W203 Lab 3: Reducing Crime

Chi Iong Ansjory, Tsung-Chin Han, Marcelo Queiroz

7/17/2018

Introduction

The motivation of this analysis is to understand the determinants of crime and to generate policy suggestions in order to reduce crime. Imagine that we have been hired to provide research for a political campaign, our data source is primarily the dataset of crime statistics for a selection of counties in North Carolina.

The Initial EDA

For this analisys, the team decided to add the car library, for scatterplot matrix comparation and the stargazer library, for improved analysis of linear models:

```
library(car)
```

```
## Loading required package: carData
```

```
library(stargazer)
```

```
##
## Please cite as:
## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

Load the cross-section data set into R and inspect it:

```
Data <- read.csv("crime_v2.csv", header=TRUE, sep=",")
summary(Data)
```

```
##      county           year        crmrte             prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6
##         prbconv      prbpris          avgsen          polpc
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6
##     density            taxpc             west            central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6
##      urban           pctmin80           wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed            wsta            wloc             mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

As we saw, the data consists of 97 observations of 25 variables, where 6 of them seems consistently missing. To make sure, we can perform:

```
Data[is.na(Data$county),]
```

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 92     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 93     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 94     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 95     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 96     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 97     NA   NA     NA     NA       `      NA     NA    NA      NA    NA
##    west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta
## 92   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 93   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 94   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 95   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 96   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 97   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
##    wloc mix pctymle
## 92   NA  NA      NA
## 93   NA  NA      NA
## 94   NA  NA      NA
## 95   NA  NA      NA
## 96   NA  NA      NA
## 97   NA  NA      NA
```

From the summary we also noticed that *prbconv* was a factor variable, and some of the variables that were supposed to be probabilities were actually greater than 1, therefore clearly wrong. In order to fix the problems exposed, we did the following:

- Convert *prbconv* from factor to numeric.
- Eliminate all missing data based *county*.
- Eliminate probability values greater than 1 from *prbarr*, *prbconv*, *prbpris*.

```
Data$prbconv = as.numeric(paste(Data$prbconv))
subcases = !is.na(Data$county) & !Data$prbarr>1 & !Data$prbconv>1 & !Data$prbpris>1
crime_data = Data[subcases, ]
```

Now, the new data frame has 81 observations which can be assessed to improve our employer policies proposal's for North Caroline. The available variable's descriptions are:

variable	label
year	1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	avg. sentence, days
polpc	police per capita
density	people per sq. mile
taxpc	tax revenue per capita
west	=1 if in western N.C.
central	=1 if in central N.C.
urban	=1 if in SMSA
pctmin80	perc. minority, 1980
wcon	weekly wage, construction
wtuc	wkly wge, trns, util, commun
wtrd	wkly wge, whlesle, retail trade
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps
mix	offense mix: face-to-face/other
pctymle	percent young male

As our employer is interested on public policies that could address the crime problem, our dependent value will be *crmrte*, or crimes commited per person. Additionally, as analizing 25 variables would be inneficient, we decided to divide our analysis in 3 steps, with the variables grouped by their nature. That said, we will have a group of variables that looks for models that explains how convictions and police enforcement explains crime rates, a separate group that looks for models that explains how etno-geographic data influences crime rates and a last one, that covers variations in wages and industries differences.

This division may be useful for figuring out the variables that may be used for a final model later, more robust and contemplating all kinds of variables. Also this was choosen in order to make the campaign decision making process easier, since policies usually have well defined areas of impact, such as housing, employment, police forces, etc.

Crime and Law Enforcement

As stated previously, the first model we want to develop is related to variables that are reflex of law enforcement policies:

variable	label
crmrte	crimes committed per person (Dependent Variable - DV)
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	avg. sentence, days
polpc	police per capita
mix	offense mix: face-to-face/other

For a first anlisys, we did a scatterplot matrix is crime rate with variables related to the nature of crime: probabilities of arrest, conviction and prison sentence, average sentence days, and log transformation of offense mix.

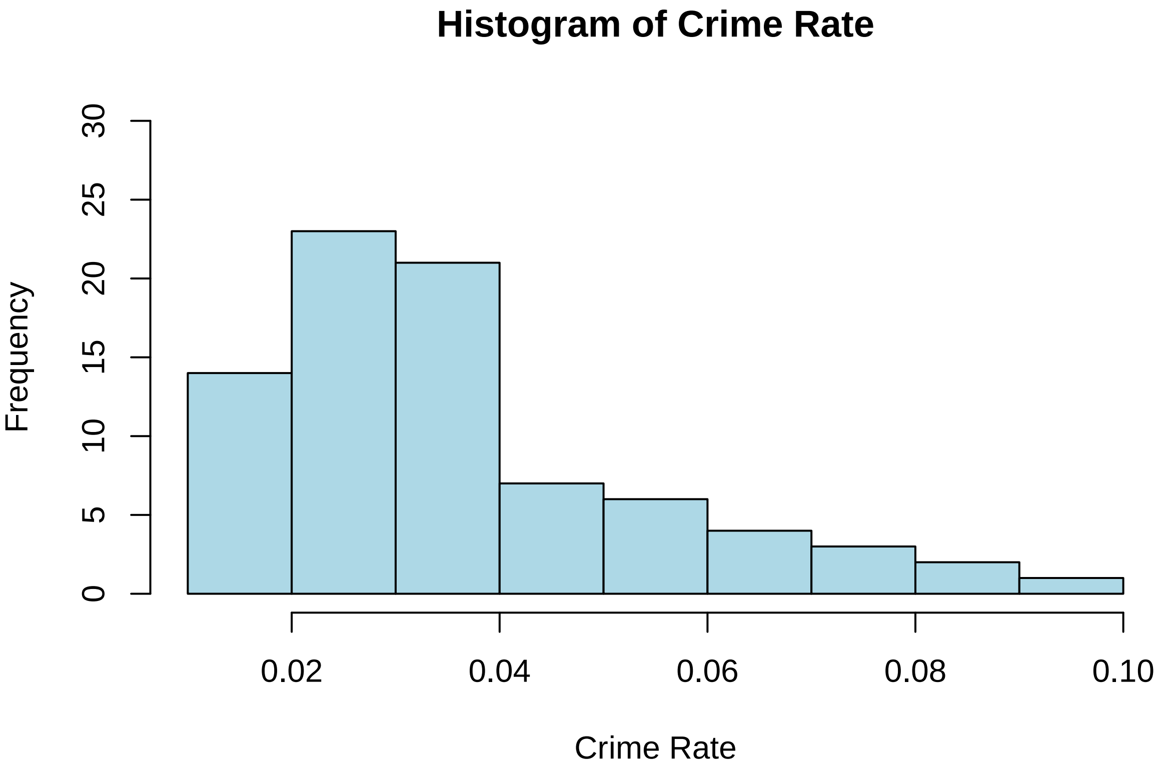
str(crime_data) names(crime_data) summary(crime_data) "'

Now, the new data frame has 81 observations. First of all, our goal is to understand the determinants of crime, crimes committed per person *crmrte* is more direct as to what we want to measure. Therefore, our dependent variable will be *crmte* (%). Let's first look at the un-transformed type.

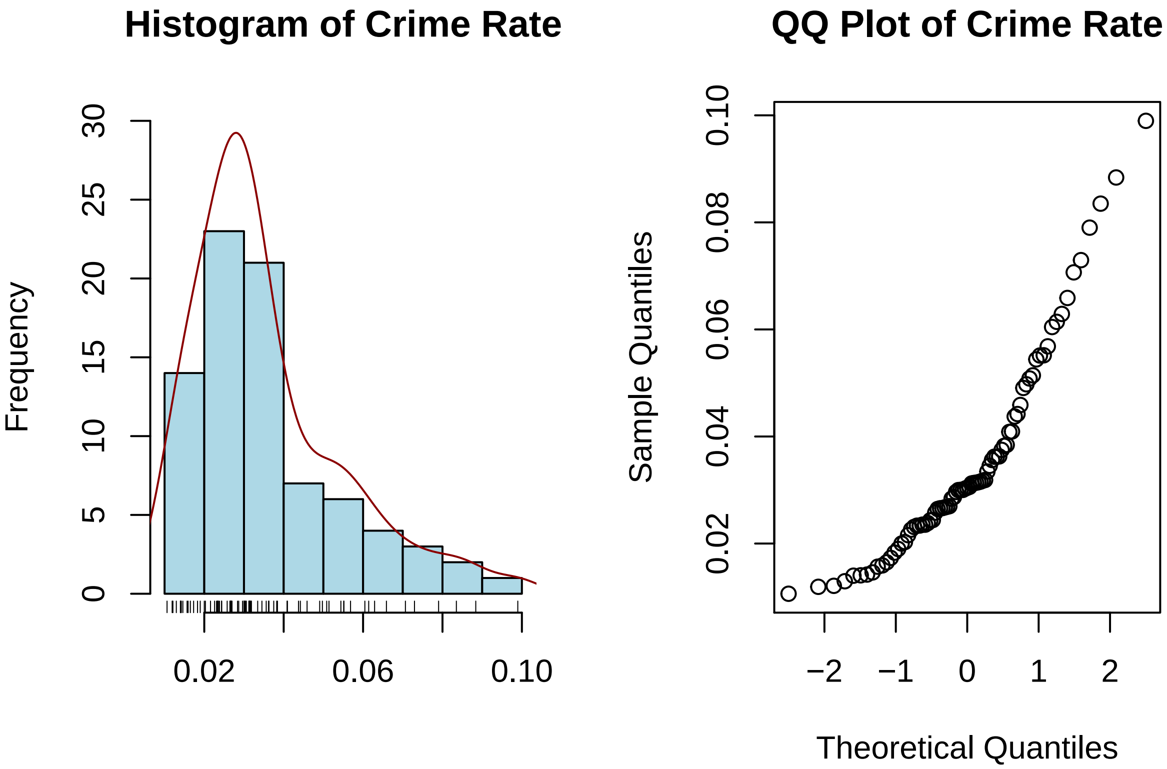
```
summary(crime_data$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02337 0.03043 0.03536 0.04374 0.09897
```

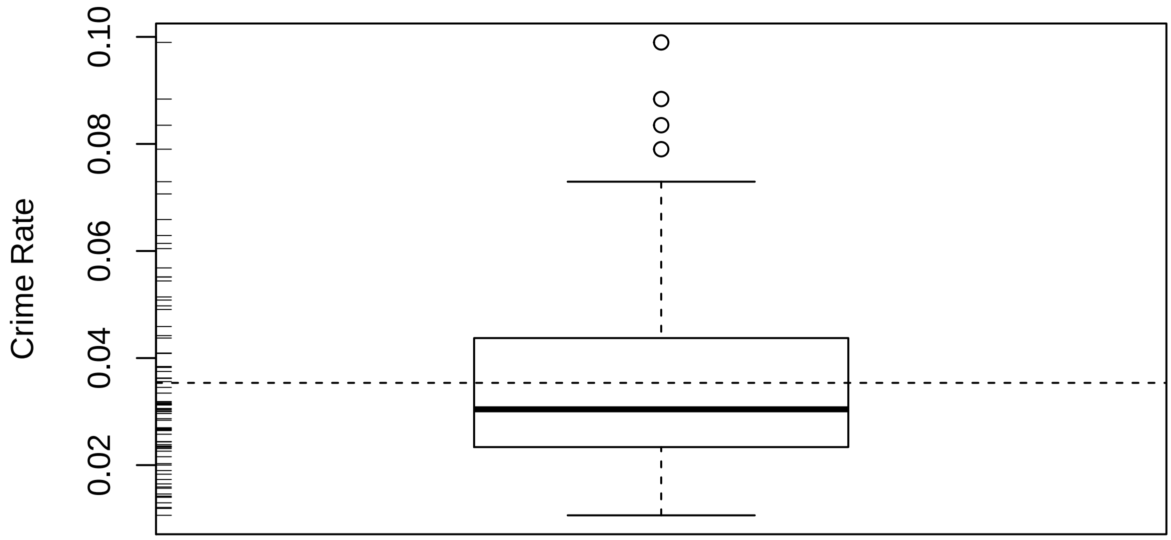
```
hist(crime_data$crmrte,
     col="light blue",
     xlab="Crime Rate", ylim=c(0,30),
     main="Histogram of Crime Rate")
```



```r
# to better understand the skewness distribution and it's spread graphically
par(mfrow=c(1,2))
hist(crime_data$crmrte, xlab="",
     col="light blue",
     main="Histogram of Crime Rate", ylim=c(0,30))
lines(density(crime_data$crmrte, na.rm=T),
      col="dark red")
rug(jitter(crime_data$crmrte))
qqnorm(crime_data$crmrte, main="QQ Plot of Crime Rate")
```



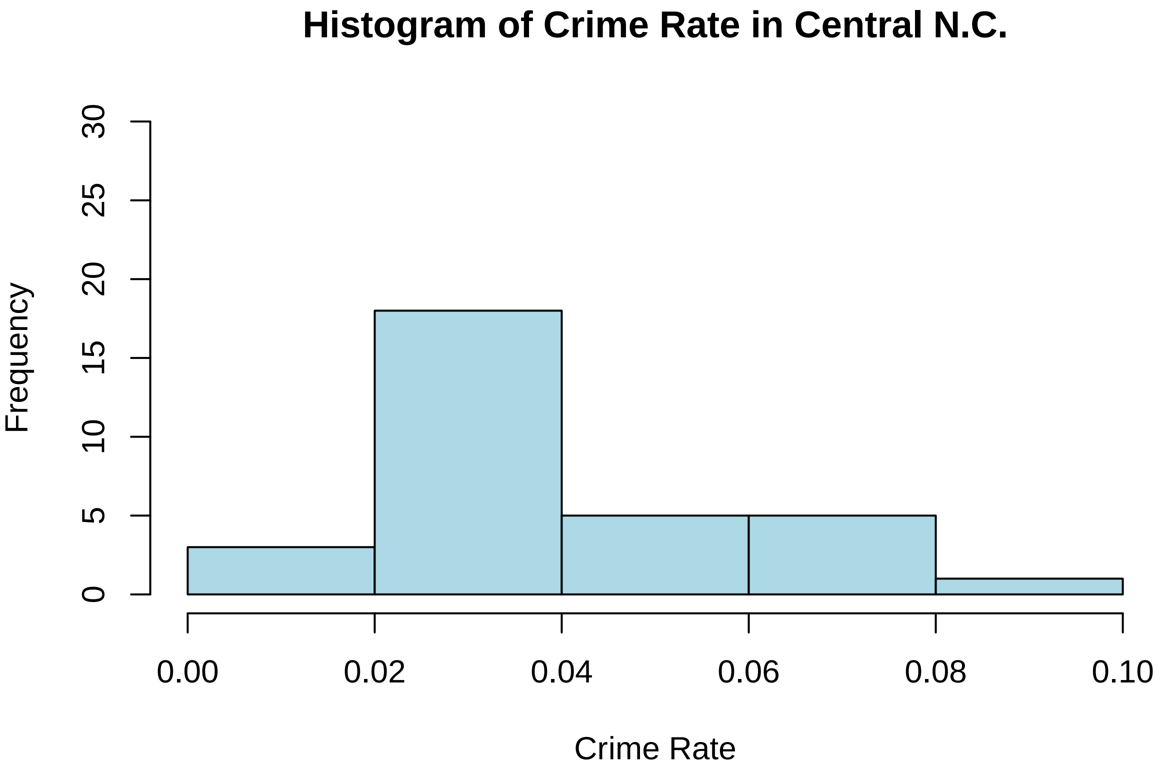
```r
par(mfrow=c(1,1))

# boxplot
boxplot(crime_data$crmrte, ylab="Crime Rate")
rug(jitter(crime_data$crmrte), side=2)
abline(h=mean(crime_data$crmrte, na.rm=T), lty=2)
```



The crime rate has right skew with the mean at 0.033, and median at 0.030. The distribution is not normally distibuted. The box plot also shows more possible outliers have distorted the value of the mean as a statistic of centrality. Also, the variable *crmrte* has a distribution of the observed values concentrated on low values, thus with a positive skew.

One last observation is central N.C. tends to have higher frequency of crime rates than west N.C. and SMSA.

```
hist(crime_data[crime_data$central == 1, ]$crmrte,
     col="light blue",
     main="Histogram of Crime Rate in Central N.C.",
     xlab="Crime Rate", ylim=c(0,30))
```

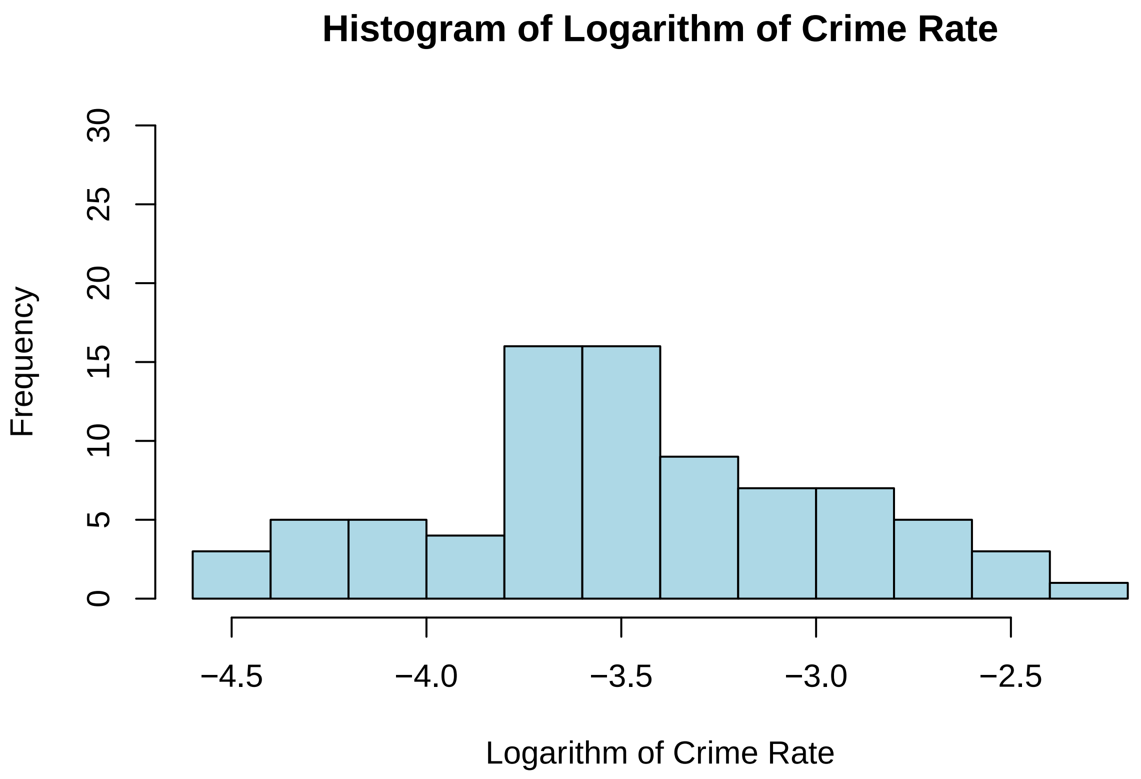


Now, let's see if we apply log transformation on the dependent variable *crmrte*.

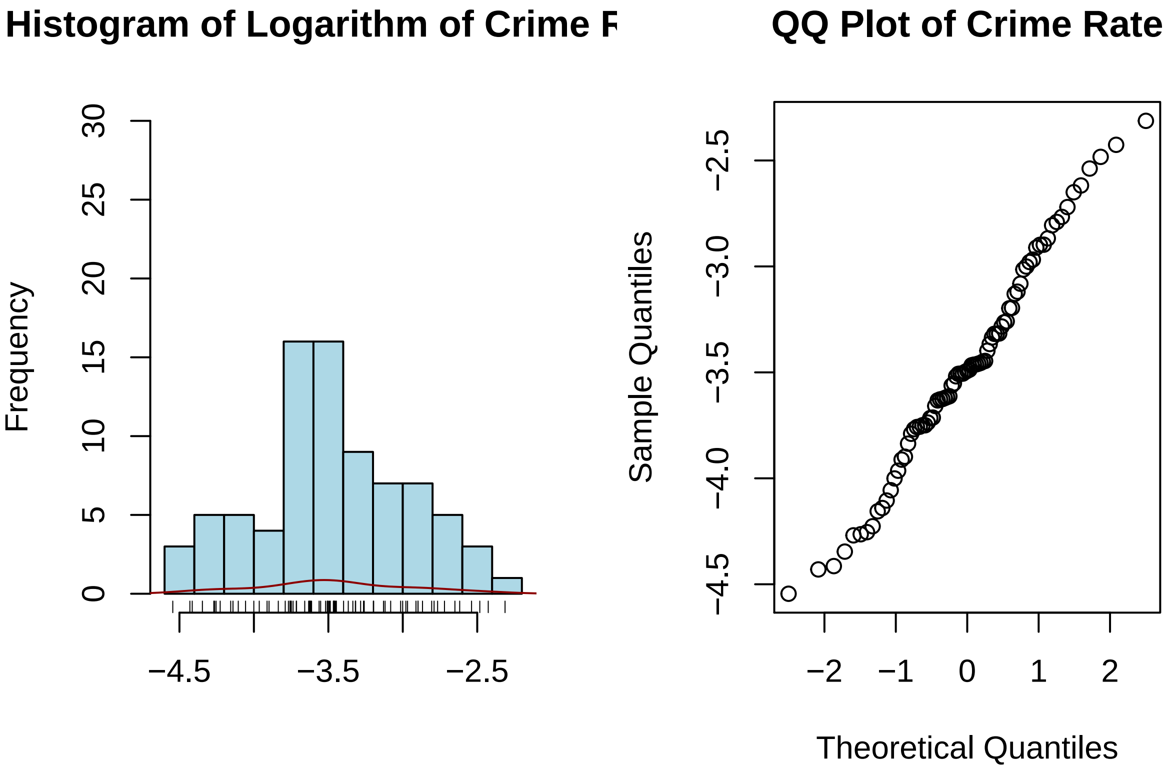
```
summary(log(crime_data$crmrte))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.545  -3.756  -3.492  -3.469  -3.130  -2.313
```

```
hist(log(crime_data$crmrte),
     col="light blue",
     xlab="Logarithm of Crime Rate", ylim=c(0,30),
     main="Histogram of Logarithm of Crime Rate")
```

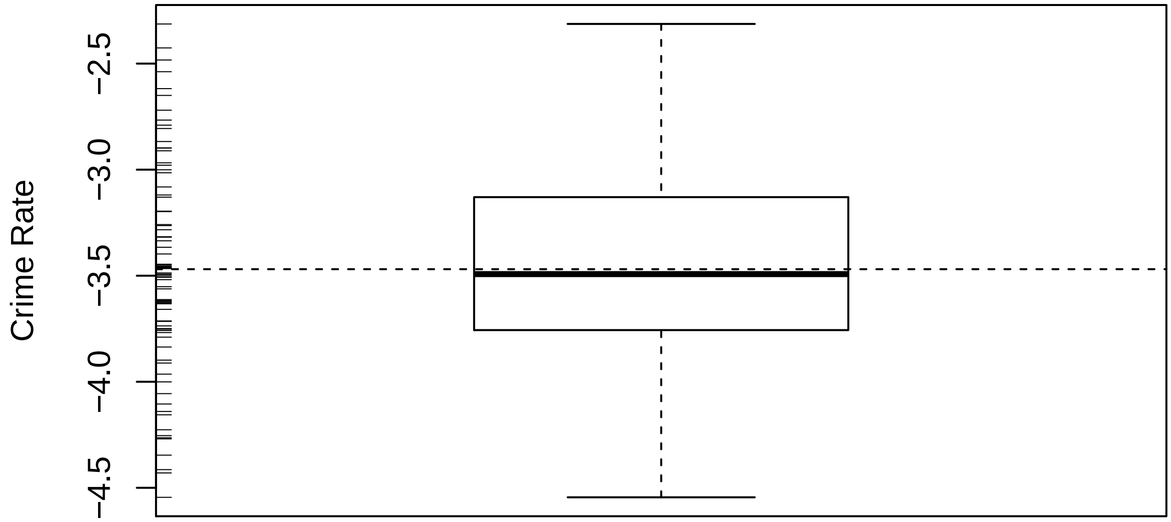


```r
# to better understand the skewness distribution and it's spread graphically
par(mfrow=c(1,2))
hist(log(crime_data$crmrte), xlab="",
     col="light blue",
     main="Histogram of Logarithm of Crime Rate", ylim=c(0,30))
lines(density(log(crime_data$crmrte), na.rm=T),
      col="dark red")
rug(jitter(log(crime_data$crmrte)))
qqnorm(log(crime_data$crmrte), main="QQ Plot of Crime Rate")
```


Histogram of Logarithm of Crime F

QQ Plot of Crime Rate

```
par(mfrow=c(1,1))

# boxplot
boxplot(log(crime_data$crmrte), ylab="Crime Rate")
rug(jitter(log(crime_data$crmrte)), side=2)
abline(h=mean(log(crime_data$crmrte), na.rm=T), lty=2)
```

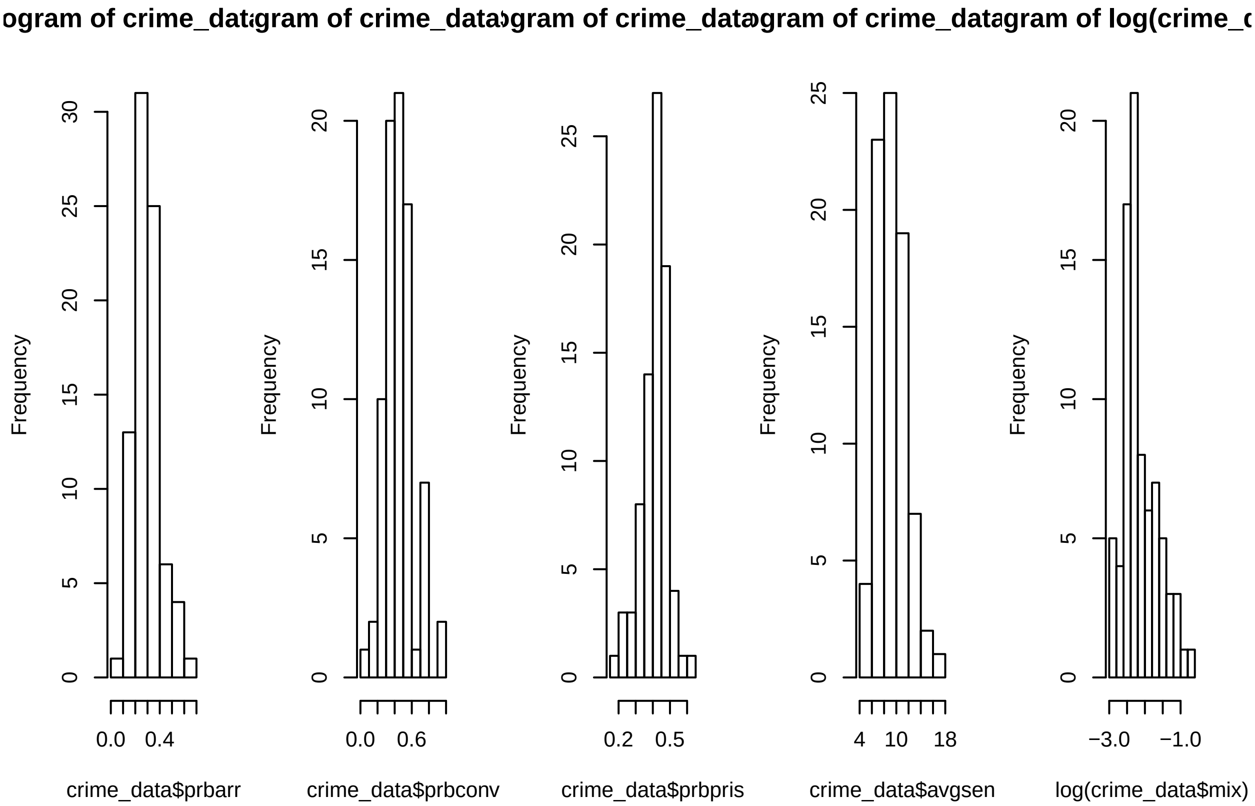


Clearly, if we apply log transformation on crime rate, our distribution becomes normally distibuted with mean and median to be very close, almost no skew and symmetric. This log transformed crime rate could be more ideal when it comes to modelling for OLS.

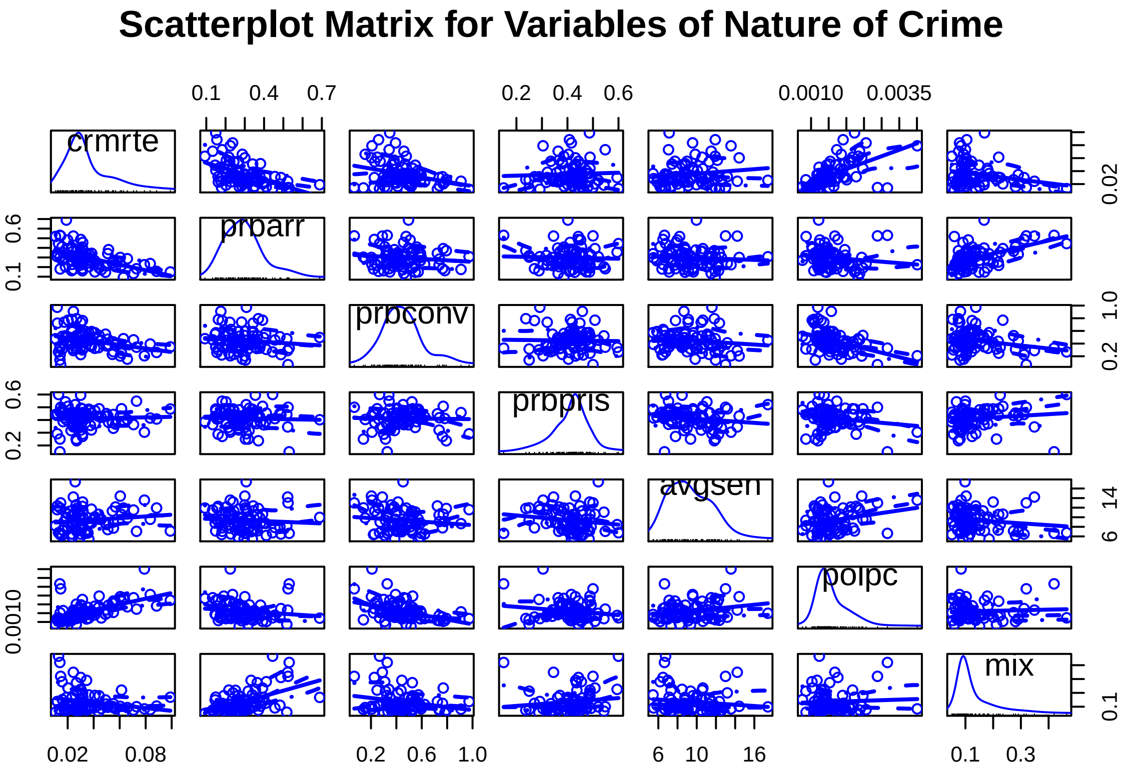
We break the variables into 3 groups to examine the relationship against crime rate.

First group is crime-related variables: $prbarr, prbconv, prbpris, avgsen, mix$. Inspecting histograms of each variable and *mix* needs to be log transformed.

```r
par(mfrow=c(1,5))
hist(crime_data$prbarr) # close to normal
hist(crime_data$prbconv) # close to normal
hist(crime_data$prbpris) # close to normal
hist(crime_data$avgsen) # close to normal
hist(log(crime_data$mix)) # close to normal
```



```r
scatterplotMatrix(~ (crmrte) + (prbarr) + (prbconv) + (prbpris) + (avgsen) + (polpc) + (mix),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Nature of Crime")
```



Scatterplot Matrix for Variables of Nature of Crime

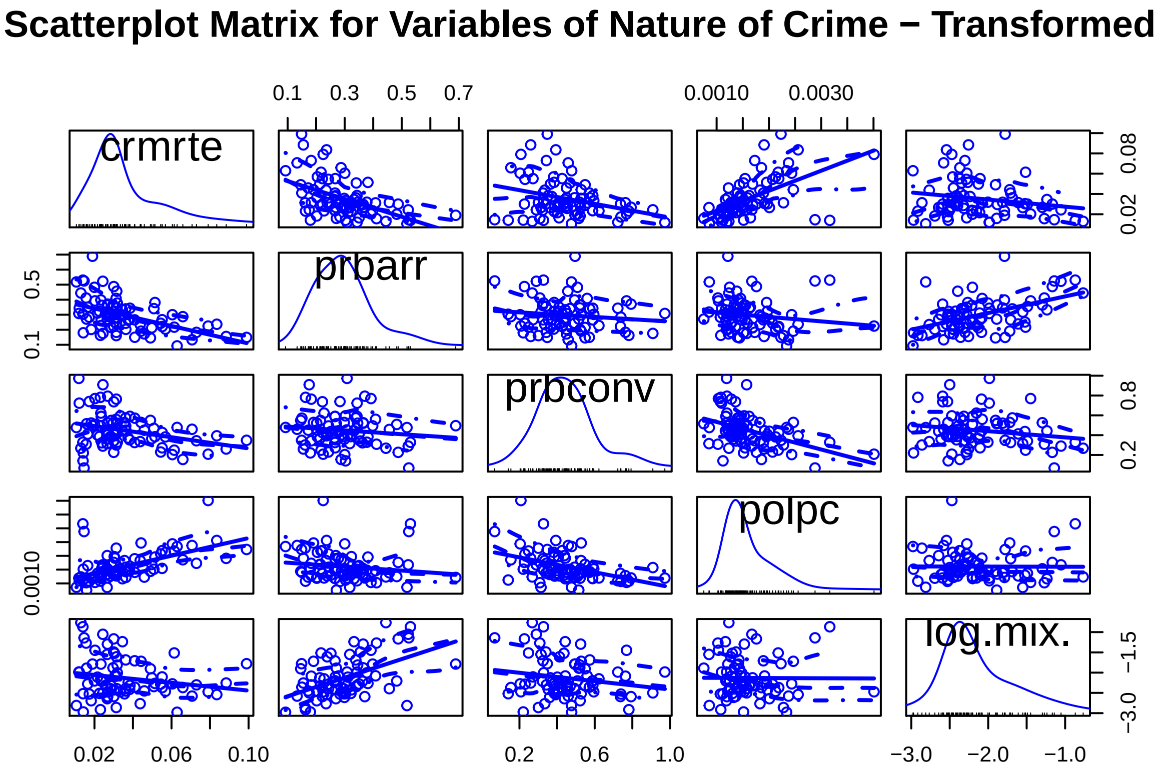
Observing the first column in the matrix, we noticed that the offense mix variable have higher spread over lower values of crime per capta and also a positevely skewed distribution, so it was decided to perform a log tranformation on that variable.

Additionally some realationship identified that are not necessarily related to our dependent variable, crime per capta, but may be useful to note for further consideration:

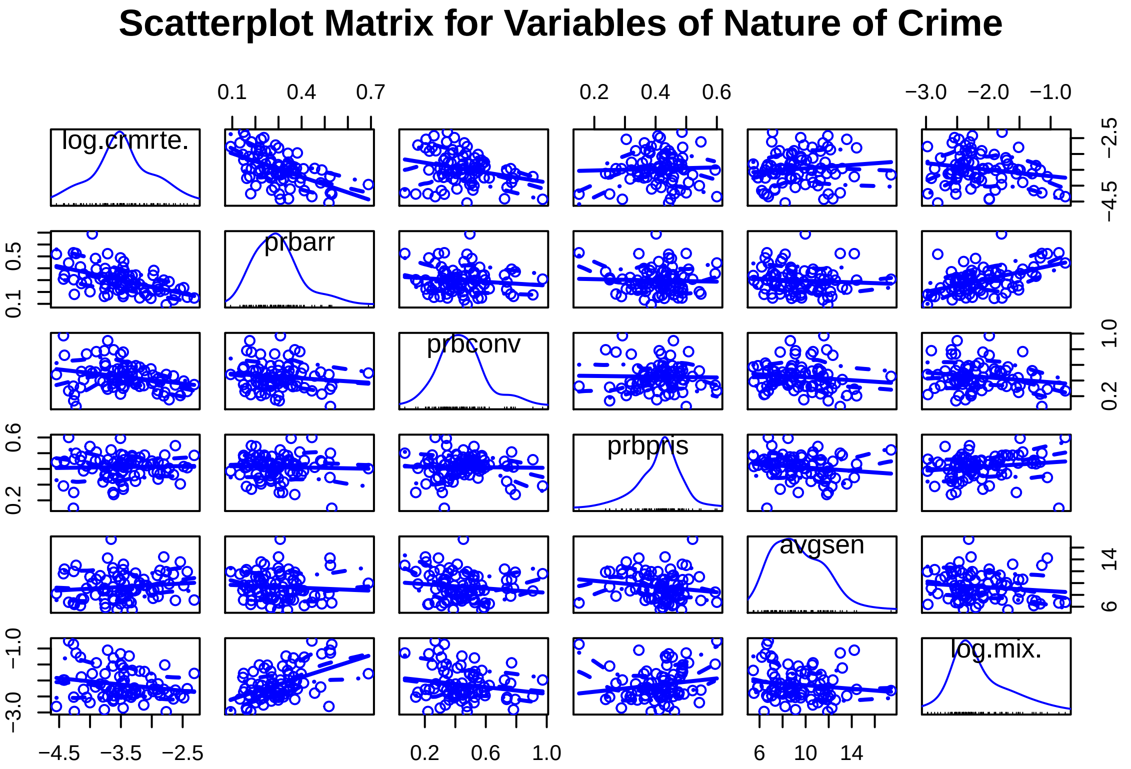
- There is strong positive relationship between probability of arrest and offense mix.
- Probability of prison sentence and average sentence days do not seem to have a strong relationship with any other variables in this group.

For an improved visualization, we dropped the variables that were hardly correlated to our DV and reploted the scatterplot matrix:

```
scatterplotMatrix(~ crmrte + prbarr + prbconv + polpc + log(mix),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Nature of Crime - Transformed")
```



```r
scatterplotMatrix(~ log(crmrte) + prbarr + prbconv + prbpris + avgsen + log(mix),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Nature of Crime")
```



```r
cor(log(crime_data$crmrte), crime_data$prbarr,
    use="complete.obs")
```

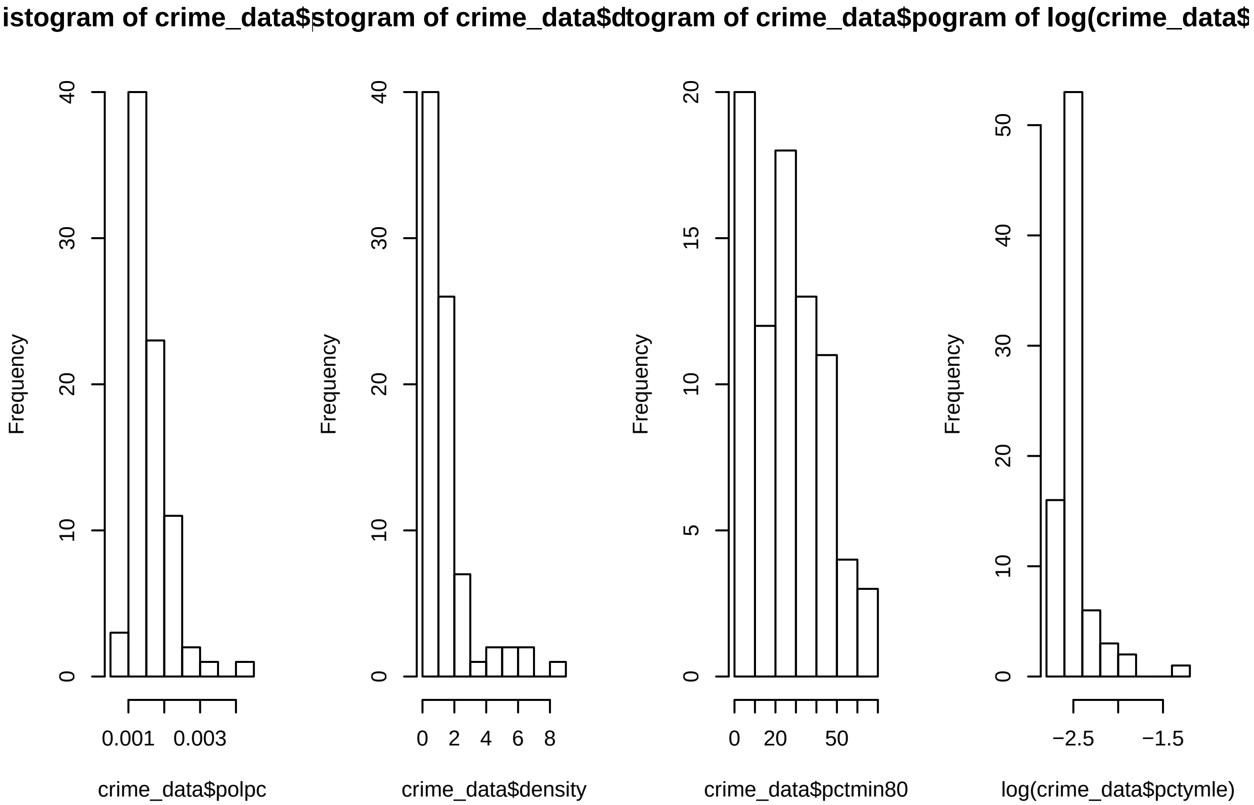
```
## [1] -0.5277865
```

```
cor(log(crime_data$crmrte), crime_data$prbconv,
    use="complete.obs")
```

```
## [1] -0.2650348
```

Second group is population-related variables: $polpc, density, pctmin80, pctymle$. Inspecting histograms of each variable and $pctymle$ needs to be log transformed.

```
par(mfrow=c(1,4))
hist(crime_data$polpc) # close to normal
hist(crime_data$density) # right skew
hist(crime_data$pctmin80) # close to normal
hist(log(crime_data$pctymle)) # right skew
```



- There are noticable negative relationship between crime rate and probability of arrest, crime rate and probability of conviction.
- More police officers per capta seems to be correlated to more crimes per capta. Although this fact seems couter-intuitive, we remember this is not a causal analisys, and more police on the streets may be an effect of a higher crime rate, not the opposite (as one could erroneously infer).
- The mix of face-to-face offences related to other seems to have a lower negative relationships, but still requires further inivestigation.

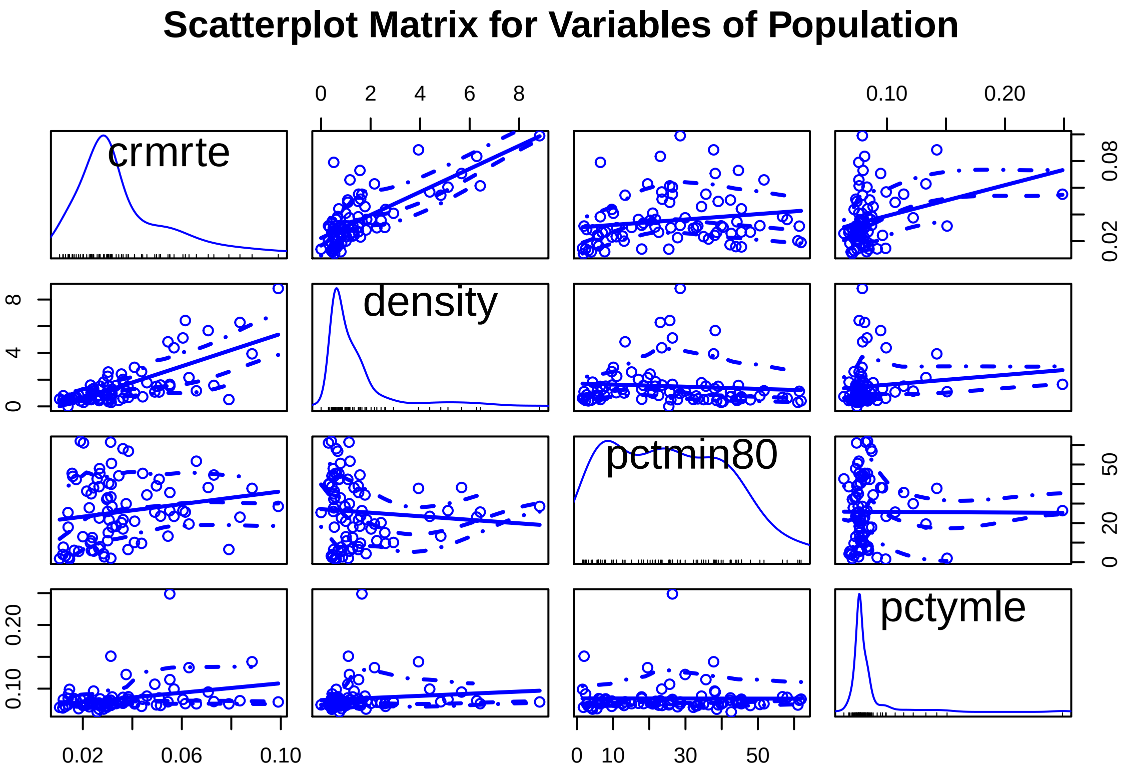
Crime and Etno-geographical variables:

Second scatterplot matrix is crime rate with variables related to population and geographical (excluding the binary variables):

Label	Description
density	people per sq. mile
pctmin80	perc. minority, 1980
pctymle	percent young male

The same scatterplot matris analisys gave us:

```
scatterplotMatrix(~ (crmrte) + (density) + (pctmin80) + (pctymle),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Population")
```

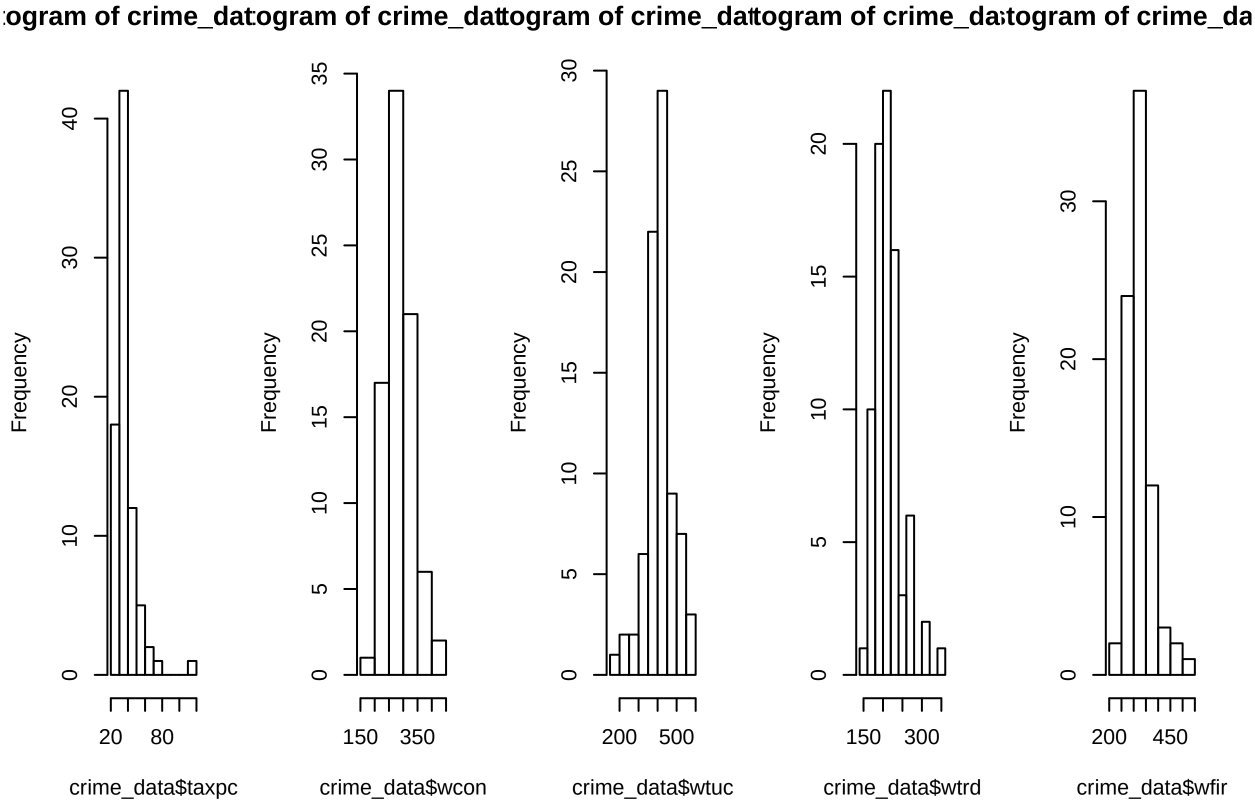


```
cor(log(crime_data$crmrte), crime_data$density,
    use="complete.obs")
```

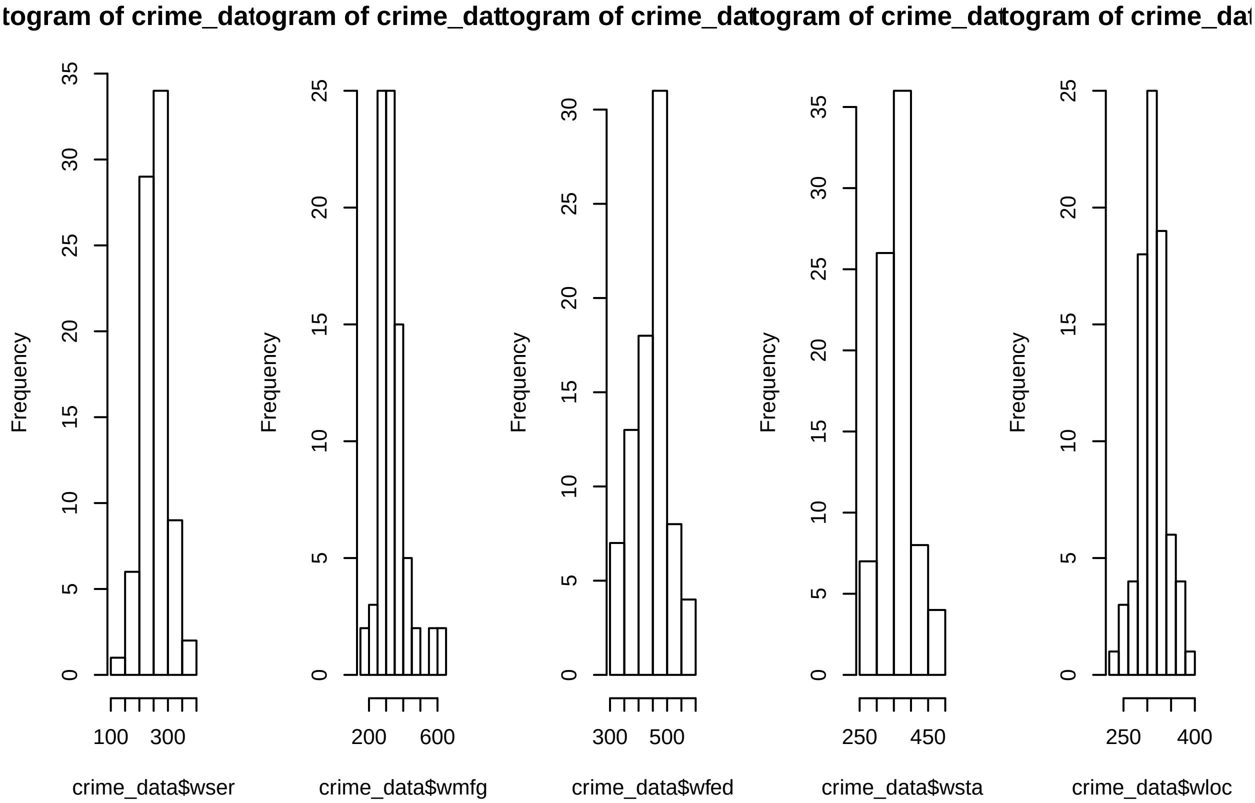
```
## [1] 0.6451216
```

Third group is economy-related variables: $taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wtoc$. Inspecting histograms of each variable.

```
par(mfrow=c(1,5))
hist(crime_data$taxpc) # right skew
hist(crime_data$wcon) # close to normal
hist(crime_data$wtuc) # close to normal
hist(crime_data$wtrd) # close to normal
hist(crime_data$wfir) # close to normal
```



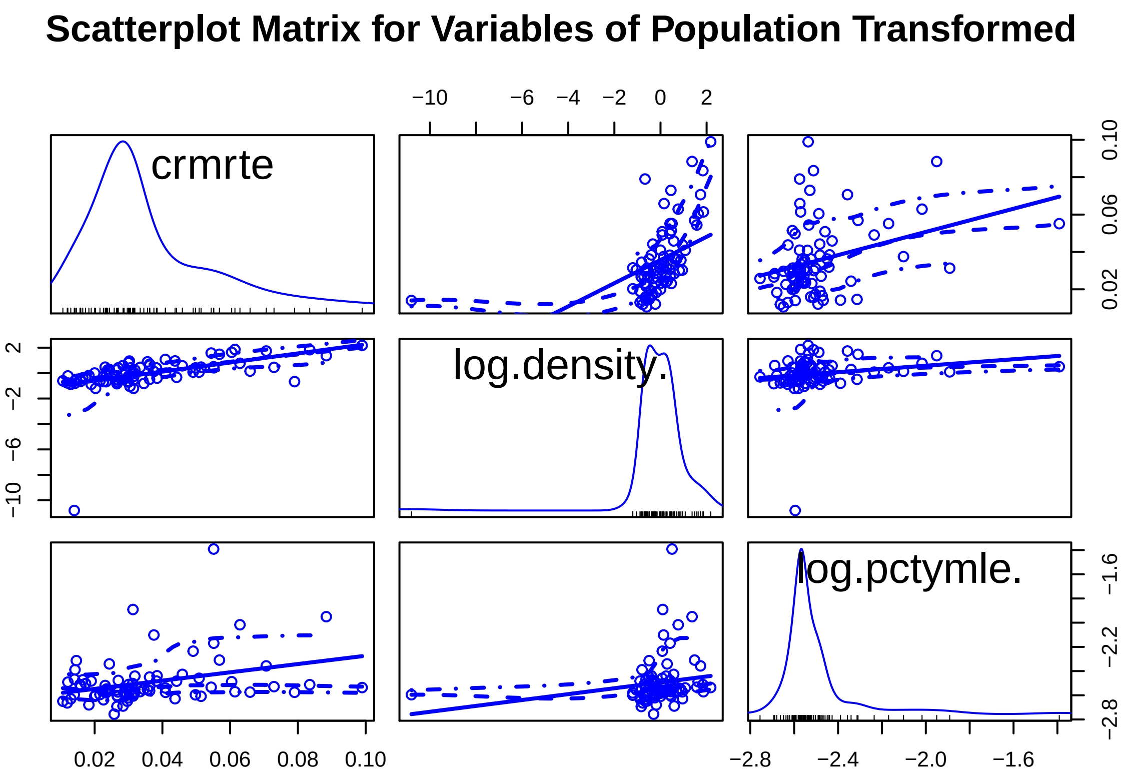
```
par(mfrow=c(1,5))
hist(crime_data$wser) # close to normal
hist(crime_data$wmfg) # close to normal
hist(crime_data$wfed) # close to normal
hist(crime_data$wsta) # close to normal
hist(crime_data$wloc) # close to normal
```



As once again the data seemed positevely skewed for pctymle and density variables, and we did a log tranformation. Also, as the percentage of minority does not seem to affect any of the other variables, this variable was dropped for this analisys (although this may be an important output of this research as well).

The new scatterplot matrix was:

```
scatterplotMatrix(~ crmrte + log(density) + log(pctymle),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Population Transformed")
```



Here is a features noticed from the matrix:

- There are noticable positive relationship between crime rate and people per sq. mi., and % young male (although the latter is still positevely skewed).

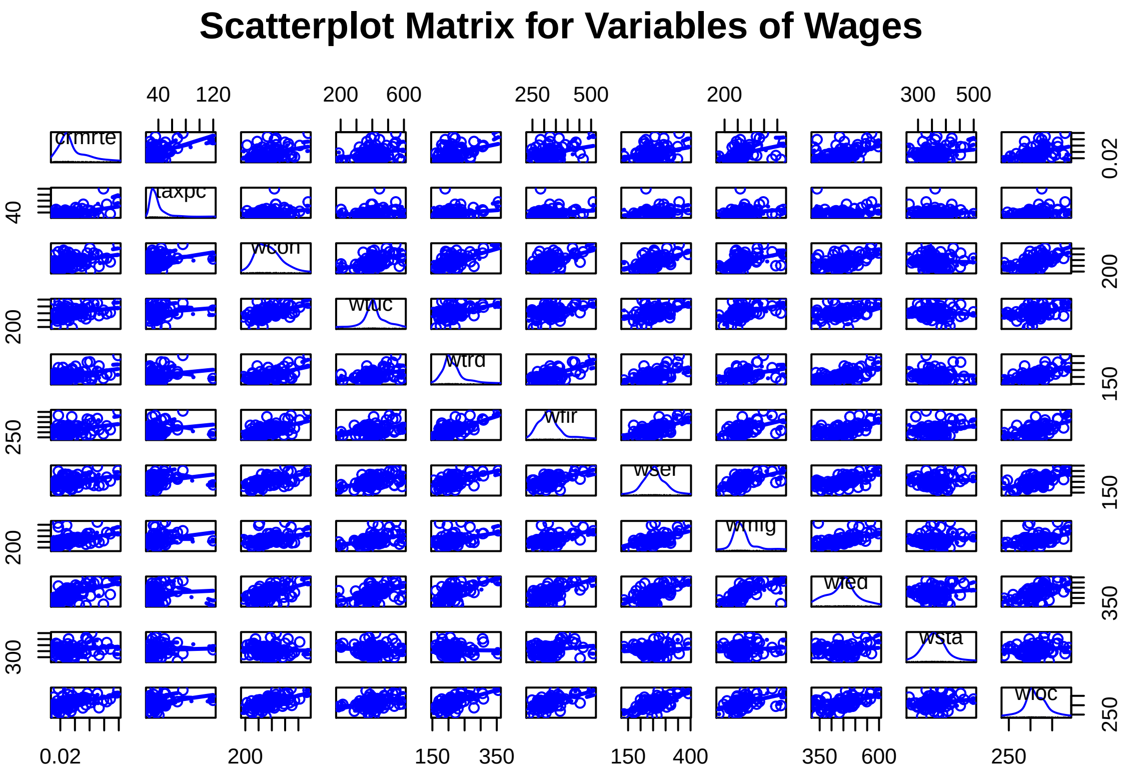
Crime and wage:

For the wage related variables, the folowing were used:

Label	Description
taxpc	tax revenue per capita
wcon	weekly wage, construction
wtuc	wkly wge, trns, util, commun
wtrd	wkly wge, whlesle, retail trade
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps

The usual scatterplot matrix was:

```
scatterplotMatrix(~ crmrte + taxpc + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc,
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Wages" )
```



```r
cor(log(crime_data$crmrte), crime_data$wcon,
    use="complete.obs")
```

```
## [1] 0.3435583
```

```r
cor(log(crime_data$crmrte), crime_data$wtrd,
    use="complete.obs")
```

```
## [1] 0.3518993
```

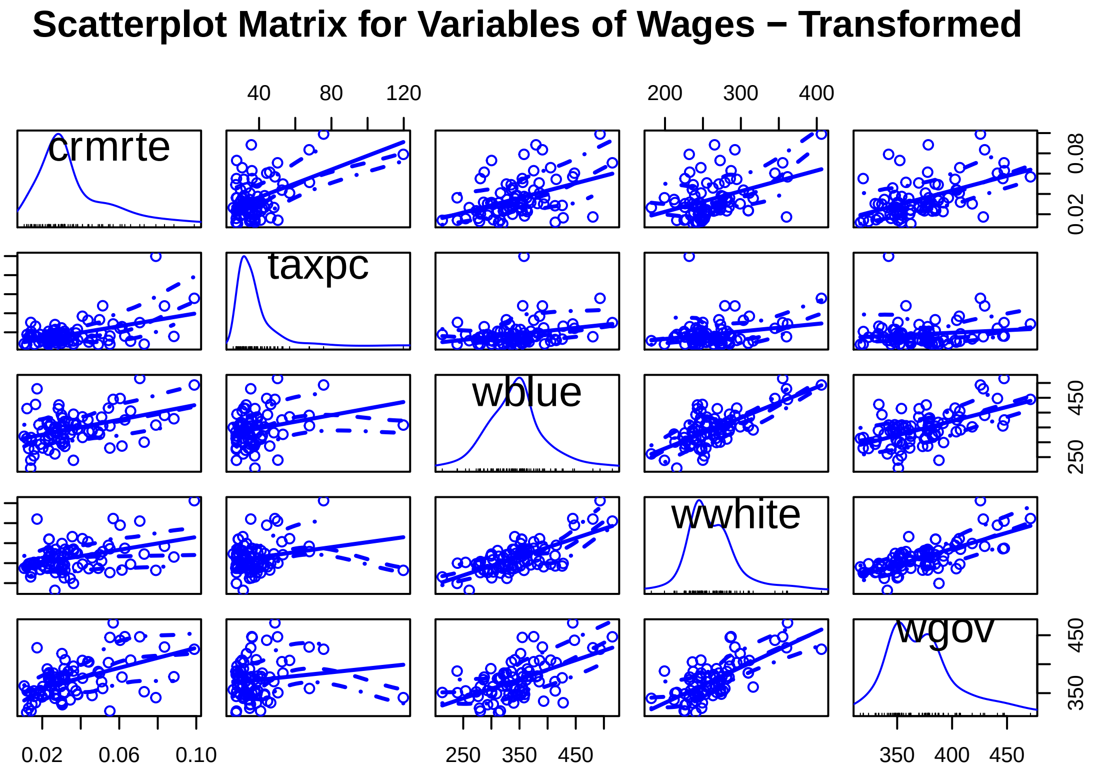
```r
cor(log(crime_data$crmrte), crime_data$wfed,
    use="complete.obs")
```

```
## [1] 0.5266092
```

<<<<<<< HEAD:Ansjory_Han_Queiroz_lab3_draft.Rmd As it seems to exists a positive relationship between wages in all studied industries and crime rates, we decided to summerize those variables into 3 more intuitive ones:

Label	Description
taxpc	tax revenue per capita
wblue	wcon + wtuc + wmfg = average Blue-collar professionals
wwhite	wtrd + wfir + wser = average White-collar professionals
wgov	wfed + wsta + wloc = average Government professional

```r
crime_data$wblue <- (crime_data$wcon + crime_data$wtuc + crime_data$wmfg)/3
crime_data$wwhite <- (crime_data$wtrd + crime_data$wfir + crime_data$wser)/3
crime_data$wgov <- (crime_data$wfed + crime_data$wsta + crime_data$wloc)/3

scatterplotMatrix(~ (crmrte) + (taxpc) + (wblue) + (wwhite) + (wgov),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Wages - Transformed" )
```



Again, we can see the correlations seems positive, indicating that higher wages are related to higher crime rates, which seems counter intuitive again, which reminds us that there are no causation realationships here. This maybe addressed on further developments if this study.

One last observation is central N.C. tends to have higher frequency of crime rates than west N.C. and SMSA.

The Model Building Process

The purpose of this analysis is to identify variables relevant to the concerns of the political campaign in order to reduce crime rate.

Those variables found correlated to crime rate from EDA as follow:

- Potentially applicable for policy suggestions: $prbarr$, $prbconv$, $taxpc$
- Not applicable for policy suggestions: $density$, $pctymle$, $w*$

The covariates that help us identify a causal effect are $prbarr$ and $prbconv$, $density$ and $pctymle$. On the other hand, the problematic covariates due to multicollinearity are $taxpc$ and $w*$ since they will absorb some of causal effect we want to measure.

We will consider building 3 model specifications:

1. Model with only the explanatory variables of key interest and no other covariates.

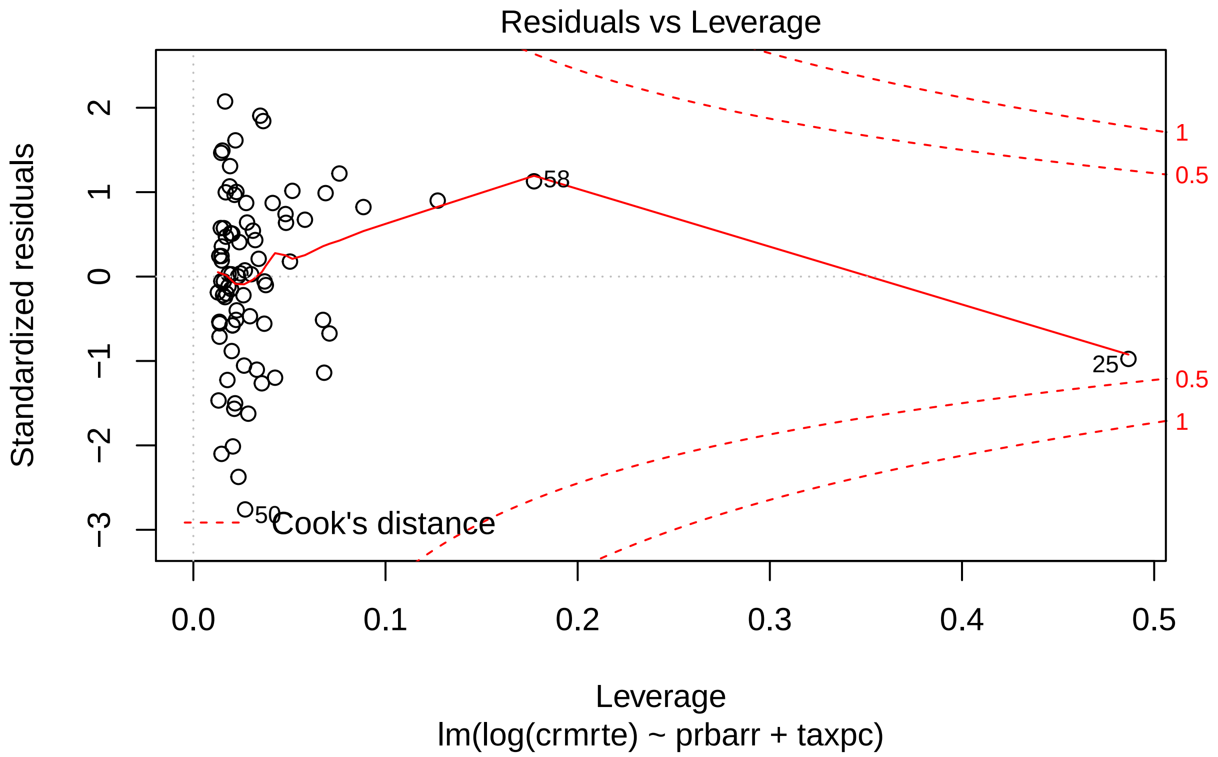
$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 taxpc + u$$

Picking variables which are only applicable for policy suggestions as the key interest with no other covariates from each variable.

```
(model1 = lm(log(crmrte) ~ prbarr + taxpc,
             data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + taxpc, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr        taxpc
##    -3.27518     -2.29379      0.01279
```

```
plot(model1, which = 5)
```



```
summary(model1)$r.square
```

```
## [1] 0.3899895
```

```
summary(model1)$adj.r.squared
```

```
## [1] 0.3743482
```

```
AIC(model1)
```

```
## [1] 86.31843
```

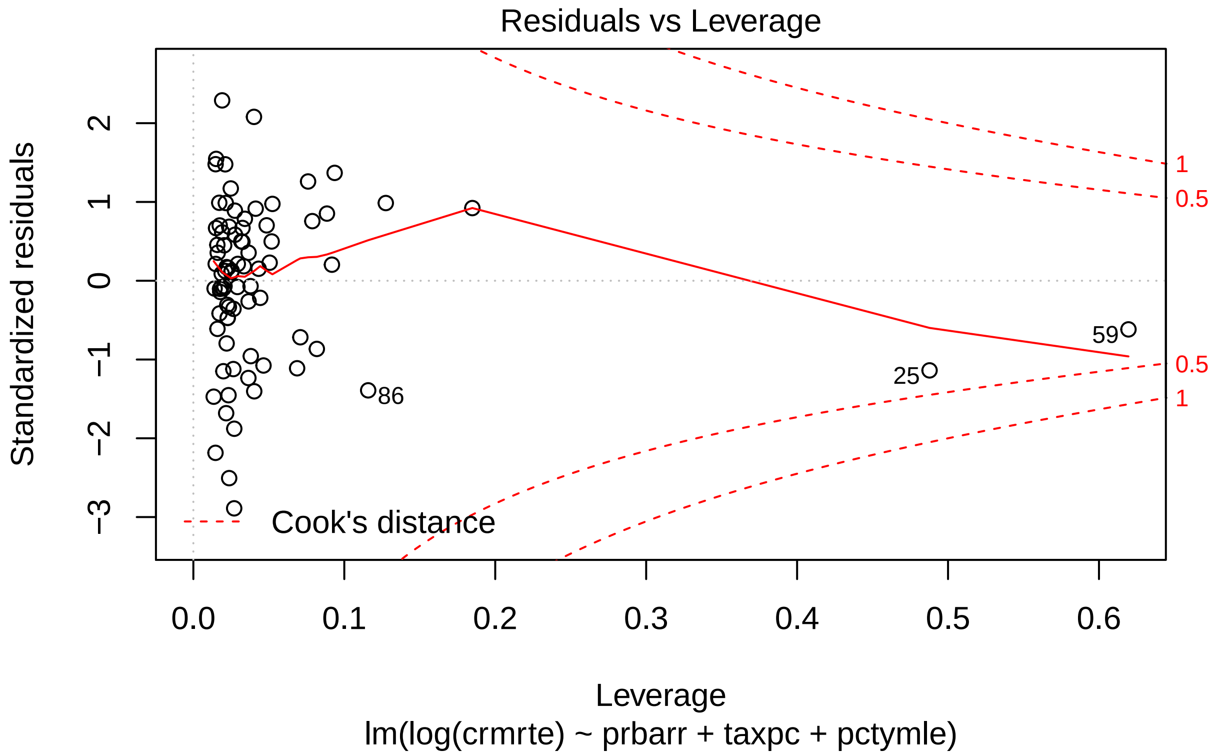
2. Model that includes key explanatory variables and only covariates that we believe increase the accuracy of your results.

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 taxpc + \beta_3 pctymle + u$$

```
(model2 = lm(log(crmrte) ~ prbarr + taxpc + pctymle,
             data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + taxpc + pctymle, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr        taxpc      pctymle
##    -3.80317     -2.05544      0.01393      4.89767
```

```
plot(model2, which = 5)
```



```
summary(model2)$r.square
```

```
## [1] 0.4404113
```

```
summary(model2)$adj.r.squared
```

```
## [1] 0.4186091
```

```
AIC(model2)
```

```
## [1] 81.33023
```

Adjusted R^2 increases by 11.8% by adding one additional variable, and AIC decreases by 5.78% to indicate improvements on parsimony. However, there is not a significant changes on accuracy when comparing the Cook's distance.

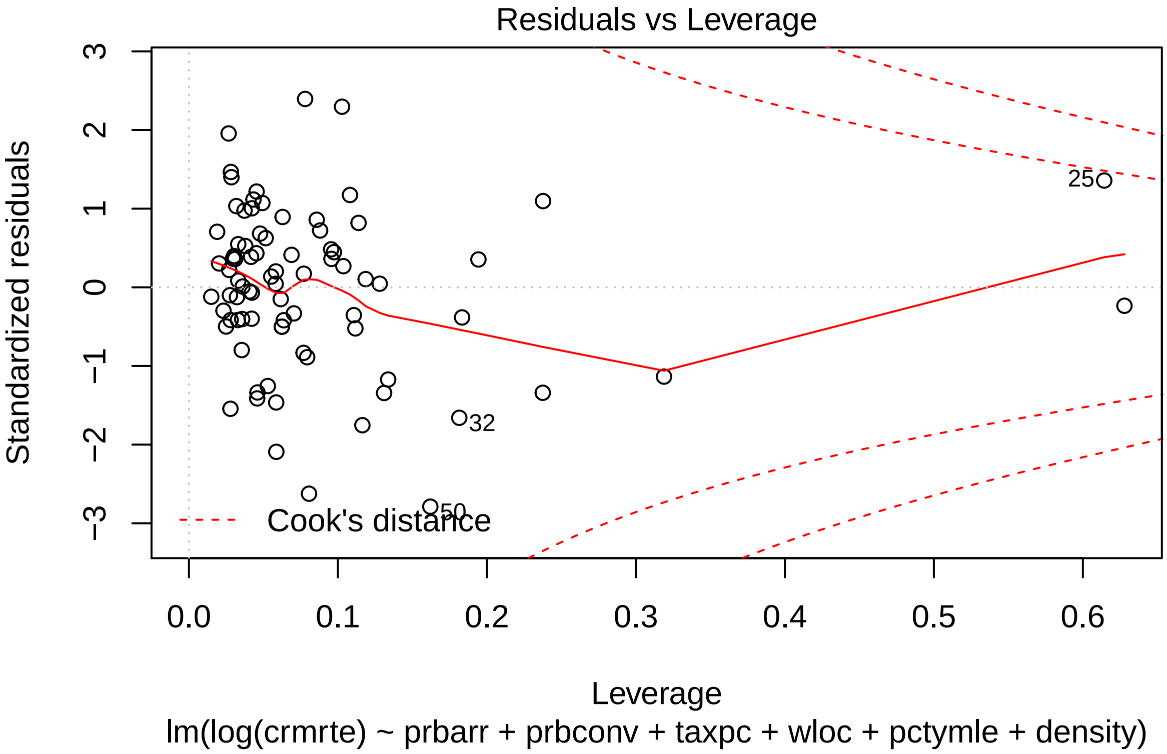
3. Model that includes the previous covariates, and most, if not all, other covariates.

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + \beta_3 taxpc + \beta_4 wloc + \beta_5 pctymle + \beta_6 density + u$$

```
(model3 = lm(log(crmrte) ~ prbarr + prbconv + taxpc + wloc + pctymle + density,
             data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv + taxpc + wloc +
##     pctymle + density, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        taxpc         wloc
##   -4.118106    -1.482461    -0.349108     0.007134     0.001581
##     pctymle      density
##    3.585714     0.117496
```

```
plot(model3, which = 5)
```



```
summary(model3)$r.square
```

```
## [1] 0.5939268
```

```
summary(model3)$adj.r.squared
```

```
## [1] 0.5610019
```

```
AIC(model3)
```

```
## [1] 61.35607
```

Adjusted R^2 increases by 34.0% by adding 3 additional variables, and AIC decreases by 24.6% to indicate further improvements on parsimony. Moreover, there is a significant changes on accuracy when comparing the Cook's distance.

The Regression Table

```
stargazer(model1, model2, model3, type = "latex",
          report = "vc",
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Jul 30, 2018 - 21:11:45

Table 6: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
prbarr	−2.294	−2.055	−1.482
prbconv			−0.349
taxpc	0.013	0.014	0.007
wloc			0.002
pctymle		4.898	3.586
density			0.117
Constant	−3.275	−3.803	−4.118
Observations	81	81	81
R^2	0.390	0.440	0.594

According to Table 1, for Model 1, increasing the probability of arrest will reduce crime rate with minimal effect from tax revenue per capita. For Model 2, on top of Model 1, decreasing % of young male will reduce crime rate. For Model 3, on top of Model 2, increasing both probabilities of arrest and conviction, decreasing people per sq. mi. will reduce crime rate.

Inference for linear regression and standard errors via statistical tests will be performed on the later draft.

The Omitted Variables Discussion

The omitted variables discussion will be based on Model 1 with *taxpc* dropped since its effect is minimal with following 5 variables omitted one at a time.

1. Omitted *taxpc*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 taxpc + u$$

$$taxpc = \alpha_0 + \alpha_1 prbarr + u$$

```
(omit1_pri = lm(log(crmrte) ~ prbarr + taxpc, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + taxpc, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr        taxpc
##    -3.27518     -2.29379      0.01279
```

```
(omit1_sec = lm(taxpc ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = taxpc ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##       41.87       -12.89
```

Since $\beta_2 = 0.01279$ and $\alpha_1 = -12.89$, then $OMVB = \beta_2\alpha_1 = -0.1649$. Since $\beta_1 = -2.2938 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

2. Omitted *prbconv*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + u$$

$$prbconv = \alpha_0 + \alpha_1 prbarr + u$$

```
(omit2_pri = lm(log(crmrte) ~ prbarr + prbconv, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv
##     -2.2442      -2.6470      -0.9807
```

```
(omit2_sec = lm(prbconv ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = prbconv ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##      0.5052      -0.1921
```

Since $\beta_2 = -0.9807$ and $\alpha_1 = -0.1921$, then $OMVB = \beta_2\alpha_1 = 0.1884$. Since $\beta_1 = -2.647 < 0$, the OLS coefficient on *prbarr* will be scaled toward zero (less negative) losing statistical significance.

3. Omitted *pctymle*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 pctymle + u$$

$$pctymle = \alpha_0 + \alpha_1 prbarr + u$$

```
(omit3_pri = lm(log(crmrte) ~ prbarr + pctymle, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + pctymle, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      pctymle
##      -3.119       -2.282        3.870
```

```
(omit3_sec = lm(pctymle ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = pctymle ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##     0.09810     -0.04568
```

Since $\beta_2 = 3.870$ and $\alpha_1 = -0.04568$, then $OMVB = \beta_2 \alpha_1 = -0.1768$. Since $\beta_1 = -3.119 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

4. Omitted *density*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 density + u$$

$$density = \alpha_0 + \alpha_1 prbarr + u$$

```
(omit4_pri = lm(log(crmrte) ~ prbarr + density, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + density, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      density
##     -3.2691      -1.5169       0.1657
```

```
(omit4_sec = lm(density ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = density ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##       3.195       -5.682
```

Since $\beta_2 = 0.1657$ and $\alpha_1 = -5.682$, then $OMVB = \beta_2 \alpha_1 = -0.9415$. Since $\beta_1 = -1.5169 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

5. Omitted *mix*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 mix + u$$

$$mix = \alpha_0 + \alpha_1 prbarr + u$$

```
(omit5_pri = lm(log(crmrte) ~ prbarr + mix, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + mix, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr          mix
##    -2.74009     -2.46742      0.02237
```

```
(omit5_sec = lm(mix ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = mix ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##      0.0190       0.3936
```

Since $\beta_2 = 0.02237$ and $\alpha_1 = 0.3936$, then $OMVB = \beta_2 \alpha_1 = 0.0088$. Since $\beta_1 = -2.4674 < 0$, the OLS coefficient on *prbarr* will be scaled toward zero (less negative) losing statistical significance.

Conclusion

Based on the analysis on several models, the determinants of crime are essentially probability of arrest, probability of conviction, and % young male. In order to reduce crime, the policy suggestions would be as follow for local government:

- Increase the probability of arrest when offense occurs.
- Increase the probability of conviction when arrest occurs.
- Decrease the % young male by allocating more police workforce to manage communities with high % of young male, especially in area of central N.C.