Capstone Project

HOTEL BOOKING ANALYSIS

Team Member:-

1) **Kuresh Chandra Tripathy**

2) **N Santosh Kumar Choudhury**

Points to Discuss:

1. Agenda
2. Data analysis
    1. Univariate Analysis
    2. Bivariate Analysis
    3. Multivariate Analysis
3. Data summary
4. Hotel wise analysis
5. Distribution Channel wise analysis
6. Booking cancellation analysis
7. Timewise
8. Correlation heatmap
9. Some important questions
10. Conclusion

Agenda

To discuss the analysis of given hotel bookings data set from 2015-2017.

We'll be doing analysis of given data set in following ways :

- Data analysis
- Hotel wise analysis
- Distribution Channel wise analysis
- Booking cancellation analysis
- Timewise analysis

By doing this we'll try to find out key factors driving the hotel bookings trends.

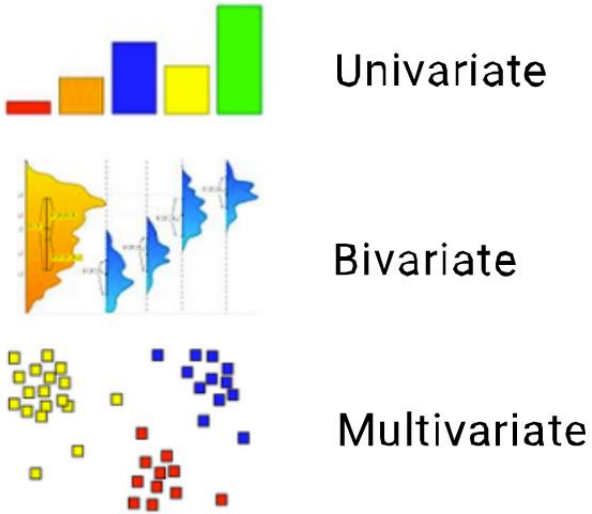
Data Analysis

Q) What is Data Analysis ?

Data Analysis is the process of collecting, cleaning, sorting, and processing raw data to extract relevant and valuable information for various organization and business Operation and Maintenance.

Types:-

1. Univariate Analysis
2. Bivariate Analysis
3. Multivariate Analysis



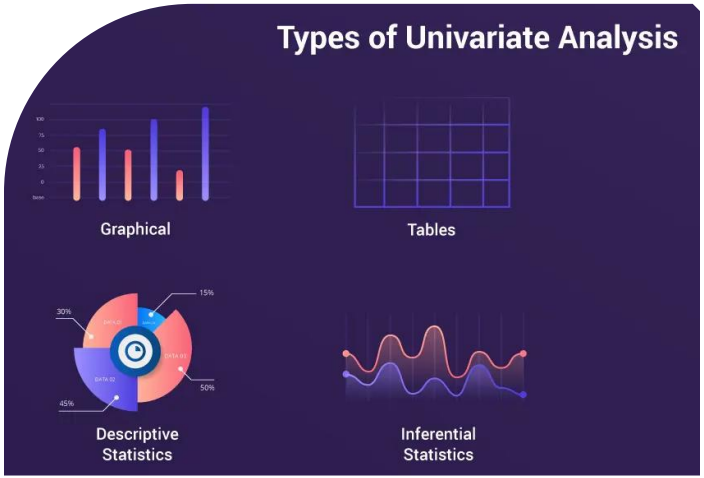
Univariate Analysis

Q) What is Univariate Analysis ?

- **Univariate** is a type of data analysis consists of observations on only a single characteristic or attribute it can be any numerical data or may be nonnumerical data (such as eye colors of brown or blue).
- In other word the examination of the distribution of cases on only one variable at a time.(e.g. Weight of ALMA BETTER Student)

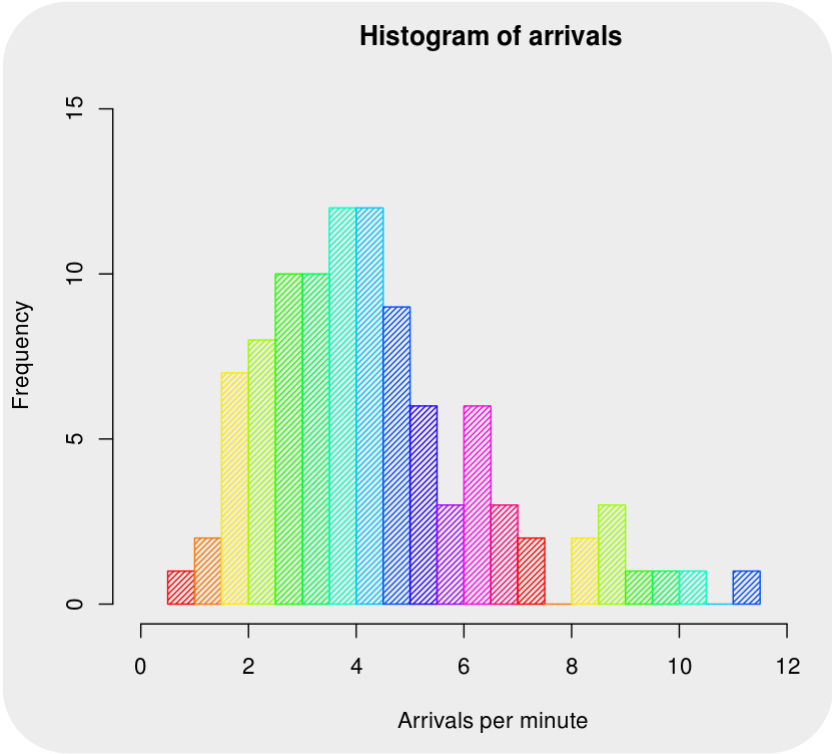
Example:-

1. Histograms
2. Frequency Distribution Tables
3. Frequency Polygons
4. Pie Charts
5. Bar Charts



Histograms

- **Histograms** is a graphical way of representing statistical or quantitative data with the bars of different height.
- In **Histograms** mostly data are compared with numerical data in the bars.
- Here mostly two dimension are there x-axis and y-axis. Where we plot our histogram.
- Here in this example X-axis represent the number of arrivals per minute and Y- axis represent Frequency.



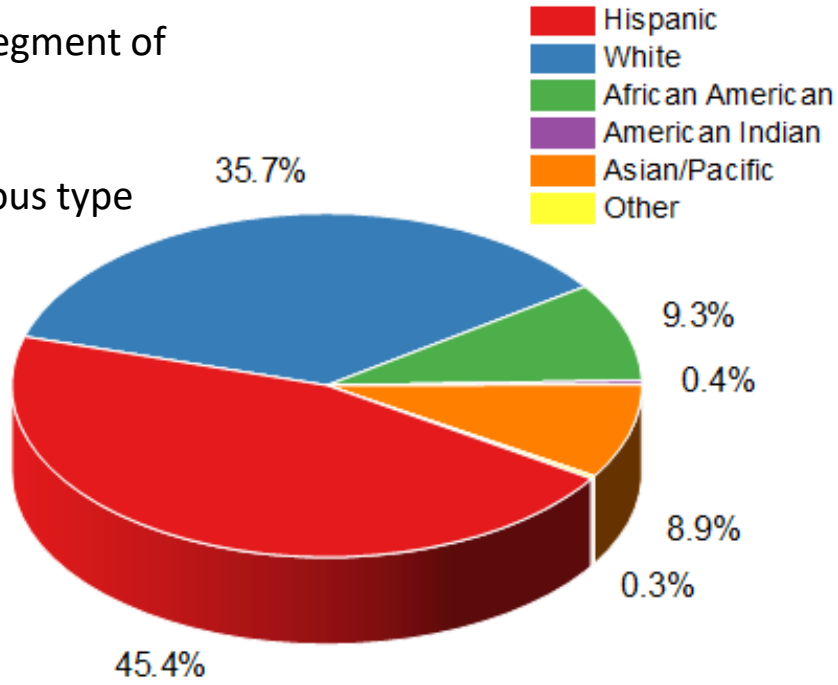
Frequency Distribution Tables

- **Frequency Distribution Tables** it show a data occurs how many number of times.
- Here in this example we can clearly see that, this table show the heights of some sample data with different frequency.
- So it is easy to understand the occurrence of any values in the FDT (Frequency Distribution Tables).

Heights M	Frequency
1.34 - 1.39	4
1.40 - 1.45	14
1.46 - 1.51	31
1.52 - 1.57	19
1.58 - 1.63	14
1.64 - 1.69	6

Pie Charts

- **Pie Charts** is a pictorial representation of a statistical data in a circular graph.
- Mostly it divided whole circle unit to different segment of circle.
- Each slices illustrate numerical proportion/ various type of proportion

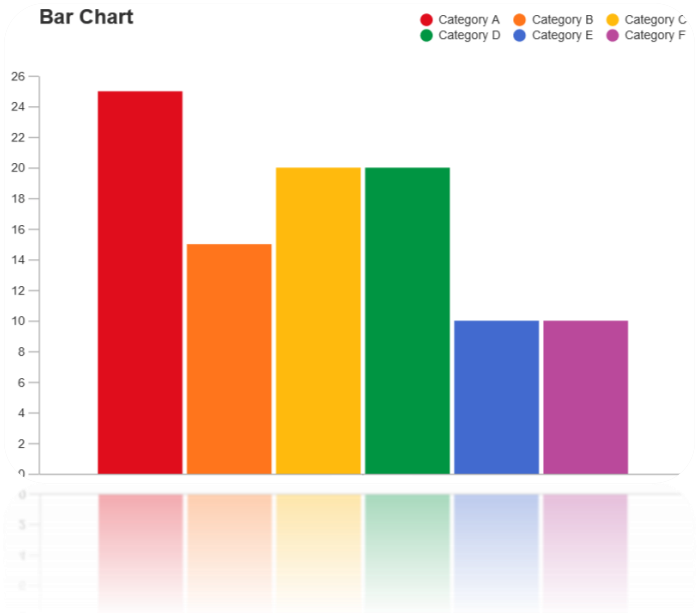


Bar Charts

- **A bar** chart provides a way of showing data values represented as vertical bars/horizontal bars. It is sometimes used to show trend data, and the comparison of multiple data sets side by side.

