Table of Contents

INTRODUCTION

Cancellation of hotel bookings has a hugely significant impact on revenue and decision-making in the hotel industry. Furthermore, an accurate forecast is needed to meditate on the impact of cancellation on revenue generation. To that end, most hotels have adopted overbooking and stringent cancellation policy strategies. These stringent measures have a negative impact on revenue and customer satisfaction, leaving bad online reviews.

The management of ESTANA Hotels in Portugal has a matter of importance have taken a strategic decision to leverage machine learning capabilities to forecast and predict cancellations. Therefore, the study aims to forecast and predict cancellation, increase revenue. To achieve this, the study will leverage descriptive analysis to uncover key factors that affect cancellations and utilize visualization to provide insights into the ESTANA dataset. Additionally, predictive models will be developed to predict cancellations accurately. Time series forecasting will be employed to forecast guest arrivals, and a regression model will be used to determine factors influencing cancellation through R Studio. Based on the above, the study seeks to achieve the following objectives.

BUSINESS CASE

ESTANA Hotel Group is one of the largest conglomerates with a chain of hotels in Lisbon, Portugal, consisting of a Resort and a City Hotel. ESTANA started operations in 1990 and has a presence in Europe and Asia, catering for over 2 million guests annually with an annual revenue of £1.2b. The hotel chain has a strong competitive position with a market share of 10% in the hospitality industry. ESTANA's robust networks and competitive pricing strengthen its competitive position and dominance.

One significant issue that ESTANA faces is booking cancellations. The rapid evolution of technology in the online booking sector has led to changes in consumer behaviour resulting in last-minute cancellations by guests, lost revenue, and decreased customer satisfaction. The management of ESTANA believes that accurate cancellation forecasts will support the management in the decision-making process and reduce the impact of cancellation on its revenue. Hence, this study utilizes machine learning models to forecast and predict cancellation. Furthermore, to demonstrate how analytics will help ESTANA Hotel Group predict cancellation and revenue management.

STUDY OBJECTIVE

To develop and find the best machine learning model that predicts hotel booking cancellation.

To forecast the number of guest bookings

To provide insights analysis of features that cause cancellations.

To determine the variables that impact cancellation.

In achieving the study's objectives, the study will utilize R Studio and the following analytical approach to enhance cancellation prediction and forecasting capabilities.

LITERATURE REVIEW

Studies on hotel booking cancellation prediction and forecasting could be more developed than those on airline cancellations (Antonio et al., 2019). However, there is an increasing interest in predicting hotel booking cancellations, and this study aims to explore the alignment between the study's objectives and past research on hotel cancellation.

Morales and Wang (2010) predicted the hotel chain cancellation rate using logistic regression, decision trees, and support vector machines. The authors' analysis used variables such as seasonality and weather in its classification and regression model. The findings indicated that support vector machines (SVM) achieved the highest performance, with accuracy rates ranging from 68.4% to 83.9% as the check-in date approached.

Recent research, including those conducted by Antonio et al. (2019) and Wu et al. (2017), has utilized artificial intelligence (AI) and guest data to predict individual cancellations in the hospitality sector. The authors achieved high accuracy and satisfactory performance in the AI-based methods. Moreso, studies by Huang et al., 2013 used artificial neural networks to forecast and classify restaurant bookings. The author's models achieved a classification accuracy of around 88%, while Tse & Poon (2017) study employed maximum-likelihood estimation to predict no-show restaurant bookings.

Furthermore, In the hotel industry, Antonio et al. (2017) investigated nine alternative classification algorithms using guest hotel booking data and achieved an accuracy rate of 90% with the XGBoost algorithm. Similarly, Falk and Vieru (2018) in Finland employed a probit model in predicting hotel booking and achieved an accuracy of 84%. Additionally, Sanchez et al. (2020) employed an artificial neural network (ANN) and Support vector machine (SVM) to predict hotel chain cancellation in Spain. The study achieved an accuracy of 71% in a seven-day lead time.

To that end, studies on factors that lead to hotel cancellations have provided valuable insights into customer satisfaction and revenue management. Hernandez-Mendez et al. (2018), for instance, argued that reservation type, length of stay and booking channel could impact hotel cancellation. Customer reviews and ratings could also result in high cancellation rates (Li et al., 2019). Also, the Competitor's revised pricing after booking online can affect guest cancellation decisions (Chen et al., 2011). Lead time between booking and check-in increases guests' cancellation rate (Chew & Jahari, 2014). According to Antonio et al. (2017), guest Culture and travel-related factors can influence hotel cancellation rates. Likewise, seasonality and room availability can impact hotel cancellation (Bigne et al., 2021)

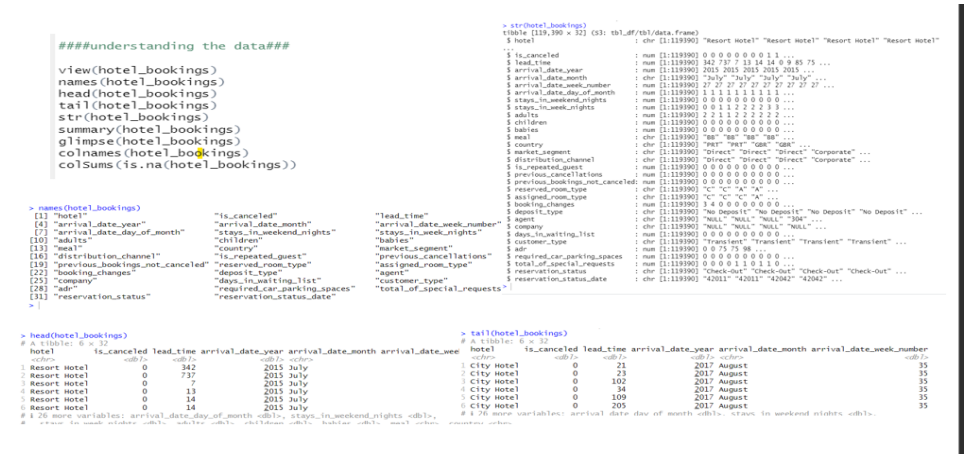
INTRODUCTION TO DATASET

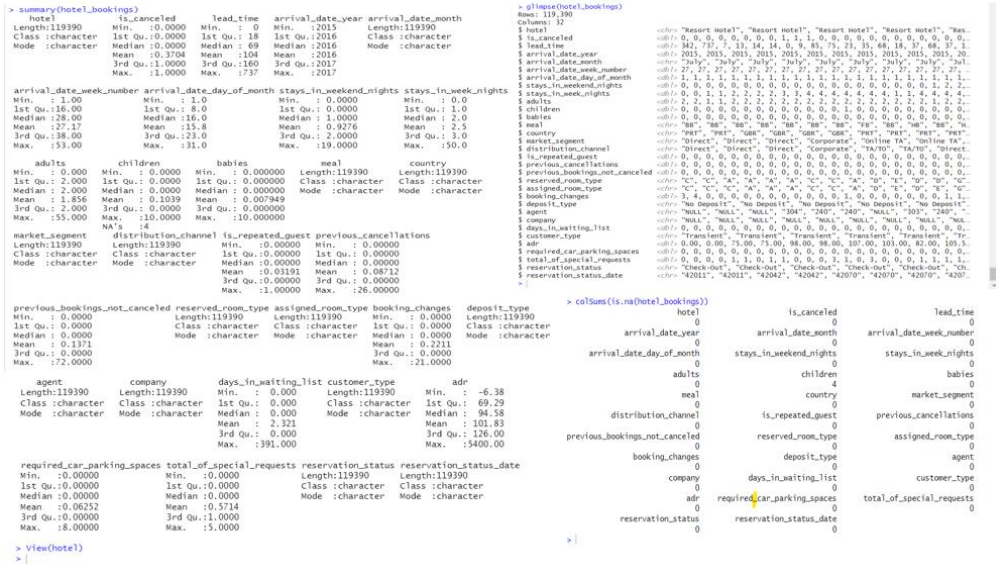
The dataset is originally from hotel bookings for a Resort and City Hotel in Portugal. It was first collected by Antonio et, al (2019) and it is available for downloads at kaggle.com/jessemostipak/hotel_bookings

UNDERSTANDING THE DATASET

The dataset has 119,210 observations and 32 features and it includes key features such as IS Cancelled, Lead time, Previous cancellation, Deposit type etc. The dependent variable is IS-CANCELLED.

understanding the variable, the dataset (view), (head), and str command to view type of variables. The data is made up of both categorical & numeric variables. Summary(hotel)command gave statistical summaries of each feature. The colsum(is.na) command was used to check for missing variables and the data has 4 missing observations in Feature called children.





DATASET KEY VARIABLE DESCRIPTION

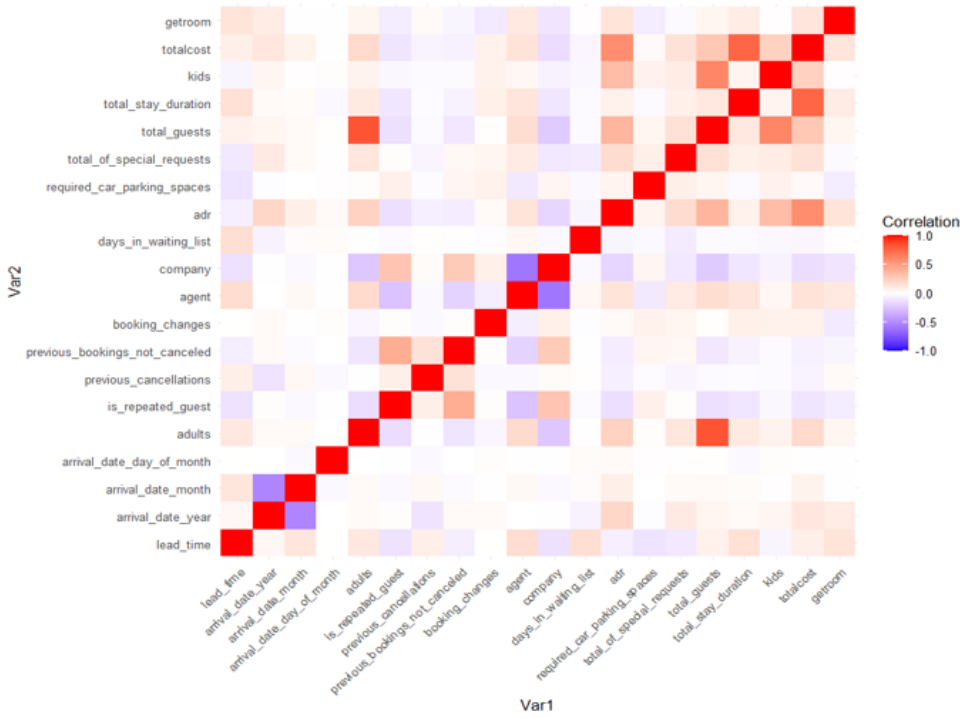
VARIABLE DESCRIPTION

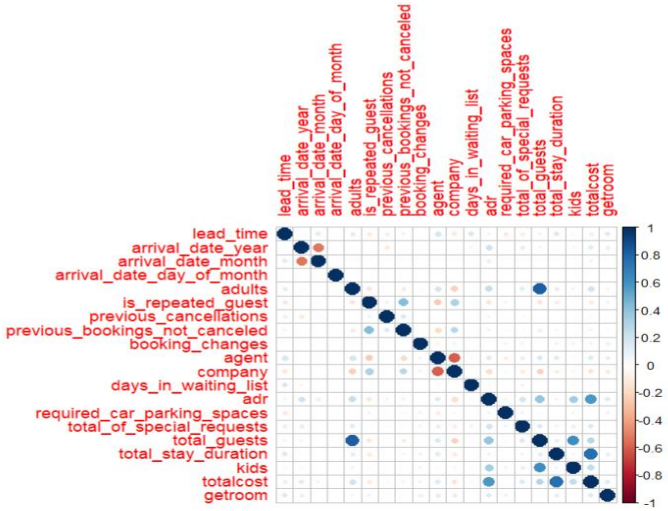
Variable	Values	Type	Description
Hotel	0,1	Numeric	(H1 = Resort Hotel or H2 = City Hotel)
is_canceled (Dependent Variable)	0,1	Numeric	Value indicating if the booking was canceled (1) or not (0)
lead_time	1,2,3,4,5,…...	Numeric	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	2015-2017	Character	Year of arrival date
arrival_date_month	Jan -Dec	Character	Month of arrival date
arrival_date_week_number	Week 1 –Week 54	Character	Week number of year for arrival date
arrival_date_day_of_month	Monday to Tuesday	Numeric	Day of arrival date

Variable	Values	Type	Description
stays_in_weekend_nights	0,1,2,3,4,5.....	Numeric	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	0,1,2,3,4,5.....	Numeric	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	1,2,3,4,5,…...	Numeric	Number of Adults
meal	Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)	Character	Type of meal booked
country	represented in the ISO 3155–3:2013 format	Character	Country of origin
distribution_channel	"TA" means "Travel Agents" and "TO" means "Tour Operators"	Character	Booking distribution channel
is_repeated_guest	0,1	Numeric	Value indicating if the booking name was from a repeated guest (1) or not (0)

Variable	Values	Type	Description
previous_cancellations	0,1,2,3,4,5.....	Numeric	Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled	0,1,2,3,4,5.....	Numeric	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	A,B,C,D,E.....L	Character	Code of room type reserved. Code is presented instead of designation for anonymity reasons.
booking_changes	1,2,3,4,5,….....	Numeric	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	No Deposit, Non-Refund, Refundable	Character	Indication on if the customer made a deposit to guarantee the booking.
days_in_waiting_list	0,1,2,3,4,5,…..	Numeric	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	0,1	Character	Type of booking, assuming one of four categories: Contract, Group, Transient and Transient party

CORRELATION PLOTS

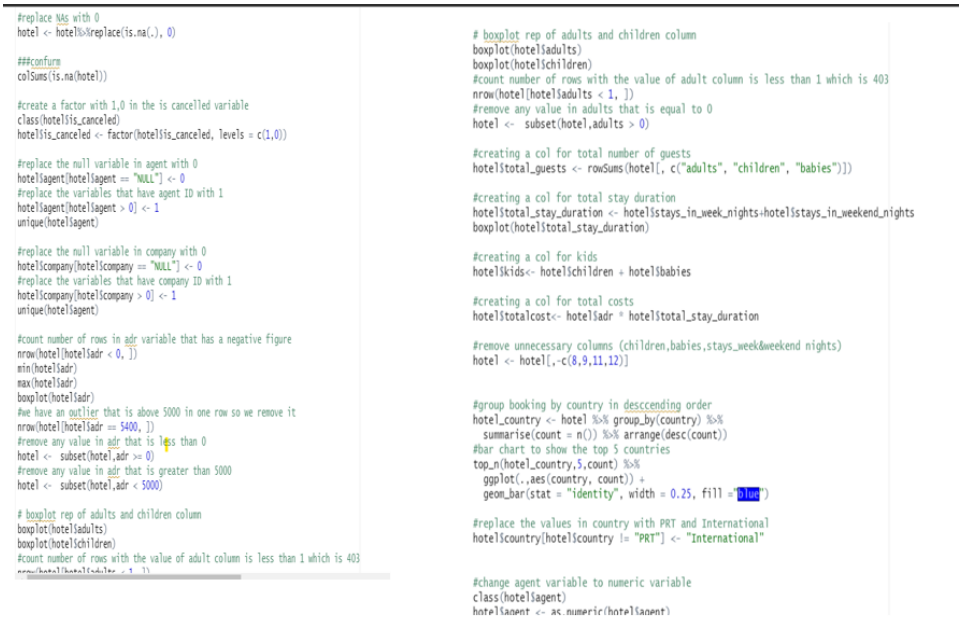




The above shows the heatmap correlation of the Variables. Variables such as total guest and adults are correlated.

