CS4850 – Senior Project
Spring 2023
SP-2-DataMining-AI Red
“Crime Mining”
April 29, 2023

Team Members
Peter McFarlane
Dev Malani
Manav Desai
Duc Ho

Professor Sharon Perry

Website
https://sites.google.com/view/crimemining/home

1 – Introduction

1.1: Abstract

This project is a web application built using the Django framework that serves as a data visualization tool for the National Crime Victimization Survey (NCVS). We used Python and its Pandas, NumPy, and Matplotlib libraries to perform additional analysis with machine learning algorithms and used Chart.js to allow users to dynamically render data from the survey onto the page.

1.2: Overview

The National Crime Victimization Survey (NCVS) is a comprehensive survey of crime victimization in the United States and is the Bureau of Justice Statistics' primary source of information on crime victimization, obtained from a sample of nearly a quarter million individuals in over 150,000 households nationwide. This concatenated dataset contains data from over five million survey responses spanning from 1992 to 2021.

The NCVS includes victims of property crimes such as thefts and burglaries, as well as victims of violent crimes such as assault and rape. The NCVS includes crimes not reported to the police, allowing for a more thorough representation of crime victimization than other crime surveys. The NCVS does not include murder, nonnegligent manslaughter, commercial crimes, or victims of crimes under 12 years old.

Our web app mines different variables from the survey and plots their prevalence against various demographics of people. These variables include property crime (theft, burglary, etc.) as well as personal crime (aggravated and sexual assault, robbery, etc.). Additionally, our project presents detailed analysis of the dataset obtained using linear regression and association rule mining.

1.3: Goals

Our goals were threefold: 1) identify which groups of people are most predisposed to being victims of crime, 2) evaluate how the prevalence of different types of crime has changed over decades, and 3) identify which crimes are less likely to be reported to police.

2 – Project Plan

2.1: Project Team

Roles	Name	Major Responsibilities	Contact
Project Owner	Peter McFarlane	Project management, application development, and data analysis	pmcfarl6@students.kennesaw.edu 770-313-4931
Developer	Dev Malani	Application design and development	dmalani@students.kennesaw.edu 678-579-4093
Technical Writer	Manav Desai	Handle documentation of the application	mdesai10@students.kennesaw.edu 551-247-7998
Developer	Duc Ho	Application design and development, data analysis	dho7@students.kennesaw.edu 470-918-8389
Advisor / Instructor	Sharon Perry	Facilitate project progress; advise on project planning and management.	sperry46@kennesaw.edu 770-329-3895

Dev Malani **Peter McFarlane** **Manav Desai** **Duc Ho**

2.2: Communication

Team meetings and other communication were done through Microsoft Teams or in person. The project owner, Peter McFarlane, took notes of the meetings.

2.3: Scheduling

We had meetings on Teams on Fridays from 7 to 7:30 PM and during class time to plan and coordinate the project.

Our Gantt chart:

Project Name: SP-2-DataMining-AI :: Crime Mining
Report Date: April 29th, 2023

					Milestone #1			*Milestone #2*				*Milestone #3*				C-Day	
Deliverable	**Tasks**	**Complete%**	**Current Status Memo**	**Assigned To**	**6-Feb**	**13-Feb**	**20-Feb**	**27-Feb**	**6-Mar**	**13-Mar**	**20-Mar**	**27-Mar**	**3-Apr**	**10-Apr**	**17-Apr**	**####**	
	Define requirements	100%	Completed	All	5												
	Review requirements with SP	100%	Completed	All	3												
	Get sign off on requirements	100%	Completed	All	1												
Project design	Define tech required	100%	Completed	All	5	10	10										
	Web app prototype	100%	Completed	Peter, Dev, Duc		10	10	10	10	10	15						
	Data mining / statistical analysis	100%	Completed	Peter, Dev, Duc			5	10	10	10	10	10	10	5			
	Website design / layout	100%	Completed	Duc, Manav				5	5	5	5						
	Document required deliverables	100%	Completed	Manav						5	5						
	Test prototype	100%	Completed	Peter						5	10						
Development	Review prototype design	100%	Completed	All							5	5	10				
	Rework requirements	100%	Completed	All							5	10	10	10			
	Document updated design	100%	Completed	Manav									5	5	5		
	Finish web app	100%	Completed	Peter, Dev, Duc								10	5	5			
	Test product	100%	Completed	Peter										5	15		
Final report	Presentation preparation	100%	Completed	All											10	5	
	Poster preparation		N/A	Manav												3	
	Final report submission to D2L and project owner	100%		Peter													1
			Total work hours	**345**	**16**	**20**	**25**	**25**	**25**	**35**	**55**	**35**	**40**	**30**	**30**	**8**	**1**

Legend	
Planned	
Delayed	
Number	Work: hours worked

2.4: Version Control

We used GitHub for version control, and documentation was kept on Microsoft Teams.

