Capstone Project

Hotel Booking Analysis

Team Members:

Vidit Ghelani

Mahesh Lakum

What is EDA?

- It is the abbreviation for Exploratory Data Analysis.
- Input: Raw dataset
- Output: Some useful conclusion

Processing Method:

- This is user defined.
- Have a look at the dataset and formulate a set of questions. These questions are representative viewpoints to a dataset. The output of entire analysis depends upon these viewpoints.
- Hence, choose wisely.

P.S.: These viewpoints are called KPIs. (Key Performance Indexes)

How to Approach the Problem:

Approach the problem in three simple steps:

1. Pre- Processing
2. Performing exploratory data analysis (EDA)
3. Answer the questions based on analysis and draw out the conclusions

Pre-Processing

In just few simple steps:

1. View the data
2. Inspecting the data
3. Cleaning the data
4. Formulate the Questions

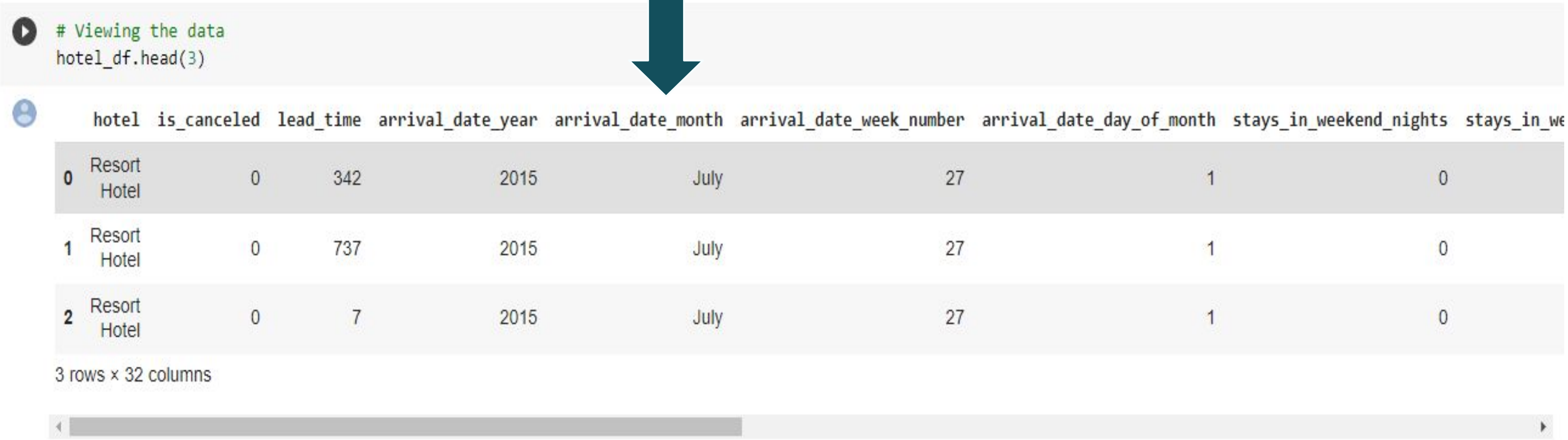
View the data

Quick look:

- **Size of data**

  (119390, 32) => 119390 rows and 32 features

- **Viewing first 3 rows**



```
# Viewing the data
hotel_df.head(3)
```

	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_we
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	

3 rows × 32 columns

View the data (Cont.)

Features:

1. hotel : Talks about the type of Hotel in the data:
   Resort hotel and City hotel
2. is_canceled: Talks about the Cancellation Status of Booking
   1 mean Canceled and 0 means Not Canceled
3. lead_time: This shows the difference of booking date and arrival date.
4. arrival_date_year : This gives the year in which the visitor arrived
   2015, 2016, 2017
5. arrival_date_month: This gives the month in which the visitor arrived
   January to December
6. arrival_date_week_number: This gives the week number of year in which the visitor arrived
   1 to 53
7. arrival_date_day_of_month: This gives the day number of month when the visitor arrived
   1 to 31
8. stays_in_weekend_nights: This gives the number of weekend nights, i.e. Saturday and Sunday
9. stays_in_week_nights: This gives the number of week nights, i.e. Monday to Friday
10. adults: This gives the number of adults per booking
11. children: This gives the number of children per booking
12. babies: This gives the number of babies per booking
13. meal: This gives the type of meal preferred.
    Undefined/SC means no meal package, BB means Bed & Breakfast, HB means Half board (i.e., breakfast & one other meal – usually dinner), FB means Full board (i.e., breakfast, lunch & dinner)
14. country: This gives the country of origin of visitor
15. market_segment: This gives the group of people based on market
    Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Aviation Where, TA: Travel Agents, TO: Tour Operators
16. distribution_channel: This mentions the type of distribution channel
    Direct, Corporate, TA/TO, Undefined, GDS

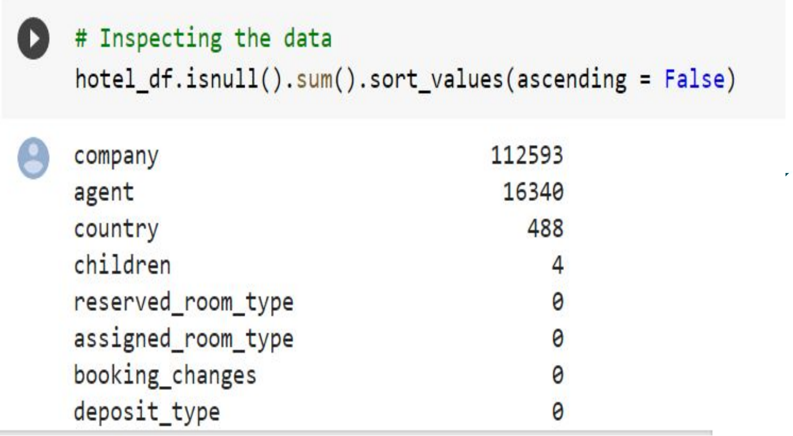
View the data (Cont.)

Features (cont.):

17. is_repeated_guest: This shows repeated customers
    1 means repeated customer, 0 means not repeated
18. previous_cancellations: This gives the number of previous bookings that were canceled by the customer prior to the current booking
19. previous_bookings_not_canceled: This gives the number of previous bookings not canceled by the customer prior to the current booking
20. reserved_room_type: This gives the type of room reserved
    'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'
21. assigned_room_type: This gives the type of room whose possession is given at the time of arrival.
    'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'
22. booking_changes: This gives the number of bookings changed
23. deposit_type: This gives the types of deposit
    No Deposit, Non Refund, Refundable
24. agent: Agent Id
25. company: Company Id
26. day_in_waiting_list: Number of days the booking was in the waiting list before confirmation
27. customer_type: Type of customer
    Contract, Group, Transient, Transient-party
28. adr: means average daily rate
29. required_car_parking_spaces: Number of car parking spaces required by the customer
30. total_of_special_requests: Number of special requests made by the customer
31. reservation_status: Status of reservation
    Canceled, Check-Out, No-Show
32. reservation_status_date: Date at which the last status was updated

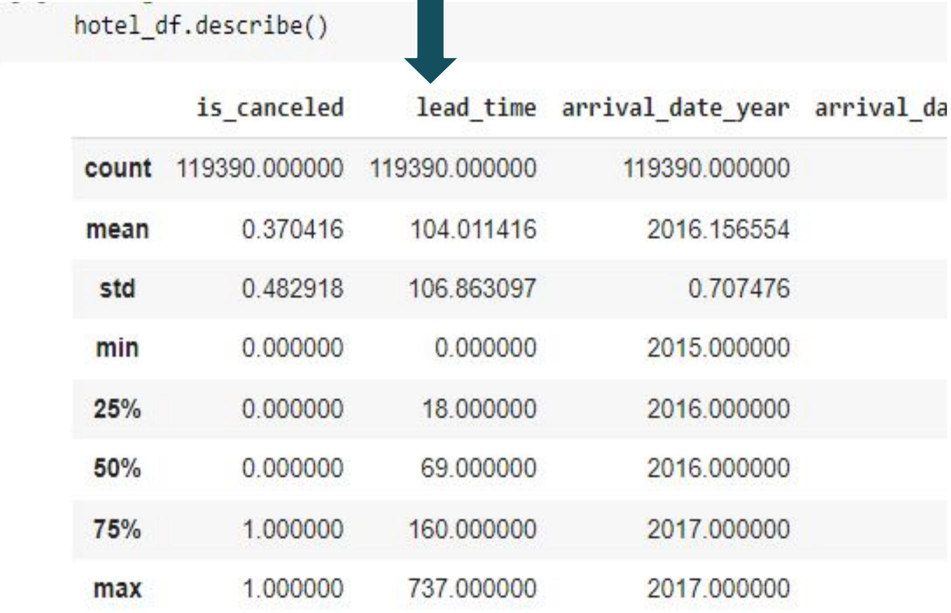
Inspecting the data

- **Inspecting the data for null values.**
- **Get the basics statistics for each feature.**



```
# Inspecting the data
hotel_df.isnull().sum().sort_values(ascending = False)
```

```
company                 112593
agent                    16340
country                    488
children                     4
reserved_room_type           0
assigned_room_type           0
booking_changes              0
deposit_type                 0
```



```
hotel_df.describe()
```

