PHASE 1

PROJECT TITLE: HOTEL BOOKING PREDICTION

MEMBER 1: **Name:** Sahithi Bakaram. **UBID:** sahithib

MEMBER 2: **Name:** Spoorthy Reddy Avancha **UBID:** savancha

1. INTRODUCTION:

Nowadays, hotels are occupying a significant place in travelers' thoughts as they go from one place to another. Hotels genuinely feel like you're getting a break from all the daily routine and helps you get a fresh perspective. People typically reserve hotels for a variety of occasions, from necessity to luxury, such as when they require a co-working area for business meetings, when they wish for a staycation/vacation, or when they must leave work early for an event and need to change.

The hotels, on the other hand, need to know when the busy times are, who tends to cancel bookings, if parking spaces are in high demand, and how many people will be staying. We want to provide a summary of the features based on the dataset in this project.

2. PROBLEM STATEMENT:

As travel demand increases, it becomes harder to get a quality hotel room. Using "Hotel Booking Demand database", we use machine learning to create an approach for hotel booking by utilizing data and booking attitudes and behaviors. The travelers are increasingly impacted by the daily trend of hotel cancellations during border times. Using "Hotels booking Demand" analysis we can know the Hotel Cancellation Predictions, thereby users may choose whether to book a hotel in advance or not. Further, it will also address some of the following questions:

Which hotel is the most popular?

What elements affect to hotel cancellations?

What types of rooms are most popular?

From the Hotel perspective, knowing that the booking has a higher probability of getting cancelled in advance will help them to handle it so that they can minimize the loss and manage their revenue. Also, having insights about the booking might help them to improve the facilities they can provide so that they can enhance the user experience. For example,

1. Knowing about the frequency of the booking type i.e., if people are visiting for a business, you can enhance their stay experience by having a meeting room
2. If the guests include kids, you can have medical facilities ready incase of an emergency. You can also have a kids play area and a meal plan designed for them.

The dataset can be used for drawing multiple insights to increase hotel's revenue.

By comprehending the business need, the data was examined, cleaned and performed out exploratory data analysis and plotted different types of relationships.

The datasets that are currently available were gathered with the intention of creating classification schemes for the likelihood of cancellation of hotel reservations. However, these

datasets use extend beyond this cancellation prediction issue because of the properties of the variables of the datasets.

These datasets can be important for research in revenue management, machine learning, or data mining, as well as in other sectors, due to the lack of real business data for scientific and educational reasons.

3. DATA SOURCES:

- The “Hotels Booking Demand” data is taken from Kaggle.com. The data was first published in the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.
- Portugal provides most of the hotel data.

4. DATASET DETAILS:

The Hotel Booking Demand Dataset consists of hotel booking data from two different hotels located in Portugal. Hotel H1 is located at the resort region of Algarve and Hotel H2 is at the city of Lisbon. The Hotel names include Resort Hotel (H1) and City Hotel (H2). The dataset has the observations of 3 years.

The dataset contains 32 columns and 119390 rows in total. Resort Hotel and City Hotel comprises of 40060 and 79330 observations respectively.

The descriptions of various features in the dataset are mentioned below:

No	Column	Data type	Description
1	hotel	object	Hotel H1 represents Resort Hotel or Hotel H2 represents City Hotel
2	is_cancelled	int64	Value indicating if the booking was cancelled (1) or not (0)
3	lead_time	int64	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4	arrival_date_year	int64	Year of arrival date
5	arrival_date_month	object	Month of arrival date with 12 categories from January to December

6	arrival_date_week_number	int64	Week number of the arrival date
7	arrival_date_day_of_month	int64	Day of the month of the arrival date
8	stays_in_weekend_nights	int64	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
9	stays_in_week_nights	int64	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
10	adult	int64	Number of adults
11	children	float64	Number of children
12	babies	int64	Number of babies
13	meal	object	Type of meal books. Categories are presented in standard hospitality meal packages: Undefined/SC: no meal package, BB: Bed & Breakfast, HB: Halfboard, FB: Fullboard
14	country	object	Country of origin
15	market_segment	object	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
16	distribution_channel	object	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
17	is_repaeted_guest	int64	Value indicating if the booking name was from a repeated guest (1) or not (0)

18	previous_cancellations	int64	Number of previous bookings that were cancelled by the customer prior to the current booking
19	previous_bookings_not _cancelled	int64	Number of previous bookings not cancelled by the customer prior to the current booking
20	reserved_room_type	object	Code of room type reserved
21	assigned_room_type	object	Code for the type of room assigned to the booking
22	booking_chnages	int64	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
23	deposit_type	object	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit, Non-Refund, and Refundable
24	agent	float64	ID of the travel agency that made the booking
25	company	float64	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
26	days_in_waiting_list	int64	Number of days the booking was in the waiting list before it was confirmed to the customer
27	customer_type	object	Type of booking, assuming one of four categories: Contract, Group, Transient, Transient-party

28	adr	float64	Average Daily Rate
29	required_car_parking_spaces	int64	Number of car parking spaces required by the customer
30	total_of_special_requests	int64	Number of special requests made by the customer (e.g. twin bed or high floor)
31	reservation_status	object	Reservation last status, assuming one of three categories: Canceled, Check-Out, No-Show,
32	reservation_status_date	object	Date at which the last status was set.

