Capstone Project-1

Hotel Booking Analysis

Team Members

Shivangi Mishra
Saksham Tripathi
Bindu Kovvada
Deepak Kumar Gautam
Satyajit Sahoo

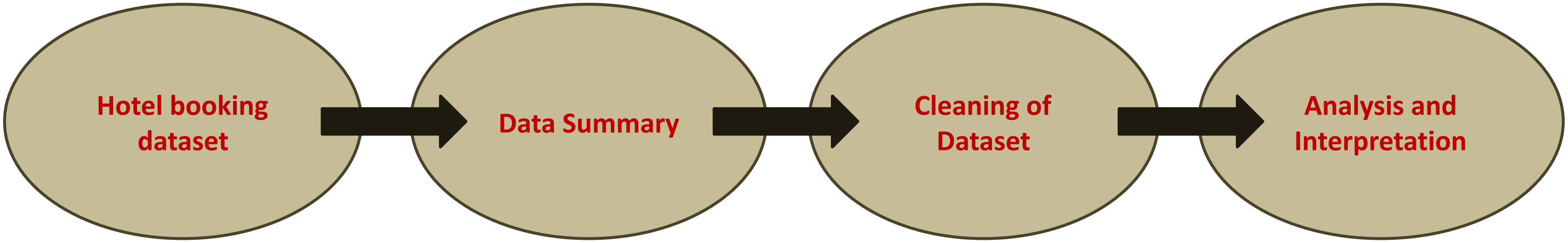
Flow of Presentation

- Agenda
- Data Summary
- Cleaning of Dataset
- Data Visualization
- Inferences
- Conclusion

Agenda

Agenda is to discuss the given problem statement i.e., Analysis of several factors influencing hotel bookings in the given hotel booking dataset.

APPROACH:



Data Summary

Data set name:

Hotel Booking Database that contains booking information for a city hotel and a resort hotel of various countries from 2015 to 2017.

Data shape:

Rows - 119390

Columns - 32

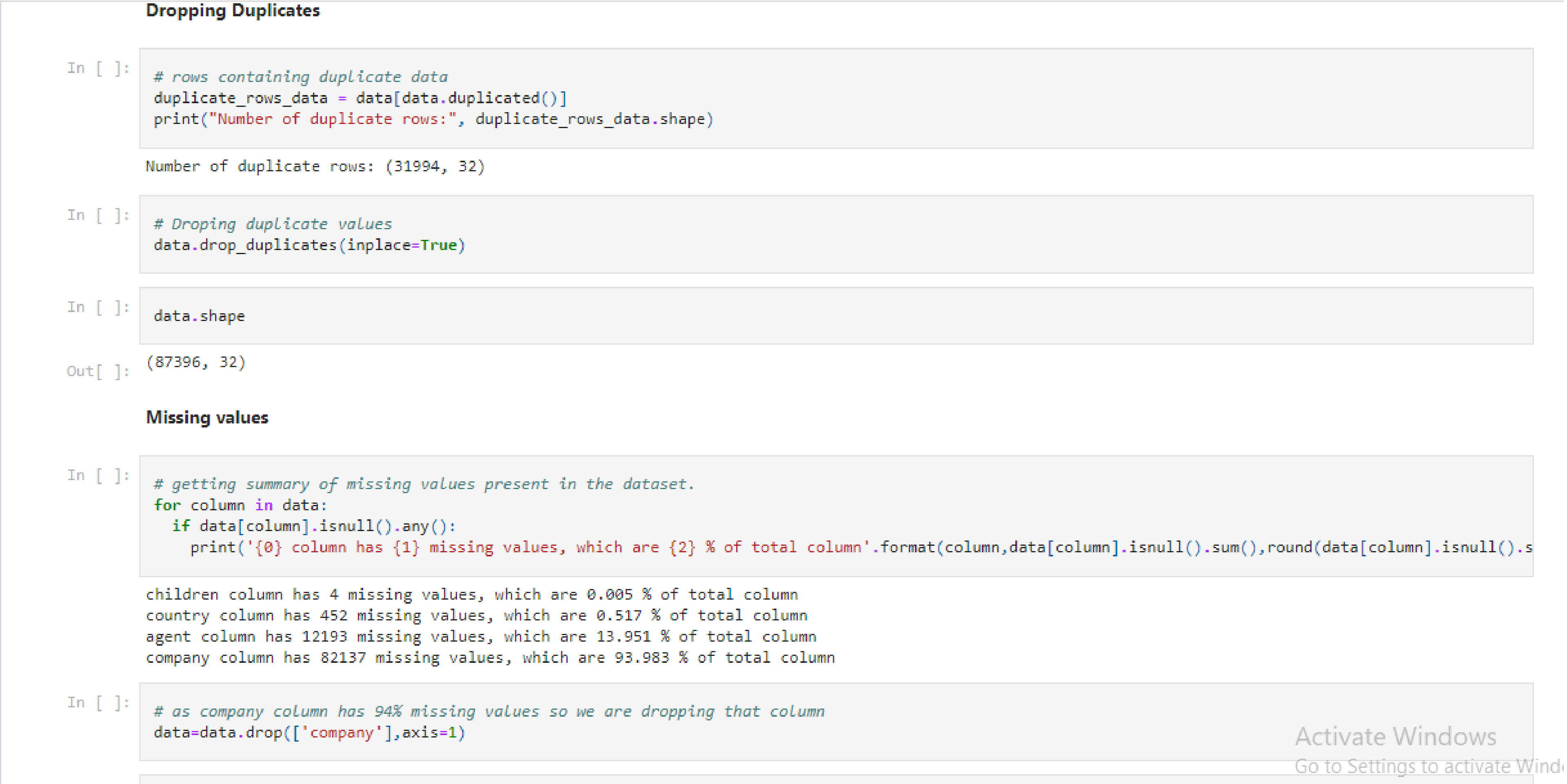
Columns used:

- hotel - name of hotel whether City Hotel or Resort Hotel
- is_canceled –(0 or1) Indicates whether booking was cancelled or not.
- lead_time – The time between reservation and actual arrival.
- 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 the guest stayed or booked to stay at the hotel.
- Stays_in_week_nights : Number of week nights the guest stayed or booked to stay at the hotel.

- adults : Total number of adults in hotel
- children : Total number of children in hotel
- babies : Total number of babies in hotel
- meal : Type of meal booked i.e. Bed & Breakfast (BB), Half Board (HB), Full Board (FB), Undefined contain no meal package
- country : Country of customers
- market segment : A group of people who share one or more common characteristics used for business
- distribution_channel : Chain of business through which a service passes until it reaches the final buyer
- previous_cancellations – (0 or 1)previous cancellation by customer
- previous_booking_not_canceled- confirmed booked by customer

- reserved_room_type – Type of Room booked
- assigned_room_type – Type of Room assigned/alloted
- booking_charges – booking charges we charged
- deposit_type – No Deposit, Non Refund, Refundable
- agent- ID of travel agency
- company- ID of the company
- days_in_waiting_list – Number of days in waiting
- adr- Average daily rate
- required_car_parking_space – Number of car parking spaces required
- total_of_special_request – Number of special request requested
- reservation_status - Canceled, Check out
- reservation_status_date – Canceled / check out
- is_repeated_guest –(0 or 1) contain data of repeated guest
- customer_type – Type of customer Contract / Group / Transient

Cleaning of Dataset



Dropping Duplicates

```
In [ ]:
# rows containing duplicate data
duplicate_rows_data = data[data.duplicated()]
print("Number of duplicate rows:", duplicate_rows_data.shape)
```

```
Number of duplicate rows: (31994, 32)
```

```
In [ ]:
# Droping duplicate values
data.drop_duplicates(inplace=True)
```

```
In [ ]:
data.shape
```

```
Out[ ]:
(87396, 32)
```

