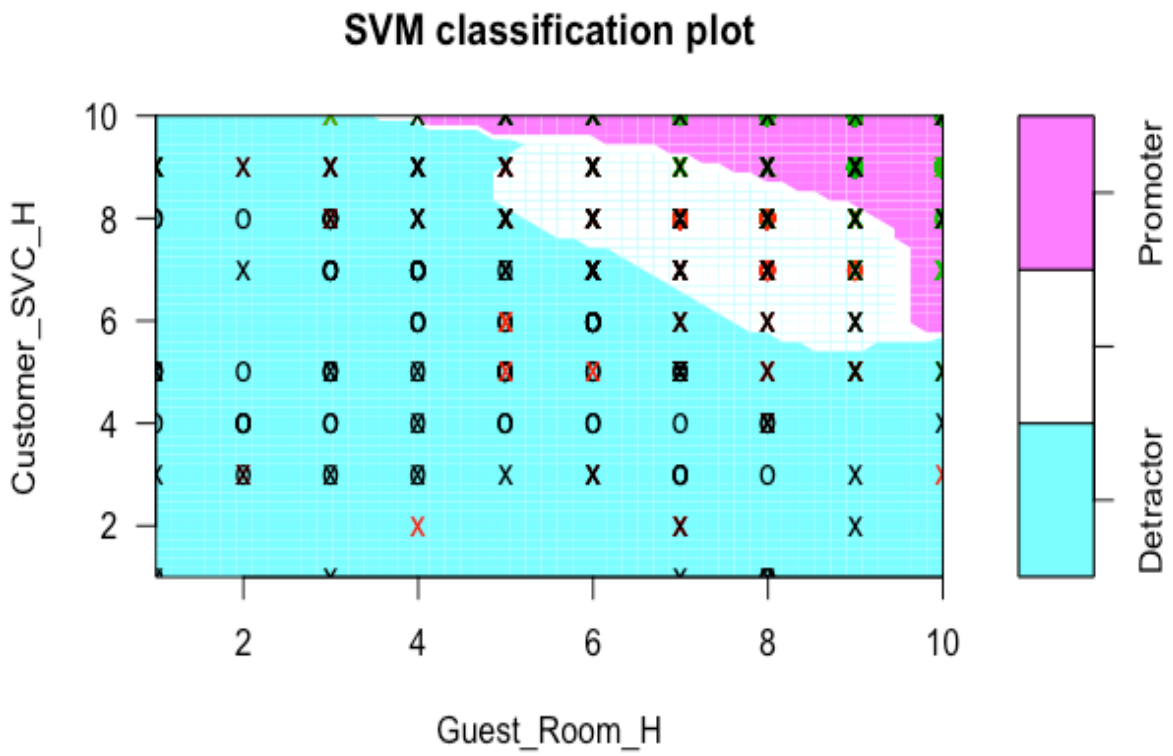
1. The aRules for detractors confirm what we found with linear regression, especially with respect to Customer_SVC_H, Guest_Room_H, Condition_Hotel_H and Staff_Cared_H.

which seem to be the most important variables that need to be focused on for areas of improvements.

2. The first set of rules found for detractors all show a very high lift, which demonstrates the reliability of the rules.

3. Support Vector Machines # STEP 12

Lastly, we used Support Vector Machines (SVMs) to confirm the predictions we found using the other methods above. By performing our predictive modeling to see if NPS_Type could be predicted by various variables in our dataset, we noticed that all the key variables—highly connected to predicting a satisfied customer—remained the same. We came to the conclusion that our other methods were sound and that our predictions were accurate. This last confirmation helped by reaffirming our findings, regardless of the model used, and ultimately supported the recommendations offered.



Below is a table of the correlation of the most important variables.

