Hotel Booking Analysis

Team Mate - Chetan Patil, Mohit Jain, Siddharth D choury, Rajesh Patil,

Alma Better Capstone Project

1. Abstract:

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, location, seasonality, days of week and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series. we have done brainstorming to find out insightful from the data by using various technique such as programming i.e. Python (Pandas , Numpy, etc.), data visualization by python libraries like seaborn, matplotlib etc.

2. Problem Statement:

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 analyse the data to discover important factors that govern the bookings.

3. Data Summary:

The main objective of this project is to find out which type of hotel is preferred by visitor with their special requirement. While analyzing we found out that in total the data set contain below rows and column.

1. The dataset has a shape of (119390, 32) which indicates that it contains approximately 1.2 lakh rows and 32 columns.
2. The Dataset has 4 columns with float64 dtype, 16 columns with int64 dtype, and 12 columns with object dtype.
3. In the Dataset, we observed null values in the following columns:
    a. 4 null values in the children column
    b. 488 null values in the country column
    c. 16,340 null values in the agent column
    d. 112,593 null values in the company column

- While analyzing we came to know that maximum Null values are present at the columns agent and company & some null values in children and Country column. So we dropped company column as it has 94.3 % values as Null. We will replace null values of other column i.e. agent, Children, Country with feature Engineering.
- Agent Column has around 13.69 % data as Null value and agent data indicates that data is of agent id so we will enrich/fill null values with **'0'** as no agent involve for booking hotel**.**. Country Column as there is total 195 contraries in the world. we will replace null values of country column to '**other**'

We have the following column names provided to us in the dataset.

1. hotel : Defines the Type of the Hotels in the Given Data Set of Hotel Booking Analysis (Resort hotel and City hotel)

2. is_canceled: Defines the Status of the Booking (Ex: Cancelled)1 refers to Canceled and 0 suggests Not Canceled

3. lead_time: Gives us the timing difference between the booking Time and the arrival from the given data set

4. arrival_date_year : Represents the Year of Arrival of the Visitor (2015, 2016, 2017)

5. arrival_date_month: Represents the month of Guest (Visitors) Arrival From Jan To Dec

6. arrival_date_week_number: This represents the Week No. of The Visitors Arrival - 1 to 53

7. arrival_date_day_of_month: This gives the day number of month when the visitor arrived - 1 to 31

8. stays_in_weekend_nights: This gives the number of weekend nights, i.e. Saturday and Sunday

9. stays_in_week_nights: This gives the number of week nights, i.e. Monday to Friday

10. adults: This gives the number of adults per booking

11. children: This gives the number of children per booking

12. babies: This gives the number of babies per booking

13. meal: This gives the type of meal preferred.

Undefined/SC means no meal package, BB means Bed & Breakfast, HB means Half board (i.e., breakfast & one other meal – usually
dinner), FB means Full board (i.e., breakfast, lunch & dinner)

14. country: This gives the country of origin of visitor

15. market_segment: This gives the group of people based on market

Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Aviation Where, TA: Travel Agents, TO: Tour Operators

16. distribution_channel: This mentions the type of distribution channel

Direct, Corporate, TA/TO, Undefined, GDS Features (cont.):

17. is_repeated_guest: This shows repeated customers 1 means repeated customer, 0 means not repeated

18. previous_cancellations: Represents the number of previous bookings that were cancelled by the customer before the current booking

19. previous_bookings_not_canceled: Represents the number of previous bookings not cancelled by the customer prior to the current booking

20. reserved_room_type: Represents the type of room reserved 'C', 'A', 'D', 'E', 'G', 'F', 'H',

'L', 'P', 'B'

21. assigned_room_type: Represents type of room whose possession is given at the time of arrival. 'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'

22. booking_changes: Represents the number of bookings changed

23. deposit_type: Represents the types of deposit No Deposit, Non Refund, Refundable

24. agent: Represents the Agent Id

25. company: Represents the Company Id

26. day_in_waiting_list: Represents the Number of days the booking was in the waiting list before confirmation

27. customer_type: This Gives us Type of customer Contract, Group, Transient, Transient-party

28. adr: means average daily rate

29. required_car_parking_spaces: Number of car parking spaces required by the customer

30. total_of_special_requests: Number of special requests made by the customer

31. reservation_status: Status of reservation Canceled, Check-Out, No-Show

32. reservation_status_date: Date at which the last status was updated

4. Steps involved in the Data Analysis: -

1. **Data Wrangling:** After loading the dataset, we performed this method by cleaning, organizing, and transforming raw data into the desired format which makes us to understand the data clearly. This process helped us to tackle the unwanted data, to produce accurate results, to make better decision.
2. **Null Value Treatment:** Our data set contains a small number of null values. we have used various method to fill null values by feature engineering.
3. **EDA analysis:** Exploratory Data Analysis is one of the most efficient methods used to Analyse the given data sets. After loading the dataset, we performed many methods by comparing our target variable that is booking analysis with other independent variables. Using the exploratory data analysis, we can summarize the characteristics of the data sets which are important, and in this EDA Capstone Project we have made use of the statistical graphs and the other visualization methods such as.
    1. Bar Plot
    2. Line Plot
    3. Count Plot
    4. Pie Chart
    5. Histogram
    6. Heatmap

