Programming For Big Data Analytics

Project Report

Chicago Crime Data Analysis

Gururaj Shinde - gss399

Kushan Singh – ks5013

Nikhila Dwarakanath – nd1556

Ritu Kuklani – rtk304

CONTENTS

INTRODUCTION:

In this world, crimes are an inseparable part of our lives. Every day we hear about them and some of us even involved in at least one of them during our life. Being cautious and improve safety is not a simple instruction anymore. We need to use modern technology and data science techniques to more wisely act against this problem. There are so many records and documentation in the police department that have been gathered during the years, which can be used as a valuable source of data for the data analytics tasks. Applying analytical task to these data bring us valuable information that can be used to increase the safety of our society and lower the crime rate.

In this project we analyze the Chicago Crime dataset (between the years 2001–2019), which is one of the richest open source data in this area, to get a better understanding about the security status of this city.

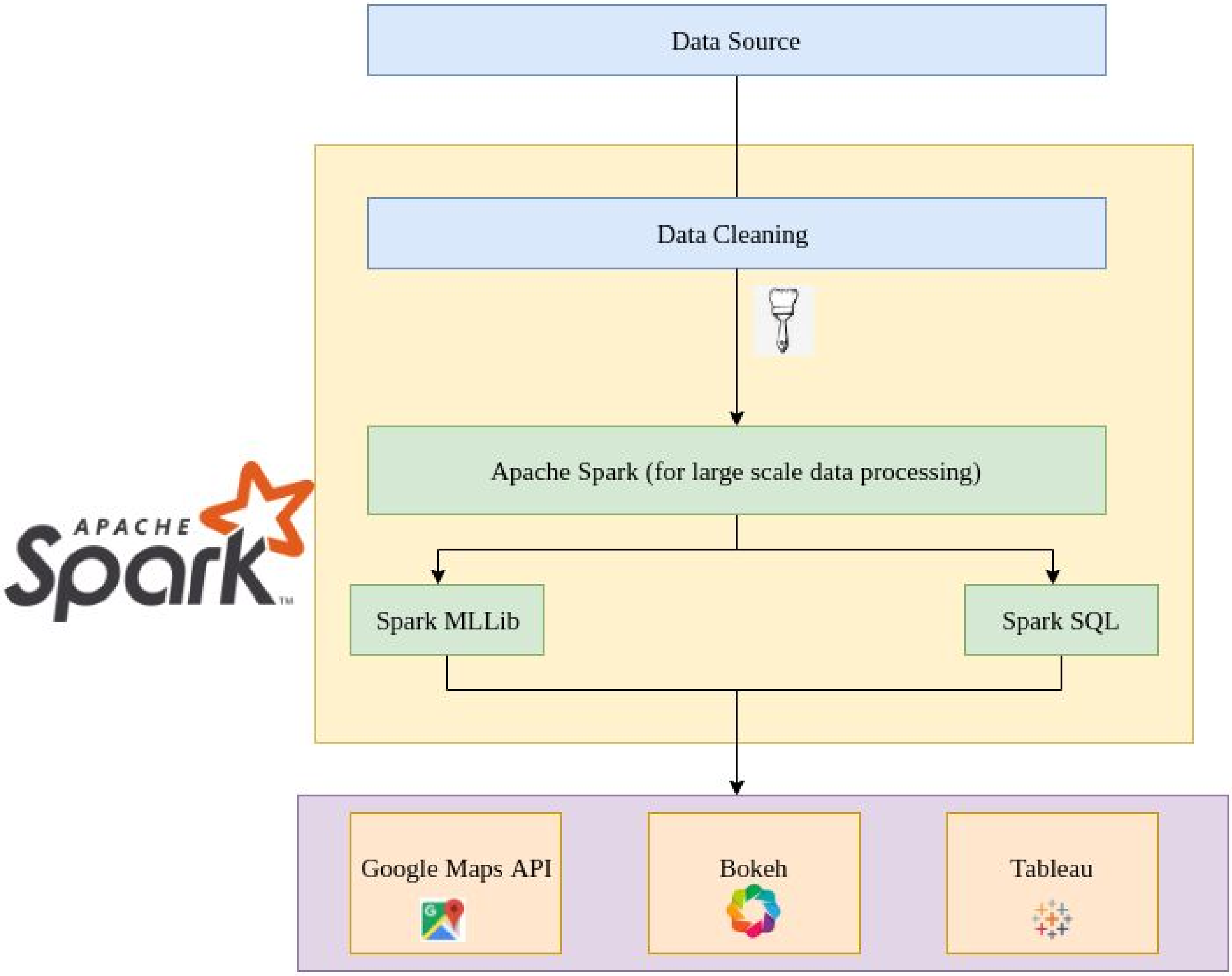
To have a perspective of the security states of the city of Chicago we defined few questions, which we answered them during our data analytic project.

Here is the list of these questions:

1. How has the number of various crimes changed over time in Chicago?
2. How have the number arrests corresponded to the crimes changed over time in Chicago?
3. Are there any trends in the crimes being committed?
4. Which crimes are most frequently committed?
5. Which locations are these frequent crimes being committed to?
6. Are there certain high crime locations for certain crimes (etc Sexual offense)?
7. How has the number of certain crimes (ex. sexual assault) changed over the years in Chicago?

