Crime Data Analysis of New York City

Anonymous

ABSTRACT

Crime has been a major concern to mankind and especially in metropolitan cities such as New York with a huge population, controlling it could be a challenging task. The New York Police Department regularly uploads its crime data in the data repository and has released this data to the public as part of the open data initiative. This study aimed to apply big data analysis techniques to understand this huge data and perform an in-depth study of this data to understand the trends in crime and the factors which affect it like time of the year, location, etc. Understanding such trends not only gives the police department a crucial edge to mitigate the crime but also understand the behaviors or patterns in crime to prevent it from taking place in the first place. This study can effectively help figure out the best way to use public resources to reduce crime levels. The model used by us not only has an exploratory but predictive analysis component that uses factors like place, time, and date into consideration. We have used various machine learning algorithms like K-Nearest Neighbor [13], Decision trees[11], etc to predict the crime. New York is a collection of various neighborhoods and has five boroughs: Manhattan, Brooklyn, the Bronx, Queens, and Staten Island and according to our study, Brooklyn has a high number of crime events while the Bronx which is considered a little unsafe has lesser crime number than Manhattan. But when crime per person is considered, Queens has the lowest average crime events We also found that Misdemeanor which has a jail time of no more than a year is the most common level of crime in every borough where Brooklyn had the highest number of such cases with 50.38 % of crime being a Misdemeanor. We also found that the second half of each year saw the most number of crimes with October having the highest crime rate. If we compare the crimes by victim's race then the Black victims were the highest and if compared by Gender then Females were subject to most crimes. Such insights can only come after a careful study of every important attribute which we wish to do with our analysis.

KEYWORDS

Crime Prediction; Crime Prevention; Crime Analysis; Prediction Methods; Data Analysis; Crime Mitigation

1 INTRODUCTION

Safety has always been a prime concern for the government. And with the increasing crime rates, it has become critical for the government to maintain law and order. The US government proposes The Open Data Government Act to make traffic, crime, and other related data accessible to the public[5]. This increases citizen participation which helps in economic growth and citizens involve in decision-making and uncover certain facts. New York City is one of the cities which has signed the Open Data initiative. The NYPD updates all the information on crime incidents that occurred within the confines of New York City. Millions of complaints had been recorded in the city-wise historical data which can be helpful to predict the crime in future.

The motivation behind taking up this topic for the project is to make the society threat safe. Although the country is becoming technologically advanced the regular occurrence of crime poses a threat to society and indulges a sense of fear within the citizens. The NYPD has made all the crime data available from the year 2006 to 2021 (current year) for public interest [1]. The crime cannot be predicted perfectly as it can be random, neither we can predict the victim but we can predict the location and the probability of occurrence of a particular level of crime. Therefore, taking into consideration the above facts, we will use big data analytics on the given historic data and analyze the patterns and trends in crime over the years. The predicted results might not be 100% accurate but they will surely help to reduce crime rates to some extent in NYC.

The basic approach to moving forward with the crime prediction and prevention model includes:

(1) Collection of sample data
(2) Data Preprocessing
(3) Pattern Recognition
(4) Prediction
(5) Visualization

The project is based on the above approach and will be implemented using Python. It is divided into two major components: Data Management and Data Analytics. For the data management part, we will use the MySql database to store the dataset that has approximately 7.5M instances and 35 features which contains all NYC complaint data. Although we can use Apache Spark for managing the real-time data. Apache Spark is used to process a big amount of data parallely which increases the performance of the application. But in this project it will not make much difference as the size of the data is not so big.

Data cleaning and transformation will be performed while taking important features into consideration and then filtering the data accordingly. The preprocessing steps would be:

(1) The data set which is an Excel file is converted into a .csv file
(2) Then this .csv file is converted into a table and loaded into structured database using MySql Query Browser
(3) The Redundant columns and rows,including null values are either dropped or replaced with a significant value using Sql queries

Data mining highly depends on the quality of input data. Missing information, redundancy, noise and outliers may lead to deviation of predicted result from the actual result. Thus, Data Preprocessing is an important component of the project. After the Data Preprocessing, we will use the improved data for data analytics.

