Responsible Data Analytics
SEN 163B

Report Part II -
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

April 9th, 2023

1. Final Problem Formulation

The dataset utilized for this report are the NYPD Complaint Data Historic, New York Street Tree Census Data and Police Precincts Map, although there is some limitation in the data in terms of availability. There's no information regarding if the complaint was followed with a charge or if the complaint was fake, which can mislead the machine learning model, also since the TREES dataset is only available for the years 1995, 2005 and 2015, it is not entirely accurate to utilize the whole data of the NYPD. Lastly, the locations given are approximated so it is not recommended to match the location of the incident to an exact address or link.

Full dataset are used after being cleaned according to the assumptions made in Part 1. However, there are two type of crimes that are going to be dropped to this analysis, namely, crimes in prison, rape and sexual crime offenses. Following the principle "Legitimize Embodiment and Affect" of the responsible data guiding principles, to protect the victims' identities and therefore build trust with the police department, rape and sex crime offenses have been located as occurring at the police station house within the precinct of occurrence. New York Street Tree Census Data, Police Precincts Map would be required to make the analysis.

After assessing the possibility of a relationship between greenery and the number of complaints, we find ourselves with another questions. Is there any public policy that can be implemented according to the examination of a machine learning model? Is it viable to 'control' the number of trees of certain region?

2. Responsible Data Guiding Principles

In the paper "Feminist Data Visualization", by D'Ignazio and Klein (D'Ignazio et al., 2016), there is enumerated six responsible data guiding principles that lead us think about the data in a broader way and help us represent the diverse realities of our world.

Regarding the principle "Rethink Binaries", which emphasizes adopting strategies to achieve more multiplicity on the categories rather than 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 principle "Embrace Pluralism" encourages to look at the data visualization from other perspectives and experiences, so there are no voices left out and the research can be more meaningful for more people. The available dataset doesn't take into account, for example people that don't identify with either the female or male gender or even a whole range of races that are not represented (about 30% of the data). Which means that there can be correlations that won't be concluded due to a narrow collection of data.

Related to the principle "Examine Power and Aspire Empowerment", there is a need to take a closer look on the way how power and privilege are distributed amongst the data collectors, data processors and data subjects. The data collector in the dataset studied is the New York Police Department, data processors are the students writing this report and data subjects are the victims that pressed complaints

and the respective subjects. The way that power is distributed in this case has a strong historical aspect that led us to another principle “Consider context”, which highlights the importance of the historical, social and economic context of the data. New York has an history of police brutality, namely when referring to Black and Hispanic people. By coincidence or not the research made on the data shows that both races are the highest number of suspects attributed to complaints, that is a warning to pay closer attention to the biases that may be in factor. This makes us think about the influence that the role of power and historical context play in this dataset.

“Legitimize Embodiment and Affect” principle pays special attention to the people’s emotions and experiences through the data visualization. The fact that the data is anonymized increases trust between victims and the police department. Also, the type of crime communicated in a non-graphic way helps the readers understand the data and its impact, acknowledging the victim’s suffrage.

