**Hotel Booking Analysis**

**Mohan Vishe, Shambhuraj Desai,**

**Gaurav Yadav, Rahul Ray**

**Data science trainees**

**GitHub link-**

[**https://github.com/MohanVishe/Data-Analysis---Hotel-Booking-**](https://github.com/MohanVishe/Data-Analysis---Hotel-Booking-)

**Abstract:**

The dataset contains information about hotel bookings in the year 2016 to 2018. With the help of EDA, we will try to insight into the data, and in this process, we will take help from various libraries such as pandas , matplotlib.pyplot

Numpy, and seaborn for better visualization.

**1. Problem Statement**

The data contains information about the booking of the Hotels, the cancellation, the adults in the customers, the children, the type of booking, services they demand, the country they demand, arrival date, and the things.

We need to explore the dataset and find the things which affect the bookings relatively.

There are so many features present in the dataset which contain various keywords. There are some important keywords in the dataset.

1. hotel: type of hotels

2. is\_canceled: canceled or not

3. lead\_time: no. of days before actual arrival in the hotel

4. arrival\_date\_year: year of booking

5. arrival\_date\_month: month of booking

6. arrival\_date\_week\_number: week number of the year in which booking 7.arrival\_date\_day\_of\_month: arrival month date

8. stays\_in\_weekend\_nights: no. of weekends guest stayed

9. stays\_in\_week\_nights: no. of weekdays guest stayed

10. meal: BB – Bed & Breakfast HB – only two meals including breakfast meal FB – breakfast, lunch & dinner.

11. market\_segment: TA: Travel agents too: Tour operators

12. previous\_cancellations: cancellation in past

13. previous\_bookings\_not\_canceled: not cancelled in the past.

14. reservation\_status: "Canceled" for booking was cancelled, "Check-Out" for the customer who has checked in but departed, "No-Show" for the customer who did not check-in and did inform the hotel of the reason why

15 .customer\_type: "Group" when the booking is associated with a group, "Transient" when the booking is not part of a group

"Contract" when the booking has an allotment/another type of contract

"Transient-party" when the booking is transient but is associated with at least other transient bookings

**2. Introduction**

Whenever we think about hotel booking the words that come To mind are money required and my budget, location, services, cleanliness, and facilities. Feels like this are more important than any other while sorting the hotels. Some things are adjustable according to price and some features can be rewarded such as free wifi and breakfast.

The Hotel Booking Dataset gives us information about whether the booking was cancelled or not. There are two types of hotels called city hotels and resort hotels which contain booking date and time, number of adults, children, and babies, duration of stay, reservation type, etc

Customers search for many things while booking the resort or hotel like the facilities they have, meal quality, room prices, assigning the class of rooms, availability of parking for the cars, etc.

The main objective of an analysis of this data is to find out is: on which period of the year we should book the slots, in which time they consider high demand and also ask for more money, whether they provide the best services or not, the quality and type of food they give us.

Much more things were answered in the analysis.

Take a look at the short map of our journey..!

i) Loading the dataset into the data frame.

ii) Cleaning the dataset and prioritizing our main objectives.

iii) Using some techniques we set our data.

iv) Using some techniques we do statistical analysis.

v) Conclusion.

**3. Steps Involved:**

**i) Loading the dataset**: There are 4 libraries used which are NumPy, pandas,matplotlib.pyplot, and seaborn. Pandas is a great library for EDA. Here Data is in '.csv' format .We use the pandas library to load the dataset into the notebook, after mounting the drive and putting the path of the dataset we then read the data." pandas.read\_csv() to convert it into a data frame. This method takes the path of a CSV file. To find the number of rows and columns of the data we use ".shape" where ".info()" gives information about columns. We have the data set of shape(119390,32) i.e., we have 119390 rows and 32 different columns. After loading the dataset we have to do some cleaning.

**ii) Cleaning the dataset:**

The cleaning process involves removing the dataset values which are unnecessary for the objectives, also it includes some columns which have some null values. We transform the dataset into a consistent format to ensure we predict some good results.

