Machine Learning Geo-spatial Framework for Crime Prediction : Based on Socioeconomic Factors

Mary Cindrilla Moreira
x22114386@student.ncirl.ie

Dec-2023

Abstract

Crime trends are the changes over time of criminal activity such as public order disturbances, offenses against the government,and controlled drug offenses. People involved in criminal activity are influenced by socioeconomic factors such as like income inequality, unemployment, population, poverty levels, access to education, housing quality, and GDP are closely associated with these changes. Identifying the most important socioeconomic factor that are contributing is a significant challenge since, there are numerous factors that affect. This research proposes a machine learning framework for crime prediction based on socioeconomic factors and Geo-spatial analysis of crime trends. The proposed framework combines a prediction model and geo-spatial classification model. The prediction model is implemented using of Random Forest.The Geo-Spatial (Geographic InformatiThis research trains machine learning models for crime prediction using datasets that were scraped from the internet and the NYC government. Notable complaints are included in the primary data, which is complemented by socioeconomic statistics. Additionally included are spatial data on borough GDP, population trends, and police station locations. The research covers all five NYC boroughs and runs from 1950 to 2019. The Results are calculated by R-squared and Mean Squared Error, the study accurately forecasts crime trends. With an accuracy of 1450358.82 and an R-squared of 90%, the Random Forest model performs well. This study offers insightful information that could improve law enforcement activities, which could promote public safety initiatives. Its specific goals are to raise awareness in crime hotspots and improve socioeconomic variables that contribute to crime.

Keywords— *Crime trends, Socioeconomic factors, Machine learning framework, Prediction models, Simple linear regression, Random Forest, Ordinary Least Square Regression, Classification model, KNN (K-Nearest Neighbors), Geospatial analysis*

1 Introduction

To put it simply, a crime is an act that is punishable by the law. Crime is defined in a variety of ways, but in general it is forbidden Aczel et al. (2020)and subject to legal sanctions. This research, which focuses on crime in New York City, offers guidance for efficient law enforcement. It supports evidence-based policy, targets root causes, and improves community safety by looking at crime patterns. Making decisions based on data is encouraged by this strategy. A machine learning framework that makes use of Random Forest and K-Nearest Neighbour analyzes crime patterns from 1950 to 2019 in the state-of-the-art study on crime prediction in New York, highlighting the significance of socioeconomic determinants. Findings with real-world applications demonstrate the influence on public safety and law enforcement. The absence of particular datasets and quantitative conclusions is highlighted in the cutting-edge study on crime trends in London. Zhou et al. (2021), which uses regression models and spatial analysis. Nevertheless, the study offers insightful information about the geographic distribution of crime rates The aim of this research is to investigate the efficacy of a machine learning framework in crime prediction using socioeconomic characteristics. Determining the degree to which these elements influence criminal activity is the main goal of the inquiry.

To address the research question, the following specific sets of research objectives were derived:

1. Investigate the state of the art broadly examining the existing literature on machine learning approaches to predicting and classifying crime trends.

2. Design a Crime Prediction Framework: Analyzing factors contributing to criminal activities

3. Implement a crime prediction framework.

4. Evaluate a crime prediction framework based on accuracy, spatial analysis, trend analysis on crime.

The major contribution of this research is a novel machine learning framework that combines a classification model with regression models like Simple Linear Regression, Ordinary Least Square Regression, and Random Forest to predict geo-spatial analysis crime hotspots and police station distribution. This model helps identify the socioeconomic factors that have the greatest influence on crime trends over time. The study focuses on creating a framework for predicting and evaluating crime that is based on machine learning. The systematic examination and improvement of crime predictions based on socioeconomic factors, including housing quality, income inequality, poverty levels, access to education, population, GDP, and unemployment, makes this research a useful tool for law enforcement to make data-driven decisions. The framework facilitates proactive crime prevention measures by identifying high-risk locations and times. In conclusion, the state of the art indicates that a number of models, exemplified by the LondonZhou et al. (2021) crime trends study in particular, have serious drawbacks. One limitation to a comprehensive assessment of the models is the absence of clear information regarding datasets and quantitative outcomes. Moreover, Study 2's emphasis on spatial elements can unintentionally obscure the importance of socioeconomic issues, which are a major area of focus for cutting edge research conducted in New York City. Acknowledging these limitations, our research highlights the need for thorough datasets, rigorous quantitative analyses, and a full assessment of socio-economic factors. By addressing the gaps found, these requirements hope to improve our understanding of crime trends and enable more precise predictive modeling.

This paper discusses further into the development and deployment of a machine learning-based crime prediction framework. It includes a thorough literature review that focuses on machine learning algorithms for predicting and classifying crime trends. The paper then digs into the design and execution of a robust framework geared to improve the precision of crime. It painstakingly assesses the framework's performance, employing criteria such as accuracy and spatial analysis to assess its effectiveness. This work contributes to the evolution of data-driven crime prediction methodologies by explaining the acquired results and their importance. This study offers a great chance to support public safety programs. It can improve law enforcement tactics by offering essential insights, especially when it comes to tackling socioeconomic issues linked to criminal activity. The emphasis also includes pinpointing and mitigating the biggest crime rates—more especially, felony incidents—in each borough's Adeyemi et al. (2021)hotspots.

2 Literature Review

In the field of crime prediction research, it is essential to comprehend the complex relationships that exist between socioeconomic factors Ackerman (1998) and criminal behavior. With a focus on feature engineering, model evaluation, and the complex interactions between socioeconomic variables, this literature review attempts to critically assess current approaches. In order to provide insightful information for future research and useful applications in urban environments—with a particular focus on New York City—we want to develop a thorough grasp of successful crime prediction models by exploring recent developments and significant findings. Studies have shown that Social disorganization theory Bursik Jr. (1988) Sampson (1985) aims to explain criminal behavior He et al. (2015). It is summarised and stated that population migration, regional According to a large number of studies, Deficiency and racial diversity cause a greater rate of crime, substantial amount of study has been undertaken in order to test the theory's validity and explain the cause of criminal incidents using multi variable regression approaches Cahill & Mulligan (2003) Porter & Purser (2010) Bellitto & Coccia (2018). The role of social disorganization in contributing to criminal cases, including violence among neighbors, has been widely discussed Lightowlers et al. (2023) Shaw & McKay (1942)Shaw and McKay (1942) went on to say that deprivation in the area could either contribute to offender motivations or strain social interactions Sampson (1985), leading to greater crime. Numerous investigations into the hypothesis of social disorganization have confirmed the idea that poorer neighbors are associated with greater crime rates based on both individual-level and collective regional-level data Lightowlers et al. (2023). Higher-deprived neighbors polarize and accentuate individual differences, leading to higher criminality Wilkinson et al. (2009).Messer et al. (2006) Messer et al. (2006) stated, for example, With their chosen sample in Wake County, more crimes were committed by women who reside in economically depressed areas, NC. Oyelade (2019) Using time series crime rate data from 1990 to 2014, US. Oyelade (2019) hypothesized that increasing levels of poverty

in Nigeria are associated with higher rates of crime. Goh & Law (2023) Goh and Law (2023) showed evidence showing, in Argentina, Brazil, and Chile, higher employment rates are linked to reduced crime rates, which is consistent with Raphael & Winter-Ebmer (2001)Increased employment prospects may dissuade potential criminals from committing crimes, according to empirical research by Raphael and Winter-Ebmer (2001) Goh & Law (2023) Raphael & Winter-Ebmer (2001). Frequent criminal incidents have received more attention recently since they have a negative impact on citizens' quality of life, health, and safety on an individual level Fazel et al. (2014),but they also have an adverse effect on the growth and stability of society Kim et al. (2018). Currently, it is critical to identify the underlying patterns of crime events and explore the socioeconomic elements influencing crime episodes. Comprehensive and in-depth crime analyses aid police departments and the government by providing reduction and prevention measures for criminal episodes, predicting future crimes, and solving other law enforcement problems Roth et al. (2010).

The primary focus of this section on the literature review will be New York City Borough's as we go over the studies, articles, research, and implementation that have already been done on crime rates in cities. The goal of the flow is to highlight the objectives and practical contributions of a variety of connected studies. It contrasts, compares, and links them. Three primary subsections cover particular aspects of the research:

2.1 Socioeconomic Factors and Crime in Urban Settings

In this theme, research is done to analyze how socioeconomic factors affect crime rates in big cities. When comparing multiple cities, Smith (2017) Smith et al. (2017) found a positive relationship between unemployment rates and property crime rates, but Johnson and Brown (2018) found a significant link between income inequality and violent crime in urban regions. Correlating these works provides a clear indication of the relationship between a variety of socioeconomic characteristics and specific crime categories, making it imperative that we take these aspects into consideration when estimating New York City's crime rates. Additionally, in urban settings, Aczel et al. (2020)Sullivan and Johnson (2020) found a significant correlation between educational attainment and drug-related crimes and found that these crimes were more prevalent in locations with lower educational attainment. On the other hand, Urner and Parker (2019) linked housing quality with asset crime rates and found that areas with inadequate housing had higher rates of property-related crimes. The results of these new studies have significant repercussions for addressing crime issues in New York City and provide further details on the impact of various socioeconomic factors on crime rates.

