

Hyatt in the Holidays

A Net-promotor score analysis for Hyatt Group of hotels

Group 1| Section M006 IST 687 | 12/6/17

Team Members:

Sagarika Naik

Tanmay Atkekar

Jim Hwang

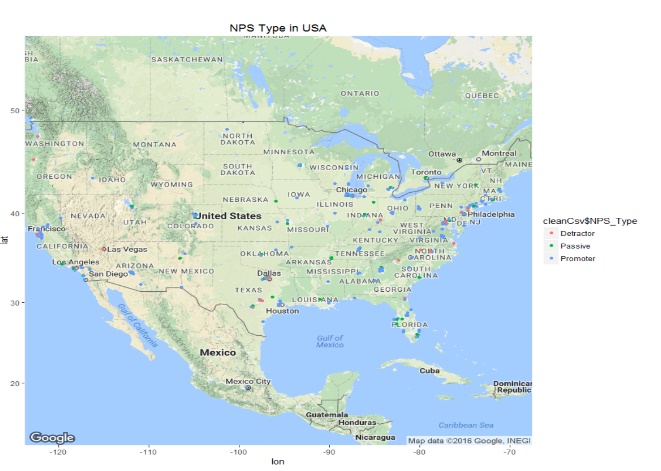
Mohit Shetty

Table of Contents

* Introduction
* Business Questions
* Data Acquisition, Cleansing, Transformation
* Descriptive statistics
* Graphs charts insights
* Use of modeling techniques
* Linear modeling
* Association rules
* SVM modeling
* Insights

# Introduction

A dataset containing twelve months of exhaustive data for the Hyatt Hotels chain collected from locations all over the world. The team was instructed to choose the business questions they would analyze, and provide actionable insights for the Hyatt Group. The underlying purpose was to analyze the Net-Promoter Scores provided by guests who answered surveys, and to provide suggestions to improve the likelihood to recommend a Hyatt Hotel stay. Upon viewing the data, we realized that the scope of the dataset was humongous. Analyzing twelve months of data would require relentless processing and a huge amount of time. As graduate students, we were short on both these resources: the processing power of our machines was limited to what paltry force our i7 machines could provide us, and time was a constraint as we had a little over ten weeks to complete the entire process which involved data collection, data cleansing, data munging, devising business questions, running statistical and descriptive analyses, Creating visualizations and reports, coming up with actionable insights and suggestions, and finally putting all our findings and graphs together in a comprehensive document and present it. We therefore decided to limit the scope of our dataset to three months spread across the financial year. However, performing analytics on these truncated datasets was becoming tedious. The files were still humongous and took more than fifteen minutes to just load on our computers. After many fruitless efforts, we decided to implement another approach: focusing on the United States, since most of the data indicated that a great percentage of the hotel locations were situated in the United States of America. To focus our analysis further, we decided to survey the dataset for the time which would be the best to perform analysis. Our criteria for deciding the month was simple: we needed a time of the year when people travelled a lot. Our second criterion was the number of surveys filled by customers. Behavioral science indicates that people are more likely to perform mundane tasks like filling customer satisfaction surveys when they are in high spirits. Keeping these things in mind, we chose the month of December- the month of holiday spirits and New Year revelry. To even out the dataset, we chose three prime locations: New York City, Los Angeles and Chicago.



The ggmap analysis performed on the United States of America indicates that a lot of reviewers in general are clustered in California, New York and Illinois. We will thus be analyzing the datasets for the hotels located in New York City, Los Angeles and Chicago, in the month of December ‘14. The team came up with numerous data questions that needed to be inspected. Our aim was to perform linear modeling and association rules modeling on the data sets to obtain rules and predictors. These models would then be used to provide actionable insights to the managers at Hyatt Corporations.

The key issues addressed in our analysis are enlisted here. We intend to perform location analysis to identify the city, state and country guests come from. This could provide an insight into trends. We will also be visualizing the results for better clarity. Another interesting metric is the trend in internet satisfaction provided by the customers. We would also perform descriptive and statistical analysis on the likelihood to recommend and all its inter-dependent metrics, in order to understand the underlying trends. Guest satisfaction and guest demographic analysis can be better understood using association rules and modeling. We would also be studying the trends of customer visits and customer satisfaction over the holidays like Christmas Eve, Christmas Day, and New Year’s Eve. The purpose of visit is also an interesting facet that would provide valuable insights for developing more promoters than usual. The analysis would also involve correlation between the age, gender and location of the guests with respect to their likelihood to recommend the Hyatt group to friends, family and colleagues. We would also be performing correlation analysis on check-in and check-out times and net revenue.

# Business Questions

The business questions that were addressed by us as a team are included in this section:

1. How are the customers reacting to the surveys?
2. Are the satisfaction metrics up-to-the mark?
3. How can the number of detractors be reduced?
4. How can the passives be converted to promoters?
5. How do we increase the likelihood to recommend Hyatt Hotels?
6. How does the guest’s location and demographic affect their satisfaction metrics?
7. How does nightly rate affect the likelihood to recommend?
8. How does purpose of visit affect likelihood to recommend?
9. How does a Free Independent Traveler react differently than a group of individuals on business meetings?
10. What is the trend of purpose of visit amongst different age groups, and how does it affect customer satisfaction?

# Data Acquisition, Cleansing, Transformation

The data acquisition process was straightforward. We were provided the complete set of data by the course instructors. We divided the data into subsets after an initial quality assignment. It was discovered that the most number of completed surveys came from large cities. We wanted to focus on a month when people of all age groups and affiliations traveled and were likely to stay in hotels. December is a time when families, corporate groups, and students all choose to travel over the holidays. Hence, the month of December was chosen for further analysis. It was a unanimous decision to focus only on the United States. The reasons behind this were purely psychological: Management in the United States is more likely to take into account suggestions made by amateur data scientists. Also, from a purely demographic point of view, it was convenient to factor in the various revenue variables in the American Currency, since most values had had to be converted into USD in the initial dataset anyway. The three cities chosen were Los Angeles, New York City and Chicago: these three locations are spread over the vast expanse of the United States and hence would help us consider the diversity in the data.

Hence, we filtered the data to retain only the observations which had “NYC”, “LA” and “Chicago” as City\_Pl. From the 238 variables that were available in the glossary, the team incorporated 33 variables in the analysis.

Performing a “str” analysis on the data set to obtain its structural information yielded these results:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 63986 obs. of  33 variables:

 $ CHECK\_IN\_DATE\_C       : POSIXct, format: "2014-11-28" "2014-12-01" "2014-12-01" ...

 $ CHECK\_OUT\_DATE\_C      : POSIXct, format: "2014-11-30" "2014-12-02" "2014-12-03" ...

 $ LENGTH\_OF\_STAY\_C      : num  2 1 2 1 1 1 1 1 1 1 ...

 $ ADULT\_NUM\_C           : num  1 1 1 1 1 1 1 1 1 2 ...

 $ CHILDREN\_NUM\_C        : num  0 0 0 0 0 0 0 0 0 0 ...

 $ POV\_CODE\_C            : chr  "BUSINESS" "BUSINESS" "BUSINESS" "BUSINESS" ...

 $ MAJOR\_MARKET\_CODE\_R   : chr  "TRANSIENT" "TRANSIENT" "CONVENTION" "TRANSIENT" ...

 $ NT\_RATE\_R             : chr  "0" "449" "415" "419" ...

 $ STATE\_R               : chr  "NJ" "AB" "AB" NA ...

 $ COUNTRY\_CODE\_R        : chr  "UNITED STATES" "CANADA" "CANADA" "HONG KONG" ...