5. DATA CLEANING / PROCESSING:

The below mentioned steps are done as part of the Data Preprocessing:

i. Deleting the duplicate values if any.
There are 31994 rows which contain duplicate values. So, after the duplicated rows are removed, we have 87396 rows in total.

ii. Handling the missing values.
The null values are present only in the company, agent, country, children in the descending order. The total count of null values are 94786. We have dropped the company and agent columns. So, there are 30 columns and 87396 rows at present. Children and country column null values are replaced by their mean and mode values respectively.

iii. Deleting the unwanted rows and columns.
We will delete the rows with total guest value equal to 0. The rows of columns adults, babies, children whose value is equal to 0 are removed from the dataset. There are 166 rows which has value equal to 0. So, the total rows are 87230.
If the total stay in weekends and weekdays are zero then there are no bookings in the hotel. So, we will delete those rows with value equal to 0. As such there are 591 rows. Dropping these rows results in 86639 rows in total.

iv. Converting all columns to proper data types and precision
- We converted the reservation_status_date from string type to datetime datatype
- We can observe that the no.of children value is of float datatype. So, converting this feature into integer datatype as count of children cannot be float
- Rounding up all the values of adr column to precision 2

v. Identifying the Outliers
We plotted the boxplot to represent each feature center thereby getting information about the distribution of values and detecting the anomalies. Calculation of outliers is done by IQR and imputed them.

vi. Shape of the Distribution
Based on the calculated value of skewness, we decided whether or not the values of the feature are left-skewed or right-skewed.
For Example, in the hist-plot, lead_time is skewed-right, adr is symmetric and unimodal, arrival_date_of_month is uniform

vii. Measures of Center: Mean, Median
Mean and Median are calculated thereby plotting the boxplot. Through which we can information about the spread of the values. For suppose, in the box plot, lead_time, stays_in_weekend_nights, stays_in_week_nights, total_stays have the similar spread i.e., first half of the values are lesser than the second half of the values as the median is towards left. The values of arrival_date_week_number, arrival_date_day_of_month, adr are spreaded correctly towards the center.

viii. Modality
Modality gives the most frequent value in the dataset. We have calculated the mode values for each feature.

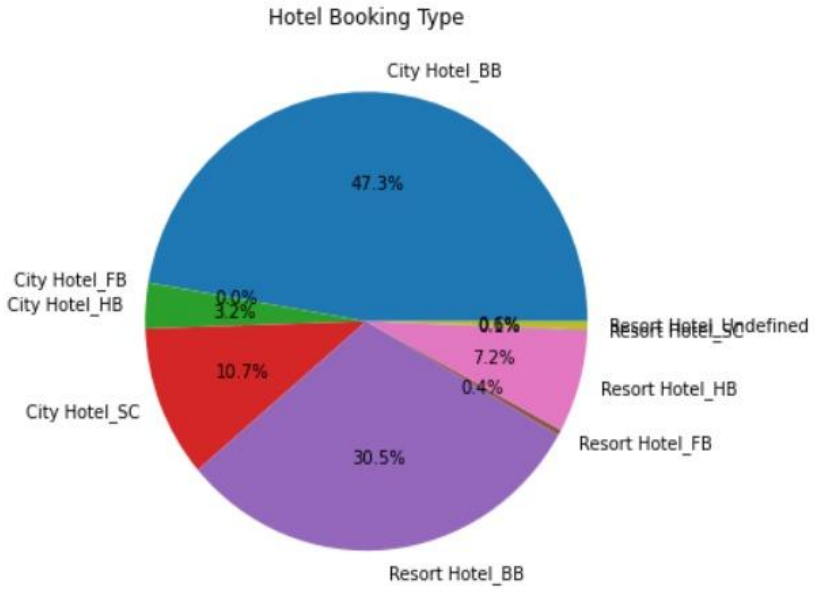
ix. Quantiles
We split the date into four parts using quantiles i.e., 25%, 50%, 75%, max.
25% refers to the first quartile, 50% refers to the median, 75% refers to the third quartile.

x. Removing the outliers
By calculating Inter Quartile Range IQR, Q1, Q3, we remove the outliers.

