**Chapter 1**

**Introduction**



* 1. **Overview**

The study of daily reports and crime to determine the location, time of day, special characteristics, and similarities to other crimes as well as any significant data that will or may identify the existence of patterns of criminal behaviour.

Crime scene analysis (crime analysis) is the analytical process of interpreting the specific features of a crime and related crime scenes. It involves an integrated assessment of the forensic evidence, forensic [victimology](https://www.sciencedirect.com/topics/social-sciences/victimology), and crime scene characteristics. The results of crime scene analysis (CSA) may be used to determine the limits of the available evidence and the need for additional investigative and forensic efforts, as in a threshold assessment (discussed shortly). When sufficient behavioural evidence is available, these same results may also be used to infer offender modus operandi (MO) and signature behaviours, evidence of crime scene staging, crime scene motive, and offender characteristics, or to assist with linkage analysis efforts.

This is done through the collection, collation, analysis of reports, and evaluation of crime data. Crimes are random and non-random foreseeable and not foreseeable and take place in high crime areas as well as low crime areas. Crime analysis becomes an essential tool for short- and/or long-term planning of personnel, budgets, facilities, and equipment. Have your numbers gone up or have they gone down? Your data are key to measurement. Data do not necessarily have to reflect public service needs, phone calls, or unlocking specific doors or cabinets. Even if crime is down, the calls for noncriminal services could be going up and up.

Crime analysis is a critical first step in determining the need for crime/loss prevention programs as well as identifying problems on the site. Crime analysis comes into play again in the evaluation of programs. Is crime up or is it down? Compared to whom and what? Is the city crime rate up and yours down? You can have a low crime rate in a high crime area. So how do you get such information? Police departments break their crime down into areas or zones. Each zone may be given a number, for example, 1 to 8 and you are located in zone six. So how does zone six compare to the other seven areas? Effective programming begins with results of relevant and reliable data/information.

* 1. **Objective**

Crime refers to an action that violates law and comes under punishable offense, any of the social, emotional or economic imbalance can lead a person to commit crime. Every day we can hear about some criminal activity, there is not a single day when there was no crime reported throughout the world. Precautions are indeed necessary but that will ensure our safety is not true. Therefore technology and data science is needed to combat this problem. The historic data of criminals and types of crime committed collected by the police department from various geographic locations can be used. Applying analytical tools to find out a trend or pattern in the behavior of crime is called crime analysis.

**Chapter 2**

**Software Development Tools**

**2.1 Tools Used**

Python – The project is developed in python.

#### Jupyter Notebook – Jupyter Lab is a web-based interactive development environment for Jupyter notebooks, code, and data. Jupyter Lab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. Jupyter Lab is extensible and modular: write plug-ins that add new components and integrate with existing ones.

Python IDLE - IDLE is integrated development environment (IDE) for editing and running Python 2.x or Python 3 programs. The IDLE GUI (graphical user interface) is automatically installed with the Python interpreter. IDLE was designed specifically for use with Python.

**2.2 Modules and their functionalities**

**NumPy** - NumPy is the fundamental package for scientific computing with Python. It contains among other things:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the BSD License, enabling reuse with few restrictions.

**Pandas** - Python has long been great for data munging and preparation, but less so for data analysis and modeling. Pandashelps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

Python has long been great for data munging and preparation, but less so for data analysis and modeling. Pandashelps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

**Matplotlib** - Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

**Word Cloud** - A Word cloud (or Tag cloud) is a visual representation of text data. It displays a list of words, the importance of each being shown with font size or colour. This format is useful for quickly perceiving the most prominent terms. [Python](https://www.python.org/) is totally adapted to draw this kind of representation, thanks to the word cloud library developed by [Andreas Mueller](http://amueller.github.io/). This page gives you a few examples showing the main features and customization you can apply.

**2.3 Requirements**

1) Hardware Requirements

• Minimum 4 GB RAM, and

• Minimum i3 processor.

2) Software Requirements

• Any windows based operating system,

• Python,

• Jupyter

**Chapter 3**

**Libraries**

**3.1 Pandas**

Pandasis a Python pa**c**kage providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexibleopen source data analysis / manipulation tool available in any language**.** It is already well on its waytoward this goal.

**Functions:**

* Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
* Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
* Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations
* Powerful, flexible groupby functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
* Make it easy to convert ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects
* Intelligent label-based slicing, fancy indexing, and subsetting of large data sets
* Intuitive merging and joining data sets
* Flexible reshaping and pivoting of data sets
* Hierarchical labeling of axes (possible to have multiple labels per tick)
* Robust IO tools for loading data from flat files (CSV and delimited), Excel files, databases, and saving / loading data from the ultrafast HDF5 format
* Time series-specific functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

**Installation:**

The best way to get pandas is via conda.

condainstallpandas

Packages are available for all supported python versions on Windows, Linux, and MacOS.

Wheels are also uploaded to PyPI and can be installed with

pip install pandas

**3.2 Matplotlib**

Matplotlib.pyplotis a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

Matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes (please note that "axes" here and in most places in the documentation refers to the *axes*part of a figure and not the strict mathematical term for more than one axis).

**Installation:**

Matplotlib and its dependencies are available as wheel packages for macOS, Windows and Linux distributions:

python-mpipinstall-Upip

python-mpipinstall-Umatplotlib

**Functions:**

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

**3.3 NumPy**

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the BSD license, enabling reuse with few restrictions.

**Functions:**

The main **benefits** of using **NumPy** arrays should be smaller memory consumption and better runtime behavior. So the more numbers you need to store - the better you do. This shows some performance numbers of operations between **Python** and **Numpy**.

**Installation:**

NumPy can be installed via following command:

pip install numpy

**3.4 Word Cloud**

A Word cloud (or Tag cloud) is a visual representation of text data. It displays a list of words, the importance of each being shown with font size or color. This format is useful for quickly perceiving the most prominent terms. [Python](https://www.python.org/) is totally adapted to draw this kind of representation, thanks to the word cloud library developed by [Andreas Mueller](http://amueller.github.io/). This page gives you a few examples showing the main features and customization you can apply.