Table 1: NYC CRIME HISTORIC SAMPLE DATASET

CMPLNT NO.	DATE	TIME	OFNS DESC	PREM TYPE	LOCATION
268384942	03/23/2012	20:40:00	Harrassment	Street	(40.814630053°,-73.908579349°)
922946565	01/19/2016	18:45:00	Petit Larceny	Street	(40.824479305°,-73.862728733°)
884332860	06/05/2014	01:56:00	Dangerous Weapons	Street	(40.667842675°,-73.894833888°)
832276444	05/09/2018	00:01:00	Theft fraud	Inside	(40.840016963°,-73.783794721°)

The sample dataset provided in Table1 is very specific to the spatial and temporal details and gives an overview of the date, time, location, and the type of offense. However, the original dataset chosen has 35 attributes and gives us detailed information about the crimes. The dataset has the potential to provide a much wider perspective about the crime and predict different patterns. The predictive analysis would have been a difficult task if the data was not provided with quality attributes which makes it rich.

The Data Analytics component will deal with all the visualization and data mining algorithms. This component focuses on two distinct parts:

(1) Exploratory Data Analysis
(2) Predictive Modelling

For the first, we intend to perform exploratory data analysis to mine patterns in crime. This analysis will take salient features into consideration and observe different existing patterns in crime throughout NYC. We can determine the crime spread in NYC considering various factors such as the levels of offense, suspects age group, race, and sex and how does it vary with the type of crime committed over a period of time. There are details regarding the crime such as the exact date and time of occurrence, the precinct in which the incident occurred, and the location coordinates. With the use of this data set, we aim to visualize any patterns and analyze the spatial and temporal relationships in the crimes being committed. Both spatial and temporal aspects will aid in determining the frequency of crime in a particular region of NYC. Visualizing all this data will easily assist to find any outliers and help towards our end goal which is to predict the crime that can take place in the future as that would be beneficial to the enforcement agencies.

Later for building a prediction model, we also intend to use classification techniques to segregate the crime data into different class labels or categories so that they can be analyzed in their own segment. As there are multiple categories of crimes we believe that a Multiclass Classification will best suit the data as in this method the data is split into multiple categories. One example of how we could make use of this is by classifying the different crime hotspots in the city or classification based on the time and day of the crime as this can provide helpful insights.

We can use the Decision tree classifier as it tests the input against only specific subsets of data, which is determined by the splitting function. Hence a lot of redundant computations can be removed. Also, we can use the feature selection algorithms to decide which features of the data are worth selecting for the classifier which makes this algorithm very flexible to use.

Using the k-Nearest Neighbor algorithm [13] we can classify our data into multiple categories by looking at the majority vote of its neighbors. Here we group those data points together that are closest together and then we predict the label of these points. The number of neighbors can be defined by the user as K or it can be based on a particular radius specified.

There are many data mining techniques available but after testing each model, a suitable technique will be chosen that will provide us our expected results. The final goal of this analysis is to help understand the spread of crime in terms of spacial and temporal factors. All these analysis will be beneficial for NYPD for the prediction of all the possible victimization areas in NYC and further helps them to mitigate the risk of a crime.

The report is further organized as follows. The next Section 2 discusses about the project goals. Most the research work related to crime analysis and our project is referred in Section 3. Section 4 describes the design and presents the architecture of our project. A detailed analysis about the crime is covered under this Section 4.4. All the Legal and Ethical issues related to our project have been considered under Section 5. Section 6 discusses about the lesson we learned from the project and the current and future scope of the project is presented in Section 7. The report is finally concluded with remarks in Section 8.

2 PROJECT GOALS

The fundamental goal of our project is to build a model that can predict the crime based on different categories such as the location, the time of occurrence, the level of crime. The model can be used to analyze the historical criminal data using descriptive or predictive analysis techniques. The descriptive analysis focuses on mining crime patterns and also tends to find a spatial and temporal relationship in the crime data. The predictive analysis helps in predicting the level of crime that is the type of crime that can be predicted on a given location at a certain time. All these studies on temporal and spatial features of criminal records from the real world historic dataset aids in predicting the most likely crime locations or hotspots. In addition, we intend to provide a detailed analysis of all the various attributes and provide a statistical and graphical relationship between the main features of the huge dataset.

3 RELATED WORK

In this section, we discuss about topics that are highly related to crime analysis. There have been a lot of studies involving the use of big data analytics to predict and analyse the crime patterns. Observing patterns in crime is complex and requires fair knowledge in data mining and analysis techniques. Some of these techniques used by researchers involve one or more datasets. Many researchers have devoted their time and attention to study crime and the various factors involved in it. They take into consideration factors like historical records, spatial and temporal patterns [16]. Musa et al. studies the relation between employment rate and crime in a smart city [9]. Other factors also revolve around crime such as education [4] ethnicity [2], and income level of an individual. The researchers have studied the crime prediction dynamics through three perspectives which can be categorized into three paradigms: time-centric, place-centric, and people-centric.