METHODOLOGY:

DATA ARCHITECTURE:



We acquired the Chicago crime dataset from the Chicago city data portal. To answer the questions mentioned above we took the four main steps of the KDD data mining pipeline which are respectively, selection, data pre-processing, analysis and post-processing.

In general, the data included information such as date/time when the crime happened, the block where the crime occurred, type of crime, location description, whether there was an arrest, and location coordinates.

DATA PREPROCESSING:

In the project, the Chicago Crime dataset requires one of the most important data pre-processing procedures which is cleaning. Our data need to be clean by:

1. Removing duplicate rows
2. Removing missing values (etc. Null/NA values) in the dataset
3. Filtering out all the features from the dataset that are not relevant to our data analysis (etc. X Coordinate, Y Coordinate, Latitude Longitude).

VISUALIZATIONS:

Finally, we visualize the trends in the data over a period of time. We have used libraries such as matplotlib and sns for these visualizations. Moreover, we use Tableau to handle the visualization by using the input as csv files obtained by exporting the spark dataframes.

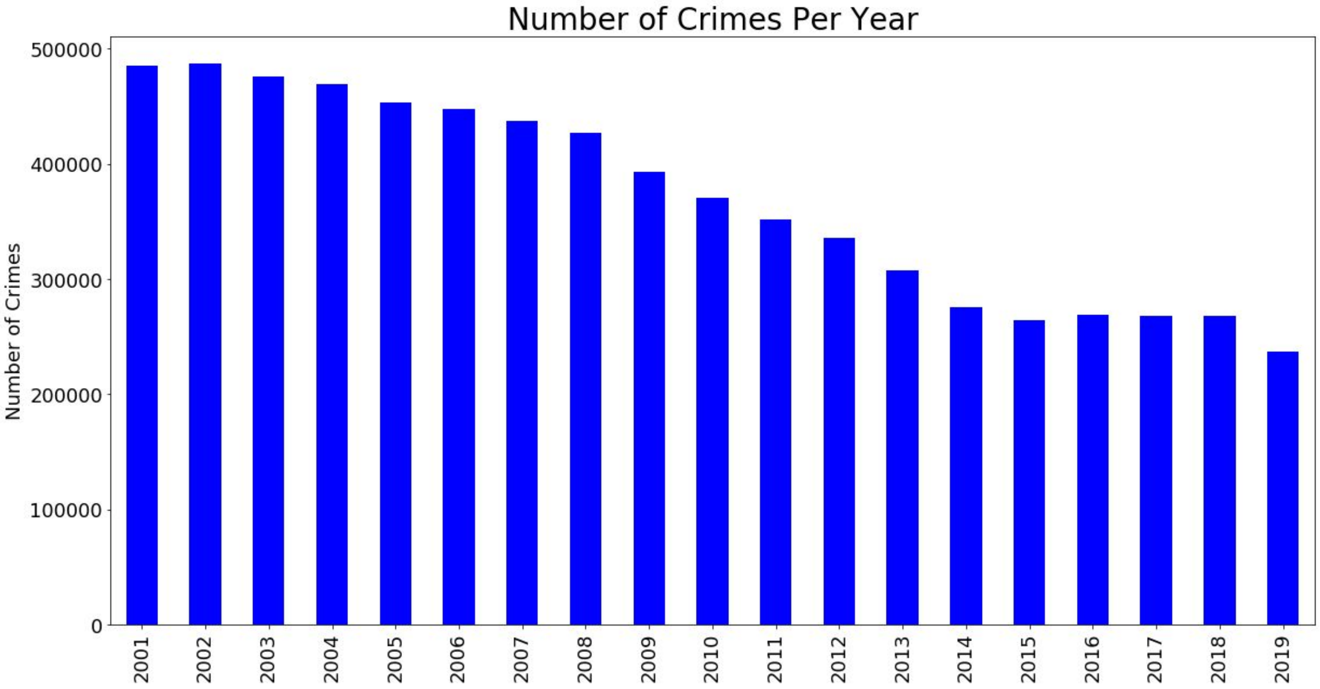


Fig 1: Number Of Crimes Reported Per Year

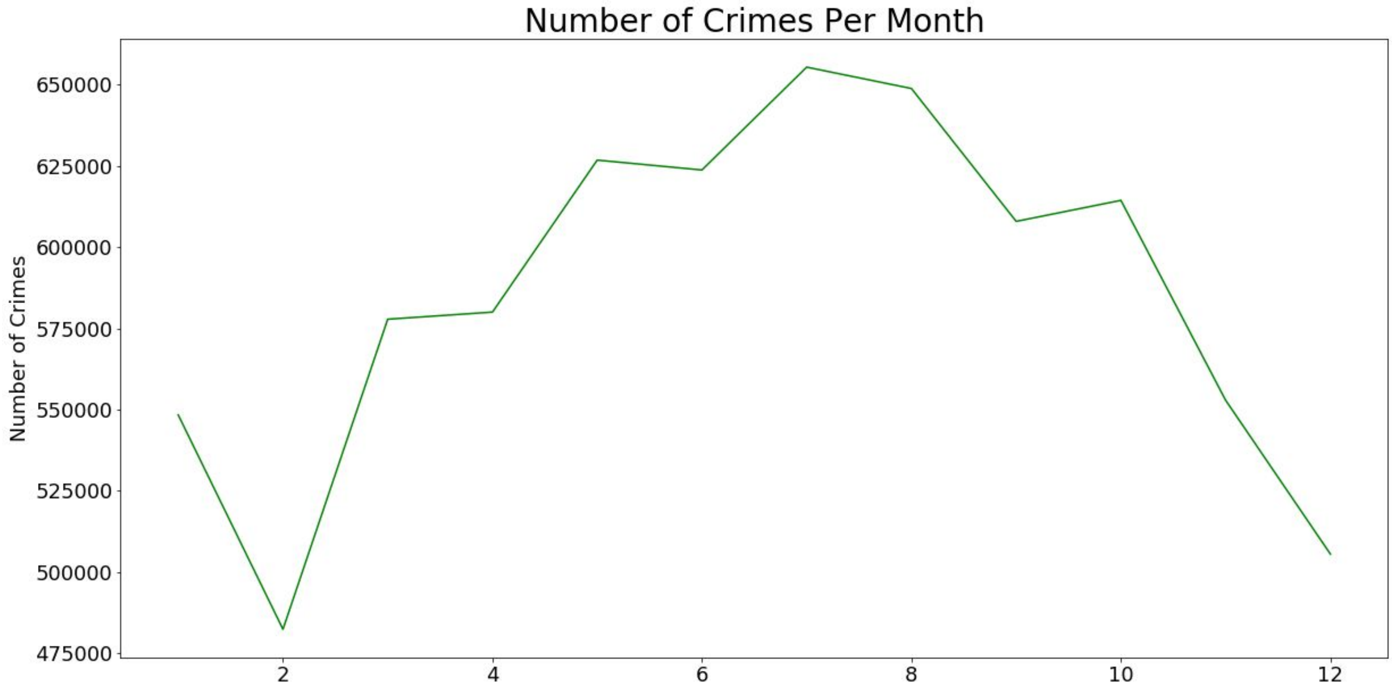


Fig 2: Number of crimes reported per month.

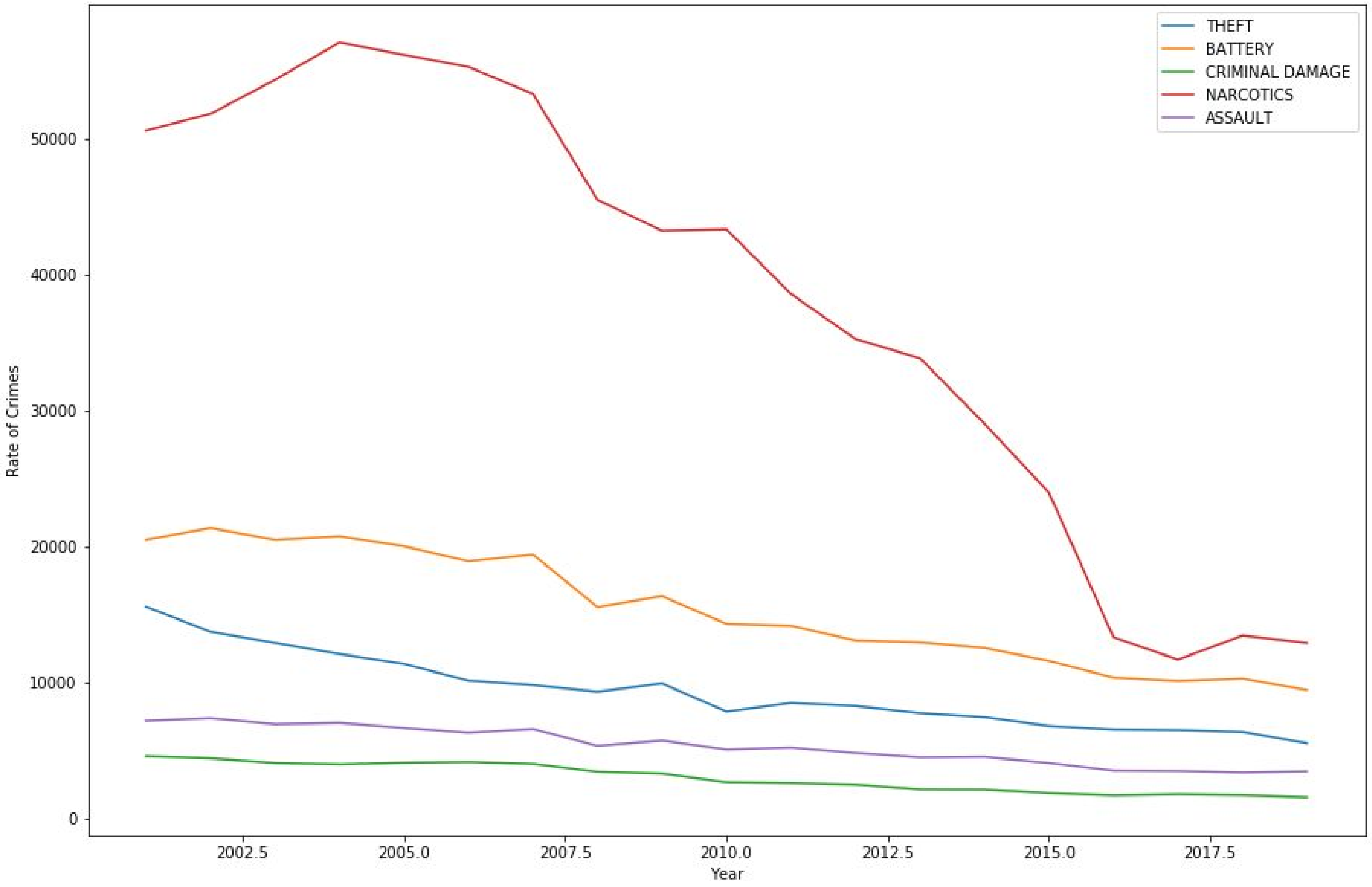


Fig 3: Trends in the top 5 primary types of crime.

