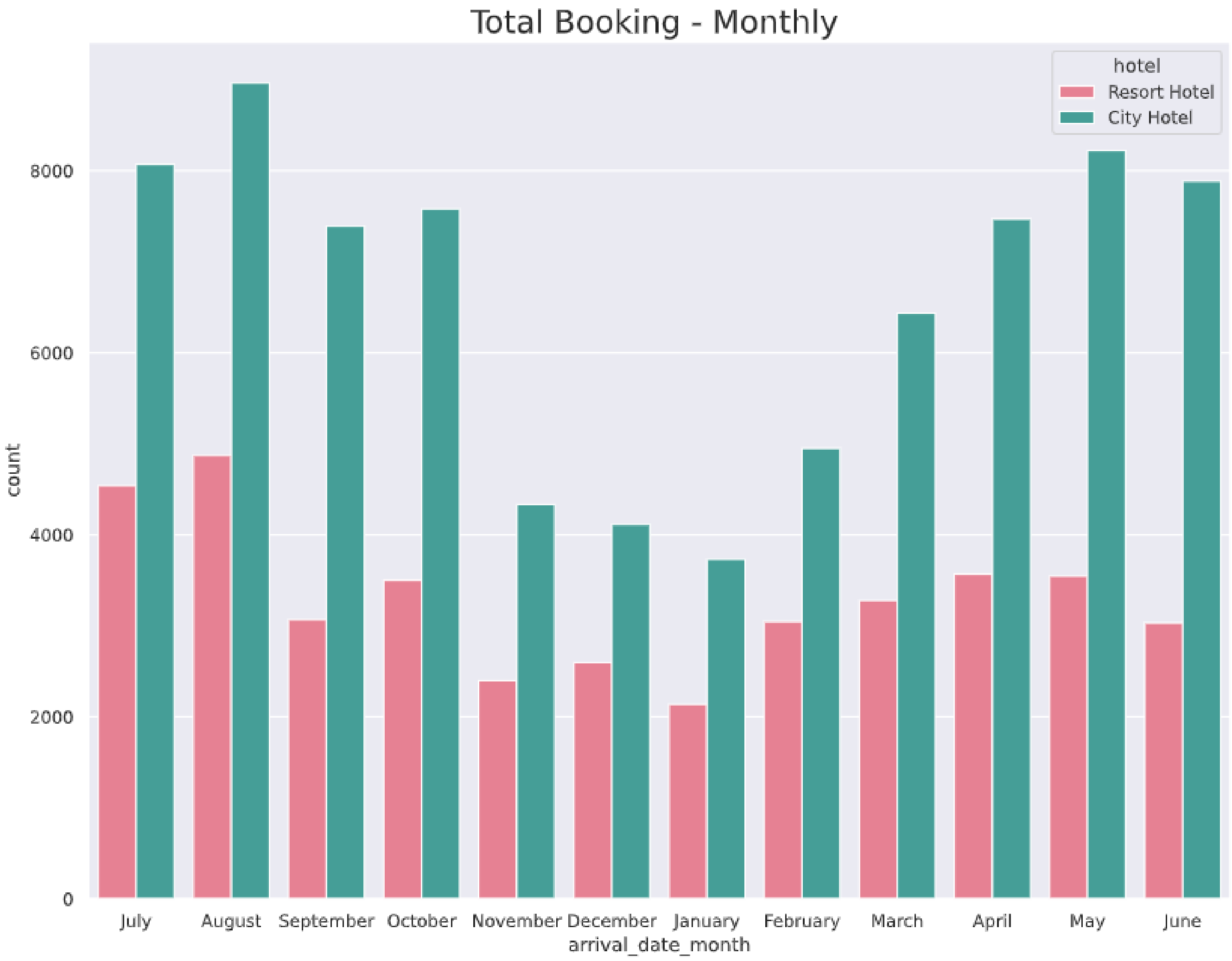
Hotel Booking Dataset

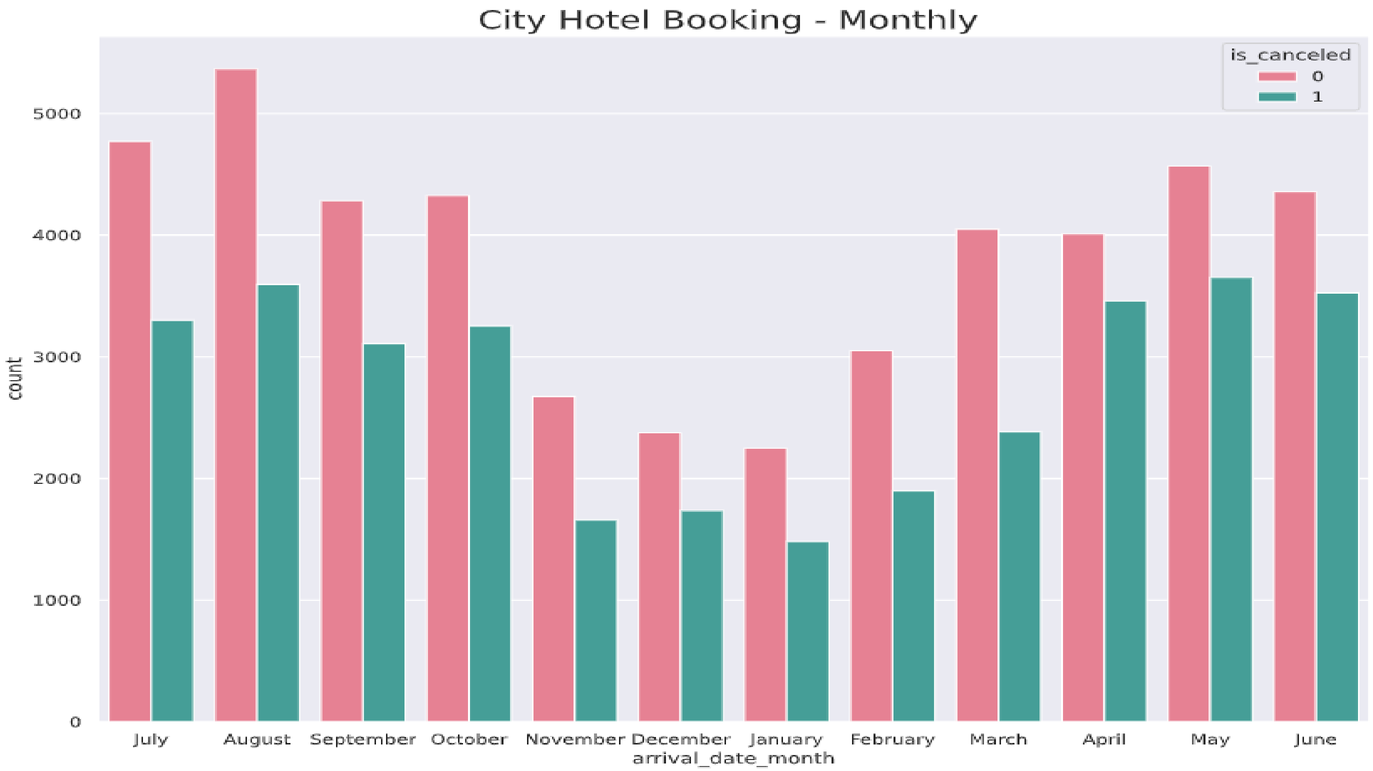
Amit Meena
19010013

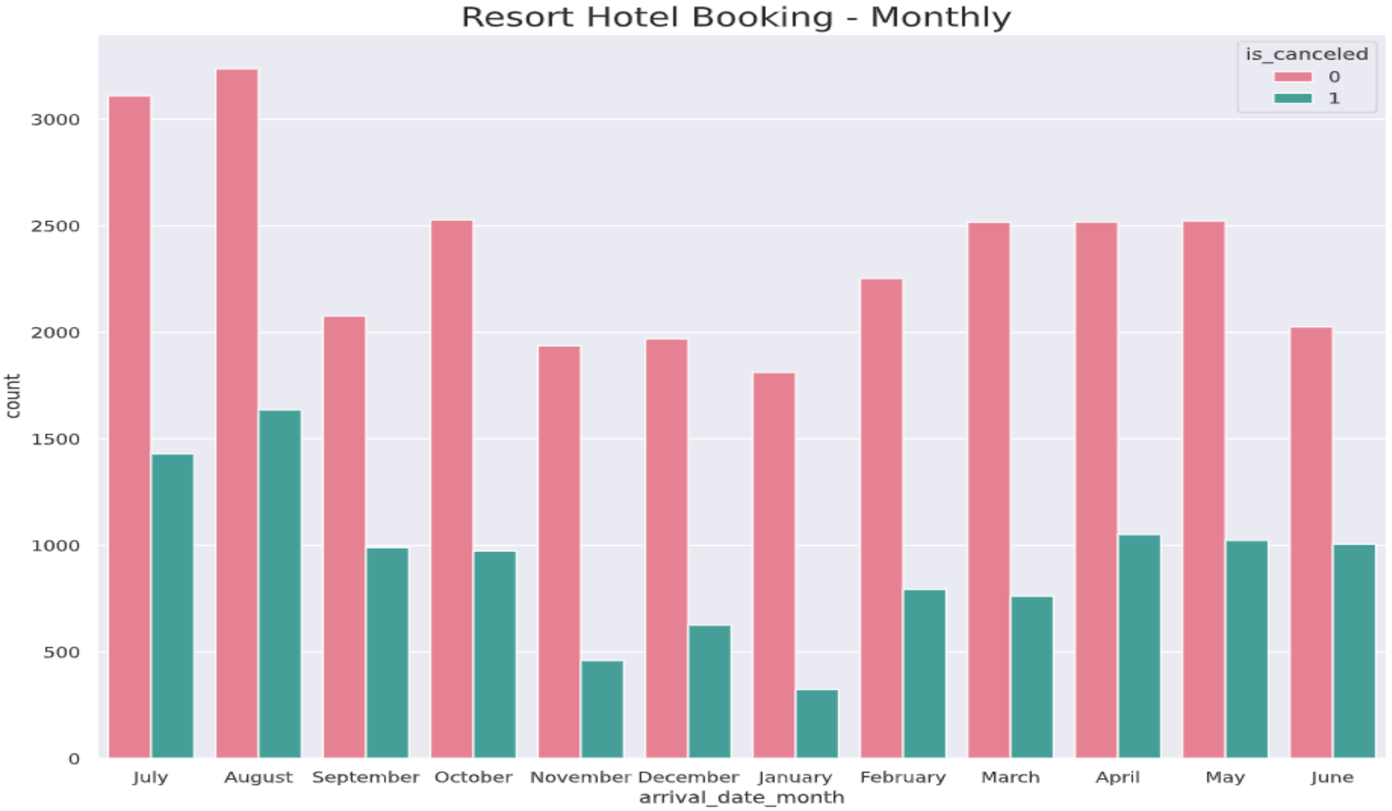
1. How many bookings does each hotel get every month? How many of these bookings are cancelled?



