**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Data Analysis on Crime

Submitted by

**Fakhruddin Tahery** 1MS12CS029

**B M Tanvi Raj** 1MS12CS137

**Arshi Khan** 1MS13CS409

**B.Nishanth Varma** 1MS12CS019

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*

>



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu/), **May 2015**

# Abstract

Data mining and machine learning play a vital role in early crime detection and therefore, its subsequent prevention. We implement the regression and polynomial regression algorithms using the same finite set of features, on the crime dataset. The main objective of the project is to provide time-based, location-based and crime-type based rate and count ,which may help the society to avoid the highly concentrated

places of crime.

Crime maps are becoming significant tools in crime and justice. Advances in the areas of information technology and Geographic Information Systems (GIS) have opened new opportunities for the use of digital mapping in crime control and prevention programs. Crime maps are also valuable for the study of the ecology and the locational aspects of crime. Maps enable areas of unusually high or low concentration of crime to be visually identified. Maps are however only pictorial representations of the results of more or less complex spatial data analyses.

A hierarchical model dealing with crime analysis is proposed and applied to the regional analysis of crime in Chicago, the model helps to identify spatial concentration of crimes in specific area (area based method). In area-based methods, crime data are aggregated into geographical areas such as blocks, precincts, and for each area, the analyst computes a measure of crime value. Multicriteria evaluation concept has been used to assess the crime rate in various blocks a discrete (part) of Tehran city. In this part we used two methods for crime density assessment:

Crime assessment based on crime per block,

Crime assessment based on density of crime per population.

After determination of hot spots based on two methods mentioned above spatial function is used to find suitable location to establish new police station or direct patrol to the hot spots to reduce of crime.

# Contents

***Declaration***

***Acknowledgements ii***

***Abstract iii***

***List of Figures***

***List of Tables***

1 **INTRODUCTION**

* 1. General Introduction………………. **<Page No>**
  2. Statement of the Problem…………..
  3. Objectives of the project……………
  4. Project deliverables……………
  5. Current Scope………………………
  6. Future Scope……………………….

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1….Introduction

3.2…Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirement 1.1

4.3.2

4.3.3

4.3.4

4.3.5

4.3.6

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.4.6 Performance

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram and Classes (represent Inheritance, Aggregation and Association)
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK <No Page Numbers>**
2. **REFERENCES**

**1. INTRODUCTION**

* 1. **General Introduction**

Many type of crimes have a chance to take place at any particular area and

there must be some way such that one can avoid the same and protect himself/herself. We are implementing a idea which can match the necessity.

* 1. **Statement of the Problem**

Recognize the types of crime that occur at different places in Chicago [USA].

* 1. **Objectives of the Project**

Predict and provide preventive measures for the citizens in order to avoid the possible

crime situations.

* 1. **Project Deliverables**

Present a list to the society the types of crime possible at particular places.

Provide type, rate and count of any crime in chronological order.

* 1. **Current Scope**

As there are many chances for different crimes to take place , this idea shall give a chance

to an individual to avoid certain heat mapped places of particular crimes.

* 1. **Future scope**

The lack of opportunities due to the heavy completion in the main stream fields have a large effect in the increase of crime rate and has a very good chance to be increasing every year and this approach shall help the society.

**2. PROJECT ORGANIZATION:**

**Software process models**

The software process model used for this project is Iterative model. First we started segregating the data from the excel sheet based on crime type, year and count. After segregation we started coding on it. After each block of code we tested to check whether the code is running properly or not. In the third iteration we took feedback of our team mates and note them down to make further changes. In the final iteration we made the required changes to the code and other things discussed during the feedback phase.

Finally we checked the data predicted by SAP Lumira and the data predicted by our code.

**3. Literature Survey**

Crime is a multifaceted concept that can be defined in legal and non-legal sense. From a legal point of view, it refers to breaches of the criminal laws that govern particular geographic areas (jurisdictions) and are aimed at protecting the lives, property and rights of citizens within those jurisdictions. Most of the crimes with which the criminal justice system is concerned involve breaches of State/Territory legislation that cover most offences relating to persons (for example, murder and sexual assault), property (for example, theft and property damage) and regulation (for example, traffic violations).

