Hotel Booking Analysis (EDA)

RAHUL KUMAR GUPTA, AFFIAN BIN NISHANT,
AMISHA KHARE, RAHUL THUMAR
Data science trainees,
AlmaBetter, Bangalore

Abstract:

This data article describes datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Datasets share the same structure, with 32 variables describing the 40060 observations of H1 and 79330 observations of H2. Each observation represents a hotel booking-Datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

Keywords:
Exploratory Data Analysis (EDA), Date set, Hotel, column.

1.Problem Statement and Objective

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? This hotel booking dataset can help you explore those questions!

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. All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that govern the bookings.

1.1OBJECTIVE

Create in-depth analysis to figure out the standard patterns of booking

Generate a report derive a strategy for the marketing team. So that the strategy can create an impact on lagging steps for business development and their growth

2. Introduction (EDA)

Today there are quite a few widespread misconceptions of exploratory data analysis (EDA). One of these misperceptions is that EDA is said to be opposed to statistical modeling. Actually, the essence of EDA is not about putting aside all modeling and preconceptions; rather, researchers are urged not to start the analysis with a strong preconception only, and thus modeling is still legitimate in EDA. In addition, the nature of EDA has been changing due to the emergence of new methods and convergence between EDA and other methodologies, such as data mining and resampling. Therefore, conventional conceptual frameworks of EDA might no longer be capable of coping with this trend. In this article, EDA is introduced in the context of data mining and resampling with an emphasis on three goals: cluster detection, variable selection, and pattern recognition. Two Step clustering, classification trees, and neural networks, which are powerful techniques to accomplish the preceding goals, respectively, are illustrated with concrete examples. EDA helps in following purpose

- Descriptive analytics can be employed to further understand patterns, trends, and anomalies in data.
- Used to perform research in different problems like: bookings cancellation prediction, customer segmentation, customer satiation, seasonality, among others.
- Researchers can use the datasets to benchmark bookings' prediction cancellation models against results already known.

3. DATA (HOTEL BOOKING ANALYSIS)

Tourism and travel related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry. However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry's specific data. Hence, this hotel datasets with demand data are shared to help in overcoming this limitation. The datasets now

made available were collected aiming at the development of prediction models to classify a hotel booking's likelihood to be canceled. Nevertheless, due to the characteristics of the variables included in these datasets, their use goes beyond this cancellation prediction problem. One of the most important properties in data for prediction models is not to promote leakage of future information. In order to prevent this from happening, the timestamp of the target variable must occur after the input variables' timestamp. Thus, instead of directly extracting variables from the bookings database table, when available, the variables' values were extracted from the bookings change log, with a timestamp relative to the day prior to arrival date (for all the bookings created before their arrival date).

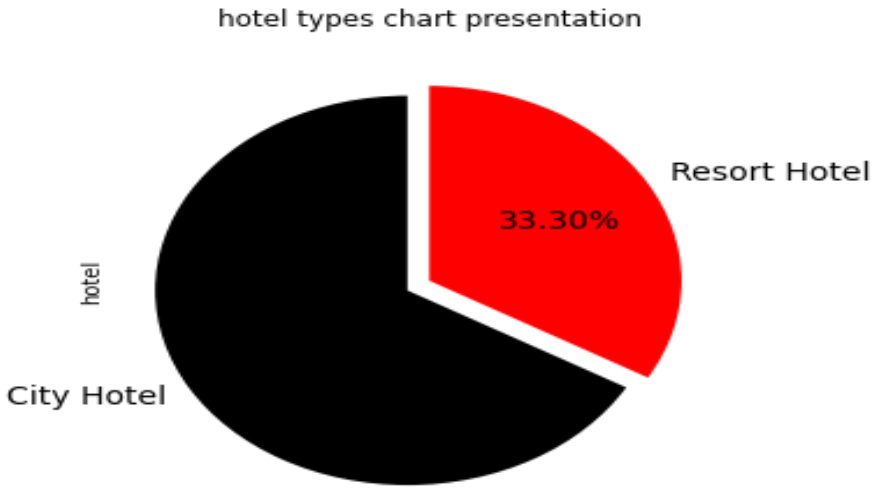
4. Steps involved:

DATA PREPERATION AND CLEANING

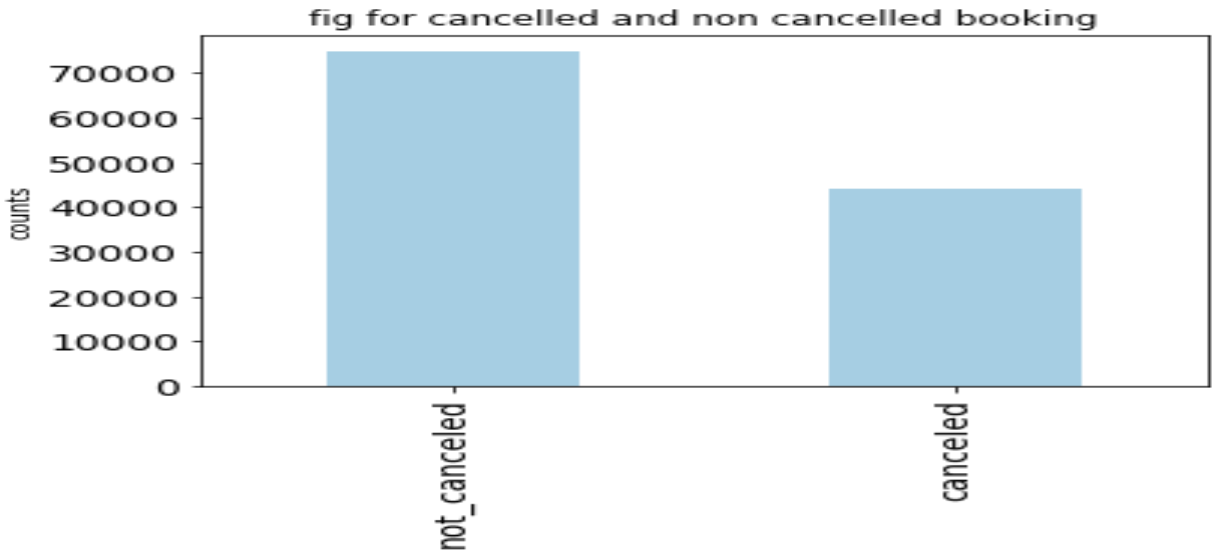
● 1) Hotel booking analysis data set having total 119390 row and 32 columns, out of which column having agent (16340-null value) and company (112593- Null value), this two column doesn't have impact on EDA so I dropped that column.
● 2) Apart from this column having country and adults have 488 and 4 Null value so I dropped that row.
● 3) Now our data is cleaned and ready for EDA.

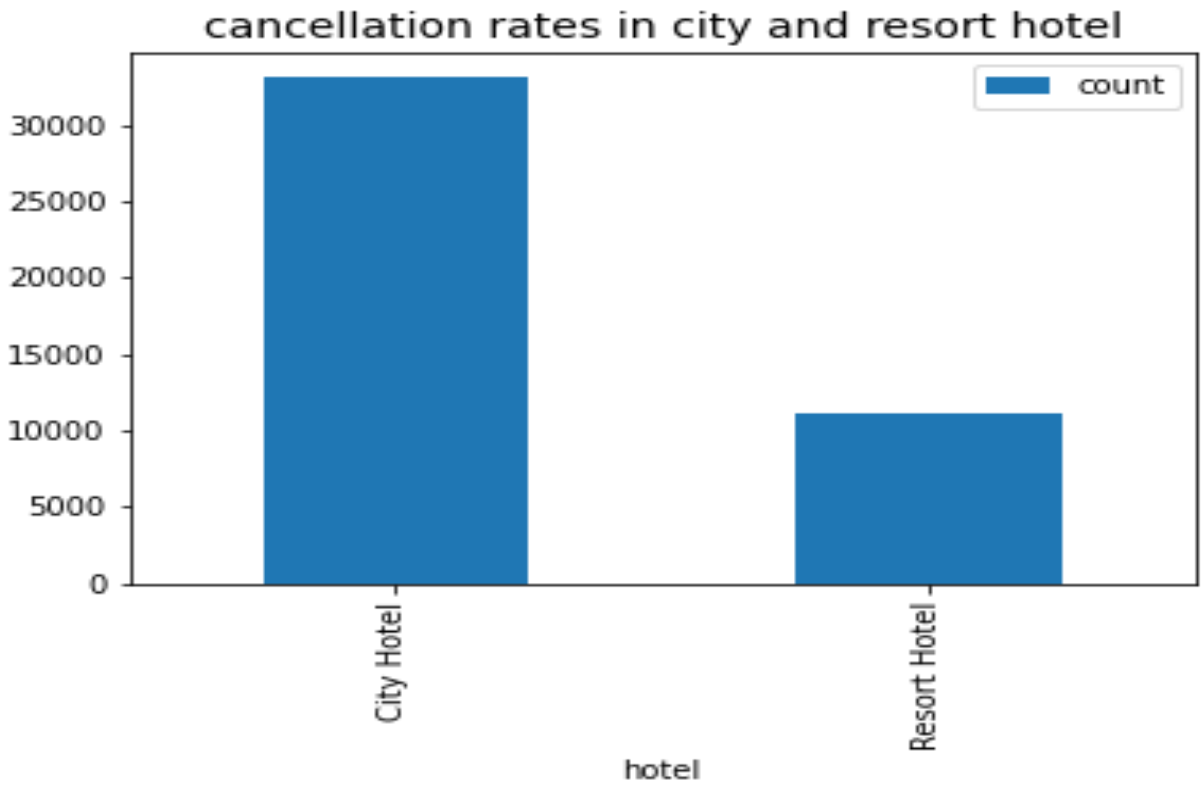
Exploratory Data Analysis

1) Types Of Hotel Present: Total 2 types of hotel booking present: Resort Type_33.30%, : Citytype_66.69%

