EDA Capstone Project – 1

Hotel Booking Analysis

Zuber Ahmad

Contents :

- PREPROCESSING THE DATASHEET
  - Data Cleaning
  - Treating Missing & Duplicate Values
- EXPLORATORY DATA ANALYSIS (EDA)
  - Data Exploration
  - Univariate Analysis
  - Hotel wise Analysis
  - Distribution Channel wise Analysis
  - Booking Cancellation Analysis
  - Time Wise Analysis
- DATA VISUALISATION TOOLS
  - Matplotlib
    - matplotlib.pylot
    - matplotlib.image
  - Seaborn.

- DATA VISUALISATION TECHNIQUES
  - Histogram Plot
  - Bar Plot
  - Pie Plot
  - Scatter Plot
  - Line Plot
  - Count Plot
  - KDE Plot

INTRODUCTION

- **Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, 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 or decision trees.**

- **We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings**

OBJECTIVES

- In this, we will perform an exploratory data analysis (EDA) in order to investigate each of the features and also come up with a conclusion for the relationship between features and variable.
- Some objectives of analysis are given below:
  - ✓ Cancelation rates in City hotel and Resort hotel.
  - ✓ What is the booking ratio between Resort Hotel and City Hotel?
  - ✓ What is preferred stay in each hotel?
  - ✓ Which hotel has higher bookings cancellation rate.
  - ✓ From which country most guests come?
  - ✓ Which hotel has high chance that its customer will return for another stay?
  - ✓ Which was the most booked accommodation type (Single, Couple, Family)?

Problem Description

- **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 analyze the data to discover important factors that govern the bookings.**

Data Description

Given data set has different columns of variables crucial for hotel bookings. Some of them are:

- **hotel: The category of hotels, which are two resort hotel and city hotel.**
- **is_cancelled: The value of column show the cancellation type. If the booking was cancelled or not. Values[0,1], where 0 indicates not cancelled.**
- **lead_time: The time between reservation and actual arrival**
- **stayed_in_weekend_nights: The number of weekend nights stay per reservation**
- **stayed_in_weekday_nights: The number of weekday nights stay per reservation.**
- **meal: Meal preferences per reservation.[BB, FB, HB, SC, Undefined]**
- **country: The origin country of guest.**
- **market_segment: This column show how reservation was made and what is the purpose of reservation. Eg , corporate means corporate trip , TA for travel agency.**
- **distribution_channel: The medium through booking was made.[Direct, Corporate, TA TO, undefined, GDS .]**
- **Is_repeated_guest: Shows if the guest is who has arrived earlier or not. Values [0,1]---->0 indicates no and 1 indicated yes person is repeated guest.**
- **days_in_waiting_list: Number of days between actual booking and transact.**
- **customer_type: Type of customers(Transient, group, etc.)**

Exploratory Data Analysis

- Exploratory data analysis popularly known as EDA is a process of performing some initial investigations on the dataset to discover the structure and the content of the given dataset. It is often known as Data Profiling. It is an unavoidable step in the entire journey of data analysis right from the business understanding part to the deployment of the models created.
- EDA can be divided into two categories: graphical analysis and non-graphical analysis. EDA is a critical component of any data science or machine learning process.

- Steps involve before EDA:
  - Importing the Python Libraries
  - Loading the Dataset in Python
  - Structured Based Data Exploration
  - Handling Duplicates
  - Handling Outliers
  - Handling Missing Values

Exploratory Data Analysis

AI

- **Data Wrangling:** Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

  Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. Pandas is an open-source library specifically developed for Data Analysis and Data Science. The process like data sorting or filtration, Data grouping, etc.

  Data wrangling in python deals with the below functionalities:

a. Data exploration: In this process, the data is studied, analyzed and understood by visualizing representations of data.

b. Dealing with missing values: Most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.

c. Reshaping data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.

d. Filtering data: Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered

e. Other: After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training etc.

Data Visualization Tools

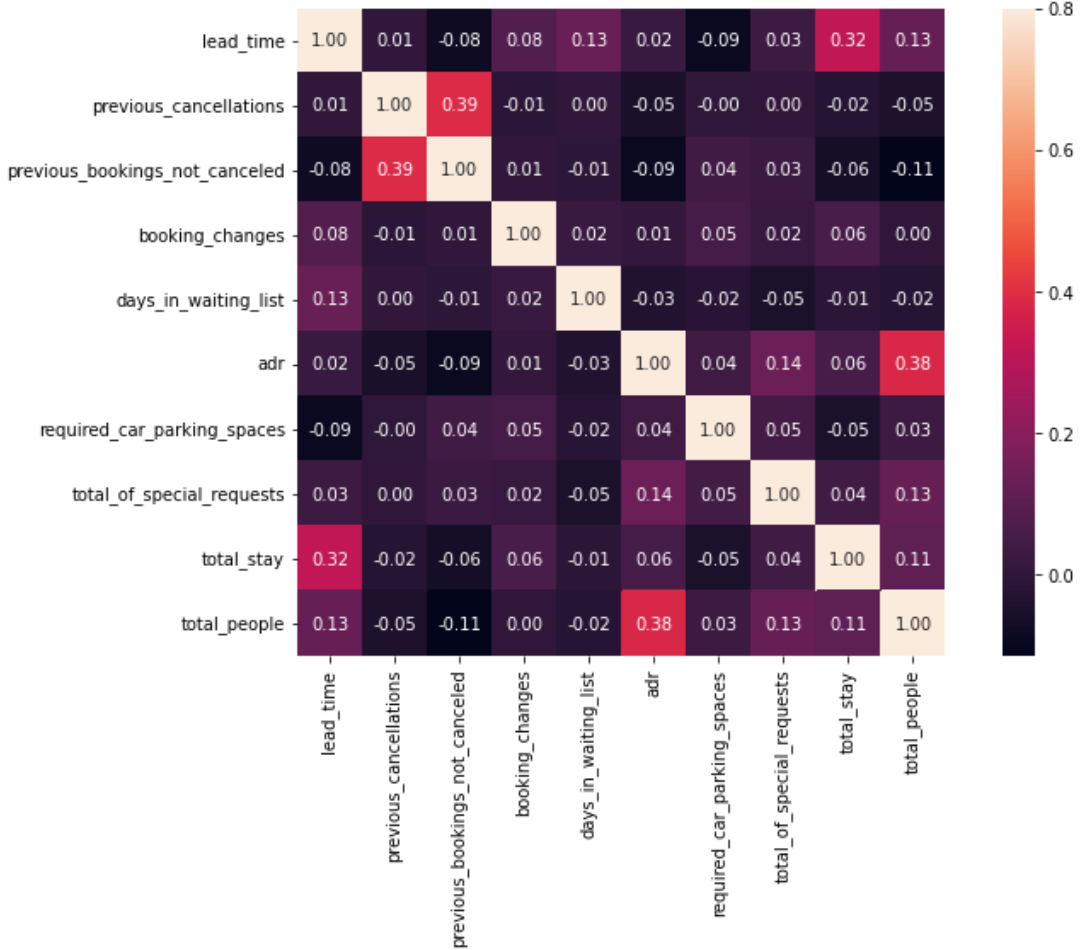
- **Matplotlib**

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Matplotlib is an amazing visualization library in Python for 2D plots of arrays.

  Main matplotlib module are:

  - matplotlib.pyplot: Pyplot is a Matplotlib module which provides a MATLAB-like interface. Matplotlib is designed to be as usable as MATLAB, with the ability to use Python and the advantage of being free and open-source. Pyplot is an API (Application Programming Interface) for Python's matplotlib that effectively makes matplotlib a viable open source alternative to MATLAB. Matplotlib is a library for data visualization, typically in the form of plots, graphs and charts.
  - matplotlib.image: The image module in matplotlib library is used for working with images in Python. The image module also includes two useful methods which are imread which is used to read images and imshow which is used to display the image.

