ReducingCrime

w203 Lab3

Harith Elrufai and Gaurav Desai

3/17/2018

Setup

```
rawCrimeData = read.csv("crime_v2.csv")
```

Data Cleanup and reformatting

```
str(rawCrimeData) # 97 obs, 25 variables. This is odd, if i open file in excel i see 92 rows with heade
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","0.068376102",..: 62 88 12 61 51 2 58 77 41 85 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

```
#summary(rawCrimeData)
```

drop NAs

```
crimeData <- rawCrimeData[!is.na(rawCrimeData$county),]
```

Convert county to factor as it is not a measurement

```
#county
crimeData$county <- as.factor(crimeData$county)
length(levels(crimeData$county)) #=> 90
```

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

This is interesting, we have 91 rows but only 90 levels. Eyeballing the data shows there are two identicle rows for county 193, same can be varifired using duplicated function.

```
crimeData[duplicated(crimeData),]
```

```
##    county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
## 89    193   87 0.0235277 0.266055 0.588859022 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 89 0.07819394
```

so lets delete the duplicate row

```
crimeData <- crimeData[!duplicated(crimeData),]
```

Now lets see if counties are exclusively marked as west or central or urban

```
nrow(crimeData[crimeData$west+crimeData$central+crimeData$urban > 1,])
```

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

```
nrow(crimeData[crimeData$west+crimeData$central+crimeData$urban == 0,])
```

```
## [1] 33
```

so there are 7 counties which are under more than 1 category and 33 without any category. Lets create new varible "region" combining these three variables

```
crimeData$region<-ifelse(crimeData$west==1,"West",NA)
crimeData$region<-ifelse(crimeData$central==1,"Central",crimeData$region)
crimeData$region<-ifelse(crimeData$urban==1,"Urban",crimeData$region)
crimeData$region<-ifelse(crimeData$west+crimeData$central+crimeData$urban > 1,"Mixed",crimeData$region)
crimeData$region<-ifelse(crimeData$west+crimeData$central+crimeData$urban == 0,"None",crimeData$region)
crimeData$region <- as.factor(crimeData$region)
summary(crimeData$region)
```

```
## Central   Mixed    None   Urban    West
##      28       7      33       2      20
```

A final check to see if there are any more NAs left in the data

```
crimeData[!complete.cases(crimeData),]
```

```
##  [1] county   year     crmrte   prbarr   prbconv  prbpris  avgsen
##  [8] polpc    density  taxpc    west     central  urban    pctmin80
## [15] wcon     wtuc     wtrd     wfir     wser     wmfg     wfed
## [22] wsta     wloc     mix      pctymle  region
## <0 rows> (or 0-length row.names)
```

Now lets convert prbconv from factor to number because it is a probability value.

```
crimeData$prbconv <- as.numeric(levels(crimeData$prbconv))[crimeData$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```
summary(crimeData$prbconv)
```

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

Now lets see if any of the probability is crossing 0 to 1 range

```
filter(crimeData, prbarr< 0 | prbarr>1 | prbconv < 0 | prbconv > 1 | prbpris < 0 | prbpris > 1) [,c("cou
```

```
##    county   prbarr  prbpris
## 1       3 0.132029 0.450000
## 2      19 0.162860 0.333333
## 3      99 0.153846 0.556962
## 4     115 1.090910 0.500000
## 5     127 0.179616 0.335616
## 6     137 0.207143 0.322581
## 7     149 0.271967 0.227273
## 8     185 0.195266 0.442857
## 9     195 0.201397 0.470588
## 10    197 0.207595 0.360825
```

We have 10 counties where prbconv is greater than 1 which means there are more convictions than arrests. Out of these 10 counties, one county 115 also has prbarr greater than 1 indicating more arrests than offences. We have two ways to clean this data, either we remove these 10 counties or we cap the max probabilities at 1. For this analysis we take second approach of capping the probabilities at 1.

```
crimeData$prbconv <- ifelse(crimeData$prbconv>1,1,crimeData$prbconv)
crimeData$prbarr <- ifelse(crimeData$prbarr>1,1,crimeData$prbarr)
summary(crimeData$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20495 0.27146 0.29423 0.34487 1.00000
```

```
summary(crimeData$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34422 0.45170 0.50799 0.58513 1.00000
```

```
summary(crimeData$prbpris)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1500  0.3642  0.4222  0.4106  0.4576  0.6000
```

