Assignment

Group 5

2023-04-10

Contents

1 Introduction

This data article describes two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted.

2 Dataset Description

The dataset used in this analysis is obtained from the website https://www.sciencedirect.com/science/article/pii/S2352340918315191 and is related to hotel bookings. The dataset contains information about various aspects of hotel bookings, including booking dates, customer demographics, booking channels, hotel features, and booking outcomes. The dataset is intended for analysis and research purposes in the field of hotel bookings.

After merging and cleaning the two datasets, we acquired a more diverse dataset for analysis. All columns retrieved from the dataset are relevant variables that can be used for our analysis. Thereafter, the dataset "hotel" contains 119386 observations with 32 variables is retained for analysis. The variables are:

1. "IsCanceled": Indicates if the booking was canceled (1) or not (0).
2. "LeadTime": Number of days between the booking date and the arrival date.
3. "ArrivalDateYear": Year of the arrival date.
4. "ArrivalDateMonth": Month of the arrival date.
5. "ArrivalDateWeekNumber": Week number of the arrival date.
6. "ArrivalDateDayOfMonth": Day of the month of the arrival date.
7. "StaysInWeekendNights": Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
8. "StaysInWeekNights": Number of weekday (Monday to Friday) nights the guest stayed or booked to stay at the hotel.

9. “Adults”: Number of adults in the booking.
10. “Children”: Number of children in the booking.
11. “Babies”: Number of babies in the booking.
12. “Meal”: Type of meal booked.
13. “Country”: Country of origin of the guest.
14. “MarketSegment”: Market segment designation for the booking.
15. “DistributionChannel”: Booking distribution channel.
16. “IsRepeatedGuest”: Indicates if the guest has stayed at the hotel before (1) or not (0).
17. “PreviousCancellations”: Number of previous booking cancellations.
18. “PreviousBookingsNotCanceled”: Number of previous bookings not canceled by the guest.
19. “ReservedRoomType”: Code of room type reserved by the guest.
20. “AssignedRoomType”: Code for the type of room assigned to the booking.
21. “BookingChanges”: Number of changes/amendments made to the booking from the initial reservation to the time of arrival.
22. “DepositType”: Type of deposit made by the guest.
23. “Agent”: ID of the travel agency that made the booking.
24. “Company”: ID of the company/entity that made the booking or responsible for paying the booking.
25. “DaysInWaitingList”: Number of days the booking was in the waiting list before it was confirmed to the guest.
26. “CustomerType”: Type of booking made by the customer.
27. “ADR”: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
28. “RequiredCarParkingSpaces”: Number of car parking spaces required by the guest.
29. “TotalOfSpecialRequests”: Number of special requests made by the guest (e.g. bed preferences, high floor, etc.).
30. “ReservationStatus”: Last status of the booking.
31. “ReservationStatusDate”: Date at which the last status was set.
32. “Hotel”: Type of hotel (City hotel or Resort hotel).

Given a presence of variables in the dataset, we have selected a subset of the most significant ones for analysis. Before proceeding with the data analysis, we first performed the following preliminary data modification and cleansing:

- Rows with NA are removed (removed 4 rows)
- Converted all the variables with data type of Character to Factor
- Combined the variables “Children” and “Babies” to one variable, “Children”
- Compared and combined the variables “ReservedRoomType” and “AssignedRoomType” to one variable, “RoomTypeMatch”, which indicates 1 if the room type matched, and 0 otherwise

After the preliminary selection and modification mentioned above, the dataset “hotel_data” contains 119,386 observations with 14 variables for analysis. In particular, we will use 13 variables to analyse our variable of interest, IsCanceled. Below is a list of variables contained in the dataset, and an overview of the dataset structure:

1. “IsCanceled”: Indicates if the booking was canceled (1) or not (0).
2. “LeadTime”: Number of days between the booking date and the arrival date.
3. “ArrivalDateMonth”: Month of the arrival date.
4. “StaysInWeekendNights”: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
5. “StaysInWeekNights”: Number of weekday nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
6. “Adults”: Number of adults in the booking. “7. Children”: Number of children in the booking.
7. “IsRepeatedGuest”: Indicates if the guest has stayed at the hotel before (1) or not (0).

8. "PreviousCancellations": Number of previous booking cancellations.
9. "PreviousBookingsNotCanceled": Number of previous bookings not canceled by the guest.
10. "BookingChanges": Number of changes/amendments made to the booking from the initial reservation to the time of arrival.
11. "ADR": Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights (we ignore the duplicated rows due to the nature of booking being possibly duplicated).
12. "Hotel": Type of hotel, either resort hotel (RH) or city hotel (CH)
13. "RoomTypeMatch": Indicates if reserved room type is the same as assigned room type (1) or not (0).

```
## 'data.frame':    119386 obs. of  14 variables:
##  $ IsCanceled                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
##  $ LeadTime                   : int  342 737 7 13 14 14 0 9 85 75 ...
##  $ ArrivalDateMonth           : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ StaysInWeekendNights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ StaysInWeekNights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ Adults                     : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ Children                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ IsRepeatedGuest            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PreviousCancellations      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousBookingsNotCanceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BookingChanges             : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ ADR                        : num  0 0 75 75 98 ...
##  $ Hotel                      : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2 2
##  $ RoomTypeMatch              : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...
```

3 Visualisation of Dataset

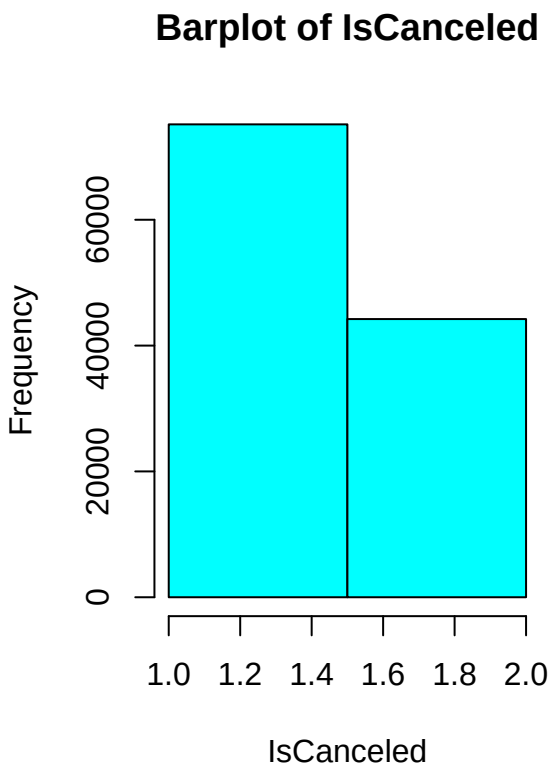
In this section, we will conduct a detailed analysis of the data and elaborate on our findings. We will examine and visualise each variable individually to detect any potential outliers. Furthermore, for highly skewed data, we will perform appropriate transformations to ensure a more symmetrical distribution, making it suitable for statistical tests that assume normality.

3.1 Summary Statistics for Main Variable of Interest (IsCanceled)

The bar plot below shows the overall distribution of our main variable is_canceled, together with a table that shows the whether a reservation is canceled.

```
##
##     0     1
## 75166 44220
```

```
## [1] 0.3703952
```



0 indicates a booking is not canceled, and 1 indicates a booking is canceled. It appears that the rate of cancellation is around 37%, which is relatively high.

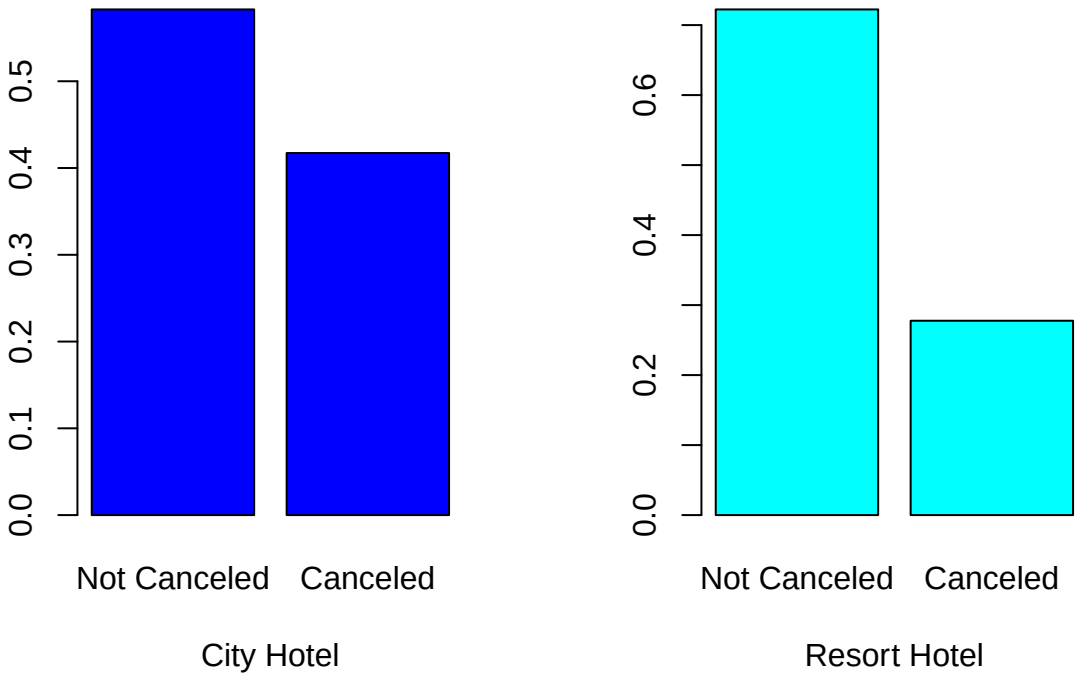
3.2 Summary Statistics for Other Variables

The remaining variables are investigated individually by first applying transformation when applicable, followed by identification and removal of possible outliers to avoid highly skewed data. The transformation (if applicable), histogram, boxplot, and the number of outliers removed from each variable are tabulated and illustrated in the following subsections below

3.2.1 Hotel Type

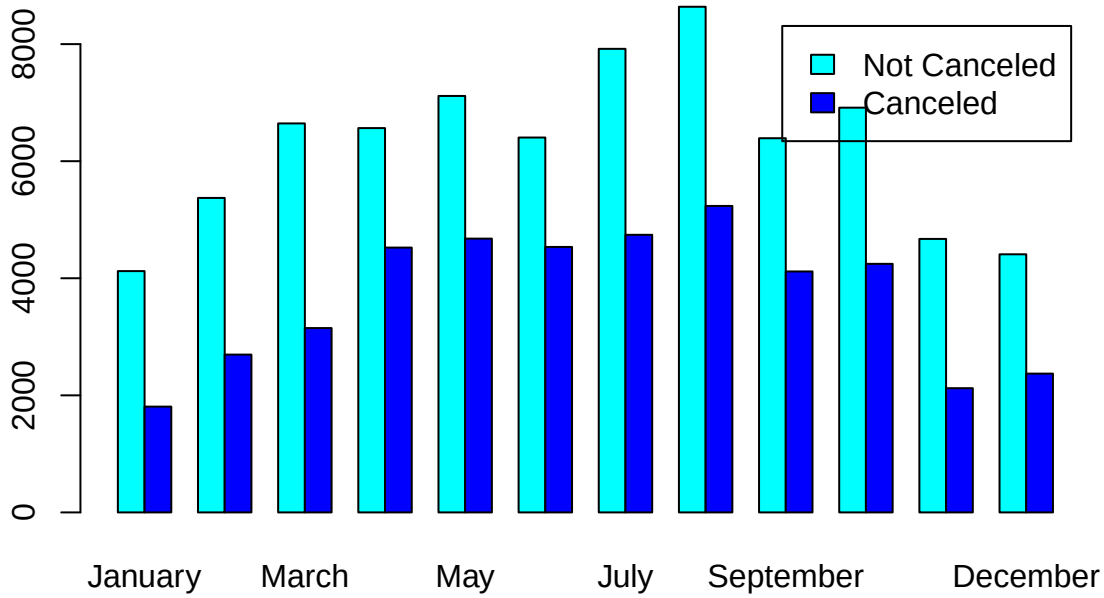
To better understand the hotel type and their rate of cancellation, a rate table is calculated, together with a bar chart to show the value for each hotel.

```
##
##                Not Canceled  Canceled
##   City Hotel      0.5827598 0.4172402
##   Resort Hotel    0.7223665 0.2776335
```



