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, 9 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.

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

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

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

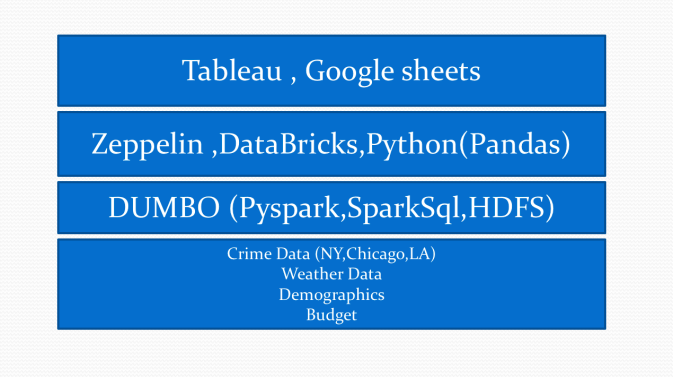
3.1 Pre-Processing

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 NYC Open Data. The Chicago Crime Data has been gathered from Crimes in Chicago (kaggle). The Los Angeles Crime Data and Budget Data has been gathered from Data LaCity. The weather data for all the three cities was collected from National Oceanic and Atmospheric Administration.

3.2 Architecture

The following architecture was adopted by us for this project.We started off by collecting all the data from various sources. Once all the data had been collected, we stored in HDFS on the DUMBO cluster. This enabled us to leverage the functionalities provided by Hadoop ecosystem such as SPARK. Apache Spark is a fast and general-purpose cluster computing system. It provides an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing and MLlib for machine learning. It also supports Resilient Distributed Datasets RDD which is a fault-tolerant collection of elements that can be operated on in parallel. With the help of Spark we were able to perform complex operations with ease. Do perform all these analysis we connected Zeppelin notebook with the HDFS backend on DUMBO while also using Databricks platform on the side. This further enabled us to view and download the results in a faster way. Finally once we were done with our analysis and aggregated our results we used Tableau

and Google Sheets for Visualizations.



3.3 Data Cleaning

Before starting with our analysis we needed to process the data as all the three cities have a different system to reporting and maintaining crime incidents. Moreover, there were lot of inconsistencies in the data such as Name of the Columns, Null Values, Unique Case number, Date formats and inconsistencies with Crime Description and code. One example of inconsistencies with Crime Description code is "Battery and Assault" and "BatteryandAssault".

3.4 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.

3.5 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.

3.6 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.

3.7 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.

3.8 Model Data

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

3.9 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.

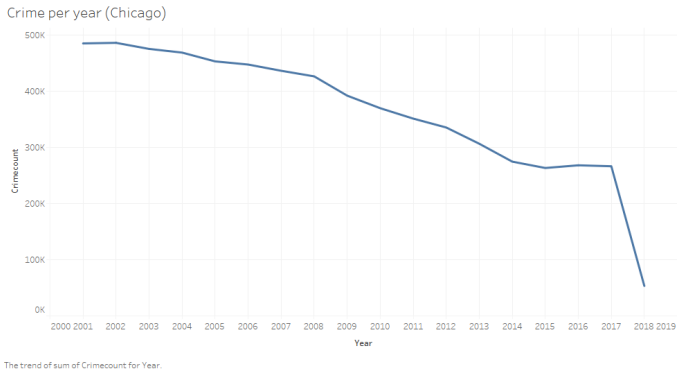
4 PRELIMINARY RESULTS

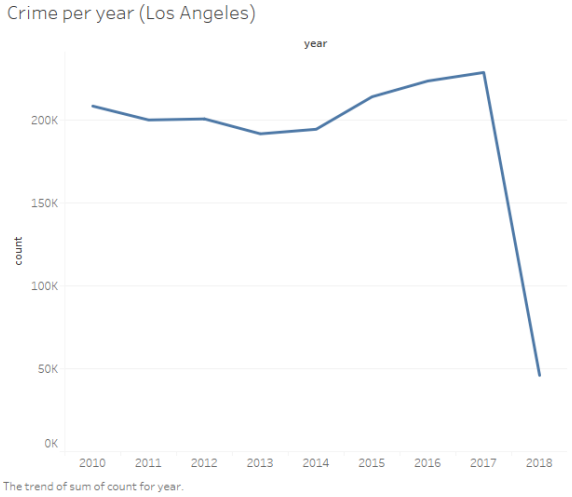
4.1 Chicago

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.





