Hotel Booking Analysis

Capstone Project

Team Person: Arshad Aafaq D, Kaveri Shende, Sakshi Chaturvedi, Vikas Kumar, Yogesh Shivraj Agre

---

Abstract:

First, it's important to understand some basic terms related to hotel booking. For example, the main factors I usually take into account are the cost per night, the hotel's proximity to restaurants and attractions, the availability of free breakfasts, the view from the room, the cleanliness of the room, and of course, the presence of free Wi-Fi. We are able to learn about various booking types using this dataset (i.e. type of hotel, cancelation, duration of stay, types of visitors, types of booking, booking, parking etc). We have carried out data analysis at every level of the dataset, from straightforward data visualizations to intricate multivariate analysis, in order to glean important industry insights. EDA is the initial and crucial step in resolving any data science issue. Additionally, it offers insightful information about the pattern and content it has to express. It is a way for analyzing data sets to highlight their key features, frequently using visual techniques. The greatest EDA provides insightful information about your data. We have adhered to every procedure that is necessary when conducting exploratory data analysis (EDA). After cleaning the data, we created a number of graphs that illustrate the relationships between various attributes using several Python tools, including matplotlib and seaborn. With it, we were able to successfully get some useful insights that can aid the sectors in securing the hotel booking market.

Key Words: EDA hotel booking, Python, Pandas, NumPy, Data Visualization, matplotlib, seaborn, Exploratory Data Analysis.

---

1. Introduction:

Hotel reservations can be made online. We will provide the user with a list of hotels based on the user's selection criteria. If there are available rooms in that particular hotel, the user can make a reservation. There are three types of user roles in the application: administrator, hotel agent, and regular user. The actions available to each user are listed below.

The primary goal of exploratory data analysis is to understand guest trends and behaviour in relation to hotel reservations. To begin, we must understand the importance of each data feature. The data table is made up of 32 columns and 119,390 rows. The data set includes columns such as hotel type, is canceled, arrival date year, arrival date month, stays in weekend nights, stays in weeknights, country,market segment, distribution channel, and so on, which allowed us to derive significant insights from the data.

2. Problem Statement:

We've been placed in the position of a marketing team for a brand-new hotel company that is curious about the hospitality sector. A dataset for the hotel sector in the area where our firm plans to launch its initial operations has been given to us. Our main objective is to comprehend the market in this particular location and apply data analytic approaches to identify the market's important characteristics. Our secondary purpose is to provide findings regarding important market factors like the cancellation rate, distribution channels, and similar topics based on our data and to provide actionable recommendations.

- what is the count of each type of Hotels?
- In which month maximum hotels were booked?
- What is the booking rate according to the population?
- Which form of distribution do customers prefer most?
- Which hotel will have long-term guests?
- Which type of food is preferred by the guest?
- Which hotel has a higher rate of returning customers?
- which type of hotel is mostly preferred by adults, children, or babies
- Which hotel will have long-term guests?
- Which hotel produces the maximum revenue?
- Which distribution route has given adr the most boost in terms of revenue?
- Which room type has highest adr?
- ADR across different market segment?
- ADR across the different month.
- Which month saw the most canceled reservations?
- Which hotel has the highest cancellation rate, the city or the resort?
- determining which countries have the most hotel cancellations in different type of hotels
- Does longer waiting period causes booking cancellation?
- What is the percentage distribution of required_car_parking_spaces?
- Which type of food is preferred by the guest?

3. Data Summary:

Based on the preliminary evaluation, we discovered that the data was essentially clean, with the exception of a few columns with some missing values. The info () method reveals the following salient facts about the data: -

1. The dataset has a shape of (119390, 32) which means that it contains approximately 1.2 lakh rows and 32 columns.
2. Our Dataset has 4 columns with float64 dtype, 16 columns with int64 dtype, and 12 columns with object dtype.