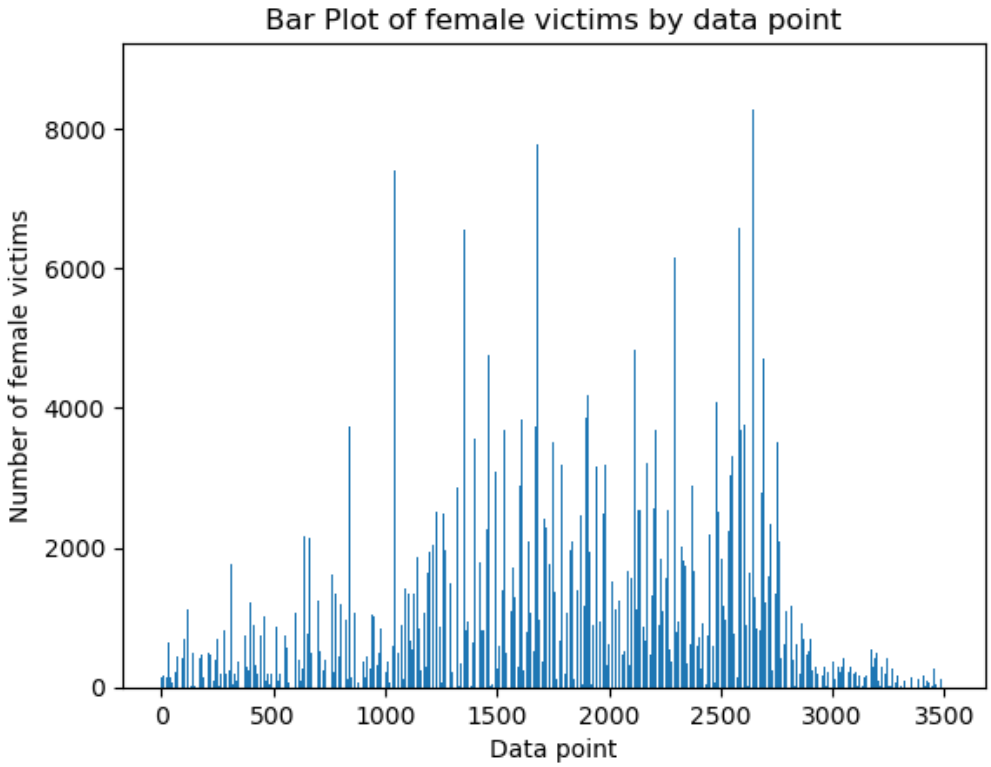
Lastly, regarding the principle “Make labor visible” to fairly attribute credit to all individuals involved in the data. We couldn’t process this data without mentioning the people in the New York Police Department responsible for collecting the data related to the complaints and, especially, the victims that had the courage to press complaints and revive the traumatic experience.

3. Predictive Analytics

A predictive model has been built to make a prediction on unseen data, based on existing data. The predictive model is a regression model and the target variable to be predicted is “the number of female victims”. This target variable is calculated based on the following variables: Longitude, Latitude, Hour of the day, Day of the week, Fair tree count, Good tree count and Poor tree count.

The existing data is treated as followed: the map of New York is divided into a grid of points and each points counts the crimes by sex and the trees by tree condition near that point. A new table is created and for each point, on each hour of the day and day of the week, the number of trees by their condition is calculated as well as the number of female victims.

The distribution of the target variable (number of female victims) over the data points is unbalanced. On average, each datapoint has 863 female victims. The data point with the highest number of female victims has over 8000 female victims while several datapoints have 0 female victims. This distribution shows that some datapoints are much more dangerous for women than other datapoints.



The dataset is then split into a test and train set. The train set is used for machine learning and trains the algorithm to predict based on existing data. The test set is used to estimate the model performance. A baseline model, using the linear regression function, is computed. This model is simple, easy to fit and is used as a reference point for performance analysis. This baseline model will be compared to the following supervised predictive models:

- Lasso
- Elastic net
- Decision tree
- Random Forest
- Neural net
- K-nearest neighbors

For the performance analysis, the mean squared error test, the mean absolute error and the mean absolute percentage error are used, and the results are shown in table 1.

	Baseline	Lasso	Elastic net	Decision tree	Random forest	Neural net	K-nearest neighbors
MSE	54.50	56.49	58.9	40.99	44.03	49.51	9.18
MAE	4.67	4.63	4.73	3.61	3.87	4.25	1.76
MAPE	$4.86*10^{15}$	$4.55*10^{15}$	$4.67*10^{15}$	$2.22*10^{15}$	$3.14*10^{15}$	$3.34*10^{15}$	$7.16*10^{14}$

Table 1. Performance analysis per predictive model

As can been seen, K-nearest neighbors scores much ~~lower~~ better on all metrics than the other supervised predictive models and the baseline model. The simple baseline model even outperforms the

Lasso and Elastic net on some metrics. As the K-nearest neighbors model scores lowest on all metrics, this model will be the predictive model used for the dataset.