- *Bar chart **vs** Histogram*

  Unlike **histograms**, the **bars** in bar charts have spaces between them to emphasize that each bar represents a discrete value, whereas histograms are for continuous data.



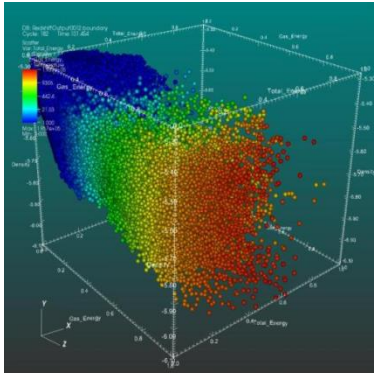
Bivariate analysis

Q) What is Bivariate analysis?

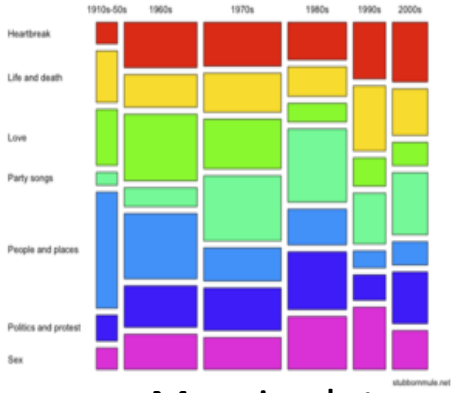
- **Bivariate** is a type of data analysis consists of analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.
- Bivariate Analysis:- The examination of two variables simultaneously.(e.g. The relationship between Gender and Weight of ALMA BETTER Student)
- Empirical relationship is nothing but a correlation that is supported by experiment and observation but not necessarily supported by theory.)

Example:-

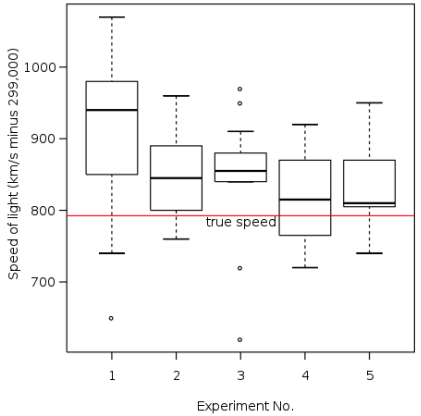
1. Scatter plot
2. Box plot
3. Mosaic plot



Scatter plot



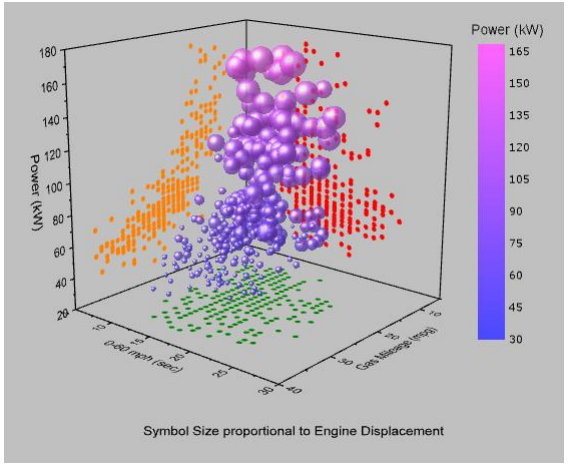
Mosaic plot

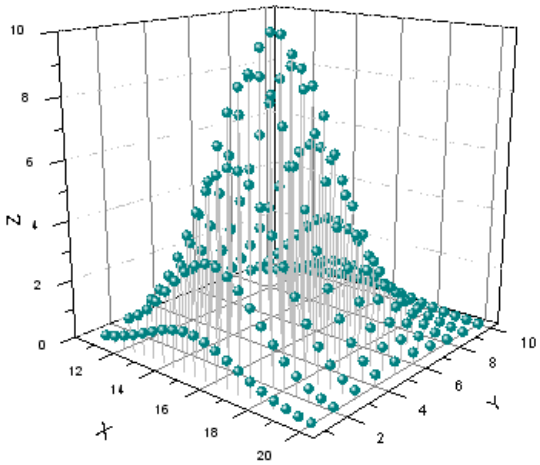


Box plot

Scatter plot

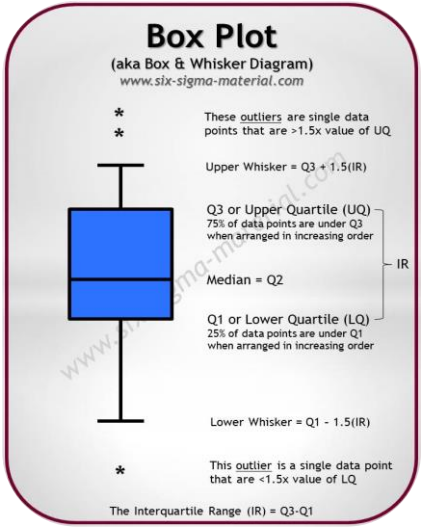
- A **scatter plot** is a graphical representation of data visualization in which it shows the relationship between different variables.
- A scatter plot used dots to represent values for two different numeric variables.
- The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

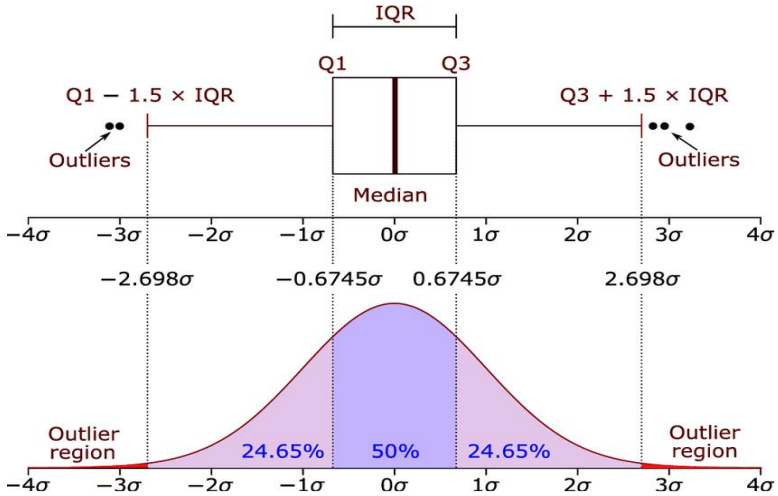




Box plot

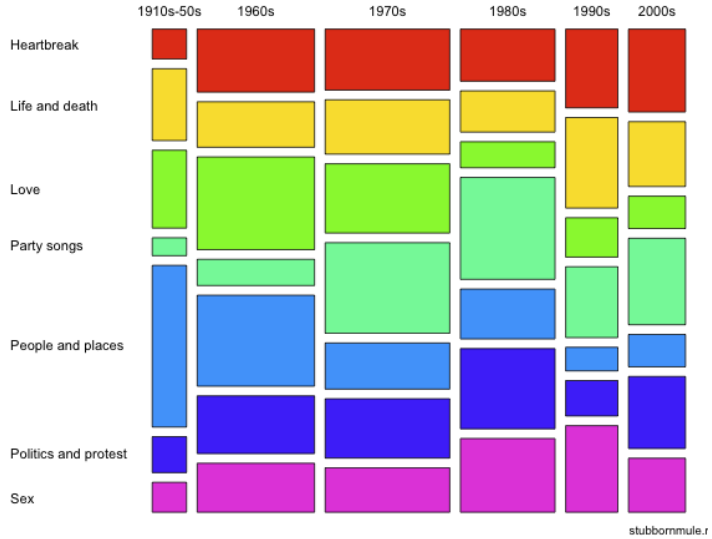
- **A box** plot also know as box and whisker plot is a summary of five point data.
- These five-number summary are the minimum, first quartile, median, third quartile, and maximum.
- In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.

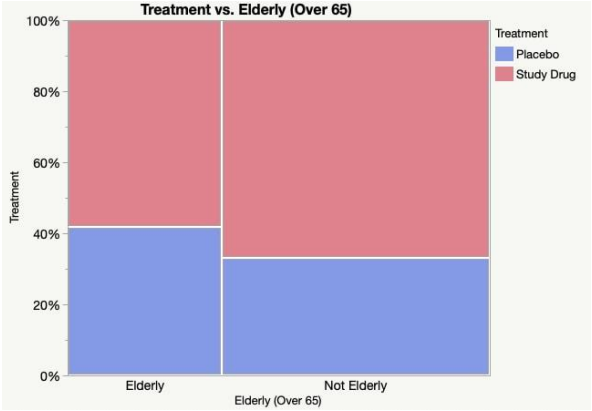




Mosaic plot

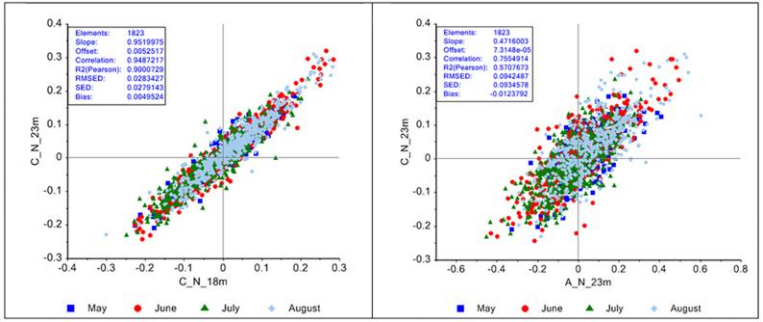
- **A Mosaic plot** is a graphical visualization of data from two or more qualitative variables.
- It gives an overview of the data and makes it possible to recognize relationships between different variables.
- It show percentages of data in groups.
- Mosaic plots are used to show relationships and to provide a visual comparison of groups.





