Capstone Project - 1
Hotel Booking Analysis

Team members:

Minaz Uddin R.

Kiran Mamtani

Content

- Summary of Data
- Data wrangling
- Data Analysis
- Data visualization
- Challenges
- Conclusion

Summary of Data

About Data - This Database consists of information between 2 different hotels over 3 years (2015 to 2017)

Size of Data:

- **Rows -** 119390
- **Columns -** 32

Source: https://www.freepik.com/

Data wrangling

Data cleaning :

- This Database consists of null values ‘Nan’.
- Replaced those null values with mean and string in two different columns.
- Removed two columns

Data preparation :

- Database consist of different types of data types.
- Data types : integer, float, object

Data Analysis

Below are some questions came to our mind before analysis and we thought to include them…

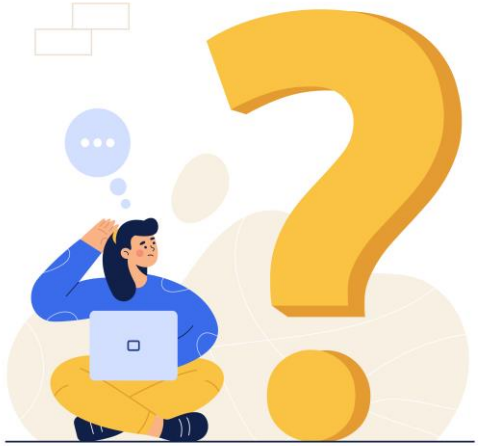
1. Which are the type of hotels booked?
2. Which type of customer are booking the hotels?
3. From which medium the maximum booking are happening?
4. Find the rate of cancellation of hotels.
5. Find the month and year that has maximum booking. Compared it with hotel type and hotel canceled.
6. Find the top 10 countries that has maximum bookings.

Continues…

Data Analysis

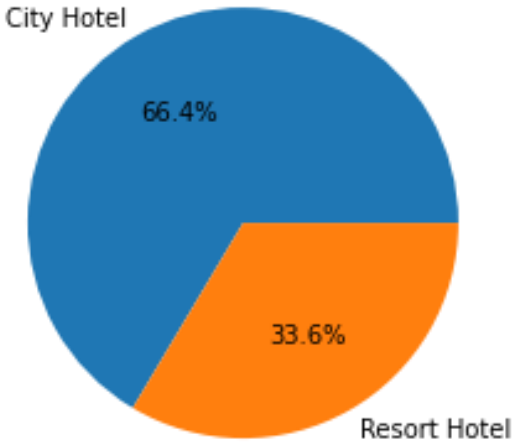
7. Find the relationship between room change and cancellations.
8. What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special request ?
9. What kind of people book the most hotel ?
10. For how long the people stay in hotel ?
11. Percentage of repeated guest in hotel ?
12. Analyze average ADR for given statement.

a) Avg ADR for each year
b) Avg ADR for resort and hotels
c) Avg ADR for each country
d) Avg ADR of top 10 countries who has highest number of bookings



Data Analysis and Visualization

1. Which are the types of hotels booked ?



Here we can observe that, out of 100%, City hotels are more booked (66.4%) compared to resort hotels (33.6%)

2. Which type of customers is booking hotels?

Customer type	Count	Customer type %
Transient	89613	75.059050
Transient-party	25124	21.043638
Contract	4076	3.414021
Group	577	0.483290

So the maximum customer type is transient (75%). It means that these are the customers who are living for the short period of time.

3. From which medium the maximum booking had happened ?

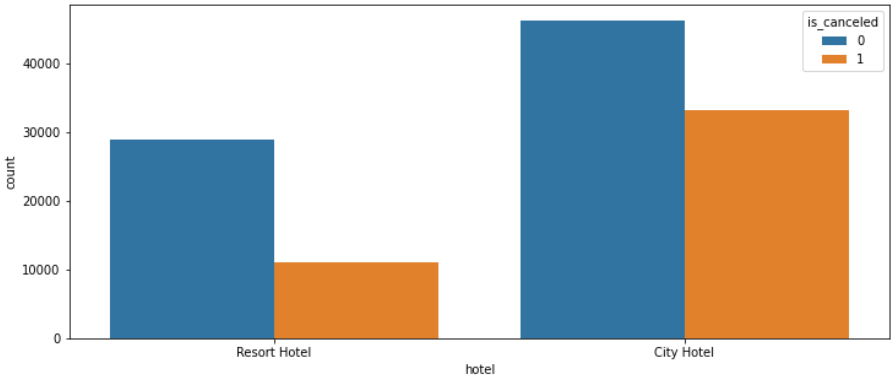
Market segment	Market segment counts	Market segment percentage
Online TA	56477	47.304632
Offline TA/TO	24219	20.285619
Groups	19811	16.593517
Direct	12606	10.558673
Corporate	5295	4.435045
Complementary	743	0.622330
Aviation	237	0.198509
Undefined	2	0.001675

So 47% of the customers are booking hotels online

4. Find the rate of cancellation of hotels?

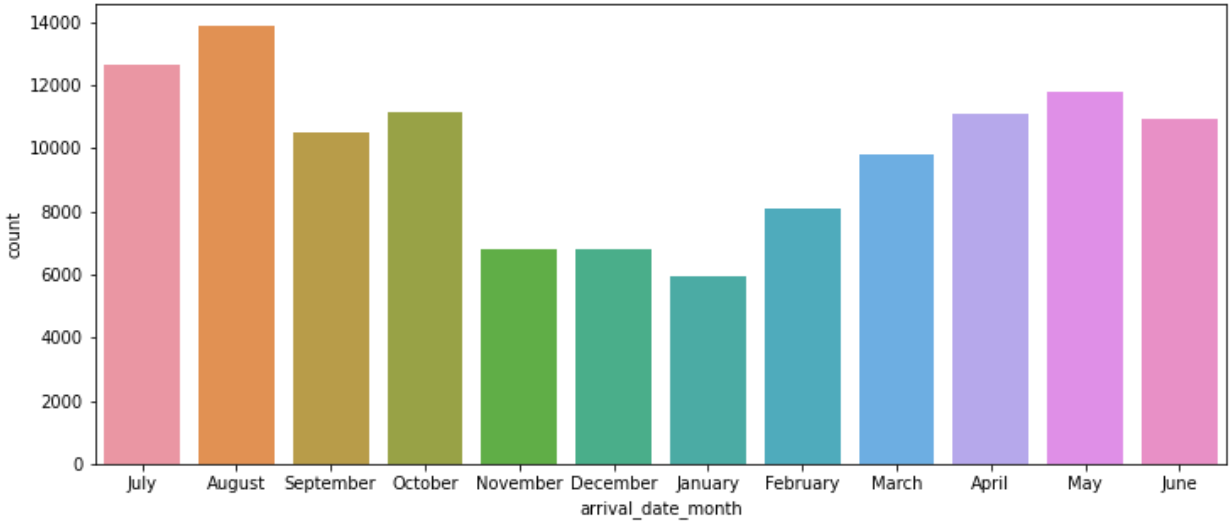
Canceled	Counts	Percentage
0 (Not canceled)	75166	62.95
1 (Canceled)	44224	37.04

