An Analysis of Homicide Rates in the US

Data Mining CSCI 4502/5502

Girish Ramkumar
102358937, 4502
University of Colorado Boulder
Boulder, Colorado, USA
gira3678@colorado.edu

Nhi Nguyen
830381981, 4502
University of Colorado Boulder
Boulder, Colorado, USA
nhng5827@colorado.edu

Deekshitha Thumma
103370339, 4502
University of Colorado Boulder
Boulder, Colorado, USA
deth5087@colorado.edu

Soham Shah
101883949, 5502
University of Colorado Boulder
Boulder, Colorado, USA
soham.shah@colorado.edu

ACM Reference Format:
Girish Ramkumar, Nhi Nguyen, Deekshitha Thumma, and Soham Shah. 2017. An Analysis of Homicide Rates in the US: Data Mining CSCI 4502/5502. In *Proceedings of Data Mining (Project Proposal).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

The goal of this report is to find interesting patterns, correlations, and possible predictions using homicide rates data from 1980 to 2014. Prior work done by the FBI and CIA with this data has been very successful. However, we want to approach this data from a different point of view. We used different Data mining techniques we learned in class like correlation, classification and clustering to investigate particular questions using the dataset. We found that we can indeed predict if a homicide will likely be solved based on the homicide conditions up to a 71% accuracy. Additionally, we found that we can predict the relationship between the victim and perpetrator with a 52% accuracy. We also discovered that a homicide is easier to solve if the victim count is more than one, and even more if the perpetrator count is more than one. Finally, we found that after adjusting for population size, the highest rate of homicides reported by a very wide margin occurs in the District of Columbia. Moreover, the rate of homicide cases submitted increases from the months of June through September. These results can be very valuable for the citizens of the United States to give them information on how to stay safe and possibly prevent being a victim of homicide. It can also be helpful for investigators who need to determine who the perpetrator was or under what conditions the homicide took place.

Project Proposal, Fall 2017, Boulder, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 INTRODUCTION

2.1 Motivation

Our team is interested in analyzing homicide data from 1980 to 2014 to see and identify the different patterns that homicide revolves around and gain insight as to whether or not we can make some predictions that may be able to save some lives. One of the original challenges behind the data was to try and detect serial killers by analyzing the data. Data is very valuable and we want to take full advantage of the open data we have to be able to try and make the field more knowledgeable for the future. Government agencies like the FBI and CIA have done extensive investigation into the data they have collected. Now, this data is being offered to the public in order to inform people about homicide so that they can also gain an accurate perspective about homicide in the US. We are interested in using this data to gain new insights and learn more about this issue.

2.2 Prior Work And Challenges

A lot of prior work has been done on the dataset using the Kaggle platform. On the platform anyone can go online and share their analysis. They share their code on how they clean the data and then show interesting concepts that they have learned. Some looked at the amount of people killed using a certain weapon, for example, a handgun. Others examined who was being killed and seeing if there was a racial component to the killings. The FBI also uses this dataset to provide information to lawmakers in order to help them craft effective laws to decrease the amount of deaths. It's also used by the Bureau of Justice make sure that law enforcement is following the law. Their research can be found in Reference [3]
The challenge with this dataset is that it is often erroneous. The FBI asks other agencies to voluntarily share their data. This means that there can be errors or missing information in the data that needs to be removed before and analysis can be done. Also, it is important to understand what each column signifies. For example in the relationship column if it says "father" does that mean the father is the perpetrator or victim. It's important to gain context behind each datapoint.

2.3 Approaches

Our approach to the data was to take several questions we were interested in and apply data mining techniques we've learned in class to explore if we could answer the question. We employed techniques like decision trees, naive Bayes classification, and correlation analysis.

2.4 Contributions

Overall, we were able to train basic models which have a good amount of accuracy in predicting if a crime is solved or not. We split up the data set with 30% as training data and the rest for testing and found that we could accurately guess if a crime would be solved 70% of the time. We also were able to correctly guess the perpetrator and victim relationship 52% of the time.

2.5 Organization of Report

Within the Methodology section, there is first information about the data set, and then the various problems we investigated. It is then split into sections for each problem. Then for each problem there is a section for Preprocessing, Data Analysis, and Design.
Within the Evaluation, it is split into sections for each problem. For each problem there are various sub-tasks for evaluation under it.
The remainder sections of this report are not split by problem; the content of these sections reflect for all the questions together.

3 RELATED WORK

With the rise of social media and blogging, crime reporting is more prevalent and reaches a huge portion of the population of the United States. However, this kind of mass reporting is often not backed by statistical analysis and leads to a skewed perception of crime. Each year, the FBI compiled a data set of all crime in the United States. They accomplish this by accruing the count of all crimes committed that are reported by various law enforcement agencies who volunteer in the Uniform Crime Reporting program. The FBI also follows the Hierarchical Rule when counting crimes committed. Only the most serious crimes committed in a multiple-offense criminal incident are included in the count. In relation to homicide data, the FBI classifies this under violent crimes and provides general statistics, ie count of crimes committed, multi-year increases/decreases in crimes committed(reported as a percentage), arrest rate etc. See Reference [1].

4 METHODOLOGY

4.1 The Data

The main data set we used is Homicide Rates in Reference [2] which gives information about homicides throughout the United States from 1980 to 2014. This data set has 638,455 objects.
The attributes that this database consist of are: The attributes that this database consists of are: Agency Code, Agency Name, Agency Type, City, State, Year, Month, Incident, Crime Type, Crime Solved, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, Perpetrator count, and Record Source.

4.2 Problem Formulation

- Can we predict if a homicide is likely to be solved given the homicide's conditions?
  - For this problem we want to create some way to predict accurately if a certain homicides in particular conditions will be solved. We want to see if certain parameters of the homicide can help predict if the homicide gets solved. These parameters include victims age, victim's sex, victim's race, month of the year, and weapon used. For example, if a homicide occurs in May to a 31 year old white female with a rifle, will it be solved? This problem was investigated by Deekshitha.
- Can we predict the relationship between the victim and the perpetrator?
  - For this question, we wanted to see if we could help guide investigators on where to look for the perpetrator based on data regarding the crime. The hope was that data could help us identify who was likely to commit the crime given the victim and the information around the scene. By splitting up the perpetrator into three sets, Family, Acquaintances and Strangers we could begin to figure out who could potentially be responsible. The success criteria for the data mining is to be better than random guessing which group is responsible which in this case would be 33% accuracy. This problem was investigated by Soham.
- Does a homicide get solved easier if the victim count and/or perpetrator count is more than one?
  - In this problem, we investigate whether the number of victims and perpetrators affect whether the crime was solved or not. More specifically, we compare 1 victim or 1 perpetrator versus any other number with regards to the crime being solved or not. This problem was investigated by Nhi.
- Is there a correlation between homicide and time of year, location?
  - To address this problem, we investigate the average rate of homicides over time across the United States based on two factors, the location (on the state level) and the month. By identifying the rate of homicide cases submitted per state within the past 5 years (normalized to a population of 10,000 people) we can identify and order the states where the most homicides occur. We can then further the analysis by focusing on when the homicides occur in the top ten states for homicides. This problem was investigated by Girish.

