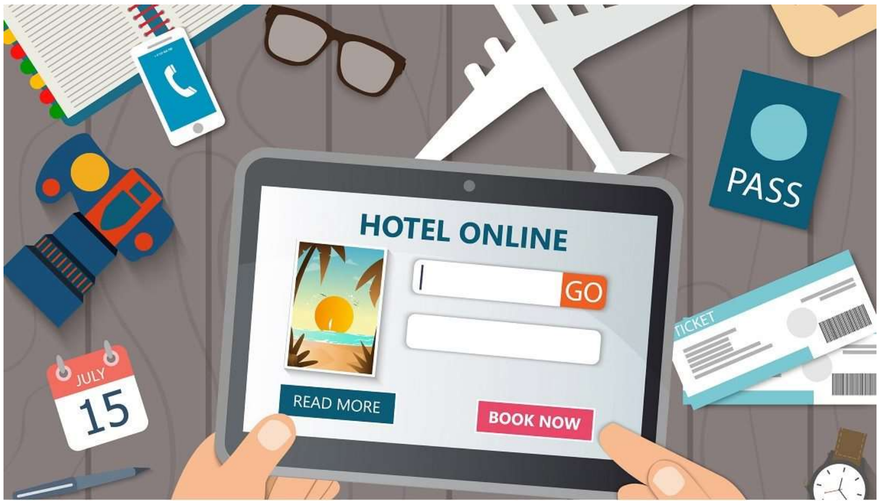


HOTEL BOOKING DEMAND

Booking Cancellation Prediction

UNIVERSITÀ degli STUDI di CATANIA
DIPARTIMENTO di ECONOMIA e IMPRESA

ABSTRACT

This data describes two datasets with hotel demand data from Portugal. One of the hotels is a resort hotel and the other is a city hotel. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July, 2015 and the 31st of August, 2017, including bookings that effectively arrived and bookings that were canceled

Presented to: Prof. Ingrassia

Data Analysis & Statistical Learning

Presented by:

Muhammad Zaffar, Syed
1000023608
Muhammad Hassan Ali
1000023619
Adeel 1000023622

Exploratory Data Analysis

The dataset for this report analysis is “**Hotel Booking Demand**,” which may be accessed in the “**Assignments for final reports**” section of MS-Teams as well as on the Kaggle platform: https://www.kaggle.com/jessemostipak/hotel-booking-demand.

This data collection comprises booking information for a city hotel and a resort hotel, including when the booking was made, duration of stay, number of adults, children, and/or babies, and number of available parking spaces, among other things.

Three datasets will be used in the analysis:

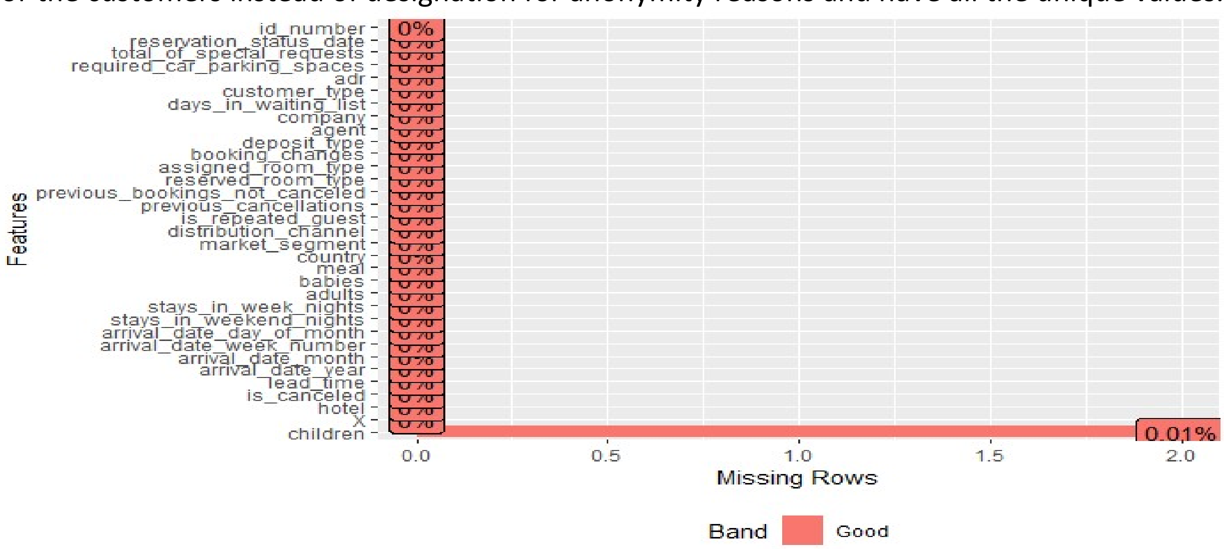
- Train Data: About 60% of the units of the original dataset.
- Validation Data: About 20% of the units of the original dataset.
- Test Data: About 20% of the units of the original dataset.

The Target Variable to be predicted: "is_canceled" in the training dataset, which consists of two classes “0” or “1”. 0 means “NO” and 1 means “Yes”.

This dataset on hotel bookings can assist us in answering those queries as below:
Have you ever thought about when the optimal time to book a hotel stay is? Or what is the best length of stay to get the greatest daily rate? What if you wanted to know whether a hotel was likely to receive an unusually large number of special requests?
Also we can utilize this dataset to conduct research on a variety of issues such as booking cancellation prediction, customer segmentation, customer satiation, seasonality, and so on.

Attribute Specification

The dataset contains initially 71633 observations of 32 variables. We will drop the first variable which is for the Id of the hotels which has all the unique values also the ID variable is presented for the customers instead of designation for anonymity reasons and have all the unique values.



Also we have 2 missing values in the children variable so we will omit these missing values from the dataset.
When the transformation has been made, now the dataset has 31 variables with 71631 observations. Following are the variable type and description to understand the nature of the variable as per the information was collected.