Null values are the data that cannot provide any information and create obstacles to reach out the goal. Data Pre-processing is important before doing EDA as there are many outliers and missing values present in the dataset. We use ".isna().sum()" to find missing values in the dataset. The dataset contains missing values in a few columns. We remove columns that are less important and have a large number of missing values by using ".drop()"

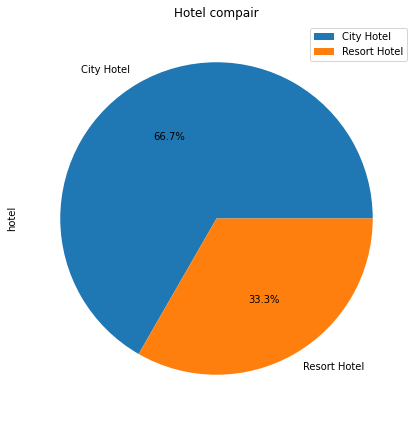
To save data there are some NaN values that are replaced by mean, median, or zero. To replace these values we use ". fillna()". There are some values converted into integers where required.

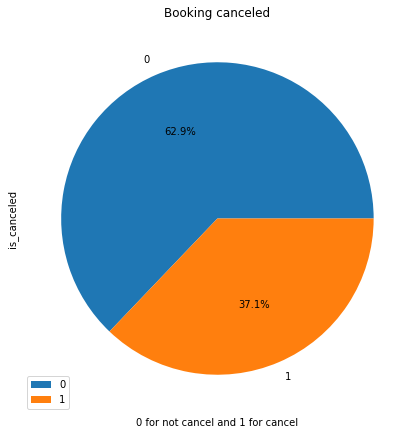
**4. Exploratory Data Analysis:**

**1. Compare Hotels:**

To compare hotels we have to access the data of the feature "hotel". For visualization, we used the pie chart for which we use ".plot .pie()". The figure size is set by "rcparams[ ]".also compared the booking cancelled and not cancelled. City hotels have a higher number of bookings than a resort hotel

The cancel rate is high we can see that more than 1/3 of bookings were canceled which is not good.



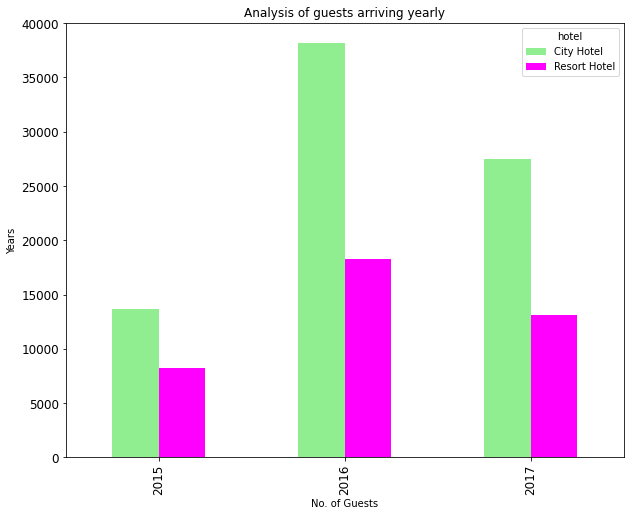


2. The annual number and percentage of arriving visitors.

For this, we have to access the data only for arrival\_date\_year. By using the bar chart we can explain the exact value of arriving the customers at the different years.

We can clearly say that city hotel have the highest booking over the three years than Resort Hotels.

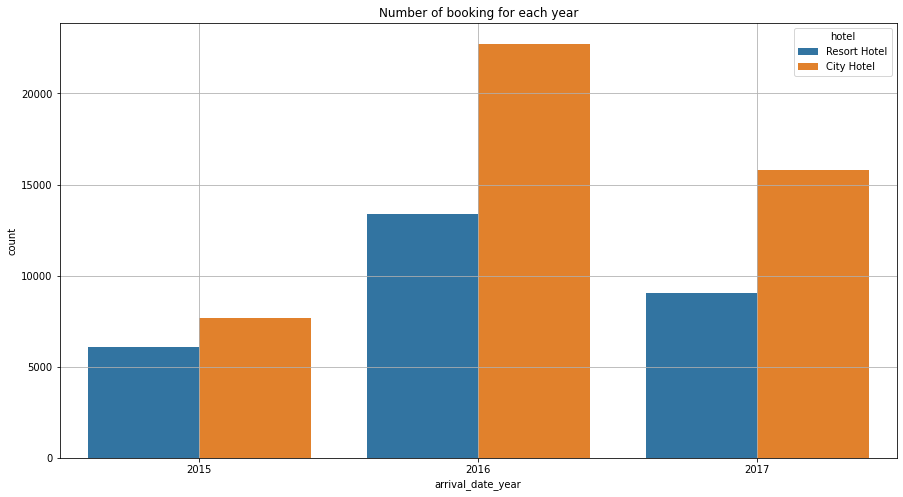
Also, we prove that year 2016 had more customers than 2015 and 2017.



