Hotel Booking Demand

IST 652

Group 11: Zeyang Zhou,He Pang,

Xiaoben Yin, Jinlong Zhou

[1. Introduction 3](#_Toc71744783)

[Abstract 3](#_Toc71744784)

[Introduction 3](#_Toc71744785)

[Data set 3](#_Toc71744786)

[2. Data set 3](#_Toc71744787)

[2.1 Dataset Overview 3](#_Toc71744788)

[2.2 Descriptive Analysis 4](#_Toc71744789)

[2.2.1 normal variables 4](#_Toc71744790)

[2.2.2 abnormal variables 5](#_Toc71744791)

[3. Data Cleaning and Preparation 6](#_Toc71744792)

[Data Preprocessing 6](#_Toc71744793)

[1. Processing Missing values: 6](#_Toc71744794)

[2. Convert datatype: 7](#_Toc71744795)

[Data transformation 8](#_Toc71744796)

[3. Handle outliers in the dataset 8](#_Toc71744797)

[4. Duplicate values: 9](#_Toc71744798)

[5. Replacing Values: 9](#_Toc71744799)

[4. Visualization EDA 9](#_Toc71744800)

[1. What is the cancellation situation? 10](#_Toc71744801)

[2. What are the monthly cancellations and monthly customers by hotel types? 10](#_Toc71744802)

[3. What is the percentage of booking for each year? 11](#_Toc71744803)

[4. Which are the busiest months for hotels? 12](#_Toc71744804)

[5. How long do people stay in the hotel? 12](#_Toc71744805)

[6. What is the most frequent accomodation type and customer type? 13](#_Toc71744806)

[7. What is the correlation between cancelled booking and other variables? 14](#_Toc71744807)

[5. Models 15](#_Toc71744808)

[a. Feature engineering 15](#_Toc71744809)

[b. Logistic regression 15](#_Toc71744810)

[c. Random forest 15](#_Toc71744811)

[d. Feature importances 16](#_Toc71744812)

[6. Conclusion 17](#_Toc71744813)

[References 18](#_Toc71744814)

# 1. Introduction

## Abstract

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? Our project objective is to solve questions like this.

## Introduction

Travel is a big part of our life and during the travel booking the hotel is always the most important process. The total number of online travel bookings made each year is around 148.3 million, which generates sales of around $755 billion per year in 2020. And it continues to increase since 2014, with an average of 10% every year. But in the real world, many people are troubled by how to book a value hotel. It’s no surprise that over half the people spend more than one week researching their hotel before the holiday. Based on those questions, we make this report to solve it.

## Data set

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

# 2. Data set

## 2.1 Dataset Overview

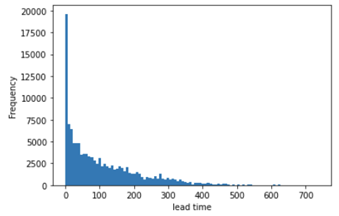
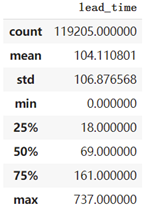
Hotel booking dataset contains 119205 rows with 31 columns.



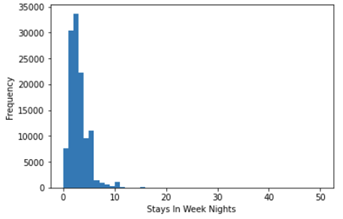
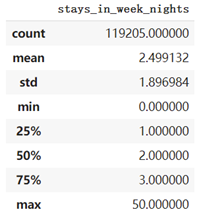
## 2.2 Descriptive Analysis

### 2.2.1 normal variables

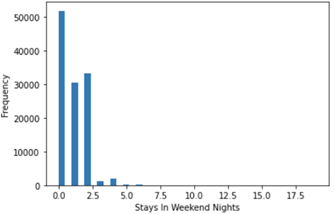
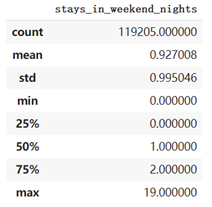
lead time



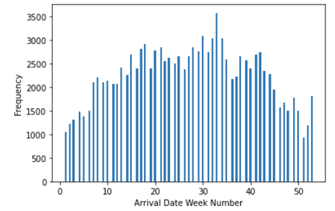
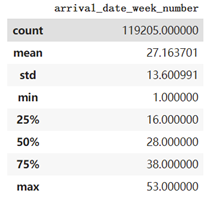
stays in week nights



stays in weekend nights

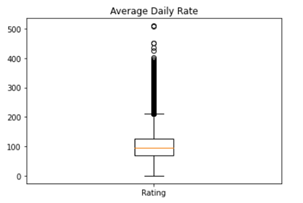
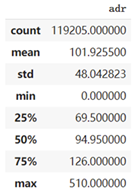


