Grand Hyatt Group of Hotels

IST 687-Data Science

Submitted to:

Prof. Jeffery Saltz
Prof. Gary Krudys
Ivan Shamshurin

Submitted by:

Sanjana Rajagopala

Table of Contents

Introduction

Data Science deals with extraction of data from various sources and generation of predictive models to gain insights. As data scientists we aim to cater to the demands of the project on hand by generating trends and patterns identifying the anomalies and finally providing recommendations.

Based on the business questions as described in the next section we provide an overview of the steps to be followed to reach our desired goals. The descriptive statistics modelling techniques and visualizations will support our findings and will suffice as the base for recommendations.

Business Questions

1. By how much does the percentage of promoters differ from the percentage of detractors in different countries?
2. Why is the percentage of promoters high in one country than the others?
3. What can be done to improvise the percentage of promoters in the countries that do not have high promoters?

Overview

Grand Hyatt group of hotels is a large chain of international hotels with voluminous data. We followed these steps to achieve the goals –

1. Data Acquisition – This involved the collection of data from the provide Excel data sheets. Considering the processing power and time involved in reading such humungous data we need to limit to reading only the necessary data set. To achieve this, we finalized on three countries, two months and extraction of only the attributes of interest.

We wanted to perform analysis and compare the results across developed and developing economies in the eastern part of the world. We chose Egypt – a developing country and Japan – a developed country. Additionally, because India to be a fast-growing nation we added it as part of the chosen countries. The months of February and December are one of the peak time of travel and vacation around the world. Thus, we chose the two months for our analysis. For our analysis and the goals, we identified the required attributes and categorized them into separate groups –

a. **Internet Satisfaction** – The provision of a hassle-free internet facility contributes greatly to the overall satisfaction of the customer. Thus, we choose this as one of our variable groups and its value is mainly determined by those of the following attributes -

i. Internet_Dissat_Lobby_H
ii. Internet_Dissat_Slow_H
iii. Internet_Dissat_Expensive_H
iv. Internet_Dissat_Connectivity_H
v. Internet_Dissat_Billing_H
vi. Internet_Dissat_Wired_H
vii. Internet_Dissat_Other_H
viii. TV_Internet_General_H

b. **Food & Beverage Experience** – The hotel experience is majorly highlighted with the experience offered in terms of food and beverages. The following are the attributes in this group –

i. F&B_FREQ_H
ii. Spa F&B offering_PL
iii. F&B_Overall_Experience_H

c. **Room Condition Experience** – There are several factors that decide the satisfaction of the customer with regards to the stay in the hotel. The attributes in this group include –

i. All Suites_PL
ii. Bell Staff_PL
iii. Boutique_PL
iv. Business Center_PL
v. Casino_PLConference_PL
vi. Convention_PL
vii. Dry-Cleaning_PL
viii. Elevators_PL
ix. Fitness Center_PL
x. Fitness Trainer_PL
xi. Golf_PL
xii. Indoor Corridors_PL
xiii. Laundry_PL
xiv. Limo Service_PL
xv. Mini-Bar_PL
xvi. Pool-Indoor_PL
xvii. Pool-Outdoor_PL
xviii. Regency Grand Club_PL
xix. Resort_PL
xx. Restaurant_PL

xxi. Self-Parking_PL
xxii. Shuttle Service_PL
xxiii. Ski_PL
xxiv. Spa_PL
xxv. Spa services in fitness center_PL
xxvi. Spa online booking_PL
xxvii. Spa F&B offering_PL
xxviii. Valet Parking_PL Country_PL

2. **Data Cleansing/Munging** - On discovering the data set we determined the data variables supporting our business questions. One of the data problems is the inconsistence in the data – presence of NA and other blank values. We ought to figure out methods to make the data consistent. To do so, we employed several data munging techniques which are explained in the next sections.
3. **Data Exploration and analysis** - This step involved performing Descriptive statistics analysis, Modelling and Comparison of results. Using descriptive statistics helped us in getting an overall picture of the available data set. The usage of LM, KSVM, LM and Naïve-Bayes models enabled us to compute prediction of the Likelihood_To_Recommend, compare and draw results from the same. This entire process is repeated for different countries to compare the percentage of promoters and detractors promoting or demoting the hotel in that particular country. Recommendations will then be given to the hotels of the country where the NPS is less.
4. **Data Visualization** – This involved the generation of relevant and easy-to-understand plots. Based on the kind of business question we intend to answer, we created relevant plots. It helped us communicate the results efficiently and clearly.

Data munging/cleaning

We identified three main factors that would contribute to Likeliness To Recommend

1.Internet satisfaction
2.Room condition
3.Food and Beverage experience

We also identified the variables that will contribute to these factors and imported the selected variables into R for the countries Japan, Egypt and India for the months December and February.

After observing large number of blanks and NA's in the imported data, we removed NA's and blanks in Internet satisfaction and Food and Beverage variables. For the Room condition variables we imputed NAs and blank values with the mean value of the respective columns.

Descriptive Statistics

Descriptive statistics are used to summarize data in a way that provides insight into the information contained in the data. This might include examining the mean or median of numeric data or the frequency of observations for nominal data. Plots can be created that show the data and indicating summary statistics.

While choosing which summary statistics are appropriate we mainly focus on the type of variable being examined. In describing or examining data we typically are concerned with variation percentage of one variable over the other.

We begin with the "str" function in order to obtain compact analysis of the given data frame. As seen below it given an overall picture about the number of rows columns and the type of each variable in the data frame.

```
> str(requiredRoomData)
'data.frame':	66209 obs. of  30 variables:
 $ All.Suites_PL                    : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Bell.Staff_PL                    : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Boutique_PL                      : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Business.Center_PL               : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 1 1 1 ...
 $ Casino_PL                        : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Conference_PL                    : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Convention_PL                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
 $ Dry.Cleaning_PL                  : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Elevators_PL                     : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Fitness.Center_PL                : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Fitness.Trainer_PL               : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Golf_PL                          : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Indoor.Corridors_PL              : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Laundry_PL                       : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Limo.Service_PL                  : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 1 1 ...
 $ Mini.Bar_PL                      : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Pool.Indoor_PL                   : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Pool.Outdoor_PL                  : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 1 1 ...
 $ Regency.Grand.Club_PL            : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 1 1 1 ...
 $ Resort_PL                        : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Restaurant_PL                    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Self.Parking_PL                  : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 1 1 1 ...
 $ Shuttle.Service_PL               : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Ski_PL                           : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Spa_PL                           : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Spa.services.in.fitness.center_PL: Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Spa.online.booking_PL            : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 1 ...
 $ Spa.F.B.offering_PL              : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 2 1 1 ...
 $ Valet.Parking_PL                 : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 1 1 ...
 $ V30                              : Factor w/ 11 levels "0","1","3","4",..: 3 11 6 10 10 10 10 7 4 4 ...
> |
```

