Hotel Bookings Project

OREM 7357

Aditya Ambre
Sanket Katoch

14 December 2022

Delegation of Tasks

- Programming:Sanket, Aditya
- Part 1 Description Analysis: Sanket
- Part 2 Predictive Analysis : Aditya, Sanket
- Report: Aditya
- Presentation: Sanket

Abstract

We have analyzed and implemented various techniques on our dataset “Hotel Bookings”.We have used different types of visualization so that one can understand the data easily and work on it efficiently.

SMU

Problem Defination

The goal is to provide and implement various techniques of visualizations for descriptive analysis on “Hotel Bookings” dataframe and also create a predictive model analysis which will help in “Hotel bookings” dataframe.Where people could take help from visualization and interpret the data.

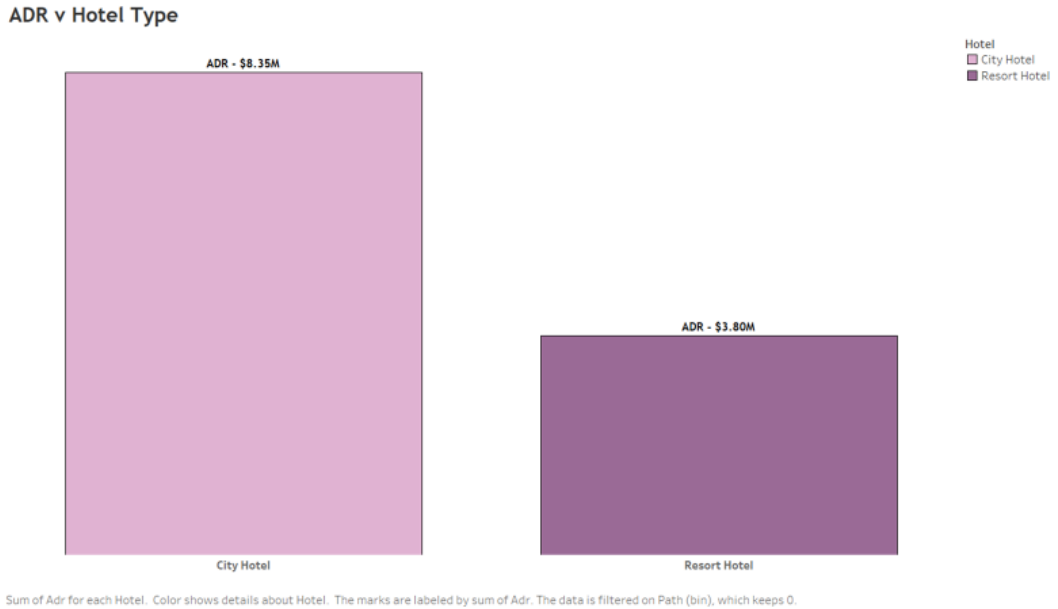
Data Understanding

To analyze and implement various we have to understand the data given. There are 2 types of Hotel in Hotel Bookings dataset City Hotel and Resort Hotel also it contains records as follows:

- Each booking
- How many days in advance of arrival was it booked
- Day of arrival
- Number of adults
- Number of children
- Number of babies
- Meal plan
- Country of residence
- Average daily rate (adr) etc.

Part 1:Descriptive Analytics

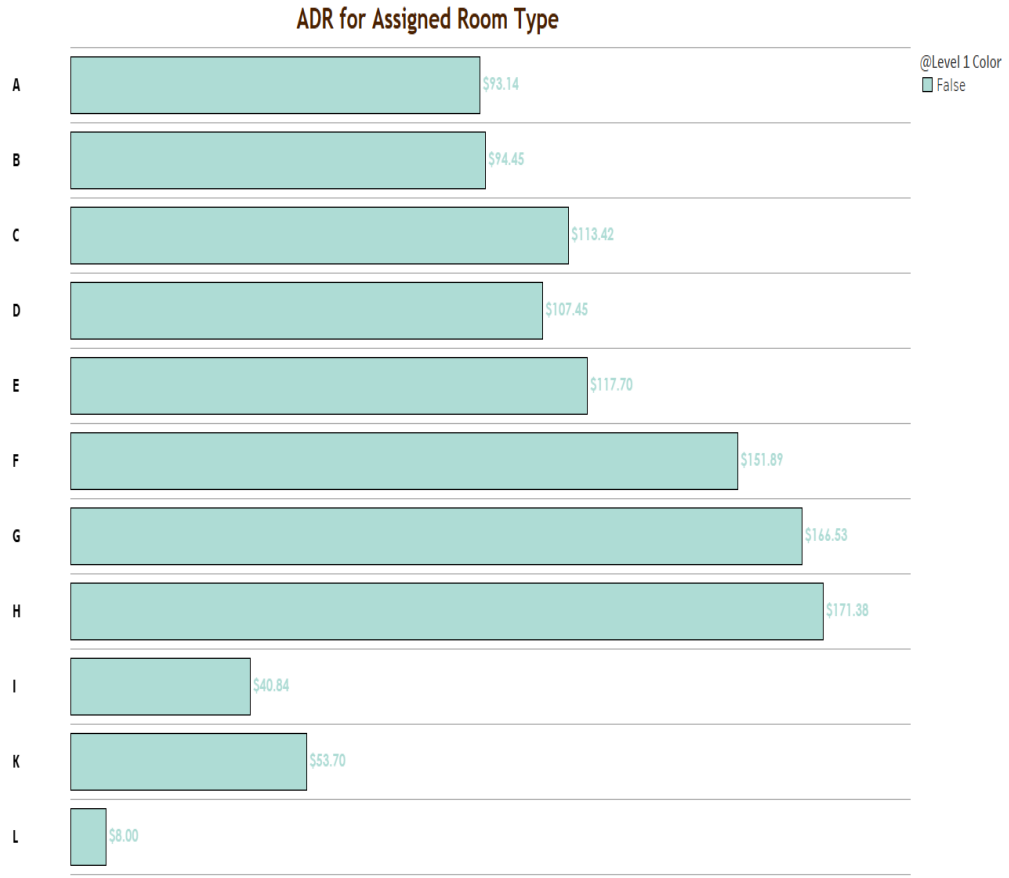
Histogram of the ADR for each Hotel type



Key takeaways from the Histogram

- ADR is average daily rate which is calculated as ADR = Room Revenue/Rooms Sold
- ADR of City Hotel is way higher than Resort Hotel where ADR of City Hotel=$8.3M ; ADR of Resort Hotel=$3.8M
- Revenue of City Hotel is higher because there are more city hotel than resort hotel
- Maximum rate of City Hotel for single day is $5400 while for Resort Hotel is $508
- The above reasons are why the Average daily rate of City Hotel was much higher than the resort Hotel.

