Task 6.1: Sourcing Open Data

Data Source

The Murder Accountability Project

Context:

Established in 2015, the Murder Accountability Project or MAP is a nonprofit group based in Alexandria, VA. Majority of its members have extensive experience when it comes to homicide investigations, as the project includes veteran homicide investigators, investigative journalists and homicide scholars. The project is intended to spread information about homicides, especially unsolved killings and serial murders committed in the United States. Given the context of the work and where the data comes from the MAP is a reliable and creditable source.

Data Collection:

The MAP gives police and the general public easy-to-use access to two datasets maintained by the FBI. The Uniform Crime Report from 1965 to the present and the Supplemental Homicide Report from 1976 to the present. These are voluntary reporting systems, meaning local police isn't required to provide any information to the FBI. However, the MAP has determined that UCR data is more than 95 percent complete and the SHR is more than 93 percent complete.

The MAP uses the Freedom of Information Act to obtain its data from justice departments around the U.S. regarding crimes related to homicide. Currently the MAP has obtained data on more than 28,000 homicides that were not reported to the justice department. This means the MAP has the most complete data sets on U.S. homicides available anywhere.

Content:

This project will be working with the Supplemental homicide Report and contains information regarding the state, the city, the agency, the year, the month, the action type, the type of homicide, the situation, everything relating to the victim and offender (age, sex, race, & ethnicity), the weapon used, and the relationship between the victim and offender.

Resources:

The Murder Accountability Project's website – Link
The Wiki related to WAP – Link
The Kaggle dataset – Link

Why I Chose This Data:

I spent a lot of time contemplating which dataset to work with and what topic to choose for this project. Ultimately, I knew I wanted to work with some kind of crime statistics, as I've always been a true crime fan. I searched a lot of different topics, but not a lot of them met the criteria for this project. I stumbled upon MAP and decided it was a perfect case study as a lot of the data was raw and need a lot of help.

Data Profile

The original data contains 827,219 rows and 31 columns

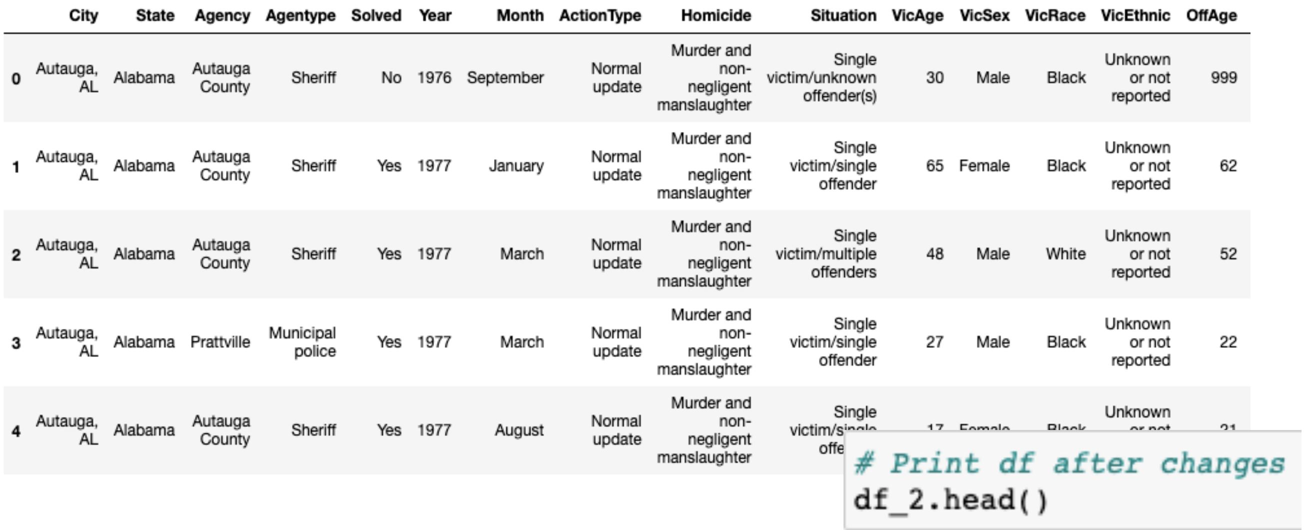
Columns:

ID, CNTYFIPS, Ori, State, Agency, Agentype, Source, Solved, Year, StateName, Month, Incident, Action Type, Homicide, Situation, VicAge, VicSex, VicRace, VicEthnic, OffAge, OffSex, OffRace, OffEthnic, Weapon, Relationship, Circumstance, Subcircum, VicCount, OffCount, FileDate, MSA

