

Capstone Project

Hotel Booking Analysis

Project Done By:

Vivek Kumar Singh

Problem Statement:

- We have got the data of a Hotel's booking system .There are two types of hotels 'resort hotel' and 'city hotel' Our goal is to find and extract meaningful insights from data. Using these insights hotels can take steps to attract more customers and thus increase their profits thereby.

Data Summary:

- **hotel** : There is two type of Hotel one is Resort Hotel and the other is City Hotel
- **is_canceled** : It gives the status if the booking wheter it has been cancelled or not
- **lead_time** : It gives the difference in days between booking and the arrival of the customer
- **arrival_date_year** : Gives year in which the visitor arrived
- **arrival_date_month** : Gives month in which the visitor arrived
- **stays_in_weekend_nights** : Number of weekend nights spent i.e. Saturday and Sunday
- **stays_in_week_nights**: Number of week nightsspent i.e. Monday to Friday
- **adults** : Number of adults per booking

Data Summary:

- **children** : Number of children per booking
- **babies** : Number of babies per booking
- **meal** : Type of meal preferred by customers
- **country** : Country of origin of visitor
- **market_segment** : group of people based on market
- **distribution_channel** : Name of the distribution channel
- **is_repeated_guest** : Tells wheter the guest is a repeat customer or first time visitor.
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking

Data Summary:

- **previous_bookings_not_canceled** : Number of previous bookings not canceled by the customer prior to the current booking
- **reserved_room_type** :Type of room reserved
- **assigned_room_type** : Type of room assigned during booking
- **booking_changes** : Number of bookings changed
- **deposit_type** : Types of deposit
- **agent** : Gives the ID of the agent
- **company** : Gives the ID of the Company related to the booking
- **day_in_waiting_list** : Number of days the booking was in the waiting list before confirmation
- **customer_type** : Type of customer

Data Summary:

- **adr : average daily rate**
- **required_car_parking_spaces : Number of car parking spaces required by the customer**
- **total_of_special_requests : Gives the number of special requests made by the customer**
- **reservation_status : Gives the status of the reservation**
- **reservation_status_date : Date at which the last status was updated**

What is EDA?

Exploratory Data Analysis is the critical process to perform initial investigations on data so as to discover patterns, to spot anomalies to check assumptions with the help of summary statistics and graphical representations.

IMPORTANCE:

It is very useful in understanding the data .The insights generated are critical and helpful in decision making.

Approach:

Approach the problem in simple steps:

1. Understanding the Data
2. Data Pre-Processing
3. Formulate Questions about the Data
4. Performing Exploratory Data Analysis (EDA)
5. Answer the questions based on analysis and draw out the conclusions.

Data Pre-Processing

Viewing Data :

- Firstly we look at all the columns as what they represent and check the overall shape of the data
- Secondly we check the data type of each column to see what type of data we are going to deal with.

Cleaning the Data:

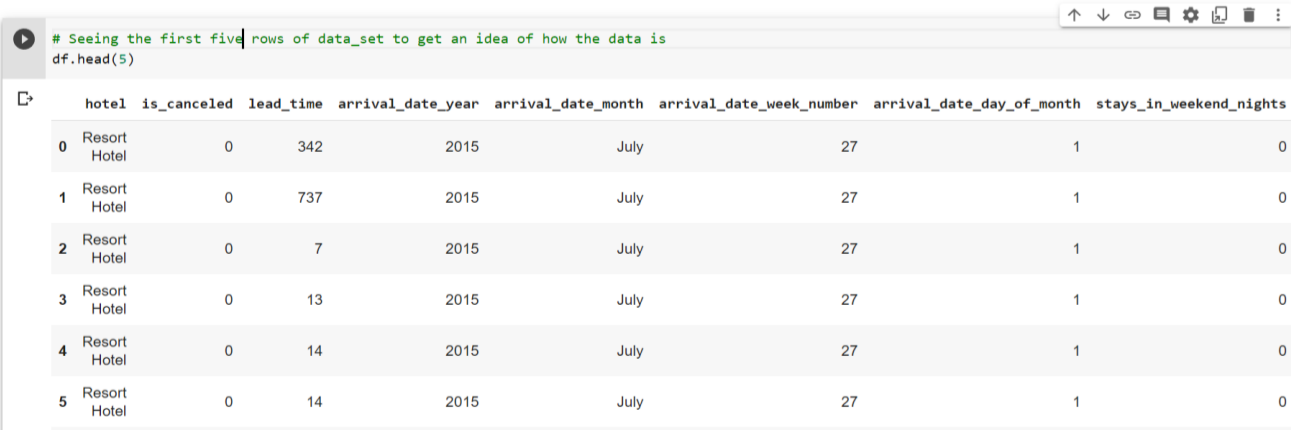
- We check the null or missing values and fill the numerical missing values with zero or mean value if the column data has arithmetical significance and categorical ones with mode.
- We delete the rows having cancelled bookings