This is further explored in detail using the summary function where the information about each of the variables with the mean median mode is displayed. A sample result is as follows –

```
> summary(requiredRoomData)
 All.Suites_PL Bell.Staff_PL Boutique_PL Business.Center_PL Casino_PL Conference_PL Convention_PL Dry.Cleaning_PL Elevators_PL
 0:66209       0:15691       0:66209     0:20259            0:66209   0:66209       0:36145       0:15458         0:16992
               1:50518                   1:45950                                    1:30064       1:50751         1:49217



 Fitness.Center_PL Fitness.Trainer_PL Golf_PL  Indoor.Corridors_PL Laundry_PL Limo.Service_PL Mini.Bar_PL Pool.Indoor_PL
 0:16712           0:16945            0:66209  0: 3760             0:15458    0:19331         0:20894     0:42609
 1:49497           1:49264                     1:62449             1:50751    1:46878         1:45315     1:23600



 Pool.Outdoor_PL Regency.Grand.Club_PL Resort_PL Restaurant_PL Self.Parking_PL Shuttle.Service_PL Ski_PL    Spa_PL
 0:27860         0:25339               0:66209   0: 3295       0:26022         0:49808            0:66209   0: 3993
 1:38349         1:40870                         1:62914       1:40187         1:16401                      1:62216



 Spa.services.in.fitness.center_PL Spa.online.booking_PL Spa.F.B.offering_PL Valet.Parking_PL      V30
 0:66209                           0:32059               0:46253             0:33601          17     :18666
                                   1:34150               1:19956             1:32608          19     :13773
                                                                                              4      : 6431
                                                                                              15     : 6131
                                                                                              18     : 5412
                                                                                              3      : 5267
                                                                                              (Other):10529
```

Apart from this we have further performed the descriptive statistics on this data by computing percentage values using certain attributes.

1. Among the RoomCondition variables, the basic facilities that are expected in any hotel include – Dry-Cleaning_PL, Elevators_PL, Indoor Corridors_PL, Laundry_PL, Restaurant_PL. In order to analyze this, we computed the percentage of the set of hotels offering these facilities

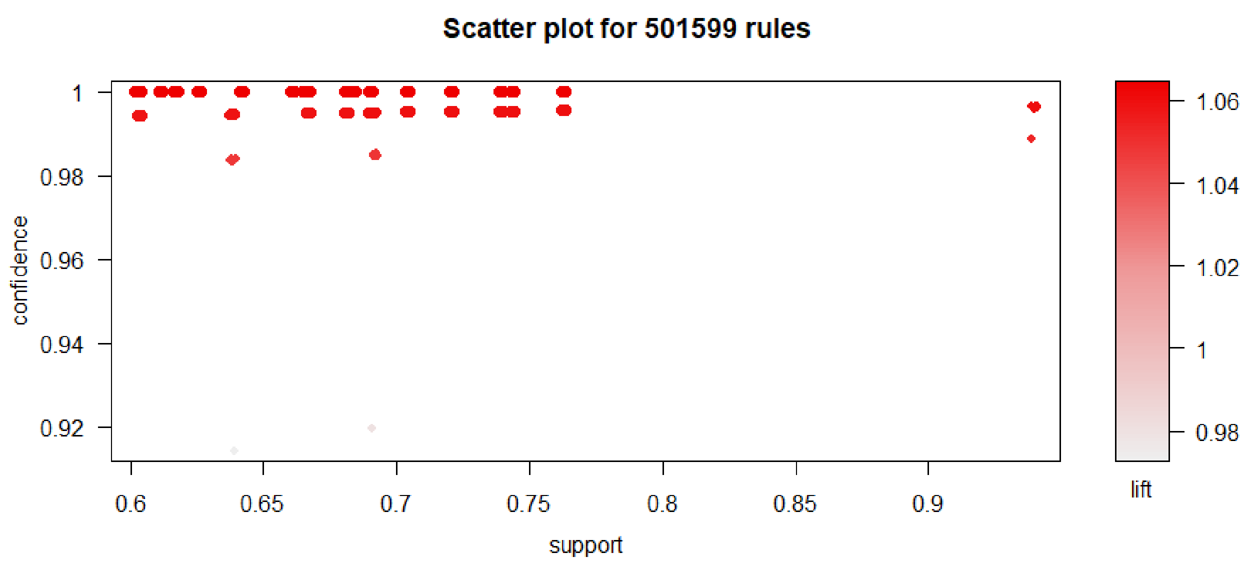
```
> #How many have the basic facilities of the amenities available in their hotels
> AmenitiesPerc<- satisfyFunction(requiredRoomData)/ length(requiredRoomData$All.Suites_PL)
> AmenitiesPerc
[1] 0.7433733
```

2. Going forward, we computed the percentage of hotels offering the Spa_ Services along with the above basic facilities. The results are as follows –

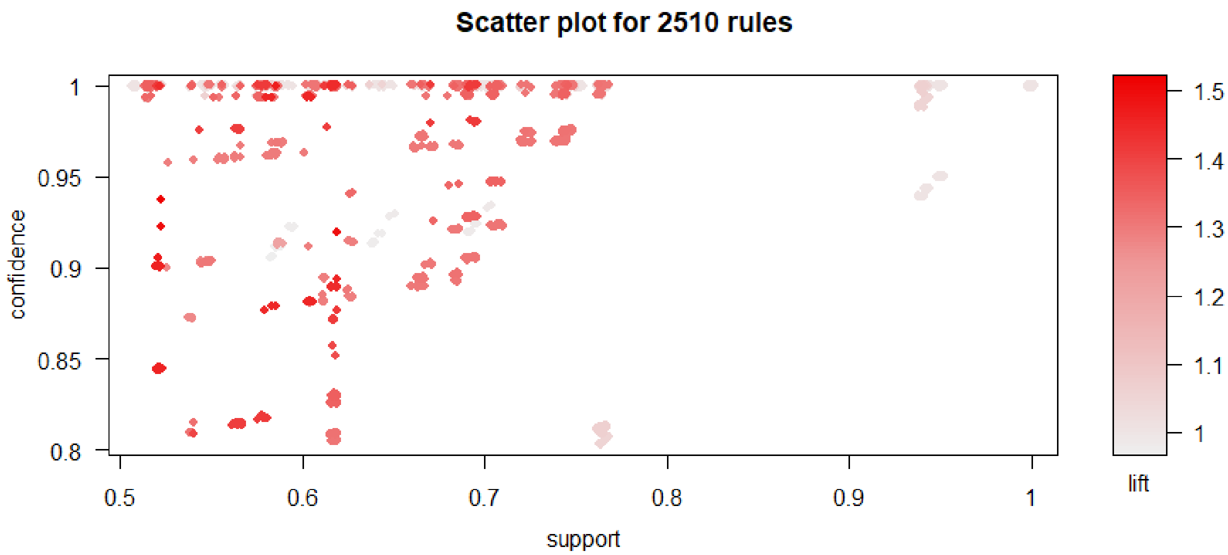
```
> SpaBasicPerc <- totalspaBasic/length(requiredRoomData$All.Suites_PL)
> SpaBasicPerc
[1] 0.7398541
>
```

The next steps were to use the aRules package and obtain set of rules from the data set.