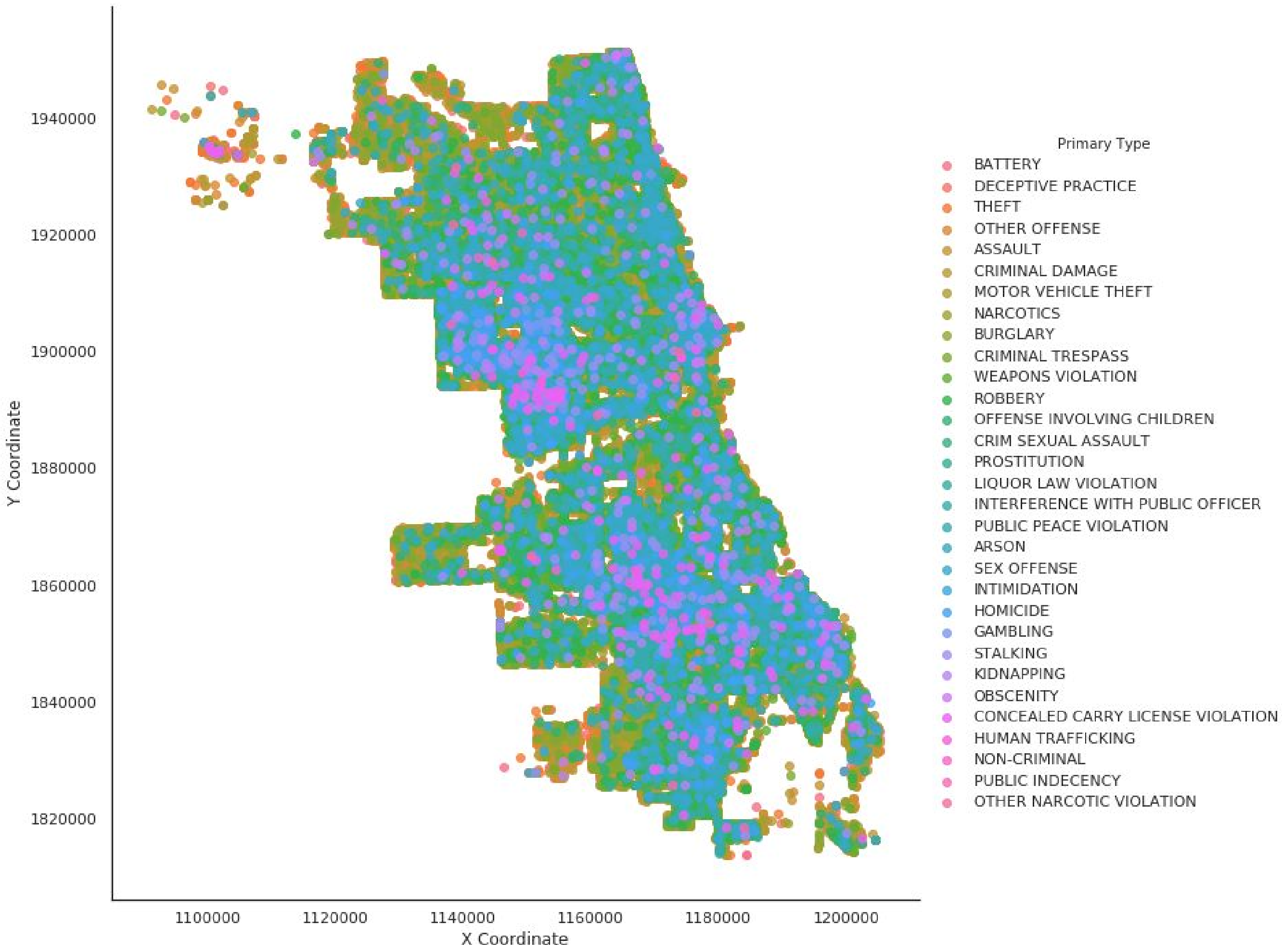


Fig 4:Heatmap of the primary type of crime.