```
#summary(crimeData)
str(crimeData)
```

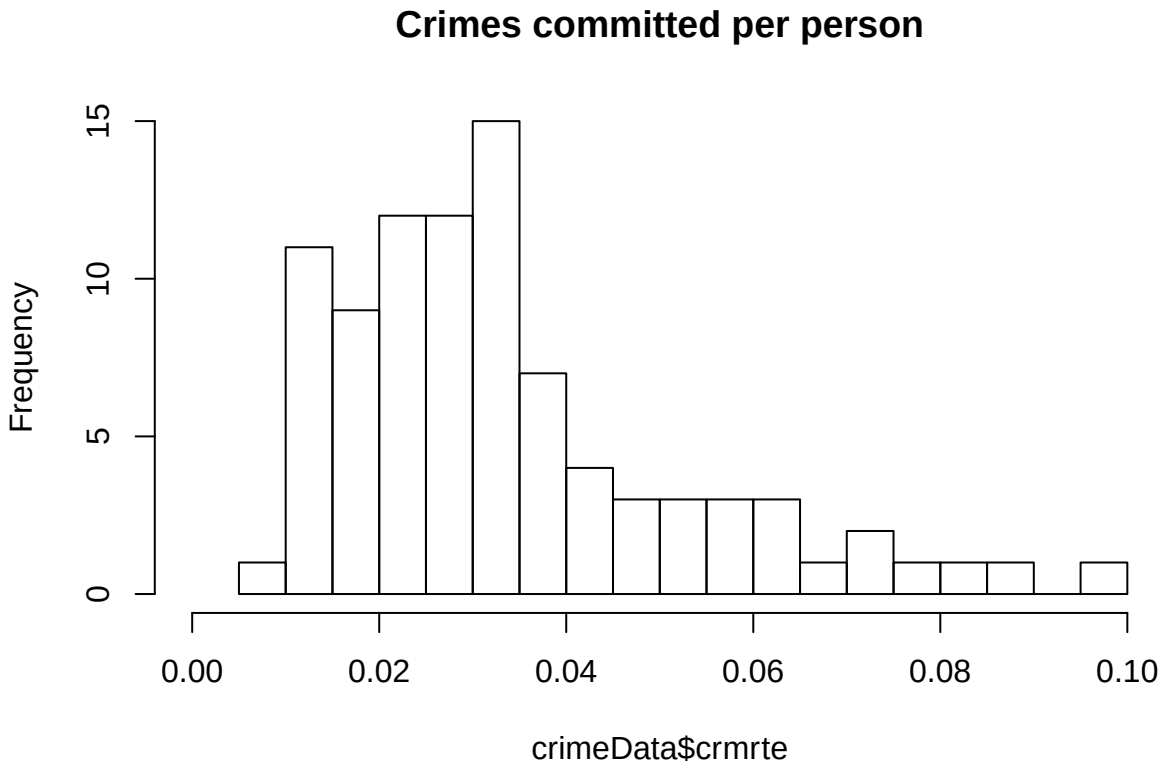
```
## 'data.frame':    90 obs. of  26 variables:
##  $ county  : Factor w/ 90 levels "1","3","5","7",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : num  0.528 1 0.268 0.525 0.477 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
```

