PROJECT REPORT

ABSTRACT

Expedia wants to predict which customers are likely to book hotels. LETS GET THEM SORTED…..!!!

Vikrant Dev Rathore(VDR180000)
Siddarth Basavanneppa Sheshagiri(SBS180000)
Laxmi Supriya Ketireddy(LXK170003)
Prarthna Vasudevamurthy(PXV180009)
Sanjana Patil(SXP180071)

Group 1

Table of Contents

List of figures and Tables

Acknowledgements

We would like to express our gratitude towards our professor, Dr.Ling Ge for her timely inputs and corrections in order for us to proceed further.

We are also grateful to Kaggle.com and Expedia for providing the dataset for us to make our project.

Vikrant Dev Rathore(VDR180000)
Siddarth Basavanneppa Sheshagiri(SBS180000)
Laxmi Supriya Ketireddy(LXK170003)
Prarthna Vasudevamurthy(PXV180009)
Sanjana Patil(SXP180071)
(Group 1)

Executive Summary

Goal: The main aim of the project is to help Expedia decide which customers are most likely to book hotels and which customers will only surf on their website(comprise of click data). This will help Expedia in designing better marketing strategies to cater to customers who are most likely to book.

Dataset: The dataset has been taken from Kaggle.com in which Expedia provides information on customer behavior like what customers searched for, how likely were they to book, whether the search result was provided in a travel package. The dataset is only a sample of the population and does not explain the overall statistics. The project uses the file **Train.csv.**

Data Description:

- **Train :** The dataset **contains data from 2013,2014** about users which were involved in booking events. Test dataset **contains 24 variables** which deal with dates, ID of the Expedia point of sale, ID of country, region and city where customer is located, Physical distance between the hotel site and the customer's location , whether user connected by mobile or not, ID of the user, whether booking was generated, ID of the marketing channel, Check in Date and Check out Date, The number of adults and children and the number of hotel rooms specified. It **contains more than 37 million records! As a result, the whole file cannot be opened in R and a random sample of the records were taken, loaded as chunks and analyzed in R.**

Algorithms:

As a part of this project we have implemented the following machine learning algorithms in order to classify customers who are likely to book and who will not book:

- Logistic Regression with threshold 0.4
- LDA
- Naïve Bayes Classifier

Project Motivation/Background

With cut throat competition in today's world, companies are looking for methods to optimize their profits. One way of achieving is to find the target audience and work on strategic marketing. With more than 300 million visits in 2013 and 2014, it is a daunting task for Expedia to find customers who are most likely to book hotels.

Expedia uses search parameters to adjust the user's hotel recommendations, but Expedia can further improve its business by finding users that are likely to book and create special discounts, better hotel packages. Not only this, Expedia can also create special offers for customers surf the website (not likely to book) like first booking discount etc. in order to increase the likelihood that these customers book and not just surf the website.

In this project, we take on the challenge for predicting for Expedia which customers are most likely to book hotels and which customers will comprise of click data. This in turn will help Expedia organize its customer base better and improve the overall functionality of their business thus optimizing profits, customer service and user experience.

Data Description

The dataset was obtained from kaggle.com in which Expedia provides information on customer behavior like what customers searched for, how likely were they to book, whether the search result was provided in a travel package. The dataset is only a sample of the population and does not explain the overall statistics. We are given the hotel clusters and each user is assigned to a hotel cluster which he is likely to book using in house algorithms that Expedia has not provided. The aim is to predict which customer will book hotels and which will not.

The file train.csv has more than 37 million records and hence cannot be opened in R due to RAM limitations. We opened the files in chunks and then created a random sample from the file 395,054 records out of which 208,299 observations for click data and 186,755 records for booking of hotels. This was expected that the booking records will be less as in the original 37 million records, nearly 2/3 of the data were records for clicks(no booking).

The dataset provided by Expedia has 24 variables

Variable Name	Description
date_time	Time stamp
site_name	ID of the Expedia point of sale
posa_continent	ID of continent associated with site_name
user_location_country	The ID of country customer is located
user_location_region	The ID of region customer is located

user_location_city	The ID of city the customer is located
orig_destination_distance	Physical distance between a hotel and a customer at the time of search.
user_id	ID of user
is_mobile	1 when a user connected from a mobile device, 0 otherwise
is_package	1 if the click/booking was generated as a part of a package, 0 otherwise
channel	ID of a marketing channel
srch_ci	Check-in date
srch_co	Check-out date
srch_adults_cnt	The number of adults specified in the hotel room
srch_children_cnt	The number of children specified in the hotel room
srch_rm_cnt	The number of hotel rooms specified
srch_destination_id	ID of the destination where the hotel search was performed

srch_destination_type_id	Type of destination
hotel_continent	Hotel continent
hotel_country	Hotel country
hotel_market	Hotel market
is_booking	1 if a booking, 0 if a click
cnt	Number of similar events in the context of the same user session
hotel_cluster	ID of a hotel cluster

Table 1: Dataset variables

Exploratory Data Analysis

The dataset had 37 million records out of which we created a sample which had 395,054 records. The first thing to check was whether our sample had missing values, so we visualized whether there were missing values in our dataset.

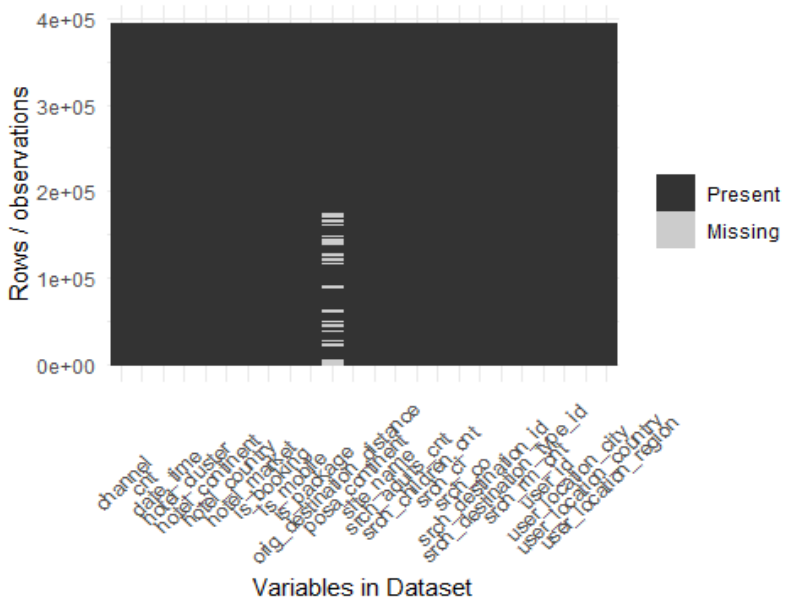


Figure 1: missing values plot

From Figure 1 we observed that the variable orig_destination_distance had missing values, so we removed the missing values from our dataset and we were left 332,106 records.

Next, we decided to check the correlation between the variables to see if any of the variables had any strong correlations.

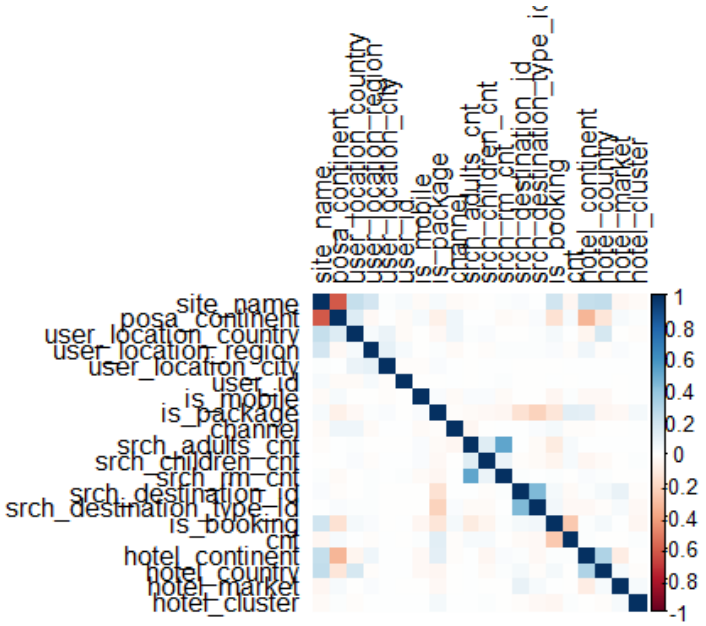


Figure 2: Correlation Plot

From figure 2, we can observe some correlation posa_continent and site_name , other than that our variables do not show any strong correlations.

