Group: Crime One
Team Members: Courtney Walker, James Bain, Keevey Song, and Andrew Pistole
Final Report
Due Friday, March 25th

The Data

Agency Data

The data were retrieved from the *2014 FBI National Incident-Based Reporting System*[1]. The most granular data set was a table of agencies that provided their respective crime counts. The following is a list of characteristics of the data set:

- 5454 Rows (each row is single reporting agency) by 70 Columns
- There were 4 independent predictor variables
    - **`State`** - which state in the US in which the agency is located
    - **`Region`** - an assigned region within the US. This feature was one in which we added by creating a dictionary in Python with the region label as the key and a list of states that belonged to each region as the values. We then looped through the States column and if the state was in the list of values for a particular region we mapped to it that key. Regions include *New England*, *Mid-Atlantic*, *Midwest (East)*, *Midwest (West)*, *South Atlantic*, *South Central (East)*, *South Central (West)*, *West Mountain*, and *West Pacific*.
    - **`Agency Type`** - a classification assigned to each agency. Labels include *Cities*, *Metropolitan Counties*, *Universities and Colleges*, *Non-Metropolitan Counties*, and *Tribal*.
    - **`Population`** - this is the population size of the agency. In some cases, just the reported sample from which the crime counts were collected or, in other cases, it was the entire population of the agency.
- There were 3 super-categories of crime, **`Crime Against Persons`**, **`Crime Against Property`**, and **`Crime Against Society`**. These super-categories could be split up into several other categories, that could be further split up into sub-categories.
    - **`Crime Against Persons`** are those offenses committed against an individual. Examples include *Simple Assault*, *Homicide*, and *Sex Offenses*. Features like *Homicide* could further be partitioned into *Murder*, *Negligent Manslaughter*, and *Justifiable Homicide*.
    - **`Crimes Against Property`** are those offenses committed against physical possessions like *Burglary*.
    - **`Crimes Against Society`** are those offenses deemed reprehensible such as *Prostitution* and *Gambling*.
- All crime counts were normalized by dividing each of them by the population of their respective agency. In other words, the crime counts became crime per capita for which we could now use to compare crime rates between agencies.

Other Data

We also wanted to incorporate Agency data from the years of 2012[2] and 2013[3] as this was the earliest that we could find from the FBI. However, data was very disparate in these files. For

[1] https://www.fbi.gov/about-us/cjis/ucr/nibrs/2014/tables/main
[2] https://www.fbi.gov/about-us/cjis/ucr/nibrs/2012/data-tables
[3] https://www.fbi.gov/about-us/cjis/ucr/nibrs/2013/data-tables

example, the 2012 data have NA values for a minimum of 35% for any given feature, with the majority of the columns having more. Therefore, the *majority* of the analyses stuck with the 2014 data.

The website also contained over 30 different summary tables that tabulated different reported crime statistics throughout the US. However, given that they were just summaries, one table couldn't be related to another. Therefore, we focused our modeling efforts on the Agency table.

Data Carpentry

Data files were originally in *xlsx* format and columns were not in a format to be directly read into Python. For example, each super-category of crime had a header above the individual components of that category. The figure below demonstrates the shift from one column header to a group of column headers nested under *Crimes Against Persons*. A function was written to

	Crimes Against Persons											
Crimes Against Society	*Assault Offenses*	*Aggravated Assault*	*Simple Assault*	*Intimidation*	*Homicide Offenses*	*Murder and Nonnegligent Manslaughter*	*Negligent Man-slaughter*	*Justifiable Homicide*	*Human Trafficking Offenses*	*Commercial Sex Acts*	*Involuntary Servitude*	
387	713	32	502	179	1	1	0	0	0	0	0	
565	493	64	355	74	2	2	0	0	0	0	0	
2,318	1,241	143	897	201	0	0	0	0	0	0	0	
1,270	1,316	419	747	150	4	3	1	0	0	0	0	
466	559	123	327	109	6	5	1	0	0	0	0	

handle this messy format in order to read the file into Python in a clean, data frame format. This function also cleaned up the column headers (ie. removed spaces with underscores, removed numbers, transformed to lowercase, etc.).

We also added some features that we thought might be helpful in our models. This includes the Region column mentioned above as well as a state code column. This state code column was created in R instead of Python as R has a built in function to transform state names to state codes. Below is the snippet of code to create the state codes from the *State* column. The only

```
df14$codes<-as.factor(state.abb[match(df14$State,state.name)] )
```

purpose for this feature was for plotting a map in the plotly package for which uses state codes to define the map boundaries.

Questions We Sought to Answer

The majority of our questions revolved around violent crime prediction, primarily homicide rates and its various forms. The following questions were explicitly investigated:

- *Can the size of the population predict the rate of murders of an agency?*

    - There are a lot of complexities involved with murder at the population level. Certain factors have been used to explain why murder rates are higher in some places than others such as inequitable distribution of resources[4]. These factors often piggyback with population size shifts, which served as the foundation of our curiosity.
- *Do particular states (at least for the data provided) have different rates of homicide?*
    - Aggregating our data by summed crime rates per state seemed like a natural place to start. States provide a nice, albeit somewhat arbitrary, boundary to group populations.
- *Are different types of homicide predicted better by population or state?*
    - For example, murder involves intent but negligent manslaughter should theoretically be accidental. If this were indeed the case, you should not see a significant effect of population size on the rates of negligent manslaughter.
- Do certain regions have varying rates of homicide rates?
    - Groups of states often share cultural and political characteristics. What's not to say that regions may also have similar crime rates particularly if these other characteristics play into it?

Models

Some of the models were better at tackling the questions above. Since our response variables were continuous, discretization was necessary in order to get meaningful output in a Naive Bayes Classifier and Support Vector Machines.

Naive Bayes

What was tried?
The questions that I attempted to answer revolved mainly around whether or not population size could explain homicide rates. Intuitively, there are a lot of factors at play in predicting the rates of homicide at the population level, but I wanted to keep the model clean and simple. Plus, we didn't have very many predictor variables to begin with. I also used state and region as predictors in some models to see if some of the variance in crime rates per capita could be explained by regional phenomena. This left me with the following models:

1. Murder ~ Population Size
2. Murder ~ Population Size + State
3. Murder ~ Population Size + Region
4. Negligent Manslaughter ~ Population Size
5. Negligent Manslaughter ~ Population Size + State
6. Negligent Manslaughter ~ Population Size + Region

[4] Daly and Wilson (1988). *Homicide*.