1. Considering the large amount of data points and the associated variables the obtained rules set consisted of more than 50,000 rules. The plot for the same rule set is as shown below –



2. To reduce the number of rules – removing the redundant rules and concentratin only on the suitable rules we adjusted the confidence and support values iteratively and reduced the number of rules to around 2000.



3. To obtain the good rules from the above rule set, we inspected and filtered the rules with lift value greated than 1.49. A glimpse of the rule set is as shown below –

```
[44] {Pool.Outdoor_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Business.Center_PL=1}              0.5651951  0.9758012 1.4060244 37421
[45] {Business.Center_PL=1,
      Pool.Outdoor_PL=1}                    => {Golf_PL=0}                         0.5651951  1.0000000 1.0000000 37421
[46] {Laundry_PL=1,
      Limo.Service_PL=1}                    => {Pool.Outdoor_PL=1}                 0.5792113  0.8180596 1.4123683 38349
[47] {Pool.Outdoor_PL=1,
      Resort_PL=0}                          => {Limo.Service_PL=1}                 0.5792113  1.0000000 1.4123683 38349
[48] {Elevators_PL=1,
      Pool.Outdoor_PL=1}                    => {Dry.Cleaning_PL=1}                 0.5560422  1.0000000 1.3045851 36815
[49] {Elevators_PL=1,
      Pool.Outdoor_PL=1}                    => {Casino_PL=0}                       0.5560422  1.0000000 1.0000000 36815
[50] {Fitness.Trainer_PL=1,
      Pool.Outdoor_PL=1}                    => {Spa.services.in.fitness.center_PL=0} 0.5756921  1.0000000 1.0000000 38116
[51] {Fitness.Center_PL=1,
      Pool.Outdoor_PL=1}                    => {Laundry_PL=1}                      0.5792113  1.0000000 1.3045851 38349
[52] {Fitness.Center_PL=1,
      Pool.Outdoor_PL=1}                    => {Conference_PL=0}                   0.5792113  1.0000000 1.0000000 38349
[53] {Bell.Staff_PL=1,
      Pool.Outdoor_PL=1}                    => {Resort_PL=0}                       0.5756921  1.0000000 1.0000000 38116
[54] {Dry.Cleaning_PL=1,
      Pool.Outdoor_PL=1}                    => {All.Suites_PL=0}                   0.5792113  1.0000000 1.0000000 38349
[55] {Laundry_PL=1,
      Pool.Outdoor_PL=1}                    => {Spa.services.in.fitness.center_PL=0} 0.5792113  1.0000000 1.0000000 38349
[56] {Pool.Outdoor_PL=1,
      Spa_PL=1}                             => {Spa.services.in.fitness.center_PL=0} 0.5756921  1.0000000 1.0000000 38116
[57] {Indoor.Corridors_PL=1,
      Pool.Outdoor_PL=1}                    => {All.Suites_PL=0}                   0.5792113  1.0000000 1.0000000 38349
[58] {Pool.Outdoor_PL=1,
      Restaurant_PL=1}                      => {Resort_PL=0}                       0.5792113  1.0000000 1.0000000 38349
[59] {Pool.Outdoor_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Conference_PL=0}                   0.5792113  1.0000000 1.0000000 38349
[60] {Pool.Outdoor_PL=1,
      Ski_PL=0}                             => {Casino_PL=0}                       0.5792113  1.0000000 1.0000000 38349
[61] {Casino_PL=0,
      Pool.Outdoor_PL=1}                    => {Conference_PL=0}                   0.5792113  1.0000000 1.0000000 38349
[62] {Fitness.Center_PL=1,
      Mini.Bar_PL=1}                        => {Self.Parking_PL=1}                 0.5845127  0.8783278 1.4470651 38700
[63] {Self.Parking_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Mini.Bar_PL=1}                     0.6034527  0.9942021 1.4526123 39954
[64] {Mini.Bar_PL=1,
      Self.Parking_PL=1}                    => {Golf_PL=0}                         0.6034527  1.0000000 1.0000000 39954
[65] {Indoor.Corridors_PL=1,
      Self.Parking_PL=1}                    => {Business.Center_PL=1}              0.5484753  0.9036256 1.3020271 36314
[66] {Elevators_PL=1
```

4. To have better perspective about the effect of the Spa_ Services as seen from the above descriptive statistics, we computed rules with following filters –

