Midterm Project: Victimization - Aidan Long

A) Introduction: The Dataset

The dataset that we are looking at deals with crime and crime victimization since 1973. The National Crime Victimization Survey Series (NCVS) collects data on personal and household victimization via surveys. This survey is funded by the US Department of Justice and Bureau of Justice Statistics. The survey is ongoing and deals with a nationally representative sample meaning this is a retrospective observational study as we are looking at data on crimes that have been committed and tallied after the offense. We are not looking at active crime despite the survey being an ongoing collection. This survey was designed to detail information about victimization, consequences of crime, number and types of crimes not reported, measures of selected types of crimes, and to permit comparisons between areas of crime. The survey is split into two categories: "personal" and "property" each with multiple qualifying factors. Factors include:

* Personal: Rape, Sexual attack, Robbery, Aggravated and Simple Assault, and Pocket-picking
* Property: Burglary, Theft, Motor vehicle theft, and Vandalism

The survey was conducted so that respondents answered a series of questions to see if they were victimized by an interview. The questioning is directed to one individual aged 12 or older who then speaks on behalf of the rest of their household.

The data surveyed includes:

* Type of crime, month, time, location crime, relationship between victim and offender, characteristics of the offender, self-protective actions taken and the results of those actions, consequences of the victimization, type of property lost, reported or unreported crime, offender use of weapons, drugs, and alcohol, and basic demographics (age, race, gender, income)

In total, the data set has collected around 8043 observations across 1811 households with 81 metrics of differentiation (columns) for each observation. Looking at what is surveyed for and how the sampling is done we can conclude that there is no sampling bias but some survey limitations. Because the US and US crime is so large, the sample size may not be a perfect representation of the nation. There's also the chance people may withhold or lie as it deals with an unfavorable subject so a lot depends on the interviewer and their questions. But outside of human error and size factors, the survey does a good job of eliminating bias by grouping and using sub-groups. They take lists of households from the Census Bureau, take a random sample from that, and then select a random person from the household so every level of grouping is randomized. The data also includes all states and types of areas so everyone in the population is represented in the data. While there may be multiple natural limiting factors, the survey does a good job of accounting for and hindering biases.

B) Characteristics of Sample

The sample of people and households is determined by a randomized process to ensure that representation is maximized. We can look further into these characteristics and see how victimization takes place by variable/characteristic. Let's look at employment types where we have a table of the distribution of total victims by their type of employment.(figure 1). We can see that people who worked for a private company were victimized the most of all employment types. We then see a drop off after out-of-universe for people who were self employed or worked for the government. This reveals victims were predominantly working more for private companies and less for the government. This figure also shows the range of employment, highlighting diversity in household response for the overall sample. A second characteristic we can look

into is age and how many respondents/victims there were by age (figure 2). We can see that the respondents were aged 12 all the way to 90, with ages 30 to 65 being the most victimized group, age 30 being the most victimized. We can see a very even distribution of age showing how everyone can be a victim regardless of age but that those who are middle aged are more commonly a victim of crime. The distribution of age we see in the graph is also another indicator of a good sample size for the data. Another interesting variable we can look at is income by person or household and how that affects crime attitudes (figure 3). Looking at the figure we can see the number of respondents by income level and see that more crime is committed to those with a higher income. The most victimized income level is 13, which if we refer to the graph of income ranges corresponding to the index number we see on the table (figure 4), is $50-75,000. Those with a lesser income are less likely to be victimized and numbers generally climb until the income is over the poverty line (~$30k). Looking at these 3 characteristics as examples we can see the diversity the sample and survey exhibits by seeing the differences and distributions of things like age, income, and employment; we have a better understanding of who is affected and the general range of those affected.

