Exploring the Relationship between Female Victims and Crime using Clustering Algorithms

Alexandra C. Coffin

Northwest Missouri State University, Maryville MO 64468, USA
S561404@nwmissouri.edu

Abstract. The purpose of this project was the examination of crime data with a focus on female crime victims to explore the effectiveness of clustering algorithms. Through the application of K-means and DBSCAN, on data collected from the New York Police Department (NYPD) and the National Crime Victim Survey (NCVS). K-means was selected as the first algorithm as it clusters data based on the distance from a centroid, a classic method for identifying patterns within crime data. DBSCAN clusters data based on density calculations to examine crimes that occur in a similar location to discover hotspots and outliers. The findings of this project reveal patterns demonstrating the intricate relationship between age, geographical location, crime type, and women in victimology. We removed sensitive information about victims and instances that did not contain age, geolocation, sex, and crime type.

Keywords: data analytics · criminal justice · female victimology · machine learning · DBSCAN · K-means

1 Introduction

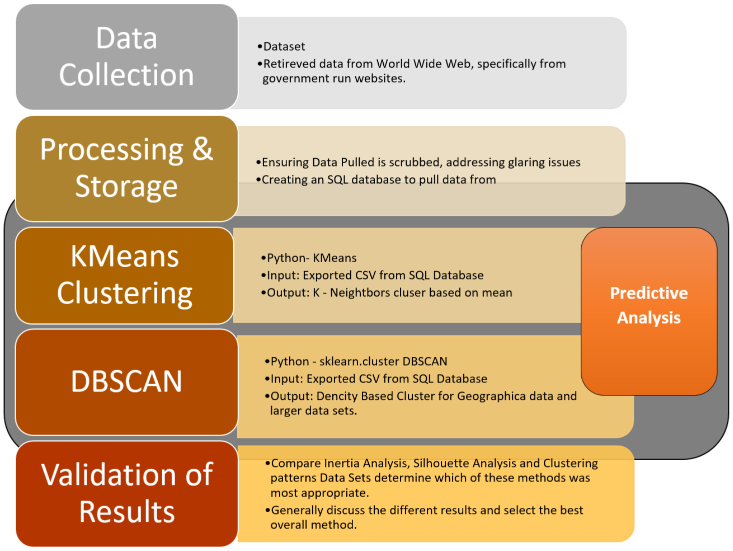
Crime is an ever-evolving issue as technology brings the world closer together - it creates an even greater complexity when discussing crime victimology. Crime victimology is a broad field of study spanning from understanding individuals and criminals to social factors and technological advancements [5]. Examining vulnerable groups enables investigators and officers to be more effective. The Criminal Justice System focuses on the actions of criminals, justice, and criminal rights, but not necessarily on the victims subjected to these actions. The result is a lack of research into the patterns of criminals when selecting their victims in combination with geographical factors. This project revolves around the analysis of female victims who were subject to personal crimes. According to the 2022 FBI Crime Data released through the CDE from 2011 to 2022, there were 469,261 female victims of crimes committed, 48.402 % of the total number of victims from that period [6].

The number of cases reported to the FBI and the National Incident-Based Reporting System (NIBRS) is immense. This project uses data from two sources: the National Crime Victim Survey[11], and the New York Police Department [14]. Each data set was selected based on the completeness of data, diversity, and ability to obtain victim data.

1.1 Objective of this Research

The primary goal of this project is to explore the applications of Machine Learning Models when predicting crimes common among women geographically and provide increased awareness of crime in America.

Section 2 explains the Methodology, 2.1 Data Collection, and 2.2 Storage and Cleaning. Section 3 Exploratory Data Analysis, 3.1 Load the Data for EDA, 3.2 NYPD Data, 3.3 NCVS Data, 3.4 Conclusion of Exploratory Analysis. Section 4 Predictive Modeling, 4.1 Pre-Processing, 4.2 K-means, 4.3 DBSCAN. Section 5 Discussion, 5.1 Limitations, 5.2 Future Work based on research findings, and Section 6 Conclusion.



- Dataset
- Retrieved data from World Wide Web, specifically from government run websites.

- Ensuring Data Pulled is scrubbed, addressing glaring issues
- Creating an SQL database to pull data from

- Python- KMeans
- Input: Exported CSV from SQL Database
- Output: K - Neightbors cluser based on mean

- Python - sklearn.cluster DBSCAN
- Input: Exported CSV from SQL Database
- Output: Dencity Based Cluster for Geographica data and larger data sets.

- Compare Inertia Analysis, Silhouette Analysis and Clustering patterns Data Sets determine which of these methods was most appropriate.
- Generally discuss the different results and select the best overall method.

Fig. 1. *Methodology applied to this project.* Created by Alexandra Coffin.

2 Methodology

The workflow of collecting and analyzing victim data is as follows. 1) Data Collection, 2) Methodology, 3) Data, 4) Exploratory Data Analysis, 5) Predictive Modeling, and 6) Validation of Results.

3 Data

The data for this project was collected from the National Crime Victims Survey(NCVS) Dashboard[11] from the Bureau of Justice Statistics and the NYPD Open Data Portal[14] as static files in CSV format.

The NCVS data is the baseline comparison because it has the lowest statistical variance and was assembled by the Bureau of Justice of Statistics. All of the data being examined is related to female crime victims. The statistical variance for all of this data is two measures of the accuracy of an estimate [10].

The data attributes utilized are Sex, Age, Location, Zip Code/County, Offense Code, and NCIC Code. Women are the focus of this project, as they are statistically more likely to report a crime than men [4]. The table below summarizes the characteristics of each data set and the date the files were updated. Since these are CSV files, the range of dates is also included in the table.

Table 1. *A summary of the characteristics of each source utilized in this project.* Created by A. Coffin.

Sources List

Source	Entries	Cl. Entries	Raw Attributes	Attributes	Dates
NYPD[14]	8.3M	2.7M	35	17	1922-2023
NCVS[10]	1344	1344	8	6	1993-2022

3.1 Attribute Selection

Crime victim data also contains entities such as businesses, organizations, government organizations, and even society. These crimes include vandalism of a business to traffic tickets and have been removed. All the attributes are collected from various sources and vary slightly depending on the source. Columns that were not pertinent to this project were removed. Names of attributes were altered for uniformity across tables. The annual NCVS report releases a summary of victims based on specific variables see table 2.

When examining crime, the dependent variables are the type of offense committed and domestic violence. Offenders will seek victims to meet their needs with a specific type of offense in mind. Independent variables include sex, age, race, geographic location, and firearms. Observing these variables provides a better understanding of who crimes affect. When interpreting national and regional statistics, the results will differ. This project analyzes the relationship between sex, age, location, and types of crimes committed. A series of models was created with the NCVS data as the basis for comparison and generating models for the NYPD data to explore applications of clustering algorithms.

Table 2. *A summary of data attributes from each dataset.* Created by A. Coffin.

Attribute List

Feature Name	Description	Data Type	data sets
RPT_NUM	Report/Case Number	String	All
RPT_DT	When the crime was reported	Date	All
RPT_FROM_DT	When the investigation/quarter began	Date	All
RPT_FROM_DTM	MS Excel number associated with date	Number	NYPD
RPT_TO_DT	When the investigation/quarter ended	Date	All
YEAR	Year of the incident	Number	NCVS
PD_CODE	Local Offense Code varies depending on the region.	string	NYPD
NCIC	Penal Code used by the Federal Justice System	Number	NCVS
OFENS_DESC	a short description of the offense	String	All
BORO_NM	Area, City, or Borough where the incident occurred.	String	NYPD
REGION	Region Data was collected	String	NCVS
REGION_M	Numerical Assignment	Number	NCVS
ZIP_CODE	Postal number associated with area	Number	NYPD
SEX	Biological Sex of the Victim	String	All
RACE	Race of the victim	String	All
AGE_GROUP	Age group that victim fell into	String	ALL
AGE_GM	Number Assigned with correlating AGE_GROUP	Number	All
VIC_NUM	Total Number of victims specific category	Number	NCVS
LOCATION	Point where the incident occurred	String	NYPD
LAT	Latitude coordinates of incident	Float	NYPD
LONG	Longitude coordinate of incident	Float	NYPD

3.2 Cleaning, Date Range & Storage

The data was loaded into Tableau Prep Initially to handle the adjustments. All data sets were modified in terms of uniform age groups and NCIC. Nulls are common in victim data as many police forces remove victim information for individual safety. An entry was kept as long as three of the four requirements, except for homicides. The requirements included sex, location, age, and type of crime. Homicides are the exception locations that are occasionally omitted to preserve crime scene integrity.

An additional column has been added to the NYPD data set to translate AGE_GROUP into a usable format for modeling. Each AGE_GROUP has been assigned an associated number based on numerical order in a new column labeled AGE_GM. The conversion of each age group is as follows: under 18 = 1, 18-24 = 2, 25-44 = 3, 45-64 = 4, 65+ = 5, Unknown = 6.

A series of tables were drawn from the NCVS Dashboard combined to create a more extensive data set through unions for a comprehensive set. Any data with a coefficient of variation greater than 50% but based on 10 or fewer samples is included for individuals under 18 and those 65 and older. In both of these, data collected is not often reliable. In the case of minors, the crime may not be reported, or the abuse is complex. Whereas with the elderly, it's a combination of lack of attention when crimes are reported, the Fear-Crime Paradox, risk, and vulnerability [4]. Both groups are considered vulnerable based on physical strength, financial dependence or limited financial power, geographical risk, and reliance on a caretaker. Types of crime data that were flagged remained in instances of sexual assault and robbery as these are not consistently reported.

