Lab3 Draft, w203: Statistics for Data Science

Avinash Chandrasekaran

March 26, 2018

1. Introduction

Our team has been hired to provide research for a political campaign. The campaign has obtained a dataset of crime statistics for a selection of counties in North Carolina. Our task is to examine the data to help the campaign understand the determinants of crime and to generate policy suggestions that are applicable to local government.

The data provided consists of 25 variables and 91 different observations collected in a given year. Moreover the dataset obtained is a single cross-section of data collected from variety of different sources. For the analysis made in this research, we will assume that the data collected from different counties in NC were randomly sampled.

Our primary analysis of data will include ordinary least squares regressions to make casual estimates and we will clearly explain how omitted variabled may affect our conclusions. We begin our research by conducting exploratory analysis of the dataset to gain a better understanding of the variables.

2. Exploratory Analysis

Data processing

```
# Read the csv file
crime_data_raw = read.csv("crime_v2.csv")

# Print out summary of the data read (excluded from the report)
#summary(crime_data_raw)
#str(crime_data_raw)
```

There appears to be 6 rows of NA's across all variables. We also notice that 'prbconv' is a factor while the rest of the variables are numeric.

```
# Remove NA rows
crime_data = na.omit(crime_data_raw)

# Print out all column names (excluded from the report)
#colnames(crime_data)

# convert factor to numeric for variable prbconv
crime_data$prbconv = as.numeric(levels(crime_data$prbconv)[crime_data$prbconv])
```

There are a total of 91 observations across 25 different variables. We will now explore each of the variables collected in the data

County and Year variables just represent the different counties and the year the data was collected. As such these don't require further analysis.

```
crime_data %>% group_by(county) %>% filter(n()>1)
```

```
## # A tibble: 2 x 25
## # Groups: county [1]
##   county  year crmrte prbarr prbconv prbpris avgsen   polpc density taxpc
##    <int> <int>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl>
## 1    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## 2    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## # ... with 15 more variables: west <int>, central <int>, urban <int>,
## #   pctmin80 <dbl>, wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>, wser
## #   <dbl>, wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,
## #   pctymle <dbl>
```

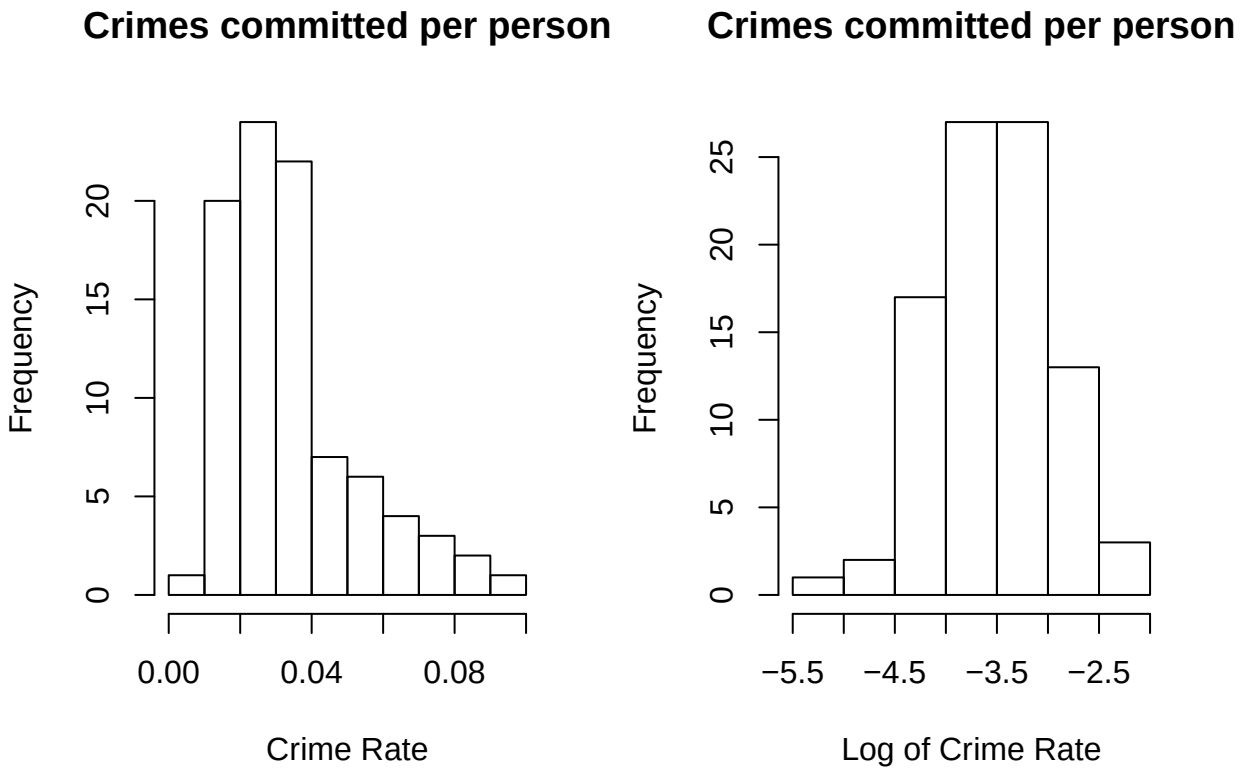
```
crime_data = distinct(crime_data)
```

We also noticed that one of the rows was duplicated in the dataset. As this could potentially affect our regression analysis, we decided to remove the duplicated entry.

Individual Variable Analysis

Crimes committed per person

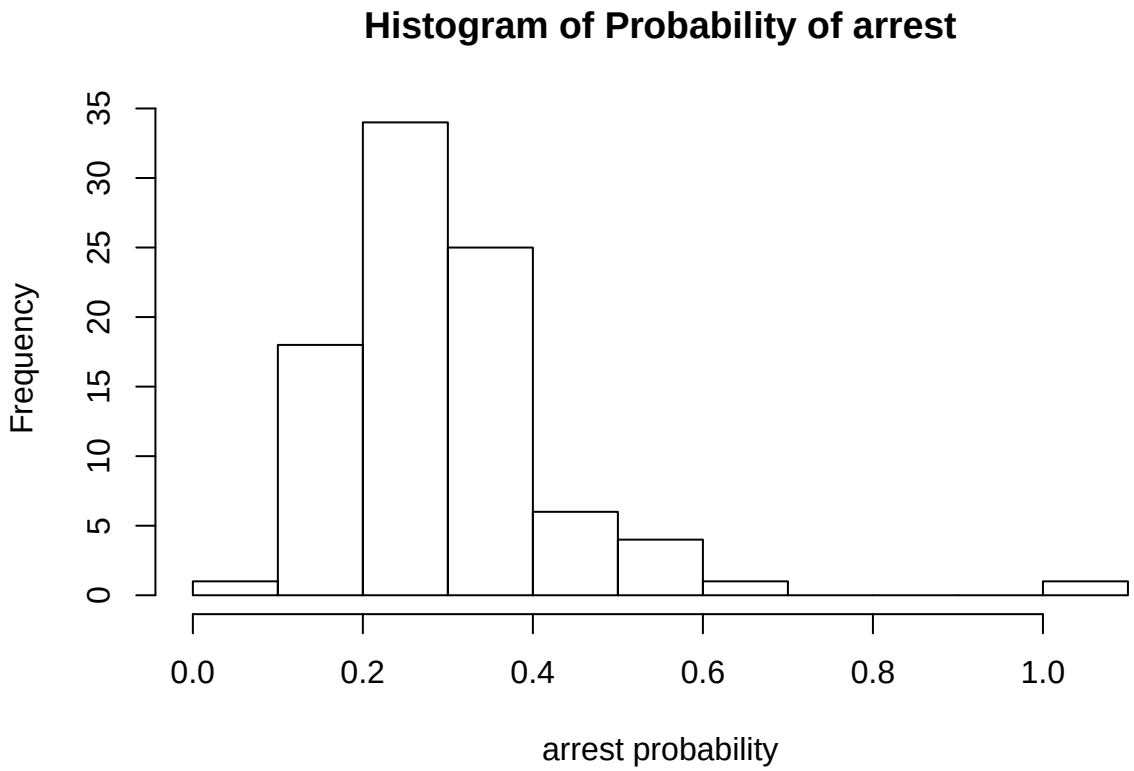
```
par(mfrow=c(1,2))
hist(crime_data$crmrte, main = "Crimes committed per person",
     xlab="Crime Rate")
hist(log(crime_data$crmrte), main = "Crimes committed per person",
     xlab="Log of Crime Rate")
```



The crime rate variable is the key dependent variable of interest. Looking at the histogram, the distribution is positively skewed to the left. We can take the log transformation which makes the variable appear more normally distributed. As a result, for our modeling, we will stick with using the log of the crime rate from here on.

Probability of arrest

```
hist(crime_data$prbarr, main = "Histogram of Probability of arrest",
     xlab="arrest probability")
```



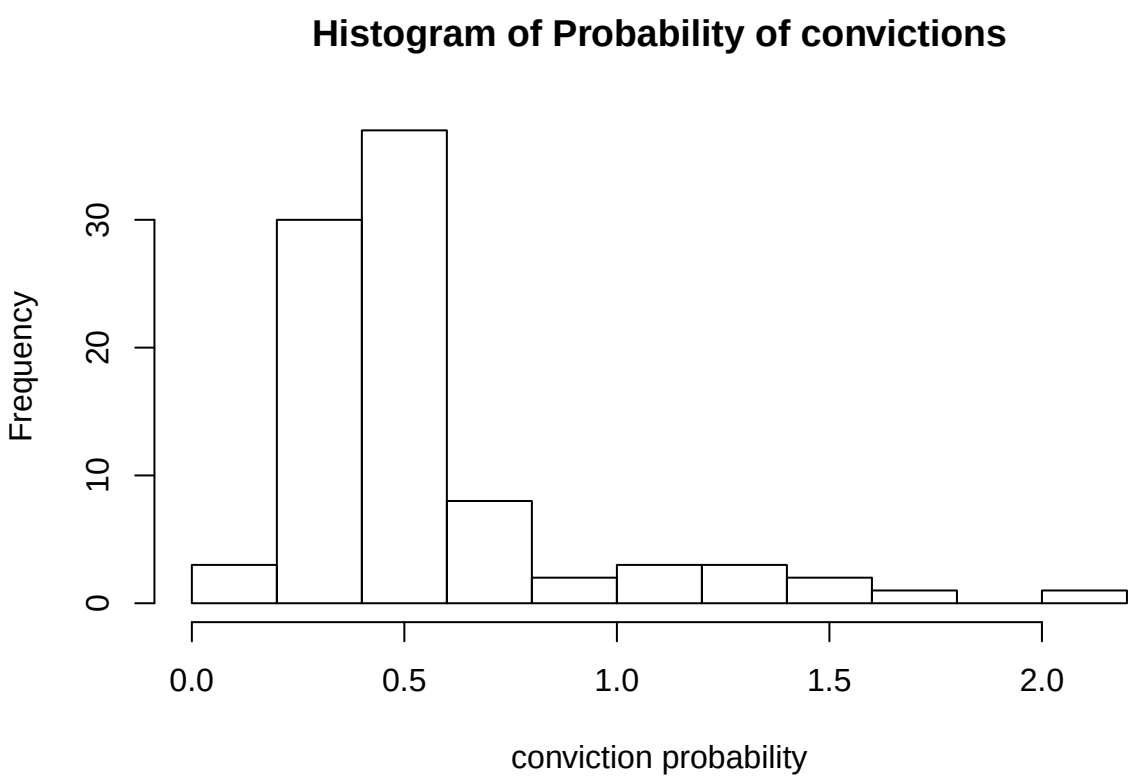

```
summary(crime_data$prbarr)
```

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

The plot looks fairly normal with some values showing above 1.0 which seems odd for a probability statistic. We likely have to pay attention to this variable in our analysis later.

Probability of conviction

```
hist(crime_data$prbconv, main="Histogram of Probability of convictions",
     xlab="conviction probability")
```




```
summary(crime_data$prbconv)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34422 0.45170 0.55086 0.58513 2.12121
```

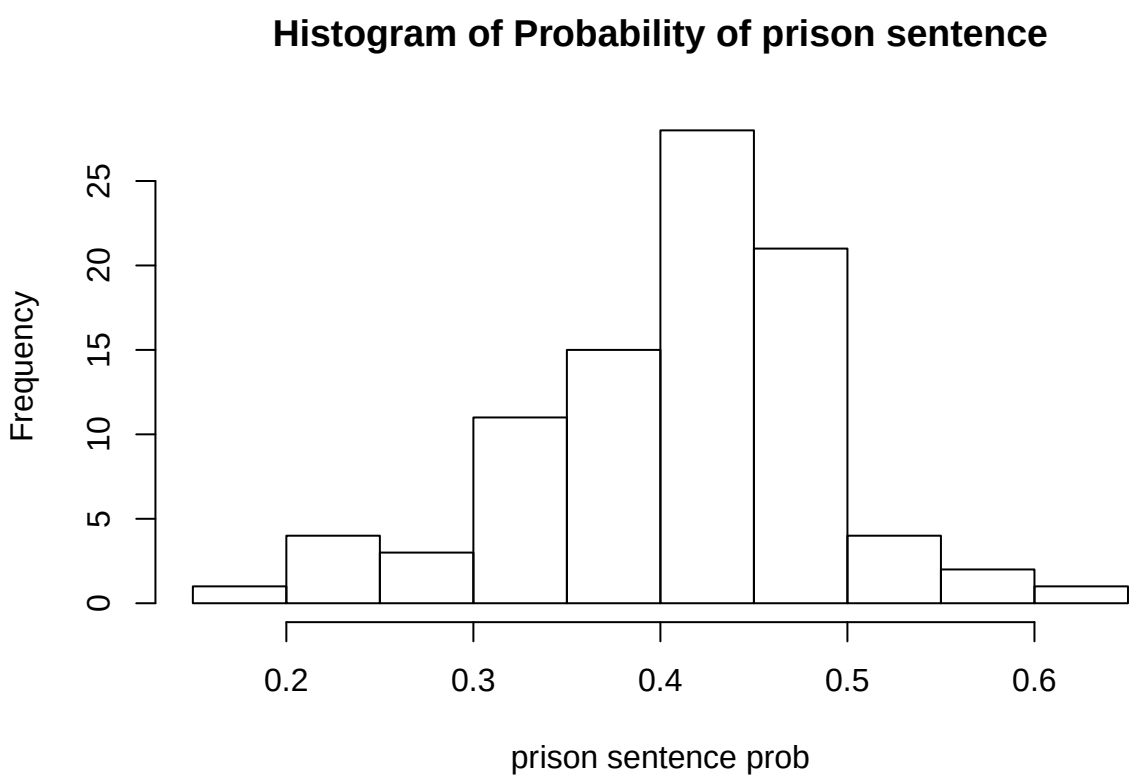
The histogram plot doesn't look normal with more positive/left skew observed in the data. Moreover, plenty of values appear above 1 which again seems odd considering this is a probability that is supposed to be between 0 and 1.

Instead of excluding all values above 1, we likely have to assume these higher values denote high changes of getting convicted in our analysis. From the definition of these terms provided, probability of arrest is proxied by the ratio of of the arrests to offenses. And probability of conviction is proxied by ration of convictions to arrests, probability of prison sentence is proxied by convictions resulting in prison to total convictions. Since these are all "ratios" and not true probabilities, we decided to not exclude and remove these values from the dataset.

Taking the log transformation of this statistics doesn't make much sense either.

Probability of prison sentence

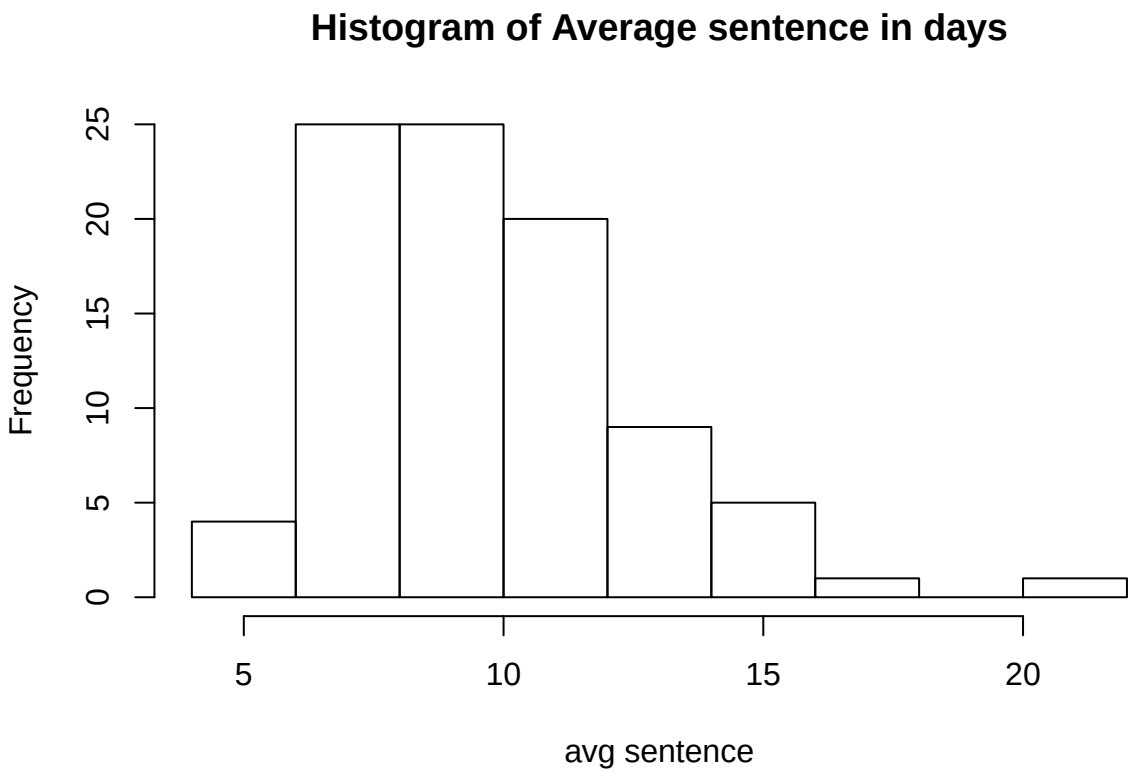
```
hist(crime_data$prbpris, main = "Histogram of Probability of prison sentence",
     xlab="prison sentence prob")
```



This histogram plot looks fairly normal and we don't observe any weird values.

Average sentence days

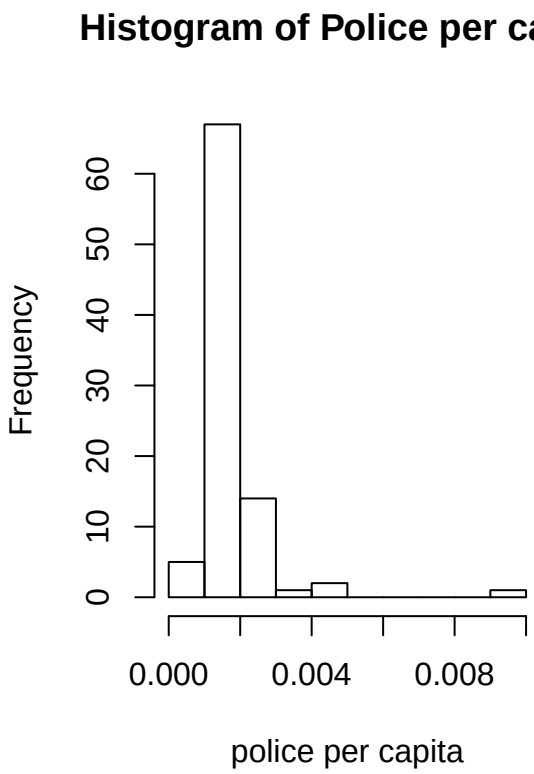
```
hist(crime_data$avgsen, main="Histogram of Average sentence in days",
     xlab="avg sentence")
```

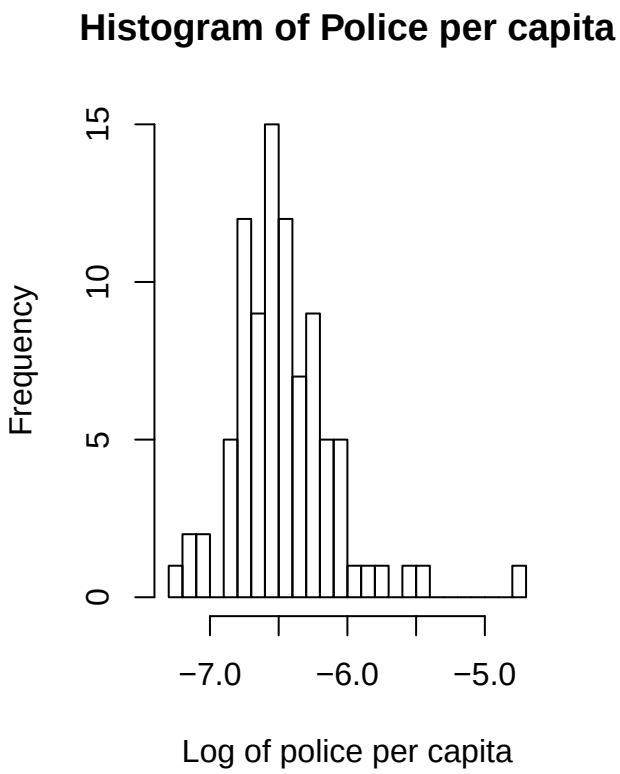


The average sentence in days looks slightly positive skewed. There appears to be an outlier with some values appearing greater than 20 days.

Police per Capita

```
par(mfrow=c(1,2))
hist(crime_data$polpc, main="Histogram of Police per capita",
     xlab="police per capita")
hist(log(crime_data$polpc), main="Histogram of Police per capita", breaks=20,
     xlab="Log of police per capita")
```





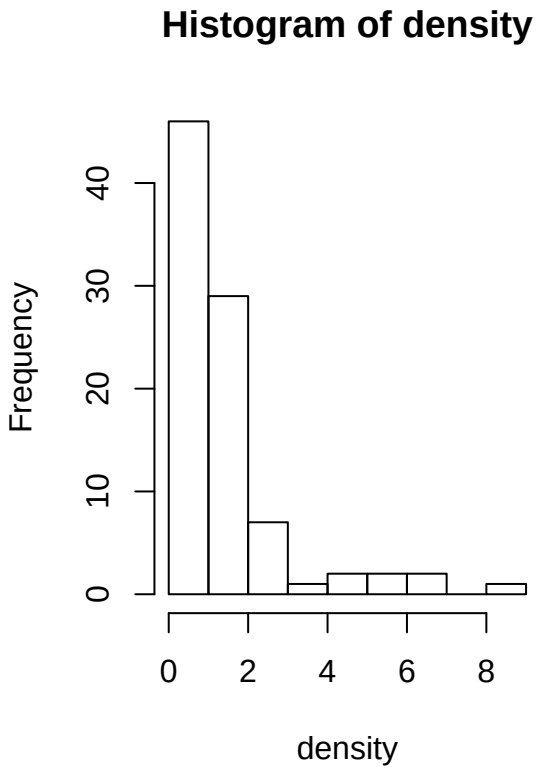

```
summary(crime_data$polpc)
```

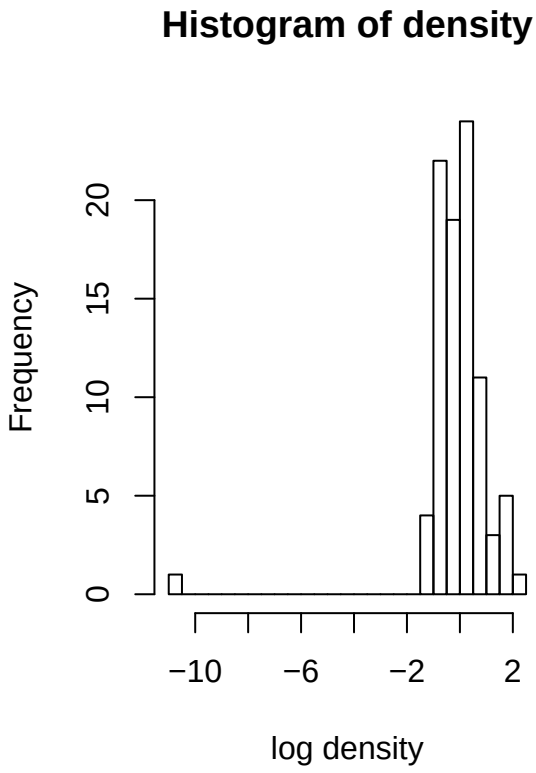
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012378 0.0014897 0.0017080 0.0018856 0.0090543
```

The histogram of police per capita appears to be postively skewed with some outlier on the far right closer to 0.01. Taking the log of the metric makes the plot look more normal. The log transformation will likely be useful to examine the effects of police presence on crime.

People per sq. mile

```
par(mfrow=c(1,2))
hist(crime_data$density, main="Histogram of density",
     xlab="density")
hist(log(crime_data$density), main="Histogram of density", breaks=20,
     xlab="log density")
```






```
summary(crime_data$density)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.54718 0.97925 1.43567 1.56926 8.82765
```

```
crime_data = filter(crime_data, density>0.01)
```

The histogram of density shows lot of positive skew. The log transformation shows a more promising normal distribution whereas it is skewed more towards the right due to a min value that seems out of place. We will continue to use the log value but pay heed to the min value for anamolies.

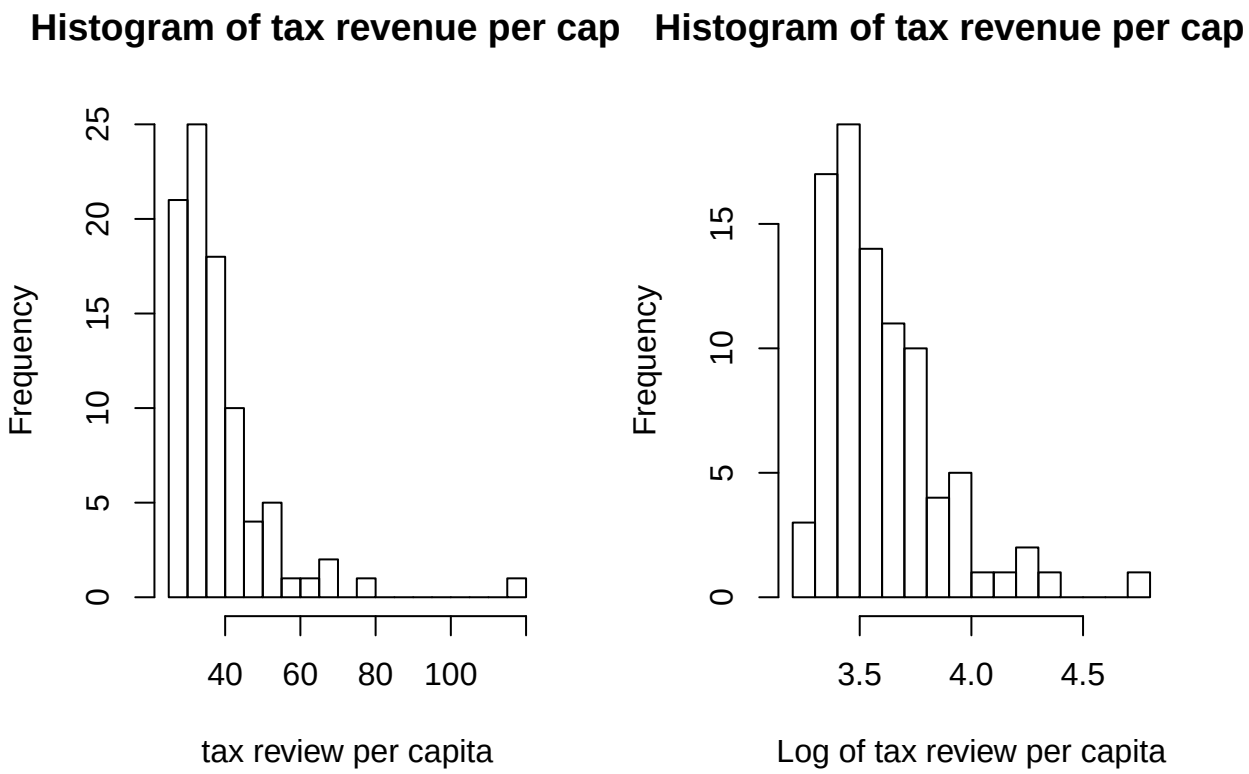
We also observe one extreme outlier value of 0.000020342. This seems like an unlikely value for people per square mile and we are thus removing it from consideration.

Tax revenue per capita

```
summary(crime_data$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.70   34.87   38.17   41.07  119.76
```

```
par(mfrow=c(1,2))
hist(crime_data$taxpc, main="Histogram of tax revenue per capita", breaks=20,
     xlab="tax review per capita")
hist(log(crime_data$taxpc), main="Histogram of tax revenue per capita",
     breaks=20,
     xlab="Log of tax review per capita")
```



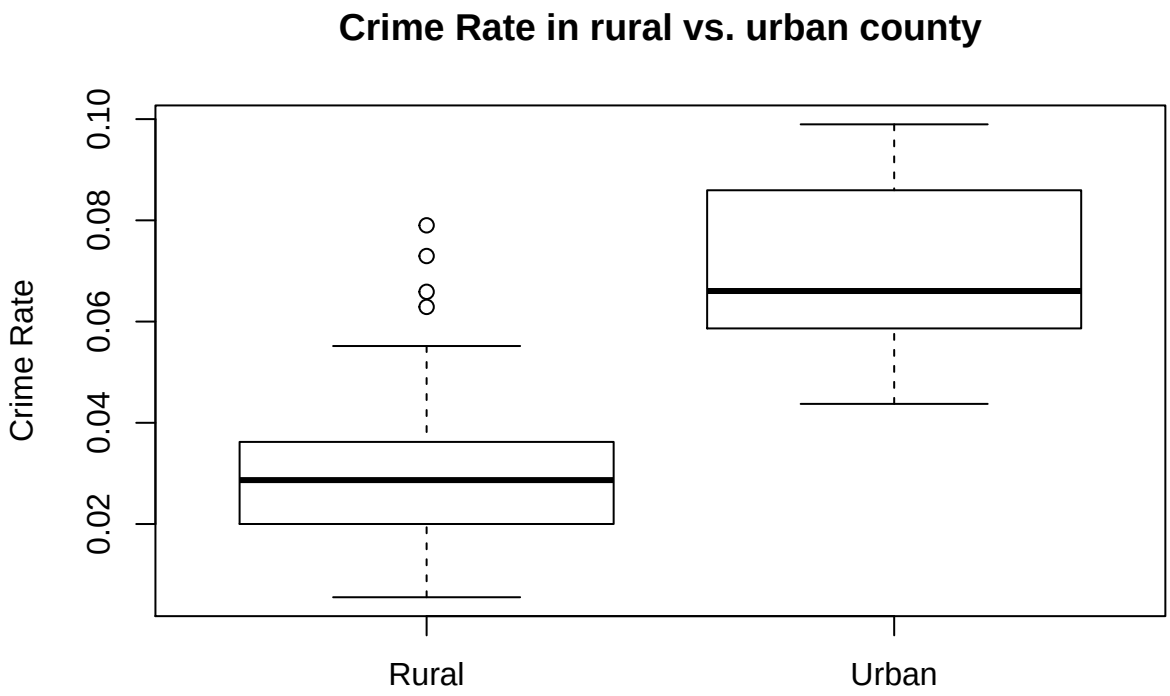
The tax revenue summary indicates a max value of 120, which might be a outlier assuming max of 100. The histogram is sligtly positively skewed on the left and the log transformation appears to be normal yet still retaining small skew towards the left

Urban population

```
sum(crime_data$urban == 1)
```

```
## [1] 8
```

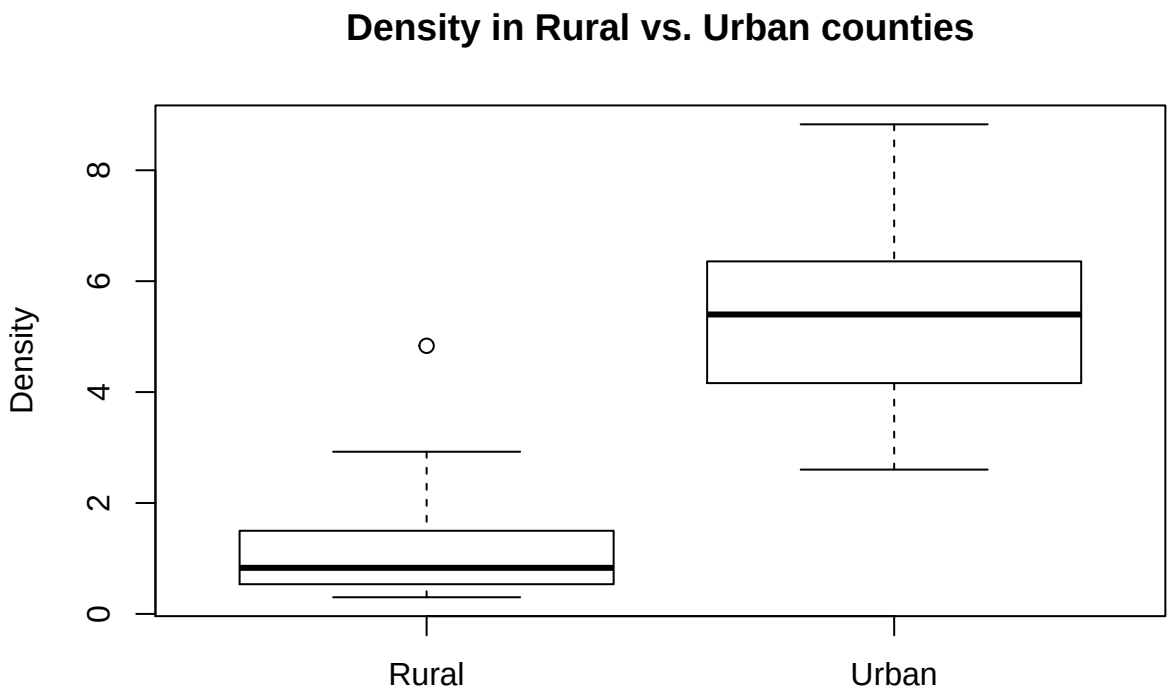
```
boxplot(crime_data$crmrte ~ crime_data$urban, ylab="Crime Rate",
        main="Crime Rate in rural vs. urban county",
        names=c("Rural", "Urban"))
```



There appears to be only 8 counties that are classified as urban in NC

When we compare the crime rates between rural and urban centers, it appears about 3 times higher in urban counties than rural counties. However there are only 8 data points for urban counties, so this might not be enough to devote serious consideration into this variable.

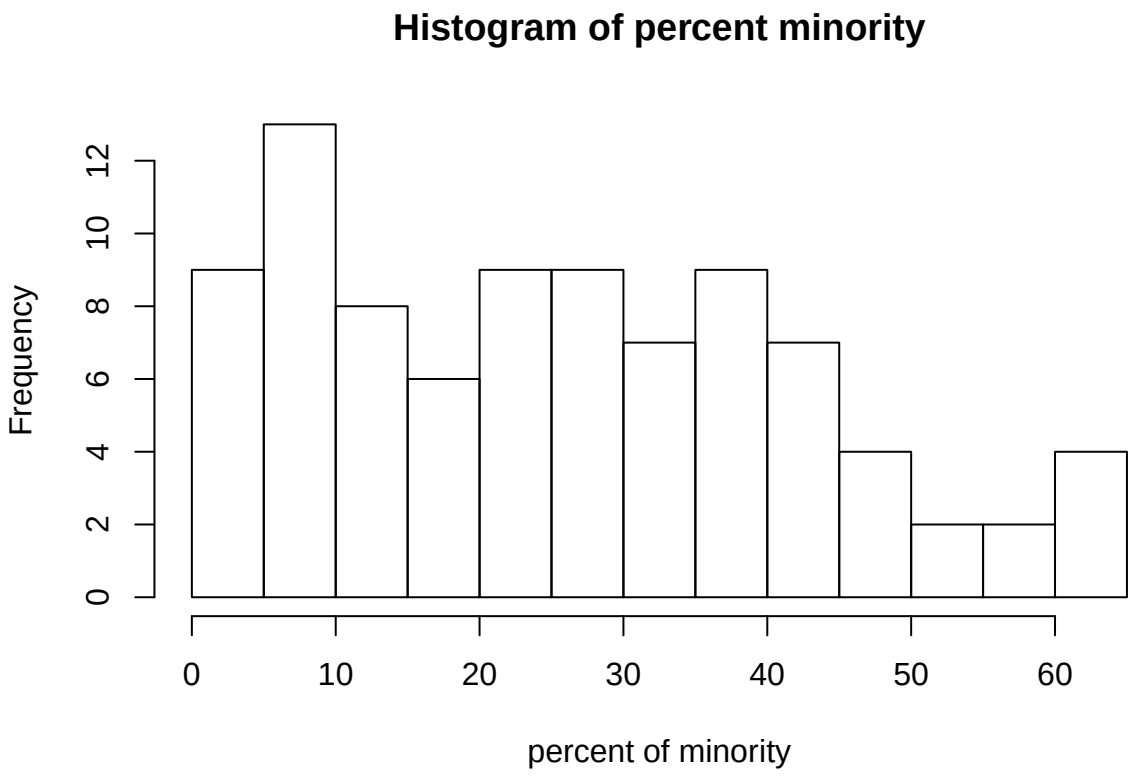
```
boxplot(crime_data$density ~ crime_data$urban, ylab="Density",
        main="Density in Rural vs. Urban counties",
        names=(c("Rural", "Urban")))
```



As expected the density of population is higher in urband couties as compared to rural counties.

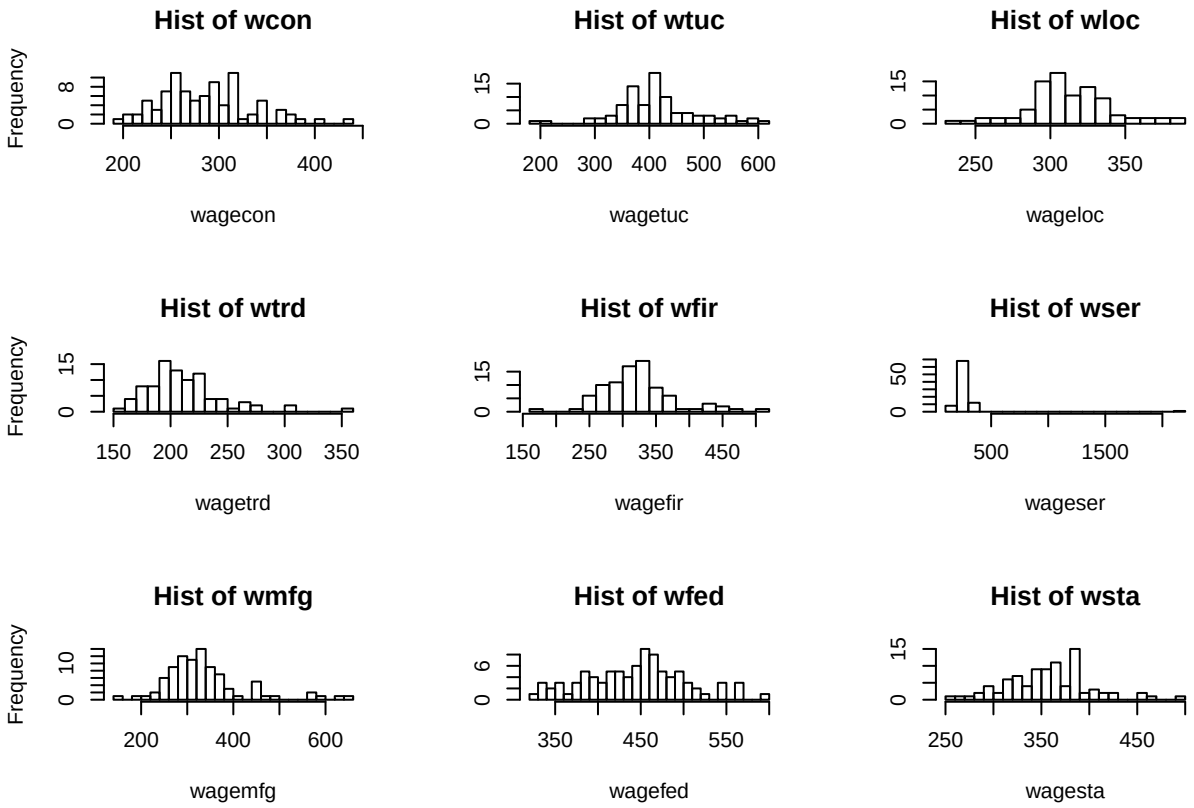
Percent minority

```
hist(crime_data$pctmin80, main="Histogram of percent minority", breaks=20,
     xlab="percent of minority")
```



There doesn't seem to be anything odd about the percent of minority as calculated in 1980. Data and plot seems as expected. ### Wage distribution

```
par(mfrow=(c(3,3)))
hist(crime_data$wcon, breaks=20,
     main="Hist of wcon",ylab="Frequency", xlab="wagecon")
hist(crime_data$wtuc, breaks=20,
     main="Hist of wtuc",xlab="wagetuc", ylab="")
hist(crime_data$wloc, breaks=20,
     main="Hist of wloc",xlab="wageloc", ylab="")
hist(crime_data$wtrd, breaks=20,
     main="Hist of wtrd",ylab="Frequency", xlab="wagetrd")
hist(crime_data$wfir, breaks=20,
     main="Hist of wfir",xlab="wagefir", ylab="")
hist(crime_data$wser, breaks=20,
     main="Hist of wser",xlab="wageser", ylab="")
hist(crime_data$wmfg, breaks=20,
     main="Hist of wmfg",ylab="Frequency", xlab="wagemfg")
hist(crime_data$wfed, breaks=20,
     main="Hist of wfed",xlab="wagefed", ylab="")
hist(crime_data$wsta, breaks=20,
     main="Hist of wsta",xlab="wagesta", ylab="")
```



Most of the wage variables conform to normal distributions.

```
summary(crime_data$wser)
```

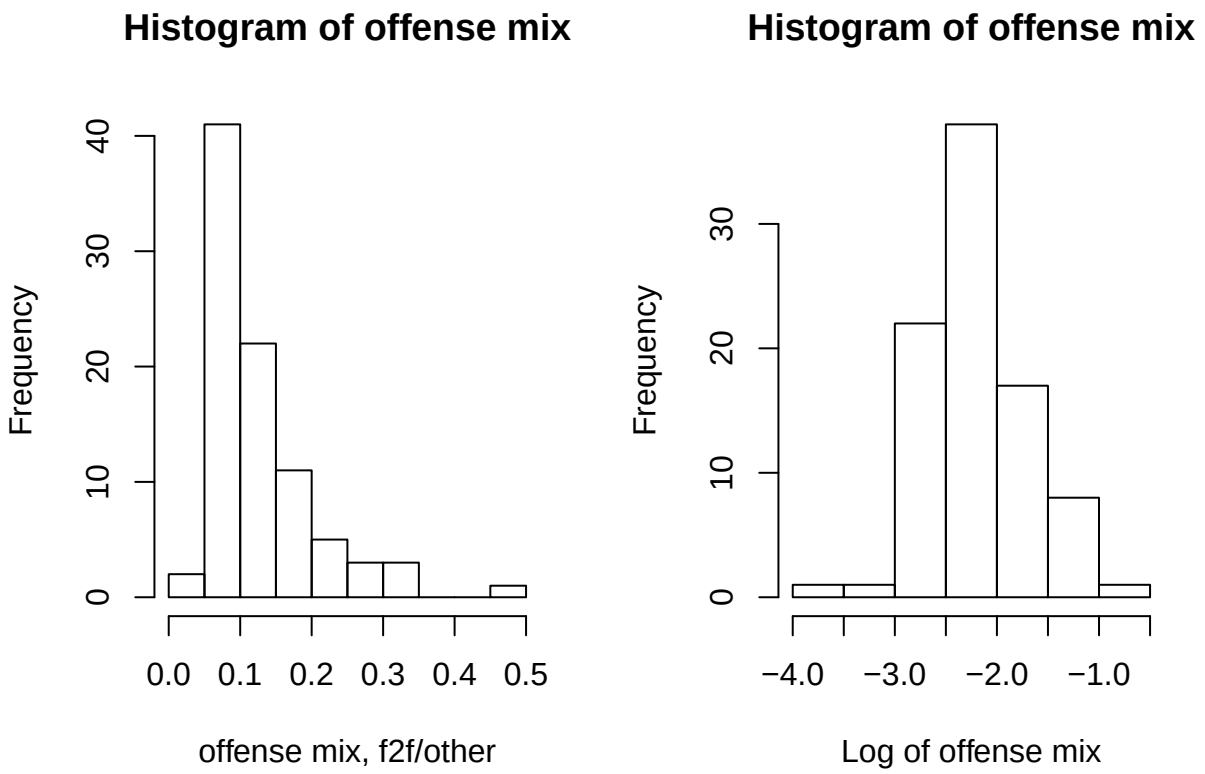
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   230.3   253.2   276.1   278.1  2177.1
```

```
crime_data= filter(crime_data, wser<2177)
```

Wage in service industry does seem to have one strange outlier that is causing some skewness in the plot. We believe this is likely an error and are thus removing it from consideration.

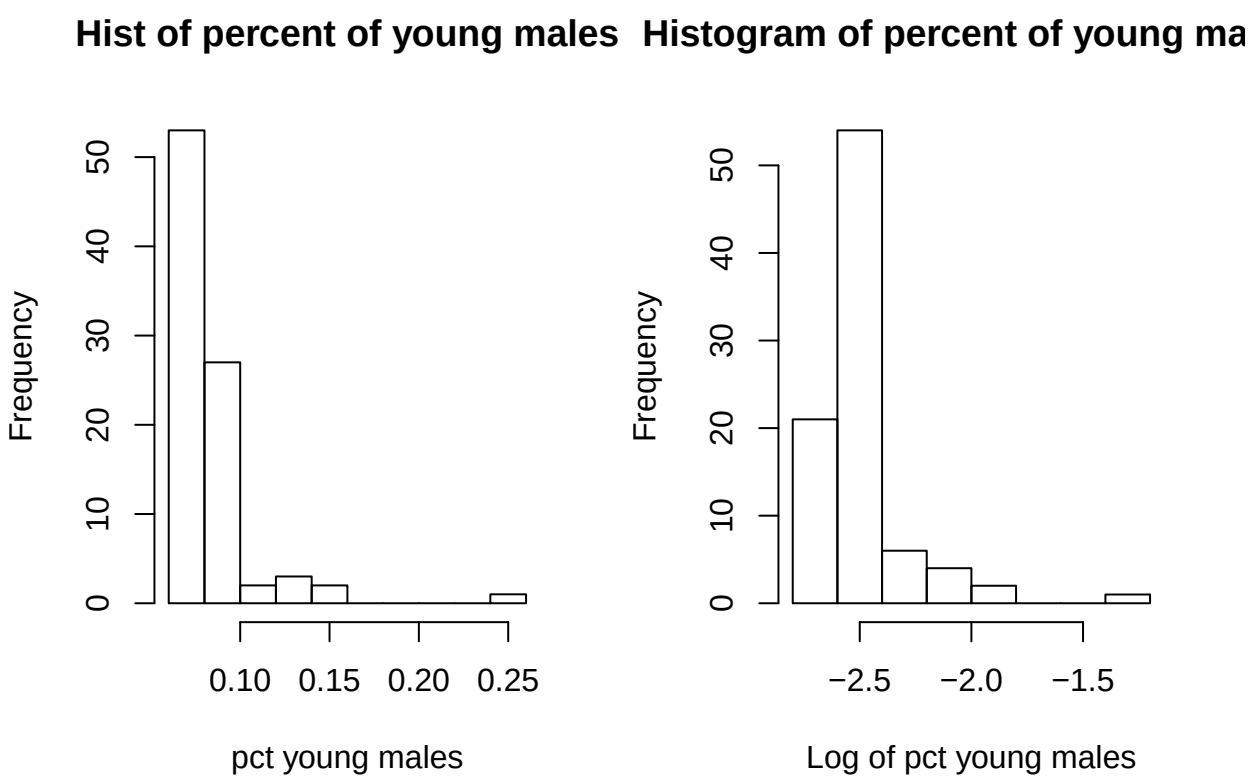
Offense Mix & Percent of young males

```
par(mfrow=c(1,2))
hist(crime_data$mix, main="Histogram of offense mix",
     xlab="offense mix, f2f/other")
hist(log(crime_data$mix), main="Histogram of offense mix",
     xlab="Log of offense mix")
```



Log transformation of offense mix is more normal while the percent of young males has a heavy left positive skew regardless of the log transformation

```
par(mfrow=c(1,2))
hist(crime_data$pctymle, main="Hist of percent of young males",
     xlab="pct young males")
hist(log(crime_data$pctymle), main="Histogram of percent of young males",
     xlab="Log of pct young males")
```



Data Transformation

Based on the univariate analysis performed above, we can opt to take the following transformations of the variables to make better analysis and judgement calls:

Log transformation of the Crime Rate, Police per Capita, Density per sq. mile, Tax revenue per capita And finally scaling the percent (percent young male) and probabilities (arrest, conviction and prison sentence) to be between 0-100

```
crime_data$log_crmrte = log(crime_data$crmrte)
crime_data$log_density = log(crime_data$density)
crime_data$log_polpc = log(crime_data$polpc)
crime_data$log_taxpc = log(crime_data$taxpc)
crime_data$adj_pctymle = crime_data$pctymle *100
crime_data$adj_prbarr = crime_data$prbarr *100
crime_data$adj_prbconv = crime_data$prbconv *100
crime_data$adj_prbpris = crime_data$prbpris *100
```

A final sumary table of our dataset with all transformation and data cleansing performed is displayed below:

```
stargazer(crime_data, title = "Descriptive Statistics", digits=1)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 28, 2018 - 10:54:28 PM

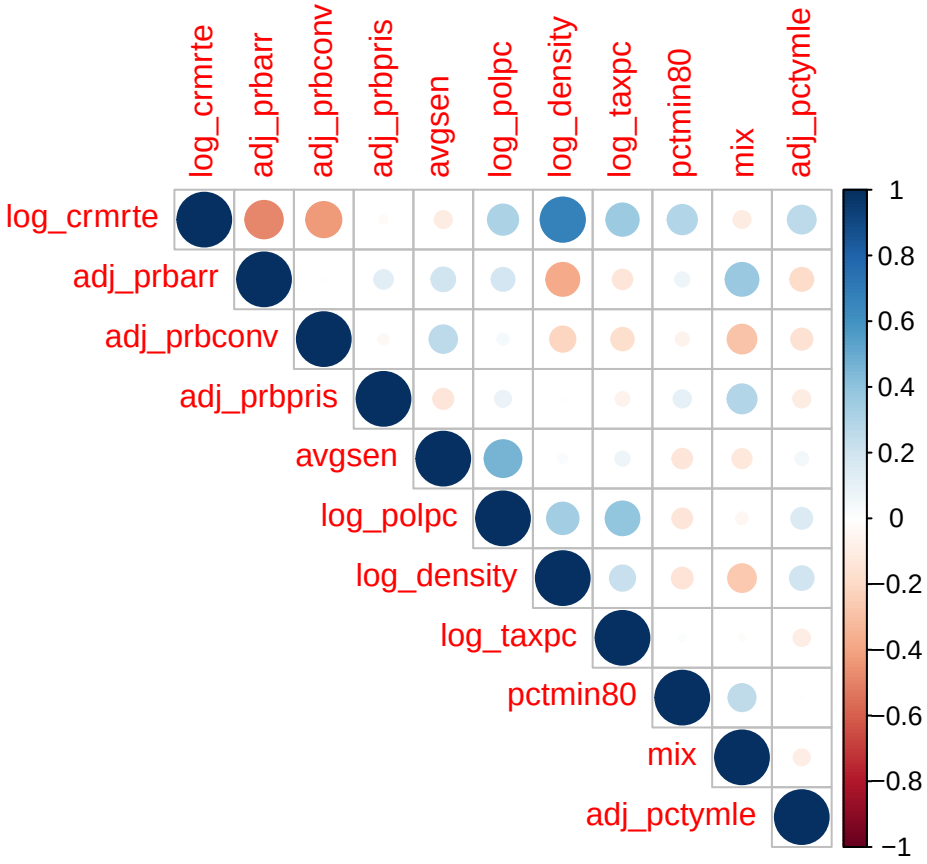
Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
county	88	98.8	57.7	1	197
year	88	87.0	0.0	87	87
crmrte	88	0.03	0.02	0.01	0.1
prbarr	88	0.3	0.1	0.1	1.1
prbconv	88	0.5	0.3	0.1	1.7
prbpris	88	0.4	0.1	0.2	0.6
avgsen	88	9.8	2.8	5.4	20.7
polpc	88	0.002	0.001	0.001	0.01
density	88	1.5	1.5	0.3	8.8
taxpc	88	38.1	13.3	25.7	119.8
west	88	0.2	0.4	0	1
central	88	0.4	0.5	0	1
urban	88	0.1	0.3	0	1
pctmin80	88	25.3	16.7	1.3	61.9
wcon	88	286.6	47.5	193.6	436.8
wtuc	88	414.0	74.8	187.6	613.2
wtrd	88	211.8	33.7	154.2	354.7
wfir	88	322.9	53.9	170.9	509.5
wser	88	254.5	44.0	133.0	391.3
wmfg	88	338.7	87.4	157.4	646.8
wfed	88	444.5	59.1	326.1	598.0
wsta	88	357.0	43.3	258.3	499.6
wloc	88	312.5	28.4	239.2	388.1
mix	88	0.1	0.1	0.02	0.5
pctymle	88	0.1	0.02	0.1	0.2
log_crmrte	88	−3.5	0.5	−5.2	−2.3
log_density	88	0.05	0.8	−1.2	2.2
log_polpc	88	−6.5	0.4	−7.2	−4.7
log_taxpc	88	3.6	0.3	3.2	4.8
adj_pctymle	88	8.4	2.4	6.2	24.9
adj_prbarr	88	29.4	13.7	9.3	109.1
adj_prbconv	88	53.6	31.5	6.8	167.1
adj_prbpris	88	41.3	7.7	22.7	60.0

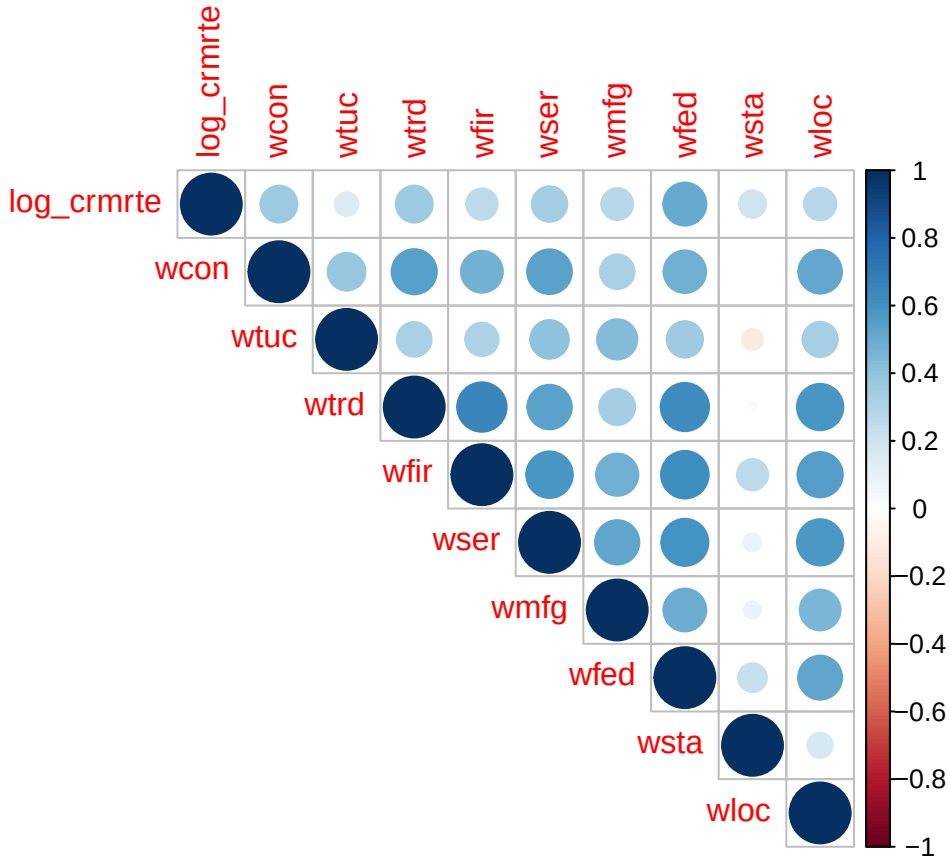
Bi-variate Analysis

The correlation plot between the different variables is as follows:

```
corrplot(cor(crime_data[,
                        c("log_crmrte", "adj_prbarr", "adj_prbconv", "adj_prbpris", "avgsen",
                          "log_polpc", "log_density", "log_taxpc", "pctmin80", "mix",
                          "adj_pctymle")]), type = "upper")
```




```
corrplot(cor(crime_data[,
                        c("log_crmrte", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed",
                          "wsta", "wloc")]), type = "upper")
```



We can see that there is a high positive correlation between:

- log of crime rate vs. log of policy per capita, log of tax revenue per capita, log of density and percent young male
- log of crime rate vs. most of the wage variables

And there is a high negative correlation between:

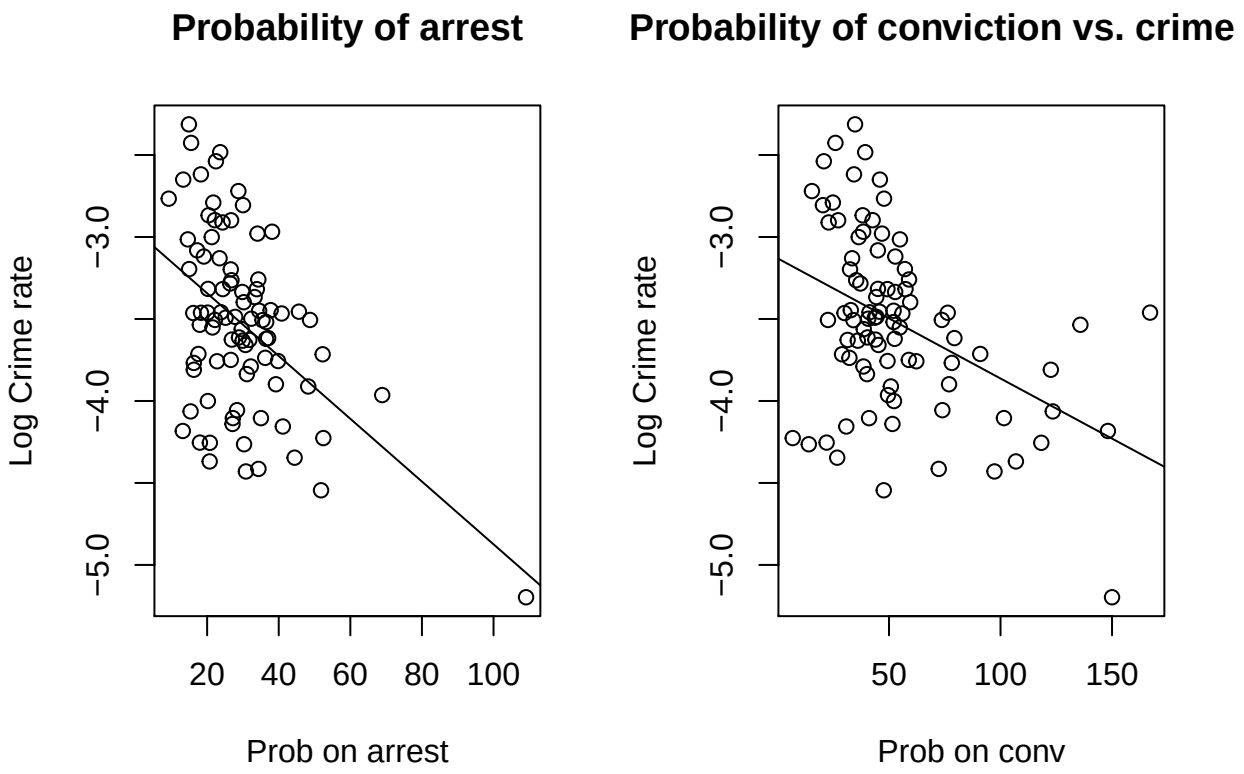
- log of crime rate vs. probability of arrests and conviction

The positive correlation observed makes sense for the following reasons:

1) More densely populated regions tends to observe more crimes
2) More wealthy areas (more wages and taxes) tend to have more crimes
3) More crimes leads to more police presence in a particular county to monitor and reduce crime rate

The negative correlations can be further observed using:

```
par(mfrow=c(1,2))
plot(crime_data$adj_prbarr, crime_data$log_crmrte,
     main="Probability of arrest", ylab="Log Crime rate", xlab="Prob on arrest")
abline(lm(crime_data$log_crmrte ~ crime_data$adj_prbarr))
plot(crime_data$adj_prbconv, crime_data$log_crmrte,
     main="Probability of conviction vs. crime rate", ylab="Log Crime rate",
     xlab="Prob on conv")
abline(lm(crime_data$log_crmrte ~ crime_data$adj_prbconv))
```



As seen above, as the probability of arrests and conviction go down, there are more criminals on the loose which leads to higher crime rates observed

TODO: Talk about other possible correlations here?

TODO: Discuss other interesting bi-variate analysis?

3. Model Specification and Assumptions

In our earlier analysis, we observed some key relationships between crime rate and other variables presented. Some of these variables had high positive correlation to crime rate while some others exhibited strong negative correlation.

For our first simple model, we will choose a subset of these variables that we believe are most important determinants of crime rate.

Model 1

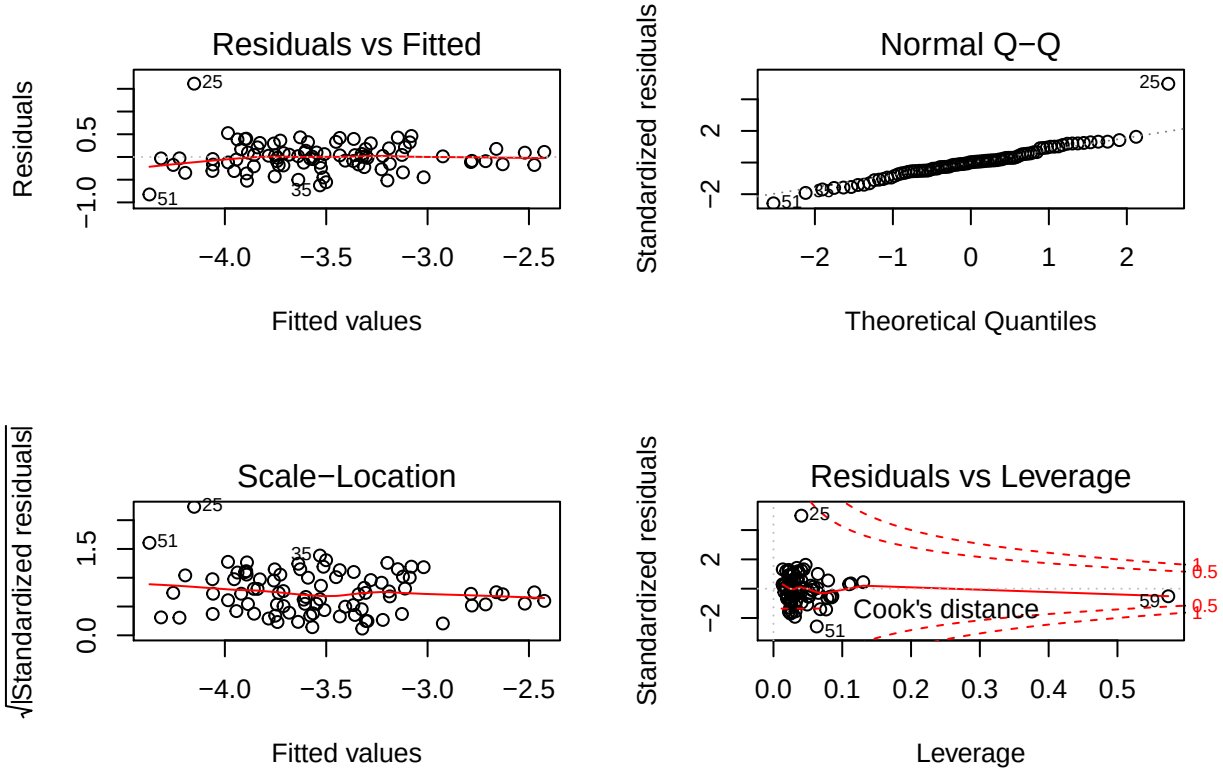
$$log(CrimeRate) = \beta_0 + \beta_1 log(Density) + \beta_2(YoungMale) + \beta_3(Minority) + u$$

It is common knowledge that areas with higher density have more crime. Therefore we include that factor in our model. Similarly we hypothesized that crime rate is high among minority and young male population, so we round off our model with that factored in as well.

```
model1 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80, data=crime_data)
model1$coefficients
```

```
## (Intercept) log_density     pctymle    pctmin80
## -4.12804608  0.50214515  3.00053687  0.01304853
```

```
par(mfrow=c(2,2))
plot(model1)
```




```
AIC(model1)
```

```
## [1] 60.87321
```

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

```
## [1] 0.6366295
```

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

```
## [1] 0.6236519
```

Model 2

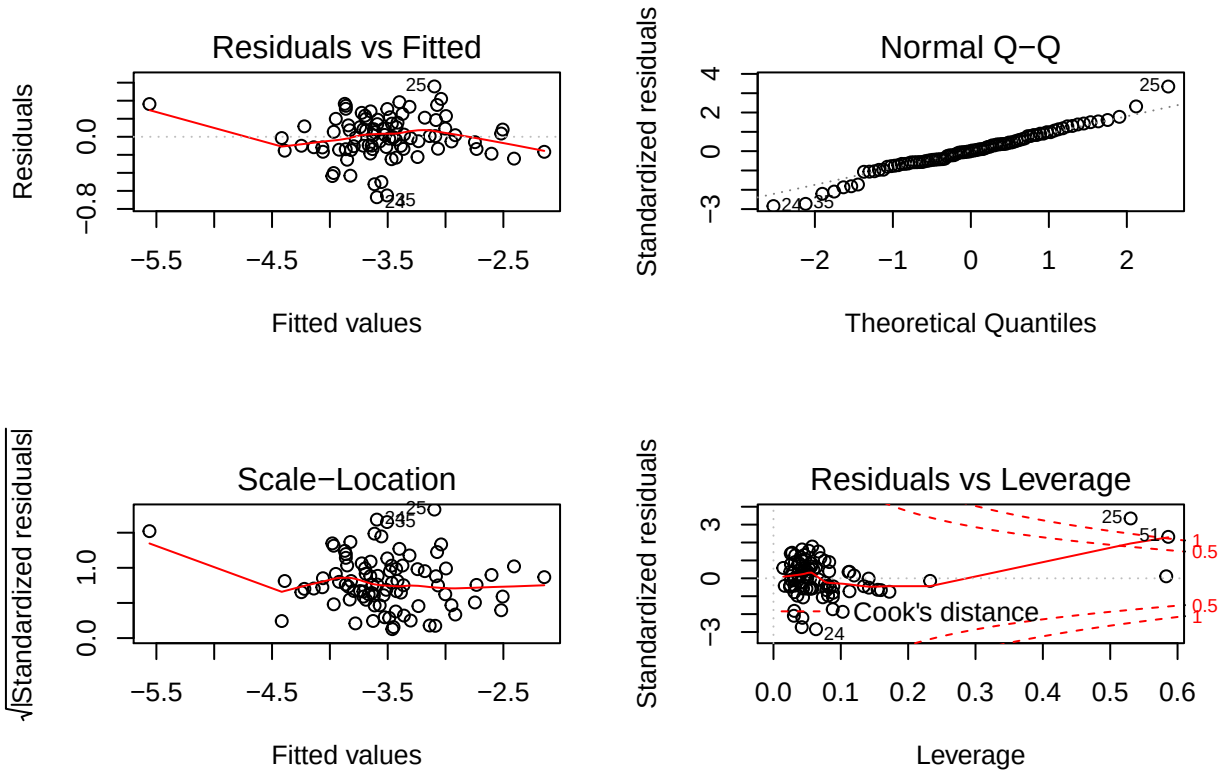
high probability of arrests and conviction act as deterrents to crime.

$$log(CrimeRate) = \beta_0 + \beta_1 log(Density) + \beta_2(YoungMale) + \beta_3(Minority) + \beta_4(Conviction) + \beta_5(Arrest) + \beta_6(Tax) + u$$

```
model2 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80+adj_prbarr+
            adj_prbconv+taxpc, data=crime_data)
model2$coefficients
```

```
##  (Intercept)  log_density      pctymle     pctmin80   adj_prbarr
## -3.854199245  0.364862877  2.332928041  0.012490294 -0.010725276
##  adj_prbconv        taxpc
## -0.004169832  0.008945281
```

```
par(mfrow=c(2,2))
plot(model2)
```




```
AIC(model2)
```

```
## [1] 9.814178
```

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

```
## [1] 0.8099998
```

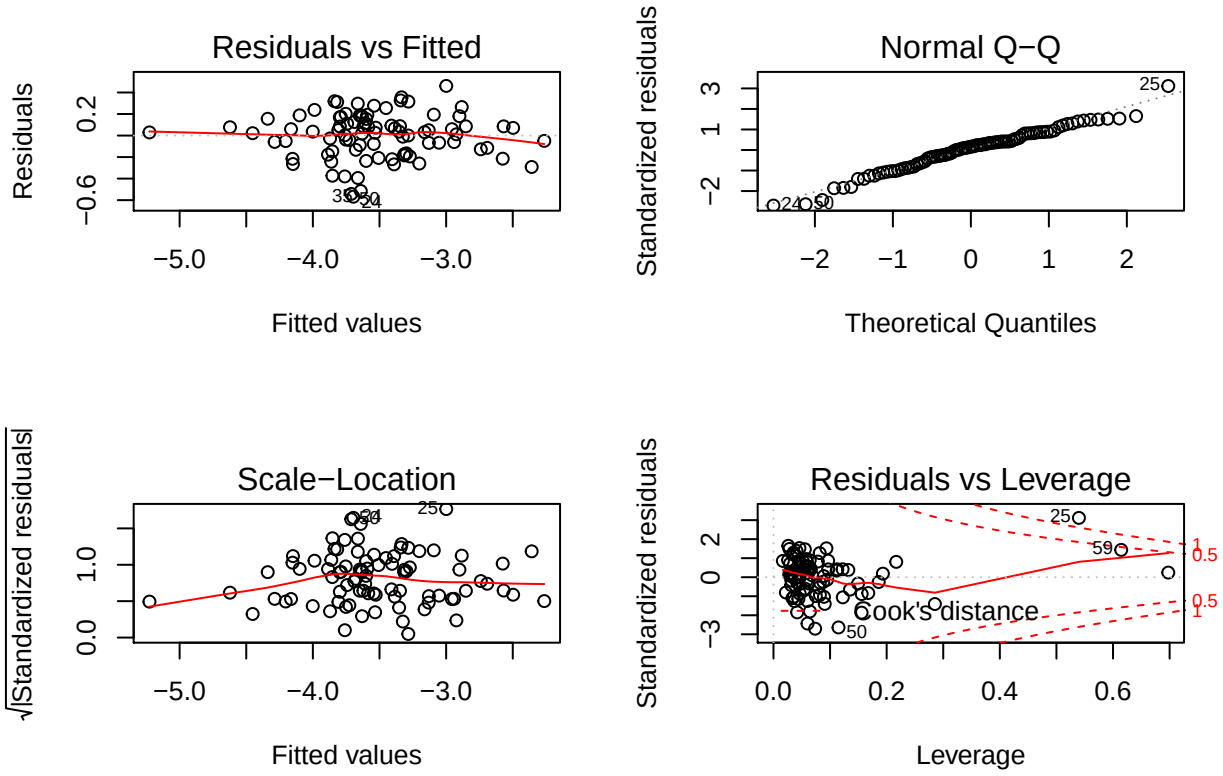
Model 3

everything

```
model3 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80+adj_prbarr+adj_prbconv+
              taxpc+log_polpc, data=crime_data)
model3$coefficients
```

```
##  (Intercept)  log_density      pctymle     pctmin80   adj_prbarr
## -0.788863315  0.283453739  0.917955741  0.013135054 -0.015585820
##  adj_prbconv        taxpc    log_polpc
## -0.005301213  0.003968981  0.396915211
```

```
par(mfrow=c(2,2))
plot(model3)
```




```
AIC(model3)
```

```
## [1] -8.691449
```

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

```
## [1] 0.8494935
```

```
stargazer(model1, model2, model3)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 28, 2018 - 10:54:32 PM

5. Discussion of omitted variables (Identify what you think are the 5-10 most important omitted variables that bias results you care about.)

Education

Unemployment

Poverty

Table 2:

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log_density	0.502^{***}	0.365^{***}	0.283^{***}
	(0.048)	(0.039)	(0.039)
pctymle	3.001^{*}	2.333^{**}	0.918
	(1.529)	(1.159)	(1.083)
pctmin80	0.013^{***}	0.012^{***}	0.013^{***}
	(0.002)	(0.002)	(0.001)
adj_prbarr		-0.011^{***}	-0.016^{***}
		(0.002)	(0.002)
adj_prbconv		-0.004^{***}	-0.005^{***}
		(0.001)	(0.001)
taxpc		0.009^{***}	0.004^{*}
		(0.002)	(0.002)
log_polpc			0.397^{***}
			(0.087)
Constant	-4.128^{***}	-3.854^{***}	-0.789
	(0.143)	(0.181)	(0.688)
Observations	88	88	88
R^2	0.637	0.810	0.849
Adjusted R^2	0.624	0.796	0.836
Residual Std. Error	0.331 (df = 84)	0.243 (df = 81)	0.218 (df = 80)
F Statistic	49.056^{***} (df = 3; 84)	57.553^{***} (df = 6; 81)	64.505^{***} (df = 7; 80)
Note:			*p<0.1; **p<0.05; ***p<0.01