Analysing Murder Data

What are we going to do in this video?

So, in this video, we are going to be analyzing the US murder data. This data contains the information about the murder incidents that happened in the United States from 1976 to 2017.

The dataset comes from an online project from the name http://www.murderdata.org/ and also this website gives one very simple algorithm for clustering the informations in the data provided based on the location, weapon used in the crime and the sex of the victim.

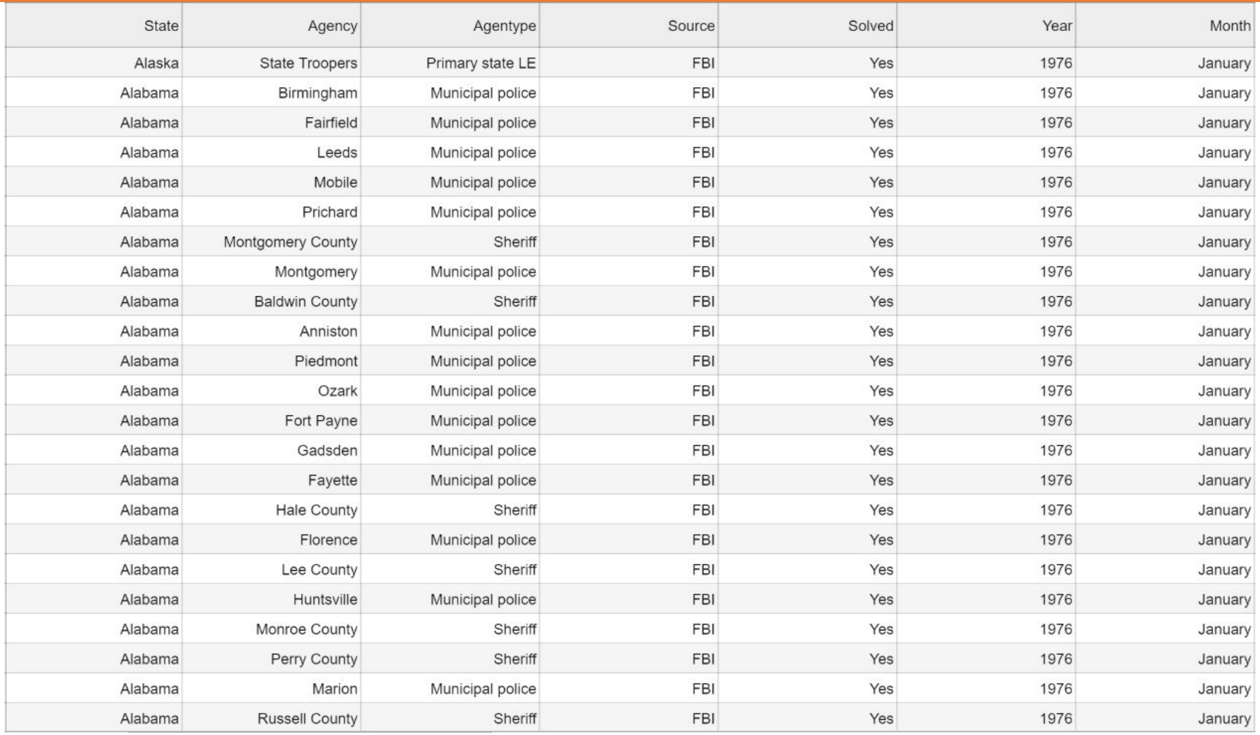
So, we will use this data as well as this algo. This algo has been implemented in SPSS but what we are going to do is implement it in the Python.

So, what is the Objective of this Mini Project?

Broadly in this project, we are going to do the following three things:

1. Some exploratory data analysis to make some hypothesis, understand the data and to answer those hypothesis.
2. Implement the murder data algorithms (SPSS one) in python.
3. Then, display the results on US map.