 $ LENGTH\_OF\_STAY\_R      : num  1 1 2 2 2 2 2 1 1 3 ...

 $ ROOM\_NIGHTS\_R         : num  1 1 2 2 2 2 2 1 1 3 ...

 $ GROUPS\_VS\_FIT\_R       : chr  "FIT" "FIT" "Groups" "FIT" ...

 $ REVENUE\_USD\_R         : num  0 449 830 838 958 ...

 $ Survey\_ID\_H           : num  NA NA NA NA NA ...

 $ Length\_Stay\_H         : num  NA NA NA NA NA NA NA 1 NA NA ...

 $ Guest\_State\_H         : chr  NA NA NA NA ...

 $ Guest\_Country\_H       : chr  NA NA NA NA ...

 $ Gender\_H              : chr  NA NA NA NA ...

 $ Age\_Range\_H           : chr  NA NA NA NA ...

 $ Net\_Rev\_H             : num  NA NA NA NA NA ...

 $ Room\_Rev\_H            : num  NA NA NA NA NA ...

 $ Likelihood\_Recommend\_H: num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Overall\_Sat\_H         : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Guest\_Room\_H          : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Tranquility\_H         : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Condition\_Hotel\_H     : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Customer\_SVC\_H        : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Staff\_Cared\_H         : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Internet\_Sat\_H        : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Check\_In\_H            : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Hotel Name-Long\_PL    : chr  "Grand Hyatt New York" "Grand Hyatt New York" "Grand Hyatt New York" "Grand Hyatt New York" ...

 $ NPS\_Type              : chr  NA NA NA NA ...

As is evident from the structure analysis, we had to deal with the NAs that were prevalent in almost every variable. We had to employ various combinations for removing or replacing NAs as and when the analysis demanded it. We replaced the NA’s in the CHILDREN\_NUM\_C by zero, as it would not impact our analysis much. Our motive behind keeping the NAs was simple: data is valuable. In a dataset where most surveys were incompletely filled or not filled at all, we did not want to lose the little data that we did have. To understand how many people from different nationalities filled the surveys, we removed the NAs. This helped us to understand that people from specific countries or cities did not fill in surveys at all. This in turn, helps us to analyze and provide an actionable insight that the surveys be provided in a manner that is more specific to these individuals, so that they would feel inclined to fill in the surveys.

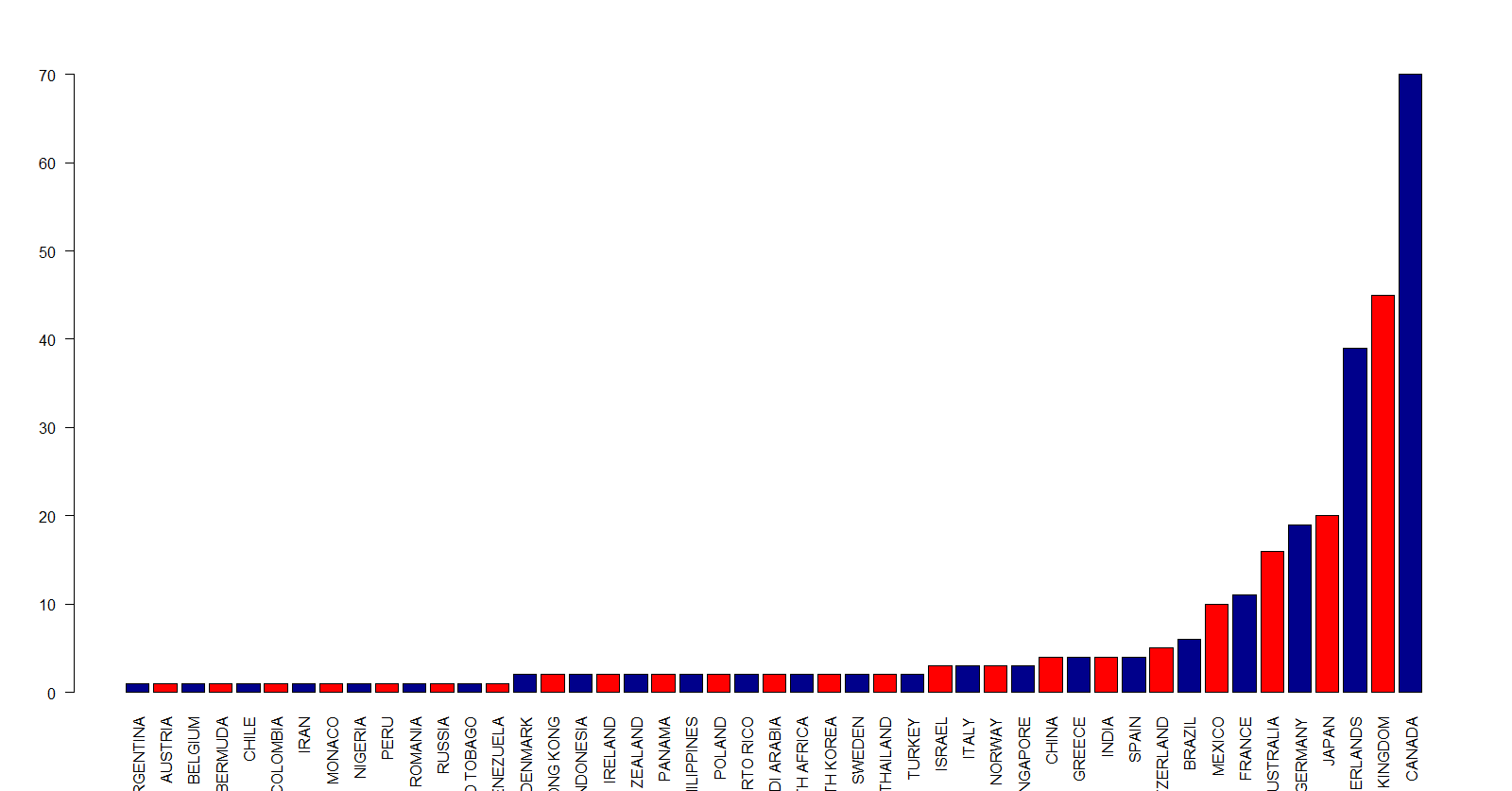
# Descriptive statistics

We plotted the number of individuals that filled complete surveys by their nationality. A plot was also employed for visitors by different countries to our three cities.

Visitors by country:



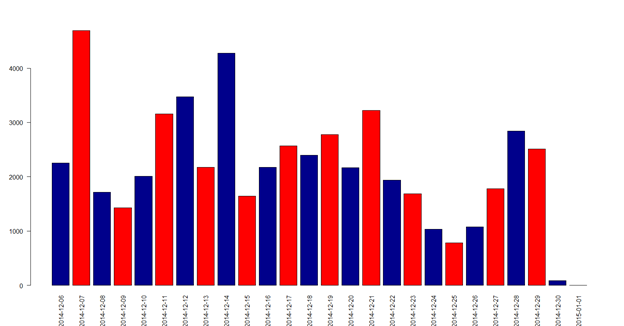
Surveys filled by Guest countries:



As is observed from the plots, most people who filled the surveys belonged to predominantly English speaking nations. Guests identifying with the Dutch nationality do not seem to be inclined towards filling in the surveys. The team thus recommends a survey in the Dutch language for these people from the Netherlands. This is because, the Hotel might be losing out on promoters due to the simple fact that they are not filling out the surveys. Also to service our East Asian customers better (particularly from Thailand, Japan, Singapore) we should make our survey in Thai, Japanese, and Chinese respectively.