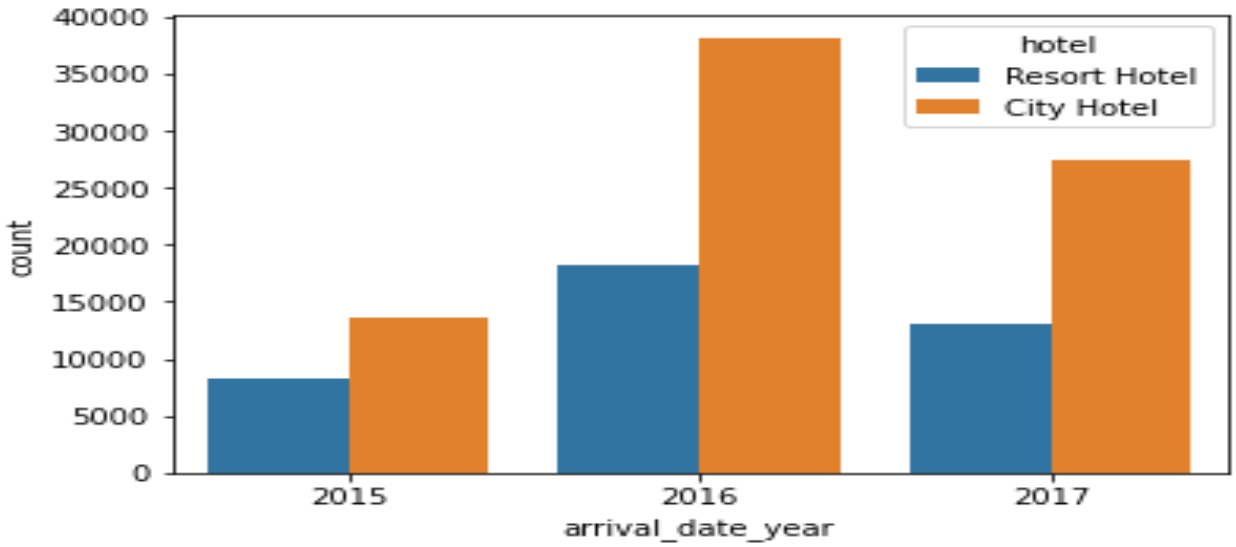


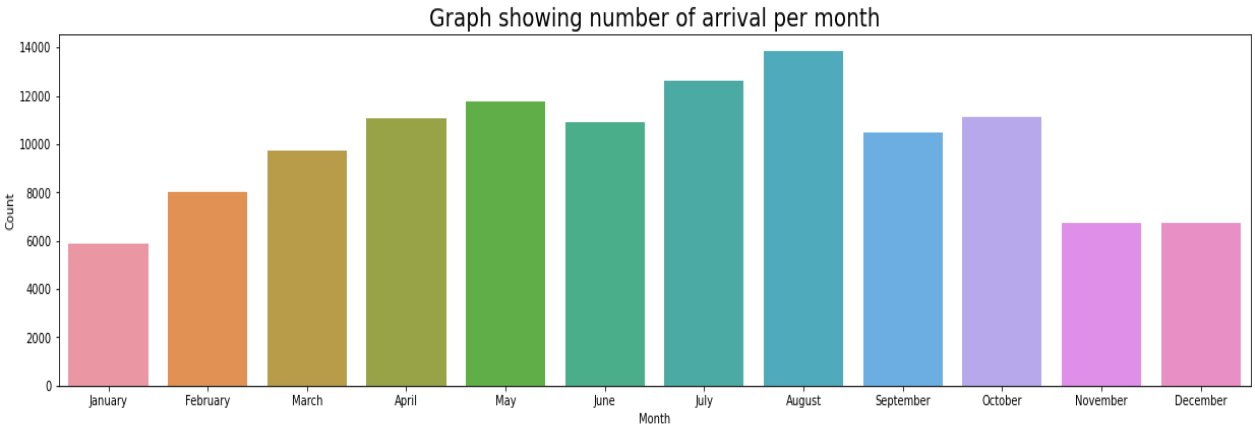
2) Analysis for canceled and non-canceled booking: it seems most of the booking are not cancelled but around 33% booking are cancelled in which most of the cancellation are from city hotel.