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.

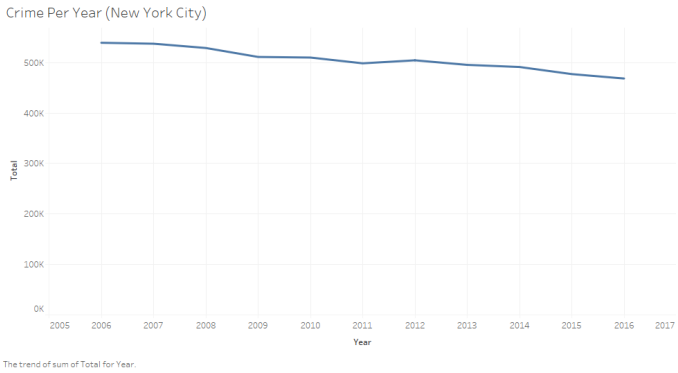
4.2.3 Initial Trends. Initial query on total data revealed Battery, Burglary from vehicle attempted, Vehicle Stolen, Burglary and Theft as top 5 crimes.

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.

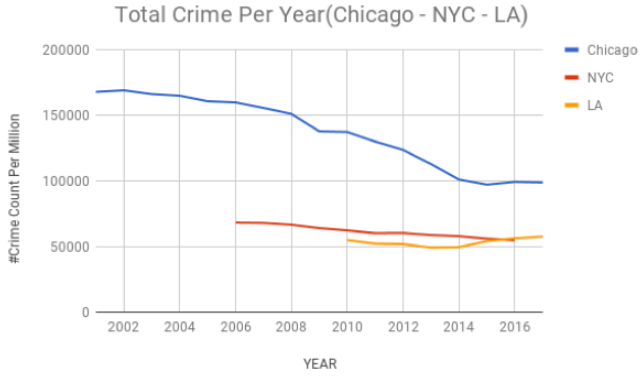
4.3.2 Initial Data Exploration and Cleaning. Significant count of null values is found in CmplntDate, Borough Name and Latitude-Longitude. Most of these columns would not be used for data analysis hence null value handling is not required at the moment except borough name. No duplicates were found in 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.