All of the data referenced was pulled from 1993 to 2022 to match the date range of the NCVS data. In cases where the date range is shorter, the entire data set is required. The only special characters remaining in the data are those for AGE_GROUP and LOCATION data in the form of a dash and parentheses. The remaining special characters have been addressed by replacing and removing attributes. LOCATION data has been transformed into the order of SQL columns for latitude and longitude added to store this data separately.

Within the NCVS data, several columns were added to ensure that the data was in a format that predictive models could interpret. This includes coding for Age Groups utilizing a number structure for each Age Range, Crime Type encoded with associated NCIC numbers, and Regions numbered 1 through 4. Encoding is covered in detail within the repository associated with this project.

All final details are listed in the Source List Table 1, including the number of entries for each cleaned data set and the number of selected attributes. The data was then loaded into an SQL database to allow for rapid queries and data selection. Queries are used to pull from larger sets and demonstrate the effectiveness of applied methods to smaller sets. The results of these queries were then loaded into a pandas data frame to allow for further analysis.

4 Exploratory Data Analysis

Exploratory Data Analysis is an essential step in any project utilizing data. During this stage, the data structure is revealed along with possible outliers. A sense of story is developed with visuals and decide on the data with the strongest correlations. As this project requires multiple sources, the NYPD[14] and the NCVS[10] the data analysis has been broken into a section for each. There are two aspects of the data to be addressed, the first being the quality of the data in terms of the shape of the data, correlation, and possible error. The second is that evaluating this data establishes a sense of known data, allowing for contextualization of the results of the predictive models.

All of the analysis can be found in the following Repository.

https://github.com/accoffin12/CrimeVictimAnalysis_Capstone/

4.1 Load the Data for EDA

All of the data was exported from the SQL database as CSV files and then loaded into Pandas Data Frames using the following code.

```
NYPD = pd.read_csv('Data/NYPDv3.csv')
NYPD_AgevCrime = pd.read_csv('Data/NYPD_AgeVCrime.csv')
```

```
NYPD_BoroCrime = pd.read_csv('Data/NYPD_BoroCrime.csv')

# Importing NCVS Files:
NCVS_RegionSeg = pd.read_csv('Data/NCVS_RegionSegv1.csv')
NCVS_Region = pd.read_csv('Data/NCVS_Regionv1.csv')
NCVS_AgeSeg = pd.read_csv('Data/NCVS_AgeSeg.csv')
NCVS_AgeType = pd.read_csv('Data/NatFlow_AgeTypev1.csv')
```

4.2 NYPD Data

For this analysis, all of the NYPD Data[14] was selected between 1993 and 2023, except a scatter plot depicting the number of crimes reported per day. Most of the Analysis was completed using a modified pull from the SQL database, where the data had been filtered based on year, nulls for crime key codes had been pulled, and only women were selected. To fully understand the data, this analysis examines data manipulated from the initial pull, except the heat map, using a separate sheet coordinating the correlation of age and crime.

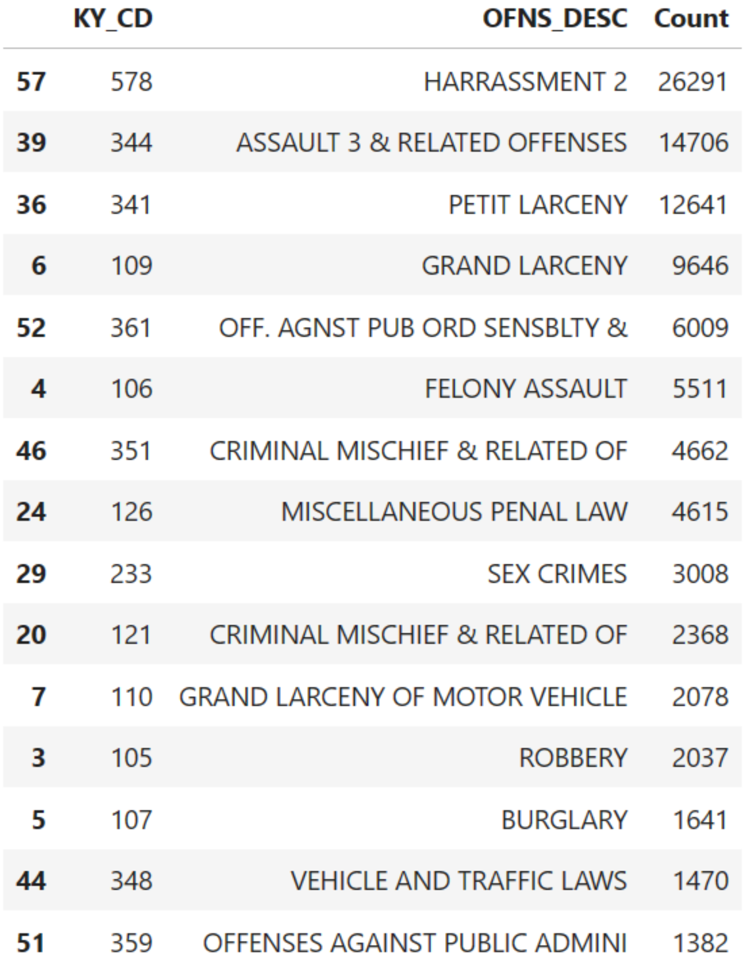
The original data for this project can be accessed from the following URL: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data
-Historic/qgea-i56i

Each case is cataloged as a single entry. The following command grouped the victims based on a specific attribute and was modified periodically to sort data into groups. This sample code was designed to group data based on numeric Key Codes associated with criminal offenses, sort them based on descending numerical order, and finally pull the top 10 most common crimes.

```
grouped_data1 =
NYPD.groupby('KY_CD')['OFNS_DESC'].value_counts().reset_index(name='Count')
sorted_d2 = grouped_data1.sort_values(by='Count', ascending=False)
top_10 = sorted_d2.head(10)
display(top_10)
```

Personal Crimes include harassment, sexual assault, robbery, aggravated assault, and murder. From 1993 to 2023, a majority of the crimes reported with female victims were considered Personal Crimes. Several crimes fall within the Public category of Criminal Activity defined by the NIRBS, and NCIC recognized that 4 of the top 10 most common crimes are considered Public or Social offenses without delving into the complexity of Penal Law. The five most common crimes to have female victims are Harassment 2 with 26,291 incidents, Assault 3 & Related Offenses with 14,706 incidents, Petit Larceny with 12,641 incidents, Offense Against Pub Order Sensibility with 6,009 cases, and Felony Assault with 5,511 cases [1].

```
grouped_data1 =
↪  NYPD.groupby('KY_CD')['OFNS_DESC'].value_counts().reset_index(name='Count')
sorted_d2 = grouped_data1.sort_values(by='Count', ascending=False)
top_15 = sorted_d2.head(15)
display(top_15)
```



	KY_CD	OFNS_DESC	Count
57	578	HARRASSMENT 2	26291
39	344	ASSAULT 3 & RELATED OFFENSES	14706
36	341	PETIT LARCENY	12641
6	109	GRAND LARCENY	9646
52	361	OFF. AGNST PUB ORD SENSBLTY &	6009
4	106	FELONY ASSAULT	5511
46	351	CRIMINAL MISCHIEF & RELATED OF	4662
24	126	MISCELLANEOUS PENAL LAW	4615
29	233	SEX CRIMES	3008
20	121	CRIMINAL MISCHIEF & RELATED OF	2368
7	110	GRAND LARCENY OF MOTOR VEHICLE	2078
3	105	ROBBERY	2037
5	107	BURGLARY	1641
44	348	VEHICLE AND TRAFFIC LAWS	1470
51	359	OFFENSES AGAINST PUBLIC ADMINI	1382

Fig. 2. *Output for creating a table with the top 15 most common crimes in NYC.* Created by Alexandra Coffin.

Crimes such as Murder and Non-Negligible Manslaughter (KY_CD 101) where women were victims were fewer in the period, with 21 cases reported with that code. It is crucial to note there are several codes used to define crimes that have a sexual component, such as 104 (Rape) and 116 (Sex Crimes). Sex Crimes is an umbrella term for a broader category that is broken into several facets-aggravated sexual assault, sexual abuse, and trafficking.

```
# New York PD, NY:
grouped_GEO = NYPD["BORO_NM"].groupby(NYPD["BORO_NM"]).count()
grouped_GEO.plot(kind="barh")
plt.xlabel("Count")
plt.ylabel("Borough")
plt.title("NYPD Incidents by Borough")
```

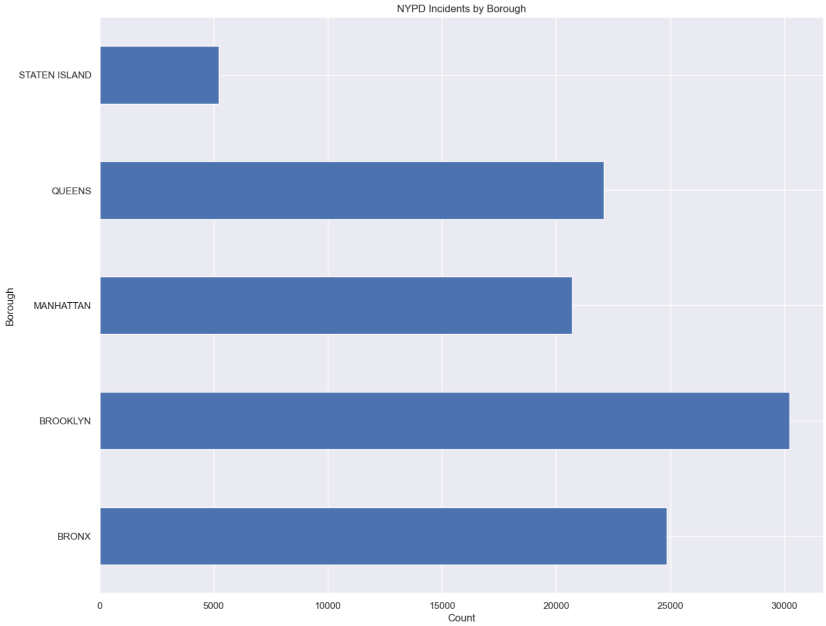


Fig. 3. *A bar graph depicting the relationship between the number of victims and the location of the crime.* Created by Alexandra Coffin.

