CRIMES IN BOSTON | A STUDY OF TRENDS FROM 2015 TO 2018

INSIGHTS INTO BOSTON CRIME

Boston's crime index rating of 18 suggests that it is safer than only 18% of the cities in United States. The number is startling considering that 89.66% of Boston's workforce is employed in white-collar jobs, which is well above the national average. Further, compared to 95% of the cities in the USA, Boston has more people who are trained to use computers ("Boston, MA Real Estate & Demographic Data"). Boston is one of the major cities of the USA and houses a population of diverse ethnic backgrounds. Further, it is considered the international center of higher education. However, the crime numbers recorded within the city are extremely high and stand at conflict with Boston's economic significance within the country.

Considering the nature of crime in Boston, several studies have been conducted to compare the city's demographics with its crime rate. Further, various statistics pertaining to Boston's crime incidents have been compared with the country's average figures at numerous instances. The studies specifically highlight the fact that crime rate in Boston is considerably higher than the national average, and Boston has a higher proportion of violent crimes per 1000 crimes ("Boston, MA Real Estate & Demographic Data").

There is high variability in the crime scenario in Boston, which directs towards a dire need to analyze and validate it crime trends. The figures require that the crime numbers be studied and analyzed against related factors. Accordingly, the city's crime data from 2015 to 2018 must be examined and scrutinized to understand the existence of crime in Boston and explain its dependence of a varying number of factors.

Such analysis is essential since it shall lend insights into the possible dependence of crime upon factors such as zip code, time, day of week, income per household, and ethnic division of population. It shall aid in understanding the significance of underlying factors in determining and predicting the crime rates within various regions of Boston and designing prevention strategies accordingly.

COMPUTATIONAL SETUP OF THE ANALYSIS

The study of Boston's crime data starts with an exploratory analysis, which uses visual methods such as box plots, histograms and scatter plots to provide a snapshot of the relationships between various groups of independent and dependent variables. Post the exploratory analysis, analytical techniques such as regression analysis and K-means clustering further help identify possible correlations and identify groups organically.

ANALYSIS OF CRIME IN BOSTON: STEP BY STEP ACCOUNT

The data relating to crime incidences recorded in Boston between 2015 and 2018 was obtained from Kaggle ("Crimes in Boston | Kaggle"), an extensive data resource reputed especially for its analytics competitions.

This dataset comprised of 303,372 observations for 17 variables including offense description, location, day, date and time of crime, et cetera. Prior to initiating the analysis of such variables, data points with missing location values were removed to create a consistency within the dataset. Subsequently, using the available information regarding the latitude and longitude of an incident, unique coordinates were identified to find zip codes through the zips library. The zip code data was then merged into the original dataset based on the location of each crime.

Following the data cleaning exercise, various data analytics techniques were applied to comprehend the data and identify the trends portrayed by the same. Such analysis can be bifurcated into two broad categories:

a. Analysis of the crime information contained within the dataset: using the Seaborn library, various pair plots of crime incidents were plotted against other particulars available in the dataset to identify existing trends. Using the information derived from such preliminary descriptive statistics, data relating to the location (i.e., zip code), date and time of the crime was chosen for an initial regression model.

b. Analysis of the dataset in conjunction with demographic information imported from external sources: the initial regression model suggested that the location (i.e., zip codes) of crime could significantly explain the occurrence of crime incidents. Therefore, to further understand the relationship between the occurrence of a crime and the related location, demographic information with respect to each zip code, including income, population characteristics, et cetera, was extracted from an external database, CDX, and merged into the original dataset to further understand the trends depicted by the crime data.

The analysis of the crime data of Boston consisted of the following:

DESCRIPTIVE ANALYSIS

a. Pair plots: crime incidents and crime rate versus population, number of businesses, and income per household;

b. Box plots: crime incidents versus districts, day of week and hour of day.