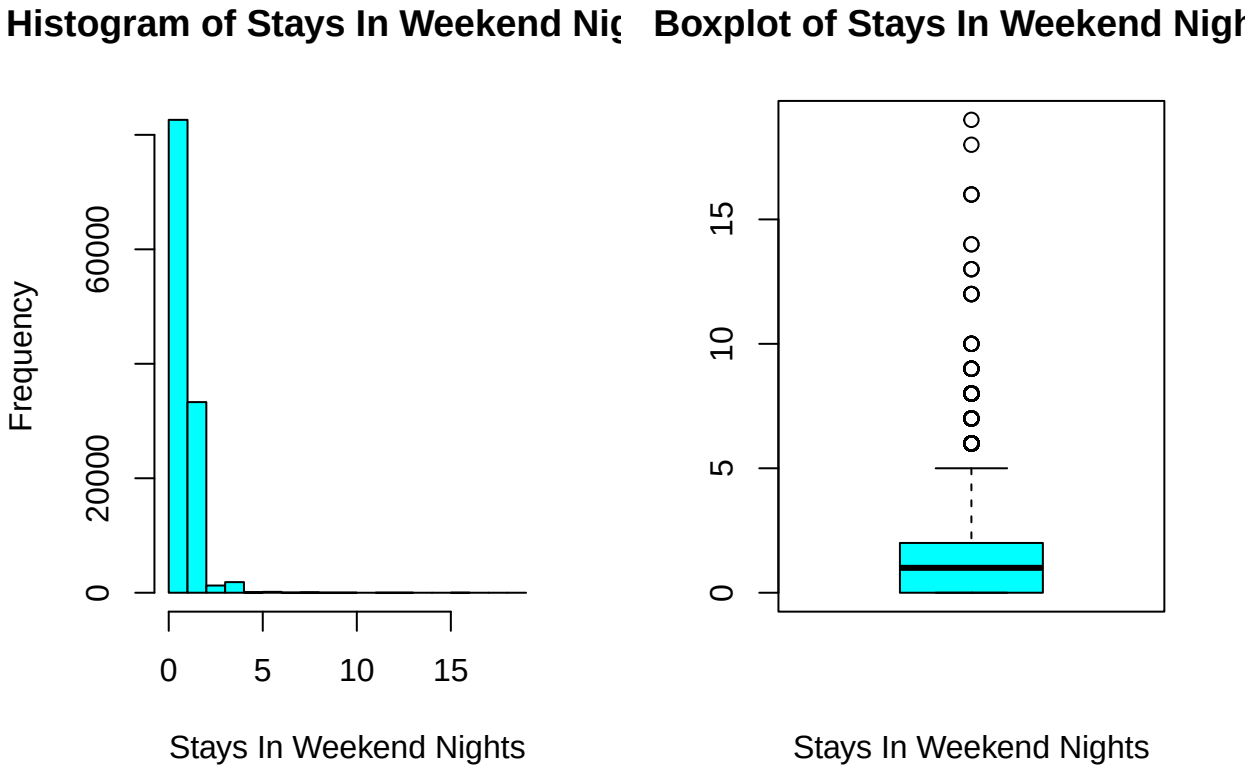
There seems to be a higher rate of cancellation for city hotels compared to resort hotel with about 41% cancellation from city hotel and 28% cancellation from resort hotel.

3.2.2 Arrival Date Month



- We observe that July and August are the peak sessions for hotel bookings, while December and January are the off-peak sessions.
- The months with the highest number of cancellations are May, July, and August, and the months with the lowest number of cancellations are November, December, and January.
- Additionally, we observe that in general, the number of cancellations appears to be much lesser than the total number of non-cancellations.

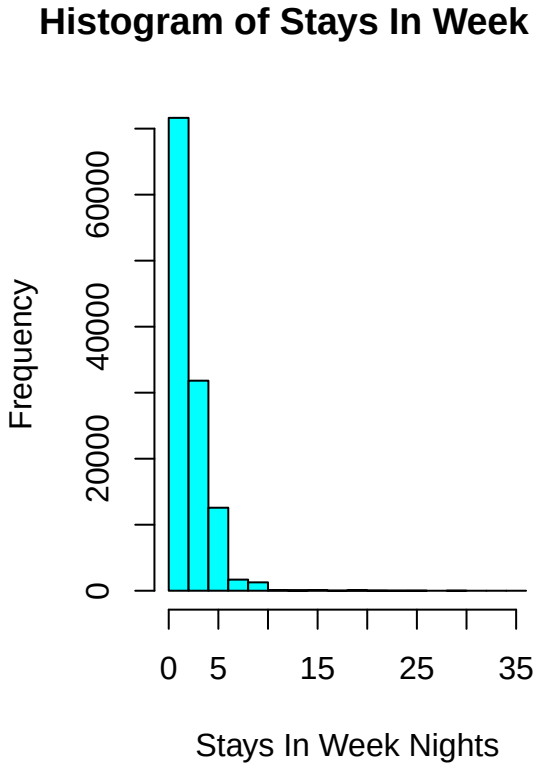
3.2.3 Stays In Weekend Nights

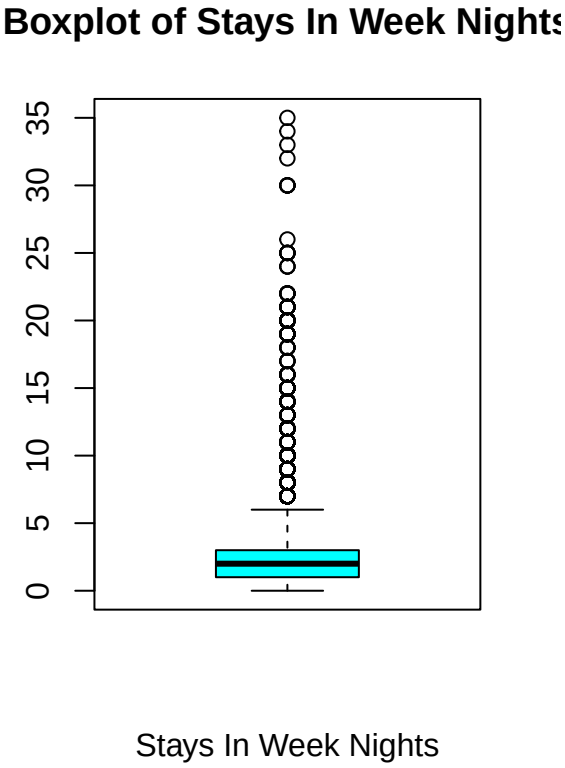



```
##
##     0     1     2     3     4     5     6     7     8     9    10    12    13
## 51996 30625 33307  1259  1855    79   153    19    60    11     7     5     3
##    14    16    18    19
##     2     3     1     1
```

- There appears to be right skewness in StaysInWeekendNights
- Most of the guests have 0-2 weekend night stays
- Due to the presence of 0 in the data, log transformation isn't applied
- Square root transformation was used
- 5 extreme outliers (>15) are removed

3.2.4 Stays In Week Nights

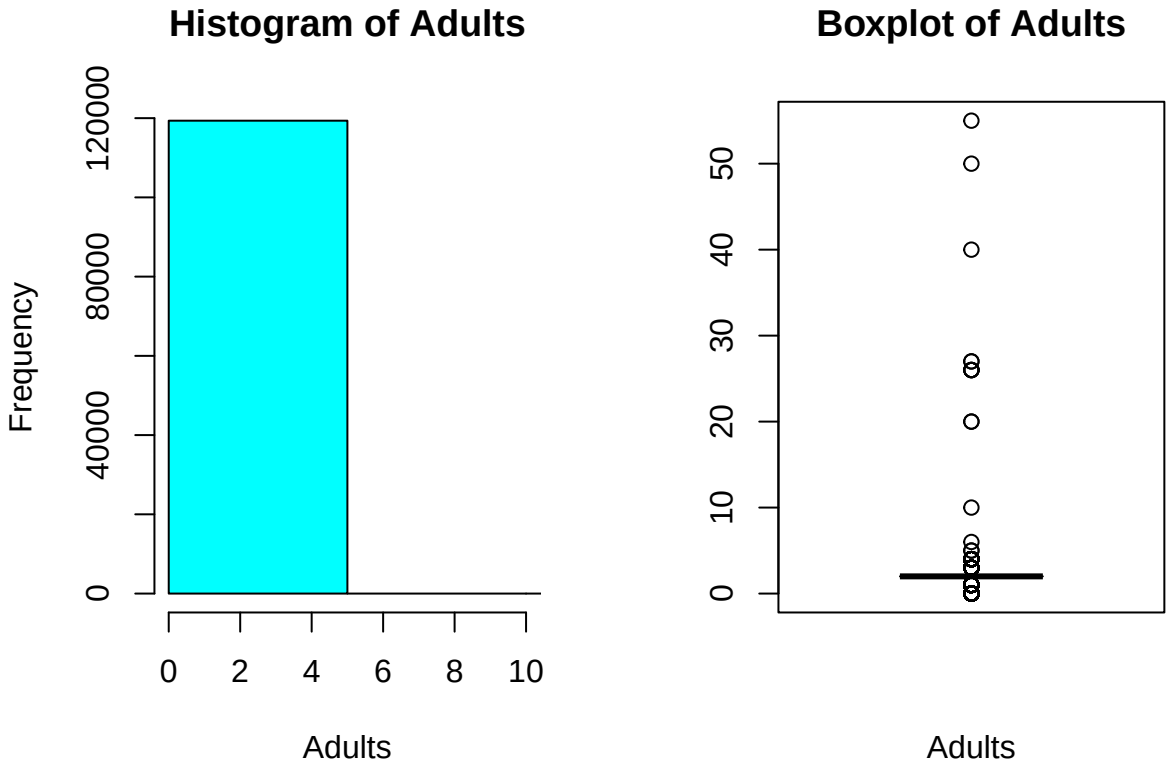





```
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
##  7644 30310 33682 22258  9563 11076  1499  1029   656   231  1036    56    42
##    13    14    15    16    17    18    19    20    21    22    24    25    26
##    27    35    85    16     4     6    44    41    15     7     3     6     1
##    30    32    33    34    35
##     5     1     1     1     1
```

- There appears to be right skewness in StaysInWeekNights
- Most of the guests have 1-3 weeknight stays
- Due to the presence of 0 in the data, log transformation isn't applied
- 4 extreme outliers (>=40) are removed

