Responsible Data Analytics
SEN 163B

Report Part I -
Exploring the Relationship between Criminal Complaints and Green cover across Police Precincts of New York City

---

Group 4

Authors:

Lala Sayyida Millati Nadhira	5844266
Kaninik Baradi	5216664
Rezzy Yolanda Wulandhari	4779487
Kelvin Engee	4664043
Philippe Almeida Mirault	5898803

March 6th, 2023

Contents

1. Problem Formulation

1.1. Problem Summary

The relationship between green spaces and crime rates is debated, with some studies suggesting that greenery can have a calming effect and attract foot traffic to deter criminal activity (Kuo & Sullivan, 2001). However, the relationship is not always clear-cut and may be influenced by various factors. In certain contexts, green spaces may even increase crime rates by providing cover for criminal activity or attracting individuals more likely to engage in crime. The distribution of green spaces and crime may not be equal across different socio-economic groups, which may lead to sensitive social and environmental justice issues. For instance, increasing green spaces in certain areas distribution may exacerbate existing inequalities.

Understanding the factors that may have contributed to the crime would be helpful in predicting the likelihood of future crimes occurring. Giuliani's 1993 New York mayoral campaign was focused on quality-of-life issues and his administration implemented the "Greenstreets" program to convert underutilized street spaces into green areas. Predictive analytics could help target such programs and anticipate their co-impacts on the residents.

Research has attempted to establish causal relations between factors, such as socio-economic and demographic as the determinants of crime rates. One of them is the usage of multifaceted robust data-driven framework for predicting county level annual felony and misdemeanor rates for the state of New York from a holistic perspective was used, with advanced statistical learning algorithms on a multidimensional publicly available data set (Ganguly & Mukherjee, 2021).

The findings show that there is a correlation between felony and misdemeanor crimes with socio-economic and demographic conditions. Areas with higher percentages of people living in poverty experience more felony and misdemeanor crimes than other areas (Ganguly & Mukherjee, 2021). Regarding the demographic component, there is a connection between the rise in crime and particular demographic predictors of vulnerability.

From these explanations, the scope of this project is reduced to the following questions:

Main Question: What is the nature and extent of the relationship between greenery and complaints rates?
Sub-Questions:

- Is there a direct relationship between green spaces and complaints rates?
- If not, what are other underlying factors that explain the correlation?

1.2. Dataset Overview

The dataset used in this project is NYPD Complaint Data Historic, taken from the New York City (NYC) Open Data website. The main dataset consists of 35 rows with 7.83 million rows, where each row represents a complaint. The dataset contains crimes in different categories (felonies, misdemeanours, or violations) that were reported to the New York City Police Department (NYPD) between 2006 and 2019. However, the number of actual crimes that happened is unknown. Thus, in this project, the crimes counted will only take the number of reported complaints into account.

Several variables are analysed to select the best case for this project, including the year of crimes, patrol borough, type of crimes, and demographics of the suspects and victims. To support the analysis, we chose the following additional datasets – from the NYC Open data website:

- Police Precincts, Map - GeoJSON
- [1995, 2005, 2015] Street Tree Census Data, Data file with coordinates

2. Descriptive Analytics

To establish that the selected data sets are fit-for-purpose, an overview is conducted. The selected data must be accurate, representative, and reliable. The following analysis ensures that the temporal and geographical distribution of the available data are as expected that malformed or mistyped values are either corrected or removed. The analysis also supports the identification of anomalies, outliers and errors that impact the hypothesis or create confounding variables.

2.1. Variables Analysis of Dataset

2.1.1. Timeline

Analysing the total number of complaints annually supports 2 goals. First to ensure that there are no years which have a skewed number of complaints recorded, and second to evaluate the change in complaints year on year. Where the date of the reported event is identified as a range, we use the initial time for analysis. There is a clear downward trend in the total number of complaints annually – with the lowest number in 2020. This dip can likely be explained by CoVID-19 linked lockdowns. (Refer: Figure 1)

2.1.2. Precincts

The primary dataset identified the police precinct inside which each complaint was recorded. In total, the dataset has 77 precincts – consistent with the current official numbers. No precincts have been added or removed in the data period. The precincts were also mapped to ensure that the data is complete spatially. (Refer: Figure 2)

2.1.3. Type of Crimes

The types of crimes are divided into categories represented by the classification code of offense (PD_CD). This code has 432 different codes, where each code represents a specific type of crime. The PD_CD column was chosen instead of the KY_CD column because the information only provides data regarding PD Code correlation with the type of crimes. From the analysis, crimes with PD Code 101 (assault crime) have the highest number of occurrences, followed by crime with PD Code 638 (harassment) on the second. However, the classification of the crime type is unclear because of different crime codes used by the dataset and the New York State Law website.