3. In our Dataset, we observed null values in the following columns:
    - 4 null values in the children column
    - 488 null values in the country column
    - 16,340 null values in the agent column
    - 112,593 null values in the company column

The dataset gives us the names of the following columns:

- hotel
- is_canceled
- lead_time
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- booking_changes
- deposit_type
- agent
- company
- days_in_waiting_list
- Customer_type
- adr
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights
- adults
- children
- babies
- meal
- country
- market_segment
- distribution_channel
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- reserved_room_type
- assigned_room_type
- reservation_status_date

4. Procedures for the Data Analysis:

i. **Formulating the questions:** - Prior to conducting any type of analysis, it is crucial to formulate the questions

we hope to answer with the help of the data. To do this, the space team used a variety of creative thinking strategies, including brainstorming, team discussions, and arguments, to eliminate any potential dataset-related queries.

ii. **Data Summary:** - Based upon the initial assessment we found that the data was pretty much clean except for some missing values in a few columns.

iii. **Cleaning of Dataset: -** As previously indicated, the Alma Better academics gave us a dataset that was essentially error-free from the beginning. By "clean," we mean that there were no incorrect dtypes in the dataframe columns or nested lists or dictionaries as row items. Before continuing with our analysis, we had to account for the fact that it had four variables with null values.

iv. **Analysis of Data:-** We refer to exploratory data analysis as EDA. In this, we examined the dataframe and chose the target variables (Important Columns) on the basis of which we would carry out additional analysis. To get insight into their relationships, we began comparing our target variable with additional independent variables (remaining columns). This improved our understanding of how the many variables impact the target variable.

v. **Data Visualization: -** We utilized the Python packages matplotlib and seaborn to visually exhibit our results after finishing the analysis of our data. To convey difficult findings in an appealing way, we made use of pie charts, bar charts, scatterplots, displots, and many other visual representations. We gained a lot of knowledge regarding the various data analysis visualization technologies that are available.

vi. **Finding Conclusions and Solutions:** - In the end, we distorted each analysis by deriving conclusions from it.

4. Cleaning of Dataset:

We used this technique after loading the dataset, cleaning, organizing, and transforming the raw data into the appropriate format that helps us understand the data. This method assisted us in dealing with the undesired data, generating accurate outcomes, and improving our decision-making.

Data cleansing is an important step before EDA since it removes unclear data that could alter the results of EDA.

We will take the following actions while cleaning the data:

1) Eliminate redundant rows

2) Taking care of missing values.

3) Change the datatypes of the columns.

4) Including significant columns

Here we find that there is the following nulls value in the dataset:

- Children = 4
- Agent = 12193
- Company = 82137
- Country = 452

Despite the limited quantity of null values in our data collection, we processed them by filling them with zeros to produce more accurate findings.

```
#Company and agent values replaced with zero
df_hotel[['company','agent']] = df_hotel[['company','agent']].fillna(0)
```

```
df_hotel['children'].fillna(df_hotel['children'].mean(), inplace = True)
```

```
df_hotel['country'].fillna('others', inplace = True)
```

```
null_detail(df_hotel)
```

After that, we converted this dataset into the data also added some required columns in the dataset.

5. EDA

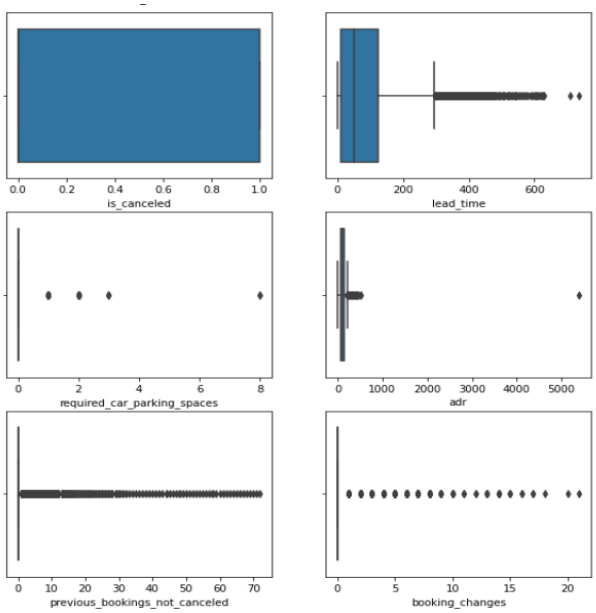
This procedure was carried out by comparing our goal variable, the booking analysis, with other independent variables after loading the dataset. Through this procedure, we were able to identify several features and connections between the goal and the independent variables. We now have a clearer understanding of how each character interacts with the target variable.

