Module 2

Capstone Project: Exploratory Data Analysis

A Comprehensive Analysis of Hotel Booking Data

Subtitle:
Uncovering Insights for Strategic Decision-Making in the Hospitality Industry

Presented By:

Dipak Gaikwad

- **Objective:**
- To explore and analyze key factors affecting hotel bookings, guiding operational and strategic decisions to enhance profitability and guest satisfaction.

- **Introduction**
  - Dataset overview and analytical focus.
- **Data Overview**
  - Key variables and initial data assessment.
- **Booking Time and Length of Stay Analysis**
  - Impact on pricing and guest decisions.
- **Seasonal and Weekly Trends**
  - Identifying peak and low demand periods.
- **Demographic Insights**
  - Guest composition and behavior analysis.
- **Parking Utilization and Influence**
  - Correlation with guest choices and hotel type.
- **Cancellation Trends and Predictive Modeling**
  - Insights and forecast model development.
- **Conclusions and Recommendations**
  - Key findings and actionable strategies.

Here I have divided this task in to Three steps:

1. Data Collection & Understanding

❑ **Data Collection:**

- Gather all necessary data files or sources.
- Verify the completeness and accuracy of the data sources.

❑ **Initial Data Exploration:**

- Perform an initial scan of the data to understand the variables and their types (numerical, categorical).
- Identify key variables that will be relevant to the analysis based on the project's objectives.

2. Data Cleaning & Manipulation

❑ Data Cleaning:

- Handle missing values by either imputing or removing them based on their impact on the dataset.
- Detect and correct errors or outliers in the data.

❑ Data Manipulation:

- Create new variables that might be necessary for deeper insights, such as combining date and time fields, or segmenting guests into new demographic groups.
- Format and transform data (e.g., converting strings to datetime objects, categorizing numerical values into bins).

❑ Data Reduction:

- Reduce the dataset to a manageable size if necessary, focusing on the most relevant attributes.
- Perform initial feature selection to remove redundant or irrelevant features.

3: Exploratory & Analysis (EDA)

❑ **Univariate Analysis:**

- Analyze single variables to understand their distribution, central tendency, and dispersion.
- Generate histograms, box plots, and summary statistics for individual variables.

❑ **Bivariate/Multivariate Analysis:**

- Explore relationships between pairs or groups of variables using scatter plots, correlation matrices, and cross-tabulations.
- Investigate potential relationships and patterns that could influence hotel bookings, such as the relationship between lead time and cancellation rates.

❑ **Insight Generation:**

- Synthesize findings from univariate and multivariate analysis to generate actionable insights.
- Prepare preliminary reports or dashboards that summarize key statistics and show visual trends.

❑ **Hypothesis Testing:**

- Formulate and test statistical hypotheses based on patterns observed in the data.
- Use tests such as t-tests, chi-square tests, or ANOVA to validate these hypotheses, focusing on relationships that are crucial for decision-making.

First Five Rows Data

index	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num	arrival_date_day_of_mo	stays_in_weekend_night	stays_in_week_nights	adults	children	babies	meal	country	market_segment	distribution_channel	is_repeated_guest	previous_cancellations	previous_bookings_not_	reserved_room_type
0	Resort	0	342	2015	July	27	1	0	0	2	0	0	BB	PRT	Direct	Direct	0	0	0	C
1	Resort	0	737	2015	July	27	1	0	0	2	0	0	BB	PRT	Direct	Direct	0	0	0	C
2	Resort	0	7	2015	July	27	1	0	1	1	0	0	BB	GBR	Direct	Direct	0	0	0	A
3	Resort	0	13	2015	July	27	1	0	1	1	0	0	BB	GBR	Corporat	Corporat	0	0	0	A
4	Resort	0	14	2015	July	27	1	0	2	2	0	0	BB	GBR	Online	TA/TO	0	0	0	A

Last Five Rows Data

index	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num	arrival_date_day_of_mo	stays_in_weekend_night	stays_in_week_nights	adults	children	babies	meal	country	market_segment	distribution_channel	is_repeated_guest	previous_cancellations	previous_bookings_not_	reserved_room_type
119385	City	0	23	2017	August	35	30	2	5	2	0	0	BB	BEL	Offline	TA/TO	0	0	0	A
119386	City	0	102	2017	August	35	31	2	5	3	0	0	BB	FRA	Online	TA/TO	0	0	0	E
119387	City	0	34	2017	August	35	31	2	5	2	0	0	BB	DEU	Online	TA/TO	0	0	0	D
119388	City	0	109	2017	August	35	31	2	5	2	0	0	BB	GBR	Online	TA/TO	0	0	0	A
119389	City	0	205	2017	August	35	29	2	7	2	0	0	HB	DEU	Online	TA/TO	0	0	0	A

Explore The Dataset

babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0

Checking Null Values In Dataset

babies	0
meal	0
country	0
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	0
company	0
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0

Replacing The Null Values With Their Mean

1. Data Collection & Understanding

The Data Contains of 119390 rows and 32 Columns

❑ **Dataset Description:**

- **hotel**: Indicates the type of hotel - either a resort hotel (H1) or a city hotel (H2).
- **is_cancelled**: Specifies whether the booking was cancelled (1) or not (0).
- **lead_time**: Number of days elapsed between the booking entry date and the arrival date.
- **arrival_date_year**: Year of the arrival date.
- **arrival_date_month**: Month of the arrival date.
- **arrival_date_week_number**: Week number of the arrival date.
- **arrival_date_day_of_month**: Day of the month of the arrival date.
- **stays_in_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
- **stays_in_week_nights**: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
- **adults**: Number of adults included in the booking.
- **children**: Number of children included in the booking.
- **babies**: Number of babies included in the booking.
- **meal**: Type of meal option opted for.
- **country**: Country code of the guest.
- **market_segment**: Market segment of the booking (e.g., corporate, direct, travel agents).

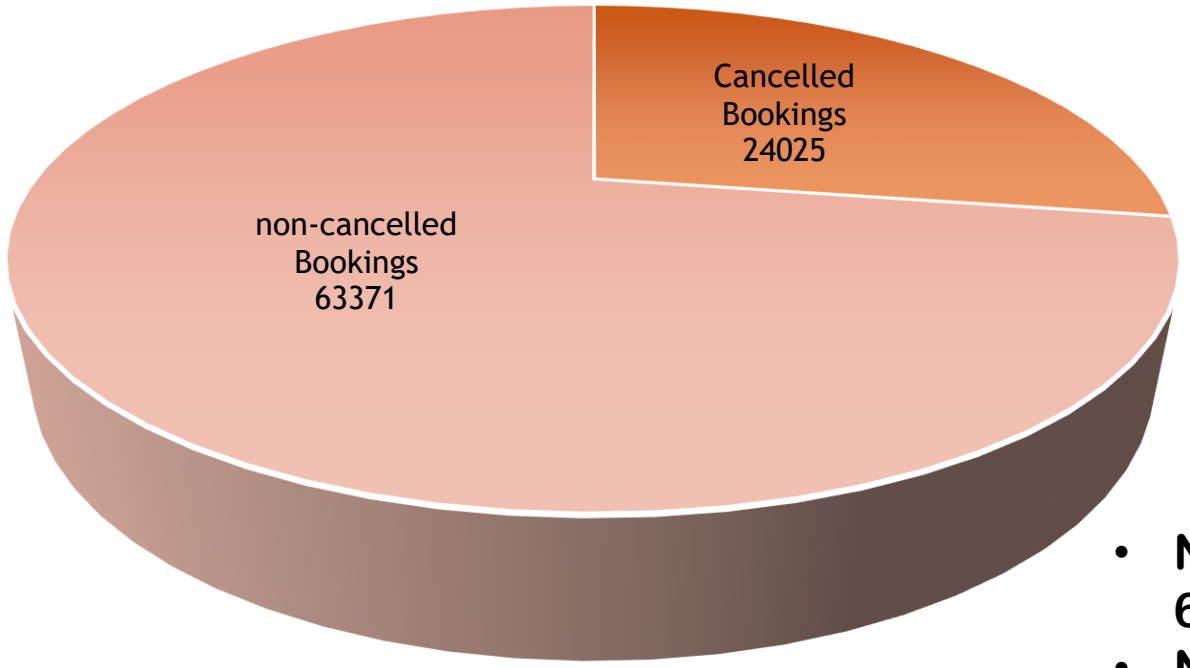
- **distribution_channel**: Channel through which the booking was made (e.g., corporate, direct, travel agents).
- **is_repeated_guest**: Indicates if the guest is a repeated guest (1) or not (0).
- **previous_cancellations**: Count of previous booking cancellations by the guest.
- **previous_bookings_not_canceled**: Count of previous bookings not canceled by the guest.
- **reserved_room_type**: Type of room reserved by the guest.
- **assigned_room_type**: Type of room assigned to the guest.
- **booking_changes**: Count of changes made to the booking.
- **deposit_type**: Type of deposit made for the booking.
- **agent**: ID of the travel agent through whom the booking was made.
- **days_in_waiting_list**: Number of days the booking was in the waiting list before it was confirmed.
- **customer_type**: Type of customer (e.g., transient, contract, group).
- **adr**: Average daily rate (i.e., average revenue per available room).
- **required_car_parking_spaces**: Number of car parking spaces required by the guest.
- **total_of_special_requests**: Number of additional special requests made by the guest.
- **reservation_status**: Status of the reservation (e.g., canceled, check-out, no-show).
- **reservation_status_date**: Date of the specific reservation status.

❑ Data Cleaning & Manipulation:

Number of duplicate rows in the dataset: 31994
Used .drop_duplicates() Method to remove Duplicate rows from datasets Number of rows after removing duplicates: 87396

Here are the columns with the most missing values along with their respective counts:
company: 82,137 missing values
agent: 12,193 missing values
country: 452 missing values
children: 4 missing values
"After identifying columns with the most missing values in the dataset, which include 'company', 'agent', 'country', and 'children', we utilized the fillna() method in pandas to replace the missing values in these columns. For 'company' and 'country', we replaced the missing values with 'Unknown' to indicate unknown or unspecified values. For 'agent', which likely represents the ID of the booking agent, and 'children', we replaced missing values with 0, assuming no children were specified for those bookings.