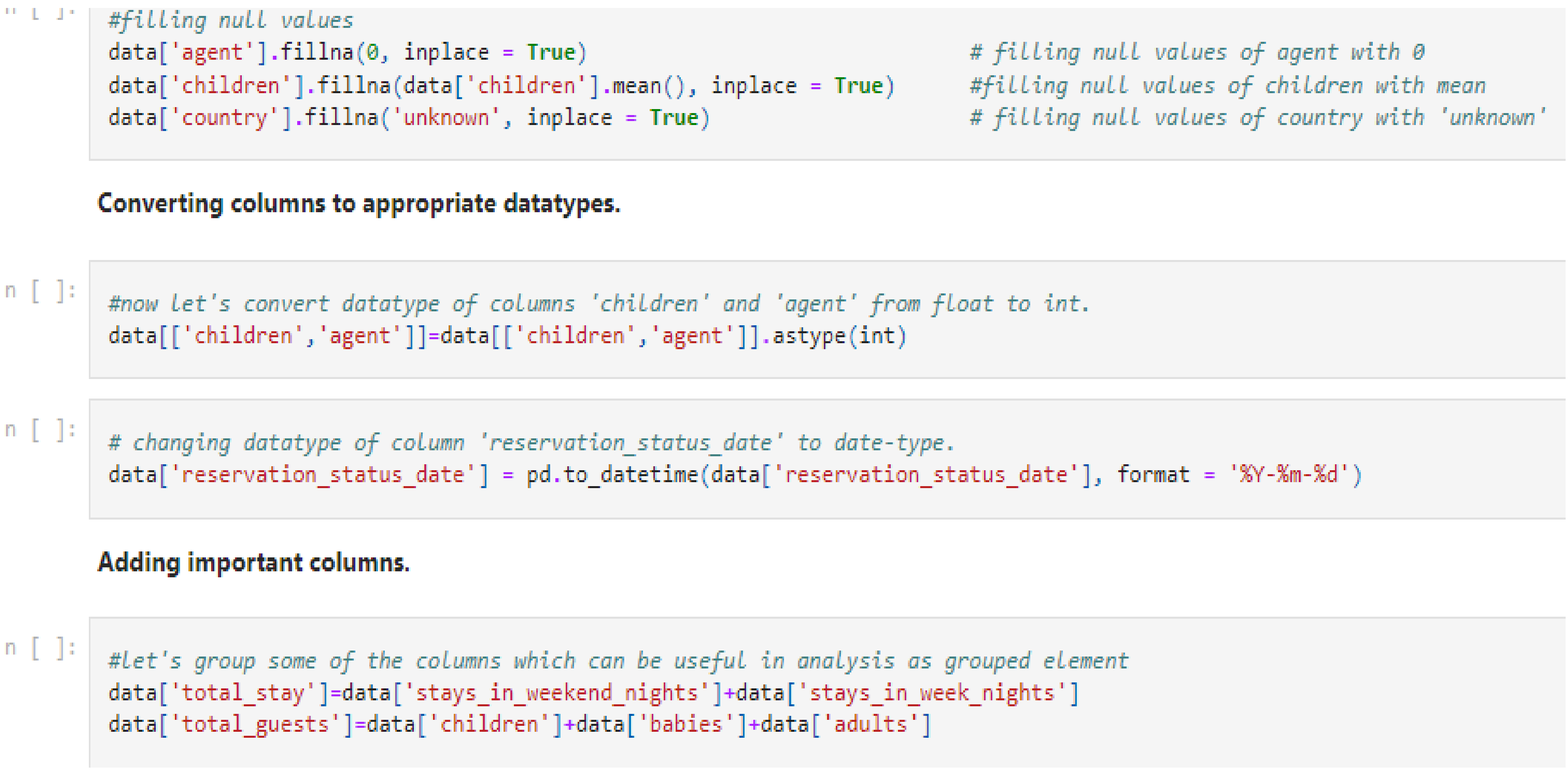
Missing values

```
In [ ]:
# getting summary of missing values present in the dataset.
for column in data:
  if data[column].isnull().any():
    print('{0} column has {1} missing values, which are {2} % of total column'.format(column,data[column].isnull().sum(),round(data[column].isnull().s
```

```
children column has 4 missing values, which are 0.005 % of total column
country column has 452 missing values, which are 0.517 % of total column
agent column has 12193 missing values, which are 13.951 % of total column
company column has 82137 missing values, which are 93.983 % of total column
```

```
In [ ]:
# as company column has 94% missing values so we are dropping that column
data=data.drop(['company'],axis=1)
```

Cleaning of Dataset



```
#filling null values
data['agent'].fillna(0, inplace = True)                                 # filling null values of agent with 0
data['children'].fillna(data['children'].mean(), inplace = True)        #filling null values of children with mean
data['country'].fillna('unknown', inplace = True)                       # filling null values of country with 'unknown'
```

Converting columns to appropriate datatypes.

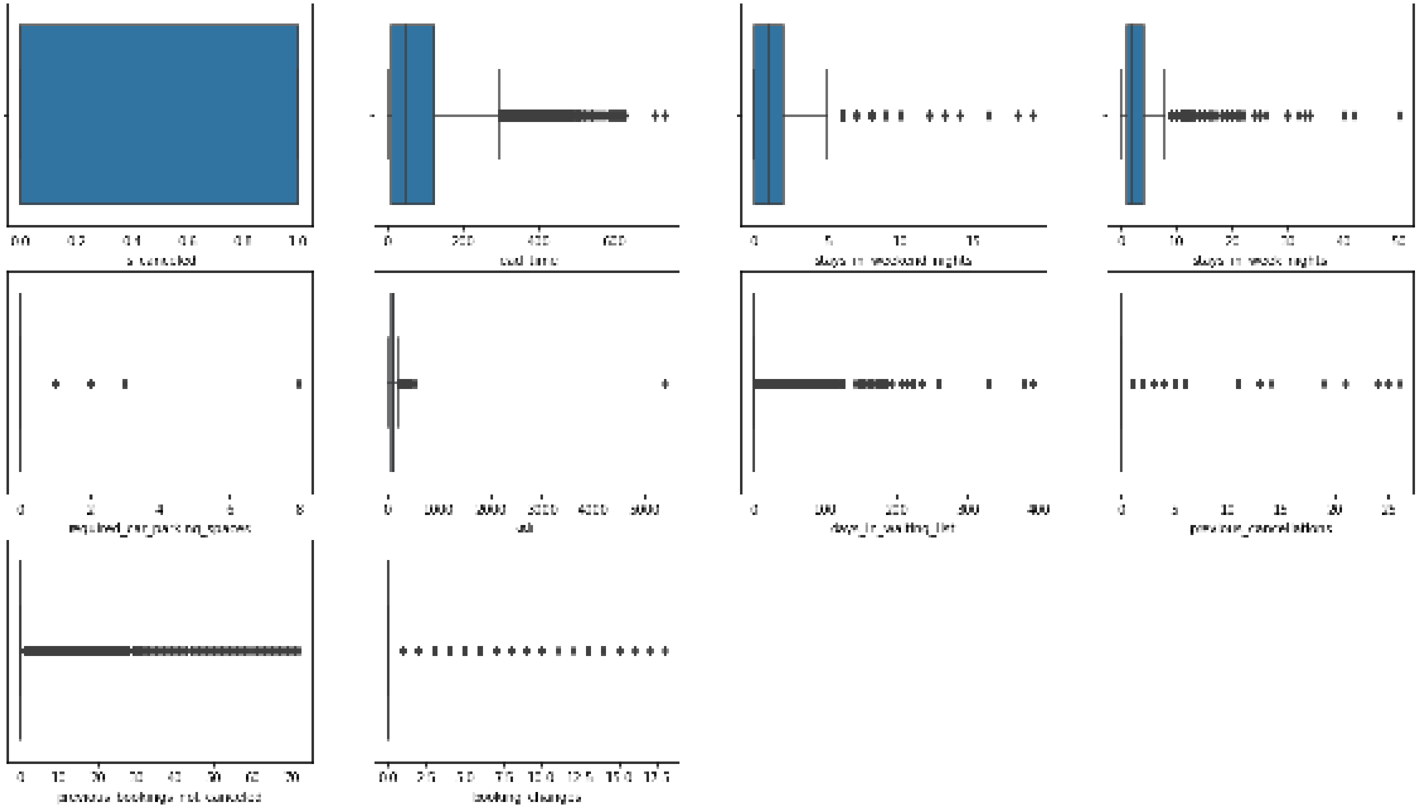
```
#now let's convert datatype of columns 'children' and 'agent' from float to int.
data[['children','agent']]=data[['children','agent']].astype(int)
```

```
# changing datatype of column 'reservation_status_date' to date-type.
data['reservation_status_date'] = pd.to_datetime(data['reservation_status_date'], format = '%Y-%m-%d')
```

Adding important columns.

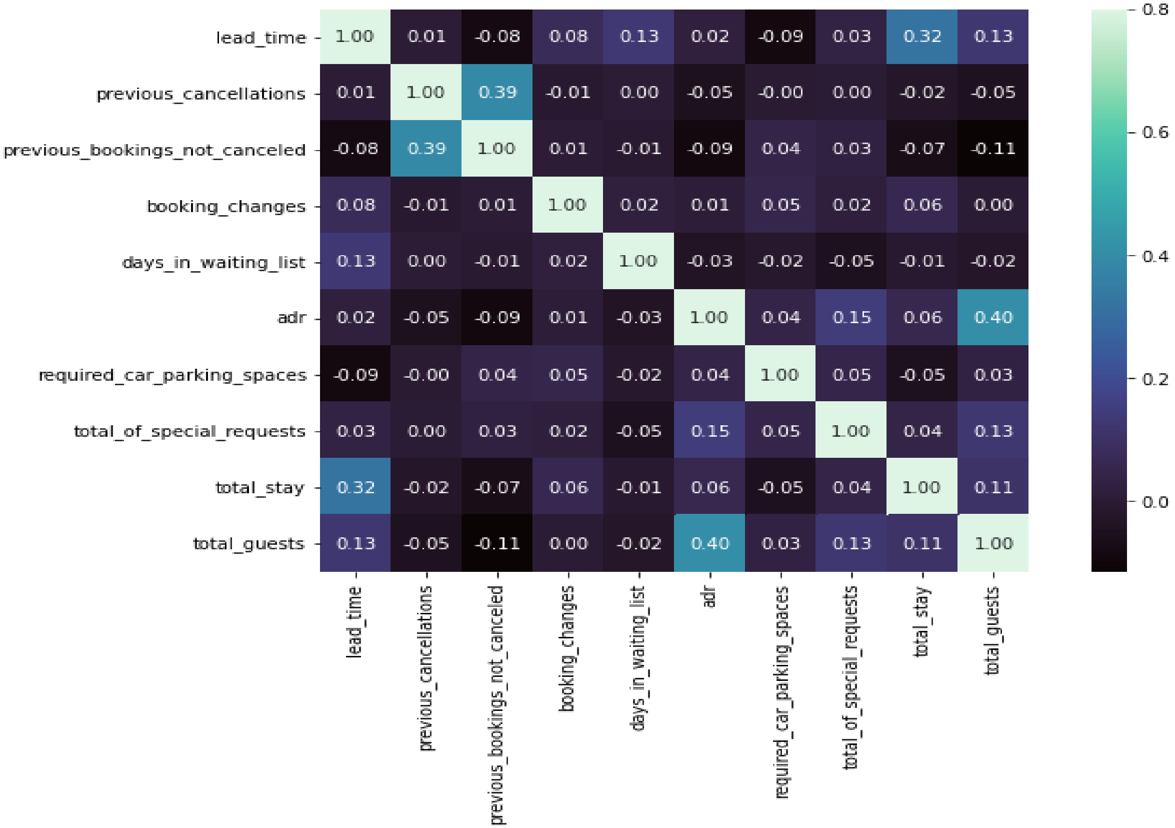
```
#let's group some of the columns which can be useful in analysis as grouped element
data['total_stay']=data['stays_in_weekend_nights']+data['stays_in_week_nights']
data['total_guests']=data['children']+data['babies']+data['adults']
```