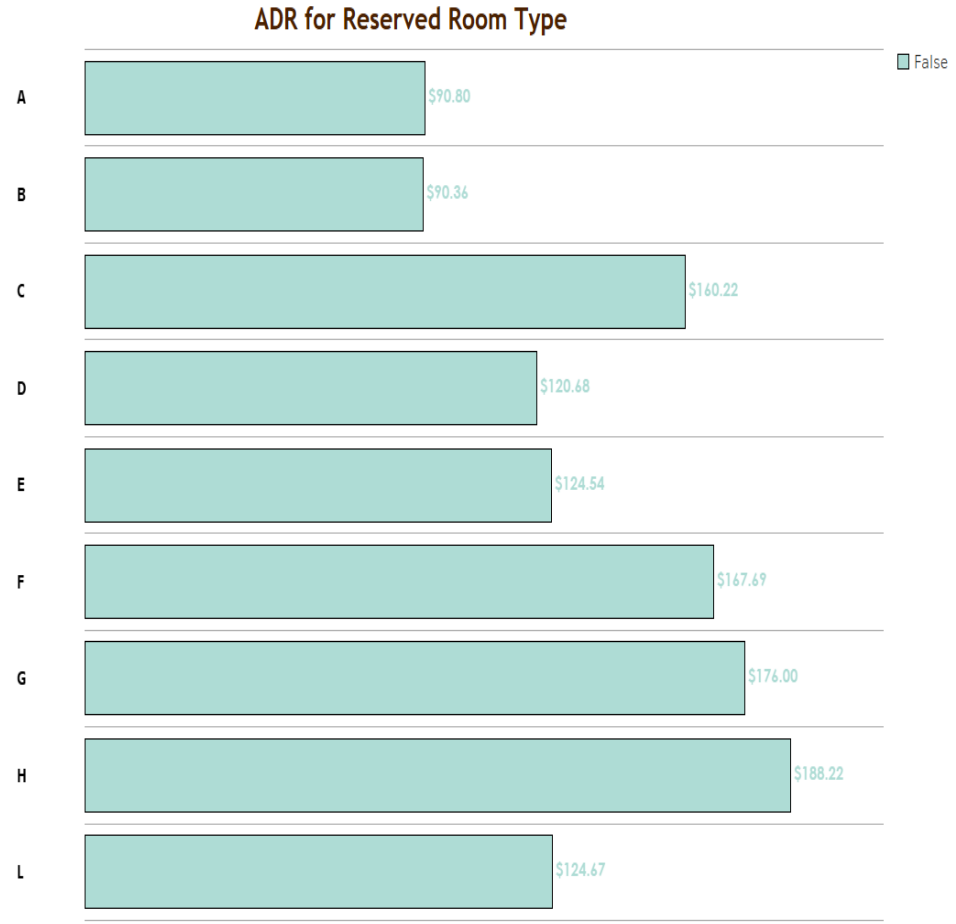
ADR For Assigned Room Type



Average of Adr for each @Level 1 broken down by Assigned Room Type and @Level 1 Header. Color shows details about @Level 1 Color. The marks are labeled by average of Adr. The data is filtered on Path (bin), which keeps 0.

- The Graph demonstrates the Average daily rate for assigned room type in city and resort hotels
- Room H has the Highest ADR revenue with $171.38
- Room L has the Lowest ADR revenue with $8.00

ADR For Reserved Room Type

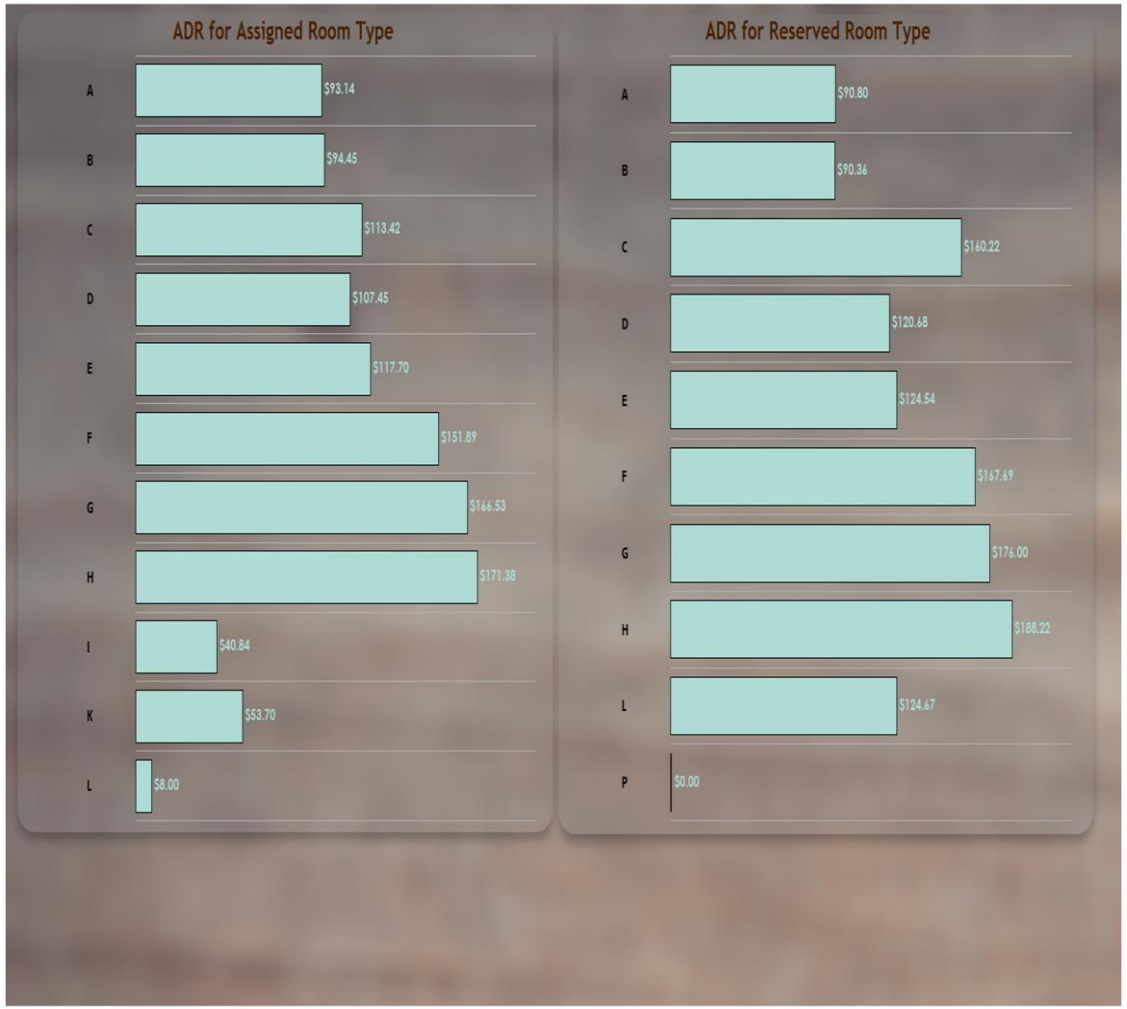


Average of Adr for each @Level 1 (copy) broken down by Reserved Room Type and @Level 1 Header (copy). Color shows details about @Level 1 Color (copy). The marks are labeled by average of Adr. The data is filtered on Path (bin), which keeps 0.

- The Graph demonstrates the Average daily rate for reserved room type in city and resort hotels
- Room H has the Highest ADR Revenue with $188.22
- Room B has the Lowest ADR Revenue with $90.36

Part 1: Descriptive Analysis

ADR for Assigned Room v ADR for Reserved Room

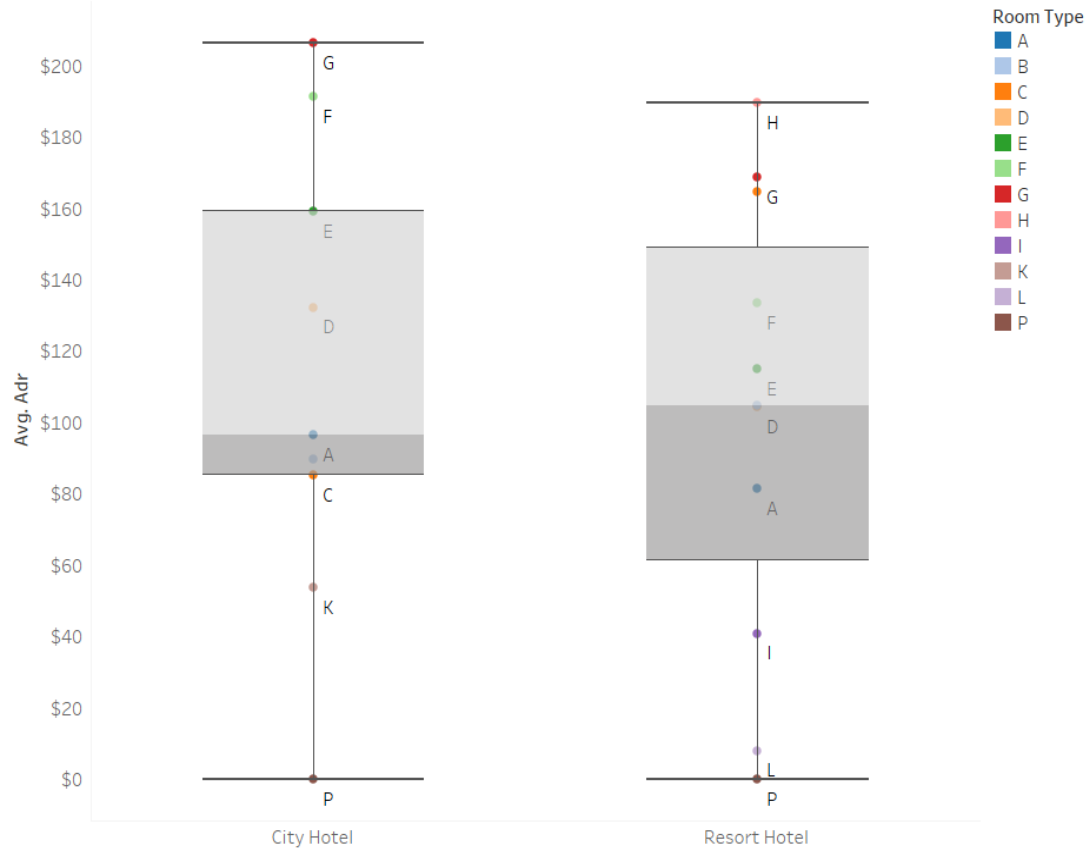


- After Comparing we can conclude that ADR for Reserved Room Type is Higher than the Assigned Room Type.
- Only in Room A and B type of Reserved Room the ADR is less than Assigned Room Type.
- Customers can be in profit if they reserved Room A and B.
- Whereas Reserving Room Type L could be a huge loss since there is such as big difference between assigned and reserved room type L.