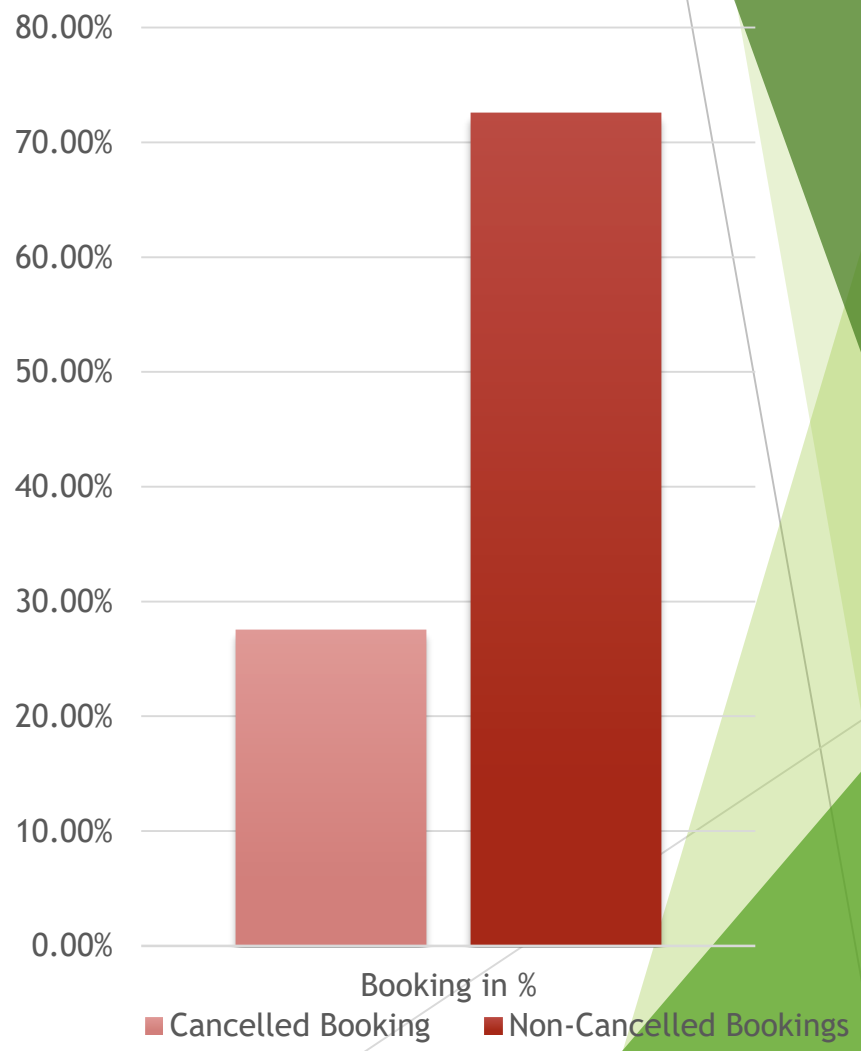
Cancelled and Non-Cancelled Bookings



- Number of non-canceled bookings: 63371
- Number of canceled bookings: 24025
- Due to this canceled bookings there will be an adverse effect on hotel business which means hotels are not able to make more profit, they are losing their customers.

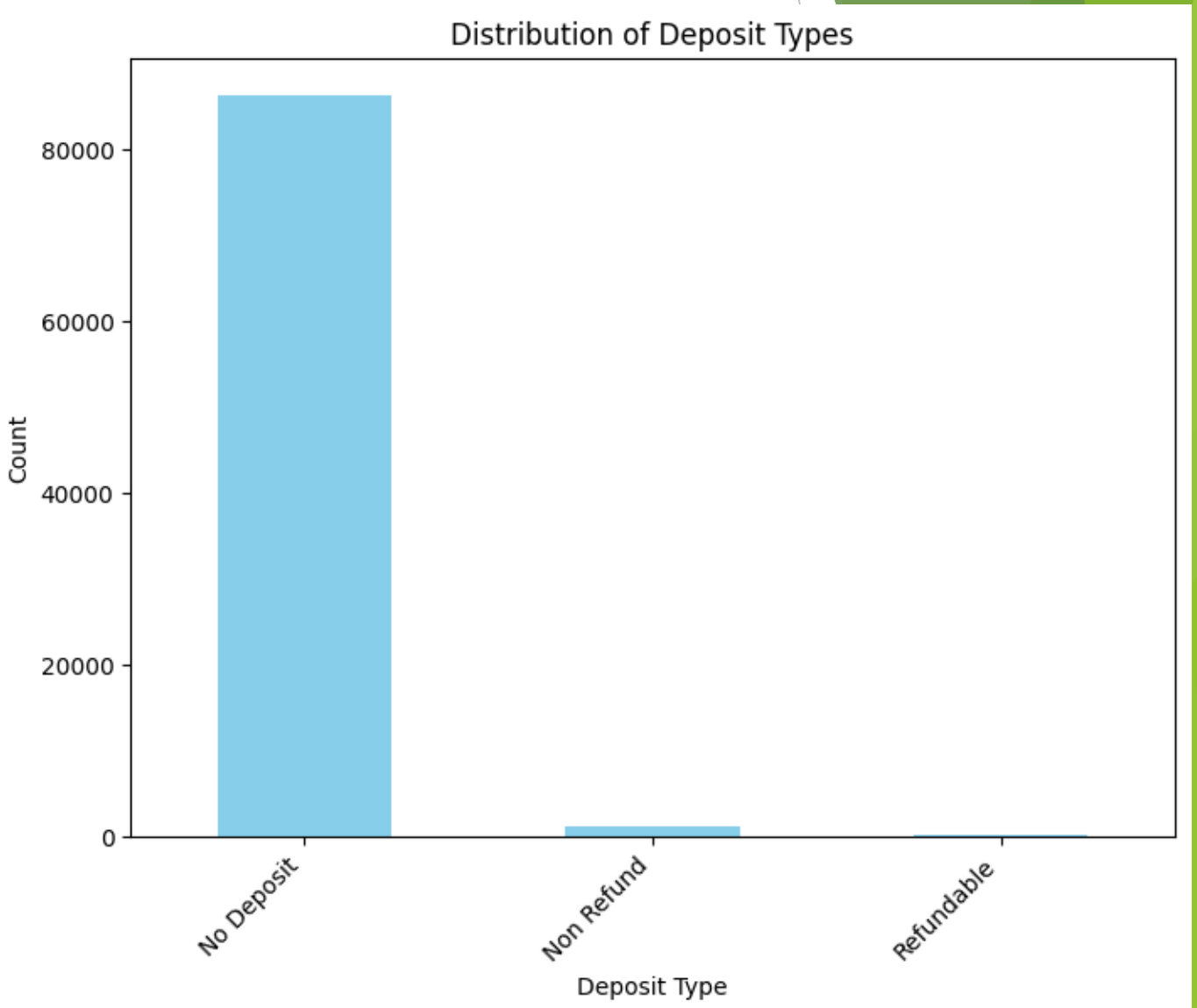
Non-Canceled vs Canceled Booking Percentage

As we saw total number of booking from our last slide, here we are going to see the same but in terms of percentage. This bar graph representing that 72.5% of customers are check-in to hotels where 27.5% of customers canceled their bookings.



▫ Deposit Policies Of Hotels

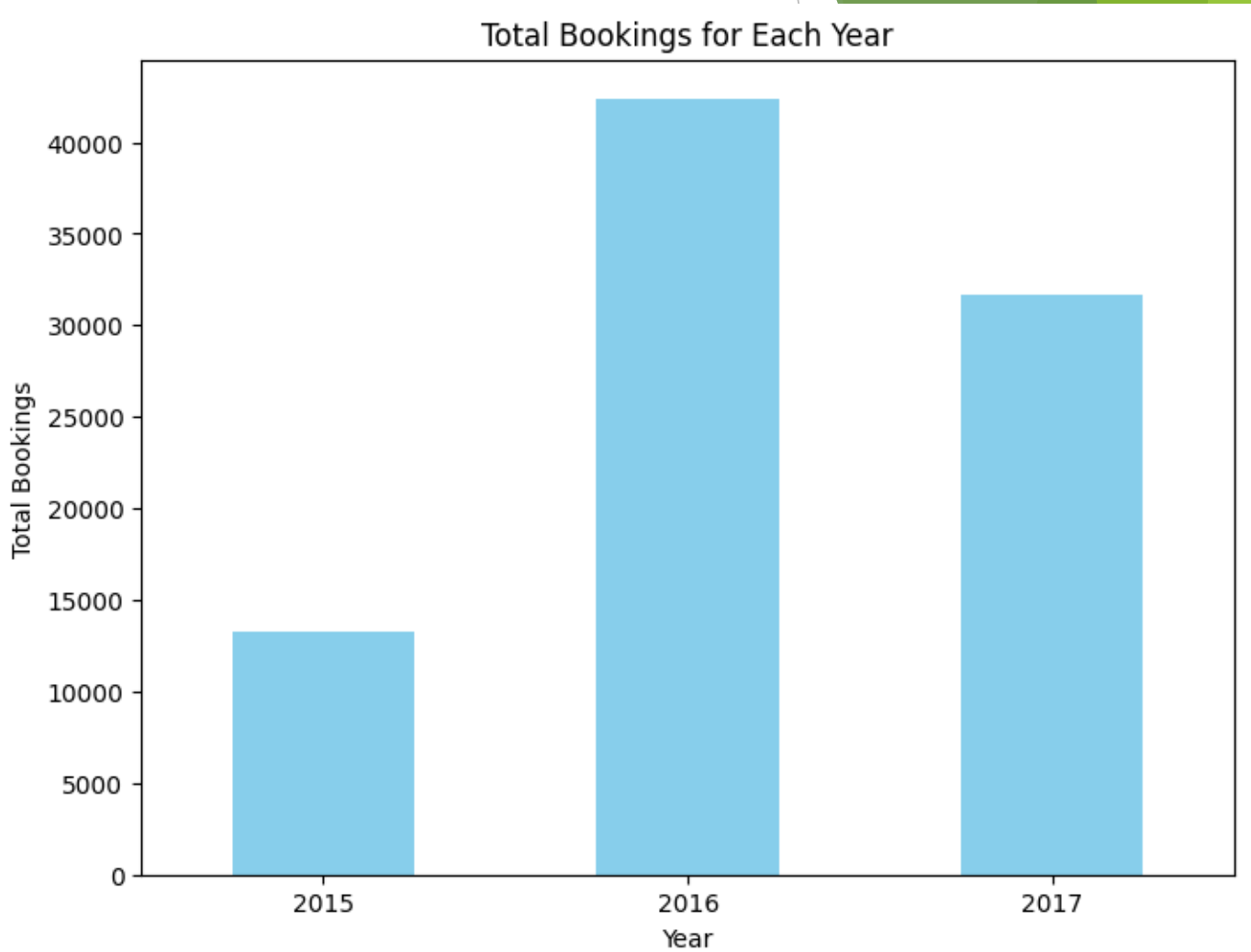
Very large amount of hotels have “No Deposit” Policy. And this may be the reason for cancellation of high amount of bookings. To avoid this booking cancellation, in account to collect more profit and customers- “No Deposit” policy should be change.



▫ Total Number Of Bookings Across Different Years

How many customers actually checked-in to the hotel across different years?