6. Exploratory Data Analysis (EDA):

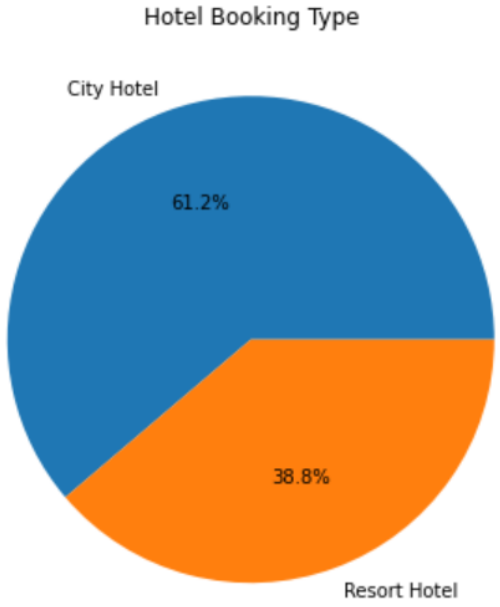
i. Pie Plot:

The below plot helps to know about the type of the Hotel Booking in various categories: City Hotel_BB, City Hotel_FB, City Hotel_HB, City Hotel_SC, Resort Hotel_Undefined, Resort Hotel HB, Resort Hotel FB, Resort Hotel BB. Of this City Hotel_BB has the highest percentage.



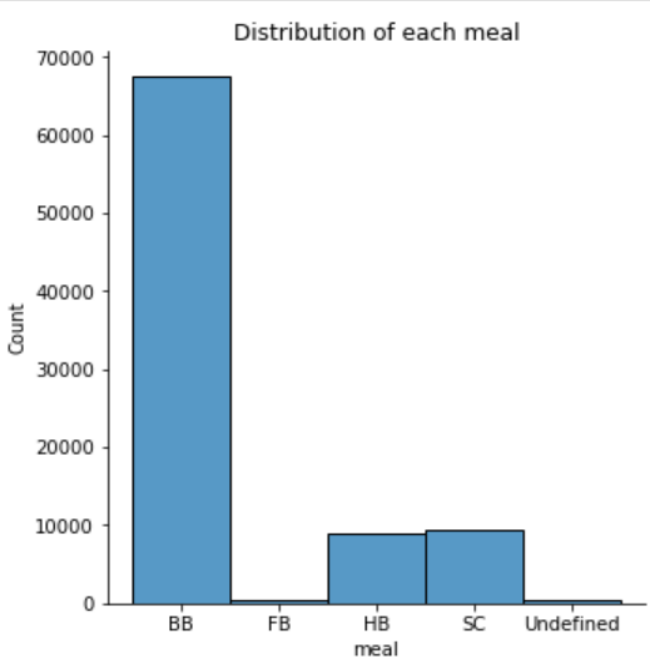
ii. Pie Chart:

On whole, City Hotel has the highest booking than Resort Hotel.

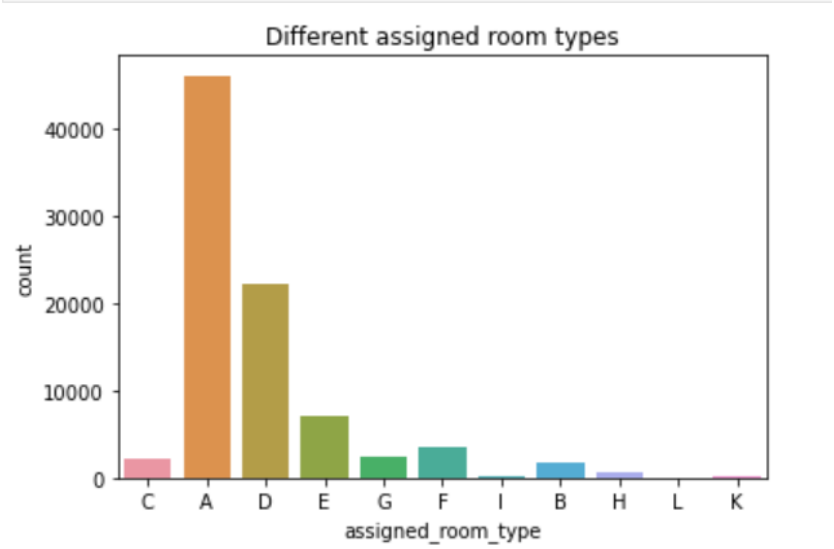


iii. Bar Graph:

The below bar graph represents the different categories of meal plan i.e., BB, FB, HB, SC, Undefined. Highest count is for BB category.

