**Analysis of Chicago Crime data from 2014 - 2018**

**Abstract (max 200 words):**

The criminal justice system has become one the key area in any state due to increase in number of crimes over the years. However, being cautious and enhancing safety is not a simple task anymore. We should use modern technologies and data science techniques to tackle this challenging problem. The police departments have documented criminal records over the years, which can be used as a data source for analysis. Therefore, applying statistical and analytics techniques on these data can extract valuable information which can be used to increase the safety of the society and lower the crime rates in the world.

In this project, I analysed the Chicago crime data between the years 2014-2018, to have a better understanding of the security status of the city. The result of my analysis showed that the frequency of crimes reported in the city has decreased over the years. Furthermore, the analysis revealed that most common crimes in Chicago are battery, assault and burglary, which usually happened in street, residence and apartments. My results show that there is significant decrease in narcotics after 2015, however, theft was at the peak in the year 2016.

**What is the question or story you are trying to tell?**

To have an overview of the security status in Chicago from the years 2014 to 2018, I have drafted the below questions, which I answered after analysing the data. Below is the list of questions:

1. How has the frequency of the overall crimes changed over the years?
2. How many numbers of criminals were arrested as compared to the total crimes committed?
3. Which of the crimes were most frequently committed in Chicago?
4. Which were the common locations where the frequent crimes have occurred?
5. How has the frequency of certain crimes changed over the years?
6. Were there any trends in the location where the crimes have occurred?
7. Were there any trends in the crime being committed?

To answer the above questions, I have utilised the knowledge gathered as part of CA682 (Data Management and Visualization) which includes data gathering, processing, analysing and presenting. How I utilised these four parts is described below in the report.

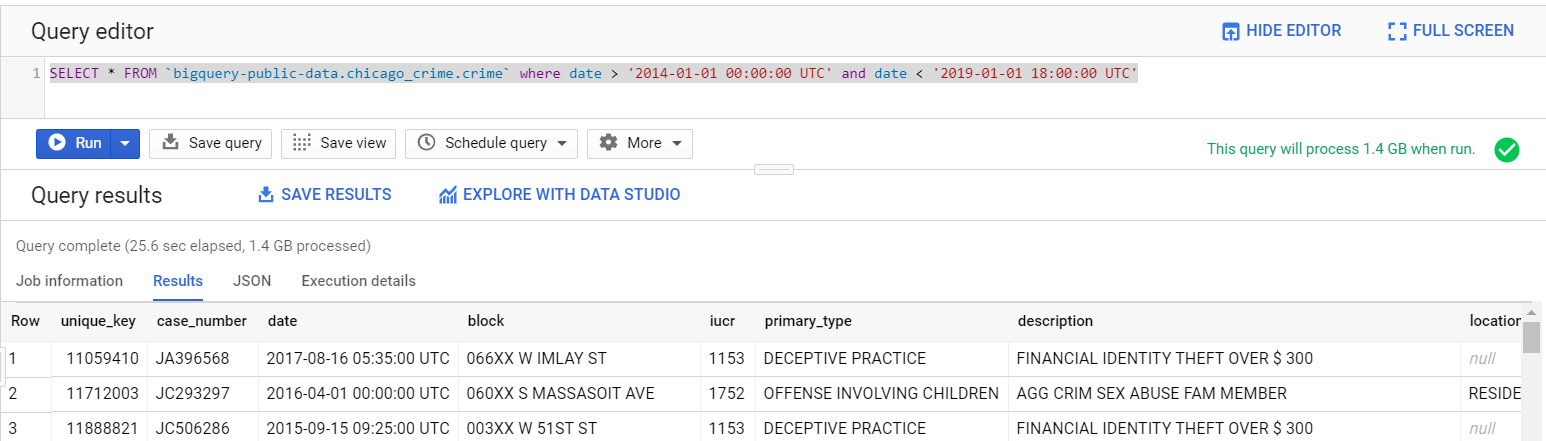
**1. Dataset:**

***Where/how did you retrieve it or them***

The data is gathered from Google Cloud Platform dataset by using big query to query the data from table. The query is as below for extracting the data from the dataset.[1]