Exploratory Data Analysis

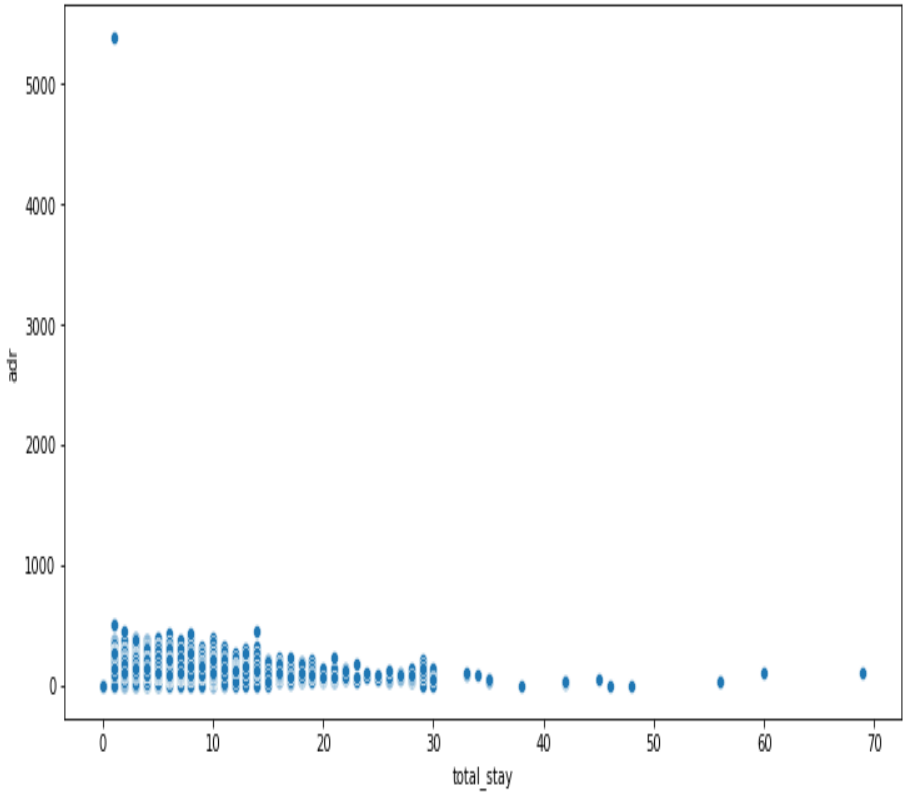


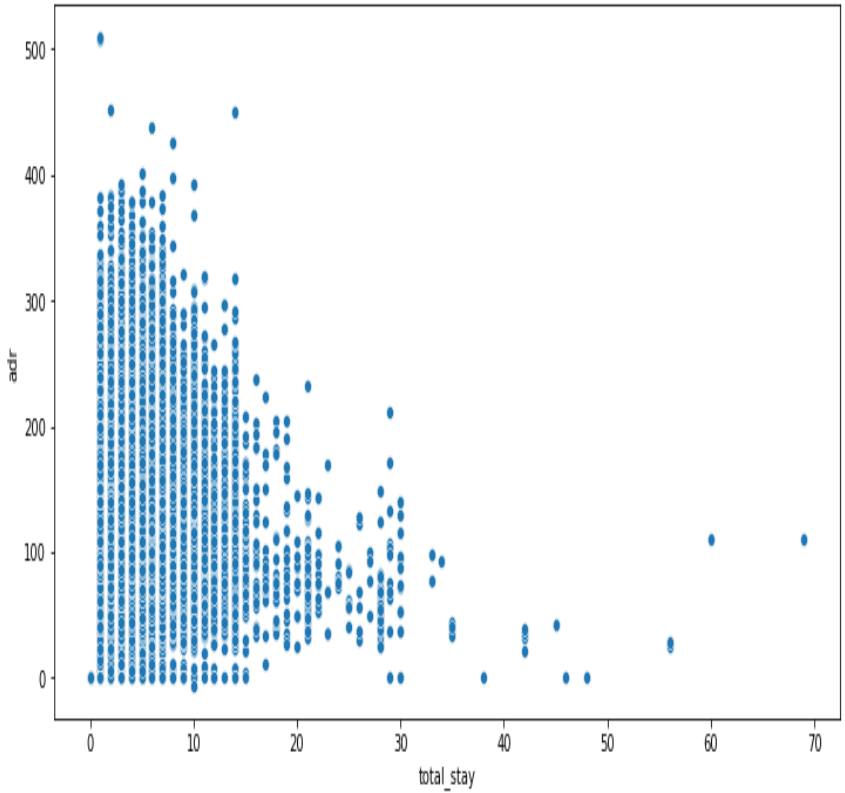
Since, columns like 'is_cancelled', 'arrival_date_year','arrival_date_week_number','arrival_date_day_of_month', 'is_repeated_guest', 'company', 'agent' are categorical data having numerical type. So we wont need to check them for correlation.
And we get following result:

1) Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the the actual arrival.

2) adr is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.

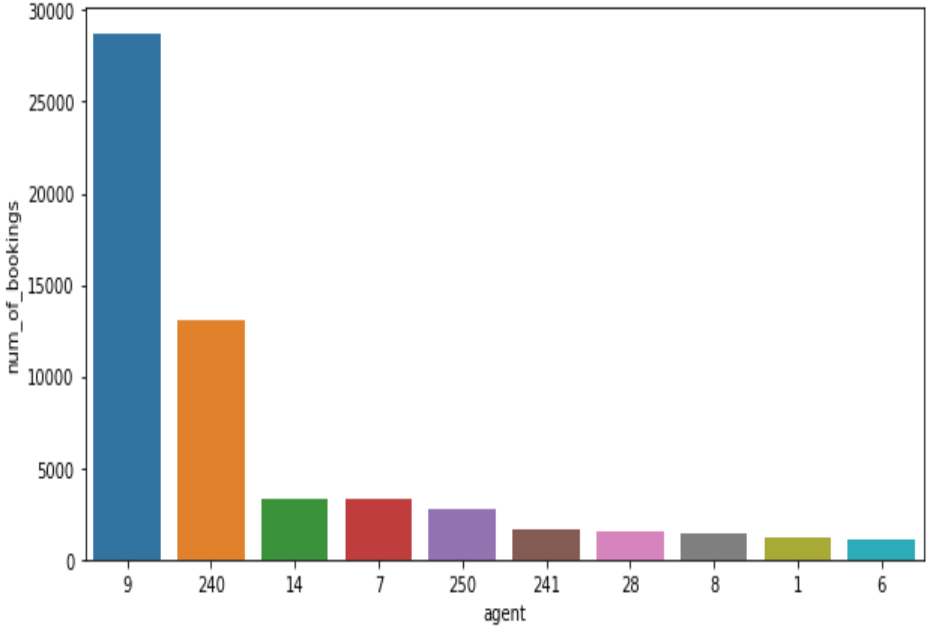
Exploratory Data Analysis

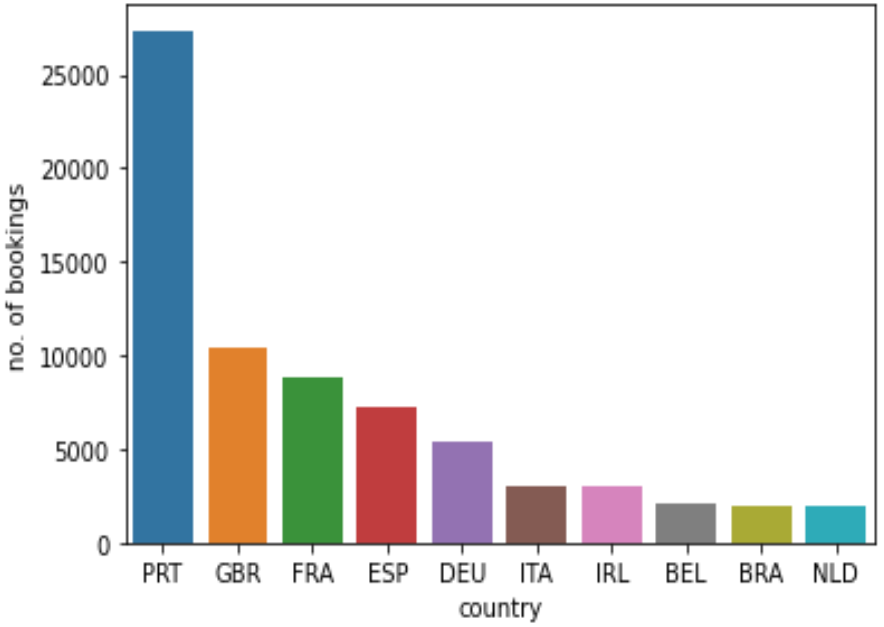




From the scatter plot we can see that as length of total_stay increases the adr decreases. This means for longer stay, the better deal for customer can be finalized.

Exploratory Data Analysis

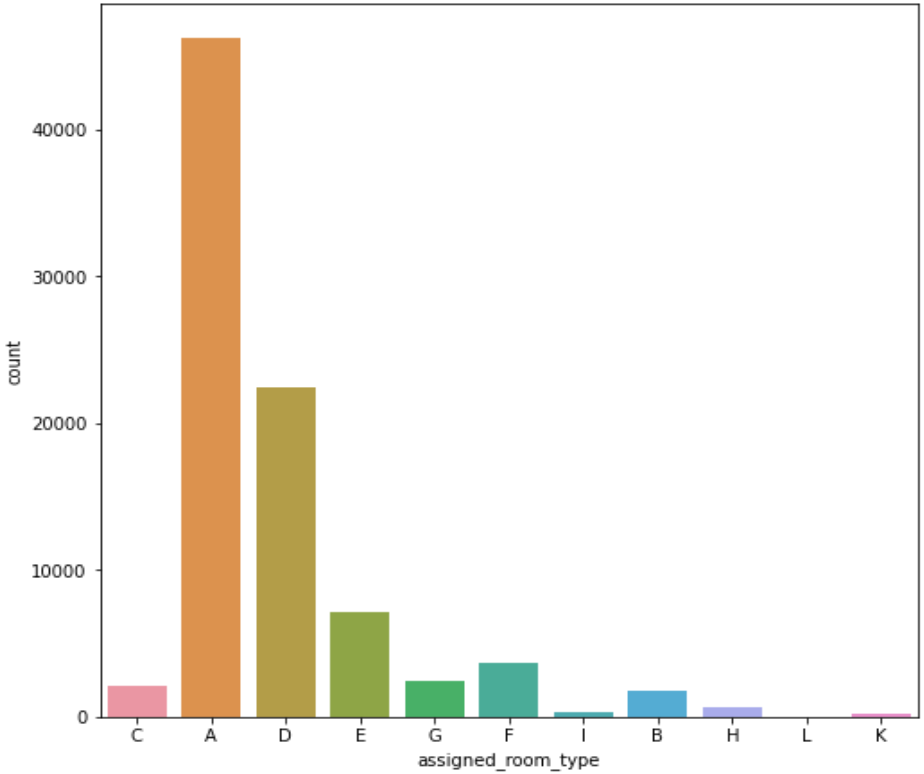


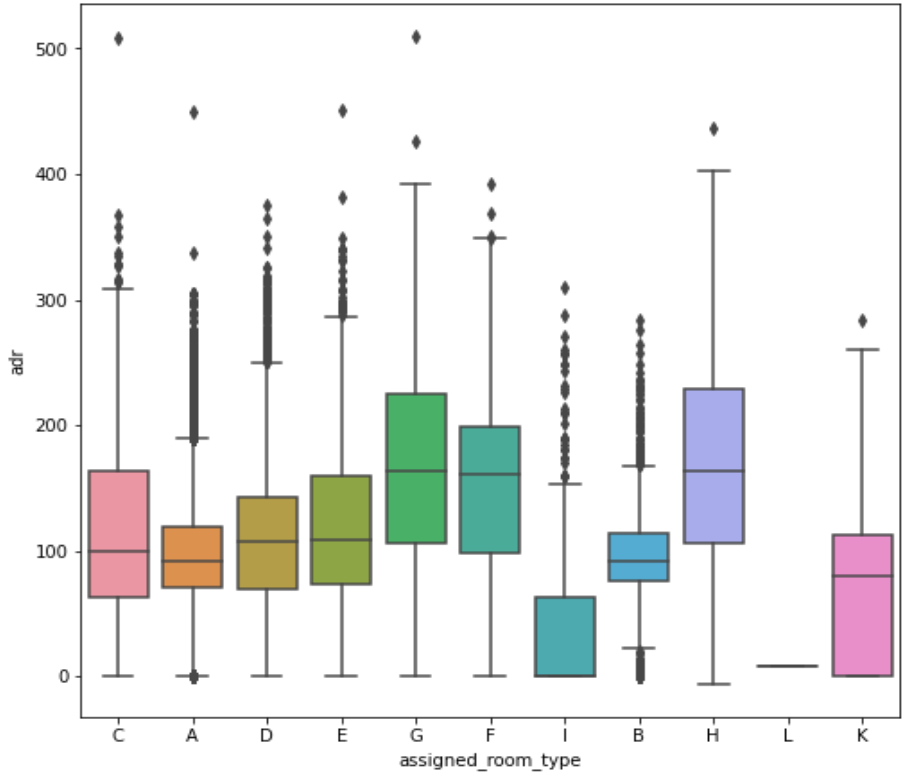


Agent no. 9 has made most no. of bookings.