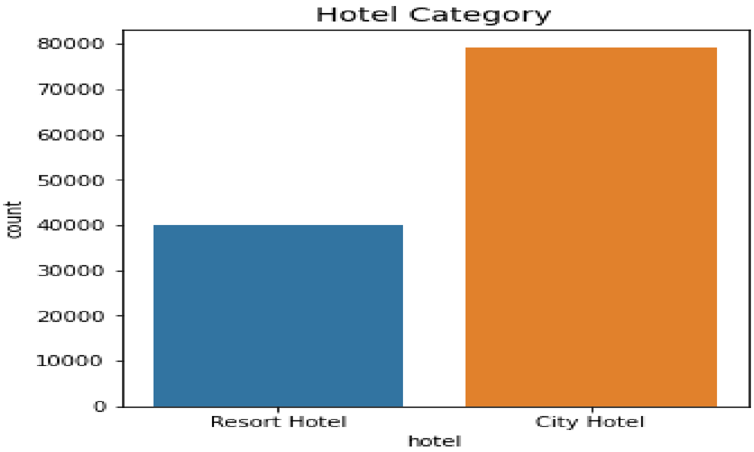
4. **Visualization**: Visualization is one of the most efficient methods of data representation in readable plots and very easy to interpreted.

5. Let’s see the analysis by using data visualization

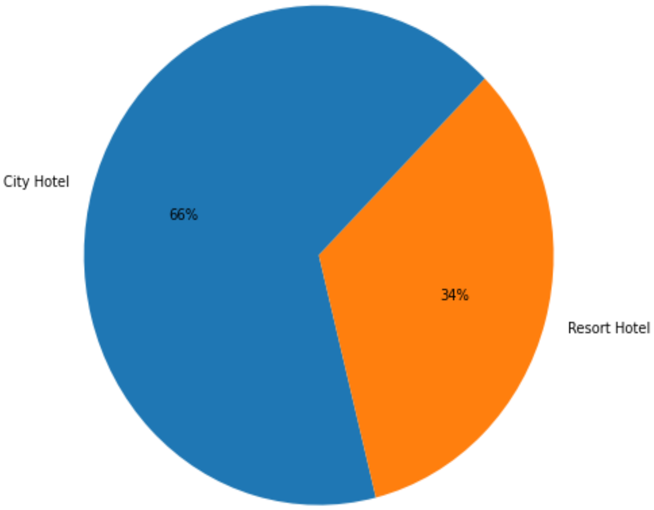
I. Count Plot by Hotel Category.

- Data representing which type of hotel Visitors prefer.

