DATA MINING FOR SPATIO-TEMPORAL RULES WITH GEOSPATIAL ANALYSIS OF CRIME PATTERNS

A PROJECT REPORT

Submitted by

VATSAL GUPTA [17104060]
ANANYA SHARMA [17104038]
ARJAV JAIN [17104070]

Under the guidance of

Mr. M. Gurve

(Assistant Professor, Department of Computer Science & IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Sector 62, Noida, Uttar Pradesh 201309

NOVEMBER 2019

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

(Deemed to be University under section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled "**DATA MINING FOR SPATIO-TEMPORAL RULES WITH GEOSPATIAL ANALYSIS OF CRIME PATTERNS** " is the bonafide work of "**VATSAL GUPTA [17104060], ANANYA SHARMA [17104038], ARJAV JAIN [17104070]**", who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mr. M. Gurve
SUPERVISOR
Assistant Professor
Dept. of Computer Science & IT

SIGNATURE

Dr. V. Saxena
HEAD OF THE DEPARTMENT
Dept. of Computer Science & IT

SIGNATURE

Ms. A. Kaur
PANEL MEMBER
Dept. of Computer Science & IT

SIGNATURE

Dr. A. Sahoo
PANEL MEMBER
Dept. of Computer Science & IT

ABSTRACT

Crime has always been a universal phenomenon. However, with the rapid urbanisation and industrialisation, there has been a surge in techno-industrial-urban complexes which offer a setting conducive to crime. Consequently, instances of crime have increased rapidly all across the globe among all segments of society. Furthermore, new forms of crime are emerging, and old forms are assuming new dimensions. Even though law enforcement agencies are doing their utmost to prevent crimes and catch criminals, they still face an uphill task. Comprehensive data is required to identify the criminals who are well organised and well equipped in using contemporary techniques that pose a severe threat to people's safety. Numerous investigations addressing this issue have generally employed disciplines of behaviour science and statistics. Recently, the data mining approach has proved to be a proactive decision-support tool in analysing, predicting and preventing crime. In this work, a framework based on clustering and association rule mining has been proposed to detect and analyse crime trend patterns from temporal and spatial crime activity data. In addition, an open-source Geographic Information System (GIS) application, QGIS, is employed to reveal the overall crime hotspots as well as the hotspots of certain violent crimes in isolation. Furthermore, time series analysis of the crime has been done in order to analyse the changing patterns of crime with time. The resultant model can support police managers in developing more appropriate law enforcement strategies, as well as enhancing the use of police duty deployment for crime prevention.

PREFACE

This project aims to employ crime analysis techniques to enable the police administrators to objectively determine the nature of criminal activities in their jurisdictions and allow them to develop tactical action plans to combat them effectively. At the same time, this project can also help the public at large in being more aware and consequently safer. Doing this project has helped us to enhance our knowledge of data mining and data analyses techniques. Besides, this project has equipped us to learn a free and open-source GIS tool named QGIS, and, has contributed to our increased fluency in the python environment and its libraries.

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to our mentor, Mr. Mahendra Gurve, for his valuable guidance, consistent encouragement, timely help, and providing us with an excellent atmosphere for doing our project. During the entire duration of our dissertation work, despite his busy schedule, he has extended cheerful and cordial support to us for completing this project work. We would also like to convey our sincere regards to all other faculty members of the department of CSE & IT, JIIT, who have bestowed their great effort and guidance at appropriate times. Our thanks and appreciations also go to people who have willingly helped in developing this project.

Contents

List of Tables

List of Figures

ABBREVIATIONS

GIS	Geographic Information System
FP	Frequent Pattern
HDD	Hard Disk Drive
SSD	Solid State Drive
IUCR	Illinois Uniform Crime Reporting
EDA	Exploratory Data Analysis
WCSS	Within Cluster Sum of Squares
SQL	Structured Query Language
TAR	Temporal Association Rules
SVM	Support Vector Machines
KNN	K Nearest Neighbour

Chapter 1

INTRODUCTION

1.1 Problem Statement

Increased population, technological advancements and heightened competition for economic resources have created various social problems. Many of these changes in the human condition have brought new challenges to the doorstep of the law enforcement profession that begs for resolution. The major challenge facing law enforcement agencies is to deal with the increased number of criminal activities effectively and efficiently. Current policing strategies work towards finding the criminals, basically after the crime has occurred. However, with the help of technological advancement, we can use historical crime data to recognise crime patterns [8]. If enforcement agencies have a prior assumption of the class of the crime, it would give them tactical advantages and help resolve cases faster. An overall study of criminal activity in a geographic area also helps in understanding the underlying pattern of the crime in that area.

1.2 Research Motivation

Criminals have been a nuisance for society in all corners of the world for a long time now. Measures are required to eradicate crimes from all over the world or at least limit its occurrence. The large volumes of crime datasets, as well as the complexity of associations between these kinds of data, have made criminology an appropriate field for the application of data mining techniques. Criminology is an area that focuses on the scientific study of crime, criminal behaviour, and law enforcement. In simpler terms, it is a process that aims to identify crime characteristics [10]. It is one of the most relevant fields where the application of data mining techniques can produce remarkable results that can help and support police forces. Using data mining and data analytics techniques, we can analyse the crime patterns that can further help the authorities to understand the underlying reasons for the occurrence of crime and can, therefore, help them to a large extent in the prevention of future crimes.

1.3 Research objectives

The primary objective of our work is to analyse criminal data based on demographics, spatial and temporal information and consequently identify useful crime patterns to aid police in preventing crimes. Towards this end, we have employed data mining and crime mapping techniques. The main objectives of our project work are summarised as follows:

1. Identifying the crime patterns based on a criminal dataset that contains the geographical location and basic details of the criminal activity.
2. Exploring data mining techniques to generate association rules for crime analysis.
3. Visualising these patterns on an open source GIS software - QGIS for better understanding of the results.

Chapter 2

LITERATURE SURVEY

Much work has been done in the direction of crime analysis to improve the activities aimed at detecting and preventing safety problems for the public. Techniques ranging from conventional data association methods to the modern approaches of data mining have been applied to this field. This section aims to summarise the work done in this regard.

In [2], automated approaches to data association to increase the accuracy of crime prediction have been proposed. Results included in the paper indicated that the employed data association methods significantly reduced the time required by manual methods while maintaining a high level of accuracy, comparable to that of experienced crime analysts. Furthermore, in contrast to existing analysis techniques that employed Structured Query Language (SQL), these methods were both faster and more accurate.

Nath (2006) proposed a clustering technique over supervised learning techniques such as classification for crime data analysis [16]. In this work, K Means clustering has been applied to identify criminal patterns and subsequently to help prevent future crimes. Furthermore, the clustering algorithm has been integrated with a geospatial plot using which a crime analyst can choose a time range and one or more types of crime from specific geography and view the result graphically.

The concept of Temporal Association Rules (TAR) was introduced in [17] to solve the problem of time series handling by including time expressions into association rules. An incremental algorithm - ITAR (Incremental TAR) has been proposed in this paper to overcome the re-scanning issue in the existing TAR algorithm for updating the dataset. This paper made use of negative border method for preserving temporal association rules with numerical attributes. The temporal negative border method proposed in this paper only retains all the past winners who become losers in subsequent rounds instead of maintaining a power set of dense base cubes. As a result, the number of losers of base cubes held by the negative temporal boundary was minimised. Preliminary results showed a significant improvement over the recurrent TAR algorithm and showed that ITAR is very stable and has good performance.

[11] examined the performance of two types of Support Vector Machines (SVM) techniques: two-class SVMs and one-class SVMs for predicting the location of crime hotspots when a predefined level of crime rate is given. The paper also compared an SVM with a neural network-based approach and a spatial auto-regression-based approach. Initially, K means clustering has been used as a data selection approach. After labelling the data, the resulting dataset has been used as the input of the SVM algorithm. In both one-class and two-class SVMs, different kernel functions were used to determine the accuracy of the classification. Their experiments have shown that one-class SVMs produces reasonable results, particularly when the size of the training set is small with more positive samples. However, for larger datasets and more negative samples, two-class SVMs were found to have a better performance.

A framework of intelligent decision-support model based on a fuzzy self-organising map (FSOM) network to detect and analyse crime trend patterns from temporal crime activity data was proposed in [14]. Besides, a rule extraction algorithm has been employed to uncover hidden causal-effect knowledge and reveal the shift around effect. As per the analysis of the experimental results, they discovered characteristics of four crime patterns, namely, typical, gradual increase, sharp increase, and Wintertime. Their results showed that their framework could help provide vital information to the police management for determining the kind of duty deployment that should be employed. However, a significant limitation of their study was the lack of evaluation metrics to evaluate the accuracy of their model.

[19] discussed the preliminary results of a crime forecasting model developed in collaboration with the police department of a United States city in the Northeast. Their datasets comprised of aggregated counts of crime and crime-related events categorised by the police department. The location and time of these events were also included in the dataset. This work employed various data classification techniques to perform crime forecasting and to determine the best classification method for predicting crime hotspots. Their results indicated that 1NN classifier modified with location constraint is better in finding similar circumstances in a neighbourhood but not in the entire city. Furthermore, their results found Naïve Bayes classifier to be the best probability predictor.

In [7], two classification methods, namely, Naïve Bayes and Decision Tree, have been applied to a crime dataset to predict 'Crime Category' for different states of the United States of America. The results obtained from the experiment indicated that the Decision Tree algorithm

outperformed the Naïve Bayesian algorithm and achieved 83.9519% Accuracy in predicting 'Crime Category' for different states of the US.

Various strategies to find spatial and temporal criminal hotspots have been employed in [1]. Freely available crime datasets of two cities, namely, Denver (Colorado) and Los Angeles (California) have been used in this paper. The crime patterns in this paper were achieved by applying the Apriori algorithm on both the Denver and the Los Angeles datasets based on a predefined threshold. 62 interesting frequent patterns were generated for Denver, whereas 59 interesting frequent patterns were generated for Los Angeles. Two classification methods, namely, Naïve Bayes and Decision Tree classifiers, were also employed for crime type prediction. Their results indicated that Naïve Bayes classifier had better accuracy (51%) as compared to the Decision Tree classifier (42%). Furthermore, this paper presented a statistical analysis of the dataset supported by several graphs.

Spatio-temporal and demographic data has been used in [3] to predict which category of crime is most likely to have occurred at a given time and place. Various classification algorithms such as Naive Bayes, Support Vector Machines, Gradient Boosted Decision Trees, and Random Forests have been applied and compared. The inputs to these algorithms were time (hour, day, month, year), place (latitude, longitude, and police district), and demographic data (population, median income, minority population, and the number of families). The output produces was the category of the crime that is likely to have occurred. Classification of blue-collar and white-collar crimes, as well as violent and non-violent crimes, was also done using the algorithms as mentioned above. The Gradient Boosted Decision Trees algorithm was found to have the best accuracy.

Linear Regression, Additive Regression, and Decision Stump algorithms were implemented in [15] on the same finite set of features. The results indicated that the linear regression algorithm has the best performance among the three selected algorithms. The relatively poor performance of the Decision Stump algorithm was attributed to a certain factor of randomness in the various crimes and the associated features; the branches of the decision trees are more stringent and give accurate results only if the test set follows the pattern modelled. On the other hand, the linear regression algorithm was able to handle the randomness in the test samples to a great extent. This paper explored the efficiency and accuracy of the machine learning algorithms in data mining research for predicting trends of violent crimes.

In [18], a model for estimating the regions with high probability of crime incidence has been proposed. Various data mining techniques, in particular, data clustering algorithms have been employed to extract unknown but useful information from unstructured data. Various clustering techniques like K - Means and Fuzzy C have been used in this paper to take into account the dynamic nature of the crimes.

In [12], Vancouver crime dataset for the last 15 years was used for the application of predictive machine learning models such as K Nearest Neighbour (KNN) and enhanced decision trees were used and gave crime prediction accuracies between 39% and 44%. Owing to the use of different approaches, performance, complexity, and training time of algorithms were found to be different. Although the predictive model used in this paper lacks reliability, it can act as a basis for further studies.

A novel methodology for forecasting crime has been applied in [9]. KNN has been used in this paper to calculate optimal values for good performance. Furthermore, a Bayesian Network was used to establish associations that benefit a range of variables. The results of the simulations indicated that the Naïve Bayes algorithm was highly precise and time-efficient.

Chapter 3

DATASET DESCRIPTION

The dataset used for the purposes of this work has been retrieved from the official government portal of the City of Chicago. The dataset reflects reported incidents of crime (except for murders) that occurred in the City of Chicago from 2001 to November of 2019. Addresses are manifested at the block level only, and specific locations are not identified to preserve the privacy of crime victims. Although the original dataset has 6.6 million rows, owing to the computational constraints, we have chosen a small fraction of this dataset. The description of all the attributes(columns) is given in the table below:

Column Name	**Description**	**Data Type**
ID	Unique Identifier for the record	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description.	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act	Checkbox
Beat	Indicates the beat where the incident occurred.	Plain Text
District	Indicates the police district where the incident occurred.	Plain Text
Ward	The ward (City Council district) where the incident occurred.	Number
Community Area	Indicates the community area where the incident occurred.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).	Plain Text
X Coordinate	The x coordinate of the location where the incident occurred.	Number
Y Coordinate	The y coordinate of the location where the incident occurred	Number
Year	Year the Incident occurred	Number
Updated On	Date and Time the record was last updated	Date & Time
Latitude	The latitude of the location where the incident occurred.	Number
Longitude	The longitude of the location where the incident occurred.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.	Location

Table 3.1: Dataset Columns

Additional information about some attributes -

1. **IUCR**: Illinois Uniform Crime Reporting (IUCR) codes are four-digit codes that are used by law enforcement agencies to classify criminal incidents when taking individual reports. In addition, these codes are used to aggregate types of cases for statistical purposes. The Chicago Police Department currently uses more than 350 IUCR codes to classify criminal offences. The list of IUCR codes is available at https://data.cityofchicago.org/d/c7ck-438e.