INFERENTIAL ANALYSIS

a. To further understand the relationship between incidence of crime in Boston and the zip code, the data was grouped by location. Subsequently, a regression model was developed to identify the relationship between crime incidents and the demographic information obtained by zip-code.

b. The first regression model analyzed the impact of all demographic parameters as independent variables to explain the variation in the dependent variable, i.e., number of crime incidents.

c. As a next step, the regression model was refined to obtain the best coefficient of determination with minimum number of independent variables by using step-wise/forward/backward selection model.

d. Also, an independent analysis on the data set was performed by using K-means clustering process. The data was categorized in different clusters based on the zip demographic information, and the average crime rates were found for each of these clusters and compared against each other.

e. Using such clusters, another regression was run to analyze the impact of the different location clusters on crime incidence.

DATA VISUALIZATION: MAP OF BOSTON

a. Using the Folium package and MarkerCluster plugin, a map of the city of Boston was created, and the number of crime incidents were recorded at each related latitude and longitude position.

b. Using the same Folium package, a heat map of the Boston city was created to show the intensity of the number of crime incidents within different regions of the city.

COMPUTATIONAL CHALLENGES AND BOTTLENECK

a. The Pandas library within Jupyter Notebook did not recognize the encoding in the original dataset (i.e., 'ISO-8859-1'), unless explicitly mentioned.

b. Geographical indicators contained within the original dataset comprised only of latitude and longitude information of the crime locations. To include the respective zip codes within the main file, the 'ZipcodeSearchEngine' class of the 'uszipcode' package was called upon within the code. The algorithm involved in this process is not a simple lookup and involves complex computations before the zip-code can be identified. Such compilation of data, hence, presented a major bottleneck in finding zip-codes for the entire dataset of approximately 287,000 data points. The code took around 15 minutes to populate the zip-code information for the entire data-set.

However, this bottleneck was resolved by creating a new data-frame for unique latitude and longitude information and running the 'ZipcodeSearchEngine' only for this data-frame. The execution time was reduced to 16 seconds. Nonetheless, this remains the slowest part of our code due to complex computations involved therein.

EXECUTION TIME & MEMORY USAGE

The entire code approximately takes a total of 150 seconds to provide the results. Zip codes are obtained for the entire dataset in 16 seconds, data visualizations, including hover chart and heat map, takes approximately 90 seconds, and clustering requires about 10 seconds.

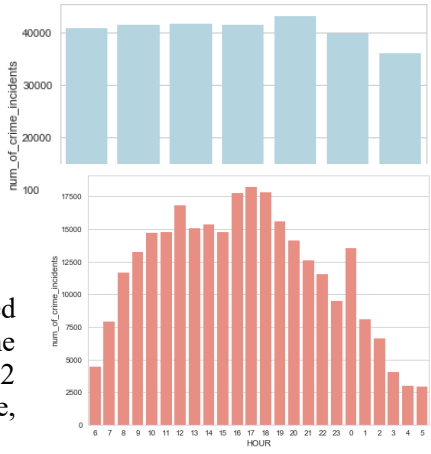
However, as the size of data increases, the time required and memory usage to execute the code increases significantly. Time and memory usage was measured for regression and clustering analysis, i.e., the major components of the code. The time and memory increase linearly with the increase in the data size (as depicted in Figures 1 and 2, and Tables 1 and 2 in the Appendix).

RESULTS

The preliminary descriptive analysis of the data provided various insights:

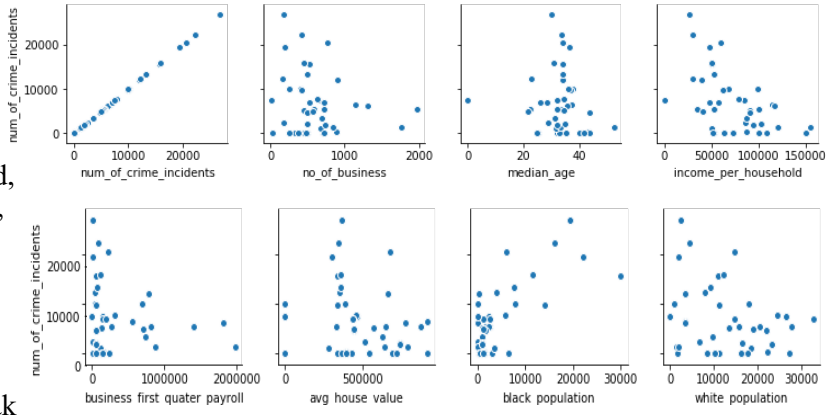
a. Over the 7 days of the week, the number of crimes recorded remain at a fairly similar level, with only a slight increase on Friday followed by a decrease on Saturday and Sunday.

b. The number of crime incidents recorded over the 24 hours of a day vary in an asymmetrical manner. From 6 AM and onwards, the crime incidents gradually increase till 11 AM, followed by a spike at 12 noon. Subsequently, a high crime period is noticed between 4 PM and 6 PM, which corresponds to a rush hour, and has a majority of road related incidents. Moving through the day, the number of crime incidents per hour keep decreasing, with a sharp increase at 12 midnight. Such increase arises since this is a rather quiet time, and crimes may be committed with a scarcity of witnesses.



c. 6 out of the 42 zip codes in Boston contribute more than 50% of the total crime incidents recorded in Boston.

d. The pair plots on the right provide the relationship between number of crime incidences and other variables including number of businesses, median age, income per household, business first quarter payroll, average house value, and black and white populations.

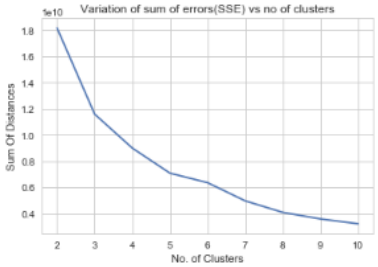


These plots suggest that the number of crimes (grouped by zip codes) have a weak negative correlation with income per household, and a weak positive correlation with black population. A significant relationship is not noticed for the other variables.

e. Following from the above exploratory analysis, an initial regression was run for the number of crimes against three categorical variables (including day of week, hour of day, and zip code). This provided a coefficient of determination (adj-R^2) of 81.5%, which suggested a significant impact of the included independent variables on the number of crime incidents.

f. Based on the results of the initial regression, several regressions were run for the number of crimes and a list of different population attributes. Following such regressions and the removal of attributes that did not suggest a strong economic and statistical significance, a final regression was run for crime incidences against day of the week, hour of the day, and certain zip code attributes (including male population, households per zip, income per household, and different ethnic groups contained within the population). Such regression returned an adjusted R^2 of 65.1% and suggested that the variables included were significant. Such value of the R^2 can further be increased if additional information relating to the specific populations (such as education levels, unemployment rates, et cetera) within each zip code becomes available.

g. Based on the results of the above linear regression, a cluster analysis was run on the significant variables contained in the regression model. This exercise was aimed at grouping zip codes having similar characteristics in relation to the included variables. Using the elbow method, an appropriate number of clusters is found to be 3.



The results of the cluster analysis are as provided in Table 2 in the appendix. This table supports the results from the regression that the crime indigence is positively correlated with number of household and Hawaiian and Indian populations, and negatively correlated with income per household, and white and black population.

CONCLUSIONS

The regression analysis gives an insight into the significance of different factors in determining the crime numbers of Boston. As described under the Results section, zip code attributes such as race-wise population statistic, income statistics, et cetera can explain 65.3% of the variation in the number of crime incidents. Further, the date and time information in the dataset depict the periodicity of crime incidents within the city.

Such results provide a beneficial outlook to the problem at hand, and aid in coming up with various measures that can be used to curb the crime numbers in Boston. Depending on the specific requirements of involved stakeholders, this information can be used to serve multiple purposes. For example, the Boston Police Department may be interested in the insights to implement effective crime control policies. The department may use the analysis to understand the causal relationships between number of crimes committed and demographic attributes of the zip code area. For example, as the analysis suggests, areas with low Hawaiian and Indian populations and high-income levels can be considered the safer places of Boston. Similarly, the results of the cluster analysis performed on data points grouped by zip codes can help the police identify the problem areas and concentrate their forces within the same.