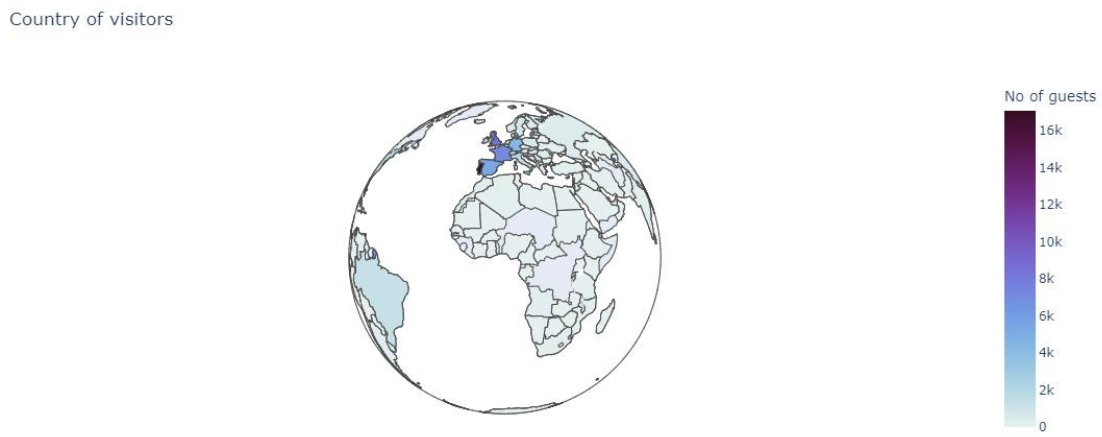


iv. The below bar graph represents the different assigned room types and its occupancy. The highest occupancy is for room type'A'

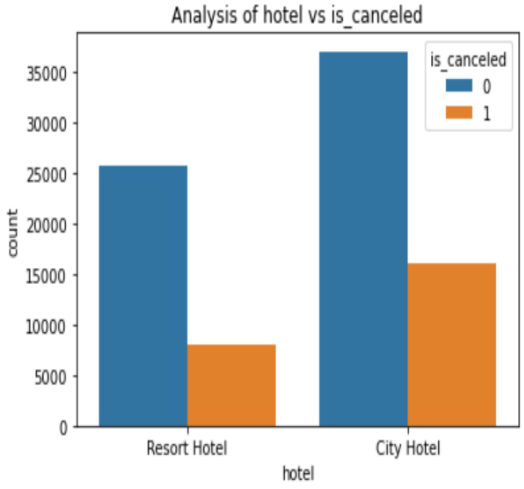


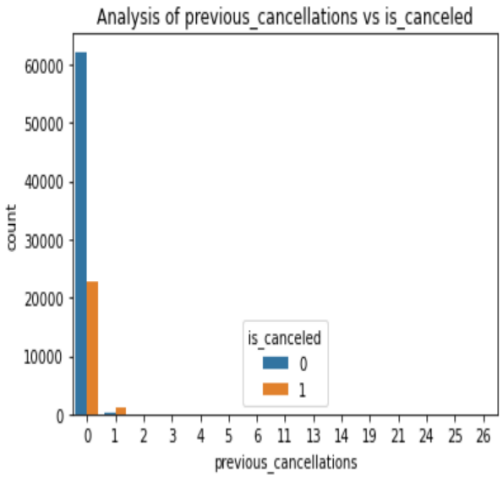
v. Map Plot:

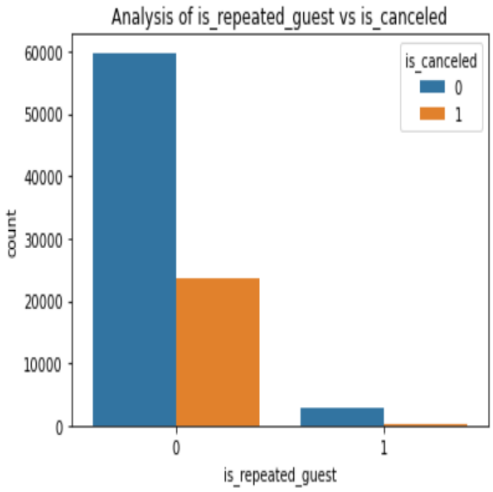
The below plot describes the number of visitors from different countries.