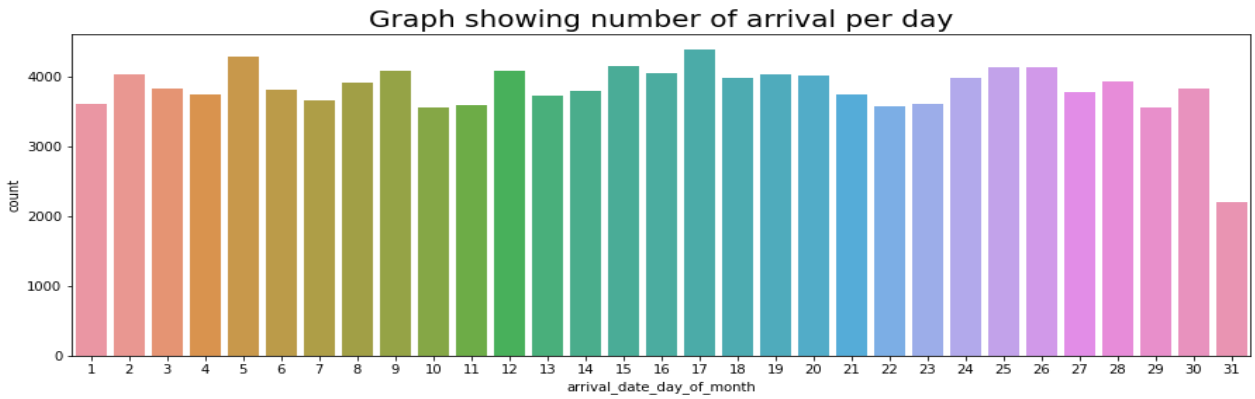




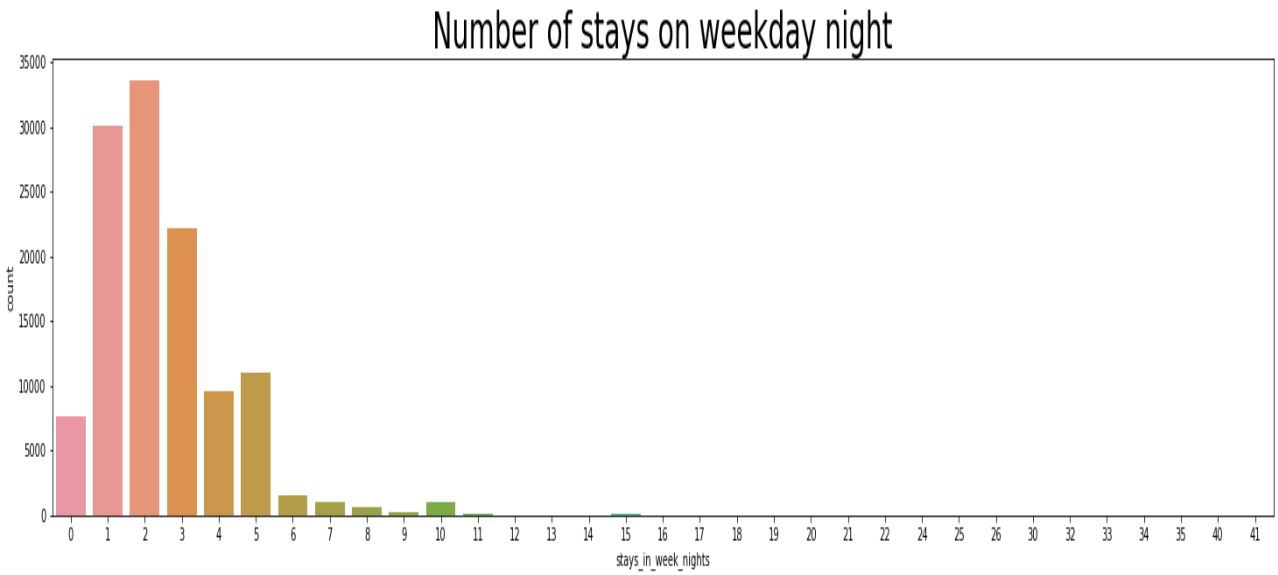
3)Analysis on Arrival Period: Highest number of arrival was 2016 and with monthly basics increasing and peak trend is May to august due to summer.