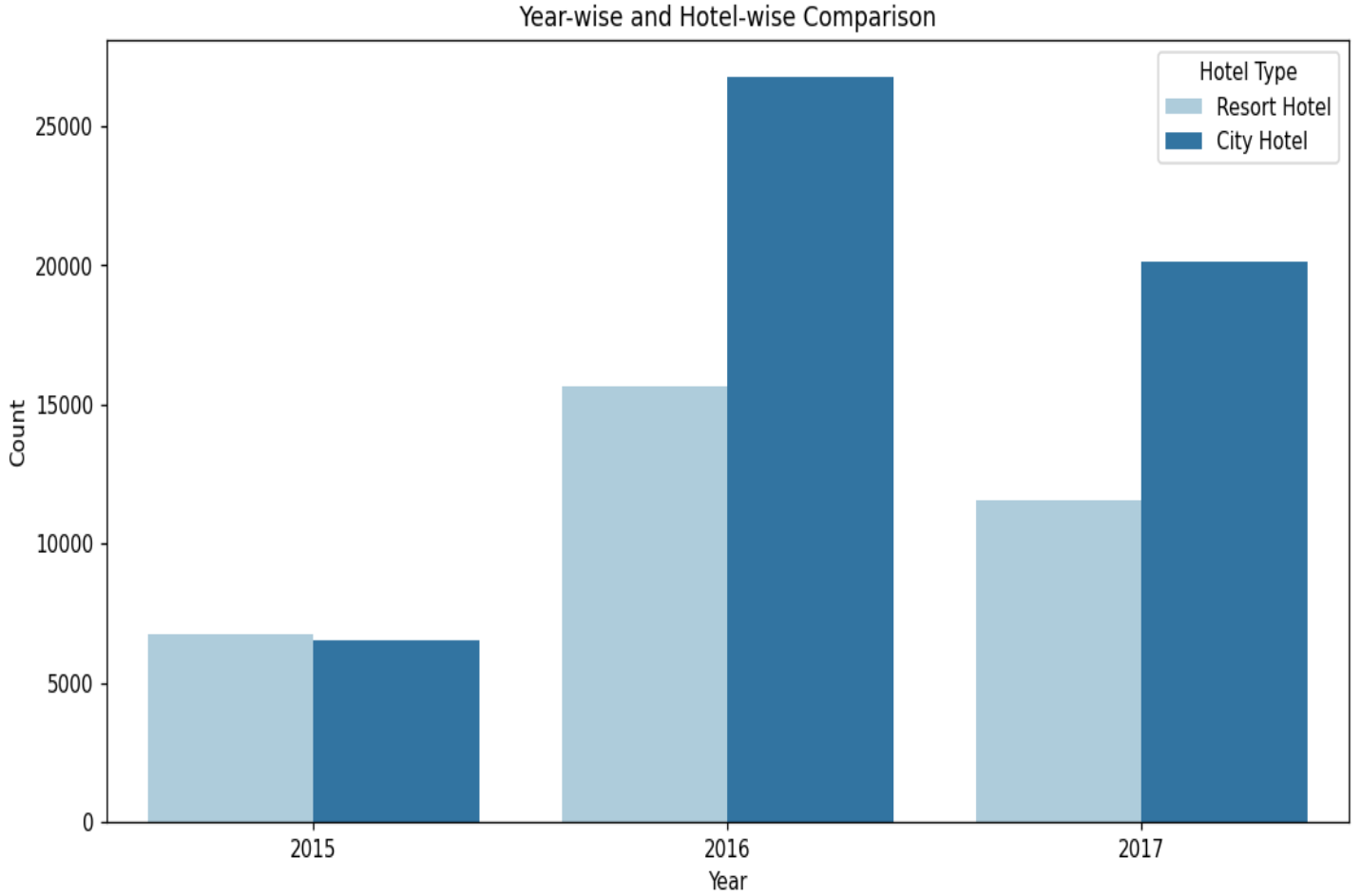
Let us find out with simple bar chart, in 2015 there are 15.23% of customers checking in. Whereas in 2016 we can see that there is increase in bookings up to 48.5%. This increase in trend did not sustain for more time, going downward in 2017 with only 36.26% bookings.



▫ Demand Trend Of Hotels Year wise

Which type of hotels customer preferred to stay in different years?

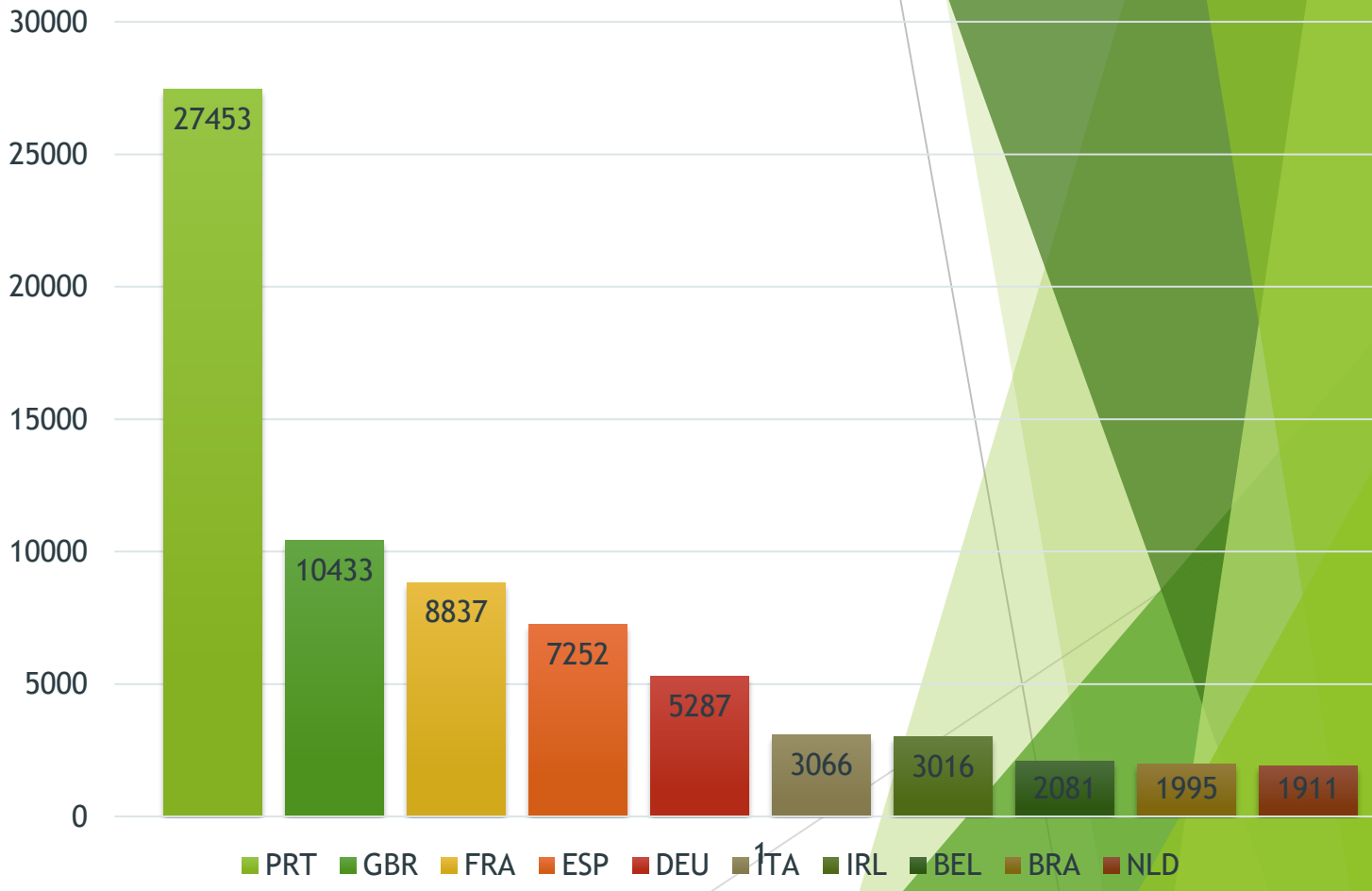
Here, we plotted a subplot for Resort hotel and City hotel. From these columns we can conclude that there is always demand of City hotels as compared to Resort hotels across three different years 2015, 2016 and 2017. As we discussed early, after increasing the booking trend it got decreased again. This happened in both cases – Resort as well as for City hotels.



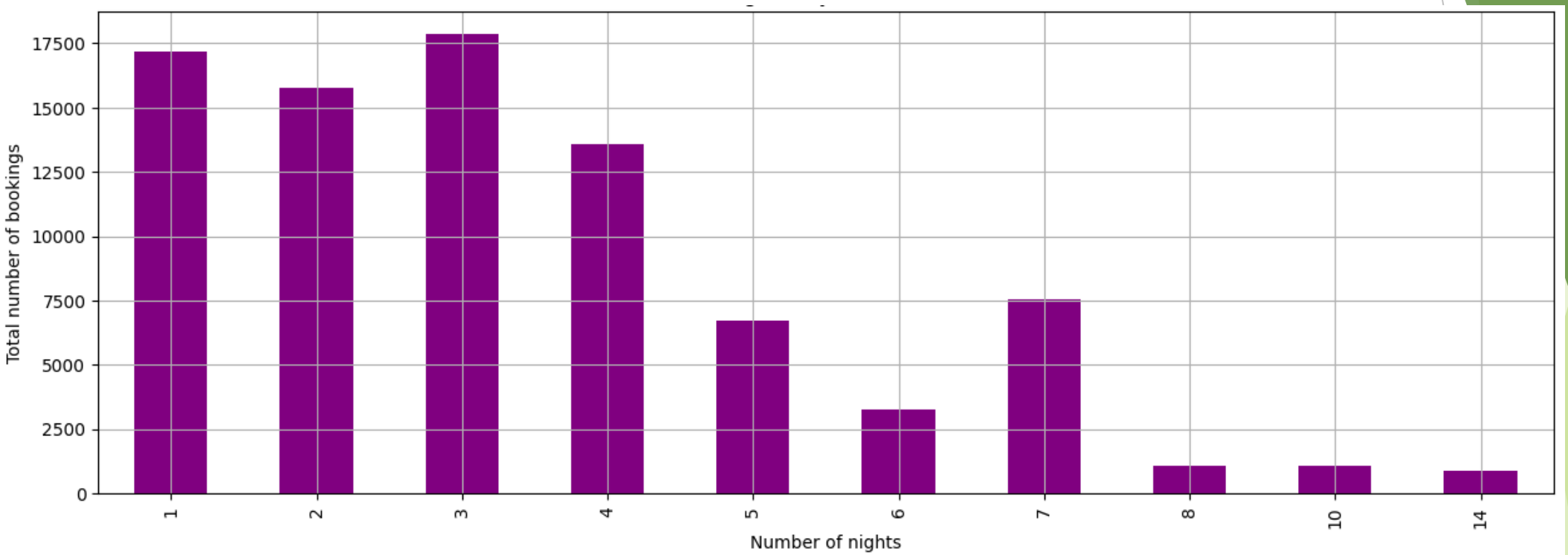
Top 10 Countries With Maximum Customers

Which are those countries giving maximum customers?

After analyzing the dataset, we found that Portugal (PRT) has the highest number of customers 72453, followed by the United Kingdom (GBR), France (FRA), Spain (ESP), Germany (DEU), Italy (ITA), Ireland (IRL), Belgium (BEL), Brazil (BRA), and the Netherlands (NLD)with 10433, 8837, 7252, 5287, 3066,3016,2081,1995,1911 respectively, in terms of the number of customers Netherlands sits back with lowest number of customers.