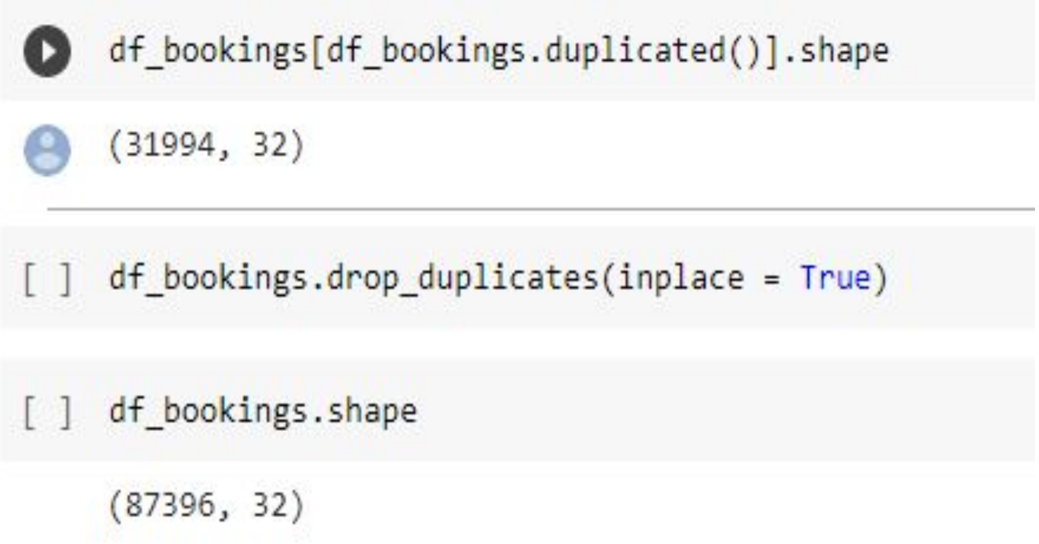
	is_canceled	lead_time	arrival_date_year	arrival_da
count	119390.000000	119390.000000	119390.000000	
mean	0.370416	104.011416	2016.156554	
std	0.482918	106.863097	0.707476	
min	0.000000	0.000000	2015.000000	
25%	0.000000	18.000000	2016.000000	
50%	0.000000	69.000000	2016.000000	
75%	1.000000	160.000000	2017.000000	
max	1.000000	737.000000	2017.000000	

Cleaning the data

Duplicate Entries in the data:

31994

Hence, we drop them



```
df_bookings[df_bookings.duplicated()].shape
```

```
(31994, 32)
```

```
[ ]  df_bookings.drop_duplicates(inplace = True)
```

```
[ ]  df_bookings.shape
```

```
(87396, 32)
```

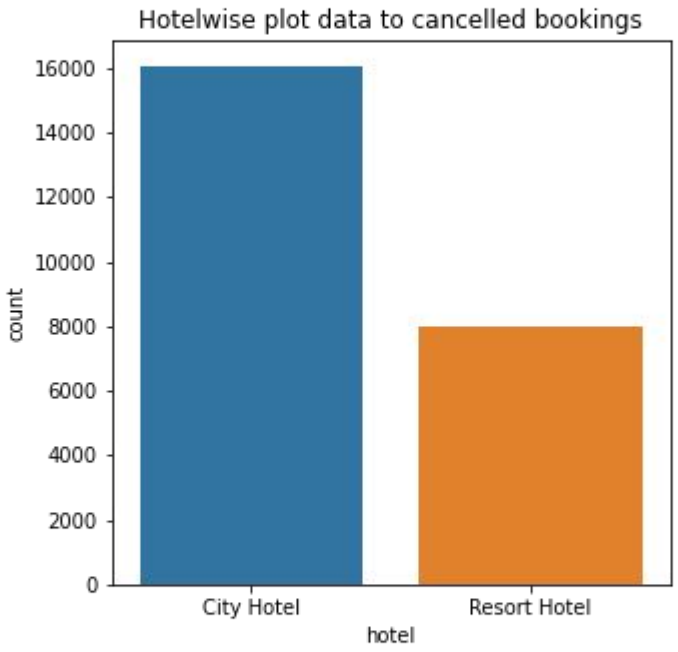
Cleaning the data (Cont.)

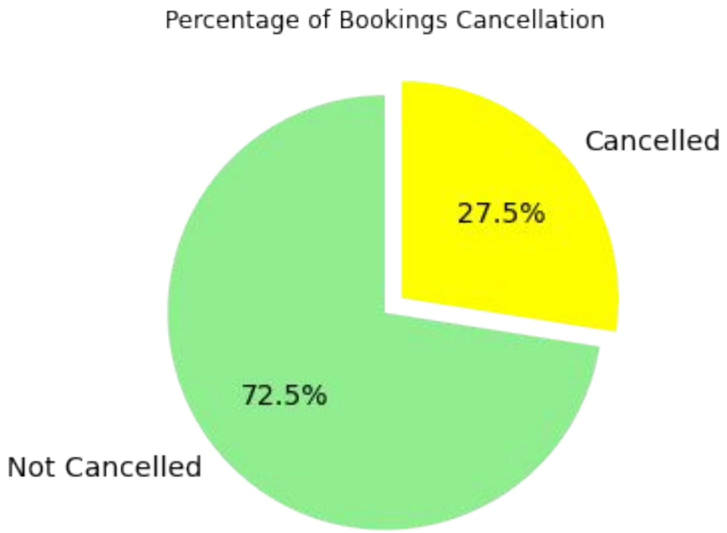
Dealing with Null Values

Features	Observation	Action (Replace nan with)
'agent'	Null value means those customers as direct to hotel, hence we need not omit them from the count .	0
'company'	Null value means those bookings are possibly not for business tours	0
'country'	Every customer must belong to a unique country. Hence this field cannot be empty. Here, we replaced it with mode because it means we will take the value with maximum occurrence in that column.	Mode of that feature (Here, PRT)
'children'	Null values means zero children.	0

Cleaning the data (Cont.)

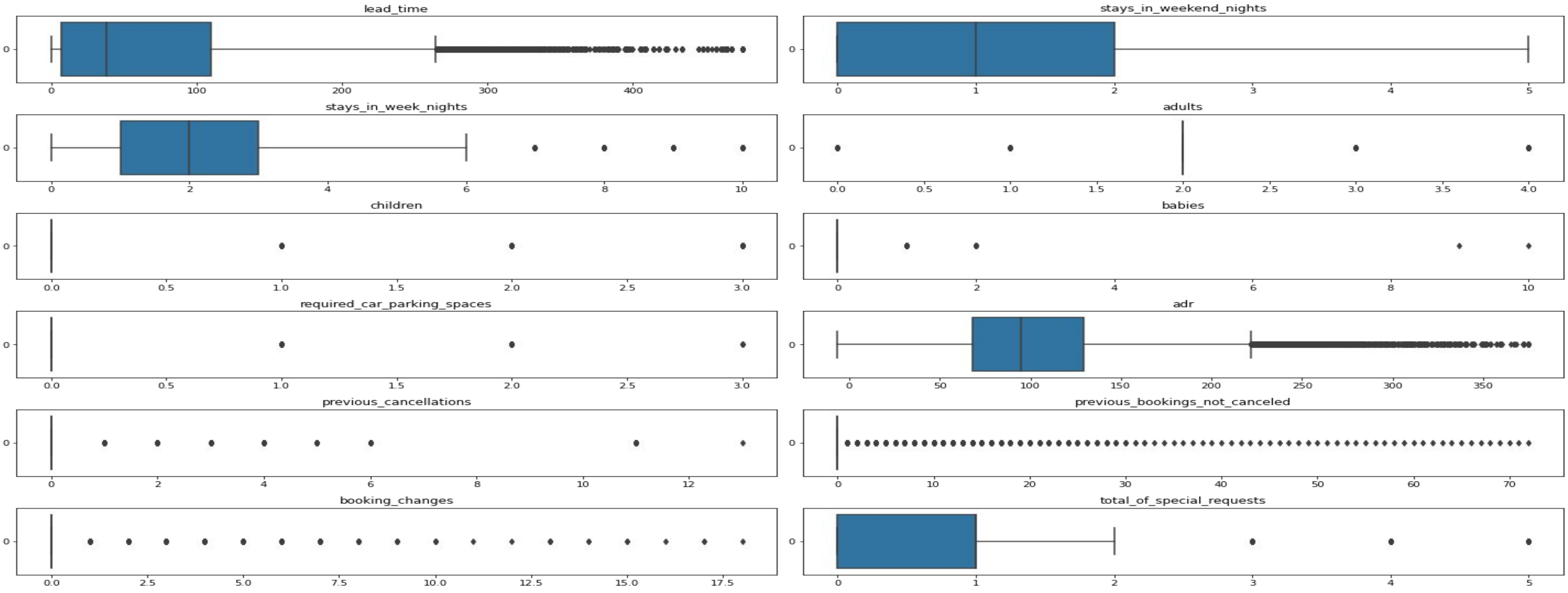
Removing the data of canceled booking





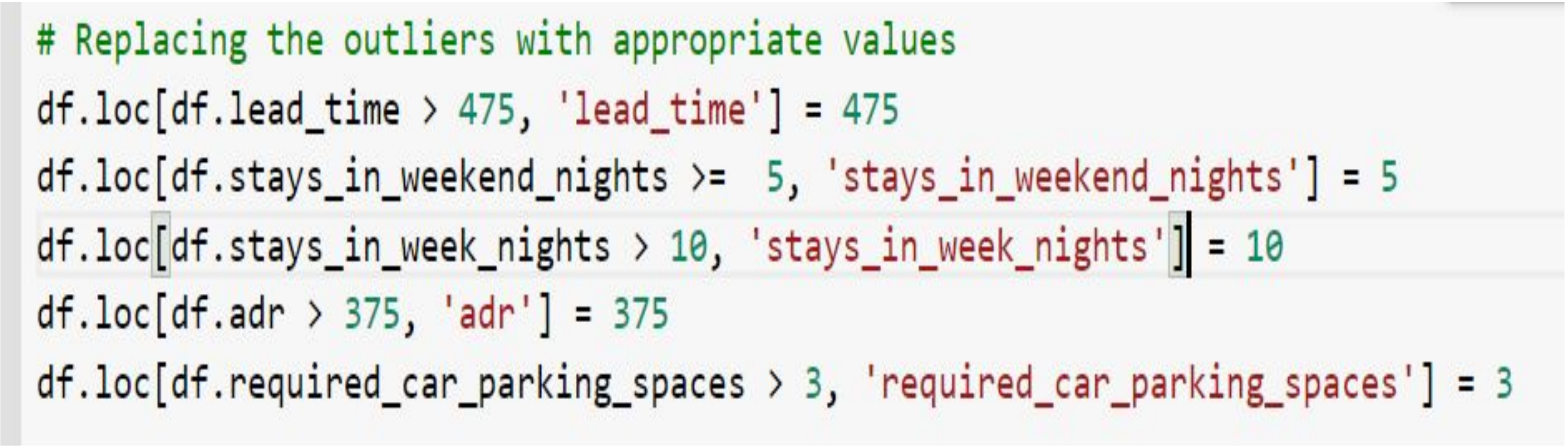
Cleaning the data (Cont.)

