**Hotel Booking Analysis**

**Manisha Dhanuka**

**Data science trainee,**

**AlmaBetter, Bangalore**

**1. Abstract:**

The hotel industry is one of the most important components of the wider service industry, catering for customers who require overnight accommodation. It is closely associated with the travel industry and the hospitality industry. In this project, I explored data related to bookings of two hotels: Resort Hotel and City Hotel.

**2. Problem Statement:**

The data consist of the booking related information, we try to explore the possible reasons for the cancellation and Average Daily rate of the rooms.

**Features:**

1. Hotel – City Hotel or Resort hotel

2. Arrival Information: Date, month, year and day

3. Market segment and distribution channel

4. Native country of the guests

5. Number of guests: Number of adults, babies, children

6. Stays in weekend nights, stays in week days

7. Cancellation status, previous cancellation and booking changes.

8. ADR, Deposit type

9. Services: Meal, Room type, car- parking space, special requests

10. Lead time

**3. Introduction:**

Hotel industry is the most investment and human resource dependent industry. With so much at stake and the dependency on the ‘Tourism Industry’, which itself is unpredictable, any decision should be on the data and not only on the intuition. The recent example of complete shutdown due to covid has proven the importance of analyzing things first. In this project I tried to find the reasons of the cancellation and how the ADR varies so that Cancellation policy and the price policy could be made taking into those things in consideration.

**4. What is Exploratory Data**

**analysis?**

*“Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.”*

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

**5. Tools Used:**

**Pandas:** Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source data analysis / manipulation tool available in any language. It is already well on its way towards this goal. The main data object we worked with in pandas are Series and DataFrame.

* **Series:** A Pandas Series is like a column in a table. It is a one-dimensional array holding data of any type.
* **DataFrame:** It is like table consisting of labelled rows and columns.

**Various in-built functions/methods of pandas used in the project are:**

***1. read\_csv():*** It is the inbuilt function of pandas to read csv (comma separated variable file) to the python. Pandas not only allow us to read csv file only, but can read other types like tsv, excel etc. Moreover, we can convert the results into csv also using to\_csv.

***2. info():*** This method allows us to preview the DataFrame, about the columns in the dataset, their data-type, non-null values of the data, the storage and many more things.

***3. head():*** It prints the starting rows of the data. It takes n i.e., number of rows you want to print as the input and by default it prints top 5 rows.

***4. tail():*** It prints the last rows of the data. It takes n i.e., number of rows you want to print as the input and by default it prints last 5 rows.

***5. describe():*** This method gives the statistical description of the data. It mentions: count, mean, std, min, max, quartiles of the numerical data and also, top, frequency, count of the categorical data if mentioned ’all’ in include parameter.

***6. transpose():*** This method interchanges row and column indices.

***7. corr():*** It displays the correlation coefficient between the numerical features of the data. If there is any feature whose value doesn’t change for any instance, it gives Nan there. By default, it displays Pearson correlation coefficient.

***8. isnull():*** It returns Boolean output as True or False depending it is NaN value or not. The data modelling doesn’t give better results if there are null values present in the data. Thus, finding them and get rid of them is a necessary thing.

***9. isduplicated():*** This also gives the True or False Boolean results if the particular instance is duplicated or not. We can give the subset as the parameter, but by default it checks if all the feature of the given rows is same or not.

***10. dropna():*** To drop the null values.

***11. fillna():*** To impute the null values.

***12. drop\_duplicates():*** To drop the duplicate values.

***13. value\_counts():*** It gives the counts of the unique values in the given series.

***14. groupby():*** It groups the data on the basis of list of columns passed as argument and then performs the aggregate functions mentioned and displays the result.

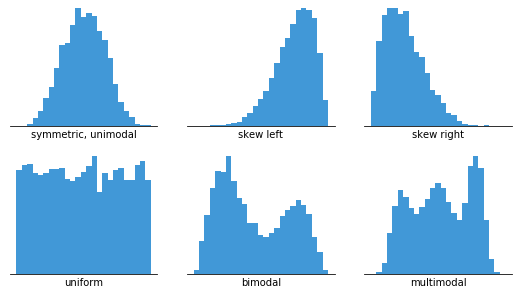
***15. pivot\_table():*** This is similar to the pivot tables we use in the Excel for better display of the results and analysis.