Boxplots of ADR depending on Room Type

Boxplots of ADR depending on Room Type

c)

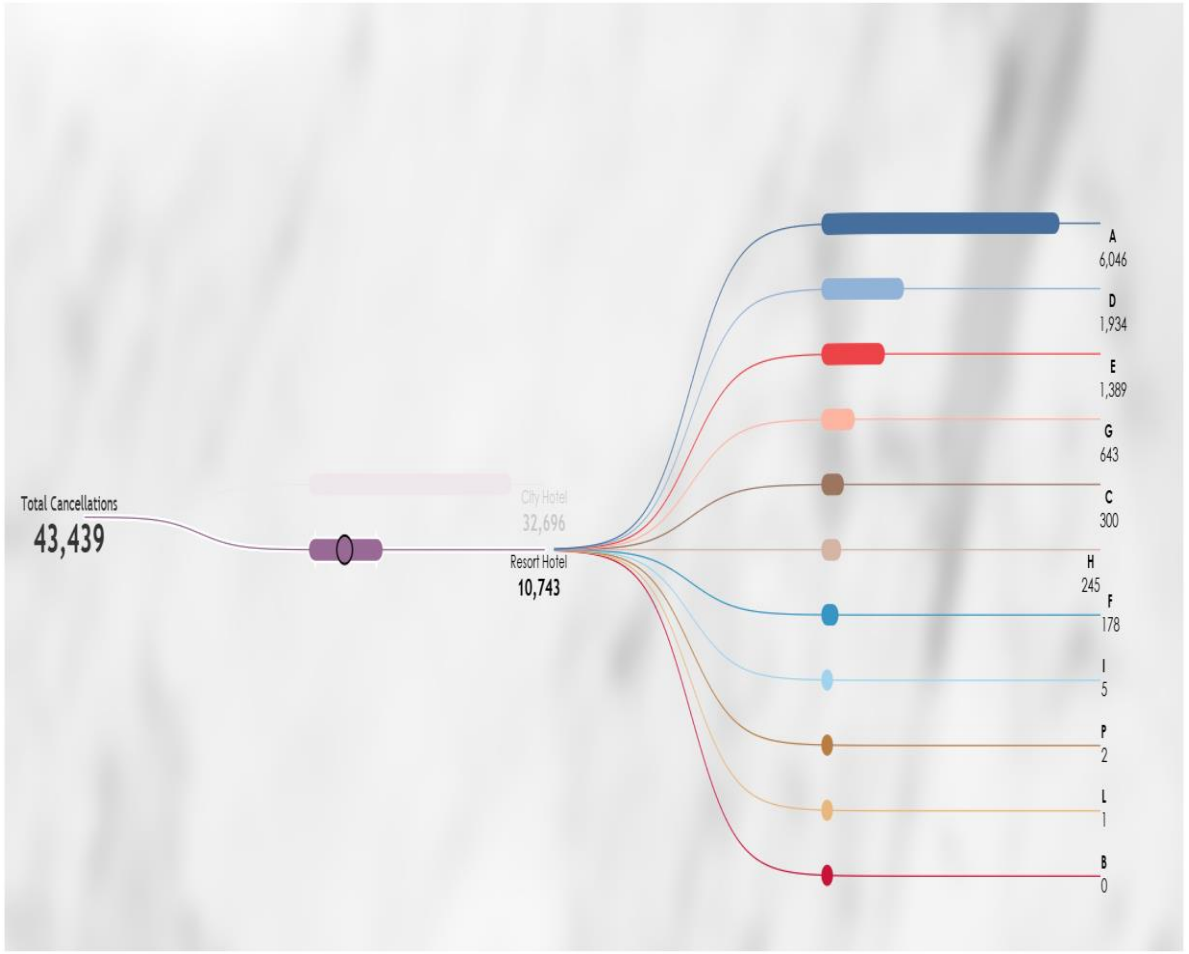


Average of Adr for each Hotel. Color shows details about @Room Type as an attribute. The marks are labeled by @Room Type. Details are shown for @Room Type. The data is filtered on @RRT Filter and Path (bin). The @RRT Filter filter keeps True. The Path (bin) filter keeps 0.

- By comparing we can city ADR of City Hotel is higher than Resort Hotel.
- Room type G has the highest average daily rate. In city hotel ($206)
- Whereas, in Resort Hotel the room type H has the highest average daily rate ($190)
- With the help of this graph customers can plan which room to take.

Cancellations based on Hotel and Room Type

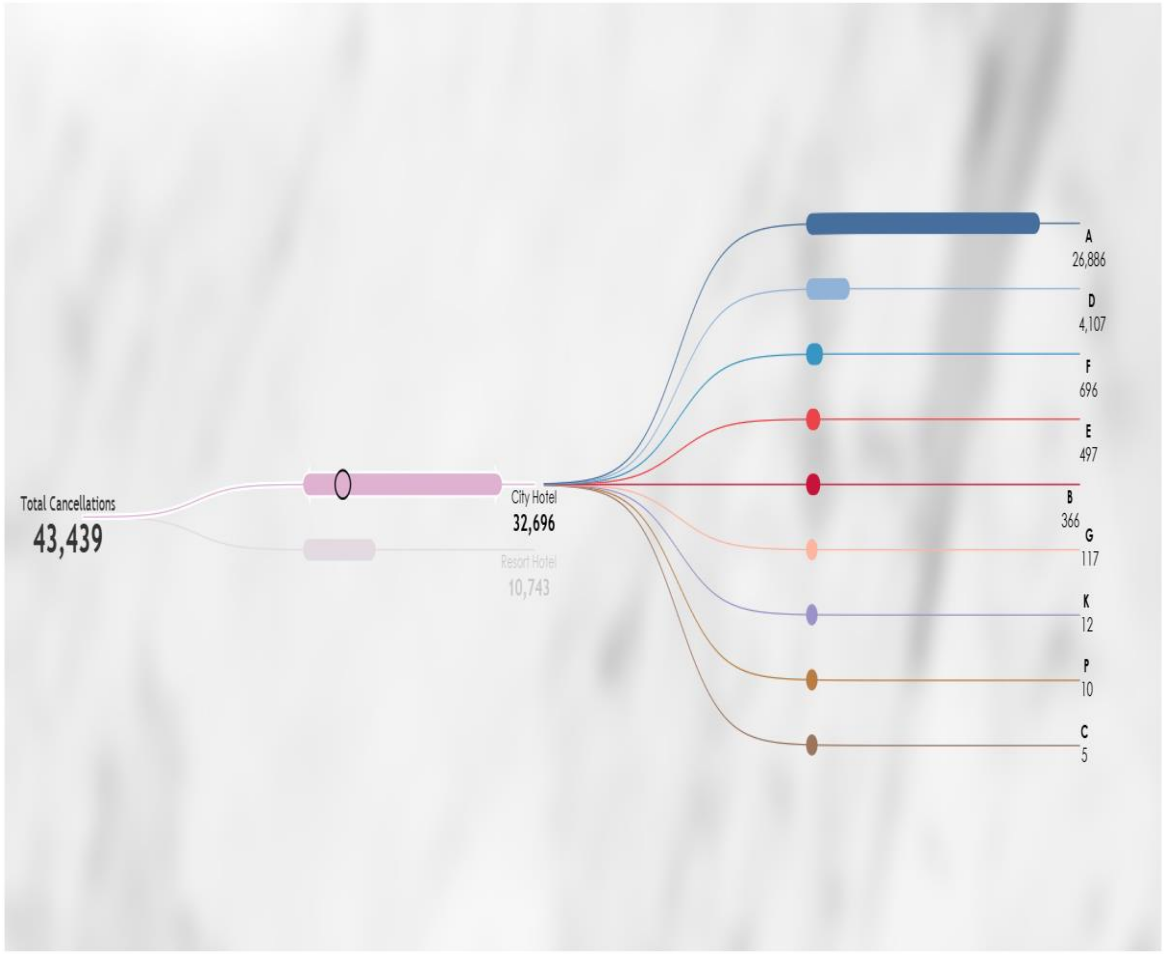
Cancellations based on Resort Hotel and Room Type



- This Graph show cancellations based on Resort Hotel varying in Room Type
- There have been 43,439 cancellations in which resort hotel has 10,743 cancellations
- Customers have cancelled Room Type A the most around 6,046 so customer should avoid booking room type A
- The room type B has 0 cancellations which shows B is the most efficient room which customer has liked the most.