**Functions:**

* **It reveals the essential.** Brand names pop and key words float to the surface.
* **They delight and provide emotional connection.** Both the creation of a word-cloud and the observation of one help to provide an overall sense of the text. The same visceral response doesn't happen when staring at a page of text.
* **They're fast.** Poring over text to develop themes from research takes time.
* **They're engaging.**Visual representation of data tends to have an impact and generates interest amongst the audience. For your client, it may stimulate more questions than it answers, but that's often a good entry point to discussion. In addition, Christy Ransom recently mentioned how word clouds are a great way to show themes for engaging your community panel members. Word clouds can allow you to share back results from research in a way that doesn't require an understanding of the technicalities.

**Installation:**

WordCloud can be installed via following command:

pip install wordcloud

## Chapter 4

## Implementation



## 4.1 IMPORT LIBRARIES

## In this project we have used python libraries like numpy, pandas, matplotlib and wordcloud. All the libraries are imported at once by following commands:

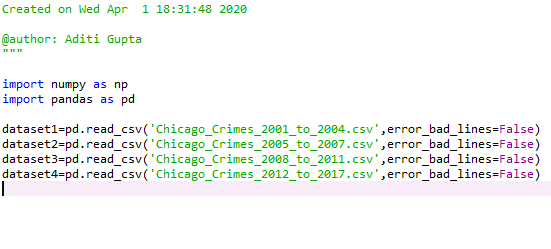
## C:\Users\Krishna Gupta\OneDrive\Desktop\mini project\mini1.PNG

## Fig: 4.1.1

## 4.2 IMPORTING DATASET

## The dataset used in this project is divided into four datasets, dataset1 is the crime record of Chicago city from year 2001-2004, dataset2 consisting of records from year 2005-2007, dataset3 from year 2008-2011 and dataset4 from year 2012-2017.

## The datasets are imported as pandas data frame as follows:



**Fig: 4.2.1**

**About the dataset:**

In this project we have used the Chicago Crime dataset (between the years 2001–2017), the dataset consist of 23 columns and 5 million rows. The date set is divided into four smaller dataset divided on the basis of time from year  2001-2004, 2005-2007, 2008-2011, 2012-2017. The 22 columns are identified as follows:

ID - Unique identifier for the record.

Case Number - The Chicago Police Department Records Division Number  which is unique to an incident.

Date - Date when the incident occurred.

Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.

IUCR - The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description.

Primary Type - The primary description of the IUCR code.

Description - The secondary description of the IUCR code, a subcategory of the primary description.

Location Description - Description of the location where the incident occurred.

Arrest - Indicates whether an arrest was made.

Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

Beat - Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.

District - Indicates the police district where the incident occurred.

Ward - The ward (City Council district) where the incident occurred.

Community Area - Indicates the community area where the incident occurred.

FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).

X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

Y Coordinate - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

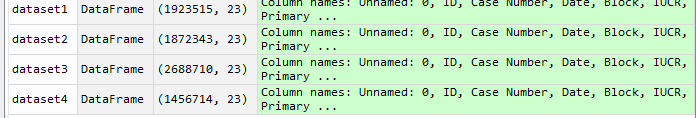
Year - Year the incident occurred.

Updated On - Date and time the record was last updated.

Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

Location - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.



## Fig: 4.2.2

## 4.3 DATA CLEANING AND PERPARATION

## 

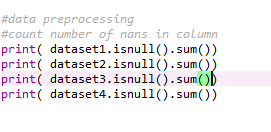
## Fig: 4.2.3

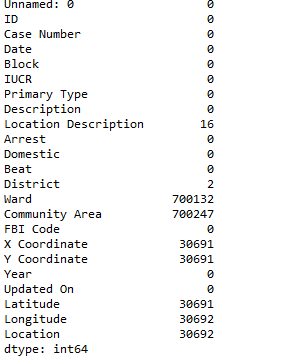
## 

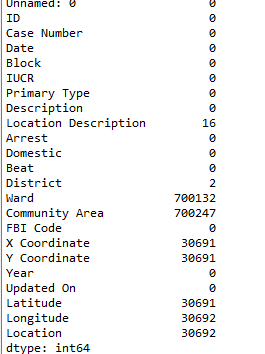
## Fig: 4.2.4

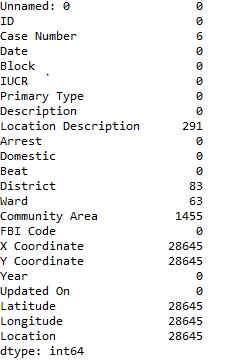
It can be observed that huge of amount of information is lost while removing rows with null value so in place of removal.

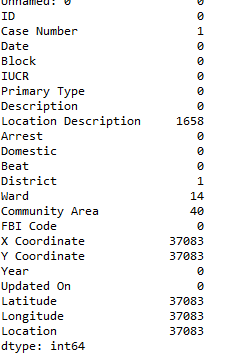
Determining the count of null and missing values in columns in all four datasets:

**Fig: 4.2.5**

1. 

**2.**

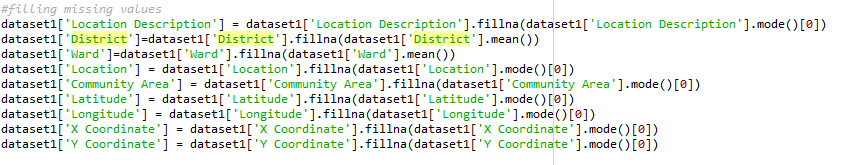
**3.**

**4.**

**Fig: 4.2.6**

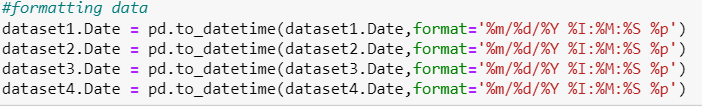
Filling missing values:

Since the count of null values in the columns Location Description, Ward, X Coordinate, Y Coordinate, Location, Community Area, Latitude and longitude is non zero therefore the values are filled with mean or mode of the corresponding column.



