Spring 3-23-2018

APPLICATION OF DATA ANALYTICS TECHNIQUES IN ANALYZING CRIMES

Christian Sunday Nwankwo
Georgia Southern University, cn02099@georgiasouthern.edu

Majeed Kayode Raji
Georgia Southern University, mr07157@georgiasouthern.edu

Etinosa Sharon Oghogho
Florida International University, eogho001@fiu.edu

APPLICATION OF DATA ANALYTICS TECHNIQUES IN ANALYZING CRIMES

Christian Sunday Nwankwo
Georgia Southern University
cn02099@georgiasouthern.edu

Majeed Kayode Raji
Georgia Southern University
mr07157@georgiasouthern.edu

Etinosa Sharon Oghogho
Florida International University
eogho001@fiu.edu

ABSTRACT

Over the past decades, data analytics have played a very important role in knowledge discovery and making decisions in different situations. Proper application of data analytics tools and techniques will help to identify important patterns and relationships in datasets. For this project, we conducted crime analysis and investigation using the Chicago crime dataset to gain more information and insights on why, when and how criminal activities are carried out. This project also focuses on identifying the relationship of socioeconomic indicators like poverty rate and unemployment to crime. We have used data analytics software like Microsoft excel, Tableau and Python for analyzing, manipulating, visualizing and implementing linear regression all geared towards achieving the goal of this project. This is an important study which can help Law enforcement agencies speed up the process of solving crimes and predict the possibility of crimes in the future.

Keywords: data analysis, socioeconomic indicators, crimes

INTRODUCTION

It is known that crime is neither orderly nor totally random. It drifts with phases of human behavior, but some places attract crime naturally. It is of upmost importance to discover the spatial and progressive patterns for a better understanding of crime events. By using data that consists of geo-data and time series data, hands-on crime prevention explanations can be established that tally with specific places and times.

This project focuses on discovering criminal patterns in the Chicago crimes dataset from 2008 to 2012. We aim to identify and unravel very important and useful information that will be very beneficial and useful for Law enforcement agencies and help in making business decisions in Chicago. Some of the valuable information will include; areas in Chicago communities with high rate of criminal activities, the type of crimes that have been committed the most and the distribution of crimes over a period.

We have also gotten a socioeconomic indicator dataset that gives us information about the livelihood of people in every community of Chicago. Some of the attributes in this dataset includes the Percentage of House Crowding, the rate of Unemployment and the Hardship Index of every community. We have been able to generate two new datasets, one which consists of the crimes data and the socioeconomic data and the second one consists of the socioeconomic data with a new column, crime rate.

With this merger, we plan to identify the relationship between criminal activities and the living condition in each community area. We will be applying linear regression on this dataset to find important correlations between the predictor variables (the socioeconomic indicators) and the response variable (crime rate). Being able to achieve a level of significance in research will help in understanding how to solve and reduce the rate of criminal activities in Chicago.

LITERATURE REVIEW

The paper by (Thongsatapornwatana, 2016), “A Survey of Data Mining Techniques for Analyzing Crime Patterns” examined data mining methods for analyzing crime patterns over the years. The aim of the paper was to review several literatures on diverse data mining applications especially applications that can be applied to solving crimes. Data mining techniques explored by the author includes, Association Rule Mining, Clustering and Classification. The paper focused on certain crime types such as; Traffic Violation and Border Control, Violent Crime, The Narcotics, and Cyber Crime. Issues that exist in data analysis includes the collection of large and unstructured data that is difficult to prepare, transform and integrate; the changes in crime pattern that has led to unpredictability, the challenge of performance which involves the developing and detecting the algorithms to increase the time detection accuracy, the difficulty in visualizing and displaying the data due to it large size. In conclusion, the changes and rise of crime has led to understanding the crime behavior, crime prediction, accurate crime detection, and managing large volumes of data obtained from various sources.

