**PROJECT: Hotel Bookings**

**EMIS73571: Analytics For Decision Support**

A picture containing text

Description automatically generated

**Group**

Aditya Ambre

Sanket Katoch

Under the guidance of Professor

Dr. Aurelie Thiele

Southern Methodist University

# Executive Summary:

In this report you will find a dataframe given called HotelBooking which contains the dataset of Hotel bookings done by customer where you will find different types of hotel and many other variables. By understanding this data and performing various tasks such as model prediction,visualization of the data which will help the customers in knowing the data as well as the hotel can predict the overbooking predictions. We have used various software such as tableau and R. With the help of Tableau we have visualized various data variables from those visualizations we could easily interpret the data. R has helped in making model which helps in prediction of various data such as hotel stays etc. Also in this report you will find how to handle missing data and various clustering techniques.

# Table of Contents

Contents

[Executive Summary: 2](#_Toc121953674)

[Table of Contents 3](#_Toc121953675)

[List of Figures 5](#_Toc121953676)

[List of Tables 6](#_Toc121953677)

[1.Data Understanding: 6](#_Toc121953678)

[2.Part a: Descriptive Analytics: 7](#_Toc121953679)

[Histogram of the ADR for each Hotel type and comment on how the hotel type affects the ADR: 8](#_Toc121953680)

[Comparing Resort Hotel ADR and City Hotel ADR we can see the following results: 9](#_Toc121953681)

[A graph that shows how the assigned room type may differ from the booked room type, perhaps depending on the ADR or some other factors 9](#_Toc121953682)

[Boxplots of the ADRs depending on room types, for each of the hotel types. Are there meaningful differences based on hotel types? 11](#_Toc121953683)

[A graph or multiple graphs showing how cancellations vary based on hotel type, lead times, room type, whether there will be children. 14](#_Toc121953684)

[Plotting graph on whether there will be children and lead time based on room type for Resort Hotel 16](#_Toc121953685)

[Plotting graph on whether there will be children and lead time based on room type for City Hotel 17](#_Toc121953686)

[Part 2 Predictive Analytics 18](#_Toc121953687)

[Create a model predicting ADR based on number of days in stays, hotel type, room type requested, number of days before arrival time (=lead time). 18](#_Toc121953688)

[Step 1: Importing the dataset 18](#_Toc121953689)

[Step 2: Handling the missing data 18](#_Toc121953690)

[Step 3: Splitting Dataset into Training and Testing Set 19](#_Toc121953691)

[Step4: Fitting Multiple Linear Regression to the Training set 19](#_Toc121953692)

[Step5: After that we predict the test results and R squared on Testing sets 20](#_Toc121953693)

[Create a model predicting cancellation (yes/no) based on hotel type, lead time, room type requested, number of days in stays 21](#_Toc121953694)

[Step 1: Remove unwanted Columns 21](#_Toc121953695)

[Step 2: Splitting Dataset into Training and Testing Set 21](#_Toc121953696)

[Step 3: Fitting Logistic Regression to the Test set 22](#_Toc121953697)

[Step 4: Visualizing the Test Results: 24](#_Toc121953698)

[Step 5: Evaluating the Model 25](#_Toc121953699)

[Clustering The hotel stay with respect to Lead time 25](#_Toc121953700)

[4.Graduate Exceptional Work: EMIS 7357 Only 29](#_Toc121953701)

[Create and evaluate two different models predicting ADR using validation sets 29](#_Toc121953702)

[Step 1: Separting data into Resort Hotel Booking and City Hotel Bookings 29](#_Toc121953703)

[Step 2: Splitting the data into train and test data 29](#_Toc121953704)

[Step3: First we use Resort Hotel dataset and apply linear Regression 29](#_Toc121953705)

[Step 4: Calculating the Accuracy of Resort Hotel 29](#_Toc121953706)

[Step 5: Calculating the Accuracy of City Hotel 30](#_Toc121953707)

