By: Dae Woong Ham, Haoyang Xu, Boan Fan

Sampling Surveys – Final Project

Section 1 Introduction

"You have young men of color in many communities who are more likely to end up in jail or in the criminal justice system than they are in a good job or in college. And, you know, part of my job, that I can do, I think, without any potential conflicts, is to get at those root causes."

- Barack Obama

This is taken from the former president Barack Obama in a former speech to address the crime issue in America. Victimization rates in general are very important because that is directly related to our safety. It will be useful and interesting to do a comprehensive analysis of the United States crime victimization rate. To be more specific, there will be different factors that affect the crime victimization rate according to many different categories. As indicated by president Obama, ethnicity may be correlated with the crime rate, but will that also be correlated with the victimization rate? These questions inspire us to analyze victimization data and seek what is truly associated with the victimization rate.

Section 2 Sampling Design

Section 2.1 - About NCVS

The NCVS is a household survey sponsored nationwide by the Bureau of Justice Statistics (BJS). It is the only source of data that includes victimizations both reported and non-reported to the police in the United States. The United States Census Bureau serves as the primary data collection organization for the NCVS.

Interviewees are asked about the number and characteristics of victimization experienced during the previous 6 months. Households are interviewed every 6 months for a total of 7 interviews over a 3-year period. Each survey consists of two stages. In the first stage which serves as a screening process, respondents are asked about the number of experiences of victimizations during the 6-month reference period. In the second stage, dates and characteristics of each experience of victimization are recorded.

The target population is U.S. residents age 12 or older residing in housing units or group quarters (GQs), such as dormitories, rooming houses, and religious group dwellings. The survey excludes persons under age 12; crew members of maritime vessels; armed forces personnel living in military barracks; the homeless; institutionalized persons, such as correctional facility inmates; U.S. citizens residing abroad; and foreign visitors to the United States (BJS).

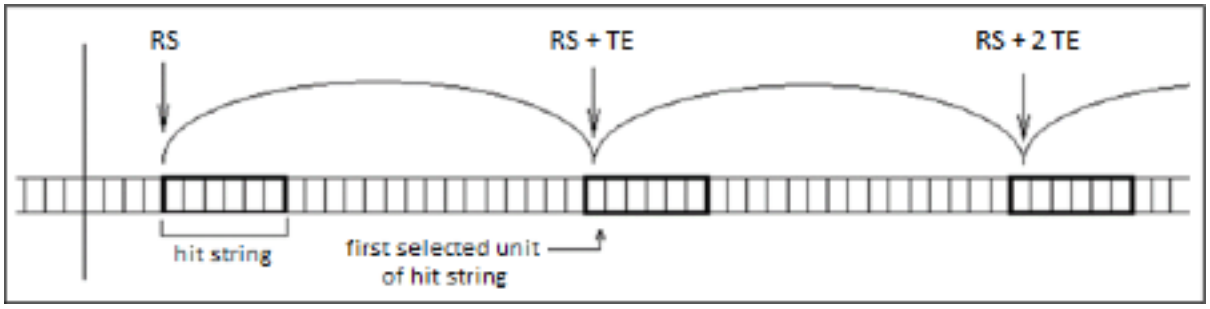
The NCVS collects information on both violent and non-violent crimes. Information about the offender, such as age, sex, race and ethnicity, victim-offender relationship; characteristics of the crime event, including time and place, use of weapons, physical injury, and economic consequences of the crime. Crimes recorded in the NCVS are associated with the addresses of the respondents instead of the locations of crimes.

Section 2.2 - Design Elements

The first stage of sampling is to divide the country into Primary Sampling Units (PSU), which consist of large metropolitan areas, counties, or groups of bordering counties. The PSUs are defined within each state with characteristics such as land area, population and natural boundaries using data from the 2010 Census. The population for a PSU is at least 7,500 persons, and the land area is at most 3,000 square miles except some very sparsely populated PSUs. Then, the PSUs will be divided into two strata based on the size of their Core Based Statistical Area (CBSA). PSUs with large CBSA are self-representing and will be automatically included in the sample. PSUs in the other stratum are known as Non-Self-Representing PSUs. NSR PSUs within sates are then grouped into strata based on decennial census demographic data to reduce inter-PSU variance within all strata within the state. NSR PSUs are sampled with probability proportional to population size and one PSU is selected from each NSR stratum.

The purpose of the second stage selection is to select a sample of housing units and Group Quarters (GQ) from the first stage sample PSUs. A systematic sample is selected from an ordered list of addresses from CQ and housing unit frames. The systematic sample design starts with a known sampling interval referred to as the "take-every" (TE) which is the inverse of the selection probability within a PSU. Next, a random start (RS) is calculated as a random number from a uniform distribution on the interval (0,TE). The RSs and TEs are used to determine the selected units of the sample.

Figure 2.1 Representation of systematic sampling



Source: Bureau of Justice Statistics, National Crime Victimization Survey.

Then, within the selected addresses of housing unit or GQ, households are grouped into clusters of four. Then, a simple random sample of the clusters will be taken within each housing unit or GQ and one cluster will be chosen. All four households in the cluster will be included in the sample. To check this idea, we get the summary statistic of the V2002 variable from the second dataset, which should be the cluster id.

Table 2.1 Checking Cluster ID

```
> table(da36828.0002$V2002)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
 19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
 37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
 55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
 73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
 91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
```

This table clearly shows that for each unit there are 4 households, which correspond to what the cluster is in the survey.

Table 2.2 Summary of Sampling Stages

Stage	Sampling Unit	Sampling Method	Stratification Criteria
1	PSUs (counties or groups of counties with population at least 7,500 and land area at most 3,000 square miles)	All self-representing PSUs are automatically selected; non-self-representing PSUs are selected proportional to population size.	Whether the PSU has large Core Based Statistical Area (CBSAs) (whether the PSU is self representing).
2	Housing unit and Group Quarter (GQ)	Systematic Sampling	
3	Households	SRS of cluster of four households; all households in the cluster chosen will be sampled.	

Section 3 Publicly Available Data

The 2016 NCVS releases the data publicly for free to everyone in the world. However, due to certain strict privacy laws they have edited several columns to mask the identity of the person and household picked for interview. For example, the stratum in which the household was picked is not reported as the real stratum code but rather a pseudo stratum is used. This goes the same for specific household ids, etc.

