Analysis of Crime Rate in New York City

Ahalya Sugumar Stony Brook University 112684060

Nagashree Angadi Chandrasekhar Stony Brook University 112683867

Udit Gupta Stony Brook University 112715403

1 INTRODUCTION

The Sustainable Development Goal(SDG) chosen for this project is Peace, Justice, and Strong Institutions. Even the world's greatest democracies face major challenges in the crimes that threaten the foundation of peaceful societies such as homicides, human trafficking, and other organized crimes. The Police department is a vital component for achieving these SDG indicators, as they directly contribute to the promotion of peace and justice.

The motivation behind this project is to aid the Police Department to strengthen its forces in specific areas of their jurisdiction to tackle a specific problem. There are vital temporal and spatial trends in the occurrence of crime, that when analyzed can be deployed as a powerful tool in preventing crime. The project also looks at the effect of socio-economic conditions such as poverty and unemployment rates, on the crime rate to ascertain whether the problem lies in the root of the society.

With the advent of technology in all walks of life, employing big data tools to perform crime analysis has become an essential tool for law enforcement's efforts to enhance public safety, identify emerging trends, allocate resources, and plan crime-prevention strategies.

2 BACKGROUND

The New York Police Department (NYPD) has taken up an initiative for data-driven crime prevention [2]. The analysis of crime data is important for enforcement agencies to deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects. In [1] a multivariate conditional autoregressive model was used to find the spatial dependence between sites and crime types. In [3] a Poisson regression model to compare differences in crime and arrest counts before and after census block groups.

[4] The stop and frisk data shows that police stop only individuals of a certain race and ethnic minority groups more often than whites relative to their proportions in the population. This undermines the target set by SGD Indicator. The SDG indicator states to promote the rule of law at the national and international levels and ensure equal access to justice for all.

3 DATA

We used multiple datasets for our analysis and below is the list of them along with their source,

1. NYPD Arrests(2006-2019) - Data.gov
2. NYPD Complaints(2006 - 2019) - Data.gov
3. Demographics(Sub-Borough) - CoreData
4. GeoJSON (Sub-Borough) - NYC Open Data

The historic arrests and complaints data formed the primary source for our analysis while the demographics data and GeoJSON were used as a supplemental data source for socio-economic analysis and spatial analysis respectively. The primary data sources are close to 11.2 GB in size and contain more than 12 million data points with 35 distinct features.

The information in this dataset includes time and date of crime, type of crime, the ethnicity of suspects and victims, and spatial information about the arrests and complaints. The demographic data includes information about various socio-economic factors like poverty rate, unemployment rate, racial diversity, and income diversity at borough and sub-borough levels. Lastly, the GeoJSON datasets include coordinate information about boroughs and sub-boroughs in NYC and helped us to perform spatial analysis.

4 DATA PREPROSSESSING

Spark was the primary tool employed for performing all the data wrangling operations like missing value imputation, data standardization, grouping and aggregations, and pivot table generation. We also used Pandas for a few operations when we had a summarized output, and for preprocessing the files lesser than 128MB such as demographics data.

The sub-borough information was imputed for crime data using the location of the crime and sub-borough GeoJSON using GeoPySprak. Later, the crime data was aggregated on the borough and sub-borough levels to perform the regression analysis and hypothesis testing.

5 METHODS

In order to find insights from our data, we used a wide variety of methods from simple EDA techniques such as data visualization to Multivariate Linear Regression and Hypothesis Testing.

1. **Exploratory Data Analysis**

   Different attributes from the dataset were explored in attempts to identify the most efficient visualization for them. The variation in data was analyzed under the following broad spectrum's,

   (a) Time-Series analysis
   (b) Spatial Analysis
   (c) Demographic Analysis
   (d) Social/Cultural Analysis

   We performed an extensive EDA with drill-down charts covering all the major aspects of the data for each crime type.

   Further, we used Correspondence Analysis to understand the relationships between different crime types vs the location of their occurrence at precinct, borough, and premises level. We also performed a similar analysis with crime type vs the racial group of the victims involved in those crimes.

2. **Stratified Sampling**

   The data has more than 11 million rows which made it difficult to analyze the data on maps using standard python visualization tools. Stratified sampling was performed to overcome this issue. In this method of sampling, we partitioned data into subgroups and picked the samples from the subgroup. This was done to ensure that each subgroup within the data receives proper representation within the sample.

