Exploratory Analysis by Aaron

August 17, 2017

Summary of Findings from EDA

Types of Variables:

1. Rates, averages, and probabilities - crmrte, prbarr, prbconv, prbpris, polpc, density, pctmin80, pctymle
2. $ variables - taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc
3. Indicator variables - west, central, urban. No base category in dataset (e.g. non-west/central, rural)
4. Not sure what this is - mix

Potential Transformations:

1. Percentage minority (pctmin80) data is between 0 - 100 whereas other percentage variables is between 0-1; We may want to transform this variable to keep things consistent from an interpretability perspective.
2. Other than this, so far no obvious variable that needs transformation. There are some variables that maybe needs log transform to be less skewed.

Potential Outliers/Data Issues:

1. Row 51 has prbarr > 1.0 which seems suspicious, given prbarr should between 0 and 1.
2. There seems to be 10 rows where prbconv is > 1.0, which again is suspicious, given that prbconv should be between 0 and 1.
3. Average Sentence in Days (avgsen) looks like it should be Average Sentence in Years.
4. Row 81 looks like an outlier where wser is extremely high.

Setup

Reading the data and loading the right libraries:

```
library(corrplot)
library(car)

setwd("C:\\Users\\aayuen\\Documents\\GitHub\\w203_lab4_kka")
data = read.csv("crime.csv")
```

There are 90 data points and 25 variables

```
nrow(data)
```

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

```
colnames(data)
```

```
##  [1] "X"        "county"   "year"     "crmrte"   "prbarr"   "prbconv"
##  [7] "prbpris"  "avgsen"   "polpc"    "density"  "taxpc"    "west"
## [13] "central"  "urban"    "pctmin80" "wcon"     "wtuc"     "wtrd"
## [19] "wfir"     "wser"     "wmfg"     "wfed"     "wsta"     "wloc"
## [25] "mix"      "pctymle"
```

There doesn't seem to be any NAs in the dataset

```
apply(!is.na(data[,]), MARGIN = 2, mean)
```

```
##        X   county     year   crmrte   prbarr  prbconv  prbpris   avgsen
##        1        1        1        1        1        1        1        1
##    polpc  density    taxpc     west  central    urban pctmin80     wcon
##        1        1        1        1        1        1        1        1
##     wtuc     wtrd     wfir     wser     wmfg     wfed     wsta     wloc
##        1        1        1        1        1        1        1        1
##      mix  pctymle
```

```
##        1        1
```

Univariate Variable Analysis

1. county - County Identifier

County is essentially a unique identifier (no duplicates).

```
length(unique(data$county))
```

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

```
length(data$county)
```

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

2. year - Only 87

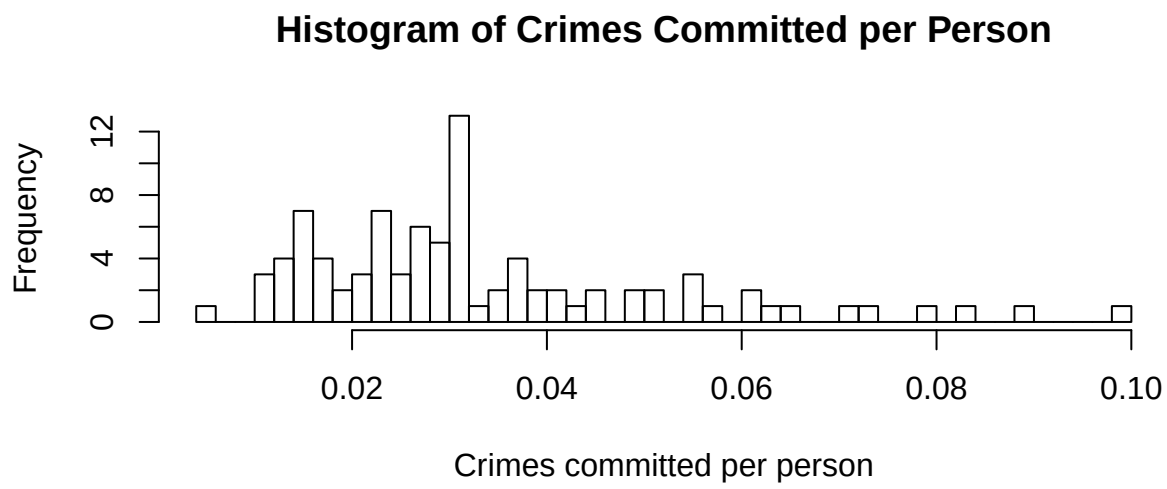
The dataset only contains data for 1987.

```
unique(data$year)
```

```
## [1] 87
```

3. crmrte - Crimes Committed per Person

```
hist(data$crmrte, breaks=50,
     main="Histogram of Crimes Committed per Person",
     xlab="Crimes committed per person")
```



4. prbarr - Probability of Arrest

Row 51 has prbarr > 1.0 which seems suspicious, given prbarr should between 0 and 1.

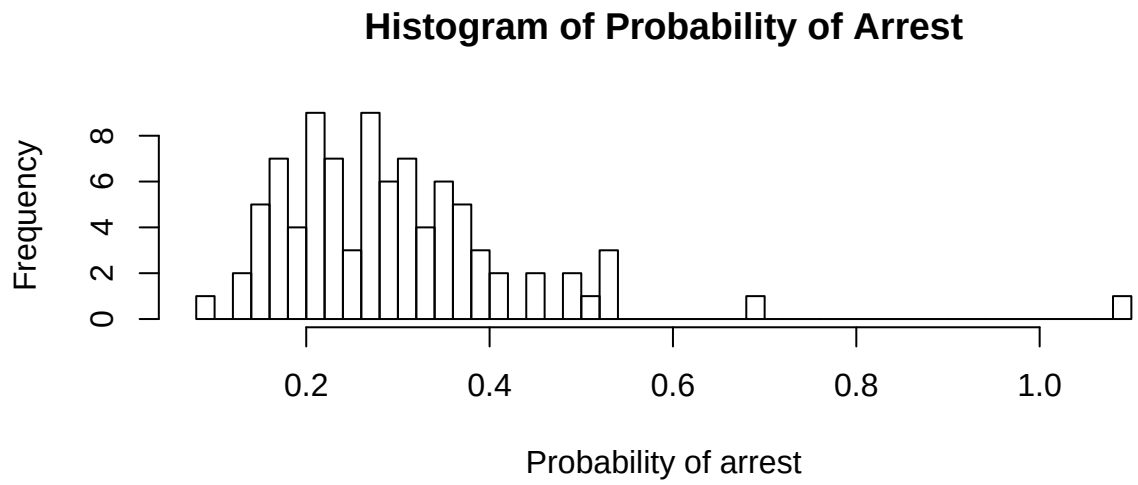
```
summary(data$prbarr)
```

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

```
data[data$prbarr > 1,]
```

```
##     X county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

```
hist(data$prbarr, breaks=50,
     main="Histogram of Probability of Arrest",
     xlab="Probability of arrest")
```



5. prbconv - Probability of Conviction

There seems to be 10 rows where prbconv is > 1.0, which again is suspicious, given that prbconv should be between 0 and 1.

