**Project Document**

**Group: Shivit & Devansh**

**I Table of Contents**

**1 PROBLEM**

**2 Data Pre-Processing: Data Extraction**

**2.1 Data Exploration**

**2.1.1 Size of data**

**2.2 Data Extraction**

**3 Analysis and Visualisation**

**4 Kmeans Clusters**

**4.1 Algorithm**

**4.2 K-Means Clustering using sklearn**

**4.2.1 STEPS**

**5 CONCLUSION**

**6 LITERATURE on CHICAGO CRIME**

**7 REFERENCES**

**Crime in Chicago**

Crime is an inextricable part of our lives in this world. We hear about them every day, and some of us have been interested in at least one of them at some point in our lives. Being vigilant and improving safety is no longer a clear order. To act more wisely against this challenge, we need to use modern technology and data science techniques. The police department has a large number of records and documentation that have been accumulated over time and can be used as a valuable source of data for data analytics activities. Applying analytical tasks to these data yields useful knowledge that can be used to improve our society's protection and reduce crime.

In this project, we examine the Chicago Crime dataset (from 2012 to 2018), which is one of the most comprehensive open-access datasets in this field, to gain a better understanding of the city's security situation. Just 28.33 percent of the offenders identified were apprehended, according to the findings of our project. and the number of crimes reported in recent years has declined significantly as compared to the number of crimes reported between 20012 and 2018. As a result, we can see that Chicago's safety has improved over the years.

As a result, we can see that Chicago's security has improved over the years. We also looked at the relationship between various forms of crimes and where they occurred in this project. The most common crimes in Chicago are theft, battery, criminal harm, and drug, which account for 65.7 percent of all crimes in the city.

**PROBLEM**

Not all crime is the same. Depending on the location and time, some forms of crimes are more likely to occur than others. We'll look at how crimes vary in different locations and at different times in this review.

To gain a better understanding of Chicago's security situation, we proposed a set of questions to which we reported during our analysis

How has the number of different crimes in Chicago increased over time?

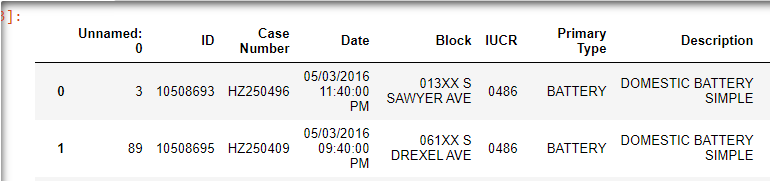
How has the number of arrests in Chicago increased over time concerning the number of crimes?

* Is there a pattern to the crimes that are being committed?
* What are the most common types of crimes?
* What are the most common places where these crimes occur?
* Are there any high-crime areas for specific offences (e.g., sexual offence)?

**2.Data Pre-Processing: Data Extraction**

**2.1 Data Exploration**

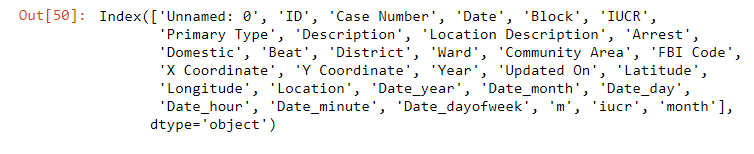
We have extracted our data directly from Chicago Police’s Department’s CLEAR system. From 2012 to 2018, this dataset represents recorded crime events in the city of Chicago.In general, the dataset contains data such as the date and time of the crime, the block where the crime occurred, the type of crime, the location description, whether an arrest was made, and the coordinates of the crime scene. In the next section, we'll go over the basic features of our results.



**2.1.1 Size of Data**

The data was a 350 MB file with 1456714 rows and 23 features/columns.