We will first find the relationship between the numerical data. 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 won't need to check them for correlation. Also, we have added total_stay and total_people columns. So, we can remove adults, children, babies, stays_in_weekend_nights, stays_in_week_nights columns.

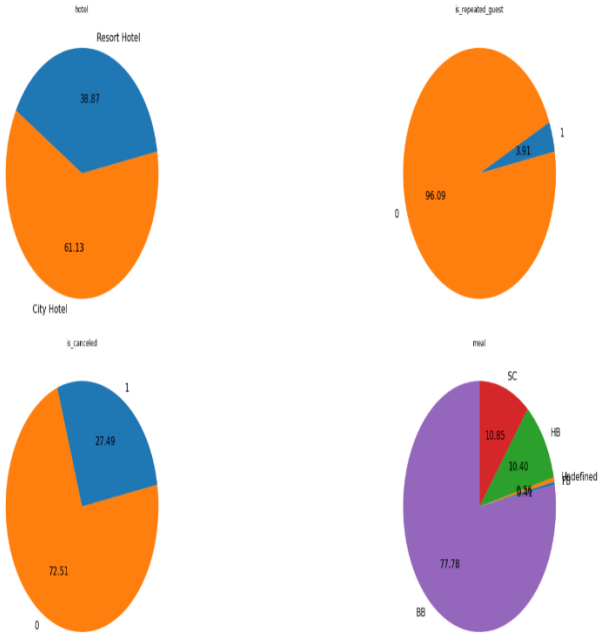
The following graphs and plots were mainly created using Matplotlib and the Seaborn library:

- **The Bar Plots**
- **Histogram.**
- **The scatter plots**
- **Pie Diagram**
- **The line plots**
- **Heatmap**
- **Boxplot**

Plotting a box plot understanding the outlier.



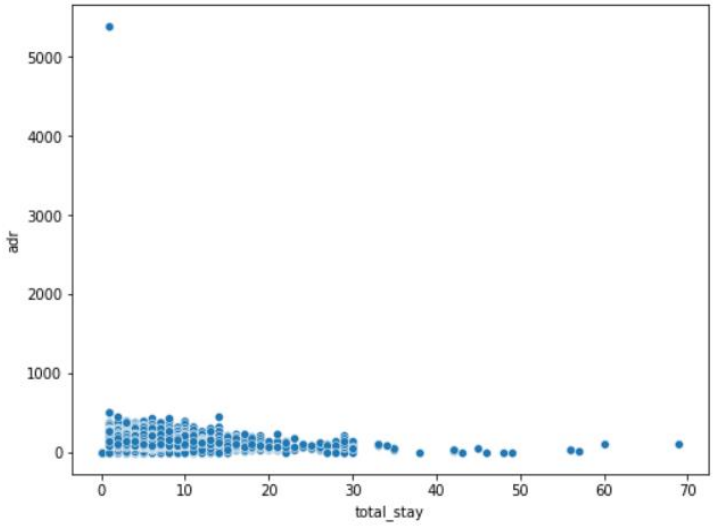
5.1 Deciding some columns to analyze:



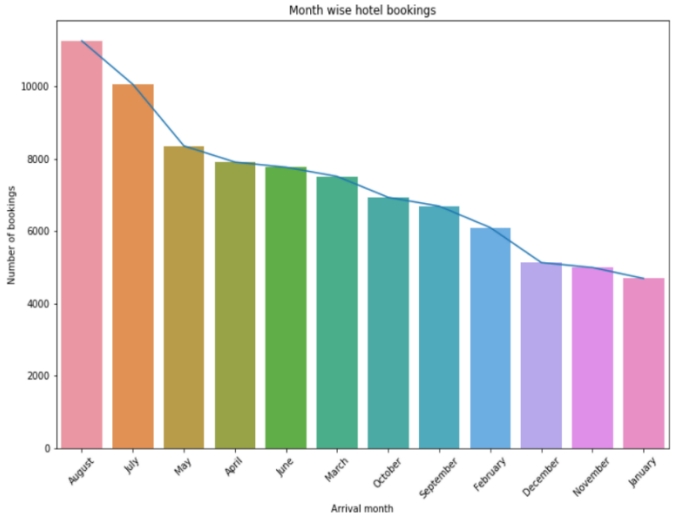
Here I took some columns to analyze by the pie chart. Here we can see that there is visible that in the hotel column there are two types of hotels (city hotel is 62.23% and resort hotel is 38.87%).