3.2.5 Adults

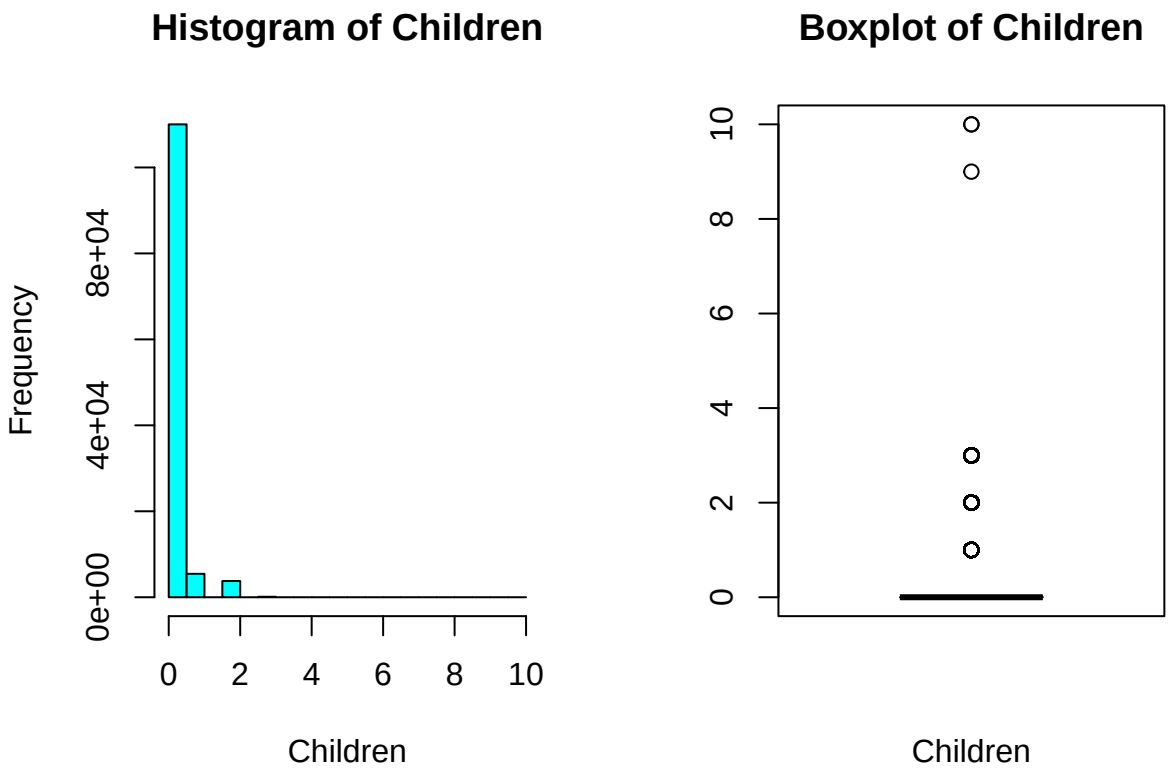



```
##
##     0     1     2     3     4     5     6    10    20    26    27    40    50
##   401 23024 89672  6201    62     2     1     1     2     5     2     1     1
##    55
##     1
```

- There appears to be right skewness in Adults
- Most of the bookings have 1-2 adults
- Due to the presence of 0 in the data, log transformation isn't applied
- 12 extreme outliers (>20) are removed

3.2.6 Children

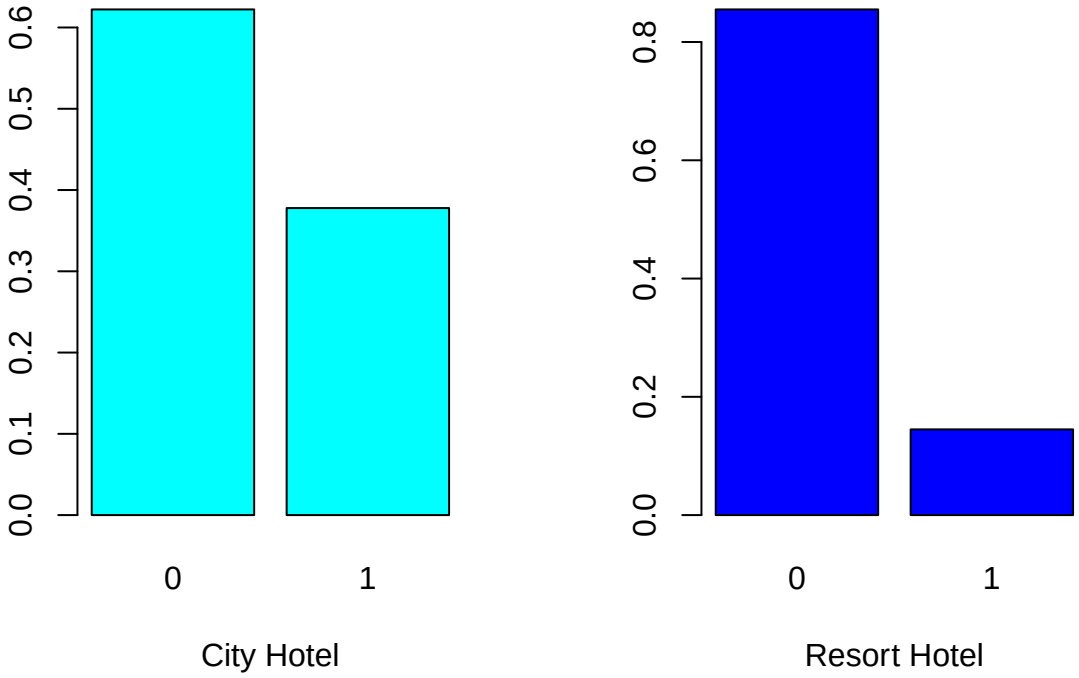
```
##
##      0      1      2      3      9     10
## 110032   5446   3772    111      1      2
```



- There appears to be right skewness in Children
- Most of the bookings have 0-2 Children
- Due to the presence of 0 in the data, log transformation isn't applied
- 3 extreme outliers (>5) are removed

3.2.7 If a guest has stayed at the hotel before (IsRepeatedGuest)

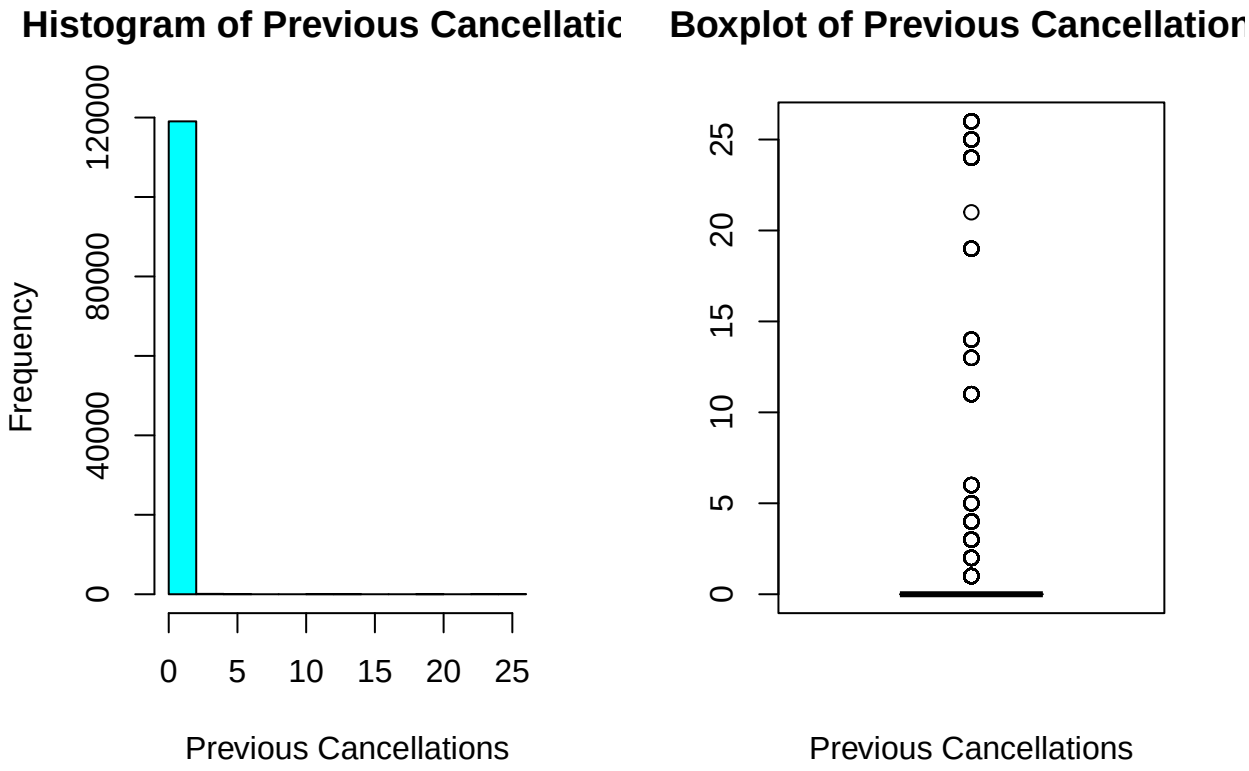
```
##
##             0         1
##   0 0.6222220 0.3777780
##   1 0.8550801 0.1449199
```



There seems to be a higher rate of repeated guest for city hotels compared to resort hotel with about 38% repeated guest from city hotel and 14% repeated guest from resort hotel.

3.2.8 Number of previous cancellation (Previous Cancellations)

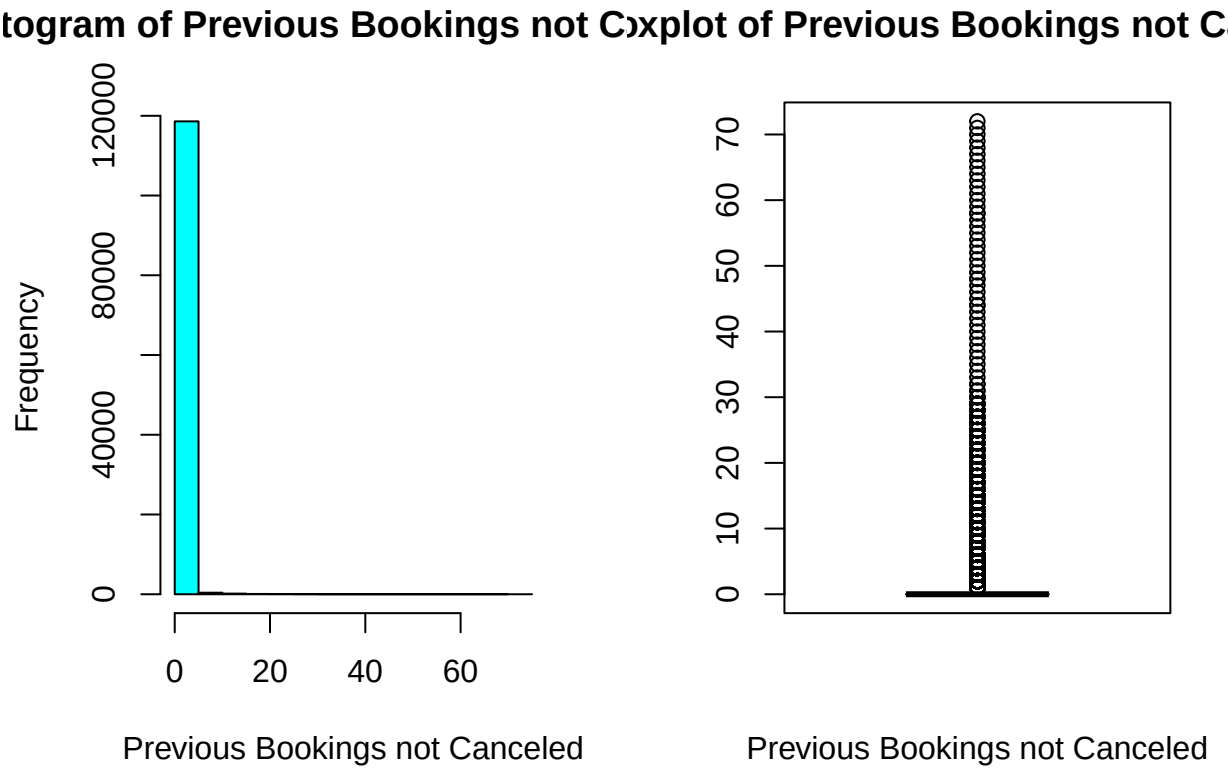
```
##
##      0      1      2      3      4      5      6     11     13     14     19
## 112877   6051    116     65     31     19     22     35     12     14     19
##     21     24     25     26
##      1     48     25     26
```



- Most of the bookings have 0-1 previous cancellation
- There appears to be right skewedness in Previous Cancellations
- Square root transform was used
- No extreme values were removed

