Capstone Project

Exploratory Data Analysis

Topic: Hotel Bookings Analysis

BY

Varsha Rani

POINTS FOR DISCUSSION

- Problem Statement & Objective
- Data Summary
- Data Exploration
- Data Cleaning
- Data Preparation
- Numerical Data
- Correlation
- Types of Hotels
- Visual Analysis of Data
- Relationship between data
- Conclusion

Problem Statement & Objective

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.

The objective of this project is to explore and analyze the data to discover important factors that govern the bookings.

Libraries Used:

- Pandas – Manipulation of tabular data into data frames
- NumPy – Mathematical operation on arrays
- Matplotlib - Visualization
- Seaborn - Visualization

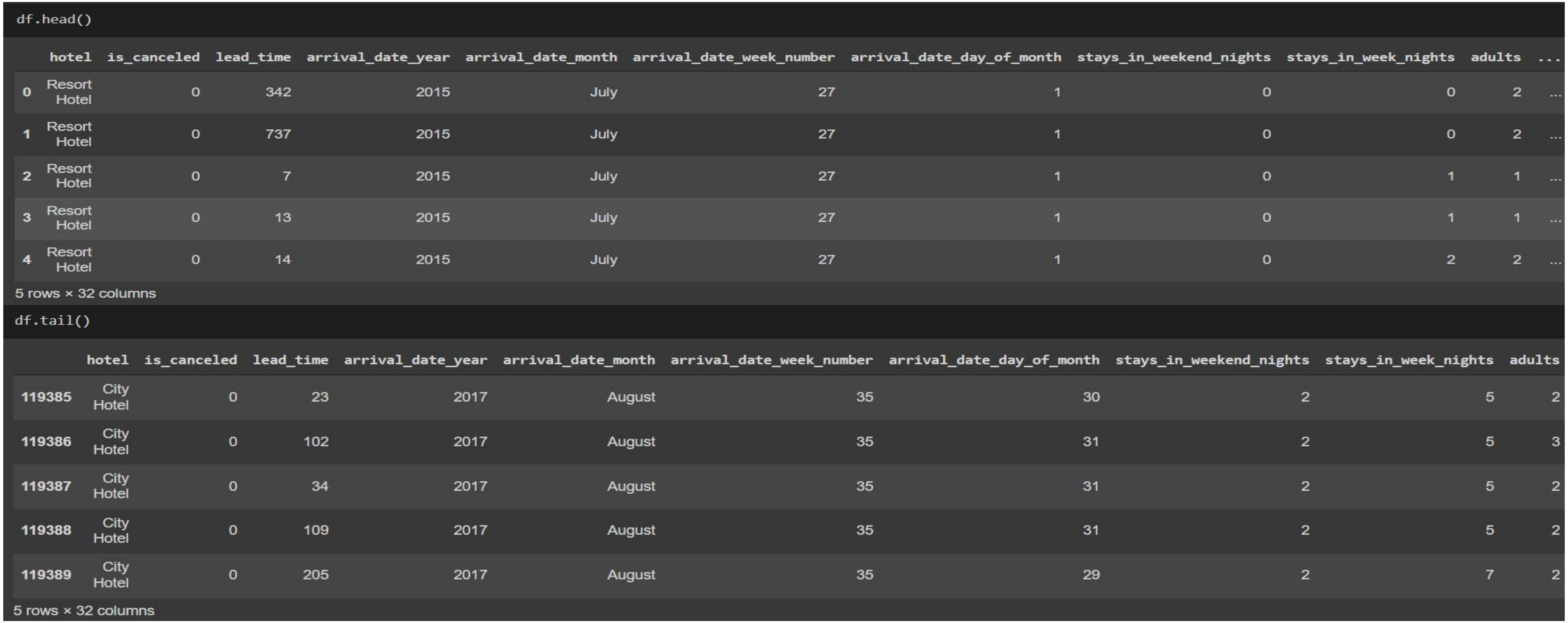
Data Summary

The Dataset contains the following columns :

1. **hotel:** type of hotels
2. **is_canceled:** canceled or not
3. **lead_time:** no. of days before actual arrival in the hotel
4. **arrival_date_year:** year of booking
5. **arrival_date_month:** month of booking
6. **arrival_date_week_number:** week number of the year in which booking was done
7. **arrival_date_day_of_month:** arrival month date
8. **stays_in_weekend_nights:** no. of weekends guest stayed
9. **stays_in_week_nights:** no. of weekdays guest stayed
10. **Adults:** number of adult guests
11. **Children:** number of children guests
12. **Babies:** number of baby guests
13. **meal:** BB – Bed & Breakfast
    HB – only two meals including breakfast meal
    FB – breakfast, lunch, and dinner
14. **country**
15. **market_segment:** TA: Travel agents
    TO: Tour operators
16. **distribution_channel**
17. **is_repeated_guest**
18. **previous_cancellations:** cancellation in past
19. **previous_bookings_not_canceled:** not cancelled in past
20. **reserved_room_type**
21. **assigned_room_type**
22. **booking_changes**
23. **deposit_type**
24. **agent**
25. **company**
26. **days_in_waiting_list**
27. **customer_type**
28. **adr:** average daily rate
29. **required_car_parking_spaces**
30. **total_of_special_requests**
31. **reservation_status**
32. **reservation_status_date**

Data Exploration

There are 119390 rows and 32 columns in the given dataset.



```
df.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...

5 rows × 32 columns