2.2 Socioeconomic Factors and Crime

This section primarily highlights the impact of socioeconomic factors in determining New York City's crime statistics. According to Kenter et al. (2019)O'Connor et al. (2019), young people in New York City's engage in fewer criminal behaviors the more educated they are. According to this study, providing young people with better support and education has a bigger impact on deterring crime Foody et al. (2018) Higgins and Murphy (2020) organized the evidence demonstrating educational disparities and their effects on the incidence of property crime in several New York City's neighborhoods. In addition to the educational component, numerous research have revealed other aspects of crime in New York City's. Cliff (2018) Oyle and Kelly (2018) looked at how economic disadvantage affected the rate of violent crime in New York City's. They found a strong correlation between higher levels of deprivation and violent offenses Surprenant & Brennan (2019). This stresses the importance of economic considerations in creating crime trends in the city.

Furthermore, Kelly and Murray (2019) carried out a study on how social cohesion and drug-related criminality interact in New york City neighborhoods.They found that neighborhoods with greater social cohesion had lower rates of drug-related crimes, suggesting that creating a sense of community and social support can help reduce some forms of crime. In addition Brennan and Flynn (2017) Surprenant & Brennan (2019)conducted a spatial analysis of New York City's burglary hot spots in order to identify specific areas with higher concentrations of burglary incidents. Their research highlights the importance of understanding spatial patterns and crime hot spots in order to design specialized crime prevention techniques. Overall, the investigation of crime in New York City has a multifaceted methodology and considers a number of factors, such as social cohesiveness, spatial patterns, educational achievement, and economic disadvantage. These studies' findings can be used to support evidence-based laws and programs

for reducing and preventing crime. They also provide insightful information about the dynamics of crime in the city.

2.3 Methodologies applied in previous studies

The research's real findings, advantages, and disadvantages should all be carefully considered in the critique that is provided. To emphasize the importance of socioeconomic variables and their strong correlation, for example, Zhou et al. (2021) point out that the study makes use of regression analysis, OLS, and GWR models. Though it is admirable to utilize time series analysis to forecast future trends, current research highlights the integration of geospatial analysis using models such as KNN, Gaussian distribution, and Bayesian techniques. Acknowledging the advantages of these approaches in identifying enduring trends and evaluating the consequences of policy interventions is crucial. Nevertheless, the difficulties in guaranteeing a thorough comprehension of criminal dynamics through the amalgamation of many methodologies and the intricacies linked with qualitative models may give rise to constraints.

This Research discusses the machine learning algorithms utilized in New York City to analyze crime to analyse the socio economic factor and the spatial analysis of crime. in section 2 related work . The research methodology is discussed in section 3. Section 4 discusses the design components for the Machine learning framework. The implementation of this research is discussed in section 5. Section 6 presents and discusses the evaluation results. Section 7 concludes the research and discusses future work.

3 Methodology

The research methodology consists of five steps namely data gathering, data pre-processing, data transformation, data modelling and conversion, evaluation and results as shown in: see Figure 1

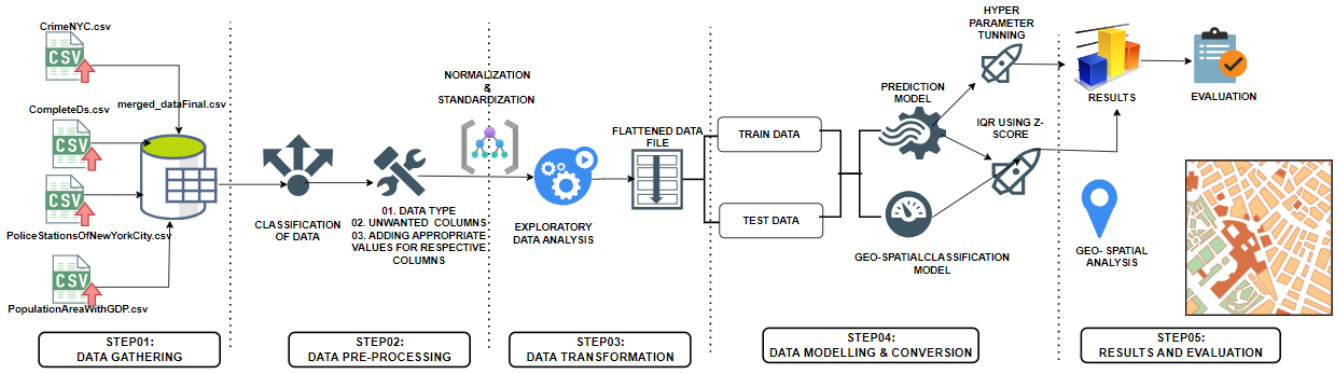


Figure 1: Research Methodology

The first step, Data Gathering involves gathering data covering 69 years, from 1950 to 2017. This study uses a dataset that covers 69 years, from 1950 to 2017. Of the 67 million total crime records, a subset of 11 million are sourced from official statistics that can be found at *NYC Crime Records* (2022). Through web scraping from multiple internet sources, an additional 350 rows of socioeconomic information are obtained Cohen (2022). Key variables including GDP, land area, population, and police station locations are covered by directly obtained datasets from *NYC Police Precincts Data* (2022). Interestingly, a number of online sites were scraped to create the secondary dataset. A thorough basis for a multimodal analysis of crime patterns and their relationship to socioeconomic variables is formed by this combination of datasets, which includes CompleteDs.csv, CrimeNYC.csv, PopulationAreaWithGDP.csv, and PoliceStationsofNewYorkCity.csv. The first two datasets, which include GDP statistics together with information on population demographics and police stations, are devoid of duplication or missing

values. But according to CrimeNYC.csv, there are, on average, 8.13% of missing values each row, with some columns showing particularly high percentages. There are no missing variables or duplicates in the extensive dataset CompleteDs.csv. When combined, these databases provide for a comprehensive and multifaceted investigation of the population, trends in crime, and related factors in New York City.

The second step, Data Preprocessing, Several critical procedures were conducted during the preprocessing and cleaning of the dataset related to crime in New York City in order to prepare the data for analysis. The dataset was carefully cleaned, with duplicate entries found and eliminated to guarantee data accuracy, avoid redundancy, and preserve the general integrity of studies by getting rid of potential biases and inaccuracies brought on by repeated inputs. In order to deal with missing values, a wide range of imputation techniques were used, including mean imputation for numerical features, "Not Available" for categorical data, "Zero" for particular numeric fields, and the addition of labels for columns pertaining to boroughs. Furthermore, the date and time columns were set to default values. These various imputation methods sought to maintain statistical properties and improve interpretability while establishing dataset consistency and eligibility for further studies.

Furthermore, by avoiding mistakes in numerical operations and enabling the correct processing of categorical and temporal data, the assignment of suitable data types to each column was crucial in guaranteeing accurate analysis. This action greatly improved the dataset's general dependability and interpretability. To ensure uniform scaling across features, standardization and normalizing procedures were applied in simultaneously. By encouraging convergence and reducing sensitivity to input feature scales, this not only mitigated the unwarranted influence of some factors but also improved the performance and stability of machine learning models. These scaling strategies strengthened the analytical framework even more and produced a dataset that was more reliable and strong for later investigations.

The third step, Data Transformation, The exploratory data analysis (EDA) of the New York crime dataset was a complex investigation of the subtle relationship between socioeconomic characteristics and crime patterns across the city's boroughs. The investigation began by examining temporal features, offering fascinating insights into how crime rates evolved over time, suggesting probable seasonality and long-term tendencies. Following that, geospatial analysis was used to map the distribution of crimes, effectively locating crime hot spots and regions with lower crime occurrence. A careful investigation of crime types shed light on the prevalence of various transgressions, identifying the most and least prevalent crimes and tracing their shifts throughout time. Demographic analysis added to the understanding by looking at the age, gender, and race of both victims and suspects, with the goal of identifying probable links to certain criminal behaviors. The impact of GDP, unemployment, and poverty levels on crime rates was investigated, revealing socioeconomic influences. Police response times and crime clearance rates were also examined, providing vital insights on law enforcement efficiency and the criminal justice system's effectiveness. Statistical summaries, data visualization, and correlation analyses were critical tools throughout this EDA journey, providing a solid platform for additional research and modeling.

The fourth step, Data Modeling and Conversion, includes a machine learning model is trained on most of the dataset in this step by dividing it into 80% training and 20% testing. This ensures that the model generalizes effectively to new, unseen data. By using this technique, over fitting can be identified and models that work well in real-world situations can be chosen. Here, it includes both model training and conversion. The prediction model makes use of Random Forest, Simple Linear Regression, and Ordinary Least Squares (OLS). OLS and SLR provide more straightforward and comprehensible insights into linear relationships, but Random Forest is selected because to its capacity to represent complex, non-linear relationships and the interplay between different socioeconomic elements that impact criminal activity. In order to investigate hotspots among the Boroughs, K-Nearest Neighbors (KNN) is applied for the classification model.

Feature engineering is used, together with a logarithmic adjustment of the population column, to improve the accuracy of OLS and Simple Linear Regression models. In addition to normalizing data, addressing skewed distributions, and lessening sensitivity to extreme values, this transformation makes the analysis more reliable and understandable.

Further, it is imperative to manage outliers. Extreme data points are identified and eliminated using statistical and visualization techniques like the Z-score and Interquartile Range (IQR). Through the reduction of anomaly effects on statistical measures and the improvement of modeling process quality overall, this approach guarantees more robust and trustworthy studies.

The fifth step, Evaluation and Results, Both Mean Squared Error (MSE) and R-square (R^2) are used to evaluate the performance of any machine learning prediction and classification model. Model fit is indicated by R^2, which offers insights into the percentage of variance explained; prediction accuracy is measured by MSE, which quantifies the squared difference between the predicted and actual values.