The time-centric study focuses on the temporal dimension of the dataset. The authors of [7] have studied the variations of crime throughout the year to see if there exists a pattern or a trend with the change of time. They studied data from different Canadian cities and discovered that crime peaks at a certain hour of the day. [8] The authors implemented a self-exciting point process model to study the temporal trends in the rate of burglary. In another study [12], researchers analyze the temporal constraints in environmental criminology. The author in their research have discovered Spatio-temporal patterns in the crime [10]. Oliveira et al. studied the one-year cycle of criminal events and observe a seasonal pattern in crime and provided findings on how the crime wave keeps traveling over time.

The place-centric paradigm deals with the spatial dimensions for predicting the location of the crime. Many kinds of research show a relation between temporal and spatial dimensions. Zhao et al. explore the Spatio-temporal crime patterns in the urban data [20]. The researcher mapped temporal patterns and spatial patterns to predict the crime. Wu et al. introduces a system that automatically mines crime-logged data through alert messages to make a university premise more secure by predicting the crime [19]. Toole et al. studied the correlation between spatial and temporal patterns based on criminal offense records [14]. All the above studies aids in predicting the location of crime incidents and there are plenty more on the exploration of the crime hotspots. The author focuses on crime places in contrast with the neighborhoods [18]. The place level explanation justifies the crime events and the neighborhood theories usually highlight the development of offenders. Emre et al. focuses on the statistical significance of crime hotspots and categorizes the two detection approaches like linear and circular hotspots detection [3].

The people-centric research focuses on an individual or collective criminal profiling. Wang et al. proposes a pattern detection algorithm called Series Finder [17]. The main aim of the study was to reduce the efforts of the analysts to look for each record and check for individual criminality. The algorithm proposed to take insights from the historic crime data and automatically identifies the crime committed by an individual or a group of offenders.

All the traditional approaches use demographic and geographic datasets to predict the crime hotspots. A lot of research in crime prediction and analysis requires a wide range of datasets. According to a study by Wang wt al. using newer features improves the accuracy of the prediction [15]. They used large-scale Point-Of-Interest data in the city of Chicago, IL, and observed a significant improvement in the performance of crime prediction. Moreover, considering new features has also proved to have a significant impact on finding a correlation between many other features.

In our proposed approach, we study the temporal and spatial patterns within the criminal records in NYC. The work described above shares a few similarities with our work such as visualization of the data, classification of crime, and prediction of crime hotspots in New York City. Data mining techniques are used to show statistical and graphical analysis of the relationship between the various features of our huge dataset. Machine learning algorithms are deployed to predict the trends in crime and provide a model with the highest accuracy.

4 DESIGN AND IMPLEMENTATION

The main objective of this project is to study the NYC crime dataset and find a relationship between the transient and spatial features of the dataset. Descriptive analysis aids in finding trends in crime while predictive analysis will help in predicting the criminal hotspots. Using the above analysis a model is structured to predict the crime category that is more like to happen in a given time range and at a particular location.

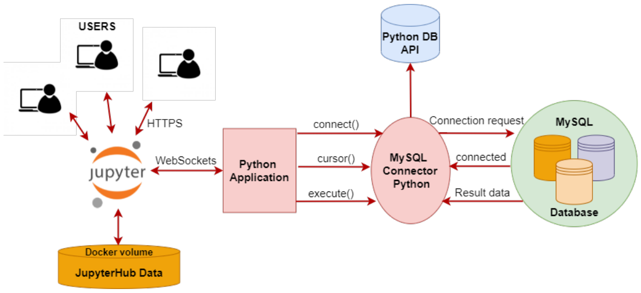

Figure 1: A high-level Architecture Diagram.

The architecture of our system is illustrated in figure 1. The diagram shows a high-level description of our proposed project idea. The user interacts through the web browser and sends HTTPS requests or establishes a connection via Websockets to the Jupyter Notebook server. The IPython kernel is the core part of the Jupyter ecosystem and handles the computation and communication within the user and frontend interfaces like notebooks. When a user request is made by the IPython kernel, the Jupyter Hub spawns, manages, and proxies the request of a single user. The Jupyter Hub is a multi-user Hub and can be run on many different infrastructures.