So 37% of total booking got canceled



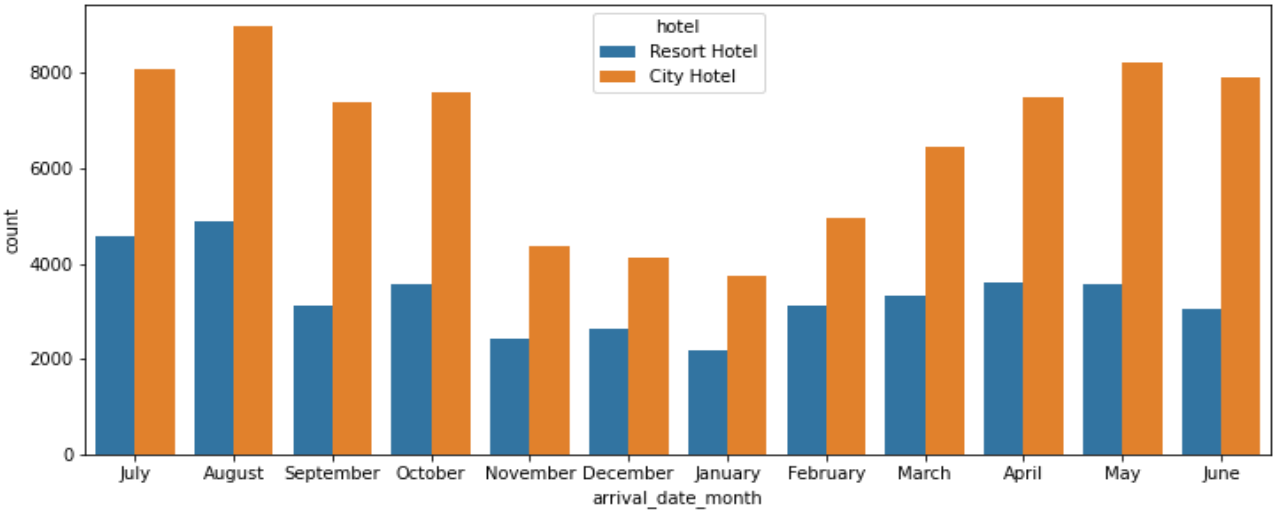
When compared with City and Resort hotels. City hotel cancellation is 42% and Resort hotel cancellation is 28%

5. Find the month and year that has maximum booking? Compare it with hotel type and hotel canceled.



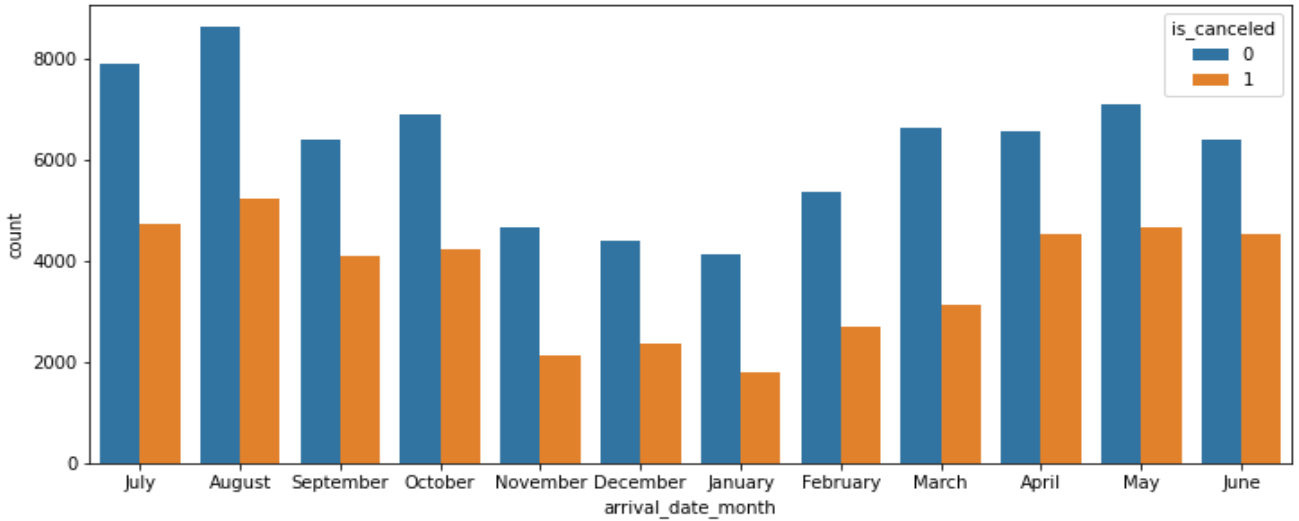
So August has highest number of bookings. But this include both resort and city hotels

Lets see resort and city hotel differently



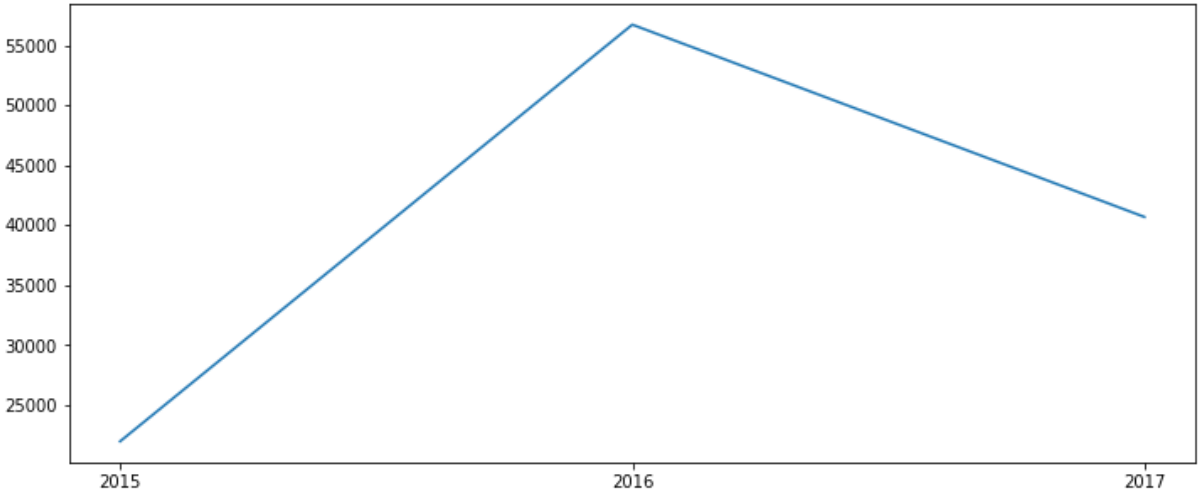
So resort and city both hotels has highest number of bookings in August month

