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 KMeans and DBSCAN, on data collected from the New York Police Department (NYPD) and the National Crime Victim Survey (NCVS). KMeans was selected as the first algorithm as it clusters data based on the distance from a centroid, which is a common method for identifying patterns within crime data. DBSCAN clusters data based on density calculations, serving as a way 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 pertaining to victims and instances that did not contain age, geolocation, sex, and crime type.

Keywords: data analytics · criminal justice · female victimology · machine learning

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 [4]. Examining the groups considered the most vulnerable 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 [5].

The number of cases reported to the FBI and the National Incident-Based Reporting System(NIBRS) is immense. This project will use a combination of data collected from two sources the National Crime Victim Survey[10] and the

New York Police Department [12]. Each data set was selected based on 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 NYPD Data, 3.2 NCVS Data, section 3.3 Conclusion of Exploratory Analysis. Section 4 Predictive Modeling, 4.1 Pre-Processing, 4.2 KMeans, 4.3 DBSCAN. Section 5 Discussion, 5.1 Limitations, and 5.2 Future Work based on research findings.

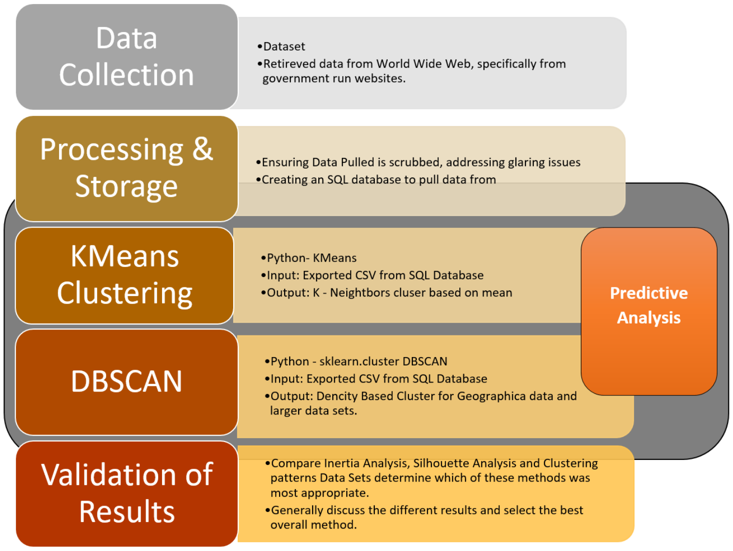

Data Collection
•Dataset
•Retireved data from World Wide Web, specifically from government run websites.
Processing & Storage
•Ensuring Data Pulled is scrubbed, addressing glaring issues
•Creating an SQL database to pull data from
KMeans Clustering
•Python- KMeans
•Input: Exported CSV from SQL Database
•Output: K - Neightbors cluser based on mean
DBSCAN
•Python - sklearn.cluster DBSCAN
•Input: Exported CSV from SQL Database
•Output: Dencity Based Cluster for Geographica data and larger data sets.
Predictive Analysis
Validation of Results
•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) Initial 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[10] from the Bureau of Justice Statistics and the NYPD Open Data Portal[12] as static files in CSV format.

The NCVS data will serve as a baseline comparison and has the least amount of statistical variance as it has been cleaned previously by the Bureau of Justice. 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 [9].

The data attributes utilized are Sex, Age, Race, Location, zip code/County, Offense Code, and NIBRS Code. Women will be the focus of this analysis as they are statistically more likely to report a crime than men [3]. 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[12]	8.3M	2.7M	35	17	1922-2023
NCVS[9]	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.

The full list of Attributes is as follows:

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 finally the use of a firearm. These variables when combined allow for a better understanding of who crimes affect. In combination with regional statistics, the results will differ from national statistics. This project analyzes the relationship between sex, age, and location to the types of crimes committed. By creating a series of models utilizing the NCVS data as the training set, and then comparing it to models training with the regional data it is possible to determine if an issue is regional or national. Additionally, through

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	NYPD
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

the applications of these models, it presents a point of comparison as to groups of women at the greatest risk geographically and socially.

3.2 Cleaning, Date Range & Storage

The data was loaded into Tableau Prep Initially to handle the larger adjustments. All data sets were modified in terms of uniform age groups and race features. Nulls are common in victim data as many police forces will remove victim information for individual safety. An entry was kept as long as three out of three of the five requirements, except for homicides. These requirements included sex, location, age, race, and type of crime. Homicides are the exception to this as in some cases location will be omitted.

An additional column has been added to the NYPD data set to translate AGE_Groups into a usable format for modeling. Each of the Age Groups 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: ¿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 and then combined to create a more extensive data set through the use of unions to create 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 groups 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 [3]. Both groups are considered vulnerable based on physical strength, financial dependence or limited financial power, geographical risk, and finally the 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. The only data set that was used with flagged data was the table detailing women filtered by race under the category other as the feature by nature is broad so variance is expected.