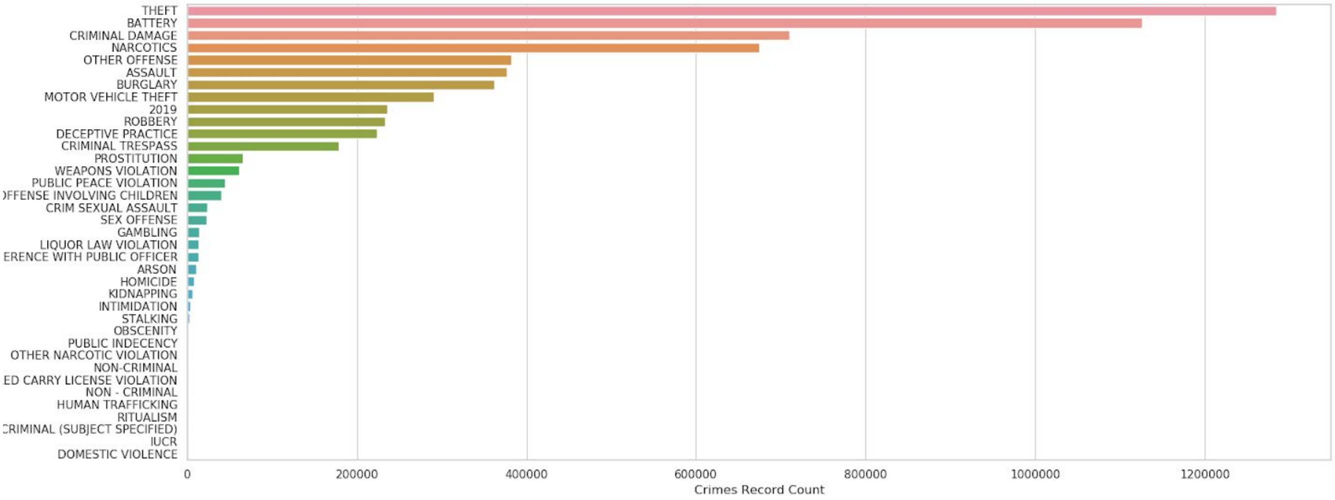


Fig 5: Count depending on the primary type of crime

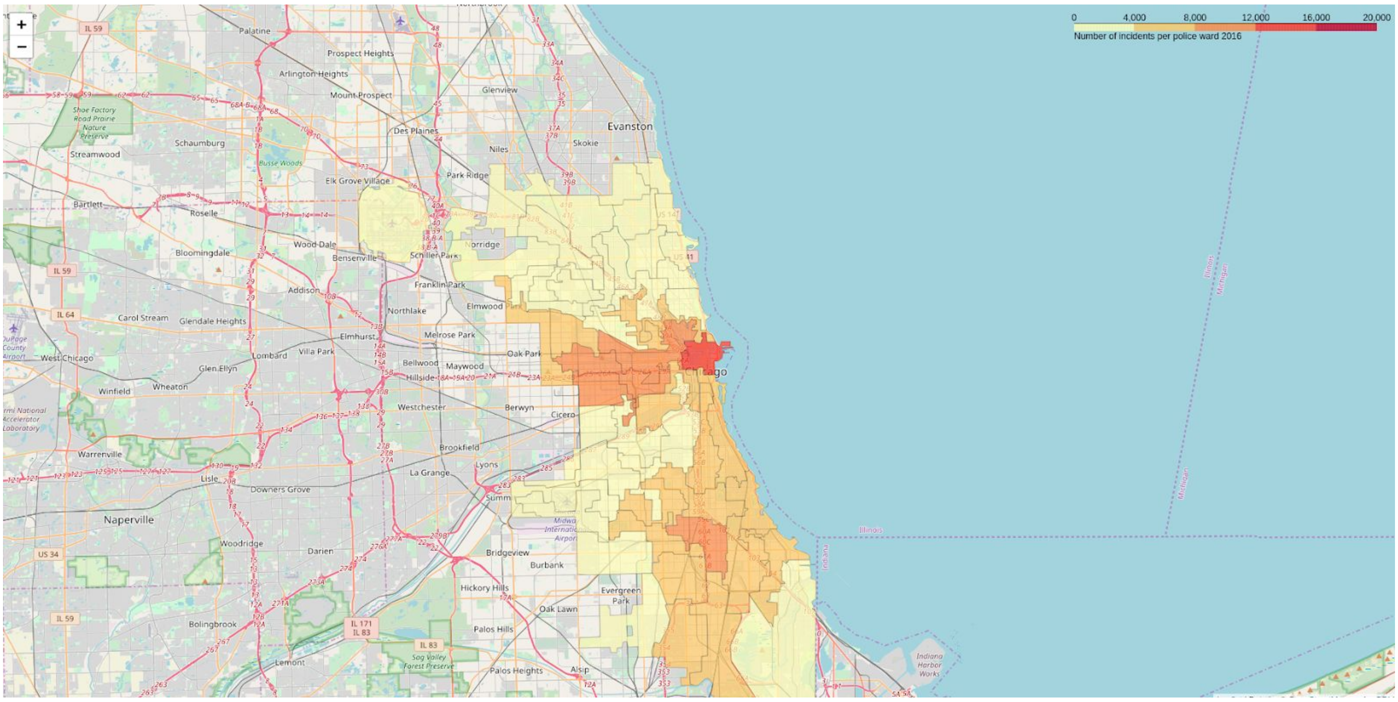


Fig 6: Number of incidents per police ward.