**Matplotlib:** Visualization is crucial to quickly understanding trends and relationships in the data. Matplotlib is one of the most popular plotting libraries for plotting with Python. Many other visualization libraries like pandas built-in visualization and seaborn are built directly off matplotlib.

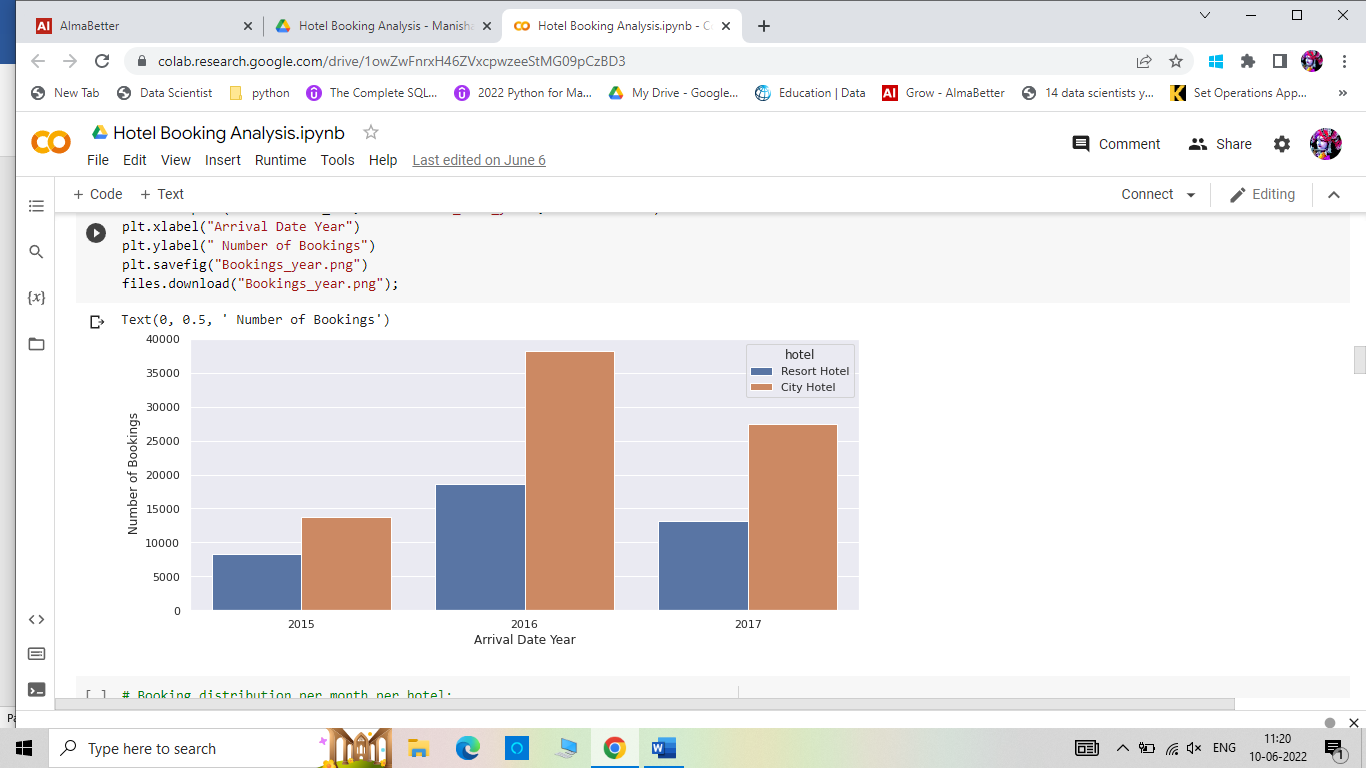
**Seaborn:** Statistical plotting library built-off on matplotlib but uses a simpler “one-line” syntax. But that comes off at a tradeoff heavy customization. We have to use matplotlib for that purpose.