4. Prescriptive Analytics

4.1. Operationalization of the Model

The proposed use of the model is on 2 levels. First to better allocate police resources to deter crimes against vulnerable groups – in this care women. The intent is that police precincts may use the model as a part of their resource allocation strategies. A precint would be provided a map that identifies areas and times where women are likely to be more vulnerable. They may choose some locations for additional, or even more visible patrols, as a deterrent. The system would be able to continuously monitor and update itself as reports change, and be updated with additional data sets to improve predictions over time.

That connects to the second level of operationalization. By integrating the presence of trees in the risk assessment, the model assigns a weight to that factor. It may be used to make policy prediction on the impact of trees on crime that are more nuanced that simple linear relationships. It can be used to judge the social impacts of greening projects, as well those of restricting access to certain spaces due to construction or redevelopment. It also serves as a proof of concept to add more spatial features to a predictive model like this, such as mass transit stations or liquor stores.

4.2. Model Validation

Given the temporal nature of the model, a temporal validation test was devised to assess model performance in the real world. The training data for the model was restricted to the years up-to and including 2020. The year 2021 was used to validate the model's predictive performance.

The performance of the model under validation is shown below:

Method	Test Performance (MAE)	Validation Performance (MAE)
Lasso	0.440036	0.465551
Elastic Net	0.447509	0.472120
Decision Tree	0.384958	0.402807
Random Forest	0.411742	0.426900
Neural Net	0.467020	0.420941
K-Nearest Neighbors	**(Best) 0.326714**	**(Best) 0.332538**
Average	0.412996	0.420142

The performance is close to the performance in the test case, implying that the model is able to make similar quality predicts going into the future.

5. Privacy, Positionality, and Bias I

5.1. Privacy Impact on the Project

Privacy plays an important role when data analysis is conducted. The data owner should have the right and control to protect their own data. However, there are spaces and times where and when we are vulnerable and we have an ethical duty to safeguard our own and others' privacy even when "the use of our data" is said to be in our best interest (Allen, 2011). This might result in a situation where certain data in the dataset needs to be protected in order to safeguard the data owner.

The NYPD Complaint Data Historic dataset used in this project contains information on various types of crimes, including murder, sexual-related crimes, fraud, and assault. To preserve the victim's privacy, the closest police station is designated as the location of sexual crimes in this dataset. This reason has an impact on the data analysis that was performed in this project.

The relation between green space and crime rates in this project is calculated by finding the correlation between the complaint, the police precinct map, and the tree location. To determine whether there is a correlation between a crime and trees nearby, the analysis requires the exact location of each crime. Because the actual location of the sexual crimes is unknown, the correlation between the trees and sexual-related crimes may be rendered invalid. As a result, sexual-related crimes are excluded from the analysis to provide more reliable and valid results.

When necessary data, as in the case of the NYPD Complaint Data Historic case, cannot be published for various reasons, the privacy clause may lead to ambiguous results in the analysis. Even though the complaint is anonymous, knowing the precise location of the crime could spark a commotion and put the victim in danger because their identity may be revealed. To protect the victim from harm, the privacy of the victim must be taken into account during data analysis and processing.

5.2. Positionality and Reflexivity

Positionality will impact our perception of certain things. The positionality necessitates the researcher consciously examining their own identity to allow the reader to assess the effect of their personal characteristics and perspectives in relation to the study population, the topic under study and the research process (Wilson et al, 2022). In this project, assessing positionality may aid in the creation of a more thorough conclusion, particularly when discussing the race, age, and gender of the suspects and victims.

Positionality is normally identified by locating the researcher's position in relationship to three areas: the topic under investigation; the research participants; and the research design, context and process (Holmes, 2020). In this project, we (as the researchers) are not closely related to NYPD crimes. Our only knowledge of NYPD crimes comes from this course. As a result, our positionality could be placed in a more objective context.

In terms of suspects and victims, the case in NYPD crimes involves certain races, ages, and genders. Some of us are closely related to specific races, that is Asian and White. Researcher positionality is commonly discussed in the literature as a clear distinction between insider (emic) and outsider (etic) perspectives

(Huberman & Miles, 2002). We positioned ourselves as outsiders because we had no idea who the suspects and victims were and had no connection to the case.

