Syracuse University
School of Information Studies

School of
Information Studies

IST 687 – Applied Data Science

Lab Section M003 | Group 3

Hotel Industry Analysis for European Hotels Group

Recommendations to Maximize Revenue and Accommodate Wider Traveler Range

EUROPEAN
HOTEL GROUP

Submitted by:

Abhijit Gokhale | Srishti Sanghvi | Ruixue Qin | Mackenzie Ess | Thejaswini Ram | Kevin Chung

Table of Contents

Description

The European Hotels Group is a hospitality supply-chain giant of hotel and resort properties spanning the European continent. Faced with possibilities of growth prospects, the group believes that the best way to maximize profits and revenues would be to acquire a number of hotel properties, situated in a variety of settings. By accommodating a large number of travelers from varying countries, the group strategizes between either expanding their base in, *resort* or *city* settings.

This project revolves around analyzing European Hotels Group's resort and city data, in order to provide the new owners of the chain recommendations on expanding and building hotels in either resort or city locations. The solidified business recommendations serve purpose to reduce risk and uncertainty and answer business questions to compare and comprehend how each property is performing individually over time, as well as how each property is performing in comparison to the other over time.

Project Scope and Objective

Scope

The scope of the project is based on two datasets acquired from the client (European Hotels Group) containing information affiliated with resort (H1) and city hotels (H2). The two datasets span 31 variables containing information that describe attributes related to a customer's stay – these attributes narrate aspects such as a guest's arrival time, number of weeknights stayed in the hotel, reservation status, etc. In particular, data affiliated with the hotel located in a resort spans 40,060 observations (i.e. *records*), and the data affiliated with the hotel located in a city spans 79,330 observations, wherein each observation and/or record represents a hotel booking.

The span of time for which data has been collected spans July 1st, 2015 through August 31st, 2017. In essence, the data includes bookings that either were successful in arrival or were cancelled.

Objective

The objective of the project is to deliver recommendations to the European Hotels Group for expanding their hotel locations in either resort or city settings, in order to best maximize the business' revenue. The workings of the project consist of using various data analysis techniques and models to support the interpreted recommendations, driven by correlational trends between the various parameters present in the data sets.

Deliverables

Starting from the first stage, the following are deliverables produced through the tenure of the project:

- Processing and cleaning of the data for selection of the most suited variables that can be used for data analysis. Each of the 31 variables provided in the data set were analyzed to gauge their use in being able to provide a recommendation of the business model to stakeholders.
- New columns were created in both data sets to address Average Daily Rate (ADR), Visitor Type, Season, and Average Revenue per Stay.
    - *ADR*: The average daily rate was calculated by summing the revenues generated by all occupied rooms and dividing the calculated sum by the total number of occupied rooms over a given period of time. Calculating the ADR for each data set allowed for comparison across varying time periods to help gauge key trends and observations for further analysis.
    - *Visitor Type*: Calculated to indicate whether the guest was single ("single"), came with their spouse ("couple"), and/or came with multiple members of their family ("family").
    - *Season*: Calculated to indicate the season in which the hotel booking was made.
    - *Average Revenue per Stay*: Calculated by multiplying the length of each booking's stay by its respective ADR value.
- Remodified/cleaned data sets consisting of addressed missing values (NAs). The missing values were either replaced by interpolated values or omitted.
- Identification and analysis of correlation and association rules using linear models and algorithms such as support vector machines (SVMs) that can yield actionable insights.
- A consolidated code repository with comments analyzing the steps taken to analyze, the data and each step taken in its coded examination.
- Recommendations for expanding the hotel locations in either resort or city settings, in order to best maximize the business revenue.

Data Requisition

The data used in the analysis of this project was made available by the course instructors, in the form of two excel files. Before proceeding with data cleaning or munging, the both data sets were read in and examined through their structures and consequently cleaned to address missing (NULL/NA values) or perform conversions of variables (numeric to categorical and vice-versa), and study any outliers. In essence, all outliers were examined to address retention and removal, based on the possibility of a bias free model – one that can clearly depicts any variability and any edge-use cases in providing data driven recommendations.

```
###### set the file directory ######
  getwd()
  setwd("C:\\Users\\abhij\\Desktop\\Lectures\\SEM 1\\IST 687\\Project")

  library(readxl)
  dataH1 <- read_excel("H1-Resort.xlsx")

  dataH2 <- read_excel("H2-City.xlsx")

###### checking data set structure ######
  str(dataH1)

  str(dataH2)
```

Data Preprocessing

- Check for missing values
- Use Interpolation to replace the missing values
- Remove unnecessary missing values because is not helpful for our model

```
# checking for NA values in the dataset H1 AND H2
colSums(is.na(dataH1))
colSums(is.na(dataH2))

# H1 - Resort #
# changing the NA values in Agent by its NA interpolated values.
dataH1$Agent <-  as.numeric(dataH1$Agent)
dataH1$Agent <- na_interpolation(dataH1$Agent)

# H2 - City #
# changing the NA values in Agent by its NA interpolated values.
dataH2$Agent <-  as.numeric(dataH2$Agent)
dataH2$Agent <- na_interpolation(dataH2$Agent)

# Remove 4 NA values from Children column in H2-city #
dataH2 <- subset(dataH2, Children!= "NA")
```

- Convert Children column from character to numeric to avoid null values
- Remove special character `
- Remove this outlier due to its extremely large value. we have different opinion on this outlier, it is too large and makes the later plot very unclear, but this outlier can help us in understanding the variability in the data and help us explore how the ADR varies.
- Remove Nulls values present under column Country from both datasets

```
# Convert Children column from character to numeric
dataH2$Children <- as.numeric(dataH2$Children)

# Diagnosing Outliers and Special characters in the dataset ###
# H1- Resort taking care of special character " #
dataH1[dataH1$Arrival_Date< dataH1$ReservationStatusDate & dataH1$ReservationStatus!= "Check-Out",]$ReservationStatus <- "Check-Out"

# H2 - City taking care of outlier ADR = 5400 #
dataH2<-subset(dataH2, ADR!= 5400)

# Removing Nulls values present under column Country from both datasets
dataH1 <- subset(dataH1, Country!= "NULL")
dataH2 <- subset(dataH2, Country!= "NULL")
```

- Created new column visitor_type so that we can observe the data set by group. For Resort, there are 13 NA records in visitor_type and we decide to remove them because we are not considering records where only children are going to stay in the hotel without parents(Adults). For City there are 390 NA records in visitor_type , we remove them because we are not considering records where only children are going to stay in the hotel without parents(Adults)
- Removing NA's from visitor_type

```
## Created new column visitor_type##
# H1 - Resort #
dataH1$visitor_type[(dataH1$Babies>= 0 | dataH1$Children>= 0)& dataH1$Adults > 0] <- "Family"
dataH1$visitor_type[(dataH1$Adults == 2)] <- "Couple"
dataH1$visitor_type[(dataH1$Adults == 1)] <- "Single"
sum(table(dataH1$visitor_type))

# H2 - City #
dataH2$visitor_type[(dataH2$Babies>= 0 | dataH2$Children>= 0) & dataH2$Adults > 0] <- "Family"
dataH2$visitor_type[(dataH2$Adults == 2)] <- "Couple"
dataH2$visitor_type[(dataH2$Adults == 1)] <- "Single"
sum(table(dataH2$visitor_type))
```

- created new column season
- convert the returned month value to numeric for easier comparison

```r
## created new column season ##
# Created below function to extract month value from the ReservationStatusDate
extractmonth = function (date) {
  month = format(date, format="%m")
  year = format(date, format="%Y")
  #monthyear = c(month, year)
  #list(day=day, month=month, year=year)
  return (month)
}
# For H1 - Resort Data #
# converting the returned month value to numeric so that it can be compared later in the code
monthonlyH1 <- as.numeric(extractmonth(dataH1$ReservationStatusDate))

dataH1$season[monthonlyH1>=1 & monthonlyH1<=3] <- "Spring"
dataH1$season[monthonlyH1>=4 & monthonlyH1<=6] <- "Summer"
dataH1$season[monthonlyH1>=7 & monthonlyH1<=9] <- "Fall"
dataH1$season[monthonlyH1>=10 & monthonlyH1<=12] <- "Winter"

# For H2 - City Data #
# converting the returned month value to numeric so that it can be compared later in the code
monthonlyH2 <- as.numeric(extractmonth(dataH2$ReservationStatusDate))

dataH2$season[monthonlyH2>=1 & monthonlyH2<=3] <- "Spring"
dataH2$season[monthonlyH2>=4 & monthonlyH2<=6] <- "Summer"
dataH2$season[monthonlyH2>=7 & monthonlyH2<=9] <- "Fall"
dataH2$season[monthonlyH2>=10 & monthonlyH2<=12] <- "Winter"
```

- Created new column AVGrevperstay
- Created New Column roomtypechanged

```r
## created new column AVGrevperstay ##
# For H1 - Resort Data #
dataH1$AVGrevperstay <- (dataH1$StaysInWeekendNights + dataH1$StaysInWeekNights) * dataH1$ADR
# For H2 - City Data #
dataH2$AVGrevperstay <- (dataH2$StaysInWeekendNights + dataH2$StaysInWeekNights) * dataH2$ADR

## Created New Column roomtypechanged ##
# For H1 - Resort Data #
dataH1$roomtypechanged[dataH1$ReservedRoomType == dataH1$AssignedRoomType] <- "No Change"
dataH1$roomtypechanged[dataH1$ReservedRoomType != dataH1$AssignedRoomType] <- "Change"
# For H2 - City Data #
dataH2$roomtypechanged[dataH2$ReservedRoomType == dataH2$AssignedRoomType] <- "No Change"
dataH2$roomtypechanged[dataH2$ReservedRoomType != dataH2$AssignedRoomType] <- "Change"
```

- Removing columns such as Arrival_Date, Company and ReservationStatusdate, to make linear model cleaner. And as assigned room type and reserved room type do not tell us whether change in room type hotel staff decision or customer decision is, we cannot decide whether it is good thing and whether is helpful to generate better ADR or not.
- Vector for Non-Numerical (categorical) columns

```r
# Removing certain columns
dataH1.1 <- dataH1[,!colnames(dataH1) %in% c("Arrival_Date","ReservationStatusDate","AssignedRoomType","ReservedRoomType","Company") ]

dataH2.1 <- dataH2[,!colnames(dataH2) %in% c("Arrival_Date","ReservationStatusDate","AssignedRoomType","ReservedRoomType","Company") ]

### Vector for Non-Numerical columns ######
datacols <- c("ReservationStatus","Country", "Meal","MarketSegment","DistributionChannel","DepositType","CustomerType","visiter_type","season","roomtypechanged")
```

- Creating countries table to present which countries are contributing
- Stort_Top_10 recorded countries from H1 in the variable

```
# Creating countries table to present which countries are contributing more in our H1
countryH1<- as.data.frame(table(dataH1.1$Country))
countryH1$Country_code = countryH1$Var1
countryH1$Var1 = NULL
countryH1$count <- countryH1$Freq
countryH1$Freq = NULL
countryH1 <- countryH1[order(-countryH1$count),]
#Storing_Top_10 recorded countries from H1 in the variable
toprecordedcountryH1 <- countryH1[1:10,]
# Creating countries table to present which countries are contributing more in our H2
countryH2<- as.data.frame(table(dataH2.1$Country))
countryH2$Country_code = countryH2$Var1
countryH2$Var1 = NULL
countryH2$count <- countryH2$Freq
countryH2$Freq = NULL
countryH2 <- countryH2[order(-countryH2$count),]
#Storing_Top_10 recorded countries from H2 in the variable
toprecordedcountryH2 <- countryH2[1:10,]
```

Initial Phase

Below we describe the key variables in our data set that were central to our analyses.

- **Visitor Type** - categorical variable that have Single, Couple and Family. we define Single when adult=1, define couple when adult=2 and define family when baby and children are greater than 0 under the condition of adult greater than 1.
- **Season** – categorical variable to indicate the time of year, there are 4 of them in total which are “Spring”, Summer”, Fall” and Winter”. we define spring when month is greater than January(1) and less than march(3), define summer when month is between April(4) and June(6), define fall when month is between July(7) and September(9) and winter when month is between October(10) and December(12).
- **ADR (Average Daily Rate)** - numeric value equals to the sum of the total revenue divided by the total occupied night.
- **Average Revenue per Stay** – numeric value equals to days stayed times ADR, but we find it not affecting that much after the linear modeling so we decided not to mention it.
- **Is_Canceled** – binary object, 0 for not canceled customer and 1 for canceled customer.
- **IsRepeatedGuest** – binary object,0 for not repeated guest and 1 for repeated guest.

Modeling

- ## Association Rules

Approach

We conducted Association Rules Mining analysis to examine frequent pairings of factors in our dataset. We examined two variables, set to the right-hand side, across the two datasets. The two variables were IsCanceled, a factor variable with two levels (0, indicating they did not cancel and 1, indicating they did cancel) and IsRepeatedGuest, a factor variable with two levels

(0, indicating that row does not contain a guest that has stayed at the hotel before, or 1, indicating that guest is a repeat customer of that hotel). We chose these two variables because we operationalized IsCanceled, or choosing to cancel a previously reserved hotel room, as a *negative* response to the hotel. Conversely, being a repeat customer (IsRepeatedGuest) represents a *positive* response to the hotel. Taken together, we expected these two variables to be highly informative in highlighting both areas of improvement, as well as areas that are strengths, for each respective hotel.

For the variables on the left-hand side, we constructed a new data frame to coerce into a transaction matrix. This data frame consisted of variables that we expected to be conceptually related to motivations for either canceling a reservation or being a repeat customer (IsCanceled and IsRepeatedGuest, respectively). These variables included: Meal, Country, Market Segment, Reserved Room Type, Agent, Company, number of Adult, Children, and Babies in the reservation, and a Customer Type classification. Critically, no continuous variables were included in the Association Rules Mining analysis. To keep the number of Adults, Childrens, and Babies factor variables, we binned these continuous variables into 3 levels: "none" (0 adults, children, or babies, respectively), "one" (1 adult, child, or baby, respectively), and "high" (2 or more adults, children, or babies, respectively).

We approached the rules mining analysis with the aim of examining the datasets separately because: 1) we were interested in examining any pairings that were similar across the two variables for the same hotel and 2) we were interested in pairings that would be different between the two hotels (to inform our business question related to each hotel's relative strengths and weaknesses).

Findings

As described above, we examined rule sets separately for the two datasets. As such, we will describe our findings by each hotel, in turn. We consider the things associated with customers not canceling their reservations as potential strengths of the hotel, or markets where the hotel is excelling. Conversely, we consider things associated with not being a repeat guest as areas where the hotel could consider for improvement.

For H1, not canceling is associated with...