Most of the customers come from western countries - Portugal, Great Britain, France and Spain.

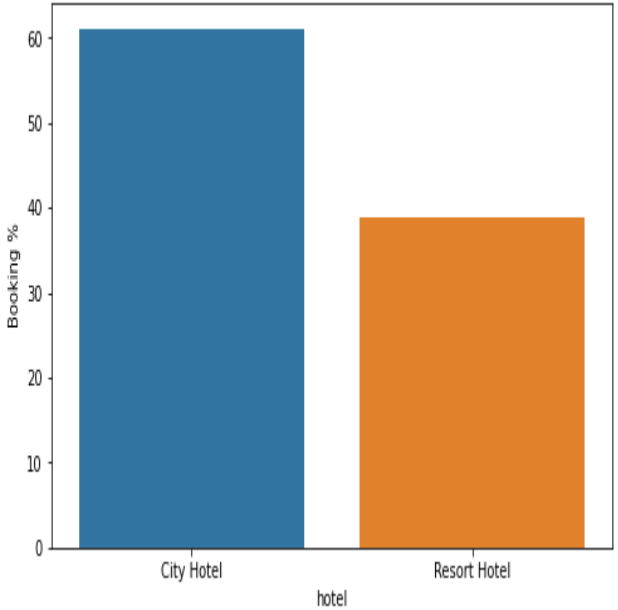
Exploratory Data Analysis



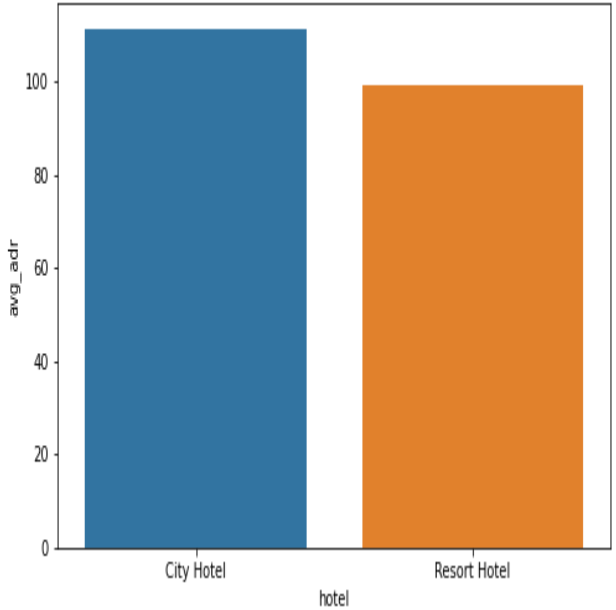


Most demanded room type is A, but better adr rooms are of type H, G and C also. Hotels should increase the no. of room types A and H to maximize revenue.

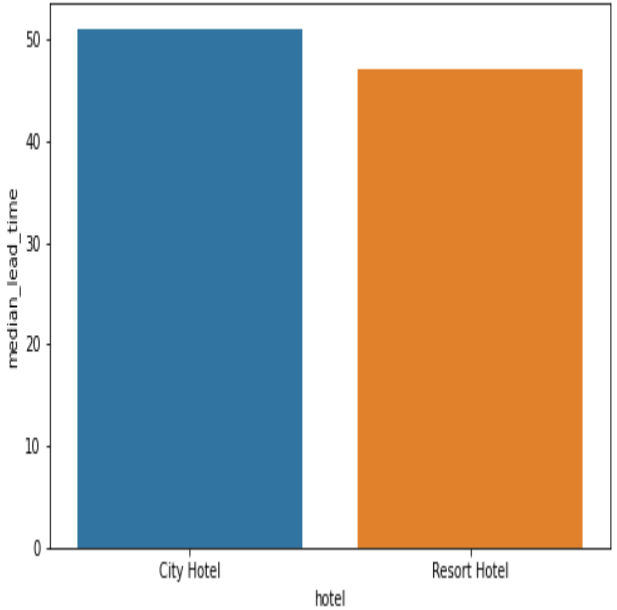
Exploratory Data Analysis



Around 60% bookings are for City hotel and 40% bookings are for Resort hotel.

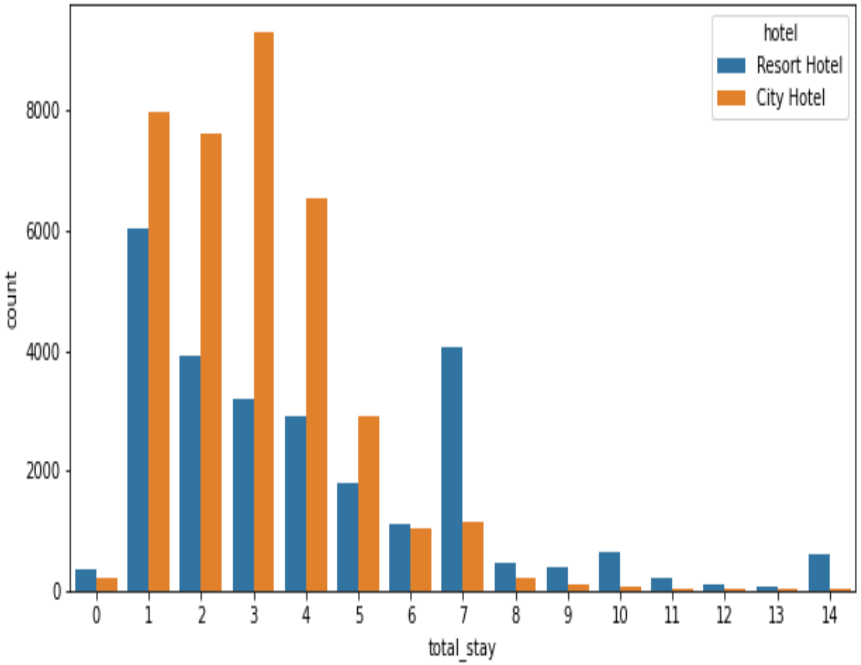


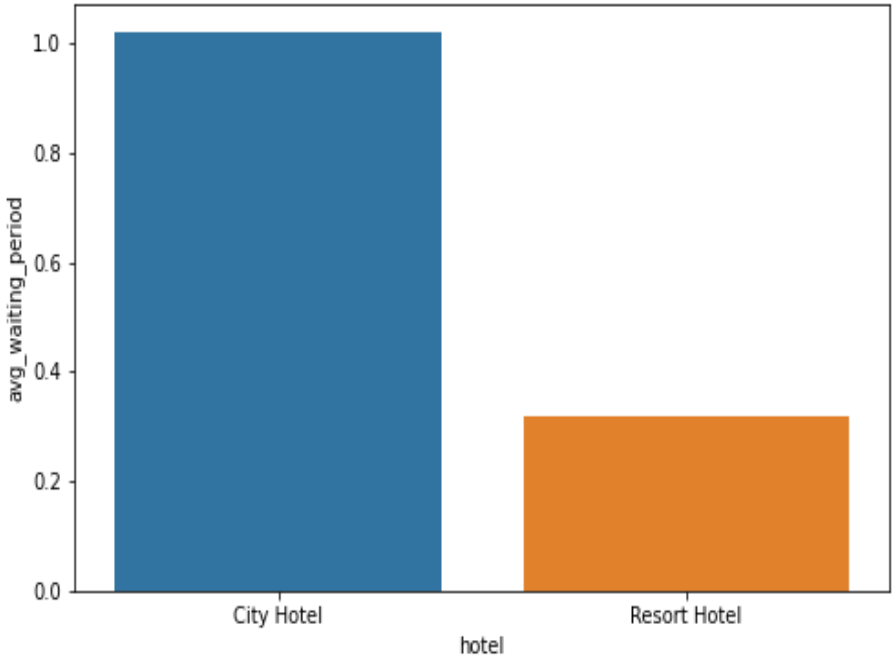
Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.



City hotel has slightly higher median lead time. Also median lead time is significantly higher in each case, this means customers generally plan their hotel visits way to early.

Exploratory Data Analysis





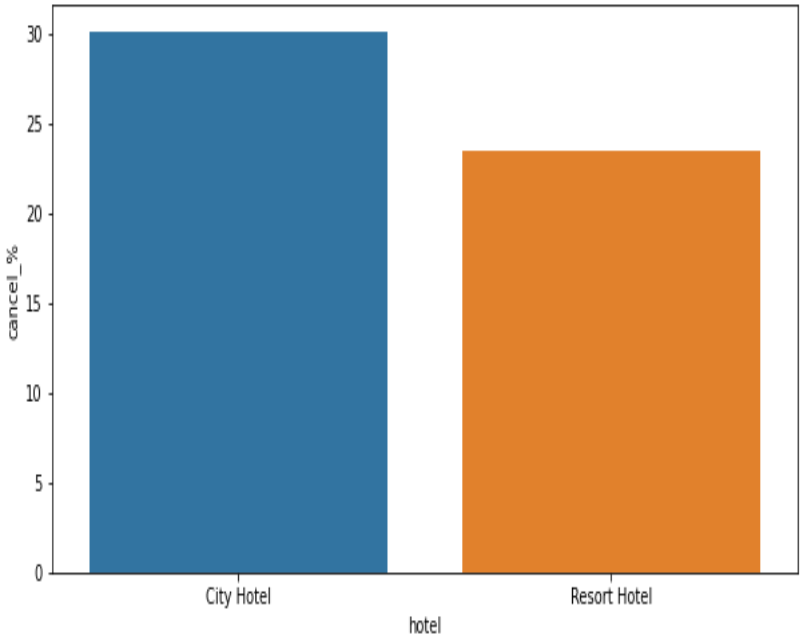
Most common stay length is less than 4 days and generally people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.