arrival date week number

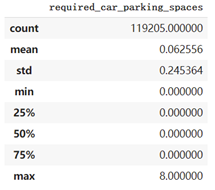


### 2.2.2 abnormal variables

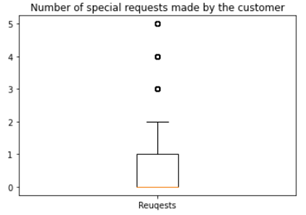
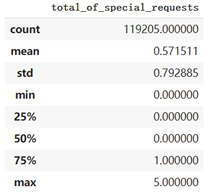
average daily rate



number of car parking spaces required by the customer



number of car parking spaces required by the customer



# 3. Data Cleaning and Preparation

## Data Preprocessing

### 1. Processing Missing values:

First, we would like to process missing values in this dataset. There are 4 columns with missing values which are “agent”, “country”, “children” and “company”.

Since “company” has too many missing values and it is less helpful to our analysis, we decide to delete all of them from the dataset. In column “children”, we fill the missing values with the mean value of “children” since it’s just 0.0035% missing values. In column ‘country’, we fill the missing values with the mode value. In column ‘agent’, we replace the missing values with 0 because these customers may select independent travel and they do not need to report the IDs of their agents.

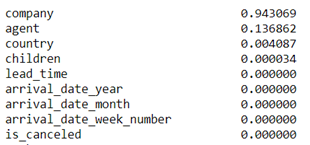


Figure Missing Value

### 2. Convert datatype:

"children" means "Number of children" and "agent" represents "ID of the travel agency that made the booking". Their data types should be "int" in the dataset. However, by hb\_new.describe(), we find that the datatype of them are both "float". So, we convert their datatypes to avoid statistical mistakes （e.g., 1.5 children）

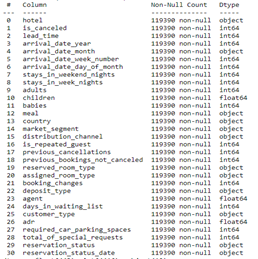


Figure Data Type

## Data transformation

### 3. Handle outliers in the dataset

Based on the dataset description, in one demand, the number of guests which is the sum of “adults”, “children”,” babies” must be bigger than 0. Therefore, we delete all rows with “guests = 0”. And, since the values in column “adr” are required to be positive, we delete all rows with negative “adr” values and an outlier that “adr” = 5400. Similarly, we delete too large values like one booking demand with 10 babies or 10 children in the dataset. It is because no family will bring more than 10 babies with booking just one room.

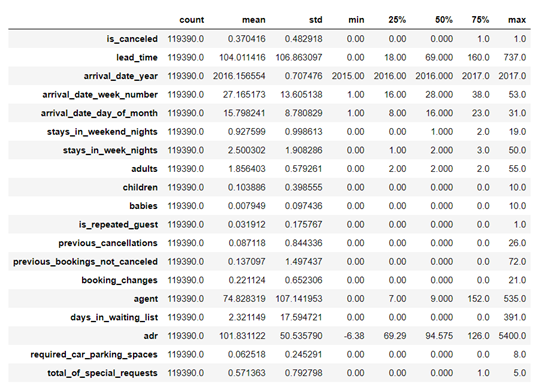


Figure Descriptive analysis

### 4. Duplicate values:

We skip this step because there is no primary key existing in the dataset which means duplicate data are allowed in this case.

### 5. Replacing Values:

In column ‘meal’, we find five different kinds of meal services. But Undefined/SC both represent no meal package, it may cause confusion in the next data analysis. Thus, we replace all Undefined with SC.