… reservations for 2 or more adults. Our association rules mining analysis revealed that not canceling was associated with reservations with 2 or more adults (rule 4), reservations with 2 or more adults and no children (rule 13), and reservations with 2 or more adults, no children, and no babies (rule 26).

… reservations without children or infants. Our association rules mining analysis revealed that not canceling was associated with reservations with no babies (rule 8), reservations with no children (rule 5), reservations with no babies and no children (rule 19), reservations with 2 or more adults and no children (rule 13), and reservations with 2 or more adults, no children, and no babies (rule 26).

… reservations that selected the BB meal. Our association rules mining analysis revealed that not canceling was associated with reservations that chose the BB meal (rule 2)

… not being a repeat guest. Interestingly, our analysis revealed that people who chose not to cancel were also people who were not repeat guests (rule 7). We interpret this as an area for improvement for H1 in fostering customer satisfaction and loyalty.

Further, for H1, not being a repeat guest is associated with...

… reservations with the "A" Reserved Room Type. Our analysis revealed that reserving the "A" room type was associated with not being a repeat guest (rule 2). As such, we would recommend pushing other room types to customers to increase customer satisfaction and retention.

… reservations that selected the BB meal. Although not canceling was associated with selecting the BB meal, it's possible that the BB meal was not satisfactory to the customer, as selecting the BB meal was also associated with not being a repeat customer (rule 4). Similar to the recommendation above, we would recommend pushing other meal types on customers to increase retention.

… being a "transient" customer type. "Transient" customer types are not associated with being repeat customers (rule 5). Therefore, we recommend that H1 either focus on other customer types or, per the discussion below, improve the experience of these customers.

… reservations for 2 or more adults and without children or infants. Similar to above, we found that not being a repeat guest is associated with reservations with 2 or more adults (rule 6), reservations with no children (rule 7), and reservations with no babies (rule 9). Given that these variables are associated with not canceling, we suggest that a strength of H1 is their ability to market to "transient," adult, non-family customers. However, an area for improvement would be enhancing the experience, satisfaction, and increasing retention for this specific market demographic.

For H2, we opted to only examine the IsRepeatedGuest factor for conciseness. For H2, not being a repeat guest is associated with...

… reservations with the "A" Reserved Room Type. Our analysis revealed that reserving the "A" room type was associated with not being a repeat guest (rule 5). As such, we would recommend pushing other room types to customers to increase customer satisfaction and retention.

… reservations that selected the BB meal. Selecting the BB meal was also associated with not being a repeat customer (rule 4). Similar to the recommendations above, we would recommend pushing other meal types on customers to increase retention.

… being a "transient" customer type. "Transient" customer types are not associated with being repeat customers (rule 3). Therefore, we recommend that H2 focus on other customer types.

… reservations for 2 or more adults and without children or infants. Our analysis revealed that not being a repeat guest was associated with reservations with 2 or more adults (rule 6), reservations with no children (rule 7), reservations with no babies (rule 9), reservations

with 2 or more adults and no children (rule 28), reservations with 2 or more adults and no babies (rule 30), and reservations with no babies and no children (rule 32).

Conclusions

Although we initially set out to find differences between H1 and H2, our analysis reveals quite clearly that H1 and H2 are very similar with respect to factors associated with canceling reservations or being a repeat guest. As such, we provide these results as a more thorough analysis of the strengths and weaknesses of both hotels together, noting that we do not find any major distinctions between the two in this analysis.

- ## Support Vector Models

Approach

"Unlike logistic regression, which defines optimality by overall probability, SVM wants the smallest distance between data points and the decision boundary to be as large as possible" (Zhang, 2019). In developing our Support Vector Model, we first established a training set and a test set within our data.

Findings

Our Support Vector Model analysis provided numerous insights, building on the findings described in our Association Rules Mining analysis.

Initially we split data to train and test set, to check how wrong the model be on external data, given the hyperparameters and kind of data that we have. There are some things that we should always need consider that the model can overfit, that is, it has a good fit to the data that is used for training, but it performs poorly on external data. The problem is that with predictive models, we usually do not want to make predictions about training data (we have it, so there is nothing to predict), but about external data, to predict something that is unknown to you. Now we test our model on the external data or test dataset to imitate how the algorithm could potentially behave on external data. Although this may look very straightforward way of treating the data, there are some potential ethical challenges this type of automated classification may pose. Dividing train and test data differently every time leads to having issues in KSVM model every time i.e., separate set of outputs (for training and cross validation error) for KSVM model. This eventually changes the prediction accuracy, error rate and confusion matrix.

In our model for H1-Resort data, KSVM model gives out 21.55% training error based on which our model has 77.23% accuracy to predict whether the customer will cancel the reservation or not correctly and 22.77% error rate.

Also, in our model for H2 – City data, KSVM model gives out 24.03% training error based on which our model has 75.31% accuracy to predict whether the customer will cancel the reservation or not correctly and 24.69% error rate.

- ## Regression Modeling

Approach

We also built and tested several models to test predictors of interest. We built the models by setting Average Daily Revenue (ADR) or Cancelation Rate as our outcome variable. We then used multiple linear regressions to identify predictors that explained more variability in those outcome variables, as measured by the Adjusted R Squared statistic. We also examined the coefficients and *p*-values for each predictor to interpret how it influenced the outcome variable. Given the arbitrary nature of the $p < .05$ threshold (Greenland et al., 2016), we opted to select a stricter significance threshold of $p < .001$ and report only findings that reached that cut off point. "The $p < .001$ findings are more significant and we can reject the null hypothesis" (2017, Stanton)(2018, Saltz & Stanton).

Findings based on Linear modeling technique to predict ADR

Our multiple linear regression Model analysis provided numerous insights in which components are significantly predicting the increase or decrease in the ADR values.

In our model for H1-Resort data, ADR value is increasing if the reservation is getting canceled or no-show [7.37 estimate]. Also, if customer is getting required car parking space [6.569 estimate] then it is affecting positively to the ADR. The meal types of FB [11.41 estimate] and HB [12.31 estimate] are contributing highly positively to predicting ADR. In addition to this if the meal is undefined [18.38 estimate], it is also contributing highly positively to ADR. All the market segments are highly positively predicting the ADR. The family [20.57 estimate] visitor type is affecting ADR positively. If the room type is not changed [6.84 estimate] then it is likely that the ADR value would increase.

On the other hand, some predictors are decreasing ADR if there is significant increase in their values. The reservation status with higher stays during weeknights [-8.393 estimate] and weekend nights [-10.88 estimate] affects inversely to the ADR value. Somehow if the customer is mostly repeated [-8.28 estimate] then it is affecting inversely to the ADR. Furthermore, undefined distribution channel [-43.50 estimate] is affecting ADR highly inversely. single [-12.34 estimate] visitor type is affecting ADR inversely. Among seasons, spring [-36.67 estimate], summer [-19.49 estimate] and winter [-37.58 estimate] are affecting inversely to the ADR.

All the above-mentioned coefficients are significant to the model. Among all the predictors with their various categories we have Babies, SC meal type, direct distribution channel, undefined distribution channel, refundable deposit, group and transient-party customer types are not significant as their P-values are greater than the assumed alpha threshold value of 0.05.

H1: Resort Data		
Significant Predictor Variable	**Estimates of Coefficients**	**Effect on ADR (Positive or Negative)**
IsCanceled 1	7.37	Positive
RequiredCarParkingSpace	6.59	Positive

FB Meal Type	11.41	Positive
HB Meal Type	12.31	Positive
Undefined Meal Type	18.38	Positive
Family Visitor Type	20.57	Positive
Room Type Not Changed	6.84	Positive
Undefined Distribution Channel	-43.50	Negative
Staysweeknights	-8.39	Negative
Staysweekendnights	-10.88	Negative
Spring Season	-36.67	Negative
Summer Season	-19.49	Negative
Winter Season	-37.58	Negative

In our model for H2-City data, ADR value increases with canceled or no-show [1.66 estimate] reservations. Children [13.04 estimate] variable is affecting highly positively to the ADR. This also reflects into visitor type as family [15.05 estimate]. In this dataset if customers get required car parking space, then it will affect positively to increase ADR. Meal types of FB [6.5 estimate] and HB [8.51 estimate] affects positively to increase ADR. Market segments, direct [4.26 estimate] and online TA [4.93 estimate] are affecting positively to increase the ADR. Nonrefundable deposits [5.048 estimate] increases ADR positively. The hotels ADR will increase positively with customer types of Groups [6.23 estimate], Transient [5.21 estimate], Transient-Party [4.17 estimate].

On the other hand, some predictors are decreasing ADR if there is significant increase in their values. The reservation with higher stays during weeknights [-19.97 estimate] and weekend nights [-20.63 estimate] affects inversely to the ADR value. If the guest is repeated [-16.79] it affects inversely to the ADR. Whereas meal type SC [-5.62 estimate] affects inversely to the ADR. Among market segments, complementary [-69.48 estimate], Corporate [-11.72 estimate], Groups [-9.744 estimate], offline TA/TO [-6.04 estimate] are affecting inversely to the ADR. Refundable deposits [-8.97 estimate] affects ADR inversely. In addition to this, spring [-7 estimate] and Winter [-6.658 estimate] seasons affects ADR inversely.

All the above-mentioned coefficients are significant to the model. Now among all the predictors in the model babies is the only variable not significant as its p value: 0.227773 > 0.05 the assumed alpha threshold value.

H2: City Data		
Significant Predictor Variable	**Estimates of Coefficients**	**Effect on ADR (Positive or Negative)**
Children	13.04	Positive
Family	15.05	Positive
FB Meal type	6.5	Positive
HB meal Type	8.51	Positive
Direct Market segment	4.26	Positive
Nonrefundable deposits	5.048	Positive
Groups Customer Type	6.23	Positive

Transient Customer Type	5.21	Positive
Transient Party Customer Type	4.17	Positive
Complementary Market segments	-69.48	Negative
Staysweeknights	-19.97	Negative
Staysweekendnights	-20.63	Negative
Isrepeated guest	-16.79	Negative
Corporate Market Segments	-11.72	Negative
Groups Market Segments	-9.74	Negative
TA/TO Market Segments	-6.04	Negative
Refundable Deposits	-8.97	Negative
Spring Season	-7	Negative
Winter Season	-6.66	Negative

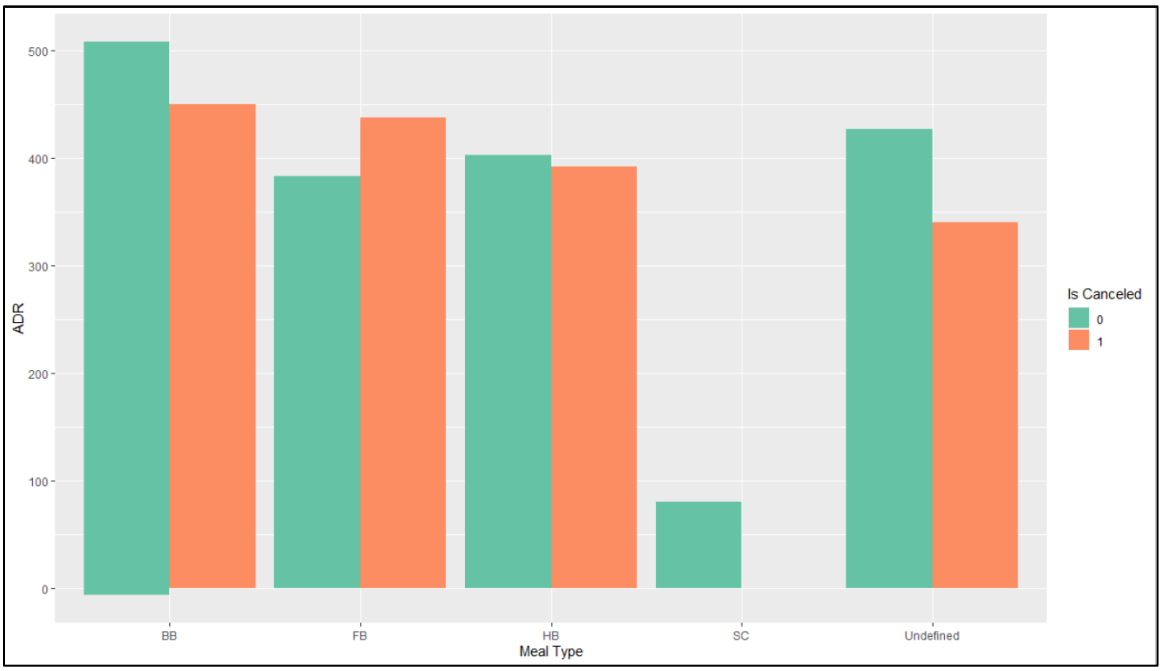
Findings based on generalized linear modeling technique to predict IsCanceled

In the Logistic Regression experiment for H1- Resort data, we can see that the coefficients have reduced the Null deviance from 46931 to Residual deviance of 31382, which suggests that the model is predicting the canceled reservations by reducing the residual errors. The normal odds are predicting whether the reservation is likely to get canceled i.e., 1. There are many significant predictors in the model. Let us understand the major effects due to predictors with highest coefficients to the Is canceled variable. For each one-unit change in Non-Refundable Deposit Type will increase the normal odds 31.43 of getting the registration canceled. Also, for each one-unit change in room type not changed data will increase the normal odds 6.96 of getting the registration canceled. Furthermore, one unit change in Transient customer type will increase the normal odds of 3.1 for the registration getting canceled. Additionally, all the one-unit change in the seasons Spring [4.043 estimate], Winter [2.5 estimate] and Summer [2 estimate] increase the reservation's cancellation by the number of coefficients.