All of the data being 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 that remain in the data are those for AGE_GROUP and LOCATION data in the form of a dash and parentheses. The rest of the special characters have been addressed by a combination of replacement and the removal of attributes utilizing these. LOCATION data has been transformed into the order of SQL, additionally, columns were added to store this data separately.

Within the NCVS data, several columns were added as a way 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 of the age ranges, Crime Type encoded with associated NCIC numbers, and Region which was given 1 through 4. This coding is covered in detail within the repository associated with this project.

All final details are listed in the Source List Table1, this includes the final number of entries for each cleaned data set as well as the number of final attributes. The data was then loaded into an SQL database to allow for rapid queries and data selection. The use of quires was utilized to pull from larger sets and demonstrated 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 structure of the data is revealed as well as possible outliers that can be discovered while determining the nature of the data. Through the generation of visuals, we can develop a sense of story, and decide on the data that has the strongest correlations. As this project requires the use of multiple sources, the NYPD[12] and the NCVS[9] the data analysis has been broken into a section for each. There are two major 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 the 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 NYPD Data

For this analysis, all of the NYPD Data[12] has been selected between 1993 and 2023, except a scatter plot depicting the number of crimes reported per day. A majority 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, it is important to note that a majority of this analysis utilizes data manipulated from the initial pull, except the heatmap which utilizes a separate sheet coordinating the correlation of age and crime. The original data for this project was pulled from: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

To coordinate this data as each case is listed in an individual line the following code was utilized to group the victims based on a a specific attribute. The following is code that was developed 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)
```

The result was the following table, a brief overview of the different types of crimes committed in the set. The repository contains a longer list which is more extensive. 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. While there are a large number of crimes that fall within the Public category of Criminal Activity as defined by the NIRBS codes, 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 top five most common crimes to have female victims are HARASSMENT 2 with 26,291 cases, ASSAULT 3 & RELATED OFFENSES with 14,706 cases, PETIT LARCENY with 12,641 cases, OFFENSE AGAINST PUB ORDER SENSIBILITY with 6,009 cases, and FELONY ASSAULT with 5,511 cases[1]. 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 also important to note that there are several codes used to define crimes that have a sexual component, such as 104 (Rape) and 116 (Sex Crimes). Sex Crimes is a broader category that is broken into several facets including aggravated sexual assault, sexual abuse, and trafficking.

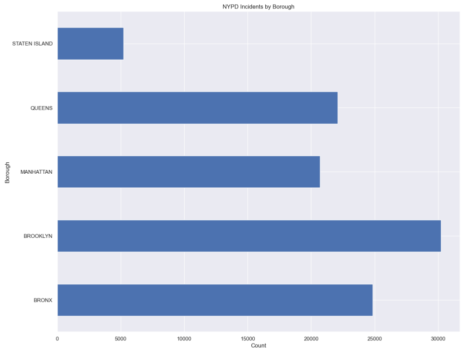


Fig. 2. *A bar graph depicting the relationship between the number of victims and the location of the crime.* Created by Alexandra Coffin.

When examining the data in terms of the Borough where the offense had occurred there appeared to be a difference between each of the five. Staten Island had the fewest female victims, with just over 5,000, whereas Brooklyn had one of the highest numbers with just over 30,000 female victims as noted in Fig. 2.

The plot below examines the most recent shift in crime, from 2021 to 2023. For the scatter plot, the years 2019 and 2021 were not included because of Covid-19. During the pandemic, New York State went into lockdown, which resulted in an alteration to 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 did result in a decrease in reported crimes as there was a removal of risk-taking behaviors and social conditions. The result is a massive increase in the number of crimes reported between 2022 and 2023, fig. 3. This is attributed to several factors including social interaction, commutes and working in offices, criminal organization activity increasing due to restrictions being lifted, and social movements.

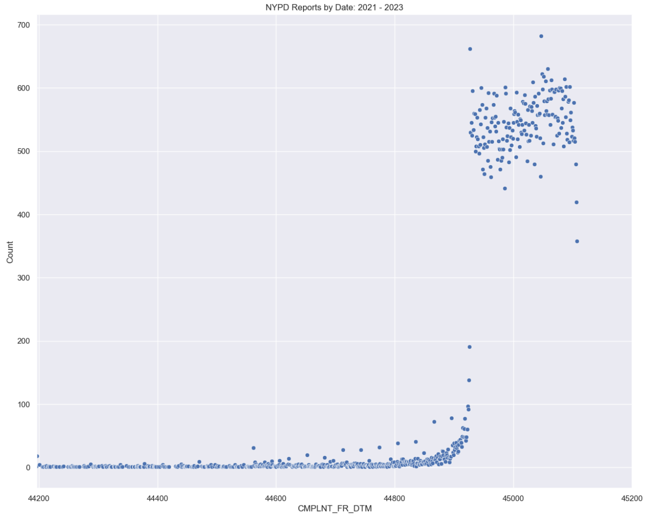


Fig. 3. *Scatter Plot depicting clustering of crime reports from 2021 to 2023.* Created by Alexandra Coffin.

Examining the relationship between age and crime has been the subject of many studies. Especially when discussing victimology, as there is a correlation between age and the types of crimes committed. For the data that has been selected from the NYPD, there is a strong correlation between victims and crime in terms of age groups, however, the association between the type of crime and victims isn't as strong. Which results in a unique correlation coefficient that is slightly negative. Women between 25-44 years old (Group 3) with 51,774 cases had the highest number, which translated to roughly 49.42% of cases reported. The ratio of minors to adults is roughly 6.67%, with a total of 6447 cases involving minors.

The code used to generate the heatmap is as follows:

```
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")
```

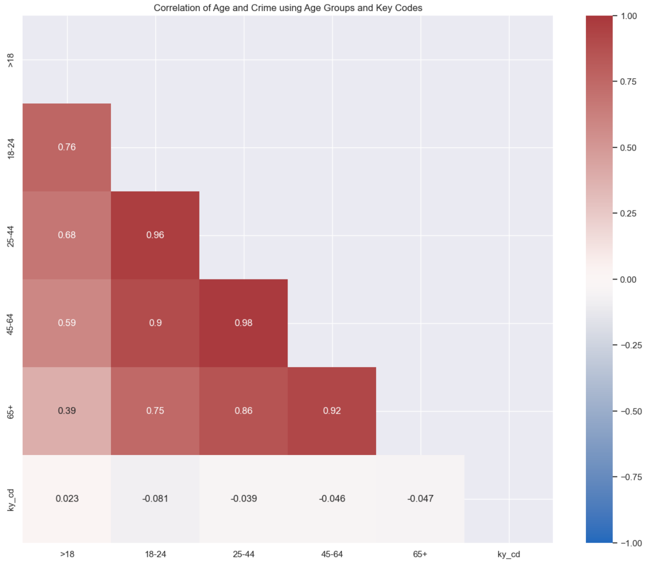


Fig. 4. *A heatmap examining the relationship between victim age and the type of crime they were subjected to.* Created by Alexandra Coffin.

4.2 NCVS Data

The NCVS Data[9] consisted of several files which were condensed into several files. These files revolve around the Type of Crime, Region, and Age Group for women who participated in these studies. It is important to note that data regarding age does not reflect the ages of women in the regions surveyed as it is not possible to 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[1].

Between 1996 and 2022 there has been a significant reduction in the amount of crime occurring in the United States. This is especially evident when examining the data for the South as the number of cases decreased from 2,052,437 victims in 1996 to 709,878 in 2011. During the period of the pandemic, there was a decreased number of victims overall reported. 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 overall has increased above the reported numbers before the pandemic in 2018.

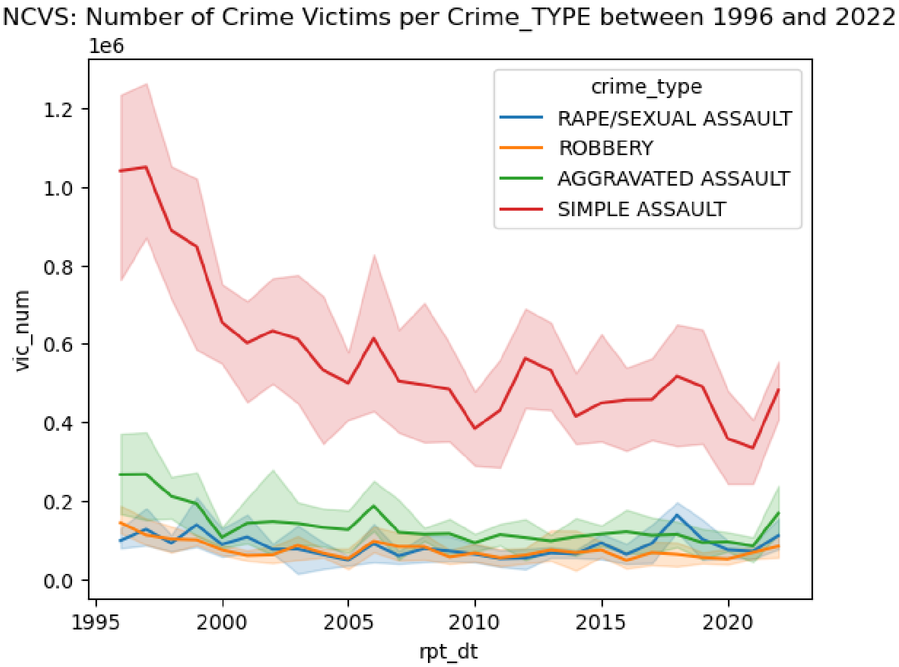


Fig. 5. *An examination of trends for each type of crime committed over the period as reported regionally by female victims.* Created by Alexandra Coffin.

In addition to the increase in the number of crimes with female victims regionally, the types of crimes being committed have also changed slightly. While the most common crime of Simple Assault has continued to be dominant. The was also an increased report of Aggrivated Assaults and Rape/Sexual Assault demonstrated in fig. 5. The trend of the data is negative as a whole, however, the concern is that for many crimes there has been an increase to levels that are higher than pre-pandemic levels, especially with more violent types of crime.

Unfortunately, the data when examining different crimes across age groups is slightly more complex. Unlike the NYPD which uses six age groups, the NCVS has a series of eight age groups. As the NCVS is a survey legally 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 the following 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 $9.60e^6$.

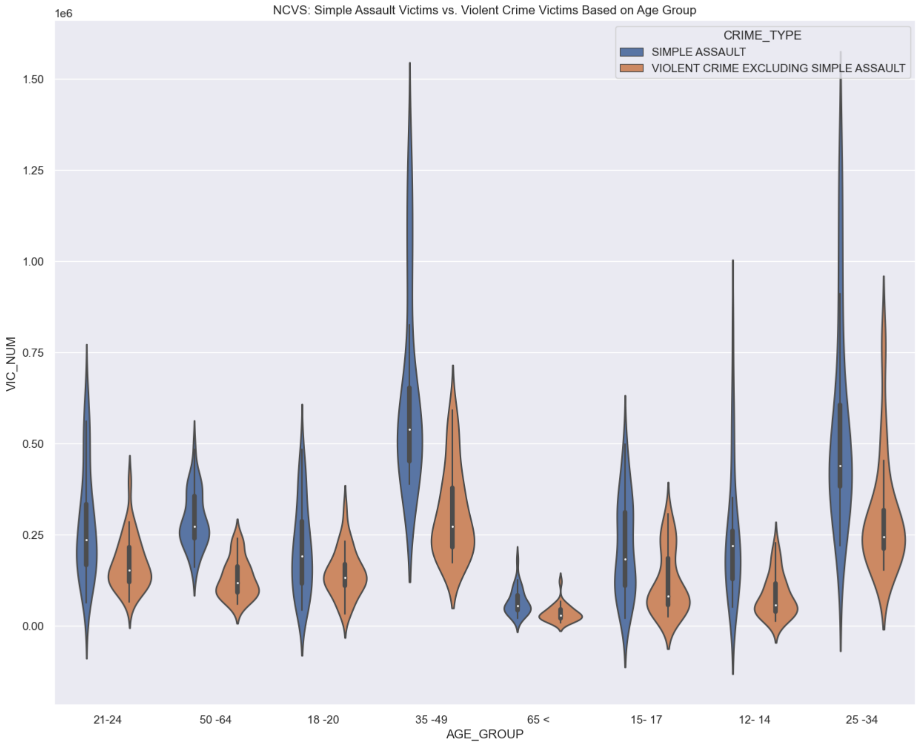


Fig. 6. *A violin plot examining the relationship between age and crime type.* Created by Alexandra Coffin.

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 year-olds see fig. 7. The data involving Sexual Assaults was lacking in the age groups of minors and those over 65 as well as the 50-64-year-old block.

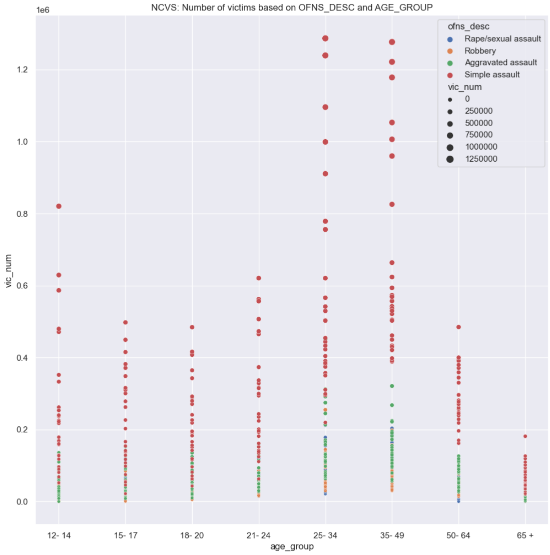


Fig. 7. *Scatter Plot examining the relationship between age groups and a more segmented data set about types of crime.* Created by Alexandra Coffin.

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 dataset is condensed. Based on what information the NCVS did obtain on the subject two groups have a high number of sexual assault victims, 25-34 year-olds and 34 to 49 year-olds noted in fig.7.

When examining data gathered on minors, the group with the highest number of reported simple assaults is actually between the ages of 12-14. 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 total participants.