Similarly, such insights can also help the business community to improve its security based on the crime trends in an area. The insights can also help real estate investors to decide which areas to invest in Boston considering safety and cost as factors.

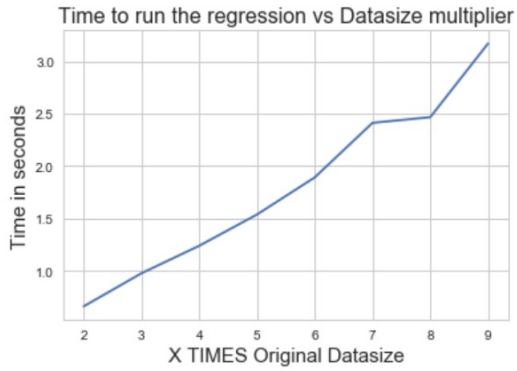
FUTURE SCOPE OF THE ANALYSIS

To further the current study, more granular population and regional attributes (including race-wise educational, employment and income levels) can be considered to analyze the related impact of crime in Boston. To further validate the relationship between number of crime incidents and the population within an area of Boston, a separate detailed analysis may also be conducted to factor in the impact of the presence of gangs in such regions. Such investigation may help substantiate the claims made based on the population mix in crime-rich areas.

Additionally, data on alcohol consumption and drug abuse in Boston can also be helpful in increasing the degree of explanation of the number of crime incidents in Boston.

APPENDIX

Figure 1:



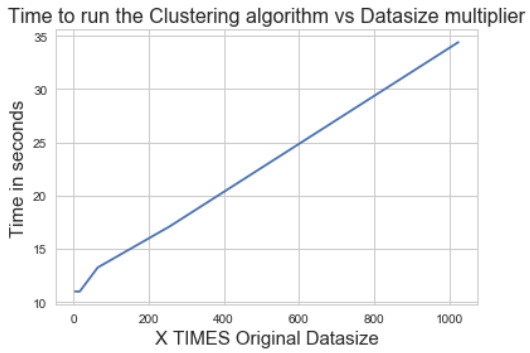


Table 1:

Memory Usage for Regression		
Dataset Size	Peak Memory (MiB)	Time Taken (seconds)
567,742	1,361.05	1.169
851,613	1,626.62	1.539
1,135,484	1,687.41	2.298
1,419,355	1,675.50	2.647
1,703,226	1,859.27	3.196
1,987,097	1,959.91	3.782
2,270,968	2,131.64	4.412
2,554,839	2,131.14	4.836

Memory Usage for Clustering		
Dataset Size	Peak Memory (MiB)	Time Taken (seconds)
164	1,538.96	10.99
656	1,594.65	10.98
2,624	1,547.79	13.24
10,496	1,648.1	17.11
41,984	1,735.38	34.39

Table 2:

Cluster	Number of Zip Codes in the Cluster	Number of Crime Incidents	Aggregate Population	Hawaiian Population Proportion	Indian Population Proportion	Income per Household ($ 000's)	Average Number of Households per Zip
0	5	3,098.00	6,102.80	0.0009	0.0040	131.76	2,518.40
1	21	9,951.29	27,160.38	0.0032	0.0117	48.98	10,387.71
2	15	3,401.20	19,275.47	0.0014	0.0061	90.58	8,626.53

WORKS CITED

"Boston, MA Real Estate & Demographic Data." *Bridgeport, CT Demographics and Population Statistics - NeighborhoodScout*, www.neighborhoodscout.com/ma/boston.

Jain, Ankur. "Crimes in Boston | Kaggle." *Countries of the World | Kaggle*, 11 July 2018, www.kaggle.com/ankkur13/boston-crime-data.