To find the number of cancellations on monthly basis



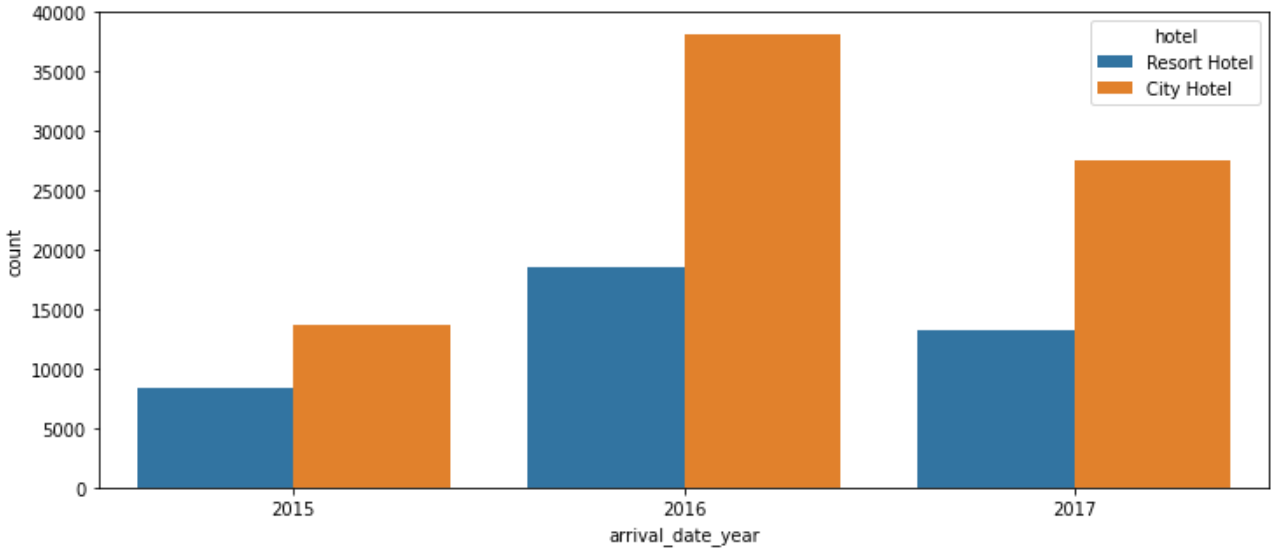
So when compared to all the months ,August month has maximum number of cancellations followed by July and may month.

Now lets see which year has the highest number bookings



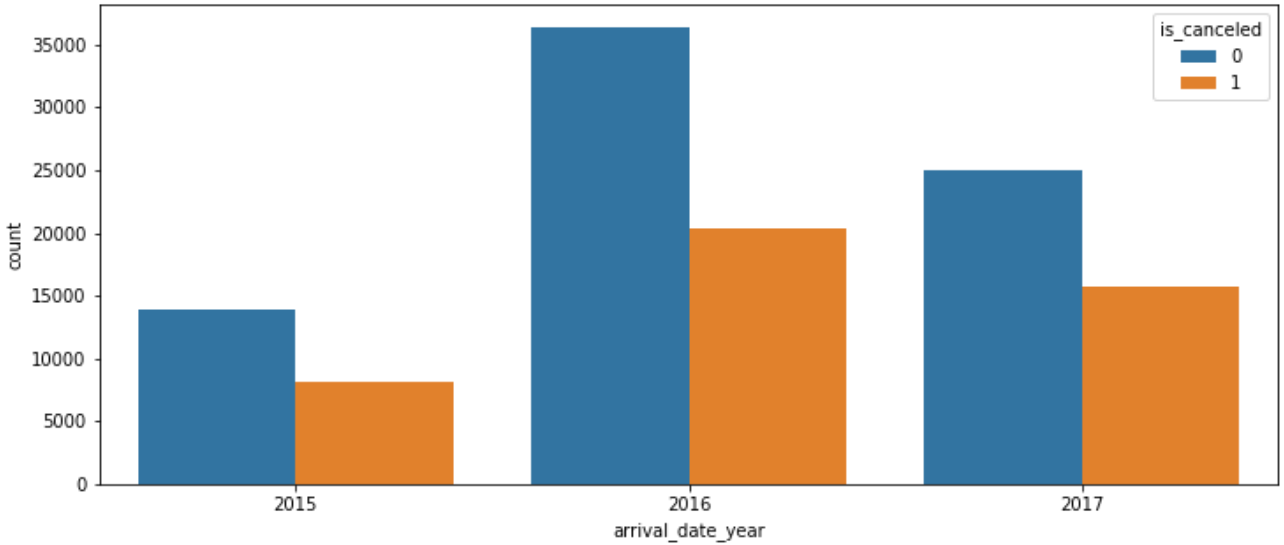
Bookings has increased in 2016 than decreased in 2017. And max booking was in 2016 which include both resort and city hotels

