PROJECT SUMMARY

In this project, the primary objective was to perform exploratory data analysis (EDA) on the Hotel Booking dataset. The project began by importing essential libraries, including NumPy, Pandas, Matplotlib, and Seaborn. The dataset was then loaded using Google Drive, and the dataframe named **"hotel _df"** was created, containing 119390 rows and 32 columns. For the safe side, I have created a copy of the dataset named **"hotel_df1"** and done the manipulations on that data. The dataset provided information about city and resort hotels. The dataset contains the following columns:

- **Hotel:** Type of hotel (City or Resort).
- **is_cancelled:** If the booking was cancelled (1) or not (0).

.

- **lead_time:** Number of days before the actual arrival of the guests.
- **arrival_date_year:** Year of arrival date.
- **arrival_date_month:** Month of arrival date.
- **arrival_date_week_number:** Week number of year for arrival date.
- **arrival_date_day_of_month:** Day of arrival date.
- **stays_in_weekend_nights:** Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests.
- **stays_in_weel_nights:** Number of weeknights (Monday to Friday) spent at the hotel by the guests.
- **adults:** Number of adults among the guests.
- **children:** Number of children.
- **babies:** Number of babies.
- **meal:** Type of meal booked.
- **country:** country of the guests.
- **market_segment:** Designation of the market segment.
- **distribution_channel:** Name of booking distribution channel.
- **is_repeated_guest:** If the booking was from a repeated guest (1) or not (0)
- **previous_cancellation:** Number of previous bookings that were cancelled by the customer prior to the current booking.

- **previous_bookings_not_cancelled:** Number of previous bookings not cancelled by the customer prior to the current booking.
- **reserved_room_type:** Code from room type reserved.
- **assigned_room_type:** Code of room type assigned.
- **booking_changes:** Number of changes made to the booking.
- **deposit_type:** Type of deposit made by the guest.
- **agent:** ID of the travel agent who made the booking.
- **company:** ID of the company that made the booking.
- **days_in_waiting_list:** Number of the days the booking was on the waiting list.
- **customer_type:** Type of customer, assuming one of four categories.
- **ADR:** Average daily rate.
- **required_car_parking_spaces:** Number of car parking spaces required by the customer.
- **total_of_special_requests:** Number of special requests made by the customer.
- **reservation_status:** Reservation status (Canceled, check-out or no-show)
- **reservation_status_date:** The date at which the last reservation status was updated.

DATA CLEANING AND MANIPULATION

Duplicate values

The dataset has 31994 duplicate values. So, these duplicate values are removed from the dataset using the drop_duplicates() function. After dropping the duplicate values shape of the dataset becomes 87396 rows and 32 columns.

Missing values/null values

The given dataset has 4 columns company, agent, country and children having missing values, so these values are replaced by using the fillna() function.

Addition of new columns

Total_people and Total_stay are two columns that are added to the given dataset. Some rows are removed from columns adults, children and babies which have 0 values.

EXPLORATORY DATA ANALYSIS

The EDA is done by using 3 analyses Univariate, Bivariate and Multivariate analysis. For the data visualization following charts are used:

1. Pie chart
2. Barplot
3. Countplot
4. Lineplot
5. Heatmap

Univariate analysis:

In univariate analysis following questions are tried to solve:

1. Which type of hotel is most preferred by the guests?
2. Which agent made the most bookings?
3. What is the percentage of repeated guests?
4. What is the most preferred room type by the customers?
5. What type of food is most preferred by the guests?
6. In which month most of the bookings happened?
7. Which distribution channel is most used for hotel booking?
8. Which year had the highest bookings?

Conclusion:

1. The City Hotel has more bookings.
2. Agent No. 9 made the most of the bookings.
3. There are very few guests booking for the same hotel again.
4. Type A rooms are the most preferred rooms.
5. BB-type food is the most preferred food.
6. August month has a maximum number of bookings.
7. The most used distribution channel is TA/TO channel.
8. 2016 has the highest bookings.

Bivariate and Multivariate analysis:

In bivariate and multivariate analysis following questions are tried to solve:

1. Which hotel type has the highest ADR?
2. Which hotel has a longer waiting time?
3. Which distribution channel contributed more to ADR to increase income?
4. What is the optimal stay length in both types of hotels?
5. The relationship between the repeated guests and previous bookings not cancelled.
6. Relationship between ADR and the total number of people?
7. Relationship between ADR and the total stay?

Conclusion:

1. The City Hotel has the highest adr.
2. City hotel has longer waiting time.
3. **GDS distribution channel** contributed more to ADR in city hotels.
4. The optimal stay length in both hotel types is less than **7 days.**
5. Repeated guests do not cancel their bookings.
6. As the number of people increases adr also increases.
7. There is a **positive** relationship between adr and total_stay and it indicates that most of the guests tend to stay for at least 7 days.

The conclusion from the correlation heatmap:

1. **arrival_date_year** and **arrival_date_week_number** columns have a negative correlation which is **-0.51.**
2. **stays_in_week_nights** and **total_stay** have a positive correlation which is **0.95**.

Overall conclusion:

1. City hotels have almost **60%** of bookings and resort hotels have **40%** of bookings.
2. **Agent no. 9** made the most bookings, and those bookings are 28721.
3. The percentage of repeated guests is just **4%.**
4. **Room type A** is the most preferred room type **46283** guests preferred A room type.
5. **BB type food** is the most preferred food type, and **67907** guests preferred this food.
6. **August** month has a maximum number of bookings, and those bookings are 11242.
7. **TA/TO** distribution channel is the preferred channel, and the bookings are 69028.
8. **2016** has 42313 bookings.
9. City hotels have the highest ADR, and the average ADR is **111.27**.
10. City hotels have longer waiting time means city hotel is busy hotel type.
11. GDS contribution channel contributed more to ADR to increase income in city hotels.
12. The optimal stay length in both hotel types is less than **7 days**.
13. Repeated customers do not cancel their bookings.
14. The number of people increases ADR increases.
15. arrival_date_year and arrival_date_week_number columns have a negative correlation which is **-0.51**.
16. stays_in_week_nights and total_stay have a positive correlation which is **0.95**.
17. Line plot is used to show the relationship between adr and total_stay and it indicates that they both have a positive correlation.

The EDA findings suggested several strategies that the client should implement to improve their pricing, marketing, and customer satisfaction approaches. These include optimizing pricing strategies, focusing efforts on improving their services, preferred room types [including larger/smaller sizes], and aligning prices with market trends. Additionally, monitoring guest satisfaction; addressing concerns promptly; utilizing positive reviews for marketing purposes; and monitoring market trends will contribute to the client's success in the competitive Hotel industry.

In conclusion, this project demonstrated the importance of EDA in gaining insights from the Hotel Booking dataset. The analysis of variables and visualization of data enabled us to formulate actionable recommendations for our client that would help them achieve their business objectives and excel in the competitive Hotel industry.