Attribute	Attribute Type	Description
hotel	Categorical	booking information for a city hotel or a resort hotel
is_canceled (Target Variable)	Categorical	Value indicating if the booking was canceled (1) or not (0)
lead_time	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Integer	Year of arrival date
arrival_date_month	Categorical	Month of arrival date with 12 categories: “January” to “December”
arrival_date_week_number	Integer	Week number of the arrival date
arrival_date_day_of_month	Integer	Day of the month of the arrival date
stays_in_weekend_nights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Integer	Number of adults
children	Integer	Number of childern
babies	Integer	Number of babies
meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board

		(breakfast and one other meal – usually dinner);
country	Categorical	Country of origin
market_segment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
distribution_channel	Categorical	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
is_repeated_guest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0
previous_cancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking
agent	Categorical	ID of the travel agency that made the booking
previous_bookings_not_canceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
assigned_room_type	Categorical	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
booking_changes	Integer	Number of changes or amendments made to the

		booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
days_in_waiting_list	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking

adr	Numeric	Average Daily Rate as defined
required_car_parking_spaces	Integer	Number of car parking spaces required by the customer
total_of_special_requests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status_date	Date	Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel

The Dataset consists of 14 variables that are categorical type, 16 variables that are of numeric type and 1 variable of Date type.
The main objective is to find out which criteria are the most successful in the forecast of cancelations of bookings. In this report, we will also classify our datasets by various techniques so that we can calculate the maximum precision for the main purpose. This helps us lower the booking cancelation after realizing which aspects have to be increased or decreased to mitigate the booking cancellation.

Data Exploration

Given Below are the descriptive statistics of the hotel booking dataset by which we can see that the mean, median, minimum, and maximum of the variables;

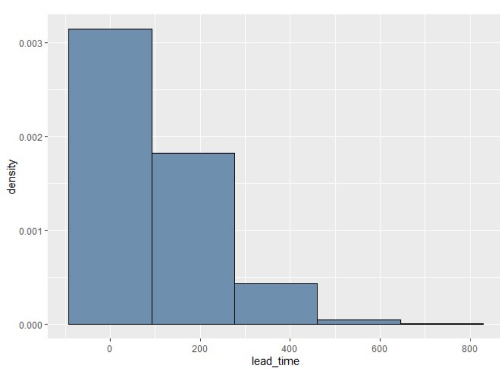
Continuous							
Variable	**Minimum**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**	**Others**
lead_time	0.0	18.0	69.0	103.8	160.0	737.0	-
arrival_date_year	2015	2016	2016	2016	2017	2017	-
arrival_date_week_number	1.0	16.0	28.0	27.19	38.0	53.0	-
arrival_date_day_of_month	1.0	8.0	16.0	15.77	23.0	31.0	-
stays_in_weekend_nights	0.0	0.0	1.0	0.924	2.0	18.0	-

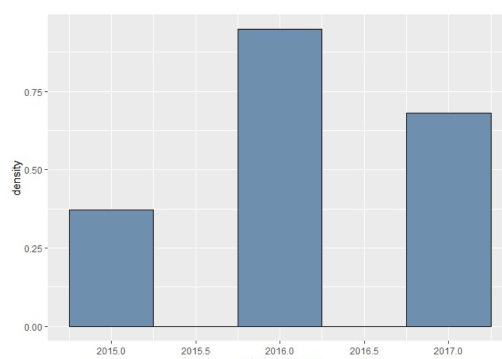
stays_in_week_nights	0.0	1.0	2.0	2.497	3.0	42.0	-
adults	0.0	2.0	2.0	1.857	2.0	55.0	-
children	0.0	0.0	0.0	0.104	0.0	10.0	-
babies	0.0	0.0	0.0	0.007	0.0	10.0	-
previous_cancellations	0.0	0.0	0.0	0.08	0.0	26.0	-
previous_bookings_not_canceled	0.0	0.0	0.0	0.136	0.0	70.0	-
booking_changes	0.0	0.0	0.0	0.223	0.0	21.0	-
days_in_waiting_list	0.0	0.0	0.0	2.35	0.0	391.0	-
adr	-6.38	69.0	94.67	101.9	126.0	5400	-
required_car_parking_spaces	0.0	0.0	0.0	0.062	0.0	8.0	-
total_of_special_requests	0.0	0.0	0.0	0.571	1.0	5.0	-
Date							
reservation_status_date	2014-10-17	2016-02-01	2016-08-06	2016-07-29	2017-02-08	2017-09-14	-

