Final Project Report

Based on the *National Crime Victimization Survey, 2016*

STAT 152

Professor: Elizabeth Purdom
GSI: Amanda Glazer

Team Members:
• Jinbo Wu 3033263170
• Yijun Xu 3033385708
• Anna Lu 3033352649

Group 4

Content:

Introduction

The project is based on 2016 National Crime Victimization Survey. This survey is to collect information of criminal incidents against household and personal around the United States to measure victimization from personal crimes and property crimes.

The complete survey data contains 5 parts of datas: address record data, personal record type, household record type, incident record type and the 2016 collection year incident-level extract file. The 2016 Collection Year Incident-Level Extract file consists of all household and personal incidents recorded in the survey in 2016. Since the file also collects many characteristics of the households and personals, this project is to find what kind of values of which characteristics would be likely to affect the chance of incidents happened to the household. After running the Chi-Square test and finishing the analysis of the survey, the result is that the income level, number of people older than 12 and younger than 12 years old, number of cars of the household are related to the chances that the household becomes a target of crimes.

Survey Design

The public data of the survey is distributed by Inter-university Consortium for Political and Social Research at Ann Arbor, Michigan. This survey is a broadly representative survey on the whole country. The design and the process of the survey are supervised by the U.S. Census Bureau (under the U.S. Department of Commerce) on behalf of the Bureau of Justice Statistics (under the U.S. Department of Justice). Since 1973, the National Crime Victimization Survey (below called NCVS) has taken place every 6 months. NCVS recorded many details about the incidents happened on personals and households such as when and where the incident happened, what kind of incident that was, was there any property lost or if anyone got hurt, if the victim

reported the incident to the police, etc. And also the domestic information such as income, living area and household structure of the offenders and victims were also recorded. The initial purpose of NCVS was to estimate the number and types of incidents that were not reported to the police, to obtain details about the incidents and to do comparisons on these crimes at different regions.

The U.S. Census Bureau takes a population census every 10 years. The population data of NCVS, 2016 is based on the current census which was taken in 2010. Therefore, the NCVS from 2006 to 2016 should be considered under different estimates about the trends and numbers. In 2006, the survey was taken under computer assisted instead of paper to pencil. In 2007, in order to minimize the spend on money and human resources, the sample of survey was reduced by 14% and computer assisted interviews at telephone centers were no longer in use. But in 2010, the U.S. Census Bureau increased the sample size by 24% to improve the precision of the survey. More precisely, increase from 8500 households per month to 10,500 per month. Besides, after the 2010 census, in order to reflect the changes in the population census and to make a more precise state and local level estimates of crime victimization rate for the 22 largest states and their metropolitan areas, the changes included new samples, new counties in the samples, etc. But that also made the NCVS of 2016 not comparable to other years.

From 2008, the data files contain respondent information which are unbounded. That means every household in the survey is newly sampled and could not be a in the next sample again in the next 6 months. This method is used for selecting samples called a rotating panel with a period of 3 years with total 7 interviews including the first time. Each year, NCVS is a complex survey with multi-stage sampling schemes. The first step is to divide the United States into 1,987 counties or county equivalent areas. These are our PSUs for the survey(There is an exception for

Alaska because Alaska contains counties with small populations but large land areas. The minimum population for a PSU in Alaska is 7,500 and maximum area is 3,000 square miles) and the process is done once every 10 years. Second, these 1,987 counties are divided into sampling stratas. During this step, some large PSUs with large population and large land areas are grouped as "self-representing PSU(SR PSU)" while other PSUs with homogeneous characteristics within a state are grouped into "non-self-representing PSU(NSR PSU)". The sample selection of PSU must contain SR PSUs from the SR strata since they are heavily representative and one PSU would be selected from the NSR PSU strata under probability proportional programming algorithm.

