Crime Data Analysis

Rishi Reddy Cheruku
Pradeep Thomas Yerramothu
Department of Computer Science, University Of Akron
rc81@zips.uakron.edu
py11@zips.uakron.edu

1. Introduction

As the crime in today's world is increasing rapidly, there is a need for the law enforcement to analyze the crime data to allocate the time and man power required accordingly. The data can also be used to analyze the trends in crime and placement of law enforcement officials in a state.

1.1 Definition of Crime Analysis:

The Qualitative and Quantitative study of crime and law enforcement information in combination with socio-demographic and spatial factors to apprehend criminals, prevent crime, reduce disorder and evaluate organizational procedures [1]. According to FBI (Federal Bureau of Investigation), crimes can be mainly categorized into two categories: Property Crimes and Violent Crimes. Property Crimes can be further classified into Burglary, Larceny-theft, Motor vehicle theft and Arson, violent crimes can be classified into Murder and non-negligent manslaughter, Forcible rape, Robbery and Aggravated assault.

2. Motivation

Importance of Crime Data Analysis and its relevance to Data Mining:

Crime Data Analysis aides police work and enable investigators to allocate time and man power to other valuable tasks as and when required [2]. The main objective of a crime analysis is to increase the efficiency of the analysis and reduce the error rate in the analysis. A major challenge is to accurately and efficiently analyze the growing volumes of crime data. Data Mining is a powerful tool that enables us to explore large databases quickly and efficiently to generate useful patterns. The knowledge discovered from the existing data leads to reveal a value-added of its information.

A criminal act can compass of wide range of activities from civil infractions such as illegal parking to a mass murder. Law enforcement agencies compile all these statistics in order to maintain some sort of stability in the law [3]. Traditional Mining Techniques such as Association analysis, classification and prediction, cluster analysis and outlier analysis identify patterns in structured data. Especially Clustering techniques group data items into classes with similar characteristics to maximize or minimize interclass similarity. For example, to identify suspects who conduct crimes in similar ways or distinguish among groups belonging to different gangs [4]. Also visualization of those clusters on a graph or a map conveys more information by generating useful patterns.

As an effort to generate some patterns from the Crime data provided by the FBI, we analyzed the crime data and the law enforcement official data to generate some important results.

3. Dataset

The dataset consists of the crime statistics and population for all the cities across United States from the year 2006-2012 except for the year 2009. The dataset also consists of the Law enforcement data for mentioned time frame in those respective cities. The crime data given basically consists of the fields like state, city, population, violent crime, murder, forcible rape, robbery, assault, property crime, burglary, theft, motor vehicle, arson. The Law data consists of the fields like state, city, population, law enforcement officer, officers and civilians. We have described the number of records in the each year for crime and law in table 1.

4. Problem Formulation

As the crime in today's world is growing fast, there is a need for analyzing the data and generate useful patterns. For our project, we have considered the crime statistics and law enforcement officer data for all the cities in USA from the year 2006-2012. We have collected the data from Federal Bureau of Investigation (FBI). The problem considered was to classify all the cities with respect to the type of Crime and number of law enforcement officials in that particular location. Our goal was to ensure if there are sufficient law enforcement officers in a state by using the distribution of crime data each year from 2006-2012. The other problem was to find the distribution of crime data with respect to each city group.

5. Data Pre-processing:

The Data preprocessing phase includes data cleaning, recording selection and production of training and testing data. Additionally, datasets can be merged or aggregated in this step. As a part of Data Cleaning process, incompleteness and inconsistency was found and removed. The inconsistent data was sorted out by comparing the raw data. We have analyzed our data set and found the following issues with the data set.

- The state name was not filed for every record in the data set so we have filled the state name for the missing records for the respective cities.

State	City
ALABAMA	Abbeville
	Adamsville
	Alabaster

State	City
ALABAMA	Abbeville
ALABAMA	Adamsville
ALABAMA	Alabaster

- Extra character was appended for few city names like Illinois was written as illinois2. So we have removed the extra character appended to the city name.

