Proposal: The Dynamics of Sex Offenses and Rape Incidents In The U.S.

DATA 450 Capstone

Zainab Makhdum

February 8, 2024

1 Introduction

According to RAINN – the Rape, Abuse, and Incest National Network – every 68 seconds an American is sexually assaulted with one out of six American women and one in thirty-three American men being the victim of attempted or completed rape in their lifetime (Rape and Network, n.d.). Due to these jarring statistics, it is important to address this issue in order to combat sexual violence by spreading awareness as well as providing mediums of support to victims of this heinous crime. Therefore, in order to unearth nuanced insights for the complex dynamics in rape and sex offenses, the purpose of this research is to analyze how demographic, geographical, and interpersonal factors influence the varying patterns of rape incidents in the United States in 2022.

Firstly, the study aims to delve deeper into geographical disparities by examining rape incidents across states and cities, as well as specific locations such as home, office, storage facilities etc. Second, the research explores the deviations in rape rates overtime (2012-2022) on a national level. Additional temporal analysis includes analyzing specific timestamps during which the crime took place. Third, the breakdown of relational factors will be taken into consideration as relationships between the victims and the individuals are examined. Fourth, demographic information i.e., age, sex, and race will be utilized to examine the composition of the victims and the arrestees. Fifth, comparative analysis will be conducted to observe the disparities between the total number of victims with the total number of individuals arrested. Lastly, unsupervised learning in the form of clustering analysis will be utilized to identify distinct groups and patterns of sex offenders in New York state.

This study is conducted while keeping in mind the discrepancies in population densities, youth concentration, economic conditions, cultural factors, effectiveness of law enforcement agencies, crime reporting practices etc., as these factors vary from place to place. Furthermore, it is important to note that the legal definitions of 'rape' as well as the age of consent differ from

state to state. According to the FBI's National Incident Based Reporting System (NIBRS), sex offenses are defined as "any sexual act directed against another person, without the consent of the victim, including instances where the victim is incapable of giving consent," while rape is defined as "(except Statutory Rape) the carnal knowledge of a person, without the consent of the victim, including instances where the victim is unable to give consent because of his/her age or because of his/her temporary or permanent mental or physical incapacity" (Federal Bureau Of Investigation 2023b). Additionally, 2018 onwards, all sexual offenses – forcible and non-forcible – are categorized under the same category i.e., sex offenses (Federal Bureau Of Investigation 2023a).

All in all, forming data-driven insights is imperative in comprehending the pivotal aspects that influence rape rates in recent years across distinct communities in the United States to combat sexual violence through shaping policies, improving support mechanisms, and eliminating victim blaming and stigma around speaking out against the perpetrators.

2 Dataset

For the purpose of this research, the data is obtained from the Federal Bureau of Investigation (FBI) through their Crime Data Explorer online tool (Federal Bureau Of Investigation 1960-2022) in the Documents and Downloads section as well as queries run on the Data Discovery Tool (Federal Bureau Of Investigation 1960-2022). According to the FBI, the Uniform Crime Reporting Program is utilized to collect data – submitted by agencies through a state UCR program or directly to the FBI's UCR program – from more than 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies.

The data comprises of a total of 17 datasets, with 6 files in .csv format and 11 files in .xlsx format that are accessed through the Crime Data Explorer website and Data Discovery Tool. Details for the files/datasets i.e., file name and location, as well as relevant variables from each dataset that are to be used in the analysis are given below:

- **File Name:** Crimes_Against_Persons_Offenses_Offense_Category_by_State_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: State, Year: 2022, Folder Name: state)
  - Total rows: 57
  - Total columns: 9
  - Variables:
    - State: Full name of U.S. State
    - Sex Offenses: Total number of sex offenses by each U.S. state in 2022
- **File Name:** Relationship_of_Victims_to_Offenders_by_Offense_Category_2022
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Relationships, Year: 2022)