```
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
##  $ region  : Factor w/ 5 levels "Central","Mixed",..: 1 1 5 1 5 5 3 3 3 3 ...
```

Analysis of Key variables

crmrte

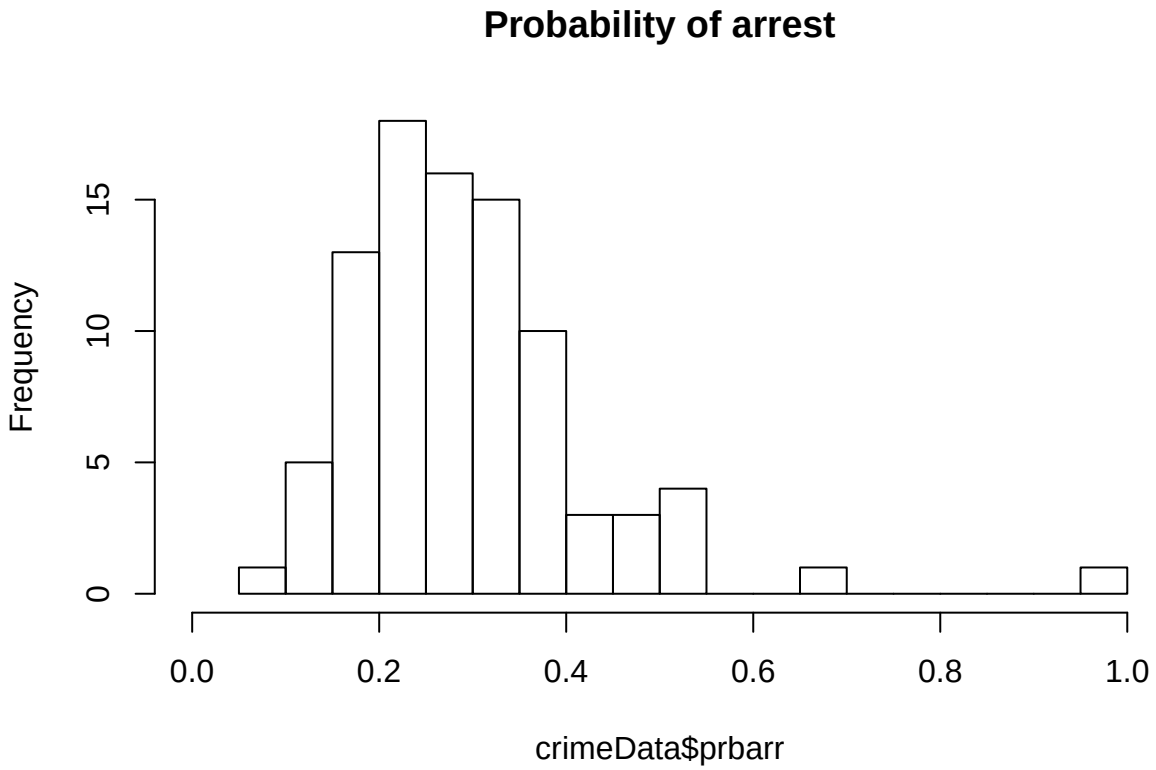
```
hist(crimeData$crmrte, breaks=20,xlim=range(0.0,.1) ,main = "Crimes committed per person")
```



Large chunk of counties have crime rate less than 4%. Nothing suspicious in here.

prbarr

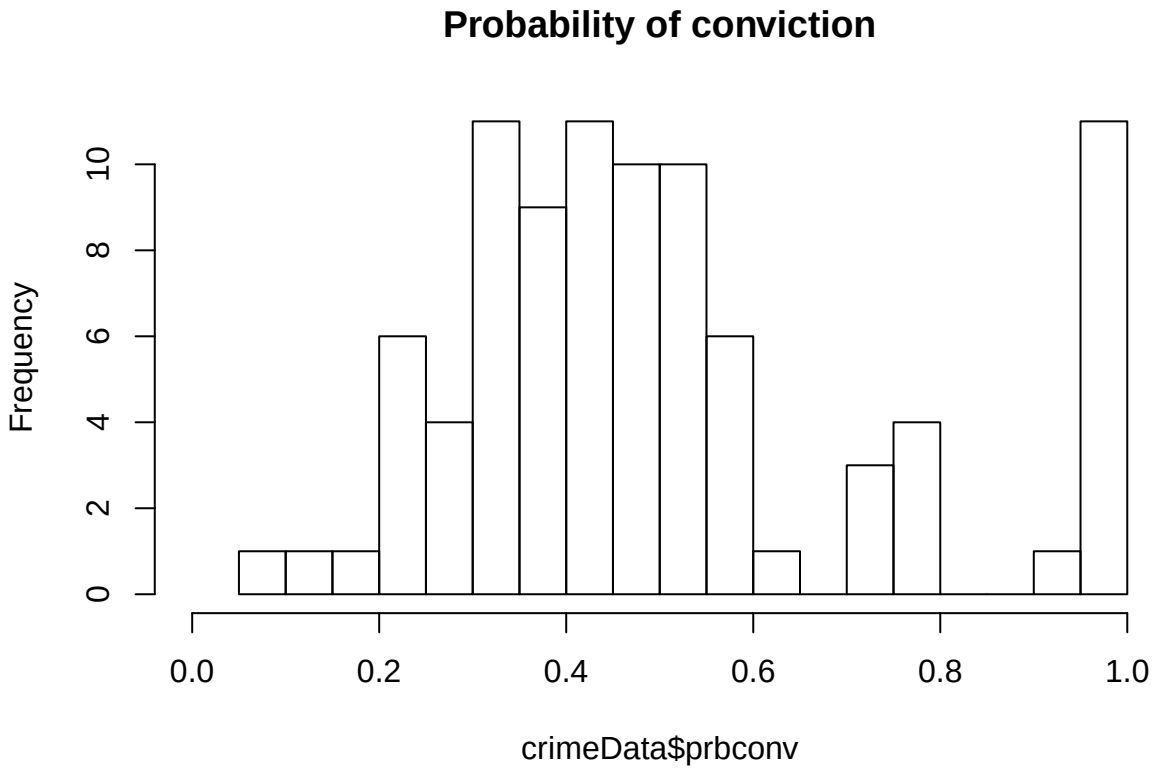
```
hist(crimeData$prbarr,main = "Probability of arrest", breaks = 20, xlim=range(0,1))
```



Majority of counties have probability of arrest less than 50% with couple of outliers at 70 and 