For this survey, there is a special rotating panel design. As mentioned before, the survey interviewed people in the sample every six-month. In order to avoid the situation where a household is interviewed repeatedly, the special rotation panel design is set to divide the sampling household into six different rotation groups, and each group would be interviewed once every 6 month for 3 year. What's more, if a new household member moves to one of the selected counties, the member will also be interviewed. That implies the number of incidents in the survey would be affected by the new come household. In order to minimize the non-response during the survey, the target household would receive an informing letter from the Bureau of Census about the survey. NCVS consists of 2 parts: a face to face interview for the first time and phone interviews for the rest of the survey. Questions for households are expected to be answered by the household representative, other personal questions are asked to all household members aged more than 12. For those non-English speakers, the interview will be conducted in another language. However, non-response can not be totally avoided. There are three types of

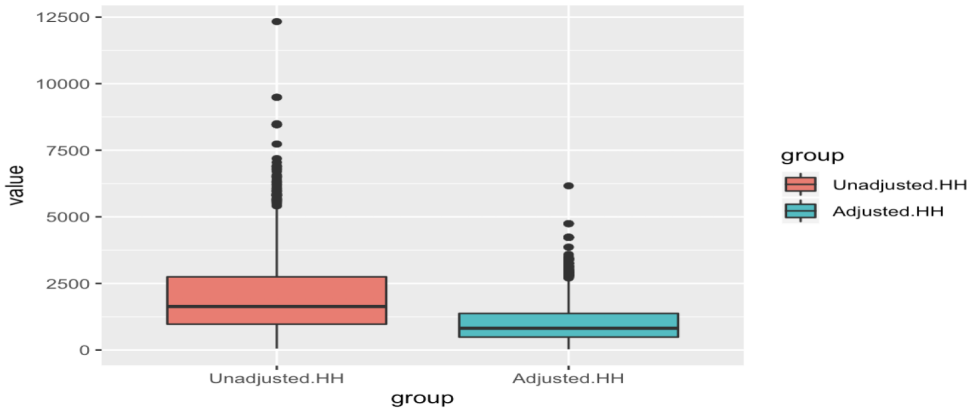
non-responses: type A is the situation which no one could be found in the household after multiple visits or the members refuse to be interviewed; type B is that the owner of the house or building has other usual residence address; type C is that the address is ineligible to be sampled.

During the data collection process, invalid responses might be recorded or the respondent chose not to answer the whole question or a specific subquestion. Under such a circumstance, these invalid responses are marked “Residue” in the file. As for some of the variables, the interviewers’ response and answers are collected but they are not suitable for putting into the survey. Then these responses and answers are collected under the original variables and these responses are being transferred into a valid form of responses into the corresponding “Allocated variables" of the “Original variables” by applying the Census Bureau recoding scheme.

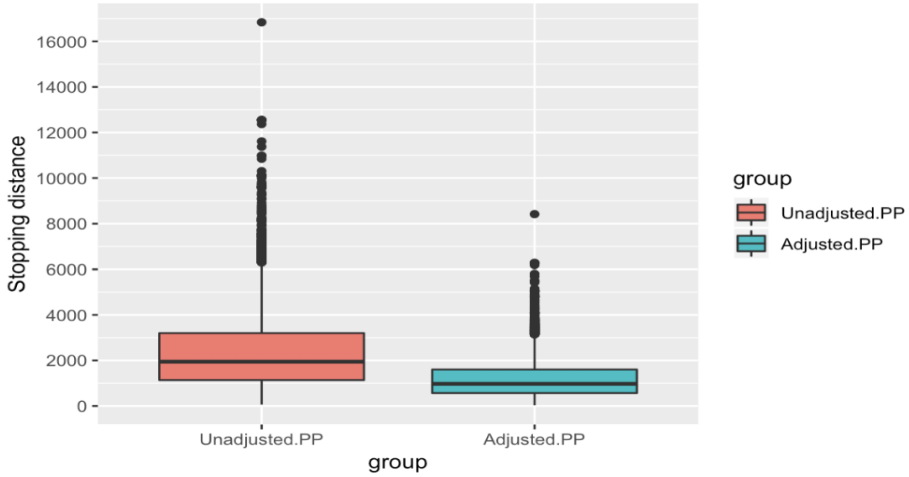
There are three types of major public files: personal record type, household record type and incident record type. Another 2016 Collection Year Incident-Level Extract File was built based on these 3 files with the variable YEARQ, IDHH, IDPER as the match keys. For household weights, the weights are calculated based on the “Principal Person” of that household. If the household is a husband-wife household(with no other people in the family), the weight is the weight of the wife. Besides, the weight is the weight of who(called “Reference Person” in the survey) owning, buying or renting the house. There are two types of incident weights in the incident record type file, personal and household. As for personal crime, the weight is derived by dividing the person weight of the victim in the crime by the total number of people suffering from the crime. The property crime weight is the same as the household weight. Within the household record type file, the variable V2116 is the household weight and the variable V3080 is the person weight which can be used for calculating the total adult population and the number of