1. Bar Plot

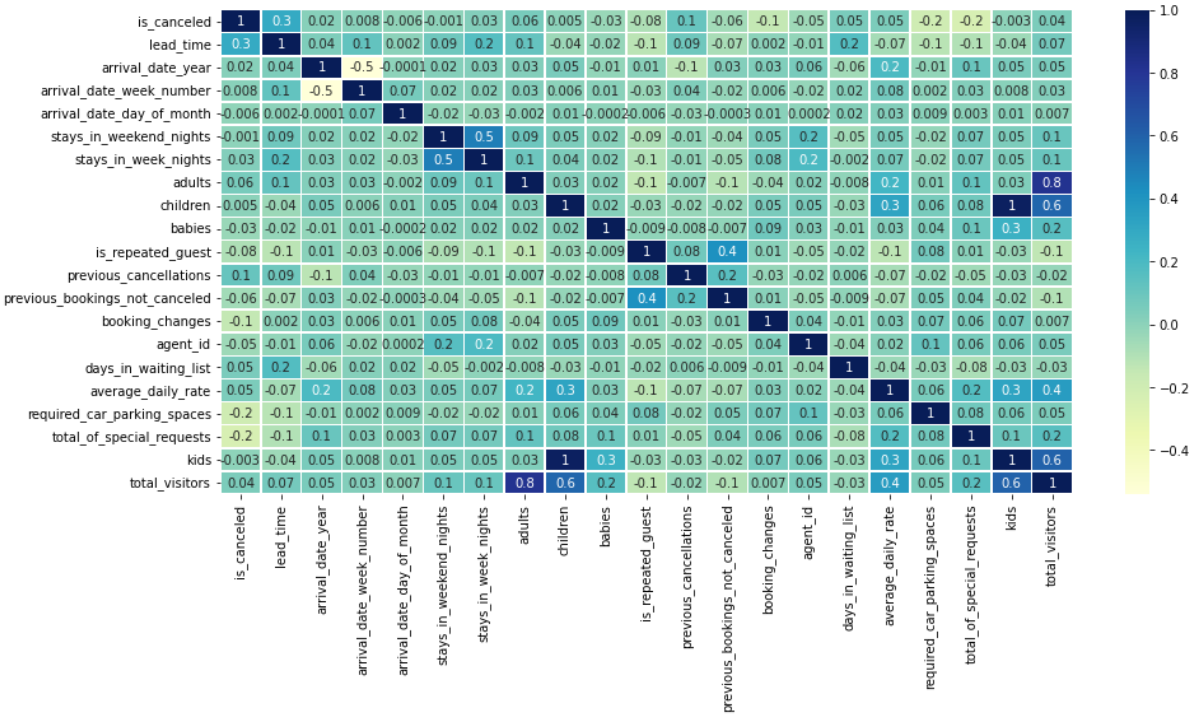


2. Pie Chart.



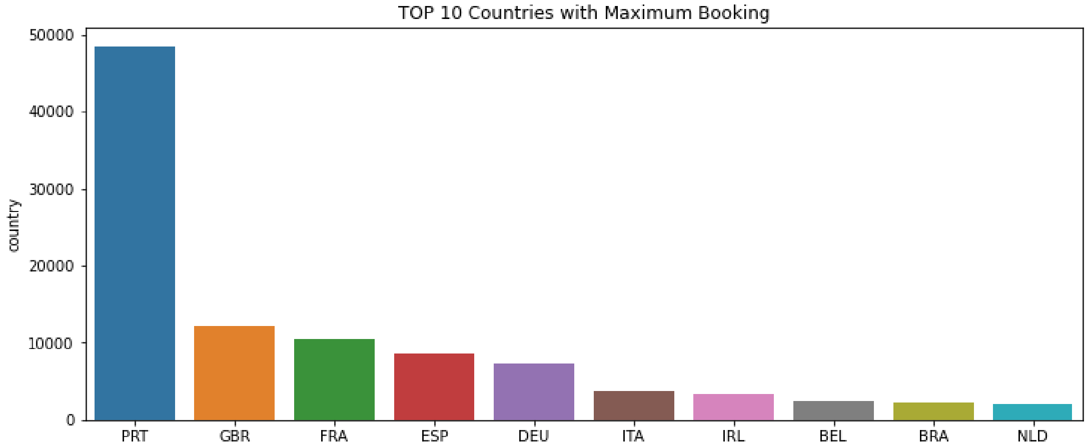
Interpretation-The Above Chart Represents the Type of the Hotel Preferred by the Visitors, after analysing the given data set, we can conclude that 66% of the visitors prefer the City Hotel as compared to 34% of the Visitors who are inclined towards the Resort Hotel.

II. Heat Map Showing the Correlation Between Given Data Sets



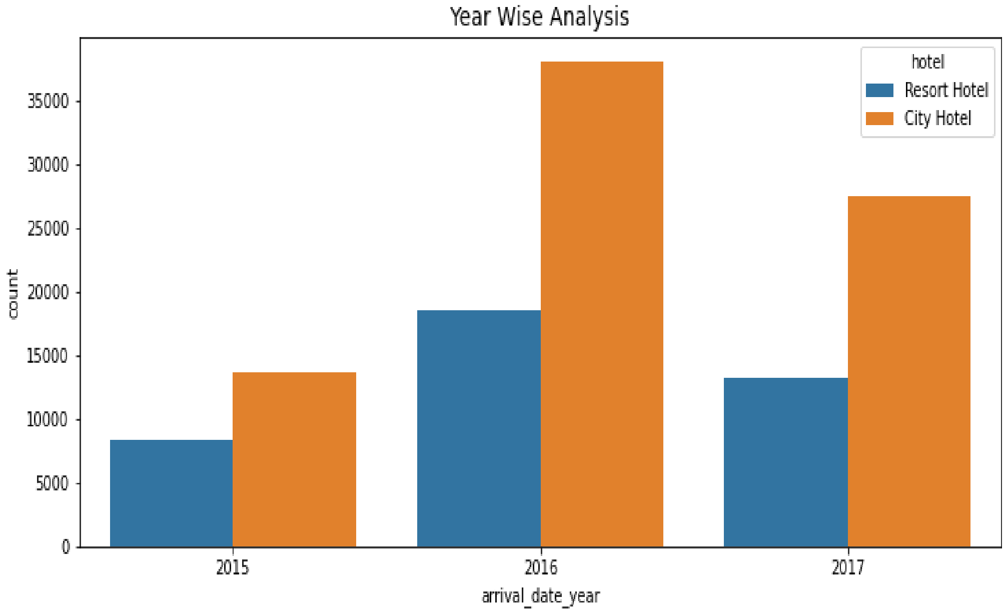
Interpretation: - As per the Heat map it is clear that there is no such strong correlation between each variable of data set. All data set variable is independent of each other.