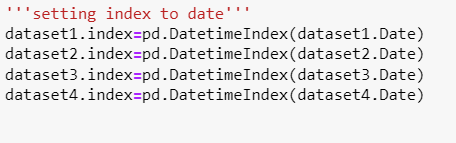
**Fig: 4.2.7**

**Data Formatting:**

To convert the given time format into pandas time format.

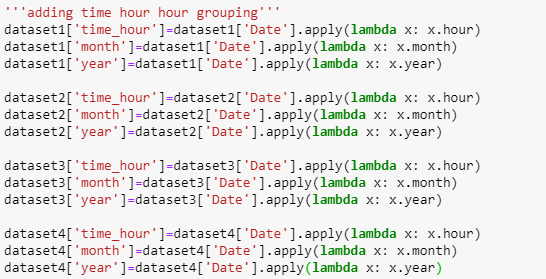
**Fig: 4.2.8**

Making the date as index in all dataset to have a better analysis on basis of time and for clear understanding of patterns(if any) in dataset:

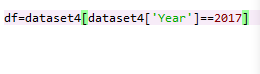


**Fig: 4.2.9**

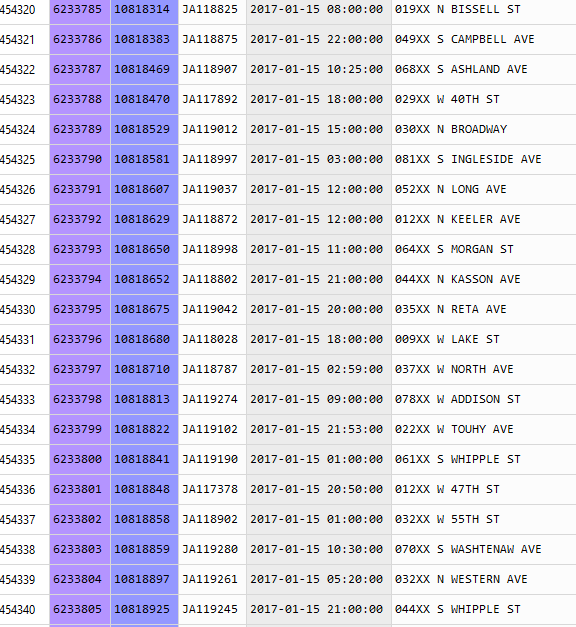
We made to make analysis on basis of crime rate per hour, crime trends in years and months, therefore for time series forecasting we need to add hour, month and year columns to our dataset.



**Fig: 4.2.10**

Selecting the dataset only for year 2017 to have understanding on the quality of data:

**Fig: 4.2.11**



**Fig: 4.2.12**

It can be seen that the year 2017 dataset only have the data of January month that will lead to wrong interpretation since a month’s data cannot determine the trend for a complete year

Therefore removing the record for the year 2017 from dataset 4

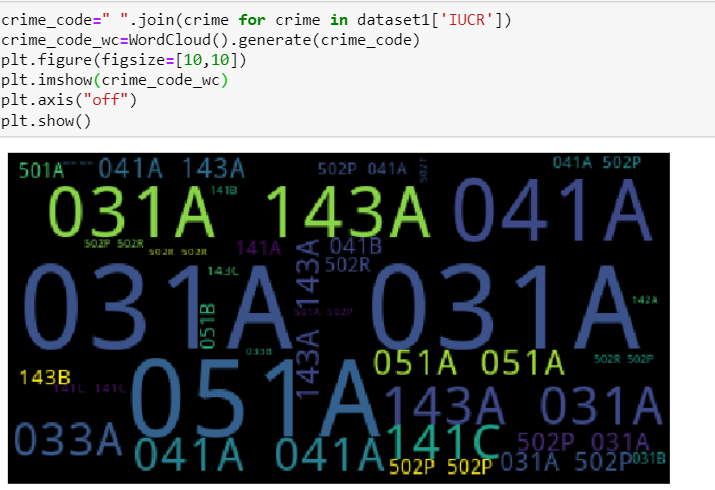


**Fig: 4.2.13**

## 4.4 DATA EXPLORATION AND VISUALIZATION

1-The most frequent crime code i.e, the visualizing most likely cime:

**Dataset 1:**



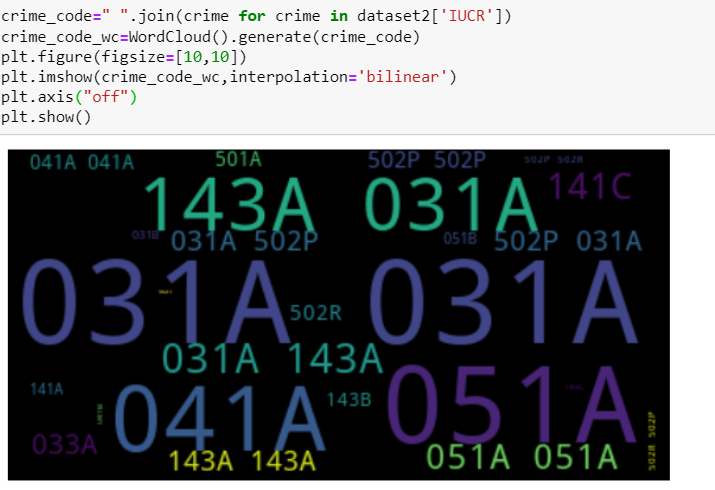
**Fig:4.3.1**

Here it can be seen that the crime with codes 031A, 143A, 051A are frequent. Crime associated with these codes are:

051A: Assault - Aggravated Handgun

031A: Robbery - Armed Handgun

143A: Weapons Violation Unlawful poss of hand gun

**Dataset 2:**

**Fig: 4.3.2**

Here also the most frequent crime codes are 143A, 031A, 041A, 051A

Associated crimes are:

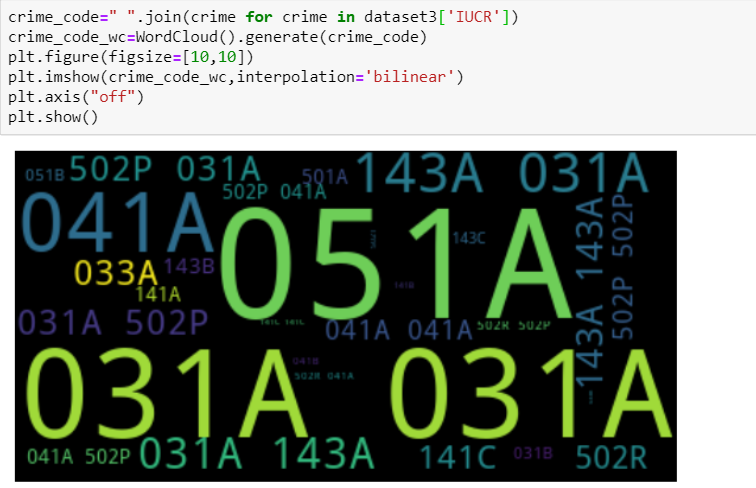
|  |  |  |
| --- | --- | --- |
| 143A | WEAPONS VIOLATION | UNLAWFUL POSS OF HANDGUN |

|  |  |  |
| --- | --- | --- |
| 051A | ASSAULT | AGGRAVATED: HANDGUN |

|  |  |  |
| --- | --- | --- |
| 031A | ROBBERY | ARMED: HANDGUN |

|  |  |  |
| --- | --- | --- |
| 041A | BATTERY | AGGRAVATED: HANDGUN |

**Fig: 4.3.3**

**Dataset 3:**

**Fig: 4.3.4**

Here the most frequent crime code are:

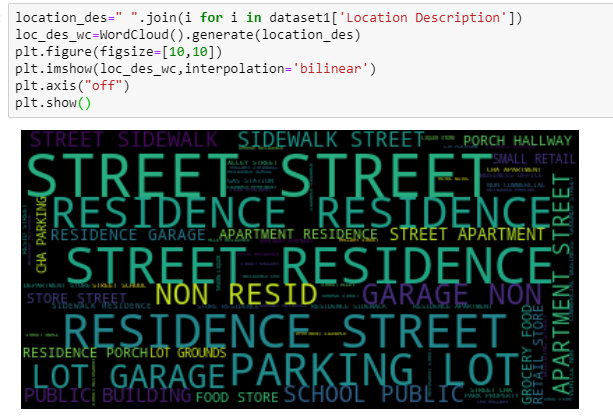
051A, 031A, 041A, 143A

## Dataset 4:

## Fig: 4.3.5

2- Locations where crimes are more likely to happen

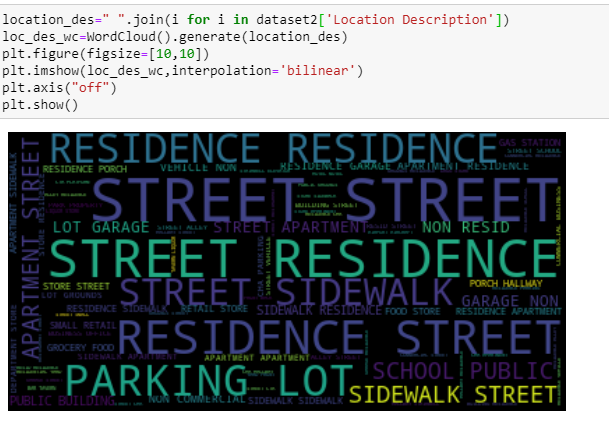
(According to location description)

**Dataset 1:**

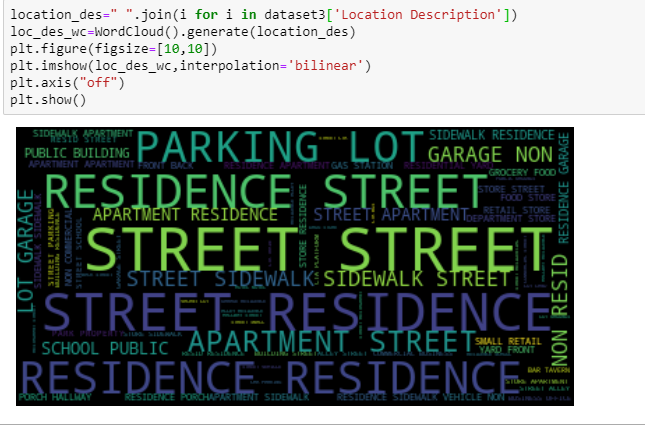
## Fig: 4.3.6

The more likely location where crime happens can be easily seen in the plot, the locations are:

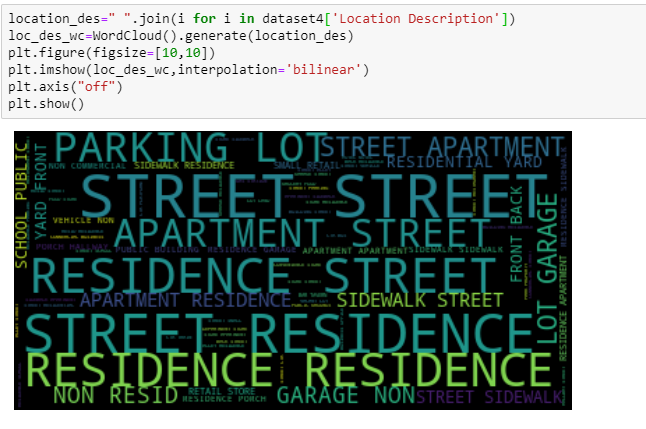
Street, Residences, Garage, Parking lot, school, public places etc.