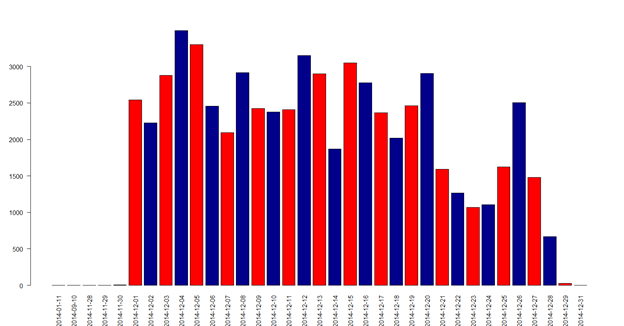
We then plotted the number of visitors checking in and checking out at all three cities over the period of one month. The check-out graph showed unusual spikes on the seventh and fourteenth of the month. On further analysis, we noticed that both these days were Sundays.

Checking out-

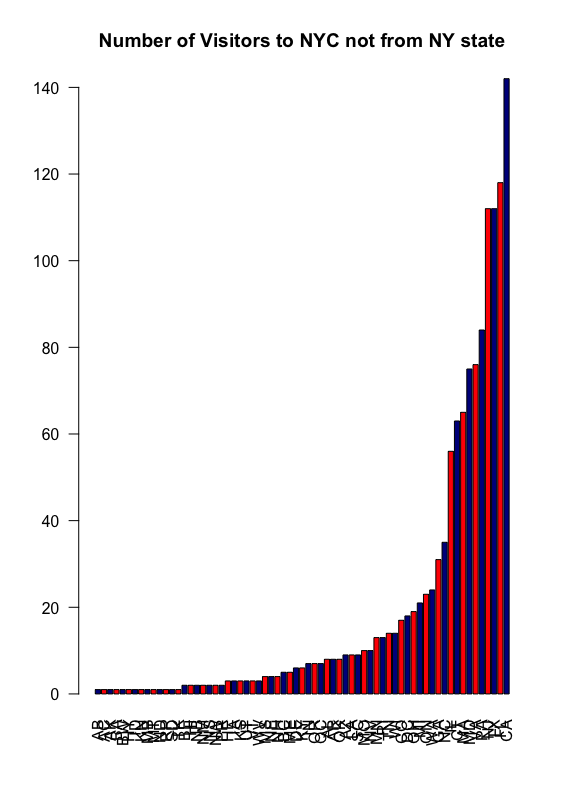
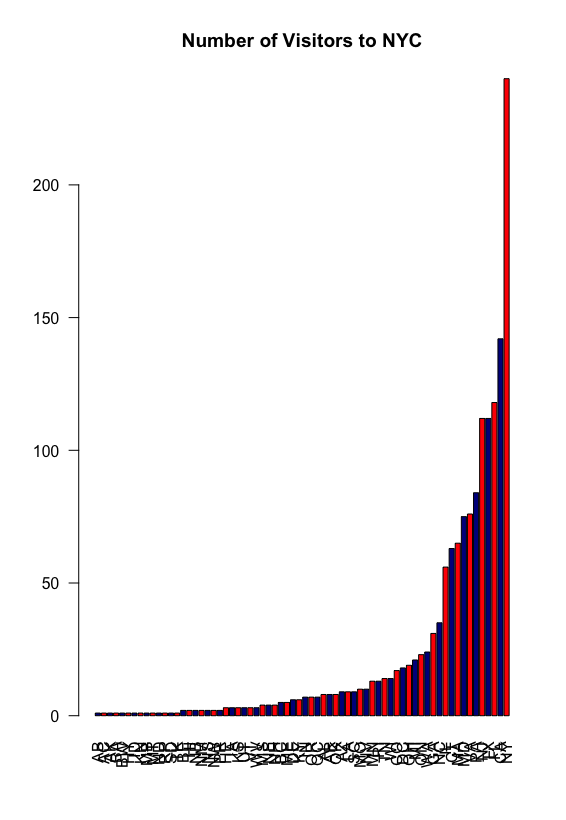


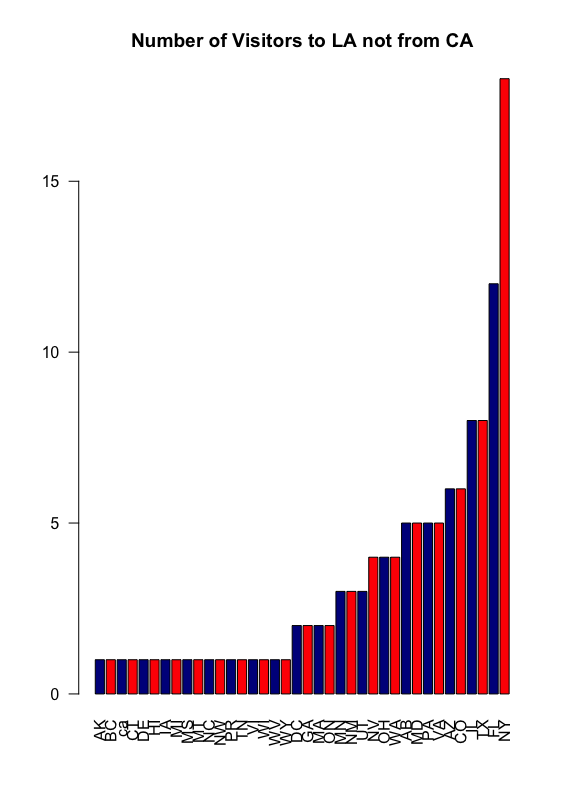
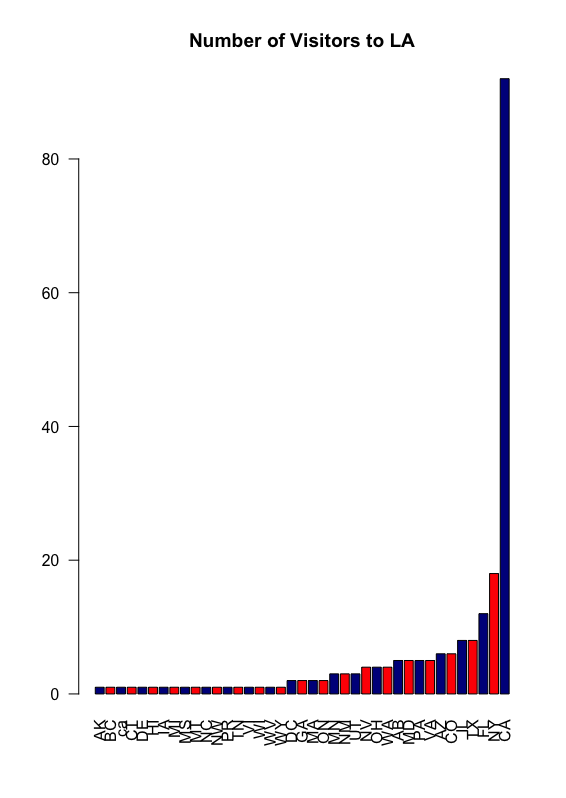
The check-in graph also provided an interesting observation: the number of visitors checking in was high throughout the month of December, with the exception of a few days. It did not come as a surprise that these days included Christmas Eve, Christmas Day, and the days preceding New Years’ Eve. Now, there could be many reasons behind this: mostly behavioral. However, the Hyatt group of hotels could introduce lucrative discounts or in-house parties to attract more customers.

