Lab 3 (Bhuvnesh Sharma, Weixin Wu)

Bhuvnesh Sharma, Weixin Wu

March 22, 2018

Introduction

Crime is huge menace in the society, there have been many attempts in past to reduce crime rates within communities in North Carolina. Traditional politicians and conventional approach has assumed that tough on crime is an effective tool to curb crime. Being tough on crime is regularly misunderstood as longer and mandatory prison sentences. This misguided strategy can lead to state's higher investment on prison infrastructure and also make laws which can promote mandatory prison sentences appear as effective crime fighting tool. The goal of this study is to uncover the real facts around the crime rates within North Carolina to develop effective state policy around to reduce crime rates. Key motivation of the report discover the real drivers and instruments which the policy makers can use and have meaningful impact on crime. Study intends to empower the state politicians , key legislative leaders with key facts which have been based on data and not on conventional empirical narratives. Study intends to discover key variables which have major impact on crime rates in North Carolina . This information would be critical for voters to understand so that they can make an informed decision on a important election issue.

Data Cleansing

```
crimeData <- read.csv("crime_v2.csv")
summary(crimeData)
```

```
##      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 shown in the summary table, there are 6 NA's in every variable. After reviewing the data, we found that all NA's are in 6 rows, so we removed those rows as they did not provide any information.

```
crimeData2 <- crimeData[complete.cases(crimeData),]
```

Variable 'prbconv' was incorrectly displayed as a text field. We converted it to numermic.

```
crimeData2 <- transform(crimeData2, prbconv = as.numeric(as.character(prbconv)))
summary(crimeData2$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

Usually the probability variable should be bound between 0 and 1. However, there is one observation with 'prbarr' (probability of arrest) higher than 1, and 10 observations with 'prbconv' (probability of conviction) higher than 1.

```
nrow(crimeData2[which(crimeData2$prbarr>1),])
```

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

```
nrow(crimeData2[which(crimeData2$prbconv>1),])
```

```
## [1] 10
```

Variable 'prbarr' is defined as the ratio of arrests to offenses. One possible explanation for 'prbarr' being greater than 1 is that multiple people who convicted a single crime together is counted as one conviction but multiple arrests.
Variable 'prbconv' is defined as the ratio of convictions to arrests. One possible explanation for 'prbconv' being greater than 1 is that one person who is convicted of multiple crimes but only arrested once.
Without further information on the variables, we could not conclude whether these values are invalid. So we left those observations in the data.

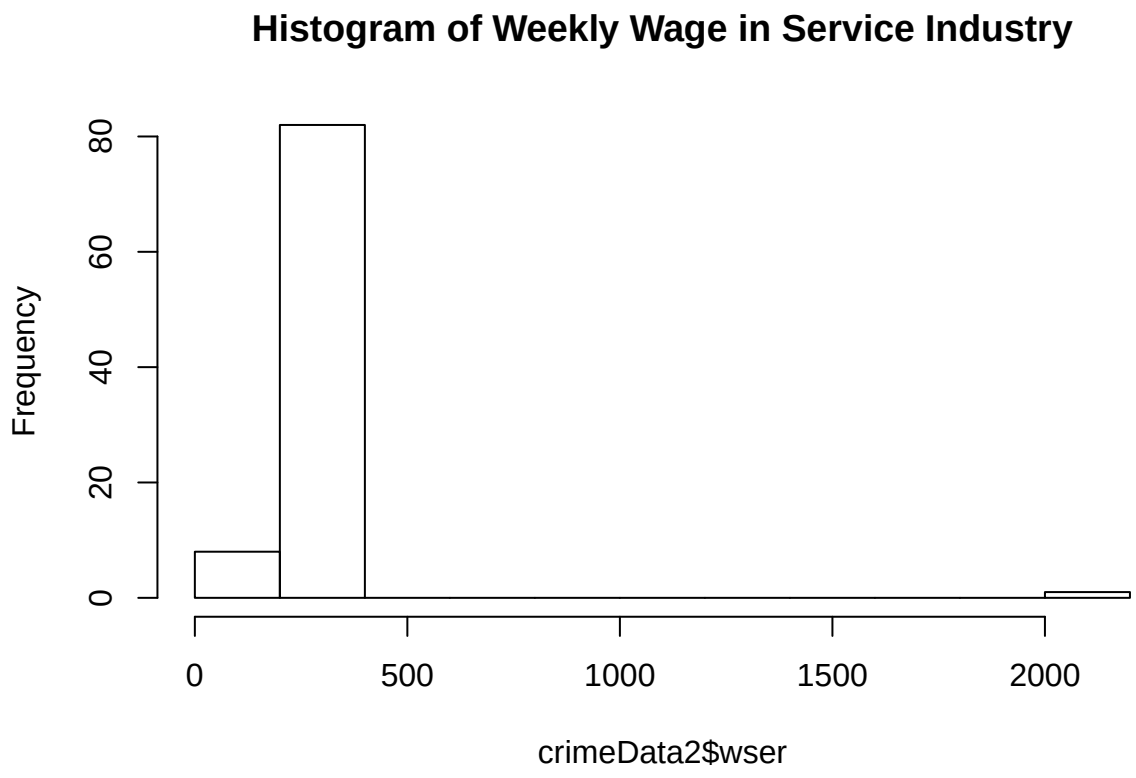
Variable 'pctmin80' (percent of minority in 1980) is expressed as percentages. We converted it into decimals to be consistent with variable 'pctymle' (percent of young male).

```
crimeData2$pctmin80_2 <- crimeData2$pctmin80/100
```

```
###################################################################### / 1000
```

The max value of variable 'wser' (weekly wage of service industry) is significantly higher than its third quartile. The histogram below shows that the max value (2177.068) is significantly higher than the rest of values.

```
hist(crimeData2$wser, main="Histogram of Weekly Wage in Service Industry")
```




```
crimeData2[which(crimeData2$wser>2000),]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##      density    taxpc west central urban pctmin80     wcon    wtuc
## 84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
##       pctymle pctmin80_2
```

```
## 84 0.07008217   0.643482
```

We examined County 185, whose wser is 2177.068. We noticed that most other weekly wage variables for County 185 are below the means. You would expect that a richer county would have weekly wage in multiple industries to be higher than the average. So it's very unlikely for a county to have lower than average weekly wage on constructure, transportation, retail, finance, etc. but extremely high weekly wage on the service industry. In addition, an average weekly wage of 2177.068 in 1987 is an unreasonable value. So we believed 2177.068 is erroneous. We removed this observation from the data.

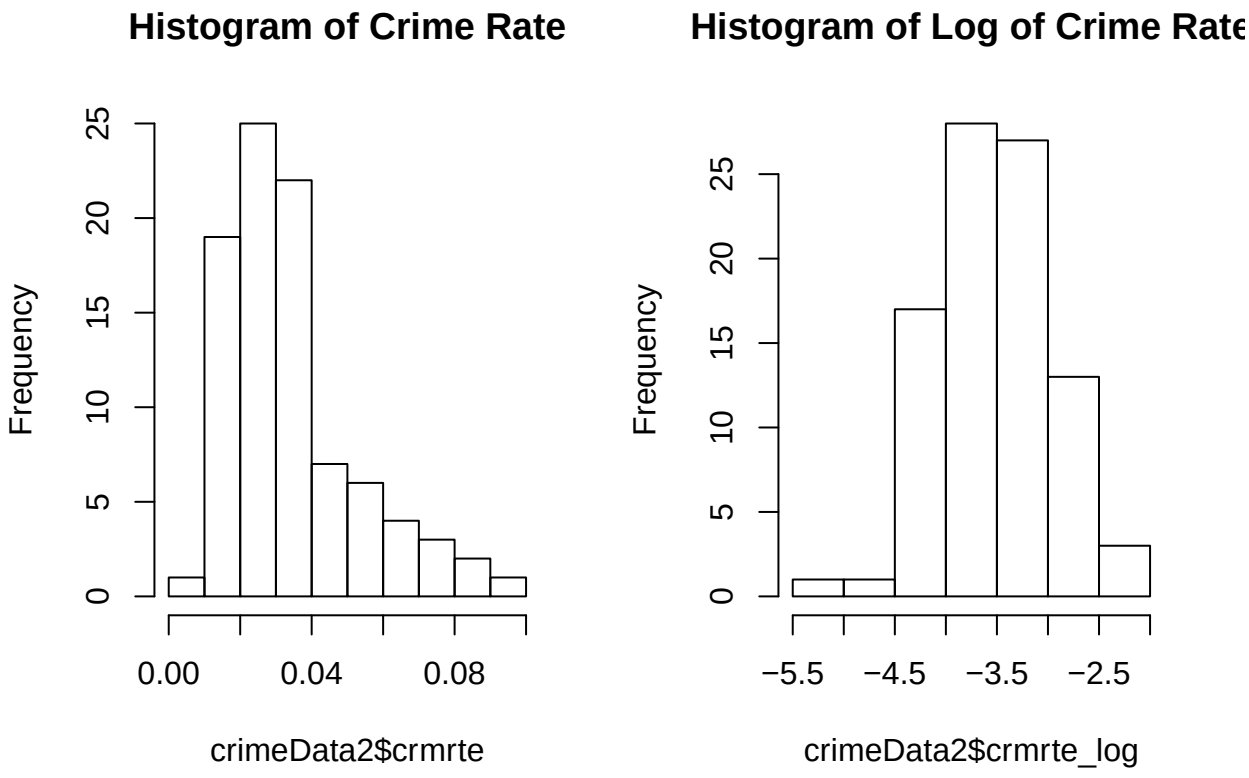
```
crimeData2 <- crimeData2[which(crimeData2$wser<2000),]
```

Exploratory Data Analysis

Crimes committed per person (crmrte)

The distribution of crime rate is skewed to the right, so we considered taking the log of crime rate. After the log transformation, the distribution of crmrte_log is closer to normal. Semilogarithmic form is interpretable later in modeling: it tells us what's the percentage change in crime rate in response to a unit change in explantory variables. Our target variable is crmrte_log.

```
par(mfrow=c(1,2))
hist(crimeData2$crmrte, main="Histogram of Crime Rate")
crimeData2$crmrte_log = log(crimeData2$crmrte)
hist(crimeData2$crmrte_log, main="Histogram of Log of Crime Rate")
```

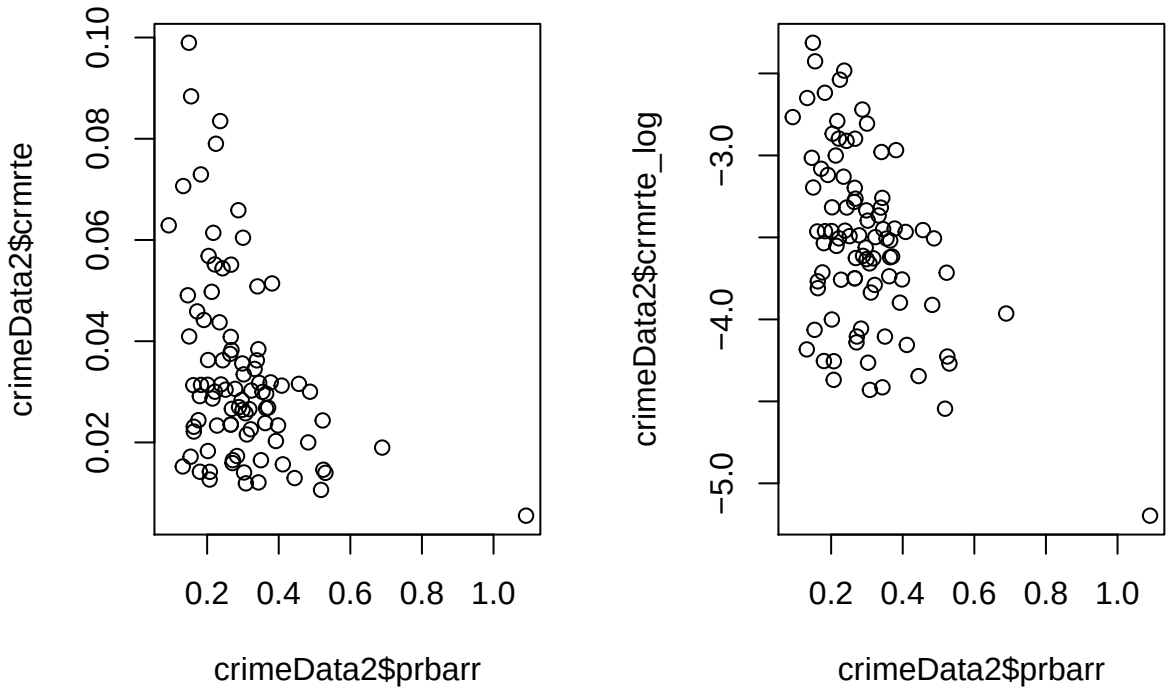


Probability of arrest (prbarr)

The scatter plot of crmrte vs. prbarr on the left shows an exponential decay trend. In addition, the variation of crmrte decreases substantially as prbarr increases. We took the log of crime rate, and then re-graph the scatter plot (shown on the right). The scatter plot of crmrte_log vs. prbarr indicates a more linear relationship and the variation of crmrte_log does not vary as much with prbarr. The correlation coefficient further supports the transformation.

* The correlation between crmrte and prbarr is -0.41
* The correlation between crmrte_log and prbarr is -0.50

```
par(mfrow=c(1,2))
plot(crimeData2$prbarr, crimeData2$crmrte)
plot(crimeData2$prbarr, crimeData2$crmrte_log)
```




```
cor(crimeData2$prbarr, crimeData2$crmrte)
```

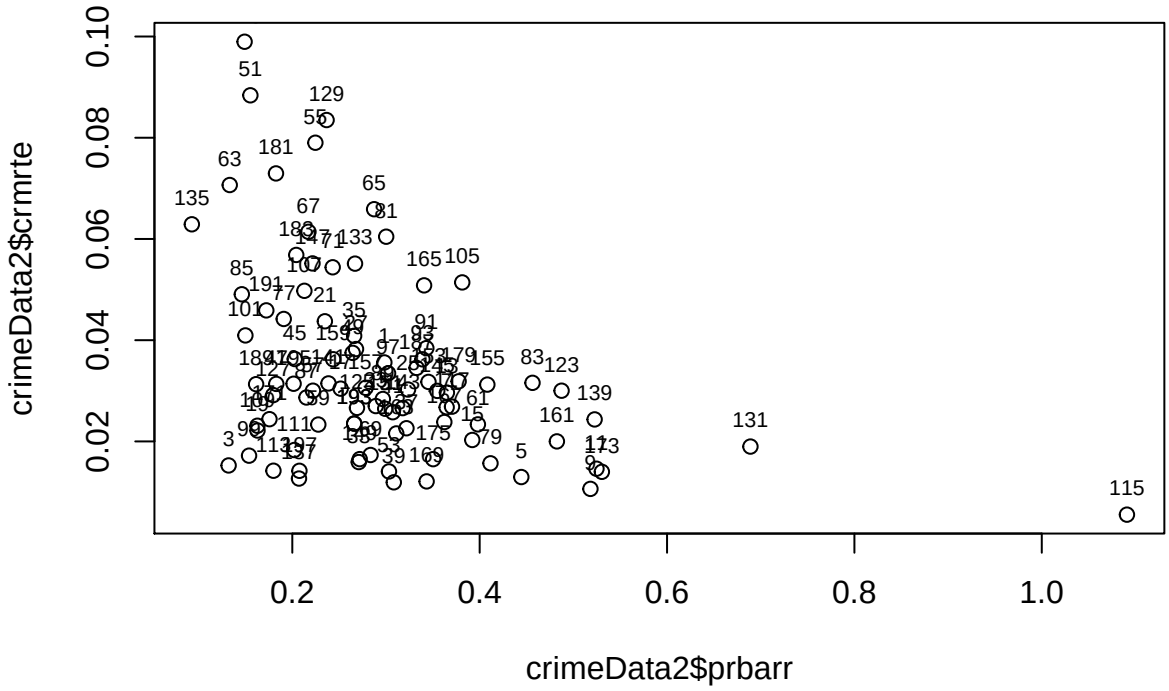
```
## [1] -0.4076239
```

```
cor(crimeData2$prbarr, crimeData2$crmrte_log)
```

```
## [1] -0.4964904
```

In addition, we noticed a leveraged data point in the graph, that's County 115. County 115 has significantly higher probability of arrest than all other counties. If we removed County 115 from the data, the correlation coefficient reduced from -0.50 to -0.39. This indicates that County 115 could be an infludential observation. Later when building the model, we will calculate Cook's distance to confirm that County 15 is an influential observation and also address the impact of influential observations to parameter estimates.

```
plot(crimeData2$prbarr, crimeData2$crmrte)
text(crimeData2$prbarr, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3)
```




```
crimeData3 <- crimeData2[which(crimeData2$county!=115),]
cor(crimeData3$prbarr, crimeData3$crmrte_log)
```

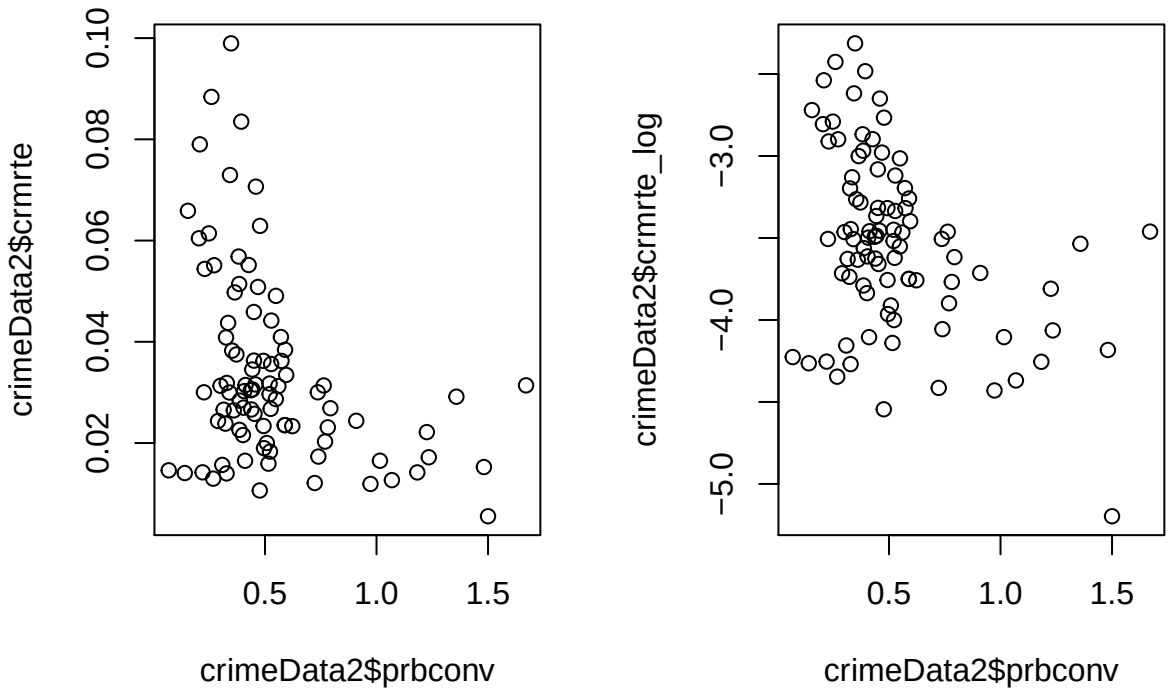
```
## [1] -0.3949839
```

Probability of conviction (prbconv)

Similar to prbarr, the scatter plot of crmrte vs. prbconv on the left shows an exponential decay trend. In addition, the variation of crmrte decreases substantially as prbconv increases. We took the log of crime rate, and then re-graph the scatter plot (shown on the right). The scatter plot of crmrte_log vs. prbconv indicates a more linear relationship and the variation of crmrte_log does not vary as much with prbconv The correlation coefficient further supports the transformation.

* The correlation between crmrte and prbarr is -0.37
* The correlation between crmrte_log and prbarr is -0.41

```
par(mfrow=c(1,2))
plot(crimeData2$prbconv, crimeData2$crmrte)
plot(crimeData2$prbconv, crimeData2$crmrte_log)
```




```
cor(crimeData2$prbconv, crimeData2$crmrte)
```

```
## [1] -0.3728922
```

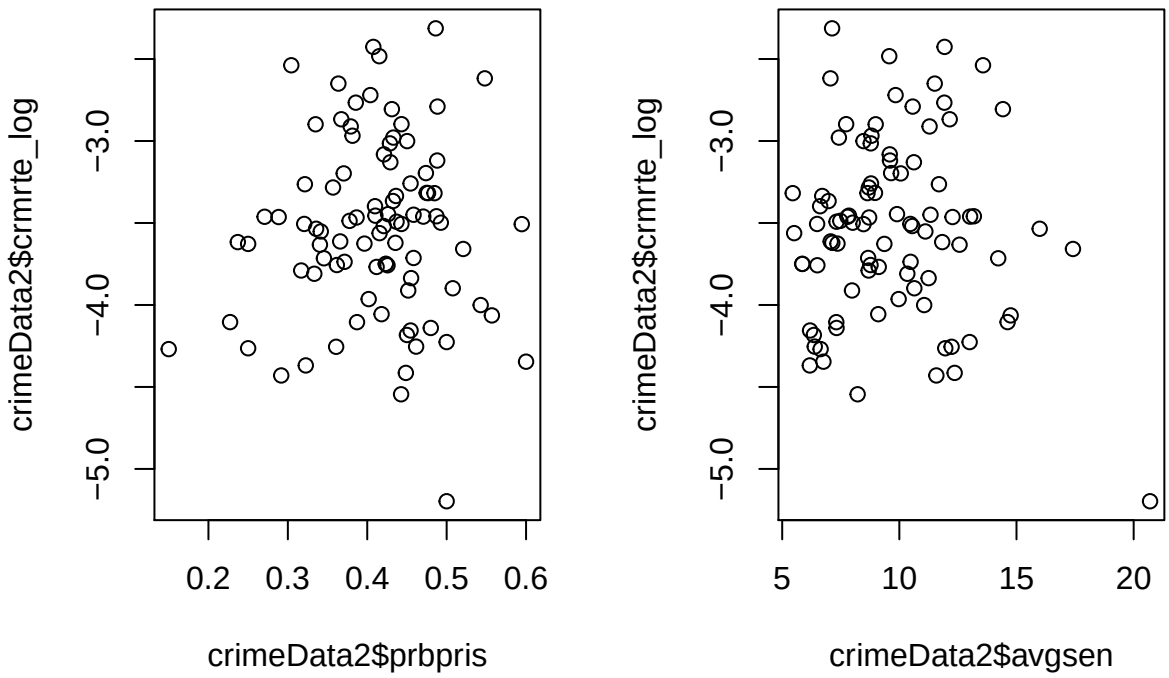
```
cor(crimeData2$prbconv, crimeData2$crmrte_log)
```

```
## [1] -0.4128166
```

Probability of prison (prbpris) | Average sentence days (avgsen)

Neither scatter plots below (prbpris vs. crmrte_log, avgsen vs. crmrte_log) shows obvious relationships. The correlation coefficients0 are only 0.03 and -0.08 respectively.

```
par(mfrow=c(1,2))
plot(crimeData2$prbpris, crimeData2$crmrte_log)
plot(crimeData2$avgsen, crimeData2$crmrte_log)
```




```
cor(crimeData2$prbpris, crimeData2$crmrte_log)
```

```
## [1] 0.02938727
```

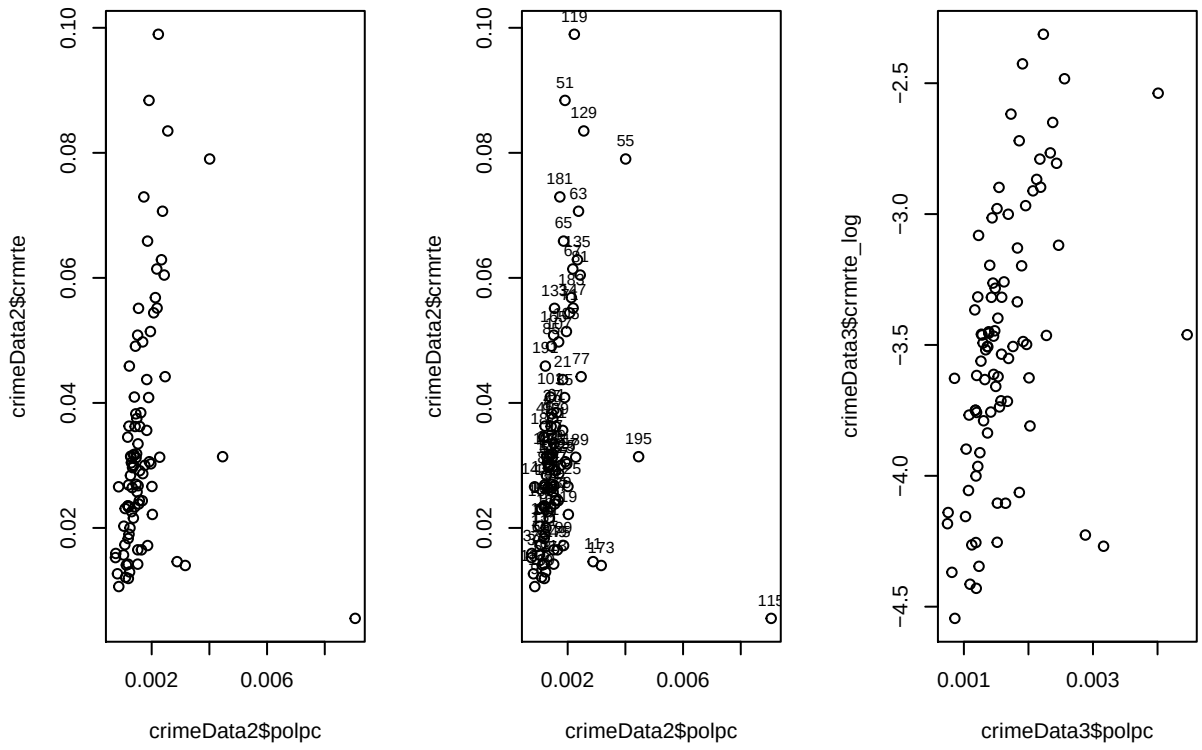
```
cor(crimeData2$avgsen, crimeData2$crmrte_log)
```

```
## [1] -0.07567514
```

Police per capita (polpc)

Similar to probabilities of arrest and conviction, we observed a linear relationship between crime rate and policy per capita in the scatter plot below. The scatter plot also shows that County 15 has significantly higher police per capital than any other counties, County 15 is a highly leveraged observation. In addition, the variation of crmrte increases as prbconv increases, which justifies taking the log of crmrte. The correlation between crmrte_log and polpc (after removing County 115) is 0.45.

```
par(mfrow=c(1,3))
plot(crimeData2$polpc, crimeData2$crmrte)
plot(crimeData2$polpc, crimeData2$crmrte)
text(crimeData2$polpc, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3)
plot(crimeData3$polpc, crimeData3$crmrte_log)
```

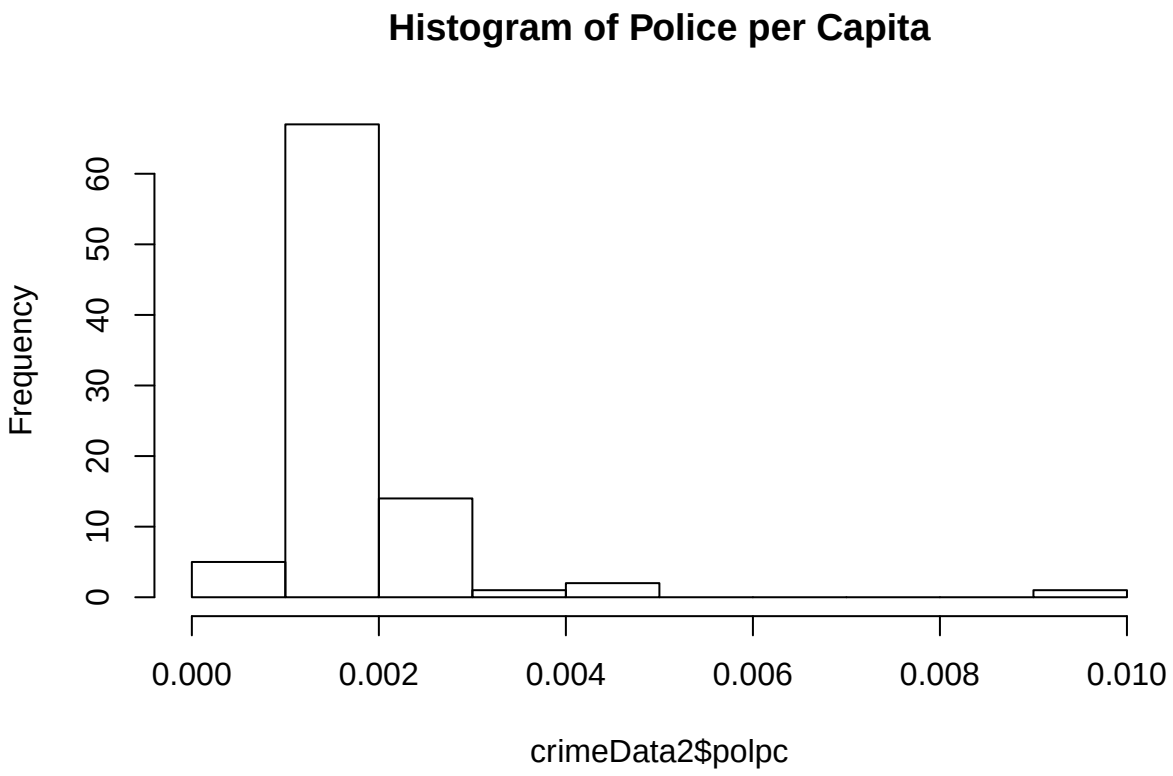



```
cor(crimeData3$polpc, crimeData3$crmrte_log)
```

```
## [1] 0.453951
```

We also noticed that the distribution of polpc is highly skewed to the right, so we took the log of polpc. The correlation coefficient (after removing County 115) increased from 0.45 to 0.54.

```
hist(crimeData2$polpc, main="Histogram of Police per Capita")
```



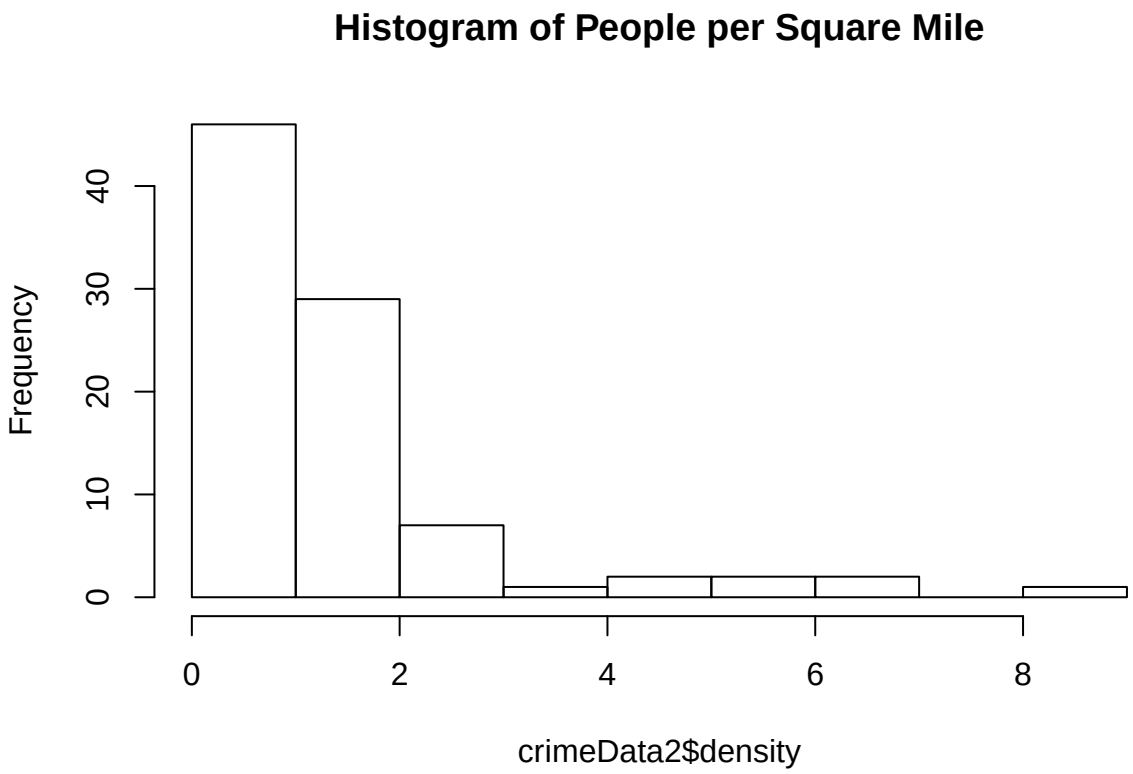

```
crimeData2$polpc_log <- log(crimeData2$polpc)
crimeData3 <- crimeData2[which(crimeData2$county!=115),]
cor(crimeData3$polpc_log, crimeData3$crmrte_log)
```

```
## [1] 0.541829
```

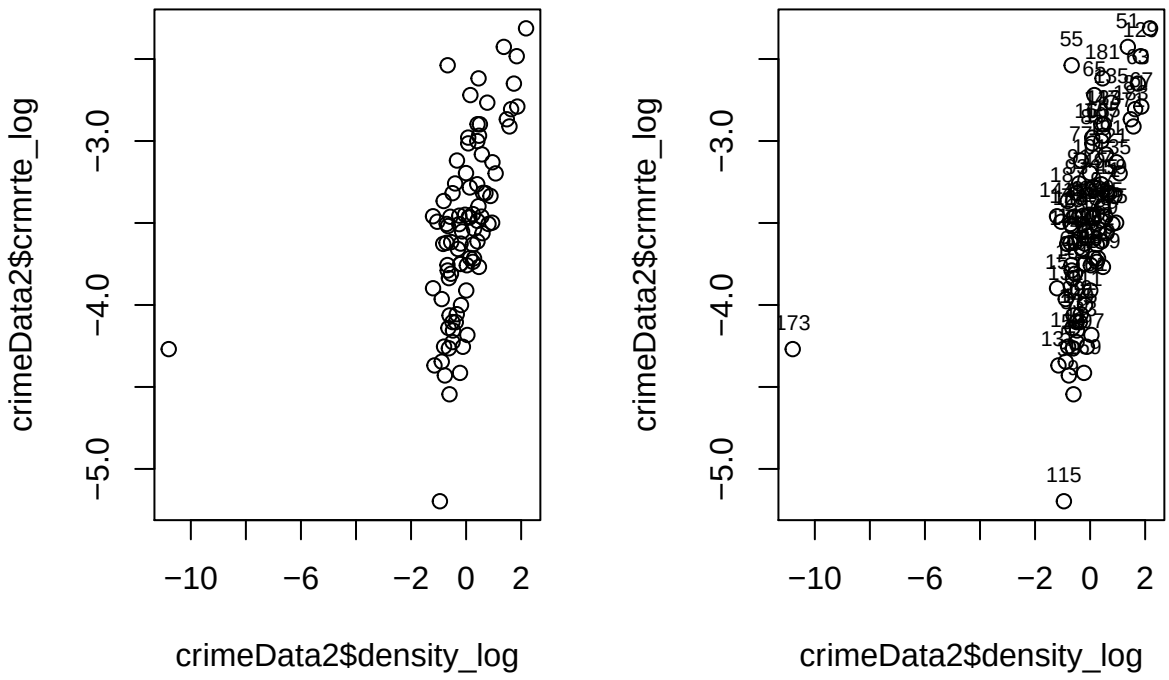
People per square mile (density) | If in SMSA (urban)

The histogram shows the distribution of density is highly skewed to the right, so we took the log of density. The scatter plot shows County 173 is highly leveraged as it has much lower population density than other counties. Removing County 173 significantly increases correlation coefficient from 0.49 to 0.68. The correlation between density and crmrte_log (without County 173) is 0.63, which is lower than the correlation between density_log and crmrte_log (without County 173) of 0.68. This further confirms that log of density has a stronger linear relationship with log of crime rate than density does.

```
hist(crimeData2$density, main="Histogram of People per Square Mile")
```




```
crimeData2$density_log <- log(crimeData2$density)
par(mfrow=c(1,2))
plot(crimeData2$density_log, crimeData2$crmrte_log)
plot(crimeData2$density_log, crimeData2$crmrte_log)
text(crimeData2$density_log, crimeData2$crmrte_log, labels = crimeData2$county, cex=0.7, pos=3)
```




```
crimeData4 <- crimeData2[which(crimeData2$county!=173),]
cor(crimeData2$density_log, crimeData2$crmrte_log)
```

```
## [1] 0.4909562
```

```
cor(crimeData4$density_log, crimeData4$crmrte_log)
```

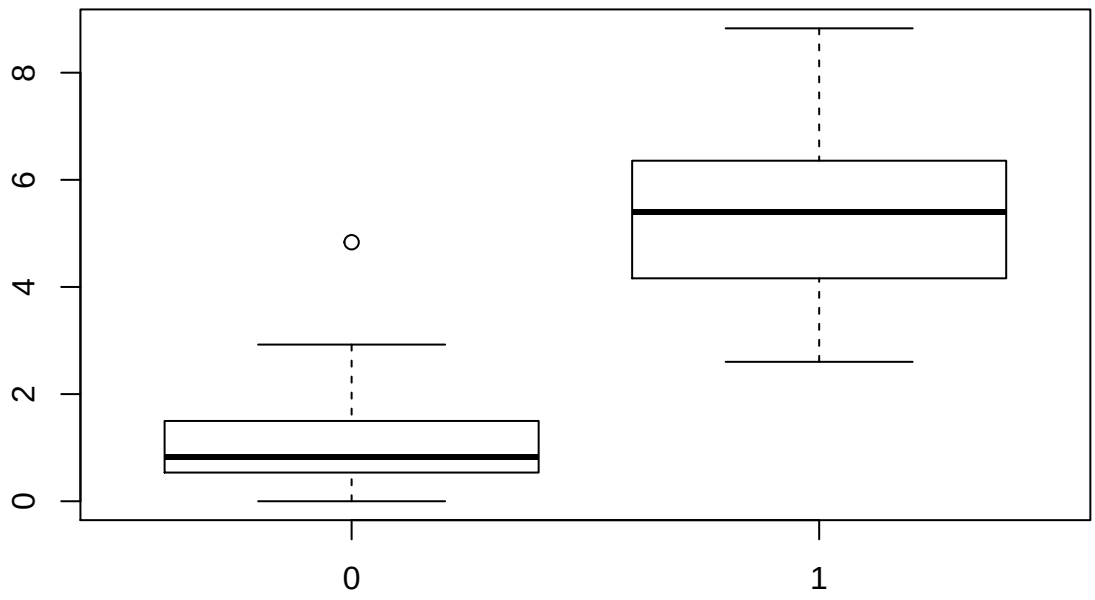
```
## [1] 0.677355
```

```
cor(crimeData4$density, crimeData4$crmrte_log)
```

```
## [1] 0.6281475
```

Urban is a binary variable. The box plot shows that the the mean and interquartile range of density is significantly different depending on whether county is in urban area or not. Log of density is highly correlated with urban with a correlation coefficient of 0.66. When building the model, we should avoid putting both variables in the model for two reasons:
1. Adding the second variable doesn't explain much additional variation of the response variable
2. High correlation can greatly increase the standard errors of parameter estimates

```
boxplot(density~urban, data=crimeData2)
```



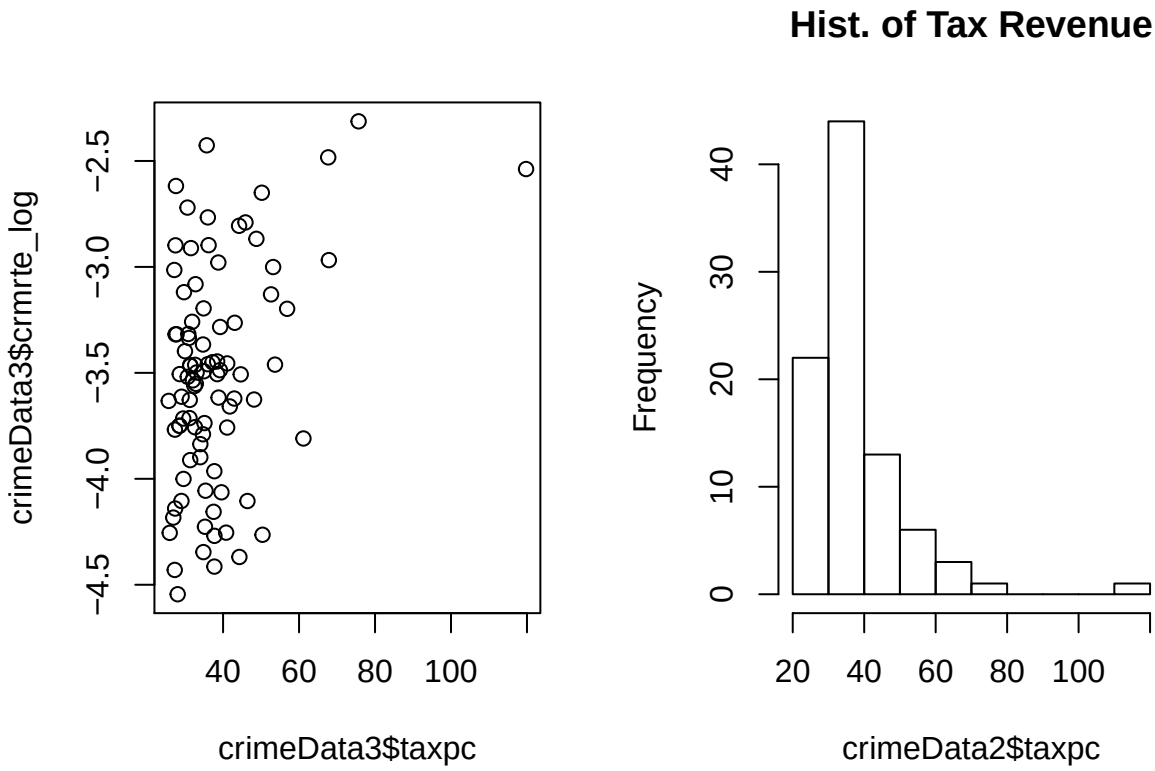

```
cor(crimeData4$urban, crimeData4$density_log)
```

```
## [1] 0.660531
```

Tax revenue per capita (taxpc)

The scatter plot indicates there may be a weak linear relationship between taxpc and crmrte_log. The histogram of taxpc is skewed to the right, so we considered taking the log of taxpc. However, the correlation between taxpc and crmrte_log (0.37) is slightly higher than the correlation between taxpc_log and crmrte_log (0.36).

```
par(mfrow=c(1,2))
plot(crimeData3$taxpc, crimeData3$crmrte_log)
hist(crimeData2$taxpc, main="Hist. of Tax Revenue")
```




```
crimeData2$taxpc_log <- log(crimeData2$taxpc)
cor(crimeData2$taxpc, crimeData2$crmrte_log)
```

```
## [1] 0.3711452
```

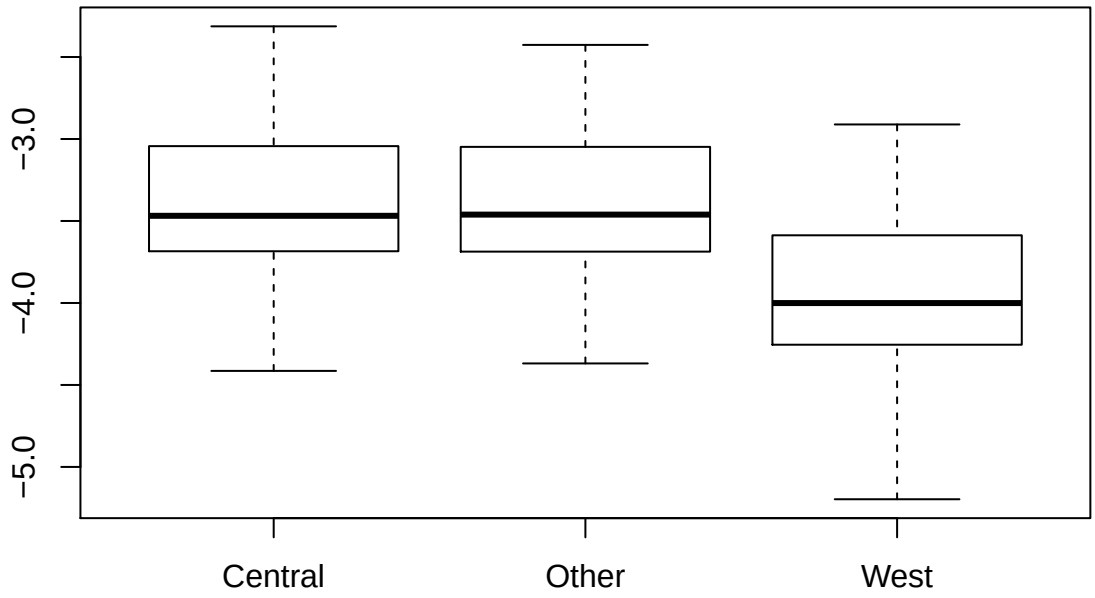
```
cor(crimeData2$taxpc_log, crimeData2$crmrte_log)
```

```
## [1] 0.3570773
```

If in western/central North Carolina

We created a variable, area, to categorize the area counties reside in. Area takes three values: West, Central, and Other. The box plot shows that the mean and interquartile range of crmrte_log is very similar between Central and Other. The crmrte_log for West area is lower than other areas.

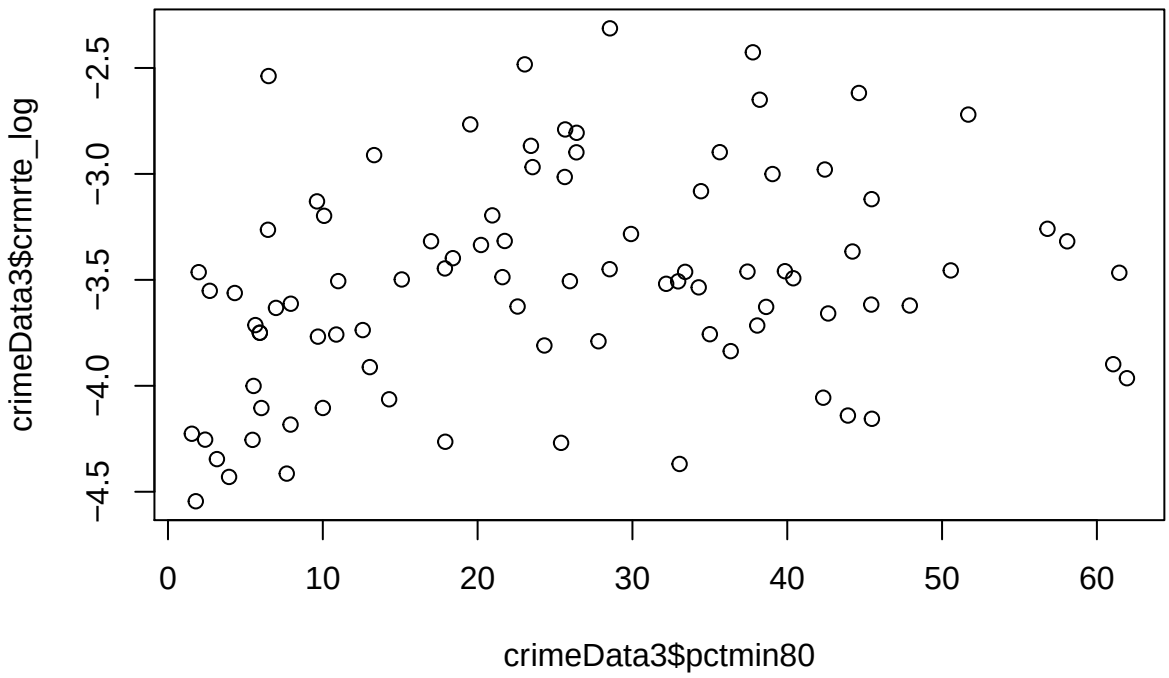
```
crimeData2$area <- ifelse(crimeData2$west==1, "West", ifelse(crimeData2$central==1, "Central", "Other"))
boxplot(crmrte_log~area, data=crimeData2)
```



Percent of minority in 1980 (pctmin80)

The scatter plot shows a weak linear relationship between log of crime rate and percent of minority. Low correlation coefficient (0.3) also confirms that.

```
plot(crimeData3$pctmin80, crimeData3$crmrte_log)
```




```
cor(crimeData2$pctmin80, crimeData2$crmrte_log)
```

```
## [1] 0.2957882
```

Weekly wages

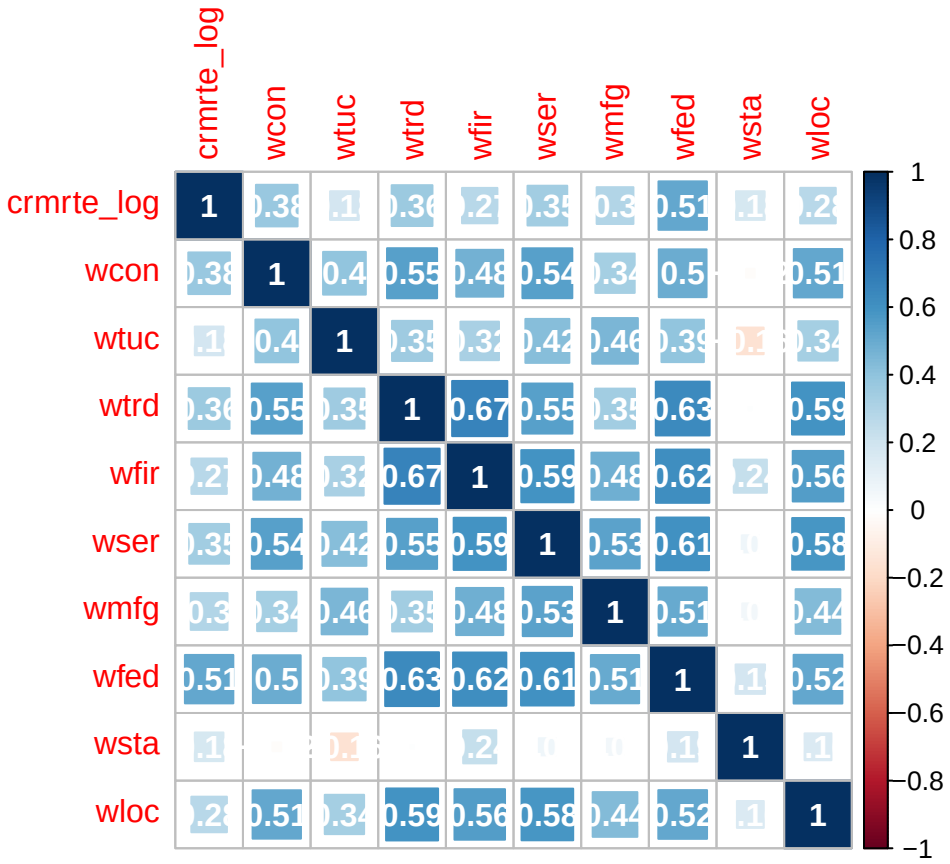
There are nine variables related to weekly wages in the data. They represent weekly wages in different industries. As the correlation matrix shows, most of the weekly wages variables are highly correlated except for wsta. When building the model, we should avoid putting all the correlated variables in the model for the same reason pointed out in the density/urban section of the EDA. We also noticed that log of crime rate has the strongest linear relationship with wfed with correlation coefficeint of 0.51.

```
crimeData_temp1 <- crimeData2[,c("crmrte_log","wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "
corr_DenUr <- cor(crimeData_temp1, use="pairwise")
library(corrplot)
```

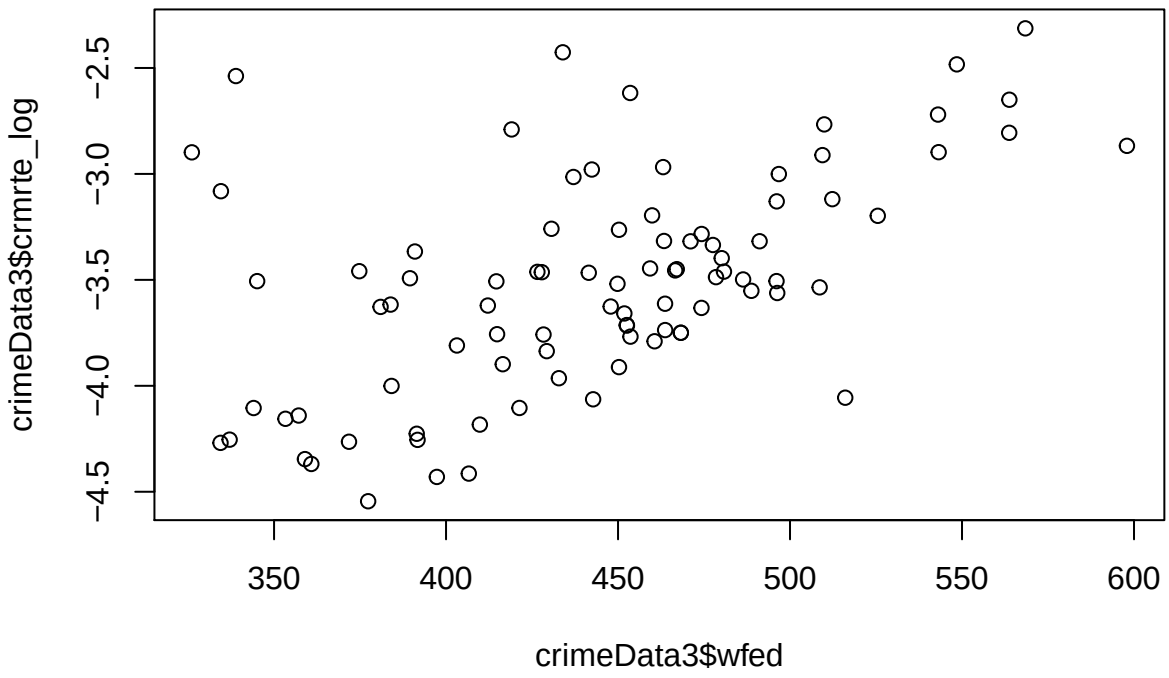
```
## Warning: package 'corrplot' was built under R version 3.4.4
```

```
## corrplot 0.84 loaded
```

```
corrplot(corr_DenUr, method="square", addCoef.col="white")
```



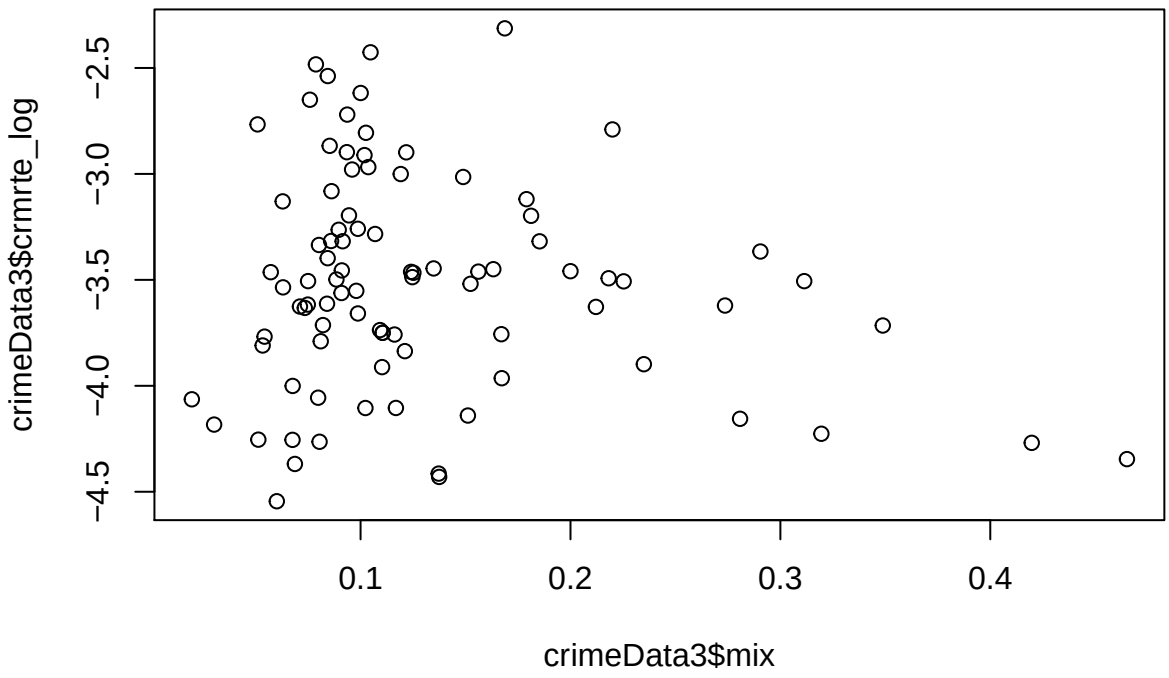

```
plot(crimeData3$wfed, crimeData3$crmrte_log)
```



Offense mix: face-to-face/other (mix)

The scatter plot doesn't indicate a strong relationship between mix and log of crime rate. The weak correlation coefficient (-0.15) also confirms that.

```
plot(crimeData3$mix, crimeData3$crmrte_log)
```



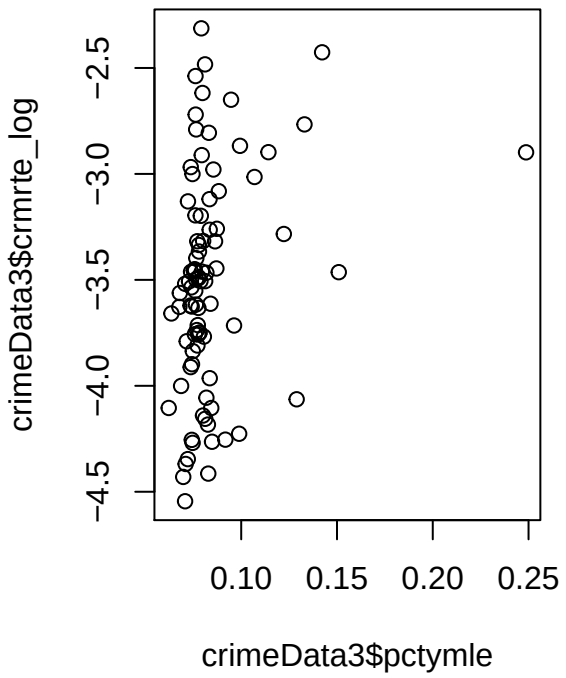

```
cor(crimeData2$mix, crimeData2$crmrte_log)
```

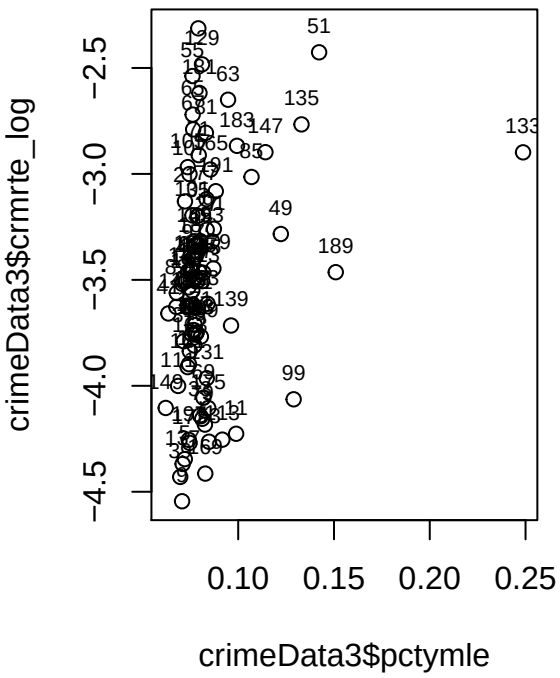
```
## [1] -0.1466527
```

Percent young male (pctymle)

The scatter plot shows that the majority of counties have 5%-10% of young male. County 133 has significantly higher male percentage than the rest of counties. Log of crime rate doesn't seem to vary by the percent of young male based on the scatter plot, which is also evidented by 0.27 correlation coefficient.

```
par(mfrow=c(1,2))
plot(crimeData3$pctymle, crimeData3$crmrte_log)
plot(crimeData3$pctymle, crimeData3$crmrte_log)
text(crimeData2$pctymle, crimeData2$crmrte_log, labels = crimeData2$county, cex=0.7, pos=3)
```






```
cor(crimeData2$pctymle, crimeData2$crmrte_log)
```

```
## [1] 0.2723973
```

Model Building 1

```
model1 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen, data=crimeData2)
model1$coefficients
```

```
## (Intercept)      prbarr     prbconv     prbpris      avgsen
## -2.92426013 -2.09740014 -0.79654556  0.42628304  0.02705387
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.4.4
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
#coeftest(model1, vcov = vcovHC)
#plot(model1)
```

```
#AIC(model1)
```

Cook's distance for Observation 51, which is County 115, is close to 1.

Model Building 2

We put all the variables with correlation higher than 0.25 in the model except for those highly correlated with each other.

```
model2 <- lm(crmrte_log ~ prbarr+prbconv+polpc_log+density_log+west+pctmin80+wfed+pctymle, data=crimeDa
model2$coefficients
```

```
##  (Intercept)       prbarr      prbconv    polpc_log  density_log
## -0.002959757 -1.876715831 -0.665762798  0.546997370  0.083469653
##         west     pctmin80         wfed      pctymle
## -0.101386251  0.010011686  0.001356532  1.111404818
```

```
#coeftest(model2, vcov = vcovHC)
#plot(model2)
#AIC(model2)
```

Model Building 3

```
model3 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen+polpc_log+density_log+taxpc+west+central+urban+
model3$coefficients
```

```
##   (Intercept)        prbarr       prbconv       prbpris        avgsen
##  0.6276444944 -1.8190558437 -0.5659797104 -0.6806825821 -0.0270524579
##     polpc_log   density_log         taxpc          west       central
##  0.6021669897  0.1027476586  0.0019230455 -0.0683115182 -0.0712727704
##         urban    pctmin80_2          wcon          wtuc          wtrd
##  0.1513316157  0.9980385344  0.0003784841  0.0002289608  0.0017024817
##          wfir          wser          wmfg          wfed          wsta
## -0.0009627851 -0.0020813035 -0.0003396119  0.0017579658 -0.0006933810
##          wloc           mix       pctymle
##  0.0016560744 -0.0504661995  2.0567468321
```

```
#coeftest(model3, vcov = vcovHC)
#plot(model3)
#AIC(model3)
```

Model Display

```
#se.model1 = sqrt(diag(vcovHC(model1)))
#se.model2 = sqrt(diag(vcovHC(model2)))
#se.model3 = sqrt(diag(vcovHC(model3)))
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.4.3
```

```
##
## 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",
          title = "Linear Models Predicting Log of Crime Rate",
          omit.stat="f",
          #se=list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          float=FALSE)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 01, 2018 - 8:44:12 PM

	Dependent variable:		
	crmrte_log		
	(1)	(2)	(3)
prbarr	-2.097^{***}	-1.877^{***}	-1.819^{***}
	(0.323)	(0.215)	(0.232)
prbconv	-0.797^{***}	-0.666^{***}	-0.566^{***}
	(0.145)	(0.082)	(0.098)
prbpris	0.426		-0.681
	(0.543)		(0.362)
avgsen	0.027		-0.027^{*}
	(0.016)		(0.011)
polpc_log		0.547^{***}	0.602^{***}
		(0.079)	(0.106)
density_log		0.083^{***}	0.103^{***}
		(0.024)	(0.030)
taxpc			0.002
			(0.003)
west		-0.101	-0.068
		(0.082)	(0.109)
pctmin80		0.010^{***}	
		(0.002)	
central			-0.071
			(0.074)
urban			0.151
			(0.116)
pctmin80_2			0.998^{***}
			(0.258)
wcon			0.0004
			(0.001)
wtuc			0.0002
			(0.0004)
wtrd			0.002
			(0.001)
wfir			-0.001
			(0.001)
wser			-0.002^{*}
			(0.001)
wmfg		23	-0.0003
			(0.0004)
wfed		0.001^{*}	0.002^{*}

Omitted Variables