**Query: -** SELECT \* FROM `bigquery-public-data.chicago\_crime.crime` where date > '2014-01-01 00:00:00 UTC' and date < '2019-01-01 18:00:00 UTC'

According to the information provided “This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago [1]. In general, the data included information such as datetime of the crime, the crime block, type of crime, location description, whether there was an arrest, district, FBI code, and location coordinates. The data table consists of 22 columns and 1748534 records, which includes categorical, numerical, Boolean, datetime and location values. The name of each column from left to right are unique\_key, case\_number, date, block, iucr, primary\_type, description, location\_description, arrest, domestic, beat, district, ward, community, fbi\_code, x\_coordinate, y\_coordinate, year, updated\_on, latitude, longitude, location. (Screenshot of the data as below)



**2. Data Exploration, Processing, Cleaning and/or Integration [½ page]**

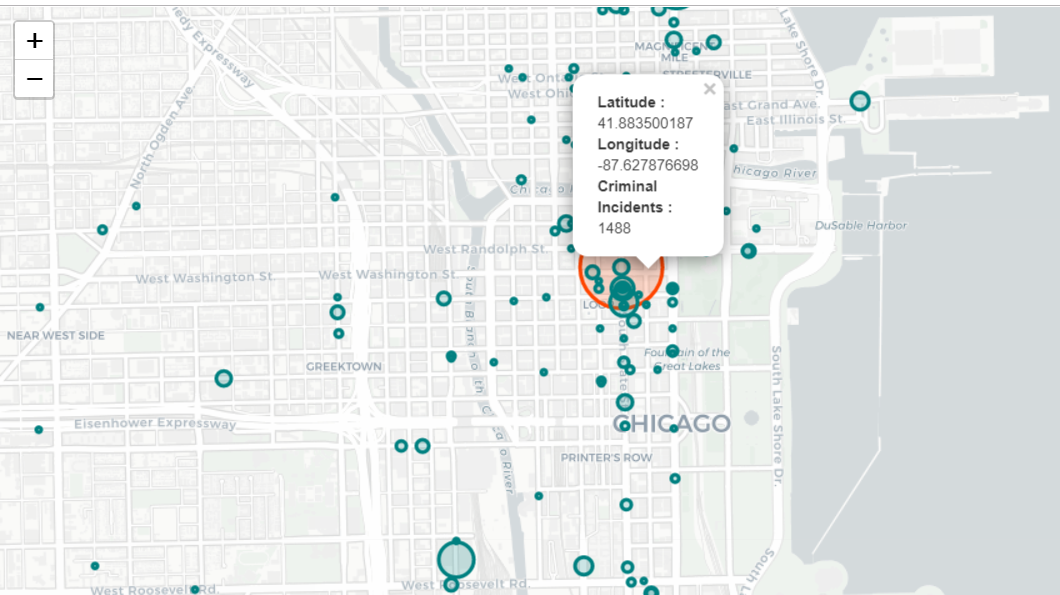
As data science community spend 70% of their time on data exploration and pre-processing, therefore, many tools and software have been developed to reduce the time and work efforts. Since, running and start using Google OpenRefine is extremely convenient (as taught in CA682 Lab -4 exercise), I have used this tool to gain insight into the Chicago crimes data. It helped me to find out that the dataset needs to be cleaned before using it for the further operations. The following are the cleaning operations performed on the dataset.

* Removed the duplicate data values
* Removed the null and NA values
* Removed the insignificant values from the data
* Removed the unwanted characters

After applying the filter operations, the size of the data reduced to 167MB and 17068436 records. After making sure that the data is cleaned, I exported the file in csv format and loaded it in the Jupyter notebook for further processing and visualization. In Jupyter, I performed statistical operations and analysed the co-relations between the attributes, which I used for displaying the data visually.