Commonwealth legislation relates primarily to matters such as trade and commerce, importation/exportation, taxation, defence and external affairs.

Non-legal point of view would define crime as acts that violate socially accepted rules of human ethical or moral behaviour. As the moral principles that underpin the notion of crime are subject to gradual change over time, the types of behaviour defined by the legal system as criminal may also change. Examples of behaviours that have been de-criminalized in some jurisdictions include prostitution, abortion, attempted suicide and homosexual intercourse. Other behaviours, such as tax evasion or credit card fraud, have been criminalized over time.

The distinction between these two points of view is important. While criminal justice agencies' crime data will be based upon the relevant legal definition of crime, data collected via victimisation surveys may be based upon an individual's interpretation of crime, rather than upon existing legal definitions. This potential disparity is more likely to affect the less serious end of the 'crime' spectrum.

**4. SOFTWARE REQUIREMENT SPECIFICATIONS**

**4.1 External Interfaces**

**User interfaces:** The user interface for the software shall be compatible to any browser such as Internet Explorer, Mozilla or Netscape Navigator by which user can access to the system .The user interface shall be implemented using Python.

**Software interfaces:** The software will generate a heat map to identify all the hotspot locations where crime is concentrated.

The software also generates line graphs to analyze the crime rates and count in different countries across the globe.

**Communication interfaces:** The software shall use the HTTP protocol for communication over the internet.

**4.2 Functional Requirements**

4.2.1 Analyze predict location-based and crime-type based rate and count.

4.2.2 To generate heat map to identify the hotspot locations where crime is concentrated.

4.2.3 To provide solutions to crime problems and formulating crime prevention strategies.

* + 1. To predict which crimes are more likely to escalate.

**4.3 Software System Attributes**

4.3.1 **Reliability**

Application will not fail without manual intervention.

The application will be error-free.

4.3.2 **Availability**

The application will be available all the time.

4.3.3 **Security**

The system’s back-end servers shall only be accessible to authenticated administrators. It provides safeguards to prevent damage to data from operator errors, simultaneous updates, module unavailability or system failures etc.

* + 1. **Portability**

The application is portable. The software can be used on different environments .It is adaptable to different specified environments without applying actions or means other than those provided for this purpose for the system. The application is portable across multiple operating systems and Browsers (including both types and versions).

4.3.5 **Maintainability**

Ability to undergo modifications and repairs.

4.3.6 **Performance**

System will provide proper adequate response for accessing each of the modules. System will show quick response in simple search and will take little bit more time for advance search and this will vary based on input conditions. There are sufficient edit and validation checks in the system.

* 1. **Performance Requirements**

Application will not fail without manual intervention.

At state level, application will be built in common technology/platform to avoid any compatibility issues.

Will provide web services.

The System will employ a single set of user interface to provide a familiar and common look and feel for the application.

The system should be designed in manner that operational data will never be lost until a manual intervention/hardware failure.

Will provide services to users in future.

The system should ensure easy scalability and extensibility through minimum effort.

* 1. **Database Requirements**

The application does not need a database because data set is used.

**4.6 Design Constraints**

4.6.1 Understandability

The final presentation layer to the end user is immediately understandable, simple, recognizable, and intuitive

4.6.2 Cost

There are no implicit and explicit costs associated with Crime Data Analysis project.

4.6.3 Speeds

Fast retrieval of query result and it feeds the user correct result in real quick time.