Finally, in conducting the research design, context, and process, we employed the tools and processes discussed throughout the course. We have no personal interest in the analysis, so the outcome can be as objective as possible.

5.3. Pre-processing Bias Analysis

5.3.1. Representation Bias

The dataset contains several variables, including the age group, race, and sex of the suspects involved in the crimes. However, these variables may be subject to representation bias. For example, the dataset shows that Black individuals are the most common suspects, followed by those of unknown race. This suggests that there may be overrepresentation of complaints against Black individuals in the dataset, which could be a result of systemic racism or other factors.

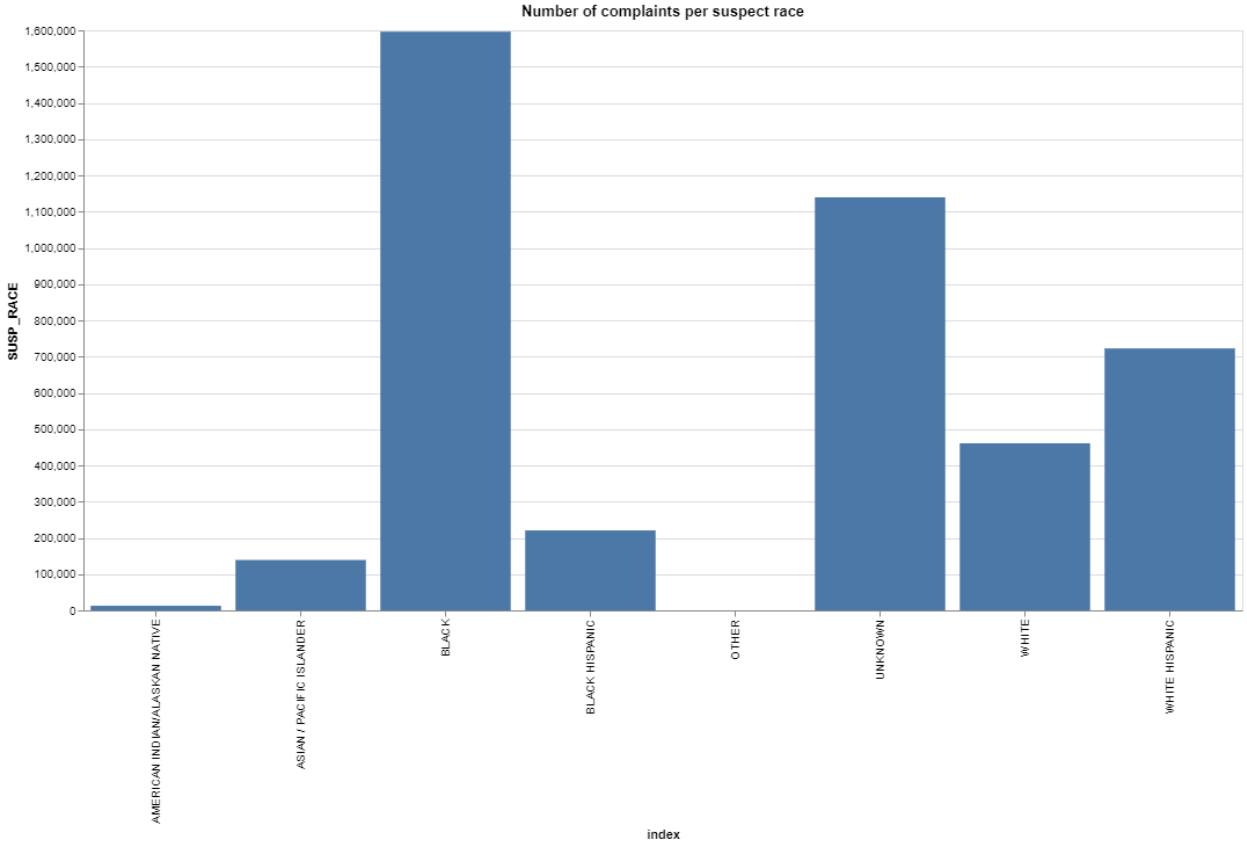


Figure 1 Number of complaints per suspect race

Another possible representation bias is related to the sex of the suspects involved in the crimes. The data shows that the majority of suspects are male, compared to a smaller proportion who are female. This could indicate that complaints against male suspects may be overrepresented in the dataset, which may be a result of gender bias or other factors.