SHR76_17.CSV data



State	Agency	Agentype	Source	Solved	Year	Month
Alaska	State Troopers	Primary state LE	FBI	Yes	1976	January
Alabama	Birmingham	Municipal police	FBI	Yes	1976	January
Alabama	Fairfield	Municipal police	FBI	Yes	1976	January
Alabama	Leeds	Municipal police	FBI	Yes	1976	January
Alabama	Mobile	Municipal police	FBI	Yes	1976	January
Alabama	Prichard	Municipal police	FBI	Yes	1976	January
Alabama	Montgomery County	Sheriff	FBI	Yes	1976	January
Alabama	Montgomery	Municipal police	FBI	Yes	1976	January
Alabama	Baldwin County	Sheriff	FBI	Yes	1976	January
Alabama	Anniston	Municipal police	FBI	Yes	1976	January
Alabama	Piedmont	Municipal police	FBI	Yes	1976	January
Alabama	Ozark	Municipal police	FBI	Yes	1976	January
Alabama	Fort Payne	Municipal police	FBI	Yes	1976	January
Alabama	Gadsden	Municipal police	FBI	Yes	1976	January
Alabama	Fayette	Municipal police	FBI	Yes	1976	January
Alabama	Hale County	Sheriff	FBI	Yes	1976	January
Alabama	Florence	Municipal police	FBI	Yes	1976	January
Alabama	Lee County	Sheriff	FBI	Yes	1976	January
Alabama	Huntsville	Municipal police	FBI	Yes	1976	January
Alabama	Monroe County	Sheriff	FBI	Yes	1976	January
Alabama	Perry County	Sheriff	FBI	Yes	1976	January
Alabama	Marion	Municipal police	FBI	Yes	1976	January
Alabama	Russell County	Sheriff	FBI	Yes	1976	January

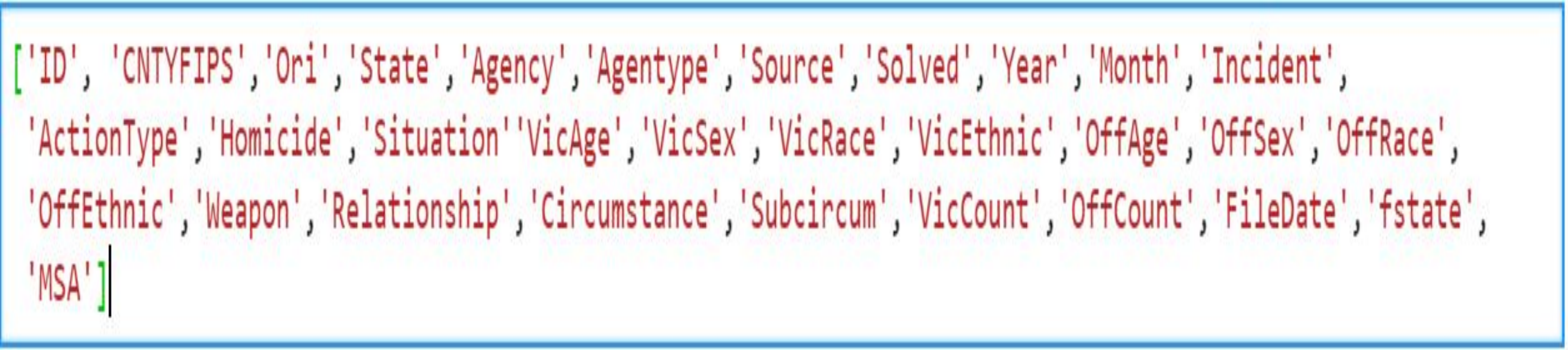
In this video,...

1. EDA (Exploratory Data Analysis).
2. Implementation of the Algorithm.

Exploratory Data Analysis

Exploratory data analysis is a way using which the data is analysed and interpreted by asking a series of questions (or making hypothesis) and verifying those questions by means of various techniques that are part of this broad field called EDA. It is also often referred to as Informal Analysis as the steps are usually hidden from the user of the project. User is only shown the final results.

Variables (Columns of the data)



```
['ID', 'CNTYFIPS','Ori','State','Agency','Agentype','Source','Solved','Year','Month','Incident',
 'ActionType','Homicide','Situation''VicAge','VicSex','VicRace','VicEthnic','OffAge','OffSex','OffRace',
 'OffEthnic','Weapon','Relationship','Circumstance','Subcircum','VicCount','OffCount','FileDate','fstate',
 'MSA']
```

Let us talk about some important features (Columns)

Features Discussion

CNTYFIPS = County in which the incident took place

MPS = Metropolitan statistical area in US

ID, STATE are self explanatory

Agency, AgentType are also self explanatory.

