IST687 Final

Hotel Analysis

Group 1

Shubham Kumar, Oluwatosin Oyediran,
Danni Pan, Aidan Surowiec

Table Of Contents

Abstract

Analysis of two hotels one located in the city and the other a resort hotel. We analyze the performance of these two hotels comparatively and how each of them perform over time. On a high level, first we are trying to compare revenues and cancellations for the hotels. Additionally, we have developed models to predict the revenue and cancellations.

Data

To read the file we used the setwd command to point to the path containing the datasets. The readxl library provides the function read_excel for easy reading of xlsx files.

library("readxl")
city<- read_excel("City.xlsx")
resort <- read_excel("Resort.xlsx")

We start with checking the dimensions and summary of the dataset.

summary(city)

The city dataset contains 79,330 rows and 28 columns. Whereas, the resort hotel dataset contains 40,060 rows and 28 columns.

```
> summary(city)
   IsCanceled        LeadTime      Arrival Date
 Min.   :0.0000   Min.   :  0.0   Min.   :2015-07-01 00:00:00
 1st Qu.:0.0000   1st Qu.: 23.0   1st Qu.:2015-10-22 00:00:00
 Median :0.0000   Median : 74.0   Median :2016-07-02 00:00:00
 Mean   :0.4173   Mean   :109.7   Mean   :2016-07-09 09:20:54
 3rd Qu.:1.0000   3rd Qu.:163.0   3rd Qu.:2017-02-20 00:00:00
 Max.   :1.0000   Max.   :629.0   Max.   :2017-08-31 00:00:00
                                  NA's   :39270
 ReservationStatusDate          ReservationStatus  StaysInWeekendNights
 Min.   :2014-10-17 00:00:00   Length:79330       Min.   : 0.0000
 1st Qu.:2016-02-05 00:00:00   Class :character   1st Qu.: 0.0000
 Median :2016-08-10 00:00:00   Mode  :character   Median : 1.0000
 Mean   :2016-07-30 17:43:49                      Mean   : 0.7952
 3rd Qu.:2017-02-06 00:00:00                      3rd Qu.: 2.0000
 Max.   :2017-09-07 00:00:00                      Max.   :16.0000
```

We see that the Arrival date has 39,270 rows with NAs. Since this is a substantial proportion of our total rows we cannot remove these rows and leave them for now. A similar summary of the resort data tells us that there are no NA values in the dataset.

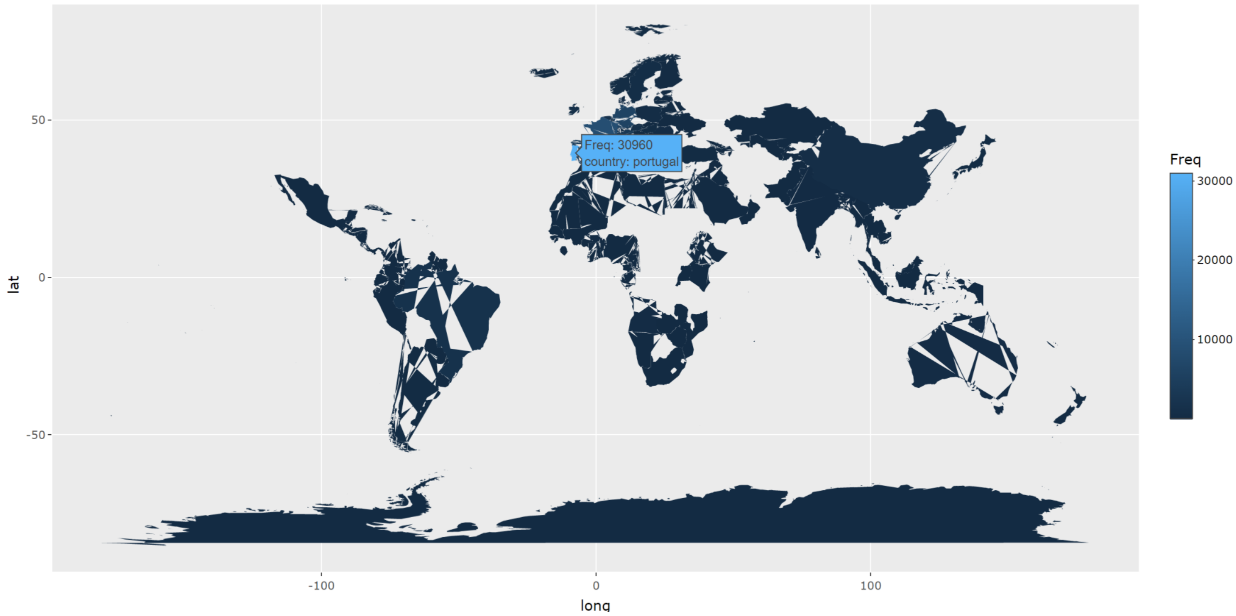
Descriptive Analysis

Customer Demographic Analysis

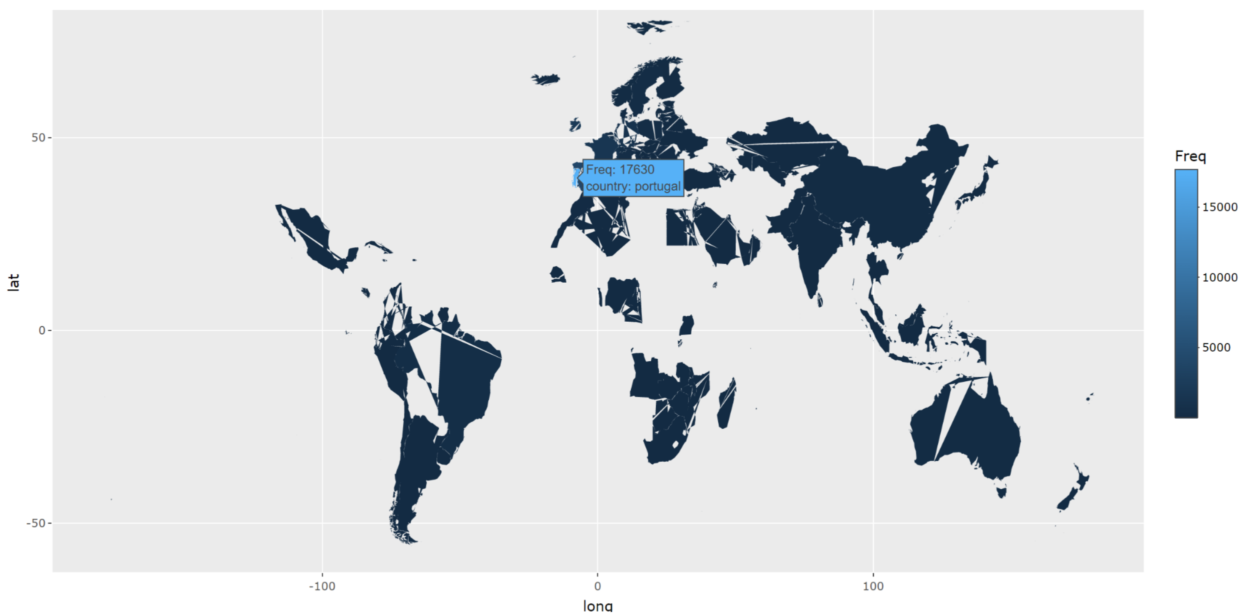
The following ggplot code is used to plot a world map of customers and

```
# map of country where visitors come from
world <- map_data("world")
# getting country names for the country code
c_code <- read_excel("ISO_codes.xlsx")
str(c_code)
c_code <- c_code[c(1,4)]
names(c_code)[1] <- 'Country_name'
names(c_code)[2] <- 'Country_code'
world$region <- tolower(world$region)
c_code$Country_name <- tolower(c_code$Country_name)

df<- as.data.frame(table(city_Country))
df <- merge(df, c_code, by.x="city_Country", by.y="Country_code")
# Retrieve the map data
customer.maps <- map_data("world", region = df$Country_name)

# Compute the centroid as the mean longitude and lattitude
# Used as label coordinate for country's names
region.lab.data <- customer.maps %>%
  group_by(region) %>%
  summarise(long = mean(long), lat = mean(lat))
customer.maps$region <- tolower(customer.maps$region)
library(plotly)
customer.maps <- merge(customer.maps, df, by.x="region", by.y="Country_name")
p<-ggplot(customer.maps, aes(x = long, y = lat, text =paste("country:", region))) +
  geom_polygon(aes( group = group, fill = Freq))
fig <- ggplotly(p)
fig
```