[Step 6: Comparing the model using validation sets 30](#_Toc121953708)

[Step 7: Checking Accuracy of Resort Model using CART 30](#_Toc121953709)

[Step 8:City Hotel Accuracy using CART 31](#_Toc121953710)

[Predicting Overbooking Threshold. 33](#_Toc121953711)

[5.Conclusion: 35](#_Toc121953712)

[References: 36](#_Toc121953713)

# List of Figures

[Figure 1:ADR V Hotel type 8](#_Toc121953714)

[Figure 2:Summaries of City Hotel and Resort Hotel 9](#_Toc121953715)

[Figure 3:ADR For Assigned Room and ADR for Reserved Room 10](#_Toc121953716)

[Figure 4:Boxplot ADR of Each Hotels 12](#_Toc121953717)

[Figure 5:Cancellations in Resort Hotel 14](#_Toc121953718)

[Figure 6:Cancellations in City Hotel 15](#_Toc121953719)

[Figure 7:Children and Lead time based Cancellations 16](#_Toc121953720)

[Figure 8:Children and lead time based cancellations in City Hotel 17](#_Toc121953721)

[Figure 9:Linear Regression Analysis 20](#_Toc121953722)

[Figure 10:After removing unwanted data 21](#_Toc121953723)

[Figure 11: Summary training set 22](#_Toc121953724)

[Figure 12: Summary Testing Set 22](#_Toc121953725)

[Figure 13:Logisitc Regression 23](#_Toc121953726)

[Figure 14:Logisitc Regression After Removing non significant variable 23](#_Toc121953727)

[Figure 15:Confusion Matrix 24](#_Toc121953728)

[Figure 16:ROC Curve 25](#_Toc121953729)

[Figure 17 Dendogram created with "Complete" Method 26](#_Toc121953730)

[Figure 18 Dendogram created with Ward.D2 method 27](#_Toc121953731)

[Figure 19:Cluster Visualization 27](#_Toc121953732)

[Figure 20:Dendogram Entanglement 28](#_Toc121953733)

[Figure 21:Optimal Cluster 29](#_Toc121953734)

[Figure 22:CART model of Resort Hotel 30](#_Toc121953735)

[Figure 23:Summary of CART Resort 30](#_Toc121953736)

[Figure 24:CART City Hotel 31](#_Toc121953737)

[Figure 25:Summary Cart of City Hotel 31](#_Toc121953738)

[Figure 26:CART Model City 32](#_Toc121953739)

[Figure 27:Mean,sd summary 33](#_Toc121953740)

[Figure 28:Actual Occupancy v Overbooking Threshold 33](#_Toc121953741)

[Figure 29:ROC Curve Overbooking Threshold 34](#_Toc121953742)

# List of Tables

[Table 1:ADR for Assigned Room Type 10](#_Toc121953743)

[Table 2:ADR For Reserved Room Type 11](#_Toc121953744)

[Table 3:ADR of each room type in City Hotel 12](#_Toc121953745)

[Table 4:ADR of each room type in Resort Hotel 13](#_Toc121953746)

[Table 5:Room Wise Cancellations in Resort Hotel 14](#_Toc121953747)

[Table 6:Room Wise Cancellations in City Hotel 15](#_Toc121953748)

[Table 7:Room Type data about children,lead time,cancellations 16](#_Toc121953749)

[Table 8:Room Type data about children,lead time,cancellations in City Hotel 18](#_Toc121953750)

# 1.Data Understanding:

The record given is a Hotel bookings dataset which consists data of two hotels: A resort Hotel and a City Hotel. This 2 types of hotels then consist record about:

Each booking: It consists of each booking record whether the booking was cancelled or not.

How many days in advance of arrival was it booked: Record also consists of data on how much earlier was the booking was made it is in the format of year,month prior to their arrival.

Day of arrival: This record consist data of on which date is the guest arriving

Number of adults: It consist data of how many adults are there in the booking made

Number of children: It consist data of how many Children are there in the booking made.