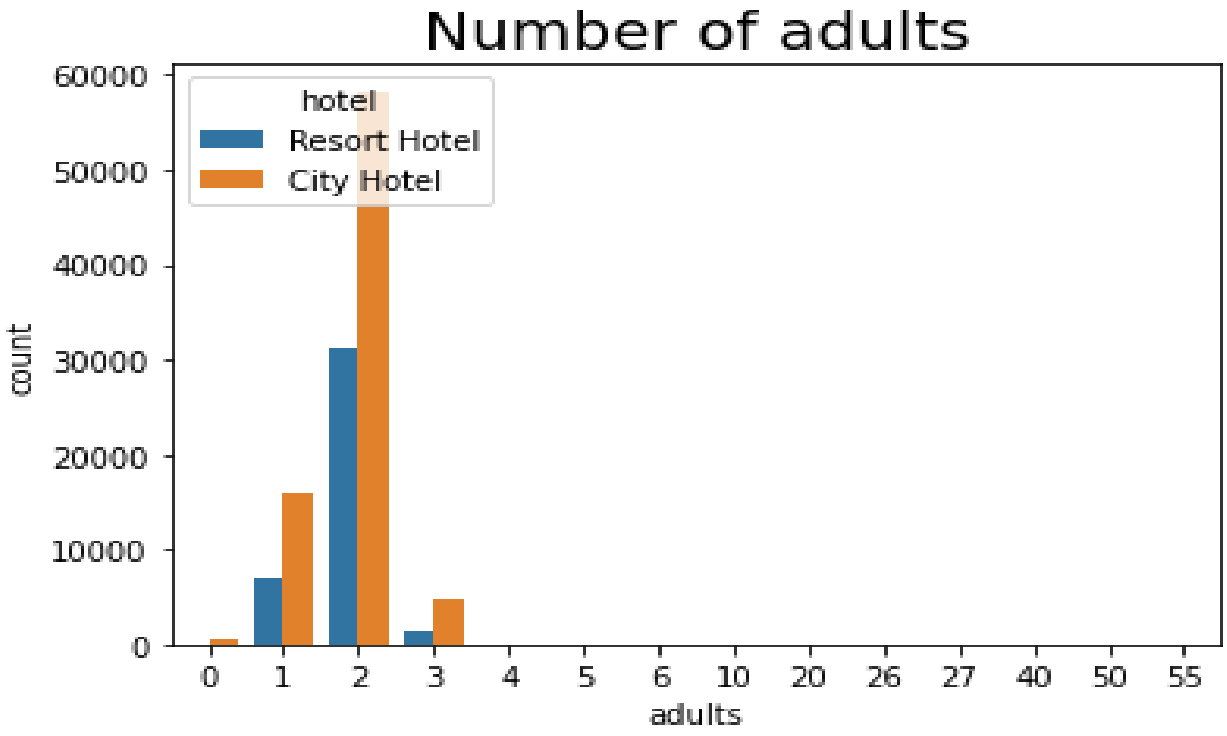


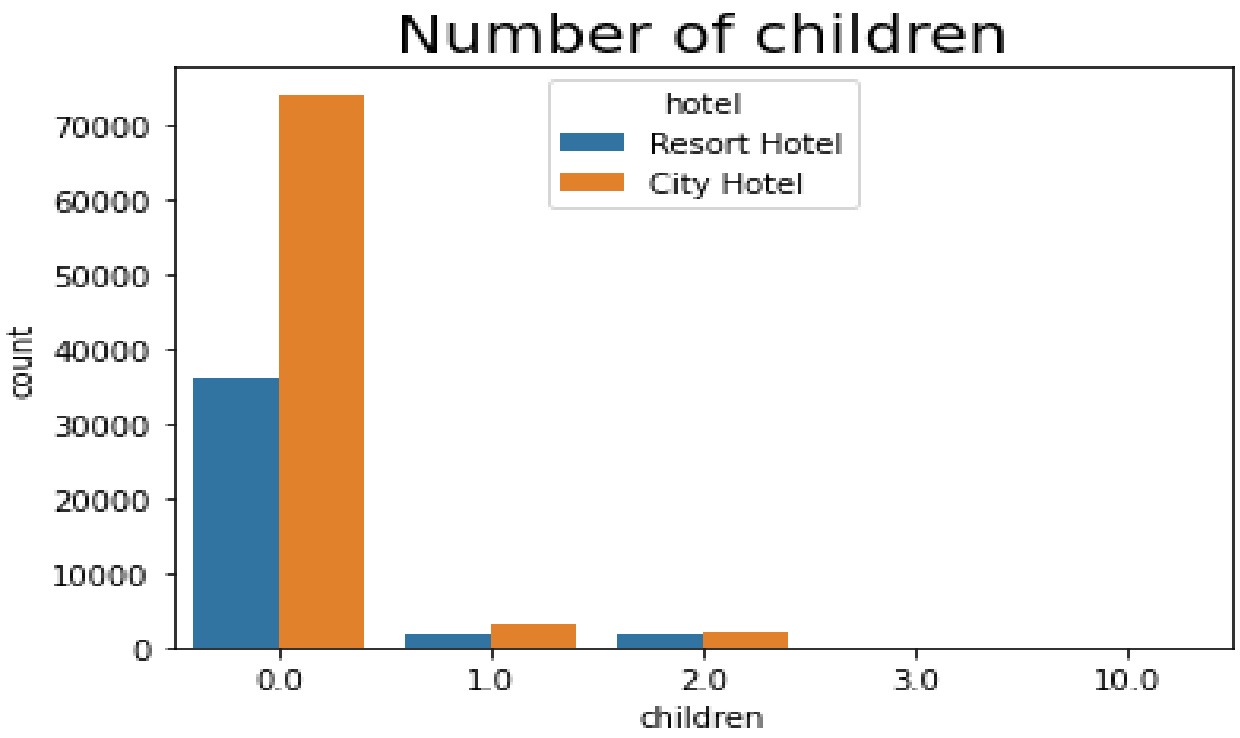


4) Stay is Over Weekend or weekdays: Random, no such Pattern for weekday and weekend.