100%. These two look interesting and should be further probed. Similarly counties with very low rate of arrest (<10%) should also be probed further.

prbconv

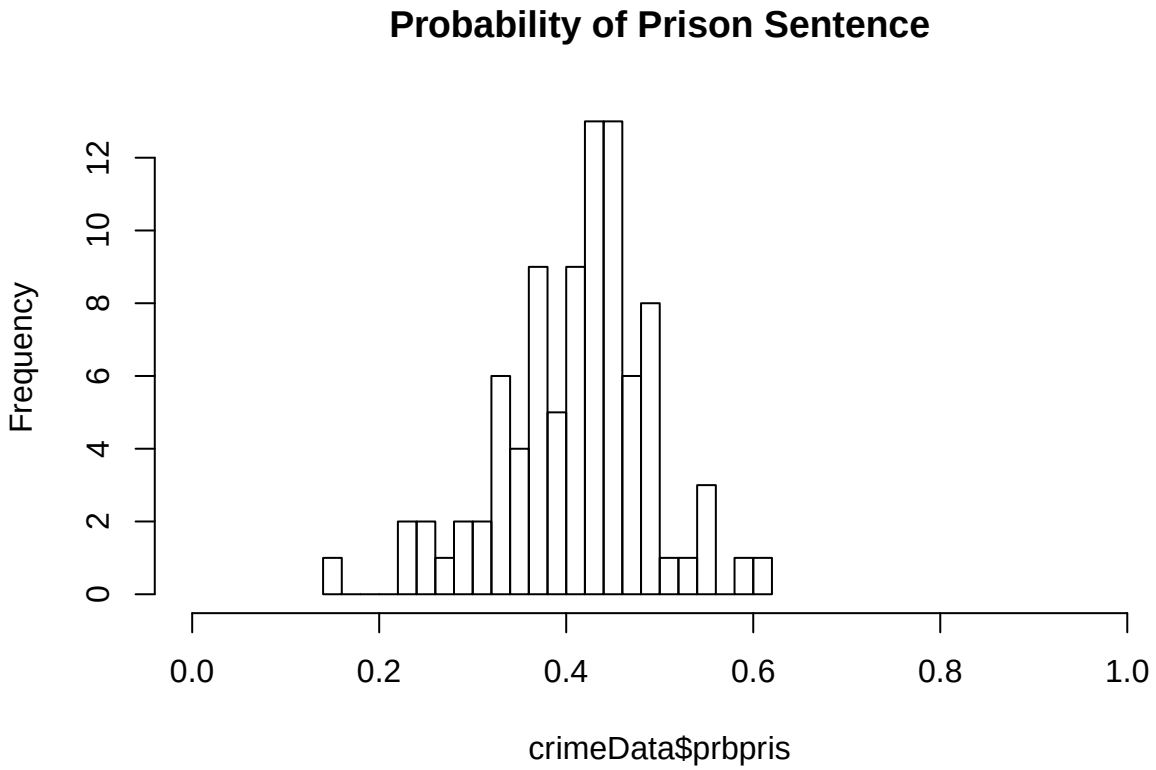
```
hist(crimeData$prbconv,main = "Probability of conviction", breaks = 20, xlim=range(0,1))
```



Again here we have couple of outliers with very high conviction rate(90 and 100%). This should be looked into. Similarly on lower end we have 3 counties with less than 20% conviction rate which again should be looked into.

prbpris

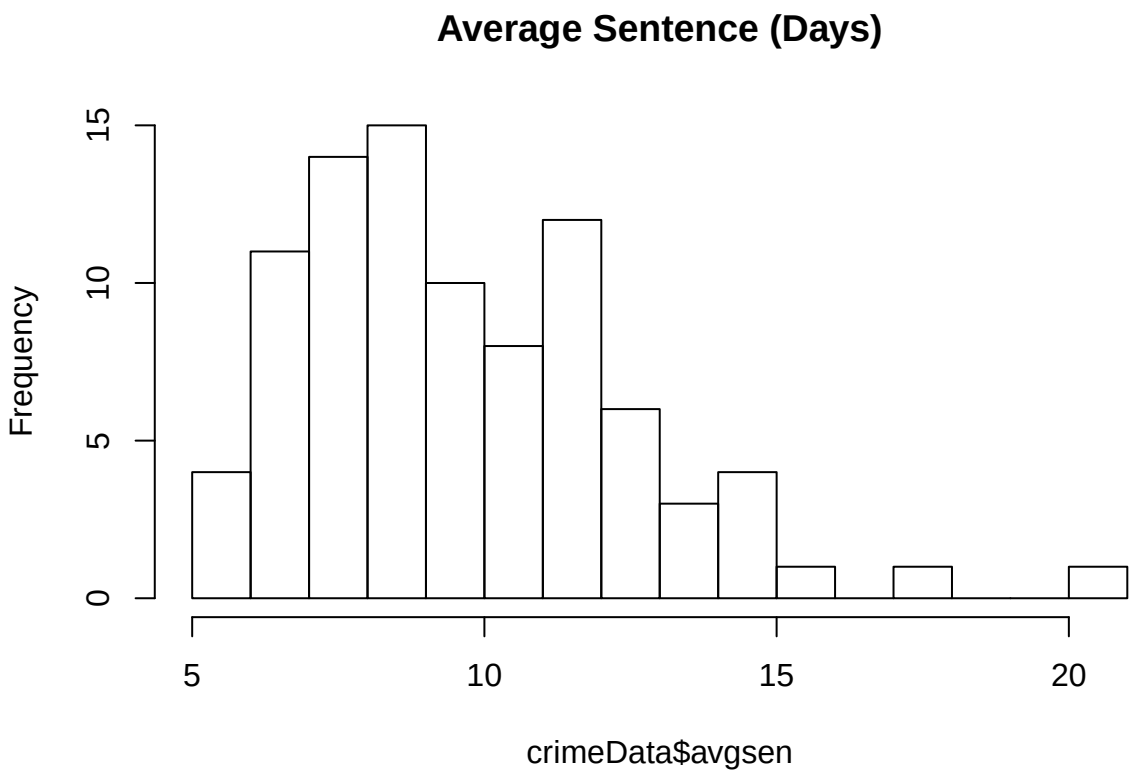
```
hist(crimeData$prbpris,main = "Probability of Prison Sentence", breaks = 20, xlim=range(0,1))