```
summary(data$prbconv)
```

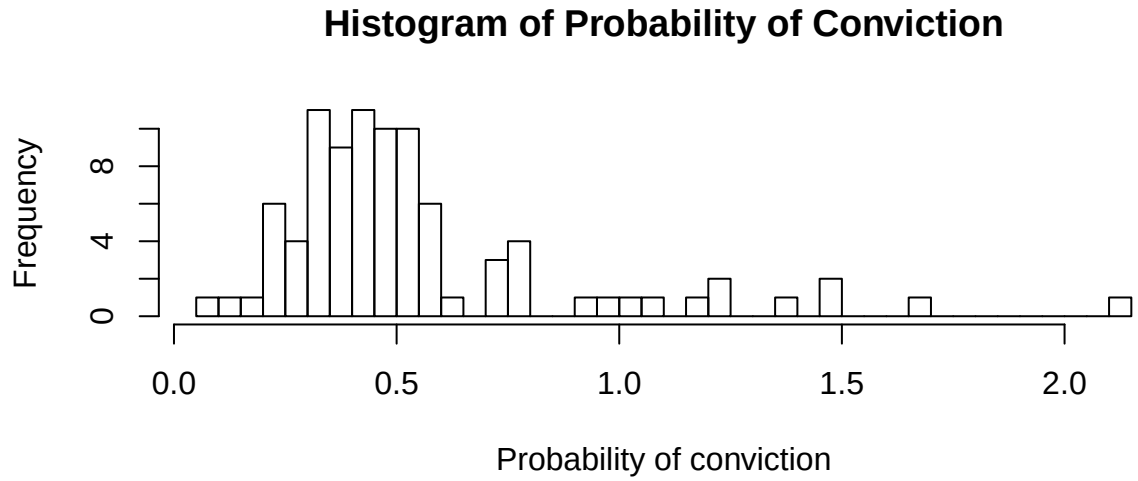
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34422 0.45170 0.55086 0.58513 2.12121
```

```
data[data$prbconv > 1,]
```

```
##     X county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 2   2      3   87 0.0152532 0.132029 1.48148 0.450000   6.35 0.00074588
## 10 10     19   87 0.0221567 0.162860 1.22561 0.333333  10.34 0.00202425
## 44 44     99   87 0.0171865 0.153846 1.23438 0.556962  14.75 0.00185912
## 51 51    115   87 0.0055332 1.090910 1.50000 0.500000  20.70 0.00905433
## 56 56    127   87 0.0291496 0.179616 1.35814 0.335616  15.99 0.00158289
## 61 61    137   87 0.0126662 0.207143 1.06897 0.322581   6.18 0.00081426
## 67 67    149   87 0.0164987 0.271967 1.01538 0.227273  14.62 0.00151871
## 84 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.00122210
## 89 89    195   87 0.0313973 0.201397 1.67052 0.470588  13.02 0.00445923
## 90 90    197   87 0.0141928 0.207595 1.18293 0.360825  12.23 0.00118573
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 2  1.0463320 26.89208    0       1     0  7.91632 255.1020 376.2542
## 10 0.5767442 61.15251    0       0     0 24.31170 260.1381 613.2261
## 44 0.5478615 39.57348    1       0     0 14.28460 259.7841 417.2099
## 51 0.3858093 28.19310    1       0     0  1.28365 204.2206 503.2351
```

```
## 56 1.3388889 32.02376    0       0     0 34.27990 290.9091 426.3901
## 61 0.3167155 44.29367    0       0     0 33.04480 299.4956 356.1254
## 67 0.6092437 29.03402    1       0     0 10.00460 223.6136 437.0629
## 84 0.3887588 40.82454    0       1     0 64.34820 226.8245 331.5650
## 89 1.7459893 53.66693    0       0     0 37.43110 315.1641 377.9356
## 90 0.8898810 25.95258    1       0     0  5.46081 314.1660 341.8803
##        wtrd     wfir      wser   wmfg   wfed   wsta   wloc        mix
## 2  196.0101 258.5650  192.3077 300.38 409.83 362.96 301.47 0.03022670
## 10 191.2452 290.5141  266.0934 567.06 403.15 258.33 299.44 0.05334728
## 44 168.2692 301.5734  247.6291 258.99 442.76 387.02 291.44 0.01960784
## 51 217.4908 342.4658  245.2061 448.42 442.20 340.39 386.12 0.10000000
## 56 257.6008 441.1413  305.7612 329.87 508.61 380.30 329.71 0.06305506
## 61 170.8711 170.9402  250.8361 192.96 360.84 283.90 321.73 0.06870229
## 67 188.7683 353.2182  210.4415 289.43 421.34 342.92 301.23 0.11682243
## 84 167.3726 264.4231 2177.0681 247.72 381.33 367.25 300.13 0.04968944
## 89 246.0614 411.4330  296.8684 392.27 480.79 303.11 337.28 0.15612382
## 90 182.8020 348.1432  212.8205 322.92 391.72 385.65 306.85 0.06756757
##       pctymle
## 2  0.08260694
## 10 0.07713232
## 44 0.12894706
## 51 0.07253495
## 56 0.07400288
## 61 0.07098370
## 67 0.06215772
## 84 0.07008217
## 89 0.07945071
## 90 0.07419893
```

```
hist(data$prbconv, breaks=50,
     main="Histogram of Probability of Conviction",
     xlab="Probability of conviction")
```

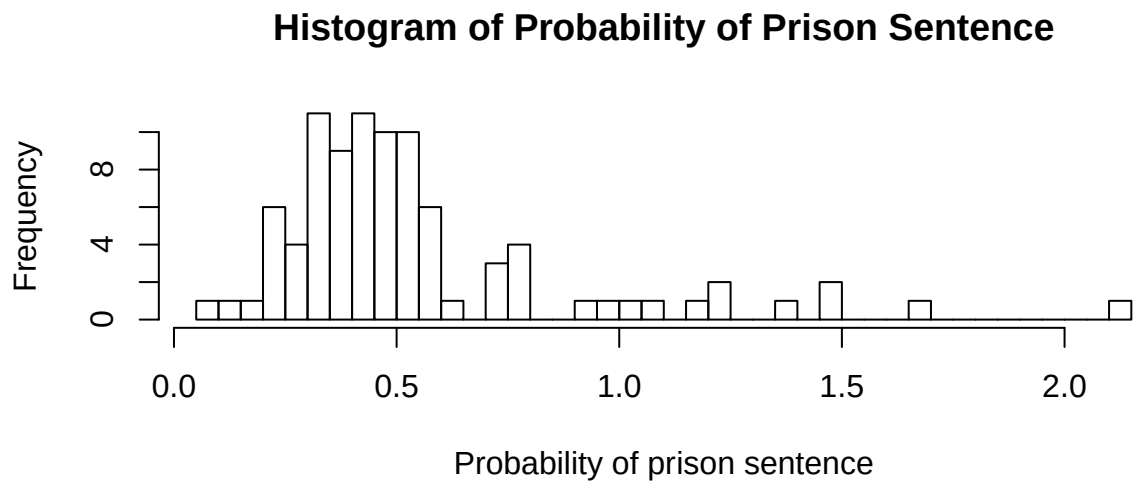


6. prbpris - Probability of Prison Sentence

```
summary(data$prbpris)
```

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

```r
hist(data$prbconv, breaks=50,
     main="Histogram of Probability of Prison Sentence",
     xlab="Probability of prison sentence")
```