City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.

Exploratory Data Analysis

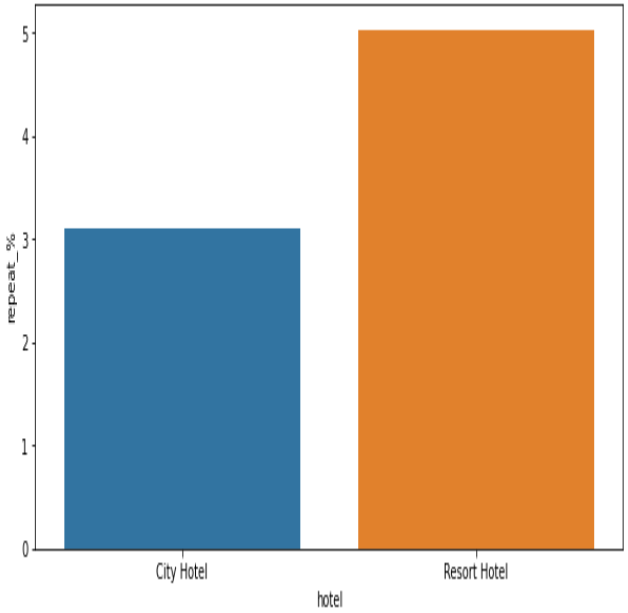
```
# Selecting and counting number of cancelled bookings for each hotel.
cancelled_data = df[df['is_canceled'] == 1]
cancel_grp = cancelled_data.groupby('hotel')
D1 = pd.DataFrame(cancel_grp.size()).rename(columns = {0:'total_cancelled_bookings'})

# Counting total number of bookings for each type of hotel
grouped_by_hotel = df.groupby('hotel')
total_booking = grouped_by_hotel.size()
D2 = pd.DataFrame(total_booking).rename(columns = {0: 'total_bookings'})
D3 = pd.concat([D1,D2], axis = 1)

# Calculating cancel percentage
D3['cancel_%'] = round((D3['total_cancelled_bookings']/D3['total_bookings'])*100,2)
D3
```

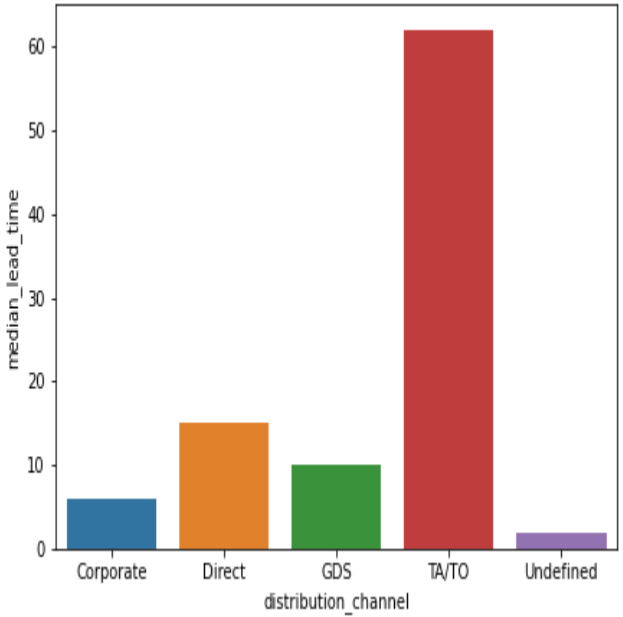
hotel	total_cancelled_bookings	total_bookings	cancel_%
City Hotel	16034	53273	30.10
Resort Hotel	7974	33956	23.48

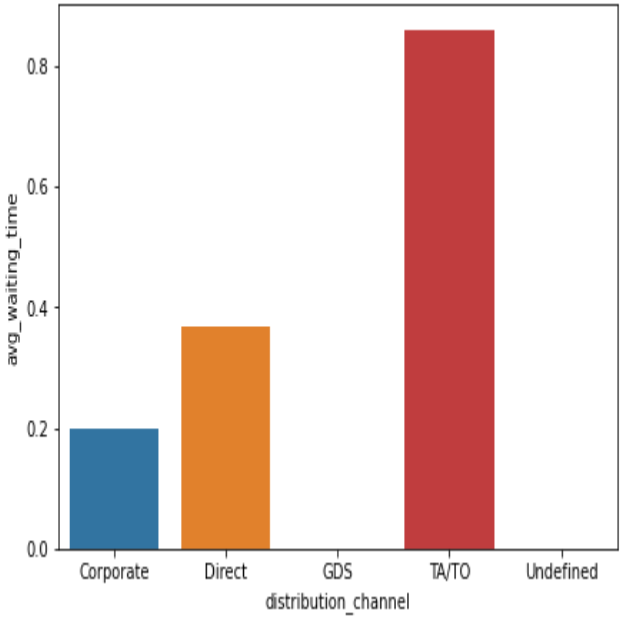


Almost 30 % of City Hotel bookings got canceled.

Exploratory Data Analysis





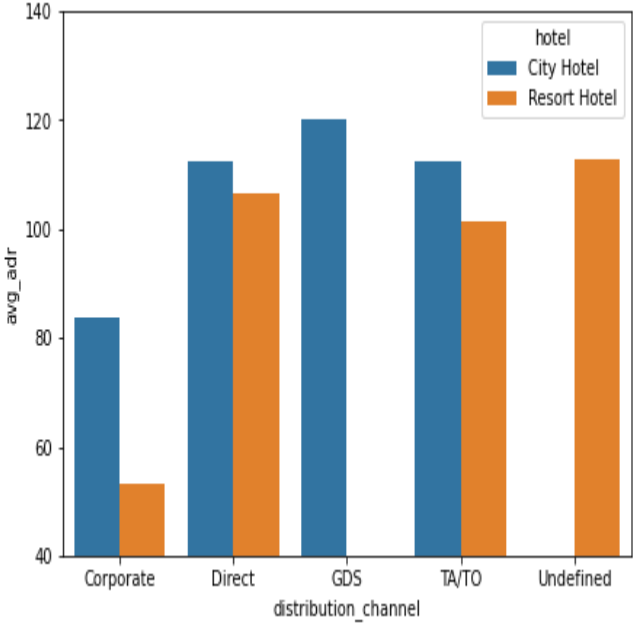


Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than City Hotel.

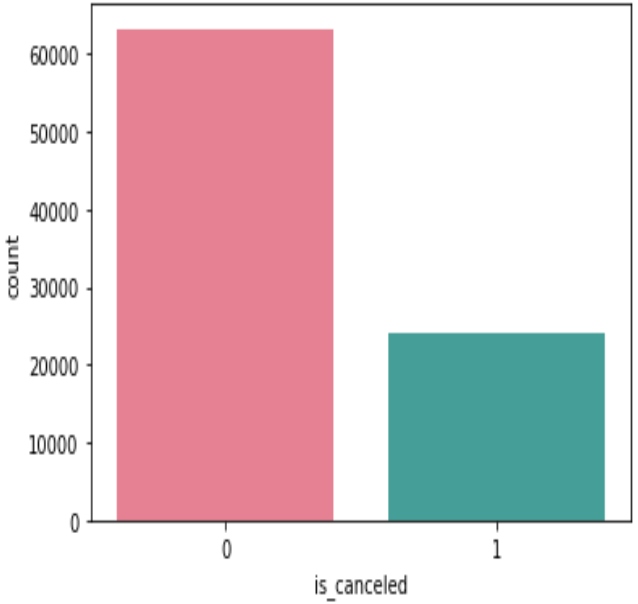
TA/TO is mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred.

While booking via TA/TO one may have to wait a little longer to confirm booking of rooms.

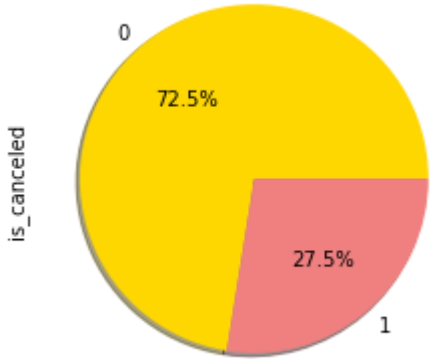
Exploratory Data Analysis



Resort hotel has more revenue generating deals by direct and TA/TO channel. Resort Hotel need to increase outreach on GDS channel to increase revenue.