Dealing with Outliers



Cleaning the data (Cont.)

Dealing with Outliers

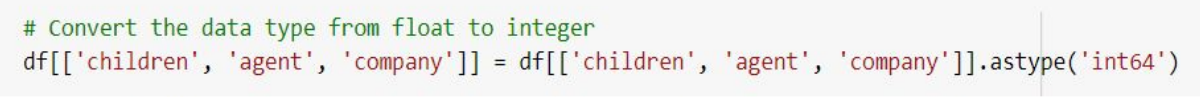


```
# Replacing the outliers with appropriate values
df.loc[df.lead_time > 475, 'lead_time'] = 475
df.loc[df.stays_in_weekend_nights >=  5, 'stays_in_weekend_nights'] = 5
df.loc[df.stays_in_week_nights > 10, 'stays_in_week_nights'] = 10
df.loc[df.adr > 375, 'adr'] = 375
df.loc[df.required_car_parking_spaces > 3, 'required_car_parking_spaces'] = 3
```

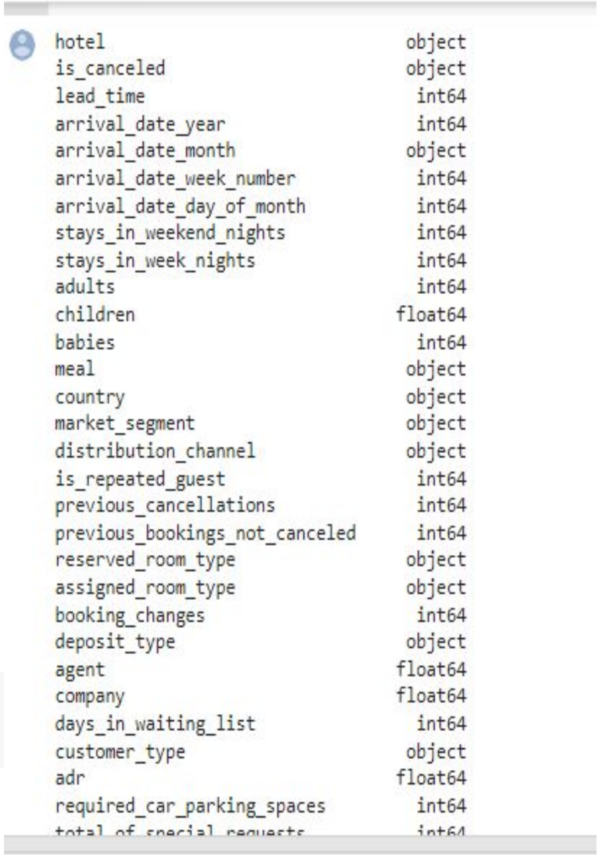
Cleaning the data (Cont.)

Changing the data types

to integer because these cannot be floating points number… ;)



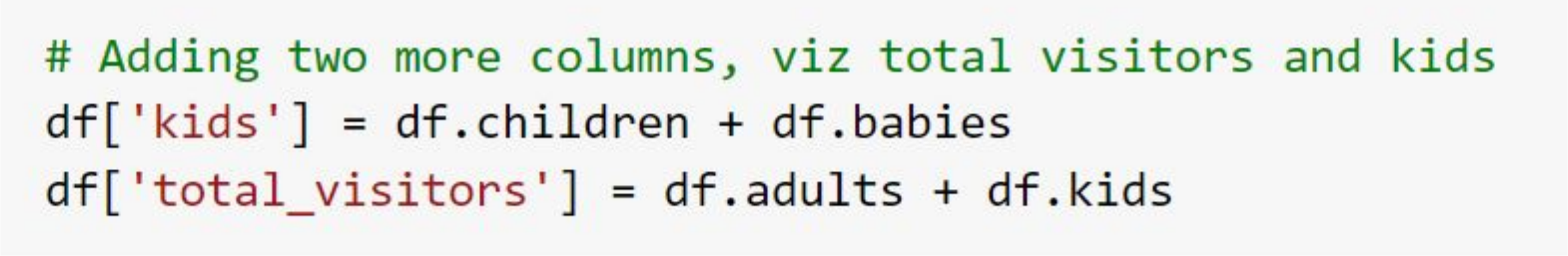
```
# Convert the data type from float to integer
df[['children', 'agent', 'company']] = df[['children', 'agent', 'company']].astype('int64')
```



```
hotel                              object
is_canceled                        object
lead_time                           int64
arrival_date_year                   int64
arrival_date_month                 object
arrival_date_week_number            int64
arrival_date_day_of_month           int64
stays_in_weekend_nights             int64
stays_in_week_nights                int64
adults                              int64
children                          float64
babies                              int64
meal                               object
country                            object
market_segment                     object
distribution_channel               object
is_repeated_guest                   int64
previous_cancellations              int64
previous_bookings_not_canceled      int64
reserved_room_type                 object
assigned_room_type                 object
booking_changes                     int64
deposit_type                       object
agent                             float64
company                           float64
days_in_waiting_list                int64
customer_type                      object
adr                               float64
required_car_parking_spaces         int64
total_of_special_requests           int64
```

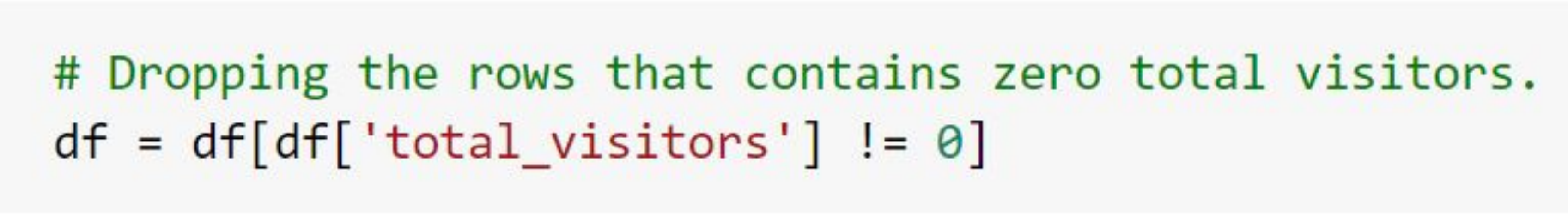
Cleaning the data (Cont.)

Derive new features



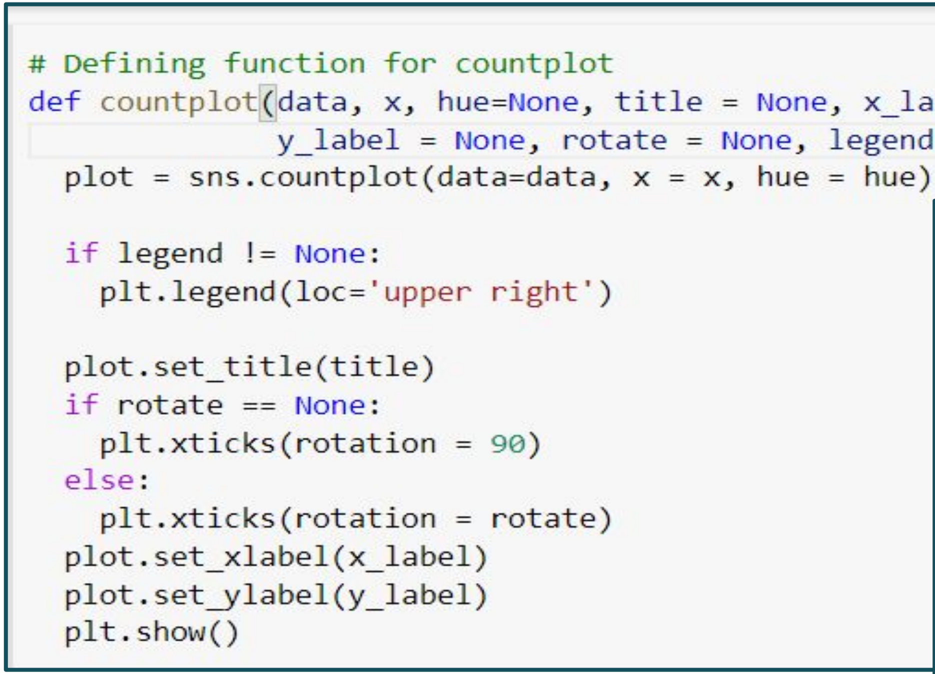
```
# Adding two more columns, viz total visitors and kids
df['kids'] = df.children + df.babies
df['total_visitors'] = df.adults + df.kids
```

Drop rows with zero total visitors



```
# Dropping the rows that contains zero total visitors.
df = df[df['total_visitors'] != 0]
```