We then decided to plot factors that might be of importance with relation to booking/non booking. The graphs are presented below

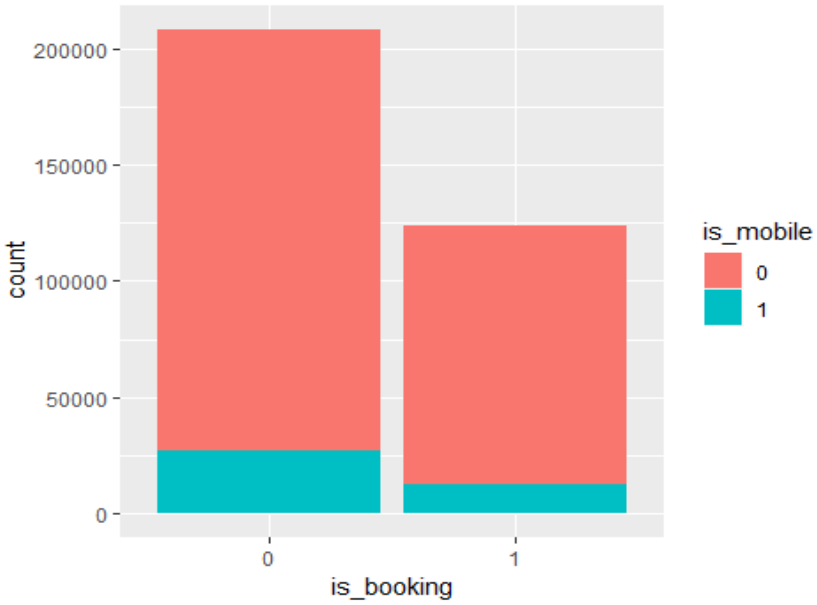


Figure 3: is_booking by mobile/non mobile devices

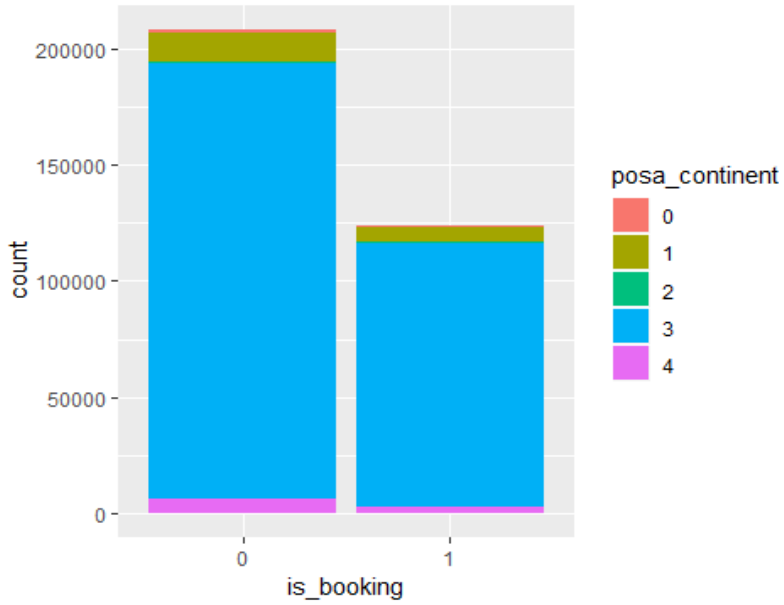


Figure 4: is_booking by posa_continent

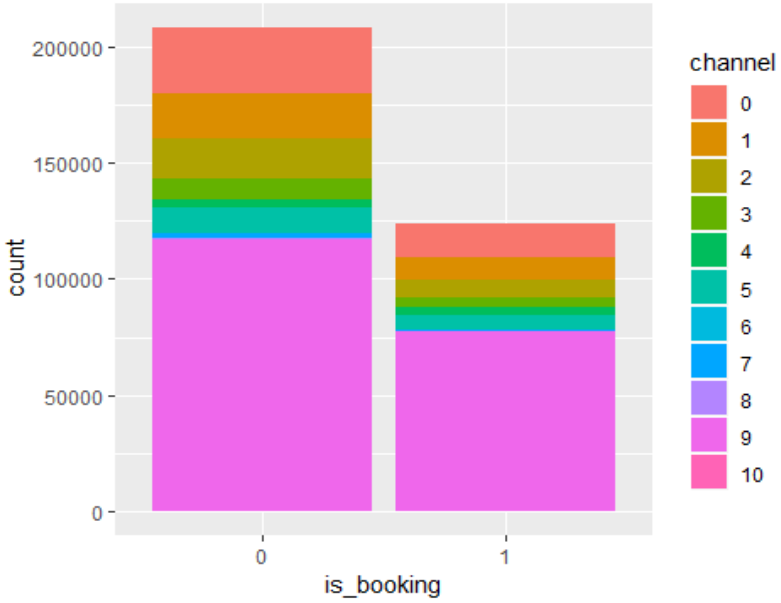


Figure 5:is_booking by channel

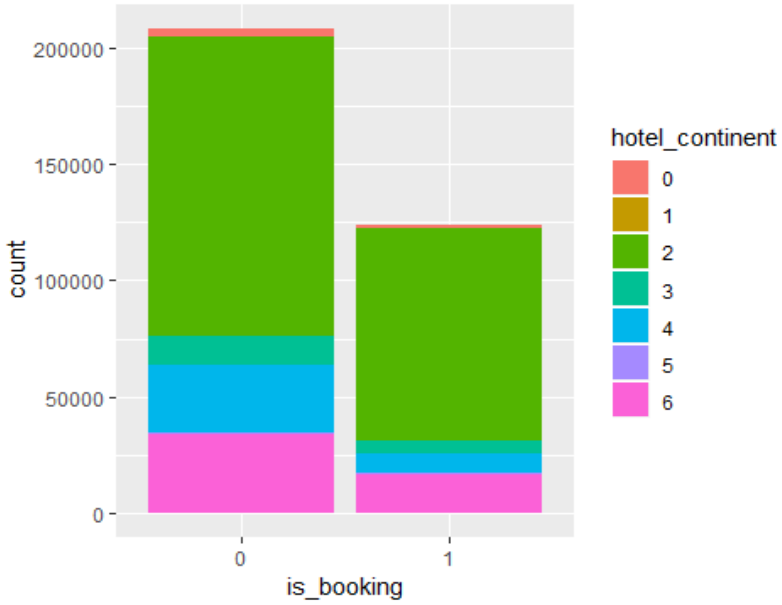


Figure 6: is_booking by hotel_continent

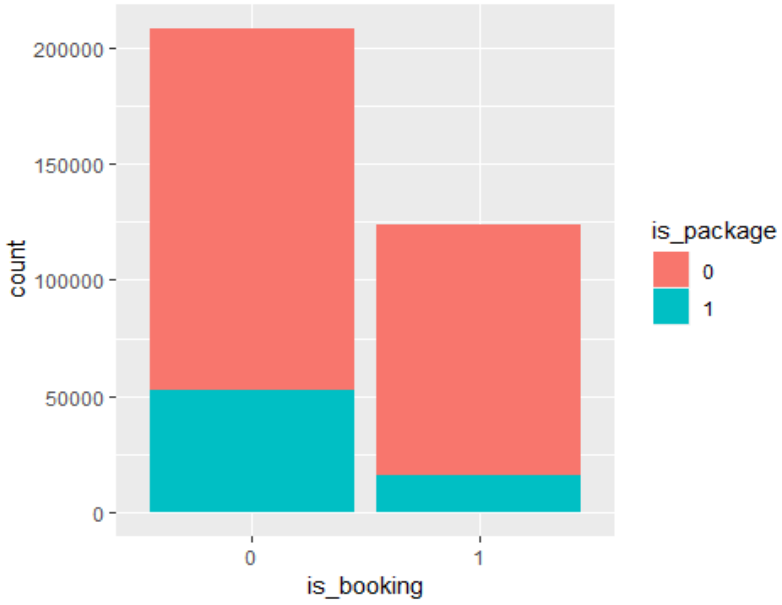


Figure 7: is_booking by packages

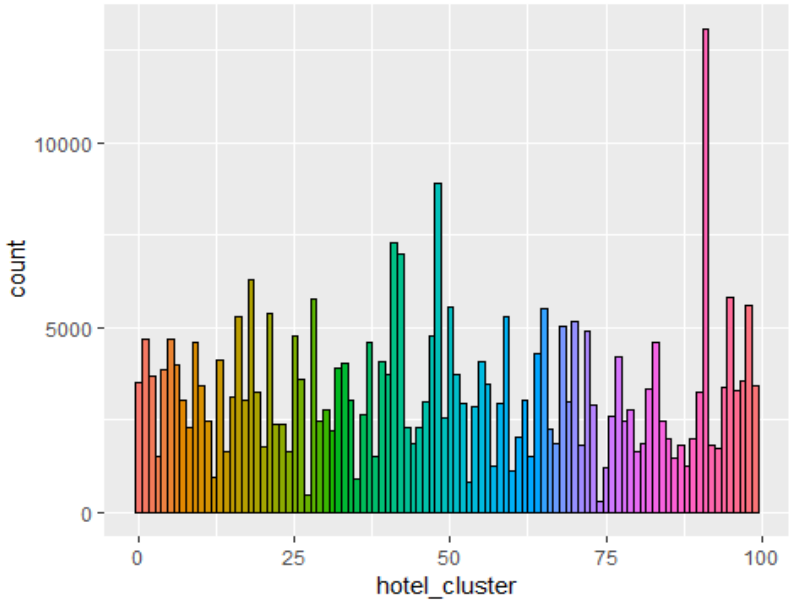


Figure 8: hotel clusters

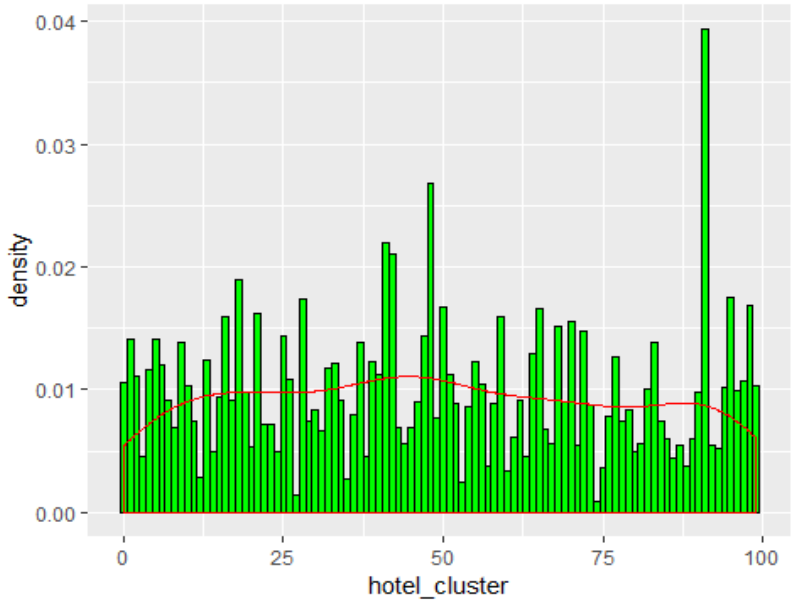


Figure 9: hotel clusters density plot

From Figure 3 we observe that most customers book/click from devices that are not mobile. Figure 4 shows that most sales, i.e. most bookings of customers takes place in continent 3. Expedia can create special offers for these customers in order to retain them. Figure 5, shows that most bookings happen through booking channel 9. From Figure 6, we conclude that the most preferred destination for users is continent 2, therefore Expedia can create special packages for this continent in order to improve their business. Figure 7 shows some shocking results! Most

customers did not book the packages Expedia had to offer, rather they booked without packages. A possible reason could be that maybe the packages offered by Expedia were not lucrative to customers, therefore Expedia needs to work on its packages in order to attract more customers.

Figure 8 and 9 shows that the customers are allocated clusters 1-100 and the plot of clusters shows that there is some skewness in the distribution of hotel clusters.

Models and Analysis

We ran logistic regression, LDA and Naïve Bayes Classifier on the dataset in order to make predictions about bookings. Since most of the variables are categorical in nature but are presented as integers in the dataset, we converted them to factors before running our models.

1. **Logistic Regression on the whole dataset**: Logistic regression was performed on the entire dataset to obtain results. The model accuracy was 63.31% and the threshold was set at 0.4. The accuracy of bookings and clicks are summarized in table 2 below:

Prediction	% correct
booking	76.23%
click	55.64%

Table 2: Logistic regression on whole dataset

From table 2 , we can see that that the model is able to predict bookings correctly to 76.23%, while clicks were predicted correctly upto 55.64%

We divided the data randomly into train and test sets such that train had 70% of the data and test had 30%. The models were built on the train dataset and then tested on the test dataset in order to find their accuracy.

2. **Logistic Regression:** This time, we used the train and test sets in to obtain the results. The model accuracy was 63.45% and the threshold was set at 0.4. The accuracy of bookings and clicks are summarized in table 3 below:

Prediction	% correct
booking	75.46%
click	56.18%

Table 3: Logistic regression on test set

The overall accuracy of the model increased. From table 3, the correct% of booking predictions fell to 75.46%, however the correct% of clicks increased to 56.18%

3. **LDA**: we performed LDA using the train and test sets to obtain the results. The overall model accuracy is 65.97%. The accuracy of bookings and clicks are summarized in table 4 below:

Prediction	% correct
Booking	24.25%
Click	91.20%

Table 4: LDA on test set

We see that the overall accuracy of the model increases, from table 4 we see however bookings are predicted correctly on upto 24.25% while Clicks are predicted correctly 91.20%! A possible reason for these observations could be because the dataset contains a lot of observations for clicks and hence the model could be biased towards clicks because of that.

4. **Naïve Bayes Classifier**: The naïve Bayes classifier was used to classify into bookings and non-bookings using the train and test sets. The overall accuracy of the model is 59.36%. The results of correct bookings and clicks are summarized in table 5 below.

Prediction	% correct
Booking	89.49%
Click	41.14%

Table 5: Naïve Bayes Classifier on test set

We see that the overall accuracy of the model decreased, however from table 5 the bookings were predicted with an accuracy of 89.49% while click was predicted with an accuracy of 41.14%.

Findings and Managerial Implications

The findings are summarized below in table 6:

Model	Accuracy %	Booking correct%	Click correct %
Logistic Regression	63.31%	76.23%	55.64%
Logistic regression on test	63.45%	75.46%	56.18%
Linear Discriminant Analysis on test	65.97%	24.25%	91.20%
Naïve Bayes Classifier on test	59.36%	89.49%	41.14%

Table 6: Summary of results

- From the table above if we were to select a model based on accuracy then LDA would be the most accurate model, however we see it is biased towards click.
- Naïve Bayes Model does an excellent at predicting which customers are likely to book with a fairly decent prediction for clicks as well.
- The logistic regression model gives a good accuracy and gives good predictions for both bookings and clicks respectively.

The choice of selection of model that Expedia should use will depend on the cost of advertising to customers who will not book and the cost of not advertising to customers who are likely to book. Therefore, the model which minimizes the advertising cost and optimizes profits should be chosen. Hence depending on the scenario, Logistic regression and Naïve Bayes Classifier works well. In any case, we should not use the LDA because of its biased results.

Conclusions