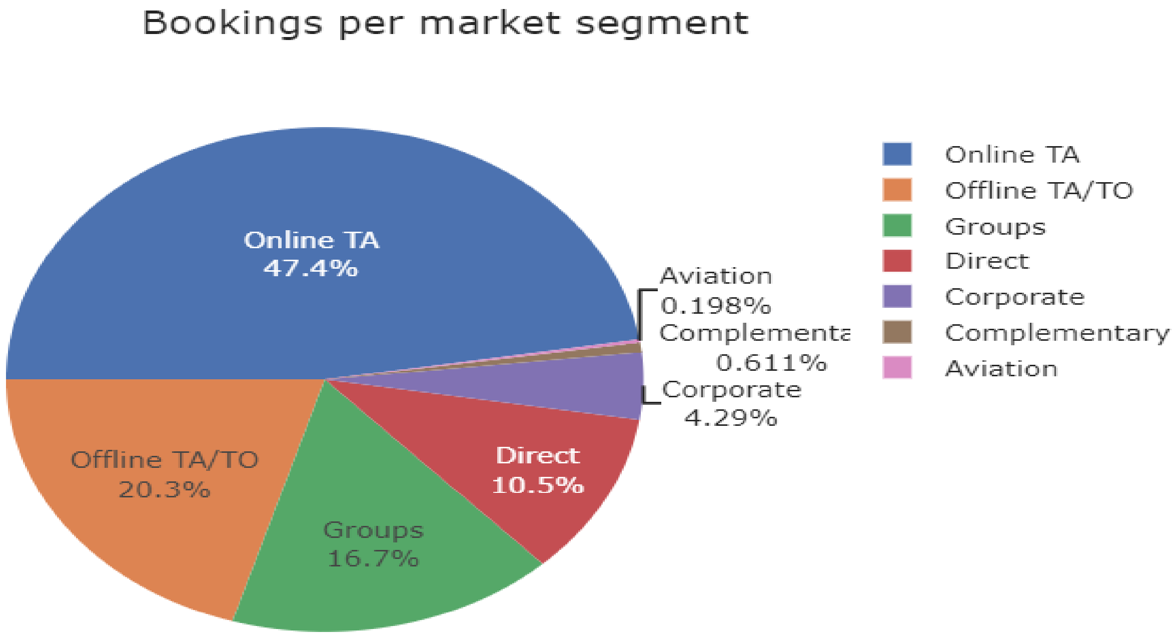
Looking at the first graph, it can be clearly seen that city hotels have more customers in all months. Both hotels have the fewest guests during the winter. The city has more guests during spring and autumn. Same for resort hotel guests during spring is more.

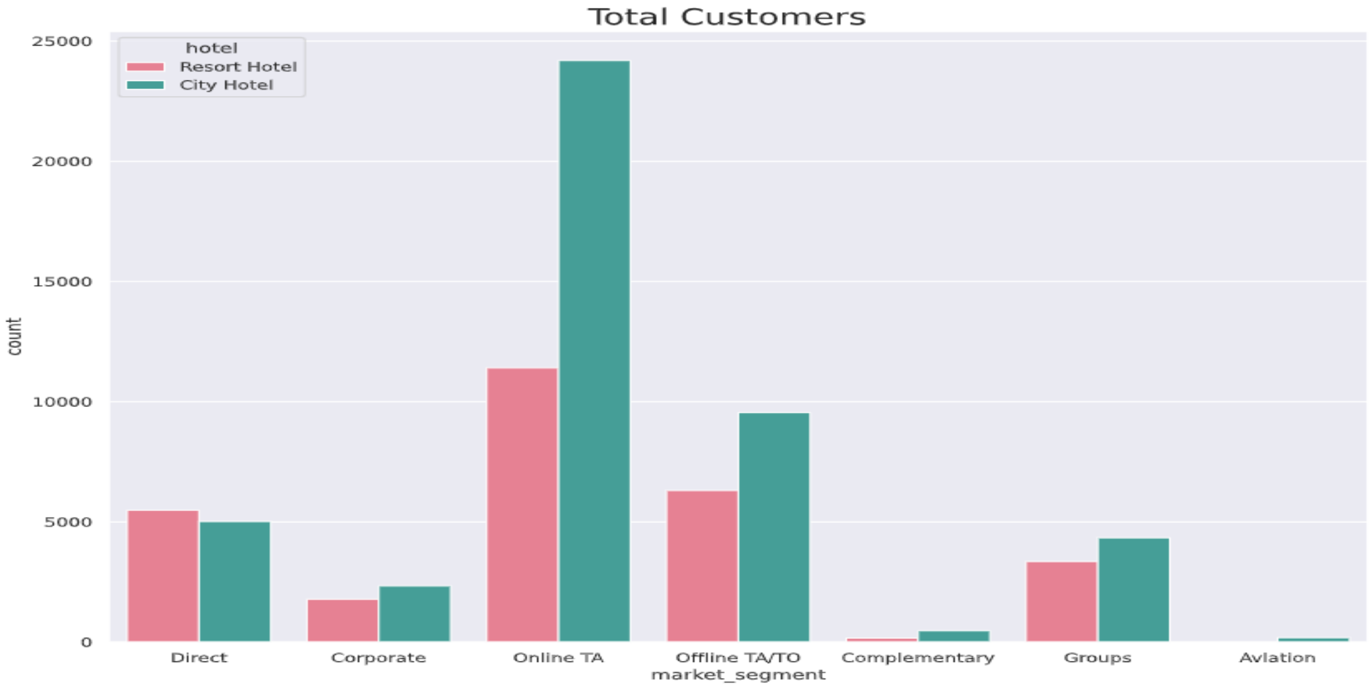
Looking at the second graph and third graph we can clearly seen that resort hotel has high cancelation rate as compare to city hotel.
For the City hotel the relative number of cancelations is around 40% throughout the year. For the Resort it is highest in the summer and lowest during the winter.

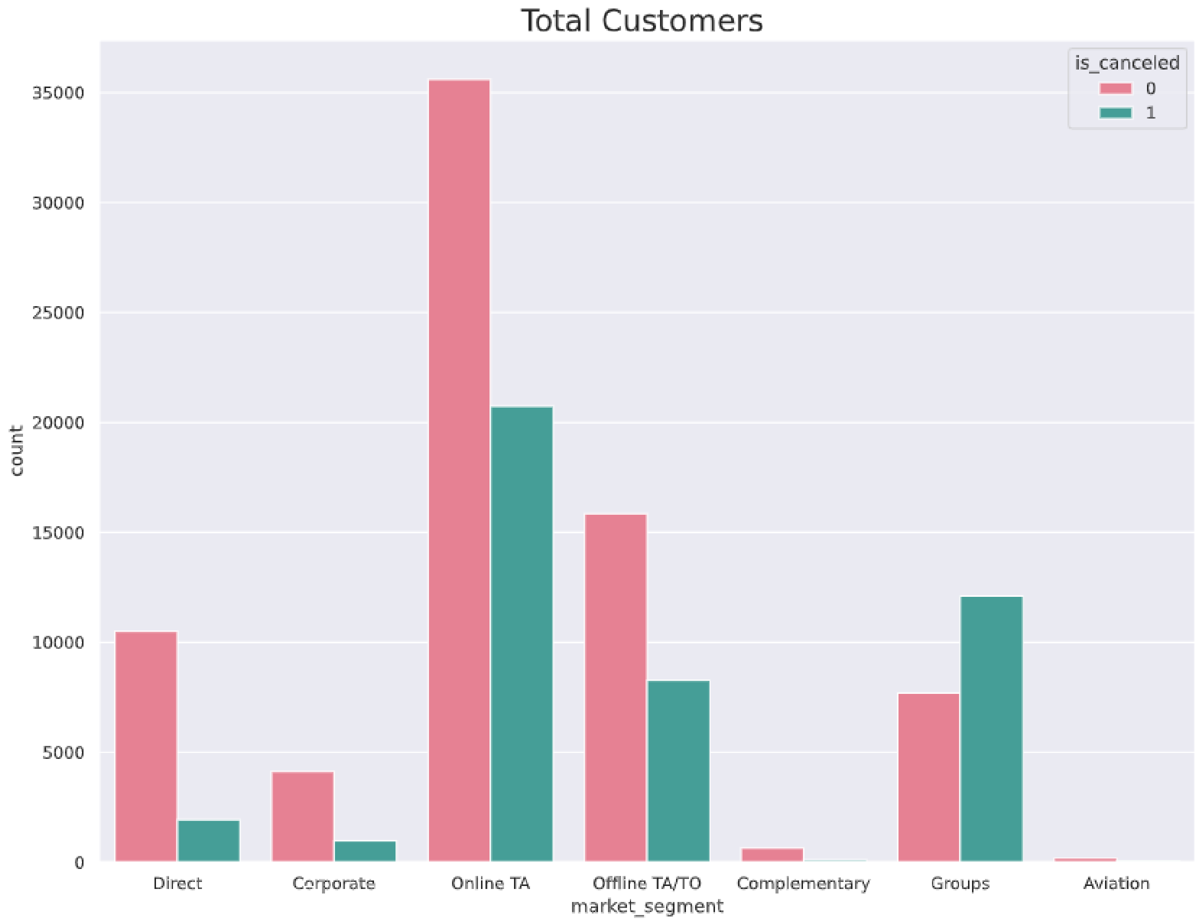

City Hotel Booking - Monthly
is_canceled
0
1
count
0
1000
2000
3000
4000
5000
July
August
September
October
November
December
January
February
March
April
May
June
arrival_date_month


Resort Hotel Booking - Monthly
is_canceled
0
1
count
0
500
1000
1500
2000
2500
3000
July
August
September
October
November
December
January
February
March
April
May
June
arrival_date_month

2.How do the number/ proportion of bookings vary with the market segment for each restaurant? Does the market segment have any correlation with the bookings getting cancelled?