Before the data is released NCVS spends 6 long months to process the data in their own way. There are two main things they do. They correct nonresponse and consequently adjust the weights. To analyze the data, it is essential to understand exactly how this process was done as will be illustrated in the following subsections.

Section 3.1 Nonresponse

Like many complex surveys, NCVS tries very hard to minimize nonresponse. The 2016 NCVS was relatively successful in dealing with nonresponse. There was in general about 77% response rate for households and more importantly since we are analyzing from the perspective of the person data, there was around an 85% response rate for people. And out of the 15% that did not respond around 7% of it did actually respond but chose not to answer any of the questions simply because they did not want to participate in the survey (it is by law in the United

States not required to answer the NCVS survey). Moreover, the remaining around 7% couldn't respond because they were never available. (Note these are unit non-response rates. Item non-response rates can be individually calculated for each variable)

The NCVS to get to that 85% response rate and also further adjust for the 15% nonresponse mainly used three methods. The first method is trying again to get the response. This is mainly done through manually calling them back or trying to get a proxy that knows the people in the household well enough to answer questions or basic questions about them. For example, if the household is a family of four but the husband is not at home; it is very likely the wife knows enough about the husband and related criminal statistics to answer about her husband.

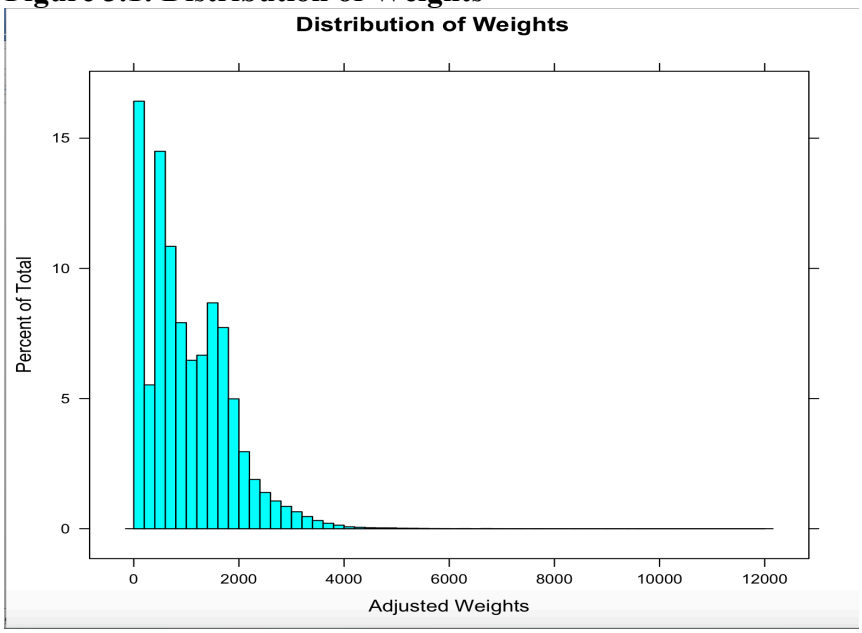
The second method to fix for nonresponse is imputation. However, it is true that NCVS did not do an imputation for ALL the columns. The NCVS has two methods of imputation. The first method is logical imputation. For example, if the household is a family and the father states that he has two sons aged 20, and for some reason the gender variable or the item of his son is completely missing, then logical imputation will be used to assign male. The variables that were selected for this logical imputation is age, gender, race, and number of crime incidents. Therefore, in the actual dataset it can be seen that there is not a single NA for these columns. Of course the number of crimes is imputed by seeing if they answered yes or no to any of the specific crimes and logically imputed that way. The other imputation method they used is the hot-deck procedure. They specifically used the sequential hot-deck imputation, basically grouping up a variable with other similar variables to impute the missing data. This was used to impute also similar variables from above that couldn't be logically imputed such as age, gender and race. Again it is important to realize that imputation was not used for all variables with missing data so one should still expect to see lots of missing data in the original dataset.

Lastly, the final method NCVS uses to account for nonresponse after all this is done is to adjust the weighting. Essentially, weight adjustments will try to adjust those who are in the sample more after setting those unit non-respondents' weights to 0. This will be further explored in the next subsection.

Section 3.2 Weight Adjustments

The weight adjustments are not only done to adjust for non-response. There are many factors that come into adjusting the weights for this complex survey, all with one common goal: to try to represent the United States age 12 or above population better. There are two variables for weights in the survey. The unadjusted weights (otherwise known as just base weights) and the adjusted weights. Here is the distribution of the adjusted weights (which we will use for the analysis).

Figure 3.1: Distribution of Weights



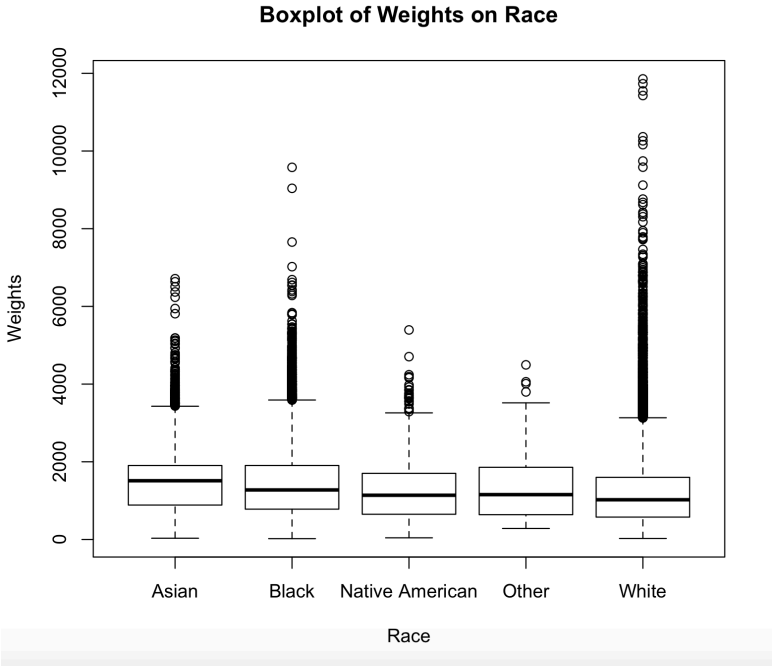
What one should immediately see is that the distribution of the adjusted weights has around 15% of values close to 0. This was purposefully done by NCVS because around 15% of people were unit non-response as mentioned above. We will now describe how these adjusted weights were calculated. The first stage is just the base-weight. Therefore, the original base-weight is of course just the inverse of the probability of selecting that house or person. This is calculated from the complex survey process and this will be different for each PSU because the sampling design throughout its stages uses unequal probability sampling (for example sampling PSUs by population size). The first weight adjustment they do is when they observe many housing units within a GQ (grouping quarters). For example, if the NCVS picks a GQ that seems to have an abnormally large number of housing units, then weighting will be adjusted to make sure these houses are not overrepresented. Then the next adjustment factor is for nonresponse. Because nonresponse will still weigh households that didn't respond if not adjusted, they use the following factor called HHNRF defined in the following way:

$HHNRF = \frac{A+B}{B}$ where A is the weighted count of interviewed households and B is the weighted count of non-interviewed households. Of course there is also within household nonresponse (people in the household not responding) and with exactly the same formula they defined WHNRF to have the following formula: $WHNRF = \frac{C+D}{D}$ where C and D is exactly the same as A and B respectively but for people. These are applied to those that did responded and those that didn't respond had immediately 0 weight as evidenced above. Note this method is the weighting class adjustment in our book, thus the NCVS must have assumed there were some covariates relevant to do this, in other words the nonresponse was missing at random given covariates.

The next step of weight adjustment is called ratio adjustment. The main idea for this is to try to get all the ratios of certain United States demographic characteristics. For example, if the survey feels like they didn't capture enough black people in the survey according to the U.S. Census data, they will try to weight black people more in the data. Therefore, one should expect

to see the weights be differently distributed among certain ethnic groups that might not have been chosen enough. We have tried to analyze whether the sample does indeed reflect this as shown in the following figures:

Figure 3.2 Adjusted Weight distribution among Race



Surprisingly there is not that much difference among the ethnicities. This must mean already that the NCVS have successfully managed to really get ethnically diverse samples. However, as small as it may be, there is still evidential differences as Asians have the highest mean weight; this makes a lot of sense because Asian Americans have a higher probability to be underrepresented while the dominant population of whites have the lowest mean weight as expected.

The last few factors NCVS added to adjust for weights is known as bounding adjustments and Time-in-Sample adjustments. All these adjustments are very small and have to do specifically with if the people interviewed fall out of the time of interest study. Then there are specific weights to account for this. The final weight computation is obviously all these factors multiplied together. This adjusted weight is put in the variable called "WGTPERCY" and it will be used for our analysis when building our survey function.

Section 4 Questions of Interest

In this project we seek to understand a general relationship between those that are affected by crime and possible correlations with other variables. The first thing we will do before starting any analysis is to do exploratory data analysis of basically asking the question: Does our data reflect the target population. After all, if our sample is not representative of the United States population 12 years or old, then all our inferential claims and associations will be wrong.

After the EDA, we will first seek to give general statistics on crime rates of the United States. For example, what is the general crime rate of the United States (including all minor

crimes such as simple stealing)? What is the crime rate of sexual crime? Then after we ask these general questions, we will focus on answering association questions about general crime rates pertaining to four covariates of interest. These are education, age, race, and gender. Some of the questions we will ask is whether or not some ethnicities experience more crime? Does male or female have any association for being more likely prone to crime? What about sexual crime? Does sexual crime have more of an association with gender? If so by how much?

One important disclaimer that we want to make before starting our analysis is that all of these inferential claims are purely claims of innocent associations! We are NOT making causal claims. This is mainly because we are not doing an experiment at all. We are just observing data and have no theoretical basis for making any causal associations.

Section 5 Methodology and Cleaning Data

5.1 Data chosen

Before talking about the EDA and analysis, we will use this section to illustrate how we cleaned the data and did the analysis. First the main dataset we used was the person dataset from NCVS. NCVS has organized its datasets into 4 main datasets. The only ones that were useful for us is the person and household datasets. From the household datasets, the only variables that we took were the pseudostratum code and cluster ids. These were essential variables that we needed to build our survey. We then merged this to the person dataset by the common variable household_ids. The variables that we extracted from the person dataset is the following: age, education, gender, race, unwanted_sex, general_theft, attacked_weapon, adjusted weights, and any crime, pseudostratum and cluster id. These are of course my own named variables, the variables listed in their codebook are all numerically coded such as V3002, etc. Here is a table listing my interested variables.

Table 5.1 – Variable Chosen

Variable name in NCVS	Variable name for us	Description	Values we chose to use
V3014	age	Age of person	Same numerical from 12-90 (note only 12+ people were interviewed)
V3020	education	Education of person coded in as factor	Factored in as either high school or below, college, or higher education
V3018	gender	Gender of person (M/F)	Male or Female
V3023A	race	Ethnicity coded in as factor	White, Asian, Black, Native American, or Other. Others contained anything that was ambiguous or multiethnic.
V3081	any_crime	Basically a Boolean to indicate whether they were affected by any crime or not in the last six months	1 or 0 indicated yes or no respectively
V3046	Unwanted_sex	Forced or unwanted sex imposed on them in the last six months	1 or 0 indicated yes or no respectively
V3034	General_theft	Any minor theft counts in this category of theft in the last six months	1 or 0 indicated yes or no respectively
V3042	Attacked_weapon	Attacked with any sort of weapon that is harmful such as a bat or scissors in the last six months	1 or 0 indicated yes or no respectively
WGTPERCY	Weights	Adjusted weights the way it was described above	Kept the same numerical value the way the histogram was shown above
V2117	Pseudo_stratum	Stratum code	Numerical integers
V3002	Cluster_id	Cluster code	Numerical integers

5.2 Imputations

As explained above not all the variables were imputed. The specific variables such as unwanted_sex, general_theft and attacked_weapon were not imputed and were left as NA. Moreover, unit-nonresponses were immediately given 0 weight so it was only worthy to do imputations on the item non-responses, which is much smaller than unit nonresponse. However, there are were still around 1% of item nonresponses in specifically the three variables mentioned above. Therefore, because we had plenty of columns, we decided to do imputation. To do this imputation, we used the mice() function in R that was versatile in that it used two methods for

imputation: logistic regression and predictive mean matching using other columns in this case age, education, gender, and race to predict for some of these values. It is important to note that because we really imputed only 1% of item nonresponses using other covariates, it really will not change the data so much.

