Detecting the Serial Killer Activity

CS989 - Joshua Eick - 201779687

Analysis of the Serial Killer Dataset

Word count: 2,164

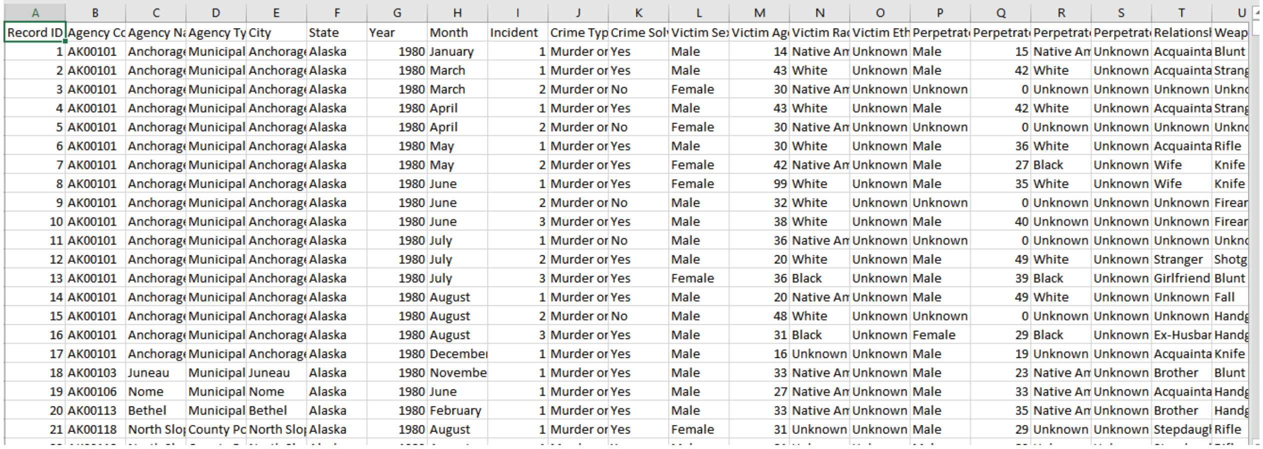
Python Version: python 3.3

Libraries:

- Numpy
- Scikit
- Math
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- import numpy as np
- plt.style.use('fivethirtyeight')
- from sklearn import metrics
- import sklearn
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.model_selection import train_test_split
- from sklearn.cross_validation import KFold
- from sklearn.linear_model import LogisticRegression
- from sklearn import metrics
- from sklearn.model_selection import cross_val_score

Introduction:

In this report we are going to understand and analyze the dataset, specially form a hypothesis and create a code separately to deduce the reason of the crimes or murder provided in the dataset. The dataset contains the distributed data about the murders/crimes in America from 1980 to 2014. The dataset is well organized as it is provided by the FBI and it contains up-to 638455 counts of Crimes. Figure 1.1 is a sample picture of the dataset.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Record ID	Agency Cc	Agency Na	Agency Ty	City	State	Year	Month	Incident	Crime Typ	Crime Solv	Victim Sex	Victim Age	Victim Rac	Victim Eth	Perpetrato	Perpetrato	Perpetrato	Perpetrato	Relationsl	Weap
1	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	January	1	Murder or	Yes	Male	14	Native Am	Unknown	Male	15	Native Am	Unknown	Acquainta	Blunt
2	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	March	1	Murder or	Yes	Male	43	White	Unknown	Male	42	White	Unknown	Acquainta	Strang
3	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	March	2	Murder or	No	Female	30	Native Am	Unknown	Unknown	0	Unknown	Unknown	Unknown	Unknc
4	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	April	1	Murder or	Yes	Male	43	White	Unknown	Male	42	White	Unknown	Acquainta	Strang
5	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	April	2	Murder or	No	Female	30	Native Am	Unknown	Unknown	0	Unknown	Unknown	Unknown	Unknc
6	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	May	1	Murder or	Yes	Male	30	White	Unknown	Male	36	White	Unknown	Acquainta	Rifle
7	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	May	2	Murder or	Yes	Female	42	Native Am	Unknown	Male	27	Black	Unknown	Wife	Knife
8	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	June	1	Murder or	Yes	Female	99	White	Unknown	Male	35	White	Unknown	Wife	Knife
9	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	June	2	Murder or	No	Male	32	White	Unknown	Unknown	0	Unknown	Unknown	Unknown	Firear
10	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	June	3	Murder or	Yes	Male	38	White	Unknown	Male	40	Unknown	Unknown	Unknown	Firear
11	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	July	1	Murder or	No	Male	36	Native Am	Unknown	Unknown	0	Unknown	Unknown	Unknown	Unknc
12	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	July	2	Murder or	Yes	Male	20	White	Unknown	Male	49	White	Unknown	Stranger	Shotg
13	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	July	3	Murder or	Yes	Female	36	Black	Unknown	Male	39	Black	Unknown	Girlfriend	Blunt
14	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	August	1	Murder or	Yes	Male	20	Native Am	Unknown	Male	49	White	Unknown	Unknown	Fall
15	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	August	2	Murder or	No	Male	48	White	Unknown	Unknown	0	Unknown	Unknown	Unknown	Handg
16	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	August	3	Murder or	Yes	Male	31	Black	Unknown	Female	29	Black	Unknown	Ex-Husbar	Handg
17	AK00101	Anchorage	Municipal	Anchorage	Alaska	1980	Decembe	1	Murder or	Yes	Male	16	Unknown	Unknown	Male	19	Unknown	Unknown	Acquainta	Knife
18	AK00103	Juneau	Municipal	Juneau	Alaska	1980	Novembe	1	Murder or	Yes	Male	33	Native Am	Unknown	Male	23	Native Am	Unknown	Brother	Blunt
19	AK00106	Nome	Municipal	Nome	Alaska	1980	June	1	Murder or	Yes	Male	27	Native Am	Unknown	Male	33	Native Am	Unknown	Acquainta	Handg
20	AK00113	Bethel	Municipal	Bethel	Alaska	1980	February	1	Murder or	Yes	Male	33	Native Am	Unknown	Male	35	Native Am	Unknown	Brother	Handg
21	AK00118	North Slo	County Pc	North Slo	Alaska	1980	August	1	Murder or	Yes	Female	31	Unknown	Unknown	Male	29	Unknown	Unknown	Stepdaug	Rifle

Figure 1.1

The dataset contains Record ID, Agency Code, Agency Name, Agency Type, City, year, Month, Incident, Crime Type, Crime Solve, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator race, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, perpetrator Count.

Agency Code, Agency Name, Agency Type describes the information related to the Agency.

State represents the area in which crime happened.

Year and month represents the time in which crime happened.

Incident is the number of Crimes happened at that time.

Crime type describes the information about the crime whether it is a murder or not.

Crime Solve is analyzed by FBI whether the case is solved or not.

Victim Sex, Victim Race, Victim Age and Victim Ethnicity is also enlisted in the dataset because it can be a useful fact in finding the relationship with the killer.

Perpetrator Sex, Perpetrator Age, Perpetrator race, Perpetrator Ethnicity are also enlisted in dataset to check whether the crimes are irregular or going on a pattern so that future crimes could be stopped after finding a pattern in perpetrator.

Relationship is mentioned between the Victim and the perpetrator whether it is Acquaintance or girlfriend or someone else. The relationship can come to handy for figuring out the reason behind the crimes.

Victim Count is mentioned in the dataset for the help of hypothesis proof so that multiple killings can be identified and can be displayed using charts.

Perpetrator Count is also mentioned in the dataset for the help of hypothesis proof. Thus, the that group killing can be identified and displayed using the charts.

Dataset:

The dataset is available at and downloaded from Kaggle.com. The link is provided below

https://www.kaggle.com/murderaccountability/homicide-reports/data

Content:

According to Kaggle (2017) discussed the content of the data:

> "The Murder Accountability Project is the most complete database of homicides in the United

> States currently available. This dataset includes murders from the FBI's Supplementary
>
> Homicide Report from 1976 to the present and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used."