Most of the bookings made by Online TA around 48% followed by Offline Ta/To around 21% followed by Groups around 17% . Aviation segment around 0.2% least among all segments.

Groups segment has cancellation rate more than 50%.
Offline Ta/TO and Online TA has cancellation rate more than 30%
Direct segment has cancellation rate less than 20%.

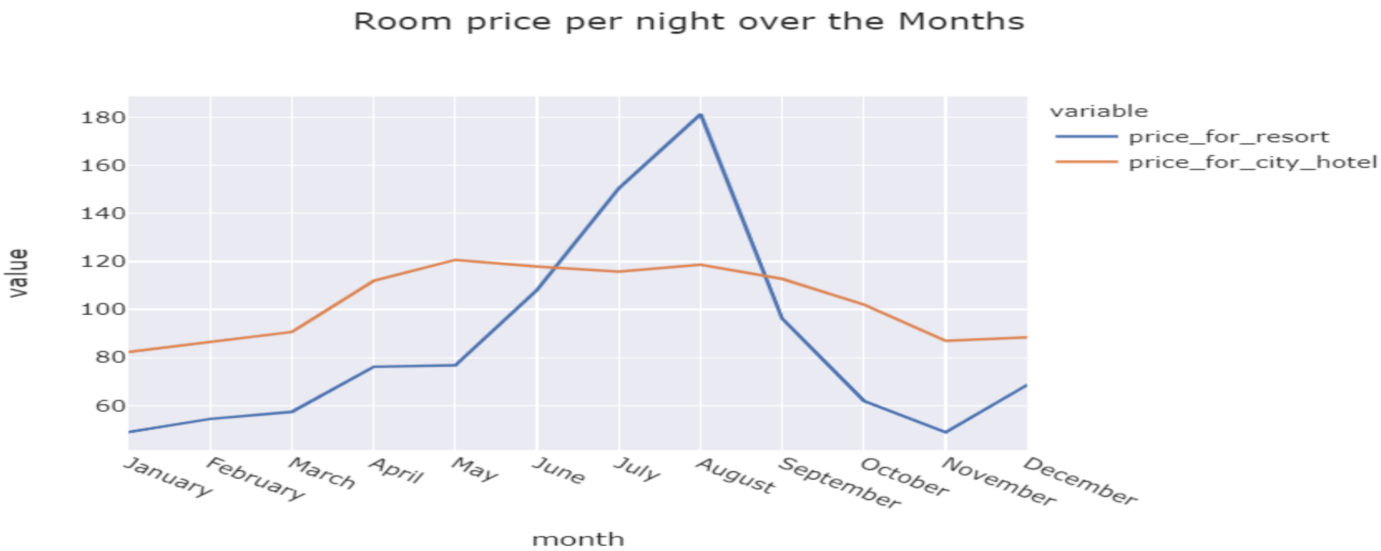
3. Do the hotels cost the same? How does their pricing scheme vary with respect to the month and the market segments? Do you note any kind of special behaviour?

Note both hotel have different room types and different meal arrangements. Seasonal factors are also important, so prices varies a lot.

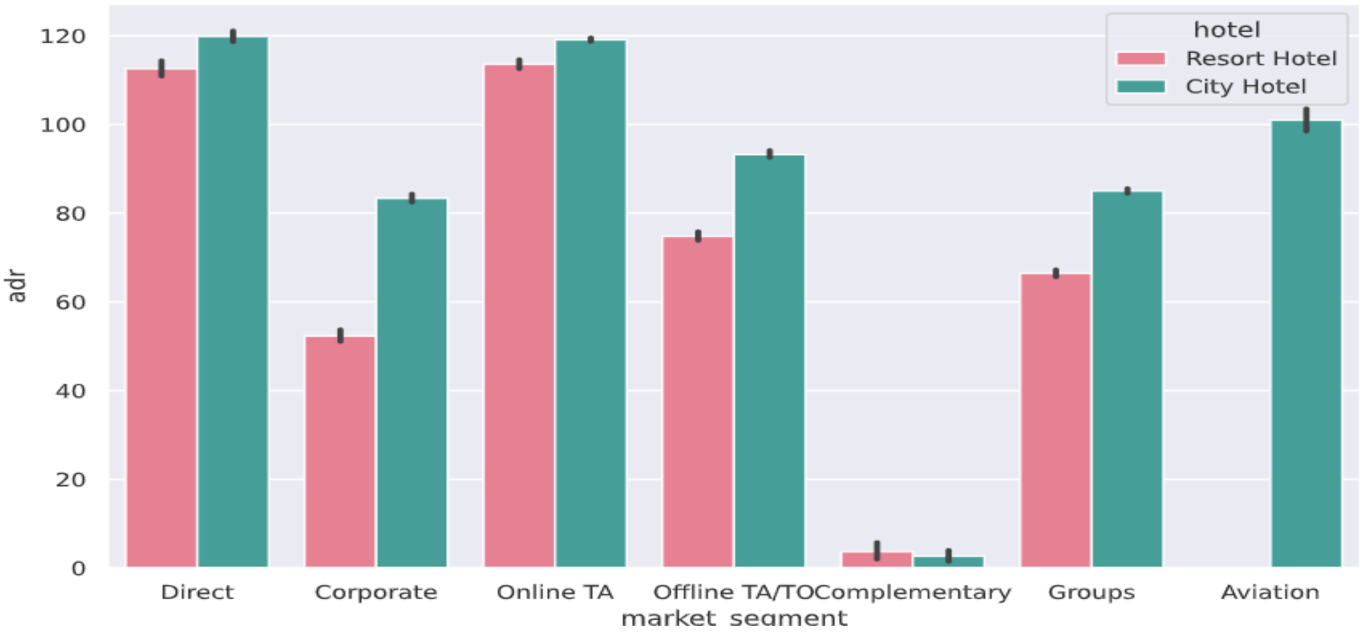
This is clearly shown that the prices in the Resort hotel are much higher during the summer.
The price of the city hotel varies less and is most expensive during spring and autumn.

And also city hotel has more guests during spring and autumn, when the prices are also highest.In July and August there are less visitors, although prices are lower.

Guests numbers for the Resort hotel go down slightly from June to August, which is also when the prices are highest.



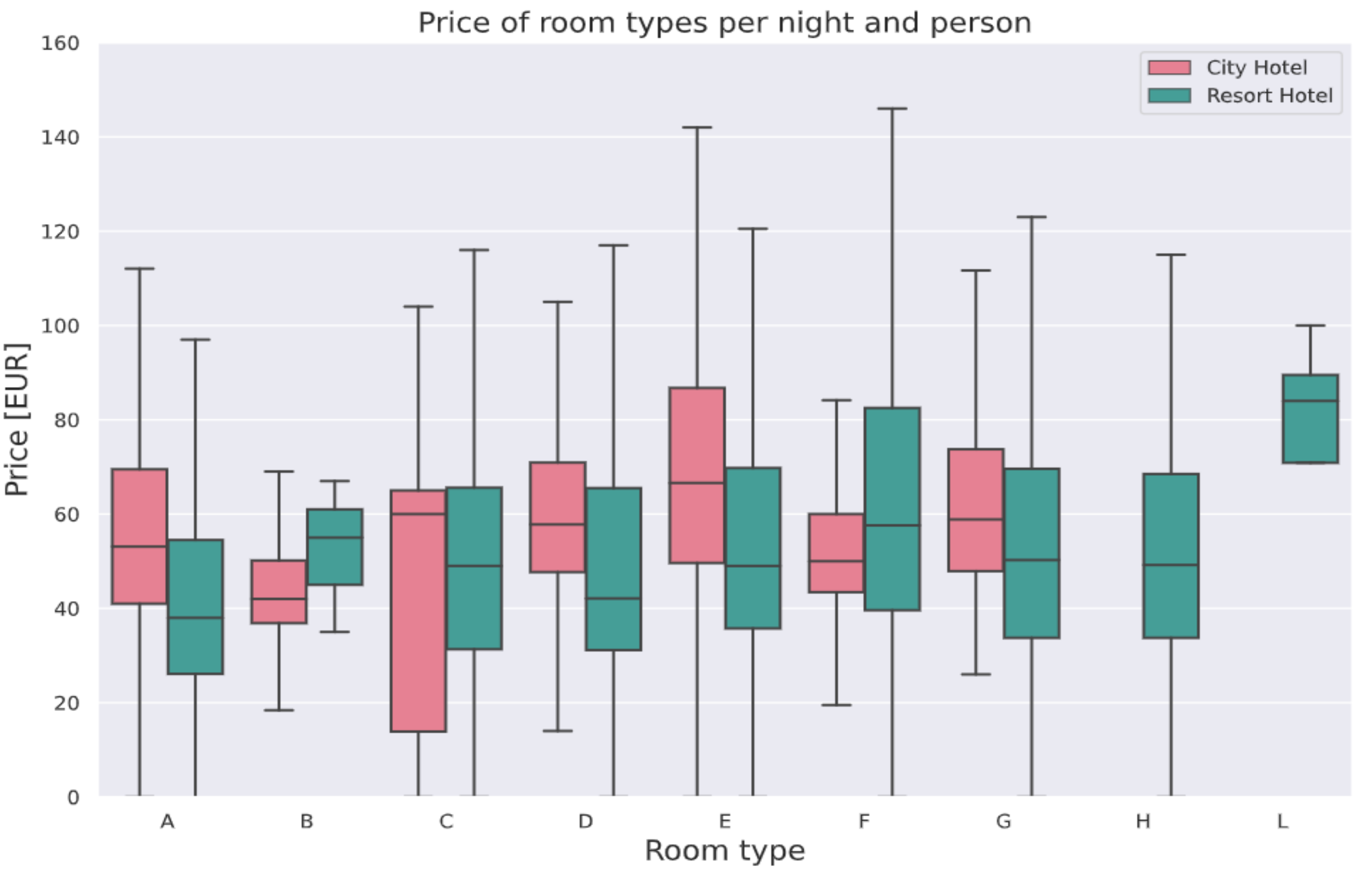
From the second graph it clearly seen For Complementary market segment prices are relatively lower. And for Direct and Online Ta market segments prices are relatively higher.