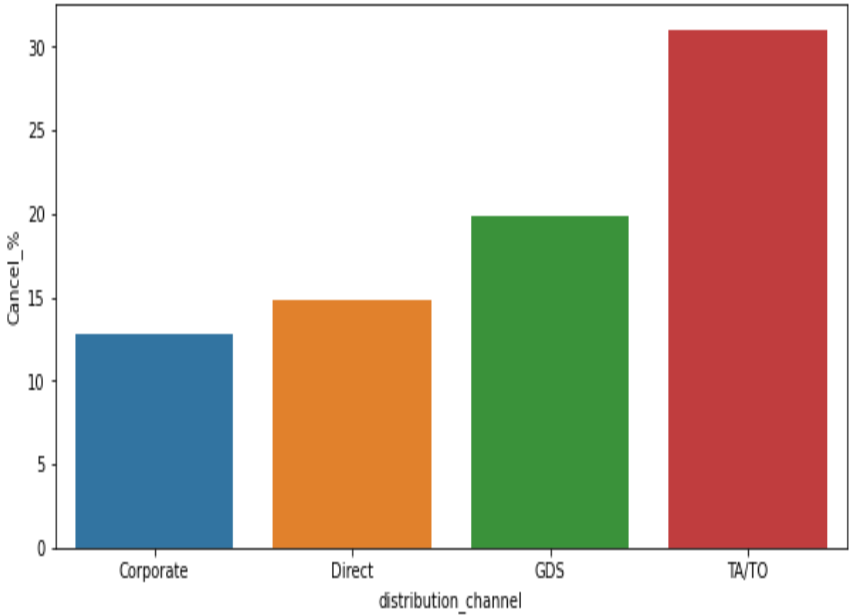


Majority of bookings were not canceled, still some half of the bookings were canceled

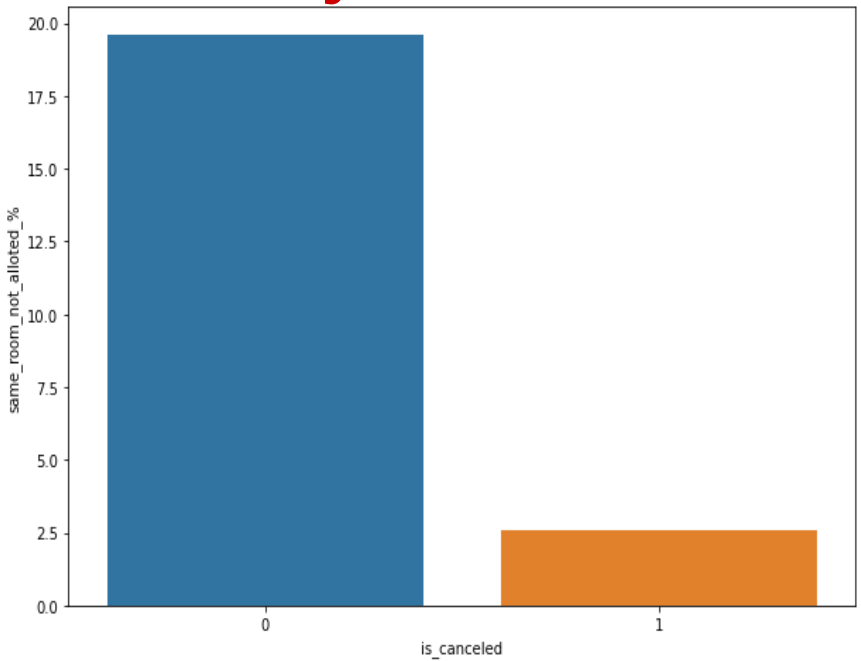


According to the pie chart, 73% of bookings were not canceled and 27% of the bookings were canceled at the Hotel.

Exploratory Data Analysis

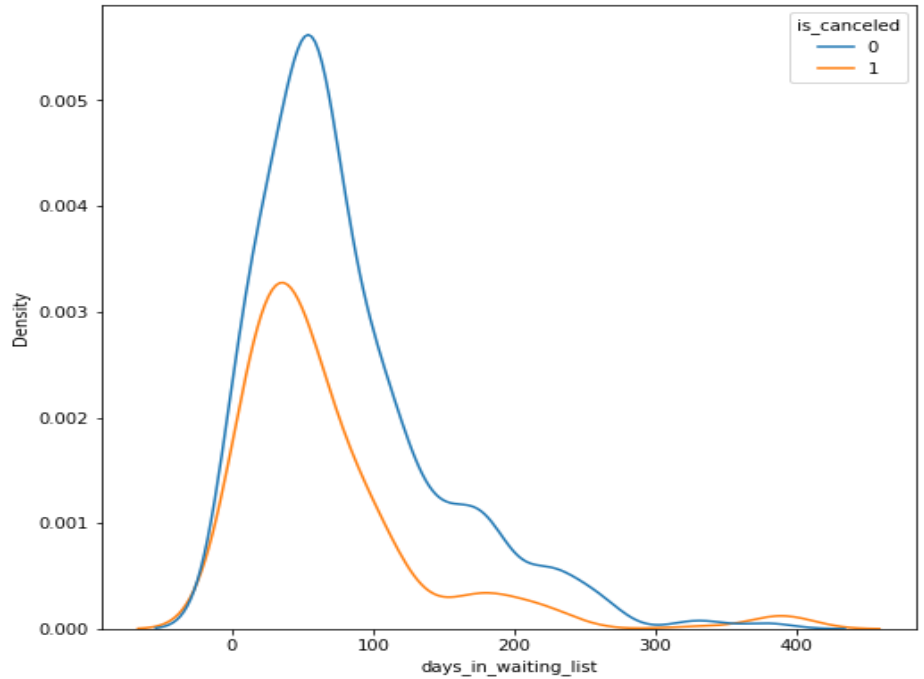


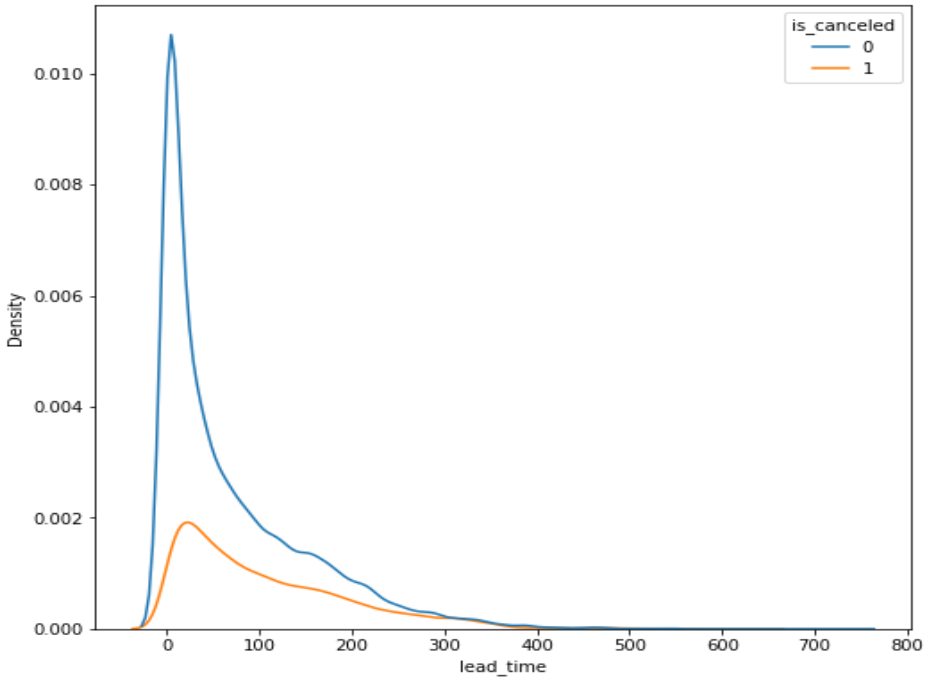
TA/TO has highest booking cancellation %. Therefore, a booking via TA/TO is 30% likely to get cancelled.



We see that not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting different room as demanded.

Exploratory Data Analysis

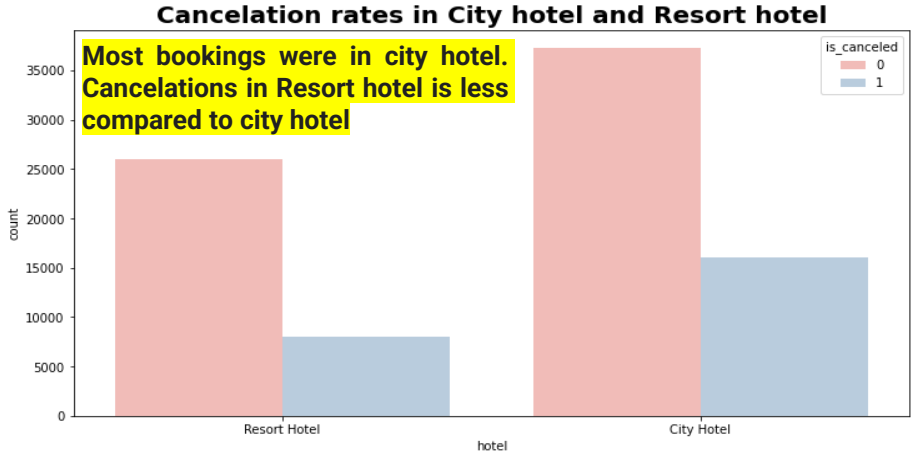


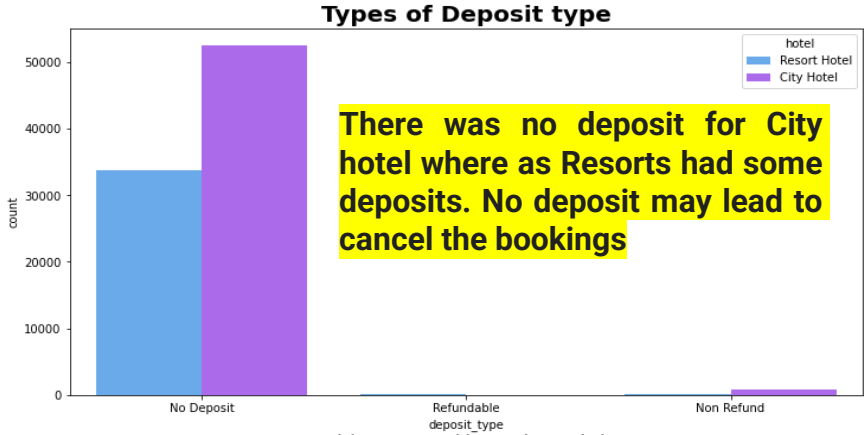


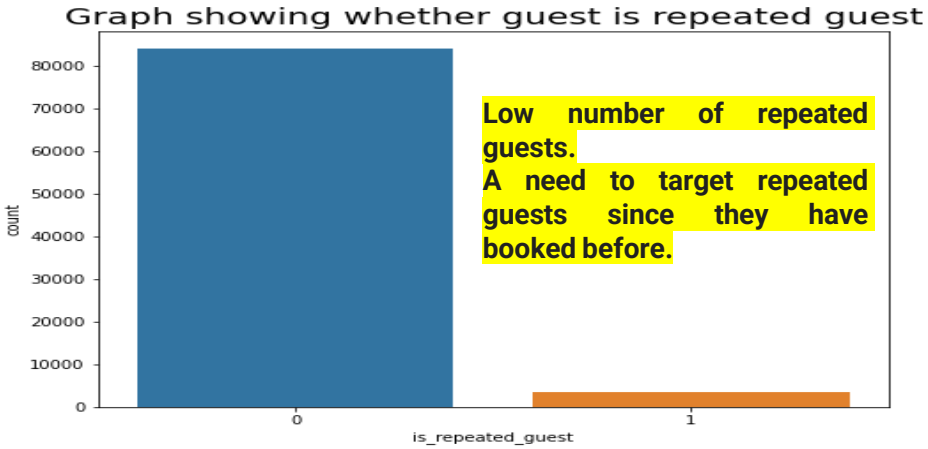
We see that most of the bookings that are cancelled have waiting period of less 150 days but also most of bookings that are not cancelled also have waiting period less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings.