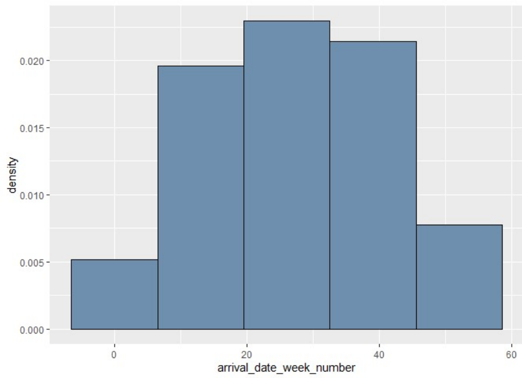
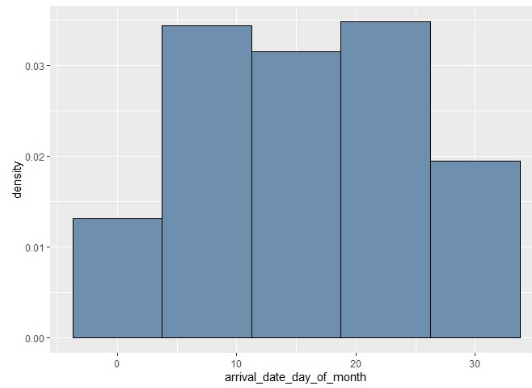
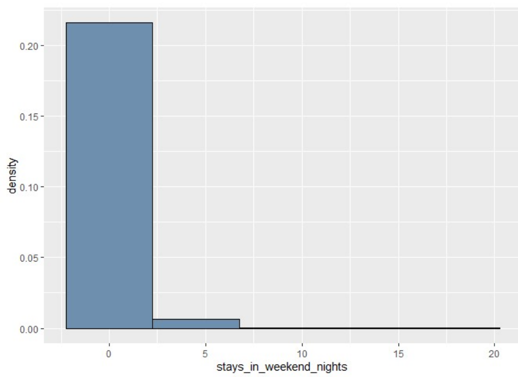
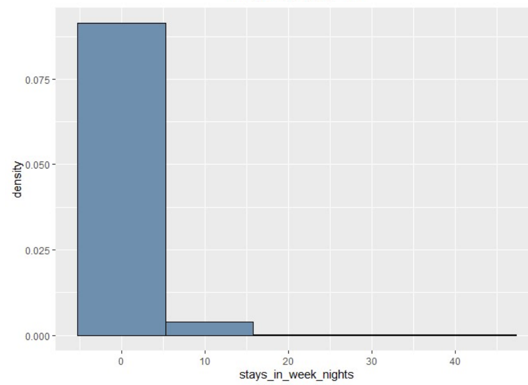
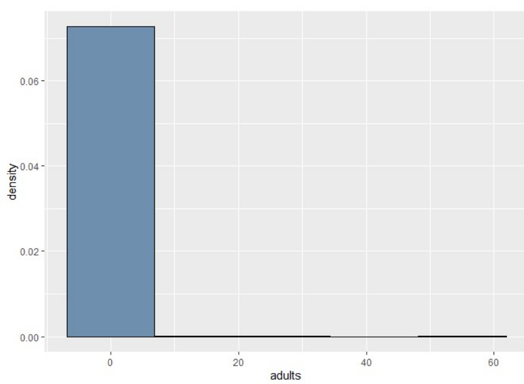
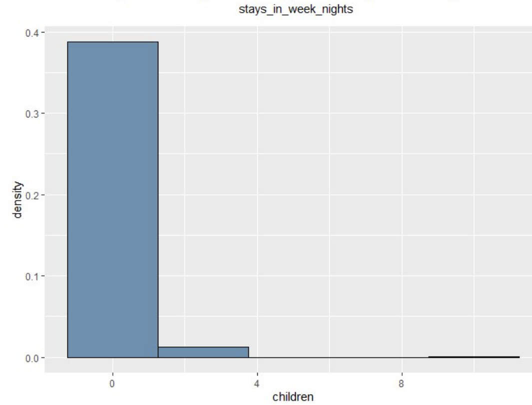
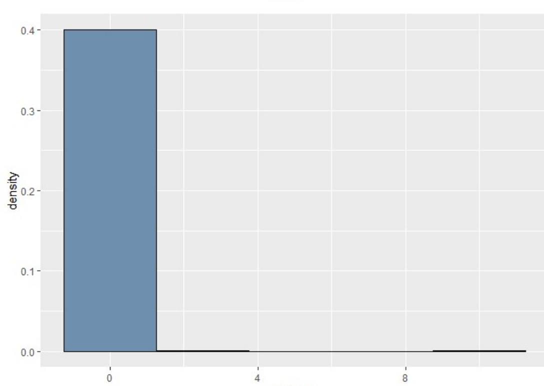
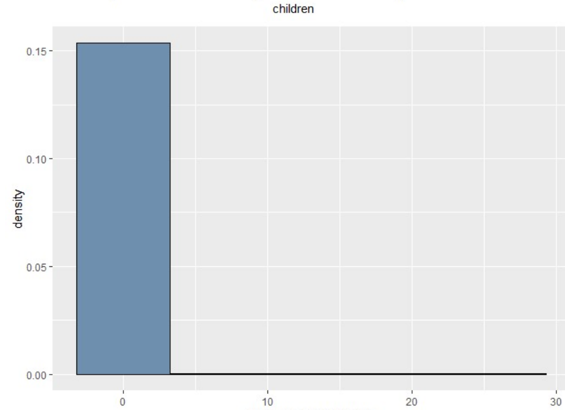
Categorical							
hotel		**arrival_date_month**		**meal**		**country**	
City Hotel	47586	**August**	8293	**BB**	55337	**PRT**	29092
Resort Hotel	24045	**July**	7632	**FB**	465	**GBR**	7257
		May	7133	**HB**	8711	**FRA**	6306
is_canceled		**October**	6699	**SC**	6397	**ESP**	5138
0	45099	**April**	6625	**Undefined**	721	**DEU**	4362
1	26532	**June**	6569			**ITA**	2223
		Others	28680	**customer_type**		**Others**	17253
deposit_type				**Contract**	2450		
No Deposit	62804	**is_repeated_guest**		**Group**	343		
No Refund	8732	**0**	69339	**Transient**	53821		

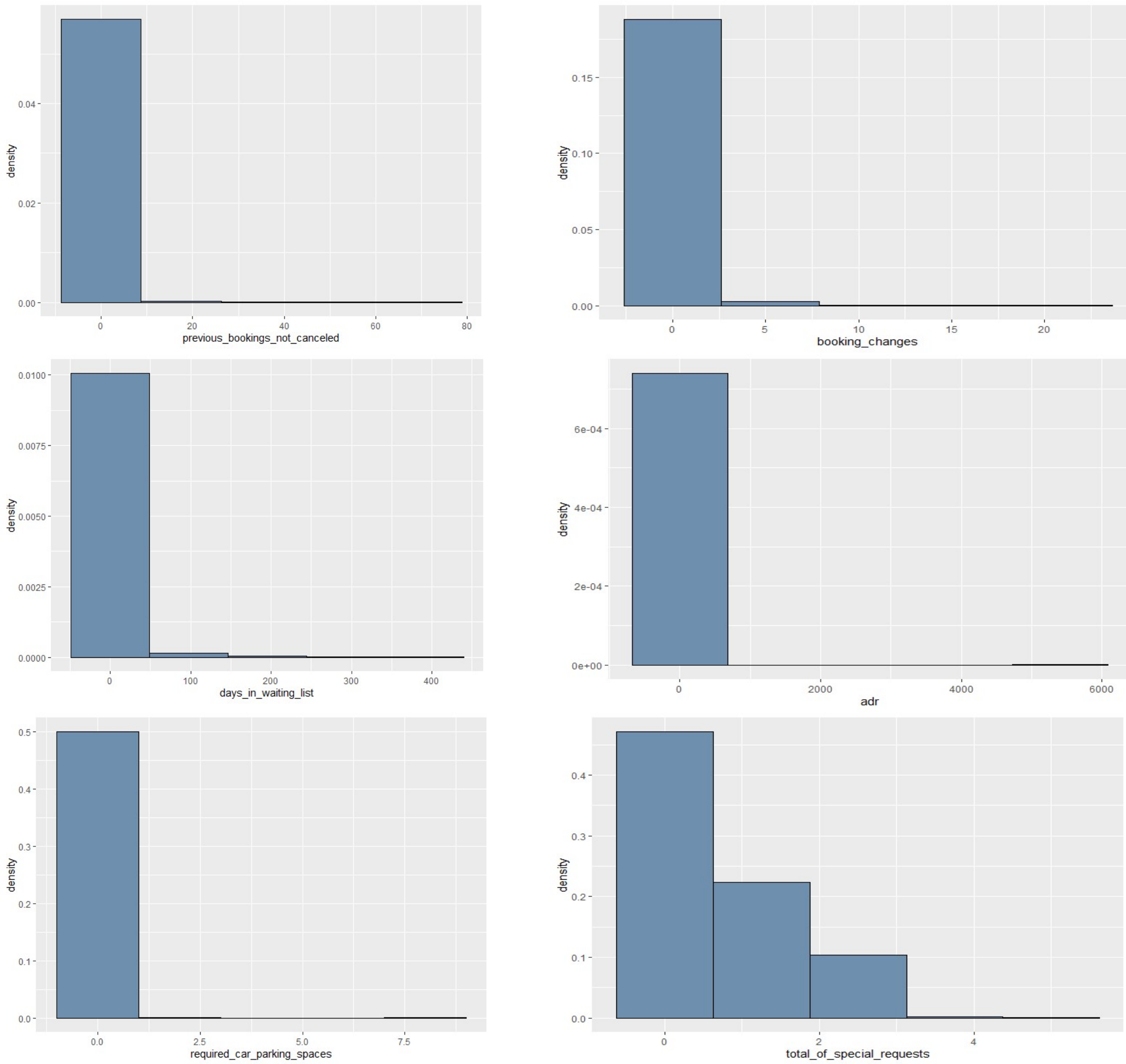
Refundable	95	**1**	2292	**Transient-party**	15017		
market_segment		**distribution_channel**		**company**			
Aviation	144	**Corporate**	4018	**NULL**	67566		
Complementary	415	**Direct**	8752	**40**	540		
Corporate	3162	**GDS**	127	**223**	464		
Direct	7608	**TA/TO**	58733	**67**	146		
Groups	11857	**Undefined**	1	**45**	141		
Offline TA/TO	14534			**153**	128		
Online TA	33911			**Other**	2646		
reserved_room_type		**assigned_room_type**		**agent**			
A	51591	**A**	44401	**9**	19252		
D	11477	**D**	15159	**NULL**	9781		
E	3952	**E**	4755	**240**	8302		
F	1763	**F**	2240	**1**	4327		
G	1259	**G**	1542	**14**	2174		
B	646	**C**	1433	**7**	2137		
Other	943	**Other**	2101	**Other**	25658		