3 – Requirements

3.1: Functional Requirements

Our site contains four pages of information:

About (landing page):

- General information about the project and its methodology.
- Information about the dataset.
- Our team and headshots.

Documents:

- A link to this document and the dataset.
- Statistical notes and caveats about the dataset.
- A summary of the technologies used.

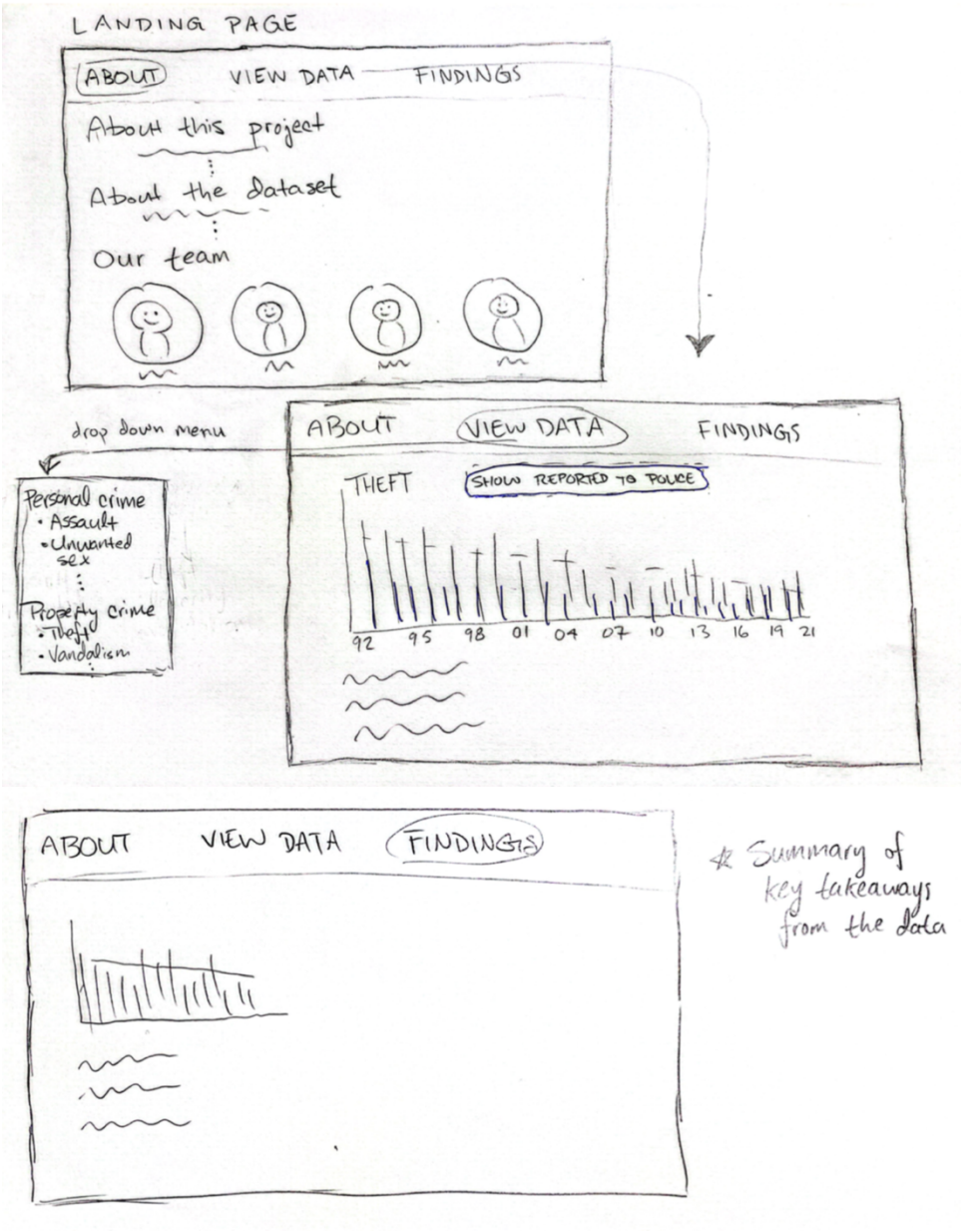
Data:

- The core functionality of the application.
- Users can generate a graph of information from the dataset, selecting variables through drop down menus.

