Michael Neman
Write-Up
ECON 8320-850, Fall 2022, Dr. White

Data collected by the National Crime Victimization Survey (NCVS) was processed for the analysis of violent crime that occurred in the workplace. Data was collected with the purpose of accumulating information on victims of crime, to estimate the number of crimes not reported to police, and to conduct comparisons over time and areas (National Crime Victimization Survey, Concatenated File, [United States], 1992-2021). Four files were downloaded and used for the data processing: study-level files, which includes a PDF detailing the scope of the study and the variables contained within the data files(codebook), concatenated household file(household file), which contains data on the household and surrounding area, concatenated person file(person file), which contains data on each household member 12 years and older that was reported, and concatenated incident file(incident file), which contains data on the incident for each household or person incident (National Crime Victimization Survey, Concatenated File, [United States], 1992-2021).

Document Reproducibility. The original data files used can be found at the ICPSR website. Python was used for the cleaning and processing of the data sets. The source code for the processing can pulled from the GitHub repository, source code. The code provided will reproduce the resulting data set that will be used for analysis, barring any formatting changes or variable assignments that were changed. The STATA file format was used for the household file, person file, and incident file. If another file format is used, the `pandas.read` function in the source code will need to be updated accordingly.

About the Data

Four files pulled from the NCVS were used for the processing. The codebook contains detailed information on variables, weights, and the scope of the study. The codebook file is part

of the study-level files from the NCVS. The Household file in STATA format; contains 226 variables relating to the household and surrounding area. Variable examples include but are not limited to, type of living quarters, location of phone, direct outside access, gated or walled community, household income. The person file, in STATA format contains 171 variables relating to each member of the household 12 years and older that was reported during the interview. Variable examples include but are not limited to, age, sex, race, gender, and employment. The incident file, in STATA format contains 1,017 variables relating to information drawn from the incident report for each household or person incident recorded during the interview (National Crime Victimization Survey, Concatenated File, [United States], 1992-2021). All 3 data files contain the columns IDHH and YEAR. IDHH is a unique row designator for the purpose of merging data frames.

The data files use four codes to represent missing data: Residue, Out of Universe/Blank, Item Valid Until YYYY Q#, and Item Invalid After YYYY Q#. Residue is represented by 8's, 9's or a combination of 8's and 9's. Two types of Residue codes can occur for multiple response variables. When the entire response was not obtained, '8' is used. If an initial response was obtained, but follow up responses were not, the initial response will have a '1' with follow up responses codes as an '8'. Out of Universe/Blank missing data is coded as a '9'. Item Invalid Until YYYY Q# is coded with a '-1'. The missing data is a result of the variable not being recorded until the date and quarter identified. Item Invalid After YYYY Q# is coded with a '-2'. The missing data is a result of the variable no longer being recorded after the year and quarter identified (National Crime Victimization Survey, Concatenated File, [United States], 1992-2021).

Preparing the Data

Data processing was performed in the Spyder IDE, using Python 3.9.

The three data sets, household file, person file and incident file were downloaded from the NCVS website in the STATA format (National Crime Victimization Survey, Concatenated File, [United States], 1992-2021). I initially attempted to download the CSV format, but the download appeared to stall on my computer. The STATA format file downloaded without issue. If another format is used, the source code to read the file will need to be updated accordingly. The three files were read using the `pandas.read_stata` function.

After reading the files the appropriate columns containing relevant variables for the scope of analysis were selected. Variable selection was time consuming given the number of variables available: 226 in the household file, 171 in the person file and 1,017 in the incident file. Several iterations of variable selection took place while focusing the scope of analysis until 54 columns were selected. The columns selected from the incident file contained data pertaining to incidents that occurred at the workplace involving violent and/-or sexual crimes. Pertinent demographic columns were selected from the person file. Household income was selected from the household file. For a full list of variables selected, see Appendix.

After selecting the relevant columns, the missing data in each data frame was converted to NaN. Due to there being a variety of missing data codes I attempted to determine the most efficient way to replace the correct codes within each column. Some of the variables had categorical designations that included ‘8’ and ‘9’. By simply replacing 8 and 9 in all columns would erase the entries where this was valid data. My solution to this is less than ideal. I unsuccessfully ran the initial code I was attempting to implement and had to make adjustments. I generated multiple lists, each one containing the column Id’s for variables that utilized the same

missing data code per the codebook. Then I looped through the columns replacing the missing data code with Nan.

The code snippet above is an example of a column list and the for loops I ran to remove the missing data codes. Ideally, I would have looped through the list of column names replacing all the relevant missing data codes in one loop, but I was unable to run the loop correctly. In the end, I ran individual loops over the column name list multiple times for each individual missing data code.

The columns were renamed to names that more accurately described the data that was contained in each column, limiting the amount of time needed to reference the Codebook during analysis. The three files were merged into one data frame using the `pandas.merge` function on the IDHH column. Rows pertaining to incidents that did not occur at work were removed.

Conclusions

Determining the scope of the analysis wanting to be performed on the dataset and selecting the appropriate variables for that scope was the most time-consuming aspect of the data cleaning process. Learning that beginning the cleaning process with research questions in mind and having the flexibility and understanding that the research questions and scope could change during the process was the most important finding for me. Beginning the process with that flexible mindset will be beneficial for future projects. While I was able to execute code that successfully created a cleaned data frame, I learned of deficiencies in my coding ability that prevented me from processing the data more efficiently.

References

National Crime Victimization Survey, Concatenated File, [United States], 1992-2021. (n.d.). Www.icpsr.umich.edu. https://www.icpsr.umich.edu/web/NACJD/studies/38430/summary

Appendix

For detailed description of variables see NCVS Codebook. (Statistics, 2022)

NCVS Variable name	Cleaned crime_data Variable Name	NCVS Variable name	Cleaned crime_data Variable Name
IDHH	IDHH	V4055	blunt_object
YEAR	YEAR	V4056	weapon_other
V2026	household_income	V4060	hit_attack
V2120	public_housing	V4061	attck_attempt
V3013	Age	V4069	sex_contact_force
V3015	marital_status	V4070	sex_contact_no_force
V3017	sex	V4078	threat_rape
V3020	ed_attain	V4079	threat_kill
V3022	race	V4081	threat_sex_assault
V3023A	race_recode	V4094	raped
V3074	job_desc	V4095	rape_attempt
V3075	emp_sector	V4096	sex_assault
V3076	work_urban_sub_rural	V4097	shot
V3078	university_employee	V4098	shot_missed
V3083	citizen_status	V4100	knife_attack
V3084	sexual_orientation	V4102	hit_object
V3085	birth_gender_ident	V4105	hit_slapped
V3086	current_gender_ident	V4106	grabbed
V4014	month_occured	V4482A	industry_code
V4015	year_occured	V4482B	occupation_code
V4049	had_weapon	V4483	job_location
V4050	weapon_type	V4484	occurred_at_work
V4051	hand_gun	V4485	work_days_nights
V4052	other_gun	V4528	crime_code_old
V4053	knife	V4529	crime_code_new