Acknowledgements:

Moreover, Kaggle (2017) outlined: "The data was compiled and made available by the Murder Accountability Project, founded by Thomas Hargrove."

From this data we would be allowed to deduce a conclusion and prove our hypothesis

Key challenges/problems:

The most difficult phase is that the data is bit ambiguous because most of the values in the data are not mentioned like relationship or victim count or weapon so instead of these liabilities we have to use this data and sort out the problem of the missing data of these columns.

In addition, the key challenge in this raw data is to find out the relation/pattern/cause of the murders. The victim age and perpetrator age can be cause of the murder but as race can also be a factor here, as racism is also cause of the clash between nations so race can also not be neglected but according to the given statistics most of the perpetrator race is not mentioned and it leads to ambiguity. Therefore, we have to keep in mind the link between races. According to

Institute of Race Relation (2014):

> "A racist incident, according to the police, is any incident, including any crime, which is perceived by the victim or any other person to be motivated by a hostility or prejudice based on a person's 'race' or perceived 'race'. In 2013/14, there were 47,571 'racist incidents' recorded by the police in England and Wales. On average, that is about 130 incidents per day."

Moving on to next part, the relation between perpetrator and the victim should also be checked as it can come to handy but through relationship the deductions can't be made as there always will be ambiguity because most of the data in the database does not have relation mentioned.

The key challenge faced during the analysis and hypothesis making is that the data in the data set is of huge diversity like the victim count and perpetrator count has large diversity in it. The diverse data is not easy to put together and get it solved while the columns with less diversity are easy to put together and deduce a conclusion.

Hypothesis:

The number of suicides (dependent variable) in each states have a relation to the variables of Penetrator Age, Victim age, Relationship, Penetrator Race, Victim race (which are the independent variables).

As we have to verify Penetrator Age, Victim age, Relationship, Penetrator Race, Victim race. Thus, we can come to a conclusion that may be the murders are related to the discrimination due to race because it involves a fact of hates.

For a unsupervised analysis method or coded proof of our hypothesis we will use k-means clustering that makes clusters of the dataset according to command. As a result, we can point out the desired area as we wish in the form of the cluster.

Methods for Analysis:

First of all, raw data is verified for making hypothesis. The hypothesis is base by checking the data. The histograms or plots can be used to get some help in the making of the hypothesis.

The histograms can be used for highlighting the age factor.it will help in pointing out the age in which most murders occurred. The year variable can also be verified by making histogram of it. Thus, that year can also be pointed out.

The pie charts will also be used for analyzing the data. Pie charts are used for pointing out the peak of the data.

Finding mean of the different values can also be helpful. Therefore, wee can come close to the value which occurs most and that is our main concern here.

K-means clustering is used so that we can verify the dataset by making clusters of most of the victims which is very helpful and can lead to final deduction that which cluster (number of murders) were caused by which reason. The bigger cluster can be deduced as the result of the analysis.

First, have to import all the necessary libraries for the Analysis of code. Second, loading the data from csv file. Afterward, checking the data in Spyder. Then, finding the Average number murder per year and finding the states with most murder per year for the hypothesis. Finally and most important a partial analysis of the data all-together.