Geographical location does affect crime, especially when discussing risk factors. Dividing the data based on the location of the incident into each of the five Boroughs there is an indication that Brooklyn and the Bronx and two of the most dangerous areas for women based on the number of incidents reported fig. 3. Whereas Staten Island has the lowest number of reports fig. 3. Geographically Manhattan is the smaller of the five, with only 22.8 square miles [13], with an estimated population of 1,694,251 according to the 2020 Census [2]. Of the five Manhattan is the most densely populated and has the largest flexing population due to the number of commuters. The most populated Borough is Queens with a population of 5,141,538 [2], within 109.7 square miles [13]. Brooklyn has a population of 2,736,074 [2], within 71 square miles, making it the second-largest borough in terms of population and area. The Bronx has a population of approximately 4,208,728 [2], situated above Manhattan it is 42.2 square miles in size. Staten Island has a smaller population with 495,747 people [2] in 35 square miles [15], making it the second smallest borough in terms of size and the smallest population.

The plot below (fig. 4) examines the most recent shift in crime from 2021 to 2023. The years 2019 and 2020 were not included because of COVID-19. During the pandemic, New York State went into lockdown, alternating first responder behavior in response to the virus. The lockdown also resulted in many areas considered public or social gathering locations being closed. This isolation caused a decrease in reported crimes as there was a removal of risk-taking behaviors and social conditions. The result is a massive increase in crimes reported between 2022 and 2023, fig. 4. Risk-taking behavior is a matrix of actions, physical attributes, or aspects of a victim's behavior that offenders seek when deciding to commit a crime. These behaviors include engaging in dangerous situations, living in areas with high crime rates, lack of security, participating in socially deviant behavior, physical limitations, emotional instability, and even economic status [5].

```
# Number of cases reported daily with female victims from 1993 to
↪  2023
#Group Cases Based on Date
grouped_dt =
↪  NYPD.groupby('CMPLNT_FR_DT')['CMPLNT_FR_DTM'].value_counts().reset_index(name='Count')

# Creating Scatter:
sns.set(style="whitegrid")
sns.set(rc={"figure.figsize": (15,12)})
sns.scatterplot(data=grouped_dt, x="CMPLNT_FR_DTM", y="Count")
plt.xlim(44197, 45200)
plt.title("NYPD Reports by Date: 2021 - 2023")
```

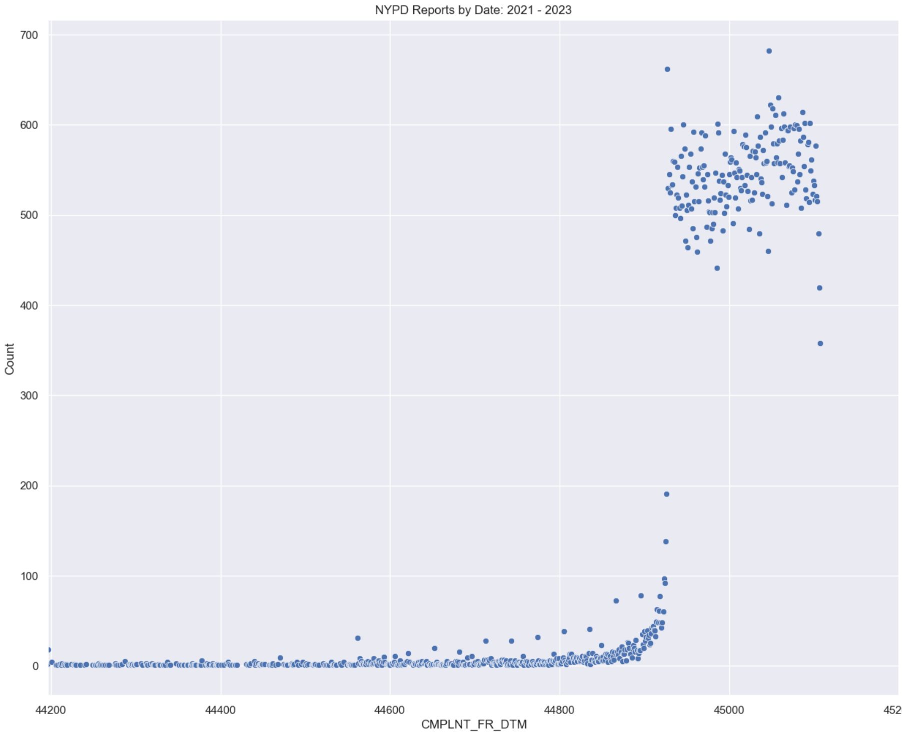


Fig. 4. *Scatter Plot depicting clustering of crime reports from 2021 to 2023.* Created by Alexandra Coffin.

Examining the relationship between age and crime has been the subject of many studies, especially when discussing victimology. There is a correlation between age and the types of crimes committed. For the data selected from the NYPD, there is a strong correlation between victims and crime in terms of Age Groups, while the association between the type of crime and victims isn't as strong (fig. 5. This results in a unique correlation coefficient that is slightly negative. Women between 25-44 years old (Group 3) with 51,774 incidents had the highest number, which translated to roughly 49.42% of cases reported.

```
#heatmap Code:
matrix = NYPD_AgevCrime[['>18', '18-24', '25-44', '45-64', '65+',
'ky_cd']].corr()
mask=np.triu(np.ones_like(matrix, dtype=bool))
sns.heatmap(NYPD_AgevCrime[['>18', '18-24', '25-44',
'45-64', '65+','ky_cd']].corr(),
annot=True, vmax=1, vmin=-1, center=0, cmap='vlag', mask=mask)
plt.title("Correlation of Age Groups and Key Codes")
```

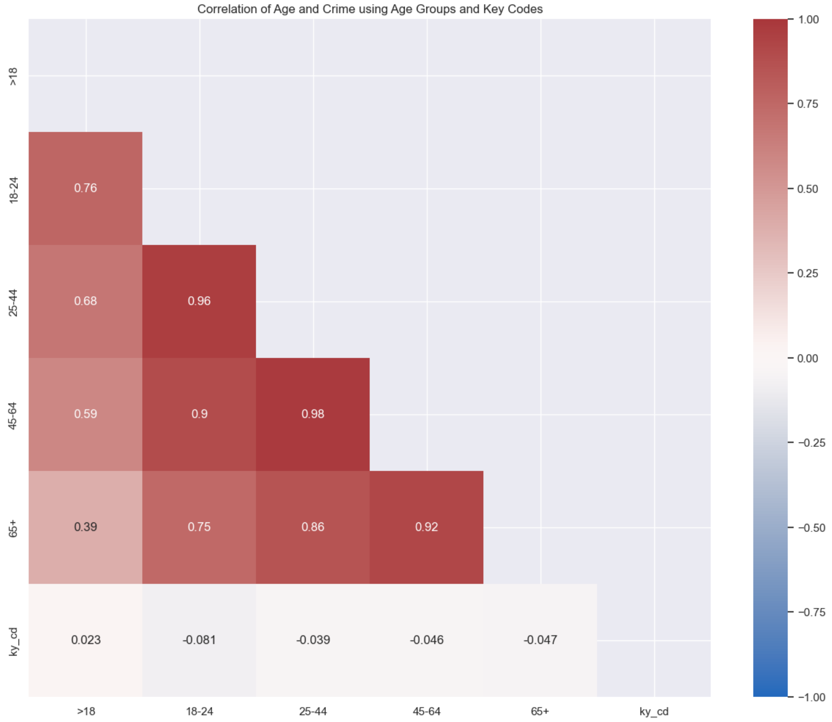
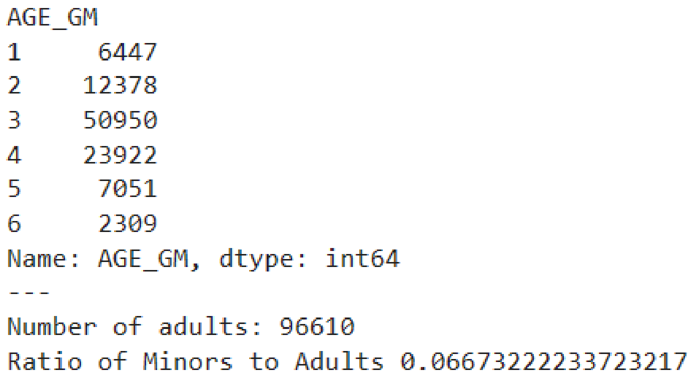


Fig. 5. *A heatmap examining the relationship between victim age and the type of crime they were subjected to.* Created by Alexandra Coffin.