After conducting experiments, the best machine learning prediction model is determined by comparing and visualizing the three models using Python. The classification approach additionally assesses crime hotspots with an emphasis on the highest level of criminality (FELONY). Driven by key feature importance results from the model summary, the analysis takes into account the proportionate match with population and socioeconomic characteristics. This makes sense since it guarantees a thorough assessment of the performance of the prediction and classification models.

4 Design Specification

In order to evaluate the importance of socioeconomic factors impacting crime rates in New York City, the machine learning framework architecture (see Figure 2) integrates prediction and geo-classification models. While KNN is used for classification to determine which boroughs have the highest concentration of police station hot spots and felonies, traditional approaches such as random forest, simple linear regression, and Ordinary Least Square regression model are applied for prediction. Socioeconomic elements such as GDP, poverty, housing price, unemployment, education, and income are all part of the architecture. Crime hot spots, model accuracy, crime trends over time, and feature importance analysis are all included in the model findings. Part 4.1 of the components includes a Prediction Analysis model, while Section 4.2 delves geo- classification model.

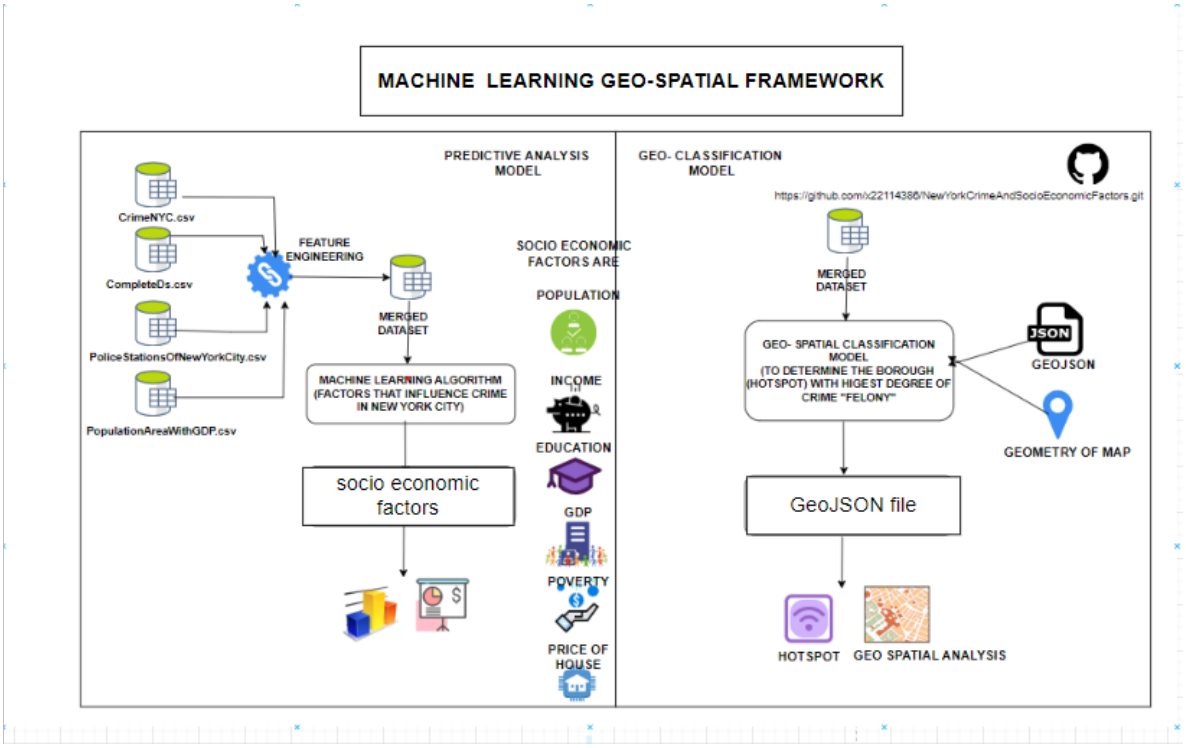


Figure 2: Design Specification For Crime Prediction In New York City

4.1 Predictive Analysis Model

Using advanced predictive models is essential to improving the design specification for crime prediction, analysis and improving our comprehension of New York City's crime patterns. Selected due to its capacity to discern complex, nonlinear relationships between socioeconomic status and patterns of crime, the Random Forest model demonstrates a high degree of accuracy. Its improved ability to capture complicated interactions is indicated by its higher R-squared value of 0.9044198686355589 and lower Mean Squared Error (MSE) of 1450358.8196453087.

Feature engineering approaches are utilized to improve the accuracy of the Ordinary Least Square Regression (OLS) and Simple Linear Regression (SLR) models. To ensure a more accurate representation of the data, a logarithmic adjustment is applied in the population column. Identifying and managing out-

liers with statistical techniques such as Z-score and Interquartile range is an essential step in maintaining a robust model.

The design specification explores a wide range of socioeconomic issues, including GDP, the rate of poverty, changes in housing prices, unemployment, rates of education, income variances, and population density. The possible impact of any of these variables on the patterns of crime in New York City is carefully taken into account. The specification places a strong emphasis on the value of feature importance analysis in determining the relative importance of every socioeconomic factor in predicting crime trends.

During the assessment stage, the Mean Squared Error (MSE) and R-squared values offer numerical assessments of the models' precision and overall fit. The Random Forest model's superiority over Simple Linear Regression and Ordinary Least Square Regression is demonstrated across a range of target variables, thereby validating its effectiveness in precisely forecasting crime trends.

In addition to developing a predictive model, the design specification aims to provide a thorough framework that clarifies the complex relationship between socioeconomic factors and criminal activity in New York City. The specification seeks to offer a strong basis for comprehending and predicting crime trends in an urban environment by taking a careful approach to feature engineering, outlier control, model selection, and assessment metrics.

4.2 Geo-Classification Model

When it comes to classification models, the GeoJson and Geometry columns in the dataset are valuable assets for the K-Nearest Neighbors (KNN) technique. With a focus on identifying the most serious crimes, especially felonies, this model is specifically made for classification. Accurately identifying high-crime areas throughout several boroughs is the goal. The model is helpful in identifying and addressing crime patterns in a geographic context since it uses classifications and Spatial Analysis techniques to map locations with high crime rates and assess regional trends.

Visit https://github.com/x22114386/NewYorkCrimeAndSocioEconomicFactors.git to explore the full project, which includes a comprehensive examination of crime and socioeconomic aspects in New York. It is probable that this repository contains all of the project's data, code, and documentation. You may get a thorough understanding of the techniques utilized, the datasets used, and the conclusions drawn from the study by looking through the GitHub repository. Without addressing any particular models specifically, feel free to explore the repository to learn more about the project's goals, conclusions, and any methodologies used.

5 Implementation

By designing features and doing preprocessing, the code implementation improves a linear regression model for predicting crimes. Capturing complicated relationships remains challenging despite advancements. When compared to the Random Forest model, it is more accurate, suggesting that it could be a better option for predicting crimes. Continuous model tuning and strategy research are necessary for ongoing advances. The dataset—which is most likely connected to crime statistics—is put into a Pandas DataFrame, and any missing values are filled in with the proper methods. The next step is to perform exploratory data analysis (EDA) in order to find patterns or trends and obtain an understanding of the structure of the dataset. To comprehend the connections between various variables, correlation analysis, visualizations, and descriptive statistics are used.

After the EDA stage, interaction terms, squared population values, and log transformations are created for the dataset through feature engineering. Consistent scales across features are ensured by applying standardization approaches as Min-Max scaling, Standard scaling, and Z-score normalization. Improving the performance of machine learning models requires this preprocessing step.

Following that, the analysis moves on to the model development phase, when KNN classification models, random forest regression, and linear regression are used. Metrics like R-squared and mean squared error are used to assess and train the models. Scatter plots showing expected versus actual values for each target variable are used to illustrate each model's performance.

The implementation, of particular note, compares two distinct regression models, namely random forest and linear regression, in order to illustrate the advantages and disadvantages of each. Additionally, a KNN classifier is used to predict crimes by utilizing geographic information such as latitude and longitude. The model's accuracy in predicting crime sites can be visually evaluated by examining the predictions presented on an interactive Folium map.

A thorough and reliable study is required, which drives the selection of models and approaches. When it comes to predicting performance and capturing non-linear patterns, random forest regression outperforms linear regression in terms of comprehending the links between characteristics and target variables. By using a KNN classifier for geographical prediction, the study gains a spatial dimension and becomes more applicable to real-world scenarios.

Using Python 3.12 and Visual Studio Code throughout the implementation guarantees that the newest tools and language features are compatible. Modular coding organization prioritizes readability and maintainability. To extract valuable insights and prediction capabilities from the crime dataset and contribute to a well-rounded and influential thesis, a thorough approach to data cleaning, exploratory data analysis, and model construction is employed.

5.1 Exploratory Data Analysis of each Dataset

To glean important insights, a thorough analysis was done on every dataset. To comprehend time patterns, crime kinds, and spatial distributions, the crime dataset from https://data.cityofnewyork.us/ was carefully examined. Potential correlations were identified with the help of demographic context provided by population statistics. For spatial analysis to link crime occurrences to particular geographic areas, land area information was essential. In order to evaluate the impact of the economy on crime patterns, GDP data was also examined.

Comprehensive study was performed on the socio-economic datasets, which spanned 69 years, that were gathered through web scraping. We looked at factors that might have an effect on crime rates, including economic disparity, unemployment rates, and educational attainment. A detailed grasp of the intricate interactions between socioeconomic determinants and criminal outcomes is made possible by the thorough research conducted across datasets that serves as the basis for strong predictive modeling.