The python notebook implemented on the Jupyter web application is further used to establish connections between the python application and MySQL database. Python DB API has all the python database access modules such as MySQL-python connector, sqlite3, etc. To establish the connection a MySQL-python connector is used. The connection request is sent to the MySQL connector using the credentials and the MySQL connector further forwards the request to the MySQL database. After the connection is established a database is created on MySQL and the data is stored, retrieved, and update through SQL queries.

4.1 Overview of Dataset

The dataset used for the project is NYPD Historic Complaint Data from 2006 till date. Our data is publicly available by New York City Police Department on the NYC Open Data website[include citation-Open Data]. The detailed analysis of the dataset is as follows:

- CMPLNT_ NUM: It is a randomly generated numerical persistent ID for each complaint
- CMPLNT_FR_TM: It is a text field and describes the exact time of occurrence for the reported event.
- OFNS_DESC: It is a text field and provides a description of offense corresponding with key code.

- CRM_ATPT_CPTD_CD: It is a text field and an indicator of whether a crime was successfully completed or attempted, but failed or was interrupted prematurely.
- LAW_CAT_CD: It is a text field and describes the Level of offense: felony, misdemeanor, a violation.
- BORO_NM: It is a text field. The name of the borough in which the incident occurred.
- PREM_TYP_DESC: It is a text field. A specific description of premises where the crime took place; grocery store, residence, street, etc.
- JURIS_DESC: It is a text field. Description of the jurisdiction code like NY Police dept, NY housing police, etc.
- X_COORD_CD: It is a numerical field. X-coordinate for New York State Plane Coordinate System.
- Y_COORD_CD: It is a numerical field. Y-coordinate for New York State Plane Coordinate System.
- SUSP_AGE_GROUP: It is a text field about Suspect's Age Group.
- SUSP_RACE: It is a text field about the Suspect's Race Description.
- Latitude: It is a numerical value. Midblock Latitude coordinates for Global Coordinate System.
- Longitude: It is a numerical field. Midblock Longitude coordinates for Global Coordinate System.
- Lat_Lon: It is a location field. Geospatial Location Point
- VIC_AGE_GROUP: It is a text field and describes Victim's Age Group.
- VIC_RACE: It is a text field and describes the victim's Race Description.
- VIC_SEX: It is a text field and describes the victim's Sex Description.

The raw dataset consists of 35 attributes and around 7.5 million criminal records. The size of the data is around 2.5GB. The dataset contains records from the year 2006 to 2021. The dataset is further processed to increase the data quality and data integrity.

4.2 Data Preprocessing

Before implementing the machine learning algorithms on the dataset to predict the crime hotspots, we first need to preprocess our data. Preprocessing the data will increase the data quality and improve the accuracy of our analysis. In the data processing step, the original dataset is modified by removing null values, replacing the unknown values, removing a few attributes, and transforming the data by adding a new column. The new dataset has 21 features and around 2 million records.

4.2.1 *Data Cleaning.*

The primary step of data cleaning is to take care of the missing data from the dataset. We find the missing percentage of each attribute and removed the attributes with the missing percentage of more than 80%. Further, we either substituted the null values with unknown or removed the records with null values depending upon the type of specific feature.

We performed outlier handling in the next step. Several features were selected from the dataset to check for the outliers. The criminal records were filtered by year and only records between 2006 till 2020 were taken into the new dataset. Moreover, we considered features like the suspect's age group, sex, and race and removed the outliers from the dataset. The same is done for victim-related attributes.

For the primary key, we check for the uniqueness of the data. According to our dataset, the complaint number must be unique to fetch data through the case number. Therefore, we removed the duplicate entries from our dataset so that the complaint number can further be used as a primary key and also improves the data quality.

4.2.2 *Data Transformation.*

One of the most important data preprocessing techniques is data transformation. The structure of the data is changed to increase the readability of the data and to facilitate analysis. In one of the attributes, the year is extracted from the date and added as a separate new column for easy analysis. Hence, the date field is used to generate the year feature which is useful for predicting the crimes over the years.

4.2.3 *Data Reduction.*

As we are trying to predict the crime and study the relationship between transient and spatial features, so it is essential to have only those features into account. We can remove features like Transit district, park name, housing PSA, and station name as these features are not related to our study about the temporal and spatial features and do not provide any additional information to help us achieve our goal.

4.3 Software and Technologies Used

The following tools and technologies have been used to implement our project.