households at the time the survey was constructed. The variable WGTHHCY is the adjusted household weight for the collection year(in this project the year is 2016), which can be the denominator while calculating the household victimization rate. Similar to the WGTHHCY, WGTPERCY is the adjusted person weight. There is also an incident weight in the incident record-type file. It is calculated by using the victim's weight divided by the total number of injuries reported by the defendant to represent the personal crimes, while using the household weight to represent the property crimes incident weight. What's more, there is a set of household replicate weights, which is located in the Household Record-Type File. This weight is only used to do the test.

The designers of the surveys give two kinds of weight, household weight and person weight, unadjusted and adjusted. The figures below show the difference between unadjusted and adjusted household weight. From the graphs, the medians of the adjusted weight are about half of the unadjusted weight, which means the estimate will be more precise by using adjusted weight instead of the unadjusted weight. Also the boxes of the unadjusted household weight are longer than the adjusted household's, which is the reason why adjusted weight is needed, since the longer boxes and whiskers indicate greater variability. There are also several outliers above the upper whisker. The range of unadjusted households is about 0 to 12500, while the range of adjusted households is about 0 to 3750, which is much smaller. Also, the range of unadjusted person weight is 0 to 17000, and the adjusted person weight is 0 to 8500.



<figure 1>



<figure 2>

Methodology

The complete data files are available for download at https://doi.org/10.3886/ICPSR36828.v1. The data file used for this project is called 2016

Collection Year Incident-Level Extract File. The 1,175 variables contains domestic informations such as income, household sex, household address, the time household being interviewed, etc. Also, the file contains other information such as the number of times incidents happened to the household, the type of the incident(personal or property), date and time when the incident happened, the offenders' geographic information and other detailed answers in the survey. During this project, not all variables are used and analysed since the purpose of this project is to determine the relationships between the chances of incidents happening to households and some of the households' domestic information.

Instead of directly working on the file, some cleaning job must be done before putting hands on it. The variables of interest used for the project are V2113(number of incidents), V2025A(living in gated or walled communities), V2026(household income intervals), V2071(number of members aged 12 and above), V2072(number of members aged 12 and below), V2078(number of vehicles owned), V2116(household unadjusted weight), WGTHHCY(household adjusted weight), SERIES_IWEIGHT(adjusted incident weight) and 160 replicate weights columns. Correspondingly, these variables are renamed as 'HH.ID', 'num.in', 'g.or.w', 'hh.income', 'num.age.L12', 'num.age.S12', 'num.car', 'hh.weight', 'WGTHHCY', 'Ad.in.wei', 'HHREPWGT1' ~ 'HHREPWGT160'

In part 2 Survey Design, if an entry is defined invalid for a question, it would be marked "Residue" in the data set, as known as non-response data. The variable 'hh.income' contains 1,497 'Residue' rows without providing the household income. However, in this column, residues are grouped together as one of the types of households because these households provided other information except the household income. This is treated as a categorical missing

data. Without predicting and implementing the NAs, there would be bias when working on the Chi-Square test and lead to an incorrect conclusion. Besides, if NAs were deleted without any adjustment on the data, the weight would not reflect the correct population. In order not to be confusing, the "(98) Residue" is modified to be an empty space and then put it back to the dataset, so that there will be a real NA showing in the dataset. Next step is to predict the values for these empty cells. The 'mice' package is used for implementing values to the NAs in the column. The missing values and categories are predictable because the other information of these households are available, and these will be the predictors for the missing values. Those 1,497 missing categorical data were categorized to different household income intervals.