We can see that in our dataset, certain variables have a mean that is higher than the median, while others have a median that is higher than the mean. So, in general, but not always, if the median is lower than the mean, we may see significant outliers at the high end of the distribution, while if the mean is lower than the median, we may see major outliers at the low end. To have a detailed overview of the variables, we will perform a univariate analysis to narrate the pattern of response to the variables.



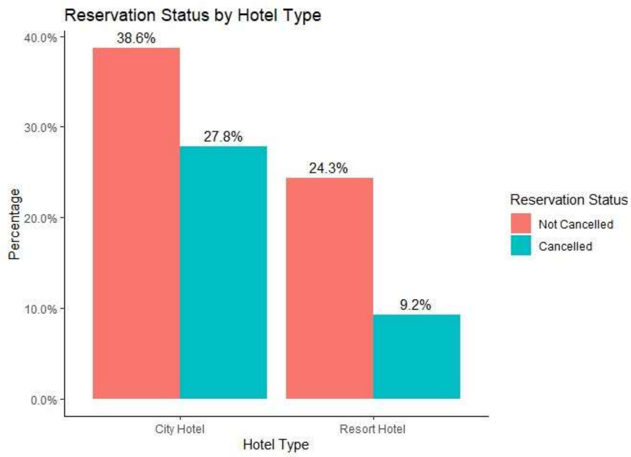



density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
arrival_date_week_number

density
0.00
0.01
0.02
0.03
0
10
20
30
arrival_date_day_of_month

density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
stays_in_weekend_nights

density
0.000
0.025
0.050
0.075
0
10
20
30
40
stays_in_week_nights

density
0.00
0.02
0.04
0.06
0
20
40
60
adults

density
0.0
0.1
0.2
0.3
0.4
0
4
8
children

density
0.0
0.1
0.2
0.3
0.4
0
4
8
babies

density
0.00
0.05
0.10
0.15
0
10
20
30
previous_cancellations

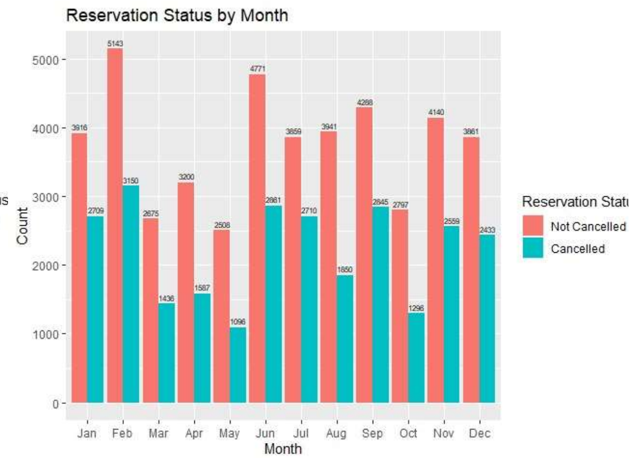


Above are the histograms shows us the distribution of the continuous data in the Hotel Booking dataset. Also we will plot the bar plot for the distribution to determine the existence of relationships between the variables and the class variable.

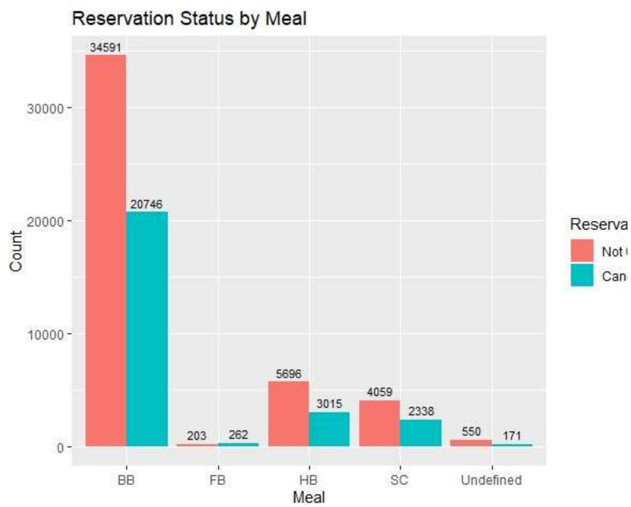
We use the bar graph to see the relationship between the variables as follows:



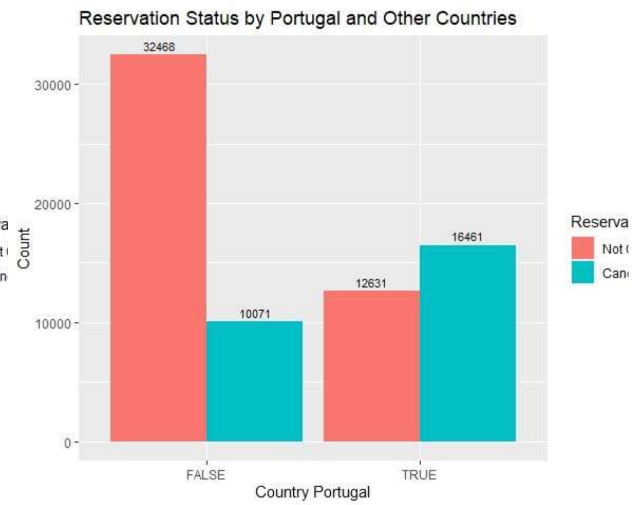
Reservation Status by Hotel Type
40.0%
30.0%
20.0%
10.0%
0.0%
Percentage
38.6%
27.8%
24.3%
9.2%
City Hotel
Resort Hotel
Hotel Type
Reservation Status
Not Cancelled
Cancelled