Creating Functions



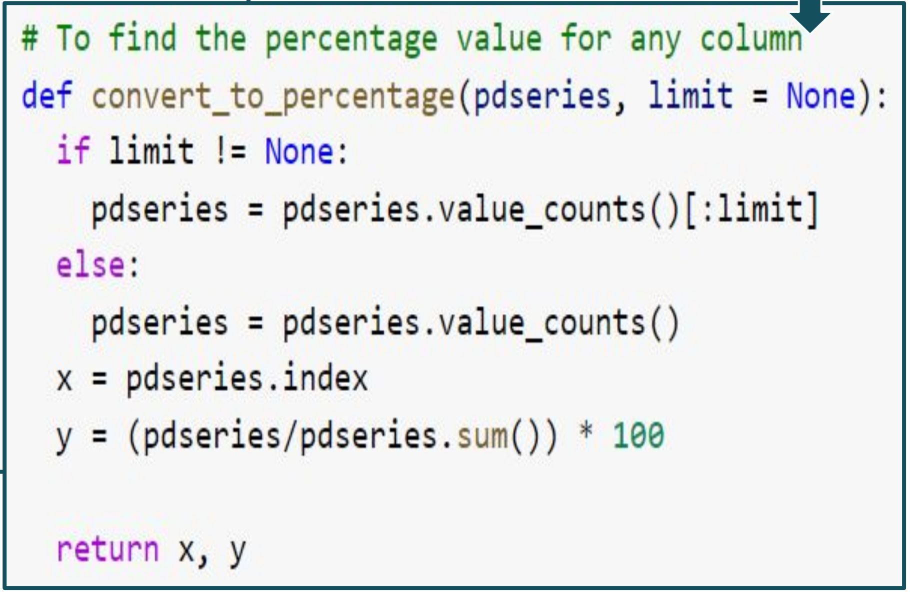
```python
# Defining function for countplot
def countplot(data, x, hue=None, title = None, x_label = None,
              y_label = None, rotate = None, legend = None):
  plot = sns.countplot(data=data, x = x, hue = hue)

  if legend != None:
    plt.legend(loc='upper right')

  plot.set_title(title)
  if rotate == None:
    plt.xticks(rotation = 90)
  else:
    plt.xticks(rotation = rotate)
  plot.set_xlabel(x_label)
  plot.set_ylabel(y_label)
  plt.show()
```

For ease to plot countplot

To get percentage of values in any column



```python
# To find the percentage value for any column
def convert_to_percentage(pdseries, limit = None):
  if limit != None:
    pdseries = pdseries.value_counts()[:limit]
  else:
    pdseries = pdseries.value_counts()
  x = pdseries.index
  y = (pdseries/pdseries.sum()) * 100

  return x, y
```

Lost the track?

Pre-Processing

In just few simple steps:

1. View the data
2. Inspecting the data
3. Cleaning the data
4. Formulate the Questions

- Duplicate entries
- Null values
- Remove irrelevant data
  (i.e. cancelled bookings)
- Outliers
- Change the data type
- Derive new features
- Drop rows with zero total visitors

The shape of final data frame with clean data is: **(63221,34)**

Formulate the Questions

1. What are the types of Hotels in the data?
2. What is the percentage of booking for each hotel?
3. What is the year wise trend of bookings for each hotel?
4. Which agent made the most number of bookings?
5. Enlist the country of origin of the majority of visitors.
6. What is the busiest time for hotels?
7. What is the proportion of weekend and weekday nights? Is there any difference between them?
8. How many bookings were previously canceled?
9. Which market segment does most visitors come from?
10. Which distribution channel does most visitors come from?
11. How many visitors are repeating?
12. Which is the most preferred meal?
13. Which is the most preferred deposit type?
14. How many visitors asked for car parking space?
15. Which month has the highest average daily rate per person?
16. What is the trend of ADR?
17. Which room Type is high in demand?
18. How likely is the hotel to receive a disproportionately high number of special requests?
19. Which hotel type has a longer waiting time for booking?
20. Which hotel type has a higher lead time for booking?

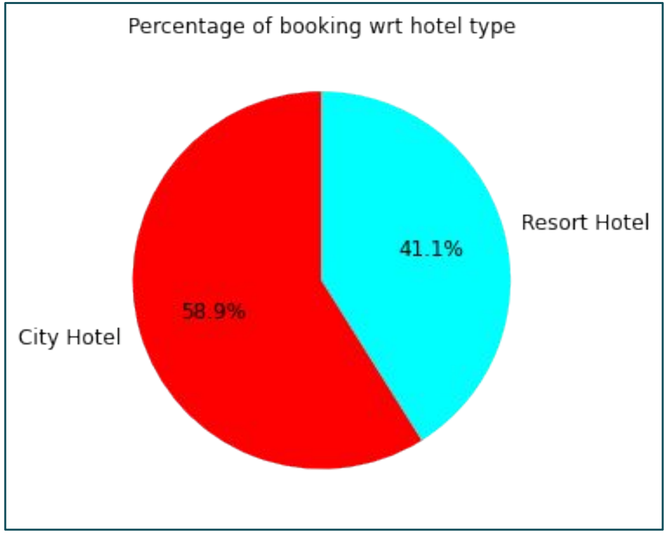
Performing the EDA

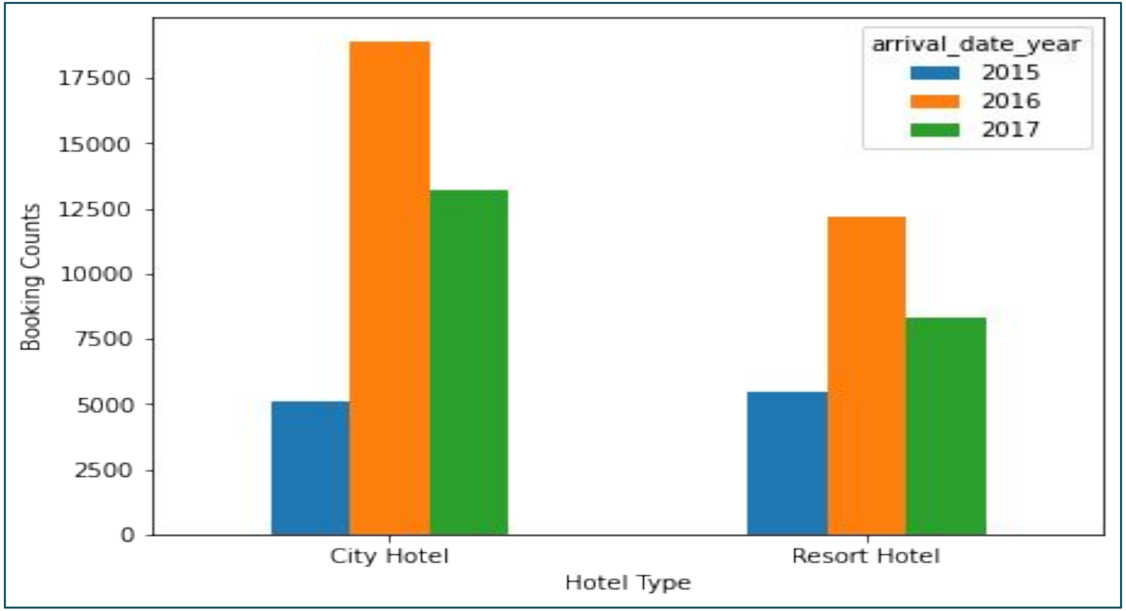
Performing the EDA

Q1. What are the types of Hotels in the data?

Q2. What is the percentage of booking for each hotel?

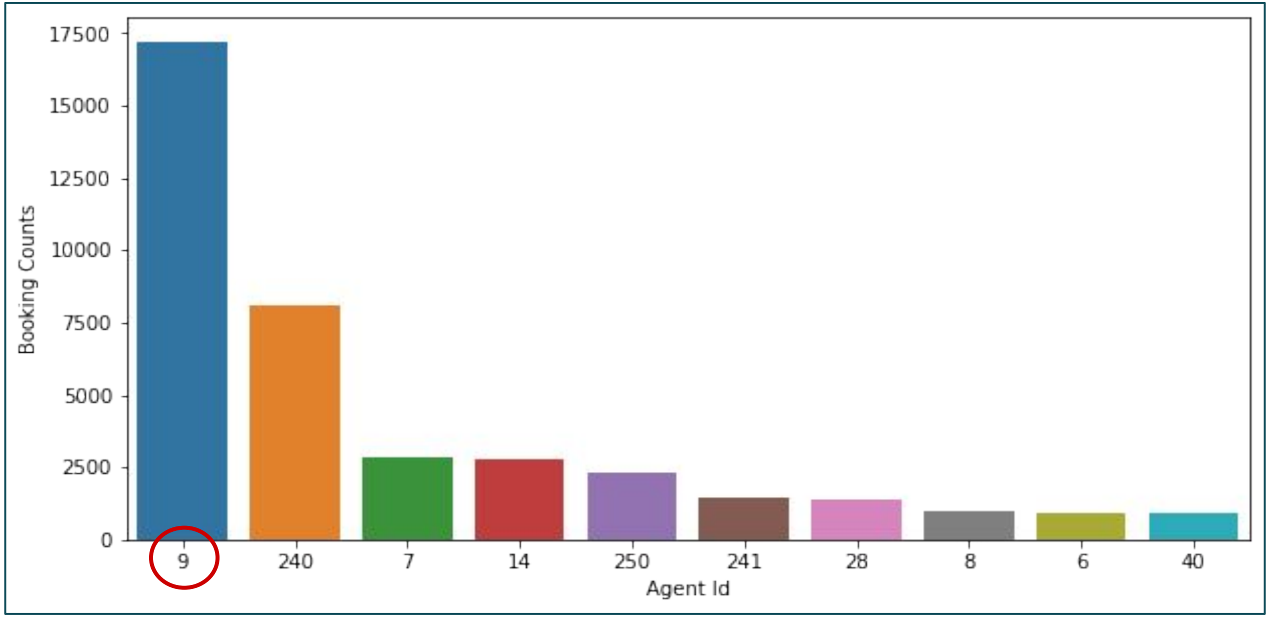
Q3. What is the year wise trend of bookings for each hotel?





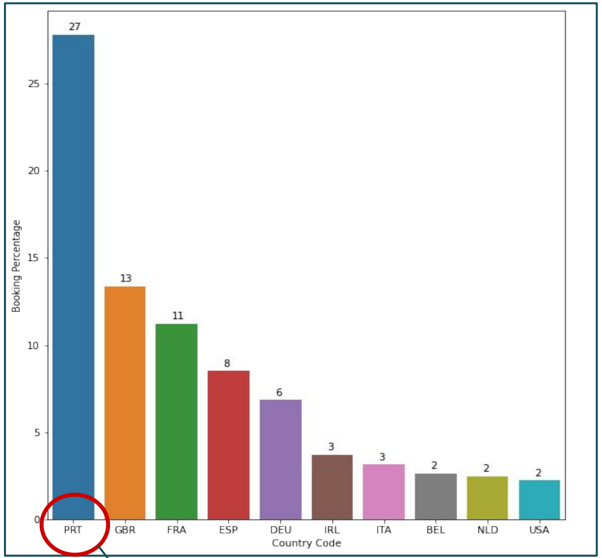
Performing the EDA (Cont.)