5.3 Cleaning the Data

There wasn't too much cleaning that was necessary. The only thing we did clean-up is some of the covariates we were interested in. For example, education had 24 factors each indicating very specific education such as elementary school grade 1, elementary school grade 2, etc. and we grouped them up into better and more compact categories. This was the same for ethnicity. Some of these ethnicities were extremely specific and only had a few people in there, so it was much more sensible to group them up into major categories. Lastly for the any_crime general variable, it was first reported as number of crimes but since we really care whether or not they were affected or not we changed anything above 0 to a 1. This is all the cleaning that we did.

5.4 Survey Design

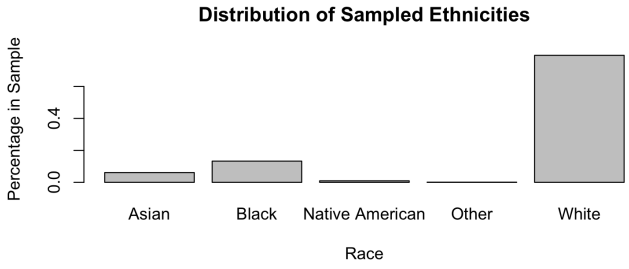
After doing all of this we finally built the survey with all the necessary variable using the "survey" package in R. This was the following line of code we used:

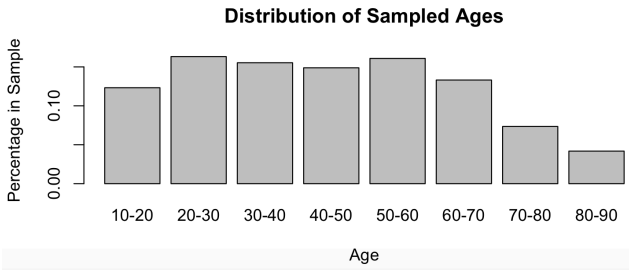
```
survey = svydesign(ids =~ cluster_id, strata =~ pseudo_stratum, weights =~ weights, data = clean_df, nest = TRUE)
```

Section 6 EDA

Our main goal in section is to check one thing only – does our data represent the target population. As stated above, we could do all analysis later on but realize it was a pointless analysis because we couldn't generalize it to the general U.S. 12 or above population. We will now check through the data whether this is true by comparing two covariates, age and race, that we know theoretical values from the U.S. Census data. Here are the following visualizations.

Figure 6.1 sample coverage of different race and age





So far this looks very promising. One can immediately see that the white population really dominates and the Asian and black populations are relatively small. Moreover, there seems to be the correct distribution in age. Note of caution: because this sample didn't choose anyone below 12 years old, the below 20 years old coverage should be much different than the theoretical one. Of course one can only tell if our sample is truly representative by comparing it to theoretical values that the U.S. census provides. This is exactly what the following table will do.

Table 6.1 sample vs theoretical coverage of race

	Sample Coverage	Theoretical Values (provided by U.S. Census)
Asian	0.061286759	0.056000000
Black	0.132882473	0.123000000
Native Americans	0.009663625	0.020000000
Others	0.001139475	NA
White	0.795027668	0.773500000

This shows that the NCVS sample really did do a good job in almost matching the correct coverage of each race, especially after their adjustment of weights. It is interesting to see how both Asians and Blacks are slightly higher than their theoretical values but also the whites too. Moreover, the Native Americans seem quite underrepresented in this sampling coverage. Now to compare the age distribution, according to the census bureau it states that those who are 15-64 cover 65.94% and 65+ covers 15.03%. From our sample it is evident that if we divide up 60-70 by half and add 70-80 and 80-90 we get around 15% therefore, showing that we seem to have the right distribution for age also.

In conclusion, from the EDA done on sampled ethnicities and sampled ages, it seems that we do indeed have a representative sample. This was quite comforting and also very expected especially from a big survey like NCVS to really get close to the target population.

Section 7 Analysis

The analysis will be organized in the following way. The first subsection will analyze general crime statistics in the United States without any consideration about covariates. Then we will talk about how these crime statistics might be associated with each of the following covariates. It is again very important to repeat this but the following analysis is not a causal inferential study but merely an associative study.

7.1 General Crime Statistics

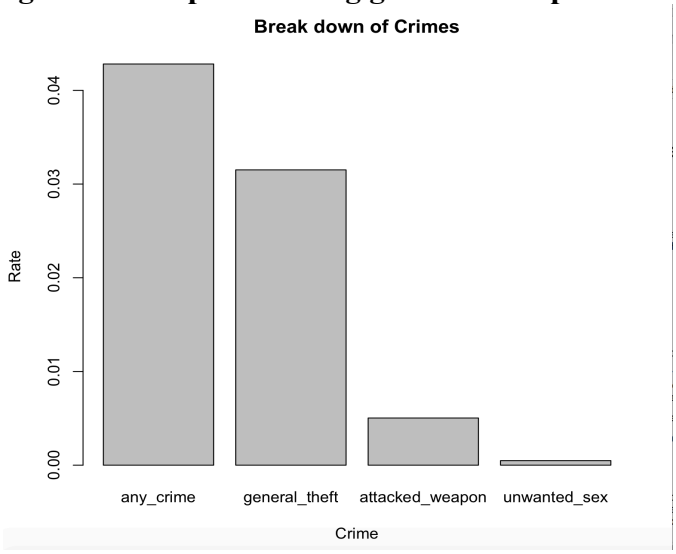
I will look at the four following variables: percentage of people in the United States who was a victim of any crime, this means of course any minor crime such as theft, etc. I will then also look at the percentage of people in the United States specifically who are affected by general theft, attacked by a weapon, and lastly rape/unwanted sex. Note all of this is basically crimes affected in the last six months. This is the following results.

Table 7.1 Showing crime rates for 12+ people in United States

	Estimated Rate	95% CI lower bound	95% CI upper bound
Any Crime	0.0428128610	0.0418107918	0.0438149301
General Theft	0.0315132032	0.0306485537	0.0323778527
Attacked with Weapon	0.0050323247	0.0046776146	0.0053870349
Rape/Unwanted Sex	0.0004853567	0.0003844735	0.0005862399

It is clear that general theft really covers up most of the crime statistic. It is also surprising that 4% of all Americans 12 years or older seems to be inflicted by some sort of crime in the last six months. Moreover, it seems sexual crime of 0.004% really isn't a very common crime when looking at the holistic rate of crime. Surprisingly, people who got attacked with a weapon seems to still be quite significant at a rate of 0.05%. All of this is easily shown in the following barplot.