Cleaning of Dataset



✔ **As we can see that this dataset has many outliers and thus we can infer that this dataset is not very reliable.**

Data Visualization



	lead_time	previous_cancellations	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_stay	total_guests
lead_time	1.00	0.01	-0.08	0.08	0.13	0.02	-0.09	0.03	0.32	0.13
previous_cancellations	0.01	1.00	0.39	-0.01	0.00	-0.05	-0.00	0.00	-0.02	-0.05
previous_bookings_not_canceled	-0.08	0.39	1.00	0.01	-0.01	-0.09	0.04	0.03	-0.07	-0.11
booking_changes	0.08	-0.01	0.01	1.00	0.02	0.01	0.05	0.02	0.06	0.00
days_in_waiting_list	0.13	0.00	-0.01	0.02	1.00	-0.03	-0.02	-0.05	-0.01	-0.02
adr	0.02	-0.05	-0.09	0.01	-0.03	1.00	0.04	0.15	0.06	0.40
required_car_parking_spaces	-0.09	-0.00	0.04	0.05	-0.02	0.04	1.00	0.05	-0.05	0.03
total_of_special_requests	0.03	0.00	0.03	0.02	-0.05	0.15	0.05	1.00	0.04	0.13
total_stay	0.32	-0.02	-0.07	0.06	-0.01	0.06	-0.05	0.04	1.00	0.11
total_guests	0.13	-0.05	-0.11	0.00	-0.02	0.40	0.03	0.13	0.11	1.00

- ✔ Lead time and total stay have slight correlation. This means when people want to stay little longer they plan little before than actual arrival.
- ✔ Total guests and average daily rate has some correlation. This means the when the number of guests increases, adr will also increase.

Data Visualization

Following insights are pulled out from this analysis:

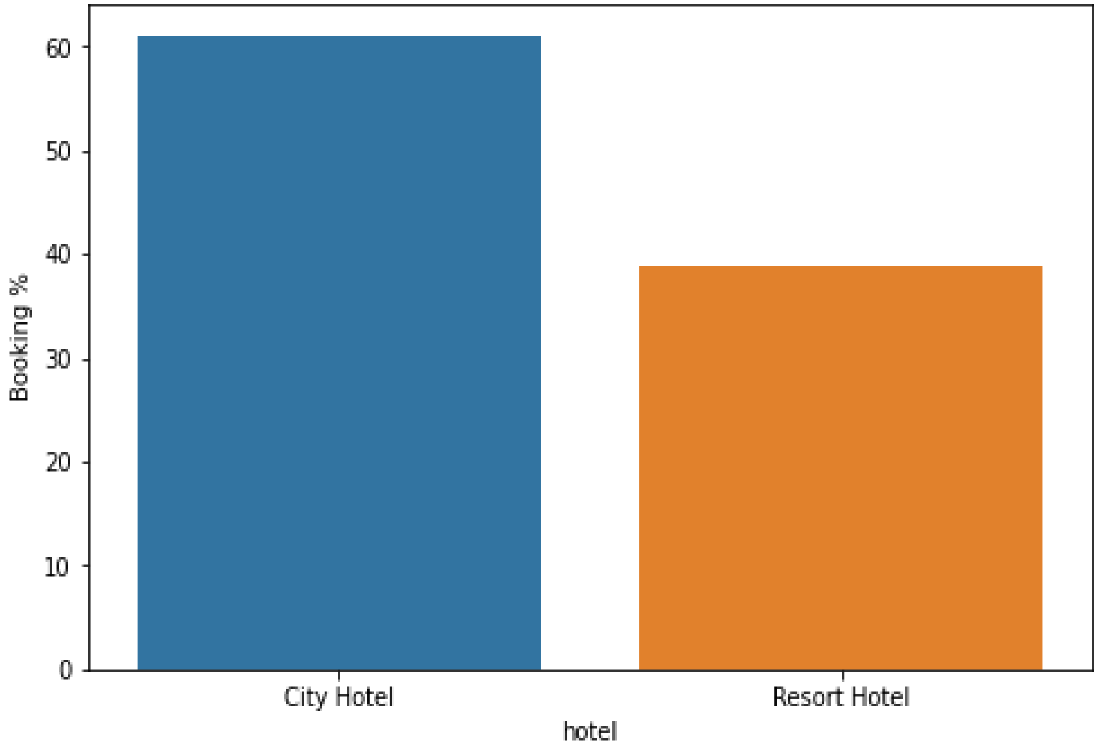
1. What is the percentage of bookings in each hotel type?
2. From which country most guests come?
3. Which is the busiest month for hotels?
4. Which room type is in most demand ?
5. Which room type generate highest adr?
6. Which meal type is most preferred meal of customers?
7. How many booking were cancelled?
8. Which type of customers are most repeated?
9. Booking cancellation and Repeated guest

Data Visualization

10. Checking whether not getting allotted the same room type as demanded is the cause of cancellation of bookings?
11. Does not getting same room affects the adr.?
12. Which is the most common channel for booking hotels?
13. Which Distribution channel has highest no. of days in waiting list?
14. Which distribution channel has highest cancellation percentage?
15. Which Market segment has highest no. of days in waiting list?
16. Which is preferred stay length in each hotel?
17. Does lead time have effect on cancellation?

Data Visualization

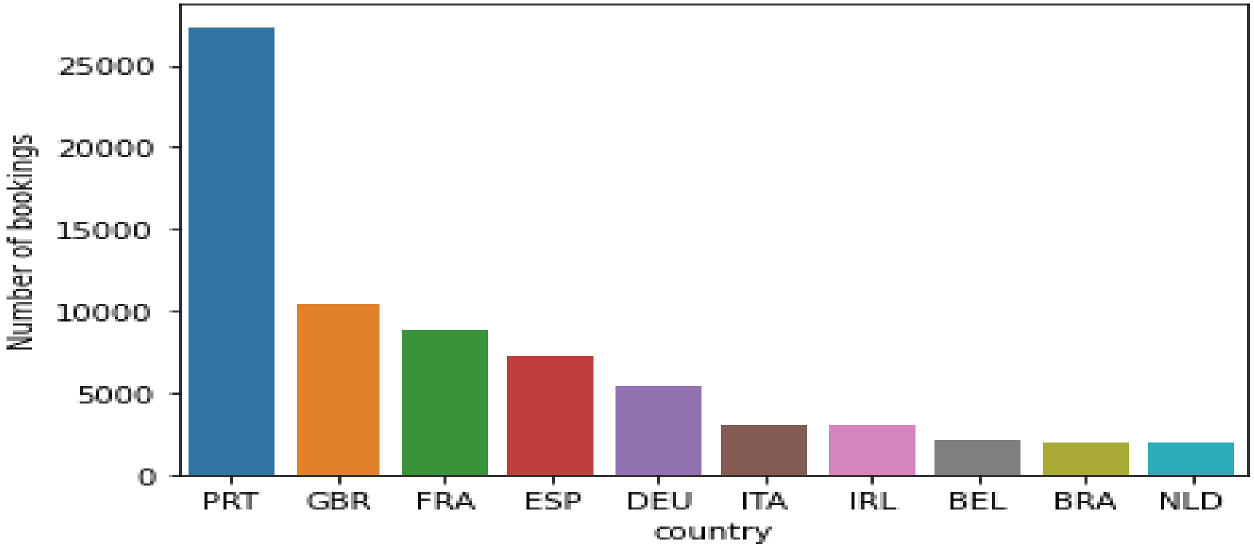
❑ **Percentage of bookings in each hotel type**



✔ Around 60% bookings are for City Hotel and 40% are for Resort Hotel.

Data Visualization

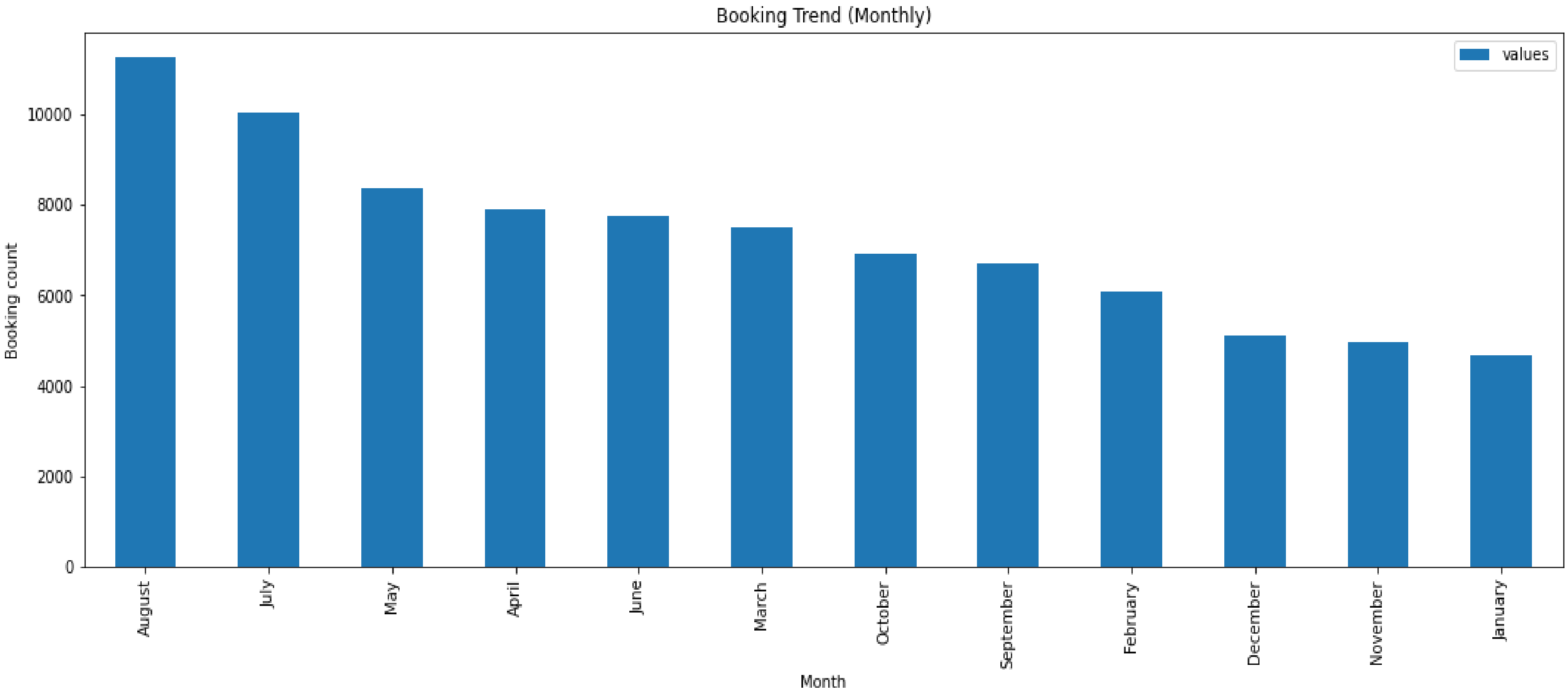
❑ **From which country most guests come?**