Q4. Which agent made the most number of bookings?

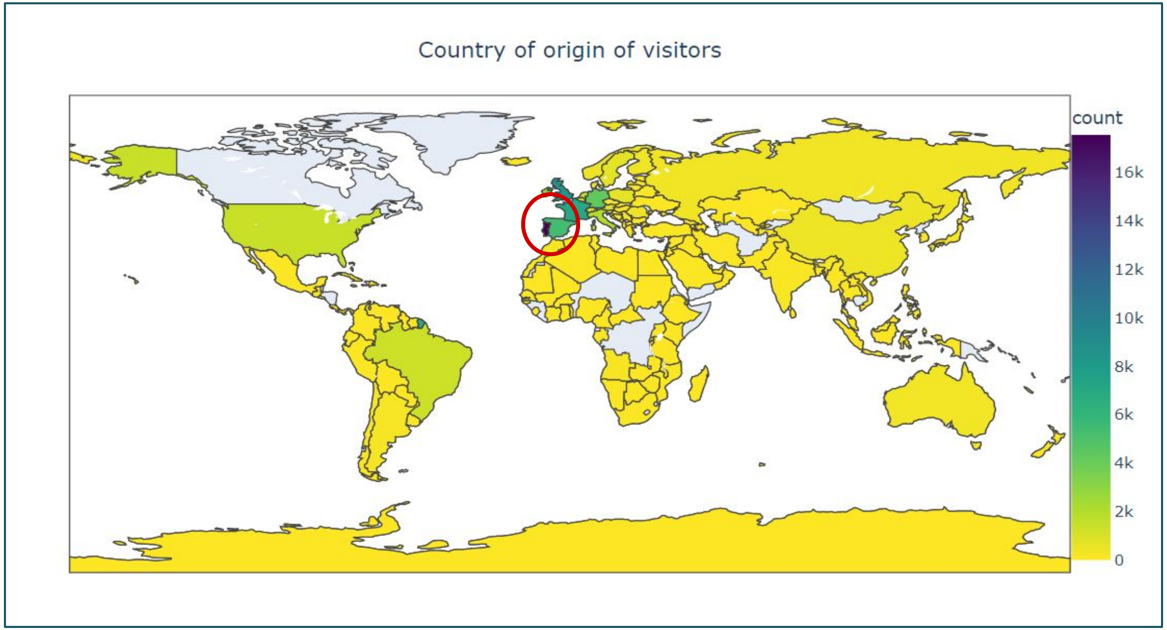


Performing the EDA (Cont.)

Q5. Enlist the country of origin of the majority of visitors.

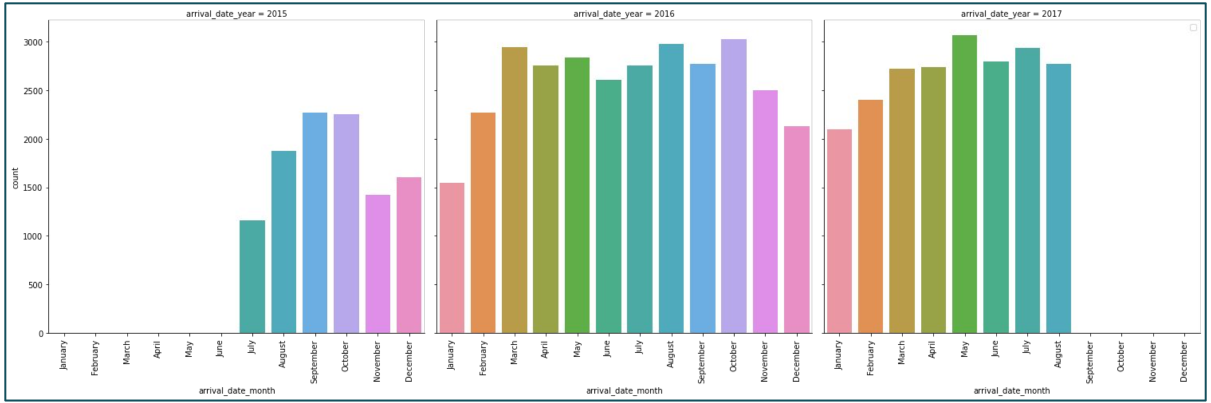


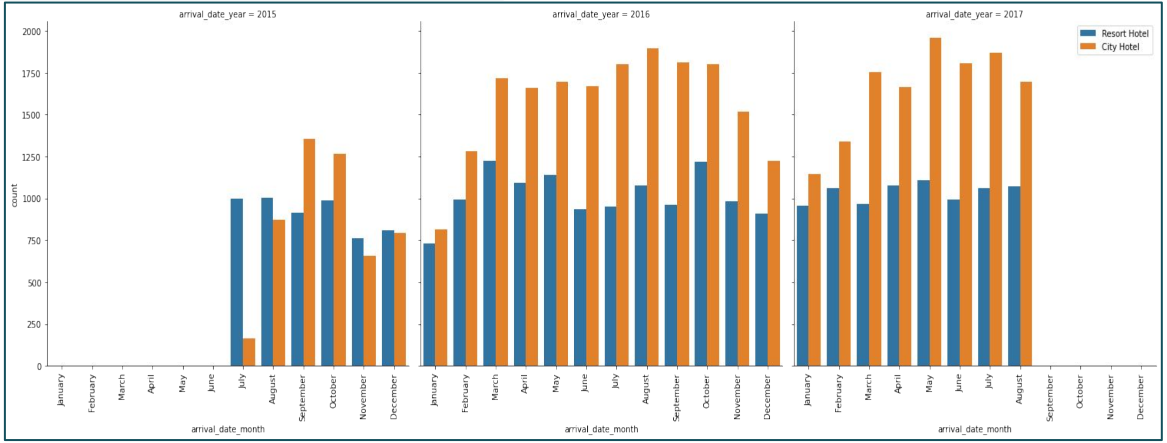
Portugal



Performing the EDA (Cont.)

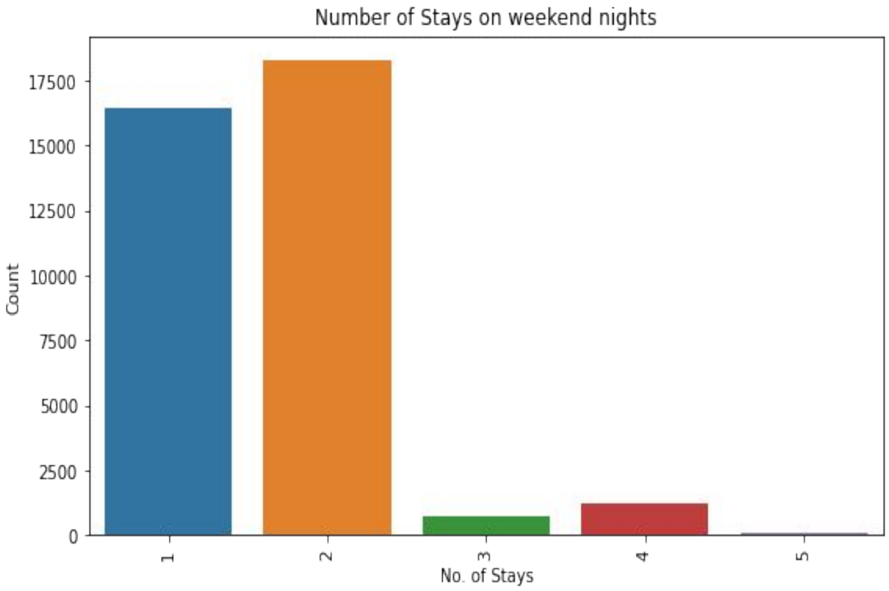
Q6. What is the busiest time for hotels?

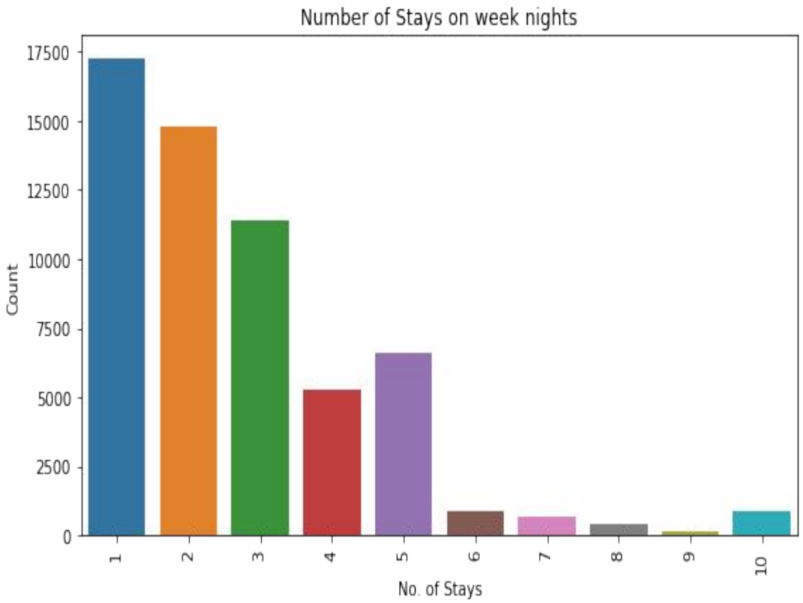




Performing the EDA (Cont.)