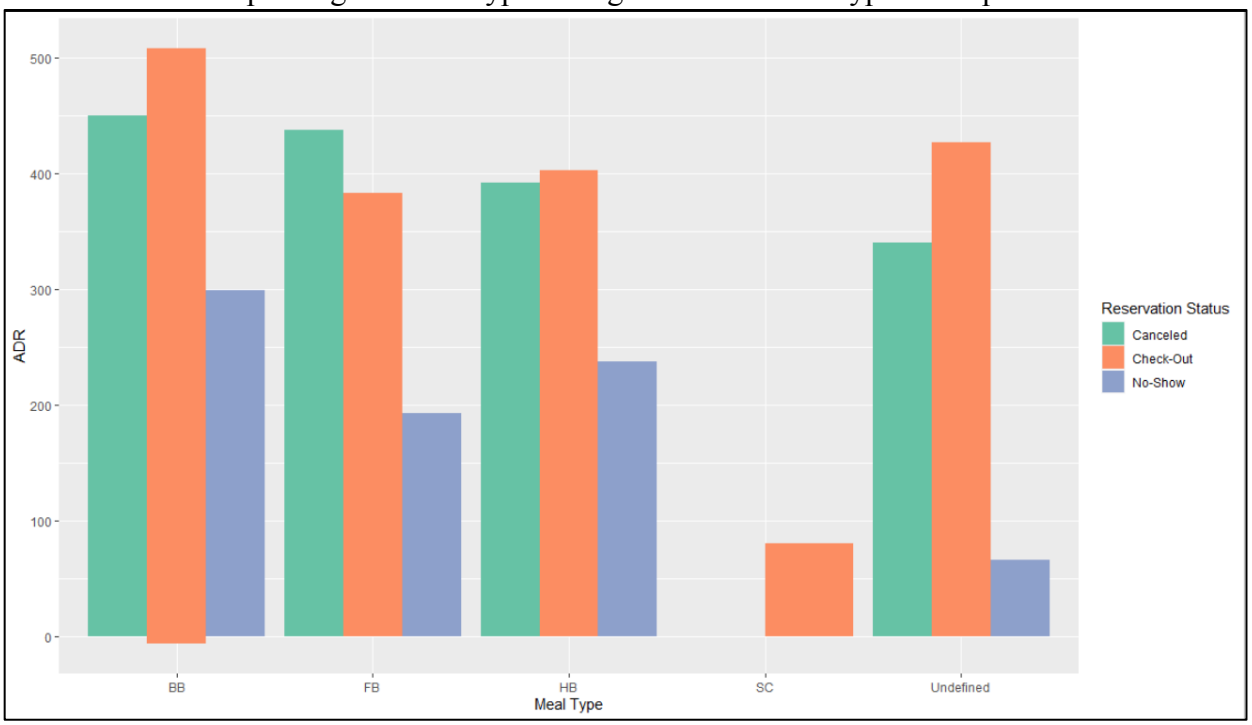
Whereas for the Logistic Regression experiment for H2- City data, we can see that the coefficients have reduced the Null deviance from 107265 to Residual deviance of 68013, which suggests that the model is predicting the canceled reservations by reducing the residual errors. The normal odds are predicting whether the reservation is likely to get canceled i.e., 1. There are many significant predictors in the model. Let us understand the major effects due to predictors with highest coefficients to the Is canceled variable. For each one-unit change in Non-Refundable Deposit Type will increase the normal odds 687.58 of getting the registration canceled. Also, for each one-unit change in meal type FB will increase the normal odds 20.82 of getting the registration canceled. Additionally, every one-unit change in Deposit Type Refundable will increase the normal odds 7.57 of getting registration canceled. If the room type is not changed then there is also increase in the normal odds 6.58 of getting registration canceled. Market segments, complementary [3.13 estimate] and online TA [2.67 estimate] are predicting that the registrations may get canceled. Finally for every one-unit change in the spring season there is normal odds of 2.26 that the reservation may get canceled.

Descriptive Analysis and Data Visualizations

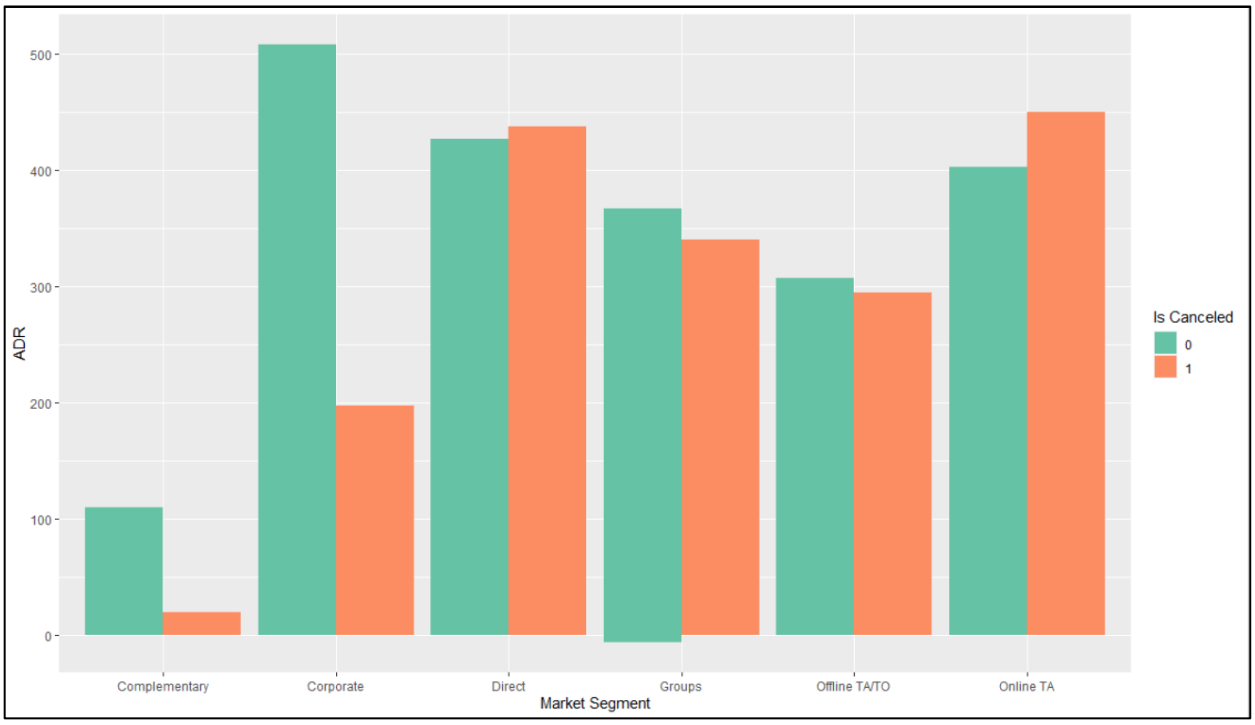
Resort Data Analysis



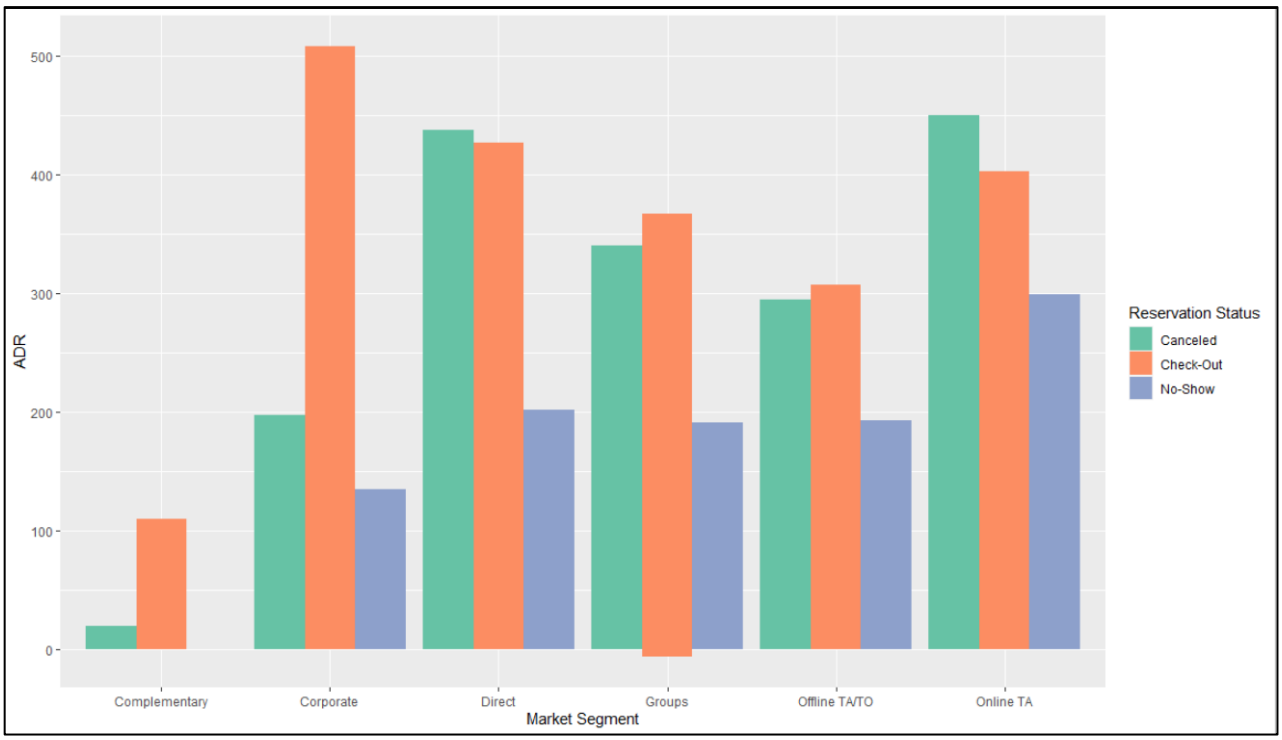
Meal Type Analysis: For meal type FB we are getting more ADR for canceled records. As the ADR values for canceled records and non-canceled records. For HB meal type are very close, we should focus on improving HB meal types along with other meal types to improve the AD.



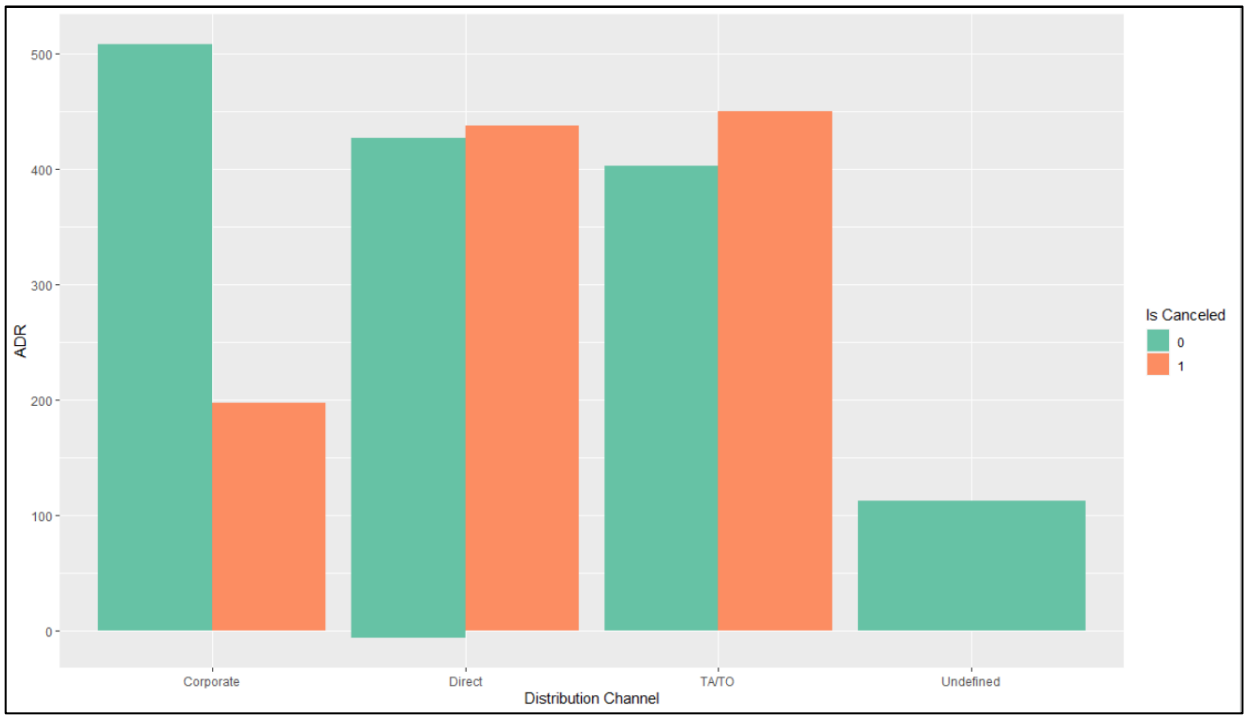
Meal Type Analysis: With BB as meal type there are more ADR values with no show and canceled records



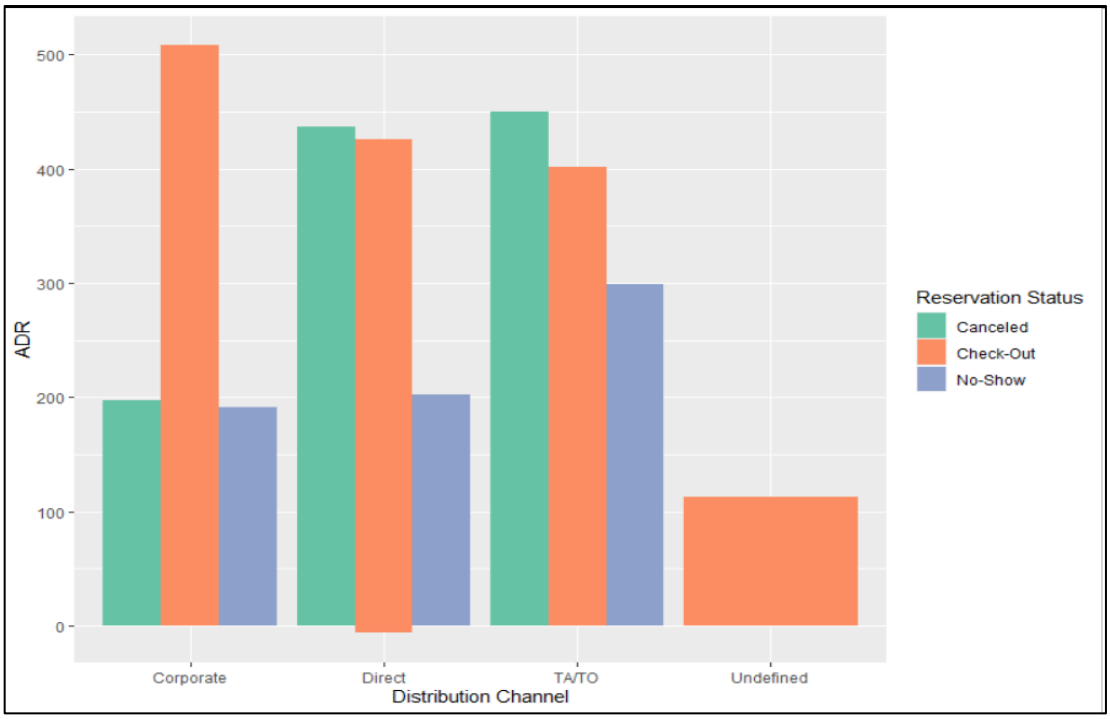
Market Segment Analysis: With Market segment as online TA and Direct we have higher ADR values for canceled records. Corporate market segment people are contributing with higher ADR values for more checkouts than other marker segments.