The hotel booking dataset provided by Expedia was analyzed using machine learning algorithms such as logistic regression, LDA and Naïve Bayes classifier in order to help predict which customers are likely to book based on the variables in the dataset. The amount of records (37 million) and the fact that most of the data is categorical makes this problem challenging. In this project, using the various machine learning algorithms, we were able to predict which customers are likely to book. We observed that the highest accuracy was given by LDA, however because of its biased results towards clicks (We are interested in bookings), the model should not be used. The logistic regression model and Naïve Bayes Classifiers give good results with respect to bookings. The choice of model selection will ultimately depend on the cost of advertising to a customer who will not book and the cost of missing out on a potential customer.

Appendix

Project

Group 1

November 4, 2018

Load libaries

```
# to add the weekly dataset
library("ISLR")

#Loading MASS for LDA
library("MASS")

#Loading Class for classification functions
library("class")

#visualizations
library("ggplot2")
library("Amelia")
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
#tidyverse
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------
---------- tidyverse 1.2.1 --
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------
---- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```r
#reshape for melt
library("reshape")

##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths

## The following object is masked from 'package:class':
##
##     condense

#correlation Plot
library("corrplot")

## corrplot 0.84 loaded

#naive bayes
library("naivebayes")

# function taken from https://www.r-bloggers.com/ggplot-your-missing-data-2/
ggplot_missing <- function(x){

  x %>%
    is.na %>%
    melt %>%
    ggplot(data = .,
           aes(x = X2,
               y = X1)) +
    geom_raster(aes(fill = value)) +
    scale_fill_grey(name = "",
                    labels = c("Present","Missing")) +
    theme_minimal() +
    theme(axis.text.x  = element_text(angle=45, vjust=0.5)) +
    labs(x = "Variables in Dataset",
         y = "Rows / observations")

}

bookingfile<-read.csv("expedia.csv")

#creating a missing free dataset
bookingstrain1<-na.omit(bookingfile)
bookingstrain1<-bookingstrain1[,c(-1,-12,-13)]
```

```
#new dataset
bookingstrain2<-bookingfile[,c(-1,-12,-13,-7)]
bookingstrain2<-cor(bookingstrain2)
bookingstrain2
```

```
##                            site_name posa_continent user_location_country
## site_name                 1.000000000   -0.616147217           0.231479868
## posa_continent           -0.616147217    1.000000000           0.147252021
## user_location_country     0.231479868    0.147252021           1.000000000
## user_location_region      0.180969268   -0.030801526           0.033707114
## user_location_city        0.010334285    0.004640173           0.084704567
## user_id                   0.038222154   -0.025670749          -0.028662572
## is_mobile                -0.024477135    0.038036173          -0.009910088
## is_package                0.032928716   -0.073725886          -0.034828049
## channel                  -0.023303704    0.065832566           0.062438471
## srch_adults_cnt          -0.013464493    0.002905233           0.015586713
## srch_children_cnt        -0.006769078    0.005545270           0.025402316
## srch_rm_cnt               0.018988038   -0.027261595          -0.006202595
## srch_destination_id       0.029621276   -0.013845732          -0.002890544
## srch_destination_type_id  0.002867725    0.029287630           0.019474155
## is_booking                0.196617452   -0.151940208           0.040549107
## cnt                      -0.042211677    0.039569908           0.002351495
## hotel_continent           0.234236542   -0.338583407          -0.050474940
## hotel_country             0.242938750   -0.136259865           0.176767076
## hotel_market             -0.047992313    0.033624502          -0.001790209
## hotel_cluster            -0.025330185    0.014860660          -0.016721485
##                          user_location_region user_location_city
## site_name                         0.180969268       0.0103342850
## posa_continent                   -0.030801526       0.0046401727
## user_location_country             0.033707114       0.0847045673
## user_location_region              1.000000000       0.1060296932
## user_location_city                0.106029693       1.0000000000
## user_id                           0.037229206      -0.0117826122
## is_mobile                        -0.004425862      -0.0141502614
## is_package                        0.013106185       0.0235545846
## channel                          -0.021271351       0.0063274453
## srch_adults_cnt                  -0.009902828      -0.0009521936
## srch_children_cnt                -0.011592178       0.0025215038
## srch_rm_cnt                       0.010987160       0.0035742093
## srch_destination_id               0.015412121      -0.0035729519
## srch_destination_type_id          0.011659862      -0.0067785836
## is_booking                        0.028221648      -0.0093806457
## cnt                              -0.010349643       0.0023505614
## hotel_continent                   0.063118074       0.0008229693
## hotel_country                    -0.016490457       0.0025461751
## hotel_market                      0.029446138       0.0038909068
## hotel_cluster                     0.002691211       0.0014228847
##                                 user_id     is_mobile   is_package
## site_name                  3.822215e-02 -0.0244771349  0.032928716
## posa_continent            -2.567075e-02  0.0380361728 -0.073725886
```