- Language: Python v3.9.6
- Database: MySQL v8.0
- Tools: Jupyter Notebook, MySQL Workbench 8.0

The project is implemented using Python version 3.9.6. Various python libraries are used throughout the project and are described below:

- Pandas: It is an open-source library and used for fast data manipulation and indexing. The data is stored in pandas data frame object which helps to prepare data and helps in machine learning algorithms.
- NumPy: It a python library that provides features for handling multidimensional arrays and perform scientific computing in Python. As pandas is built on top of NumPy, it is used as a collaborator for pandas to perform their mathematical functions.
- MatPlotlib: It is a plotting library used for statistical and graphical analysis of the dataset. It is useful in plotting interactive data visualizations such as bar charts, line charts, scatter plots, etc.
- Seaborn: It is a statistical data visualization library built on top of MatPlotlib. It presents data visualization plots that are more attractive and informative for data analysis.
- Scikit-learn : It is a very useful machine learning library in Python to implement the various machine learning algorithms. It provides various regression, clustering, and classification algorithms and can be significantly used in prediction models.

4.4 Data Analysis

In this section, we focus on two major components of data analytics that is the Exploratory Data Analysis (EDA) and Predictive Analysis. Several graphical techniques are used for effective data visualization such as scatter plots, bar charts, line charts, etc. We used the Pandas and Seaborn library to plot more attractive and informative graphs for exploratory data analysis. These graphs further help in mining patterns in crime and make our analysis easier.

We have divided our exploratory analysis into three parts:

(1) Crime Information based on Date & Time. By understanding how crime varies with day and time. We can gain useful stats about the happenings of crime. We plotted our graphs based on the time, month, and year of crime and these were the outcomes of it.

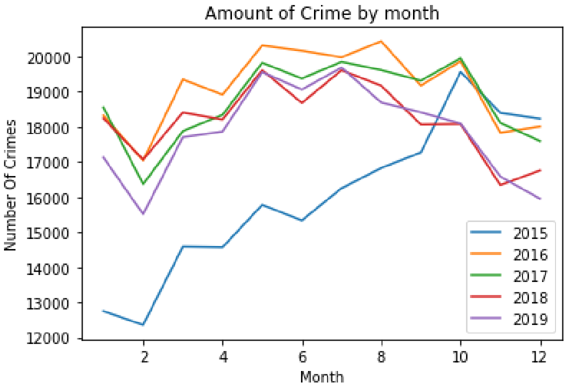

Figure 2: Monthly Crime Rate

Figure 2 plot tells us the variation in crime rates from 2015-2019. The line of 2015 is most inclined as it starts from the lowest crime and reaches above all by the end of the year. Year 2016 shows maximum number of crime almost throughout the year while year 2019 is the only year in which the crime rate declined after mid of the year and reached the minimum of all.

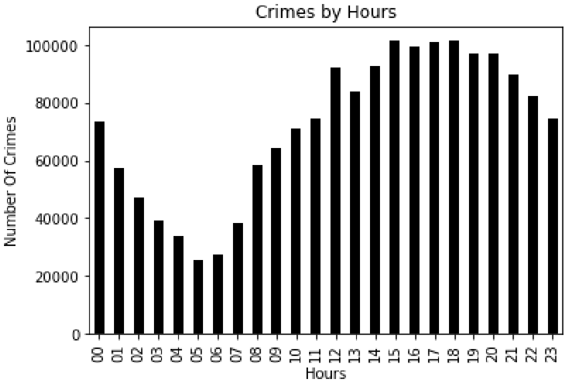

Figure 3: Hourly Crime Rate

Figure 3 illustrates the crime rate in hourly manner. Points worth noticing are that the maximum crimes occur during day time around 2PM to 4PM which was least expected as most people are awake and active during this time and surprisingly crime rate reduces at night continuously till 5AM. The safest time of the day is from 3AM to 7AM at which the crime rate is lesser comparatively.

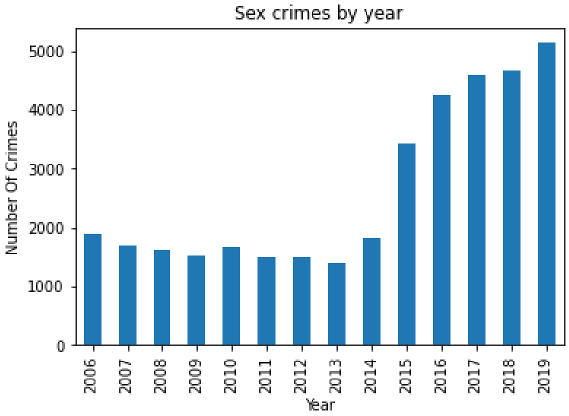

Figure 4: Yearly Crime Rate

Figure 4 tells us the variation in crime rates from 2015-2019. The line of 2015 is most inclined as it starts from the lowest crime and reaches above all by the end of the year. Year 2016 shows maximum number of crime almost throughout the year while year 2019 is the only year in which the crime rate declined after mid of the year and reached the minimum of all.