✔ **Portugal** is the country from where most guests come. Around 31.36 % of guests come from Portugal, followed by Great Britain and France.

Data Visualization

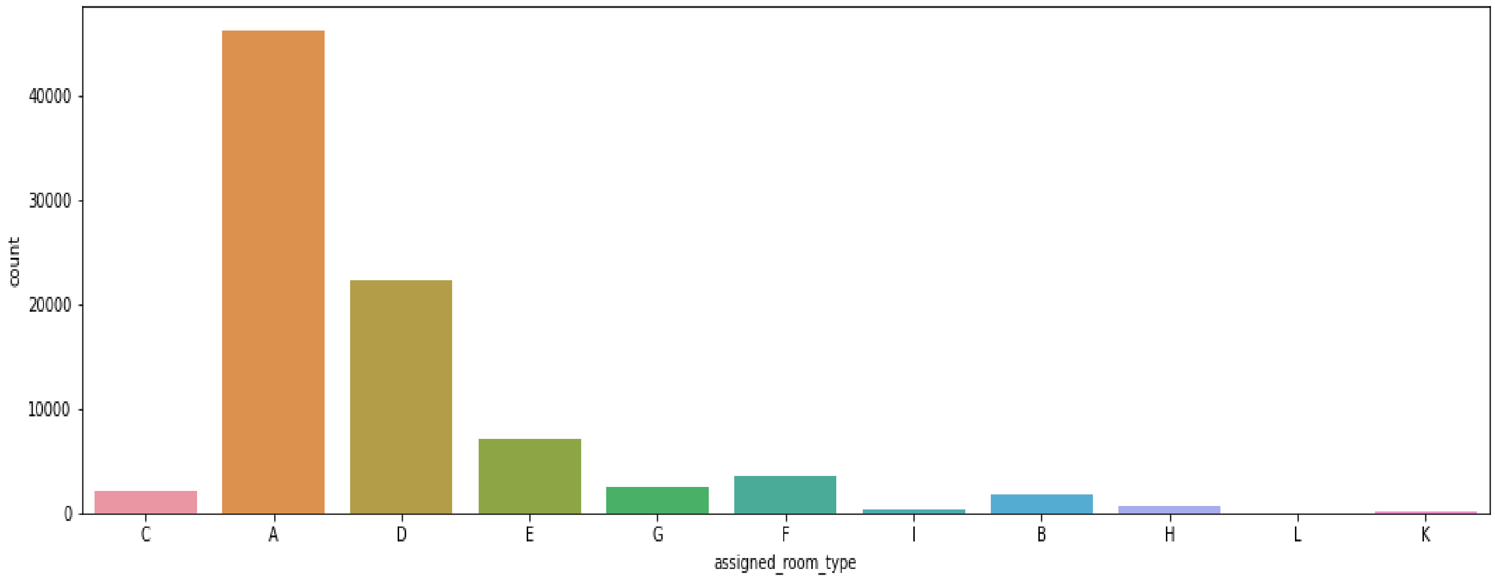
- ❑ **Which is the busiest month for hotels?**



✔ As we can see most bookings were made from **July to August**. And the least bookings were made at the start and end of the year.

Data Visualization

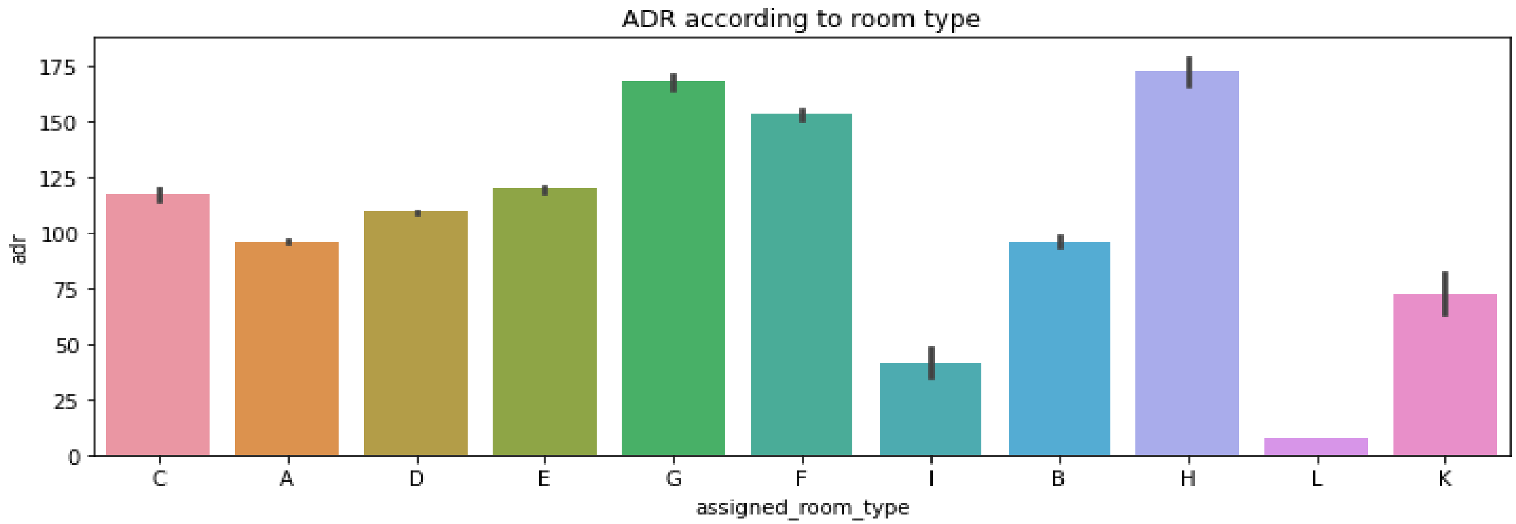
- ❑ **Which room type is in most demand ?**



✔ **Most demanded room** type is ’A’. Hotel should increase room type A to increase the revenue.

Data Visualization

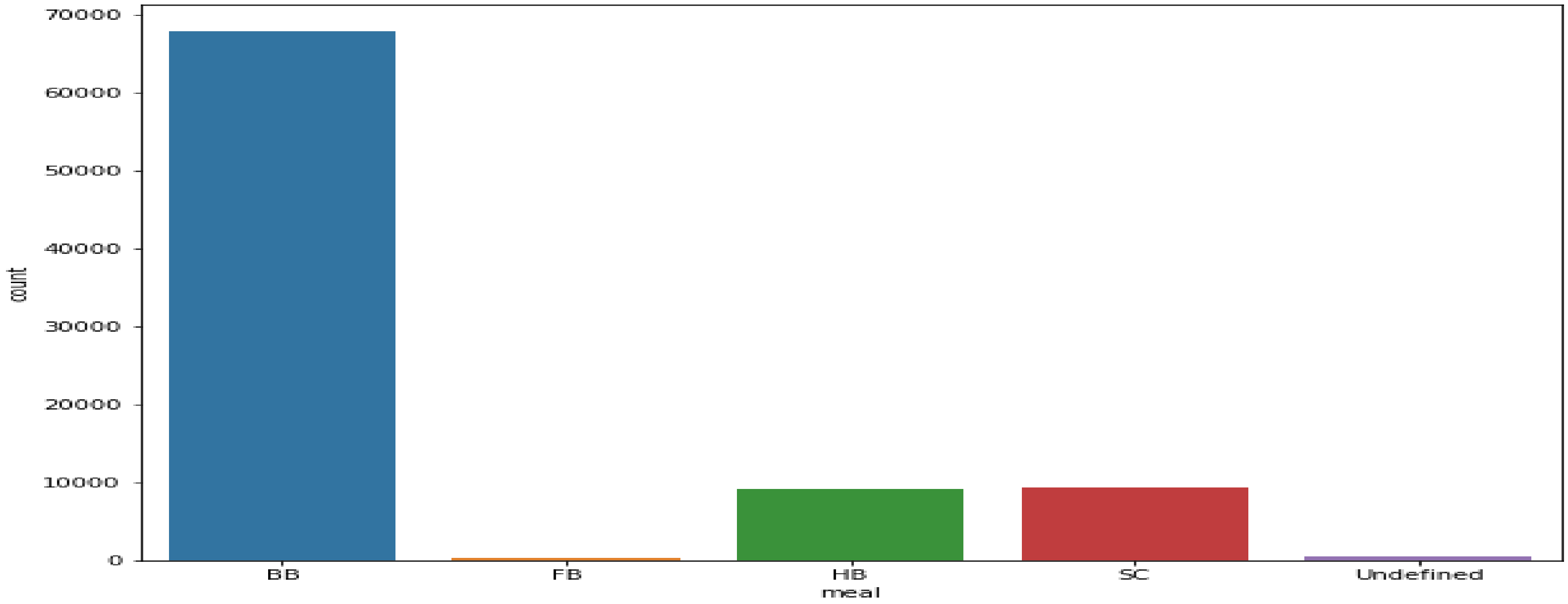
- ❑ **Which room type generate highest adr?**