Viewing the Data :

Quick look:

- Size of data : (119390, 32) => 119390 rows and 32 features
- Viewing first 5 rows

UNDERSTANDING THE STRUCTURE OF DATA



```
# Seeing the first five rows of data_set to get an idea of how the data is
df.head(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights
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	Resort Hotel	0	13	2015	July	27	1	0
4	Resort Hotel	0	14	2015	July	27	1	0
5	Resort Hotel	0	14	2015	July	27	1	0

Data Cleaning:

- Checking The Missing or Null values:

```
#We want to check the missing values or null values
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	0
previous_cancellations	0
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0
reservation_status	0
previous_bookings_not_canceled	0
is_repeated_guest	0
is_canceled	0
distribution_channel	0
market_segment	0
meal	0
babies	0

Data Cleaning:

Checking and Dropping Duplicate Rows:

```
✓ [33] #Removing Duplicate Entries  
0s df[df.duplicated()].shape  
  
(31994, 32)
```

```
✓ [34] df.drop_duplicates(inplace = True)
```

Removing Cancelled bookings from our Dataset


```
# Removing Cancelled Bookings from our Data Set as it is not needed.  
df = df[df['is_canceled'] == 0]
```

Data Cleaning:

Dropping Rows without any visitor:

- ✓ [38] # Adding a Column to keep count of total visitors in each bookings
`df['total_visitors'] = df.adults + df.children + df.babies`
- ✓ [42] # Drop rows having zero total visitors
`df = df[df.total_visitors != 0]`

Imputing Data :

- ✓ [43] # Fill the Null values of agent and company column with zero
`df[['agent', 'company']] = df[['agent', 'company']].fillna(0.0)`
- ✓ [44] #For the missing values in the children column we replace it with mean value
`df['children'].fillna(df.children.mean(), inplace=True)`
- ✓  #For the missing values in the country column we replace it with mode value
`df['country'].fillna(df.country.mode().to_string(), inplace=True)`

Problem Formulation:

1. What is the booking ratio between the two types of Hotels?
2. What is the trend of Hotel Bookings by Months?
3. What is the trend of Hotel Bookings by Year?
4. Top ten countries by most visitors in Hotels
5. Which type of hotel is most preferred Hotel For Babies?
6. Distribution of lead time for Hotel Bookings
7. What is the Preferred Meal type?
8. Which Market Segment contribute most to Hotel Bookings?
9. Which type of room is allotted most?
10. Who are the Top ten performing agents?
11. What is the share of Distribution channel in Hotel Bookings?
12. What is the Average Daily Rate of Hotels by Months?
13. What Number of Visitors are Repeat Guests and how many are Non Repeat Guests?
14. What is the Correlation between the Variables?

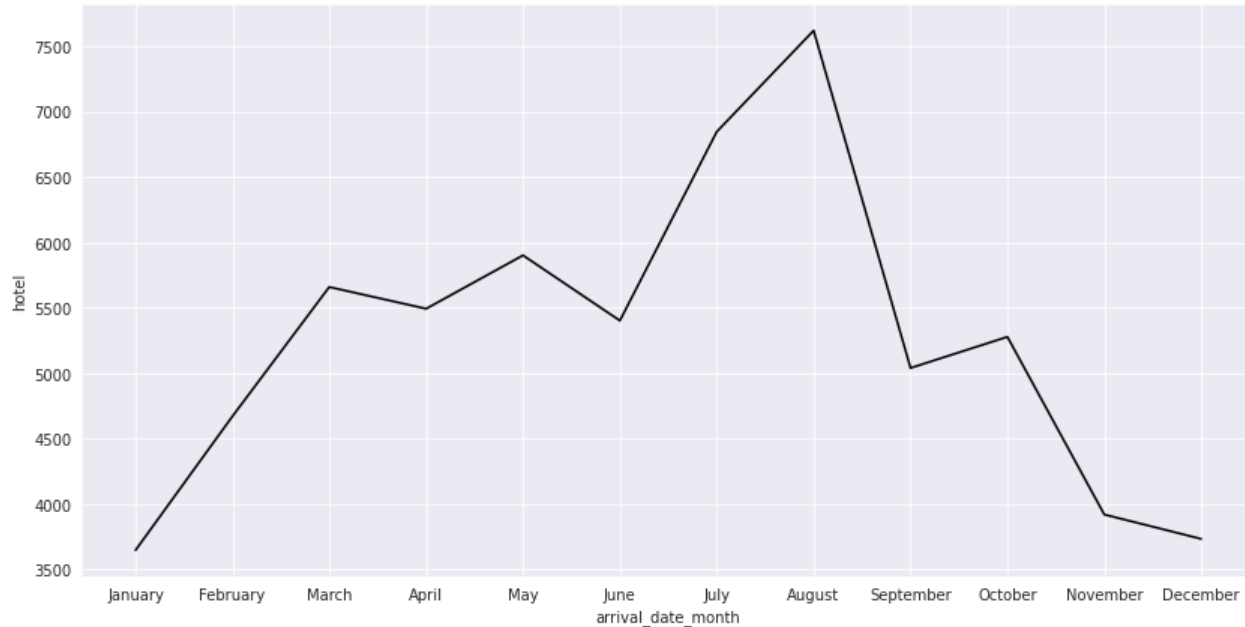
Exploratory Data Analysis(EDA)