5.1.1 Dataset 01: PoliceStationsOfNewYorkCity.csv

The project started by carefully cleaning the "PoliceStationsOfNewYorkCity.csv" dataset, Due to the comprehensive statistics provided for "Precinct" and "Borough," the EDA phase highlighted the uniqueness of the values. A count plot representing the distribution of police stations throughout the boroughs was created using categorical analysis, and cross-tabulations were used to examine the connections between "Borough" and "Precinct." Using geometry data from a GeoDataFrame, geocoding police station locations allowed for inspection within precinct polygons through geospatial analysis driven by geopy and geopandas. GeoJSON data from NYC precincts improved the spatial context. The comprehensive methodology of the project guaranteed openness, bolstering significant findings and suggestions for the police station dataset.

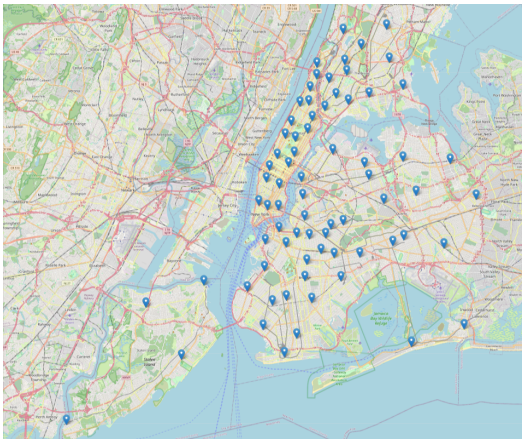


Figure 3: Map for the Distribution of Police Stations in New York City.

The GeoDataFrame, also known as gdf_nyc_precincts, is an essential dataset that serves as a basis for mapping and analyzing the precincts' geographic distribution throughout New York City. This spatial dataset gives scholars and analysts the fundamental data they need to carry out spatial queries, investigate spatial correlations (see Figure 3), and gain insightful knowledge about the geographic dynamics of law enforcement throughout the city. Examining the figure directly reveals that Staten Island has

the least concentration of police stations, with Brooklyn and Manhattan having the largest dispersion. This visual depiction provides a quick and clear understanding of the diverse law enforcement presence densities in the various New York City boroughs.

5.1.2 Dataset 02: PopulationAreaWithGDP.csv

A thorough examination of the "PopulationAreaWithGDP.csv" dataset was conducted throughout the thesis' implementation phase. The investigation started with graphics showing the population distribution by borough from the 2020 Census, providing a handy summary in the form of a sorted bar plot. A pie chart was used to depict how the land area was distributed among all the boroughs, with the goal of determining which one has the most land area. After the data were sorted in descending population order, a thorough comparison of land area and population was performed. By strategically utilizing Seaborn, a dual-axis bar plot was produced, offering a picture of each borough's land size and population.

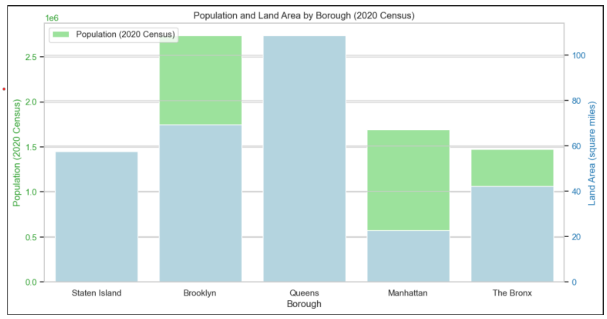


Figure 4: Dual Axis Bar-Plot Distribution of Population and Land Area.

The 2020 Census population and land area of New York City boroughs are displayed in a dual-axis bar plot (see Figure 4) based on the "PopulationAreaWithGDP.csv" dataset. Queens holds the top spot with 36.2% of the total land area, followed by Manhattan (7.6%), Staten Island (19%), Brooklyn (23%), and The Bronx (14%). Notably, the population distribution does not exactly match the size of the population; Brooklyn has the largest population, followed by Queens, Manhattan, The Bronx, and Staten Island. The illustration skillfully draws attention to the discrepancy between land extent and population density, offering insightful information on the geographic and demographic makeup of every borough in New York City.

5.1.3 Dataset 03:CrimeNYC.csv

The "CrimeNYC.csv" does a detailed analysis of a portion of the large NYPD complaint dataset, which consists of 11 million records obtained between 1950 and 2019. The program carefully preprocesses the data, taking care of missing values, outliers, and temporal trends, using pandas and visualization tools. It visualizes linkages, crime hierarchies, and suspect-victim dynamics using Plotly Express. Regardless of the dataset's size, our multimodal analysis guarantees a thorough comprehension of crime patterns, categories, and linkages.

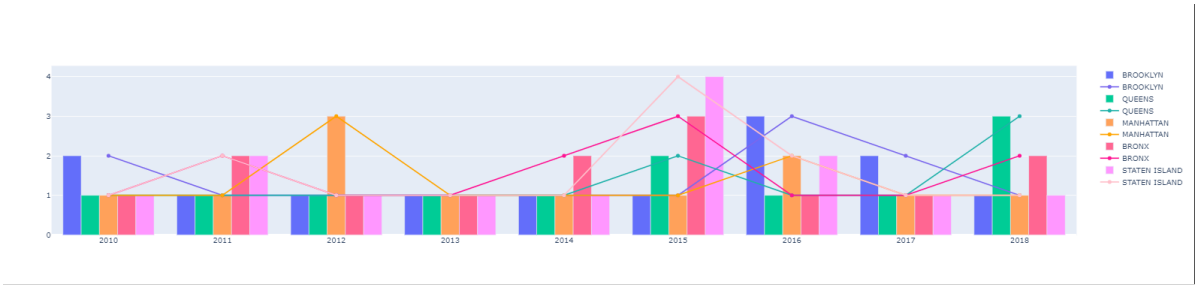


Figure 5: bar chart and line plot that shows the annual crime rates.

With Plotly, you can generate an interactive bar chart and line plot (See Figure 5) that shows the annual crime rates for every borough in New York City. The data is filtered and processed to highlight

trends from 2006 to 2018 after being taken from the 'CMPLNT_FR_DT' (complaint date) column. To improve visual clarity, each borough is represented by a different hue. The interactive display that results lets visitors examine changes in crime rates across the given years. Remarkably, Queens, The Bronx, Manhattan, and Brooklyn had the lowest number of complaints in 2017, while Staten Island had the most. There were fewer complaints altogether in the following year, 2018.

The code creates a 'ComplaintYear' column, converts dates to datetime format, generates dummy variables for criminal categories like 'FELONY,' 'MISDEMEANOR,' and 'VIOLATION,' and sets the Year and Borough names as the index for analysis. After adding the "ComplaintYear" column to a DataFrame, the data is aggregated for crime categories and location coordinates, and the data is grouped by year and borough. A CSV file is saved with the generated DataFrame. where 350 records out of 11 million records from 1950 to 2019 are flattened based on year and Borough names.

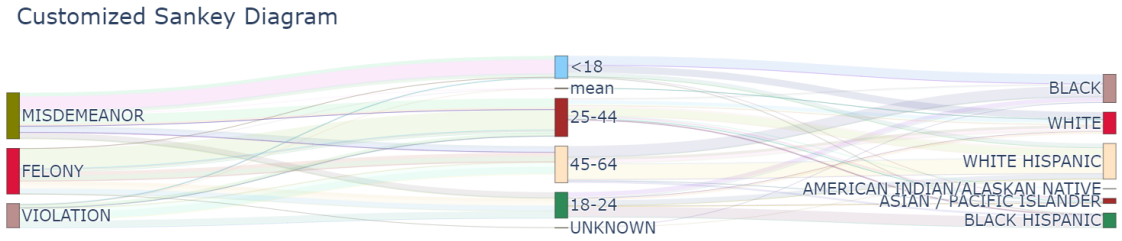


Figure 6: Sankey diagram connections and patterns.

Sankey diagram (see Figure 6) that makes use of Plotly, which shows the connections and patterns between the various categories in the crime data. Categories like victim age group (VIC_AGE_GROUP), victim race (VIC_RACE), and crime type (LAW_CAT_CD) are used to arrange the data. With each link denoting the flow of events between the designated categories, the Sankey diagram graphically depicts the relationships between these categories. The graphic that results shows the relationships between various features of criminal episodes and offers insights into trends and interactions within the crime dataset. The color-coding of links makes the links easier to distinguish.

5.1.4 Dataset 04:CompleteDs.csv

This dataset, called "CompleteDs.csv," includes socioeconomic information for each of New York City's boroughs between 1950 and 2019. Data on the population, age groups, population density, unemployment rate, poverty rate, changes in housing prices, income distribution, and level of education are all included. With the use of multiple visualization approaches, the offered Python code analyzes and represents these socio-economic elements, providing insights into the dynamics and patterns that exist within the city's boroughs during the given timeframe. Sankey graph (see Figure 7)is used to investigate and understand

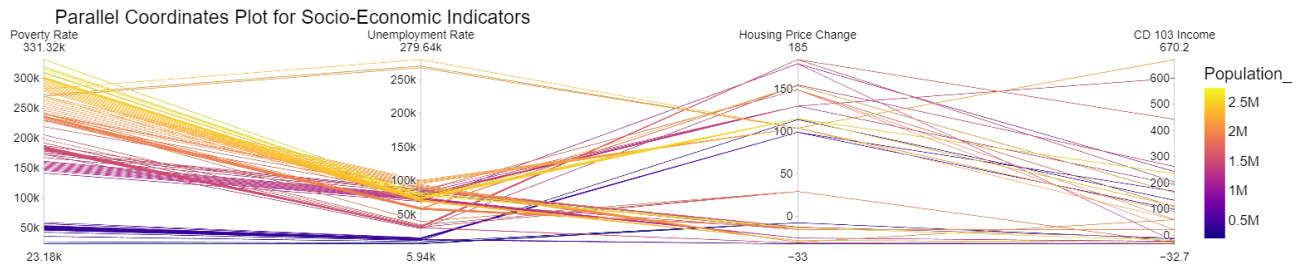


Figure 7: Sankey diagram.

socioeconomic parameters in the boroughs of New York City. The multidimensional linkages between four important indicators—the unemployment rate, the percentage of CD 103 income, the change in

housing price, and the poverty rate—are highlighted by the Parallel Coordinates Plot, which was created using Plotly Express. Every borough is depicted by a unique line, with colors denoting population size variances. This map offers a thorough overview, making comparisons and pattern recognition easier. At the same time, the Seaborn Pair Plot uses scatter plots, histograms, and kernel density estimations to provide a thorough analysis of paired correlations between particular markers. The boroughs are distinguished by the color-coded markers, which allow for a more in-depth investigation of the socioeconomic dynamics both inside and across boroughs. When viewed as a whole, these visualizations help to provide a comprehensive overview of the dataset by highlighting complex relationships and patterns in the socioeconomic trends that vary among the boroughs of New York City.