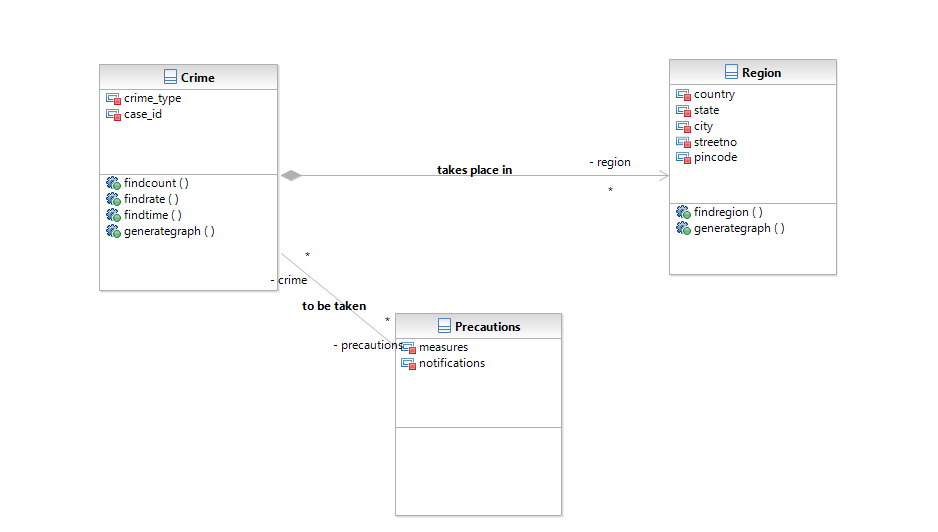
4.6.4 Data Relevance

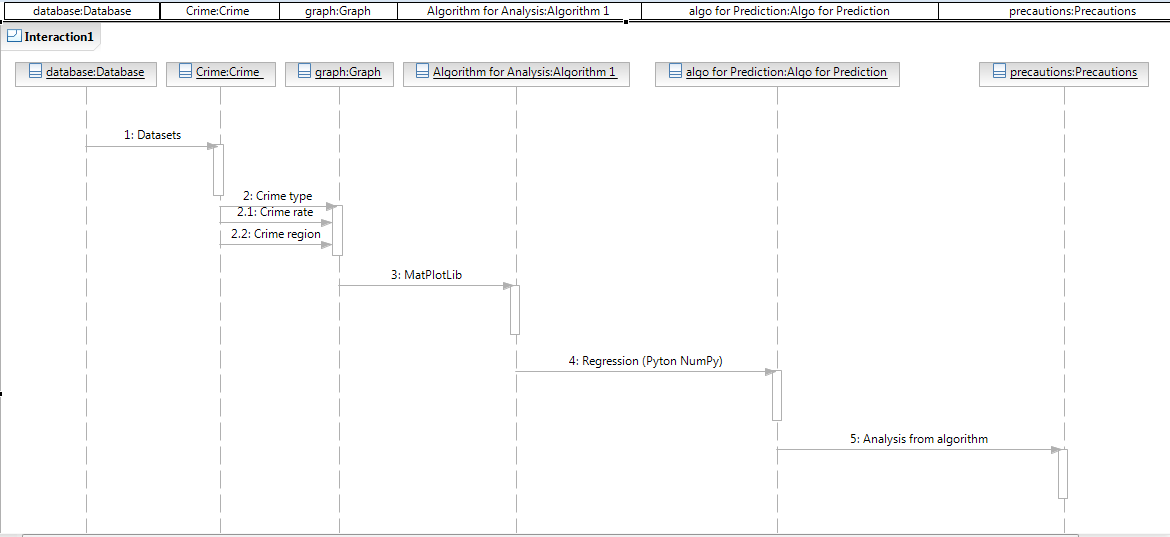
Data is relevant. It is exactly what the end- user wants.

4.6.5 Avoid Undue Relay

The application is designed in such a way that product is deployed within given time.

**5. Design**

**Class Diagram**



**Sequence Diagram**

**6. IMPLEMENTATION**

6.1 Tools Introduction

SAP Lumira

SAP Lumira (formerly called SAP Visual Intelligence) is a self-service, [data visualization](http://searchbusinessanalytics.techtarget.com/definition/data-visualization) application .Here in this analysis we use this application for plotting

A graph for consecutive years over a period of time.

Techniques used for the analysis and prediction of data is SAP Lumira tool, Linear and polynomial regression.

Linear Regression:

In statistics, **linear regression** is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple **linear regression**.

