Crime Data Analysis

(New York, Los Angeles and Chicago)

Arun Kodnani
New York University
New York, NY
ak6384@nyu.edu

Vaibhav Lodha
New York University
New York, NY
vl1015@nyu.edu

Vatsal Shah
New York University
New York, NY
vds254@nyu.edu

ABSTRACT

Crime data analysis is one of the highly focused activities for any law enforcement agency in the world. An exhaustive study of historical crime data could help the law enforcement department to understand the shifting trends, causes and various other factors that affect crime. The major problem here is there are a lot of factors that affect the crime in different ways. Some of those factors are weather, population, education, growth, policies etc. In this project, we aim at studying the trends in crime data and try to find a few factors that affect the crime in a particular area. To do so we have included the study of three different cities viz. New York, Los Angeles, Chicago. Our aim would be to find trends that are similar in all the cities and understanding the reasons that make up those trends.

ACM Reference Format:
Arun Kodnani, Vaibhav Lodha, and Vatsal Shah. 2018. Crime Data Analysis: (New York, Los Angeles and Chicago). In *Proceedings of ACM Conference (Conference'17),* Arun Kodnani, Vaibhav Lodha, and Vatsal Shah (Eds.). ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Crime analysis has become one of the most vital activities of the modern world due to the high magnitude of crimes which is a result of several factors combined. These factors vary a lot, from the socio-economic structure of society to law enforcement department in the state, from natural factors like temperature and precipitation to policies formulated by the state government. The study of crime becomes even more complicated as the impact of each factor changes over the period of time. Also, the definition of crime varies for each state. This makes the relative study of crime across different cities difficult. In the United States, state law enforcement agencies collect data as is no central agency that does so. The unified crime reporting (UCR) program does the task of collecting data from each state and producing unified data in a standard format. But this data is in the aggregated format and is not feasible for a granular study.

2 PROBLEM FORMULATION

Reducing the crime rate within the country is one of the highest priority of any federal government. In order to make policies to tackle crime, the officials need to have a wholistic idea of crime rate/pattern across the entire country. But, for a country like the United States, this becomes a huge problem as every state has there own definitions of different crimes. And to further add to the problem, different states have different platforms for reporting the crime. The Uniform Crime Reporting(UCR) initiative by the Federal Bureau of Investigation (FBI) is a great effort by the federal government to unify the crime reporting system and terminologies, but the system still hasn't been operationalized by all the states. Our project aims to provide broader insights into crime patterns by using the publicly available dataset for New York, Los Angeles and Chicago, all of which follows their own platform for crime reporting.

3 METHODOLOGY

3.1 Pre-Processing

We have gathered the data of Criminal Activities and incidents of Three cities mentioned above are from below sources. The New York Crime Data has been gathered from SOURCE. The Chicago Crime Data has been gathered from SOURCE. The Los Angeles Crime Data has been gathered from SOURCE.

3.2 Data Mining

Retrieving initial Crime Analytics across all the cities individually and finding to similarity. For all the trends and similarity, We will try to enhance and go into the detail in more granular level.

3.3 Results Validation

Execute data crawling and observe the patterns lying beneath hidden. If there are any abnormalities, irregularities, or inconsistencies, certainly there are some introspection and observations have to be done.

3.4 Understand Data

The Crime domain is widely diverse in terms of rules and regulations and even in terms of definition. Sometime the violation considered at one place is not recognized as the same at other which will create discrepancies in between cities because some of the definition will bring huge differences in the state. To analyze the data, trends, similarity, and correlation we have to create and consider the hypothesis carefully and analyze.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.5 Prepare Data

To prove or correlate the hypothesis exclusiveness, we have to gather the supporting datasets to assert the relations and provide concrete back-support for the hypothesis analysis.

3.6 Model Data

Applied supporting datasets to universal crime database and got s list of statistical significance to support or revoke the hypothesis credibility.

3.7 Evaluation

Previously conducted studies similar to the ones mentioned in the references would be used to verify the correctness of the analysis. Other evaluation metrics include discovering factors that greatly affect crime and the ease with which our work could be reproduced.