Features include :



**2.2 Data Extraction**

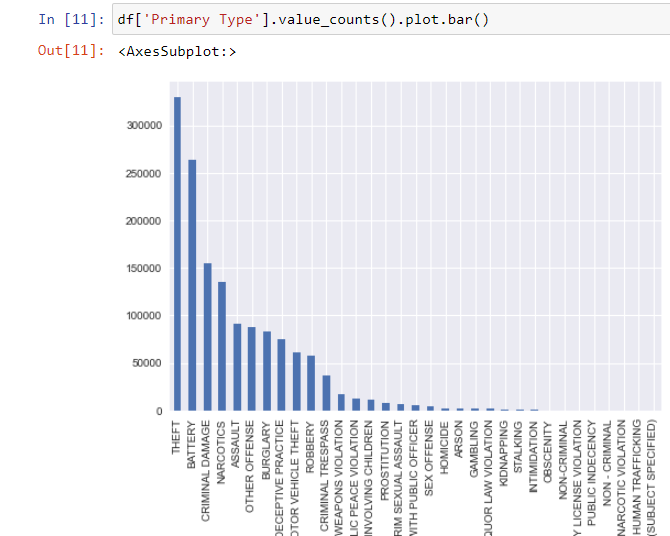
We extracted selected features for our analysis using the Pandas library.

For this analysis, we dropped Isnull() values.

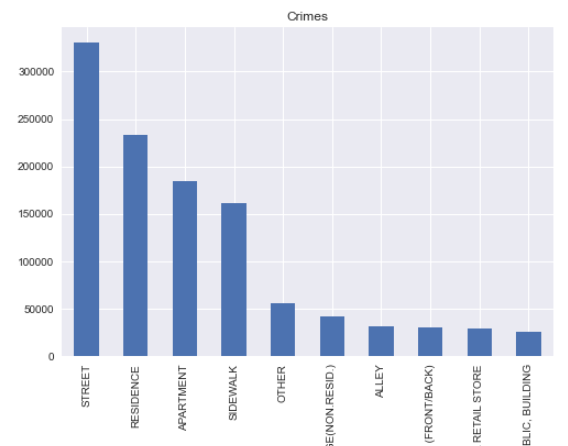
Not all the features were used as a part of the analysis.

**3 Analysis and Visualisation**:

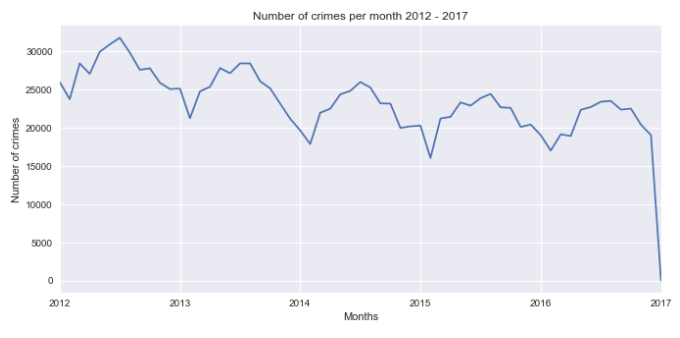
* We analysed the primary type of crime committed that lead to an arrest.
* Theft/robbery was the most committed crime amongst the accused, followed by stealing battery appliances in second.

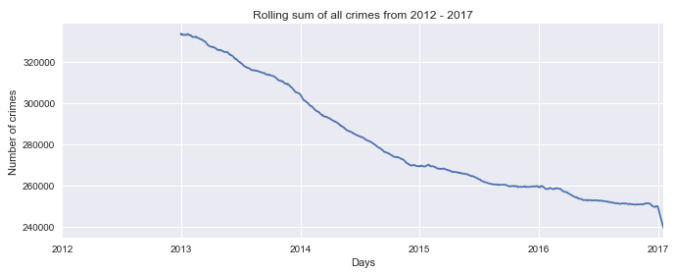


* Most of the crime depending upon the “Location Description” were committed on the streets.

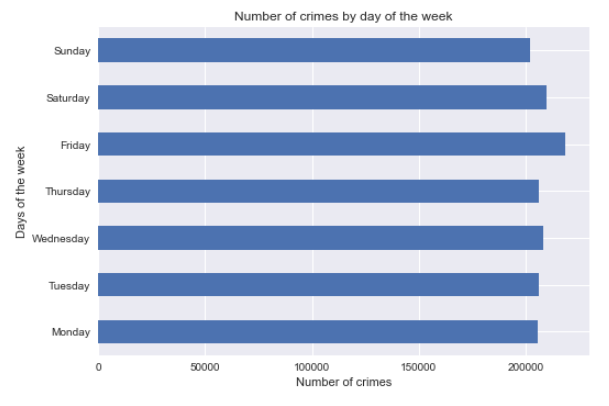


* Surprisingly “Residence” faced second most frequently committed crimes while least of all crimes committed in a Block and Building type environment.
* Now we explore the frequency of crimes committed each month and year from 2012-2017.

