EDA

On performing EDA:

1. December showed the least number of cases by a large margin. This could show that people are less rage prone around the time of Christmas, or that the law enforcement is much more negligent during the festive month.
2. We also see that the victims of black and white races are equal in numbers, but the case is that in terms of total population, the white people outnumber the black people in the ratio of 5:1, which shows that the people of darker skin colour are more likely to be a victim of murder.
3. Next we see that though number of Asian victims are low, there is a considerably large spike in the number of Asians in the plot of perpetrators. This could mean any number of things from Asians not being given an equal chance, or being discriminated against to Asians having a shorter temper compared to the others.
4. We also see that the the number of people between the age of 20 and 40 comprise of more than half of the victims. This doesn’t say much but we could interpret a meaning that this age group is more prone to dangerous endeavors, but since motives are not included in the database it is completely uncertain as to why there is such a spike in this age group.
5. We can also see that the number of perpetrators are considerably high in the range of 10-20. On further analysis, we can see that most perpetrator below the age of 20 are of white race. This could mean that the white people are much more hot-tempered and more prone to make such rash decision. This is against the common belief that black children are more likely to commit crimes since the grow up in ‘bad neighborhoods’, or can even show that white parents are not very attentive of their children’s behaviour. Once again, due to lack of motivations in the database, it is no more than speculation rather than analysis.
6. We next see that the the average number of crimes is higher pre-2000 than post-2000. We also see a rapid drop in the number of killings during they years of 1995 to 2000. This might have a number of reasons, many of which are mentioned in **Journal of Economic Perspectives** (https://pubs.aeaweb.org/doi/pdf/10.1257/089533004773563485)

It factors it down to four major reasons:

1. Increases in the Number of Police
2. The Rising Prison Population
3. The Receding Crack Epidemic
4. The Legalization of Abortion

The journal also summaries the changes in criminal law that were done during the said period as well.

**Q1]Which City has recorded the highest number of killings?**

This shows that the highest number of crimes have occured in Los Angeles and New York. While Cook and Wayne still have a higher number of crimes than most others, though not as high as New York and Los Angeles.

These are the top ten cities recorded to have the highest number of killings:

1. Los Angeles :44511
2. New York :38431
3. Cook :22383
4. Wayne :19904
5. Harris :16331
6. Philadelphia :12851
7. Dallas :11377
8. Jefferson :9573
9. Baltimore city :9336
10. Dade :9077

**Q2]Percentage of killing were the perpetrator was related to the victim.**

Since the data had certain entries where the relation was unknown, those were removed from consideration.

The percentages were found to be as follows:

Percentage of known relations : 73.57 %

Percentage of unknown relations/no relations : 26.43 %

SUMMARY

1. VMP: Verzeo Major Project

Our aim was to predict if a certain case would be solved or not.(Target feature: Crime Solved)

The selected features were:

Year,Month,Crime\_Type,Victim Sex,Victim Age,Perpetrator Age,Crime Solved,Relationship

There were no requirements for feature engineering in the chosen dataset.

We used 5 methods in our ensemble model:

1. Logistic Regression
2. KNN Method
3. KMeans Method
4. Random Forest Method
5. Support Vector Machines

(Since SVM takes much time computing only 50000 or less data is given to it. It appears that SVM had the highest number of accuracy most of the times, other times it was taken over by Random Forest.)

Using 40000 or more records would result in accuracy above 90% and the accuracy would decrease with decrease in number of records used.

\*The area related data were ignored as the attempt was to take into account the statistics of the country as a whole.

1. MLA and UsingMLA

The MLA contains algorithm made from scratch. UsingMLA imports each class relating to the specific algorithm and executes them. The created algorithm includes:

1. Linear Regression
2. Logistic Regression
3. K Nearest Neighbours
4. K Means

The algorithms were successful and were quite accurate( Logistic regression showed lower accuracy than the other algorithms and could be improved upon)

The algorithm of KMeans was not made in the conventional style but had a good accuracy nevertheless

Overall:

1. We used the scikitlearn package to make an ensemble model to predict if a case was solved using five algorithms.
2. We made 4 algorithm from scratch which had good accuracy.