Year, Month = Year, Month in which incident took place

VicAge, VicSex, VicRace, VicEthinicity = Information about the Victim

OffAge, OffSex, OffRace, OffEthinicity = Information about the Offender

Solved = Whether the incident has been solved or not.

Weapon used for murder.

Features Discussion (Continued)

Relationship = Relationship of the Victim and the Offender

Circumstance = In what circumstance the incident took place.

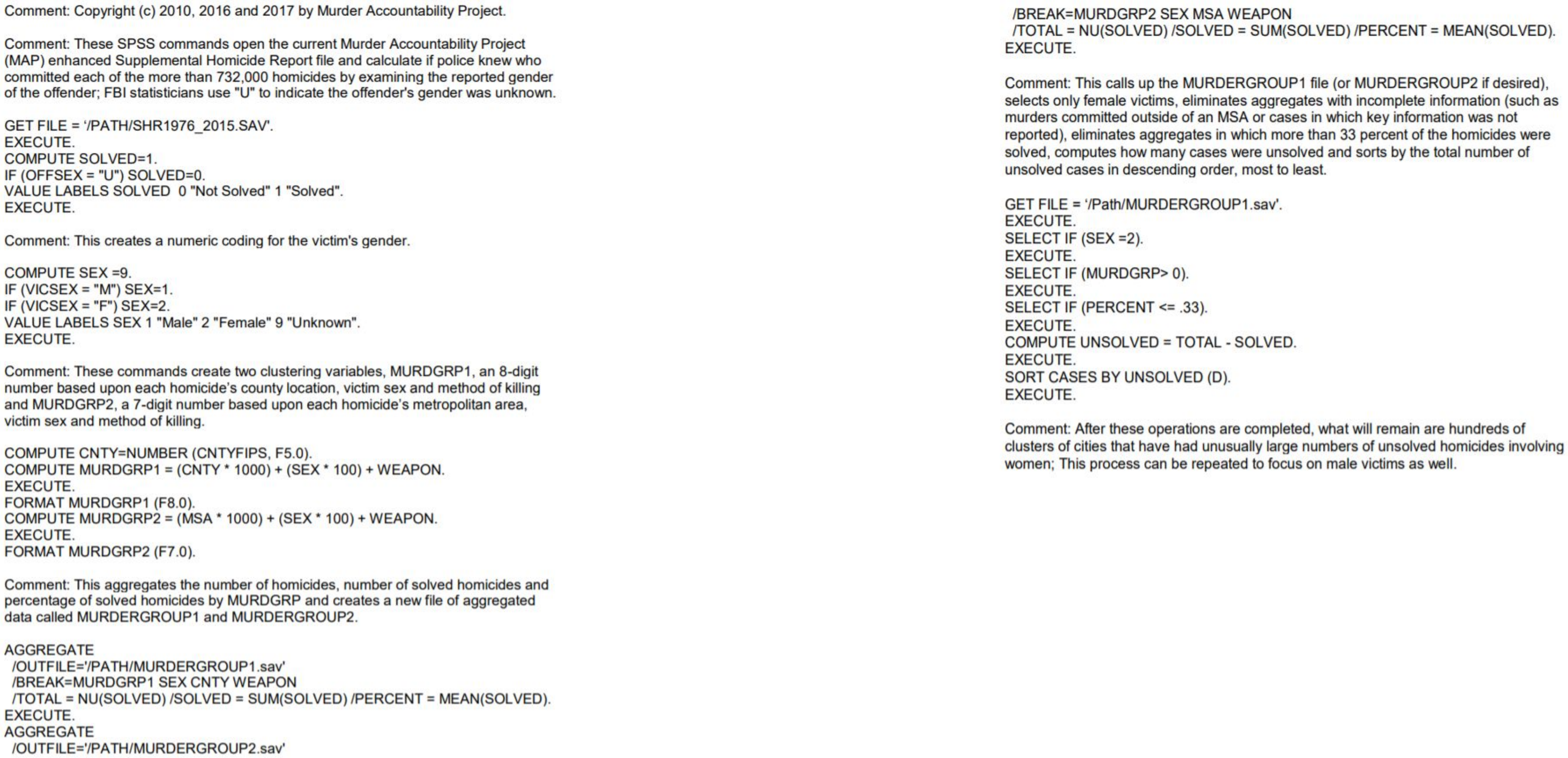
ActionType (what type of action that was taken) , Homicide and Situation (whether multiple victim or multiple offenders)