```
df.tail()
```

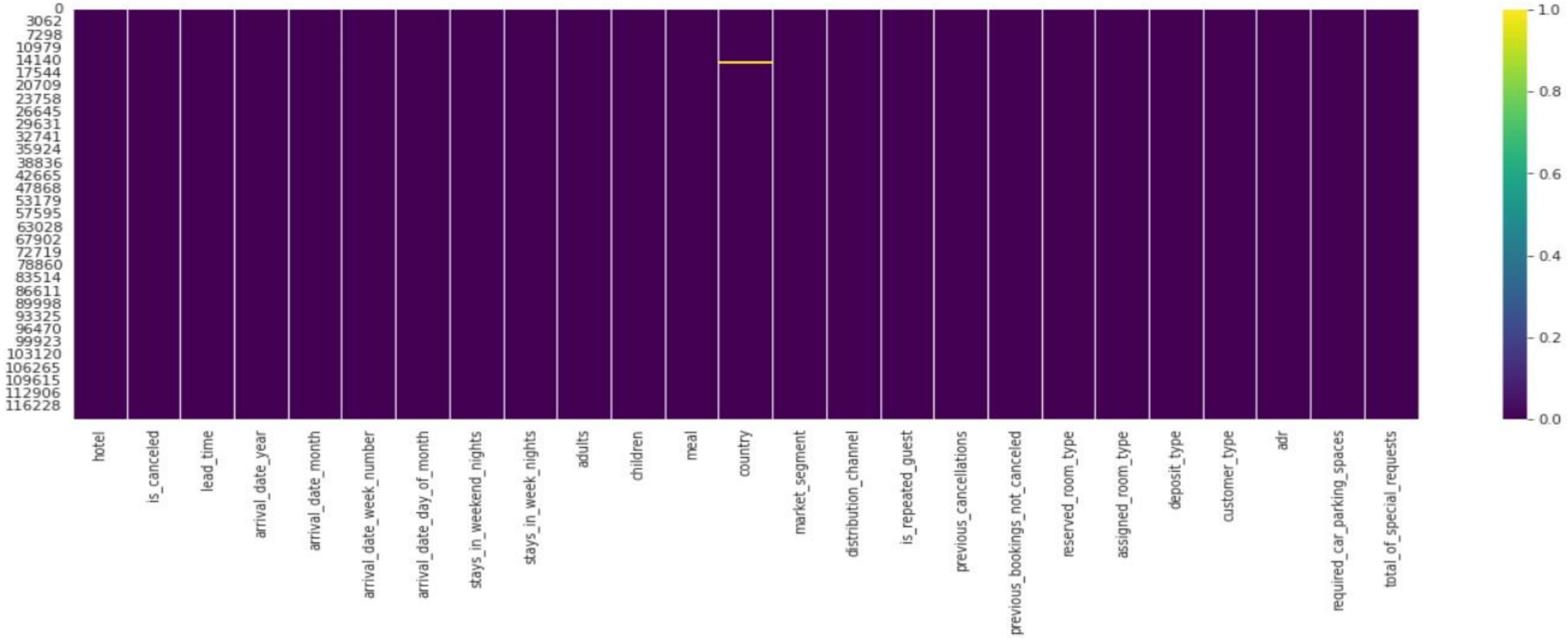
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults
119385	City Hotel	0	23	2017	August	35	30	2	5	2
119386	City Hotel	0	102	2017	August	35	31	2	5	3
119387	City Hotel	0	34	2017	August	35	31	2	5	2
119388	City Hotel	0	109	2017	August	35	31	2	5	2
119389	City Hotel	0	205	2017	August	35	29	2	7	2

5 rows × 32 columns

- Checked the name of columns.
- Extracted number of unique values in each column.
- Viewed data according to arrival_date_month and arrival_date_year.

Data Cleaning

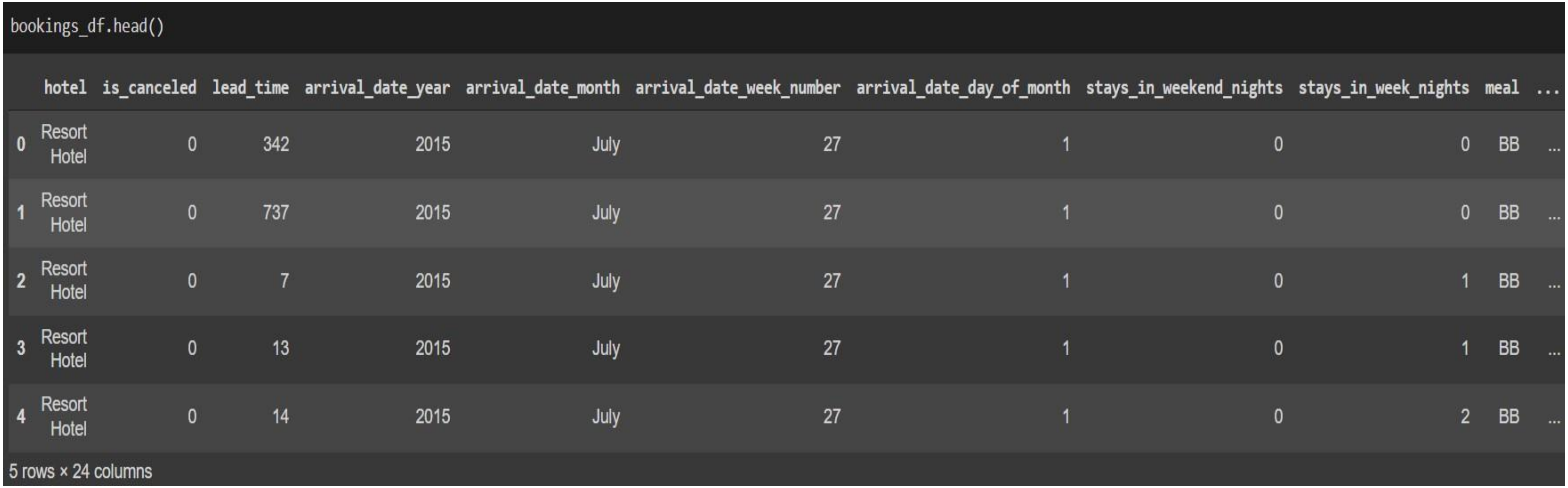
- Dropped duplicates.
- Stored the selected columns for further analysis in a new variable.
- Checked for null values.



Heatmap showing null values

Data Preparation

- Handled null values
- Merged ‘adult’ and ‘children’ column to get total number of guests and created a new column named ‘Total_Guests’. Then dropped the ‘adult’ and ‘children’ column.
- Excluded rows having 0 guests in ‘Total_Guests’ column.
- Now there are 87230 rows and 24 columns in the dataset.

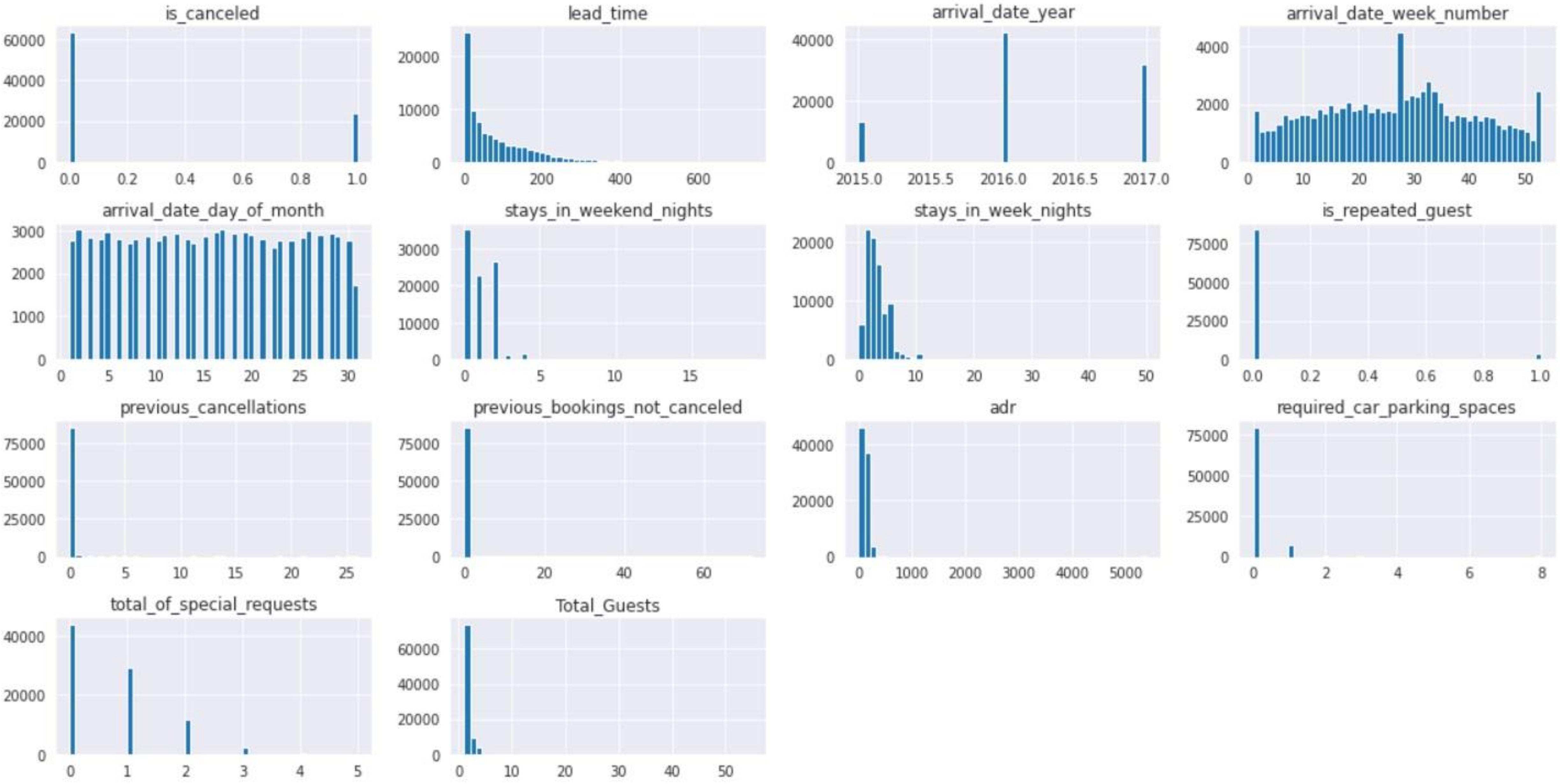