```



There is one county with less than 20% probability of sentence. This should be looked into for further investigation. There are couple of counties with more than 50% prison sentence. May be we can have alook for further probe.

avgsen

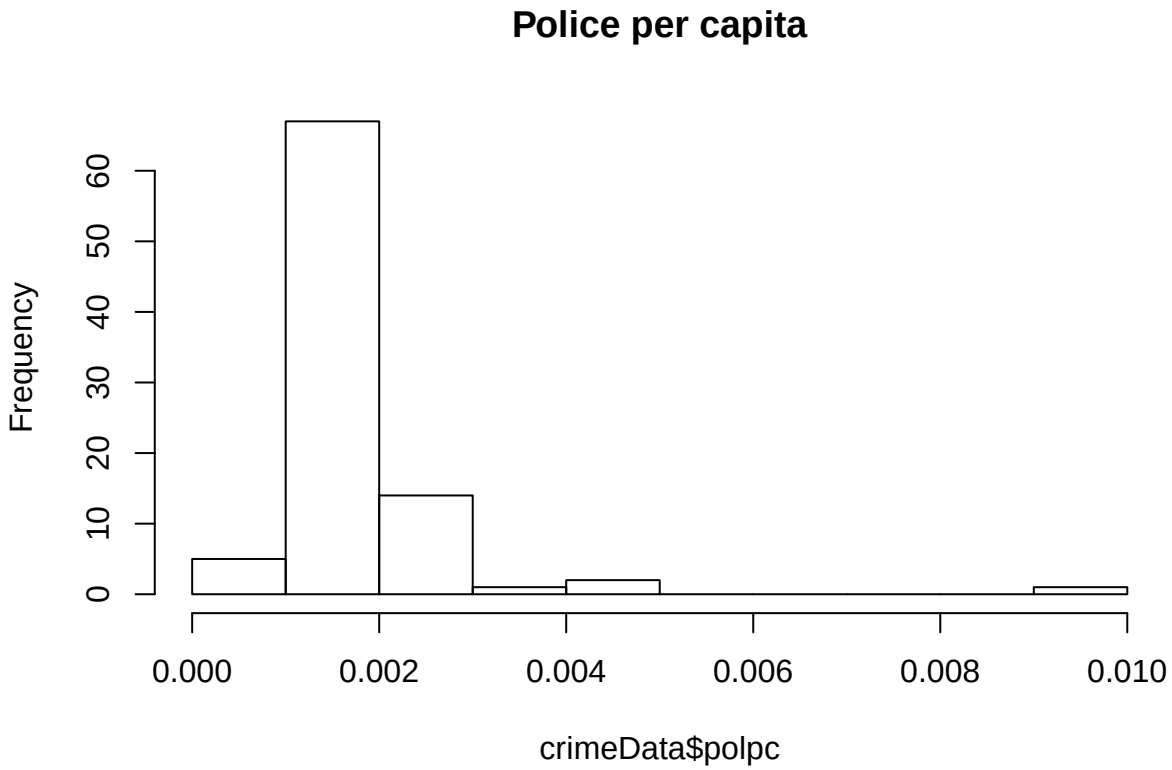
```
hist(crimeData$avgsen,main = "Average Sentence (Days)", breaks=20, xlim=range(5,21))
```



There is one outlier with 20 days of average sentence, much higher than rest of the counties.

polpc

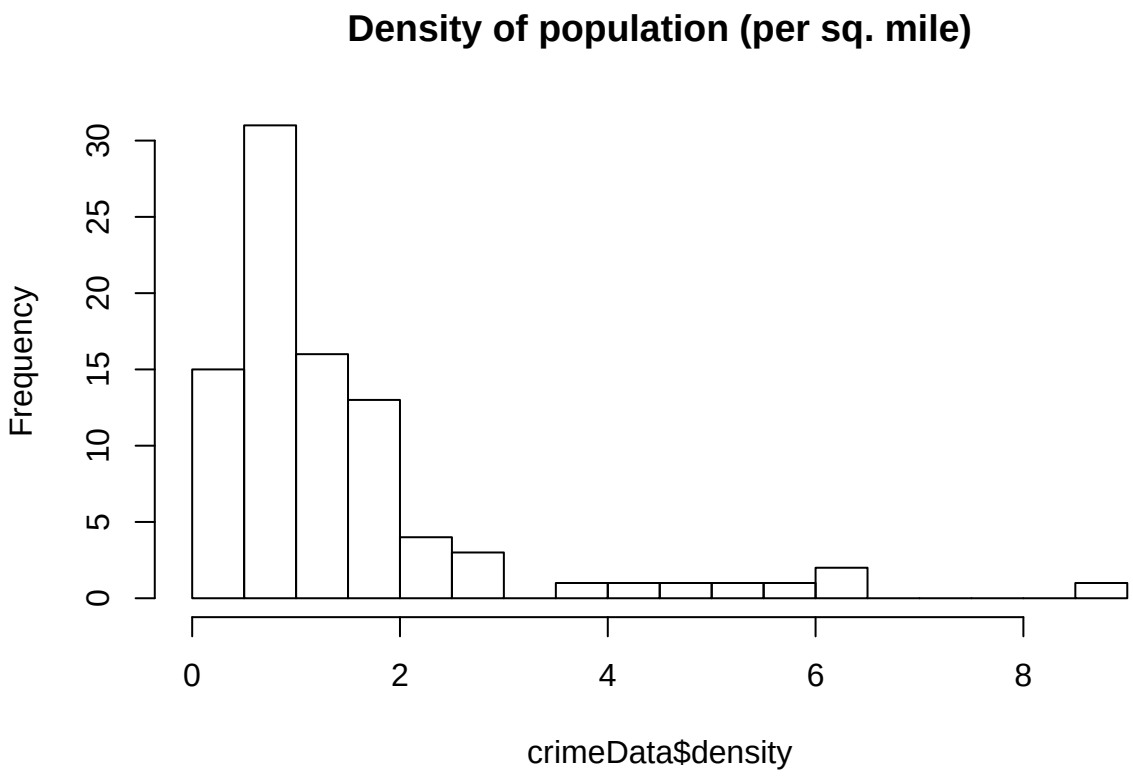
```
hist(crimeData$polpc,main = "Police per capita")
```