✔ Although room type A was oh high demand and most booked, Highest revenue was produced by room type H followed by G. We can say that room type **H and G are premium rooms.**

Data Visualization

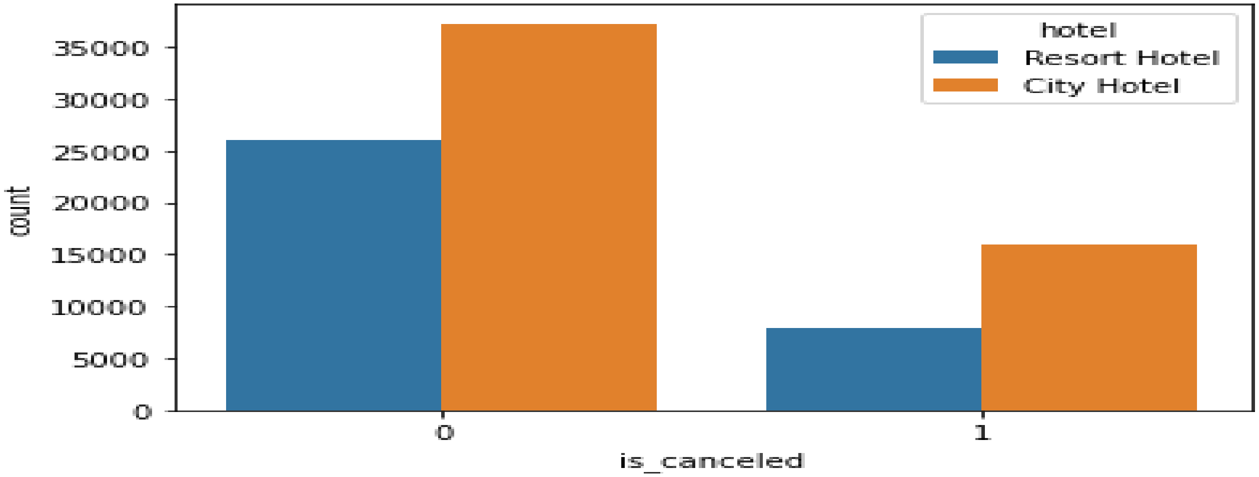
- ❑ **Which meal type is most preferred meal of customers?**



✔ Most preferred meal type is BB i.e., **Bed and breakfast**

Data Visualization

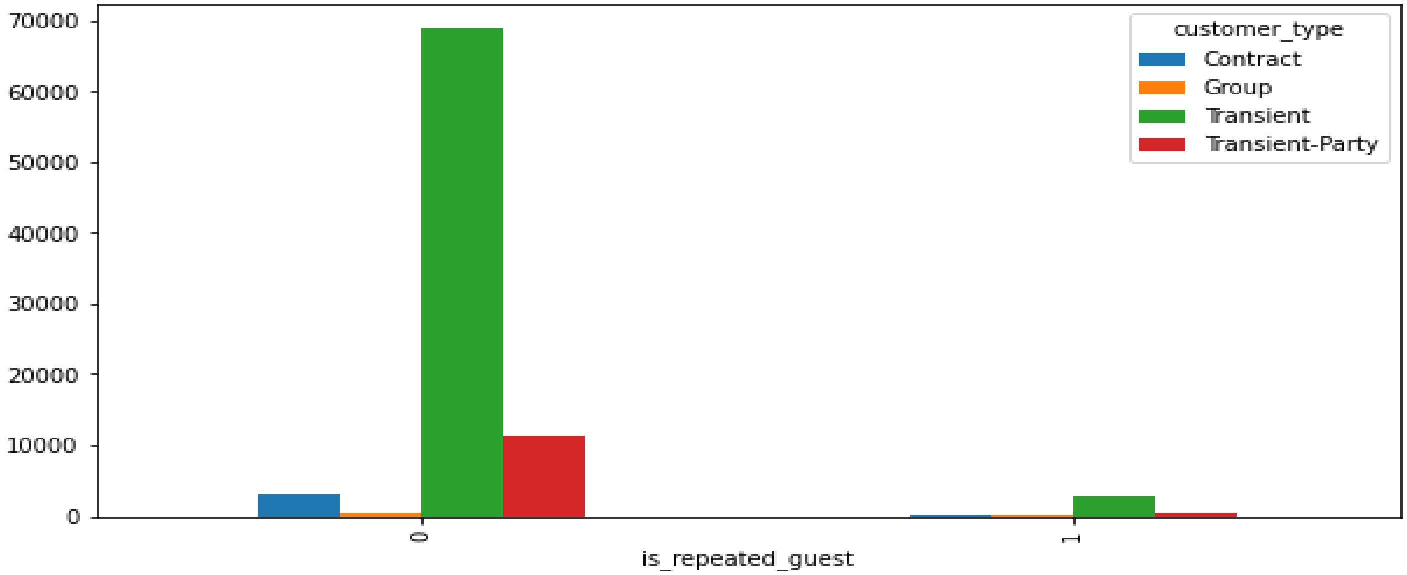
- ❑ **How many booking were cancelled?**



✔ Around 8000 **Resort Hotels** and 16000 **City Hotels** got cancelled.

Data Visualization

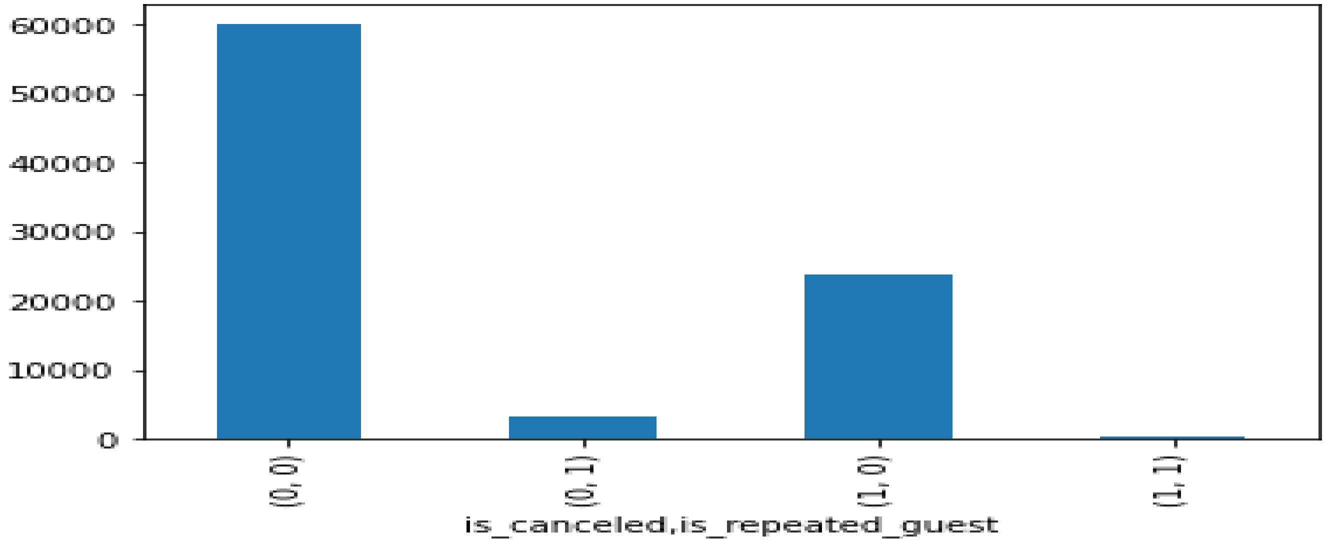
- ❑ **Which type of customers are most repeated?**



✔ Here, we can see that the maximum number of repeated guests are "**Transient type**" i.e., the "**Short-time customers**"

Data Visualization

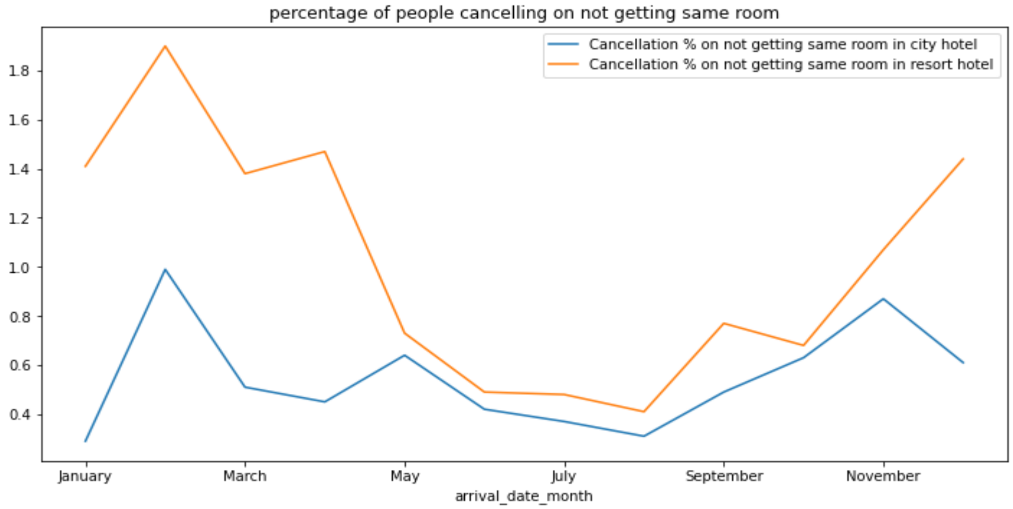
❑ **Booking cancellation and Repeated guest**