The ratio of minors to adults is roughly 6.67%, or 6,447 cases involving minors (fig. 6).

```
# Determining the ratio of adults to minors:
print(grouped_AGE)
print("---")
NYAdults = 12378 + 50950 + 23922 + 7051 + 2309
print("Number of adults:", NYAdults)
RatioAMinor = 6447/NYAdults
print("Ratio of Minors to Adults", RatioAMinor)
```



```
AGE_GM
1     6447
2    12378
3    50950
4    23922
5     7051
6     2309
Name: AGE_GM, dtype: int64
---
Number of adults: 96610
Ratio of Minors to Adults 0.06673222233723217
```

Fig. 6. *Output based on age group to determine the ratio of adults to minors and total count of individuals per age group.* Created by Alexandra Coffin.

4.3 NCVS Data

The NCVS Data[10] consisted of several files that were condensed into several files. These files revolve around the Type of Crime, Region, and Age Group of female participants. It is important to note that data regarding age does not reflect the ages of women in the regions surveyed, as a user can not pull data from the N-Dash with more than two parameters. Regional Data spans from 1996 to 2022 and consists of the four Major Regions of the United States: the Northeast, South, Midwest, and West. All Data Related to the N-Dash can be accessed from the following URL: https://ncvs.bjs.ojp.gov/multi-year-trends/crimeType

The mean for the number of victims across the entire set is $8.613e^8$, with a standard deviation of $1.123e^8$, and all the data was within an acceptable distance from the mean. When examining Segmented Data collected Regionally from the N-Dash, there was a greater distance with the standard deviation being $2.420e^8$ and the mean for that set as $2.158e^8$. While the segmented data is further from the mean, the total number of cases for this dataset is $4.320e^5$, which is higher when compared to the Regional Data Set containing $1.080e^5$ surveys ??.

Between 1996 and 2022, there has been a significant reduction in crime in the United States. This is evident when examining the data for the South - the number of cases decreased from 2,052,437 victims in 1996 to 709,878 in 2011. During the COVID-19 pandemic, the number of victims decreased overall. There was an increase in crime across all regions in 2013, which continued until the artificial decrease due to lockdowns from 2019 to 2020. Unfortunately, for each region, the number of crimes has increased above the reported numbers before the COVID-19 pandemic.

```
# Example of Crime Trends divided by Type over time:
sns.lineplot(data=NCVS_RegionSeg, x='rpt_dt', y='vic_num',
↪  hue='crime_type')
plt.title("NCVS: Number of Crime Victims per Crime_TYPE between
↪  1996 and 2022")
```

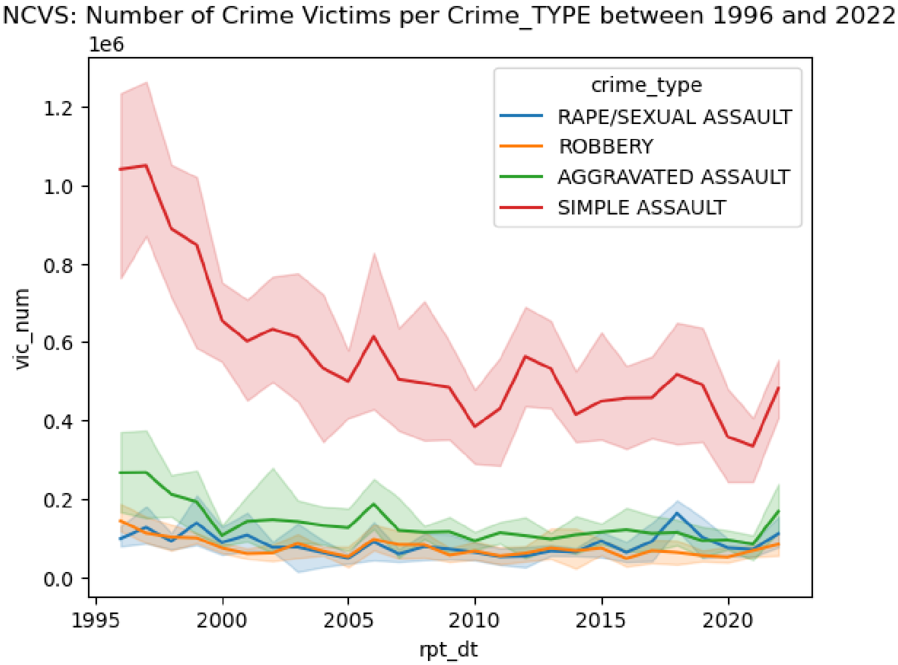


Fig. 7. *An examination of trends for each type of crime committed over the period as reported regionally by female victims. The x-axis represents report data, and the y-axis represents the number of victims* Created by Alexandra Coffin.

In addition to the increase in the number of crimes with female victims regionally, certain aspects of crime have changed. Simple assault has continued to be the dominant crime type, but there was an increase in reports of Aggravated Assaults and Rape/Sexual Assaults demonstrated in fig. 7. Despite the negative trend, there is concern that violent crime may return to levels that were higher than before the pandemic.

Unfortunately, examining different crimes across age groups is slightly more complex. Unlike the NYPD, which uses six age groups, the NCVS has eight groups. As the NCVS is a survey, children are unable to participate, and data is only collected on victims as young as 12, limiting the scope in terms of crimes with children as victims. The NCVS places victims into these Age Groups: 12-14, 15-17, 18-20, 21-24, 25-34, 35-49, 50-64, and 65+. The total number of victims that participated in the age portion of the survey was 9.60e^6.

The Age Group with the largest number of victims of simple assault crimes is between the ages of 25-34 - it is also the group that experienced the highest number of violent crimes as well. The second group that has the largest number of victims of violent crimes is the 35-49 Age Group, which also has the second highest number of victims of simple assaults. In terms of violent crimes that exclude simple assault, the Age Group most likely to experience Aggravated Assault was between 35-49, according to the NCVS. The age group with the highest number of larceny victims was the 25-34 group see fig. 9. The data involving Sexual Assaults was lacking in the age groups of minors and those over 65 as well as the 50-64 group.

```
# Violin plot to clearly visualize the amount of Simple Assaults
↪  versus other types of Violent Crime (See fig. 8):
sns.violinplot(data=NCVS_AgeType, x='AGE_GROUP', y='VIC_NUM',
↪  hue='CRIME_TYPE')
plt.title("NCVS: Simple Assault Victims vs. Violent Crime Victims
↪  Based on Age Group")
```

Due to the nature of Sexual Assaults and the stigmas surrounding reporting these incidents, the data gathered was closer to the 85% Confidence Interval. It has been included in the analysis to establish that there are cases of Sexual Assault for those years, as the data set is condensed. Based on what information the NCVS obtained, two groups have a high number of sexual assault victims, 25-34 and 34-49 noted in fig. 9.

```
# Scatter to demonstrate Crime Types and their occurrences
↪  regionally (See fig. 9):
sns.set(rc={"figure.figsize": (12,12)})
sns.scatterplot(
    data=NCVS_RegionSeg, x="rpt_dt", y="vic_num",
    ↪  hue="crime_type",
    size="vic_num", sizes=(20,200))
plt.title("NCVS: Number of victims reported between 1995 to 2023
↪  from Regional Data")
```

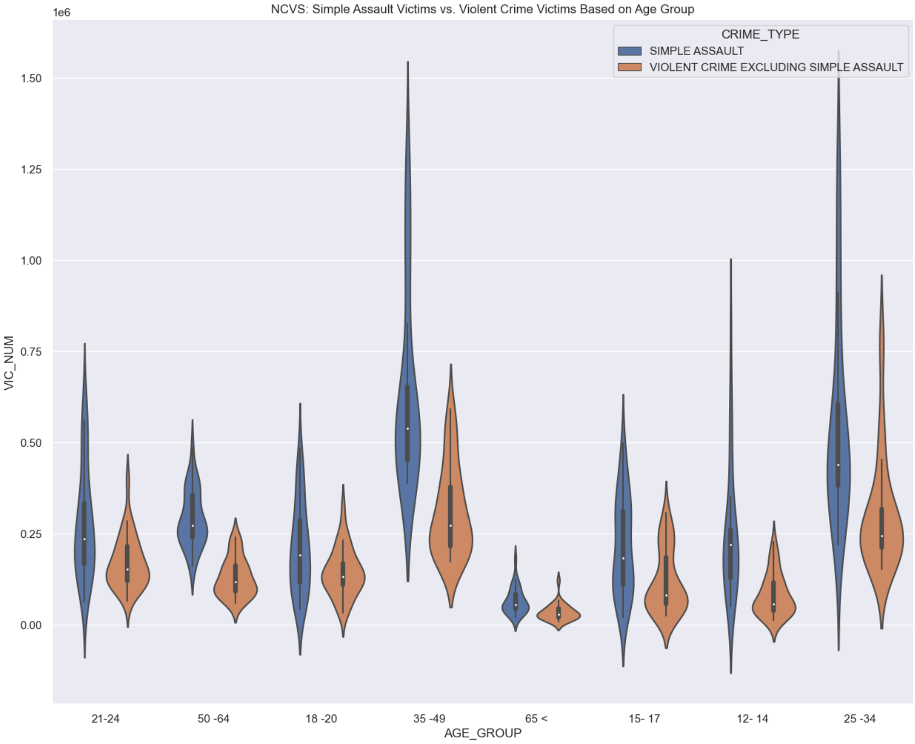


Fig. 8. *A violin plot examining the relationship between age and crime type.* Created by Alexandra Coffin.