Figure 7.1 Barplot showing general and specific crime rates

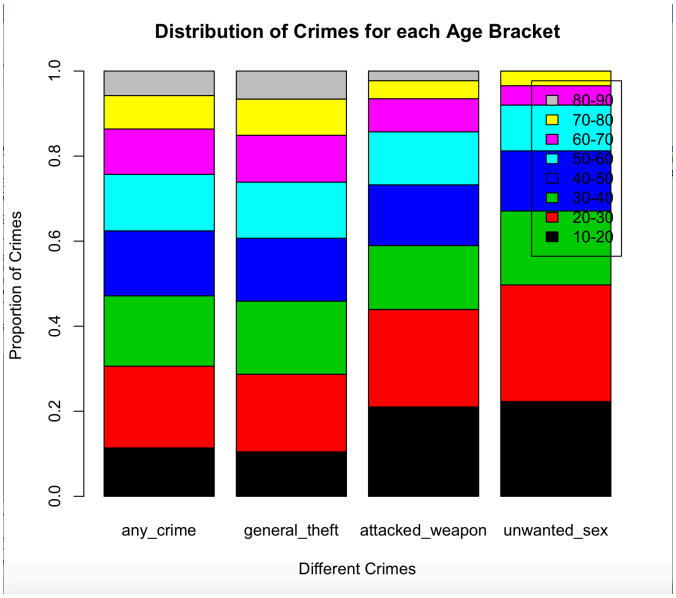


From here it is really evidential that everything I said above is true. We will not answer the more interesting questions of associations and correlation between each of these crimes and general crime statistics grouped into different covariates of race, age, etc. This is exactly what we will show next.

7.2 Crime statistics for Age

In this section we will show how each crime is broken down in terms of age brackets. We hope to see some associations that differ across each age bracket that we constructed in intervals of 10. We do not necessarily care about the actual values of crime in each bracket for example, we do not care if 0.05% of people between 10-20 got affected by general theft. What we care more about is among all the thefts, what percent of those general theft crimes is by those between 10-20 years old. This is exactly what the following barplot will show:

Figure 7.2 Barplot showing proportion of each crime per age bracket



The way to read this barplot is quite simple. Each area represents the proportion in that corresponding age bracket affected by that crime. No matter which crime category one looks at, the above figure really does show massive differences in each age bracket. It is very clear that younger people, especially from 20-40, have consistently much bigger areas and thus seems more likely to be affected by crime. It also seems quite evident that the older you get it seems the less likely one is to be affected by crime. Moreover, when looking at weapon attack and sex crimes, it seems people between 10-30 years old get more affected, which definitely makes sense in terms of social youth.

7.3 Crime statistics for Education

Another possible covariate that could be potentially interesting to study the distribution of crimes is education. The analysis was done in exactly the same way as above. Because we cared more about the proportion affected for each education bracket, we did not actually care about the true contingency table dividing it up into actual counts and yes/no for the crimes. Therefore, the following table will only represent the final proportion of those that were affected by crime broken up into the education brackets. The rest of the analysis will follow this logic.

Table 7.3 Breakdown of crimes among education brackets.

	Any Crime	General Theft	Attacked with Weapon	Unwanted Sex
High School or Lower	0.3142656	0.3027372	0.3545479	0.3662406
College	0.3699603	0.3733701	0.3507259	0.4115302
Higher Education (Ph.D. Masters)	0.3157740	0.3238927	0.2947263	0.2222292

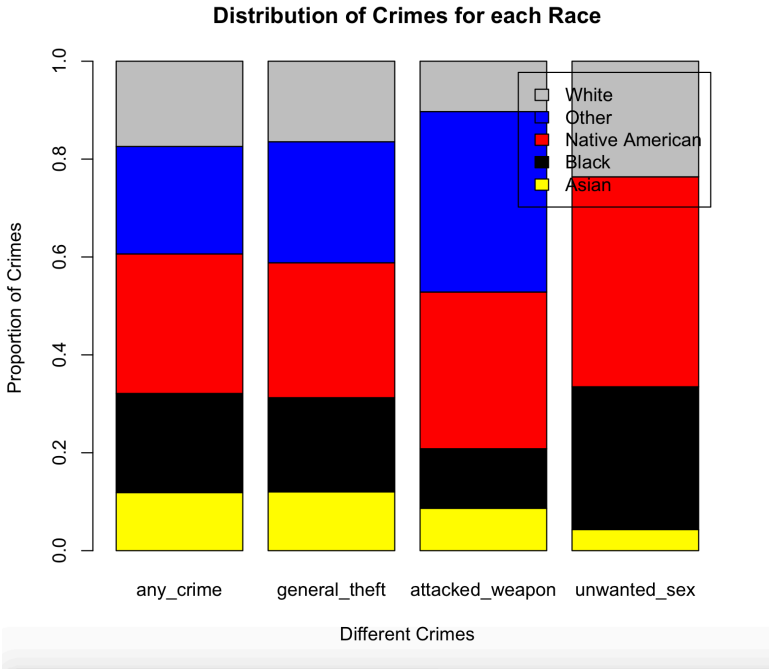
There definitely does seem to be significant differences in distribution for each education bracket but not necessarily in any obvious or consistent pattern like the age bracket above. For example, surprisingly people who went to only high school or lower is the lowest group affected with any sort of general crime in the past six months, but at the same time the highest group attacked with a weapon in the past six months is also those who have the lowest educational attainment. Moreover, some of these statistics are quite similar for two groups such as high school or lower and higher educated people seem to have very close statistics in any general crime.

However, it does not seem that there is necessarily any pattern or obvious social associations in here. This is not too surprising as there are no real obvious social reasons why people who are more or less educated are affected with more or less crime of a specific sort in the past six months. One never knows what the person is doing right now despite their educational backgrounds. Perhaps this data shows exactly that story.

7.4 Crime Statistics for Race

In such a modern time where racial issues and the goal for a colorblind society is pressing, it seemed like an unavoidable covariate to analyze crime statistics. Despite saying this over two times again, we want to emphasize that any inferential claims made here are associative. Because this is a sensitive topic, we do not want to make any racially inappropriate comment saying that because one is in a specific race they are more prone to be in a certain crime. We will just merely tell the story of associatiton. Here are the following results in a barplot.