- What is the booking ratio between the two types of Hotels?



Exploratory Data Analysis(EDA)

- What is the trend of Hotel Bookings by Months?



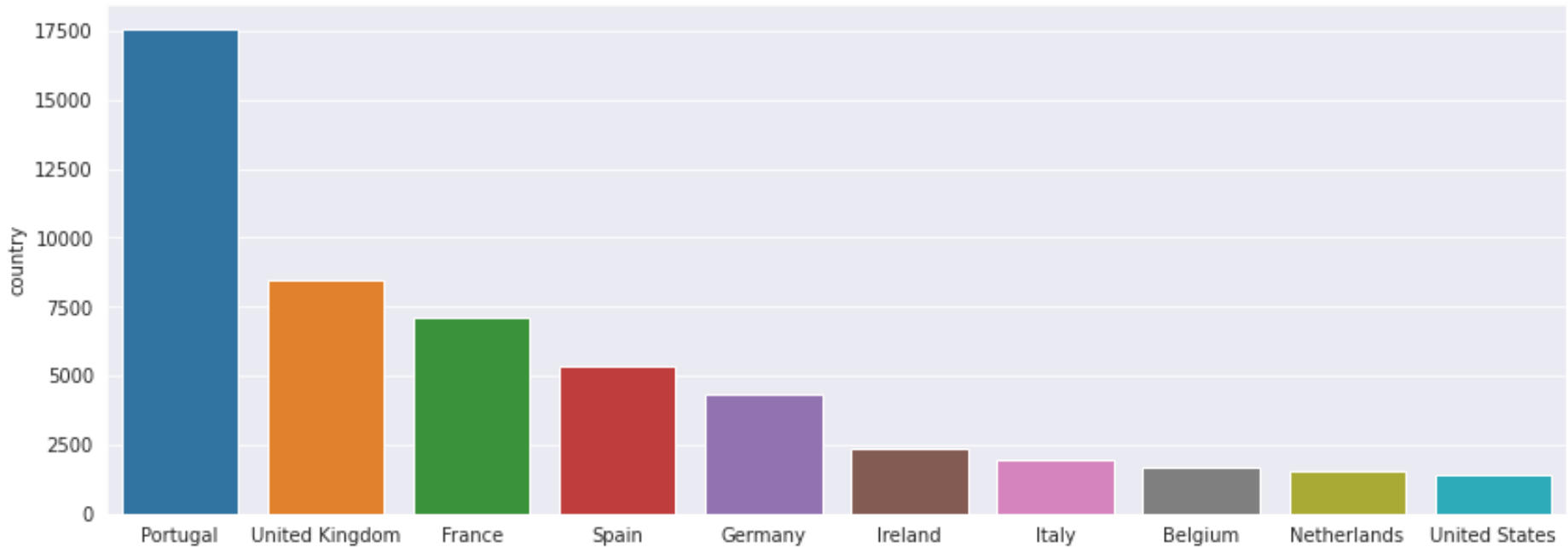
Exploratory Data Analysis(EDA)

- What is the trend of Hotel Bookings by year?



