# 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 following queries: 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

In the start we need to import our train dataset for the exploratory data analysis.

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 of 71633 in total.

The dataset contain 2 missing values in the children variable so we will omit these missing values from the dataset. When above 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.

**hotel:** Categorical, booking information for a city hotel or a resort hotel  
**is\_canceled:** 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 children  
**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 canceled 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 canceled 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 cancellations of bookings. In this report, we will also classify our dataset by various techniques so that we can calculate the maximum precision for the main purpose. This helps us lower the booking cancellation 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 train dataset;

## hotel is\_canceled lead\_time arrival\_date\_year  
## City Hotel :47586 0:45099 Min. : 0.0 Min. :2015   
## Resort Hotel:24045 1:26532 1st Qu.: 18.0 1st Qu.:2016   
## Median : 69.0 Median :2016   
## Mean :103.8 Mean :2016   
## 3rd Qu.:160.0 3rd Qu.:2017   
## Max. :737.0 Max. :2017   
##   
## arrival\_date\_month arrival\_date\_week\_number arrival\_date\_day\_of\_month  
## August : 8293 Min. : 1.00 Min. : 1.00   
## July : 7632 1st Qu.:16.00 1st Qu.: 8.00   
## May : 7133 Median :28.00 Median :16.00   
## October: 6699 Mean :27.19 Mean :15.77   
## April : 6625 3rd Qu.:38.00 3rd Qu.:23.00   
## June : 6569 Max. :53.00 Max. :31.00   
## (Other):28680   
## stays\_in\_weekend\_nights stays\_in\_week\_nights adults   
## Min. : 0.0000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.0000 1st Qu.: 1.000 1st Qu.: 2.000   
## Median : 1.0000 Median : 2.000 Median : 2.000   
## Mean : 0.9247 Mean : 2.497 Mean : 1.857   
## 3rd Qu.: 2.0000 3rd Qu.: 3.000 3rd Qu.: 2.000   
## Max. :18.0000 Max. :42.000 Max. :55.000   
##   
## children babies meal country   
## Min. : 0.0000 Min. : 0.000000 BB :55337 PRT :29092   
## 1st Qu.: 0.0000 1st Qu.: 0.000000 FB : 465 GBR : 7257   
## Median : 0.0000 Median : 0.000000 HB : 8711 FRA : 6306   
## Mean : 0.1043 Mean : 0.007874 SC : 6397 ESP : 5138   
## 3rd Qu.: 0.0000 3rd Qu.: 0.000000 Undefined: 721 DEU : 4362   
## Max. :10.0000 Max. :10.000000 ITA : 2223   
## (Other):17253   
## market\_segment distribution\_channel is\_repeated\_guest  
## Aviation : 144 Corporate: 4018 0:69339   
## Complementary: 415 Direct : 8752 1: 2292   
## Corporate : 3162 GDS : 127   
## Direct : 7608 TA/TO :58733   
## Groups :11857 Undefined: 1   
## Offline TA/TO:14534   
## Online TA :33911   
## previous\_cancellations previous\_bookings\_not\_canceled reserved\_room\_type  
## Min. : 0.0000 Min. : 0.0000 A :51591   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 D :11477   
## Median : 0.0000 Median : 0.0000 E : 3952   
## Mean : 0.0832 Mean : 0.1365 F : 1763   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 G : 1259   
## Max. :26.0000 Max. :70.0000 B : 646   
## (Other): 943   
## assigned\_room\_type booking\_changes deposit\_type agent   
## A :44401 Min. : 0.0000 No Deposit:62804 9 :19252   
## D :15159 1st Qu.: 0.0000 Non Refund: 8732 NULL : 9781   
## E : 4755 Median : 0.0000 Refundable: 95 240 : 8302   
## F : 2240 Mean : 0.2232 1 : 4327   
## G : 1542 3rd Qu.: 0.0000 14 : 2174   
## C : 1433 Max. :21.0000 7 : 2137   
## (Other): 2101 (Other):25658   
## company days\_in\_waiting\_list customer\_type adr   
## NULL :67566 Min. : 0.00 Contract : 2450 Min. : -6.38   
## 40 : 540 1st Qu.: 0.00 Group : 343 1st Qu.: 69.00   
## 223 : 464 Median : 0.00 Transient :53821 Median : 94.67   
## 67 : 146 Mean : 2.35 Transient-Party:15017 Mean : 101.94   
## 45 : 141 3rd Qu.: 0.00 3rd Qu.: 126.00   
## 153 : 128 Max. :391.00 Max. :5400.00   
## (Other): 2646   
## required\_car\_parking\_spaces total\_of\_special\_requests reservation\_status\_date  
## Min. :0.00000 Min. :0.0000 Min. :2014-10-17   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:2016-02-01   
## Median :0.00000 Median :0.0000 Median :2016-08-06   
## Mean :0.06298 Mean :0.5715 Mean :2016-07-29   
## 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:2017-02-08   
## Max. :8.00000 Max. :5.0000 Max. :2017-09-14   
##

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 narrates the pattern of response to the variables.