In linear regression, the relationships are modeled using [linear predictor functions](https://en.wikipedia.org/wiki/Linear_predictor_function) whoseunknown model parameters are [estimated](https://en.wikipedia.org/wiki/Estimation_theory) from the [data](https://en.wikipedia.org/wiki/Data). Such models are called [*linear models*](https://en.wikipedia.org/wiki/Linear_model). Most commonly, the [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of *y* given the value of *X* is assumed to be an [affine function](https://en.wikipedia.org/wiki/Affine_transformation) of *X*; less commonly, the [median](https://en.wikipedia.org/wiki/Median) or some other [quantile](https://en.wikipedia.org/wiki/Quantile) of the conditional distribution of *y* given *X* is expressed as a linear function of *X*. Like all forms of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), linear regression focuses on the [conditional probability distribution](https://en.wikipedia.org/wiki/Conditional_probability_distribution) of *y* given *X*, rather than on the [joint probability distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution) of *y* and *X*, which is the domain of [multivariate analysis](https://en.wikipedia.org/wiki/Multivariate_analysis).

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

* If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
* Given a variable *y* and a number of variables *X*1, ..., *Xp* that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the *Xj*, to assess which *Xj* may have no relationship with *y* at all, and to identify which subsets of the *Xj* contain redundant information about *y*.

Polynomial Regression:

In [statistics](https://en.wikipedia.org/wiki/Statistics), polynomial regression is a form of  [in which the relationship between the independent variable](https://en.wikipedia.org/wiki/Linear_regression) *x* and the dependent variable *y* is modelled as an *n*th degree [polynomial](https://en.wikipedia.org/wiki/Polynomial). Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of *y*, denoted E(*y* | *x*), and has been used to describe nonlinear phenomena such as the growth rate of tissues,the distribution of carbon isotopes in lake sediments,and the progression of disease epidemics.Although *polynomial regression* fits a nonlinear model to the data, as a [statistical estimation](https://en.wikipedia.org/wiki/Estimation_theory) problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown [parameters](https://en.wikipedia.org/wiki/Parameter) that are estimated from the [data](https://en.wikipedia.org/wiki/Data). For this reason, polynomial regression is considered to be a special case of multiple linear regression.

The goal of regression analysis is to model the expected value of a dependent variable *y* in terms of the value of an independent variable (or vector of independent variables) *x*. In simple linear regression, the model



is used, where ε is an unobserved random error with mean zero conditioned on a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) variable *x*. In this model, for each unit increase in the value of *x*, the conditional expectation of *y* increases by *a*1 units.

In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improves by increasing amounts for each unit increase in temperature. In this case, we might propose a quadratic model of the form



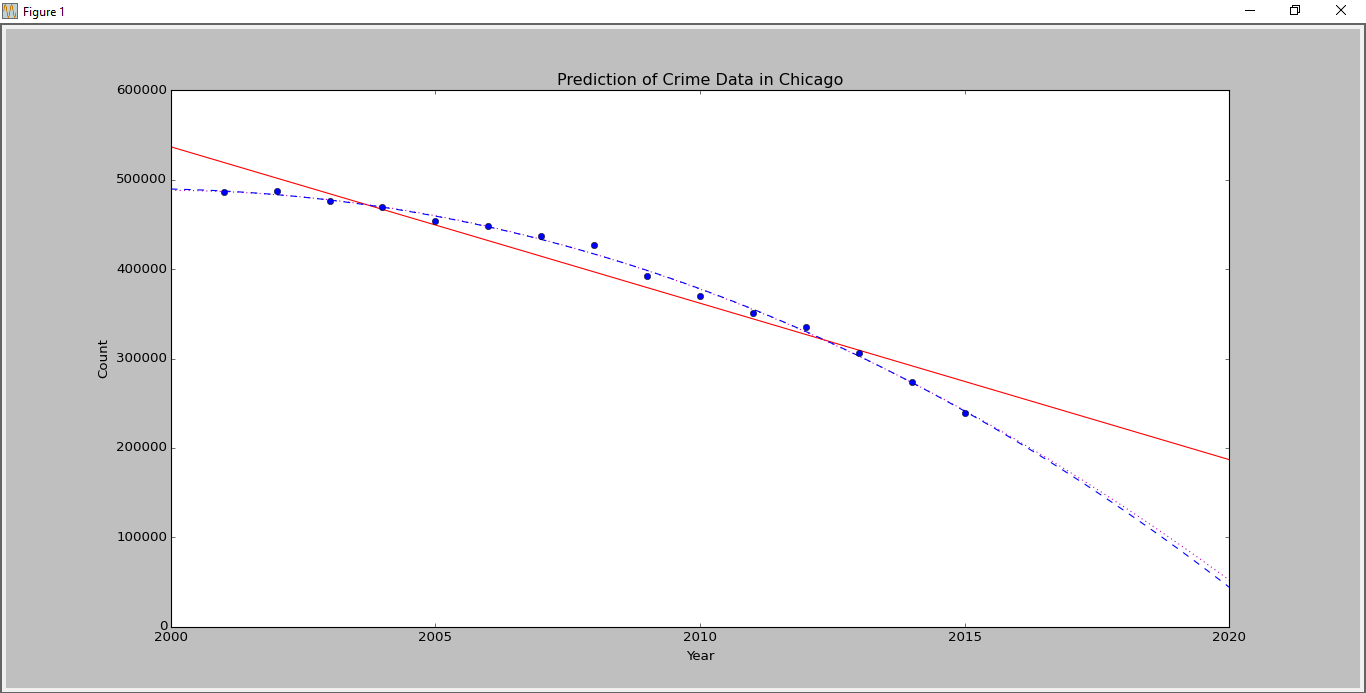
In this model, when the temperature is increased from *x* to *x* + 1 units, the expected yield changes by *a*1 + 2*a*2*x*. The fact that the change in yield depends on *x* is what makes the relationship nonlinear (this must not be confused with saying that this is nonlinear regression; on the contrary, this is still a case of linear regression).