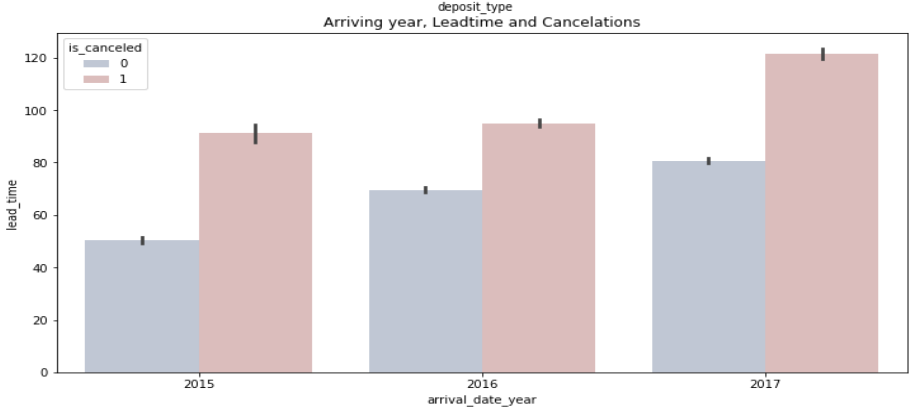
Also, lead time has no affect on cancellation of bookings, as both curves of cancelation and not cancelation are similar for lead time too.

Exploratory Data Analysis



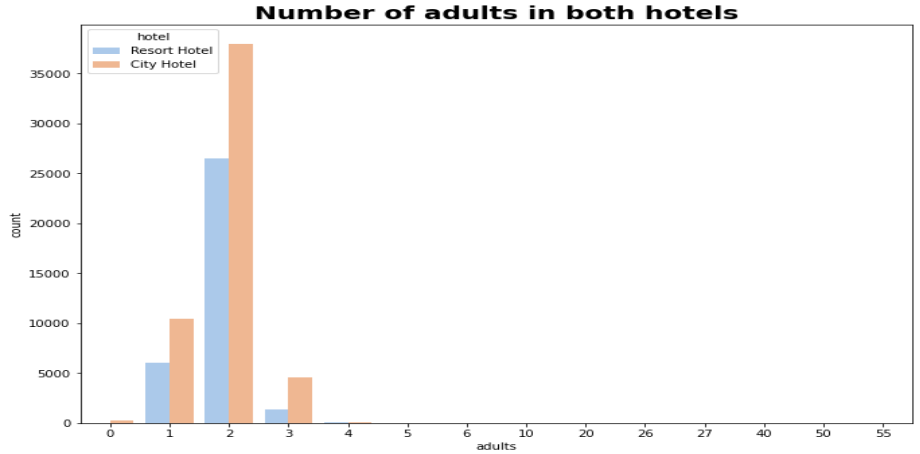


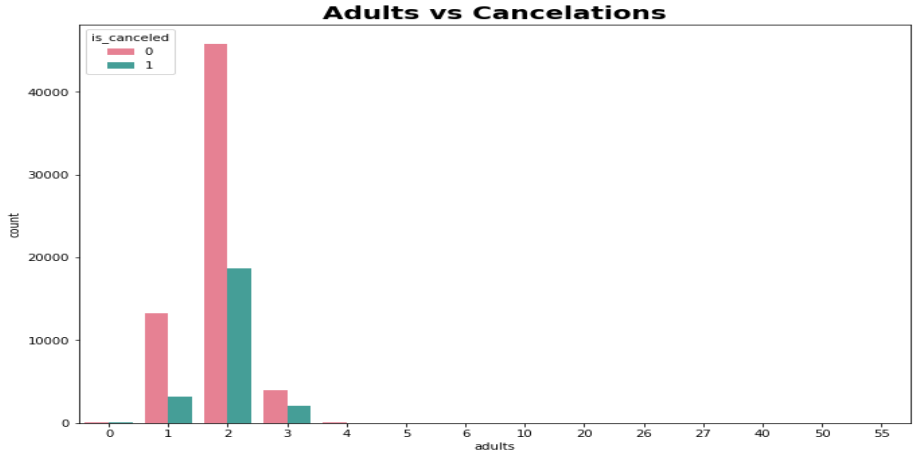


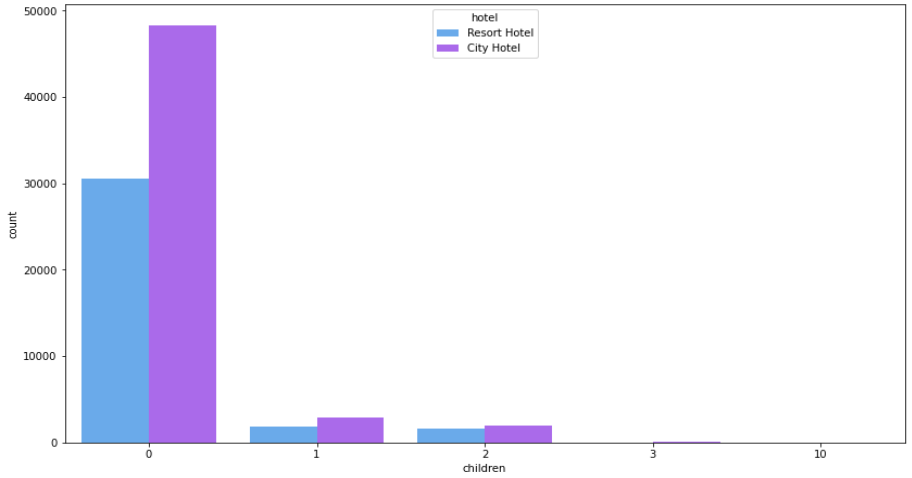


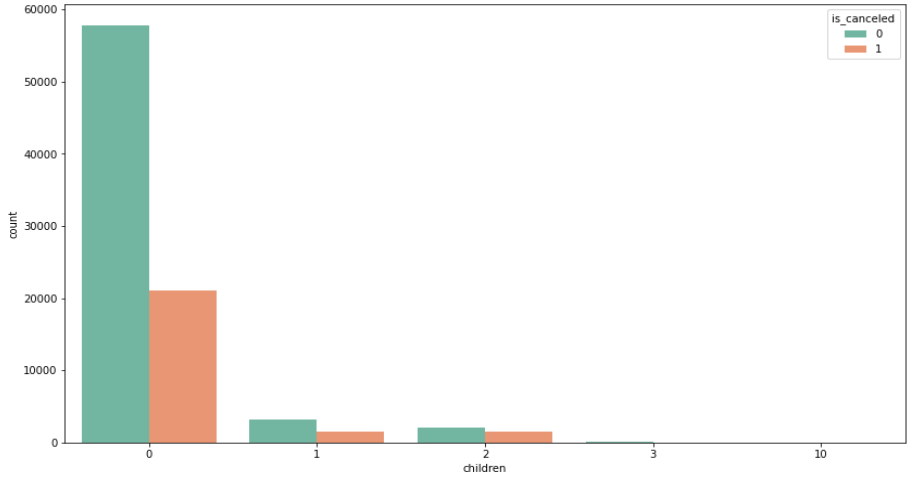
For all the 3 years, bookings with a lead time less than 100 days have fewer chances of getting canceled, and lead time more than or near 100 days have more chances of getting canceled.