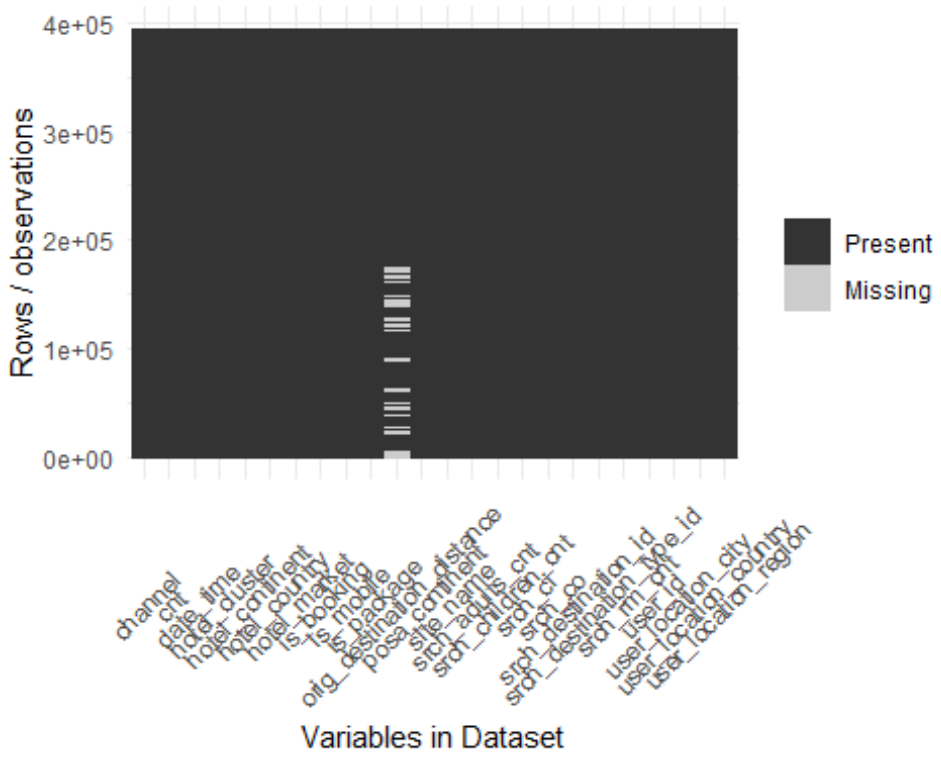
```
## user_location_country    -2.866257e-02 -0.0099100881 -0.034828049
## user_location_region      3.722921e-02 -0.0044258621  0.013106185
## user_location_city       -1.178261e-02 -0.0141502614  0.023554585
## user_id                   1.000000e+00 -0.0099717634 -0.004723798
## is_mobile                -9.971763e-03  1.0000000000  0.029113921
## is_package               -4.723798e-03  0.0291139209  1.000000000
## channel                   3.496742e-03 -0.0289515699 -0.018959450
## srch_adults_cnt          -3.961441e-03  0.0240574415 -0.023008799
## srch_children_cnt         9.569875e-03  0.0252655653 -0.030530594
## srch_rm_cnt               1.069193e-02 -0.0272624865 -0.046021732
## srch_destination_id      -7.020308e-04 -0.0083008416 -0.154644299
## srch_destination_type_id -5.846939e-05 -0.0057686953 -0.223607916
## is_booking                2.107161e-03 -0.0478112413 -0.139871653
## cnt                      -1.784824e-03  0.0107201892  0.125171660
## hotel_continent           5.534905e-03 -0.0321231690  0.113013847
## hotel_country             6.359686e-03 -0.0307925900 -0.041625904
## hotel_market             -7.634260e-04 -0.0004929748 -0.033641299
## hotel_cluster             9.323106e-03  0.0042807518  0.042931770
##                                 channel srch_adults_cnt srch_children_cnt
## site_name                -2.330370e-02   -0.0134644926      -0.006769078
## posa_continent            6.583257e-02    0.0029052334       0.005545270
## user_location_country     6.243847e-02    0.0155867135       0.025402316
## user_location_region     -2.127135e-02   -0.0099028279      -0.011592178
## user_location_city        6.327445e-03   -0.0009521936       0.002521504
## user_id                   3.496742e-03   -0.0039614410       0.009569875
## is_mobile                -2.895157e-02    0.0240574415       0.025265565
## is_package               -1.895945e-02   -0.0230087994      -0.030530594
## channel                   1.000000e+00   -0.0435641592      -0.009375388
## srch_adults_cnt          -4.356416e-02    1.0000000000       0.130609990
## srch_children_cnt        -9.375388e-03    0.1306099903       1.000000000
## srch_rm_cnt               8.640886e-03    0.5266862507       0.088928065
## srch_destination_id       2.420879e-03    0.0016557753      -0.015881175
## srch_destination_type_id  2.548254e-02   -0.0277050562      -0.021496900
## is_booking                4.440275e-02   -0.1036238475      -0.057869976
## cnt                      -1.985759e-02    0.0337975611       0.034740754
## hotel_continent          -1.836889e-02    0.0020598936      -0.040389287
## hotel_country            -7.442732e-04    0.0020736117      -0.026445535
## hotel_market             -6.215601e-05    0.0047555343      -0.008310337
## hotel_cluster             2.086062e-03    0.0050439178       0.010627758
##                             srch_rm_cnt srch_destination_id
## site_name                 0.0189880383        0.0296212756
## posa_continent           -0.0272615953       -0.0138457320
## user_location_country    -0.0062025953       -0.0028905436
## user_location_region      0.0109871598        0.0154121213
## user_location_city        0.0035742093       -0.0035729519
## user_id                   0.0106919251       -0.0007020308
## is_mobile                -0.0272624865       -0.0083008416
## is_package               -0.0460217318       -0.1546442995
## channel                   0.0086408859        0.0024208795
## srch_adults_cnt           0.5266862507        0.0016557753
```

```
## srch_children_cnt          0.0889280651       -0.0158811752
## srch_rm_cnt                1.0000000000        0.0136756231
## srch_destination_id        0.0136756231        1.0000000000
## srch_destination_type_id   0.0071698886        0.4445360560
## is_booking                 0.0197868545        0.0435069228
## cnt                       -0.0062354454       -0.0277402556
## hotel_continent            0.0207650970        0.0199119971
## hotel_country              0.0100616125        0.0458149683
## hotel_market              -0.0007558751        0.0911849807
## hotel_cluster             -0.0099386323       -0.0129705728
##                          srch_destination_type_id   is_booking
## site_name                       2.867725e-03  0.196617452
## posa_continent                  2.928763e-02 -0.151940208
## user_location_country           1.947415e-02  0.040549107
## user_location_region            1.165986e-02  0.028221648
## user_location_city             -6.778584e-03 -0.009380646
## user_id                        -5.846939e-05  0.002107161
## is_mobile                      -5.768695e-03 -0.047811241
## is_package                     -2.236079e-01 -0.139871653
## channel                         2.548254e-02  0.044402754
## srch_adults_cnt                -2.770506e-02 -0.103623847
## srch_children_cnt              -2.149690e-02 -0.057869976
## srch_rm_cnt                     7.169889e-03  0.019786855
## srch_destination_id             4.445361e-01  0.043506923
## srch_destination_type_id        1.000000e+00  0.056311951
## is_booking                      5.631195e-02  1.000000000
## cnt                            -3.094777e-02 -0.269204473
## hotel_continent                -3.825705e-02  0.018492808
## hotel_country                  -1.547242e-02  0.049183171
## hotel_market                    3.123389e-02  0.001668716
## hotel_cluster                  -3.149616e-02 -0.040164351
##                                       cnt hotel_continent hotel_country
## site_name                  -0.042211677    0.2342365424  0.2429387498
## posa_continent              0.039569908   -0.3385834070 -0.1362598651
## user_location_country       0.002351495   -0.0504749397  0.1767670761
## user_location_region       -0.010349643    0.0631180736 -0.0164904575
## user_location_city          0.002350561    0.0008229693  0.0025461751
## user_id                    -0.001784824    0.0055349053  0.0063596862
## is_mobile                   0.010720189   -0.0321231690 -0.0307925900
## is_package                  0.125171660    0.1130138469 -0.0416259037
## channel                    -0.019857591   -0.0183688937 -0.0007442732
## srch_adults_cnt             0.033797561    0.0020598936  0.0020736117
## srch_children_cnt           0.034740754   -0.0403892875 -0.0264455348
## srch_rm_cnt                -0.006235445    0.0207650970  0.0100616125
## srch_destination_id        -0.027740256    0.0199119971  0.0458149683
## srch_destination_type_id   -0.030947773   -0.0382570463 -0.0154724161
## is_booking                 -0.269204473    0.0184928084  0.0491831709
## cnt                         1.000000000    0.0168369034 -0.0076017841
## hotel_continent             0.016836903    1.0000000000  0.3099013669
## hotel_country              -0.007601784    0.3099013669  1.0000000000
```