- Few city names have already comma between them like "Richfield Township, Genesee County" so while converting the file into comma separated the word after "," are getting appended into next field. So we have replaced the "," with "|".

Then we have converted the given .xls file to comma separated file. We have migrated into an Oracle database by using a .ctl file and SQLLDR. SQLLDR is a bulk loader utility used for

moving data from external files into the oracle database. After inserting the crime data into crime tables and laws enforcement officer's data into law table we have joined the two tables with year, state and city. The below table gives the number of records per year in crime and law officers and the records we have considered.

Year	Crime Data	Law enforcement officers data	Records considered
2006	8252	8933	6394
2007	8435	9340	6537
2008	8772	9415	6820
2010	9150	9508	7293
2011	9314	9980	7413
2012	9491	10607	7642
			42099

Table 1: The above table gives the number of records for crime and law officer's records we have per year and the number of records we have considered for that year.

Then we have grouped the data into five categories based on the population. The below figure illustrates the City group, population and Number of records in each city.

Type of City	Population	Number of records
Metropolitan (M)	≥ 250000	434
Urban (U)	≥ 100000 and < 250000	1074
Town (T)	≥ 25000 and < 100000	6411
Rural – 1 (R1)	≥ 15000 and < 25000	4687
Rural – 11 (R2)	< 15000	29493

Table 2: The above table gives the cities classified by population and number of records in each city.

6. Modeling Data cube:

Today's information services are published through Web systems, and massive log-data are output from the systems, a data cube can be used to extract knowledge from this data-set, a human analyst sets up a search space of the "when/where/what" conditions, and then he must issue many aggregate queries and data-mining queries under different constraints in the space[8]. For the crime data analysis we moved the preprocessed data into SQL server database. Then we have created a data source and a data source view. Then a star schema with one fact table and two dimension tables is created. The fact table actually contains the numeric data, such as violent crime, property crime. The following figure depicts how the crime data analysis looks when converted to star schema.

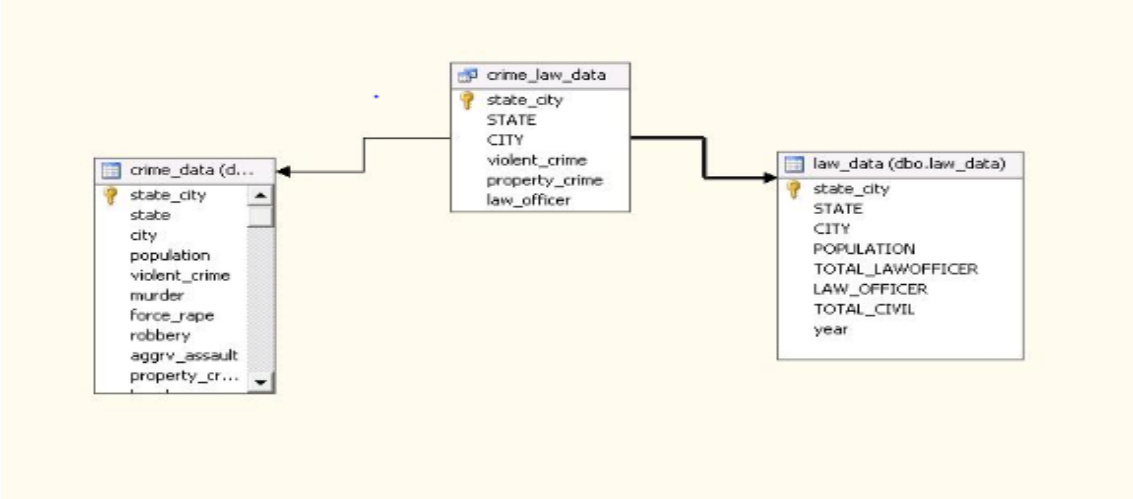

Figure 1: The above figure illustrates the star schema with a fact table and dimension tables.

The central table is the fact table and has two dimension tables. We have built two dimensions violent crime and property crime and built a data cube using these tables. A data cube consists of a lattice of cuboids, each corresponding to a different degree of summarization of the given multidimensional data. Then we have created violent crime hierarchy and property crime hierarchy for the created dimensions. Concept Hierarchies organizes the values of attributes/dimensions into gradual abstraction levels. We have two abstraction levels one is violent crime and other is property crime.