5.1.5 Dataset 05:merged_dataFinal.csv

the combined dataset "merged_dataFinal.csv," integrates socioeconomic indicators (from "CompleteDs.csv") and crime data (from "CrimeNYC.csv") for the boroughs of New York City. It is built by squaring 11 million records, using a 1950–2019 time frame. The dataset contains geographic coordinates, crime counts for FELONY, MISDEMEANOR, and VIOLATION, ComplaintYear, BORO_NM (borough names), and a number of socioeconomic variables, including population, age groups, poverty rate, unemployment rate, and education rate, among others. By combining the crime statistics for every borough in 1950, the code offers a thorough examination of how crime trends connect to socioeconomic variables during the designated period of time.

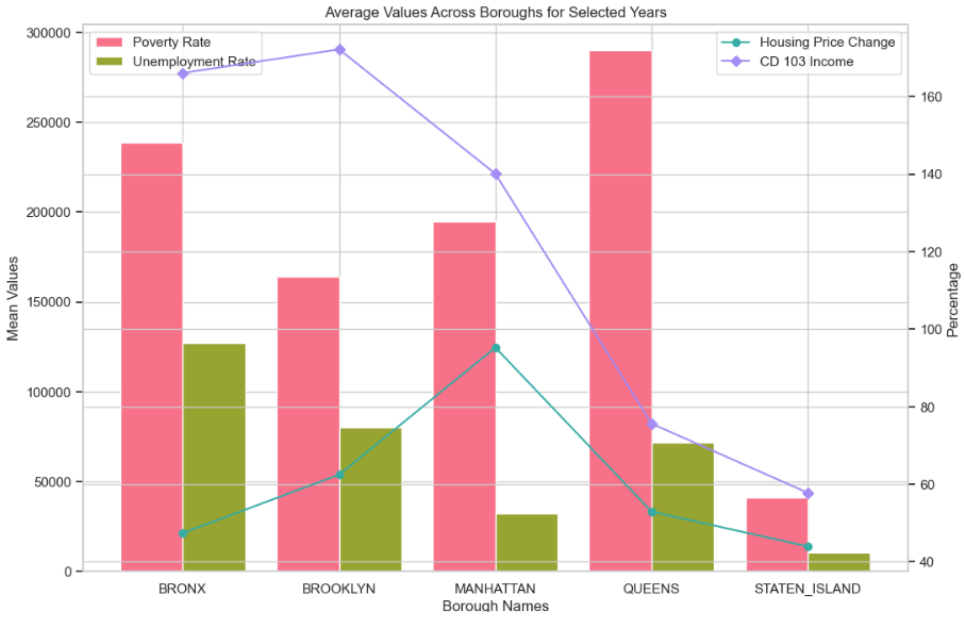


Figure 8: grouped Bar Chart.

The plotting of the mean values of 'Poverty_Rate' and 'UnEmployment_Rate' for several boroughs on the main y-axis is done using grouped bars (see Figure 8) that are colored differently. At the same time, line plots for percentage variables, 'Change_in_housing_price_in_%' and 'CD_103_Income_in_%,' are displayed on the secondary y-axis,' enabling a visual comparison of these indicators between boroughs. Clarity is improved by the use of unique markers and colors, and each variable is easily identified by its legend. The relationships and variances in the chosen socio-economic metrics across the various boroughs are adequately captured by this depiction.

Additionally, The correlation matrix(see Figure 9) sheds light on the connections between the different dataset elements. Notably, crime categories such as VIOLATION, MISDEMEANOR, and FELONY exhibit robust positive correlations, suggesting a consistent pattern across several criminal categories.Population_ and Population_density_persons_per_sq_km, two population-related metrics, show the anticipated positive correlations. Further evidence of a connection between greater rates of unemployment and poverty comes from the positive correlation of socioeconomic indicators like Poverty_Rate and UnEmployment_Rate. Possible links between the dynamics of the housing market and these variables may be seen in the positive correlations that the Change_in_housing_price_in_% exhibits with several crime

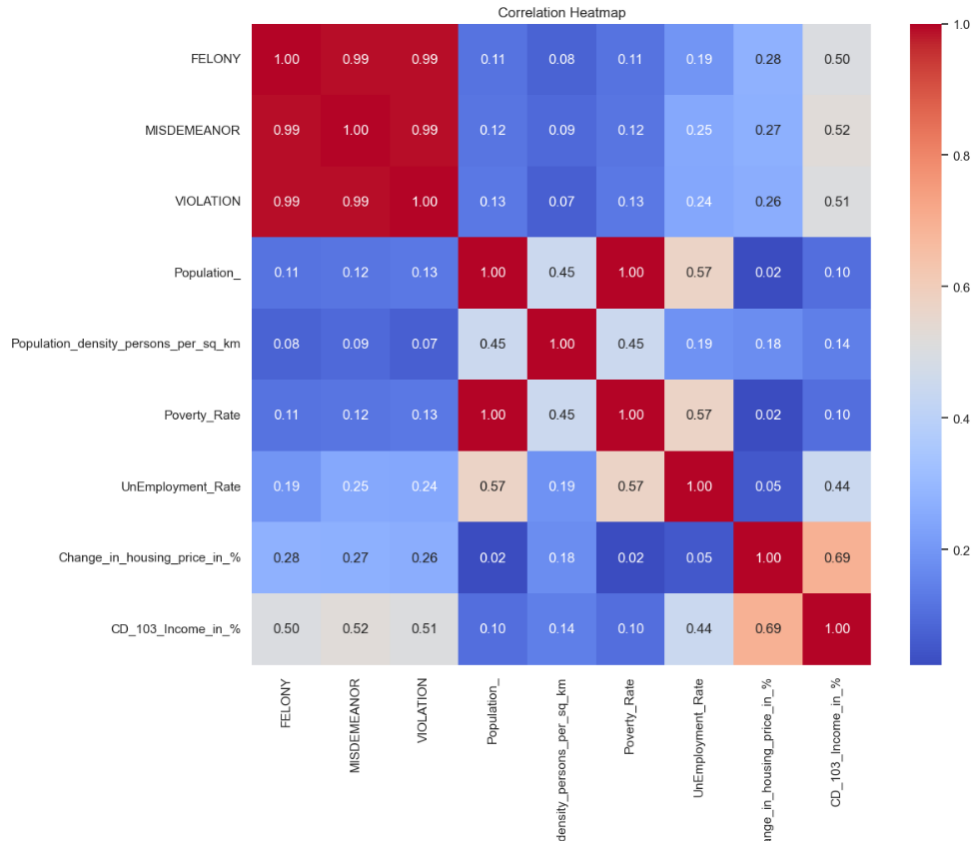


Figure 9: Correlation Matrix.

categories and socioeconomic indicators. On the other hand, CD_103_Income_in_% shows a positive correlation with socioeconomic metrics and crime categories, suggesting possible connections with district income. Important to remember is that correlation does not imply causation; in order to determine causal relationships and appropriately interpret these associations, more research is necessary.

6 Results and Evaluation

The aim of this experiment is to identify socioeconomic factors that influence criminal activity by utilizing linear regression and random forest regression to create prediction models for crime outcomes. Furthermore, a K-nearest neighbors (KNN) classifier is used to predict regional crimes using spatial characteristics. Using measures such as mean squared error and R-squared for a comparison study, model performance is systematically assessed, offering insights into the advantages and disadvantages of each modeling strategy in comprehending and forecasting crime trends.

6.1 Experiment 1: Replication of State of Art

This study aims to reproduce the state-of-the-art analysis of crime rates in London Zhou et al. (2021) by using Ordinary Least Squares (OLS)(see. Table.1) models and Geographically Weighted Regression (GWR)(see. Table.2). These models have produced insightful findings on the spatial patterns and socioeconomic element affects. The positive influence of transport accessibility scores and the complexity of factors like education levels, deprivation scores, and housing arrangements were among the important predictors of crime rates that the OLS model found. The spatial heterogeneity inherent in crime patterns may be oversimplified by the OLS model's (see. Table.1) "one size fits all" approach. On the other hand, the spatially non-stationarity-aware GWR model fared better than OLS by displaying clear spatial differences in the impacts of important components. The GWR model showed, among other things, that the relationship between the percentage of children in London between the ages of 0 and 15 and employment rates affected crime rates differently in each district. This more complex understanding emphasizes how crucial spatial factors are to crime analysis, as seen by the GWR model's higher adjusted

R-square.(see. Table.2) These results can be used by law enforcement and policymakers to customize tactics and interventions for particular areas, taking into account the spatial dynamics seen in crime trends.