3. Booking for each year separated by hotel type Where an order is not canceled

We create a dataset that contains values where a booking is not canceled

for 2015 the booking was less than 10000 whereas for the year 2016 bookings were at pick crossing 22500. There is a higher jump in city hotel booking as compared to a resort hotel.



4. People stay in a particular hotel:

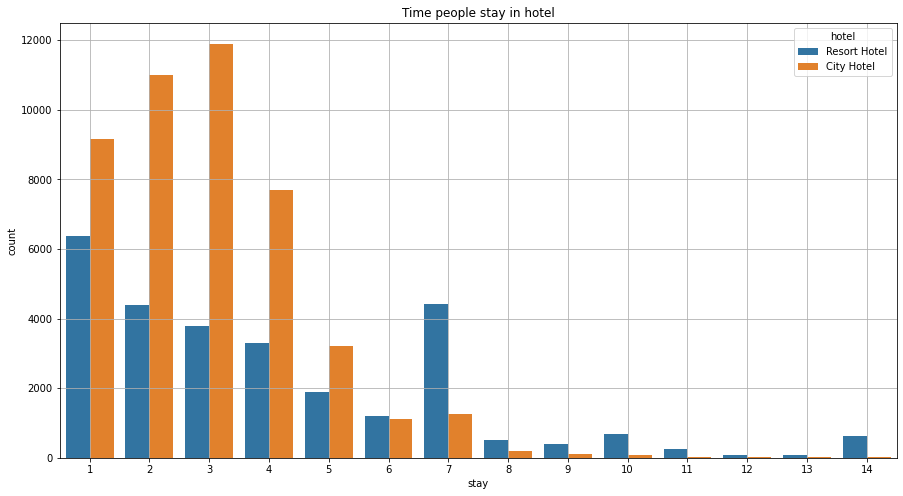
Combine the stay\_in \_weekend\_nights and stays\_in \_week\_nights to form

a stay. To compare the bar plots of both hotels we use "hue='hotel' ".

City hotel shows pick for 3 days stay which is around 12000 and the maximum number of people stay below 5 days

Resort hotels show a decrease in number as days increase but for a week or 2-week stay, there is a considerable increase in count.

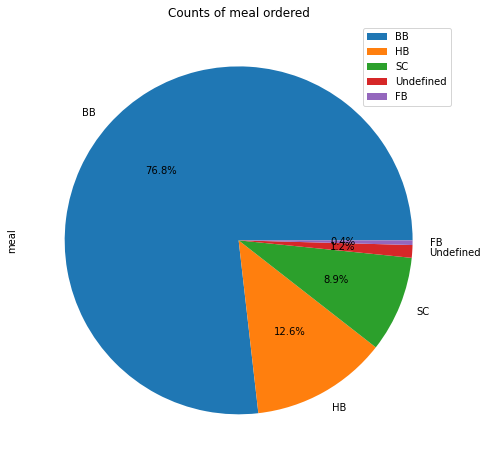
For day 1 Resort hotels show pick which is greater than 6000.



5. Compare meal bookings:

We used a pie chart for comparing meals and to compare the hotels we used a bar plot. Around 76.8% of people Book (BB) means Bed and Breakfast

HB(Half board) and SC Booking are 12.6% and 8.9% respectively and FC(Full Board) Booking are negligible



6. Correlation Heatmap:

A heatmap is a data visualization technique that is used to find the correlation between the two variables and whether they affecting on the dataset. By using the Seaborn library we can create the Correlation heatmap. Here we use sns.heatmap(“a”,annot=True/False) to find out the correlation between the features.



The heatmap shows a positive correlation, negative correlation, and No-correlation.

