The main idea of clustering was to determine the similar areas in the city bases on the type of crime that occurred, the time when it occurred and whether it occurred on a weekend or a weekday. We have removed few unnecessary crimes which occurred very few number of times and put few important crimes in the “Other” crime category. Though these crimes occurred few times, they are major crimes.

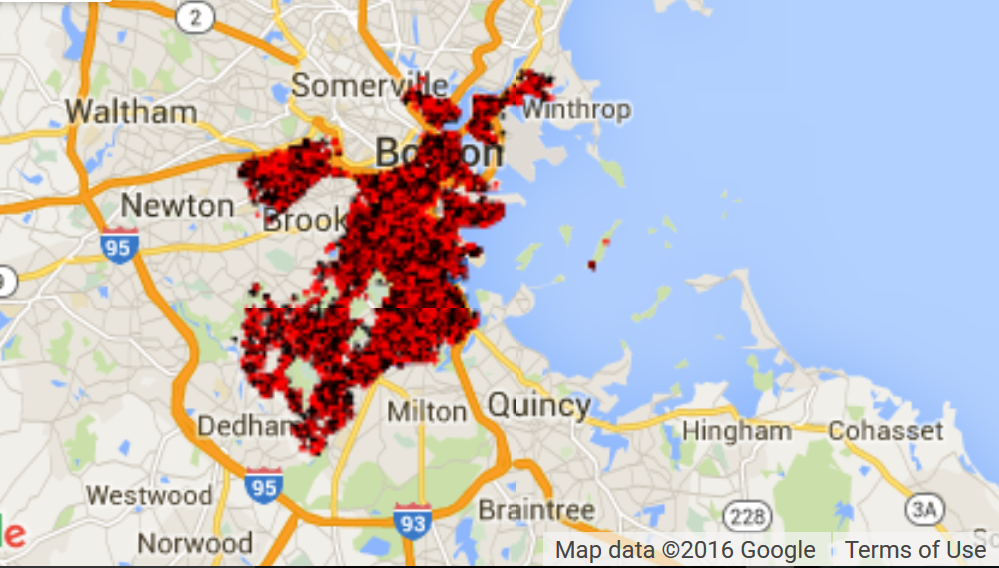
For clustering, we have used K-means++ algorithm. To determine the number of clusters, we computed the clustering and their associated error using k=1 through k=30. Using the output of the graph above, 15 is the number where error ceases to decrease by a significant amount. Hence, we have taken the number of clusters to be 15.

We copied the dataframe with original values of latitude and longitude, normalized latitude and longitude into a new dataframe. We then actual latitude and longitude into the data frame and also the kmeans label. Based on the kmeans label, we created subsets of dataframes and used gmplot to plot the clusters on the google map.

Below are the crime areas, where crimes occurred during the mid-night time 9pm – 12 am.

The black spots indicate the areas where crimes occurred during weekend.

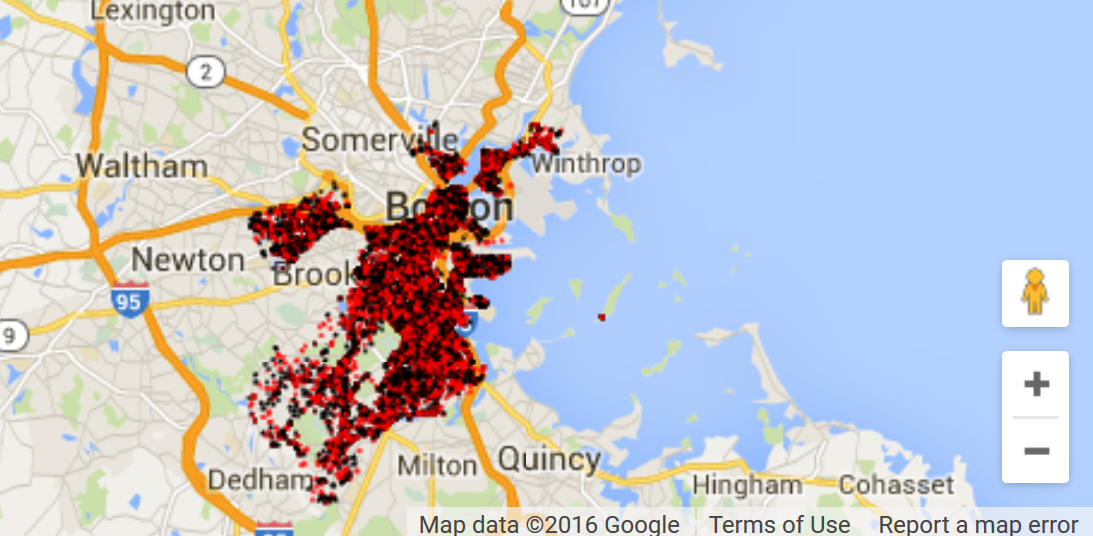
The red spots indicate the areas where crimes occurred during weekday.



Below are the crime areas, where crimes occurred during the early morning time 1am – 4 am.

The black spots indicate the areas where crimes occurred during weekend.

The red spots indicate the areas where crimes occurred during weekday.



The below areas have Assault crimes predominantly as compared to the other areas.