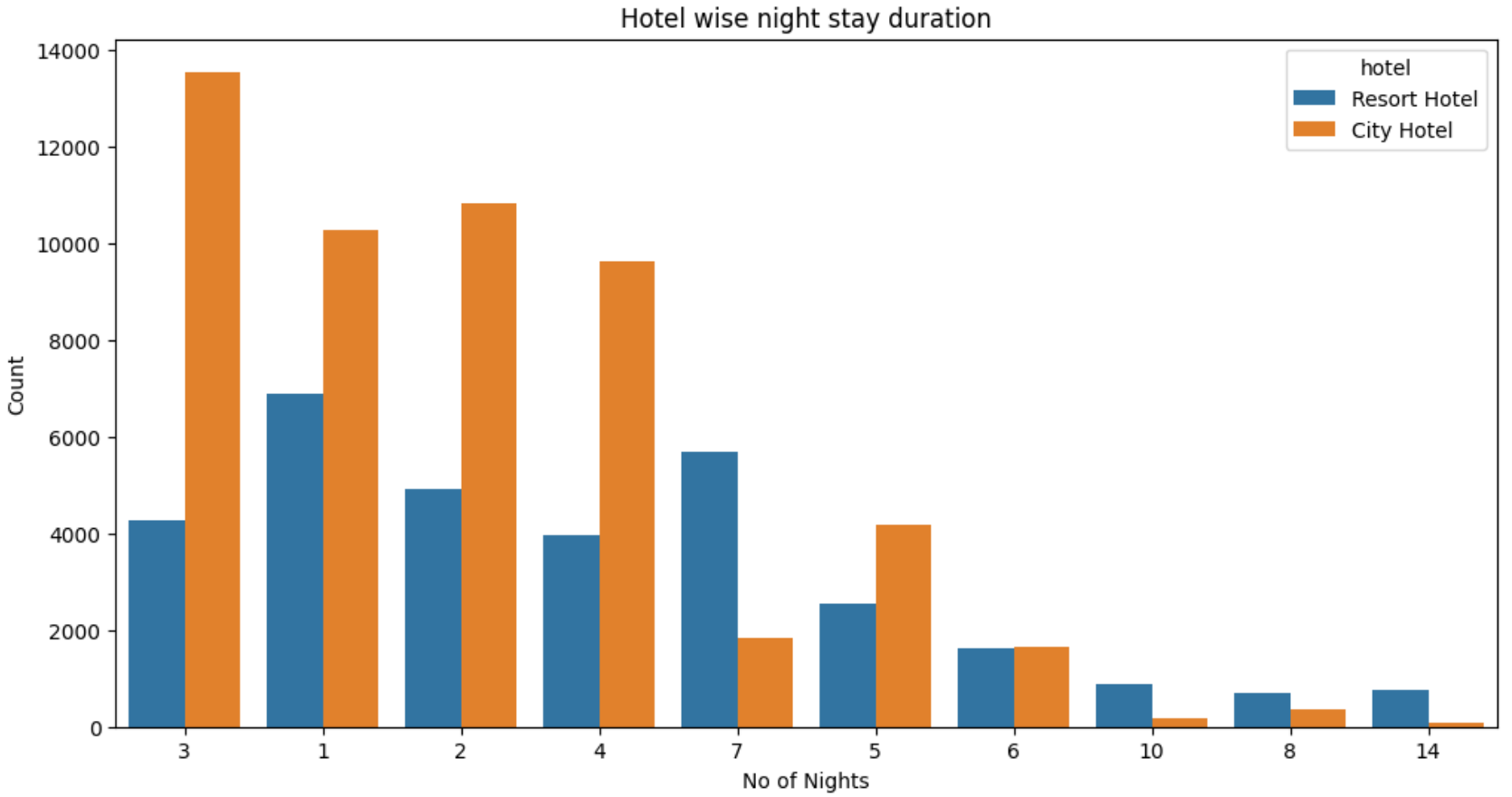
▫ Night Stay Duration



By combining the two columns of stays_in_week_nights and stays_in weekend_nights we got total number of nights. Hence, we can say that more customers like to spend 2 – 3 nights where some customer prefer to stay for 1 – 4 nights. Very few customers are there who are interested to stay for more than 5 days.

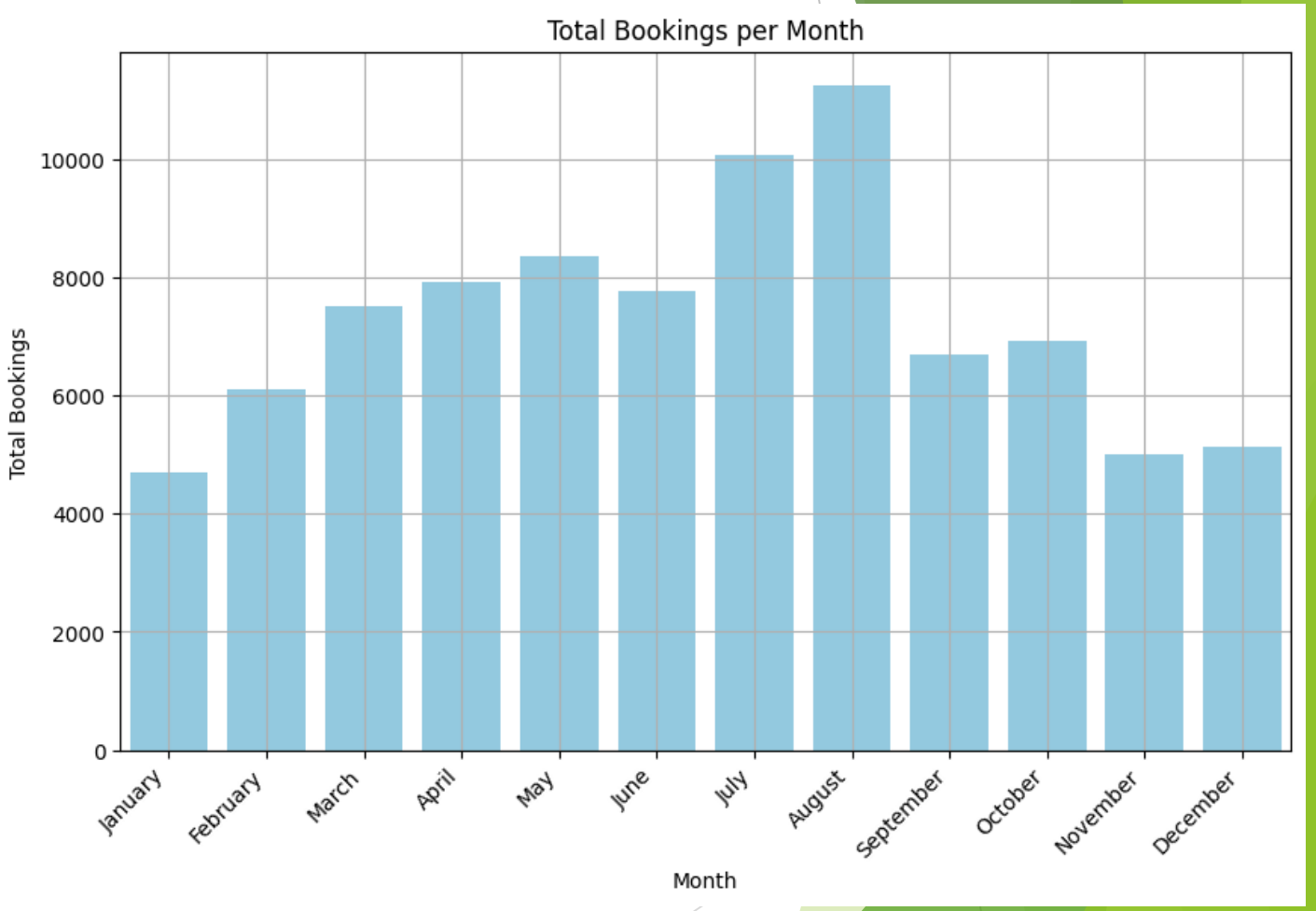
Hotel wise Night Stay Duration

Now we are going to track night stay duration of customers according to Resort hotels and City hotels. As we already aware that customers loves to stay in City hotels, here also City hotels have large amount of bookings for 2-3 night stay duration and then 1 night stay and 4 night stay customers are there for City hotels. In Resort hotels, 1 night stay customers are more and then 7 night stay customers comes in focus. Very few customers likely to stay for 8 night or more than it for both type of hotels.



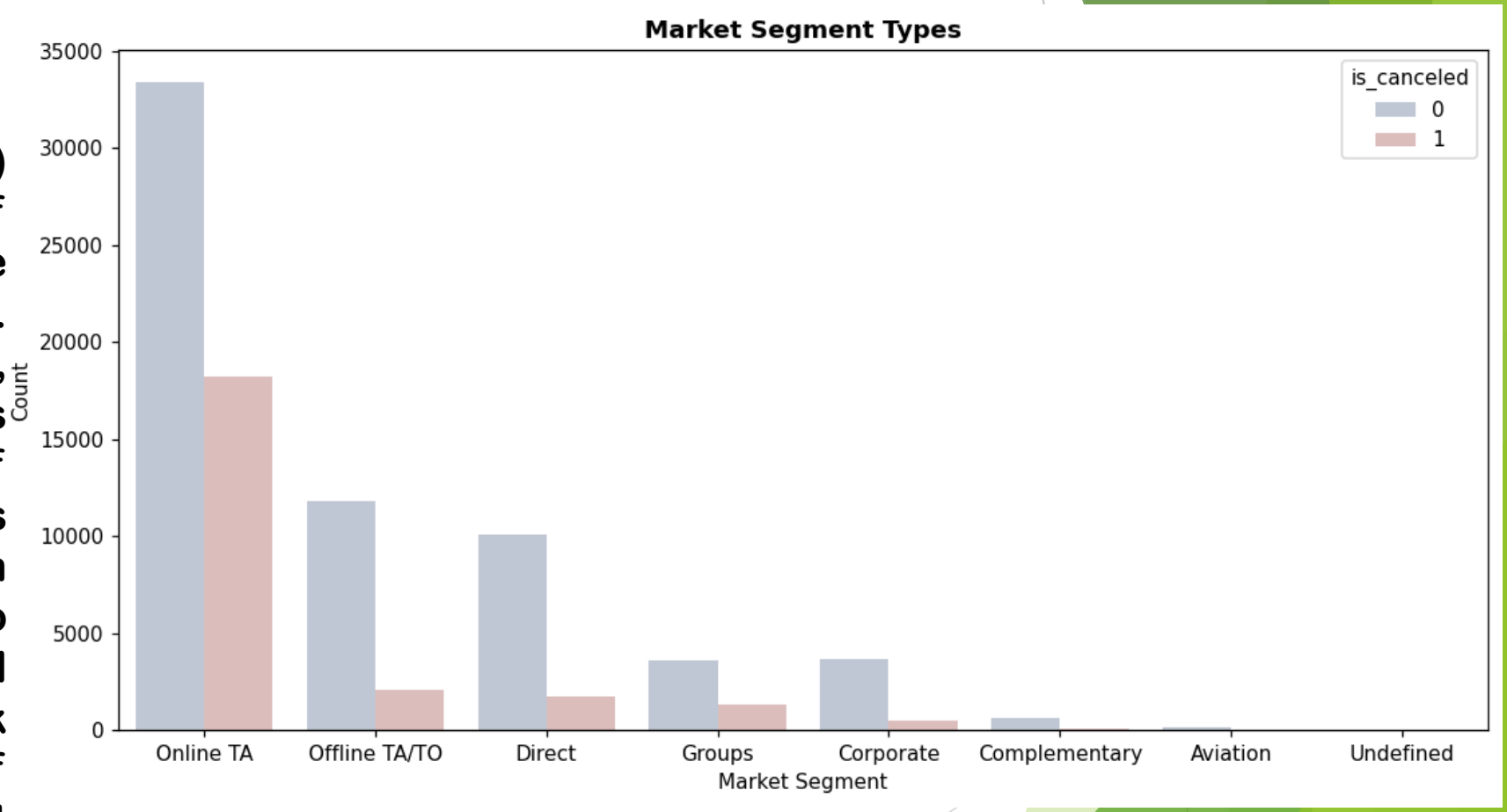
Booking Trend Through The Year

If we go through booking data along with different months, we found out that August has the highest number of bookings throughout the year then July is at second place where January has the lowest number of bookings i.e., we can assume that January will be the best month for booking to get the best rate on daily basis where booking in month of August will not be economical since it has high demand of room bookings obvious that the cost will also be high.



Total Number Of Bookings Across Various Market Segment

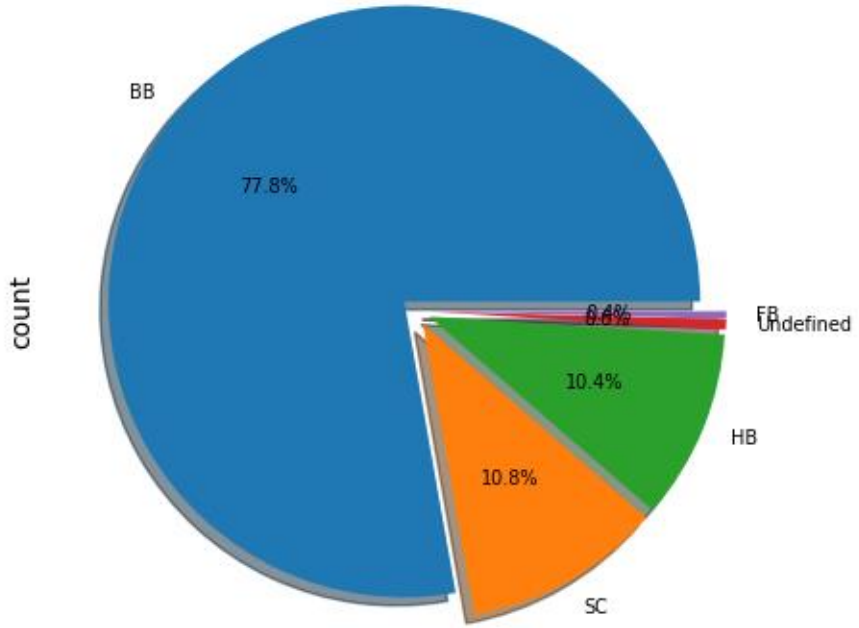
Online TA (Travel Agency) segment gives high amount of customers and then Offline TA/TO, Groups, Direct etc. respectively. Complementary, Aviation and Undefined has the lowest amount of customers. 8. So , from this we conclude that We can target our marketing area to be focus on these travel agencies website and work with them since majority of the visitors tend to reach out to them



▫ Meal Category vs Count Of Booking

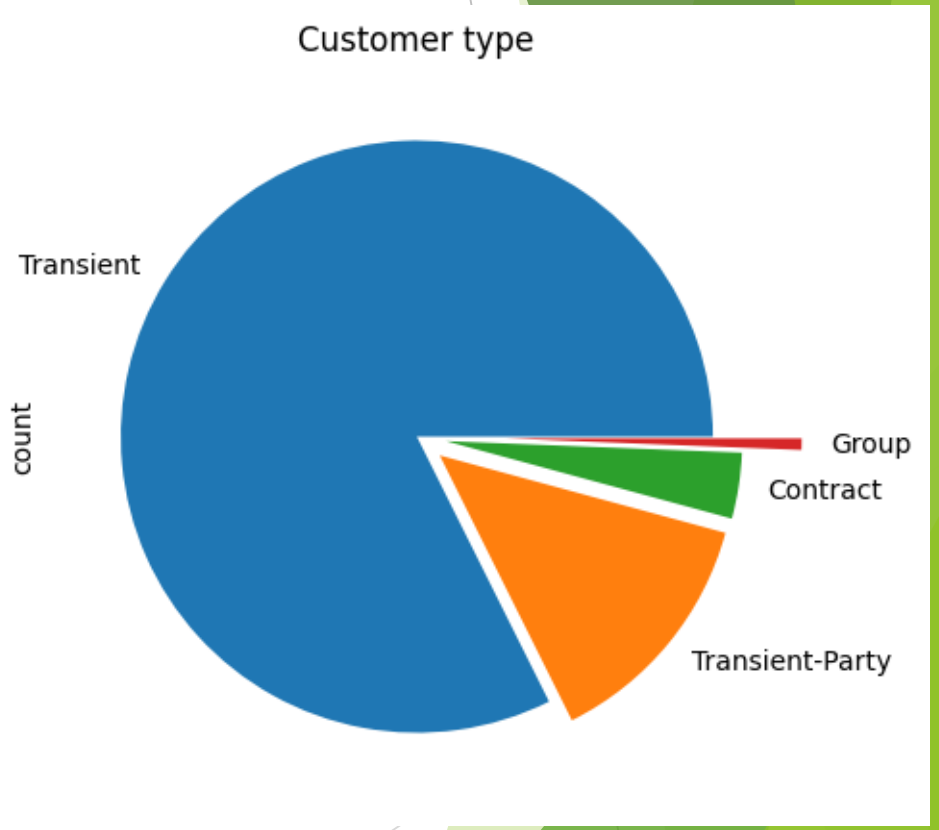
- Undefined/SC — no meal package
- BB — Bed & Breakfast
- HB — Half board (breakfast and one other meal — usually dinner)
- FB — Full board (breakfast, lunch and dinner)

Maximum of the bookings are made with bed and breakfast .So,
BB type of meal category is the most preferable in all type of customers, where negligible bookings are made with FB type of meal.

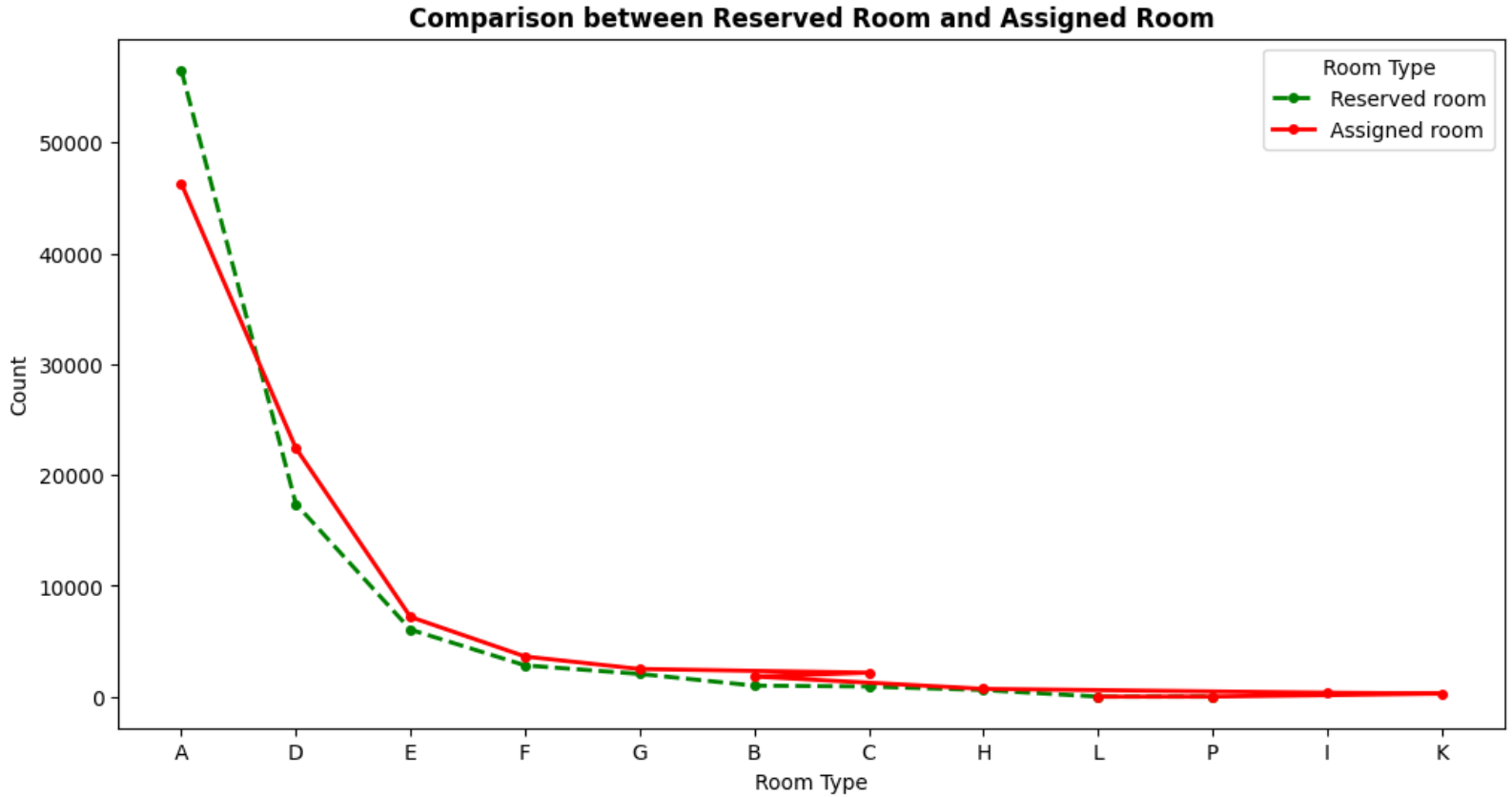


Booking vs Customer Type

Contract — when the booking has an allotment or other type of contract associated to it 2. Group — when the booking is associated to a group 3. Transient — when the booking is not part of a group or contract, and is not associated to other transient booking 4. Transient-party — when the booking is transient, but is associated to at least other transient booking • This means that the booking is not part of a group or contract. With the ease of booking directly from the website, most people tend to skip the middleman to ensure quick response from their booking. • Transient type of customer is the main source of booking because 75% of booking coming from this side after that Transient-Party, Contract and Group are coming in the focus.



Booking Trend With Respect To Room Type



A - type of room is the most favorite in all types of customers covering all the market about more than 85% ,the D – type of room is at second place in que while negligible customers are there which are ready to stay in L - type and P – type of room. 8. So we need to upgrade L- type and P – type of room to attract more customers so that no one should be in waiting list and do not search any other hotel which results in increasing the profit of hotels as more customers will book the rooms in the hotels.

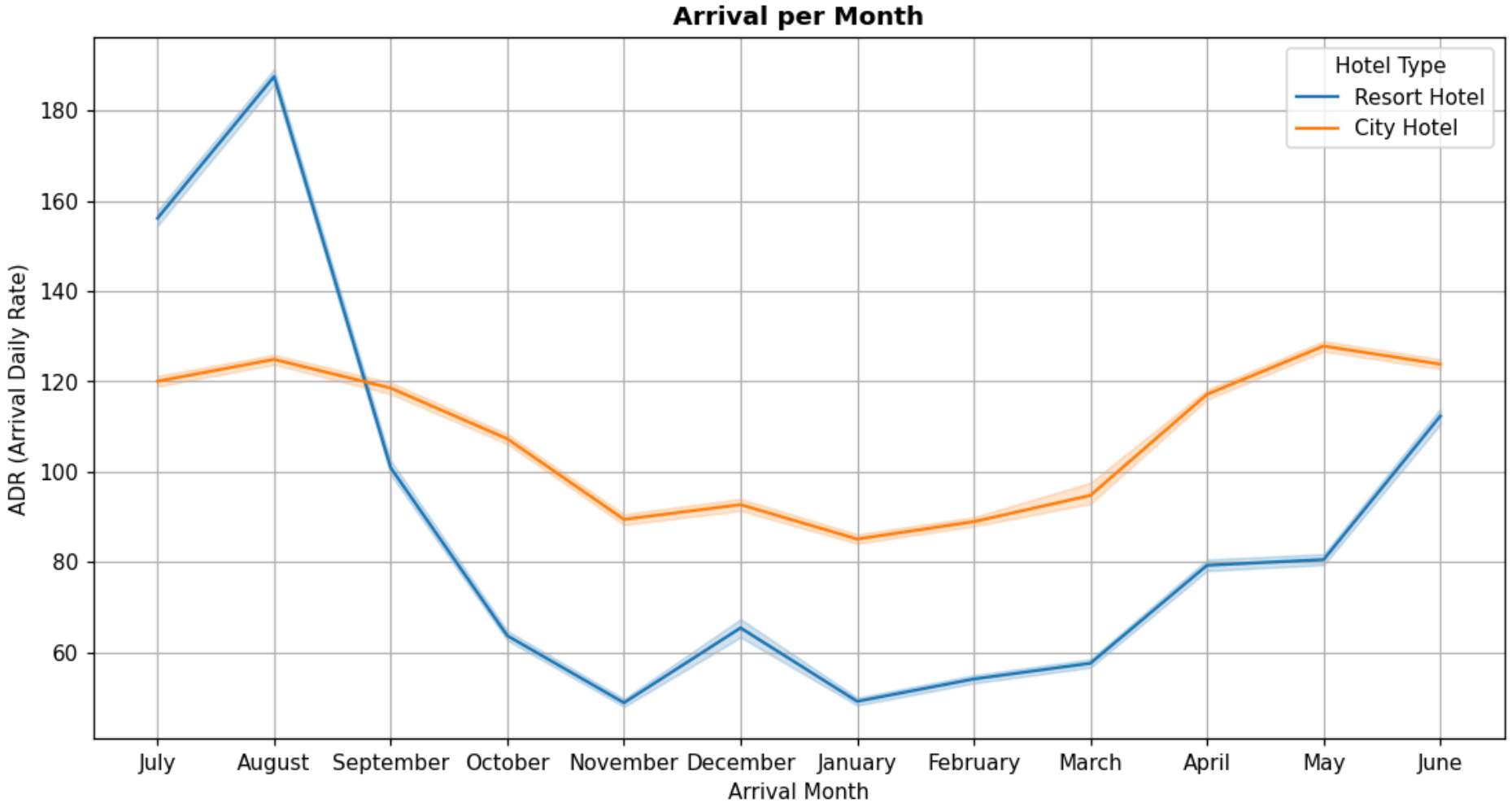
Variation In ADR Across Different Months

```
# Grouping the arrival accoridng to the month and finding the mean of ADR
df.groupby(['arrival_date_month','hotel'])['adr'].mean().unstack()
```

hotel	City Hotel	Resort Hotel
arrival_date_month		
April	117.156250	79.283805
August	124.901601	187.566659
December	92.717339	65.409093
February	88.945304	54.081107
January	85.092612	49.131584
July	120.055385	156.166914
June	123.836342	112.340141
March	94.763375	57.569213
May	127.851240	80.551101
November	89.454120	48.823434
October	107.304166	63.676313
September	118.546566	100.892331

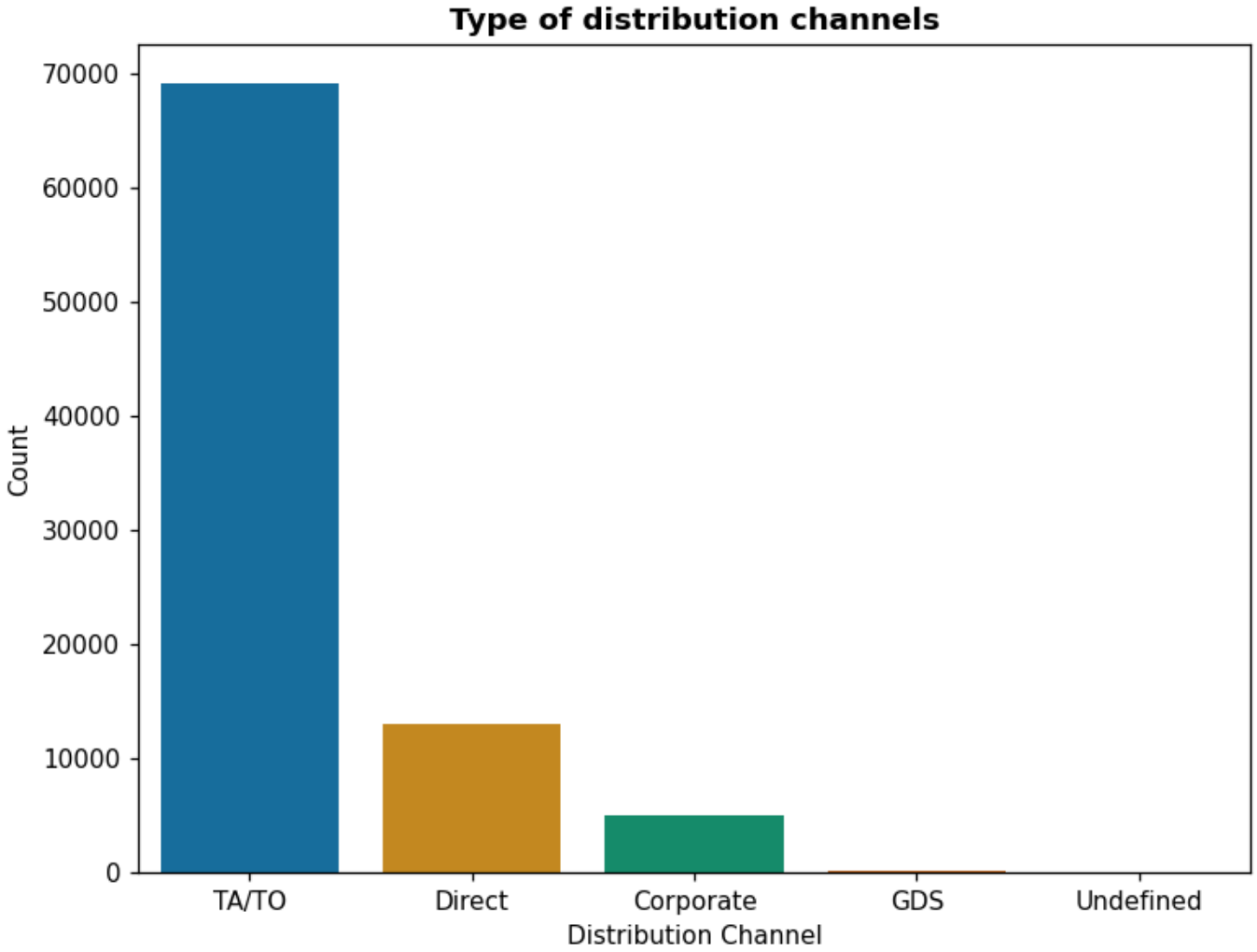
- For resort hotels, the Average Daily Rate (ADR) is more expensive during August, July, June and September where it is lower for January and November.
- For city hotels, the Average Daily Rate (ADR) is more expensive during August, July, May and June where it is lower also for January and November.

Variation In ADR Across Different Months



. So overall Average Daily Rate of both city hotels and resort hotels are more expensive between May and September.

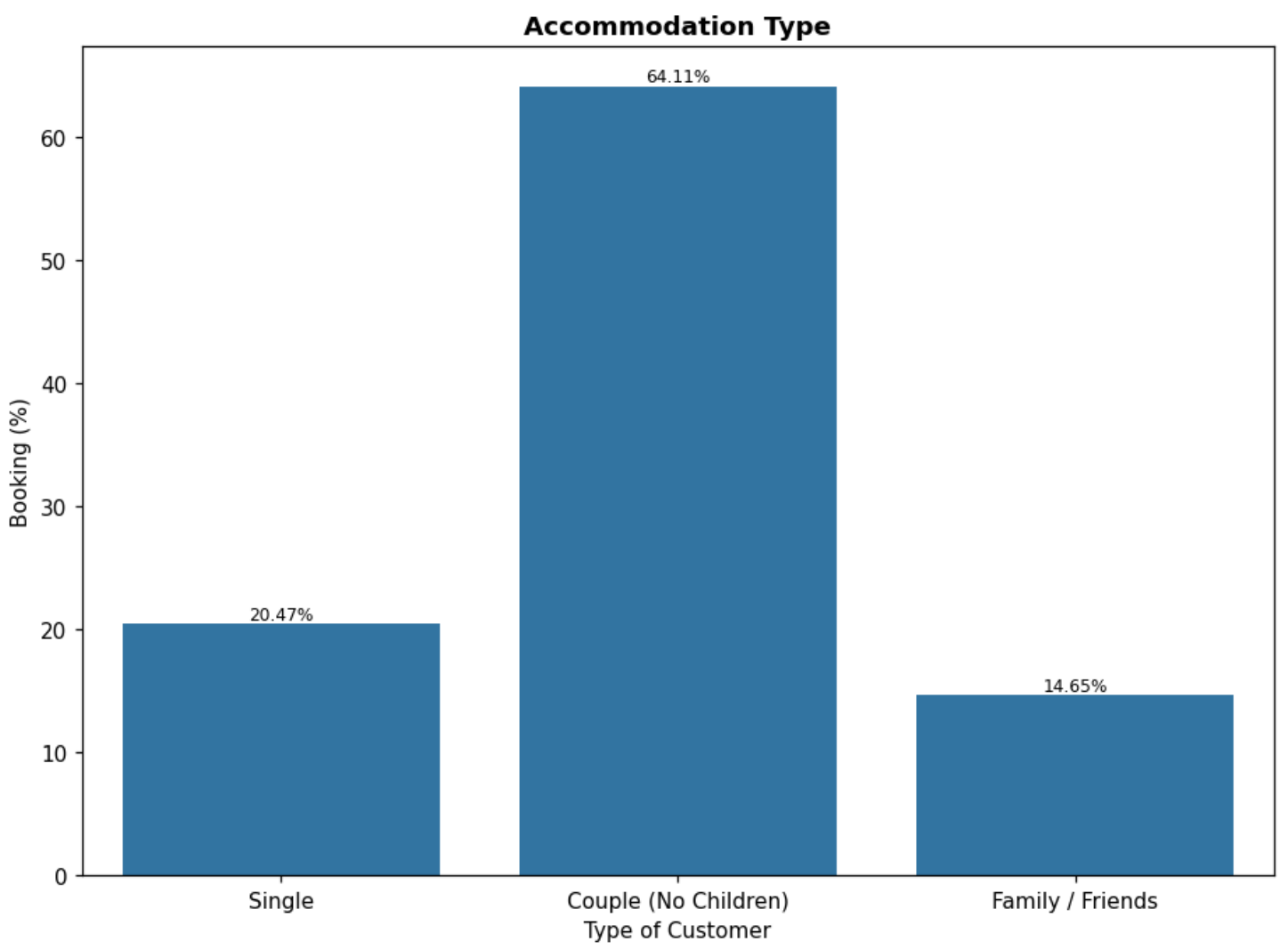
Distribution Channel



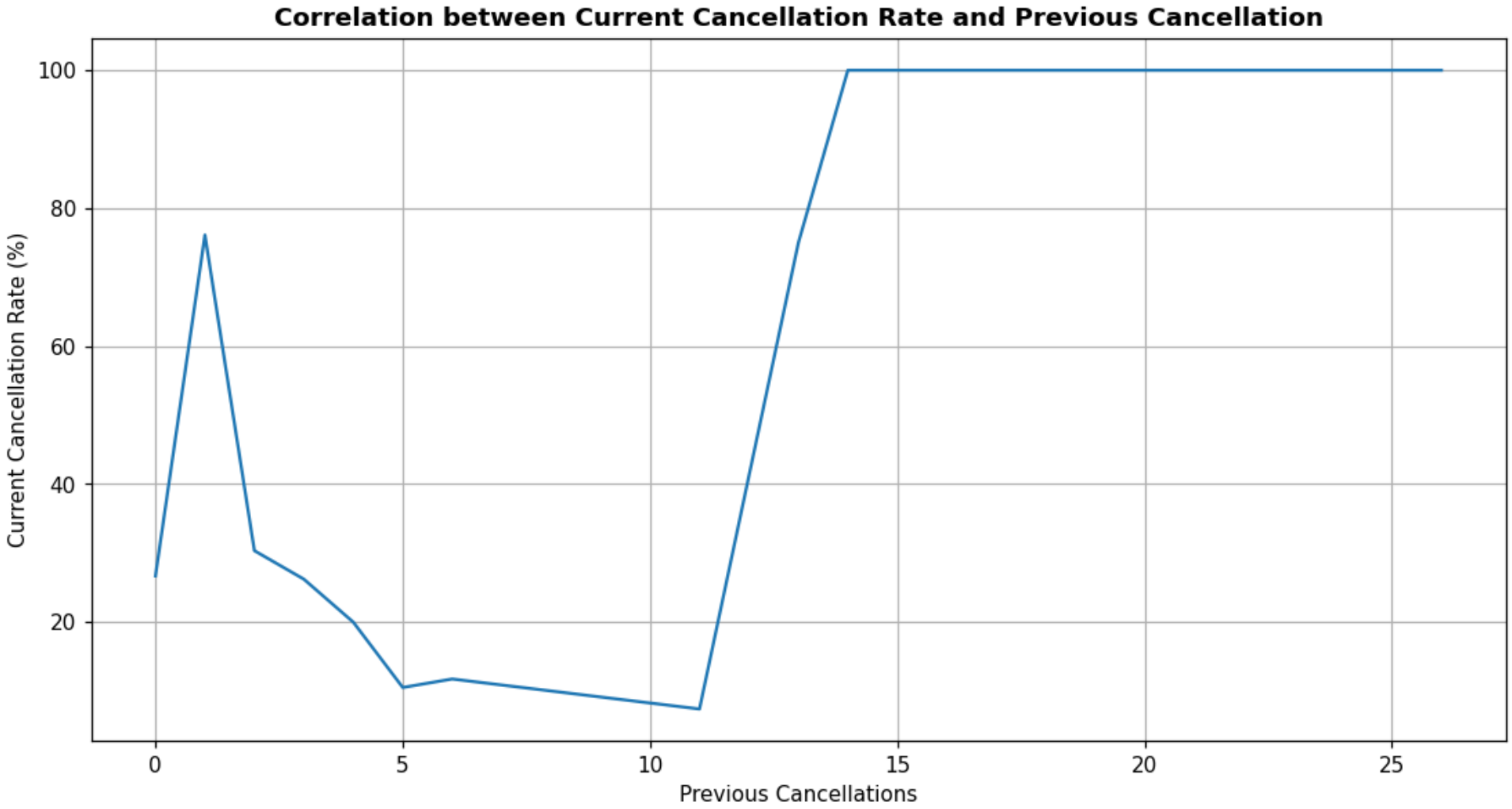
- From the side graph we observed that most of the customer preferred to book the hotels through TA/TO (Travel agent / Tour operators).

Accommodation Type – Single, Couple & Family

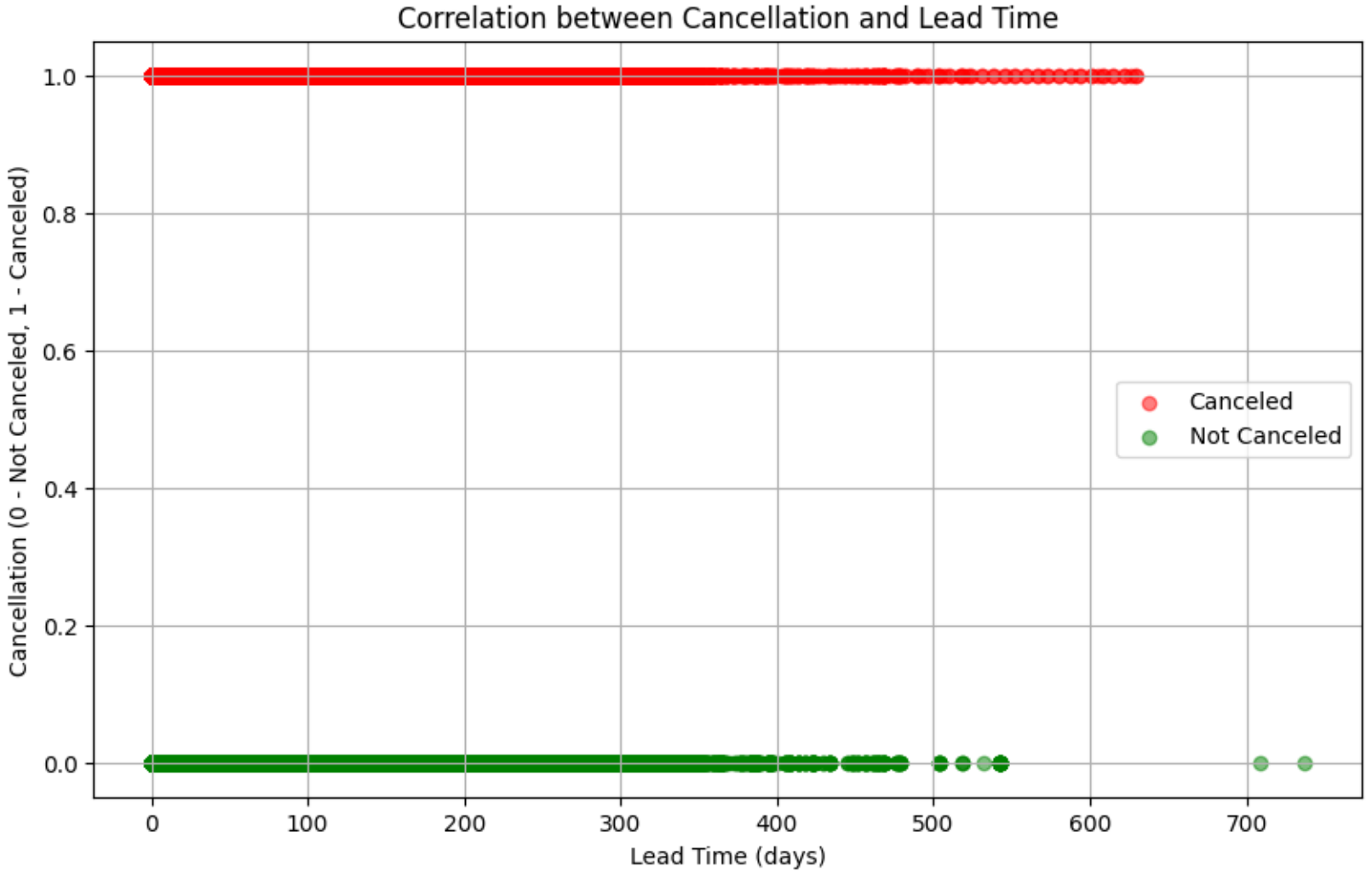
- From graph, it is seen that most number of customer are containing 2 adults customer.
- And those who have more than 2 either containing adults, children & babies have the lowest number of customer.