(2) Crime information based on Race and gender. By understanding the crime variation we can figure out the minority or majority of crime victims by their Race or gender.

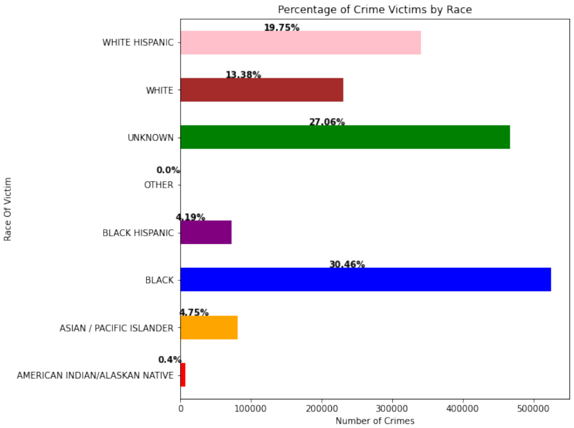

Figure 5: Crime Rate on basis of Race

Determining the crime on basis of victims as shown in figure 5 by race we noticed that black people are more prevalent towards crime while American Indian/Alaskan native are least prone to crime. This difference is also because oh the difference of population in two groups as number of American Indian is far lower than Black people. Number of victims of White Hispanic and White people are in between of black and American Indians.

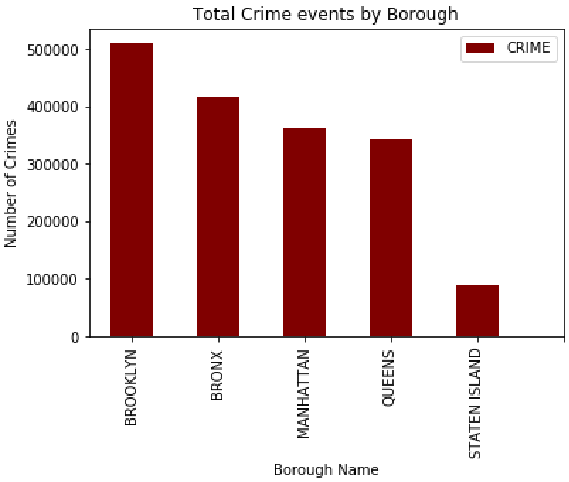

Figure 6: Crime Rate in each Borough

(3) Crime information based on borough and Level of crime. By understanding the borough crime rates we can figure which areas have higher crime rates and what is the level of crime seen over there. Hence this information helps us make our predictive model efficient.
As per the graphs in figure 6 it can be concluded that Staten Island has least occurrence of crime while Brooklyn tops the list followed by Manhattan, Bronx and Queens.

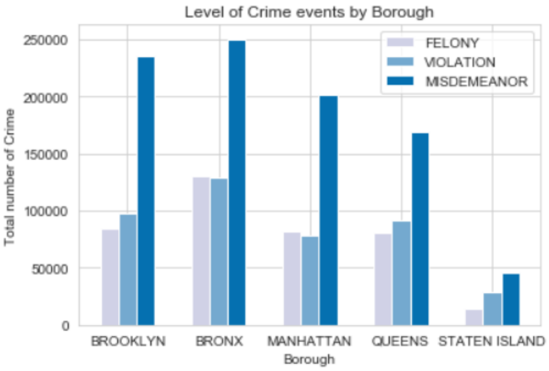

Figure 7: Level of crime in each Borough

Crime is classified by the level into three categories namely felony , misdemeanor, and violation amongst which felony is most serious and violation is least severe of the crimes. From figure 7 we can conclude that in all the boroughs Misdemeanor is most prevalent while number of crimes of felony and violation are almost equal to each other and almost half to the misdemeanor. Staten Island has least number of crime incidents across all boroughs while the maximum number can be seen in Bronx borough.

We take the above exploratory data analysis into consideration and tends to find patterns in the dataset. With the help of above data visualizations, relationship between temporal and spatial features of the dataset can be used to study the trends in crime and further a model can be built to predict the crime.