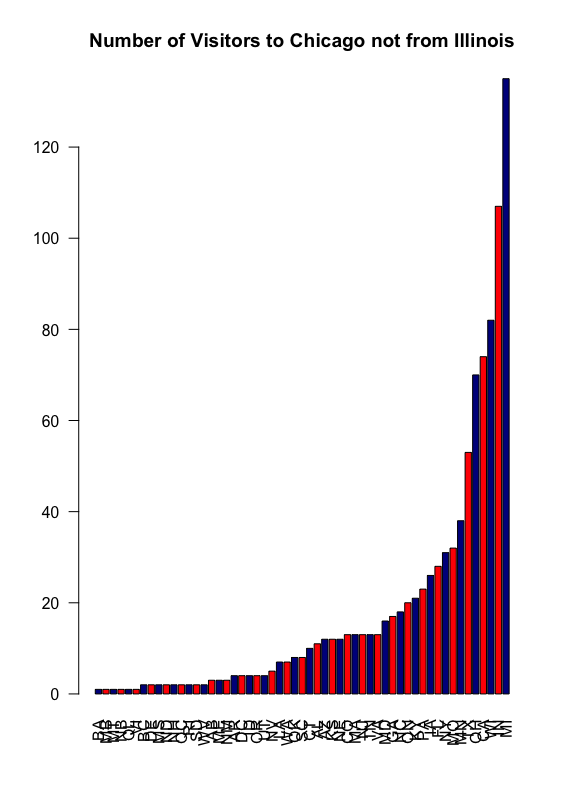
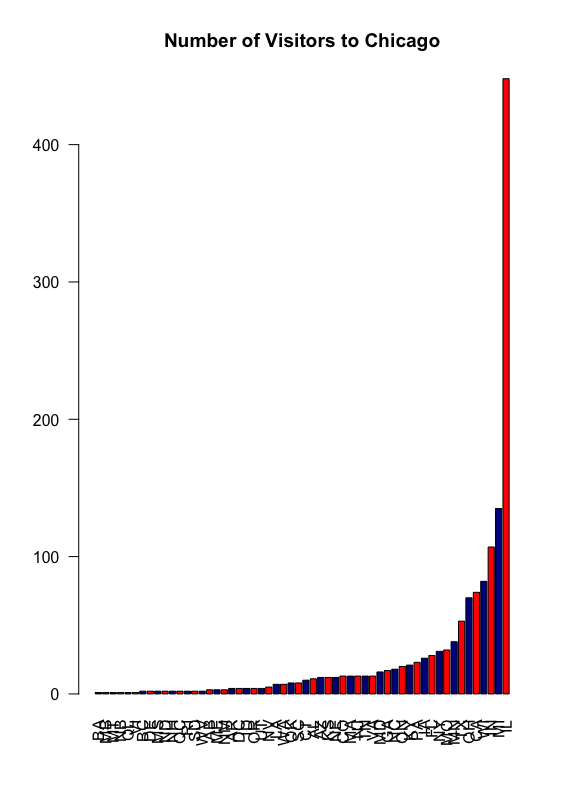
The Checking in graph is shown below.



Next, we plotted the number of visitors in each state separately. We also tried out an analysis just to see if the results would impact the analysis much. We plotted the number of visitors that were not natives of that state. For New York City and Los Angeles, a large number of visitors belonged to the home state. This is a surprising finding, as one would assume that people would want to travel farther in the holidays. These cities also had a lot of visitors from states spread all across the United States. In order to ensure guest satisfaction, the hotels should be designed in a cosmopolitan palette. In Chicago however, the guests come from the nearby states. Hence hotels in Chicago should have a mid-western flair to themselves. Molding themselves according to the customers’ tastes would ensure an increase in the likelihood to recommend.

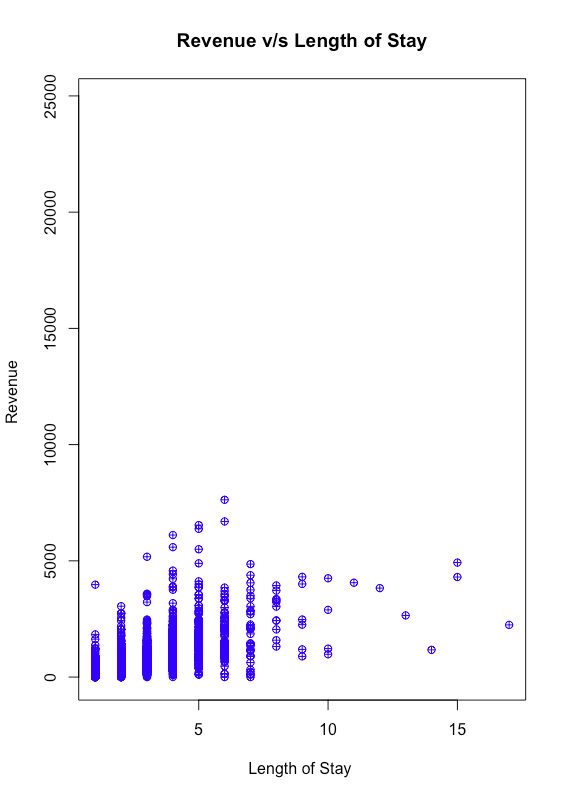




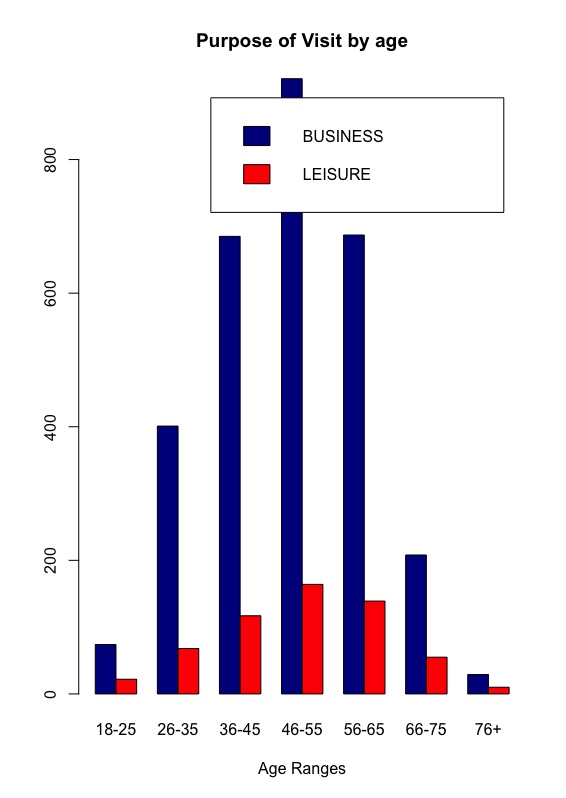


Although the graphs might indicate otherwise, on scrutiny it can be observed that the scales differ and hence, the insight is practicable.

A scatter plot to depict the length of stay and the revenue earned is depicted below. We can observe that stays lesser than or equal to the duration of a week generated a steady revenue in a huge number. This is indicative of the number of guests who are present in the hotels for attending conferences and conventions that are short three to five day events. They could also be guests who are free independent travelers passing by and in the need of accommodation for the short span of a night or two. The group could focus on improving the overall experience for these customers to enhance the net promoter scores.



We also plotted a bar plot to show purpose of visit against age. Its results are depicted below. Contrary to our estimates that the number of people traveling for leisure would be considerably higher than that of people travelling for business, the bar plot showed us the opposite. It was observed that the number of people travelling for business was highest in the age range of 46-55. This taught us a valuable lesson in data analytics: to never estimate any value or insight without concrete evidence to support it. So, instead of performing further descriptive analysis we decided to use modeling techniques for further exploration.



# Use of modeling techniques

* Linear modeling

We performed linear modeling to study the guests by their nationalities as predictors to likelihood to recommend. This included their state and their city.