4.3 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 amount the crime reported was a simple assault which in New York State is considered Harassment 2. Additionally, women between the ages of 25 to 44 years old 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 did vary greatly from the NYPD data. The NYPD reported a lower number of cases, as a whole except for the increase between 2022 to 2023. Another difference is that when comparing the amount of data collected involving minors. Out of 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 through the comparison of two different clustering models: KMeans and DBSCAN. Data from the NCVS and the NYPD were fitted to each of these models to explore the variables of age, location, and type of crime.

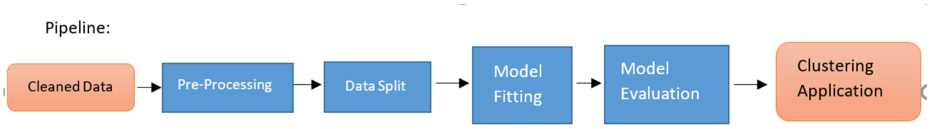


Fig. 8. *Pipelines to be applied to the data as the models are created and executed.* Created by Alexandra Coffin.

5.1 Pre-Processing

Crime data by nature is categorical, in the sense that while it is possible to count the number of victims and offenders, the data is often reduced to an understanding of categories. As such several adjustments have been made to ensure that there is uniformity between the two sets. In comparison to the size of the data set, there are fewer entries within the NCVS data than in the NYPD. The data is also simplified from the various codes that can be noted in the NYPD to two main classes of crime: Simple Assault and Violent Crime Excluding Simple Assault. When analyzing the data more in-depth the data provided focuses on Personal crimes, specifically Simple Assault, Assault, Robbery, and Sexual Assault. The limited number of codes does influence the behavior of the algorithm during clustering. The NYPD data contains 55 different 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 dataset 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 National Crime Information Center created code numbers that correspond to specific offences. These are different from NIRBS codes as they are entirely numeric. NIBRS codes are alphanumeric and as such are treated as strings, which many models can not process. As the NCVS data pulled from the N-Dash is 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 (fig.3). When sorting the data most of the crimes with the KY_CD 361 were Aggravated Harassment, which is considered another form of Harassment 2 according to the New York State. As such it was placed in the same category as Simple Assault.

Two major crime types were removed 359 and 355. Code 359 was removed as it is used to note Violations of Order of Protection, Criminal Contempt, and Resisting Arrest. While this data is important, it 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.

NCVSAgeMod

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/Aggrivate Assault	2
1313	344, 578, 230, 355, 361	Simple Assault/Related	1

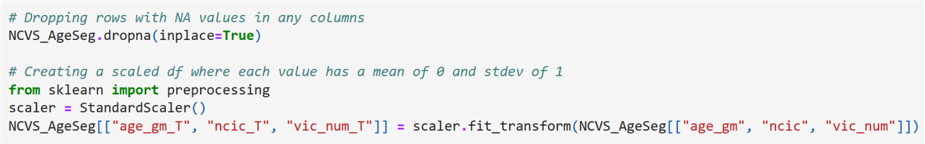
Age groups are also bundled differently between the two sets, and to avoid misinterpretation of classes the age groups see fig.4.

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

NCVSAgeMod

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+

After each of these modifications had been made the following files were 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. As both of the models selected are unsupervised, scaling had to occur in order to prevent the data from becoming skewed by different measures. For example, the coding used for age groups as compared to the number of victims reported.