Figure 7.3 Barplot showing different crimes for race

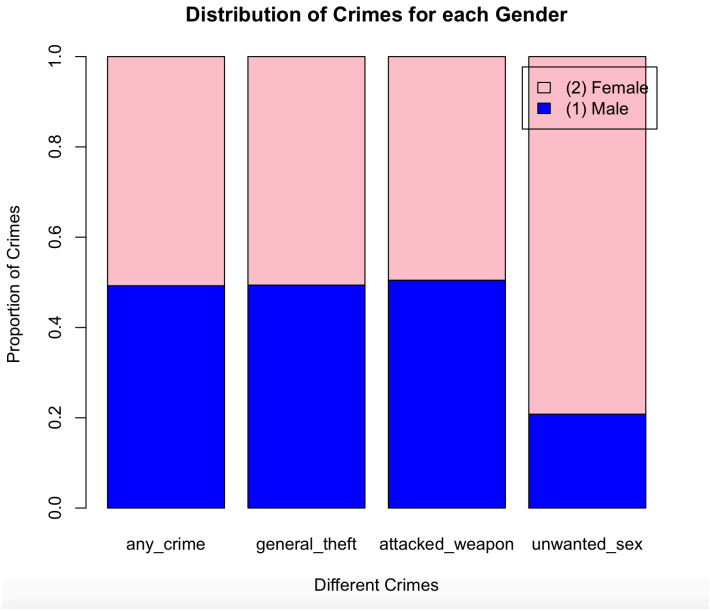


This plot, unlike the education covariate, does have a certain trend across any crime. The most obvious trend here is that Native Americans are consistently shown as the highest affected group of people for any sort of crimes. This is in fact not surprising. Although there is a lack of media attention on them, Native American reservation life is filled with poverty and thus consequently crime. Moreover, Asians and whites in no matter which category seems to be the lowest. The African Americans seems to be the second group to have the highest crime except for being attacked by a weapon. However, it is important to note that the "other" racial category really only had a few data point because these were really the multiethnic and unsure racial categories that were just grouped into one category. Another point worthy of mention is that in the crime category of rape, it's evidently clear that blacks seem to have quite a significant amount of rapes compared to their other crimes, while Asians in general have a significantly lower rate for being affected by rape.

7.5 Crime Statistics for Gender

The final and perhaps the most surprising part of the analysis is gender. One might always wonder if being a male or female might have a huge impact on certain crimes or crime in general. This is explored in this part of the analysis in the following barplot

Figure 7.4 Barplot showing different Crimes for Gender



This plot is by far the most interesting plot in all of the analysis. This is the only plot that every single category showed almost no significant difference between the groups except only one crime category – rape/unwanted sex. This is absolutely not surprising because definitely female in reality are way more likely to be raped than males. In fact, we will further explore this in chi-square tests and see if there were any significant differences between any general crime for gender and then for unwanted sex.

Table 7.4 Chi-square test for any crime between male and female

	Affected by any Crime	Not Affected by any crime
Male	5584645	126906058
Female	6067906	133615475

Chi-square test results: F = 1.7193, ndf = 1, ddf = 132610, p-value = 0.1898

Note these chi-square tests and tables were all created with the survey design considered from the survey package using the default method of Rao-Scott approximations. This was tested on a degree of freedom of 1 with an extremely high number of data points. Therefore because of the extreme amount of data points, the chi-square test should pick up any even small differences. However, despite that, one can clearly see that male and females do not have a significant difference for any general crime as p-value is clearly much bigger than 0.05. However, one cannot say the same for rape between male and female as the following will show.

Table 7.5 Chi-square test for unwanted sex/rape between male and female

	Affected by rape/unwanted sex	Not affected by rape/unwanted sex
Male	26326	132408481
Female	105716	139513219

Chi-square test results: F = 25.169, ndf = 1, ddf = 132610, p-value = 5.258e-07

This p-value is pretty much 0, and thus very significantly different. One does not really have to do a chi-square test to see that but it is very evidential that male and females do not really have a lot of different crime statistics in general but when dealing with rape, it is almost always disadvantaged for females. This social problem is quite obvious but the staggering statistics should be a gentle reminder to society such a discrepancy or existence is really not acceptable.

Section 8 Discussion and Conclusion

In general, through our EDA we learned that the NCVS sample is representative of the US population by the factors of race and age, which showed that the sampling frame really did well to match the target population. We firstly study the overall crime statistics in the United States from a holistic point of view. About 4.28% of Americans 12 years or old are affected by some sort of crime in the past six month, with general theft being the most common, contributing around 74% to all crimes. Attacks with weapons are also relatively significant, with an infliction rate at around 0.05%. Rape is the least common crime type. Around 1% of all crimes are reported as rape.

We then investigate the relationship between victimization rate and covariates: age, education, race and gender. This analysis was, as explained above, an analysis of association and not causation. Therefore, it is important to emphasize all our following conclusions are NOT arguments of causation but merely attempts of explanation. Overall, younger people tend more likely to be victimized than older people. In fact, younger people suffer the most from weapon attacks and rape, making up around 45% of the victimized population. This can be explained by the problems of social youth in the country and the sort of crimes that are related to them. Education, based on the statistics, does not seem to be an important factor in victimization. Although people with lower education (High School or Lower) have the lowest victimization rate, the differences are not significant.

Our findings in the two covariates of gender and race are rather interesting and socially significant. Among all racial groups, Native Americans are most likely to be affected by any crimes in every crime category. Despite the lack of media coverage, the social disadvantages experienced by Native Americans are truly astounding in and out of reservation land. This finding confirms the urgency for the government to divert resources to improve the lives of Native Americans. Asians and whites seem to be least affected by crimes in every category. The African Americans seems to be the second group to have the highest crime except for being attacked by a weapon. However, black people have a relatively high victimization rate in rape. This also deserves more attention in schools and work environment. At last, in terms of general crimes, gender does not seem to play a big role in victimization rate. However, women are much more likely to be affected by rape than men. This is in line with the depressing reality, and requires further efforts by both government and people in this country to resolve.