Now lets see resort and city hotels differently in each year



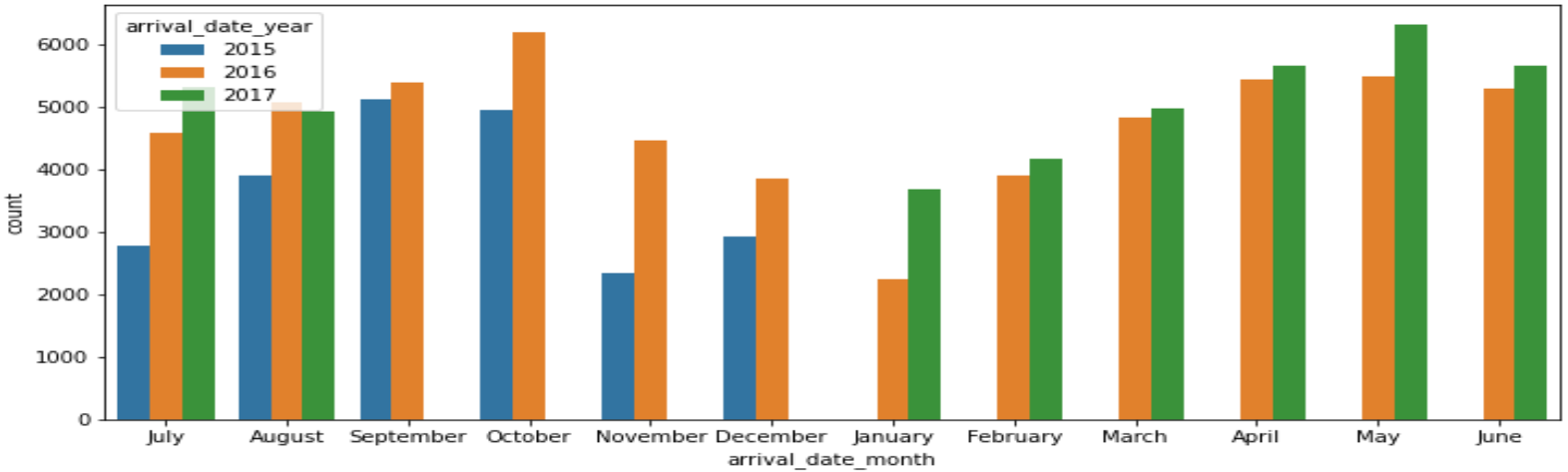
So resort and city both hotels has highest number of bookings in 2016

Lets see which year has maximum cancellation



We could find that maximum number of booking are been cancelled in the year 2016

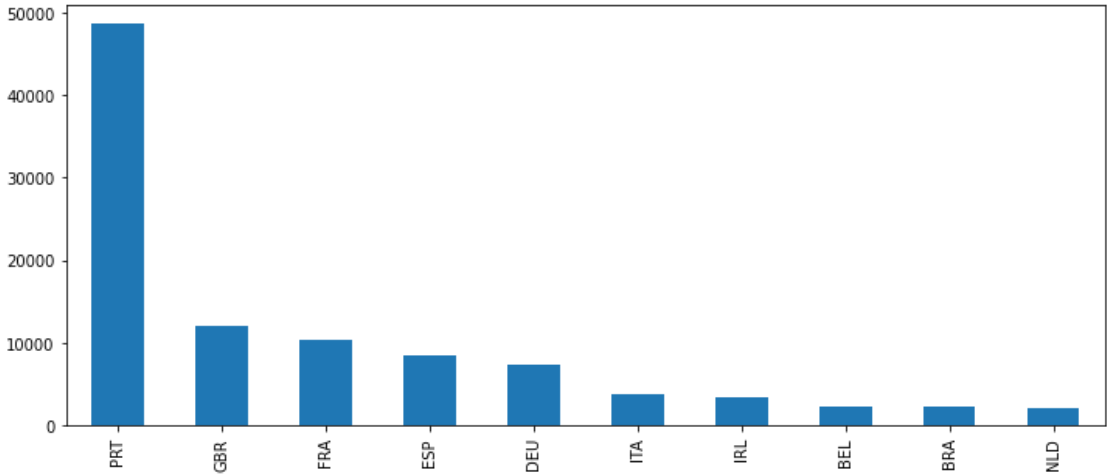
Now lets see year wise booking of each month



- As we can see in above graph, we have data from July 2015 till August 2017.
- Earlier we have seen August is the month who has maximum booking but that was cumulative result of all three years. August is repeating in every year.
- Here in year 2015, the maximum booking was in September. In year 2016, it is October and in year 2017, it is May.

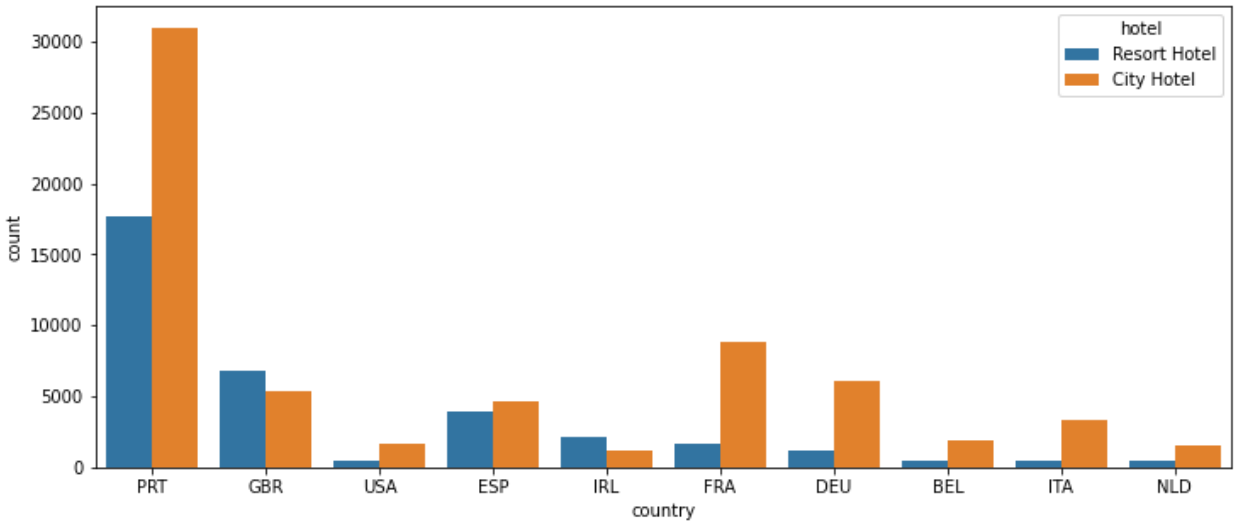
- In none of the year we have got August.
- So we can't say August has maximum booking in each year, it has maximum booking but in the given period of time in our data base.
- If our data was from January to December or from July 2015 till June 2017 then we can get accurate number in the outcome

