Hotel Booking Cancellation

Manal AlSaleh

Agenda

- Problem Statement
- Data Description
- Exploratory of Data Analysis (EDA)
- Data Cleaning
- Modeling

Problem Statement

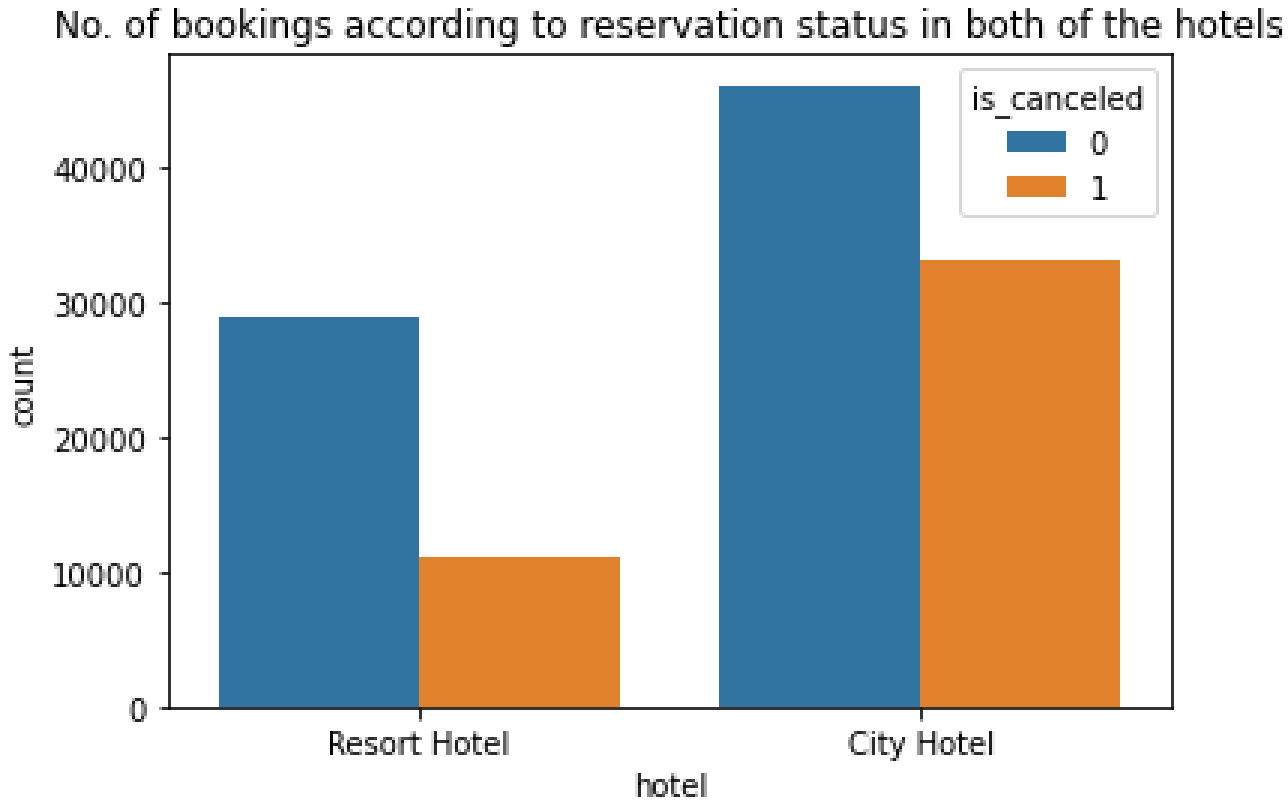
- Hotel Booking Cancellation Impact:
  - High Distribution Cost
  - High Opportunity Cost
  - Lost of revenue

Data Description

- Data for 119,390 bookings
- 32 variables (12 Categorical columns and 20 Numerical columns)
- Period 2015 - 2017