In conclusion, our findings give a rather grim view of victimization rates in the United States. As a developed country, America has a relatively high victimization rate, with more infliction among the disadvantaged and marginalized in our society. Moreover, according to different covariates such as race, age, and especially gender, some crimes are too statistically different among this factors. We hope our conclusions can be an aid in attracting more attention into those affected by crimes across the different factors. For example, rape among females are too high or general victimization rates are unbalanced across the young people. These need to be seriously considered by the government.

Appendix - R Code

Extract data/Cleaning data

```
#Get household data
load(file = "~/Downloads/ICPSR_36828/DS0002/36828-0002-Data.rda")
household_ncvs = da36828.0002
#stratums are in V2117
in_df = household_ncvs[, c("IDHH", "V2117")]
in_df = in_df[order(in_df$IDHH), ]
in_df = in_df[!duplicated(in_df), ]

#Get person data
load(file = "~/Downloads/ICPSR_36828/DS0003/36828-0003-Data.rda")
person_ncvs = da36828.0003
clean_df = person_ncvs[ , c("IDHH" , "V3002","V3014", "V3020", "V3018", "V3023A", "V308
1", "V3046", "V3034", "V3036", "V3042", "WGTPERCY")]
colnames(clean_df) = c("IDHH", "cluster_id","age","education", "gender", "race", "crime_
reports", "unwanted_sex", "general_theft", "broken_in", "attacked_weapon", "weights")
clean_df = merge(x = clean_df, y = in_df, by = c("IDHH"))
```

Item-Nonresponse - Imputations

```
#Trying to check how many nonresponse is unit and item
N = nrow(clean_df)
num_NA = function(x){
  sum(is.na(x))
}

for (i in 7:10) {
  clean_df[, i][c(which(clean_df[, i] == "(8) Residue"))] = NA
}

table(clean_df$unwanted_sex)
table(clean_df$general_theft)

#Taking only the rows without the unit nonresponses to do imputations
index = which(is.na(clean_df$unwanted_sex) & is.na(clean_df$general_theft) & is.na(clean
_df$broken_in) & is.na(clean_df$attacked_weapon))

clean_df = subset(clean_df, select = -c(broken_in))
clean_df = clean_df[-index, ]

#imputing with mice package
library("mice")
interested_df_1 = clean_df[, c(2, 3, 4, 5, 8)]
head(interested_df_1)
imputed_df_1 = mice(interested_df_1)

interested_df_2 = clean_df[, c(2, 3, 4, 5, 9)]
head(interested_df_2)
imputed_df_2 = mice(interested_df_2)

interested_df_3 = clean_df[, c(2, 3, 4, 5, 10)]
head(interested_df_3)
imputed_df_3 = mice(interested_df_3)

a = complete(imputed_df_1); b = complete(imputed_df_2); c = complete(imputed_df_3)

a$general_theft = b$general_theft
a$attacked_weapon = clean_df$attacked_weapon
a$weights = clean_df$weights
a$pseudo_stratum = clean_df$V2117
a$cluster_id = clean_df$cluster_id
a$race = clean_df$race; View(a)
a$crime_reports = clean_df$crime_reports
clean_df = a
```

Cleaning up variables

```
#Clean up education data into factors we care about
edu_digits = substr(as.character(clean_df$education), 2, 3)
edu_digits = as.numeric(edu_digits)
edu_digits[edu_digits <= 12 | edu_digits == 28 | edu_digits == 27] = "High School or Low
er"
edu_digits[(edu_digits > 12 & edu_digits <= 27) | edu_digits == 40 | edu_digits == 41 |
 edu_digits == 42] = "College"
edu_digits[edu_digits == 43 | edu_digits == 44 | edu_digits == 45] = "Higher Education"
clean_df$education = factor(edu_digits)
clean_df$education[clean_df$education == 98] = NA
clean_df$education = factor(clean_df$education)

clean_df$gender = factor(clean_df$gender)

#Make age into factor
clean_df$age[clean_df$age < 11] = 1
clean_df$age[clean_df$age <20 & clean_df$age >= 11] = 2
clean_df$age[clean_df$age <30 & clean_df$age >= 20] = 3
clean_df$age[clean_df$age <40 & clean_df$age >= 30] = 4
clean_df$age[clean_df$age <50 & clean_df$age >= 40] = 5
clean_df$age[clean_df$age <60 & clean_df$age >= 50] = 6
clean_df$age[clean_df$age <70 & clean_df$age >= 60] = 7
clean_df$age[clean_df$age <80 & clean_df$age >= 70] = 8
clean_df$age[clean_df$age <= 90 & clean_df$age >= 80] = 9
clean_df$age = factor(clean_df$age, labels = c("10-20", "20-30", "30-40", "40-50", "50-6
0", "60-70", "70-80", "80-90"))


#Clean up race variable into factors we care about
race_digits = substr(as.character(clean_df$race), 2, 3)
race_digits = as.numeric(race_digits)
#1 white 2 black 3 native american 4 asian 5 Other
race_digits[race_digits == 7 | race_digits == 9 | race_digits == 1] = "White"
race_digits[race_digits == 6 | race_digits == 10 | race_digits == 11 | race_digits == 12
 | race_digits == 2] = "Black"
race_digits[race_digits == 5 | race_digits == 13 | race_digits == 3] = "Native American"
race_digits[race_digits == 8 | race_digits == 14 | race_digits == 4] = "Asian"
race_digits[race_digits == 15 | race_digits == 16 | race_digits == 17 | race_digits == 1
8 | race_digits == 19 | race_digits == 20] = "Other"
clean_df$race = factor(race_digits)

#Clean up crime reports only want basically yes or no 0 or 1
clean_df$crime_reports[clean_df$crime_reports > 0] = 1

#Make the responses more easy to interpret

clean_df$unwanted_sex = as.numeric(clean_df$unwanted_sex)
clean_df$unwanted_sex[clean_df$unwanted_sex == 2] = 0
clean_df$unwanted_sex[clean_df$unwanted_sex ==3] = NA

clean_df$general_theft = as.numeric(clean_df$general_theft)
clean_df$general_theft[clean_df$general_theft == 2] = 0
clean_df$general_theft[clean_df$general_theft ==3] = NA
```

```
clean_df$attacked_weapon = as.numeric(clean_df$attacked_weapon)
clean_df$attacked_weapon[clean_df$attacked_weapon == 2] = 0
clean_df$attacked_weapon[clean_df$attacked_weapon ==3] = NA
clean_df$any_crime = clean_df$crime_reports
save(clean_df, file = "cleanedup_df")
```