Dataset2:

**Fig: 4.3.7**

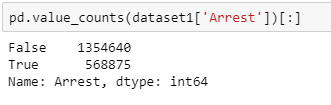
**Dataset3:**

**Fig: 4.3.8**

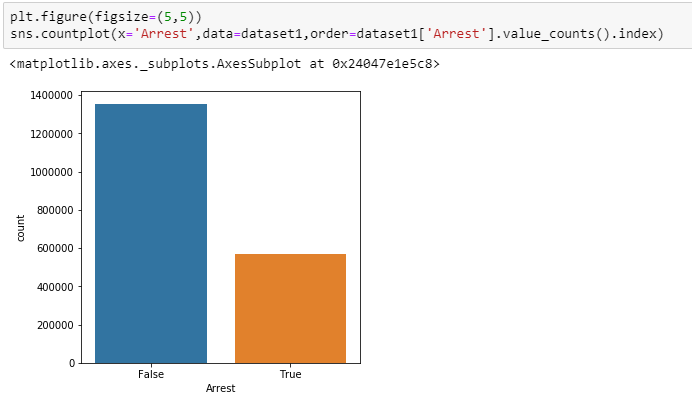
**Dataset4:**

**Fig: 4.3.9**

3-Total number of arrest made which to determine the capability of Chicago police to tackle with crime:



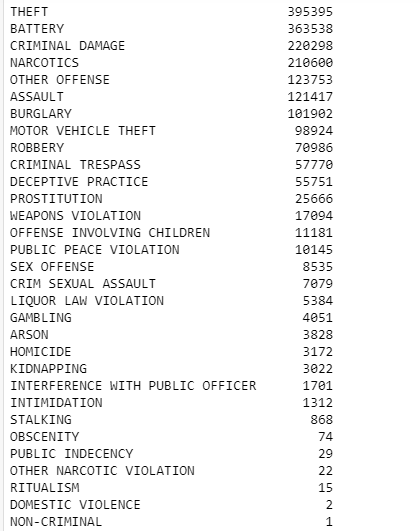
**Fig: 4.4.10**



**Fig: 4.4.11**

The count plot shows that out of 1,923,515 cases 568875 arrest were made for the time period between 2004-2007 which is roughly equal to 30% of cases.

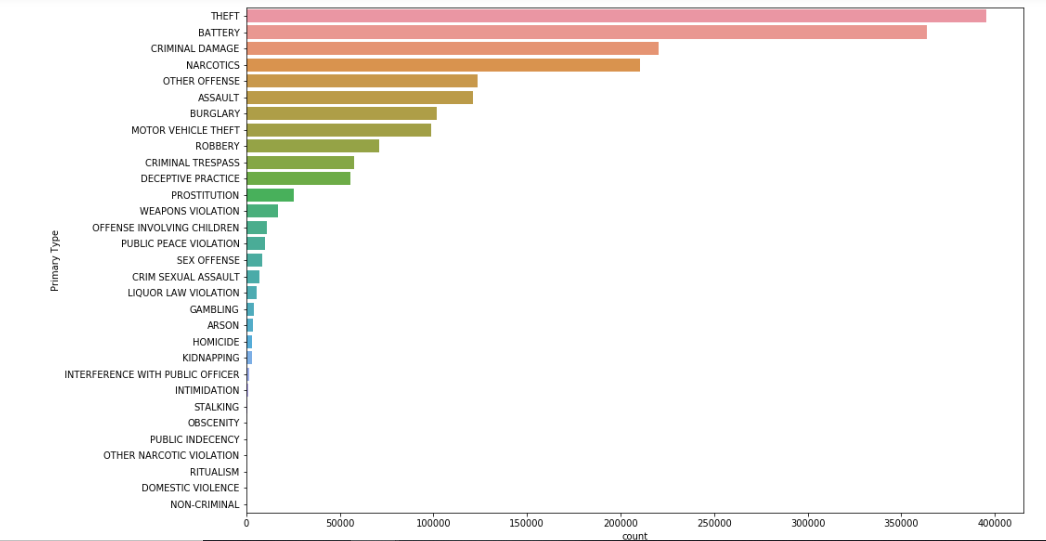
4- Frequent Crimes

Dataset1:

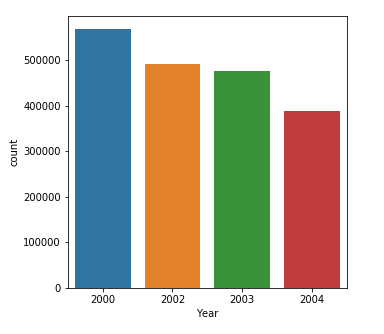
**Fig: 4.3.12**

It can be seen that crime types like public indecency, obscenity, other narcotic violations have very low count therefore the data can be further cleaned and these crime types with little impact can be removed whereas the crimes like theft, battery, narcotics, criminal damage, burglary are very dominant in Chicago.

For better understanding we have plotted a bar plot:

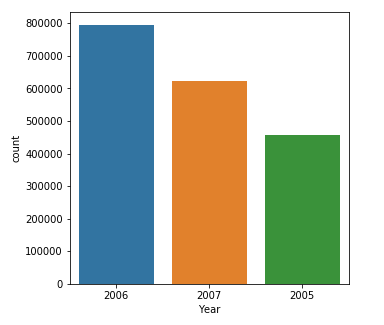


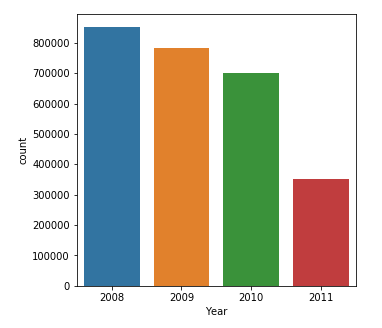
**Fig: 4.3.13**

5- Yearly count and comparison for crime

**Fig: 4.3.14**

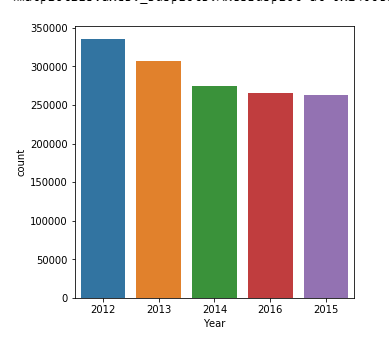
In the period of four years 2000-2004, Crime rate is highest in the year 2000 (>500000) while least in year 2004(<400000)