Similarly, we can also check in the columns of is_repetated_guests so it indicates that how many percent the guest arrive in the same before or not.

5.2 Scattter Plot for column “adr” w.r.t “total_stay”.

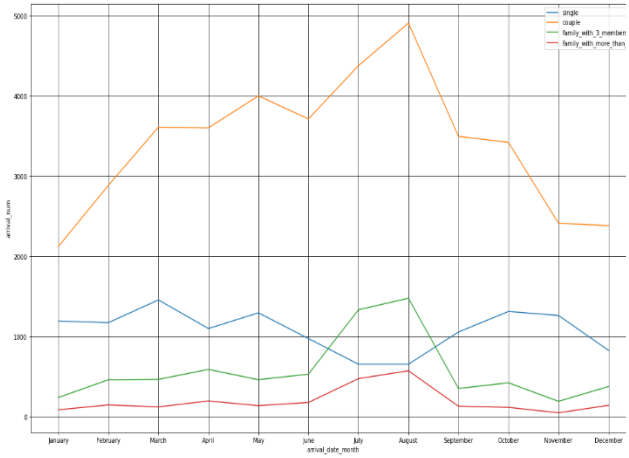


5.3 In which month maximum hotels were booked?



Here, we see the output visualization. Line plots and bar plots are included in the visualization, and it demonstrates that the month of August saw the highest hotel occupancy, but the month of January saw the lowest. However, if we look at the hotel's bookings for each month and use visualization, we can quickly determine which months have the most bookings.

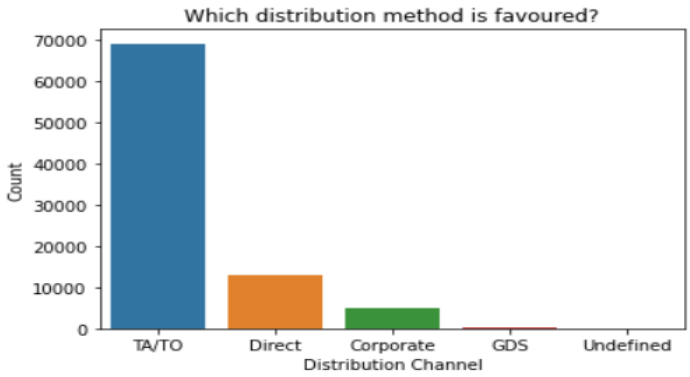
5.4Booking rate according to the population: -



Although it appears that couples made the majority of reservations, we cannot be certain that they are a couple because the data contains no information specifically about couples or families. Additionally, as was

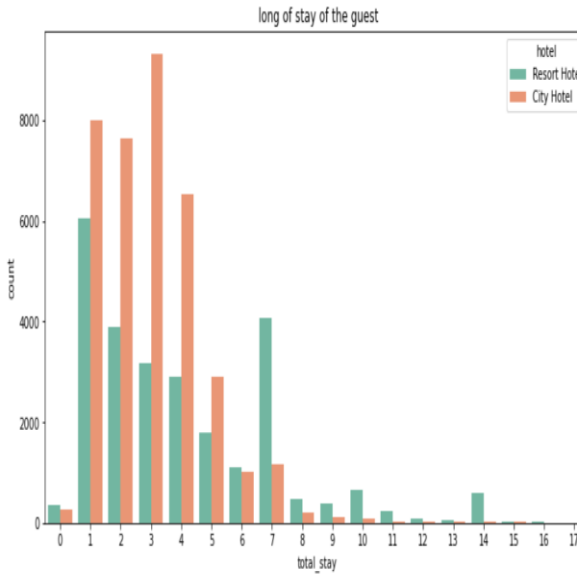
the months of June, July, and August. Bookings for families with three or more members are the least expensive.