Number of babies: It consist data if how many infants are there in the booking made.

Meal plan: This record describes what is the meal plan each and every guest has chosen for them (BB=Bed and Breakfast, FB = Full Board = 3 meals a day, HB = Half Board = breakfast and evening meals)

Country of residence: This record describes in which country does the guest resides

Whether the customer is a repeated guest: If the guest has previously stayed in the hotel

Whether there have been previous cancellations: If they guest had any previous cancellations

Reserved room type: What type of room is reserved by the guest

Assigned room type: What type of room has been assigned to the guest by hotel staff

Booking changes: If there was any booking changes made

Deposit type: What type of payment and deposit was done by the guest

Agent: The Agent that helped the guest make reservation

Days in waiting list: How many days is the guest in the waiting list

Customer type: What type of customer is the guest

Average daily rate (adr): Average daily rate of people in the hotel

required car parking spaces: The required parking spaces for the guest

total of special requests: Total number of special requests asked the guest

reservation status: what is the status of the reservation made by the guest whether it is confirmed or not.

# 2.Part a: Descriptive Analytics:

## Histogram of the ADR for each Hotel type and comment on how the hotel type affects the ADR:

Average daily rate is one of the key hotel performance indicators. It is the measure of the average paid for rooms sold in a given time period. The metric covers only revenue-generating guestrooms. ADR is calculated by dividing room revenue by rooms sold. The metric is of course applicable for any currency.

ADR = Room Revenue/Rooms Sold.

We have compared 2 types of hotels: City Hotel and Resort Hotel and below is the Histogram of Average Daily Rate of each hotel type.

Chart, bar chart

Description automatically generated

Figure 1:ADR V Hotel type

The average daily rate of city hotel is $8.35M and average daily rate of Resort Hotel is $3.80M

As we can see from the above histogram the average daily rate of city hotel is way higher than average daily rate of resort hotel.

Graphical user interface, text, application

Description automatically generated

Figure 2:Summaries of City Hotel and Resort Hotel

To show why the average daily rate of the city hotel is higher than average daily rate of resort hotel we ran a summary command on adr of both the hotel and the above image is the description of it.

## Comparing Resort Hotel ADR and City Hotel ADR we can see the following results:

* Data of city hotel is much higher than resort hotel There are 79330 data entries of City hotel whereas there is 40060 data entries of resort hotel
* Mean and Median of the City hotel is Higher than the resort hotel
* Since Max value of City Hotel is way higher than the resort hotel there are multiple days when the rate of city hotel is way higher than the resort hotel. Max rate of city hotel one day was $5400 whereas the max rate of resort hotel one day was $508.

The above reasons are why the Average daily rate of City Hotel was much higher than the resort Hotel.

## A graph that shows how the assigned room type may differ from the booked room type, perhaps depending on the ADR or some other factors

We have designed a graph which compares the average daily rate for assigned vs the average daily rate for reserved room type.

Chart, bar chart

Description automatically generated

Figure 3:ADR For Assigned Room and ADR for Reserved Room

|  |  |
| --- | --- |
| Room Type | ADR |
| A | $93.14 |
| B | $94.45 |
| C | $113.42 |
| D | $107.45 |
| E | $117.70 |
| F | $151.89 |
| G | $166.53 |
| H | $171.38 |
| I | $40.84 |
| K | $53.70 |
| L | $8.00 |

Table 1:ADR for Assigned Room Type

|  |  |
| --- | --- |
| Room Type | ADR |
| A | $90.80 |
| B | $90.36 |
| C | $160.22 |
| D | $120.68 |
| E | $124.54 |
| F | $167.69 |
| G | $176.00 |
| H | $188.22 |
| L | $124.67 |

Table 2:ADR For Reserved Room Type

After comparing the average daily rate of Assigned room and Reserved room type we can conclude that mostly the ADR of the reserved room type is much higher than the ADR of the Assigned room type except for the room type A,B in which the Assigned daily rate of the Assigned room is slightly higher than Average daily rate for reserved room type A and B.