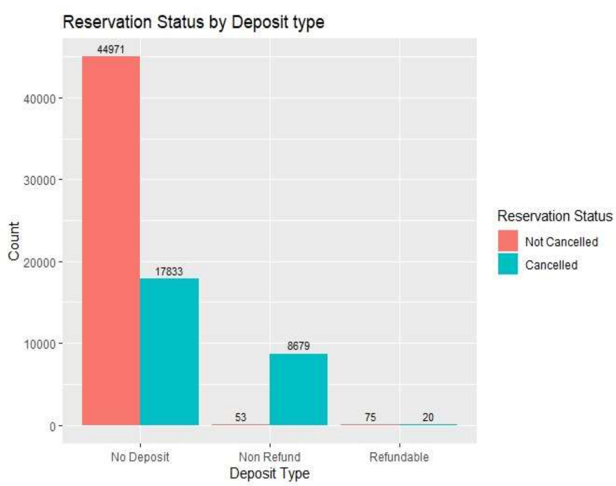
Reservation Status by Month
5000
4000
3000
2000
1000
0
Count
3916
2709
5143
3150
2675
1436
3200
1587
2508
1096
4771
2861
3859
2710
3941
1850
4268
2845
2797
1296
4140
2559
3861
2433
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Not Cancelled
Cancelled



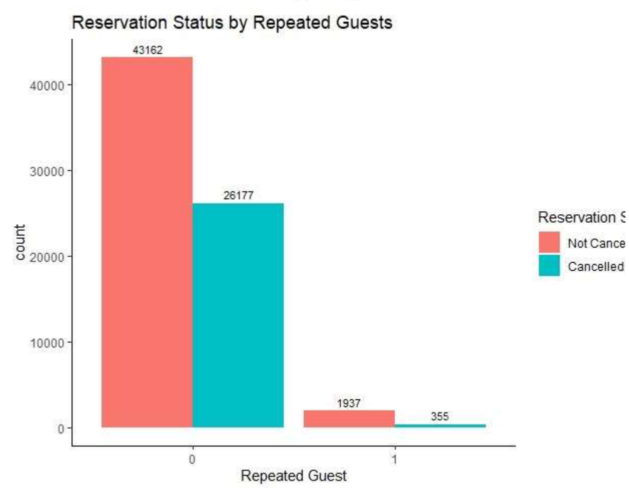
Reservation Status by Meal
30000
20000
10000
0
Count
34591
20746
203
262
5696
3015
4059
2338
550
171
BB
FB
HB
SC
Undefined
Meal



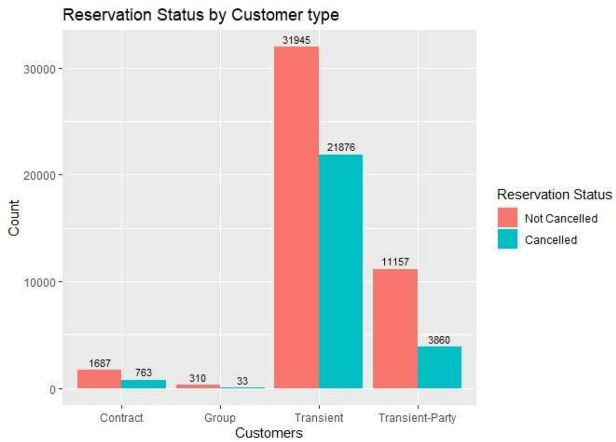
Reservation Status by Portugal and Other Countries
30000
20000
10000
0
Count
32468
10071
12631
16461
FALSE
TRUE
Country Portugal

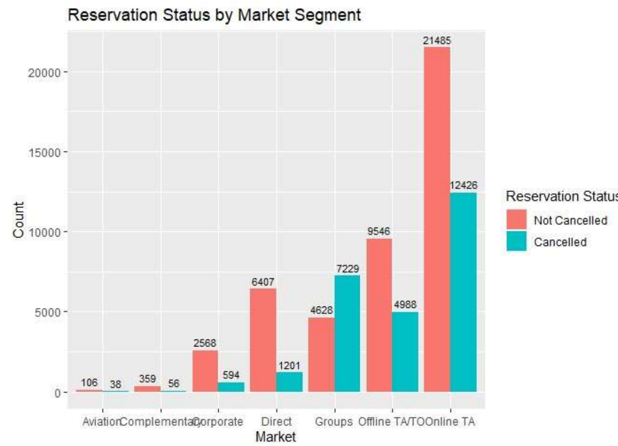


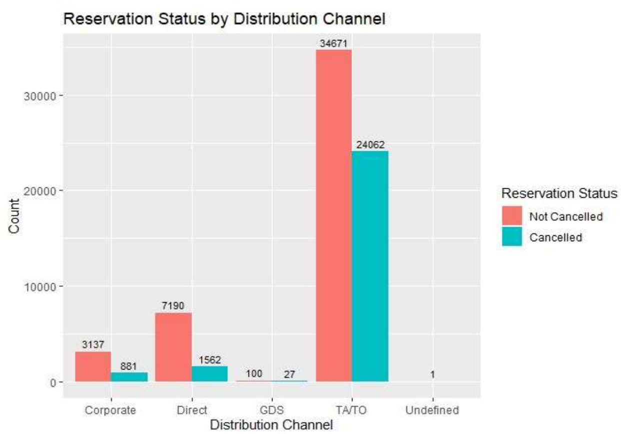
Reservation Status by Deposit type
40000
30000
20000
10000
0
Count
44971
17833
53
8679
75
20
No Deposit
Non Refund
Refundable
Deposit Type
Reservation Status
Not Cancelled
Cancelled