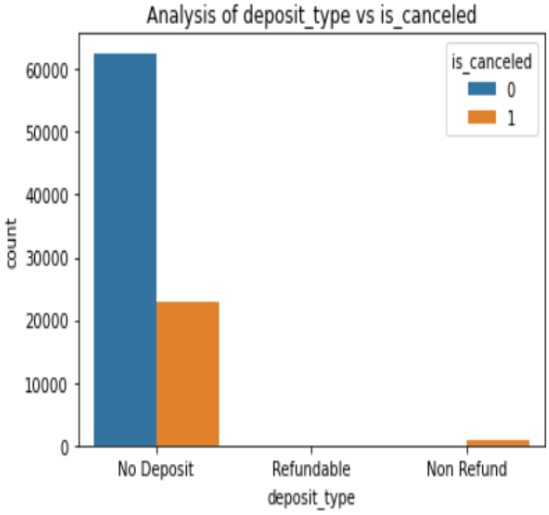


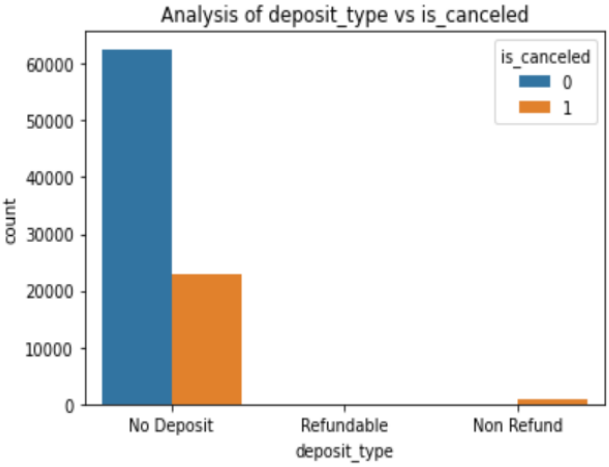
vi. Multivariate Analysis for predicting is_cancelled based on the other independent features.

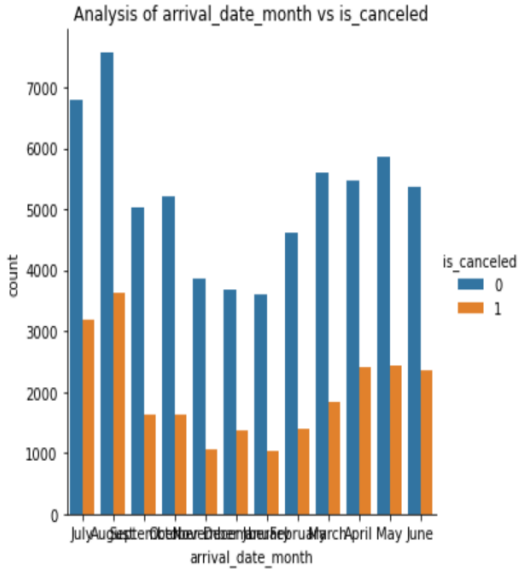

Analysis of hotel vs is_canceled
is_canceled
0
1
count
hotel
Resort Hotel
City Hotel

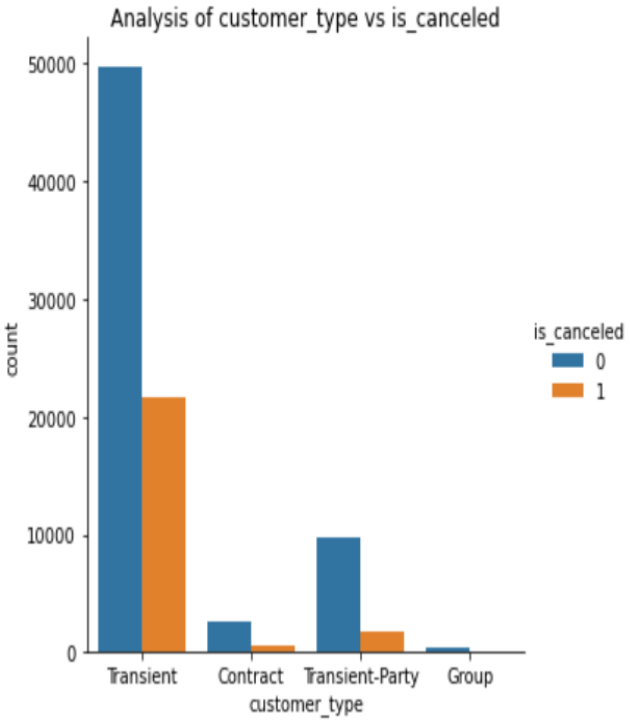

Analysis of previous_cancellations vs is_canceled
is_canceled
0
1
count
previous_cancellations