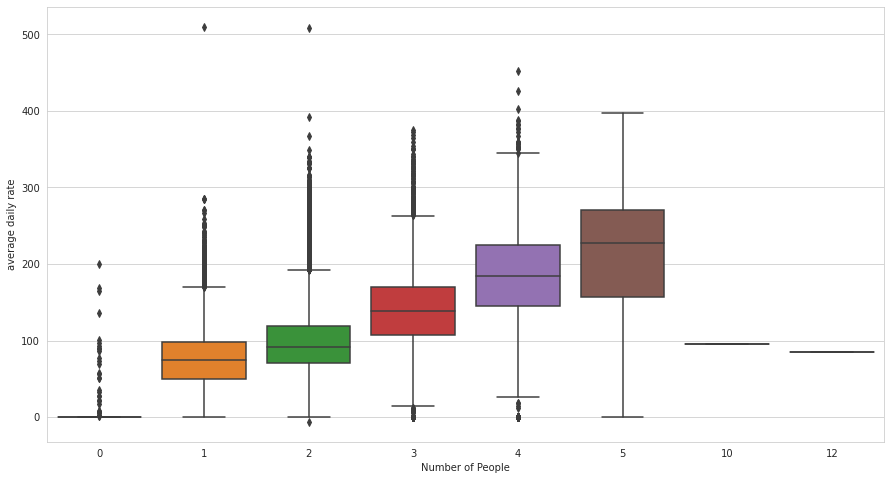
A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases or one variable decreases while the other decreases.

A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. A zero correlation exists when there is no relationship between two variables. By plotting the seaborn heatmap correlation we got to know that there are three features in the given hotel booking data set which are highly correlated with each other.

7. ADR vs the number of people per booking:

In the following method of creating the box-plot visualization, we can exactly find out the main reason for increasing the ADR(Average Daily Rate) as the number of customers increases to the resort or hotels.

By using the seaborn and the command sns.boxplot(x=’ ‘,y=’ ‘, data = ) we create the following chart.



We made some insights from it,

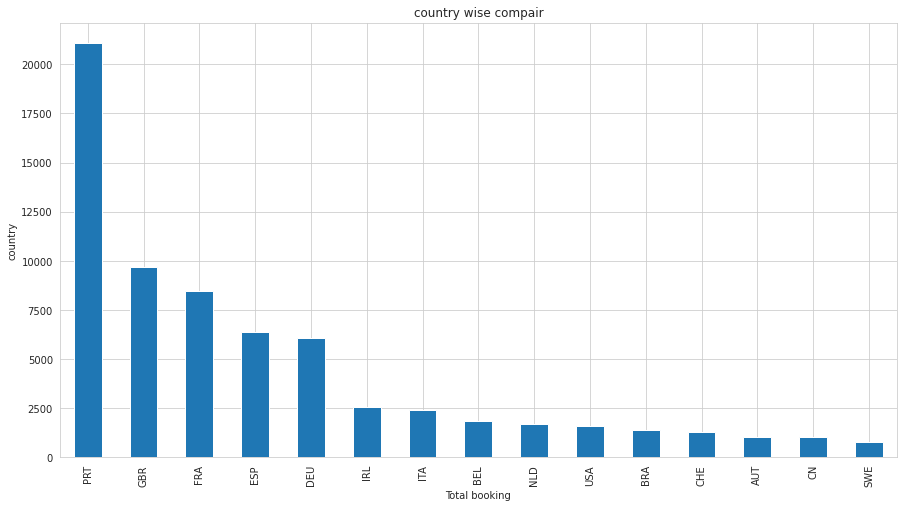
i) Median value of ADR is increasing as the number of customers increases

ii) We clearly see the outliers for the first 4 bars i.e., if the customers are arriving at the rate of more than 5 daily, ADR found normally distributed data.

'ADR' is correlated with people as the number of people increases earnings increase and so does ADR.

Total\_stay is correlated with lead\_time means longer stays have a plan for a long time before arrival

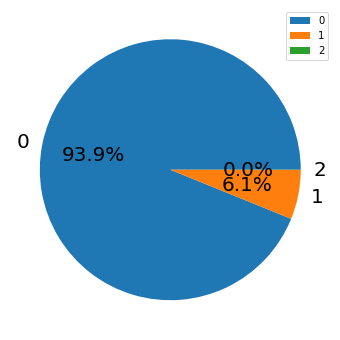
8. See the area from which costumer come to the hotel: We took only the top 15 values for analysis and plot them. The country "PRI" has top booking above 20000. Where "GBR" and "FRA" are in second and 3rd position.