* Linear model for state:

state\_likelihood <- lm(formula = dfall1$Likelihood\_Recommend\_H~dfall1$STATE\_R, data = dfall1)  
summary(state\_likelihood)

Call:  
lm(formula = dfall1$Likelihood\_Recommend\_H ~ dfall1$STATE\_R,   
    data = dfall1)  
Residuals:  
    Min      1Q  Median      3Q     Max   
-7.8896 -0.6164  0.7317  1.2738  2.7143   
   
Coefficients:  
                 Estimate Std. Error t value Pr(>|t|)      
(Intercept)        7.5556     0.6650  11.362  < 2e-16 \*\*\*

* Linear model for country-

summary(country\_likelihood)

 country\_likelihood <- lm(formula = dfall1$Likelihood\_Recommend\_H~dfall1$COUNTRY\_CODE\_R, data = dfall1)  
summary(country\_likelihood)

Call:  
lm(formula = dfall1$Likelihood\_Recommend\_H ~ dfall1$COUNTRY\_CODE\_R,   
    data = dfall1)  
Residuals:  
    Min      1Q  Median      3Q     Max   
-7.7000 -0.6564  0.3436  1.3436  2.8000   
Coefficients:  
                                           Estimate Std. Error t value Pr(>|t|)      
(Intercept)                               1.000e+01  2.011e+00   4.973 6.95e-07 \*\*\*

These values told us that location alone does not have a significant impact on likelihood to recommend. It is worth noting though that state has a higher R-squared than Country. We then used linear modeling on guest satisfaction metrics, likelihood to recommend and net promoter score.

> df\_lm<-dfall1[,c("Likelihood\_Recommend\_H","Net\_Rev\_H","Overall\_Sat\_H","Guest\_Room\_H","Tranquility\_H","Condition\_Hotel\_H",

+                  "Customer\_SVC\_H","Staff\_Cared\_H","Internet\_Sat\_H","Check\_In\_H")]

> View(df\_lm)

>

> lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$Guest\_Room\_H+df\_lm$Tranquility\_H,data=df\_lm)

> summary(lm\_combined)

Call:

lm(formula = df\_lm$Likelihood\_Recommend\_H ~ df\_lm$Guest\_Room\_H +

    df\_lm$Tranquility\_H, data = df\_lm)

Residuals:

    Min      1Q  Median      3Q     Max

-8.6426 -0.1330  0.1474  0.4470  5.6763

Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)          0.78716    0.15662   5.026 5.52e-07 \*\*\*

df\_lm$Guest\_Room\_H   0.69653    0.02190  31.801  < 2e-16 \*\*\*

df\_lm$Tranquility\_H  0.21002    0.01982  10.596  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 1751 degrees of freedom

  (62232 observations deleted due to missingness)

Multiple R-squared:  0.6031, Adjusted R-squared:  0.6026

F-statistic:  1330 on 2 and 1751 DF,  p-value: < 2.2e-16

>

> #Use-64%

> lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$+df\_lm$Condition\_Hotel\_H,data=df\_lm)

Error: unexpected '+' in "lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$+"

> summary(lm\_combined)

Call:

lm(formula = df\_lm$Likelihood\_Recommend\_H ~ df\_lm$Guest\_Room\_H +

    df\_lm$Tranquility\_H, data = df\_lm)

Residuals:

    Min      1Q  Median      3Q     Max

-8.6426 -0.1330  0.1474  0.4470  5.6763

Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)          0.78716    0.15662   5.026 5.52e-07 \*\*\*

df\_lm$Guest\_Room\_H   0.69653    0.02190  31.801  < 2e-16 \*\*\*

df\_lm$Tranquility\_H  0.21002    0.01982  10.596  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 1751 degrees of freedom

  (62232 observations deleted due to missingness)

Multiple R-squared:  0.6031, Adjusted R-squared:  0.6026

F-statistic:  1330 on 2 and 1751 DF,  p-value: < 2.2e-16

>

> #Use-69%

> lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$Guest\_Room\_H+df\_lm$Customer\_SVC\_H,data=df\_lm)

> summary(lm\_combined)

Call:

lm(formula = df\_lm$Likelihood\_Recommend\_H ~ df\_lm$Guest\_Room\_H +

    df\_lm$Customer\_SVC\_H, data = df\_lm)

Residuals:

    Min      1Q  Median      3Q     Max

-8.9114 -0.2929  0.0886  0.5186  5.0944

Coefficients:

                     Estimate Std. Error t value Pr(>|t|)

(Intercept)          -0.29623    0.10381  -2.854  0.00435 \*\*

df\_lm$Guest\_Room\_H    0.55620    0.01242  44.796  < 2e-16 \*\*\*

df\_lm$Customer\_SVC\_H  0.46457    0.01394  33.335  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.078 on 3467 degrees of freedom

  (60516 observations deleted due to missingness)

Multiple R-squared:  0.6994, Adjusted R-squared:  0.6992

F-statistic:  4033 on 2 and 3467 DF,  p-value: < 2.2e-16

>

> #Use-67%

> lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$Guest\_Room\_H+df\_lm$Staff\_Cared\_H,data=df\_lm)

> summary(lm\_combined)

Call:

lm(formula = df\_lm$Likelihood\_Recommend\_H ~ df\_lm$Guest\_Room\_H +

    df\_lm$Staff\_Cared\_H, data = df\_lm)

Residuals:

    Min      1Q  Median      3Q     Max

-8.5051 -0.1922  0.0678  0.4949  4.5060

Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)         -0.48554    0.15772  -3.079  0.00211 \*\*

df\_lm$Guest\_Room\_H   0.61469    0.01816  33.847  < 2e-16 \*\*\*

df\_lm$Staff\_Cared\_H  0.42709    0.01948  21.925  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.128 on 1767 degrees of freedom

  (62216 observations deleted due to missingness)

Multiple R-squared:  0.6707, Adjusted R-squared:  0.6703

F-statistic:  1799 on 2 and 1767 DF,  p-value: < 2.2e-16

>

> #Use-67%

> lm\_combined<-lm(formula=df\_lm$Likelihood\_Recommend\_H~df\_lm$Guest\_Room\_H+df\_lm$Staff\_Cared\_H,data=df\_lm)

> summary(lm\_combined)

Call:

lm(formula = df\_lm$Likelihood\_Recommend\_H ~ df\_lm$Guest\_Room\_H +

    df\_lm$Staff\_Cared\_H, data = df\_lm)

Residuals:

    Min      1Q  Median      3Q     Max

-8.5051 -0.1922  0.0678  0.4949  4.5060

Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)         -0.48554    0.15772  -3.079  0.00211 \*\*

df\_lm$Guest\_Room\_H   0.61469    0.01816  33.847  < 2e-16 \*\*\*

df\_lm$Staff\_Cared\_H  0.42709    0.01948  21.925  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

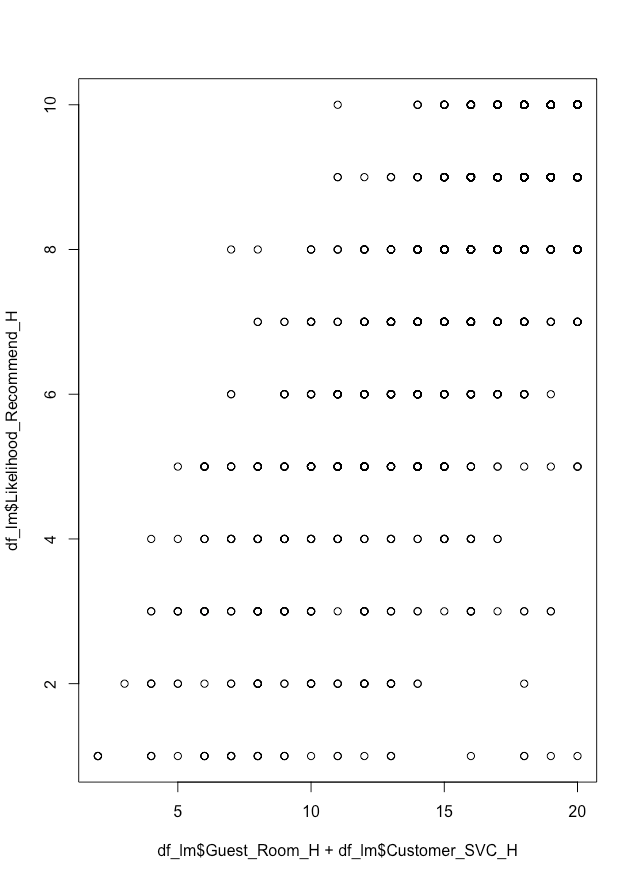
Residual standard error: 1.128 on 1767 degrees of freedom