2. **Beat**: The smallest police geographic area is known as a beat – each beat has a dedicated police beat car. Three to five beats constitute a police sector, and three sectors further constitute a police district. There are a total of 22 police districts in the Chicago Police Department.

3. **Community Areas**: Chicago has a total of 77 community areas.

4. **Coordinates**: All the coordinates, including latitude and longitude, are projected in State Plane Illinois East NAD 1983 projection. Furthermore, the coordinates (including latitude and longitude) map to a location that is shifted from the actual location for partial redaction but falls on the same block to protect the privacy of the crime victim.

3.1 Summary

(a) **Dataset Name**: Crimes - 2001 to present

(b) **Number of Rows**: 6.6 million

(c) **Number of Columns** = 22

(d) **Dataset Source** - Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

(e) **Dataset Avalaibility** - Publicly Available for free

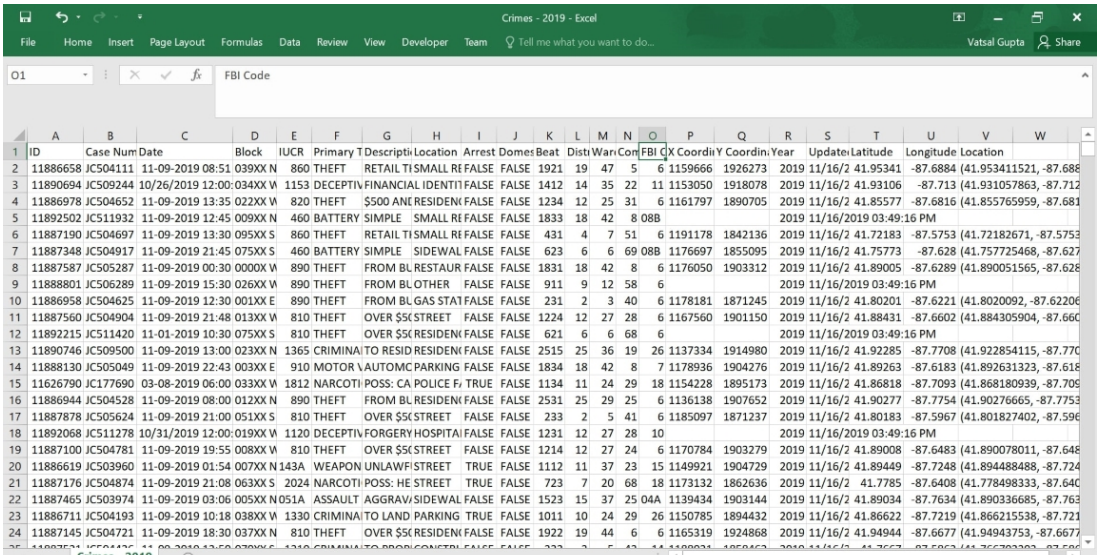


Figure 3.1: CSV Dataset Snapshot

Chapter 4

RESEARCH METHODOLOGY

Data mining techniques include supervised learning methods such as regression and classification algorithms and unsupervised learning methods such as clustering algorithms and association rule mining algorithms. In this work, to analyse the data and observe the crime patterns, three unsupervised learning algorithms, namely, 1) K Means (Clustering), 2) Apriori and, 3) FP Growth (Association Rule Mining) have been used. These algorithms have been applied to a segment of the crime dataset available freely on the official website of the Chicago Police Department. Furthermore, to validate the results obtained from the application of the algorithms as mentioned above, an open-source GIS software - QGIS, has been used to map the crime dataset. In addition to this, various graphical analysis techniques such as hotspot analysis, cluster & outlier analysis and, time series analysis have been performed in order to extract meaningful patterns and characteristics from the crime data. All of the algorithms and techniques, as mentioned above, have been further elaborated in the following sections.

4.1 Clustering

Clustering is an unsupervised data mining strategy to group the relevant data into desired clusters. The centroid of each cluster represents the collection of feature values that define the resulting groups. By analysing in which cluster the data points lie, clustering enables us to gain valuable insights about our data. A clustering technique has been chosen over any other supervised technique such as classification since crimes widely vary in nature, and crime databases are often filled with unsolved crimes. Therefore, a classification technique that will rely only on the existing and known solved crimes will not give good predictive quality for future crimes.

4.1.1 K Means Clustering

K-means clustering is the simplest and most commonly used clustering algorithm owing to its less computational complexity. It is used when the data available is unlabelled (i.e., data

without defined categories or groups) or when the data sets are large. The algorithm works iteratively to assign each data point to one of the K groups based on the features that are provided. Data points are clustered based on feature similarity.

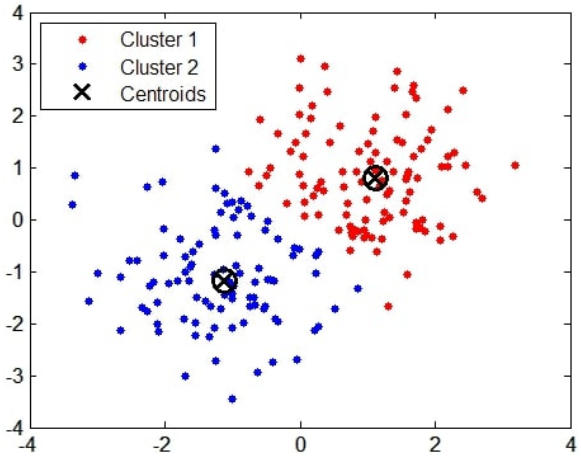


Figure 4.1: K Means Clustering

4.2 Association Rule Mining

Association rule mining is a technique for recognizing frequent patterns and associations among a set of objects. Association rules are used to help uncover relationships between seemingly unrelated data in a set of transactions. An association rule has two components, an antecedent or the LHS (if) and a consequent or the RHS(then). An antecedent is an item that exists in the list of transactions. A consequent is an item that is found in combination with the antecedent. Consider A and B to be two items present in a list of N transactions, where A is the antecedent and B is the consequent, then:

Support: Support of item A is the ratio of the number of transactions in which item A appears to the total number of transactions.

$$Support\,(A) = \frac{frequency\,(A)}{N} \tag{4.1}$$

Confidence: Confidence measures the percentage of times item B is purchased, given that item A was purchased. Values of confidence range from 0 to 1, where 0 indicates that B is never purchased when A is purchased, and 1 indicates that B is always purchased whenever A is purchased.

$$Confidence\,(A \rightarrow B) = \frac{frequency\,(A, B)}{frequency\,(A)} \tag{4.2}$$

Lift: The lift of a rule measures the percentage of times item B is purchased when item A is purchased while controlling the popularity of item A. A lift value of 1 implies that there is no association between the items.

$$Lift\,(A \rightarrow B) = \frac{Support\,(A, B)}{Support\,(A) * Support\,(B)} \tag{4.3}$$

Conviction: It represents by what factor the correctness of the rule would reduce if the antecedent (in this case - A) and the consequent (in this case - B) of the rule were independent. Higher is the confidence; higher is the conviction of a rule.

$$Conviction\,(A \rightarrow B) = \frac{1 - Support\,(B)}{1 - Confidence\,(A \rightarrow B)} \tag{4.4}$$

4.2.1 Apriori

Apriori is the most commonly used association rule mining algorithm to find frequent itemsets. It takes advantage of the fact that any subset of a frequent itemset is also a frequent itemset. The algorithm, thereby, reduces the number of candidates being considered by only exploring the item-sets whose support count is greater than the minimum support count. Steps to implement the Apriori algorithm are as follows:

1. Set minimum support and confidence.
2. Take all the subsets in transactions having higher support than minimum support.
3. Take all the rules of the subsets having higher confidence than minimum confidence.
4. Sort rules by decreasing lift.

4.2.2 FP Growth

FP Growth is a tree based association rule mining algorithm. Unlike Apriori, FP Growth uses a pattern growth approach rather than a candidate generation approach to find frequent itemsets. FP Growth is faster and more efficient as compared to Apriori as it scans the dataset only twice, whereas Apriori scans the dataset multiple times in order to generate frequent itemsets.

4.3 QGIS

A GIS is a system designed for the collection, storage, manipulation, analysis, management, presentation and analysis of spatial or geographical data. GIS applications allow users to analyse geospatial information, edit data in maps and present the results of all these operations. QGIS is one such free and open source cross-platform desktop application. QGIS supports shapefiles, coverages, personal geodatabases, dxf, MapInfo, PostGIS, and other formats. It also supports Web services, including Web Map Service and Web Feature Service, to allow the use of data from external sources. QGIS has a plethora of plugins that enable a dynamic analysis of data including hotspot analysis, time series analysis, cluster & outlier analysis.

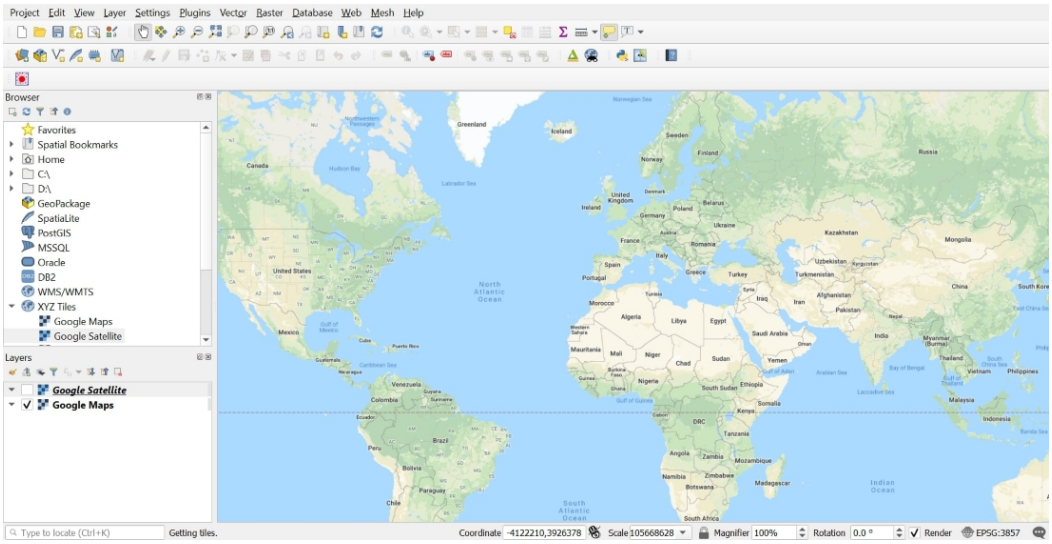


Figure 4.2: QGIS Desktop Snapshot

Chapter 5

RESEARCH DESIGN

5.1 Data Acquisition

The dataset under scrutiny has been acquired from the official website of the Chicago City Government. The dataset contains the information of all the crime incidents from 2001 to 2019 with the exception of murders. A detailed description of the dataset has already been presented in Chapter 3.

5.2 EDA

EDA is a systematic way of visualisation and transformation to explore and summarise the main characteristics of the available data. The main objectives of an EDA include:

- Generating questions about the available data.
- Searching for answers by visualising, transforming, and modelling the data.
- Using the findings to generate new questions.

In our project, the EDA was carried out on the crime data for the current year, i.e., 2019. The following criteria were considered while performing the EDA:

- Number of crimes by day of the week. (Figure 5.1)
- Number of crimes by month of the year. (Figure 5.2)
- Number of crimes by type. (Figure 5.3)
- Frequency of different types of crime in different locations. (Figure 5.4)