**3. Visualisation**

1. **Graph 1 (Interactive Map): -** Criminal incidents in the Chicago City



**Tools: -**

* Jupyter Notebook

**Chart Type: -** Interactive folium Map

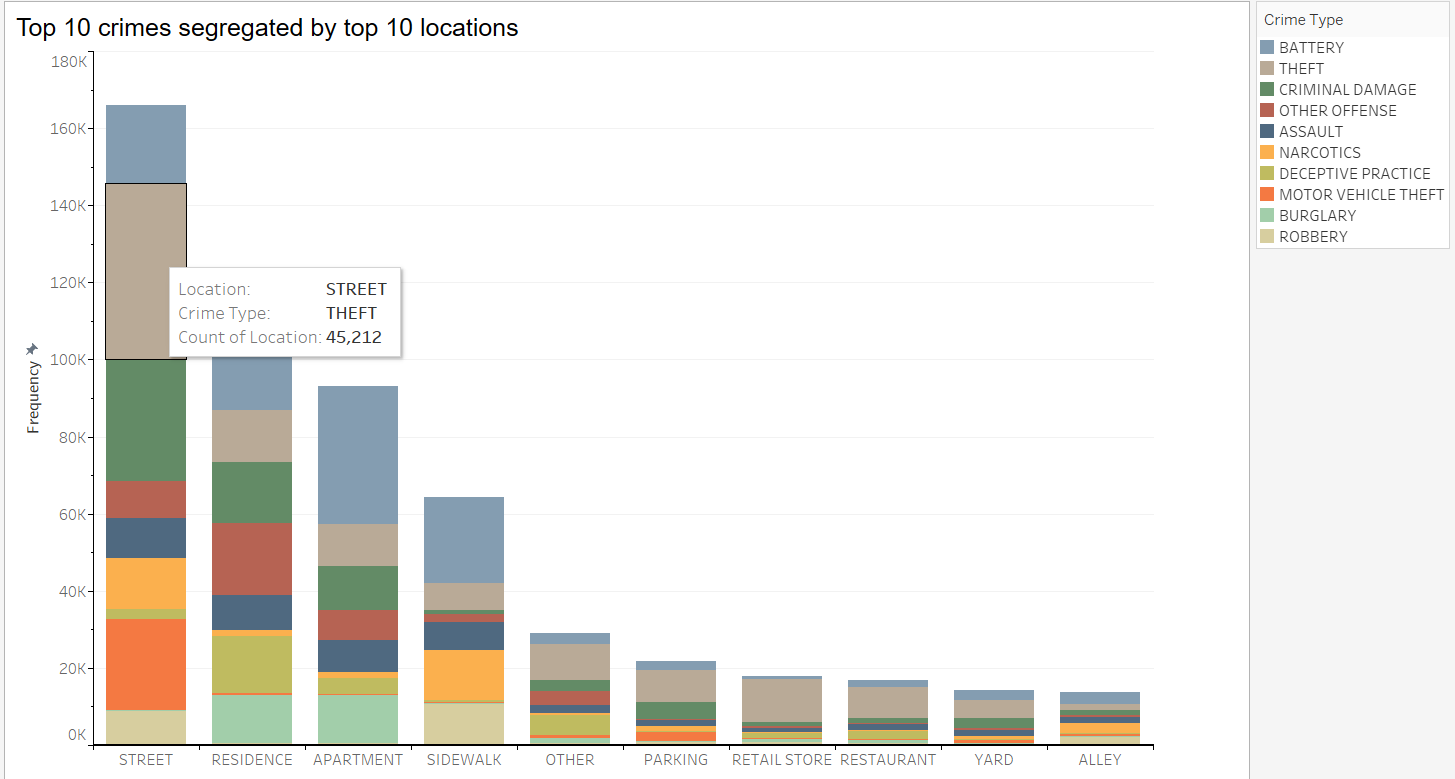
**Libraries: - Below** are the python libraries used:

* **Numpy** – library for python programming
* **Pandas** – library for data manipulation and analysis
* **Os** – Library to make Jupyter notebook compatible with the system
* **Matplotlib**.pyplot – Library for plotting 2D figures
* **Seaborn** – Library based on matplot to for plotting interactive figures
* **Folium** – Library for creating leaflet maps

**Choice of graph: -**

As the data set consists of the location coordinates (latitude and longitudes) of the crimes, I was able to get the count of crimes in the data along with the location. I utilised this information to create an interactive leaflet map using python’s folium library which can be zoomed in and out. The map can even be dragged left and right to display the information. I have used bubbles as the marker. Also, the size of the bubbles increases depending on the frequency of crimes at a particular location. The map style chosen” cartodbpositron” and the locations are displayed in the form of bubbles. Furthermore, I have made “latitude”, “longitude” and “Criminal Incidents” labels in bold using html tags to separate it from the values.