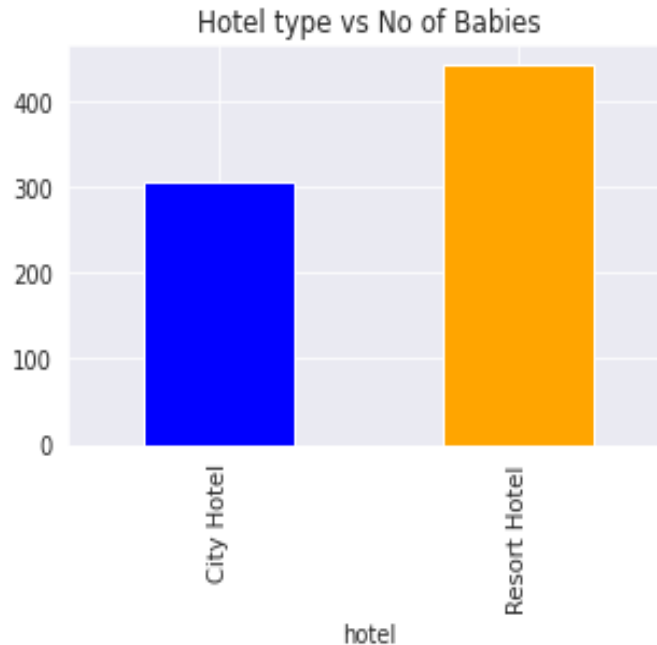
Exploratory Data Analysis(EDA)

- Top ten countries with most visitors



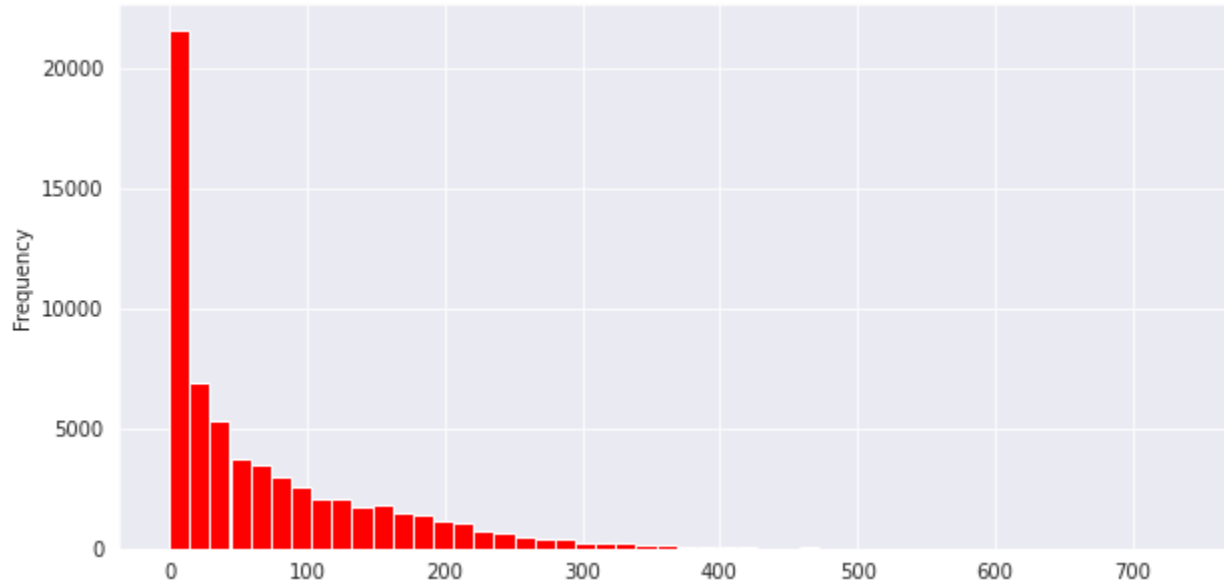
Exploratory Data Analysis(EDA)

- Which type of hotel is most preferred Hotel For Babies?



