Lab 3: Reducing Crime

Namburi, Pouchepanadin, Reyes

April 2, 2018

Introduction

For this lab, a data science consultancy has been hired by the challenging candidate running for North Carolina's governorship. Our group is acting as data scientists assigned to this particular project. Our primary task is to analyze the provided data to identify determinants of crime to help develop crime control policies. Thus, the motivating question behind our research is: "How does crime punishment and demographics relate to the amount of crimes committed?"

Exploratory Data Analysis

```
setwd('C:/Users/apouchepanad/Desktop/Berkeley/Courses/Statistics/Unit 11')
crime <- read.csv("crime_v2.csv")
```

General Analysis

The file is appended with 6 rows that are mostly empty. One row has a ' character, which does not make much sense within its respective column (prbconv). We believe it is safe to remove these rows from the dataset. We declare a second dataset "crime_fix" in which we insert our subsetted data. Moving forward, all transformations and manipulations will be performed on our new dataset.

```
crime_fix <- crime[1:91,]
```

The extra character within the removed rows caused column "prbconv" to load as a factor instead of a numeric.

```
crime_fix$prbconv <- as.numeric(as.character(crime_fix$prbconv))
```

There also appears to be a duplicate row within our dataset. Applying 'unique' shows that there exists two rows that contain the same county identifier.

```
length(crime_fix$county)
```

```
## [1] 91
```

```
length(unique(crime_fix$county))
```

```
## [1] 90
```

We can list out the county IDs in order to make it easier to identify the duplicate ID. Browsing through the data, we find that county 193 is duplicated. Displaying both rows of data shows that they are exactly alike across all columns.

```
sort(crime_fix$county)
```

```
##  [1]   1   3   5   7   9  11  13  15  17  19  21  23  25  27  33  35  37
## [18]  39  41  45  47  49  51  53  55  57  59  61  63  65  67  69  71  77
## [35]  79  81  83  85  87  89  91  93  97  99 101 105 107 109 111 113 115
```

```
## [52] 117 119 123 125 127 129 131 133 135 137 139 141 143 145 147 149 151
## [69] 153 155 157 159 161 163 165 167 169 171 173 175 179 181 183 185 187
## [86] 189 191 193 193 195 197
```

```
crime_fix[crime_fix$county == 193,]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 88    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 88 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 88 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 88 0.07819394
## 89 0.07819394
```

One of the duplicate rows should be removed.

```
crime_fix <- unique(crime_fix)
```

prbarr - Probability of Arrest

The prbarr column contains a probability value > 1

```
summary(crime_fix$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20495 0.27146 0.29524 0.34487 1.09091
```

```
crime_fix$prbarr[crime_fix$prbarr > 1]
```

```
## [1] 1.09091
```

Traditionally in statistics, probability ranges from 0 to 1. It could be argued that a person could be arrested twice for the same offense. If both arrests were counted, it could lead to a ratio greater than 1. However, even if this were the case, it would still be a stretch to observe a probability greater than 1. Either all offenses have lead to an arrest or there exists offenses that lead to many arrests. Either reasoning seems unlikely. Lastly, looking at the rest of the data, this row lies far off from all other values. We believe this value is erroneous and that this row should be omitted.

```
crime_fix <- crime_fix[crime_fix$prbarr <= 1,]
```

prbconv - Probability of Conviction

The prbarr column contains probability values > 1

```
summary(crime_fix$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34302 0.45057 0.54020 0.57394 2.12121
```

```
crime_fix$prbconv[crime_fix$prbconv > 1]
```

```
## [1] 1.48148 1.22561 1.23438 1.35814 1.06897 1.01538 2.12121 1.67052 1.18293
```

Again, traditionally in statistics, probability ranges from 0 to 1. However, unlike the case with prbarr, these values are not extreme outliers when compared against the other rows. This makes it slightly more difficult to dismiss these values. One strong argument we can make is that the 5th amendment clearly states that a person cannot be tried twice for the same crime. We feel that is is sufficient to assert that a single crime cannot lead to multiple convictions and that these values are erroneous.

```
crime_fix <- crime_fix[crime_fix$prbconv <= 1,]
```

west/central - Geographic Indicators

There is a row that has both the west and central indicator set to 1.

```
crime_fix[crime_fix$west == 1 & crime_fix$central == 1,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 33     71   87 0.0544061 0.243119 0.22959 0.379175  11.29 0.00207028
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 33 4.834734 31.53658    1       1     0   13.315 291.4508 595.3719
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 33 240.3673 348.0254 295.2301 358.95 509.43 359.11 339.58 0.1018608
##       pctymle
## 33 0.07939028
```

Since there is no way of knowing which indicator is accurate and which one is not, we just have to remove the row entirely.

```
crime_fix <- crime_fix[crime_fix$county != 71,]
```

There are several rows that do not have the west or central indicator flagged. From this observation, we can most likely infer that these counties are located in eastern North Carolina, as that is the ordinal geographic region that best fits the data.

```
crime_fix[crime_fix$west == 0 & crime_fix$central == 0,]$county
```

```
##  [1]  13  15  17  41  47  49  51  53  55  61  65  79  83  85  91  93 101
## [18] 107 117 129 131 133 139 141 143 147 155 163 165 187 191
```

To make this indication more clear, we can add a third geographic flag for counties in east North Carolina

```
crime_fix$east <- ifelse(crime_fix$west == 0 & crime_fix$central == 0, 1, 0)
```

