Crime Rate Analysis

Submitted by

Arushi Gupta

130905372(7 'B')

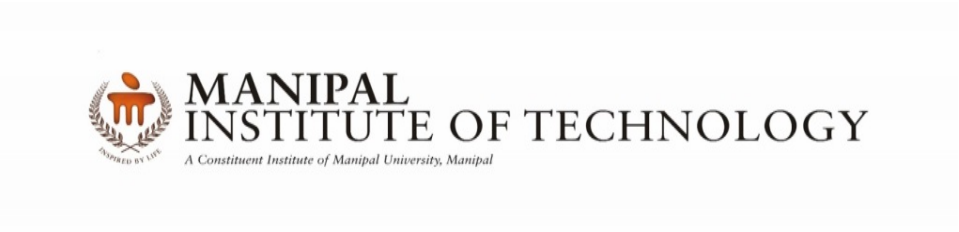
Sanjana Kumar

130905146 (7 'B')

Arnav Pradhan

130905025(7 'A')

Department of Computer Science & Engineering



1. ABSTRACT

Project Goals:

The goal of the project is to create an automated tool (or dashboard) to analyse crime patterns over the past 6 years.

Key deliverables:

Understand patterns and get insight into crimes in the DC metro police system to show trends over a certain period of time.

Time Line of Project:

Data Collection – 0.5 months

Data Cleaning – 0.5 months

Structure and design of dashboard – 0.5 months

Plotting of data – 0.5 months

Proposed Approach:

Initially, two weeks will be spent on the building of the dataset of all of the crimes in the DC metro police system ranging from Theft, Arson, Assault, Homicide, Sex Abuse, Robbery, and Burglary. We will then start cleaning the data set to get a complete and well-structured data set. Data will be mapped and trends will be extracted. Following that, we will design the structure of the dashboard using various widgets such as pie charts, bar graphs, filter headers and scatterplots. The final part will be to complete the plotting of the data.

Alternative Approach:

An alternative approach would be to carry out the dataset collection as the first step (same as previous approach) using an SQL database.

Project Scope:

The project scope includes attributes in the database such as offence, method, year, month, week, hour, shift, start date, end date. The project will be based solely for analysing the trends over the given period.

Project out of scope:

The project shall not be considering every single crime in every region for the given time period.

Transit stations are acknowledged as particularly criminogenic settings. Transit stations may serve as crime "generators," breeding crime because they bring together large volumes of people at particular geographies and times. They may also serve as crime "attractors," providing well-known opportunities for crimes. To reflect the trends in crime, crime counts are stratified into three temporal groups: peak hours, off-peak day hours, and off-peak night hours. The DC metro police system ranging from Theft, Arson, Assault, Homicide, Sex Abuse, Robbery, and Burglary had a large collection of police information reports (PIR) that were filled out by officers at the time of recording incidents over a period of six years. The main portion of each PIR holds a text description of the incident. This description includes when and where the incident took place, at what shift, the type of offence and the method used.

2. MOTIVATION

Building on the success of employing the analysis of structured data to help solve and prevent crimes, Law Enforcement and Government organizations are seeking to expand the scope of their analysis to include unstructured text data. Today, new data and text mining technologies provide a next generation of tools for the analysis and visualization of both structured data and text. Such tools help increase the quality and productivity of the analysis and reduce the latency period between recording raw data and obtaining key knowledge necessary for making informed decisions.

3. OBJECTIVES

The department was seeking a capability to identify historical crime patterns from a large volume of unstructured data. There are many questions investigators needed to get answered quickly. Manual analysis of all PIRs was a cumbersome, time-consuming process prone to errors and biases. New automated text analysis could help the agency quickly and consistently discover important patterns in crime occurrences such as:

Are there correlations between the crime type and the location of the incident?

Are there correlations between the type of crime, weapon employed, and the location of the incident?

What is the most typical weapon?

Are there correlations between the shift and no. of crimes?