## Boxplots of the ADRs depending on room types, for each of the hotel types. Are there meaningful differences based on hotel types?

There are 2 types of hotel: Resort Hotel and City Hotel where with the help of boxplots we have plotted the average daily rate of each room type which is as follows:

Chart, box and whisker chart

Description automatically generated

Figure 4:Boxplot ADR of Each Hotels

As we can see from the boxplot graph in city hotel the room type G has the highest average daily rate. Whereas, in Resort Hotel the room type H has the highest average daily rate.

Below is the average daily rate of each room in city and resort hotel:

|  |  |
| --- | --- |
| Room Type | ADR |
| A | $97 |
| B | $90 |
| C | $85 |
| D | $132 |
| E | $159 |
| F | $191 |
| G | $206 |
| K | $54 |

Table 3:ADR of each room type in City Hotel

|  |  |
| --- | --- |
| Room Type | ADR |
| A | $81 |
| B | $105 |
| C | $164 |
| D | $104 |
| E | $115 |
| F | $133 |
| G | $169 |
| H | $190 |
| I | $41 |
| L | $8 |

Table 4:ADR of each room type in Resort Hotel

In City Hotel the room with highest average daily rate is Room G with $206 and the room with lowest average daily rate is Room K with $54

Meanwhile, In Resort Hotel the highest average daily rate is of room H whereas Room L has the lowest Average daily rate with $8

## A graph or multiple graphs showing how cancellations vary based on hotel type, lead times, room type, whether there will be children.

Diagram

Description automatically generated

Figure 5:Cancellations in Resort Hotel

The above graph is the total cancellations based on room type we can see that there are 43,439 cancellations in which there are 32,696 cancellations in city hotel whereas there are 10,743 cancellations in resort hotel. The cancellations based on room type in resort hotel are as follows:

|  |  |
| --- | --- |
| Room Type | Cancellations |
| A | 6046 |
| B | 0 |
| C | 300 |
| D | 1934 |
| E | 1389 |
| F | 178 |
| G | 643 |
| H | 245 |
| I | 5 |
| L | 1 |
| P | 2 |

Table 5:Room Wise Cancellations in Resort Hotel

From the above table we can see that Room A is the room with the highest cancellations with 6046.

Diagram

Description automatically generated

Figure 6:Cancellations in City Hotel

The above graph is cancellations in city hotel based on room type.

The cancellations based on room type are as follows:

|  |  |
| --- | --- |
| Room Type | Cancellations |
| A | 26886 |
| B | 366 |
| C | 5 |
| D | 4107 |
| E | 497 |
| F | 696 |
| G | 117 |
| K | 12 |
| P | 10 |

Table 6:Room Wise Cancellations in City Hotel

From the above table we can see that in city hotel also the room with the highest cancellations is room type A with 26886 cancellations.

## Plotting graph on whether there will be children and lead time based on room type for Resort Hotel

Graphical user interface

Description automatically generated

Figure 7:Children and Lead time based Cancellations

Here the above graph depicts whether there will be children based on room type in resort hotel in order with cancellations and also the above part of the graph depicts lead time in order with cancellations.

Resort Hotel:

|  |  |  |  |
| --- | --- | --- | --- |
| Room Type | Children | Cancellations | Lead time |
| A | 405 | 6046 | 1,798,719 |
| B | 0 | 0 | 25 |
| C | 1125 | 300 | 71,817 |
| D | 291 | 1934 | 675,574 |
| E | 174 | 1389 | 455,884 |
| F | 81 | 178 | 59,975 |
| G | 2137 | 643 | 141,295 |
| H | 572 | 245 | 46,467 |
| I | 54 | 5 | 24,181 |

Table 7:Room Type data about children,lead time,cancellations

From the above data we can see that in resort hotel the room with highest children is Room G with 2137. Whereas as seen previously Room A has the highest cancellations also the room with highest lead time is Room A.

