**1.Necessary information of the code:**

The python code contained in the file named “project\_code\_file.py” consists of nearly 300 lines of code. This code does the task of “Predicting hotel cluster groupings with user search queries” for the Expedia.com commercial web site. The entire code performs data manipulation in a systematic manner step by step basis as follows

**Line 17 to line 31**(These line numbers are with reference to the file “project\_code\_file.py):

Generating 20000 instances from training set (i.e. from file train .csv) with missing values into file ‘F20kwmiss.csv’. The actual training set consists of 3000693 instances which is too large.

**Line 37 to line 50:**

Generating 500 instances from test set (i.e. from file test .csv) with missing values into file ‘F500wmiss\_test.csv’. The actual testing set consists of 2528243 instances which is too large.

**Line 64 to line 93:**

The code of this region creates two different .csv files one for training & other for test data set namely 'nomisngval .csv' and ‘nomisngval\_test .csv’ respectively. These two files are generated from files ‘F20kwmiss.csv’ and‘F500wmiss\_test.csv’ respectively. The two newly created files at this region of code consists of records without missing values for their attributes. The file 'nomisngval .csv' has 11857 records with 24 attributes and the file ‘nomisngval\_test .csv’ has 281 instances with 22 attributes.

**Line 101 to line 143:**

In these lines of code, the actual training & test data set has been generated. The file 'train10k.csv' is generated from the file 'nomisngval .csv' and will act as training data set for the project, similarly the file 'test50.csv' is generated from the file ‘nomisngval\_test .csv’ and will act as testing data set for the project. These files consists of 10000 and 50 instances respectively.

**Line 156 to line 171:**

The code of this section computes the correlation between the attributes of the training data set.

**Line 173 to line 240:**

The code of this section eliminates string valued attributes from our training and test data set. The new files after elimination are ‘rmcoltrain10knoatt.csv’ & ‘rmcoltest50noatt.csv’

**Line 243 to line 306:**

In this section, the K-NN algorithm has been implemented for the data set contained in ‘rmcoltrain10knoatt.csv’ & ‘rmcoltest50noatt.csv’. Here, we can change the values of K to obtain different size groups of hotel cluster numbers .

**2.Software Platform Used:**

Used purely ‘Python’ platform to implement the project.

**3.Configuration:**

Used ‘Anaconda2(32-bit)’ , windows7- 32 bit version to implement the project. To run the project, simply open the file ‘project\_code\_file.py’ through python s/w

& run it. By default the program will run for K=1.

**4.Data sets:**

The data set used here in initially taken from ‘[www.kaggle.com](http://www.kaggle.com/)’ . The train and test sets are in the files namely ‘train.csv’ and ‘test.csv’.