Figure 1

Employment Type	Victim Count
Private Company	3808
Out of Universe	2617
State, County, or Local Government	681
Self-Employed	673
Federal Government Employee	191
Residue	73

Figure 2

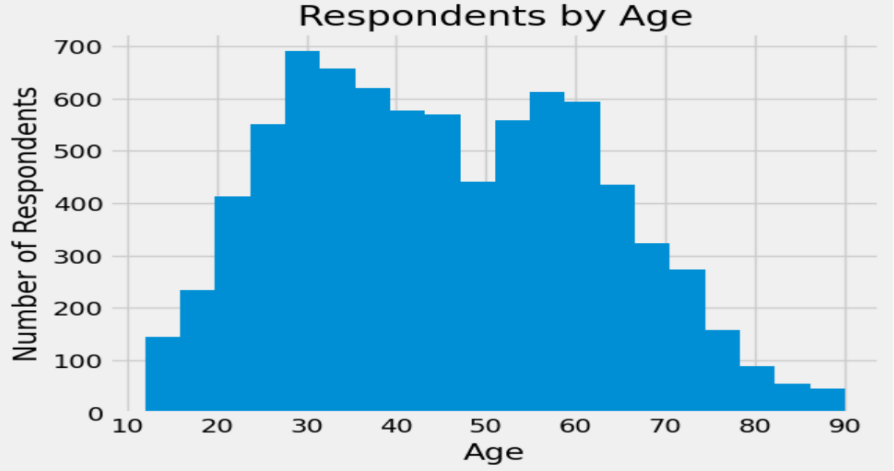


Figure 3

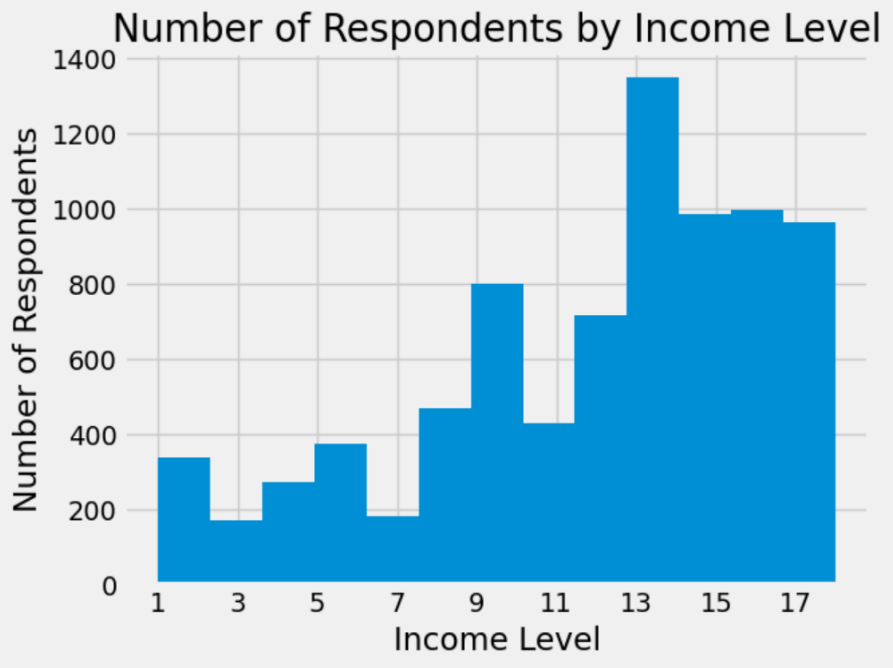


Figure 4

Index Number	Income Range
01	Less than $5,000
02	$5,000 to $7,499
03	$7,500 to $9,999
04	$10,000 to $12,499
05	$12,500 to $14,999
06	$15,000 to $17,499
07	$17,500 to $19,999
08	$20,000 to $24,999
09	$25,000 to $29,999
10	$30,000 to $34,999
11	$35,000 to $39,999
12	$40,000 to $49,999
13	$50,000 to $74,999
15	$75,000 to $99,999
16	$100,000-$149,999
17	$150,000-$199,999
18	$200,000 or more