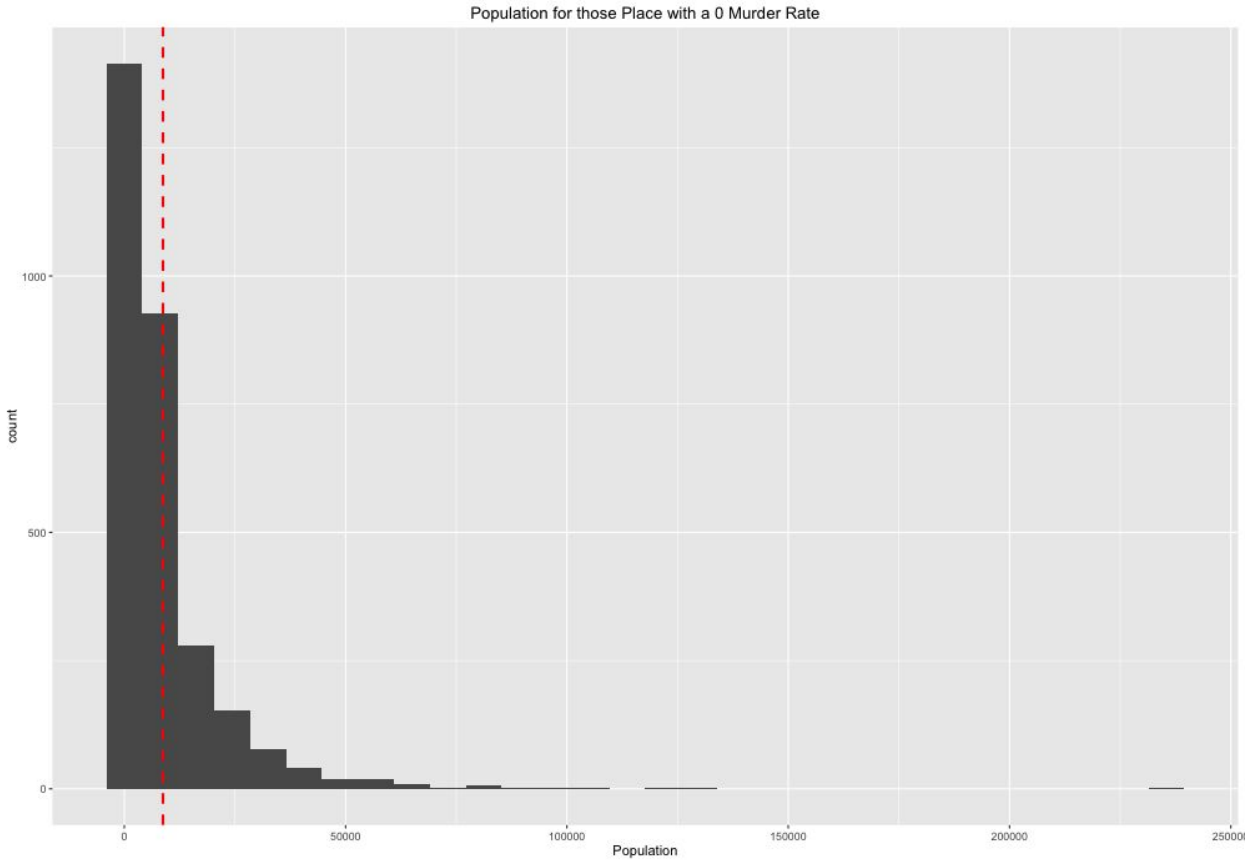
Further data carpentry was need in order to get that data in a form usable by a Naive Bayes classifier. First, NAs were removed from the variables of interest. This wasn't done in earlier phases of data carpentry because there were some crimes in which the majority of agencies didn't report any numbers. If I had removed the NAs at this point, the majority of the data would have been removed.

I also needed to discretize the response variable (Murder and Negligent Manslaughter) into bins of equal size that represented the homicide rate as a range. One bin would start at 0 and go to x. The next would start at x and go to y, and so on and so forth, for 5 bins. In order to do that, I borrowed a custom function[5] that split the variable up into bins of equal size and provided the ranges of each bin. 5 bins was simply decided based off of model optimization. Anything higher proved to be less predictive and anything lower wasn't as meaningful.

Problems and Resolutions

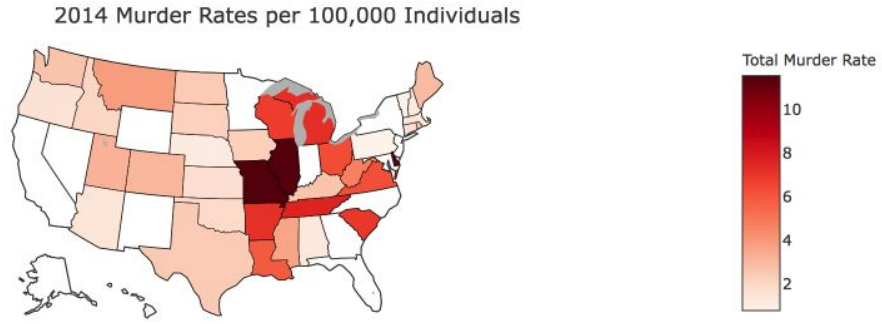
- Discretization of the response variable:
    - Problem: At first, I ran the discretization function on the entire **`Murder`** column and split it into 5 bins of equal size. Then I trained the model on the data and found that it had about an 85% accuracy, which I initially thought was wonderful. However, when I went back to see what the bin ranges were after discretization, I realized that 4 out of the 5 bins were between the values of 0 and 0 and the last bin was between 0 and the max value for murder per capita in the data. It turned out that 2956 agencies (85% of the data) had reported 0 for the for their murder count, which explains why there were four bins of only zeros in original classification model.
    - Solution: I decided to change the question a little bit, since I wasn't really interested in those places that didn't report murders, so I removed those agencies, which changed the classification problem to predicting murder rates for those places that reported at least 1 murder. This, however, does not mean that they are not important to the overall question, which is why I ran a summary on those places who reported 0 murders.
        - The mean population size for these places was 8,733.275
        - `Min: 8` (Lakeside, CO), `1Q: 1914`, `Median: 4351.5`, `3Q: 10,370`, `Max: 235,400` (Gilbert, AZ)
    - Below is a histogram of those places that reported 0 for murder rate. The graph is skewed right and the dotted line is the mean population size. Many of these rates

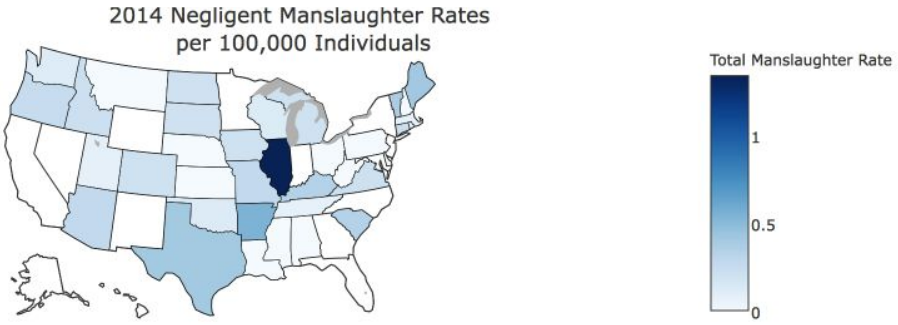
[5] Function provided by *Joris Meys* at http://stackoverflow.com/questions/5731116/equal-frequency-discretization-in-r



are obviously true. For example, Lakeside, CO, population of 8, probably didn't have a murder for the year 2014. However, Gilbert, AZ, for whatever reason doesn't have a value when indeed they most likely had some. Therefore, some of these points are most likely noise.