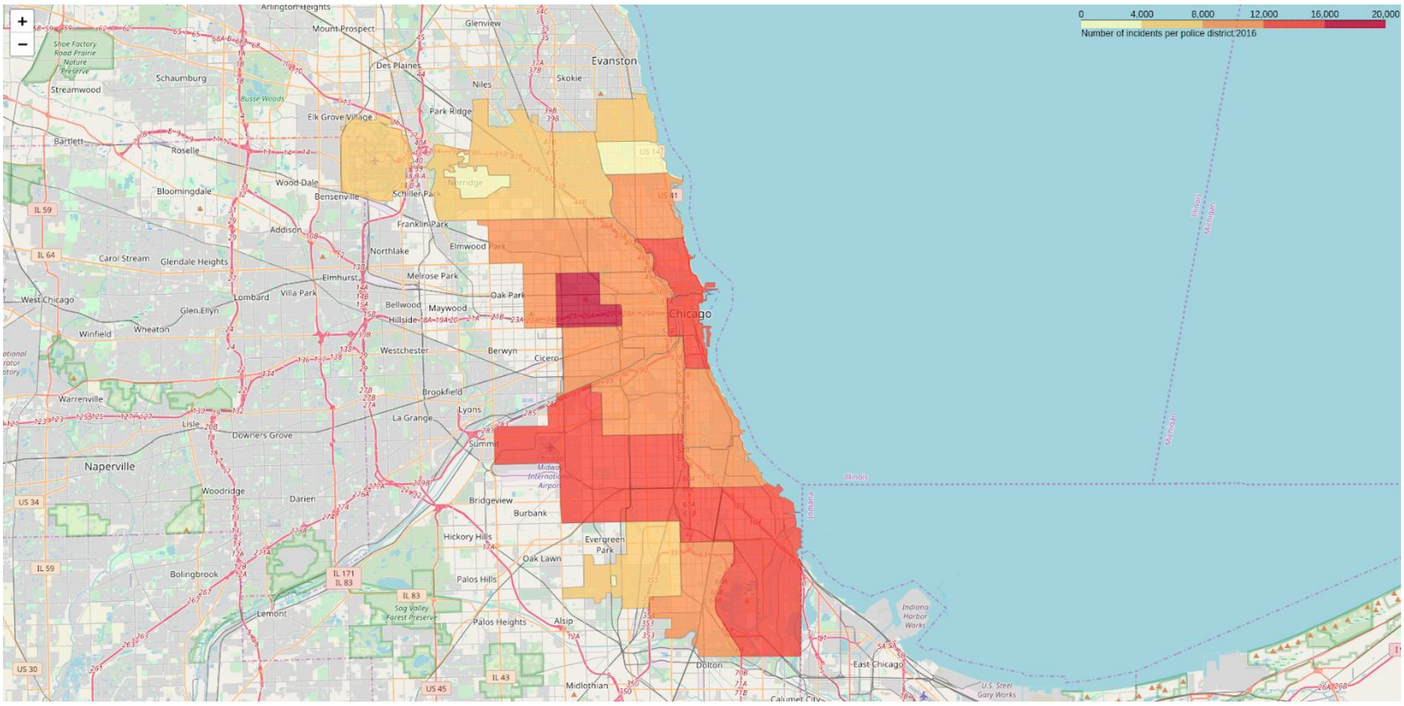


Fig 7: Number of incidents per district

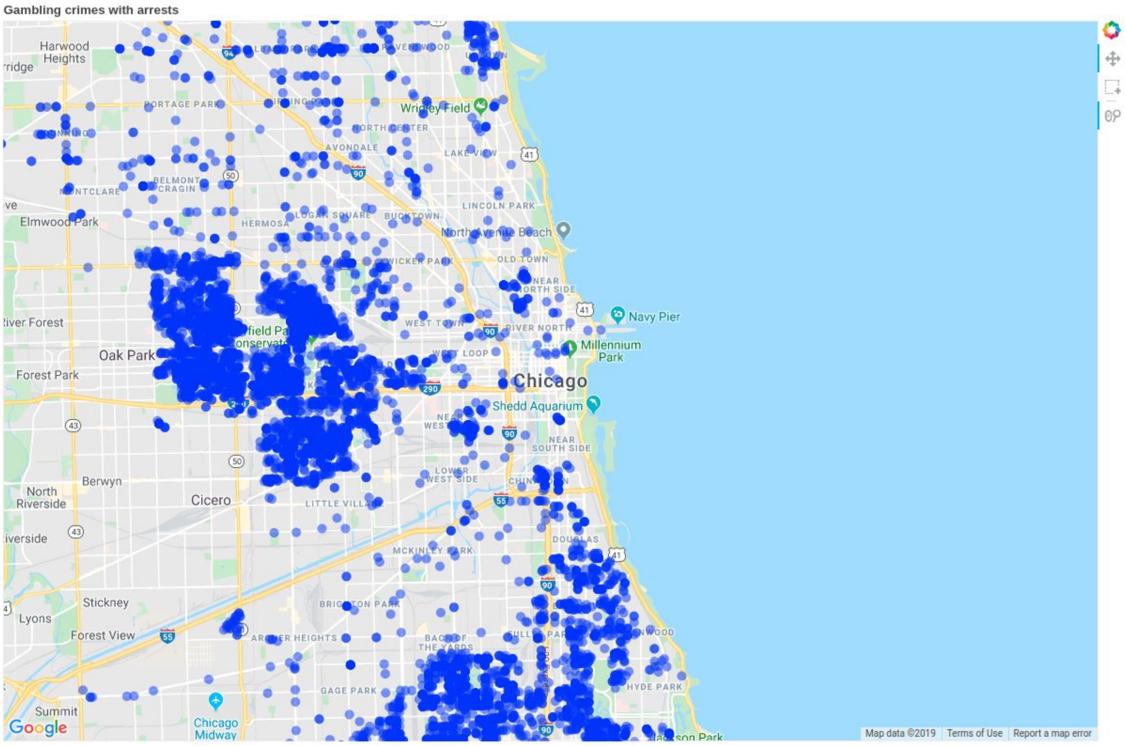
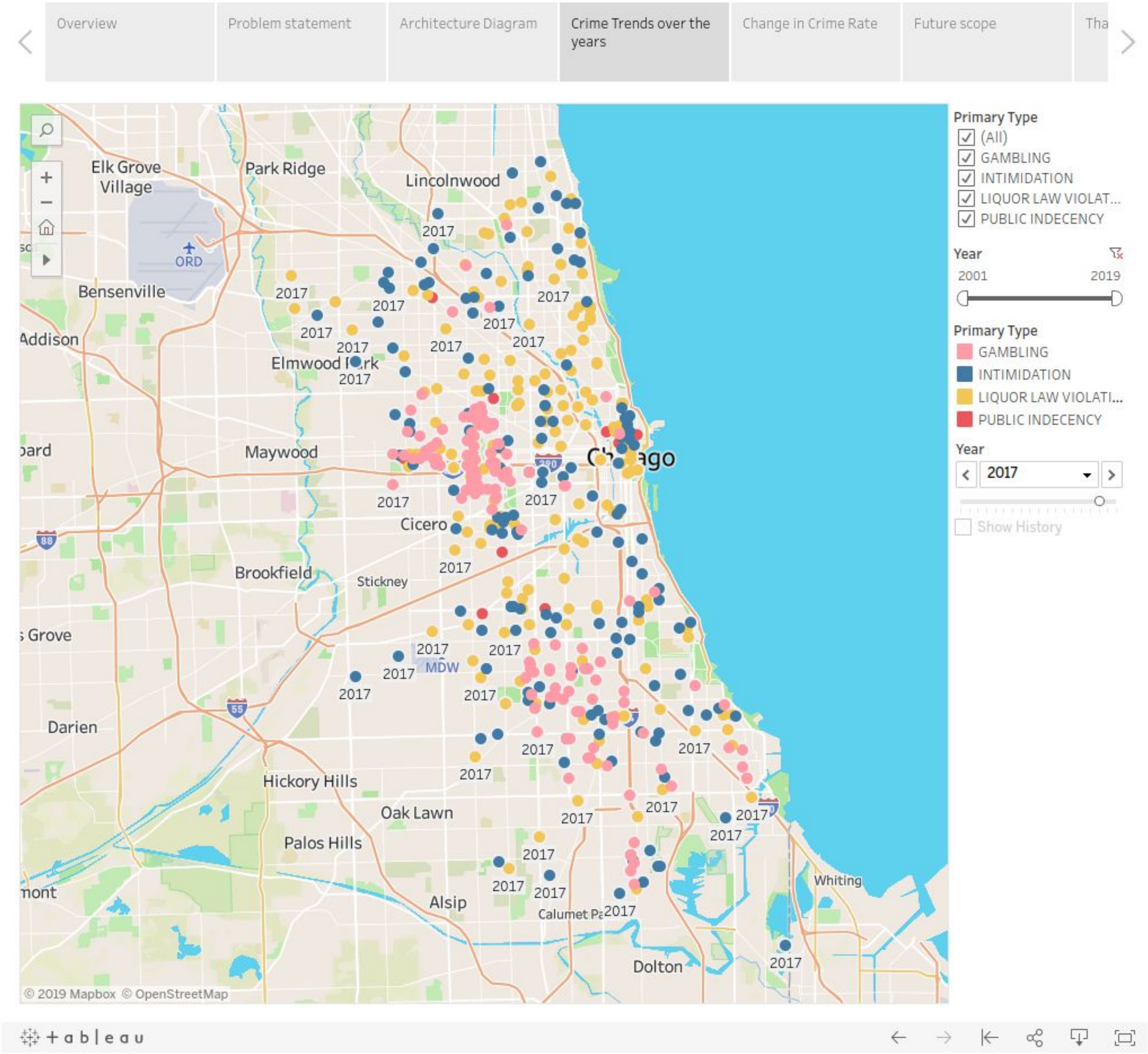


Fig 8:Locations of gambling crimes where arrests were made with Bokeh



Overview | Problem statement | Architecture Diagram | Crime Trends over the years | Change in Crime Rate | Future scope | Tha

Primary Type

☑ (All)
☑ GAMBLING
☑ INTIMIDATION
☑ LIQUOR LAW VIOLAT...
☑ PUBLIC INDECENCY

Year

2001 2019

Primary Type

🟥 GAMBLING