Multivariate Analysis

- The statistical study of data where multiple measurements are made on each experimental unit.
- Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle.It gives an overview of the data and makes it possible to recognize relationships between different variables.



Important Libraries:-

- **Pandas** library is one of the most widely used tools for Data Science and Machine learning, it used for data cleaning and analysis.
- Handling data using pandas is very fast and effective by using pandas Series and data frame.
- It supports multiple file formats. It can read or load data in many formats like CSV, Excel, SQL, etc.,

- **Numpy** library is widely used mostly for the mathematical work like large, multi-dimensional arrays and matrices.
- Adds support for large collection of high-level mathematical functions to operate on these arrays.

- **Folium** library is used for graphical presenting map data

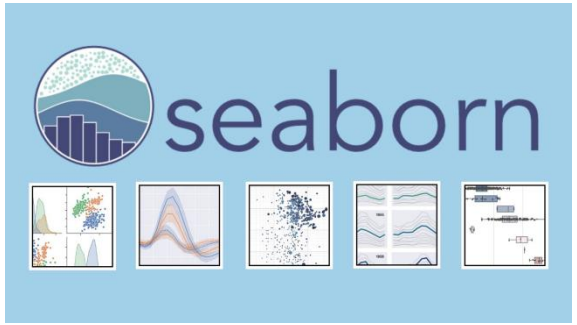
Important Libraries:-

- **Matplotlib** is a python libraries which is used to plot 2-d graphs and plots by using python.
- It has a module named pyplot which makes things easy for plotting by providing feature like controlling line styles, font size and style, formatting axes etc.

- **Seaborn** is one of the most advance than matplotlib Libraries which gives only basic 2-d graphs and plot where as seaborn give more faster and very attractive visualization of 2-d and 3-d graph and plot.
- Seaborn make visualization as central point of exploring and understanding data in data analysis.
- It is also closely integrated with panda library also.

Important Libraries we import for our project

- import pandas as pd
- import numpy as np
- import matplotlib
- import matplotlib.pyplot as plt
- import seaborn as sns
- from tabulate import tabulate
- import folium
- import plotly.express as px
- %matplotlib inline (It's a magic function



Data Summary

Given data set has different columns of variables crucial for hotel bookings. Some of them are:

hotel: The category of hotels, which are two resort hotel and city hotel.

is_cancelled : The value of column show the cancellation type. If the booking was cancelled or not. Values[0,1], where 0 indicates not cancelled.

lead_time : The time between reservation and actual arrival.

stayed_in_weekend_nights: The number of weekend nights stay per reservation

stayed_in_weekday_nights: The number of weekday nights stay per reservation.

meal: Meal preferences per reservation.[BB,FB,HB,SC,Undefined]

Country: The origin country of guest.

Data Summary(contd..)

market_segment: This column show how reservation was made and what is the purpose of reservation. Eg, corporate means corporate trip, TA for travel agency.

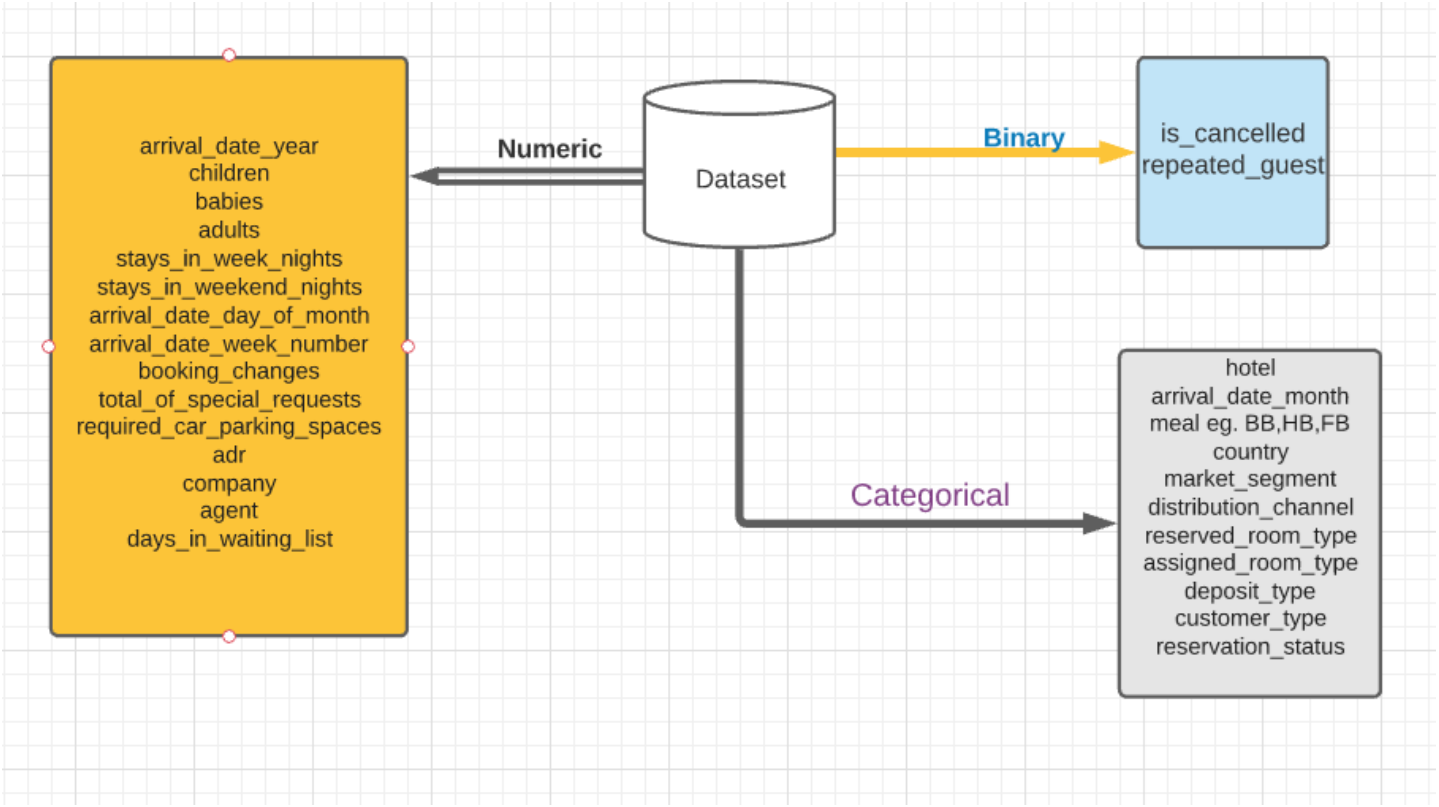
distribution_channel: The medium through booking was made.[Direct,Corporate,TA/TO,undefined,GDS.]

Is_repeated_guest: Shows if the guest is who has arrived earlier or not.Values[0,1]-->0 indicates no and 1 indicated yes person is repeated guest.

days_in_waiting_list: Number of days between actual booking and transact.

customer_type: Type of customers(Transient, group, etc.)

Data Summary

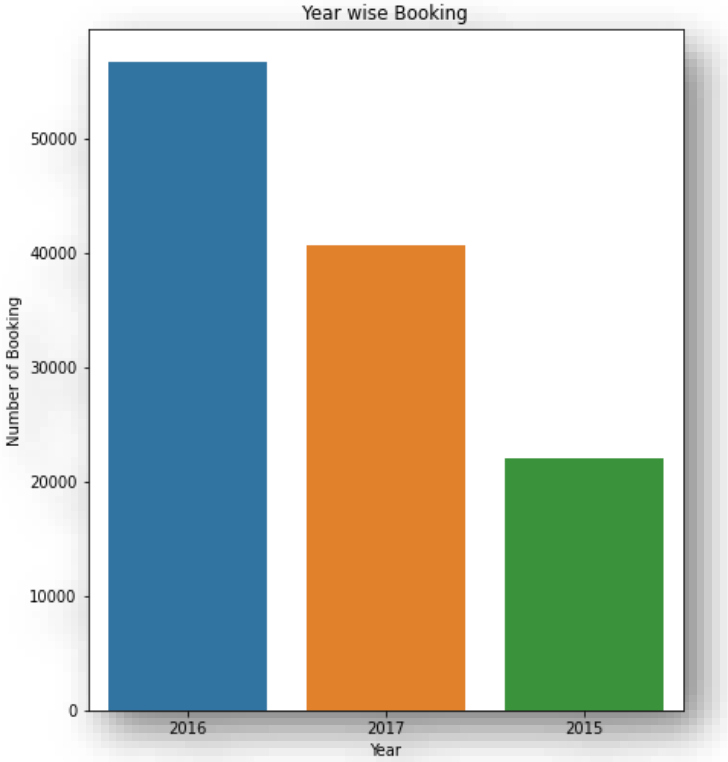


EDA----->(Exploratory_Data_Analysis)