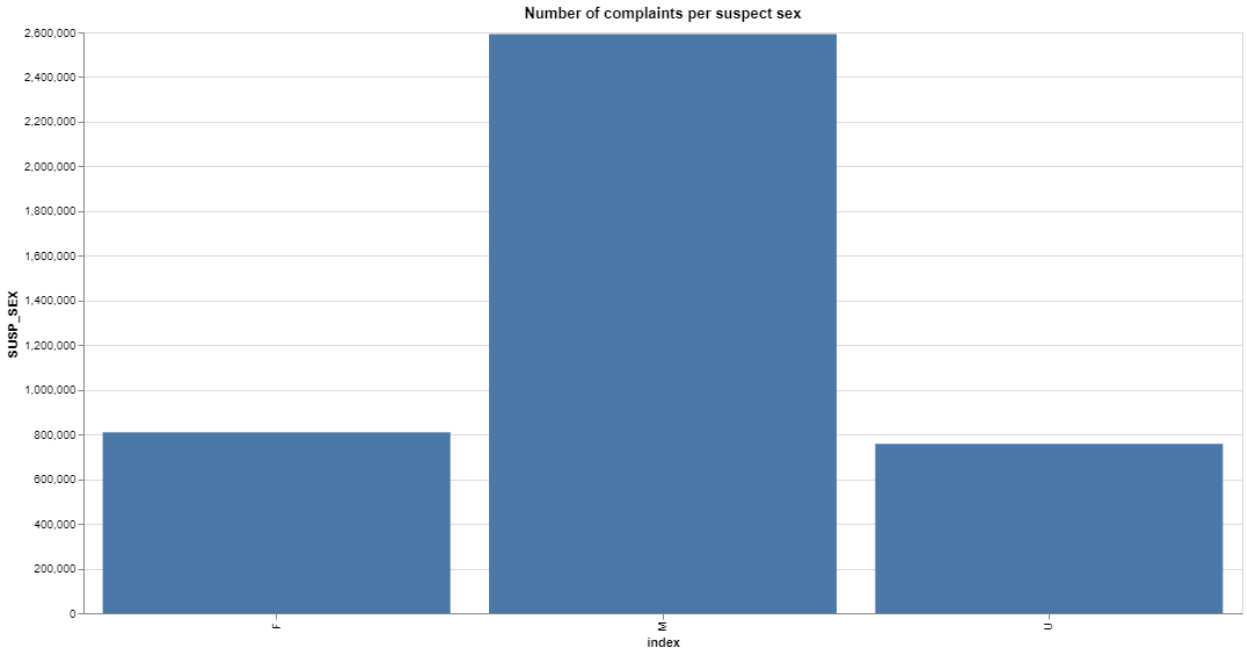


Figure 2 Number of complaints per suspect sex

Additionally, the age group of the suspects involved in the crimes may also be subject to representation bias. The data shows that suspects in the 25-44 age group are the most common, followed by those in the 18-24 age group. This may suggest that complaints against younger and middle-aged suspects may be overrepresented in the dataset, which could be a result of age bias or other factors.