  - Total rows: 18
  - Total columns: 7
  - Variables:
    * Offense Category: Types of offence categories i.e., crimes.
    * Total Victims: Total number of victims for each offence category
    * Family Member: Victims were related to all offenders, whether one or more, of the reported offense.
    * Family Member and Other: Victims were related to at least one of the multiple offenders of the reported offense.
    * Known To Victim and Other: Victims knew, but were not related to, one or more of the multiple offenders of the reported offense.
    * Stranger: Victims did not knew and were not related to the offender.
    * All Other: Victims were mutual combatants (victim was offender) or had an unknown relationship with a single offender.

- **File Name:** Table_8_Offenses_Known_to_Law_Enforcement_by_State_by_City_2022.xlsx
  - Location: Crime in the United States Annual Reports (Collection: Offenses Known to Law Enforcement, Year: 2022, Folder Name: offenses-known-to-le-2022)
  - Total rows: 7,882
  - Total columns: 13
  - Variables:
    * State: Full name of U.S. State
    * City: Name of city located in a specific U.S. state
    * Population: Recorded population for each city in 2022
    * Rape: Instances of rape by each city in 2022

- **File Name:** Crimes_Against_Persons_Offenses_Offense_Category_by_Location_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Location, Year: 2022, Folder Name: location)
  - Total rows: 52
  - Total columns: 7
  - Variables:
    * Location: Specific location where crime took place.
    * Sex Offenses: Number of sexual offenses by each location in 2022.

- **File Name:** Victims_Age_by_Offense_Category_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Victims, Year: 2022, Folder Name: victims)
  - Total rows: 26
  - Total columns: 16
  - Variables:
    * Offense Category: Types of offence categories i.e., crimes.

        * 10 and Under: number of victims 10 and under
        * 11-15: number of victims between ages 11-15
        * 16-20: number of victims between ages 16-20
        * 21-25: number of victims between ages 21-25
        * 26-30: number of victims between ages 26-30
        * 31-35: number of victims between ages 31-35
        * 36-40: number of victims between ages 36-40
        * 41-45: number of victims between ages 41-45
        * 46-50: number of victims between ages 46-50
        * 51-55: number of victims between ages 51-55
        * 56-60: number of victims between ages 56-60
        * 61-65: number of victims between ages 61-65
        * 66 and Over: number of victims for ages 66 and over
        * Unknow Age: number of victims with unknown age

- **File Name:** Victims_Race_by_Offense_Category_2022.xlsx
    - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Victims, Year: 2022, Folder Name: victims)
    - Total rows: 26
    - Total columns: 8
    - Variables:
        * Offense Category: Types of offence categories i.e., crimes.
        * White: number of white victims
        * Black or African American: number of black or African American victims
        * American Indian or Alaska Natives: number of American Indian or Alaska Natives victims
        * Asian: Number of Asian victims
        * Native Hawaiian or Other Pacific Islander: number of Native Hawaiian or Other Pacific Islander victims
        * Unknown Race: Number of Victims with unknown race

- **File Name:** Victims_Sex_by_Offense_Category_2022.xlsx
    - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Victims, Year: 2022, Folder Name: victims)
    - Total rows: 26
    - Total columns: 5
    - Variables:
        * Offense Category: Types of offence categories i.e., crimes.
        * Male: Number of male victims.
        * Female: Number of female victims.
        * Unknown Sex: Number of victims with unknow sex.