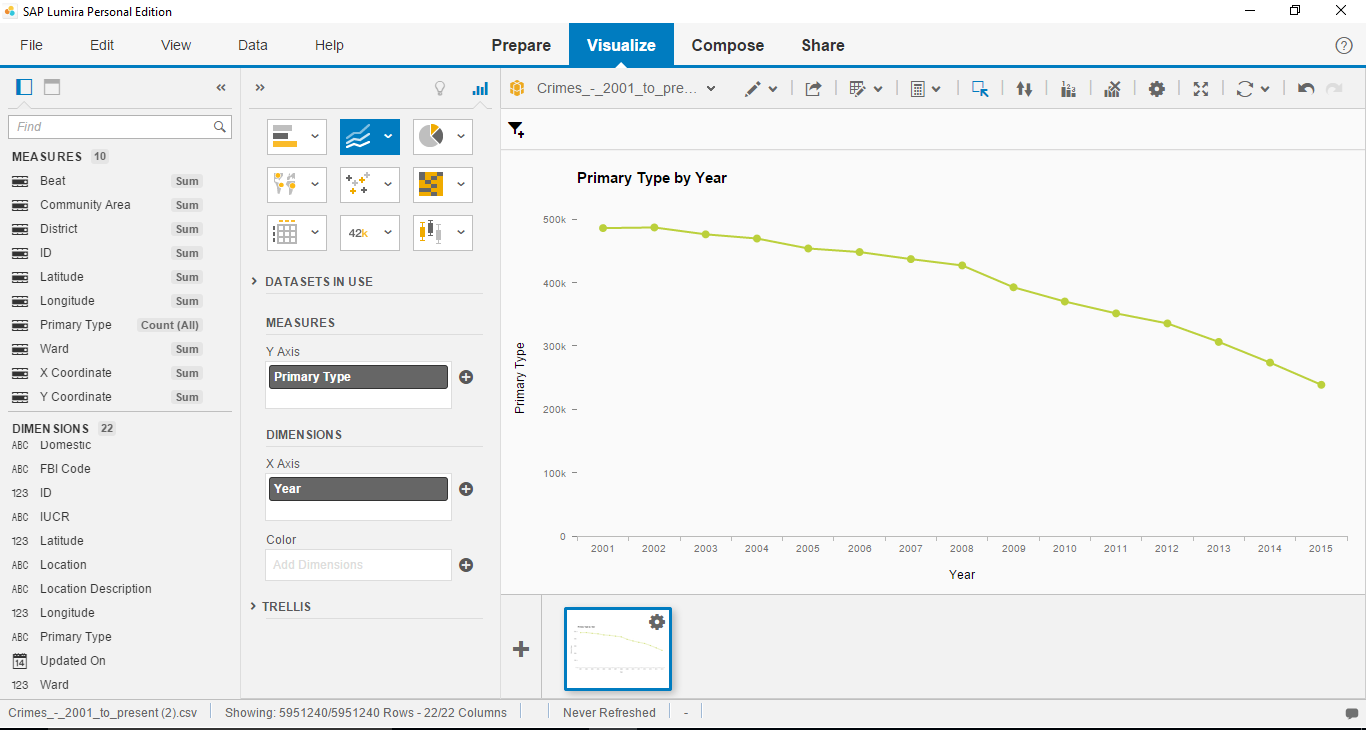
In general, we can model the expected value of *y* as an *n*th degree polynomial, yielding the general polynomial regression model