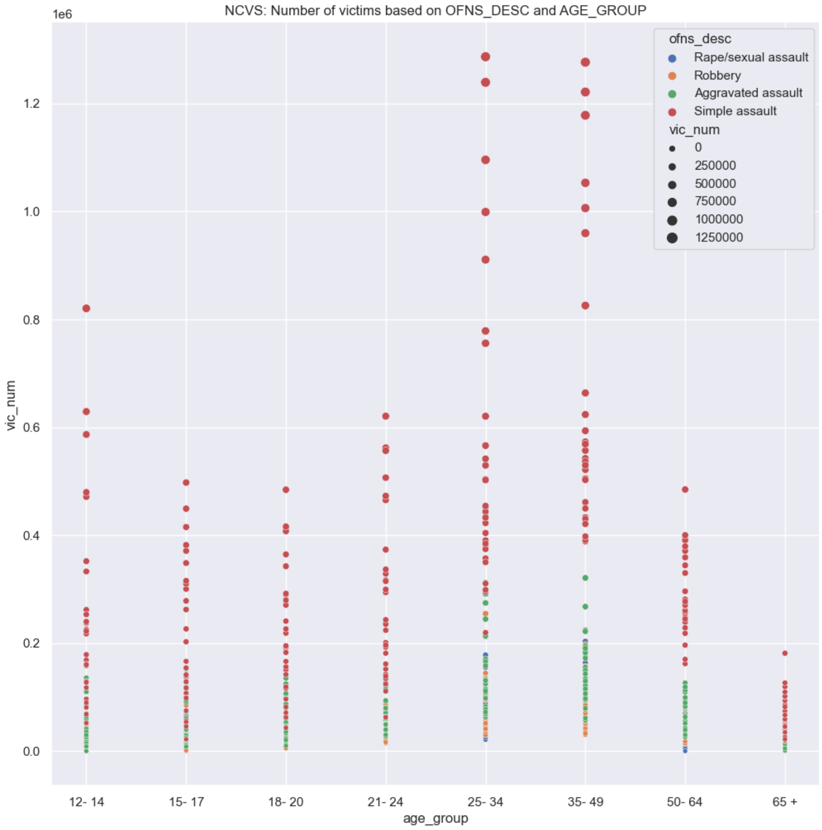
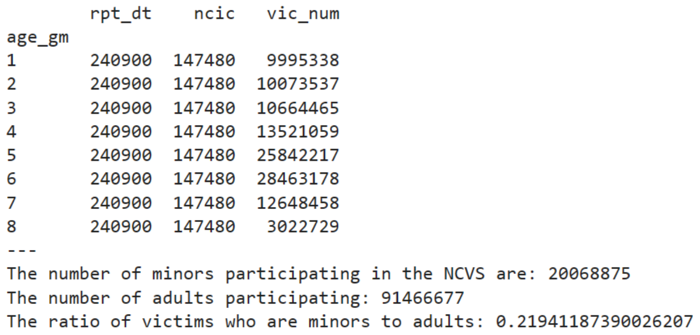


Fig. 9. *Scatter plot examining the relationship between age groups and a more segmented data set about types of crime.* Created by Alexandra Coffin.

When examining data gathered on minors, the group with the highest number of reported simple assaults is actually between the ages of 12-14. In violent crimes involving minors, the age groups most affected are between 15-17 years old. The total number of victims who are considered minors that participated in the survey is 20,068,875, which is 29.94% of participants fig. 10.

```
#Determining the number of victims per age group that participated
↪  in the NCVS
NCVS_AgeSegGroup = NCVS_AgeSeg.groupby(['age_gm']).sum()
print(NCVS_AgeSegGroup)
print("---")
num_minors=9995338 + 10073537
print("The number of minors participating in the NCVS are:",
↪  num_minors)
num_adults = 10664465 + 13521059 + 13521059 + 28463178 + 12648458
↪  + 12648458
print("The number of adults participating:", num_adults)
ratio_adulttom = num_minors/num_adults
print("The ratio of victims who are minors to adults:",
↪  ratio_adulttom)
```



```
        rpt_dt    ncic   vic_num
age_gm
1       240900  147480   9995338
2       240900  147480  10073537
3       240900  147480  10664465
4       240900  147480  13521059
5       240900  147480  25842217
6       240900  147480  28463178
7       240900  147480  12648458
8       240900  147480   3022729
---
The number of minors participating in the NCVS are: 20068875
The number of adults participating: 91466677
The ratio of victims who are minors to adults: 0.21941187390026207
```

Fig. 10. *Output of code to determine the ratio of adults to minors using NCVS_AgeSeg.* Created by Alexandra Coffin.

4.4 Conclusion of Exploratory Data Analysis

Between the two data sets, there are some shared characteristics, as well as differences. For example, the most common crime type reported was a simple assault, which in New York State is considered Harassment 2. Additionally, women between the ages of 25 - 44 were most likely to experience both a simple assault or other violent crimes. When examining the amount of crime, specifically the number of cases reported with female victims, the Regional Data varied greatly from the NYPD data. The NYPD recorded fewer cases, except for the increase between 2022 and 2023. Another difference is that when comparing the amount of data collected involving minors. Based on the NCVS data collected, 21.94% of crimes reported involved a minor as a victim. Victims that are classified as minors made up 6.67% of crimes reported.

5 Predictive Modeling:

The project is a foray into developing a predictive algorithm to assist in crime prevention. This is accomplished by comparing two different clustering models: K-means and DBSCAN. Data from the NCVS and the NYPD were fitted to each model to explore the variables of age, location, and type of crime.

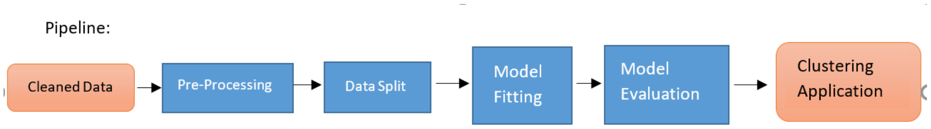


Fig. 11. *Pipelines to be applied to the data as the models are created and executed.* Created by Alexandra Coffin.

5.1 Pre-Processing

Crime data is categorical and is often reduced to an understanding of categories. Several adjustments have been made to ensure categorical uniformity between the two sets. In comparison, there are fewer entries within the NCVS data than in the NYPD. The data is also simplified from the various codes noted in the NYPD to two main classes of crime: ”Simple Assault” and ”Violent Crime Excluding Simple Assault.” When analyzing in-depth data provided focused on Personal crimes, specifically, simple assault, Assault, Robbery, and Sexual Assault. The limited number of codes influences the behavior of the algorithm during clustering. The NYPD data contains 55 Key Codes, increasing variation within the dataset. The result is that the clustering models from the NYPD data were run twice on two different variations of the data set. The first is the filtered data set containing crimes with female victims between 1994 and 2023, which was also utilized for exploratory analysis. The second variation of the data set was coordinated to decrease the dispersion, which resulted in decreased standard deviation.

Several Types of crime have been removed which do not fall within the scope of the NCVS data. There were also two columns added, NCIC and CT_M. NCIC, or the National Crime Information Center, created code numbers corresponding to specific offenses. These are different from NIRBS codes as they are entirely numeric. NIBRS codes are alphanumeric and are treated as strings, which many models can not process. As the NCVS data pulled from the N-Dash related to personal crimes, crimes such as traffic, child abandonment, drug possession, and disorderly conduct have been removed from the data set for this section. The CT_M column refers to whether or not the crime is considered one of two crime types classified by the NCVS: ”Simple Assault” = 1 or ”Violent Crime Excluding Simple Assault” = 2 (table 3). When sorting the data, most crimes with the KY_CD 361 were Aggravated Harassment, considered another form of Harassment 2 according to the New York State. It was placed in the same category as simple assault.

Two major crime types were removed, Codes 359 and 355. Code 359 was removed as it is used to note Violations of Order of Protection, Criminal Contempt, and Resisting Arrest. This data does not meet the objective of the model for Personal Crimes. Whereas 355 focuses on Custodial Issues, Unlawful Imprisonment 2, Reckless Endangerment, and Custodial Interference.

Table 3. *KY_CD Group and CT_M Assignments for NYPD Data Set* Created by A. Coffin.

Key Code Grouping Modifications

NCIC Number	KY_CD Grouped	OFNS_DESCs	CT_M
1101	104, 115, 116, 235	Sex Crimes	2
1201	105, 107, 109-113, 313, 340-343	Robbery/Fraud	2
1301	101, 103, 106, 114	Homicide/Aggravated Assault	2
1313	344, 578, 230, 355, 361	simple assault/Related	1

Age groups are also bundled differently between the two sets, and to avoid misinterpretation of Classes of the age groups, see table 4. Each of these modifications was made to the following files and added to the Repository as an export from the SQL database: NCVS_AgeSegML.csv, NCVS_RegionSegML.csv, NYPD_AgeSegML.csv, and NYPDv4ML.csv. The difference between the NYPDv4ML file and the NYPD_AgeSegML file is that the latter was formatted as an annual summary.

```
#Importing NCVS Data:
NCVS_AgeSeg = pd.read_csv('Data/ML_PreProcess/NCVS_AgeSegML.csv',
↪  index_col= 'rpt_dt')
NCVS_Region =
↪  pd.read_csv('Data/ML_PreProcess/NCVS_RegionSegML.csv',
↪  index_col='rpt_dt')

# Import NYPD Data:
NYPD = pd.read_csv('Data/ML_PreProcess/NYPDv4ML.csv', index_col=
↪  'rpt_num')
NYPD_AgeSeg = pd.read_csv('Data/ML_PreProcess/NYPD_AgeSegML.csv')
NYPD_AgeSeg.drop(['age_group'], axis=1, inplace=(True))
```

Since both models selected are unsupervised, scaling had to occur to prevent the data from becoming skewed by different measures.

```
# Dropping rows with NA values in any columns
NCVS_AgeSeg.dropna(inplace=True)

# Creating a scaled df where each value has a mean of 0 and stdev
↪  of 1
from sklearn import preprocessing
scaler = StandardScaler()
NCVS_AgeSeg[["age_gm_T", "ncic_T", "vic_num_T"]] =
↪  scaler.fit_transform(NCVS_AgeSeg[["age_gm", "ncic",
↪  "vic_num"]])
```

Table 4. *NCVS Data Age Group Modifications* Created by A. Coffin.