3. **Linear Regression**

   Linear regression was used to estimate the relationship between the crime count and various demographic factors such as poverty rate, unemployment rate, racial diversity, and income diversity. The crime data was aggregated at two spatial levels (i.e. borough and sub-borough) for this purpose. The regression was performed on the aggregated data with and without the population as the control variable. The population was chosen as a control variable to remove its effects from the regression equation.

4. **Hypothesis Testing**

   Hypothesis testing was performed to check the significance of the relationship between crime count and demographic factors at various spatial levels. The significance threshold set for the test is 0.05. The null hypotheses set for the test are as follows,

   (a) The demographic factors do not affect the crime count at the borough level.
   (b) The demographic factors do not affect the crime count controlled by the population at the borough level.
   (c) The demographic factors do not affect the crime count at the sub-borough level.
   (d) The demographic factors do not affect the crime count controlled by the population at the sub-borough level.

   Following linear regression, if the p-value obtained is less than 0.05 (chosen significance threshold), then the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected.

6 RESULTS

The goal of our analysis is to empower the authorities with actionable plans specific to certain crime types in the city. As a result, we propose a 4 step plan for the authorities to tackle any selected crime type in the city.

1. **Temporal Analysis**

In this task, we looked at the time/date aspect of the arrests/complaints in order to mine underlying hidden patterns for different crime types. We looked at Yearly level, Seasonal level, Day of Month level, and Hour of Day level to understand when different crimes happen and if there is any pattern in those occurrences. We employed Spark for data aggregation and Matplotlib for generating time series plots.The visualization in Figure 1,Figure 2 and Figure 3 shows that sex crimes in the city have risen over the past 6 years and the visuals show that the crime happens above the average during 2nd week of the month and during the night hours.

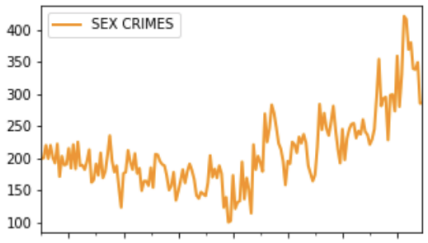

Figure 1: Sex Crimes Over Years

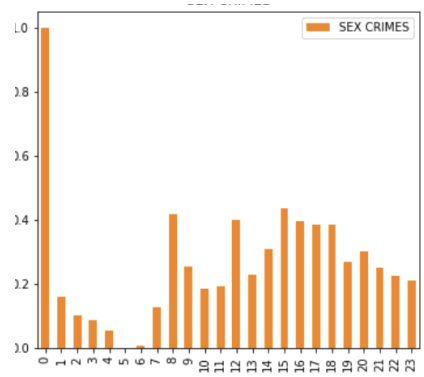

Figure 2: Sex Crimes vs Day of Month

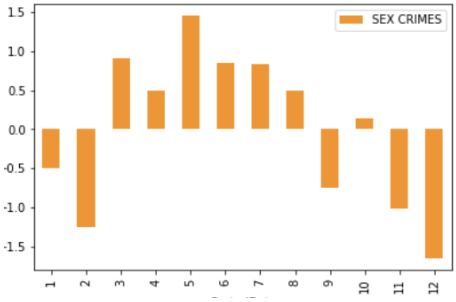

Figure 3: Sex Crimes vs Hour of Day

2. **Spatial Analysis**

In this study, we looked at the location of arrests/complaints at borough, precinct, and premises level to understand the locations in the high with high crime rate for each individual crime type. We used Folium and Seaborn libraries to generate point level heat maps and density plots respectively.