2.2. Exploratory Data Analysis

2.2.1. Complaints per Precinct

The average number of complaints recorded in each precinct annually ranges from 2000 to 15000 complaints. The proportion of complaints recorded in each borough remains constant over the period of the data set. (Refer: Figure 3). The number of trees in each precinct was also computed using the available spatial data. Using this, we can also infer the average density of greenery. (Refer: Figure 4)

2.2.2. Complaints by victim demographic

Looking at the Complaints per victim demographic of the years, the order of the distribution described in 2.1 remains the same. The number of victims in most of the demographic groups seem to follow the same decreasing trend over the years, especially in 2020, which could be related to COVID, as a small increase in 2021 follows it. Because of the small size of some demographic groups, they can't be seen in the graphs. These are the Unknown genders, Other race and American Indian/Alaskan Native race. The number of victims in some demographic groups don't seem to decrease, in contrast to the other demographic groups. The number of victims in these demographic groups seems to be stable. These demographic groups are Black Hispanic, Asian/Pacific Islander and Businesses. (Refer: Figure 9)

2.3. Dataset Limitation

For the research question – the dataset has a few limitations stemming from the missing data points, as well as from some decisions taken to limit the privacy risks. The data is only a record of complaints logged—and may not correlate consistently to actual crimes committed. The literature reviewed shows that the correlation is sufficient for the proposed analysis but limits the predictive power of the inferences.

The classification codes for the complaints also leave room for interpretation. The classification codes do not perfectly match the information available at the New York State Law website. The dataset has a PD_CD column that represents the the complaint code, while on the New York State Law website, the code uses five digits to classify its complaint levels. We may select a subset of the data over which the classifications may be aligned, if required.

The New York Street Tree Census data is available with a decadal frequency for the years 1995, 2005, and 2015. The rate of change of the trees in the data set needs to be analysed – to infer whether we can consider the number of trees in each precinct to be constant during the intervening time span.

Using this data set also limits the definition of greenery to trees covered by the census, and neglects shrubs, potted plants, vines, and miscellaneous lawns and public green spaces.

The robustness of the current analysis with demographic information from the New York Census. A preliminary check has been conducted for racial demographics of the victims using census data.

2.4. Data visualization principles

Related to the principles in paper by D'Ignazio and Klein (D'Ignazio et al., 2016), specifically "Rethink Binaries", the categories for race for both suspects and victims have been taken for granted. It is unknown how mixed races have been categorized, as they do not fit the currently stated categories. This leads to the over- or underrepresentation of certain races, depending on how they have been categorized, which again leads to the mistreatment of certain population groups. If the identification of these victims and suspects is known and their demographics are stored in a database, the dataset can be corrected for wrong categorization. If not, the only solution is communicating the limits of these categories.

The classification of crime complaints also doesn't fit the current categories. If multiple offenses have been committed, only the most serious offense is reported (NYPD Complaint Data Historic | NYC Open Data, 2022), leading to an underrepresentation of the less serious offenses. If all committed crimes are known, a problem still remains in correcting the dataset. If, in a row, all the data is copied except for the committed crime category, then the other features will be overrepresented. Leaving out all the remaining

data, however, leads to very little information about the crime, as it isn't connected to other features. Adding features, mentioning the number of crimes in a complaint or mentioning the other crimes, could help solve this problem.

3. Diagnostics Analytics

3.1. Missing Data Analysis

The largest sections of missing data are linked to Victim and Suspect demographics. In (x number) of cases, the missing victim data is because the victim is a business/organization or the victim is identified as 'the people of the State of New York'.

