SUGGESTING SAFETY MEASURES FOR CITY OF VANCOUVER

Submitted by:

Gokani Chintan Harishbhai

Jyotishi Anit

Guided by:

Prof. Sujing Wang

PROPOSAL

COSC 4301/5340 Data Science and Big Data Analysis

Final Project

Due: Monday 5/9/2016

GROUP MEMBERS:

Anit Jyotishi (L20404856)

Chintan Harishbhai Gokani (L20398278)

1: Topic:

Area wise safety measures prediction and analysis.

2: Data Sets: What data will you use? Where will you get it?

We will use official data for Canadian Government Website. Data can be downloaded from the following link:

http://data.vancouver.ca/datacatalogue/crime-data.htm

One of two data sets that you decide to use and analysis.

Provide the link that you download from

3: Description/Summary of your datasets:

The following table gives a description of our data-set:

Data set description	This is a dataset of crime data on a year-by-year basis beginning in 2003.
Attributes	• TYPE • YEAR • MONTH • HUNDRED_BLOCK • NEIGHBOURHOOD • X • Y

4: Questions/tasks that you plan to answer/perform based on the data sets. What is the problem you are solving?

Predicting safety measures for police and residents according to the type of crimes committed in a locality.

5: What data science techniques/algorithms introduced in our class that you plan to use?

We are planning to use K means and DBScan clustering along with various visualization techniques discussed in class.

6: Implementation of your project: hardware and software needed

Software Specifications: RStudio

Hardware Specifications: Windows 8, 8GB RAM

7: Output of your project (What do you expect to submit at the end of the semester?):

e.g. your source code, project report, presentation and demonstration of your project.

We expect to submit the following documents:

- Source Code
- Project Report
- Presentation
- Demonstration of The Project

TABLE OF CONTENTS

Introduction

- In today's world, security is an aspect which is given higher priority by all political and government worldwide and aiming to reduce crime incidence. As data analysis is the appropriate field to apply on high volume crime dataset and knowledge gained from data mining approaches will be useful and support police force. So in this crime analysis is done by performing k-means clustering on crime dataset using rapid miner tool.

- Crime Analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects.

- Crime analysis also plays a role in devising solutions to crime problems, and formulating crime prevention strategies. Quantitative social science data analysis methods are part of the crime analysis process; though qualitative methods such as examining police report narratives also play a role.

1. Pre Processing

Data preprocessing is necessary for data cleaning and data modification for prepare raw data for further processing. It is necessary to get accurate result of data analysis.

a) Sampling:

We sampled our data twice in the entire project:

- First was when we selected the records only for year 2016 from data file which contains records from year 2003 to 2016 with 500000 objects.
- Second instance of sampling is selecting 2/3rd (~70%) of the above 2016 sample for outlier detection through DBScan.

b) Handling missing values:

Details of some crimes were not recorded in order to protect victim's identity. These were generally crimes against a person. Generally such crimes are not affected by measures like Geographic Location, but by the circumstances of the victim. Best course of action is to eliminate records with missing attributes.

We achieved this by removing the records which have 0.0 value for X co-ordinate, As it representing crime type "Offense against person" in our dataset.

c) Convert xy coordinates:

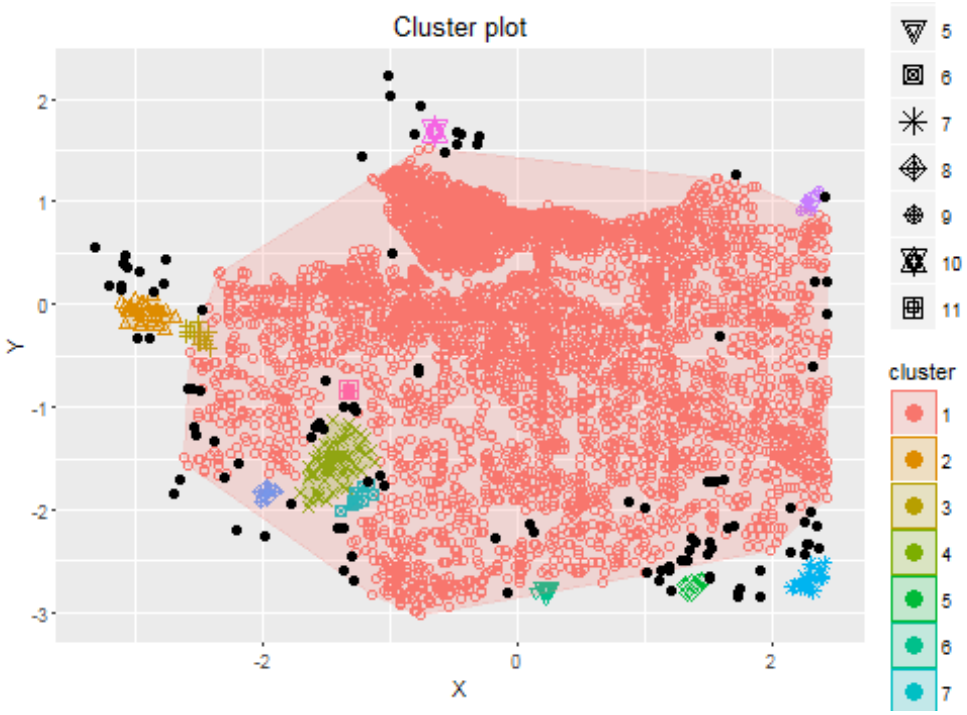
In Our dataset xy coordinates in the UTM format and we converted it into standard degree format as we see in maps. Because we applied "earth.dist" function of "fossil" package for calculation of distance matrix to apply on DBScan algorithm. Which actually accept degree format of xy coordinates.

d) Outlier Detection

This step is very important because K-means might not give accurate result with noisy points. If numbers of outliers are more, then that would change the centers of our clusters and affect our resulting clusters in a massive way. We do not want that to happen. Hence, detecting and removing outliers was something that needed to be done before applying k means clustering.

Here, we used two visualization techniques are Box plot and DBScan. Both are useful for outlier detection. For getting basic idea about outliers, we applied Box plot visualization technique and get noisy points with x and y attributes. Hence, it doesn't quite fit the sort of data we are handling. And some time it might not give appropriate output for outliers.

Thus, to get accurate result for outlier detection in our dataset we had applied DBScan algorithm. We had applied DBScan algorithm for 7000 objects (~70% of 2016 year data). And we detected around 70 outliers as a result.



e) Aggregation

This is a major operation in the workflow of the entire process. We have to prepare raw data to perform clustering. In order to cluster neighborhoods according to crime types, we need an aggregate data set which has a neighborhood as the key and total number of each crime type as attributes. As we want the similar type of neighborhood with similar crime type.

The transition of the data went through the following steps:

- Step 1:
  The data set looked like:

Type	Year	Month	Hundred_block	Neighborhood	N_hood	X	Y

- Step 2:
  The data set looked like:

Type	Neighborhood

  Here, we just selected column 1 and column 5 from the data set.

- Step 3:
  Finally, the aggregated data set looked like this:

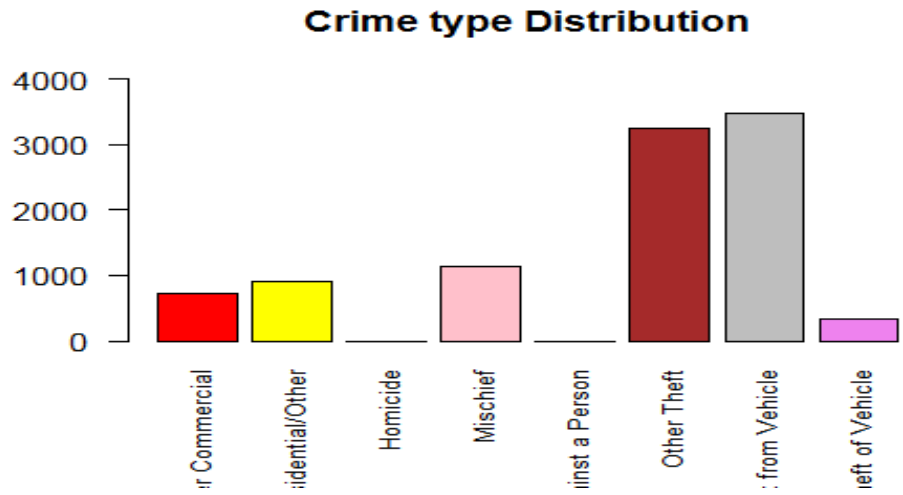
Key	Crime Type 1	Crime Type 2	…	Crime Type 8
Neighborhood 1				
Neighborhood 2				
…				
Neighborhood 25				

2. Visualization

Visualization is basically graphical representations of data. It gives a general idea about the objects of data and its attributes. In our project, we have applied the following visualizations for data analysis.

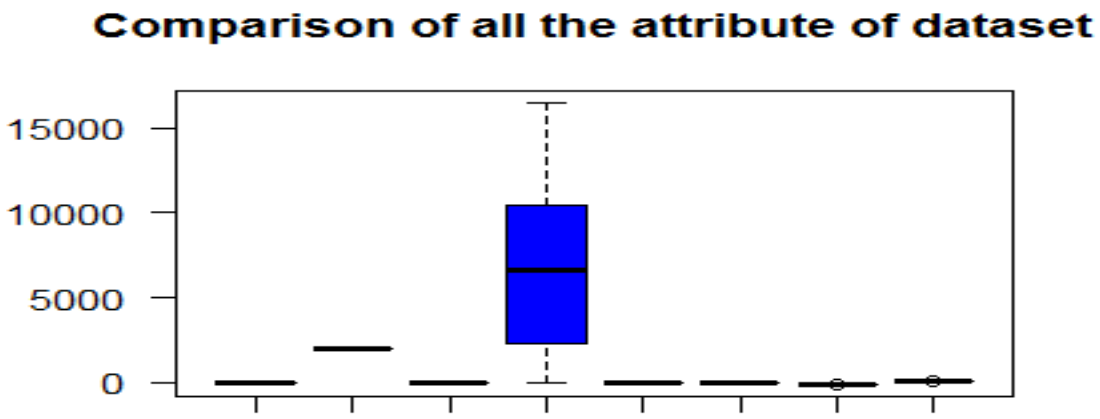
a) Bar Chart

Bar chart provides us the count of each crime occurs in year 2016. We applied bar chart visualization on attribute “type”. From the output of bar chart we found two crime types, “other theft” and “theft from vehicle”, are mostly occurred crimes in the year 2016.



b) Box Plot

Box plot is useful to compare two or more variables. We have applied it to get the idea about noisy points and value distribution among different attributes. From this visualization technique, we found outliers on “x” and “y” attributes.


Comparison of all the attribute of dataset
15000
10000
5000
0

3. Clustering

Clustering is unsupervised technique for grouping similar objects within the same cluster.

a) Selecting Clustering Technique

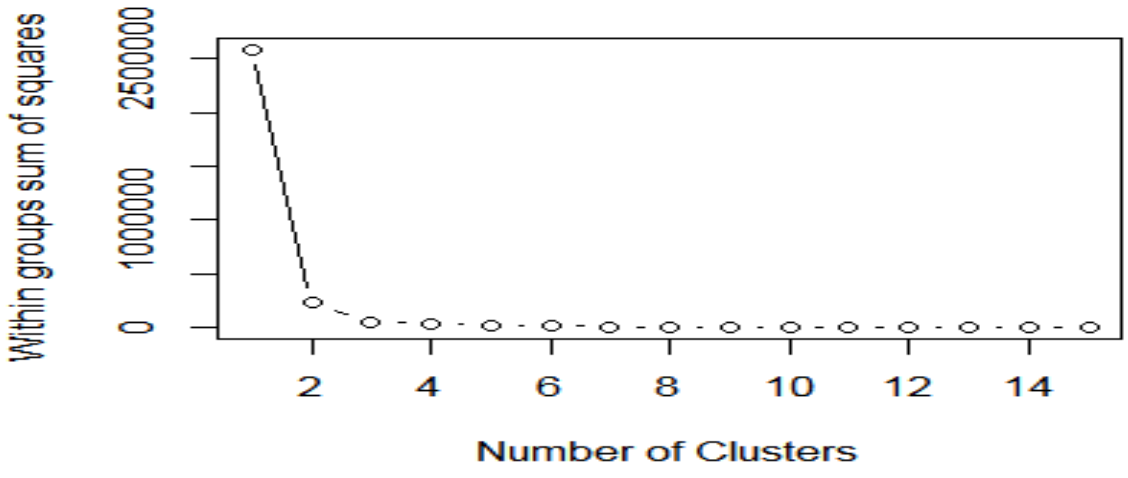
We selected K-means due to following reasons:

- K means is traditional clustering algorithm, which works well with large dataset.
- K means is less expensive
- K means is simple method of clustering with predefined value of number of clusters decided as output.

b) Selecting Value of K

We calculated the value of K by calculating Sum of Squared Errors (SSE) for the all points (25 Neighborhoods) for all possible values of K. We decided to go with K = 3, as K = 3 gives the least SSE.

Below is a graphical representation of how SSE varied with the increasing value of K:



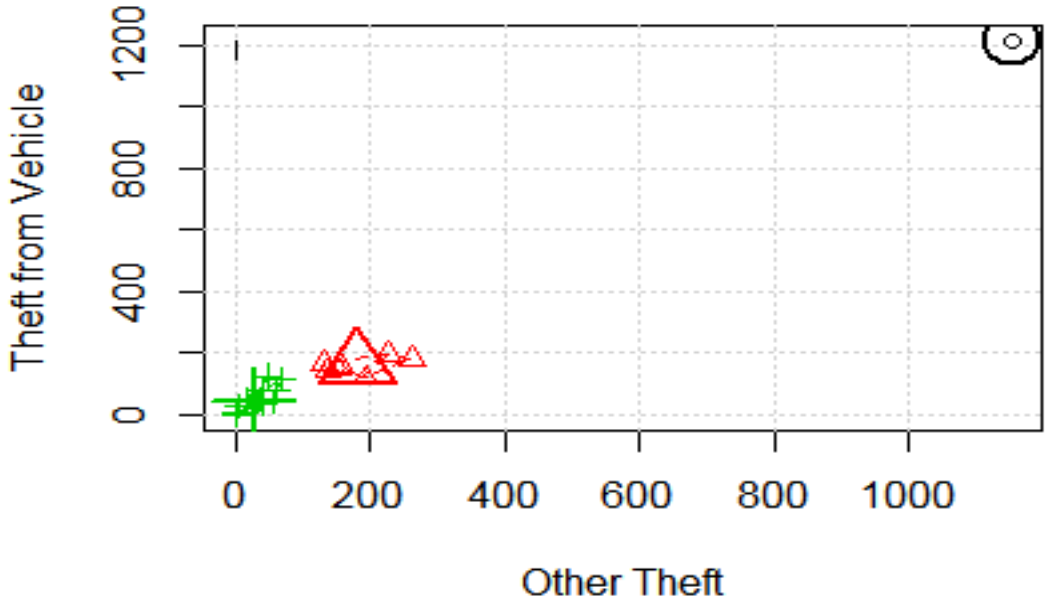
c) Characteristics of a Cluster:

Below are the characteristics of one cluster:

- Few adjacent neighborhoods can form one cluster.
- An adjacent neighborhood is always susceptible to an act of crime similar to the one in adjacent neighborhood.
- Basically, we are classifying neighborhoods according to risk (Crime types).

d) Choosing a Distance Function:

- Euclidian Distance came off as a natural choice for distance measure. So, we have considered Euclidian distance function for our dataset.
- Each crime type has been assigned a weight of 1.0.



4. Conclusion

We came to a conclusion that the most occurring crimes are in the adjacent neighborhood as seen in the cluster as follows-

- ✓ On comparing the statistical analysis of "**theft from vehicles**" and "**other thefts**", the occurrence of crime rate across the neighborhood within the same cluster seen is almost the same ratio.

- ✓ On the contrary, the cluster in the top right corner shown in '**black**' records the highest crime rate with single neighborhood.

- ✓ This on a broader platform will help in establishing **better safety norms** using this easy pictorial survey. Determining if particular crimes are increasing; identifying the hot spot locations where crime is concentrated; understanding the temporal trends of offending and analyzing potential reasons for crime trends will be critical features of crime data analysis. Hence applied!

5. References

- http://www.r-tutor.com/r-introduction/data-frame/data-import
- https://www.youtube.com/watch?v=3GorGZgTTEk
- http://www.inside-r.org/r-doc/stats/kmeans
- http://data.vancouver.ca/datacatalogue/crime-data.htm
- http://www.r-bloggers.com/r-functions-for-earth-geographic-coordinate-calculations/