Exploratory of Data Analysis

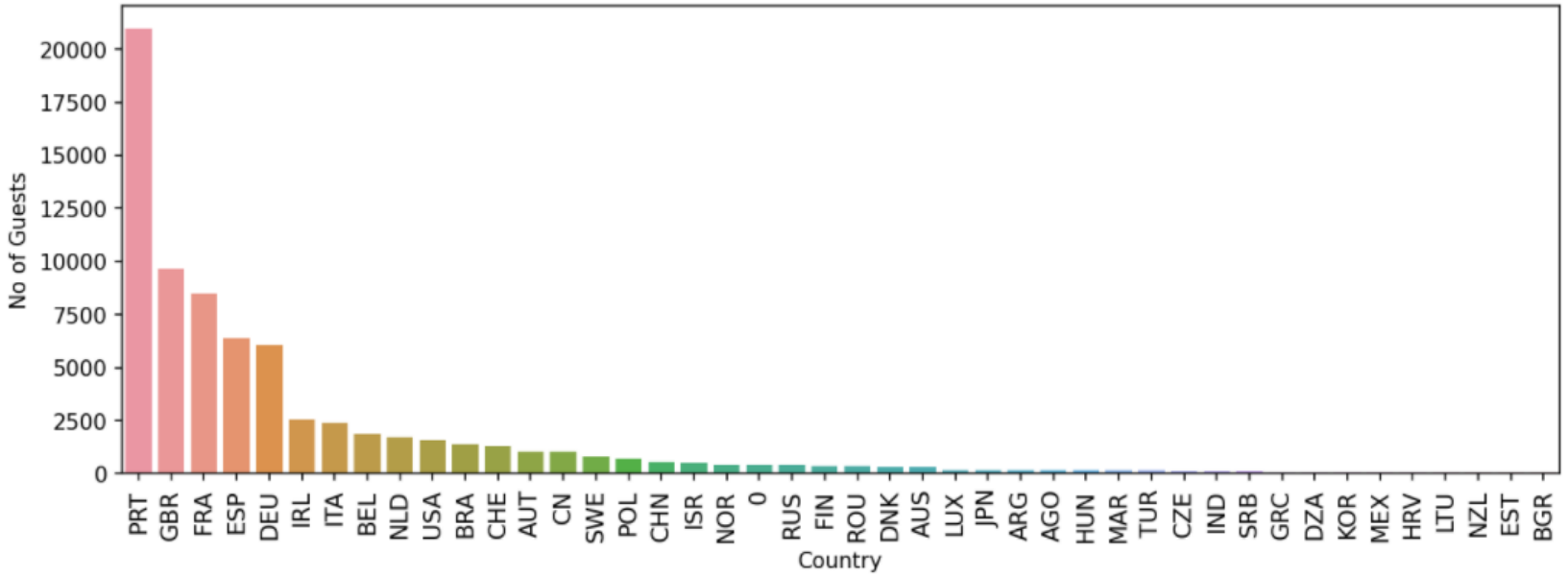
- Is Canceled : Balanced Dataset



Exploratory of Data Analysis

- From where the most guests are coming?

Majority of booking are from PRT (Portugal)



Exploratory of Data Analysis

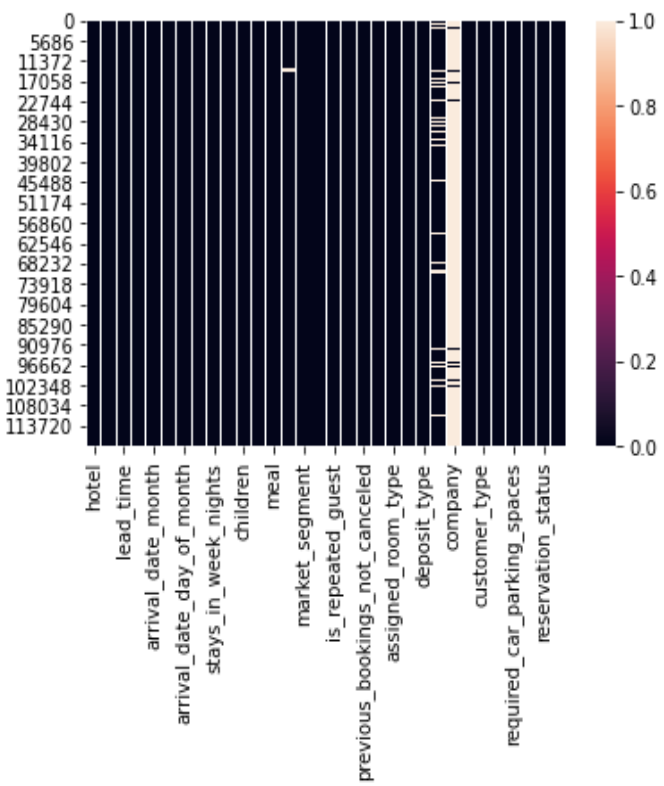
- Which are the busiest month?

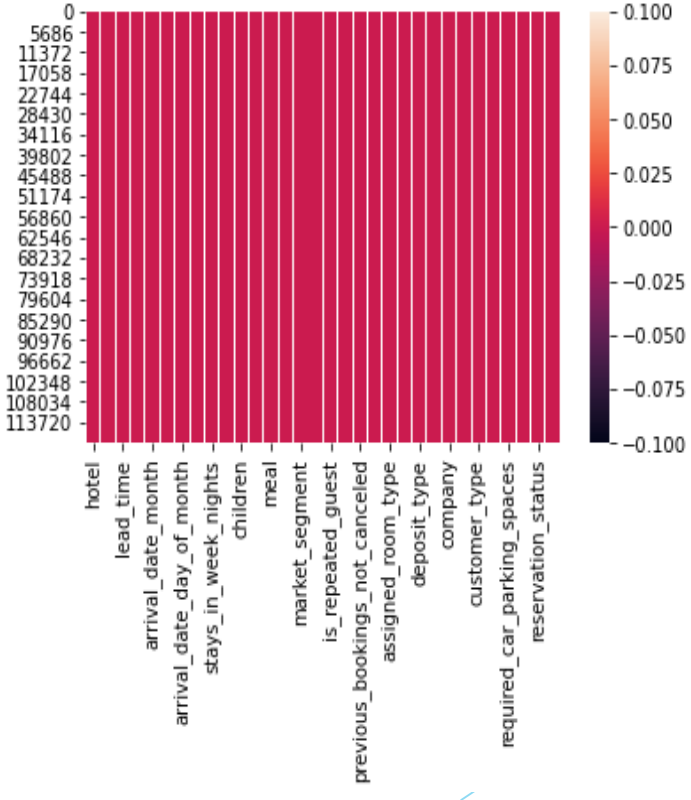
The more bookings in the summer.