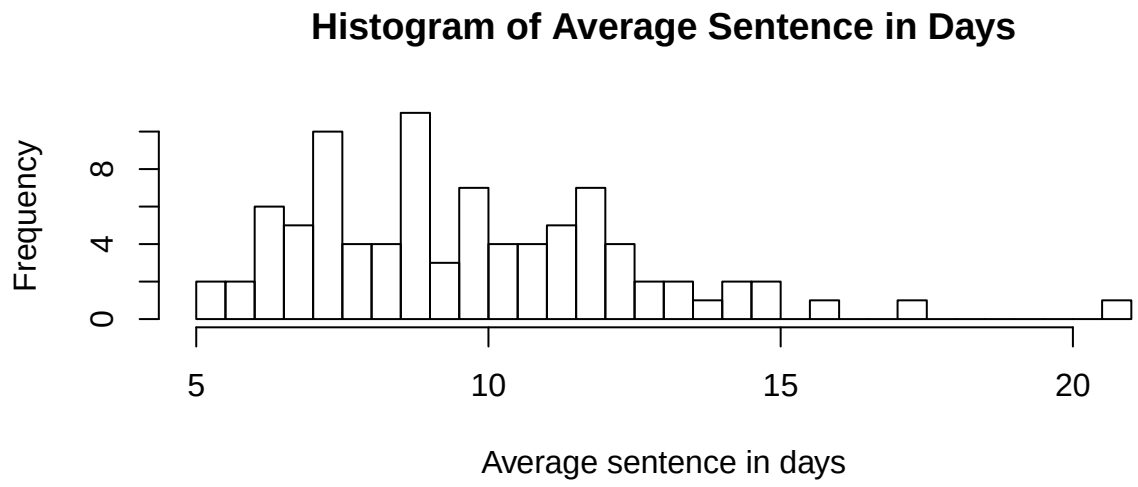
7. avgsen - Average Sentence in Days

Really? Days? This looks more like average sentence in years...

```r
summary(data$avgsen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.375   9.110   9.689  11.465  20.700
```

```r
hist(data$avgsen, breaks=50,
     main="Histogram of Average Sentence in Days",
     xlab="Average sentence in days")
```



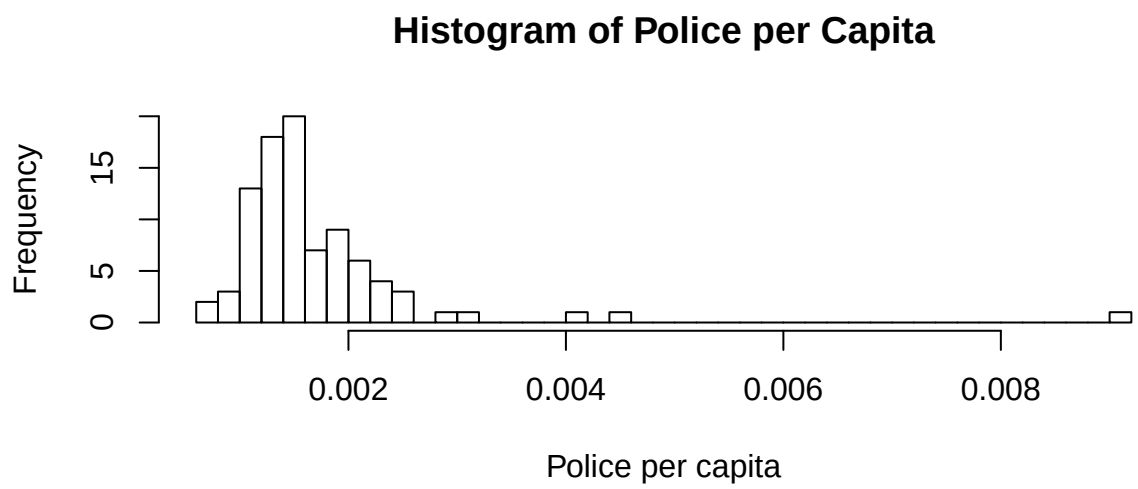
8. polpc - Police per Capita

Data looks expected.

```r
summary(data$polpc)
```

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

```r
hist(data$polpc, breaks=50,
     main="Histogram of Police per Capita",
     xlab="Police per capita")
```



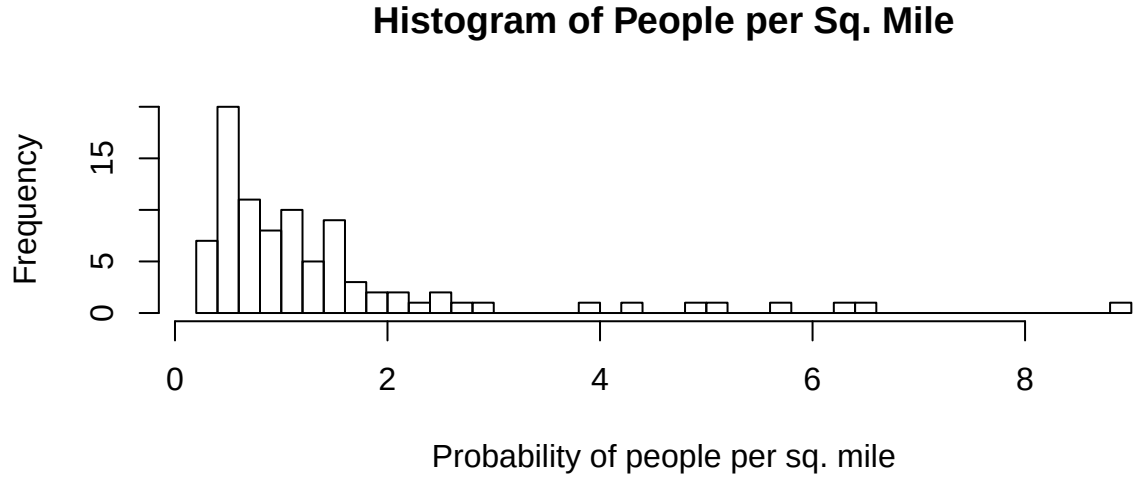
9. density - People per Sq. Mile

Data looks expected.

```r
summary(data$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2034  0.5472  0.9792  1.4379  1.5693  8.8277
```

```r
hist(data$density, breaks=50,
     main="Histogram of People per Sq. Mile",
     xlab="Probability of people per sq. mile")
```



