**Hotel Booking Analysis**

**Souvik Bhattacharyya**

**Data Science Trainees,**

**Almabetter, Bangalore**

**Abstract:**

This project deals with the trends of market in the perspective of hotel booking and how it is affected by variables such as date, time of booking, size of family etc. The study comprises of extensive research on hotel bookings in two sectors, namely Resorts and City hotels. The work done is aimed at identifying various aspect and features which customers would prefer in either of those aforementioned options like fooding. The results that were generated from this research aim at finding the best period for booking a hotel room as per the convenience of guests. The proper implementation of the analysis would help guests find their best deals in terms of hotel bookings.

***Keywords: libraries, pricing, date & time, guests, visualization and trends & curves.***

**Problem Statement**

This project deals with questions like “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 the guests this wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset would help the said guests to explore those questions and help them find a suitable hotel to stay.

This analysis of data set contains booking information for a city hotel and resorts, 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.

***\*For the safety of the guest all personally identifying information has been removed from the data.***

Some of the questions include:

Q1. What is the booking percentage between Resort Hotel and City Hotel?

Q2. Finding the yearly bookings for each given in the data frame?

Q3. Bookings that were not cancelled.

Q4. Monthly footfall in different types of hotels.

Q5. Nationality of the guests.

Q6. ADR vs Arrival Month.

Q7. ADR vs Arrival Days of Month.

Q8. Which was the most booked accommodation type (Single, Couple, Family)?

Q9. Source of bookings.

Q10. What is the relationship between lead time and cancellation?

Q11. Does market segment have any impact on repeated customer?

Q12. Number of Required car parking space.

Q13. How long people stay in a hotel?

Q14. What are the preferred meal types?

Q15. Total special requests for different types of hotels.

**INTRODUCTION**

In this project we attempt to understand the trends so that we could estimate the best time to book a room at a hotel and the optimal length of stay. we also calculated the factors that affect the monthly footfall of a hotel using the Hotel Booking Dataset provided to us. This data set contained booking information for different City hotels and Resorts and included information such as when the booking was made, length of stay, the number of adults, children, and/or babies, etc. All Personal Identification information were removed from the dataset so as to protect the privacy of the people. We explored and analysed the data to discover the Key Performance Indicators (KPI) that govern the bookings and prove us with the necessary information behind the underlying principles.

**SEGREGATION**

The dataset could be classified in different aspects/ lights with respect to multiple characteristics such as:

* **Types of hotels**
  + Resorts
  + City hotels
* **Number of guests per occupancy**
  + Single
  + Couple
  + Family
* **Modes of booking**
  + Online TA
  + Offline TA
  + Corporate
  + Groups
  + Direct**,** etc
* **Types of meals**
  + BB- Breakfast& Bed
  + HB- Half Board
  + SC- Self Catering
  + FB- Full Board

And many more.

**Libraries we used are**

* *Pandas*

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license

* *Numpy*

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

* *Matplotlib.pyplot*

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

* *Seaborn*

Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

The whole project has been divided into two phases.

**PHASE- I**

**DATA EXPLORATION & DATA PRE-PROCESSING**

**Data Exploration** is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more.

Then we proceed towards **Data Pre-processing** which includes data cleaning, data transformation and data reduction which will enhance the performance and understanding of a data frame.

* **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
* **Data Transformation:** This step is taken in order to transform the data in appropriate forms suitable for mining process.
* **Data Reduction:** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

**PHASE- II**

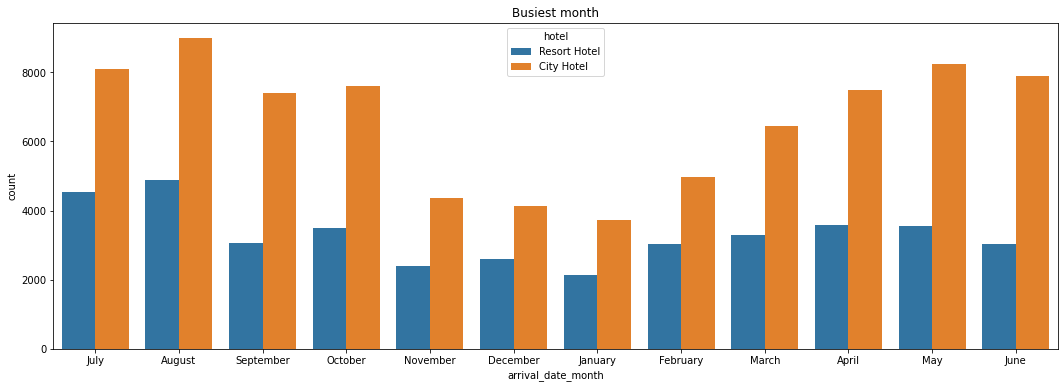
Phase II deals with challenges of simplifying the preprocess data show that it could make sense of what the data represents. This phase strictly includes questionnaire that are to be answered with the help of the newly created processed data for the specific circumstance

The tools that were involved were used for Data Engineering, Data Integration and Data Visualization.

For Data Visualization we use different forms of Charts such as

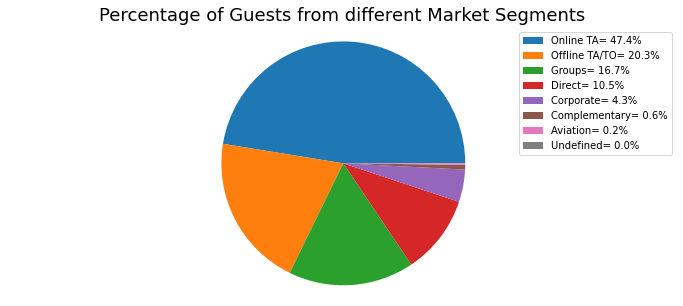
**Bar graphs**

Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars is proportional to the measure of data. For e.g. (This is a bar graph plot for monthly footfall in different types of hotels)



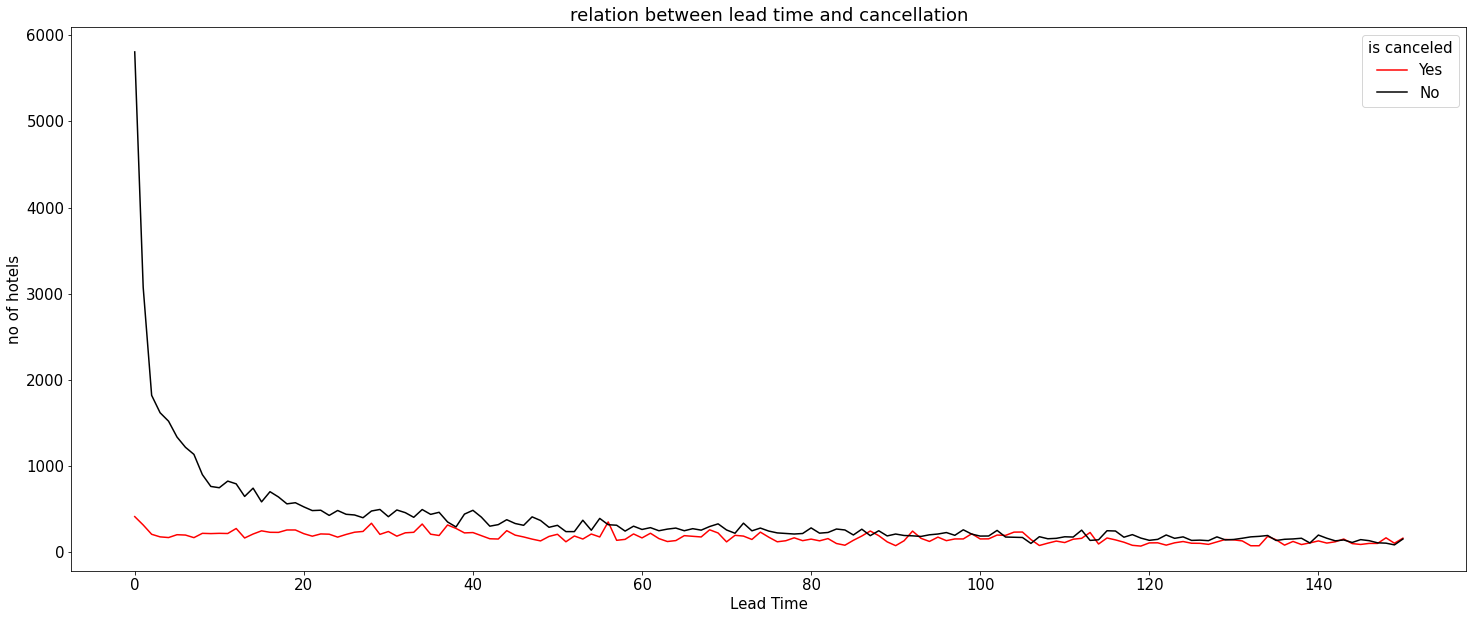
**Pie chart**

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents. For e.g. (This is a pie chart representation for source of `bookings)



**Line chart**

A line chart or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. For e.g. (This line plot represents the relationship between lead time and cancellation?)



**Conclusion**

The research concluded with the creation of data-points for guests to book suitable rooms for themselves by monitoring the trends. This data, being interactive improves the user experience and allows them to use this system as per their convenience. The background research aided in the formulation and validation of the problem statement and helped us create an information architecture.

This study allowed us to acquire insight into the processes that are followed while creating a booking token. Overall, the research accomplished our goals of comprehending how the hotel systems work and how we could avail them at an appropriate price point with suitable amenities. The implementation of the system would help guests better understand the structure involved in booking hotels on a regular basis.