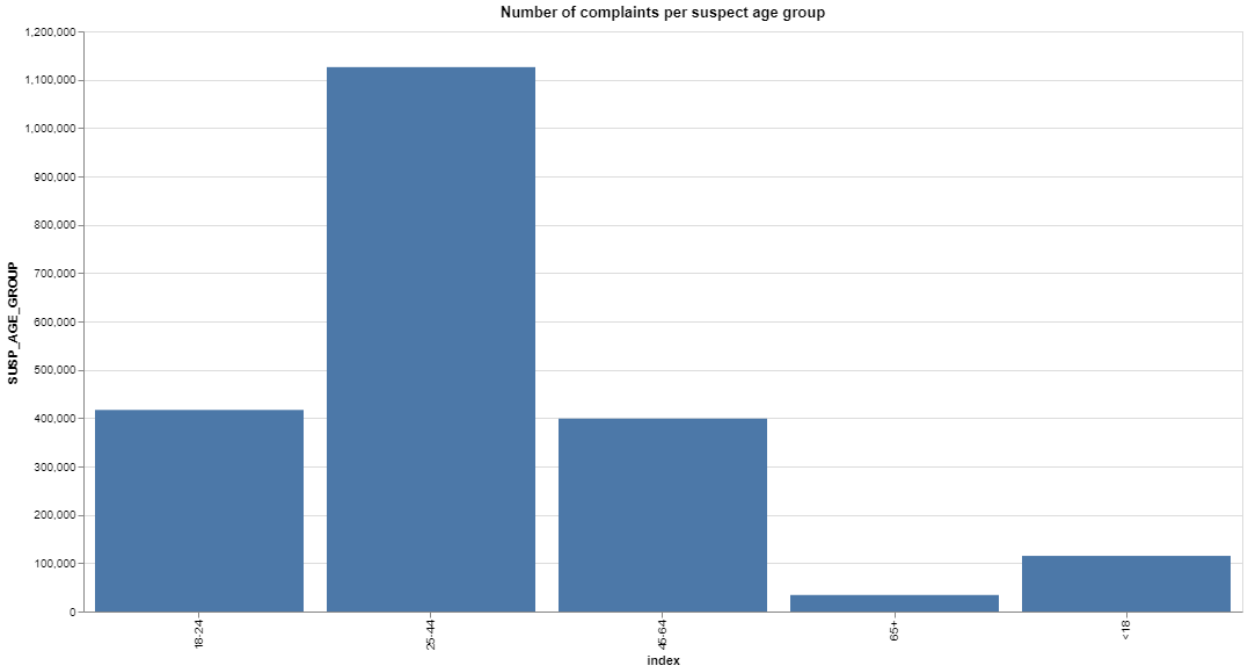


Figure 3 Number of complaints per suspect age group

It is important to consider the context in which the crimes occurred and to use additional data sources to test for representation bias.

Furthermore, it is important to note that our predictive model is not using the age group, race, or sex of the suspects involved in the crimes. We will focus on other variables, such as the crime location, type of crime, and time of day, to make predictions. By doing so, we can avoid perpetuating any biases present in the data and work towards creating a more equitable and just criminal justice system.

6. Fairness (Bias II)

When conducting data analytics projects, it is important to consider fairness in decision-making. Fairness can be achieved by ensuring that different groups of people are treated equally and not discriminated against based on certain attributes, such as race or gender. However, achieving fairness can be challenging, especially when using machine learning algorithms that may be biased due to historical data.

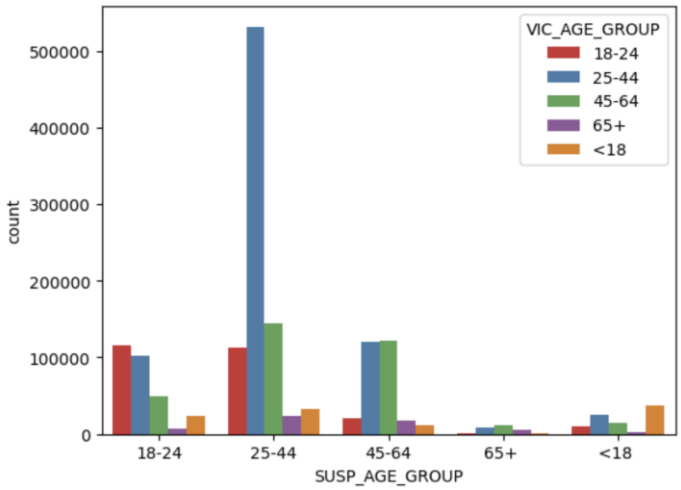
One way to address this challenge is by using the observational fairness framework, which seeks to equalize observable outcomes across different groups. This framework involves defining sensitive attributes and proxies, and then equalizing metrics such as accuracy or false positive rates across these groups (Friedler et al., 2019).