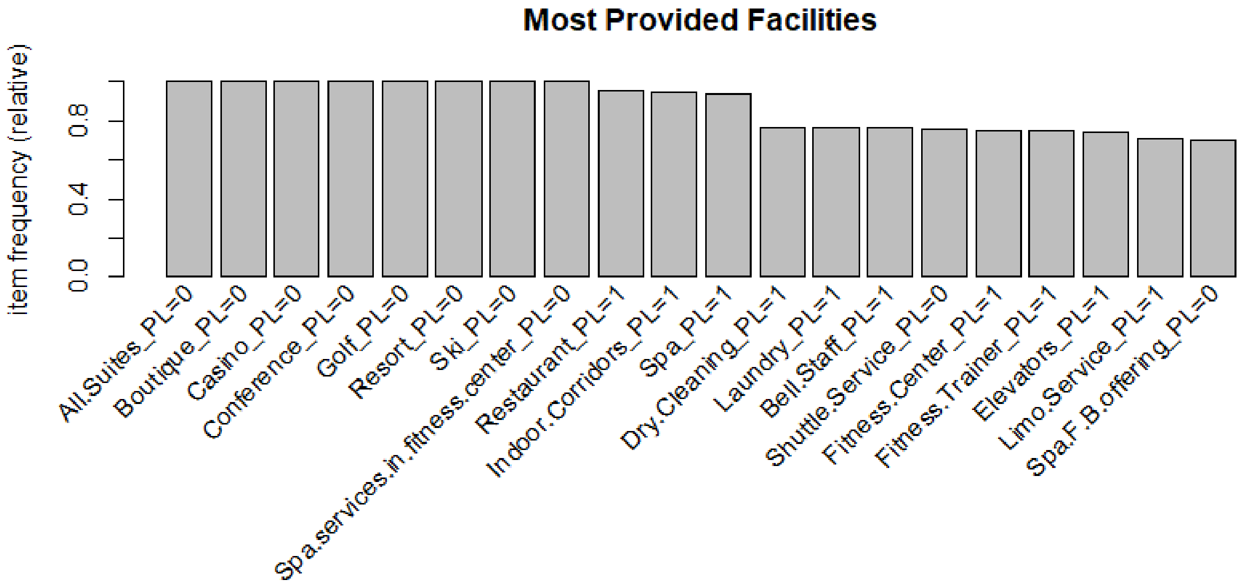
```
> rulesset_3<-apriori(room_Cond_matrix, parameter=list(support=0.5, confidence=0.2, maxlen=10),
+ appearance = list (default="rhs",lhs='Spa_PL=1'))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.2    0.1    1 none FALSE            TRUE       5     0.5      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: 33104

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[61 item(s), 66209 transaction(s)] done [0.16s].
sorting and recoding items ... [29 item(s)] done [0.02s].
creating transaction tree ... done [0.10s].
checking subsets of size 1 2 done [0.00s].
writing ... [54 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
> inspect(rulesset_3)
```

The resulting rules set is as follows –

```
[44] {Pool.Outdoor_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Business.Center_PL=1}                0.5651951  0.9758012 1.4060244 37421
[45] {Business.Center_PL=1,
      Pool.Outdoor_PL=1}                   => {Golf_PL=0}                           0.5651951  1.0000000 1.0000000 37421
[46] {Laundry_PL=1,
      Limo.Service_PL=1}                   => {Pool.Outdoor_PL=1}                   0.5792113  0.8180596 1.4123683 38349
[47] {Pool.Outdoor_PL=1,
      Resort_PL=0}                         => {Limo.Service_PL=1}                   0.5792113  1.0000000 1.4123683 38349
[48] {Elevators_PL=1,
      Pool.Outdoor_PL=1}                   => {Dry.Cleaning_PL=1}                   0.5560422  1.0000000 1.3045851 36815
[49] {Elevators_PL=1,
      Pool.Outdoor_PL=1}                   => {Casino_PL=0}                         0.5560422  1.0000000 1.0000000 36815
[50] {Fitness.Trainer_PL=1,
      Pool.Outdoor_PL=1}                   => {Spa.services.in.fitness.center_PL=0} 0.5756921  1.0000000 1.0000000 38116
[51] {Fitness.Center_PL=1,
      Pool.Outdoor_PL=1}                   => {Laundry_PL=1}                        0.5792113  1.0000000 1.3045851 38349
[52] {Fitness.Center_PL=1,
      Pool.Outdoor_PL=1}                   => {Conference_PL=0}                     0.5792113  1.0000000 1.0000000 38349
[53] {Bell.Staff_PL=1,
      Pool.Outdoor_PL=1}                   => {Resort_PL=0}                         0.5756921  1.0000000 1.0000000 38116
[54] {Dry.Cleaning_PL=1,
      Pool.Outdoor_PL=1}                   => {All.Suites_PL=0}                     0.5792113  1.0000000 1.0000000 38349
[55] {Laundry_PL=1,
      Pool.Outdoor_PL=1}                   => {Spa.services.in.fitness.center_PL=0} 0.5792113  1.0000000 1.0000000 38349
[56] {Pool.Outdoor_PL=1,
      Spa_PL=1}                            => {Spa.services.in.fitness.center_PL=0} 0.5756921  1.0000000 1.0000000 38116
[57] {Indoor.Corridors_PL=1,
      Pool.Outdoor_PL=1}                   => {All.Suites_PL=0}                     0.5792113  1.0000000 1.0000000 38349
[58] {Pool.Outdoor_PL=1,
      Restaurant_PL=1}                     => {Resort_PL=0}                         0.5792113  1.0000000 1.0000000 38349
[59] {Pool.Outdoor_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Conference_PL=0}                     0.5792113  1.0000000 1.0000000 38349
[60] {Pool.Outdoor_PL=1,
      Ski_PL=0}                            => {Casino_PL=0}                         0.5792113  1.0000000 1.0000000 38349
[61] {Casino_PL=0,
      Pool.Outdoor_PL=1}                   => {Conference_PL=0}                     0.5792113  1.0000000 1.0000000 38349
[62] {Fitness.Center_PL=1,
      Mini.Bar_PL=1}                       => {Self.Parking_PL=1}                   0.5845127  0.8783278 1.4470651 38700
[63] {Self.Parking_PL=1,
      Spa.services.in.fitness.center_PL=0} => {Mini.Bar_PL=1}                       0.6034527  0.9942021 1.4526123 39954
[64] {Mini.Bar_PL=1,
      Self.Parking_PL=1}                   => {Golf_PL=0}                           0.6034527  1.0000000 1.0000000 39954
[65] {Indoor.Corridors_PL=1,
      Self.Parking_PL=1}                   => {Business.Center_PL=1}                0.5484753  0.9036256 1.3020271 36314
[66] {Elevators_PL=1
```

5. The motivation to extract possible specifics from the data set caused us to dig deeper. By using the "Eclat" function where in the number of variables is limited to 20 and the resulting plot shows the set of variables occuring the most number of times in the given data set.

For example, All_Suites_PL=0 with the value greater than 0.8 means most of the hotels do not offer all the possible suites whereas Spa_PL=1 at a further point in the plot means that there are fairly decent number of hotels offering Spa.



Modelling techniques

Modelling techniques are the basis of predictive analysis. Predictive analysis is nothing but the area of statistics which helps in identifying different trends and patterns in the data. Obviously, if prediction is accurate, then required measures can be taken by businesses beforehand. The models we are using to predict the trends and patterns in the Hyatt dataset are as follows:

1. **Linear Modeling**

Linear models describe a continuous response variable as a function of one or more predictor variables. They can help you understand and predict the behavior of complex systems or analyze experimental, financial, and biological data. Linear regression is a statistical method used to create a linear model. This model is used to establish a linear relationship between attributes. Based on the variables required for prediction, we build an lm model. This model is then used to predict the response variable (Linear model).