Cancellations based on Hotel and Room Type

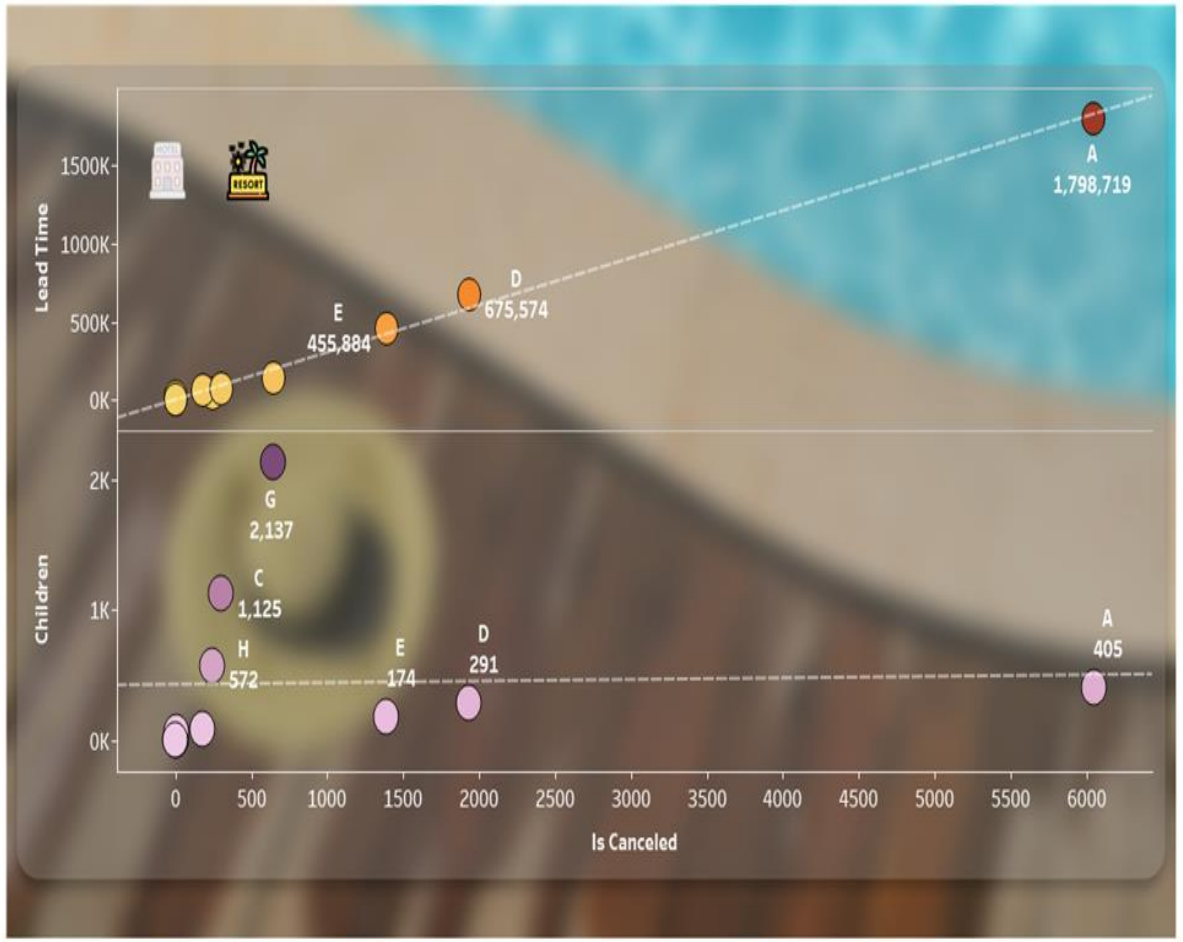
Cancellations based on City Hotel and Room Type



- This Graph show cancellations based on City Hotel varying in Room Type
- There have been 43,439 cancellations in which resort hotel has 32,696 cancellations
- Customers have cancelled Room Type A the most around 26,886 so customer should avoid booking room type A
- The room type C has 5 cancellations which shows C is the most efficient room which customer has liked the most.

Which Room Type has the most children and Lead time?

Lead time and Children based on Room type in Resort Hotel

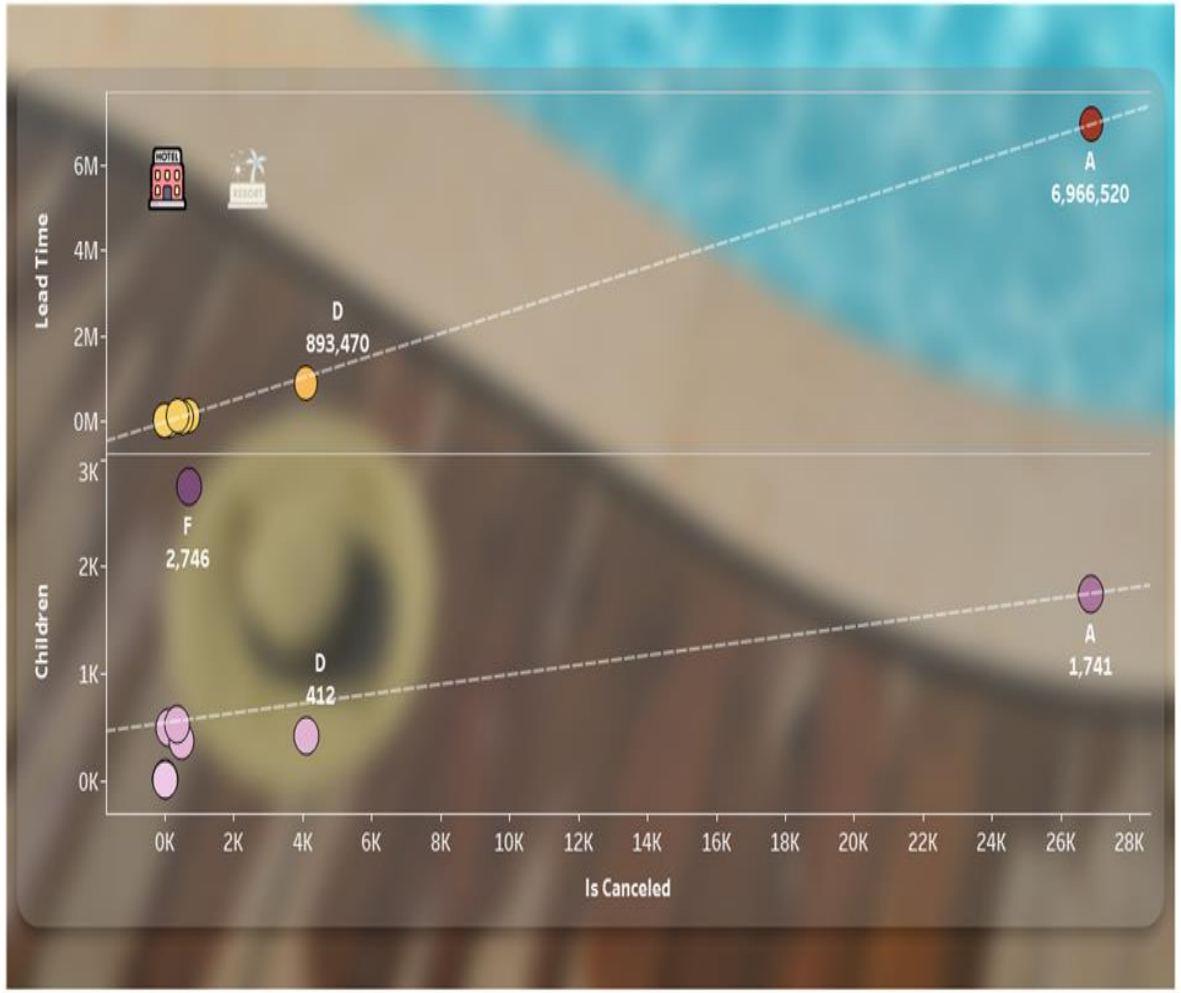


- The Below graph shows the room with most children with respect to most cancellations
- In resort hotel the room type with most children is Room G which can be interpreted as the room which is more safe for childrens.
- The above graph shows the lead time with respect to most cancellations
- In resort hotel the room type with most lead time as well as most cancellations is Room A.