4.3 Can we predict if a homicide is likely to be solved given the homicide's conditions?

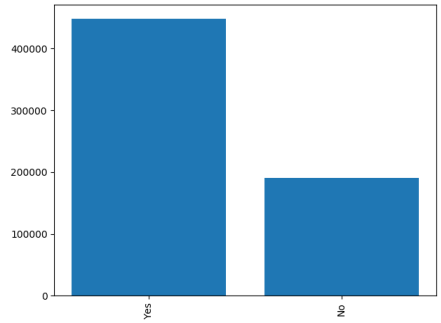
4.3.1 Preprocessing.

- Data cleaning: checked for missing data - For binary or nominal data, if the object was "Unknown" it was discarded from calculations. For numerical data (ages), if the object was 0 or > 100 it was discarded from calculations.
- Data reduction: removed attributes which do not pertain to this problem. The attributes kept include the following -

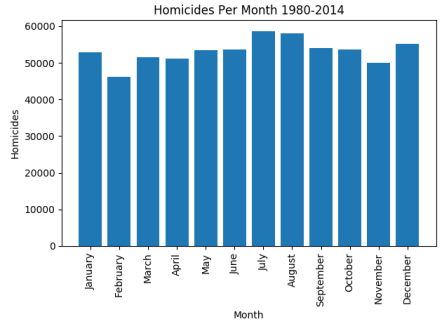
Crime Solved, Month, Victim Age, Victim Sex, Victim Race, and Weapon.
- Outlier Analysis: Global outliers for numerical date (Victim Age attribute) were found using a statistical approach of if an object was < Q1 - 3*IQR or > Q3 + 3*IQ and then discarded.
- Correlation analysis on Victim Age vs Crime Solved:the correlation coefficient is 0.037, which means negatively correlated. Had to transform the Crime Solved attribute from nominal to binary.

4.3.2 *Data Analysis.*

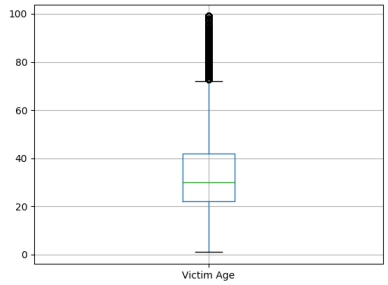
- Statistical summary for the attributes:
  - Crime Solved (binary): 70.17% Yes, 29.8% No
    The histogram below shows the number of homicides that were solved ("Yes") and unsolved ("No"), with objects with "unknown" for this attribute removed.



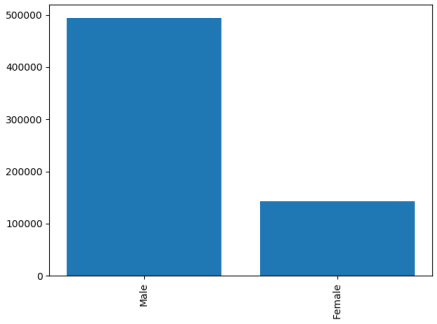
  - Month (nominal): 8.29% January, 7.22% February, 8.06% March, 8.02% April, 8.36% May, 8.40% June, 9.19% July, 9.09% August, 8.48% September, 8.40% October, 7.83% November, 8.64% December
    The histogram below shows the number of homicides that occurred in each month of the year, with objects with "unknown" for this attribute removed.



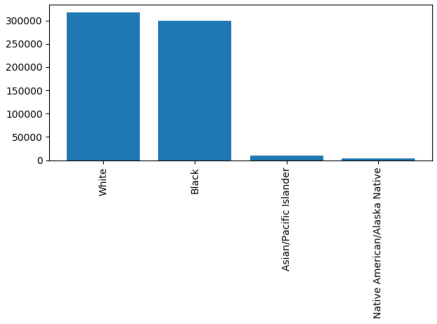
  - Victim Age (numeric): N - 629036, Min - 1, Max - 99, Mean - 34.013, Standard Deviation - 17.48, Q1 - 22.0, Median - 30.0, Q3 - 42.0, IQR - 20.0
    The boxplot below shows the distribution of victim ages with objects of <= 0 or > 100 for this attribute removed.



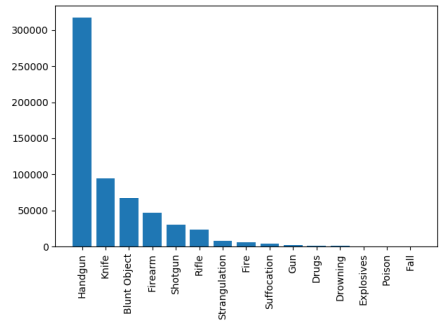
  - Victim Sex (nominal): 77.51% Male, 22.49% Female
    The histogram below shows sex of the victim, with objects with "unknown" for this attribute removed.



  - Victim Race (nominal): 50.24% White, 47.47% Black, 1.57% Asian/Pacific Islander, .72% Native American/Alaska Native
    The histogram below shows race of the victim, with objects with "unknown" for this attribute removed.



  - Weapon Type (nominal/ordinal): 52.45% Handgun, 15.69% Knife, 11.13% Blunt Object, 7.76% Firearm, 5.08% Shotgun, 3.86% Rifle, 1.34% Strangulation, 1.02% Fire, .656% Suffocation, .364% Gun, .257% Drugs, .199% Drowning, .0887% Explosives, .0750% Poison, .0314% Fall
    The histogram below shows weapon for the homicide, with objects with "unknown" for this attribute removed.



–

4.3.3 Design. For this question, we concluded that classification would be the major task since this is a classic prediction type problem. Decision tree is the best method of classification for this question because the question involves different attributes weighting into the prediction, so having a this top down method is perfect. Additionally, with the decision tree, two methodologies for attribute selection were used and compared: Information Gain and Gini Index.