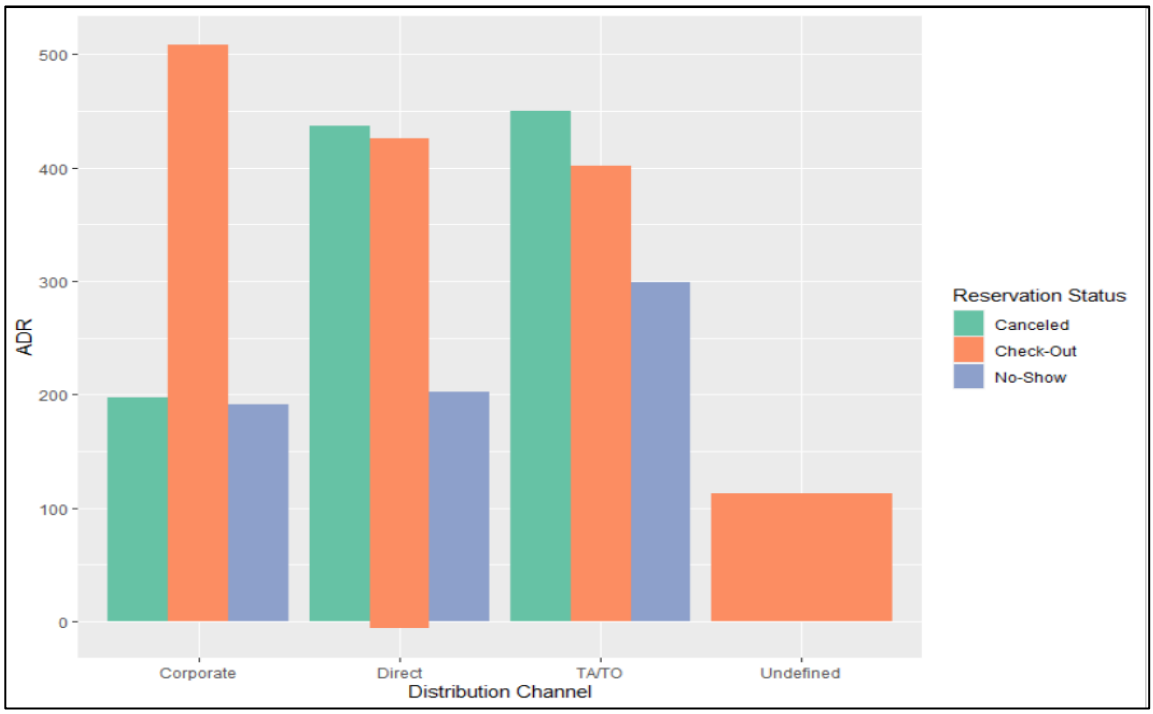
Market Segment Analysis: People are not showing up for their reservations with online TA as market segment.



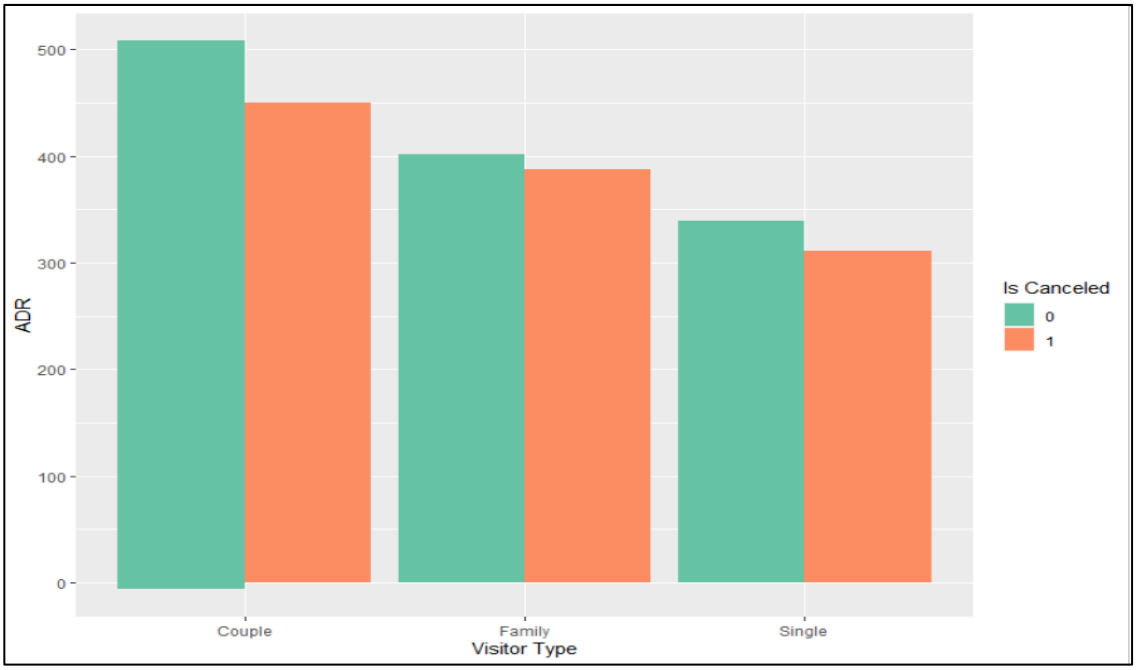
Distribution Channel Analysis: The undefined distribution channel is contributing in ADR with all customers without cancellations. Corporate distribution channel is also contributing in high ADR values for customers without cancellations. We are getting cancellations where distribution channel is Direct and TA/TO.



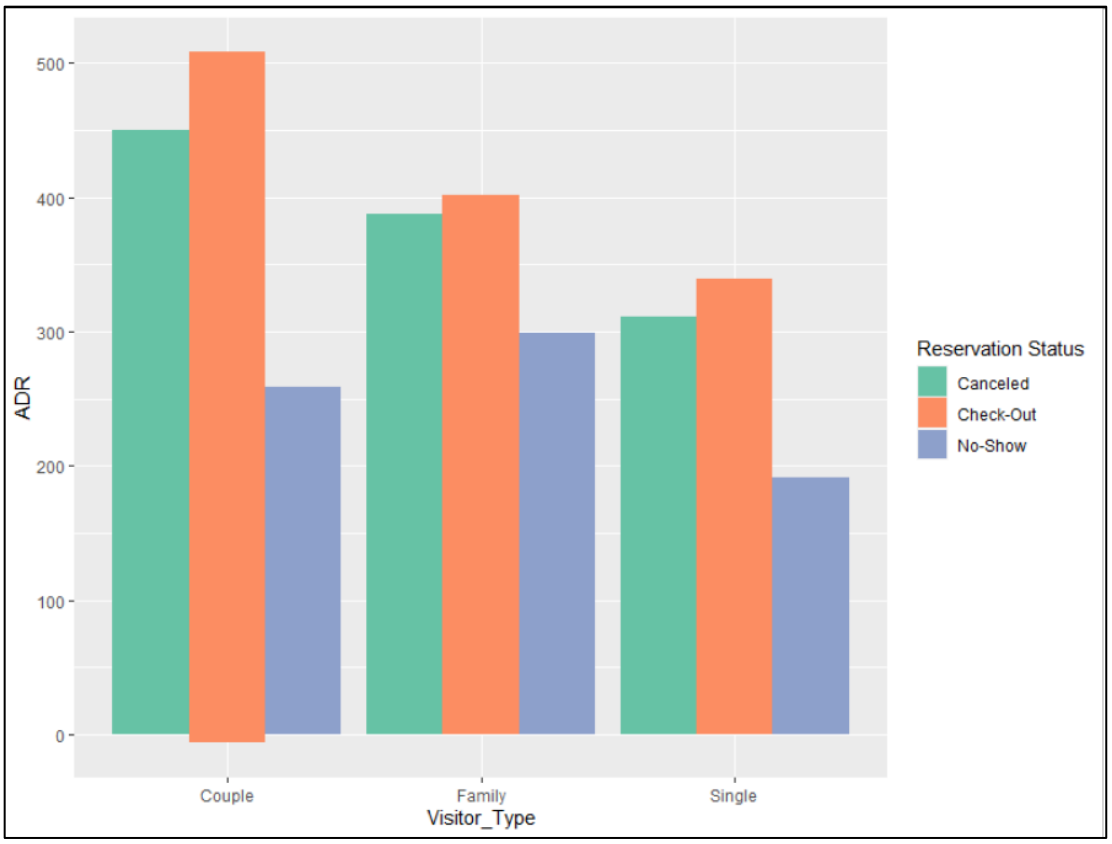
Distribution Channel Analysis: ADR values are high for customers not showing up where Distribution channel is TA/TO.



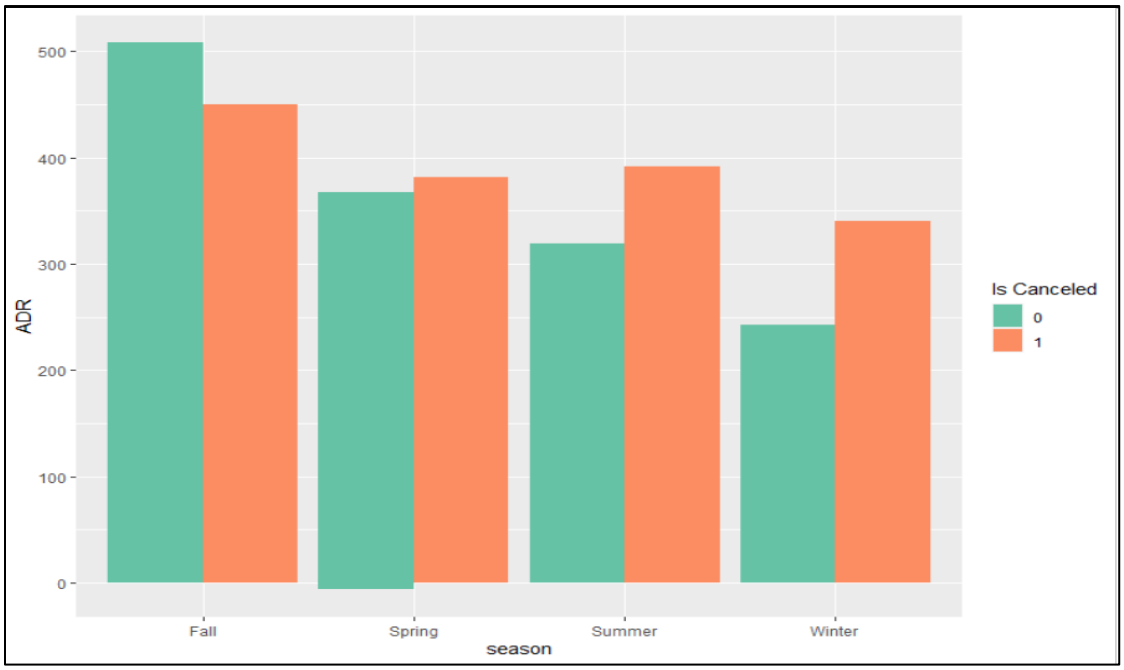
Distribution Channel Analysis: ADR values are high for customers not showing up where Distribution channel is TA/TO.



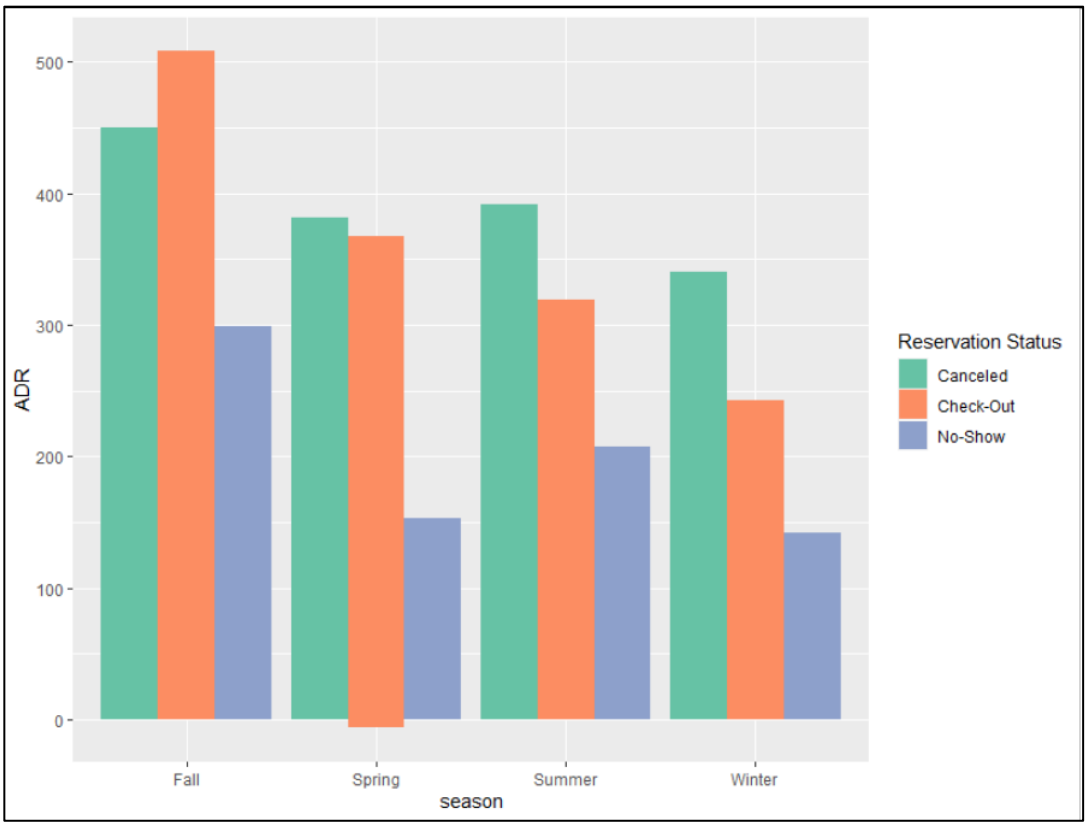
Visitor Type Analysis: We are getting high ADR without cancellations where Visitor type is Couple.



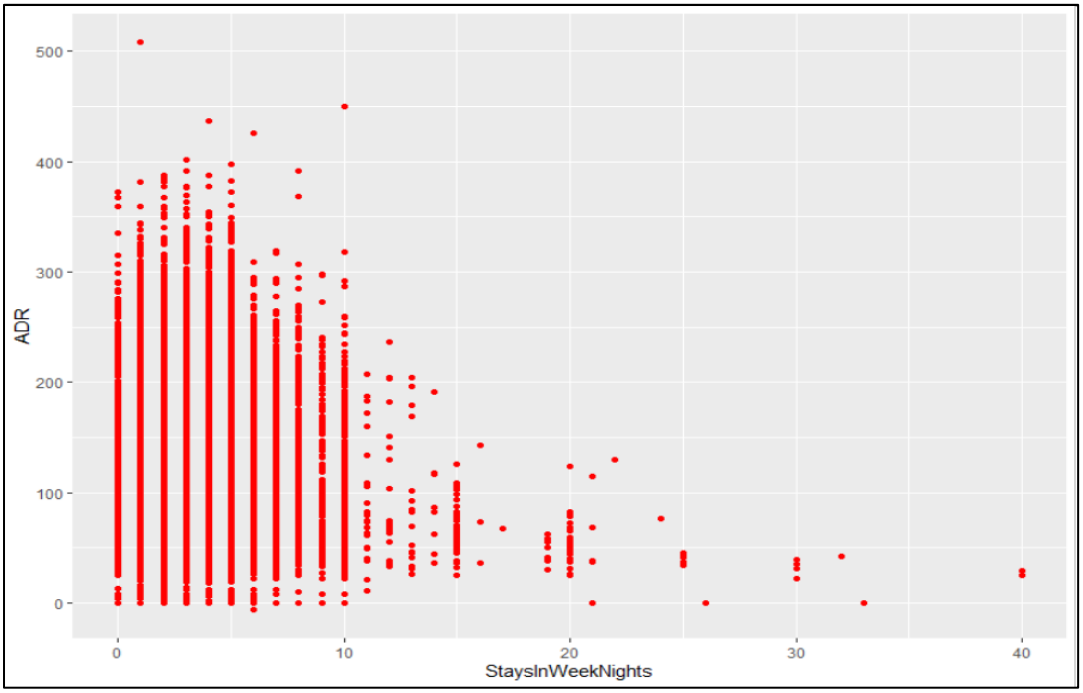
Visitor Type Analysis: For visitor type as family, we are getting high ADR with family not showing up.