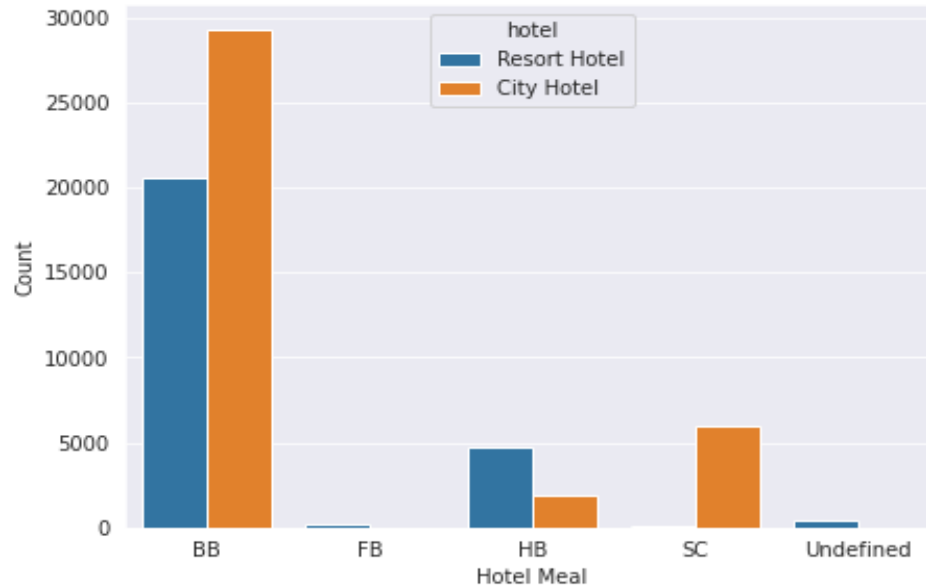
Exploratory Data Analysis(EDA)

- Histogram to show the distribution of lead time for Hotel Bookings



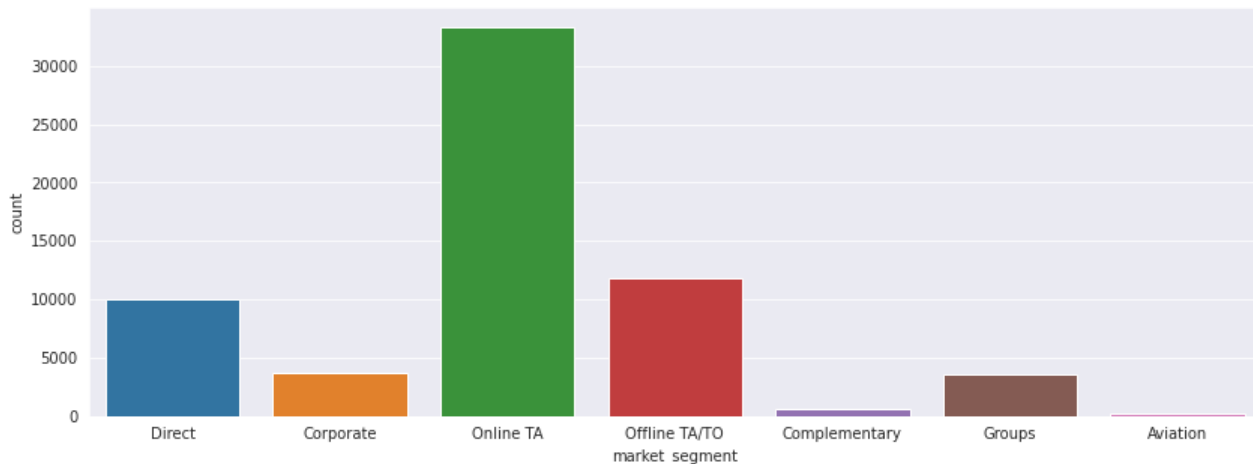
Exploratory Data Analysis(EDA)

- What is the Preferred Meal type?



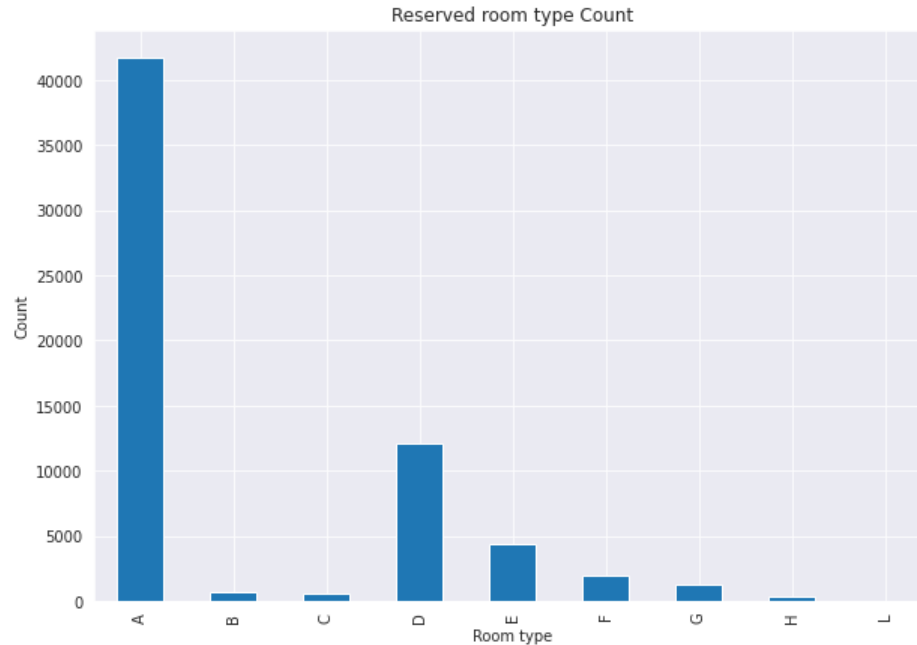
Exploratory Data Analysis(EDA)