**Fig: 4.3.15 Fig: 4.3.16**

Between the years 2005-2007, the crime rate is highest for year 2006 and lowest for year 2005



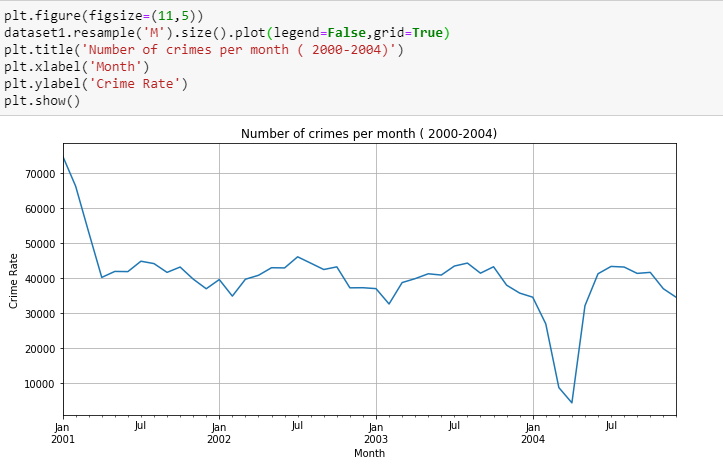
**Fig: 4.4.2**

## Fig: 4.3.16

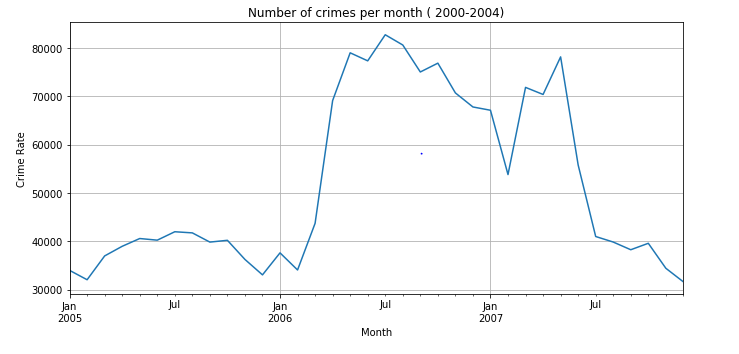
**Fig: 4.3.17**

In the period of five years between 2012-2016 the crime rate is highest for year 2012 and lowest in year 2015

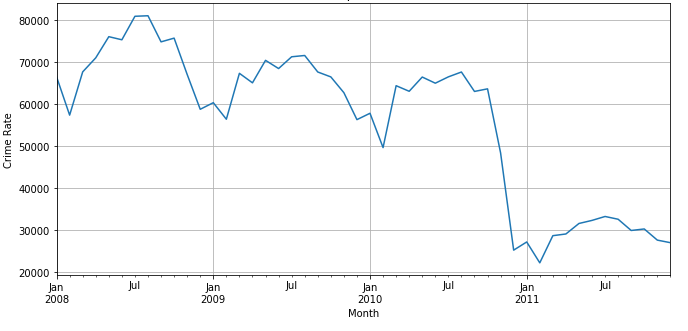
6-Monthly Analysis

For dataset 1(2000-2004):

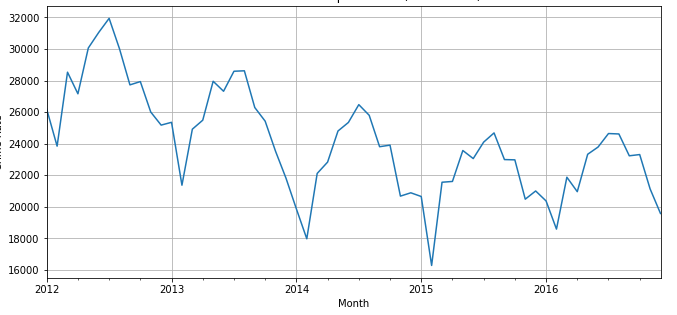
**Fig: 4.3.18**

For dataset 2(2005-2007):

**Fig: 4.3.19**

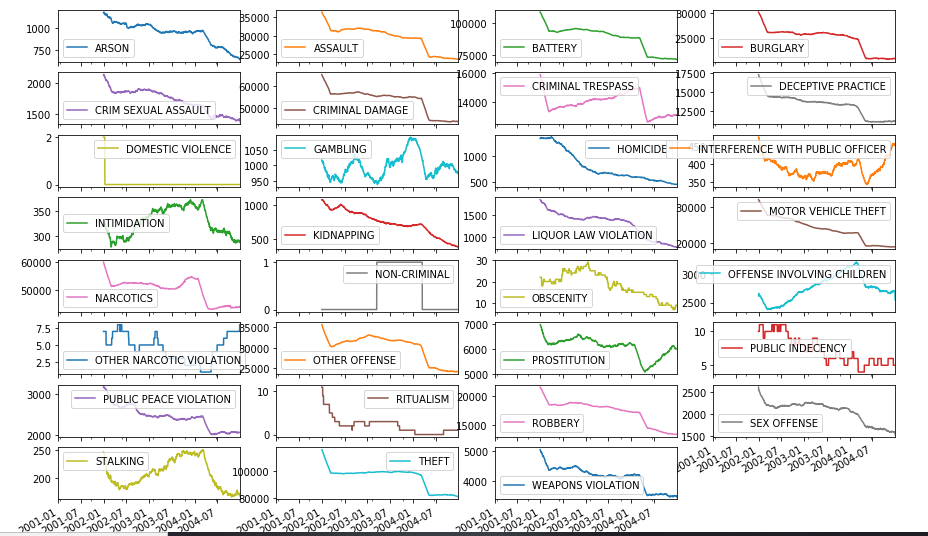
For dataset 3(2008-2011):