3.2.9 Number of previous non-cancellation (Previous Bookings not Canceled)

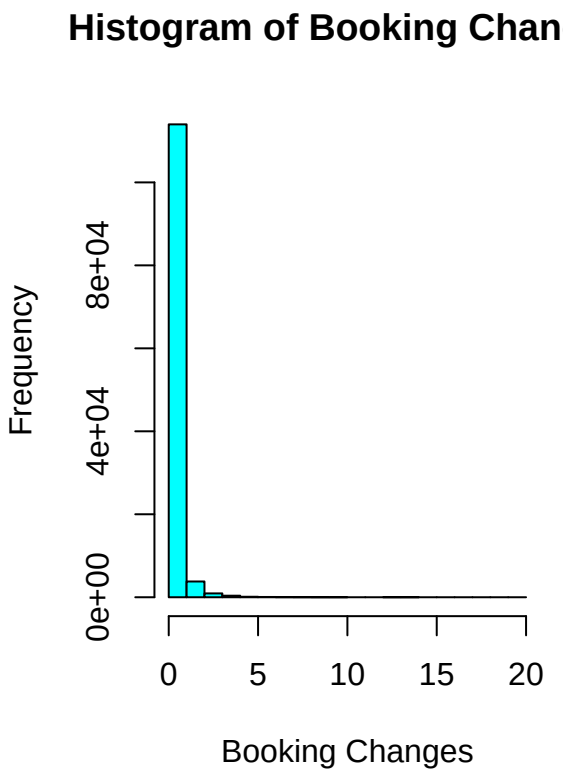
```
##
##      0      1      2      3      4      5      6      7      8      9     10
## 115743   1540    580    333    229    181    115     88     70     60     53
##     11     12     13     14     15     16     17     18     19     20     21
##     43     37     30     28     21     20     16     14     13     12     12
##     22     23     24     25     26     27     28     29     30     31     32
##     10      7      9     17      7      9      7      6      4      2      2
##     33     34     35     36     37     38     39     40     41     42     43
##      1      1      1      1      1      1      1      1      1      1      1
##     44     45     46     47     48     49     50     51     52     53     54
##      2      1      1      1      2      1      1      1      1      1      1
##     55     56     57     58     59     60     61     62     63     64     65
##      1      1      1      2      1      1      1      1      1      1      1
##     66     67     68     69     70     71     72
##      1      1      1      1      1      1      1
```

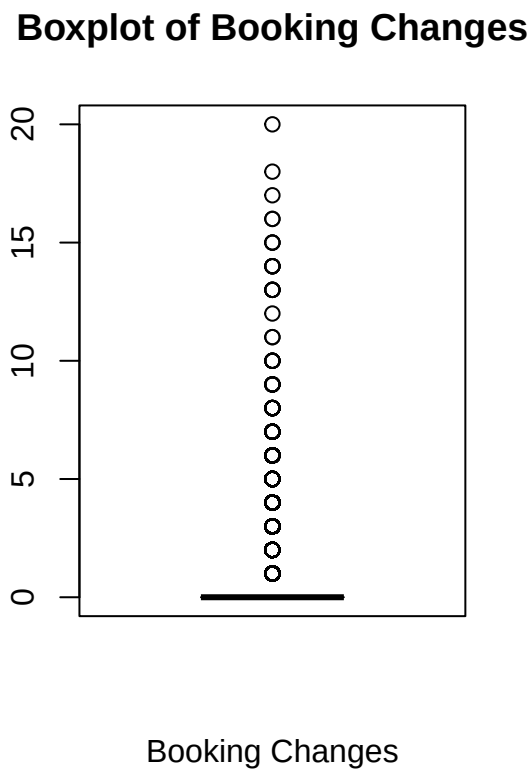


- There appears to be right skewness in PreviousBookingsNotCanceled
- Most of the bookings have 0-1 previous non-cancellation
- Due to the presence of 0 in the data, log transformation isn't applied
- No outlier is removed

3.2.10 Number of booking change (Booking Changes)

```
##
##      0      1      2      3      4      5      6      7      8      9     10
## 101295  12698   3802    927    376    118     62     31     17      8      6
##     11     12     13     14     15     16     17     18     20
##      2      1      5      5      3      2      1      1      1
```



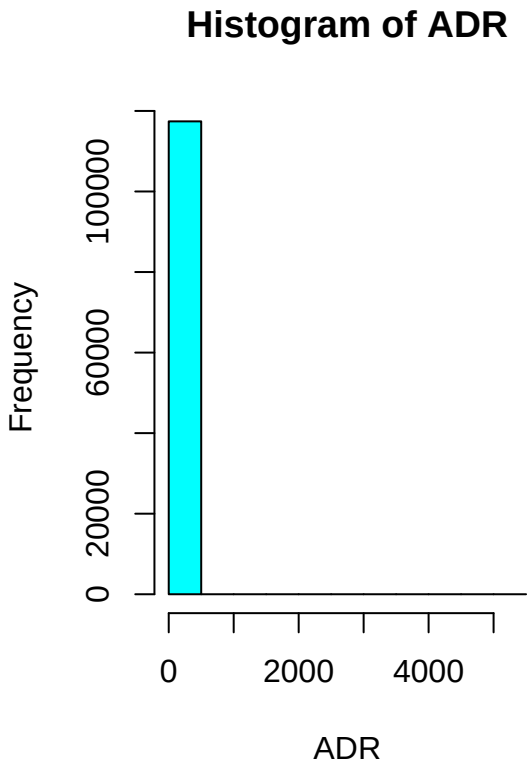


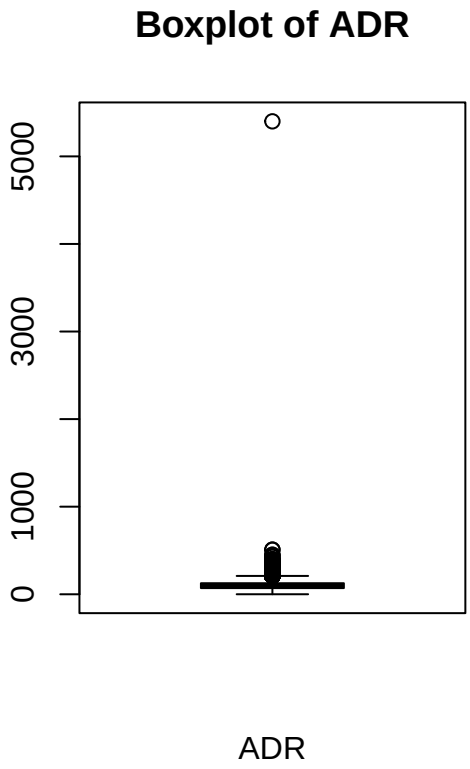
- Most of the bookings have 0-2 booking changes
- There appears to be right skewedness in Booking Changes
- Square root transformation is applied
- 1 extreme values (>=20) was removed

3.2.11 Average Daily Rate (ADR)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -6.38   69.29   94.63  101.85  126.00 5400.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.26   70.53   95.00  103.54  126.00 5400.00
```



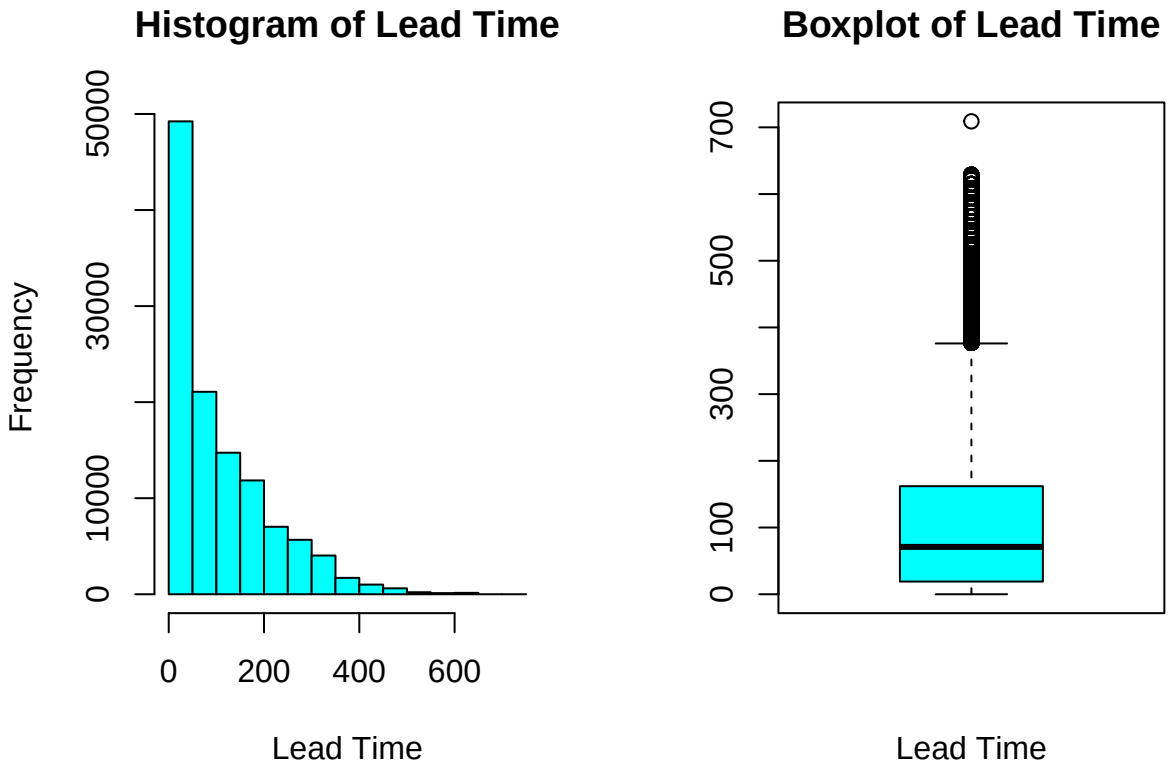


We removed the negative values of ADR as we are not interested in the special cases of getting paid to stay at hotels in cases like mystery shoppers or hotel reviewers and hotel brand ambassadors.

- There appears to be right skewedness in ADR
- Log transformation (base e) was applied
- 1 extreme values (>5000) was removed

3.2.12 Number of days between booking and arrival (Lead Time)

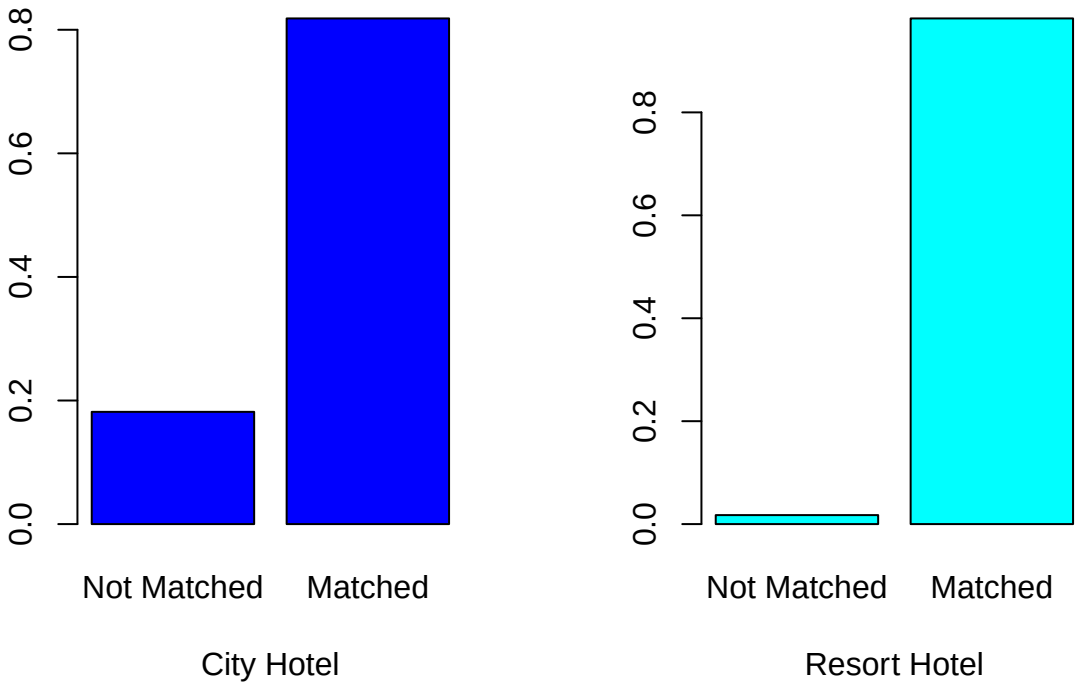
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    19.0    71.0   105.1   162.0   709.0
```