DATA CLEANING, PREPARATION AND MANIPULATION

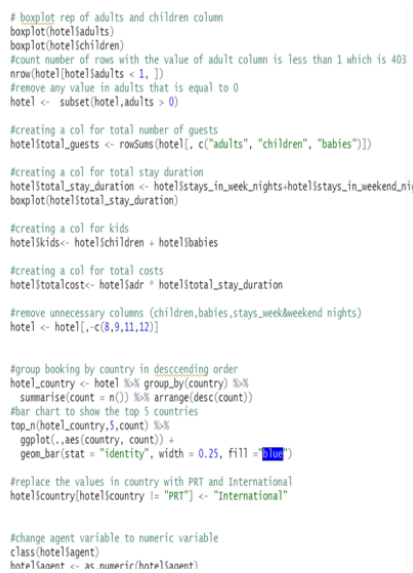
The missing variables were replaced with 0. The dependent variable was assigned with 1, 0. The null variables in agents and company were replaced with 0. An outlier was removed in one row of ADR. New columns for the total number of guests, total costs, and kids were created. Unwanted columns such as babies and children were removed. The group y function was to group country and agent, company were changed to numeric variable. The variable month was changed to integer using the unite() function. The select () was used to isolate numeric features. Dummy variables were created using the fast dummies () to change character to numeric variables.



```
#replace NAs with 0
hotel <- hotel%>%replace(is.na(.), 0)

###confurm
colSums(is.na(hotel))

#create a factor with 1,0 in the is cancelled variable
class(hotel$is_canceled)
hotel$is_canceled <- factor(hotel$is_canceled, levels = c(1,0))

#replace the null variable in agent with 0
hotel$agent[hotel$agent == "NULL"] <- 0
#replace the variables that have agent ID with 1
hotel$agent[hotel$agent > 0] <- 1
unique(hotel$agent)

#replace the null variable in company with 0
hotel$company[hotel$company == "NULL"] <- 0
#replace the variables that have company ID with 1
hotel$company[hotel$company > 0] <- 1
unique(hotel$agent)

#count number of rows in adr variable that has a negative figure
nrow(hotel[hotel$adr < 0, ])
min(hotel$adr)
max(hotel$adr)
boxplot(hotel$adr)
#we have an outlier that is above 5000 in one row so we remove it
nrow(hotel[hotel$adr == 5400, ])
#remove any value in adr that is less than 0
hotel <-  subset(hotel,adr >= 0)
#remove any value in adr that is greater than 5000
hotel <-  subset(hotel,adr < 5000)

# boxplot rep of adults and children column
boxplot(hotel$adults)
boxplot(hotel$children)
#count number of rows with the value of adult column is less than 1 which is 403
```

```
# boxplot rep of adults and children column
boxplot(hotel$adults)
boxplot(hotel$children)
#count number of rows with the value of adult column is less than 1 which is 403
nrow(hotel[hotel$adults < 1, ])
#remove any value in adults that is equal to 0
hotel <-  subset(hotel,adults > 0)

#creating a col for total number of guests
hotel$total_guests <- rowSums(hotel[, c("adults", "children", "babies")])

#creating a col for total stay duration
hotel$total_stay_duration <- hotel$stays_in_week_nights+hotel$stays_in_weekend_nights
boxplot(hotel$total_stay_duration)

#creating a col for kids
hotel$kids<- hotel$children + hotel$babies

#creating a col for total costs
hotel$totalcost<- hotel$adr * hotel$total_stay_duration

#remove unnecessary columns (children,babies,stays_week&weekend nights)
hotel <- hotel[,-c(8,9,11,12)]


#group booking by country in desccending order
hotel_country <- hotel %>% group_by(country) %>%
  summarise(count = n()) %>% arrange(desc(count))
#bar chart to show the top 5 countries
top_n(hotel_country,5,count) %>%
  ggplot(.,aes(country, count)) +
  geom_bar(stat = "identity", width = 0.25, fill ="blue")

#replace the values in country with PRT and International
hotel$country[hotel$country != "PRT"] <- "International"


#change agent variable to numeric variable
class(hotel$agent)
hotel$agent <- as.numeric(hotel$agent)
```

```
#replace NAs with 0
hotel <- hotel%>%replace(is.na(.), 0)

###confurm
colSums(is.na(hotel))

#create a factor with 1,0 in the is cancelled variable
class(hotel$is_canceled)
hotel$is_canceled <- factor(hotel$is_canceled, levels = c(1,0))

#replace the null variable in agent with 0
hotel$agent[hotel$agent == "NULL"] <- 0
#replace the variables that have agent ID with 1
hotel$agent[hotel$agent > 0] <- 1
unique(hotel$agent)

#replace the null variable in company with 0
hotel$company[hotel$company == "NULL"] <- 0
#replace the variables that have company ID with 1
hotel$company[hotel$company > 0] <- 1
unique(hotel$agent)

#count number of rows in adr variable that has a negative figure
nrow(hotel[hotel$adr < 0, ])
min(hotel$adr)
max(hotel$adr)
boxplot(hotel$adr)
#we have an outlier that is above 5000 in one row so we remove it
nrow(hotel[hotel$adr == 5400, ])
#remove any value in adr that is less than 0
hotel <-  subset(hotel,adr >= 0)
#remove any value in adr that is greater than 5000
hotel <-  subset(hotel,adr < 5000)

# boxplot rep of adults and children column
boxplot(hotel$adults)
boxplot(hotel$children)
#count number of rows with the value of adult column is less than 1 which is 403
```



```
# boxplot rep of adults and children column
boxplot(hotel$adults)
boxplot(hotel$children)
#count number of rows with the value of adult column is less than 1 which is 403
nrow(hotel[hotel$adults < 1, ])
#remove any value in adults that is equal to 0
hotel <-  subset(hotel,adults > 0)

#creating a col for total number of guests
hotel$total_guests <- rowSums(hotel[, c("adults", "children", "babies")])

#creating a col for total stay duration
hotel$total_stay_duration <- hotel$stays_in_week_nights+hotel$stays_in_weekend_nights
boxplot(hotel$total_stay_duration)

#creating a col for kids
hotel$kids<- hotel$children + hotel$babies

#creating a col for total costs
hotel$totalcost<- hotel$adr * hotel$total_stay_duration

#remove unnecessary columns (children,babies,stays_week&weekend nights)
hotel <- hotel[,-c(8,9,11,12)]


#group booking by country in desccending order
hotel_country <- hotel %>% group_by(country) %>%
  summarise(count = n()) %>% arrange(desc(count))
#bar chart to show the top 5 countries
top_n(hotel_country,5,count) %>%
  ggplot(.,aes(country, count)) +
  geom_bar(stat = "identity", width = 0.25, fill ="blue")

#replace the values in country with PRT and International
hotel$country[hotel$country != "PRT"] <- "International"


#change agent variable to numeric variable
class(hotel$agent)
hotel$agent <- as.numeric(hotel$agent)
```

DESCRIPTIVE ANALYSIS

```r
hotel %>%
  summarise(
    Portugal = round(mean(country == "PRT"), 3) * 100,
    Other_countries = round(mean(country != "PRT"), 3) * 100
  ) %>%
  pivot_longer(
    cols = c(Portugal, Other_countries),
    names_to = "Region",
    values_to = "value"
  ) %>%
  mutate(
    Region = str_replace_all(Region, "_", " "),
    lab.ypos = cumsum(value) - 0.5 * value
  ) %>%
  arrange(value) %>%
  ggplot(aes(
    x = 2,
    y = value,
    fill = Region,
    label = value
  )) +
  geom_bar(stat = "identity", color = "white") +
  geom_text(aes(y = lab.ypos), color = "white") +
  coord_polar(theta = "y", start = 0) +
  theme_void() +
  xlim(0.5, 2.5) +
  ggtitle("Booking Rate by Country of Origin") +
  scale_fill_hue(c = 50, l = 40)


library(ggplot2)
library(scales)

# Check the distribution of hotel type for cancellation
hotel_cancellation_table <- table(hotel$is_canceled, hotel$hotel)

# Visualize the cancellation by hotel type
ggplot(data = as.data.frame(hotel_cancellation_table),
       aes(x = Var2,
```

```r
library(ggplot2)

library(ggplot2)

# Convert arrival_date_month to a factor with ordered levels
hotel$arrival_date_month <- factor(hotel$arrival_date_month, levels = month

# Visualize Hotel traffic on Monthly basis
ggplot(data = hotel, aes(x = arrival_date_month)) +
  geom_bar(fill = "lightblue") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  coord_flip() +
  labs(title = "Month Wise Booking Request",
       x = "Month",
       y = "Count") +
  theme_classic()

###booking status by month

ggplot(hotel, aes(arrival_date_month, fill = factor(is_canceled))) +
  geom_bar() + geom_text(stat = "count", aes(label = ..count..), hjust = 1)
  coord_flip() + scale_fill_discrete(
    name = "Booking Status",
    breaks = c("0", "1"),
    label = c("Cancelled", "Not Cancelled")
  ) +
  labs(title = "Booking Status by Month",
       x = "Month",
       y = "Count") + theme_bw()
```

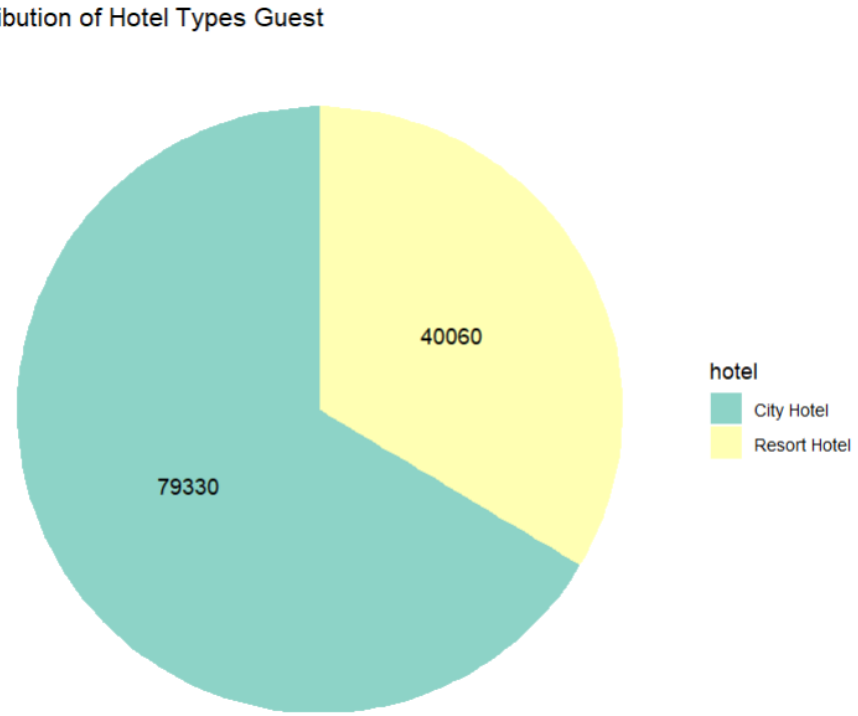


Figure 1 ESTANA HOTEL TYPE

The pie chart shows that ESTANA Hotels have two kinds of hotel. The Resort and the City hotel. The pie chart also shows the total number of Guest.

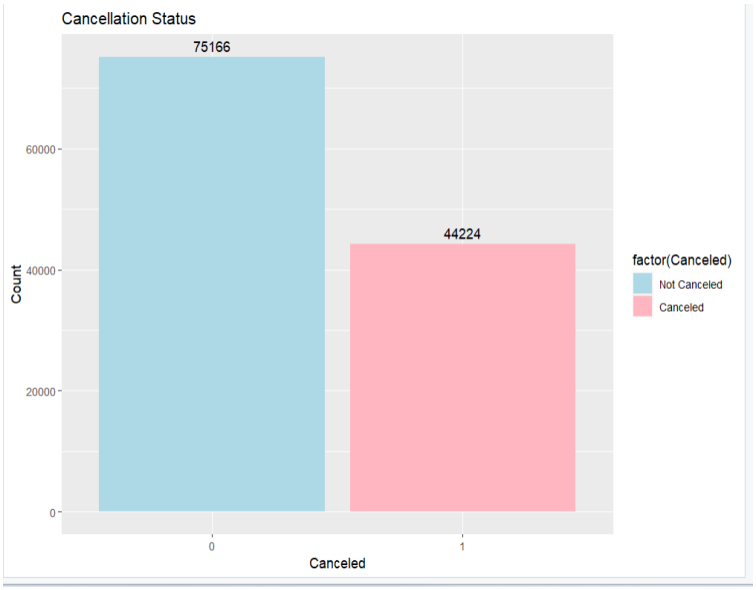


Figure 2 ESTANA CANCELLATION STATUS

The above chart represents the total number of bookings for both City and Resort ESTANA hotels. 44,224 bookings were cancelled, representing about 37% of the total bookings, while 75,166 bookings were not cancelled.