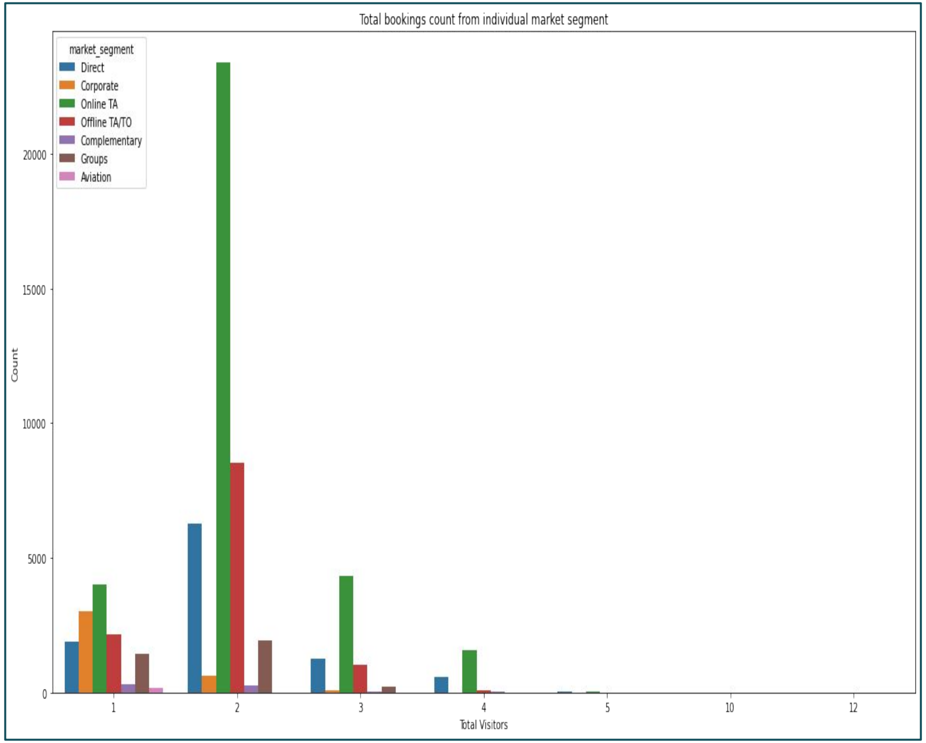
Q7. What is the proportion of weekend and weekday nights? Is there any difference between them?

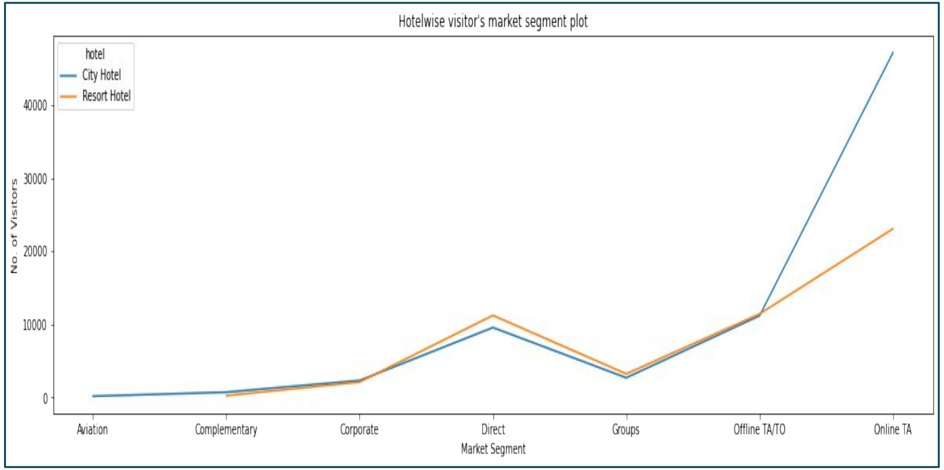


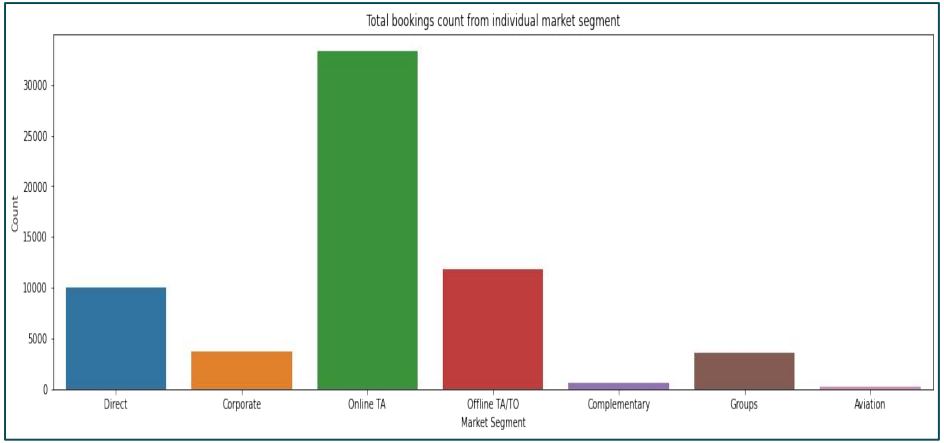


Performing the EDA (Cont.)

Q9. Which market segment does most visitors come from?

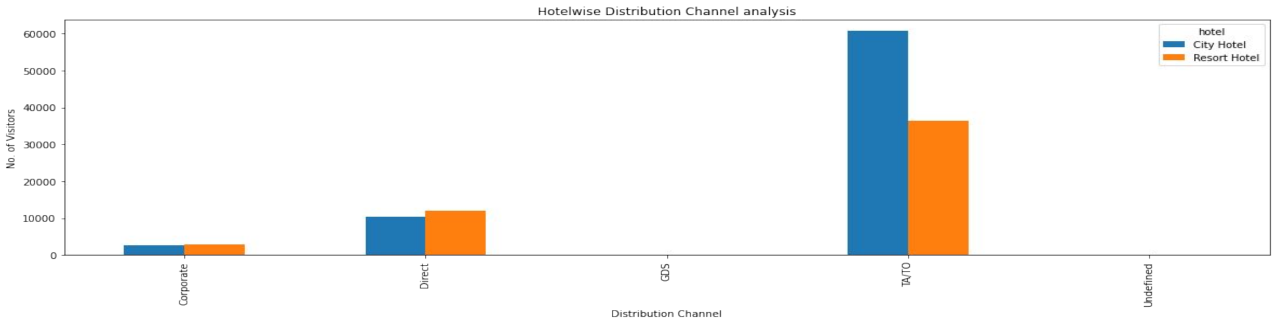


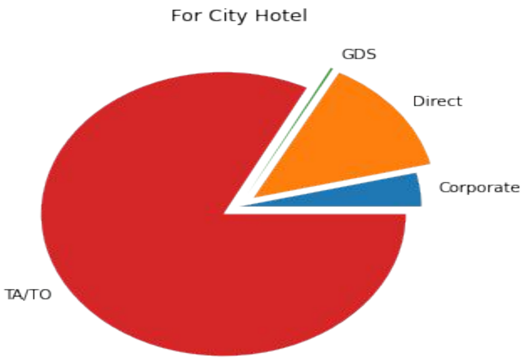


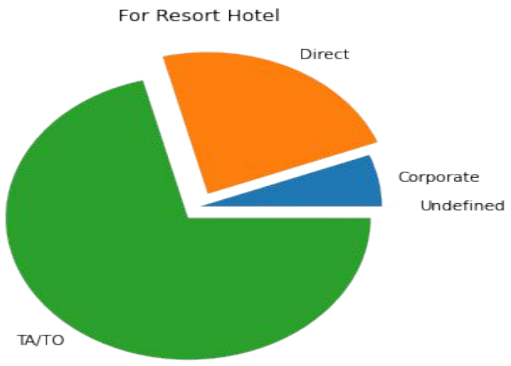


Performing the EDA (Cont.)

Q10. Which distribution channel does most visitors come from?

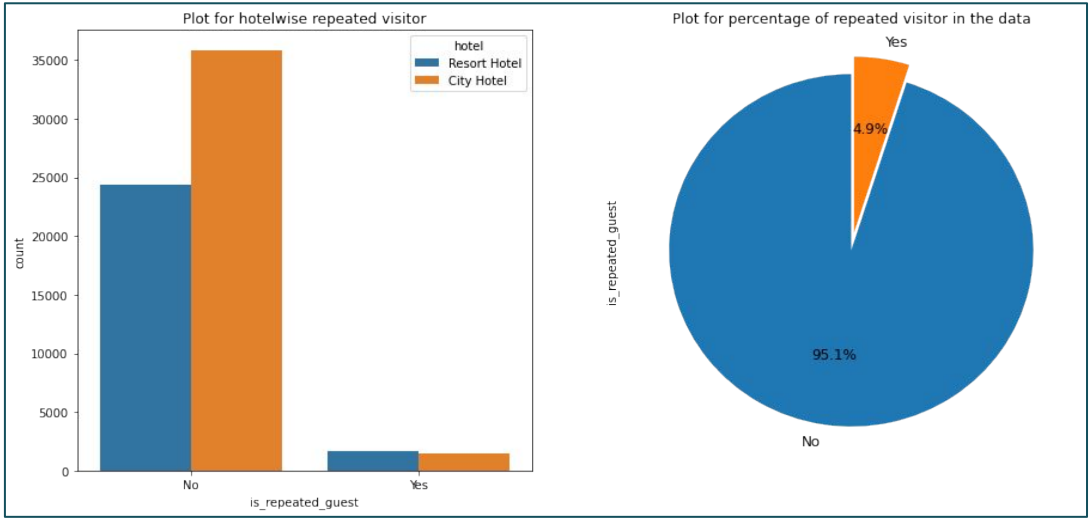




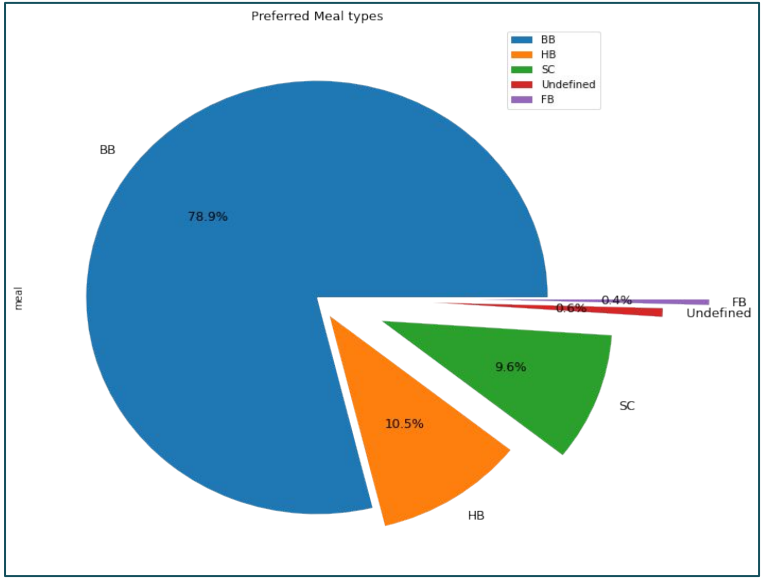


Performing the EDA (Cont.)