NCVS Age Modifications

New AGE_GM	NCVS AGE_GROUPS combined	New AGE_GROUP
1	12-14, 15-17	18
2	18-20, 21-24	18-24
3	25-34, 35-49	25-49
4	50-64	50-64
5	65+	65+

5.2 K-Means

K-means are often used to examine crime patterns, especially those across large populations in different areas. K-means functions to evaluate data based on the distance of a point from a calculated center[8]. It explores the association between data points based on distance calculations between groups. The model operates under the assumption that the cluster is spherical, equally in size, and has similar densities. K-means has limitations, as it performs poorly when clusters are irregular in size, shape, and density. As the algorithm itself is based on distance calculations from a centroid, it can be sensitive to the initial placement of a cluster and even interpret outliers with considerable impact[8].

For this project, K-means was applied in several ways. The first was constructing an analysis based on crime type and age, and the second was a region and crime type for all four data sets. Before creating each model, a function to Optimize K-means was formulated and resulted in an Elbow Plot. This plot examines the inertia or Sum of Squared Error of the set fig. 12. If Inertia decreases, the clusters become less effective, resulting in over-fitting [8]. The number of clusters found through this analysis was applied to the initial model.

```
# Estimating the number of clusters using KMeans:
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
%matplotlib inline

def optimise_k_means(data, max_k):
    means = []
    inertias = []

    for k in range(1, max_k):
        kmeans=KMeans(n_clusters=k)
        kmeans.fit(data)

        means.append(k)
        inertias.append(kmeans.inertia_)

    #Generate the plot
    fig = plt.subplots(figsize=(10, 5))
    plt.plot(means, inertias, '|-')
    plt.title('NCVS AgeSeg Elbow Method')
    plt.xlabel('Number of Clusters')
    plt.ylabel('Inertia')
    plt.grid(True)
    plt.show()
```

After each iteration of the models based on the results of the inertia graphs, several subplots were created to explore the effect that multiple iterations of k had on each data set, based on the specific attributes selected.

```
# Splitting using different K values
for k in range(1, 6):
    kmeans=KMeans(n_clusters=k)
    kmeans.fit(NCVS_AgeSeg[['age_gm_T', 'vic_num_T']])
    NCVS_AgeSeg[f'KMeans_{k}'] = kmeans.labels_
```

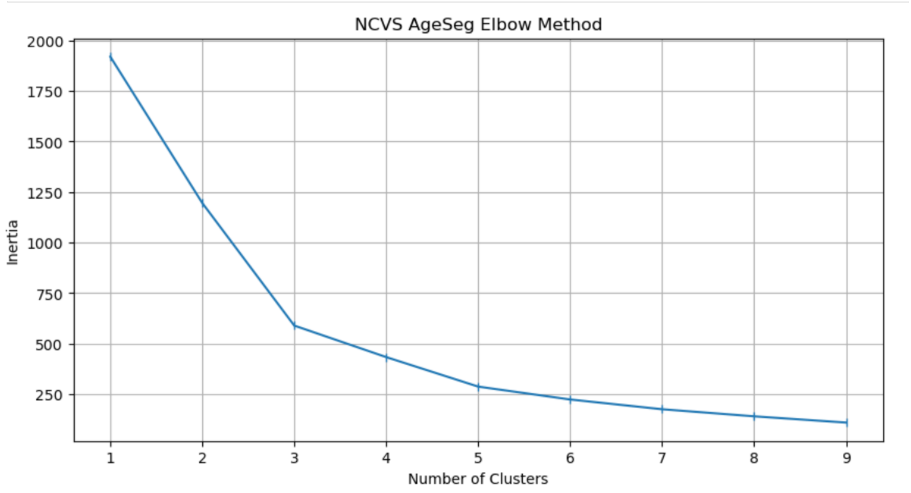


Fig. 12. *Sample of the script created to determine the number of clusters using inertia.* Created by Alexandra Coffin.

Crime data is multidimensional, and evaluating performance with the number of vectors increasing appropriately addressed the issue of assigning centroids. Models were adjusted using kmeans++ when determining the number of clusters resulting from a second model deployed to produce a silhouette.

```
# Establishing x by selecting the columns 5 through 7 holding
↪  transformed data
x = NCVS_Region.iloc[:, 3:5].values
y = NCVS_Region.iloc[:, 5]


model = KMeans(n_clusters = 4, init='k-means++', max_iter=300,
↪  n_init=10, random_state=0)
y_means = model.fit_predict(x)
labels = model.labels_
```

Finally, the results of the models were validated through the use of the Calinski-Harabasz Index (CHI), Davies-Bouldin Index (BDI), and the average Silhouette Coefficient (SS).

```
# Calinski-Harabase Index:
ch_score = calinski_harabasz_score(x, labels)

# Davies Boudlin
db_score = davies_bouldin_score(x, labels)

#Silhouette Score:
silhouette_score_average = silhouette_score(x, model.predict(x))

print(f"The Calinski-Harabasz Score for the NCVS_Region set is
↪  {ch_score}.")
print(f"The Davies-Bouldin Score for the NCVS_Region set is:
↪  {db_score}")
```

NCVS Data: Based on multiple sub-plots, the optimal number of k was selected and then used within a silhouette analysis to determine the performance of a model on predicted data. After the analysis, the optimal k for each NCVS data set is NCVS_AgeSeg = 4 and NCVS_Region = 3. After comparing the initial results with a silhouette score, the number of clusters was increased to address the complexity of the data. Each cluster is associated with a vector that is k dimensions, increasing the number of clusters decreases the dimensions of the data. However, the NCVS_Region set consistently required the same number of clusters.

A Silhouette Analysis (fig. 13) determined four clusters for NCVS_AgeSeg was optimal with an average of 0.521 SS. Cluster 1 for this set doesn't perform as strongly, but it does contain data above the Average Silhouette Score. The number of clusters for this particular analysis was performed optimally with 4 clusters. This aligns with the general split of the data because a large concentration of instances fell within the first cluster. The results of the NCVS_Region data performed better, with all four bars containing data above the average 0.567 SS. The number of clusters used in the NCVS_Region model coincides with the number assigned.

NYPD: The NYPD data, both the NYPD_AgeSeg and NYPDv4ML performed poorly. The initial model used $k = 3$, which generated a lot of overlap between the clusters. This iteration also had the lowest Silhouette Average Coefficient. The overlap combined with a 0.2817 SS indicated inaccuracy in clustering and a possible presence of multiple subsets within the set. This was made apparent when the Silhouette data was graphed. Even though each blade was thick, the negative coefficient and the overlapping of some clusters resulted in improper clustering. When the model was hyper-tuned using $k = 5$, the result was a larger number of integers that fell outside of clusters and 0.277 SS. Additionally, the CHI increased as k increased, with 5 clusters providing the best CHI.

When examining the NYPD_AgeSeg data, 0.475 SS. The use of $k = 3$ produced three clusters of equal width, indicating that the sorted silhouette coefficients fell consistently within the cluster.

```
#Silhouette Score
# Establishing x and y
x = NCVS_AgeSeg.iloc[:,3:5].values


model = KMeans(n_clusters = 4, init='k-means++', max_iter=300,
↪  n_init=10, random_state=0)
y_means = model.fit_predict(x)
labels = model.labels_

# Sihouette score
silhouette_score_average = silhouette_score(x, model.predict(x))
print(silhouette_score_average)

# Creating a visual of Silhouette:
visualizer = SilhouetteVisualizer(model, colors='yellowbrick')
visualizer.fit(x)
visualizer.show()
```

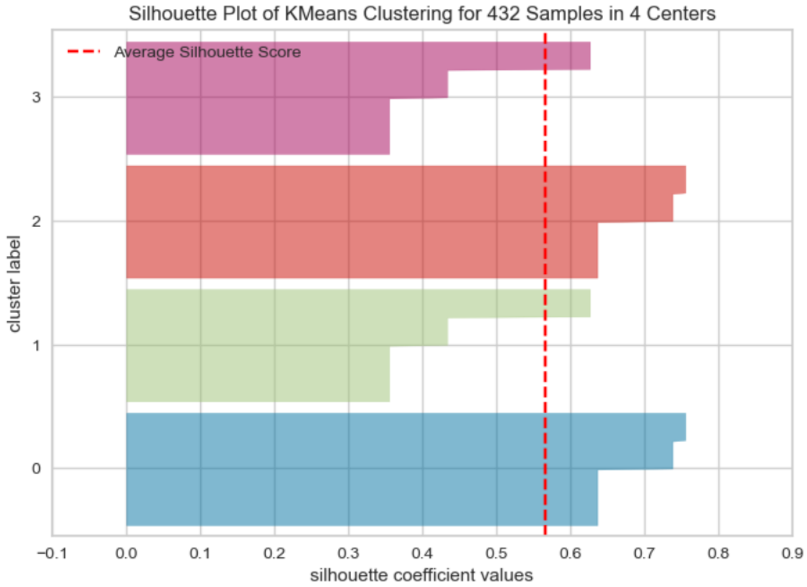


Fig. 13. *Script to Create Silhouette analysis and resulting graph for NCVS_AgeSeg Data.* Created by Alexandra Coffin.