Exploratory Data Analysis

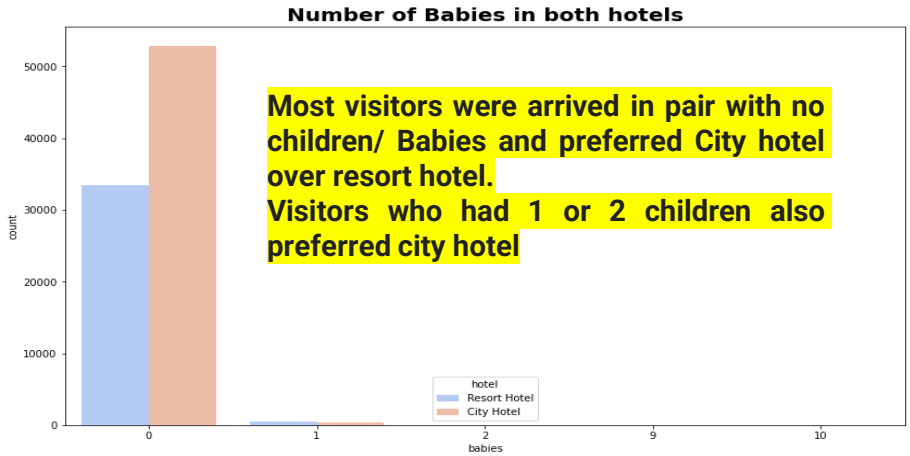


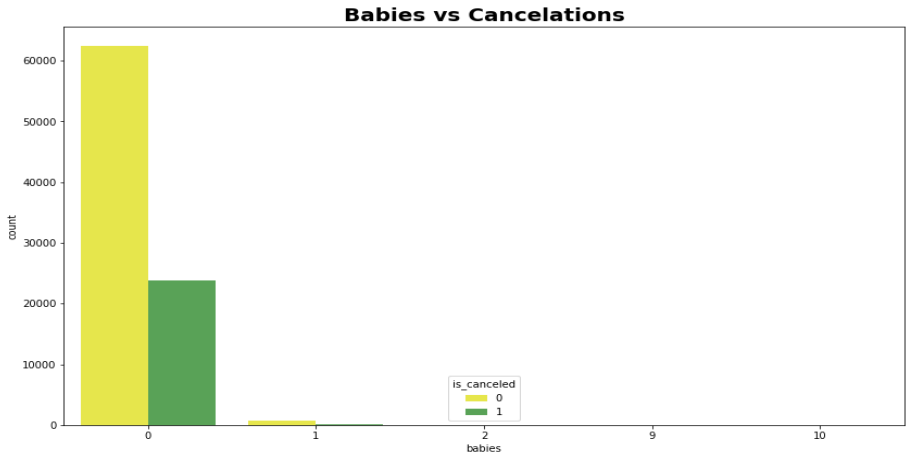


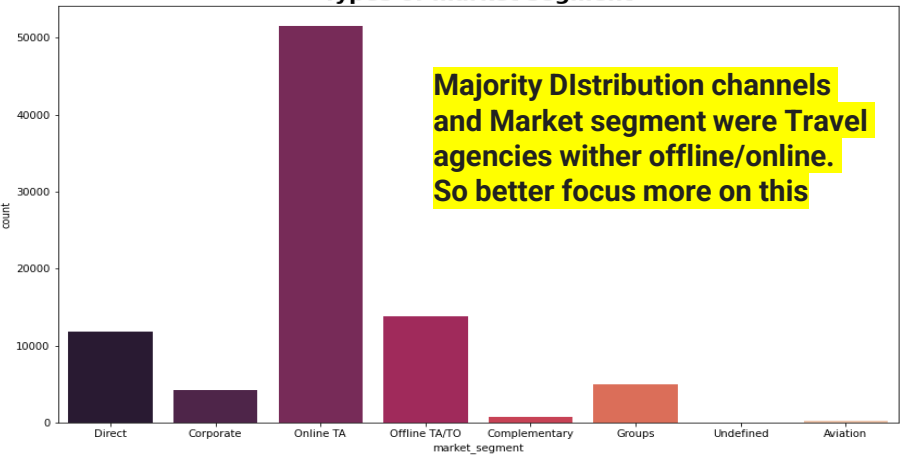


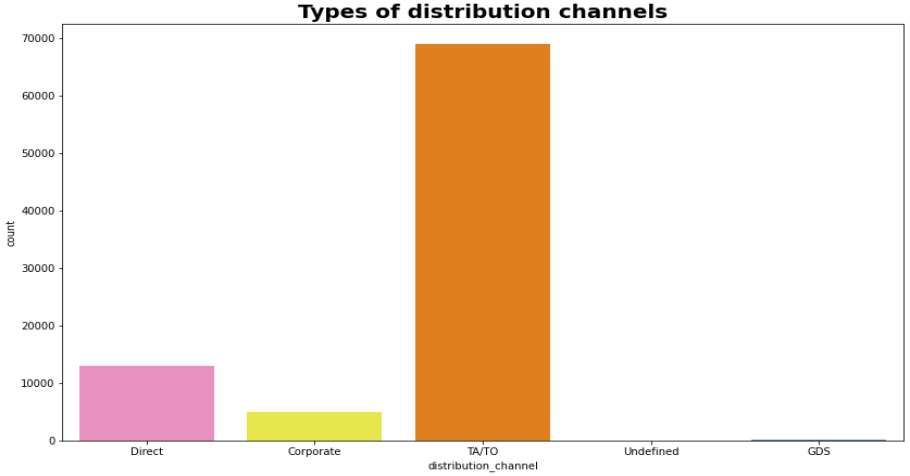


Exploratory Data Analysis

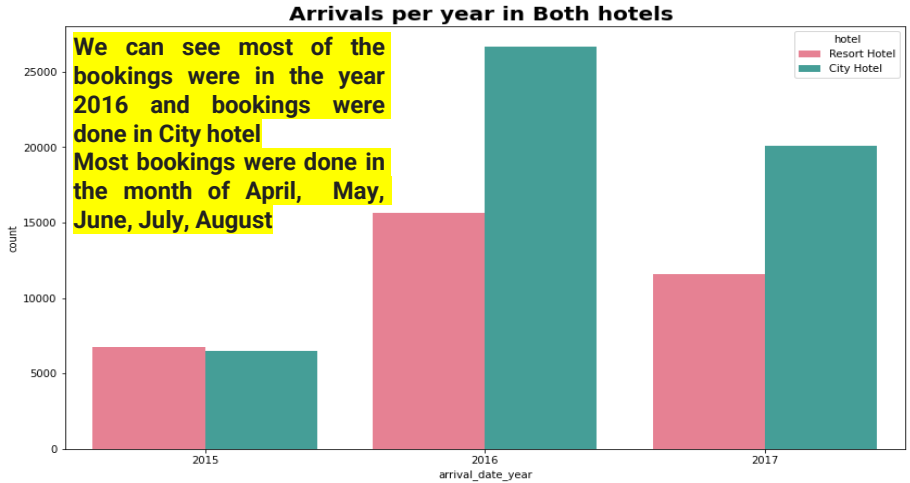


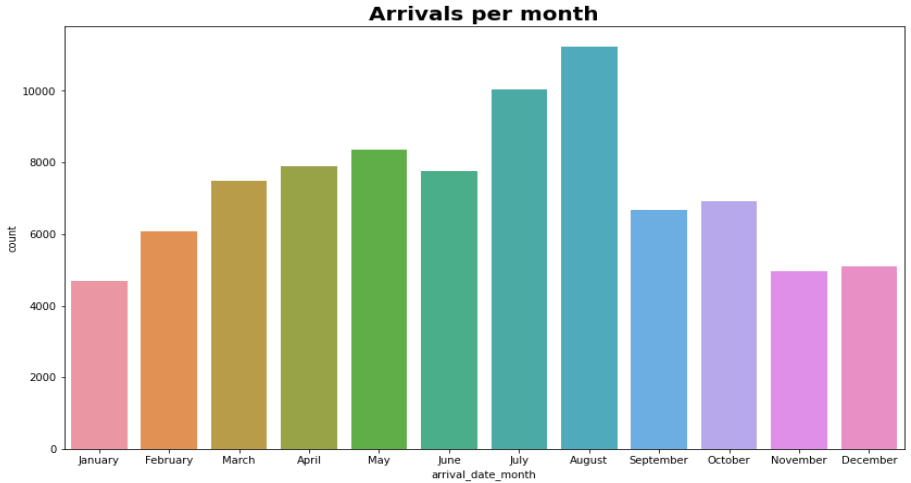


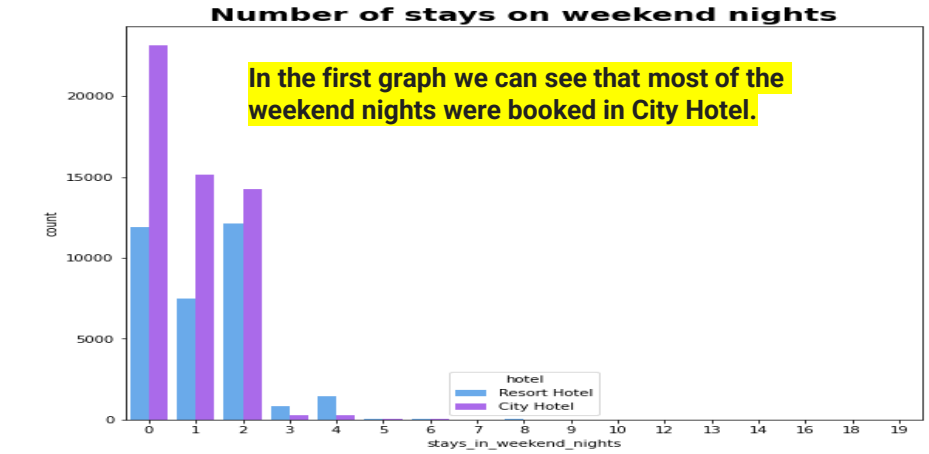


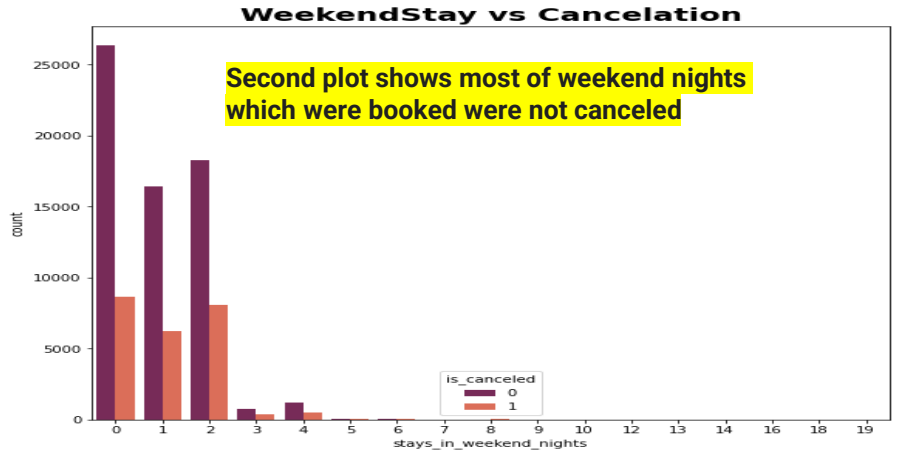


Exploratory Data Analysis

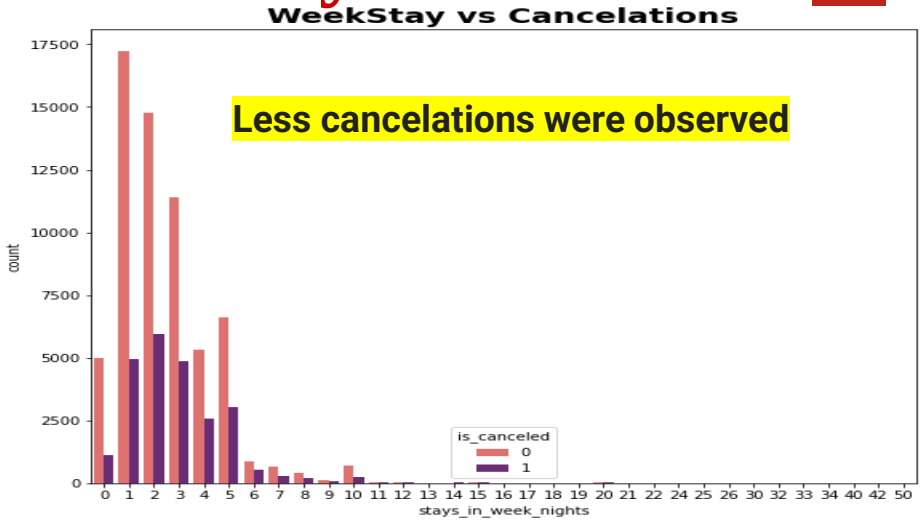


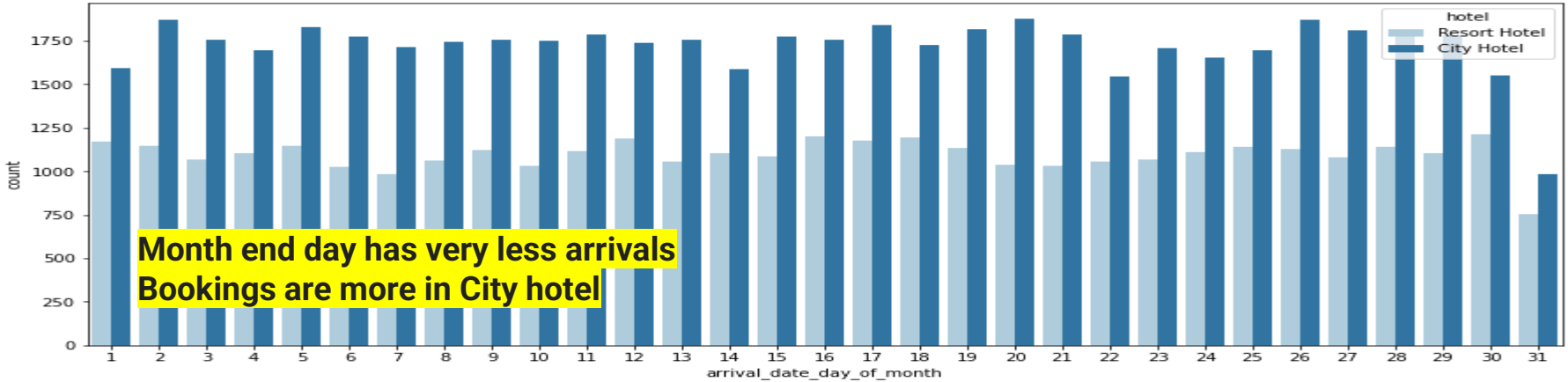




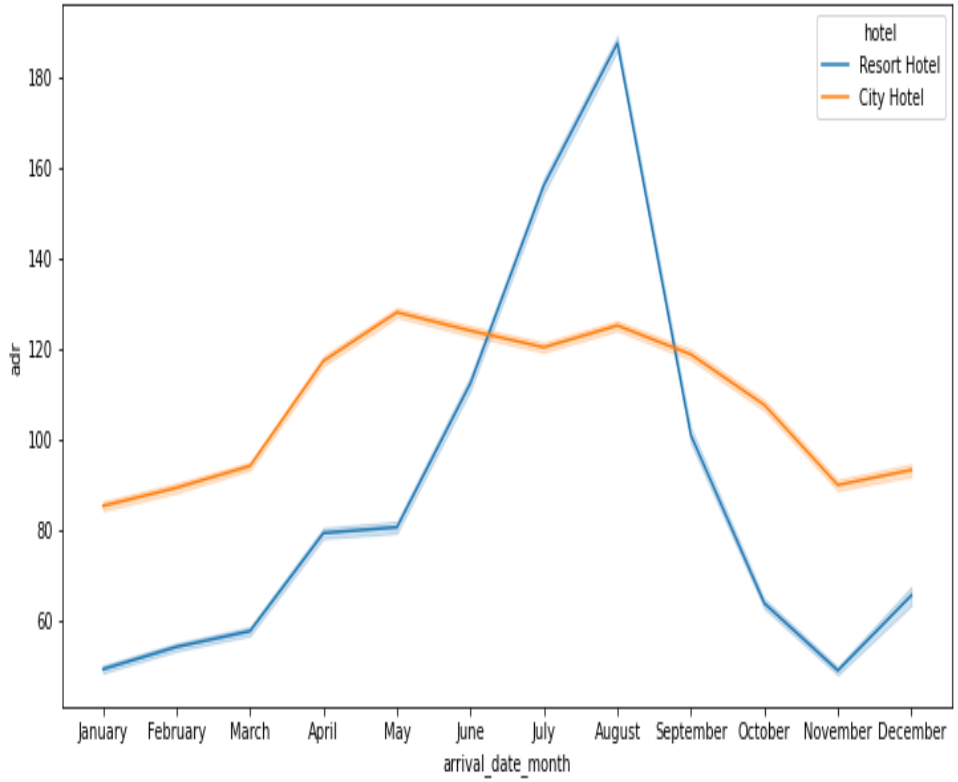


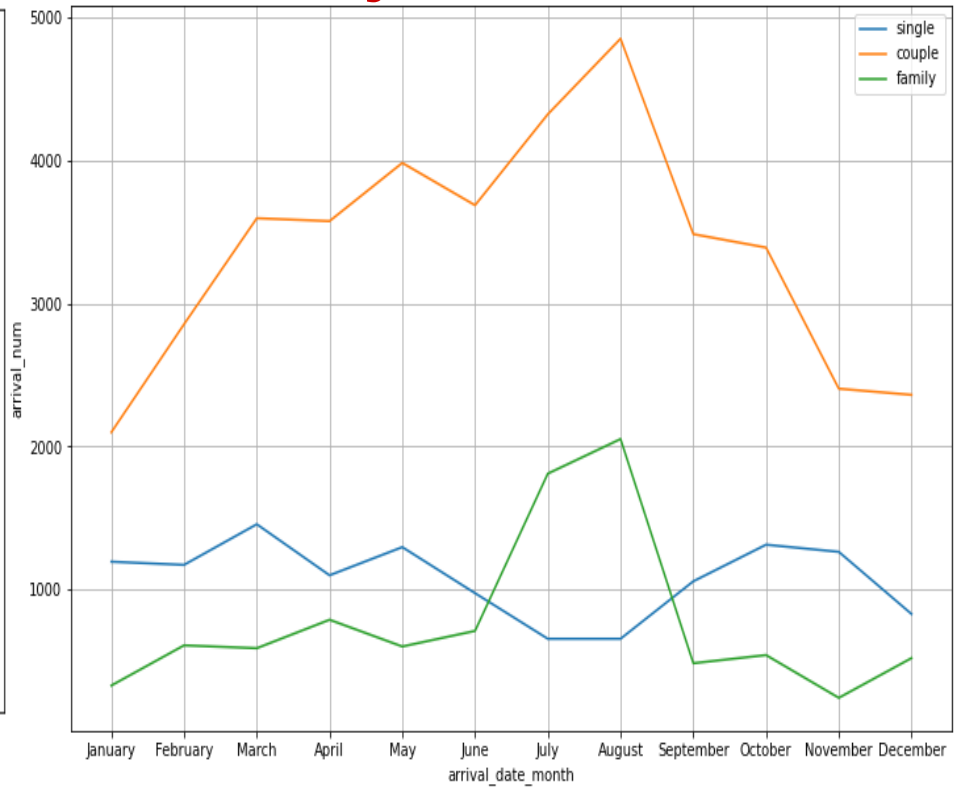
Exploratory Data Analysis





Exploratory Data Analysis





For Resort Hotel, ADR is more expensive during July, August & September and for City Hotel, ADR is slightly more during April, May & August.

Mostly bookings are done by couples(although we are not sure that they are couple as data doesn't tell about that).
It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.

EDA - Data Visualization Techniques

- **Scatter Plot**

  The scatterplot is a plot with many data points. It is one of the many plots seaborn can create.
  Seaborn is a Python module for statistical data visualization. Seaborn can create this plot with the scatterplot() method. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis. Matplotlib can create this plot with plt.scatter() method.

- **Histogram Plot**

  A histogram is basically used to represent data in the form of some groups. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency. The hist() function is used to compute and create a histogram.

- **Bar Plot**

  A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. It can be created using the bar() method.

- **Pie Plot**

  With Pyplot, we can use the pie() function to draw pie charts.

- **Line Plot**

  Line Plot in Seaborn plotted using the lineplot() method. In this, we can pass only the data argument also.

- **Count Plot**

  A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. **seaborn.countplot()** method is used to Show the counts of observations in each categorical bin using bars.