"Victim Race" however contains 2.553.116 unknowns, while "Suspect Race" contains 1.140.797 unknowns. "Victim Age Group" also contains 1.634.196 unknown data points, while "Suspect Age Group" contains 4.863.526 unknown data points. The problem with these unknown data points is that it leads to misleading results. Missing information about the Suspect seems to be linked to the information not being known at the time of the complaint being recorded.

The malformed values create more unknowns in the age columns. About 200,000 values are negative, or in the implausibly large. While some of these may be possible to infer as birth years, correcting these values would introduce a lack of repeatability and hence their data points will be considered "unknown". This, however, adds to another limitation of the dataset. Most unknown age values are from before the year 2016 – showing that this is potentially caused by a lack of standardisation in the data entry. (Refer: Figure 5)

3.2. Correlation Analysis

There is a visible downward trend in the number of complaints annually. Pearson's correlation was used to test this relationship. A coefficient of –0.955 and a p-value of 8.87e-09 were obtained. There is enough statistical evidence to prove that the number of complaints depends on when the year that it occurred, resulting in a decrease of the number of complaints along each year.

Relationship between reported complaints and number of trees in each precinct:

A series of correlation tests were done to examine the association between the number of complaints and number of trees in each precinct. An additional examination was conducted of the amount of grievances and trees per sq.km of the region of each precinct. The investigation was conducted for the sum of grievances throughout all years and only those from 2015, when the data for the quantity of trees would be up to date.

Both the direct, and area-normalised analysis show a positive correlation between the presence of trees, and incidences of complaints. (Refer: Figure 6, Figure 7)

ANALYSIS	PEARSON'S R	P-VALUE
DIRECT	0.20	$p < 0.08$
AREA NORMALISED	0.46	$p < 2.0\ e\text{-}5$

Factoring in the lack of demographic data in NYC, only the grievances in 2010 will be utilized for this examination. By conducting a correlation test between complaints and victim demographic, it can be seen

that there is a strong correlation of 1, as all of the data objects in the dataset are related to a complaint made by a victim. Accordingly, an assessment will be conducted to determine if the demographic groups in the dataset are proportionate to the demographic distribution of NYC. Not all categories will be examined, as the bases data for those categories are unavailable. Categories which have been removed from this test are the Unknown Sex, “The people of the state New York”, the Unknown Age Group and the Other Race. Since the demographics are categorical, a chi- square test will be done. For all categories in victim age, victim sex and victim race, the chi-square test produces large statistics and near 0 p values, indicating that there is a significant deviation from the expected distribution of the victim demographics and the actual distribution.

3.3. Risks and Bias

As mentioned before, the dataset contains a lot of “unknown” and “NaN”. This missing data is non-ignorable, as results can be misleading, if ignored (Bock, 2020). If we take demographics, for example, misleading results can lead to the mistreatment of certain population groups. Without mitigation strategies, people would make false conclusions. The representation of complaint by population groups can be mentioned, but the report must explicitly mention that this representation is misleading (Bock, 2020).

Historical bias could also be an issue. The dataset contains data from the 2008 financial crisis, the Covid epidemic and the BLM movement. During the 2008 financial crisis, an increase in robberies can be seen (Reuters, 2009). During the Covid epidemic, Asians were victims of hate crime (Health Affairs, 2022). During the BLM movement, many protesters were arrested (Forbes, 2021). Also, as some precincts are policed more than other precints, these more policed precincts naturally have more complaints compared to other precincts (Washington Post, 2016). Though the data has been collected accurately, some of the data has been collected from years where historic events took place, which makes it unrepresentable for the normal world (Suresh et al., 2021). Other data comes from places which are policed more, which leads to an overrepresentation of crimes compared to other precincts (Suresh et al., 2021). Data from years where historic events took place could be removed from the dataset and be used to predict crime during historic events. For the policed precincts, a solution would be to add a feature, mentioning the amount of police per precinct.

4. Conclusion

This project's research question is to determine the relationship between greeneries and complaint rates in New York City. It was discovered that there is a positive relationship between the existence of trees and the occurrence of complaints by analyzing several datasets such as NYPD Complaint Data History, Police Precincts Map, and Street Tree Census Data. There is an overrepresentation of black victims in the complaints; however, in order to answer the sub-question about the relationship between other factors and complaints, a test to see if the distribution of demographic groups in the dataset corresponds to the actual distribution of demographic groups in NYC is required. This allows for a more meaningful analysis of the relationship between other underlying factors and complaints.