5.5 Distribution do customers prefer most:



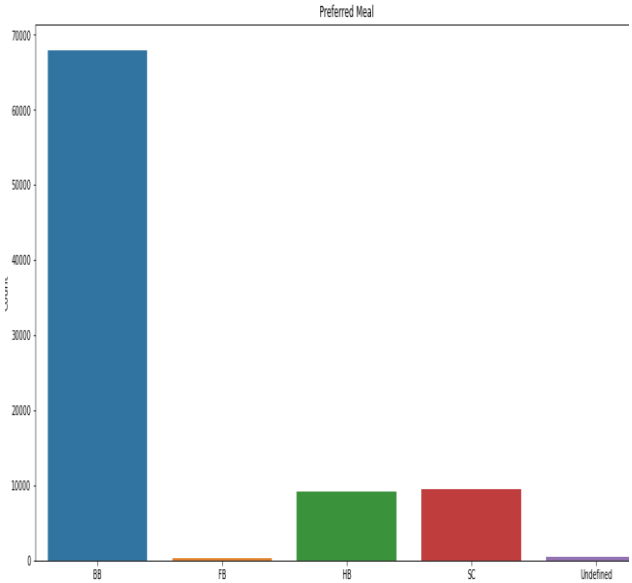
TA/TO are the customers' chosen distribution channels. In order to grow their business, hotels might partner with these agents and operators or promote using them as a medium.

5.6 Which hotel will have long-term guests: -



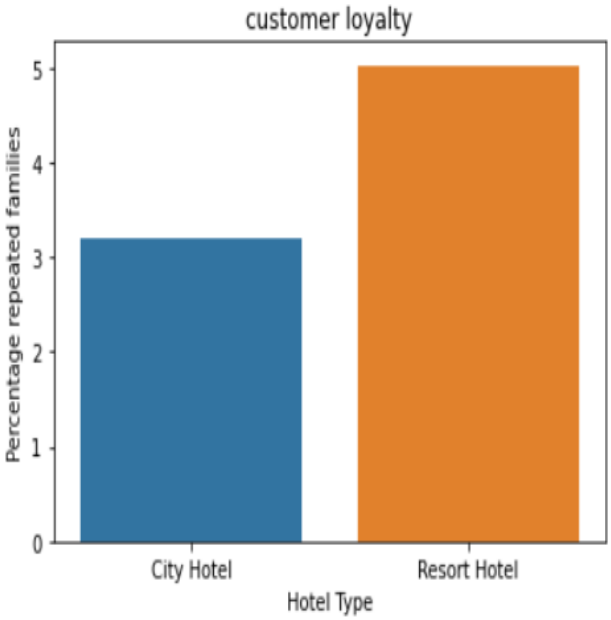
Here we have taken the column of total_stay and count that we observe that most visitors of the resort hotel stayed for one day, however most city hotel guests spent anywhere between one and seven days.

5.7 Food is preferred by the guest: -



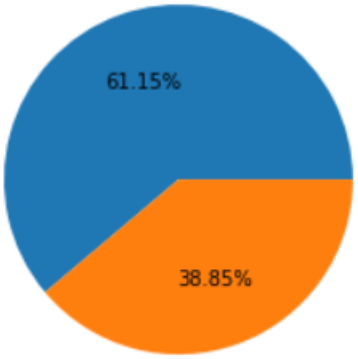
From the above bar plot result, we observed that Meal types in hotels: BB - (Bed and Breakfast) HB- (Half Board) (Half Board) SC- (Supplemental Committee) FB- (Full Board) (Self Catering) As a result, the most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering).

5.8Hotel has a higher rate of returning customers: -

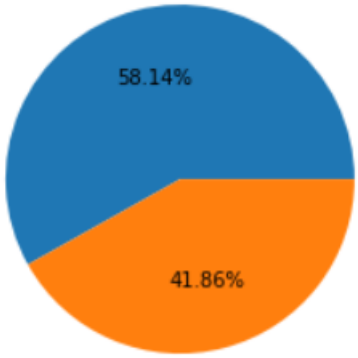


From the above graph it is clear that highest rate of returning customers are from the resort hotel.

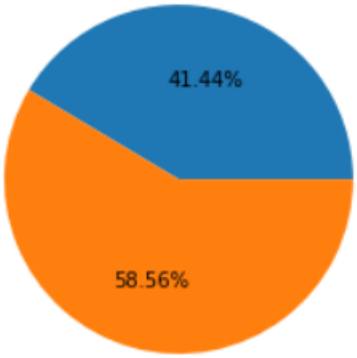
5.9 Type of hotel is mostly preferred by adults, children or babies:-



- In this pie chart, adults' numbers in these hotels are
- City Hotel -100247
- Resort Hotel -63688
- pie chart it is clearly mentioned that 61.15 % of adults prefer city hotels.

