Midterm Report: Predictive Analysis of City Crime Hotspots

Sahil Deshpande, Vinit Kumar
{*sahilsd, vinitn*}@*bu.edu*

***Abstract*—With increased digitization of data around the world, various type of data is now easily available. This includes criminal reports for most of the major cities in the US. In order to constructively use this data and the available data mining tools in Python, we are trying to analyze crime hotspots and predict next most likely crime location.**

I. Introduction

Crimes have become a common part of city life that seriously affect quality of life and economic growth of a society. Universal organizations are spending a lot of resources trying to identify safest and most dangerous cities to help local authorities manage their workforce. As more and more people shift to the city, this concentrated population makes it important to find safer places. Using the open-source data from official websites and some of the basic data mining approaches, we are attempting to help this process [1].

From past analysis of various cities and their neighbourhoods, it has been shown that certain parts of the city are more prone to criminal activities than others making them a criminal hotspot. Even though there doesn't seem to be an intuitive pattern for criminals, they tend to favor certain areas that makes it possible to predict such malicious activities. Using the information about past activities, law enforcements can effectively serve their duties.

To achieve this predictive analysis, we have looked at various datasets available for free on the Internet[2-3]. After some research on which cities have most relevant and most updated data, we have decided to work with Boston, Los Angeles and Raleigh. The first part of this project is to get the data in a specified format (in this case json) and parse it to extract meaningful information. Using these parsed data structures, we are interested in finding the most relevant crime types such as assault, robbery, hit and run, to name a few. We select the types that affect a particular area the most using regression techniques[4]. After trimming the dataset, we have applied k-means clustering to plot these various types of activities using GMplot[5]. We now plan to use classification methods in order to predict most vulnerable parts of the city. Later sections describe the results we've seen so far.

II. Technique

First we searched on the Internet for reliable data sources for criminal activities in major cities in US. After going through large number of datafiles, we decided to work with data from [3]. After getting the data in json format, we parsed it to get

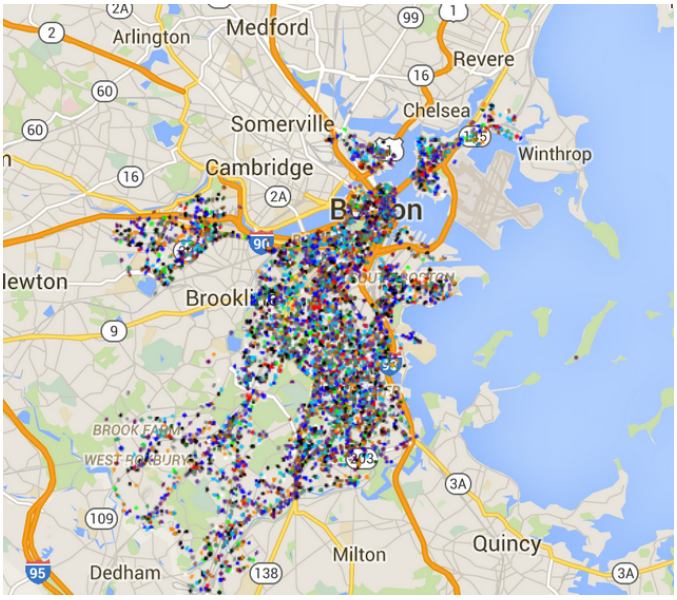

Fig. 1. Overall crime distribution across Boston shows concentrated crime data.

attributes such as crime location, latitude and longitude, time of crime, day of the week, type of the crime. Since these attributes help us distinguish between most of the important crime activities, we plan to use these attributes for further analysis.

While parsing the data, we have also divided the timestamps into six categories and day of the week as weekday or weekend. This further enables more meaningful clustering because coarse datasets are clustered more effectively.

In order to reduce this large data, have to used linear and logistic regression to find which crime types affect an area the most. Here we can calculate an average score for one area based on crime types (as we did in homework) and get top five coefficients that will denote the most common crime types. As we know, more data helps in better analysis and prediction we wanted to minimize this data reduction and thought this technique would be most effective.