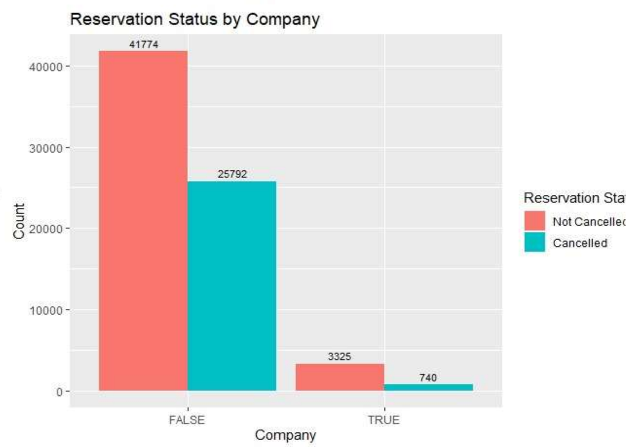


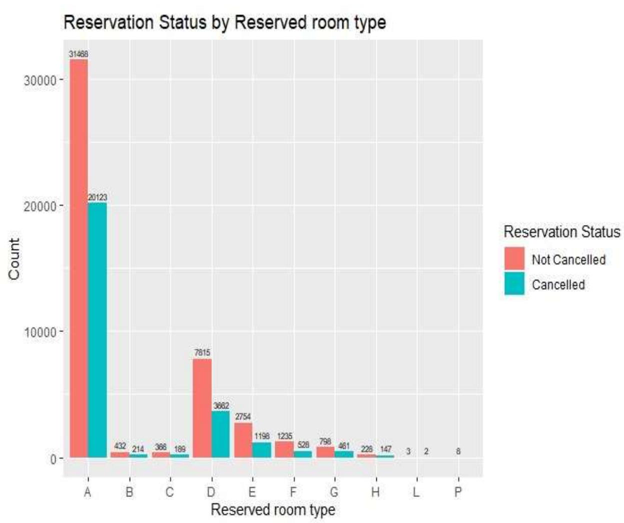
Reservation Status by Repeated Guests
40000
30000
20000
10000
0
count
43162
26177
1937
355
0
1
Repeated Guest
Cancelled

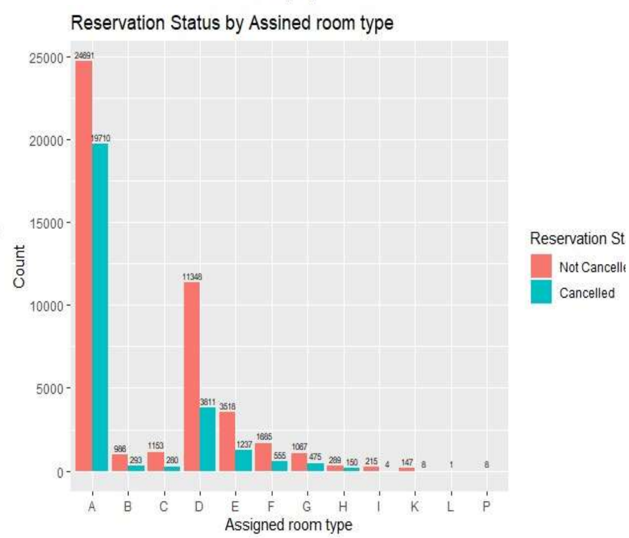

Reservation Status by Customer type
Count
Customers
Contract
Group
Transient
Transient-Party
1687
763
310
33
31945
21876
11157
3860
Reservation Status
Not Cancelled
Cancelled

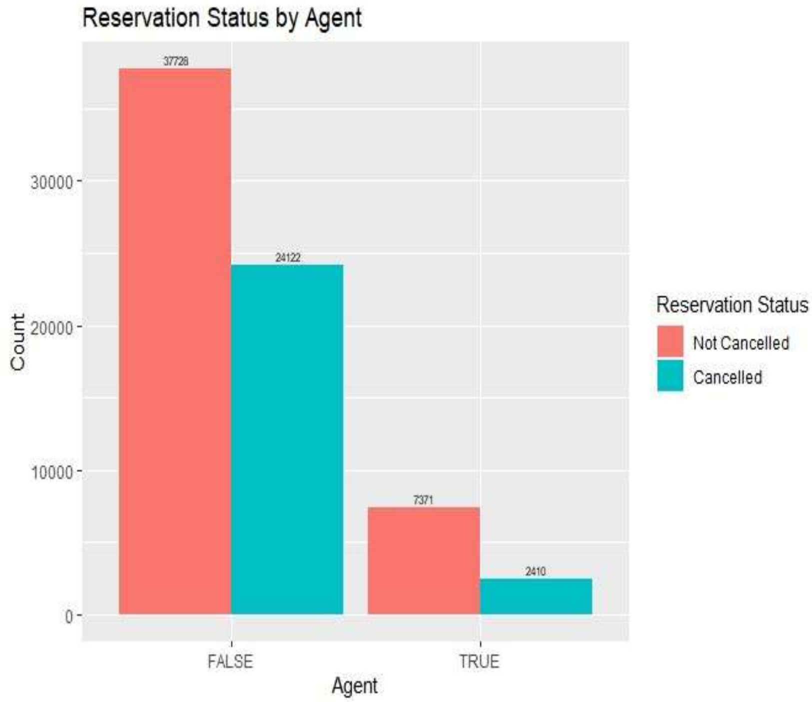

Reservation Status by Market Segment
Count
Market
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
106
38
359
56
2568
594
6407
1201
4628
7229
9546
4988
21485
12426
Reservation Status
Not Cancelled
Cancelled


Reservation Status by Distribution Channel
Count
Distribution Channel
Corporate
Direct
GDS
TA/TO
Undefined
3137
881
7190
1562
100
27
34671
24062
1
Reservation Status
Not Cancelled
Cancelled


Reservation Status by Company
Count
Company
FALSE
TRUE
41774
25792
3325
740
Reservation Status
Not Cancelled
Cancelled