4. Has there been any change in the demographics (country of origin) visiting those hotels over the years?

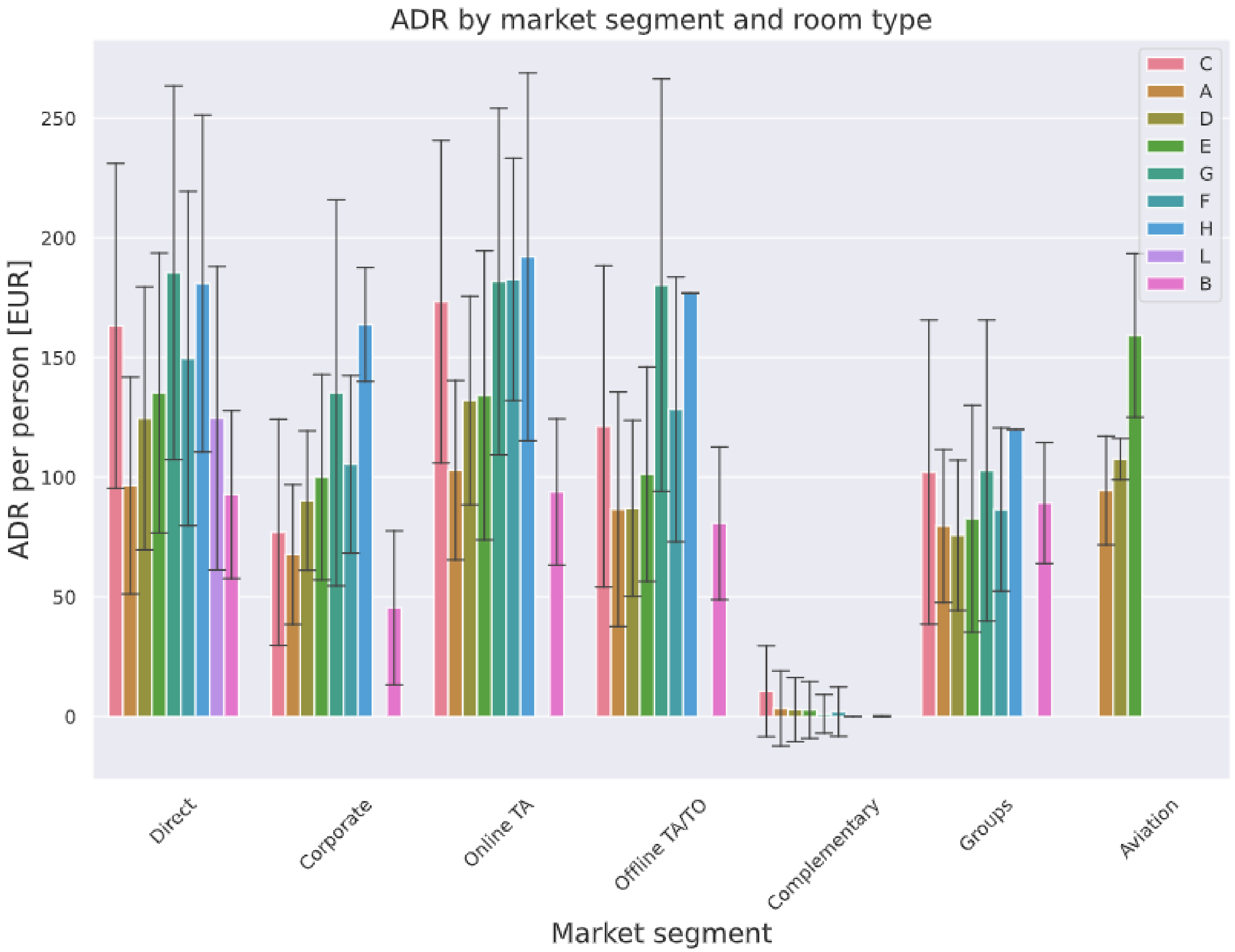
Dataset is insufficient to answer as datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017.

5. How does the price charged by the hotel vary for the different kind of rooms offered for each hotel? Or even for different market segments?



This figure shows the average price per room, depending on its type and the standard deviation.
But note that data anonymization rooms with the same type letter may not necessarily be the same across hotels.

On average, groups get the best prices

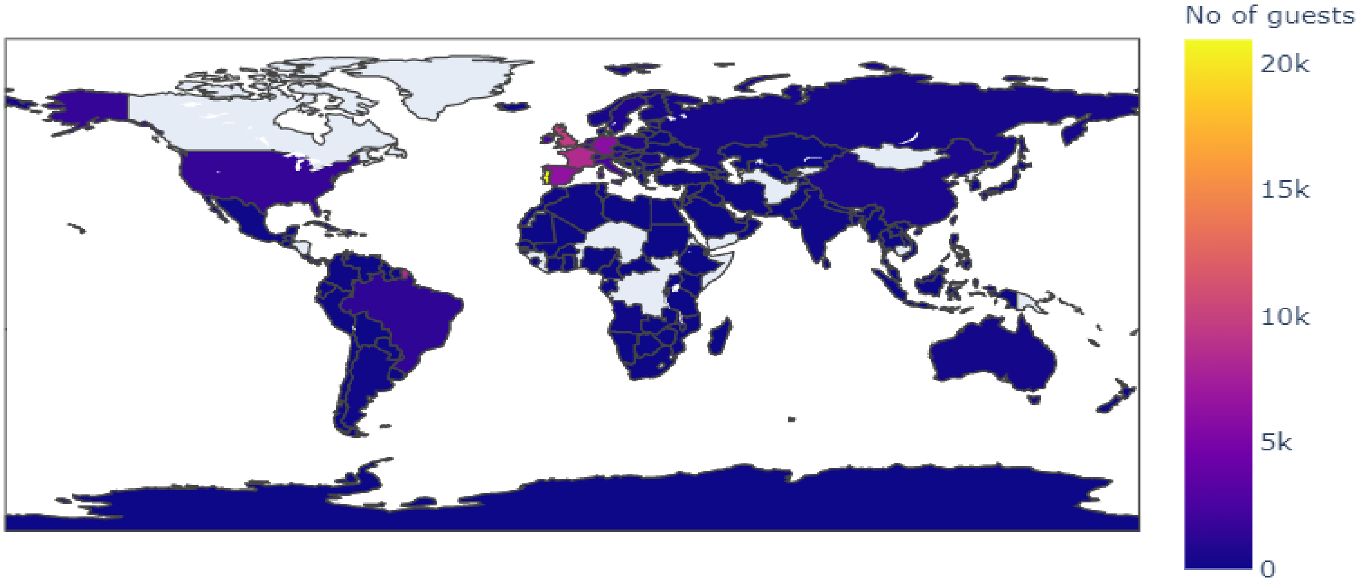


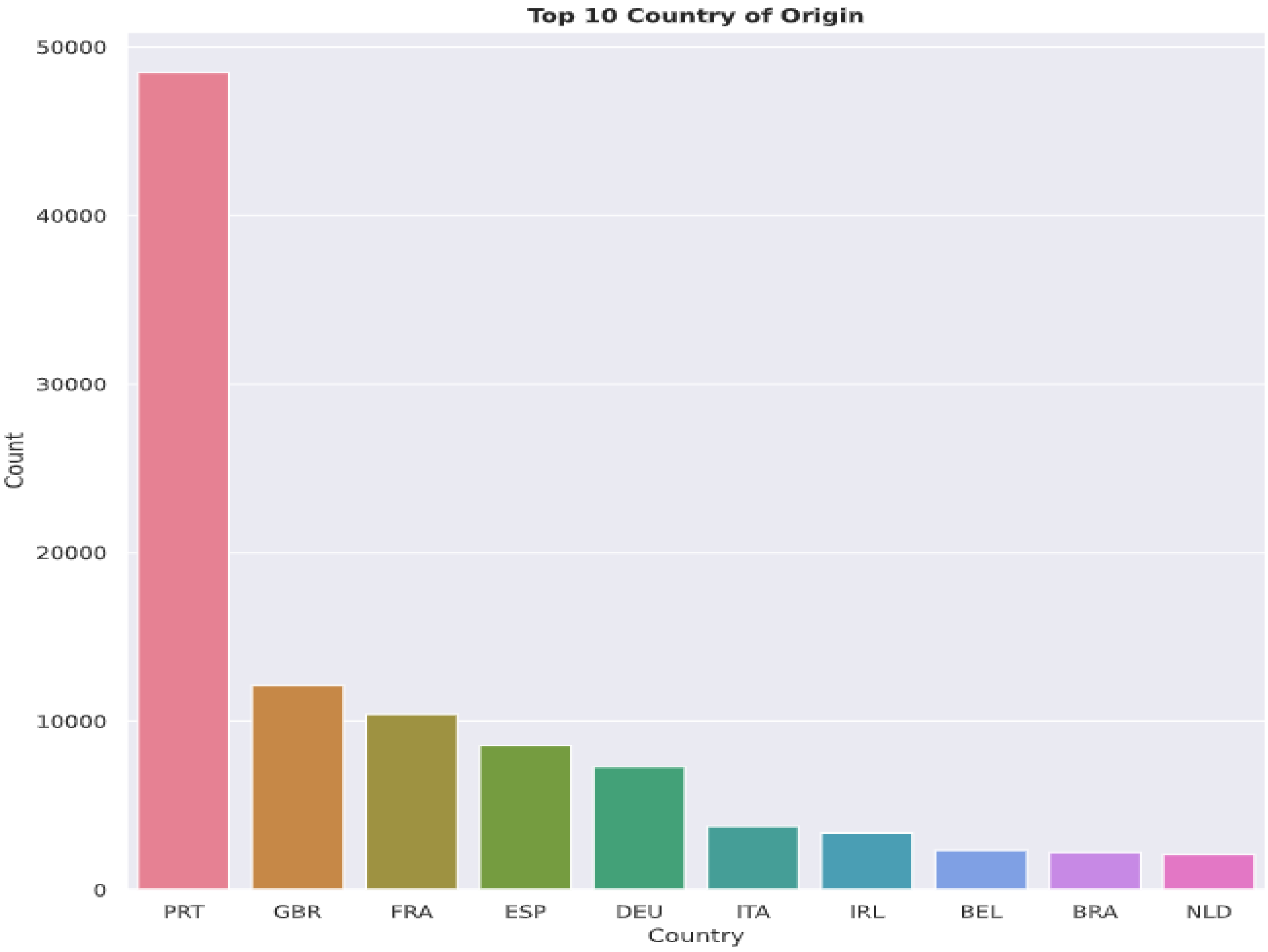
#Additional question answers

1.which are the three most common and uncommon countries of origin ?