Findings:

- A summary of general takeaways from the data.
- Results of association rule mining and linear regression.

Our initial drafts of the site:



3.2: Nonfunctional Requirements

3.2.2: Capacity

While the time it takes for operations to complete may slightly vary based on the amount of data being plotted in each operation, we expect that it should take no longer than 5 seconds to generate a graph. The site should be able to handle over 30 concurrent users at this 5 second threshold.

3.2.3: Usability

Users will be able to provide feedback to the developers via the email address located at the bottom of the About page. We will consider 90% satisfaction or greater an acceptable satisfaction rate.

3.3: User Interface Requirements

Our application has a very simple user interface with each page easily navigable from the top of the screen and a generally monotone color palette, reserving color for key UI elements and data on the graph. The site was designed with HTML and CSS.

3.4: Software and Communication Interface Requirements

Our webpage uses the Django web framework, which includes HTML and CSS to style the webpage, a web server provided by Django to run the webpage and to keep our DataFrame. It is assumed that the user is using a machine that can run web applications.

4 – Development

4.1: Development Process

The process we used to build our application was to split the work into cycles where we would first discuss what goals we need to meet for the week and then do the designing, developing, and testing. We took around 5 – 7 weeks (about 1 and a half months) to develop the whole application.

We took the first two weeks planning and designing our application based on how it is going to work, where we are going to write the codes, what features we needed to add, what algorithms we should implement, and how we can distribute work among the team. We decided to take on the approach of a data visualization tool where it can extract the data from the NCVS dataset and plot out the statistics of the crimes with graphs and diagrams.

Next, we took one week to gain knowledge on the tools that are needed to build the application. We learned about linear regression and other helpful Pandas functions from the Anaconda website, and we made heavy use of the Pandas, NumPy, and Matplotlib documentation in building the application.

After we accumulated enough knowledge in a week, we decided to build our application for about 2 weeks. We first decided to build the frontend of the application and then proceeded to build out the backend. After building our application, we took another two weeks to test and refine our application and presented it on March 13, 2023 in class.

In the subsequent weeks, we continued to refine the application, smoothing over any rough edges and ensuring the requirements were sufficiently met.

4.1.1: Handling the Dataset in Python

The concatenated NCVS dataset contains three large text files: a household file, an incident file, and a person file. Our app uses the person file; this file is a record of individuals. Each line of the file represents a person responding to the survey and their responses to each question, along with additional information, are encoded as a number in a long sequence of numbers. The accompanying codebook tells us which characters in the line correspond to what question.

The person weight of each response provides an estimate of the number of people in the population represented by a survey response. By removing all lines of the file with a person weight of 0, we reduced the size of the file from 5.7 million lines to just over 5 million.

4.1.2: Building the Application

We developed this project using the Visual Studio Code IDE. Visual Studio Code is a free open-source code editor that is refined and optimized for building, testing, and debugging modern web and cloud applications. We set up our Django project using the terminal and installing the necessary packages.

In Django, views are Python functions responsible for returning information to an HTML page. The primary view that handles the generation of graphs is the “data” view. This view takes information submitted via the drop-down menu form and uses that information to dynamically render a graph to the page. Constructing this view was a large part of the project.

The “datafunctions.py” file handles the construction of the 2D list of data from the file, the construction of the DataFrame from the 2D list, and generates lists of X and Y values for the selected graph. These values are passed into the “data” view and rendered in the graph using Chart.js.