There are few counties with less than 0.001 Police per capita. This is alarming and should be checked in conjuction with other variables like crime rate, density etc.

density

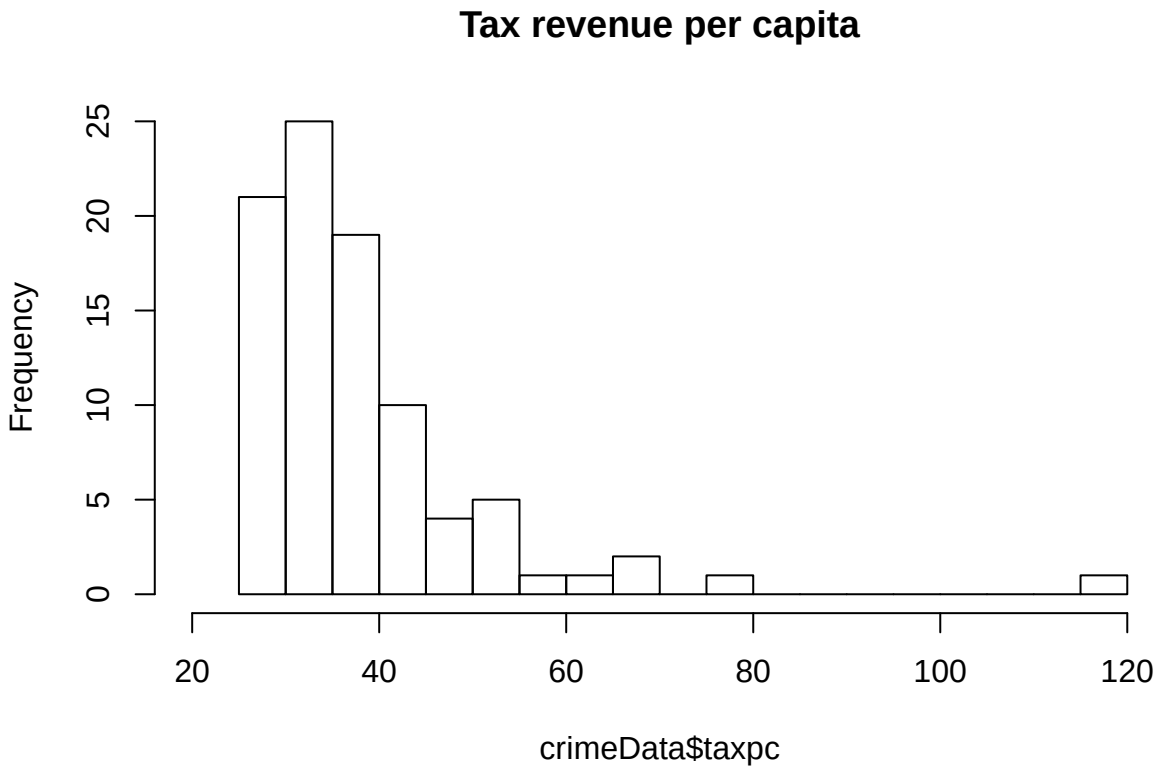
```
hist(crimeData$density,main = "Density of population (per sq. mile)", xlim=range(0,9), breaks=20)
```



Outliers with very high population density should be checked against other key variables like polic eper capita and crime rate.

taxpc

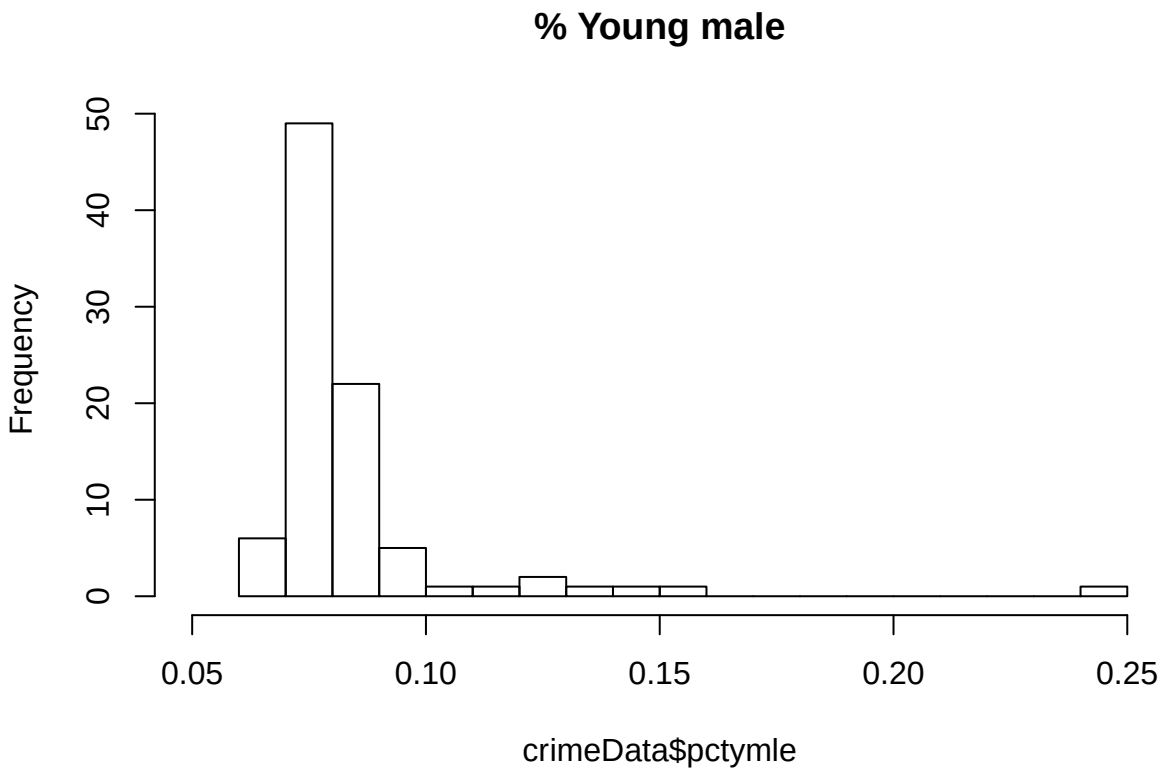
```
hist(crimeData$taxpc,main = "Tax revenue per capita", xlim=range(20,120), breaks=20)
```



There is one outlier county with very high tax revenue per capita. It will be interesting to see crime details for this county.