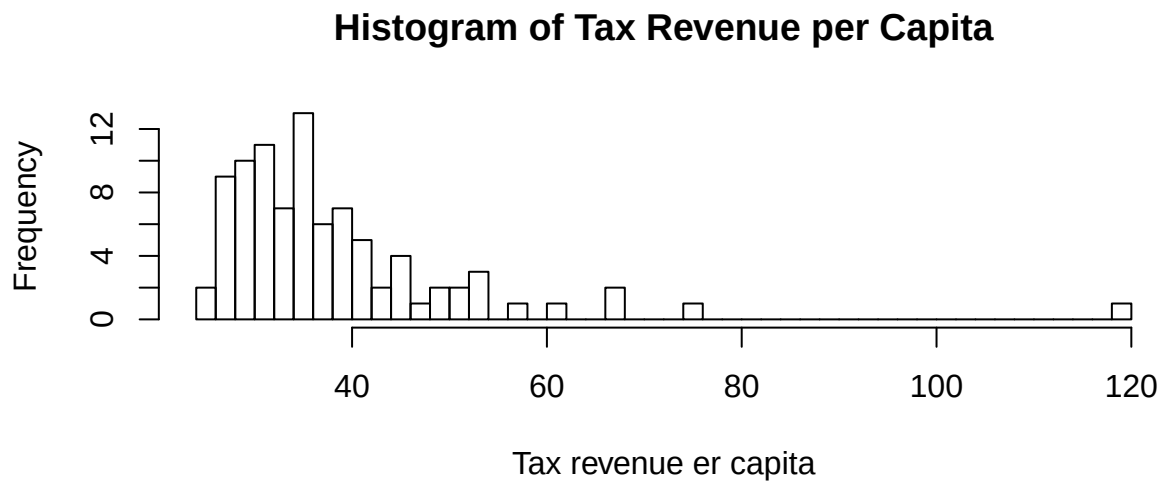
10. taxpc - Tax Revenue per Capita

Data looks expected.

```
summary(data$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.73   34.92   38.16   41.01  119.76
```

```
hist(data$taxpc, breaks=50,
     main="Histogram of Tax Revenue per Capita",
     xlab="Tax revenue er capita")
```



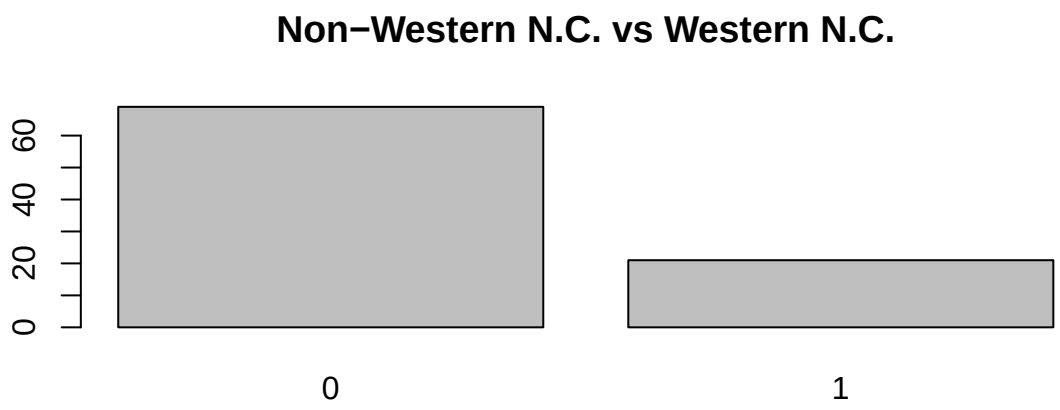
11. west - Indicator of Western N.C.

Data looks expected.

```
summary(factor(data$west))
```

```
##  0  1
## 69 21
```

```
barplot(table(data$west),
        main="Non-Western N.C. vs Western N.C.")
```



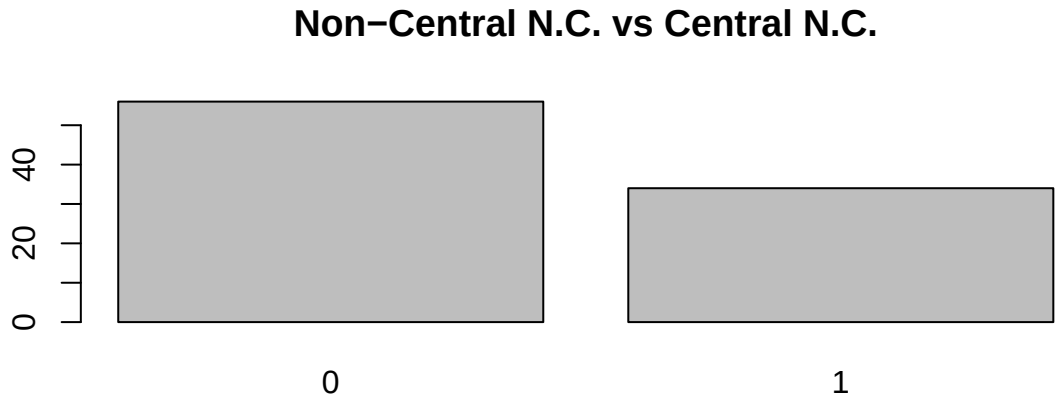
12. central - Indicator of Central N.C.

Data looks expected.

```
summary(factor(data$central))
```

```
##  0  1
## 56 34
```

```
barplot(table(data$central),
        main="Non-Central N.C. vs Central N.C.")
```



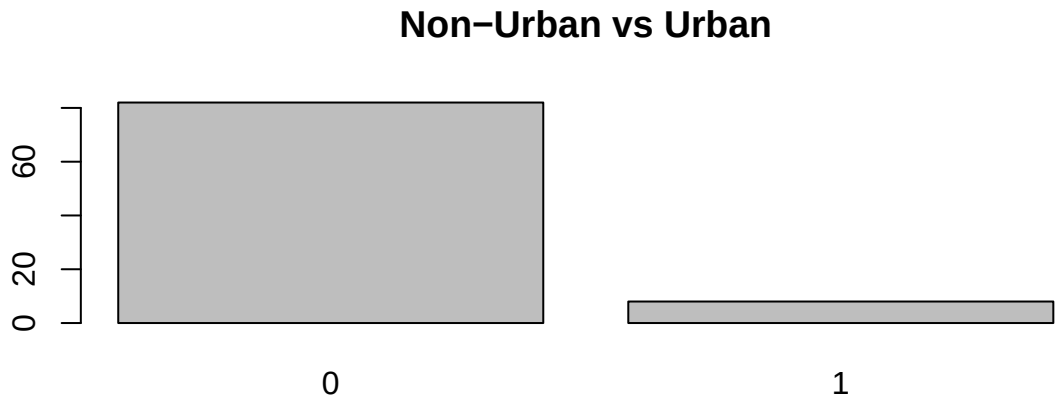
13. urban - Indicator of whether in SMSA