- Region as a predictor
    - Problem: Regions of the United States turned out to be a lousy predictor of homicide rates. The understanding behind it was that perhaps regions undergo similar economic and cultural phenomena, which may interplay with crime rates.
    - Solution: The simplest solution was to rid it from my models.
    - Below are some maps of the United States with each state having an aggregated homicide rate per 100,000 individuals. Purple is the category of *homicide* and is broken up into murder and negligent manslaughter, which are red and blue respectively. You can see that while there are differences between states, these differences are rarely regional.





Naive Bayes that Worked

After taking out the those places that didn't have a murder rate, the best predictive model was `murder ~ population + state`, although `murder ~ population` also fared alright with about a 35% accuracy.

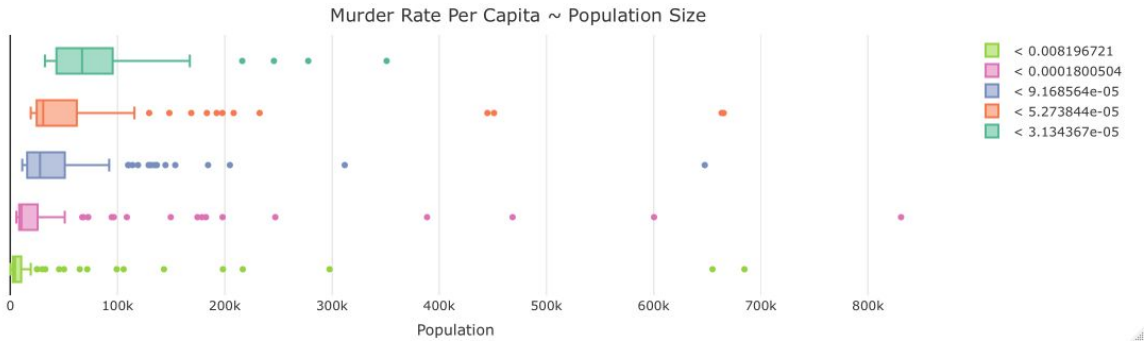
- I was left with about 500 rows to run my model on. I trained the model on 60% of these data and evaluated its accuracy on the other 40%.
- The output of the model is below. The accuracy was about 47%. This means that population along with state could partially predict murder rates.
- Notice that on the confusion matrix below the output, predictions are generally thrown into the correct bin. However, those that aren't in the correct bin are often thrown into a bin directly next to it.

usekernel	Accuracy	Kappa	Accuracy SD	Kappa SD
FALSE	0.2940286	0.1150967	0.0689399	0.08713416
TRUE	0.4753757	0.3432475	0.1051546	0.13187373

```
> table(predict(model$finalModel,xTest)$class,yTest)
                yTest
                 < 3.134367e-05 < 5.273844e-05 < 9.168564e-05 < 0.0001800504 < 0.008196721
  < 3.134367e-05             19             12              5              1             1
  < 5.273844e-05              7             14              6              5             1
  < 9.168564e-05              7             11             14              7             4
  < 0.0001800504              4              5             13             22            12
  < 0.008196721               5              4              4              3            26
```

Interpretation

By itself, the output does not provide a lot of information on what the effect of population is on the crime rate. However, the plot below gives us a lot better idea of why the model was able to predict based off of population size. As mentioned earlier, bins are discretized based on murder rate ranges. The colors of these boxplots are the separate bins where the legend shows the



bins from greatest murders per capita to lowest from top to bottom. The label by the bin is the max value of that bin and the min value is represented by the max value of the bin below it. Notice how the highest crime rates per capita usually belong to the the smaller populations, while crime appears to go down as population increases.

It is important to keep in mind that this is not considering those places that have no murder. Given our results for linear regression, as well as analyzing the population size of those places, it appears that these agencies tend toward the smaller size.

State also explained some of the variance in murder rates within the population (*reference the maps above*). This suggests that some places are prone to higher incidents of murder per capita

while others are not. However, some of the states could be skewed by the inadequate sample provided. For example, Illinois, according to the map, has one of the highest murder rates, but this number is derived from a single agency. It is hard to tell whether or not this agency is representative of the entire state, however, given its high value in comparison to the rest of the country, I am led to believe that it is an outlier, but one cannot be certain of this.

Linear Regression

The first thing I did is to find which two of those variables have linear relationship. My approach is to throw two variables into linear model and show summary. We can find that the more stars it gives us, the more significant relationship those two variable would have.

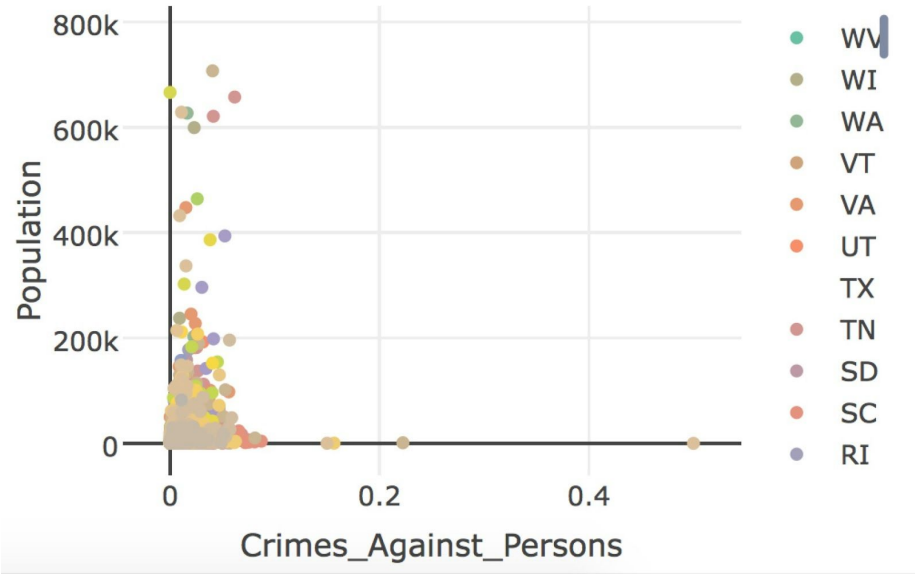
```
Call:
lm(formula = prop12$Theft_From_Building + prop12$Theft_From_Coi
    prop12$Theft_From_Motor_Vehicle + prop12$Theft_of_Motor_Vehi
ies ~
    prop12$Population)

Residuals:
      Min        1Q    Median        3Q       Max
-0.011514 -0.004677 -0.002006  0.002512  0.072244

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.139e-02  3.344e-04   34.06   <2e-16 ***
prop12$Population 5.744e-09  3.589e-09    1.60     0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007922 on 701 degrees of freedom
  (4532 observations deleted due to missingness)
Multiple R-squared:  0.00364,   Adjusted R-squared:  0.002219
F-statistic: 2.561 on 1 and 701 DF,  p-value: 0.11
```