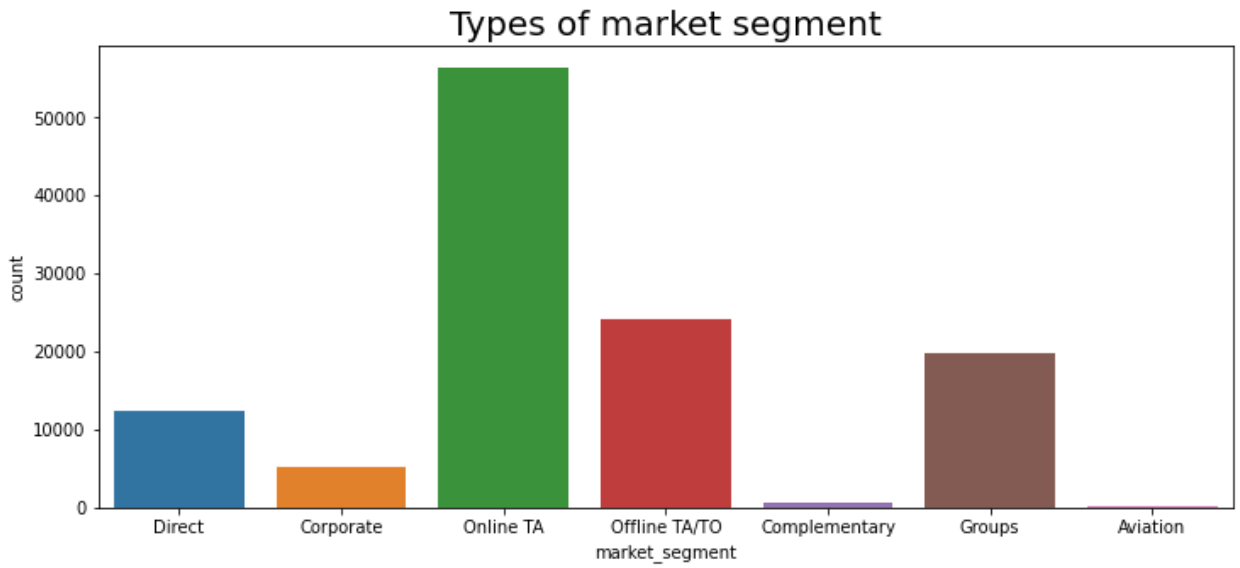
5)Types Of Visitors: It seems that majority of the visitors travel in pair. Those that travel with children or babies have no specific preference for the type of hotel. We do see that those bringing babies along prefer resort hotels.

