**Summary Report**

**Professor: Fadi Alsaleem**

Student: Pritam Shrestha

Date: 07/01/2020

Language: Python and R

File1: Hotel\_recommendation\_python( used for python)

File2: Hotel\_recommendation\_r( used notebook for r command for better visualization)

For this assignment, I have used two programming languages Python and R.

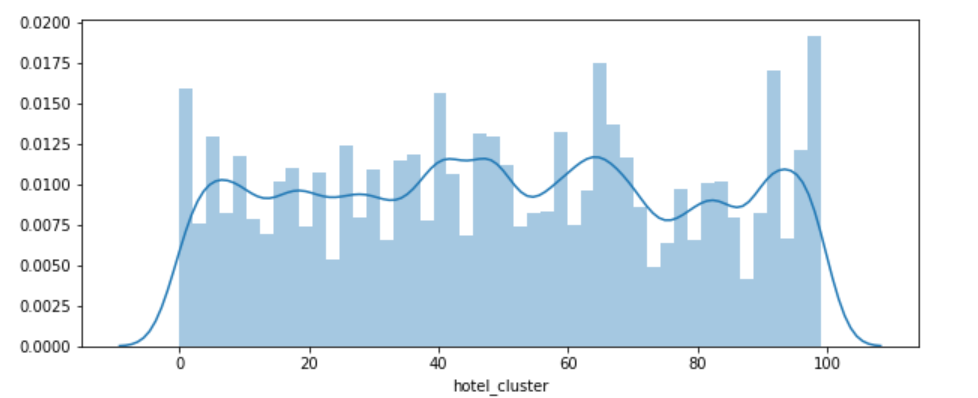
The given datasets are massive so I have encountered so many problems while downloading and reading but finally read 100000 rows. To complete this assignment, I read those train and destination files on Jupiter notebook using r. Using R, I have loaded the file and saw the structure of the file using a summary and also plotted a couple of plots. After viewing its overview and structure, I have concluded that I will do EDA and modeling in python because my Jupiter r file crashed more than 50 times.

After analyzing the data structure and summary of the data, I have completed some EDA and modeling on python. Here is some overview of the completed steps.

1. EDA :

After analyzing the hotel\_cluster variable, there are 100 different clusters in the data.

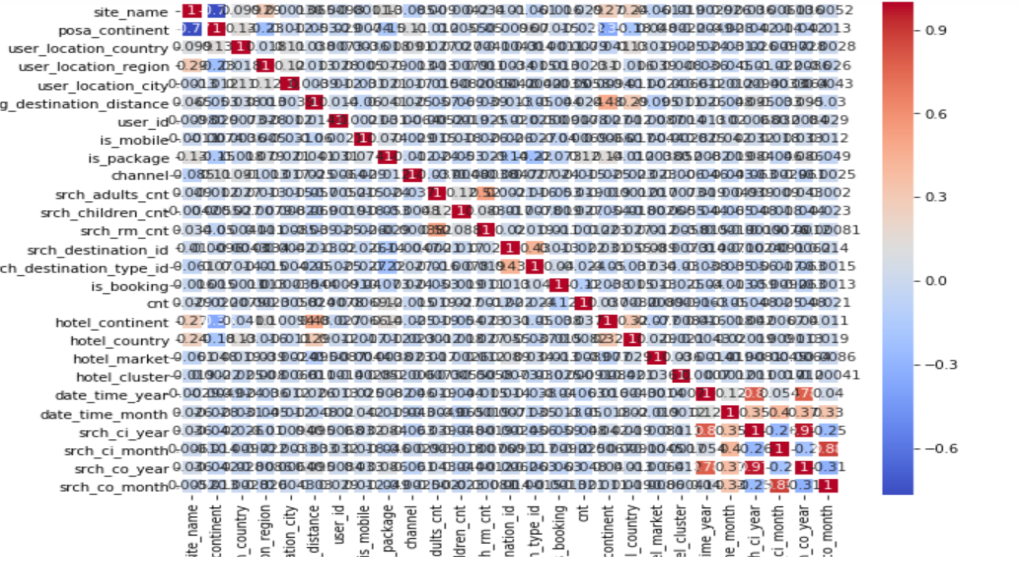
There are two distinct values for is\_booking, 0 and 1, where one means booking and zero means click.



From the above figure, it is clear that the data is well-distributed over 100 clusters with little skewness.

**Correlations of the entire dataset**

Correlation is the best way to find the relationship between variables, so I have plotted a heat map to visualize it.



I went through each variable of the training dataset to extract the insights for clustering and completed all required stapes such as data transformation and missing handling, dropping, etc.

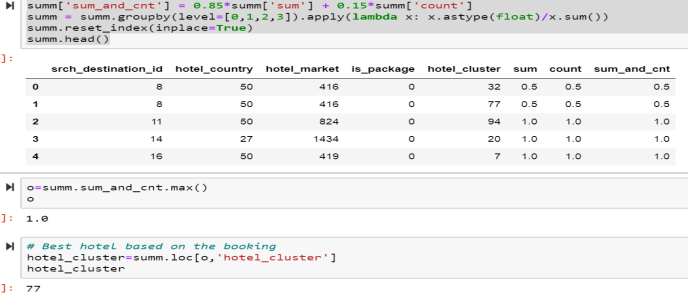
**Algorithms (Modeling)**

In this section, I have separated the target variable and predictor variable and implemented four different algorithms for modeling.

1. **Support Vector Machine (SVM):** I have used this method to predict the cluster and also calculated the cross value score of the model. And found **0.440689914218773 %** accuracy.
2. **Logistic regression**: This model also used to find prediction, and its efficiency is 0.3850620429664882%.
3. **Naive Bayes classifier:** The efficiency of this model is 0.08713381802499252%.
4. **K-Nearest Neighbor classifier:** This is the last model I have used to predict and found 0.38058519319445827% efficiency.

**Conclusion:**

Hence, I have chosen four models based on the understanding of the data from the preliminary analysis. Based on the cross-validation, none of them are useful because they have only less than 50% accuracy on predicted data. If I have to choose among them, I will select the SVM model because it has higher accuracy among them.



And finally, I have found cluster 77 is the best hotel based on the user information and booking data.