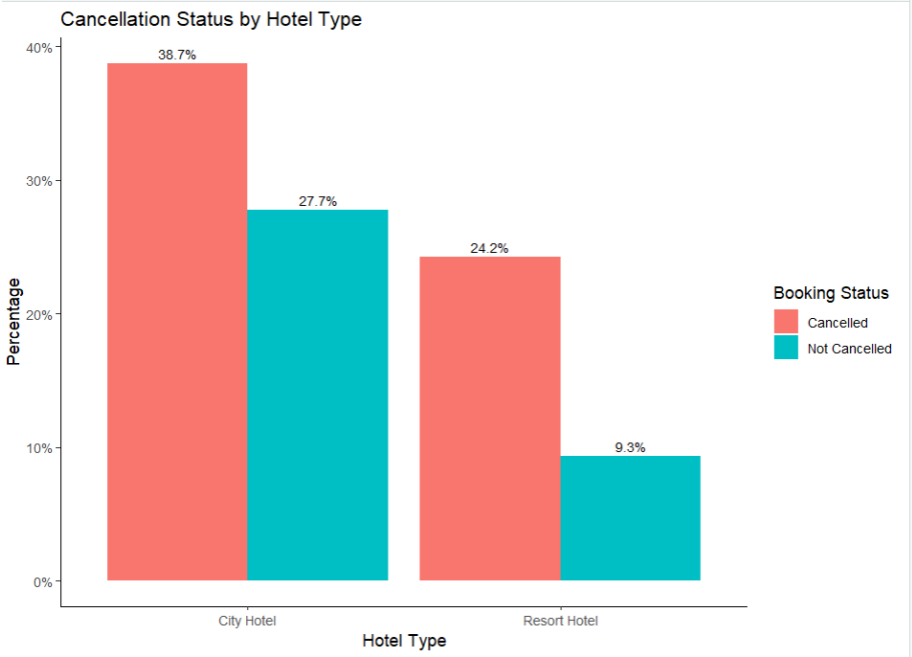


Figure 3 ESTANA CANCELLATION BY HOTEL

The bar chart reveals that 27.7% of bookings received at city hotels were not cancelled. Conversely, city hotels had 38.7% cancelled bookings compared to resorts, with 24.2% cancelled bookings.

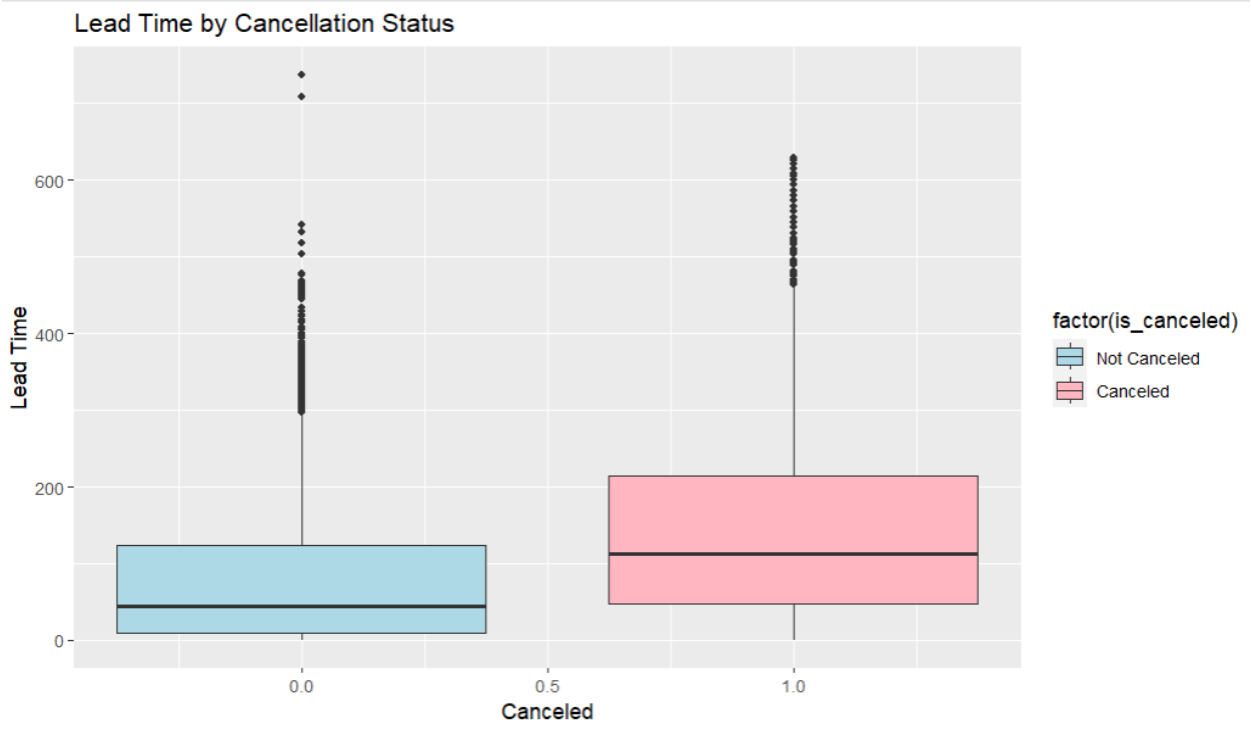


Figure 4 ESTANA LEAD TIME VS CANCELLED STATUS

Lead time affects cancellations. The higher the lead time, the higher the likelihood of cancellation. Lead time indicates the days between booking and date of arrival into ESTANA hotels.

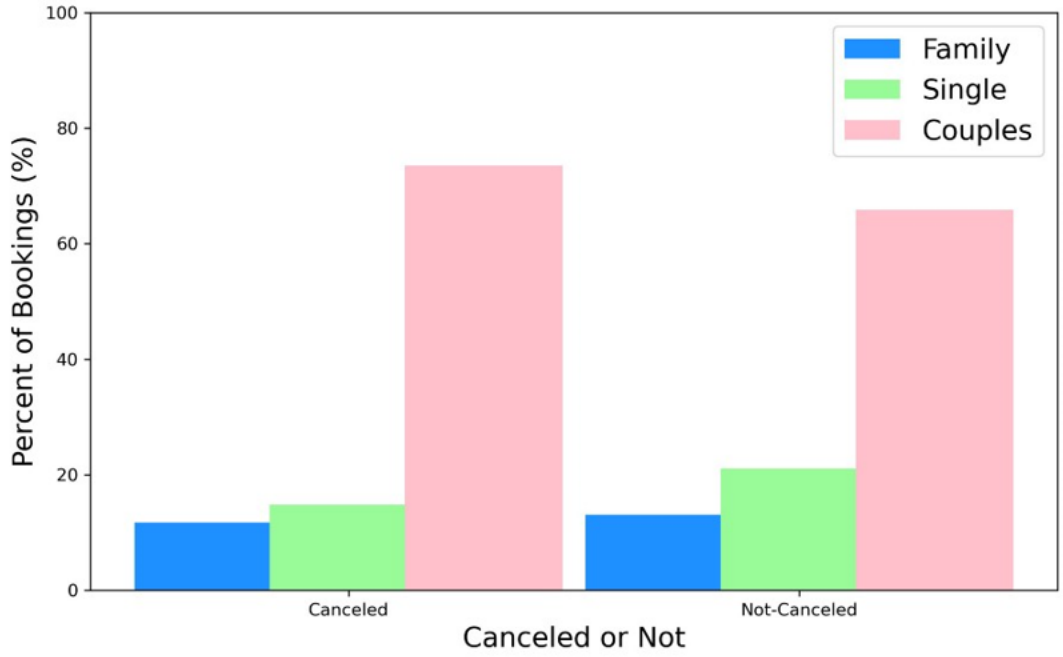


Figure 5 ESTANA CANCELLATION STATUS

From the chart, majority of ESTANA hotel Guest are couples. Moreso, the cancellation rate is higher among the couples followed by singles.

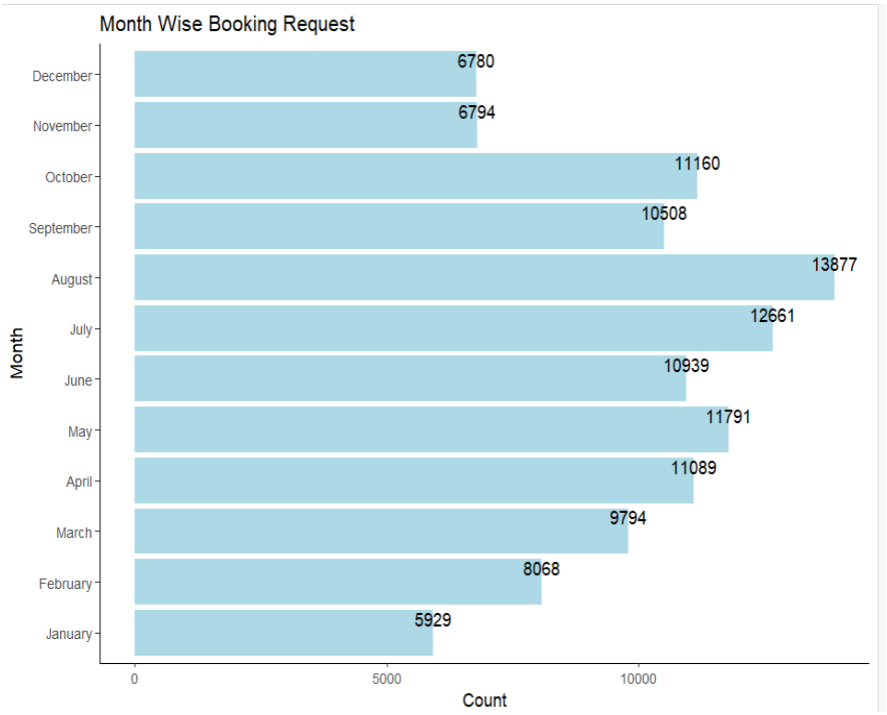


Figure 6 ESTANA BOOKINGS BY MONTHS

ESTANA received more bookings between April to August. While it received the lowest bookings between November to January. The highest booking was in August with 13,877 bookings. This could as a result of the summer period.

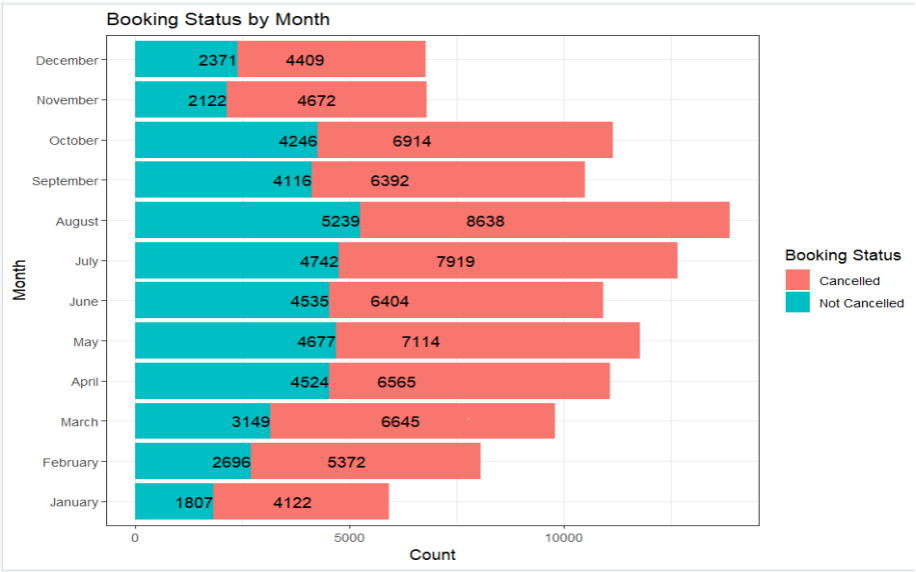


Figure 7 ESTANA CANCELLATION BOOKINGS BY MONTH

From the chart above, 8,638 bookings of the 13877 bookings received in August were cancelled. August had the highest cancelled bookings January had the lowest cancelled bookings.

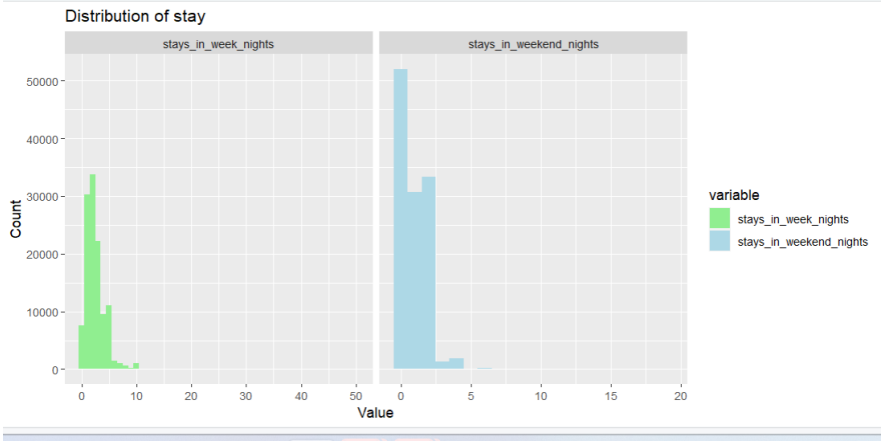


Figure 8 ESTANA BOOKINGS STAY DISTRIBUTION

The graph above shows that more bookings were received from guest who intend to stay through the weekend .

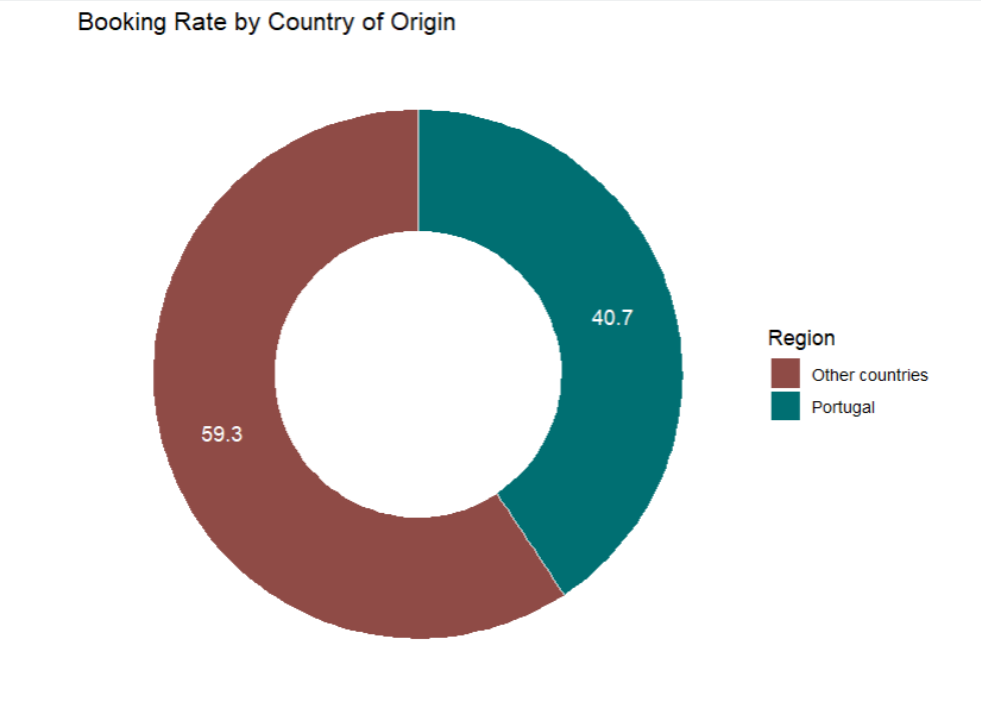


Figure 9 ESTANA COUNTRY BY BOOKING

The countries has been divided into two. Portugal and others (This include the rest of the world). However, ESTANA received 40.7% bookings from Portugal and 59.3% from others.