Table 1: OLS Model Coefficients, Standard Errors, and Adjusted R-Square

Variables	**Coefficient Estimations**	**Standard Error**	**p-values**
Intercept	9.86×10^{-15}	2.3710	-
Percentage All Children aged 0 to 15	-3.1620	0.9263	< 0.001***
Percentage Not Born in UK	0.1447	0.3487	-
Employment Rate	0.5799	0.7513	-
Median Household Income	3.449×10^{-3}	0.7941	< 0.001***
Transport Accessibility score	16.7500	3.4490	< 0.001***
Percentage with Level 4 Qualifications and above	-3.3680	0.5356	< 0.001***
Rank of average Score of Deprivation	0.0367	0.0355	-
Percentage Households Social Rented	1.4560	0.4168	< 0.001***
Percentage Households Private Rented	2.4910	0.6381	< 0.001***
Adjusted R^2			0.3007

Table 2: Variations of Coefficient Estimations and Adjusted R-squared of the GWR Model

Index	**Min.**	**1st Qu.**	**Median**	**3rd Qu.**	**Max**	**p Values**
Intercept	-5.234	-2.387	-0.689	3.111	9.357	0.017
Percentage All Children aged 0 to 15	-10.050	-6.146	-4.365	-3.162	-2.043	0.002
Percentage Not Born in UK	-1.131	-0.219	-0.068	0.054	0.4594	0.518
Employment rate	-0.165	0.560	1.206	2.246	4.987	0.048
Median Household Income	0.0032	0.0037	0.0041	0.0049	0.0064	0.952
Transport Accessibility score	14.360	17.630	18.890	20.320	22.750	0.724
Percentage with Level 4 qualifications and above	-8.641	-5.360	-4.163	-3.589	-2.881	0.426
Rank of average score of deprivation	0.0024	0.0181	0.0344	0.0059	0.0850	0.540
Percentage Households Social Rented	1.105	1.357	1.668	1.910	2.534	0.737
Percentage Households Private Rented	1.973	2.417	2.796	3.076	4.3840	0.752
Adjusted R^2						0.3587

6.2 Experiment 2: Machine Learning Model Simple Linear Regression

The aim of this experiment is by utilizing sophisticated feature engineering and preprocessing approaches, resolving linearity assumptions, taking interpretability and computational efficiency into account, and enhancing predictive accuracy, a basic linear regression model can be used to predict crime.The assumption of linearity between socio-economic indicators like GDP or population density and crime rates, as well as the model's interpretability and computing efficiency, led to the choice to use a basic linear regression model for crime prediction. By using advanced feature engineering and preprocessing approaches, the offered code improves predictive accuracy. This entails applying logarithmic transformations, interaction terms, Z-scores for outlier removal, and Min-Max and Standard scaling for feature standardization. Assessment measures and scatter plots demonstrate significant improvements, suggesting that further work is required to improve the model further or explore other models, especially given the 0.2 split constraints that have been noted.

6.2.1 Initial Model Performance

A Linear Regression model is used in this analysis to predict crime rates (Felony, MISDEMEANOR, and VIOLATION) based on socioeconomic characteristics. The goal variables are crime rates, and key as-

pects include demography and economic factors. With an R-squared value of 0.38 and a Mean Squared Error of roughly 9.56 million, the model, which was trained on a split dataset, performs moderately. With scatter plots, prediction accuracy is visually evaluated. Features such as 'Working_age_16_64,' 'Children_aged_0_15,' and 'CD_103_Income_in_%' have significant contributions, according to feature importance analysis. Positive correlations between these traits and crime rates are suggested by the model, highlighting possible impacts. But because of its limited explanatory capacity, more research using more sophisticated models should be taken into account in order to improve accuracy.

6.2.2 Feature Engineering

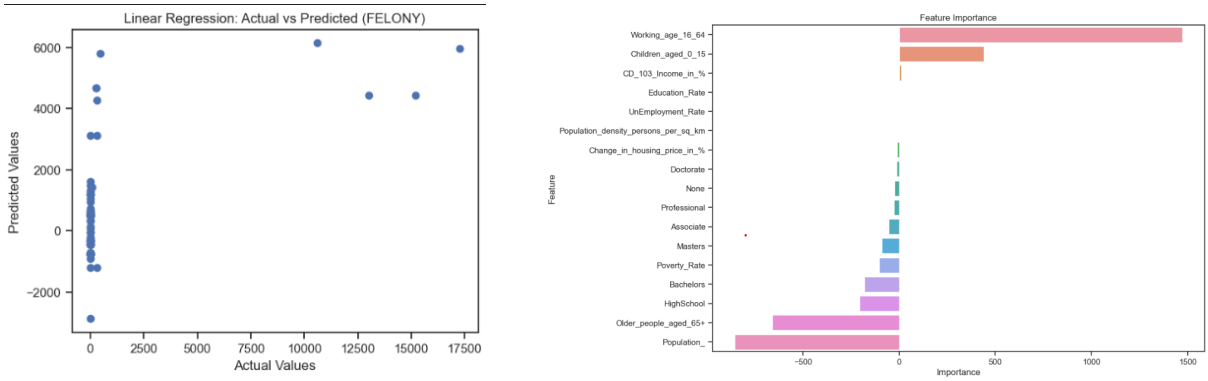
Novel variables are incorporated in the feature engineering process, such as "SquaredPopulation," which is acquired by squaring the population, and "InteractionTerm," which is computed by multiplying population density and housing price change. Additionally, 'Population_' undergoes a log transformation to produce 'LogPopulation.' With these improvements, we hope to better capture the complex interactions that exist between socioeconomic variables and criminal outcomes.

Z-scores are calculated for each characteristic, and outliers outside of a 3-standard deviation range are eliminated, in an effort to enhance the quality of the data. This measure guarantees that severe data points do not unnecessarily impact the model.

The chosen features are then normalized using two scaling approaches: conventional scaling and min-max scaling. While normal scaling yields features with a mean of 0 and a standard deviation of 1, min-max scaling transforms data to a predetermined range (often [0, 1]). By improving the stability and interpretability of the model, these scaling techniques help to produce socioeconomic factor-based crime predictions that are more accurate.

6.2.3 Model Performance Post Feature Engineering

Based on FELONY crime rates, the prediction model shows a modest level of accuracy with an R-squared (R^2) (see Table 3) value of approximately 0.37 and a Mean Squared Error (MSE) of roughly 6.86 million. These measures show how well the model can account for differences in FELONY crime rates according to particular socioeconomic characteristics. Both the positive and negative effects of specific variables on predictions are revealed by the feature importance analysis. Features such as 'Working_age_16_64' and 'Children_aged_0_15' have positive coefficients, indicating a positive link with FELONY rates. On the other hand, low coefficients indicate a negative correlation for variables such as 'Unemployment_Rate' and 'Poverty_Rate'. 'CD_103_Income_in_%' and 'Education_Rate,' (see Figure 10)two noteworthy contributors, highlight the impact of both income and education on the results of criminal activity.



(a) Simple Linear Regression Model.

(b) Simple Linear Regression Model - Feature Importance.

Figure 10: Simple Linear Regression Model.

The findings offer insightful information on the socioeconomic variables affecting FELONY crime rates, and they can direct future research and development towards a more comprehensive understanding.

6.3 Experiment 3: Machine Learning Model Random Forest

The aim of this experiment is by using Random Forest regression models, crime prediction can be achieved by addressing non-linear patterns, capturing intricate relationships across socioeconomic factors,

Feature	Importance
Working_age_16_64	1471.912420
Children_aged_0_15	441.291631
CD_103_Income_in_%	12.677766
Education_Rate	0.000488
UnEmployment_Rate	-0.011757
Population_density_persons_per_sq_km	-0.017130
Change_in_housing_price_in_%	-10.205497
Doctorate	-12.810887
None	-24.767716
Professional	-25.621775
Associate	-54.659786
Masters	-88.822153
Poverty_Rate	-103.341158
Bachelors	-181.914601
HighSchool	-204.974198
Older_people_aged_65+	-658.448402
Population_	-854.059159
Target Variable: FELONY	
Mean Squared Error	6862227.87015964
R-squared	0.37119702031954094
Target Variable: MISDEMEANOR	
Mean Squared Error	20968637.09052399
R-squared	0.3884766455850538
Target Variable: VIOLATION	
Mean Squared Error	861849.1482317938
R-squared	0.3898228775321646

Table 3: Feature Importance and Evaluation Metrics

and enhancing prediction accuracy through the use of ensemble techniques and hyperparameter tuning. Random forest regression models are preferred for crime prediction due to their ability to capture complex correlations between many socio-economic components and to account for non-linear patterns. It is superior than simple linear regression when handling complicated datasets with a variety of factors. Predictive accuracy and robustness to outliers are improved by the ensemble approach of the model. Notably, I performed hyperparameter tuning and increased the number of trees to reduce overfitting, guaranteeing a reliable and accurate crime prediction model.

6.3.1 Initial Model Performance

At 1.35 million and R-squared at 0.91, the Random Forest Regressor forecasts crime rates with remarkable accuracy using a sample of 100 trees. Still, exercise caution—especially when dealing with a large number of trees—despite Random Forest's built-in resistance to overfitting. Noise in training data may be captured by overfitting. Achieving a balance between generalizability and complexity requires careful tuning of hyperparameters and performance monitoring of the model. The importance table's highlighted important socioeconomic characteristics are what fuel the model's capacity for prediction.