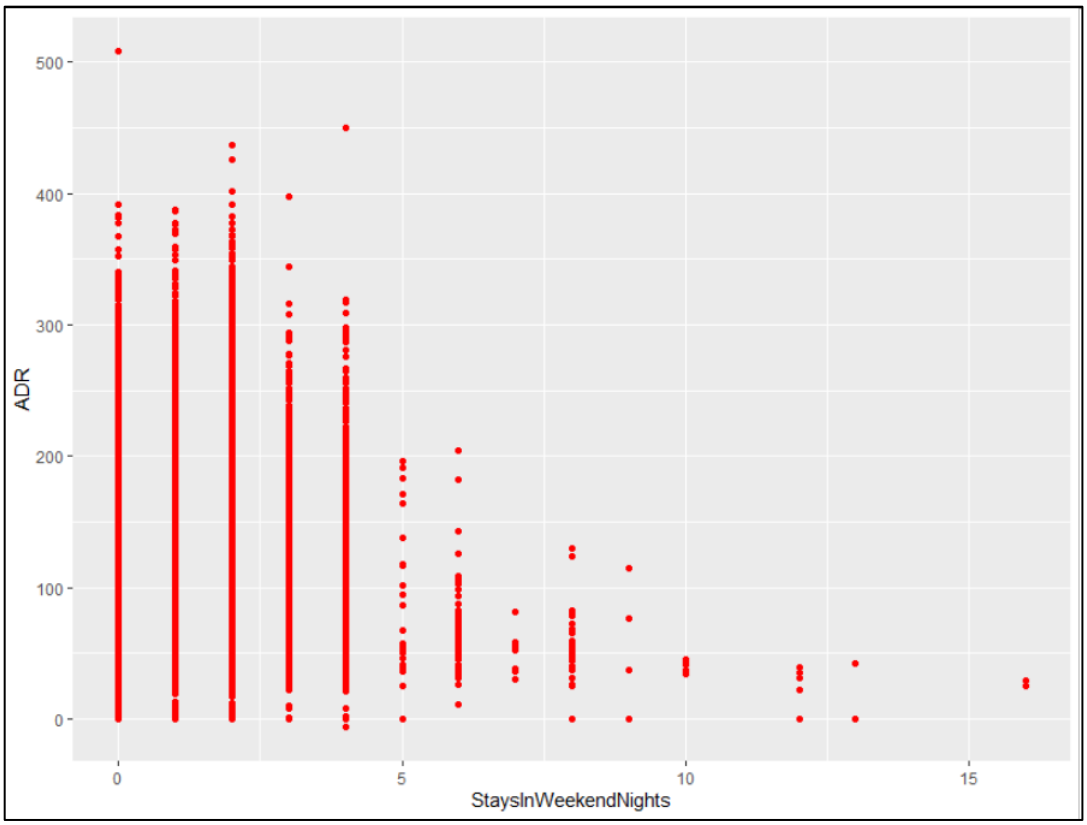


Season Analysis: In fall season the ADR values are highest and no cancellations. Whereas, for all other categories, the ADR values are highest with cancellations.

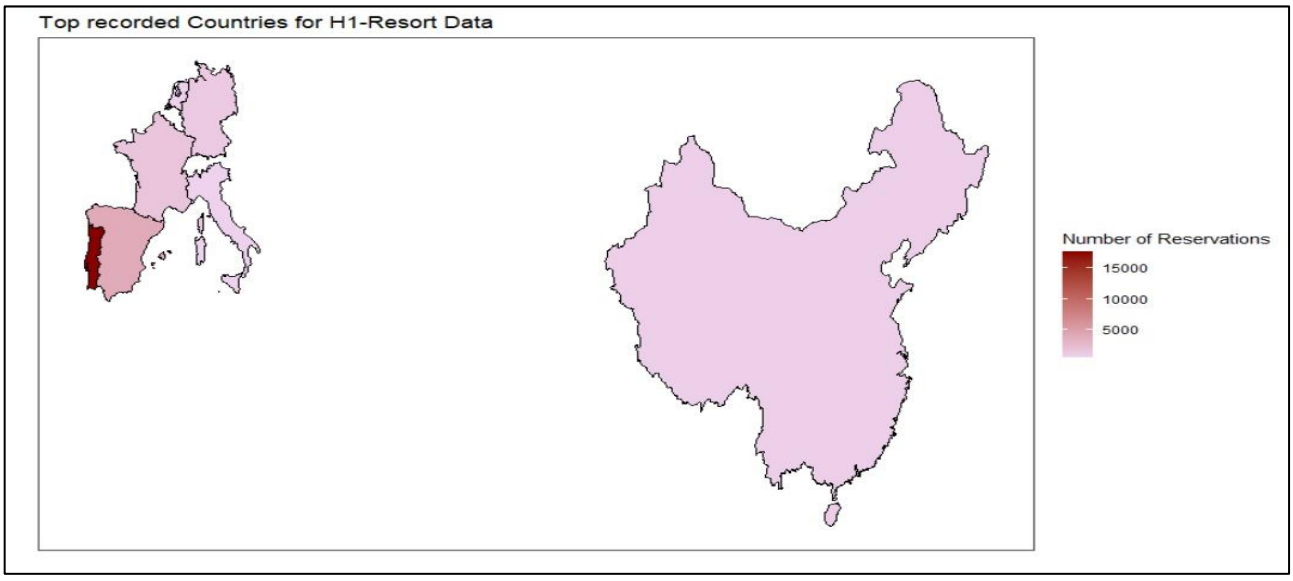


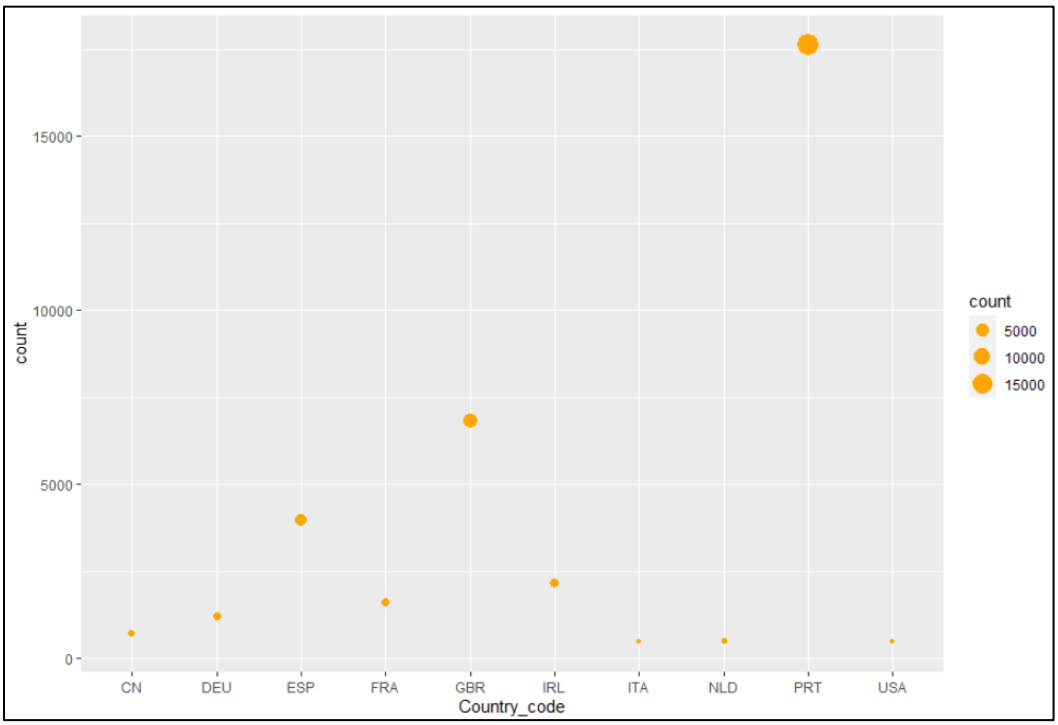
Season Analysis: ADR values are high because customers are not showing up in Fall season.





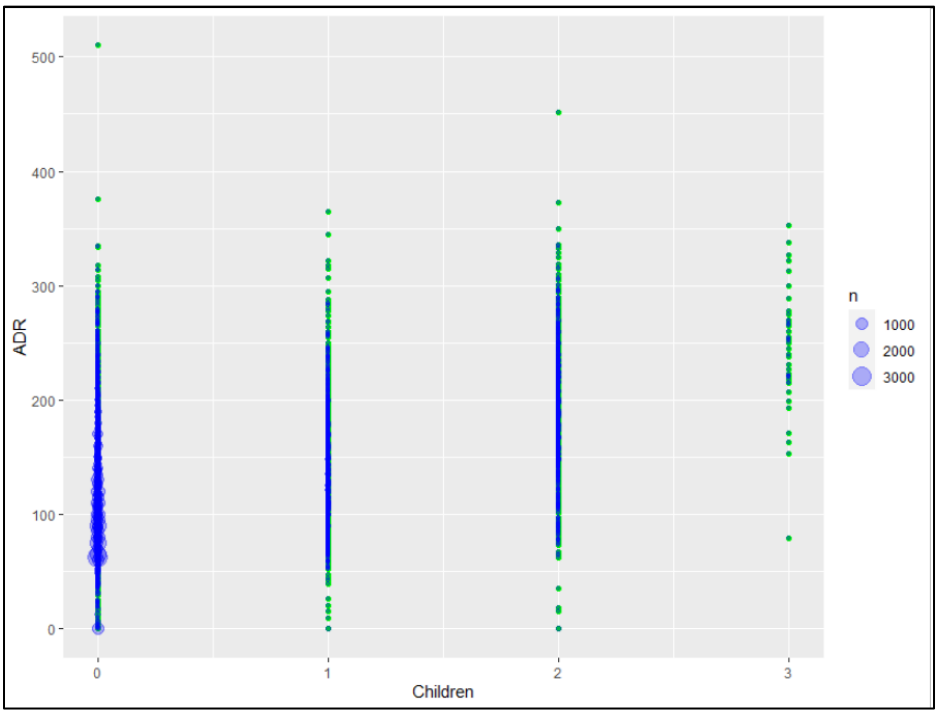
Stays Analysis: As we can see from the scatter plots above, with increase in the stay in week-nights and the weekend nights, the ADR value is decreasing gradually.



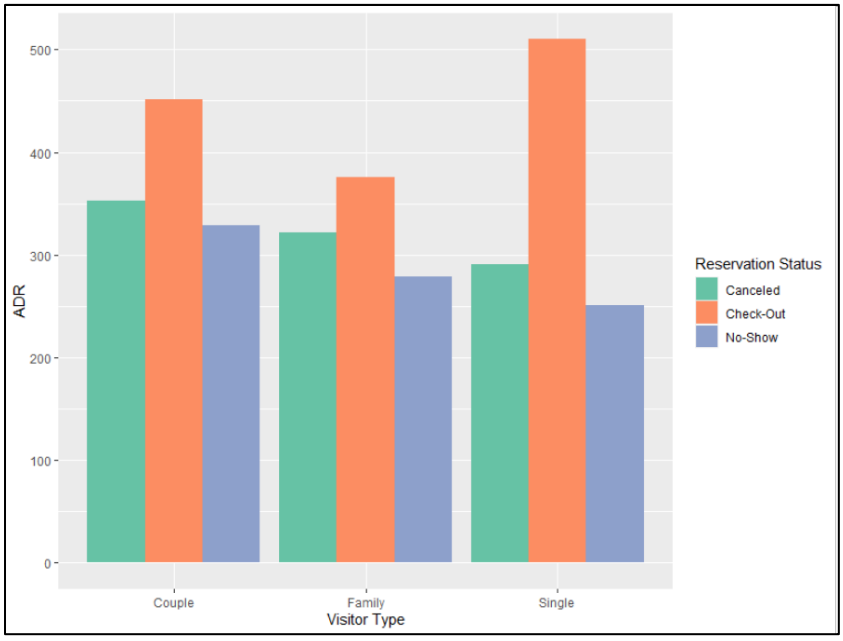


Countries Analysis: Among all the countries, PRT has maximum records 17622 then comes GBR with 6813 records, ESP with 3956 records, IRL with 2166 records, FRA with 1610 records and DEU with 1203 records. This tells us that PRT is contributing highest to get the ADR values (2020, *ISO 3166 - Country Codes*).