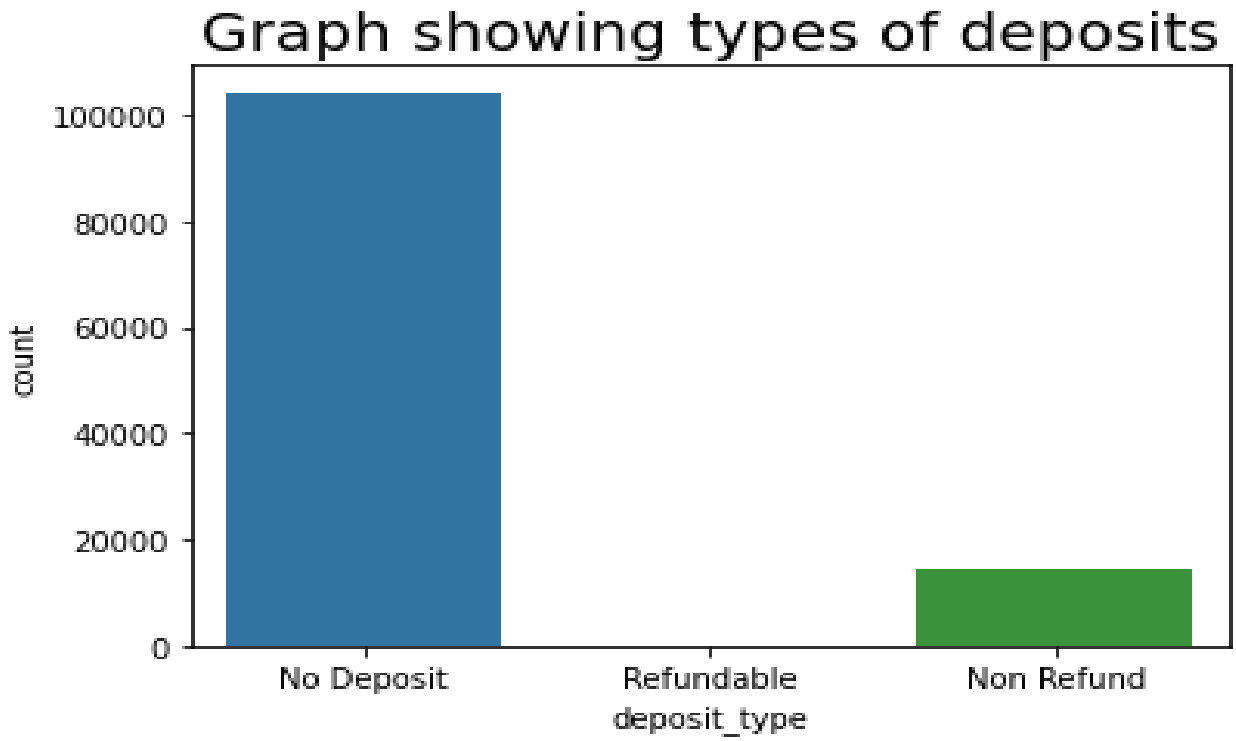


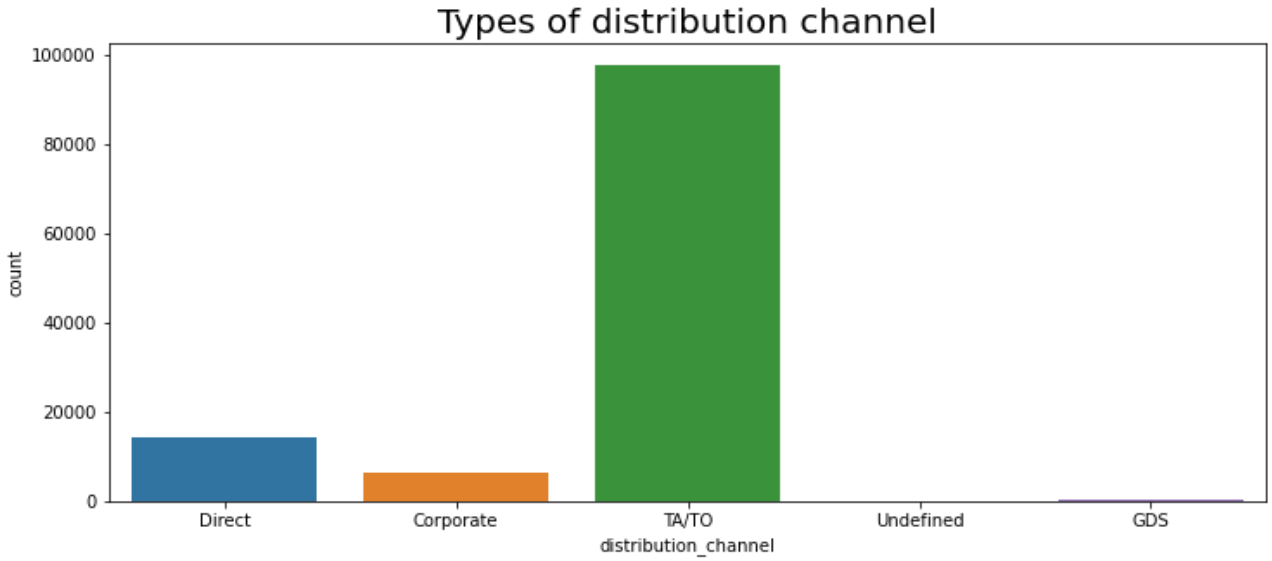


6) Visitors Country: We have a huge number of visitors from western Europe, namely France, UK and Portugal being the highest.

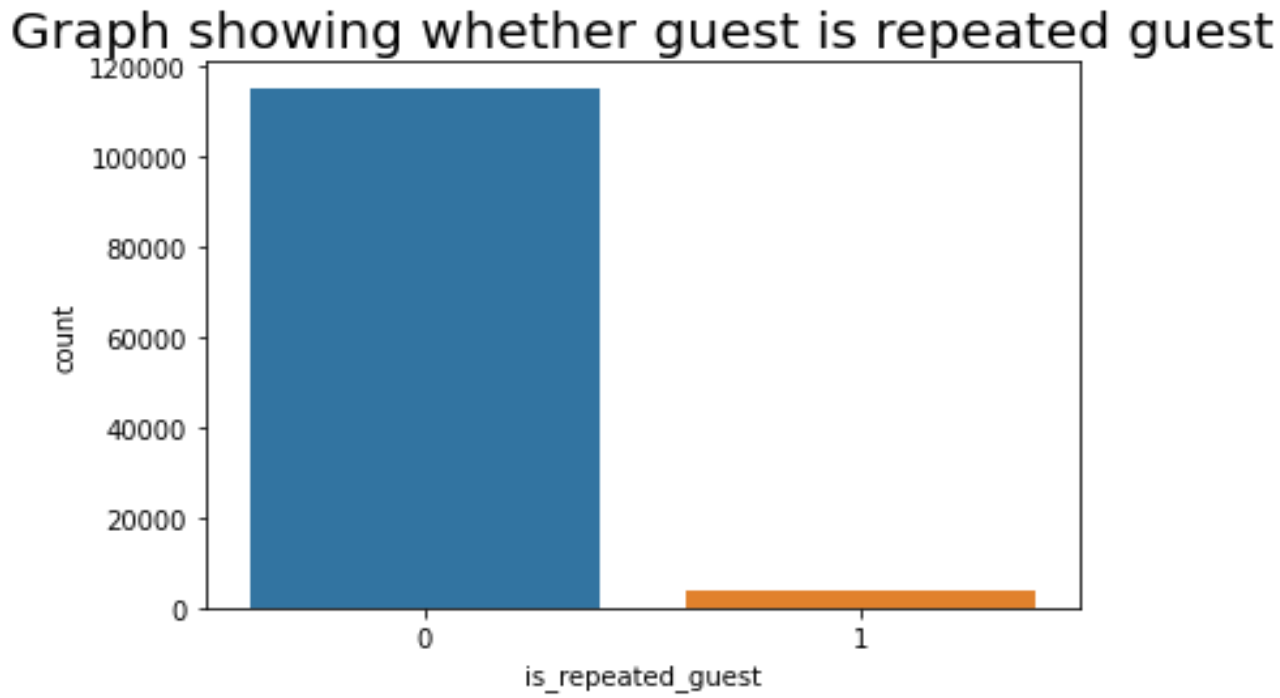
7) market segments and distribution channel: Most of the booking was done through online, with TA/TO distribution channel with that non depository types, which may cause one of the biggest reason for most of the cancellation.