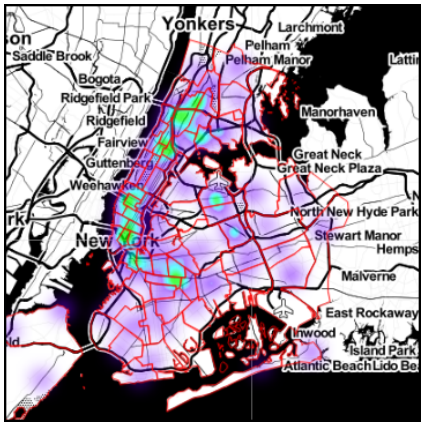

Figure 4: Heat Map of Sex Crimes

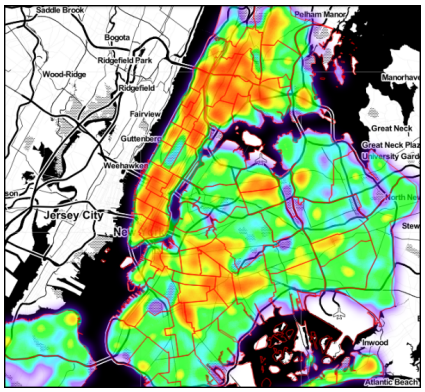

Figure 5: Heat Map of Assault in NYC

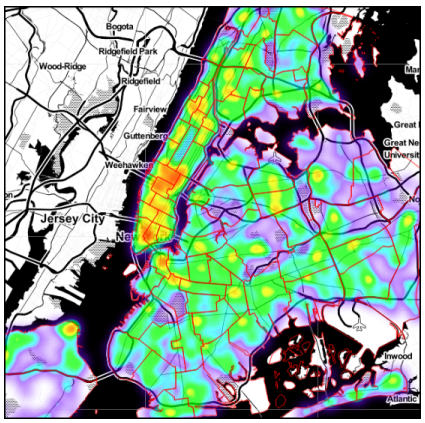

Figure 7: Heat Map of Grand Larceny

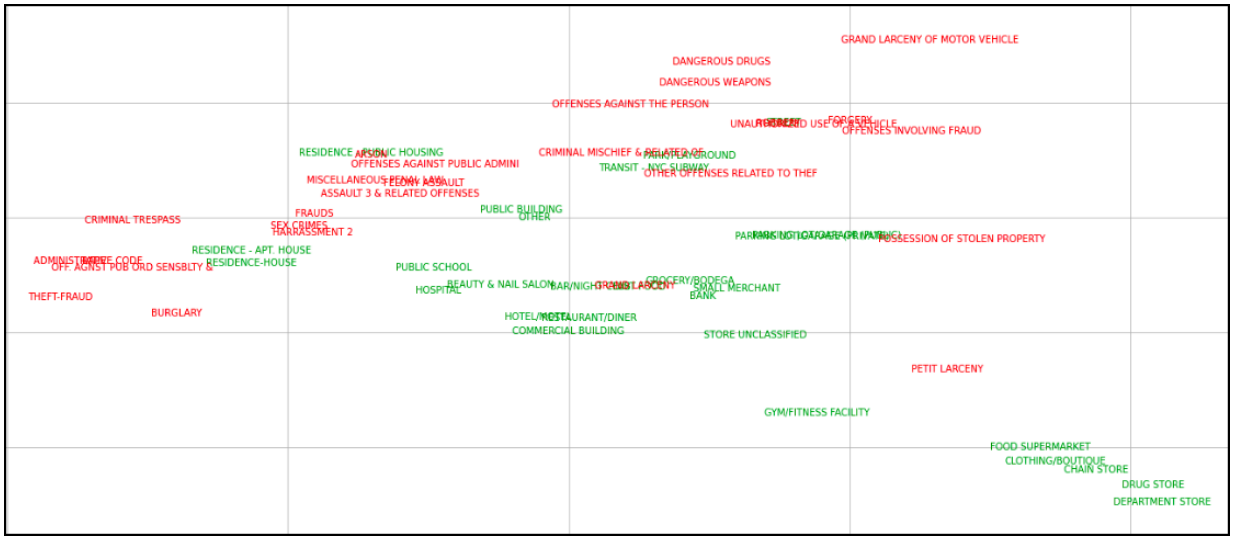

Figure 6: Correspondence Analysis on Crime Type vs Premises

3. **Correspondence Analysis**

   In order to find insights at the premises level in Figure 6, we used the method of Correspondence Analysis which gives us a linear transformation of two different attributes on a new coordinate scale for comparison.

4. **Social and Demographic Analysis** Going a step further, we decided to analyze the ethnicity information of the suspects and victims to understand their relationship in terms of crime type and locations in the city in Figure 8. The correspondence analysis in Figure 9 shows what racial groups are most susceptible to certain crime types and vice versa.