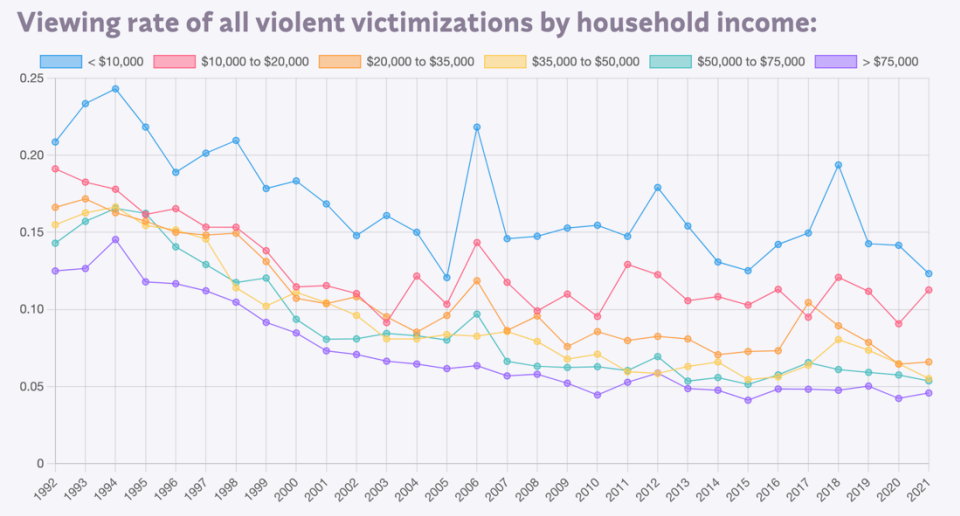
The more sophisticated machine learning algorithms were employed using Jupyter Notebook. For linear and polynomial regression, we used NumPy's polyval and polyfit functions, and for association mining, we computed the support, confidence, and lift values directly without the use of an external library.

4.2: Findings

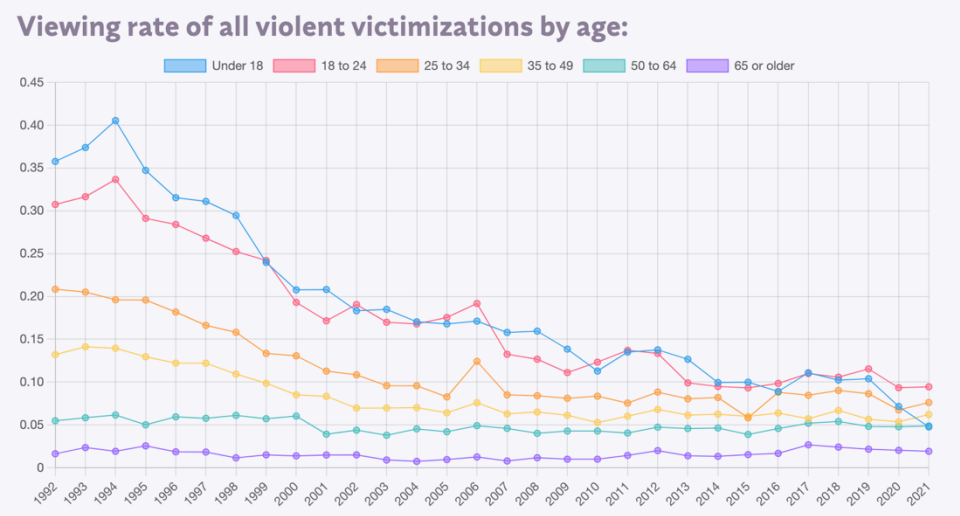
4.2.1: General Findings from the Visualization Tool

Through just a cursory exploration of the tool, we stumbled upon several interesting bits of information:

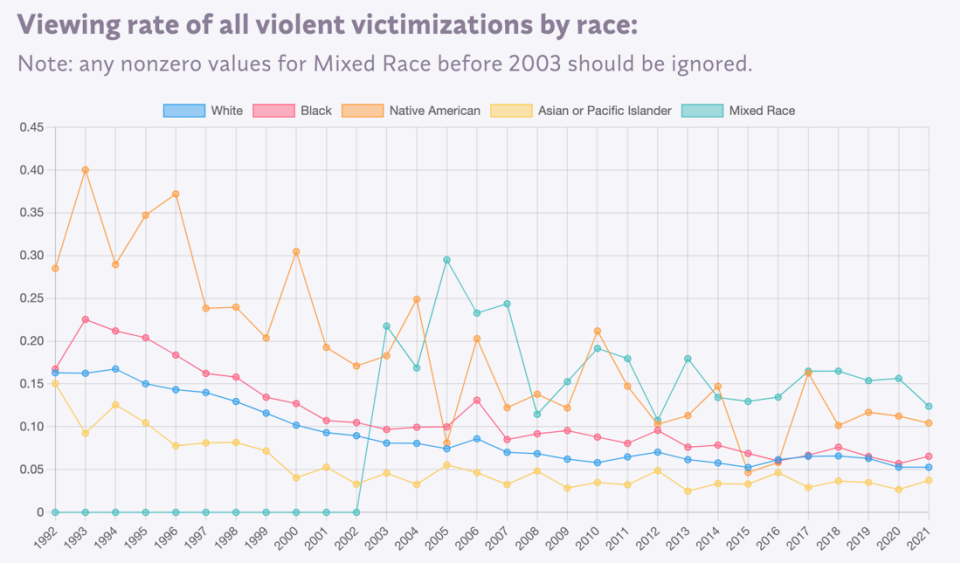
Poor people are more heavily victimized by both violent and property crime. It is perhaps unsurprising, but across all crime categories, those with an income of under $10,000 are more likely to report having been the victim of a crime.