There are some other variables too. But we have only discussed around 23 which are mostly important ones for analyzing this data. Other 8 variables, are not that much important or you are curious, you can go ahead and analyze the remaining ones too.

Some preliminary Analysis(EDA) that can be done..

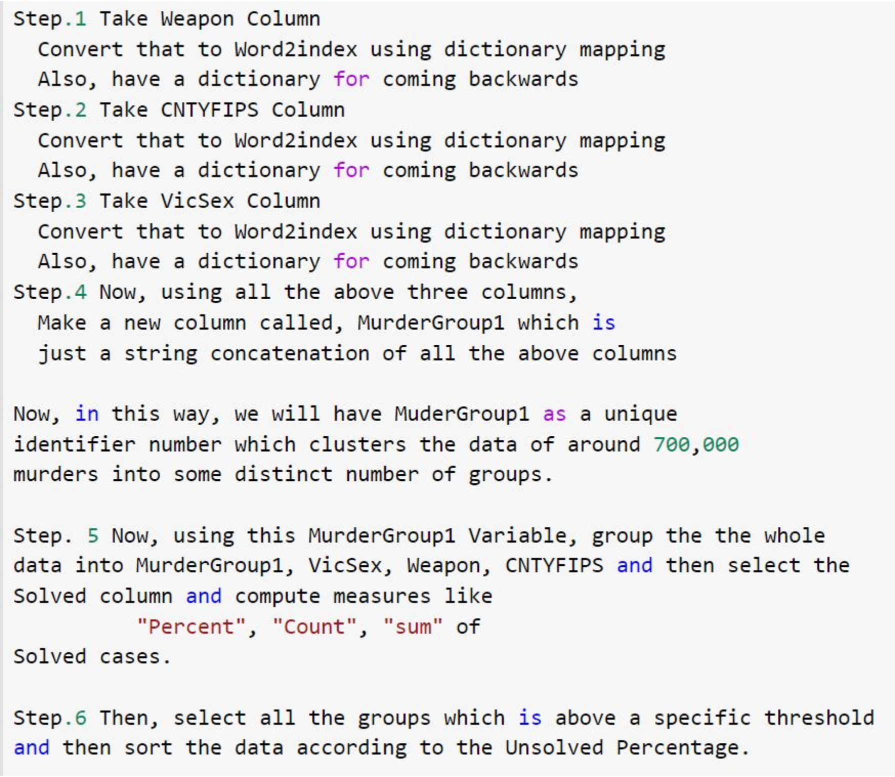
1. Are female victims are attacked by males or females more?
2. What is the type of weapons used by both the male and female offenders?
3. What is the distribution of victim's age?
4. What is the distribution of offender's age?
5. Distribution of victim's Race?
6. Distribution of offender's Race?
7. Joint Distribution of Victim and Offender's Race?
8. Age Distribution of Male and Female Offenders
9. Age Distribution of Male and Female Victims
10. In which year, most of the incidents took place? Is there any pattern?
11. What is the most common relationship between victim and offender?
12. In which month of the year, more incidents are likely to take place?
13. And so on…….