C) Relationship Between Variables

While individual characteristics give a lot of information into the sample and data, we can gain a lot more insight by comparing multiple variables and characteristics from the data. Take forced sex as a crime category and race as a household characteristic and we can see how the variables interact (figure 5). The bar graph shows that the race most victimized by forced sex were white people, followed by black people. Other races include asians, and various mixed races show up the list is dominated by (mainly) white and black people as victims of forced sex. We can presume from the bar graph that white and black people are more likely to be victims of forced sex from a correlation standpoint. Next, the data file gives us multiple forms of theft including vehicle, burglary, and basic stolen goods. In order to gain a new look at theft in general I made a function that added all types of theft and sorted that function into a new column that I added to the dataset. I then took total thefts and plotted it against age to see thefts by age (figure 6). We can see from the line graph that most thefts take place at age 30 similar to our age and crime in general graph. The age range of thefts is highest from 30 to 60 and then dips all the way down to age 90. We can see that the max thefts at age 30 is up to 40,000 giving us an idea of how many people experience theft of some sort. Even kids as young as 12 and elderly people as old as 90 still are victims of theft in counts of thousands. We can also conclude from the graph that 30 is the age most likely to be robbed in some sort of way and this chance decreases once you hit 60.

Figure 5

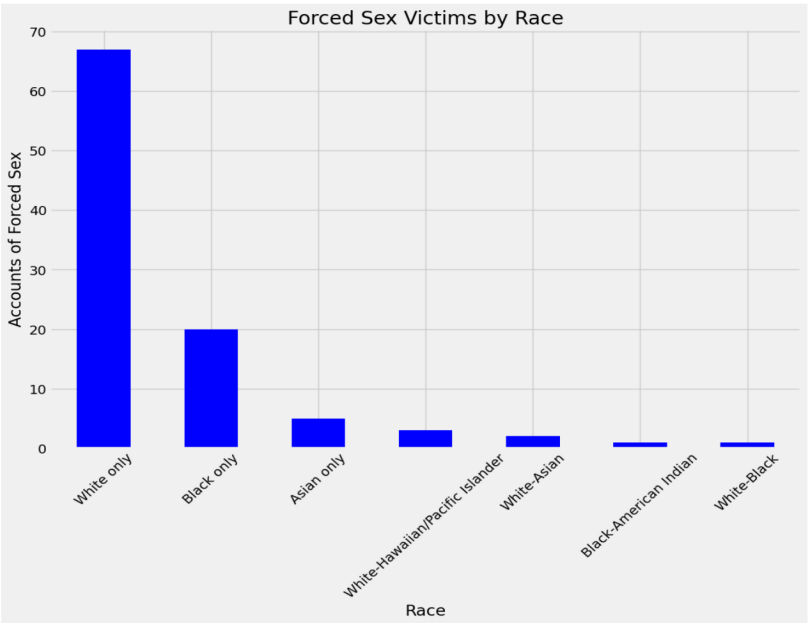
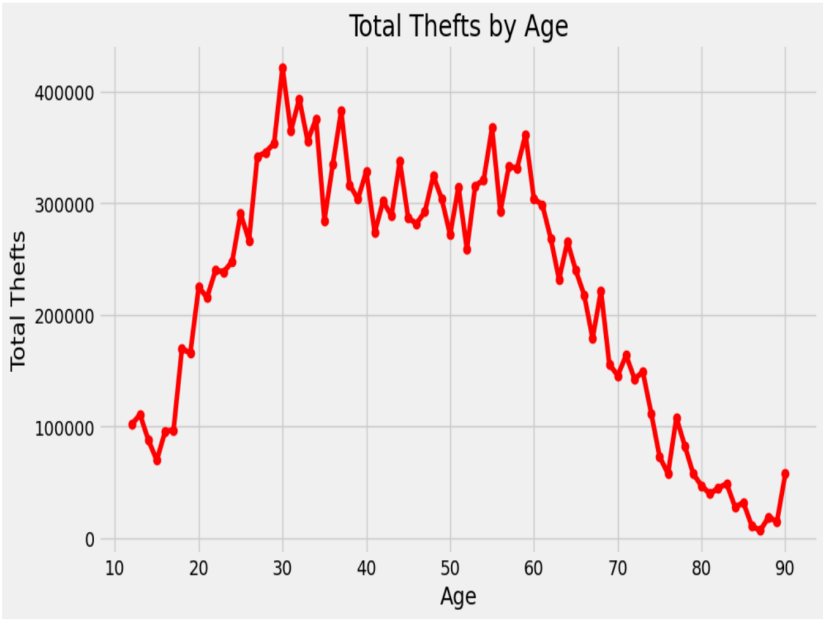