```
# 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"]])
```

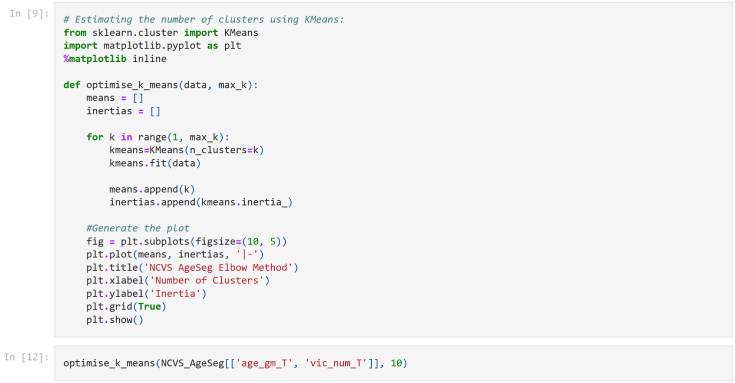
Fig. 9. *Script designed to scale data to ensure unit integrity.* Created by Alexandra Coffin.

5.2 KMeans

KMeans are often used when attempting to understand crime patterns, especially those across large populations in different areas. By nature, KMeans functions to evaluate data based on the distance of a point from a calculated center[7]. By doing so it explores the association between data points based on distance calculations between groups. The model operates under the assumption that the cluster is spherical in nature, 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 greater impact than necessary[7].

For this project KMeans 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. Prior to creating each model, a function was created to Optimize Kmeans and generate an Elbow Plot. This plot examines the inertia or Sum of Squared Error of the data set fig.10. The reason is that as inertia decreases clusters become less effective and result in over-fitting [7]. The number of clusters found through the use of this analysis was applied to the initial model.

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 fig.11. As crime data is multidimensional, the approach of evaluating performance with the number of vectors increasing appropriately addressed the issue of assigning centroids.



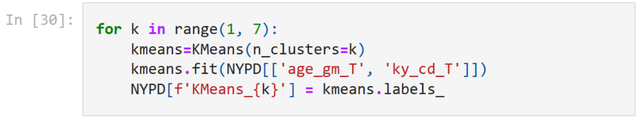
```
In [9]:  # Estimating the number of clusters using KMeans:
         from sklearn.cluster import KMeans
         import matplotlib.pyplot as plt
         %matplotlib inline