# 4. Visualization EDA

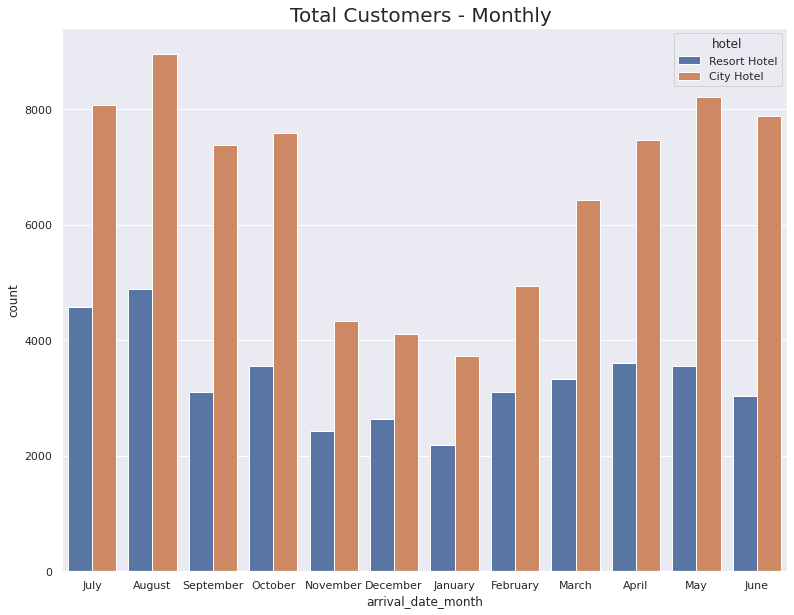
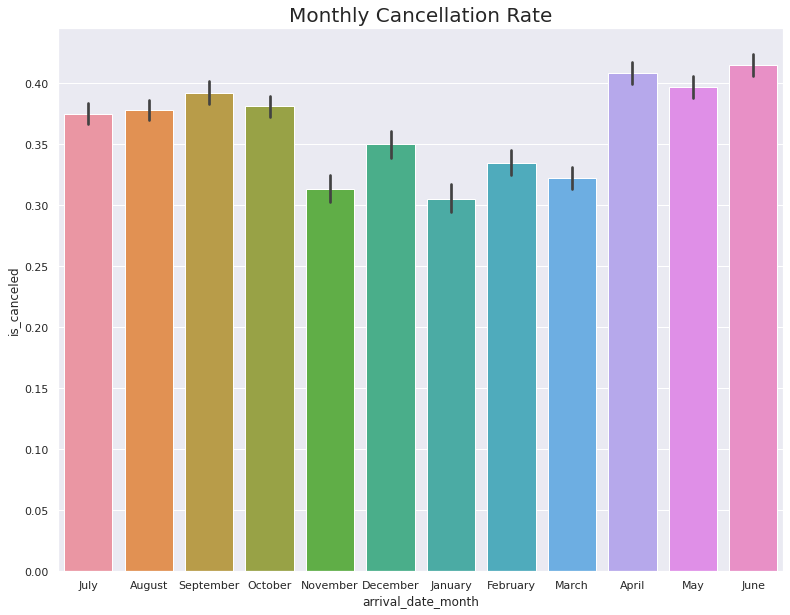
Before building models, we performed exploratory data analysis on our dataset to better understand the pattern of our dataset. The exploratory data analysis mainly focused on answering the questions that were addressed in the inference of project proposal.

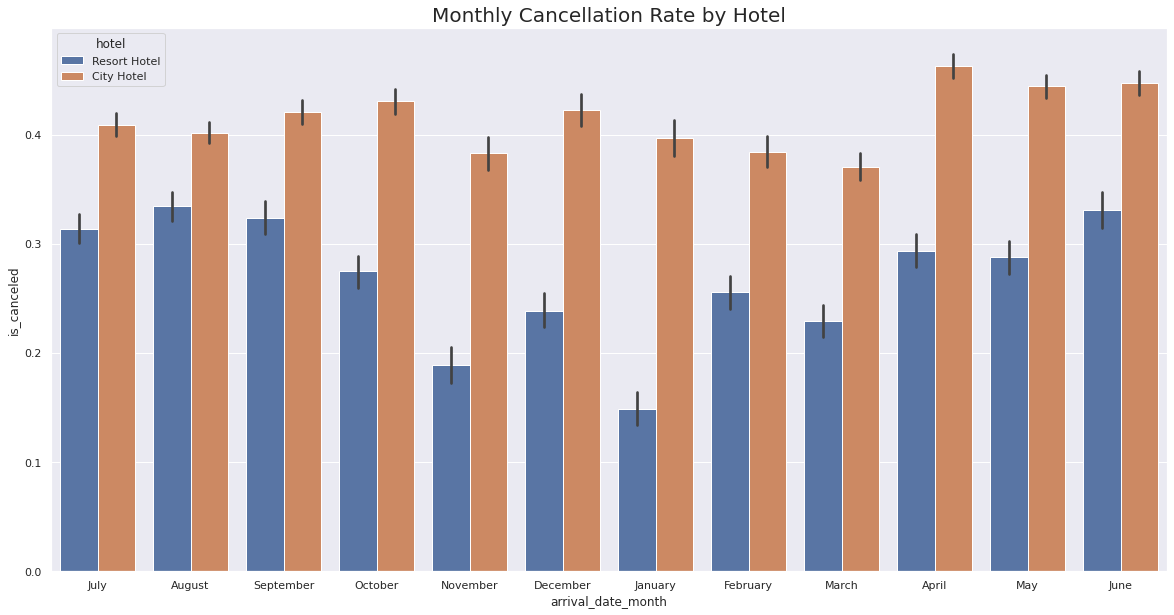
## What is the cancellation situation?



In this dataset, there are 79,160 records of city hotels and 40,045 records of resort hotels. The bar graph shown above depicts the cancellation situation of both hotels. The orange bars represent city hotels and the blue bars represent the resort hotel. In both hotel types, the none-cancellation rates are higher than the cancellation rates. City hotel has the highest canceled bookings and also the highest not canceled booking rate. The ratio of canceled booking is 37.08%, and the ratio of not canceled is 62.92%.