Figure 6



D) Providing Context

For the relationship between variables we analyzed to make sense, we need to think about the context behind the data. In figure 5 we see the number of forced sex victims by their respective race. We can see that white and black people populate most of the figures, implying that white and black are more likely to be victims of forced sex; other races follow in smaller amounts but barely appear. While there is correlation this isn't necessarily a causal relationship as white people make up a majority of the US population followed by black people so this could be proportionate to population. Some confounding variables might be status, income, location, and access to resources as all of these differ based on race.

Depending on the mix of the variables I just listed people would be less (or more) likely to report these crimes and get help/prevent forced sex. This might skew the data that we see more white people or affect the data in other ways. A question I would like to answer would be: are forced sex victims targeted based on their race? However I can't answer that question without data on the offender and the relationships between the victims and the offenders. If we had the opposite parties data we could form more in depth analyses of the relationship between the two groups. In figure 6 we have total thefts by age. Victimization by age on the other hand has lots of data to support that people are more likely to be targeted by age. We know the age most targeted is 30 and the highest range is from 30 to 60 before it dips. This makes sense as a casual relationship because the age distribution for people under 60 is similar so people would be targeted on a number of factors that we can refer to as the situation's confounding variables. These confounding variables could be income, material belongings, and location. Thinking about it probabilistically, people who are 30 are more likely to own more belongings, property, and travel out to places where they can be robbed and attacked more easily. When people are 60+ there are less of them but they also put themselves in less situations where crime can be carried out. Young people don't own enough to take so crime is more targeted to middle aged people; it makes the most sense logistically. A question I would like to ask would be: which age bracket commits the thefts? Are people stealing from their own age brackets or others? However I would not be able to answer these questions without information on the offender which once again we don't presently have.

E) Conclusion/Self Assessment

To conclude my findings on the victimization data it was interesting to see how much data could be called and compared from raw statistics. I learned a lot from exploring the data file and the codebook for corresponding information. This was the first time I have ever dealt with such a large amount of data in one centralized place so to see how it was organized was very helpful. I haven't really thought before about how surveys could be so large and somewhat accurate with all the biases and limitations that one would expect a survey of that size to have. The codebook was interesting because you got to see how raw data corresponds to actual terms. I wouldn't have considered for example that you put the actual gender in the columns rather than a "1" for man and "2" for women. The organization strategy was intriguing and thought provoking in a way. It was also beneficial to just explore the dataset and try to see the relationship between all the variables. While we only looked at a few variables and noted relationships between some, there are so many combinations of variables you could use to generate more data. Overall, it was very helpful and insightful to look at large scale code and attempt to navigate it.

Bibliography

United States. Bureau of Justice Statistics. National Crime Victimization Survey, [United States], 2020. Inter-university Consortium for Political and Social Research [distributor], 2021-10-18. Cited: 11/10/2023 from https://doi.org/10.3886/ICPSR38090.v1

Ritchie, Hannah, and Max Roser. "Age Structure." *Our World in Data*, 10 Nov. 2023, https://ourworldindata.org/age-structure#:~:text=The%20global%20population%20breakdown%20by,bracket%20between%2025%20and%2065.

U.S. Census Bureau Quickfacts: United States, https://www.census.gov/quickfacts/fact/table/US/PST045222