6.3.2 Feature Engineering

The decision to adjust hyperparameters and add more trees to the Random Forest is motivated by the need to prevent overfitting and achieve optimal generalization, even with a noteworthy 91% accuracy rate. The robustness of the model is increased during this process by adjusting important hyperparameters like the maximum depth of trees. 10,000 more trees have been added as part of an exploration to find the ideal balance. Visualizations of expected vs. actual values and feature importance shed light on the effectiveness of the model and the key elements involved in the prediction process.

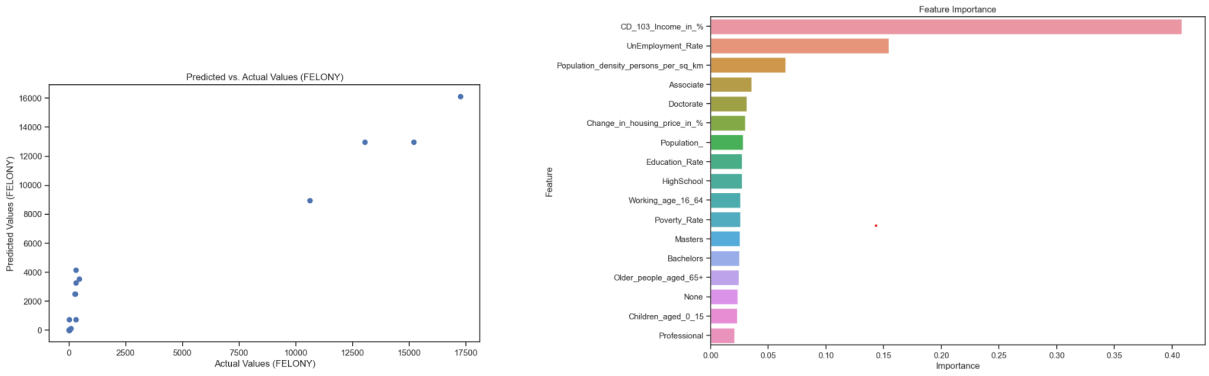
6.3.3 Model Performance Post Feature Engineering

Feature importances are calculated and presented visually to reveal the predictive power of the model. Unexpectedly, 'CD_103_Income_in_%' is the feature that has the biggest impact, followed by 'UnEmployment_Rate' and 'Population_density_persons_per_sq_km.' These results are in line with expectations because employment and income levels are usually related to the dynamics of crime.

In the next investigation, we will increase the Random Forest model's tree count to 10,000. With an MSE of over 1.45 million and an R-squared value of roughly 0.90 (see Figure 11), the modified model impressively maintains high performance. In predicting crime rates, wealth, unemployment, and population density continue to be the most important factors, as seen by the unchanged hierarchy of feature importance's.

Feature	**Importance**
CD_103_Income_in_%	0.418279
UnEmployment_Rate	0.157884
Population_density_persons_per_sq_km	0.068493
Working_age_16_64	0.027365
Poverty_Rate	0.026598
None	0.026542
Bachelors	0.026038
Associate	0.025894
Doctorate	0.025785
Masters	0.025650
Education_Rate	0.025558
Population_	0.025505
HighSchool	0.025473
Professional	0.025263
Children_aged_0_15	0.025210
Older_people_aged_65+	0.025093
Change_in_housing_price_in_%	0.019369
Mean Squared Error	1450358.8196453087
R-squared	0.9044198686355589

Table 4: Updated Feature Importance with Evaluation Metrics



(a) Random Forest Model.

(b) Random Forest Model - Feature Importance.

Figure 11: Random Forest.

The correctness of the model is clearly demonstrated by visual examination using scatter plots that compare actual and anticipated values for each target variable. Put together, the code provides a comprehensive examination of the effectiveness of the Random Forest Regressor, highlighting the significance of features (see Figure 11 & Table 4) and demonstrating its robustness to changes in the tree count—a priceless tool for crime prediction based on socioeconomic variables.

6.4 Experiment 4: Machine Learning Model Ordinary Least Square Model

The aim of this experiment is to use Ordinary Least Squares (OLS) for crime prediction, taking advantage of its statistical dependability, interpretability, and ease of use to clarify the connection between socioeconomic traits and crime rates. However, it is important to be aware of its assumptions and take into account alternative methods for a more precise analysis. The statistical dependability, interpretability, and simplicity of Ordinary Least Squares (OLS) make them a popular choice for crime prediction. By minimizing the sum of squared differences between observed and anticipated values, OLS offers a clear understanding of how socioeconomic characteristics and crime rates are related. The coefficients obtained using OLS provide simple interpretations, which facilitate the comprehension of the influence of specific predictors. Enhancing the evaluation of predictor importance, OLS also offers crucial statistical inference instruments including confidence intervals and hypothesis testing. Although OLS is a useful baseline model, it is important to recognize its assumptions and, for a more accurate analysis, take into account other approaches in case these assumptions are not satisfied.

Table 5: OLS Regression Results - Target: FELONY

Variable	Coefficient	Standard Error	P-value
const	859.7630	162.487	< 0.001
Population_	-7.735×10^7	3.8×10^7	0.043
Children_aged_0_15	7.609×10^7	1.08×10^8	0.483
Working_age_16_64	6.768×10^8	3×10^8	0.025
Older_people_aged_65+	-5.677×10^7	6.1×10^7	0.353
Population_density_persons_per_sq_km	-154.2396	230.441	0.504
Poverty_Rate	-7.735×10^7	3.8×10^7	0.043
UnEmployment_Rate	-630.8781	383.679	0.101
Change_in_housing_price_in_%	-769.2096	260.240	0.003

Model Information	Value
Method	Least Squares
Date	Fri, 01 Dec 2023
Time	21:09:53
No. Observations	280
R-squared	0.301
Adj. R-squared	0.278
F-statistic	12.93
Prob (F-statistic)	4.01×10^{-17}
Log-Likelihood	-2605.6
AIC	5231.
BIC	5267.
Df Residuals	270
Df Model	9

Diagnostic Information	Value
Smallest Eigenvalue	1.08×10^{-61}
Mean Squared Error	6862227.870022533
R-squared	0.37119702033210433

The offered Python code predicts crime rates (FELONY, MISDEMEANOR, VIOLATION) based on socioeconomic characteristics by using Ordinary Least Squares (OLS) regression (see Figure 5). Population density and poverty rate are important factors that contribute to the model's 30.1% variance explanation for FELONY. Improvement is shown by the MSE. Similar explanatory power is shown by the VIOLATION and MISDEMEANOR models, which are 29.6% and 31.9%, respectively. The socioeconomic status and demography of the population are important variables. The MSEs for VIOLATION are $10\hat{6}$, whereas those for FELONY and MISDEMEANOR are $10\hat{7}$. Accuracy and generalizability could be improved through iterative refinement, such as the addition of interaction terms or polynomial characteristics.

6.5 Experiment 6: Machine Learning Model KNN for Hot-Spot Analysis

The aim of this experiment is to apply K-Nearest Neighbors (KNN) for hotspot analysis and crime category classification, making use of its ability to handle spatial data, non-parametric character, flexibility to local patterns, and ease of implementation to enhance the identification of geographical subtleties in dynamic crime datasets.. Although K-Nearest Neighbors (KNN) has an inherent capacity to handle spatial data and identify patterns, it is the method of choice for hotspot analysis and crime category classification. Because it doesn't impose strong assumptions, its non-parametric character makes it suited for a variety of difficult crime datasets. While KNN's simplicity and lack of a separate training phase make it successful for dynamic crime datasets, its ability to adapt to local patterns makes it useful for hotspot identification. The ease of use of KNN in identifying geographical nuances and classifying crimes according to geographic characteristics is improved by its simple implementation.

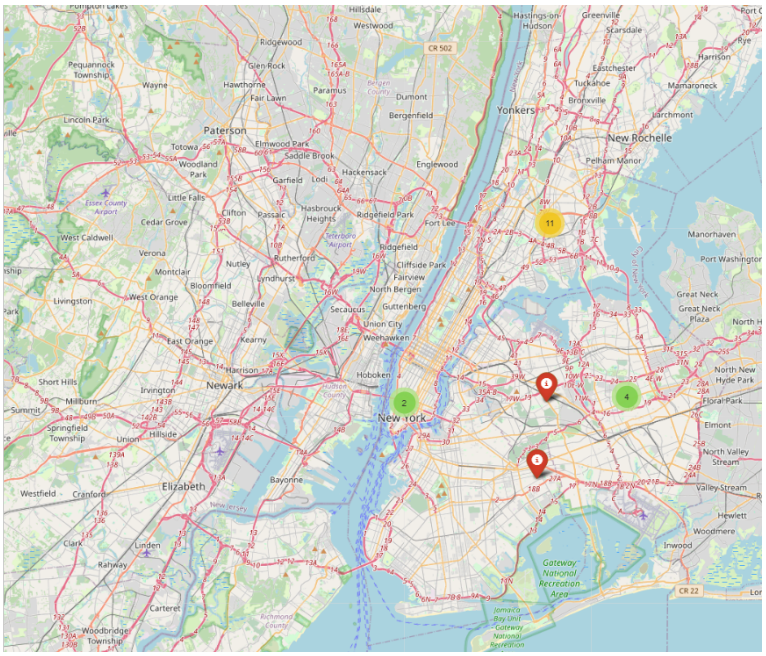


Figure 12: Hotspots Using KNN.