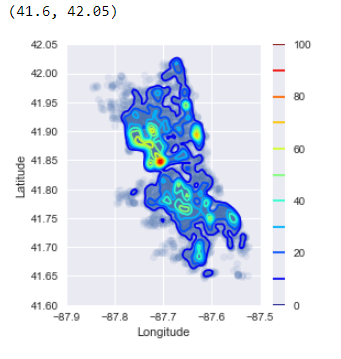


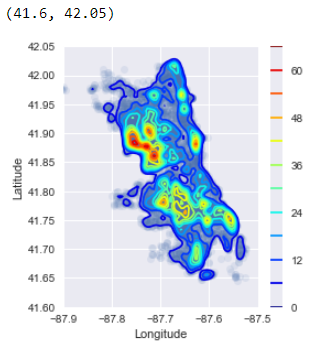


* Over several years, this graph reveals a consistent periodic trend in the crimes. I suppose this regular pattern is one of the reasons why crime is such a predictable occurrence. We see the line decreasing through the years at different rates. There was a decent decline in crimes from 2013-2015. Although the decline was almost not noticeable from 2015-2016. Then took a sudden dip in 2017 where the crime rate went at its lowest.
* Next, we'll look for is if there's a disparity in the number of crimes committed on different days of the week.



* No such difference was noticed. So we looked at crimes by month of the year.
* There was a peak in the crime rate during the summer season.
* A heat map on the map of the city shows where most of the crimes occurred in which region. based upon the type of crime. The below graphs are for the type- Liquor law violation and sex offenders.





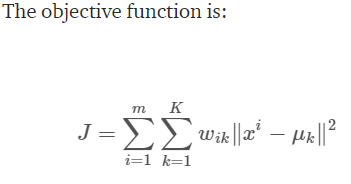
**4. Kmeans Clusters**

Clustering is a common exploratory data analysis technique for gaining a better understanding of the data's structure. Its role is to define subgroups in data such that data points within the same subgroup (cluster) are very similar while data points within different clusters are very different.

Clustering, unlike supervised learning, is an unsupervised learning approach since there is no ground truth to compare the clustering algorithm's output to the true labels to test its accuracy. We just want to look at the data's structure by grouping the data points into distinct subgroups.

**4.1 Algorithm**

The Kmeans algorithm is an iterative algorithm that attempts to partition a dataset into K distinct non-overlapping subgroups (clusters), each of which contains only one data point. It attempts to make intra-cluster data points as close as possible while keeping clusters as distinct (far) as possible. It assigns data points to clusters in such a way that the number of the squared distances between them and the cluster's centroid (arithmetic mean of all the data points in that cluster) is as small as possible. Within clusters, the less variance there is, the more homogeneous (similar) the data points are.

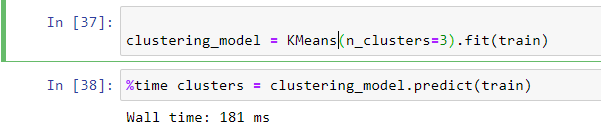


**4.2 K-Means Clustering using sklearn**

#Imports: from sklearn import preprocessing

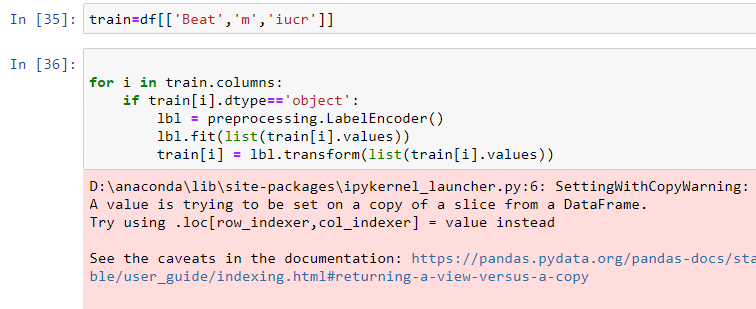
from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score



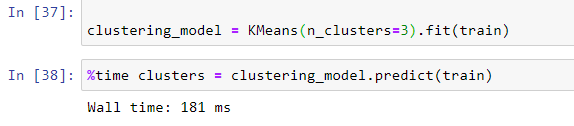
**4.2.1 STEPS:**

As we saw before the crime patterns were different by location and type. Therefore, we choose three means “date hour” and “beat”, i.e the smallest police geographic area, for our analysis in Kmean clusters.