5. References

Bock, T. (2020, 7 december). Whate are the Different Types of Missing Data? -Displayr. Displayr. https://www.displayr.com/different-types-of-missing-data/

D'Ignazio, C. & Klein, L.F. (2016). Feminist Data Visualization.

Forbes. (2021, 9 maart). Hundres Arrested At Black Lives Matter Protests PLan To Sue NYPD Over Violence: City Official. https://www.forbes.com/sites/jemimamcevoy/2021/03/09/hundreds-arrested-at-black-lives-matter-protests-plan-to-sue-nypd-over-violence-city-official/?sh=36a7d518dd08

Ganguly, Prasangsha & Mukherjee, Sayanti. (2021). A multifaceted risk assessment approach using statistical learning to evaluate socio-environmental factors associated with regional felony and misdemeanor rates. Physica A: Statistical Mechanics and its Applications. 574. 125984. https://doi.org/10.1016/j.physa.2021.125984.

Health Affairs. (2022). COVID-19 Has Driven Racism And Violence Against Asian Americans: Perspectives From 12 National Polls. Health Affairs. https://doi.org/10.1377/forefront.20220411.655787

Kuo, Ming & Sullivan, William. (2001). Environment and Crime in the Inner City: Does Vegetation Reduce Crime?. Environment and Behavior - ENVIRON BEHAV. 33. 343-367. HYPERLINK "https://doi.org/10.1177/00139160121973025"https://doi.org/10.1177/00139160121973025.

NYPD Complaint Data Historic | NYC Open Data. (2022, 9 juni). https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Reuters. (2009, 27 januari). U.S. recession fuels crime rise, police chiefs say. U.S. https://www.reuters.com/article/us-usa-economy-crime-idUSTRE50Q6FR20090127

Suresh, H. & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305

University of Washington. (2010, June 28). *Crime & Public Safety.* https://depts.washington.edu/hhwb/Thm_Crime.html. Green Cities: Good Health.

Washington Post. (2016, 25 juli). Does more policing lead to less crime – or just more racial resentment? https://www.washingtonpost.com/news/monkey-cage/wp/2016/07/25/does-more-policing-lead-to-less-crime-or-just-more-racial-resentment/

Wolch, Jennifer & Byrne, Jason & Newell, Joshua. (2014). Urban green space, public health, and environmental justice: The challenge of making cities `just green enough'. Landscape Urban Plann. 125. https://doi.org/10.1016/j.landurbplan.2014.01.017.

6. Appendix

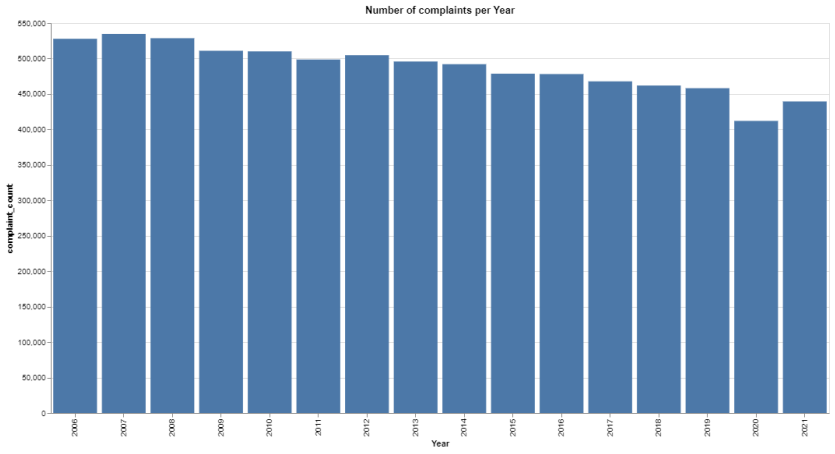

Figure 1 Number of Complaints per Year

In the graph above it is shown that, even though the number of complaints doesn't change that much from year to year, there is a clear tendency of decreasing over the years. (Refer: Figure 1)