III. Data Representing the Countries Booking Status



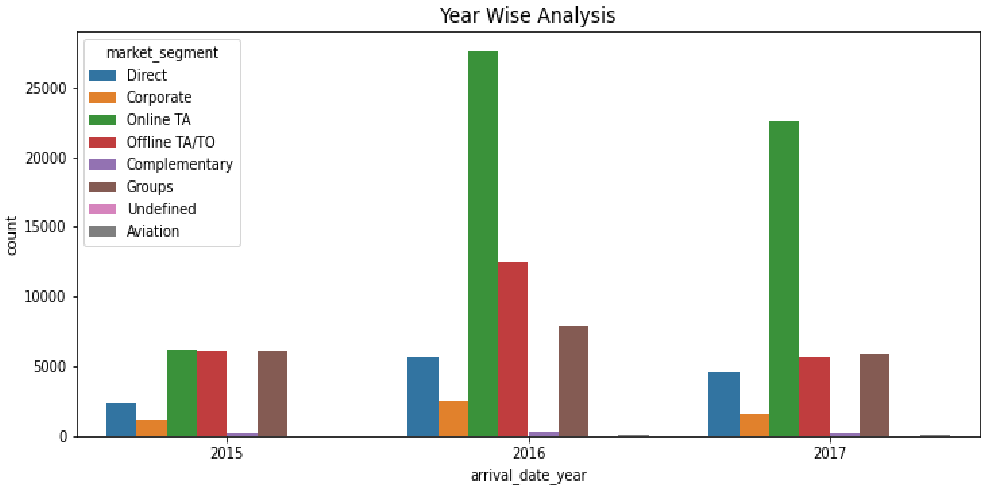
Interpretation-The Above Chart Represents the Country having Maximum bookings, as per the given data set. Portugal is the country which has close to 50000 bookings. There is a significant difference between booking status when compared the other countries. Portugal is having maximum bookings followed by the Great Britain with 10000+, France, Spain etc.

IV. Data Representing Year Wise Booking Status



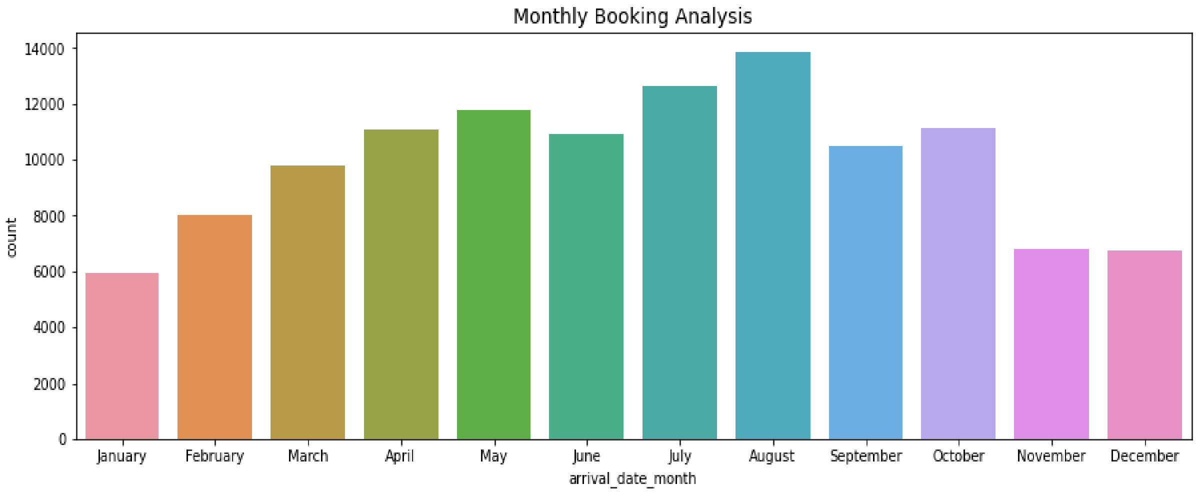
Interpretation-The Above Chart Represents the Type of the Hotel Preferred by the Visitors, the given graph is and Year wise comparison, where in we can find in 2016, the bookings were maximum as compare to year 2017 and 2015, With Year 2015 saw the least number of bookings as compared to the other two years.

V. Data Representing Yearly Booking Status – Segment Wise



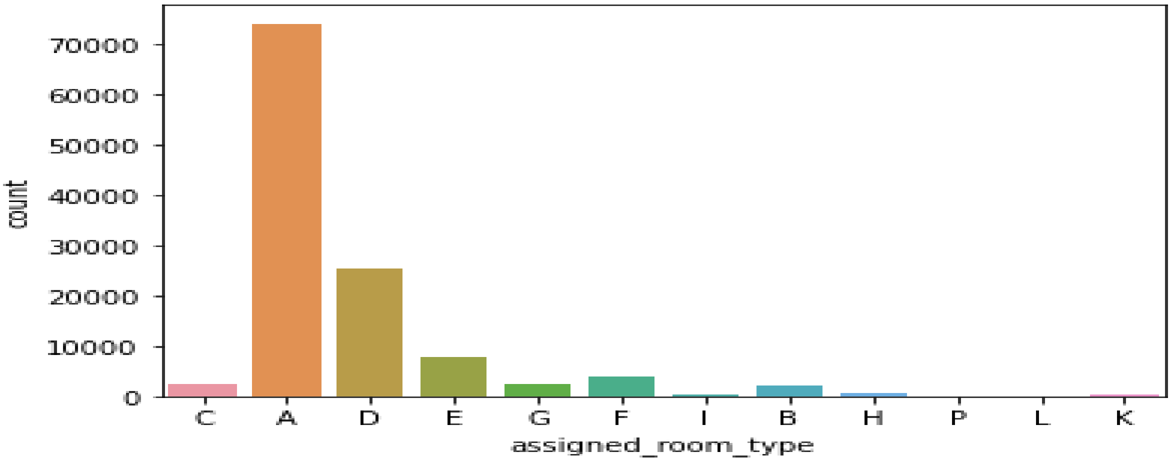
Interpretation-The Above Chart Represents the Yearly booking status as per the Market Segment, from the graph we can clearly see that the Corporate and Direct Consumers are having the least share as compared to the Online TA and Offline TA/TO