         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()
```

```
In [12]: optimise_k_means(NCVS_AgeSeg[['age_gm_T', 'vic_num_T']], 10)
```

Fig. 10. *Sample of the script created to determine the number of clusters using inertia.* Created by Alexandra Coffin.

Additionally, models were adjusted through the use of kmeans++ when determining the number of clusters resulting from a second model that was run to produce a silhouette. 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).

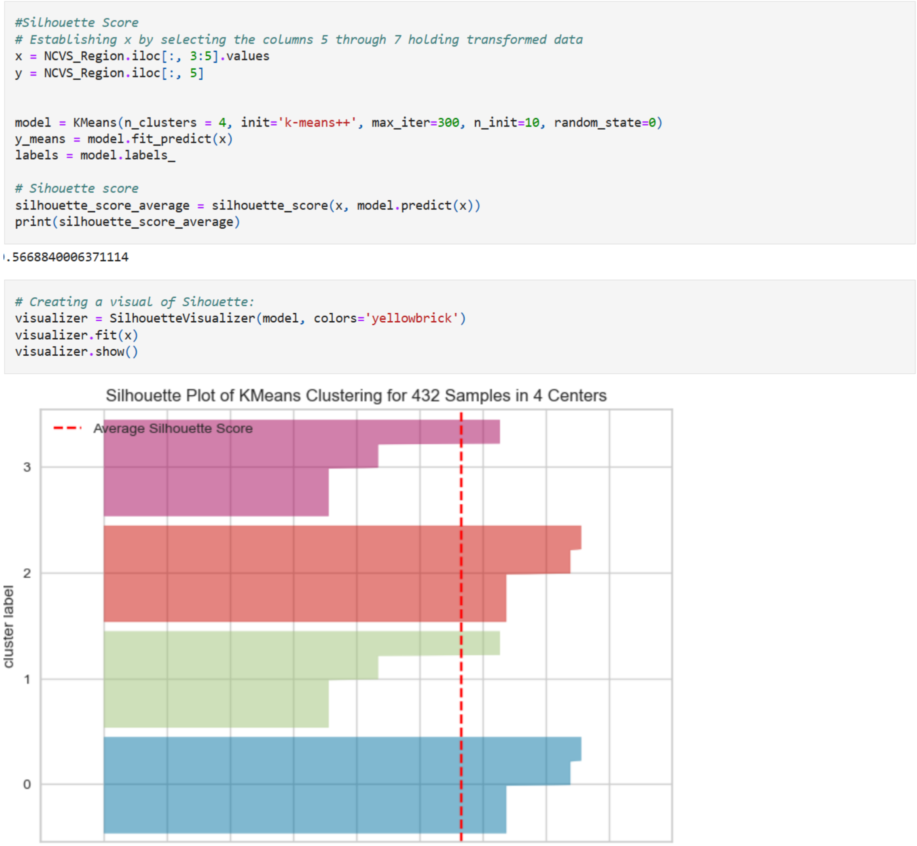
NCVS Data: Based on these subplots the optimal number of k was selected and then used within a silhouette analysis to determine the performance of a model on predicted data. The optimal number of clusters at the conclusion of the analysis for each NCVS data set is NCVS_AgeSeg = 4 and NCVS_Region = 3. After comparing the initial results with a silhouette score, the number of clusters was increased to address the complexity of the data. As each of the clusters is associated with a vector that is k dimensions, the addition of the cluster decreased the dimensional of the data. However, in the case of the NCVS_Regional data the number of clusters remained the same.



```
In [30]:  for k in range(1, 7):
              kmeans=KMeans(n_clusters=k)
              kmeans.fit(NYPD[['age_gm_T', 'ky_cd_T']])
              NYPD[f'KMeans_{k}'] = kmeans.labels_
```

Fig. 11. *Sample of the script created to demonstrate clustering using different k values.* Created by Alexandra Coffin.

The NCVS_Region data that did the best with this model performed the best as a whole when comparing predictions using a silhouette analysis fig.12. Completion of the Silhouette Analysis determined that using four clusters for NCVS_AgeSeg was optimal with an average of 0.521 SS. Cluster 1 for this set doesn't perform as strongly as the remaining 3, it does contain data that is above the average silhouette score. The number of clusters for this particular analysis performed at optimal with 4 clusters. This aligns with the general split of the data as a majority of the 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. When clustering was selected the number of clusters used in the NCVS_Region model coincides with numbers that the regions were assigned.



```
#Silhouette Score
# 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_

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

```
.5668840006371114
```

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

Silhouette Plot of KMeans Clustering for 432 Samples in 4 Centers
Average Silhouette Score
cluster label
3
2
1
0

Fig. 12. *Script to Create Silhouette analysis and resulting graph for NCVS_AgeSeg Data.* Created by Alexandra Coffin.