## Plotting graph on whether there will be children and lead time based on room type for City Hotel

Graphical user interface

Description automatically generated with medium confidence

Figure 8:Children and lead time based cancellations in City Hotel

Here the above graph depicts whether there will be children based on room type in city hotel in order with cancellations and also the above part of the graph depicts lead time in order with cancellations.

City Hotel:

|  |  |  |  |
| --- | --- | --- | --- |
| Room Type | Children | Cancellations | Lead Time |
| A | 1741 | 26,886 | 6,966,520 |
| B | 518 | 366 | 107,070 |
| C | 2 | 5 | 704 |
| D | 412 | 4,107 | 893,470 |
| E | 370 | 497 | 95,241 |
| F | 2746 | 696 | 129,455 |
| G | 485 | 117 | 24,140 |

Table 8:Room Type data about children,lead time,cancellations in City Hotel

From the above data we can conclude that the room with highest children is room F whereas the cancellations and lead time the room with highest is Room A.

We can conclude after observing the data that the Room with the highest cancellations also has the highest lead time.

# Part 2 Predictive Analytics

## Create a model predicting ADR based on number of days in stays, hotel type, room type requested, number of days before arrival time (=lead time).

To create a model predicting ADR based on number of days in stays, hotel type, room type requested, number of days before arrival time (=lead time) we have to perform some steps.

### Step 1: Importing the dataset

To create a model predictor we have to import the dataset “Hotel Bookings” which have data values such as hotel type, room type, assigned room, reserved time, average daily rate, Arrival days etc. After importing the data we could see that there are 11930 variables.

### Step 2: Handling the missing data

Missing values, as the name implies, are elements that are unknown. NA and NaN are reserved words in the R programming language that denote a missing value for undefined q arithmetic operations. The modeling functions in R language acknowledge a na.action argument which provides instructions to the function regarding its response if NA comes in its way and hence this way the function calls one of the missing value filter functions. Missing Value Filter Functions alter the data set and in the new data set the value of NAs has been changed. The default Missing Value Filter Function is na.omit. It omits every row containing even one NA.

Some other Missing Value Filter Functions are:

* na.omit– omits every row containing even one NA
* na.fail– halts and does not proceed if NA is encountered
* na.exclude– excludes every row containing even one NA but keeps a record of their original position
* na.pass– it just ignores NA and passes through it

### Step 3: Splitting Dataset into Training and Testing Set

Split data into train and test in r, It is critical to partition the data into training and testing sets when using supervised learning algorithms. We have divided the dataframe hotel bookings into 2 different parts Train data and Test data and have split ratio = 0.80 which is 80% where divided 80% is set as training data and 20% is set as test data.

### Step4: Fitting Multiple Linear Regression to the Training set

Multiple linear regression is an extension of simple linear regression that uses multiple distinct predictor variables to predict an outcome variable (y) (x).

The following equation expresses the prediction of y using three predictor variables (x):

y = b0 + b1\*x1 + b2\*x2 + b3\*x3

The "b" values are referred to as the regression weights (or beta coefficients). They assess the relationship between the predictor variable and the outcome. "b j" can be defined as the average effect on y of a one-unit increase in "x j," with all other predictors held constant.

The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary.

Table

Description automatically generated

Figure 9:Linear Regression Analysis

In our dataset, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero.

### Step5: After that we predict the test results and R squared on Testing sets

To compute we have to calculate SSE and SST

**SST:** SST stands for the sum of squared differences between the observed dependent variable and its mean. Consider this to be the dispersion of observed variables around the mean, similar to variance in descriptive statistics. Here we calculate SST as **SST = sum((test\_set$adr - mean(training\_set$adr))^2).**

**SSE:** SSE term is the sum of squares error, or SSE. The error is the difference between the observed value and the predicted value. We normally try to keep the error to a minimum. The lesser the error, the higher the regression's estimate power. Here we calculate SSE as **SSE = sum((test\_set$adr - y\_pred)^2).**

**R-Squared:** The coefficient of determination, R-squared, is defined as the ratio of the sum of squares explained by a regression model to the "total" sum of squares around the mean.

