Toronto Crime Data Analysis using Unsupervised Learning

Kunal Taneja
Department of ECE
University of Waterloo
Waterloo, Ontario, Canada
kunal.taneja@uwaterloo.ca

Prateek Gulati
Department of ECE
University of Waterloo
Waterloo, Ontario, Canada
p3gulati@uwaterloo.ca

Ganesh Rajasekar
Department of ECE
University of Waterloo
Waterloo, Ontario, Canada
g3rajasekar@uwaterloo.ca

Anmol Sharan Nagi
Department of SYDE
University of Waterloo
Waterloo, Ontario, Canada
asnagi@uwaterloo.ca

***Abstract*—Public safety and protection is the need of the hour in large cities like Toronto. Law enforcement agencies in large cities have this uphill task of identifying criminal activities, and a lot of resources and time is wasted in identifying such crime hot spots in the form of surveillance, investigations and man-hunt. Recently, modern techniques such as Data Analysis and Knowledge discovery have been playing a major role in the process of extracting unknown patterns and understanding hidden relationships of the data for many applications. With the exponential increase in the size of the crime dataset every year, the need to process this data becomes essential in order to extract meaningful information out of it. Cluster analysis is the method of classifying a large pool of data items into smaller groups which share similar properties.This papers aims to apply different clustering techniques such as K-Means, Agglomerative and DBSCAN to Toronto's Major Crime Indicator (MCI) Dataset and identify violent and non-violent neighbourhoods in the city of Toronto. It intends to perform data analysis to find out which crimes occurs at what time of the day and the geographic location associated with crime.**

***Index Terms*—Criminal Activities, Data Analysis, Knowledge Discovery, Clustering, K-Means, Agglomerative, DBSCAN.**

I. Introduction

Toronto, the capital city of Ontario as well as the most populous city in Canada is considered to be a global city since it represents a major role in the global economy in terms of trade, employment and immigration. City accounts for nearly one quarter of the country's employment in finance, health, manufacturing, trade and many more. Toronto's economy attracts a large volume of skilled workers around the global. In 2006 about one-third of Torontonians had a university degree compared to the national average of eighteen percent [1]. Toronto's global nature is also evident in the cultural diversity of its population, as about one-quarter of recent immigrants live in the city of Toronto and nearby areas. These facts and figure describe the importance of Toronto both on a national and international scale. Cities with such an impact need to keep a check on disturbing elements which may pose a threat to their glorious economies. Even though Toronto has a relatively low global crime rate but it is important to understand its crime patterns so as to curb unethical activities.

In the year 2006, Toronto had the highest number of offences overall Canada. In 2005 Toronto entitled itself with the "year of gun" because of the significance of firearm homicides committed [1]. Upon examination of local crime rates it was found that the rates of violent crimes are higher near the downtown core and east and northwest areas of the city.

Data suggests that as the population of Toronto seen a rise, their has been an increase in crime rates. With these increasing crime rates, their is an increased pressure on the Toronto police and regulating agencies to enforce law and order and to maintain the integrity and safety of Torontonians [1]. As the amount of crime grows, so does the amount of data logged upon these crimes. This collected data can be used to extract meaningful information which can help both the police department and the public to maintain safe surroundings. Identifying crime and predicting dangerous hotspots at a certain time and place could provide a better visualization for both public and authorities, and aid in terms of proper planning and safety measure to stop the antisocial activities from happening in the community. Cluster analysis is the method of classifying a large pool of data items into smaller groups which share similar properties. This paper tries to apply and compare different clustering techniques such as k-mean, agglomerative and DBSCAN in order to cluster the neighbourhoods of Toronto into violent and non-violent categories. To validate the results of the clustering algorithms we use several internal validation criteria that measure how well the data-points define the purity of the cluster. Based on these criteria, we select the best clustering algorithm and try to visualize the clusters formed by this algorithm over Toronto's map using Google Map API. This would help in clearly showcasing which neighbourhood are dangerous and require more focus of police agencies. It would also supplement to the general public's knowledge for their own well-being and safety.

II. Literature review

[2] introduces the notion of MCDM (Multiple Criteria Decision Making) algorithms. MCDM describes how an algorithm cannot be evaluated or selected based on only a single factor. A combination of factors like space complexity and time complexity are taken in consideration while designing an

algorithm for a problem. [2] proposes an MCDM approach to evaluate the clustering results in financial risk analysis. Evaluation of clustering algorithms involves more than one criterion such as Entropy, Dunn's Index, Silhouette Score and computation time. Therefore, it can be easily modelled as MCDM problem. [2] choose 6 clustering algorithms, 1 validity measure and 3 MCDM methods to validate the evaluation of clusters found. The results showed than no algorithm could achieve the best performance on all measurements and it was necessary to utilize more than a single performance measure to evaluate the clustering algorithms. Therefore, a combination of validity measures are required to rank the clustering methods.

[3] discusses how visualizing the results of clustering on a 2D plane can be challenging when we have a large number of features or attributes in the data. The paper explains how applying PCA (Principal Component Analysis) prior to k-means clustering can help in getting better clusters and also improved running times. Therefore, after applying PCA the number of dimensions of the data can be reduced to have a better visualization of the clusters formed by the clustering algorithms.

[4] addresses the issue of selection of value of K for K-means clustering algorithm. It discusses several methods for selecting K and also analyses the factors influencing the selection of K. In some applications value of K can be understood from application's point of view. For example, a problem which tries to group a set of customers based on their interest in a product can have only two clusters 'interested' and 'not interested'. [4] explains the statistical methods which can be used to decide a proper value of K based on some threshold values. Notions of elbow technique and silhouette score analysis are examined.

III. PROBLEM DEFINITION

The study aims at identifying violent and non-violent neighbourhoods in the city of Toronto, while providing a better visualization for the public. An attempt is made to model a relationship between several criminal patterns, the behaviour and degree of crime. The paper tries to cluster the crime prone areas with respect to different major crimes that have occurred in the past. The major challenge is to understand the versatile data available from Toronto Police public portal and employ different pattern recognition techniques to provide a better crime heat map. Data analysis is performed to find the temporal and spatial distribution of the crimes over the day. These findings are performed using different clustering techniques . The results of each clustering algorithm are compared against several internal validation measures. At the end an attempt is made to showcase the clustering results over the map of Toronto. The plot tries to present the types of crimes which happen at various times of the day and week, and based on geographical locations.

IV. METHODOLOGY

A. K-Means

One of the most famous clustering algorithm is k-means. The algorithm tries to cluster data by separating the samples in n groups of equal variance, minimizing the within-cluster sum of squares. This technique scales very well to problems with large datasets and has been used across many different fields of study and research.

The k-means algorithm divides a set of $\boldsymbol{N}$ samples into $\boldsymbol{K}$ disjoint clusters $\boldsymbol{C}$ each described by the mean μ_j of the samples in the cluster. These means are called centroids. The algorithm tries to choose centroids such that the within-cluster sum of squared error could be minimized. Which can be given as:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (|x_i - \mu_j|^2) \tag{1}$$

The algorithm can be understood easily as in three steps. In the first step, k samples from the dataset are chosen randomly as centroids. After this initiation, the second step calculates the inter-sample distance and assigns each sample to the nearest centroid. The third step tends to create new centroids by taking the mean of all the samples which are assigned to the same centroid. The difference between new and old centroids is calculated and the second and third step continue till this value is less than a threshold.

The convergence of the algorithm is highly dependent on the initialization of the centroids. Therefore, the algorithm is run several times, with different initializations to check result stability and escape any local minima. In high-dimensional spaces, due to the curse of dimensionality, Euclidean distance is not able to clearly identify between clusters, so running a dimensionality reduction algorithm such as PCA on the dataset before running the k-means clustering algorithm is advisable [5].

B. Agglomerative:

Hierarchical clustering assigns objects into tree-like structures, where clusters can have a data point or representation of low-level cluster. The root of the tree is a unique cluster with all samples with leaves being clusters with only one sample [6]. Agglomerative is a type of hierarchical clustering and it follows a bottom-up approach. It begins with each object being its own cluster and then merges these clusters iteratively until all the objects are merged into one cluster. For merging, the object can find the cluster which is closest to it according to the following strategies:

- **Ward** minimizes the sum of squared error within all the clusters. It is a variance-minimizing approach.
- **Complete Linkage** minimizes the maximum distance between the pair of clusters.
- **Single Linkage** minimizes the distance between the closest observation pairs of the clusters

C. DBSCAN:

DBSCAN (Density-Base Spatial Clustering of Applications with Noise) is a popular unsupervised learning method which is used in machine learning problems. It is used to separate clusters with high density from clusters with low density. DBSCAN can differentiate between clusters of any shape as opposed to k-means which assume a convex shape. The concept of DBSCAN is based on Core points (samples which are in the areas of high density), Border points (samples which are not Core-points themselves but within the vicinity of a Core-point defined by a threshold) and Noise points (samples which are not the part of the cluster or farther than a Core-point beyond a threshold distance) [7]. The algorithm needs two parameters,threshold and a minimum number of points within the threshold a point should be surrounded by to be called a Core-point.

Working Principle of DBSCAN:

- It divides the dataset into **n** dimensions. DBSCAN form an n-dimensional shape around the data point for each point and then counts how many points fall in that shape.
- It then names this shape as a cluster and iteratively expands the cluster by going over each point in the cluster and counting the number of points nearby. It does so until no more points are nearby and then starts forming the next cluster.

D. Cluster Validation Measures

Cluster validation plays an important and vital step in cluster analysis. After Clustering all our data the main step is to validate the results of the clustered data in terms of accuracy and its performance and to measure its accuracy and the quality there are two types of validity indices: external indices and internal indices. An external index is a measure of agreement between two partitions where the first partition is the a priori known clustering structure, and the second results from the clustering procedure [8]. Internal indices are used to measure the goodness of a clustering structure without external information [9].

In External Indices, results are evaluated of an algorithm based on the structure of clusters labels and for Internal Indices results are evaluated using features and quantities inherited in the data set. Internal validity index can also validate the optimal number of cluster that can be used for clustering the data set.

In our experiment, we have used some internal indices since we had no prior clustering structure i.e. ground truth labels are not known. Therefore, we have used four of the internal indices as mentioned and explained below.

- **Dunn Index:** : It is a metric for assessing a clustering algorithm . It measures the compactness that is, measuring maximum distance between the data points of the cluster and its separation (minimum distance between the clusters) [10].

Mathematical Representation:
Dunn index is defined as the quotient of ratio of d_{min} to d_{max} for a cluster C where d_{min} is the minimal distance between points of different clusters and d_{max} is the largest within-cluster distance. The distance between clusters C_k and $C_{k'}$ is measured by the distance between their closest points:

$$d_{kk'} = \min_{i \in I_k, j \in I_{k'}} ||{M_i}^k - {M_j}^{k'}|| \quad (2)$$

d_{min} is the smallest of these distances d_{kk} :

$$d_{min} = \min_{k \neq k'} d_{kk'} \quad (3)$$

For each cluster C_k , D_k is denoted as the largest distance that separates two distinct points in the cluster (also known as the diameter of the cluster)

$$D_k = \max_{i,j \in I_k, i \neq j} ||{M_i}^k - {M_j}^k|| \quad (4)$$

For Distance D_k , d_{max} is the largest.
Dunn index(c) is defined as:

$$c = \frac{d_{min}}{d_{max}} \quad (5)$$

- **Silhouette Coefficient :** It is a method of validation and interpretation to examine the consistency within the clusters of data. Its value gives an information of how well the data point is classified. It is a measure of similar a data sample is to its own cluster known as cohesion compared to other clusters, that is, separation. To calculate its value mean inter-cluster distance and mean nearest-cluster distance for each data sample is used.

Mathematical Representation:
For each object x_i we define:
s_i - mean distance to objects in the same cluster.
d_i - mean distance to objects in the next nearest cluster.
Silhouette Coefficient for x_i:

$$Silhouette_i = \frac{d_i - s_i}{\max(d_i, s_i)} \quad (6)$$

Silhouette Coefficient for x_i,....,x_N:

$$Silhouette = \frac{1}{N} \Sigma_{i=1}^{N} \frac{d_i - s_i}{\max(d_i, s_i)} \quad (7)$$

The value of the coefficient ranges from -1 to 1 , 1 being the best and -1 (indicating that the sample is assigned to the wrong cluster, as the different cluster is more similar to which it is clustered) being the worst value for the clustered data. And if we get its value as 0 then it indicated that the two clusters are overlapped onto one another. But it also has a drawback its value is generally higher for convex clusters as compared to clusters obtained using density based clustering (DBSCAN).

- **Calinski Harabaz Score:** It is also known the Variance Ratio Criterion based on the concept of dense and

well-separated clusters because it compares the variance between within-class dispersion and between-class dispersion. It is used to evaluate the optimal number of clusters.

Mathematical Representation:

$$s(k) = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-k}{k-1} \tag{8}$$

For K clusters , The C-H score (s) is calculated as the ratio of mean of between cluster dispersion and within cluster dispersion.

B_k is the between group dispersion matrix and W_k is within cluster dispersion matrix.

$$W_k = \Sigma_{q=1}^{k} \Sigma_{x \in C_q} (x - c_q)(x - c_q)^T \tag{9}$$

$$B_k = \Sigma_q n_q (c_q - c)(c_q - c)^T \tag{10}$$

N is the number of points in our data set, C_q set of points in the cluster q, c_q centre of cluster q, c the centre of E, n_q number of points in cluster q.

Calinski Harabaz score is fast to compute and is higher when the cluster is well-separated and dense. But it favours convex clusters.

- **Davis Bouldin Score:** Same as Dunn Index, silhouette score and Calinski-Harabasz index, Davis Bouldin (D-B) score is also based on the cluster itself rather than external labels. And is calculated using the ratio between the within cluster distance and between cluster distance and then computes the overall average of the clusters. It is simple to compute as compared to the other scores or indexes. It is bounded from 0 to 1 and low Davis Bouldin score is considered as a better score. As it measures the distance between clusters centroids it is confined to using Euclidean distance function [11].

Mathematical Representation:

$$DB = \frac{1}{n} \Sigma_{i=1}^{n_c} R_i \tag{11}$$

$$R_i = \max_{j=1,..n_c, i \neq j} (R_{ij}), i = 1, .., n_c \tag{12}$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{13}$$

$$d_{ij} = d(v_i, v_j), s_i = \frac{1}{||c_i||} \Sigma_{x \in c_i} d(x, v_i) \tag{14}$$

where d(x,y) is the Euclidean distance between x and y, c_i is the cluster i, v_i is the centroid of cluster c_i, $||c_i||$ refers to the norm of c_i

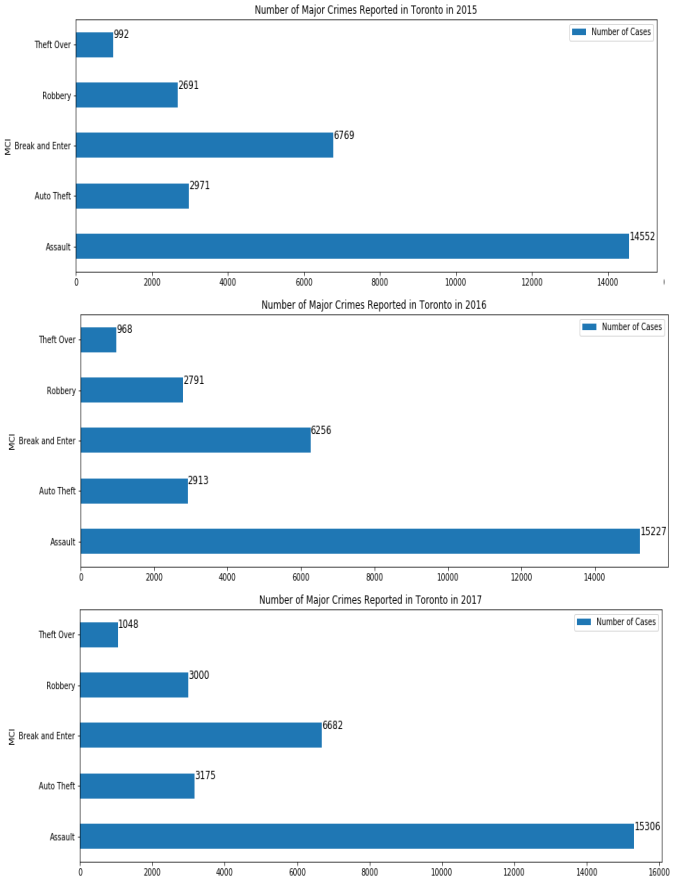


Fig. 1: Major Crime Indicators in year 2015, 2016 and 2017 respectively.

V. Experimentation and Results

A. Data Exploration

The data is retrieved from Toronto Police Public Portal [12]. The data-set has over 131000 entries of reported crimes with 29 different parameters. The data spans from year 2000 - 2017. For each reported major crime indicators there exists a unique ID, time and place of occurrence. Upon initial exploration it was found that their were 17165 duplicate unique IDs in the data-set which were removed. The information given in various features had considerable overlap regarding the nature, place and time of crime. For our analysis we consider these features as redundant features. Their were some features which gave information regarding the time and date of reporting. For our analysis we consider these features as irrelevant features. Hence, both these redundant and irrelevant features were removed from the data-set which further reduced the total number of features in the data-set. Moreover, it was found that high frequency of crimes occurred during the years 2015, 2016 and 2017. Hence, a separate analysis is made on criminal occurrences happened during these years individually in the remaining part of the paper.

Fig. 1 shows the different number and types of crime occurred during years 2015, 2016 and 2017 respectively. It was

MCI	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
Neighbourhood					
Agincourt North (129)	63.0	26.0	55.0	30.0	6.0
Agincourt South-Malvern West (128)	83.0	26.0	61.0	19.0	9.0
Alderwood (20)	37.0	16.0	26.0	6.0	4.0
Annex (95)	245.0	14.0	127.0	44.0	26.0
Banbury-Don Mills (42)	60.0	17.0	82.0	11.0	11.0
Bathurst Manor (34)	48.0	27.0	44.0	8.0	7.0
Bay Street Corridor (76)	382.0	18.0	117.0	30.0	23.0
Bayview Village (52)	89.0	16.0	36.0	5.0	8.0
Bayview Woods-Steeles (49)	37.0	7.0	33.0	1.0	1.0
Bedford Park-Nortown (39)	36.0	35.0	65.0	6.0	12.0

Fig. 2: Reformatted data-set according to MCI.

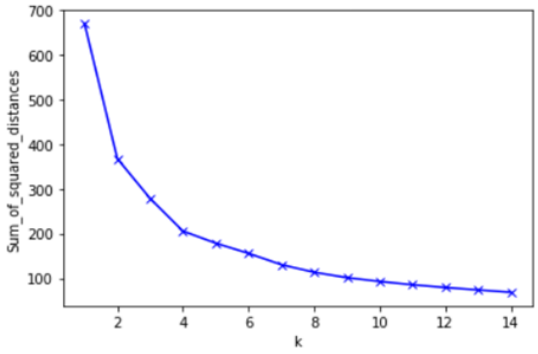


Fig. 3: Plot of Sum of Squared Distances and k for k-means clustering.

observed that Assault had highest number of occurrences for all the three years. The data-set was restructured in a way that major crime indicators were grouped on the basis of Toronto's neighbourhoods. Fig. 2 shows the reformatted data-set. This was done in order to produce a mapping between the crime types and occurrences with different regions of the city.

B. K-Means Clustering:

K-Means is one of the most powerful and easy to implement unsupervised clustering technique. The motive here is to assign a cluster to each neighbourhood. It partitions neighbourhoods into k clusters such that each neighbourhood belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The only challenge in this technique is to decide upon the value of k that is number of clusters must be known prior to applying the clustering algorithm. Even the number of clusters in this problem is well understood as there are only two groups of neighbourhood named violent and non-violent, still elbow method analysis and silhouette score study was applied to the data-set to come up with suitable value of k.

Fig. 4 shows the result of elbow method. As it can be seen there is no clear elbow (point where sum of squared distances drops suddenly for a value of k) in the graph therefore, silhouette score analysis is applied as a extension to elbow method. Fig. 4 displays the findings of silhouette score experimentation and suggest k = 2 as the best value of k for k-means clustering.

Number of clusters	Average Silhouette Score
2	0.6816
3	0.5116
4	0.4350
5	0.3422

TABLE I: Average Silhouette scores for various number of clusters.

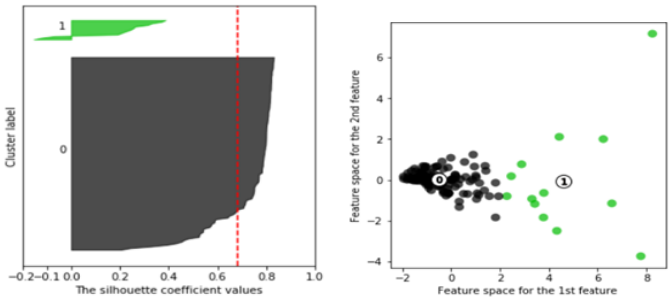


Fig. 4: Silhouette Analysis for k-means clustering with two clusters

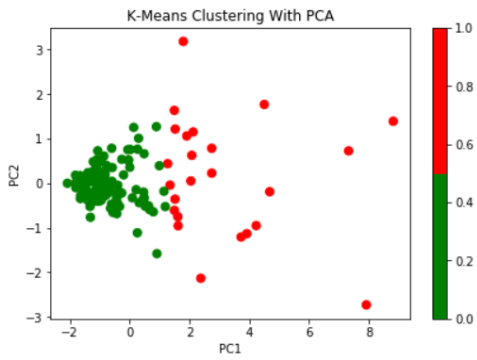
[2] explains how applying PCA prior to k-means clustering can help in getting better clusters and also improved running times. This paper also tries to implement and apply PCA to the data prior to clustering for better visualization of produced clusters over the 2D plane. PCA captures the features with highest point of variance and tries to reduce the dimensionality by extracting only these features. 3 principal components are selected based on highest eigenvalues such that maximum variation of the data is captured. Results obtained after clustering are plotted along 2 principal components. Fig. 5 provides and compares clustering results with application of PCA.

Table II summarises the results of various internal validation measures applied over the clusters formed by K-Means clustering.

C. Agglomerative Clustering

This clustering technique builds nested clusters by merging or splitting them successively. A Hierarchy of clusters represented as dendrograms (tree structures) is formed where the root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. The advantage of this method over k-means is that no prior knowledge of number of clusters is required. Hence, it becomes suitable for the applications where number of clusters are unknown and then dendograms can suitably define the number of clusters. Basically, agglomerative clustering makes clusters (by merging) based on linkage strategy. Ward linkage strategy was used in this experiment to make clusters because it gave the best structure of clusters when compared to other linkage techniques.

Fig. 7 presents and compares the results of agglomerative clustering with application of PCA. It can be seen from the figure that neighbourhoods are grouped in two major clusters namely violent and non-violent.



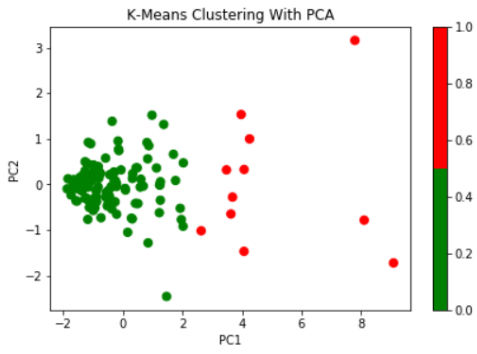
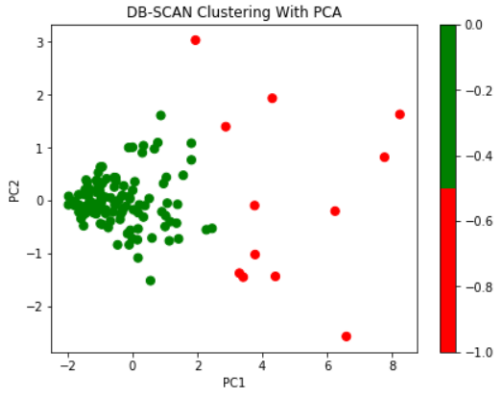
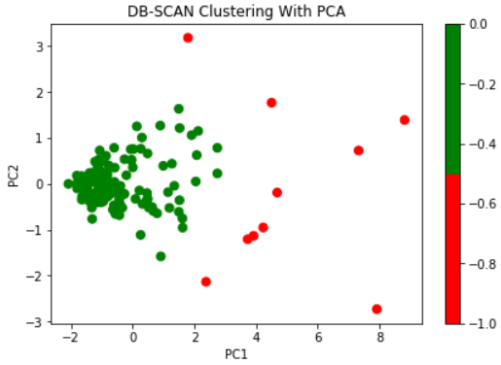


Fig. 5: K-Means for years 2015, 2016, 2017





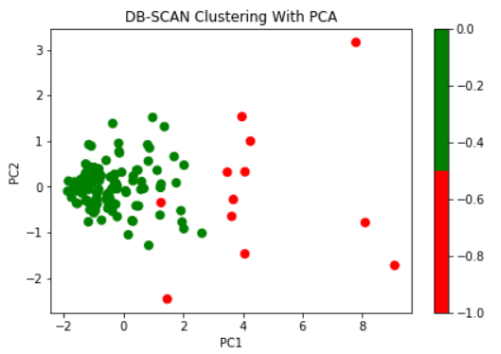
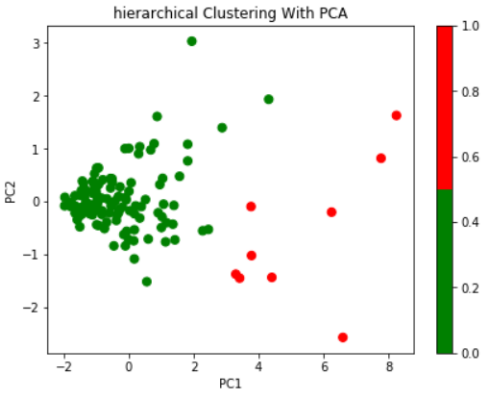


Fig. 6: DBSCAN for years 2015, 2016, 2017



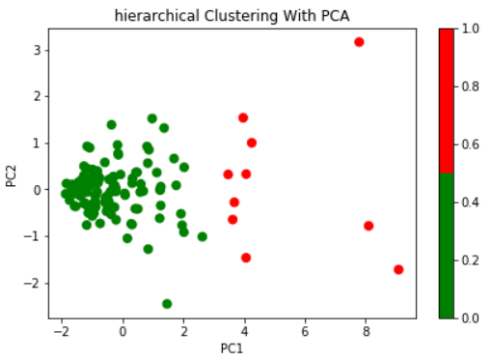


Fig. 7: Hierarchal for years 2015, 2016, 2017

K MEANS			
Validation	**2015**	**2016**	**2017**
Silhouette	0.681	0.587	0.703
Dunn Index	0.088	0.051	0.091
C-H Score	121.4	110.4	116.7
D-B Score	0.798	0.905	0.783

TABLE II: Internal measures K-Means

DB-SCAN			
Validation	**2015**	**2016**	**2017**
Silhouette	0.691	0.693	0.683
Dunn Index	0.120	0.151	0.119
C-H Score	115.2	95.52	102.8
D-B Score	0.838	0.883	0.867

TABLE III: Interval measures DBSCAN

Hierarchal (Agglomerative)			
Validation	**2015**	**2016**	**2017**
Silhouette	0.703	0.705	0.721
Dunn Index	0.120	0.151	0.150
C-H Score	100.9	101.0	115.5
D-B Score	0.766	0.815	0.775

TABLE IV: Internal measures Hierarchal

Fig 8: Most violent regions of Toronto for the year 2015, 2016 and 2017 respectively.

2015 Violent Regions: Annex, Bay street corridor, Church- Yonge corridor, Islington-City Centre West, Kensington-Chinatown, Water-front Communities-The Island

2016 Violent Regions: Annex, Bay street corridor, Church- Yonge corridor, Islington-City Centre West, Moss Park, South Riverdale, Water-front Communities-The Island

2017 Violent Regions: Annex, Bay street corridor, Church- Yonge corridor, Islington-City Centre West, Moss Park, South Riverdale, Water-front Communities-The Island, Woburn, York University Heights

Table IV summarises the results of various internal validation measures applied over the clusters formed by Agglomerative clustering.

D. DBSCAN

It views clusters as areas of high density separated by areas of low density. DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

Parameters:

- **eps:** the minimum distance between two points. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbours.
- **minPoints:** the minimum number of points to form a dense region.

For this experiment value of 1.2 was chosen as eps and value 4 as minPoints. These values gave the best structure of the clusters as the end result. Fig. 6 shows and compares the results of DBSCAN clustering with application of PCA.

Table III summarises the results of various internal validation measures applied over the clusters formed by DBSCAN clustering.

Comparing the validation scores for all metrics of every clustering method, hierarchical clustering is ranked as the most suitable clustering technique for this data-set. In the next section, the results from the hierarchical clustering are plotted over the Toronto city map such that better visualization for violent neighbourhoods can be made for both police authorities and the general public

Conclusion and Future Scope

The spatial analysis of crime in the city of Toronto demonstrates interesting relationships between police-reported crime and neighbourhoods associated with them. Outcomes of this analysis have shown how certain neighbourhood characteristics are related to a higher degree of crime rates. Data analysis techniques such as clustering have been used extensively to extract hidden relationships in the data. Clustering methods such as K-means, Agglomerative , DBSCAN were applied post to application of PCA on the dataset. Hierarchical clustering was ranked as the most suitable method for clustering this data based on several internal validation measures. Several visualisation techniques were used in order to represent the cluster profiles. Different areas of Toronto city were grouped into two clusters namely violent and non-violent based on the year and location of criminal occurrences. On the course of years from 2015 to 2017 the number of violent neighbourhoods increase from 7 to 10 in the consecutive years. The neighborhoods of Annex, Bay Street Corridor, Church-Yonge Corridor, Islington-City Centre West and Waterfront Communities-The Island were always included in the list of violent neighbourhoods as visualized in the graphs whereas Woburn and York University Heights were newly added to the list of violent neighborhoods in 2017. Finally, based on the information retrieved, a heatmap was plotted which graphically superimposes the clusters on the actual map of Toronto City. This study will aid in identifying crime and predicting dangerous hotspots at a certain time and place and also in proper planning and safety measures to stop the antisocial activities from happening in the community.

In the future, based on the results got for most violent neighbourhoods we can find the reasons and map relationships for the violent activities. A study can be done on the affect of population and socio-economic status of the neighbourhoods on crime happenings. We can also scale our study on province level and group violent and non violent neighbourhoods for other cities of Ontario to get a broader image of criminal activities.

References

[1] Charron, Mathieu. "Neighbourhood characteristics and the distribution of police-reported crime in the city of Toronto. Statistics Canada" , 2009.

[2] Kou, Gang, Yi Peng, and Guoxun Wang. "Evaluation of clustering algorithms for financial risk analysis using MCDM methods." Information Sciences 275 (2014): 1-12.

[3] Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[4] Pham, Duc Truong, Stefan S. Dimov, and Chi D. Nguyen. "Selection of K in K-means clustering."Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219.1 (2005): 103-119.

[5] Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

[6] A. Jain, R.C. Dubes, "Algorithms for clustering Data" in , Prentice Hall, 1988.

[7] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[8] Dudoit, Sandrine, and Jane Fridlyand. "A prediction-based resampling method for estimating the number of clusters in a dataset." Genome biology 3.7 (2002): research0036-1.

[9] Thalamuthu, Anbupalam, et al. "Evaluation and comparison of gene clustering methods in microarray analysis." Bioinformatics 22.19 (2006): 2405-2412.

[10] Desgraupes, Bernard. "Clustering indices." University of Paris Ouest-Lab ModalX 1 (2013): 34.

[11] Davies, David L., and Donald W. Bouldin. "A cluster separation measure." IEEE transactions on pattern analysis and machine intelligence 2 (1979): 224-227.

[12] http://data.torontopolice.on.ca/