City

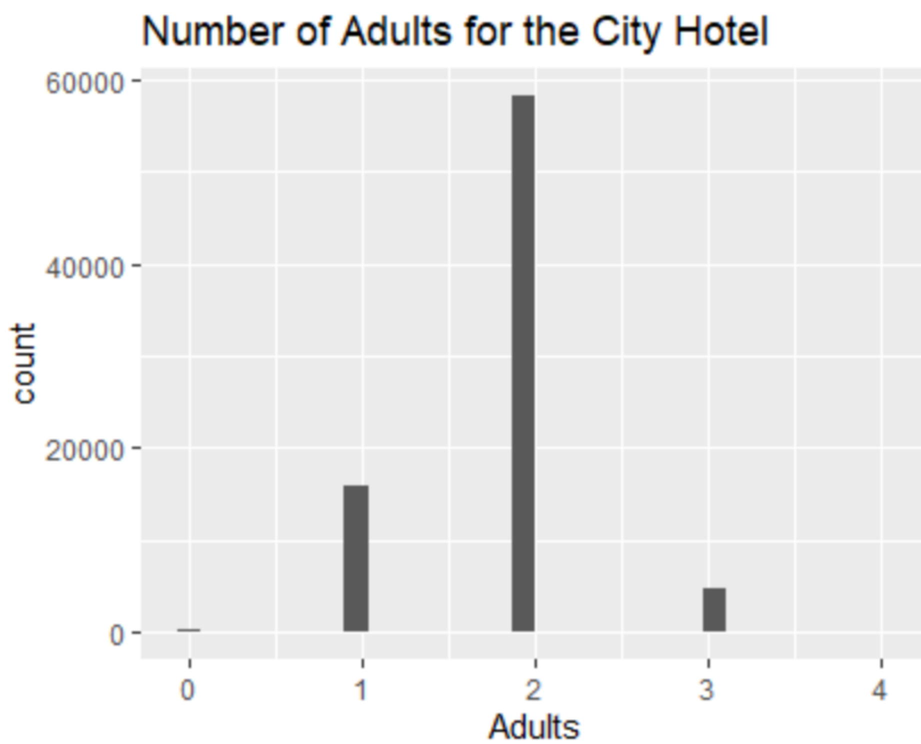


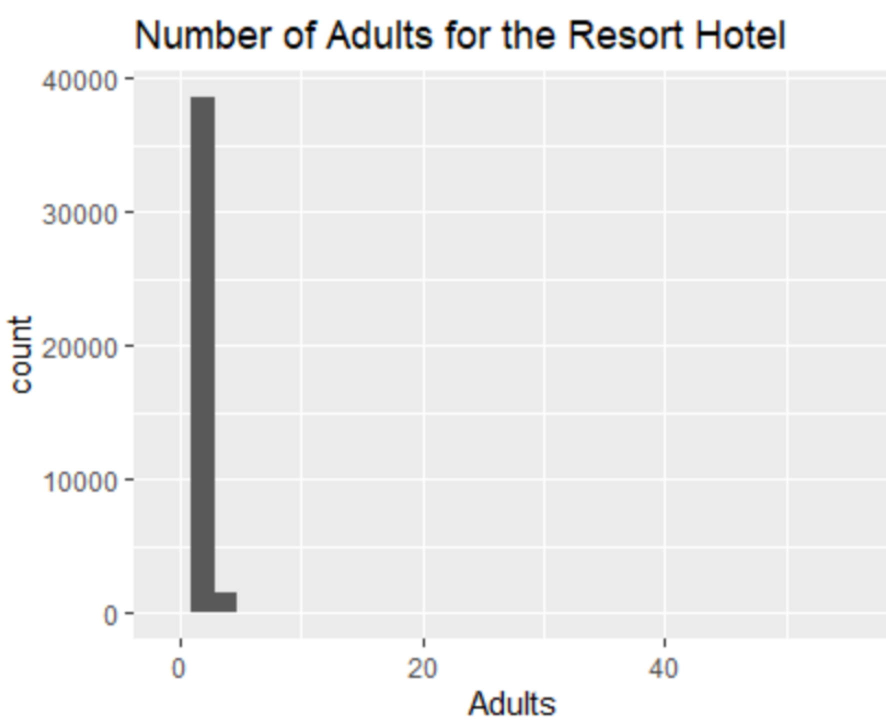
Resort



From the above maps it is clear that most of the people who book the city or resort hotel are from portugal. The color coding shows the minimum to maximum number of customers.

We will try to explore some of the demographics of customers to segment them into groups. First, let's look at the variables, Adult, Children and babies. We will use the ggplot2 library to plot the histogram of these variables. Histograms can tell us the frequency of distribution of the variable.

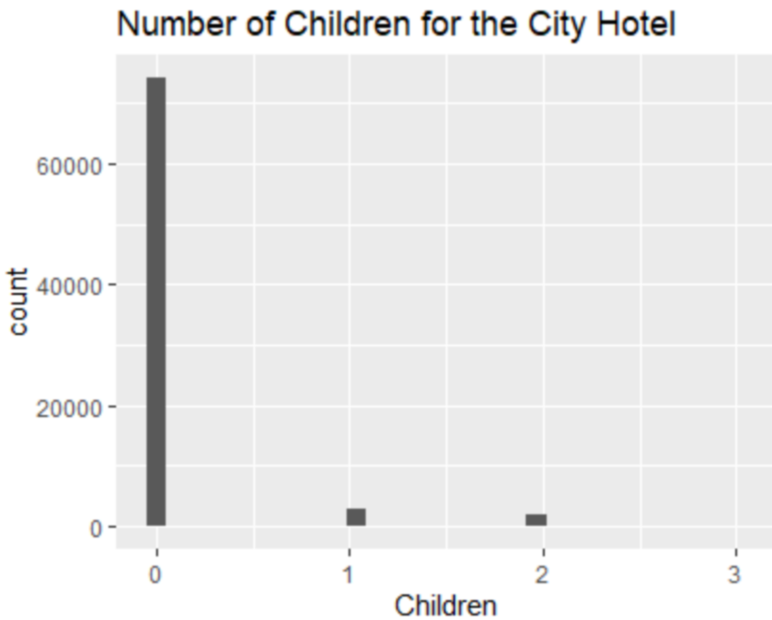


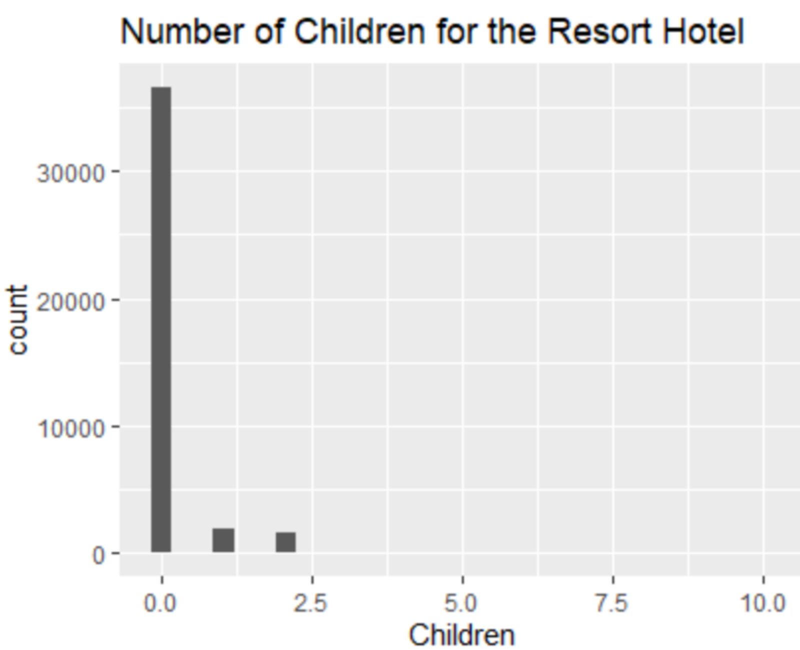



```
> table(resort$Adults)

    0     1     2     3     4     5     6    10    20    26    27    40    50    55
   13  7148 31425  1427    31     2     1     1     2     5     2     1     1     1
```

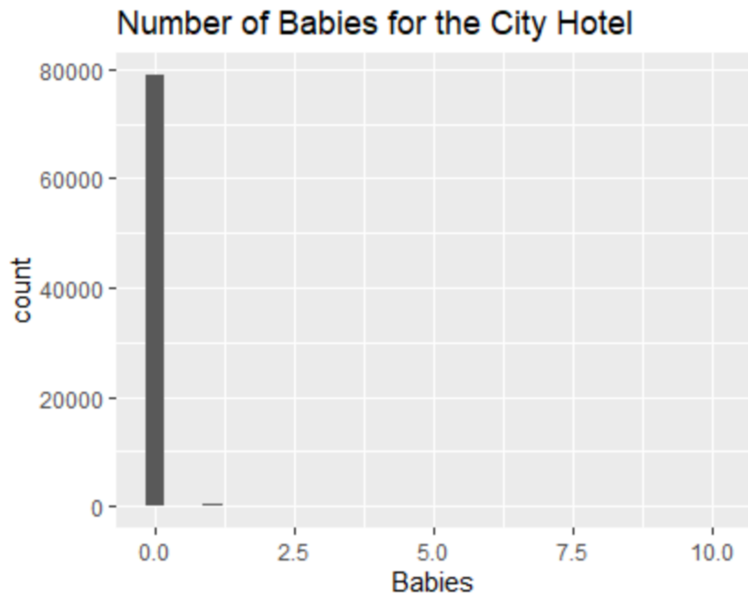
For the city hotel, most reservations have two adults and the number of Adults takes on a minimum of 0 and a maximum of 3. For the resort hotel most reservations have 2 Adults as well. However, one thing that caught my eye is that the number of Adults goes upto 55 for the resort hotel (Maybe a convention was in town!).
Similarly, the code below is to draw a Histogram of the number of babies. Before running that we need to transform the variable into numeric from character.

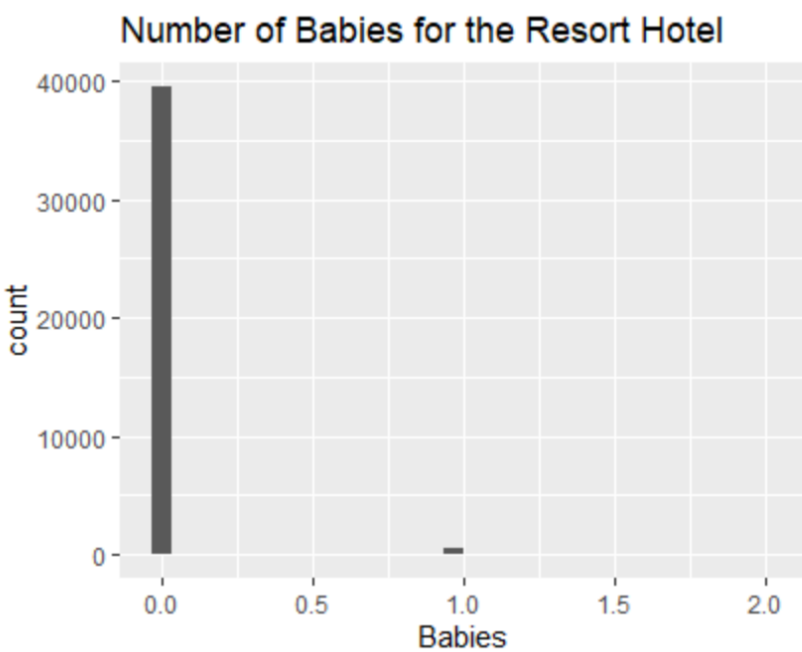
```
city$Children <- as.numeric(city$Children)
```





We can see that most reservations have no children.

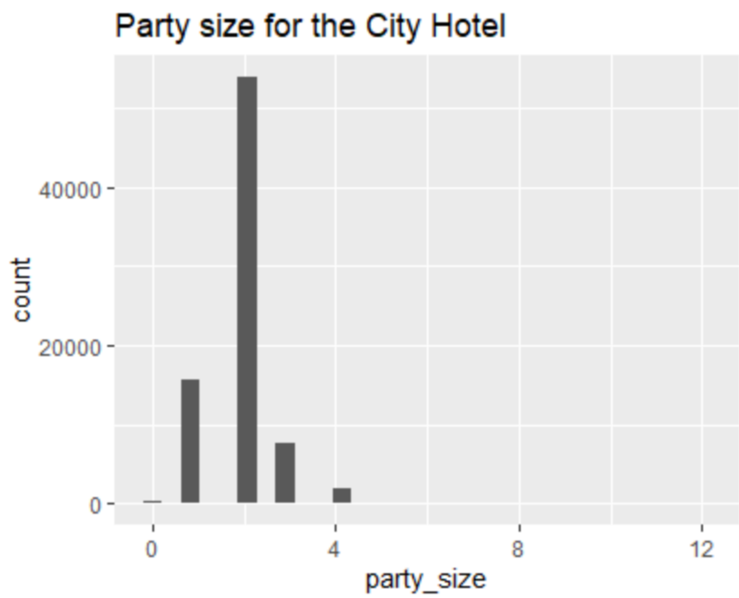


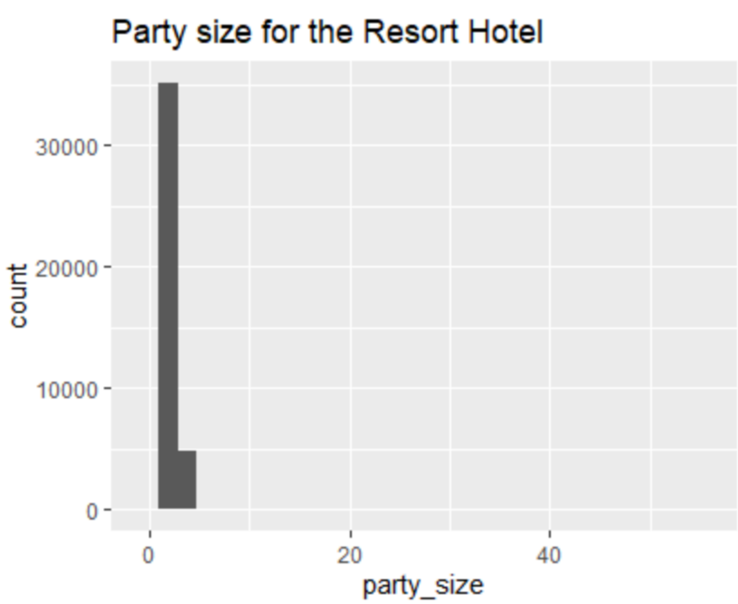


The figure above tells us that most reservations do not have babies. We can create a new variable which contains the family size for each reservation. This will help us later in defining the type of booking party.

```
city$Children <- as.numeric(city$Children)
table(city$Babies)
table(city$Children)
table(city$Adults)
city$party_size <- city$Adults +city$Children + city$Babies
```