ID – Id # of the case (DROP)
CNTYFIPS – County FIPS codes
Ori – Zip Code for Police Departments (DROP)
State – The State it happened in
Agency – Investigative Agency
Agencytype – Type of Agency
Source – Who supplied the information (DROP)
Solved – Whether the case was solved
Year – The year it happened
StateName – Abbreviated state names (Drop)
Month – The month it happened
Incident – Number of incidents that occurred (DROP)
ActionType – Whether the case has been updated (DROP)
Homicide – What happened
Situation – The Situation
VicAge – Victims age
VicSex – Victims sex
VicRace – Victims race
VicEthnic – Victims ethnicity
OffAge – Offenders age
OffSex – Offenders sex
OffRace – Offenders race
OffEthnic – Offenders ethnicity
Weapon – Weapon used
Relationship – Offenders and Victims relationship
Circumstance – What lead to the attack
Subcircum – What happened to after the attack
VicCount – Number of additional Victims
OffCount – Number of additional Offenders
FileDate – Date filed into database (DROP)
MSA – place/location (DROP)

Data Wrangling:

I'll be dropping the ID, Ori, ActionType, FileDate, Incident columns as they provide no value to my analytical study. I'll also drop the Source, StateName as they don't add or subtract from the study. Lastly, I'm dropping MSA as I'm not 100% sure what it represents and is more of duplicate data. I Changed the name of column header CNTYFIPS to City as it aligns with the city in each state.



	City	State	Agency	Agentype	Solved	Year	Month	ActionType	Homicide	Situation	VicAge	VicSex	VicRace	VicEthnic	OffAge
0	Autauga, AL	Alabama	Autauga County	Sheriff	No	1976	September	Normal update	Murder and non-negligent manslaughter	Single victim/unknown offender(s)	30	Male	Black	Unknown or not reported	999
1	Autauga, AL	Alabama	Autauga County	Sheriff	Yes	1977	January	Normal update	Murder and non-negligent manslaughter	Single victim/single offender	65	Female	Black	Unknown or not reported	62
2	Autauga, AL	Alabama	Autauga County	Sheriff	Yes	1977	March	Normal update	Murder and non-negligent manslaughter	Single victim/multiple offenders	48	Male	White	Unknown or not reported	52
3	Autauga, AL	Alabama	Prattville	Municipal police	Yes	1977	March	Normal update	Murder and non-negligent manslaughter	Single victim/single offender	27	Male	Black	Unknown or not reported	22
4	Autauga, AL	Alabama	Autauga County	Sheriff	Yes	1977	August	Normal update	Murder and non-negligent manslaughter	Single victim/single offe	17	Female	Black	Unknown or not	21

```
# Print df after changes
df_2.head()
```

Consistency Checks:

- There were mixed data types. (Either fixed or removed)
- There were missing values in the Subcircum. (Removed Subcircum from the df)
- There were 2302 rows of duplicate data. (Removed)
- Victim Age and Offender Age both had max outlier of 999. (Used np.nan on them)
- Inputted mean values for missing nan values in age columns.
- New max values for both Victim Age and Offender Age is 99.
- Deleted any offenders age min <= 2, new min is 3, as I'm setting the limit of someone who can commit homicide to the age of 3.
- Everything else was clean.

Basic descriptive statistics:

817709 rows
23 columns

Year	VicAge	OffAge	VicCount	OffCount
Min – 1976	Min – 0	Min – 3	Min – 0	Min – 0
Max – 2020	Max – 99	Max – 99	Max – 21	Max – 40
Average – 1997	Average – 33	Average – 31	Average – 0.122701	Average – 0.185788

Limitations and ethical considerations

The limitations of this project have a lot to do with the missing values of the offenders and victims ages. I struggled to input the mean value as I felt that was corrupting the data. However, if I had not done so, it would have deleted more than 100,000 rows of data. I felt it was best to input the mean value. Another limitation or ethical concern is the offender's age. I chose to go with three years old as it is possible for someone so young to mishandle a dangerous object, such as a knife or a gun. I'll also be curious to see if there is any racial ethic issues, as this data comes from police stations from around the United States.

Questions:

- Which States have the highest reported homicides documented?
- What are the percentages for Age, Race, and Sex when it comes to homicide in the US?
- What is the relationship status between Victim and Offender? Is there a correlation between family and higher homicide?
- Does year and month play a role in homicide cases? Is there an increase and decrease in homicide cases in a specific time of year?
- What is the difference between solved and unsolved cases? Has there been a change over the years?
- Is there a specific region within the US that has more unsolved homicide cases?