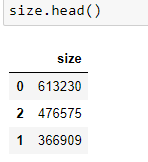


We pre-process the date using labelEncoder to convert categorical data to the numeric type.

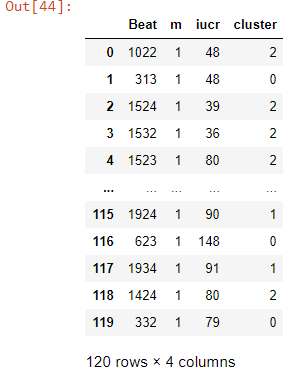
Then we train our columns:



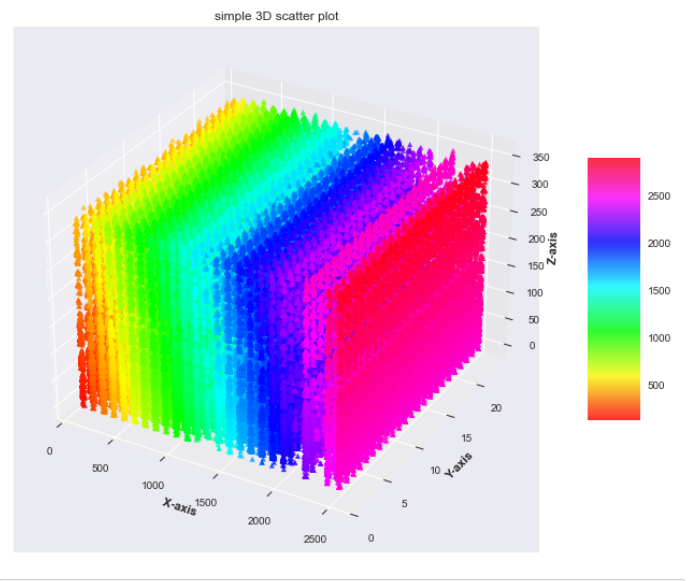
We use the Silhouette metric to see how each point of the data is closely matched to its cluster.



Looking at the distribution, we can see 3 clusters are made



**Visualising the clusters**



**5 CONCLUSION**

We hope that this data analytic project can provide us with a scientific perspective on Chicago's security situation and crime rate. We can see the most commonly occurring crimes and the frequently occurring places where crimes occurred based on the research results and visualisation. The most common crimes were burglary, battery, criminal harm, and drugs, accounting for 65.7 percent of all recorded crimes. The most popular places for crimes to occur are on the highway, sidewalks, residences, and apartments, which are all places where people congregate. We looked at various crime forms, such as robbery, murder, and sexual offences, to see how they've evolved. Even though there were many recorded crimes in Chicago each year, the arrest rate was not even close to 50%, leading us to conclude that Chicago's police arrest or investigation procedures were ineffective. We assume that if our data analytics can provide us with all of this knowledge about Chicago's security status, a larger data analytics project can provide even more useful information that can be used as a powerful source for taking wise steps to improve our cities' security status.

**6 Literature on Chicago Crime**

Law enforcement has changed in recent years as a result of better tactics, computer-assisted technology, and resource efficiency, among other things. As a result of these factors, the crime rate in the United States has dropped dramatically in recent years. For information, law enforcement has turned to data science (ranging from reports, corrective analysis and behaviour modelling). In recent years, crime rates in Chicago have decreased overall. When compared to previous decades, these rates are at an all-time low. This paper describes historical patterns, observations, and other findings in Chicago from 1965 to 2018 using the criminal dataset contained at “data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2,” rather than assigning some casual interpretation of the vang from 1965 to 2018 in Chicago, and not to give any arbitrary interpretation to the vanguards of crime rates during that period. For preparation and crime prediction, the K-Nearest Neighbor (KNN) classification is used. There are also discussions about potential investigations. The accuracy of the proposed model is 83.2 percent.

7 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].

[9]Anon, 2018. *crimeanalyze.firebaseapp.com*. Available at: <https://crimeanalyze.firebaseapp.com/> [Accessed December 14, 2018].