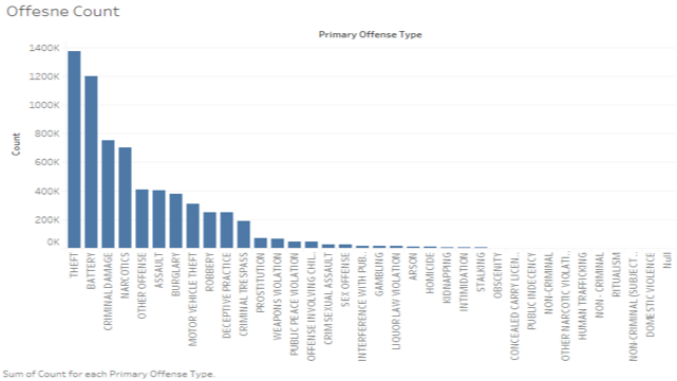


4.4 Overall Trends

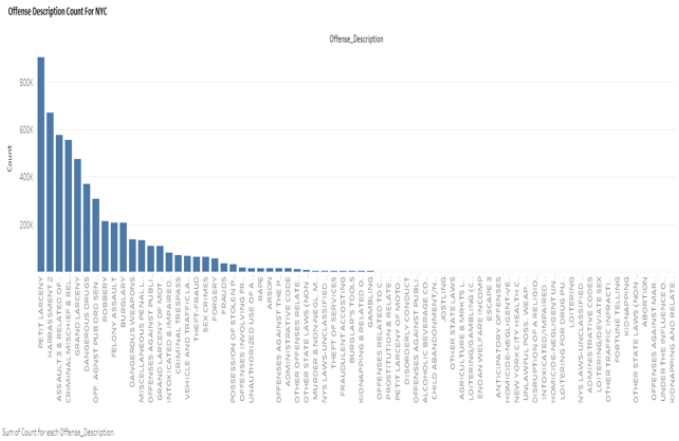
We started with plotting the crime count per million for each city over the year to understand the general trend. Since the graph shows crime per million the first thing that we observed here is that Crime in Chicago is quite worse as compared to New York and Los Angeles. The second important point that we observed is that there is a decreasing trend in crime, except for Los Angeles which has seen an increase in the crime count in recent years.



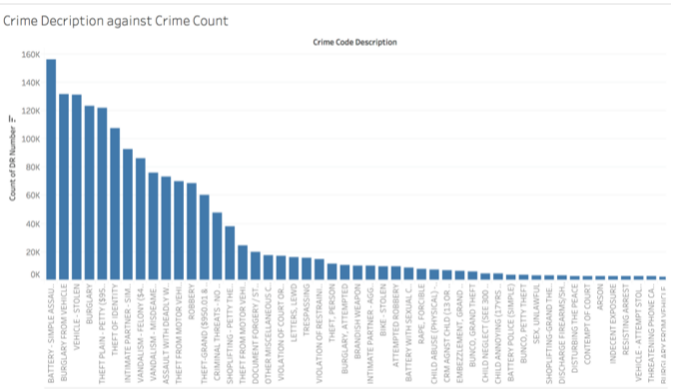
We plotted the crime category distribution for Chicago and observed that the top 5 crime for Chicago are: Theft, Battery, Criminal Damage,Narcotics and Other offense.



We plotted the crime category distribution for Chicago and observed that the top 5 crime for New York are: Petty Larceny, Harassment, Assault, Criminal Mischief and Grand Larceny.



We plotted the crime category distribution for Chicago and observed that the top 5 crime for Los Angeles are: Battery, Burglary-From Vehicle, Vehicle Stolen, Burglary and Theft.



But the main observation here is that all these cities have different definition of Crime, which becomes a challenge when we try to do a comparative study of crime trends across these cities.

5 COMPARATIVE STUDY

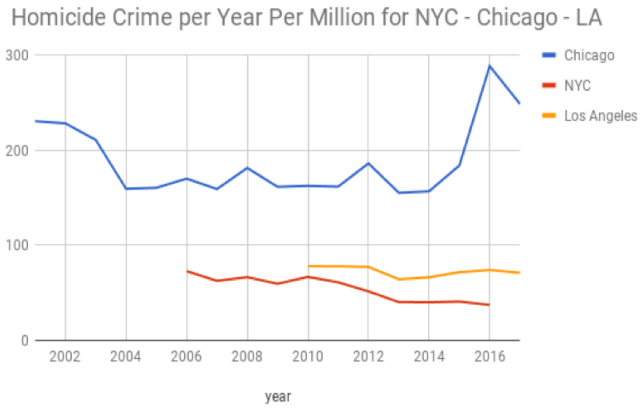
5.1 Major Challenge

As mentioned in the last section the crime description is different for every city and it was a challenge while doing a comparative study of Crime. To counter this we took reference from Uniform Crime Reporting which is a FBI Initiative to consolidate crime across a common definition. With the help of UCR definitions and the respective city crime department definitions we grouped certain crimes into one umbrella as shown below.