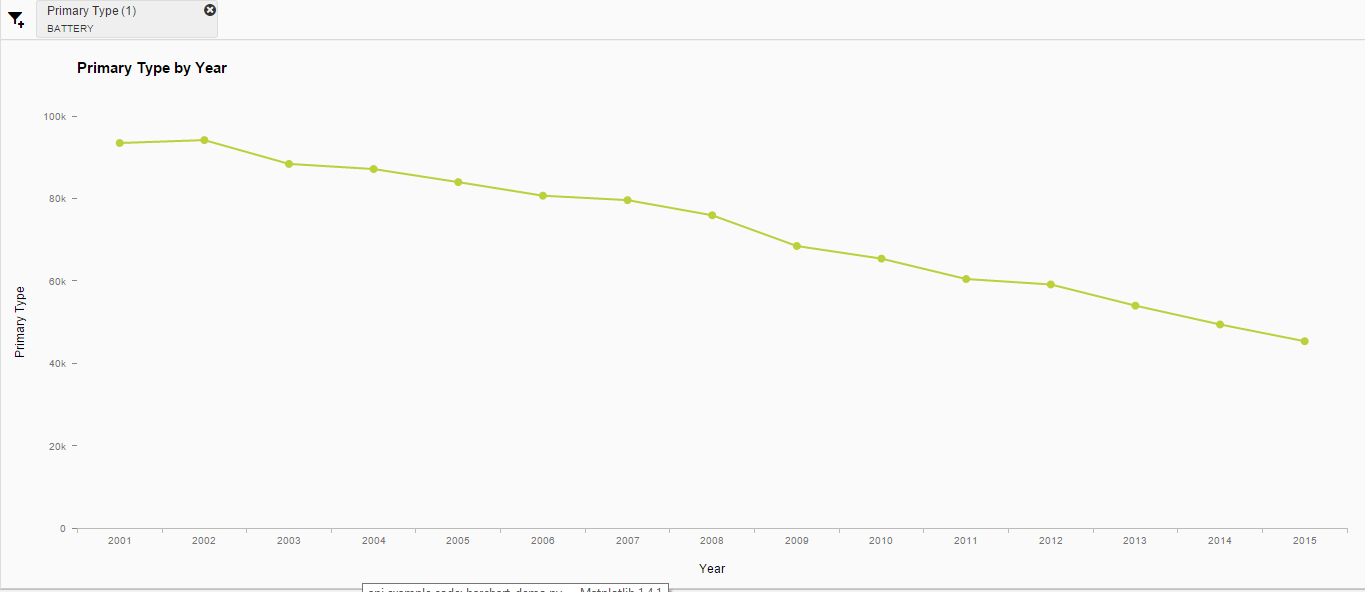


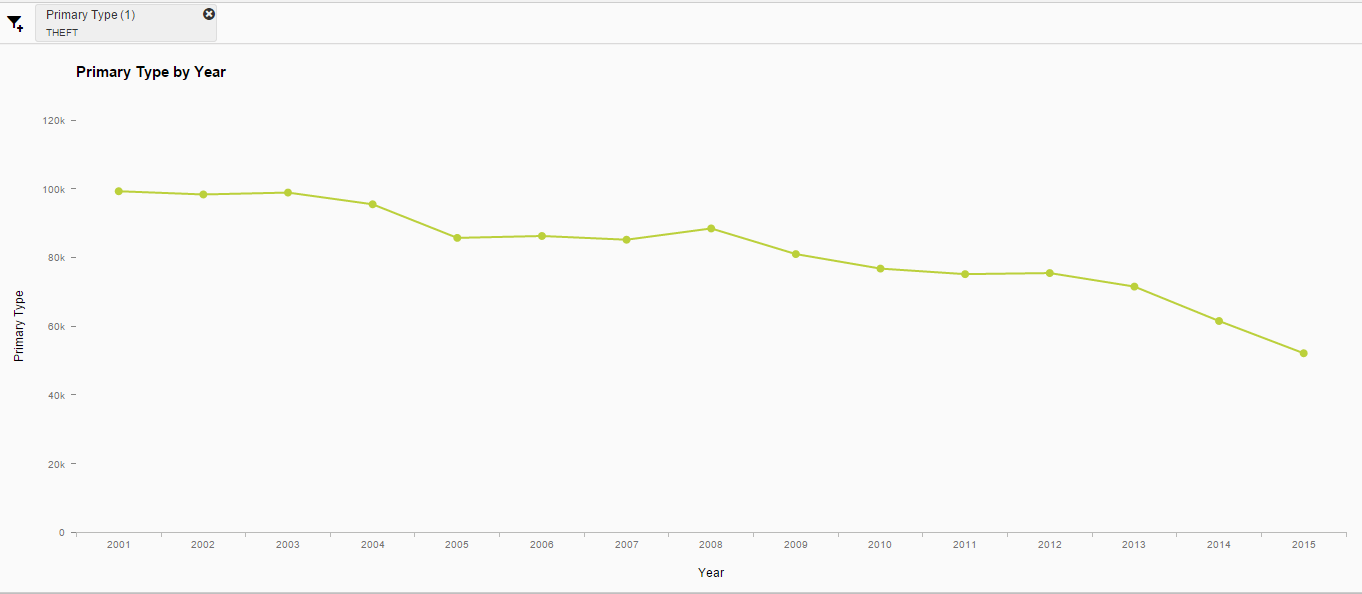
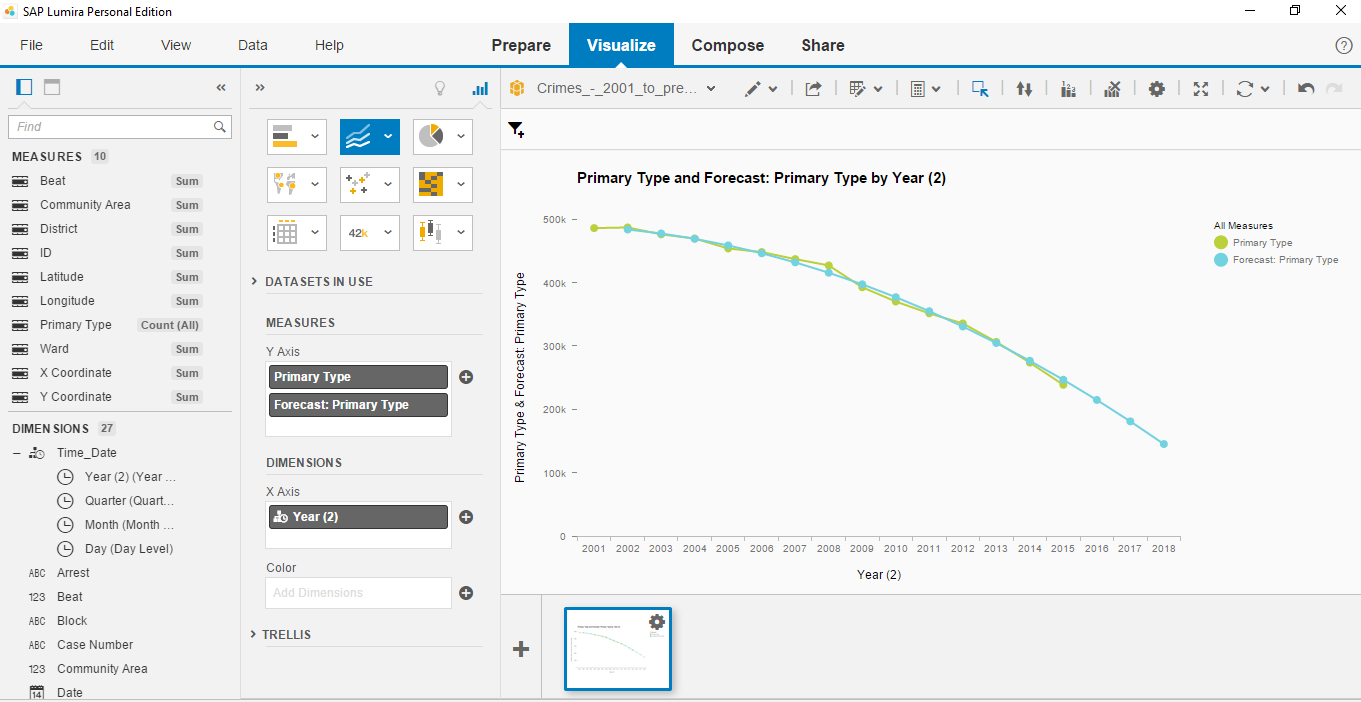
Conveniently, these models are all linear from the point of view of [estimation](https://en.wikipedia.org/wiki/Estimation_theory), since the regression function is linear in terms of the unknown parameters *a*0, *a*1, .... Therefore, for [least squares](https://en.wikipedia.org/wiki/Least_squares) analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of [multiple regression](https://en.wikipedia.org/wiki/Linear_regression). This is done by treating *x*, *x*2, ... as being distinct independent variables in a multiple regression model.