**1-SSE/SST** is the formula for R-Squared by using this formula on our dataset we get R-Squared as **0.4794058.**

## Create a model predicting cancellation (yes/no) based on hotel type, lead time, room type requested, number of days in stays

Ans: To Create model predicting based on hotel type, lead time, room type requested, number of days in stays we followed these steps:

### Step 1: Remove unwanted Columns

We removed some of the columns which are not necessary in our dataset. The columns that were removed were arrival\_date\_year, stays\_in\_week\_nights, babies, previous\_bookings\_not\_canceled, assigned\_room\_type, customer\_type, required\_car\_parking\_spaces, reservation\_status\_date.

After removing the columns these are the remaining variables in the dataset

A picture containing text

Description automatically generated

Figure 10:After removing unwanted data

### Step 2: Splitting Dataset into Training and Testing Set

Then we split the Hotel bookings into training and testing set with split ratio 0.80.After splitting the data we created the feature scaling. Feature scaling is the process of eliminating units of measurement for variables within a dataset, and is often carried out to boost the accuracy of a machine learning algorithm. We created feature scaling on test and train data and labelled them as scaled\_training\_set and scaled\_testing\_set. Then, we Attached non-numeric variables to the scaled data sets and these were the results:

**Summary(training\_set)**

A screenshot of a computer

Description automatically generated with medium confidence

Figure 11: Summary training set

**Summary(testing\_set)**

A picture containing table

Description automatically generated

Figure 12: Summary Testing Set

### Step 3: Fitting Logistic Regression to the Test set

Using the package caret, we create a training set (“Train”) and a testing set (“Test”) out of our data set. The training set will be 80% of the full data set, and the testing set will be 20%. Next, we carry out the logistic regression on the training set. For that we use the glm function, where glm stands for general linear model.

Table

Description automatically generated

Figure 13:Logisitc Regression

After that we remove the non-significant columns and the result are as follows:

Table

Description automatically generated

Figure 14:Logisitc Regression After Removing non significant variable

We removed the total\_stays.1 column. The goal is to see how well the logistic regression model performs in comparison to the testing set. We must quantify the model's outcomes in some way. When we apply the predict function to the model and the training set, we will get a probability for each observation. We save those probabilities to the testing set under the variable "prob\_pred" and convert the results in binary results and create a confusion matrix.

A confusion matrix is a table that categorizes predictions based on their actual values. It has two dimensions, one of which will show the anticipated values and the other will show the actual values and the results are as follows:

A picture containing text

Description automatically generated

Figure 15:Confusion Matrix

### Step 4: Visualizing the Test Results:

ROC visualizes two metrics as follows — Sensitivity / true positive rate: It measures the proportion of actual positives that are correctly identified. Specificity / true negative rate: It measures the proportion of actual negatives that are correctly identified. When a model is built, ROC curve — Receiver Operator Characteristic Curve can be used for checking the accuracy of the model. The area under the ROC curve is called as AUC -Area Under Curve. AUC ranges between 0 and 1 and is used for successful classification of the logistics model.

Chart, line chart

Description automatically generated

Figure 16:ROC Curve

### Step 5: Evaluating the Model

The Accuracy for logistic Regression with cut-off=0.5 is 66.92% =0.6692

The Accuracy for logistic Regression with cut-off=0.7 is 64.79% =0.6479

The Accuracy for logistic Regression with cut-off=0.2 is 55.51%=0.5551

We can conclude that as the cut-off is low the accuracy of the model also decreases.

# Clustering The hotel stay with respect to Lead time

Lead time in the data is described as the number of days before arrival time. Since lead time is one the important variable we have used hierarchical clustering and visualized the clusters using R.