An online analytical processing (OLAP) can be performed in data warehouses/marts using the multi-dimensional data model. Few of the OLAP operations are roll up, drill down, slice, pivot and dice. We have created two OLAP reports one for violent crime and other for property crime with drill down features. Below figure illustrates the violent crime and property crime OLAP reports.

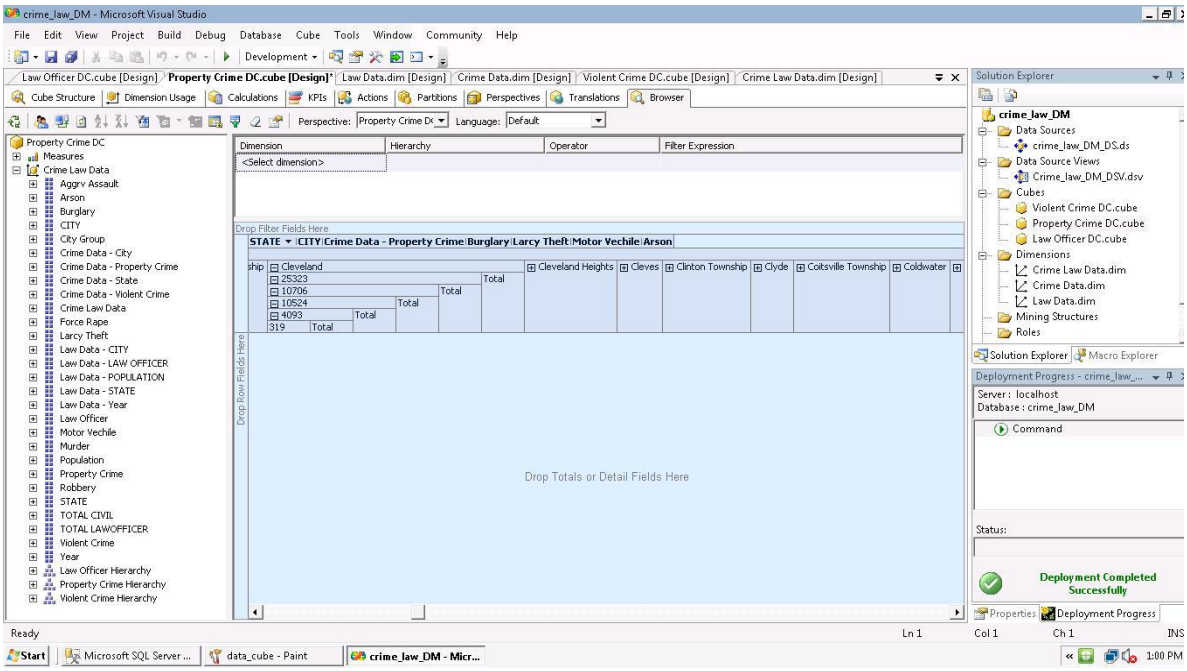

Figure 2: OLAP Report for property crime with drill down. This gives the details for all six years property crime data for Cleveland.

