Making Neighbourhoods Safer with Statistical Inference

Divya Raj
Student
School of Software Engineering, Tongji University

Yan Liu
Professor
School of Software Engineering, Tongji University

Abstract

As the crime rates shoot up year by year, in this era of data deluge, more and more data is getting piled up across the world. This study tries to find relationships between crime victims, arrests and neighbourhoods and dig out patterns that exist in the available data sets and help the police departments with statistically significant information. Using which the police can rearrange the man powers and fight crimes more effectively. The departments can use many such studies to keep the neighbourhoods safer.
Keywords: Statistical Inference, Crime, Victims, Arrests, Neighbourhoods, Chi - squared test, Multinomial Logistic Regression

I. INTRODUCTION

Technology has been helping the law enforcement officers to improve public safety in several ways. For instance, proper maintenance of criminal records, closed-circuit televisions as a witness, technical assistance to ensure that technology is deployed in a manner that facilitates information-sharing across agencies. Today, most of that criminal analysis work is getting easier, credits to the new emphasis on data analysis in crime fighting departments. The Gen X police is tapping into complicated patterns and often under-utilized data in an effort to make high-crime neighborhoods safer. This transition to a data powered crime fighting possesses great potential for life saving and secure world.

A. Objectives of the Study:

- To study the relationship between the targeted neighborhoods and the crime type.
- To study the relationship between the targeted neighborhoods and month of crime.
- To study the relationship between age group of criminal and the arrest location.
- To study the relationship between the incident offence and the arrest location.

II. HYPOTHESIS TESTING

H0: There is no significant relation between crime type and crime locations.
H0: There is no significant difference between crime locations and the month of crime.
H0: There is no significant relation between criminal age group and neighborhood.
H0: There exists no significant relation between incident offence and the arrest location.

A. Target Population/ Sample Frame:

The populations of this study consist of the crime victims in the city of Baltimore and the people arrested.

B. Sample Size:

The study utilizes two data sets from the Baltimore city, Arrest Dataset and victim data sets. Arrest data consisting of 115728 records with 15 variables defined while, victim data consisting of 253754 records and 11 variables. For the analysis 10% of the records were utilized from each set.

C. Sampling Techniques:

For the purpose of sample selection, simple random sampling was adopted.

III. INSTRUMENT FOR DATA COLLECTION

The data utilized in this study is from open datasets made available by Baltimore city government. Namely, BPD Arrests, which represents the top arrest charge of those processed at Baltimore's Central Booking & Intake Facility and BPD Victim Based Crime Data. The datasets mainly consists on nominal categorical type data.

A. The Research Environment:

This is a observational study in the crime domain.

1) Data Analysis:

The data collected are processed using R programming language. R is widely used statistical environment used for data analyses. The research tools that the researcher intends to apply in this thesis are Frequency Distribution, Chi-square Test, and Multinomial Logistic Regression.

2) Frequency Distribution:

Frequency distributions were obtained for all the classification variables for example race, gender, age, locations of crime.

3) Descriptive Statistics Analysis:

A descriptive statistics has been described is a statistics that describes the phenomenon of interest. It is used to know the average score when a set of figures are involved as well as extend of variability in the set. With descriptive statistics we are simply describing what is or what the data shows. We have used some descriptive statistics like Frequency Distributions, Cross tabs and charts.

4) Frequency Distribution:

A table that lists all the categories or classes and the numbers of values that belong to each of these categories or classes is called frequency distribution. A frequency gives the numbers of observations or classes fall into each group or category.

5) Chi - Squared Test:

A chi-squared test, also referred to as χ^2 test, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. A chi-squared test can then be used to reject the null hypothesis that the data are independent. The purpose is to determining whether there is a relationship between independent variables and dependent variable. If the probability value (P-Value) is smaller than 0.05 ($p<0.05$), the result will be significant, which means that there is a relationship between independent variable and dependent variable. But if the probability value (P-Value) is greater than 0.05 ($p>0.05$), it means that there is no relationship between independent variable and dependent variable. to the values or categories of the other variables.