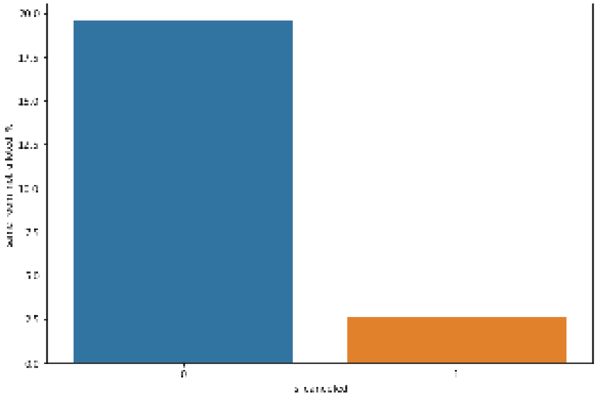


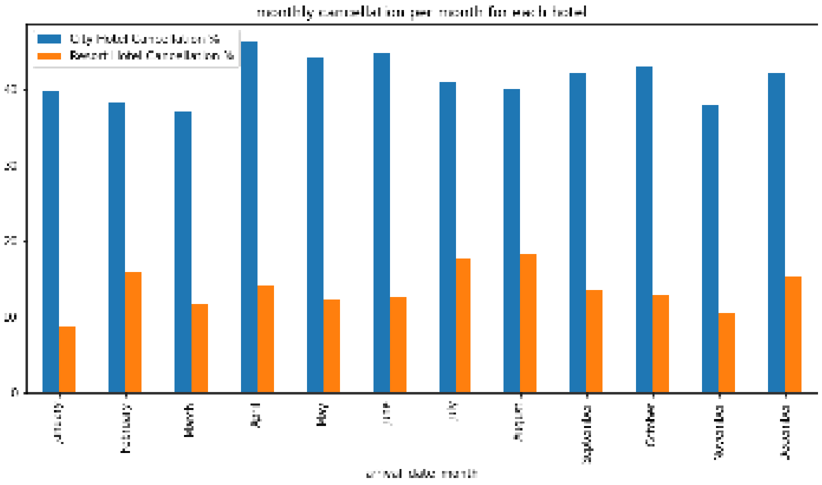
✔ We see that when a hotel booking is cancelled and the customer is a repeated guest, the entries are almost '0', which means that **repeated guest is very less likely to cancel his booking with the hotel**.

Data Visualization

- ❑ **Checking whether not getting allotted the same room type as demanded is the cause of cancellation of booking?**





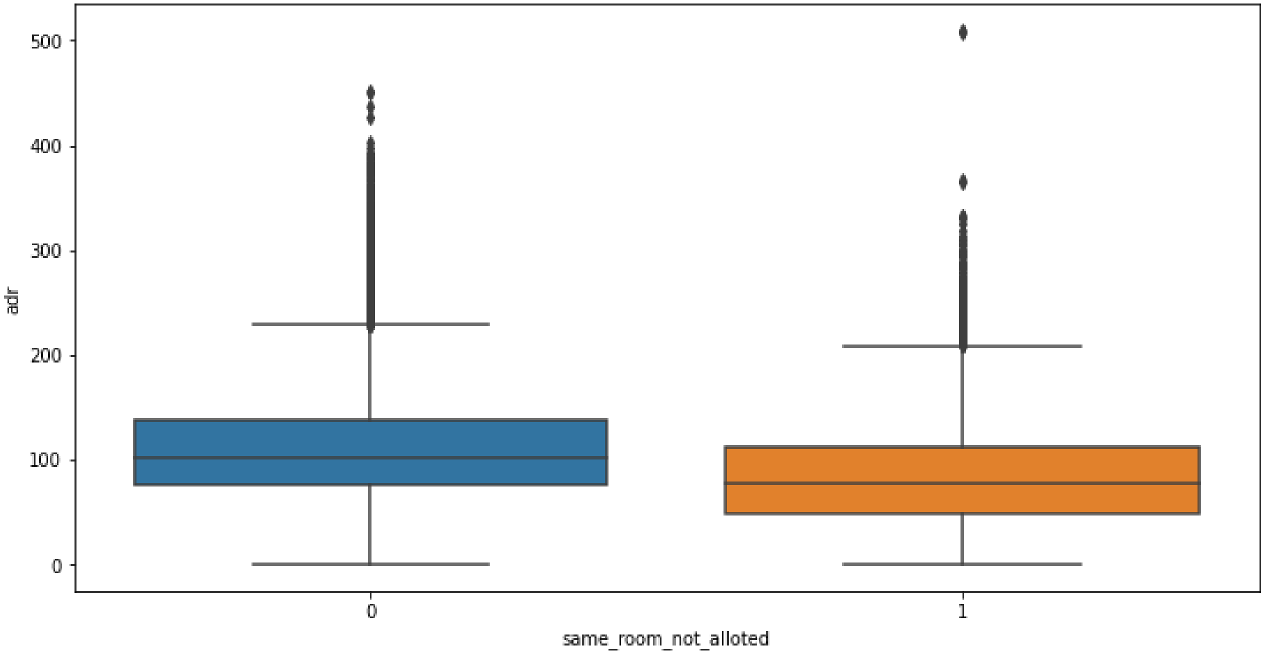


✔ Here we see that not getting same room as demanded is not a case of room cancellation. A significant percentage of bookings are not cancelled even after getting different room as demanded.

✔ we see that the cancellation are very less for resort and city hotels, which is quite significant. **Guests are willing to keep their reservation even if they do not get the chosen accommodation.**

Data Visualization

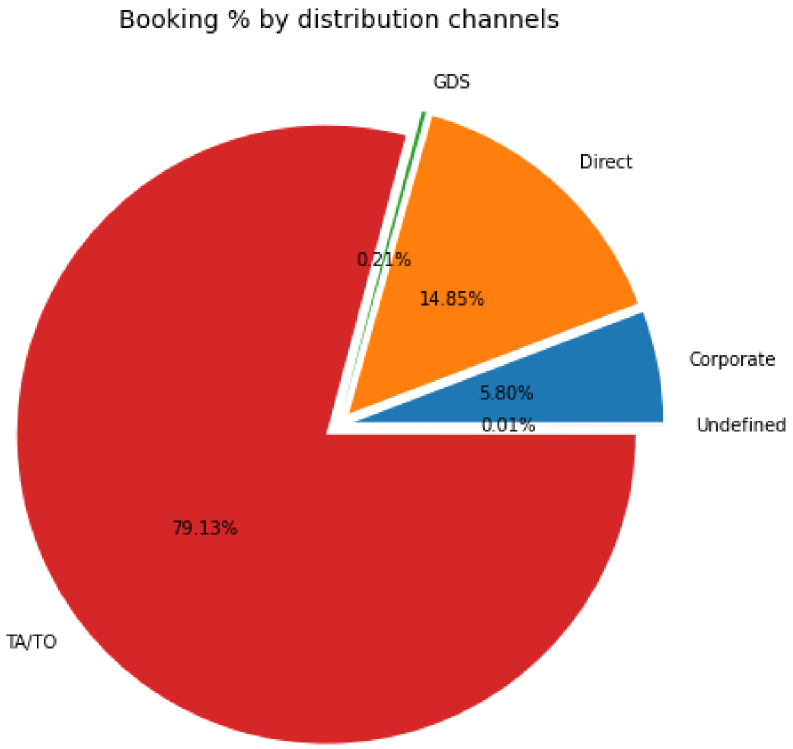
- Does not getting same room affects the adr?



✔ Not getting same room **do affects the adr**, if a guest does not get the desired room, he will pay a bit less.

Data Visualization

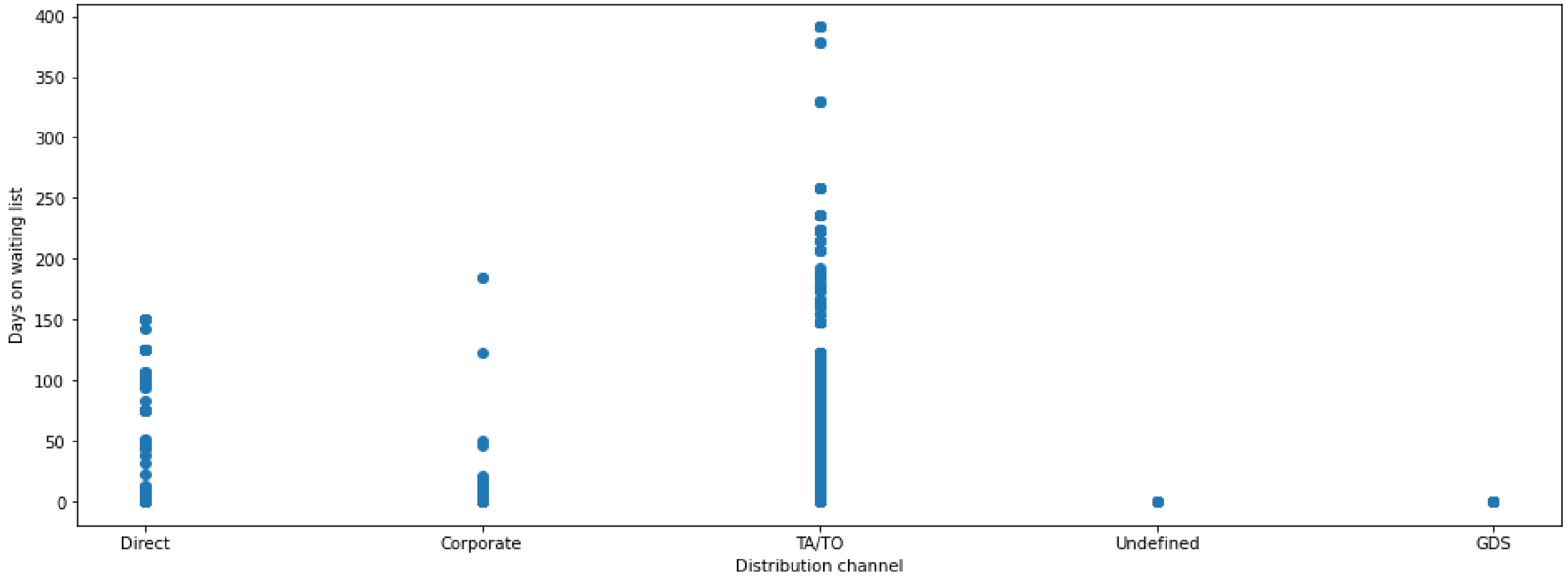
❑ **Which is the most common channel for booking hotels?**