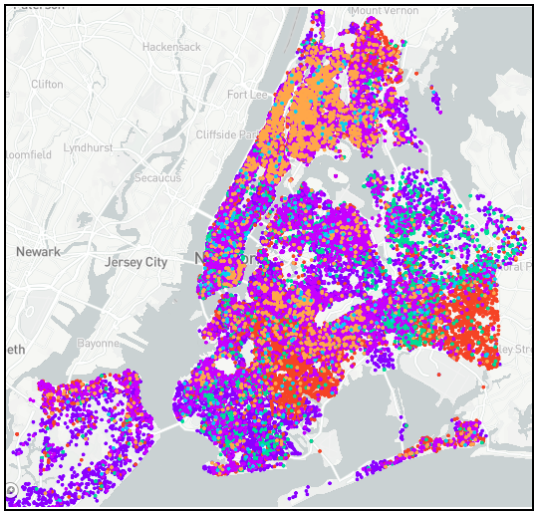

Figure 8: Location of victims and their ethnicity

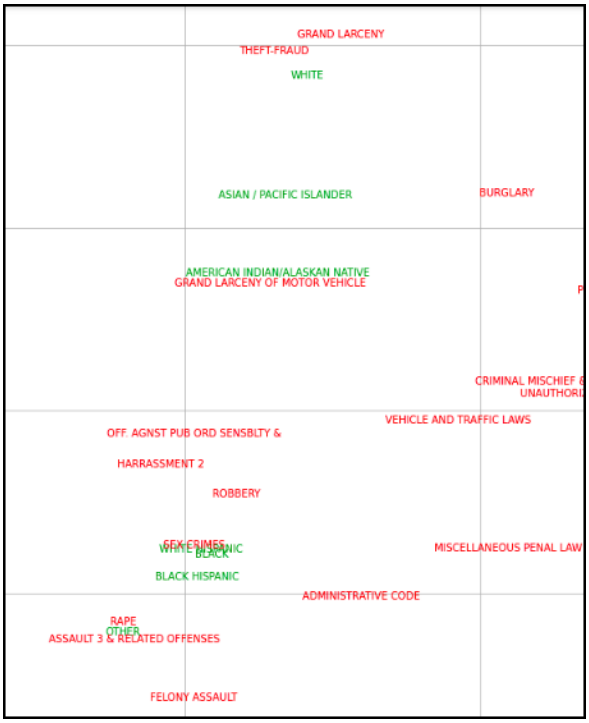

Figure 9: Correspondence Analysis of victims vs crime type

5. **Hypothesis Testing**

   As stated in the methods above, hypothesis Testing was performed to ascertain whether various socio-economic factors in a region affect the crime rate. The analysis was performed at Borough and Sub-borough levels with and without control by the population in

Level	Controlled/Not Controlled	Poverty Rate	Unemployment Rate	Income Diversity	Racial Diversity
Borough	No Control Variable	0.065	**0.036**	**0.008**	0.054
	Controlled By Population	0.18	**0.04**	**0.0002**	**0.01**
Sub Borough	No Control Variable	**0.0057**	**0.00045**	**0.0125**	**0.0219**
	Controlled By Population	**6.75e-07**	**9.776e-07**	**0.002**	**0.007**

Figure 10: The p-values for crime count vs demographic factors

the region. On setting the significance threshold at 0.05, we observe that socio-economic factors play a significant role in affecting the crime rate of a region (highlighted cells showcase the significant factors under each hypothesis test performed). At the Borough level, we see that the Unemployment Rate and Income Diversity are significant counterparts that affect the crime rate in the borough. We also observe that on controlling by population variable, our regression results showcase that Racial Diversity also contributes to the crime rate in the Borough. On delving deeper to the Suborough level, the regression results showcase that poverty rate, Unemployment rate, Income, and racial diversity are all important components at play while evaluating the crime rate of a region at the Sub-borough level.

7 CONCLUSION

Crime Rate is influenced by a large number of development indicators such as poverty rate, racial diversity, gender diversity, and unemployment rate. The temporal, spatial, and socio-economic trends/factors that seem to influence the crime rate in a region over time. The police department could employ these decisions to deploy there forces effectively based on the intensity of crime at that particular jurisdiction thereby achieve lower crime rates in their jurisdiction. The temporal trend can help the institution and law authorities in making proactive decision making.

The reduction in crime rate directly improves peace and justice in society. Thereby, reducing the crime rate in the city promotes peaceful and inclusive societies for sustainable development, provides access to justice for all, and aids in building effective, accountable, actionable, and inclusive institutions at all levels.

References

[1] Crime Risk Maps: A Multivariate Spatial Analysis of Crime Data

[2] New York City Police Department Burrough and Precinct level Statistics

[3] The Effects of Local Police Surges on Crime and Arrests in New York City

[4] An Analysis of the New York City Police Department's Stop-and-Frisk Policy in the Context of Claims of Racial Bias

[5] Demographic data - maintained by New York University

[6] Crime data - maintained by NYPD