BI's Position on crime rate

- Utilizing previous data sets and advanced statistics from past crimes to research and predict the occurrence of future crimes

- Characterize crime and patterns (day, month, area, economic factors, payday, weather, etc.)
- Use software and modeling techniques.
- Quick response and reports to crime.
- Acknowledge where to deploy proper amounts of police force.
- Law enforcements agencies collecting too much data, but no recognizing the information.
- While, large agencies are utilizing BI software techniques and data mining, smaller agencies employ the use of data mining and predictive modeling to assist in preventing crime

Techniques Used

- **Predictive Analytics**
    - Analyzing data of past crimes to forecast when and where crimes are most likely to occur next.
- **Data Mining**
    - Used to identify motives of individuals when shopping, likewise when locating criminals.
    - Discovers hidden patterns and relationships throughout large amounts of information.
- **Tactical Analysis**
    - Creating models that represent a crime or crimes that can be connected to identify cases to locate suspects.
    - Breaks down crime based on day and time and other variables.
- **Behavioral Analysis**
    - Predict future crime based on relationships or behavior of criminals.
    - Using past criminal records to categorize performance.

4. INTRODUCTION

Crime data analysis is fundamental to effective crime prevention. Knowing as much as you can about crime will help significantly in its prevention. It is generally accepted that there are significant limitations to recorded crime data. There are many factors that limit an understanding of the 'true' picture of crime. Many offences will never be reported to the police. Only about one in three assaults, attempted burglaries and robberies of the person are ever reported to the police Even fewer sexual assaults are reported to police, with data from victim surveys suggesting that only 15 to 20 per cent of sexual assaults are reported to police. Consequently, any analysis of crime data should recognise the limitations of the raw data, due to the uneven reporting rates of particular crimes. Determining if particular crimes are increasing; identifying the hot spot locations where crime is concentrated; understanding the temporal trends of offending and analysing potential reasons for crime trends will be critical features of crime data analysis.

The study setting is the Washington DC, Metro. Metro provides service for more than 700,000 customers a day throughout the Washington, DC area. It is the second busiest rail system in the United States, serving 91 stations in District of Columbia, Maryland, and Virginia.

Crime has declined both in the District and the suburbs in recent years. In fact, the influx of more affluent new residents in the city has not led to an uptick in robberies or property crimes. There was an average of 11 robberies each day across the District of Columbia in 2006, which is far below the levels experienced in the 1990s.

5. LITERATURE REVIEW

6. METHODOLOGY

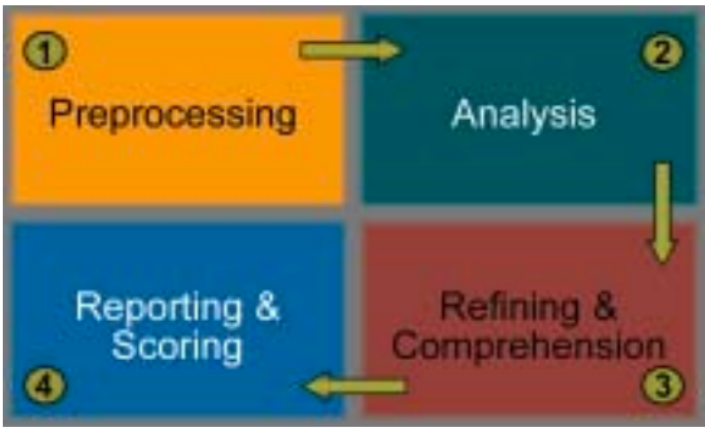


Figure 1

The developed approach consists of a series of steps:

- Preprocess data to the format suitable for further analysis
- Extract important concepts and terms through text-mining
- Analyze patterns and co-occurrences of identified concepts
- Develop an automated solution for crime pattern analysis

7. RESULTS

Data Preprocessing

The first step in creating an analytical solution involves understanding data and transforming it to a convenient format