5.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) applies a vector array to find core samples of high density and then generate clusters from these cores [3]. DBSCAN emphasizes density, making it ideal in plots containing geospatial data. To create a DBSCAN model, several factors are taken into account. The first is epsilon or the maximum distance between two samples for one to be considered a neighbor[8]. The next is min_samples, the total weight in a neighborhood for a point to be considered a core value. If min_samples is set inaccurately, it can result in density variation across clusters and create discrepancies in the data. Creating labels_ is pivotal with this model because each label dictates a point position within the plot. For example, labels assigned to outliers or noise have a value of -1 once the algorithm is executed. DBSCAN is not deterministic, so the collection of labels is crucial to analysis as the algorithm may change when run.

```
def get_scores_and_labels(combinations, X):
    scores = []
    all_labels_list = []

    for i, (eps, num_samples) in enumerate(combinations):
        dbscan_cluster_model = DBSCAN(eps=eps,
        ↪   min_samples=num_samples).fit(X)
        labels = dbscan_cluster_model.labels_
        labels_set = set(labels)
        num_clusters = len(labels_set)
        if -1 in labels_set:
            num_clusters -= 1
        if (num_clusters < 2) or (num_clusters > 50):
            scores.append(-10)
            all_labels_list.append('bad')
            c = (eps, num_samples)
            print(f"Combination {c} on iteration {i+1} of {N} has
            ↪   {num_clusters} clusters.")
            continue

        scores.append(ss(X, labels))
        all_labels_list.append(labels)
        print(f"Index: {i}, Score: {scores[-1]}, Labels:
        ↪   {all_labels_list[-1]}, NumCluster{num_clusters}")

    best_index = np.argmax(scores)
    best_parameters = combinations[best_index]
    best_labels = all_labels_list[best_index]
    best_score = scores[best_index]
    return{'best_epsilon': best_parameters[0],
           'best_min_samples': best_parameters[1],
           'best_labels': best_labels,
           'best_score': best_score}

best_dict = get_scores_and_labels(combinations, X)
```

NCVS: The NCVS Data for DBSCAN to be applicable. There was a sample run to demonstrate the effect that the size of a data set has on DBSCAN. The best score was a -10, as there wasn't enough data to process.

NYPD: As DBSCAN requires a large amount of memory, the data for the NYPD data set was filtered to contain three months' worth or 32,349 instances. This was done through a filter that selected only data from February 2023 to June 2023. Several iterations were completed to determine the epsilon and number of samples with a grid search. Values for epsilon were generated as a range between 0.01 and 1 with 15 instances. Whereas the min_samples were created using a range with step=3. The grid was to run through the generated lists for epsilon and samples, generating 90 combinations. A function was applied, which completed two tasks. The first enumerated different combinations until it could settle on a combination that produced clusters above -1, and under 50 that scored poorly were not selected. The second task was to create a list of scores and labels generated from these results. Based on the list of scores and labels, an index was devised containing the best parameters based on the function "get_scores_and_labels(combinations, X)".

The result for the geographical data was an epsilon = 0.01, min_samples = 17, best_score of 0.324, and an array of labels that was inserted into the Data Frame as another column. The model identified 65 outliers, a 0.02% error. One cluster had a higher density than the rest. When inspected based on geographical shape, the result was clustering in Brooklyn, Queens, and Manhattan. Based on these results, the area with the least amount of crime was Staten Island fig. 14

```
# A more detailed map
fig = px.density_mapbox(NYPD1, lat='lat', lon='lon', z='cluster',
↪   radius=5, center=dict(lat=NYPD1.lat.mean(),
↪   lon=NYPD1.lon.mean()), zoom=9, mapbox_style='open-street-map',
↪   height=900)
fig.show()
```

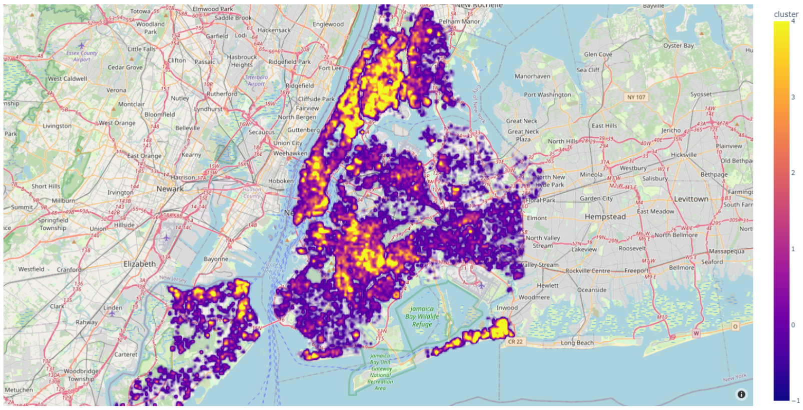


Fig. 14. *Geo-plot modeled with DBSCAN using geo-location data related to reports.* Created by Alexandra Coffin.

Another model was crafted to investigate the relationship between age and Precinct numbers as a comparison. Each column: 'addr_pct_cd' and 'age_gm', was scaled before being fed into the function created for the initial analysis (fig. 15. The resulting features were epsilon = 0.151, min_samples = 2, and 0.674 SS. This proved to have a higher performance than the K-means model when predicting clustering.

```
# Creating the graph comparing Precinct Address and Age_Gm:
fig = px.scatter(x=X2_scaled[:,0], y=X2_scaled[:,1],
↪  color=best_dict3['best_labels'])
fig.show()
```

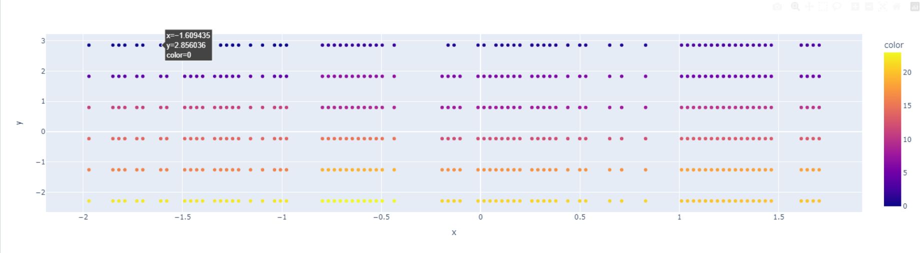


Fig. 15. *Scatter plot using DBSCAN comparing Precinct Number to Age Groups* Created by Alexandra Coffin.

6 Discussion:

The ability of a clustering algorithm to perform accurately depends on the size of the data scope and the number of instances found within the set. Smaller data sets similar to the NCVS_AgeSeg data performed the best using K-means, 0.629 SS. Each of the NCVS sets performed the best when examining clustering with K-means.

Table 5. *Table exploring Result Metrics for best-performing models.* Created by A. Coffin.

Performance Metrics

Data	Model	Calinski-Harabasz	Davies-Bouldin	Avrg. Silhouette Score
NCVS_AgeSeg	K-means k=4	1120.853	0.792	0.521
NCVS_Region	K-means k=4	623.390	0.737	0.567
NYPDv4ML	K-means k=3	30876.41	1.322	0.282
NYPDv4ML	K-means k=4	30876.410	1.322	0.281
NYPDv4ML	K-means k=5	30539.139	1.121	0.277
NYPD_AgeSeg	K-means k=3	197.687	0.814	0.475
NYPDv4ML Geo	DBSCAN e=0.01, min=17	3055.368	1.718	0.324
NYPDv4ML KY_Cd	DBSCAN e=0.1514, min=2	4842.86	23.62	0.6753
NCVS_AgeSeg	DBSCAN	NA	NA	NA

While the sizable NYPDv4ML set performed better utilizing a density-based model, DBSCAN. In terms of silhouette score average, the DBSCAN was only slightly better. When examining the clustering of the data, DBSCAN was more successful. When the model was applied to two other aspects of the data, specifically Precinct Numbers compared to Age Groups, the Average Silhouette Score increased to 0.674. Consequently, when discussing the scope of the data, simplified data sets that condense features perform better.

When processing the NYPD data, K-means required more clusters to decrease the dimensions of the data. Similar to grouping crimes based on the type of crime, when utilizing Key Codes, not all key codes are structured in numerical groups or based on the severity of the crime. This results in the improper clustering of data based on crime type. When these results were compared to NYPD_AgeSeg where the crimes were combined based on type, K-means was more accurate. Fewer clusters were required between the pair, resulting in on-target clustering in terms of the Calinski-Harabasz Index and the performance of the Davies-Bouldin. Between the pair, the NYPD_AgeSeg model had the best Silhouette Score, 0.475, indicating that most of the data will cluster within the selected clusters. The Silhouette Score and the 1.121 DBI demonstrated the K-means model performed better than the DBSCAN Model with a 1.718 DBI table 5.