2. **KSVM**

Support Vector Machines are an excellent tool for classification, novelty detection, and regression. ksvm supports the well-known C-svc, nu-svc, (classification) one-class-svc (novelty) eps-svr, nu-svr (regression) formulations along with native multi-class classification formulations and the bound-constraint SVM formulations. ksvm also supports class-probabilities output and confidence intervals for regression.

3. **SVM**

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. svm is used to train a support vector machine. It can be used to carry out general regression and classification (of nu and epsilon-type), as well as density-estimation.

4. **Naive Bayes**

The reason it is termed “naive” is because there is independence between attributes when they may be dependent in some way. Below is the Naive Bayes’ Theorem:
P(A | B) = P(A) * P(B | A) / P(B)
The Naive Bayes classifier has proven to be highly effective and is commonly deployed in email spam filters.

We used Linear Modeling and KSVM to predict Likelihood to recommend value (scale of 10) from three variables- Internet Satisfaction, F&B satisfaction and Room Satisfaction. On the other hand, we used SVM and Naive Bayes models to predict the categorical values which is nothing but the NPS type.

INDIA

KSVM

```
> ksvmOutput_india
Support Vector Machine object of class "ksvm"
SV type: eps-svr  (regression)
parameter : epsilon = 0.1  cost C = 5
Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  1.3091387111737
Number of Support Vectors : 36
Objective Function Value : -169.9751
Training error : 0.964925
Cross validation error : 0.380265
Laplace distr. width : 0
> rmse_india
[1] 0.3849052
> step(lmModel_india, data = India_ltr_data, direction = 'backward')
```

```
Start:  AIC=-171.67
LTR ~ FB + Internet + RoomCond
           Df Sum of Sq   RSS     AIC
- Internet  1 0.0005226 43.958 -173.67
- FB        1 0.0090389 43.967 -173.64
- RoomCond  1 0.0278865 43.986 -173.57
<none>                  43.958 -171.67
Step:  AIC=-173.67
LTR ~ FB + RoomCond
           Df Sum of Sq   RSS     AIC
- FB        1 0.0085929 43.967 -175.64
- RoomCond  1 0.0277061 43.986 -175.57
<none>                  43.958 -173.67
Step:  AIC=-175.64
LTR ~ RoomCond
           Df Sum of Sq   RSS     AIC
- RoomCond  1  0.033076 44.000 -177.53
<none>                  43.967 -175.64
Step:  AIC=-177.53
LTR ~ 1
Call:
lm(formula = LTR ~ 1, data = trainData_india)
Coefficients:
(Intercept)
          9
```

LM

```
> summary(lmModel_india)
Call:
lm(formula = LTR ~ FB + Internet + RoomCond, data = trainData_india)
Residuals:
    Min      1Q  Median      3Q     Max
-4.0246  0.0032  0.0082  0.0108  1.0126
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.2363213  4.6381800   1.776   0.0779 .
FB           0.0898424  0.5221070   0.172   0.8636
Internet     0.0008807  0.0212845   0.041   0.9671
RoomCond    -0.0023621  0.0078152  -0.302   0.7629
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.5525 on 144 degrees of freedom
Multiple R-squared:  0.0009589,   Adjusted R-squared:  -0.01985
F-statistic: 0.04607 on 3 and 144 DF,  p-value: 0.9868
```

> rmse_lm_india
[1] 0.385603

SVM
> perc_svm - INDIA
[1] 0.9411765

NB
> perc_nb
[1] 0.7733333

Confusion Matrix and Statistics

```
          Reference
Prediction  0  1  2
         0  0  0  0
         1  0 58  8
         2  0  9  0
```

Overall Statistics

Accuracy : 0.7733
95% CI : (0.6621, 0.8621)

The step function used with linear model signifies the importance of all the variables on which LTR depends. Here, we see that the svm model is best for recommending suggestions to Indian Hotels as the accuracy obtained is 94% which is very reliable.

JAPAN

KSVM
ksvmOutput_japan
Support Vector Machine object of class "ksvm"
SV type: eps-svr (regression)
parameter : epsilon = 0.1 cost C = 5
Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 2.09076109707378
Number of Support Vectors : 35
Objective Function Value : -36.8723
Training error : 0.121189
Cross validation error : 0.083162
Laplace distr. width : 0

LM
> summary(lmModel_japan)
Call:
lm(formula = LTR ~ FB + Internet + RoomCond, data = trainData_japan)
Residuals:

```
       Min       1Q   Median    3Q      Max
-0.51581 -0.47225  0.00822  0.45131  0.52914
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.801604   2.577538   1.475   0.1450
FB       0.705696   0.292218   2.415   0.0185 *
Internet        0.008714   0.024094   0.362   0.7188
RoomCond     -0.032119   0.007657  -4.195 8.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4152 on 66 degrees of freedom
Multiple R-squared:  0.2634,          Adjusted R-squared:  0.2299
F-statistic: 7.866 on 3 and 66 DF,  p-value: 0.0001456
> rmse_lm_japan
[1] 0.7438377
> step(lmModel_japan, data = Japan_ltr_data, direction = 'backward')
Start:  AIC=-119.19
LTR ~ FB + Internet + RoomCond
       Df Sum of Sq   RSS     AIC
- Internet  1   0.02254 11.398 -121.05
<none>                     11.376 -119.19
- FB      1   1.00521 12.381 -115.26
- RoomCond  1   3.03273 14.409 -104.65
Step:  AIC=-121.05
LTR ~ FB + RoomCond
       Df Sum of Sq   RSS     AIC
<none>                     11.398 -121.05
- FB      1         1.0907 12.489 -116.66
- RoomCond  1         3.0106 14.409 -106.65
Call:
lm(formula = LTR ~ FB + RoomCond, data = trainData_japan)
Coefficients:
(Intercept)               FB       RoomCond
       3.66814     0.72399   -0.03185
```

SVM

```
> perc_svm
[1] 0.6571429
```

```
> confMatrix
```

Confusion Matrix and Statistics

```
      Reference
Prediction  1  2
       1  5 11
       2  1 18
```

```
            Accuracy : 0.6571
              95% CI : (0.4779, 0.8087)
 No Information Rate : 0.8286
 P-Value [Acc > NIR] : 0.996223

               Kappa : 0.2734
Mcnemar's Test P-Value : 0.009375

         Sensitivity : 0.8333
         Specificity : 0.6207
      Pos Pred Value : 0.3125
      Neg Pred Value : 0.9474
          Prevalence : 0.1714
      Detection Rate : 0.1429
Detection Prevalence : 0.4571
   Balanced Accuracy : 0.7270

    'Positive' Class : 1
```

The step function used with linear model signifies the importance of F&B and Room Condition on which LTR depends. Here, we see that the svm model is best for recommending suggestions to Japanese Hotels as the accuracy obtained is 65% which is very reliable as compared to the error thrown by other models.