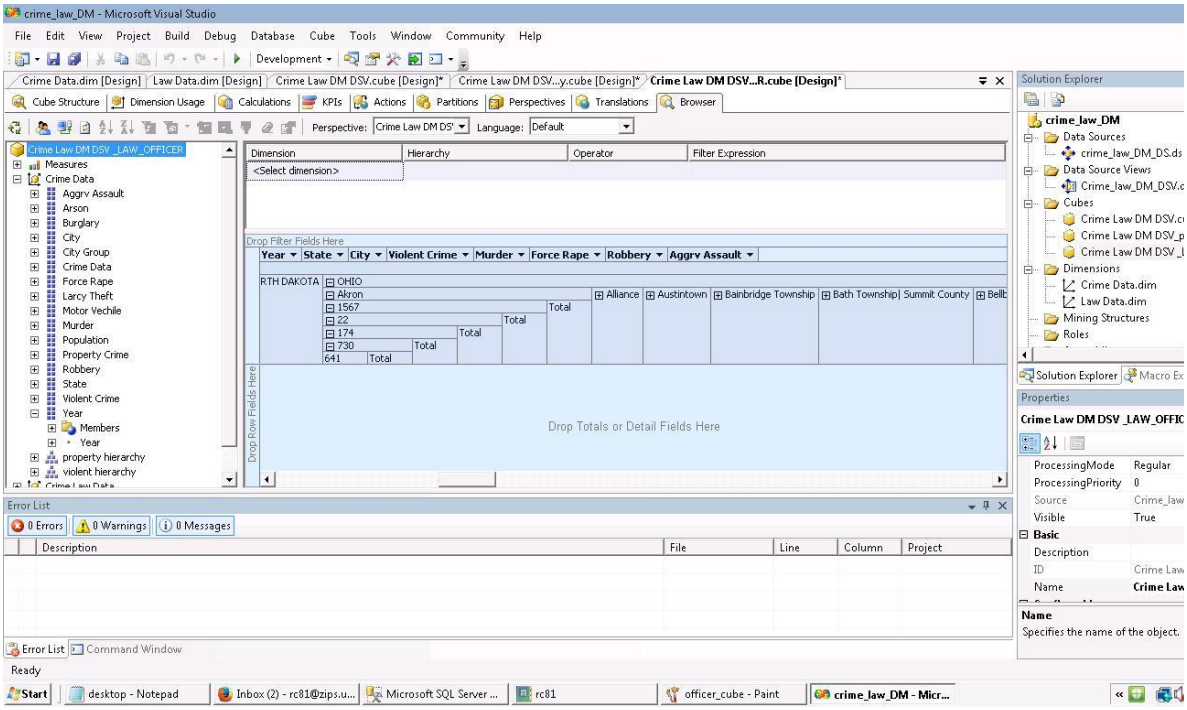

Figure 3: OLAP Report for violent crime with drill down. This gives the violent crime in Akron.

7. Cluster Analysis:

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. This process can be termed as clustering. A distance measure is usually used as data partitioning function. We have used Manhattan distance measure for implementing our project. Cluster Analysis can be done in different ways. We adapted a partitioning method called K-means Clustering to perform the clustering process.

K- Means clustering Algorithm [6] is one of the simplest unsupervised learning algorithms that can be used to solve the clustering problem. The procedure follows a way to classify the given dataset through a 'k' number of clusters. The main idea is to define k centroids, one for each cluster. The centroid is the mean of the cluster. Given a Dataset D containing n objects in Manhattan space, we divide the objects into k clusters $C_i \cup C_j = \emptyset$. Each data point in a given dataset is associated with the nearest centroid. This process is repeated until no points can be assigned to the cluster. The main objective of this clustering technique is to minimize the intra cluster similarity and maximize the inter cluster similarity.

WEKA is a tool with collection of machine learning algorithms for data mining tasks. The data which has been preprocessed is clustered using WEKA. Before running the algorithm on our data, we chose the distance measure as Manhattan distance and number of clusters as 5. We have considered the attributes **property crime per law officer** and **violent crime per officer** and **city group** for cluster analysis where city group is the nominal attribute. After the cluster analysis was performed, we have visualized the clusters to observe some patterns.