- **File Name:** Arrestees_Age_by_Arrest_Offense_Category_2022.xlsx

  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Arrestees, Year: 2022, Folder Name: arrestees)
  - Total rows: 40
  - Total columns: 16
  - Variables:
    * Arrest Offense Category: Types of arrest offence categories i.e., crimes.
    * 10 and Under: number of victims 10 and under
    * 11-15: number of arrestees between ages 11-15
    * 16-20: number of arrestees between ages 16-20
    * 21-25: number of arrestees between ages 21-25
    * 26-30: number of arrestees between ages 26-30
    * 31-35: number of arrestees between ages 31-35
    * 36-40: number of arrestees between ages 36-40
    * 41-45: number of arrestees between ages 41-45
    * 46-50: number of arrestees between ages 46-50
    * 51-55: number of arrestees between ages 51-55
    * 56-60: number of arrestees between ages 56-60
    * 61-65: number of arrestees between ages 61-65
    * 66 and Over: number of arrestees for ages 66 and over
    * Unknow Age: number of arrestees with unknown age

- **File Name:** Arrestees *Race_by* Arrest_Offense _Category_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Arrestees, Year: 2022, Folder Name: arrestees)
  - Total rows: 40
  - Total columns: 8
  - Variables:
    * Arrest Offense Category: Types of arrest offence categories i.e., crimes.
    * White: number of white individuals arrested
    * Black or African American: number of black or African American individuals arrested
    * American Indian or Alaska Natives: number of American Indian or Alaska Natives individuals arrested
    * Asian: Number of Asian individuals arrested
    * Native Hawaiian or Other Pacific Islander: number of Native Hawaiian or Other Pacific Islander individuals arrested
    * Unknown Race: Number of arrestees with unknown race

- **File Name:** Arrestees_Sex_by_Arrest_Offense_Category_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Arrestees, Year: 2022, Folder Name: arrestees)
  - Total rows: 40

  - Total columns: 4
  - Variables:
    * Arrest Offense Category: Types of arrest offence categories i.e., crimes.
    * Male: Number of males arrested.
    * Female: Number of females arrested.

- **File Name:** Crimes_Against_Persons_Incidents_Offense_Category_by_Time_of_Day_2022.xlsx
  - Location: National Incident-Based Reporting System (NIBRS) Tables (Tables: Time of Day, Year: 2022, Folder Name: timeOfDay)
  - Total rows: 34
  - Total columns: 7
  - Variables:
    * Time of Day: Time rages (am and pm) during which the crime took place.
    * Sex Offenses: Number of sexual offenses by each timestamp.

- **File Name:** National_Rape_Ten_Year_Trend.csv
  - Location: Data Discovery Tool – National or State Query (From 2012 – To 2022, Crime: Rape, Location: United States, Format: Table)
  - Total rows: 5
  - Total columns: 12
  - Variables:
    * Years: 2012-2022
    * Query1: Total number of rape instances for each year across the nation.

- **File Name:** NY_Rape_Ten_Year_Trend.csv
  - Location: Data Discovery Tool – National or State Query (From 2012 – To 2022, Crime: Rape, Location: New York, Format: Table)
  - Total rows: 5
  - Total columns: 12
  - Variables:
    * Years: 2012-2022
    * Query1: Total number of rape instances for each year in New York State.

- **File Name:** NIBRS_OFFENSE_TYPE.csv
  - Location: Crime Incident-Based Data By State (State: New York, Year: 2022, File Name: NY-2022)
  - Total rows: 86
  - Total columns: 8
  - Variables:
    * offense_code: unique code associated with each type of crime.
    * offense_category_name: description of crime.

- **File Name:** NIBRS_OFFENDER.csv
  - Location: Crime Incident-Based Data By State (State: New York, Year: 2022, File Name: NY-2022)
  - Total rows: 161034
  - Total columns: 11
  - Variables:
    - offender_id: Unique identifier for offender.
    - incident_id: Identifier associated with a crime. One incident_id can be associated with multiple victims/offenders.
    - age_num: age of offenders.
    - sex_code: sex of offenders.
    - race_id: race of offenders.
- **File Name:** NIBRS_OFFENSE.csv
  - Location: Crime Incident-Based Data By State (State: New York, Year: 2022, File Name: NY-2022)
  - Total rows: 174259
  - Total columns: 8
  - Variables:
    - incident_id: Identifier associated with a crime. One incident_id can be associated with multiple victims/offenders.
    - offense_code: specific code associated with each type of offense.
    - attempt_complete_flag: status for whether a crime has be completed (C) or attempted (A).
    - location_id: specific location where crime took place.
- **File Name:** REF_RACE.csv
  - Location: Crime Incident-Based Data By State (State: New York, Year: 2022, File Name: NY-2022)
  - Total rows: 12
  - Total columns: 7
  - Variables:
    - race_id: unique ID associated with each race.
    - race_desc: description of each race e.g., White, Asian, African American etc.