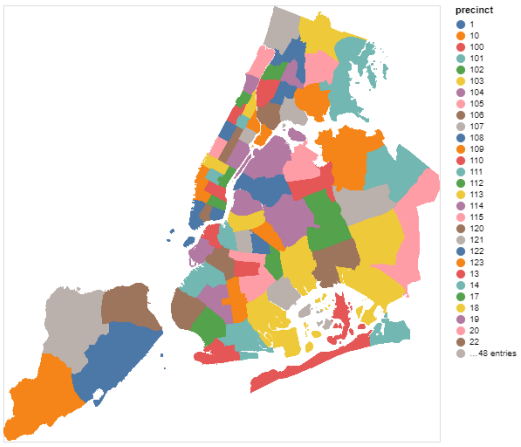

Figure 2 Precinct Map

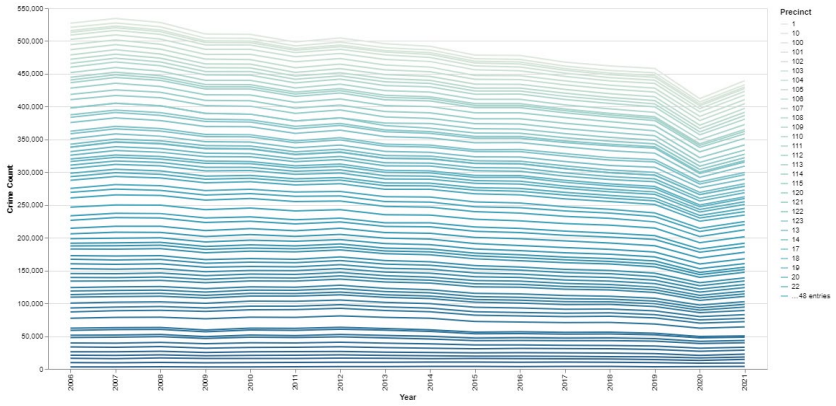

Figure 3 Complaints per year by Precinct

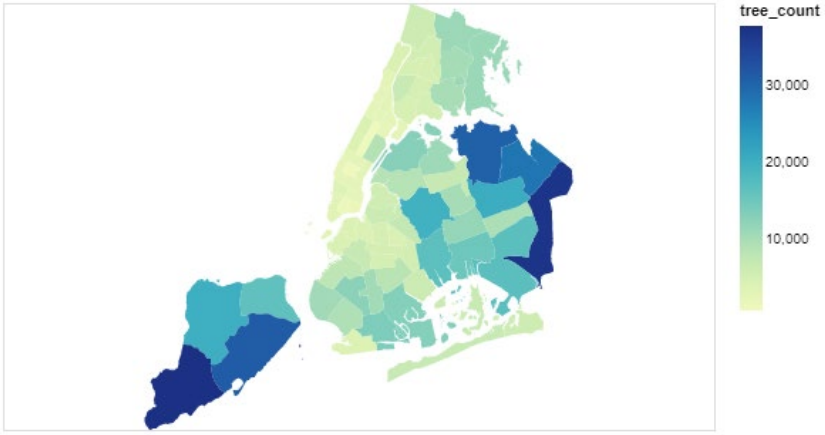

Figure 4 Number of Trees by Precinct

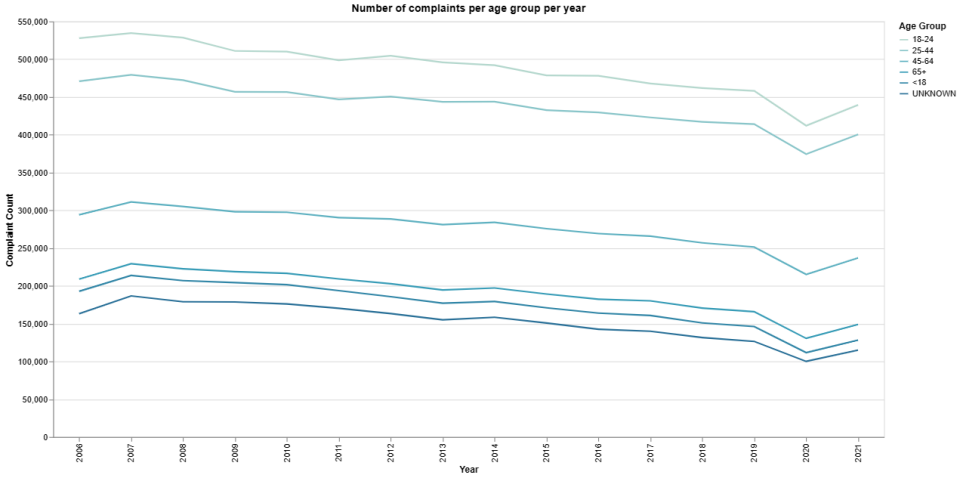

Figure 5 Victim complaints per year by Age Group

6.1. Descriptive of Suspects & Victims demographic

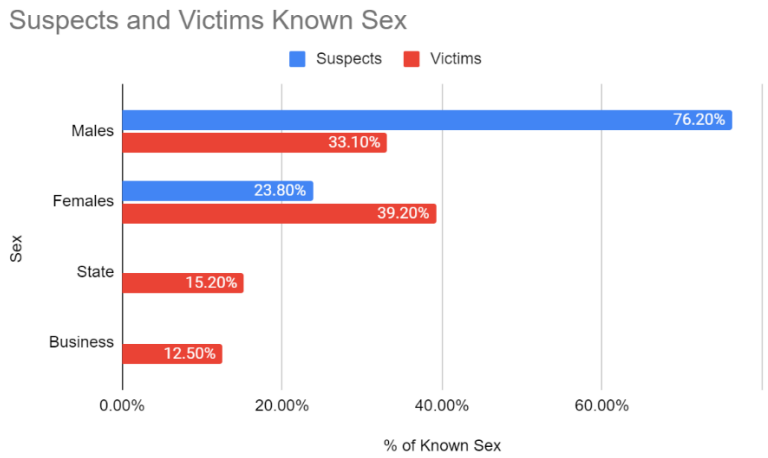

Figure 6 Suspects & Victims' Sex