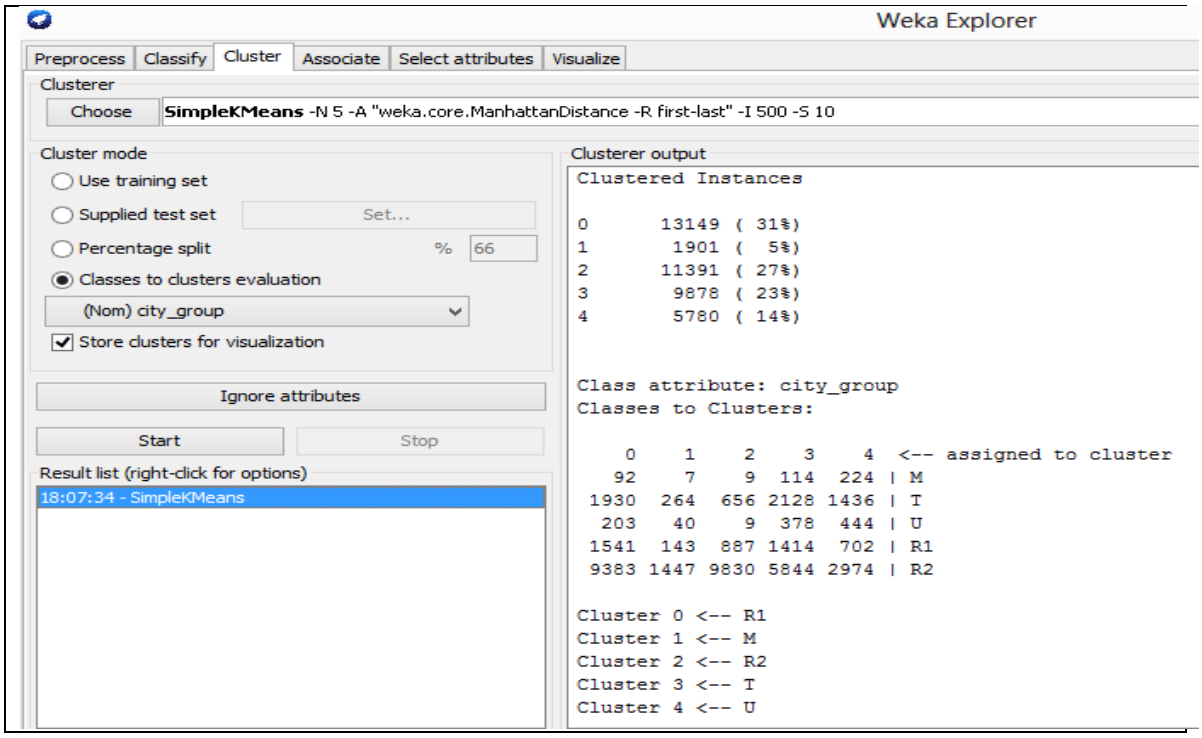

Figure 4: The Five clusters that are created using K-Means (Manhattan Distance) with help of WEKA.