9. car parking space

We can say that more than 93% of customers do not prefer parking

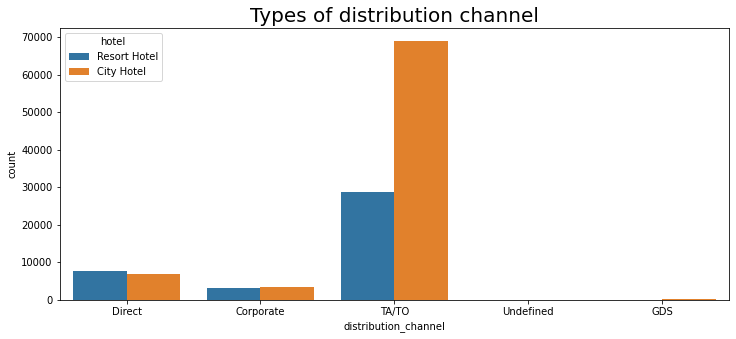
so we can commit that most of the customers prefer public transport



10. Sector should be targeted for our advertisement

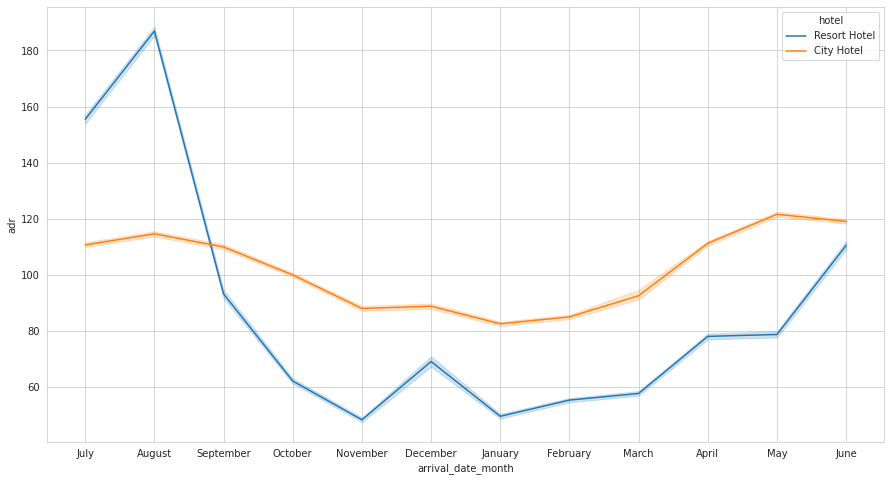
From the above, we see that the majority of the distribution channels and market segments involve travel agencies (online or offline).

We have to focus on targeting these travel agencies' websites and work with them since the majority of the visitors tend to reach out to them.



11. Months Vs Booking rates

We used a line plot to see the change in rates per month. For both city and resort hotels, Nov to Jan has cheaper average monthly rates. Hotels should focus on the months of summer as there is more rate in that time.



**8. Conclusion:**

A. There is less number of features with missing values and outliers so dealing with them is easy.

B. After the peak in 2016, there is a fall in bookings in the year 2017. After 2015 bookings increased in 2016 then decreased in 2017.

C. City hotels have a higher number of bookings so we have to focus on them and May to Aug period is targeted as there is the peak of the summer period.

D. 1/3 of bookings were canceled Which is worse.

E. As the booking rate for countries PRT, GBR, and FRA is high so we have to focus on advertisement in that country.

F. As the average daily rate increases with an increase in the number of people so we have to encourage a large group of people/families to book by giving offers and discounts.

G. Hotels should focus on the months of summer as there is more rate in that time.

H. We can say that more than 93% of customers do not prefer the parking, so we can commit that most of the customers reach the hotel by transport.

I. The hotel should be near stops of public transport

J. We should give less space for parking

K. From the above, we see that the majority of the distribution channels and market segments involve travel agencies (online or offline). We have to focus on targeting these travel agencies' websites and work with them since the majority of the visitors tend to reach out to them.

L. For both city and resort hotels, November to January has a cheaper average monthly rate.

**References-**

1. pandas.pydata.org
2. GeeksforGeeks
3. Seaborn.pydata.org
4. Matplotlib.org