- **KDE Plot**

  Kdeplot is a Kernel Distribution Estimation Plot which depicts the probability density function of the continuous or non-parametric data variables i.e. we can plot for the univariate or multiple variables altogether. Using the Python Seaborn module, we can build the Kdeplot with various functionality added to it.

Conclusion

- Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
- We also realize that the high rate of cancellations can be due high no deposit policies.
- We should also target months between May to Aug. Those are peak months due to the summer period.
- Majority of the guests are from Western Europe. We should spend a significant amount of our budget on those area.
- Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
- Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the actual arrival.
- Average Daily Rate (ADR) is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.
- Low number of repeated guests. A need to target repeated guests since they have booked before.
- Most of the bookings that are cancelled have waiting period of less 150 days but also most of bookings that are not cancelled also have waiting period less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings.
- Travel Agencies are mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred.

Conclusion

- GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via TA/TO. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals. Resort hotel has more revenue generating deals by direct and TA/TO channel. Resort Hotel need to increase outreach on GDS channel to increase revenue.
- As length of total stay increases the adr decreases. This means for longer stay, the better deal for customer can be finalized.
- For Resort Hotel, ADR is more expensive during July, August & September and for City Hotel, ADR is slightly more during April, May & August.
- There is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.

References

1. Matplotlib.org
2. Stacksoverflow.com
3. Geeks for Geeks
4. Kaggle.com
5. analyticsvidhya.com