**Various Plots used in our project are:**

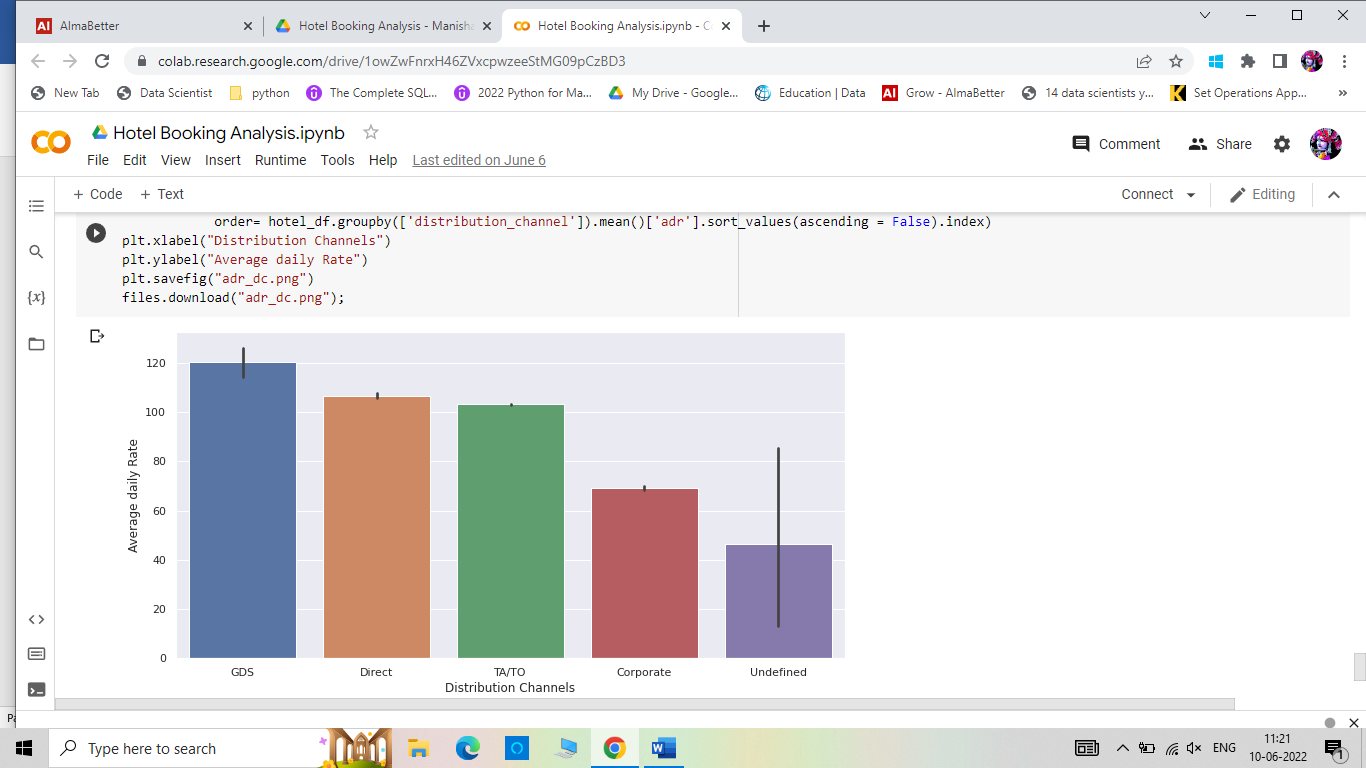
***1.******Histogram (Distribution plot)*** *:* Displays a single continuous feature and help visualize properties such as deviation and average values. It makes bins from the continuous values and display the bars based on the frequency of the bin.