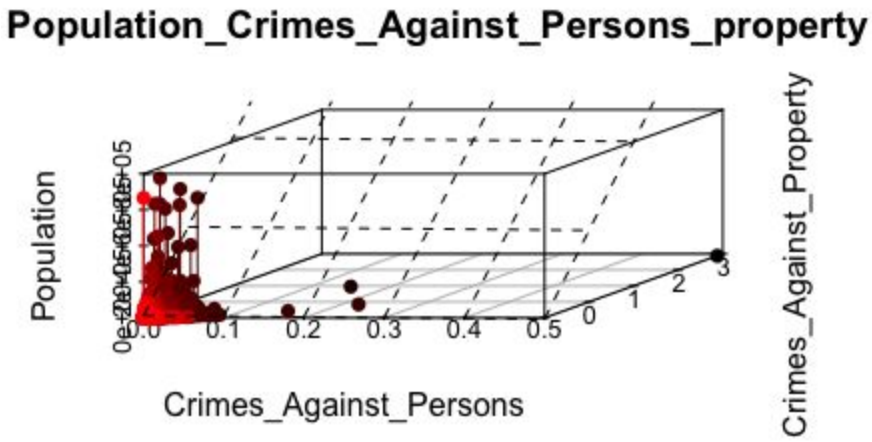
```
> e<- lm(prop12$Simple_Assault~prop12$Population)
> summary(e)

Call:
lm(formula = prop12$Simple_Assault ~ prop12$Population)

Residuals:
      Min        1Q    Median        3Q       Max
-0.013942 -0.005319 -0.002128  0.002851  0.241877

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.123e-03  1.691e-04  48.033  < 2e-16 ***
prop12$Population 1.650e-08  3.656e-09   4.514  6.6e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008958 on 3211 degrees of freedom
  (2022 observations deleted due to missingness)
Multiple R-squared:  0.006305,  Adjusted R-squared:  0.005995
F-statistic: 20.37 on 1 and 3211 DF,  p-value: 6.603e-06
```

Before I got to the right spot, I tried several different plots to interpret our questions, like using plotly or 3d linear model.

```
plot_ly(prop12, x = Crimes_Against_Persons, y = Population, text
= paste("Agency: ", Agency_Name), mode = "markers", color =
codes)

fit <- lm(Population ~ Homicide_Offenses, data=prop12)
#abline(Population ~ Homicide_Offenses)
plot(fit)

attach(prop12)
plot3d_mk <-scatterplot3d(Crimes_Against_Persons,   # x axis
                          Crimes_Against_Property,     # y axis
                          Population,    # z axis
                          pch=16, highlight.3d=TRUE,
                          type="h",
main="Population_Crimes_Against_Persons_property")
fit_mk <- lm(Population ~ Crimes_Against_Property +
Crimes_Against_Persons)
```

```
plot3d_mk$plane3d(fit_mk)
```

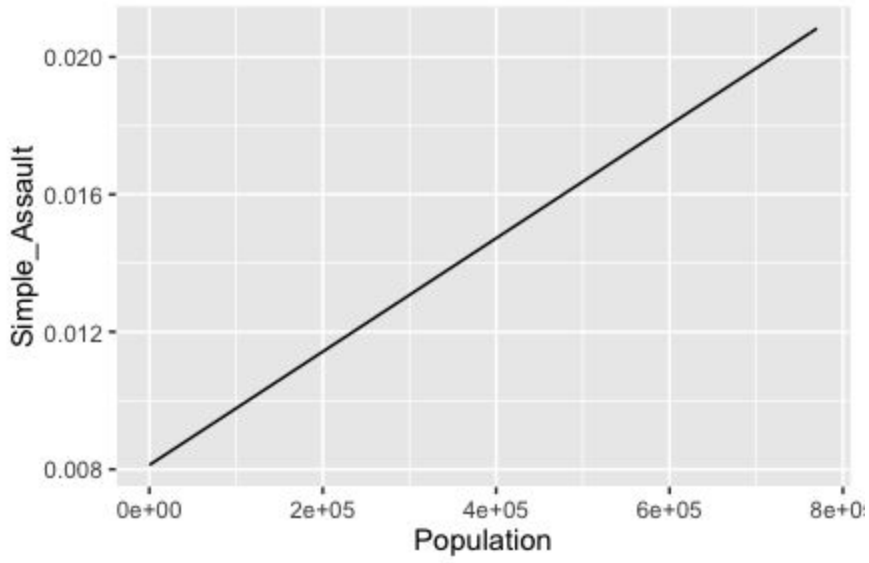




But i found them are not predicting very well and eventually I tried ggplot to help me do the linear regression.

And I also tried to draw the line for the plots. But it still didn't work very well.

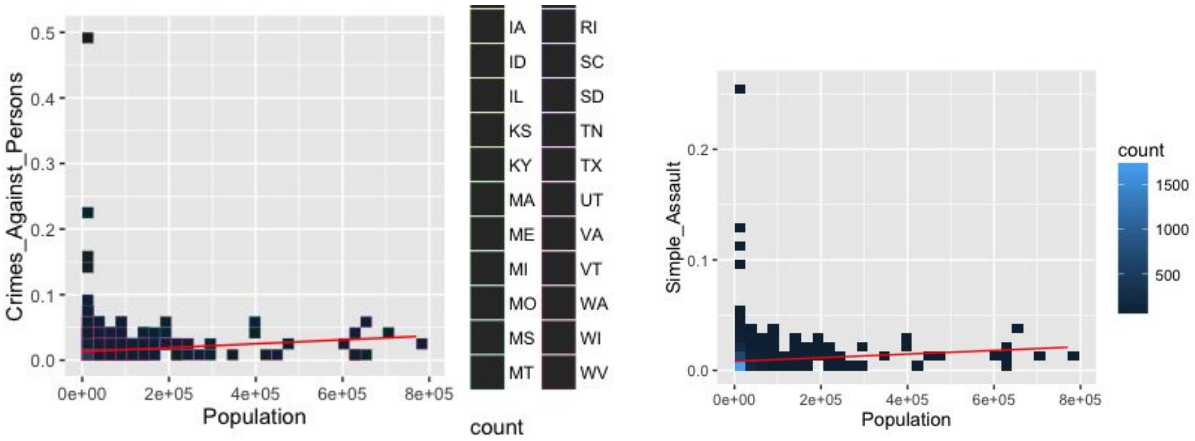
```
fun1 <-  function(x){8.123e-03 + (x*1.650e-08)}
coef <- lm(prop12$Simple_Assault~prop12$Population)$coefficients
ggplot(prop12,
aes(x=Population,y=Simple_Assault))+stat_function(fun =
fun1)+geom_abline(slope=coef["x:Simple_Assault"],
intercept=coef["(Intercept)"])
```



As the plot shows, it did a pretty good job about drawing the line fitting into the linear regression, but it doesn't make much sense because I don't even know what it is based on to draw that line. Eventually, James helped me to figure out the way to do the linear regression on the right way.

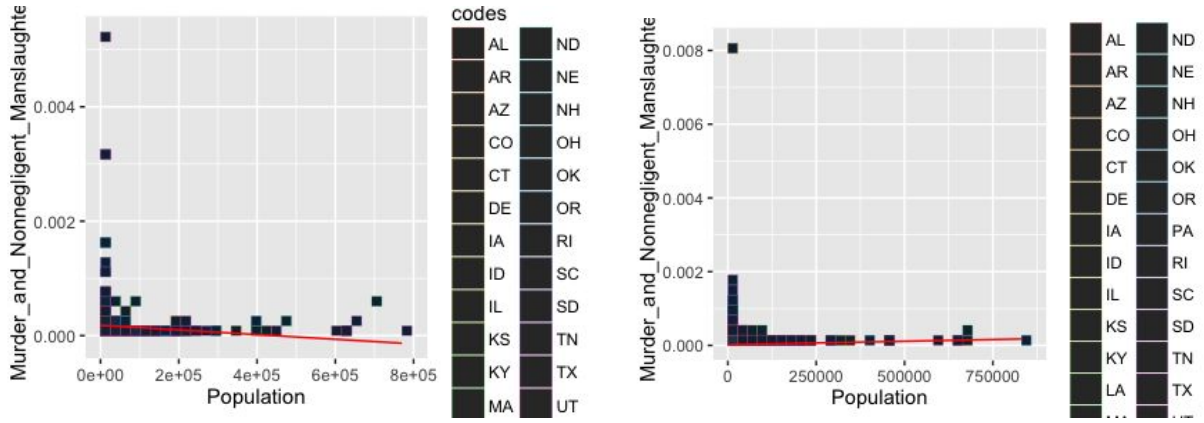
```
fun6 <-  function(x){1.924e-05 + (x*1.865e-10)}
ggplot(prop14,aes(x=Population,y=Murder_and_Nonnegligent_Manslaughter, color=codes)) + geom_point() +stat_function(fun = fun6, colour='red')
```

After that, things were going pretty well, I threw those correlated variable that we focused on to the ggplot and tried to interpret it.
Those two variables (crimes against person and simple assault) both go the positive way versus the population size.

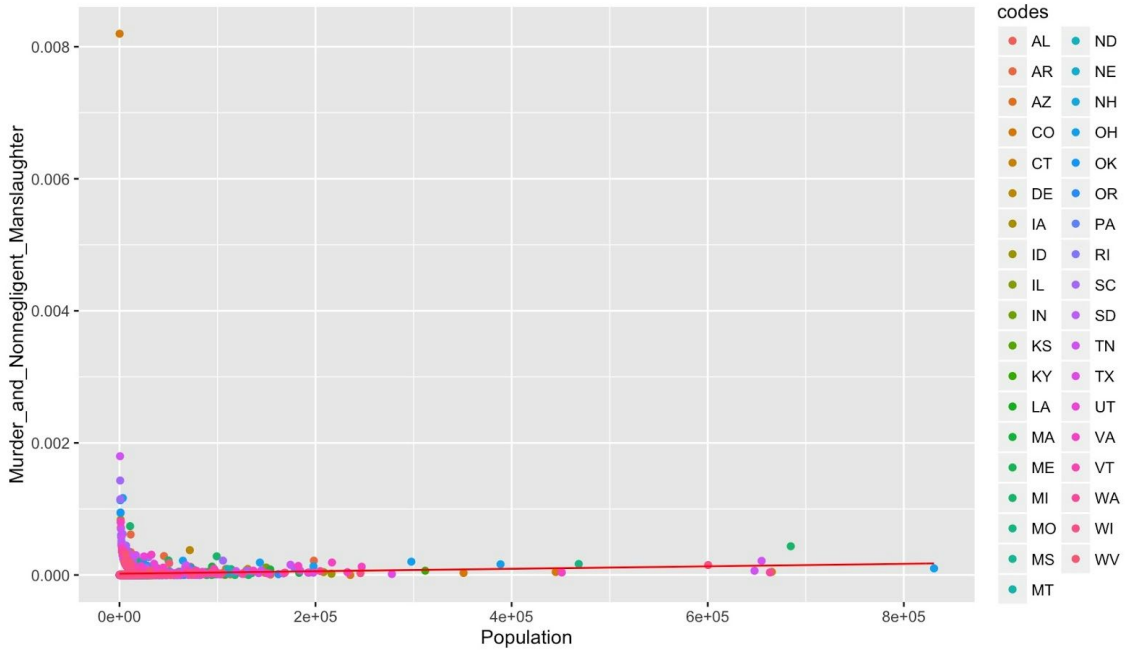


Compare to the 2012 dataset and the conclusion that James's naive bayes model gave us, I found something weird. The left picture posted below is extracted from the 2012 dataset, which predicts the same result with James's naive bayes model. the bigger population size is, the less murder would happen. And the right picture is from the 2014 dataset, but why does it show the

opposite result with the 2012 one? We figure out maybe because all those zero values James mentioned before.



we can see more clearly from this plots.



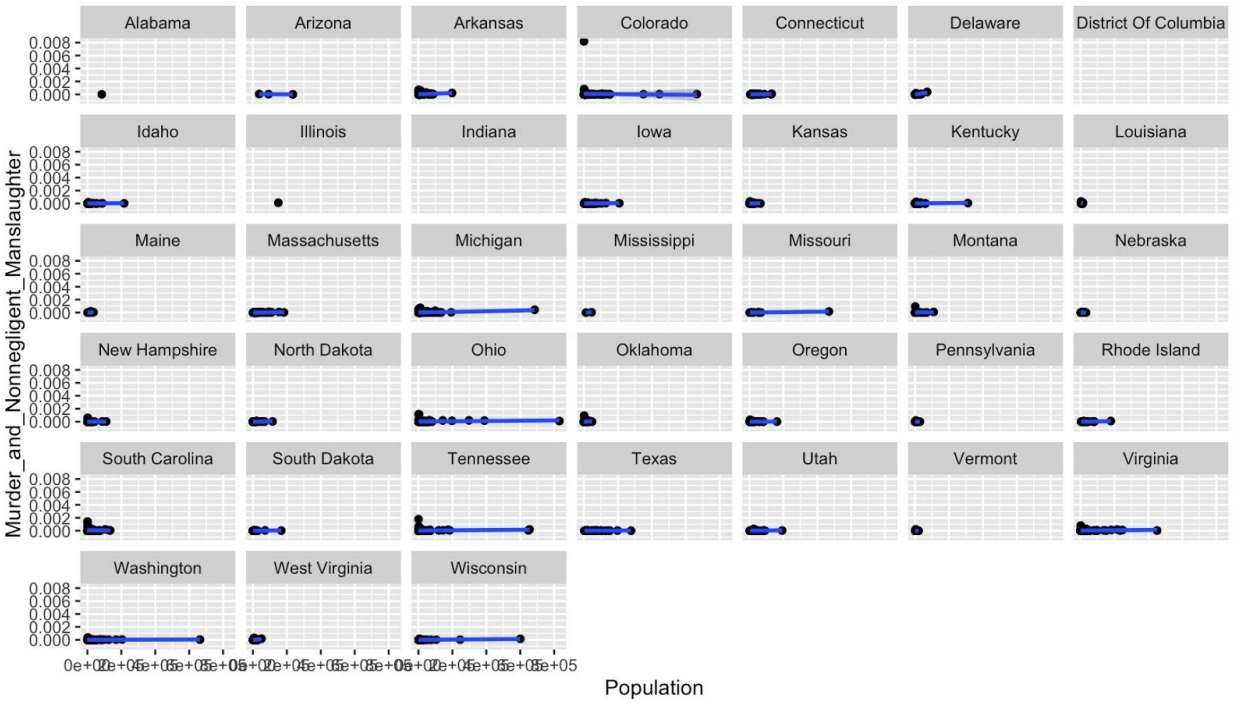
Because of all the zero values around the corner, it goes positive way. If we don't have that many zero values, the line will probably start from a higher spot so that it will probably go negative way like we predicted before. But we can't simply remove all the zero values, because it is still telling us some stories.

We focus on the Homicide_Offenses, which includes Murder_and_Nonnegligent_Manslaughter and Negligent_Manslaughter which means killing people by accident. After showed the summary of those three variable versus population, we found there is significant correlation between murder/homicide offenses and population while there is apparently no correlation between negligent manslaughter and population, which does make sense. Since if it is truly accidental then it shouldn't have an effect no matter where you are, no matter what the

population size is.That's why although both are considered homicide, but only one should be affected by population size.
And we can also find the similar conclusion with James' maps, map of homicide offenses and map of murder are almost the same however map of negligent manslaughter is much different from them.

Here is a graph of the murder rate predicted by population in each state.

```
ggplot(prop14, aes(x=Population,y =
Murder_and_Nonnegligent_Manslaughter)) + geom_point() +
facet_wrap(~State) + stat_smooth(method = lm)
```



Decision Tree

Decision tree was not the best model choice for the crime data set; or at least not the question I was trying to solve. My question was whether or not the population size had any effect on how high the crime rate was. When we first received our data set we noticed it was only for crime that happened during the year 2014. During class one day Grant Scott had told us if we wish to find more data to add to ours than we could. We all decided it might help our plots and be able to expand our questions if we added to it so we added the years 2012 and 2013. I went about trying several ways to answer my question there was lots of trial and error. The first problem I ran into was when I ran this code:

```
frmla = Crimes_Against_Persons ~ .
fit = rpart(frmla, method="class", data=df)
```

```
printcp(fit)
plotcp(fit)
```

My computer literally stopped working because it did not narrow down the data enough to run quickly. This code neither printed nor plotted the information. I learned how to do a decision tree by looking at the examples given to us in canvas and also by going to www.rdatamining.com/examples/decision-tree and looking at their code and data they plotted. Using the code and examples from canvas did not want to work with the data set I was using. That is why I had to look on the web for other solutions. After trying the code above since it would not run I tried to narrow down the information I was trying to have it find by using this code:

```
frmla = Crimes_Against_Persons ~ Population + Crimes_Against_Society
fit = rpart(frmla, method="class", data=df)
printcp(fit)
plotcp(fit)
```

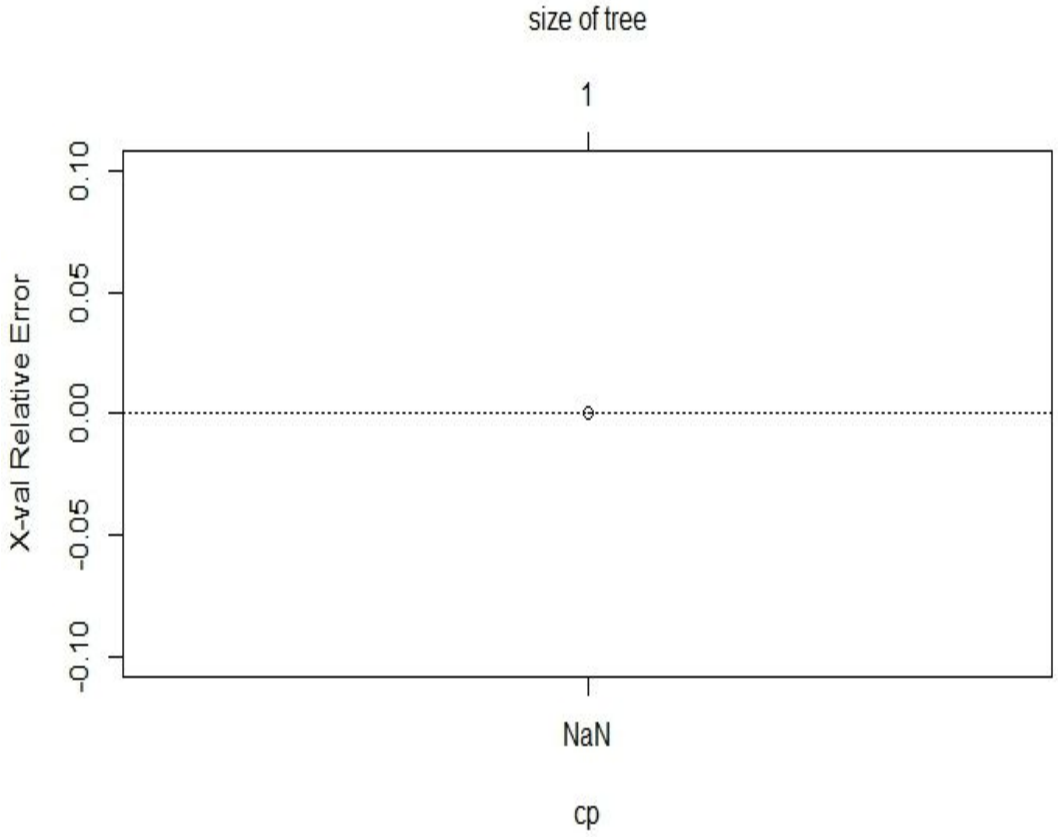
Here is the printed version -

```
Response:  Crimes_Against_Persons
Inputs:  Population, Crimes_Against_Society
Number of observations:  7005
1) Crimes_Against_Society <= 0.04492546; criterion = 1, statistic =
6108.11
2) Crimes_Against_Society <= 0.005668934; criterion = 1, statistic =
1703.113
3) Crimes_Against_Society <= 0.001697186; criterion = 1, statistic =
467.977
4) Crimes_Against_Society <= 0.0007163324; criterion = 1, statistic =
86.342
5) Population <= 4757; criterion = 0.996, statistic = 9.377
6) Crimes_Against_Society <= 0.0003427005; criterion = 0.98,
statistic = 6.585
7)*  weights = 451
6) Crimes_Against_Society > 0.0003427005
8)*  weights = 100
5) Population > 4757
9) Population <= 43762; criterion = 1, statistic = 101.39
10)*  weights = 224
9) Population > 43762
11)*  weights = 8
4) Crimes_Against_Society > 0.0007163324
12)*  weights = 695
3) Crimes_Against_Society > 0.001697186
```

```
13) Crimes_Against_Society <= 0.003546099; criterion = 1, statistic =
48.319
14)*  weights = 1100
13) Crimes_Against_Society > 0.003546099
15) Population <= 62514; criterion = 1, statistic = 17.003
16)*  weights = 954
15) Population > 62514
17)*  weights = 60
2) Crimes_Against_Society > 0.005668934
18) Crimes_Against_Society <= 0.01351706; criterion = 1, statistic =
286.528
19) Population <= 148236; criterion = 1, statistic = 58.853
20) Crimes_Against_Society <= 0.01064538; criterion = 1, statistic =
30.003
21) Population <= 6676; criterion = 0.953, statistic = 5.102
22)*  weights = 821
21) Population > 6676
23)*  weights = 822
20) Crimes_Against_Society > 0.01064538
24)*  weights = 530
19) Population > 148236
25)*  weights = 43
18) Crimes_Against_Society > 0.01351706
26) Population <= 5241; criterion = 1, statistic = 27.511
27) Crimes_Against_Society <= 0.03437575; criterion = 0.982,
statistic = 6.775
28) Population <= 2661; criterion = 0.953, statistic = 5.127
29)*  weights = 274
28) Population > 2661
30)*  weights = 199
27) Crimes_Against_Society > 0.03437575
31)*  weights = 63
26) Population > 5241
32) Crimes_Against_Society <= 0.02019775; criterion = 1, statistic =
24.896
33) Population <= 62021; criterion = 0.999, statistic = 12.099
34)*  weights = 328
33) Population > 62021
35) Population <= 184362; criterion = 0.983, statistic = 6.882
36)*  weights = 42
35) Population > 184362
37)*  weights = 8
32) Crimes_Against_Society > 0.02019775
```

```
38)*  weights = 207
1) Crimes_Against_Society > 0.04492546
39) Crimes_Against_Society <= 0.195; criterion = 1, statistic =
70.216
40) Crimes_Against_Society <= 0.0801105; criterion = 0.972, statistic
= 6.003
41)*  weights = 62
40) Crimes_Against_Society > 0.0801105
42)*  weights = 7
39) Crimes_Against_Society > 0.195
43)*  weights = 7
```

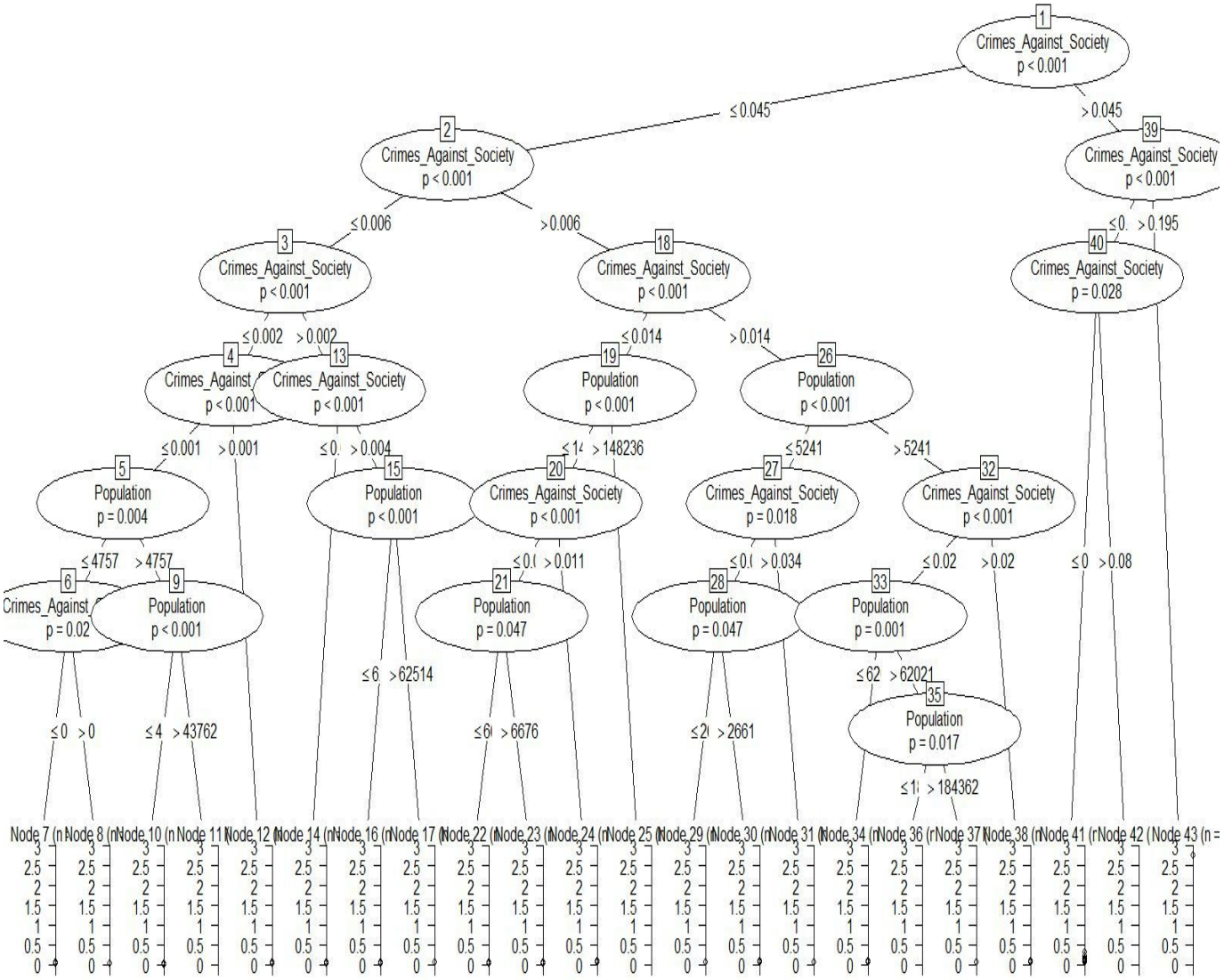
The plot to this code however, did not seem to work. Here is the plot -



I believe this plot was not able to read all the 0's which, came from the states who did not share their information on crime. The next set of code I tried to run finally seemed to work with plotting something. The working code:

```
fit<- ctree(Crimes_Against_Persons ~ Population +
Crimes_Against_Society, data = df)
print(fit)
plot(fit)
```

Which, printed version looks the same as the previous code I tried but the plot did not look so great it was very messy. The printed version was easier to read and understand what the plot is showing us. Here is the plot -



After all the trail and errors of working with decision tree we decided this was not the best model to work with for this data set. It could have also been the question needed to be narrowed down to make the plot look better. I was not able to even get the whole plot in one picture.

Support Vector Machines

Stepping back and taking a look at the data given can be quite a daunting first look as there is much to absorb. Finding questions to answer from the data can be even harder. The first thing that came to mind was trying to answer if guessing a population based on x and y crime rates was even possible. However, if you logically think about this - what good would it

do? You're literally guessing the question: "What is crime rates based off of population?" But this isn't meaningful at all. In fact the inverse question is the one to be answered: "What is the population based off the crimes rates, more specifically murder rates?" Now this is a question. Taking the combined data of from 2012, 2013 and 2014 as

Population	Total_Offenses	Crimes_Against_Persons	Crimes_Against_Property	Crimes_Against_Society

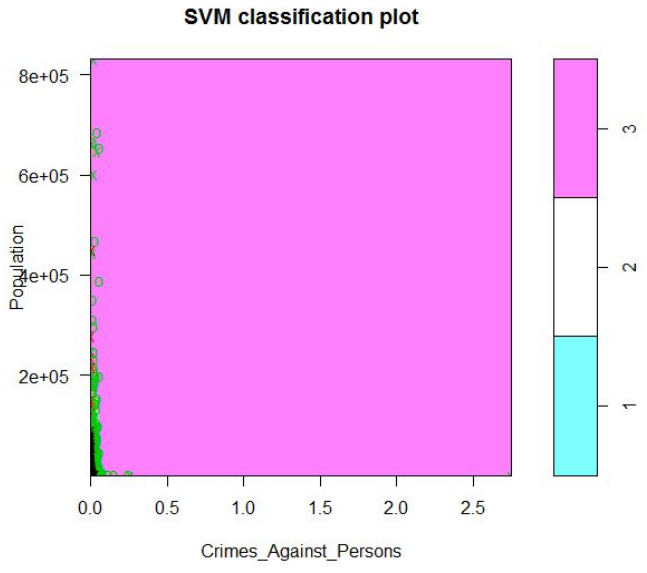
And running:

```
Call:
svm(formula = Population ~ ., data = prost)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.25
    epsilon:  0.1


Number of Support Vectors:  91
```

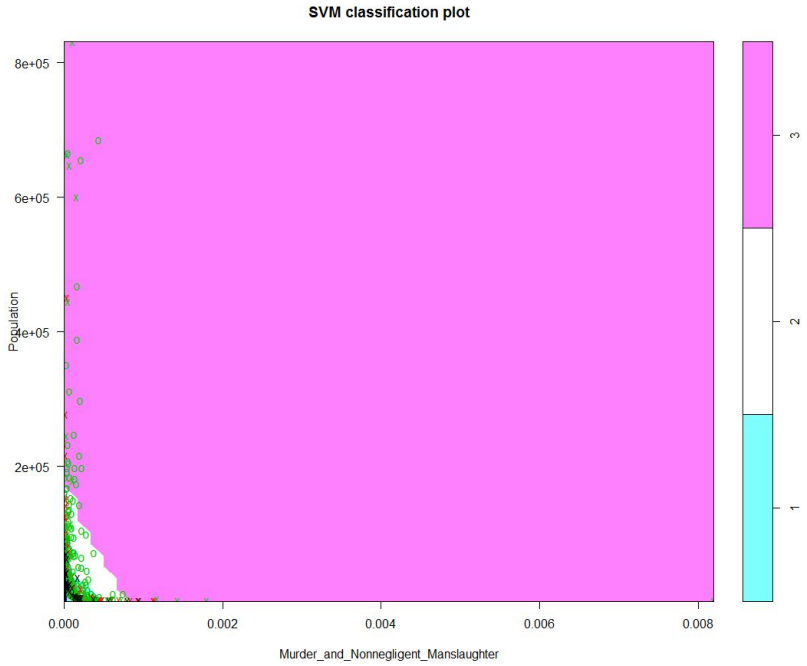
Well that's completely unhelpful. With a total of 108 objects and 91 being support vectors tells us short of nothing except that it's a bad model to run. Instead of attempting to narrow down the data so much that a model cannot tell us anything, let us take a step back. Taking all three data years and combining them expanding the columns from the previous example and running a model on it we get:

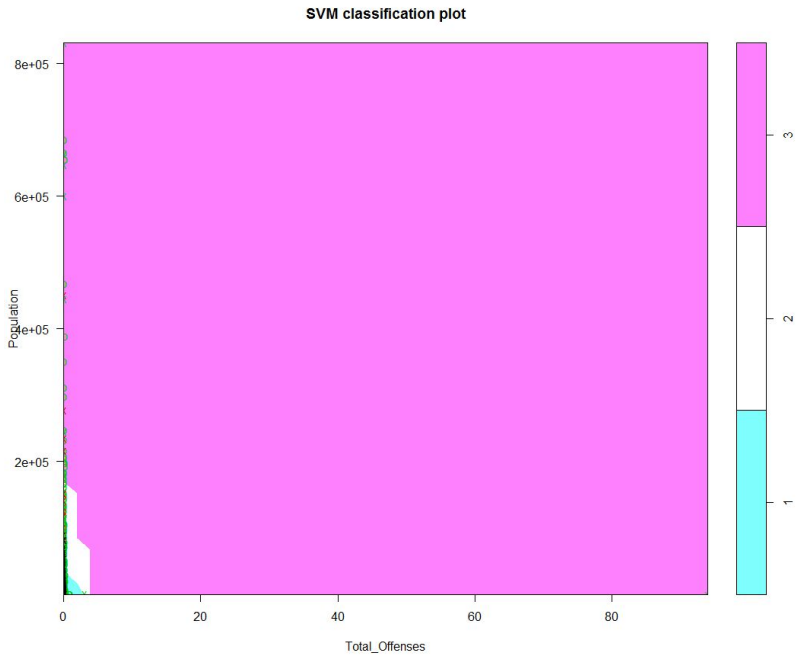
```
subset(p12, select=c('Agency_Type','codes','Region','Population','Crimes_Against_Persons',
                     'Crimes_Against_Property','Crimes_Against_Society','Total_Offenses',
                     'Homicide_Offenses', 'Murder_and_Nonnegligent_Manslaughter',
                     'Negligent_Manslaughter','Justifiable_Homicide'))
```

```
> newSVM <- svm(discr~., data=df)
> summary(newSVM)

Call:
svm(formula = discr ~ ., data = df)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.01612903

Number of Support Vectors:  3020

 ( 1159 877 984 )


Number of Classes:  3

Levels:
 1 2 3
```



Still, the model concludes 3020 support vectors out of 3485 objects which is slightly worse model than the previous one. If you continue plotting various independent variables against one another, the trend is generically the same.





And so on it goes… In conclusion, SVM is unable to specify any information that has value in regards to the questions asked. When analyzing our data set and examining the decision tree, it would make logical sense that linear regression models would represent. and better yet,

model the data given. Perhaps the outcomes of SVM would have been different had each state actually reported to the FBI. Since California and New York have some of the densest populations, their representation in this report could have have a major outcomes on the questions we seeked to answer.

Conclusion

There were several key findings that we found throughout our various models:

1. For 2014, homicide rates tended to increase as population increased. The slope of the line appears to be heavily influenced by the high density of agencies reporting no homicide.
2. For those places that did report murder rates, the trend appears to be exactly the opposite of the first key finding, murder rates tend to increase as population increases. There are several potential explanations for this finding but most likely it has to do with the interplay between multiple factors. Among the explanations could be that people are less likely to commit murder when there are more people around to report such an incident. Perhaps population density is another variable that could be considered if this analysis were to be carried out even further in order to test this hypothesis.
3. State appears to have an effect on the homicide rate. This effect is particularly strong in murder rates. This could suggest that statewide phenomena (politics, economics, etc.) may be at play.
4. Population had a significant effect on murder rate but not negligent homicide. This is probably one of the most intuitive, albeit fascinating findings of the entire project. If manslaughter were truly negligent then you would expect for the rate per capita to be distributed similarly across all populations.

There were also several things to keep in mind when running our analysis:

1. Not every state had agencies that reported crime rates. These states are indicated in white on the maps above. Saddly, this excludes several of the largest cities in the nation including the top three: New York, Los Angeles and Chicago.
2. Some states reported a lot of agencies while other reported only one or two. This makes predicting on state rather messy as one value could become incorrectly indicative of the entire state's crime rate.