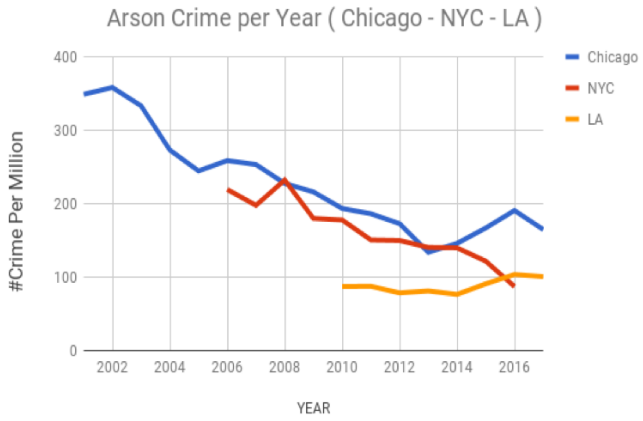
CHICAGO	NEW YORK	LOS ANGELES
ARSON	ARSON	ARSON
CRIM SEXUAL ASSAULT	LOITERING/DEVIATE SEX	CHILD PORNOGRAPHY
SEX OFFENSE	SEX CRIMES	HUMAN TRAFFICKING - COMMERCIAL SEX ACTS
		INCEST (SEXUAL ACTS BETWEEN BLOOD RELATIVES)
		SEX, UNLAWFUL
		SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ FRGN OBJ0059
		SEXUAL PENETRATION W/FOREIGN OBJECT
		SEXUAL PENTRATION WITH A FOREIGN OBJECT
		SODOMY/SEXUAL CONTACT B/W PENIS OF ONE PERS TO ANUS OTH 0007=02
ASSAULT	ASSAULT 3 & RELATED OFFENSES	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER
BATTERY	FELONY ASSAULT	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
		BATTERY - SIMPLE ASSAULT
		BATTERY ON A FIREFIGHTER
		BATTERY POLICE (SIMPLE)
		BATTERY WITH SEXUAL CONTACT
		CHILD ABUSE (PHYSICAL) - AGGRAVATED ASSAULT
		CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT
		INTIMATE PARTNER - AGGRAVATED ASSAULT
		INTIMATE PARTNER - SIMPLE ASSAULT
		OTHER ASSAULT
BURGLARY	BURGLARY	BURGLARY
		BURGLARY FROM VEHICLE
		BURGLARY FROM VEHICLE, ATTEMPTED
		BURGLARY, ATTEMPTED
HOMICIDE	HOMICIDE-NEGLIGENT	CRIMINAL HOMICIDE
	HOMICIDE-NEGLIGENT-VEHICLE	MANSLAUGHTER, NEGLIGENT
	MURDER & NON-NEGL. MANSLAUGHTER	
KIDNAPPING	KIDNAPPING	KIDNAPPING
	KIDNAPPING & RELATED OFFENSES	KIDNAPPING - GRAND ATTEMPT
	KIDNAPPING AND RELATED OFFENSES	

5.2 Results

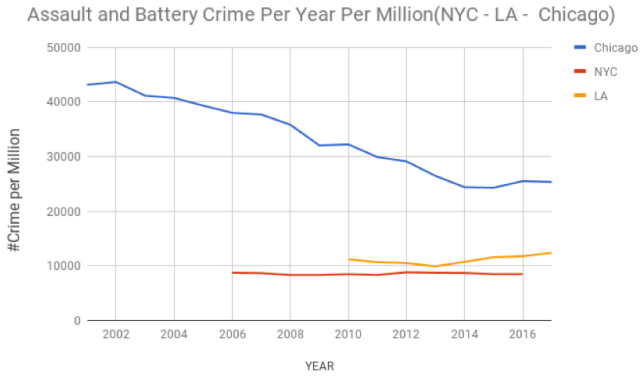
After making these grouping we started with plotting the results fro homicide to understand its trend. What we observed from this is that there was a sudden spike in the Chicago regarding homicide crime. After doing some background study about this event we found out that this might be caused by a video that was released by the Chicago Police Department which resulted in huge protests across the city.