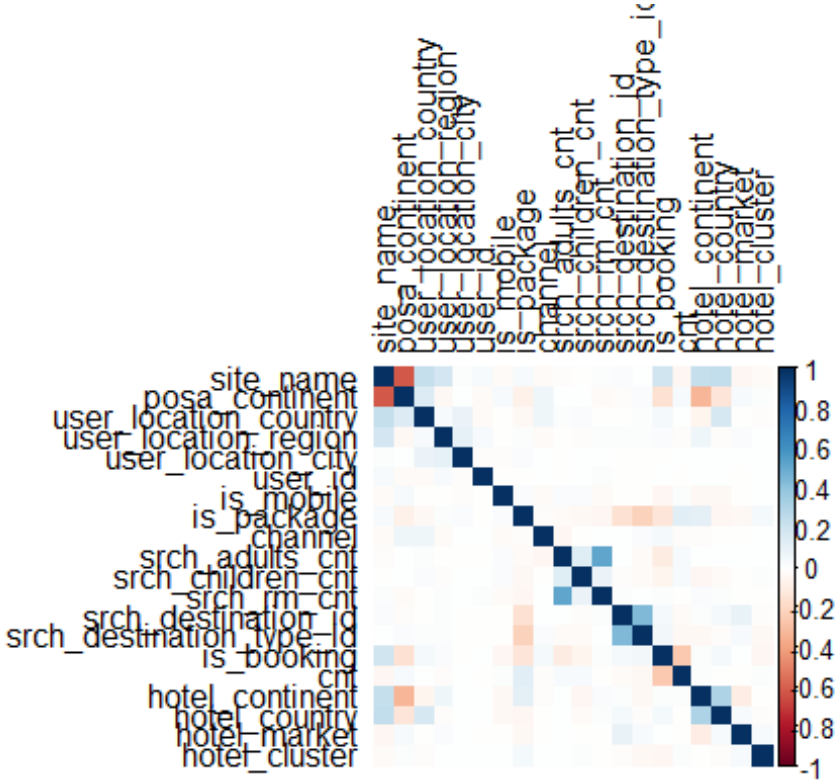
```
## hotel_market              -0.008592501   -0.0924215911  0.0329854480
## hotel_cluster              0.011807482   -0.0059817849 -0.0198183704
##                            hotel_market hotel_cluster
## site_name                 -4.799231e-02  -0.025330185
## posa_continent             3.362450e-02   0.014860660
## user_location_country     -1.790209e-03  -0.016721485
## user_location_region       2.944614e-02   0.002691211
## user_location_city         3.890907e-03   0.001422885
## user_id                   -7.634260e-04   0.009323106
## is_mobile                 -4.929748e-04   0.004280752
## is_package                -3.364130e-02   0.042931770
## channel                   -6.215601e-05   0.002086062
## srch_adults_cnt            4.755534e-03   0.005043918
## srch_children_cnt         -8.310337e-03   0.010627758
## srch_rm_cnt               -7.558751e-04  -0.009938632
## srch_destination_id        9.118498e-02  -0.012970573
## srch_destination_type_id   3.123389e-02  -0.031496162
## is_booking                 1.668716e-03  -0.040164351
## cnt                       -8.592501e-03   0.011807482
## hotel_continent           -9.242159e-02  -0.005981785
## hotel_country              3.298545e-02  -0.019818370
## hotel_market               1.000000e+00   0.030896246
## hotel_cluster              3.089625e-02   1.000000000
```

Exploratory Analysis

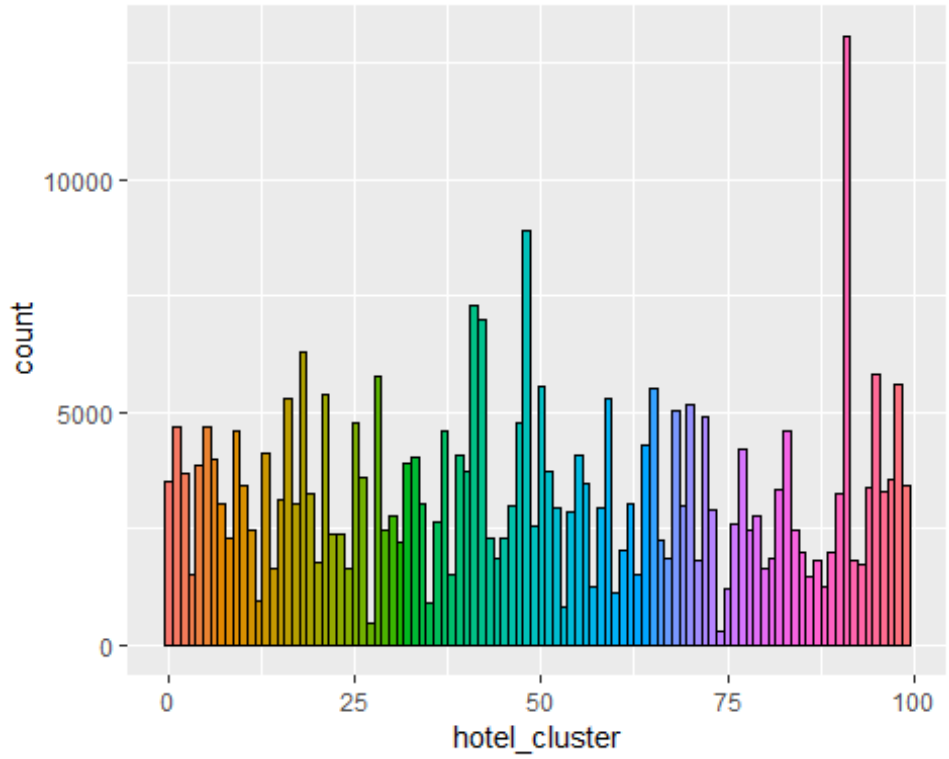
```
#finding missing values in the dataset
ggplot_missing(bookingfile)
```



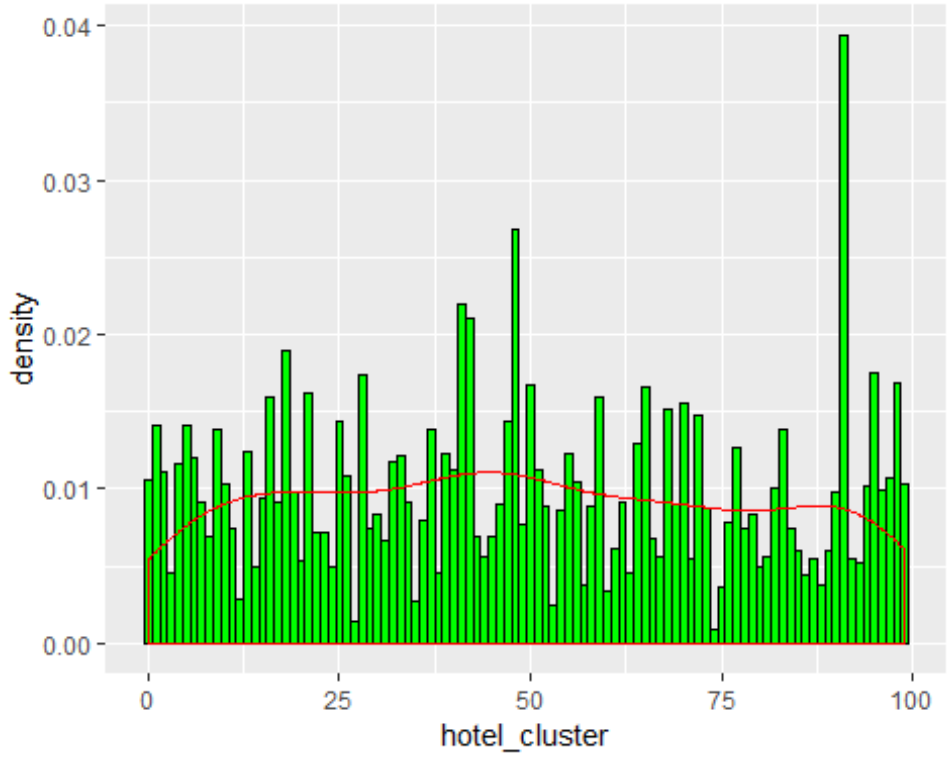

```
#correlation Matrix
corrplot(bookingstrain2, method="color",na.label = "square", na.label.col = "
orange",tl.col = "black")
```