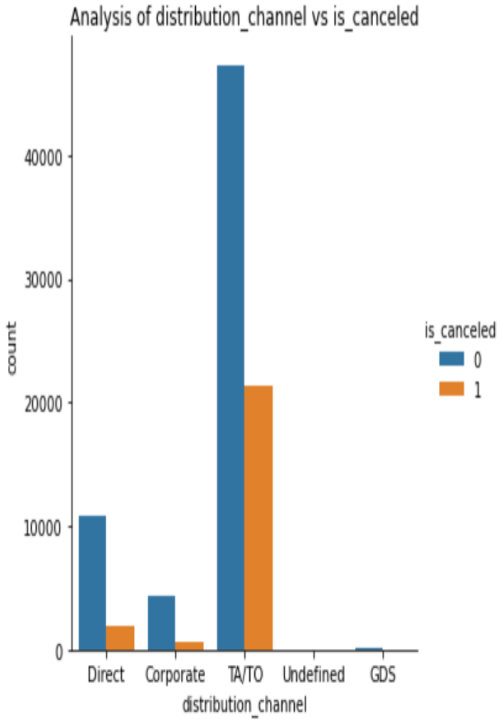

Analysis of is_repeated_guest vs is_canceled
is_canceled
0
1
count
is_repeated_guest


Analysis of deposit_type vs is_canceled
is_canceled
0
1
count
No Deposit
Refundable
Non Refund
deposit_type


Analysis of deposit_type vs is_canceled
is_canceled
0
1
count
No Deposit
Refundable
Non Refund
deposit_type

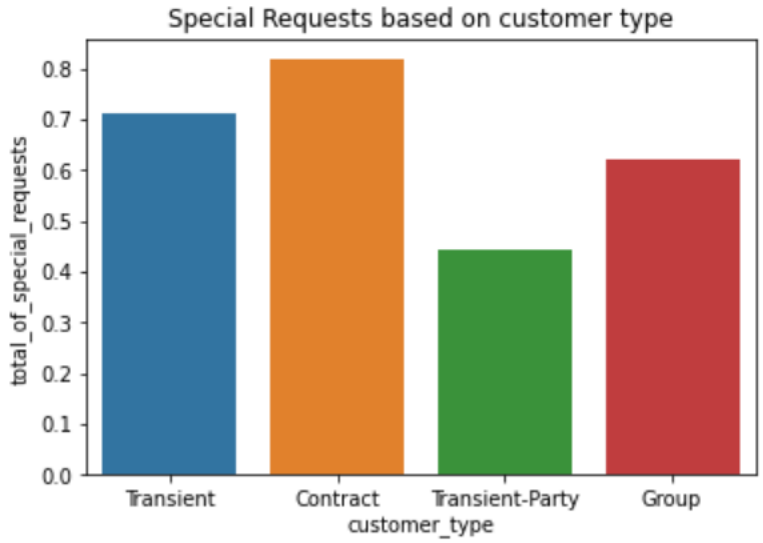

Analysis of arrival_date_month vs is_canceled
count
arrival_date_month
is_canceled
0
1
July
June


Analysis of customer_type vs is_canceled
count
customer_type
Transient
Contract
Transient-Party
Group
is_canceled
0
1


Analysis of distribution_channel vs is_canceled
count
distribution_channel
Direct
Corporate
TA/TO
Undefined
GDS
is_canceled
0
1

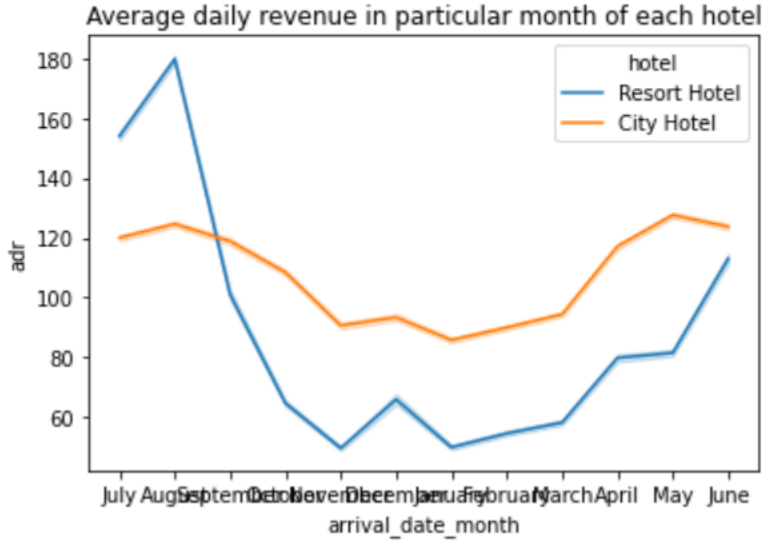
vii. Bar Graph

The below graph represents the special requests based on the customer type.