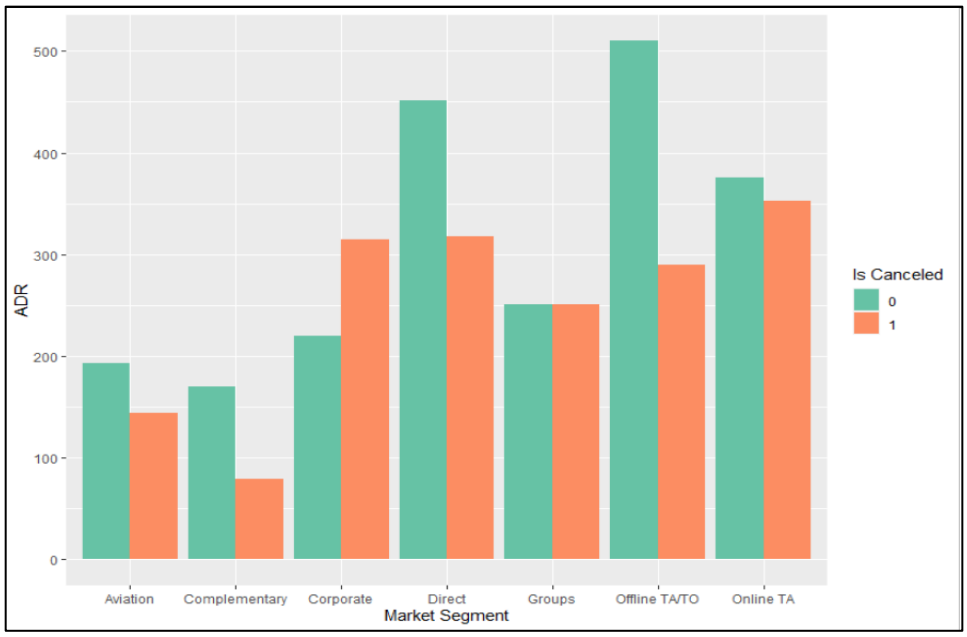
City Data Analysis



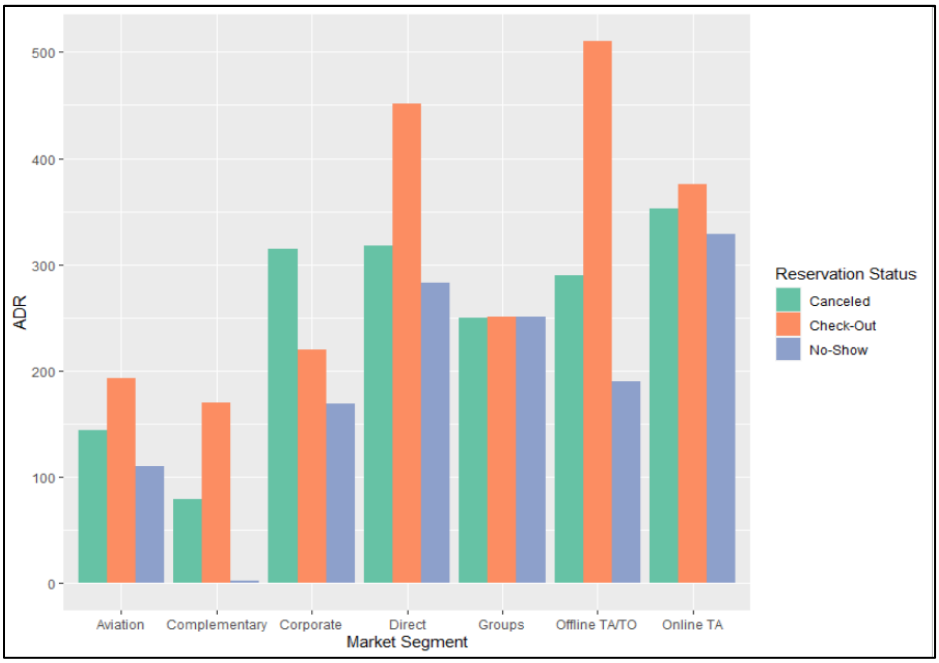
Children Analysis: when children =0, ADR cluster around 0-300 and particular around 60. In the above scatter plot, we can see the increase in ADR with increase in children numbers(shown by the green dots). Also, the blue points represent the concentrations of children counts to the ADR values.



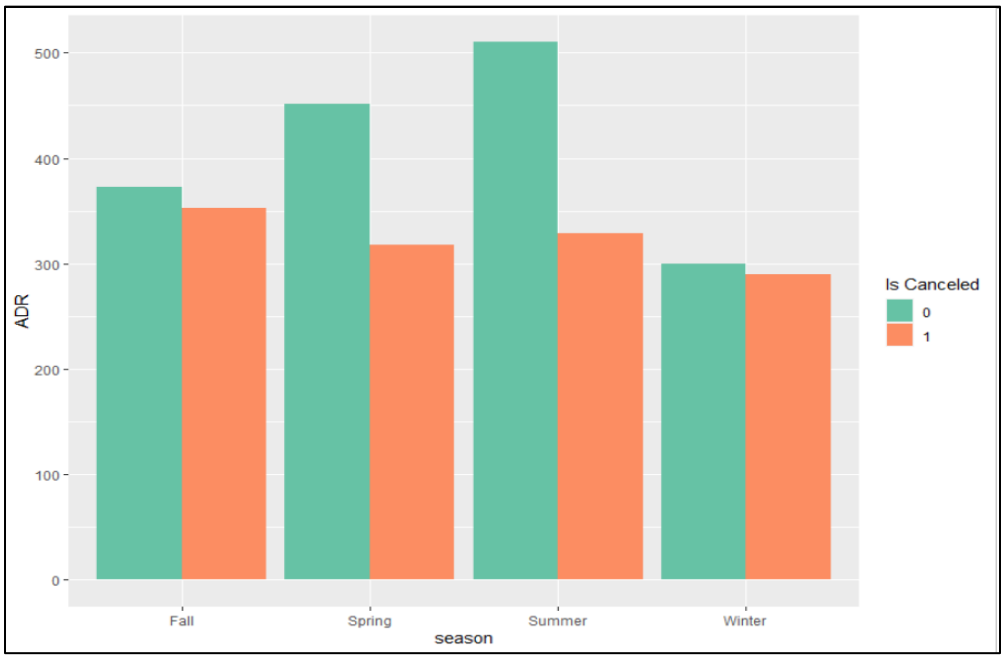
Visitor type Analysis: Couples are not showing up and giving higher ADR values overall. Single customers are contributing more in giving higher ADR values without cancellations. Couples are contributing in giving higher ADR values in general with cancellations.



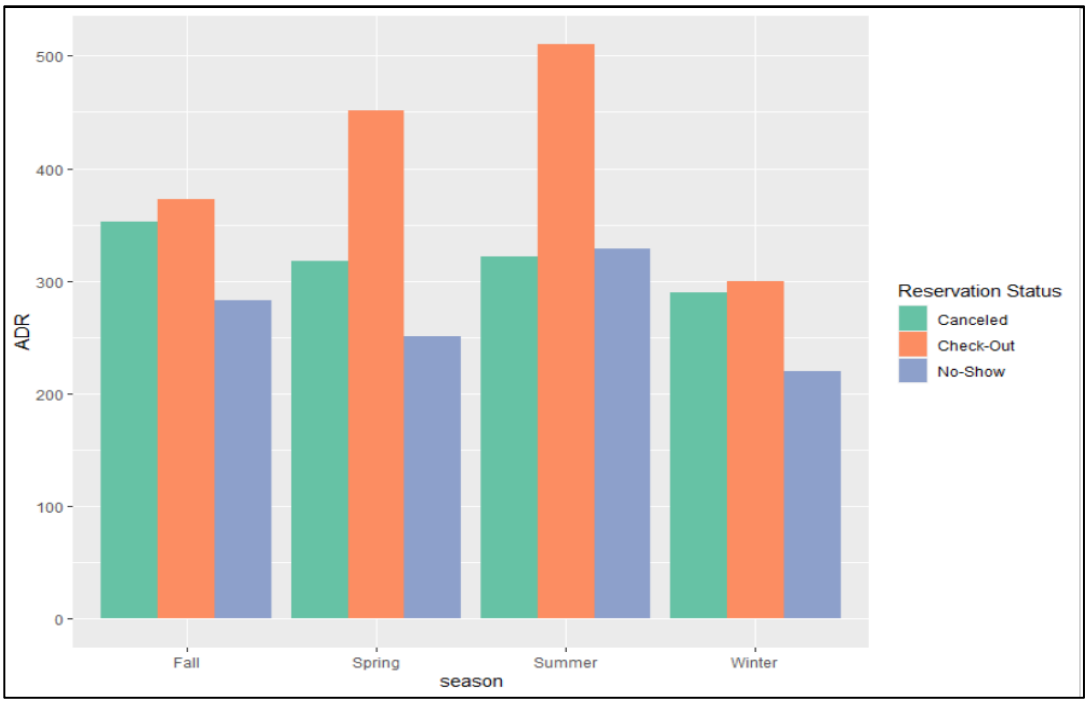
Market Segment Analysis: ADR values are higher for Direct and offline TA/TO market segment without cancellations. ADR values are highest for online TA as market segment with cancellations. Also, for corporate Market segment ADR values are affected because more records are canceled.



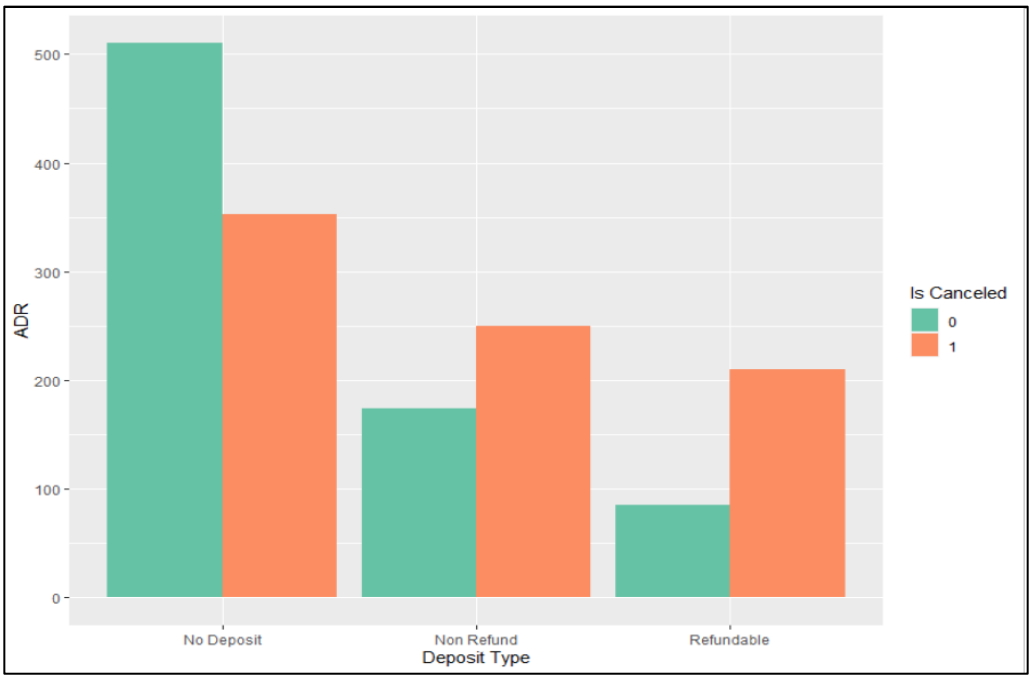
Market Segment Analysis: ADR values are same for all three categories for Groups market segment. Also, for online TA market category we have more ADR values overall where customers are no showing up.



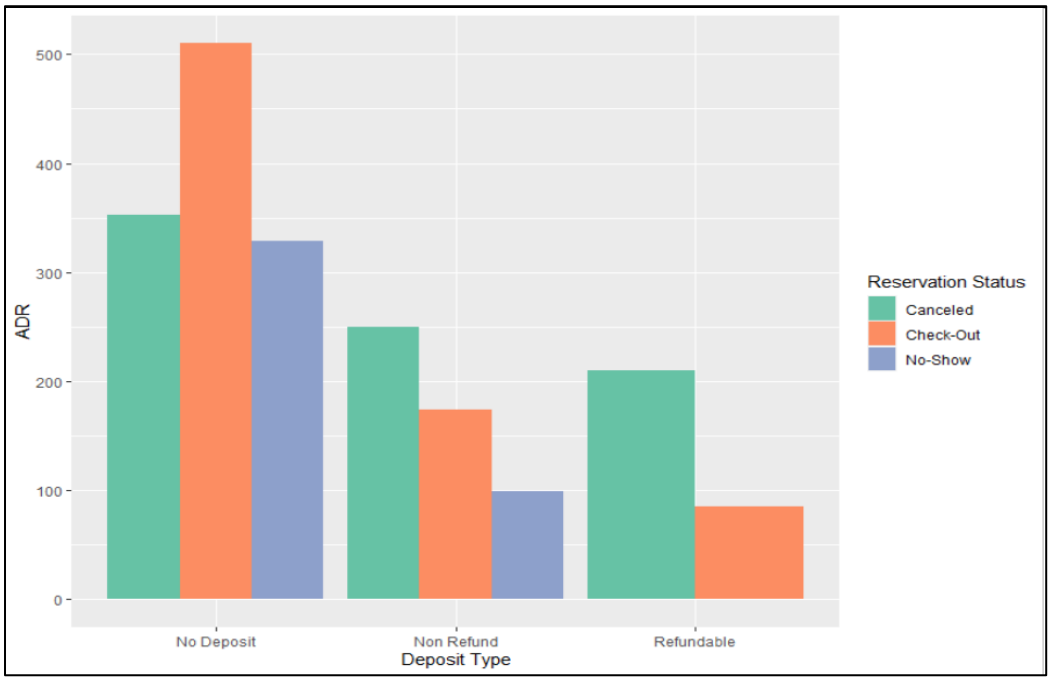
Season Analysis: In summers we are getting highest ADR without cancellations. Also in Fall, we are getting higher ADR values with cancellations in general.



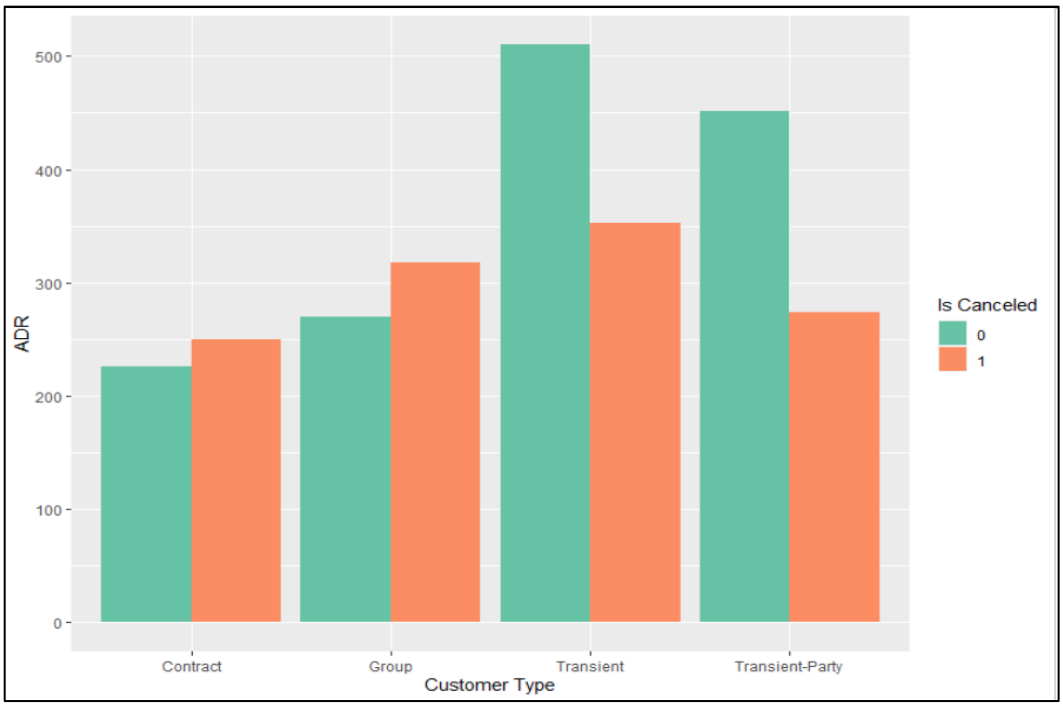
Season Analysis: Customers which are not showing up in general are also contributing to higher ADR values.