For the predictive analysis, we will further analyse the data visualizations and perform various analysis on the features to further build our model only on the relevant features of the dataset. For feature selection we can use following techniques:

- Pair plots : There are various attributes in our data which are correlated with each other and selecting the best ones is the first step in any kind of analysis. We will use a pair plot from the Seaborn library which tells us how correlated different attributes are. If two attributes are highly correlated ,then one can be removed from the predictive analysis as each one of them have the same effects in our model.
- Correlation Matrix: A correlation matrix is a matrix which is symmetric and shoes the relation between various attributes using numbers rather than graph format. This will be used to find how correlated our variables are and we will further use a heat map style matrix to understand the correlation.

After feature selection, a model will be built to predict crime based on various factors like location, time, neighborhood, etc. Various machine learning algorithms will be used to train the model. All the transient and spatial factors will be considered by an algorithm and a suitable model with high accuracy will be built.

5 LEGAL AND ETHICAL CONSIDERATIONS

Privacy is a fundamental right and while working on the NYC open data, we made careful considerations that our algorithms do not affect any race, gender, community, or life form in any manner. We have made sure that the privacy of each individual is maintained by not mentioning any particular person's name, address, or contact thus maintaining utmost discretion with such elements of our data. The machine learning models used are free of any form of bias by randomly sampling the data for every model we use so that no particular attribute is given more emphasis thus creating a non-biased and fair model to predict different crimes. As we have used the k-fold cross-validation technique [6], every data point gets added in a validation set exactly once, and gets added in a training set k-1 times. This significantly reduces bias as we are using most of the data for fitting, and also reduces variance as most of our data is also being used in the validation set. The data used is owned by NYC open data and any other third party, all rights are reserved. We ensure that the data will not be used for any commercial or profitable purposes. Also, any of our findings through this paper does not give out final conclusions about any person's guiltiness or innocence as only the Court holds the right to prosecute.

6 LESSONS LEARNED

In the making of this project, we have come across various aspects of both data management and data analysis and the handling of both sides of the coin in data science gave us a holistic growth of knowledge in the field of data science and big data. In the data management part, we learned how to store, retrieve and access big data using RDBMS and algorithms used for batch processing. We learned how to keep our data consistent and maintain its quality by proper store procedures. We also managed to interface with different technologies like python and MySQL and were able to retrieve and work on our data in an efficient manner. For the data analysis part of our project, we learned how to use various visualization tools such as PANDAS and MATPLOTLIB. We learned data cleaning and preparation by interpolation and filling null values. We learned how to correlate our data and use it efficiently use it in machine learning algorithms to make predictions about crime. Lastly, we learned about data privacy issues and how we have to be mindful about the use of our algorithms as they can impact the lives of different people.

7 CURRENT STATUS & FUTURE WORK

We have done our data analysis on 1 million rows of data and we have been accomplished this by batch processing this data using the MYSQL limit feature. We figured out ways to only take a subset of data which was sufficient for the application of data mining algorithms. The data was cleaned by removing any null values by either taking the mean of the available data and replacing it or by filling it with 0. We have also applied data mining algorithms on the data like k-nearest neighbors to predict the crime in various neighborhoods of New York. We believe that in the future we can do analysis on the complete 7.5 million data as the current data consistency didn't allow us to do so. This data can be uploaded on big data handlers like Hadoop to increase efficiency in the future using map-reduce. Our main goal was to provide insights into the trends in crime so that law enforcement and the general public can be made aware and hence this model can be replicated for crime data of different cities and countries in the world by other developers to help mitigate crime.

8 CONCLUSIONS

Crime data analytics using big data technologies helps us to convert plain data into meaningful information which can be acted upon by the judiciary and law enforcement agencies. Crime prediction and detection play a major role in keeping our societies safe and secure and thus using past data about crime helped us predict the future trends in crime. With the increase in mass shootings, gun-law abuse, and police brutality cases, studies like such can help understand and acknowledge the shortcomings of our systems and laws. Using the visualization and prediction tools on big data helps us to make an informed decision about how to control and mitigate crime in our cities. For example, if we could predict that street racing is common in a particular area at some given point of time every day then surveillance in that spot can be increased for that particular duration. Such moves and tactics can give our police forces an advantage over the wrongdoers and make our cities more peaceful.

REFERENCES

[1] Police Department (NYPD). [n.d.]. NYPD Complaint Data Historic. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i, year = 2021, howpublished= NYPD.