✔ Guests use different channels for making bookings out of which most preferred way is **TA/TO** (i.e., Travel Agents and Tour Operators)

Data Visualization

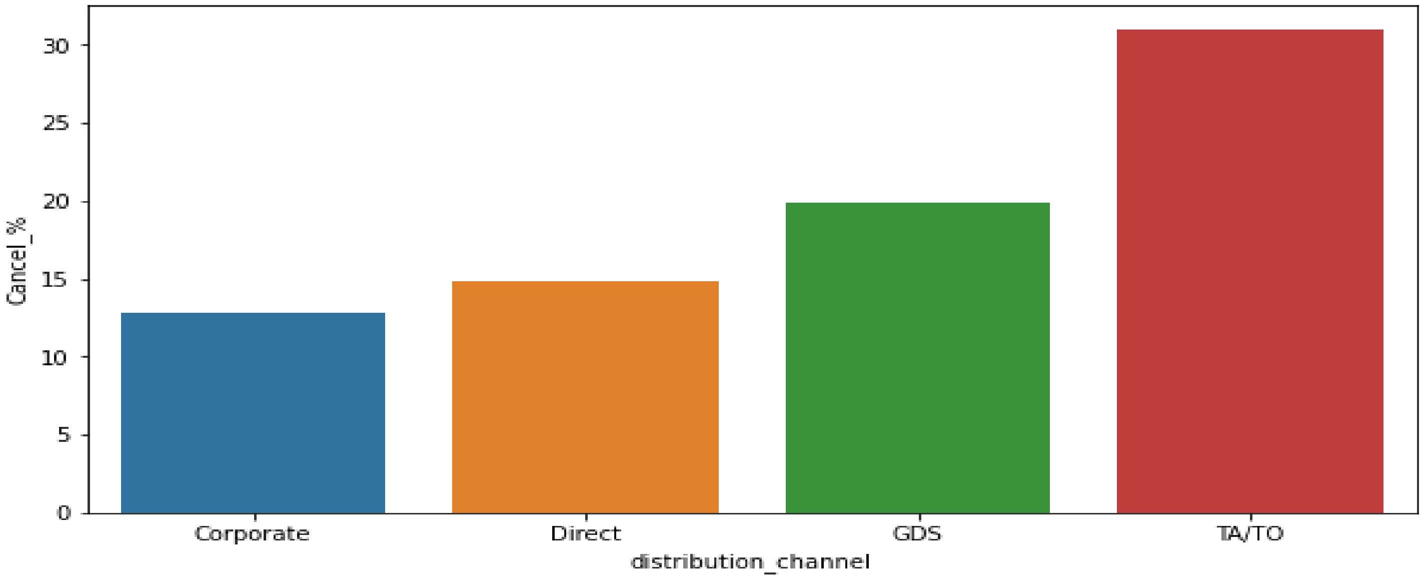
❑ **Which Distribution channel has highest no. of days in waiting list?**



✔ We see that the **'Travel Agent' and 'Tour Operators'** are the distribution channels for which the highest number of days are there on the waiting list.

Data Visualization

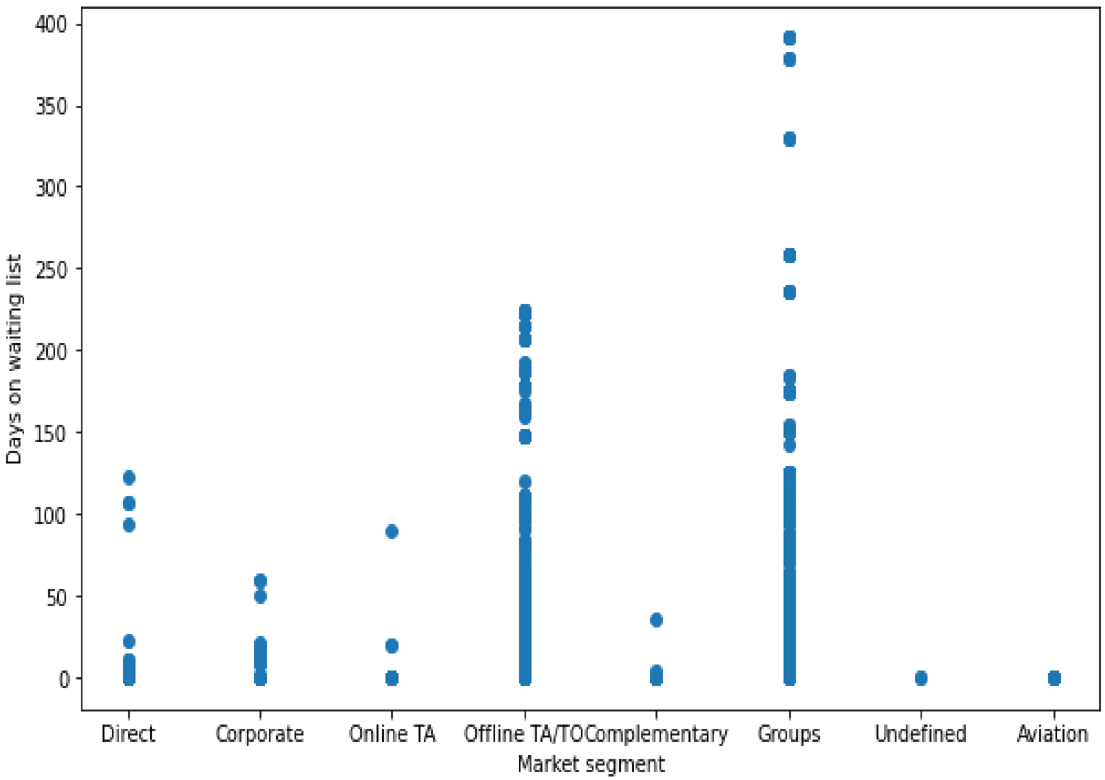
❑ **Which distribution channel has highest cancellation percentage?**



✔ TA/TO has highest booking cancellation %. As a result, **booking made through TA/TO are 30% more likely to be cancelled.**

Data Visualization

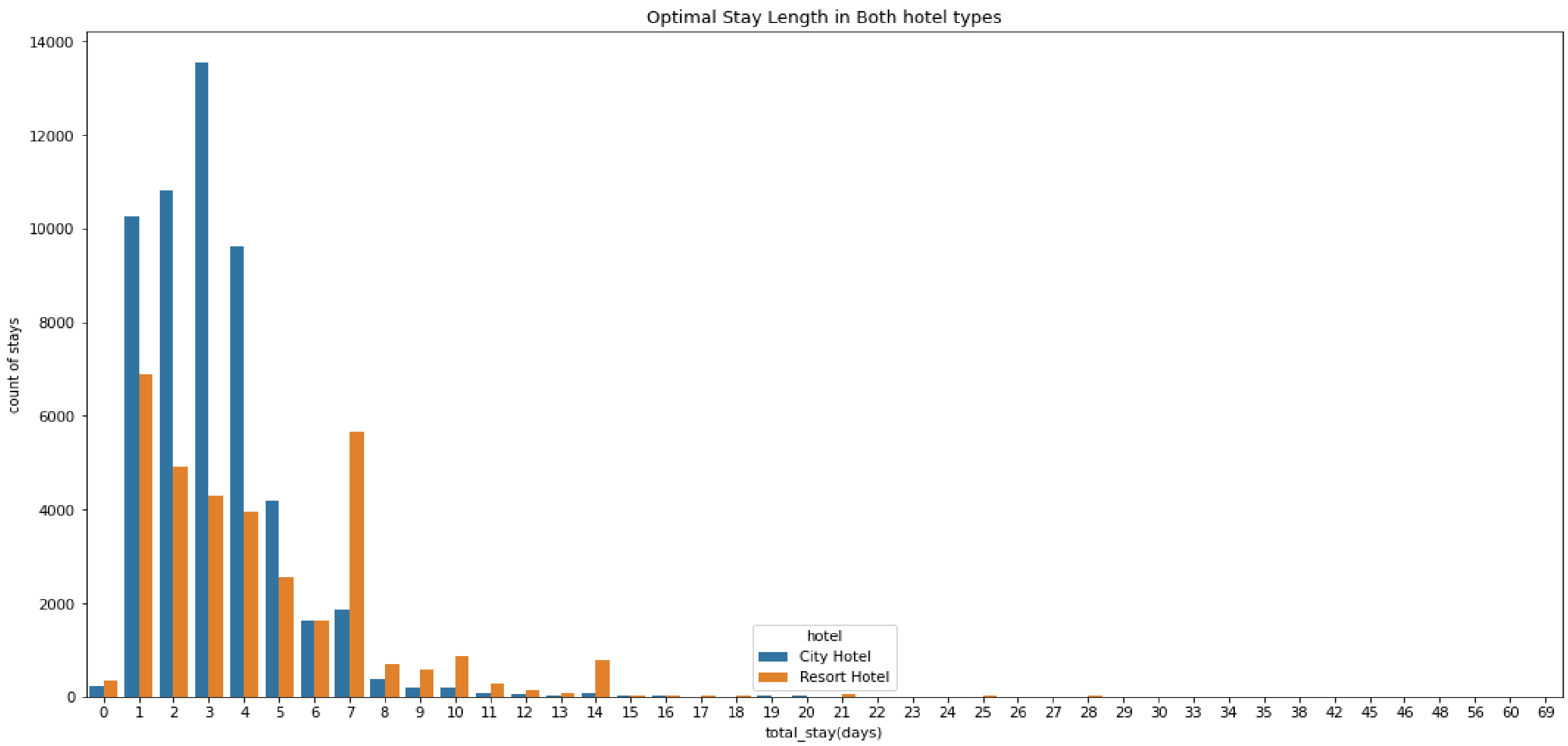
❑ **Which Market segment has least no. of days in waiting list?**