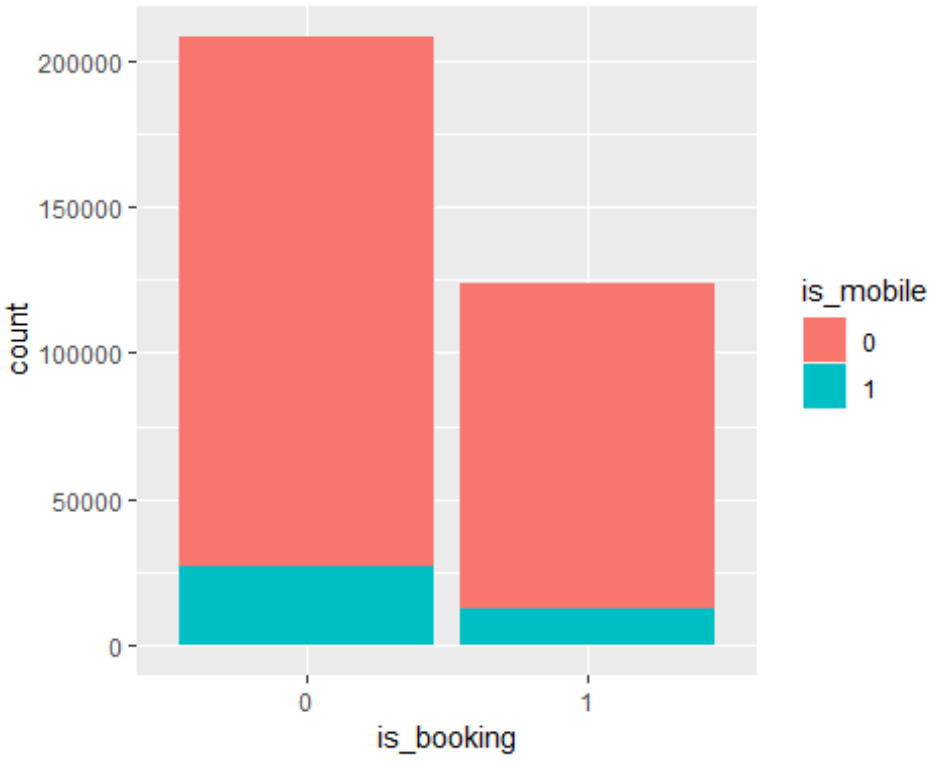
```
#hotel cluster histogram
bookingfile<-na.omit(bookingfile)
ggplot(data=bookingfile,aes(x=hotel_cluster,fill=cut(hotel_cluster,100)))+geo
m_histogram(color="black",bins=100 )+theme(legend.position = "none")
```



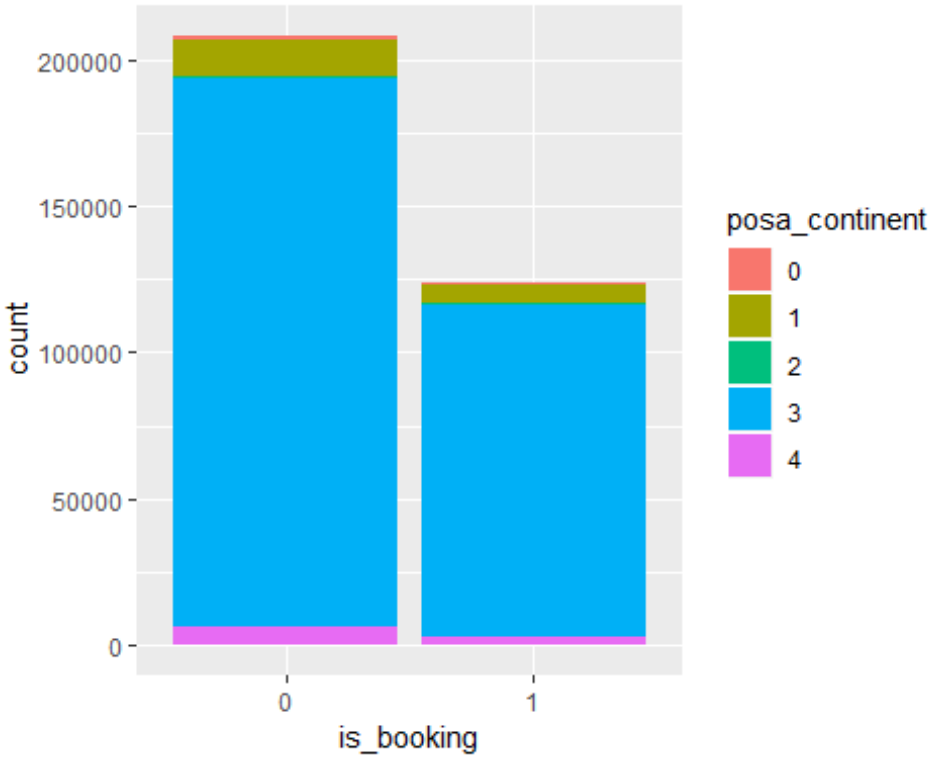

```
#cluster density plot
ggplot(data=bookingfile,aes(x=hotel_cluster),)+geom_histogram(aes(y=..density
..),bins=100 ,color="black",fill="green")+theme(legend.position = "none")+geo
m_density(adjust=4,color="red")
```



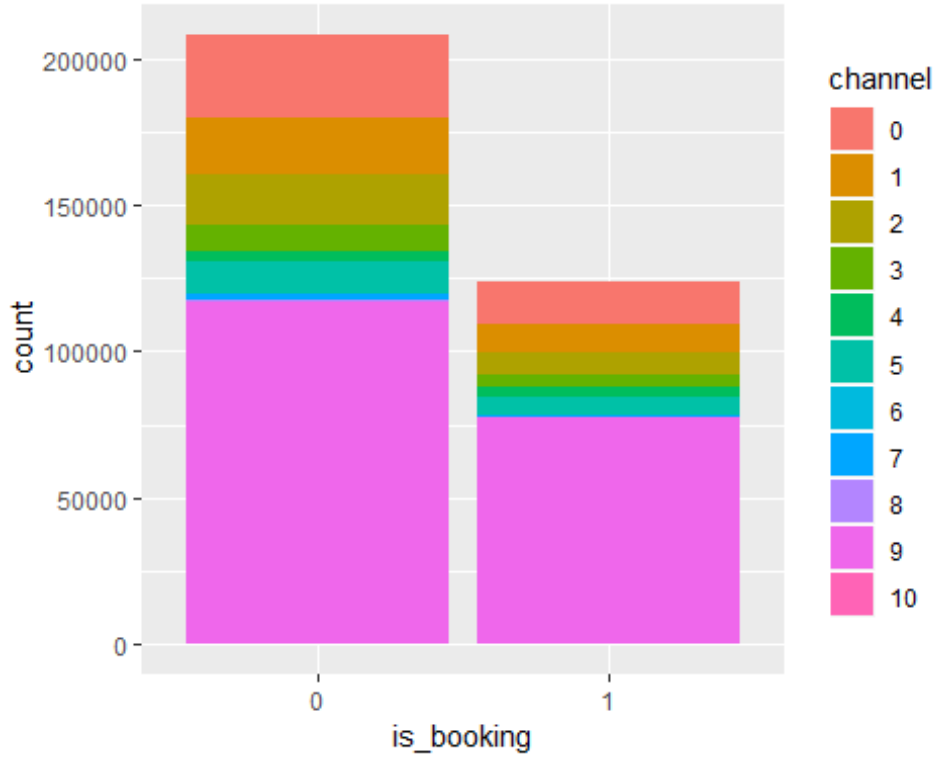

```
#bar graphs
bookingfile$is_booking<-as.factor(bookingfile$is_booking)
ggplot(data=bookingfile,aes(x=is_booking))+geom_bar(aes(fill=factor(is_mobile
)))+ guides(fill=guide_legend("is_mobile"))
```



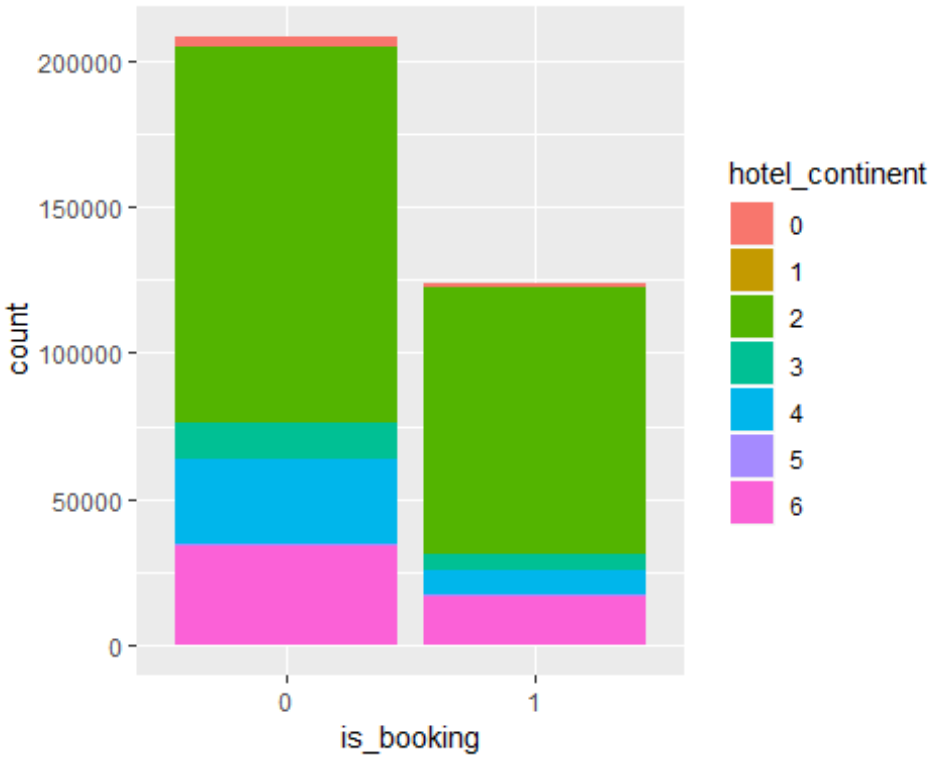

```
ggplot(data=bookingfile,aes(x=is_booking))+geom_bar(aes(fill=factor(posa_cont
inent)))+ guides(fill=guide_legend("posa_continent"))
```