- There appears to be right skewedness in Lead Time
- Square root transformation was applied
- 2 extreme values were removed

3.2.13 If reserved and assigned room type is the same (RoomTypeMatch)

```
##
##              0          1
##   0 0.18166215 0.81833785
##   1 0.01745336 0.98254664
```



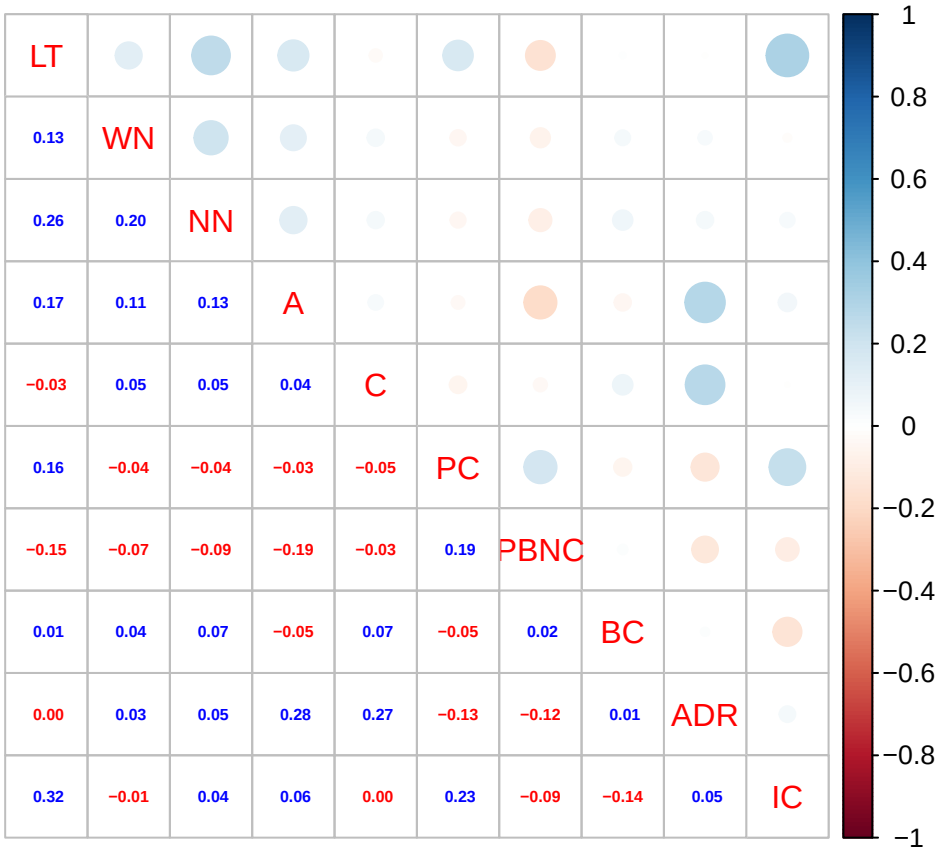
There seems to be a higher rate of room type match for resort hotels compared to city hotel with about 42% repeated guest from resort hotel and 5% repeated guest from city hotel.

##3.3 Final Dataset for Analysis

After the above analysis, the dataset is further reduced to 117,415 observations with the suggested log-transformation (base e) applied to Average Daily Rate (ADR) and square root transformation applied to StaysInWeekendNights, StaysInWeekNights, PreviousCancellations, PreviousBookingsNotCanceled and LeadTime.

4 Statistical Analysis

4.1 Correlation between *IsCanceled* against other Continous Variable



- There appears to be positive correlations between LeadTime and IsCanceled
- Stay In Weekend Nights is positively correlated to Stay in Weekday Nights

4.2 Statistical Test

4.2.1 Relation between *ADR* and *IsCanceled*

ANOVA (Analysis of Variance) is a statistical method used to test for differences in means between two or more groups. In this case, we can use ANOVA to investigate if there is a significant difference in the average daily rate (ADR), after log transformation was applied, of hotel bookings between canceled and non-canceled bookings. ANOVA is appropriate for this analysis because we have two groups (canceled and non-canceled bookings), and we want to test if there is a significant difference in the mean ADR between these groups.

ANOVA can help us to understand if the difference in the ADR between canceled and non-canceled bookings is due to random chance or if there is a real difference between the two groups. By using ANOVA, we can determine if the observed difference in ADR between canceled and non-canceled bookings is statistically significant or not, and this can help us to make more informed decisions regarding pricing and revenue management strategies for the hotel.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
```

```
## IsCanceled       1     53   53.37   246.6 <2e-16 ***
## Residuals   117412  25412    0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test output shows that there is a significant difference between the means of the groups, as indicated by the very small p-value (<2e-16) for the IsCanceled variable. This suggests that the IsCanceled variable has a significant impact on the average daily rate (ADR) in the hotel dataset. The F value of 271.4 also suggests a strong difference between the groups. The sum of squares for IsCanceled and the residuals are 53 and 25,412, respectively. This indicates that the variation within groups (residuals) is much higher than the variation between groups (IsCanceled). Overall, these results suggest that ADR is a significant predictor of IsCanceled in the datas

4.2.2 Relation between *Hotel* and *IsCanceled*

In this section we try to answer the question "Is there any association between type of hotel and IsCanceled". As both variables are categorical variables, we can use a 2-way (contingency) table, as shown below.

```
##
##                     0     1
##   City Hotel   45146 32968
##   Resort Hotel 28265 11035
```

To find any association between the type of wine and the quality of wine, we can perform the Pearson's Chi-squared test.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ctable
## X-squared = 2226, df = 1, p-value < 2.2e-16
```

In this test we obtain a very low p-value (<2.2e-16), which indicates strong evidence against the null hypothesis of independence. Therefore, we can conclude that there is a significant association between the Hotel Type and whether a booking will be canceled.

4.2.3 Relation between *PreviousCancellations* and *IsCanceled*

In this section we determine whether the booking cancelation depends on whether the guest made a previous cancellation. This variable represents the number of previous cancellations by the customer. It may have a significant impact on the cancellation rate, as customers who have a history of cancelling may be more likely to cancel their reservation in the future. A t-test and variance test can help determine if this variable is a significant predictor of cancellation.

```
##
##  Welch Two Sample t-test
##
## data:  PreviousCancellations by IsCanceled
## t = -66.028, df = 48039, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.1431027 -0.1348517
## sample estimates:
## mean in group 0 mean in group 1
##     0.008825835     0.147803034
```

```
##
##  F test to compare two variances
##
## data:  PreviousCancellations by IsCanceled
## F = 0.075925, num df = 73410, denom df = 44002, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07466572 0.07720341
## sample estimates:
## ratio of variances
##         0.07592531
```

Based on the results of the Welch two-sample t-test, we can infer that there is a significant difference in the mean values of PreviousCancellations between IsCanceled=0 and IsCanceled=1 groups. The mean value of PreviousCancellations for IsCanceled=1 is much higher than that for IsCanceled=0, with a difference of approximately 0.14.

Additionally, the F test to compare two variances indicates that the ratio of variances between the two groups is significantly different from 1, with a p-value $< 2.2e\text{-}16$. This suggests that the variance of PreviousCancellations for the IsCanceled=1 group is significantly different from that of the IsCanceled=0 group.

Overall, these results suggest that PreviousCancellations may be a significant predictor of IsCanceled.

4.2.4 Relation between *PreviousBookingsNotCanceled* and *IsCanceled*

In this section we determine whether the booking cancelation depends on whether the guest made a previous cancellation. This variable represents the number of previous bookings that were not cancelled by the customer. It may have a significant impact on the cancellation rate, as customers who have a history of not cancelling may be less likely to cancel their reservation in the future. A t-test and variance test can help determine if this variable is a significant predictor of cancellation.

```
##
##  Welch Two Sample t-test
##
## data:  PreviousBookingsNotCanceled by IsCanceled
## t = 39.288, df = 100128, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.06447779 0.07124883
## sample estimates:
## mean in group 0 mean in group 1
##     0.075911343     0.008048029
```

```
##
##  F test to compare two variances
##
## data:  PreviousBookingsNotCanceled by IsCanceled
## F = 7.9392, num df = 73410, denom df = 44002, p-value < 2.2e-16
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  7.807498 8.072855
## sample estimates:
## ratio of variances
##           7.939209
```

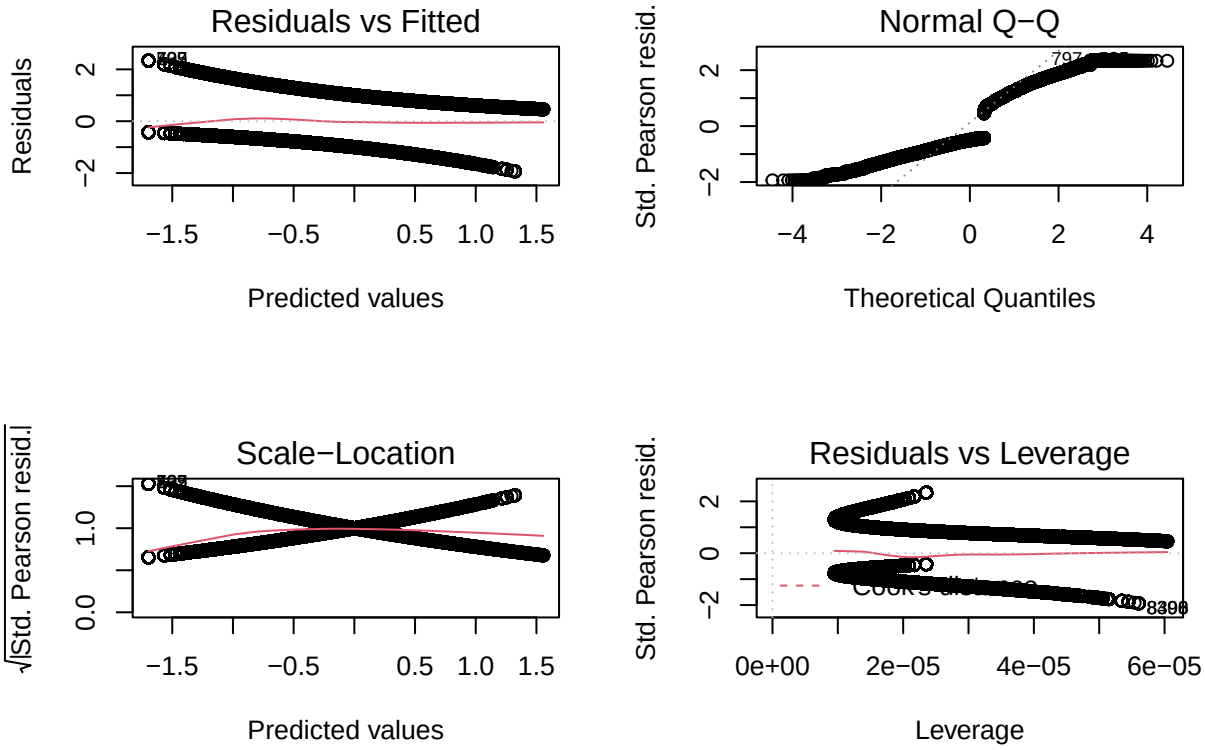
The Welch two-sample t-test result suggests that there is a significant difference in the means of "PreviousBookingsNotCanceled" variable between the "IsCanceled" groups. The p-value is less than 0.05, indicating strong evidence against the null hypothesis that there is no difference in means between the two groups.

Furthermore, the F-test result to compare two variances indicates that the variances of the two groups are significantly different. The p-value is less than 0.05, indicating strong evidence against the null hypothesis that the variances of the two groups are equal. This suggests that a Welch t-test, which does not assume equal variances, is an appropriate test to compare the means of the two groups.

4.2.5 Relation between *LeadTime* and *IsCanceled*

In this section we determine whether the booking cancelation depends on the leading time. We perform a simple linear regression between LeadTime and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ LeadTime, family = "binomial", data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7659  -0.9371  -0.6704   1.1671   1.9310
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.695977   0.013403  -126.5   <2e-16 ***
## LeadTime     0.129678   0.001247   104.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 143227  on 117412  degrees of freedom
## AIC: 143231
##
## Number of Fisher Scoring iterations: 4
```