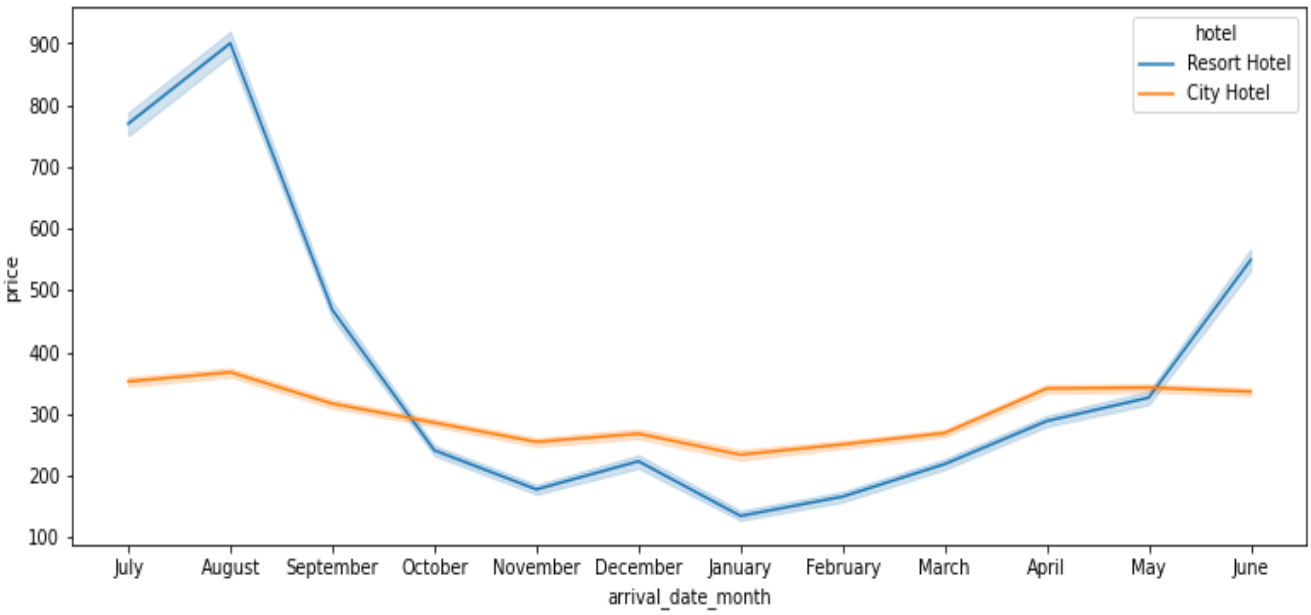




8) Repeat of Guest : Most of the guest are non repeated we can work on this, such as some coupon for repeated guest and some discount for them.



9) Price Distribution with Respective Month: price during Oct to March is Very low. Prices of resort hotel are much higher. It seems that that is definitely, the case since resort hotels specialize in that. Prices of city hotel do not fluctuate that much.



8) SUMMARY

1. Majority of the hotels booked are city hotel.
2. The high rate of cancellations can be due high no deposit policies.
3. Highest arrival will be May to August, summer period
4. Majority of the guests are from Western Europe
5. Nov to Jan Month is least price for hotel
6. Majority of guest are non-repeating, we can use advertisement and some discount for repeated guest.

REFERENCE

[1] N. Antonio, A. Almeida, L. Nunes, Predicting hotel bookings cancellation with a machine learning classification model, in: Proc. 16th IEEE Int. Conf. Mach. Learn. Appl., IEEE, Cancun, Mexico, 2017: pp. 1049–1054. doi:10.1109/ICMLA.2017.00-11.

[2] International Civil Aviation Organization, Guidelines on Passenger Name Record (PNR) data, (2010). https://www.iata.org/iata/passenger-data-toolkit/assets/doc_library/04-pnr/New%20Doc%209944%201st%20Edition%20PNR.pdf (accessed February 17, 2016).

[3] D. Abbott, Applied predictive analytics: Principles and techniques for the professional data analyst, Wiley, Indianapolis, IN, USA, 2014.

[4] Microsoft, SQL Server Management Studio (SSMS), (2017). https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms (accessed March 24, 2018).

[5] American Hotel & Lodging Association, Uniform system of accounts for the lodging industry, 11th Revised edition, Educational Institute, New York, 2014.