## What are the monthly cancellations and monthly customers by hotel types?

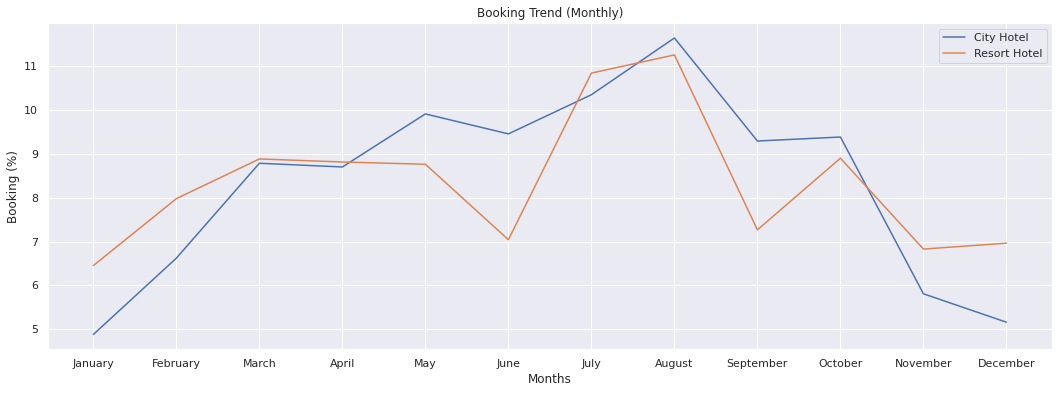


Looking at the top left graph, we can see that in all months city hotels have more customers than resort hotels. When considering the percentage, the resort hotel in summer is closer to city hotels. Looking at the three plots together, an interpretation can be made that fewer customers come in winter months, so it makes sense that the cancellation rates are lower in winter months. The fact that the total cancellation rates of the winter months are low is that the cancellation rates of the resort hotels are low in these months. Shortly, the possibility of cancellation of resort hotel in winter is very low.

## What is the percentage of booking for each year?

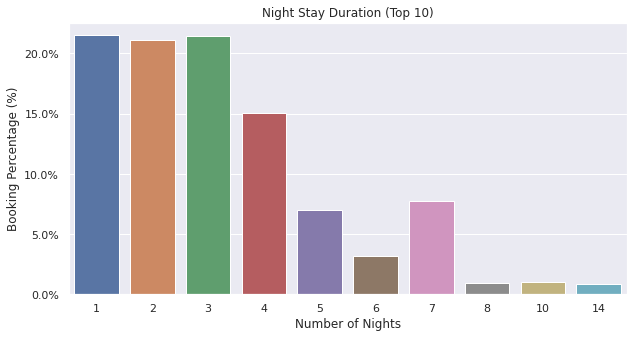
It can be seen from the plot on the left that more than double bookings were made in 2016 compared to 2015. In 2017, the total booking rate decreased by approximately 15%.

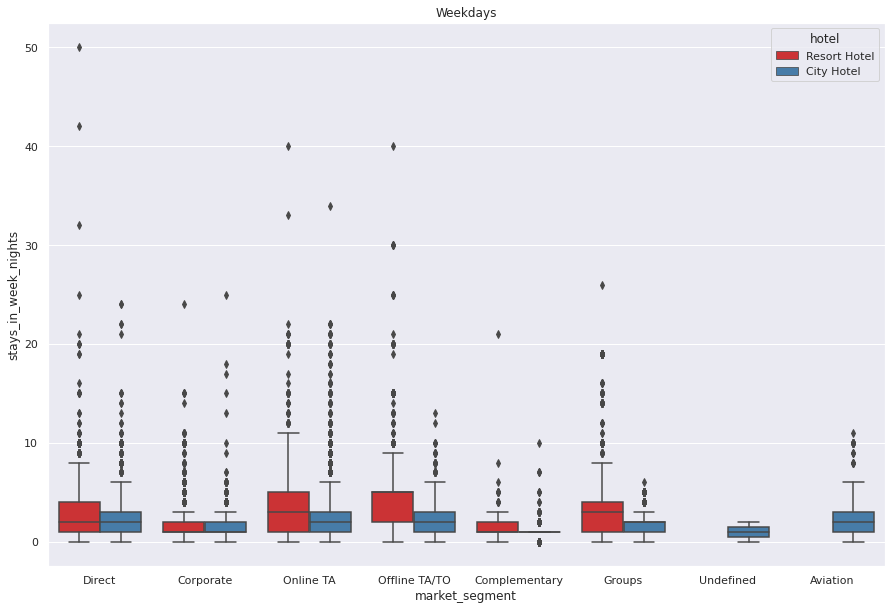
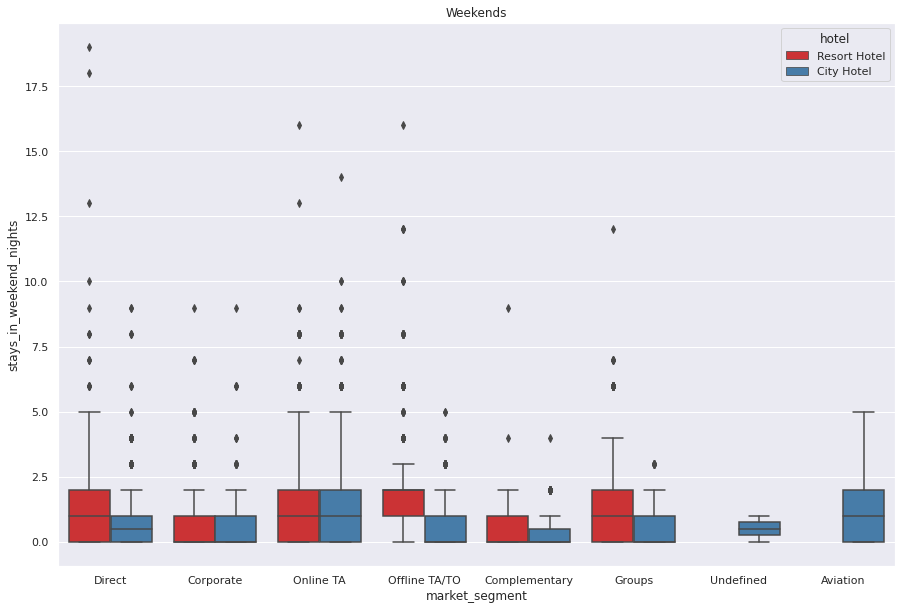
## Which are the busiest months for hotels?



From the plot above, we can see that the busiest months are July and August. I believe the reason could be there are beautiful sceneries and stunning sunshine in summer, when people love to hang out with families and friends. The least bookings were made from January, November and December, which are the coldest months. A specific pattern that attracts my attention is that in the least bookings months, the booking rate of resort hotels is higher than city hotel. However, in June and September, which are near blossom traveling months, the booking rate of resort hotel suddenly declines.

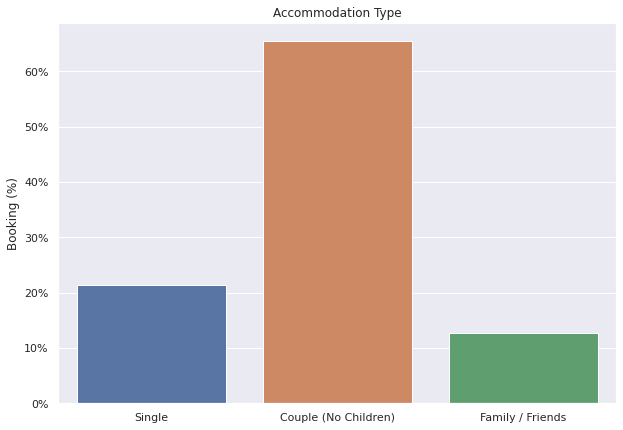
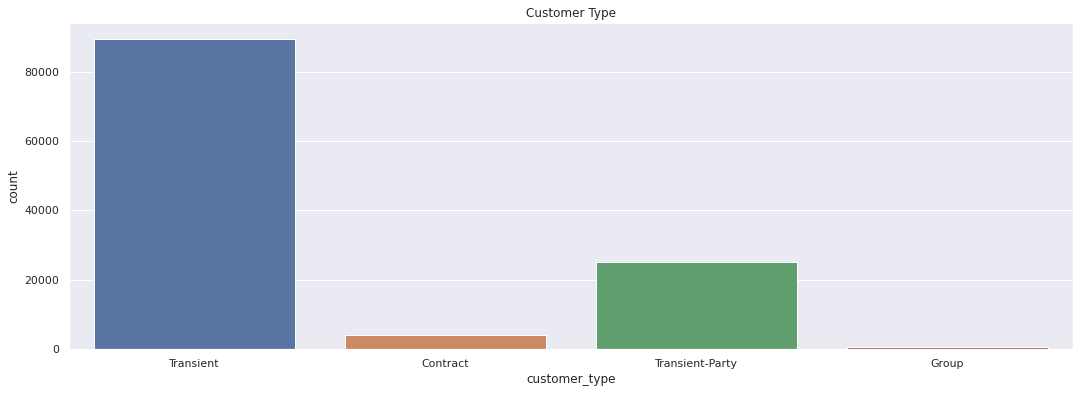
## How long do people stay in the hotel?

Most people stay for one, two, or three days, no matter what type of hotel. 

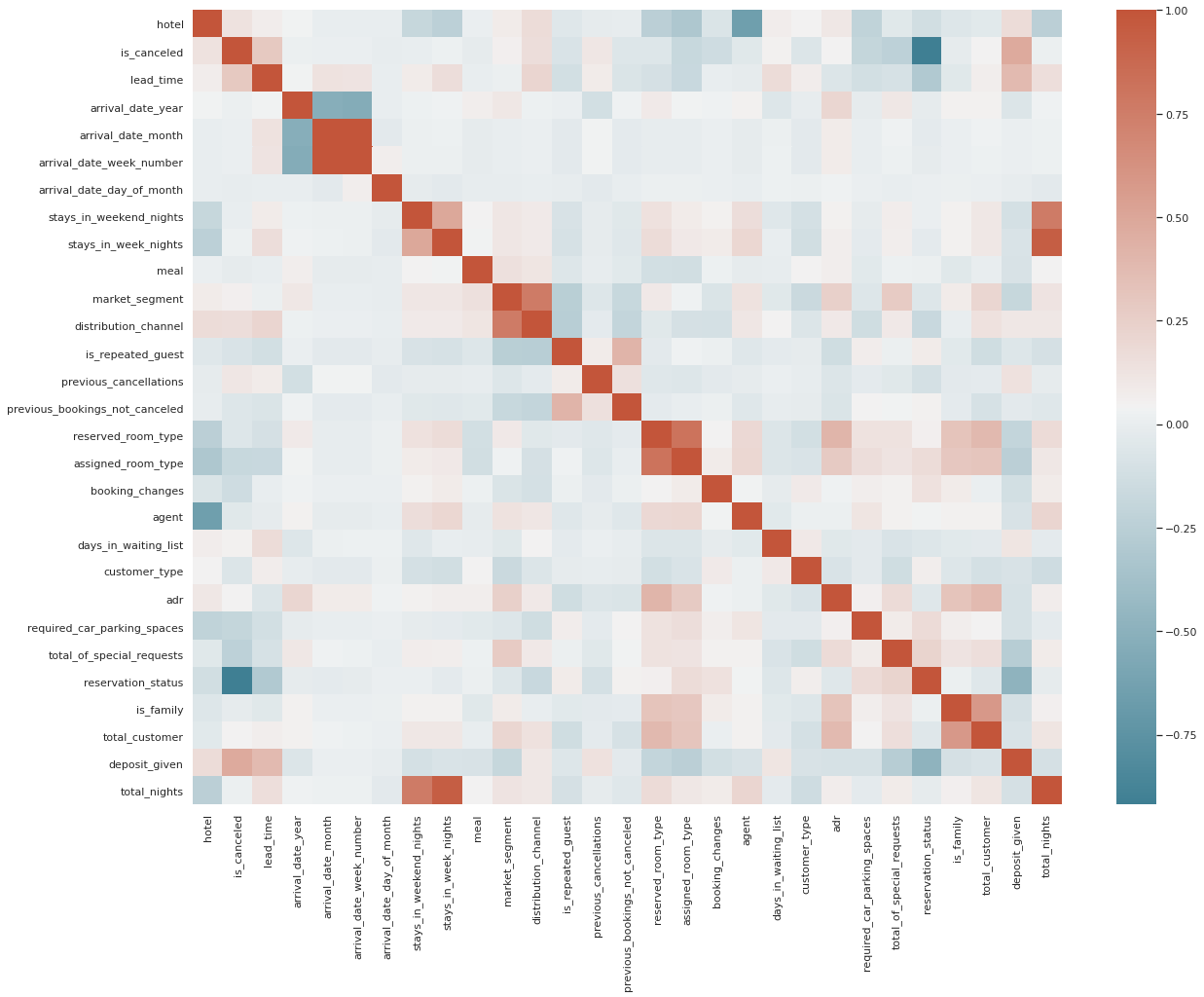
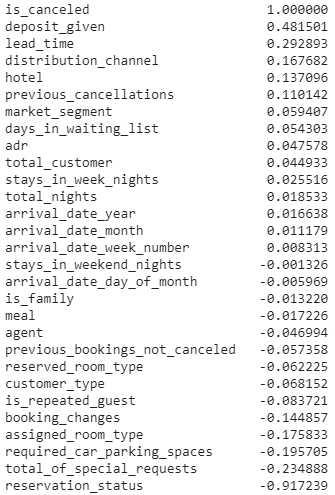
It can be seen that most of the groups are normally distributed, some of them have high skewness. Looking at the distribution, most people do not seem to prefer to stay at the hotel for more than 1 week. But it seems normal to stay in resort hotels for up to 12-13 days. As it turns out, customers from Aviation Segment do not seem to be staying at the resort hotels and have a relatively lower day average. Apart from that, the weekends and weekdays averages are roughly equal. Customers in the Aviation Segment are likely to arrive shortly due to business. Also probably most airports are a bit away from the sea and it's most likely to be closer to city hotels.

## What is the most frequent accomodation type and customer type?

The accommodation type in this dataset can be divided into three types: single, couple and family or friends. The most popular accomodation type is couples, so we could suggest hotels make accordingly plans for couples. There are four customer types: contract, when the booking has an allotment or other type of contract associated with 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 with other transient booking; and transient-party, when the booking is transient, but is associated with at least other transient booking. Transient accounts for about 75% of all customers, followed by transient-party, which accounts for about 21%. Most customers are Transient, meaning they are walk-in guests, last-minute or bookers, or simply people that require a very short-term stay in your facility. Transient customers are one of the major market segments consisting of individuals or groups.

## What is the correlation between cancelled booking and other variables?

We made a correlation heatmap on the left, and the list on the right shows the correlation scores. From the result, we can see that the most correlative variables are whether or not the deposit is given and whether the booking is canceled or not. The following factors are lead time, distribution channel, hotel type and previous cancellations situation, which accounts for approximately 20% or 10%.

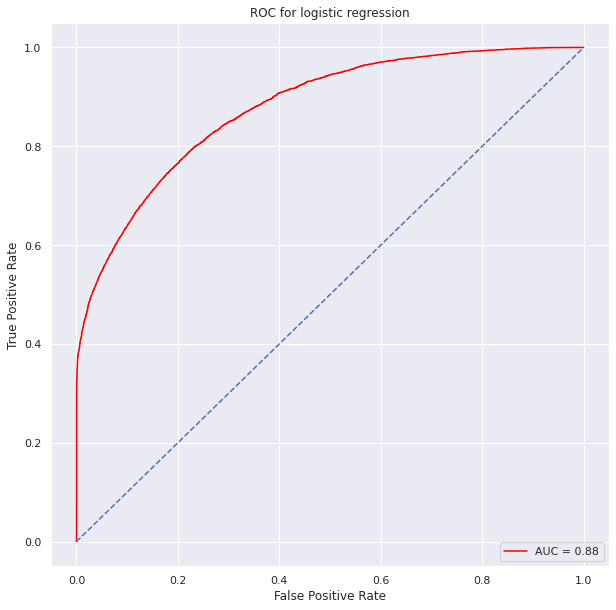
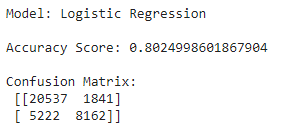
# 5. Models

## Feature engineering

Before building models, we did feature engineering to some of the columns. First, we turned the hotel type into dummy variables: 0 represents the resort hotel and 1 represents city hotel. Then, we transfer all the categorical months into numbers. After that, we divide adults, children and babies into groups based on their amount. Also, for the deposit column, we assigned 0 for those who don’t pay for deposit in advance and 1 for those who paid the deposit. For other columns without ordinal, we performed one-hot encoding. The columns include meal, market segment, distribution channel, reserved room type, assigned room type, customer type and reservation status.

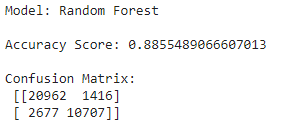
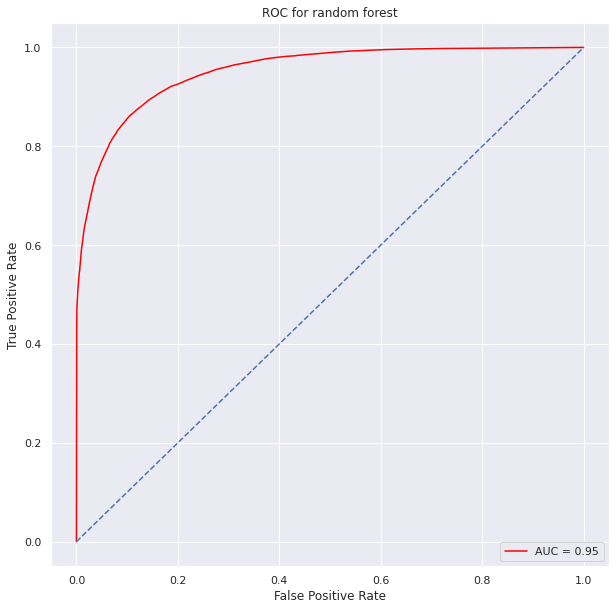
## Logistic regression

Since our target variable is either resort hotel or city hotel, we believe that logistic regression is one of the best suitable models. We used the accuracy score to evaluate the performance, and the accuracy of the logistic regression model is 80.25%. We also plot the ROC curve for logistic regression, and the AUC score is 0.88, which is pretty high.