VI. Data Representing - Monthly Booking Analysis



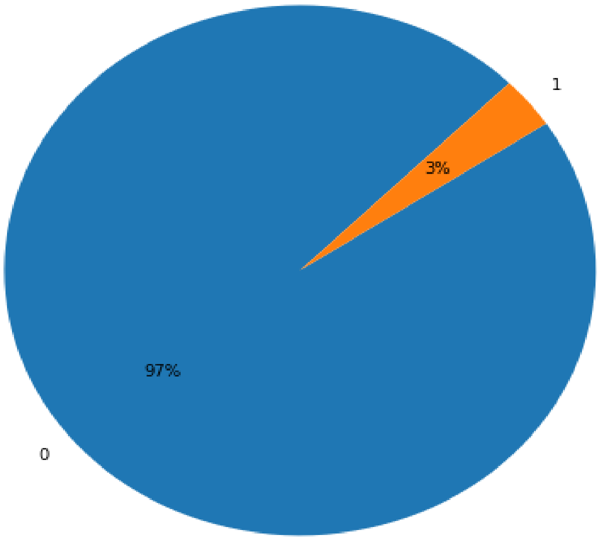
Interpretation-The Above Chart Represents the Monthly Booking Analysis of the Hotel, August month shows the maximum booking, and January month shows the least booking.

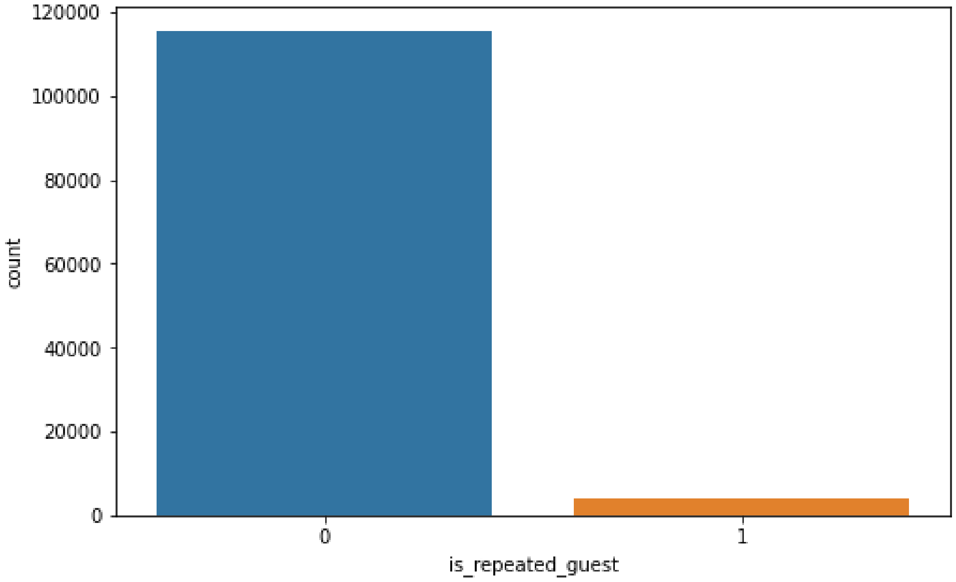
VII. Data Representing the Room Type Preferred by the Visitors



Interpretation: - From the Above graph we can clearly conclude that Room ‘A’ is preferred most of the time by the visitors followed by Room D and Room E.

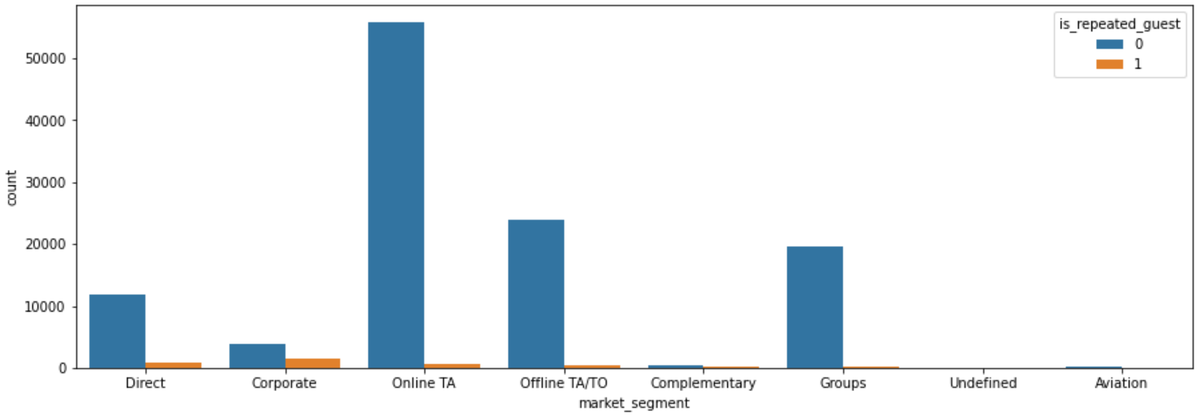
VIII. Data Representing the No. of Repeating Customers