**Fig: 4.3.20**

For dataset 4(20012-2017):

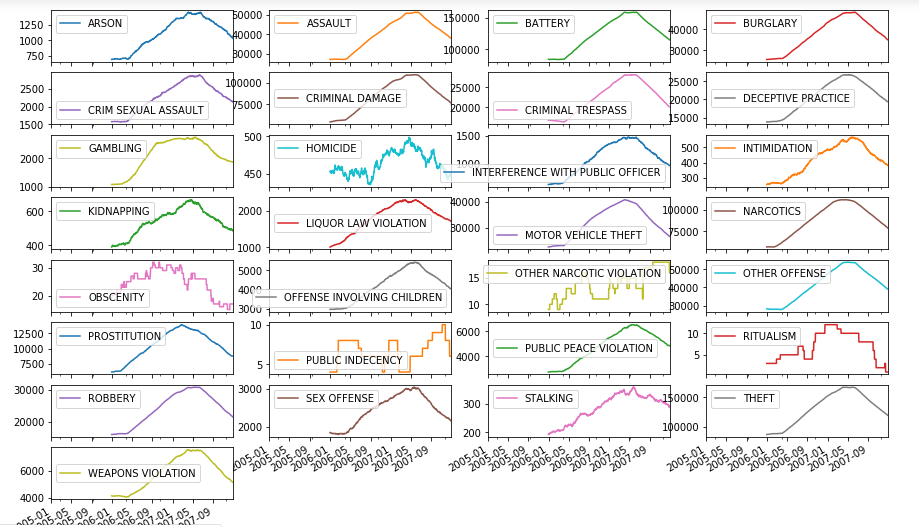
## Fig: 4.3.21

7-Individual Crime type trend

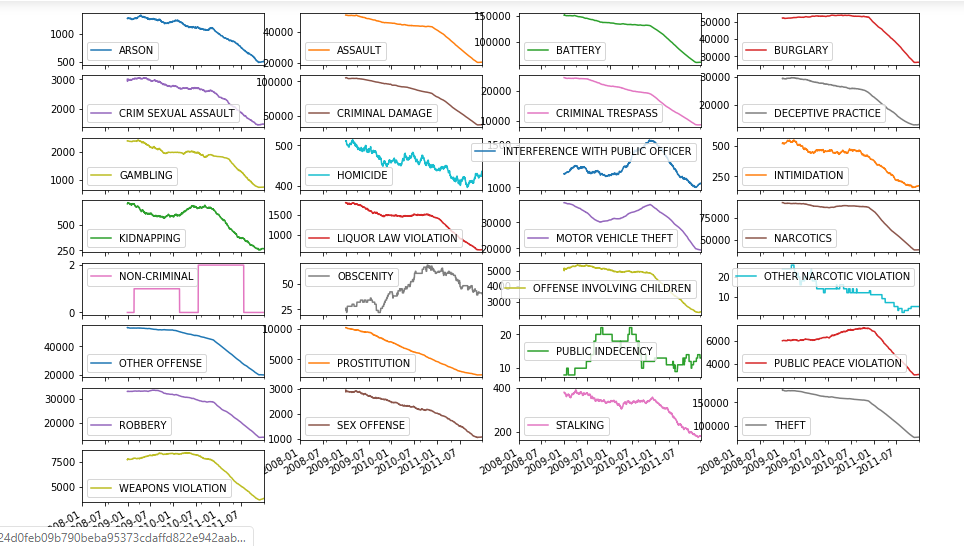
D**ataset 1:**

**Fig: 4.3.22**

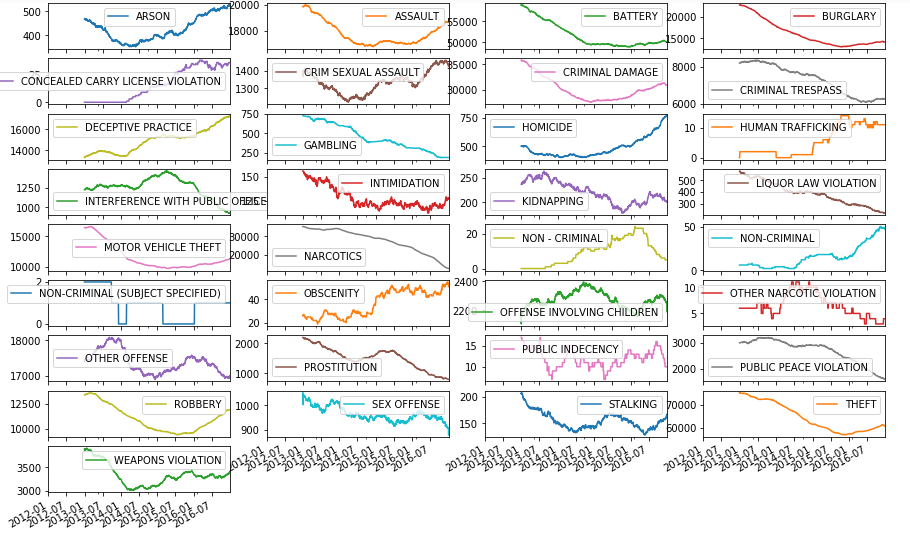
**Dataset 2:**



**Fig: 4.3.23**

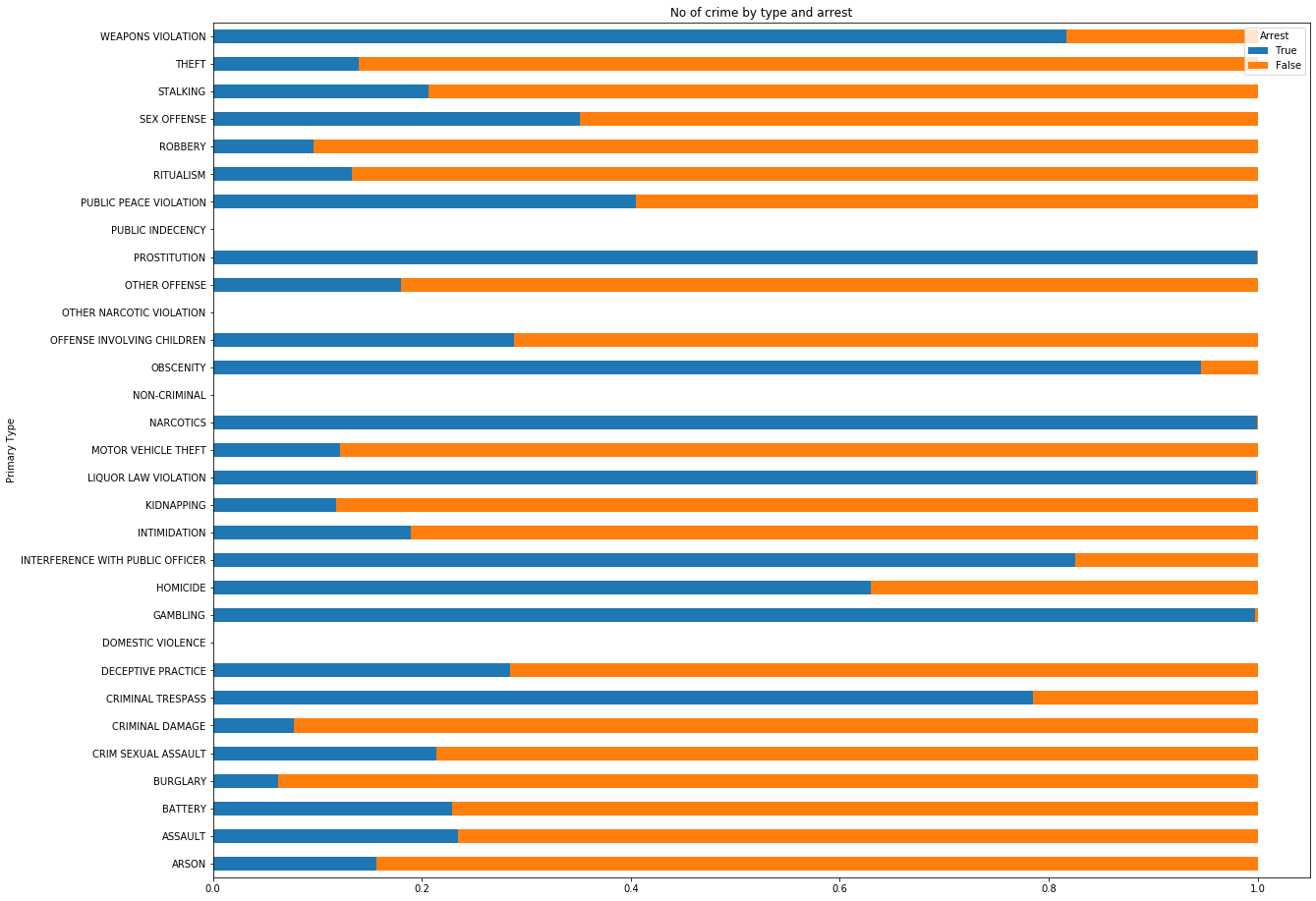
**Dataset 3:**

**Fig: 4.3.24**

**Dataset 4:**

**Fig: 4.3.25**

8-Crime Type vs Arrest



## Fig: 4.3.26

## Chapter 5

## Conclusion

In this project we analyze the Chicago Crime dataset (between the years 2001–2017), which is one of the richest open source data in this area, to get a better understanding about the security status of this city. The result of our project shows that only 28.33 percent of the criminals that were reported, were arrested and the amount of crime reported in recent years had considerably decreased in comparison to a number of crimes between the years 2008 to 2010 which it shows us that the security of Chicago has increased during these years. In this project, we also analyzed the relationship between different types of crimes and the location they happened. The results showed that, the most common crimes in Chicago which are: theft, battery, criminal damage, and narcotics, made up of 65.7 percent of the crimes in Chicago. Also, our results show these crimes usually happen in apartment, residence, sidewalks, and streets which are very common places. In our research, we also analyzed the statistics of a few specific crimes: theft, homicide, and sexual harassment. Our results display that there is a steady decline in the number of