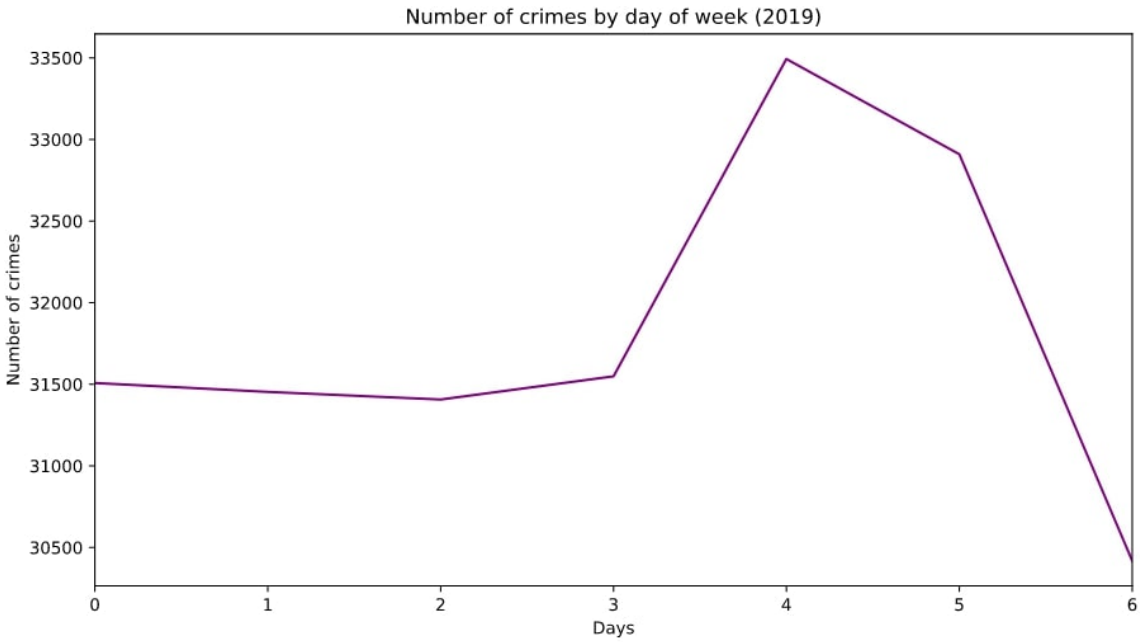


Figure 5.1: Number of crimes by day of the week

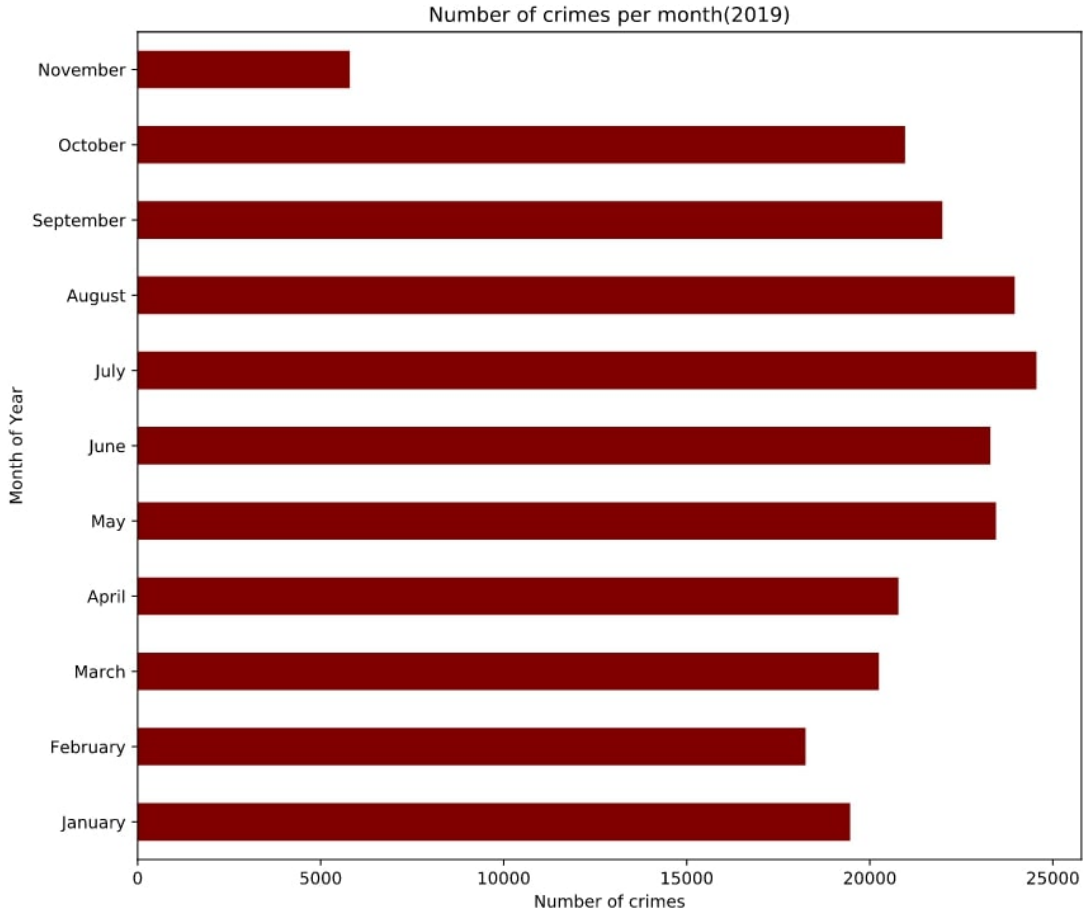


Figure 5.2: Number of crimes by month of the year

Note: November's crime count is less due to the availability of data only until the first week of November.

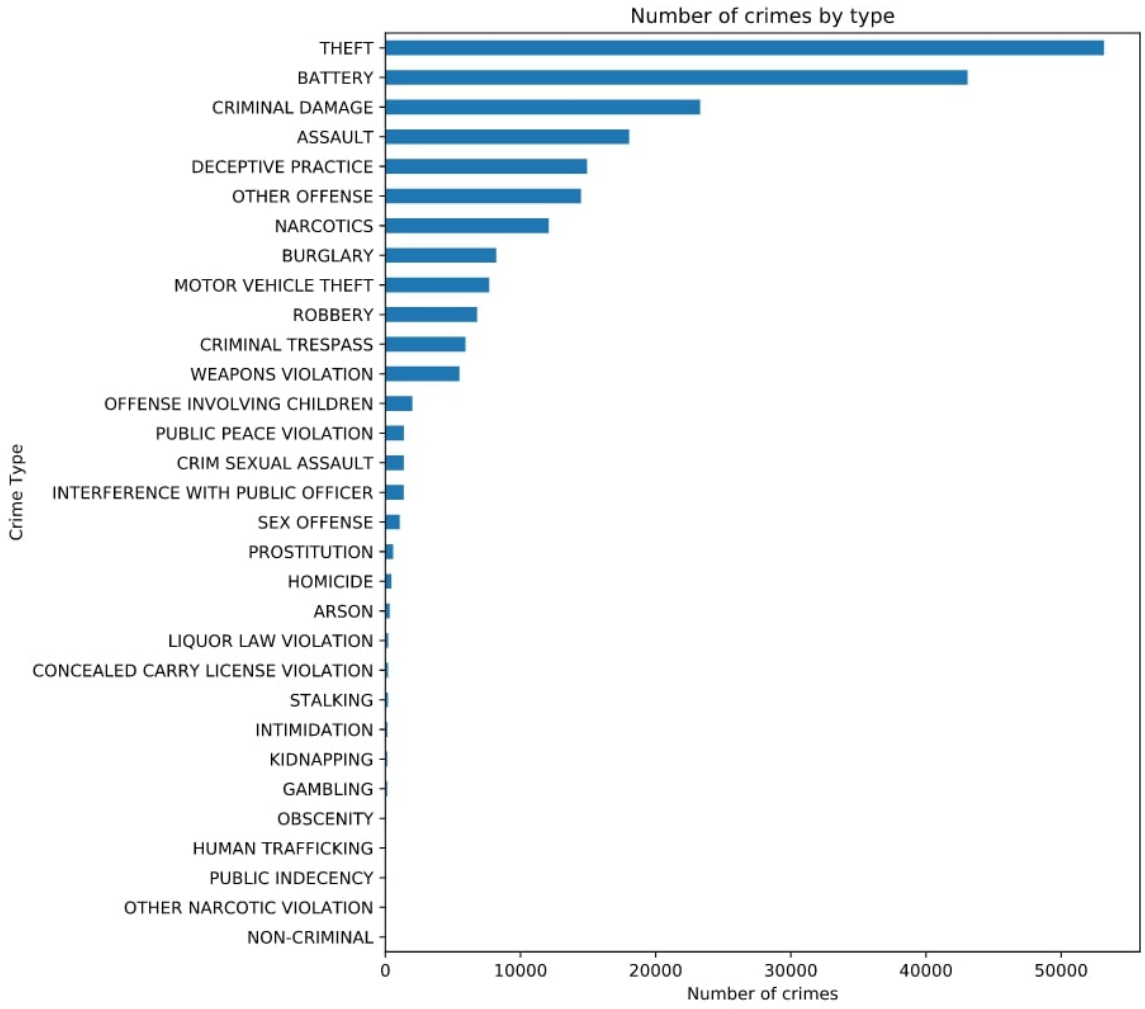


Figure 5.3: Number of crimes by type

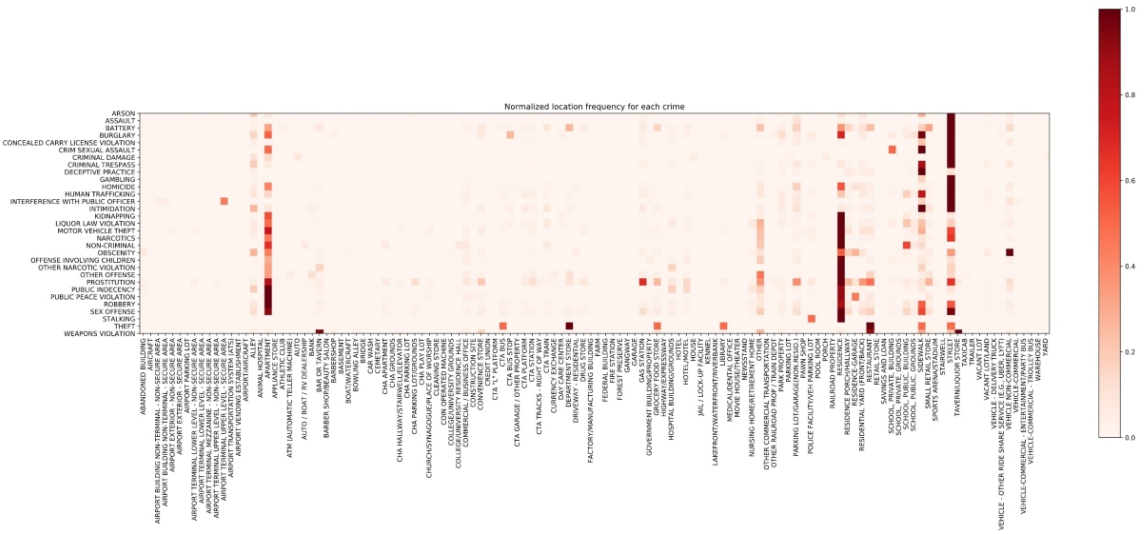


Figure 5.4: Frequency of different types of crimes in different locations

5.3 Data Pre-processing

Data pre-processing is a data mining procedure that involves converting raw data into an understandable format. Real-world data is often incomplete, inconsistent and lacking in certain behaviours or trends and is likely to contain several errors. Data pre-processing is a proven method of addressing these issues. Data Pre-processing steps involved in our project:

1. Dropping redundant columns such as 'X Coordinate', 'Y Coordinate', Location (since 'Latitude' and 'Longitude' are available), 'Updated On' and more.
2. Converting the date and time given in the dataset to pandas DateTime format.
3. For clustering, it was necessary to normalise the time, district and the IUCR code values in order to form clusters based on crime type, time and location. The formula adopted for normalisation is given below:

$$\mathbf{x_{normalised}} = \frac{\mathbf{x} - \mathbf{x_{minimum}}}{\mathbf{x_{maximum}} - \mathbf{x_{minimum}}} \tag{5.1}$$

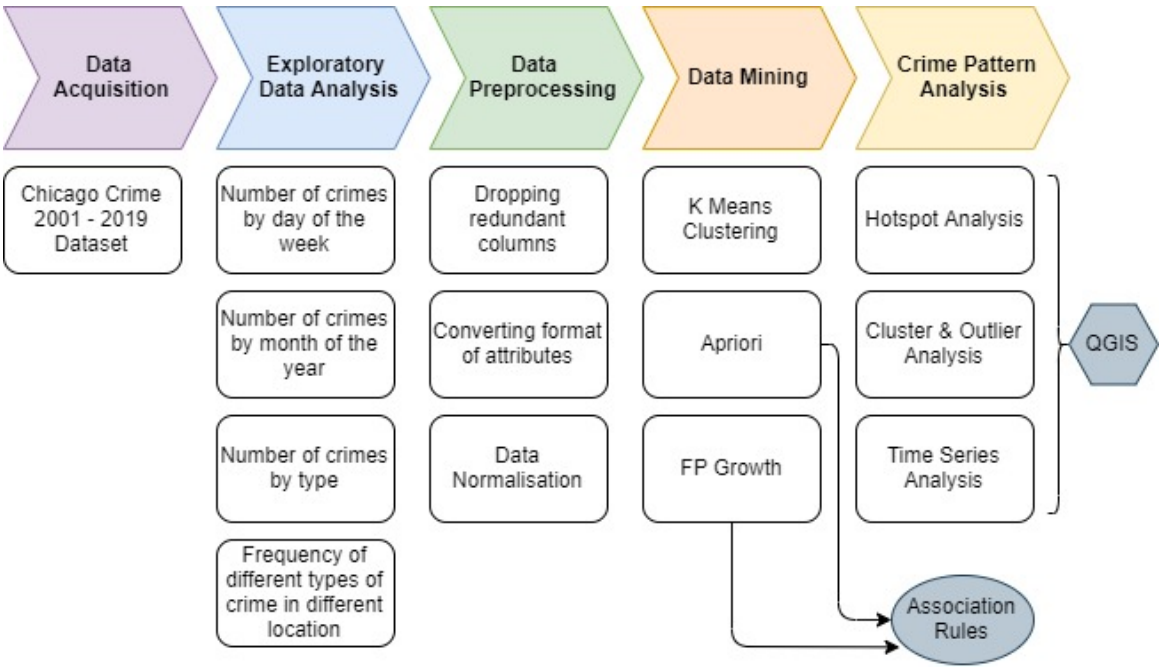


Figure 5.5: Research Framework of our Project

5.4 Data Mining

5.4.1 K Means Clustering

The steps involved while performing K Means clustering on the Chicago crime dataset are given in the figure 5.6.

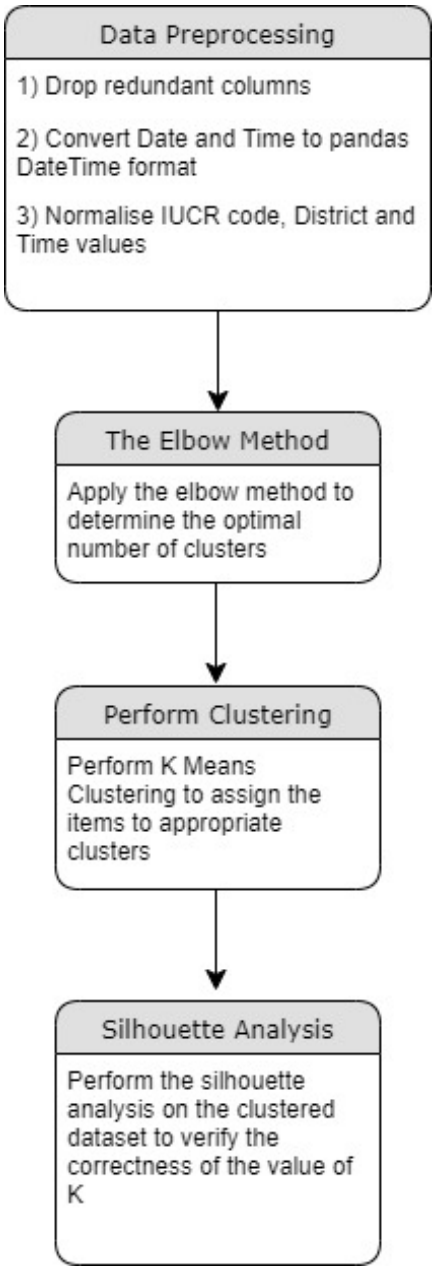


Figure 5.6: Steps involved in K Means Clustering

5.4.2 Apriori

The steps involved in applying the Apriori algorithm have been discussed in Chapter 4.

5.4.3 FP Growth

The flowchart below (figure 5.7) reveals the steps involved in generating association rules from our dataset.

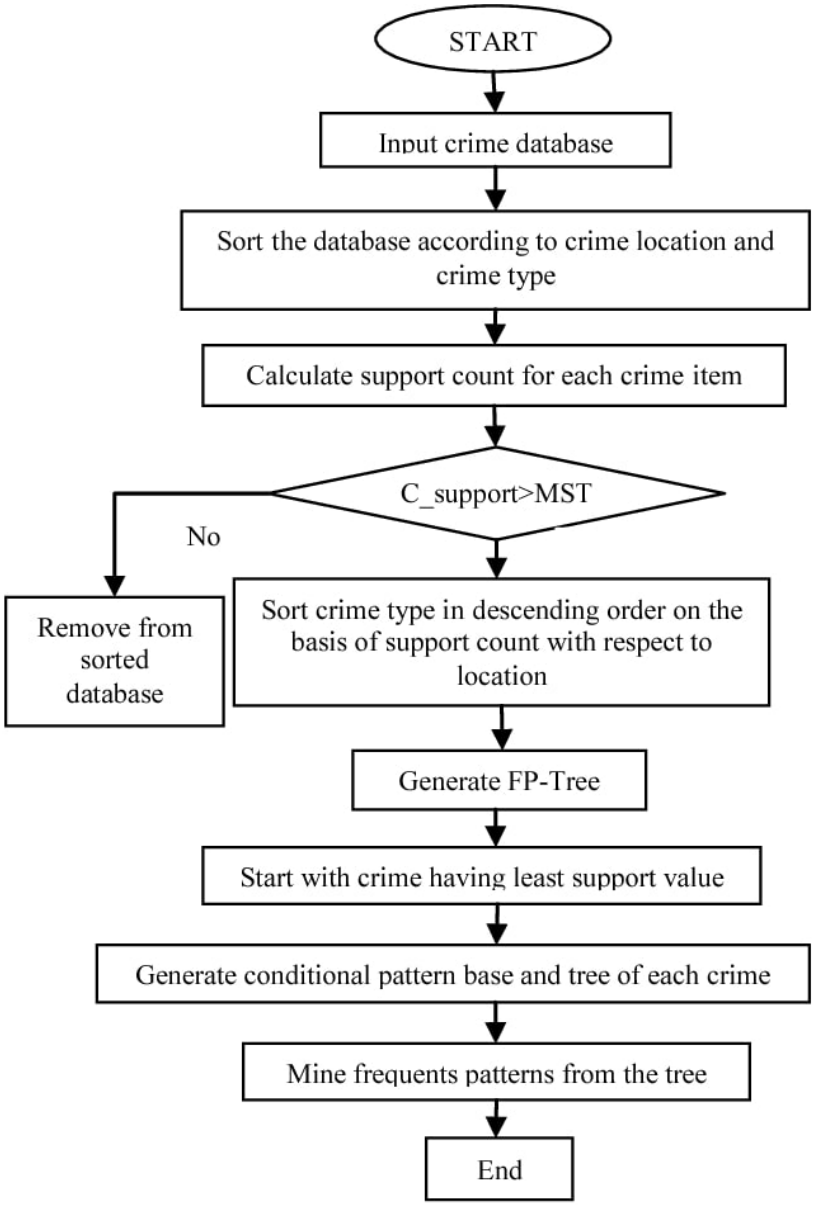


Figure 5.7: Steps involved FP Growth

Chapter 6

PROJECT SPECIFICATIONS

6.1 Hardware Specifications

6.1.1 Recommended Configurations

- **CPU Speed:** 1.4 GHz or higher recommended (per core)
- **Processor:** Intel Core i5-6100 (6th Gen) Dual Core or above
- **Memory/ RAM:** 8GB or higher
- **Monitor/ Display:** 13" LCD monitor, resolution of 1920 x 1080 or better
- **Storage:** 200 GB Hard Disk Drive (HDD)/ 128 GB Solid State Drive (SSD) or above

6.1.2 Minimum Configurations

- **CPU Speed:** 1.2 GHz (per core)
- **Processor:** Intel Core i3-6100 (6th Gen) Dual Core
- **Memory/ RAM:** 4GB
- **Monitor/ Display:** 13" LCD monitor, resolution of 1600 x 900
- **Storage:** 100 GB HDD/ 64 GB SSD

6.2 Software Specifications & Requirements

- **Operating System Used:** Microsoft Windows 10 Version 10.0.17763 64 bit
- **Programming Language Used:** Python 3.7.3 64 bit
- **IDEs/ Environments Used:**
  Spyder 3.3.6
  VS Code 1.40.1
  Jupyter Notebook 6.0.2
- **GIS Application Used:** QGIS Desktop 3.10.0 A Coruña based on Python 3.7.0

6.2.1 Project Compatibility

- **Operating Systems:**
  Microsoft Windows 7 or above
  Mac OS El Capitan or above
  Linux Flavours - Debian/Ubuntu, Fedora, Mandriva, Slackware, ArchLinux, Flatpak, openSUSE, RHEL
- **Python Version:** Python 3.x.x
- **IDEs:** Compatible with all Python IDEs provided all the required packages and modules are installed
- **Recommended IDEs:** Spdyer 3.3.x and VS Code 1.33.1 or above

6.2.2 Python Packages

- pandas
- numpy
- matplotlib
- efficient_apriori
- tkinter
- Ipython
- sklearn
- mpl_toolkits
- plotly
- pyfpgrowth
- yellowbrick
- pysal (for QGIS)
- rtree (for QGIS)
- gdal (for QGIS)
- geopandas (for QGIS)

6.2.3 Python Modules

- time
- datetime
- apriori (from efficient_apriori)
- __future__
- pylab
- Axes3D (from mpl_toolkits)
- webbrowser
- warnings
- SilhouetteVisualizer (from yellowbrick)
- fp (from pyfpgrowth)
- esda (from pysal)
- spreg (from pysal)

6.2.4 QGIS Plugins

- TimeManager
- HotspotAnalysis
- Qgis2threejs
- MapSwipe Tool

How to install the modules and packages in Python?

- In case of standalone python IDE:
  - Using pip3 installer in terminal.
- In case of anaconda environment:
  - Using conda install or pip3 installer in anaconda prompt.

How to install the plugins in QGIS?

1. Download plugins in zip format from https://plugins.qgis.org/
2. Open QGIS Desktop Application and open the 'Plugins' tab
3. Choose 'Manage and Install Plugins'
4. A new window would open, click on 'Abort Fetching'
5. After the completion of the previous step, another window would open. Go To 'Install from ZIP'
6. Select your zip file
7. Click on 'Open'
8. Click on 'Install Plugin'
9. You can verify that your plugin has been installed in the 'Installed' option.

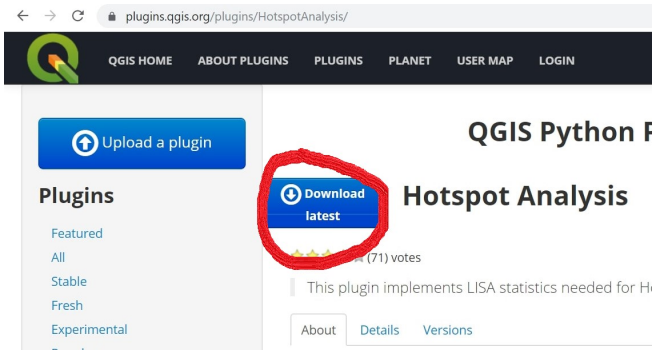


Figure 6.1: Step 1

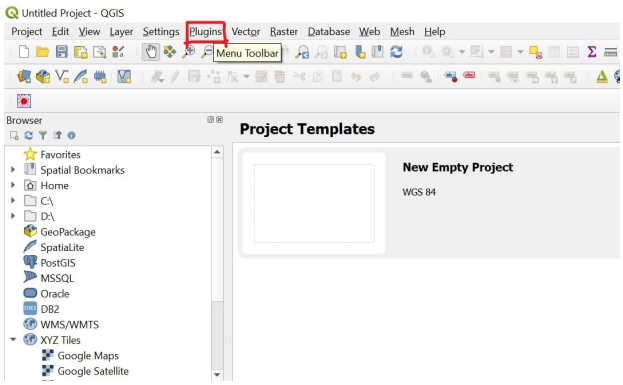


Figure 6.2: Step 2

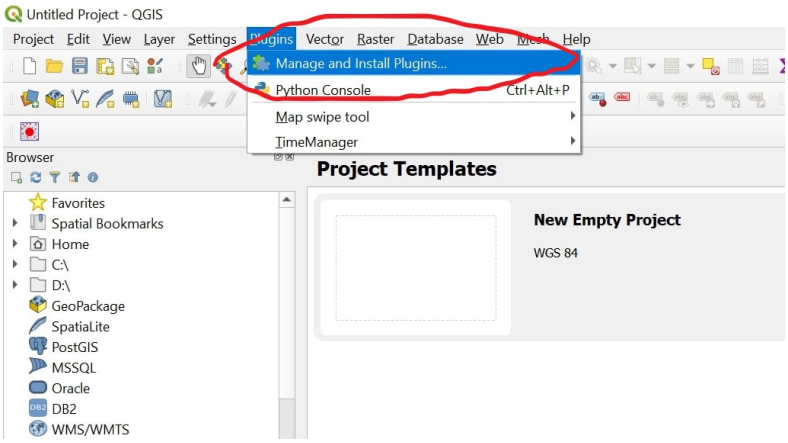


Figure 6.3: Step 3

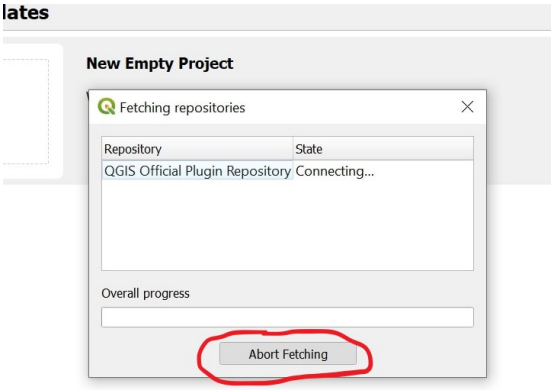


Figure 6.4: Step 4

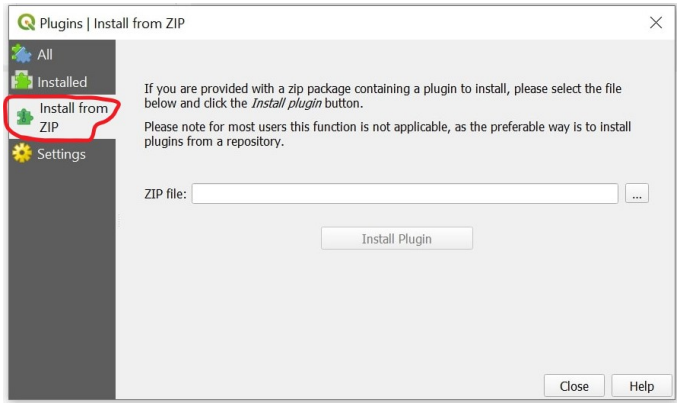


Figure 6.5: Step 5

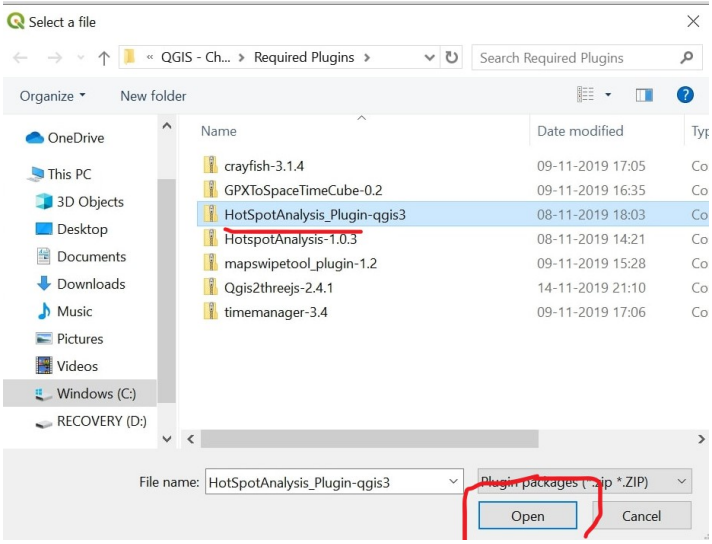


Figure 6.6: Step 6 & 7

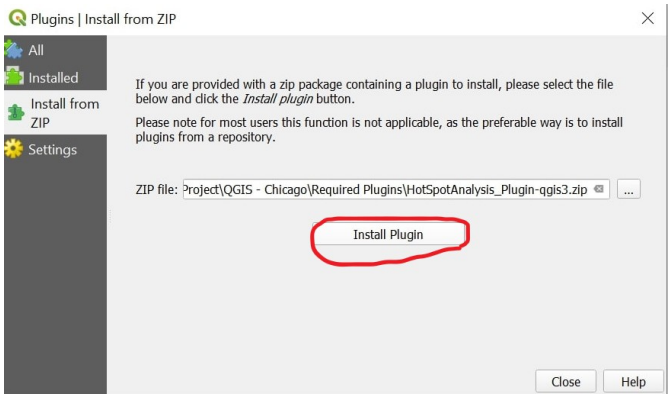


Figure 6.7: Step 8

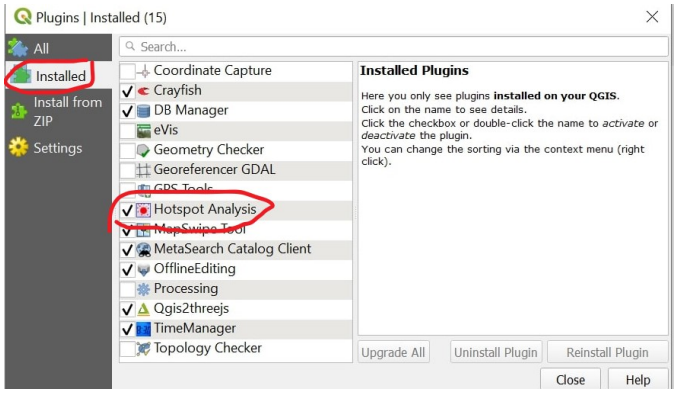


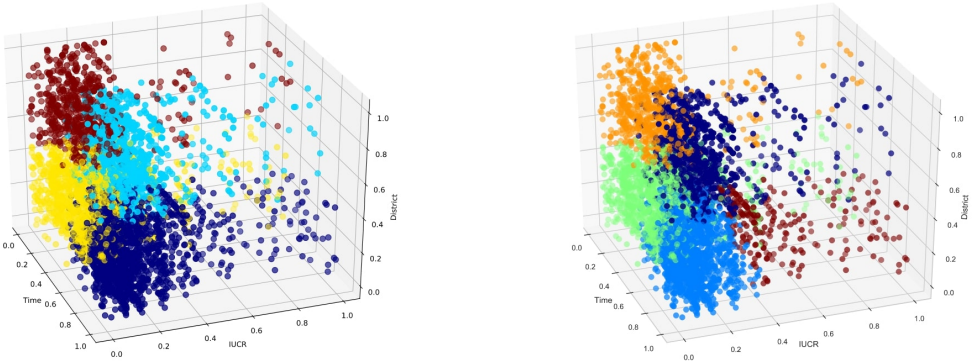
Figure 6.8: Step 9

Chapter 7

ANALYSIS AND RESULTS

7.1 K Means Clustering

Figure 7.1 shows the clustering results obtained using K Means Clustering.