PREDICTIVE RESULT ANALYSIS

UNSUPERVISED LEARNING -CLUSTERING

CLUSTER PLOT

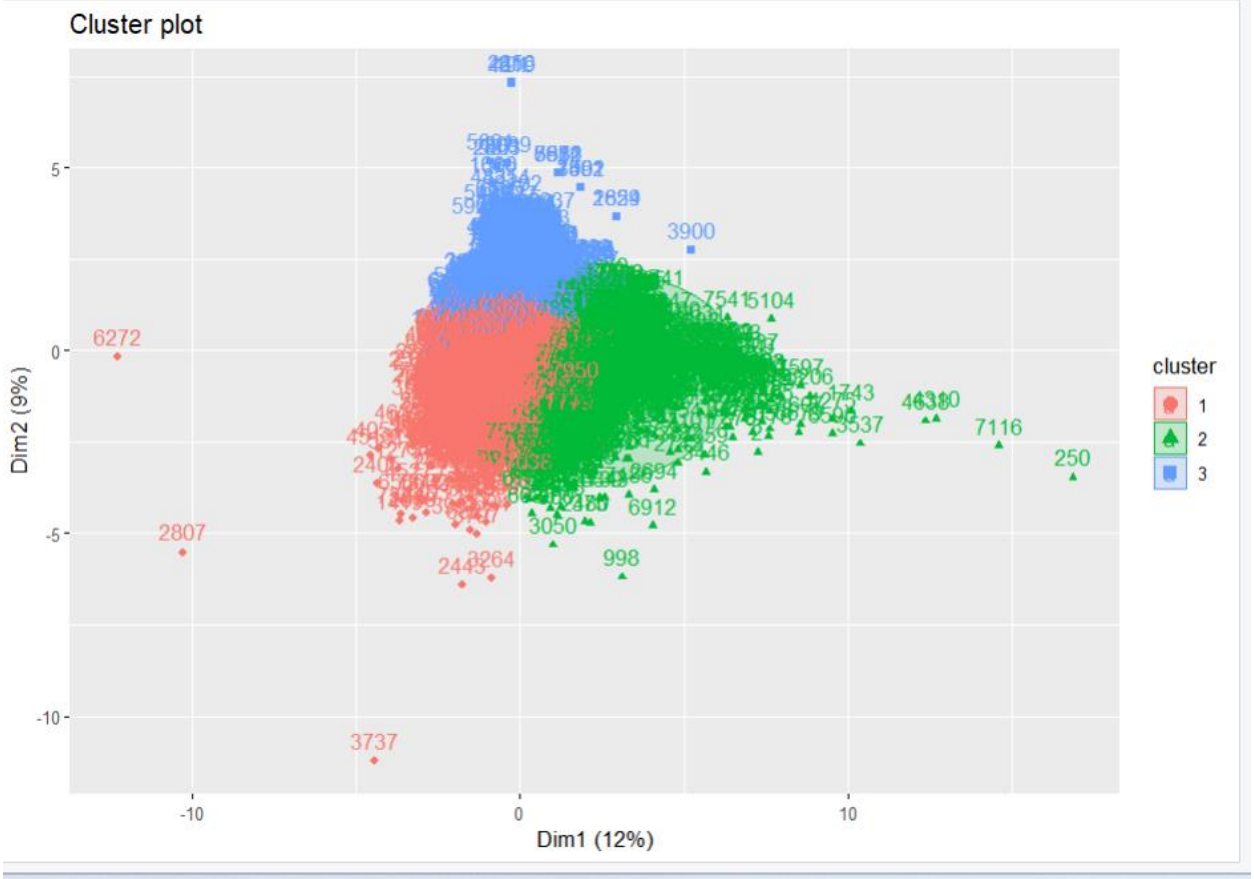


Figure 10 ESTANA COUNTRY BY BOOKING

The cluster plot with an assumed centroid of 3 shows the similarities among the datapoint. Although there are some overlaps, the hotel cluster points share some similarities. This is a benchmark.

OPTIMUM K USING ELBOW

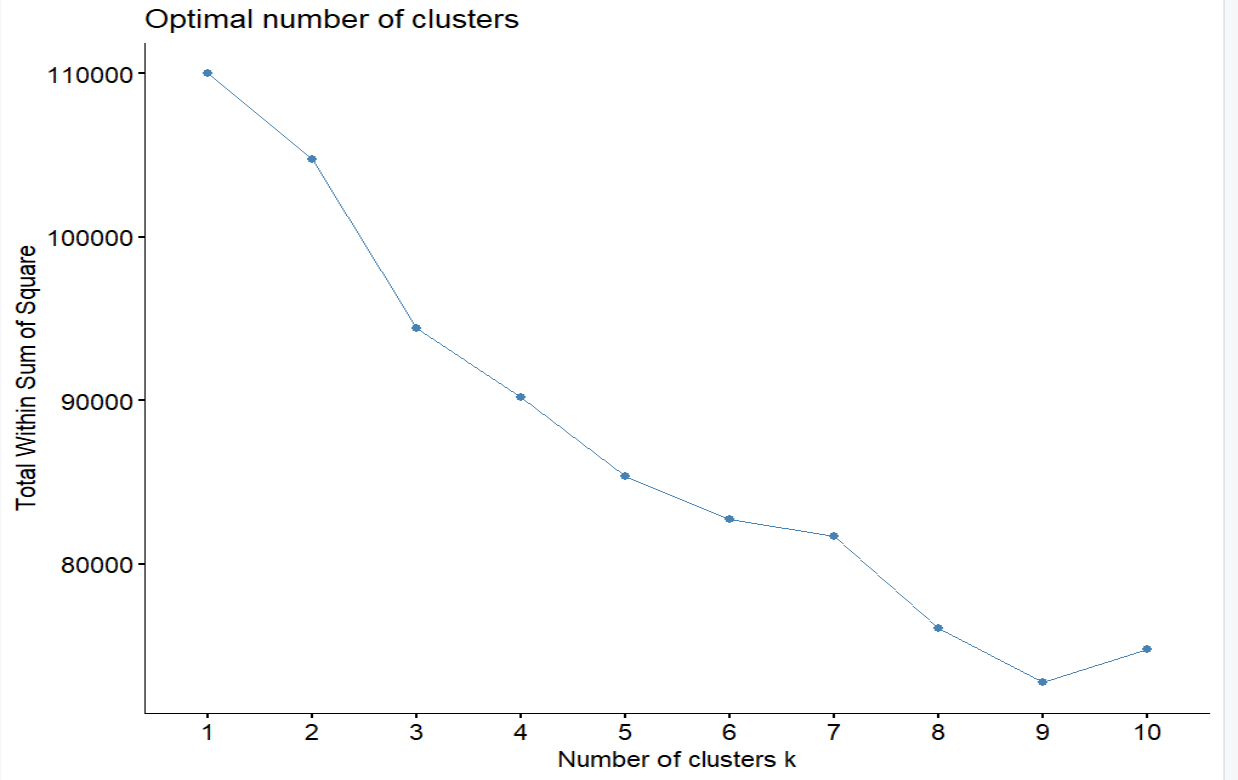


Figure 11 optimum cluster Elbow

From the graph, the bend like an elbow shows the optimum number clusters which is of K =3. As the sum of squares reduces, the value of K increases from 1-3 like an elbow. Also, beyond K=3, adding more clusters does not provide any significant improvement to the segmented clusters. Therefore, the optimum number of clusters is 3 based on the elbow method.

OPTIMUM NUMBER OF CLUSTER USING SILOUHETE

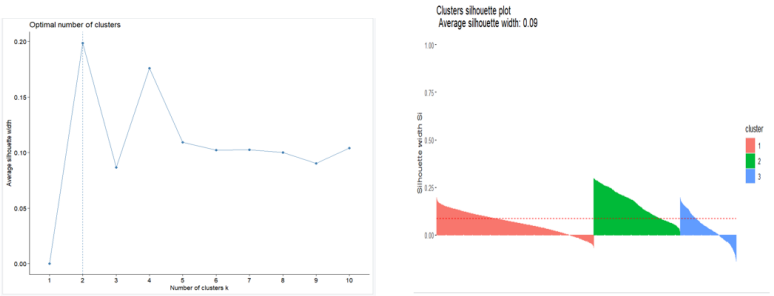
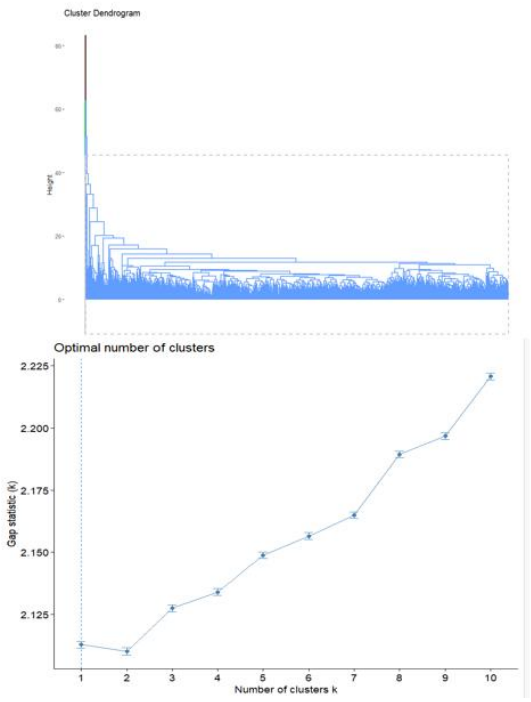


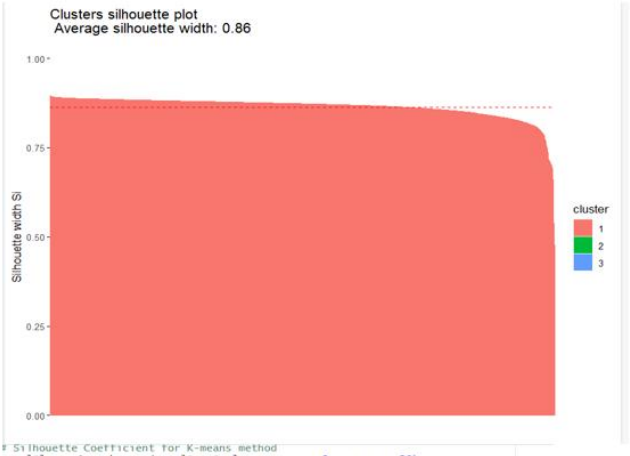
Figure 9 ESTANA COUNTRY BY BOOKING

The optimum number of clusters is 2 according to the average silhouette. The clustering algorithm divided the variables in hotel data into 2 clusters. The average silhouette coefficient is 0.09 which

indicates a moderate quality of clustering capacity. This further indicates that the data points are not well separated with respective clusters, but it achieved some level of separation using this method.

HIERACHICAL DENDOGRAM, SILHOUTTE HIERACHICAL AND GAP METHOD






```
# Silhouette Coefficient for K-means method
HotelCluster1 <- kmeans(HotelDataScale, centers = 3, nstart = 20)
sil <- silhouette(HotelCluster1$cluster, dist(HotelDataScale))
fviz_silhouette(sil)  # Average Silhouette width

# Apply hierarchical clustering method using complete linkage
hc.cut <- hcut(HotelDataScale, k = 3, hc_method = "complete")

# Show cluster dendrogram
fviz_dend(hc.cut, show_labels = FALSE, rect = TRUE)

fviz_dend(hc.cut, show_labels = FALSE, rect = TRUE, guides = "none")


# Silhouette Coefficient for hierarchical clustering method
fviz_silhouette(hc.cut)  # Average Silhouette width: 0.67

# Gap Statistic Method
gap_stat <- clusGap(HotelDataScale, FUN = kmeans, nstart = 25, K.max = 10, B = 50)

# Print the result
print(gap_stat, method = "firstmax")
fviz_gap_stat(gap_stat)
```

The silhouette hierarchical shows coefficient of 0.86, which is closer to 1 which indicates a strong quality of clustering capacity. The hierarchical silhouette coefficient agrees with the elbow method that the optimum number of clusters is 3. However, the Gap method suggests 1

SUPERVISED LEARNING -CLASSIFICATION

KNN MODEL

```
[1] "Accuracy: 0.907425305710804"
> # Calculate error percentage
> error <- 1 - accuracy
> print(paste("Error percentage:", error * 100, "%"))
```

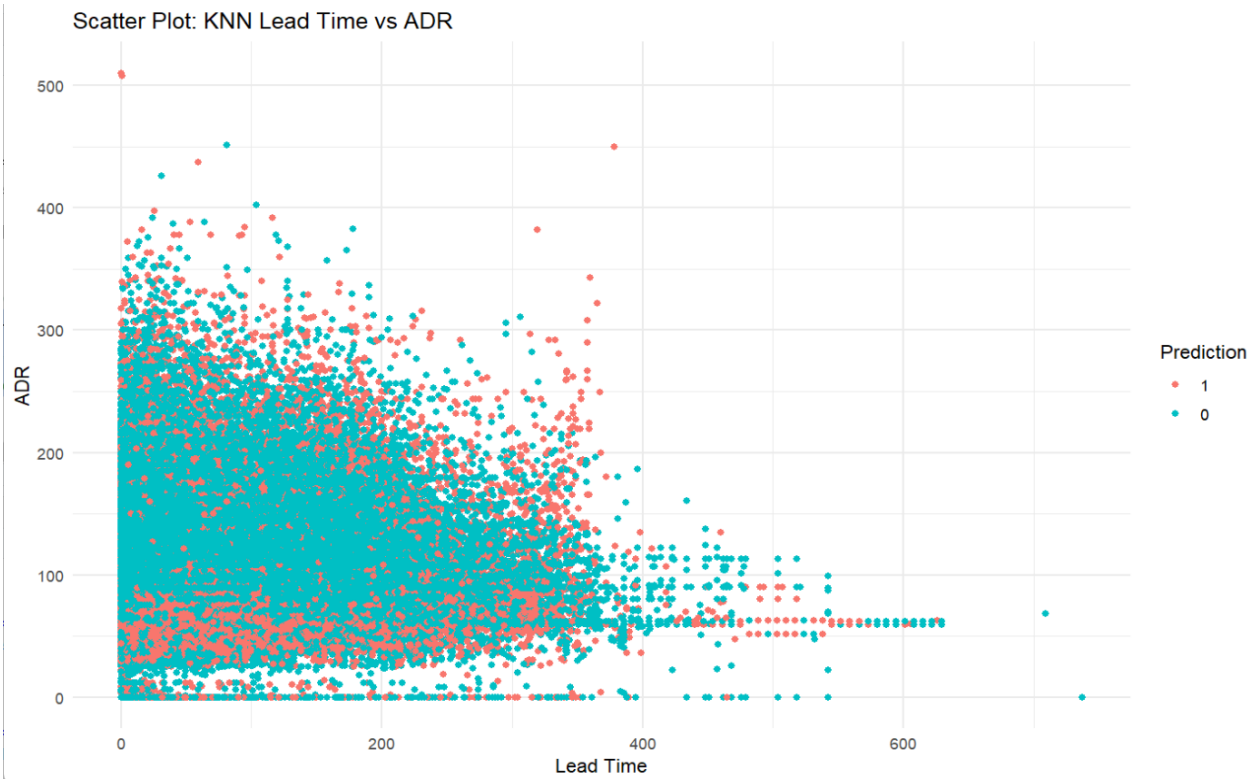
```
# Calculate accuracy
accuracy <- sum(test_hotel_labels == knn_model) / length(test_hotel_labels)
print(paste("Accuracy:", accuracy))

# Calculate error percentage
error <- 1 - accuracy
print(paste("Error percentage:", error * 100, "%"))
```