homicides during the years after 2001 and 2016, with the average number of 482 homicide per year. However, there is a sharp increase in the number of homicides in 2016. Similarly, there is an increase in the number of sexual harassment and theft in the year 2016.

## Chapter 6

## References

[1]Currie32, 2017. Crimes in Chicago. *RSNA Pneumonia Detection Challenge | Kaggle*. Available at: <https://www.kaggle.com/currie32/crimes-in-chicago> [Accessed December 14, 2018].

[2]Anon, Wrangler is an interactive tool for data cleaning and transformation. Data Wrangler. Available at: <http://vis.stanford.edu/wrangler/> [Accessed December 7, 2018].

[3]Anon, Tutorials. *pandas: powerful Python data analysis toolkit — pandas 0.23.4 documentation*. Available at: <http://pandas.pydata.org/pandas-docs/version/0.15/tutorials.html> [Accessed December 14, 2018].

[4]Anon, Welcome! *openrefine.github.com*. Available at: <http://openrefine.org/> [Accessed December 14, 2018].

[5]Stevestein, SQL Server Management Studio (SSMS). *Microsoft Docs*. Available at: <https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-2017> [Accessed December 14, 2018].

[6]Anon, MapReduce Tutorial. *Apache Hadoop*. Available at: <https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html> [Accessed December 14, 2018].

[7] Anon, Spark Overview. *Apache Spark™ — Unified Analytics Engine for Big Data*. Available at: <https://spark.apache.org/docs/latest/> [Accessed December 14, 2018].

[8]Anon, *Google*. Available at: <https://developers.google.com/chart/interactive/docs/spreadsheets> [Accessed December 14, 2018].