(a) Clustering with K=4 **(b)** Clustering with K=5

Figure 7.1: Clustering Results

Figure 7.2 depicts the cluster allotment. Here, "Normalized_time" denotes the value of time between 0 and 1. Lower values in this column would indicate midnight to early morning, medium

	Block	IUCR	Primary Type	District	Latitude	Longitude	date	time	Normalized_time	Clusters
0	086XX S BISHOP ST	560	ASSAULT	6	41.736989	-87.659450	2019-09-17	23:55:00	0.997220	2
1	025XX S KEDZIE AVE	470	PUBLIC PEACE VIOLATION	10	41.845342	-87.705064	2019-09-17	23:53:00	0.995830	2
2	062XX S INDIANA AVE	810	THEFT	3	41.781514	-87.620614	2019-09-17	23:50:00	0.993746	2
3	082XX S MARSHFIELD AVE	486	BATTERY	6	41.744069	-87.664495	2019-09-17	23:50:00	0.993746	2
4	077XX W HORTENSE AVE	810	THEFT	16	NaN	NaN	2019-09-17	23:45:00	0.990271	1
...	...	...	...	...	...	...	...	...	...	...
4664	068XX W HIGGINS AVE	1750	OFFENSE INVOLVING CHILDREN	16	41.979458	-87.799274	2019-06-04	18:00:00	0.750521	1
4665	014XX N MAYFIELD AVE	560	ASSAULT	25	41.906611	-87.773086	2019-06-04	18:00:00	0.750521	1
4666	082XX S SOUTH SHORE DR	820	THEFT	4	41.745976	-87.547931	2019-06-04	18:00:00	0.750521	2
4667	005XX E 80TH ST	820	THEFT	6	41.749358	-87.610797	2019-06-04	18:00:00	0.750521	2
4668	053XX W CONGRESS PKWY	560	ASSAULT	15	41.873905	-87.758057	2019-06-04	18:00:00	0.750521	1

4669 rows × 10 columns

Figure 7.2: Cluster Allotment

values would indicate the afternoon sessions, and high values would indicate the evening and night time.

7.2 Apriori

For a minimum support of 0.003 and minimum confidence of 0.005, Apriori generated only two rules (refer figure 7.3), whereas for a minimum support of 0.002 and the same minimum confidence, Apriori generated 70 rules (refer figure 7.4). The antecedent (LHS) of the rules generated using both, Apriori and FP Growth, have 3 or more of the following attributes: ["Block", "Location Description", "District", "Timeslot", "Month"], whereas the consequent (RHS) is always "Primary Type".

```
[1] # Applying Apriori in Chicago Dataset...

    =====================================
    2019-11-24 22:47:12 - Start Program
    =====================================

    Time elapsed for generating rules using Apriori :
    -2.4742727279663086
    ('APARTMENT', 'Grand Crossing District', 'Midnight') -> ('BATTERY',)
    Support is : 0.0034891835310537334
    Confidence is : 0.7142857142857143
    Lift is : 3.735662148070907
    Conviction is : 2.830774588836183
    ########################################################
    ('APARTMENT', 'Early Morning', 'February') -> ('BATTERY',)
    Support is : 0.0034891835310537334
    Confidence is : 0.8333333333333334
    Lift is : 4.3582725060827245
    Conviction is : 4.852756425872995
    ########################################################
```

Figure 7.3: Apriori results with minimum support = 0.003

7.3 FP Growth

In contrast to Apriori, FP Growth generated a large number even for relatively higher values of minimum support. However, to compare the performance of these two algorithms, the minimum support and confidence were kept the same. For the minimum confidence value of 0.005,

```
#####################################################
('August', 'Late Afternoon', 'SMALL RETAIL STORE') -> ('THEFT',)
Support is : 0.00209351011863224
Confidence is : 0.6
Lift is : 2.456571428571429
Conviction is : 1.8893928773421143
#####################################################
('Central District', 'DEPARTMENT STORE', 'N STATE ST') -> ('THEFT',)
Support is : 0.00209351011863224
Confidence is : 0.6
Lift is : 2.456571428571429
Conviction is : 1.8893928773421143
#####################################################
('Central District', 'February', 'Late Afternoon') -> ('THEFT',)
Support is : 0.00209351011863224
Confidence is : 0.6
Lift is : 2.456571428571429
Conviction is : 1.8893928773421143
#####################################################
('Central District', 'Late Afternoon', 'March') -> ('THEFT',)
Support is : 0.00209351011863224
Confidence is : 0.6
Lift is : 2.456571428571429
Conviction is : 1.8893928773421143
#####################################################
```

Figure 7.4: Apriori results with minimum support = 0.002

a total of 450 rules were generated when the value of minimum support was kept 0.003 (refer figure 7.5).

7.4 Hotspot Analysis

An additional plugin named "Hotspot Analysis" has to be downloaded using the steps shown in figures 6.1 to 6.8 to perform hotspot analysis in QGIS. Furthermore, additional python packages, as mentioned in Chapter 6 (Section 6.2.2 - Python Packages), need to be installed.

Hotspot analysis is a spatial analysis and mapping method interested in the identification of clustering of spatial events. These spatial events are depicted as points in a map and refer to locations of events or objects. In our work, hotspot analysis has been performed to identify locations with high occurrences of crime. Hotspot analysis has been carried out in our project to identify areas of high crime incidents. Hotspot analysis was carried out on the basis of both count (figure 7.6) and concentration (figure 7.7) of crime (crime count/area of the location). By carefully looking at the figures 7.6 and 7.7, it can be observed that certain census tracts were classified as areas of low crime counts in figure 7.6, even though they were classified as areas of high crime density. This is because, although these areas have less crime count compared to other census tracts, they have high crime count to area ratio due to their small areas.

```
[1] # Applying Frequent Pattern Growth (FP Growth) Algorithm on Chicago dataset...

RESIDENCE ,
-> BATTERY
Support is : 0.14584787159804605
Confidence is : 0.75
Lift is : 5.142344497607656
Conviction is : 3.416608513607816
########################################################
Harrison District ,
RESIDENCE ,
Night ,
-> BATTERY
Support is : 0.1297976273551989
Confidence is : 0.6666666666666666
Lift is : 5.136200716845877
Conviction is : 2.610607117934403
########################################################
February ,
Harrison District ,
RESIDENCE ,
-> BATTERY
Support is : 0.19120725750174458
Confidence is : 1.0
Lift is : 5.2299270072992705
Conviction is : inf
########################################################
```

Figure 7.5: FP Growth results with minimum support = 0.003

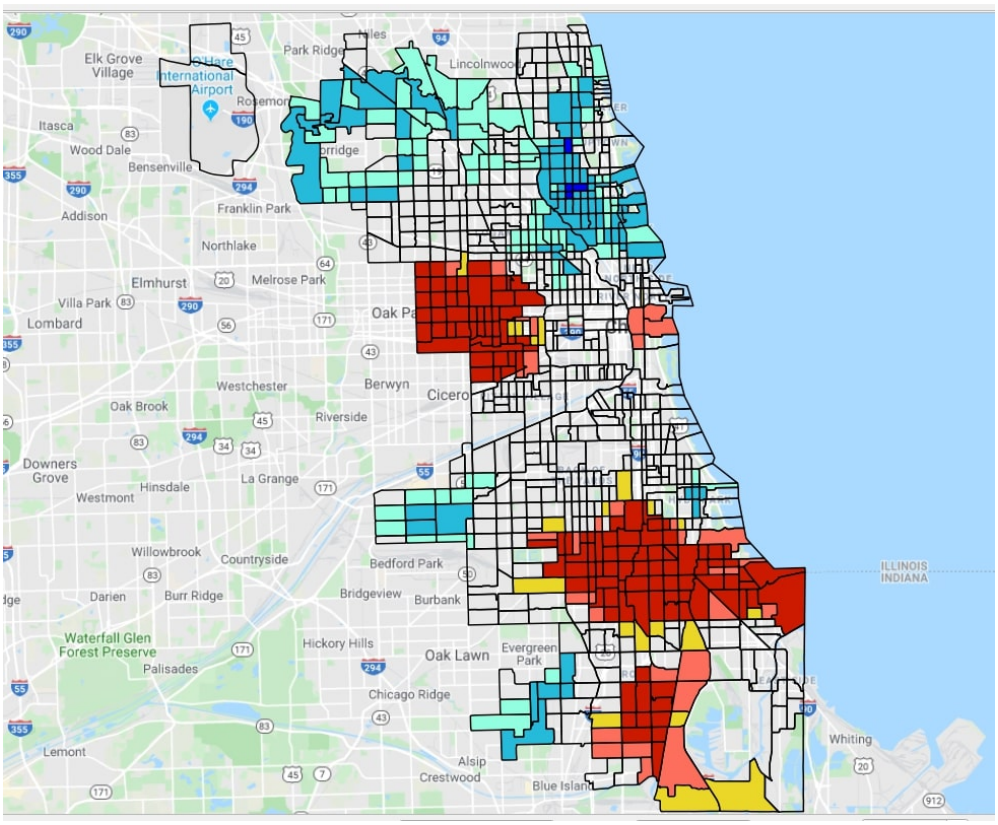


Figure 7.6: Hotspot Analysis on the basis of count

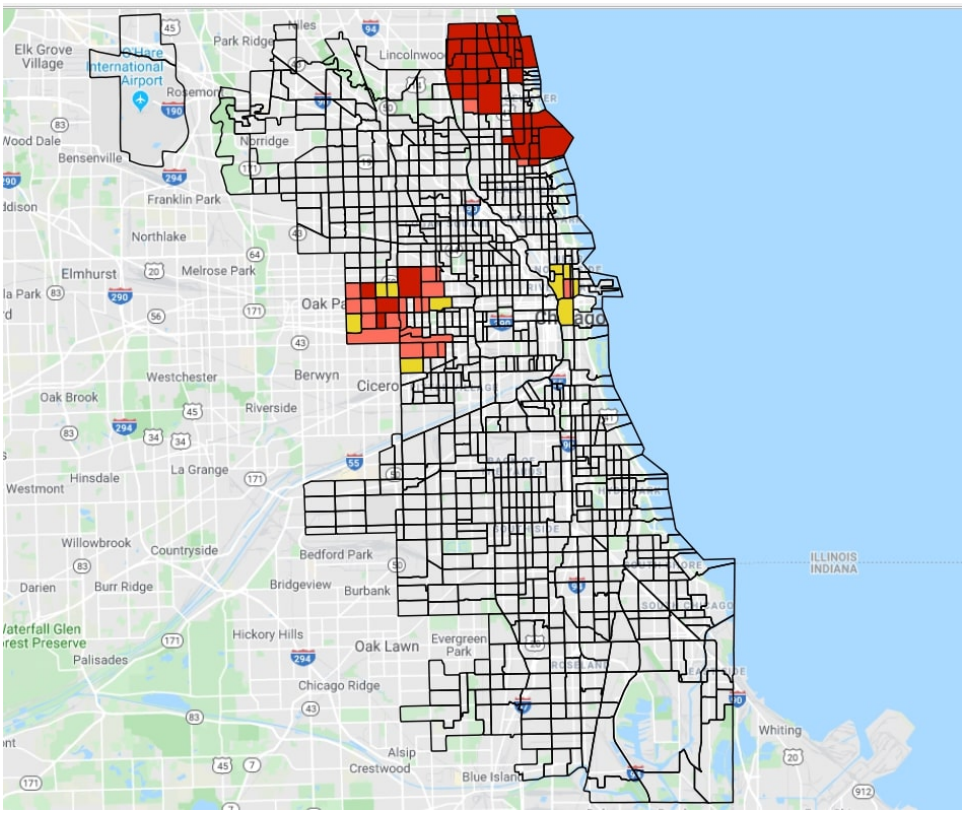


Figure 7.7: Hotspot Analysis on the basis of density

7.5 Cluster & Outlier Analysis

Cluster & Outlier Analysis is performed to identify spatial clusters of features with high or low values as well as spatial outliers. To do this, a Cluster & Outlier tool that calculates a local Moran's I value, a z-score, a pseudo p-value is used. [5] The cluster/outlier type (COType) field differentiates between a statistically significant cluster of low values (LL), cluster of high values (HH), outlier in which a low value is surrounded mostly by high values (LH), and an outlier in which a high value is surrounded mostly by low values (HL).

7.5.1 Interpretation of local Moran's I value

A positive I value indicates that a feature has neighbouring features with similar values (either high or low), i.e., this feature is part of a cluster. A negative I value indicates that a feature has neighbouring features with different values, i.e., this feature is an outlier.

The Local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{X})$$

where x_i is an attribute for feature i, $\bar{X}$ is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{X})^2}{n - 1}$$

with n equating to the total number of features.

The z_{I_i}-score for the statistics are computed as:

$$z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{\mathrm{V}[I_i]}}$$

where:

$$\begin{aligned} \mathrm{E}[I_i] &= -\frac{\sum_{j=1, j \neq i}^{n} w_{ij}}{n - 1} \\ \mathrm{V}[I_i] &= \mathrm{E}[I_i^2] - \mathrm{E}[I_i]^2 \end{aligned}$$

Figure 7.8: Calculation of local Moran's I value, p-value and z-score

7.5.2 Interpretation of p-value

The p-value measures the possibility that the observed spatial pattern was created by some random process. A low p-value signifies the unlikeliness of the spatial pattern being generated by some random process.

7.5.3 Interpretation of z-score

- A high positive z-score for a feature signifies that the surrounding features have similar values (either high or low). The COType field in the Output Feature Class will be LL for a statistically significant cluster of low values and HH for a statistically significant cluster of high values respectively. [6]
- A low negative z-score for a feature signifies a statistically significant spatial data outlier.

Depending on the values of the surrounding features, The COType field in the Output Feature Class will either be LH (low feature value surrounded by features with high values) or HL (high feature value surrounded by features with low values).

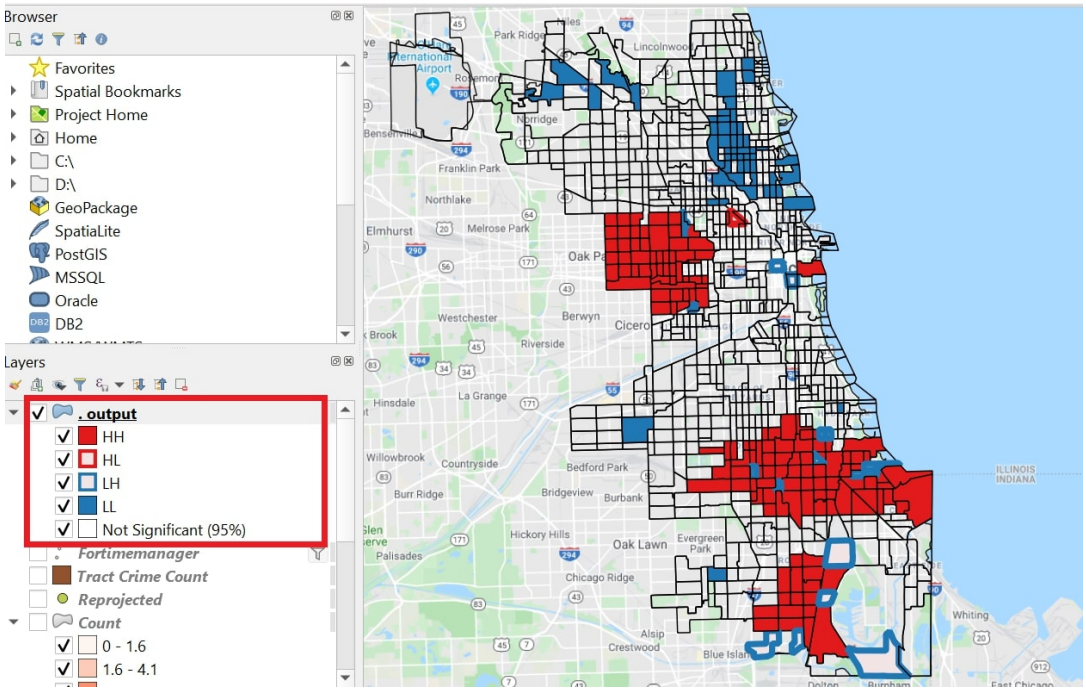


Figure 7.9: Cluster & Outlier Analysis on Chicago Dataset

7.6 Time Series Analysis

Time series analysis comprises techniques for analysing time series data in order to extract meaningful statistics and other characteristics of the data. In our project, we have performed a time series analysis using QGIS. Instead of forecasting crimes for the next time-period, we have limited our project to just the analysis of previous trends.

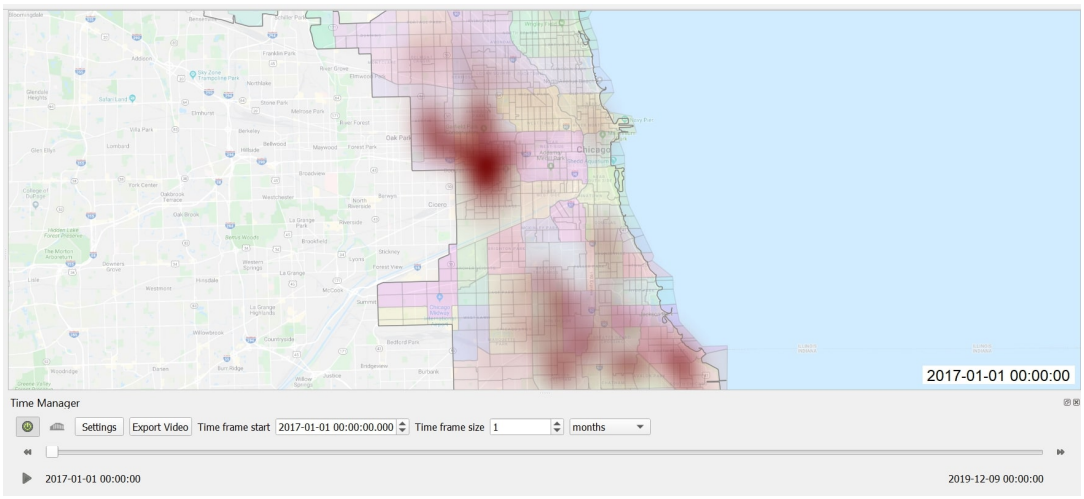


Figure 7.10: Heatmap - January 2017

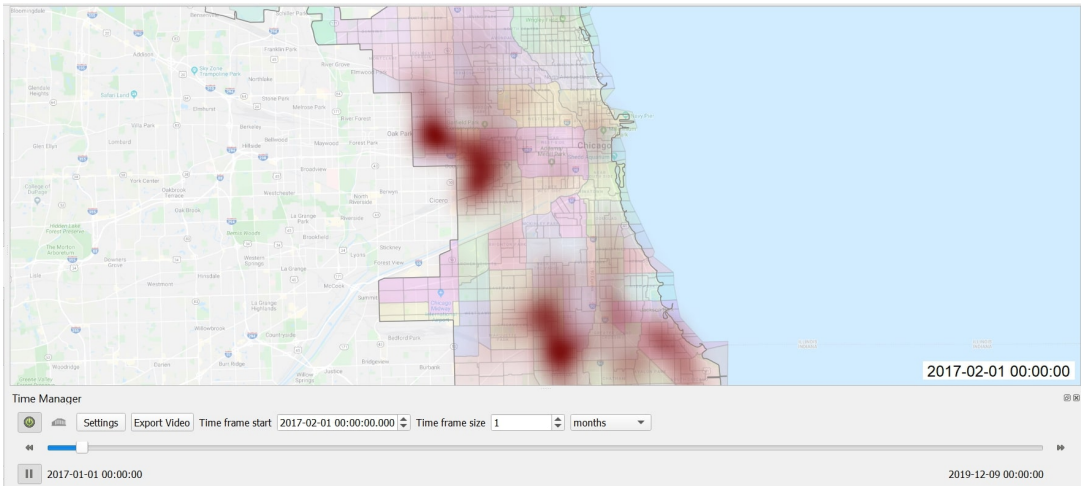


Figure 7.11: Heatmap - February 2017

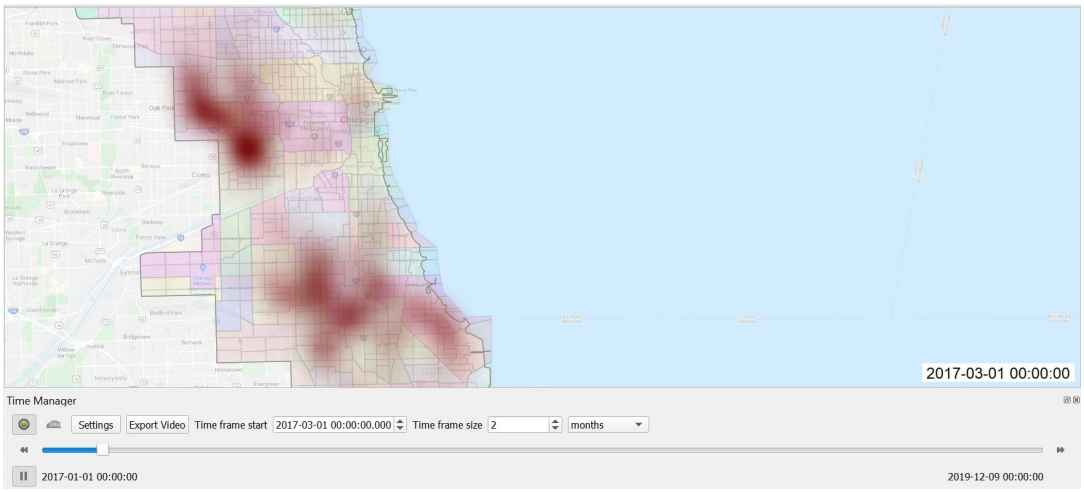


Figure 7.12: Heatmap - March 2017

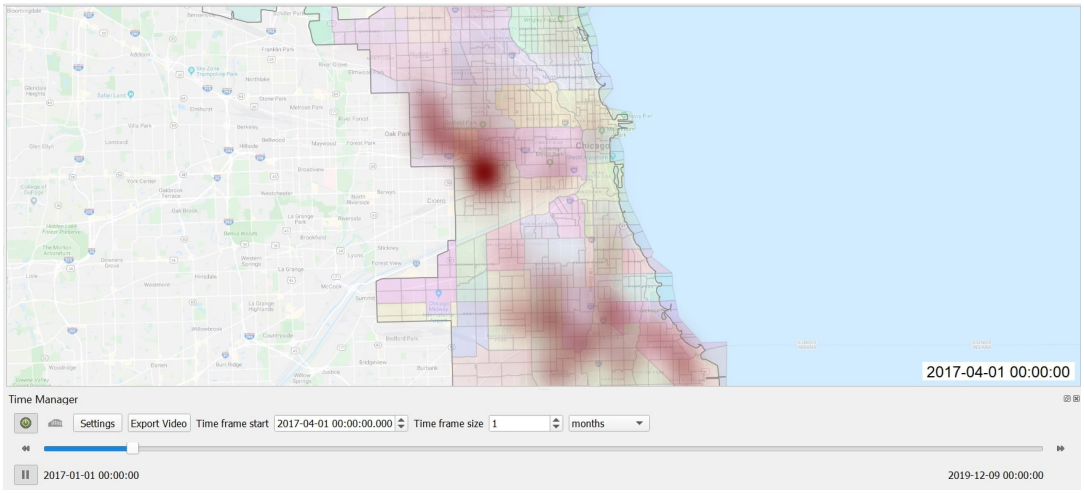


Figure 7.13: Heatmap - April 2017

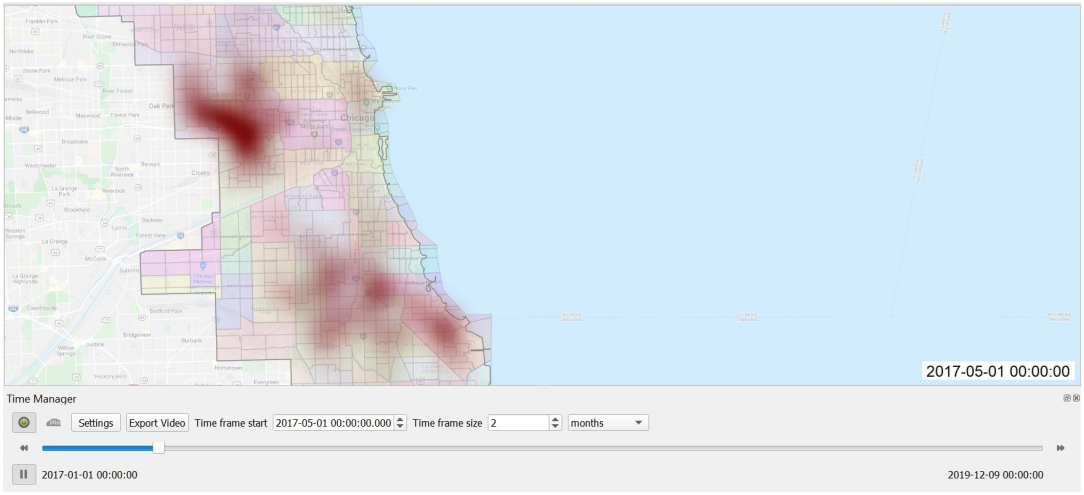


Figure 7.14: Heatmap - May 2017

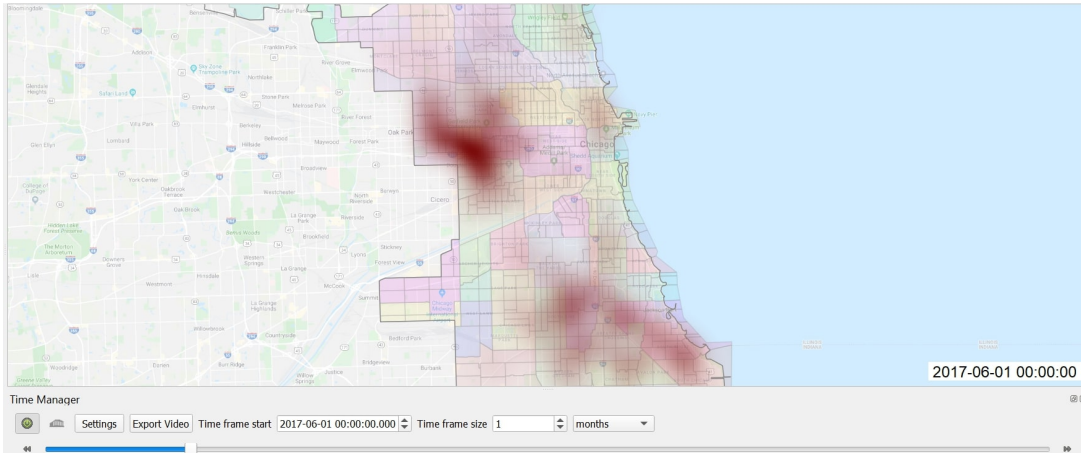


Figure 7.15: Heatmap - June 2017

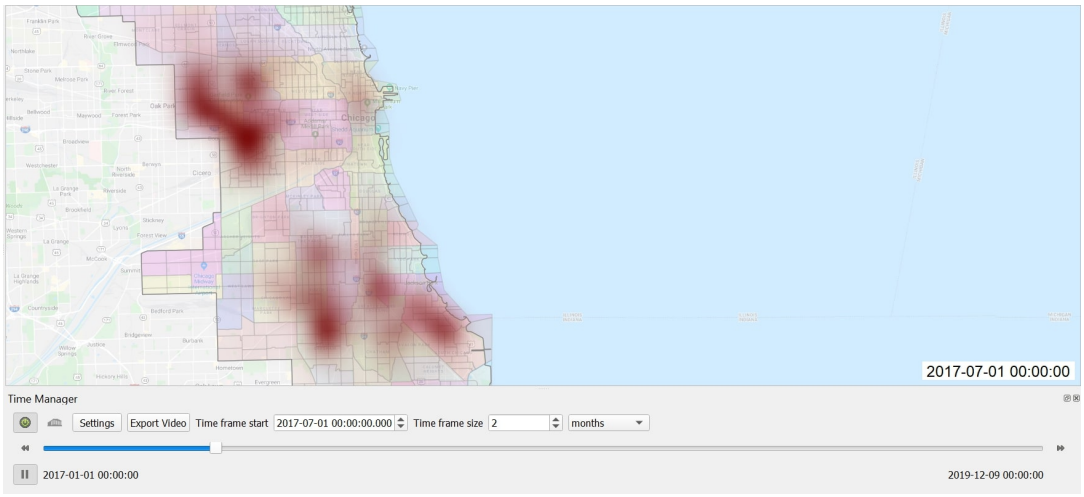


Figure 7.16: Heatmap - July 2017

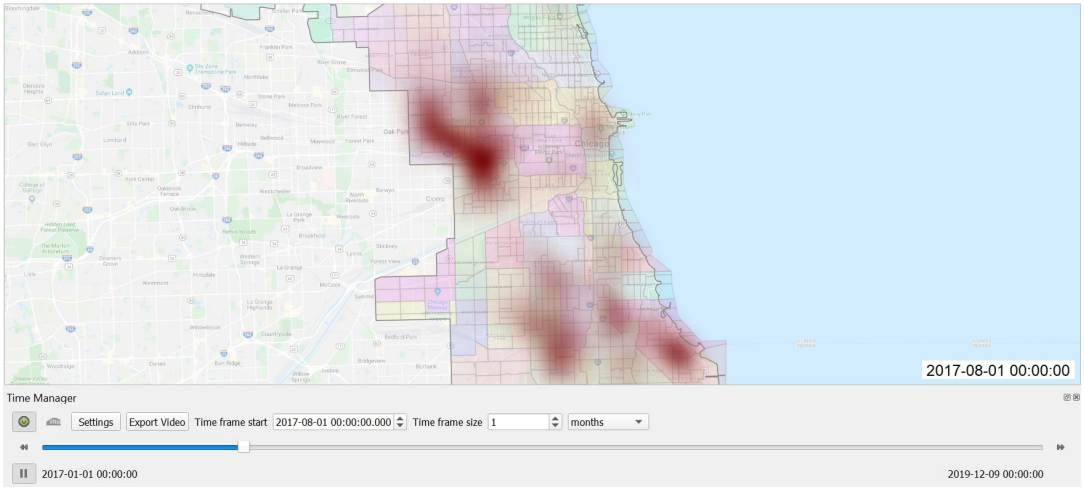


Figure 7.17: Heatmap - August 2017

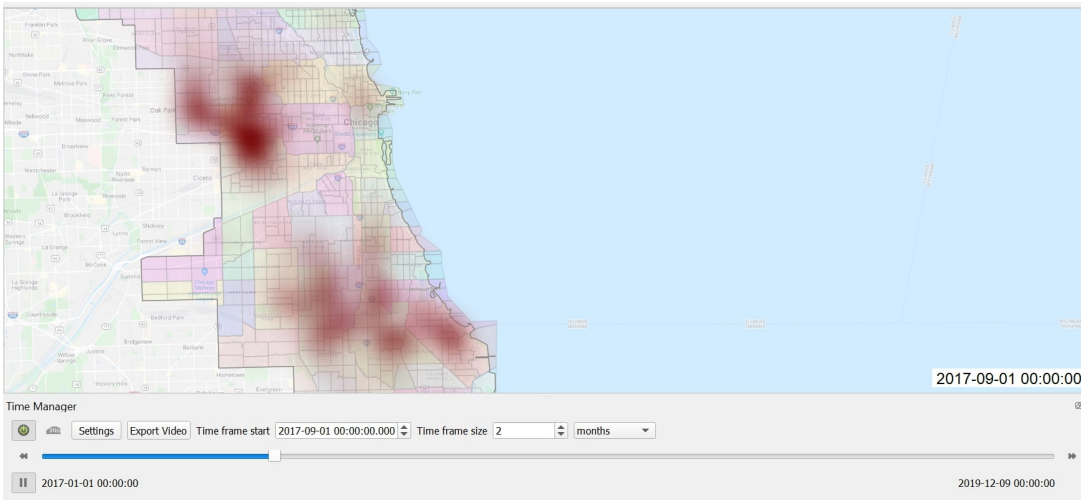


Figure 7.18: Heatmap - September 2017

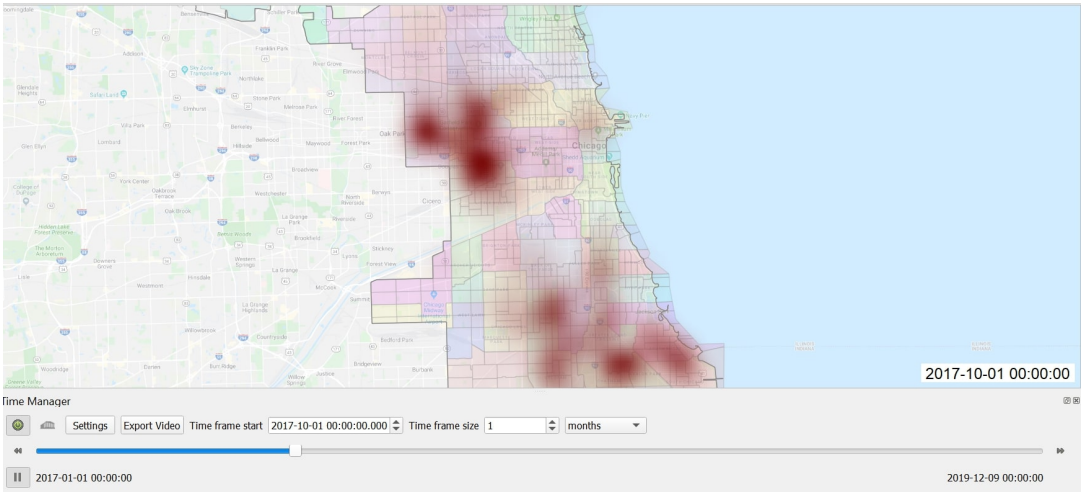


Figure 7.19: Heatmap - October 2017

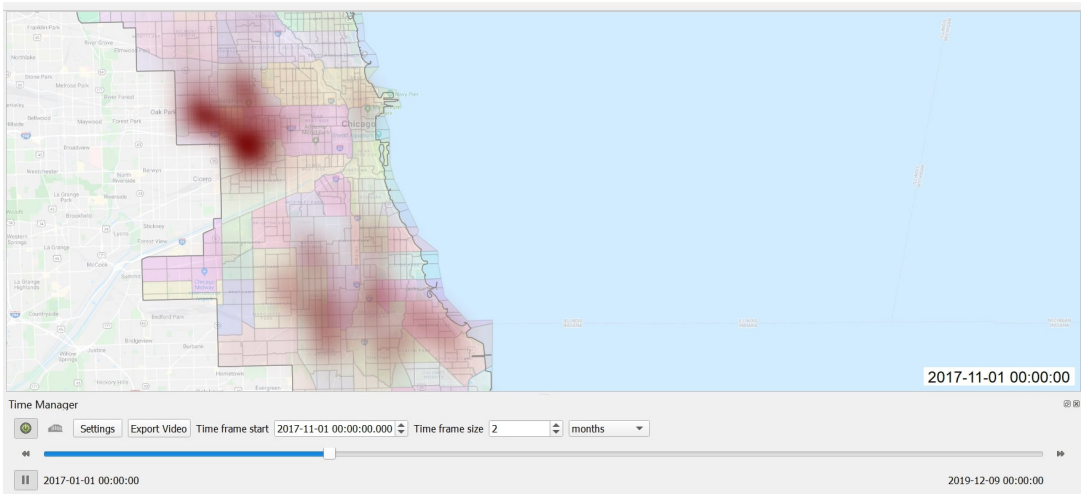


Figure 7.20: Heatmap - November 2017

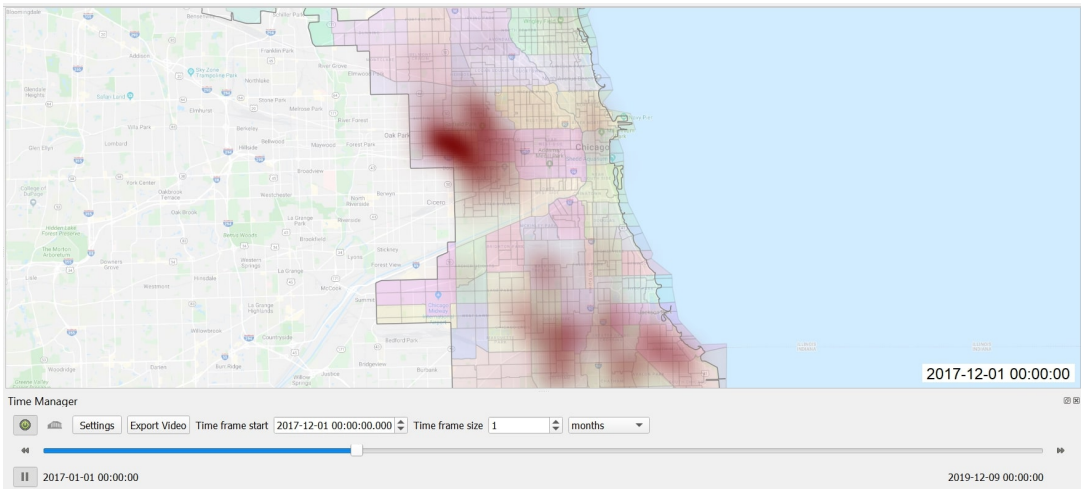


Figure 7.21: Heatmap - December 2017

Chapter 8

EVALUATION METRICS

Evaluation metrics are used to verify a model's performance. It is crucial to choose proper metrics in order to evaluate how well an algorithm is performing. The evaluation metrics used for different models are briefly explained in the following sections:

8.1 Clustering Evaluation Metrics

In comparison to the supervised learning methods, where we have the ground truth to evaluate the model's performance, clustering analysis does not have a robust evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since K-Means requires K as input and does not learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem. Sometimes domain knowledge and intuition may help, but usually, that is not the case [4]. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modelling. In our work, we have used two such metrics that may give us better intuition about K:

- Elbow Method
- Silhouette Analysis

8.1.1 Elbow Method

The elbow method is one of the most common and technically robust methods for determining a good value of K for clustering. It provides us with an idea of what a good number of clusters would be on the basis of the Within Cluster Sum of Squares (WCSS) between data points and their assigned clusters' centroids. For $K = 3$, WCSS can be computed as follows:

$$\text{WCSS} = \sum_{\text{P}_\text{i} \text{ in Cluster 1}} \text{distance}(\text{P}_\text{i}, \text{C}_1)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 2}} \text{distance}(\text{P}_\text{i}, \text{C}_2)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 3}} \text{distance}(\text{P}_\text{i}, \text{C}_3)^2$$

Figure 8.1: Computing WCSS

While clustering performance as measured by WCSS might increase (i.e. WCSS decreases) with an increase in K, the rate of increase usually decreases. So, we choose that that value of K where WCSS starts to flatten out and to form an elbow, i.e., where the rate of increase in performance starts to diminish.

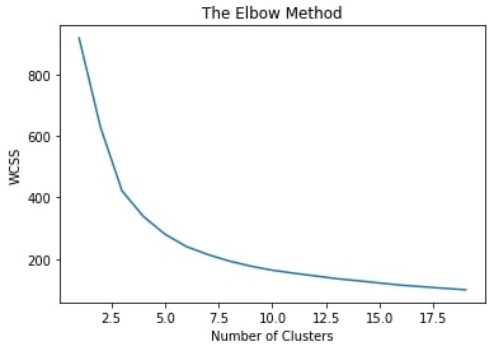


Figure 8.2: The Elbow method

The graph above shows that $K = 4$ and $K = 5$ are excellent choices for the value of K for our dataset. Although the elbow method provides useful insights for choosing the value of K, sometimes it is difficult to figure out a good number of clusters to use by looking at the elbow curve because the curve is monotonically decreasing and may not show a distinct point where the curve starts flattening out. In such cases, silhouette analysis can be of significant help.