Troublingly, **children are historically the primary victims of violent crimes** such as assaults, although this seems to be less the case more recently.

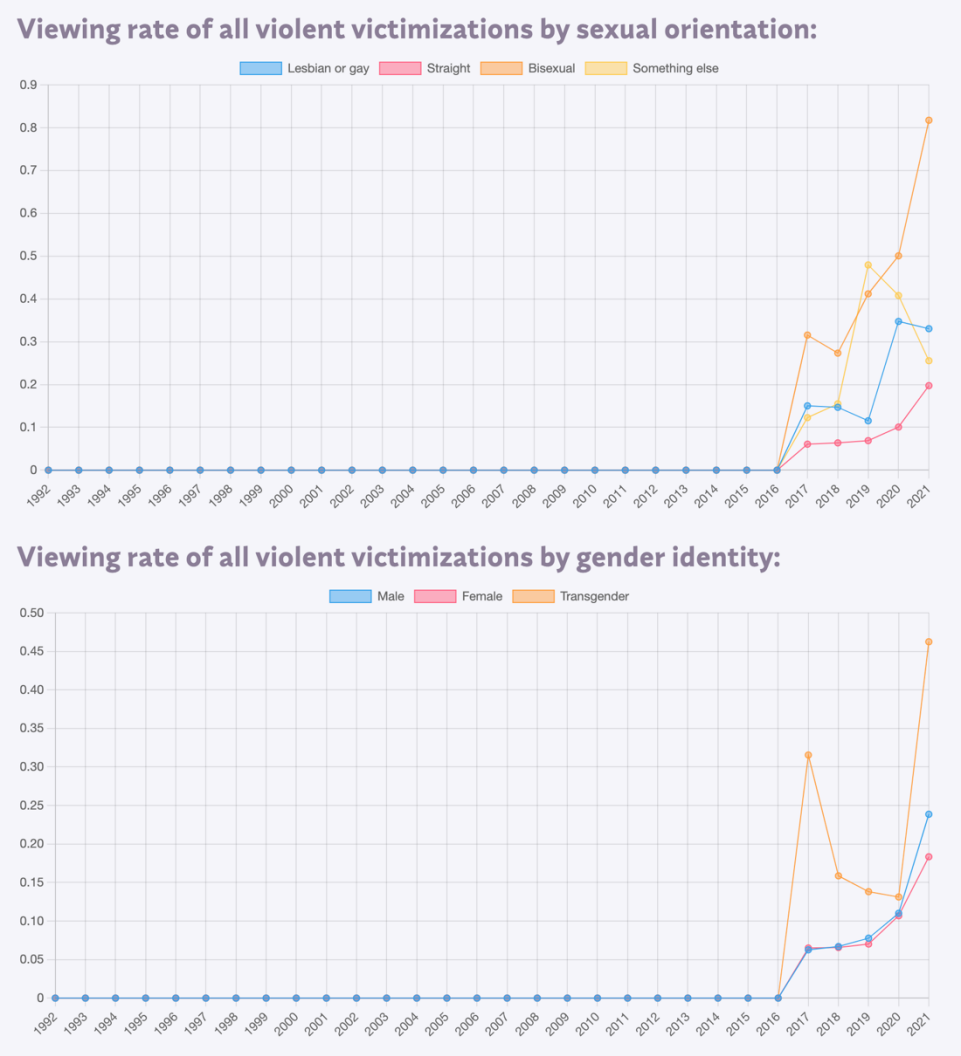


When viewing crime rates by race, we found that **Native Americans and mixed-race individuals are the most disproportionately victimized by violent and property crimes**. Asians and Pacific islanders are the least likely to report being the victim of a crime, followed by white people and black people. For visual clarity, the “mixed race” category collapses more than a dozen different categories of mixed-race people. We were surprised to find that the rates of victimization for mixed-race individuals to be consistently markedly higher than those of Americans who identify as just black.