```
bookings_df.head()
```

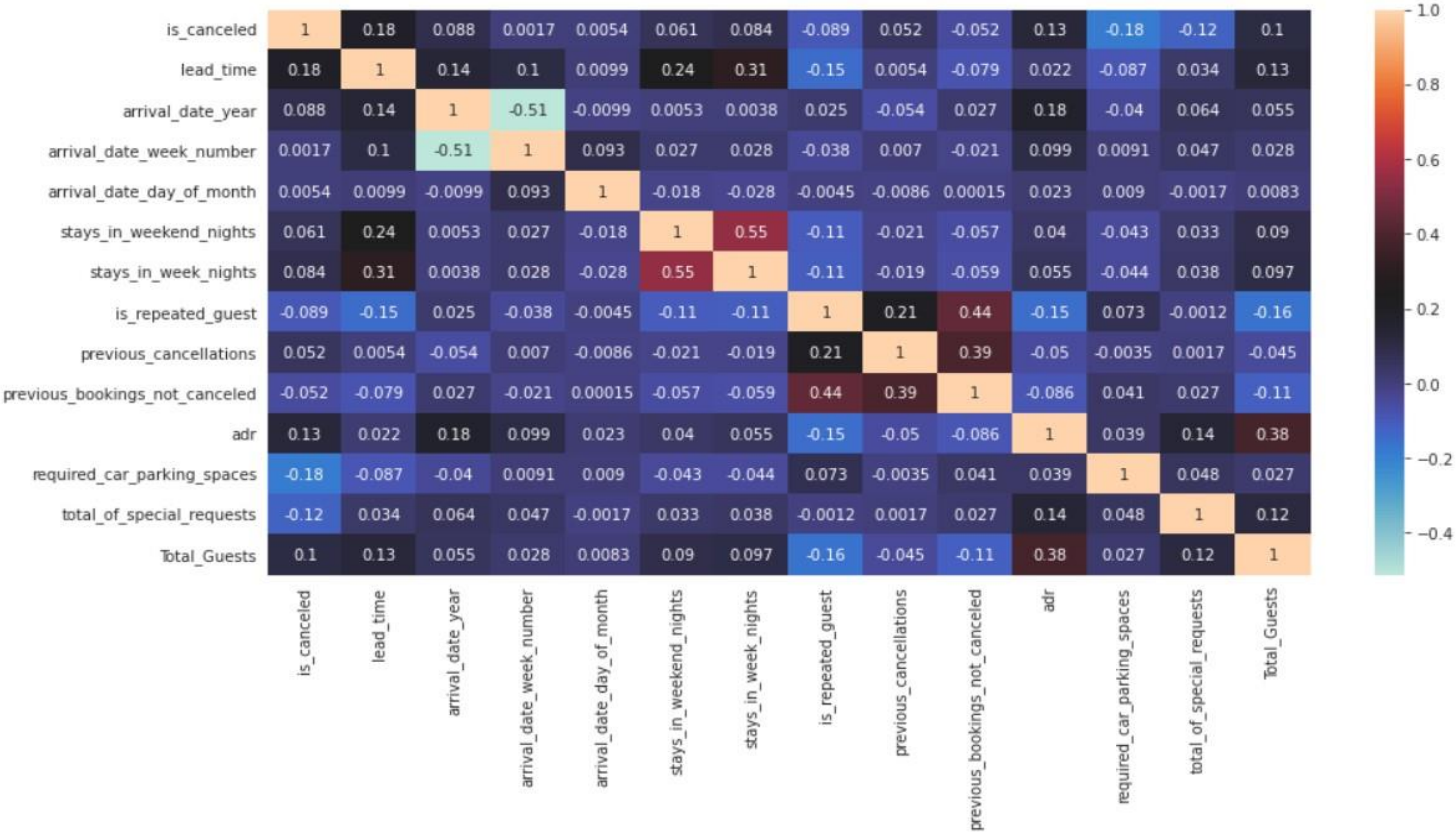
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	meal	...
0	Resort Hotel	0	342	2015	July	27	1	0	0	BB	...
1	Resort Hotel	0	737	2015	July	27	1	0	0	BB	...
2	Resort Hotel	0	7	2015	July	27	1	0	1	BB	...
3	Resort Hotel	0	13	2015	July	27	1	0	1	BB	...
4	Resort Hotel	0	14	2015	July	27	1	0	2	BB	...

5 rows × 24 columns

Each Numerical Data Count



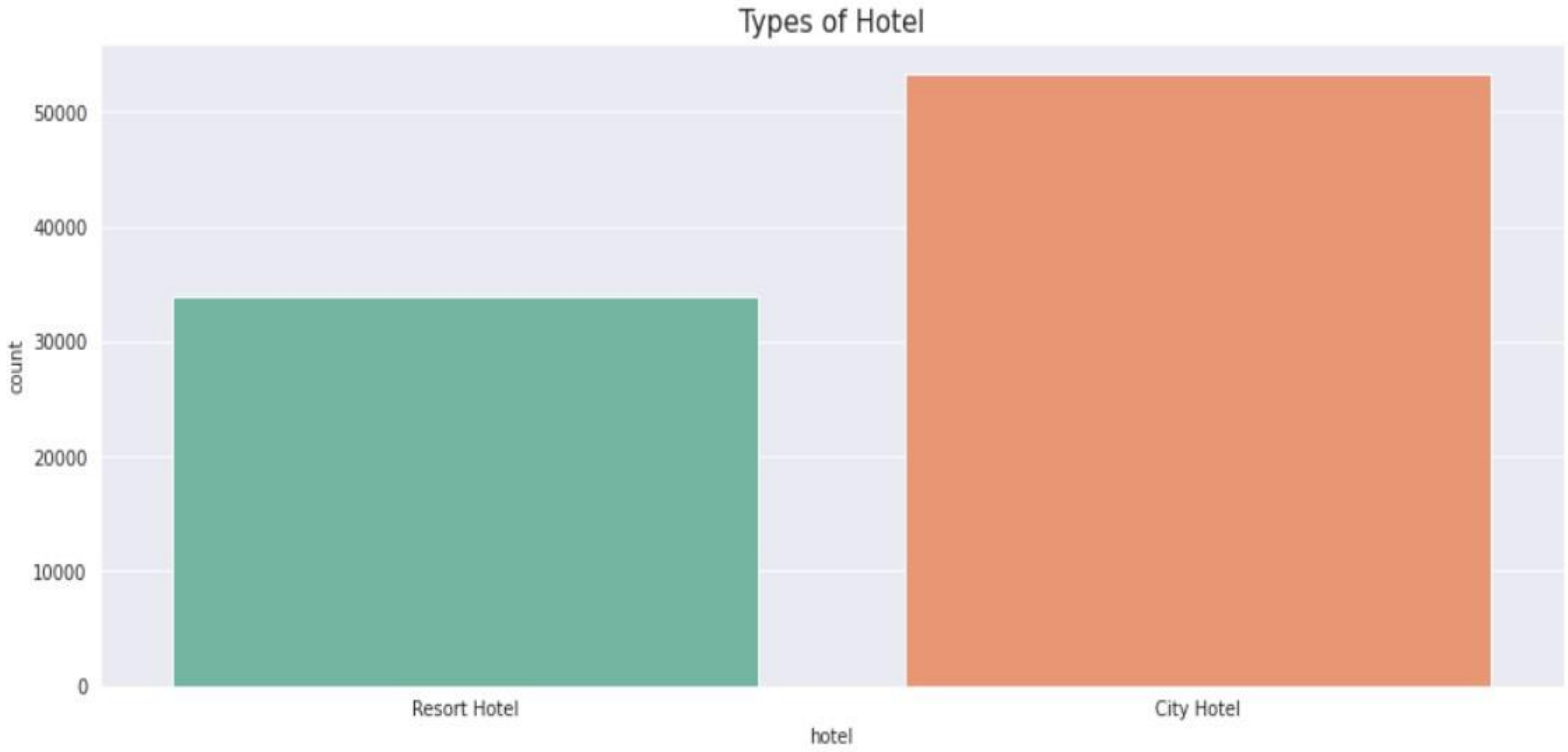
Correlation Between Data



	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	adr	required_car_parking_spaces	total_of_special_requests	Total_Guests
is_canceled	1	0.18	0.088	0.0017	0.0054	0.061	0.084	-0.089	0.052	-0.052	0.13	-0.18	-0.12	0.1
lead_time	0.18	1	0.14	0.1	0.0099	0.24	0.31	-0.15	0.0054	-0.079	0.022	-0.087	0.034	0.13
arrival_date_year	0.088	0.14	1	-0.51	-0.0099	0.0053	0.0038	0.025	-0.054	0.027	0.18	-0.04	0.064	0.055
arrival_date_week_number	0.0017	0.1	-0.51	1	0.093	0.027	0.028	-0.038	0.007	-0.021	0.099	0.0091	0.047	0.028
arrival_date_day_of_month	0.0054	0.0099	-0.0099	0.093	1	-0.018	-0.028	-0.0045	-0.0086	0.00015	0.023	0.009	-0.0017	0.0083