Hierarchical Clustering: For identifying groupings in a dataset, hierarchical clustering is an alternative to k-means clustering. It does not require us to provide the number of clusters to be produced, like the k-means approach requires.

To Find the clusters we have followed these steps:

Data Preparation: In order to run a cluster analysis in R, the data should be prepared as follows:

* The rows represent observations (individuals), while the columns represent variables.
* Any missing data value must be eliminated or estimated.
* To make variables comparable, the data must be normalized (i.e., scaled).

Here we have used HotelBookings Dataset and did all the steps written above

Agglomerative Hierarchical Clustering-Firstly we performed Agglomerative Hierarchical Clustering, With hclust, we may do agglomerative HC. We first compute the dissimilarity values with dist, then feed these values into hclust and define the agglomeration method (e.g., "complete", "average", "single", "ward.D"). The dendrogram can then be plotted.

Chart, histogram, waterfall chart

Description automatically generated

Figure 17 Dendogram created with "Complete" Method

We also used Agnes function to get the agglomerative coefficient which shows the closer the number to 1 the stronger clustering structure. The coefficient turned out to be 0.96.This suggest a strong cluster structure.

Similary we used ward’s method to plot the dendogram and form 4 clusters around the data which resulted in this:

Diagram

Description automatically generated with medium confidence

Figure 18 Dendogram created with Ward.D2 method

Here we can see 4 clusters coloured as red,green,dark blue and light blue.

To help it visualize it better we plotted it in a different way which resulted in a graph like these:

Chart

Description automatically generated

Figure 19:Cluster Visualization

Here we can see 4 different clusters as Red, Blue, Purple and green.

We used 2 different method in clustering

hc1 <- hclust(res.dist, method = "complete")

hc2 <- hclust(res.dist, method = "ward.D2")

and visualized it to see difference between 2 dendograms

Diagram, schematic

Description automatically generated

Figure 20:Dendogram Entanglement

The entanglement =0.09 shows that there is not much difference between the 2 methods of clustering. A lower entanglement coefficient corresponds to a good alignment.

We used silhouette method to determine the cluster and the result was this:

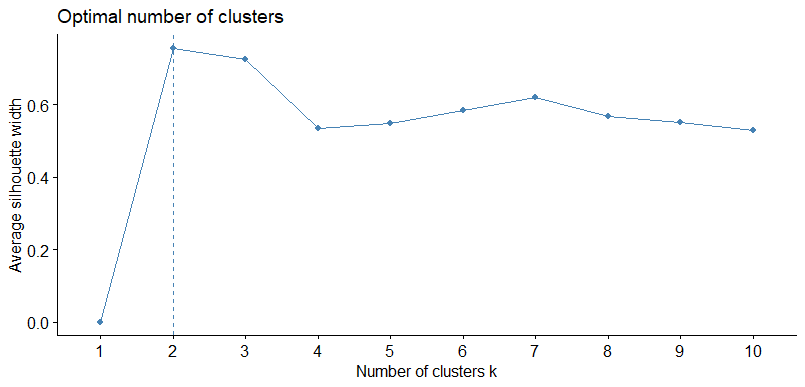


Figure 21:Optimal Cluster

**So the optimal number of Clusters should be 2**

# 4.Graduate Exceptional Work: EMIS 7357 Only

# Create and evaluate two different models predicting ADR using validation sets

## Step 1: Separting data into Resort Hotel Booking and City Hotel Bookings

We have separated the data into 2 section Hotel\_Booking\_Resort and Hotel\_Bookings\_City in one data all will be of City Hotel and other one will be of Resort Hotel.

## Step 2: Splitting the data into train and test data

We split the data into train and test data with split ratio=0.8

## Step3: First we use Resort Hotel dataset and apply linear Regression

As done in 2a we use the same method to calculate the linear regression for the resort hotel bookings and calculate the regressor and since after calculating the regressor we predict it on test data

## Step 4: Calculating the Accuracy of Resort Hotel

We firstly calculate the SSE then we calculate the SST and to calculate the Accuracy we use the formula 1-SSE/SST which results in 0.72. Therefore the Accuracy of the model is 72%

## Step 5: Calculating the Accuracy of City Hotel

Similarly we follow all the above steps and calculate the accuracy of City Hotel which results in 1-SSE/SST =0.477. Therefore the Accuracy of the model is 47%.