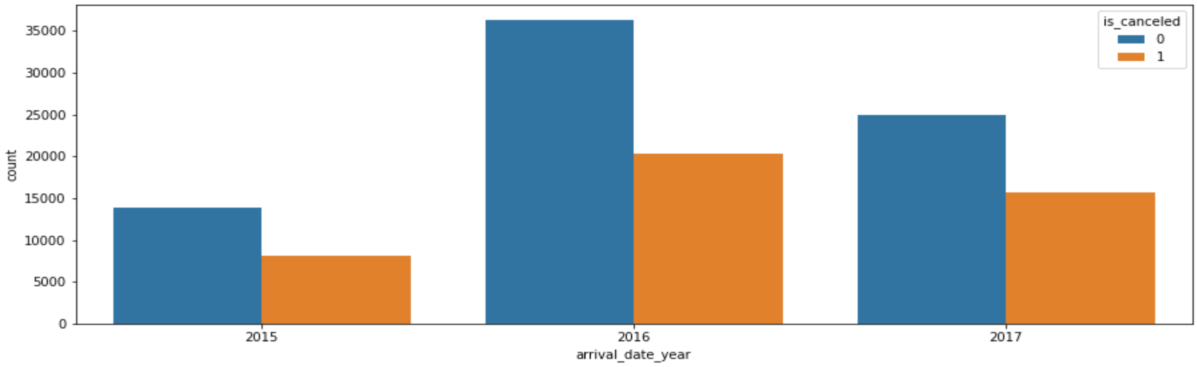
Interpretation: This data represents the No. of the customers repeated, as from the graph it clearly indicated that 97% of the times the customers are not repeated and only 3% of the people come under the category of the repeated guest.

IX. Data Representing Repeated Guest – Segment Wise



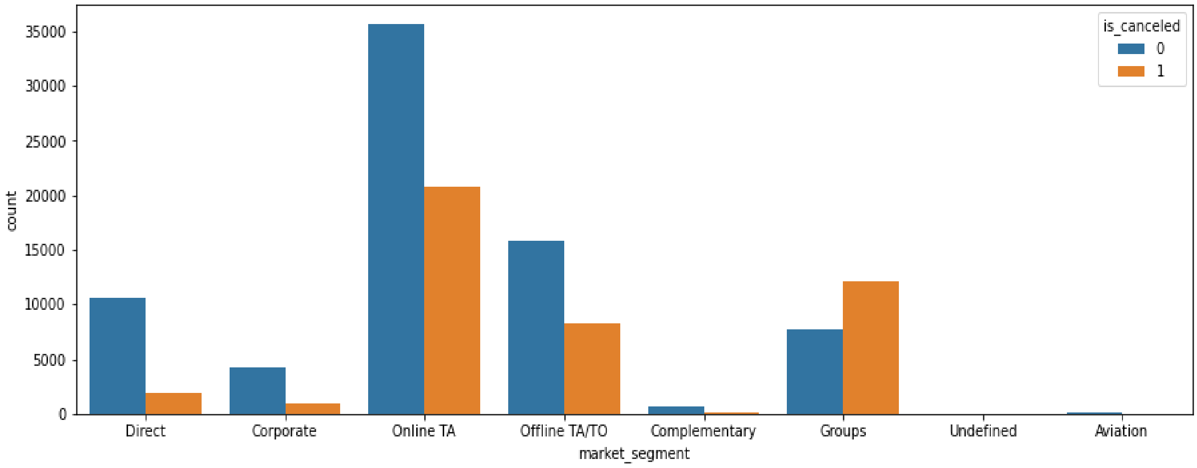
Interpretation: From the Above chart we can conclude that, Repeated Guest are mostly falling under the category of corporate segment followed by Direct and Online /Offline TA/TO. Whereas Undefined and Aviation don't have repeated Guest.