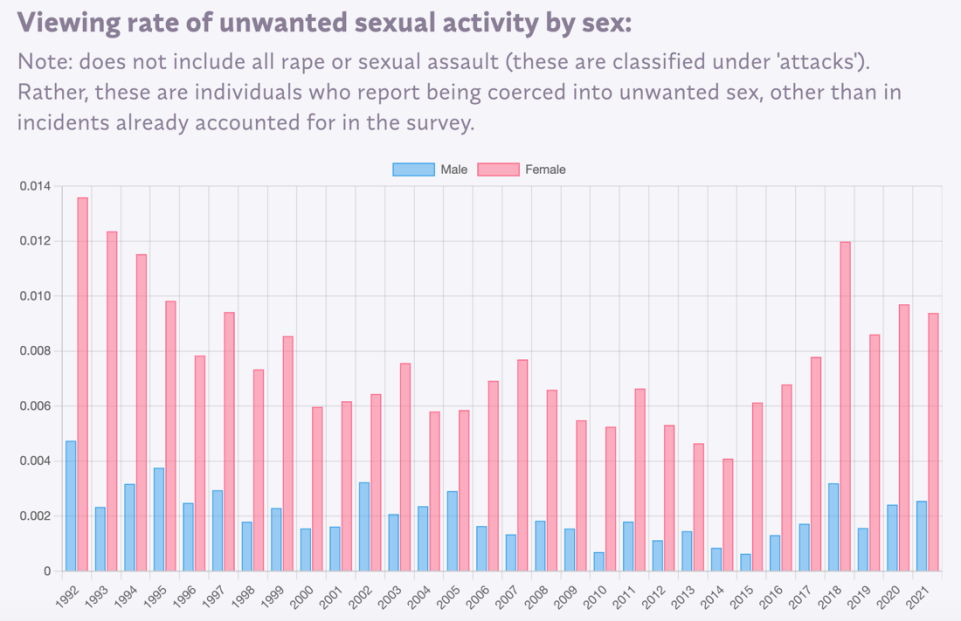


In 2017, the NCVS started surveying people’s sexual orientation and gender identity. Although we have comparatively little data for these variables, there are some stark conclusions that can be drawn from the data we do have. Curiously, when sorting for sexual orientation, **bisexuals report the highest rates of victimization for violent and property crime by a considerable margin** with heterosexuals reporting the lowest rates.

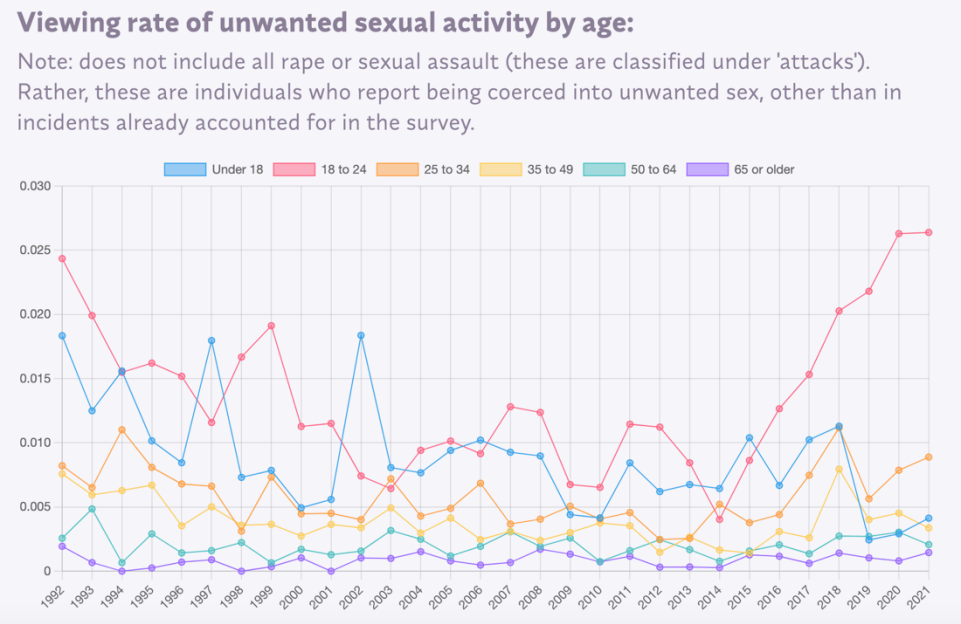
When sorting for gender identity, we found that **transgender individuals are much more likely to be victims of crime** than cisgender individuals; this gulf is even more pronounced when sorting specifically for attacks and unwanted sexual activity. This data underscores the risks that gender nonconforming people and sexual minorities face in American society. It will be interesting to see how the data changes as time goes on.



Perhaps the most fascinating results came from analyzing reports of unwanted sexual activity (individuals who report having been coerced or pressured into sex they did not want to have, excluding rapes and other sexual assaults classified as attacks). When sorting by sex, **we found a spike in reports of unwanted sex in 2018 across both sexes**. We hypothesize that this may be the result of the Harvey Weinstein sexual abuse scandal in late 2017 and the subsequent MeToo movement; individuals were perhaps more inclined to report their own incidents of sexual abuse following the high-profile case that brought attention to the issue.



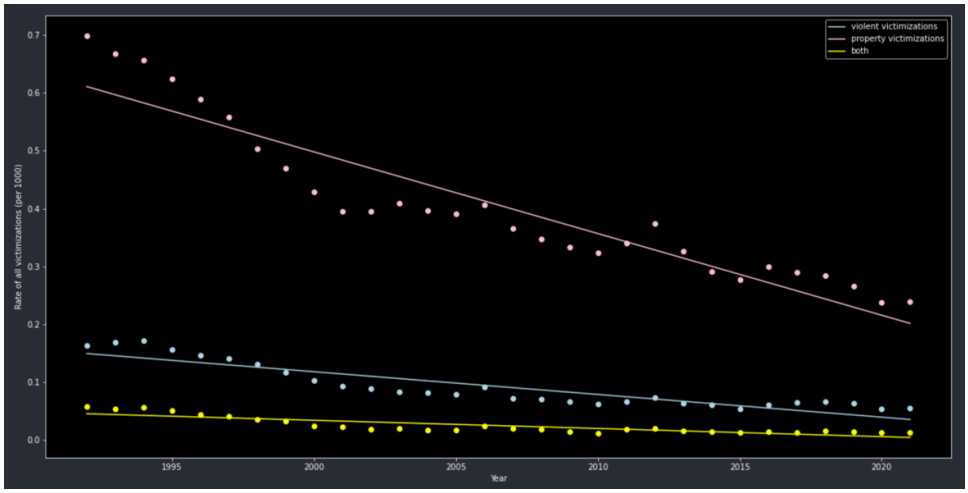
When sorting unwanted sexual activity by age, the picture becomes even clearer. The 18-24 demographic (people most likely to be active on social media and paying attention to the MeToo movement) has reported increasing rates of unwanted sex for the past several years, with the steepest increase between 2017 and 2018. There is a similar spike in children under 18 reporting unwanted sex in 2002, perhaps related to the Catholic Church sex abuse scandal that broke in January of that year. The minor spike among 18–24-year-olds in 1998 and 1999 could possibly be in response to the Monica Lewinsky scandal.



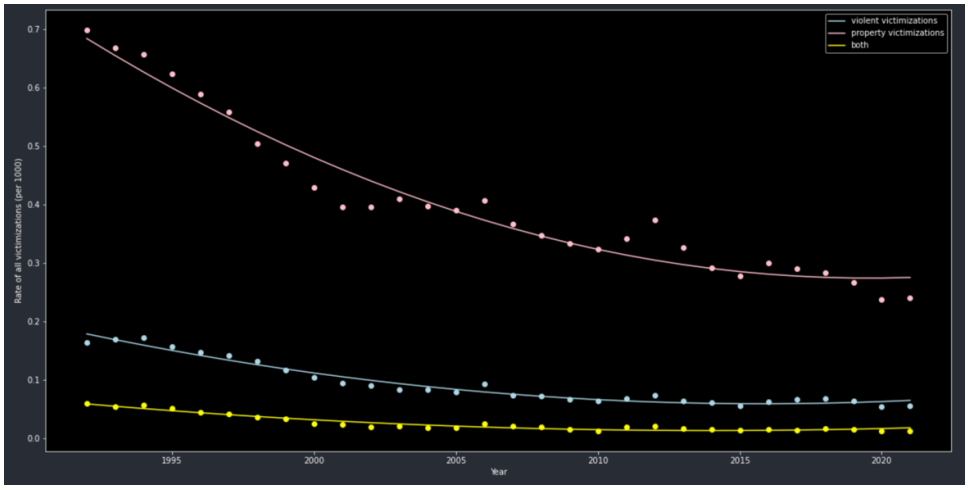
While it's difficult to prove without more data, it seems that **there is some circumstantial evidence that high-profile sex abuse stories in the media increase survey participants' willingness to report their own abuse**.