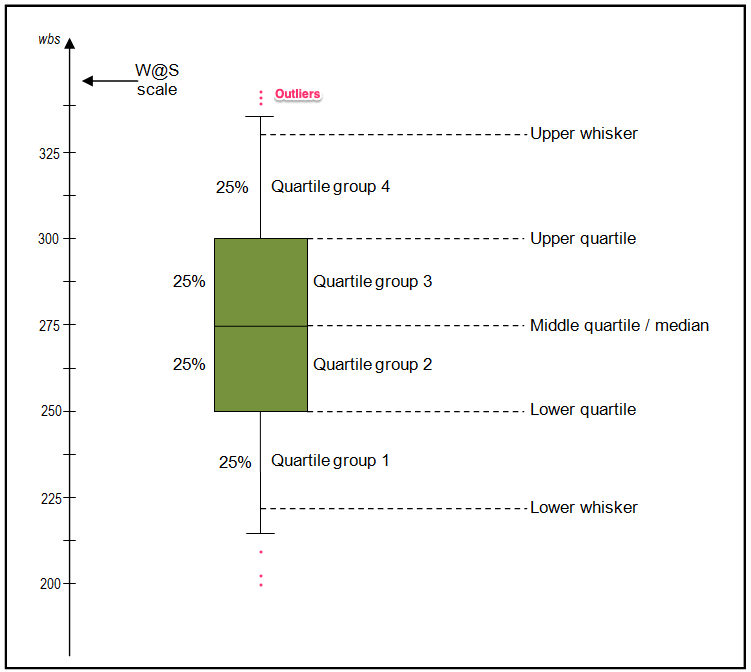
***2. Count plot:*** Counts number of rows per category. Thus, a kind of categorical plot.



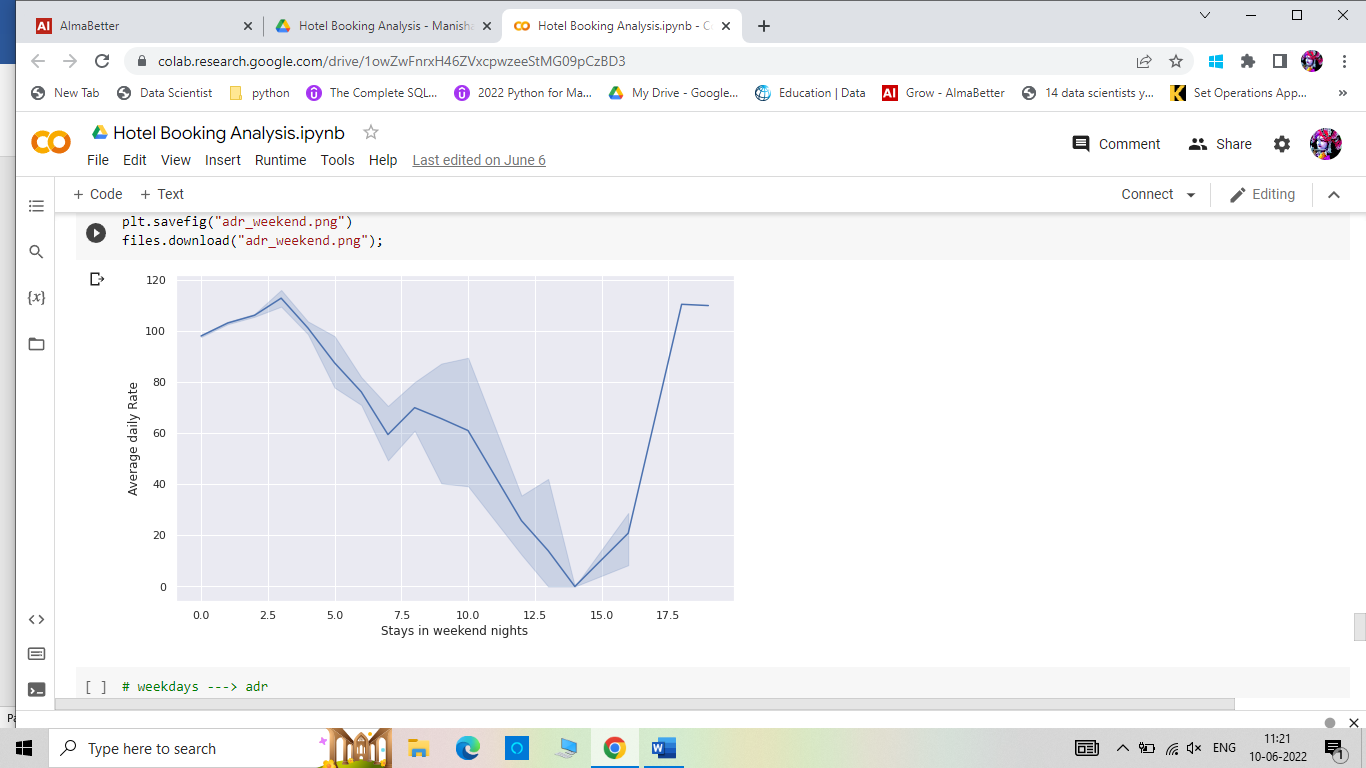
***3. Bar plot:*** Displays any chosen metric ( mean, standard-deviation) per category. Along with that it also displays confidence interval.