  (62216 observations deleted due to missingness)

Multiple R-squared:  0.6707, Adjusted R-squared:  0.6703

F-statistic:  1799 on 2 and 1767 DF,  p-value: < 2.2e-16

An important observation is that no one factor alone (other than overall satisfaction) is dominant enough to drive Likelihood to recommend. When we pair two metrics together, they improve the correlation between them and LTR dramatically. As we can see from the data, the customer service and guest room metrics play a major role in higher likelihood to recommend. This model shows a high degree of correlation and is highlighted among the other models.



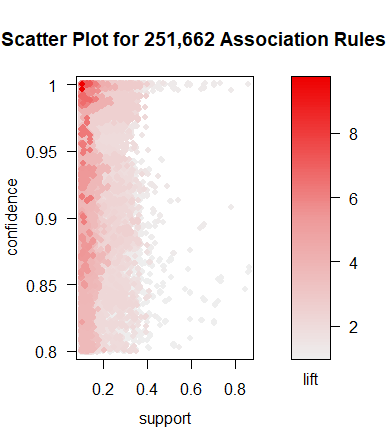
Apart from this scatter plot, all the plots were uniform and irrelevant to an extent.

* Association rules

We created an association model for predicting Net Promoter Score. The model yielded forty five good rules which are depicted here.

* > inspect(head(goodrules,50))
* lhs rhs support confidence lift count
* [1] {LENGTH\_OF\_STAY\_C=4} => {Length\_Stay\_H=4} 0.1024265 1.0000000 9.718821 439
* [2] {Length\_Stay\_H=4} => {LENGTH\_OF\_STAY\_C=4} 0.1024265 0.9954649 9.718821 439
* [3] {GROUPS\_VS\_FIT\_R=Groups} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1402240 1.0000000 7.131448 601
* [4] {MAJOR\_MARKET\_CODE\_R=CONVENTION} => {GROUPS\_VS\_FIT\_R=Groups} 0.1402240 1.0000000 7.131448 601
* [5] {GROUPS\_VS\_FIT\_R=Groups,
* City\_PL=Chicago} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1070929 1.0000000 7.131448 459
* [6] {MAJOR\_MARKET\_CODE\_R=CONVENTION,
* City\_PL=Chicago} => {GROUPS\_VS\_FIT\_R=Groups} 0.1070929 1.0000000 7.131448 459
* [7] {COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=Groups} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1269249 1.0000000 7.131448 544
* [8] {MAJOR\_MARKET\_CODE\_R=CONVENTION,
* COUNTRY\_CODE\_R=UNITED STATES} => {GROUPS\_VS\_FIT\_R=Groups} 0.1269249 1.0000000 7.131448 544
* [9] {POV\_CODE\_C=BUSINESS,
* GROUPS\_VS\_FIT\_R=Groups} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1399907 1.0000000 7.131448 600
* [10] {POV\_CODE\_C=BUSINESS,
* MAJOR\_MARKET\_CODE\_R=CONVENTION} => {GROUPS\_VS\_FIT\_R=Groups} 0.1399907 1.0000000 7.131448 600
* [11] {GROUPS\_VS\_FIT\_R=Groups,
* Guest\_Country\_H=USA} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1278581 1.0000000 7.131448 548
* [12] {MAJOR\_MARKET\_CODE\_R=CONVENTION,
* Guest\_Country\_H=USA} => {GROUPS\_VS\_FIT\_R=Groups} 0.1278581 1.0000000 7.131448 548
* [13] {POV\_CODE\_C=BUSINESS,
* GROUPS\_VS\_FIT\_R=Groups,
* City\_PL=Chicago} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1068595 1.0000000 7.131448 458
* [14] {POV\_CODE\_C=BUSINESS,
* MAJOR\_MARKET\_CODE\_R=CONVENTION,
* City\_PL=Chicago} => {GROUPS\_VS\_FIT\_R=Groups} 0.1068595 1.0000000 7.131448 458
* [15] {POV\_CODE\_C=BUSINESS,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=Groups} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1266916 1.0000000 7.131448 543
* [16] {POV\_CODE\_C=BUSINESS,
* MAJOR\_MARKET\_CODE\_R=CONVENTION,
* COUNTRY\_CODE\_R=UNITED STATES} => {GROUPS\_VS\_FIT\_R=Groups} 0.1266916 1.0000000 7.131448 543
* [17] {COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=Groups,
* Guest\_Country\_H=USA} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1259916 1.0000000 7.131448 540
* [18] {MAJOR\_MARKET\_CODE\_R=CONVENTION,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_Country\_H=USA} => {GROUPS\_VS\_FIT\_R=Groups} 0.1259916 1.0000000 7.131448 540
* [19] {POV\_CODE\_C=BUSINESS,
* GROUPS\_VS\_FIT\_R=Groups,
* Guest\_Country\_H=USA} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1276248 1.0000000 7.131448 547
* [20] {POV\_CODE\_C=BUSINESS,
* MAJOR\_MARKET\_CODE\_R=CONVENTION,
* Guest\_Country\_H=USA} => {GROUPS\_VS\_FIT\_R=Groups} 0.1276248 1.0000000 7.131448 547
* [21] {POV\_CODE\_C=BUSINESS,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=Groups,
* Guest\_Country\_H=USA} => {MAJOR\_MARKET\_CODE\_R=CONVENTION} 0.1257583 1.0000000 7.131448 539
* [22] {POV\_CODE\_C=BUSINESS,
* MAJOR\_MARKET\_CODE\_R=CONVENTION,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_Country\_H=USA} => {GROUPS\_VS\_FIT\_R=Groups} 0.1257583 1.0000000 7.131448 539
* [23] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [24] {COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_State\_H=IL} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [25] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_State\_H=IL} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [26] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [27] {COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [28] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA} => {STATE\_R=IL} 0.1101260 0.9915966 6.682364 472
* [29] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* NPS\_Type=Promoter} => {STATE\_R=IL} 0.1017266 0.9909091 6.677730 436
* [30] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA,
* NPS\_Type=Promoter} => {STATE\_R=IL} 0.1017266 0.9909091 6.677730 436
* [31] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* City\_PL=Chicago} => {STATE\_R=IL} 0.1250583 0.9889299 6.664392 536
* [32] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA,
* City\_PL=Chicago} => {STATE\_R=IL} 0.1250583 0.9889299 6.664392 536
* [33] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL} => {STATE\_R=IL} 0.1409239 0.9885434 6.661788 604
* [34] {COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA} => {STATE\_R=IL} 0.1409239 0.9885434 6.661788 604
* [35] {POV\_CODE\_C=BUSINESS,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL} => {STATE\_R=IL} 0.1110593 0.9855072 6.641327 476
* [36] {POV\_CODE\_C=BUSINESS,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_State\_H=IL,
* Guest\_Country\_H=USA} => {STATE\_R=IL} 0.1110593 0.9855072 6.641327 476
* [37] {STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_Country\_H=USA,
* City\_PL=Chicago} => {Guest\_State\_H=IL} 0.1250583 0.9745455 6.319065 536
* [38] {STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* City\_PL=Chicago} => {Guest\_State\_H=IL} 0.1250583 0.9710145 6.296170 536
* [39] {STATE\_R=IL,
* Guest\_Country\_H=USA,
* City\_PL=Chicago} => {Guest\_State\_H=IL} 0.1250583 0.9710145 6.296170 536
* [40] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_Country\_H=USA} => {Guest\_State\_H=IL} 0.1101260 0.9672131 6.271521 472
* [41] {STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_Country\_H=USA} => {Guest\_State\_H=IL} 0.1101260 0.9672131 6.271521 472
* [42] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* GROUPS\_VS\_FIT\_R=FIT,
* Guest\_Country\_H=USA} => {Guest\_State\_H=IL} 0.1101260 0.9672131 6.271521 472
* [43] {STATE\_R=IL,
* COUNTRY\_CODE\_R=UNITED STATES,
* Guest\_Country\_H=USA,
* NPS\_Type=Promoter} => {Guest\_State\_H=IL} 0.1017266 0.9667406 6.268457 436
* [44] {STATE\_R=IL,
* City\_PL=Chicago} => {Guest\_State\_H=IL} 0.1250583 0.9640288 6.250873 536
* [45] {MAJOR\_MARKET\_CODE\_R=TRANSIENT,
* STATE\_R=IL,
* Guest\_Country\_H=USA} => {Guest\_State\_H=IL} 0.1103593 0.9633401 6.246408 473