The relationship between the physical attributes of victims and the types of crimes experienced is complex. When examining the relationship between the number of victims per age group, it leans toward younger women being at a higher risk. When comparing the density between the NYPD and NCVS_Age Seg data, there was a noticeable pattern of younger victims being more common. Crime overall was documented to have a higher number of victims across clusters 1 and 2 when comparing the NYPD data to the NCVS, indicating that there is a possible difference between regional and national data (fig. 16).

Especially when observing the geographical locations where crimes occur in New York City see fig. 14. Crime hot spots in New York City are primarily in Manhattan, the Bronx, and the upper side of Queens. The Bronx has the most substantial hot spot according to the DBSCAN model. Areas such as Staten Island have the lowest density of crime reports. When comparing this information to the map generated with K-means, the clustering was more extreme, resulting in a high concentration of crime across New York City.

When comparing the capabilities of K-means Clustering to DBSCAN in creating a density map to examine the amount of crime reported across NYC, DBSCAN performed the best in terms of Silhouette Score. It failed both the comparison of Calinski-Harabasz and Davies-Bouldin Index scores. The NYPDv4ML data set performed the best utilizing 5 clusters according to metrics (fig. 5), despite several instances becoming outliers.

The NCVS data set, when fit to a DBSCAN model, failed. As DBSCAN is a density-based algorithm, the lack of points within the same region resulted in the algorithm identifying each instance as noise. The points did cluster when calculating K-means using a series of 4 clusters. However, there was a lack of clustering in terms of density. The data from both NCVS data sets produced scores that fell around -10 for the silhouette score, even with an epsilon of 0.01. The NCVS data applied to the K-means Models resulted in statistically more stable results. The NCVS_Region set applied to K-means clustering produced four uniform blades in Silhouette Visualization, with a 0.567 SS and a DBI of 0.737. The Davies-Boulding Score is high for the regional data, falling closer to one, but is still lower than the results generated for the NYPD_AgeSeg set using K-means where $k = 4$.

```
# Bar chart comparing 4=k on data
# Creating the model to analyze the data: Examining Date and key
↪  code relationship
kmeans= KMeans(n_clusters=4)

# fitting data to model based on crime code and age
kmeans.fit(NYPD[['ky_cd_T', 'age_gm_T']])

# Establishing kmeans column
NYPD['kmeans_4'] = kmeans.labels_
NYPD
ax1 = plt.subplot2grid((1,1), (0,0))
NYPD['kmeans_4'].plot(kind='hist', ax=ax1, color='b', alpha=0.4)
```

```
#NYPD['kmeans_4'].plot(kind='kde', ax=ax1, secondary_y=True,
↪  label='distribution', color='g', lw=2)
NCVS_AgeSeg['KMeans_4'].plot(kind='kde', ax=ax1, secondary_y=True,
↪  label='distribution', color='r', lw=2)
sns.displot(NYPD['kmeans_4'], color='b')
```

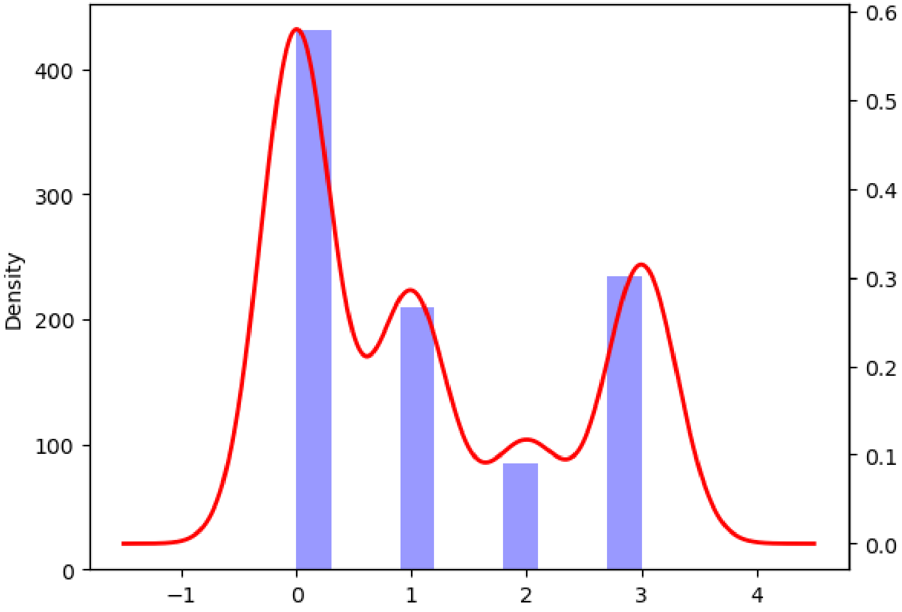


Fig. 16. *A bar chart comparing K-means 4 between NYPD and NCVS_AgeSeg data. The red line represents the NCVS_AgeSeg data, and the blue bars indicate the NYPD results.* Created by Alexandra Coffin.

6.1 Limitations

There have been various methods applied to collecting data on victims, including the use of the National Incident-Based Reporting System (NIBRS), the National Crime Victimization Survey (NCVS), and the Uniform Crime Reports (UCR). As of January 1, 2021, the UCR has been officially folded into the NIBRS to obtain more involvement among Agencies and centralize data on both criminals and victims [7]. These limitations also extend into the nature of crime, and many crimes go unreported.

As this data has been cleaned of personal information, it is impossible to determine if a victim has experienced repeat victimization. Repeat victimization is "the repeated occurrence of a crime involving either the same victim or the same location"[4]. This concept can be extended into series victimization, where separate victims can describe the same or similar events to an interviewer [4]. Series victimization is only frequent with some times of crime; the repetition of the act does create a secondary instance within the data set and is used to analyze victim and criminal behavior.

Clustering models are powerful, especially when examining crime data, as they are ideal for identifying complex patterns. Machine learning models currently can not interpret external conditions such as environmental context, demographics, economics, unemployment, and seasonality, which are external variables when examining crime[12]. Unsupervised methods have several limitations when analyzing data as layered as victim data.

K-means, when applied to noisy data, can produce irregular sizes and different densities, both of which are not optimal for this form of analysis. [9] K-means also requires a specification of the number of clusters before running the model, in addition to doing so multiple times to improve accuracy. Crime data is complex and often incomplete, forming a lot of noise that is difficult to clear entirely from the data. This is addressed by adding another layer to the analysis, DBSCAN.

DBSCAN does have some limitations, in this case, as we are dealing with crime data where it is possible to have points that belong to more than one cluster - DBSCAN can become confused and will not cluster when there are massive differences in density as it relies on distances [9]. This results in the model not being traditionally used on larger data sets as it doesn't scale well when there are regions with low density around clusters.

6.2 Future Work

Crime changes based on geographical region. Iterative learning offers a new opportunity to create a model trained on National Data and fit localized data to a model. Forging a baseline provides scope to a model when examining State or City data. Algorithms such as CURE and mini-batch K-means. Both of these methods allow for similar types of data to be fitted to a model. There is a connection between age, sex, crime type, and location - an investigation further into linear relationships within the data could provide an explanation involving socioeconomic variables such as income or population densities to observe shifts in crime rates.

7 Conclusion

The most successful method was K-means because the NCVS data could not be run through DBSCAN as its Silhouette Score and label combination were poorly performing. DBSCAN provided a different insight into the actual density of crime within the Boroughs, but in terms of comparing age groups, the clustering accuracy was less. When examining the relationship between age and female victims, the data demonstrated that there was a connection between age and crime. This relationship between the NYPD and the NCVS yielded a negative trend, indicating that younger individuals are at a higher risk. Future work is required to create a model capable of clustering and predicting national data for research into crime types. Crime is complex, but through machine learning, it is possible to discover patterns facilitating a deeper understanding of the attributes in female victimology and even into aspects of criminology.

References

1. Coffin, A.: Crimevictimanalysis_capstone, https://github.com/accoffin12/CrimeVictimAnalysis_Capstone/tree/main
2. Commision, N.P.: Nyc population finder, https://popfactfinder.planning.nyc.gov/#10.67/40.7198/-73.9515
3. scikit-learn developers: sklearn.cluster.dbscan, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn-cluster-dbscan
4. Doerner, W.G., Lab, S.P.: Measuring Criminal Victimization. Routledge (2017)
5. Doerner, W.G., Lab, S.P.: The Scope of Victimology. Routledge (2017)
6. FBI: Crime data explorer, https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/crime-trend
7. FBI: National incident-based reporting system (nibrs), https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr/nibrs
8. Geron, A.: Unsupervised Learning Techniques. O'Reilly (2017)
9. Geron, A.: Unsupervised Learning Techniques. O'Reilly (2023)
10. of Justice Statistics, B.: Custom graphics: Multi-year trends: Crime type, https://ncvs.bjs.ojp.gov/multi-year-trends/crimeType
11. of Justice Statistics, B.: Data collection: National crime victimization survey (ncvs), https://www.bjs.gov/index.cfm/content/pub/ascii/content/data/index.cfm?ty=dcdetail&iid=245
12. Kang, H.W., Kang, H.B.: Prediction of crime occurrence from multi-modal data using deep learning. PLOS ONE **12**(4), e0176244 (apr 2017). https://doi.org/10.1371/journal.pone.0176244, https://doi.org/10.1371%2Fjournal.pone.0176244
13. Kiprop, V.: The boroughs of new york city – nyc boroughs map, https://www.worldatlas.com/articles/the-boroughs-of-new-york-city.html
14. NYPD: Nypd complaint data historic, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
15. Petruzzello, M.: Staten island, https://www.britannica.com/place/Staten-Island