The paper by (Rasheed et al., 2015), "A Tool for analysis and visualization of Criminal Networks" proposed a demonstration tool called the PEVNET tool. The aim of the paper was to carry out a comprehensive demonstration of the PEVNET tool. The tool is used to support crime analysts in analyzing the intra-network criminal activities. The study centered on certain aspects of criminal network analysis and discovered definite issues. The authors mentioned that by resolving the issues and providing analysts with a more analytical interface, support for decision making can be improved to a vast extent. PEVNET is a transformation of crime investigation from the traditional live analysis on whiteboard to dynamic visualization, and this tool displays its utilization with the viewpoint of visualization. The investigators provided a case study of how analysts may gain from the tools, and how it can develop into a prediction tool in investigative analysis. The analysts are provided with advanced drag and drop facilities. In conclusion, this study showed a criminal network visualization tool. With diverse network visualizations and clustering features, the crime analysts are provided with improved support for decision making. Utilizing the proposed qualities would enable a more dynamic criminal network analysis as it would afford the analysts more time to focus on consequent concerns rather than in the past where they had to deal with difficulties.

(Ying, 2016) paper titled, "Analysis of Crime Factors Correlation Based on Data Mining Technology" focused on how the data mining technology can be used in public service to improve crime prevention and control, information science and the construction of public security. Description and prediction are the categories of data mining. Methods of data mining includes; Association analysis which identifies the associated networks and rules hidden in databases, Analysis and Prediction; which helps to predicts trends of the data, cluster analysis that is based on the clustering features of things in order discover laws and typical patterns, and Outlier mining used for mining data that are inconsistent with the general behaviors of the models. The paper also discussed how in public security, data mining can be applied in areas such as file system of the population, motor vehicle registration and criminal databases for case investigation, as well as the field of crime prevention and control. Several challenges exist in the application of data mining in the public security business, and they include; the problem of storing and maintaining large scale dataset, the issue of using a standard structured query language in unstructured data, and serving as a threat to people's privacy and data security. In summary, this research provides a rational mining mode and satisfactory outcomes from the methods employed.

In the paper, "Cluster Analysis for Reducing City Crime Rates", the authors, (Alkhaibari et al., 2017) made use of different clustering algorithms such as K-Means clustering and agglomerative clustering. K-Means clustering algorithm is an iterative clustering algorithm. Agglomerative clustering is one category of hierarchical clustering algorithms; the other category of hierarchical clustering algorithms is called divisive clustering. These clustering algorithms were studied and applied to the New York Police Department (NYPD), for analyzing the location of the crime and stopped people using the reason to reduce city crime rates. Several clustering techniques were applied to get the most suited outcomes that can aid police officers to enhance their work. Clustering helps identify various clusters discovered in databases using diverse procedures of interestingness. First, the investigators organized the selected features to use them in creating the clusters, and then used some measures to determine the optimal number of clusters for each algorithm. Different clusters were created by using different methods. Lastly, several visualization techniques were used to represent the clusters' profiles. Findings from this study revealed that K-Means algorithm is the best clustering algorithm and that good features are paramount to ensuring that the models are beneficial.

The paper by (Cozens, 2008) on "Public health and the potential benefits of Crime Prevention Through Environmental Design". Is about how crime can be prevented through environmental design, Crime and the fright of crime are not evenly dispersed throughout the city either spatially or temporally, and the notion of 'hot spots' of crime (where/when crime and/or the fear of crime are highly concentrated) has received rising attentiveness in recent years. Crime and public health issues have similar roots and effects can be enhanced using similar methods (e.g. improving socio-economic conditions and enhancing social capital). The design and structure of the built environment on crime and public health depicts it can serve as an effective planning tool. However, communities that feel unsafe are more likely to be inactive residents. This paper argues that Crime Prevention Through Environmental Design (CPTED) has possible public health advantages in delivering local crime risk assessments and safer environments, which can support active living, walkable communities and public health. Criminology and public health have in the past focused on the behavior and characteristics of the individual. Nevertheless, the fields of injury prevention and crime prevention both now identify the significance of examining the characteristics of the event itself and this is where data mining can be utilized.

METHODS, ANALYSIS AND RESULTS

Data Visualizations

In this section, we visualized the Chicago crime dataset using the Tableau software. This was done to discover the crime patterns between years 2008 to 2012. More so, it helped in unraveling valuable information that will be useful to the law enforcement agencies and help in making business decisions in Chicago. We were able to look at the crime rates with respect to day and month and made conclusions based on our visualizations.