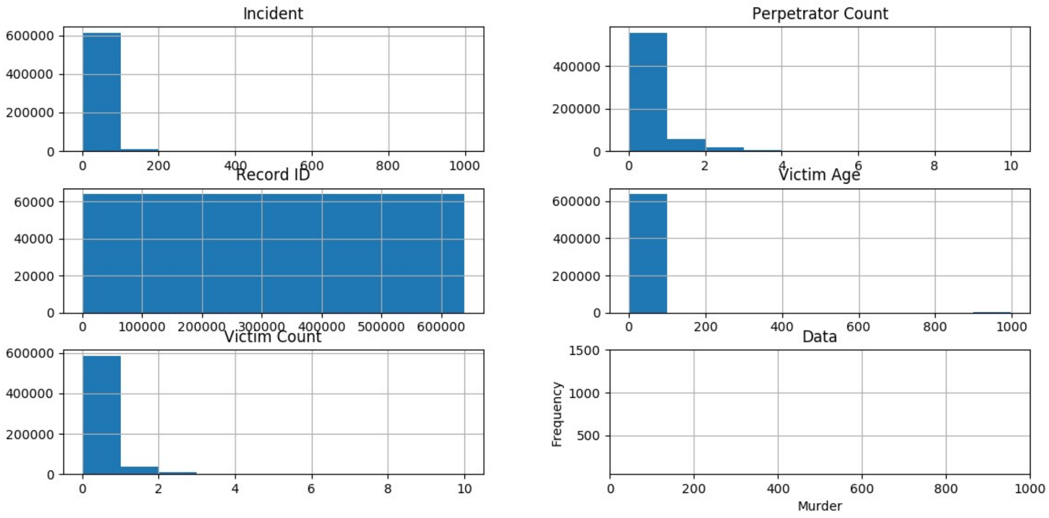


Figure 1.2

The graphs (figure 1.2) are just random presentation of the data from the dataset to just check how the values are plotted and what is the diversity of our dataset.

The Victim count is plotted (figure 1.2) and represent that there is whole lot of diversity in the dataset for Victim count. As a result, it will not be beneficial for our approach to start with the victim count. We have to select for the variable which is independent and has far more less diversity it will be good for us because in k-mean clustering we can differentiate easily and use less diverse data set easily.

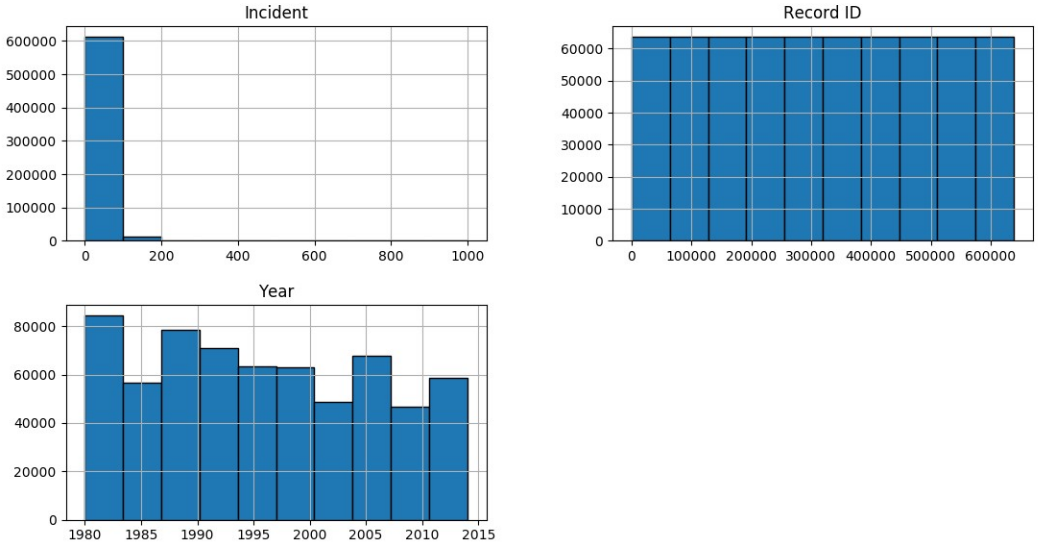


Figure 1.3

The figure 1.3, the year represents the values, which were entered at what year. As 2010 has the smallest peak then it shows that least data was entered in year 2010 and the values represents crimes. Thus, in 2010 least crimes were recorded.so far for the others. The year 1980 has the highest peak so it means most data was entered in the year 1980. Therefore, most of the crimes happened in 1980.

In figure 1.3, the incident graph just plots the raw data of the Incident column in the dataset.it shows the values going to peak like crime increasing.

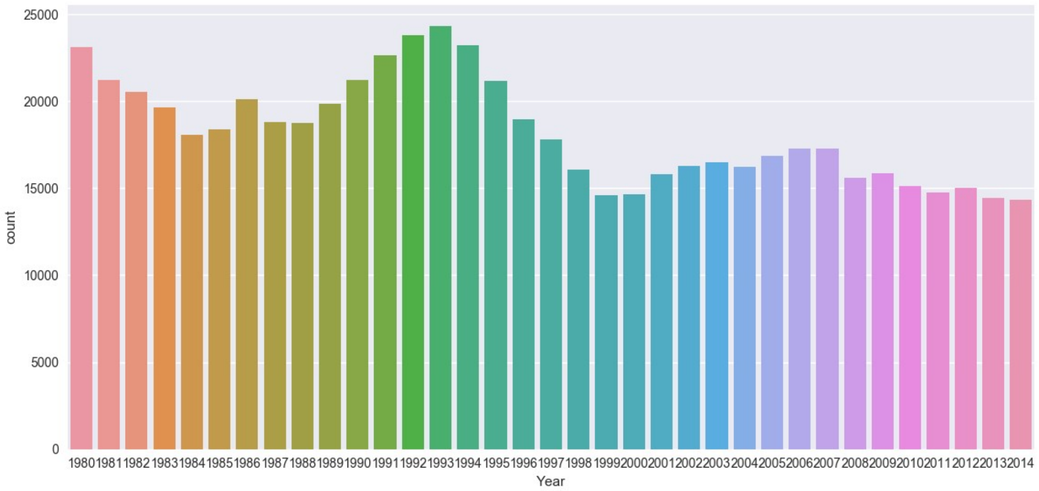


Figure 1.4

The above plotting (figure 1.4) is used for the most of the victims' counts in the dataset and the numbers of victims per year in the dataset. Thus, the most of the Victims/crimes were in 1994, which is close to our mean. The mean value for the year is 1995, which is close to 1994. As result, mean checking method is also suitable here to get close to deduction

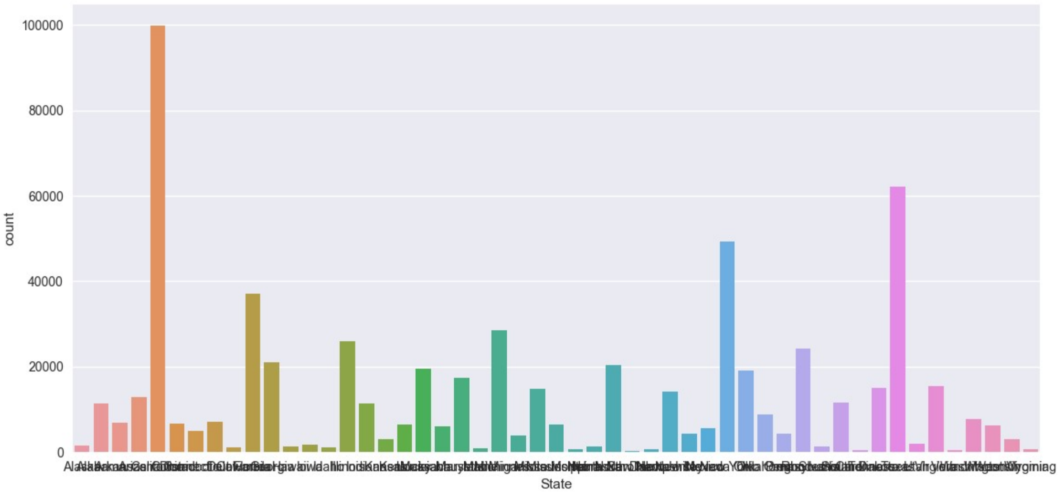


Figure 1.5

The figure 1.5 is the plotting of the states per crime. The lowest peak is the city with the low crime rate. The highest peak is the city with the highest crime rate. In the above scenario (figure 1.5), the Arizona state has the highest peak. Arizona state has the highest crime rate with respect to other cities and we have to update the focus point from New York to Arizona because it has the highest crime or murder rate in the whole New York according to our plot and it can lead to our final conclusion.

Correlation of data using heat map (other analysis method)

We have to find the relation between the values of the datasets. Indeed, heatmap is an efficient way to do this. We are going to have to check the relation between the independent variables. Therefore, we do have to verify the heatmap of the focused variables. Figure1.6 is the plot of different variables, which take part in our investigation.

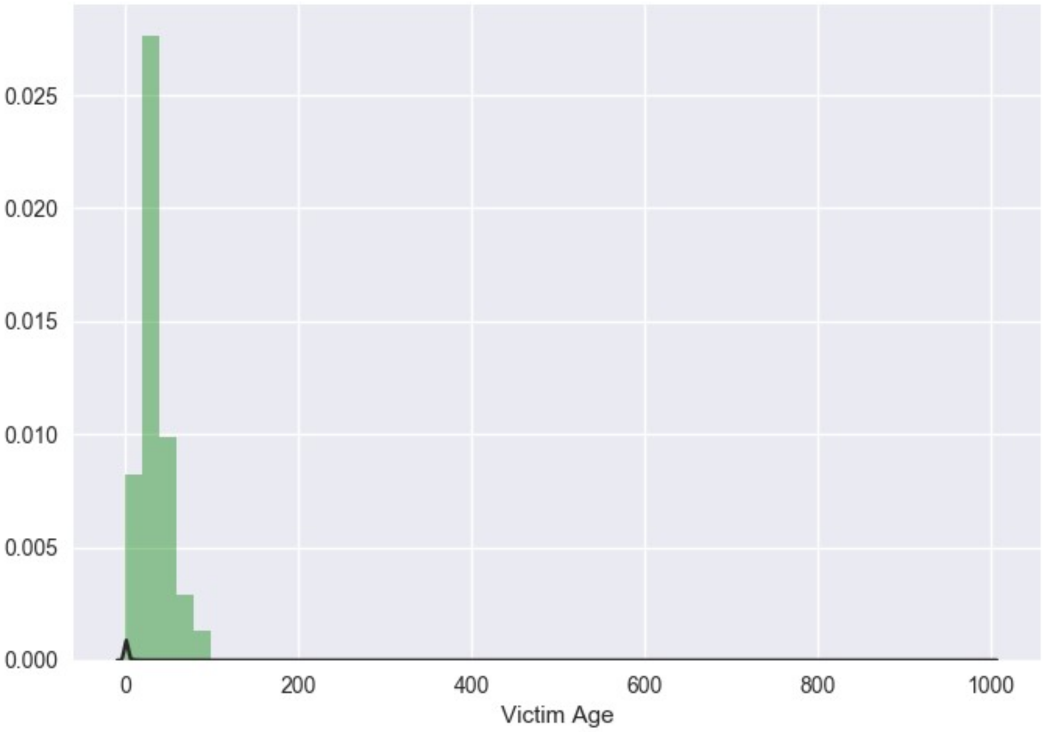


Figure 1.6

The figure 1.6 will plot the victim age values into visualization. By looking at above graph we can see that there is a huge diversity in the graph with respect to victim age. Therefore, we can also neglect the age factor that maybe we can assume that Victim Age is not the key to the murders that has happened in years.

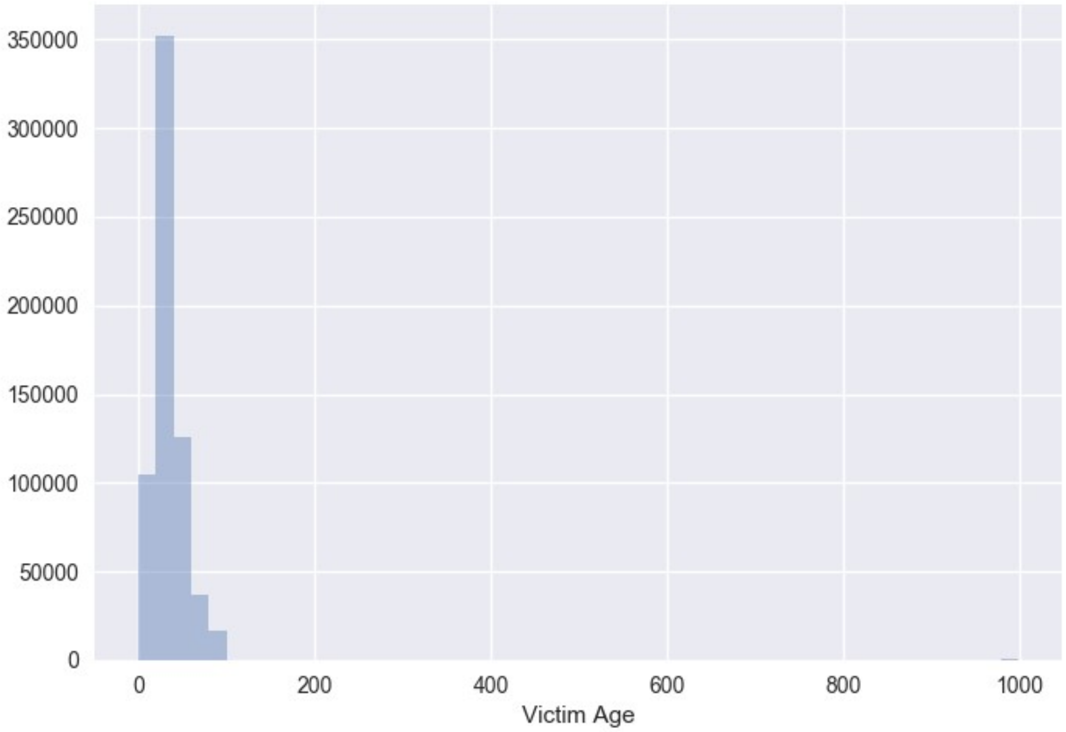


Figure 1.7

The above plotting (figure 1.7) is to counter check the perpetrator age vs. the victim age. We can verify whether there is a connection between the age or there is not.

K-means (unsupervised analysis method):

On the other hand, Wikipedia (2017) describe k-means clustering is an unsupervised analysis method for making clusters from the data, it is originated from signal processing and it is a popular method for cluster analysis in data mining. Wikipedia (2017) highlighted K-means clustering aims to make k number of clusters using n number of observations in which each cluster belong to observation with nearest mean which implies that the cluster should be focused and should be took forward for the further analysis. Thus, in our analysis k-means is very help full because we have to relate between different data of the dataset. We can use these k-means clusters in order to deduce our result. The k-means clustering uses plotly libraries and some other mentioned above.

Conclusion:

Finally, by applying the above mentioned strategies, one simple conclusion can be made and keeping in mind that our data was ambiguous enough due to some missing values in the data set like most of the values from weapon, relationship, Victim race and Perpetrator race are missing or unknown. Therefore, this can also affect our conclusion that perpetrator age, the victim age, perpetrator race and victim race can be the cause of the most murders/crimes in the New York. Most of the crimes have happened in the Arizona state and by keeping in mind that Arizona is our focal point here. Because Arizona most of the crimes have been lead to the race discrimination as the quote about race discrimination is mentioned above.

By applying k-mean clustering, from the clusters we can also deduce that race discrimination is our main point here, which cannot be neglected.

References:

Institute of Race Relation (2014) Racial Violence Statistics Available at: http://www.irr.org.uk/research/statistics/racial-violence/ (Accessed: 20 October 2017).

Kaggle (2017) Homicide Reports, 1980-2014 Available at: https://www.kaggle.com/murderaccountability/homicide-reports/data (Accessed: 28 September 2017).

Wikipedia (2017) K-means clustering Available at: https://en.wikipedia.org/wiki/K-means_clustering (Accessed: 22 October 2017).

Appendix:

Python Version: python 3.3

Libraries:

- Numpy
- Scikit
- Math
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- import numpy as np
- plt.style.use('fivethirtyeight')
- from sklearn import metrics
- import sklearn
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.model_selection import train_test_split
- from sklearn.cross_validation import KFold
- from sklearn.linear_model import LogisticRegression
- from sklearn import metrics
- from sklearn.model_selection import cross_val_score