X. Data Representing Cancellation Status after Booking



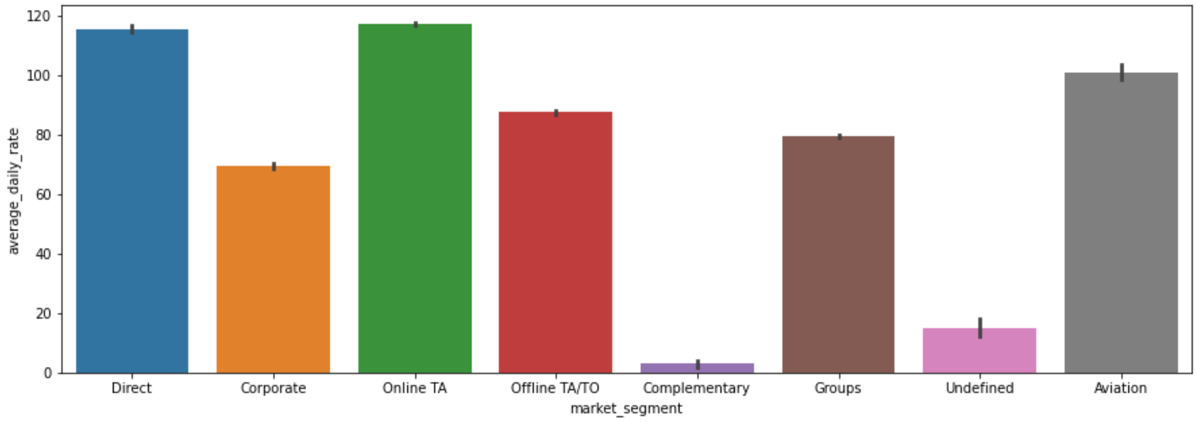
Interpretation: The above graph represents the cancellation happened after the booking, Year 2016 and 2017, shows the maximum booking cancellation, and the least cancellations were happened in the year 2015.

XI. Graph represents the cancellation happened after the booking – Segment Wise



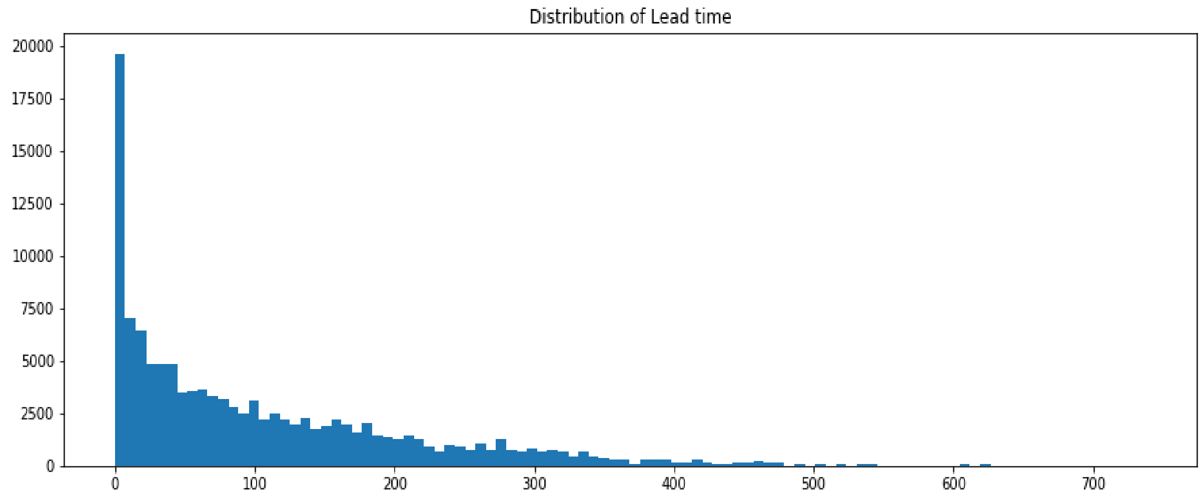
Interpretation: The above graph represents the cancellation happened after the booking, according to the segment, Online TA and Offline TA followed by the Groups forms the basis for the maximum cancellations.

XII. Graph Representing the Average Daily Rate



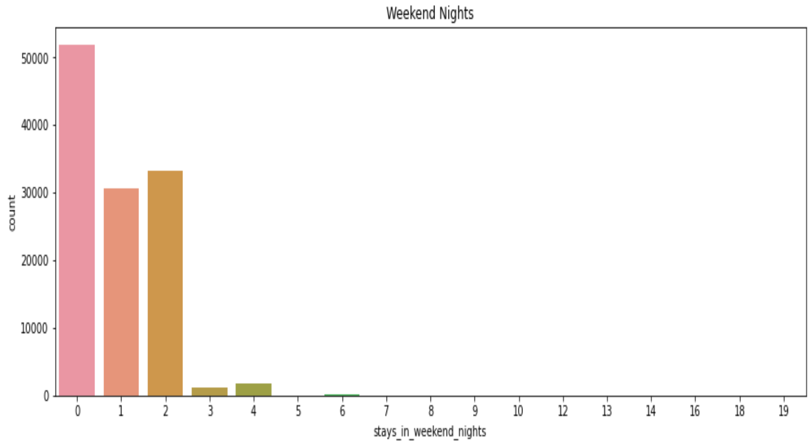
Interpretation: This graph represents the Average Daily rate according to the market segment, where Direct and Online TA forms the basis for maximum Average rate and Complementary are the least.