🟦 INTIMIDATION
🟨 LIQUOR LAW VIOLATI...
🟥 PUBLIC INDECENCY

Year

2017

□ Show History

© 2019 Mapbox © OpenStreetMap

Fig

Fig 9: Plots using geospatial co-ordinates for 4 types of primary crimes

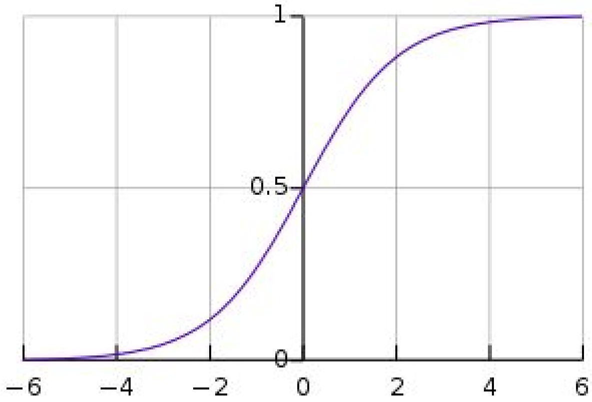
PREDICTIONS:
Machine Learning models used:

LOGISTIC REGRESSION:

Logistic regression is one of the most fundamental and widely used Machine Learning Algorithms. Logistic regression is not a regression algorithm but a probabilistic classification model.