3 Data Acquisition and Processing

For the majority of the files, the data was acquired by accessing them through the Crime Data Explorer online tool in the Documents and Downloads section. On this page, there are several sub-sections, however, only three of them were utilized for obtaining the data for this research:

3.1 Data Acquisition

- National Incident-Based Reporting Systems (NIBRS) Tables
  - For this section, data can be obtained by accessing the drop-down menu for tables and choosing a specific topic. Topics for this research in this section include Victims, Arrestees, State, Relationships, Location, Time of Day, and Completed and Attempted. For the drop-down menu for year, 2022 is selected due to the most recent data being available for this year. Once these options have been selected, the files are downloaded in xlsx format.
- Crime in the United States Annual Reports
  - In this section, the only collection from which data was extracted is Offenses Known To Law Enforcement for 2022. Similar to files from National Incident-Based Reporting Systems (NIBRS) Tables, this data folder can be easily downloaded through which relevant files in xlsx format can be accessed.
- rime Incident-Based Data By State
  - Data from this section is extracted by selecting a specific state – in our case it's New York – and the year i.e., 2022. Through this, a folder containing several xlsx files are downloaded from which only the four relevant ones are selected e.g., NIBRS Offenses, Offenders, Offence Type, and Race Reference.
- Data Discovery Tool
  - In this section, queries are run on the Data Discovery Tool page by selecting 'Crime Data' as the dataset and setting the query level to 'National or State.' Two queries are run on this page with the same year range i.e., 2012-2022 and same crime i.e., rape. However, the first query consists of selecting the location as 'United States' as we want to access rape incidents on a national level. Whereas the second query consists of selecting the location for a specific state i.e., New York as one of the sub-research questions consists of analyzing state-level data for NY and ultimately conducting clustering analysis. Once the query has been submitted, we can switch over to the table tab and download the total rape incidents for each year as a .csv file.

3.2 Data Processing

In terms of data processing, the first step will be to transform the files (both csv and xlsx) into pandas dataframes. This would involve going beyond simply reading the files as some of them contain comments and additional textual information that is irrelevant to the analysis. The second step involves modifying column headers as many of them contain spaces, are too long, and do a poor job of explaining the observations specific to the column. Therefore, for the

purpose of consistency and to make it easy to reference columns, all of them will be converted to snake_case. Another integral part of data processing for this research involves merging some of the files into a singular dataframe as well as performing data transformation and manipulation i.e., transposing rows and columns. The data will also be filtered for sex offenses/rape instances as the research focuses on this crime specifically. Moreover, the datatypes for each variable will be checked to ensure that they are correct. For instance, variables consisting of dates and times must have 'datetime64' data type.

Missing values will be detected and handled based on the specific datatype for each variable. For instance, in case of qualitative/categorical data, the missing values will be replaced by the mode whereas for quantitative/numeric data, the missing values will be replaced by the mean if there are no outliers. Otherwise, they will be imputed by the median in the presence of outliers. As for outliers, they will be detected through visualizing boxplots and any present outliers will be removed if they are greater than three times the standard deviation. Since one component of this study includes clustering analysis, feature engineering will be performed. In case of ordinal qualitative variables, dummy variables will be created while one-hot encoding will be used for nominal qualitative variables. Prior to the modelling stage, the data will be standardized and will be split into a 30% testing and 70% training split.