pctymle

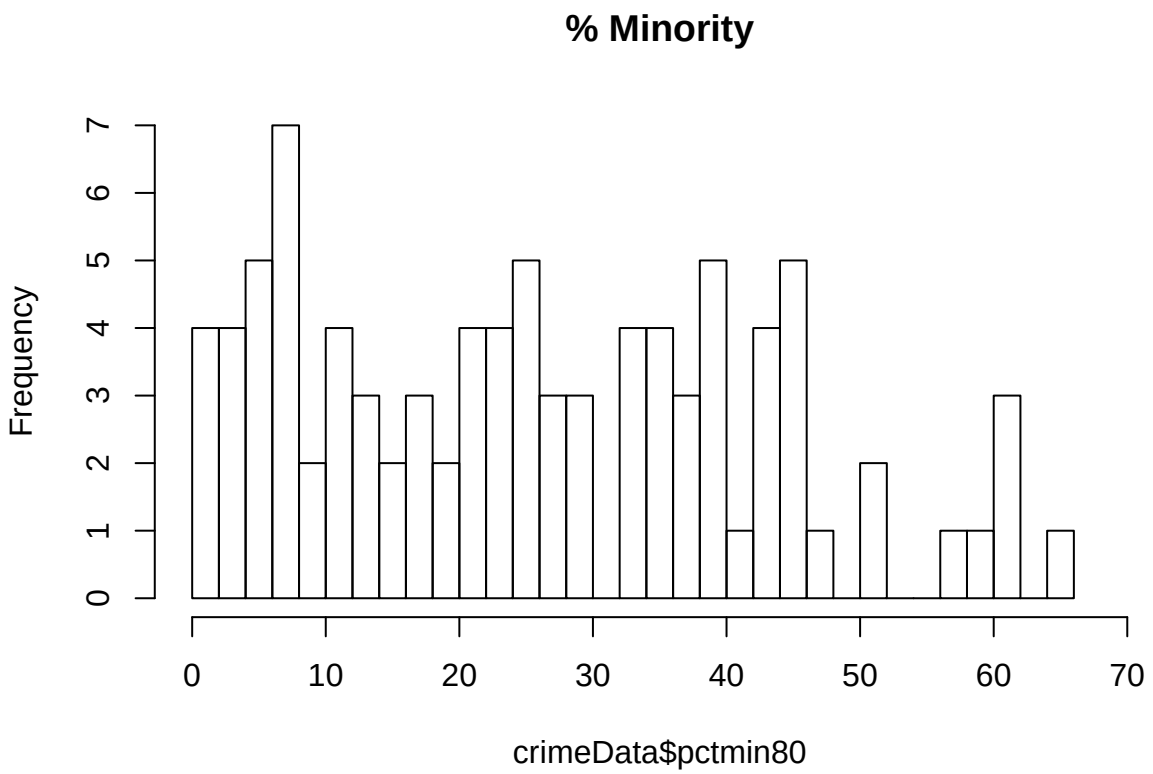
```
hist(crimeData$pctymle,main = "% Young male", breaks=20, xlim=range(.05,.25))
```



There is one outlier with almost 25% young male population. It will be interesting to see the effect on wages and crime rate.

pctmin80

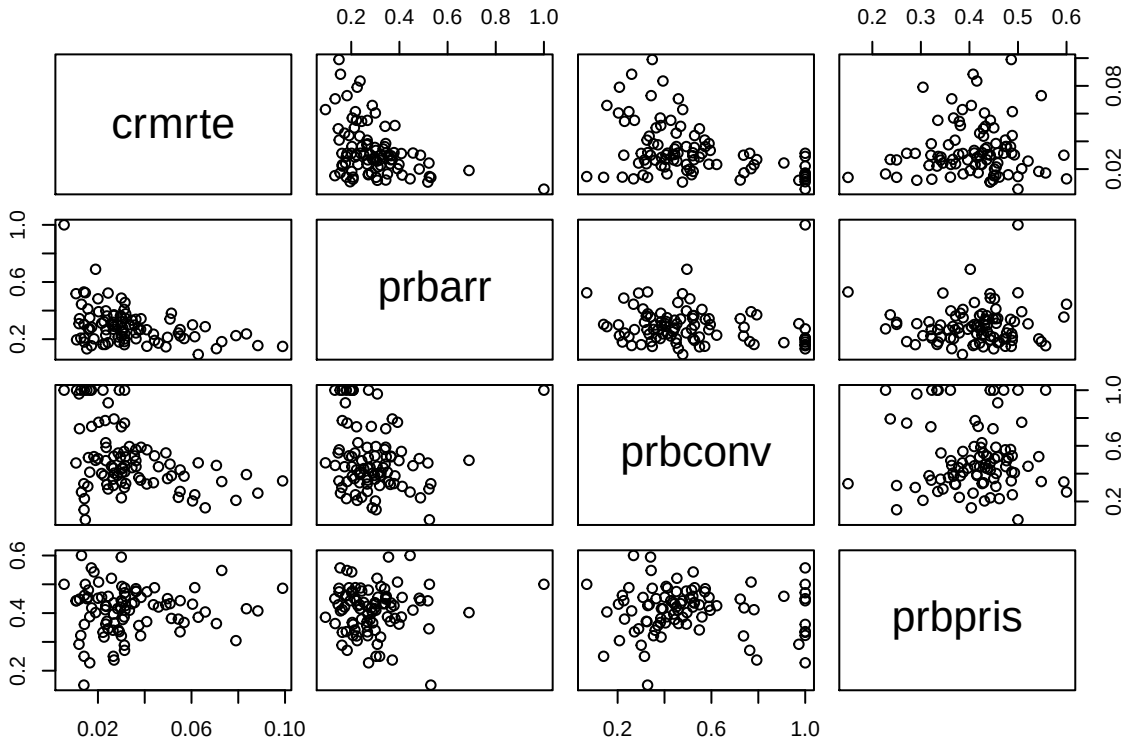
```
hist(crimeData$pctmin80,main = "% Minority", breaks=33, xlim=range(0,70))
```



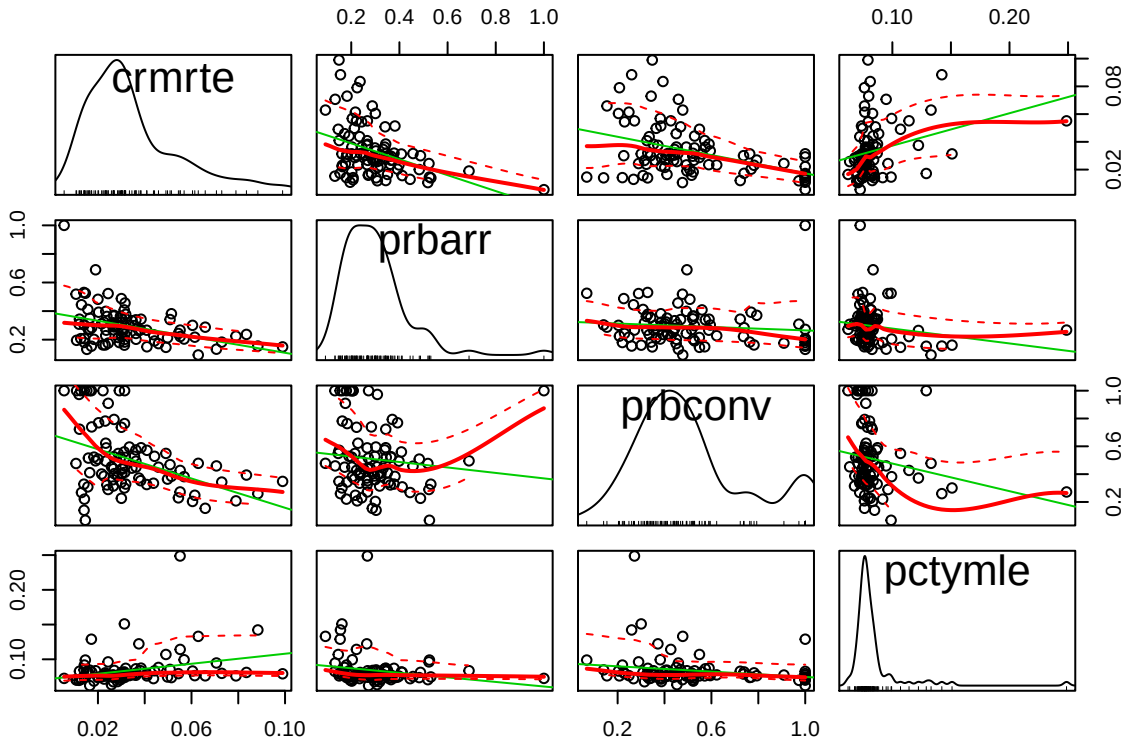
Counties with more than 50% monirity and less than 5% minority can be scrutinized further wrt wages and crime variables to see if there is any relation.

Model 1