From the coefficient estimates, we can infer that for every one unit increase in LeadTime, the log-odds of cancellation (IsCanceled being 1) decreases by 1.695977

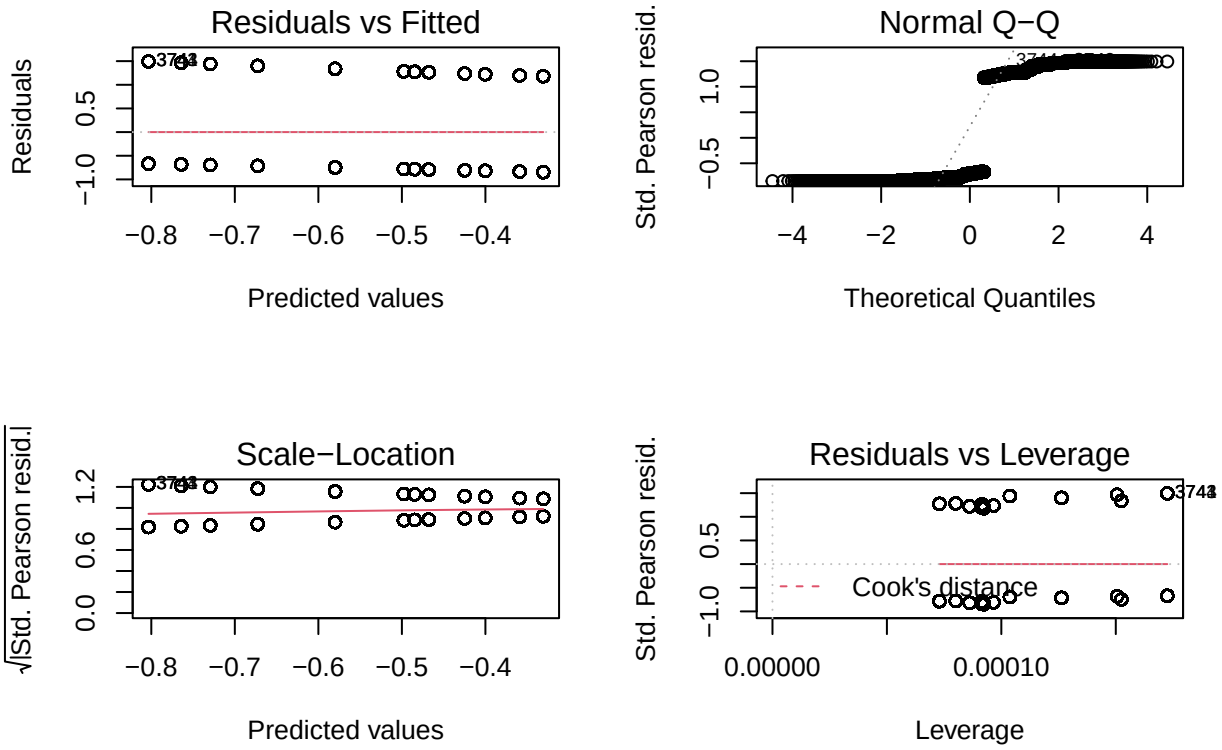
The p-value for LeadTime is 2e-16, indicating that it is a statistically significant predictor of IsCanceled.

4.2.6 Relation between *Arrival Date Month* and *IsCanceled*

In this section we determine whether the booking cancelation depends on the arrival month. We perform a simple linear regression between ArrivalDateMonth and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ ArrivalDateMonth, family = "binomial",
##     data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0406  -0.9863  -0.8871   1.3623   1.5320
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -0.80340    0.02842 -28.273  < 2e-16 ***
## ArrivalDateMonthFebruary   0.13083    0.03704   3.532 0.000412 ***
## ArrivalDateMonthMarch      0.07393    0.03578   2.066 0.038786 *
## ArrivalDateMonthApril      0.44444    0.03442  12.913  < 2e-16 ***
## ArrivalDateMonthMay        0.40318    0.03415  11.807  < 2e-16 ***
```

```
## ArrivalDateMonthJune       0.47266    0.03446  13.717  < 2e-16 ***
## ArrivalDateMonthJuly       0.30575    0.03388   9.024  < 2e-16 ***
## ArrivalDateMonthAugust     0.31872    0.03342   9.537  < 2e-16 ***
## ArrivalDateMonthSeptember  0.37858    0.03481  10.876  < 2e-16 ***
## ArrivalDateMonthOctober    0.33569    0.03455   9.715  < 2e-16 ***
## ArrivalDateMonthNovember   0.03915    0.03876   1.010 0.312442
## ArrivalDateMonthDecember   0.22343    0.03834   5.828 5.61e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 154758  on 117402  degrees of freedom
## AIC: 154782
##
## Number of Fisher Scoring iterations: 4
```



The logistic regression model shows that the month of arrival has a significant impact on the likelihood of cancellation. The intercept (-0.80340) represents the log-odds of cancellation for the reference month (January), and the estimated coefficients for the remaining months indicate how much the log-odds change relative to January.

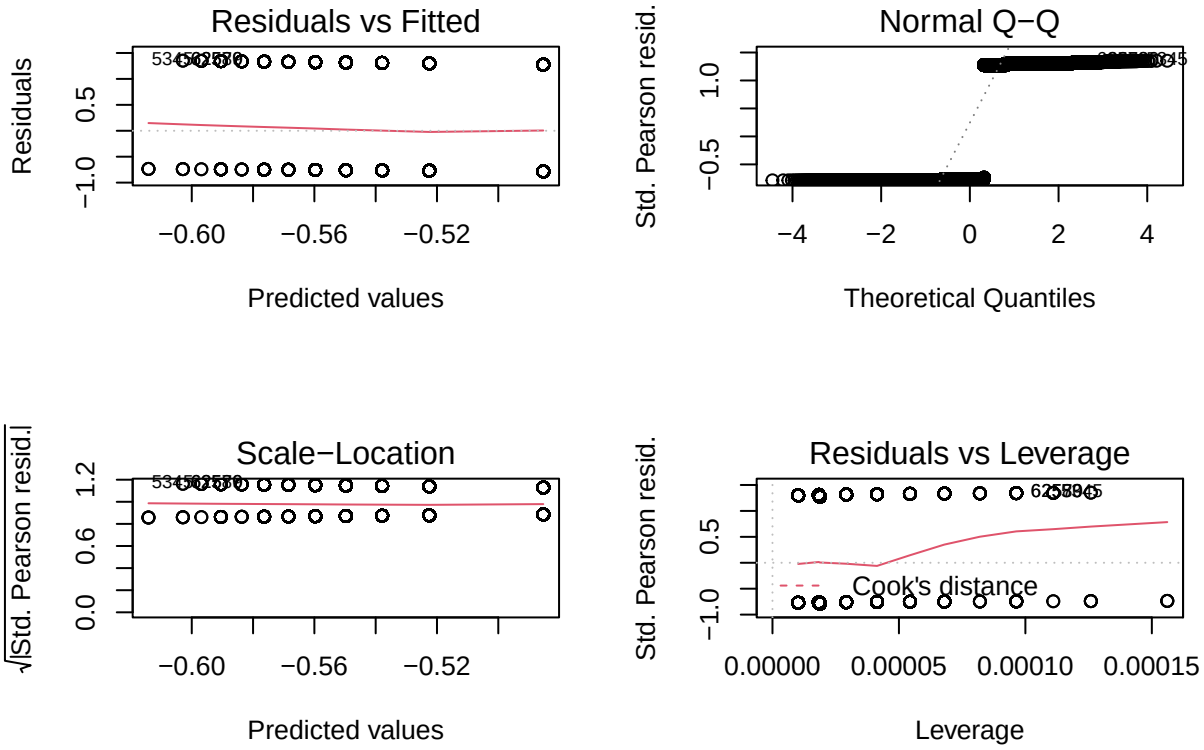
From the output, we can see that all months, except November, have a significant impact on the likelihood of cancellation. The positive coefficients indicate an increase in the log-odds of cancellation relative to January. For example, guests arriving in April have a 0.44444 increase in the log-odds of cancellation, holding all other variables constant.

It's worth noting that the coefficient for November is not statistically significant (p-value = 0.31), indicating that there is no significant difference in the log-odds of cancellation for guests arriving in November compared to January.

4.2.7 Relation between *Stays In Weekend Nights* and *IsCanceled*

In this section we determine whether the booking cancelation depends on how many nights they stay in the weekend. We perform a simple linear regression between StaysInWeekendNights and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ StaysInWeekendNights, family = "binomial",
##     data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9794  -0.9794  -0.9591   1.3892   1.4418
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.485322   0.008920 -54.411  < 2e-16 ***
## StaysInWeekendNights -0.037183   0.009251  -4.019 5.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 155310  on 117412  degrees of freedom
## AIC: 155314
##
## Number of Fisher Scoring iterations: 4
```



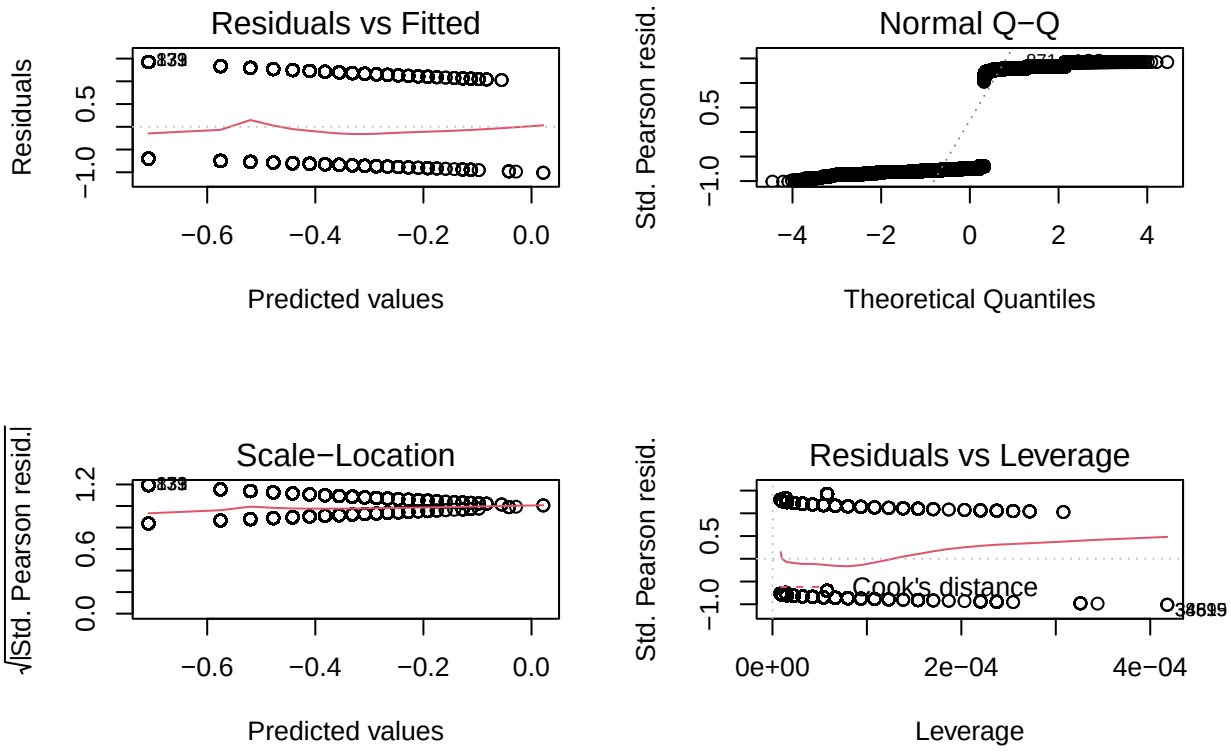
- The logistic regression model shows that the variable "StaysInWeekendNights" is statistically significant in predicting the probability of a booking being canceled, as the p-value 5.83e-05 is lower than the typical significance level of 0.05.
- The coefficient estimate (-0.037183) suggests that the odds of cancellation decrease slightly as the number of weekend nights increases.
- Overall, this variable may not be a strong predictor of cancellation behavior on its own, and it may need to be considered in combination with other variables to better understand its relationship with cancellations.