- Which Market Segment contribute most to Hotel Bookings?



Exploratory Data Analysis(EDA)

- Which Room type is allotted most ?



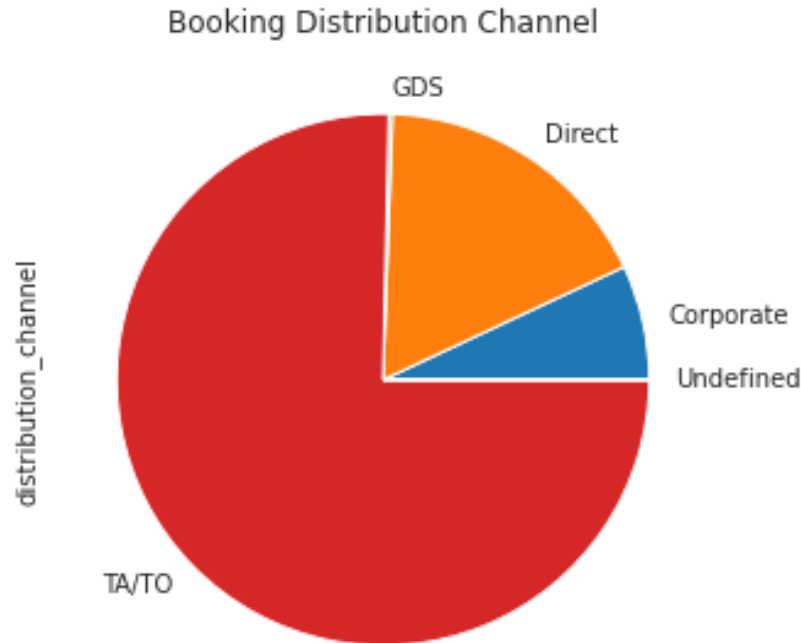
Exploratory Data Analysis(EDA)

- Who are the Top ten performing agents?



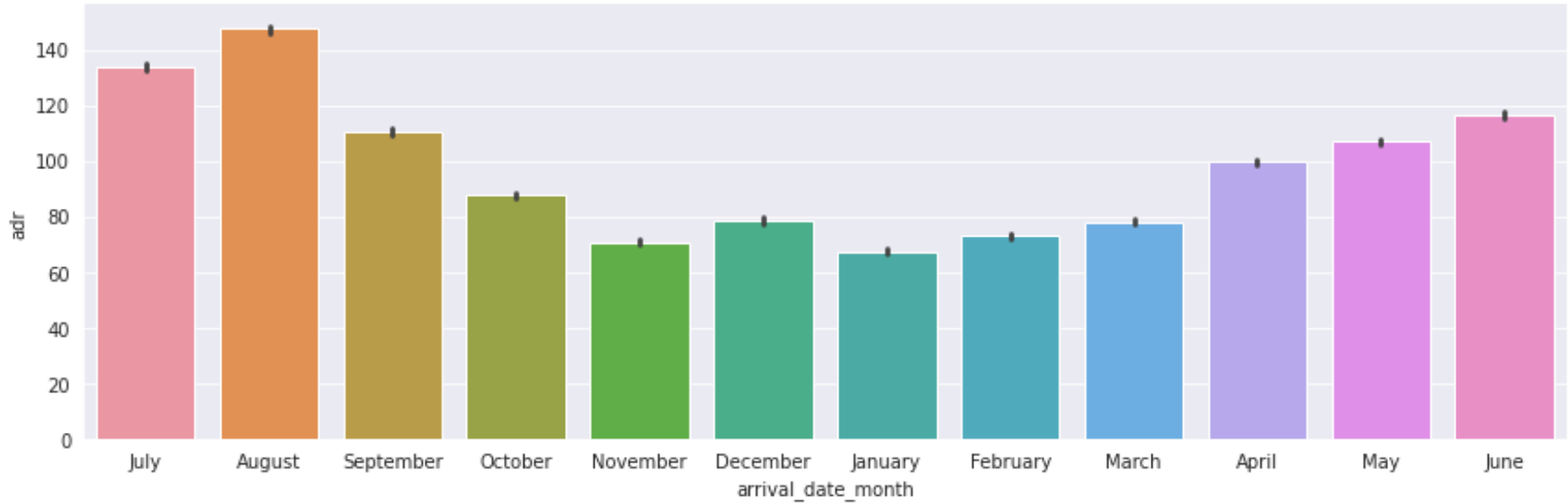
Exploratory Data Analysis(EDA)