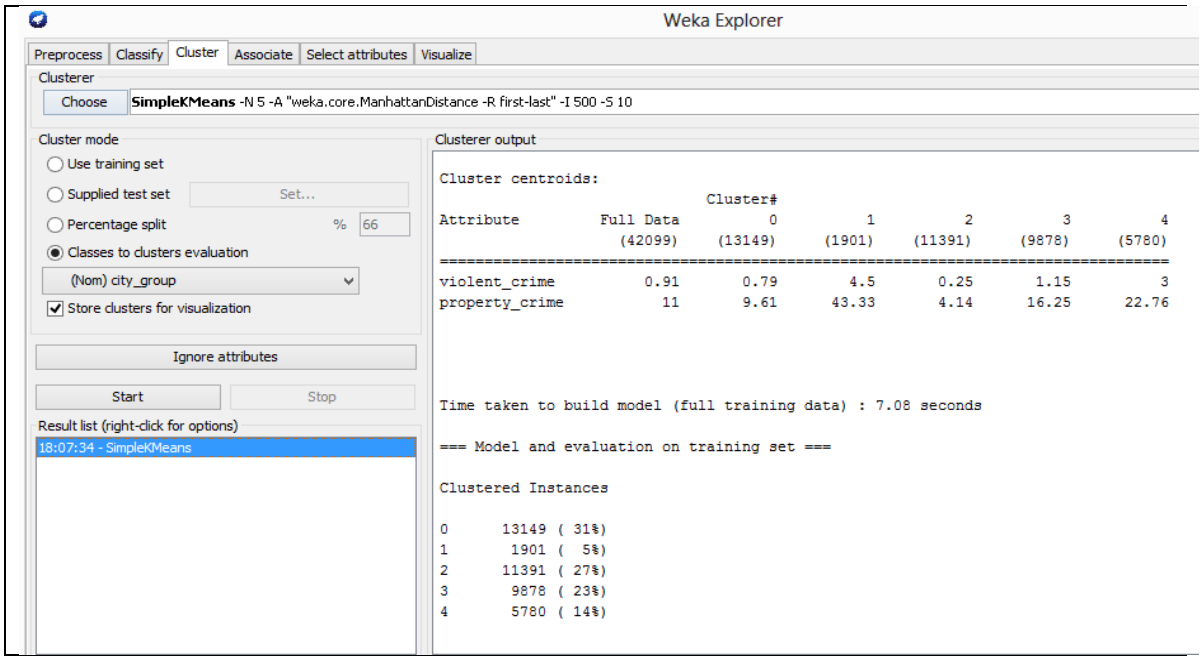

Figure 5: This figure illustrates mean among each cluster for violent and property crime. This gives an average of each kind of crime for that particular cluster.

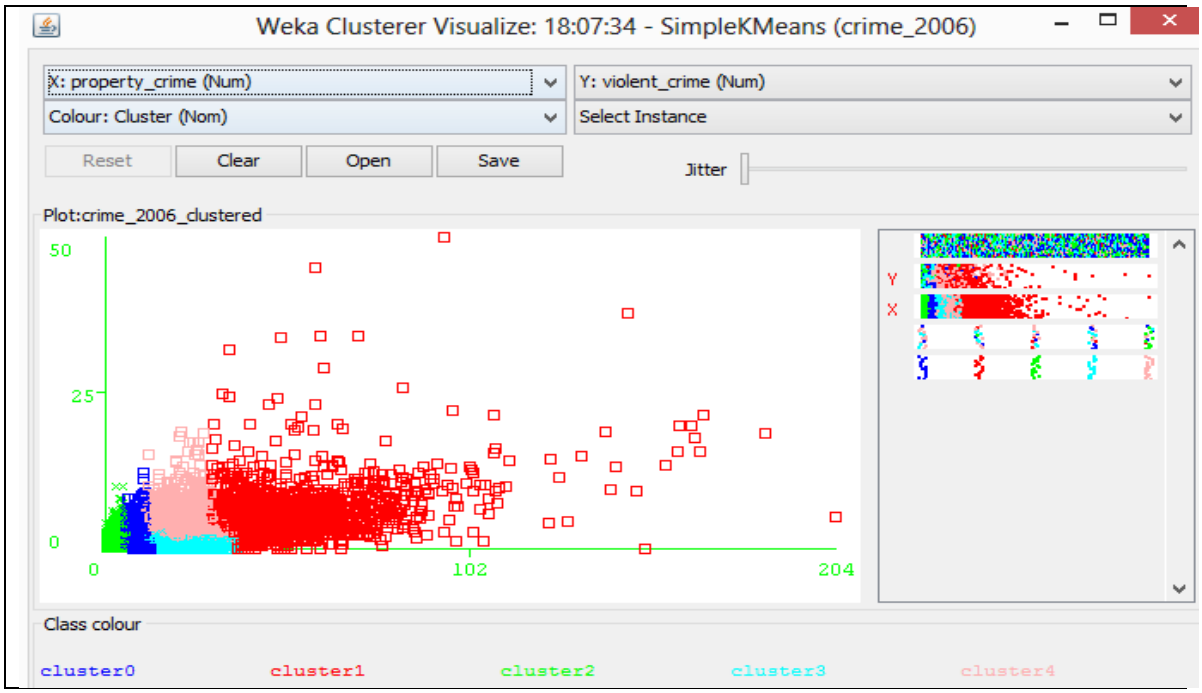

Figure 6: This figure illustrates the clusters for Property crime by violent crime.

Figure 5 illustrates there are over 4.5 violent crime and 43.33 property crime per law officer in metropolitan cities. The table gives the details about the remaining cities to. The bar graph between the crimes gives us an exact cluster differentiation and gives us each luster dominated by which set of population.

8. Data Transformation

The next step after clustering is to do the data transformation in order to perform visualization on the data set. We have transformed the data by using various factors that affect the Crime data. In order to visualize the crime data with respect to each year, we have added a weight factor to the data. The details given below actually illustrate the weight for each data.