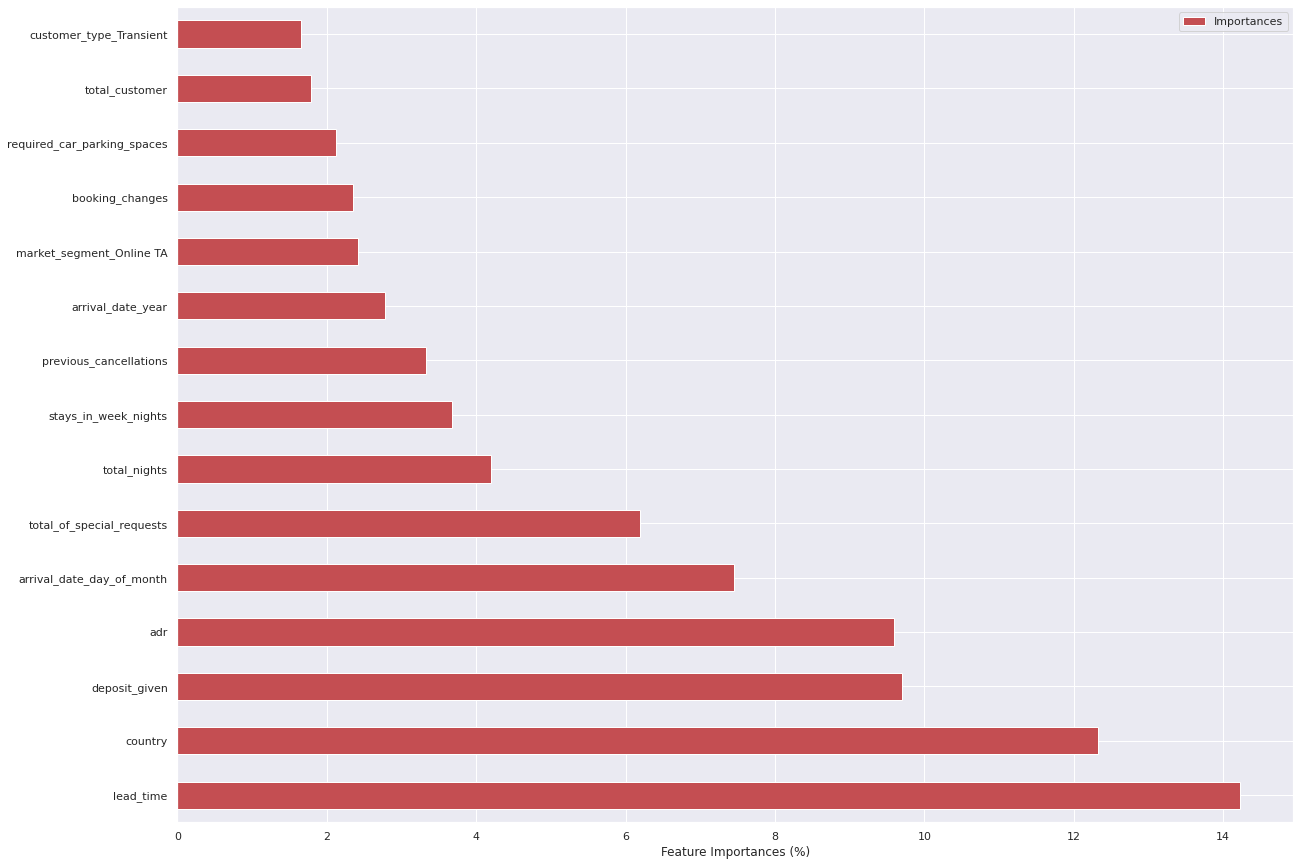
## Random forest

Apart from the logistic regression model, we built a random forest model as well. The accuracy of random forest is even better than logistic regression, which equals 88.55%. The AUC score of random forest is 0.95.

## Feature importances

Based on the accuracy scores, we decide to gain feature importances from higher performance model, random forest model. The graph below shows the percentage of feature importances. Similar to the ranking in correlation list, lead time accounts for about 14% of the variation in determining hotel type, followed by country, deposit, address, arrival date of month. The rest of variables account for very little of the changes in our target.



# 6. Conclusion

* Resort hotels tend to have less bookings in comparison to city hotels so they need to work on their marketing strategy and promote the hotels more, especially on social media.
* Resort hotels could also reduce prices to increase booking percentages.
* May-August happens to be the busiest months but so the hotels should target more customers and try to do more business during these times.
* Although city hotels have more bookings, they also tend to have more cancellations so to prevent this they could take advance money during vacation. This would ensure most bookings to not be cancelled. They could also apply no-refund policies or make the refund policies rather strict so the customers choose not to cancel.
* It is quite clear most customers travel in pairs and bringing children or babies along are very rare so the hotels could advertise in ways that attract couples more and also business travellers.
* Most guests do not return but as these customers have already visited once, advertisements should be targeted in such ways so they are bound to return the next time they visit. The customers could also be offered special benefits if they do return to stay.

# 

# References

*Hotel booking demand*. (n.d.). Retrieved May 8, 2021, from<https://kaggle.com/jessemostipak/hotel-booking-demand>

Soomro, A. Q. (2020, April 20). *Exploratory Data Analysis of the Hotel Booking Demand with Python*. Medium.<https://medium.com/analytics-vidhya/exploratory-data-analysis-of-the-hotel-booking-demand-with-python-200925230106>

*Rfordatascience/tidytuesday*. (n.d.). GitHub. Retrieved May 8, 2021, from<https://github.com/rfordatascience/tidytuesday>?>

Soomro, A. Q. (2021). *Aaqibqadeer/Hotel-booking-demand* [Jupyter Notebook].<https://github.com/aaqibqadeer/Hotel-booking-demand> (Original work published 2020)