```
pairs(crmrte~prbarr+prbconv+prbpris, data=crimeData)
```



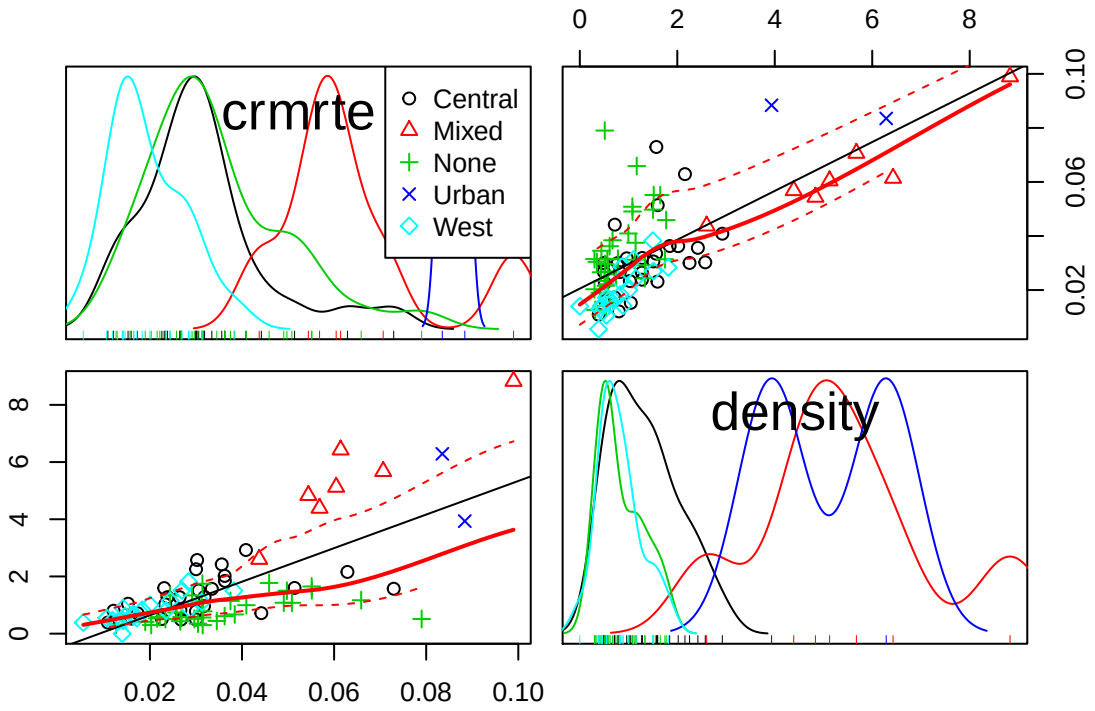

```
scatterplotMatrix(~crmrte+prbarr+prbconv+pctymle, data=crimeData)
```




```
#We can clearly see the relation betwene probbaility of arrest and conviction on crime rate.
# % of male population also affects crime rate but it appears independent on convition and arrest varia
```

Model 2

```
scatterplotMatrix(~crmrte+density|region, data=crimeData)
```




```
#density increases the region is urban where more crime
```