Correlation between previous and current cancellation

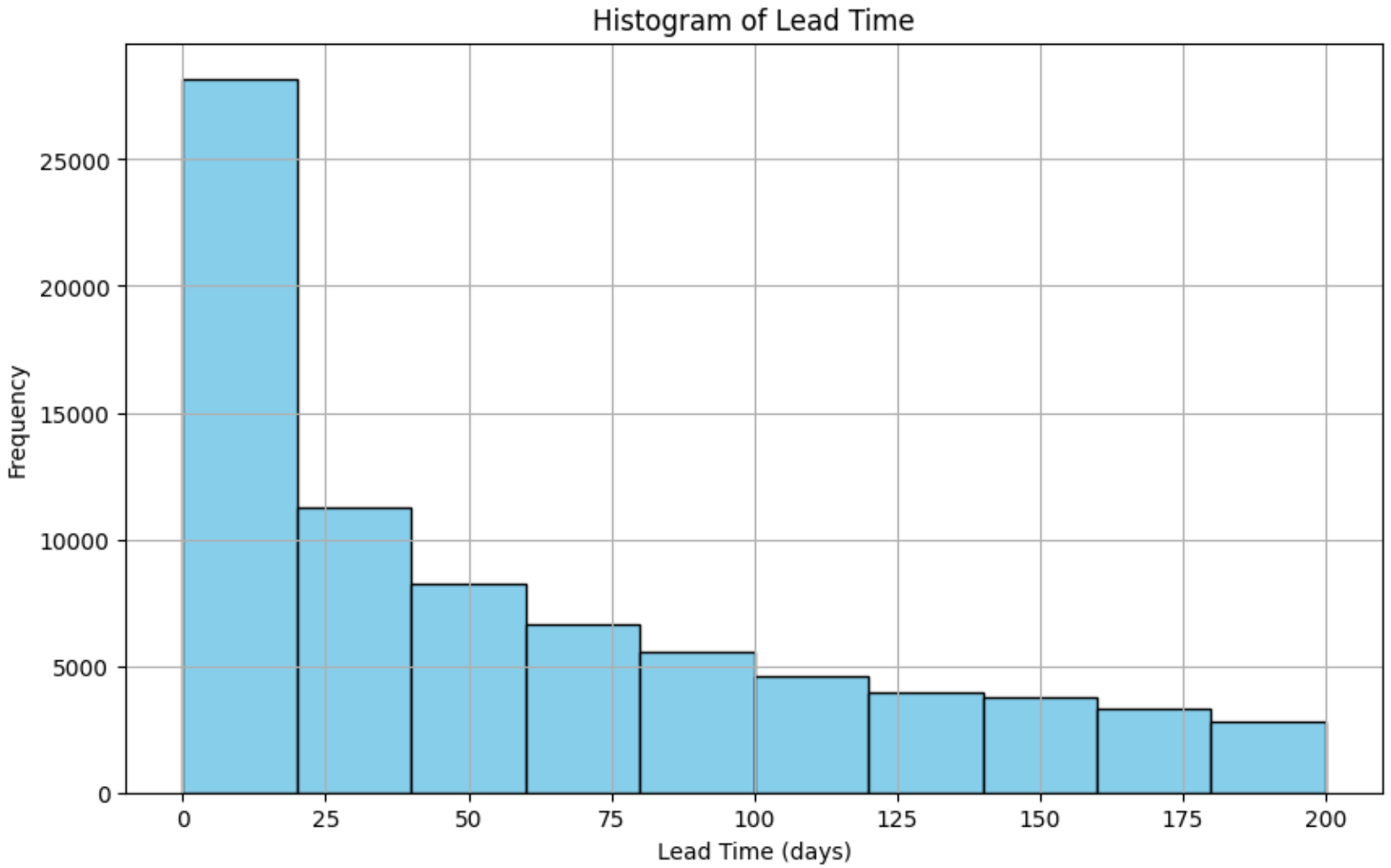


Correlation between cancellation and lead time

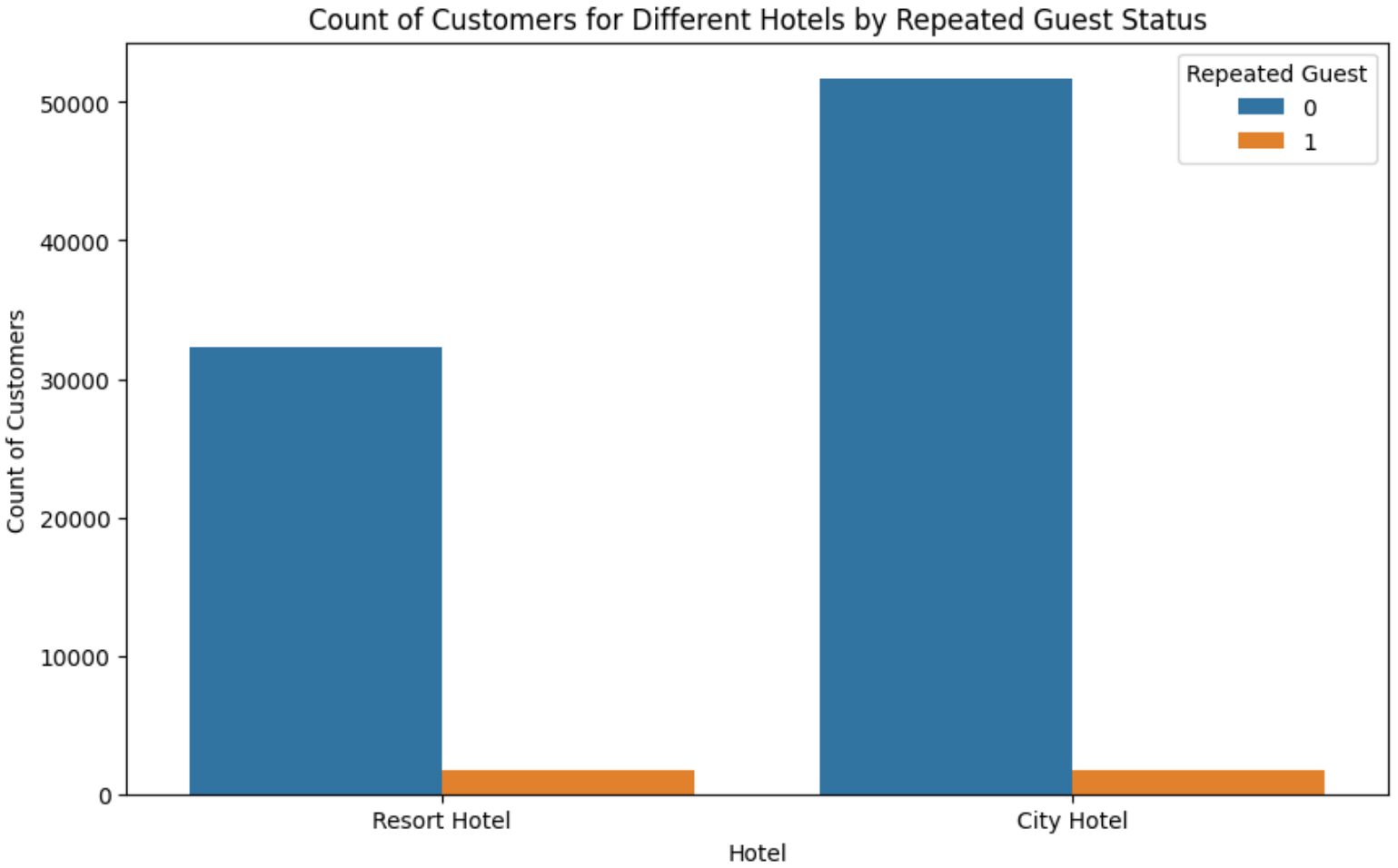


- When lead time increases cancellation of bookings also increases.
- Positive Correlation between cancellation and lead time.

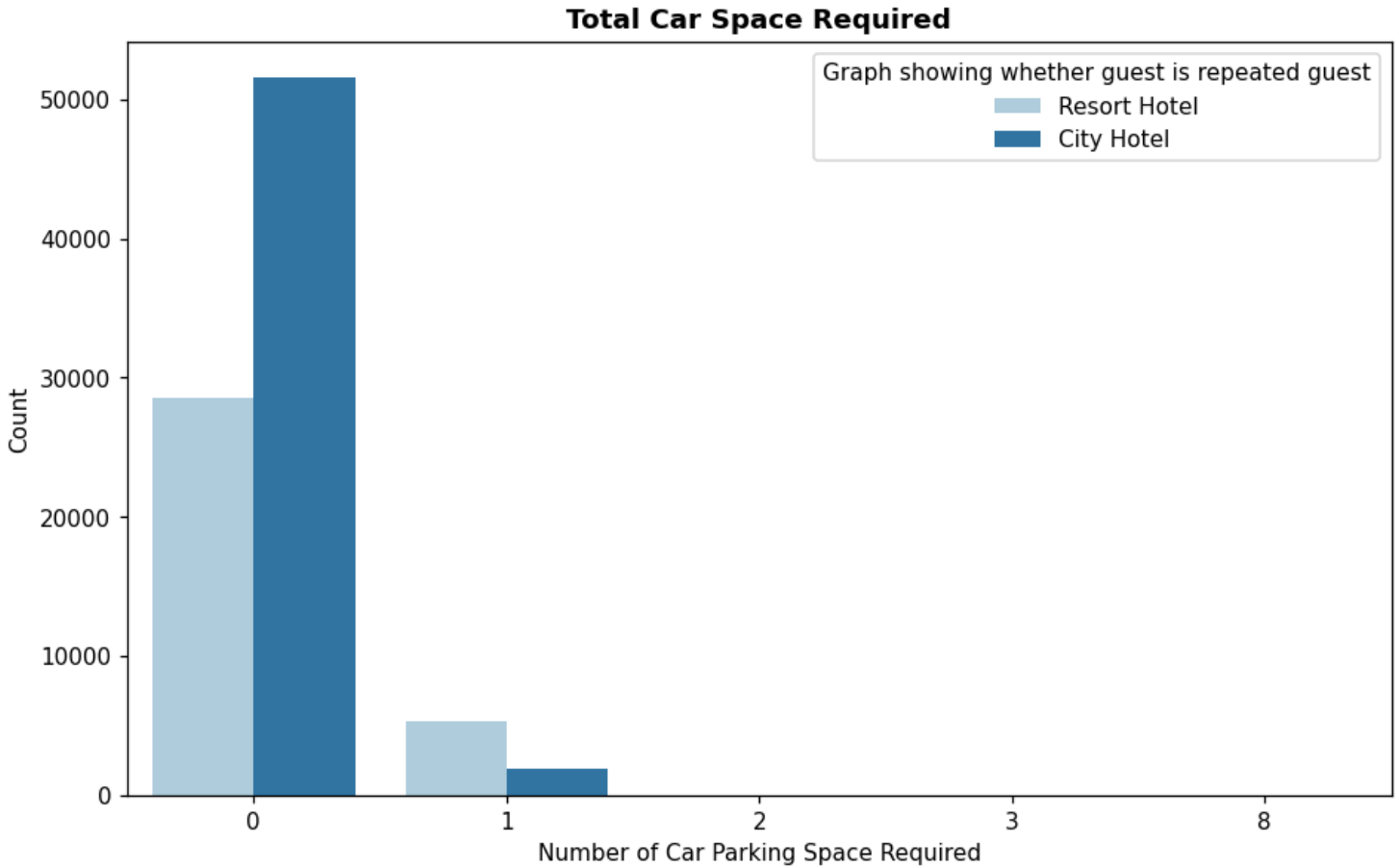
Lead Time Distribution



Bins of this map is from (0,200,10).



The chart reveals that both city hotels and resort hotels attract a substantial number of non-repeated guests. However, there may be variations in the proportion of repeated guests between the two hotel types. For instance, if resort hotels consistently attract more repeated guests than city hotels, it suggests that resort hotels may have a stronger appeal to returning customers, possibly due to their amenities, location, or overall guest experience.



This graphs shows that 93.8% of customer doesn't required any parking spaces.
And the maximum number of parking spaces required is 1 for 6.1% of the customers.
From this 6.1% most people are staying in Resort hotel as compare to City hotel

Unstacking of the Data

```
# We can find the mean of ADR across market segment per day
df.groupby(['arrival_date_day_of_month','market_segment'])['adr'].mean().unstack()
```

arrival_date_day_of_month

1	68.333333	5.309524	63.418906	122.489736	68.245862	80.807772	118.481339	NaN
2	96.000000	0.310345	66.490000	113.406917	67.155032	79.169010	113.920610	NaN
3	108.625000	1.458333	62.293554	107.901215	72.965229	81.183242	115.540622	12.0
4	95.000000	4.767391	64.667687	111.929026	70.341358	75.534286	114.001549	NaN
5	110.250000	0.200000	69.340000	112.230708	62.509458	77.400042	116.554532	18.0
6	95.000000	5.133333	68.597833	107.142579	67.455513	86.271620	117.076435	NaN
7	105.893636	3.391000	67.594074	117.351381	62.448077	76.485915	122.712641	NaN
8	93.916667	0.000000	66.546928	116.171108	72.334805	80.600046	120.119616	NaN
9	106.666667	3.761905	68.295605	107.552439	73.921708	82.964989	118.138858	NaN
10	99.090909	1.784483	65.538413	118.889029	69.644622	81.699649	118.881547	NaN
11	101.000000	0.321429	70.765034	113.092127	69.816036	80.930952	117.302451	NaN
12	99.909091	2.692308	64.241135	111.189860	67.386547	76.465946	116.569166	NaN
13	83.308333	0.300000	70.666953	120.328677	90.065046	81.176995	119.076046	NaN
14	102.500000	0.000000	65.512000	117.019153	78.730111	79.495878	117.771944	NaN
15	93.750000	14.903226	67.675694	123.585515	69.707602	82.512989	124.195855	NaN

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 because these are peak months.
- 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 more repeated guests, we should target our advertisement on guests to increase returning guests.

References

1) https://pandas.pydata.org/
2) https://matplotlib.org/
3) https://seaborn.pydata.org/
4) Geeks for Geeks
5) Ama Better
6) Stack overflow.
7) Kaggle.
8) https://github.com/KK-Niraj

THANK YOU !