NYPD: The NYPD data, both the NYPD_AgeSeg and NYPDv4ML had performed poorly when this method was applied. When initially applying the suggested 3 clusters there was a lot of possible overlap between each cluster, in addition to the lowest Silhouette coefficient. The overlap in combination with an 0.2817 SS indicated inaccuracy in clustering and a possible presence of a large number of subsets within the set. This was made apparent when the Silhouette data was graphed. Even though each of the blades is thick in nature, it's the negative coefficient as well as the overlapping of some clusters that creates a large amount of error. When the model was run again using 5 clusters, 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 three clusters did produce three clusters of equal width, an indication that the sorted silhouette coefficients fell consistently within the cluster.

5.3 DBSCAN

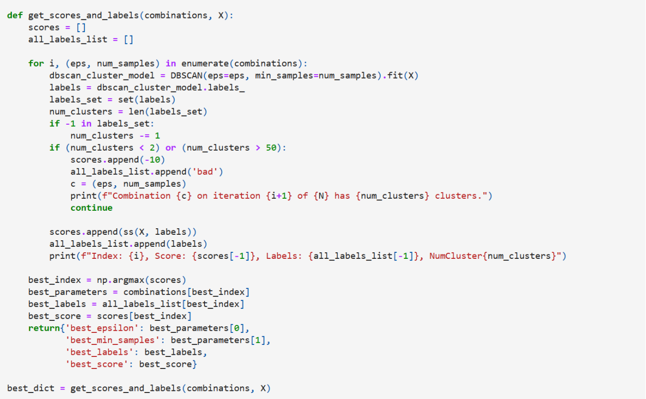
DBSCAN or Density-Based Spatial Clustering of Applications with Noise utilizes a vector array to find core samples of high density and then generate clusters from these cores [2]. As DBSCAN emphasizes density, it is ideal to create plots utilizing geospatial data. In order to create a DBSCAN model several factors are taken into account. The first is epsilon or the maximum distance between two samples for one to be considered a neighbor[7]. The next is min_samples, which is the total weight in a neighborhood for a point to be considered a core value. If min_samples is set incorrectly it can result in density variation across clusters and create discrepancies in the data. Finally understanding labels_ is critical as labels that are considered outliers or noise are given a label 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.

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 that it was able to produce 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 roughly 3 months' worth of data or 32,349 instances. This was done through a filter process that selected only the instances from February 2023 to June 2023. For this method, several iterations were completed to determine the epsilon and number of samples through the use of 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, which generated a total of 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 that were above -1, and under 50 clusters that scored poorly were not selected. The second task was to create a list of scores and labels that were generated from these results. Finally, an index was generated, allowing for the selection of the best parameters based on the function fig.13.

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 DataFrame as another column. The model did identify 65 outliers, a 0.02% error when compared to the rest of the sample. There was one cluster that was denser than the rest, however, when this is examined based on geographical shape, the result was clustering in Brooklyn, Queens, and Manhattan. In terms of density, the area with the least amount of crime was actually Staten Island fig. 14

A second analysis was performed to investigate the relationship between age and geographical location using precinct numbers to compare the data. Each of the columns, 'addr_pct_cd' and 'age_gm' were scaled prior to being fed into the function that had been created for the initial analysis. The resulting features were epsilon = 0.151, min_samples = 2, and 0.674 SS. This proved to have a higher performance than the KMeans model when predicting clustering.



```python
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)
```

Fig. 13. *Function utilized to determine clusters, labels, min_sample and silhouette score for the best model in DBSCAN.* Created by Alexandra Coffin.

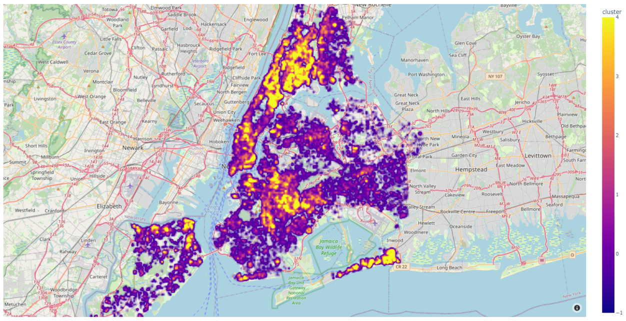


Fig. 14. *Geo-plot modeled with DBSCAN using geo-location data related to reports.* Created by Alexandra Coffin.

6 Discussion:

The ability of a clustering algorithm to perform accurately on data is dependent both on the size of the data scope as well as number of instances found within the set. Smaller data sets such as the NCVS_AgeSeg data performed the best using KMeans, 0.629 SS. Each of the NCVS data sets performed the best when examining clustering with KMeans, especially when the data was plotted based on the number of victims found per set within the period.

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	KMeans k=4	1120.853	0.792	0.521
NCVS_Region	KMeans k=4	623.390	0.737	0.567
NYPDv4ML	KMeans k=3	30876.41	1.322	0.282
NYPDv4ML	KMeans k=4	30876.410	1.322	0.281
NYPDv4ML	KMeans k=5	30539.139	1.121	0.277
NYPD_AgeSeg	KMeans 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 larger NYPDv4ML set performed better utilizing a density-based model, DBSCAN. In terms of silhouette score average, the DBSCAN was only slightly better. However, when examining the clustering of the data, DBSCAN was more successful. When examining two other aspects of the data, specifically precinct numbers compared to age groups the silhouette score increased to 0.674. Consequently, when discussing the scope of the data, smaller data sets that condense features resulting in simplified data perform better.