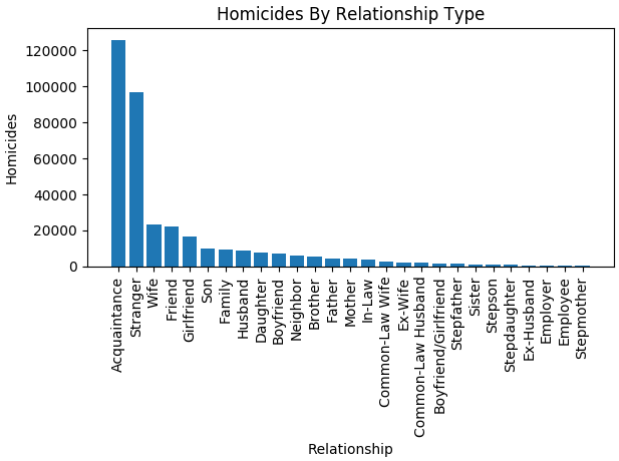
4.4 Can we predict the relationship between the victim and perpetrator?

4.4.1 Preprocessing.

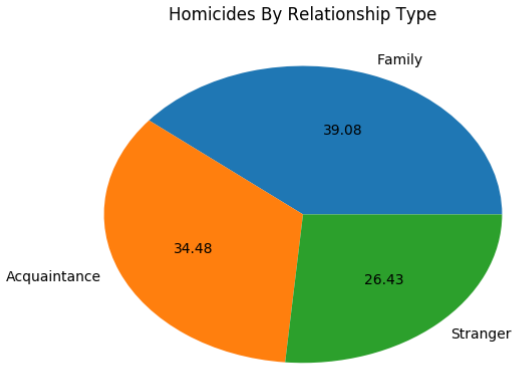
- Primarily, this task required cleaning up unnecessary data in the dataset. A good example of this was that for much of the data, the relationships were "unknown". Of the 638454 rows, 273013 were listed with an unknown relationship. This means that a little over 40% of the data needed to be thrown out. If this would be considered, then it would throw off the rest of the calculations. Therefore it was removed from the dataset.
- Another consideration that was taken into account in the process is figuring out what each relationship means and if that is valuable to the question. An example is that the dataset makes a distinction between a common-law-husband and a husband. Also, family is a separate relationship type on it's own. Naturally the question becomes, do we keep these relationships separate, merge them, or fold them under a bucket like family and not family. There was a long tail of different relationship types which will be shown in some of the later analysis. However, in order to simplify the problem statement, we chose to merge the relationships into three buckets: Family, Acquaintance, and Stranger.

4.4.2 Data Analysis.

- Here's a graph showing the relationship between victim and attacker for each of the homicides in the dataset.



- As you can see, there's a large amount of acquaintance and stranger homicides, and a long tail for the different familial relationships. It becomes hard to understand the significance of each relationship, so all the relationships were collapsed into three buckets. This leads to some very interesting results.



- It turns out that familial homicides are actually more common than Acquaintance homicides which are more common than Stranger homicides. The long tail actually was a larger sum than the first 2 bars. Here is how the count and percentages break down.

# Relationship	Homicides	Percent (%)
Family	142830	39.08
Acquaintance	126018	34.48
Stranger	96593	26.43
Total	365441	100

- Interestingly, it seems that if the relationship is known, there is a high likelihood that the crime is solved. So the question becomes, can we understand where to look for suspects, given certain factors? This would be very helpful in solving crimes. Basically, we wanted to examine if certain factors have more weighting than others to determine relationship. That's the analysis done in the next sections.

4.4.3 *Design.*

- Classification
  - Different Classification methods are evaluated to see which one is the most effective.
- Decision tree
  - A decision tree was constructed using different variables to see if the relationship between the victim and the perpetrator could be predicted.
- Correlation Analysis
  - Correlation analysis that was used which focused on finding the pearson correlation coefficient for each of the variables.

4.5 Does a homicide get solved easier if the victim count and/or perpetrator count is more than one?

4.5.1 *Preprocessing.*

- The Crime Solved attribute is a simple nominal case in which the entries can only foster yes or no, so no data preprocessing was needed for this.
- The Victim and Perpetrator Count attributes are described to be given as the number of *additional* victims or perpetrators (so 1 victim shows in the database as 0), thus some data pre-processing was needed here. 1 was added to every entry under both attributes.

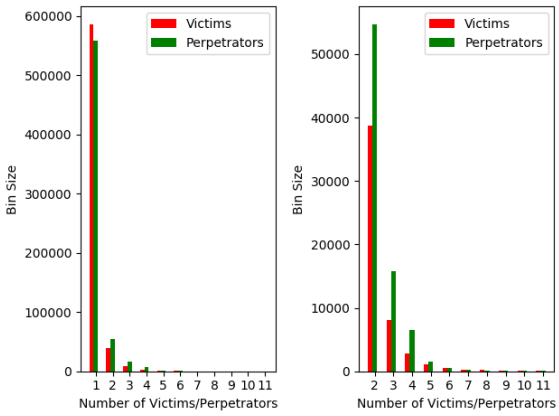
4.5.2 *Data Analysis.*

- Statistical summary for the attributes:

	Victims	Perpetrators
N	638454	638454
min	1	1
max	11	11
mean	1.123	1.185
Std dev	0.538	0.585
Median	1	1
Q1	1	1
Q3	1	1
IQR	1	1

  - The number of solved crimes in the database is 448,172 and the number of unsolved crimes is 190,282.
  - The numerical analysis for the victims and perpetrators isn't complex. The reason for this is because of the commonality of there only being 1 victim or 1 perpetrator. Below is a table that breaks down the number of entries in the database that contain 1-11 victims or perpetrators.

Number of Victims and Perpetrators



# Vics/Perps in crime	Occurrence-Vics	Occurrence-Perps
1	586059	558838
2	38750	54745
3	8156	15777
4	2847	6531
5	1084	1489
6	510	592
7	286	207
8	168	129
9	144	52
10	290	52
11	160	42

The results of the breakdown are quite interesting. It is shocking to see the magnitude of crimes that happen with 2-11 victims or perpetrators, especially on the higher end of victims/perpetrators. A visual depiction of the breakdown shows how insignificant 2+ victims or perpetrators are to 1. The graph on the left is the full breakdown that was shown in the table. The graph on the right is the removal of 1 victim and 1 perpetrator. 2 victims or 2 perpetrators still are much more common in the recorded crimes than 3+. Something also interesting to note is how the number of perpetrators (2-5) in the graph on the right seem much higher than the victims, which means that the perpetrators feel more comfortable with an accomplice.

4.5.3 *Design.*

- Correlation Analysis
  - The correlation analysis that was done focused on finding the lift coefficient to determine whether the number of victims of perpetrators affected if a crime was solved or not. In the lift analysis, 1 victim is compared to the "rest", or 2-11 victims. The same goes for the perpetrator count lift analysis. An additional test was performed checking 2 victims or 2 perpetrators to the rest, to compare to 1 victim or 1 perpetrator.
- Classification Analysis

– For classification, the classes used were solved and not solved, and X for one analysis was X = (1 victim, 1 perpetrator) and the for the other was X = (!1 victim, !1 perpetrator).

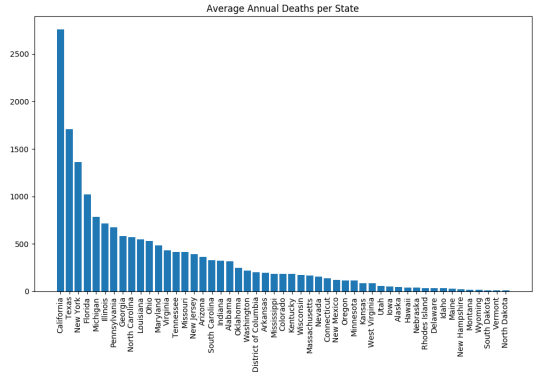
4.6 Is there a correlation between homicide and time of year, location?

4.6.1 *Preprocessing.*

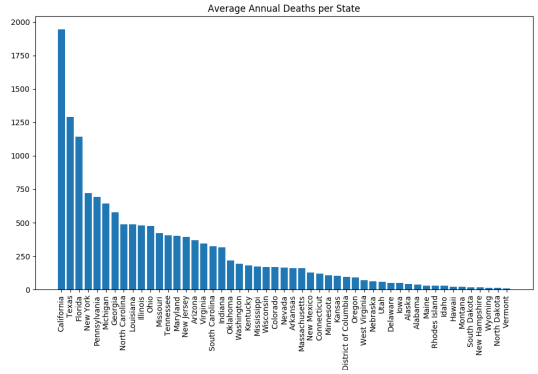
- The preprocessing task was mainly comprised of eliminating data that did not pertain to the question. This data reduction eliminated most of the variables within the dataset since variables regarding the victim, perpetrator and the agency were removed. These variables did not pertain to identifying a trend between location, time of year and homicides. The primary attributes kept to conduct analysis on include: Month, State, and Year.
- Another task in the preprocessing was to aggregate each case that was submitted. Since each row of the dataset was initially a description of each case submitted to the FBI, in order to get the count of cases submitted, these rows had to be condensed.
- Data integration was also a part of the preprocessing tasks. We incorporated a data set from the United States Census Bureau to use the population data that was missing from the kaggle dataset. From this we removed the geographical identification columns since we just required the numerical population data.

4.6.2 *Data Analysis.*

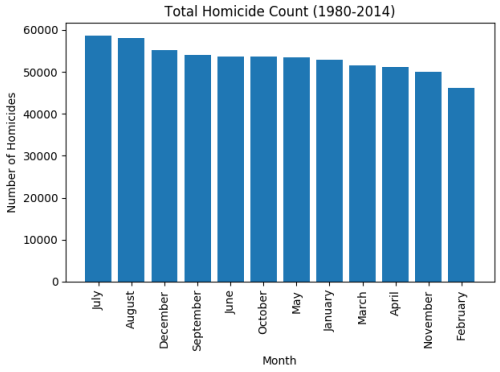
- The goal of the initial analysis was to identify long-term and short-term trends in the annual homicides recorded per state as well as total homicides recorded per month.
- The image below shows the annual homicides per state on a 34 year basis.



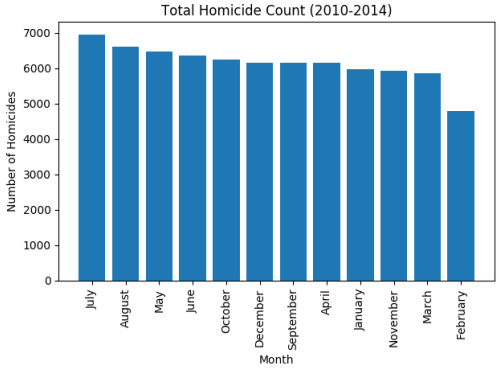
- The image below shows the annual homicides per state on a 5 year basis.



- Although the average annual rate of homicides was taken over a wide range of time, little difference is shown between the two graphs. The primary reason for this is due to the population size of each state. More homicides occur where there is a larger population so the states that have the most people are the most likely to have increased homicide counts. This does not hold much information so we will have to look at homicides per a certain number of people.
- The other trend that was investigated in the beginning was the total number of homicides recorded per month. This is primarily to see which months have the most homicide cases submitted.



- The image above is a 34 year summation of every homicide case submitted over 34 years whereas the image below is a 5 year summation of the same data.



- Whether over 34 years or just 5 years, we do see a vague pattern where the highest number of cases recorded is usually in July and August and the lowest number of cases is in February. However, while this does indicate that there could be a trend, an in depth evaluation is required to produce conclusive results.

4.6.3 *Design.*

- Trend Identification
  - For this question, the goal was to identify the time and location where homicides tend to be reported the most. In order to provide an accurate visualization of the most recent trend in homicides, we used the five most recent years in the data set (2010-2014). To prevent skewing of the homicides reported due to population, a census data set was used to normalize homicides per 10,000 people.

4.7 Implementation

Each member of the team contributed equally to the analysis. A common Github repository was used which includes the dataset and all our scripts written for the analysis. The original data is stored in a compressed form on Github. However, each team member stored it locally uncompressed in order to speed up computations without compromising the original data. This was a useful approach because it allowed us to share approaches we found effective while giving us the autonomy to work on the project when it was convenient to us. For example, the team worked off a basic python template which automatically pulled in the data from the CSV to Pandas and then creates an image which is saved into the plots folder. This way, each person just needed to modify the code for their question. A lot of this involved customizing the preprocessing involved in regards to the question itself. In the case the question on relationships, the data points where the relationship was undefined could be thrown out. However, this information was left in for other analysis. The main language we used was Python and the various libraries we implemented are: Numpy, Pandas, scikit-learn, graphviz, and matplotlib. We also used some functionality of Excel. Scripts were ran either locally or using Anaconda.

5 EVALUATION

5.1 Can we predict if a homicide is likely to be solved given the homicide's conditions?

The investigation of this question began with simple statistical analysis, then correlation analysis and then now finally classification. The attributes that will be used for this problem include: Crime Solved, Month, Victim Age, Victim Sex, Victim Race, and Weapon. All objects of the entire cleaned and preprocessed data base was used.

5.1.1 *Classification.*

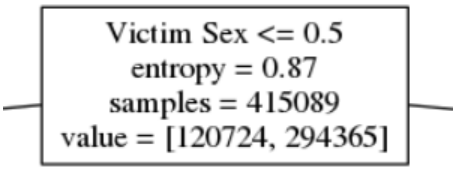
- For this problem, a decision tree was constructed. This tree using the below attributes and class labels allows us to predict if a homicide will be solved based on certain conditions.
  - The class label is Crime Solved
  - The attributes used are: Month, Victim Sex, Victim Age, Victim Race, and Weapon
  - Nominal data had to be converted to numerical for our code/script, their conversions are as follows:
    - Month: January -> 0, February -> 1 ... December -> 12
    - Victim Sex: Male -> 0, Female -> 1
    - Victim Race: White -> 0, Black -> 1, Asian/Pacific Islander -> 2, Native American/Alaska Native -> 3
    - Weapon: Handgun -> 0, Knife -> 1, Blunt Object -> 2, Firearm -. 3, Shotgun -> 4, Rifle -> 5, Strangulation -> 6, Fire -> 7, Suffocation -> 8, Gun -> 9, Drugs -> 10, Drowning -> 11, Explosives - 12, Poison -> 13, Fall -> 14
  - The maximum depth for the tree we chose was 6. This was chosen to prevent over-fitting of the data.
- Decision Tree using Information Gain:
  - The first decision tree was created using entropy or information gain as the attribute selection measure:

$$Gain(A) = Info(D) - Info_A(D)$$

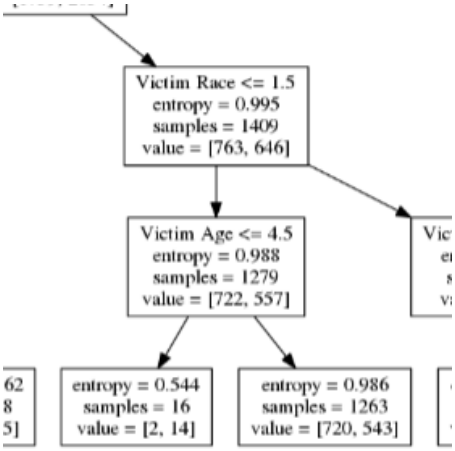
  - The training set was 70% from the whole dataset and was randomly chosen. The test set was remaining 30% of the dataset.
  - The information gain method had an accuracy measure of 71.1269%
  - Below is the confusion matrix for this method:

	Predicted Unsolved	Predicted Solved
True Unsolved	4341	47545
True Solved	3819	122191

  - Below shows the first node of the Decision Tree.



In the entire tree, the first attribute chosen was Victim Sex, followed then by Weapon and so on (entire tree is shown in source code). The "value" in the leaf nodes of the tree show if it would be classified as a No or Yes for the class label of Crime Solved. For example here we are zoomed into a part of the decision tree:



The root and some other nodes proceeding the top node in the image are not shown in image. The top node shows that if the Victim is White (numerical value of 0) or Black (numerical value of 1) it branches to the left while if the victim is any other race it branches to the right. Then, the next layer we have Victim Age. The left Victim Age node shows that if the age is less than or equal to 4.5 it branches to the left and if greater than 12.5 to the right. At the leaves we see the left leaf has a "value" of [2,14] which means it is classified as "Yes" for Crime solved since [a,b] stands for a amount of "No"s and b amount of "Yes"s. However, the right leaf shows "value=[720,543]" which means is Age was greater than 4.5, it would be classified as "No" for the class label Crime Solved.

- Decision Tree using Gini Index:
  - The second decision tree was created using gini index as the attribute selection measure:

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

  - The training set was once again 70% from the whole dataset and was randomly chosen. The test set was remaining 30% of the dataset.
  - The gini index method also had an accuracy measure of 71.12%!
  - The confusion matrix is also the same as the information gain confusion matrix.
- Predictions: After training and testing the decision tree, we can successfully predict the solved status of homicides with about 71% degrees of accuracy! The following are some situations we ran through the information gain decision tree to predict:
  - Given Month = May, Victim Age = 31, Victim Sex = Female, Victim Race = White, and Weapon = Rifle: prediction is Yes, crime is solved.
  - Given Month = July, Victim Age = 2, Victim Sex = Female, Victim Race = Black, and Weapon = Knife: prediction is Yes, crime is solved.
  - Given Month = January, Victim Age = 30, Victim Sex = Male, Victim Race = White, and Weapon = Handgun: prediction is No, crime is not solved.
  - Given Month = November, Victim Age = 72, Victim Sex = Male, Victim Race = Asian, and Weapon = Explosives: prediction is Yes, crime is solved.

5.2 Can we predict the relationship between the victim and perpetrator?

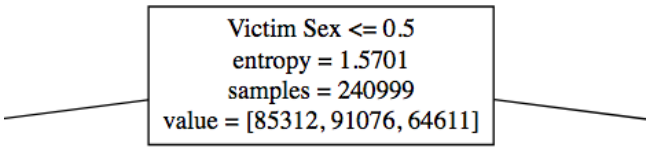
5.2.1 Correlation. The first thing that we did was a correlation analysis. The cleaned data was cast to integers using the process and code described above where string input is mapped to integers. We used the Pearson Correlation Coefficient which uses the following mathematical formula.

$$PearsonCorrelationCoefficient = \frac{cov(x,y)}{\sigma x \sigma y}$$

Here are the correlations for each variable to the relationship. The closest is Crime Solved. This makes sense because if a relationship is known then it is highly likely that the relationship is known. The next highest is Victim Sex which is interesting but also makes sense.

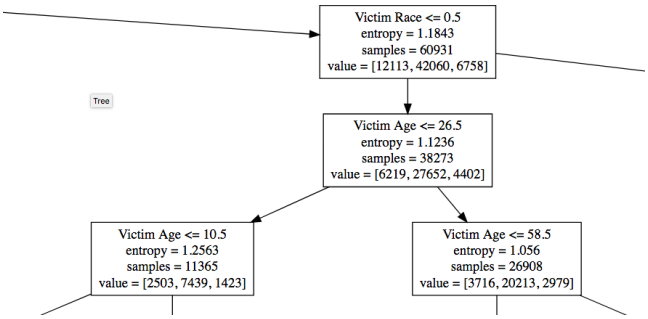
Variable	Correlation to Relationship
Crime Solved	-0.21
Victim Sex	-0.19
Weapon	-0.09
Race	-0.03
Victim Age	-0.022
Month	0.00

5.2.2 Decision Tree. A decision tree was constructed to see if that would help identify the relationship between the Victim and Perpetrator. The same process was followed as detailed above. Each metric was converted into numerical values in order to run the analysis. The accuracy for the Information gain decision tree was 52.56%, while the accuracy for the Decision tree constricted with the Gini coefficient was 52.4%. This is actually a pretty good accuracy considering the decision tree was determining between 3 different factors. It was trying to weigh whether a the relationship between the victim and perpetrator was that of family, acquaintance, and a stranger. Picking at random would generate about a 33% accuracy or lower based on the propensity for each type of crime. That the decision tree got 52% accuracy means that it is more accurate than random guessing by a factor of about 20%. Here's the first node of the decision tree.

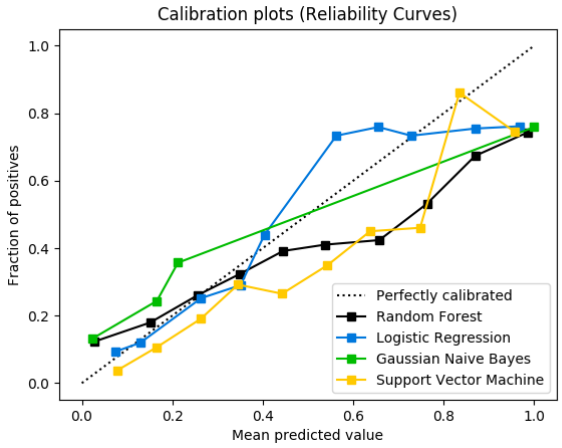


Interestingly, as with the earlier question, the Victim Sex appears to offer the most insight into the relationship between the victim and perpetrator. This is certainly something that could be worth

looking into in further analysis. The next nodes of the tree are whether the crime was solved and the victim age.



5.2.3 Classification. After that, we decided to compare multiple classification approaches to see which one was most effective using the Scikitlearn library. This is pretty straightforward to do. First the data is partitioned into test and traning data using the train_test_split_() function in scikitlearn. Then the different The specific classification methods are used and evaluated using the predict_proba() function. The classification methods used are, Random Forest, Logistic Regression, Gaussian Naive Bayes, and Support Vector Machines. After that, scikitplot was used to plot the effectiveness of each model. Here is the diagram showing the comparison of all the classification methods and their reliability curves.

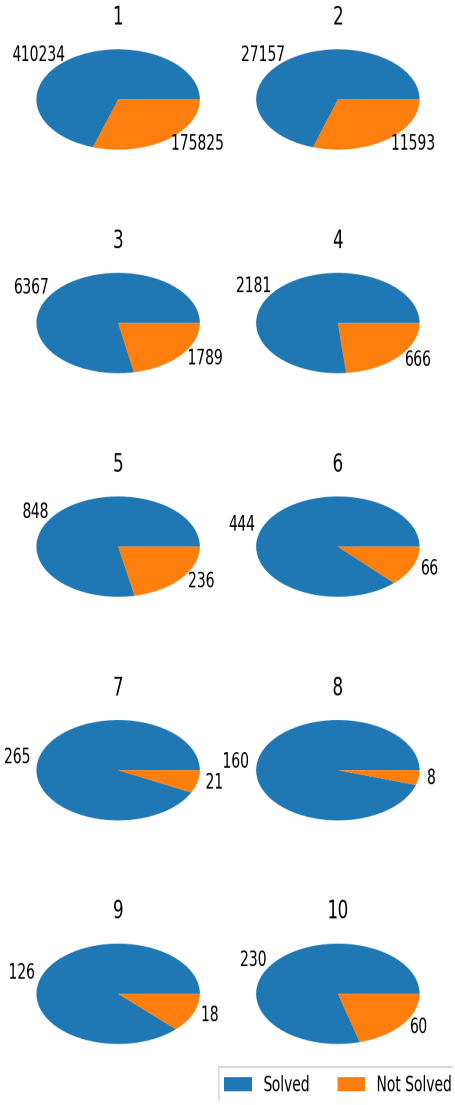


It appears that the each method converges to the same average without being tuned more.

5.3 Does a homicide get solved easier if the victim count and/or perpetrator count is more than one?

5.3.1 Correlation.

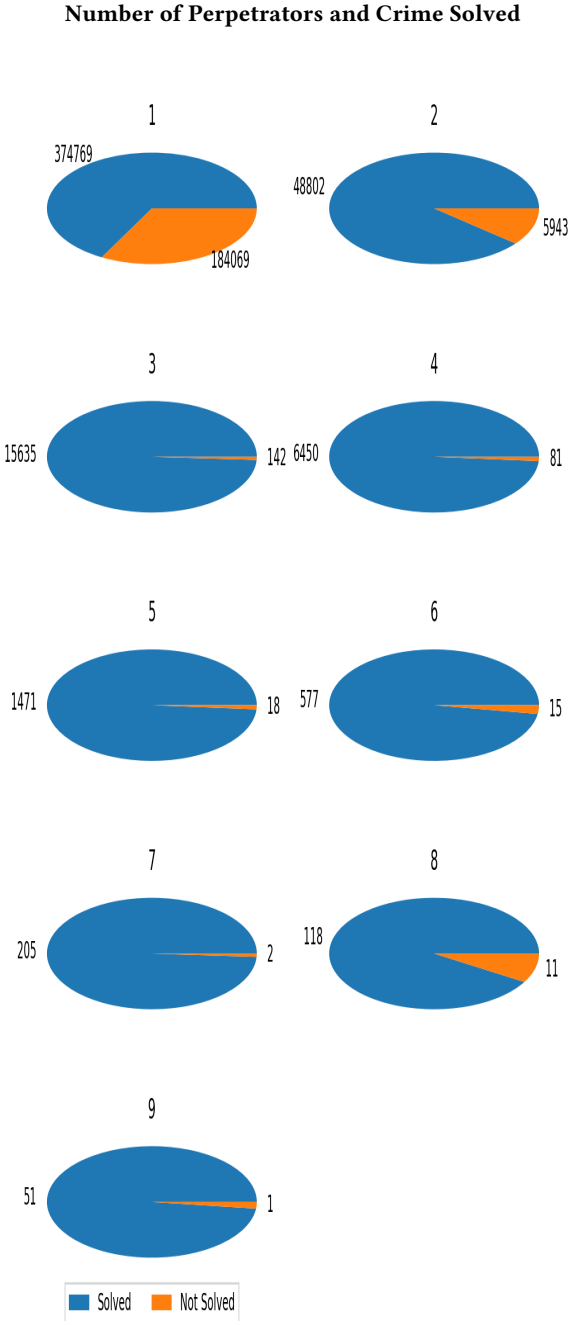
- Prior to the correlation analysis, it is important to notice the breakdown of victims and perpetrators and whether the crime was solved or not. Below, is the victim and crime solved pie chart visualization.



- The lift coefficient for 1 victim was solved using the following table.

	Solved	Not Solved	Totals
1 Vic	410234	175825	586059
Not 1	37938	14457	52395
Totals	448172	190282	638454

- Using this table and the lift coefficient formula, the results are as follows.
  - lift(1 Vic,Solved) = 0.997, negatively dependant
  - lift(1 Vic,Not Solved) = 1.007, positively dependant
  - lift(Not 1,Solved) = 1.032, positively dependant
  - lift(Not 1,Not Solved) = 0.926, negatively dependant



- The perpetrator and crime solved pie chart visualization is below (with the removal of 10 and 11 perpetrators with 52 and 42 solved respectively and 0 unsolved for both:
- The lift coefficient for 1 perpetrator was solved using the following table.

	Solved	Not Solved	Totals
1 Perp	374769	184069	558838
Not 1	73403	6213	79616
Totals	448172	190282	638454

- Using this table and the lift coefficient formula, the results are as follows.
  - lift(1 Perp,Solved) = 0.955, negatively dependant
  - lift(1 Perp,Not Solved) = 1.105, positively dependant
  - lift(Not 1,Solved) = 1.313, positively dependant
  - lift(Not 1,Not Solved) = 0.262, negatively dependant
- Conclusions we can draw are that the correlations match up whether or not a it is a victim or a perpetrator. 1 victim/perpetrator and not solved are positively correlated, as is not 1 victim/perpetrator and solved. This again lines up with the previous analysis conducted in which it is more likely that a crime gets solved when there is not just 1 victim or perpetrator. The lift value for not 1 perpetrator and with the crime not being solved is the farthest from 1 and is the most negatively dependant comparison.
- As an additional interesting study, the lift evaluation was calculated for 2 victims and 2 perpetrators.

	Solved	Not Solved	Totals
2 Vic	27157	11593	38750
Not 2	421015	178689	599704
Totals	448172	190282	638454

- And the lift results are:
  - lift(2 Vic,Solved) = 0.998, negatively dependant
  - lift(2 Vic,Not Solved) = 1.004, positively dependant
  - lift(Not 2,Solved) = 1.0001, positively dependant
  - lift(Not 2,Not Solved) = 0.9975, negatively dependant
- The lift values for the evaluation of 2 victims are *nearly* independent of each other. Here, we see that 2 victims were negatively correlated with the crime being solved.
- For 2 perpetrators:

	Solved	Not Solved	Totals
2 Perp	48802	5943	54745
Not 2	399370	184339	583709
Totals	448172	190282	638454

  - lift(2 Perp,Solved) = 1.267, positively dependant
  - lift(2 Perp,Not Solved) = 0.364, negatively dependant
  - lift(Not 2,Solved) = 0.179, negatively dependant
  - lift(Not 2,Not Solved) = 1.060, positively dependant
- These numbers are some of the farthest from the baseline of 1 that have been tested thus far. It is very positively correlated for 2 perpetrators to be busted by the FBI.

5.3.2 *Classification.*

- Bayes classifier was used to answer the question. For this evaluation, the 2 classes were from Crime Solved,

$$C_1 : Yes$$

$$C_2 : No$$

$$and \quad X :$$

$$X = (VictimCount = 1, \quad PerpetratorCount = 1)$$

$$P(Solved = "Yes") = 448172/638454$$

$$P(Solved = "No") = 190282/638454$$

$$P(VictimCount = 1|Solved = "Yes") = 410234/448172$$

$$P(VictimCount = 1|Solved = "No") = 175825/190282$$

$$P(PerpetratorCount = 1|Solved = "Yes") = 374769/448172$$

$$P(PerpetratorCount = 1|Solved = "No") = 184069/190282$$

$$P(X|CrimeSolved = "Yes") = 0.5373$$

$$P(X|CrimeSolved = "No") = 0.266$$

- What this tells us is that it is much more probable for the crime to be solved with 1 victim or perpetrator than for it to not be solved. This analysis was also performed on the "rest" of the dataset, which is from 2-11 victims or perpetrators.

$$P(Solved = "Yes") = 448172/638454$$

$$P(Solved = "No") = 190282/638454$$

$$P(VictimCount = 1|Solved = "Yes") = 37938/448172$$

$$P(VictimCount = 1|Solved = "No") = 14457/190282$$

$$P(PerpetratorCount = 1|Solved = "Yes") = 73403/448172$$

$$P(PerpetratorCount = 1|Solved = "No") = 6213/190282$$

$$P(X|CrimeSolved = "Yes") = 0.0097$$

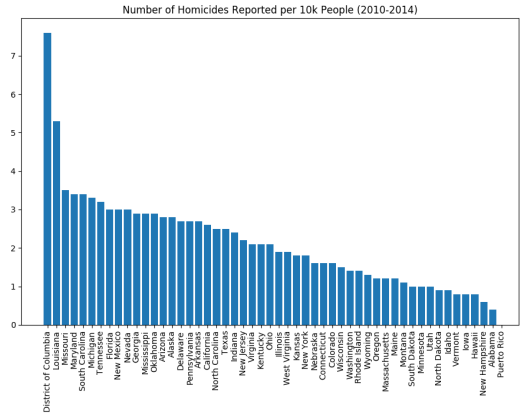
$$P(X|CrimeSolved = "No") = 0.0007$$

- Bayes classification tells us that it is far more probable for the crime to be solved with 2-11 victims or perpetrators than for it to not be. This aligns well with the correlation analysis done above.

5.4 Is there a correlation between homicide and time of year, location?

5.4.1 *Trend Analysis.*

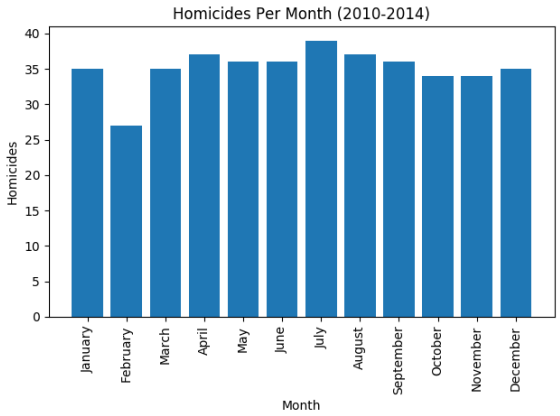
- Although the initial plots that were graphed for average annual homicides per state showed a reoccurring trend over the past 34 years, this result was not accurate. We did not account for population size when generating these graphs. The higher the population size, the more homicides are likely to be reported in that state, thus the data is skewed. We can confirm this by noting that California is consistently the state where the most homicides occur.
- To remedy this bias, we used a data set obtained from the United States Census Bureau [4] that contained the population of each state from 2010-2014. Using this data set, we can calculate the homicides committed per 10,000 people. The results are shown in the image below.



- As expected, the states with the most homicides reported have drastically changed, California is no longer the reigning state for homicides. These homicide values were rounded to the nearest whole number so there is a slight uncertainty associated with certain states.
- After correcting for the skewed population data, we can now use the accurate homicide estimates to investigate the effect of the time of year on homicide cases reported.
- The image below shows the average number of homicide cases submitted per month for the District of Columbia, the state with the highest homicides reported per 10,000 people.

Month	0	month_number
January	7.0	1
February	6.0	2
March	6.0	3
April	6.0	4
May	8.0	5
June	11.0	6
July	9.0	7
August	7.0	8
September	11.0	9
October	9.0	10
November	8.0	11
December	7.0	12

- From these results we can see that the number of homicides submitted spikes from June - September. However, these are the results for only the District of Colombia. The image below shows an aggregation of the top ten states for homicide cases submitted per month.



- From the image, we can see that the trend we identified in the District of Colombia persists throughout the top ten states. Homicides reported peaks during June,July, and August.
- This result supports the historical data shown above by confirming that the months June, July, August and September consistently report the most homicide cases submitted throughout the year for a wide of range of years.

6 DISCUSSION

6.1 Lessons

- Since the data set was quite large, we had to make sure that our code was optimized and did not allocate too many arrays for example. Before our scripts were optimized, some scripts resulted in because Memory Exception Errors. Thus, we had to make sure our code was efficient.
- Another lesson learned was that decision trees can be very large and can over-fit the data if you do not set a limit on the depth of the tree. So, we had to make sure to set a limit on the depth to get good results when creating decision trees.
- In regards to the problems of "Can we predict if a homicide is likely to be solved given the homicide's conditions?", We could not use any perpetrator info in the classification and prediction because of course if the crime was solved, all the perpetrator information is known but if the crime was unsolved, the perpetrator info was unknown. So, at first we included the "Perpetrator Age" attribute in the classification task but realized that once we removed all the "unknown" objects for that attribute, all the remaining objects were objects with "Yes" for Crime Solved. Thus, we decided not use any perpetrator attributes for classification, which then worked out great.
- We also learned that different questions/problems lend themselves to different parts of data mining. Not every problem will need every possible subtask of data mining.
- One important lesson learned was that prepossessing the data is very essential. The data is noisy, incomplete, and inconsistencies and if we had not cleaned, reduced, and transformed the data, all our our classification and correlation techniques would have failed and not been accurate!
- We learned that it's important to ask the right question. One of our original questions was, is there a correlation between the relationship of the victim and the perpetrator. This sounds like a meaningful question, but it really isn't. This is because the relationship operates in both directions. So, of course the if the perpetrator is the father the victim is the child. This is a 1-1 relationship. Therefore the question was modified.

6.2 Limitations

- Limitations of our work are the fact that it is entirely probabilistic. Because we have found trends in our data, it isn't always going to be the case that the specific relationship, conditions, or number of victims/perpetrators are going to decide whether a crime is solved or not. Factors such as the FBI unit, location, and the people involved in solving the crimes are going to vary the results.
- Another limitation is that this data is confined to a small range of data of the data that is available. The data set contains data between 1980 and 2014. While this is over three decades of data, it would be interesting to see how the trends have changed prior to 1980.

6.3 Possible Future Work and Applications

In the future, we would like to explore this data further by analyzing the homicide trends by time per state. We were not able to do this because of time and technology constraints. Analyzing each individual state by time would take a more resources than we currently needed for the problems we chose to pursue but would give some great findings.

Many homicides as we found get solved if the perpetrator was related to the victim. However, research and our work shows that if the victim was chosen at random, the homicide is much harder to solve. Possible future work could be investigating the data set further to really dig deep into the cases where the perpetrator and victim were not related and try to find some pattern in there. This could possible help solve current homicides and stranger crimes.

Some future applications of the discoveries we made include helping people avoid becoming a victim and helping spread awareness to certain demographics to be more cautious.

7 CONCLUSION

Overall, we find that we were able to successfully answer the questions we set out to answer using this data set. While there are certainly improvements that we could have made and additional things that we could explore, given more time, we feel that we did a sufficient job answering each question to a high degree of quality.

8 REFERENCES

REFERENCES

[1] "Offenses Known to Law Enforcement." FBI, FBI, 25 Aug. 2017. https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/expanded-offense

[2] Samaha, Kheirallah. âĂIJHomicide Reports, 1980-2014 .âĂİ Kaggle, 10 Feb. 2017. https://www.kaggle.com/murderaccountability/homicide-reports

[3] âĂIJHomicide.âĂİ Bureau of Justice Statistics (BJS),âĂİ Kaggle, 10 Feb. 2017. https://www.bjs.gov/index.cfm?ty=tptid=311

[4] United States, Congress, Population Division. âĂIJAnnual Estimates of the Resident Population: April 1, 2010 to July 1, 2014.âĂİ Annual Estimates of the

Resident Population: April 1, 2010 to July 1, 2014, 2014. factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

9 APPENDIX

9.1 Honor Code

On my honor as a University of Colorado at Boulder student I have neither given nor received unauthorized assistance on this work.

9.2 Contributions

9.2.1 *Deekshitha Thumma.*

- Abstract
- Motivation (written all together)
- All subtasks and work for the problem "Can we predict if a homicide is likely to be solved given the homicide's conditions?"
  - Cleaning
  - Preprocessing
  - Statistical Analysis
  - Outlier Analysis
  - Correlation Analysis
  - Classification: decision tree (two attribute selection methods)
    - Wrote script to generate decision tree and output it as an image
- Lessons
- Possible Future Work and Applications
- References

9.2.2 *Soham Shah.*

- Project report
- All subtasks and work for the problem, "What factors can we use to predict the relationship between the victim and perpetrator?"
  - Preprocessing
  - Statistical Analysis
  - Correlation Analysis
  - Decision Tree
- Set up the Github
- Imported the database into pandas for manipulation
- Wrote base script to read data, remove unwanted rows, merge into information, and generate image
- Generated images with MatPlotLib

9.2.3 *Nhi Nguyen.*

- Motivation (written all together)
- All subtasks and work for the problem "Does a homicide get solved easier if the victim count and/or perpetrator count is more than one??"
  - Preprocessing
  - Statistical Analysis
  - Correlation Analysis
  - Bayes Classification
- Limitations

9.2.4 *Girish Ramkumar.*

- Motivation (written all together)
- All subtasks and work for the problem "Is there a correlation between homicide and time of year, location?"
  - Preprocessing
  - Data Integration
  - Initial Statistical Analysis
  - Trend Analysis
- References