Egypt

KSVM

```
> ksvmOutput_egypt
Support Vector Machine object of class "ksvm"
SV type: eps-svr  (regression)
parameter : epsilon = 0.1  cost C = 5
Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  0.84212296242615
Number of Support Vectors : 14
Objective Function Value : -10.3975
Training error : 0.094657
Cross validation error : 0.221392
Laplace distr. width : 0
> rmse_egypt
[1] 0.6126318
```

LM

```
> summary(lmModel_egypt)
Call:
lm(formula = LTR ~ FB + Internet + RoomCond, data = trainData_egypt)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.8695  0.0045  0.0356  0.1264  0.1458
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 342.424798  21.766274  15.732 1.97e-15 ***
FB          -37.780986   2.462338 -15.344 3.70e-15 ***
Internet      0.004779   0.029788   0.160 0.874
RoomCond     -0.005720   0.013664  -0.419    0.679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3654 on 28 degrees of freedom
Multiple R-squared:  0.9053,	Adjusted R-squared:  0.8952
F-statistic: 89.27 on 3 and 28 DF,  p-value: 1.919e-14
> rmse_lm_egypt
[1] 0.8181145
> step(lmModel_egypt, data = Egypt_ltr_data, direction = 'backward')
Start:  AIC=-60.7
LTR ~ FB + Internet + RoomCond
           Df Sum of Sq    RSS     AIC
- Internet  1    0.0034  3.742 -62.672
- RoomCond  1    0.0234  3.762 -62.502
<none>                   3.739 -60.701
- FB        1   31.4377 35.177   9.029
Step:  AIC=-62.67
LTR ~ FB + RoomCond
           Df Sum of Sq    RSS     AIC
- RoomCond  1     0.022  3.765 -64.482
<none>                   3.742 -62.672
- FB        1    35.757 39.500  10.738
Step:  AIC=-64.48
LTR ~ FB
       Df Sum of Sq    RSS     AIC
<none>               3.765 -64.482
- FB    1    35.735 39.500   8.738
Call:
lm(formula = LTR ~ FB, data = trainData_egypt)
Coefficients:
(Intercept)           FB
     343.30       -37.89
```

SVM

```
> perc_svm
[1] 0.9411765
```

NB

```
>perc_nb
[1]    0.8
```

Confusion Matrix and Statistics

```
          Reference
Prediction  1  2
         1 15  1
         2  6 13

               Accuracy : 0.8
                 95% CI : (0.6306, 0.9156)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.01017

                  Kappa : 0.6067
 Mcnemar's Test P-Value : 0.13057

            Sensitivity : 0.7143
            Specificity : 0.9286
         Pos Pred Value : 0.9375
         Neg Pred Value : 0.6842
             Prevalence : 0.6000
         Detection Rate : 0.4286
   Detection Prevalence : 0.4571
      Balanced Accuracy : 0.8214

       'Positive' Class : 1
```

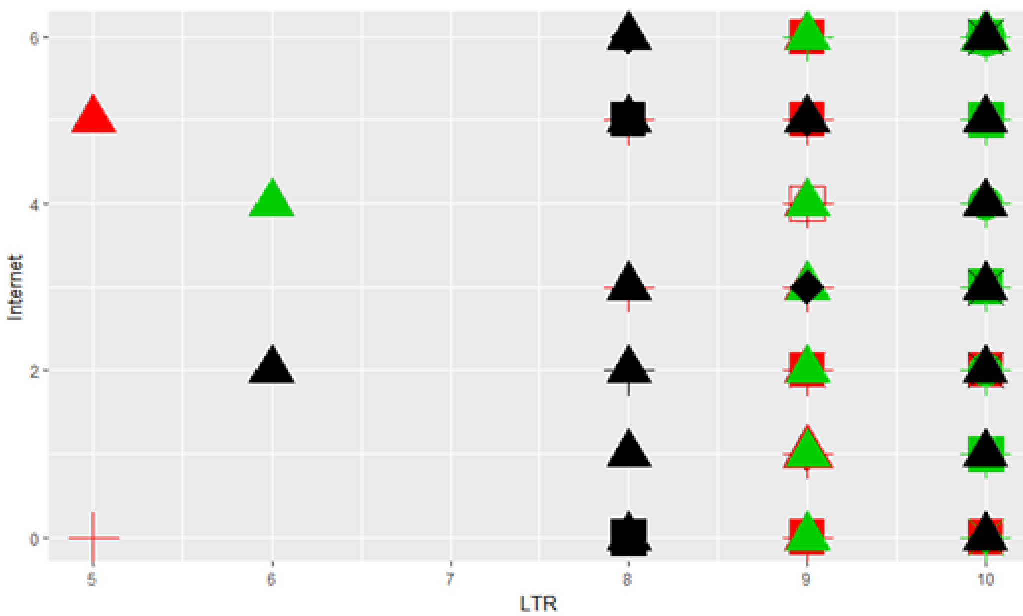
The step function used with linear model signifies the importance of F&B on which LTR depends. Here, we see that the svm or nb models are best for recommending suggestions to Egyptian Hotels as the accuracy obtained is 94% from svm and 80% from nb which are very reliable as compared to the error thrown by other models.

Visualization

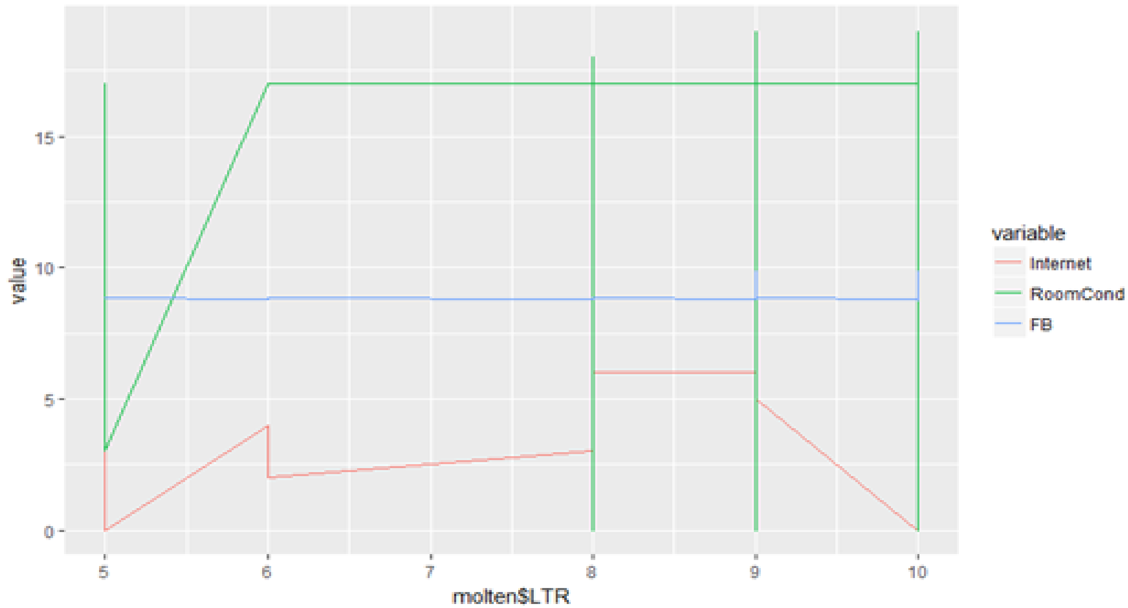
It is easier for a human brain to visualize large amounts of data using charts and graphs instead of reports or spreadsheets. In this project, the variable Likeliness_To_Recommend is determined based on the consolidated data obtained from three major factors for the countries India, Egypt and Japan