What does "SMSA" mean? Otherwise, data looks expected.

```
summary(factor(data$urban))
```

```
##  0  1
## 82  8
```

```
barplot(table(data$urban),
        main="Non-Urban vs Urban")
```



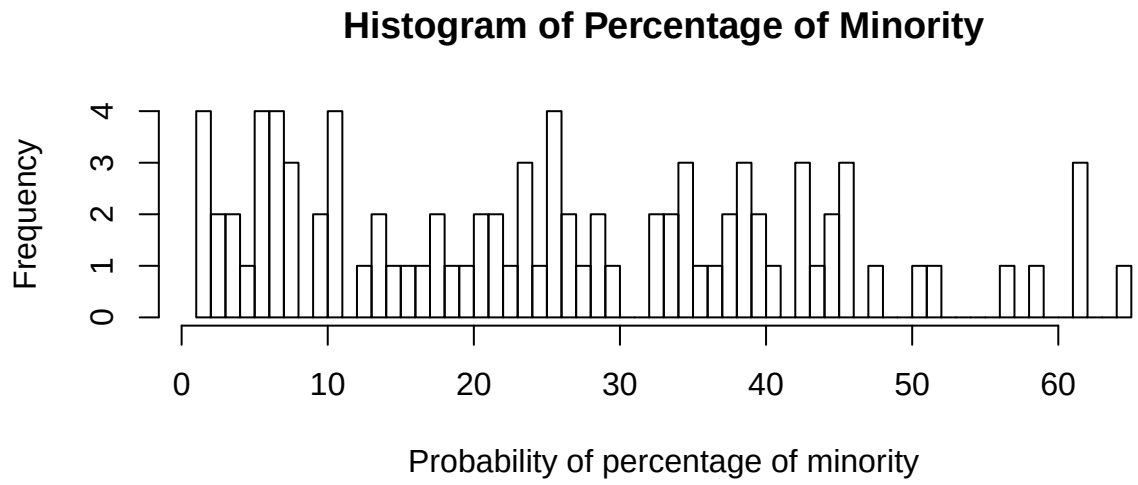
14. pctmin80 - Percentage of Minority, 1980

Data looks expected. However, data is between 0 - 100 whereas other percentage variables is between 0-1; We may want to transform this variable to keep things consistent from an interpretability perspective.

```
summary(data$pctmin80)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.284  10.024  24.852  25.713  38.183  64.348
```

```
hist(data$pctmin80, breaks=50,
     main="Histogram of Percentage of Minority",
     xlab="Probability of percentage of minority")
```



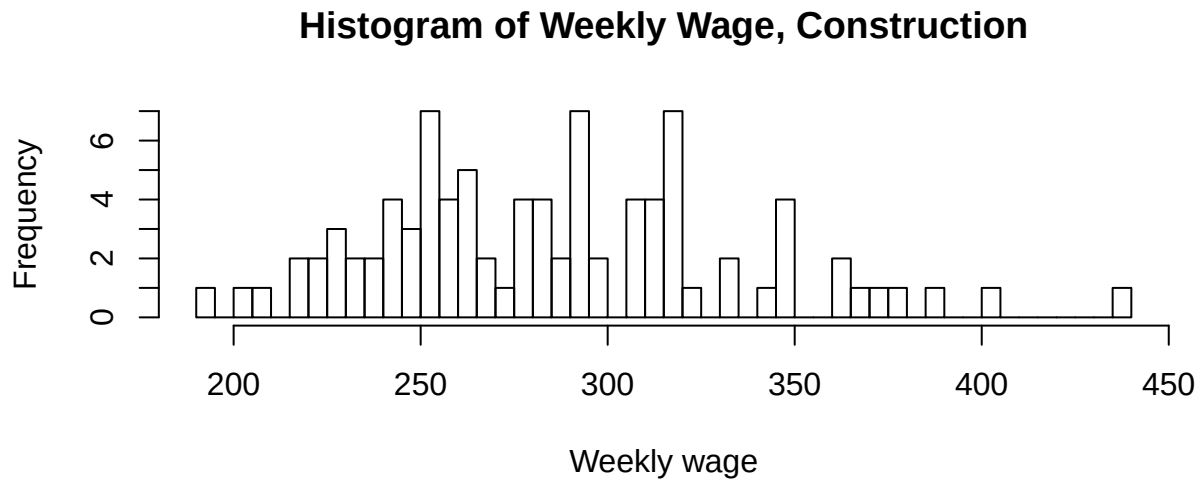
15. wcon - Weekly Wage, Construction

Data looks expected.

```r
summary(data$wcon)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   193.6   250.8   281.2   285.4   315.0   436.8
```

```r
hist(data$wcon, breaks=50,
     main="Histogram of Weekly Wage, Construction",
     xlab="Weekly wage")
```



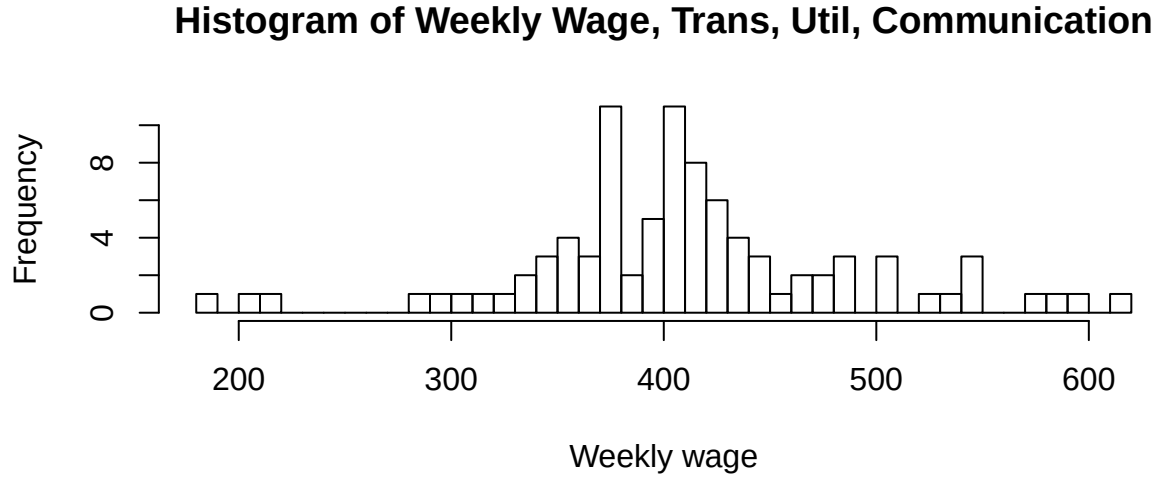
16. wtuc - Weekly Wage, Trans, Util, Communication