Crime Distribution by Month

We decided to look through the Chicago crime data to see the months that crimes are always on the increase and months when they are less frequent. Our focus is placed on knowing the top 4 months with the highest crime rate and top 4 months with the lowest crime rates. Based on the result of our visualization, we can conclude from the line chart in Figure 1 that criminals like warm weather which led to crime rates being high between May, June, July and August. More so, criminals seem to dislike cold weather thereby resulting into crime rates being low between November, December, January and February.

Crime Rate by Day

In this scenario, we looked at the day in each year crime occurred the most in the Chicago crime dataset. From the result in Figure 2, we noticed that crimes were committed the most on the first day (New Year Day) of each year. This will generate awareness so that the police department can strategies on how to stay alert on these days.

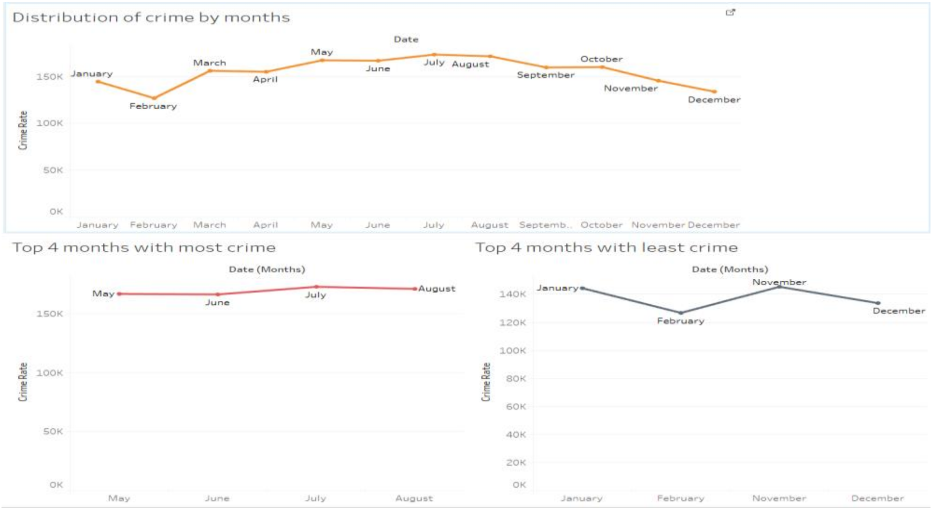

Figure 1. Rate of Crime Distribution by Month

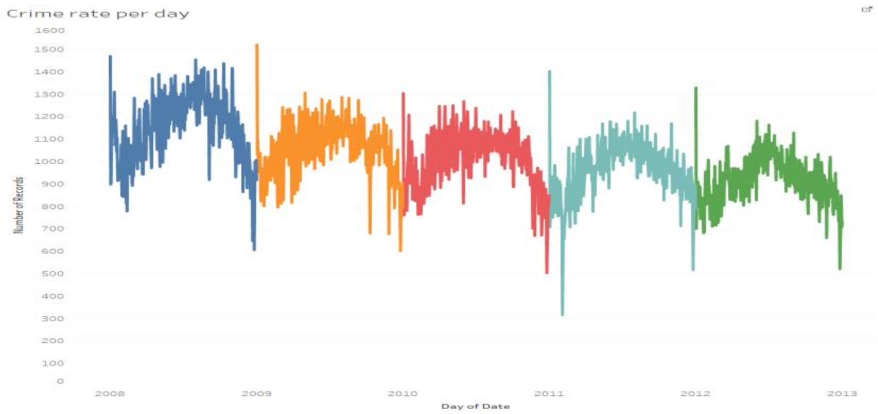

Figure 2. Rate of Crime Distribution by Day

Data Collection

For this project, we used the Chicago crimes data created from records of crimes cases by the Chicago Police Department and made available for public access and use. The second dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index’ created by the Epidemiology and Public Health Informatics, Chicago Department of Public Health (CDPH). We were able to create a new dataset from the existing ones by merging both datasets and adding new measures to the dataset.

The first dataset was used in data visualization to help us gain more information about the dangerous areas, periods with more criminal activities and the rate at which arrests are made in different community areas. The second dataset was used for descriptive and statistical analysis to show the roles of the socioeconomic factors in crimes. This dataset contains information about the socioeconomic status of each community area in Chicago from 2008 to 2012. Before merging the two datasets, we made sure they both had the same time periods, so we removed the records in the crimes dataset that was not in the range of 2008 to 2012. We used the python library known as Pandas to merge both datasets. The datasets were merged on the community area number as it was the only attribute common to both.

Descriptive Analysis

It is important to perform descriptive analysis on the dataset to help us gain more insights on the distribution of each variables. The socioeconomic dataset is used for this phase and since the dataset contains only continuous variables, we calculated the summary statistics values and scatter plots of the independent variables against the dependent variable.

<table>
<tr><th colspan="6">Summary Statistics Results</th></tr>
<tr><th>Variables</th><th>N</th><th>Min</th><th>Max</th><th>Mean</th><th>Std. Deviation</th></tr>
<tr><td>Percent of housing crowded</td><td>77</td><td>.3</td><td>15.8</td><td>4.923</td><td>3.6829</td></tr>
<tr><td>Percent households below poverty</td><td>77</td><td>3.3</td><td>56.5</td><td>21.766</td><td>11.5300</td></tr>
<tr><td>Percent aged 16+ unemployed</td><td>77</td><td>4.7</td><td>35.9</td><td>15.373</td><td>7.5434</td></tr>
<tr><td>Percent aged 25+ without high school diploma</td><td>77</td><td>2.5</td><td>54.8</td><td>20.342</td><td>11.8232</td></tr>
<tr><td>Percent aged under 18 or over 64</td><td>77</td><td>13.5</td><td>51.5</td><td>35.747</td><td>7.3277</td></tr>
<tr><td>Per capita income</td><td>77</td><td>8201</td><td>88669</td><td>25563.17</td><td>15293.098</td></tr>
<tr><td>Hardship index</td><td>77</td><td>1</td><td>98</td><td>49.51</td><td>28.691</td></tr>
<tr><td>Crime rate</td><td>77</td><td>1831</td><td>120501</td><td>24141.13</td><td>20372.191</td></tr>
</table>

Table 1. Summary Statistic Results

After running the descriptive analysis on the dataset, we created scatter plots of each independent variable against the target variable and fitted a regression line through the scatter plots to see if they are correlated and detect outliers. An example of the scatter plot is shown below.

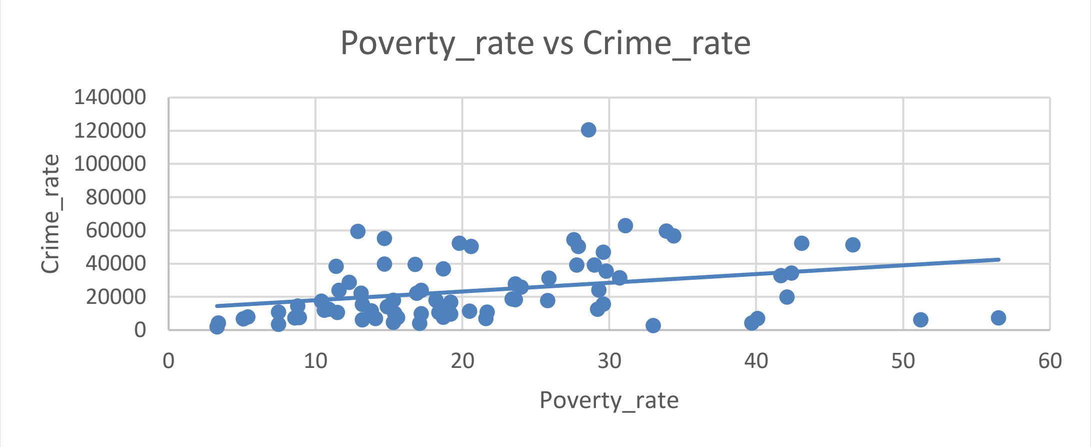

Figure 3. Scatter plot of Poverty rate against Crime rate

Linear Regression

Linear regression is a simple and one of the most common modeling techniques used in data analysis. It is very useful because it can be used to establish very important relationships between independent and dependent variables. It is also easy to understand and interpret the results from the model.