The Unknown category of Victims'sex constitutes 0% of the dataset, and 18.18% for Suspects' dataset.

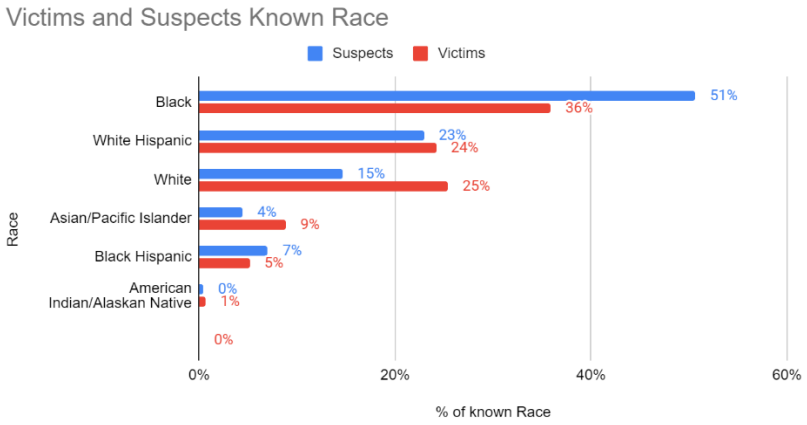

Figure 7 Suspects & Victims' Race

The Unknown category of Victims' race constitutes 32.7% of the dataset, and 26.6% for Suspects' race dataset.

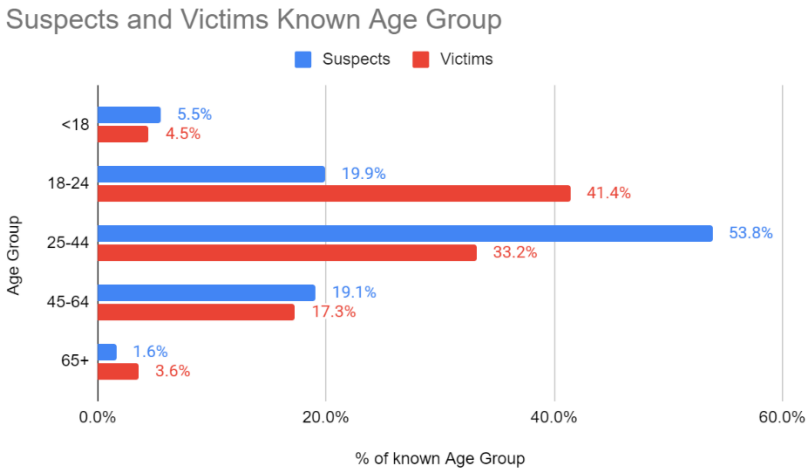

Figure 8 Suspects & Victims' Age Group

The Unknown Victims Age Group is 29% of the total dataset and the Unknown Victims Age Group constitutes 31.2% of the dataset.