We first categorize the data in types like assault, robbery, disturbance, white collar crime and others. But this can be easily replaced with the types we get from regression results so we are first focusing on analysis part. Using these categories, we have applied K-means clustering from Python sklearn. The clustering results are explained in the next section. We have normalized the latitude and longitude during clustering so that

```
--------------------------------
Assault : 12228
DrugAlcohol : 0
Othercrimes 0
PublicDisorder 0
Theft 0
WhiteCollarCrime 0
T1 926
T2 845
T3 2399
T4 3107
T5 2937
T6 2014
Weekday 12228
Weekend 0
--------------------------------
```

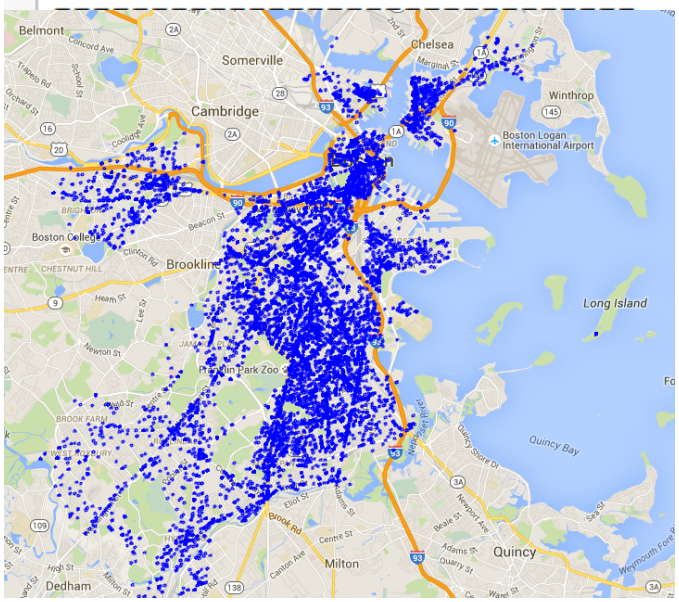

Fig. 2. Cluster showing only Assault crimes and only on weekdays.

they don't add any extra weight. While plotting the clusters on the map, we have used the original latitudes and longitudes to give correct crime hotspots.

So this has given us a basic idea of malicious activities and their patterns in Boston. We can easily identify which areas show more assaults or weekend crime. This example already starts to unfold the advantages after seeing which areas are relatively safer and which would need more attention of local law enforcement. Using this clustered data, we are now exploring classifier methods in python which were discussed in the class. Right now we have come up with two approaches on how we can predict a crime but this part needs more work. First approach is to directly use latitude/longitude for predicting most likely crime activity but this has very fine granularity which makes it less favorable. On the other hand,

```
(13, 12)
(13, 1)
                            OLS Regression Results
==============================================================================
Dep. Variable:            TotalCrimes   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 5.158e+09
Date:                Mon, 25 Apr 2016   Prob (F-statistic):           1.09e-05
Time:                        01:05:21   Log-Likelihood:                0.94924
No. Observations:                  13   AIC:                             22.10
Df Residuals:                       1   BIC:                             28.88
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Arrest         4.2771      0.086     49.758      0.013         3.185     5.369
Assault        2.5170      0.015    169.775      0.004         2.329     2.705
Burglary       2.6332      0.005    555.342      0.001         2.573     2.693
DrugAlochol    2.9349      0.008    381.251      0.002         2.837     3.033
Harassment    -7.6406      0.049   -156.432      0.004        -8.261    -7.020
Larcency       3.0397      0.001   4190.941      0.000         3.031     3.049
PublicDisorder 1.3134      0.027     48.773      0.013         0.971     1.656
Sex            2.8042      0.016    174.472      0.004         2.600     3.008
Theft          3.2933      0.005    615.779      0.001         3.225     3.361
Vandalism      4.2427      0.014    313.097      0.002         4.071     4.415
WhiteCollar    3.9988      0.008    471.484      0.001         3.891     4.107
Other          2.9259      0.001   2574.802      0.000         2.911     2.940
==============================================================================
Omnibus:                       35.803   Durbin-Watson:                   1.065
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.295
Skew:                           3.156   Prob(JB):                     5.96e-13
Kurtosis:                      11.006   Cond. No.                     3.67e+03
==============================================================================
```

Fig. 3. Linear Regression output for different types of crimes.

we are planning to utilize area codes in the data set. This divide the city in certain code like A4, G1, etc. We can use these and one of the six timestamps and try to get a tuple similar to, <type of crime, area code, time category, day category>. We think this will be more clear after the Amazon prediction assignment. We plan to revisit this once more.