According to the codebook, the Household Record-Type File contains replicate weights for each household. There are totally 160 replicate weights for each household. These weights are extracted as a 160 x 9,164 (total 9,164 unique households in the incidents extract file) matrix and merged into the data frame with variables of interest, from column 11 to column 170. In order to use the Chi-Square test to test the dependency between variables, changing numerical variables to character variables makes the test more precise. The Chi-Square test requires these variables to be characteristic data and then observed their frequencies to calculate the P-value. The variables of interests 'num.in', 'hh.income' and 'num.car' were inverted to characters. Since there are zero values in the table of variables "num.in" and "hh.income", that makes it not available to do chi-square tests. It is necessary grouping them into a smaller number of categories, which makes every cell have at least 5 observations. Variable "num.in" has categories 1 to 12, and value 1 and 2 have the highest frequency. The other values only take a small partial of the total frequency. Therefore, all of the original categories are grouped into two groups. One

includes values below mean(1.34), and is assigned as “1”. The other includes values above mean, and is assigned as “2”.

The observing dataset contains the columns of interest and 4 kinds of weight, household weight, adjusted household weight, adjusted incident weight and replicate weight namely. Interested variables are V2113(number of incidents), V2025A(living in gated or walled communities), V2026(household income intervals), V2071(number of members aged 12 and above), V2072(number of members aged 12 and below), and V2078(number of vehicles owned). Replicate weights should be used for running the Chi-Square test and variance estimation according to the codebook. Therefore, the ‘svrepdesign’ was used instead of the ordinary ‘svydesign’. For the variables argument in the below code, the data frame only contains the columns of interest. Using the sampling household weight and replicate weight data frame to satisfy the weights and repweights argument.

```
########################
#Using the svrepdesign
#chisq.d <- svrepdesign( variables = data_1[,-c(8:170)],
#                weights = data_1$hh.weight,
#                repweights = data_1[,c(11:170)] )
########################
```

Results

Many houses in the United States are single houses with an independent area. Unlike any townhouse community and apartment buildings, these areas are not protected by gates and walls

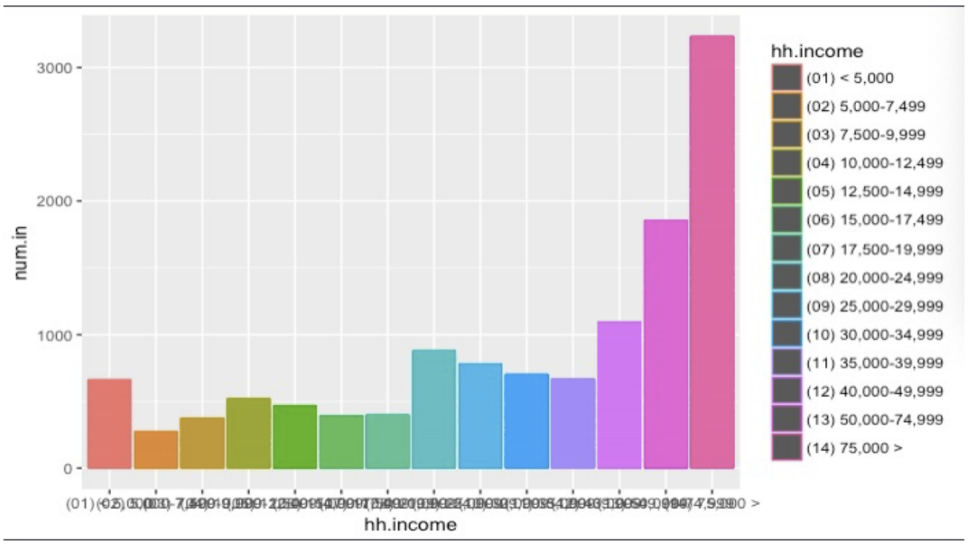
unless the owner of the property builds it for his or her own. In our survey data, 91.7% of incidents happened in households without gates or walls. No doubt, the Chi-Square result would have a reasonably small P-value to reject the null(the two variables are independent) which proves there is a significant relationship between them. This is a dummy variable to the analysis because the household can only have 2 categories: yes or no. An obvious conclusion, as well as the common sense proved, is to say a household without gates and walls around the property would have great impact on the chances of being a target.

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~g.or.w + num.in, chisq.d)
## F = 28.103, ndf = 1, ddf = 159, p-value = 3.791e-07
```

<figure 3: *Result showing the Chi-Square test*>

The variables of interest are the number of cars in a household (num.car), the total income level of a household (hh.income), the number of residents in a household that larger than 12 years old (num.age.L12), the number of residents in a household that younger than 12 years old (num.age.S12), the household has gate or wall (g.or.w) and the number of incidents that a household had (num.in). The principle variable is the number of incidents of a household. Since the purpose of this project is to find the relationship between characteristics of household and the number of incidents, then the null hypothesis is that other variables are expected to have no relationship with the principle variable.

In common sense, income has a strong relationship with crimes. For people who with higher income might have higher chances of being the target of property crime. There are about 9164 households that had at least one incident in 2016 from the 200,000 households, which are sampled by cluster sampling method from the United States. Overall, the share of higher-income households that have incidents is significantly higher than that of lower-income households. In particular, the last two categories, those with incomes of 50,000 to 75,000 and those with incomes above 75,000 make up a very large proportion. Also, those with incomes above 75,000 are basically twice as large as those with incomes between 50,000 and 75,000, and it is 27.6% of the total population of families involved in the incidents. Therefore, here comes the assumption: The chance of a household getting incident is related to household income level.



<figure 4>

After applying the Chi-square test to the incidents number and the household income columns, the P-value is minuscule, which means that the incidents number does have the significant relationship with household income level. Therefore, higher income level can be one of the reason that the household has incident.

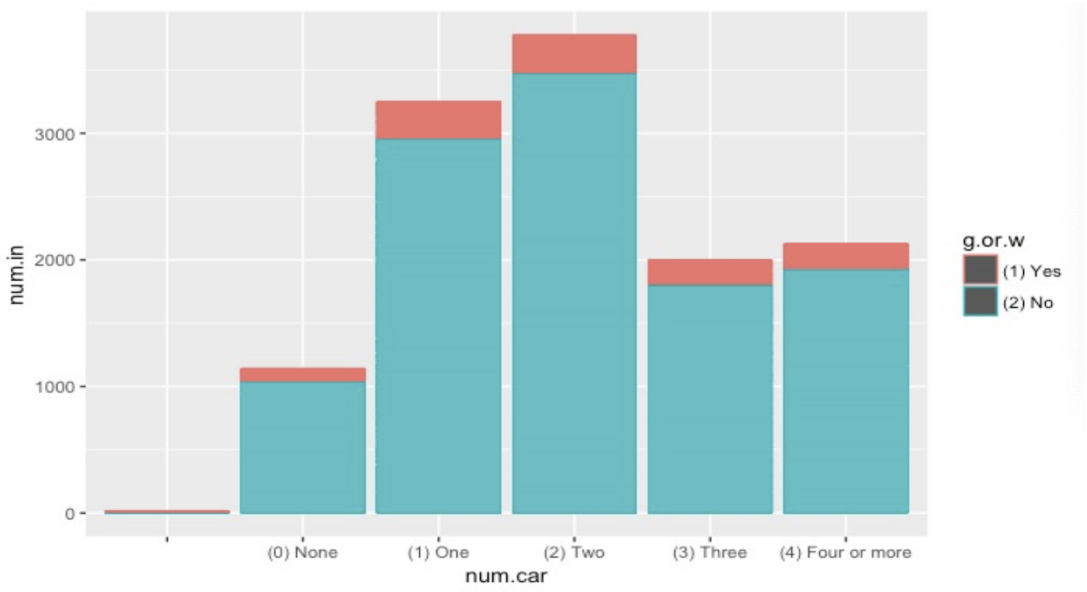
```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~hh.income + num.in, chisq.d)
## F = 5.9342, ndf = 11.547, ddf = 1836.000, p-value = 6.28e-10
```

<figure 5>

To some extent, the number of cars in a household relates to the income of the household. It makes sense to assume the relationship between the number of cars in a household and the number of incidents is similar to the relationship between the household income and the number of incidents. There are 5 categories in variable “num.car”, which are none, one, two, three, and four and more. After seeing them deeper by applying the Chi-square test, the result is to reject the null since the p value is around 0.004 and it is much smaller than 0.05. The result implies that these two variables have a relationship, but they are neither positive nor negative related. Based on the fact that the highest frequency occurs in the households with two cars and the household with one car no matter what the number of incidents is. Although 0.004 can theoretically imply the number of cars of a household and number of incidents are dependent, compared to the p-values from the chi-square tests done between the other interested variables and number of incidents, 0.004 is the biggest p-value and it is much higher than the others. That means the number of cars in a household has the weakest relationship with the number of incidents among all the variables. The conclusion about the relationship of the number of cars owned and the incident is that households with one or two cars would be the majority of the target in a crime.

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~num.car + num.in, chisq.d)
## F = 3.8714, ndf = 3.8677, ddf = 614.9600, p-value = 0.004545
```

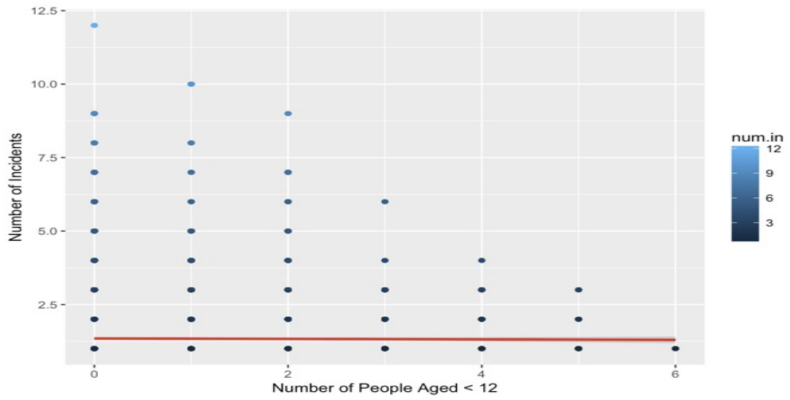
<figure 6>



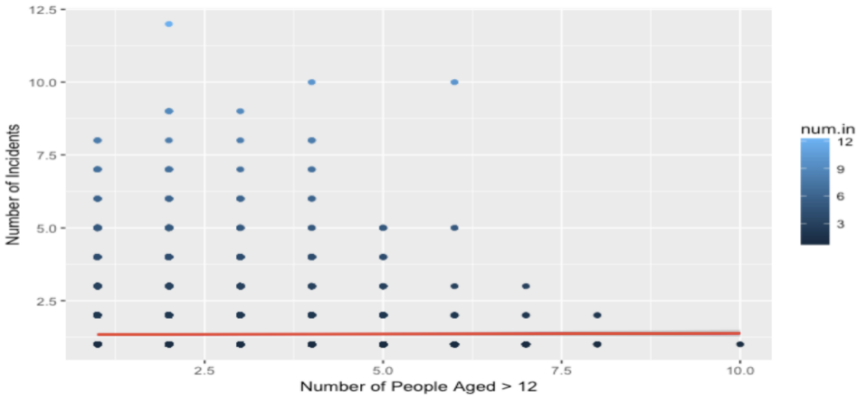
<figure 7>

There are two variables measuring the number of household members which are aged greater than 12 and younger than 12. The Chi-Square test between the number of incidents and the number of household members older than 12 and number of household members younger than 12 showed that these variables are strongly related together. The relationship is that if the greater the number of household members older or younger than 12, the chances of being a victim of a criminal incident is smaller. In general, the number of incidents of all households in this survey is between 1 and 2. In other words, since all the households had been the victim of

criminal incidents at least once, it makes more sense to focus on the distribution of the number of household members older or younger than 12. This number would not really affect the number of incidents of households because a linear relationship does not exist among the two variables in the survey data. The number of households older or younger than 12 can only be related to the chances a household would be the target of criminals. But the issue was that the type of incidents were ignored. There are many kinds of crimes that could be connected to children such as kidnapping, domestic violences and pedophilia. The result can only be a roughly reference to the distribution and can never be a tool for estimating the number of crimes.



<figure 8>



<figure 9>

Conclusion

In a word, one household's income, number of vehicles and the number of household members living together would have effects on whether a household would become a target to the criminals, but the result is not implying that the exact number of incidents have relation to these variables. The only variables that have a positive relationship to the incidents is the household income. The survey data can be used for more deeper studies such as estimating the number of different types of crimes in the country. Since the survey data file for this project only contains the households with at least one incident, the estimated proportion of the households encountered with incidents could not be calculated. The emphasis of the variables picking process is the household's homographic information. Therefore the variables about the details of the incidents are filtered. The conclusion of this project can only be used for references to the probabilities of a household being the target for criminals but could not decide what kind of crimes and how many times of such a crime the household would encounter. For further research, people can go deep to think more about personal crime, such as what characteristic of a person does affect the chance or the number of times of personal crime incidents occur.

Code Appendix

```
################################################################
```

Loading the data files and picking variables of interest

```
#load("NationalCrime.rda")

#load("HH_type")

#library(survey)

#library(dplyr)

#data=select(da36828.0005,IDHH,V2113,V2025A,V2026,V2071,V2072,V2078,V2116,

WGTHHCY,SERIES_IWEIGHT) ## Adjust dataset

#colnames(data)[1] <- "HH.ID" ##household ID in the survey

#colnames(data)[2] <- "num.in" ##number of incidents recorded

#colnames(data)[3] <- "g.or.w" ##gated or walled community

#colnames(data)[4] <- "hh.income" ##household income

#colnames(data)[5] <- "num.age.L12" ##number of members aged 12 and above

#colnames(data)[6] <- "num.age.S12"##number of members aged 12 and younger

#colnames(data)[7] <- "num.car" ##number of cars owned

#colnames(data)[8] <- "hh.weight" ##household weight

#colnames(data)[9] <- "WGTHHCY" ##adjusted household weight

#colnames(data)[10] <- "Ad.In.Wei" ##adjusted incident weight

################################################################
```

Printing the boxplot for the household unadjusted weight and adjusted weight, figure 1

```
#a = data.frame(group = "Unadjusted.HH", value = da36828.0005$V2116)
```

```
#b = data.frame(group = "Adjusted.HH", value = da36828.0005$WGTHHCY)

#plot.data = rbind(a,b)

#ggplot(plot.data,aes(x=group,y=value,fill=group))+geom_boxplot()+

scale_y_continuous(name="Household Weight", breaks = seq(0,200000,2000))
```

Printing boxplot for the personal adjusted and unadjusted weight, figure 2

```
#c = data.frame(group = "Unadjusted.PP", value = da36828.0005$V3080)

#d = data.frame(group = "Adjusted.PP", value = da36828.0005$WGTPERCY)

#plot.data = rbind(c,d)

#ggplot(plot.data,aes(x=group,y=value,fill=group))+geom_boxplot()+

scale_y_continuous(name="Personal Weight", breaks = seq(0,200000,2000))

#########################################################################
```

Implementing NA in the household income column.

```
#library(mice)

#data_1<-data[,c(-1)]

#data_1$hh.income[data_1$hh.income=="(98) Residue"]<-""

#h<-mice(data_1,m=5, method = "cart")

#completed_not_Na<-complete(h,1)

#########################################################################
```

Implementing NA in the number of cars in a household column.

```
#data_2<-completed_not_Na[,-c(7:9)]

#completed_not_Na$num.car<-as.character(completed_not_Na$num.car)

#data_2$num.car[data_2$num.car=="(8) Residue"]<-""
```

```
#k<-mice(data_2,m=5)

#################################################################
```

Combining the replicate weights with the data frame

```
#da36828.0002<-da36828.0002[!duplicated(da36828.0002$IDHH), ]

#RepliW_1 = da36828.0002[,c(-1,-2,-4:-131)]

#colnames(RepliW_1)[1] <- "HH.ID"

#data_1 = merge(data,RepliW_1, by = 'HH.ID')

#data_1$hh.income<-completed_not_Na$hh.income

#data_1$num.car<-data_2$num.car

#################################################################
```

Turning variables into characters and finishing survey design

```
#data_1$hh.income<-as.character(data_1$hh.income)

#data_1$num.car<-as.character(data_1$num.car)

#data_1$g.or.w<-as.character(data_1$g.or.w)

#data_1$num.in[which(data_1$num.in > 2)] <- 2

#summary(xtabs(~num.in+hh.income, data_1))

#Survey design object

#chisq.d<-svrepdesign(variables = data_1[,-c(8:170)],

#               weights = data_1$hh.weight,

#                repweights = data_1[,c(11:170)]

#               )

################################################
```

Running Chi-Square test for num.in and g.or.w, figure 3

svychisq(~g.or.w+num.in, chisq.d)

##

Running Chi-Square test for num.in and hh,income and figure 5

svychisq(~hh.income+num.in, chisq.d)

#################################

Running Chi-Square test for num.car and num.in, figure 6

svychisq(~num.car+num.in, chisq.d)

##

Running Chi-Square test for num.age.L12

svychisq(~num.age.L12+num.in, chisq.d)

##

Running Chi-Square test for num.age.S12

svychisq(~num.age.S12+num.in, chisq.d)

##

Printing bar plot for household income and incident number, figure 4

ggplot(data_1,aes(x=hh.income,y=num.in))+geom_col(aes(color=hh.income))

##

Printing the bar plot for number of cars owned and gated or walled, figure 7

#ggplot(data_1, aes(num.car))+

geom_bar(aes(fill=g.or.w), width = 0.5) +

theme(axis.text.x = element_text(angle=65, vjust=0.6)) +

labs(title="Histogram on Categorical Variable",

subtitle="Number of cars across Gated/Walled")

##

Printing plots & expected regression line for number of household members

ggplot(data_1, aes(x=num.age.L12,y=num.in))+geom_col(aes(color=num.age.L12)) **#figure 8**

ggplot(data_1, aes(x=num.age.S12,y=num.in))+geom_col(aes(color=num.age.S12)) **#figure 9**

###

Citation

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. *National Crime Victimization Survey, 2016*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-12-14. https://doi.org/10.3886/ICPSR36828.v1

Jinbo Wu

3033263170

STAT-152

Project Self Evaluation

Me, Yijun Xu and Jiaying Lu worked together for a month to finish this project together. In the first stage of this project, both of us chose a survey data for the project. After discussion, we made an agreement to analyse the NCVS, 2016 data files. This file was found by Jiaying Lu on https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/36828/summary.

We met everyday for a week to discuss and worked on the details of the project for stage 4. At first, both of us read the code book, totally understood the survey design and figured out what the PSUs and stratas were. Then I wrote the first draft for the survey design for the report. Yijun and Jiaying were responsible for reviewing and giving me feedback on how to revise the words to make it more precise in order to correctly reflect the codebook materials.

For the methodology part, we've discussed distribution of our job. Jiaying was responsible for taking care of the data file and found a method for solving NA in the variables. Yijun was working on the survey package and Chi-Square test results. I was working on the data visualization such as boxplots, histograms and bar plots. After we finished, Yijun was writing up the methodology part. Jiaying and I were there giving her advice. At the same time, Jiaying was working on the results based on our topic and analysing direction. We analyzed the results from the outputs in R and gave Jiaying our ideas and outlines so that Jiaying had a basic guide to work. As for the conclusion, like the second part, I was responsible for the writing with advices

from Yijun and Jiaying. At last, Jiaying and Yijun were working on the general layout and the code appendix.

At last, I would like to say that I am very glad to have two smart, beautiful, responsible and helpful teammates to work with. Together we made our last project one of the best things I will remember as a student in Berkeley, forever.