DATA 603

Final Project proposal report on the Chicago crime dataset from 2001-present

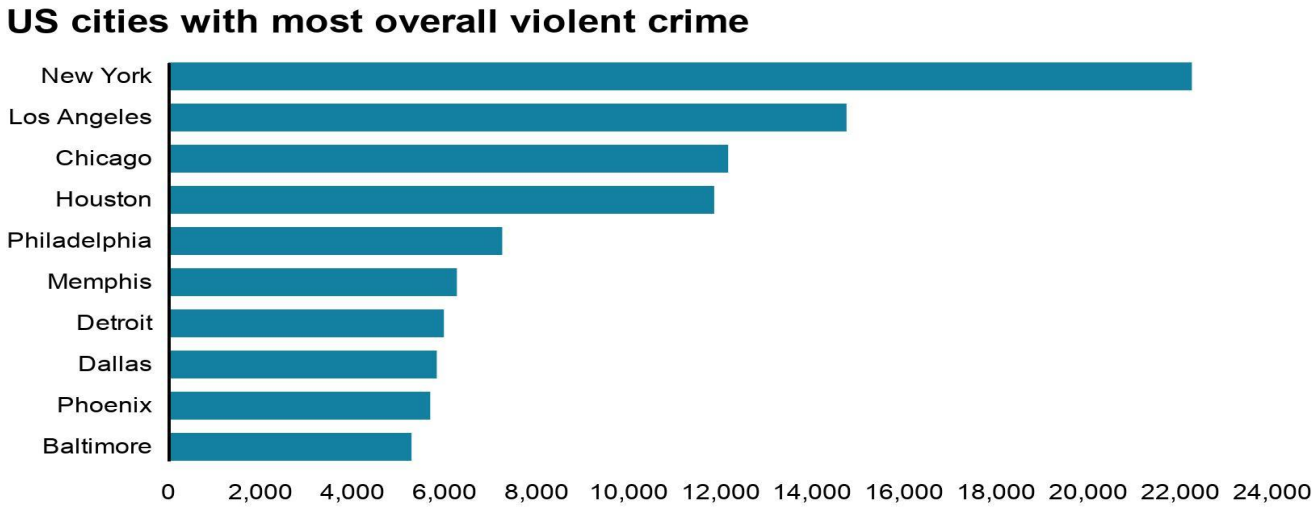
Udveg Reddy Jukanti

02.03.2022

CampusID :- FU63933

INTRODUCTION

The crime rate is a prevalent concern which is skyrocketing in USA for many years For my final project I have chosen a dataset about the crimes in chicago from the year 2001 to present, the dataset was pretty interesting as it had real time information about the crimes happening in chicago, moreover the data is updated everyday. The dataset was collected from the 'city of chicago' official website which has many public datasets. The main reasons why I chose this dataset was because there were many columns which provided a comprehension idea about the dataset and moreover many interesting patterns and insights can be drawn from the dataset. The main reason why I chose Chicago is because it is considered as one of the cities in Illinois with the highest crime rate.



Crime rates vary for different locations. It can be due to population or previous crime rate at that location, accounting these findings are very important and difficult too, but if we have knowledge about the underlying process we can implement changes which would help to mitigate the crime rate. The main idea behind this project is to find out how have crime rates in Chicago changed overtime, how does crime differ by geographic location, and what type of crime has occurred more and did it change based on the location. Rates of violent crime may influence public views of neighborhood safety. Criminal Investigative Analysis, or the behavioral analysis of violent crime, may be operationalized using data analysis and predictive analytics. Behaviorally segmenting crime based on kind, nature, and purpose can yield innovative, operationally relevant, and actionable data, similar to how advanced analytics is used in other fields. These techniques can help in implementing a model which could be helpful for crime prevention and investigation.

About the Dataset

The dataset consists of 22 columns and 7.5 Million rows (dataset size 1.78GB). The dataset has information about the reported crimes from 2001 in the city of chicago except for the murders. The data comes from the CLEAR (Citizen Law Enforcement Analysis and Reporting) system of the Chicago Police Department. Addresses are only published at the block level to preserve the anonymity of crime victims, and individual locations are not revealed.

The dataset link-
https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

While the dataset has many columns it also has a column called (IUCR) Codes which are basically Illinois Uniform Crime Reporting codes; these are four digit codes that law enforcement agencies use to classify criminal incidents when taking individual reports. These codes are also used to aggregate types of cases for statistical purposes. In Illinois, the Illinois State Police establish IUCR codes, but the agencies can add codes to suit their

individual needs. The Chicago Police Department currently uses more than 400 IUCR codes to classify criminal offenses, divided into “Index” and “Non-Index” offenses. Index offenses are the offenses that are collected nation-wide by the Federal Bureaus of Investigation’s Uniform Crime Reports program to document crime trends over time (data released semi-annually), and include murder, criminal sexual assault, robbery, aggravated assault & battery, burglary, theft, motor vehicle theft, and arson. Non-index offenses are all other types of criminal incidents, including vandalism, weapons violations, public peace violations, etc. I am not sure whether I will use this particular dataset in depth or not however in the case of use I would like to attach the link for this dataset as well.

The Dataset Link-
https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

PROBLEM STATEMENT

1. To investigate the sorts of crimes that happened more often.
2. What types of offenses were deemed serious, and whether or not they were punished.
3. Based on the geographic location of crimes, looking for the hotspots where crime occurs more often.
4. By using IUCR dataset and combining the crime dataset to classify the criminal offenses.
5. Finding out the most common crimes across chicago.

   These are some of the issue statements that made sense when looking at the data; however, as we gain a better understanding of the data through EDA and analysis, there may be additional fascinating patterns and insights through which we might discover more effective problem statements.

SOLUTION

There are various ways to find answers while looking at difficulties, but given the magnitude of the data, we must be sure to use a strategy that gives solutions quickly and effectively.

Pyspark supports programming abstraction called dataframe, moreover pandas API on spark gives us access to deal with dataframes which will allow us to work with dataframes using pandas. Apache Spark's streaming functionality allows for strong interactive and analytical applications to be built on both streaming and historical data, while retaining Spark's ease of use and fault tolerance.

As defined in the problem statement, wherever machine learning pipelines are required we can use the MLib library that provides high level API and machine learning modules.

As the spark core is the underlying engine for the spark platform it provides many big data functionalities one of the most important is Resilient Distributed Dataset (RDD) allowing mapreduce functionalities efficiently and effectively.

For the time being, I've figured out these concepts, but looking at the dataset, there's a lot I'd like to do, and because pyspark is simple to understand and use, plus it provides great visualization, it would allow me to do more.

REFERENCES

1. https://www.chicago.gov/city/en.html
2. https://www.sciencedirect.com/topics/social-sciences/crime-rate
3. https://spark.apache.org/docs/latest/api/python/
4. https://www.bbc.com/news/world-us-canada-53991722
5. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019