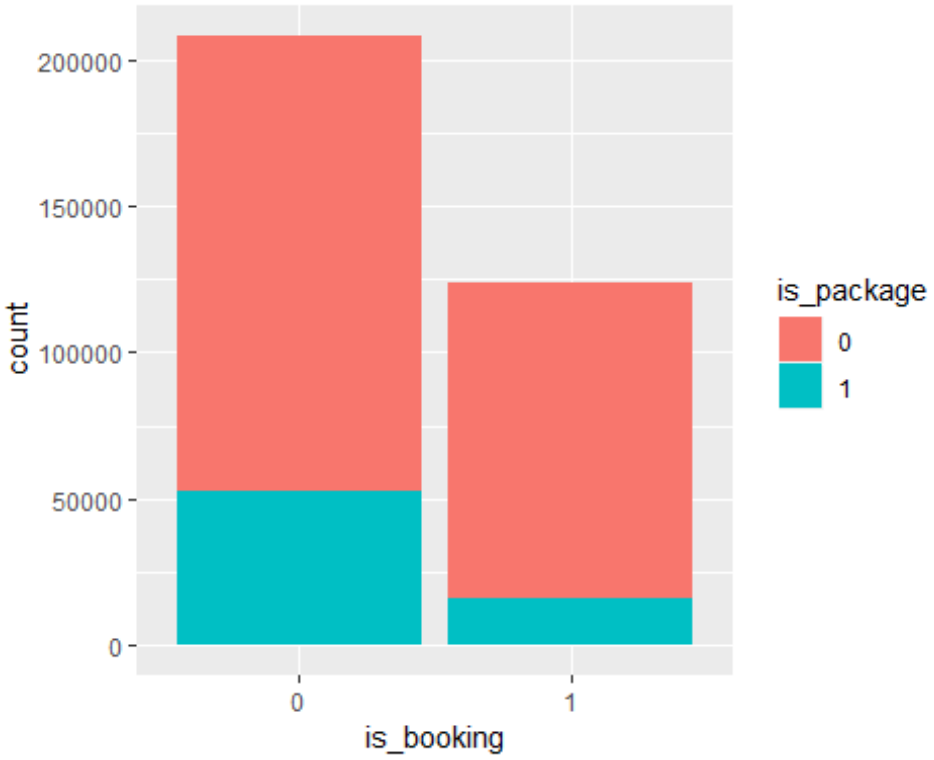
```
ggplot(data=bookingfile,aes(x=is_booking))+geom_bar(aes(fill=factor(channel))
)+ guides(fill=guide_legend("channel"))
```




```
ggplot(data=bookingfile,aes(x=is_booking))+geom_bar(aes(fill=factor(hotel_con
tinent)))+ guides(fill=guide_legend("hotel_continent"))
```




```
ggplot(data=bookingfile,aes(x=is_booking))+geom_bar(aes(fill=factor(is_packag
e)))+ guides(fill=guide_legend("is_package"))
```



```
## changing to factors
bookingfile<-na.omit(bookingfile)
bookingfile$site_name<-as.factor(bookingfile$site_name)
bookingfile$posa_continent<-as.factor(bookingfile$posa_continent)
bookingfile$user_location_city<-as.factor(bookingfile$user_location_city)
bookingfile$user_location_region<-as.factor(bookingfile$user_location_region)
bookingfile$is_mobile<-as.factor(bookingfile$is_mobile)
bookingfile$is_package<-as.factor(bookingfile$is_package)
bookingfile$channel<-as.factor(bookingfile$channel)
bookingfile$hotel_continent<-as.factor(bookingfile$hotel_continent)
bookingfile$hotel_country<-as.factor(bookingfile$hotel_country)
bookingfile$hotel_market<-as.factor(bookingfile$hotel_market)
bookingfile$hotel_cluster<-as.factor(bookingfile$hotel_cluster)
bookingfile$srch_destination_type_id<-as.factor(bookingfile$srch_destination_
type_id)

#removing columns
bookingfile1<-bookingfile[,c(-1,-12,-8,-13,-2,-4,-5,-6,-17,-21,-22,-23,-24)]
```

We observe that the variables do not show a very strong correlation between them.

Logistic Regression

```
# performing logisitic regression and summarizing the results
booking_mod<-glm(is_booking~., data=bookingfile1, family="binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(booking_mod)

##
## Call:
## glm(formula = is_booking ~ ., family = "binomial", data = bookingfile1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6181  -1.0400  -0.3111   1.1440   8.4904
##
## Coefficients:
##                              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                 3.153e+00  5.912e-02   53.332  < 2e-16 ***
## posa_continent1            -4.165e-01  5.332e-02   -7.812 5.65e-15 ***
## posa_continent2            -6.285e-02  8.209e-02   -0.766  0.44386
## posa_continent3            -2.585e-01  5.098e-02   -5.071 3.95e-07 ***
## posa_continent4            -3.299e-01  5.602e-02   -5.889 3.88e-09 ***
## orig_destination_distance -6.265e-05  1.841e-06  -34.036  < 2e-16 ***
## is_mobile1                 -2.443e-01  1.254e-02  -19.479  < 2e-16 ***
## is_package1                -6.770e-01  1.089e-02  -62.173  < 2e-16 ***
## channel1                   -7.648e-02  1.729e-02   -4.423 9.72e-06 ***
## channel2                   -2.507e-01  1.844e-02  -13.597  < 2e-16 ***
## channel3                   -1.808e-01  2.319e-02   -7.794 6.51e-15 ***
```

```
## channel4                   7.017e-01  2.903e-02   24.174  < 2e-16 ***
## channel5                   1.395e-01  2.143e-02    6.513 7.39e-11 ***
## channel6                  -7.993e-01  1.472e-01   -5.431 5.59e-08 ***
## channel7                  -5.360e-01  5.318e-02  -10.080  < 2e-16 ***
## channel8                  -4.508e-01  6.545e-02   -6.887 5.69e-12 ***
## channel9                   1.688e-01  1.208e-02   13.966  < 2e-16 ***
## channel10                  8.272e-01  3.721e-01    2.223  0.02621 *
## srch_adults_cnt           -3.081e-01  5.431e-03  -56.730  < 2e-16 ***
## srch_children_cnt         -1.194e-01  5.647e-03  -21.152  < 2e-16 ***
## srch_rm_cnt                3.859e-01  1.104e-02   34.935  < 2e-16 ***
## srch_destination_type_id3  1.841e-01  1.485e-02   12.394  < 2e-16 ***
## srch_destination_type_id4  1.760e-02  2.054e-02    0.857  0.39152
## srch_destination_type_id5  3.231e-01  1.655e-02   19.518  < 2e-16 ***
## srch_destination_type_id6  8.142e-02  9.508e-03    8.563  < 2e-16 ***
## srch_destination_type_id7  7.544e-01  5.733e-01    1.316  0.18825
## srch_destination_type_id8 -1.953e-01  6.108e-02   -3.197  0.00139 **
## cnt                       -2.771e+00  2.460e-02 -112.639  < 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: 438663  on 332105  degrees of freedom
## Residual deviance: 375915  on 332078  degrees of freedom
## AIC: 375971
##
## Number of Fisher Scoring iterations: 8
```

From the summary, we observe that hotel_market and hotel_country are not significant.

(c).Confusion matrix for current model

```
# Prediction
bookingfile1$prob<-predict(booking_mod,type="response")
View(bookingfile1)

bookingfile1$predbook="0"
bookingfile1$predbook[bookingfile1$prob>0.4]="1"
table(bookingfile1$predbook, bookingfile1$is_booking)

##
##          0      1
##   0 115896  29427
##   1  92403  94380

# % correct
mean(bookingfile1$predbook==bookingfile1$is_booking)*100

## [1] 63.31593
```

```
#% correct when customer books and model also predicts cusomer will book
((94380)/(29427+94380))*100
```

```
## [1] 76.23155
```

```
#% correct when customer does not book and model also predicts cusomer will n
ot book
((115896/(92403+115896)))*100
```

```
## [1] 55.63925
```