6. Find the top 10 countries that had maximum bookings?



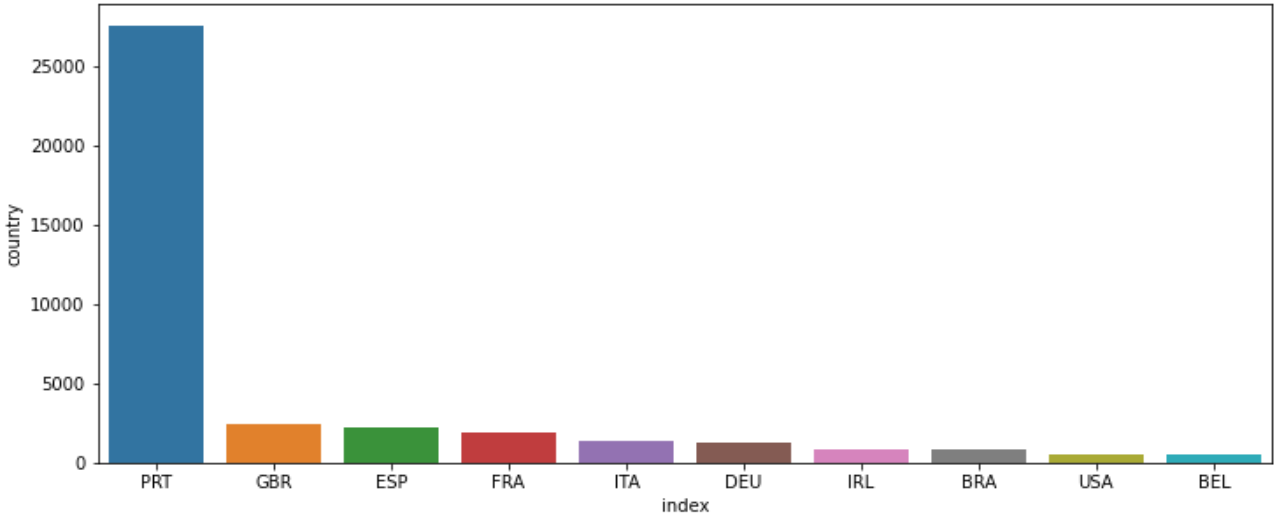
Portugal has highest bookings followed by Germany and France

Lets compare country wise bookings with hotel types



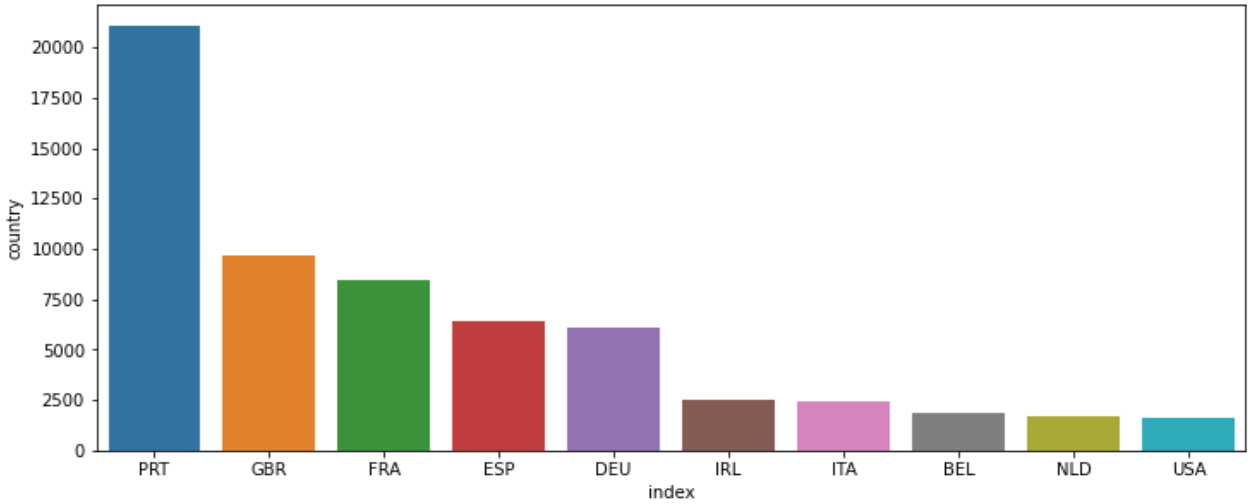
- **Portugal has max booking in resort as well as city hotel.**
- **The 2nd highest booking is in Germany for both resort hotel & city hotels.**
- **The 3rd highest booking for resort hotel is in Spain and for city hotel is in France.**

Now lets see which countries has highest cancellation



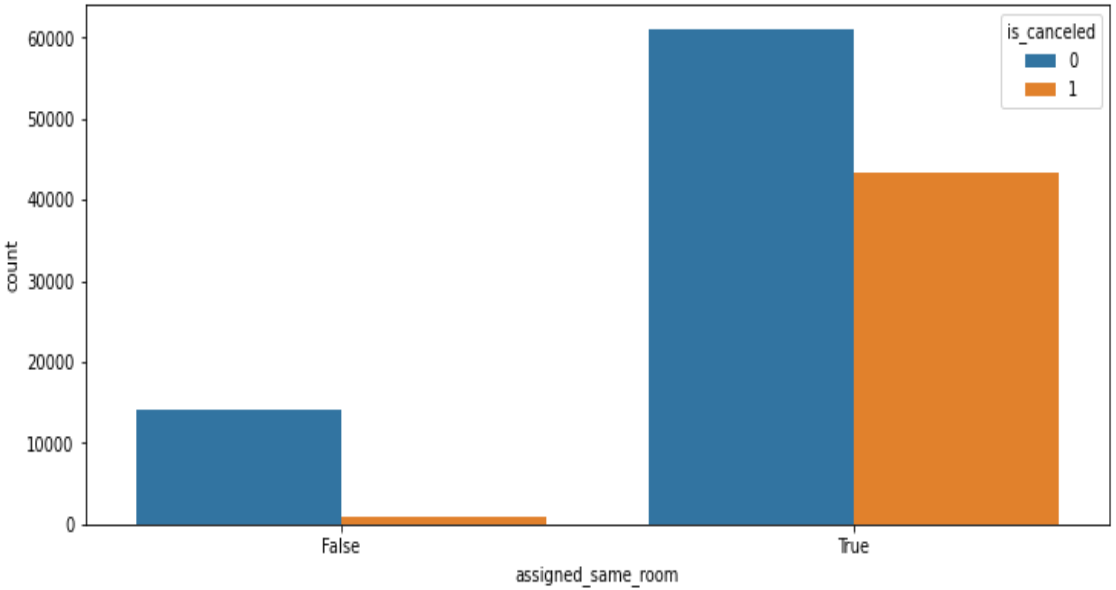
Portugal has highest cancellation followed by Germany and Spain