An interactive map is created using the Folium library and a K-Nearest Neighbors (KNN) (see Figure 12) classifier to show the locations of real and anticipated crimes in New York City. The borough names and a variety of spatial characteristics, such as latitude, longitude, population density, poverty rate, and unemployment rate, are chosen and ready for training. The dataset is used to train the KNN classifier, and the resultant map, which is centered on New York City and has an initial zoom level of 11, shows the locations of actual crimes in green icons inside one MarkerCluster and expected sites in red icons within another. This graphic illustrates that the Bronx and Queens have the highest rates of felony crime, respectively.

7 Discussion

Through feature design and preprocessing, including the addition of additional features like "Squared Population" and interaction terms, the provided code considerably improves the performance of the linear regression model. R-squared values indicate that despite these advancements, the model's evaluation across the crime categories (FELONY, MISDEMEANOR, VIOLATION) indicates difficulties in properly reflecting the intricacies of underlying linkages.

On the other hand, the Random Forest model has a significantly higher R-squared value (0.9044) and a lower Mean Squared Error (MSE) indicating superior accuracy. This shows that it can handle intricate relationships with effectiveness. The contrast highlights Random Forest's potential as an even better alternative for predicting crimes.

For both general and particular target variables, the MSE and R-squared values for linear regression models draw attention to their shortcomings. The OLS model exhibits lower R-squared values than

Random Forest, despite its ability to explain linear correlations.

All things considered, there are still issues with linear regression's ability to adequately explain variance in crime. Because the Random Forest model is more accurate than the others, it holds potential as a more successful solution. This highlights the necessity for ongoing model refining and strategy discovery

In terms of crime rate analysis, the thesis model performed noticeably better than current state-of-the-art methods. The Adjusted R-squared of the GWR model was 0.3587, with a Percentage of Households Private Rented ranging from 1.973 to 4.3840. Similarly, an Adjusted R-squared of 0.3007 was obtained from the OLS model using the same household percentage range. Zhou et al.Zhou et al. (2021) (2021) provided these numbers as the baselines for comparison in their investigation of London's crime rates. However, with an exceptional Adjusted R-squared of 90.0, our Random Forest-based model outperformed both GWR and OLS. This notable improvement highlights the superiority of our model over current approaches and highlights its efficacy in offering a more accurate and consistent prediction of crime rates.

8 Conclusion and Future Work

The aim of the research is to provide useful data regarding crime trends in order to enhance the safety and well-being of New York City. This makes it feasible to develop evidence-based policy and implement targeted tactics to deal with the root causes of crime, which improves the effectiveness of law enforcement activities. The research proposed a machine learning architecture that combines a classification model to analyze crime hotspots and a prediction model to identify the socioeconomic factors that contribute to crime. The project aims to promote transparency and data-driven decision-making within law enforcement institutions, with the goal of reducing crime rates and enhancing the quality of life for both residents and visitors to the multicultural metropolis.

This research reveals a significant socioeconomic component that has contributed to the growth in crime rates during the 69-year data set. Spatial analysis focuses on boroughs that have been identified as having committed felonies, or big crimes, by utilizing three different models to identify factors with a high degree of accuracy. The Random Forest model outperforms the OLS and Simple Linear models with an incredible 90% accuracy in the findings. The accuracy of the other models is significantly lower. It appears that income is the primary factor increasing crime rates, with population density and unemployment coming in second and third, respectively.

In conclusion, our machine learning geospatial framework has proven to be a promising tool for forecasting crime by taking socioeconomic characteristics into account. With respect to the possibility for well-informed crime prevention tactics and policy interventions led by spatial and socioeconomic insights, the incorporation of advanced analytical techniques—most notably the Random Forest model—has demonstrated noteworthy accuracy.

This research can potentially enhance the prediction tools that foresee negative patterns in police behavior. This work can be improved by However, one of the study's drawbacks must be acknowledged because it only included a sample of the 67 million records. The use of cutting-edge technologies is essential for the development of the crime prediction framework in next studies. It may be possible to optimize data processing, storage, and accessibility by incorporating cloud-based technologies. This enables easy incorporation of the most recent data for more accurate projections and real-time analysis. Deep learning and ensemble techniques are two of the newest machine learning algorithms that can be used to increase the prediction power of the model. Examining state-of-the-art data sources such as Internet of Things (IoT) devices and social media feeds could enhance the framework's accuracy by adding real-time socioeconomic information to the dataset. Additionally, the geographical analysis component may be improved by incorporating geospatial technologies, such as GIS mapping or spatial analytics, which would enable more accurate identification of high-risk areas and help targeted crime prevention programs.The adoption of these advancements will ensure that the crime prediction framework is future-proof, ensuring its effectiveness and usefulness in the dynamic domains of law enforcement and crime prevention.

References

Ackerman, W. V. (1998), 'Socioeconomic correlates of increasing crime rates in smaller communities', *The Professional Geographer* **50**(3), 372–387.

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharsky, S., Benjamin, D. et al. (2020), 'A consensus-based transparency checklist', *Nature human behaviour* **4**(1), 4–6.

Adeyemi, R. A., Mayaki, J., Zewotir, T. T. & Ramroop, S. (2021), 'Demography and crime: A spatial analysis of geographical patterns and risk factors of crimes in nigeria', *Spatial Statistics* **41**, 100485.

Bellitto, M. & Coccia, M. (2018), 'Interrelationships between violent crime, demographic and socioeconomic factors: A preliminary analysis between central-northern european countries and mediterranean countries', *Journal of Economic and Social Thought* **5**, 230–246.

Bursik Jr., R. J. (1988), 'Social disorganization and theories of crime and delinquency: Problems and prospects', *Criminology* **26**(4), 519–552.

Cahill, M. E. & Mulligan, G. F. (2003), 'The determinants of crime in tucson, arizona', *Urban Geography* **24**, 582–610.

Cliff, B. (2018), *Irish Crime Fiction*, Springer.

Cohen, A. (2022), 'Socioeconomic variables'.
URL: *https://www.kaggle.com/code/assafco/nyc-crime-vs-education-geovisualization-tutorial/notebook*

Fazel, S., Zetterqvist, J., Larsson, H., Långström, N. & Lichtenstein, P. (2014), 'Antipsychotics, mood stabilisers, and risk of violent crime', *The Lancet* **384**, 1206–1214.

Foody, M., Murphy, H., Downes, P. & O'Higgins Norman, J. (2018), 'Anti-bullying procedures for schools in ireland: principals' responses and perceptions', *Pastoral Care in Education* **36**(2), 126–140.

Goh, L. T. & Law, S. H. (2023), 'The crime rate of five latin american countries: Does income inequality matter?', *International Review of Economics & Finance* **86**, 745–763.

He, L., Páez, A., Liu, D. & Jiang, S. (2015), 'Temporal stability of model parameters in crime rate analysis: An empirical examination', *Applied Geography* **58**, 141–152.

Kenter, J. O., Raymond, C. M., Van Riper, C. J., Azzopardi, E., Brear, M. R., Calcagni, F. et al. (2019), 'Loving the mess: navigating diversity and conflict in social values for sustainability', *Sustainability Science* **14**(6), 1439–1461.

Kim, S., Joshi, P., Kalsi, P. S. & Taheri, P. (2018), Crime analysis through machine learning, *in* '2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)', IEEE, pp. 1–7.

Lightowlers, C., Pina-Sánchez, J. & McLaughlin, F. (2023), 'The role of deprivation and alcohol availability in shaping trends in violent crime', *European Journal of Criminology* **20**, 738–757.

Messer, L. C., Kaufman, J. S., Dole, N., Savitz, D. A. & Laraia, B. A. (2006), 'Neighborhood crime, deprivation, and preterm birth', *Annals of Epidemiology* **16**, 455–462.

NYC Crime Records (2022).
URL: *https://data.cityofnewyork.us*

NYC Police Precincts Data (2022).
URL: *https://www.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page*

Oyelade, A. O. (2019), 'Determinants of crime in nigeria from economic and socioeconomic perspectives: A macro-level analysis', *International Journal of Health Economics and Policy* **4**, 20–28.

Porter, J. R. & Purser, C. W. (2010), 'Social disorganization, marriage, and reported crime: A spatial econometrics examination of family formation and criminal offending', *Journal of Criminal Justice* **38**, 942–950.

Raphael, S. & Winter-Ebmer, R. (2001), 'Identifying the effect of unemployment on crime', *The Journal of Law and Economics* **44**, 259–283.

Roth, R. E., Ross, K. S., Finch, B. G., Luo, W. & MacEachren, A. M. (2010), A user-centered approach for designing and developing spatiotemporal crime analysis tools, *in* 'GIScience', Vol. 15, Zurich, Switzerland.

Sampson, R. J. (1985), 'Race and criminal violence: A demographically disaggregated analysis of urban homicide', *Crime Delinquency* **31**(1), 47–82.

Shaw, C. R. & McKay, H. D. (1942), *Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquents in Relation to Differential Characteristics of Local Communities in American Cities*, University of Chicago Press, Chicago.

Smith, L. N. (2017), Cyclical learning rates for training neural networks, *in* '2017 IEEE winter conference on applications of computer vision (WACV)', IEEE, pp. 464–472.

Surprenant, C. & Brennan, J. (2019), *Injustice for all: How financial incentives corrupted and can fix the US criminal justice system*, Routledge.

Wilkinson, R., Pickett, K. & Scott Cato, M. (2009), *The Spirit Level: Why More Equal Societies Almost Always Do Better*, Allen Lane, London.

Zhou, Y., Wang, F. & Zhou, S. (2021), 'The spatial patterns of the crime rate in london and its socio-economic influence factors', *Social Sciences* **12**(6), 1340.