In the 29th result, STATE\_R=IL is 6.68 times more tendency to be with NPS\_Type=Promoter,COUNTRY\_CODE\_R=UNITED STATES, Guest\_State\_H=IL. This means that people in IL prefer spending time in Chicago to rather spending it in New York or Los Angeles. The diagram below depicts a scatter plot for the rules.



Further, we performed association rules on Net promoter score, likelihood to recommend and guest satisfaction metrics.

Apriori-

> arules\_df$Check\_In\_H <- as.factor(arules\_df$Check\_In\_H)

> arules\_df<- dfall1[,c(24:31,33)]

>

> na.omit(arules\_df)

# A tibble: 1,116 x 9

   Overall\_Sat\_H Guest\_Room\_H Tranquility\_H Condition\_Hotel\_H Customer\_SVC\_H Staff\_Cared\_H Internet\_Sat\_H Check\_In\_H  NPS\_Type

           <dbl>        <dbl>         <dbl>             <dbl>          <dbl>         <dbl>          <dbl>      <dbl>     <chr>

 1             9            9            10                 8             10            10              7          8  Promoter

 2             7            8             8                 8              8             8              2          5   Passive

 3             6            6             6                 6              9             9              7          9 Detractor

 4            10           10            10                10             10            10              9         10  Promoter

 5             7            9             9                 9              7             3              9          7 Detractor

 6            10           10            10                10             10             9             10         10  Promoter

 7             9            8             9                 8             10            10              8         10  Promoter

 8             9            8            10                10             10            10             10         10  Promoter

 9             2            3             3                 4              4             7              1         10 Detractor

10             8            9             9                 9              8             8              7          4   Passive

# ... with 1,106 more rows

>

> str(arules\_df)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 63986 obs. of  9 variables:

 $ Overall\_Sat\_H    : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Guest\_Room\_H     : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Tranquility\_H    : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Condition\_Hotel\_H: num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Customer\_SVC\_H   : num  NA NA NA NA NA NA NA 10 NA NA ...

 $ Staff\_Cared\_H    : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Internet\_Sat\_H   : num  NA NA NA NA NA NA NA NA NA NA ...

 $ Check\_In\_H       : num  NA NA NA NA NA NA NA NA NA NA ...

 $ NPS\_Type         : chr  NA NA NA NA ...

>

> arules\_df$NPS\_Type <- as.factor(arules\_df$NPS\_Type)

>

> arules\_df$Overall\_Sat\_H <- as.factor(arules\_df$Overall\_Sat\_H)

>

> arules\_df$Guest\_Room\_H <- as.factor(arules\_df$Guest\_Room\_H)

>

> arules\_df$Tranquility\_H <- as.factor(arules\_df$Tranquility\_H)

>

> arules\_df$Condition\_Hotel\_H <- as.factor(arules\_df$Condition\_Hotel\_H)

>

> arules\_df$Customer\_SVC\_H <- as.factor(arules\_df$Customer\_SVC\_H)

>

> arules\_df$Staff\_Cared\_H <- as.factor(arules\_df$Staff\_Cared\_H)

>

> arules\_df$Internet\_Sat\_H <- as.factor(arules\_df$Internet\_Sat\_H)

>

> arules\_df$Check\_In\_H <- as.factor(arules\_df$Check\_In\_H)

>

> names(arules\_df)

[1] "Overall\_Sat\_H"     "Guest\_Room\_H"      "Tranquility\_H"     "Condition\_Hotel\_H" "Customer\_SVC\_H"    "Staff\_Cared\_H"

[7] "Internet\_Sat\_H"    "Check\_In\_H"        "NPS\_Type"

>

> guestsatisfy <-  apriori(arules\_df,parameter = list(support=.02,confidence=.5))

Apriori

Parameter specification:

 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext

        0.5    0.1    1 none FALSE            TRUE       5    0.02      1     10  rules FALSE

Algorithmic control:

 filter tree heap memopt load sort verbose

    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1279

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[83 item(s), 63986 transaction(s)] done [0.03s].

sorting and recoding items ... [5 item(s)] done [0.00s].

creating transaction tree ... done [0.01s].

checking subsets of size 1 2 3 done [0.00s].

writing ... [30 rule(s)] done [0.00s].

creating S4 object  ... done [0.01s].

> summary(guestsatisfy)

set of 30 rules

rule length distribution (lhs + rhs):sizes

 2  3

18 12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

    2.0     2.0     2.0     2.4     3.0     3.0

summary of quality measures:

    support          confidence          lift           count

 Min.   :0.02015   Min.   :0.5877   Min.   :24.07   Min.   :1289

 1st Qu.:0.02079   1st Qu.:0.7562   1st Qu.:25.51   1st Qu.:1330

 Median :0.02147   Median :0.8710   Median :28.68   Median :1374

 Mean   :0.02210   Mean   :0.8417   Mean   :29.00   Mean   :1414

 3rd Qu.:0.02252   3rd Qu.:0.9329   3rd Qu.:32.82   3rd Qu.:1441

 Max.   :0.02766   Max.   :0.9948   Max.   :34.89   Max.   :1770

mining info:

      data ntransactions support confidence

 arules\_df         63986    0.02        0.5

> inspect(guestsatisfy)

     lhs                                         rhs                    support    confidence lift     count

[1]  {Overall\_Sat\_H=10}                       => {Condition\_Hotel\_H=10} 0.02025443 0.8950276  32.81905 1296

[2]  {Condition\_Hotel\_H=10}                   => {Overall\_Sat\_H=10}     0.02025443 0.7426934  32.81905 1296

[3]  {Overall\_Sat\_H=10}                       => {Customer\_SVC\_H=10}    0.02089520 0.9233425  30.72335 1337

[4]  {Customer\_SVC\_H=10}                      => {Overall\_Sat\_H=10}     0.02089520 0.6952678  30.72335 1337

[5]  {Overall\_Sat\_H=10}                       => {NPS\_Type=Promoter}    0.02247367 0.9930939  25.96817 1438

[6]  {NPS\_Type=Promoter}                      => {Overall\_Sat\_H=10}     0.02247367 0.5876584  25.96817 1438

[7]  {Guest\_Room\_H=10}                        => {Condition\_Hotel\_H=10} 0.02186416 0.9234323  33.86060 1399

[8]  {Condition\_Hotel\_H=10}                   => {Guest\_Room\_H=10}      0.02186416 0.8017192  33.86060 1399

[9]  {Guest\_Room\_H=10}                        => {Customer\_SVC\_H=10}    0.02041071 0.8620462  28.68377 1306

[10] {Customer\_SVC\_H=10}                      => {Guest\_Room\_H=10}      0.02041071 0.6791472  28.68377 1306

[11] {Guest\_Room\_H=10}                        => {NPS\_Type=Promoter}    0.02253618 0.9518152  24.88878 1442

[12] {NPS\_Type=Promoter}                      => {Guest\_Room\_H=10}      0.02253618 0.5892930  24.88878 1442

[13] {Condition\_Hotel\_H=10}                   => {Customer\_SVC\_H=10}    0.02316132 0.8492837  28.25911 1482

[14] {Customer\_SVC\_H=10}                      => {Condition\_Hotel\_H=10} 0.02316132 0.7706708  28.25911 1482

[15] {Condition\_Hotel\_H=10}                   => {NPS\_Type=Promoter}    0.02552121 0.9358166  24.47044 1633

[16] {NPS\_Type=Promoter}                      => {Condition\_Hotel\_H=10} 0.02552121 0.6673478  24.47044 1633

[17] {Customer\_SVC\_H=10}                      => {NPS\_Type=Promoter}    0.02766230 0.9204368  24.06828 1770

[18] {NPS\_Type=Promoter}                      => {Customer\_SVC\_H=10}    0.02766230 0.7233347  24.06828 1770

[19] {Overall\_Sat\_H=10,Condition\_Hotel\_H=10}  => {NPS\_Type=Promoter}    0.02014503 0.9945988  26.00752 1289

[20] {Overall\_Sat\_H=10,NPS\_Type=Promoter}     => {Condition\_Hotel\_H=10} 0.02014503 0.8963839  32.86878 1289

[21] {Condition\_Hotel\_H=10,NPS\_Type=Promoter} => {Overall\_Sat\_H=10}     0.02014503 0.7893448  34.88053 1289

[22] {Overall\_Sat\_H=10,Customer\_SVC\_H=10}     => {NPS\_Type=Promoter}    0.02078580 0.9947644  26.01185 1330

[23] {Overall\_Sat\_H=10,NPS\_Type=Promoter}     => {Customer\_SVC\_H=10}    0.02078580 0.9248957  30.77503 1330

[24] {Customer\_SVC\_H=10,NPS\_Type=Promoter}    => {Overall\_Sat\_H=10}     0.02078580 0.7514124  33.20433 1330

[25] {Guest\_Room\_H=10,Condition\_Hotel\_H=10}   => {NPS\_Type=Promoter}    0.02108274 0.9642602  25.21420 1349

[26] {Guest\_Room\_H=10,NPS\_Type=Promoter}      => {Condition\_Hotel\_H=10} 0.02108274 0.9355062  34.30333 1349

[27] {Condition\_Hotel\_H=10,NPS\_Type=Promoter} => {Guest\_Room\_H=10}      0.02108274 0.8260870  34.88977 1349

[28] {Condition\_Hotel\_H=10,Customer\_SVC\_H=10} => {NPS\_Type=Promoter}    0.02245804 0.9696356  25.35476 1437

[29] {Condition\_Hotel\_H=10,NPS\_Type=Promoter} => {Customer\_SVC\_H=10}    0.02245804 0.8799755  29.28035 1437

[30] {Customer\_SVC\_H=10,NPS\_Type=Promoter}    => {Condition\_Hotel\_H=10} 0.02245804 0.8118644  29.76960 1437

We found thirty good rules by applying the model, out of which six were relevant due to high lift values. The output provided us with the inference that there are higher number of promoters for the guest satisfaction metric of Guests’ rooms, Condition of the hotel and the overall satisfaction metric. We inferred that if a customer is a promoter then they are satisfied by both the condition and quality of service metrics. Also since we have a lot of values for the Lift parameter in the range from 25-30 as compared to the maximum value, it can be said that there is a higher probability for promoters for almost all metrics.

* SVM modeling:

We performed svm modeling on the dataset to arrive at conclusions.

> trainData<-df.fake[randIndex[1:cutPoint2\_3],]

> testData<-df.fake[randIndex[(cutPoint2\_3+1):dim(df.fake)[1]],]

> goodOzone<-ifelse(trainData$NPS\_Type=="Promoter",2,

+ ifelse(trainData$NPS\_Type=="Passive",1,0))

> trainData<-data.frame(trainData,goodOzone)

> goodOzone<-ifelse(testData$NPS\_Type=="Promoter",2,

+ ifelse(trainData$NPS\_Type=="Passive",1,0))

> testData<-data.frame(testData,goodOzone)

> trainData$goodOzone <- as.factor(trainData$goodOzone)

> testData$goodOzone <- as.factor(testData$goodOzone)

> str(trainData)

'data.frame': 2857 obs. of 8 variables:

$ NPS\_Type : Factor w/ 4 levels "","Detractor",..: 2 4 4 4 4 3 4 4 4 4 ...

$ Likelihood\_Recommend\_H: int 6 9 10 9 10 8 9 10 10 9 ...

$ Overall\_Sat\_H : num 7 9 10 9 10 8 9 10 10 9 ...

$ Guest\_Room\_H : num 9 9 10 9 10 8 9 9 8 10 ...

$ Condition\_Hotel\_H : num 6 9 10 9 10 8 9 10 10 10 ...

$ Customer\_SVC\_H : num 9 9 10 9 10 7 9 10 10 10 ...

$ Staff\_Cared\_H : num 8 8.84 8.84 9 8.84 8.84 9 8.84 8.84 9 ...

$ goodOzone : Factor w/ 3 levels "0","1","2": 1 3 3 3 3 2 3 3 3 3 ...

> svmGood<-ksvm(goodOzone~., data=trainData, kernel="rbfdot", kpar="automatic", C=5, cross=13, prob.model=TRUE)

> svmGood

Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)

parameter : cost C = 5

Gaussian Radial Basis kernel function.

Hyperparameter : sigma = 0.579863731292562

Number of Support Vectors : 413

Objective Function Value : -48.7018 -36.2381 -29.1193

Training error : 0

Cross validation error : 0.005253

Probability model included.

> goodPred <- predict(svmGood, testData)

> compGood1 <- data.frame(testData[,8], goodPred)

> colnames(compGood1) <- c("test","Pred")

> perc\_ksvm <- length(which(compGood1$test==compGood1$Pred))/dim(compGood1)[1]

> perc\_ksvm

[1] 0.8222533

All models created by three different methods are working well. Among, ksvm is the best model for this data because the percentage of accurate cases for ksvm is high and it provides more insights.

# Insights

As a team of five members from diverse backgrounds with no prior experience in R, we have gained a fair understanding of the language and its functionalities by now. We started with a humongous dataset and an undefined scope, to arrive at the visualizations and insights. Picking the data questions that we wanted to address enabled us to think like data analysts, albeit at the amateur level. We also realized that estimates cannot be made until there is concrete evidence backed by data and visualizations or prediction models to support those said estimates. We aim to utilize this knowledge for enhancing our capabilities with respect to data science.