Table 1

	year	month	week	hour	day	SHIFT	OFFENSE	METHOD	DISTRICT
0	2011	1	1	6	0	MIDNIGHT	SEX ABUSE	OTHERS	6
1	2011	1	1	17	0	EVENING	SEX ABUSE	OTHERS	5
2	2011	1	1	19	0	EVENING	SEX ABUSE	GUN	6
3	2011	1	2	0	1	MIDNIGHT	HOMICIDE	GUN	7
4	2011	1	2	0	1	MIDNIGHT	HOMICIDE	GUN	4

Concept Extraction

The first type of text analysis starts with capturing key concepts and terms in text descriptions present with the help of a text mining engine(OLAP cubes). This engine can run in an unsupervised mode, when clusters of unusually frequently occurring terms are automatically discovered by the system, or a supervised mode when the user focuses the analysis performed by a text mining engine to only primary topics of interest to the user.

Put the olap cube pivot diagrams sanjana sent you. Put two tables

Pattern Analysis

In the next step of the analysis, all extracted terms were used for tagging individual reports, allowing further usage of these terms as new structured attributes together with the original structured data.

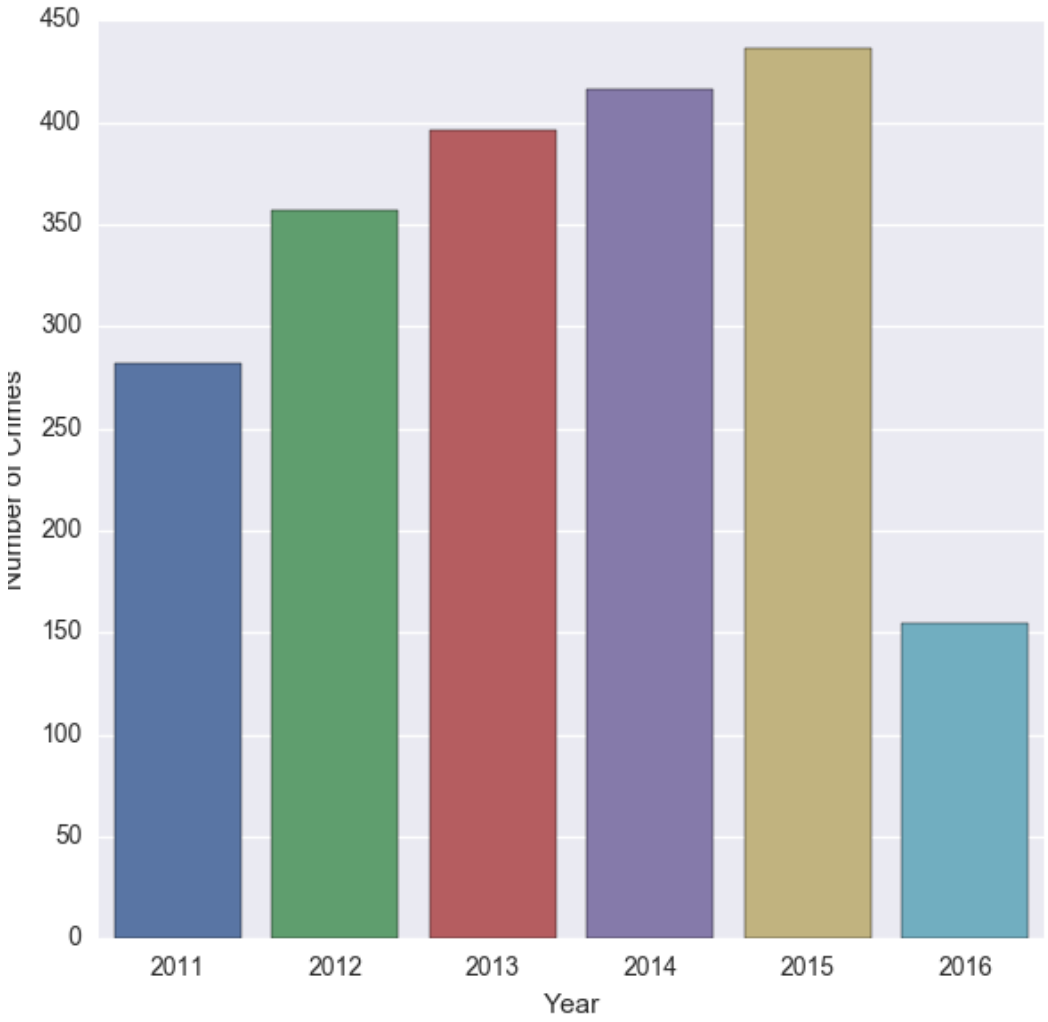


Figure 2

KPI 1: No. of crimes taking place every year

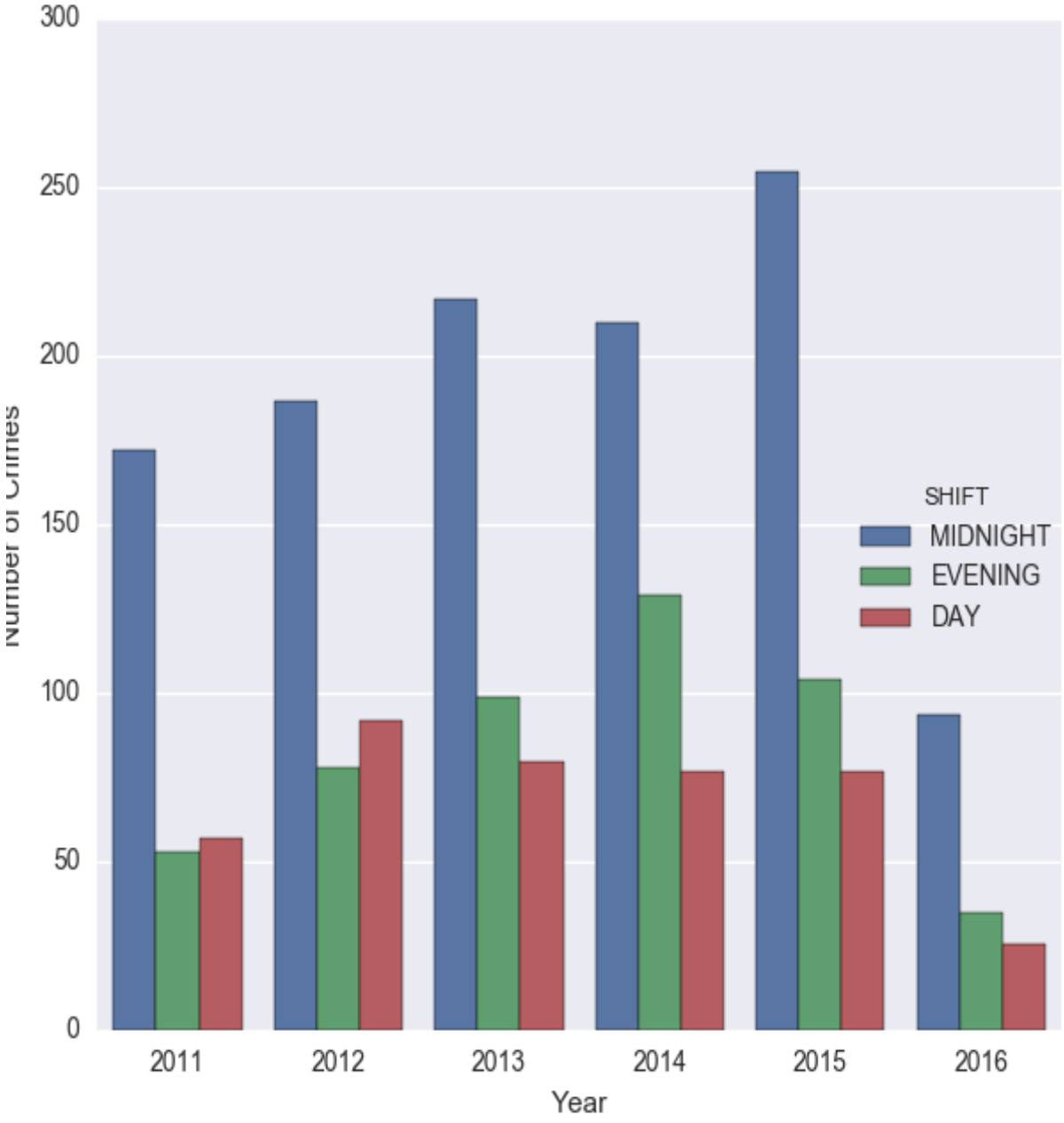


Figure 3

KPI 2: Number of crimes taking place during different shifts of the day

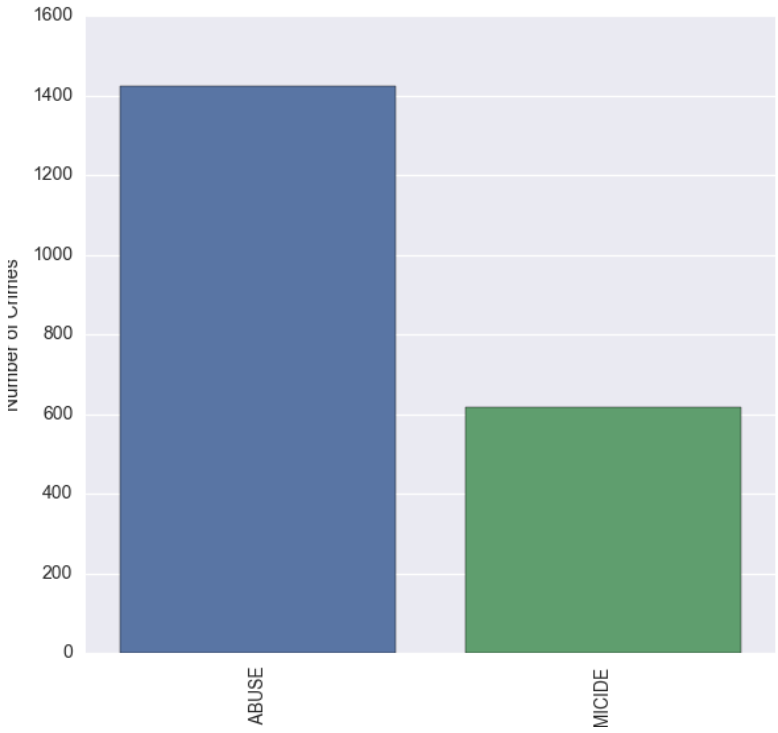


Figure 4

KPI 3: Number of instances of each offense over 6 years

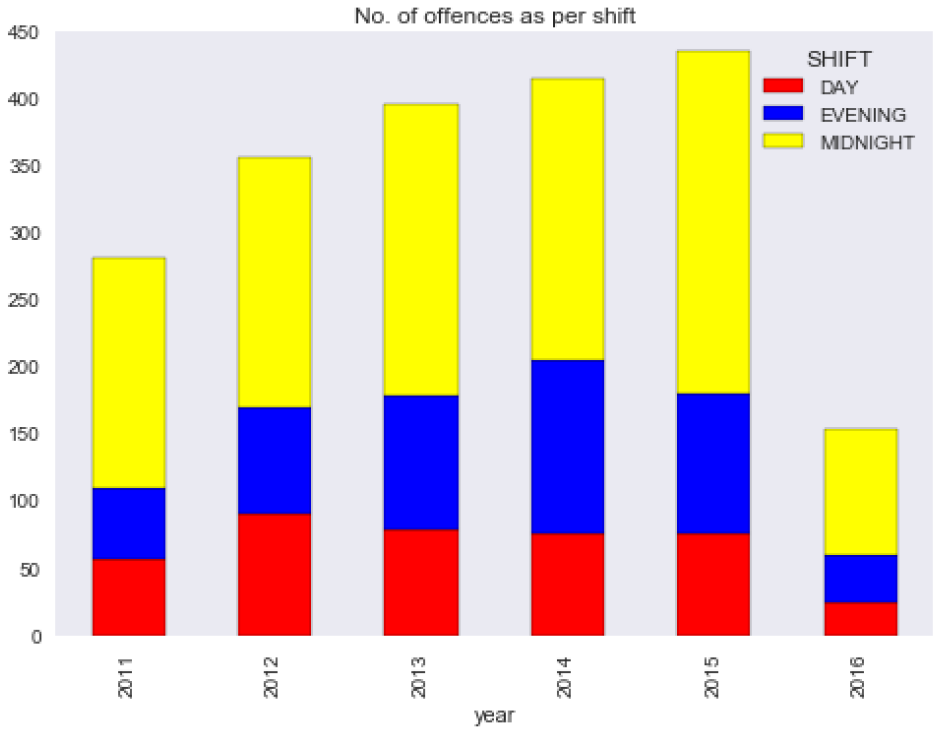


Figure 5

KPI 4: Number of offenses as per shift

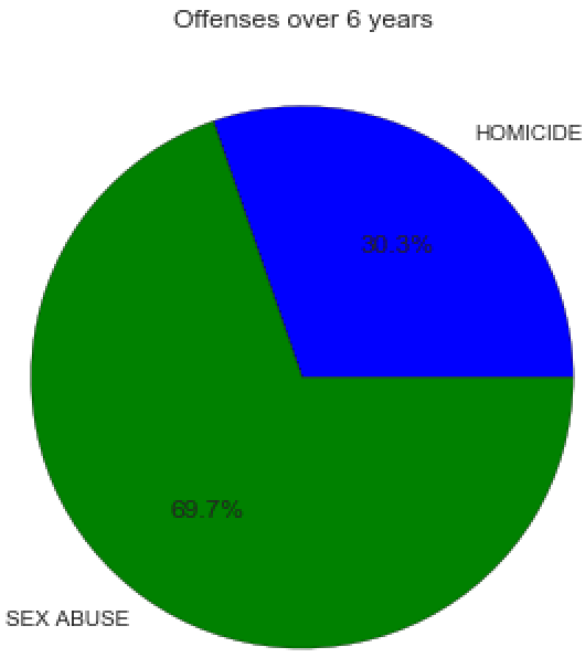


Figure 6

KPI 5: Percentage of offenses over 6 years

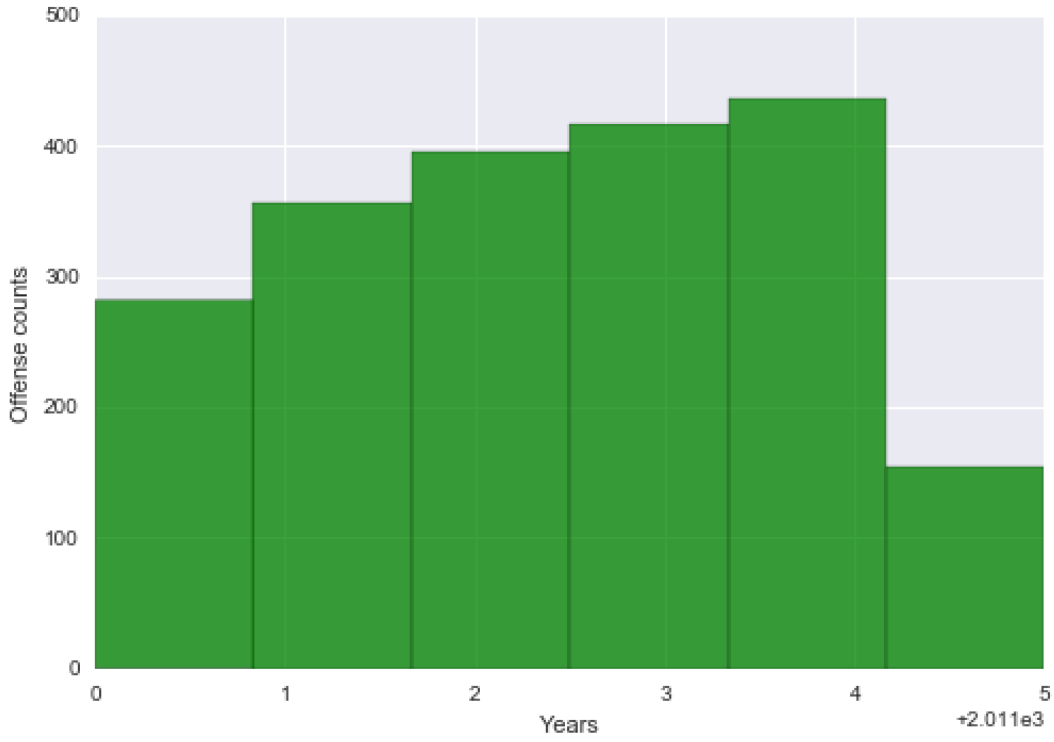


Figure 7

KPI 6: Number of offenses over 6 years

Drill-down and reporting

Law enforcement agencies need to be able to validate conclusions made on the basis of visual interpretation of the results of Link Analysis. A drill down to the selected patterns of interest helps in isolating the relevant records and validating our conclusions.

Put the slicing table sanjana sent you

Source Code:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
