Task 6.1 – Sourcing Open Data

**Data Source**

Data Source Summary:

* Summarize the data source. Include whether it’s internal or external data, who owns the data, and how trustworthy it is.

I’ve downloaded two data sets: The Uniform Crime Report (UCR) and the Supplementary Homicide Report (SHR). The source of the data is the Murder Accountability Project ([www.murderdata.org](http://www.murderdata.org)), a nonprofit group that collects and publishes homicide information from federal, state, and local governments. These data sets have information for 50+ years of homicides across the United States.

The data is external, acquired from the Federal Bureau of Investigation (FBI). This information is publicly available in accordance with local Freedom of Information Acts.

This data is trustworthy as the organization exists to help local police departments solve murder cases by providing these databases to the public, and to educate the public about the country’s unsolved murder problem. Therefore, there is no conflict of interest. The Murder Accountability Project (MAP) is run by Thomas Hargrove, a retired investigative journalist. His email and phone number are available on the website to take questions from police departments and the press. The Board of Directors consists of former police officers, former federal agents, criminologists, forensic psychologists, etc.

From the website:

“This website gives police and the public easy-to-use access to two datasets maintained by the Federal Bureau of Investigation: the Uniform Crime Report from 1965 to the present and — more useful for investigators — the Supplementary Homicide Report from 1976 to the present. The Murder Accountability Project, using the Freedom of Information Act, has obtained data on more than 30,500 homicides that were not reported to the Justice Department. This means the information at www.murderdata.org is the most complete data on U.S. homicides available anywhere.”

* Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

The data was collected through police departments entering information about cases into a database. It is administrative data. The SHR data is typically entered at the outset of the crime and is often not updated if a case is solved later. 5-10% of homicides reported as unsolved on the SHR are solved later. The UCR has less information (columns), but it is updated when at least one offender has been arrested, charged, and tried, and should be used to calculate murder clearance rates. The data is collected manually, as the information from each case must be entered separately. The latest data on these two sets is from December 2019, so there is at least a year time lag.

(I might also incorporate population data from the U.S. Census Bureau that was used on the previous medical staffing project, but am not sure yet.)

* Write an overview of the data contents. What variables are included?

The UCR has 8 columns and 166,225 rows. The variables are:

ORI – numbered code for the police department (Originating Agency Identifier) e.g. FL01300

Name – coded name of the police department e.g. MIAMI-DADE

Year – year in which the homicides took place e.g. 2019 (Range 1965 – 2019)

MRD – number of homicides e.g. 90

CLR – number of solved (cleared) homicides e.g. 45

State – state of the police department e.g. Florida

County – county of the police department e.g Miami-Dade, FL

Agency – name of the police department e.g. Miami-Dade

The SHR has 31 columns and 804,751 rows. The variables are:

ID – case identification number e.g. 201901001NY02900

CNTYFIPS – police station state and county e.g. Nassau, NY

Ori – police station code (Originating Agency Identifier) e.g. NY02900

State – state of homicide e.g. New York

Agency – police agency e.g. Nassau County

Agentype – type of police e.g. Sheriff

Source – Information source e.g. FBI

Solved – Has the case been solved? e.g. No

Year – year of case e.g. 2019 (Range 1976 – 2019)

StateName – name of state (this column is completely blank)

Month – month of case e.g. January

Incident – A unique number that distinguishes one incident from another within the ORI e.g. 1 (Ranges from 0 – 999)

ActionType – What kind of data update? Normal or adjusted? e.g. Normal update

Homicide – Type of homicide e.g. Murder and non-negligent manslaughter

Situation – description of the number of victims/offenders and if the offender is known/unknown

e.g. Single victim/unknown offender(s)

VicAge – Victim age e.g. 70

VicSex – victim gender e.g. Male

VicRace – victim race e.g. White

VicEthnic – victim ethnicity (Hispanic or not or unknown) e.g. Not of Hispanic origin

OffAge – offender age e.g. 999 (used for unknown offenders)

OffSex – offender gender e.g. Unknown

OffRace – offender race e.g. Unknown

OffEthnic – offender ethnicity (Hispanic or not or unknown) e.g. Unknown or not reported

Weapon - Homicide weapon e.g. Knife or cutting instrument

Relationship – relation of victim to offender e.g. Relationship not determined

Circumstance – circumstances surrounding homicide e.g. Circumstances undetermined

Subcircum – details in cases of justifiable homicide (mostly blank)

VicCount – Number of additional victims e.g. 0 (mostly 0, Ranges from 0 – 21)

OffCount – Number of additional offenders e.g. 0 (mostly 0, Ranges from 0 – 21)

FileDate – date report was filed e.g. 43020

MSA – Metropolitan Statistical Area e.g. New York-New Jersey-Long Island, NY-NJ

Why have I chosen this Data Set?

I have chosen this data set because it meets nearly all the requirements outlined in the project brief— It contains geographical data, quantitative data, qualitative data, and time data. The SHR has 10+ columns, including 5 qualitative and 4 quantitative. There are 5000+ rows in each set. The columns are non-anonymized. The data is less than 3 years old. I believe the data to be trustworthy as it is a non-profit organization whose goal is education and assisting law enforcement, with contact information available from the person who runs the website. The data is open-source under the Freedom of Information Act. On a personal level, I’ve chosen this data set because I am fascinated by true crime and am interested in the insights contained within the data.

**Data Cleaning Steps**

Names of columns changed in both sets to be more clear, intuitive, and consistent.

Cleaning the UCR

There are no full duplicates in the UCR.

Looking through the data in Excel, I noticed that the UCR has some discrepancies in the murders and clearances: Occasionally, the number of clearances is listed as greater than the number of murders and this doesn’t make any sense, unless the clearances count for each person tried. I looked at a particular city—Wichita, TX, 1988 which had 15 murders and 16 clearances according to the UCR. Correspondingly, the SHR listed that there were 15 victims and 16 offenders in total. Therefore, the clearances must count each offender separately.

The UCR has “0” listed for some murders and a number greater than 0 for the corresponding clearance which definitely makes no sense. The number of murders has been changed to match the number of clearances in these cases. 3547 rows were affected.

Data types have been changed to the appropriate type for each column.

Rhodes Island changed to Rhode Island.

Cleaning the SHR

Columns deleted: state\_name in the SHR, because it was completely blank.

The subcircumstance column has lots of missing values, but this is accurate because it only lists information if the original circumstance was considered a “justifiable homicide.”

999 missing codes investigated:

* 1.5 % of victim ages are listed as ‘999’ (missing)
* 33% of offender ages are listed as ‘999’ (missing)

This makes sense because the MAP website states that about one-third of all homicides are not cleared, meaning they do not know who the offenders are.

Offender ages of 0, 1, and 2 appear to be incorrect. Starting at the age of 3, “Children playing with gun” is typically listed as the circumstance. It is highly unlikely that anyone younger than that is an offender.

After discussion with my mentor, I have imputed the missing ages, and the ages listed as 0, 1, or 2 with the median age.

* Less than 1% of the file dates are missing. This is not enough to have a measurable effect on the analysis, and this column will most likely not be important for the analysis. I am leaving it alone for now, unless this column is required at a later time in the analysis.

There are 2202 duplicates in the SHR data set. (Less than 1%)

It appears that most or all of these cases contain additional victims or additional offenders, which can cause legitimate duplicates, if the people involved happen to have the same race/gender/age, or if the information is unknown for more than one of the people involved. Upon inspection of a few cases, this is what is happening.

12 of these contain no additional offenders/victims.

15 of these are listed as single victim/single offender.

Many of these overlap, and I have deleted 16 rows in total.

Data types have been changed to the appropriate type for each column.

Rhodes Island changed to Rhode Island.

**Statistical Summary/Exploration**

Descriptive statistics calculated in the Python notebook. Some new columns were derived, and a few data frame subsets of interest were created.

**Limitations and Ethics**

Ethics:

Some individual names of victims/offenders can be identified by searching for the place, year, and number of victims online, but it is not a privacy issue since these homicides were in the news.

Limitations:

Not every single case gets reported, as many police departments don’t send unsolved homicide data to the FBI.

A lot of offender data is missing because from the SHR because they simply have not found the offenders, or the offenders were identified long after the SHR was updated. The UCR makes up for this by having up-to-date clearance rates.

**Questions to Explore**

1. How has the number of homicides fluctuated over the years?

2. Who is most likely to end up a victim in “Children playing with gun” circumstances?

3. Are men and women murdered at the same rate?

4. How common is homicide towards family members? Friends?

5. Where are cold cases most common?

6. Are homicides more likely to be interracial, or within the same race group?

7. Are homicides more common in certain states?

8. Are wives more likely to murder their husbands, or are husbands more likely to murder their wives?

9. Where are cases with multiple victims likely to occur?

10. Do ‘Red’ and ‘Blue’ states have similar homicide rates where guns are used as a weapon?

11. Is there a correlation between the number of victims and the age of the offender?

12. Which gender is most likely to be a victim in a “Lovers triangle” circumstance?