Creating test and train data and building model

```
set.seed(2)
train.X<-sample(c(1:dim(bookingfile1)[1]), dim(bookingfile1)[1]*0.7)
train<-bookingfile1[train.X,]
test<-bookingfile1[-train.X,]

#building model
testglm = glm(is_booking~.-prob-predbook, data= train, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(testglm)
```

```
##
## Call:
## glm(formula = is_booking ~ . - prob - predbook, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6300  -1.0375  -0.3126   1.1440   5.9081
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 3.180e+00  7.106e-02  44.750  < 2e-16 ***
## posa_continent1            -4.688e-01  6.406e-02  -7.318 2.52e-13 ***
## posa_continent2            -1.765e-02  9.876e-02  -0.179 0.858175
## posa_continent3            -2.988e-01  6.130e-02  -4.874 1.09e-06 ***
## posa_continent4            -3.967e-01  6.741e-02  -5.884 4.00e-09 ***
## orig_destination_distance -6.363e-05  2.205e-06 -28.859  < 2e-16 ***
## is_mobile1                 -2.538e-01  1.504e-02 -16.873  < 2e-16 ***
## is_package1                -6.800e-01  1.303e-02 -52.174  < 2e-16 ***
## channel1                   -7.705e-02  2.073e-02  -3.716 0.000202 ***
## channel2                   -2.354e-01  2.205e-02 -10.672  < 2e-16 ***
## channel3                   -1.544e-01  2.772e-02  -5.568 2.57e-08 ***
## channel4                    6.901e-01  3.473e-02  19.868  < 2e-16 ***
## channel5                    1.528e-01  2.565e-02   5.957 2.57e-09 ***
## channel6                   -9.019e-01  1.801e-01  -5.007 5.54e-07 ***
## channel7                   -5.215e-01  6.339e-02  -8.227  < 2e-16 ***
## channel8                   -4.011e-01  7.792e-02  -5.148 2.63e-07 ***
```

```
## channel9                    1.757e-01  1.446e-02  12.147  < 2e-16 ***
## channel10                   9.177e-01  4.621e-01   1.986 0.047050 *
## srch_adults_cnt            -3.101e-01  6.503e-03 -47.682  < 2e-16 ***
## srch_children_cnt          -1.123e-01  6.750e-03 -16.640  < 2e-16 ***
## srch_rm_cnt                 3.914e-01  1.322e-02  29.606  < 2e-16 ***
## srch_destination_type_id3   1.775e-01  1.780e-02   9.976  < 2e-16 ***
## srch_destination_type_id4  -3.894e-03  2.462e-02  -0.158 0.874332
## srch_destination_type_id5   3.219e-01  1.978e-02  16.278  < 2e-16 ***
## srch_destination_type_id6   8.713e-02  1.137e-02   7.665 1.79e-14 ***
## srch_destination_type_id7   9.184e-01  7.100e-01   1.293 0.195853
## srch_destination_type_id8  -1.957e-01  7.251e-02  -2.698 0.006970 **
## cnt                        -2.773e+00  2.948e-02 -94.073  < 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: 306634  on 232473  degrees of freedom
## Residual deviance: 262760  on 232446  degrees of freedom
## AIC: 262816
##
## Number of Fisher Scoring iterations: 7
```

```
test$prob<-predict(testglm,type = "response",newdata=test)
test$predbook="0"

test$predbook[test$prob>0.4]="1"
table(test$predbook, test$is_booking)
```

```
##
##         0     1
##   0 34878  9213
##   1 27202 28339
```

```
# % correct
mean(test$predbook==test$is_booking)*100
```

```
## [1] 63.4505
```

```
#% correct when customer books and model also predicts cusomer will book
((28339)/(28339+9213))*100
```

```
## [1] 75.46602
```

```
##% correct when customer does not book and model also predicts cusomer will
not book
(( 34878/(27202+34878)))*100
```

```
## [1] 56.18235
```

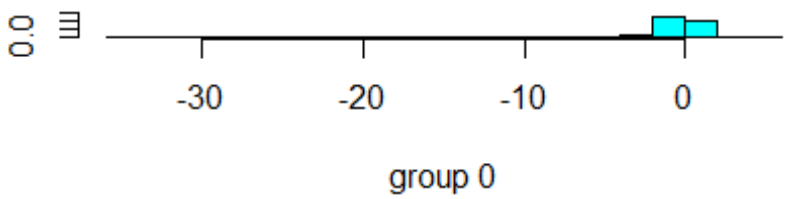
LDA

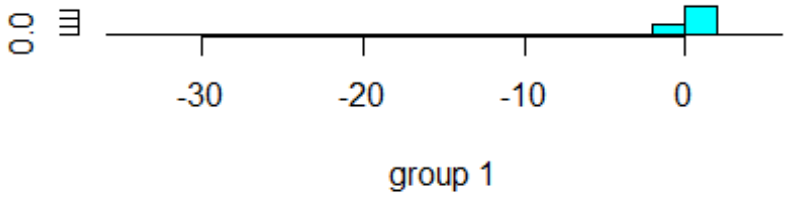
```
#building model
lda_mod = lda(is_booking~.-prob-predbook, data= train)
lda_mod

## Call:
## lda(is_booking ~ . - prob - predbook, data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.6289693 0.3710307
##
## Group means:
##   posa_continent1 posa_continent2 posa_continent3 posa_continent4
## 0      0.06085392     0.003043380       0.9006832      0.02990035
## 1      0.04934207     0.004034549       0.9167005      0.02376674
##   orig_destination_distance is_mobile1 is_package1   channel1   channel2
## 0                  2046.575 0.12824599   0.2513695 0.09451576 0.08136425
## 1                  1708.013 0.09733928   0.1270767 0.07712017 0.05963712
##    channel3   channel4   channel5     channel6    channel7    channel8
## 0 0.0453703 0.01681040 0.05132712 0.0012105130 0.008152155 0.004855730
## 1 0.0318938 0.02784766 0.05053620 0.0004753348 0.004382355 0.002967944
##    channel9    channel10 srch_adults_cnt srch_children_cnt srch_rm_cnt
## 0 0.5602008 7.522962e-05        2.052093         0.3541537    1.112667
## 1 0.6256913 1.391224e-04        1.872483         0.2773173    1.127564
##   srch_destination_type_id3 srch_destination_type_id4
## 0                0.07018240                0.03769688
## 1                0.08524723                0.03832821
##   srch_destination_type_id5 srch_destination_type_id6
## 0                0.04832477                 0.2323364
## 1                0.07673758                 0.2626514
##   srch_destination_type_id7 srch_destination_type_id8      cnt
## 0              2.051717e-05               0.004534294 1.511178
## 1              6.956118e-05               0.003802678 1.013877
##
## Coefficients of linear discriminants:
##                                    LD1
## posa_continent1           -5.812640e-01
## posa_continent2            5.649140e-02
## posa_continent3           -3.735673e-01
## posa_continent4           -4.719815e-01
## orig_destination_distance -9.585694e-05
## is_mobile1                -3.727572e-01
## is_package1               -9.061274e-01
## channel1                  -1.068415e-01
## channel2                  -3.230614e-01
## channel3                  -2.359116e-01
## channel4                   9.774220e-01
## channel5                   1.663218e-01
## channel6                  -1.157924e+00
```

```
## channel7                   -6.557129e-01
## channel8                   -5.542226e-01
## channel9                    2.715647e-01
## channel10                   1.717935e+00
## srch_adults_cnt            -4.344161e-01
## srch_children_cnt          -1.852383e-01
## srch_rm_cnt                 5.229937e-01
## srch_destination_type_id3   2.676853e-01
## srch_destination_type_id4  -3.377979e-02
## srch_destination_type_id5   5.124938e-01
## srch_destination_type_id6   1.225858e-01
## srch_destination_type_id7   1.746678e+00
## srch_destination_type_id8  -2.807033e-01
## cnt                        -7.230656e-01
```

```
plot(lda_mod)
```






```
#prediction
lda.pred<-predict(lda_mod,type="response",newdata = test)
lda.class <- lda.pred$class
table(lda.class, test$is_booking)
```

```
##
## lda.class     0     1
##         0 56621 28442
##         1  5459  9110
```

```
#% correct
mean(lda.class==test$is_booking)*100

## [1] 65.97378

#% correct when customer books and model also predicts cusomer will book
((9110)/(9110+28442))*100

## [1] 24.25969

##% correct when customer does not book and model also predicts cusomer will
not book
((56621/(56621+5459)))*100

## [1] 91.20651
```

Naive Bayes Classifier

```
r<-naive_bayes(is_booking~.-prob-predbook,data=train)
r

## ====================== Naive Bayes ======================
## Call:
## naive_bayes.formula(formula = is_booking ~ . - prob - predbook,
##     data = train)
##
## A priori probabilities:
##
##         0         1
## 0.6289693 0.3710307
##
## Tables:
##
## posa_continent           0           1
##              0 0.005519119 0.006156165
##              1 0.060853925 0.049342067
##              2 0.003043380 0.004034549
##              3 0.900683222 0.916700481
##              4 0.029900355 0.023766738
##
##
## orig_destination_distance        0        1
##                       mean 2046.575 1708.013
##                       sd   2283.804 2171.250
##
##
## is_mobile          0          1
##         0 0.87175401 0.90266072
##         1 0.12824599 0.09733928
##
##
## is_package         0         1
```

```
##          0 0.7486305 0.8729233
##          1 0.2513695 0.1270767
##
##
## channel            0            1
##      0  1.361177e-01 1.193090e-01
##      1  9.451576e-02 7.712017e-02
##      2  8.136425e-02 5.963712e-02
##      3  4.537030e-02 3.189380e-02
##      4  1.681040e-02 2.784766e-02
##      5  5.132712e-02 5.053620e-02
##      6  1.210513e-03 4.753348e-04
##      7  8.152155e-03 4.382355e-03
##      8  4.855730e-03 2.967944e-03
##      9  5.602008e-01 6.256913e-01
##      10 7.522962e-05 1.391224e-04
##
## # ... and 7 more tables
```

```
predm<-predict(r,test)
table(predm,test$is_booking)
```

```
##
## predm     0     1
##     0 25543  3944
##     1 36537 33608
```

```
# % correct
mean(predm==test$is_booking)*100
```

```
## [1] 59.36948
```

```
#% correct when customer books and model also predicts cusomer will book
((33608)/(3944+33608))*100
```

```
## [1] 89.49723
```

```
##% correct when customer does not book and model also predicts cusomer will
not book
((25543 /(25543+36537)))*100
```

```
## [1] 41.1453
```

References

1. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani(2009). An Introduction to Statistical Learning. Springer

2. Expedia Hotel Recommendations(2016) retrieved from https://www.kaggle.com/c/expedia-hotel-recommendations