For this phase of the project, we will be applying linear regression to the geographical dataset to find out if the predictor attributes have a relationship with the dependent variable. The predictor variables include;

1. Percentage of House Crowding (POHC): This is the percent of housing units with more than one person per room (i.e., crowded housing)
2. Percentage of Households below Poverty rate (PHBL): the percent of households living below the federal poverty level
3. Percentage of Unemployed People Aged 16+ (PAU16): the percent of persons aged 16 years or older in the labor force that are unemployed
4. Percentage Aged 25+ without High School Diploma (PA25WHD): the percent of persons aged 25 years or older without a high school diploma
5. Percentage Aged under 18 or Over 64 (PAUO): the percent of the population under 18 or over 64 years of age (i.e., dependency)
6. The per capital Income: the average income measures the average income earned per person in each area
7. Hardship Index (HI): a score that incorporates each of the selected socioeconomic indicators

The table below shows the results gotten from implementing OLS Regression on the Python programming language. Vital results from the output includes the R-squared, adj. R-squared, F-stat, coef, standard error, the t and p values of the intercept and independent variables.

OLS Regression Results						
Dep. Variable:	Crime rate				R-squared:	0.303
Model:	OLS				Adj. R-squared:	0.232
Method	Least Squares				F-statistic:	4.284
No. Observations	77				AIC:	-65.66
Df Residuals	69				BIC:	-46.91
Df Model:	7					
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2899	0.149	1.95	0.055	-0.007	0.586

POHC	0.2205	0.173	1.278	0.206	-0.124	0.565
PHBL	-0.3202	0.23	-1.391	0.169	-0.779	0.139
PAU16	0.2025	0.199	1.017	0.313	-0.195	0.6
PA25WHD	-0.66	0.284	-2.322	0.023	-1.227	0.093
PAUO	-0.6161	0.192	-3.209	0.002	-0.999	-0.233
PCI	-0.1857	0.201	0.924	0.358	-0.215	0.586
HI	0.84	0.418	2.011	0.048	0.007	1.673

Table 2: OLS Regression Output

From the output shown above, the percentage of people under 18 and over 64 (0.002), the Percentage of people Aged 25+ without High School Diploma (0.023) and Hardship Index (0.048) have significant p-values and are positively associated with Crime rate.

To evaluate the fitness of the model, we consider the R-squared value which is the proportion of variance explained. This signifies that the proportion of variability in the observed data is explained by the model. The value of R-squared is between 0 and 1 and a high value means that more variance is explained by the model.

The value of R-squared in this model is 0.303 which means that around 30% of the variation in crime rate is explained by the model.

Another approach for implementing the multiple linear regression will be to introduce the predictor attributes in a stepwise manner by trying each independent variable individually and including it in the model if it is statistically significant.

CONCLUSION

According to the geographic map, it is observed that Near North Side has more number of theft followed by Loop. Also, theft crimes are recorded more compared to the other types of crime. More so, the highest arrests in Austin shows more police surveillance is present in that community area. Crimes are at a maximum from 6pm to Midnight and occurred frequently on Fridays. We also noticed that crimes are committed more on Jan 1st of each year. From the visualization, we recommend that areas with least arrests need more police force to stop various types of crimes in those community areas. Also, time and day visualization aid the police department to allocate their force in various communities at peaks times and days.

The Linear regression analysis has shown that the percentage of people under 18 and over 64, the percentage of people Aged 25+ without High School Diploma and the Hardship level have an influence in the rate of criminal activities in Chicago. This socioeconomic issue that have been identified from this project can give government officials ideas on the kinds of projects to execute in reducing the crime rate in the state. This can also be a guide to entrepreneurs in making business decisions like the kind of business to embark on and the location of the business.

REFERENCES

1. Alkhaibari, A. A., and Chung, P.-T. (2017) Cluster analysis for reducing city crime rates, In *Systems, Applications and Technology Conference (LISAT), 2017 IEEE Long Island*, 1-6.
2. Cozens, P. (2008) Public health and the potential benefits of crime prevention through environmental design, *New South Wales public health bulletin*, 18, 2, 232-237.
3. Rasheed, A., and Wiil, U. K. (2015) A Tool for Analysis and Visualization of Criminal Networks, *17th International Conference on Modelling and Simulation (UKSim-AMSS)*, 97-102.
4. Thongsatapornwatana, U. (2016) A survey of data mining techniques for analyzing crime patterns, In *Second Asian Conference on Defense Technology (ACDT)*, 123-128.
5. Ying, Z. (2016) Analysis of Crime Factors Correlation Based on Data Mining Technology, *International Conference on Robots & Intelligent System (ICRIS)*, 103-106.