As We can see in graph, Portugal tops the lists with 48,586 of the cases, followed by the Great Britain with 12,129 and France with 10,414.

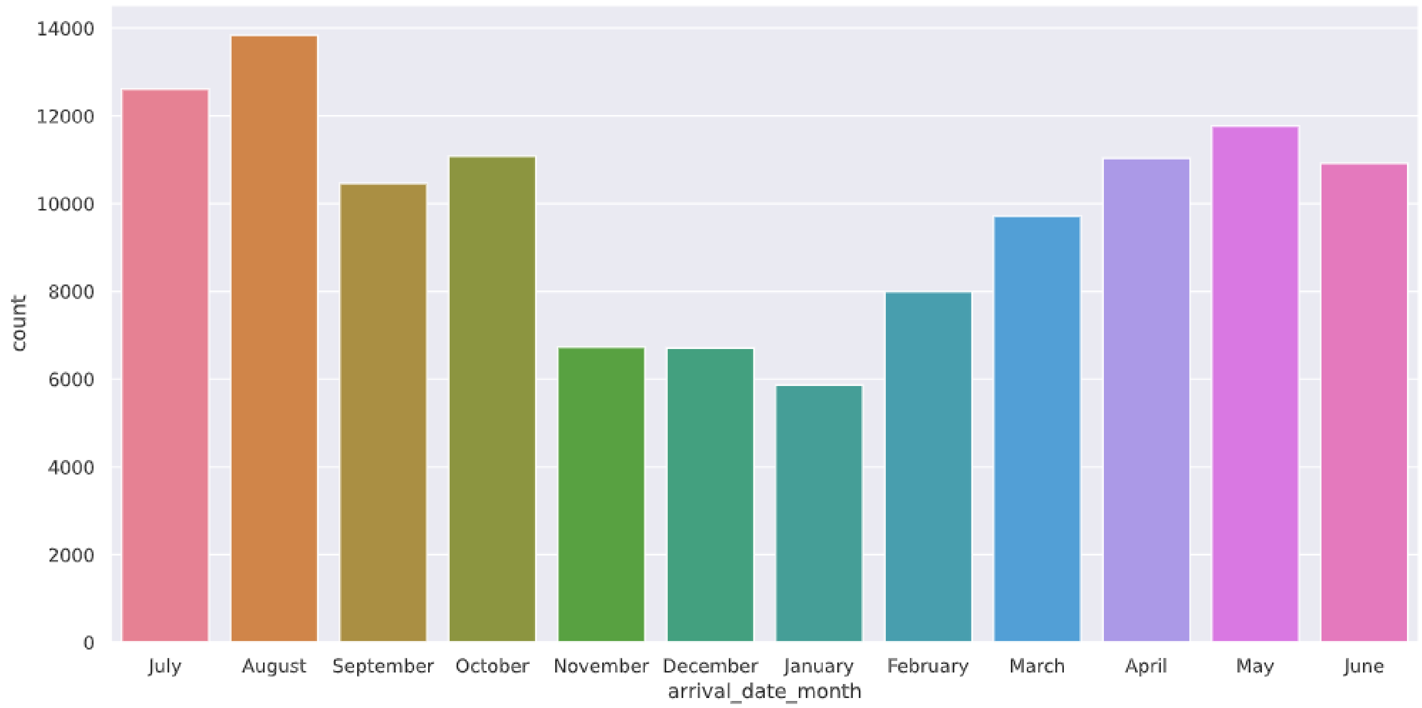
There are 30 countries that can be considered as the most uncommon country of origin, with 1 guests per country. Amont these, we have Madagascar, Dominica, Mali, Guyana, Palau, Kiribati, Sudan, etc.


No of guests
20k
15k
10k
5k
0


Top 10 Country of Origin
Count
50000
40000
30000
20000
10000
0
PRT
GBR
FRA
ESP
DEU
ITA
IRL
BEL
BRA
NLD
Country

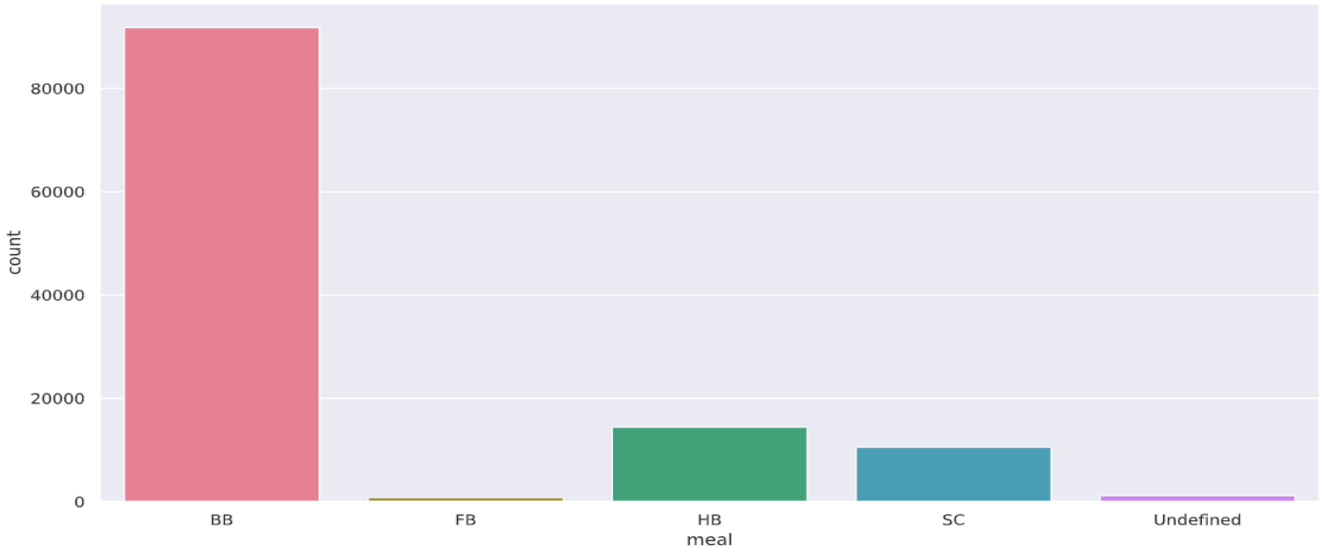
2. Which are the month of highest and least occupation?

The month of highest occupation is august with 11.65% of the reservations. The month of least occupation is January with 4.94% of the reservations.



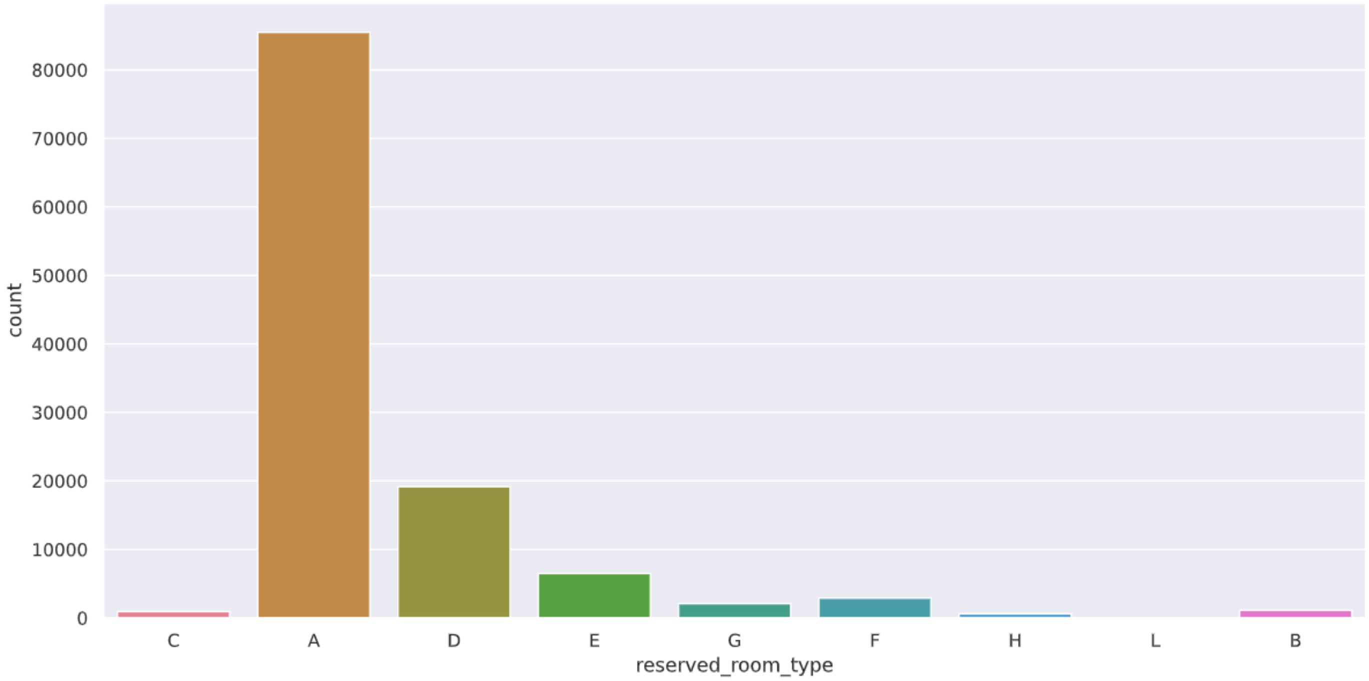
3. What is the most popular meal package?

The Bed & Breakfast option is the most popular, with frequency of 77.26%.