Now lets see the actual people stayed and enjoyed their holidays



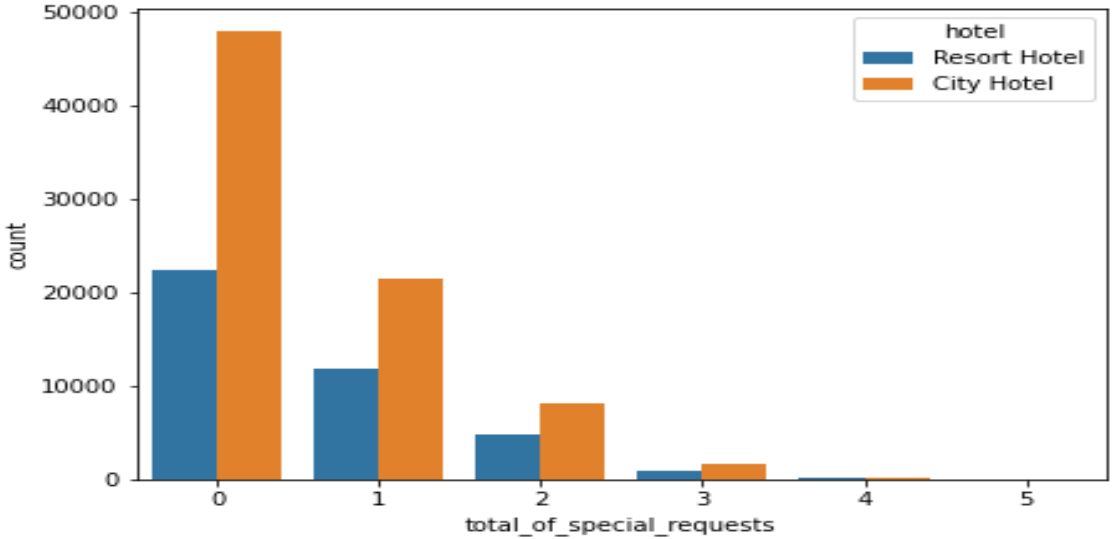
So after looking at the actual booking data, USA has taken place in top 10 and BRA (Brazil) got removed from top 10

7. Find the relationship between room change and cancellation?



Customer who has not changed room type has cancellation more and 5.7% of people cancelled room after they had assigned different room from booking room.

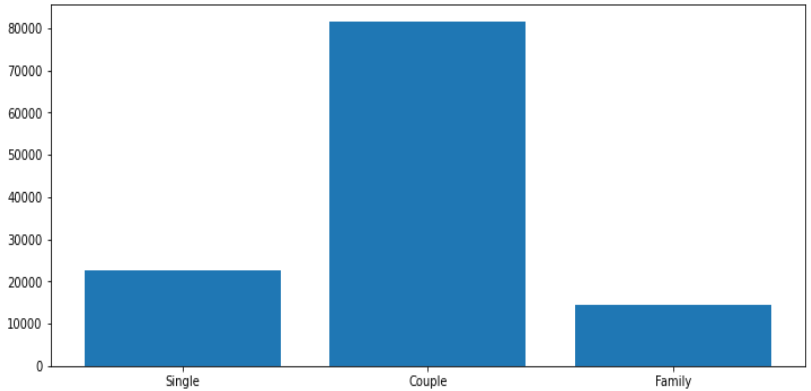
8. What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?



So the probability of 0 special request is highest

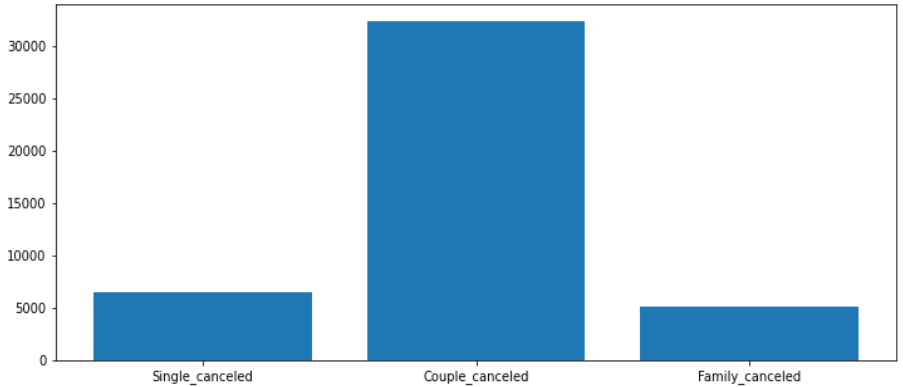
9. What Kind of People Book the Most Hotels?

Type of people has booked more



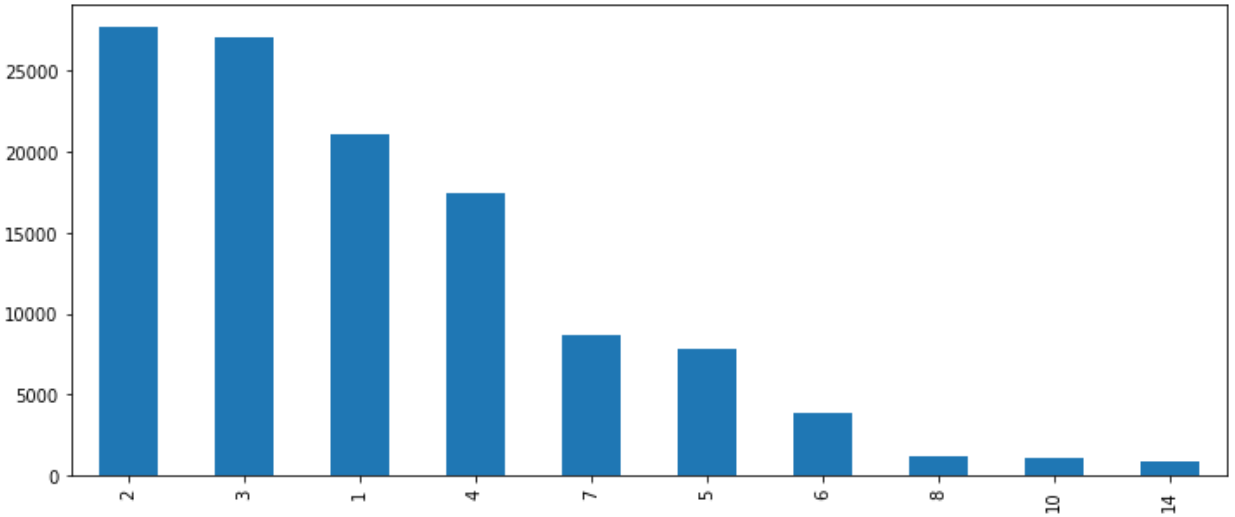
So as it is seen max booking was done by couples