**Graph 2: -** Top 10 crimes segregated by top 10 locations



**Tools Used: -**

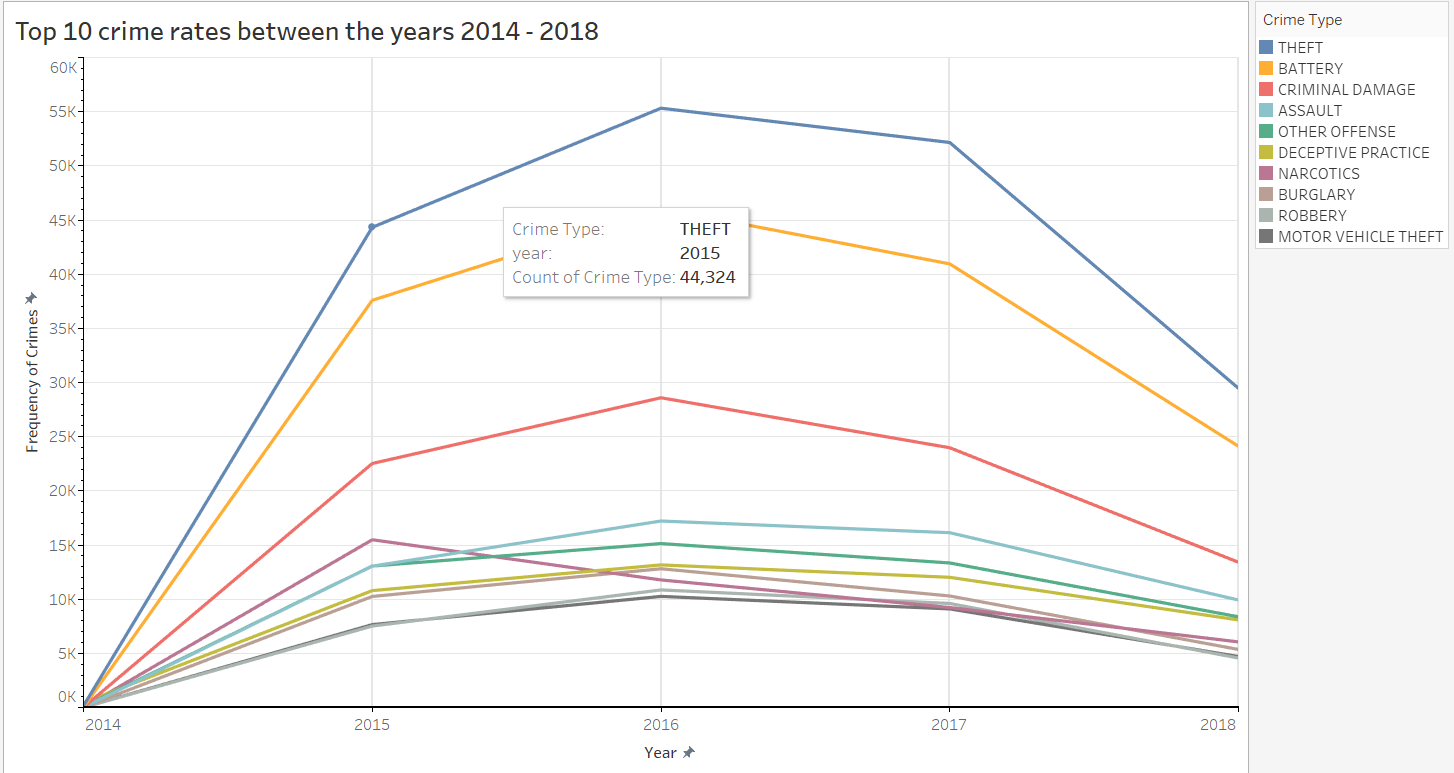
* Tableau

**Chart Type: -** Stacked Bar graph

**Choice of graph: -**

The data consists of columns named as primary\_type and location\_description, describing the different types of crimes happened in different locations of the city respectively. Therefore, I decided to display the data as stacked bar with top 10 locations as the column values and frequency rows. Since, both the data columns have categorical values, I have chosen tableau to be a better choice. In tableau it is convenient to count the two categorical values and corelate between them. The overall bar shows the type of location where the crime occurred and the frequency of different crimes occurred at that particular location. The stacked bar sections are labelled as “Location”, “Crime type” and “Count of the Location” displaying the crime type by location. The colour palette used on “primary\_type” field is “Miller Stone” as the colours available in this palette differentiate the individual bar from the others accurately. Legend is placed in the upper right corner with label as “Crime Type” to categorize the crime type by colour filter. The axis of the graph is highlighted in black (including major and minor tick marks) to make reading and understanding easier. Furthermore, the gridline is made visible in the graph to compare the values between different bars.

**Graph 3: -** Top 10 crimes rates between the years 2014 – 2018



**Tools Used: -**

* Tableau

**Chart Type: -** Multiline graph

**Choice of graph: -**

The data have different crimes happened between the years 2014 – 2018 (with crime value as type categorical and a year column). I have used tableau to make an interactive multiline graph displaying the frequency of top 10 crimes between the year range from 2014 - 2018. The colour palette used is “superfishel stone” as the colours of the line can be easily distinguished from the others. Legend labelled as “Crime Type” is placed in the upper right section for the colour referencing in the graph. The axis and tick marks of the graph are highlighted in black to make is easier to read and gridlines are made visible to check the coordinates of the value in the graph at any point of time.

4. **Questions Answered using the analysis and visualization: -**

1. **How has the frequency of the overall crimes changed over the years?**

Analysis shows that the frequency of the crimes has decreased over the years.

1. **How many numbers of criminals were arrested as compared to the total crimes committed?**

21.24% of the criminals were arrested in the years between 2014 – 2018.

1. **Which of the crimes were most frequently committed in Chicago?**