When processing the NYPD data, KMeans resulted in creating a larger number of clusters to decrease the dimensions of the data. While similar to the process of 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 the NYPD_AgeSeg where the crimes were combined based on type, KMeans was more accurate. A fewer number of clusters was required between the pair, resulting in more accurate clustering both in terms of the Calinski-Harabasz Index and the overall 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. When this was combined with the 1.121 DBI, it indicated that the KMeans model performed better than the DBSCAN Model with a 1.718 DBI fig. 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, the relationship 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. **??**).

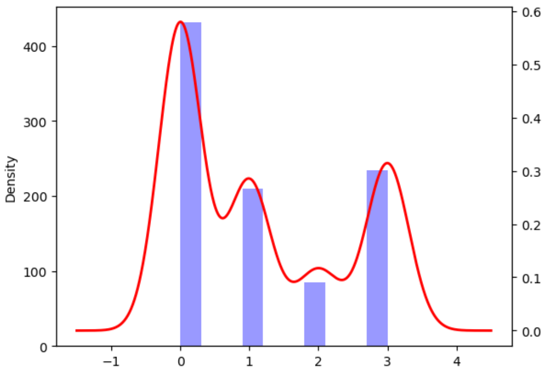


Fig. 15. *A bar chart comparing KMeans 4 between NYPD and NCVS_AgeSeg data. The red line represents the NCVS_AgeSeg data, and the blue bars indicate the NYPD results.* Created by Alexandra Coffin.

Especially when observing the geographical locations where crimes occur in New York City fig. 14. Crime hot spots in New York City are primarily in Manhattan, the Bronx, and the upper side of Queens. The borough with the largest hot spot is actually the Bronx. While areas such as Staten Island have the lowest density of crime reports. When comparing this information to the map generated through the use of KMeans, the clustering was more extreme resulting in a high concentration of crime across Manhattan as a whole.

When comparing the capabilities of KMeans Clustering to DBSCAN in creating a density map to examine the amount of crime reported across DBSCAN performed the best in terms of silhouette score, but failed in both the comparison of Calinski-Harabasz and Davies-Bouldin Index scores. The NYPDv4ML data set performed the best utilizing 5 clusters according to metrics5, 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 failing as each of the points was considered noise. While the points did cluster when calculating KMeans using a series of 4 clusters, 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. However, when the NCVS data was applied to the KMeans Models, the results were statistically more stable and accurate. The Regional NCVS data when run through KMeans clustering produced four uniform blades in Silhouette Visualization, with a Silhouette Average Score of 0.567 and a DBI of 0.737. While the Davies-Boulding Score is high for the regional data, falling closer to one, is still lower than the results generated for the NYPD_AgeSeg set using KMeans where k = 4.

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 in an attempt to obtain more involvement among agencies as well as centralize data on both criminals and victims [6]. However, 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 defined as the repeated occurrence of a crime involving either the same victim or the same location [3]. This concept can be extended into series victimization, where separate victims are capable of describing the same or similar events to an interviewer [3]. Series victimization is only common among some times of crime, however, the repetition of the act does create a secondary instance within the data set, that can be 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[11]. Unsupervised methods have several limitations when analyzing data as layered as victim data.

KMeans, when applied to noisy data, can produce irregular sizes and different densities, both of which are not optimal for this form of analysis. [8] 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. As crime data is complex and often incomplete, there is 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 [8]. This results in the model not being traditionally used on larger datasets as it doesn't scale well when there are regions with low density around clusters.

6.2 Future Work

Crime is complex, and even 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. In a sense using a larger baseline to provide scope to a model when examining State or City data. Algorithms such as CURE and mini-batch KMeans. Both of these methods allow for similar types of data to be fitted to a model. While 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 overall most successful method was the use of KMeans, the NCVS data could not be run through DBSCAN as its silhouette score and label combination were poorly performing. DBSCAN did provide 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 both 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 the use of machine learning its possible to discovered patterns allowing for 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. scikit-learn developers: sklearn.cluster.dbscan, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn-cluster-dbscan
3. Doerner, W.G., Lab, S.P.: Measuring Criminal Victimization. Routledge (2017)
4. Doerner, W.G., Lab, S.P.: The Scope of Victimology. Routledge (2017)
5. FBI: Crime data explorer, https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/crime-trend
6. FBI: National incident-based reporting system (nibrs), https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr/nibrs
7. Geron, A.: Unsupervised Learning Techniques. O'Reilly (2017)
8. Geron, A.: Unsupervised Learning Techniques. O'Reilly (2023)
9. of Justice Statistics, B.: Custom graphics: Multi-year trends: Crime type, https://ncvs.bjs.ojp.gov/multi-year-trends/crimeType

10. 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
11. Kang, H.W., Kang, H.B.: Prediction of crime occurrence from multimodal 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
12. NYPD: Nypd complaint data historic, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i