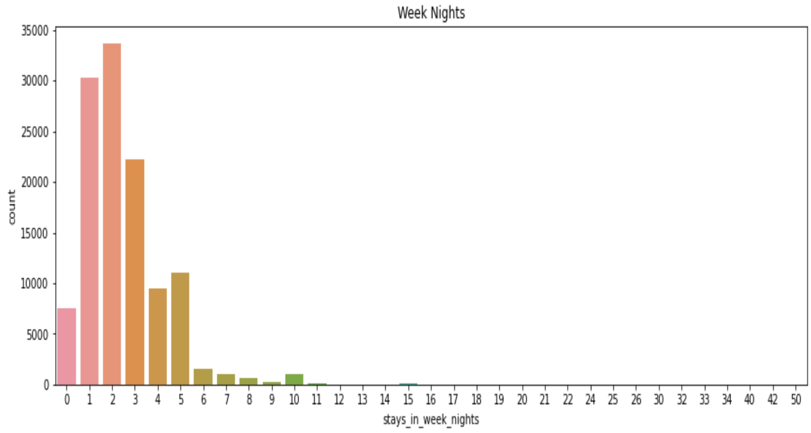
XIII. Data Chart Representing the Lead time Distribution



Interpretation: The above graph represents the Lead time of Booking. Mostly we don’t have to wait for the hotel booking but some hotel required very much time to booking.

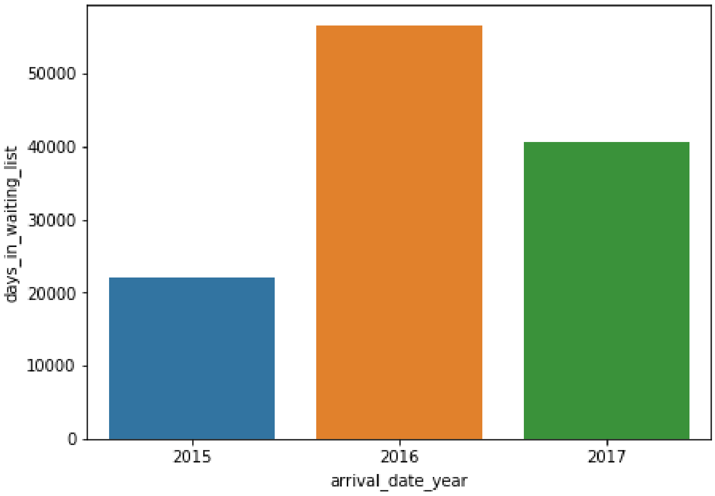
XIV. Data Representing Visitors Preference to stay





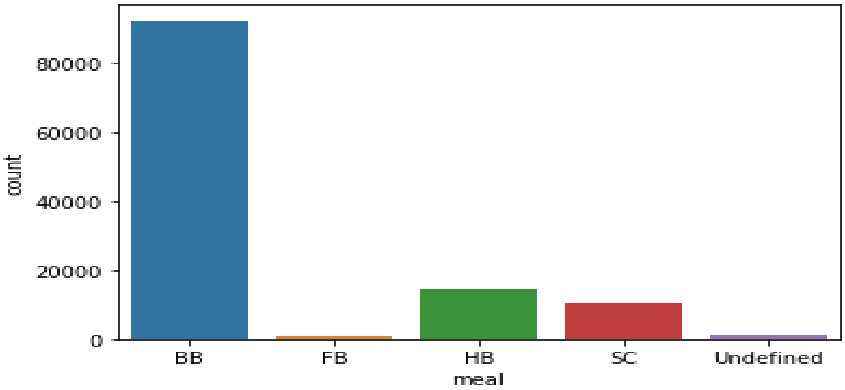
Interpretation: - The above graph represents the visitor's preference of stay, maximum of the visitors like to stay for 0-2 days maybe it is week day or week end.

XV. Data Representing the Waiting Time Period – Year wise



Interpretation: The above graph represents the waiting period according to the year; we can clearly conclude that year 2016 has a maximum waiting period and year 2015 with the least waiting period.

XVI. Data Representing type of Meal preferred by Visitor.



- **Interpretation :** The above graph represents that breakfast meal is most preferred while selection and booking hotel.

6. The Conclusive Findings of the Study

- City hotels are most busy hotels all the years and in 2016 it's on peak around 40000 customer booked city hotel
- Portugal is having maximum bookings followed by the Great Britain with 10000+, France, Spain etc.
- Majority of booking happed through online travel agent so we can provide great offers to TA
- August month has the maximum booking, and January month has the least booking number and majorly room type 'A' preferred by customers and room type 'P','L','K' is less preferred by customers
- One of our analyses showing many customers are not go repeat to same hotel only 3% retention rate of hotels
- the cancellation happened after the booking, Year 2016 and 2017, shows the maximum booking cancellation, and the least cancellations were happened in the year 2015.
- Direct and online TA gives most ADR so its helps to improve revenue of hotels
- 2016 is the busiest year because this year had most waiting list and 2015 least waiting year

References

- Google
- Alma better capstone project sample project section
- Think python