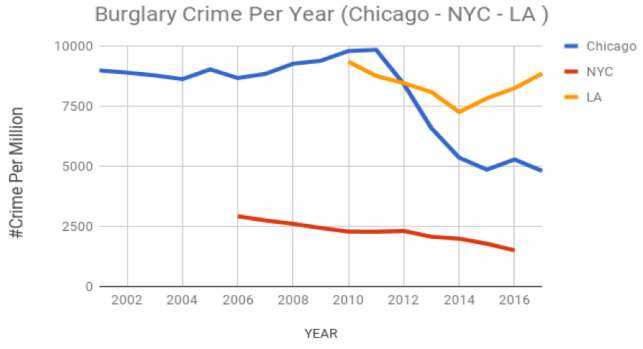
While plotting the trends for Arson we observed that one good thing to note is that the incidents of arson are on a decline, except in the case of Los Angeles, where not only Arson but the Crime count in general is on an increase. We believe this might be due to sudden increase in the population of Los Angeles over the years.



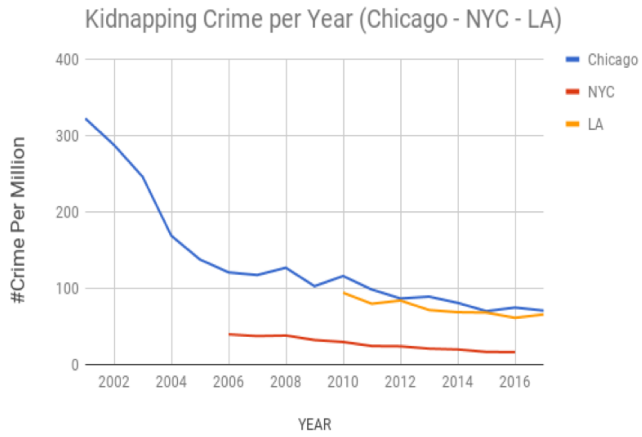
While plotting the trends for Assault we observed that one good thing to note is that the incidents of Assault are on a decline. But what we can clearly see is that Crime rate in Chicago in terms of Assault is way higher as compared to other cities.



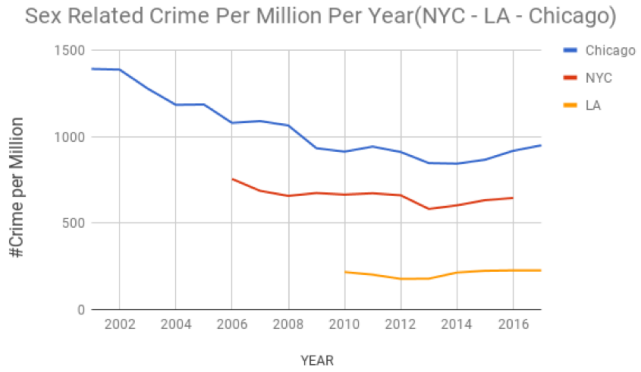
While plotting the trends for Burglary we observed that there is extremely good decline in Burglary rate in Chicago. But on the Contrary Burglary incidents in Los Angeles are the highest amongst all the three cities.



While plotting the trends for Kidnapping we observed that this might be the most positive trend as it has been on decrease in all the three cities. Especially the decline in the Chicago is recommendable.



While plotting the trends for Sex related we observed that on the contrary Sex Related crime in New York is comparatively high as compared to Los Angeles.