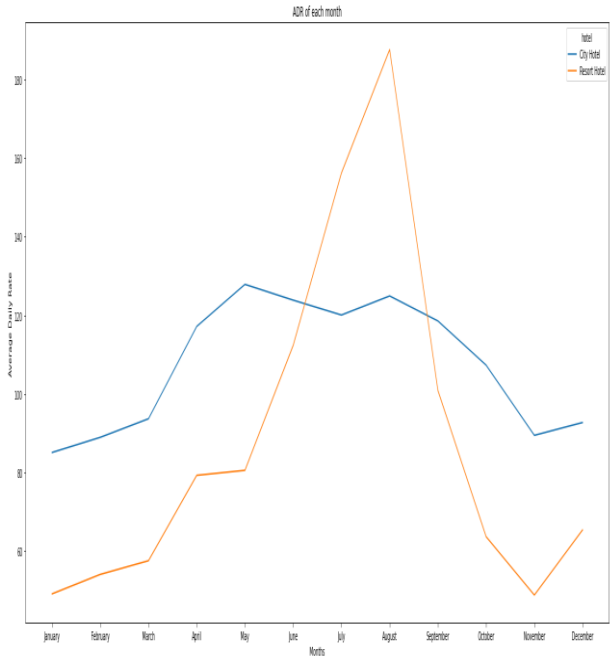


- In this pie chart, Children numbers in these hotels are
- City Hotel -7044
- Resort Hotel -5072
- it is clearly mention that 58.14 % of children prefer city hotel



- In this pie chart, babies numbers in these hotels are
- City Hotel -392
- Resort Hotel -554
- it is clearly mentioned that 41.44% of babies prefer city hotels which are less than resort hotels.

5.10 Hotels have the highest ADR:-

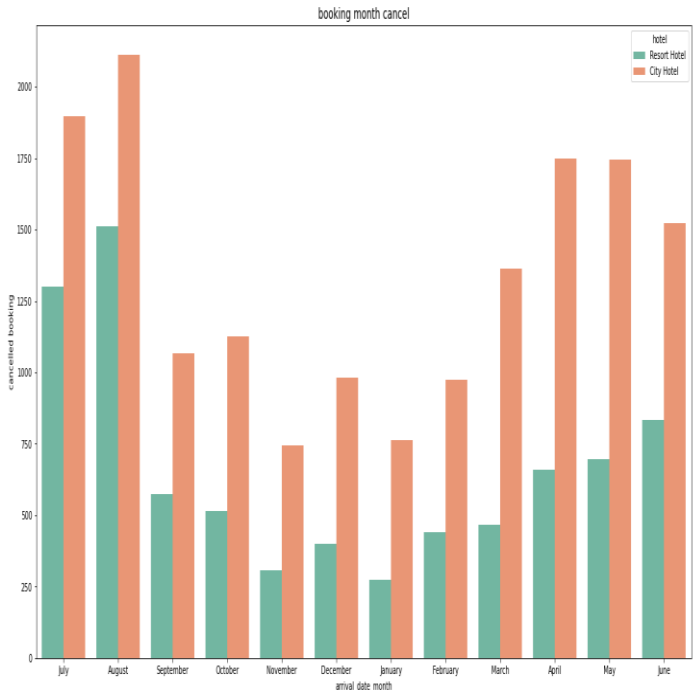


In comparison to City Hotels, the ADR for Resort Hotel is higher in the months of June, July, and August. Perhaps clients/people wish to vacation in resort hotels this summer.

January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.