If Crime in 2007 > 2006 then count = 1

Else if Crime in 2007=2006 then count = 0

Else crime in 2007<2006 then count = -1

Similarly we have given count for other years to. So based on these factors we got the count for each kind of crime and illustrated on Google map. Then we have used three different color combinations to distinguish the crime rates. The factor which we have used is -5 to -2 is green, -1 to 2 is yellow and 2 to 5 is red.

9. Data Visualization

Data Transformation is the method of consolidating data into one collective, illustrative graphic. Visualization is about Discovery, Discerning Patterns, and Disseminating Information. The goal of data visualization is to make the presented data informative. As a part of the mining process, we have visualized the dataset of crimes per law officer over 6 years, number of property crimes per person in each state, population per violent crime over 6 years. We have used the Google Fusion Tables [7], which is a web service provided, by Google Inc. The process is to upload the dataset on to the Google drive and create a fusion table. This service takes in the city name and maps its associated value on a Google Map. It also gives us the flexibility of grouping the data after it is uploaded to put them on a map.

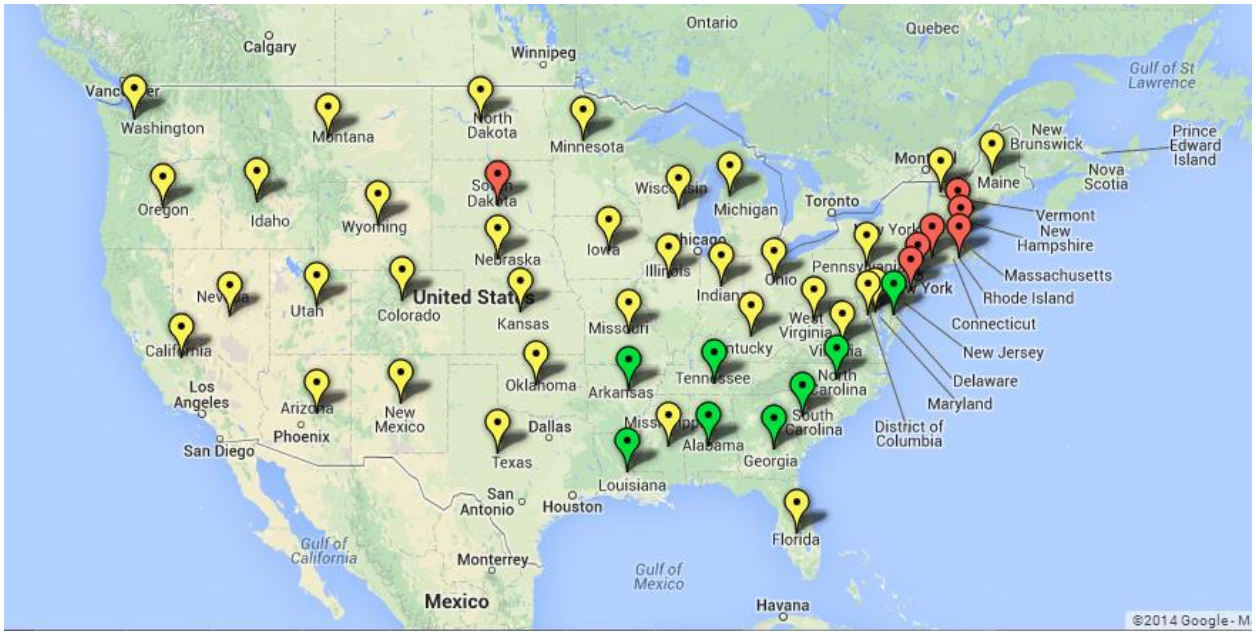

Figure 7: Property crime per person in each state. The red indicates that there is more crime in those states and green is neutral and yellow is good.