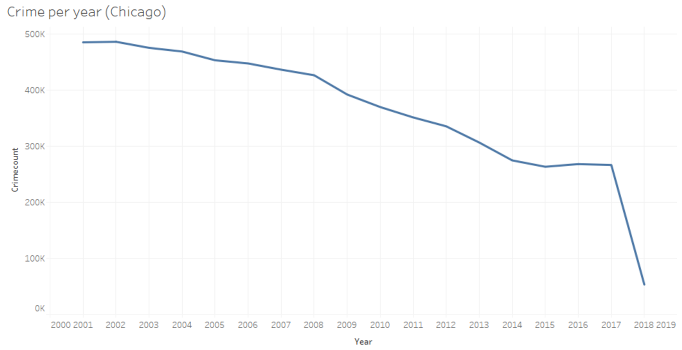
4 PRELIMINARY RESULTS

4.1 Chicago

4.1.1 Data Summary. Crime data for Chicago city is obtained from catalog.data.gov website. Data found is in CSV format. This file has data dated from 2001 to present. Total number of reported crimes during this period are 6.568 Million. Total size of the data is 1.51 GB. Total number of features in the dataset is 22.There is a row for each crime which has a unique case number. Columns in the dataset are listed below: Arrest, Beat, Block, Case Number, Community Area, Date, Description, District, Domestic, FBI Code, ID, IUCR, Latitude, Location, Description, Location, Longitude, Number of Records, Primary Type, Updated On, Ward, X Coordinate, Y Coordinate, Year This dataset has 2 important features to independently identify a particular crime, they are IUCR (Illinois Uniform Crime Reporting code) and FBI Code. This could be really helpful in the standardization of the crime.

4.1.2 Initial Data Exploration and Cleaning. Significant count of null values is found in Location, Description, Ward, Community Area, X Coordinate, Y Coordinate, Latitude, Longitude, and Location. Most of these columns would not be used for data analysis hence null value handling is not required at the moment.No duplicates were found in case number column. This shows that each case is unique.Date formatting is required. Also, conversion of date in timestamp format is required to use the field for queries. Primary Type and Description is unique and well formulated for this dataset. Hence this could be actively used for data analysis and does not require any cleaning. Initial data exploration revealed that Chicago is divided into 22 districts. Earlier there used to be 24 districts but 13 and 23 were shut recently. There is some discrepancy in district column which needs further cleaning.

4.1.3 Initial Trends. Initial query on total data revealed Theft, Battery, Criminal Damage, narcotics and other offense as top 5 crimes. All the above-listed crimes have a decreasing trend over the years. Further exploration of data revealed some of the crimes that have shown a recent increase, they are sexual assault, deceptive practices, homicide, robbery, theft. weapon violation. On the other hand, certain other crimes have shown a recent or sudden decrease, they are burglary, gambling, liquor law violation, narcotics, prostitution, public peace violation. This exploration opens the doors to finding reasons that have led to these sudden changes. This part of the project requires some more analysis and would be a part of final project review.



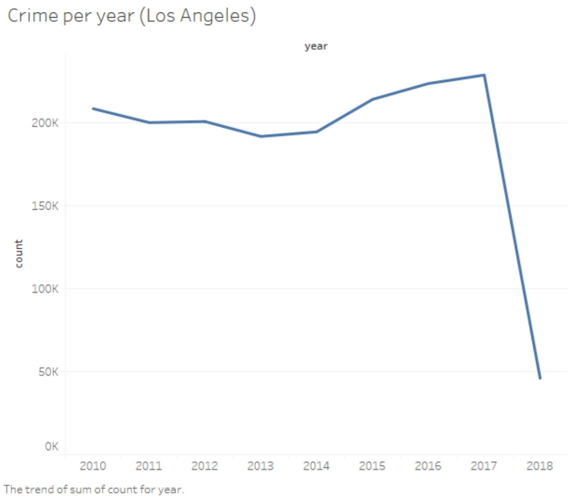
The trend of sum of Crimecount for Year.

4.2 Los Angeles

4.2.1 Data Summary. Crime data for Chicago city is obtained from data.lacity.org website. Data found is in csv format. This file has data dated from 2010 to present. Total number of reported crimes during this period are 1.72 Million. Total size of the data is 400 MB. Total number of features in the dataset is 26.There is a row for each crime which has a unique Division of Records number (DR Number). Columns in the dataset are listed below: DR Number, Date Reported, Date Occurred, Time Occurred, Area ID, Area Name, Reporting District, Crime Code, Crime Code Description, MO Codes, Victim Age, Victim Sex, Victim Descent, Premise Code, Premise Description, Weapon Used Code, Weapon Description, Status Code, Status Description, Crime Code 1, Crime Code 2, Crime Code 3, Crime Code 4, Address, Cross Street, Location

4.2.2 Initial Data Exploration and Cleaning. Significant count of null values is found in MOCodes, Victim Age, Victim Sex, Victim Descent, Premise Code, Premise Description, Weapon Used Code, Weapon Description, Crime Code 2, Crime Code 3, Crime Code 4, and Cross Street. Null values in most of these columns are justified for example in most of the cases of crime victim is not involved, or weapon is not used in every kind of crime. No duplicates were found in DR Number column. This shows that each case is unique.Date formatting is required. Also, conversion of date in timestamp format is required to use the field for queries.Primary Type and Description is unique and well formulated for this dataset. Hence this could be actively used for data analysis and does not require any cleaning.