KNN OPTIMAL VALUE OF K

After a thorough iterative process range of 305-315, the optimal value of K was determined at 305 with an accuracy prediction of 90.74%. The KNN algorithm achieved a high level of accuracy in predicting the class labels for 305 neighbors in ESTANA HOTEL data. Moreso, the sensitivity is 92% and specific is 90% which indicate the model can predict both cancellation as well as non-cancellation.

```
> confusionMatrix(table(test_hotel_labels, knn_model))
Confusion Matrix and Statistics

                knn_model
test_hotel_labels     1     0
                1  7260  1590
                0   613 14334

               Accuracy : 0.9074
                 95% CI : (0.9037, 0.9111)
    No Information Rate : 0.6692
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7973

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9221
            Specificity : 0.9002
         Pos Pred Value : 0.8203
         Neg Pred Value : 0.9590
             Prevalence : 0.3308
         Detection Rate : 0.3051
   Detection Prevalence : 0.3719
      Balanced Accuracy : 0.9111

       'Positive' Class : 1
```

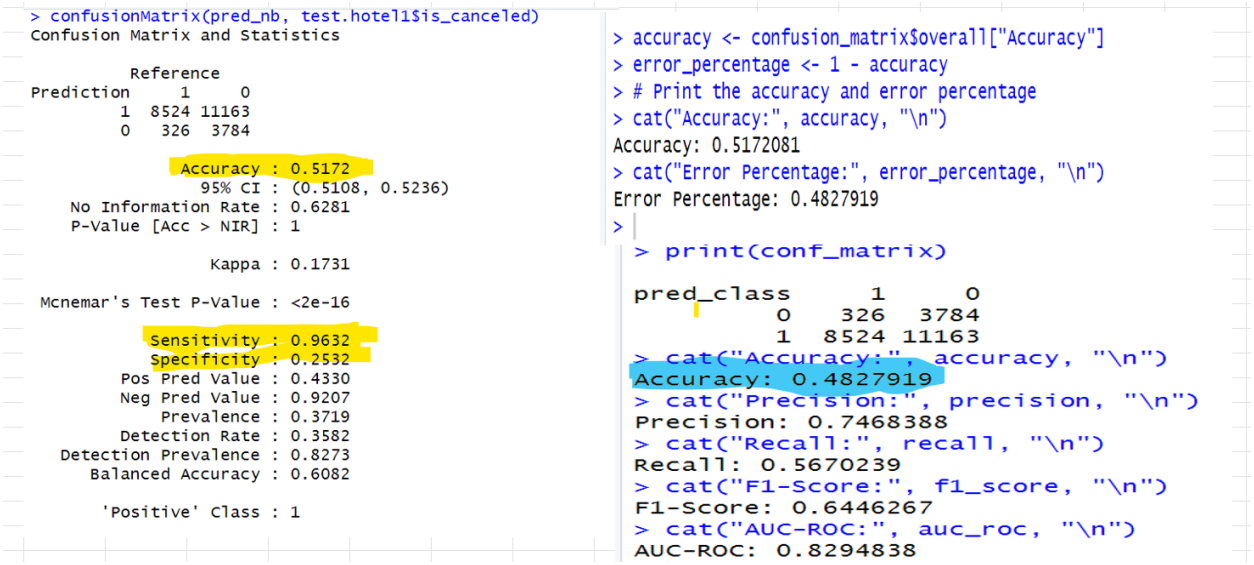
```
> k.optm <- numeric(0)
> for (i in 305:315) {
+   knn.mod <- knn(train = train.hotel, test = test.hotel, cl = train_hotel_labels,
= i)
+   k.optm[i] <- 100 * sum(test_hotel_labels == knn.mod) / NROW(test_hotel_labels)
+   cat(i, "=", k.optm[i], "\n")
+ }
305 = 90.74253
306 = 90.72152
307 = 90.71312
308 = 90.71312
309 = 90.70891
310 = 90.6837
311 = 90.62067
312 = 90.62067
313 = 90.58705
314 = 90.57864
315 = 90.52822
```



Graph shows KNN neighbors and lead time and Arrival date.

NAÏVE BAIYE

Naive bayes achieved an accuracy performance of 51.72% as shown in yellow painted area which indicates that the model can moderately classify the cancellations correctly. To further evaluate the model predictive capabilities, the model was tuned and other metrics such as AUC ROC, were used to measure how effective it identifies TRUE positives and FALSE negatives. The model achieved an accuracy of 48.3% as shown in the blue painted area indicating it can't effectively predict cancellation.



```
> confusionMatrix(pred_nb, test.hotel1$is_canceled)
Confusion Matrix and Statistics

          Reference
Prediction     1     0
         1  8524 11163
         0   326  3784

               Accuracy : 0.5172
                 95% CI : (0.5108, 0.5236)
    No Information Rate : 0.6281
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1731

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9632
            Specificity : 0.2532
         Pos Pred Value : 0.4330
         Neg Pred Value : 0.9207
             Prevalence : 0.3719
         Detection Rate : 0.3582
   Detection Prevalence : 0.8273
      Balanced Accuracy : 0.6082

       'Positive' Class : 1
```

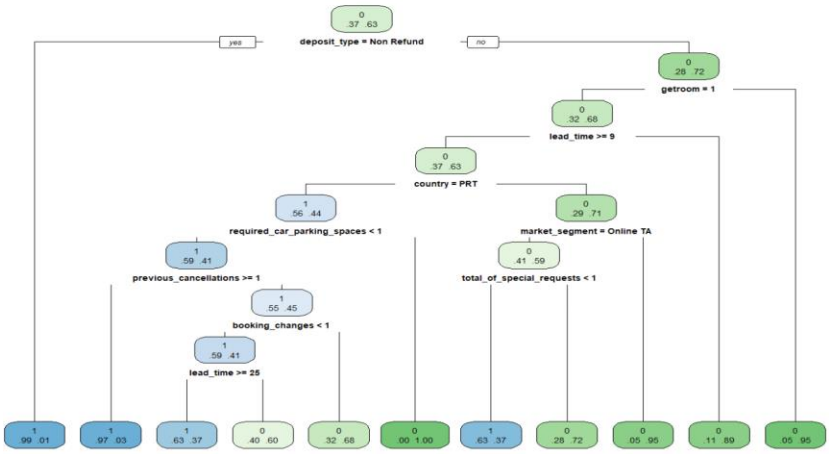
```
> accuracy <- confusion_matrix$overall["Accuracy"]
> error_percentage <- 1 - accuracy
> # Print the accuracy and error percentage
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.5172081
> cat("Error Percentage:", error_percentage, "\n")
Error Percentage: 0.4827919
>
```

```
> print(conf_matrix)

pred_class     1     0
         0   326  3784
         1  8524 11163
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.4827919
> cat("Precision:", precision, "\n")
Precision: 0.7468388
> cat("Recall:", recall, "\n")
Recall: 0.5670239
> cat("F1-Score:", f1_score, "\n")
F1-Score: 0.6446267
> cat("AUC-ROC:", auc_roc, "\n")
AUC-ROC: 0.8294838
```

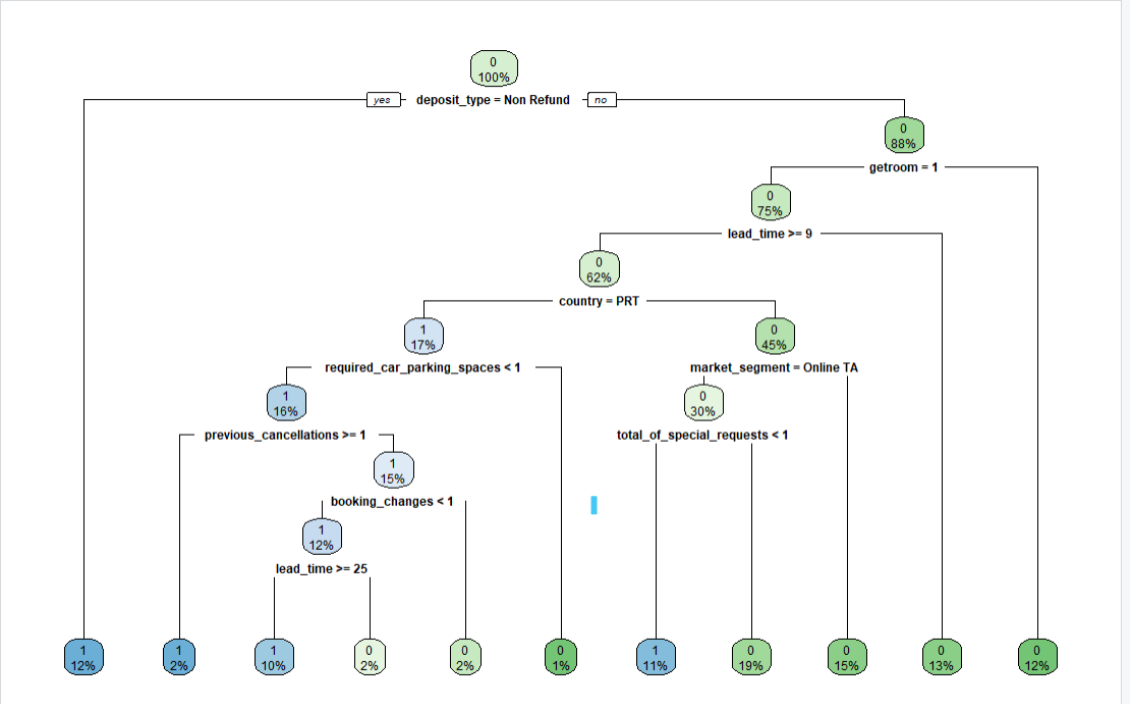
DECISION TREE MODEL

The tree below represents the first tree built by the algorithm. To avoid overfitting and model accuracy, the tree was further pruned to reduce its nodes.

INITIAL TREE WITH MANY LEAFNODE



PRUNNING THE TREE

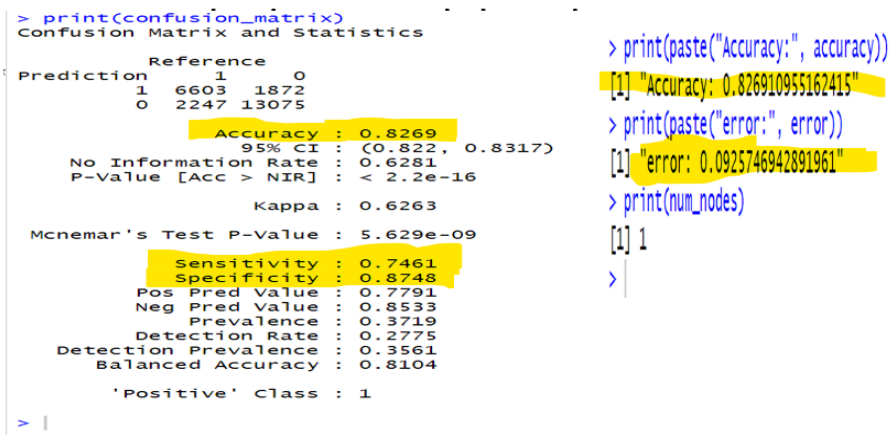


The tree above was pruned to remove unnecessary nodes, avoid overfitting, and improve the predictive accuracy of the built model.

From the pruned tree, the top node shows whether the DEPOSIT = NO REFUND. If the guest says yes, there is a 12% chance of cancellation. If no, there is an 88% probability of cancellation. Node 2, get room=1 is no; there is a 12% probability of cancellation. If yes, the probability of cancellation = 75%. Node 3, lead time >=9; if no, there is a 13% probability that the guest would cancel, and if yes, there is a 62% chance of cancellation. Node 4, if country = Portugal, if no, there is a 45% probability of cancellation; if yes, there is a 17% probability of cancellation. If the country = is Portugal, book via an online travel agent; if no, there is a 15% chance of cancellation; if yes, there is a 30% probability of cancellation. The guest has a special request <1; if yes, there is an 11% probability of cancellation and if no, there is a 19% chance of cancellation.

Furthermore, if country = Portugal is no, the guest requires parking space <1; if no, there is a 1% probability of cancellation and if yes, there is a 16% chance of cancellation. Next node, guest previous cancellation>=1; if no, there is a 2% chance of cancellation. If yes, there is a 15% probability of cancellation. Next node, booking changes <1; if no, there is a 2% probability that the guest will cancel; if yes, there is a 12% probability cancellation rate. If the lead time is >25, if no, there is a 2% probability cancellation rate. If yes, yes is a 10% cancellation probability.

