Predictive Data Analysis

Ukamaka Nkechi Oragwu

2023-02-07

## 1. Load and use Packages

2. Load Data and View Data

3. Data Preparation- Quality check and Cleaning

4. Data Transoformation, Feature extraction/selection

5. Data Exploration- Statistical, Graphical, Principal Component Analysis

#1. Load and use Packages Loading necessary libraries

```
library(ggplot2)
library(validate)

##
## Attaching package: 'validate'

## The following object is masked from 'package:ggplot2':
##
##     expr
```

2. Load and View Data

```
hotel_reservation <- read.csv("hotel reservation randomised.csv")

#making hotel_reservation data a data frame
hotel_reservation <- data.frame(hotel_reservation)

#Viewing the data
head(hotel_reservation)

##   Booking_ID no_of_adults no_of_children no_of_weekend_nights no_of_week_n
ights
```

```
## 1   INN00001            2              0                    1
2
## 2   INN00002            2              0                    2
3
## 3   INN00003            1              0                    2
1
## 4   INN00004            2              0                    0
2
## 5   INN00005            2              0                    1
1
## 6   INN00006            2              0                    0
2
##   type_of_meal_plan required_car_parking_space room_type_reserved lead_tim
e
## 1       Meal Plan 1                          0        Room_Type 1       22
4
## 2      Not Selected                          0        Room_Type 1        
5
## 3       Meal Plan 1                          0        Room_Type 1        
1
## 4       Meal Plan 1                          0        Room_Type 1       21
1
## 5      Not Selected                          0        Room_Type 1        4
8
## 6       Meal Plan 2                          0        Room_Type 1       34
6
##   arrival_year arrival_month arrival_date market_segment_type repeated_gue
st
## 1         2017            10            2             Offline             
0
## 2         2018            11            6              Online             
0
## 3         2018             2           28              Online             
0
## 4         2018             5           20              Online             
0
## 5         2018             4           11              Online             
0
## 6         2018             9           13              Online             
0
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
## 1                            0                                    0
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            0                                    0
## 6                            0                                    0
##   no_of_special_requests booking_status avg_price_per_room
## 1                      0   Not_Canceled              65.00
## 2                      1   Not_Canceled             106.68
## 3                      0       Canceled              60.00
## 4                      0       Canceled             100.00
## 5                      0       Canceled              94.50
## 6                      1       Canceled             115.00
```

```
summary(hotel_reservation)
```

```
##   Booking_ID         no_of_adults   no_of_children    no_of_weekend_nights
##  Length:36275       Min.   :0.000   Min.   : 0.0000   Min.   :0.0000
##  Class :character   1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Mode  :character   Median :2.000   Median : 0.0000   Median :1.0000
##                     Mean   :1.845   Mean   : 0.1053   Mean   :0.8107
##                     3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000
##                     Max.   :4.000   Max.   :10.0000   Max.   :7.0000
##
##  no_of_week_nights type_of_meal_plan  required_car_parking_space
##  Min.   : 0.000    Length:36275       Min.   :0.00000
##  1st Qu.: 1.000    Class :character   1st Qu.:0.00000
##  Median : 2.000    Mode  :character   Median :0.00000
##  Mean   : 2.204                       Mean   :0.03099
##  3rd Qu.: 3.000                       3rd Qu.:0.00000
##  Max.   :17.000                       Max.   :1.00000
##
##  room_type_reserved   lead_time       arrival_year  arrival_month
##  Length:36275       Min.   :  0.00   Min.   :2017   Min.   : 1.000
##  Class :character   1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000
##  Mode  :character   Median : 57.00   Median :2018   Median : 8.000
##                     Mean   : 85.23   Mean   :2018   Mean   : 7.424
##                     3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000
##                     Max.   :443.00   Max.   :2018   Max.   :12.000
##
##   arrival_date  market_segment_type repeated_guest
##  Min.   : 1.0   Length:36275        Min.   :0.00000
##  1st Qu.: 8.0   Class :character    1st Qu.:0.00000
##  Median :16.0   Mode  :character    Median :0.00000
##  Mean   :15.6                       Mean   :0.02564
##  3rd Qu.:23.0                       3rd Qu.:0.00000
##  Max.   :31.0                       Max.   :1.00000
##
##  no_of_previous_cancellations no_of_previous_bookings_not_canceled
##  Min.   : 0.00000             Min.   : 0.0000
##  1st Qu.: 0.00000             1st Qu.: 0.0000
##  Median : 0.00000             Median : 0.0000
##  Mean   : 0.02335             Mean   : 0.1534
##  3rd Qu.: 0.00000             3rd Qu.: 0.0000
##  Max.   :13.00000             Max.   :58.0000
##
##  no_of_special_requests booking_status     avg_price_per_room
##  Min.   :0.0000         Length:36275       Min.   :  0.00
##  1st Qu.:0.0000         Class :character   1st Qu.: 80.30
##  Median :0.0000         Mode  :character   Median : 99.45  
##  Mean   :0.6197                            Mean   :103.42  
##  3rd Qu.:1.0000                            3rd Qu.:120.00  
##  Max.   :5.0000                            Max.   :540.00  
##                                            NA's   :1
```

```
str(hotel_reservation)
```

```
## 'data.frame':    36275 obs. of  19 variables:
##  $ Booking_ID                          : chr  "INN00001" "INN00002" "INN00
003" "INN00004" ...
##  $ no_of_adults                        : int  2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights                : int  1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights                   : int  2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan                   : chr  "Meal Plan 1" "Not Selected"
"Meal Plan 1" "Meal Plan 1" ...
##  $ required_car_parking_space          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                  : chr  "Room_Type 1" "Room_Type 1"
"Room_Type 1" "Room_Type 1" ...
##  $ lead_time                           : int  224 5 1 211 48 346 34 83 121
44 ...
##  $ arrival_year                        : int  2017 2018 2018 2018 2018 201
8 2017 2018 2018 2018 ...
##  $ arrival_month                       : int  10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                        : int  2 6 28 20 11 13 15 26 6 18 .
..
##  $ market_segment_type                 : chr  "Offline" "Online" "Online"
"Online" ...
##  $ repeated_guest                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_special_requests              : int  0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status                      : chr  "Not_Canceled" "Not_Canceled
" "Canceled" "Canceled" ...
##  $ avg_price_per_room                  : num  65 106.7 60 100 94.5 ...
```

The variables of our data set were read in correctly except for the following: - type_of_meal_plan - room_type_reserved - market_segment_type - booking_status They were read in as characters instead of factors, however, these variables will be recoded to integers during data transformation so we will leave them as characters for now.

3. Quality Check and Cleaning

a. Detecting missing values
b. Detecting Duplicates
c. Data Validation
d. Data Cleaning

- Dealing with missing values

- Dealing with duplicates
- (Simple) outlier detection

```
#a. Detecting missing values
colSums(is.na(hotel_reservation))
```

```
##                           Booking_ID                         no_of_adults
##                                    0                                    0
##                       no_of_children                 no_of_weekend_nights
##                                    0                                    0
##                    no_of_week_nights                    type_of_meal_plan
##                                    0                                    0
##           required_car_parking_space                   room_type_reserved
##                                    0                                    0
##                            lead_time                         arrival_year
##                                    0                                    0
##                        arrival_month                         arrival_date
##                                    0                                    0
##                  market_segment_type                       repeated_guest
##                                    0                                    0
##         no_of_previous_cancellations no_of_previous_bookings_not_canceled
##                                    0                                    0
##               no_of_special_requests                       booking_status
##                                    0                                    0
##                   avg_price_per_room
##                                    1
```

```
#b. Detecting duplicate values
dim(hotel_reservation)
```

```
## [1] 36275    19
```

```
dim(unique(hotel_reservation))
```

```
## [1] 36275    19
```

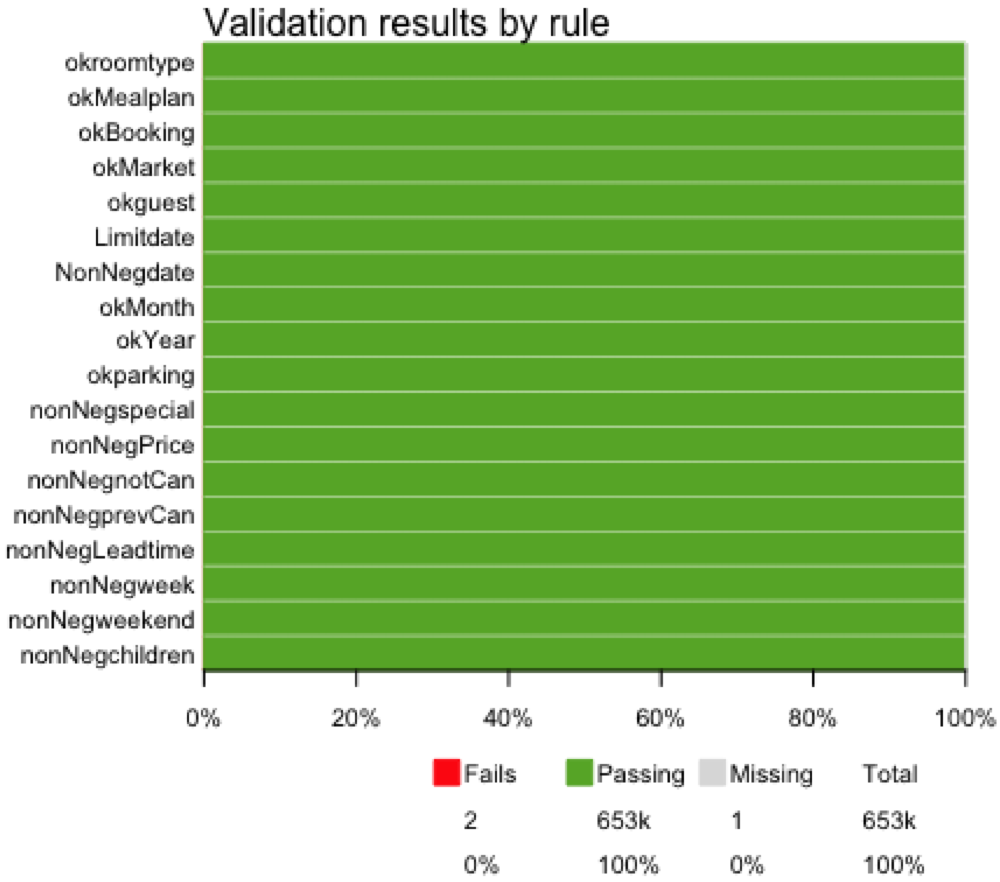
```
sum(duplicated(hotel_reservation))
```

```
## [1] 0
```

Missing Values: We recorded 1 missing value in avg_price_per_room variable. Duplicate Instances: There are no duplicates in our dataset.

```
#c. Validation Process

Mydf.Rules <- validator(
  nonNegchildren = no_of_children >=0,
  nonNegweekend = no_of_weekend_nights>=0,
  nonNegweek = no_of_week_nights>=0,
  nonNegLeadtime = lead_time>=0,
  nonNegprevCan = no_of_previous_cancellations>=0,
  nonNegnotCan = no_of_previous_bookings_not_canceled>=0,
  nonNegPrice = avg_price_per_room>=0,
  nonNegspecial = no_of_special_requests>=0,
  okMealplan= is.element(type_of_meal_plan,c("Not Selected","Meal Plan 1","Me
al Plan 2", "Meal Plan 3")),
  okparking= is.element(required_car_parking_space,c("0","1")),
  okroomtype= is.element(room_type_reserved,c("Room_Type 1","Room_Type 2", "R
oom_Type 3", "Room_Type 4", "Room_Type 5","Room_Type 6", "Room_Type 7")),
  okYear = arrival_year >= 2017 & arrival_year <=2018,
  okMonth = arrival_month> 0 & arrival_month <=12,
  NonNegdate = arrival_date>0,
  Limitdate = arrival_date<=31,
  okguest= is.element(repeated_guest,c("0","1")),
  okMarket = is.element(market_segment_type, c("Aviation", "Complementary", "
Corporate", "Offline", "Online")),
  okBooking = is.element(booking_status, c("Canceled","Not_Canceled")))

qual.check <- confront(hotel_reservation,Mydf.Rules)
summary(qual.check)

##               name items passes fails nNA error warning
## 1  nonNegchildren 36275  36275     0   0 FALSE   FALSE
## 2   nonNegweekend 36275  36275     0   0 FALSE   FALSE
## 3      nonNegweek 36275  36275     0   0 FALSE   FALSE
## 4  nonNegLeadtime 36275  36275     0   0 FALSE   FALSE
## 5   nonNegprevCan 36275  36275     0   0 FALSE   FALSE
## 6    nonNegnotCan 36275  36275     0   0 FALSE   FALSE
## 7     nonNegPrice 36275  36274     0   1 FALSE   FALSE
## 8   nonNegspecial 36275  36275     0   0 FALSE   FALSE
## 9      okMealplan 36275  36274     1   0 FALSE   FALSE
## 10      okparking 36275  36275     0   0 FALSE   FALSE
## 11     okroomtype 36275  36274     1   0 FALSE   FALSE
## 12         okYear 36275  36275     0   0 FALSE   FALSE
## 13        okMonth 36275  36275     0   0 FALSE   FALSE
## 14     NonNegdate 36275  36275     0   0 FALSE   FALSE
## 15      Limitdate 36275  36275     0   0 FALSE   FALSE
## 16        okguest 36275  36275     0   0 FALSE   FALSE
## 17       okMarket 36275  36275     0   0 FALSE   FALSE
## 18      okBooking 36275  36275     0   0 FALSE   FALSE
##
expression
## 1
no_of_children - 0 >= -1e-08
## 2
no_of_weekend_nights - 0 >= -1e-08
## 3
no_of_week_nights - 0 >= -1e-08
## 4
lead_time - 0 >= -1e-08
## 5
no_of_previous_cancellations - 0 >= -1e-08
```

```
## 6
no_of_previous_bookings_not_canceled - 0 >= -1e-08
## 7
avg_price_per_room - 0 >= -1e-08
## 8
no_of_special_requests - 0 >= -1e-08
## 9                                               is.element(type_of_meal_pl
an, c("Not Selected", "Meal Plan 1", "Meal Plan 2", "Meal Plan 3"))
## 10
is.element(required_car_parking_space, c("0", "1"))
## 11 is.element(room_type_reserved, c("Room_Type 1", "Room_Type 2", "Room_Ty
pe 3", "Room_Type 4", "Room_Type 5", "Room_Type 6", "Room_Type 7"))
## 12
arrival_year - 2017 >= -1e-08 & arrival_year - 2018 <= 1e-08
## 13
arrival_month > 0 & arrival_month - 12 <= 1e-08
## 14
arrival_date > 0
## 15
arrival_date - 31 <= 1e-08
## 16
is.element(repeated_guest, c("0", "1"))
## 17                                          is.element(market_segment_type
, c("Aviation", "Complementary", "Corporate", "Offline", "Online"))
## 18
is.element(booking_status, c("Canceled", "Not_Canceled"))

plot(qual.check, xlab="")
```



Here we see that there are 2 rules that failed our validation test and these are in the room_type_reserved and meal_plan_type variables. And 1 missing value in the avg_room_price variable.

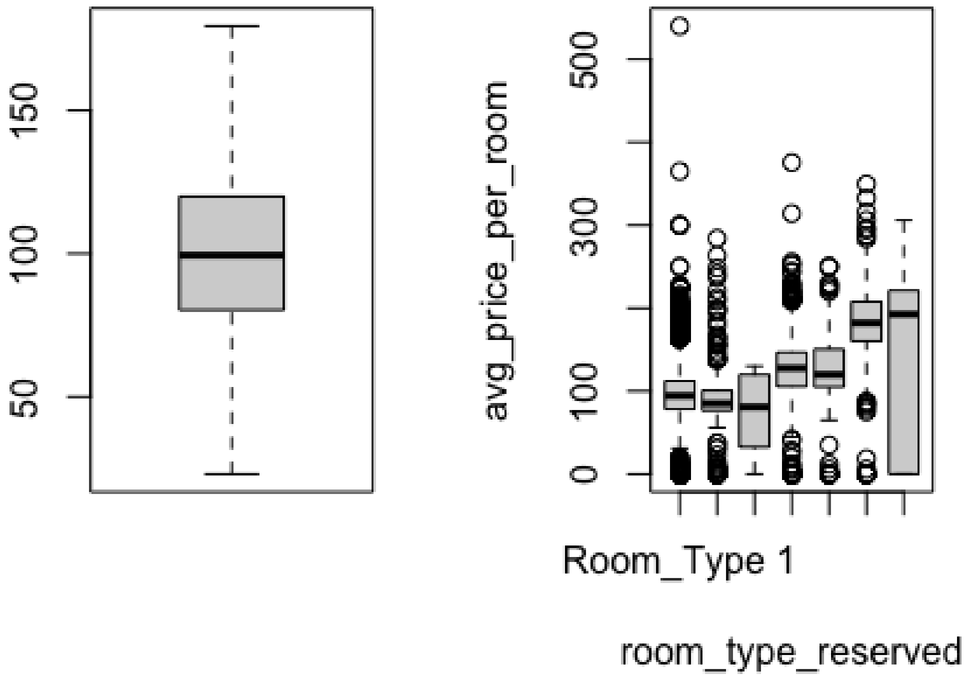
```
#investigating the failure in room_type_reserved and meal_plan_type
table(hotel_reservation$type_of_meal_plan)

##
##   Meal Plan 1  Meal Plan 2  Meal Plan 3   MealPlan 1 Not Selected
##         27834         3305            5            1         5130

table(hotel_reservation$room_type_reserved)

##
## Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5 Room_Type 6
##       28129         692           7        6057         265         966
## Room_Type 7  RoomType 1
##         158           1
```

We can see that there was a wrong spelling for Meal Plan 1 as MealPlan 1, and Room_Type 1 was misspelt as RoomType1.

```
# d. Data Cleaning
# Fixing the wrong spelling
hotel_reservation$room_type_reserved[hotel_reservation$room_type_reserved ==
"RoomType 1"] <- "Room_Type 1"
table(hotel_reservation$room_type_reserved)

##
## Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5 Room_Type 6
##       28130         692           7        6057         265         966
## Room_Type 7
##         158

hotel_reservation$type_of_meal_plan[hotel_reservation$type_of_meal_plan == "M
ealPlan 1"] <- "Meal Plan 1"
table(hotel_reservation$type_of_meal_plan)

##
##  Meal Plan 1  Meal Plan 2  Meal Plan 3 Not Selected
##        27835         3305            5         5130

# Fixing the missing value
hotel_reservation$avg_price_per_room <- as.numeric(hotel_reservation$avg_pric
e_per_room)
hotel_reservation$avg_price_per_room[hotel_reservation$avg_price_per_room ==
" "] <- NA #Recoding missing value

hotel_reserve_noNA <- hotel_reservation #Creating new data frame before imput
ing
hotel_reserve_noNA$avg_price_per_room[is.na(hotel_reserve_noNA$avg_price_per_
room)] <- median(hotel_reserve_noNA$avg_price_per_room, na.rm = T)
summary(hotel_reserve_noNA$avg_price_per_room)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   80.30   99.45  103.42  120.00  540.00
```

Our variables are correctly represented now

```
# (Simple) Outlier Detection for the target variable

# inspect the Fare distribution using summary statistics
summary(hotel_reserve_noNA$avg_price_per_room)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   80.30   99.45  103.42  120.00  540.00

# generate a boxplot of the avg_price_per_room variable
#png(file = "hotelreserve boxplot_price.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(1,2))
boxplot(hotel_reserve_noNA$avg_price_per_room, outline=FALSE)
boxplot(avg_price_per_room ~ room_type_reserved, data = hotel_reserve_noNA)
```



```
par(opar)
#dev.off()
```

We do not see any outliers in the average room price alone but when compared with other variables we see that one price instance is significantly different from the rest and may be skewing the data. We will take a closer look at this outlier in the EDA before deciding if to take it out.

#3.Feature Selection/extraction The booking_ID column does not add significant information to our dataset so we will be dropping this column

```
#dropping booking_ID column
hotel_reserve_noNA <- hotel_reserve_noNA[,-1]
head(hotel_reserve_noNA)

##   no_of_adults no_of_children no_of_weekend_nights no_of_week_nights
## 1            2              0                    1                 2
## 2            2              0                    2                 3
## 3            1              0                    2                 1
## 4            2              0                    0                 2
## 5            2              0                    1                 1
## 6            2              0                    0                 2
##   type_of_meal_plan required_car_parking_space room_type_reserved lead_tim
e
```

```
## 1       Meal Plan 1                                    0        Room_Type 1       22
4
## 2      Not Selected                                    0        Room_Type 1
5
## 3       Meal Plan 1                                    0        Room_Type 1
1
## 4       Meal Plan 1                                    0        Room_Type 1       21
1
## 5      Not Selected                                    0        Room_Type 1        4
8
## 6       Meal Plan 2                                    0        Room_Type 1       34
6
##   arrival_year arrival_month arrival_date market_segment_type repeated_gue
st
## 1         2017            10            2             Offline
0
## 2         2018            11            6              Online
0
## 3         2018             2           28              Online
0
## 4         2018             5           20              Online
0
## 5         2018             4           11              Online
0
## 6         2018             9           13              Online
0
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
## 1                            0                                    0
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            0                                    0
## 6                            0                                    0
##   no_of_special_requests booking_status avg_price_per_room
## 1                      0   Not_Canceled              65.00
## 2                      1   Not_Canceled             106.68
## 3                      0       Canceled              60.00
## 4                      0       Canceled             100.00
## 5                      0       Canceled              94.50
## 6                      1       Canceled             115.00
```

#4. Exploratory Data Analysis a. Statistical Exploration

```
summary(hotel_reserve_noNA) #summary of our cleaned data
```

```
##   no_of_adults    no_of_children    no_of_weekend_nights no_of_week_nights
##  Min.   :0.000   Min.   : 0.0000   Min.   :0.0000       Min.   : 0.000
##  1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000       1st Qu.: 1.000
##  Median :2.000   Median : 0.0000   Median :1.0000       Median : 2.000
##  Mean   :1.845   Mean   : 0.1053   Mean   :0.8107       Mean   : 2.204
##  3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000       3rd Qu.: 3.000
##  Max.   :4.000   Max.   :10.0000   Max.   :7.0000       Max.   :17.000
##  type_of_meal_plan  required_car_parking_space room_type_reserved
##  Length:36275       Min.   :0.00000            Length:36275
##  Class :character   1st Qu.:0.00000            Class :character
##  Mode  :character   Median :0.00000            Mode  :character
##                     Mean   :0.03099
##                     3rd Qu.:0.00000
##                     Max.   :1.00000
##    lead_time       arrival_year  arrival_month     arrival_date
##  Min.   :  0.00   Min.   :2017   Min.   : 1.000   Min.   : 1.0
##  1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000   1st Qu.: 8.0
##  Median : 57.00   Median :2018   Median : 8.000   Median :16.0
##  Mean   : 85.23   Mean   :2018   Mean   : 7.424   Mean   :15.6
##  3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000   3rd Qu.:23.0
##  Max.   :443.00   Max.   :2018   Max.   :12.000   Max.   :31.0
##  market_segment_type repeated_guest    no_of_previous_cancellations
##  Length:36275        Min.   :0.00000   Min.   : 0.00000
##  Class :character    1st Qu.:0.00000   1st Qu.: 0.00000
##  Mode  :character    Median :0.00000   Median : 0.00000
##                      Mean   :0.02564   Mean   : 0.02335
##                      3rd Qu.:0.00000   3rd Qu.: 0.00000
##                      Max.   :1.00000   Max.   :13.00000
##  no_of_previous_bookings_not_canceled no_of_special_requests booking_statu
s
##  Min.   : 0.0000                      Min.   :0.0000         Length:36275
##  1st Qu.: 0.0000                      1st Qu.:0.0000         Class :charac
ter
##  Median : 0.0000                      Median :0.0000         Mode  :charac
ter
##  Mean   : 0.1534                      Mean   :0.6197
##  3rd Qu.: 0.0000                      3rd Qu.:1.0000
##  Max.   :58.0000                      Max.   :5.0000
##  avg_price_per_room
##  Min.   :  0.00
##  1st Qu.: 80.30
##  Median : 99.45
##  Mean   :103.42
##  3rd Qu.:120.00
##  Max.   :540.00
```

```
str(hotel_reserve_noNA)
```

```
## 'data.frame':    36275 obs. of  18 variables:
##  $ no_of_adults                        : int  2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights                : int  1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights                   : int  2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan                   : chr  "Meal Plan 1" "Not Selected"
"Meal Plan 1" "Meal Plan 1" ...
```

```
##  $ required_car_parking_space          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                  : chr  "Room_Type 1" "Room_Type 1"
"Room_Type 1" "Room_Type 1" ...
##  $ lead_time                           : int  224 5 1 211 48 346 34 83 121
44 ...
##  $ arrival_year                        : int  2017 2018 2018 2018 2018 201
8 2017 2018 2018 2018 ...
##  $ arrival_month                       : int  10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                        : int  2 6 28 20 11 13 15 26 6 18 .
..
##  $ market_segment_type                 : chr  "Offline" "Online" "Online"
"Online" ...
##  $ repeated_guest                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_special_requests              : int  0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status                      : chr  "Not_Canceled" "Not_Canceled
" "Canceled" "Canceled" ...
##  $ avg_price_per_room                  : num  65 106.7 60 100 94.5 ...
```

```r
head(hotel_reserve_noNA)
```

```
##   no_of_adults no_of_children no_of_weekend_nights no_of_week_nights
## 1            2              0                    1                 2
## 2            2              0                    2                 3
## 3            1              0                    2                 1
## 4            2              0                    0                 2
## 5            2              0                    1                 1
## 6            2              0                    0                 2
##   type_of_meal_plan required_car_parking_space room_type_reserved lead_tim
e
## 1       Meal Plan 1                          0        Room_Type 1       22
4
## 2      Not Selected                          0        Room_Type 1
5
## 3       Meal Plan 1                          0        Room_Type 1
1
## 4       Meal Plan 1                          0        Room_Type 1       21
1
## 5      Not Selected                          0        Room_Type 1        4
8
## 6       Meal Plan 2                          0        Room_Type 1       34
6
##   arrival_year arrival_month arrival_date market_segment_type repeated_gue
st
## 1         2017            10            2             Offline
0
## 2         2018            11            6              Online
0
## 3         2018             2           28              Online
```

```
0
## 4         2018              5           20              Online
0
## 5         2018              4           11              Online
0
## 6         2018              9           13              Online
0
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
## 1                            0                                    0
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            0                                    0
## 6                            0                                    0
##   no_of_special_requests booking_status avg_price_per_room
## 1                      0   Not_Canceled              65.00
## 2                      1   Not_Canceled             106.68
## 3                      0       Canceled              60.00
## 4                      0       Canceled             100.00
## 5                      0       Canceled              94.50
## 6                      1       Canceled             115.00
```

```
# creating label vectors for numerical and categorical variables
hotel_reservation_num <- c("no_of_adults", "no_of_children", "no_of_weekend_n
ights", "no_of_week_nights","lead_time", "no_of_previous_cancellations", "no_
of_previous_bookings_not_canceled", "avg_price_per_room", "no_of_special_requ
ests")

hotel_reservation_cat <- c("type_of_meal_plan", "required_car_parking_space",
"room_type_reserved",  "market_segment_type","repeated_guest", "booking_statu
s","arrival_date","arrival_year","arrival_month")
```

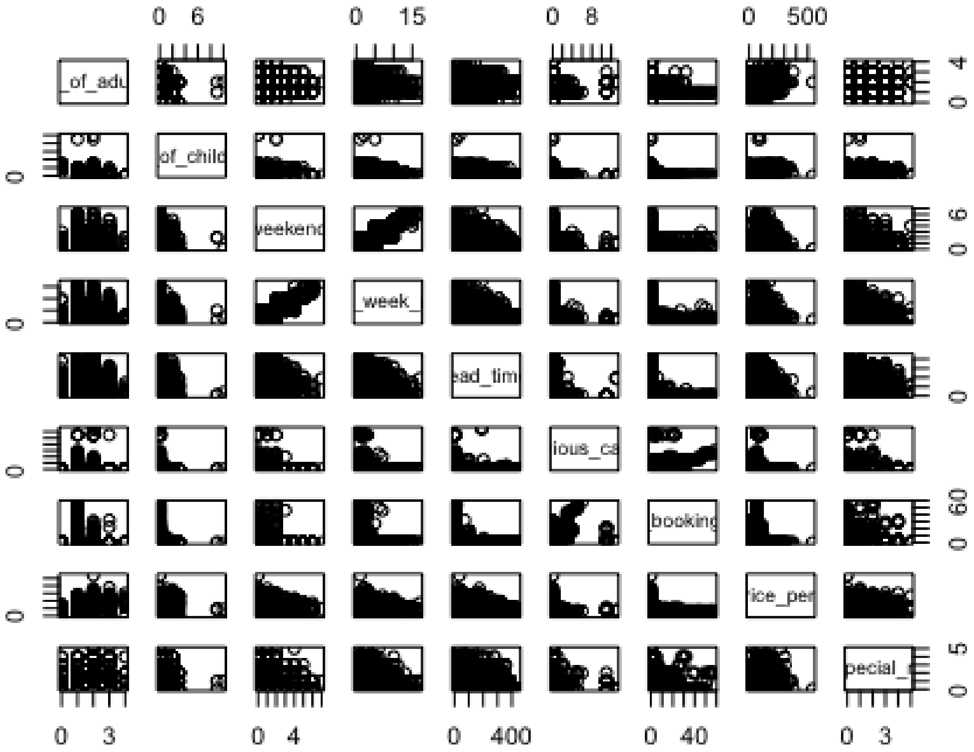
Visualizing Numerical Data

```
# exploring relationships among features: correlation matrix
hotel_reservation_num_cor <- cor(hotel_reserve_noNA[hotel_reservation_num])

# visualize the correlation matrix
hotel_reservation_num_cor
```

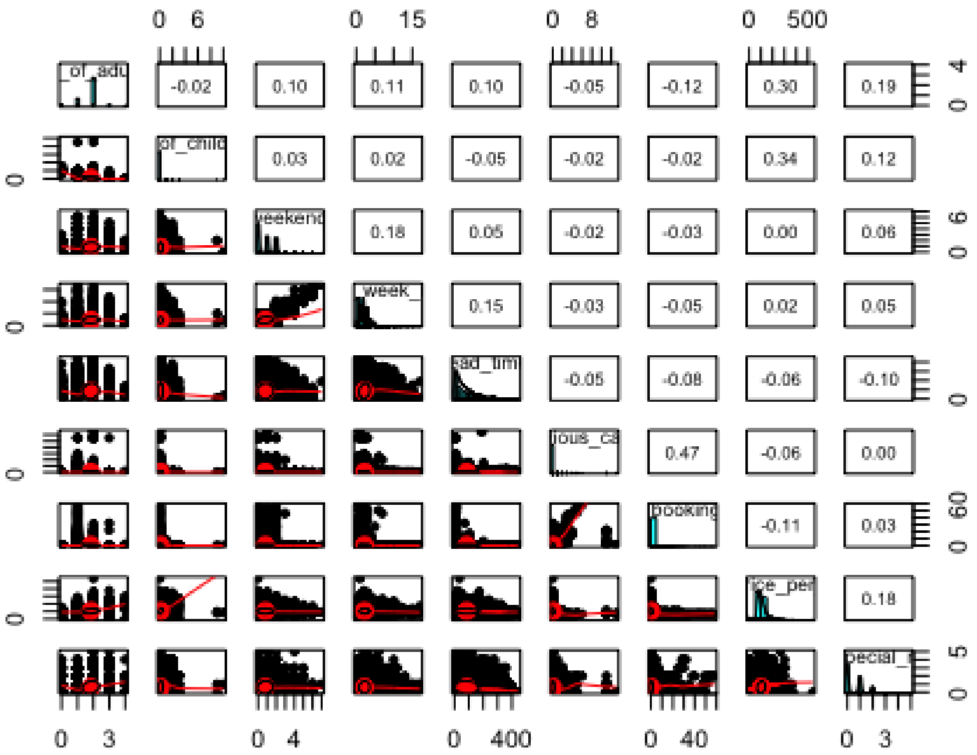
```
##                                      no_of_adults no_of_children
## no_of_adults                           1.00000000    -0.01978707
## no_of_children                        -0.01978707     1.00000000
## no_of_weekend_nights                   0.10331578     0.02947758
## no_of_week_nights                      0.10562190     0.02439811
## lead_time                              0.09728651    -0.04709128
## no_of_previous_cancellations          -0.04742575    -0.01638958
## no_of_previous_bookings_not_canceled  -0.11916579    -0.02118896
## avg_price_per_room                     0.29688259     0.33773135
## no_of_special_requests                 0.18940095     0.12448619
##                                      no_of_weekend_nights no_of_week_night
```

```
s
## no_of_adults                                0.103315775        0.1056219
0
## no_of_children                              0.029477584        0.0243981
1
## no_of_weekend_nights                        1.000000000        0.1795767
6
## no_of_week_nights                           0.179576764        1.0000000
0
## lead_time                                   0.046595440        0.1496501
6
## no_of_previous_cancellations               -0.020690482       -0.0300804
0
## no_of_previous_bookings_not_canceled       -0.026311984       -0.0493437
4
## avg_price_per_room                         -0.004513731        0.0227626
7
## no_of_special_requests                      0.060592526        0.0459936
5
##                                        lead_time no_of_previous_cancellati
ons
## no_of_adults                          0.09728651                 -0.047425
747
## no_of_children                       -0.04709128                 -0.016389
584
## no_of_weekend_nights                  0.04659544                 -0.020690
482
## no_of_week_nights                     0.14965016                 -0.030080
402
## lead_time                             1.00000000                 -0.045722
982
## no_of_previous_cancellations         -0.04572298                  1.000000
000
## no_of_previous_bookings_not_canceled -0.07813666                  0.468146
833
## avg_price_per_room                   -0.06260275                 -0.063339
719
## no_of_special_requests               -0.10164497                 -0.003317
358
##                                      no_of_previous_bookings_not_canceled
## no_of_adults                                                  -0.11916579
## no_of_children                                                -0.02118896
## no_of_weekend_nights                                          -0.02631198
## no_of_week_nights                                             -0.04934374
## lead_time                                                     -0.07813666
## no_of_previous_cancellations                                   0.46814683
## no_of_previous_bookings_not_canceled                           1.00000000
## avg_price_per_room                                            -0.11368297
## no_of_special_requests                                         0.02737658
##                                      avg_price_per_room no_of_special_requ
```

```
ests
## no_of_adults                              0.296882590            0.18940
0951
## no_of_children                            0.337731352            0.12448
6186
## no_of_weekend_nights                     -0.004513731            0.06059
2526
## no_of_week_nights                         0.022762671            0.04599
3653
## lead_time                                -0.062602751           -0.10164
4974
## no_of_previous_cancellations             -0.063339719           -0.00331
7358
## no_of_previous_bookings_not_canceled     -0.113682967            0.02737
6578
## avg_price_per_room                        1.000000000            0.18437
5523
## no_of_special_requests                    0.184375523            1.00000
0000

# plot the relationships among features - scatterplot matrix
pairs(hotel_reserve_noNA[hotel_reservation_num])
```



```
# plot a more informative scatterplot matrix
#png(file = "hotelreserve pairs plot.png")
psych::pairs.panels(hotel_reserve_noNA[hotel_reservation_num])
```



```
#dev.off()
```

There are no significant correlations between the numerical variables

```
head(hotel_reserve_noNA)

##   no_of_adults no_of_children no_of_weekend_nights no_of_week_nights
## 1            2              0                    1                 2
## 2            2              0                    2                 3
## 3            1              0                    2                 1
## 4            2              0                    0                 2
## 5            2              0                    1                 1
## 6            2              0                    0                 2
##   type_of_meal_plan required_car_parking_space room_type_reserved lead_tim
e
## 1       Meal Plan 1                          0        Room_Type 1       22
4
## 2      Not Selected                          0        Room_Type 1
5
## 3       Meal Plan 1                          0        Room_Type 1
```

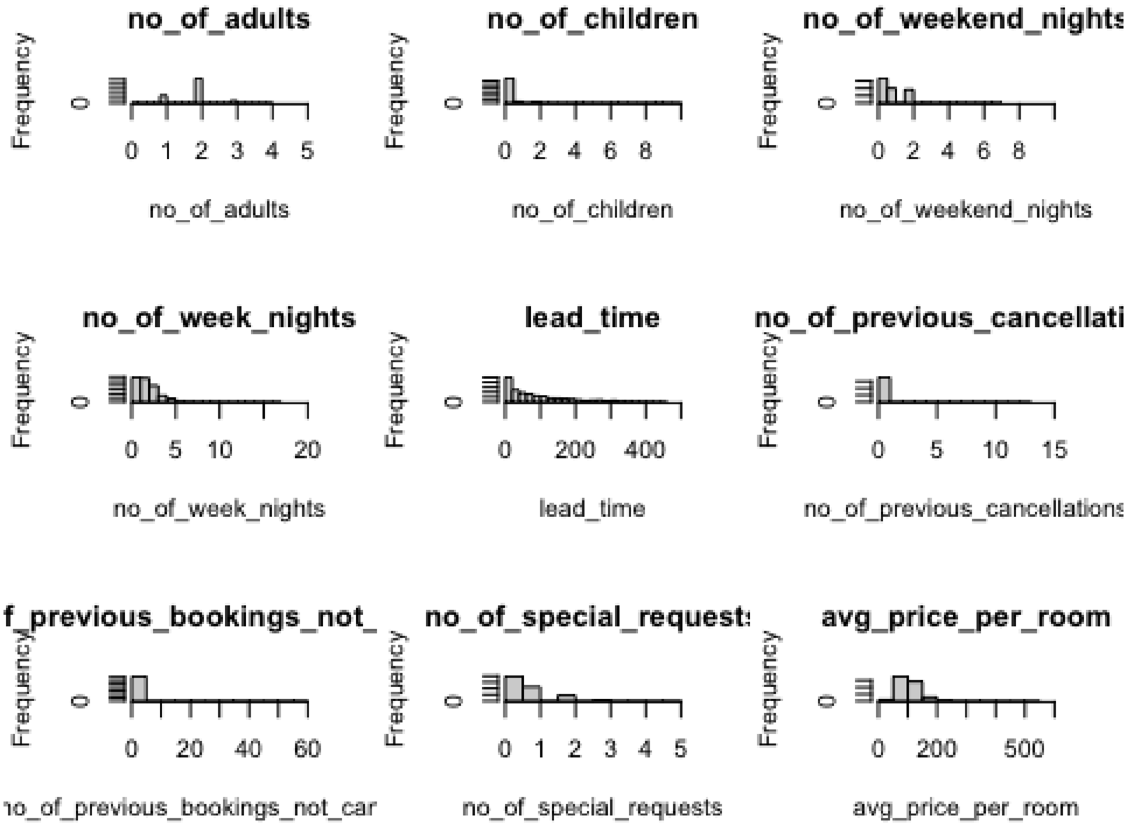
```
1
## 4       Meal Plan 1                          0        Room_Type 1       21
1
## 5      Not Selected                          0        Room_Type 1        4
8
## 6       Meal Plan 2                          0        Room_Type 1       34
6
##   arrival_year arrival_month arrival_date market_segment_type repeated_gue
st
## 1         2017            10            2             Offline
0
## 2         2018            11            6              Online
0
## 3         2018             2           28              Online
0
## 4         2018             5           20              Online
0
## 5         2018             4           11              Online
0
## 6         2018             9           13              Online
0
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
## 1                            0                                    0
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            0                                    0
## 6                            0                                    0
##   no_of_special_requests booking_status avg_price_per_room
## 1                      0   Not_Canceled              65.00
## 2                      1   Not_Canceled             106.68
## 3                      0       Canceled              60.00
## 4                      0       Canceled             100.00
## 5                      0       Canceled              94.50
## 6                      1       Canceled             115.00
```

```
summary(hotel_reserve_noNA)
```

```
##   no_of_adults   no_of_children    no_of_weekend_nights no_of_week_nights
##  Min.   :0.000   Min.   : 0.0000   Min.   :0.0000       Min.   : 0.000  
##  1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000       1st Qu.: 1.000  
##  Median :2.000   Median : 0.0000   Median :1.0000       Median : 2.000  
##  Mean   :1.845   Mean   : 0.1053   Mean   :0.8107       Mean   : 2.204  
##  3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000       3rd Qu.: 3.000  
##  Max.   :4.000   Max.   :10.0000   Max.   :7.0000       Max.   :17.000  
##  type_of_meal_plan  required_car_parking_space room_type_reserved
##  Length:36275       Min.   :0.00000            Length:36275      
##  Class :character   1st Qu.:0.00000            Class :character  
##  Mode  :character   Median :0.00000            Mode  :character  
##                     Mean   :0.03099                              
##                      3rd Qu.:0.00000
##                      Max.   :1.00000
##    lead_time       arrival_year  arrival_month     arrival_date
##  Min.   :  0.00   Min.   :2017   Min.   : 1.000   Min.   : 1.0
##  1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000   1st Qu.: 8.0
##  Median : 57.00   Median :2018   Median : 8.000   Median :16.0
##  Mean   : 85.23   Mean   :2018   Mean   : 7.424   Mean   :15.6
##  3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000   3rd Qu.:23.0
##  Max.   :443.00   Max.   :2018   Max.   :12.000   Max.   :31.0
##  market_segment_type repeated_guest    no_of_previous_cancellations
##  Length:36275        Min.   :0.00000   Min.   : 0.00000
##  Class :character    1st Qu.:0.00000   1st Qu.: 0.00000
##  Mode  :character    Median :0.00000   Median : 0.00000
##                      Mean   :0.02564   Mean   : 0.02335
##                      3rd Qu.:0.00000   3rd Qu.: 0.00000
##                      Max.   :1.00000   Max.   :13.00000
##  no_of_previous_bookings_not_canceled no_of_special_requests booking_statu
s
##  Min.   : 0.0000                      Min.   :0.0000         Length:36275
##  1st Qu.: 0.0000                      1st Qu.:0.0000         Class :charac
ter
##  Median : 0.0000                      Median :0.0000         Mode  :charac
ter
##  Mean   : 0.1534                      Mean   :0.6197
##  3rd Qu.: 0.0000                      3rd Qu.:1.0000
##  Max.   :58.0000                      Max.   :5.0000
##  avg_price_per_room
##  Min.   :  0.00
##  1st Qu.: 80.30
##  Median : 99.45
##  Mean   :103.42
##  3rd Qu.:120.00
##  Max.   :540.00
```

Independent graphical views of the numeric variables:

```
#png(file = "hotelreserve histogram plots.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(3,3)) #since we have 9 plots to show we use a 3x3 matrix
hist(hotel_reserve_noNA[, 1], main = names(hotel_reserve_noNA)[1], xlab = nam
es(hotel_reserve_noNA)[1], xlim = c(0,5))
hist(hotel_reserve_noNA[, 2], main = names(hotel_reserve_noNA)[2], xlab = nam
es(hotel_reserve_noNA)[2], xlim = c(0,10))
hist(hotel_reserve_noNA[, 3], main = names(hotel_reserve_noNA)[3], xlab = nam
es(hotel_reserve_noNA)[3], xlim = c(0,10))
hist(hotel_reserve_noNA[, 4], main = names(hotel_reserve_noNA)[4], xlab = nam
es(hotel_reserve_noNA)[4], xlim = c(0,20))
hist(hotel_reserve_noNA[, 8], main = names(hotel_reserve_noNA)[8], xlab = nam
es(hotel_reserve_noNA)[8], xlim = c(0,500))
hist(hotel_reserve_noNA[, 14], main = names(hotel_reserve_noNA)[14], xlab = n
ames(hotel_reserve_noNA)[14], xlim = c(0,15))
hist(hotel_reserve_noNA[, 15], main = names(hotel_reserve_noNA)[15], xlab = n
ames(hotel_reserve_noNA)[15], xlim = c(0,60))
hist(hotel_reserve_noNA[, 16], main = names(hotel_reserve_noNA)[16], xlab = n
ames(hotel_reserve_noNA)[16], xlim = c(0,5))
hist(hotel_reserve_noNA[, 18], main = names(hotel_reserve_noNA)[18], xlab = n
ames(hotel_reserve_noNA)[18], xlim = c(0,600))
```



```
par(opar)
#dev.off()
```

For the average price per room, our histogram looks skewed to the right. This will be investigated further when the price is compared to other variables.

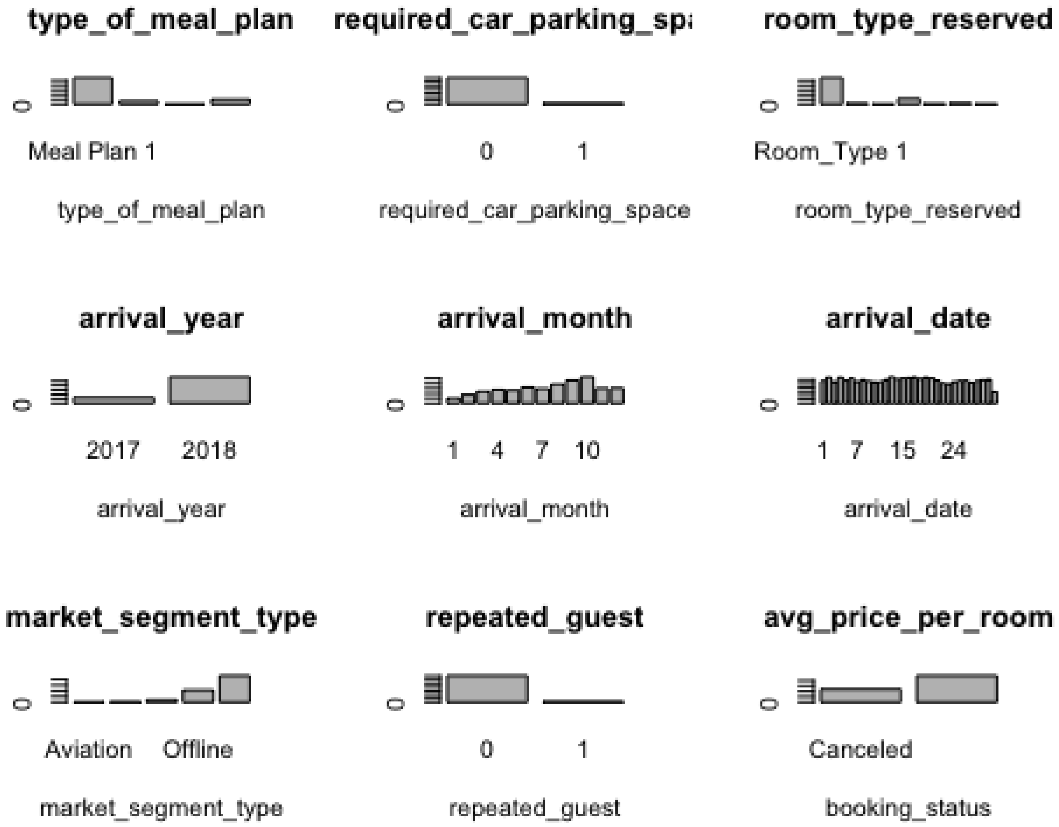
Categorical Data

```
#  frequency tables for each categorical variable
hotel_reservation_cat_table <- apply(hotel_reserve_noNA[,c("type_of_meal_plan
", "required_car_parking_space", "room_type_reserved", "arrival_year","arriva
l_date","arrival_month", "market_segment_type", "repeated_guest", "booking_st
atus")], 2, table)

# visualize the table
hotel_reservation_cat_table
```

```
## $type_of_meal_plan
##
##  Meal Plan 1  Meal Plan 2  Meal Plan 3 Not Selected
##        27835         3305            5         5130
##
## $required_car_parking_space
##
##     0     1
## 35151  1124
##
## $room_type_reserved
##
## Room_Type 1 Room_Type 2 Room_Type 3 Room_Type 4 Room_Type 5 Room_Type 6
##       28130         692           7        6057         265         966
## Room_Type 7
##         158
##
## $arrival_year
##
##  2017  2018
##  6514 29761
##
## $arrival_date
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
16
## 1133 1331 1098 1327 1154 1273 1110 1198 1130 1089 1098 1204 1358 1242 1273
1306
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31
## 1345 1260 1327 1281 1158 1023  990 1103 1146 1146 1059 1129 1190 1216  578
##
## $arrival_month
##
##    1    2    3    4    5    6    7    8    9   10   11   12
## 1014 1704 2358 2736 2598 3203 2920 3813 4611 5317 2980 3021
##
## $market_segment_type
##
##      Aviation Complementary     Corporate       Offline        Online
##           125           391          2017         10528         23214
##
## $repeated_guest
##
##     0     1
## 35345   930
##
## $booking_status
##
##     Canceled Not_Canceled
##        11885        24390
```

Bar plots for to analyze categorical variables individually

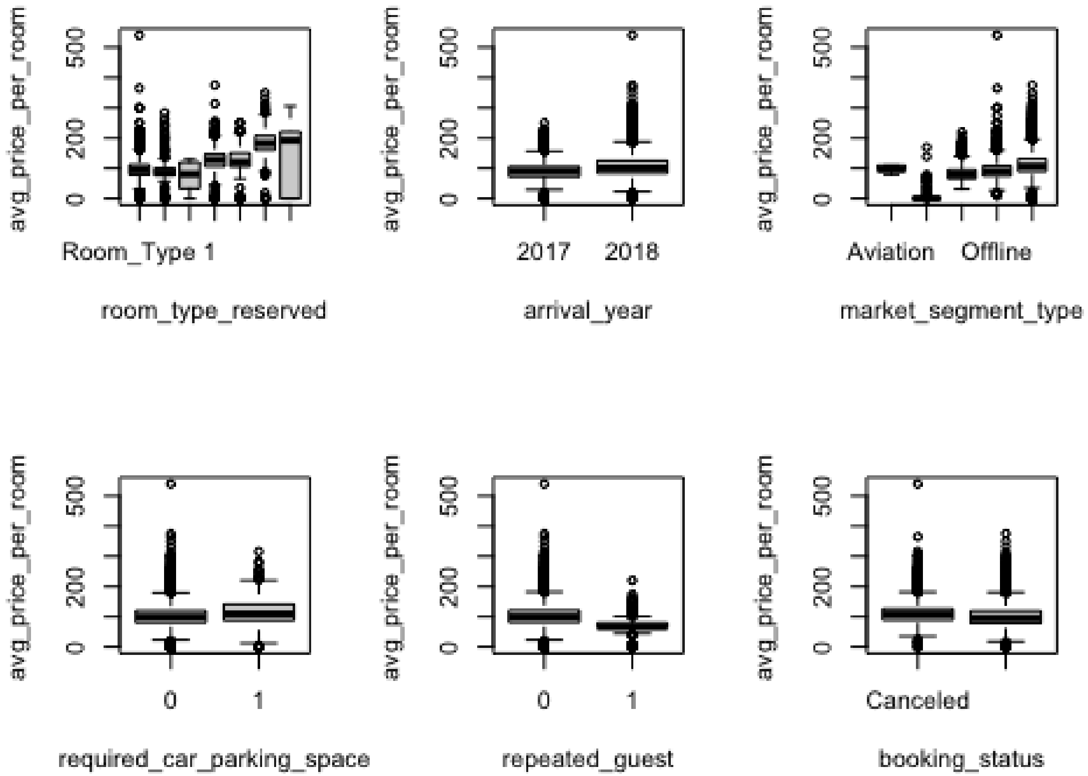
```
#png(file = "hotelreserve bar plots.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(3,3)) #since we have 7 plots to show we use a 3x3 matrix
barplot(table(hotel_reserve_noNA[, 5]), main = names(hotel_reserve_noNA)[5],
xlab = names(hotel_reserve_noNA)[5])
barplot(table(hotel_reserve_noNA[, 6]), main = names(hotel_reserve_noNA)[6],
xlab = names(hotel_reserve_noNA)[6])
barplot(table(hotel_reserve_noNA[, 7]), main = names(hotel_reserve_noNA)[7],
xlab = names(hotel_reserve_noNA)[7])
barplot(table(hotel_reserve_noNA[, 9]), main = names(hotel_reserve_noNA)[9],
xlab = names(hotel_reserve_noNA)[9])
barplot(table(hotel_reserve_noNA[, 10]), main = names(hotel_reserve_noNA)[10]
, xlab = names(hotel_reserve_noNA)[10])
barplot(table(hotel_reserve_noNA[, 11]), main = names(hotel_reserve_noNA)[11]
, xlab = names(hotel_reserve_noNA)[11])
barplot(table(hotel_reserve_noNA[, 12]), main = names(hotel_reserve_noNA)[12]
, xlab = names(hotel_reserve_noNA)[12])
barplot(table(hotel_reserve_noNA[, 13]), main = names(hotel_reserve_noNA)[13]
, xlab = names(hotel_reserve_noNA)[13])
barplot(table(hotel_reserve_noNA[, 17]), main = names(hotel_reserve_noNA)[18]
, xlab = names(hotel_reserve_noNA)[17])
```



```
par(opar)
#dev.off()
```

Comparing relationships between the average room price and other variables

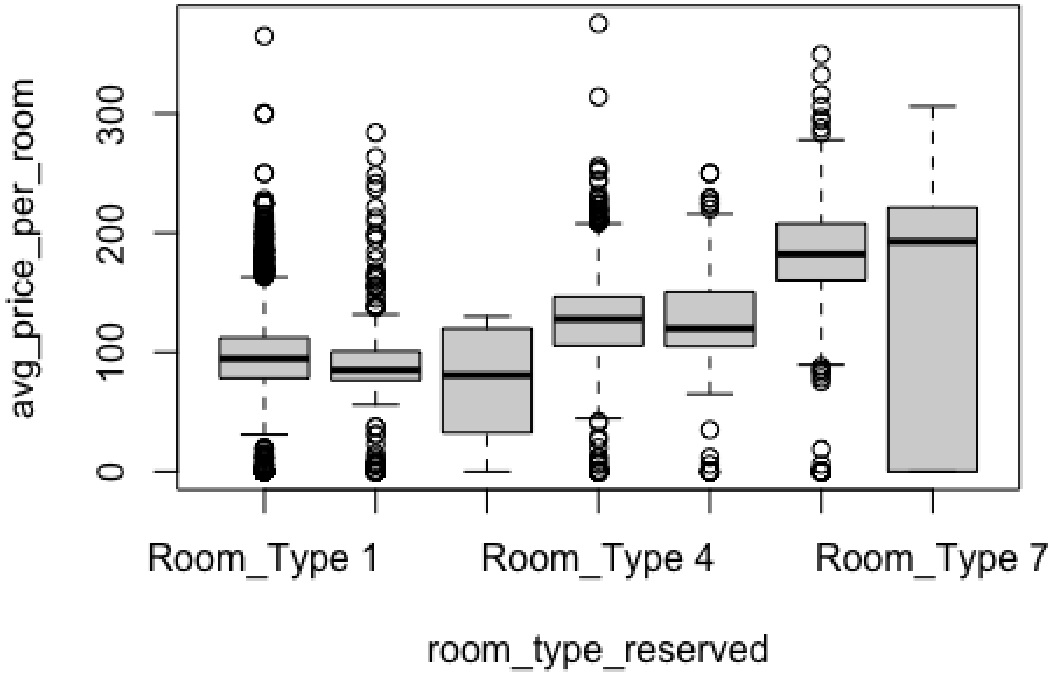
```
# plot avg_price_per_room distribution by group of categorical variables - bo
xplot
#png(file = "hotelreserve box plots price.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(2,3))
boxplot(avg_price_per_room ~ room_type_reserved, data = hotel_reserve_noNA)
boxplot(avg_price_per_room ~ arrival_year, data = hotel_reserve_noNA)
boxplot(avg_price_per_room ~ market_segment_type, data = hotel_reserve_noNA)
boxplot(avg_price_per_room ~ required_car_parking_space, data = hotel_reserve
_noNA)
boxplot(avg_price_per_room ~ repeated_guest, data = hotel_reserve_noNA)
boxplot(avg_price_per_room ~ booking_status, data = hotel_reserve_noNA)
```



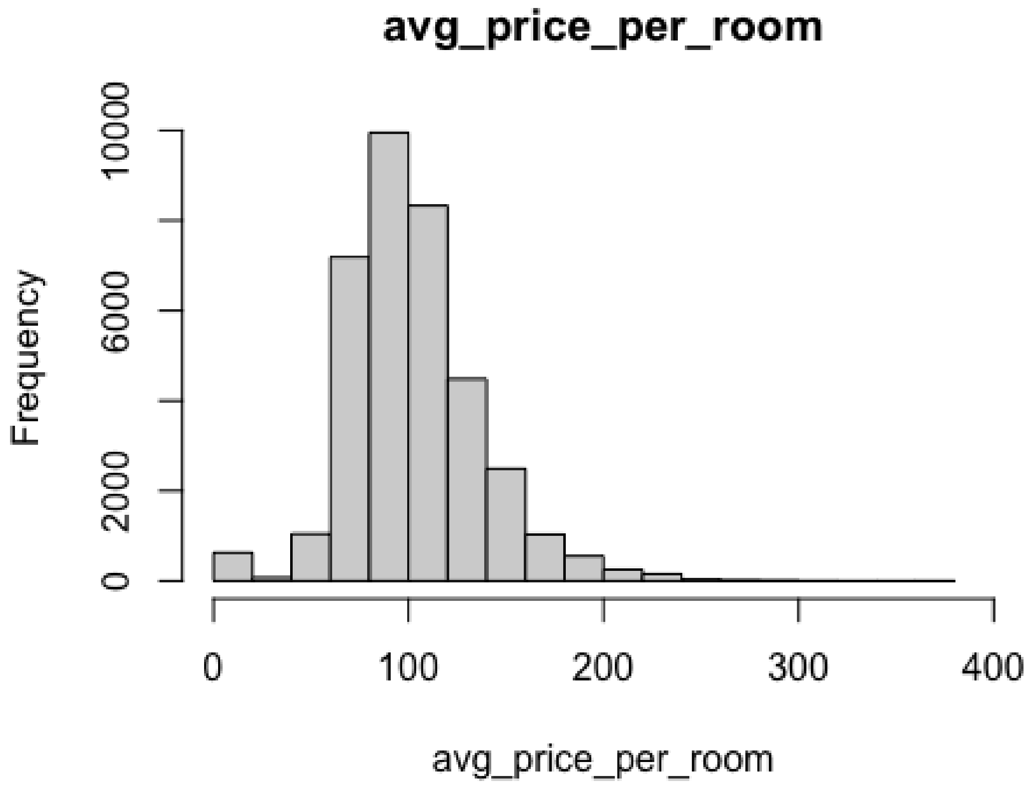
```
par(opar)
#dev.off()
```

From here we can see consistently that there are outliers. However the outlier that looks most plausible is the price above 500 which is significantly distant from the rest of the points. We will take this point out but we do not have sufficient reason to remove the other outliers as they are most likely part of our data.

```
#Removing the outlier (instance in average price greater than 500)
# outliers rows can be extracted by conditional selection
hotel_reserve_noOut <- hotel_reserve_noNA[hotel_reserve_noNA$avg_price_per_ro
om <= 500, ]
boxplot(avg_price_per_room ~ room_type_reserved, data = hotel_reserve_noOut)
```

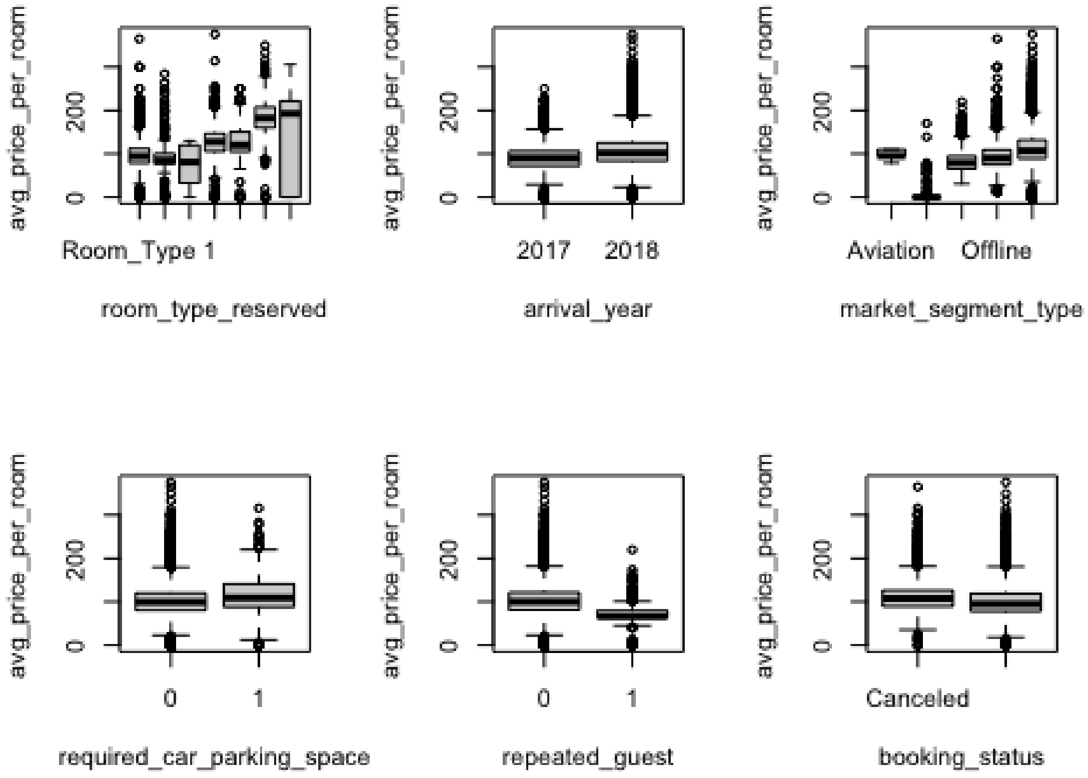


```
#visualizing the average price per room
summary(hotel_reserve_noOut$avg_price_per_room)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   80.30   99.45  103.41  120.00  375.50

hist(hotel_reserve_noOut[, 18], main = names(hotel_reserve_noOut)[18], xlab =
names(hotel_reserve_noOut)[18], xlim = c(0,400))
```



The average price per room is still a little skewed to the right but atleast better than before the outlier was removed.

```r
#comparing the relationship of price with other variables without the outlier
#png(file = "hotelreserve no Out box plots price.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(2,3))
boxplot(avg_price_per_room ~ room_type_reserved, data = hotel_reserve_noOut)
boxplot(avg_price_per_room ~ arrival_year, data = hotel_reserve_noOut)
boxplot(avg_price_per_room ~ market_segment_type, data = hotel_reserve_noOut)
boxplot(avg_price_per_room ~ required_car_parking_space, data = hotel_reserve
_noOut)
boxplot(avg_price_per_room ~ repeated_guest, data = hotel_reserve_noOut)
boxplot(avg_price_per_room ~ booking_status, data = hotel_reserve_noOut)
```

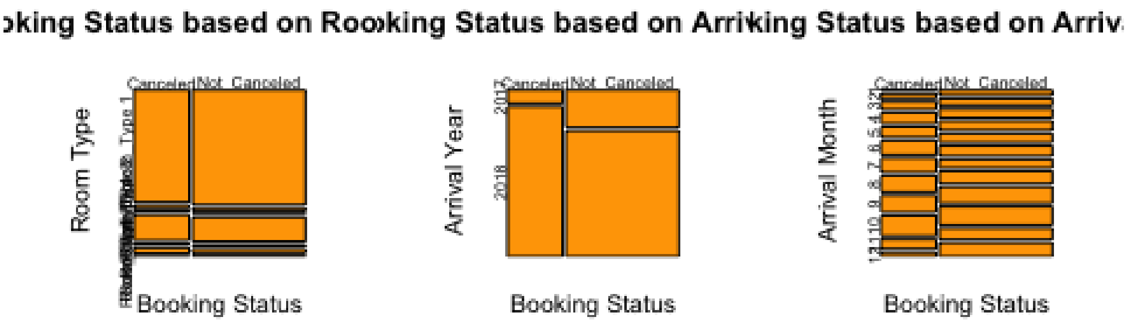


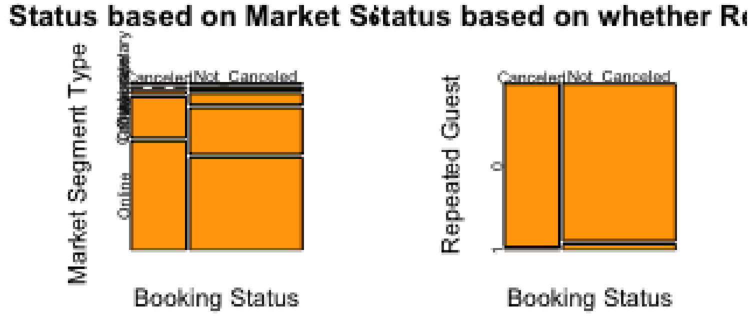
```r
par(opar)
#dev.off()
```

Our data looks good to proceed with.

Mosaic Plots - Categorical Variables against each other

```r
#png(file = "hotelreserve_noOut mosaic plots .png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(2,3))
counts <- table(hotel_reserve_noOut$booking_status, hotel_reserve_noOut$room_
type_reserved)
mosaicplot(counts, xlab='Booking Status', ylab='Room Type',main='Booking Stat
us based on Room Type', col='orange')

counts <- table(hotel_reserve_noOut$booking_status, hotel_reserve_noOut$arriv
al_year)
mosaicplot(counts, xlab='Booking Status', ylab='Arrival Year',main='Booking S
tatus based on Arrival Year', col='orange')

counts <- table(hotel_reserve_noOut$booking_status, hotel_reserve_noOut$arriv
al_month)
mosaicplot(counts, xlab='Booking Status', ylab='Arrival Month',main='Booking
Status based on Arrival Month', col='orange')

counts <- table(hotel_reserve_noOut$booking_status, hotel_reserve_noOut$marke
t_segment_type)
mosaicplot(counts, xlab='Booking Status', ylab='Market Segment Type',main='Bo
oking Status based on Market Segment Type', col='orange')

counts <- table(hotel_reserve_noOut$booking_status, hotel_reserve_noOut$repea
ted_guest)
mosaicplot(counts, xlab='Booking Status', ylab='Repeated Guest',main='Booking
Status based on whether Repeated Guest', col='orange')
par(opar)
```





```
#dev.off()
```

Data Transformation

We will now be re-coding the variables below to enable easy manipulation of our data in the following sections (PCA and Modelling) - type_of_meal_plan - room_type_reserved - market_segment_type - booking_status

```r
hotel_reserve_noOut$type_of_meal_plan[hotel_reserve_noOut$type_of_meal_plan =
= "Not Selected"] <- 0
hotel_reserve_noOut$type_of_meal_plan[hotel_reserve_noOut$type_of_meal_plan =
= "Meal Plan 1"] <- 1
hotel_reserve_noOut$type_of_meal_plan[hotel_reserve_noOut$type_of_meal_plan =
= "Meal Plan 2"] <- 2
hotel_reserve_noOut$type_of_meal_plan[hotel_reserve_noOut$type_of_meal_plan =
= "Meal Plan 3"] <- 3

hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 1"] <- 1
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 2"] <- 2
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 3"] <- 3
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 4"] <- 4
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 5"] <- 5
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 6"] <- 6
hotel_reserve_noOut$room_type_reserved[hotel_reserve_noOut$room_type_reserved
== "Room_Type 7"] <- 7

hotel_reserve_noOut$market_segment_type[hotel_reserve_noOut$market_segment_ty
pe == "Aviation"] <- 1
hotel_reserve_noOut$market_segment_type[hotel_reserve_noOut$market_segment_ty
pe == "Complementary"] <- 2
hotel_reserve_noOut$market_segment_type[hotel_reserve_noOut$market_segment_ty
pe == "Corporate"] <- 3
hotel_reserve_noOut$market_segment_type[hotel_reserve_noOut$market_segment_ty
pe == "Offline"] <- 4
hotel_reserve_noOut$market_segment_type[hotel_reserve_noOut$market_segment_ty
pe == "Online"] <- 5

hotel_reserve_noOut$booking_status[hotel_reserve_noOut$booking_status == "Can
celed"] <- 1
hotel_reserve_noOut$booking_status[hotel_reserve_noOut$booking_status == "Not
_Canceled"] <- 2
```

Visualize the encoded data

```r
summary(hotel_reserve_noOut) #summary of our cleaned data
```

```
##   no_of_adults   no_of_children    no_of_weekend_nights no_of_week_nights
##  Min.   :0.000   Min.   : 0.0000   Min.   :0.0000       Min.   : 0.000
##  1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000       1st Qu.: 1.000
##  Median :2.000   Median : 0.0000   Median :1.0000       Median : 2.000
##  Mean   :1.845   Mean   : 0.1053   Mean   :0.8107       Mean   : 2.204
##  3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000       3rd Qu.: 3.000
##  Max.   :4.000   Max.   :10.0000   Max.   :7.0000       Max.   :17.000
##  type_of_meal_plan  required_car_parking_space room_type_reserved
##  Length:36274       Min.   :0.00000            Length:36274
##  Class :character   1st Qu.:0.00000            Class :character
##  Mode  :character   Median :0.00000            Mode  :character
##                     Mean   :0.03099
##                     3rd Qu.:0.00000
##                     Max.   :1.00000
##    lead_time       arrival_year  arrival_month     arrival_date
##  Min.   :  0.00   Min.   :2017   Min.   : 1.000   Min.   : 1.0
##  1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000   1st Qu.: 8.0
##  Median : 57.00   Median :2018   Median : 8.000   Median :16.0
##  Mean   : 85.23   Mean   :2018   Mean   : 7.424   Mean   :15.6
##  3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000   3rd Qu.:23.0
##  Max.   :443.00   Max.   :2018   Max.   :12.000   Max.   :31.0
##  market_segment_type repeated_guest    no_of_previous_cancellations
##  Length:36274        Min.   :0.00000   Min.   : 0.00000
##  Class :character    1st Qu.:0.00000   1st Qu.: 0.00000
##  Mode  :character    Median :0.00000   Median : 0.00000
##                      Mean   :0.02564   Mean   : 0.02335
##                      3rd Qu.:0.00000   3rd Qu.: 0.00000
##                      Max.   :1.00000   Max.   :13.00000
##  no_of_previous_bookings_not_canceled no_of_special_requests booking_statu
s
##  Min.   : 0.0000                      Min.   :0.0000         Length:36274
##  1st Qu.: 0.0000                      1st Qu.:0.0000         Class :charac
ter
##  Median : 0.0000                      Median :0.0000         Mode  :charac
ter
##  Mean   : 0.1534                      Mean   :0.6197
##  3rd Qu.: 0.0000                      3rd Qu.:1.0000
##  Max.   :58.0000                      Max.   :5.0000
##  avg_price_per_room
##  Min.   :  0.00
##  1st Qu.: 80.30
##  Median : 99.45
##  Mean   :103.41
##  3rd Qu.:120.00
##  Max.   :375.50
```

```
str(hotel_reserve_noOut)
```

```
## 'data.frame':    36274 obs. of  18 variables:
##  $ no_of_adults                        : int  2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights                : int  1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights                   : int  2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan                   : chr  "1" "0" "1" "1" ...
##  $ required_car_parking_space          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                  : chr  "1" "1" "1" "1" ...
##  $ lead_time                           : int  224 5 1 211 48 346 34 83 121
44 ...
##  $ arrival_year                        : int  2017 2018 2018 2018 2018 201
8 2017 2018 2018 2018 ...
##  $ arrival_month                       : int  10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                        : int  2 6 28 20 11 13 15 26 6 18 .
..
##  $ market_segment_type                 : chr  "4" "5" "5" "5" ...
##  $ repeated_guest                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_special_requests              : int  0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status                      : chr  "2" "2" "1" "1" ...
##  $ avg_price_per_room                  : num  65 106.7 60 100 94.5 ...
```

```
head(hotel_reserve_noOut)
```

```
##   no_of_adults no_of_children no_of_weekend_nights no_of_week_nights
## 1            2              0                    1                 2
## 2            2              0                    2                 3
## 3            1              0                    2                 1
## 4            2              0                    0                 2
## 5            2              0                    1                 1
## 6            2              0                    0                 2
##   type_of_meal_plan required_car_parking_space room_type_reserved lead_tim
e
## 1                 1                          0                  1       22
4
## 2                 0                          0                  1        
5
## 3                 1                          0                  1        
1
## 4                 1                          0                  1       21
1
## 5                 0                          0                  1        4
8
## 6                 2                          0                  1       34
6
##   arrival_year arrival_month arrival_date market_segment_type repeated_gue
st
## 1         2017            10            2                   4            
0
## 2         2018            11            6                   5            
0
## 3         2018             2           28                   5            
0
## 4         2018             5           20                   5            
0
## 5         2018             4           11                   5            
0
```

```
## 6         2018             9           13                   5
0
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
## 1                            0                                    0
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            0                                    0
## 6                            0                                    0
##   no_of_special_requests booking_status avg_price_per_room
## 1                      0              2              65.00
## 2                      1              2             106.68
## 3                      0              1              60.00
## 4                      0              1             100.00
## 5                      0              1              94.50
## 6                      1              1             115.00
```

The factors below have been successfully encoded but they are still being read as character so we will convert them to numerical: - type_of_meal_plan - room_type_reserved - market_segment_type - booking_status

```
hotel_reserve_noOut$type_of_meal_plan <- as.numeric(hotel_reserve_noOut$type_
of_meal_plan)
hotel_reserve_noOut$room_type_reserved <- as.numeric(hotel_reserve_noOut$room
_type_reserved)
hotel_reserve_noOut$market_segment_type <- as.numeric(hotel_reserve_noOut$mar
ket_segment_type)
hotel_reserve_noOut$booking_status <- as.numeric(hotel_reserve_noOut$booking_
status)
str(hotel_reserve_noOut)
```

```
## 'data.frame':    36274 obs. of  18 variables:
##  $ no_of_adults                        : int  2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights                : int  1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights                   : int  2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan                   : num  1 0 1 1 0 2 1 1 1 1 ...
##  $ required_car_parking_space          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                  : num  1 1 1 1 1 1 1 4 1 4 ...
##  $ lead_time                           : int  224 5 1 211 48 346 34 83 121
44 ...
##  $ arrival_year                        : int  2017 2018 2018 2018 2018 201
8 2017 2018 2018 2018 ...
##  $ arrival_month                       : int  10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                        : int  2 6 28 20 11 13 15 26 6 18 .
..
##  $ market_segment_type                 : num  4 5 5 5 5 5 5 5 4 5 ...
##  $ repeated_guest                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_special_requests              : int  0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status                      : num  2 2 1 1 1 1 2 2 2 2 ...
##  $ avg_price_per_room                  : num  65 106.7 60 100 94.5 ...
```

Principal Component Analysis

```
# Performing PCA on all the variables except our target variable avg_price_pe
r_room
pc_hotel_reservation <- prcomp(hotel_reserve_noOut[,c(1,2,3,4,5,6,7,8,9,10,11
,12,13,14,15,16,17)], center = T, scale. = T)
attributes(pc_hotel_reservation)
```

```
## $names
## [1] "sdev"     "rotation" "center"   "scale"    "x"
##
## $class
## [1] "prcomp"
```

```
summary(pc_hotel_reservation)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6     PC
7
## Standard deviation     1.5419 1.3473 1.25973 1.18635 1.1132 1.02109 1.0037
1
## Proportion of Variance 0.1399 0.1068 0.09335 0.08279 0.0729 0.06133 0.0592
6
## Cumulative Proportion  0.1399 0.2466 0.33997 0.42276 0.4957 0.55699 0.6162
5
##                            PC8     PC9    PC10    PC11   PC12    PC13    P
C14
## Standard deviation     0.98560 0.96307 0.91029 0.89449 0.8429 0.73831 0.71
683
## Proportion of Variance 0.05714 0.05456 0.04874 0.04706 0.0418 0.03207 0.03
023
## Cumulative Proportion  0.67340 0.72795 0.77670 0.82376 0.8656 0.89762 0.92
785
##                           PC15    PC16    PC17
## Standard deviation     0.66718 0.63490 0.61507
## Proportion of Variance 0.02618 0.02371 0.02225
## Cumulative Proportion  0.95403 0.97775 1.00000
```

Visual Analysis of PCA results

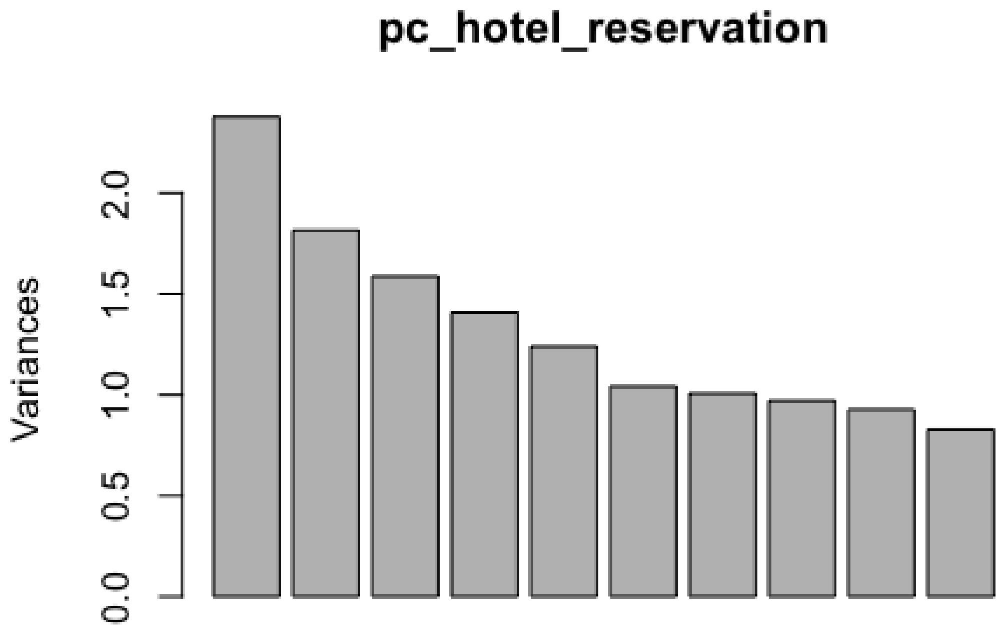
```
# calculate the proportion of explained variance (PEV) from the std values
pc_hotel_reservation_var <- pc_hotel_reservation$sdev^2
pc_hotel_reservation_var
```

```
##  [1] 2.3774818 1.8151355 1.5869288 1.4074283 1.2392624 1.0426320 1.0074375
##  [8] 0.9714132 0.9275011 0.8286243 0.8001041 0.7105618 0.5451076 0.5138486
## [15] 0.4451277 0.4030969 0.3783084
```

```
pc_hotel_reservation_PEV <- pc_hotel_reservation_var / sum(pc_hotel_reservati
on_var)
pc_hotel_reservation_PEV
```

```
##  [1] 0.13985187 0.10677268 0.09334875 0.08278990 0.07289779 0.06133129
##  [7] 0.05926103 0.05714195 0.05455889 0.04874261 0.04706495 0.04179776
## [13] 0.03206516 0.03022639 0.02618398 0.02371158 0.02225344
```

```
# plot of the variance per PC
#png(file = "hotelreserve_noOut PC PEV .png")
plot(pc_hotel_reservation)
```



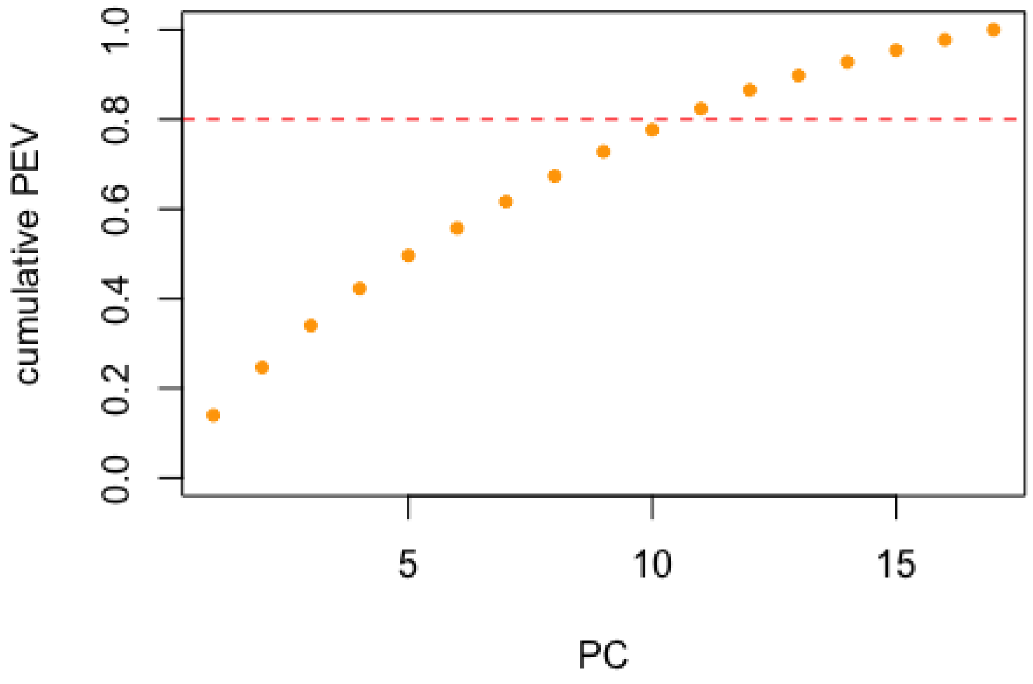
```
#dev.off()
```

Plot of the cumulative value of PEV for increasing number of additional PCs

We added an 80% threshold line to inform the feature extraction

according to the plot the first 10 PCs should be selected

```r
#Scree Plot
#png(file = "hotelreserve PC Scree Plot.png")
opar <- par(no.readonly = TRUE)
plot(
  cumsum(pc_hotel_reservation_PEV),
  ylim = c(0,1),
  xlab = 'PC',
  ylab = 'cumulative PEV',
  pch = 20,
  col = 'orange'
)
abline(h = 0.8, col = 'red', lty = 'dashed')
```



```
par(opar)
#dev.off()
```

From here we can see that 10 PC's contribute to 80% of the information in the dataset.

Getting and inspecting the loadings for each PC

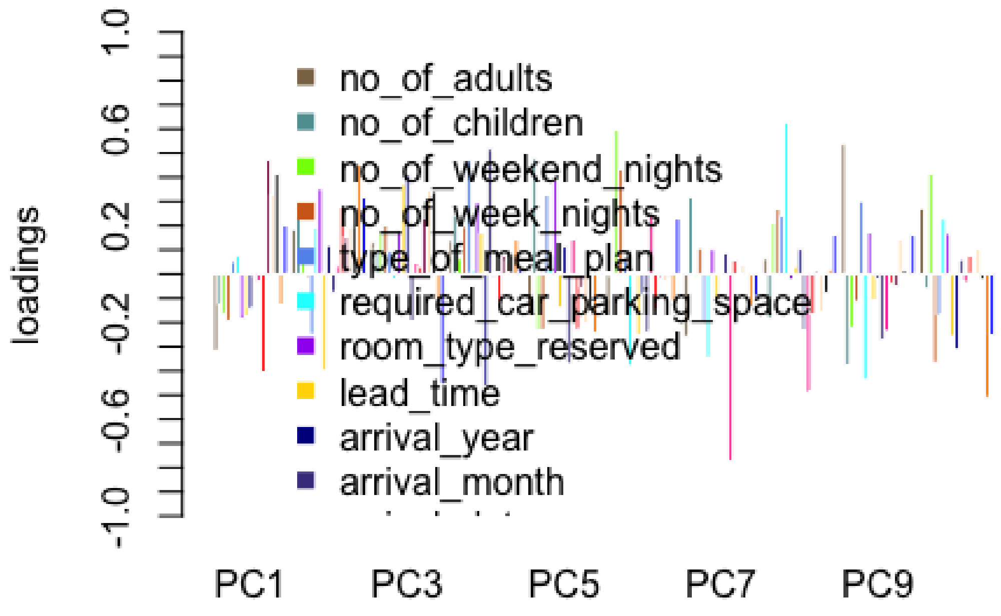
```
pc_hotel_reservation_loadings <- pc_hotel_reservation$rotation
pc_hotel_reservation_loadings

##                                               PC1         PC2         PC3
## no_of_adults                         -0.318138429  0.18543746  0.13078203
## no_of_children                       -0.124688084  0.26627099  0.10812053
## no_of_weekend_nights                 -0.162511112  0.06984672  0.16969632
## no_of_week_nights                    -0.191893804 -0.01470706  0.20288398
## type_of_meal_plan                     0.052635707 -0.24721632  0.09235529
## required_car_parking_space            0.074971865  0.19439823 -0.01984771
## room_type_reserved                   -0.183831167  0.35123646  0.17880244
## lead_time                            -0.174796728 -0.40044758  0.37141487
## arrival_year                         -0.142571687  0.11845760  0.40933380
## arrival_month                        -0.008699844 -0.07505234 -0.19664085
## arrival_date                         -0.032255212  0.03341048  0.05025817
## market_segment_type                  -0.406045402  0.27919308  0.03251631
## repeated_guest                        0.471261525  0.15816283  0.23215253
## no_of_previous_cancellations          0.329235034  0.18969988  0.34858479
## no_of_previous_bookings_not_canceled  0.415169135  0.19886326  0.34161367
## no_of_special_requests               -0.130620240  0.45498221 -0.10779374
## booking_status                        0.202590240  0.31733093 -0.45702734
##                                               PC4          PC5          PC
6
## no_of_adults                          0.139062372 -0.170244369 -0.16763746
1
## no_of_children                        0.244433135  0.477617571 -0.01262745
9
## no_of_weekend_nights                  0.063865881 -0.232273431  0.59843690
5
## no_of_week_nights                     0.197951229 -0.231956691  0.43390929
6
## type_of_meal_plan                     0.473615023  0.326948622  0.06884571
8
## required_car_parking_space            0.014775416  0.072199595 -0.38030673
6
## room_type_reserved                    0.301579553  0.389344816  0.00853787
3
## lead_time                             0.171411125 -0.138446609 -0.24997160
9
## arrival_year                         -0.464851942  0.113042989 -0.08350759
2
## arrival_month                         0.516795299 -0.370775684 -0.24037230
4
## arrival_date                         -0.008422843  0.147330830  0.24432280
```

```
7
## market_segment_type                  -0.115239470 -0.230370803 -0.15001775
9
## repeated_guest                        0.074343277 -0.056981065 -0.01372752
4
## no_of_previous_cancellations          0.057358836 -0.204740635 -0.02894783
2
## no_of_previous_bookings_not_canceled  0.081344366 -0.154559747 -0.00923741
2
## no_of_special_requests                0.142607203 -0.241174280 -0.11200190
4
## booking_status                        0.022418476 -0.005357345  0.22977178
4
##                                                PC7         PC8         PC9
## no_of_adults                         -2.576871e-01  0.06253300  0.53623875
## no_of_children                        3.188647e-01 -0.18603162 -0.37744305
## no_of_weekend_nights                 -1.234539e-02  0.21545706 -0.22071742
## no_of_week_nights                     1.091667e-01  0.26872327 -0.11319443
## type_of_meal_plan                    -2.062806e-01  0.24024691  0.30158056
## required_car_parking_space           -3.471048e-01  0.62814343 -0.43751839
## room_type_reserved                    1.010936e-01 -0.02329001  0.17375986
## lead_time                            -1.114517e-01  0.03180831 -0.10952407
## arrival_year                         -4.660257e-03  0.10342381 -0.02344023
## arrival_month                         8.737237e-02 -0.23166391 -0.27359370
## arrival_date                         -7.717505e-01 -0.48852713 -0.23673740
## market_segment_type                   5.873144e-02 -0.16884536 -0.03790930
## repeated_guest                       -3.083058e-05  0.01321782 -0.04999741
## no_of_previous_cancellations          3.840226e-02 -0.15622512  0.14739955
## no_of_previous_bookings_not_canceled  1.735078e-03 -0.07647624  0.01845996
## no_of_special_requests               -1.318109e-01  0.02173276 -0.01086007
## booking_status                       -9.285647e-02  0.16302419  0.16224873
##                                              PC10          PC11        PC1
2
## no_of_adults                          0.271553866 -0.0992698287  0.1630078
1
## no_of_children                       -0.060810205  0.0976260468 -0.1776378
3
## no_of_weekend_nights                  0.414246539  0.4983077744  0.0689334
0
## no_of_week_nights                    -0.364129199 -0.6308920564 -0.0444133
7
## type_of_meal_plan                    -0.170728050  0.2601480727 -0.2783815
9
## required_car_parking_space            0.227776323 -0.1605667740 -0.1222725
2
## room_type_reserved                    0.168413483 -0.1677102296  0.3232154
7
## lead_time                            -0.256917868  0.1978472300  0.0189685
9
## arrival_year                         -0.311269301  0.1475909995  0.3423062
```

```
8
## arrival_month                         0.061314986 -0.0085797573  0.3395286
5
## arrival_date                         -0.034225918 -0.1443026269  0.0133316
8
## market_segment_type                   0.078672594 -0.0093559845 -0.4169311
8
## repeated_guest                        0.001532874 -0.0001567824  0.2865185
9
## no_of_previous_cancellations          0.105324700 -0.0612670416 -0.4950697
8
## no_of_previous_bookings_not_canceled -0.023725267  0.0215423919  0.0440420
6
## no_of_special_requests               -0.509959304  0.3506110229 -0.0253020
9
## booking_status                       -0.251484429  0.0603070324  0.0405516
7
##                                             PC13        PC14         PC15
## no_of_adults                          0.10874509 -0.51014627 -0.018181672
## no_of_children                        0.15768988 -0.49451759 -0.043603285
## no_of_weekend_nights                  0.05280095  0.02048826  0.016667824
## no_of_week_nights                    -0.03818085 -0.05197669  0.012875564
## type_of_meal_plan                    -0.16015907  0.20296160 -0.100698360
## required_car_parking_space            0.04854765  0.04258691 -0.032653527
## room_type_reserved                    0.01520774  0.46383276  0.061498932
## lead_time                             0.24845721 -0.13301102  0.016883295
## arrival_year                          0.27022556  0.20764406 -0.119894668
## arrival_month                         0.21535256  0.22227342 -0.119327025
## arrival_date                          0.02005782  0.01878586  0.004549866
## market_segment_type                  -0.26779222  0.21040643 -0.163163829
## repeated_guest                       -0.31273228 -0.17042197  0.469680208
## no_of_previous_cancellations          0.49549654  0.19735791  0.265954638
## no_of_previous_bookings_not_canceled -0.24634787 -0.08373985 -0.723458287
## no_of_special_requests               -0.23489974  0.01793716  0.258845041
## booking_status                        0.46732814 -0.06139036 -0.212297395
##                                             PC16          PC17
## no_of_adults                         -0.10430110  1.398448e-01
## no_of_children                       -0.10899316  8.433810e-02
## no_of_weekend_nights                  0.01715400 -5.044112e-03
## no_of_week_nights                    -0.04840367  6.583829e-02
## type_of_meal_plan                    -0.35330841  1.558329e-01
## required_car_parking_space            0.03131787  2.881246e-02
## room_type_reserved                    0.28016064 -2.621052e-01
## lead_time                             0.21568156 -5.503984e-01
## arrival_year                         -0.34647040  2.541261e-01
## arrival_month                        -0.28136573  2.209904e-01
## arrival_date                         -0.02405811  4.145549e-03
## market_segment_type                  -0.41683112 -3.706844e-01
## repeated_guest                       -0.41490973 -2.927040e-01
## no_of_previous_cancellations          0.05864546  1.478964e-01
## no_of_previous_bookings_not_canceled  0.20136174  3.423441e-05
## no_of_special_requests                0.32334020  2.112424e-01
## booking_status                       -0.17195415 -4.113460e-01
```

Plotting first 10/17 PCs as barplots

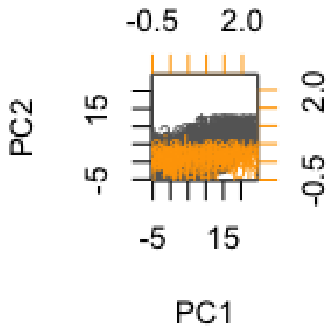
```
#png(file = "hotelreserve PC loadings.png")
opar <- par(no.readonly = TRUE)
colvector = c('burlywood4', 'cadetblue', 'chartreuse', 'chocolate', 'cornflow
erblue', 'cyan','purple','gold','darkblue','darkslateblue', 'deeppink', 'red'
, 'deeppink4', 'bisque','black','darkorange','blue')

labvector = c('PC1', 'PC2', 'PC3', 'PC4', 'PC5','PC6',"PC7","PC8","PC9","PC10
")
barplot(
  pc_hotel_reservation_loadings[,c(1:10)],
  beside = T,
  yaxt = 'n',
  names.arg = labvector,
  col = colvector,
  ylim = c(-1,1),
  border = 'white',
  ylab = 'loadings'
)
axis(2, seq(-1,1,0.1))
legend(
  'topright',
  bty = 'n',
  col = colvector,
  pch = 15,
  row.names(pc_hotel_reservation_loadings)
)
```

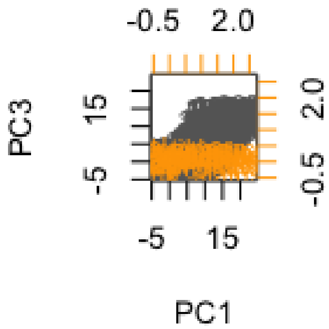


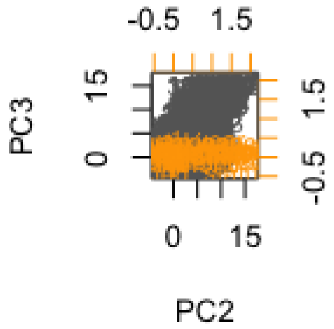
```
par(opar)
#dev.off()
```

Generating a biplot for each pair of important PCs (and show them on the same page)

```
# generate a biplot for each pair of important PCs (and show them on the same
page)
#   note: the option choices is used to select the PCs - default is 1:2
#png(file = "hotelreserve PC biplot.png")
opar <- par(no.readonly = TRUE)
par(mfrow = c(2,2))
biplot(
  pc_hotel_reservation,
  scale = 0,
  col = c('grey40','orange')
)
biplot(
  pc_hotel_reservation,
  choices = c(1,3),
  scale = 0,
  col = c('grey40','orange')
)
biplot(
  pc_hotel_reservation,
  choices = c(2,3),
  scale = 0,
  col = c('grey40','orange')
)
par(opar)
```







```
#dev.off()

#Hotel_reservation_cleaned <- write.csv(hotel_reserve_noOut, "HotelReservatio
nClean2.csv")
```

Creating a new data frame for the significant PC's and the average price per room

```
df2<- pc_hotel_reservation$x[,c(1,2,3,4,5,6,7,8,9,10)]
head(df2)
```

```
##             PC1         PC2        PC3        PC4        PC5        PC6
## 1  0.52037127 -1.66856375 -0.9274377  1.5478862 -0.9653901 -0.3405382
## 2 -0.61736771  1.22439969 -0.7916546 -0.6958705 -2.1282103  0.6014270
## 3  0.06452015 -0.55908129  0.6575124 -2.0316992  0.9921149  1.4371093
## 4 -0.71763592 -1.45500405  1.3331134 -0.8383802  0.1975531 -1.0226342
## 5 -0.51028576 -0.09989665  0.4996816 -2.3764327 -0.3540033 -0.4853107
## 6 -1.03426819 -2.14480542  1.6753671  1.2787078 -0.2469431 -1.9233033
##          PC7        PC8         PC9       PC10
## 1  0.9032258  0.6435880  0.41065273  0.6089206
## 2  1.2616942  0.2504536 -0.59996076  0.4971390
## 3 -0.6538572 -0.6776372 -0.96578912  1.0647277
## 4 -0.5255071 -0.5630776  0.17643233 -0.1580429
## 5  0.7901308 -0.4882149 -0.08424167  1.4339735
## 6 -0.5658988  0.1042361  0.45169554 -1.4585034
```

```
df3 <- cbind(df2,hotel_reserve_noOut$avg_price_per_room)
head(df3)
```

```
##           PC1         PC2        PC3        PC4        PC5        PC6
## 1  0.52037127 -1.66856375 -0.9274377  1.5478862 -0.9653901 -0.3405382
## 2 -0.61736771  1.22439969 -0.7916546 -0.6958705 -2.1282103  0.6014270
## 3  0.06452015 -0.55908129  0.6575124 -2.0316992  0.9921149  1.4371093
## 4 -0.71763592 -1.45500405  1.3331134 -0.8383802  0.1975531 -1.0226342
## 5 -0.51028576 -0.09989665  0.4996816 -2.3764327 -0.3540033 -0.4853107
## 6 -1.03426819 -2.14480542  1.6753671  1.2787078 -0.2469431 -1.9233033
##          PC7        PC8         PC9       PC10
## 1  0.9032258  0.6435880  0.41065273  0.6089206  65.00
## 2  1.2616942  0.2504536 -0.59996076  0.4971390 106.68
## 3 -0.6538572 -0.6776372 -0.96578912  1.0647277  60.00
## 4 -0.5255071 -0.5630776  0.17643233 -0.1580429 100.00
## 5  0.7901308 -0.4882149 -0.08424167  1.4339735  94.50
## 6 -0.5658988  0.1042361  0.45169554 -1.4585034 115.00
```

```
colnames(df3)
```

```
##  [1] "PC1"  "PC2"  "PC3"  "PC4"  "PC5"  "PC6"  "PC7"  "PC8"  "PC9"  "PC10"
## [11] ""
```

```
colnames(df3)[colnames(df3) == ""] <- "avg_price_per_room"
colnames(df3)
```

```
##  [1] "PC1"                "PC2"                "PC3"               
##  [4] "PC4"                "PC5"                "PC6"               
##  [7] "PC7"                "PC8"                "PC9"               
## [10] "PC10"               "avg_price_per_room"
```

```
head(df3)
```

```
##           PC1         PC2        PC3        PC4        PC5        PC6
## 1  0.52037127 -1.66856375 -0.9274377  1.5478862 -0.9653901 -0.3405382
## 2 -0.61736771  1.22439969 -0.7916546 -0.6958705 -2.1282103  0.6014270
## 3  0.06452015 -0.55908129  0.6575124 -2.0316992  0.9921149  1.4371093
## 4 -0.71763592 -1.45500405  1.3331134 -0.8383802  0.1975531 -1.0226342
## 5 -0.51028576 -0.09989665  0.4996816 -2.3764327 -0.3540033 -0.4853107
## 6 -1.03426819 -2.14480542  1.6753671  1.2787078 -0.2469431 -1.9233033
##          PC7        PC8         PC9       PC10 avg_price_per_room
## 1  0.9032258  0.6435880  0.41065273  0.6089206              65.00
## 2  1.2616942  0.2504536 -0.59996076  0.4971390             106.68
## 3 -0.6538572 -0.6776372 -0.96578912  1.0647277              60.00
## 4 -0.5255071 -0.5630776  0.17643233 -0.1580429             100.00
## 5  0.7901308 -0.4882149 -0.08424167  1.4339735              94.50
## 6 -0.5658988  0.1042361  0.45169554 -1.4585034             115.00

#This data set will be used for both machine learning and deep learning metho
ds in python
Hotel_reservation_PC <- write.csv(df3, "HotelReservationPC2.csv")
```