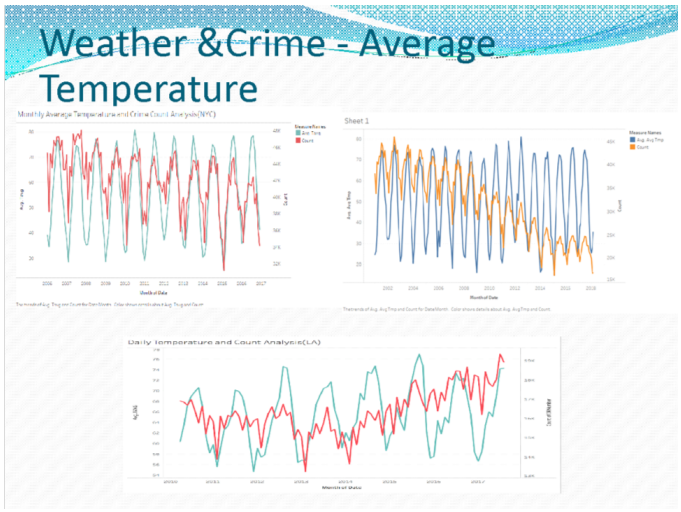


6 CRIME CORRELATION

For further in depth understanding of crime patterns, we are trying to find various factors that might affect crime. For this paper we have included 4 different datasets for study. Here we are studying weather (which includes temperature, wind, snow and precipitation: for all 3 cities), demographics (for NYC and Chicago), budget (for Los Angeles) and Restaurants (for NYC).

6.1 Weather and Crime

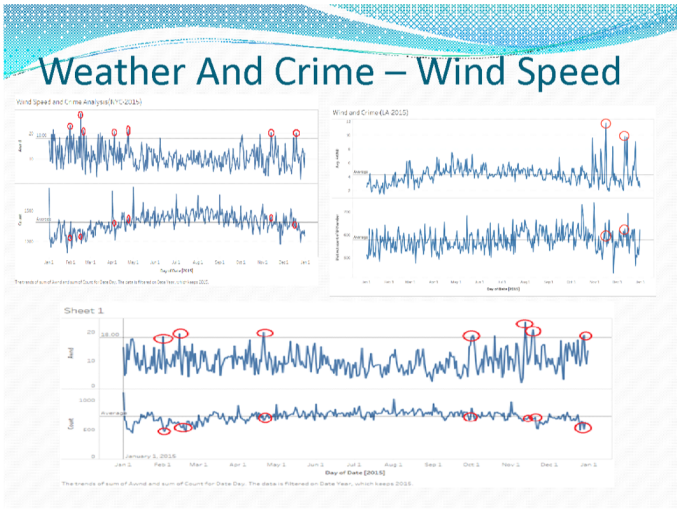
6.1.1 ***Average Temperature and Crime.*** Below shown are the plots of average temperature and crime for New York, Chicago and Los Angeles grouped together per month. An interesting pattern observed here is the ?peaks and the depths for average temperature and crime match?. Also this trend is continuous throughout the years and for all the cities. This is a strong indication that crime depends on average temperature. Evidently from the above graphs, we can hypothesize that with the increase in average temperature crime rises.



To prove the above hypothesis we here present p-value and a few correlation values for each of the cities. It is worth observing here is for all the cities we have a low p-value and a positive correlation. For New York and Chicago the observed values are significant but correlation values for Los Angles are weak.
With the evidence above we can conclude that crime is dependent on average temperature and is positively correlated.

	Chicago	Los Angeles	New York
P-value	<0.0001	<0.0001	<0.0001
Pearson corr.	0.37	0.152	0.408
Spearman corr.	0.35	0.256	0.418

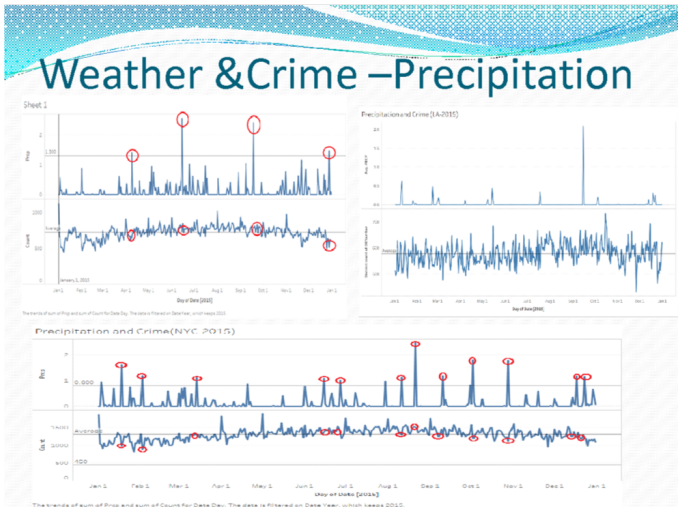
6.1.2 ***Wind Speed and Crime.*** Similar to average temperature we have plotted graphs for correlation between wind speed and crime. One key observation here is: for all days where wind speed is exceptionally high the crime for that day is lower than average crime per day. Hence we can hypothesize that with increase in wind speed there is a decrease in crime. This trend is evident from the above graphs and this hypothesis holds true for all the cities.