MODEL ACCURACY AND CONFUSSION MATRIX/ SENSITIVTY

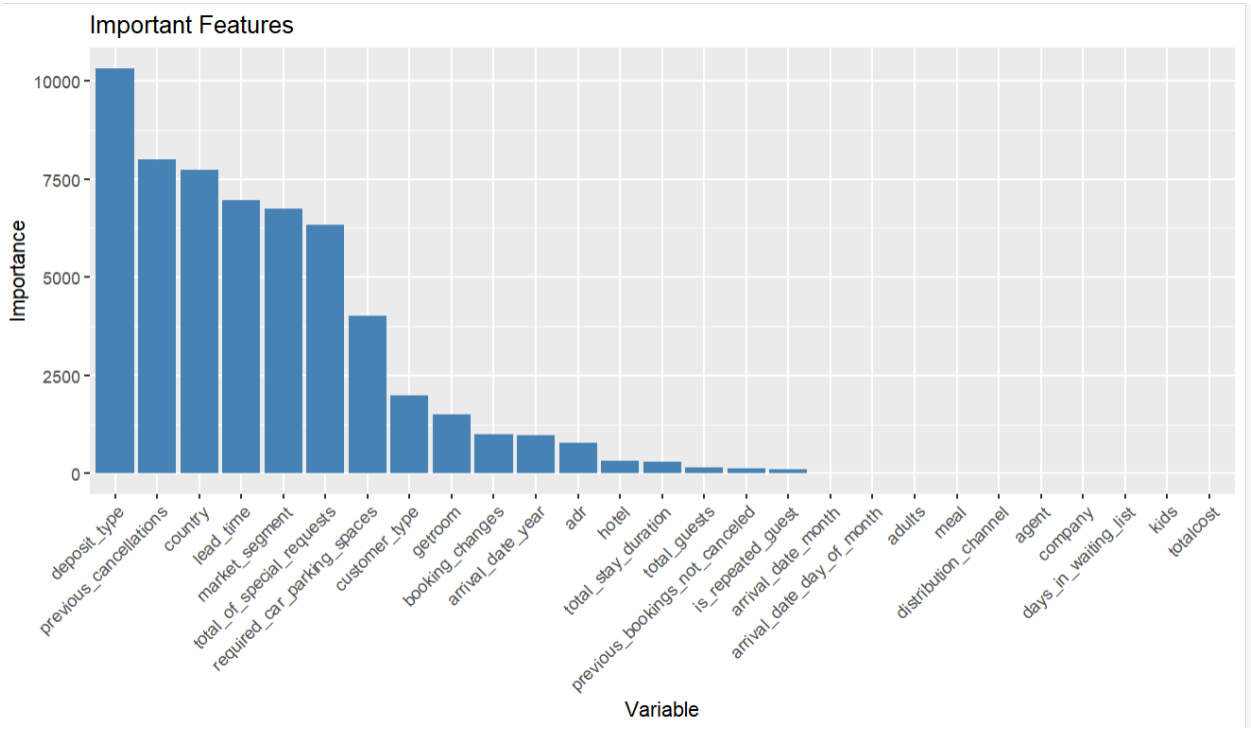


The pruned decision tree model achieved an accuracy of 82.69% of correctly predicting booking cancellation. The sensitivity of 74.6% indicates the model has high ability to determine TRUE POSITIVES

The node print is 1, which indicates that all the outcomes belong to a class and no further splitting could be done because all outcomes belong in the same outcome and a pure node. Similarly, the model was pruned using cp parameter of 0.01which helped to prevent overfitting.

MODEL IMPORTANT FEATURES

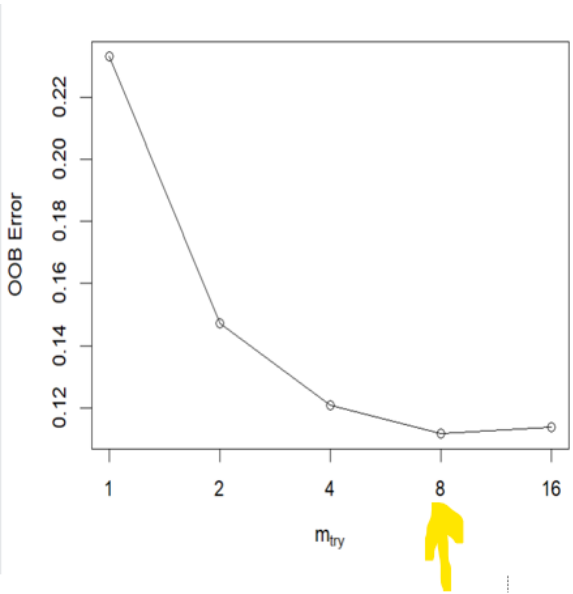
In building the model, the decision tree considered deposit type the most essential feature. The following important feature is previous cancellation. Country occupies the third, and lead time is placed fourth. The importance feature graph considers customer demographics such as age and gender as having less importance.



RANDOM FOREST MODEL

BEST MTREE

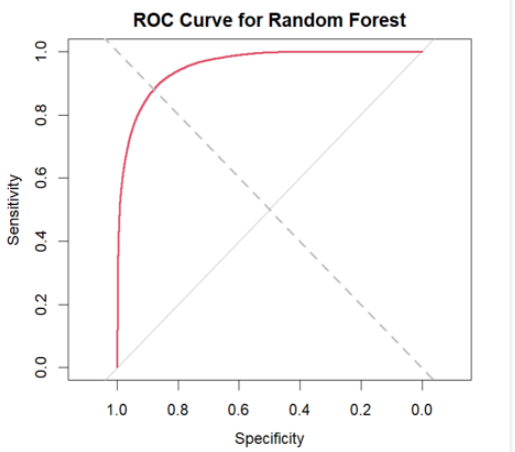
```
mtry = 2  OOB error = 14.72%
Searching left ...
mtry = 1        OOB error = 23.31%
-0.5833333 0.05
Searching right ...
mtry = 4        OOB error = 12.08%
0.1796518 0.05
mtry = 8        OOB error = 11.17%
0.07531745 0.05
mtry = 16       OOB error = 11.36%
-0.0174003 0.05
> best.m <- best_mtry[best_mtry[, 2] == min(best_mtry[, 2]), 1]
> print(best_mtry)
       mtry  OOBError
1.00B     1 0.2331386
2.00B     2 0.1472455
4.00B     4 0.1207925
8.00B     8 0.1116948
16.00B   16 0.1136383
> print(best.m)
[1] 8
```



The OOBError of 11.17% indicates incorrect prediction error rate out of bag, suggesting the model can predict whether a booking will be cancelled or not with about 88.83% prediction accuracy.

CONFUSSION MATRIX AND MODEL ACCURACY/ROC

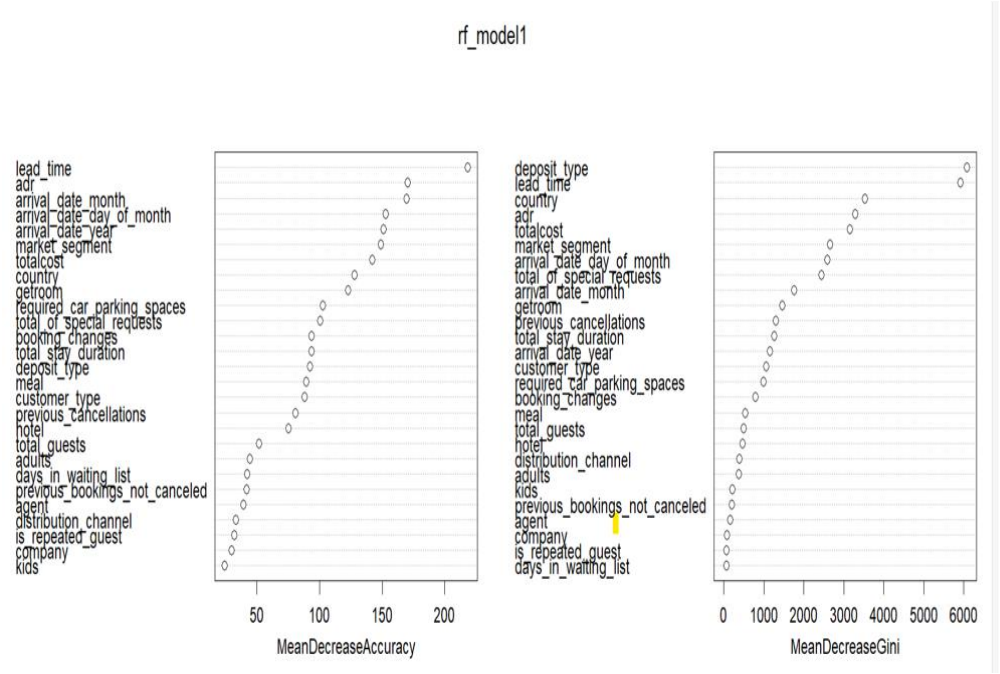
```
Confusion Matrix and Statistics

          Reference
Prediction     1     0
         1  7321  1035
         0  1529 13912

               Accuracy : 0.8923
                 95% CI : (0.8882, 0.8962)
    No Information Rate : 0.6281
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7667

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8272
            Specificity : 0.9308
         Pos Pred Value : 0.8761
         Neg Pred Value : 0.9010
             Prevalence : 0.3719
         Detection Rate : 0.3076
   Detection Prevalence : 0.3511
      Balanced Accuracy : 0.8790

       'Positive' Class : 1

>
```



The random forest model achieved an accuracy rate of 89.23% in predicting hotel booking cancellation or not correctly. Further tuning of the model achieved a ROC value of 0. 09 closer to the top left corner

of the curve, which indicates the better performance of the model in assigning values to its predicted classes correctly. The model has a sensitivity of 82.72% and a specificity of 93.08%.

FEATURES IMPORTANCE



The graph shows that features such as Lead time, ADR, deposit type, country and arrival date months are among the most important features that the random forest model used in predicting hotel booking cancellations. These variables are like the decision tree variables. However, in the decision tree features, deposit type, lead time are important in predicting booking cancellation.

XBOOST MODEL

The xboost model gave an accuracy of 1 with zero percentage error and AUC-ROC of 82.95% to predict hotel booking cancellation.

```
> print(conf_matrix)
      pred_classes
test.y     0
     0 35695
> cat("Accuracy:", accuracy, "\n")
Accuracy: 1
> cat("Error Percentage:", error_percentage, "\n")
Error Percentage: 0
> cat("AUC-ROC:", auc_roc, "\n")
AUC-ROC: 0.8294838
> cat("Recall:", recall, "\n")
Recall: 1
>
```

DISCUSSION AND MODEL COMPARISON ANALYSIS

CURRENT STUDY MODEL COMPARISON

MODEL	ACCURACY	ERROR
KNN	90.74%.	9.26%
NAÏVE BAYES	48.27%	51.73%
DECISION TREE	82.69%	17.31%
RANDOM FOREST	89.23%	10.77%
XBOOST	100%	0%

BENCHMARK MODEL

BENCHMARK	MODEL	ACCURACY
Antonio et al. (2017)	XBOOST	91%

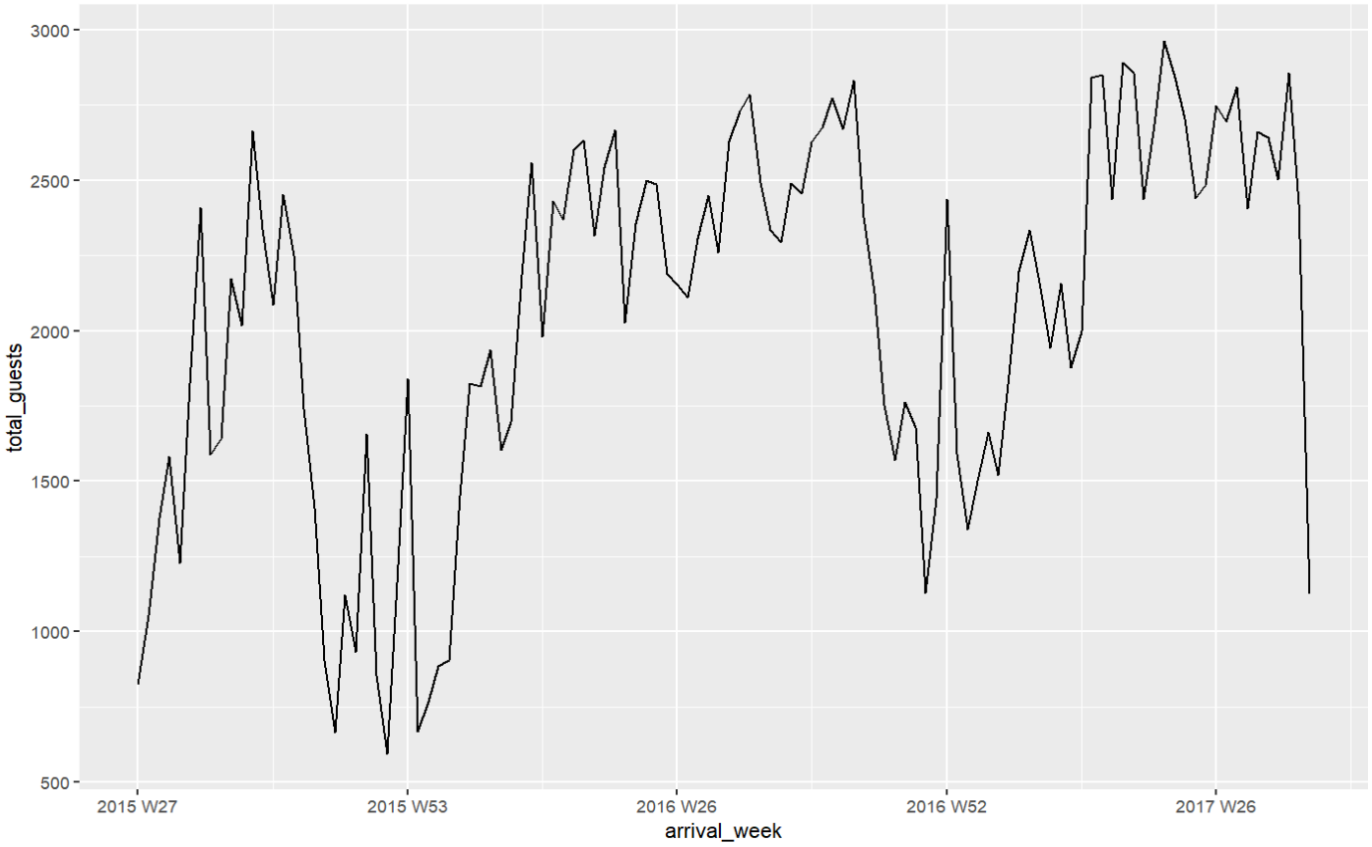
KNN in this study achieved an accuracy of 90.74% with an error rate of 9.26%. The model performed close to the benchmark model and predicted hotel booking cancellations with 90% accuracy. In contrast, to the benchmark, the Naive Bayes model achieved a low accuracy of 48.27% with a high error rate. The result implies that the Naive Bayes model did not perform well in predicting hotel booking cancellation. Moreso, the Decision Tree model, with an accuracy of 82.69%, falls slightly below the benchmark model accuracy. However, it did perform well in predicting hotel booking cancellations.

The Random Forest model achieved an accuracy of 89.69%, indicating that it effectively predicted ESTANA booking cancellations and came close to the benchmark model. Surprisingly, the XBOOST model achieved an accuracy of 100% with no errors. The result indicates that the model outperformed the bench march model and could predict booking cancellations without errors. However, it is essential to note that there may be overfitting in the model since further tuning could not be done.