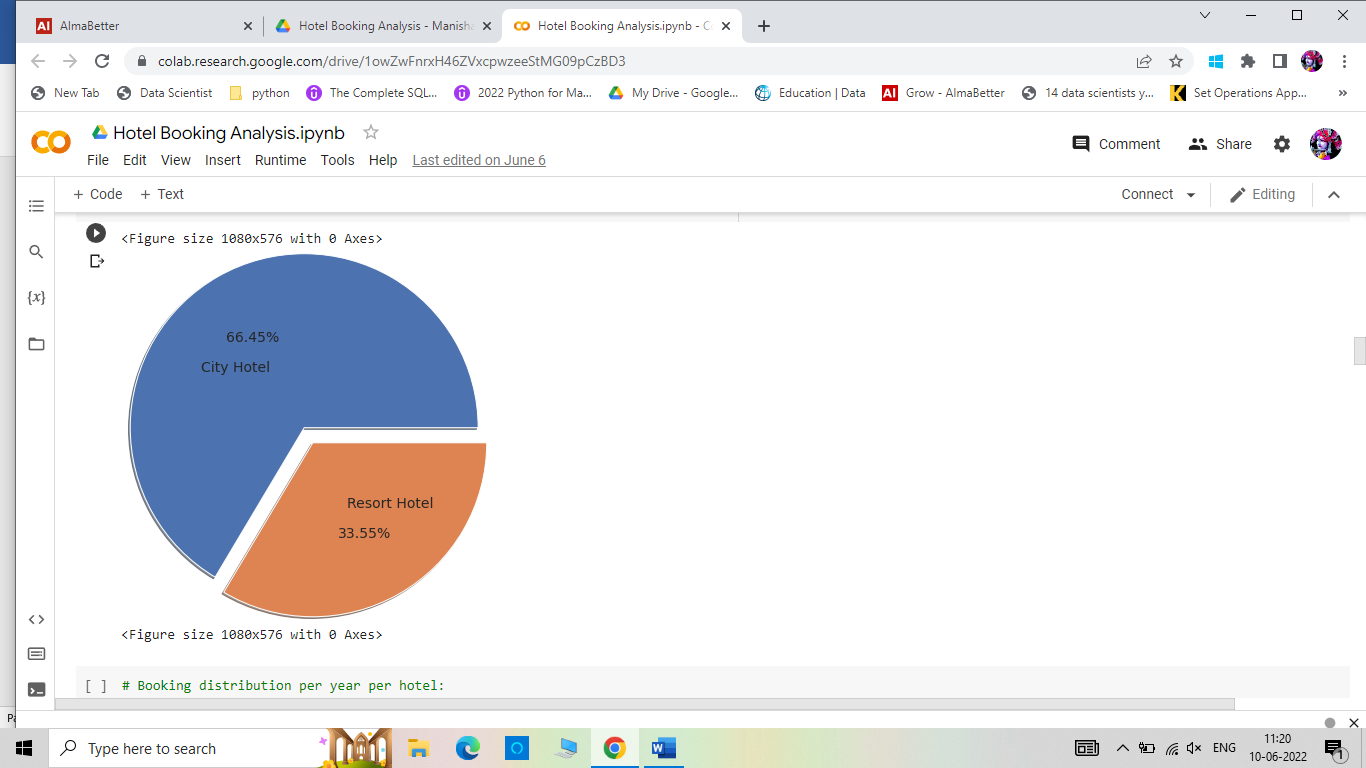
***4. Box plot***: It helps us to visualize distribution of continuous variable within categories. The box represents the quartiles Q1 and Q3. The middle line represents the median and the whiskers are displayed at extremes of Inter-quartile range. Thus, this plot helps in finding the outliers in the data.