stays_in_weekend_nights	0.061	0.24	0.0053	0.027	-0.018	1	0.55	-0.11	-0.021	-0.057	0.04	-0.043	0.033	0.09
stays_in_week_nights	0.084	0.31	0.0038	0.028	-0.028	0.55	1	-0.11	-0.019	-0.059	0.055	-0.044	0.038	0.097
is_repeated_guest	-0.089	-0.15	0.025	-0.038	-0.0045	-0.11	-0.11	1	0.21	0.44	-0.15	0.073	-0.0012	-0.16
previous_cancellations	0.052	0.0054	-0.054	0.007	-0.0086	-0.021	-0.019	0.21	1	0.39	-0.05	-0.0035	0.0017	-0.045
previous_bookings_not_canceled	-0.052	-0.079	0.027	-0.021	0.00015	-0.057	-0.059	0.44	0.39	1	-0.086	0.041	0.027	-0.11
adr	0.13	0.022	0.18	0.099	0.023	0.04	0.055	-0.15	-0.05	-0.086	1	0.039	0.14	0.38
required_car_parking_spaces	-0.18	-0.087	-0.04	0.0091	0.009	-0.043	-0.044	0.073	-0.0035	0.041	0.039	1	0.048	0.027
total_of_special_requests	-0.12	0.034	0.064	0.047	-0.0017	0.033	0.038	-0.0012	0.0017	0.027	0.14	0.048	1	0.12
Total_Guests	0.1	0.13	0.055	0.028	0.0083	0.09	0.097	-0.16	-0.045	-0.11	0.38	0.027	0.12	1

Correlation is used to find the relationship between two variables which is important in real life because we can predict the value of one variable with the help of other variables, who is being correlated with it.