1. **Which Year most room booking happen?**

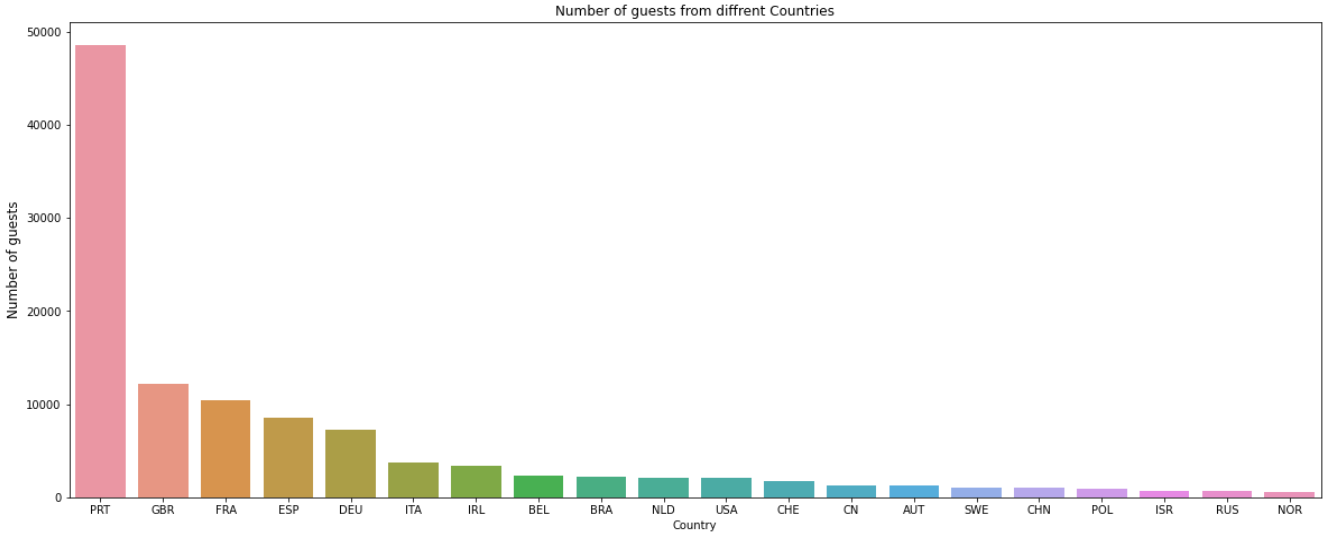
- In our data frame there are only 3 year data is present, so from the give data we can do a analysis on highest number of booking.

- This graph represent above analysis and giving information that in 2016 highest number(56,707) of room booking happen and in 2015 lowest number(21,996) of booking happen.
- From here we can say in the year 2016 room booking number is more than double as compare to in the year 2015 which is lowest among all.

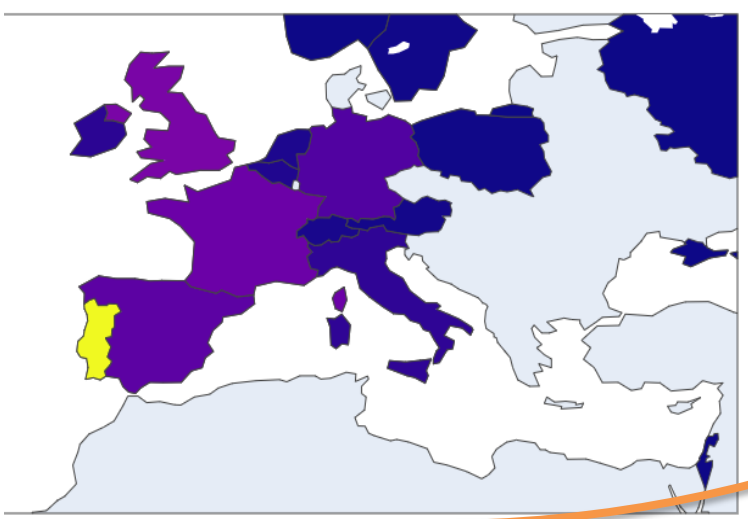


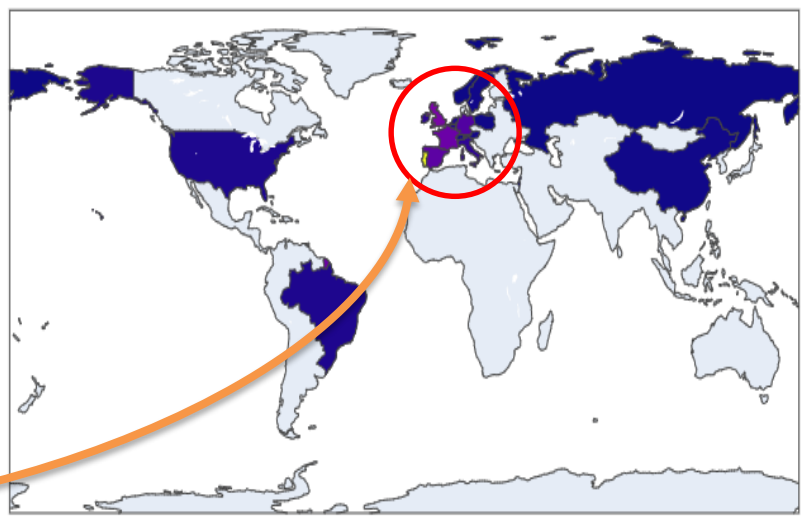
2. Country With Highest Number of Booking ?

- Form the above graph it is clearly showing that high number of booking happen in Portugal, Great Britain, France, Spain etc country.
- It show that as compare to any country Portugal has the highest booking number(48590). More than four times as compare to 2nd highest Great Britain(12129).



3. Map wise room booking density graph



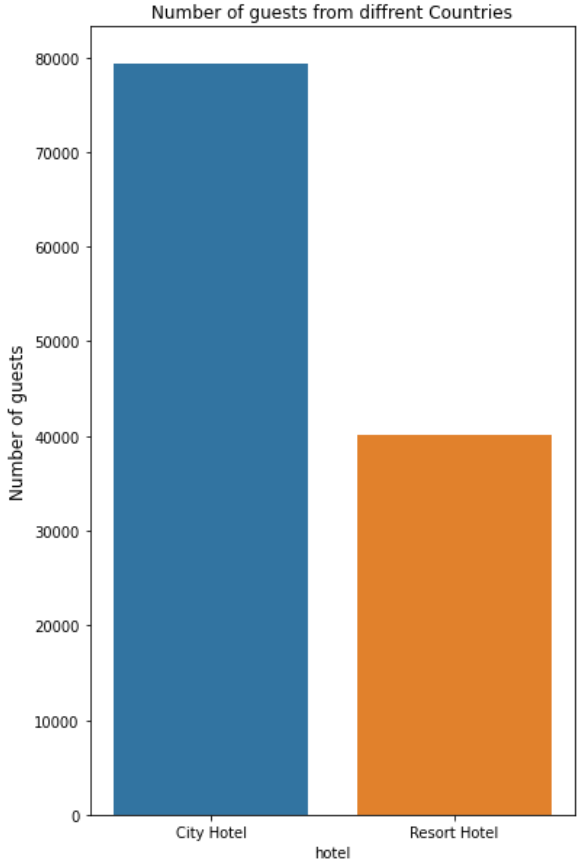


Map wise room booking density in Europe

Map wise room booking density in World Map

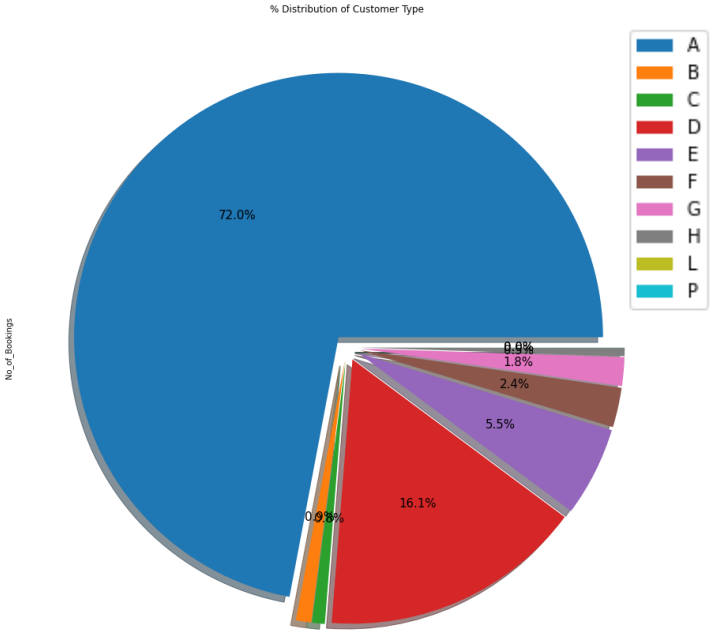
4. Which type of Hotel has highest number of booking ?

- Our Data frame has 2 types of hotel data is present, so from the give data we can do a analysis on highest number of booking in which type of hotel.
- This graph represent two type of hotel type like City Hotel and Resort Hotel.
- Here we can clearly get an idea that Number of guests in City Hotel are more as compare to Resort Hotel



5. Which room type booked in highest Number?

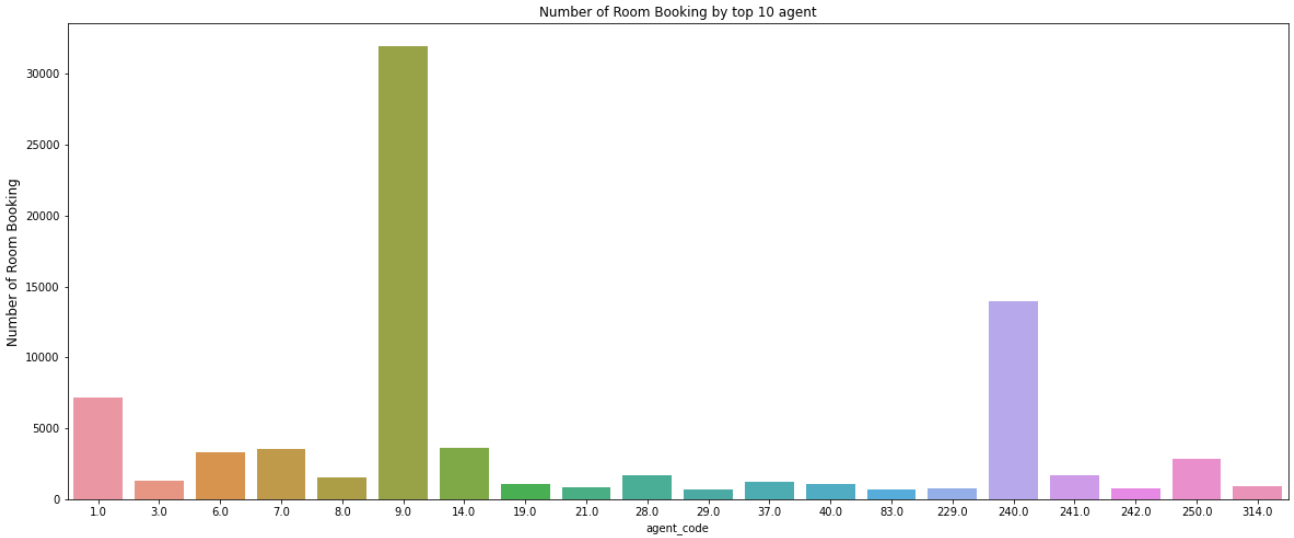
- There are 10 types of room are present in hotels.
- Out of that most room are from type '**A**' or type '**D**' types
- '**A**' type room has the highest reserved room booking type having 56552 number.
- Where as both '**L**' and '**P**' type room has lowest reserved room booking type having each 6 number of booking.



Id	reserved_room_type	No_of_Bookings
0	A	56552
1	B	999
2	C	915
3	D	17398
4	E	6049
5	F	2823
6	G	2052
7	H	596
8	L	6
9	P	6

6. Number of room booking by top 10 Agent

- In this analysis we finding top 10 agent having highest number of booking. Here in this graph, we can clearly see that mostly 3 agent have done most of the booking.
- Agent id- 9.0 have done highest number of booking.
- Mostly agent 1.0,9.0 and 240.0 have done most of the booking.

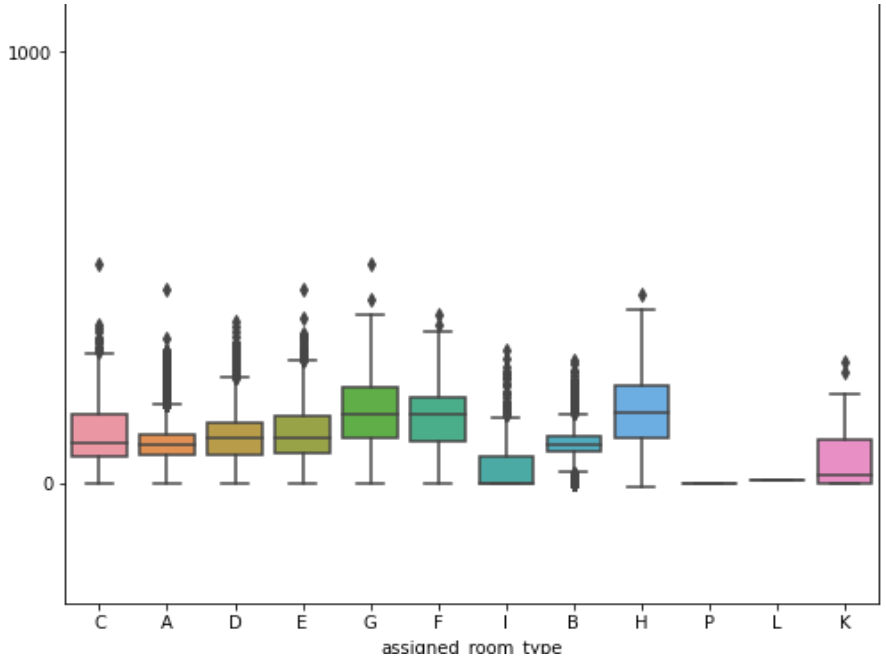


id	agent	No_of_Bookings
0	1	7191
2	3	1336
5	6	3290
6	7	3539
7	8	1514
8	9	31961
13	14	3640
17	19	1061
19	21	875
26	28	1666
27	29	683
35	37	1230
38	40	1039
72	83	696
168	229	786
173	240	13922
174	241	1721
175	242	780
182	250	2870
227	314	927

7. which room type generates highest adr?

- Most demanded room type is A, but better adr rooms are of type H, G and C also. Hotels should increase the no. of room types A and H to maximize revenue.

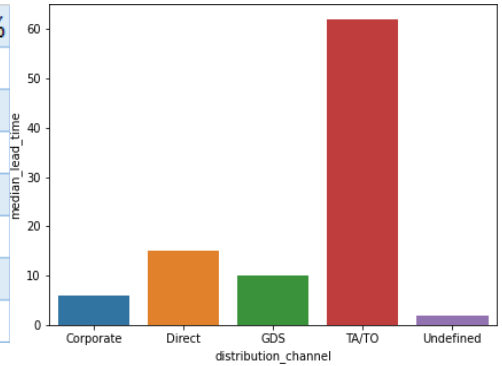
Id	reserved_room_type	No_of_Bookings
0	A	56552
1	B	999
2	C	915
3	D	17398
4	E	6049
5	F	2823
6	G	2052
7	H	596
8	L	6
9	P	6

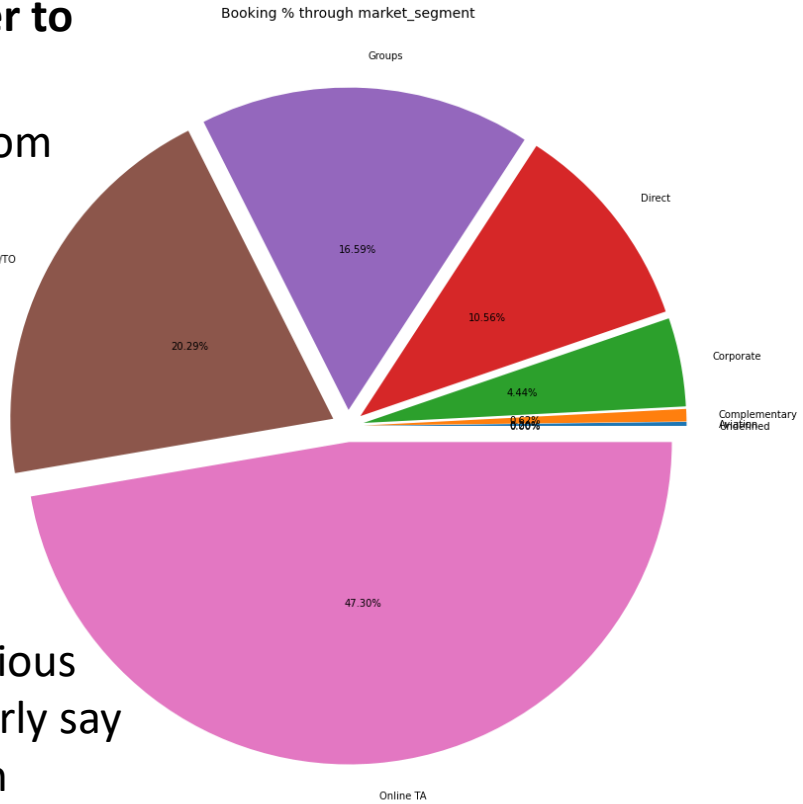


8. Find is the most common market segment prefer to booking hotels?

- In this data analysis here we get to know that from Offline TA/TO.

market_segment	Booking_%
Aviation	0.2
Complementary	4.44
Corporate	10.56
Direct	16.59
Groups	20.29
Offline TA/TO	47.3
Online TA	0





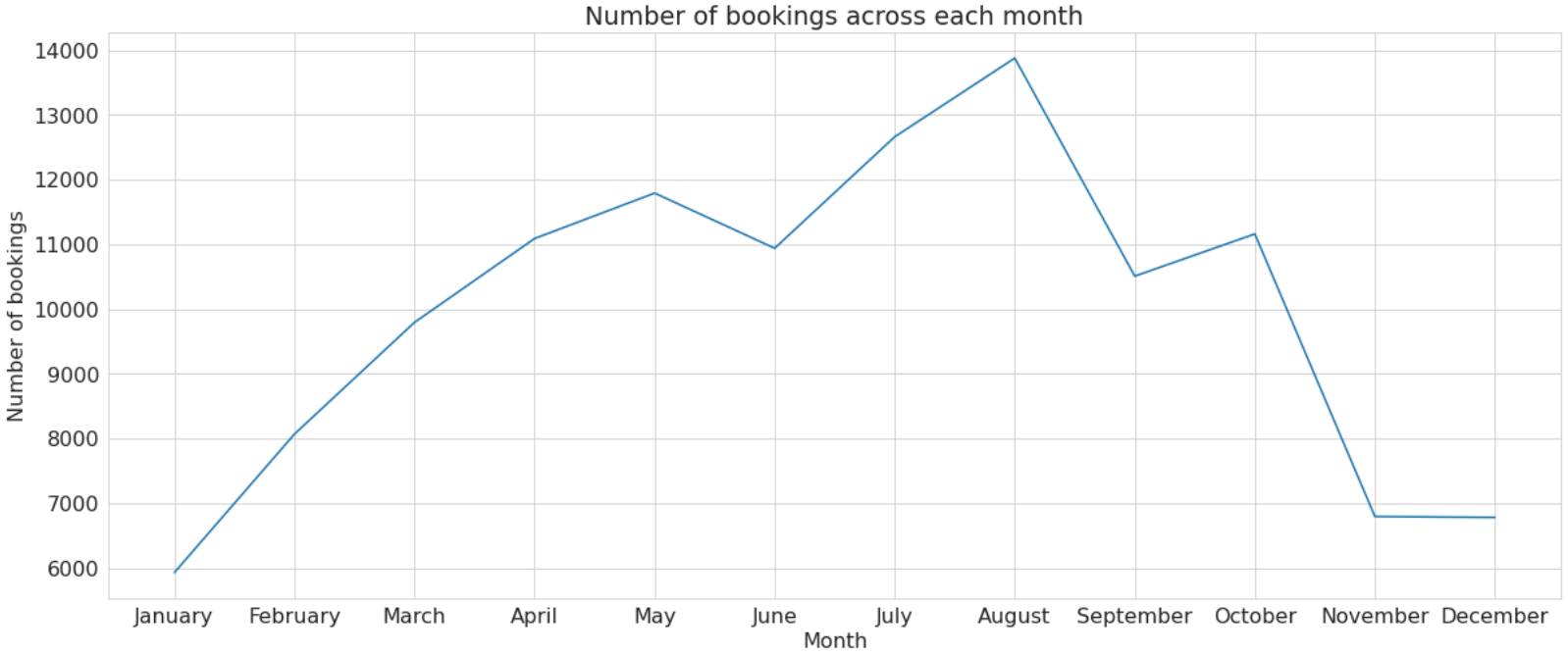
- Market segment is well distributed between various segment. Here in this dada analysis we can clearly say that aviation and online TA has very less share in booking hotel rooms.

9. In which month most of the bookings happened?

- In this table we can determine that in the month of august highest room booking happened .
- Also we can see up to may the demand for booking in increase steadily. In the month of June suddenly the trend go towards downward.
- From June to till august again a consistently demand increase in the room booking. After that mostly the tread goes downward trend.
- Lastly we can say towards end of the year the demand of booking is reduce drastically.

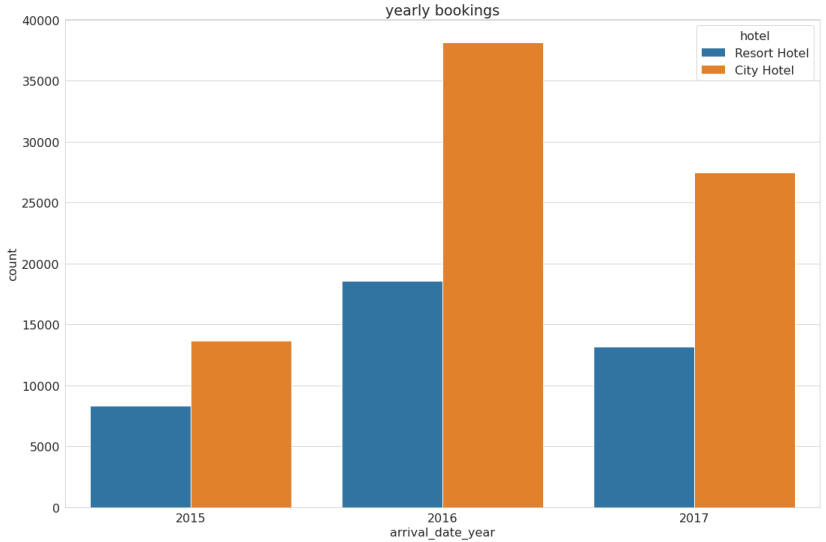
arrival_date_month	Counts
January	5929
February	8068
March	9794
April	11089
May	11791
June	10939
July	12661
August	13877
September	10508
October	11160
November	6794
December	6780

Area plot graph for month most of the bookings happened?



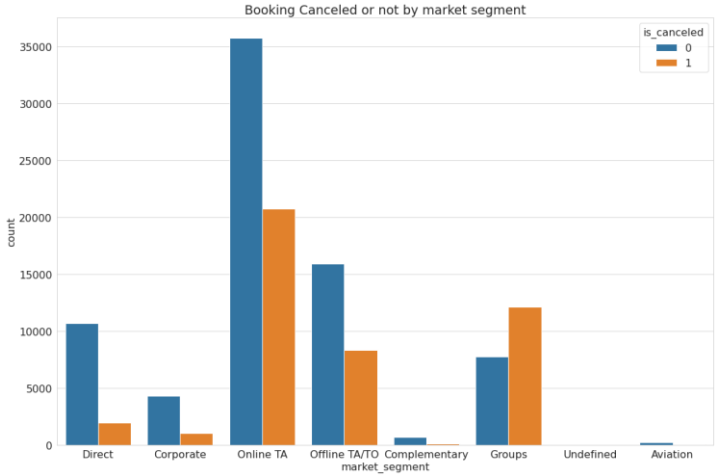
10) Visualizing Hotel wise yearly bookings

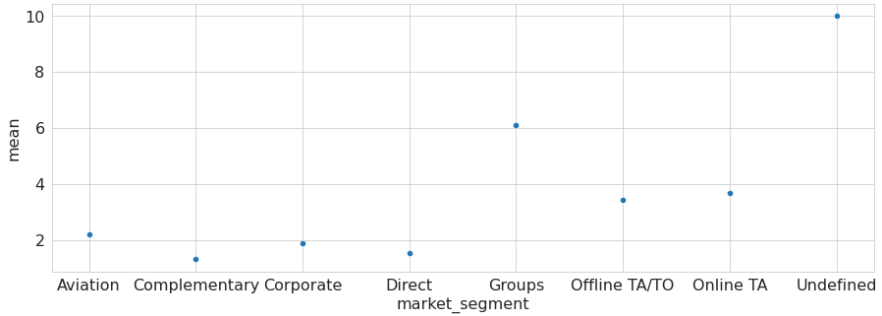
- This bar count plot graph show the yearly booking of room with respect to hotel wise data.
- The data show in this slide show that the booking trend go from low to high in the year 2016 to 2017.and then again slowly decrease in the year 2017.
- Both hotel type Resort hotel and city Hotel has same type of trend but city hotel has more demand as compare to resort hotel



11) What is the relationship between market segment and cancellation?

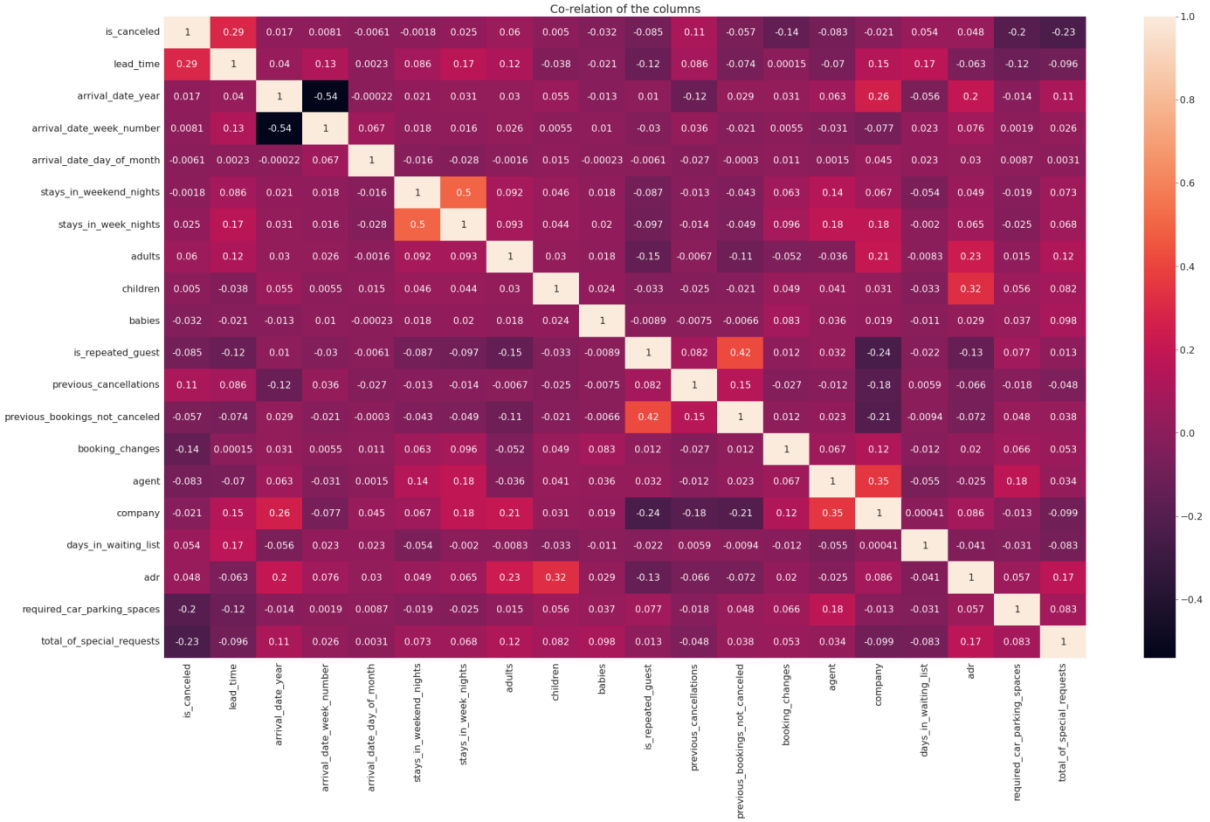
- This count plot graph show the relation between market segment and booking trend.
- This graph show that in Online TA high number of booking as well as high number of cancelation happen.
- One of the interesting key note is complementary and aviation booking rooms are least chances of cancellation.
- High cancellation happen when booking happen in group. Because high chance is there plan for tours not executed as per schedule.