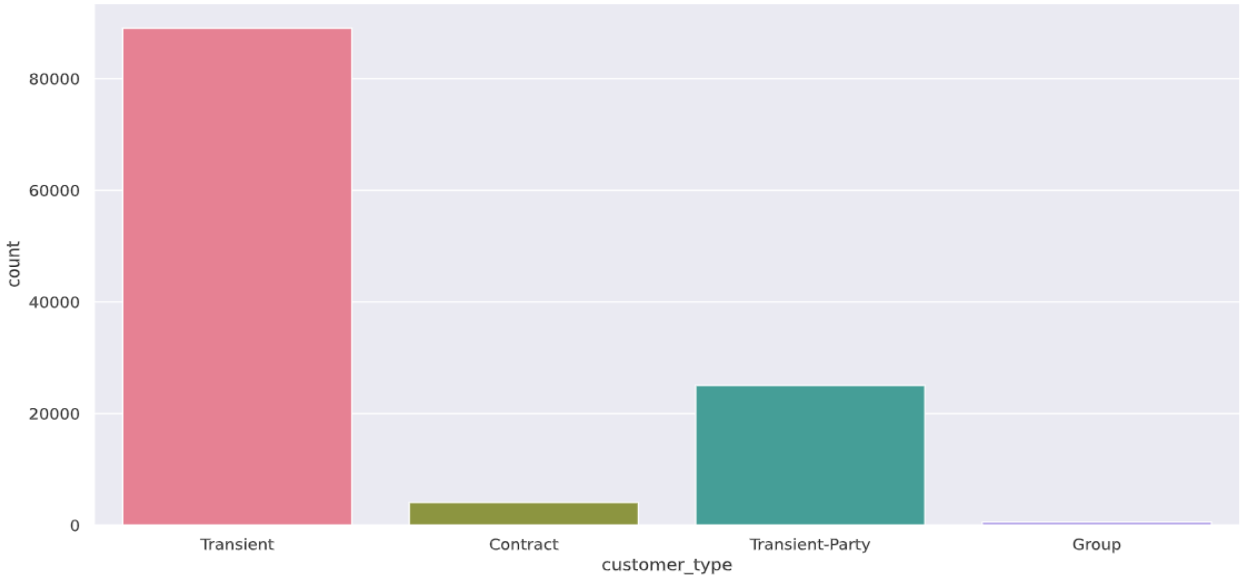
4 . Which is the most reserved room type?

The “A” room type is the most popular among the clients, with 71.99% of the reservations.



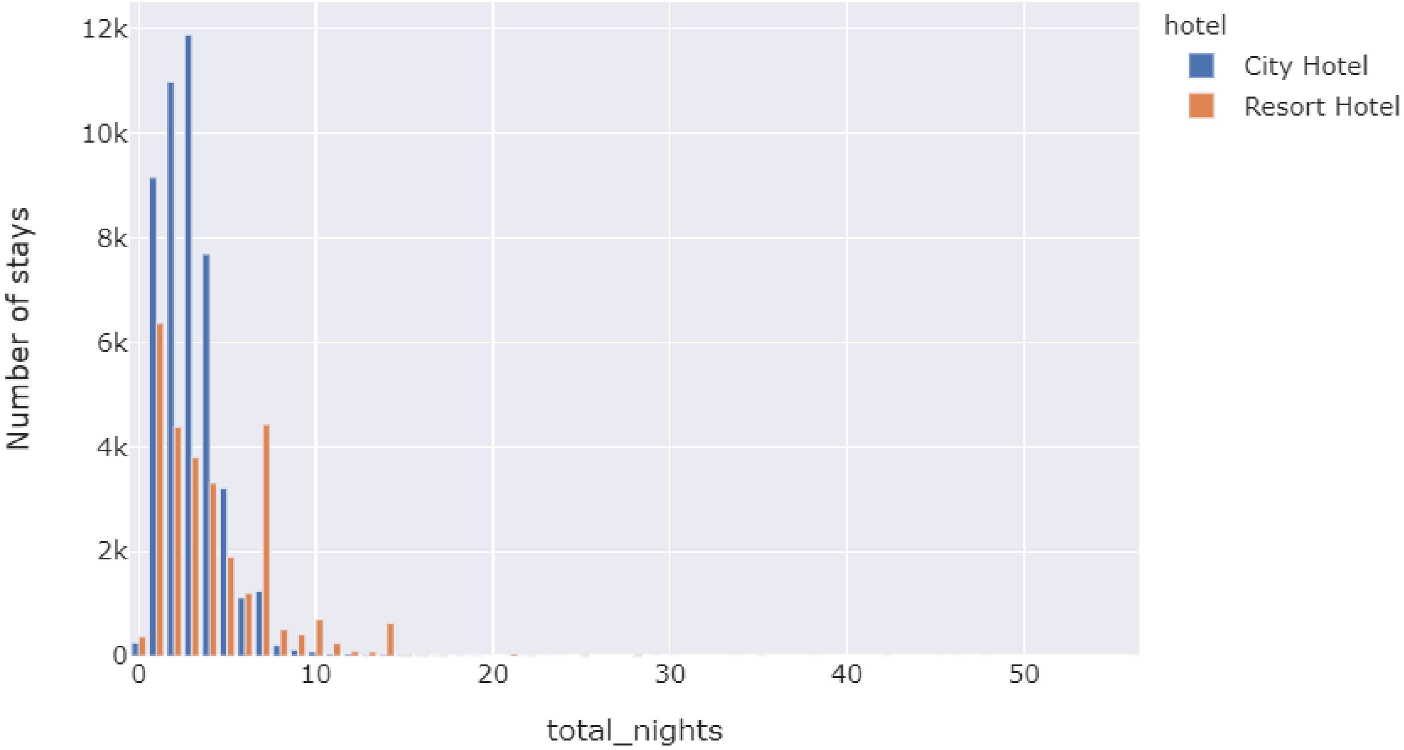
5.Which is most common customer type?

Trasients are the most common customer type, they represent 75% of the total customers.



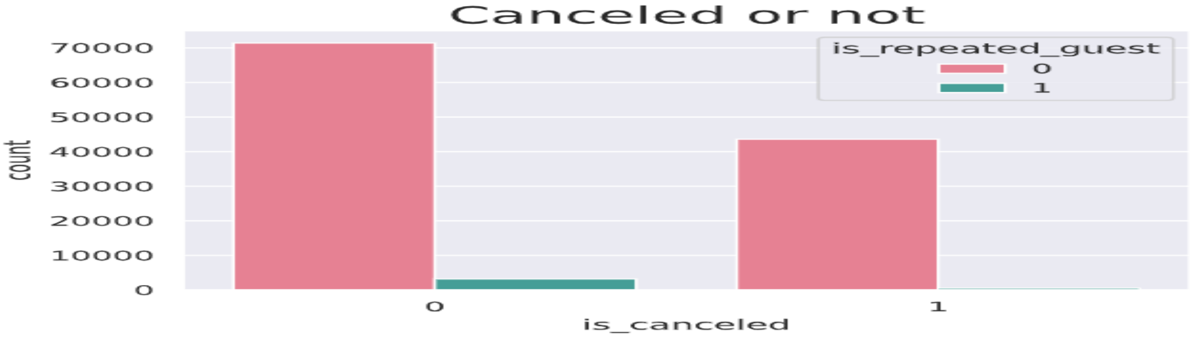
6.How long do people stay at the hotels?

For the city hotel there is a clear preference for 1-4 nights.
For the resort hotel, 1-4 nights are also often booked, but 7 nights also stand out as being very popular.



7.Repeated guest effect on cancellations.

Most of the customers are not repeated guests. There is on surprise that repeated guests do not cancel their reservations. Of course there are some exceptions.



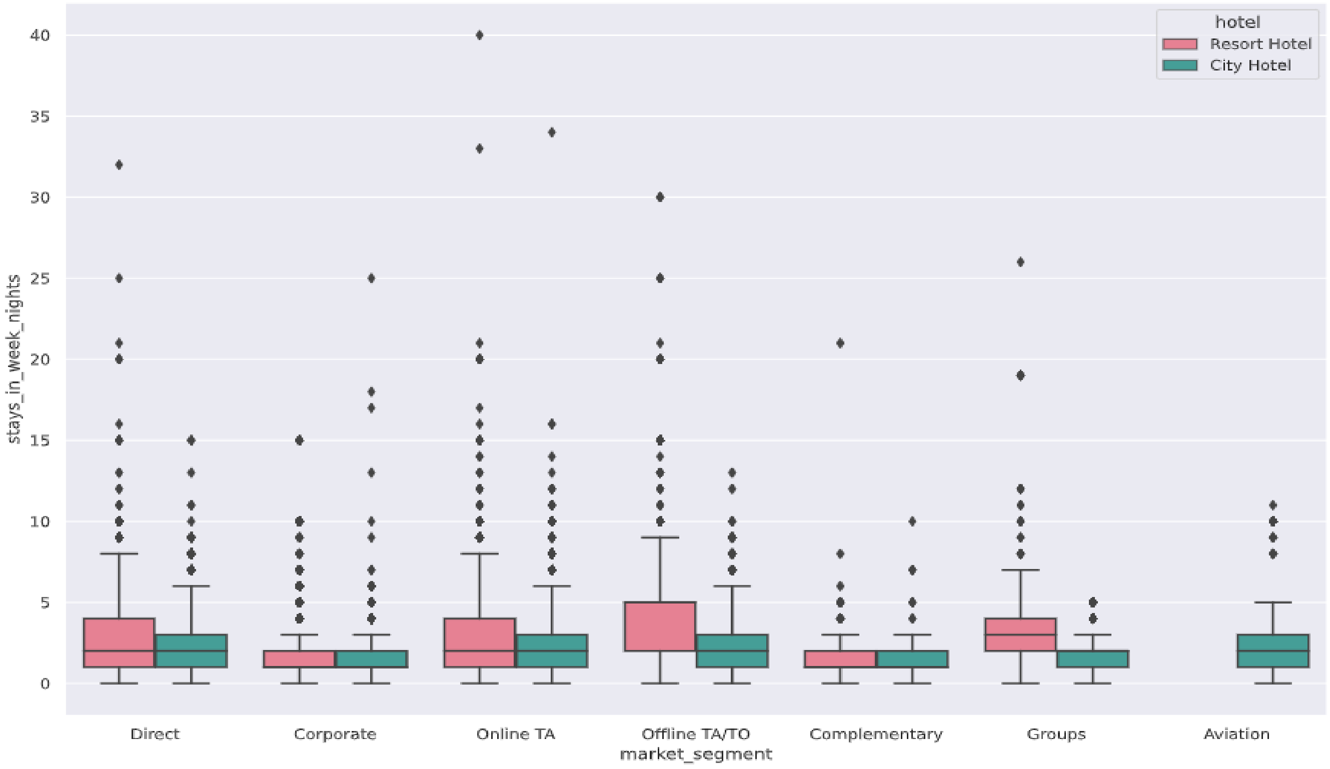
8.Hotel type with more time spent.

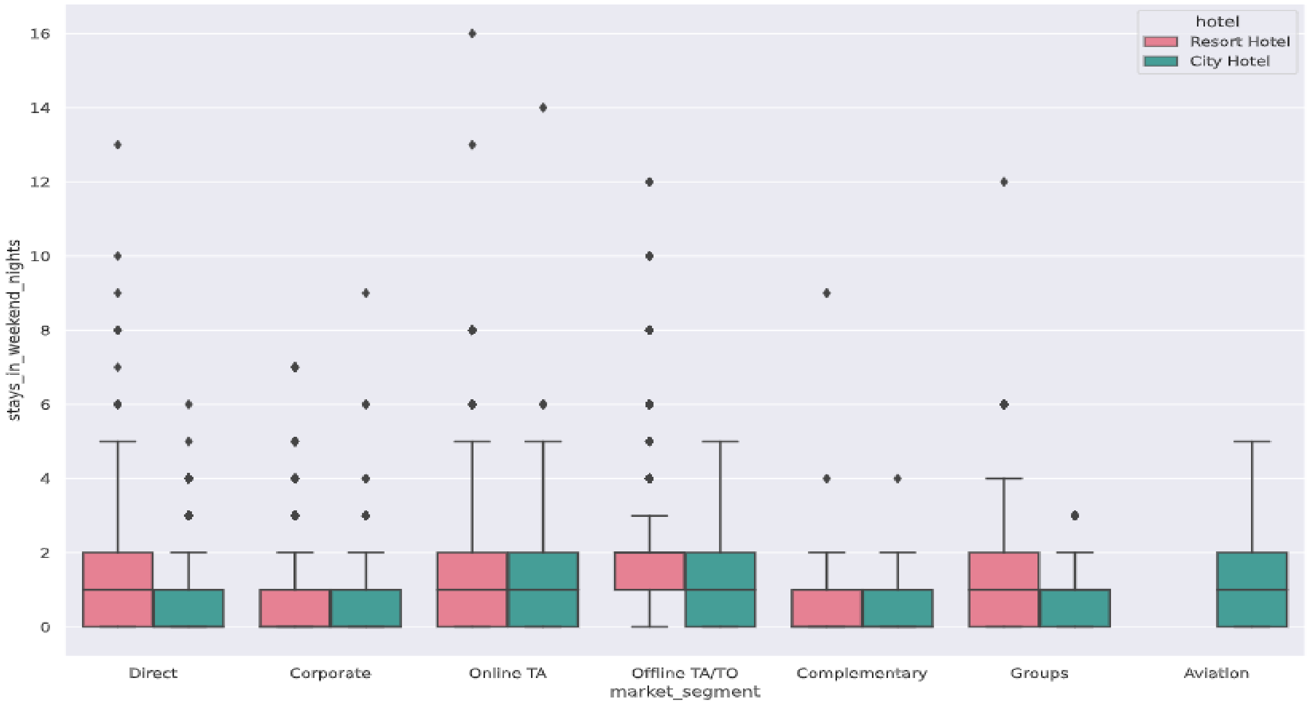
It can be seen that most of the group are normal distributed. Looking at the distribution, most people do not seem to prefer to stay at the hotel for more than 1 week, but it seems normal to stay in resort hotel for up to 12-13 days.

As it turns out, customers from Aviation Segment do not seem to be staying at the resort hotels and have a relatively lower day average.

This can be because Customers in the Aviation Segment are likely to arrive shortly due to business. Also probably most airports are a bit away from sea and its most likely to be close to city hotels.

It is obvious that when people go to resort hotels, they prefer to stay more.





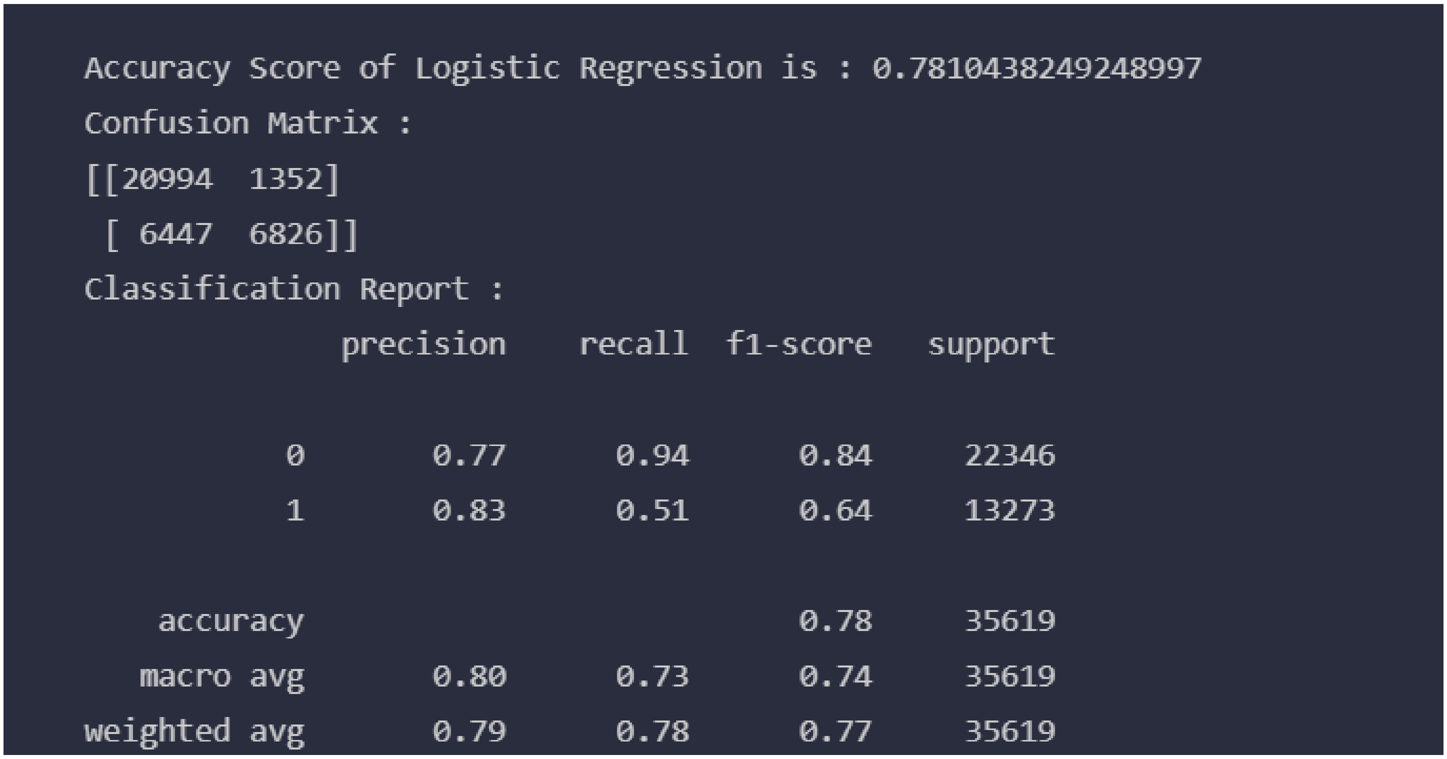
Inferences and Conclusion

- The majority of guests come from western europe countries.
- the majority of reservations are for city hotels.
- The number of repeated guests is too low.

Model Building

Logistic Regression :-

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analyzing the association of all variables together. In this article, we explain the logistic regression procedure using examples to make it as simple as possible. After definition of the technique, the basic interpretation of the results is highlighted and then some special issues are discussed.

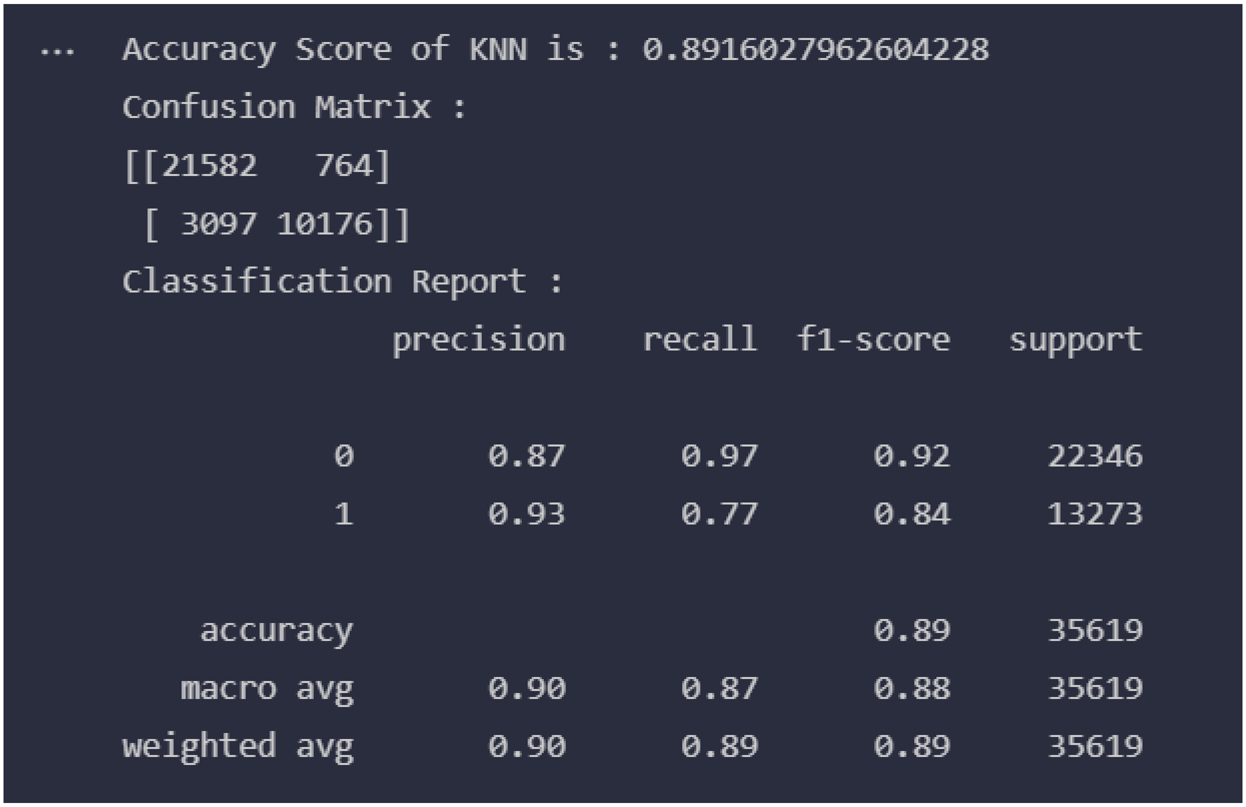


```
Accuracy Score of Logistic Regression is : 0.7810438249248997
Confusion Matrix :
[[20994  1352]
 [ 6447  6826]]
Classification Report :
              precision    recall  f1-score   support

           0       0.77      0.94      0.84     22346
           1       0.83      0.51      0.64     13273

    accuracy                           0.78     35619
   macro avg       0.80      0.73      0.74     35619
weighted avg       0.79      0.78      0.77     35619
```

As we can see Accuracy Score of Logistic Regression is : 78%

KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. Pause! Let us unpack that.

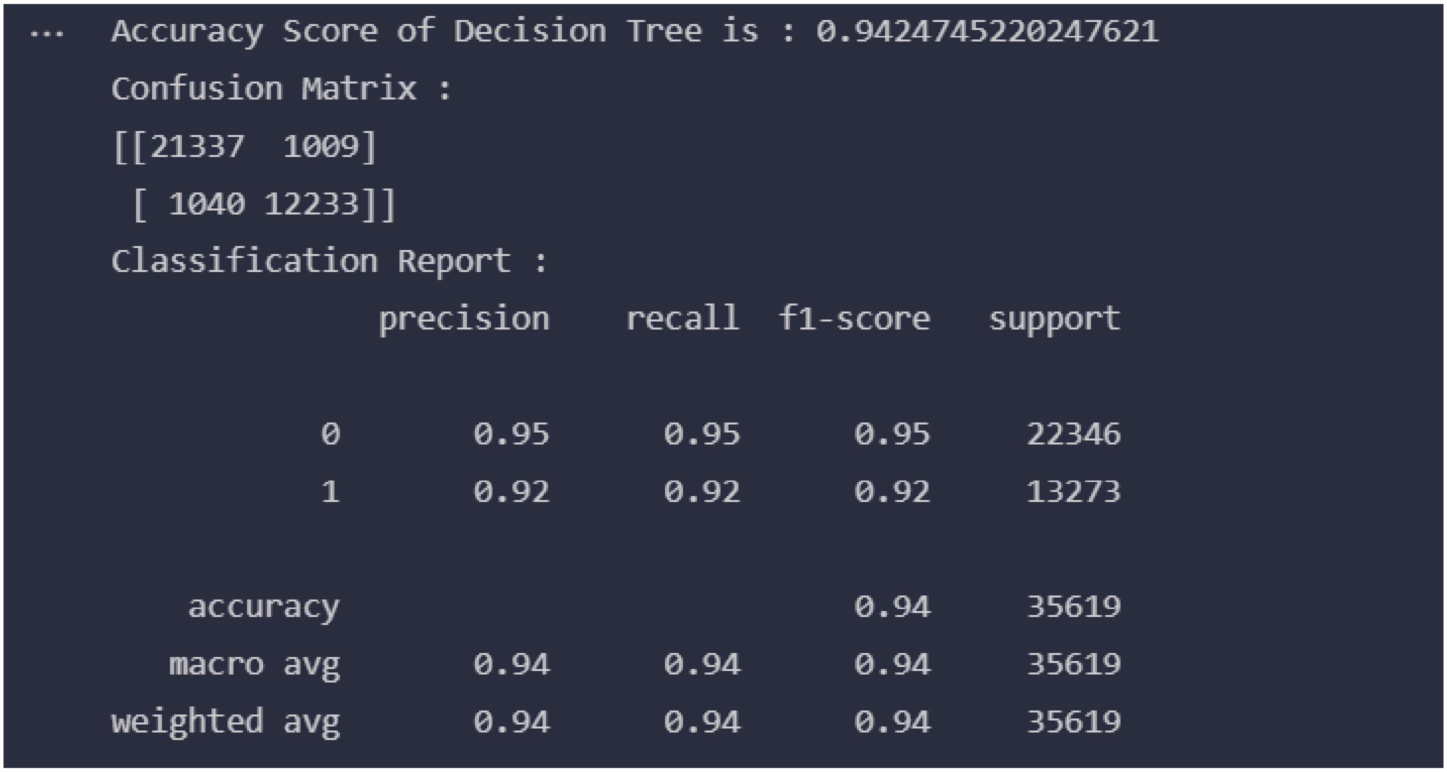


```
...  Accuracy Score of KNN is : 0.8916027962604228
     Confusion Matrix :
     [[21582   764]
      [ 3097 10176]]
     Classification Report :
                   precision    recall  f1-score   support

                0       0.87      0.97      0.92     22346
                1       0.93      0.77      0.84     13273

         accuracy                           0.89     35619
        macro avg       0.90      0.87      0.88     35619
     weighted avg       0.90      0.89      0.89     35619
```

As we Can See Accuracy Score of KNN is 89% not that bad :p

Decision Tree Classifier

Decision Trees are also used in tandem when you are building a **Random Forest classifier** which is a culmination of multiple Decision Trees working together to classify a record based on majority vote. A Decision Tree is constructed by asking a serious of questions with respect to a record of the dataset we have got.

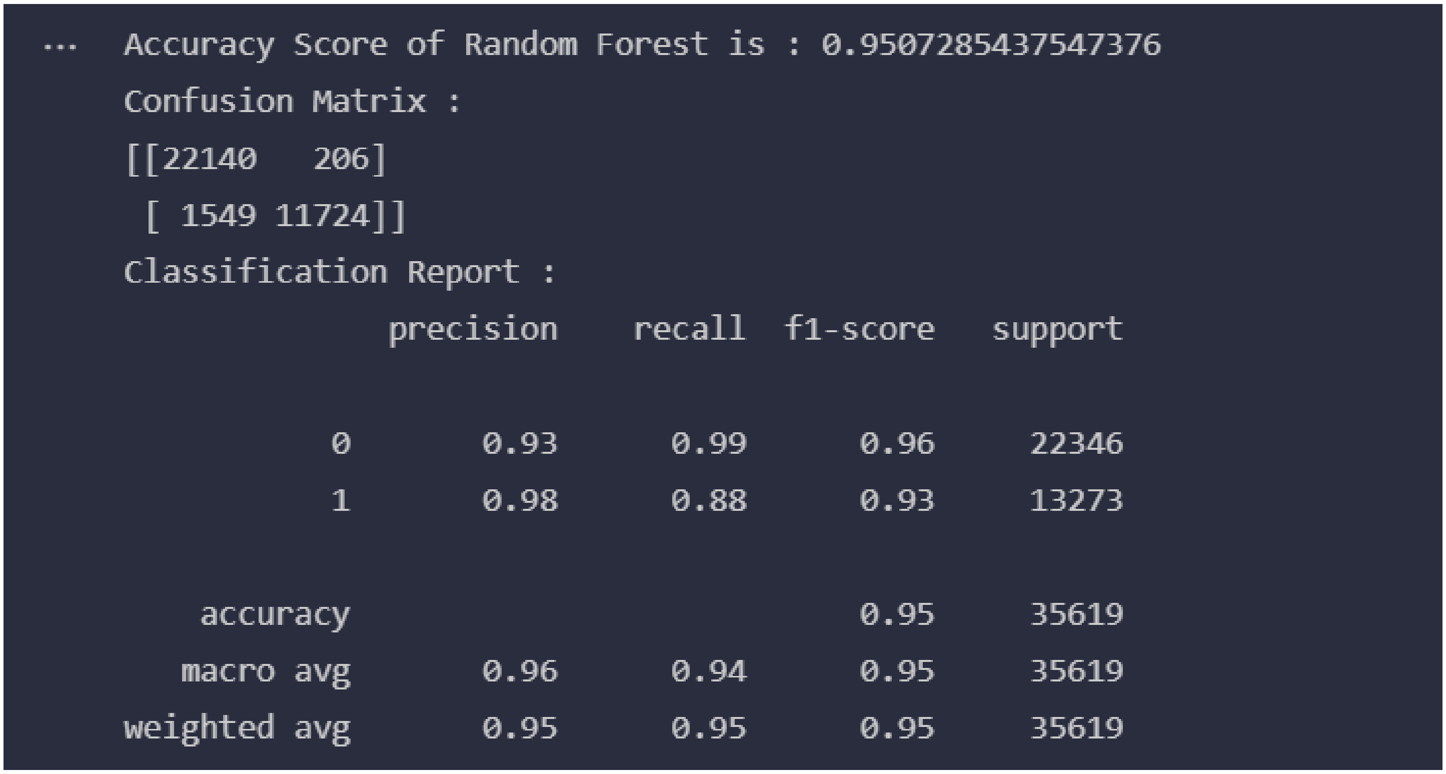


```
...  Accuracy Score of Decision Tree is : 0.9424745220247621
     Confusion Matrix :
     [[21337  1009]
      [ 1040 12233]]
     Classification Report :
                   precision    recall  f1-score   support

                0       0.95      0.95      0.95     22346
                1       0.92      0.92      0.92     13273

         accuracy                           0.94     35619
        macro avg       0.94      0.94      0.94     35619
     weighted avg       0.94      0.94      0.94     35619
```

As we can see Accuracy Score of Decision Tree is 94.24% which is fairly good

Random Forest Classifier

The random forest is a classification algorithm consisting of many decisions trees. It **uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee** is more accurate than that of any individual tree.



```
...   Accuracy Score of Random Forest is : 0.9507285437547376
      Confusion Matrix :
      [[22140   206]
       [ 1549 11724]]
      Classification Report :
                    precision    recall  f1-score   support

                 0       0.93      0.99      0.96     22346
                 1       0.98      0.88      0.93     13273

          accuracy                           0.95     35619
         macro avg       0.96      0.94      0.95     35619
      weighted avg       0.95      0.95      0.95     35619
```

As we can see Accuracy Score of Random Forest is

95.07% which is fairly reasonable number..!!