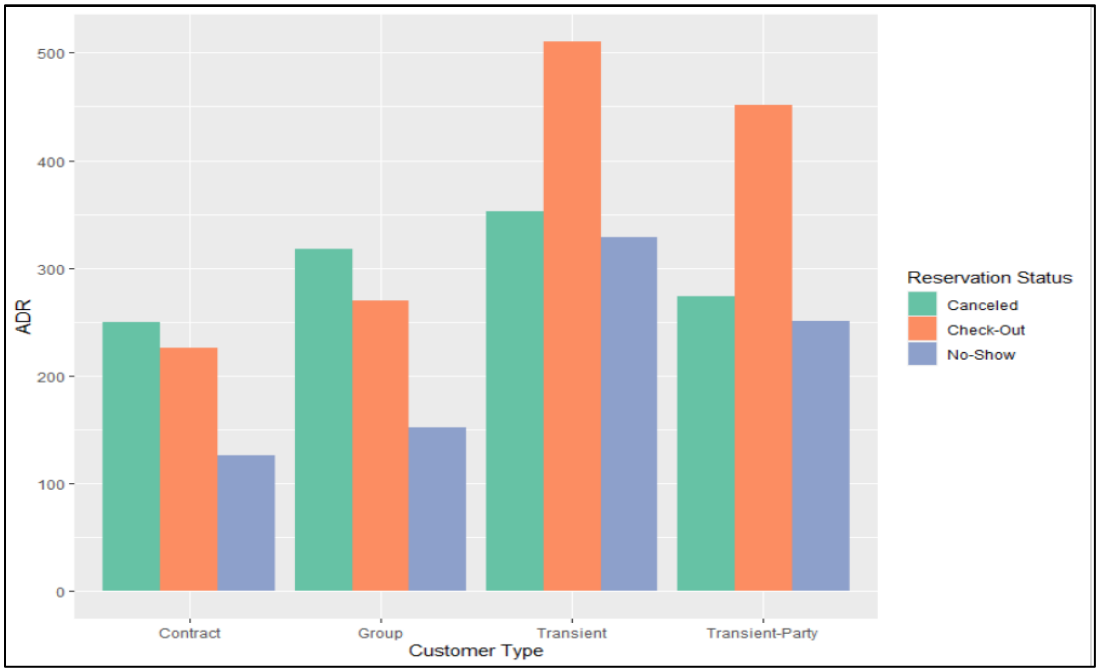
Deposit Type Analysis: ADR values are highest and no cancellations where the reservation is done with no deposit. Whether deposit type is non refundable or refundable we are getting high ADR values for cancellations.



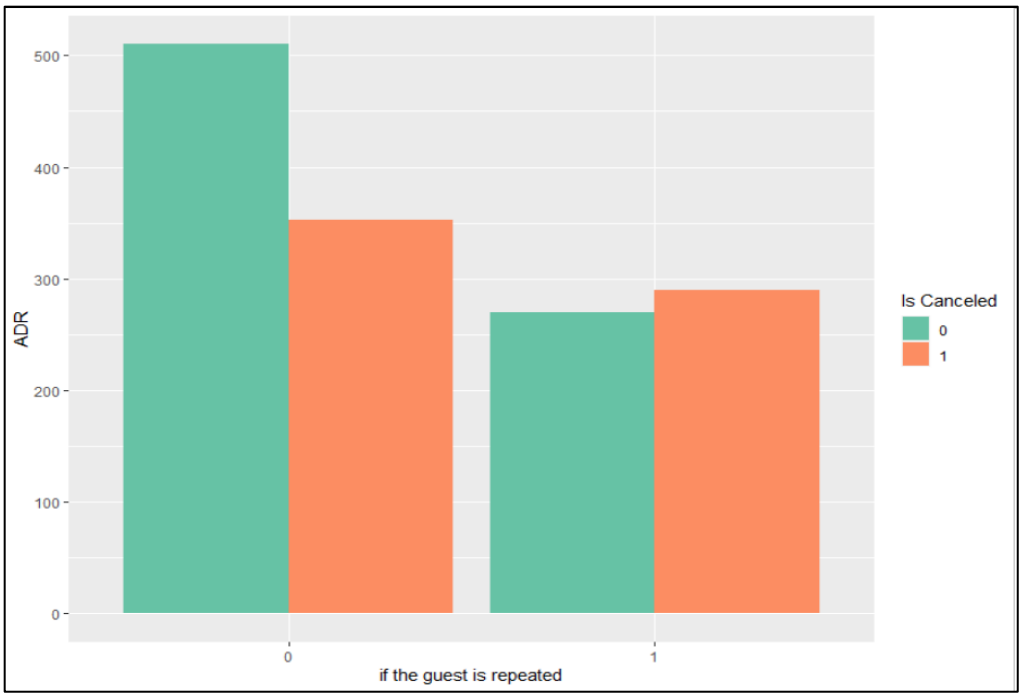
Deposit Type Analysis: ADR values are affected with customers not showing up when there is no deposit. Also, ADR values are affected with more cancellations and there is no case where customers are not showing up.

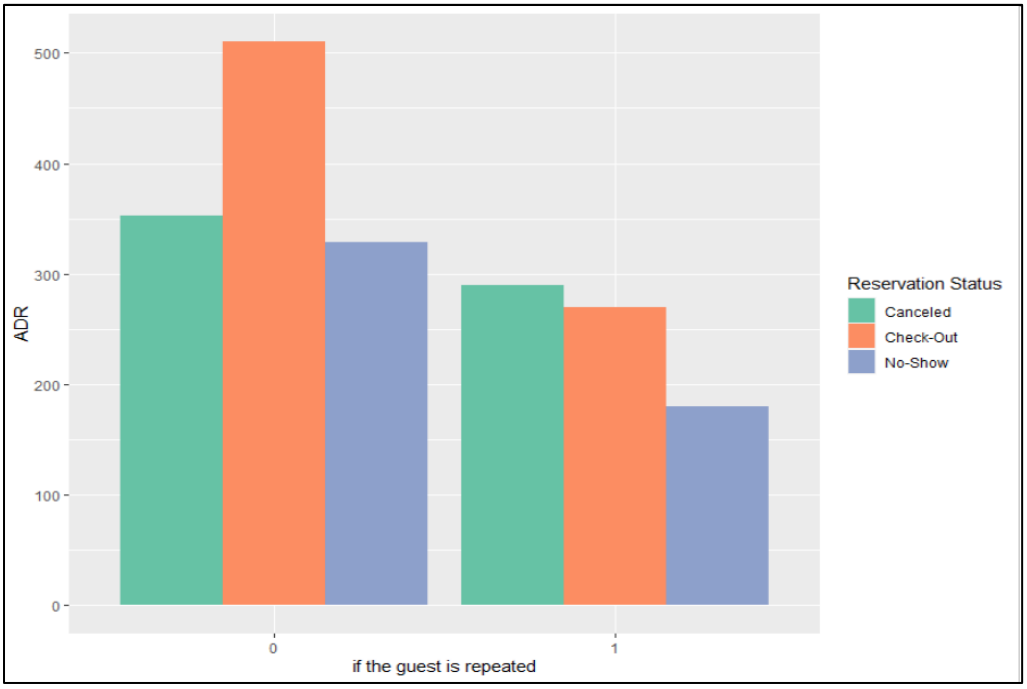


Customer Type Analysis: Transient and transient party are giving out high ADR values without cancellations. Also, we are getting high cancellations overall where transient is category.

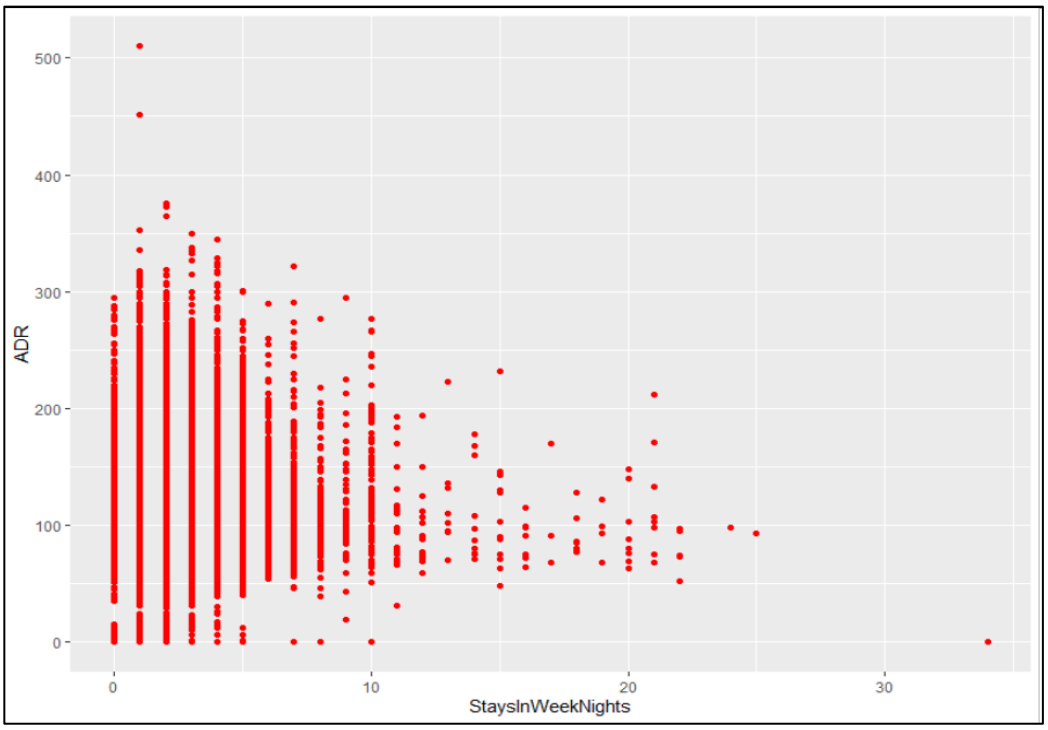


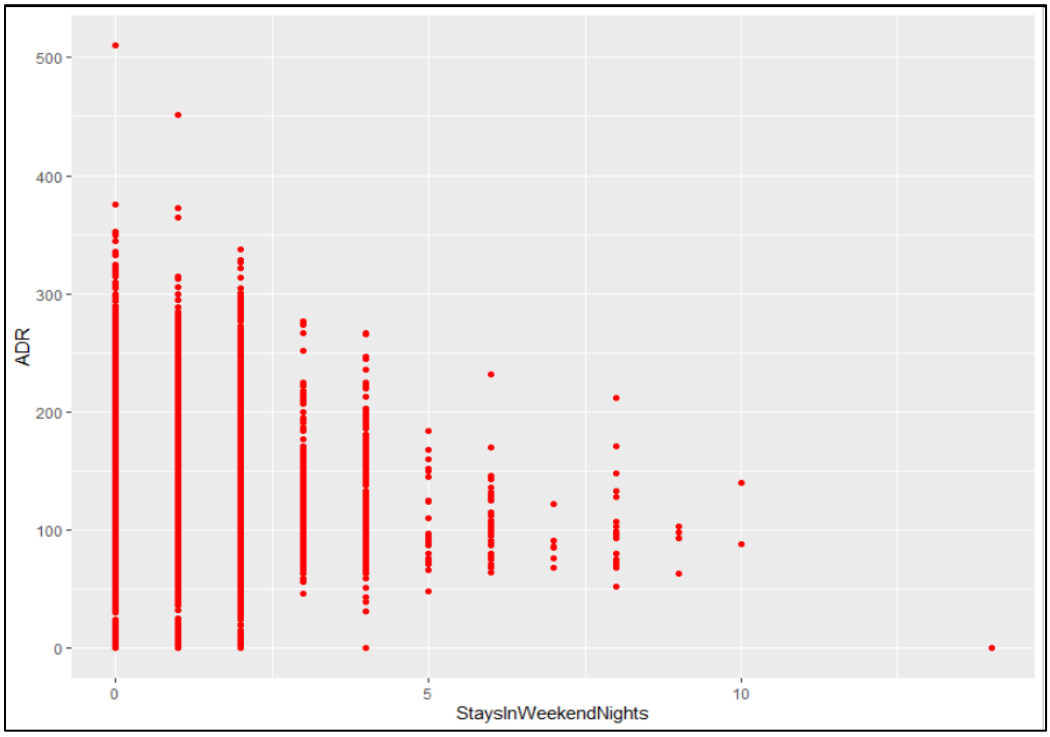
Customer Type Analysis: ADR values are getting affected with people not showing up in transient category.