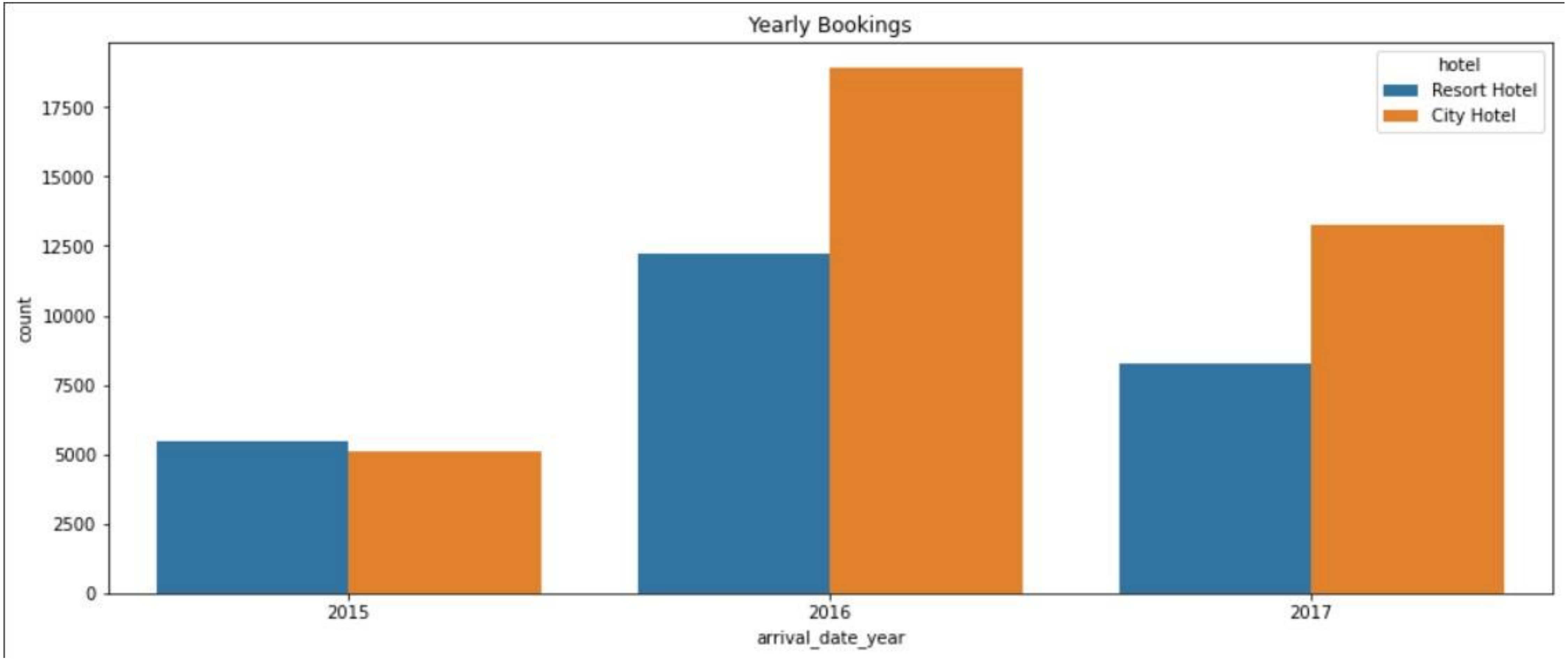
Types of Hotels



There two types of hotels:

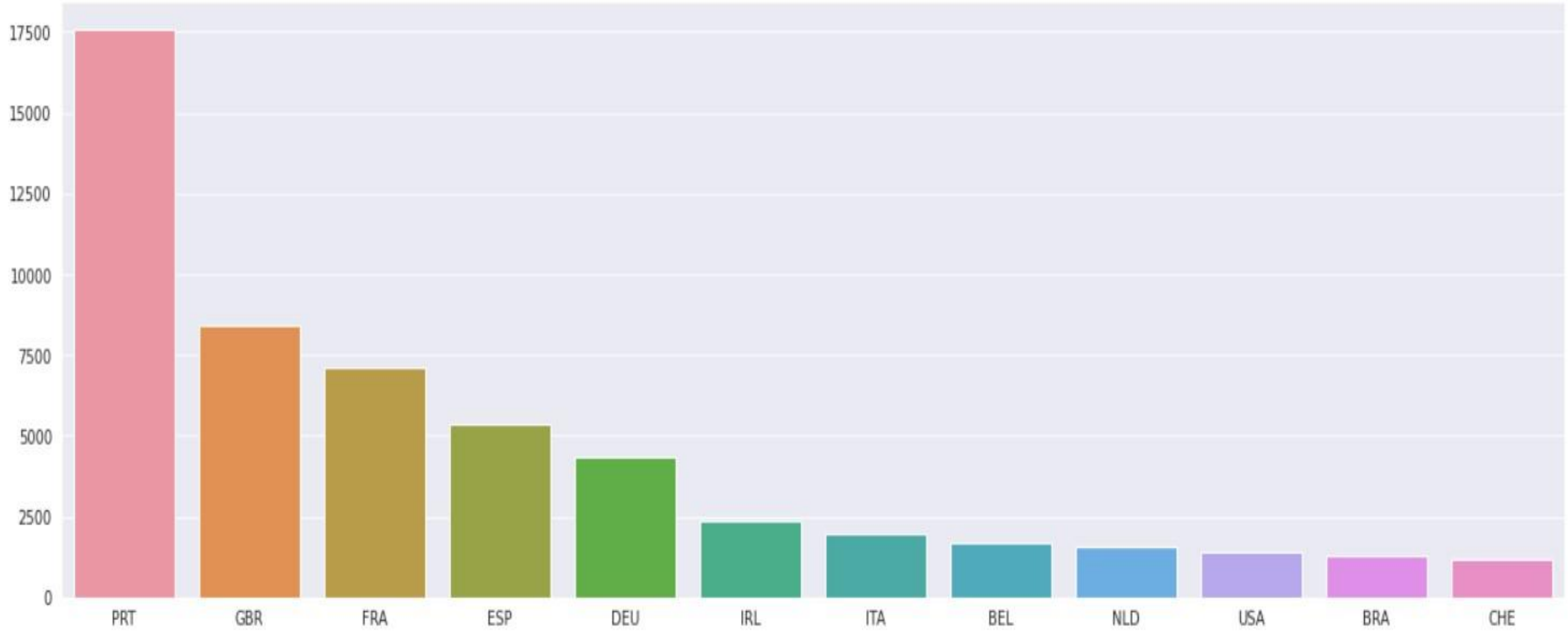
- ✓ Resort Hotel
- ✓ City Hotel

Hotel wise Yearly Bookings



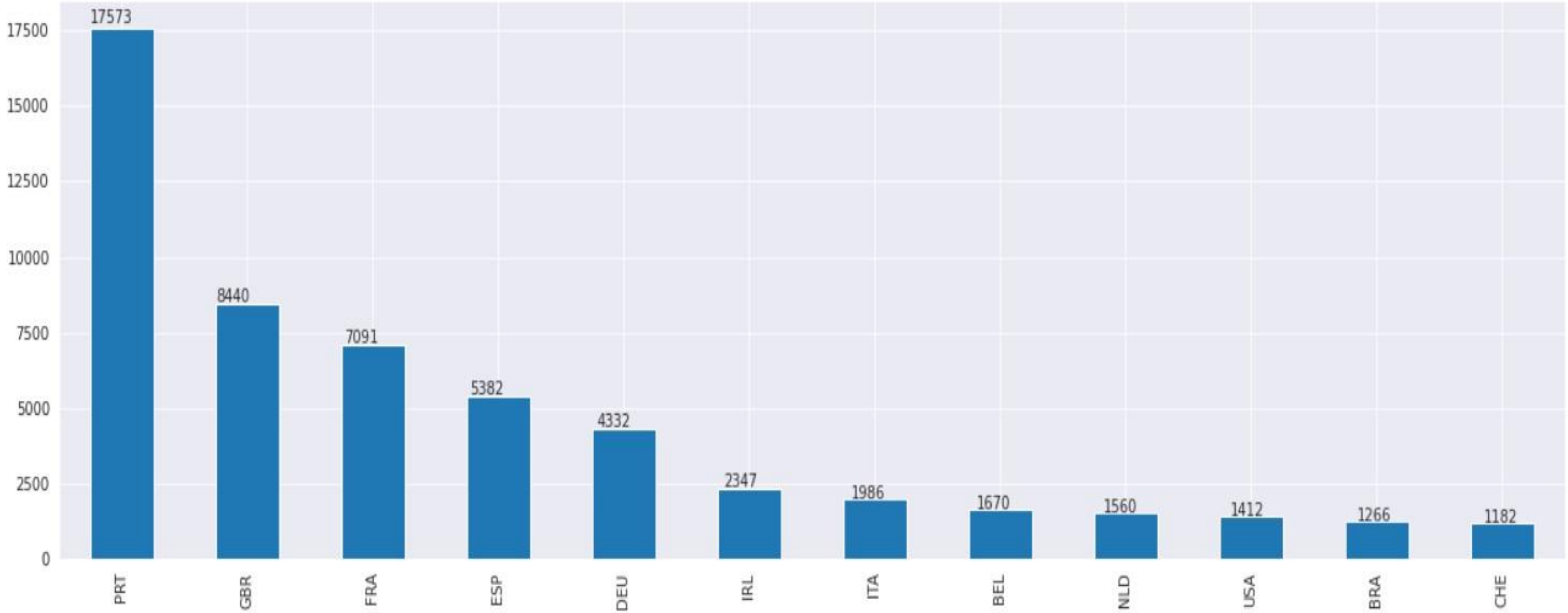
Bookings across the years 2016 and 2017 is higher for City Hotel compared to Resort Hotel and do not increase proportionately over the years.

Top Countries From Where The Most Guests Are Coming



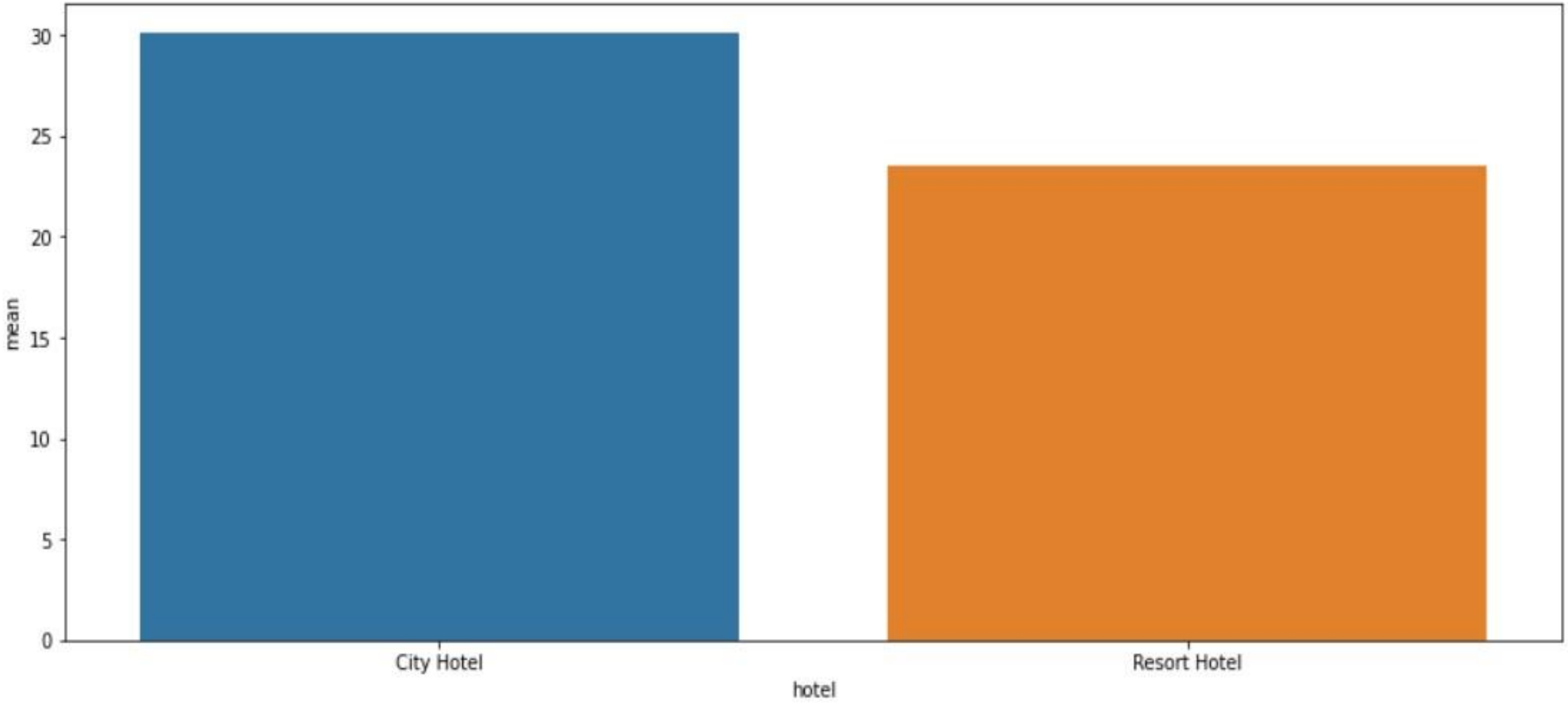
The above graph shows the names of top 12 countries from where the most guests are coming.

Top Countries With Values



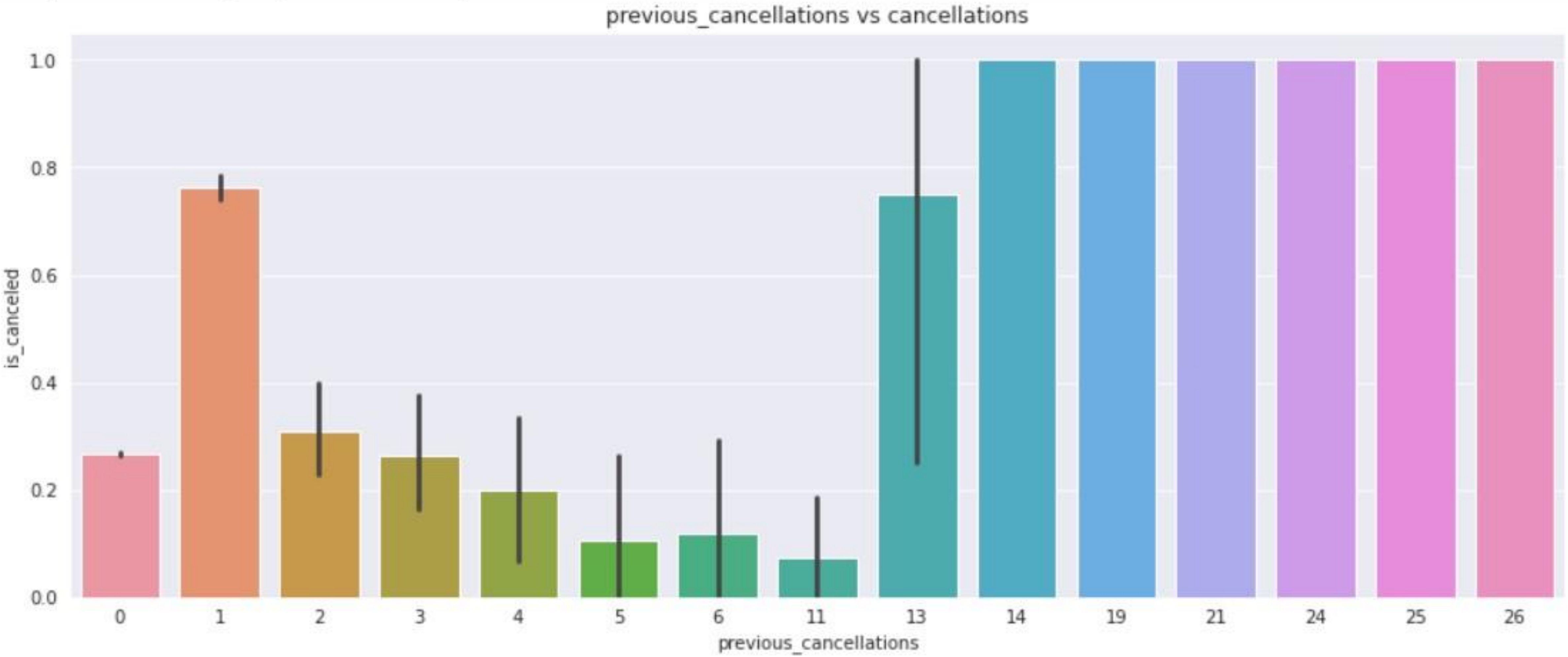
The above graph shows the names of top 12 countries with number of guests from where the most guests are coming.

Proportion of Booking Cancellation



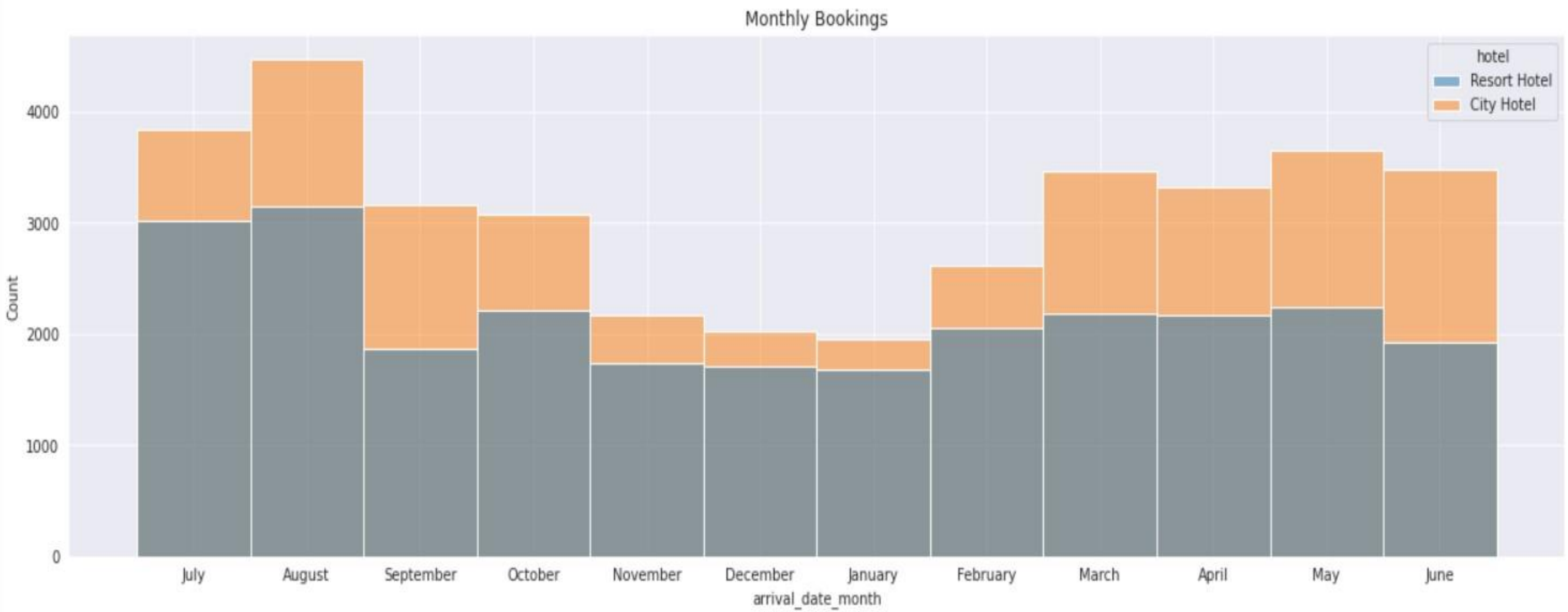
Around 30% of bookings were cancelled in City Hotels and 25% in Resort Hotels.

Previous Cancellations vs Cancellations



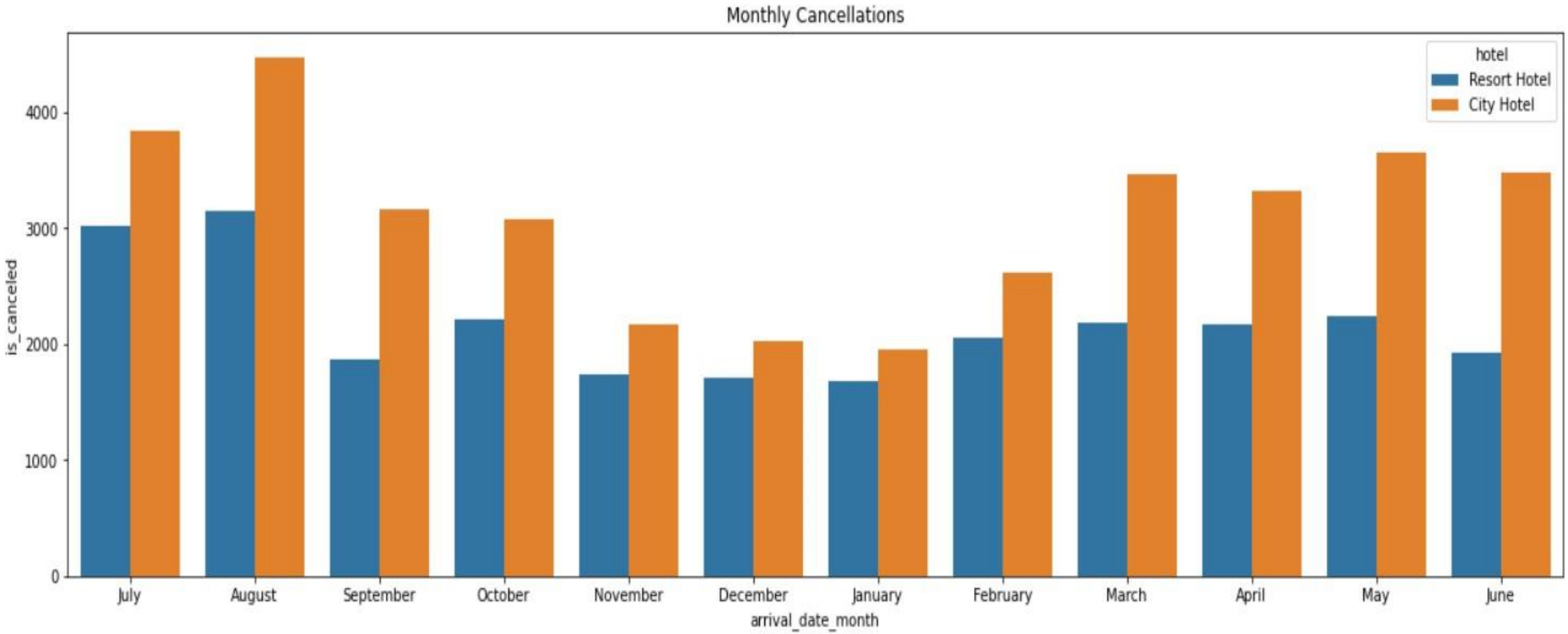
History of previous cancellations increases chances of cancellation.

Hotel wise Monthly Bookings