TIME SERIES FORECASTING

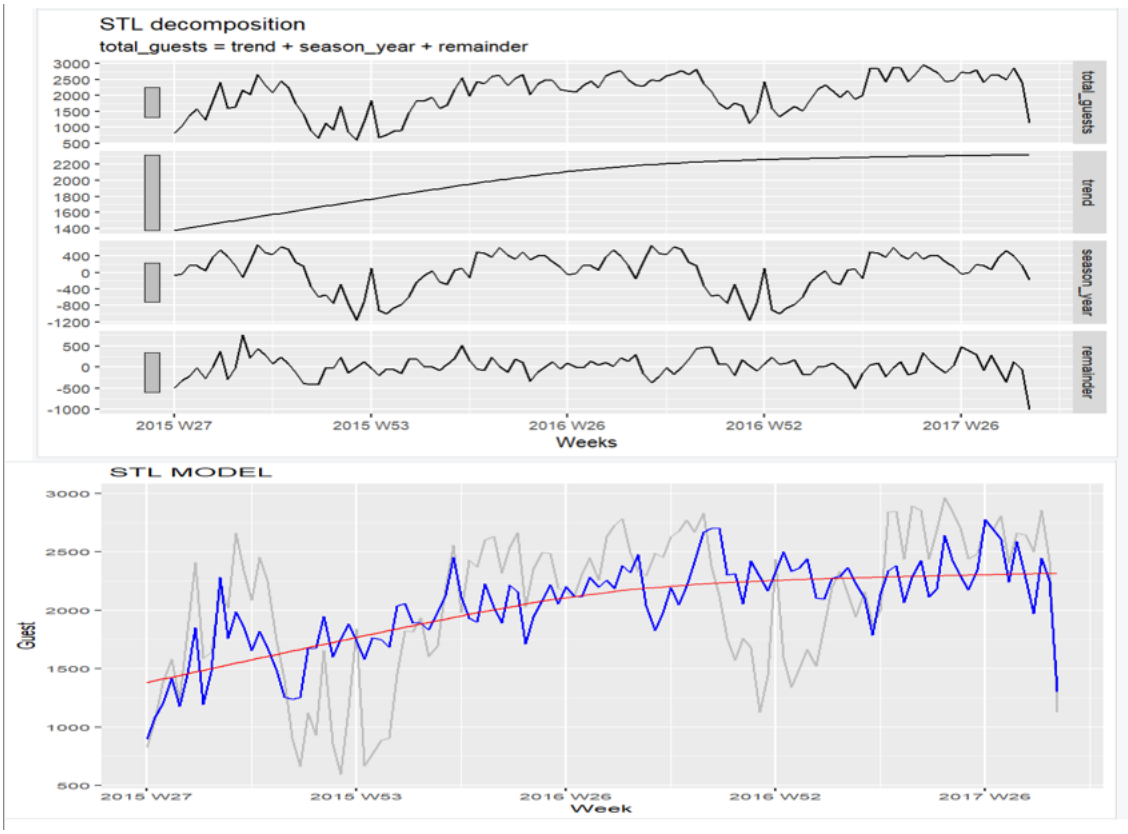
The Graph below shows the weekly guest arrival time series plot for ESTANA HOTEL. It shows that there is an observed trend and seasonality, but it shows some irregularities.

TIME SERIES PLOT




```
#Plotting the time series with weekly frequency
wk_tsb %>% ggplot(aes(x=arrival_week,
                      y=total_guests))+geom_line(aes(group=1))
```

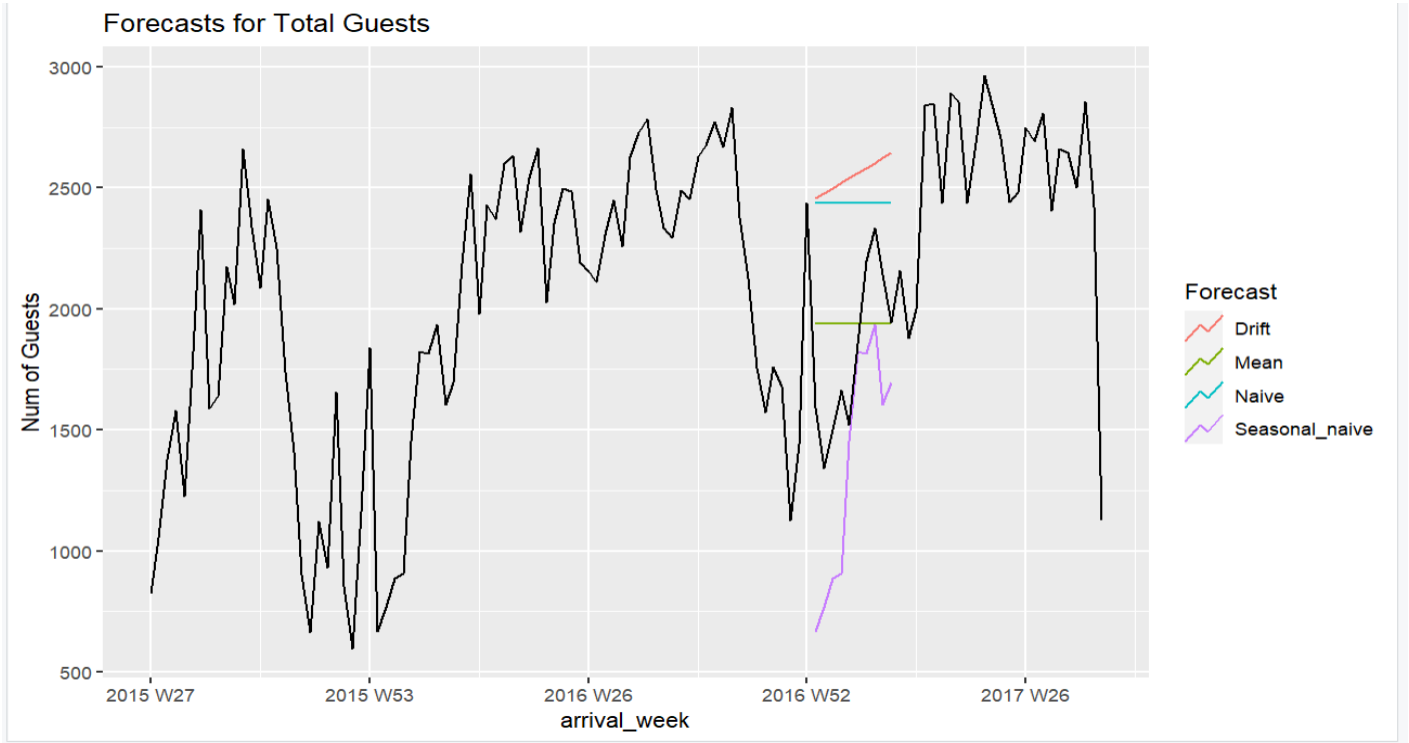
STL DECOMPOSTION



The Graph was decomposed using the STL method. The decomposed graph shows elements of seasonality, trend, and the remainder of irregular components. The STL forecast was also fitted into model. As seen in the second graph with blue markings, it shows a good fit to forecast arrival guest.

```
#STL Model
wk_tsb %>%
  model(stl = STL(total_guests))

dcmp <- wk_tsb %>%
  model(stl = STL(total_guests))

components(dcmp)
components(dcmp) %>% autoplot() + xlab("Weeks")

wk_tsb %>%
  autoplot(total_guests, color='gray') +  autolayer(components(dcmp), season_adjust, color='blue') + autolayer(components(dcmp), trend, color='red')
```

MEAN, DRIFT, NAÏVE AND SEASONAL NAÏVE MODELS




```
> #Measuring Accuracy
> accuracy(arrival_fc, wk_tsb)
# A tibble: 4 x 10
  .model         .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
  <chr>          <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Drift          Test  -742.  791.  742. -45.0  45.0 1.04  0.953 0.689
2 Mean           Test  -130.  345.  302. -10.6  18.3 0.423 0.416 0.743
3 Naive          Test  -628.  705.  628. -39.0  39.0 0.880 0.849 0.743
4 Seasonal_naive Test   453.  528.  453.  26.5  26.5 0.634 0.636 0.245
>
```

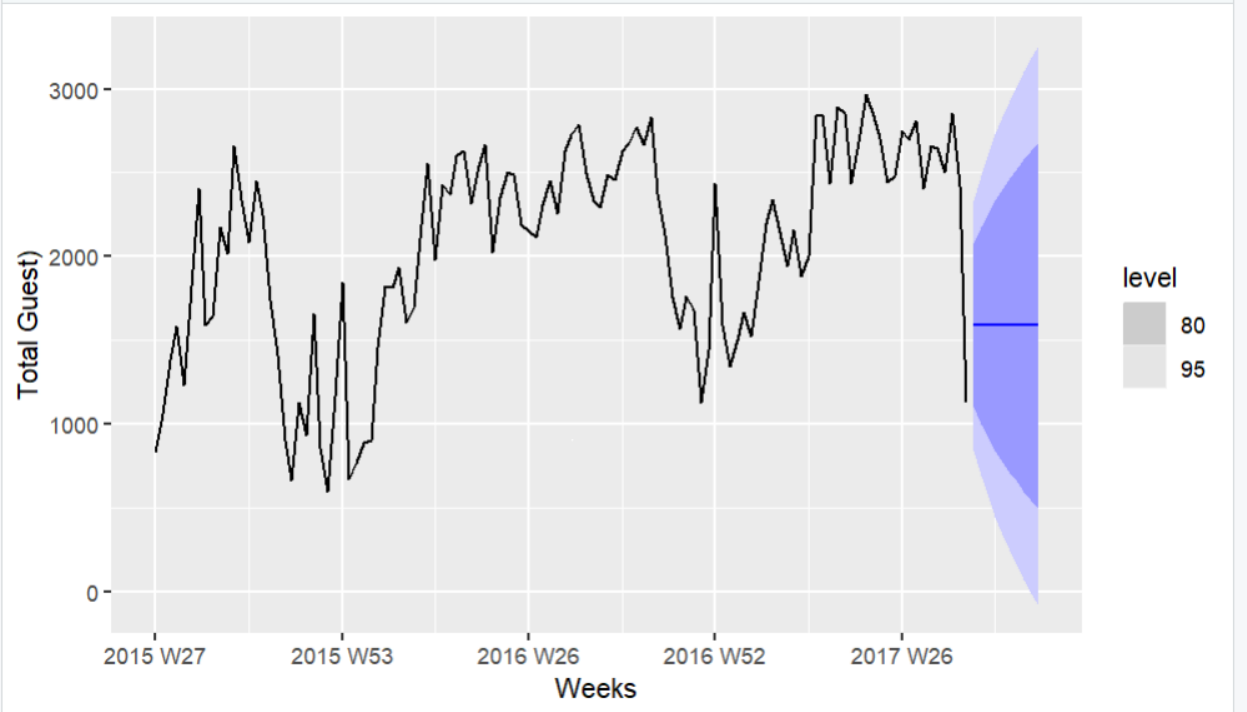
MEAN FORECAST AS BENCHMARK VS OTHER MODELS

The above graph shows the four models and the mean model being the benchmark model. From the above graph, the seasonal naïve forecast has the least MAPE, MASE, RMSE Errors. However, it does not perform well as the Mean forecast benchmark model which has lower MAE, RMSE, MASE error.

ETS ADDITIVE MODEL

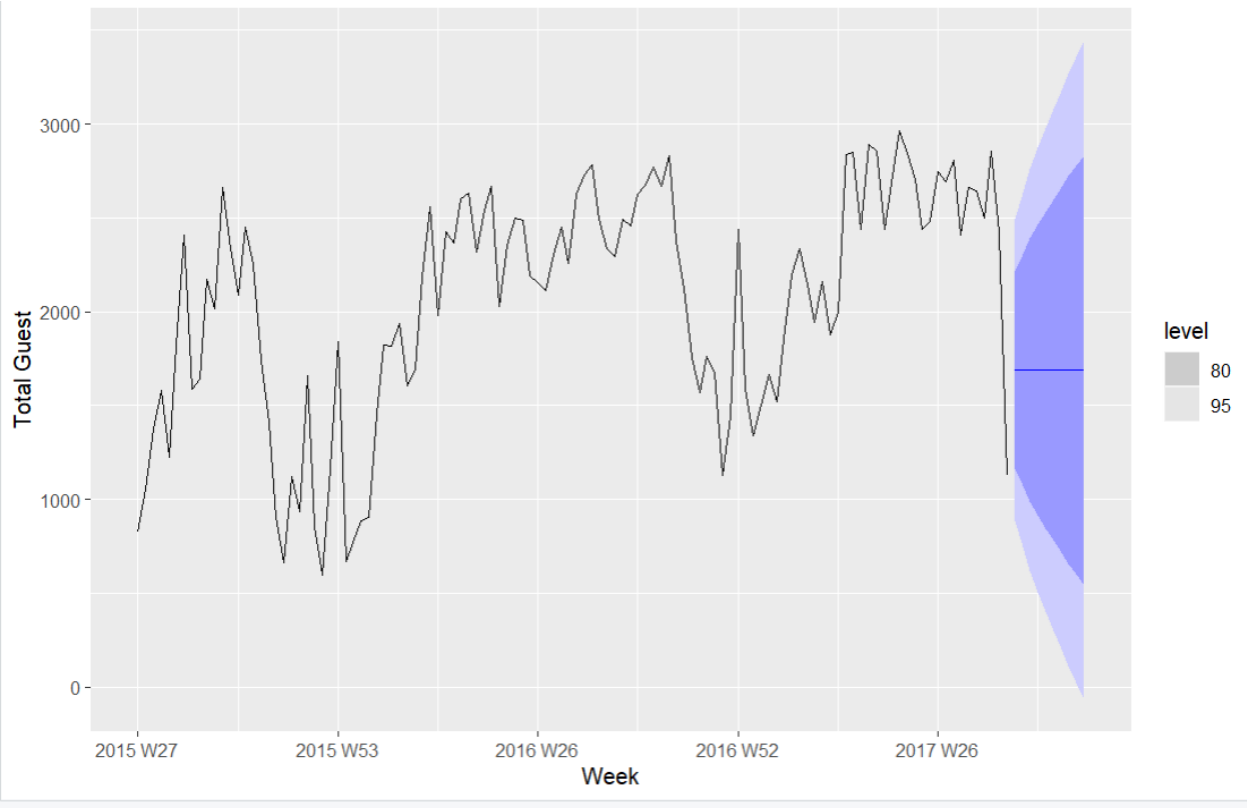
```
> report(fit)
Series: total_guests
Model: ETS(A,N,N)
  Smoothing parameters:
    alpha = 0.6714875

  Initial states:
     l[0]
 940.9779

  sigma^2:  143907.3

     AIC     AICc      BIC
1897.878 1898.096 1906.087
>
```

```
#ETS(A,N,N)
fit <- wk_tsb %>% model(ANN = ETS(total_guests ~ error("A") + trend("N") + season("N")))
report(fit)

fit %>%
  forecast(h = 10) %>%  autoplot(wk_tsb) +
  ylab("Total Guest)") + xlab("Weeks")
```