It is important to note that achieving fairness can be difficult due to various factors, including measurement biases and misreported attributes (Friedler et al., 2019). Moreover, different stakeholders may have different perspectives on what constitutes fairness. For instance, judges may be more interested in positive predictive parity, while convicts may prioritize false negative error rate balance (Friedler et al., 2019). Society, on the other hand, may value demographic parity to ensure that certain groups are not unfairly targeted (Friedler et al., 2019).

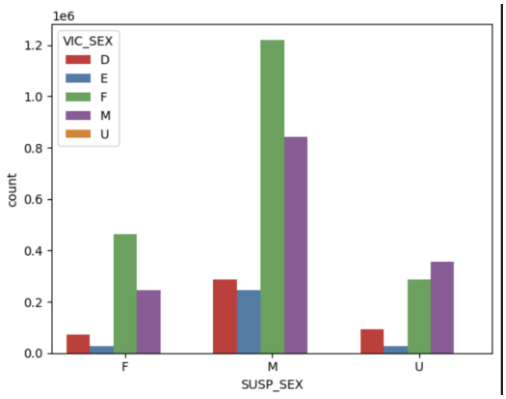
In addition to considering fairness, data analytics projects should also take into account distributive justice, which involves ensuring that resources and benefits are allocated fairly among different groups (Longley, 2022). The capability approach, proposed by Nussbaum (2003), can be a useful framework for achieving distributive justice by focusing on ensuring that individuals have the necessary capabilities to lead fulfilling lives.

Overall, achieving fairness and distributive justice in data analytics projects requires a nuanced understanding of different frameworks and perspectives, as well as a willingness to engage with stakeholders and address potential biases in the data and algorithms used.

From the analysis, there is disparities between victims and suspects in race and age group. Most suspects will threat victims with the same age group with them. The highest victims are on the age of 25-44.



It also happens when gender of the victims are being grouped. Most of the victims are female for all of types of suspects.



7. Transparency and Explainability

Within the last years, transparency has become a must in data analytics projects. The requirement of transparency came along with the need to understand and share with all the stakeholders how decisions are made, how to make them fair and just and to attribute accountability to those decisions. However, transparency in the context of machine learning often requires explainability, it shouldn't be its only focus.

Transparency should go beyond explaining patterns and decisions made by machine learning models. Even though, explainability already gives the stakeholders involved a sense of trust, transparency involves providing not only the context of the data but also all the limitations, biases and risks that the data entails in order to maximize fairness.

Managing to achieve transparency is a really demanding process due to the utilization of complex models and the need of large datasets. Large amounts of data are a step to increase accuracy and for that, external stakeholders need to get involved. Key words like safety and reliability are essential for this process because within the exchanging of information there needs to be regulation to safeguard possible special and sensitive data.

In the context of this report transparency beyond explainability plays an important role because data limitations and biases related to demographic information have severe consequences. To avoid these problems there could be implemented measures that pay close attention when registering complaints to ensure that the type of crimes and the demographic of either the victim or the suspect are accurately registered.

We must not forget to communicate with the involved stakeholders to make sure that the data is used in an ethnically and morally corrected way and this way can be used to try to develop ways of informing a possible crime in New York City.

8. References

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2019). The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Proceedings of the Conference on Fairness, Accountability, and Transparency, 1-16.

Longley, R. (2022). What is Distributive Justice. ThoughtCo. Retrieved from https://www.thoughtco.com/distributive-justice-373317.

Nussbaum, M. C. (2003). Capabilities as fundamental entitlements: Sen and social justice. Feminist economics, 9(2-3), 33-59.

Pollock, Meagan (2021 March 24). What is positionality?. Engineer Inclusion. https://engineerinclusion.com/what-is-positionality/

Wilson, C., Janes, G., & Williams, J. (2022). Identity, positionality and reflexivity: relevance and application to research paramedics. *British paramedic journal*, *7*(2), 43–49. https://doi.org/10.29045/14784726.2022.09.7.2.43

Holmes A. (2020). Researcher positionality. A consideration of its influence and place in qualitative research: A new researcher guide. *Shanlax International Journal of Education*, 8, 1–10.

Huberman A. M. & Miles M. B. (2002). *The qualitative researcher's companion*. SAGE Publications.