- What is the share of Distribution channel in Hotel Bookings?



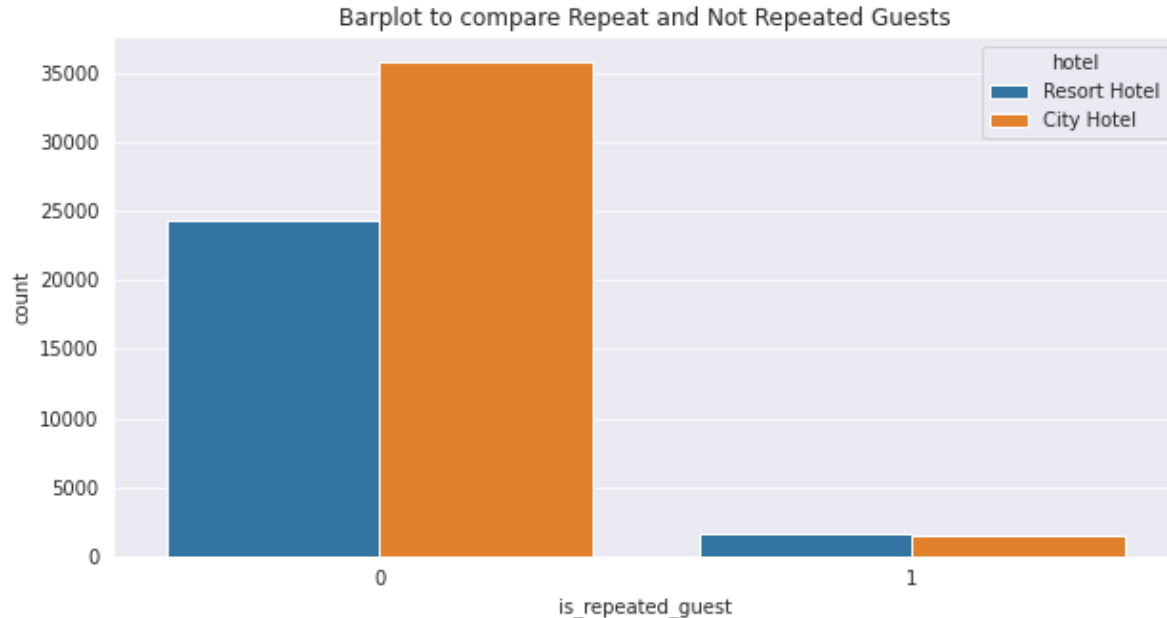
Exploratory Data Analysis(EDA)

- What is the Average Daily Rate of Hotels by Months?



Exploratory Data Analysis(EDA)

- What Number of Visitors are Repeat Guests and how many are Non Repeat Guests?