III. Analysis of Crimes

A. Boston vs Los Angeles

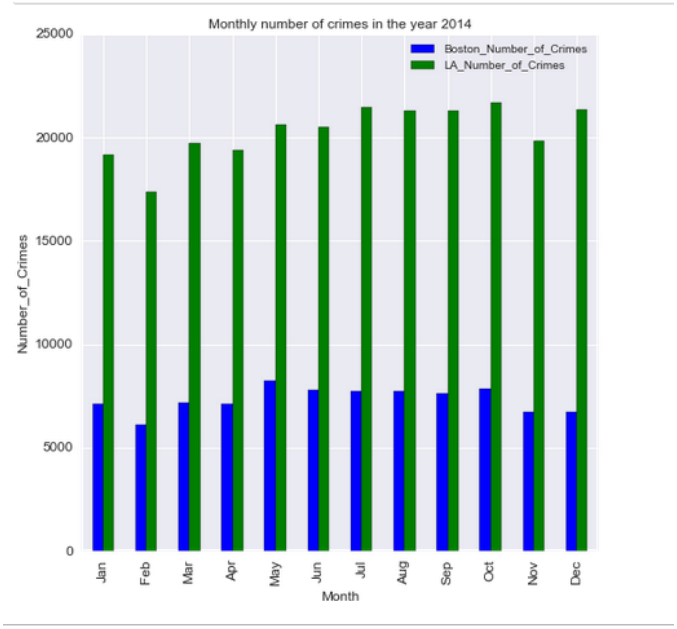

Fig. 4. Total number of crimes in Boston and Los Angeles.

This graph clearly shows that Boston is a much safer city to live in as compared to LA and for every month, the number of crimes experienced for both the cities is pretty much constant. Even though this is not normalized with population, this seems to give some idea about crimes occurring in these two cities. In the month of February, the peak winter probably shows a slight dip in the number of crimes.

B. Clustering

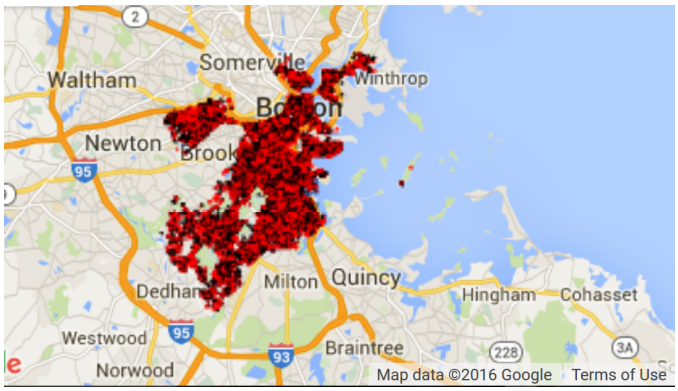

Fig. 5. Clustering data based on weekday and weekend crimes.

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.

C. Clustering

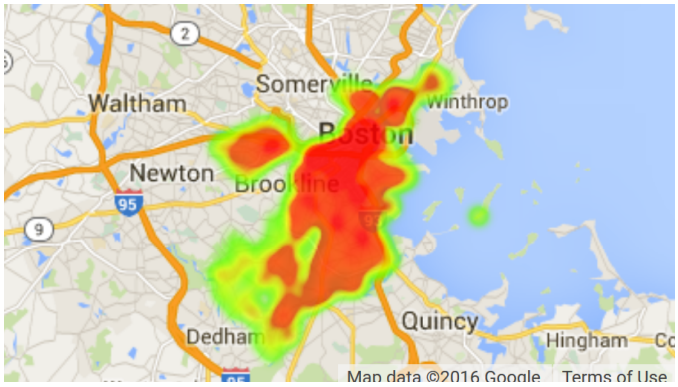

Fig. 6. Heatmap of Larceny and Assault crimes in Boston.