5.11 Waiting period results in canceled bookings

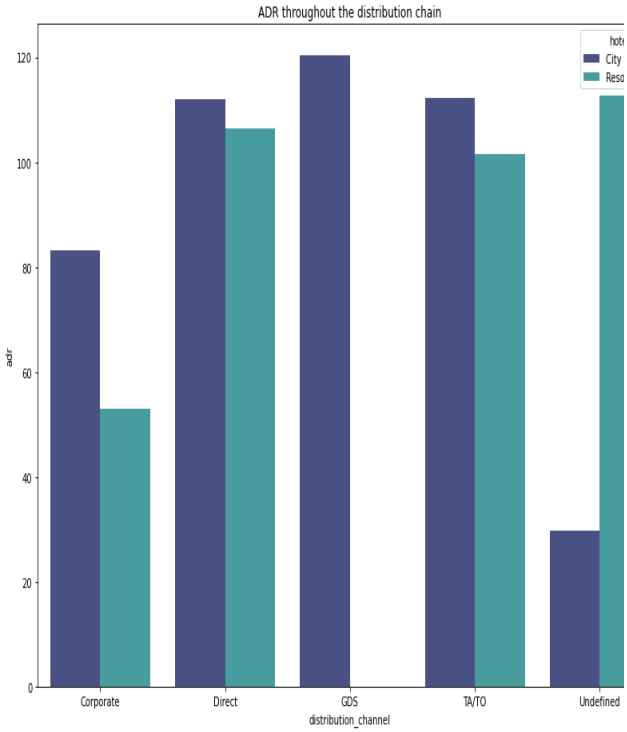
For hotels in cities, the majority of cancellations occurred in the months of October, but for hotels in resort areas, the majority occurred in the month of August. Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotel has greater cancellation of reservations overall.



5.12 Hotel produces maximum revenue :-

According to the above graph, the average revenue of city hotels is higher than that of resort hotels Observation.

5.13 Distribution route has given adr the most boost in terms of revenue?



Corporate - These are companies that help businesses make hotel reservations.

GDS-GDS - serves as a global link between travel agents and suppliers, including hotels and other lodging establishments. It enables automated transactions and provides real-time product, price, and availability data to travel agencies and internet booking engines. Direct - refers to making reservations with specific hotels directly. TA/TO - Bookings are made through travel agents or travel operators. Undefined – Reservations are not defined. Maybe customers made their reservations when they arrived.

Inference

- In both types of hotels, "Direct" and "TA/TO" have contributed to adr about equally.
- GDS made a significant contribution to ads of the "City Hotel" type.
- GDS contributes little to hotel reservations at resorts.

CONCLUSION: -

- There are a total of 2 hotels, a city hotel, and a resort hotel.
- From the above 2 hotels city hotels were preferred feed by the customer.
- Maximum number of bookings were done in the month of august so we can put in more offers during this month.
- As compared to couples, single and family with more than 3 members are less expensive.
- TA/TO distribution channel is more preferred by the customer as compared to others, In order to grow their business, hotels might partner with these agents and operators or promote using them as a medium.
- Most visitors of the resort hotel stayed for one day, however most city hotel guests spent anywhere between one and seven days.
- The most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering). From the above graph, it is clear that the highest rate of returning customers is from the resort hotel.
- From above it is clear that city hotels are mostly preferred by babies, adults and children.
- More than 25000 people, or the majority of the attendees, are from

- Portugal. the average revenue of city hotels is higher than that of resort hotels.
- H type has the highest Average daily rate followed by G type.
- January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.
- For hotels in cities, the majority of cancellations occurred in the month of October, but for hotels in resort areas, the majority occurred in the month of August. Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotels had greater cancellations of reservations overall.
- For hotels in cities, the majority of cancellations occurred in the month of October, but for hotels in resort areas, the majority occurred in the month of August. Additionally, similar cancellations of reservations for both hotel types occurred in the month of August. City hotels had greater cancellations of reservations overall.

- About 30% of hotel reservations for city hotels and 24% for resort hotels are canceled.
- here we can say that PRT country made the highest number of cancelations under the city hotel
- here we understand that the PRT country has made a large number of cancelations in resort-type hotels.
- There is no direct correlation between a longer waiting period and booking cancellation, as can be seen from the fact that the majority of reservations that had less than 100 days on the waiting list were canceled. However, reservations that had more than 100 days on the waiting list were also canceled at a slightly higher rate.
- 91.6 % of guests do not require a parking space. only 8.3 % of guests required parking space.
- As a result, the most popular meal type among guests is BB (Bed and Breakfast), followed by HB (Half Board) and SC (Self Catering).

References:

1. GeeksforGeeks
2. Analytics Vidhya
3. AlmaBetter Class material
4. Pandas and Numpy libraries
5. Stack overflow
6. YouTube
7. Researchgate.net
8. W3schools.com