SMU

Which Room Type has the most children and Lead time?

Lead time and Children based on Room type in City Hotel



- The Below graph shows the room with most children with respect to most cancellations
- In City hotel the room type with most children is Room F which can be interpreted as the room which is more safe for childrens.
- The above graph shows the lead time with respect to most cancellations
- In resort hotel the room type with most lead time as well as most cancellations is Room A.

SMU

Part 2:Predictive Analytics

To perform predictive analysis we created a model which predicted ADR based on various factors and followed these steps.

- Step1:Importing the Dataset-To create a model predictor we have to import the dataset “Hotel Bookings”.
- Step2:Handling the missing data- With the help of na.omit we took care of the missing data
- Step3:Splitting Dataset into Training and Testing Set- Splitted the data set into training set and testing set with split ratio=0.8
- Step4:Fitting Multiple Linear Regression to the Training set- We used the Linear Regression technique on the training set

SMU

Part 2:Predictive Analytics

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
total_stays                 6.341e-01  5.335e-02  11.887  < 2e-16 ***
hotel0                      5.376e+01  6.584e-01  81.645  < 2e-16 ***
hotel1                      7.419e+01  6.298e-01 117.801  < 2e-16 ***
reserved_room_typeB        -2.698e+00  1.314e+00  -2.054  0.03998 *
reserved_room_typeC         6.198e+01  1.442e+00  42.976  < 2e-16 ***
reserved_room_typeD         2.335e+01  3.579e-01  65.231  < 2e-16 ***
reserved_room_typeE         3.591e+01  5.779e-01  62.134  < 2e-16 ***
reserved_room_typeF         6.808e+01  8.205e-01  82.976  < 2e-16 ***
reserved_room_typeG         8.143e+01  9.816e-01  82.952  < 2e-16 ***
reserved_room_typeH         9.869e+01  1.792e+00  55.067  < 2e-16 ***
reserved_room_typeL         2.193e+01  1.734e+01   1.265  0.20591
reserved_room_typeP        -1.092e+02  1.169e+01  -9.338  < 2e-16 ***
lead_time                  -8.719e-02  1.364e-03 -63.915  < 2e-16 ***
customer_type2             -4.263e+00  3.341e-01 -12.759  < 2e-16 ***
customer_type3             -1.268e+01  1.852e+00  -6.848 7.51e-12 ***
customer_type4             -9.225e+00  7.333e-01 -12.580  < 2e-16 ***
arrival_date_monthAugust    4.728e+01  5.685e-01  83.158  < 2e-16 ***
arrival_date_monthDecember -2.263e+00  7.016e-01  -3.226  0.00126 **
arrival_date_monthFebruary -2.758e+01  6.383e-01 -43.208  < 2e-16 ***
arrival_date_monthJanuary  -3.188e+01  7.046e-01 -45.242  < 2e-16 ***
arrival_date_monthJuly      3.314e+01  5.770e-01  57.442  < 2e-16 ***
arrival_date_monthJune      1.786e+01  5.875e-01  30.405  < 2e-16 ***
arrival_date_monthMarch    -1.987e+01  6.022e-01 -32.996  < 2e-16 ***
arrival_date_monthMay       9.720e+00  5.750e-01  16.905  < 2e-16 ***
arrival_date_monthNovember -8.240e+00  6.954e-01 -11.849  < 2e-16 ***
arrival_date_monthOctober   1.002e+01  6.238e-01  16.058  < 2e-16 ***
arrival_date_monthSeptember 2.790e+01  6.373e-01  43.779  < 2e-16 ***
arrival_date_year2016       1.907e+01  3.966e-01  48.082  < 2e-16 ***
arrival_date_year2017       3.518e+01  4.724e-01  74.466  < 2e-16 ***
reservation_status2         6.951e+00  2.841e-01  24.463  < 2e-16 ***
reservation_status3         3.926e-01  1.267e+00   0.310  0.75663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.75 on 95481 degrees of freedom
Multiple R-squared:  0.8843,    Adjusted R-squared:  0.8843
F-statistic: 2.354e+04 on 31 and 95481 DF,  p-value: < 2.2e-16
```

- Linear Regression on Training Set-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.
- 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.

SMU

Part 2:Predictive Analytics

- After that we predict the test results and R squared on Testing sets and the results are as follows:
- SSE:SST stands for the sum of squared differences between the observed dependent variable and its mean
- SST:SSE term is the sum of squares error, or SSE. The error is the difference between the observed value and the predicted value.
- Rsquared is 1-SSE/SST =**0.4794058. Therefore Accuracy of the model is 47.94%.**

SMU

Part 2:Predictive Analytics

Create a model predicting cancellation (yes/no) based on various factors and testing the model. For this model predictor we followed various steps such as:

- Step 1: Remove unwanted Columns-We removed columns which were not necessary for creating a model predictor.After removing the columns these are the remaining variables in the dataset

```
> str(hotel_bookings)
'data.frame':   119390 obs. of  8 variables:
 $ hotel              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ is_canceled        : Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 2 2 ...
 $ lead_time          : int  342 737 7 13 14 14 0 9 85 75 ...
 $ adults             : int  2 2 1 1 2 2 2 2 2 2 ...
 $ children           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ reserved_room_type: Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
 $ adr                : num  0 0 75 75 98 ...
 $ total_stays        : int  0 0 1 1 2 2 2 2 3 3 ...
> 
```

- The above image describes the variables remaining in the dataset

SMU

Part 2:Predictive Analytics

- Step 2:Splitting Dataset into Training and Testing Set- Splitted the data set into training set and testing set with split ratio=0.8 and carried out feature scaling on both data sets.
- Step 3: Fitting Logistic Regression to the Test set-We used the glm function, where glm stands for general linear model.
- Step 4: Removing non significant column-After performing Logistic Regression we removed the non significant column and ran the regression again and got these result.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.854644   0.015641 -54.640  < 2e-16 ***
hotel1           0.525149   0.016822  31.218  < 2e-16 ***
reserved_room_typeB -0.425521   0.075301  -5.651 1.60e-08 ***
reserved_room_typeC -0.007960   0.083717  -0.095   0.9242
reserved_room_typeD -0.280612   0.021075 -13.315  < 2e-16 ***
reserved_room_typeE -0.236828   0.034773  -6.811 9.71e-12 ***
reserved_room_typeF -0.366949   0.050045  -7.332 2.26e-13 ***
reserved_room_typeG  0.006155   0.056548   0.109   0.9133
reserved_room_typeH  0.210827   0.101528   2.077   0.0378 *
reserved_room_typeL  0.646793   0.871984   0.742   0.4582
reserved_room_typeP 12.151173  39.514248   0.308   0.7585
lead_time            0.619654   0.007594  81.594  < 2e-16 ***
adults               0.060703   0.009026   6.725 1.75e-11 ***
adr.1                0.168605   0.008995  18.745  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125918  on 95511  degrees of freedom
Residual deviance: 115564  on 95498  degrees of freedom
AIC: 115592

Number of Fisher Scoring iterations: 9
```


SMU

Part 2:Predictive Analytics

- Step 5:Creating 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:

```
> #Making confusion Matrix
> cm = table(test_set[,2], y_pred)
> cm
     y_pred
         0     1
  Yes 13233  1800
  No   6143  2702
> 
```