7. Screenshots and Results:









**8. CONCLUSION & SCOPE FOR FUTURE WORK**

We have successfully completed the analysis and prediction on crime data of Chicago. The analysis of the data is done using a tool called SAP Lumira.SAP Lumira (formerly called SAP Visual Intelligence) is a self-service, [data visualization](http://searchbusinessanalytics.techtarget.com/definition/data-visualization) application for business users.  SAP Lumira grew out of [SAP Business Explorer](http://searchsap.techtarget.com/definition/BEx), the company's previous, but more basic, self-service data visualization tool. According to SAP, the two are  meant to go hand-in-hand. For example, analysts using SAP Lumira can alter data structures and correlations in whatever way they want, without the help of their IT department, then push the data back into the system so it can then be consumed by more casual users in tools like Business Explorer.

The prediction of the data is done using Python. Library used for predicting the data is Scipy, Numpy, Matplotlib and Pandas. Matplotlib is used for plotting graphs.

**Scope:**

The deliverables of the project is a python code and sap lumira visualization which will run without any errors. The code takes care of all the null values and other nan data.

**9. References:**

[1]ViolentCrime.http://www2.fbi.gov/ucr/cius2009/offenses/violent\_crime/.

[2]Murder.http://www2.fbi.gov/ucr/cius2009/offenses/violent\_crime/murder\_homicide.html.

[3]ForcibleRape. <http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/forcible_rape.html>.

[4]Robbery.http://www2.fbi.gov/ucr/cius2009/offenses/violent\_crime/robbery.htm

[5]Assault.http://www2.fbi.gov/ucr/cius2009/offenses/violent\_crime/aggravated\_assault.html.

[6]Mississippi Crime Rates and Statistics - NeighborhoodScout. Mississippi Crime Rates and Statistics - NeighborhoodScout. Accessed February 17, 2015. <http://www.neighborhoodscout.com/ms/crime/>.

[7] S. M. Nirkhi, R.V. Dharaskar and V.M. Thakre. "Data Mining : A Prospective Approach for Digital Forensics," International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 6 pp. 41-48, 2012.

[8] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/.

Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015 12

[9] E. W. T. Ngai, L. Xiu, and D. C. K. Chau. "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," Expert Systems with Applications, pp. 2592–2602, 2008.

[10] J. McCarthy, "Arthur Samuel: Pioneer in Machine Learning," AI Magazine, vol. 11, no. 3, pp. 10-11, 1990.