1.Internet Satisfaction
2.Food and Beverage Experience
3.Room condition

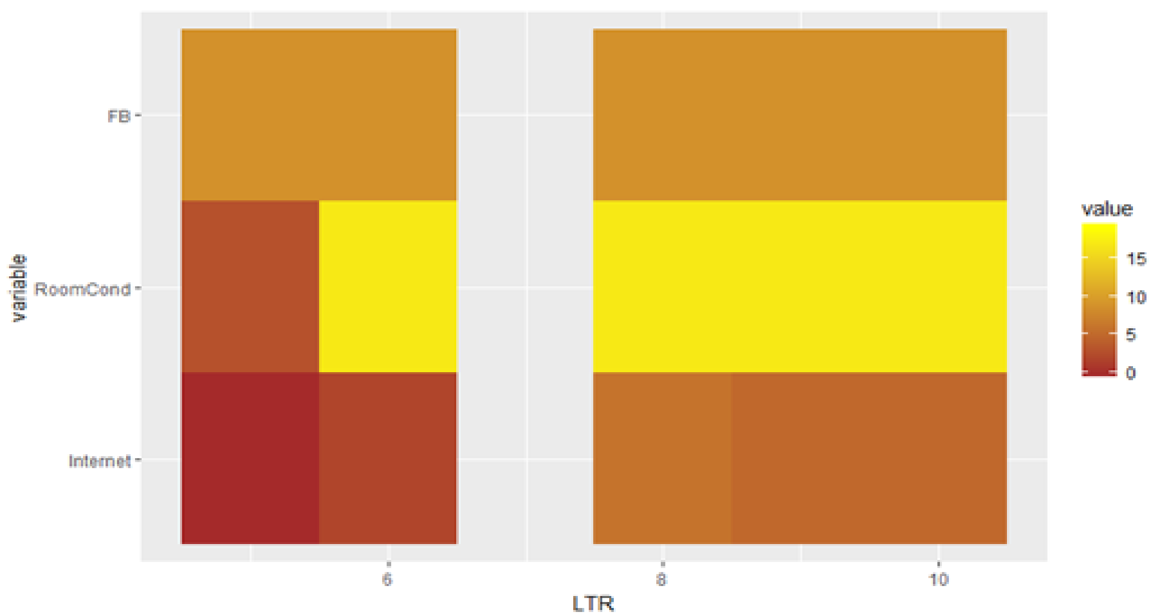
The following is a scatter plot which shows all the three factors that can influence LTR.
This plot has x axis as LTR, y axis as Internet_H,Size of the point based on the F&B value,,shape of the point based on the Room_condition_H and color of the point based on the country's name
Scatter plot:



Line chart:
From this we can interpret that F&B is almost similar in all the three countries.

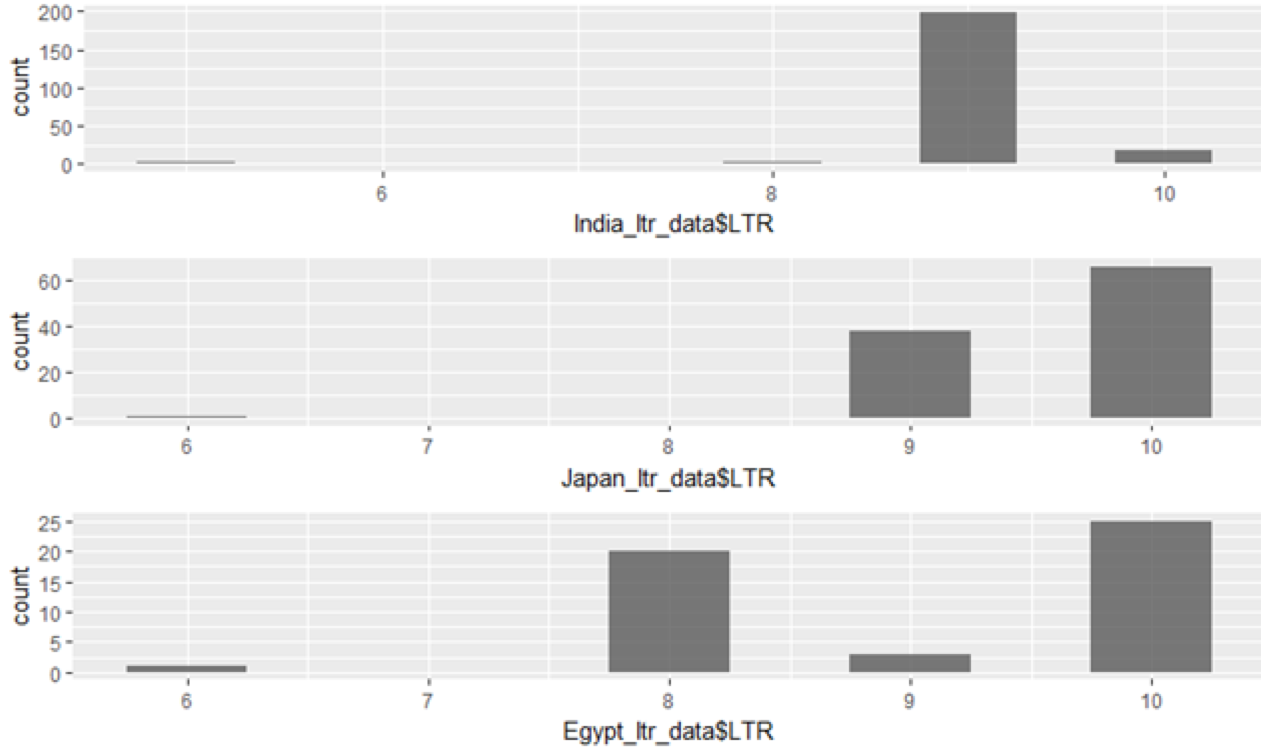


Heat map:
FB is of a constant shade in the heat map which again signifies that F&B is same across the three countries.