SMU

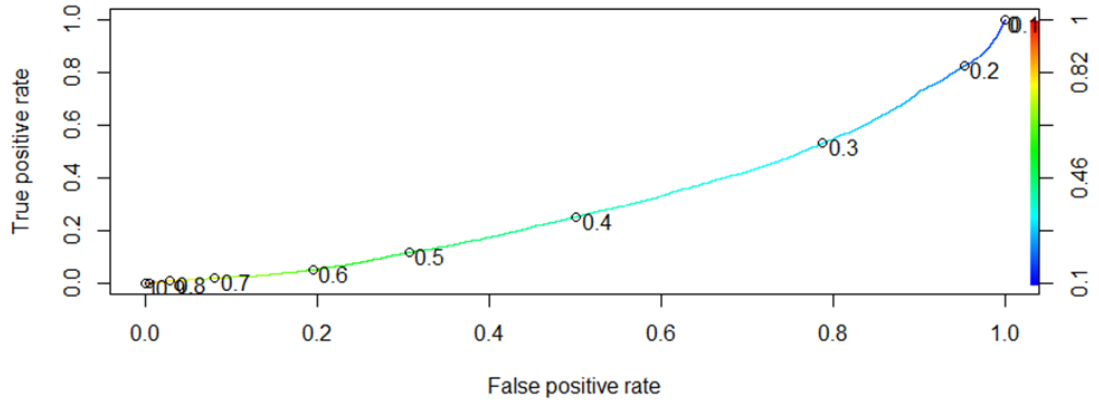
Part 2:Predictive Analytics

Step 6:Visualizing and Evaulating the Results-

- 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

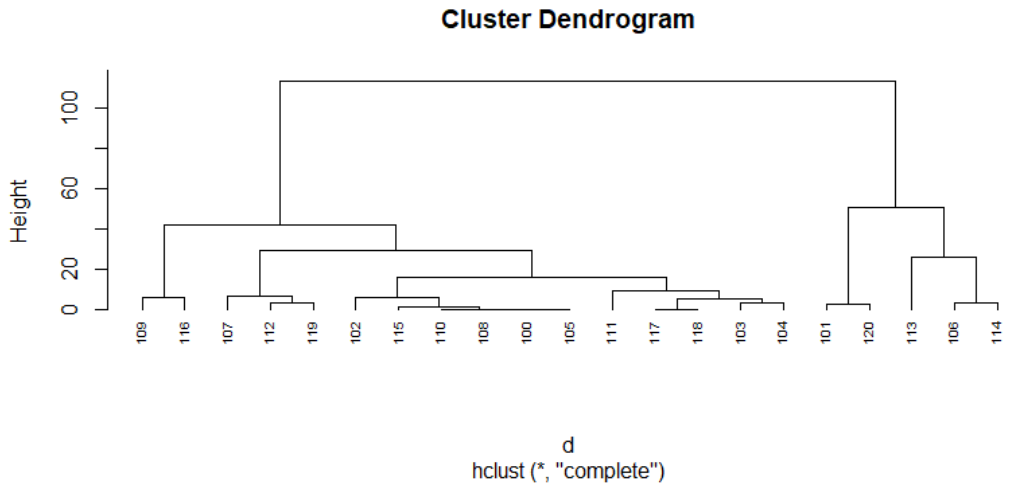
We can conclude that as the cut-off is low the accuracy of the model also decreases.

Roc Curve



Part 2:Predictive Analytics

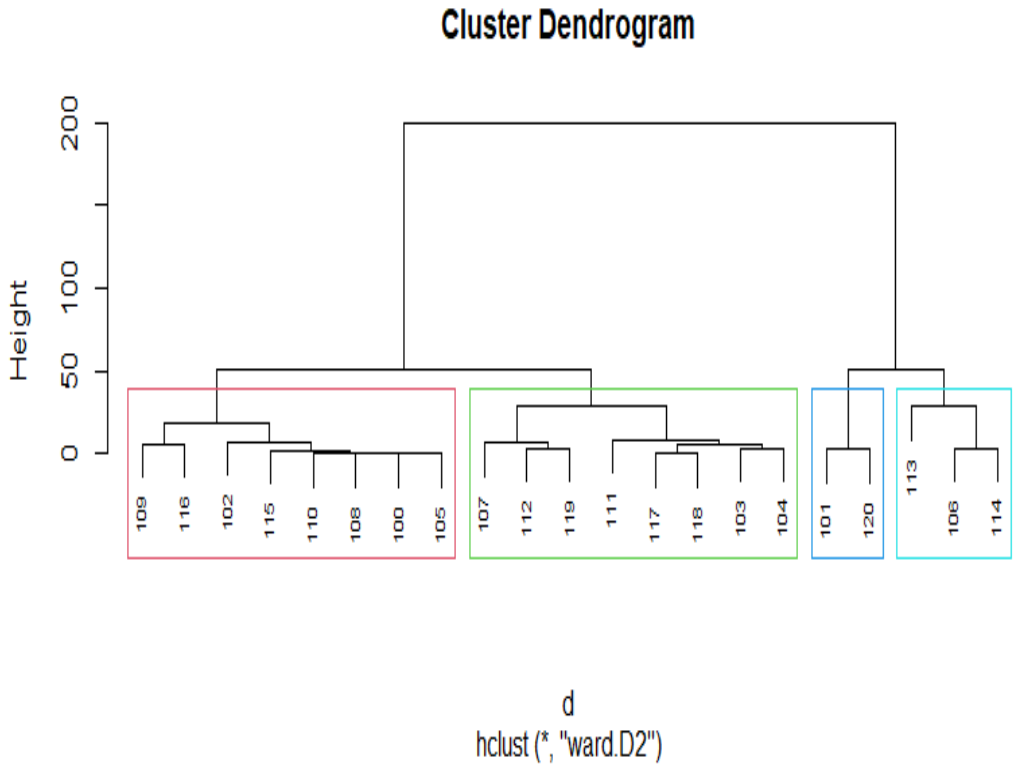
Clustering The hotel stay with respect to Lead time



- We used variable lead time and hotel stay to calculate the clusters
- Performed Agglomerative Hierarchical Clustering with “Complete” and “ward” method
- The given dendogram is the obtained from complete method.

Part 2:Predictive Analytics

Clustering The hotel stay with respect to Lead time

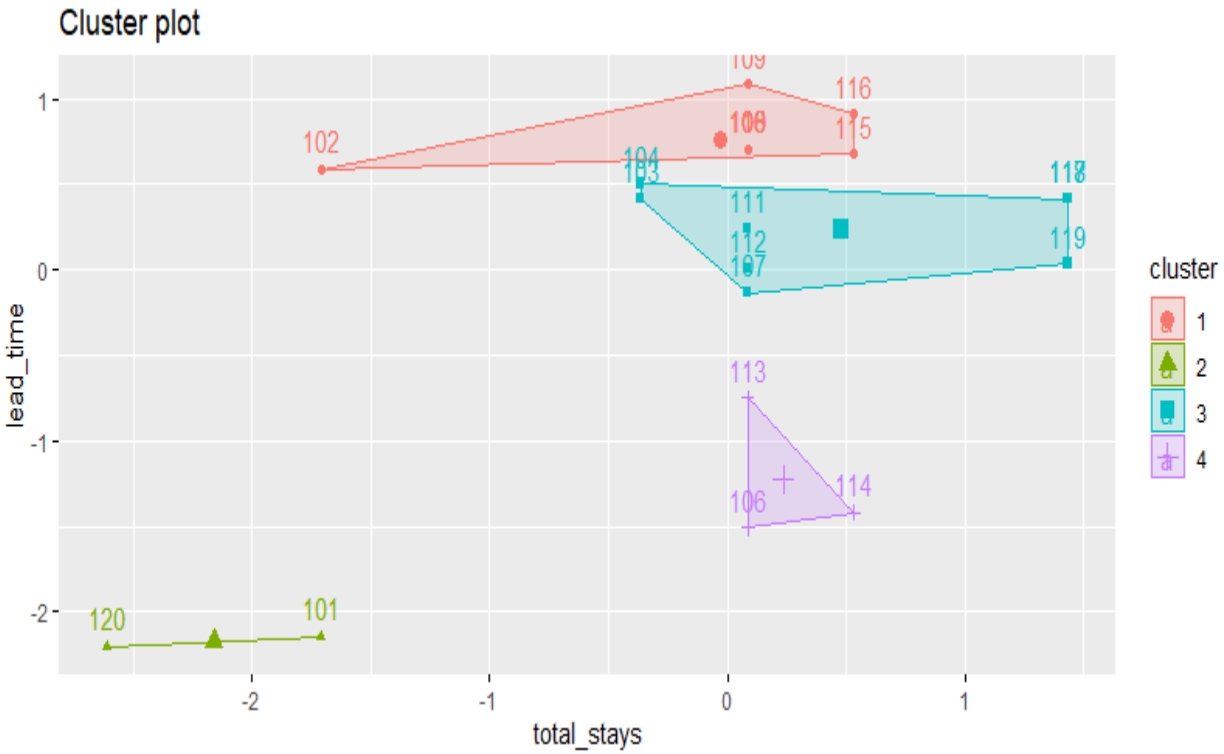


- This dendogram was obtained by ward.D2 method
- We divided the data and clustered it into 4 parts
- With help of clusters the hotel can look up the lead time of the customer as well as how many days are they staying

SMU

Part 2:Predictive Analytics

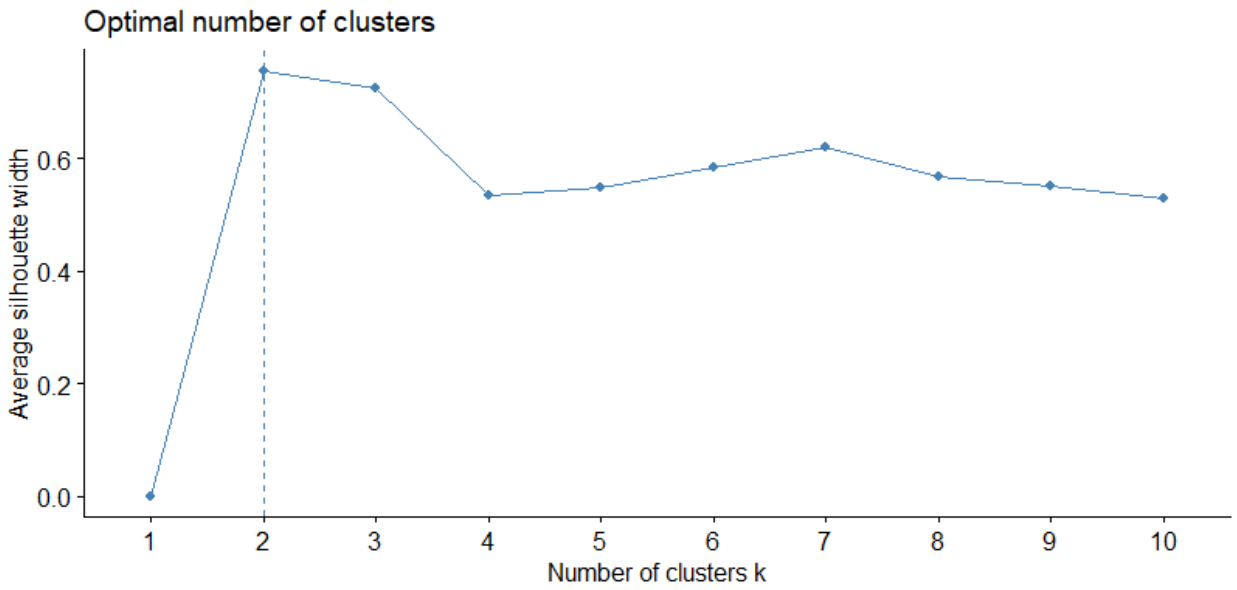
Clustering The hotel stay with respect to Lead time



- Here we could see the cluster with respect to the lead time
- The cluster is divided into 4 parts Red, Green, Purple and Blue

Part 2:Predictive Analytics

Clustering The hotel stay with respect to Lead time



- With the help of silhouette method we calculated the optimal number of clusters
- From the graph we could see the optimal numbers of clusters should be 2
- Therefore K=2

SMU

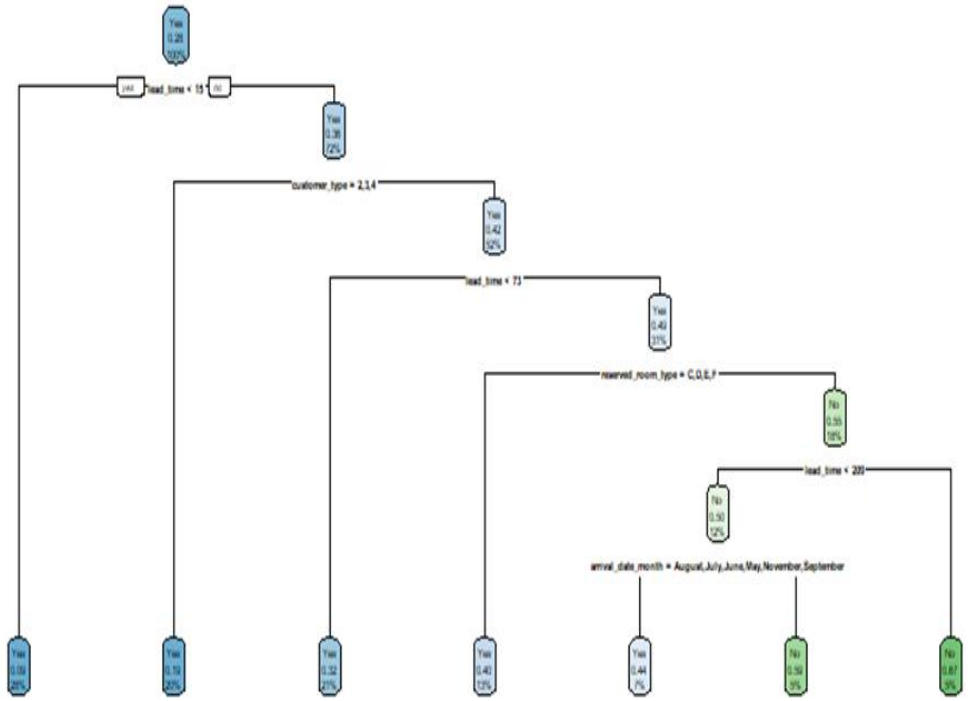
Graduate Exceptional Work (EMIS 7357 only)

Create and evaluate two different models predicting ADR

- Step 1: Separting data into Resort Hotel Booking and City Hotel Bookings
- Step 2: Splitting the data into train and test data
- Step3: First we use Resort Hotel dataset and apply linear Regression
- 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%.

Graduate Exceptional Work (EMIS 7357 only)

Checking Accuracy of Resort Model using CART



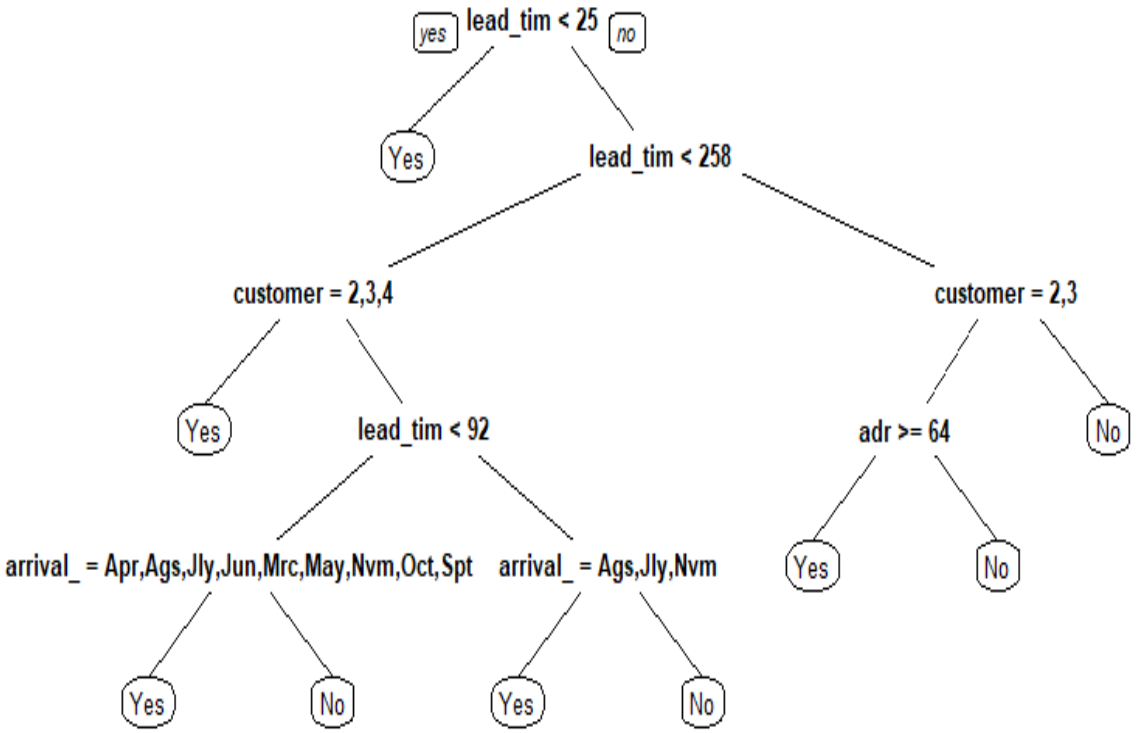
Accuracy of the resort model =69% i.e 0.6900