	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_visitors
is_canceled	1	0.12	0.1	0.014	0.28	0.36	0.16	0.02	-0.0012	-0.15	-0.045	-0.083	0.12	0.077	-0.063	0.14	0.028	-0.066	0.073	0.13	
lead_time	0.12	1	-0.53	0.0044	0.0470	0.0087	0.042	0.03	-0.016	0.053	0.011	0.039	0.019	-0.0025	0.034	-0.033	0.16	-0.028	0.088	0.047	
arrival_date_year	0.1	-0.53	1	-0.087	0.038	0.039	0.03	0.016	0.01	-0.052	-0.03	-0.026	0.012	0.017	-0.026	0.018	0.11	0.011	0.044	0.034	
arrival_date_week_number	0.014	-0.0044	0.087	1	-0.015	-0.0240	0.00710	0.018	-0.00010	0.00540	0.00640	0.000310	0.00670	0.00560	0.00180	0.0047	0.024	0.012	-0.0025	0.011	
arrival_date_day_of_month	0.28	0.0047	0.038	-0.015	1	0.54	0.13	0.0091	0.013	-0.12	-0.033	-0.062	0.05	0.15	-0.094	-0.031	0.039	-0.039	0.047	0.11	
stays_in_weekend_nights	0.36	-0.0087	0.039	-0.024	0.54	1	0.14	0.0074	0.014	-0.12	-0.032	-0.062	0.09	0.18	-0.064	0.01	0.05	-0.036	0.052	0.11	
stays_in_week_nights	0.16	0.042	0.03	-0.00071	0.13	0.14	1	0.037	0.028	-0.22	-0.07	-0.16	-0.04	0.049	-0.23	-0.019	0.33	0.032	0.19	0.77	
adults	0.02	0.03	0.016	0.018	0.00910	0.0074	0.037	1	0.021	-0.045	-0.018	-0.031	0.045	0.02	-0.051	-0.018	0.32	0.063	0.084	0.64	
children	-0.0012	-0.016	0.01	-0.0001	0.013	0.014	0.028	0.021	1	-0.016	-0.0059	-0.011	0.083	0.027	-0.014	-0.007	0.031	0.03	0.096	0.21	
babies	-0.15	0.053	-0.052	-0.0054	-0.12	-0.12	0.22	-0.045	-0.016	1	0.25	0.45	-0.0007	-0.07	0.2	-0.015	-0.17	0.06	-0.014	-0.19	
is_repeated_guest	-0.045	0.011	-0.03	-0.0064	-0.033	-0.032	-0.07	-0.0180	0.0059	0.25	1	0.54	-0.002	0.042	0.04	0.0046	0.05	0.009	0.014	-0.064	
previous_cancellations	-0.083	0.039	-0.0260	0.000310	0.062	-0.062	-0.16	-0.031	-0.011	0.45	0.54	1	0.000670	0.065	0.13	-0.008	-0.096	0.033	0.022	-0.14	
previous_bookings_not_canceled	0.12	0.019	0.012	0.0067	0.05	0.09	-0.04	0.045	0.083	0.0007	0.00020	0.0067	1	0.024	0.089	0.033	0.022	0.038	-0.0038	0.013	
booking_changes	0.077	-0.0025	0.017	0.0056	0.15	0.18	0.049	0.02	0.027	-0.07	-0.042	-0.065	0.024	1	-0.14	-0.01	-0.0099	0.15	0.022	0.053	
agent	-0.063	0.034	-0.026	-0.0018	-0.094	-0.064	-0.23	-0.051	-0.014	0.2	0.04	0.13	0.089	-0.14	1	-0.0098	-0.15	0.028	-0.13	-0.2	
company	0.14	-0.033	0.018	0.0047	-0.031	0.01	-0.019	-0.018	-0.007	-0.015	-0.0046	-0.008	0.033	-0.01	-0.0098	1	-0.031	-0.019	-0.048	-0.026	
days_in_waiting_list	0.028	0.16	0.11	0.024	0.039	0.05	0.33	0.32	0.031	-0.17	-0.05	-0.096	0.022	-0.0099	-0.15	-0.031	1	0.08	0.21	0.45	
adr	-0.066	-0.028	0.011	0.012	-0.039	-0.036	0.032	0.063	0.03	0.06	0.009	0.033	0.038	0.15	0.028	-0.019	0.08	1	0.031	0.067	
required_car_parking_spaces	0.073	0.088	0.044	-0.0025	0.047	0.052	0.19	0.084	0.096	-0.014	0.014	0.022	-0.00380	0.022	-0.13						

Conclusions:

1. The City hotel is booked much more than the Resort Hotel
2. The Bookings consistently increases from January to May ,then decreases in June and consistently increases till August and the sale reaches its peak and then almost consistently decreases then after.
3. We observe that the sales were lowest in year 2015 increased in 2016 but again decreased in 2017.
4. Most Visitors came to Hotel from Portugal.
5. The number of babies in city hotel vs resort hotel, and we conclude that the number of babies arriving in the city hotel is much less though the city hotel has much more overall bookings.
6. Advance Bookings are much less compared to immediate bookings.
7. BB type meal is most preferred.
8. The online market segment contributes most to the Hotel Bookings
9. Room type 'A' is allotted most.
10. The agent with agent id '9' performs best.
11. The share of TA/TO distribution channel is most
12. The average daily rate of August is most?
13. The Repeat Customers are significantly minute in comparison to Non Repeating Customers
14. The correlation matrix plot shows the correlation between every variable pair.