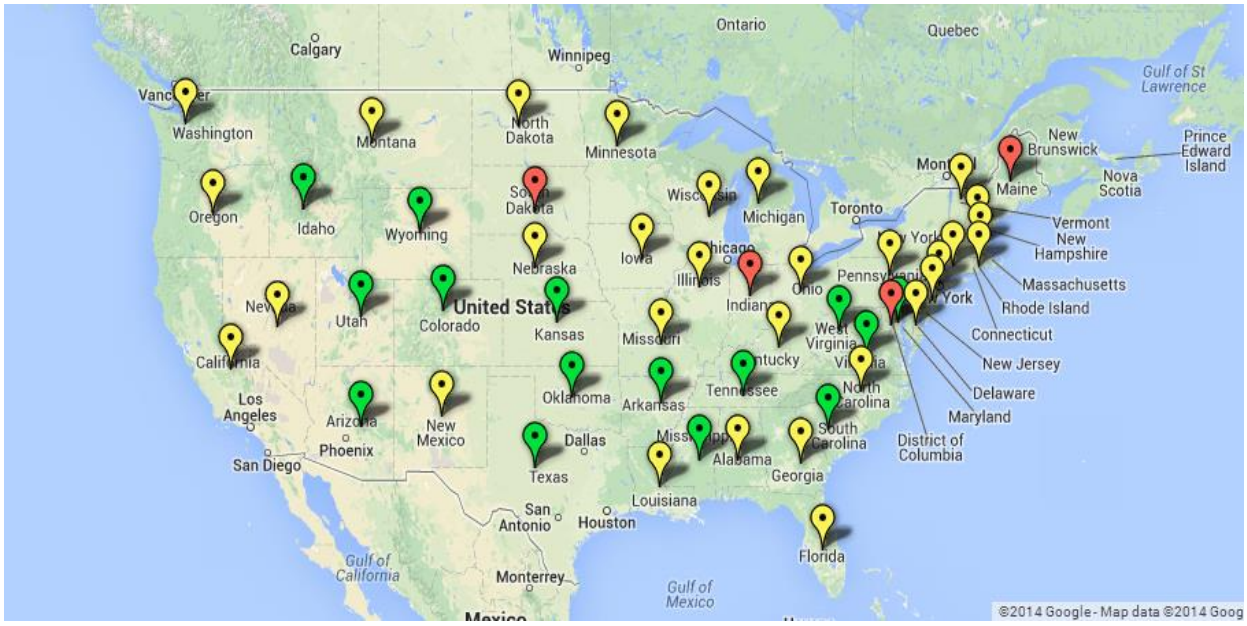

Figure 8: Crime per law officer in six years in each state. The red indicates that there is more crime in those states and green is neutral and yellow is good.

10. Conclusion:

Many efforts are taken by the governments to reduce the crime rates. In this process we have done analysis on the crime rate and illustrated with detail explanation and got few interesting patterns like where the crime is increasing and which states are the law enforcement needs to increase.

11. References:

[1].Crime Analysis,http://www.cops.usdoj.gov/Publications/introguidecrimeanalysismapping.pdf , (Accessed: 24 April 2014)

[2]. Hsinchun Chen; Chung, W.; Xu, J.J.; Wang, G.; Qin, Y.; Chau, M., "Crime data mining: a general framework and some examples," Computer, vol.37, no.4, pp.50, 56, April 2004 doi: 10.1109/MC.2004.1297301

[3]. Thongtae, P.; Srisuk, S., "An Analysis of Data Mining Applications in Crime Domain," Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on , vol., no., pp.122,126, 8-11 July 2008

[4]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[5]. Cities Classification based in the population, http://nces.ed.gov, (Accessed: 10 April, 2014)

[6]. Clustering, http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html, (Accessed: 29 April 2014)

[7]. Fusion Tables, https://support.google.com/fusiontables/answer/2527132?hl=en&topic=2573107&ctx=topic, (Accessed: 20 April 2014)

[8]. Ohmori, T.; Naruse, M.; Hoshi, M., "A New Data Cube for Integrating Data Mining and OLAP," Data Engineering Workshop, 2007 IEEE 23rd International Conference on , vol.,no., pp.896, 903, 17-20 April 2007 doi: 10.1109/ICDEW.2007.4401082