## Step 6: Comparing the model using validation sets

To compare the model using validation sets we first have to calculate the validation set.So, Firstly we calculate the Validation set.

## Step 7: Checking Accuracy of Resort Model using CART

Diagram

Description automatically generated

Figure 22:CART model of Resort Hotel

Text, letter

Description automatically generated

Figure 23:Summary of CART Resort

From the above figure we could the summary of the CART algorithm where Lead\_time is one of the most important variables.

**Accuracy of the resort model =69% i.e 0.6900**

## Step 8:City Hotel Accuracy using CART

Diagram

Description automatically generated

Figure 24:CART City Hotel

Text, letter

Description automatically generated

Figure 25:Summary Cart of City Hotel

Even in city Hotel we can see lead time is the most important variable

Diagram

Description automatically generated

Figure 26:CART Model City

**Accuracy of the city model =70% i.e 0.7015905**

# Predicting Overbooking Threshold.

Step1: To predict overbooking Threshold we firstly couple all the bookings by their year,month and day .

Step2: Then calculate Hotel capacity which is calculated by all the assigned rooms in the hotel addition with all the cancellations.

Step3: Calculate Predicted occupancy , overbooking threshold and actual occupancy

Predicted occupancy is calculated by all the assigned rooms in the hotel whereas overbooking threshold is calculated by hotel booking capacity- hotel booking cancelled and actual occupancy is calculated by assigned room – bookings cancelled.

Step4: Calculate mean and sd of all the occupancy

Text

Description automatically generated

Figure 27:Mean,sd summary

Step 5: Plotting the actual occupancy and overbooking threshold

Chart, line chart

Description automatically generated

Figure 28:Actual Occupancy v Overbooking Threshold

Step 6: Plot ROC Curve of Overbooking Threshold

Chart

Description automatically generated

Figure 29:ROC Curve Overbooking Threshold

Threshold value nearest to 1, i.e 1.5

# 5.Conclusion:

From all the research,tutorial,knowledge gathered we have successfully learned and implemented data and visualization techniques using Rstudio and also gained an insight into the data and various techniques used for different types of visualization. We developed relationship between different attributes which helped in plotting them on a single graph. Also we used the subset function in a dataset where we reduced the number of data and analyzed it from different perspectives. We used the Hotel Bookings dataset and perfomed various tasks successfully (For Eg: Calculating ADR which could help customers in reserving the room type) and also created various model which can predict various attributes that can help Hotels.

# References:

Sites

1. <https://cran.r-project.org/web/packages/covid19.analytics/vignettes/covid19.analytics.html> [Accessed on 12-2-22]
2. <https://rkabacoff.github.io/datavis/DataPrep.html#cleaning-data> [Accessed on 12-6-22]
3. <https://uc-r.github.io/hc_clustering#algorithms> [Accessed on 12-10-22]
4. <https://uc-r.github.io/kmeans_clustering> [Accessed on 12-13-22]

Books

1. R: Data Analysis and Visualization, by Tony Fischetti, Brett Lantz, Jaynal Abedin, Released June 2016, Publisher(s): Packt Publishing, ISBN: 9781786463500
2. Visualizing Data: Exploring and Explaining Data with the Processing Environment, by Ben Fry, "O'Reilly Media, Inc."

Papers

1. Principles of Effective Data Visualization, Stephen R. Midway, Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
2. Big Data Visualization and Visual Analytics of COVID-19 Data, Carson K. Leung, Yubo Chen , Calvin S.H. Hoi , Siyuan Shang , Yan Wen , Alfredo Cuzzocrea. Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada iDEA Lab, University of Calabria, Rende, CS, Italy