Type of people has canceled more



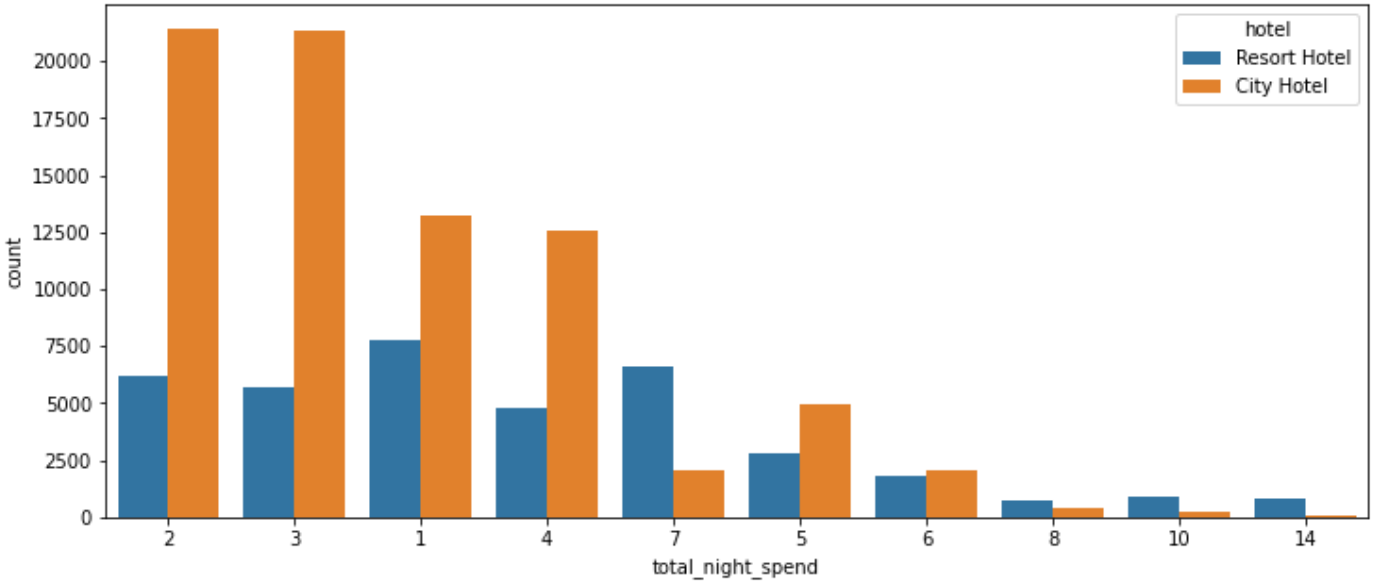
So couples has canceled maximum booking followed by single then family respectively

10. For how long people stay in hotel?



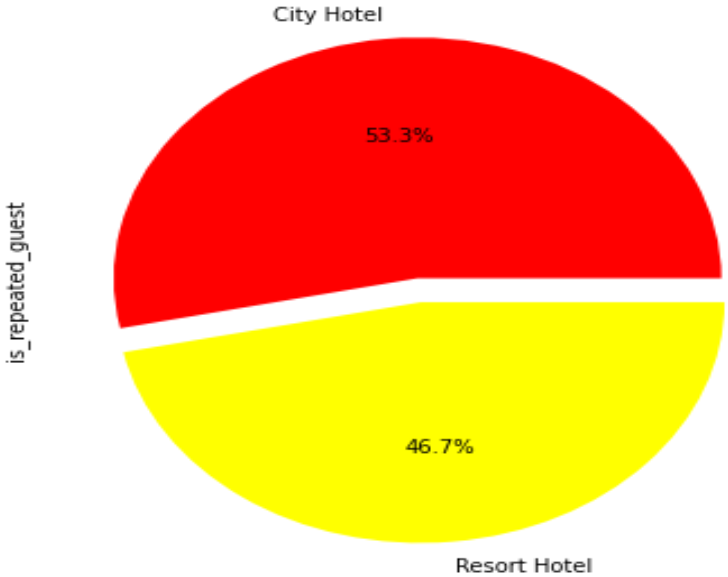
So most of people staying in hotel for 2 days follow by 3 days and 1 day.

Now lets compare resort and city hotel stays...



For city hotel most popular stay in 2,3,1 days respectively.
And for resort its 1,7,2 days respectively

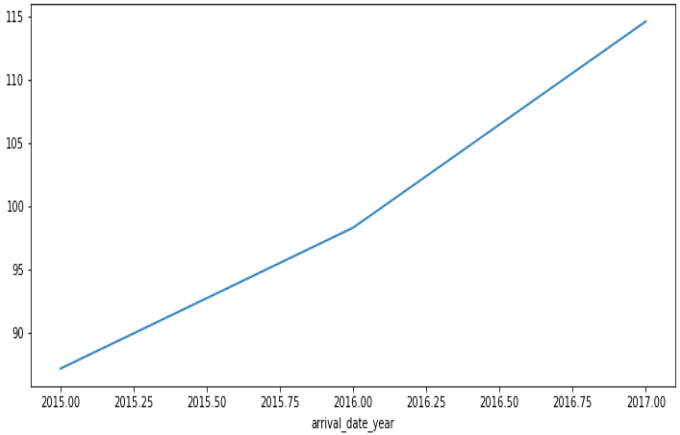
11. Percentage of Repeated Guests in hotels.



In city hotel 53.3% people are repeated guest while in resort the percentage is 46.7%.

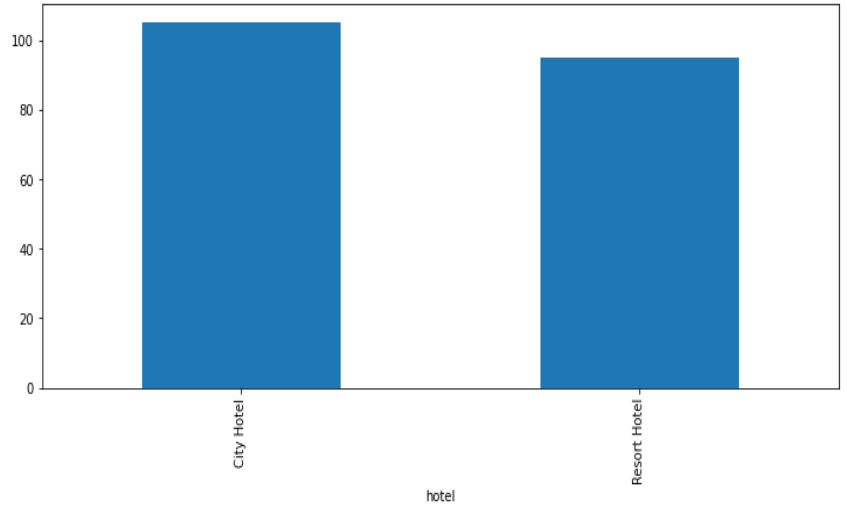
12. Calculate average ADR for given statements ?

a) Avg ADR for each year



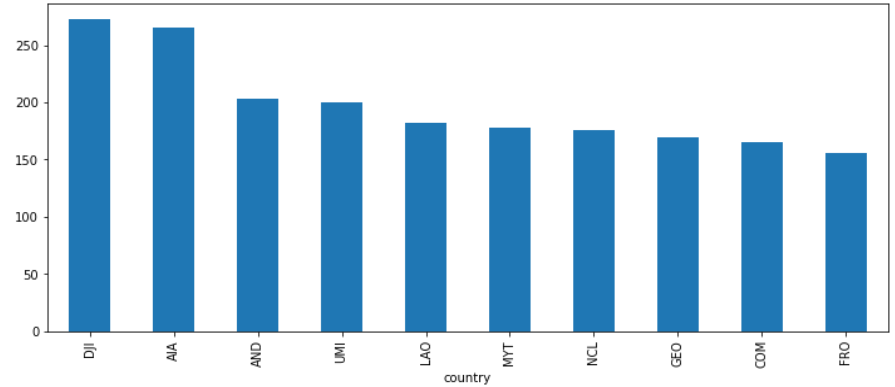
As we can see in above graph the avg ADR is gradually increasing from 2015 to 2017. So in 2017 the avg ADR is max

b) Avg ADR for hotel and resort



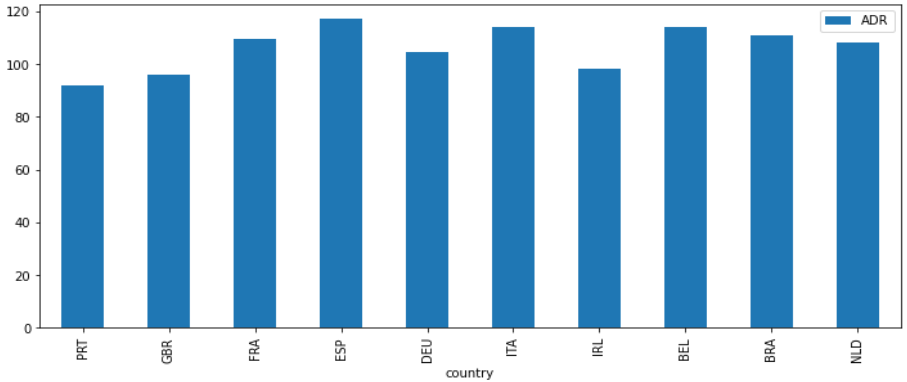
As the ADR between city hotel and resort are the Slight change. when compare to city hotel is more

c) Avg ADR for each country



So the highest ADR is of Djibouti (DJI) . But the above 10 countries are not our top 10 countries who has highest booking. So lets see our top 10 countries with highest booking and their ADR

d) Avg ADR of top 10 countries who has highest number of bookings



So as we know Portugal has highes t booking and its ADR is 92. And th e highest ADR from top ten countr ies is of Spain i.e. 116.99 ~ 117

Conclusion

- In our analysis of hotel booking we have tried different methods and visualized it by graphs.
- We started with asking few question to ourself and drafted the same to get the best solution.
- The first and most important step was cleaning the data. And then our analysis of data started.

Here is the conclusion of all the solution we got from our analysis:

- People are preferring city hotel booking more than the resort hotels.
- The preferred booking mode is online booking
- People preferred to stay for shorter period of time. For city hotel the popular stay is 2,3,1 days respectively and for resort hotel it is 1,7,2 days respectively.
- Couples books as well as cancels more hotel than single and family.
- 37% of people are cancelling their bookings (City & Resort). Resort hotel has less cancellation rate (28%) in compare to city hotels (42%)

Continues…

Conclusion

Now our data is from July 2015 till August 2017. But we have analyzed for year and months so it will not give us accurate results.

- 2016 has highest booking in both hotel & resort
- August has highest number of booking in both hotel & resort
- In year 2015, the maximum booking was in September. In year 2016, it was October and in year 2017, it was May
- Also, 2016 and August has highest number of cancellation
- The top three countries who has highest number of bookings are Portugal, Germany and France
- If we analyze hotel wise: Portugal has max booking in resort as well as city hotel. 2nd highest booking is in Germany. The 3rd highest booking for resort hotel is Spain and for city hotel is France.

Continues…

Conclusion

- Portugal, Germany and Spain has maximum cancellation
- We analyzed relationship between room changed and cancellation: We conclude that Customer who has not changed room type has cancellation more
- 5.7% of people cancelled room after they had assigned different room from booking room.
- The highest probability of 0 special request is highest
- The highest average ADR is for 2017 i.e. 114.63
- The average ADR for city hotel is 105.30 and for resort hotel is 94.95
- The highest average ADR is for Djibouti (DJI) country i.e. 273.
- And the highest average ADR for Portugal is 92.04

Summation

• Q2 & Q3 are having higher number of Bookings So hotels can avail some discount offers in Q1 and Q4 as Bookings rate is less in these quarters.

• Hotels can promote their business online as most of the Bookings are coming from online portals.

• Special discount can be given to repeated guests.

• City Hotel can come up with special plans for long staying guests.

• Resort Hotel can come up with special plans for short staying guests

Challenges

• Most difficult part was to decide how to clean data. Should we remove the columns? Or should we replace it with mean? Or should we replace it with 0.

• Second was should we remove all the canceled rows then continue with the analysis. Or we should continue with data and compare each analysis with canceled column.

• At the starting point of the project, got confused in groupby method.

• It was also difficult to choose between what to include or what not.

Thank you!