4.2.3 Initial Trends. Initial query on total data revealed Battery, Burglary from vehicle attempted, Vehicle Stolen, Burglary and Theft as top 5 crimes.

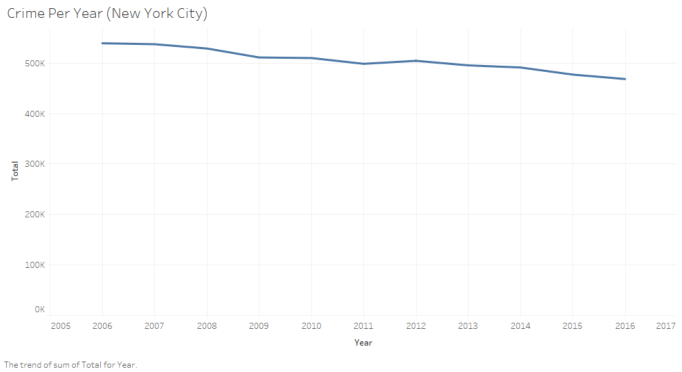


4.3 New York

4.3.1 Data Summary. Crime data for New York city is obtained from NYPD Complaint Data Historic | NYC Open Data website. Data found is in CSV format. This file has data dated from 2006 to 2016. Total number of reported crimes during this period are 5.58 Million. Total size of the data is 1.43 GB. Total number of features in the dataset is 24.There is a row for each crime which has a unique case number.

4.3.2 Initial Data Exploration and Cleaning. Significant count of null values is found in CmplntDate, Borough Name and Latitude-Longitude. Most of these columns would not be used for data analysis hence null value handling is not required at the moment except borough name. No duplicates were found in case number column. This shows that each case is unique. Date formatting is required. Also, conversion of date in timestamp format is required to use the field for queries. Primary Type and Description is unique and well formulated for this dataset. Hence this could be actively used for data analysis and does not require any cleaning.

4.3.3 Initial Trends. Initial query on total data revealed Assault, Harassment, Petit larceny, grand larceny, criminal mischief as top 5 crimes. All the above-listed crimes have a decreasing trend over the years. Further exploration of data revealed total number of crime has been insignificantly decreasing which is a good steps for consideration but there still much work to do reduce the number of criminal violations. This exploration opens the doors to finding reasons that have led to these sudden changes. This part of the project requires some more analysis and would be a part of final project review.



5 TECHNICAL DEPTH AND INNOVATION

5.1 Innovation

In the part of the innovation, It is evident that there are majorly only single city wise analysis case study and none of them are inter-relating different cities analysis. Thus, We are going to combine and analyzing the trends similarity. The conjecture is that for some factors there should be similarity which could have affected the crime incidents. By achieving the result of similarity, federal departments can take preventive steps to cease. We are also aimed to discover the underlying factors of several other dataset like weather, real-estate, census, Household data and find the consequences that this factors have on the Crime Incidents and violations. The effective vital steps can be carried out by this discovery which can really be impactful.

5.2 Technical Depth

For this Project, we are using and adopting various technologies explained below.

5.3 Spark

The functionality of Spark In-memory data processing is the paramount reason we are using this. We are employing the spark functionalities by making different resilient distributed dataset for cleaning and clustering the data.

5.4 PySpark

The another functionality of Apache spark as PySpark opens the data to quickest and most elegant way of initiating analysis. We are incorporating the pySpark functionalities for organizing and retrieving data, and faster ways of moving the data into an analytics framework.

5.5 DataBricks

The Apache Spark Unified Analytics System is helping us to harness the power of truly unified approach providing highly elastic cloud services and effective scale using different interactive notebook support for python and SQL with 10x faster performance. OpenRefine: OpenRefine is granting us powerful tool for working with messy, noisy and corrupted data. We are using it for cleaning, transforming it, and extending it for analytics system.

5.6 Tableau

This tool for translating information to insight and visualization of datasets and various discrete results.

6 CODE REPOSITORY

Code for this project could be found at the following github repo:
https://github.com/ArunKodnani/Crime-Data-Analysis
There is a separate directory for each city. For each city directory there are 4 subfolders viz. data, plots, results and src.
data folder has some sample data and the link for actual dataset
plots folder has some plots from preliminary data exploration results
results folder has some output files
src folder has relevant code for the module

7 REFERENCES

1. http://www-cs-faculty.stanford.edu/
2. https://nycdatascience.com/blog/student-works/correlation-between-weather-condition-and-the-type-of-crime/
3. https://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7112359
4. https://data.cityofchicago.org/
5. https://opendata.cityofnewyork.us/
6. https://data.lacity.org/