Above figure shows 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.

D. Regression

For Regression Analysis, we divided the data for each area code and got the total number of crime for each crime type, time of the day and type of the day. The aim of our regression analysis was to find to out what are the most important factors for the total number of crimes occurring in an area and what are the factors favorable for an area to be safe.

```
Optimization terminated successfully.
         Current function value: 0.250140
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:              Safe_Area   No. Observations:                   13
Model:                          Logit   Df Residuals:                        7
Method:                           MLE   Df Model:                            5
Date:                Mon, 25 Apr 2016   Pseudo R-squ.:                 0.5947
Time:                        01:05:27   Log-Likelihood:                -3.2518
converged:                       True   LL-Null:                       -8.0241
                                        LLR p-value:                   0.08921
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Arrest         0.1016      0.140      0.726      0.468        -0.173     0.376
Harassment     0.6262      0.641      0.977      0.328        -0.629     1.882
Theft         -0.0186      0.033     -0.564      0.573        -0.083     0.046
WhiteCollar   -0.0171      0.022     -0.769      0.442        -0.061     0.026
Larcency      -0.0036      0.004     -0.933      0.351        -0.011     0.004
Vandalism     -0.0027      0.013     -0.203      0.839        -0.028     0.023
==============================================================================
```

Fig. 7. Logistic Regression output for different types of crimes.

In Figure 3 we observe that the number of Harassment crimes has a negative correlation with the total number of crimes. This means the areas which have high number of Harassment crimes have a less total number of crimes. From Figure 4, we can see that for the safe areas the only crime type that occurs in a significant number is Harassment. Arrest and Vandalism have a high positive correlation as compared to other types. This implies the areas which have more Arrest and Vandalism crimes have more chances to be unsafe areas.

IV. Predicting Crimes

For prediction techniques, we explored many options and chose Naï"ve Bayes Classifier and Decision Tree Classifier. We found that these two are easiest to use with Python libraries and faster compared to other manual crime predicting algorithms. We predict type of crime for a given area, for a given day and a given time slot. This model is used by dividing the entire city data into training, validation and test data to compare and analyze accuracy of different classifiers.

A. Naïve Bayes Classifier

Nave Bayesian classifier is a supervised learning algorithm, which is effective and widely used. It is a statistical model that predicts class membership probabilities based on Bayes theorem . It assumes the independent effect between attribute values. While our selected crime features have an independent effect on each other, this classifier was an ideal choice. We constructed this model using ScikitLearn that provides a set of open source data-mining tools for Python. We applied Multinomial Naïve Bayes, which is used for multinomial distributed data that conforms to the categorical features in our datasets. The crime features contain (month, day, time, location) of the crime while we selected the crime type to represent the class label.

B. Decision Tree Classifier

Decision Tree classifier is our second method used supervised learning algorithm. It creates a model to predict the class label values by learning simple decision rules implied from the data features.We created this model for both datasets using ScikitLearn another library tool allocated for decision tree induction. To measure the quality of the split, we applied the entropy function for the information gain.

V. Results

The outcome of this project is two-fold. We first state all the results from analyzing dataset within Boston as well as comparison between Boston and Los Angeles. We also found some interesting patterns by this analysis. We then move on to discuss trends found using prediction techniques. We also compared the two classifiers for accuracy, execution time and ease of use.

A. Results from Analysis

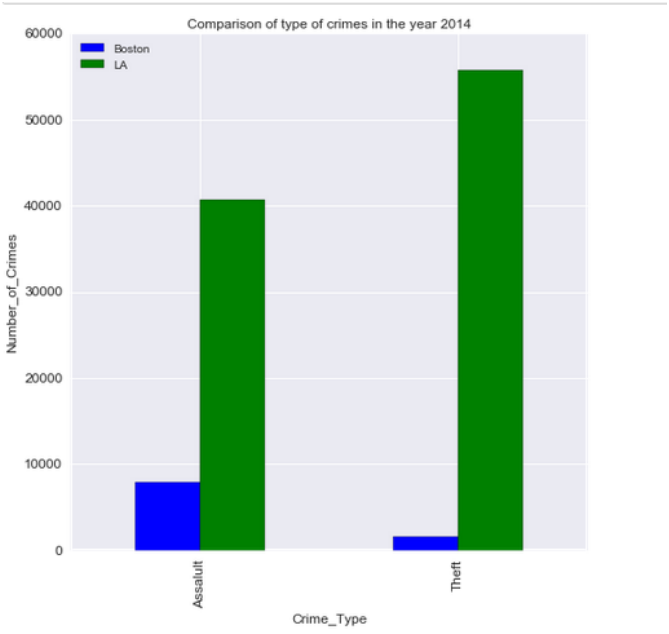

Fig. 8. Comparing Assault and Thefts in Boston and Los Angeles.

By looking at sheer number of crimes occurring in Boston and Los Angeles, we can say Boston is relatively safer city. We have taken 2 most common type of crimes in both the cities and plotted the above graph. It is clearly visible that LA has much more crime rate for these crimes as compared to Boston.

By looking at different timeslots, we can see that the time T5 has a higher correlation as compared to other times. This implies if an area has more number of crimes during the time period 5pm to 9pm, then that area has more chances to be unsafe.

```
(13, 6)
(13, 1)
                            OLS Regression Results
==============================================================================
Dep. Variable:            TotalCrimes   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 6.676e+05
Date:                Mon, 25 Apr 2016   Prob (F-statistic):           8.73e-20
Time:                        01:05:22   Log-Likelihood:                -74.395
No. Observations:                  13   AIC:                             160.8
Df Residuals:                       7   BIC:                             164.2
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
T1             3.1397      0.103     30.625      0.000         2.897     3.382
T2             2.9730      0.145     20.439      0.000         2.629     3.317
T3             3.0901      0.108     28.735      0.000         2.836     3.344
T4             2.5674      0.176     14.608      0.000         2.152     2.983
T5             3.6401      0.156     23.294      0.000         3.271     4.010
T6             2.5717      0.159     16.208      0.000         2.197     2.947
==============================================================================
Omnibus:                        1.869   Durbin-Watson:                   2.337
Prob(Omnibus):                  0.393   Jarque-Bera (JB):                0.881
Skew:                           0.052   Prob(JB):                        0.644
Kurtosis:                       1.729   Cond. No.                         69.2
==============================================================================
```

Fig. 9. Finding predominant timeslot using regression.

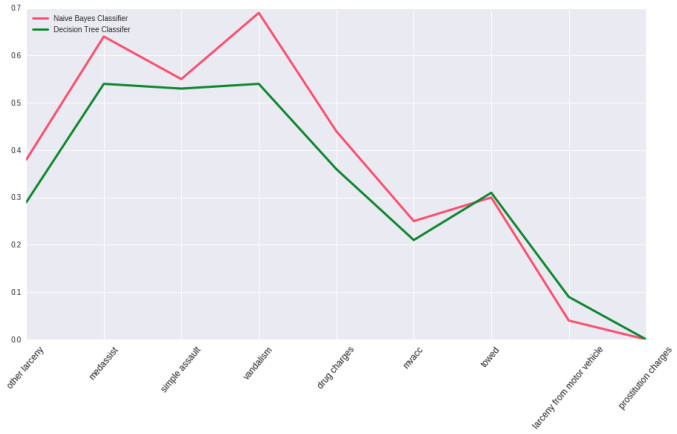

Fig. 10. Output accuracy comparison for NB and DT classifiers.

B. Results from Prediction

Regarding different classifiers, the Naïve Bayesian classifier, it achieves an accuracy of 51% in Boston crime prediction while it reaches 54% for Los Angeles crime prediction. On the other hand, decision tree classifier reports less prediction accuracy with 42 % for Boston and 43% for Los Angeles. Moreover, the decision tree model created a very complex tree that cannot generalize the data for both cities. However, the two classifiers have the same performance in terms of their running time. Above figure also shows trends for different districts within Boston. Each point stands for a prediction in certain area for certain day for certain type of crime. These seven lines passing through a single point says that type of crime is predominant in corresponding area irrespective of day of the week. E18 and C6 districts fir this description.

VI. Conclusion

We have now completed collecting required data, parsed it to extract required information for Boston. We have then used various analysis methods for finding patterns in criminal activities. Using different classifiers, we've also tried to use these patterns for predicting crimes. The primary are encouraging enough to pursue further techniques to get more insight on most likely next crime. We also found which would be the

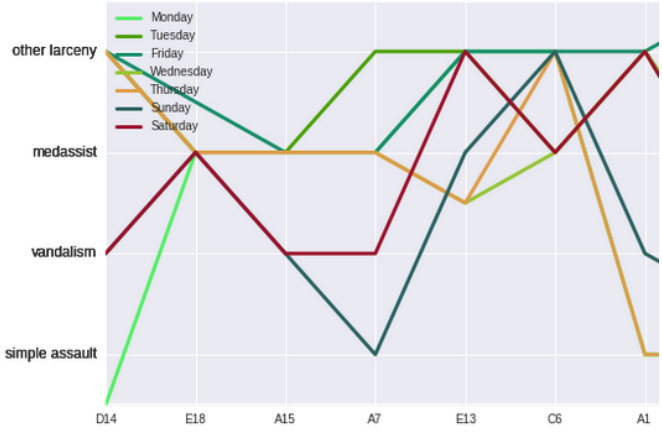

Fig. 11. Per day trends for different areas and different crime types.

:

	AreaCode	SafePercentage	Safe_Area
0	A1	88.800052	0
1	A15	97.901372	1
2	A7	95.132359	1
3	B2	85.222284	0
4	B3	91.269420	0
5	C11	87.115265	0
6	C6	92.166446	0
7	D14	92.661242	0
8	D4	85.145783	0
9	E13	94.365433	0
10	E18	94.781819	0
11	E5	95.441388	1
12	HTU	99.997137	1

Fig. 12. Boston safe areas.

safe areas in Boston. One such area is 'HTU which is the safest of all areas in Boston. This consists of streets PARK PLAZA, BRAGDON ST, STUART ST, WASHINGTON ST and COMMONWEALTH AV.

As of now we can conclude from the Logistic Regression results, we can see that for the safe areas the only crime type that occurs in a significant number is Harassment. Also, Boston looks to be a safer city for majority of crime types in terms of frequency. But this comparison will be more accurate if we can normalize it based on population density.

Naïve Bayes Classifier performs more accurately than Decision Tree Classifier. This accuracy can be further increased by more social data discussed in future work.

VII. Future Work

We believe the analysis will be more accurate for multi-city data if we consider population density. Since number of crimes will vary with this factor, our analysis is not yet complete. Also, the prediction accuracy can be improved by applying external data trends. We can also see if average income, age distribution and other social aspects affect the rate and type of crimes. This can aid our prediction model while predicting crimes in specific parts of a city.

References

[1] CRIME PREDICTION BASED ON CRIME TYPES AND USING SPATIAL AND TEMPORAL CRIMINAL HOTSPOTS Tahani Almanie, Rsha Mirza and Elizabeth Lor Department of Computer Science, University of Colorado, Boulder, USA.
[2] Crimereports.com, 2015. [Online]. Available: https://www.crimereports.com. [Accessed: 20- May-2015].
[3] 'Crime Datasets - US City Open Data Census', Us-city.census.okfn.org, 2015. [Online]. Available: http://us-city.census.okfn.org/dataset/crime-stats. [Accessed: 20- May- 2015].
[4] Regression techniques reference. http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
[5] GMPlot reference. https://pypi.python.org/pypi/gmplot/1.0.5
[6] GitHub code repository. https://github.com/sahilsd/591-crime-project