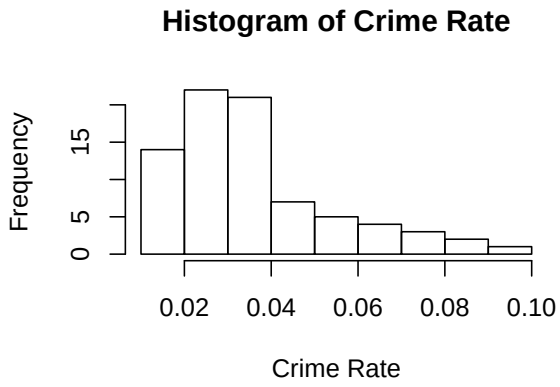
Univariate Analysis

Analyzing Crime Rate - Crimes committed per person

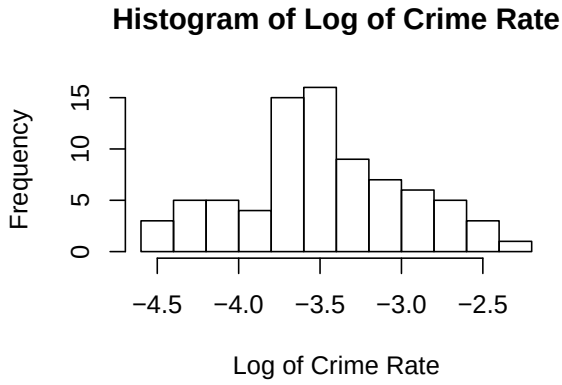
```
summary(crime_fix$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02335 0.03043 0.03527 0.04234 0.09897
```

```
hist(crime_fix$crmrte, main="Histogram of Crime Rate", xlab = 'Crime Rate')
```




```
hist(log(crime_fix$crmrte), main="Histogram of Log of Crime Rate", xlab = 'Log of Crime Rate')
```



Crime rate in counties ranges from 1% to 9.8% with a median of 3% and a median of 3.5%. 75% of the counties have crime rates less than equal to 4.23%, while 25% of the counties have crimes rates higher than 4.37% and less than 9.8%. We also observe that Crime Rate is right skewed, thus deviating from the normality assumption required for linear model specification.

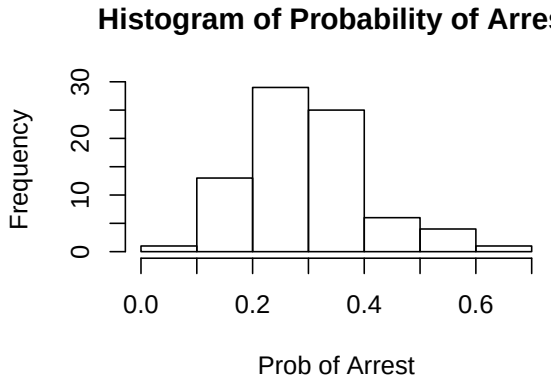
Log transformations are an effective way of battling skewness while not violating our assumptions. In this case, we will not transform the Crime Rate variable, given there are 79 observations (after EDA/data transformations) and it meets the 'n=30' requirement of the Central Limit Theorem.

Analyzing Probability of Arrest

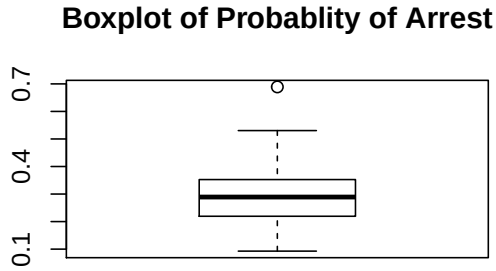
```
summary(crime_fix$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.21938 0.28912 0.29779 0.35254 0.68902
```

```r
hist(crime_fix$prbarr, main = "Histogram of Probability of Arrest", xlab = 'Prob of Arrest')
```




```r
boxplot(crime_fix$prbarr, main = "Boxplot of Probablity of Arrest")
```



The probability of arrest has a mean and median of 29%. The boxplot highlights an outlier at 7% probability. There is no significant skew in the variable that would affect our analysis.

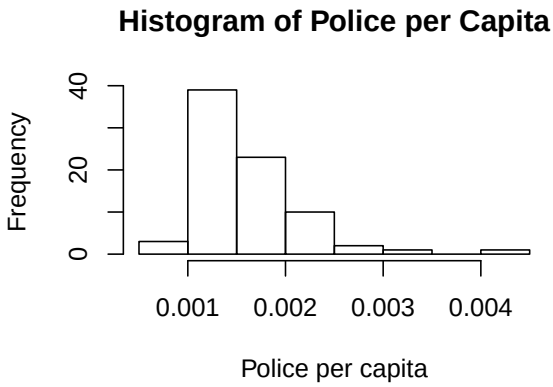
Analyzing Police per Capita

```r
summary(crime_fix$polpc)
```

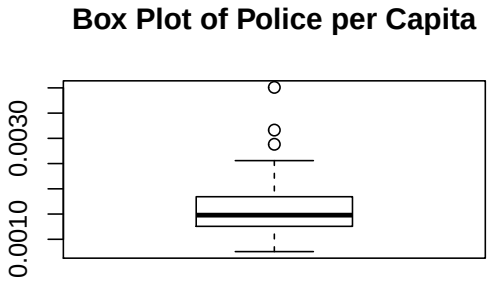
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007559 0.0012564 0.0014782 0.0016099 0.0018435 0.0040096
```

The number of Police officers n a given county range from as low as 7 officers per 10,000 people to as high as 40 officers in 10,000.As highlighted by the boxplot, the maximum observation of 40 per 10,000 is an outlier. There is a significant right skew in the data, violating the normality condition required for linear model specification. The log transformation of 'police per capita' is closer to normal distribution. Given the number of observations we will not transform the variable in our analysis.

```r
hist(crime_fix$polpc, main = "Histogram of Police per Capita", xlab = 'Police per capita')
```




```
boxplot(crime_fix$polpc, main = "Box Plot of Police per Capita")
```

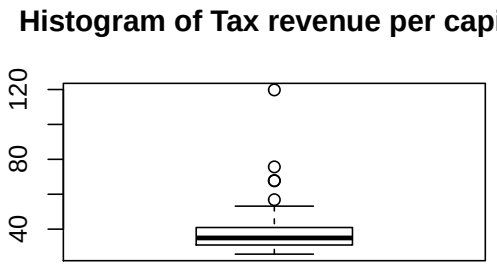


Analyzing taxpc - Tax Revenue per capita

```
summary(crime_fix$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.92   34.96   38.24   40.94  119.76
```

```
boxplot(crime_fix$taxpc, main ="Histogram of Tax revenue per capita", xlab = 'Tax revenue per capita')
```



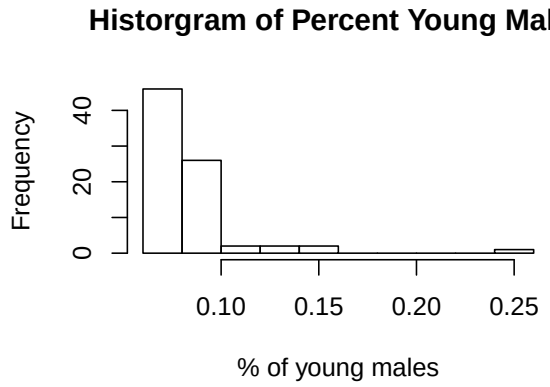
The median of Tax revenue per capita is at ~34.9 while the mean is higher at 38.2 due to outliers. Taxpc has a very high value outlier at 119.76. The Interquartile range is $10 from 30.9 to 40.9. We will assess the impact of outliers while specifying the linear models.
Although the value of this outlier is quite larger than the rest, it is not outside the realm of possibility. There could be local county policy that places higher taxes on its resident. Alternatively, there could have been a major public project underway that required higher taxes for that year.

Analyzing pctymle - Percentage of Young Males

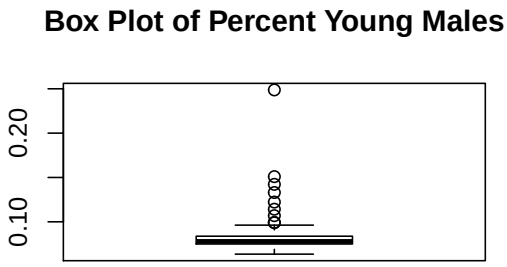
```
summary(crime_fix$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06356 0.07497 0.07787 0.08469 0.08377 0.24871
```

```
hist(crime_fix$pctymle, main ="Historgram of Percent Young Males", xlab = '% of young males')
```




```
boxplot(crime_fix$pctymle, main ="Box Plot of Percent Young Males")
```



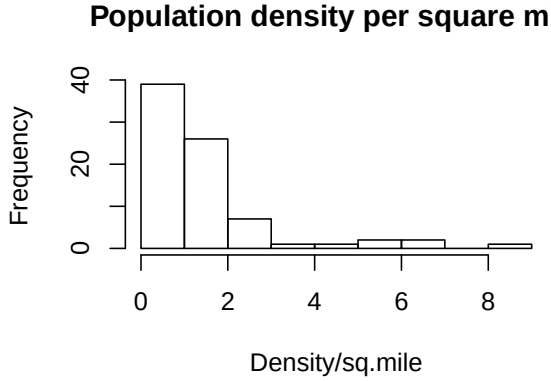
The median is at 7.8% with a higher mean at 8.4%. The boxplot highlights an outlier at ~25%. Though this seems improbable, it can be explained by a community like a mining town with a lot of younger working males. As this could pose some issues with our analysis, we will analyze the impact of the outlier while specifying the linear model.

Analyzing density - Population density

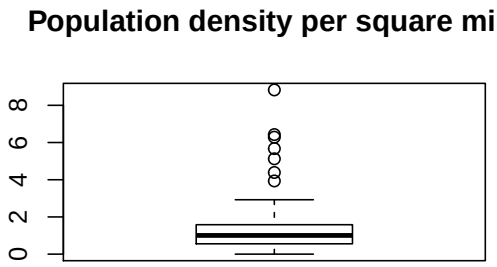
```
summary(crime_fix$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.55546 1.00528 1.47506 1.58212 8.82765
```

```
hist(crime_fix$density, main ="Population density per square mile", xlab = 'Density/sq.mile')
```




```
boxplot(crime_fix$density, main ="Population density per square mile")
```



The minimum in this case is very small at 0.00002, and therefore an outlier. The data is heavily right skewed, but the log transformation does not help with the normality requirement.

Bivariate Analysis

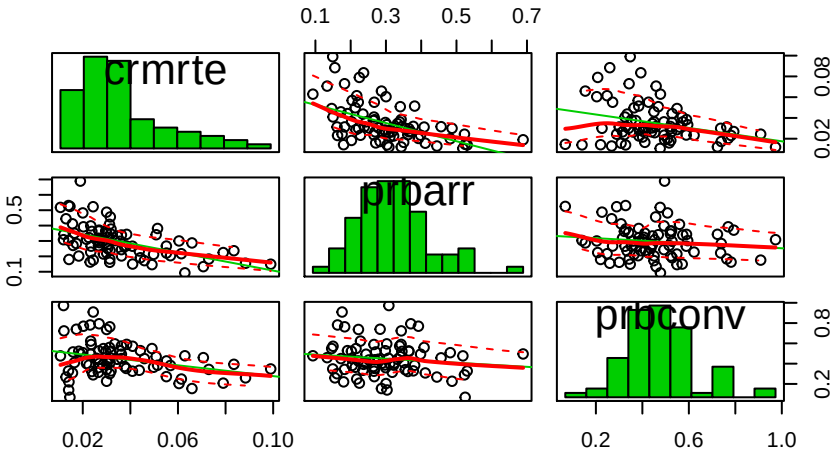
Establishing our Hypotheses

As stated earlier, our research question is: "How does crime punishment and demographics relate to the amount of crimes committed?"

Hypothesis 1

Our group hypothesizes that a county can expect a lower Crime Rate , given a higher probability of arrest. Similarly, we also believe that the certainty of criminal punishment (probability of conviction) would also deter crime in a given county. To begin the analysis, we can generate a scatterplot matrix of the variables in question for a high-level view.

```
library(car)
scatterplotMatrix(~ crmrte + prbarr + prbconv, data=crime_fix, diagonal="histogram")
```



We observe a strong negative correlationship between Crime Rate and the Probability of arrest. We also see that there is a negative relationship between Crime Rate and the certainty of conviction.

```
## [1] "Correlationship between Crime Rate and Probability of Arrest"
```

```
## [1] -0.5076381
```

```
## [1] "Correlationship between Crime Rate and Probability of Conviction"
```

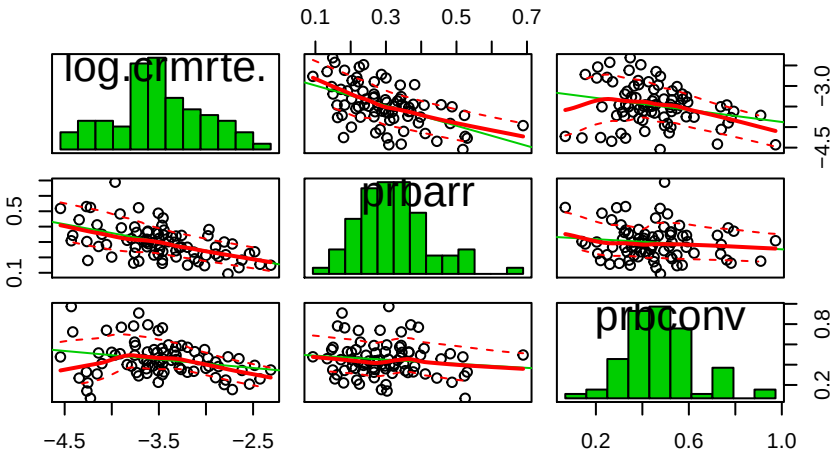
```
## [1] -0.2927517
```

```
## [1] "Correlationship between Probability of Arrest and Probability of Conviction"
```

```
## [1] -0.1282592
```

Given Crime Rate is right skewed, we would like to assess the relationships with the log of Crime Rate and Probability of arrest and conviction.

```
library(car)
scatterplotMatrix(~ log(crmrte) + prbarr + prbconv, data=crime_fix, diagonal="histogram")
```

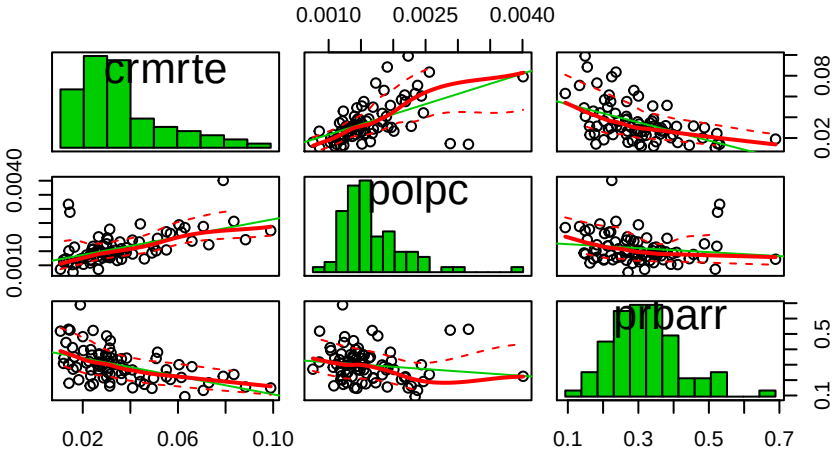


The log of crime Rate does not have a significant skew and we now observe that the negative relationships with the probability of arrest and conviction are more pronounced.

Hypothesis 2

We hypothesize that higher police per capita will increase the probability of arrest thus deterring Crime Rate in a given country. Given the right skew in Crime Rate, we are analyzing the relationships with the log of Crime Rate.

```
scatterplotMatrix(~ crmrte + polpc + prbarr, data=crime_fix, diagonal="histogram")
```

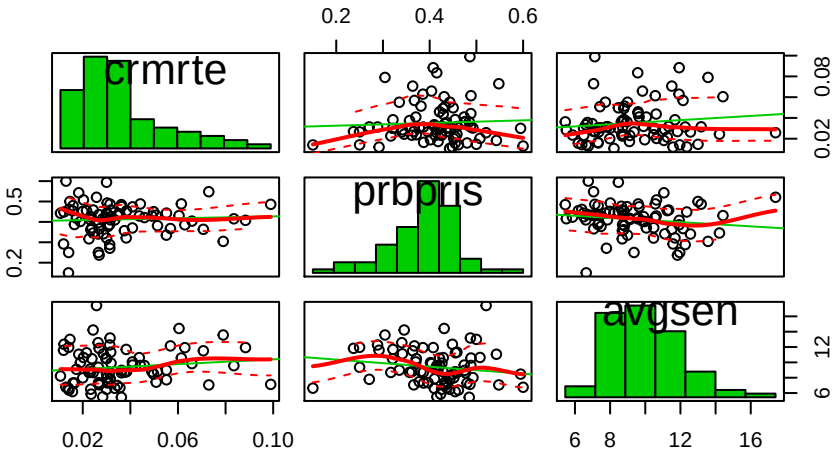


As against our expectations, we observe that there is a strong positive relationship between 'crime rate and police per capita'. This indicates that higher police per capita indicate higher crime rates, possibly implying that more number of police officers are required in counties with higher crime rates. We will further validate our assumptions as we specify our linear models. There does not seem to be a definite relationship between police per capita and probability of arrest.

Hypothesis 3

We hypothesize that "higher the probability of imprisonment and higher the average sentence in a given county would imply lower crime rate". The severity of punishment would deter criminals.

```
scatterplotMatrix(~ crmrte + prbpris + avgsen, data=crime_fix, diagonal="histogram")
```

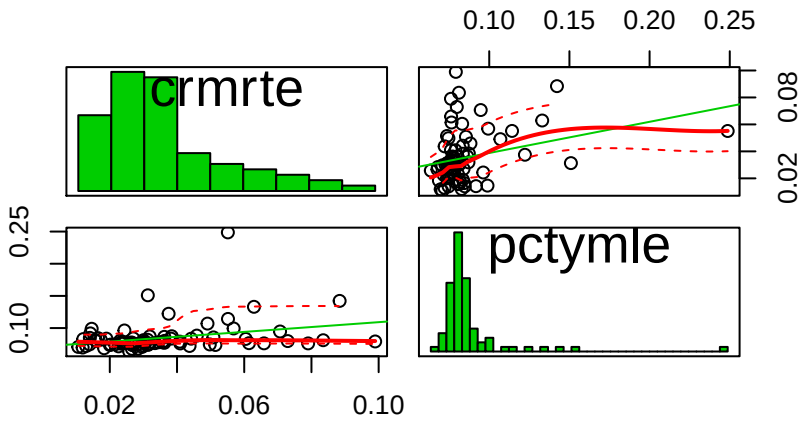


On observing the scatterplot above, we observe that there does not seem to be a significant relationship between the Crime Rate and the Average Sentence as well as the probability of imprisonment.

Hypothesis 4

We hypothesize that "higher the percentage of young males in a county, the higher the crime rate". We will assess the relationship using the scatterplot below.

```
scatterplotMatrix(~ crmrte + pctymle, data=crime_fix, diagonal="histogram")
```



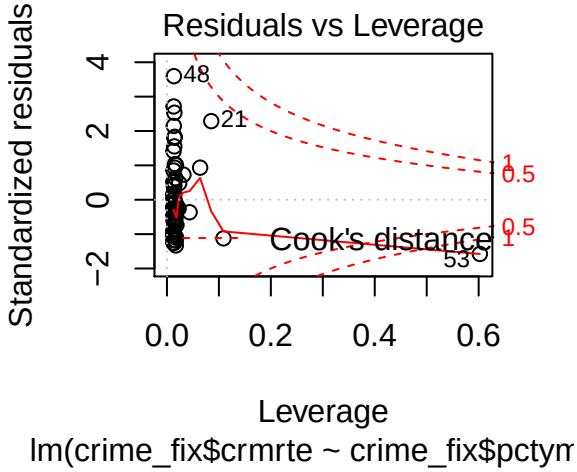
We don't observe a strong positive relationship in the graph above. We would also like to assess the influence of the outlier, if any.

```
m2 <- lm(crime_fix$crmrte ~ crime_fix$pctymle)
m2$coefficients
```

```
##       (Intercept) crime_fix$pctymle
```

```
##         0.01568209          0.23131498
```

```
plot(m2, which=5)
```

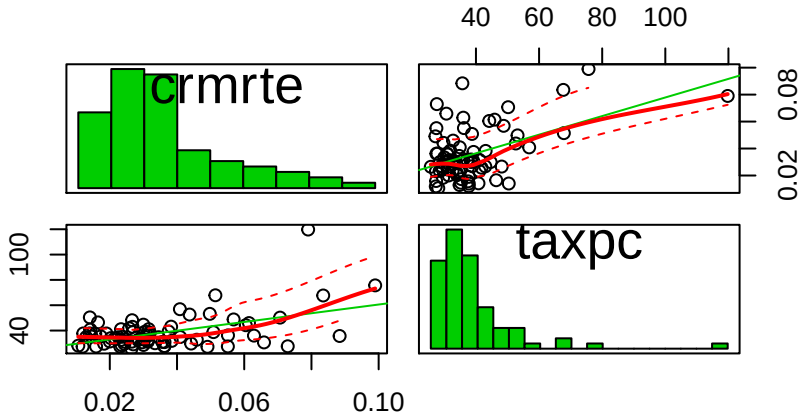


There is an observation with the Cook's distance greater than 1, but this is not the observation we expected (i.e. pctymle = 25%). This observation has a very crime rate of ~9.8%, even though the percentage of young males is ~8%.

Hypothesis 5

We hypothesize that "higher the tax revenue in a county, the lower the crime rate". We will assess the relationship using the scatterplot below.

```
scatterplotMatrix(~ crmrte + taxpc, data=crime_fix, diagonal="histogram")
```



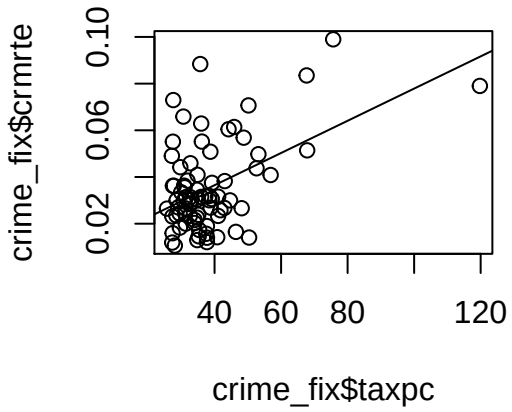
The graph above does not strengthen our hypothesis, indicating that higher the tax revenue in a given county, the higher the crime rate. We will now assess the linear model of Crime Rate and taxpc and also assess the

effect of the outlier value (119.76).

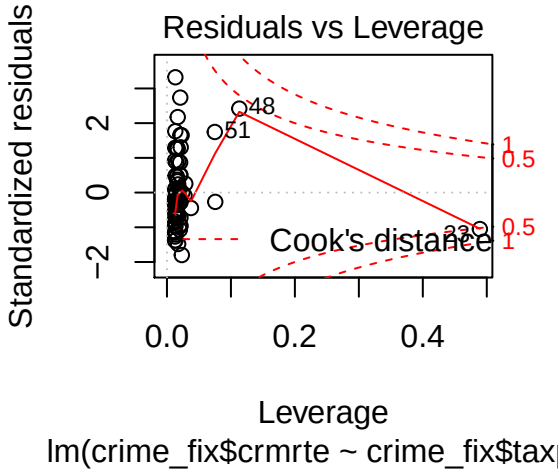
```
m1 <- lm(crime_fix$crmrte ~ crime_fix$taxpc)
m1$coefficients
```

```
##     (Intercept) crime_fix$taxpc
##    0.0088912534    0.0006898154
```

```
plot(crime_fix$taxpc, crime_fix$crmrte)
abline(m1)
```




```
plot(m1, which=5)
```

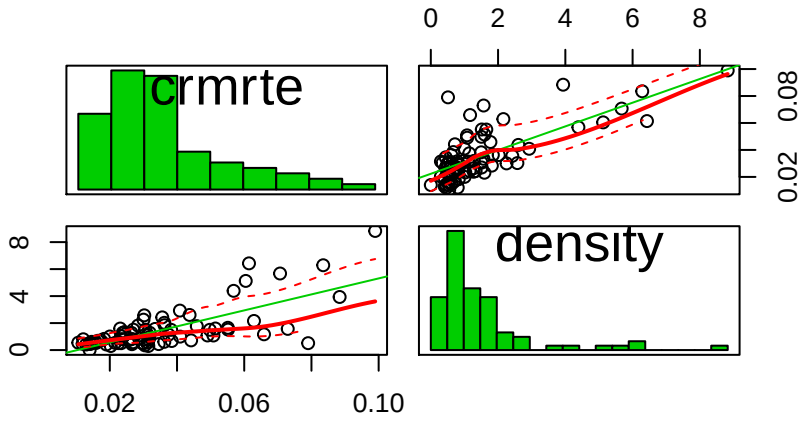


This would imply that $1 increase in taxpc, would increase crime rate by ~0.0007. We also observe that all the data points have a Cook's distance of less than 0.5 and greater than -0.5. Therefore we can conclude, the outlier does not have the influence to move the regression line significantly.

Hypothesis 6

We hypothesize that "higher the population density, higher the crime rate". We will assess the relationship using the scatterplot below.

```
scatterplotMatrix(~ crmrte + density, data=crime_fix, diagonal="histogram")
```

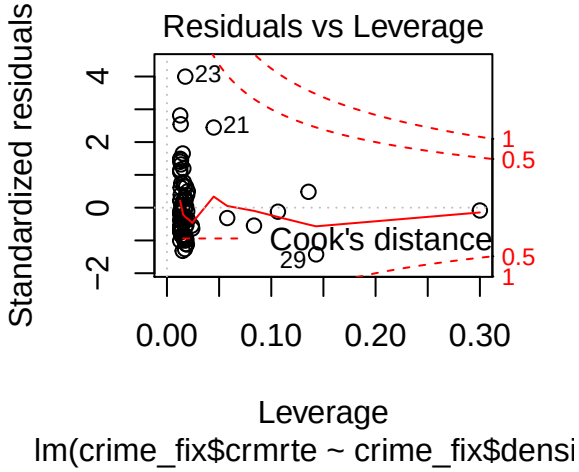


There is a strong positive relationship observed in the scatterplot above. Assessing influence of outliers as we observed a very small value in the univariate analysis.

```
m3 <- lm(crime_fix$crmrte ~ crime_fix$density)
m3$coefficients
```

```
##       (Intercept) crime_fix$density
##        0.02230438        0.00879209
```

```
plot(m3, which=5)
```



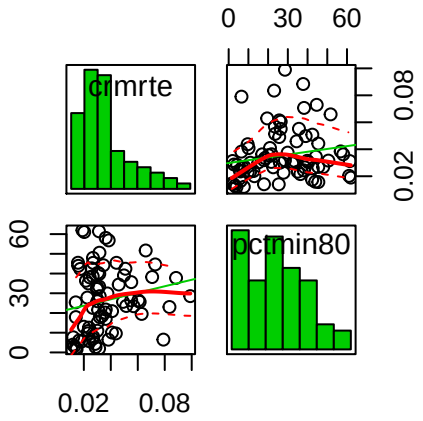
We observe that there are no observations with Cook's value greater than 1, thus implying no observation

has the influence to shift the regression line significantly.

Hypothesis 7

We hypothesize that "higher the minority population in a given county, higher the crime rate". We will assess the basic relationship using the scatterplot below.

```
scatterplotMatrix(~ crmrte + pctmin80, data=crime_fix, diagonal="histogram")
```



Based on the plot above, we don't see an obvious relationship here.

Analyzing correlationships with Crime rate

Let's take a look at the correlations of Crime Rate with the other variables.

```
knitr::kable(cor(subset(crime_fix
, select = -c(county,west,central,year,urban)))[1,])
```

	x
crmrte	1.0000000
prbarr	-0.5076381
prbconv	-0.2927517
prbpris	0.0537413
avgsen	0.1256901
polpc	0.5558263
density	0.7217443
taxpc	0.4874887
pctmin80	0.1854219
wcon	0.3637529
wtuc	0.2534870
wtrd	0.3994104
wfir	0.3455101
wser	0.3303221
wmfg	0.3892113

	x
wfed	0.4712225
wsta	0.1904091
wloc	0.4256231
mix	-0.2320608
pctymle	0.2956789
east	0.1654490

We are using the above table to analyze the correlations with the wages in different industries as well as average federal/local wages. We hypothesize that "Higher the average wage in a given county, lower the Crime Rate". Higher wages would imply more skilled workers, better education system and thus lower Crime Rate.

The above table does not validate our hypothesis as there is a slightly positive relationship with all measures of average wage. The positive relationship with average weekly wage of federal workers (wfed) is the most pronounced.

Model Specifications:

Based on our inital EDA, univariate and bivariate relationships, we now have a better understanding of the variables in our cleaned up dataset. Looking back at our research question, we need to come up with recommendations on what policies would affect crime rate in the counties of interest.

To help inform our recommendations, we have specified three multiple regression models that could affect crime rate in the counties. While we are trying to establish a relationship between crime rate and our other variables, we are aware that there is likelihood of several key variables that are not included in our dataset to also be causal factors to crime rate. We will address these issues in the following section about omitted variable bias. For now, while building our models, we will work off the assumption that our dataset has enough variables to establish causality with crime rate.

We are proposing three model specifications: one that has covariates that we feel has the most explanatory power, one that has additional covariates that might help address the unexplained variation from the first model and the final model that has all of the covariates from the dataset.

The variables in our dataset are broken out into two types that we are calling crime variables and demographic variables. The crime variables include crime rate (dependent) and the probabilities of arrest, conviction and imprisonment, police per capita and average sentence length. The demographic variables account for the remaining covariates.

For the first model, we feel that the crime variables have the most explanatory power for our model. In addition, we will also add a demograhic variable that is the most important in terms of explanatory power. While creating the first model, it is important to recall the outcome variables should always be on the left. When we look at prbarr, prbconv and prbpris, we notice a staggered effect where prbconv is linked to prbarr and prbconv is linked to prbpris. For this reason, we plan on just using prbarr for the first model and disregarding prbconv and prbpris. Additonally, avgsen is likely also an outcome of crime rate. We anticipate that crime rate drives avg sentence time in that counties with higher crime rates migh be stricter in the sentence times. For this reason, we do not include avgsen in the first model. Police per capita we feel plays a strong role in influencing crime rate. Counties with strong police presence might deter crime and lead to a lower crime rate. Hence we use prbarr and polpc as the two key explanatory variables for the first model. Additionally, we also add one demographic variable to the first model. Looking at all the demographic variables we have available to us, we feel that density does a good job of covering demographic information at a high level. The number of people per square mile would be a good control to account for geographic location, wages, age and culutural mix. For this reason, we feel density would be a good control (not necessarily an explanatory variable) to add to the first model.

For the second model, in addition to the covariates from the first model, we add two more covariates; pctmin80 and pctymle. The reasoning here is to add more control variables to our first model. While the first model focuses heavily on the crime variables for the explanatory power, having density as the sole demographic variable might overstate the effect that density has on crime rate. For this reason, we add two more demographic variables that we feel control for demographic variation that aren't covered by density. We hypothesise that counties that have higher proportions of minorities might be more inclined to have higher crime rates. Similarly, we also hypothesise that counties with higher proportions of young males might also have higher crime rates.

For the third model, we add all the covarites we have able in our dataset. The reason we do this is to test the robustness of our inital two models. How much more of the variation is explained by having a model that has 20+ covariates versus a model that has 3 or 5 covariates? How do the AIC scores look for these three models.

Our first model has an adjusted R squared of 66%, the second 78% and the third 85%. The AIC scores are -482, -514 and -532 respectively. In the interest of parsimony and explained variance, we feel our second model reflects the optimal specification. Adding 2 covariates to the first model increases our R2 squared by 12% while adding 15+ more covariates to the second model only increase our R2 squared by 7%. While the AIC score for the second model is higher than the first model, we feel the trade off with increased R squared is worth it.

```
model1 <- lm(crmrte ~ prbarr + density + polpc, data = crime_fix)

model2 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + pctymle, data = crime_fix)

model3 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + pctymle + county +
             prbconv + prbpris + avgsen + taxpc + west + central + urban + wcon + wtuc +
             wtrd + wfir + wser + wmfg+ wfed + wsta + wloc + mix, data = crime_fix)
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(model1, model2, model3, type = 'latex',
          report = 'vc',
          title = 'Linear Models Predicting Crime rate',
          keep.stat = c('n','rsq','adj.rsq','aic'),
          omit.table.layout = 'n', intercept.top = T, intercept.bottom = F,
          add.lines = list(c('AIC',round(AIC(model1),2),round(AIC(model2),2),round(AIC(model3),2))),
          header = F)
```

Omitted Variables Discussion

While our model specifications do a good job of explaining variation in crime rate, we feel that the estimates for the coefficients of our independent variables might be biased by some variables that were omitted / not present in our dataset. Below are some variables we feel would affect crime rate. While we are aware that it is hard to determine direction of bias in the presence of multiple independent variables, we will explain the bias of our omitted variables in the context of having one independent variable.

We feel that employment rates is a good variable in trying to explain crime rate. We would expect that counties that have higher emplyment rates might have lower crime. Having a job and steady income stream might negate the need to engage in criminal activities. In the context of our second model, we feel that not having

Table 2: Linear Models Predicting Crime rate

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
Constant	0.022	0.005	0.012
prbarr	−0.046	−0.054	−0.052
density	0.006	0.006	0.005
polpc	11.492	12.473	10.791
pctmin80		0.0004	0.0003
pctymle		0.091	0.145
county			0.00000
prbconv			−0.007
prbpris			0.011
avgsen			−0.001
taxpc			0.0002
west			−0.005
central			−0.006
urban			0.004
wcon			0.00003
wtuc			0.00001
wtrd			0.0001
wfir			−0.00005
wser			−0.0001
wmfg			−0.00000
wfed			0.00004
wsta			−0.0001
wloc			0.00004
mix			−0.023
AIC	-481.62	-514.49	-532
Observations	79 19	79	79
R^2	0.671	0.794	0.895
Adjusted R^2	0.658	0.780	0.852

employment rates might be biasing the impact of pctymle. The coefficient for pctymle might be overstated since it might be trying to explain the lack of employment in younger males in the county. Employment rate could also be overstating the coefficient on police per capita if there are budgetary restrictions on the number of police officers by county.

Similar to employment rate, we think education level would also be an interesting variable to have in the dataset. We would expect counties with higher education levels to have lower crime. While one could agrue that education and employment are correlated, the counter argument would be that having a good education doesn't translate to having a job especially in times of economic crises. Similar to employment rate, we feel that education level could be overstating the effect that pctymle and polpc have in our second model.

Abortion stance is another interesting omitted variable that could explain crime rate. Based on what we know about Roe vs. Wade, the effect of abortion on crime in the future could be an interesting effect that is not being captured in our model. Counties that have had pro life regulation in the past could be experiencing higher crime rates currently from unwanted children. This variable could also be overstating the impact the the pctymle variable has on crime rate.

Re-employment rates for ex-convicts might also be an omitted variable in the model. When ex-convicts find it hard to find a job, this might make it easier for them to revert back to crime to support themselves. It is hard to determine which of our variables in the second model might be biased by this variable and as such the direction of bias might be harder to identify.

The presence of capital punishment might also be interesting to have in our data set. Counties that have death sentences might see lower rates of crime and this variable might be biasing the polpc variable. If death sentences are a stronger deterrant to crime than the presence of police, our estimate for polpc might be overstated and account for the effect of death sentences as well. A similar bias argument could be made for the prbarr variable as well. Since the probability of arrest would affect the sentence type, some of the death sentence effect could also reside in the estimate for prbarr thereby biasing this estimate upwards.

The presence of homeless people might also be a good indicator of crime rate. We think that counties that have higher levels of homelessness might also have higher levels of crime. This could again be linked to the employment rate and education level discussions we went over in this section. This ommitted variable might not have that big of an effect if it were explained by employment rate and education level. In the absence of these two variables, however, the homeless people per capita variable might overstate the effect of some of our demographic variables like pctymle, pctmin80 and density.

Conclusion

In conclusion, we feel that given our dataset, our recommendation would revolve strongly around policies for optimizing police forces across counties and policies around arrests, convictions and sentencing. Given access to the omiited variables mentioned in the section above, we feel that we have additional suggestions for policies around economic welfare, education and employment. Since our demoraphic variables in the suggested model are mainly included for control and we can't affect them through policy reform, our recommendation would be to optimise over staffed police offices in counties with lower crime rates and reallocte the officers to counties that appear under-staffed based on our model. In addition, we would recommend stricter enforcement of laws whereby proportionally more offenses would lead to arrests than is present currently.