Classification in Machine Learning is a technique of learning, where an instance is mapped to one of many labels. The machine learns patterns from data in such a way that the learned representation successfully maps the original dimension to the suggested label/class without any intervention from a human expert.

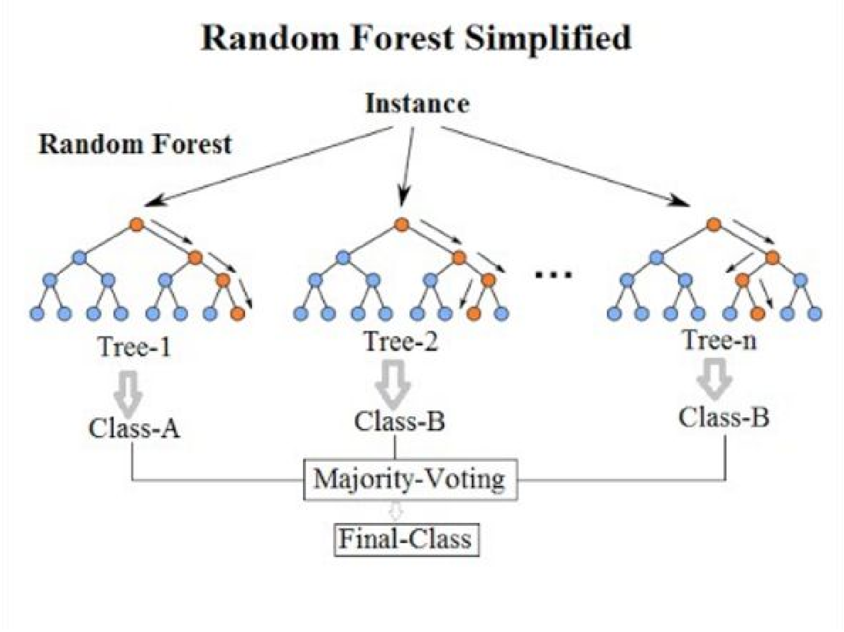
Logistic regression has a sigmoidal curve.



Multiclass classification with logistic regression can be done either through the one-vs-rest scheme in which for each class a binary classification problem of data belonging or not to that class is done, or changing the loss function to cross- entropy loss.

In the multi class logistic regression python Logistic Regression class, multi-class classification can be enabled/disabled by passing values to the argument called ‘‘multi_class’ in the constructor of the algorithm. In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the ‘multi_class’ option is set to ‘ovr’ and uses the cross-entropy loss if the ‘multi_class’ option is set to ‘multinomial’. In our project we have multi-class to ‘multinomial’.

RANDOM FOREST:



If we take predictions from hundreds or thousands of decision trees, some of which are high and some of which are low, and decided to average them together,then we create a random forest. The fundamental idea behind a random forest is to combine many decision trees into a single model. Individually, predictions made by a single decision tree may not be accurate, but combined together, the predictions will be closer to the mark on average.

In our project, we predicted the primary type of crime using Logistic Regression and Random Forest to predict the primary type of crime. The features used were “Arrest”, “Domestic, “Beat”, “District”, “Location Description”, “District”, “Ward”, “Community Area”, “FBI code” and "hour”. We could achieve an accuracy of 51.57%. We achieved the maximum accuracy for prediction of thefts with an accuracy of 92.18%. We also implemented predictions using Random Forest which gave us an accuracy of 76% using the same features to predict the primary type of crime.

FUTURE WORK:

The future scope of our project includes a real-time integration to the spark pipeline using Kafka streams. This can help us provide analysis on real-time data so that we can take the necessary preventive measures. Also, we could implement the same model to different city datasets and compare the results to take necessary preventive measures. Moreover, we could use a scalable dataset such as MongoDB in order to store the results of our analytics features.

CONCLUSION:

We believe this data analytic project give us a scientific view about the security status and crime rate of the Chicago city. According to the analysis result and visualization, we can view the most frequently occurring crimes and the frequent occurring locations where crimes happened. From these reports, the most occurred crimes were theft, battery, criminal damage and narcotics which is 65.7% of all the crimes reported. We specifically looked into certain crime types to view how they have changed over the years, such as theft, battery, criminal damage, narcotics and sexual assault crimes. Even though there were a lot of reported crimes in Chicago each year, the arrest rate was not even as high as 50% for each year letting us believe that Chicago's police arrest or investigation methods were not effective enough. We believe if our data analytics can give us all these information about the security status of the Chicago city, a bigger data analytics project will provide much more valuable information which can be used as a powerful source for taking wise actions that increases the security status of our cities.

CODE:

The code used for this project is attached as a zip file with this report. It is also available on Github at https://github.com/kushan22/BigData-Project

REFERENCES:

[1] Chicago data portal

[2]My notes on Chicago crime data analysis. **https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20**

[3]Chicago crime mapping **https://hackernoon.com/chicago-crime-mapping-magic-of-data-science-and-python-f2ecad74a597**

[4] Pyspark documentation

[5]Tableau: Business Intelligence And Analytics Software

[6]Interactive Maps with Bokeh https://automating-gis-processes.github.io/2017/lessons/L5/interactive-map-bokeh.html

[7]Machine Learning with Spark and Mlib https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa