Crime in New York City

Riju Khatri
Computer Science
New York University
New York, USA
rk3766@nyu.edu

Niharika Sinha
Computer Science
New York University
New York, USA
ns4451@nyu.edu

Shobhit Sinha
Computer Science
New York University
New York, USA
ss13881@nyu.edu

***Abstract*—With the diverse population of New York City, and varying trends in schooling, health, and consumption of alcohol and drugs across NYC's five boroughs, predicting crime is a challenging problem. This project aims to throw insights into how each of these boroughs differs across these aspects and find trends between what are the significant factors affecting each of the different kinds of crime happening around. The authors combine crime, mental health, alcohol, and school data to make predictions of the types of crimes happening at each borough level.**

***Index Terms*—New York City, crimes, arrests, borough, alcohol, drugs, mental health, classification**

I. Introduction

Having a crime index of 19, New York City is safer than 19% of American cities. A 1000 people, there are 25.80 crimes committed. There are 5.80 violent crimes per 1,000 persons on average, which is 1.80 more than the average for the country. According to statistics, the most frequent crime 20 years ago was burglary. Over the next ten years, it evolved into robbery, and today it is criminal assault. Many elements of these crime statistics may have been updated as a result of changes in the city's demography, judicial system, and population.

In recent years, the most commonly misused substances in New York City have been alcohol, heroin, cocaine, marijuana, and prescription opioids. Approximately one in twelve people in New York City seeking treatment for drug abuse and addiction named cocaine or crack as their main substance of abuse, according to reports from 2015. Heroin was the most commonly mentioned substance of abuse, and it has long been a problem in New York City. On the other hand, marijuana has seen its greatest rise in recent years. The effects of binge drinking and high alcohol intake were examined in a report by the Behavioral Risk Factor Surveillance System (BRFSS). A higher risk of developing a number of chronic diseases and ailments is linked to excessive alcohol consumption, whether it takes the form of heavy drinking or binge drinking. Numerous cancers, including those of the liver, colon, rectum, oral cavity, and pharynx have been associated with excessive alcohol consumption. According to research, a person's chance of acquiring alcohol-related cancer increases with the amount of alcohol they consistently consume over time.

The prevalence, commonality, and disability of mental illnesses persist. 1 in 10 adults and kids encounter mental health issues each year that are severe enough to impair functioning in daily life, including at work, at home, and in school. Nine years on the average pass in our country before someone seeks medical attention. As a result, there are fewer cases that are formally registered for mental health programs, which means that there is a lack of statistics in this area.

The dispute over school budget cuts in New York City has become a contentious issue in recent years, with decreased student enrollment being a significant subtopic of discussion. Since the pandemic started, K–12 enrolments have decreased by 9.5% overall. Before the sharp declines during the pandemic, enrollment numbers had been decreasing gradually since the 2015–16 school year. The accessibility, quality, and affordability of the educational options available have a significant impact on student enrollment.

In order to study the inter- and intra-domain patterns of these domains and their correlation to crime, this paper thoroughly examines these three domains.

The portions of the paper are as follows. The authors' review of the literature is discussed in Section II. The datasets used are then described in Section III. Section IV digs deeply into the paper's approach. Section V discusses the outcomes after that. The authors address their conclusions in Section VI, while Section VII talks about the work's future scope followed by Acknowledgments.

II. Literature Survey

The paper [1] focuses on evidence that suggests that policies designed to increase educational attainment and improve school quality can significantly reduce crime rates. It discusses evidence about the effects of educational attainment, school type, and school quality on subsequent criminal outcomes, and the ways in which altering youths' School attendance is likely to affect their contemporaneous engagement in crime. It finally concludes with discussions of important policies regarding education and crime. The paper [2] used the dataset from the Police Department of Kayseri. The idea is to analyze the effect of education, income, policies, and unemployment on crime rates. The findings show that the increase in income and education leads to a decrease in crime. Also, higher education has a lower impact compared to lower education which has a higher impact. Unemployment has a marginal effect on crime. It is also shown that Income, education, and security policies reduce crimes but do not give any evidence about what kind of policies.

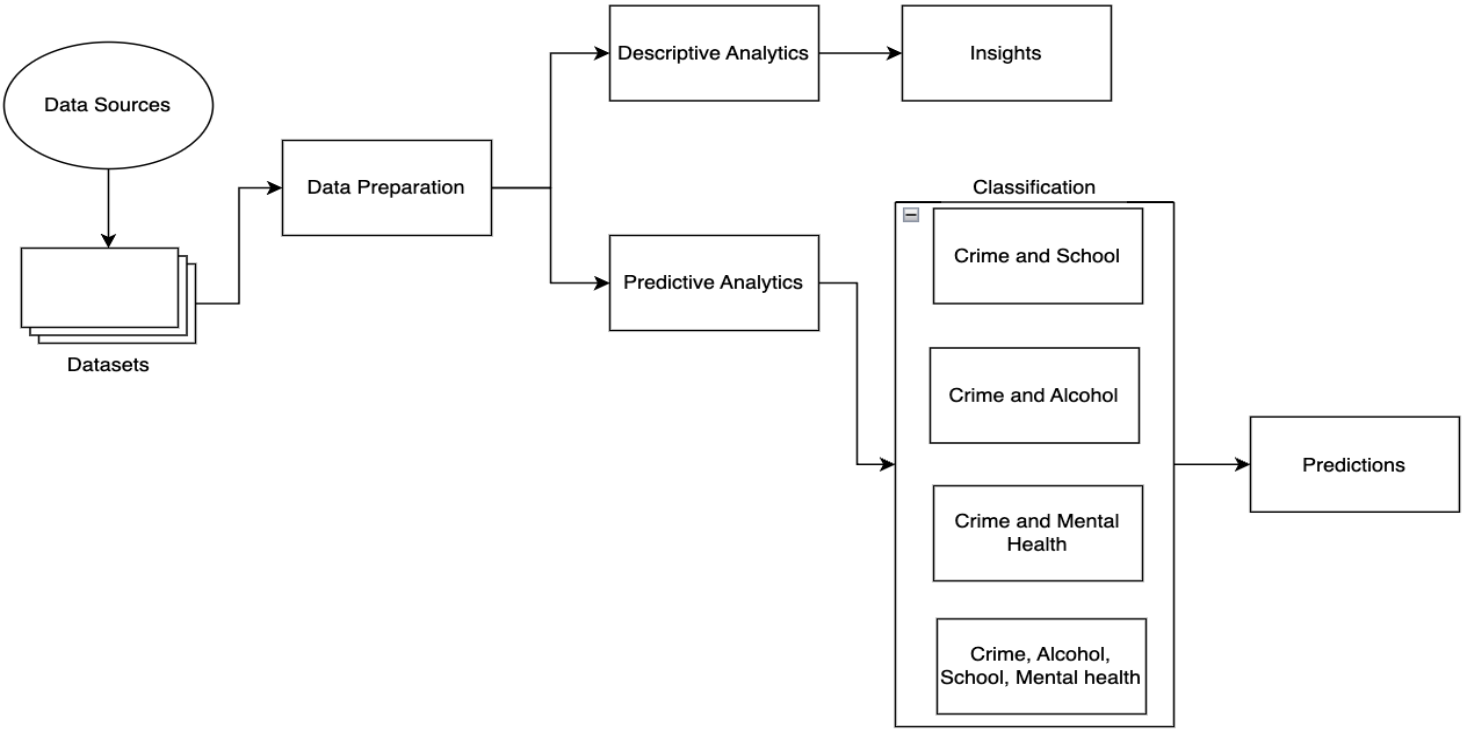

Fig. 1: Design Diagram

Another paper that we studied was about Education and Crime across America: Inequity's Cost. This paper combines the FBI's uniform crime reporting database and district finance data and studies the longitudinal relationship between crime in every town/city (whose police department has reported crime statistics) and its school district spending in the years 2003 to 2018. They find patterns in how the school district spending has changed over the years affecting the rate and type of crimes. The paper takes numbers and pieces of evidence from various past research to confirm or extrapolate their findings. The aim of this paper is to demonstrate a relationship between education funding and reduced crime across America with regard to the amount spent per student as well as equity in spending. With respect to the dataset, they combined over 213 datasets to control for population, density, wealth, education, employment, cost-of-living, race, law enforcement, voting history, teacher salary, teacher engagement, and student chronic absenteeism. The paper [3] focuses on the factors that cause crime in various states in India and uses regression to predict which of them have a higher probability of crime occurrences. The dataset used contains information about all states in India; the number of males, and females divided upon age groups; and the different kinds of crimes in that state.

Concerns have grown that expenditure on education and welfare is diminishing as federal, state, and local governments continue to devote a large portion of their budgets to law enforcement and corrections. The effectiveness of public spending in preventing crime must be evaluated given the rising fiscal pressure in the United States. In paper [4] the effectiveness of government spending on welfare and education versus that on criminal justice and corrections is compared. Results of linear regression with panel-corrected standard errors and GMM estimation using panel data from 50 U.S. states from 1994 to 2014 show that public welfare and education spending may be able to reduce rates of violent and property crime, but law enforcement spending can only deter property crime.

Crime has played a significant role in the health of each and every individual. This paper [5] studied the relationship between crime and health in the state of turkey. For empirical analysis, this study integrates data at the individual, household, and regional levels from a nationwide household survey. The connections between a perceived neighborhood crime indicator and an individual-level health status index are quantified using a multilevel estimating methodology.

In this research paper [6] studied the relationship between school attendance and crime rates. There are two major conclusions. First, tighter rules requiring compulsory education have a noticeable and long-lasting negative impact on crime. Second, there is a tenuous or nonexistent connection between these regulations and educational achievement. Because of this, it is difficult to identify meaningful causal estimates of the education-crime association for the more recent time, however, it is possible for some populations with lower educational

levels (in particular, for blacks).

III. Datasets

A. NYPD Arrests Data (Historic)

A breakdown of each arrest made in NYC by the NYPD from 2006 through the end of the previous calendar year is provided below. The type of crime, the location, and the time of enforcement are all listed in each record, which indicates an arrest the NYPD made in NYC. Additionally, details about suspect demographics are also presented. Every three months, the Office of Management Analysis and Planning manually extracts this data, reviews it, and then posts it on the NYPD website. The dataset has 5.3M rows and 19 columns [7]

B. NYPD Shooting Incident Data (Historic)

The shooting incidents in NYC from 2006 through the end of the previous calendar year are broken down in this table. Before being released on the NYPD website, this information is personally gathered each quarter and examined by the Office of Management Analysis and Planning. Each row denotes a shooting that happened in New York City and contains information about the occurrence, including the time and date of the shooting. Additionally, information on the demographics of the suspect and victim is given. The dataset includes 19 columns and 25.6K rows [8]

C. NYPD Hate Crimes

Hate crimes are offenses committed against a person primarily because of that person's characteristics or religious convictions. All confirmed hate crime incidences in New York City are included in the NYPD hate crimes dataset. There are 1.8K rows and 14 columns in the dataset. The columns include details on the borough, a description of the prejudice motive, the PD code, the offense category, and others. The Law code category provides information on the severity of the offense, such as a Misdemeanor and a Felony [9]

D. Demographic Snapshot

This dataset contains all the registered schools till 2021-22 and their enrollment count. The enrollment count is further sub-categorized into Grades K-12. Other columns include information regarding ethnic diversity, poverty percentage and economic need index. The October 31 Audited Register is the basis for the enrollment figures for the 2017–18 to 2019–20 academic years. Enrollment statistics are based on the November 13 Audited Register for 2020–21 and the November 12 Audited Register for 2021–22 to take into account the delayed start of the school year. The Economic Need Values of the school's students are averaged to create the Economic Need Index. The percentage of students that are struggling financially is estimated by the Economic Need Index (ENI). [10]

E. Hospital Inpatient Discharges

Mental health awareness and reluctance to report mental health problems lead to limited availability of data in this domain. The authors faced numerous challenges to identify and gather significant data at the borough level. Discharge-level information on patient characteristics, diagnoses, treatments, services, and charges can be found in the Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File. Basic record-level information for the discharge is contained in this data file. There is no PHI (protected health information) in the de-identified data file as defined by HIPAA. The health information cannot be used to identify any specific person because all data elements deemed identifiable have been redacted. For instance, the day and month of a date are not part of the direct identifiers for that date. The authors found the data to be divided by year having about 2.3M rows and 30+ columns. The columns have information regarding ethnicity, race, Diagnosis description, and code along with the severity of illness and others. [11] [12] [13] [14] [15]

F. Liquor Authority Current List of Active Licenses

The dataset contains the Liquor Authority's current list of all active licensees in New York City. It has about 50K rows and 21 columns. The columns have information regarding the license type and code along with the mode of operation. The mode of operation gives insights into the availability and affordability of alcohol in the neighborhood. [16]

IV. Methodology

A. Data Understanding

To analyse how the shifting patterns of crime in New York City boroughs get affected by factors like the states of education, mental health and alcohol consumption in each of the boroughs, we studied and explored around 7 datasets related to the same.

1) Crime datasets: As part of analyzing crime in New York City, it was important to understand the trends in crime the city has shown over the years. For this we took statistics about a few specific crimes like hate crimes and shooting crimes, as well as data that could depict patterns in major felony offenses, non-major felony offenses, misdemeanor offenses and violation offenses. Most of the crime datasets that we have used for this project are from NYC Open Data and the NYPD website. [17] [18]

The NYPD releases overall citywide statistics weekly. It also releases incident-level data related to criminal activity. It records reported crime and offense data in accordance with the New York State Penal Law and other New York State laws. For statistical presentation purposes, the numerous law categories and subsections are summarized by law class: felony, misdemeanor, and violation. These legal categories are then subdivided into broad crime and offense categories, e.g., Felonious Assault, Grand Larceny, Misdemeanour Criminal Mischief, etc. The NYPD also provides a historical view of crime data starting from 2000 to 2021. [19]

The Arrests dataset contains labeled crime data for a borough. Since the objective of the project is to predict the level of offense (the type of crime) occurring in the 5 boroughs, the information in this dataset, combined with the NYPD crime statistics, was ideal for running classification models. This dataset is a breakdown of every arrest made by the NYPD from 2006 to 2021. Before being released on the NYPD website, this information is manually gathered each quarter and examined by the Office of Management Analysis and Planning. Each record corresponds to each arrest that the NYPD made in NYC and contains details not only on the crime, the scene, and the time of enforcement, but also the suspect and victim demographics. This dataset had more than 5 million rows. [20]

Datasets related to shooting and hate crimes were also analyzed for gathering statistics. The NYPD shooting incidents dataset has shooting records from 2006 to 2021 and has around 25,000 rows. The NYPD hate crime dataset is relatively smaller since it records the hate crimes only from 2019-2021 and has about 1500 rows. Both datasets contain information about the crime, as well as suspect and victim demographics. [21] [22]

2) Education datasets: With education, we aimed to gain some insight into the schooling and dropout statistics of each borough. The NYC Department of Education dataset was used. It had all demographic data about schools in each borough, as well as data on a granular level about enrolments in each class of the school. The dataset also has a column featuring the school's Economic Need Index or the ENI. ENI estimates the percentage of students facing economic hardship. The 2014-15 school year is the first year these estimates were provided, hence education data that we have covered only the last 5 years. The ENI metric is calculated as follows:

- The student's Economic Need Value is 1.0 if either of these is true for the student:
  - The student is eligible for public assistance from the NYC Human Resources Administration
  - The student lived in temporary housing for the past four years
  - The student is in high school has a home language other than English and entered the NYC DOE (NYC Department of Education) for the first time within the last four years.
- Otherwise, the student's Economic Need Value is based on the percentage of families (with school-age children in the student's census tract whose income is below the poverty level, as estimated by the American Community Survey 5-Year estimate (2020 ACS estimates were used in calculations for 2021-22 ENI).

The student's Economic Need Value equals this percentage divided by 100.
This dataset had a total of 9251 rows and 44 columns, with an almost equal number of rows of enrollment for each year. Every row in the dataset represents a school in NYC for a particular enrollment year. [23]

3) Mental Health datasets: Combining mental health with crime was one of the most crucial and challenging parts of this problem statement. That was majorly because mental health data is not very commonly available unless it is a survey, which doesn't often cover a large demographic region, or aggregated data which is used to spread awareness and create programs for people. Interestingly the data we used is not only comprehensive but also covers all boroughs. We have used data from Hospital In-patient Discharges from hospitals in all boroughs, across 5 years from 2016 to 2020. Each row of this dataset describes the patient and a description of their diagnosis. Due to the large number of health concerns for which one may visit a hospital, this dataset had an average of 2.5 million records for every year. [24]

4) Alcohol Consumption datasets: We first took a look at an aggregate dataset about Drug and Alcohol Consumption from the NYC Health website, which covers the consumption of several drugs and the percentages of its consumers over a 4-year window from 2017 to 2020. To further analyze the alcohol provider and user patterns, we used a dataset called the Alcohol Licence Statistics from the New York State government website. It contains information on all current and past Alcohol licenses held by restaurants, bars, grocery stores, etc in all boroughs of New York City. This dataset consisted of approximately 50,000 rows in total, where each row represented an alcohol license that was issued, and columns describe the license and the entity holding it. [25] [26]

B. Data Preparation

1) Crime datasets:

- Arrests Data - In the arrests data the borough information was available as initials. We did the mapping of each initial to a borough. The one-letter abbreviation was - M for Manhattan, Q for Queens, K for Brooklyn, B for Bronx, and S for Staten Island. After plotting the graphs for visual inspection the authors dropped all the columns related to the coordinates of the arrests.
- Hate Crimes - Some of the columns like Month Number, Record create date, Arrest date, and id were dropped as they don't serve any valuable purpose.
- Shooting Crimes - Columns related to location coordinates, occurrence time, and jurisdiction code were dropped.

2) Education datasets: The NYC Department of Education dataset is quite comprehensive. Information about which borough the school lies in was missing. So, after the dataset was pre-processed and normalized, we utilized the DBN column to create a new column that would represent the Borough. DBN stands for District Borough Number, and it is a combination of the district number, the letter code for the borough and the number of the school. Every school in New York City has a District Borough Number. With the borough in place now, the dataset was first segregated into bins according to the

enrolment numbers of each school in NYC to analyse schools at a larger level.

The dataset at a granular level depicts the number of students in each grade for every school. To make the dataset more manageable, we grouped data for several grades together into elementary school, middle school and high school. This gave us a good high-level view of elementary, middle and high school distribution across all boroughs and also helped us find relevant correlations between all features.

3) Mental Health datasets: The Hospital In-patient Discharges datasets from 2016 to 2020 are huge with 2.5 million records for each year. This is a state-wide dataset, but since we were only interested in statistics about New York City, we reduced the dataset based on the hospital's service area. The large volume of data was further reduced which we reduced by considering only those rows which had any association to mental health. Across all years from 2016 to 2020, Mental Diseases and Disorders were one of the top 8 diagnoses amongst all.

4) Alcohol Consumtion datasets: The Alcohol Licence Statistics dataset required several data wrangling steps. After dropping columns that weren't relevant to this problem statement, we processed the dataset to remove columns that had missing values above a certain threshold of 50 percent. Most features of this dataset that were relevant for this problem were nominal in nature, therefore we performed One hot encoding on those columns, followed by correlation analysis among them, and the observations are in line with the intuition. This dataset had several date columns relevant to alcohol licenses (License Issued Date, License Expiration Date, Effective Date and Original Date). Since the purpose of using this dataset for this problem is with the assumption that the alcohol licenses of a borough speak about the kind of consumption, Original Date which corresponds to the first date the license was issued, and Expiration date, which is the date current license expires, were the only 2 columns used. Using these 2 columns helped create an interval about alcohol consumption. These intervals were then split to only extract the year, since we are only dealing with years in all other datasets of this problem as well. With the year, we now exploded the dataset to create new entries of alcohol license, one for each year it was held. With this, our drugs and alcohol dataset is ready for modelling.

C. Data Analysis

1) Crime datasets:

- *Sum of law classes of crime with time for NYC*
  To understand crime in New York City better, we analysed each of major felony offenses, non-major felony offenses, misdemeanour offenses and violation offenses separately. The trend in the occurrence of these crimes in a 22 year range from 2000 to 2021 is as observed in their corresponding graphs as elaborated below.
  The trend of major felony offenses like murder, rape, robbery, assault, burglary, and larceny show that their numbers have been more or less stable across all years from 2000 to 2021 as seen in Fig 2, with a slight dip observed in a few types from 2000 to 2003. The 7 types of felony offenses dipped in total count significantly from 2000 to 2009, and then stayed constant with a few bumps till 2021.

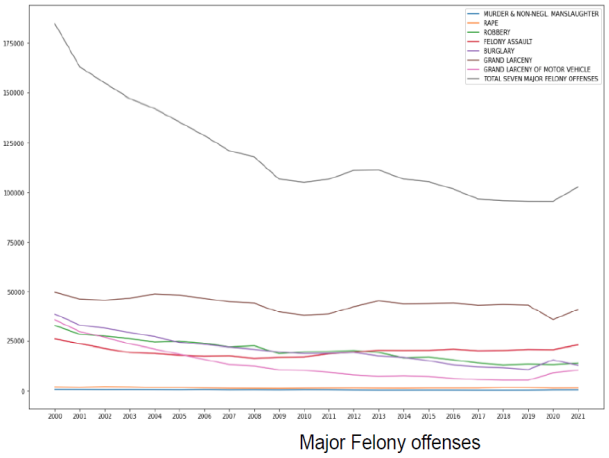

Fig. 2: Major Felony Offenses reported in NYC from 2000 - 2021

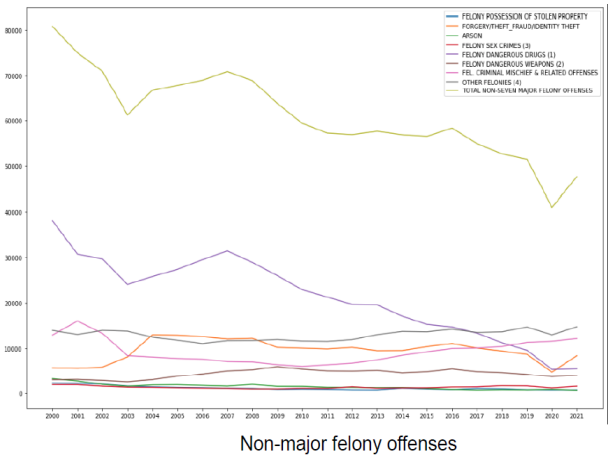

Fig. 3: Major non-Felony Offenses reported in NYC from 2000 - 2021

Major non felony offenses as seen in Fig 3 like possession of stolen property, forgery, identity theft, arson, sex crimes, drugs, and criminal mischief show variations in patterns across all years from 2000 to 2021. Crimes related to drugs were the highest among non-felony offenses from 2000 to 2016. Criminal mischief showed sudden and steep drop in 2002. Sex crimes and arson incidents have stayed constant throughout the years. The total non-felony related crime incidents have shown a drop across the years, with a slight increase in 2007.

Misdemeanour drug related crime incidents have dropped significantly from 2000 to 2021 as seen in 4. Misdemeanour fraud incidents were almost constantly the most frequent among this set. The total incidents related to misdemeanour offenses have shown an overall decrease

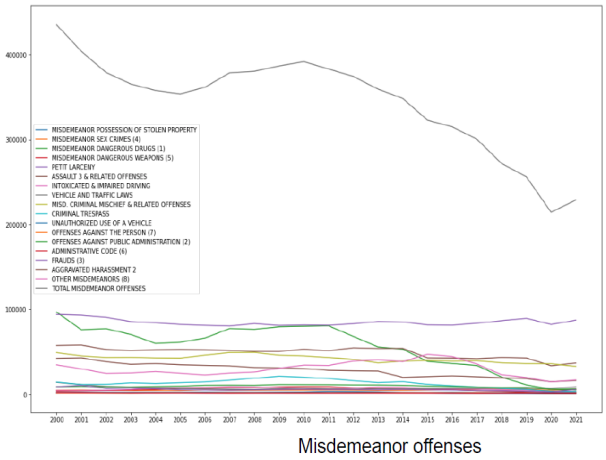

Fig. 4: Misdemeanor Offenses reported in NYC from 2000 - 2021

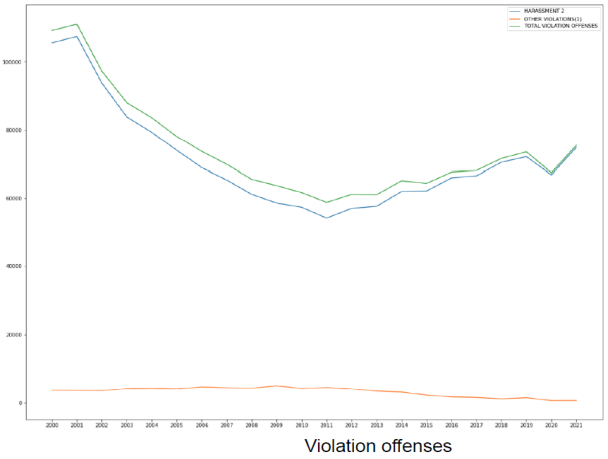

Fig. 5: Violation Offenses reported in NYC from 2000 - 2021

across all years from 2000 to 2021.

Strangest trend of all, 5 shows harassment related incidents show steep decrease from 2000 to 2012, but start increasing gradually after that till 2019, and took a sharp downturn in 2020 (probably due to Covid). The trend of the total violation incidents also show a decrease from 2000 to 2021.

As is clear from the maps of the 5 boroughs in fig 10 in plotted with offense distribution, dangerous drug related crimes are most widespread in all the boroughs. Manhattan has significant cases of assault and traffic violations. Brooklyn reports high numbers on incidents related to assault and dangerous drugs. Assault and traffic violation cases are most frequent in Queens. All the 5 boroughs show similar trends in terms of the most common reported incidents of crime in the boroughs. Crimes related to dangerous drugs, larceny, assault and traffic violations are common across all 5 boroughs. Additionally, crimes related to miscellaneous penal laws are very high in Brooklyn, Queens and Brooklyn.

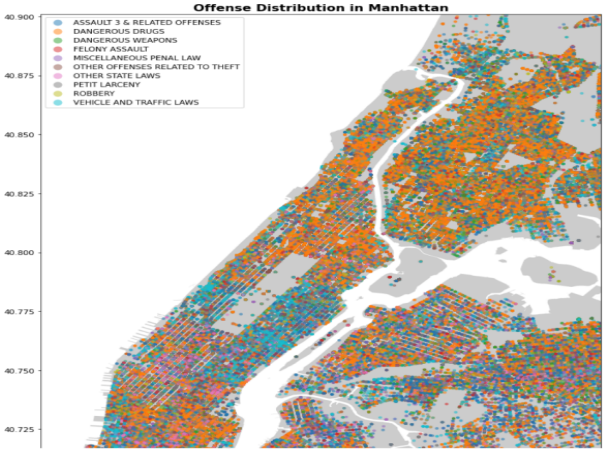

Fig. 6: Offense distribution in Manhattan

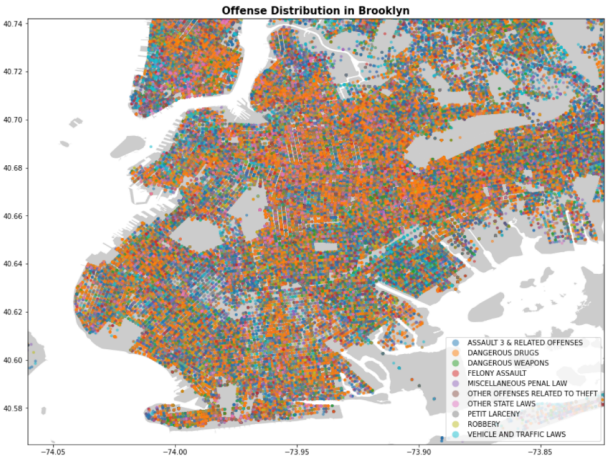

Fig. 7: Offense distribution in Brooklyn

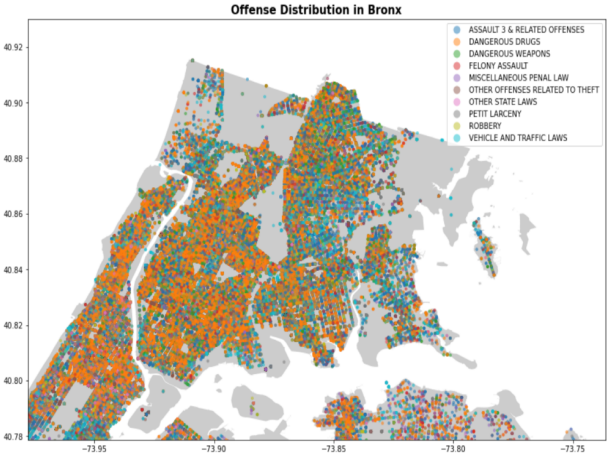

Fig. 8: Offense distribution in Bronx

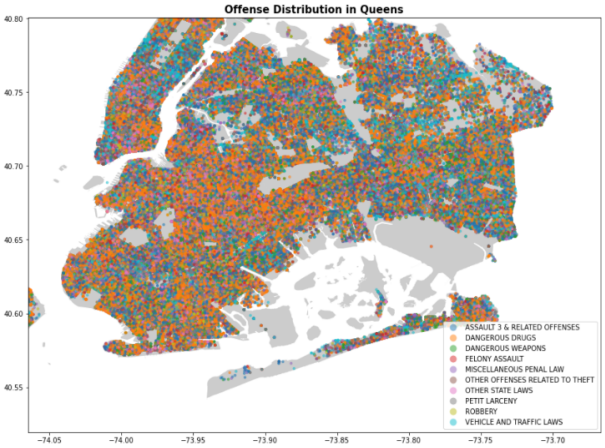

Fig. 9: Offense distribution in Queens

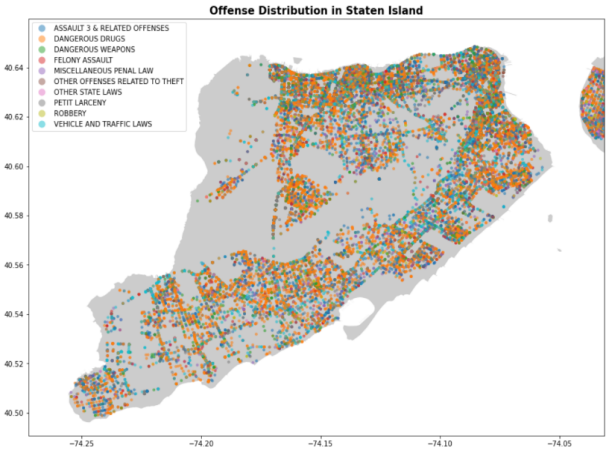

Fig. 10: Offense distribution in Staten Island

In terms of broad categories, misdemeanour offenses are clearly more widespread than violation and felony offenses in all the 5 boroughs.

- *Hate crime trends for NYC*
  NYPD Hate Crime dataset was used to plot number of hate crimes reported against the religion and sexual orientation of the victim from the years 2019 to 2021. The results are as shown in fig 11. The most important analysis from this graph is the bias motive description, where most of the hate crimes clearly show a bias motive towards anti-Jewish sentiments. Hate crimes against Asians, homosexuals and African Americans follow next.

- *Shooting crime trends for NYC*
  NYPD shooting statistics dataset was used to plot the total number of shooting incidents that took place in each of the 5 NYC boroughs from 2006 to 2021. Brooklyn reports the highest number of shooting crimes among the 5 boroughs as seen in 12.
  Fig 13 shows the total number of shooting crimes that

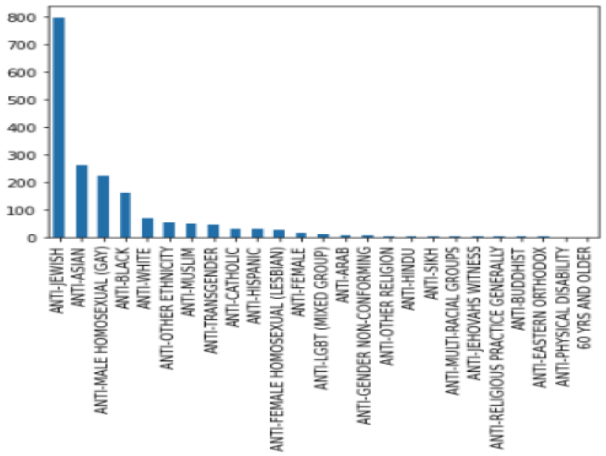

Fig. 11: Number of shooting incidents in the 5 NYC boroughs from 2006 to 2021..

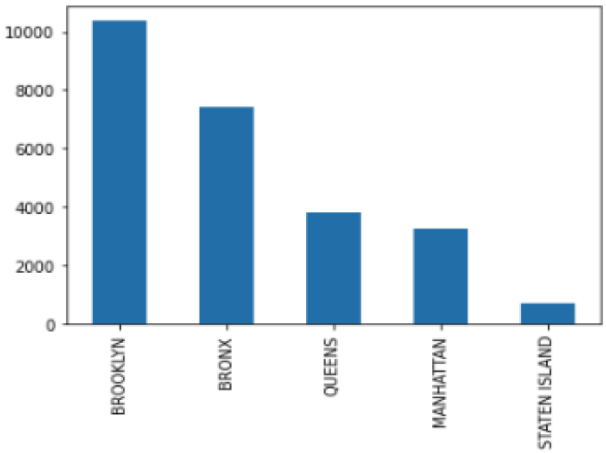

Fig. 12: Number of hate crimes reported based on their bias from 2019- 2021.

were reported in all the 5 boroughs of NYC from 2006 to 2021. It shows a gradual decrease in shooting incidents till 2017, and remains constant after that till 2020. Shooting crimes report high numbers for 2020 and 2021. Fig 14 plots the locations of all the reported shooting incidents and distinguishes them by colour based on the ethnicity of the victim. We see that majority of the shooting crimes have been done against African Americans, followed next by American Hispanics. Shooting incidents against Asians and Black Hispanics are lesser compared to the first 2, and the number against White are low.

2) Education datasets:

- *Enrollment pattern and Economic Need Index Trends in NYC*
  With the borough in place now after pre-processing, the NYC Education dataset was segregated into bins according to the enrolment numbers of each school in

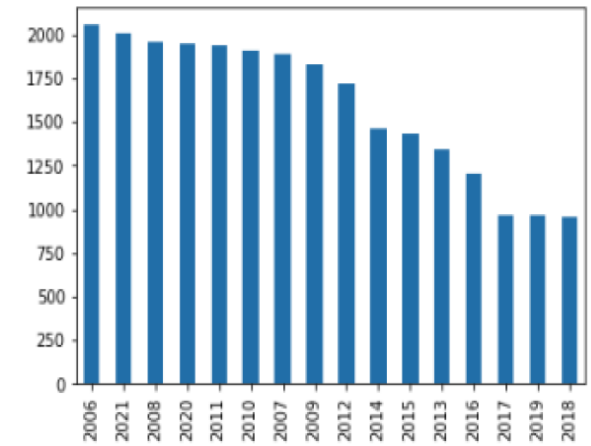

Fig. 13: Reported number of shooting crime incidents in NYC from 2006 – 2021.

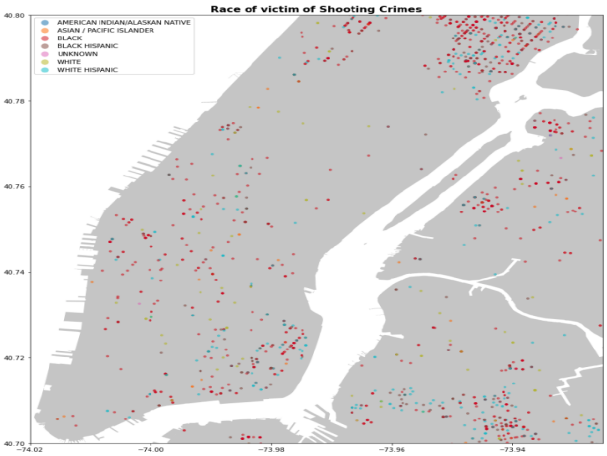

Fig. 14: Race of shooting crime victims in NYC.

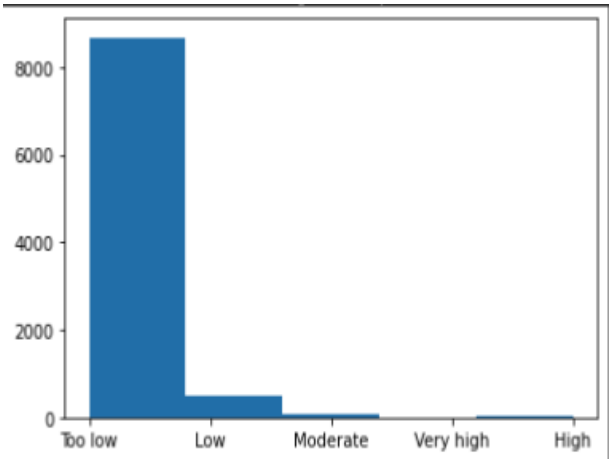

Fig. 15: Enrolment pattern of students in NYC schools from 2017-2021.

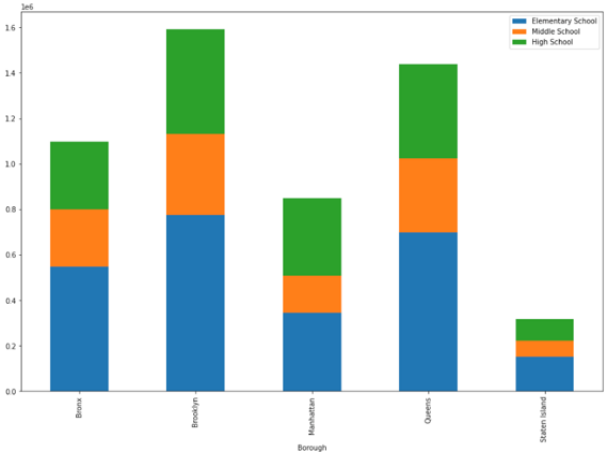

Fig. 16: Enrolment pattern of students by their grades from 2017-2021.

NYC. The following class labels were assigned to analyse the enrolment pattern of students in schools:

- Too low: 0-1200
- Low: 1200-2400
- Moderate: 2400-3600
- High: 3600-4800
- Very high: 4800-6000

As seen in fig 15, there are very few schools with Moderate to High enrolment numbers, and a majority of them lie in the Too Low region, which means most schools have their enrollment numbers less than 1200 in every academic year.

Figure 16 depicts the number of students enrolled in the different sections of schooling. To plot this graph, the data were grouped into various buckets based on the grades. The elementary school enrolment number is higher than the other two, this is because there are more grades included in this bucket compared to the other two. Brooklyn shows the highest enrolment of students in schools, followed by Queens.

Figure 17 shows the total enrollment across all boroughs and the aggregate economic need index of all schools of each borough. There is a slight dip observed in the total enrolment over the years. Brooklyn shows the highest enrollment number. While the Bronx ranks third in total enrolment, it has the highest economic need index across all boroughs.

Figure 18 shows the distribution of race across the enrolment patterns of students in the schools of all 5 boroughs. The highest enrolment shown is by Hispanics, and African Americans and Native Americans compete for a second.

- *Correlation Matrix*

The correlation matrix between all features of the education dataset was plotted. It gave some really good insights into the schooling system of the 5 boroughs:

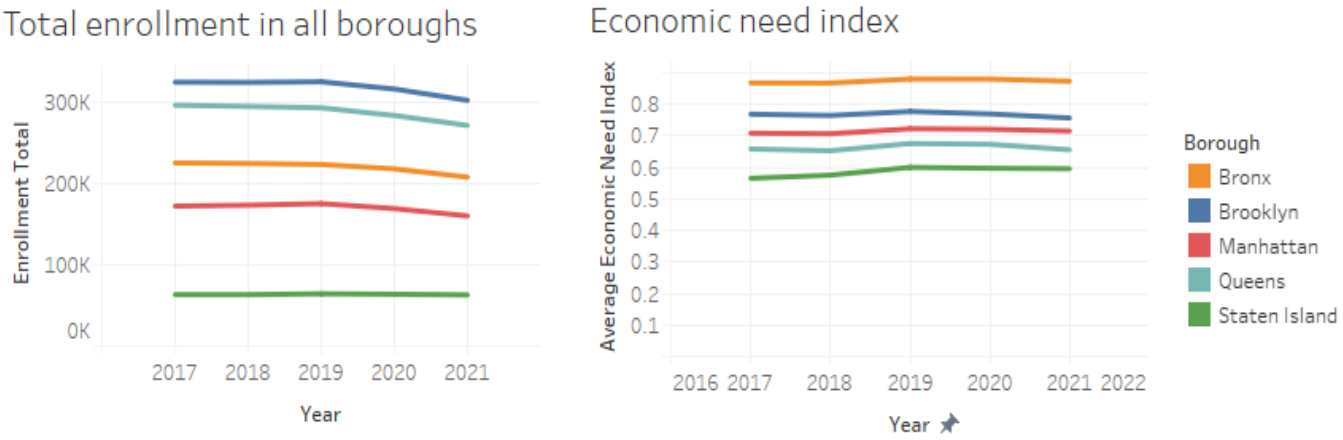

Fig. 17: Total enrollment and Economic Need Index across NYC boroughs.

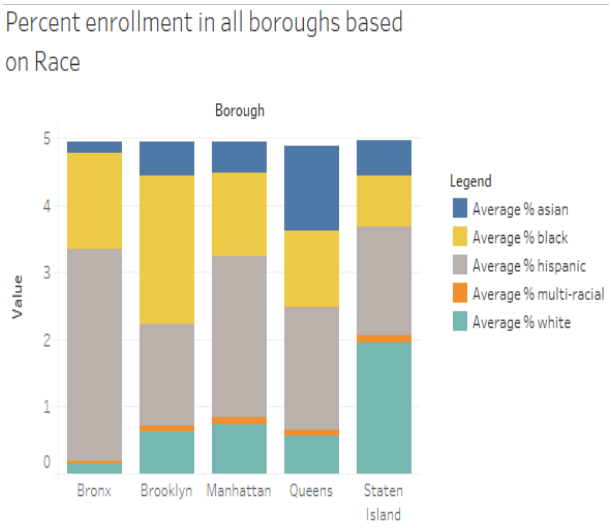

Fig. 18: Distribution of race among enrolled students in NYC boroughs.

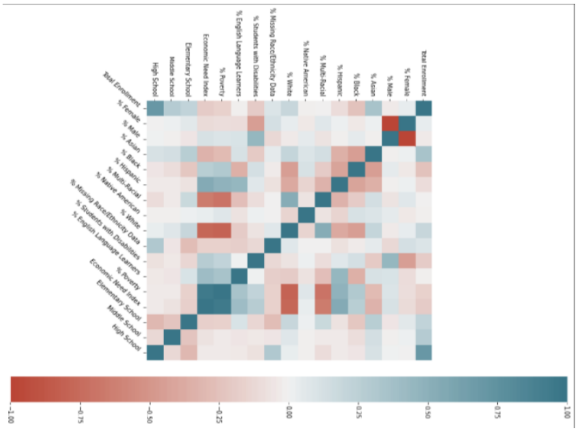

Fig. 19: Correlation Matrix of all features of education dataset.

- Percent Poverty is very positively correlated to Economic Need Index
- Percentage of Multi-racial and white students have a strong negative correlation to Poverty and Economic Need Index
- Students with disabilities are more positively correlated to the percent of males, and negatively correlated to the percentage of female

3) Mental Health datasets: We combined the datasets for Hospital In-patient Discharges from 2016 to 2020 into a single dataset to understand the mental health trend across all the years in the 5 boroughs. We made a plot of the different mental health disorders with their total count in terms of cases reported. Through this, we got a list of the most common mental health-related disorders seen in the 5 boroughs. The aggregate of the top 10 diagnoses for mental health amongst the in-patient data is shown in figure 20.

Alcohol-related disorders, although not the highest themselves, are still significantly higher than other disorders in all 5 boroughs. This indicates a strong correlation between mental health and alcohol consumption, which gives us confidence in analyzing the crime in NYC with mental health and alcohol consumption. This is a very interesting trend observed. After alcohol, schizophrenia and related psychotic disorders rank second in all boroughs except for Staten Island. Mood disorders and substance-related disorders also report high cases in all 5 boroughs. Overall, the mental health for the entire NYC shows widespread disorders related to alcohol consumption, schizophrenia, mood and substance-related disorders.

4) Alcohol Consumption datasets:

- Usage of Drugs across NYC
  The drug and alcohol consumption dataset indicate the total types of drugs used across the years. We plotted the usage of various drugs stored in the dataset from 2017 to 2020 as shown in fig 21. The graph above throws insight on how the consumption of *Fentanyl* has been growing for the past three years, whereas the consumption of Benzodiazepines has declined. Other drugs are

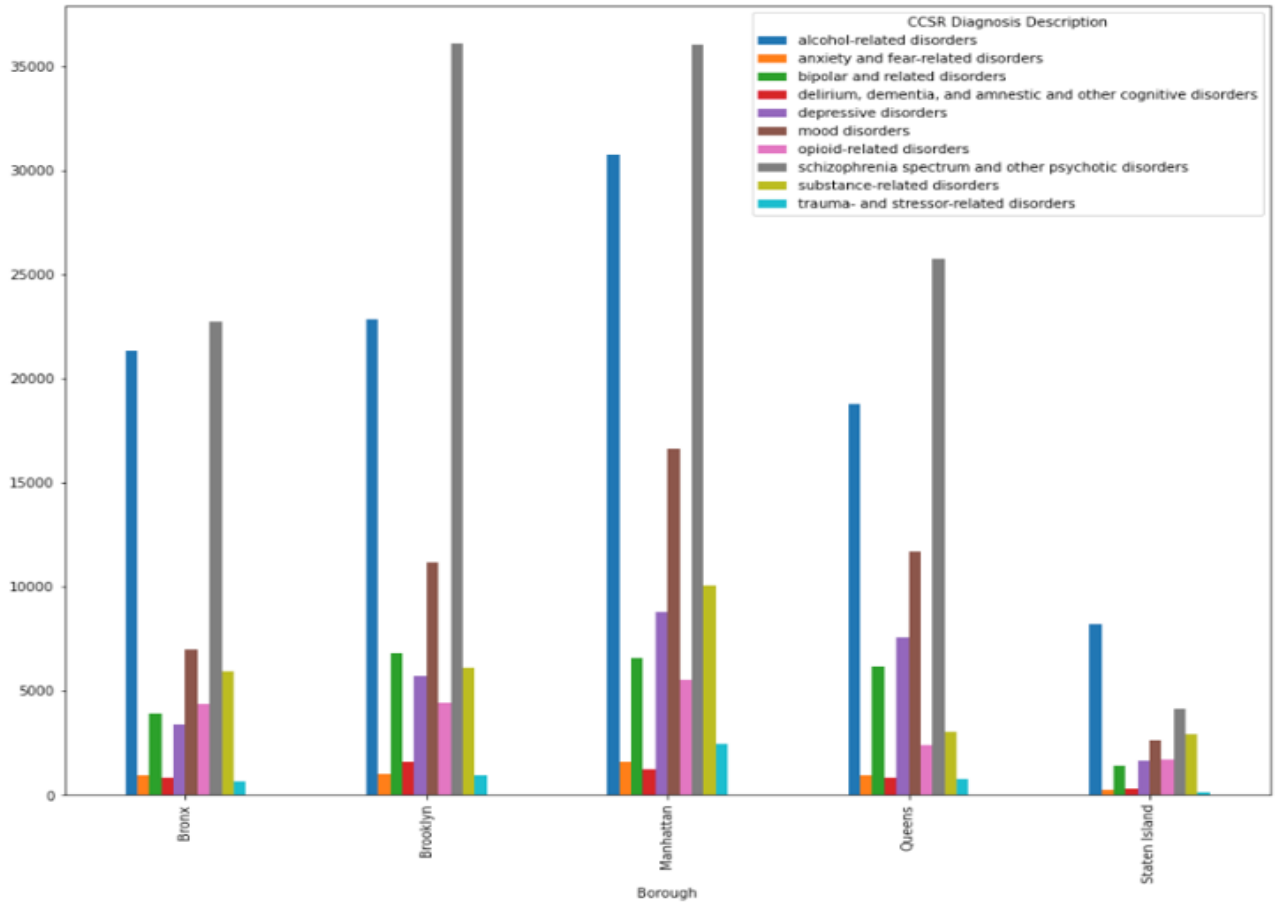

Fig. 20: Dignosis Description Mental Health

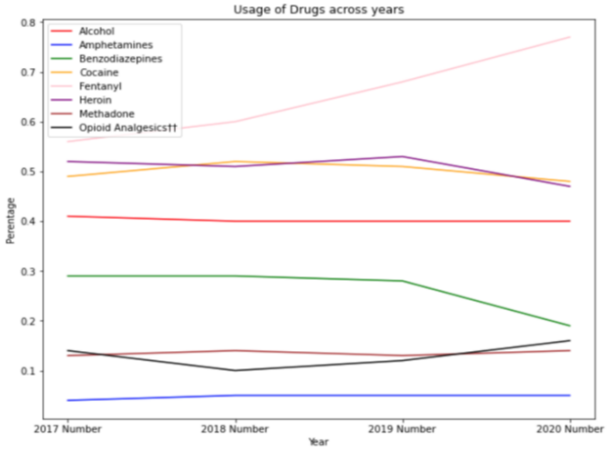

Fig. 21: Drug usage

showing a neutral or minimal shift over the years, but interestingly the consumption of alcohol has more or less been constant.

- Understanding alcohol consumption from Alcohol sources in NYC boroughs

The alcohol license dataset had information about the major alcohol distributors in all the boroughs. We have made an assumption here that the alcohol licenses of a borough speak about the kind of consumption patterns in it. Using this data, we plotted the major sources of getting alcohol in the 5 boroughs as seen in fig 22.

The graph shows highest alcohol consumption in Manhattan. It is to be noted however that the highest number reported is of mainly type on premise liquor. This implies sources such as clubs, pubs and restaurants. This could be due to the huge popularity of Manhattan even among tourists. This is also expected as Manhattan is a richer neighbourhood than the rest. Hotel and restaurant liquor is significantly high as a source in Manhattan, which further corroborates this statement. Cheaper sources such as grocery chains and grocery stores are more readily available in other boroughs. Consumption from sources such as drug chains, serving only beer and wine is relatively low across all the 5 boroughs.

D. Modelling

In this section, our efforts were to be able to predict what type of crime could occur in each borough, based on several

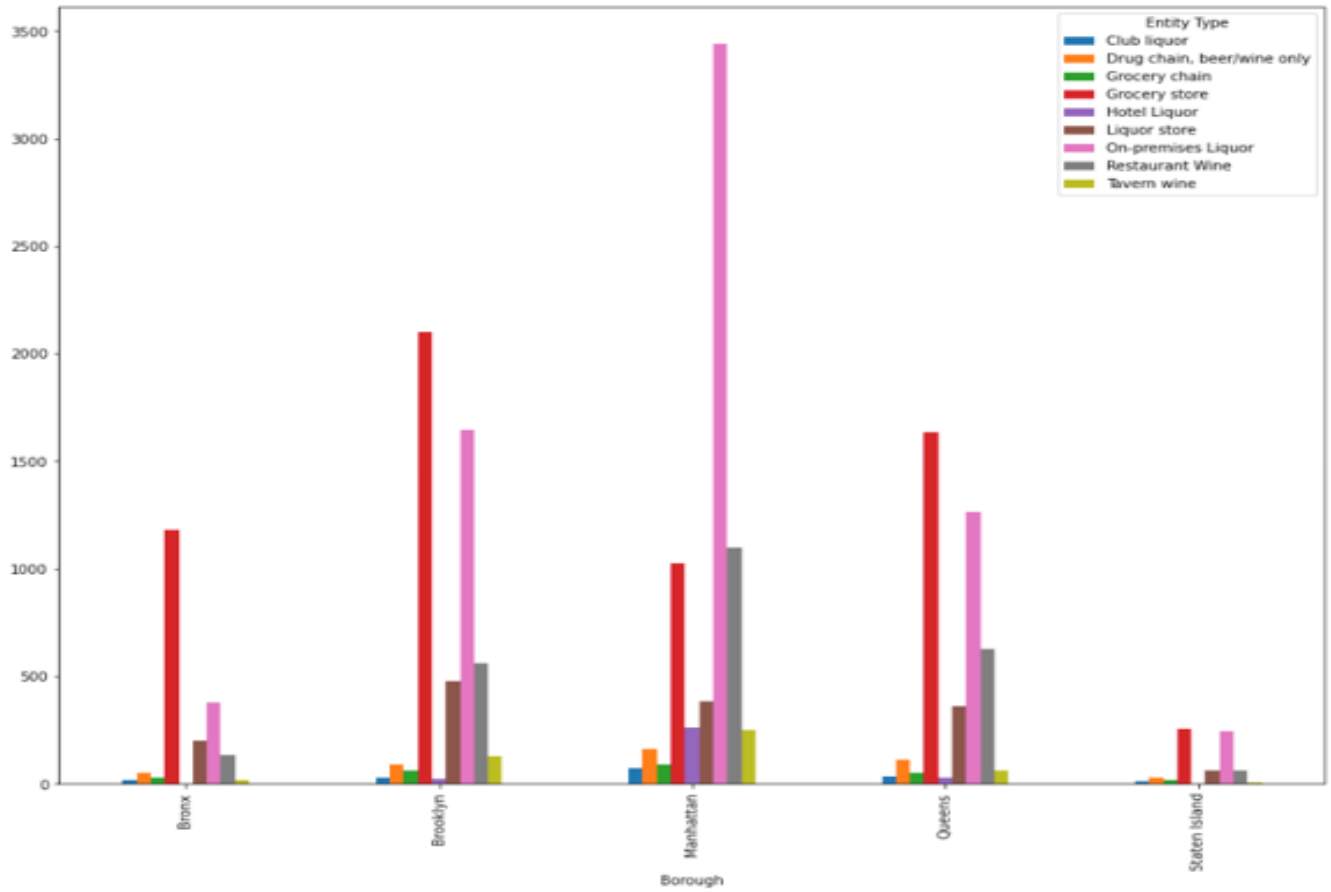

Fig. 22: Alcohol Entity type

education, mental health, alcohol distribution factors. For this, of all the crime types, we picked the 10 most frequent offenses to be predicted by the models.

1) Education and Crime: The education dataset was compressed to have each row represent an entry corresponding to a borough and year. Since all columns in education represented number of students or number of enrollments in the school, and average was taken over the years to have a single value for a feature corresponding to the borough and year.

The crime dataset consisting of about 800,000 rows was joined with this compressed education dataset on 'Year' and 'Borough', in order to be able to make predictions about crime using statistics from education.

With this merged dataset, first step was to encode all categorical variables. Models only work with numerical values. For this reason, it was necessary to convert the categorical values of the features into numerical ones, so the algorithms can learn from those data and gives the right model. First, we used One-hot encoding for all categorical variables, since One hot encoding makes data more useful and expressive, and can be re-scaled easily. This type of encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category. A great advantage of this type of encoding is it does not create any relationship between categorical variables where none previously existed.

Following encoding, we applied PCA on the dataset for dimensionality reduction. Since the strength of principal components decreased largely after the fourth principal component, we only picked the top 4 principal components to fit our model. We then applied 4 classification models on this dataset: Decision-tree classifier, XGBoost classifier, Multi-class Logistic Classifier and Naive Bayes classifier. We enhanced the performance of these classification models using k-fold cross validation with k ranging from 5 to 10, and found k=10, to give us the best performance with a Decision Tree Classifier.

2) Mental Health and Crime: The mental health dataset was compressed to have each row represent the most prevalent mental health diagnosis in a particular borough, for a particular year. Since mental health data consisted of features that were mainly categorical, during compression the categorical variable with the highest frequency for each feature was picked.

The crime dataset consisting of about 800,000 rows was joined with this compressed mental health dataset on 'Year' and 'Borough', which now in the merged form represented different crimes and each borough along with the most prevalent mental health disorder of the region.

Since this dataset consisted of a lot of categorical variables,

we tried 2 encoding techniques followed by different classification models to make predictions. The first one was one-hot encoding, and the next was leave-one-out-encoding where the encoded column is not a conventional dummy variable, but instead is the mean response over all rows for this categorical level, excluding the row itself. This has an advantage of having a one-column representation of the categorical while avoiding direct response leakage.

With each encoding, we applied PCA on the dataset for dimensionality reduction. Since the strength of principal components decreased largely after the second principal component, we only picked the top 2 principal components to fit our model. We then applied 4 classification models on this dataset: Decision-tree classifier, XGBoost classifier, Multi-class Logistic Classifier and Naive Bayes classifier. We enhanced the performance of these classification models using k-fold cross validation with k ranging from 5 to 10, and found k=10, to give us the best performance with a Decision Tree Classifier, after categorical variables were One-hot encoded.

While the training accuracy with Leave One Out Encoding was incredibly higher than the accuracy with One-hot encoding, classifiers models seemed to overfit with this, giving us an accuracy of over 90

3) Alcohol and Crime: The alcohol dataset was compressed to have each row represent the most frequent modes of selling alcohol in each borough for the entire year range. This dataset conveyed information about the top 5 modes of alcohol distribution.

The crime dataset was joined with this compressed alcohol dataset on 'Year' and 'Borough', which now in the merged form represented different crimes and each borough along with the top 5 most frequenct alcohol distribution channels of the region.

Since this dataset did not contain a lot of categorical variables, apart from the ones in the crime dataset, hence we only used One-hot encoding here.

For this dataset, since all features that were added as part of the alcohol dataset conveyed information about the distribution of alcohol, we did not need to apply PCA here. We under-sampled the dataset to have equal number of rows for all crimes and ran 4 classification models on this dataset: Decision-tree classifier, XGBoost classifier, Multi-class Logistic Classifier and Naive Bayes classifier. We enhanced the performance of these classification models using k-fold cross validation with k ranging from 5 to 10, and found k=10, to give us the best performance with a Decision Tree Classifier, after categorical variables were One-hot encoded. .

4) Education, Mental Health, Alcohol and Crime: In this part of modelling, each of the education, mental health and alcohol datasets were merged together with the crime dataset on 'Year' and 'Borough'. Here each row represented all the information , which now in the merged form represented different crimes and each borough along with the top 5 most frequent alcohol distribution channels of the region.

Since this dataset did not contain a lot of categorical variables, apart from the ones in the crime dataset, hence we only used One-hot encoding here.

Following this, we applied PCA on the dataset for dimensionality reduction. Since the strength of principal components decreased largely after the fourth principal component, we only picked the top 4 principal components to fit our model. We then applied 4 classification models on this dataset: Voters Classification, Decision Tree Classification, Multi-class Logistic Classification and Naive Bayes classifier. We enhanced the performance of these classification models using k-fold cross validation with k ranging from 5 to 10, and found k=10, to give us the best performance with a Decision Tree Classifier, after categorical variables were One-hot encoded.

V. Result

A. Descriptive Analytics

1) Crime:

- The overall crime in NYC has reduced over the years, keeping felony, misdemeanour and violation cases in mind. As observed in each of these graphs, the total number of criminal incidents reported reduced over the years.
- Misdemeanour incidents are the highest reported among these classes of crimes, with larceny being the most common offense. Drug related felony and misdemeanour incidents were pretty high in 2000, but have seen a significant decrease over the years. Harassment incidents decreased till 2012, but started increasing after that.
- All boroughs show similar trend in terms of the most widespread type of crime. Crimes related to dangerous drugs, assault, larceny and penal law violations are highest across all the boroughs.
- Shooting incidents were reducing in NYC gradually from 2006 to 2018, but this stopped 2019 onwards. 2020 saw a sharp rise in the number of shooting incidents reported in NYC.
- Brooklyn has reported more incidents of shooting crimes than Manhattan, Staten Island and Queens combined from 2006 to 2021.
- Majority of the shooting crimes reported in NYC are targeted towards African Americans.

2) Education, Mental Health, Alcohol Consumption and Crime:

- There is no direct correlation with total enrolment and economic index according to the data. Brooklyn has the highest school enrolment, but is second in terms of economic index need. Bronx ranks third in total enrolment and first in economic need index.
- The highest population enrolled in schools in all 5 boroughs by race is Hispanics, followed by African Americans.
- Majority of the students that have disabilities enrolled in schools across all boroughs are males.
- Mental health disorders that can be attributed to alcohol usage are most prevalent across all 5 NYC boroughs.

Dataset	Accuracy
Crime and Alcohol	63.7%
Crime and Mental Health	51.1%
Crime and Education	42.2%

TABLE I: Classification Accuracy: Type of Crime

Schizophrenia is the biggest cause of mental health disorders as reported. Along with mood disorders, these three are the biggest contributors of mental health disorders in NYC boroughs

- Alcohol consumption statistics have stayed constant in New York City in the last 4 years.
- Manhattan leads in alcohol consumption, but a major portion of it can be attributed to on-premise liquor. Cheaper sources like grocery chains are more common sources of alcohol in boroughs other than Manhattan.
- Top crimes in Manhattan are petty larceny, dangerous drugs, assault, vehicle and traffic violations, and grand larceny.
- Top crimes in Brooklyn are assault, dangerous drugs, petty larceny, penal law violation and felony assault.
- Top crimes in Queens are assault, dangerous drugs, vehicle and traffic violations, penal law violations, and petty larceny.
- Top crimes in Bronx are assault, dangerous drugs, vehicle and traffic violations, petty larceny, and felony assault.
- Top crimes in Staten Island are dangerous drugs, assault, petty larceny, penal law violation and vehicle and traffic violations.

B. Predictive Analytics

Predictive analytics was conducted in two phases. Predicting the type of crime like Felony, Misdemeanor, Violation, or Traffic infraction was the former, and predicting the exact crime like Grand larceny or Forgery was the latter. As expected the former showed good results due to fewer variations in the prediction column. Since there are only 4 categories to predict from, the model performed better on minor fine-tuning and encodings. The authors were motivated to further expand the research to predict the exact crime. The authors believe that predicting the exact crime did not result in very good accuracy due to:

- Less data - The amount of data available for a particular crime, like Grand Larceny, is not adequate enough to be able to build a model. Since the data covers all types of crimes thus the number of crimes increases exponentially reducing the number of per crime.
- Data Preprocessing - Since the number of crimes is exponential the authors decided to restrict the prediction to the top 10 crimes. This resulted in removing a lot of data in the data preprocessing stage.

1) Type of crime prediction: The authors used a Decision tree classifier on pairs of datasets, keeping crime constant, to classify a given record into either Felony, Misdemeanor, violation, or traffic infraction. The table I contains the accuracy of the decision classifier. As the results state, the highest accuracy was achieved when predicting alcohol followed by mental health and education. This can be attributed to the correlation of criminal activities with alcohol and mental health issues.

Classifier	Training Accuracy	Testing Accuracy
Decision Tree	20%	20%
XGBoost	23%	22%
Multiclass Logistic	14%	14%
Naive Bayes	15%	14%

TABLE II: Classification Accuracy : Crime and Alcohol

Classifier	Training Accuracy	Testing Accuracy
Decision Tree	23%	22%
XGBoost	25%	23%
Multiclass Logistic	13%	12%
Naive Bayes	14%	14%

TABLE III: Classification Accuracy : Crime and Mental Health

2) Crime prediction: The following are the observations for predicting the exact crime. The results are not on the very positive side but there are powerful insights about the classifiers. The Random Forest classifier and Voting Classifier took a lot more time compared to others so the authors could not report the findings due to computation limitations. Amongst the rest, the Decision tree and XGBoost classifiers stood out to be the best in most cases. The XGBoost performed better than the Decision tree in 3 out of the 4 cases. The table II depicts the accuracy of crime combined with alcohol. The output of the crime dataset combined with Education is present in the table IV. We have reported the accuracy of crime merged with the mental health dataset in the table III. Finally, all the datasets are combined and the accuracy reported by the classifiers is present in table V.

Classifier	Training Accuracy	Testing Accuracy
Decision Tree	24%	22%
XGBoost	26%	23%
Multiclass Logistic	14%	13%
Naive Bayes	13%	13%

TABLE IV: Classification Accuracy : Crime and Education

Classifier	Training Accuracy	Testing Accuracy
Decision Tree	24%	22%
XGBoost	21%	19%
Multiclass logistic	16%	16%
Naive Bayes	14%	13%

TABLE V: Classification Accuracy: Crime, Alcohol, Mental Health, and Education

VI. Conclusions

This research successfully understands the trends and relationship of crime with education, alcohol, and mental health in different boroughs of New York City. The authors did a comprehensive analysis of each of the domains and found concrete trends. Multiple classifiers with different types of encodings were used on pairs of datasets, crime being constant, to build a robust model which is capable of predicting the type of crime, namely Felony, Misdemeanor, Violation, and traffic infractions. Further, the authors combined all the datasets together and make the prediction more granular by predicting the exact crime.

VII. Future Work

The authors believe this research can be extended to other cities in the United States. With adequate data, high computation power, and fine-tuning, higher accuracy can be achieved. Another advancement could be to further granularize the location by using coordinates to predict crimes inside boroughs as well. This will help collate predictions at the precinct level to optimize police workforce distribution across precincts of New York City. In the future, the prediction of the time of the crime can also be very insightful.

Acknowledgments

Professor Anasse Bari's efforts and prompt responses to all queries regarding this study are acknowledged by the authors. His expertise was essential in helping the authors establish their method for analyzing crime data and build a classifier that incorporated various dataset types. Finally, we thank our TAs and Graders for the conversations we had with them on how to improve the project.

References

[1] R. Hjalmarsson and L. Lochner, "The impact of education on crime: International evidence," *CESifo DICE Report*, vol. 10, no. 2, pp. 49–55, 2012. [Online]. Available: http://hdl.handle.net/10419/167078

[2] R. Yıldız, O. Öcal, and E. Yildirim, "The effects of unemployment, income and education on crime: Evidence from individual data," *International Journal of Economic Perspectives*, vol. 7, 01 2013.

[3] B. S, P. R. Misal. A, and K. S, "Crime rate prediction using data mining algorithms," *Journal of Data Mining and Knowledge Engineering*, vol. 4, no. 1, pp. 21–29, Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2629982

[4] D. Hazra and J. Aranzazu, "Crime, correction, education and welfare in the U.S. – What role does the government play?" *Journal of Policy Modeling*, vol. 44, no. 2, pp. 474–491, 2022. [Online]. Available: https://ideas.repec.org/a/eee/jpolmo/v44y2022i2p474-491.html

[5] T. Kose and N. Orak, "Perceived neighborhood crime and health: a multilevel analysis for turkey," *Safer Communities*, vol. 21, no. 4, pp. 243–259, Jan 2022. [Online]. Available: https://doi.org/10.1108/SC-08-2021-0034

[6] B. Bell, R. Costa, and S. Machin, "Crime, compulsory schooling laws and education," *Economics of Education Review*, vol. 54, pp. 214–226, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272775715001144

[7] P. D. (NYPD), "Nypd arrests data (historic): Nyc open data," Jun 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

[8] ——, "Nypd shooting incident data (historic): Nyc open data," Jun 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

[9] ——, "Nypd hate crimes: Nyc open data," Oct 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78

[10] N. Y. C. D. of Education, "2017-18 - 2021-22 demographic snapshot: Nyc open data," Jun 2022. [Online]. Available: https://data.cityofnewyork.us/Education/2017-18-2021-22-Demographic-Snapshot/c7ru-d68s

[11] Health.data.ny.gov, "Hospital inpatient discharges (sparcs de-identified): 2016," Feb 2021. [Online]. Available: https://healthdata.gov/State/Hospital-Inpatient-Discharges-SPARCS-De-Identified/nff8-2va3

[12] N. Y. S. D. of Health, "Hospital inpatient discharges (sparcs de-identified): 2016: State of new york," Sep 2019. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/gnzp-ekau

[13] ——, "Hospital inpatient discharges (sparcs de-identified): 2017: State of new york," Oct 2019. [Online]. Available: https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-z7e7

[14] ——, "Hospital inpatient discharges (sparcs de-identified): 2018: State of new york," Sep 2022. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/yjgt-tq93

[15] ——, "Hospital inpatient discharges (sparcs de-identified): 2019: State of new york," Sep 2022. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/4ny4-j5zv

[16] N. Y. S. L. Authority, "Liquor authority current list of active licenses: State of new york," Dec 2022. [Online]. Available: https://data.ny.gov/Economic-Development/Liquor-Authority-Current-List-of-Active-Licenses/hrvs-fxs2

[17] N. O. D. : City of New York, "Nyc open data." [Online]. Available: https://opendata.cityofnewyork.us/data/

[18] "Statistics." [Online]. Available: https://www.nyc.gov/site/nypd/stats/stats.page

[19] "Citywide crime statistics." [Online]. Available: https://www.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page

[20] "Nypd arrests data (historic)." [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/data

[21] P. D. (NYPD), "Nypd hate crimes: Nyc open data," Oct 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78

[22] ——, "Nypd shooting incident data (historic): Nyc open data," Jun 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

[23] N. Y. C. D. of Education, "2017-18 - 2021-22 demographic snapshot: Nyc open data," Jun 2022. [Online]. Available: https://data.cityofnewyork.us/Education/2017-18-2021-22-Demographic-Snapshot/c7ru-d68s

[24] N. Y. S. D. of Health, "Hospital inpatient discharges (sparcs de-identified): 2016: State of new york," Sep 2019. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/gnzp-ekau

[25] "Community health survey public use data." [Online]. Available: https://www.nyc.gov/site/doh/data/data-sets/community-health-survey-public-use-data.page

[26] "Liquor authority current list of active licenses." [Online]. Available: https://data.ny.gov/Economic-Development/Liquor-Authority-Current-List-of-Active-Licenses/hrvs-fxs2/data