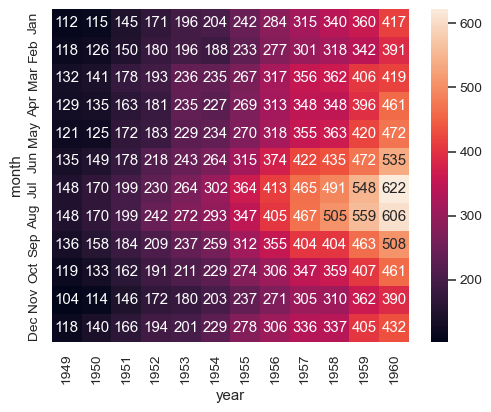
***5. Line plot:*** Displays continuous value variable in the form of line chart. It is the best chart for time- series analysis.



***6. Pie plot:*** Displays the category distribution as the area of the circle.



***7. Heatmap:*** A kind of matrix plot. It is visual equivalent of a pivot table.



**Steps involved:**

***1. Importing the packages:*** First we import the libraries like NumPy, pandas, matplotlib, seaborn and give them aliases.

***2. Read the data:*** Since our project’s data was a csv file, we used pd.read\_csv() method to load the data from the drive.

***3. Preview the Data:*** columns names, non-null values , data-types were previewed using .info(), .head() and .tail() method. Along with used .describe() method for statistical summary.

***4. Correlation check:*** Then calculated the Pearson’s coefficient of correlation and visualize it through heatmap. Most of the variables were not that much related.

***5. Data Cleaning:*** Null values were checked and then imputed and also checked for the duplicates. But as the personal information was omitted and no unique identifier was there, we decided to go along with the duplicates.

***6. Univariate Analysis and Bivariate Analysis:*** Used categorical plots , distribution plots , pivot tables, groupby to analyze the data.

***7. Cancellation possible reasons:*** For finding the possible reasons of cancellation reason, we worked upon these questions:

* Are number of days in waiting list is linked to cancellation or not?
* If most of the cancellations were from repeated guests or the new guests ?
* The number of tourists who have previously cancelled the bookings too
* Tourist Cancellation with respect to deposit type.

***8. ADR Analysis:*** Hotel average daily rate (or ADR) is a calculation for how much money hotel brings in per day in room rentals. This is not a comprehensive calculation; ADR only accounts for the amount the hotel is making per day in room bookings. It does not consider earnings from ancillary revenue such as extra amenities and services hotel offers — add-ons such as spa treatments, pool access, room service, vending machines, etc.

ADR is the basis for understanding the average rate of a stay at your hotel. It’s important to consistently and periodically calculate ADR, as it will change with the seasons. ADR is affected by factors such as time of year, location, and local events guests stay to attend.

To calculate ADR, divide the room revenue by the number of rooms sold. Thus, if the room revenue is not increasing but ADR is increasing, then it is a concern for the hotel as it points towards low occupancy.

For analysis of ADR, we checked for:

* Do bookings on weekends vs on week days affect the revenue
* How ADR is varying over the years.
* ADR with respect to various room types
* Contribution of various market segments to ADR
* Contribution of various distribution channels to ADR

**Conclusion:**

ADR is increasing by year and it is more for weekend nights than the weekdays. Also, Online Travel Agents are contributing more to ADR. Refund deposit type is not the concern for cancellation.