Below tables contains p-value and correlation for wind speed with crime. For all the cities, a very low p-value is observed. There is a negative correlation but there values are very weak and could not be used to bolster our hypothesis that wind speed leads to less crime.

	Chicago	Los Angeles	New York
P-value	<0.0001	<0.0001	<0.0001
Pearson corr.	-0.156	0.028	-0.208
Spearman corr.	-0.159	0.1028	-0.172

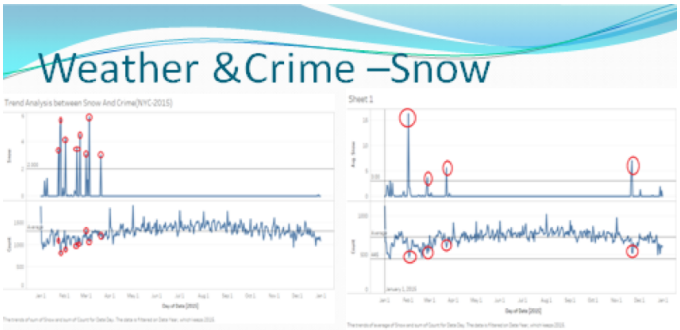
6.1.3 ***Precipitation and Crime***. Below shown is the graph of precipitation and crime. Similar to wind speed the key observation here is: for all days where precipitation is exceptionally high the crime for that day is lower than average crime per day. Hence we can hypothesize that with increase in precipitation there is a decrease in crime. This trend is evident from the above graphs and this hypothesis holds true for all the cities.



Below tables contains p-value and correlation for precipitation with crime. For all the cities, a very low p-value is observed. There is a negative correlation but there values are very weak and could not be used to support our hypothesis that precipitation leads to less crime.

	Chicago	Los Angeles	New York
P-value	<0.0001	0.03	<0.0001
Pearson corr.	-0.02	-0.039	-0.147
Spearman corr.	-0.62	-0.04	-0.129

6.1.4 ***Snow and Crime***. Below shown is the graph of snowfall and crime. Here we skipped Los Angeles because of there isn?t much snow in Los Angeles. Similar to wind speed and Precipitation, the key observation here is: for all days where snowfall is exceptionally high the crime for that day is lower than average crime per day. Hence we can hypothesize that with increase in snowfall there is a decrease in crime. This trend is evident from the above graphs and this hypothesis holds true for both the cities.
We also have p-value and correlation for snow and crime. For both the cities, a very low p-value is observed. There is a negative correlation but there values are weak and could not be used to support our hypothesis that precipitation leads to less crime.

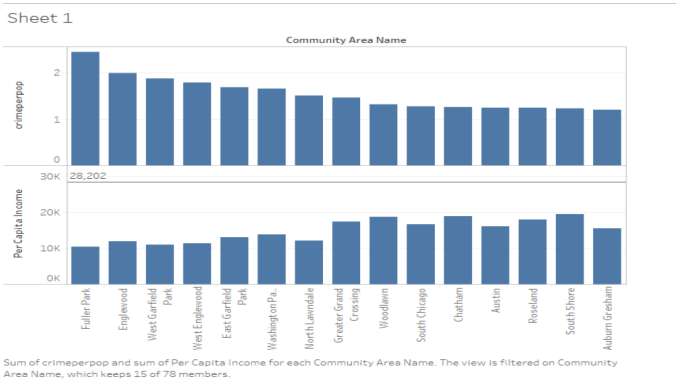


	Chicago	New York
P-value	<0.0001	<0.0001
Pearson corr.	-0.155	-0.223
Spearman corr.	-0.185	-0.218

6.1.5 ***Weather and Crime conclusion.*** From the above study we have observed that average temperature is positively and highly correlated with crimes whereas precipitation, wind speed and snowfall are negatively and weakly correlated with crimes. These results are found in all the cities which makes them concrete and reliable.