	Overall_Sat_H	Guest_Room_H	Tranquility_H	Condition_Hotel_H	Customer_SVC_H	Staff_Cared_H	Internet_Sat_H	Check_In_H	F.B_Overall_Experience_H
Overall_Sat_H	1.0000000	0.7385044	0.6088581	0.6801870	0.7445344	0.6784438	0.2401780	0.5003796	0.4065557
Guest_Room_H	0.7385044	1.0000000	0.6476170	0.7282980	0.5577648	0.5129001	0.2172725	0.3907463	0.3398665
Tranquility_H	0.6088581	0.6476170	1.0000000	0.5717308	0.5104585	0.4869508	0.1999145	0.3675522	0.3052259
Condition_Hotel_H	0.6801870	0.7282980	0.5717308	1.0000000	0.5815353	0.5516152	0.1950891	0.4594733	0.3627324
Customer_SVC_H	0.7445344	0.5577648	0.5104585	0.5815353	1.0000000	0.8367528	0.2118969	0.5685682	0.4154054
Staff_Cared_H	0.6784438	0.5129001	0.4869508	0.5516152	0.8367528	1.0000000	0.2243039	0.5357339	0.4247022
Internet_Sat_H	0.2401780	0.2172725	0.1999145	0.1950891	0.2118969	0.2243039	1.0000000	0.1703789	0.2194686
Check_In_H	0.5003796	0.3907463	0.3675522	0.4594733	0.5685682	0.5357339	0.1703789	1.0000000	0.2742025
F.B_Overall_Experience_H	0.4065557	0.3398665	0.3052259	0.3627324	0.4154054	0.4247022	0.2194686	0.2742025	1.0000000

VI. DATA GENERALIZATION

There is no specific STEP in our R code for this section. We have re-imported the full dataset, cleaned it to keep all the hotels from the United States, and re-run our algorithms on the new data. Below is the code we used to do so.

dataset <- data.frame(dataset[dataset$Country_PL == "United States",]) # to only keep the U.S. data. **The new dataset contained 115,585 rows instead of 3,253 (NYC dataset), which would give us more accurate results.**

After talking with Professor Saltz, we decided to re-run our total code on the full U.S. dataset as a last step. Our objective for running the code on the whole U.S. data was to validate our findings and recommendations. If the results were the same, our NYC findings would be validated. On the other hand, if the results were different to the NYC outcomes, it would be emphasis that NYC was not a representative of the Hyatt Hotels Corporation operations throughout the U.S.

1. Linear Regression

```
> summary(MostPar1) #Display summary of info

Call:
lm(formula = Likelihood_Recommend_H ~ ., data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0307 -0.1110  0.0276  0.4255  6.0733

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.869289   0.025695 -72.749   <2e-16 ***
Guest_Room_H             0.302767   0.003227  93.837   <2e-16 ***
Tranquility_H            0.116784   0.002286  51.090   <2e-16 ***
Condition_Hotel_H        0.187557   0.003543  52.944   <2e-16 ***
Customer_SVC_H           0.350175   0.004382  79.917   <2e-16 ***
Staff_Cared_H            0.113714   0.003764  30.213   <2e-16 ***
Internet_Sat_H           0.024311   0.001600  15.195   <2e-16 ***
Check_In_H              -0.001997   0.002776  -0.719    0.472
F.B_Overall_Experience_H  0.096690   0.002077  46.548   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.004 on 115576 degrees of freedom
Multiple R-squared:  0.6591,    Adjusted R-squared:  0.659
F-statistic: 2.793e+04 on 8 and 115576 DF,  p-value: < 2.2e-16

> summary(MostPar1)$r.squared
[1] 0.6590579
> summary(MostPar1)$adj.r.squared
[1] 0.6590343
```

Observations: We found the variables Internet_Sat_H & Check_In_H were still not good predictors of likelihood to recommend yet Internet_Sat_H was now considered significant. Everything else, including adjusted R-squared remained very similar to that of the predicted data using only the New York dataset.

2. Association Rules

We ran the same code to get aRules on the U.S. dataset. Below are the results.

Promoters:

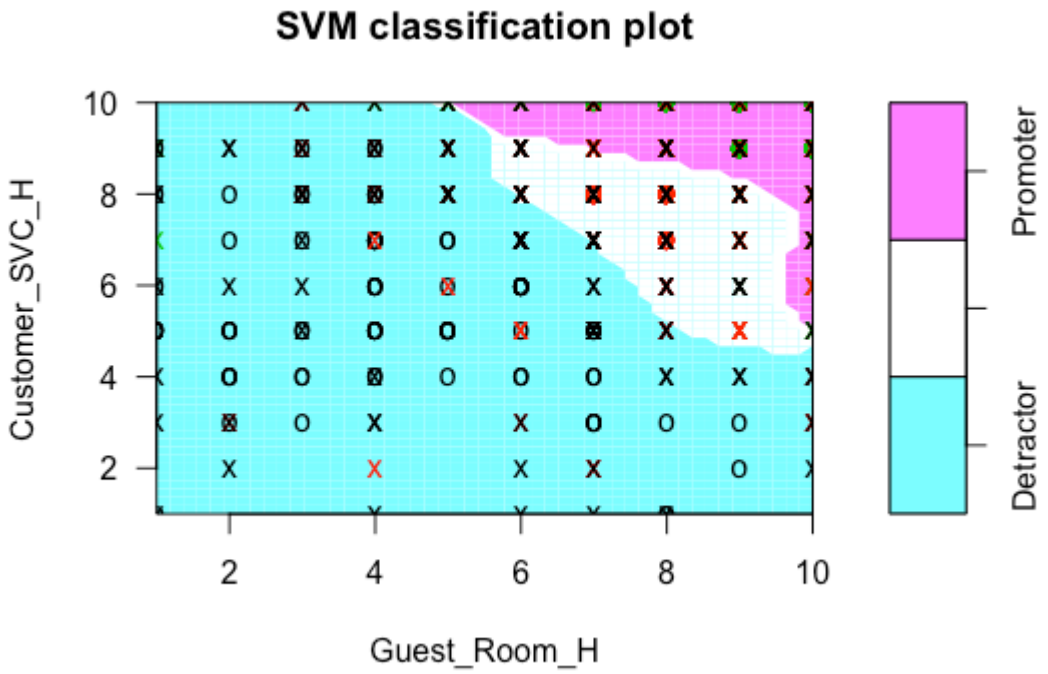
```
> inspect(RulesPromoters)
     lhs                                                            rhs                    support confidence     lift count
[1]  {F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.6682961  0.8020788 1.091945 77245
[2]  {Tranquility_H=Tranquility_H_HIGH}                       => {NPS_Type=Promoter} 0.6852879  0.8157467 1.110552 79209
[3]  {Guest_Room_H=Guest_Room_H_HIGH}                         => {NPS_Type=Promoter} 0.7110092  0.8194681 1.115618 82182
[4]  {Internet_Sat_H=Internet_Sat_H_HIGH,
      F.B_Overall_Experience_H=F.B_Overall_Experience_H_HIGH} => {NPS_Type=Promoter} 0.5865467  0.8288121 1.128339 67796
[5]  {Tranquility_H=Tranquility_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.6006575  0.8410399 1.144986 69427
[6]  {Guest_Room_H=Guest_Room_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.6215945  0.8448414 1.150161 71847
[7]  {Staff_Cared_H=Staff_Cared_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.6237488  0.8253409 1.123613 72096
[8]  {Condition_Hotel_H=Condition_Hotel_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.6281957  0.8261463 1.124710 72610
[9]  {Customer_SVC_H=Customer_SVC_H_HIGH,
      Internet_Sat_H=Internet_Sat_H_HIGH}                     => {NPS_Type=Promoter} 0.6333607  0.8186230 1.114468 73207
[10] {Internet_Sat_H=Internet_Sat_H_HIGH,
```

Detractors:

```
> inspect(RulesDetractors)
    lhs                                            rhs                    support confidence      lift count
[1] {Customer_SVC_H=Customer_SVC_H_LOW}         => {NPS_Type=Detractor} 0.01274387  0.8588921  9.735711  1473
[2] {Condition_Hotel_H=Condition_Hotel_H_LOW} => {NPS_Type=Detractor} 0.01442229  0.8764458  9.934686  1667
[3] {Guest_Room_H=Guest_Room_H_LOW}           => {NPS_Type=Detractor} 0.02228663  0.8157061  9.246190  2576
[4] {Guest_Room_H=Guest_Room_H_LOW,
     Condition_Hotel_H=Condition_Hotel_H_LOW} => {NPS_Type=Detractor} 0.01064152  0.9360731 10.610572  1230
[5] {Guest_Room_H=Guest_Room_H_LOW,
     Tranquility_H=Tranquility_H_LOW}         => {NPS_Type=Detractor} 0.01222477  0.8842303 10.022924  1413
```

Observations: As you can see, from the 5 rules generated, the results are mostly similar to the NYC results. However, the remainder of the rules generated for detractors from the U.S. dataset lacks the Staff_Cared_H variable. The difference of metrics in how staff cared might emphasize a difference in staff performance in N.Y.C. This should be further investigated by the Hotel chain.

3. Support Vector Machines



Observations: We re-ran SVMs on the full United States dataset, and we found similar results as before.

VII. DATA VALIDATION

Data validation was an important step throughout our project. We constantly needed to make sure that our results were meaningful and accurate. As a result, we used several techniques.

1. First of all, we checked one another's work and code on a regular basis as we had a shared folder on DropBox and Google Drive. We were able to download everyone's work, run the code, understand it and double-check it.
2. We also used different tools to make sure our calculations were always right (e.g. Microsoft Excel, calculators, and different formulas within R).
3. Our ggplot2 visualizations were always verified through Tableau Desktop, due to the ease of use of the program (i.e. lots of drag and dropping, and quick preliminary results.) We made compared the results and made sure our Tableau and R visualizations were similar.
4. Lastly, we also made sure to ask Professor Salz and Professor Krudys for guidance throughout the entire project, and always took their recommendations into consideration.

VIII. RESULTS

1. Final Conclusions

The analysis we conducted was thorough and we ensured our results were meaningful by putting forth a tireless effort to reaffirm our findings and ensure the information we received using advanced algorithms was both accurate and insightful by running a multitude of predictive models against the dataset.

2. Final Recommendations

Primary *(for quality purposes)*

Here are final recommendations to the Hyatt Hotels Corporation management:

- Focus on all of the variables relating to customer service but put the most emphasis specifically on the top two variables: Customer_SVC_H and Guest_Room_H.
- Put less emphasis on Tranquility_H and F.B_Overall_Experience_H as they have less of an effect, but do not disregard them altogether as they do have a little bit of influence.
- Do not focus on variables such as Internet_Sat_H and Check_in_H as they don't have much if any influence on customer satisfaction.
- Based on SVM Classification analysis Hyatt hotel should pay close attention on Customer service.
- Provide more customer service training for management and staff.
- Cultivate and maintain a strong customer-oriented experience by way of providing exceptional personal (positive employee-to-customer interaction) service.

Secondary *(for marketing purposes)*

- NYC is an international tourist hub, as such it is disappointing that the majority of the customers are U.S.-based travellers. The Hyatt Hotel Corporation should try to attract more international travellers.
- Hyatt Hotel Corporation should consider a wider variety of demographics:
    - Female travelers
    - Leisure travelers
    - Younger and older travelers besides (40-60)

3. Final Notes

Finally, we will add that the Hyatt Hotels Corporation should:

- …incentivise or invite the customers to be more responsive to its previous hotel stay surveys to overtly improve the hotel metrics.
- … refine their surveys and make sure to get more input in the specific areas where customers are unsatisfied (i.e. that make customers unsatisfied), as these areas are negatively impact the NPS. This will help them better target the unsatisfied guests and understand the reasons of their dissatisfaction.
- … always make sure to better organize their surveys and avoid having duplicates for the same question categories.
- ... keep in mind that collecting high quality data is the most important step in the analysis.