import sqlite3


df = pd.read_csv('Data_BI.csv')
print df.head(5)

g = sns.factorplot(x='year', data=df, kind='count', size=6)
g.set_axis_labels('Year', 'Number of Crimes')

g = sns.factorplot(x='year', hue="SHIFT", data=df, kind='count', size=6)
g.set_axis_labels('Year', 'Number of Crimes')

g = sns.factorplot(x='OFFENSE', data=df, kind='count', size=6)
g.set_axis_labels('Offense', 'Number of Crimes')
g.set_xticklabels(rotation=90)


var = df.groupby(['year','SHIFT']).OFFENSE.count()
var.unstack().plot(kind='bar',stacked=True,  color=['red','blue', 'yellow'], grid=False)
plt.title("No. of offences as per shift")
plt.show()

va=df.groupby(['OFFENSE']).sum().stack()
temp=va.unstack()
x_list = temp['year']
label_list = temp.index
plt.axis("equal")
plt.pie(x_list,labels=label_list,autopct="%1.1f%%")
plt.title("Offenses over 6 years")
plt.show()

plt.hist(df.year, bins=6, facecolor='green', alpha=0.75)
plt.plot(6, df.OFFENSE.count())
plt.axis([2011, 2016, 0, 500])
plt.xlabel("Years")
plt.ylabel("Offense counts")
plt.show()
```

8. LIMITATIONS AND POSSIBLE IMPROVEMENTS

This study did not test the influence of station design and management characteristics on crime outcomes as the design and management characteristics were uniform for Metro stations. Future studies on crime at and around metro stations can further explore the effect of this by a thorough examination of new design and management characteristics at Metro stations.

Learn from historical crime patterns and enhance crime resolution rate. Preempt future incidents by putting in place preventive mechanisms based on observed patterns. Reduce the training time for officers assigned to a new location and having no prior knowledge of site-specific crime patterns. Increase operational efficiency by optimally redeploying limited resources (like personnel, equipment, etc.) to the right place at the right time

9. CONCLUSION

The considered case illustrates an overall process for implementing a text mining solution and proves the feasibility and value of performing simultaneous analyses of both text and structured data within the same software system. The discovered results help investigators identify hidden patterns through the automated analysis of historical police reports. Till date, this knowledge was largely dependent on local expertise (so called 'local veterans'). Moreover, the new approach to the analysis delivers a much more comprehensive and objective overall picture of the incidents as it involves evaluating both structured and textual portions of the database. Law enforcement agencies and government organizations can benefit from this combination of text mining and pattern analysis technologies by achieving:

- Conclusion Improved crime resolution rate
- Optimal resource allocation based on dynamically changing patterns
- Faster and more up to date results from raw data
- Reduced officer training time and costs

- Better crime prediction and prevention of offences

10. REFERENCES