4.2.4 Relation between *Stays In Week Nights* and *IsCanceled*

In this section we determine whether the booking cancelation depends on how many nights they stay in the week. We perform a simple linear regression between StaysInWeekNights and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ StaysInWeekNights, family = "binomial",
##     data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1867  -0.9824  -0.9448   1.3858   1.4893
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)       -0.70867    0.01617  -43.82   <2e-16 ***
## StaysInWeekNights  0.13336    0.01012   13.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 155152  on 117412  degrees of freedom
## AIC: 155156
##
## Number of Fisher Scoring iterations: 4
```



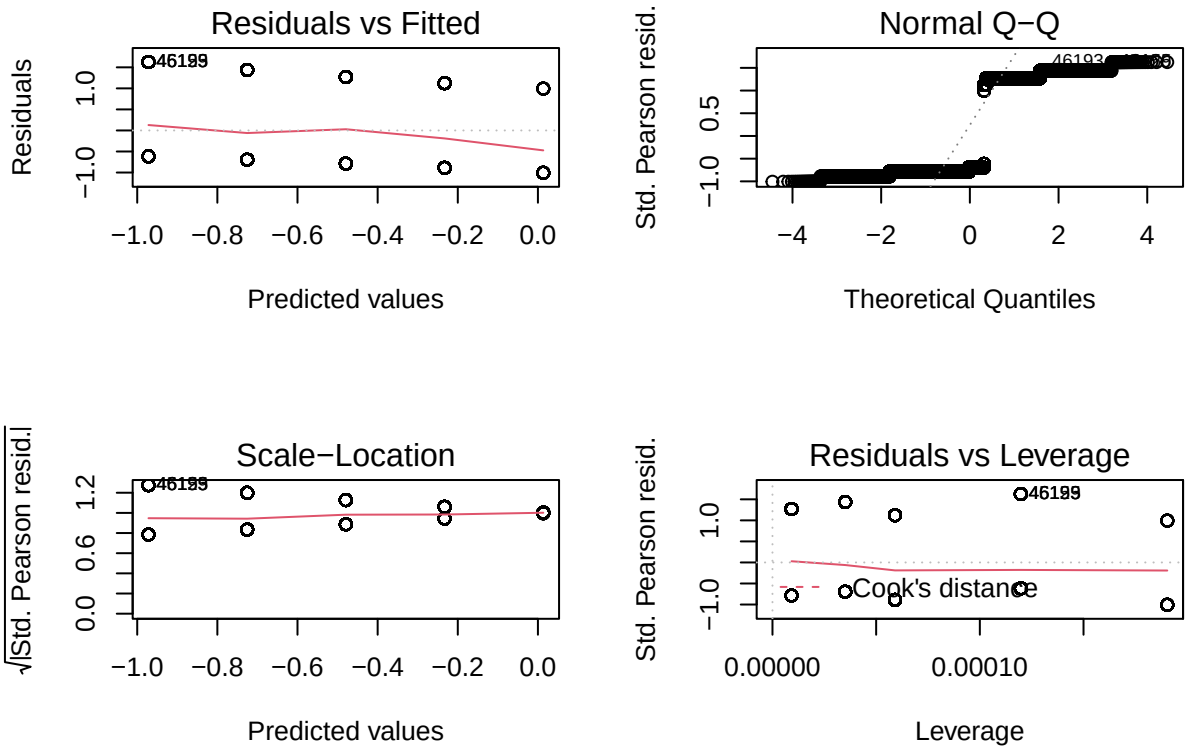
StaysInWeekNights has a coefficient of 0.13336, the p-value of <2e-16 is lesser than the typical significance level of 0.05 which shows that this variable is statistically significant. This suggests that the number of week nights stayed does have a significant effect on the likelihood of a booking being canceled, with an increase in the number of week nights leading to an increase in the likelihood of cancellation.

4.2.8 Relation between *Adults* and *IsCanceled*

In this section we determine whether the booking cancelation depends on how many adults are staying. We perform a simple linear regression between Adults and IsCanceled

```
##
## Call:
```

```
## glm(formula = IsCanceled ~ Adults, family = "binomial", data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1831  -0.9818  -0.8885   1.3865   1.6080
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.97195    0.02453  -39.63   <2e-16 ***
## Adults       0.24633    0.01267   19.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 154944  on 117412  degrees of freedom
## AIC: 154948
##
## Number of Fisher Scoring iterations: 4
```

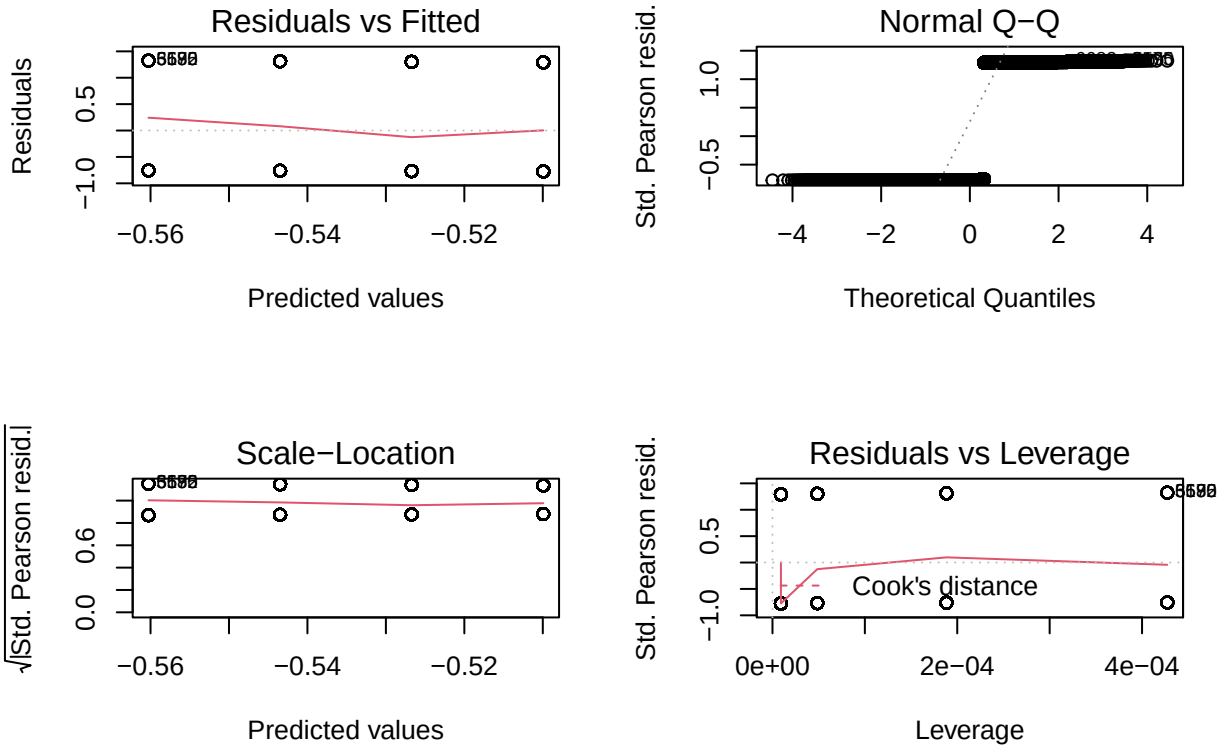


The output shows that the coefficient estimate for Adults is 0.24633, which means that for each additional adult in the booking, the log-odds of cancellation increase by 0.24633 The p-value of <2e-16 indicates that this effect is statistically significant.

4.2.9 Relation between *Children* and *IsCanceled*

In this section we determine whether the booking cancelation depends on how many childrens are staying. We perform a simple linear regression between Children and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ Children, family = "binomial", data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9699  -0.9699  -0.9699   1.4002   1.4227
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.509941   0.006248 -81.612   <2e-16 ***
## Children    -0.016774   0.014731  -1.139    0.255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 155324  on 117412  degrees of freedom
## AIC: 155328
##
## Number of Fisher Scoring iterations: 4
```



The p-value for the variable is 0.255, which indicates that the variable "Children" is not statistically significant in predicting the response variable "IsCanceled". Therefore, it is likely that the variable "Children" is not a strong predictor of cancellation behavior in this dataset.

4.2.10 Relation between *IsRepeatedGuest* and *IsCanceled*

In this section we determine whether the booking cancelation depends on whether it was a repeated guest. We perform a simple linear regression between IsRepeatedGuest and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ IsRepeatedGuest, family = "binomial",
##     data = hotel_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9792  -0.9792  -0.9792   1.3895   1.9090
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.485830   0.006095  -79.70   <2e-16 ***
## IsRepeatedGuest1 -1.159989   0.047876  -24.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 154588  on 117412  degrees of freedom
## AIC: 154592
##
## Number of Fisher Scoring iterations: 4
```

The p-value for the coefficient is <2e-16 which indicates that the variable "IsRepeatedGuest" is statistically significant in predicting the response variable "IsCanceled". Therefore, it is likely that the variable "IsRepeatedGuest" is a strong predictor of cancellation behavior in this dataset.

4.2.11 Relation between *BookingChanges* and *IsCanceled*

In this section we determine whether the booking cancelation depends on whether it was affected by the number of changes requested by the guest. We perform a simple linear regression between BookingChanges and IsCanceled

```
##
## Call:
## glm(formula = IsCanceled ~ BookingChanges, family = "binomial",
##     data = hotel_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.018  -1.018  -0.753   1.346   4.904
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.388081   0.006394  -60.70   <2e-16 ***
## BookingChanges -0.727233   0.015052  -48.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155326  on 117413  degrees of freedom
## Residual deviance: 152119  on 117412  degrees of freedom
## AIC: 152123
##
## Number of Fisher Scoring iterations: 4
```

The coefficient estimate for BookingChanges is -0.727233, which suggests that as the number of booking changes increases, the log-odds of cancellation decrease. The intercept coefficient (-0.388081) represents the estimated log-odds of cancellation when BookingChanges is zero.

The p-values associated with both coefficients are very small (<2e-16), indicating that both variables are statistically significant predictors of IsCanceled.

4.2.12 Relation between *RoomTypeMatch* and *IsCanceled*

In this section we try to answer the question "Is there any association between type of RoomTypeMatch and IsCanceled". As both variables are categorical variables, we can use a 2-way (contingency) table, as shown below.

```
##                   IsCanceled
## RoomTypeMatch        No   Yes
##   Different Types 13336   768
##   Same Types      60075 43235
```

To find any association between the type of wine and the quality of wine, we can perform the Pearson's Chi-squared test.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ctable
## X-squared = 7017.4, df = 1, p-value < 2.2e-16
```

The output indicates that a Pearson's Chi-squared test with Yates' continuity correction was performed on the ctable contingency table. The test statistic is 7017.4, with 1 degree of freedom, and the p-value is <2.2e-16, indicating a highly significant association between the variables RoomTypeMatch and IsCanceled.