6) Contingency coefficients:

The strength of the relationship approved by chi-squared test can be figured out depending on whether the variable are nominal/ ordinal or mix of the two. The nominal measures of association take values from 0 to 1, with 0 meaning no association and 1 meaning perfect association.

For the nominal data, contingency coefficients method was employed to measure the strength of association.

7) Multinomial Logistic regression :

Multinomial logistic regression is a classification method that generalizes logistics regression to multi class problems i.e. with more than two possible discrete outcomes. This model is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables, which may be real-valued, binary-valued, categorical-valued, etc. It is used when the dependent variable in question is nominal (equivalently categorical, meaning that it falls into any one of a set of categories which cannot be ordered in any meaningful way) and for which there are more than two categories.

Multinomial logistic regression is a particular solution to the classification problem that assumes that a linear combination of the observed features and some problem-specific parameters can be used to determine the probability of each particular outcome of the dependent variable.

IV. STATISTICAL INFERENCE ON THE CRIME DATA AND RELATED VARIABLES

Random sampling technique was used in this observational study. Sample size was set to be 10% of the original set. For this analysis, the significance level is 0.05 was set to conduct the chi-square test for independence.

A. On Victim Data:

1) Crime Type vs Crime LOCATION

1) Null Hypothesis

H_0: Crime Type and Crime Location are independent.

H_a: Crime Type and Crime Location are not independent.

2) Results:

On performing chi squared test and the test of association, results obtained are as follows:

Pearson's Chi-squared test: p-value =7.8018E-78

Nominal by Nominal : Contingency Coefficient = 0.795

3) Test Statistics Analysis:

There is very strong evidence of a relationship between neighborhood and the crime type with a p-value < 0.05 and the strength of association is 0.795. So, the crime types and the neighborhood are statistically related.

4) Interpretation:

Majority of the crimes are location dependent and are carried out targeting victims in a particular location and attacked. It is seen that Northeastern District accounts for a 15.8 %, highest number of crimes in the city. Some Neighborhoods are more likely to be targeted for certain types of crimes. Using multinomial logistic regression, the logs odds of crime type were calculated and many prominent results were found. Results show there exists a strong pattern with the victim selection. In the Downtown area, one is the most likely to be a victim of common assaults, while a house in Belair - Edison is more prone to burglary compared to other neighborhoods.

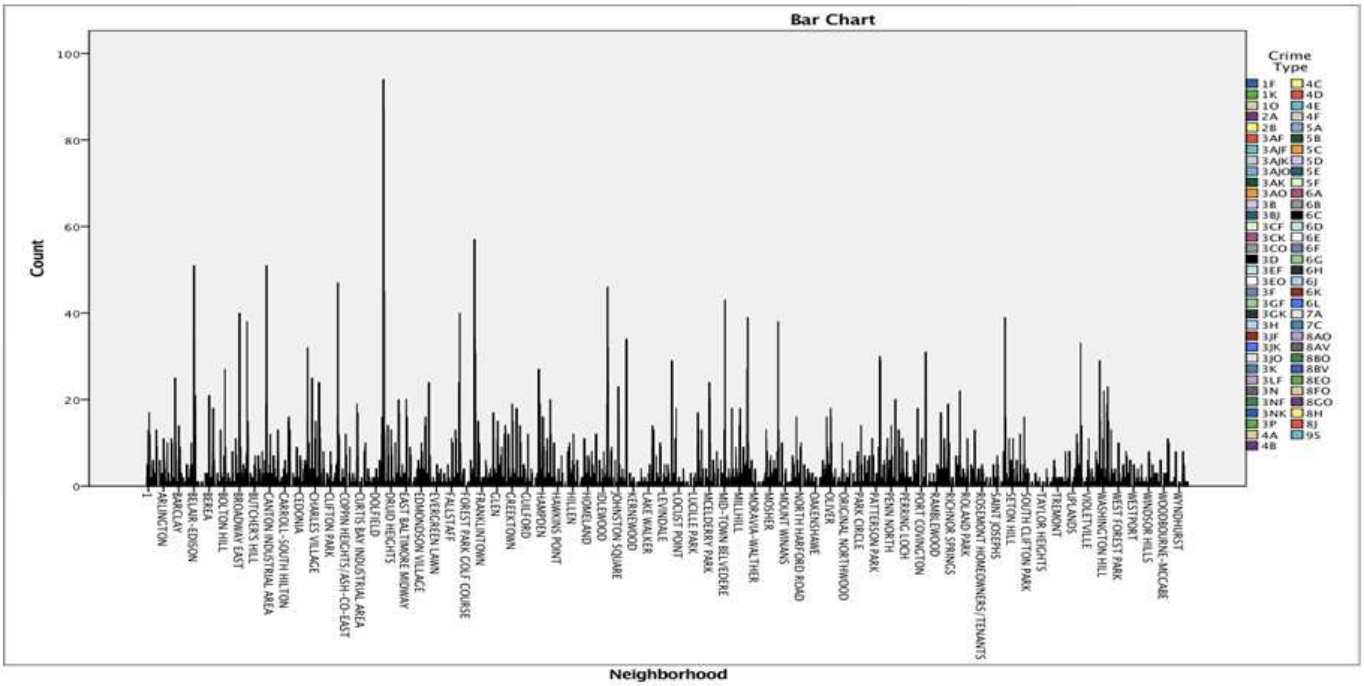

Fig. 1: Bar chart of Neighborhood vs. Crime Type

B. Crime Location vs Crime MONTH

1) Null Hypothesis

H_0: Crime Location and Month of crime are independent.

H_a: Crime Location and Month of crime are not independent.

2) Results:

On performing chi squared test and the test of association, results obtained are as follows:

Pearson's Chi-squared test: p-value = 0.008761

Nominal by Nominal : Contingency Coefficient = 0.499

3) Test Statistics Analysis:

There is very strong evidence of a relationship between neighborhood and the month of the crime with a p-value < 0.05 and the strength of association is 0.499. So, the month of crime and the neighborhood are statistically related with significant association.

4) Interpretation:

The crimes committed are mostly day and month specific. They are carried out with proper planning with neighborhoods at target. Frankfort accounts for about highest 2.4% of the crimes in the city of Baltimore. The crime peaks around the mid of June across most neighborhoods. In Downtown, the month of major crimes is October, whereas in Curtis Bay area, March is more preferred by criminals

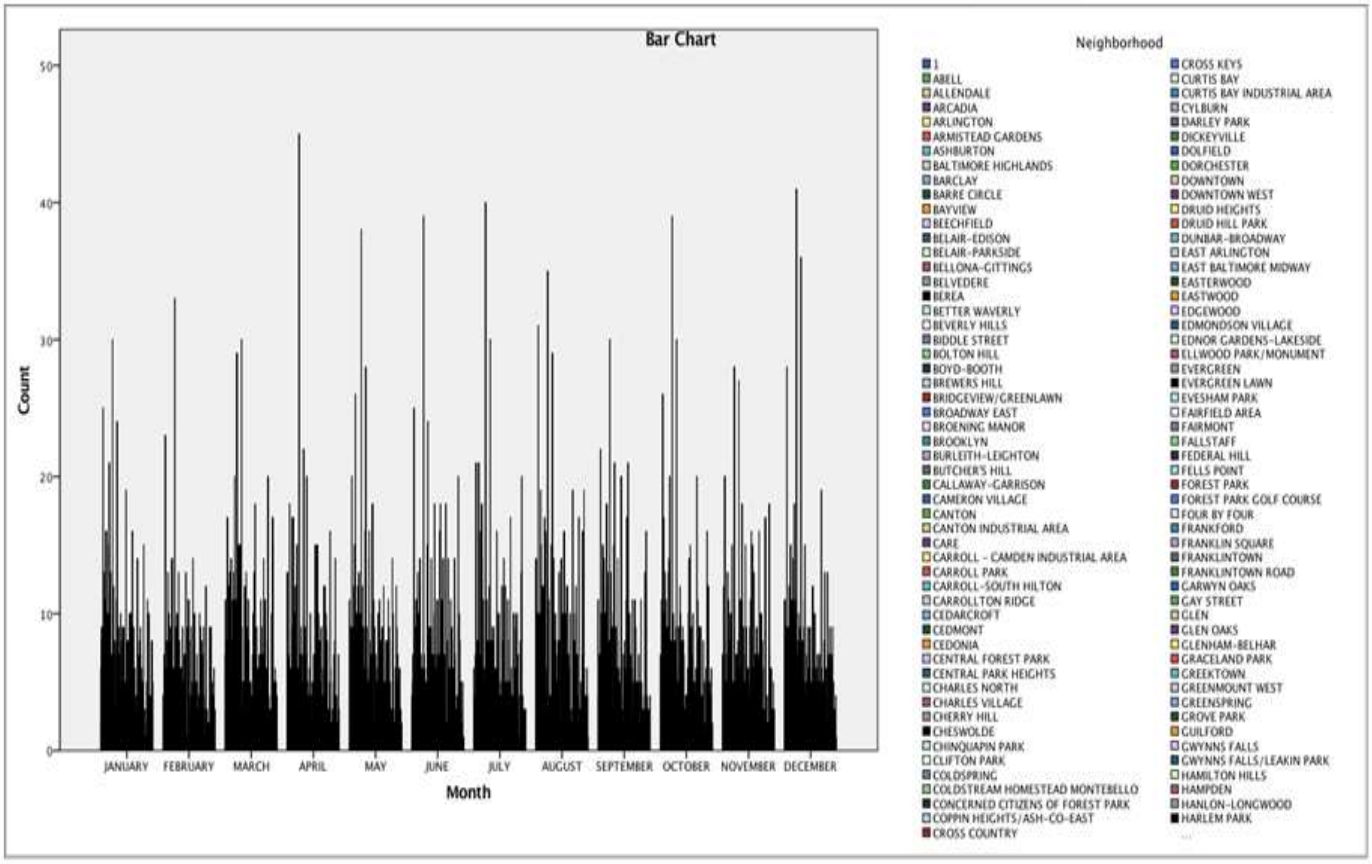

Fig. 2: Bar chart of Month of Crime vs. Neighborhood

C. *On arrest data*

1) CRIME LOCATION vs AGE GROUP

1) Null Hypothesis

H_0: Criminal age group and Crime location are independent.

H_a: Criminal age group and Crime location are not independent.

2) Results:

On performing chi squared test and the test of association, results obtained are as follows:

Pearson's Chi-squared test : p-value = 0.000036

Nominal by Nominal : Contingency Coefficient = 0.497

3) Test Statistics Analysis:

There is very strong evidence of a relationship between age group of the criminal and the crime location with a p-value < 0.05. And the strength of association is 0.497. So, the criminal age groups and the neighborhood are statistically related with significant association.

4) Interpretation:

The age group between 20 - 40 comprises 68.6 % of the arrests , followed by 40 -60 age group with a total of 25.7 % whereas the juveniles form about 4.3 %. The 40 - 60 age group are found in the Eastern region whereas 20 - 40 in the districts of southern region. Also the juveniles are spread in the southern neighborhoods.

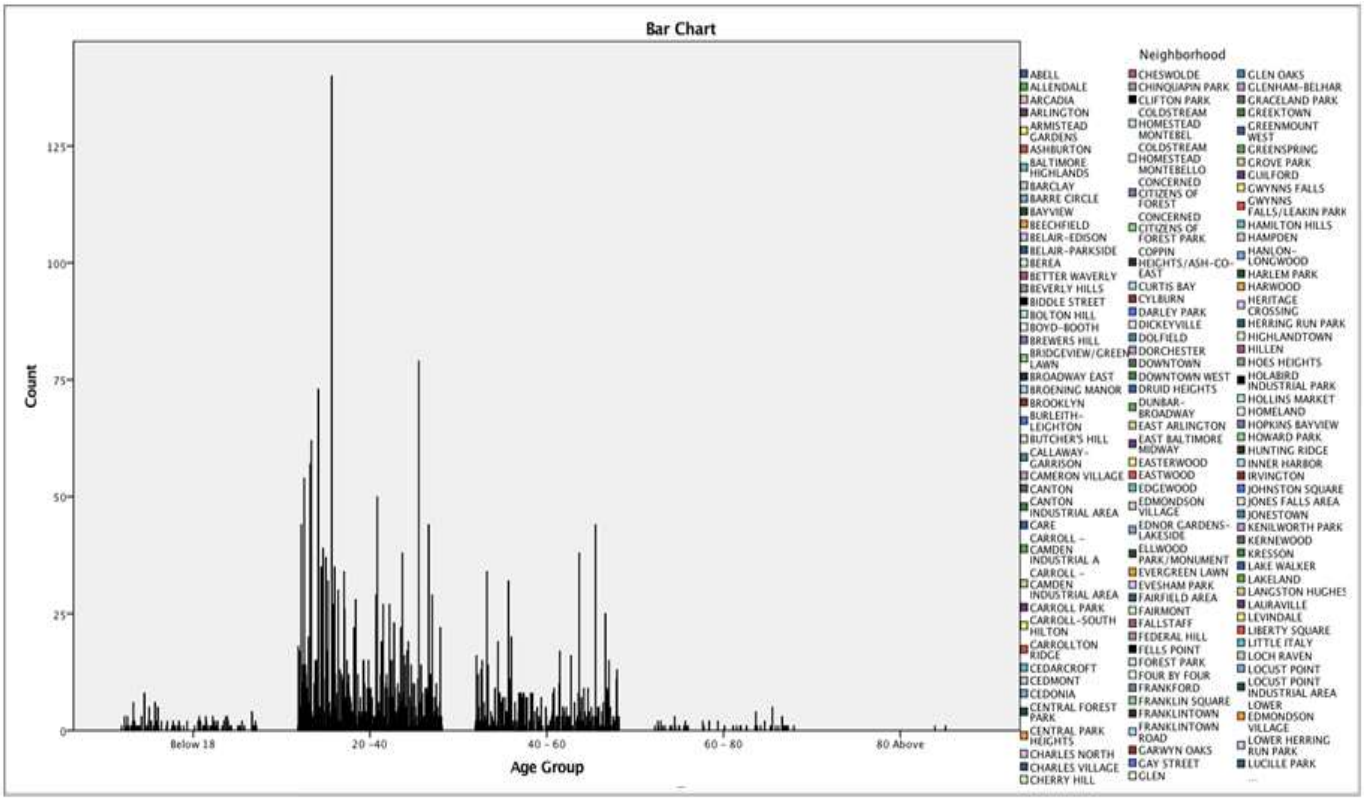

Fig. 3: Bar chart of Criminal Age Group vs. Neighborhood

D. Incident offence vs. Arrest Location

1) Null Hypothesis

H_0: Incident offence and arrest location are independent.

H_a: Incident offence and arrest location are not independent.

2) Results:

On performing chi squared test and the test of association, results obtained are as follows:

Pearson's Chi-squared test: p-value = 2.2643E-133

Nominal by Nominal : Contingency Coefficient = 0.931

3) Test Statistics Analysis:

There is very strong evidence of a relationship between incident offence and the arrest location with a p-value < 0.05. And the strength of association is 0.497. So, the criminal age groups and the neighborhood are statistically related with highly significant association.

4) Interpretation:

The incident offences and the arrest locations seem to be strongly associated. The most likelihood of getting arrested for narcotics lies in Broadway East and Downtown region. Belair - Edison accounts for most larceny of shoplifting type arrests. Eastern and north eastern see the arrests for common assaults. It tells us that, there exists a strong pattern amongst various neighborhoods for the arrests of different types of incident offences.

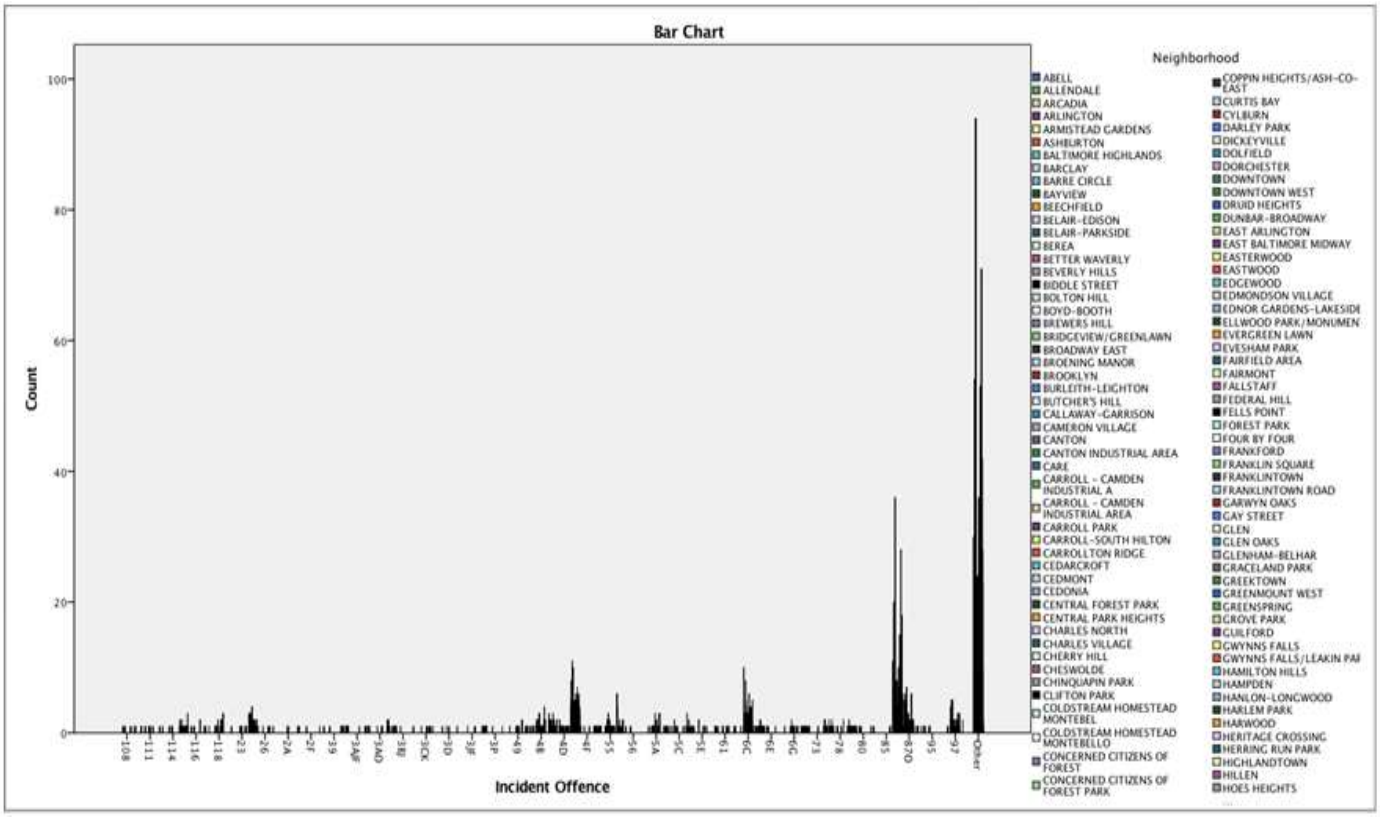

Fig. 4: Bar chart of Incident Offence vs Neighborhood.

V. CONCLUSION

As the complexity of the criminal networks and patterns is expected to continue to challenge the crime fighters for many years, data analysis with statistical inference methods hold tremendous promise for determining the critical relationships governing crimes. This can be realized through better extraction of information from all possible sources and a more active experimentation driven by active learning. A lot of statistically significant information can be extracted from the data available and thereby making neighborhoods safer. The result will be a better and safer society making the world a better place to live and development will be dramatically improved by the ability to assess effects of potential discoveries more comprehensively. The study shows a significant relationship exists between crime locations, the crime types and other related variables. Clearly much work remains to be done, not least of which is to convince practitioners of the value of ceding some important decisions to machines.

REFERENCES

[1] Brian Caffo, Jeff Leek, Roger Peng , Statistical Inference for data science , Johns Hopkins Bloomberg , 2014
[2] James A.Fox , Jack Levin , Elementary Statistics in Criminal Justice research: The essentials, 2005.
[3] Croissant, Y. (2011). Package 'mlogit'. http://cran.r-project.org/web/packages/mlogit/index.html