✔ Here, we see that **Aviation industry has the minimum number of days on the waiting list.** The reason for this could be that when a flight has to land at the location, it has to provide immediate accommodation to all of its working staff such as pilots and air hostages therefore they do not entertain hotels that have a long waiting list. So, in general, the hotel management sees to it that their needs are satisfied immediately and that they have almost no days on the waiting list.

Data Visualization

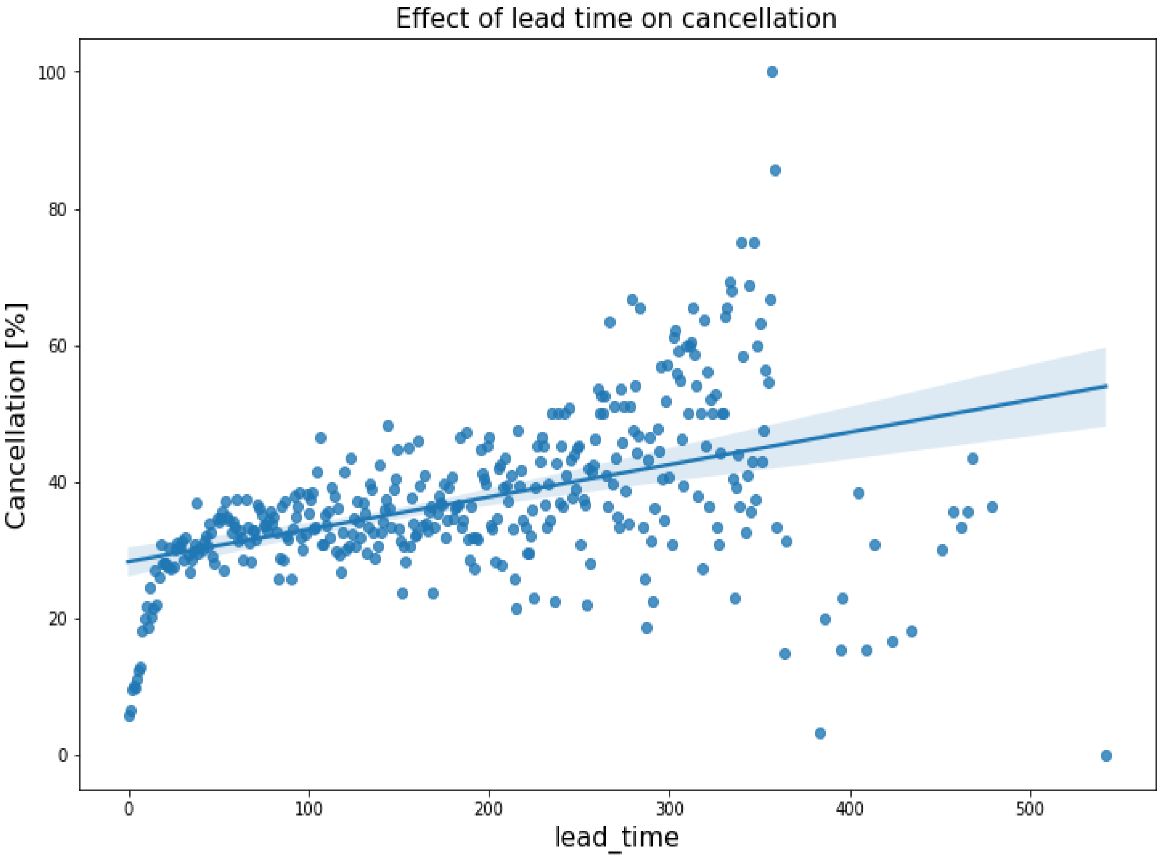
❑ Which is preferred stay length in each hotel?



✔ Most people prefer to stay in hotels **for at least 5 days in both types of hotel**. However, people tend to stay longer at resort hotels.

Data Visualization

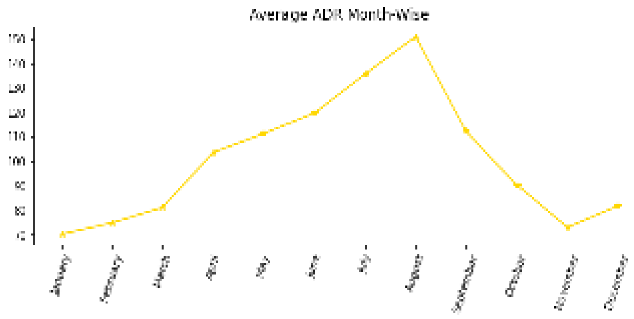
❑ **Does lead time have effect on cancellation?**

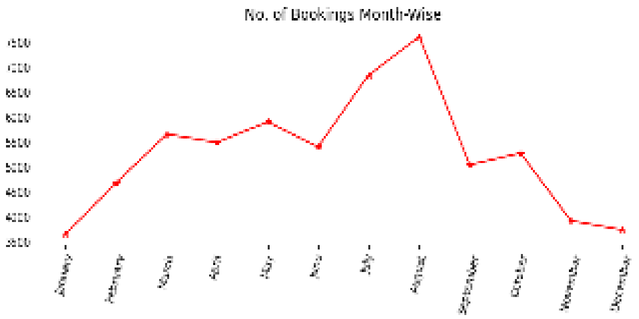


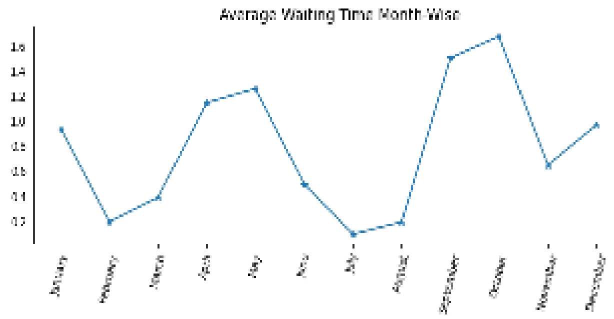
✔ We see that the Effect of Lead time on Cancelation is **POSITIVE** which means that as the lead time increases, so does the number of cancelation.

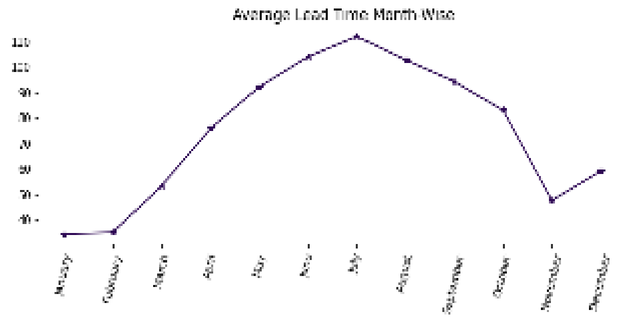
Data Visualization

❑ Month-wise Analysis









When compared to previous months, we can see that the price and number of bookings are reduced from October to February. During this time, the average lead time is also very low.

The only concern is the waiting time, which is slightly longer from September to October, whereas the waiting time is the lowest and the price is the highest in July and August.

So the best time to book a hotel would be between October to February to get cheaper price and some privacy.

Inferences

1. The most popular hotel type is **‘City hotel’** which is preferred by **61.07% of total visitors.**
2. **96.15% of guests were not repeated guests**.
3. **27.08%** customers have **cancelled their booking**.
4. Market was widely captured by distribution channels such as TA or TO, especially **“Online TA” which captured 59% percent of the market.**
5. Around **53%** customers were assigned A room type and **65%** has reserved it. This also indicates the fact that **room type A is in higher demand.**
6. Almost **82.4% of customers are transient,** with very few booking as a group.

Conclusion

1. Higher lead time has higher chance of cancellation.
2. July-August are the busiest and most profitable months for both the hotels.
3. The best time to book a hotel would be between October to February to get for cheaper price.
4. Most people prefer to stay in the hotels for at least five days in both types of hotel. However, people tend to stay longer at resort hotels.
5. Maximum number of repeated guests are "Transient type" i.e., the "Short-time customers"
6. City hotels receive around 60% of bookings, while Resort hotels receive 40%, hence city hotels are busier than Resort hotels.
7. City hotel’s total adr is slightly higher than Resort hotel’s.
8. Most of the guests were from European countries, with the most coming from Portugal followed by Great Britain and France.
9. Guests use different channels for making bookings out of which most preferred way is TA/TO.
10. Almost 30% of bookings via TA/TO are cancelled.
11. Bookings are not affected if they do not receive the same room as reserved. Although different room allotment does lessen the adr.

THANK YOU