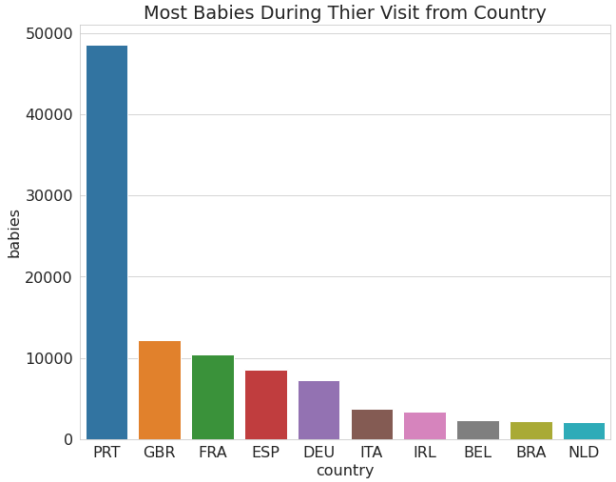


12) Correlation of the columns

- This Heat map show the relation between the various column of the data given
- The graph show the relation of each column with each other either strongly bonded or weakly bonded.
- If they strongly bonded then it show light white color where as if it is very dark then it show loosely bonded to each other.



13) Which top 10 Country have most babies during their visit?

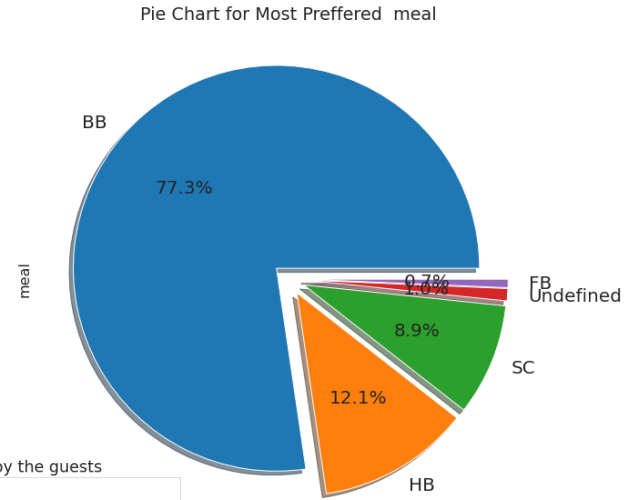


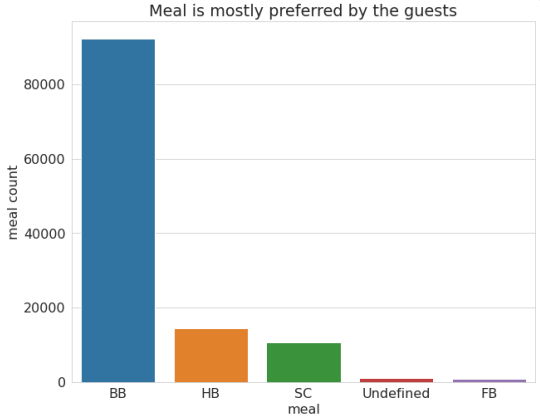
Country	Babies
PRT	48590
GBR	12129
FRA	10415
ESP	8568
DEU	7287
ITA	3766
IRL	3375
BEL	2342
BRA	2224
NLD	2104

- In the above chart and graph we can clearly visualize that Portugal is country where highest number of visitor book hotel room similarly this trend continue with most number of babies visit as well.
- So mostly western Europe country's babies visit hotel rooms with their parents mostly.

14) Which type of Meal is mostly preferred by the guests during their visit?

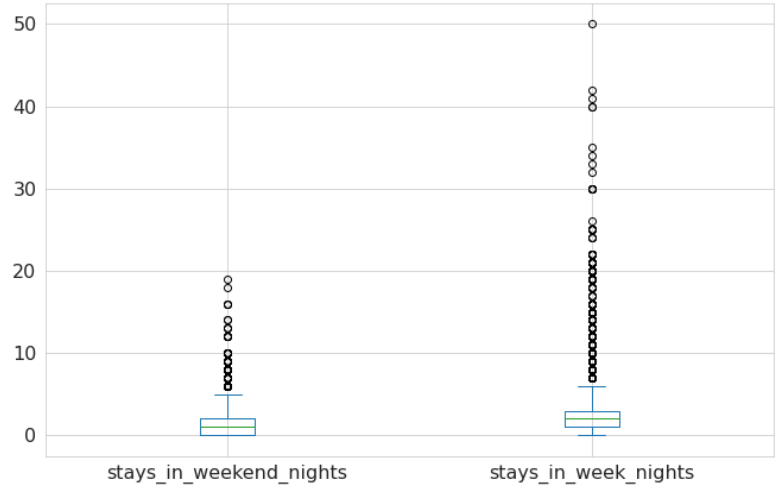
- In the above pie chart we can easily determine that ‘BB’ meal is most preferred by the visitor.
- Mostly more than ¾ of visitor preferred ‘BB’ meal type.
- Generally the visitor prefer either ‘BB’/’HB’/’SC’ type of meal.
- ‘FB’ is the least prefer meal type.



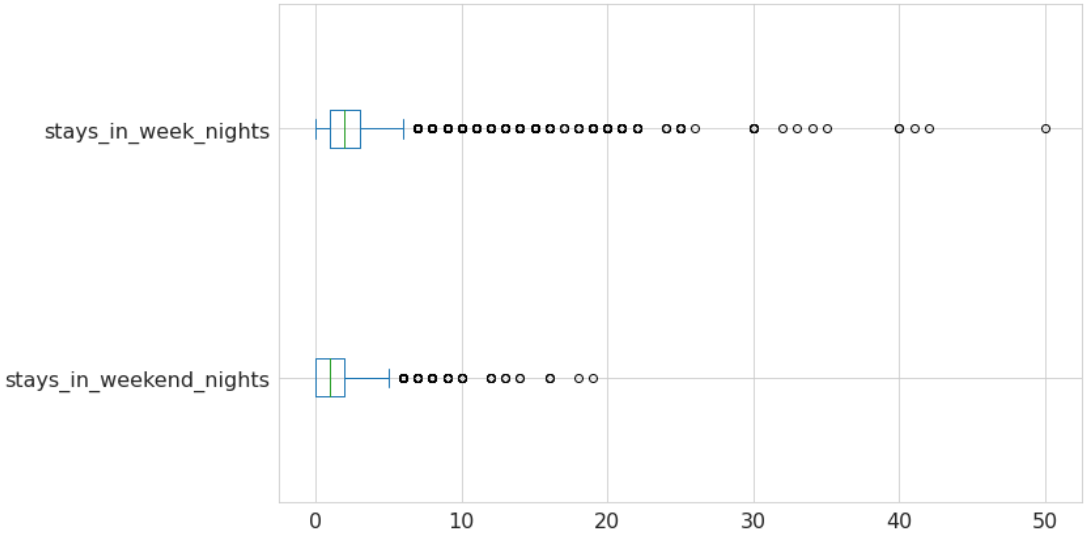


15) Draw the boxplots of the two columns of stays in weekend nights and stays in week nights in a single plot using the following one-liner code. The method we use is plot box.

- **From the Graph It is clear that customer stay in week nights is higher than weekend night.**



15) Draw the boxplots of the two columns of stays in weekend nights and stays in week nights in a single plot using the following one-liner code. The method we use is plot box.