Q11. How many visitors are repeating?

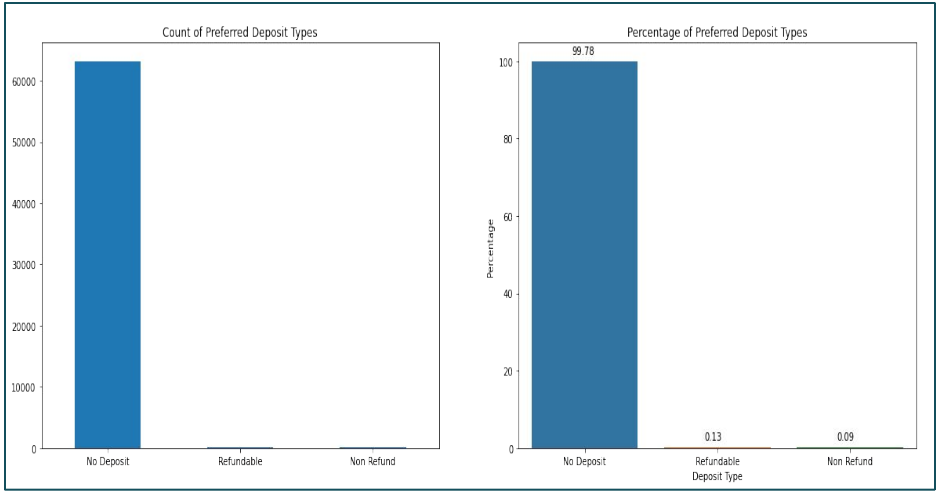


Q12. Which is the most preferred meal?

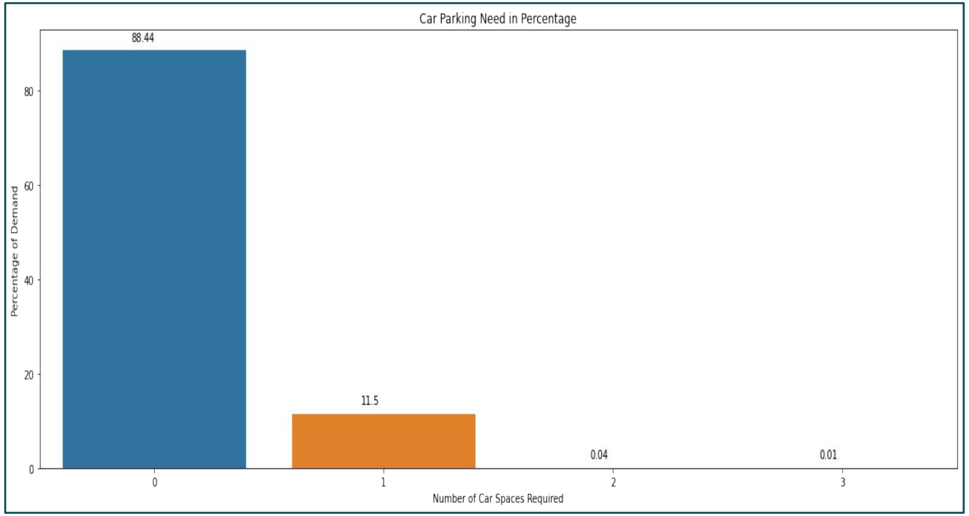


Performing the EDA (Cont.)

Q13. Which is the most preferred deposit type?

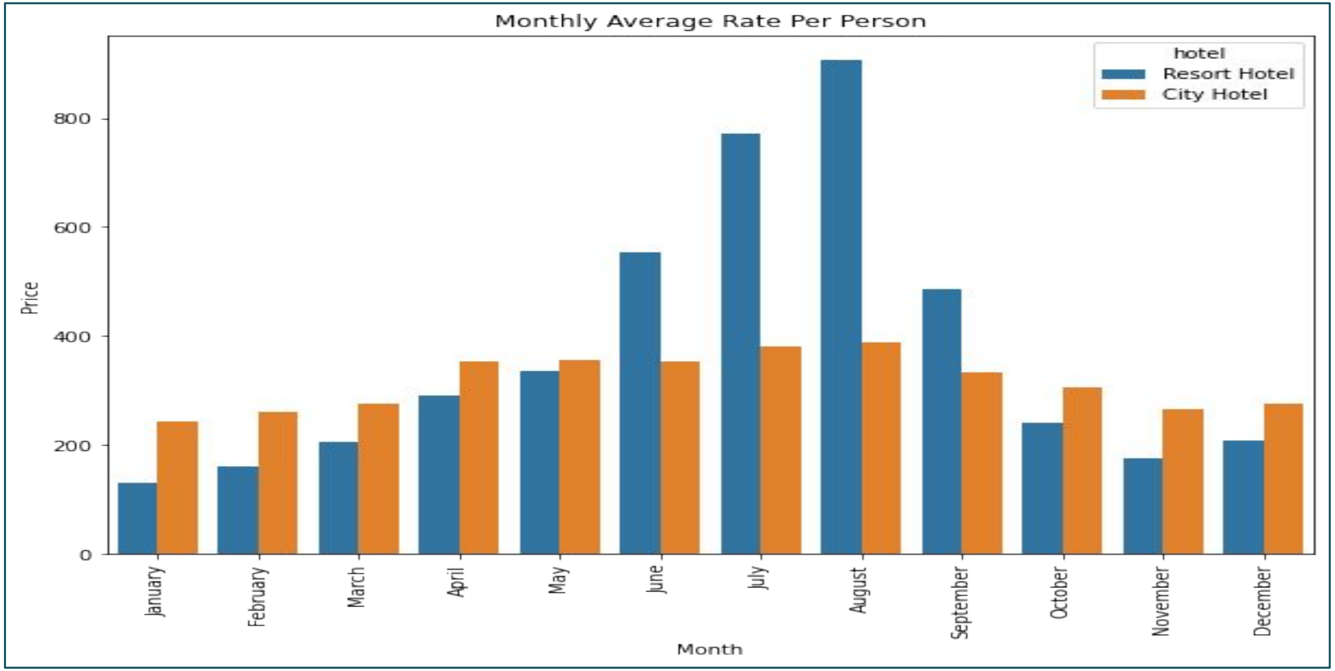


Q14. How many visitors asked for car parking space?



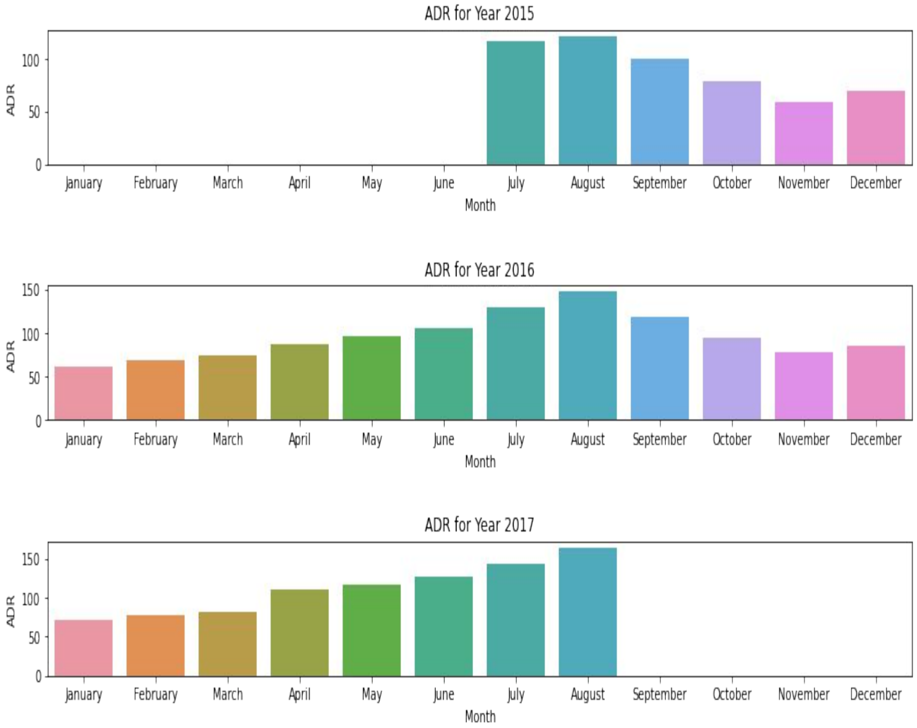
Performing the EDA (Cont.)