4.2.2: Linear and Polynomial Regression

We used NumPy's polyfit function to perform linear and polynomial regression on data for rates of violent victimizations, property victimizations, and individuals reporting to have been victims of both. After establishing a line of best fit for the data, we can see that crime has generally been going down since the 1990s, particularly for property crimes.



By increasing the degree of the polynomial from 1 to 2, we can get a better sense of the rate of change of the data; crime dropped more precipitously during the 1990s and has been steadily plateauing since the 2000s. Increasing the polynomial degree also allows us to get a better sense of the outliers in the dataset; as shown below, 2006 and 2012 are outlier years that fall far afield of the trend lines.



4.2.3: Association Rule Mining

Since the dataset deals primarily in categorical data rather than quantitative data, we used association rule mining to determine individuals' propensity to be victimized by a crime given that they are the victim of a different crime or possess some specified personal characteristic. The most interesting results are displayed below.

To do this, we calculated the support (the frequency with which a crime or characteristic appears in the dataset), the confidence (the strength of association between the X and Y variables), and the lift (the strength of association between X and Y, taking into account their relative frequency of occurrence). A lift value greater than 1 indicates a strong association, meaning that if an individual is the victim of X, they are more likely to be a victim of Y.

Using reports of theft or attempted theft as the X variable and reports of violent attacks as the Y variable, we can see that an individual is much more likely to be the victim of an attack given they are a victim of theft. Individuals who report being forced or coerced into sex are over 16 times as likely to also report being the victim of a violent attack.

```
X → victim of theft or attempted theft
Y → victim of violent attack
Support:  0.003261259049524368
Confidence:  0.09287742829384918
Lift:  5.281082917406366
```

```
X → reported unwanted sex
Y → victim of a violent attack
Support:  0.00014411599404016276
Confidence:  0.29291294026815756
Lift:  16.655257941071582
```

When selecting for personal characteristics rather than crimes, the lift values get lower. An individual is slightly less likely to report a violent victimization if they are a woman, and slightly more likely to be the victim of an attack if they are a man. An individual living in an urban area is slightly more likely to be the victim of a property crime, whereas for rural Americans, property victimizations are considerably less likely.

```
X → individual is male
Y → reported violent victimization
Support:  0.00622207058168025
Confidence:  0.012801534519308589
Lift:  1.1266045640205458
```

```
X → individual is female
Y → reported violent victimization
Support:  0.005140864496111038
Confidence:  0.010002479674622252
Lift:  0.8802725357616583
```

```
X → individual lives in an urban area
Y → victim of a property crime
Support:  0.00036695452614398457
Confidence:  0.03991998556844823
Lift:  1.010472207095268
```

```
X → individual lives in an rural area
Y → victim of a property crime
Support:  0.00020990337137236748
Confidence:  0.01469017440776232
Lift:  0.3718441463605727
```

Association rule mining provides a useful lens for understanding the degree to which a crime is associated with another crime. The main takeaway from these results is that being the victim of one crime is a better predictor of whether an individual will be the victim of another crime more than any given personal characteristic.

5 – Challenges and Assumptions

5.1: Assumptions

This project assumes that, unless otherwise stated, the NCVS contains accurate information from survey participants answering questions honestly. This is a necessary but flawed assumption to make. As noted earlier, external or internal factors may inhibit a person from being upfront about having been the victim of a crime, even in the context of a survey where the participant's identity is protected, because talking about one's victimization can be difficult.

5.2: Complications with the survey methods

The NCVS acquires its data by asking participants a series of survey questions. Crimes are ranked according to severity, with the most severe crimes, such as assault, asked first. Some incidents can be classified as multiple crimes, in which case the survey classifies it as the more severe crime.

For example, the survey may ask if the participant was attacked or threatened, including incidents of rape or sexual attack. Then, it may ask: "other than incidents already mentioned…" and inquire about any incidents of forced or coerced sex from the interviewee. Because of this, it was difficult to get an accurate count of victims of certain crimes.

After cross-referencing our results with those of the NCVS' own graphics, we are confident that the *shape* of our data is, more or less, correct. However, the number and rate of some crimes in our tool are lower than one might expect they should be.

5.3: Managing the size of the dataset

Dealing with a text file over five million lines long, with each line being hundreds of characters long, provided a challenge. It often took over a minute to create a DataFrame

with Pandas, and so requiring the DataFrame to be generated every time a user visits the site proved to be untenable.

To solve this, we saved our DataFrame to a parquet, a file format that allows for very fast storage and retrieval of data. Additionally, the parquet file was small enough to be pushed to the GitHub repo, whereas the raw data was too large.