4 Research Questions and Methodology

1. Are there disparities between the number of arrestees and the number of victims as well as the number of sex offenses completed compared to the number of sex offenses attempted? To answer this question, I'll create two separate bar charts – one for comparing the no. of victims and arrestees and the other for comparing completed and attempted sex offenses.
2. How do rape rates differ across different states, cities, and towns in 2022? To answer this, I will create two individual choropleth maps using the plotly library – one map to display the overall rape incidents in the states and another to display the rape incidents by each city and town. By hovering over a specific area i.e., a state or a city, a toolkit would appear, showing the recorded population as well as the rape incidents. The locations with darker colors would indicate higher instances of rape whereas the areas with lighter colors would indicate lower instances of rape in 2022.
3. How do rape rates change over time on a national level between 2012-2022? To answer this, I will create a line plot depicting the fluctuations in rape rates over time. The x-axis would consist of the years i.e., 2012-2022 while the y-axis would represent the instances of rape.
4. How do the demographics differ for the victims and the offenders? To answer this question, I will break it down into two parts. First, I will visualize the age ranges for the victims and perpetrators in the form of a grouped bar chart. In this case, the x-axis would represent the age ranges and different colors will be used for the victims and the

offenders to get a side-by-side comparison. Next, I will create two heatmaps to display the sex and race of the victims and the offenders. In this case, higher color intensity would indicate higher rape instances.

5. Does a relationship exist between the victims and offenders? To answer this question, I will create a pie chart as there are only a few categories for the relationship between the two groups e.g., family member, known to victim, stranger etc.
6. What specific locations (home, university, office etc.) are hotspots for higher rape incidents/sex offenses? To answer this question, I will create a word cloud where location names with a greater font size would indicate greater frequency of rape incidents in that specific location.
7. During which time periods throughout the day were incidents of rape highest and lowest? To answer this question, I will create a line graph ranging from midnight to 11:59 pm. This plot would include the timestamps as the x-axis while the y-axis would include the number of rape cases.
8. What are the distinct patterns and profiles of sex offenders in New York State in 2022? To answer this question, clustering analysis will be utilized – with algorithms such as K-means clustering, hierarchal clustering (single, mean, and average linkage), and DBSCAN, as well as an ensemble learning algorithm. The variables that will be utilized in this analysis include offense_category_name (offense type), age_num (age of offenders), sex_code (sex of offenders), race_code (race of offenders), and attempt_complete_flag (attempted or completed sex offense). Furthermore, to check the performance of the clustering algorithms, silhouette score will be calculated.

5 Work plan

Week 4 (2/12 - 2/18):

- Data tidying and recoding (5 hours)
- Questions 1-3 (4 hours)

Week 5 (2/19 - 2/25):

- Questions 4-7 (10 hours)

Week 6 (2/26 - 3/3):

- Question 8 (12 hours)

Week 7 (3/4 - 3/10):

- Presentation prep and practice (4 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).*

- Poster prep (4 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)

Week 10 (4/1 - 4/7):

- Edits and revision for code for reproducibility Q1-5 (4 hours)

Week 11 (4/8 - 4/14):

- Edits and revision for code for reproducibility Q6-8 (6 hours)

Week 12 (4/15 - 4/21):

- Gathering articles and other literature for supporting arguments (1 hours)
- Citing resources (2 hours)
- Initial write-up for draft blogpost (3 hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.*

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)

References

Federal Bureau Of Investigation, Dept. of Justice. 1960-2022. "Crimes Data Explorer," 1960-2022. https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/downloads.

———. 1960-2022. "Data Discovery Tool," 1960-2022. https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/query.

———. 2023a. "Crimes Against Persons,property,and Society." *Uniform Crime Reporting Program*, 1.

———. 2023b. "NIBRS Offense Definitions." *Uniform Crime Reporting Program*, 7–8.

Rape, Abuse, and Incest National Network. n.d. "Scope of the Problem: Statistics." https://www.rainn.org/statistics/scope-problem.