HOTEL BOOKING ANALYSIS

Sushwet Kumar Pandey, Data Science Trainee, AlmaBetter, Bangalore

1.ABSTRACT: This hotel booking dataset contains booking information for city and resort hotel. Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both 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 cancelled. 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.

2.INTRODUCTION: In 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, two 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 cancelled. 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).

3.**PROBLEM STATEMENT:** We are here to explore a hotel booking dataset to discover important factors that govern the bookings, which contain booking information for a city hotel and a resort hotel. We will analyze some important aspects of hotel bookings which will helps us identify major loopholes and give us insights which will be helpful to run profitable hotel business are as follows:

- The time of year to book a hotel room?
- Optimal length of stay to get the best daily rate?
- To predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

4.FEATURE DESCRIPTION: The data feature in this dataset respectively:

- ADR (Numeric) Average Daily Rate as defined.
- Adults (Integer) Number of adults.
- Agent (Categorical) ID of the travel agency that made the booking.
- ArrivalDateDayOfMonth (Integer) Day of the month of the arrival date.
- ArrivalDateMonth (Categorical) Month of arrival date with 12 categories: “January” to “December”.
- arrival_date_week_number (Integer) Week number of year for arrival date.
- arrival_date_year (Integer) Year of arrival date.
- Babies (Integer)number of babies in count.
- Children (Integer) number of children.
- Company (Integer) ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons.
- Country(object) Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- customer_type(categorical) Type of booking, assuming one of four categories: Contract - when the booking has an allotment; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract; Transient-party –

when the booking is transient, but is associated to at least other transient booking.

- distribution_channel(categorical) Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- days_in_waiting_list (Integer) Number of days the booking was in the waiting list before it was confirmed to the customer.
- Hotel(categorical) Hotel (H1 = Resort Hotel or H2 = City Hotel).
- is_canceled (Integer) Value indicating if the booking was canceled (1) or not (0).
- is_repeated_guest (Integer) Value indicating if the booking name was from a repeated guest (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.
- Meal(categorical) Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal.
- market_segment(categorical) Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- previous_cancellations(categorical) Number of previous bookings that were cancelled by the customer prior to the current booking.
- previous_bookings_not_canceled (Integer) Number of previous bookings not cancelled by the customer prior to the current booking.
- reservation_status(categorical) Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out.
- reservation_status_date (Date) Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel.
- 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.

- total_of_special_requests (Integer) Number of special requests made by the customer (e.g., twin bed or high floor).

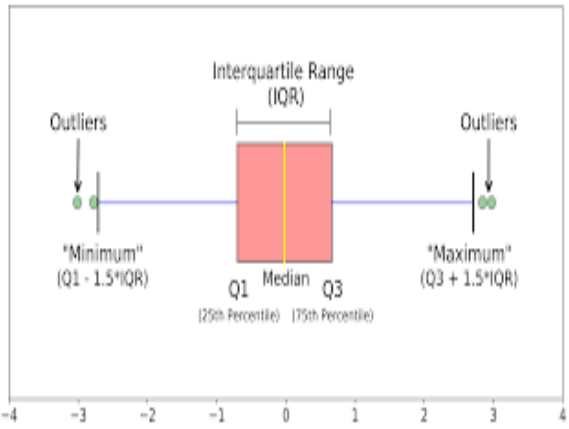
5.EXPLORATORY DATA ANALYSIS:

- **DATA PREPARATION:** Firstly, we imported libraries and dataset, some of the libraries used are NumPy, pandas, matplotlib, seaborn, warnings. Once the data is collected, process of analysis begins. But data has to be translated in an appropriate form. This process is known as Data Preparation.
- **Validate data**
- **Clean the data set**
- **Checking and deleting the duplicate values**
- **Statically adjust the data**
- **Store the data set for analysis**
- **Analyze the data**

MISSING VALUES AND OUTLIER TREATMENT: Used three different concept to treat the missing values and outlier. There are different ways and methods of identifying outliers, but we are only going to use some of the most popular techniques:

- Visualization: by boxplot or histogram plot
- Skewness: The skewness value should be within the range of -1 to 1 for a normal distribution, any major changes from this value may indicate the presence of outliers.
- Interquartile Range: IQR
- Standard Deviation: It shows the variability distribution of the data.
- Flooring or capping
- Trimming

Firstly, we demonstrate and remove the outlier based up on own understanding by setting up the threshold limit. And in terms of outlier, we used IQR, In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the mid spread, middle 50%, or H-spread. It is defined as the difference between the 75th and 25th percentiles of the data.



And lastly, we used quantile-based technique to treat the outlier, Capping is replacing all higher side values exceeding a certain theoretical maximum or upper control limit (UCL) by the UCL value.

DATA PREPROCESSING: A dataset may contain noise, missing values, and inconsistent data, thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

- **CLEANING AND MANIPULATING THE DATASET:**

CLEANING:

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system. Irregularities are of different types of data.

- Missing Values
- Incorrect Format
- Incorrect Headers
- Anomalies

MANIPULATING:

Data Manipulation: Manipulation of data is the process of manipulating or changing information to make it more organized and readable. Made some new features with the help of column present in the datasets .

UNIVARIATE ANALYSIS:

In Univariate Analysis, we choose a single feature from the data and try to determine what the output or the target value is ,i.e., one feature/variable at a time.

- Understand the trends and patterns of data
- Analyze the frequency and other such characteristics of data
- Know the distribution of the variables in the data.
- Visualize the relationship that may exist between different variables.

BIVARIATE ANALYSIS:

In a Bivariate Analysis, we try to analyze two features instead of one, and finally determine the classification of output we are looking for. It is a methodical statistical technique applied to a pair of variables (features/ attributes) of data to determine the empirical relationship between them. In order words, it is meant to determine any concurrent relations. There are three main types of bivariate analysis. They are as follows:

- Scatter Plots - It makes use of dots to represent the values for two different numeric variables.
- Regression Analysis- This involves a wide range of tools that can be utilized to determine just how the data points might be related. It tends to provide us with an equation for the curve/line along with giving us the correlation coefficient.
- Correlation Coefficients - This shows how one particular variable moves about with relation to another.

MULTIVARIATE ANALYSIS:

- Multivariate analysis deals with such complex set of data with more than two feature and variables. There are two types of multivariate analysis techniques: Dependence techniques, which look at cause-and-effect relationships between variables, and interdependence techniques, which explore the structure of a dataset.

CHALLENGES:

- Dealing with such big dataset is quite difficult sometimes , lots of missing values made things some more complicated , defining a function which is used to annotate the histogram percent according to their respective count taken a big notch of this obstacle part .Coming to the visualization part , more or less makes our challenges addresses to code in such a way to visualize the graphs as per rows and columns with fixed figure size to retain as per the subplots.

CONCLUSION:

Our analysis, would be capable of helping prospective guests in choosing the right hotel, right stay duration and much more for their

stay and moreover, would also be introspecting for hotel management in bringing out changes in their services for the guests.

- City Hotel is the most booked hotel with 62 percent not_canceled bookings.
- Resort Hotel has been preferred over City Hotel by larger group of guests or families.
- One out of every three bookings are cancelled.
- Direct bookings have very less cancellation%.
- Most preferred meal is BB (Bread and Breakfast.
- Online marketing is the best way to attract customers.
- People do not want to pre-deposit the money for booking.
- Only 10% of people require parking space.
- Resort hotel is preferred mostly for longer stay, day time stays. and when the parking space is needed.
- More than 15 days advance bookings have high chances of cancellation.
- Assigning different room is not a reason for cancellation.

As for the prediction of cancellations concerns. it is clear that better results can be achieved in a more exhaustive machine learning process, that includes more models into consideration. Besides, this data is somewhat limited (only two years). A wider time window and more features, which sure will be at the hands of every hotelier in the business, better results could be obtained.