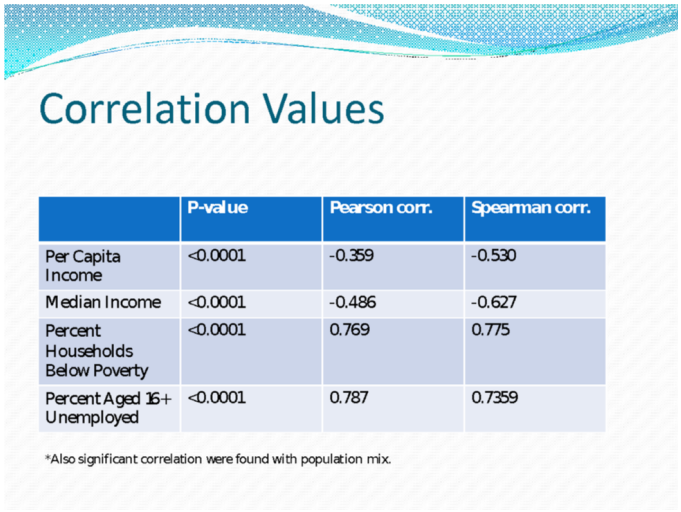
6.2 Demographics and Crime

6.2.1 ***Chicago.*** This section describes the study of correlation between demographics and crime. Chicago is divided into 77 community areas. This study is based on demographics data for each community area between years 2008-2012.
Following is the list of columns that were used for this study:
Area (in sq miles), Community Area Name, crimeperpop (crime per person), Community Area Number, Hardship Index, Per Capita Income, Percent Aged 16+ Unemployed, Percent Aged 25+ Without High School Diploma, Percent Aged Under 18 Or Over 64, Percent Households Below Poverty, Percent Of Housing Crowded, Asian (Percentage), Black (Percentage), Crime Count, Hispanic (Percentage), Median Income, Other (Percentage), Popdensity (population density), total Population, White (Percentage).
We calculated correlation values between all the features and crimeperpop (crime per person).



Sum of crimeperpop and sum of Per Capita Income for each Community Area Name. The view is filtered on Community Area Name, which keeps 15 of 78 members.

Above is a bar chart where the top plot crimeperpop and community area and the plot below is per capita income. The plot contains only first 15 community areas ordered descending w.r.t crime count. The crime in these 15 areas constitutes about 37 (Percentage) of the total crime in Chicago between the years of study.
The key observation here is for all these 15 community areas the per capita income is less than average per capita income in Chicago. These could be the poorest of neighborhoods in Chicago. Here we can hypothesize that areas with low per capita income are more prone to crime.



	P-value	Pearson corr.	Spearman corr.
Per Capita Income	<0.0001	-0.359	-0.530
Median Income	<0.0001	-0.486	-0.627
Percent Households Below Poverty	<0.0001	0.769	0.775
Percent Aged 16+ Unemployed	<0.0001	0.787	0.7359

*Also significant correlation were found with population mix.

Below is the table for a few more features and their correlation values. Each feature has a very low p-value and display high correlation.
As discussed with the plot, it is evident now that Per capita income has strong impact on crime and we can say both statistically and intuitively that areas with low per capita income are prone to crime. Also we can observe that number of households under poverty and unemployment rate have significant impact on the number of crimes happening in the area.

7 TECHNICAL DEPTH AND INNOVATION

7.1 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.

7.2 Technical Depth

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

7.3 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.

7.4 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.

7.5 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.

7.6 Tableau

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

8 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

9 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/