* Data Modeling – Approach
* Main objective of the program is
  1. To identify schools (currently non-feeder schools) with academically qualified students, which would benefit from training programs. Currently, at least some students in these schools are taking SHSAT but without good results.(list1)
  2. To identify schools where the awareness of SHSAT is less and there are not many test takers. In such schools PASSNYC can do road shows and immersion programs to popularize the program and its benefits.(list2)
* For all schools,
  1. Need\_Training\_Score = 'number of offers received in a school' **/**

‘total number of students in the school',

with a lower score indicating a for training requirement for the school.

* + Need\_Awareness\_Score = 'number of test takers in the school' **/**

‘total number of students in the school' .

* Training\_Need\_Score and Awareness\_Need\_Score are later classified into
* 0 or 1 based on score value to make the target variable a classifier.

So for the two lists , either Need\_Training\_Score or Need\_Awareness\_Score would be the dependent variables and their classification is done based on the values of a set of feature variables in the data set.

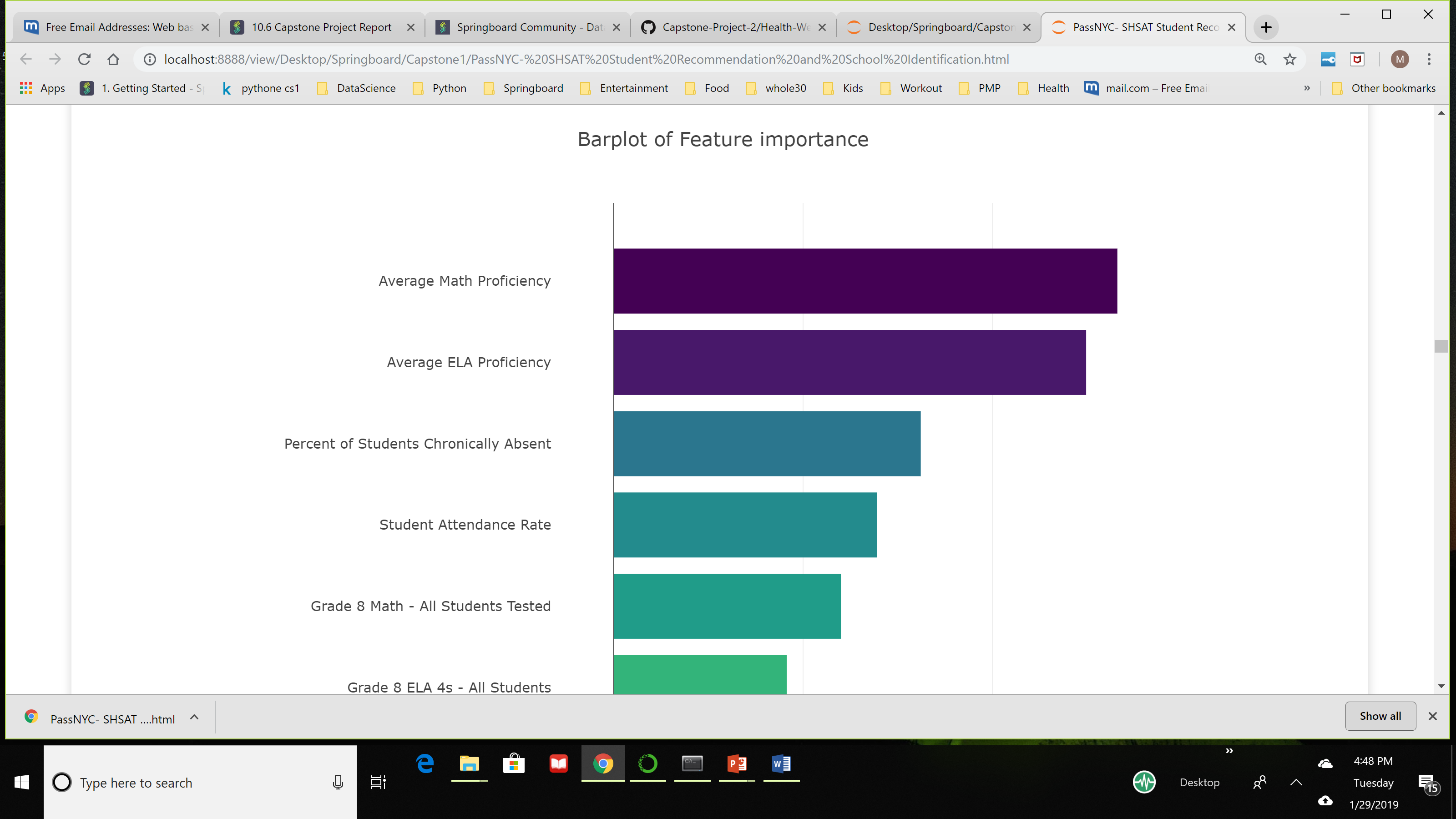
This dataset has close to 50 feature variables and plotting a co-relation for each of these variables would be impractical.

This is where an ensemble method like Random Forest can help.Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is very simple to use but at the same time very effective.

Feature Importance

With random forest algorithm, it is very easy to measure the relative importance of each feature on the prediction. It measures feature importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1.

From the feature importance score, we can decide which features to drop, and which ones to keep.



Using the Random Forest Classifier gives a Mean Absolute Error of 0.08 which is very minimal and the cross validation score is 0.9, which ensures a good prediction, as the model ensures a 91% accuracy in prediction.

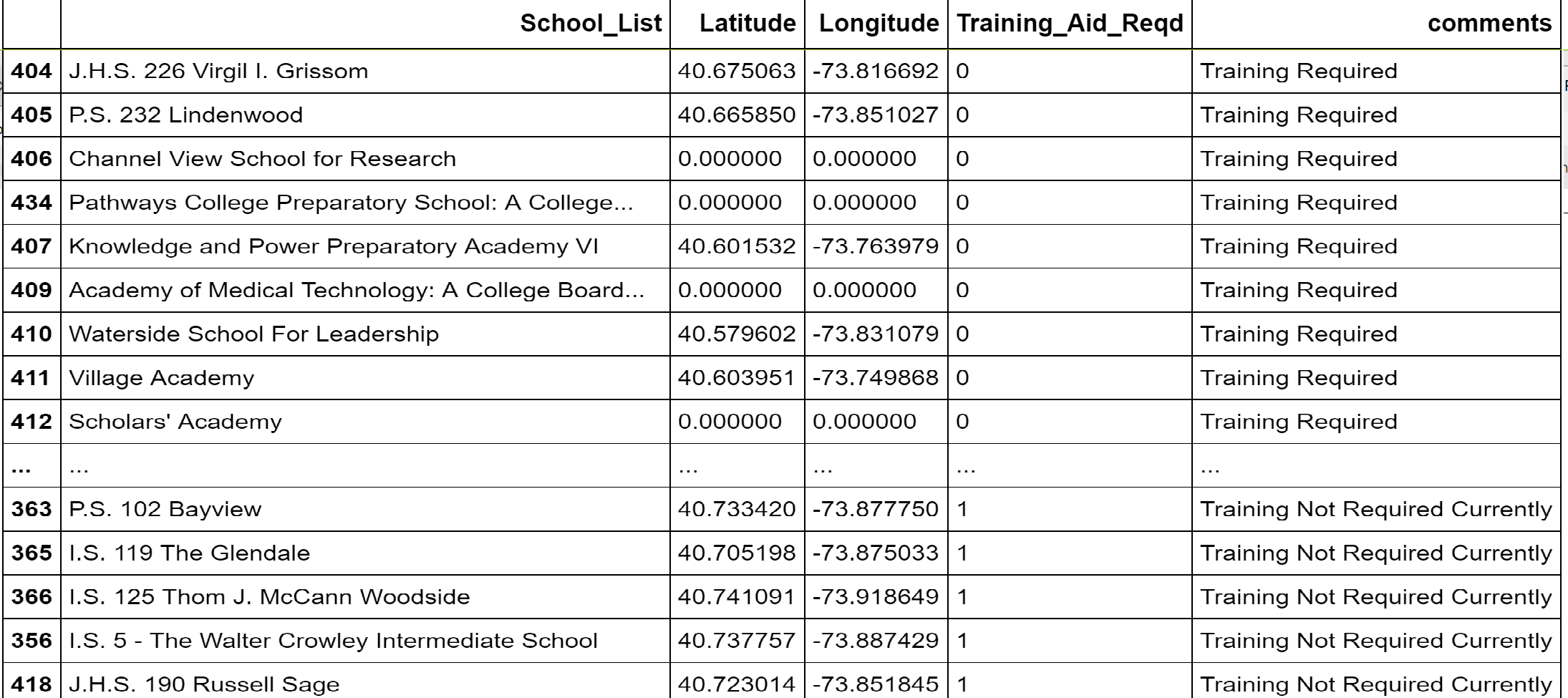
We can use this Classifier to create two separate list of schools that need help with

1. training and
2. awareness creation

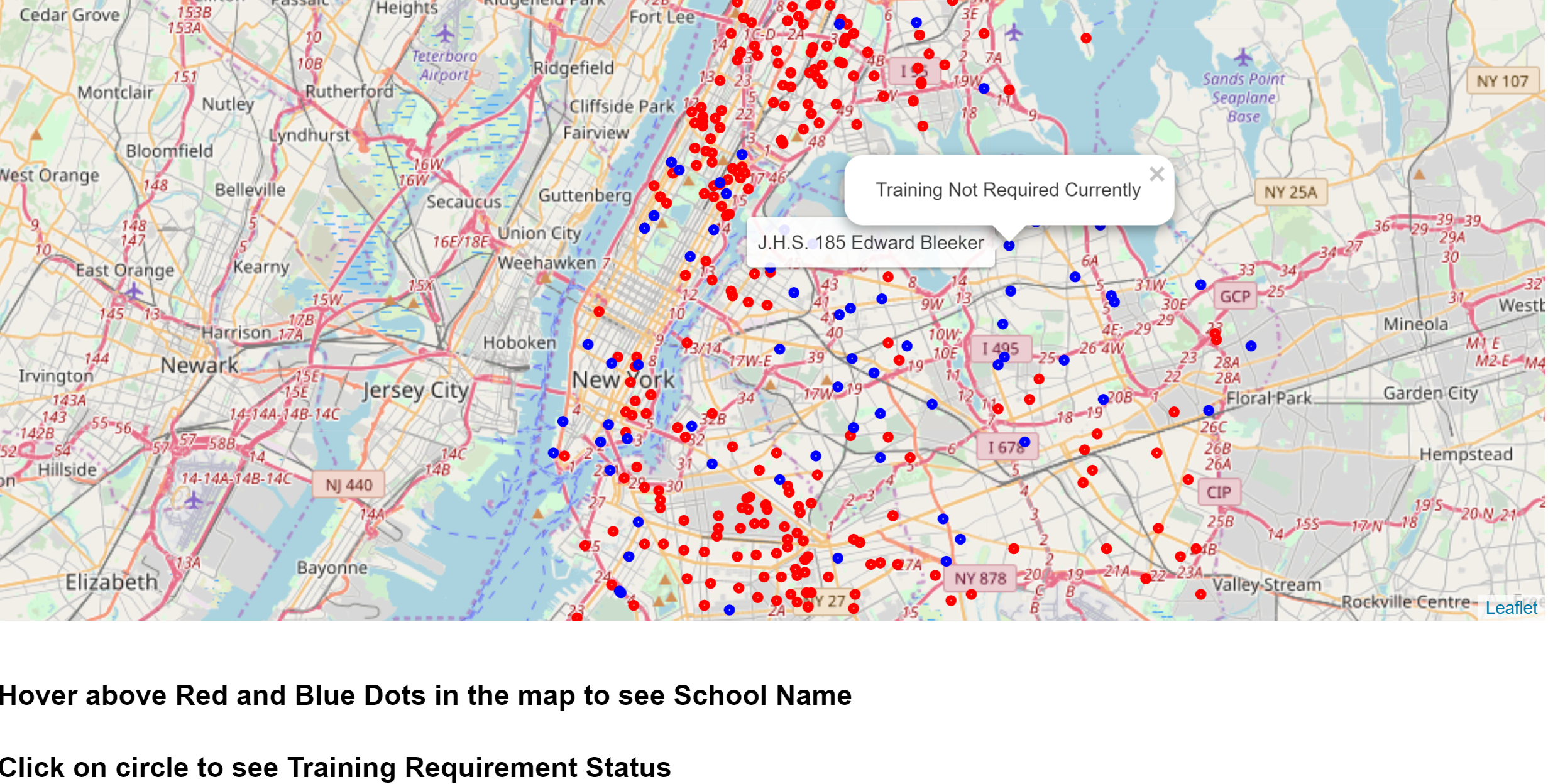
respectively, from the 'School Test' data.

Using this prediction model, we can predict whether any public middle school in NYC would make it to the list of PASSNYC's training\_required or Awareness\_required lists, if we have schools' students' performance data and other feature data. Below, I have taken a new data set (schools\_testdata\_df), that has not been used in training, for prediction.

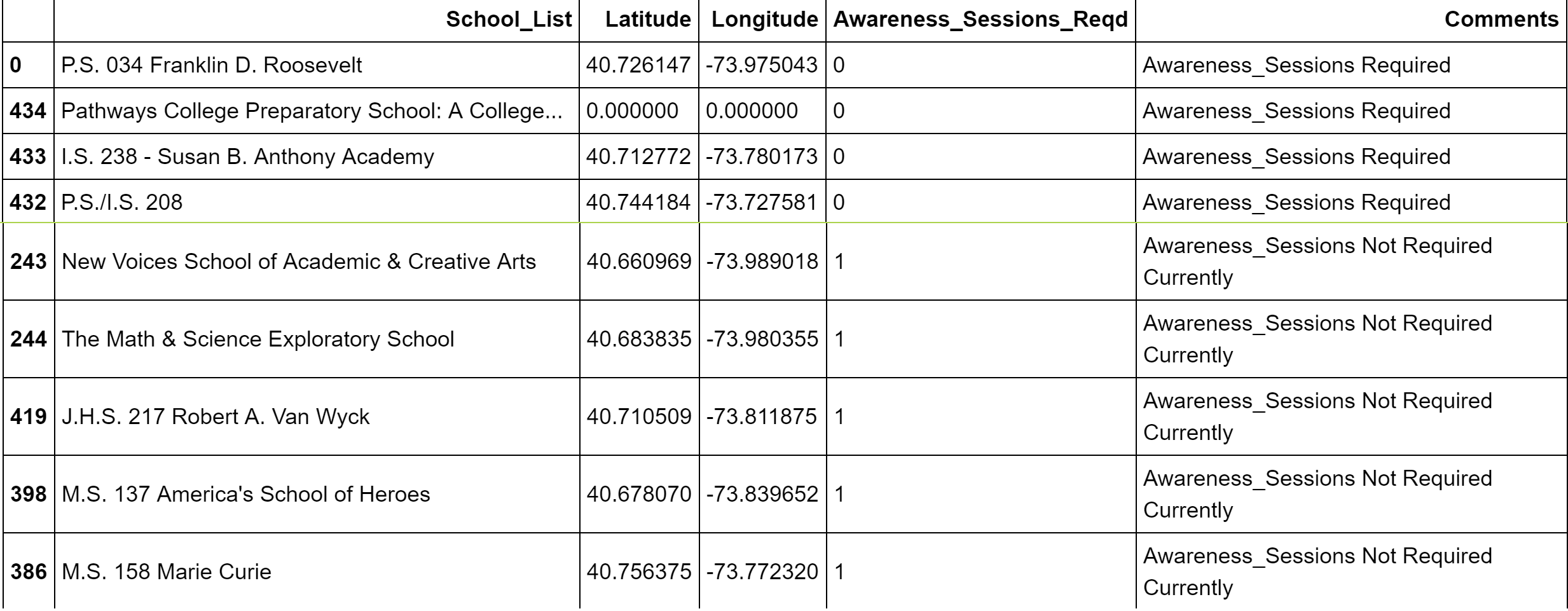
Training Required or not List



School list with option to display Training on OpenStreetMap



Awareness Sessions required List



School list with option to display Training on OpenStreetMap