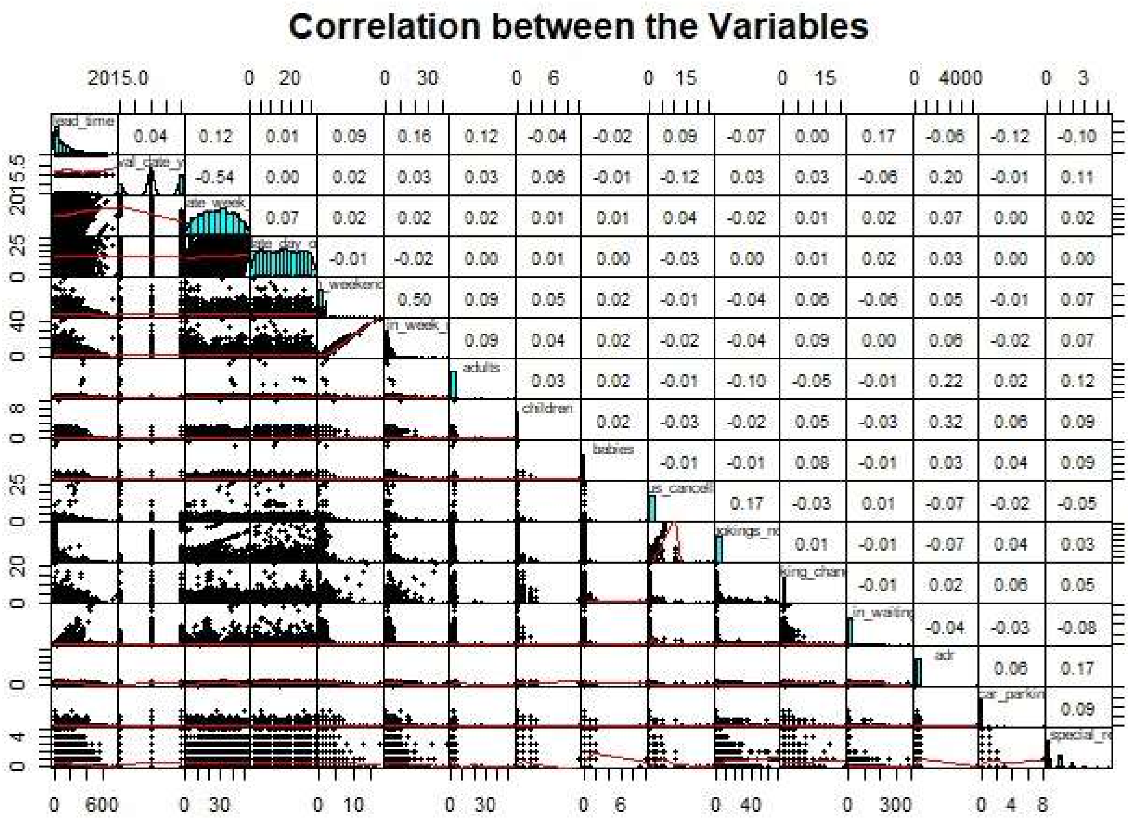

Reservation Status by Reserved room type
Count
Reserved room type
A
B
C
D
E
F
G
H
L
P
31468
20123
7815
3662
Reservation Status
Not Cancelled
Cancelled


Reservation Status by Assined room type
Count
Assigned room type
A
B
C
D
E
F
G
H
I
K
L
P
24891
19710
11348
3811
3518
Reservation Status
Not Cancelled
Cancelled

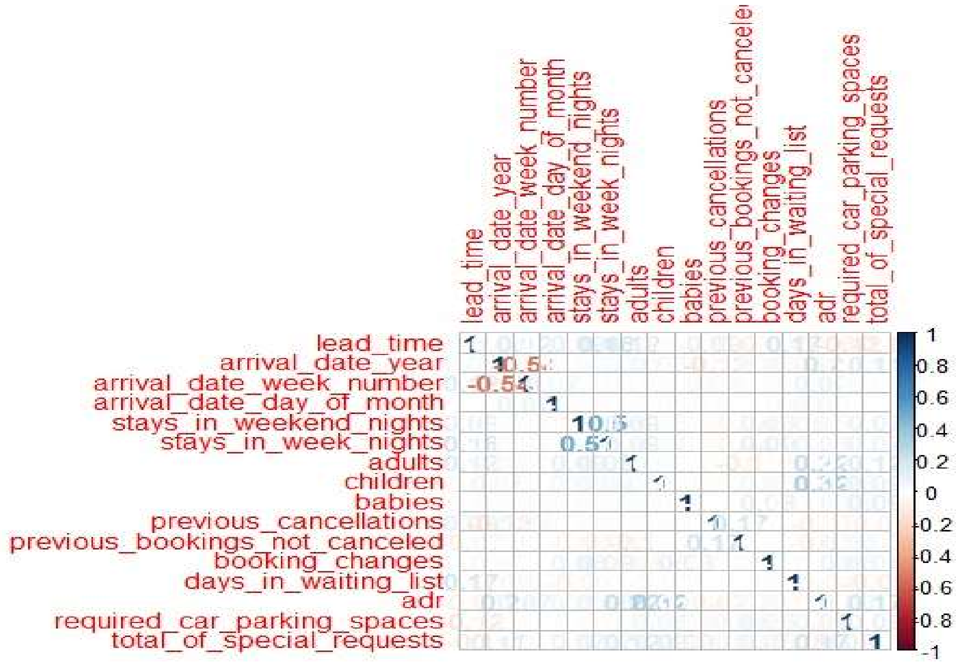


Above bar graph gives us the following information for the given data for both hotels:

- ➢ The City Hotel received around 65 percent of all booking inquiries, while the Resort Hotel received 35 percent.
- ➢ We can see from the graph that the months of April to October had the highest amount of hotel booking requests.
- ➢ The reservation has been made mostly for the BB – bread & breakfast for both hotels.
- ➢ The reservation for both the hotels mostly made from the peoples, who don't belong to Portugal.
- ➢ Most of the tourist didn't pay any deposit at the time of reservation.
- ➢ Very few of the tourists have visited previously in both the hotels.
- ➢ Most of the customer are from "Transient" and "Online TA" also the distribution channel was "TA/TO".
- ➢ At the time of reservation, mostly the reserved room and the assigned room at arrival was A.
- ➢ Most of the tourists didn't have any agent and company at the time of reservation.



This is a preliminary analysis of the data in the original space, the upper triangle of the matrix there are the coefficients of correlation between variables. Specifically, if we look at the plot, we can see that there is a correlation between some variables also there is no correlation between variables as we can see that the correlation between children and arrival_date_week_number is 0.01 which is nearly 0, arrival_date_year & arrival_date_week_number, which is -0.54 have a negative correlation also stays_in_weekend_nights & stays_in_week_nights, which is 0.50 having a positive correlation.



As we can observe that the positive correlation for stays_in_weekend_nights & stays_in_week_nights is having a positive correlation but arrival_date_year & arrival_date_week_number having a negative correlation.