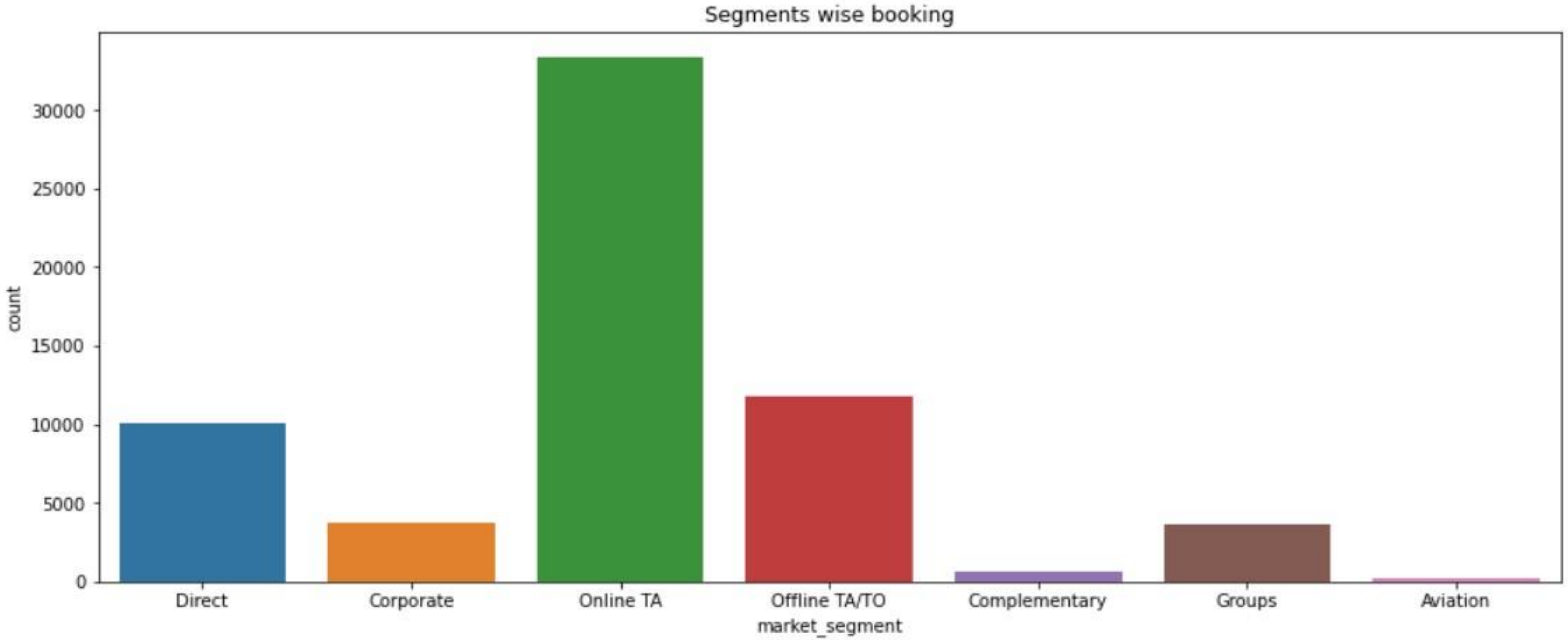
March, April, May, June, July, August, September and October are the months with higher bookings.

Hotel wise Monthly Cancellations



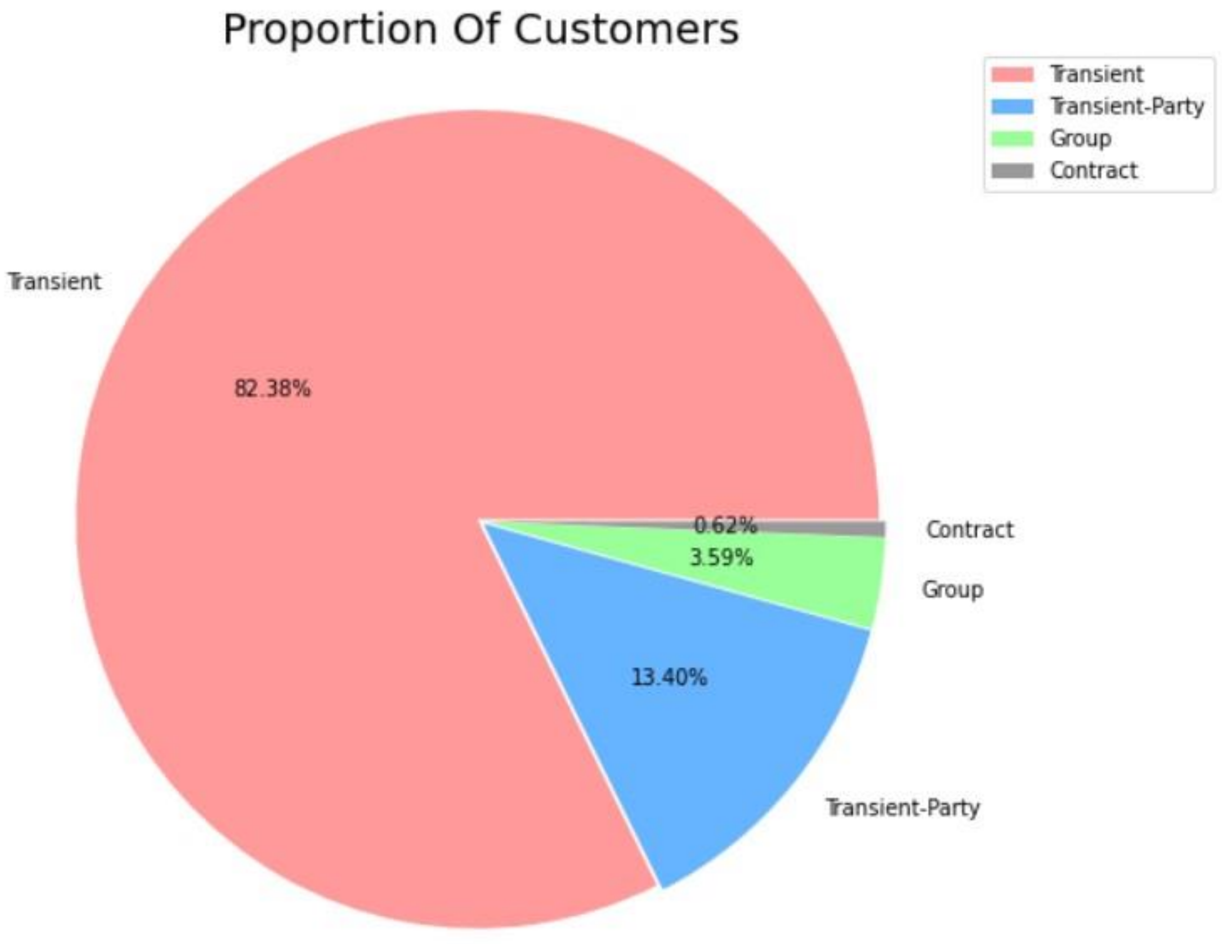
In case of city hotel, months with high bookings (March, April, May, June, July, August, September, October) also witnessed more cancellations. Both hotels have the fewest guests during the winter.

Market Segment vs Bookings



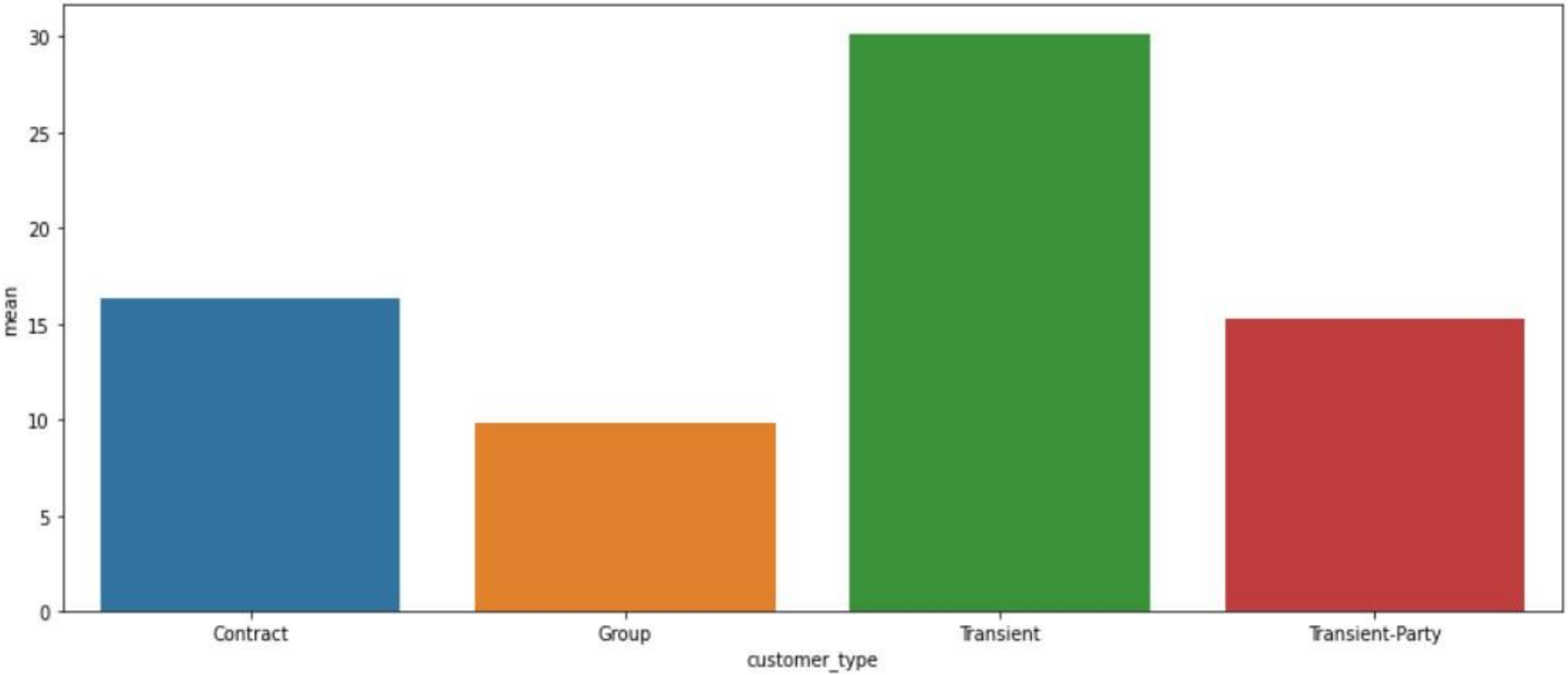
Indirect bookings through online and offline travel agents are higher compared to direct, corporate, groups, complementary and aviation.

Types of Customers



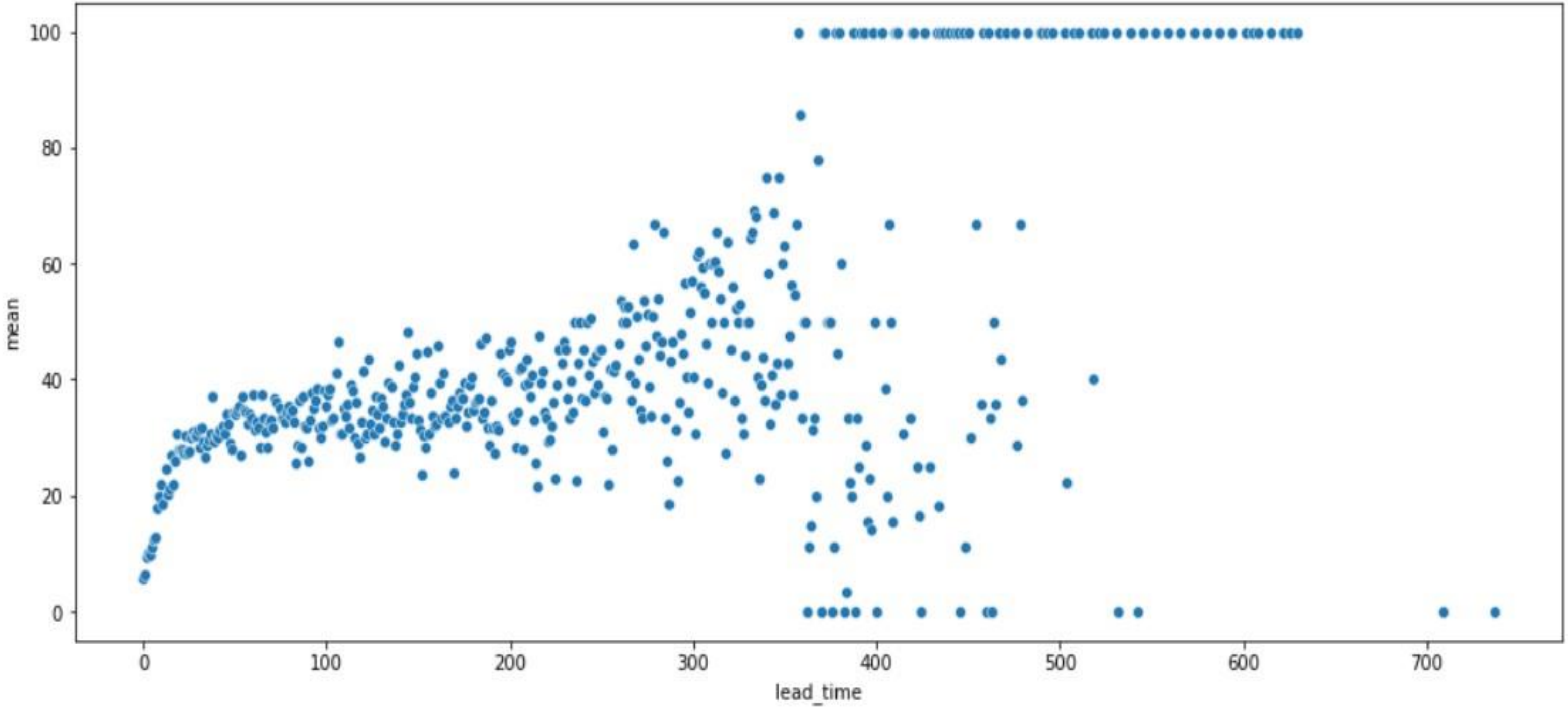
The pie chart shows proportion of different types of customers. Most number of customers are of Transient type and least number of customers are of Contract type.

Customer Type vs Bookings Cancellation



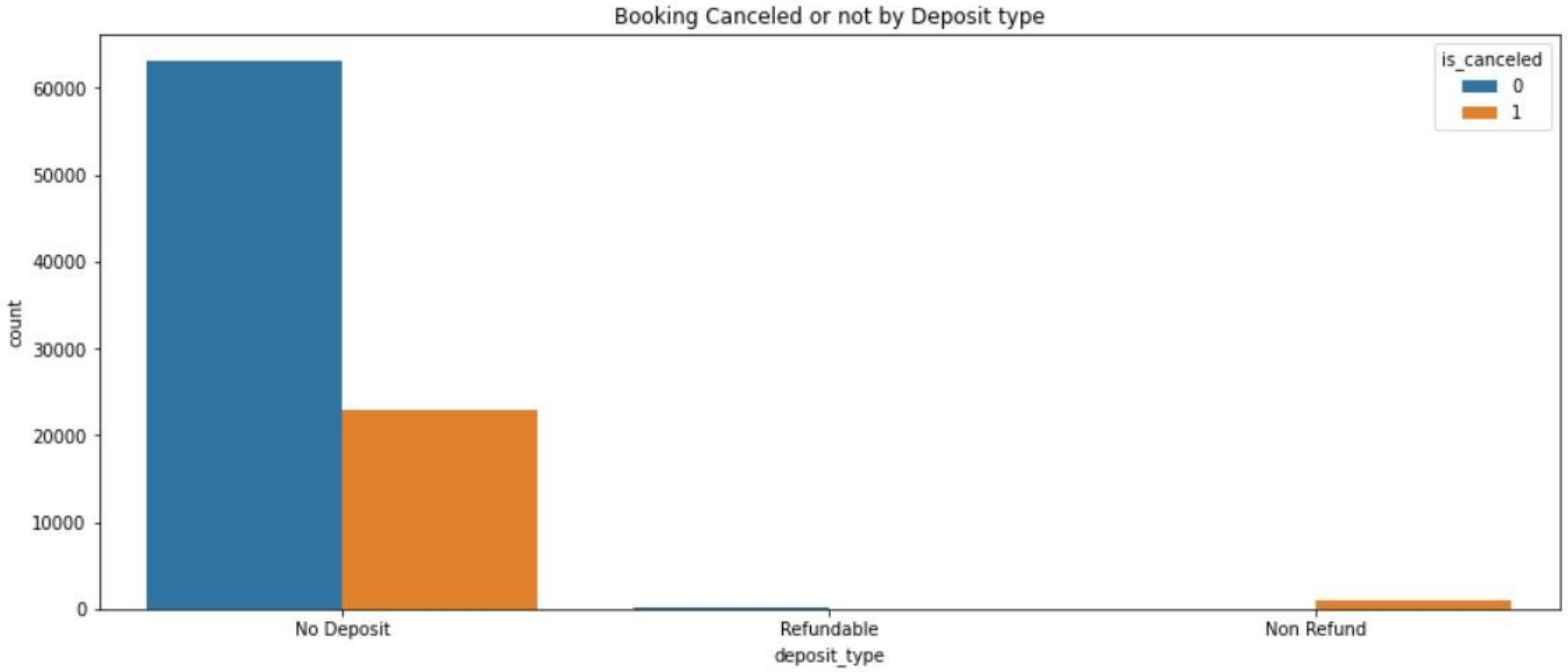
Transient customer types have higher cancellations.

Lead Time vs Cancellation



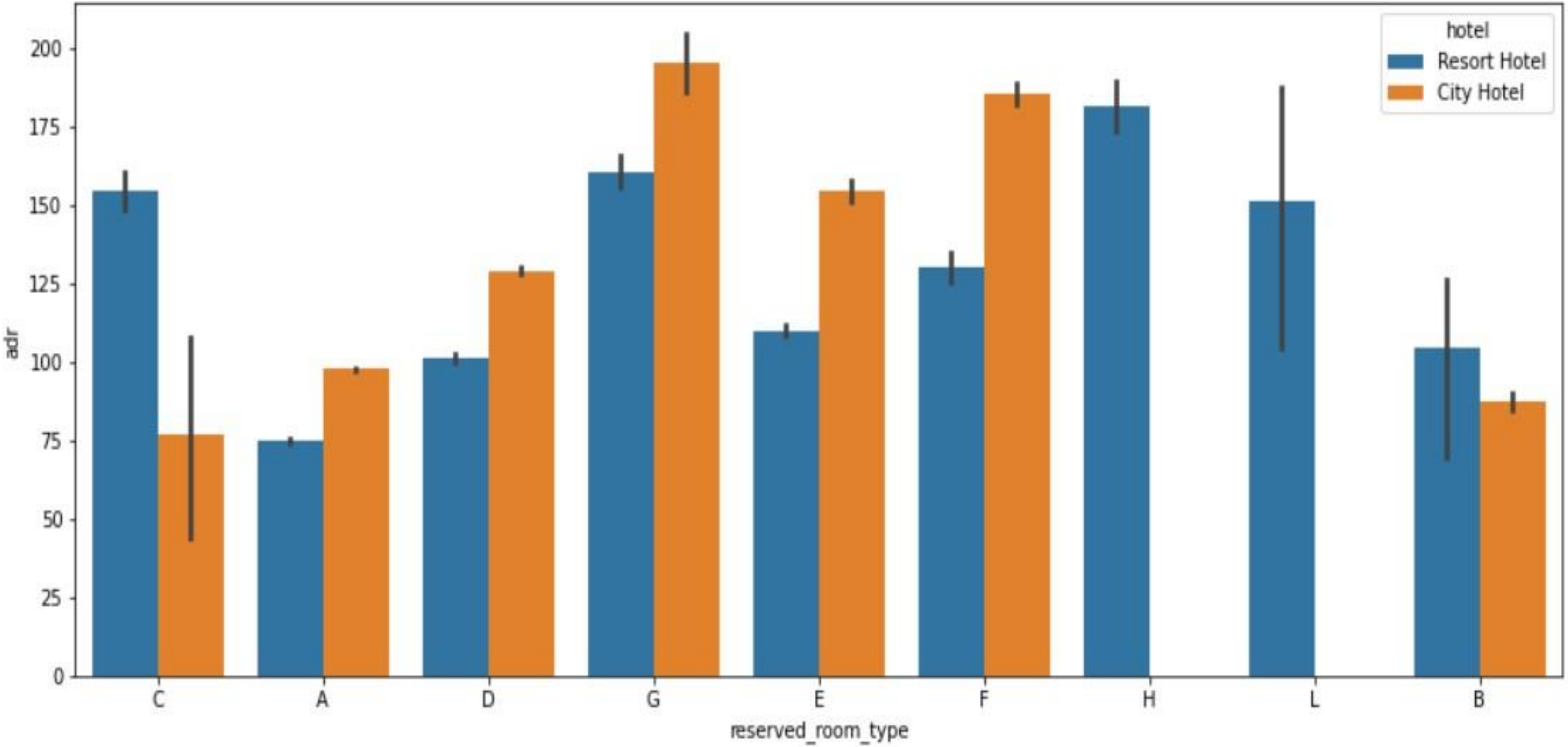
Lead time has a positive correlation with cancellation.

Deposit Type vs Cancellation