8.1.2 Silhouette Analysis

Silhouette analysis is employed to interpret and validate the consistency within clusters of data [**?**]. The silhouette value is a measure of how similar an entity is to its cluster (cohesion) compared to other clusters (separation). It can be used to determine the degree of separation between clusters. For each sample, the steps involved in silhouette analysis are as follows:

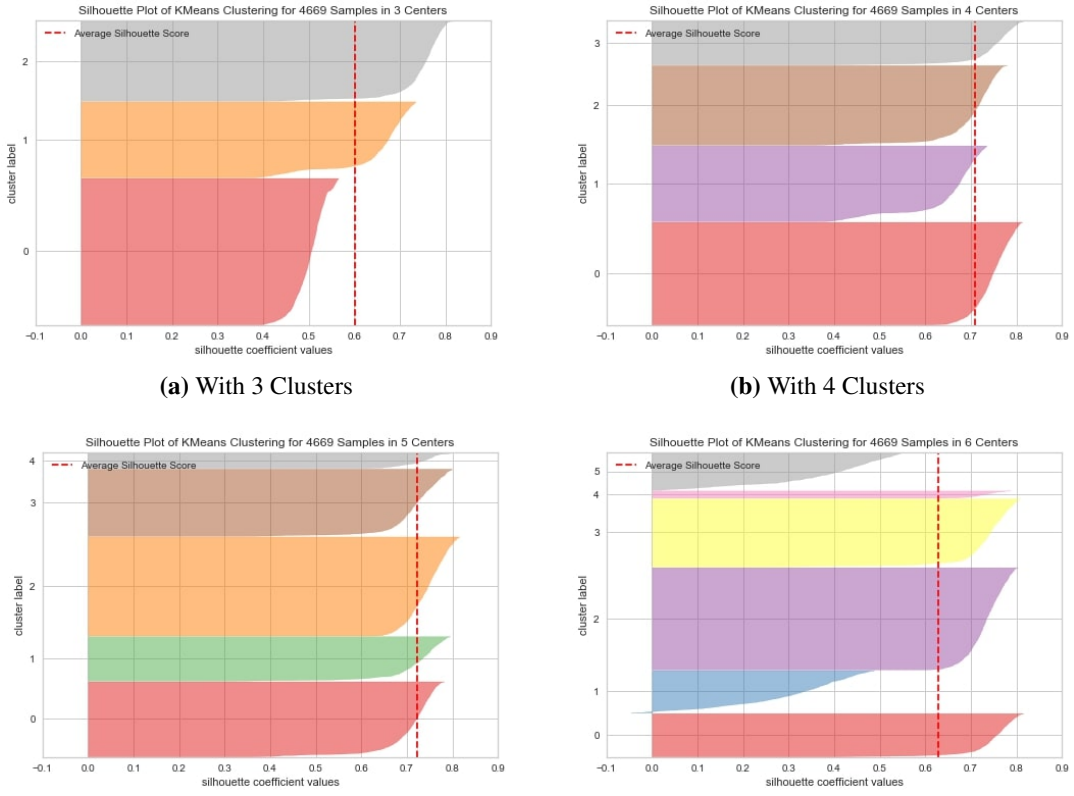
1. The mean distance from all data points in the same cluster (a^i) is calculated.
2. The mean distance from all data points in the closest cluster (b^i) is computed.

3. The silhouette coefficient is computed using the equation given below:

$$\textbf{Coefficient} = \frac{b^i - a^i}{\max\left(a^i, b^i\right)} \tag{8.1}$$

The coefficient can take values in the interval [-1, 1]:

- If it is 0 –> the sample is very close to the neighbouring clusters.
- If it is 1 –> the sample is far away from the neighbouring clusters.
- If it is -1 –> the sample is assigned to the wrong clusters.



(a) With 3 Clusters

(b) With 4 Clusters

(c) With 5 Clusters

(d) With 6 Clusters

Figure 8.3: Silhouette Analysis

For a good value for the number of clusters, the following conditions should be satisfied:

1. The mean value of silhouette coefficients should be as close to 1 as possible.
2. The silhouette plot of each cluster should be above the mean silhouette value as much as possible. Any plot region below the mean value is not desirable.
3. The width of the plot should be as uniform as possible.

From the fig. 8.3, it can be established that $K = 4$ and $K = 5$ are the optimal values for K. For $K = 3$ and $K = 6$, not only is the mean silhouette value low, the silhouette plot of many individual clusters is below that the average silhouette line.

For our dataset, Silhouette analysis has further corroborated the results obtained from the elbow method. This demonstrates the effectiveness of our implementation of K Means clustering algorithm.

8.2 Association Rule Mining Evaluation Metrics

Like clustering, association rule mining algorithms lack properly defined evaluation metrics. However, support, confidence, lift, and conviction of the rules, as established in eqns. 4.1, 4.2, 4.3 and 4.4 respectively, can offer insights pertaining to the correctness of the rules. The evaluation metrics for association rule mining algorithms can broadly be classified into two categories:

- **Symmetric Measures:** Support, Lift/Interest etc.
- **Asymmetric Measures:** Confidence, Conviction etc.

8.2.1 Comparison of FP Growth and Apriori

Since FP Growth and Apriori have been used for the same purpose, i.e., extracting rules concerning crime characteristics from our dataset, it becomes imperative to compare their performance. We have evaluated and compared the performance of these two algorithms on the basis of two parameters:

1. The values of lift and conviction (for the same values of minimum support and confidence).
2. Time and space complexity.

Lift and Conviction

Conviction of a rule measures by what factor the correctness of the rule would reduce if the antecedent and the consequent of the rule were independent. For example, if the conviction

value of a rule is 1.2, then, the rule would be incorrect 20% more often if the association between antecedent and consequent was purely by chance. Some interesting points to note:

- If items are independent: lift = conviction = 1
- Higher is the value of confidence, higher is the value of conviction. If confidence of a rule is equal to 1, then conviction of that rule would be infinity.

From the rules generated using Apriori and FP Growth, the average conviction of the rules obtained using FP Growth was found to be higher than the average conviction values of the rules obtained using Apriori.

Time & Space Complexity

FP Growth makes only two passes over the dataset as compared to Apriori, which scans the dataset multiple times. Reduced number of scans make FP Growth a more time-efficient algorithm. Furthermore, FP Growth uses a tree data structure for storing the generated rules, which subsequently make it more space-efficient as well.

Table 8.1: Comparison of Apriori and FP Growth

	Apriori	FP Growth
Time Complexity	O(2^d)	O(n+d)
Space Complexity	O(2^d)	O(nd)

Table 8.1 compares the time and space complexity of the two algorithms [13]. Here, n is the number of transactions, i.e., rows in our dataset, while d denotes the number of unique items in our dataset.

To determine the time taken by both the algorithms, namely, FP Growth and Apriori for rule generation, a timing module was used. For the same dataset, it was observed that FP Growth was significantly faster than Apriori in generating rules. Apriori took around 6.28 seconds to generate rules (refer figure 8.4), whereas, FP Growth took only 0.25 seconds (refer figure 8.5). This discrepancy in the time taken by these algorithms would be more substantial for larger datasets.

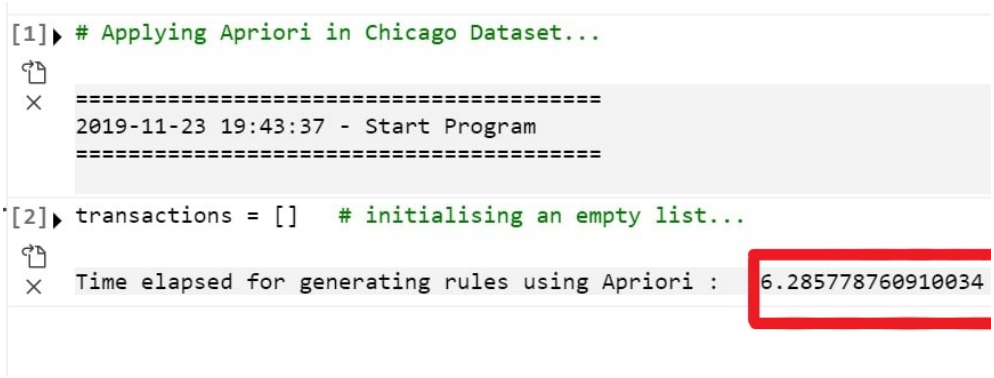


Figure 8.4: Time Taken using Apriori

Figure 8.5: Time Taken using FP Growth

Chapter 9

FUTURE DIRECTION AND CONCLUSION

9.1 Future Work

Taking a close look at the problems discussed, we can say that our model is pretty much expandable to other domains as well. However, there exist certain limitations in the proposed model. One limitation of our model is its sensitive nature to the quality of input data that may be inaccurate or have missing information. Another limitation of the current work is the difficulty of evaluating the accurate performance of the proposed model, so future works could be aimed at finding a way to use quantified methods to evaluate the degree of crime reduction achieved, the improvement in duty deployment, and the impact on society. Moreover, owing to computational constraints, algorithms have only been applied to data for the current year. If the period under consideration could be increased, the results produced would be more reliable.

As a future extension of our work, supervised learning models such as classification models can be used to predict the type of crime. It is also a useful extension for our study to consider neighbourhood income information in order to see if there is a relationship between the income level of neighbourhoods and their crime rate. Additionally, the ability to search suspect description in regional, FBI databases, traffic violation databases from various states to help detect crime patterns will also add value to this crime detection paradigm.

9.2 Conclusion

The society we live in is a complicated and culturally revolutionised one, where crime problems are rising in an endless stream, and their prevention has become of utmost importance for the police and the government. In this project, we have applied data mining strategies, in particular, association rule mining strategies towards public security index requirement to support decision making for police departments and authorities.

Initially, an EDA was performed to demonstrate the baseline for understanding the Chicago crime dataset. Our EDA found many exciting results and statistics that prompted us to apply rule mining algorithms on our dataset to uncover crime trends from the Chicago crime dataset. Before applying association rule mining, K Means clustering was carried out to understand the relationship between the three most important characteristics of the crime dataset, namely - location, time and crime type. Proper understanding of the graphs generated by clustering further warranted the appropriateness of association rule mining for this dataset. The most widely used association rule mining algorithm, Apriori, was first applied. Although the rules obtained using Apriori were reliable and had satisfactory values of lift/ interest and conviction, the application of Apriori was both time and computation intensive. To this end, a better association rule mining algorithm, FP Growth, was applied, which not only produced better results (rules with better conviction and lift) but also proved to be more time-efficient.

Finally, a GIS application, QGIS, was used to analyse the changing locations of hotspots with each week, month and year. Furthermore, a cluster& outlier analysis was conducted to identify low crime rate locations in high crime count areas to provide the police department with further insight into the reasons behind the crimes. In addition, a crime hotspot analysis was done to help in the deployment of police at most likely places of crime for any given window of time, to allow the most effective utilisation of police resources. Various results and graphs presented in this report corroborate the effectiveness of our model in identifying crime patterns.

Bibliography

[1] Almanie, T., Mirza, R., and Lor, E. (2015). "Crime prediction based on crime types and using spatial and temporal criminal hotspots." *arXiv preprint arXiv:1508.02050*.

[2] Brown, D. E. and Hagen, S. (2003). "Data association methods with applications to law enforcement." *Decision Support Systems*, 34(4), 369–378.

[3] Chandrasekar, A., Raj, A. S., and Kumar, P. (2015). "Crime prediction and classification in san francisco city." *URL http://cs229. stanford. edu/proj2015/228 {_} report. pdf*.

[4] Dabbura, I. "K-means clustering: Algorithm, applications, evaluation methods, and drawbacks [online]. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Accessed On: Nov. 21 ,2019.

[5] esri. "How cluster and outlier analysis (anselin local moran's i) works [online]. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm. Accessed On: Nov. 24 ,2019.

[6] esri. "What is a z-score? what is a p-value? [online]. http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/what-is-a-z-score-what-is-a-p-value.htm.

[7] Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., and Khanahmadliravi, N. (2013). "An experimental study of classification algorithms for crime prediction." *Indian Journal of Science and Technology*, 6(3), 4219–4225.

[8] Jain, V., Sharma, Y., Bhatia, A., and Arora, V. (2017). "Crime prediction using k-means algorithm." *Global Research and Development journal for engineering Volume2, issue5*.

[9] Jangra, M. and Kalsi, M. S. (2019). "Crime analysis for multistate network using naive bayes classifier.

[10] Keyvanpour, M. R., Javideh, M., and Ebrahimi, M. R. (2011). "Detecting and investigating crime by means of data mining: a general crime matching framework." *Procedia Computer Science*, 3, 872–880.

[11] Kianmehr, K. and Alhajj, R. (2008). "Effectiveness of support vector machine for crime hot-spots prediction." *Applied Artificial Intelligence*, 22(5), 433–458.

[12] Kim, S., Joshi, P., Kalsi, P. S., and Taheri, P. (2018). "Crime analysis through machine learning." *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE. 415–420.

[13] Kosters, W. A., Pijls, W., and Popova, V. (2003). "Complexity analysis of depth first and fp-growth implementations of apriori." *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, Springer. 284–292.

[14] Li, S.-T., Kuo, S.-C., and Tsai, F.-C. (2010). "An intelligent decision-support model using fsom and rule extraction for crime prevention." *Expert Systems with Applications*, 37(10), 7108–7119.

[15] McClendon, L. and Meghanathan, N. (2015). "Using machine learning algorithms to analyze crime data." *Machine Learning and Applications: An International Journal (MLAIJ)*, 2(1), 1–12.

[16] Nath, S. V. (2006). "Crime pattern detection using data mining." *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, IEEE. 41–44.

[17] Ng, V., Chan, S., Lau, D., and Ying, C. M. (2007). "Incremental mining for temporal association rules for crime pattern discoveries." *Proceedings of the eighteenth conference on Australasian database-Volume 63*, Australian Computer Society, Inc. 123–132.

[18] Vaidya, O., Mitra, S., Kumbhar, R., Chavan, S., and Patil, R. (2018). "Comprehensive comparative analysis of methods for crime rate prediction.

[19] Yu, C.-H., Ward, M. W., Morabito, M., and Ding, W. (2011). "Crime forecasting using data mining techniques." *2011 IEEE 11th international conference on data mining workshops*, IEEE. 779–786.