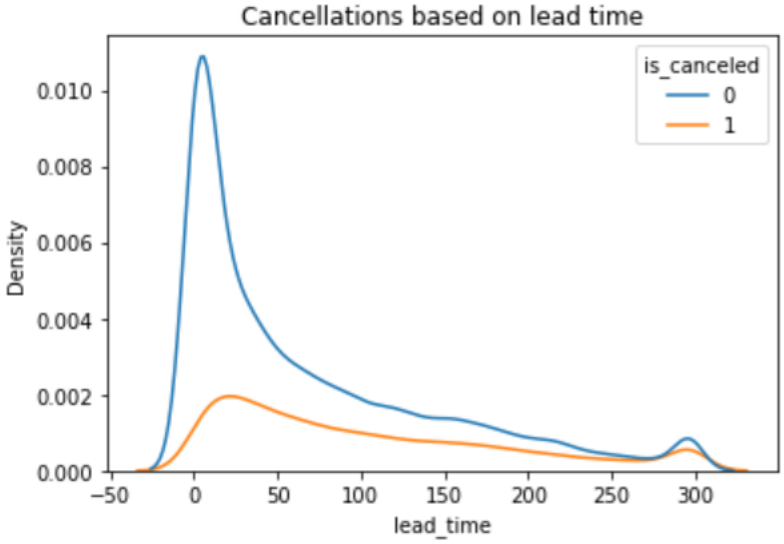
viii. Line Graph

The below graph represents the Average daily revenue in particular month of each hotel for a year. The highest is for the Resort Hotel in the month of September. And the lowest is for the Resort Hotel in 2 different months.



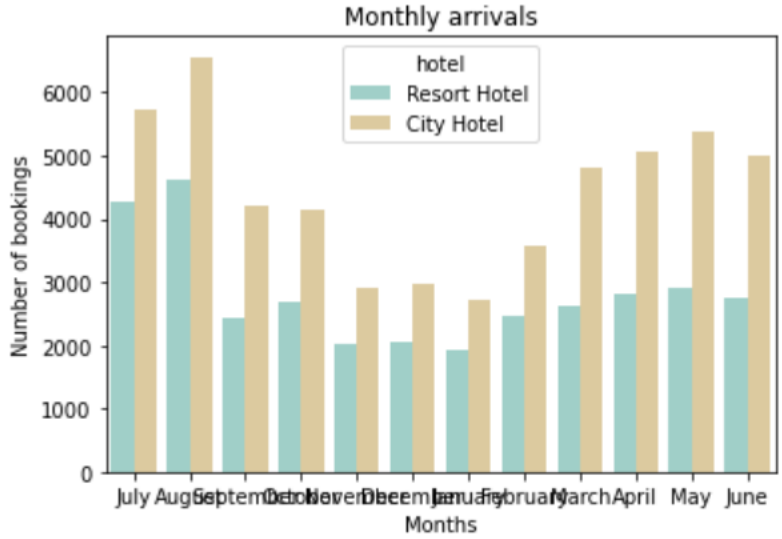
ix. Line Graph

The below plot describes the cancellations based on the lead time.



x. Histogram

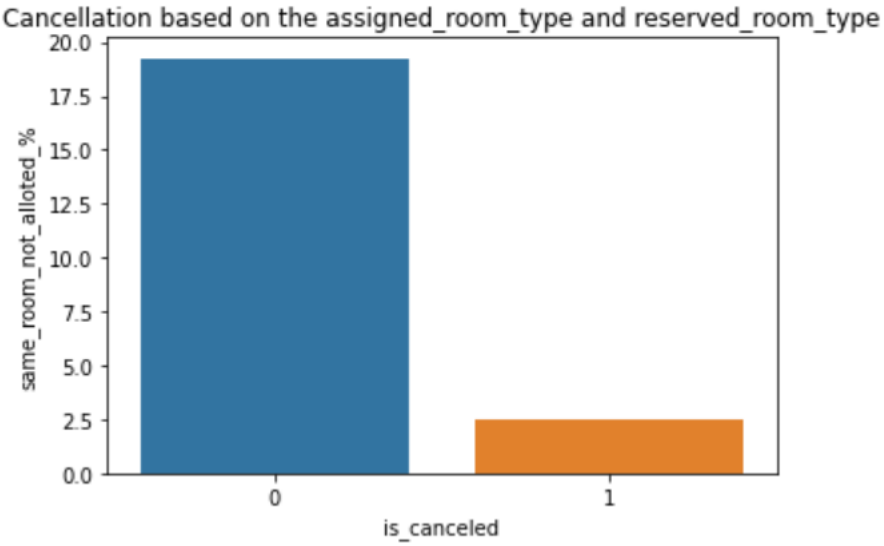
The below histogram describes the Monthly Bookings for 12 different months of both the hotels. Th highest booking is for the City Hotel and lowest booking is for the Resort Hotel.