Survey design

```
library("survey")
survey = svydesign(ids =~ cluster_id, strata =~ pseudo_stratum, weights =~ weights, data
 = clean_df, nest = TRUE)
```

EDA does our data look alright?

```
####Figure 3.1####
summary(person_ncvs$WGTPERCY)
histogram(person_ncvs$WGTPERCY, breaks = 50, xlab = "Adjusted Weights", main = "Distribu
tion of Weights")

#boxplots about weight
par(mfrow = c(1, 1))

####Figure 3.2####
boxplot(weights~race, data = clean_df, xlab = "Race", ylab = "Weights", main = "Boxplot
 of Weights on Race")

#EDA

####Figure 6.1####
par(mfrow = c(2,1))
a = svytable(~race, survey)/sum(svytable(~race, survey))
barplot(a, xlab = "Race", ylab = "Percentage in Sample", main = "Distribution of Sampled
 Ethnicities")

b = svytable(~age, survey)/sum(svytable(~age, survey))
barplot(b, xlab = "Age", ylab = "Percentage in Sample", main = "Distribution of Sampled
 Ages")

####Table 6.1####
#Creating tables comparing theoretical values
t_1 = as.table(as.matrix(data.frame(sample_estimate = as.numeric(a), theoretical = c(0.0
56, 0.123, 0.02, NA, 0.7735)))); row.names(t_1) = row.names = c("Asian", "Blacks", "Nati
ve Americans", "Others", "Whites"); t_1
svytable(~race, survey)
confint(svytable(~race, survey))

as.table(as.matrix(data.frame(sample_estimate = as.numeric(b))))
```

Analysis of general crime rates

```
#First question What is the general crime rate in the United States
general_a = svymean(~any_crime, survey, na.rm = TRUE)

general_b = svymean(~general_theft, survey, na.rm = TRUE)

general_c = svymean(~attacked_weapon, survey, na.rm = TRUE)

general_d = svymean(~unwanted_sex, survey, na.rm = TRUE)

####Table 7.1####
a_1 = as.table(as.matrix(data.frame(estimated_rate = c(general_a[1], general_b[1], gener
al_c[1], general_d[1]), lower_CI = c(confint(general_a)[1], confint(general_b)[1], confi
nt(general_c)[1], confint(general_d)[1]), upper_CI = c(confint(general_a)[2], confint(ge
neral_b)[2], confint(general_c)[2], confint(general_d)[2])))); a_1

####Figure 7.1####
par(mfrow = c(1,1))
estimated_rate = c(general_a[1], general_b[1], general_c[1], general_d[1])
barplot(estimated_rate, xlab = "Crime", ylab = "Rate", main = "Break down of Crimes")
```

Age Analysis

```
age_a = svyby(~any_crime, ~age, survey, svymean)

age_b = svyby(~general_theft, ~age, survey, svymean, na.rm = TRUE)

age_c = svyby(~attacked_weapon, ~age, survey, svymean, na.rm = TRUE)

age_d = svyby(~unwanted_sex, ~age, survey, svymean, na.rm = TRUE)

age_table = as.table(as.matrix(data.frame(any_crime = getProp(age_a[2]), general_theft =
 getProp(age_b[2]), attacked_weapon = getProp(age_c[2]), unwanted_sex = getProp(age_d[2
])))))

####Figure 7.2####
barplot(age_table, legend.text = TRUE, xlab = "Different Crimes", ylab = "Proportion of
 Crimes", main = "Distribution of Crimes for each Age Bracket", col = 1:8)
```

Education analysis

```
education_a = svyby(~any_crime, ~education, survey, svymean)

education_b = svyby(~general_theft, ~education, survey, svymean, na.rm = TRUE)

education_c = svyby(~attacked_weapon, ~education, survey, svymean, na.rm = TRUE)

education_d = svyby(~unwanted_sex, ~education, survey, svymean, na.rm = TRUE)
#Make into proportion
getProp = function(vec) {
  vec/sum(vec)
}

####Table 7.3####
education_table = as.table(as.matrix(data.frame(any_crime = getProp(education_a[2]), gen
eral_theft = getProp(education_b[2]), attacked_weapon = getProp(education_c[2]), unwante
d_sex = getProp(education_d[2]))))
education_table
```

Race analysis

```
race_a = svyby(~any_crime, ~race, survey, svymean)

race_b = svyby(~general_theft, ~race, survey, svymean, na.rm = TRUE)

race_c = svyby(~attacked_weapon, ~race, survey, svymean, na.rm = TRUE)

race_d = svyby(~unwanted_sex, ~race, survey, svymean, na.rm = TRUE)

race_table = as.table(as.matrix(data.frame(any_crime = getProp(race_a[2]), general_theft
 = getProp(race_b[2]), attacked_weapon = getProp(race_c[2]), unwanted_sex = getProp(race
_d[2]))))

####Figure 7.3####
barplot(race_table, legend.text = TRUE, col = c("yellow", "black", "red", "blue", "grey"
), xlab = "Different Crimes", ylab = "Proportion of Crimes", main = "Distribution of Cri
mes for each Race")
```

Gender analysis

```
gender_a = svyby(~any_crime, ~gender, survey, svymean)

gender_b = svyby(~general_theft, ~gender, survey, svymean, na.rm = TRUE)

gender_c = svyby(~attacked_weapon, ~gender, survey, svymean, na.rm = TRUE)

gender_d = svyby(~unwanted_sex, ~gender, survey, svymean, na.rm = TRUE)

gender_table = as.table(as.matrix(data.frame(any_crime = getProp(gender_a[2]), general_t
heft = getProp(gender_b[2]), attacked_weapon = getProp(gender_c[2]), unwanted_sex = getP
rop(gender_d[2]))))

####Figure 7.4####
barplot(gender_table, legend.text = TRUE, col = c("blue", "pink"), xlab = "Different Cri
mes", ylab = "Proportion of Crimes", main = "Distribution of Crimes for each Gender")

#Chi-square tests
####Tables 7.4 and Table 7.5####
svytable(~any_crime + gender, survey)
svychisq(~any_crime + gender, survey)

svytable(~unwanted_sex + gender, survey)
svychisq(~unwanted_sex + gender, survey)
```