Algorithm



```
Comment: Copyright (c) 2010, 2016 and 2017 by Murder Accountability Project.

Comment: These SPSS commands open the current Murder Accountability Project
(MAP) enhanced Supplemental Homicide Report file and calculate if police knew who
committed each of the more than 732,000 homicides by examining the reported gender
of the offender; FBI statisticians use "U" to indicate the offender's gender was unknown.

GET FILE = ‘/PATH/SHR1976_2015.SAV’.
EXECUTE.
COMPUTE SOLVED=1.
IF (OFFSEX = "U") SOLVED=0.
VALUE LABELS SOLVED  0 "Not Solved" 1 "Solved".
EXECUTE.

Comment: This creates a numeric coding for the victim's gender.

COMPUTE SEX =9.
IF (VICSEX = "M") SEX=1.
IF (VICSEX = "F") SEX=2.
VALUE LABELS SEX 1 "Male" 2 "Female" 9 "Unknown".
EXECUTE.

Comment: These commands create two clustering variables, MURDGRP1, an 8-digit
number based upon each homicide’s county location, victim sex and method of killing
and MURDGRP2, a 7-digit number based upon each homicide’s metropolitan area,
victim sex and method of killing.

COMPUTE CNTY=NUMBER (CNTYFIPS, F5.0).
COMPUTE MURDGRP1 = (CNTY * 1000) + (SEX * 100) + WEAPON.
EXECUTE.
FORMAT MURDGRP1 (F8.0).
COMPUTE MURDGRP2 = (MSA * 1000) + (SEX * 100) + WEAPON.
EXECUTE.
FORMAT MURDGRP2 (F7.0).

Comment: This aggregates the number of homicides, number of solved homicides and
percentage of solved homicides by MURDGRP and creates a new file of aggregated
data called MURDERGROUP1 and MURDERGROUP2.

AGGREGATE
  /OUTFILE='/PATH/MURDERGROUP1.sav'
  /BREAK=MURDGRP1 SEX CNTY WEAPON
  /TOTAL = NU(SOLVED) /SOLVED = SUM(SOLVED) /PERCENT = MEAN(SOLVED).
EXECUTE.
AGGREGATE
  /OUTFILE='/PATH/MURDERGROUP2.sav'
  /BREAK=MURDGRP2 SEX MSA WEAPON
  /TOTAL = NU(SOLVED) /SOLVED = SUM(SOLVED) /PERCENT = MEAN(SOLVED).
EXECUTE.

Comment: This calls up the MURDERGROUP1 file (or MURDERGROUP2 if desired),
selects only female victims, eliminates aggregates with incomplete information (such as
murders committed outside of an MSA or cases in which key information was not
reported), eliminates aggregates in which more than 33 percent of the homicides were
solved, computes how many cases were unsolved and sorts by the total number of
unsolved cases in descending order, most to least.

GET FILE = ‘/Path/MURDERGROUP1.sav’.
EXECUTE.
SELECT IF (SEX =2).
EXECUTE.
SELECT IF (MURDGRP> 0).
EXECUTE.
SELECT IF (PERCENT <= .33).
EXECUTE.
COMPUTE UNSOLVED = TOTAL - SOLVED.
EXECUTE.
SORT CASES BY UNSOLVED (D).
EXECUTE.

Comment: After these operations are completed, what will remain are hundreds of
clusters of cities that have had unusually large numbers of unsolved homicides involving
women; This process can be repeated to focus on male victims as well.
```

Algorithm in Python



```
Step.1 Take Weapon Column
  Convert that to Word2index using dictionary mapping
  Also, have a dictionary for coming backwards
Step.2 Take CNTYFIPS Column
  Convert that to Word2index using dictionary mapping
  Also, have a dictionary for coming backwards
Step.3 Take VicSex Column
  Convert that to Word2index using dictionary mapping
  Also, have a dictionary for coming backwards
Step.4 Now, using all the above three columns,
  Make a new column called, MurderGroup1 which is
  just a string concatenation of all the above columns

Now, in this way, we will have MuderGroup1 as a unique
identifier number which clusters the data of around 700,000
murders into some distinct number of groups.

Step. 5 Now, using this MurderGroup1 Variable, group the the whole
data into MurderGroup1, VicSex, Weapon, CNTYFIPS and then select the
Solved column and compute measures like
          "Percent", "Count", "sum" of
Solved cases.

Step.6 Then, select all the groups which is above a specific threshold
and then sort the data according to the Unsolved Percentage.
```

Let's Implement the algo now.

We will graph the results and interpret the results in the nxt Video

Thank you