Data looks expected.

```r
summary(data$wtuc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.6   374.3   404.8   410.9   440.7   613.2
```

```r
hist(data$wtuc, breaks=50,
     main="Histogram of Weekly Wage, Trans, Util, Communication",
     xlab="Weekly wage")
```



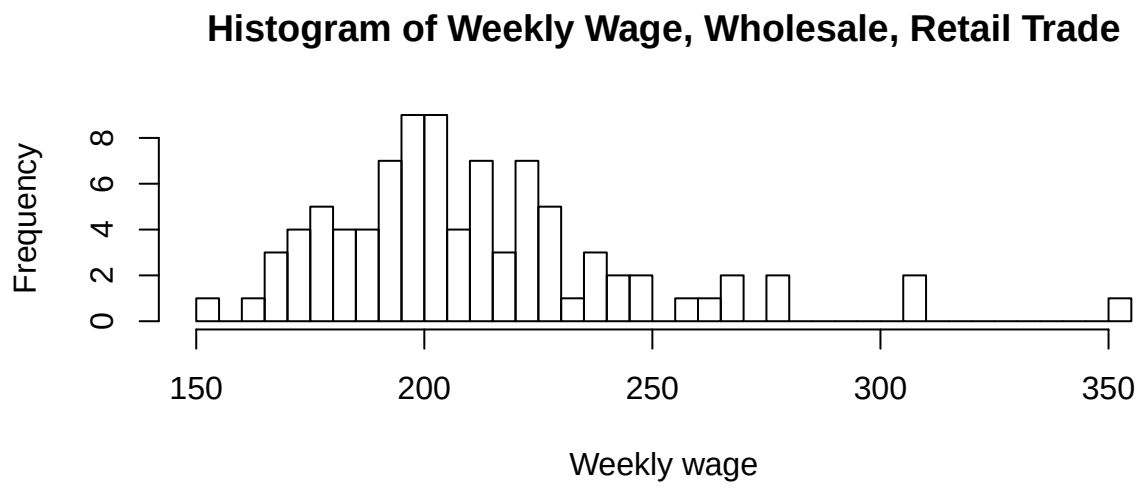
17. wtrd - Weekly Wage, Wholesale, Retail Trade

Data looks expected.

```
summary(data$wtrd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   154.2   190.7   203.0   210.9   224.3   354.7
```

```
hist(data$wtrd, breaks=50,
     main="Histogram of Weekly Wage, Wholesale, Retail Trade",
     xlab="Weekly wage")
```



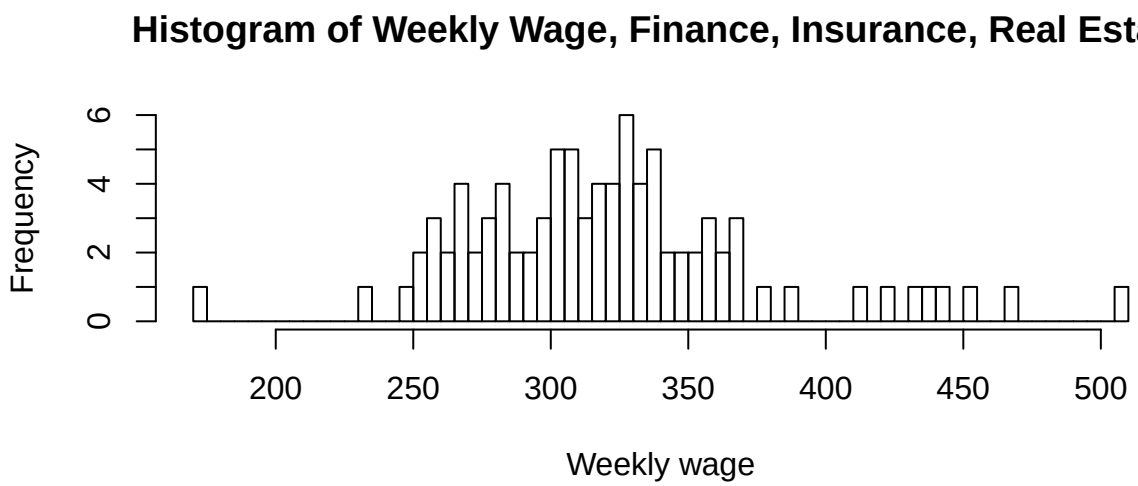
18. wfir - Weekly Wage, Finance, Insurance, Real Estate

Data looks expected.

```
summary(data$wfir)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   170.9   285.6   317.1   321.6   342.6   509.5
```

```
hist(data$wfir, breaks=50,
     main="Histogram of Weekly Wage, Finance, Insurance, Real Estate",
     xlab="Weekly wage")
```



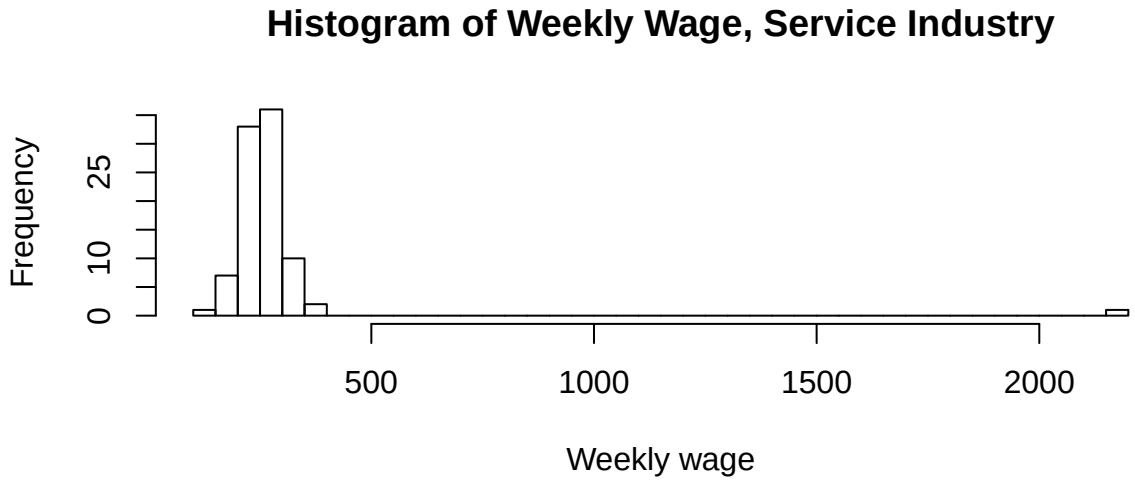
19. wser - Weekly Wage, Service Industry