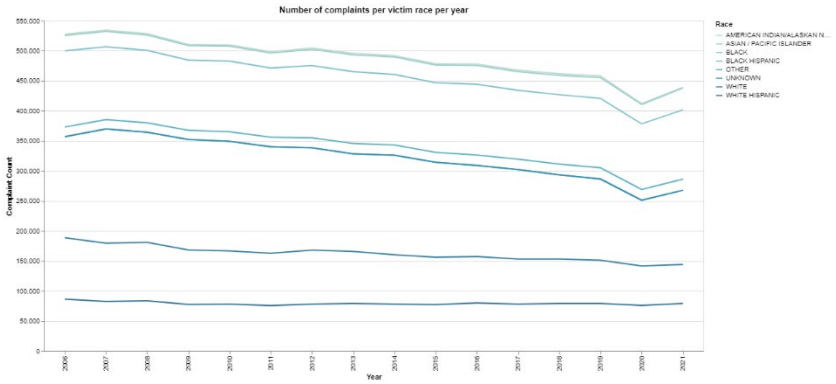

Figure 9 Complaints per Year by Victim Race

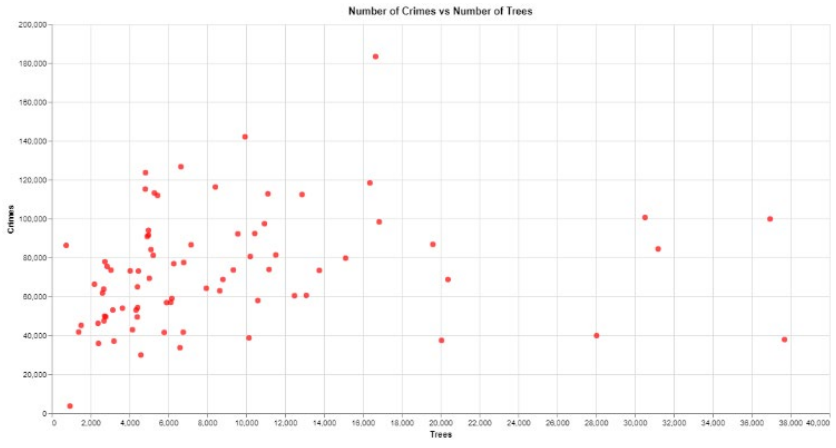

Figure 10. Number of Complaints versus Trees by Precinct – Scatterplot

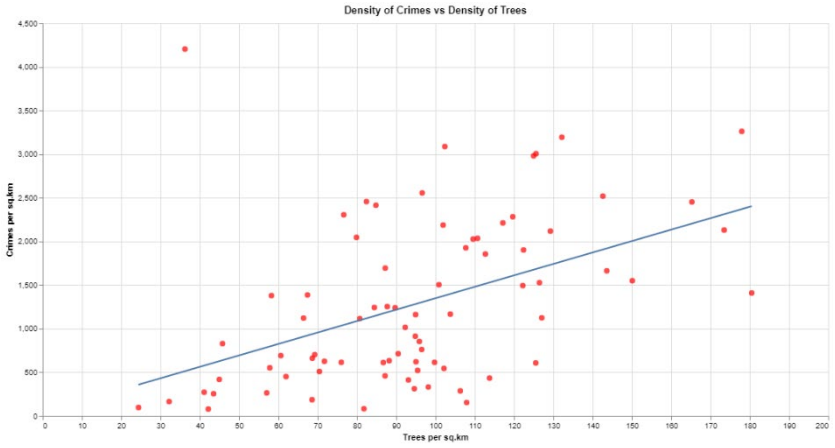

Figure 11. Complaints versus Trees, by Precinct, normalised by precinct area.