Most frequent crimes types in Chicago were assault, battery, burglary, criminal damage, deceptive practice, motor vehicle theft, narcotics, robbery, and theft

1. **Which were the common locations where the frequent crimes have occurred?**

Most common crime locations as per the analysis were street, residence, apartments, sidewalk, retail store, restaurant, yard, and alley.

1. **How has the frequency of certain crimes changed over the years?**

I chose theft to see if its increasing or decreasing over the years. As per the analysis theft is the most occurred crime in Chicago and was at its peak in the year 2016.

1. **Were there any trends in the location where the crimes have occurred?**

As per the analysis street was the most common location for the crime incidents.

1. **Were there any trends in the crime being committed?**

Theft, Battery and criminal damage remained the top three crimes between the years 2014 – 2018.

**4. Conclusion**

This data visualization project gives a scientific view about the security status and crime rates in Chicago city between the years 2014 – 2018. According to the analysis and visualization result, most frequent crime and most frequent locations can be easily viewed. Most frequent crime locations were street, residence, apartments, sidewalk, retail store, restaurant, yard, and alley. The most frequent crimes occurred at these locations were assault, battery, burglary, criminal damage, deceptive practice, motor vehicle theft, narcotics, robbery, and theft. Even though Chicago crime rate was high over the years, the arrest rate was just 21.24%, making us believe that there were flaws in the Chicago security department. Therefore, by analysing the data efficiently can provide valuable information, which can be used to increase the efficiency of the security systems.

**Were there effects or functionality that you were technically unable to achieve?**

I was not able to plot an interactive stacked bar graph using python libraries between two columns having categorical values. Therefore, I switched to tableau to create an interactive stacked bar graph between locations and crimes.

**References:**

1. Google Cloud Platform, Chicago Crime Data ,Available at: <https://console.cloud.google.com/marketplace/details/city-of-chicago-public-data/chicago-crime> , [Accessed December 8, 2019].
2. Dominodatalab, Creating interactive crime maps with Folium, Available at:

<https://blog.dominodatalab.com/creating-interactive-crime-maps-with-folium/>

[Accessed December 10, 2019].

1. Yuanjieli, EDA of Crime in Chicago (2005 - 2016), Available at:

<https://www.kaggle.com/yuanjieli/chicago-crime-data-visualization-exercise>

[Accessed December 10, 2019].