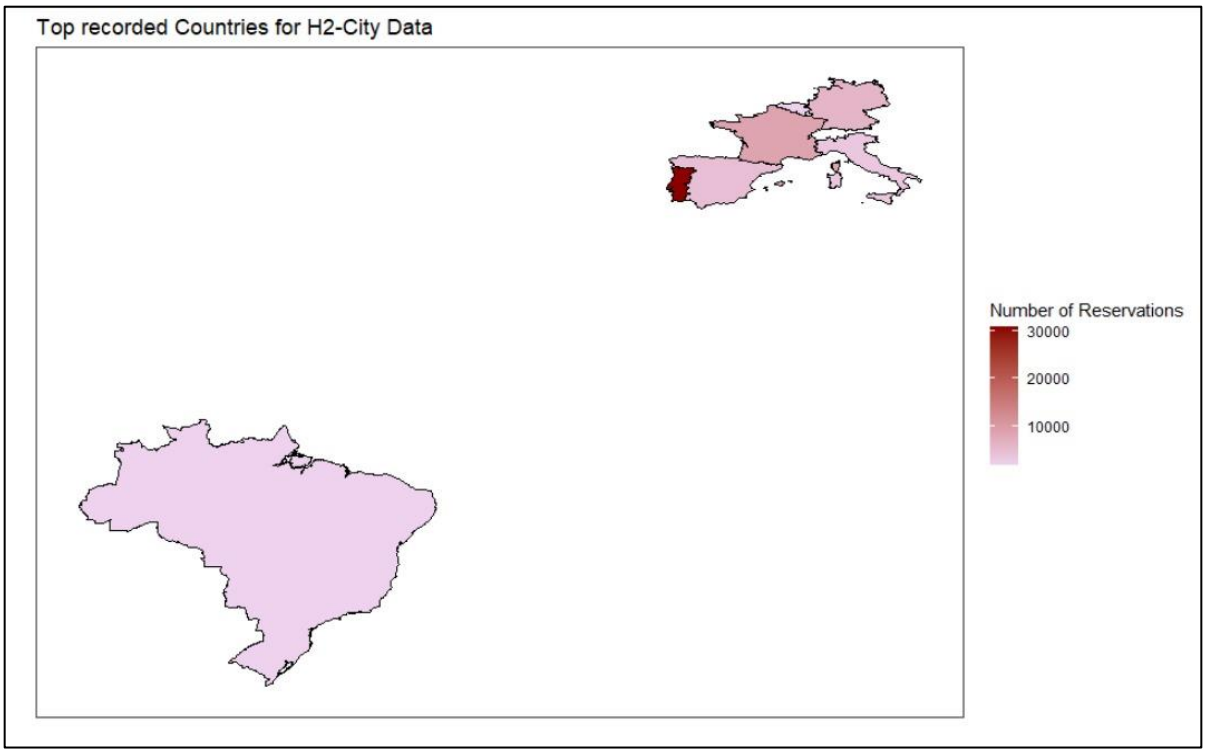


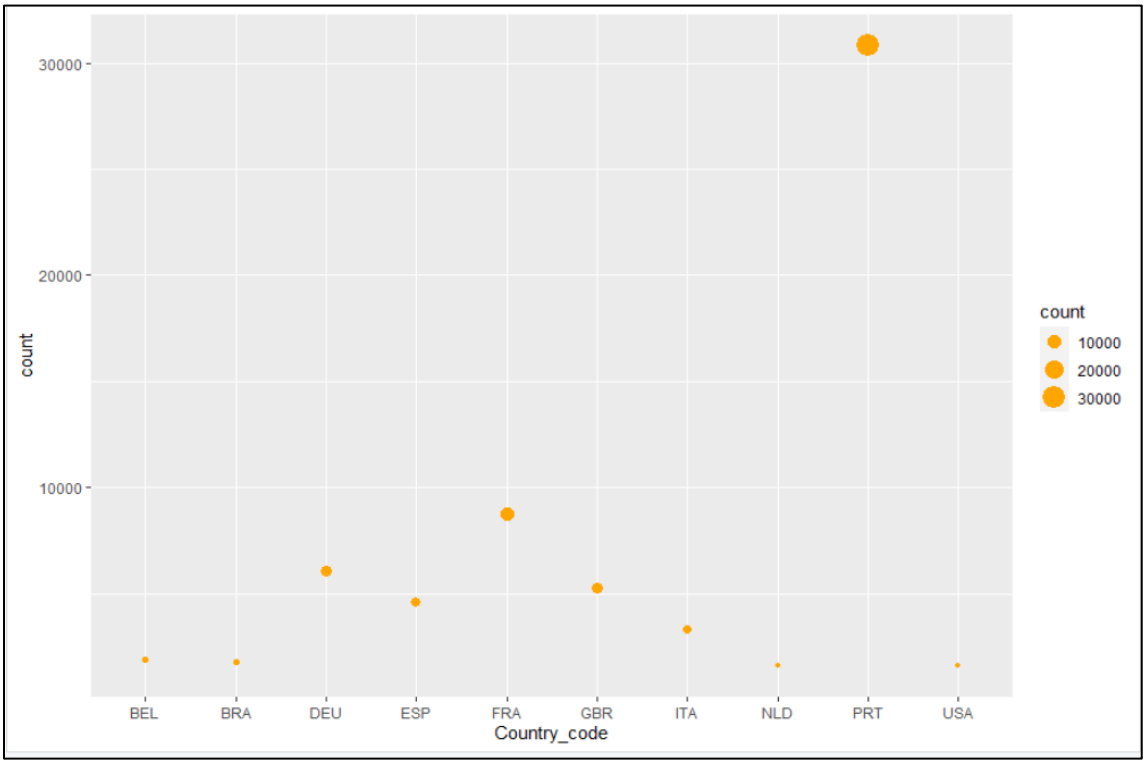
Is Repeated Analysis: As we can see from the plots above, if the guest is repeated it is most likely that the ADR value will decrease because of more cancellations or no-show.





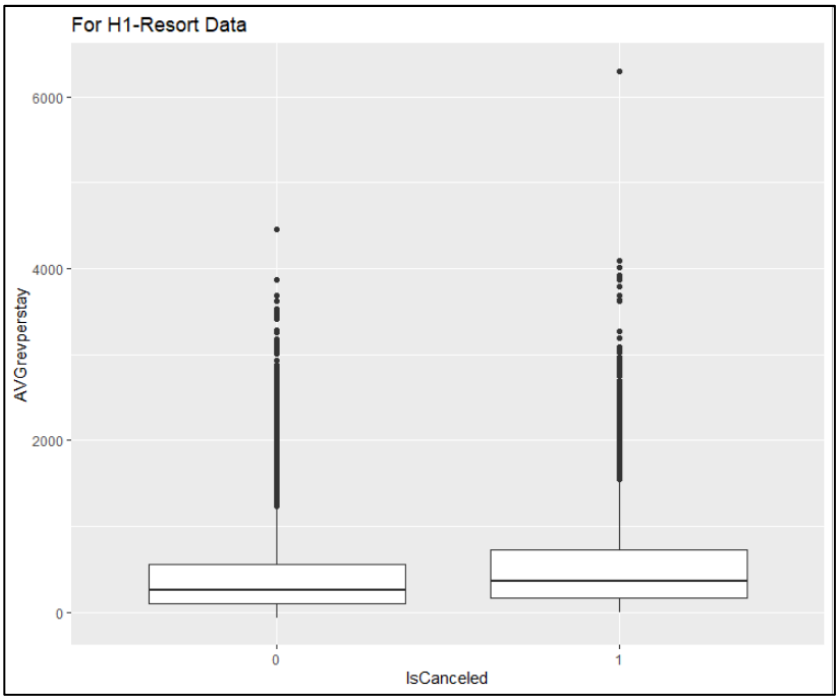
Stays in weekend Analysis: As we can see from the scatter plots above, with increase in the stay in weeknights and the weekend nights, the ADR value is decreasing gradually.

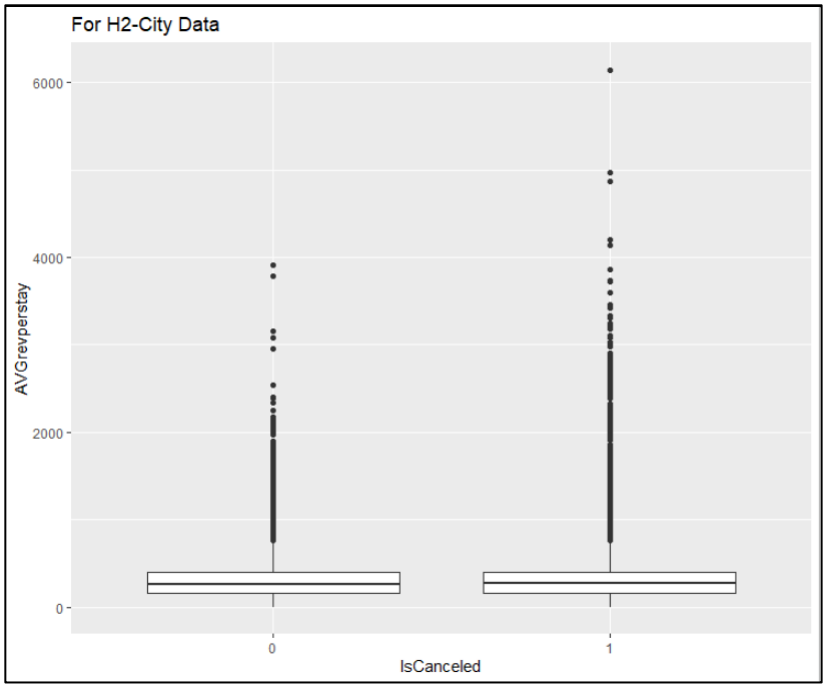




Country Analysis: Among all the countries, PRT has maximum records 30813 then comes GBR with 6813 records, FRA with 8766 records, DEU with 6068 records, GBR with 5292 records, ESP with 4590 records and ITA 3293 records. This tells us that PRT is contributing highest to get the ADR values (2020, *ISO 3166 - Country Codes*).

Average Revenue per Stay Analysis for Resort and City Data





As we can see from the boxplots above for both H1-Resort and H2-City data, the average revenue per stay shows no clear distinction between reservation is canceled or not. This is very important parameter to show the combined effect of ADR and stay data on cancellations. This suggests that average revenue per stay is not an informative variable for the available data, in other words the business model needs some changes to get the better average revenue.

Results

To inform our results, we take together insights gathered from each stage of the data analytic process. Here, we emphasize conclusions drawn from our modeling, as they proved most informative.

1) Results obtained from Association Rules Mining Techniques

The association rules mining results suggest that both hotels are performing similarly with respect to our categorical variables: number of children, infants, and adults, room type (namely, "A"), meal type (namely, "BB"), and customer type (namely, "transient"). Here, we highlight how, for H1, many of the same variables were associated with IsCanceled as IsRepeatedGuest. We conceptualized factors associated with IsCanceled=0, or, in other words, a customer choosing to not cancel a reservation, as potential strengths. Conversely, things associated with IsRepeatedGuest=0, or, in other words, things associated with not returning, as potential areas to improve. Finding that many of these associations were the same, we interpret this to mean that in the cases of both the hotel and the resort, the marketing team is very skilled at capturing their desired market demographic, possibly through highlighting the amenities that are attractive to these clients (attractive enough to ensure a reservation). However, the experience of those customers may be quite negative, as they tend not to return. Therefore, we recommend maintaining a strategy of marketing to adult customers without children but improving customer experience and retention.

2) Results obtained from Linear Modelling Techniques

The below table shows us numerous factors are significantly affecting the hotels ADR value at Resort and City locations based on the results obtained from the linear modelling technique. Despite the ADR has some outliers in it, our model is considering some of them to get the unbiased results. The ADR value is getting affected in the same manner whether the reservation is canceled or not. The strengths and weaknesses reveals which factors are giving positive coefficients and negative coefficients respectively.

	Strengths	Weaknesses

Sr. No	*H1-Resort Data*	
1	Family, Children	**Spring, Summer, Winter season**
2	Meal Type: HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner) Undefined – no meal package	**Stays during week and weekend nights**
3	**Required Car Parking Space**	Undefined distribution channel
4	**No change in the room type**	
	H2-City Data	
1	Family, Children	**Complementary market segment** is reducing the ADR value significantly. Other Market Segments: **Corporate, group and TA/TO market segments** are affecting ADR negatively. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators".
2	**Meal Type**: **HB** – Half board (breakfast and one other meal – usually dinner); **FB** – Full board (breakfast, lunch and dinner)	**Stays during week and weekend nights** are not satisfactory
3	**Customer Type**: **Group** – when the booking is associated to a group; **Transient** – when the booking is not part of a group or contract, and is not associated to other transient booking;	Repeated guests are somehow not good to the hotel's ADR value.
4	Deposit Type: **Non-Refund – a deposit** was made in the value of the total stay cost	We are seeing lower ADR values during **Spring and Winter seasons.**
5	**Required Car Parking Space**	

3) Results obtained from Generalized Linear Modelling Techniques

The below table shows us numerous factors are significantly affecting hotel's reservations cancellation status (i.e., whether they are canceled – IsCanceled: 1 or whether they are not canceled – IsCanceled: 0) at Resort and City locations based on the results obtained from the Generalized Linear Modelling technique. Even though the reservation is canceled, it does not mean that it s currently affecting badly to ADR.

Sr. No.	Significant Components, can cause Reservations to either get canceled or no-show
H1-Resort Data	
1	DepositType: Non Refundable
2	**No change in the room type**
3	**Seasons: Spring, Winter and Summer**
4	Customer Type Transient and Group
5	Meal Type FB
H2-City Data	
1	DepositType: Non-Refundable, Refundable
2	Market Segment: Complementary, Online Travel Agents, Group, Direct, Offline Travel Agents
3	**Season: Spring, Winter**
4	**If the guest is repeated**
5	Meal Type: FB – Full board (breakfast, lunch and dinner)
6	**No change in the room type**

4) ***Results obtained from Data Mining Technique Support Vector Machine (SVM)***

The data mining using SVM is done on the whole data because the business would be interested in knowing how all these columns are affecting the business model. The SVM is designed with 60% of data and able to predict IsCanceled Column by testing 40% of data. Below are the results:

Predicting IsCanceled Column	**Accuracy (%)**	**Error Rate (%)**
H1- Resort Data	77.23	22.77
H2- City Data	75.31	24.69

Conclusion and Recommendations

Following recommendations are based on, how to improve the average daily rate i.e., average revenue by increasing the number of reservations along with customer satisfactions and not with number of cancellations:

1) Services are lacking when customers are staying during week and weekend nights at both locations, make sure to increase employee night shifts or some necessary adjustments in the available services which are essential during night.

2) Ensuring that the following meal packages: FB (Full Board), HB (Half Board) and Undefined (Meal is not in the reservation package) are available in all seasons.

3) Ensure that the decorum is maintained for the Family with Children and Babies.

4) Try to give some offers to single and couples during the festival and vacation seasons.

5) Reduce the efforts on complementary services at both locations.

6) Our analyses highlighted potential weaknesses in customer dissatisfaction (see linear modeling and logistic regression). Further, our exploratory association rules mining analysis suggested that customer experience and retention could be improved. We recommend obtaining customer reviews for the travel agents or tour operators and based on that some improvements can be done in better marketing and advertising strategies. This technique can be applied to bring variations and improvements in the Full Board meal type category.

7) We recommend ensuring that each hotel operates with enough car parking spaces to provide for group reservations, as our analyses reveal positive associations between these factors.

8) Also, if a reservation is from corporate, make sure to allot room types and amenities which could meet specific corporate needs. With this recommendation, we aim to improve customized experiences and customer retention.

9) The city hotel is performing well in the Summer and Fall seasons. However, the resort hotel struggles with respect to ADR in the Spring, Summer, and Winter seasons.

- As such, we would recommend expanding the city hotels. Further, we would recommend investing any additional resources into improving current offerings at the resort hotel.

References

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31,* 337–350. https://doi.org/10.1007/s10654-016-0149-3

Buteikis, A. (n.d.). *Practical Econometrics and Data Science*. 4.5 Multicollinearity. http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html.

ISO 3166 - Country Codes. ISO. (2020, March 25). https://www.iso.org/iso-3166-country-codes.html.

Zhang, Z. (2019, August 7). *Support Vector Machine Explained*. Medium. https://towardsdatascience.com/support-vector-machine-explained-8bfef2f17e71.

Stanton, J. M. (2017). *Reasoning with data: an introduction to traditional and Bayesian statistics using R*. The Guilford Press.

Saltz, J. S., & Stanton, J. M. (2018). *An introduction to data science*. SAGE Publications, Inc.