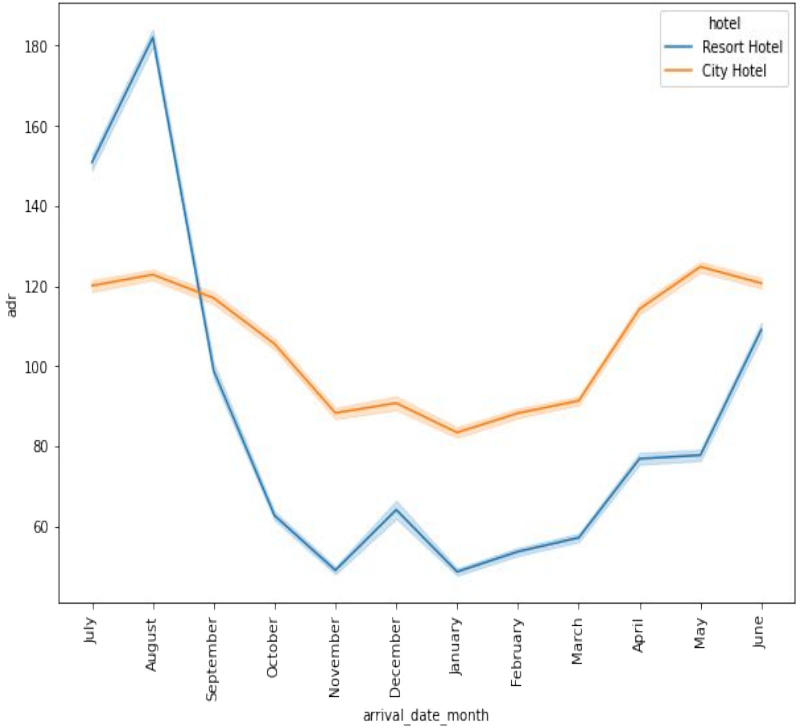
Q15. Which month has the highest average daily rate per person?



Performing the EDA (Cont.)

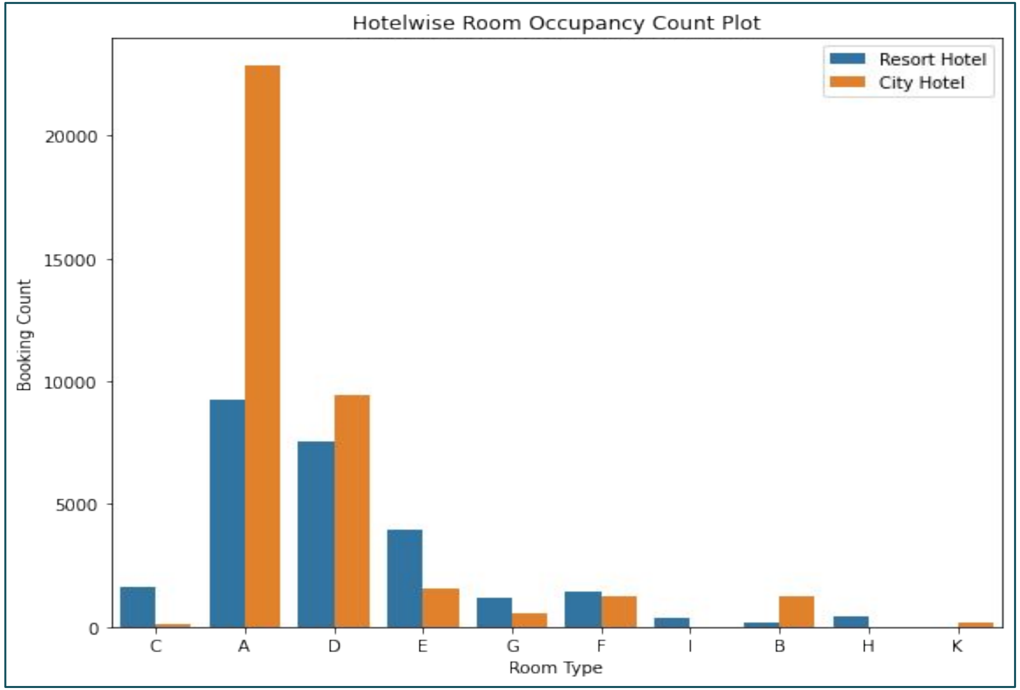
Q16. What is the trend of ADR?

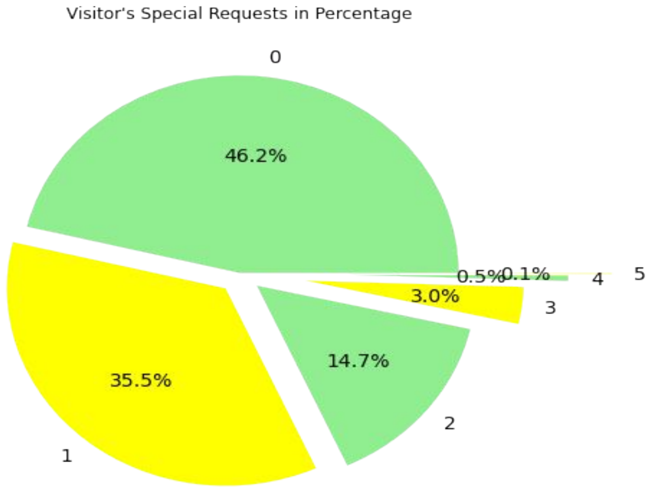




Performing the EDA (Cont.)

Q17. Which room Type is high in demand?

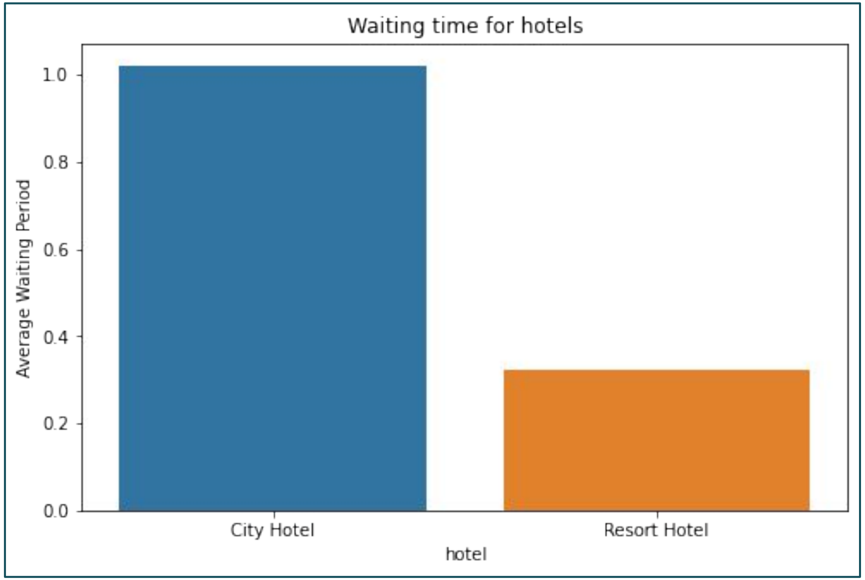




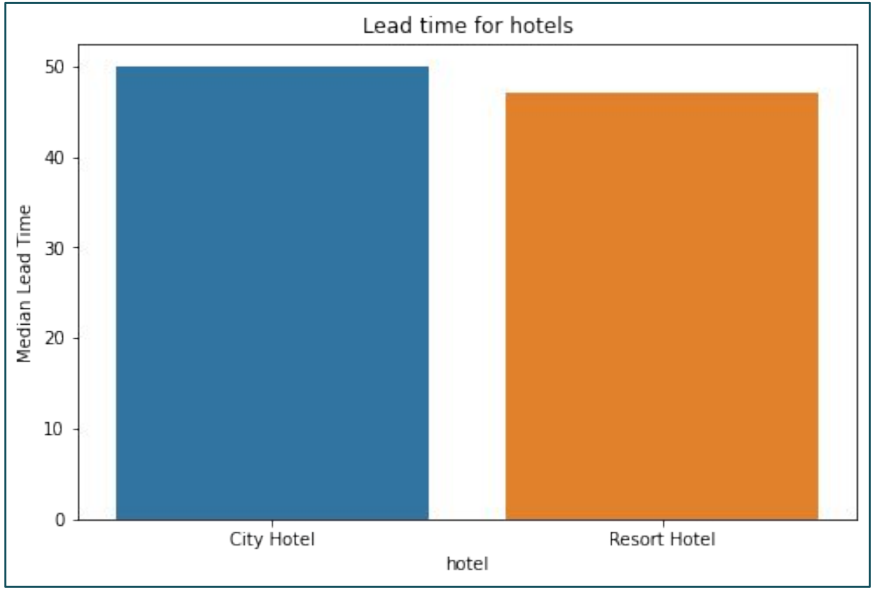
Q18. How likely is the hotel to receive a disproportionately high number of special requests?

Performing the EDA (Cont.)

Q19. Which hotel type has a longer waiting time for booking?



Q20. Which hotel type has a higher lead time for booking?



Conclusions:

From the EDA we can conclude that:

1. Majority of the booking came for City Hotel, i.e. 58.9%, which City Hotel is more preferred.
2. Booking trend for both the hotels is nearly the same. However, talking about the volume of bookings, it is the same in 2015, but for 2016, City Hotel received more bookings.
3. Agent with Id number: 9, made the most bookings.
4. Majority of the visitors arrived from Portugal.
5. Occupancy of hotels:

   2015 - September and October are busiest

   2016 - August followed by July, September and October are busiest

   2017 - May and July are the busiest.
6. Single visitors preferred weekday stays, while visitors traveling in pairs preferred weekend stays more. Possibly, they are couples. ;)
7. Majority of the visitors arrived from online travel agents (TA) market segment. The same applies to distribution channels.
8. Majority of the time booking for visitors traveling in pairs arrived via online travel agents (TA) .
9. Majority of the visitors preferred meal type BB (Bed & Breakfast).
10. Only 4.9% of customers are repeated.
11. Of the arriving customers, a total 538 bookings were previously canceled.
12. Majority of the customers do not prefer to pay a deposit amount.
13. About 88.4% of visitors did not require car parking space.
14. August has the highest average daily rate per person.
15. ADR for resort hotel types is quite fluctuating compared to that of city hotels. When checked yearly for months, the ADR forms a bell shaped curve with August at the center. The month of January has the least ADR value.
16. Room Type A is high in demand.
17. 46.2% of visitors do not have any special request.35.5% visitors have 1 special request.
18. City Hotel takes longer to confirm booking status.
19. City Hotel has slightly higher lead time compared to the resort hotel.

Thank you

References of Images:

1. Vector image and graphics of the muse: shutterstock.com.
2. Lost the path: hovercraftdoggy.com
3. EDA vector image: medium.com