ETS MULTIPLICATIVE MODEL

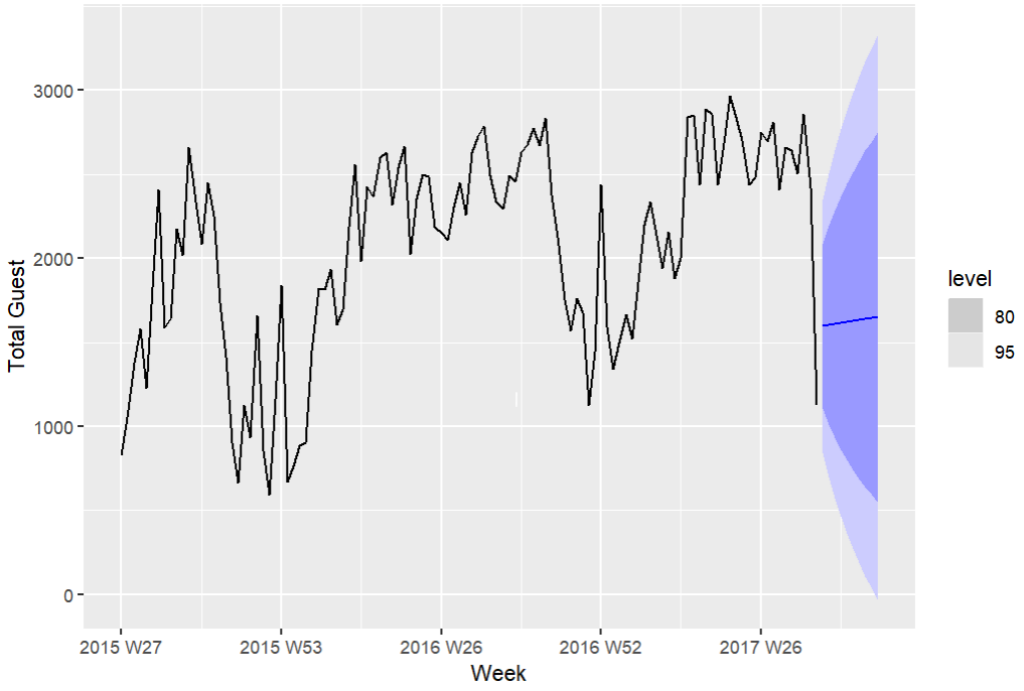
```
> fit_multiplicative <- wk_tsb %>%
+   model(ETS_multiplicative = ETS(total_guests ~ error("M") + trend
("N") + season("N")))
> report(fit_multiplicative)
Series: total_guests
Model: ETS(M,N,N)
  Smoothing parameters:
    alpha = 0.6032805

  Initial states:
     l[0]
 954.4916

  sigma^2:  0.0585

     AIC     AICc      BIC
1947.038 1947.256 1955.247
> fit_multiplicative %>%
+   forecast(h = 10) %>% autoplot(wk_tsb) +  ylab("Total Guest") + x
lab("Week")
>
```



ETS HOLT MODEL

```
> report(fit)
Series: total_guests
Model: ETS(A,A,N)
  Smoothing parameters:
    alpha = 0.6714059
    beta  = 0.0001000338

  Initial states:
     l[0]     b[0]
 896.6191 6.077898

  sigma^2:  146483.5

     AIC     AICc      BIC
1901.847 1902.402 1915.528
> |
```

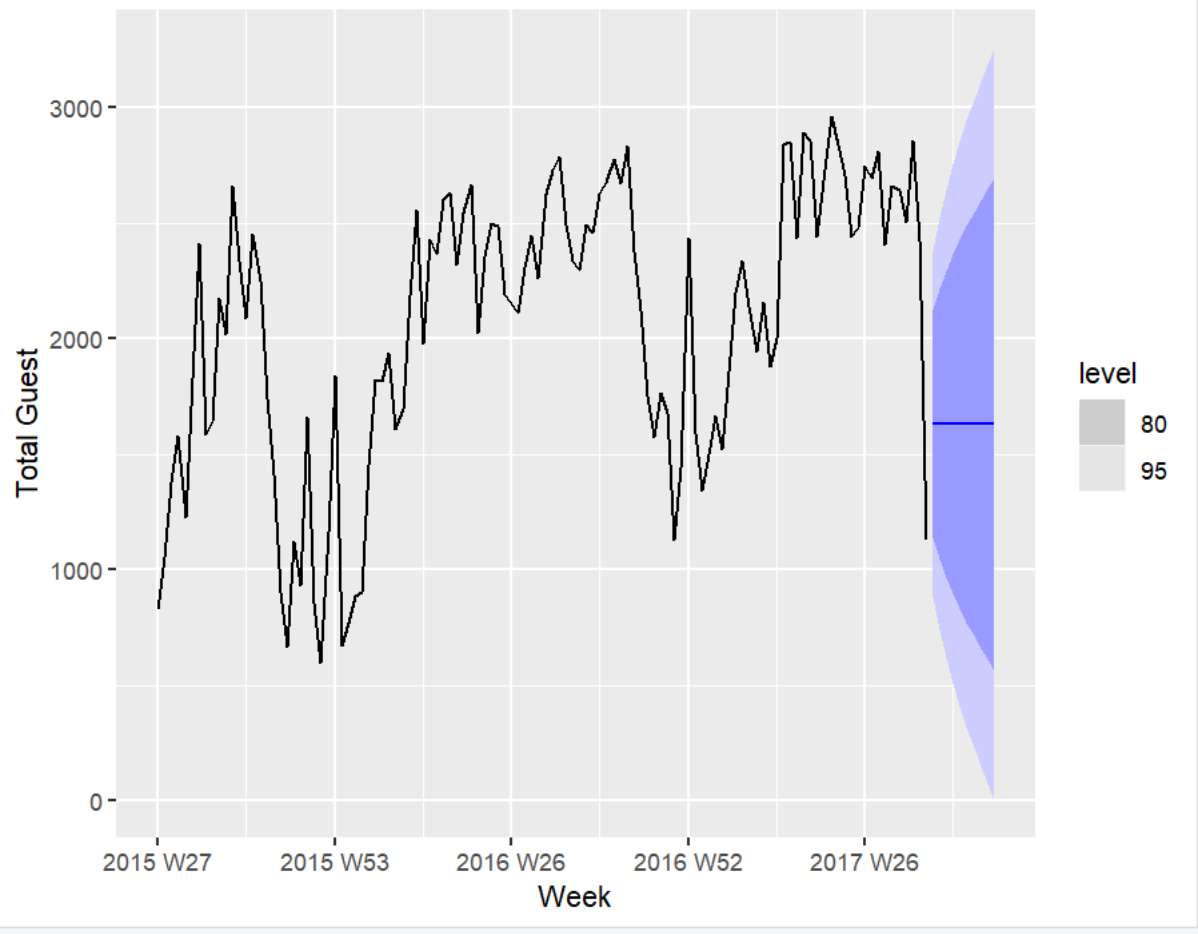
```
#ETS(A,A,N) - Holt's linear model
fit <- wk_tsb %>%
  model(AAN = ETS(total_guests ~ error("A") + trend("A") + season("N")))

report(fit)
fit %>%
  forecast(h = 10) %>%  autoplot(wk_tsb) +  ylab("Total Guest") + xlab("Week")
```



DAMPED MODEL

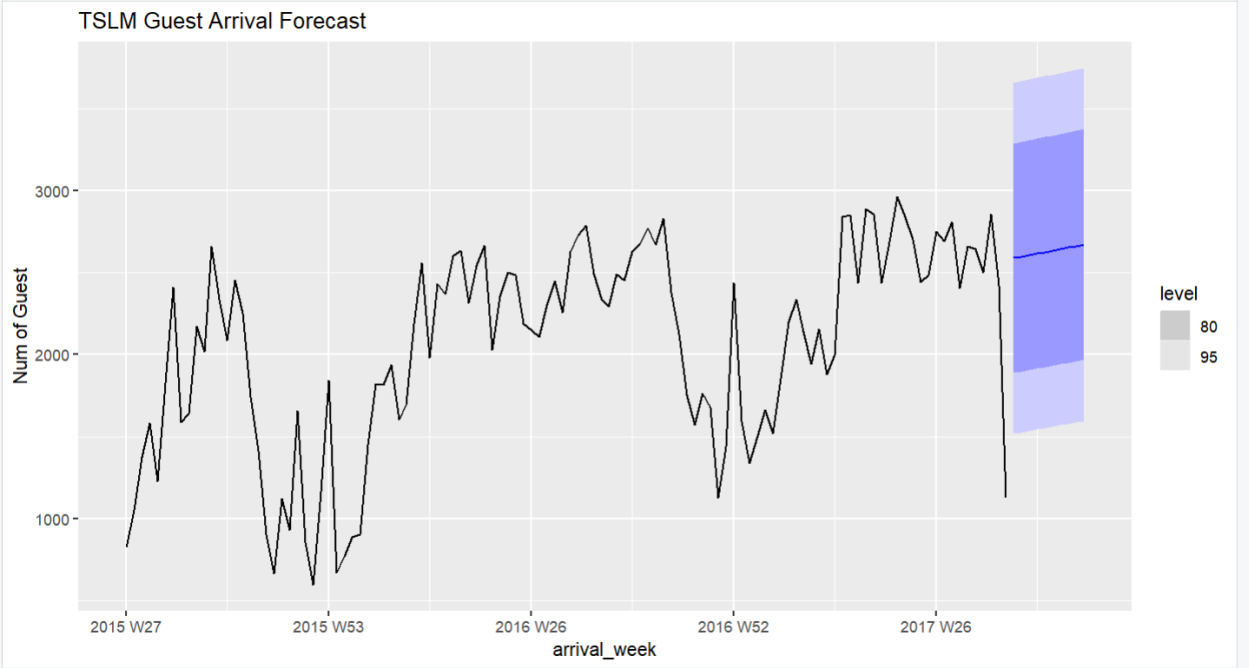
```
> report(fit)
Series: total_guests
Model: ETS(A,Ad,N)
  Smoothing parameters:
    alpha = 0.6430396
    beta  = 0.0001002854
    phi   = 0.8681099

  Initial states:
     l[0]     b[0]
 979.6807 123.5943

  sigma^2:  146198.6

     AIC     AICc      BIC
1902.584 1903.369 1919.001
> |
```

```
#damped trends model
fit <- wk_tsb %>%
  model(holt = ETS(total_guests ~ error("A") + trend("Ad") + season("N")))
report(fit)

fit %>%
  forecast(h = 10) %>% autoplot(wk_tsb) +  ylab("Total Guest") + xlab("Week")
```



TSLM MODEL

```
#Forecasting and visualizing
tslm_fit %>% forecast(h = 10)  %>%  autoplot(wk_tsb) +
  ggtitle("TSLM Guest Arrival Forecast") + ylab("Num of Guest")


#OTHER FORCASTING METHOD
```



REGRESSION ANALYSIS

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	getroom, total_of_special_requests, previous_cancellations, booking_changes, is_repeated_guest, required_car_parking_spaces, total_guests, lead_time, agent, adr, previous_bookings_not_canceled[b]	.	Enter

a. Dependent Variable: is_canceled

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.464[a]	.216	.215	.428

a. Predictors: (Constant), getroom, total_of_special_requests, previous_cancellations, booking_changes, is_repeated_guest, required_car_parking_spaces, total_guests, lead_time, agent, adr, previous_bookings_not_canceled

The R-squared of 21.6% explains the degree of variability in the dependent variable(Is cancelled) caused by the independent variables listed

Variable entered/Removed shows the independent variables that are being predicted to cause the increase in the attrition rate in the company.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6000.900	11	545.536	2981.662	.000[b]
	Residual	21841.673	119377	.183		
	Total	27842.573	119388			

a. Dependent Variable: is_canceled

b. Predictors: (Constant), getroom, total_of_special_requests, previous_cancellations, booking_changes, is_repeated_guest, required_car_parking_spaces, total_guests, lead_time, agent, adr, previous_bookings_not_canceled

H0: b1 = b2 = 0 H1 : at least one of the above bi ≠ 0

Test statistic is F (with df= k; n-(k+1)) F11,1458 = 2981.662

p<.05 Reject H0. At least one of the independent variables carries significant information to help explain the variable Is Cancelled in ESTANA hotel.

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.009	.006		1.551	.121
lead_time	.001	.000	.222	82.448	.000
is_repeated_guest	-.025	.008	-.009	-3.155	.002
previous_cancellations	.045	.001	.079	30.333	<.001
previous_bookings_not_canceled	-.004	.001	-.013	-4.649	<.001
booking_changes	-.078	.002	-.106	-40.756	.000
agent	.037	.004	.026	9.490	<.001
adr	.001	.000	.072	25.366	<.001
required_car_parking_spaces	-.259	.005	-.132	-50.298	.000
total_of_special_requests	-.127	.002	-.209	-78.372	.000
total_guests	.019	.002	.029	10.213	<.001
getroom	.252	.004	.173	65.296	.000

ependent Variable: is_canceled

H0: b1 = 0

H1: b1 ≠ 0

p<.05 Reject H0.

X1,2,3,4,6,8,9,10,11 (lead time, is repeated guest, previous cancelation, previous booking not cancelled, booking changes adr,, required car parking space, total special request, total guests, getroom)carries unique information about Y(IS cancelled) over and above the other variables in the model.

CONCLUSION AND DISCUSSION

The study's objective was to develop models to predict hotel booking cancellations and to understand the factors that lead to cancellation. To achieve this, descriptive analysis visualization was used to provide insights into the data of ESTANA hotels. Moreso, predictive models were built to predict ESTANA's hotel booking cancellation by dividing the dataset into 80% train and test. After a series of models, evaluation, tuning and pruning of various models, XBOOST achieved 100% accuracy, while KNN achieved 90.74%, Random Forest 89.23% and Decision Tree with 82.69% accuracy in predicting hotel booking cancellation. The naive Bayes model could have performed better in predicting cancellation. Moreso, the model was compared to a benchmark model by Antonio et al. (2017) XBOOST model, which achieved 91% accuracy in predicting cancellation. The Decision Tree and Random Forest models showed that lead time, deposit type, previous cancellations and country were among the variables that impact booking cancellation. The Elbow method and Hierarchical methods segmented the features into three data clusters. The clustering model was used to segment ESTANA data into clusters.

To further achieve the study's objectives, Time series forecast models were developed to forecast guest arrival dates. The models benchmarked against the Mean forecast. However, the lowest forecast with the least MAPE, MAE, was a seasonal naive model. A Regression model was developed to understand the features that affect hotel cancellation. As opined by Chew & Jahari (2014), lead time and deposit type, the country was among the important features the model identified as those causing cancellations.

The results support the literature reviewed by Morales and Wang (2010) and Antonio et al. (2017). Thus, ESTANA hotels can adopt the machine learning models above to predict hotel cancellations and forecast guest arrival week.

LIMITATION AND FUTURE WORK

There were limitations to computing Unsupervised clustering with the entire data set because of memory warning. Future research could use all the observations. There will be a need to further tune the XBOOST parameters to see if it gives the same result.

ETHICAL CONSIDERATION

The following ethics of secondary data analysis were taken into consideration:

- The hotel data was freely data available, therefore no form of permission was required to obtain the data.
- There was no personal or identifying information on the data.
- The data does not carry any potential harm.