Some important questions

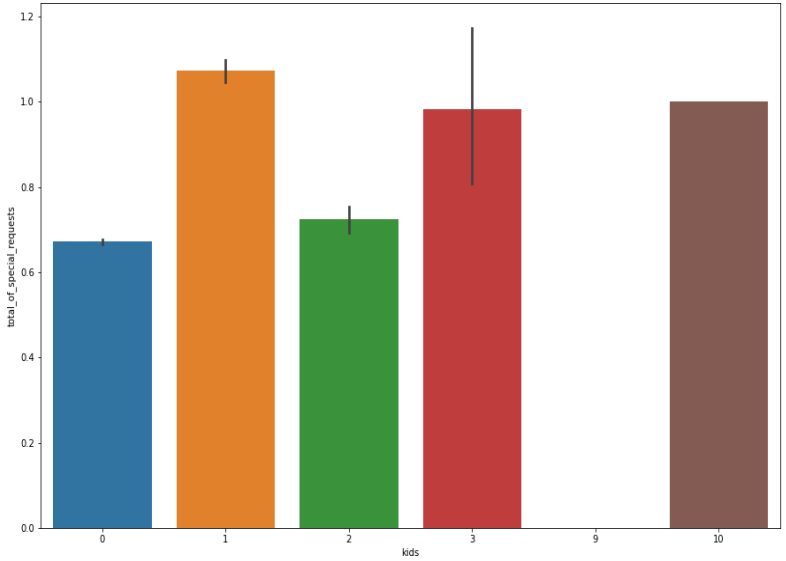
Some other analysis are also done, which are as follows:

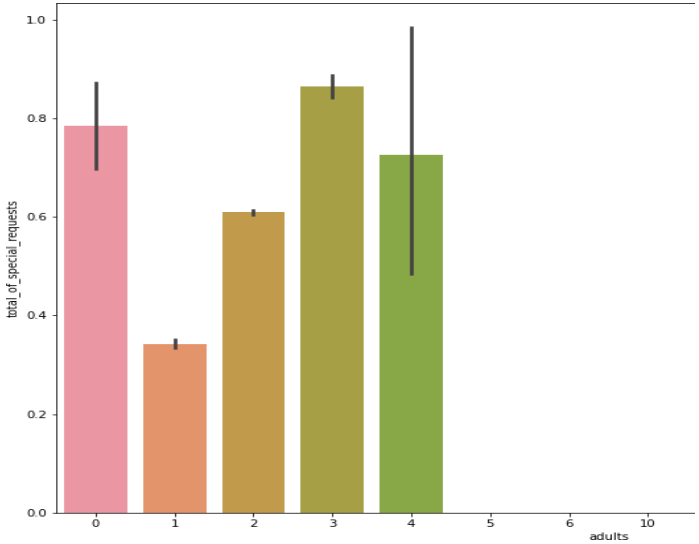
(1) What are the different reason for special requests

(2) What is the optimal stay length for better deal for customers

(3) How adr is affected by total staying period in hotels.

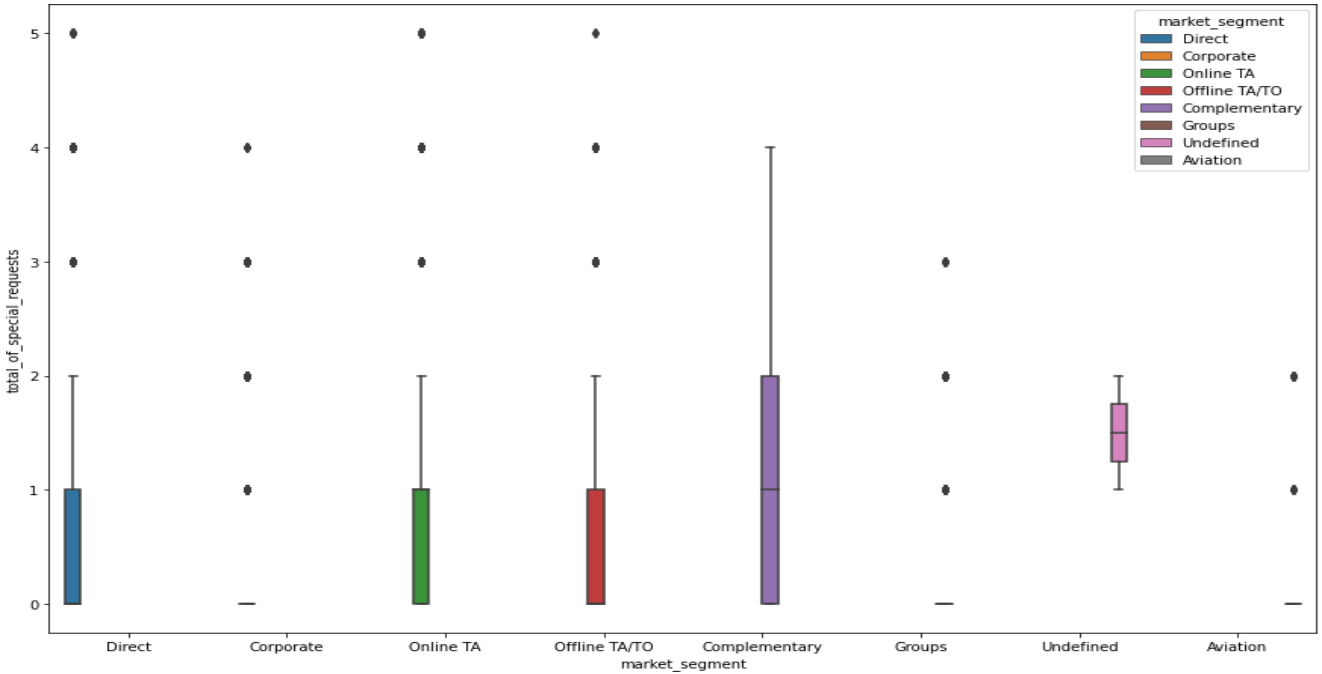
(1) What are the different reason for special requests





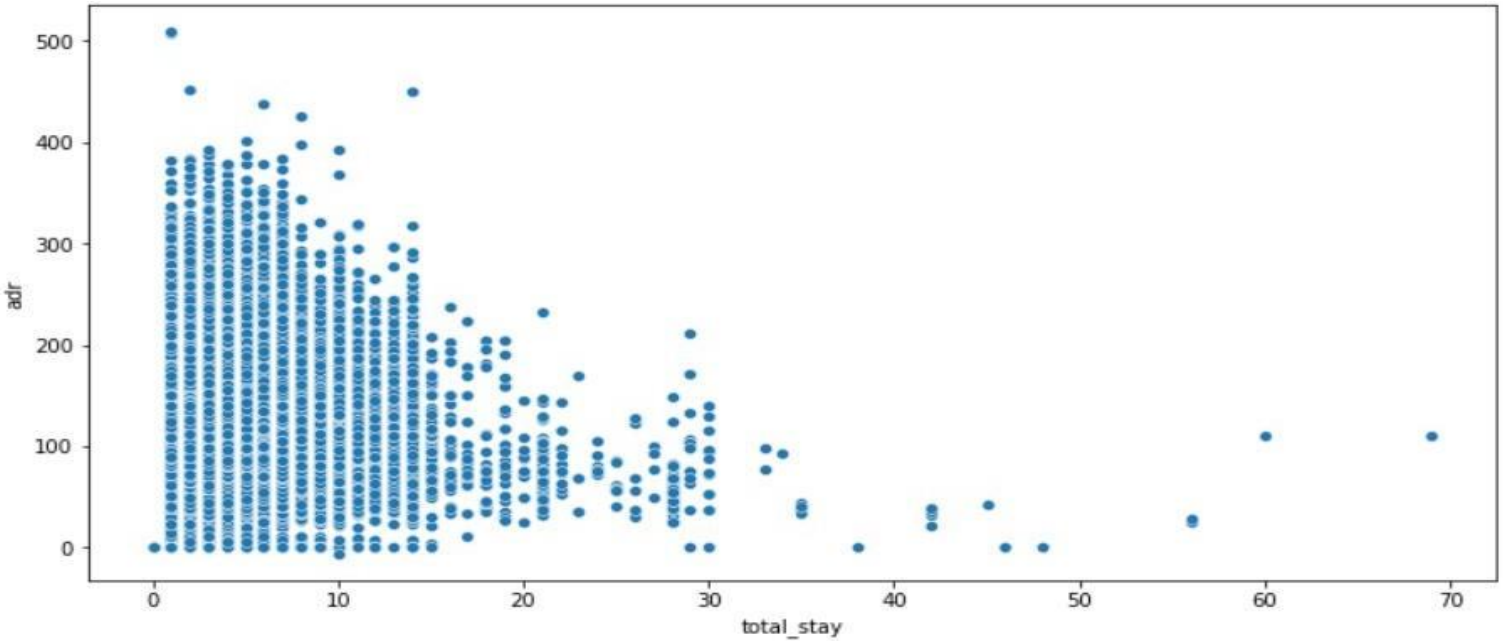
- The number of special request are almost the same in the kids section. But, we can see that if the adults are more than 2, there are more chances that hotels will receive more special requests.

Reasons for special requests(cont.)



- Here we can see that all market segment mostly have special request.
- There is one segment which is complementary, having more than average number of special request.

Optimal stay length for better deals in average daily rate adr(average daily rate)



For shorter stays the adr(average daily rate varies greatly) but for longer stays (> 15 days) adr is comparatively very less. Therefore, customers can get better deal for longer stays more than 15 days.

Conclusion

(1) Customer visiting from Country Portugal has most number of Babies.
(2) Most of Booking done in year 2016 i.e. 56,707 number of booking.
(3) Meal is mostly preferred by the guests during their visit is BB type which is 77.8% of all type of meal.
(4) Most number of booking coming from Country Portugal .
(5) Most number of booking for stays_in_week_nights.
(6) High number of booking happens in Western Europe country.
(7) Most number of booking done in month of AUGUST.
(8) City Hotel has highest number of booking i.e. 79330 numbers.
(9) TA type of market-segment has most number of cancellations.
(10) 'A 'Type room has highest number of booking i.e. 72%.
(11) Agent no-9 is most valuable agent.
(12) Most demanded room type is A, but better adr rooms are of type H, G and C also.
Hotels should increase the no. of room types A and H to maximize revenue
(13) Bookings made via complementary market segment and adults have on average high no. of special request.
(14) For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr(average daily rate).


Thank you