xi. Bar Graph

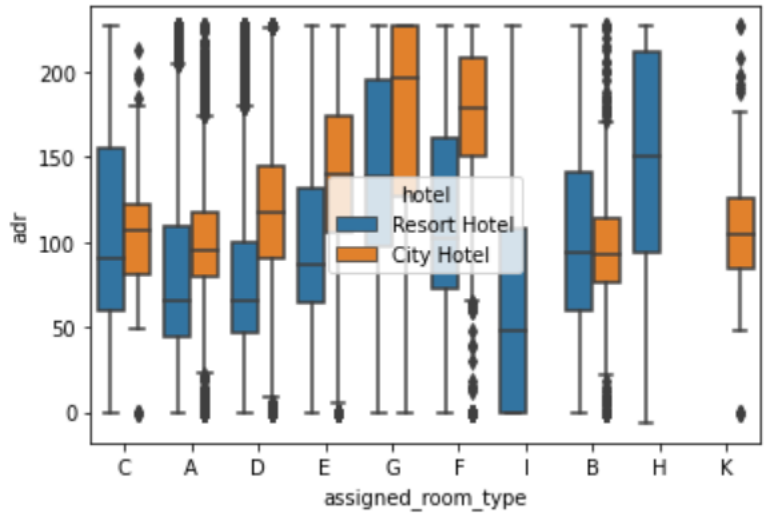
The below graph describes the Cancellations based on the assigned_room_type and

reserved_room_type.



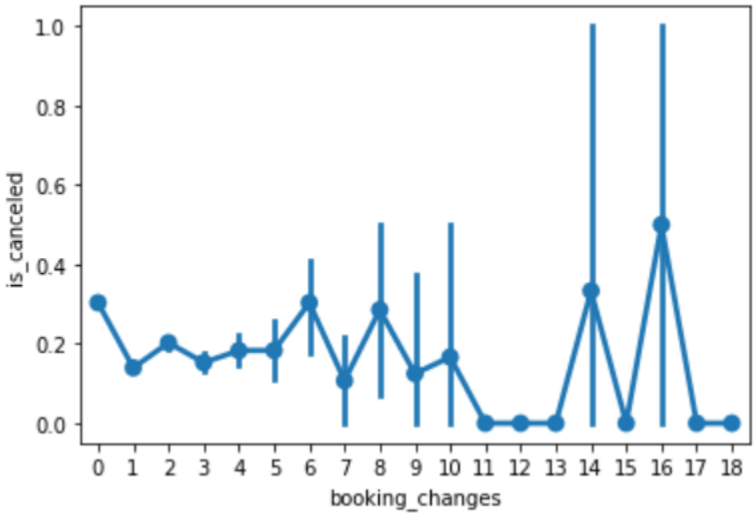
xii. Box Plot

The below plot analyzes the variation of assigned_room_type w.r.t adr while using hotel as hue variable. This helps us compare the trend for different aggregation indices.



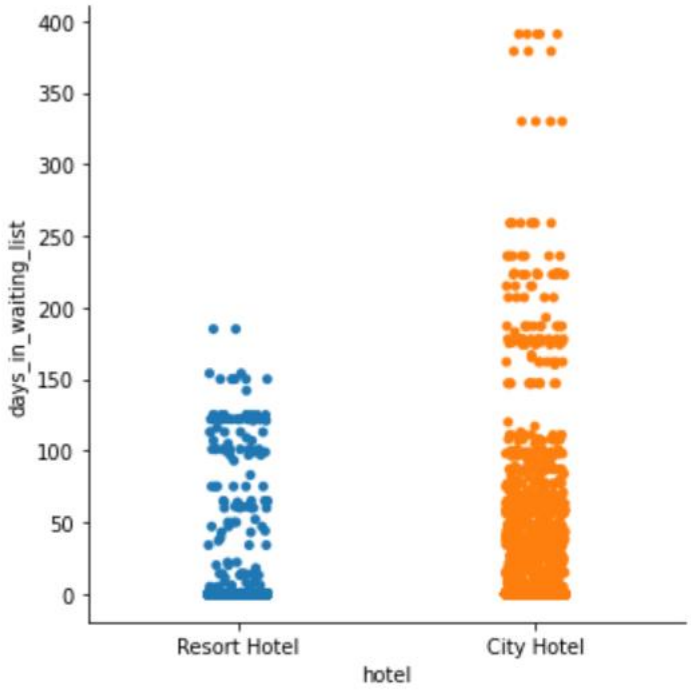
xiii. Point Plot

The below plot analyzes how booking_changes is varying w.r.t is_cancelled feature.



xiv. Scatter Plot

The below plot describes the there are people who are in the waiting list for a longer time in City Hotel rather than Resort Hotel.



After Exploratory Data Analysis, we found out the correlations between is_cancelled and other independent variables. This correlations helps us to make improvements in the Hotel Management System by improving techniques, thereby cancellation of bookings can be minimized.

7. REFERENCES

- https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand