Lab 3 - Predicting Crime (DRAFT)

W203 Statistics for Data Science

Group 4 - Daniel Alvarez, Anup Jha, Mumin Khan, Peter Wang

November 17th, 2018

Contents

Introduction

In this report, our team seeks to help a political campaign identify the key determinants of crime and recommend policy changes in relation to these key factors. The examination of data starts with the given data set of crime statistics for 90 unique counties and 24 variables. We will then use EDA and other techniques to find the key variables and covariates that contribute to crime statistics. Based on these variables and covariates, we seek to generate useful and effective policy recommendations for the political campaign. The dataset is described in the following table.

Number	Variable	Description
1	county	county identifier
2	year	1987
3	crmrte	crimes committed per person
4	prbarr	'probability' of arrest
5	prbconv	'probability' of conviction
6	prbpris	'probability' of prison sentence
7	avgsen	avg. sentence, days
8	polpc	police per capita
9	density	people per sq. mile
10	taxpc	tax revenue per capita
11	west	=1 if in western N.C.
12	central	=1 if in central N.C.
13	urban	=1 if in SMSA
14	pctmin80	perc. minority, 1980
15	wcon	weekly wage, construction
16	wtuc	weekly wage, transportation, utilities, communication
17	wtrd	weekly wage, wholesale, retail trade
18	wfir	weekly wage, finance, insurance, real estate
19	wser	weekly wage, service industry
20	wmfg	weekly wage, manufacturing
21	wfed	weekly wage, fed employees
22	wsta	weekly wage, state employees
23	wloc	weekly wage, local gov employees
24	mix	offense mix: face-to-face/other
25	pctymle	percent young male

Anomaly and Outlier detection

```
crimeData <- read.csv("crime_v2.csv", stringsAsFactors = F)
```

We see from the summary that there are 6 rows with all values as NA. Which means we have missing data in the file. So we will remove these rows from the data set. We also see that prbconv, probability of conviction, is a character variable while it should be numeric so we will convert to numeric. Also one of the rows seems to be duplicated. So we need to get rid of that row.

```
#remove NA
crimeData <- crimeData[!is.na(crimeData$county),]
#Treat prbconv as numeric
crimeData$prbconv <- as.numeric(crimeData$prbconv)
#Remove duplicate row
crimeData <- crimeData[!duplicated(crimeData),]
```

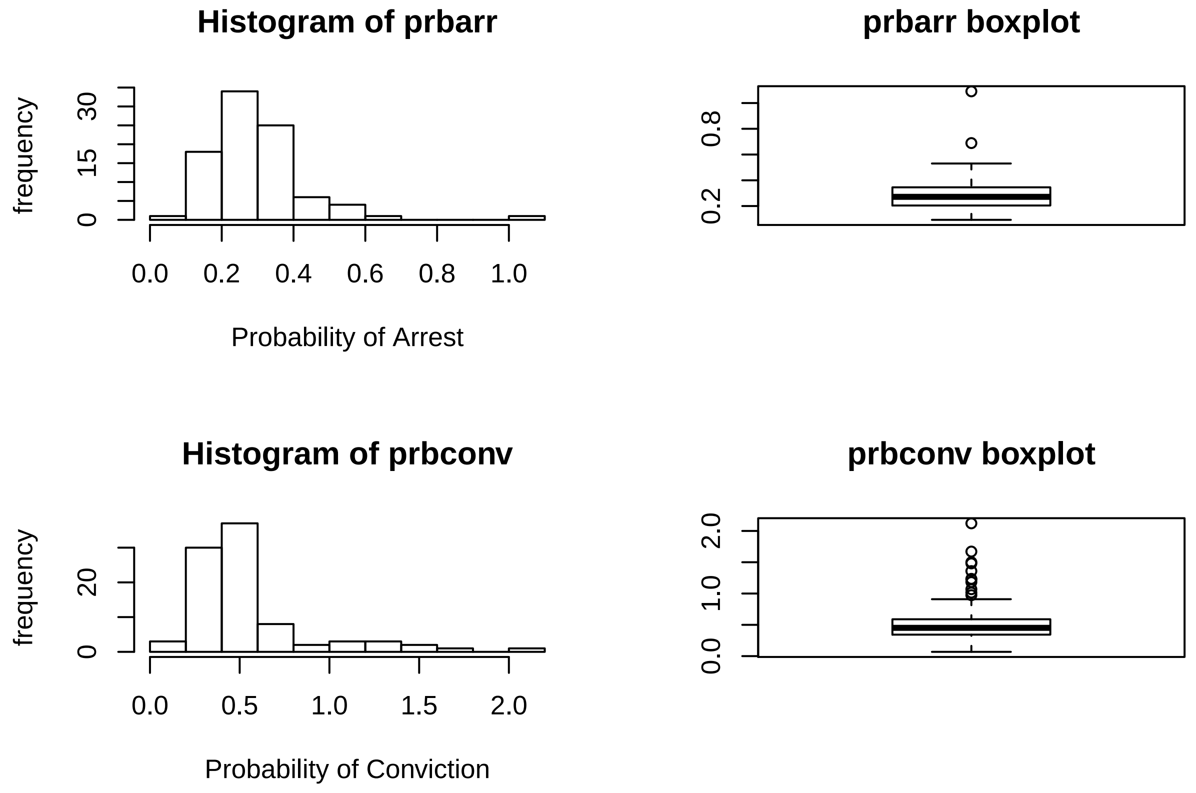
```r
#number of rows
nrow(crimeData)
```

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

So we have sample size of 90 county-level observations in the state of North Carolina. We know that North Carolina is comprised by 100 counties; therefore, 10 county-level observations are missing from the entire population.

We looked at several variables for outliers and anomalies. Following are a few salient ones:

```r
par(mfrow=c(2,2))
hist(crimeData$prbarr,main="Histogram of prbarr", xlab="Probability of Arrest",ylab = "frequency")
boxplot(crimeData$prbarr, main ="prbarr boxplot")
hist(crimeData$prbconv,main="Histogram of prbconv",xlab="Probability of Conviction",
     ylab = "frequency")
boxplot(crimeData$prbconv, main ="prbconv boxplot")
```

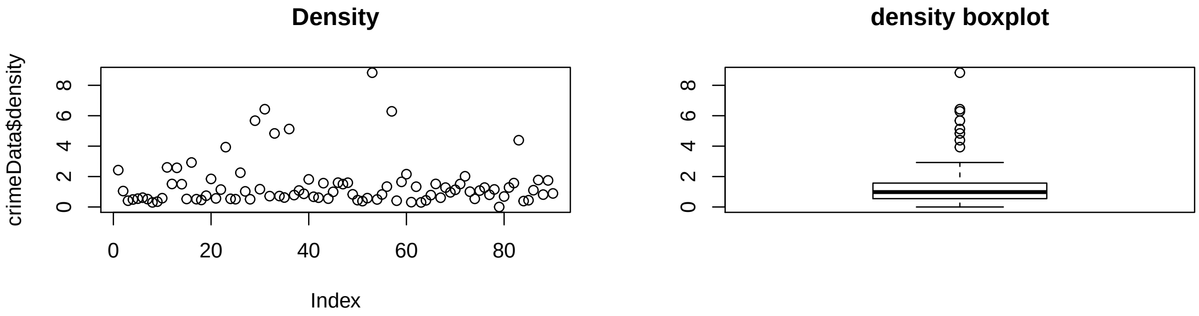


For the variable prbarr, ‘probability’ of arrest, we see that there is one data point which has probability value greater than one which is not possible if it was true probability. However, we determine that since the prbarr variable is a proxy variable of the ratio of arrests to offenses, it can be possible to have a probability value greater than one where arrests are more than offenses reported. The histogram and boxplots of the prbarr variable shows that it is right-skewed, yet we do not have large concerns with violations from normality with the large sample size.

Similarly, the prbconv variable, ‘probability’ of conviction, has ten data points where the probability value is greater than one which is not possible if it was a true probability measure. Again, we determine that since the prbconv is a proxy variable for the number of convictions to arrests, it can be possible to have a probability value greater than one where convictions are more than arrests reported. The histogram and boxplot of the prbconv variable shows a high positive skew as a result of these outlier observations.

We don't have much information about how were these variables were calculated other than that prbarr is equal to the ratio of arrest to offenses and prbconv is equal to the number of convictions to arrests. We know that the data is from the year 1987 so it might have happened for some counties the with different police or legal policies that allow for arrests to be made without a noted offense reported or for convictions without arrests required. It could also be due to timing whereby arrests or convictions were made in 1987 for offenses or arrests that occurred in the previous year so the proxy variables are greater than 1. We don't omit this data as it doesn't seem erroneous but just that its a proxy variable so can have value greater than 1 for certain situation. Greater than 1 just represents higher probability.

```
par(mfrow=c(1,2))
plot(crimeData$density,main="Density")
boxplot(crimeData$density, main ="density boxplot")
```

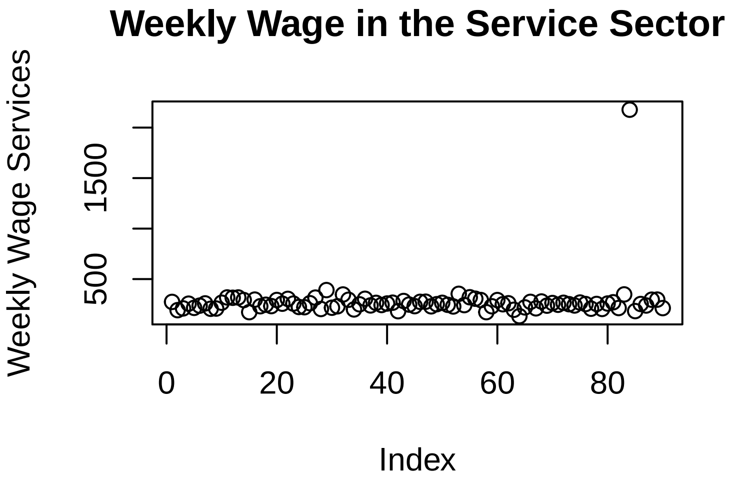


```
nrow(crimeData[which(crimeData$density>4),])
```

```
## [1] 7
```

We see that there are a few outliers in the positive side for the density with seven values greater than a density of 4 people per square mile.

```
plot(crimeData$wser,ylab="Weekly Wage Services",main="Weekly Wage in the Service Sector")
```



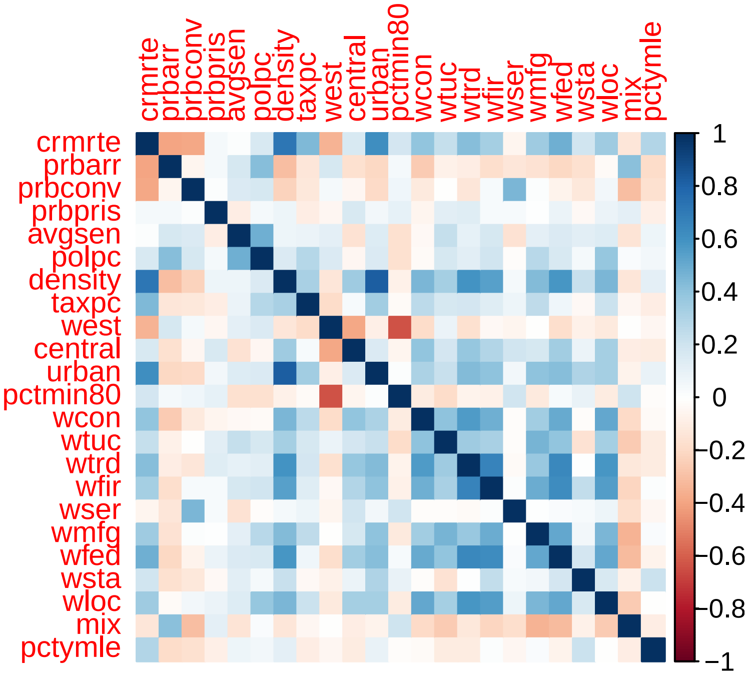
We see that variable wser, weekly wage in the service industry, has an outlier. Index 84 which has wser value of 2177.0681 which is more than 4 times the mean value of wser for all other counties. We are going to ignore this row if we are modeling using the wser variable.

EDA

Correlation Matrix

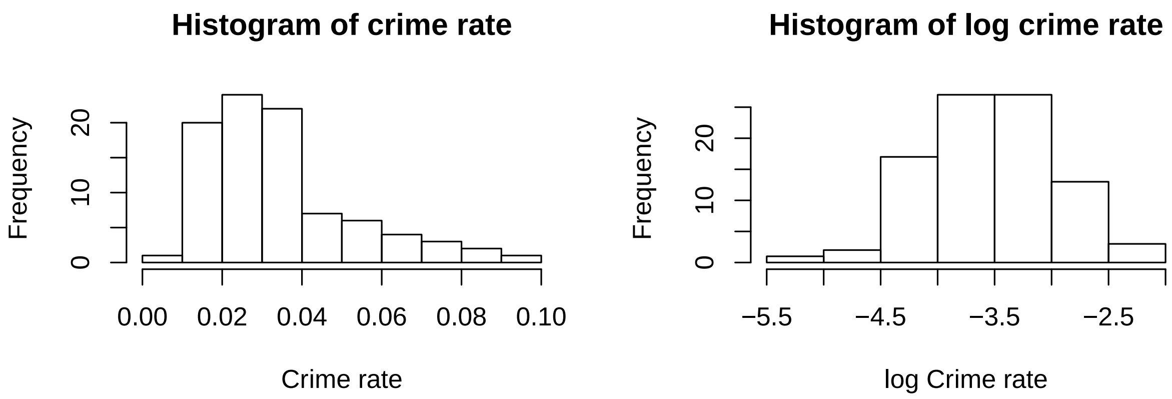
We are trying to model dependent variable crmrte, crimes committed per person. As a first step, we show the correlation matrix and plot it to check the collinear relationships visually through a heat map. This will guide us in our EDA and model building. We can ignore the year and county column as year is a constant since all data comes from the same year and county identifiers are just the serial numbers and do not provide any meaningful information.

```
corrmatrix <- cor(crimeData[,3:25])
corrplot(corrmatrix, method = "color",tl.cex = 0.90)
```



We see from the overall picture that important variables which seems to be related to crime rate are prbarr, prbconv, prbpris, polpc, density, taxpc, pctymle, and almost all the wage variables.

```
par(mfrow=c(1,2))
hist(crimeData$crmrte,main="Histogram of crime rate",xlab="Crime rate",ylab="Frequency")
hist(log(crimeData$crmrte),main="Histogram of log crime rate",xlab="log Crime rate",
     ylab="Frequency")
```

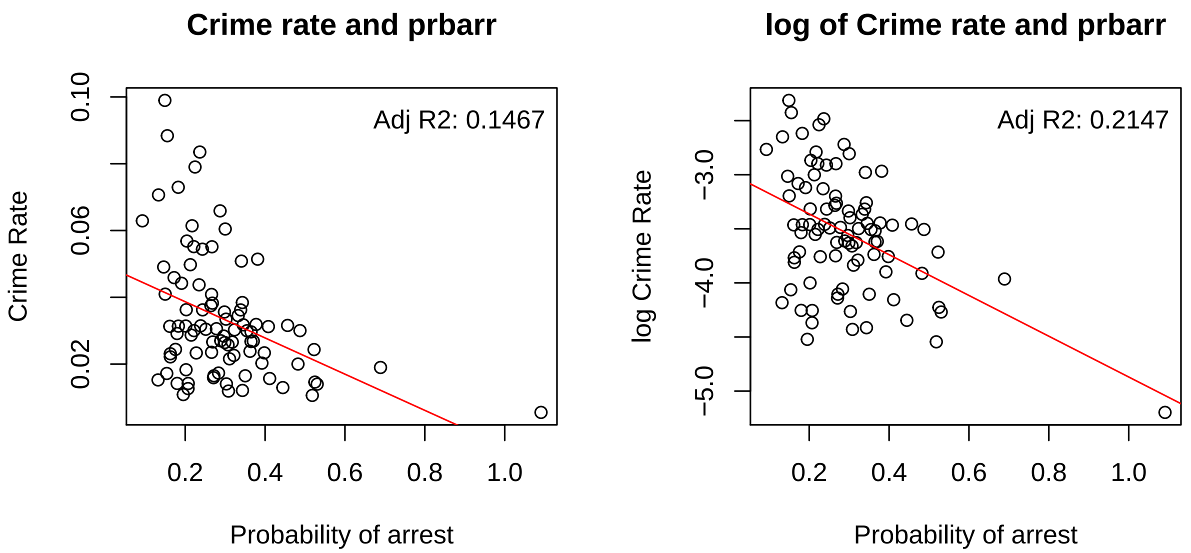


This shows that crime rate has positive skew and log transformation would be appropriate. The histogram of the log-transformed variable shows an approximate normal distribution. Also log transformation in the model would give us way to analyze the percentage increase or decrease in crime rate based on different predictors rather than value change of crime rate, which will be useful to explain policy formation.

Let us now examine the relation between crmrte and some key variables to check what transformations can be beneficial.

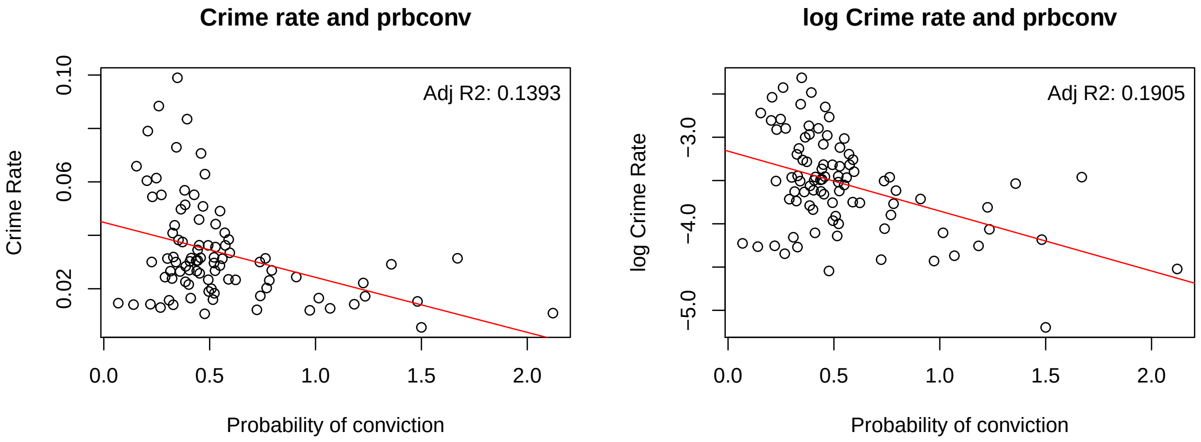
Key Variables

```
par(mfrow=c(1, 2))
plot(crimeData$prbarr,crimeData$crmrte,main="Crime rate and prbarr",
     xlab="Probability of arrest",ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ prbarr, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$prbarr,log(crimeData$crmrte),main="log of Crime rate and prbarr",
     xlab="Probability of arrest",ylab ="log Crime Rate")
abline(fit <- lm(log(crmrte) ~ prbarr, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
```



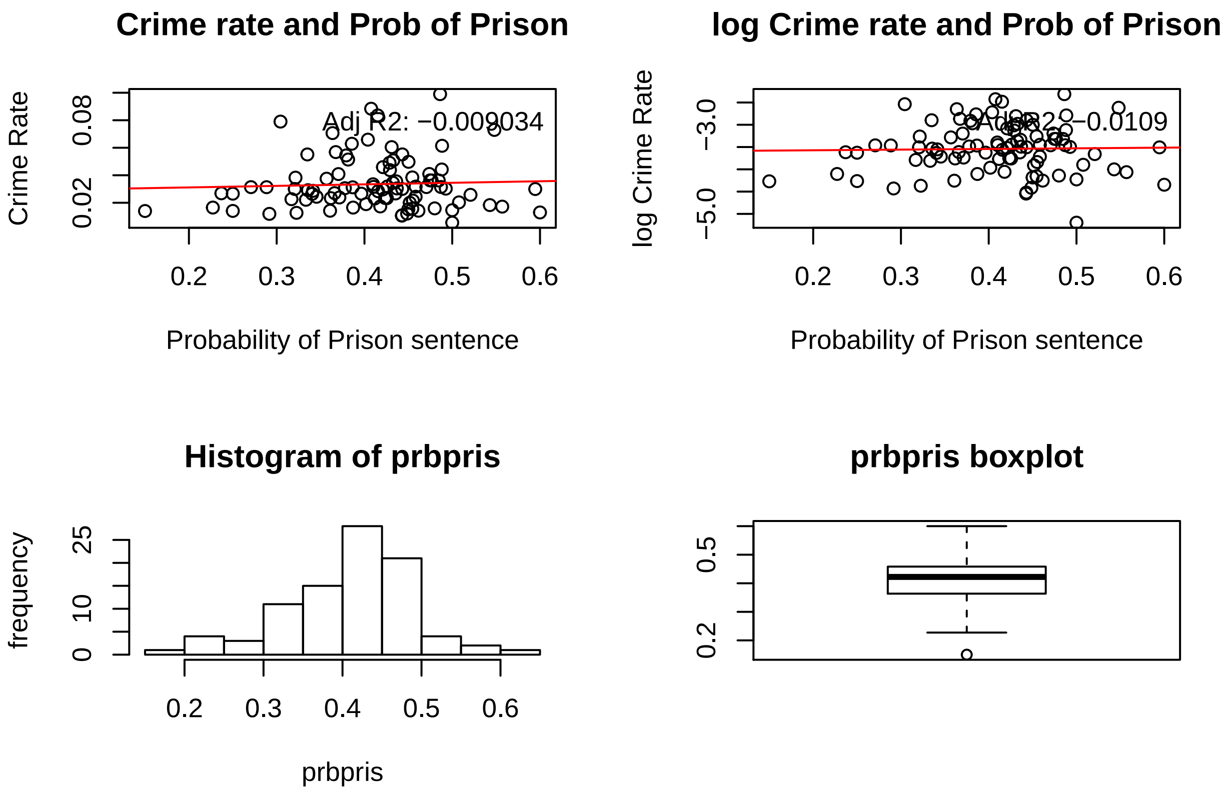
We see that the crime rate declines exponentially with the probability of arrest variable. The log transformation of the dependent variable, crime rate, gives a better fit to the OLS regression, which we observe from the plot. We see that the adjusted R-square from of the log transformed crime rate on probability of arrest is greater than adjusted R-square from the non-transformed crime rate on probability of arrest.

```
par(mfrow=c(1, 2))
plot(crimeData$prbconv,crimeData$crmrte,main="Crime rate and prbconv",
     xlab="Probability of conviction",ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ prbconv, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$prbconv,log(crimeData$crmrte),main="log Crime rate and prbconv",
     xlab="Probability of conviction",ylab ="log Crime Rate")
abline(fit <- lm(log(crmrte) ~ prbconv, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
```



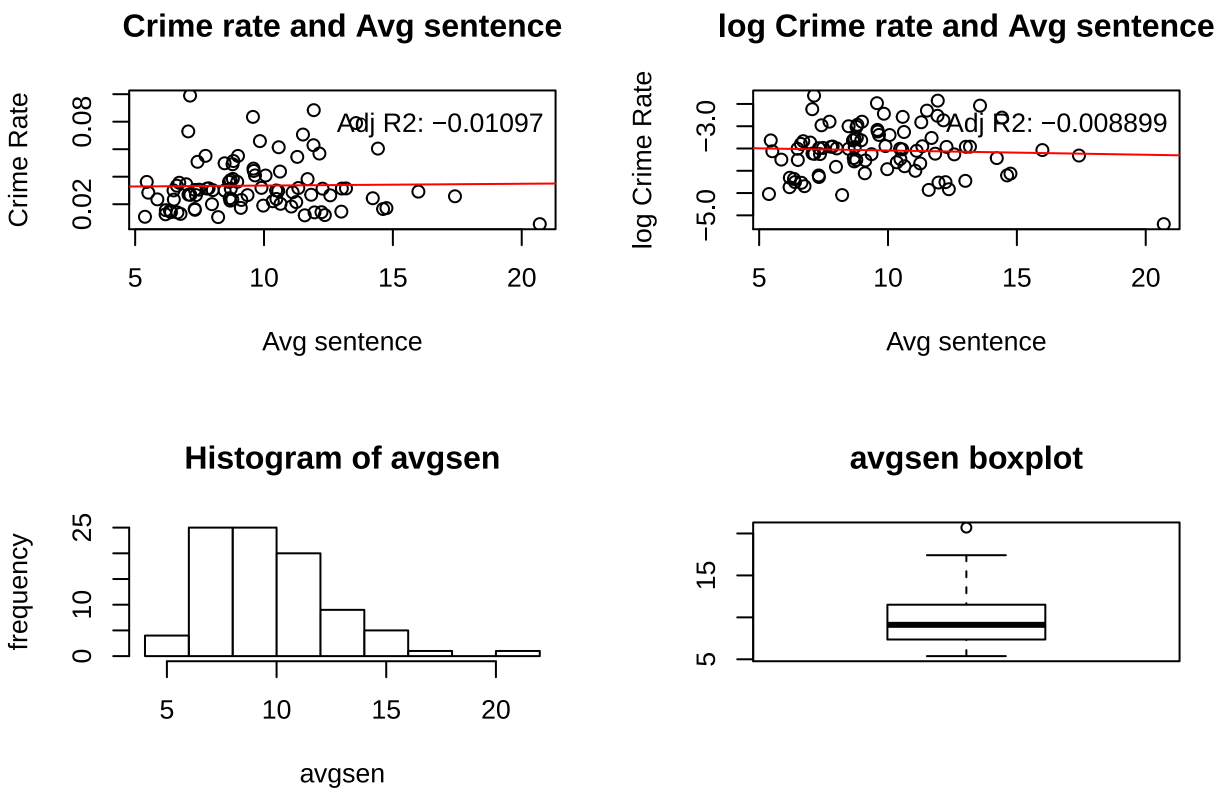
We see that, in general, the probability of conviction has a negative relation with crime rate. Similar to the relationship between crime rate and probability of arrest, we also observe that log transformation of crime rate gives a higher adjusted R-square for the OLS line.

```
par(mfrow=c(2, 2))
plot(crimeData$prbpris,crimeData$crmrte,main="Crime rate and Prob of Prison",
     xlab="Probability of Prison sentence",ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ prbpris, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$prbpris,log(crimeData$crmrte),main="log Crime rate and Prob of Prison",xlab="Probability
abline(fit <- lm(log(crmrte) ~ prbpris, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
hist(crimeData$prbpris,main="Histogram of prbpris", xlab="prbpris",ylab = "frequency")
boxplot(crimeData$prbpris, main ="prbpris boxplot")
```



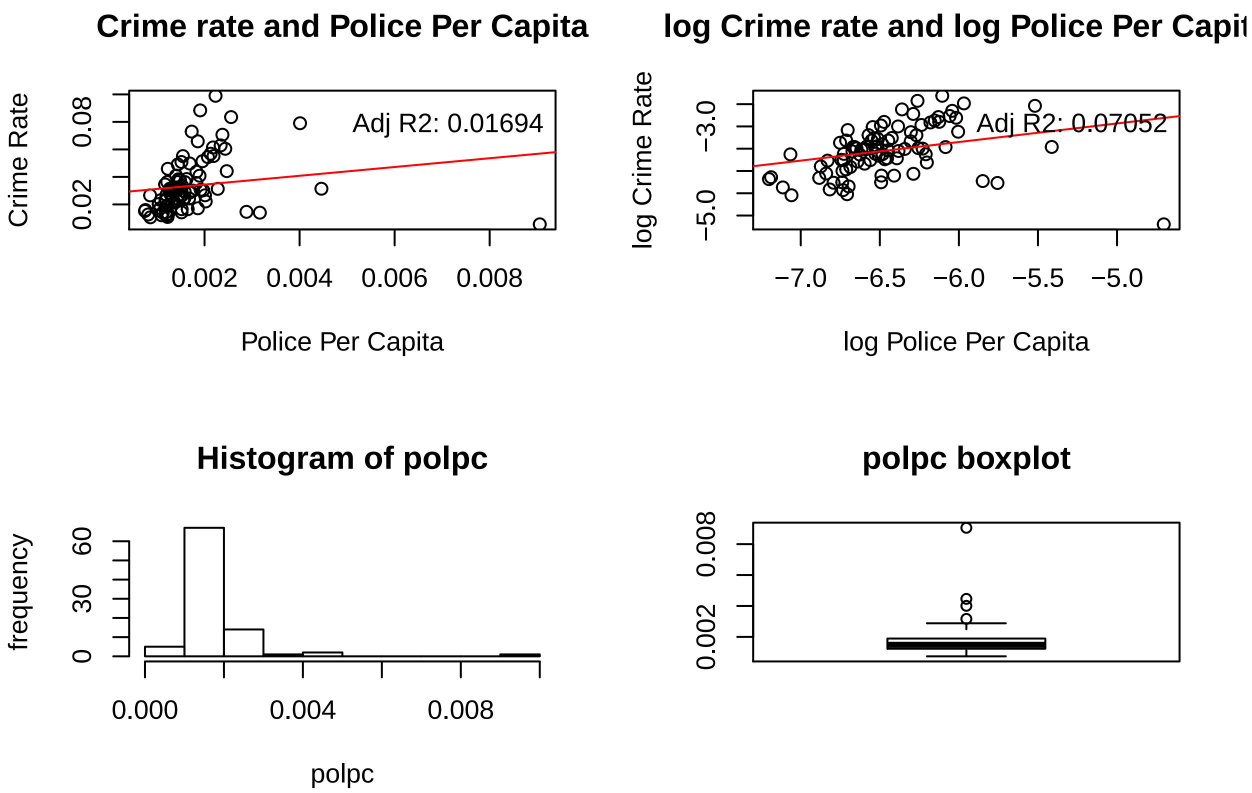
We see that the probability of prison sentence doesn't seem to have much affect on the crime rate, which would indicate that the probability of prison sentence is not acting as a strong deterrent to crime incidence. However, we may surmise that criminals tend to be deterred of being arrested and convicted, yet not so much about a potential prison sentence. We also see that the variable prbpris is about normal. Box plot shows one outlier in the negative side.

```
par(mfrow=c(2, 2))
plot(crimeData$avgsen,crimeData$crmrte,main="Crime rate and Avg sentence",
     xlab="Avg sentence",ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ avgsen, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$avgsen,log(crimeData$crmrte),main="log Crime rate and Avg sentence",xlab="Avg sentence",
abline(fit <- lm(log(crmrte) ~ avgsen, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
hist(crimeData$avgsen,main="Histogram of avgsen", xlab="avgsen",ylab = "frequency")
boxplot(crimeData$avgsen, main ="avgsen boxplot")
```



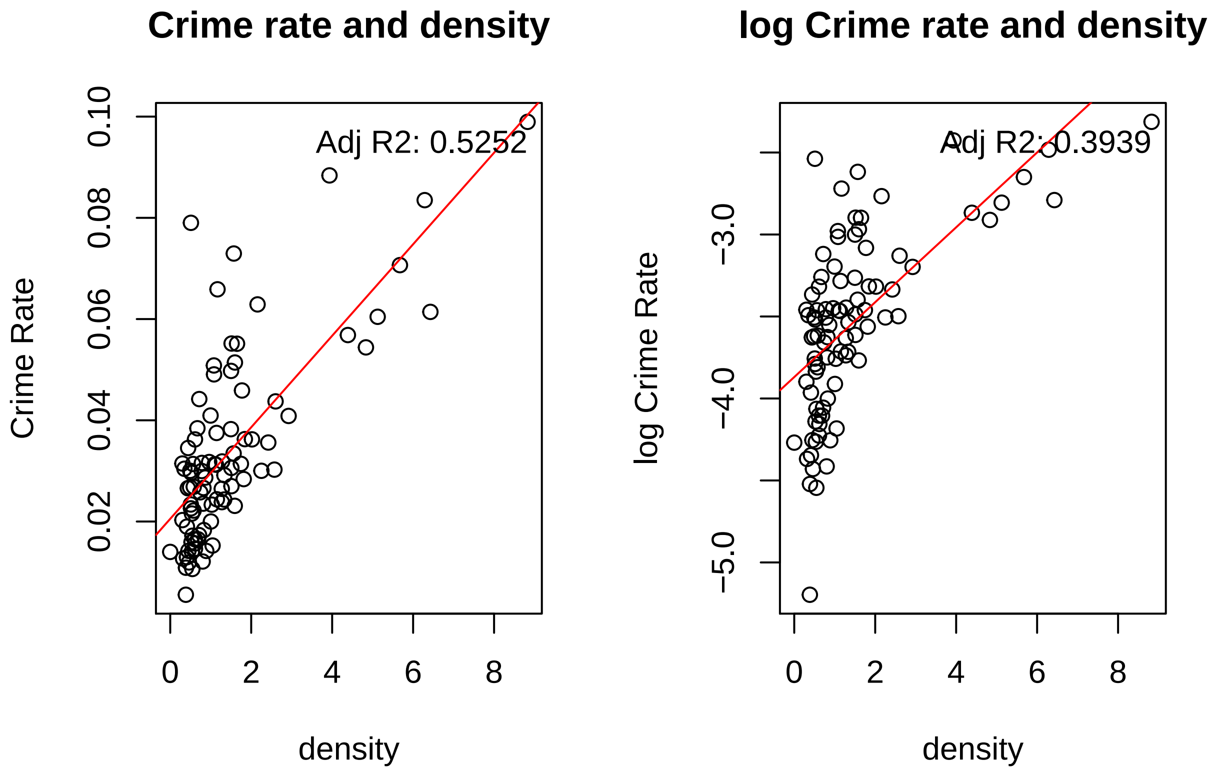
We again see that average sentence doesn't have much of an affect on the crime rate. The average sentence has a positive skew distribution as shown by the histogram and boxplot with one outlier.

```
par(mfrow=c(2, 2))
plot(crimeData$polpc,crimeData$crmrte,main="Crime rate and Police Per Capita",
     xlab="Police Per Capita",ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ polpc, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(log(crimeData$polpc),log(crimeData$crmrte),main="log Crime rate and log Police Per Capita",xlab="l
abline(fit <- lm(log(crmrte) ~ log(polpc), data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
hist(crimeData$polpc,main="Histogram of polpc", xlab="polpc",ylab = "frequency")
boxplot(crimeData$polpc, main ="polpc boxplot")
```



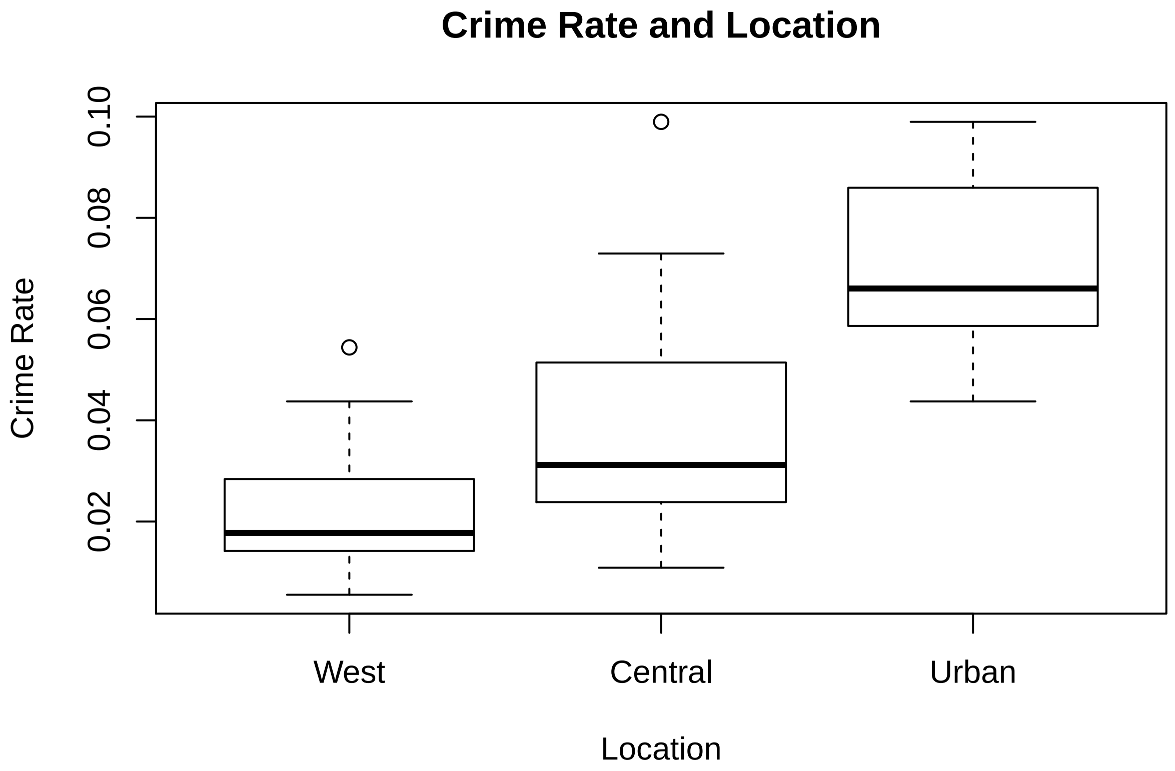
We see that police per capita has a positive relation with crime rate, which may indicate an endogenous relationship with larger police presence in areas with higher crime. We observe a better fit in taking log transformation of both the crime rate and per capita police. The regression line of the log-log bivariate model would show the affect of percentage change in police per capita on percentage change in crime rate. This will be beneficial in terms of explaining the a policing policy. We also see that police per capita has a positive skew and there is one extreme outlier for per capita police which has value of .01, where the crime rate is also the lowest.

```
par(mfrow=c(1, 2))
plot(crimeData$density,crimeData$crmrte,main="Crime rate and density",xlab="density",
     ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ density, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$density,log(crimeData$crmrte),main="log Crime rate and density",xlab="density",
     ylab ="log Crime Rate")
abline(fit <- lm(log(crmrte) ~ density, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
```



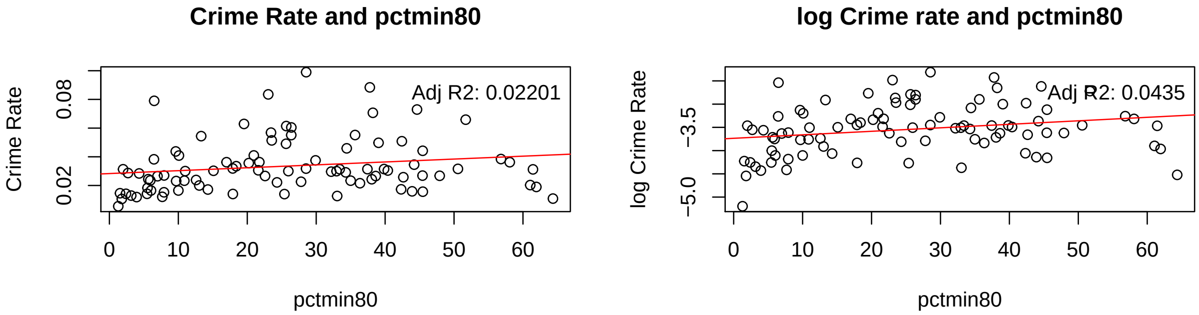
We see that density has strong affect on the crime rate. Density itself has a positively skewed distribution. We may surmise that lower density areas are associated with lower crime rates. There are a few outliers in the positive side for the density variable and for these outliers the crime rate is also high.

```
boxplot(crimeData$crmrte[crimeData$west==1],crimeData$crmrte[crimeData$central==1],
        crimeData$crmrte[crimeData$urban==1], names=c("West","Central","Urban"),
        main="Crime Rate and Location", xlab ="Location",ylab="Crime Rate")
```



We see from the above boxplots that the distribution of the crime rate variable is higher in the Urban areas as compared to the West and Central variables. This makes the Urban area variable a perfect candidate for inclusion in our model specification.

```
par(mfrow=c(1, 2))
plot(crimeData$pctmin80,crimeData$crmrte,main="Crime Rate and pctmin80",xlab="pctmin80",
     ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ pctmin80, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$pctmin80,log(crimeData$crmrte),main="log Crime rate and pctmin80",xlab="pctmin80",ylab =
abline(fit <- lm(log(crmrte) ~ pctmin80, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
```



The percent minority, pctmin80, has a weak, positive relation with crime rate. The pctmin80 itself is spread evenly in all counties and doesn't have outliers.

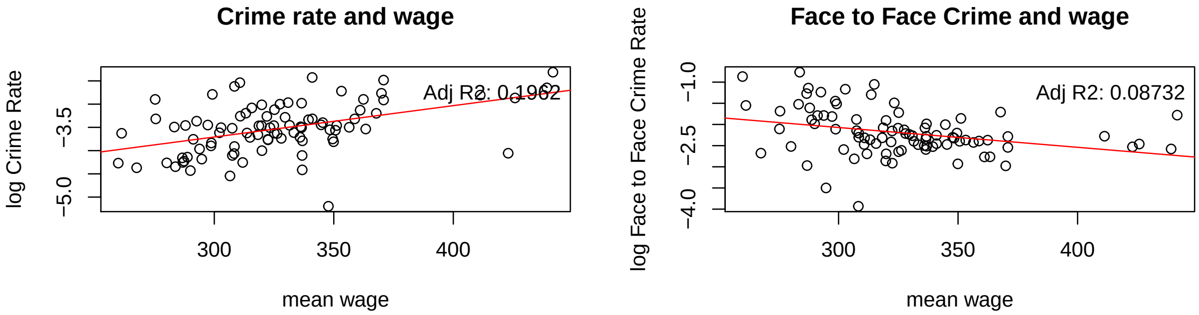
From the correlation plot above we see that the crime rate has positive relation with wage indicators. We can surmise that counties which have higher wages and tax revenue per capita have higher crime rate.

However, the offense mix variable (ratio of face-to-face verses other offenses) has a negative correlation with the wage indicators. We can surmise that counties which have higher wages, the more affluent ones, have lower levels of the face to face crimes (such as battery, robbery) relative to other crimes. A policymaker may be inclined to think that higher wages can reduce the relative incidence of violent crimes.

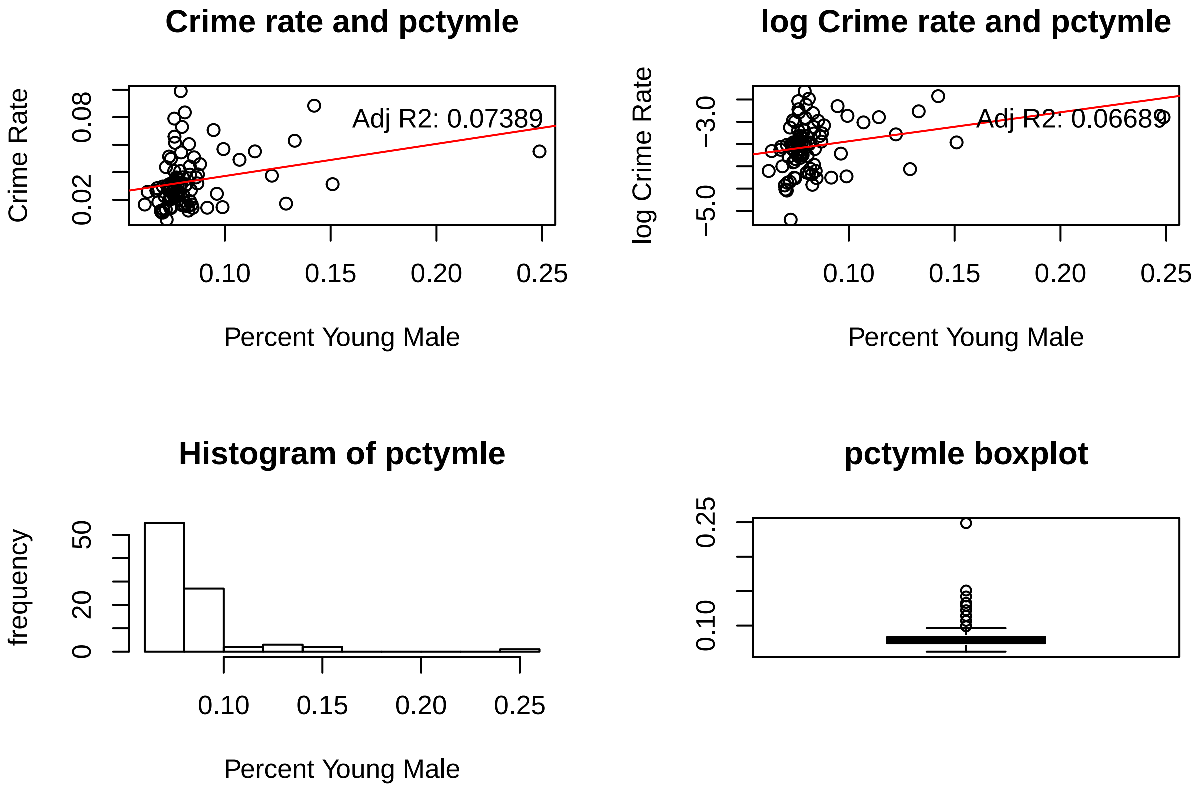
From the plots below, we observe that crime rate increases with average wage but violent crimes decreases with wage.

```
par(mfrow=c(1, 2))
#create Avg wage column don't consider data point 84
crimeData$avgWage <- rowSums(crimeData[,c("wcon","wtuc","wtrd","wfir","wser",
                                          "wmfg","wfed","wsta","wloc")])/9
plot(crimeData$avgWage[-84],log(crimeData$crmrte[-84]),main="Crime rate and wage",
     xlab="mean wage",ylab ="log Crime Rate")
abline(fit <- lm(log(crmrte[-84]) ~ avgWage[-84],data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))

plot(crimeData$avgWage[-84],log(crimeData$mix[-84]),main="Face to Face Crime and wage",
     xlab="mean wage",ylab ="log Face to Face Crime Rate")
abline(fit <- lm(log(mix[-84]) ~ avgWage[-84],data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
```



```
par(mfrow=c(2, 2))
plot(crimeData$pctymle,crimeData$crmrte,main="Crime rate and pctymle",xlab="Percent Young Male",
     ylab ="Crime Rate")
abline(fit <- lm(crmrte ~ pctymle, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
plot(crimeData$pctymle,log(crimeData$crmrte),main="log Crime rate and pctymle",
     xlab="Percent Young Male",ylab ="log Crime Rate")
abline(fit <- lm(log(crmrte) ~ pctymle, data=crimeData), col='red')
legend("topright", bty="n", legend=paste("Adj R2:",
       format(summary(fit)$adj.r.squared, digits=4)))
hist(crimeData$pctymle,main="Histogram of pctymle", xlab="Percent Young Male",ylab = "frequency")
boxplot(crimeData$pctymle, main ="pctymle boxplot")
```



We observe that the relationship between percent young male and crime rate is weak and positive. The percent male variable is positively skewed. We see that there are quite a few outliers but the crime rate for those outliers are not high which makes us believe that young percent male in a county might not be a major driver of high crime rates.

EDA Summary

Table 2: Summary of EDA

Variable	Label	Expectation	Corr.Coef	EDA.Findings
crmrte	Crimes committed per person	None	NA	Crime rate has positive skew and log transformation can be used
prbarr	Probability of arrest	Negative relation with Crime rate	-0.395	Confirms the expectation
prbconv	Probability of conviction	Negative relation with Crime rate	-0.386	Confirms the expectation
prbpris	Probability of prison sentence	Negative relation with Crime rate	0.048	Doesn't have much affect.
avgsen	Avg. sentence, days	Negative relation with Crime rate	0.020	Doesn't have much affect.
polpc	Police per capita	Negative relation with Crime rate	0.167	Reverse affect than expectation
density	People per sq. mile	Positive relation with Crime rate	0.728	Confirms the expectation
west	1 if in western N.C.	None	-0.346	West has least crime rate
central	1 if in central N.C.	None	0.166	Central has average crime rate
urban	1 if in SMSA	None	0.615	Urban location has highest crime rate
pctmin80	Perc. minority, 1980	Not clear as what consitiutes minority is not given	0.182	Not much affect on crime rate.
Economic variables such as wages and tax		Negative relation with Crime rate	NA	Positive affect on crme rate . But negative affect on violent crimes.
pctymle	Percent young male	Positive relation with Crime rate as Crime is mostly associated with yoing males.	0.290	Positive affect but outliers don't have high crime rate.

Now we have a fair idea about the key variables. We would be creating three models. One with the variables which has been shown by our EDA to have strong relation with crime rate and no other covariates. Another which includes a few more covariates that we believe increases the accuracy of our results without introducing substantial bias. A third model includes all the variables to check if the model is robust to model specification.

Model Building

Model 1

In this model we will consider the crmrte with log transformation as our outcome variable and regressors as prbarr, prbconv, and density. We get from our EDA that these can be the top predictors. We are doing the log transformation of the outcome variable as it fits better in the linear model and also gives us the way to interpret the coefficients as a unit increase associated with a percentage change in crime rate which would be beneficial in policy making. We would also convert prbarr and prbconv in scale of 100 which would our model interpretation would be in percentage change in prbarr and prbconv instead of values of change in prbarr and prbconv. We call these prbarr_100 and prbconv_100 variable names. The population model we are considering is:

$$log(crmrte) = \beta 0 + \beta 1 * prbarr\_100 + \beta 2 * prbconv\_100 + \beta 3 density + u_i$$

```
crimeData$prbarr_100 <- 100*crimeData$prbarr
crimeData$prbconv_100 <- 100*crimeData$prbconv
model1 <- lm(log(crmrte)~prbarr_100+prbconv_100+density, data= crimeData)
coef(model1)
```

```
##  (Intercept)   prbarr_100  prbconv_100      density
## -3.033518362 -0.014339667 -0.005684893  0.159027074
```

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

```
## [1] 0.599675
```

```
AIC(model1)
```

```
## [1] 70.90214
```

```
Summary_Model1 <- read.csv("Model1_Summary - Sheet1.csv",
                          header = TRUE, sep = ",", quote = "\"",
                          allowEscapes = TRUE)
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote =
## quote, : incomplete final line found by readTableHeader on 'Model1_Summary
## - Sheet1.csv'
```

```
kable(Summary_Model1, "latex", longtable = TRUE, booktabs = TRUE,
      caption = "Model1 Coeffecient Explanations")%>%kable_styling(full_width = TRUE,
      latex_options =c("HOLD_position", "striped", "repeat_header"),
      row_label_position = 1)%>%column_spec(1,width="7em")%>%column_spec(2,width="7em")
```

Table 3: Model1 Coeffecient Explanations

Variable	Coefficient	Explanation
Intercept	-3.0335184	
prbarr_100	-0.0143397	One unit of increase in prbarr_100(about 1% change in prbarr) decreases the predicted value of crime rate by about 1.4% when everything else is kept same. We also saw from EDA that this variable has correlation coefficient of -0.395. This seems to be practically significant variable .
prbconv_100	-0.0056849	One unit of increase in prbarr_100(about 1% change in prbarr) decreases the predicted value of crime rate by about .5% when everything else is kept same.We also saw from EDA that this variable has correlation coefficient as -0.386.This seems to be practically significant variable
density	0.1590271	One unit of increase in density (1 person per square mile) increases the predicted value of crime rate by about 15.9% when everything else is kept same.We also saw from EDA that this variable has correlation coefficient as 0.728. This seems to be practically significant variable

We see from the adjusted R squared values that our model1 is able to explain 59.9% of the variations. Also the AIC score is 70.90214

Model 2

In this model we are going to add log(polpc) and pctymle and pctmin80 as other covariates. We believe that police per capita increases the probability of arrest and so including this as one of the covariates variables would get us the estimate of the coefficients for prbarr much closer to the real value and will have lesser bias. Also, since pctymle and pctmin80 have a positive association with crime rate, these variables would absorb some of the bias in the coefficients of prbarr,prbconv and density and bring the latter coefficients closer to their true, unbiased parameter estimates. The population model would be:

$$log(crmrte) = \beta 0 + \beta 1 prbarr\_100 + \beta 2 prbconv\_100 + \beta 3 density + \beta 4 log(polpc) + \beta 5 pctymle + \beta 6 pctmin80 + u_i$$

```
model2 <- lm(log(crmrte)~prbarr_100+prbconv_100+density+log(polpc)+pctymle+pctmin80,
             data= crimeData)
coef(model2)
```

```
##  (Intercept)   prbarr_100  prbconv_100      density   log(polpc)
## -0.020199123 -0.019299733 -0.006454535  0.110056897  0.486513215
##      pctymle     pctmin80
##  1.110446161  0.011459496
```

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

```
## [1] 0.783899
```

```
AIC(model2)
```

```
## [1] 18.21878
```

```
Summary_Model2 <- read.csv("Model2_Summary - Sheet1.csv",
                           header = TRUE, sep = ",", quote = "\"",
                           allowEscapes = TRUE)
kable(Summary_Model2, "latex", longtable = TRUE, booktabs = TRUE,
      caption = "Model2 Coeffecient Explanations")%>%kable_styling(full_width = TRUE,
      latex_options =c("HOLD_position", "striped", "repeat_header"),
      row_label_position = 1)%>%column_spec(1,width="7em")%>%column_spec(2,width="7em")
```

Table 4: Model2 Coeffecient Explanations

Variable	Coefficient	Explanation
Intercept	-0.0201991	
prbarr_100	-0.0192997	One unit of increase in prbarr_100(about 1% change in prbarr) decreases the predicted value of crime rate by about 1.9% when everything else is kept same. We also saw from EDA that this variable has correlation coefficient as -0.395. This seems to be practically significant variable
prbconv_100	-0.0064545	One unit of increase in prbarr_100(about 1% change in prbarr) decreases the predicted value of crime rate by about .6% when everything else is kept same. We also saw from EDA that this variable has correlation coefficient as -0.386.This seems to be practically significant variable

Table 4: Model2 Coeffecient Explanations *(continued)*

Variable	Coefficient	Explanation
density	0.1100569	One unit of increase in density (1 person per square mile) increases the predicted value of crime rate by about 11% when everything else is kept same.We also saw from EDA that this variable has correlation coefficient as 0.728. This seems to be practically significant variable
log(polpc)	0.4865132	1% increase in police per capita increases the predicted value of the crime rate by about 48% when everything else is kept same. We also saw from EDA that this variable has correlation coefficient as 0.167. This seems to be practically significant variable
pctymle	1.1104462	One unit of increase in pctymle(1% increase in percent young male population) increases crime rate by about 111% when everything else is kept same. We also saw from EDA that this variable has correlation coefficient as 0.290. This seems to be practically significant
pctmin80	0.0114595	One unit of increase in pctmin80(1% increase in minority population) increase predicted value of crime rate by about 1.1 % when everything else kept same. We also saw from EDA that this variable has correlation coefficient as 0.182. This also seems to be practically significant

We see from the adjusted R squared values that our model1 is able to explain 78.3% of the variation. Also the AIC score is 18.21878 which means that overall this model is better fit than model1 accounting for parsimony too.

Model 3

In this model we take almost all the variables available as predictor and check the model. Here, we drop the 84th observation given that it represents an extreme outlier in terms of wage in the service industry and might be suspect.

```
#Make sure to exclude data point for 84 as it has erroneous wage in service sector
model3 <- lm(log(crmrte)~prbarr_100+prbconv_100+density+log(polpc)+prbpris+avgsen+taxpc+
               urban+west+central+pctmin80 +pctyme + wcon +
               wtuc+wtrd+wfir+wser+wmfg+wfed+wloc+wsta,
             data=crimeData[-84,])
coef(model3)
```

```
##   (Intercept)    prbarr_100   prbconv_100       density    log(polpc)
## -1.0276175846 -0.0177873749 -0.0057278087  0.1102238390  0.4291253033
##       prbpris        avgsen         taxpc         urban          west
## -0.1138944757 -0.0122642827  0.0021372363 -0.1308776293 -0.1773651906
##       central      pctmin80       pctymle          wcon          wtuc
## -0.1365835081  0.0076426512  2.7422346679  0.0004897294  0.0002750242
##          wtrd          wfir          wser          wmfg          wfed
##  0.0006223103 -0.0009148068 -0.0018465547 -0.0000767829  0.0023177186
##          wloc          wsta
##  0.0015139071 -0.0010865219
```

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

```
## [1] 0.8192076
```

```
AIC(model3)
```

```
## [1] 11.97263
```

We see from the adjusted R-square and AIC of the model 3 that it seems to be a better fit than model1 and model2. While there are some small impacts on the coefficients of the variables included in model1 and model2, the directionality of the coefficients are preserved suggesting that the bias introduced by the inclusion of the variables is not strong. That said, we need to see the standard deviation of the errors to check if the coefficients are really statistically significant or not. That work will be done in Stage3.

Regression Table

Let us now create a regression table which will compare the three models. We will use the stargazer library.

```
stargazer(model1,model2,model3,
          report="vc",
          title="Linear Models Predicting log Crime rate",
          keep.stat = c("rsq","adj.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: Tue, Nov 27, 2018 - 1:28:50 PM

We see that across all the models the variables prbarr,prbconv,density and log(polpc) retain their direction and strength of association and therefore, model2 seems to be robust to changes in model specification.

Table 5: Linear Models Predicting log Crime rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
prbarr_100	−0.014	−0.019	−0.018
prbconv_100	−0.006	−0.006	−0.006
density	0.159	0.110	0.110
log(polpc)		0.487	0.429
prbpris			−0.114
avgsen			−0.012
taxpc			0.002
urban			−0.131
west			−0.177
central			−0.137
pctymle		1.110	2.742
wcon			0.0005
wtuc			0.0003
wtrd			0.001
wfir			−0.001
wser			−0.002
wmfg			−0.0001
wfed			0.002
wloc			0.002
wsta			−0.001
pctmin80		0.011	0.008
Constant	−3.034	−0.020	−1.028
Observations	90	90	89
R^2	0.613	0.798	0.862
Adjusted R^2	0.600	0.784	0.819

Omitted Variables

Table 6: Omitted Variables

Variable	Description	Bias.Inference	Possible.Proxy
Drugs Usage	Drugs in community has been stated as one of the reasons for incidence of crime, either directly due to the criminal sale and possession of illicit drugs or indirectly through drug-related criminal actions (i.e. robbery or assault for funds to purchase drugs). This variable can be defined as drug usage per capita which will be ration of count of drug users to population.	Drugs usage in general should have a positive relationship with crime rate . Drug usage might also have a positive relationship with percent young male, pctymle. So, with this variable omitted the coefficient on pctymle in our model specifications will have a positive bias. We may surmise that drug usage might also have negative relation with density if the propensity of drug usage is higher in suburban and rural counties, hence having a negative bias effect on the the density coefficient.	There is no proxy in the data set available
Poverty level	Poverty is generally thought one of the motivating reasons for crime. So the percentage of population under poverty would be a variable of interest.	Poverty level will have in general positive relation with crime rate and negative relation with tax per capita (taxpc). Therefore, there will be negative bias on the coefficient of taxpc due to this omitted variable.	A possible proxy would be the ratio of taxpc/density but that will be a weak proxy
Unemployment level	Unemployment is also thought of as one of the motivating reasons for crime. So the percentage of unemployed population would be a variable of interest.	Unemployment level could have a negative association with the wage and taxpc variables and, possibly, a negative relationship with density if we think unemployment is higher in rural counties. Unemployment level would have positive relation with crime rate. Hence the bias for the coefficients on density, wage variables and taxpc would be negative and possibly positive for the coefficient on urban.	There is no proxy available in the dataset.

Table 6: Omitted Variables *(continued)*

Variable	Description	Bias.Inference	Possible.Proxy
Youth employment opportunities	Limited youth employment opportunities is thought of as one of the motivating reasons for crime. Therefore, the number of youth employment opportunities per capita of young people would be a variable of interest.	Higher employment opportunities per youth capita will have a negative association with crime rate and possibly, positive association with the percent young male, wage, density and taxpc variables. Therefore, the bias on the coefficients for wage, density, pctymle and taxpc variables would likely be negative.	There is no proxy available in the dataset.
Education level	Education level is thought to have a negative impact on the crime rate. Education level can be calculated as years of education per capita or percentage of population with a college degree.	Education would have a negative relation with crime rate but positive relation with wages and taxpc. So if the education level is an omitted variable the bias on the coefficients of taxpc and wage variables would be negative.	There is no proxy available but taxpc and wage variables give a perspective but doesn't give full picture of education level
Social organizations working with at-risk people	Social organizations that provide supportive features for at-risk people in the community can serve as a deterrent to crime. This can be calculated as density of number of social organizations, such as religious institutions, YMCAs, or educational institutions that work with at-risk young people.	Social organizations working with at-risk youth are believed to have a negative relation with the crime rate and would have a positive relation with density and percent young male, pctymle. So, if this variable is omitted it will have negative bias on the density and pctymle coefficients.	There is no proxy available in the dataset which can be used for this purpose
Number of gangs	Gangs are thought of as a strong factor in the incidence of crime. Number of gangs in a county can be a variable which will have positive relation with crime rate	The number of gangs could have a postive relationship with density and a positive relation with the offense mix variable and percent young male. So, if this variable is omitted from the population model then the coefficients for density, offense mix and percent young male will have positive bias.	There is no proxy available in the dataset which can be used as gang activity. We can infer some from mix but that will not be sufficient.

Table 6: Omitted Variables *(continued)*

Variable	Description	Bias.Inference	Possible.Proxy
Aggressive police tactics policy	Aggressive police tactics policy might differ by county and represents how aggressive police are in issuing arrests or profiling certain segments of the population (i.e. young minority males). This could be a regime indicator variable taking on a value of 1 for a say the enforcement of a "broken windows" style policy that targets individuals more readily and 0 for no aggressive policing policy.	Aggressive police tactics policy might be positively associated with 'probability' of arrest, police per capita, percent young male and percent minority. Therefore, we might see a strong positive bias in the prbarr, polpc, pctymle and pctmin80.	We can infer a proxy through prbarr, yet that might be a weak proxy as it may not capture all of the impacts of aggressive police tactics policy.
Residential turnover	Residential turnover in highly transient communities is thought to be associated with higher crime rates. Percentage of population that turns over, or moves, would be a variable of interest.	High residential turnover might be positively associated with density and percent young male. Therefore, we might see a modest positive bias in the coefficients for the density and percent young male variables due to the residential turnover.	There is no proxy available in the dataset.
Number of "hot spots" (micro places)	"Hot spots" or micro places can be described as street intersections or segments (two block faces on both sides of a street). These can be tied to bus stops, a busy street or lack of lighting that lend themselves to obscurity.	The number of 'hot spots' would have a positive relation with crime rate. This could also have a positive relationship with density, urban and offense mix. Therefore, there should be a modest to substantial positive bias on the density, urban and offense mix variables' coefficients due to the number of 'hot spots'.	There is no proxy available in the dataset.

Table 6: Omitted Variables *(continued)*

Variable	Description	Bias.Inference	Possible.Proxy
Immigration levels	Immigration, particularly recent or relatively new arrivals of immigrants, tends to be associated with lower incidence of crime. One potential reason might be motivation to work and greater ambition to seek quality of life improvements.	Immigration levels would have a negative relation with crime rate. This could also have a positive relationship with density and wage variables, assuming immigrants tend to immigrate to where relative wages are higher. Therefore, there should be modest positive bias on density and wage variable due to the omission of immigration levels.	There is no proxy available in the dataset.

Conclusion

From our analysis we offer the following policy perspectives:

1) We see that the probability of arrest is a deterrent to crime. So we propose policy to use police more effectively. The police should increase its presence in the "hot spot" areas where crime is more prone to occur and at times of the day when the incidence of crime is greater. This could be increasing presence in busy street intersections or not well lit street segments where particular type of crime might occur. This approach uses Data Science to make vigilance more effective.

2) We also see that the probability of conviction is deterrent to crime. We suggest policy which makes judiciary system to be swift in convicting criminals such as increasing the number of session judges and increasing the working hours for judiciary staff from 8 hours to 9 hours a day.

3) The population density is a large predictor of crime rates. Therefore, we propose policies where different counties can be made equally attractive to live in, such as making sure that jobs are evenly spread across counties. However, we think the coefficient on density is likely biased negatively by omitted variables such as gang activity and residential turnover and biased positively by social organizations that work with at-risk youth. Therefore, policies that promote social organizations that work with at-risk youth, promote residential stability and target the formation of gangs would also reduce crime rates through the population density variable.

4) We see that the crime rate is highest in urban area and has highest density too. We propose an effective transportation system so that people can travel to urban areas to work while living in sub urban areas which will negative effect on density in urban area and hence would effectively control crime. However, as mentioned above, we think the coefficient on density is likely biased negatively by omitted variables such as gang activity and residential turnover and biased positively by social organizations that work with at-risk youth.

5) To truly identify how policing effects crime rates, we might want to explore how counties differ in their policing aggressiveness policies (e.g. 'broken windows' policies). Exogenous differences in policing regimes could serve as a potential instrumental variable as part of an identification strategy to tease out the endogeneity and measure a causal effect from police per capita on crime rates.

6) Following the discussion in the omitted variables section, policies to promote lower drug usage, poverty alleviation, greater youth employment opportunities, higher levels of education among the population, greater formation of social organizations working with at-risk youth, target incipient gang formation, promote residential stability and higher immigration levels can all serve to reduce crime rates.