There's a very extreme outlier here, need to take a closer look.

```
summary(data$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

```
hist(data$wser, breaks=50,
     main="Histogram of Weekly Wage, Service Industry",
     xlab="Weekly wage")
```




```
data[data$wser > 2000, ]
```

```
##     X county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84 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
```

```
## 84 0.07008217
```

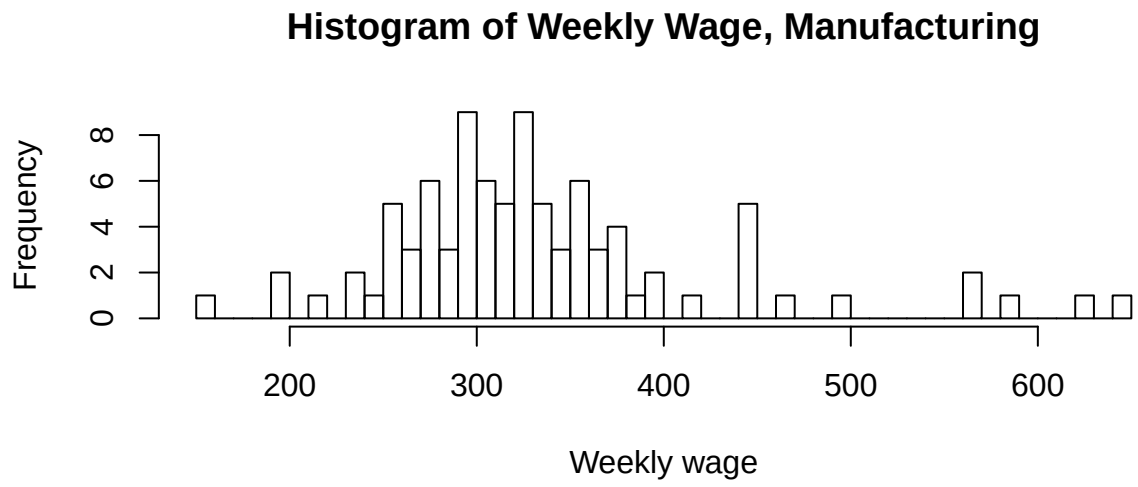
20. wmfg - Weekly Wage, Manufacturing

Data looks expected.

```
summary(data$wmfg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   157.4   288.6   321.1   336.0   359.9   646.9
```

```
hist(data$wmfg, breaks=50,
     main="Histogram of Weekly Wage, Manufacturing",
     xlab="Weekly wage")
```



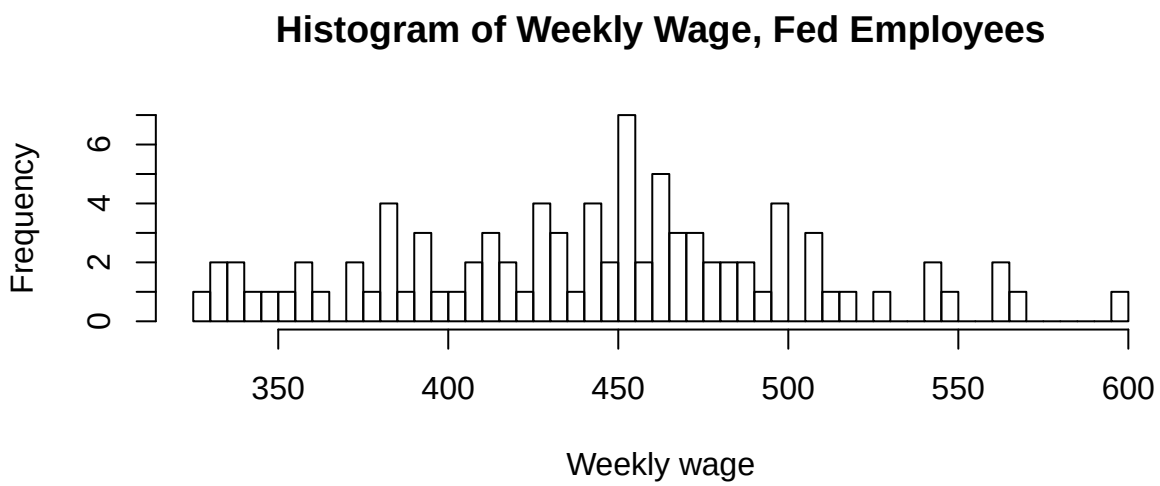
21. wfed - Weekly Wage, Fed Employees

Data looks expected.

```
summary(data$wfed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   326.1   398.8   448.9   442.6   478.3   598.0
```

```
hist(data$wfed, breaks=50,
     main="Histogram of Weekly Wage, Fed Employees",
     xlab="Weekly wage")
```



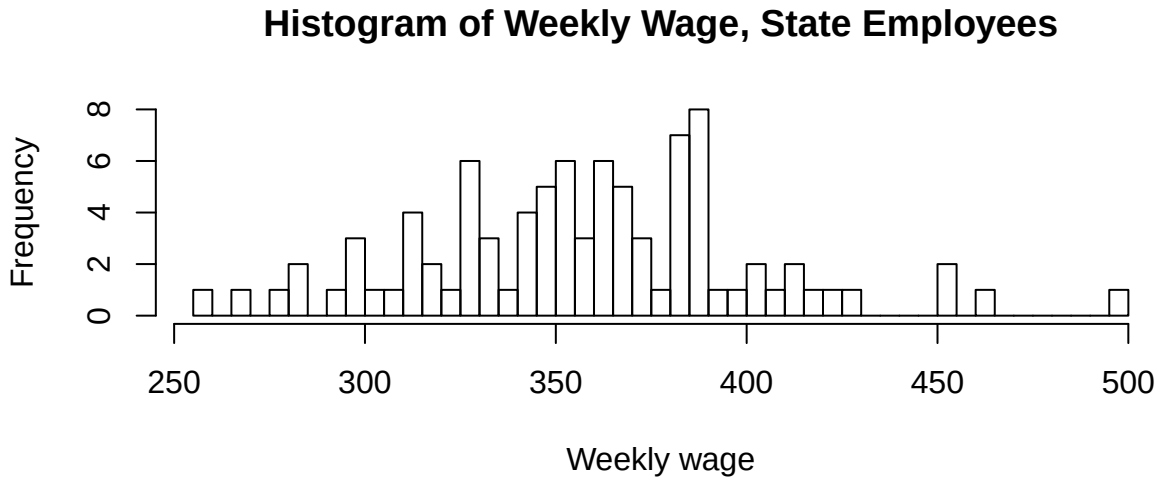
22. wsta - Weekly Wage, State Employees

Data looks expected.