SMU

Graduate Exceptional Work (EMIS 7357 only)

Checking Accuracy of City Model using CART

Accuracy of the City model =70% i.e 0.7015905



Graduate Exceptional Work (EMIS 7357 only)

Predicting Overbooking Threshold.

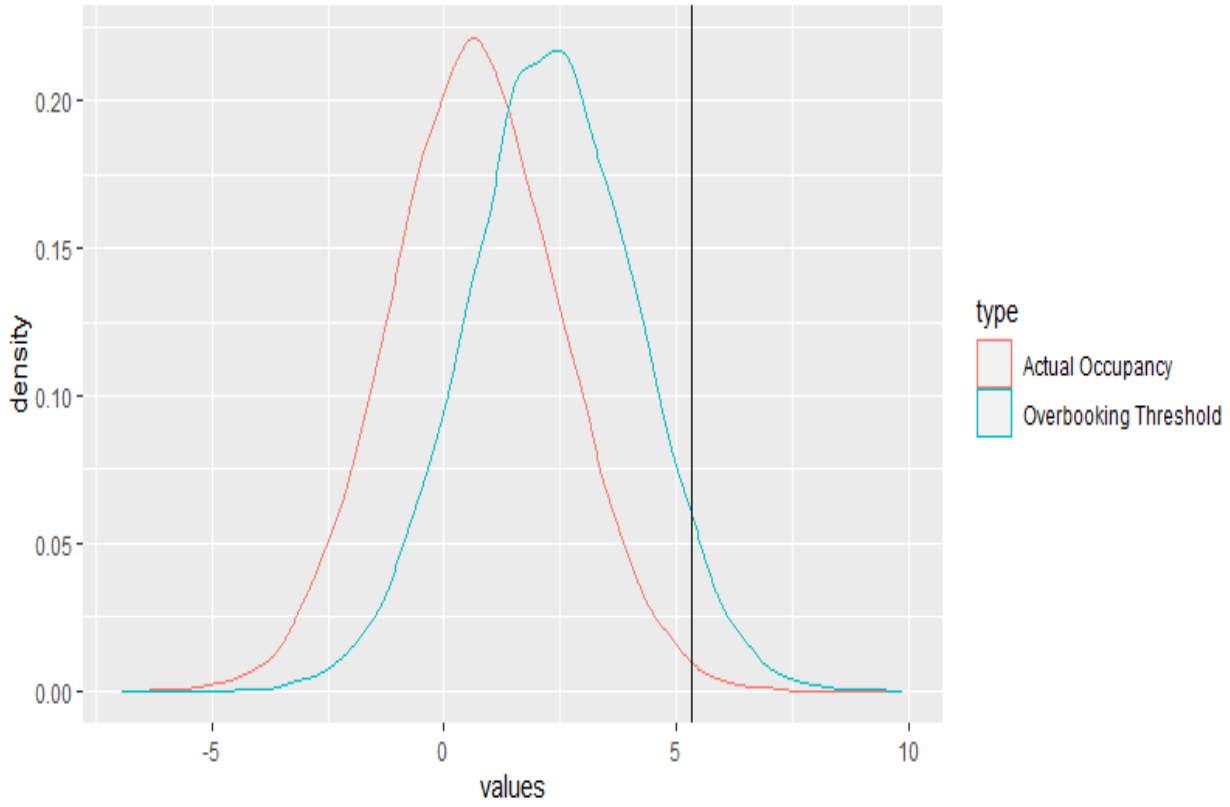
- Step1: To predict overbooking Threshold we firstly couple all the bookings by their year,month and day.
- Step2: Then calculate Hotel capacity which is calculated by all the assigned rooms in the hotel addition with all the cancellations.
- Step3: Calculate Predicted occupancy , overbooking threshold and actual occupancy
- Step4: Calculate mean and sd of all the occupancy

```
> print(mean(hotel_bookings_threshold$overbooking_threshold))
[1] 2.309766
> print(sd(hotel_bookings_threshold$overbooking_threshold))
[1] 1.844781
> print(mean(hotel_bookings_threshold$actual_occupancy))
[1] 0.6823502
> print(sd(hotel_bookings_threshold$actual_occupancy))
[1] 1.824139
```

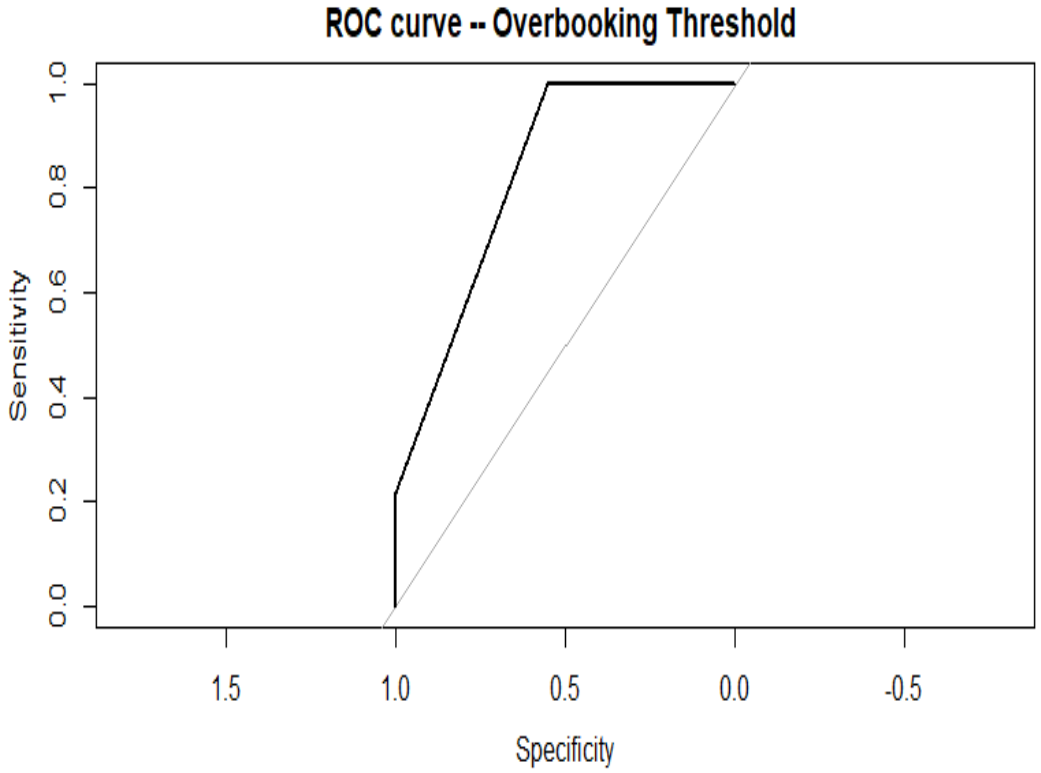

SMU

Graduate Exceptional Work (EMIS 7357 only)

Actual Occupancy V Overbooking



ROC Curve- Overbooking Threshold



Threshold value nearest to 1, i.e 1.5

SMU

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

SMU