Modeling

Logistic Regression

```
[1] 71631
[1] 23880
glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = is_canceled ~ hotel + lead_time + arrival_date_year +
    arrival_date_month + stays_in_weekend_nights + arrival_date_day_of_month +
    stays_in_week_nights + arrival_date_week_number + adults +
    children + babies + meal + market_segment + distribution_channel +
    is_repeated_guest + previous_cancellations + previous_bookings_not_canceled +
    booking_changes + deposit_type + days_in_waiting_list + customer_type +
    adr + total_of_special_requests + required_car_parking_spaces,
    family = binomial, data = train_hb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.7893  -0.4677   0.3458   5.6286
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  -4.348e+02  4.205e+01 -10.341  < 2e-16 ***
hotel                        -4.590e-02  2.228e-02  -2.060 0.039390 *  
lead_time                     2.622e-03  1.155e-04  22.699  < 2e-16 ***
arrival_date_year             2.115e-01  2.087e-02  10.130  < 2e-16 ***
arrival_date_month            3.479e-01  1.069e-01   3.254 0.001140 ** 
stays_in_weekend_nights       4.188e-02  1.124e-02   3.727 0.000194 ***
arrival_date_day_of_month     1.228e-02  3.696e-03   3.324 0.000887 ***
stays_in_week_nights          3.571e-02  5.931e-03   6.021 1.73e-09 ***
arrival_date_week_number     -7.625e-02  2.455e-02  -3.106 0.001898 ** 
adults                        1.199e-01  2.020e-02   5.937 2.90e-09 ***
children                      1.764e-01  2.374e-02   7.431 1.08e-13 ***
babies                        2.329e-01  9.677e-02   2.407 0.016092 *  
meal                         -1.432e-02  8.573e-03  -1.670 0.094859 .  
market_segment                5.487e-01  1.609e-02  34.091  < 2e-16 ***
distribution_channel         -3.159e-01  2.131e-02 -14.821  < 2e-16 ***
is_repeated_guest            -5.927e-01  1.030e-01  -5.752 8.80e-09 ***
previous_cancellations        3.071e+00  7.929e-02  38.730  < 2e-16 ***
previous_bookings_not_canceled -4.674e-01  3.177e-02 -14.709  < 2e-16 ***
booking_changes              -3.881e-01  1.940e-02 -20.007  < 2e-16 ***
deposit_type                  4.469e+00  7.970e-02  56.079  < 2e-16 ***
days_in_waiting_list         -7.671e-04  6.055e-04  -1.267 0.205225    
customer_type                -2.562e-02  1.837e-02  -1.394 0.163187    
adr                           4.206e-03  2.436e-04  17.266  < 2e-16 ***
total_of_special_requests    -6.217e-01  1.412e-02 -44.041  < 2e-16 ***
required_car_parking_spaces  -2.361e+04  9.883e+05  -0.024 0.980939    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 94434  on 71630  degrees of freedom
Residual deviance: 64757  on 71606  degrees of freedom
AIC: 64807

Number of Fisher Scoring iterations: 14


glm.pred2_L     0     1
        No  42854 12412
        Yes  2245 14120

glm.pred2_L     0     1   Sum
        No  42854 12412 55266
        Yes  2245 14120 16365
        Sum 45099 26532 71631
[1] 20.46181
```

```
glm.pred2_T     0     1
        No  14341  4046
        Yes   693  4800

glm.pred2_T     0     1   Sum
        No  14341  4046 18387
        Yes   693  4800  5493
        Sum 15034  8846 23880
[1] 19.84506
```

We can see the p value of all coefficients are significantly small so we can reject null hypothesis and realize these variables are completely related to the model.
Observing the above confusion matrix, we can see that the diagonal elements indicate correct predictions, while other represents incorrect predictions. In general 14341 + 4800 are true-positive & true-negative and 4046+693 false-positive and false-negative predicted. While delta is 19.84 which is quite higher.

Neural Network

```
                    Length Class      Mode
call                    6  -none-     call
response             1000  -none-     numeric
covariate           24000  -none-     numeric
model.list              2  -none-     list
err.fct                 1  -none-     function
act.fct                 1  -none-     function
output.act.fct          1  -none-     function
linear.output           1  -none-     logical
data                   25  data.frame list
exclude                 0  -none-     NULL
net.result              5  -none-     list
weights                 5  -none-     list
generalized.weights     5  -none-     list
startweights            5  -none-     list
result.matrix         280  -none-     numeric
                           [,1]         [,2]         [,3]         [,4]
error                115.05950078 115.059500019 113.214853602 115.059500577
reached.threshold      0.00910566   0.001418172   0.006267145   0.007819014
steps                 28.00000000  23.000000000  30.000000000  33.000000000
Intercept.to.1layhid1 -0.50219235   1.897465700  -0.875869629  -0.261016314
V1.to.1layhid1         0.13153117  -2.271925486  -0.363100999  -0.642269499
V2.to.1layhid1        -0.07891709   0.980464139   1.247008646  -0.340968618
                            [,5]
error                115.059500922
reached.threshold      0.009882676
steps                 18.000000000
Intercept.to.1layhid1 -1.330034111
V1.to.1layhid1        -0.850580314
V2.to.1layhid1        -1.788830742
```

```
[1] 115.0595 115.0595 113.2149 115.0595 115.0595
[1] 3
     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
[5,]    0
[6,]    0

          [,1]
[1,] 0.3491061
[2,] 0.3491061
[3,] 0.3491061
[4,] 0.3491061
[5,] 0.3491061
[6,] 0.3491061
     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
[5,]    0
[6,]    0
          yhat_test
y_valid_hb     0     1   Sum
       0   14794   240 15034
       1    8371   475  8846
       Sum 23165   715 23880
[1] 0.6394054
```

Observing the above confusion matrix, we can see that the diagonal elements indicate correct predictions, while other represents incorrect predictions. In general 14794 + 475 are true-positive & true-negative and 240+8371 false-positive and false-negative predicted. While accuracy is better than the logistic regression but still not good enough.

Random Forest

```
Call:
 randomForest(formula = is_canceled ~ hotel + lead_time + arrival_date_year +
arrival_date_month + stays_in_weekend_nights + arrival_date_day_of_month +
stays_in_week_nights + arrival_date_week_number + adults +      children + babies + meal
+ market_segment + distribution_channel +      is_repeated_guest + previous_cancellations
+ previous_bookings_not_canceled +      booking_changes + deposit_type +
days_in_waiting_list + customer_type +      adr + total_of_special_requests +
required_car_parking_spaces,      data = train_hb, mtry = 24, ntree = 10)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 24

          Mean of squared residuals: 0.1280721
                    % Var explained: 45.08
```

```
 [1] 0.1085559
 [1] 0.3294782
```

```
1 2 3 4 5 6
0 0 1 0 0 0
           yhat_test
y_valid_hb     0     1   Sum
         0 13653  1381 15034
         1  2201  6645  8846
         Sum 15854  8026 23880
[1] 0.85
```

Observing the above confusion matrix, we can see that the diagonal elements indicate correct predictions, while other represents incorrect predictions. In general 13653 + 6645 are true-positive & true-negative and 1381+2201 false-positive and false-negative predicted. While accuracy is better than the both neural-network and logistic regression.

We can clearly observe that the better model for the hotel booking prediction is Random Forest.