[2] John Braithwaite. 1989. *Crime, Shame and Reintegration.* Cambridge University Press. https://doi.org/10.1017/CBO9780511804618

[3] Emre Eftelioglu, Shashi Shekhar, and Xun Tang. 2016. *Crime hotspot detection: A computational perspective.* IGI Global, 82–111. https://doi.org/10.4018/978-1-5225-0463-4.ch004 Publisher Copyright: © 2016 by IGI Global. All rights reserved. Copyright: Copyright 2017 Elsevier B.V., All rights reserved.

[4] Isaac Ehrlich. 1975. On the Relation between Education and Crime. In *Education, Income, and Human Behavior.* National Bureau of Economic Research, Inc, 313–338. https://EconPapers.repec.org/RePEc:nbr:nberch:3702

[5] Federal Government of United States. [n.d.]. OPEN GOVERNMENT. https://www.data.gov/open-gov/.

[6] Jason Brownlee. 2021. K-fold Cross validation. https://machinelearningmastery.com/k-fold-cross-validation/.

[7] Shannon Linning, Martin Andresen, and Paul Brantingham. 2016. Crime Seasonality: Examining the Temporal Fluctuations of Property Crime in Cities With Varying Climates. *International Journal of Offender Therapy and Comparative Criminology* 61 (03 2016). https://doi.org/10.1177/0306624X16632259

[8] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. 2011. Self-Exciting Point Process Modeling of Crime. *J. Amer. Statist. Assoc.* 106, 493 (2011), 100–108. https://doi.org/10.1198/jasa.2011.ap09546 arXiv:https://doi.org/10.1198/jasa.2011.ap09546

[9] Sam Musa. 2018. Smart Cities-A Road Map for Development. *IEEE Potentials* 37, 2 (2018), 19–23. https://doi.org/10.1109/MPOT.2016.2566099

[10] Marcos Oliveira, Eraldo Ribeiro, Carmelo Bastos-Filho, and Ronaldo Menezes. 2018. Spatio-temporal variations in the urban rhythm: the travelling waves of crime. (09 2018). https://doi.org/10.1140/epjds/s13688-018-0158-4

[11] Prashant Gupta. 2021. Decision Trees. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

[12] Jerry H. Ratcliffe. 2006. A Temporal Constraint Theory to Explain Opportunity-Based Spatial Offending Patterns. *Journal of Research in Crime and Delinquency* 43, 3 (2006), 261–291. https://doi.org/10.1177/0022427806286566 arXiv:https://doi.org/10.1177/0022427806286566

[13] Srishti Sawla. [n.d.]. K-nearest algorithm. https://medium.com/@srishtisawla/k-nearest-neighbors-f77f6ee6b7f5, year = 2021,.

[14] Jameson L. Toole, Nathan Eagle, and Joshua B. Plotkin. 2011. Spatiotemporal Correlations in Criminal Offense Records. *ACM Trans. Intell. Syst. Technol.* 2, 4, Article 38 (July 2011), 18 pages. https://doi.org/10.1145/1989734.1989742

[15] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. 2016. Crime rate inference with big data. In *KDD 2016 - Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining).* Association for Computing Machinery, 635–644. https://doi.org/10.1145/2939672.2939736 Funding Information: The work was supported in part by NSF award 1544455, 1054389, and funding from NICHD R24-HD044943. The views and conclusions contained in this paper are those of the authors and should not be interpreted as reprinting any funding agencies Publisher Copyright: © 2016 ACM. Copyright: Copyright 2017 Elsevier B.V., All rights reserved.; 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016 ; Conference date: 13-08-2016 Through 17-08-2016.

[16] Shuai Wang, Xiao Wang, P. Ye, Yong Yuan, Shuo Liu, and F. Wang. 2018. Parallel Crime Scene Analysis Based on ACP Approach. *IEEE Transactions on Computational Social Systems* 5 (2018), 244–255.

[17] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. 2013. Learning to Detect Patterns of Crime. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 515–530.

[18] David Weisburd and Cody Telep. 2010. The Efficiency of Place-Based Policing. *SSRN Electronic Journal* 17 (07 2010). https://doi.org/10.2139/ssrn.2630369

[19] Shela Wu, John Male, and Eduard Constantin Dragut. 2017. Spatial-temporal campus crime pattern mining from historical alert messages. *2017 International Conference on Computing, Networking and Communications (ICNC)* (2017), 778–782.

[20] Xiangyu Zhao and Jiliang Tang. 2017. Exploring Transfer Learning for Crime Prediction. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW).* IEEE, 1158–1159.