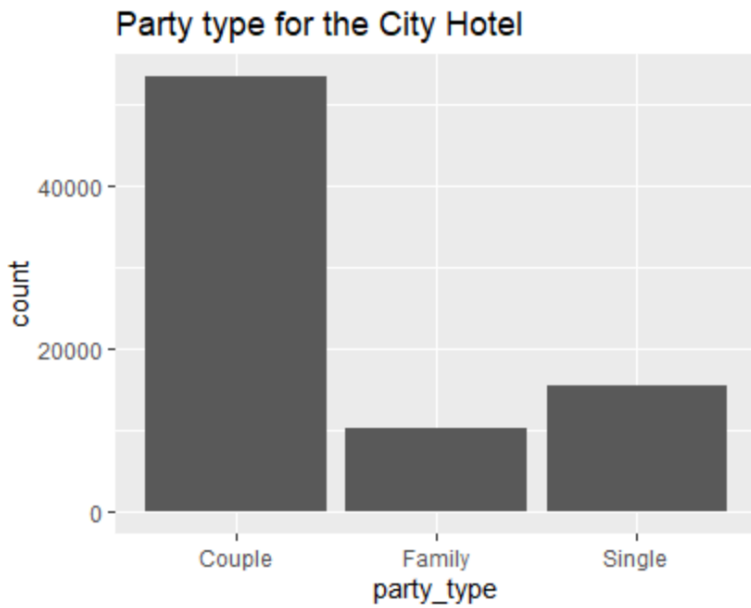
Now we try to plot the distribution of the booking party size. In the figure below, we can see that most bookings have two people.

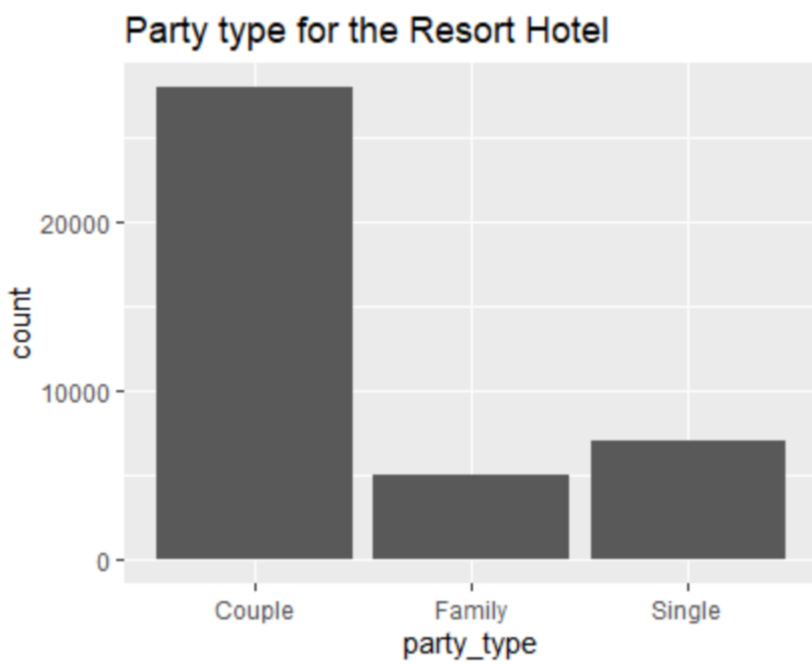




We can try to segment the customers by whether they are a Family/Single/Couple.
We have assumed that “Couple” means a group of just 2 Adults. “Single” party means when there is one person in the booking. All other types of combinations are treated as a “family”.
The following code helps us to create the partition.

```
city$party_type <- "Family"
city$party_type[city$party_size == 1] <- "Single"
city$party_type[city$Adults == 2 & city$party_size == 2] <- "Couple"
city$party_type <- as.factor(city$party_type)
```





The above image tells us that couples are the most common group of customers. We will subset these Couples and study them more closely.

s_city<-city[city$party_type=="Couple",]

The diagram below tells us that only 1.2% of these couples are repeat guests, which is not good.

```
> prop.table(table(s_city$IsRepeatedGuest))

         0          1
0.98798647 0.01201353
```

A look into the cancellations of couples, tells us that 45% of them cancel, which accounts for 30% of the total cancellations.

```
> prop.table(table(s_city$IsCanceled))

        0         1 
0.5499692 0.4500308 
```

Further we look at the deposit type these couples made and we see that 80% of couples do not make a deposit. This could be an indication of why these customers can cancel their reservation at whim.

```
> prop.table(table(s_city$DepositType))

  No Deposit   Non Refund   Refundable 
0.8071670123 0.1926087850 0.0002242027 
```

The above table tells us that only 7% of parties out of these couples have had to face a situation where the requested room was different from the assigned room.

Comparing cancellations

```
> prop.table(table(resort$IsCanceled))

        0         1 
0.7223665 0.2776335 
> prop.table(table(city$IsCanceled))

        0         1 
0.5827304 0.4172696 
```

The figure above shows that cancellations(1) are more frequent in the city hotel than at the resort hotel. The city hotel has 41.7% of bookings as cancellations while the resort hotel has 27.7% cancellations.

Cancellations by Party Type

The City hotel has, for each party_type(Couple/Single/Family) a higher proportion of people who cancel. The highest of these is Couples, they are the highest, out of the three to cancel. While in the resort hotel the highest group is Family.

```
> prop.table(table(city$IsCanceled, city$party_type), margin=2)

       Couple    Family    Single
  0 0.5499692 0.6442035 0.6549332
  1 0.4500308 0.3557965 0.3450668
> prop.table(table(resort$IsCanceled, resort$party_type), margin=2)

       Couple    Family    Single
  0 0.7027181 0.6798324 0.8313133
  1 0.2972819 0.3201676 0.1686867
```

Cancellations by Customer Type

Customer type is an important segmentation, where we can compare cancellations. We see that most Contract customers are the lowest cancelling group in the resort hotel. Whereas, Contract customers at the city hotel have 48 percent of cancellations. Group customers are the lowest proportion of cancellations in the city hotel. Overall, the resort hotel has percentage wise lower cancellations in all groups.

```
> prop.table(table(resort$IsCanceled, resort$CustomerType), margin=2)

     Contract     Group Transient Transient-Party
  0 0.9115991 0.8943662 0.6883048       0.8050314
  1 0.0884009 0.1056338 0.3116952       0.1949686
> prop.table(table(city$IsCanceled, city$CustomerType), margin=2)

      Contract      Group  Transient Transient-Party
  0 0.51956522 0.90102389 0.54383543      0.71903306
  1 0.48043478 0.09897611 0.45616457      0.28096694
```

Cancellations by Market Segment

Market segment Complementary is a group which has the lowest proportion of cancellations for the cityhotel. Whereas, 'Groups' have the highest percentage of cancellations for both hotels.

```
> prop.table(table(city$IsCanceled, city$MarketSegment), margin=2)

     Aviation Complementary Corporate    Direct    Groups Offline TA/TO Online TA Undefined
  0 0.7805907     0.8819188 0.7853315 0.8266864 0.3114132     0.5716845 0.6260194 0.0000000
  1 0.2194093     0.1180812 0.2146685 0.1733136 0.6885868     0.4283155 0.3739806 1.0000000
> prop.table(table(resort$IsCanceled, resort$MarketSegment), margin=2)

    Complementary Corporate    Direct    Groups Offline TA/TO Online TA
  0     0.8358209 0.8479861 0.8651927 0.5760795     0.8476981 0.6475831
  1     0.1641791 0.1520139 0.1348073 0.4239205     0.1523019 0.3524169
```

Cancellations by season

Looking at the cancellations by seasons we see that the city hotel has 90% cancellations out of the total Spring bookings, and 83% cancellation in all Summer bookings.

```
prop.table(table(city$IsCanceled, city$season), margin=2)

      summer     spring     winter     autumn
0 0.16180198 0.09186116 0.86878571 0.37300895
1 0.83819802 0.90813884 0.13121429 0.62699105
prop.table(table(resort$IsCanceled, resort$season), margin=2)

     summer    spring    winter    autumn
0 0.6739130 0.7290556 0.7795821 0.7313187
1 0.3260870 0.2709444 0.2204179 0.2686813
```

For the resort hotel we see less percent cancellations in all the seasons and the highest cancellation percent in the summer season of 32% of all summer bookings.

Comparing the Average Revenue per Stay

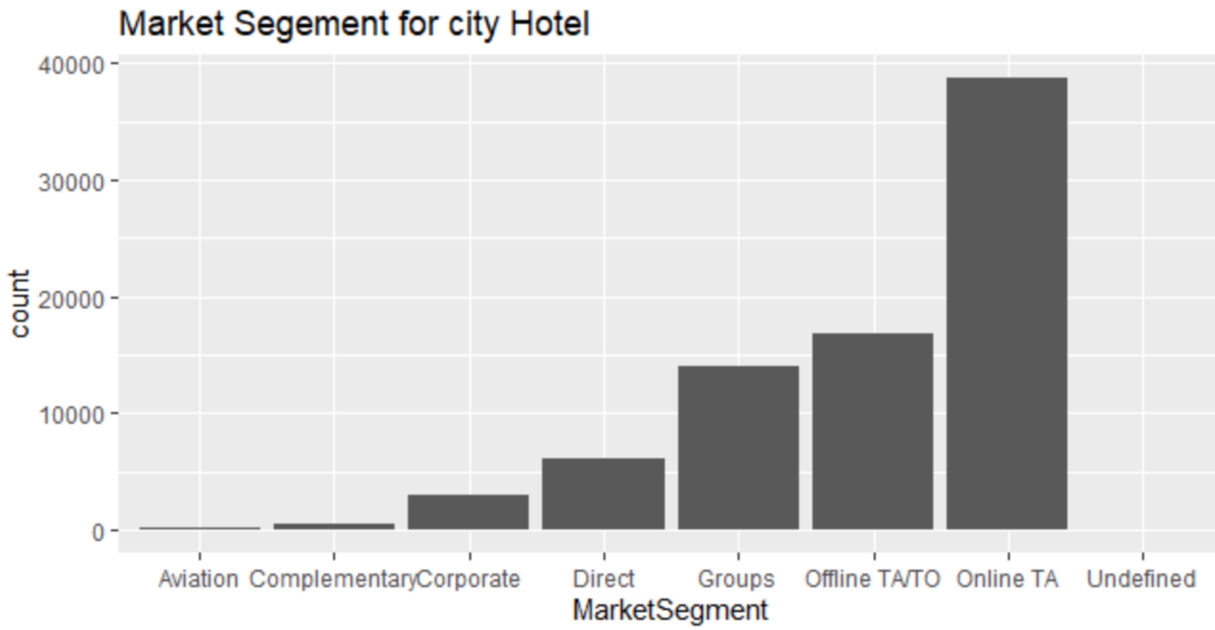
Based on the Average Daily Revenue(ADR), we create a variable called AvgRevPStay which means "Average Revenue Per Stay". The code is shown below for calculating the AvgRevPStay using the ADR and total number of days a person stays in the hotel.

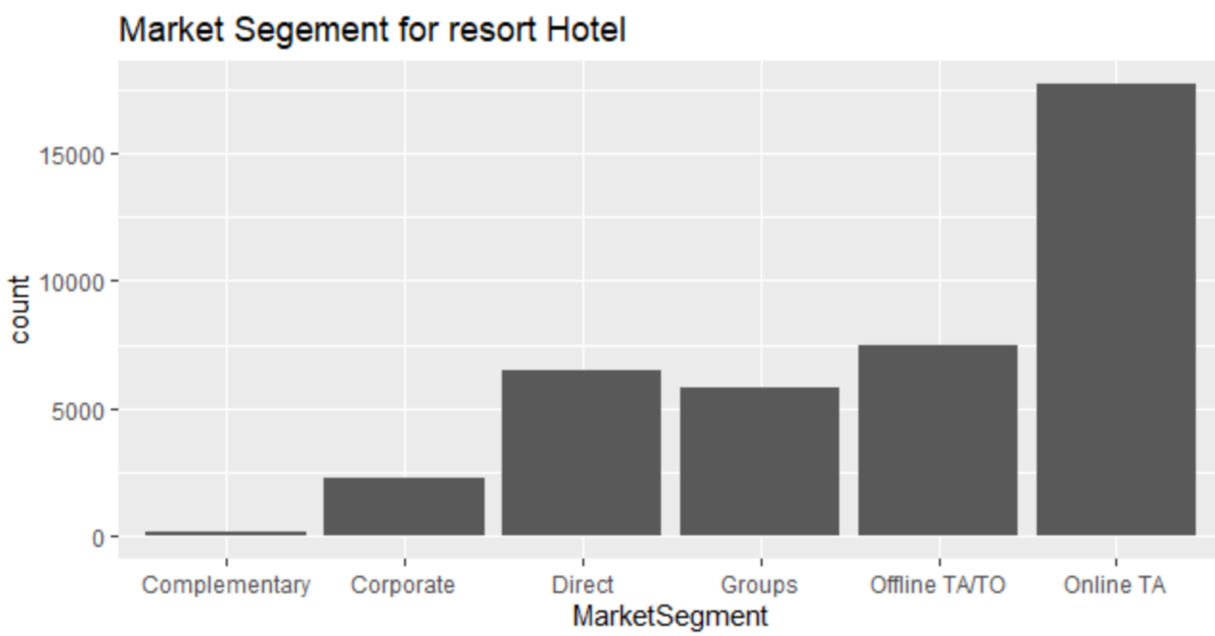
```
resort$StayNights <- resort$StaysInWeekendNights + resort$StaysInWeekNights
resort$AvgRevPStay <- resort$StayNights*resort$ADR
```

We do a summary of the newly created variable and see that mean Average revenue per stay is higher for the resort hotel. The minimum AvgRevPStay is negative for the resort hotel on account of negative ADR.

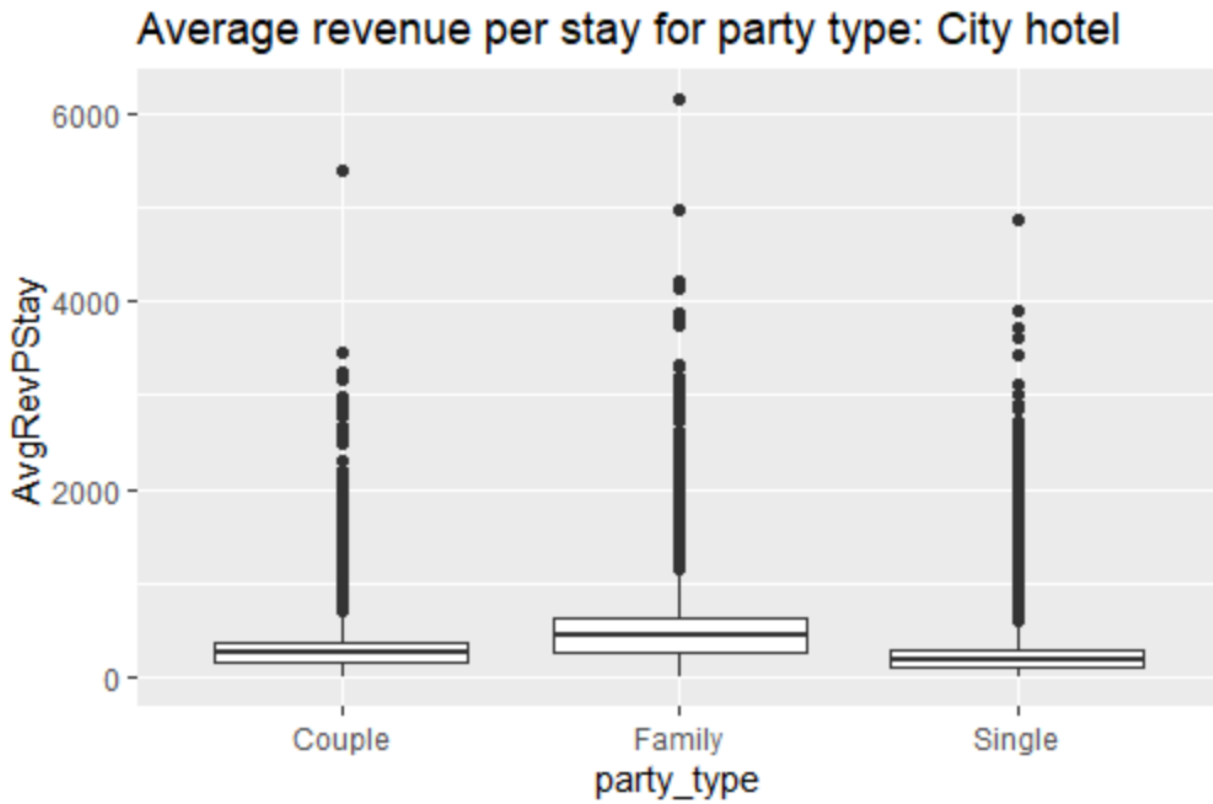
```
> summary(resort$AvgRevPStay)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -63.8   117.0   273.0   435.4   593.0  7590.0
> summary(city$AvgRevPStay)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   160.0   264.0   318.7   401.2  6148.0
```

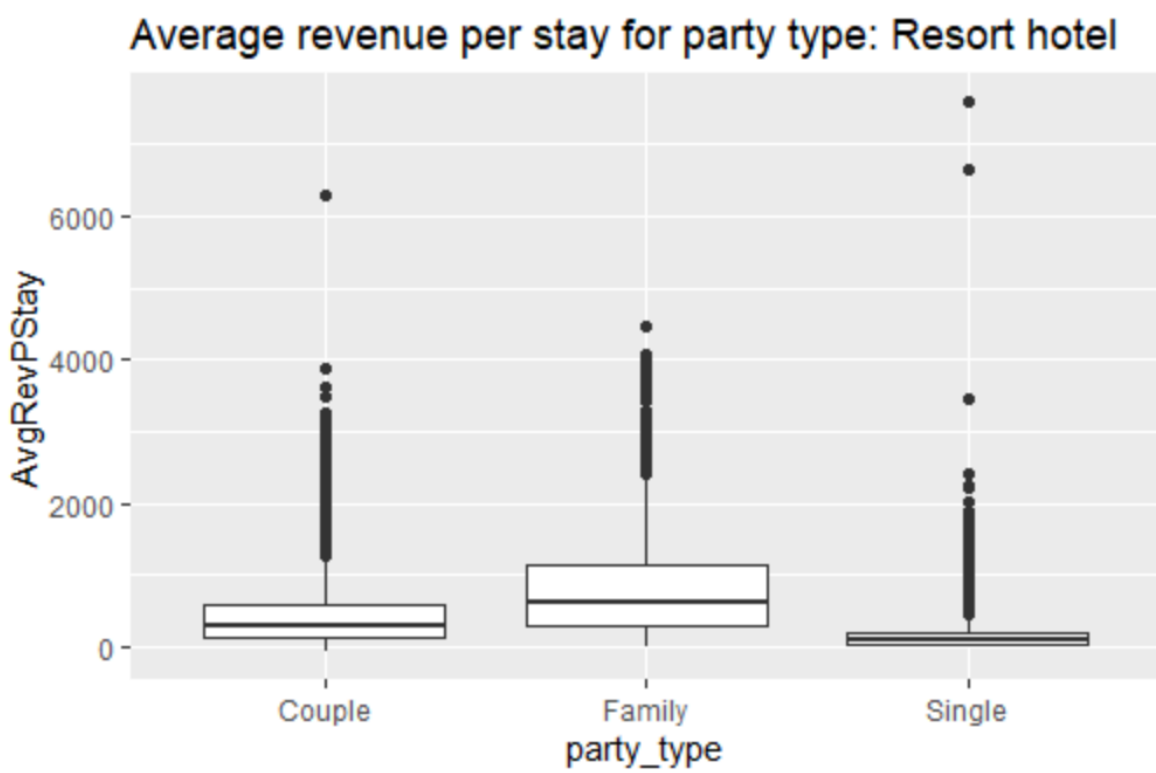
We do a market segmentation of the City and resort hotel below using ggplot.



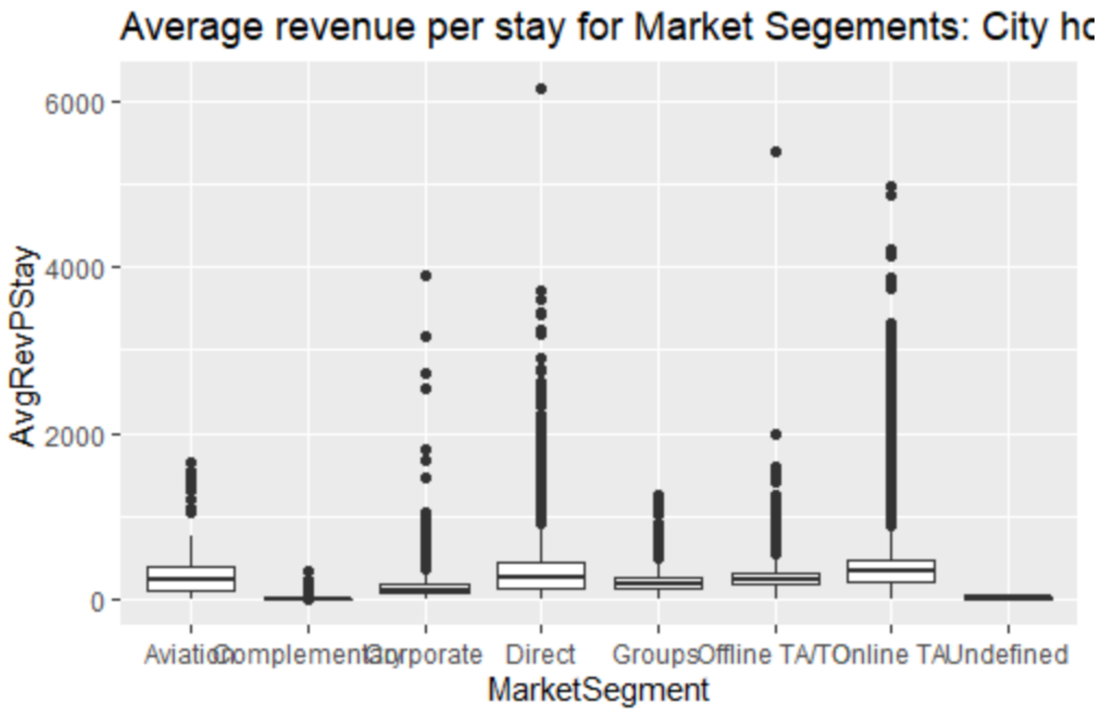


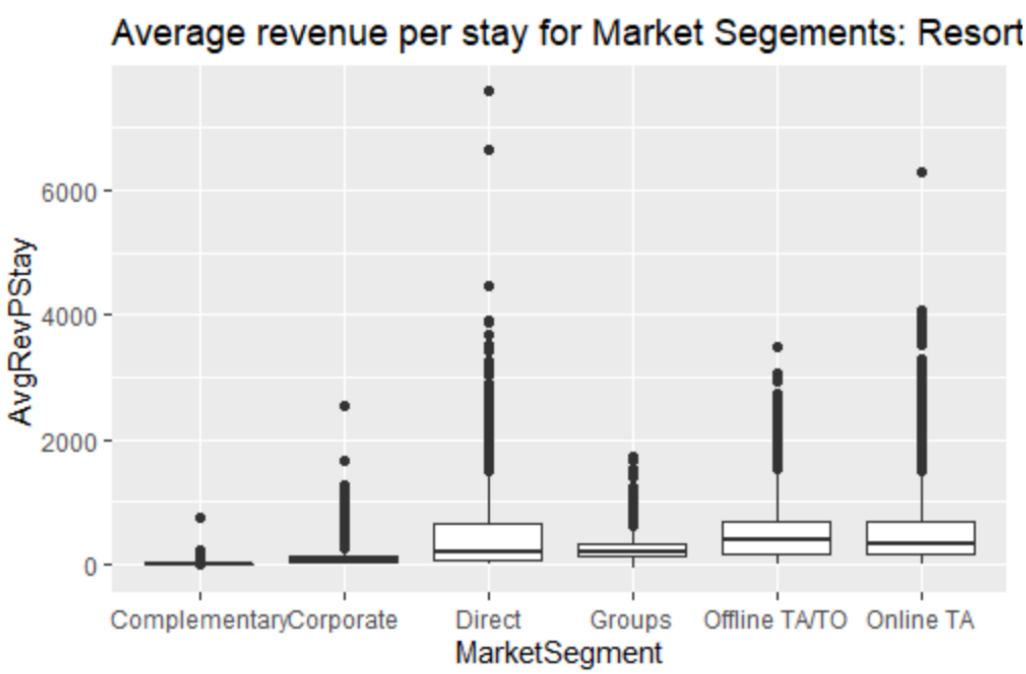
We can see that as compared to direct bookings, group bookings are more likely in the city hotel. The most frequent market segment is Online Travel Agent for both of them.


Average revenue per stay for party type: City hotel
AvgRevPStay
6000
4000
2000
0
Couple
Family
Single
party_type

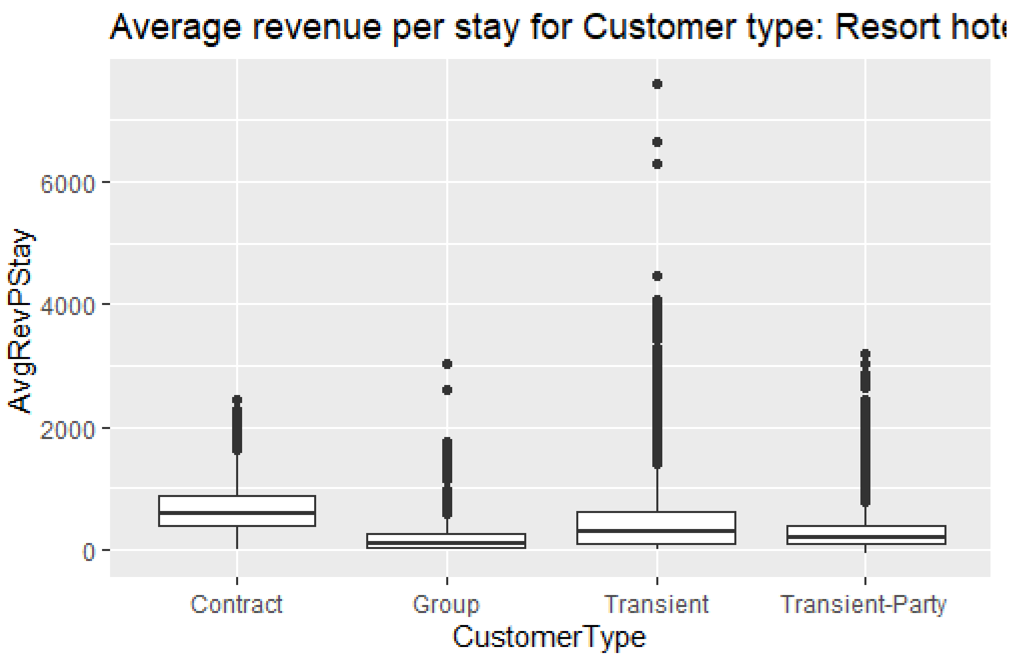

Average revenue per stay for party type: Resort hotel
AvgRevPStay
6000
4000
2000
0
Couple
Family
Single
party_type

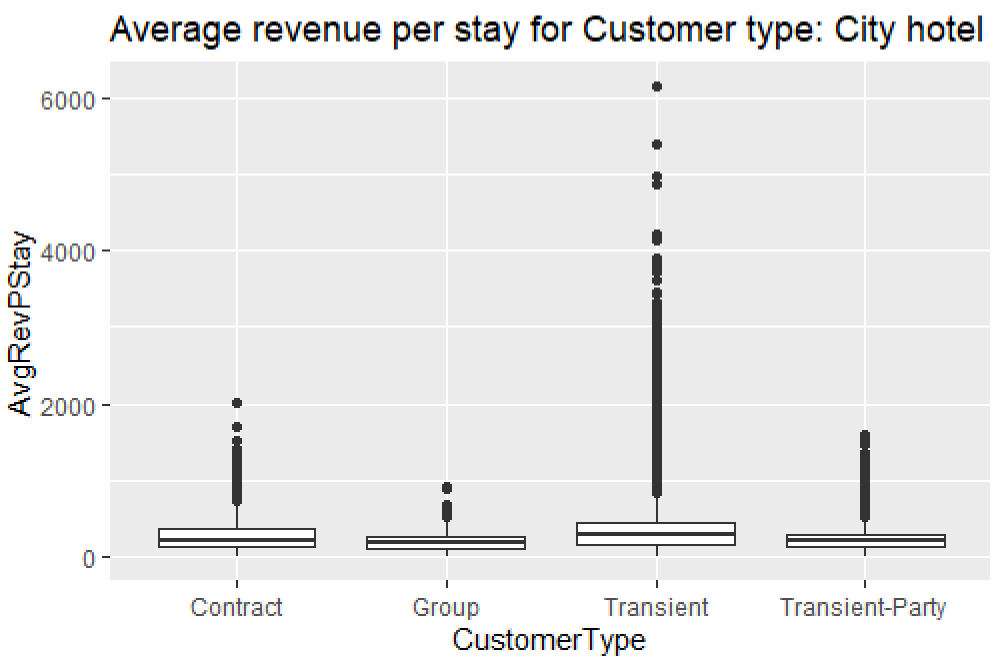
The above diagrams show the revenue per stay grouped by part type(Couple/Family/Single). We see that Family type bookings have the highest average revenue per stay and singles have the lowest average revenue per stay for both resort and city hotels.





The above diagrams show the revenue per stay grouped by Market Segment. We see that ‘Complementary’ segment customers have the lowest average revenue per stay in both the city and resort hotels.





From the above two figures we can infer that 'Group' Customers have the lowest Average Revenue per stay. We move to an apriori analysis of the cancellations to try and find structured rules that can be correlated with cancellations.

Apriori Analysis of Cancellations

We look at the cancellations for both the hotels using apriori rules. This algorithm requires factors as the input. We started off with creating new factors to categorize some of the numeric variables in the dataset.
For example, BookingTime denotes the time that elapsed between the entering date of the booking into the PMS and the arrival date. We have divided the variable Lead time into 4 groups to calculate BookingTime. The lowest value meaning small lead time and 4 meaning the longest of lead time.

```
city$BookingTime <- 4
city$BookingTime[city$LeadTime < 23] <- 1
city$BookingTime[city$LeadTime >= 23 & city$LeadTime < 74] <- 2
city$BookingTime[city$LeadTime >= 74 & city$LeadTime < 163] <- 3
city$BookingTime <- as.factor(city$BookingTime)

quantile(resort$LeadTime, probs=c(0.25, 0.5, 0.75))
resort$BookingTime <- 4
resort$BookingTime[resort$LeadTime < 10] <- 1
resort$BookingTime[resort$LeadTime >= 10 & resort$LeadTime < 57] <- 2
resort$BookingTime[resort$LeadTime >= 57 & resort$LeadTime < 155] <- 3
resort$BookingTime <- as.factor(resort$BookingTime)
```

Another transformation we did was to the Average Revenue Per stay variable by dividing it into 4 groups. Group 1 meaning lowest AvgRevPStay and Group 4 meaning the highest.

```
city$AvgRevPS <- 4
city$AvgRevPS[city$AvgRevPStay < 160] <- 1
city$AvgRevPS[city$AvgRevPStay >= 160 & city$AvgRevPStay < 264] <- 2
city$AvgRevPS[city$AvgRevPStay >= 264 & city$AvgRevPStay < 401] <- 3
city$AvgRevPS <- as.factor(city$AvgRevPS)

quantile(resort$AvgRevPStay, probs=c(0.25, 0.5, 0.75))
resort$AvgRevPS <- 4
resort$AvgRevPS[resort$AvgRevPStay < 117] <- 1
resort$AvgRevPS[resort$AvgRevPStay >= 117 & resort$AvgRevPStay < 273] <- 2
resort$AvgRevPS[resort$AvgRevPStay >= 273 & resort$AvgRevPStay < 593] <- 3
resort$AvgRevPS <- as.factor(resort$AvgRevPS)
```

To capture seasonality we have also captured the season of customer’s arrival time for these hotels. Dividing the seasons majorly into Summer/Spring/Winter/Autumn. Based on the month we have made the assumption that Winter is from December to February. Spring is from March to May. Summer is from June to August, and Autumn is from September to November.

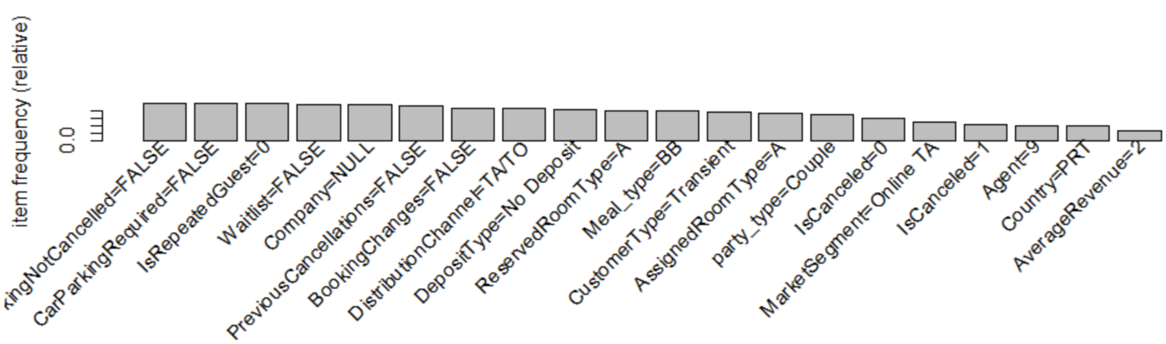
Season of Arrival City

```
city$`Arrival Date`<-strptime(city$`Arrival Date`,format="%Y-%m-%d")
city$month<-as.numeric(format(city$`Arrival Date`,"%m"))
city$season <- "winter"
city$season[city$month>=6&city$month<=8]<-"summer"
city$season[city$month>=9&city$month<=11]<-"autumn"
city$season[city$month>=3&city$month<=5]<-"spring"
city$season<-factor(city$season,levels=c("summer","spring","winter","autumn"))
```

Season of Arrival Resort

```
resort$`Arrival Date`<-strptime(resort$`Arrival Date`,format="%Y-%m-%d")
resort$month<-as.numeric(format(resort$`Arrival Date`,"%m"))
resort$season<-"winter"
resort$season[resort$month>=6&resort$month<=8]<-"summer"
resort$season[resort$month>=9&resort$month<=11]<-"autumn"
resort$season[resort$month>=3&resort$month<=5]<-"spring"
resort$season<-factor(resort$season,levels=c("summer","spring","winter","autumn"))
```

Armed with these new variables and several categorical variables from the existing dataset we create a subset data frame containing only these variables and transform the data frame into a transactions table. From the transactions table we get the item frequencies and plot it sorted. Thus we see that BookingNotCancelled=FALSE is the most common item.



City Hotel

For a 40% confidence and 40% support threshold we found further rules like:

- {IsRepeatedGuest=0, Company=NULL} => {IsCanceled=1}
- {IsRepeatedGuest=0, Company=NULL, CarParkingRequired=FALSE} => {IsCanceled=1}

- {IsRepeatedGuest=0, Company=NULL, CarParkingRequired=FALSE, PreviousBookingNotCancelled=FALSE} => {IsCanceled=1}
- lhs rhs support confidence coverage lift count
-
- {Company=NULL} => {IsCanceled=1} 0.4073869 0.4272551 0.9534980 1.023930 32318
- {IsRepeatedGuest=0}=> {IsCanceled=1} 0.4117106 0.4225336 0.9743855 1.012615 32661
- {CarParkingRequired=FALSE}=> {IsCanceled=1} 0.4172696 0.4276523 0.9757217 1.024882 33102
- {PreviousBookingNotCancelled=FALSE} => {IsCanceled=1} 0.4157444 0.4242366 0.9799824 1.016697 32981

Most of these rules focus on whether the customer has previously cancelled, whether they require a car parking, are they a repeated guest and did they forget to fill out the company info (or for whatever reasons the company value might be null).

The apriori rules for 25% support and 60% confidence are displayed below.

```
    lhs                                                                    rhs            support   confidence coverage  lift     count
[1] {Country=PRT}                                                       => {IsCanceled=1} 0.2531325 0.6486111  0.3902685 1.554417 20081
[2] {Country=PRT,CarParkingRequired=FALSE}                              => {IsCanceled=1} 0.2531325 0.6631113  0.3817345 1.589167 20081
[3] {Country=PRT,PreviousBookingNotCancelled=FALSE}                     => {IsCanceled=1} 0.2518341 0.6730679  0.3741586 1.613029 19978
[4] {Country=PRT,PreviousBookingNotCancelled=FALSE,CarParkingRequired=FALSE} => {IsCanceled=1} 0.2518341 0.6833590  0.3685239 1.637692 19978
> 
```

We observe that Customers from Portugal, who have previously cancelled a booking are likely to cancel with a 67.3% confidence and 25.2% support. The last rule from the above diagram is a further segment of customers from the above category, Portugal customers, who have previously cancelled a booking and who do not require parking are 68.3% likely to cancel and these incidents are supported by 25.1% cases.

We ran further models to identify more unique rules which might have even higher confidence but lower occurrences in our data.

To get more unique rules, we use 20% support and 80% confidence. These rules are really important as they have high confidence and just enough support to use these rules.

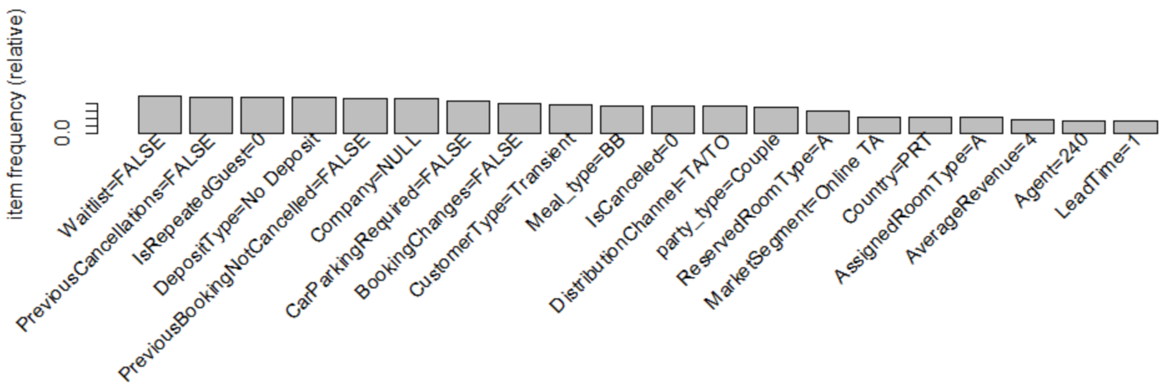
```
     lhs                                        rhs              support confidence  coverage     lift count
[1]  {DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2171940  0.8083509 0.2686878 1.937239 17230
[2]  {Meal_type=BB,
      IsRepeatedGuest=0,
      AssignedRoomType=A,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2001891  0.8049572 0.2486953 1.929106 15881
[3]  {Meal_type=BB,
      AssignedRoomType=A,
      Country=PRT,
      PreviousBookingNotCancelled=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2037817  0.8064049 0.2527039 1.932575 16166
[4]  {ReservedRoomType=A,
      DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2171436  0.8100635 0.2680575 1.941343 17226
[5]  {DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Company=NULL,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2149880  0.8095600 0.2655616 1.940136 17055
[6]  {IsRepeatedGuest=0,
      DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2135132  0.8100430 0.2635825 1.941294 16938
[7]  {DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Country=PRT,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2171940  0.8123910 0.2673516 1.946921 17230
```

```
[8]  {DistributionChannel=TA/TO,
      AssignedRoomType=A,
      Country=PRT,
      PreviousBookingNotCancelled=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2168789  0.8104098 0.2676163 1.942173 17205
[9]  {ReservedRoomType=A,
      Meal_type=BB,
      IsRepeatedGuest=0,
      AssignedRoomType=A,
      Country=PRT,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2000882  0.8069239 0.2479642 1.933819 15873
[10] {ReservedRoomType=A,
      Meal_type=BB,
      AssignedRoomType=A,
      Country=PRT,
      PreviousBookingNotCancelled=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2036808  0.8084255 0.2519476 1.937417 16158
[11] {Meal_type=BB,
      IsRepeatedGuest=0,
      AssignedRoomType=A,
      Country=PRT,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2001891  0.8114143 0.2467162 1.944580 15881
[12] {Meal_type=BB,
      IsRepeatedGuest=0,
      AssignedRoomType=A,
      Country=PRT,
      PreviousBookingNotCancelled=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2000504  0.8054202 0.2483802 1.930215 15870
[13] {Meal_type=BB,
      AssignedRoomType=A,
      Country=PRT,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}                  => {IsCanceled=1} 0.2037817  0.8128111 0.2507122 1.947928 16166
```

Based on these rules, the important factors at play with the City hotels are whether the Distribution Channel is Travel Agent/Tour Operators. Assigned room type A comes up in a lot of these rules and so does the country portugal. Whether customers make changes to their bookings, what is the meal type, whether they have previously cancelled, are they a repeated guest. When a combination of these factors exist together, we have a 80+ percent confidence of a likely cancellation.

Resort Hotel

For the resort hotel we applied a similar analysis and found that the most common feature here was waitlist=FALSE.



We did not find any rules with more than 50% confidence. The best rules we found were for 20% support and 40% confidence threshold. There were 14 rules which can be seen below.

```
[1]  {CustomerType=Transient,
      Company=NULL,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2058163  0.4050403 0.5081378 1.458903  8245
[2]  {CustomerType=Transient,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2106091  0.4128095 0.5101847 1.486886  8437
[3]  {IsRepeatedGuest=0,
      CustomerType=Transient,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2104843  0.4099174 0.5134798 1.476469  8432
[4]  {CustomerType=Transient,
      Company=NULL,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2051922  0.4156133 0.4937094 1.496985  8220
[5]  {IsRepeatedGuest=0,
      CustomerType=Transient,
      Company=NULL,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2047678  0.4151946 0.4931852 1.495477  8203
[6]  {CustomerType=Transient,
      Company=NULL,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2056665  0.4049445 0.5078882 1.458557  8239
[7]  {IsRepeatedGuest=0,
      CustomerType=Transient,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2099601  0.4151940 0.5056915 1.495475  8411
[8]  {CustomerType=Transient,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2104593  0.4127178 0.5099351 1.486556  8431
[9]  {IsRepeatedGuest=0,
      CustomerType=Transient,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2103345  0.4098249 0.5132302 1.476136  8426
[10] {IsRepeatedGuest=0,
      CustomerType=Transient,
      Company=NULL,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2045931  0.4179287 0.4895407 1.505325  8196
[11] {CustomerType=Transient,
      Company=NULL,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2050424  0.4155200 0.4934598 1.496649  8214
[12] {IsRepeatedGuest=0,
      CustomerType=Transient,
      Company=NULL,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2046181  0.4151010 0.4929356 1.495140  8197
[13] {IsRepeatedGuest=0,
      CustomerType=Transient,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2098103  0.4151027 0.5054418 1.495146  8405
[14] {IsRepeatedGuest=0,
      CustomerType=Transient,
      Company=NULL,
      PreviousBookingNotCancelled=FALSE,
      CarParkingRequired=FALSE,
      Waitlist=FALSE,
      BookingChanges=FALSE}              => {IsCanceled=1} 0.2044433  0.4178358 0.4892911 1.504990  8190
```

We also ran apriori rules with a lesser support but higher confidence to find more unique associations. Even if the support is 10% - meaning that it is a low occurrence in out data set - we can see which variables have an extremely likely chance to end up cancelling. A 67% confidence of cancellation is something that should be addressed, or at least attempted to counteract to some degree.

```
    lhs                              rhs             support confidence  coverage     lift count
{party_type=Couple,
     DistributionChannel=TA/TO,
     Country=PRT,
     CarParkingRequired=FALSE,
     BookingChanges=FALSE}           => {IsCanceled=1} 0.1077883  0.6700807 0.1608587 2.413544  4318
{party_type=Couple,
     DistributionChannel=TA/TO,
     Company=NULL,
     Country=PRT,
     CarParkingRequired=FALSE,
     BookingChanges=FALSE}           => {IsCanceled=1} 0.1074638  0.6710834 0.1601348 2.417155  4305
{party_type=Couple,
     DistributionChannel=TA/TO,
     Country=PRT,
     PreviousBookingNotCancelled=FALSE,
     CarParkingRequired=FALSE,
     BookingChanges=FALSE}           => {IsCanceled=1} 0.1077384  0.6804351 0.1583375 2.450839  4316
{party_type=Couple,
     IsRepeatedGuest=0,
     DistributionChannel=TA/TO,
     Country=PRT,
     CarParkingRequired=FALSE,
     BookingChanges=FALSE}           => {IsCanceled=1} 0.1065402  0.6812450 0.1563904 2.453756  4268
```

For the resort, we see that variables like waitlist, party types being couples, customers from Portugal, and transient customer types are are fairly common occurrences with other variables. This indicates that these attributes may lead to higher chances of cancellations, and we can use this information to accommodate or focus efforts on high risk variables like these. Based on our findings from these rules, we believe that prioritizing strategies towards transient customers, couples, those not from a company, and those who are not a repeated guests could lead to improvements in the overall cancellation rates in the resort setting.

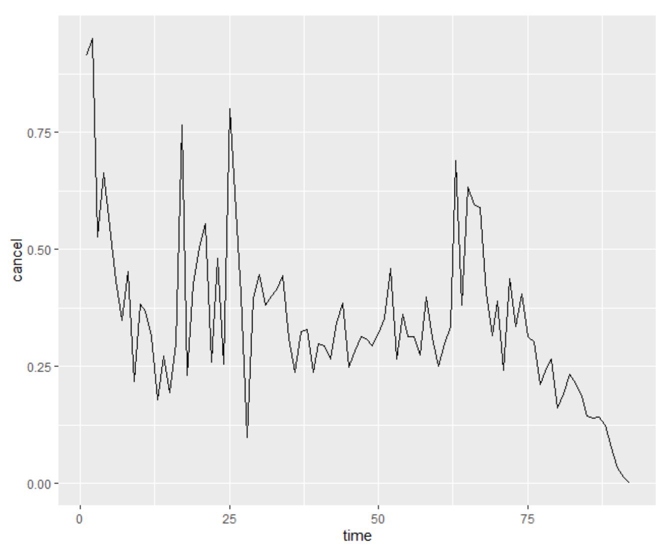
Linear Model of Cancellations

We use weekly and daily data for our analysis. Weekly data smooths the data, because sometimes daily data fluctuate dramatically, causing outliers that will interfere with the accuracy of the whole model.

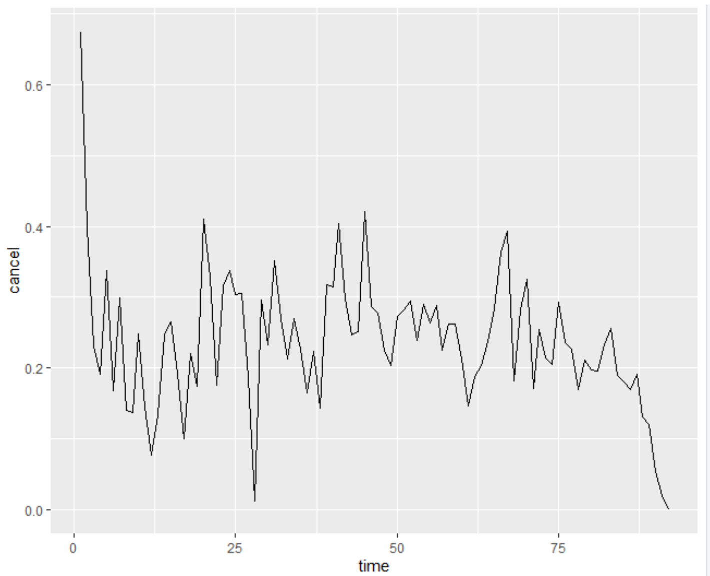
Variable Name	Description

Average ADR (aveADR)	Average revenue by week
LeadTime	Average weekly number of days between the booking and the arrival date
Seasonality (Spring, Summer, Fall, Winter)	Dummy variables indicating when these reservations are made according to "ReservationStatusDate"
Customer Age Group (Adults, Children, Babies)	Average number of customer of different ages
A Room Type	Value indicating if the reservation of room is type A (1) or others (0)
Customer Type	Value indicating type of reservation
Distribution Channel (DCCorporate, DCDirect, DCTATO, DCGDS)	Values indicating booking distribution channel
Deposit Type (No Deposit, Non Refund, Refundable)	Values indicating if the customer made a deposit to guarantee the booking

Cancellation in City hotels:



Cancellation in Resort hotels:



We tried to draw a line plot and looking at the plot, we found that the cancellation rate decreased with time, so the first variable we added to the linear model was time. We also saw that the plot fluctuated by season, and added seasonality into the model. We found that some seasonality variables are insignificant like summer, so we dropped it. After dropping the summer variable, the total adjusted R-Squared increases, so we accepted it.

We tried to add other variables like lead time, customer type and room type, and then repeated the steps to test p-value and adjusted R-Squared.

We tried to work with the distribution channel variable, but unfortunately it wasn't significant and the adjusted R-squared didn't increase, so finally we took it out.

In the end, we got the final linear model because the adjusted R-Squared was higher and all the variables were significant.

City Hotel

```
lmOutF<-lm(formula=cancel~LeadTime+Winter+time+Contract+roomtype,data=cityWeekC)
summary(lmOutF)
```

```
#take customer type into consideration
lmOutE <- lm(formula=cancel~LeadTime+Summer+Fall+Winter+time+Contract+Group+Transient+TransientParty
,data=cityWeekC)
summary(lmOutE)
lmOutF <- lm(formula=cancel~LeadTime+Winter+time+Contract+roomtype,data=cityWeekC)
summary(lmOutF)

#This model shows that cancellation rate is about 34.03% with other variables consistent.
#Over time, cancellation rate decreases 0.46% per week.
#With other variables consistent, leadtime increases 1 day, cancellation rate increases 0.06%.
#Only in Winter, cancellation rate increases 17.99%.
#Only for contract type reservation, cancellation rate increases 0.23%.
#Only for A room type reservation, cancellation rate increases 0.03%.

#Adjusted R^2 is 0.655, which means 65.5% of the data fits the linear model.
#p-value: < 2.2e-16 means the whole model is significant.
```

105:41 (Top Level) ÷

Console Terminal × Jobs ×

E:/Education/MSBA/DS/Team Project/

```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.3403518  0.0440490   7.727 1.88e-11 ***
LeadTime     0.0006175  0.0003101   1.991   0.0496 *  
Winter       0.1799203  0.0341056   5.275 9.76e-07 ***
time        -0.0045993  0.0005074  -9.065 3.63e-14 ***
Contract    -0.0022550  0.0004625  -4.875 4.92e-06 ***
roomtype     0.0003389  0.0000547   6.196 1.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1034 on 86 degrees of freedom
Multiple R-squared:  0.674,	Adjusted R-squared:  0.655 
F-statistic: 35.55 on 5 and 86 DF,  p-value: < 2.2e-16
```

Resort Hotel

lmOutI <- lm(formula=cancel~Fall+Winter+time+Group+Transient,data=resortWeekC)
summary(lmOutI)

```
lmOutH <- lm(formula=cancel~LeadTime+Summer+Fall+Winter+time,data=resortWeekC)
summary(lmOutH)
lmOutI <- lm(formula=cancel~Fall+Winter+time+Group+Transient,data=resortWeekC)
summary(lmOutI)
#The cancellation in resort in such more difficult to predict because adjusted R^2 always too low.

#This model shows that cancellation rate is about 15.54% with other variables consistent.
#Over time, cancellation rate decreases 0.09% per week.
#In Fall, cancellation rate decreases 2.72%.
#In Winter, cancellation rate increases 5.02%.
#For transient type reservation, cancellation rate increases 0.06%.

#Adjusted R^2 is 0.1937, which means 19.37% of the data fits the linear model.
#p-value: < 0.05 means the whole model is significant.
```

63:17 (Top Level) ÷

Console Terminal × Jobs ×

E:/Education/MSBA/DS/Team Project/

```
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.1554862  0.0470763   3.303  0.00140 **
Fall        -0.0272439  0.0199143  -1.368  0.17486   
Winter       0.0502346  0.0289996   1.732  0.08681 . 
time        -0.0009197  0.0003743  -2.457  0.01602 * 
Group       -0.0065659  0.0044247  -1.484  0.14148   
Transient    0.0005818  0.0001713   3.395  0.00104 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08628 on 86 degrees of freedom
Multiple R-squared:  0.238,	Adjusted R-squared:  0.1937 
F-statistic: 5.372 on 5 and 86 DF,  p-value: 0.0002403
```

In conclusion, we know that overtime, the average cancellation rate decreased and the general cancellation rate in City hotels were higher than that of Resort hotels. We anticipate that Winter will lead to an increase in cancellation rate in both City and Resort hotels while Fall would lead to a decrease in cancellation. Variable CustomerType influences both City and Resort hotels while Variable RoomTypeA and LeadTime only influence cancellation in City hotels.

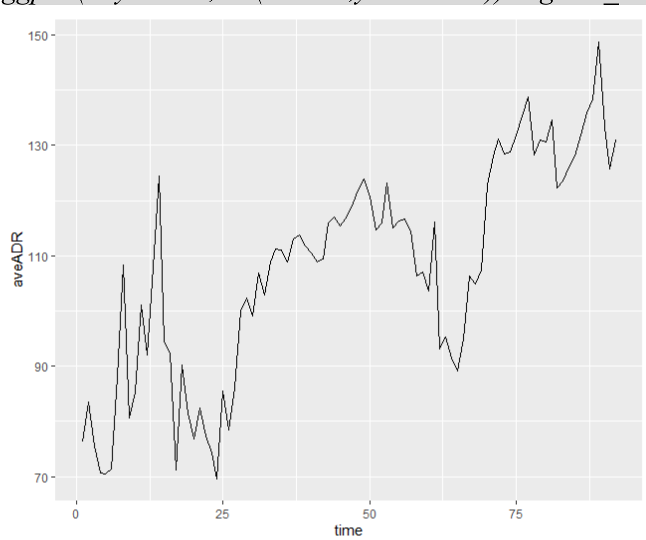
Linear Model of Revenue

We also drew a line plot and found out that the revenue increases with time and fluctuates regularly, so we added time and seasonality as variables to the model. We also found that some variables like summer were insignificant, so we dropped it and added significant ones like LeadTime. We believe how many people in different age groups checking in the hotel, such as the number of adults and babies etc, will influence the revenue so we took the customer age group as a variable to the model.

City Hotel

Average ADR with time for the City Hotel

ggplot(cityWeekC,aes(x=time,y=aveADR)) + geom_line()



Linear Model of Average ADR for the City Hotel

```
lmOutA <- lm(formula=aveADR~LeadTime+Fall+Winter+time+children,data=cityWeekC)
summary(lmOutA)
#This model shows that the average ADR is about $89.82 at the beginning of spring and summer.
#Over time, average ADR increase 0.52 per week.
#With other variables consistent, leadtime increases 1 day, average ADR decreases 0.06.
#In Fall, average ADR decreases 5.26.
#In winter, average ADR decrease 21.78.
#Every one child increases average ADR by 43.76.

#Adjusted R^2 is 0.7789, which means 77.89% of the data fits the linear model.
#p-value: < 2.2e-16 means the whole model is significant.
```

59:1 (Top Level)

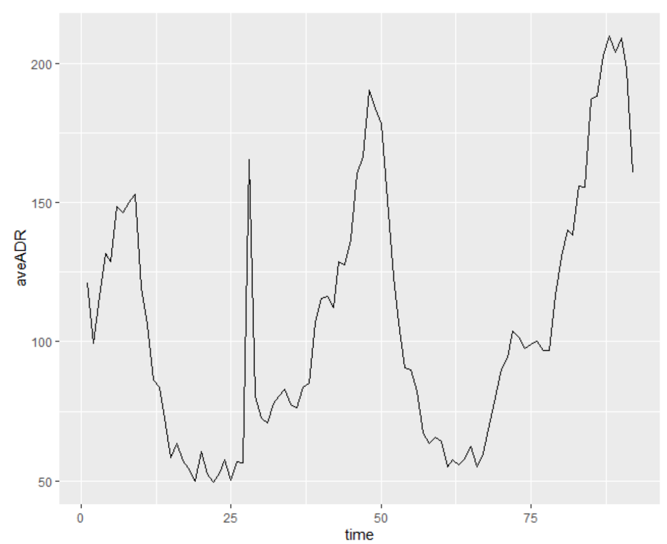
Console Terminal Jobs

E:/Education/MSBA/DS/Team Project/

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.82362    3.56416  25.202  < 2e-16 ***
LeadTime      -0.06039    0.02355  -2.565   0.0121 *
Fall          -5.25992    2.25932  -2.328   0.0223 *
Winter       -21.78372    3.06270  -7.113 3.17e-10 ***
time           0.52339    0.04250  12.315  < 2e-16 ***
children      43.76350   19.05402   2.297   0.0241 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.231 on 86 degrees of freedom
Multiple R-squared:  0.7911,	Adjusted R-squared:  0.7789
F-statistic: 65.13 on 5 and 86 DF,  p-value: < 2.2e-16
```

Resort Hotel

Average ADR with time for the Resort Hotel



Linear Model of Average ADR for the resort Hotel

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -128.29442   56.73758  -2.261 0.026687 *  
LeadTime          0.26685    0.07381   3.615 0.000545 ***
Summer           10.97922    5.05040   2.174 0.032909 *  
Winter          -17.19527    5.70446  -3.014 0.003526 ** 
Contract        -32.05892   13.67291  -2.345 0.021727 *  
Group           -30.52707   13.71877  -2.225 0.029115 *  
Transient       -31.94643   13.67679  -2.336 0.022213 *  
TransientParty  -31.95263   13.67381  -2.337 0.022160 *  
adults           94.96314   32.31497   2.939 0.004394 ** 
children        286.65374   38.93302   7.363 2.06e-10 ***
babies          395.02332  188.22815   2.099 0.039261 *  
BB               -0.19360    0.07003  -2.764 0.007195 ** 
Groups           -0.30180    0.10071  -2.997 0.003714 ** 
OfflineTATO      -0.38460    0.13671  -2.813 0.006278 ** 
OnlineTA         -0.28635    0.12504  -2.290 0.024875 *  
DCCorporate      32.26931   13.66935   2.361 0.020876 *  
DCDirect         32.26133   13.66418   2.361 0.020860 *  
DCTATO           32.27841   13.67165   2.361 0.020862 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.05 on 74 degrees of freedom
Multiple R-squared:  0.9322,	Adjusted R-squared:  0.9167
```

In conclusion, we can assume that overtime the average revenue increases naturally. Customer age group and seasonality influence hotels' revenue in both city and resort but to different degrees. All three groups affect average revenue in Resort but only the number of children helped increase average revenue in City. Similarly, lead time just has an impact on hotels in the city.

<u>SVM Model for Cancellation</u>

For the SVM model we are using Cancellations as our target variable and trying to predict it using the following variables.

```
city_svm <- data.frame(Cancel=city$IsCanceled,
                       LeadTime=city$LeadTime,
                       StaysWkn=city$StaysInWeekendNights,
                       StaysWk=city$StaysInWeekNights,
```

```
                Adults=city$Adults,
                Children=as.numeric(city$Children),
                Babies=city$Babies,
                RepeatedGuest=city$IsRepeatedGuest,
                PrevCancellations=city$PreviousCancellations,
                PrevBookingsNotCancelled=city$PreviousBookingsNotCanceled,
                BookingChanges=city$BookingChanges,
                DaysWaitingList=city$DaysInWaitingList,
                ADR=city$ADR,
                ReqParkingSpaces=city$RequiredCarParkingSpaces,
                NumSpecialRequests=city$TotalOfSpecialRequests,
                StayNights=city$StayNights,
                AvgRevStay=city$AvgRevPStay
)
```

We used a 60-40 proportion for train-test split, Cross-Validation of 5 folds, and cost as 5. For the city hotel, we have not included seasons as it has NA values and SVM doesn't make any predictions for cases where any independent variable is an NA. The summary of the model is presented in the figure below.

City Hotel

```
> csvmOut
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.102755229730916

Number of Support Vectors : 24589

Objective Function Value : -112537.8
Training error : 0.213799
Cross validation error : 0.227918
Probability model included.
```

The city hotel SVM model has a 21.3% training error, and a cross validation error of 22.79% against the 5 fold cross validation. These are accurate numbers, and we can use this model relatively confidently when trying to predict the cancellation outcome.

```
Confusion Matrix and Statistics


csvmPred     0     1
       0 15969  4615
       1  2522  8624

               Accuracy : 0.7751
                 95% CI : (0.7704, 0.7797)
    No Information Rate : 0.5828
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5268

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

            Sensitivity : 0.8636
            Specificity : 0.6514
         Pos Pred Value : 0.7758
         Neg Pred Value : 0.7737
             Prevalence : 0.5828
         Detection Rate : 0.5033
   Detection Prevalence : 0.6487
      Balanced Accuracy : 0.7575

       'Positive' Class : 0
```

Resort Hotel

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0795104953160874

Number of Support Vectors : 11697

Objective Function Value : -52562.55
Training error : 0.19861
Cross validation error : 0.214586
Probability model included.
>
```

For the resort SVM model, we can see that a 19% training error, and a 21.4% cross validation errors is produced with the variables listed above. These are both solid error numbers, and we can use this model with confidence when attempting to predict cancellation based on the factors in those categories.

```
> resort_confusion
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10913  2792
         1   662  1656

               Accuracy : 0.7844
                 95% CI : (0.778, 0.7908)
    No Information Rate : 0.7224
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3696

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

            Sensitivity : 0.9428
            Specificity : 0.3723
         Pos Pred Value : 0.7963
         Neg Pred Value : 0.7144
             Prevalence : 0.7224
         Detection Rate : 0.6811
   Detection Prevalence : 0.8553
      Balanced Accuracy : 0.6576

       'Positive' Class : 0
```

We can see the accuracy for the resort hotel is 78.4 % whereas, for the city hotel, it is 77.5%. This is the best model that we found accuracy wise, after experimenting with multiple variables. These models enhance the validation and trust we can have in our linear models, majority of the variables overlap, so we can trust the trends we see in the linear models for cancellation.

Recommendations

City

- Create repeated guest incentive program - discounts on future stays
- Offers for contract workers to book in the winter seasons - lower cancellations
- Increase booking prices for fall and winter - 56% of bookings are in winter - avg rev decreases in the winter - high demand will pay higher prices
- Capitalize on high average revenues of Family group types - develop summer vacation package that includes HB/FB meals for families specifically in rooms of type C/D
- For corporate or groups Market segment bookings with large lead times, bump them to room type D, or Meal HB

Resort

- Incentivize contract and group type customers to bring their children. Partnering with some companies which can give kids a nice experience.

- Focus on Distribution channels like Corporate, Direct in the winter season.
- Create a family holiday package for the winter season - to counteract decrease in revenue and increase in cancellations
- Capitalize on high average revenues of Family group types - develop packages that includes Meal FB and HB or bumping them to Room P
- Require non-refundable deposit for bookings made by transient customers who do not require a parking pass, have not previously cancelled a booking, and are not a repeated guest.

Conclusion

The hotel industry is very competitive and revenue and cancellations can be quite tricky to predict. Data Analysis can be a good way for businesses to gain a competitive advantage. In this report we have tried to do a comparative analysis of two hotels(City and Resort). We have compared the customer demographics and based on these categories we have compared Average revenue per stay and Cancellations. From our analysis, the city hotel is facing the problem of more cancellations and lower average revenues. In our report we have also tried to find recommendations which the management of these two hotels can use to likely lower cancellations and increase their revenue.