Cairo University - Faculty of Engineering
Computer Engineering Department
CMP461 Big Data

BDA Project

Final Delivery
Hotel Booking Demand

Submitted to:

Eng. Omar Samir

By:

Andrew Tawdros Hanna	Sec: 1	BN: 14
Mark Medhat Abdou	Sec: 2	BN: 12

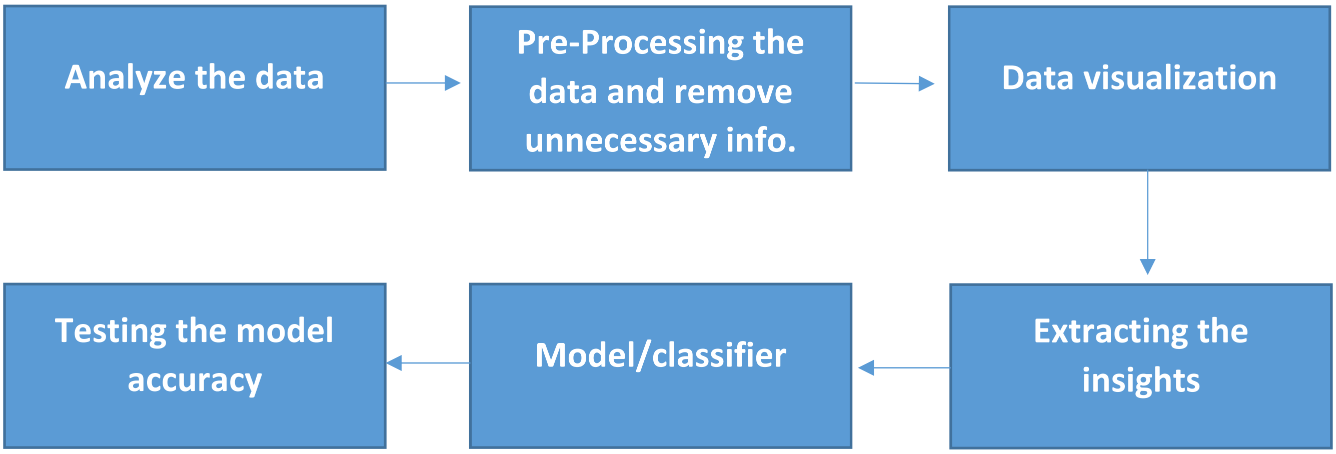
Academic Year

2021-2022

1) Brief problem description

The cancellation rate for booking hotels online is high that creates discomfort for many hotels and create a desire to take precautions. Therefore, predicting reservations that can be cancelled will create a surplus value for hotels and hotels can take action to prevent these cancellations.

2) Project pipeline



3) Analysis and solution of the problem:

a. Data pre-processing

- Get the structure of the data

```
> str(HotelBooking)
'data.frame':	119390 obs. of  32 variables:
 $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel" ...
 $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
 $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
 $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ arrival_date_month            : chr  "July" "July" "July" "July" ...
 $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
 $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ stays_in_week_nights          : 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 ...
 $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ meal                          : chr  "BB" "BB" "BB" "BB" ...
 $ country                       : chr  "PRT" "PRT" "GBR" "GBR" ...
 $ market_segment                : chr  "Direct" "Direct" "Direct" "Corporate" ...
 $ distribution_channel          : chr  "Direct" "Direct" "Direct" "Corporate" ...
 $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ reserved_room_type            : chr  "C" "C" "A" "A" ...
 $ assigned_room_type            : chr  "C" "C" "C" "A" ...
 $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
 $ deposit_type                  : chr  "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...
 $ agent                         : chr  "NULL" "NULL" "NULL" "304" ...
 $ company                       : chr  "NULL" "NULL" "NULL" "NULL" ...
 $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ customer_type                 : chr  "Transient" "Transient" "Transient" "Transient" ...
 $ adr                           : num  0 0 75 75 98 ...
 $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
 $ reservation_status            : chr  "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...
 $ reservation_status_date       : chr  "2015-07-01" "2015-07-01" "2015-07-02" "2015-07-02" ...
```

- Adjust the data types

```
# Adjust data type (Int -> Factor)
HotelBooking$is_repeated_guest <- as.factor(ifelse(HotelBooking$is_repeated_guest==1, "Yes", "No"))
HotelBooking$arrival_date_year <- as.factor(HotelBooking$arrival_date_year)
HotelBooking$arrival_date_week_number <- as.factor(HotelBooking$arrival_date_week_number)
HotelBooking$arrival_date_day_of_month <- as.factor(HotelBooking$arrival_date_day_of_month)
HotelBooking$is_canceled=as.factor(HotelBooking$is_canceled)
```

- Deal with missing values

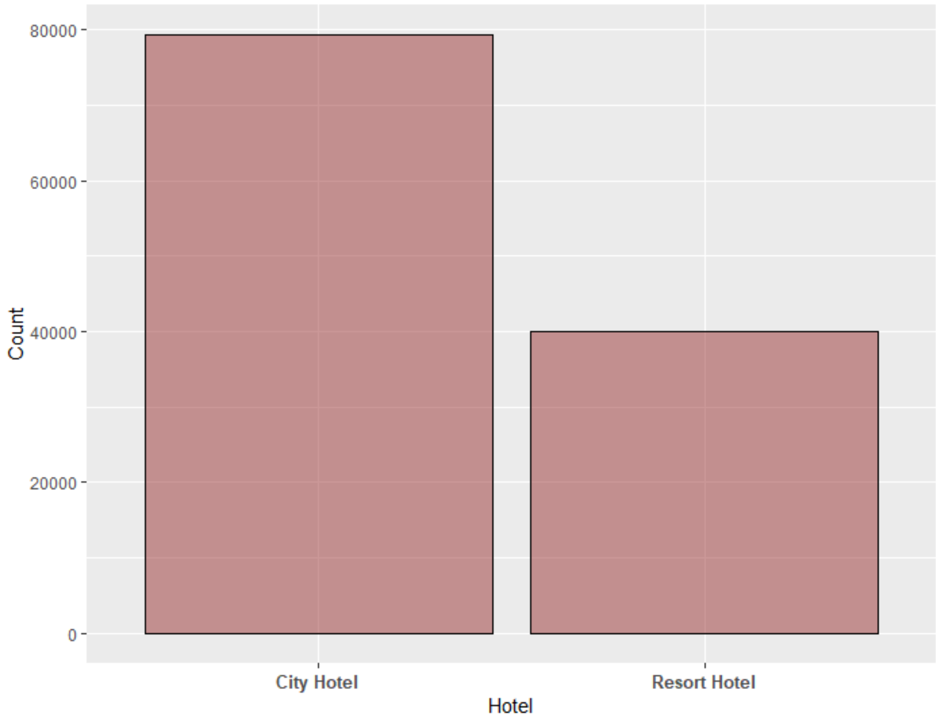
```
> colSums(is.na(HotelBooking)) #( Column Children has 4 missing value)
                         hotel                    is_canceled                      lead_time 
                             0                              0                              0 
             arrival_date_year             arrival_date_month       arrival_date_week_number 
                             0                              0                              0 
     arrival_date_day_of_month        stays_in_weekend_nights           stays_in_week_nights 
                             0                              0                              0 
                        adults                       children                         babies 
                             0                              4                              0 
                          meal                        country                 market_segment 
                             0                              0                              0 
          distribution_channel              is_repeated_guest         previous_cancellations 
                             0                              0                              0 
previous_bookings_not_canceled             reserved_room_type             assigned_room_type 
                             0                              0                              0 
               booking_changes                   deposit_type                          agent 
                             0                              0                              0 
                       company           days_in_waiting_list                  customer_type 
                             0                              0                              0 
                           adr    required_car_parking_spaces      total_of_special_requests 
                             0                              0                              0 
            reservation_status        reservation_status_date 
                             0                              0 
```

- Drop some unnecessary columns

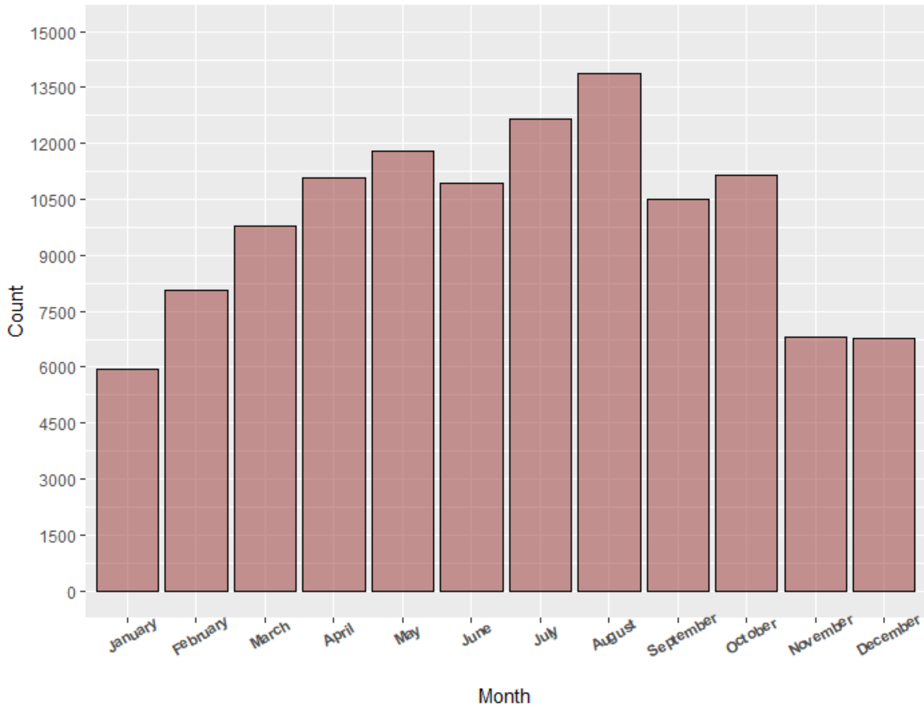
```
#Remove unwanted columns
drops <- c("company","country","adr_pp","company","reservation_status_date","agent","reservation_status")
HotelBooking<-HotelBooking[ , !(names(HotelBooking) %in% drops)]
```

b. Data visualization

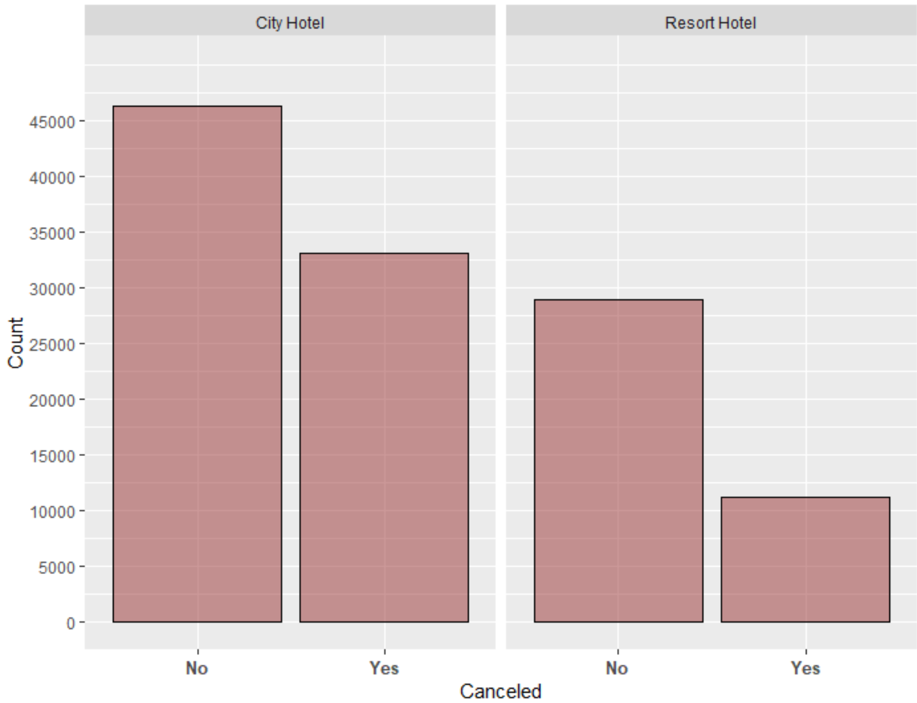
- City hotels vs Resort hotels



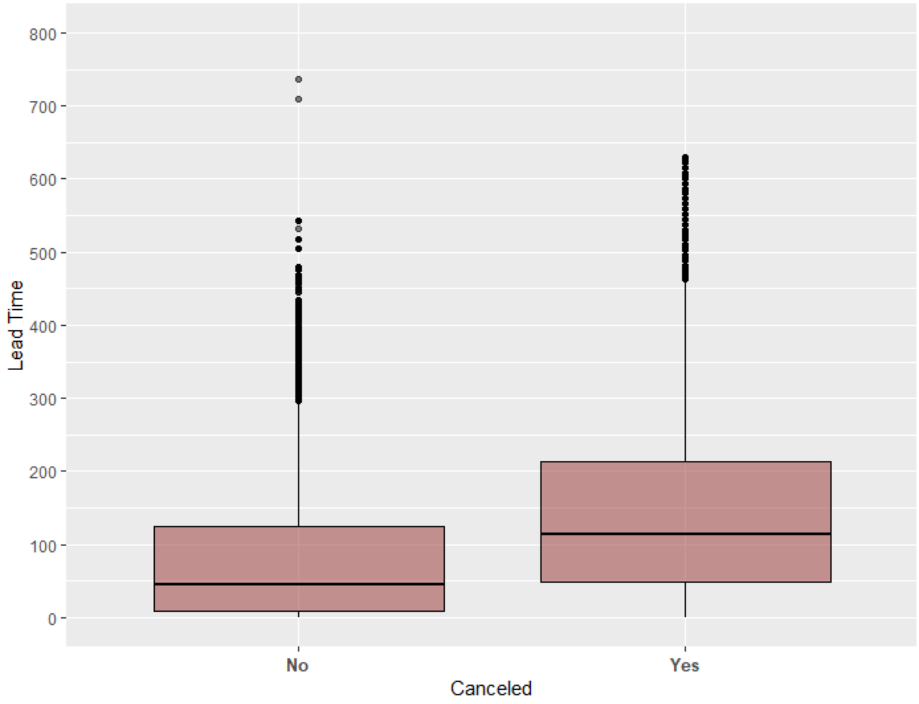
- Number of arrival Date by Month



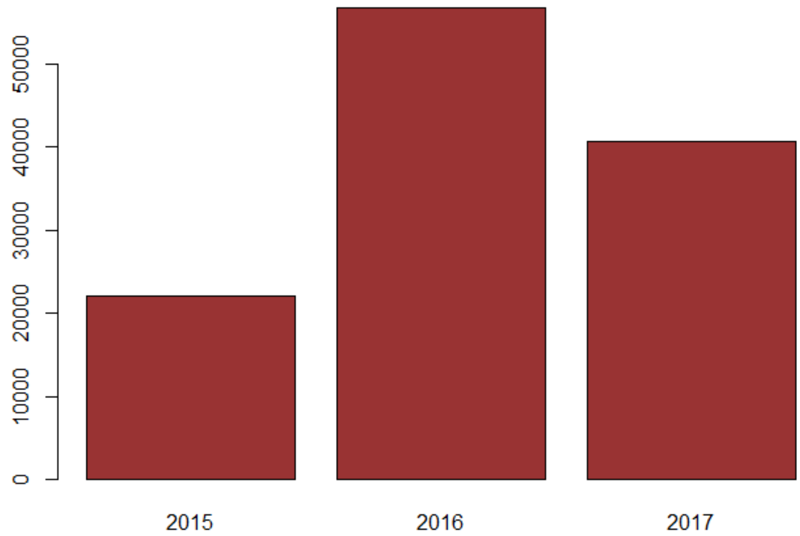
- Number of city hotel and Resort Hotel cancelled or not cancelled



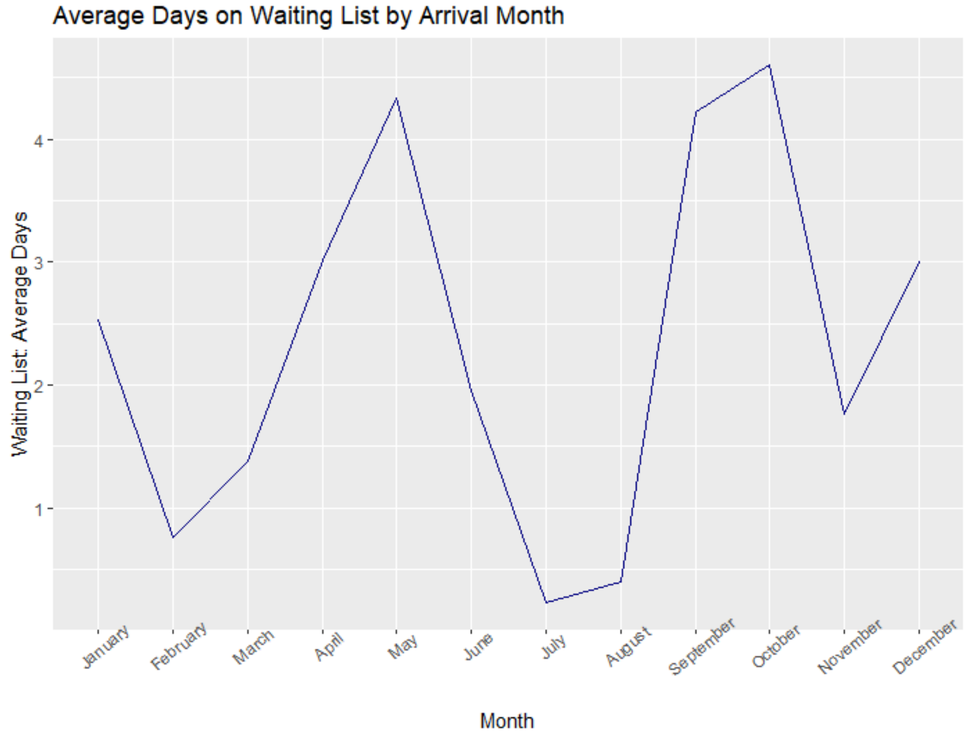
- Relation between Cancelled booking and Lead time



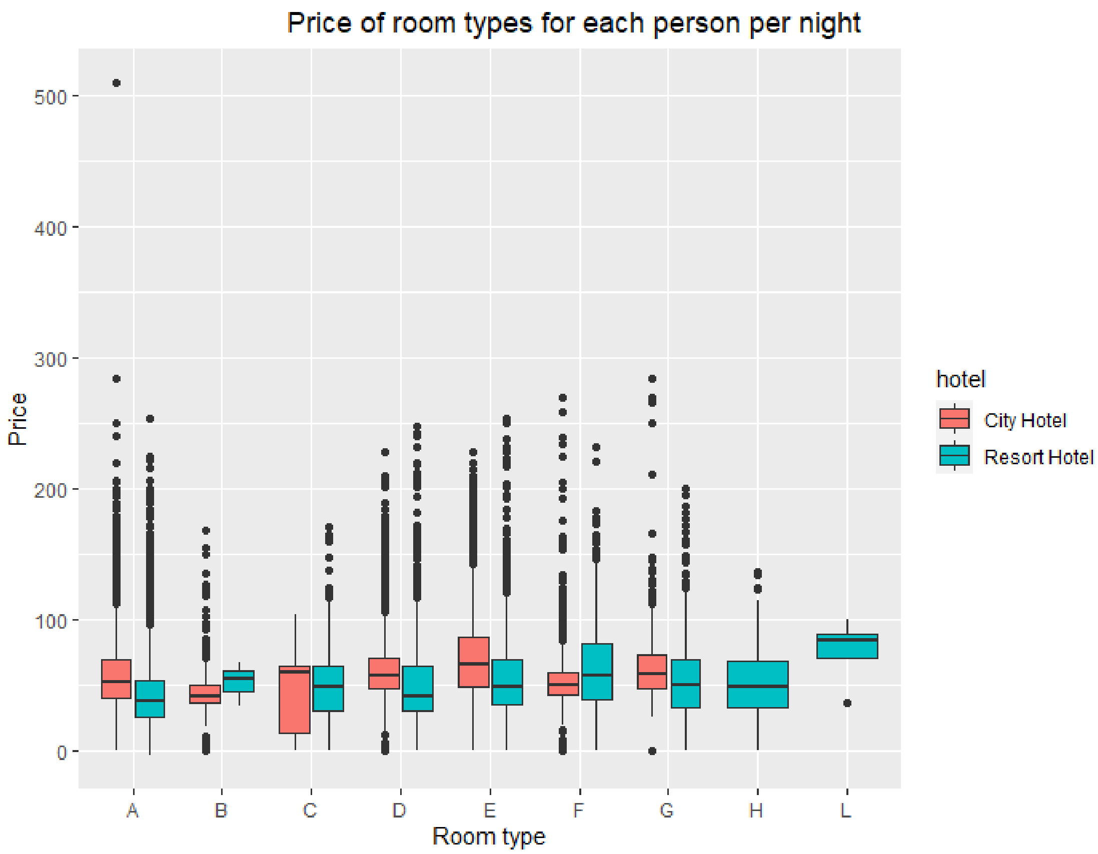
- No. of arrivals per year



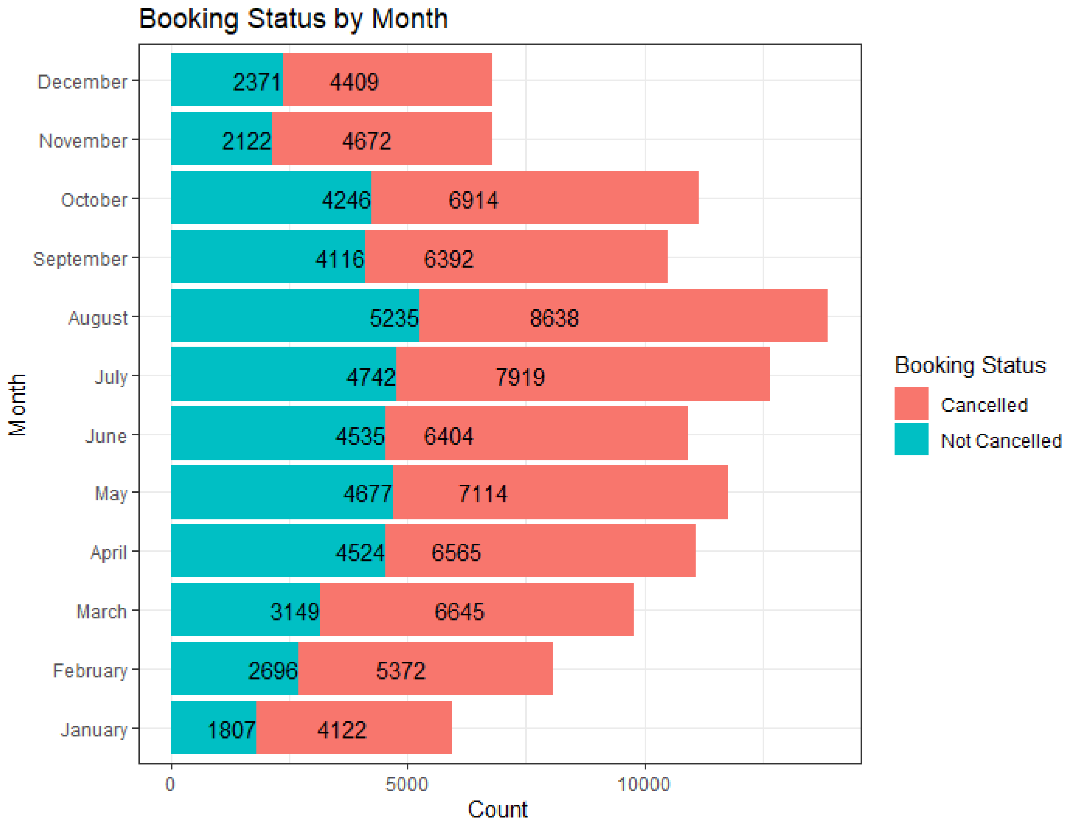
- May and October have the highest waiting times; these months represent the times right before and after peak reservation months (respectively)



- Price of room types (which was cancelled) for each person per night



- Cancellation per months



c. Extracting insights from data

- The most arrival months are August and July

```
   ArrivalDateMonth     N
1           January  5929
2          February  8068
3             March  9794
4             April 11089
5               May 11791
6              June 10939
7              July 12661
8            August 13873
9         September 10508
10          October 11160
11         November  6794
12         December  6780
```

- Online market segment's cancellation is more

```
  `HotelBooking$market_segment` `length(is_canceled)`
  <chr>                                         <int>
1 Aviation                                        237
2 Complementary                                   743
3 Corporate                                      5295
4 Direct                                        12605
5 Groups                                        19811
6 Offline TA/TO                                 24219
7 Online TA                                     56476
```

- City Hotel Cancellation is more

```
# A tibble: 2 x 2
  `HotelBooking$hotel` `length(is_canceled)`
  <chr>                                <int>
1 City Hotel                           79326
2 Resort Hotel                         40060
```

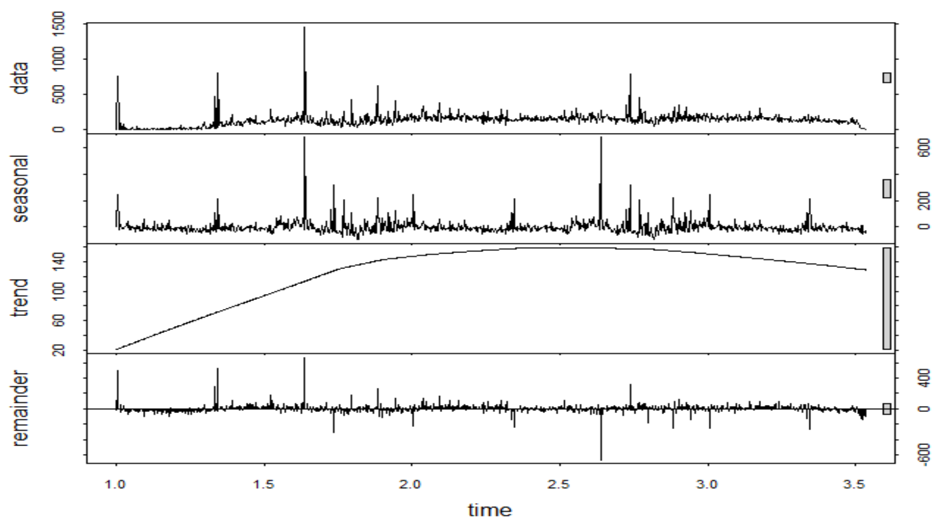
- Couples booking cancellation is more

```
# A tibble: 14 x 2
   `HotelBooking$adults` `length(is_canceled)`
                   <int>                 <int>
 1                     0                   403
 2                     1                 23027
 3                     2                 89677
 4                     3                  6201
 5                     4                    62
 6                     5                     2
 7                     6                     1
 8                    10                     1
 9                    20                     2
10                    26                     5
11                    27                     2
12                    40                     1
13                    50                     1
14                    55                     1
```

- 'A' type room cancellation is higher

```
# A tibble: 10 x 2
   `HotelBooking$reserved_room_type` `length(is_canceled)`
   <chr>                                             <int>
 1 A                                                 85994
 2 B                                                  1114
 3 C                                                   932
 4 D                                                 19201
 5 E                                                  6535
 6 F                                                  2897
 7 G                                                  2094
 8 H                                                   601
 9 L                                                     6
10 P                                                    12
```

- Time series as hotel is strongly dependant on seasonality



d. Model/Classifier training

- We have used two classifiers which are:
    a. Logistic regression
    b. Random Forest
    Finally, Random Forest was considered

4) Results and Evaluation.

- Accuracy of Random Forest = 84.2%

5) Unsuccessful trials that were not included in the final solution.

- Logistic regression classifier with accuracy less than Random Forest = 83.3%

6) Any Enhancements and future work

Enhancements as:

- This dataset has 32 variables. So, using PCA to reduce dimension is a good next step.

- Tuning the parameters of Random Forest Classifier to get more efficient accuracy.
- Analyze logically more correlated variables to get better efficiency.

Future work as:

- Q1. when is the best time of year to book a hotel room?
- Q2. What if you want to predict whether or not a hotel is likely to receive a huge number of special requests?