	Month	No of Guests in Resort Hotel	No of Guest in City Hotel
0	August	3257	5367
1	July	3137	4770
2	October	2575	4326
3	March	2571	4049
4	April	2550	4010
5	May	2535	4568
6	February	2308	3051
7	September	2102	4283
8	June	2037	4358
9	December	2014	2377
10	November	1975	2676
11	January	1866	2249

Data Cleaning

To check if we have missing values, then we will have to replace those missing values with the mean of that feature if that variable is numeric or the constant if it is a categorical feature.

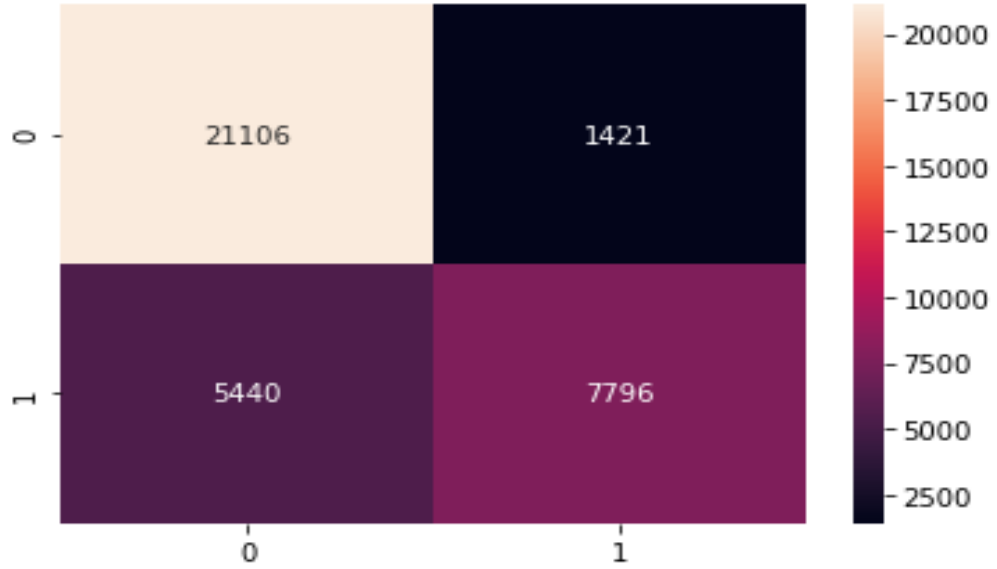




Modeling

- **Logistic Regression**

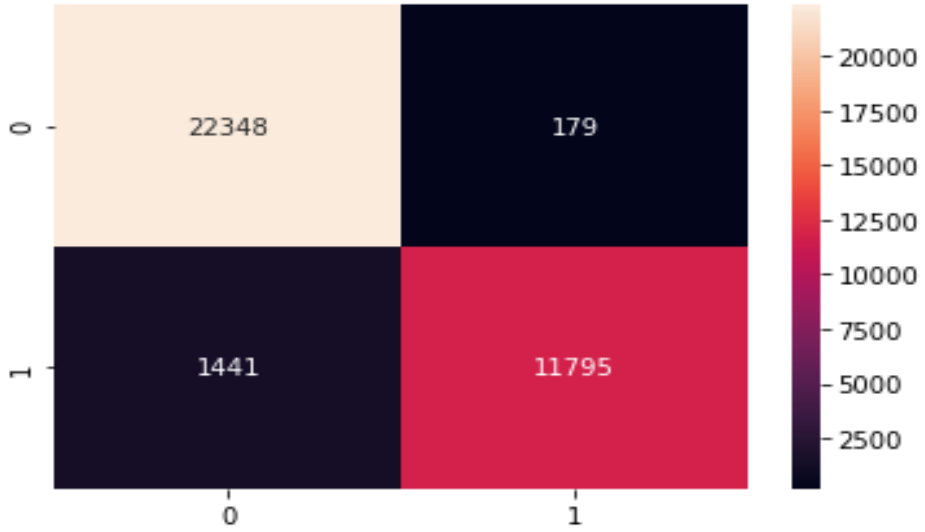
```
Accuracy Score of Logistic Regression is : 0.808153678382686
Confusion Matrix :
[[21106  1421]
 [ 5440  7796]]
precision_score :
0.8458283606379516
Recall Score :
0.5889996977938954
f1_score :
0.6944283614661738
<AxesSubplot:>
```



Modeling

- **Random Forest**

```
Accuracy Score of Logistic Regression is : 0.9547017867628554
Confusion Matrix :
[[22348   179]
 [ 1441 11795]]
precision_score :
0.9850509437113747
Recall Score :
0.8911302508310668
f1_score :
0.935739785799286
<AxesSubplot:>
```



Thank You