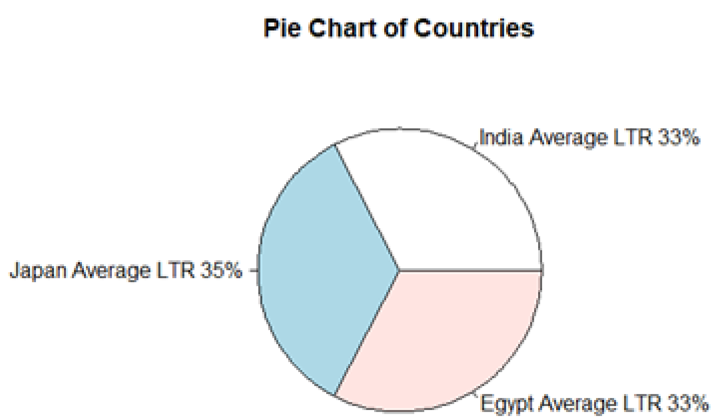


LTR Distribution:

The following graph shows the LTR distribution in India, Japan and Egypt. We can observe that no. of passives and detractors are more in Egypt than the other countries.



Pie chart:



The mean LTR of India,Japan and Egypt is plotted in a pie chart as slices.From this we can observe that Japan has the highest mean LTR.

Interpretation of results/ Actionable Insights

On analyzing the dataset and obtaining the answers for business questions which is nothing but the deliverables we have come across certain observations. Based on the observations we would like to recommend a few solutions to hotels to increase the net promoter score for that hotel. Following are the recommendations that we would like to give:

- **Improvement in Food and Beverages department in Egypt Grand Hyatt group of Hotels**

It is observed that the number of passive and detractors is comparatively higher in Egypt. Based on the modelling done on Egypt's data it can also be determined that the p-value of F&B department which is contributing to Overall satisfaction is the least. Hence it can be considered as the most significant factor contributing towards Egypt's LTR. The recommendation would be to improvise or provide the Food and Beverages services by the F&B department.

- **Improvisation in Room Condition department in India's Grand Hyatt group of Hotel**

Based on the modelling done on India's data it can be observed that the p-value of Room Condition contributing to Overall Satisfaction is the least. Hence it can be considered as the most significant factor contributing towards India's LTR. The recommendation would be to improvise the Room Condition in Indian Hotels.

- **Improvisation in Room Condition and F&B department in Japans Grand Hyatt group of Hotel**

Based on the modelling done on India's data it can be observed that the p-value of Room Condition is less than the F&B which is contributing to Overall Satisfaction. Room Condition has higher significance than F&B. Hence both can be considered as the significant factors contributing towards Japans LTR. The recommendation would be to improvise the Room Condition in Japanese Hotels.

- **Incorporation of Spa Services to enhance Room Condition experience**

Based on Association rules and the obtained rulesets it was noticed that the Spa Services attribute occurs with some meaningful frequency in the dataset. On exploring the same we determined that the presence of Spa services have a considerable impact on the Net Promoter Score of the country's Hotel.

By Spa Services attribute we mean Spa Online Booking, Spa F&B Offering and Spa Fitness Center.

The provision of these services will majorly affect the overall room satisfaction. Hence in general for all three countries it is recommended to improvise spa services.

Validation

Validation is important step in data modelling process.

Accuracy:

For measuring accuracy of our results, we relied upon Root-mean -square value of the models we built. RMSE is a frequently used measure of the differences between values predicted by a model and the values observed**.** We used KSVM and lm models to predict Likeliness to recommend for countries India, Japan and Egypt.

India

We used lm model and ksvm model for predicting LTR. Svm model and NB model were used to obtain the percentage good for calculating NPS

LM MODEL:

RMSE value:0.385603

KSVM model:

RMSE value:0.3849052

The KSVM model's RMSE value has the lowest RMSE value.so KSVM model is better for predicting LTR

SVM model:

Percentage good:0.9411765

NB model:

Percentage good:0.7733333

we see that the svm model is best for calculating NPS of Indian Hotels as the percentage good obtained is 94% which is very reliable based on the confusion matrix

Similarly, validation techniques were followed for other countries

Japan

LM MODEL:

RMSE value:0.755689

KSVM model:

RMSE value:0.7438377

The KSVM model's RMSE value has the lowest RMSE value.so KSVM model is better for predicting LTR

SVM model:

Percentage good:0.6571429

NB model:

Percentage good:0.612345

we see that the svm model is best for calculating NPS of Japan Hotels as the percentage good obtained is 65% which is very reliable based on the confusion matrix

Egypt

KSVM MODEL:
RMSE value:0.6126318

LM model:
RMSE value:0.8181145

The KSVM model's RMSE value has the lowest RMSE value.so KSVM model is better for predicting LTR

SVM model:
Percentage good:0.9411765
NB model:
Percentage good:0.8

We see that the svm model is best for calculating NPS of Egypt Hotels as the percentage good is 94% which is very reliable based on the confusion matrix.

Conclusion

After the entire process of Data extraction, transformation (munging), loading (consolidation), analysis, modelling and predicting, we came to a few conclusions. We identified trends and patterns in different countries and how these trends might affect the Hotels in near future. We answered our business questions by projecting the percentage of promoters in different countries. Japanese Hotels have high LTR as compared to Indian and Japanese Hyatt Hotels. We thus, have achieved our goal of recommending solutions to the low LTR problems by improving in a certain domain.

Code

https://github.com/Sanjana-Rajagopala

References

- "Linear Model." *MATLAB & Simulink*, https://www.mathworks.com/discovery/linear-model.html
- "Kernlab." *Function | R Documentation*, www.rdocumentation.org/packages/kernlab/versions/0.9-25/topics/ksvm.
- "e1071." *Function | R Documentation*, www.rdocumentation.org/packages/e1071/versions/1.6-8/topics/svm.
- "Naive Bayes Classification in R (Part 2)." *R-Bloggers*, 17 Feb. 2017, www.r-bloggers.com/naive-bayes-classification-in-r-part-2/.
- Saltz, J. and Standon, J (2017). *Introduction to data science*