Mike Peyton, Logan Cuccia, Kyle Cummings, Jimit Bhalavat
CS435 Term Project Proposal
Dr. Sangmi Pallickara
October 21, 2021

Analyze Crime Rates to Help Mitigate Crimes in Chicago.

Chicago was once a pillar of prosperity and trade where companies housed their headquarters, where business partners formed coalitions, and where impactful business decisions were made. Today, it is still a major capital for trade, and some areas are very prosperous. Other sectors are rampant with crime, murder, theft, break-ins, and kidnappings. The crimes within some Chicago districts are so high that officers often don't know whether they will return home that day to their families. Although there are countless factors contributing to the growing crime rate and law enforcement procedures, crimes are overwhelmingly complex and have to do with politics, state, and local government actions. The question arises how has crime changed over the years? Is it possible to predict where or when a crime will be committed? Which areas of the city have evolved over this time span?

The goal of this project is to tackle how we can use data from these crimes to potentially stop them from becoming so frequent. It will also provide data summarizations such as the location and frequency of the crimes, and the nature of the crimes which can ultimately help mitigate crimes and help law enforcement agencies pinpoint where these crimes are more likely to occur, and which districts should have a higher abundance of police officers. This information can not only save the lives of citizens but also police officers who risk their lives every day to better their communities. It also ensures that these criminals have a higher likelihood of being caught since so many crimes in these districts go unsolved. The dataset contains attributes such as date, coordinates of the incident, district, the nature of the crime, etc. We will attempt to use these attributes to generate summarizations that will assist in the fight against crime as well as assist in the city's economic future. This is an interesting Big Data problem because "With the fast development of positioning technology and prevalence of mobile devices, a large amount of modern urban data have been collected and such big data can provide new perspectives for understanding crime" (Wang, Hongjian).

In order to solve this problem and help aid law enforcement agencies in Chicago, we will aim to analyze this dataset and run a variety of different processing techniques and models in order to interpret results and answer the questions mentioned above. This dataset is large and contains various attributes, so it is complex to analyze through normal methods. In order to process this data effectively and efficiently, we will use MapReduce to process the data and provide numerical summarizations such as the accuracy rates of prediction through Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN). Other numerical summarizations will include the proportions of arrests made by the crimes committed, whether the trends in crimes have increased over the years, and if yes, what crimes are more likely to be committed.

In this project, we aim to analyze the relationship between different types of crimes and the location they happened. In our research, we also aim to analyze the statistics of a few specific crimes: theft, homicide, and sexual harassment, and whether these crimes have declined or increased over the years. We aim to begin with KNN and try to predict the relationship between the types of crimes and the location they occurred. We also plan to use Decision Trees and Logistic Regression to try and predict the accuracy of our model and aim to predict if the crimes have changed over the years. After selecting the best model, we plan to provide data visualizations so that it is easily readable and interpretable. We aim to use a few of the packages available in Python3 such as Pandas (Dataframe), Numpy (Math), Seaborn and Matplotlib (Data Visualization), and Sklearn (Algorithms). The framework we plan to use is Hadoop or Spark.

The dataset, which can be found here, comes from Kaggle and reflects the reported incidents of crime in the City of Chicago from 2001 to 2017. The set excludes murders where data exists for each victim. This dataset is withdrawn from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting System. The entire dataset spans around 2GB. In order to span the 16 years of data, the entire dataset has been broken down into four different CSV files. The entries of the files have attributes such as a unique identifier, block where the incident occurred, description of the location where the incident occurred, whether the incident was domestic-related, police district where the incident occurred, latitude, and longitude of the incident, to name a few. To protect the privacy of crime victims, the exact addresses of the crimes are not shared, but the location data included will allow us to pinpoint the general location of the crime. We can use the description attributes to categorize what type of crime was committed, and also filter based on other attributes such as whether an arrest was made, or whether the case was domestic.

Week	Task	Task Description	Team Member
Week 10	Refine Dataset Part 1	Clean the dataset and only keep the attributes of interest. Data preprocessing: Data Exploration and Data Extraction (Part 1)	Mike Peyton Kyle Cummings
Week 10	Refine Dataset Part 2	Clean the dataset and only keep the attributes of interest. Data preprocessing: Data Exploration and Data Extraction (Part 2)	Logan Cuccia Jimit Bhalavat
Week 11	Implementing MapReduce	Use MapReduce for Data Exploration and Data Extraction. Extract data and attributes that are necessary to the project.	Kyle Cummings Jimit Bhalavat Logan Cuccia Mike Peyton

Week 12	Implementing K-Nearest Neighbors (KNN)	Implement KNN using the location the crime occurred, and looking at a time series the crimes take place in specific months of the year to predict what crime can take place at what location. In the end, we aim to achieve higher prediction accuracy rates using a KNN approach.	Mike Peyton Jimit Bhalavat Logan Cuccia Kyle Cummings
Week 13	Implementing Logistic Regression	Implement Logistic Regression to predict the types of crimes at a particular location.	Kyle Cummings Mike Peyton
Week 13	Implementing Decision Trees	Implement Decision Trees in order to come to a conclusion of the types of crimes committed at a location in the coming months and years.	Jimit Bhalavat Logan Cuccia
Week 14	Data Visualization	Provide Data Visualization using different aspects of the dataset, such as which crimes have declined or increased over the years. We also aim to provide a correlation between different attributes and decision trees visualization as well.	Mike Peyton Logan Cuccia
Week 14	Conclusion	Conclude the project with the results achieved through different strategies such as KNN, Logistic Regression, and Decision Trees. Mention the prediction accuracy rate and how the project is of importance to the law enforcement agencies.	Kyle Cummings Jimit Bhalavat

Bibliography

Currie32. “Crimes in Chicago.” Version 1, Kaggle, 28 Jan. 2017, https://www.kaggle.com/currie32/crimes-in-chicago

Wang, Hongjian et al. “Crime Rate Inference with Big Data.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 635–644. Web.