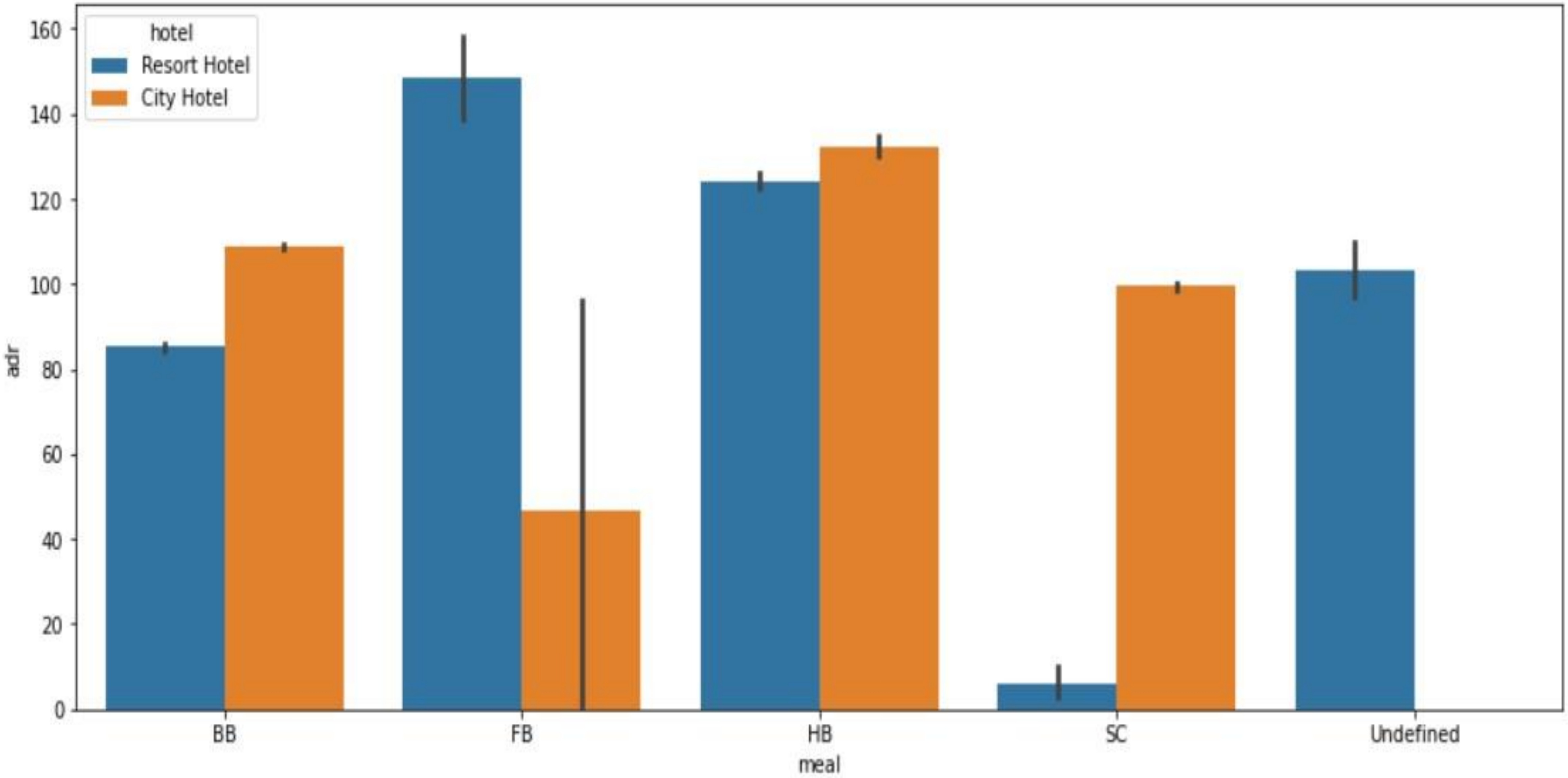


Most of the bookings were cancelled by guests with no deposit. Also it is interesting to note that non-refundable deposits had more cancellation than refundable deposits.

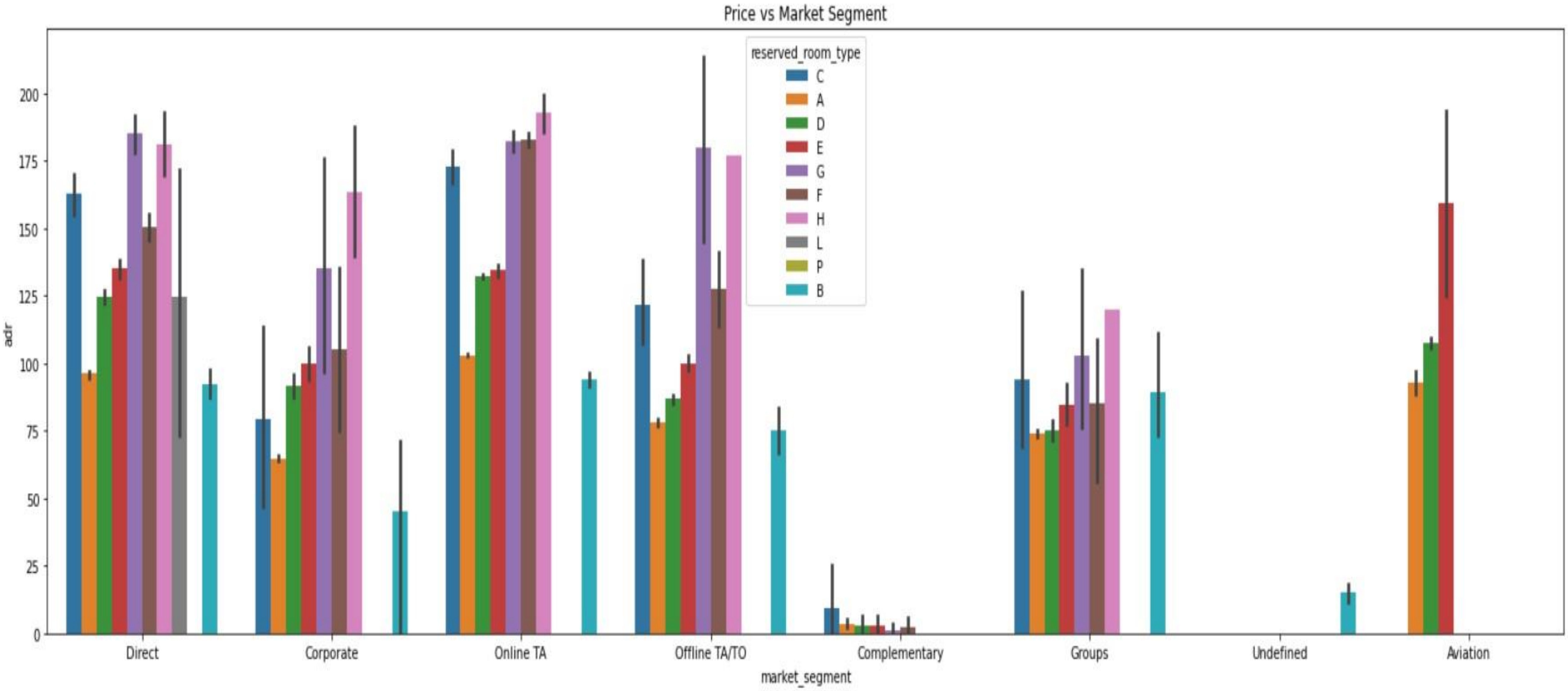
Average Daily Rate vs Room Type



Average Daily Rate vs Meal



Average Daily Rate vs Market Segment Based On Room Type



Conclusion

As per the EDA, the important factors derived from the given dataset are :

- There are two different types of hotels namely - 'Resort Hotel' and 'City Hotel'.
- Bookings across the years 2016 and 2017 are higher for City Hotels compared to Resort Hotels and do not increase proportionately over the years.
- Top 10 countries from where guests are coming are - PRT, GBR, FRA, ESP, DEU, IRL, ITA, BEL, NLD and USA. Most number of guests i.e. 17573 are coming from PRT.
- Focusing on bookings cancellation, around 30% of bookings were cancelled in City Hotels and 25% in Resort Hotels.
- The City hotel has more guests during spring and autumn, when the prices are also highest, In July and August there are less visitors, although prices are lower. Thus, customers can get good deal on bookings in July and August in city hotel.
- Guest numbers for the Resort hotel go down slightly from June to September, which is also when the prices are highest. Thus, these months should be avoided for bookings.
- Broadly, April to August is the peak season of bookings. Both hotels have the fewest guests during the winter.
- There are four different types of customers namely - Transient, Transient-Party, Group and Contract. Transient customer types have higher cancellations.
- Higher lead time has higher chance of cancellation. Also, history of previous cancellations increases chances of cancellation.
- No deposit cancellations are high compared to other categories but these should not be discouraged per se as bookings in this category are also very high compared to non refundable type bookings.
- Cancellations are high when done through agents compared to direct bookings. Hotels need to do marketing and give special incentives for direct bookings as these may establish personal one to one relationships promoting customer loyalty.

Thank You!!

☺