4.3 Multiple Linear Regression

In this Section, we attempt to build a multiple linear regression model for IsCanceled based on all the variables

```
##
## Call:
## glm(formula = IsCanceled ~ ., family = binomial, data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1666  -0.9060  -0.4466   1.0529   5.0251
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.09381    0.11482 -44.365  < 2e-16 ***
## LeadTime                     0.12568    0.00175  71.825  < 2e-16 ***
## ArrivalDateMonthFebruary     0.01732    0.04736   0.366 0.714582
## ArrivalDateMonthMarch       -0.21156    0.04536  -4.664 3.10e-06 ***
## ArrivalDateMonthApril       -0.14830    0.04443  -3.338 0.000844 ***
## ArrivalDateMonthMay         -0.33798    0.04480  -7.544 4.55e-14 ***
## ArrivalDateMonthJune        -0.38758    0.04555  -8.509  < 2e-16 ***
## ArrivalDateMonthJuly        -0.75862    0.04613 -16.446  < 2e-16 ***
## ArrivalDateMonthAugust      -0.66570    0.04597 -14.482  < 2e-16 ***
## ArrivalDateMonthSeptember   -0.64256    0.04717 -13.622  < 2e-16 ***
## ArrivalDateMonthOctober     -0.43953    0.04560  -9.638  < 2e-16 ***
## ArrivalDateMonthNovember    -0.22703    0.04984  -4.555 5.24e-06 ***
## ArrivalDateMonthDecember    -0.19321    0.05031  -3.840 0.000123 ***
## StaysInWeekendNights        -0.08696    0.01230  -7.069 1.56e-12 ***
## StaysInWeekNights           -0.04765    0.01420  -3.355 0.000793 ***
## Adults                      -0.07782    0.01750  -4.446 8.75e-06 ***
## Children                     0.06341    0.01912   3.315 0.000915 ***
## IsRepeatedGuest1            -0.67620    0.13717  -4.930 8.24e-07 ***
## PreviousCancellations        4.12476    0.10129  40.723  < 2e-16 ***
```

```
## PreviousBookingsNotCanceled -1.71626    0.08547 -20.081  < 2e-16 ***
## BookingChanges              -0.62956    0.01743 -36.128  < 2e-16 ***
## ADR                          0.48904    0.02389  20.466  < 2e-16 ***
## HotelResort Hotel           -0.19338    0.01846 -10.474  < 2e-16 ***
## RoomTypeMatch1               2.08102    0.04353  47.801  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124244  on 93930  degrees of freedom
## Residual deviance:  99530  on 93907  degrees of freedom
## AIC: 99578
##
## Number of Fisher Scoring iterations: 7
```

```
## [1] 0.703445
```

We get a mean accuracy of 70.3%

4.3.1 Fine Tuning the Multiple Linear Regression Model

We will now try to fine tune the model by using the step function. We use a backward elimination method to select the most appropriate model.

```
## Start:  AIC=99578.11
## IsCanceled ~ LeadTime + ArrivalDateMonth + StaysInWeekendNights +
##     StaysInWeekNights + Adults + Children + IsRepeatedGuest +
##     PreviousCancellations + PreviousBookingsNotCanceled + BookingChanges +
##     ADR + Hotel + RoomTypeMatch
##
##                                Df Deviance    AIC
## <none>                              99530  99578
## - Children                      1    99541  99587
## - StaysInWeekNights             1    99541  99587
## - Adults                        1    99550  99596
## - IsRepeatedGuest               1    99556  99602
## - StaysInWeekendNights          1    99580  99626
## - Hotel                         1    99640  99686
## - ADR                           1    99963 100009
## - ArrivalDateMonth             11   100288 100314
## - PreviousBookingsNotCanceled   1   100442 100488
## - BookingChanges                1   101221 101267
## - RoomTypeMatch                 1   103254 103300
## - PreviousCancellations         1   105052 105098
## - LeadTime                      1   105203 105249
```

```
##
## Call:  glm(formula = IsCanceled ~ LeadTime + ArrivalDateMonth + StaysInWeekendNights +
##     StaysInWeekNights + Adults + Children + IsRepeatedGuest +
##     PreviousCancellations + PreviousBookingsNotCanceled + BookingChanges +
##     ADR + Hotel + RoomTypeMatch, family = binomial, data = train.data)
```

```
##
## Coefficients:
##                 (Intercept)                     LeadTime
##                    -5.09381                      0.12568
##    ArrivalDateMonthFebruary        ArrivalDateMonthMarch
##                     0.01732                     -0.21156
##       ArrivalDateMonthApril          ArrivalDateMonthMay
##                    -0.14830                     -0.33798
##        ArrivalDateMonthJune         ArrivalDateMonthJuly
##                    -0.38758                     -0.75862
##      ArrivalDateMonthAugust    ArrivalDateMonthSeptember
##                    -0.66570                     -0.64256
##     ArrivalDateMonthOctober     ArrivalDateMonthNovember
##                    -0.43953                     -0.22703
##    ArrivalDateMonthDecember         StaysInWeekendNights
##                    -0.19321                     -0.08696
##           StaysInWeekNights                       Adults
##                    -0.04765                     -0.07782
##                    Children             IsRepeatedGuest1
##                     0.06341                     -0.67620
##       PreviousCancellations  PreviousBookingsNotCanceled
##                     4.12476                     -1.71626
##              BookingChanges                          ADR
##                    -0.62956                      0.48904
##           HotelResort Hotel                RoomTypeMatch1
##                    -0.19338                      2.08102
##
## Degrees of Freedom: 93930 Total (i.e. Null);  93907 Residual
## Null Deviance:       124200
## Residual Deviance: 99530     AIC: 99580


##
## Call:
## glm(formula = IsCanceled ~ LeadTime + ArrivalDateMonth + StaysInWeekendNights +
##     StaysInWeekNights + Adults + Children + IsRepeatedGuest +
##     PreviousCancellations + PreviousBookingsNotCanceled + BookingChanges +
##     ADR + Hotel + RoomTypeMatch + I(LeadTime^2) + I(PreviousCancellations^2) +
##     I(PreviousBookingsNotCanceled^2) + I(ADR^2), family = binomial,
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4224  -0.9163  -0.4393   1.0438   4.6924
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.3913265  0.3533503  -1.107 0.268089
## LeadTime                        0.2412227  0.0057934  41.638  < 2e-16 ***
## ArrivalDateMonthFebruary        0.0023311  0.0479012   0.049 0.961186
## ArrivalDateMonthMarch          -0.2653539  0.0458251  -5.791 7.01e-09 ***
## ArrivalDateMonthApril          -0.2327712  0.0449662  -5.177 2.26e-07 ***
## ArrivalDateMonthMay            -0.4162291  0.0452394  -9.201  < 2e-16 ***
## ArrivalDateMonthJune           -0.4572274  0.0459308  -9.955  < 2e-16 ***
## ArrivalDateMonthJuly           -0.8306651  0.0464965 -17.865  < 2e-16 ***
```

```
## ArrivalDateMonthAugust           -0.7622574  0.0466400 -16.343  < 2e-16 ***
## ArrivalDateMonthSeptember        -0.6791078  0.0474498 -14.312  < 2e-16 ***
## ArrivalDateMonthOctober          -0.4821257  0.0459510 -10.492  < 2e-16 ***
## ArrivalDateMonthNovember         -0.2727087  0.0501903  -5.433 5.53e-08 ***
## ArrivalDateMonthDecember         -0.2375852  0.0508511  -4.672 2.98e-06 ***
## StaysInWeekendNights             -0.1210632  0.0124858  -9.696  < 2e-16 ***
## StaysInWeekNights                -0.0956553  0.0144844  -6.604 4.00e-11 ***
## Adults                           -0.1131356  0.0177012  -6.391 1.64e-10 ***
## Children                          0.0077087  0.0197140   0.391 0.695776
## IsRepeatedGuest1                 -0.5888830  0.1532805  -3.842 0.000122 ***
## PreviousCancellations             5.8486657  0.1659786  35.237  < 2e-16 ***
## PreviousBookingsNotCanceled      -2.3388703  0.1293756 -18.078  < 2e-16 ***
## BookingChanges                   -0.6391060  0.0174805 -36.561  < 2e-16 ***
## ADR                              -1.7555136  0.1564731 -11.219  < 2e-16 ***
## HotelResort Hotel                -0.2000950  0.0187182 -10.690  < 2e-16 ***
## RoomTypeMatch1                    2.0870050  0.0436921  47.766  < 2e-16 ***
## I(LeadTime^2)                    -0.0055404  0.0002678 -20.687  < 2e-16 ***
## I(PreviousCancellations^2)       -1.0654776  0.0394372 -27.017  < 2e-16 ***
## I(PreviousBookingsNotCanceled^2)  0.1825438  0.0197684   9.234  < 2e-16 ***
## I(ADR^2)                          0.2525419  0.0178025  14.186  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124244  on 93930  degrees of freedom
## Residual deviance:  98590  on 93903  degrees of freedom
## AIC: 98646
##
## Number of Fisher Scoring iterations: 7

## [1] 0.7004642
```

4.3.2 Comparison with another model

We have a slight improvement in accuracy after fine tuning but it is not very significant. In this section we will try to explore other prediction models instead. We will attempt to use Random Forest, a machine learning algorithm that works by constructing multiple decision trees during training time, and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree in the random forest is built using a random subset of the training data and a random subset of the features in the dataset. During prediction, each decision tree in the random forest independently predicts the class of the input data point and the class with the most votes (classification) or mean prediction (regression) is outputted as the final prediction of the random forest model.

```
## Random Forest accuracy: 0.7855896
```

We can see that the accuracy is now much higher at 78.6% as opposed to the multi linear regression model of around 70%. This is because Random Forests can perform better than a linear regression model in situations where the relationship between the predictors and the outcome variable is complex and nonlinear, or when there are interactions between the predictors. This is because Random Forests can capture complex relationships between variables through non-linear combinations of the predictors and can handle interactions between variables more effectively. In contrast, linear regression models assume a linear relationship between the predictors and the outcome variable and can struggle to capture complex and non-linear relationships.

5 Conclusion and Discussion

In conclusion, we attempted to build a multiple linear regression model to predict hotel cancellations based on various variables. In the multiple linear regression model for predicting hotel cancellations, after using the step function for fine-tuning, the significant predictors were identified based on their p-values in the model summary. The significant predictors were found to be:

- LeadTime: The number of days between the booking date and the arrival date. A longer lead time was found to be associated with a higher likelihood of cancellation, indicating that customers may be more likely to cancel bookings made further in advance.
- PreviousCancellations: The number of previous cancellations by the customer. A higher number of previous cancellations was found to be associated with a higher likelihood of cancellation, indicating that customers who have cancelled bookings before may be more likely to cancel future bookings.
- PreviousBookingsNotCanceled: The number of previous bookings that were not canceled by the customer. A lower number of previous bookings that were not canceled was found to be associated with a higher likelihood of cancellation, suggesting that customers who have a history of canceling bookings may be more likely to cancel future bookings.
- ADR: The average daily rate, which represents the average price per night paid by the customer. A higher ADR was found to be associated with a lower likelihood of cancellation, indicating that customers who are paying a higher price for their bookings may be less likely to cancel.

These predictors were found to be statistically significant in predicting hotel cancellations based on their p-values in the multiple linear regression model. Understanding these significant predictors can help hotel management to identify potential risk factors for cancellations and take appropriate measures to mitigate them, such as offering incentives for customers with longer lead times, addressing concerns of customers with previous cancellations, or providing value-added services to customers with higher ADRs to reduce the likelihood of cancellations.

After fine-tuning the model using backward elimination, we achieved a modest improvement in accuracy. However, we also explored an alternative approach using Random Forest, a machine learning algorithm, which resulted in a significantly higher accuracy of 78.6% compared to the linear regression model. This demonstrates that Random Forest is a more effective approach for predicting hotel cancellations in this case, as it can capture complex and non-linear relationships between variables and handle interactions more effectively. Therefore, considering the higher accuracy of the Random Forest model, it may be a better choice for predicting hotel cancellations in practice.