```
summary(data$wsta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   258.3   329.3   358.4   357.7   383.2   499.6
```

```
hist(data$wsta, breaks=50,
     main="Histogram of Weekly Wage, State Employees",
     xlab="Weekly wage")
```



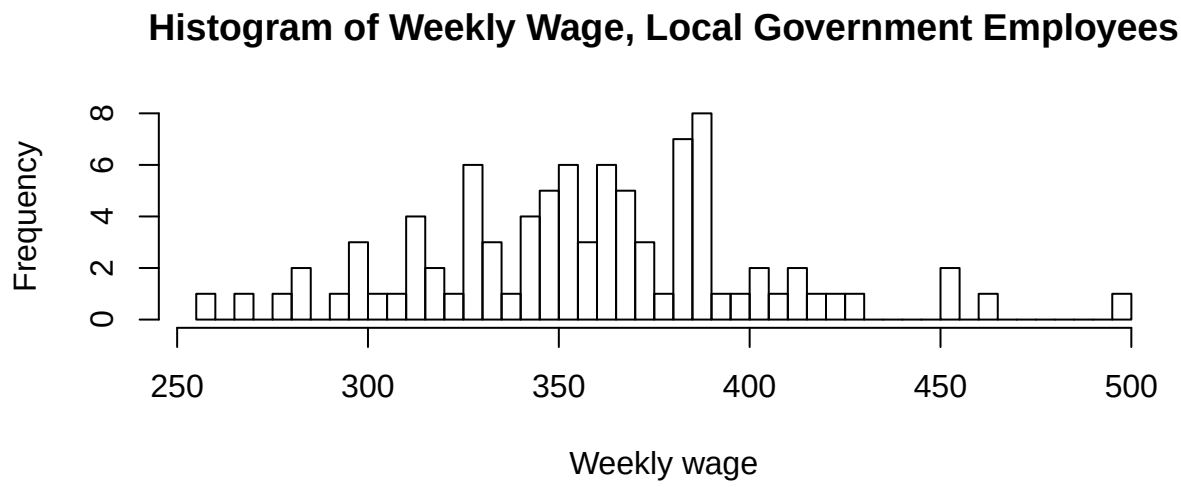
23. wloc - Weekly Wage, Local Government Employees

Data looks expected.

```
summary(data$wsta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   258.3   329.3   358.4   357.7   383.2   499.6
```

```
hist(data$wsta, breaks=50,
     main="Histogram of Weekly Wage, Local Government Employees",
     xlab="Weekly wage")
```



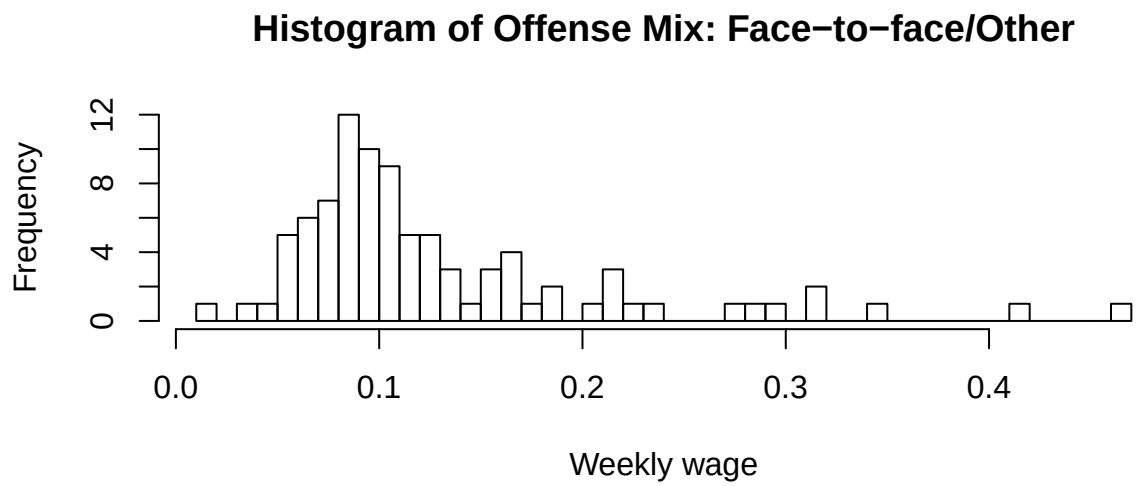
24. mix - Offense Mix: Face-to-face/Other

Not really sure what this variable means...

```
summary(data$mix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01961 0.08060 0.10095 0.12905 0.15206 0.46512
```

```
hist(data$mix, breaks=50,
     main="Histogram of Offense Mix: Face-to-face/Other",
     xlab="Weekly wage")
```



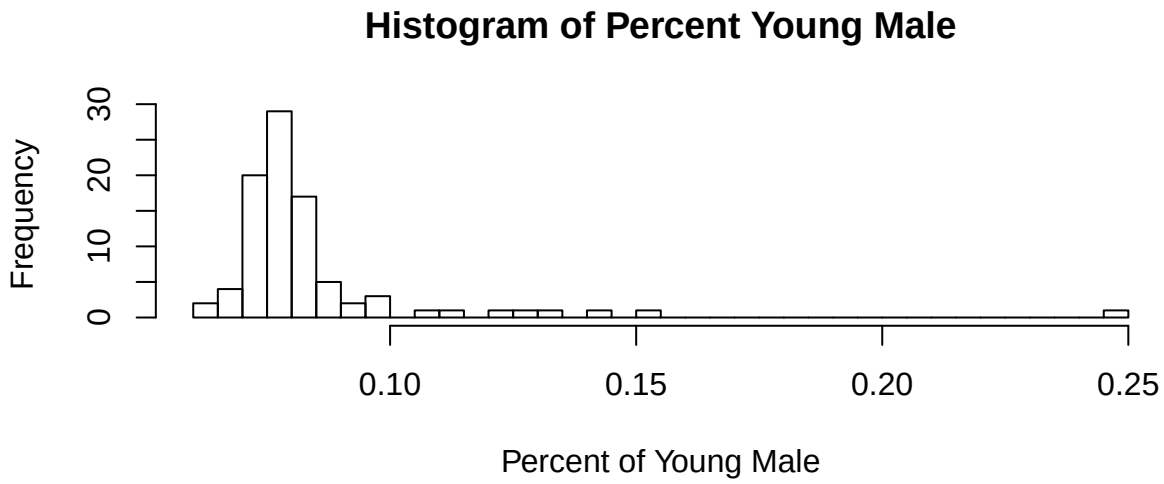
25. pctymle - Percent Young Male

Data looks expected.

```r
summary(data$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871
```

```r
hist(data$pctymle, breaks=50,
     main="Histogram of Percent Young Male",
     xlab="Percent of Young Male")
```



Bi-variate Variable Analysis

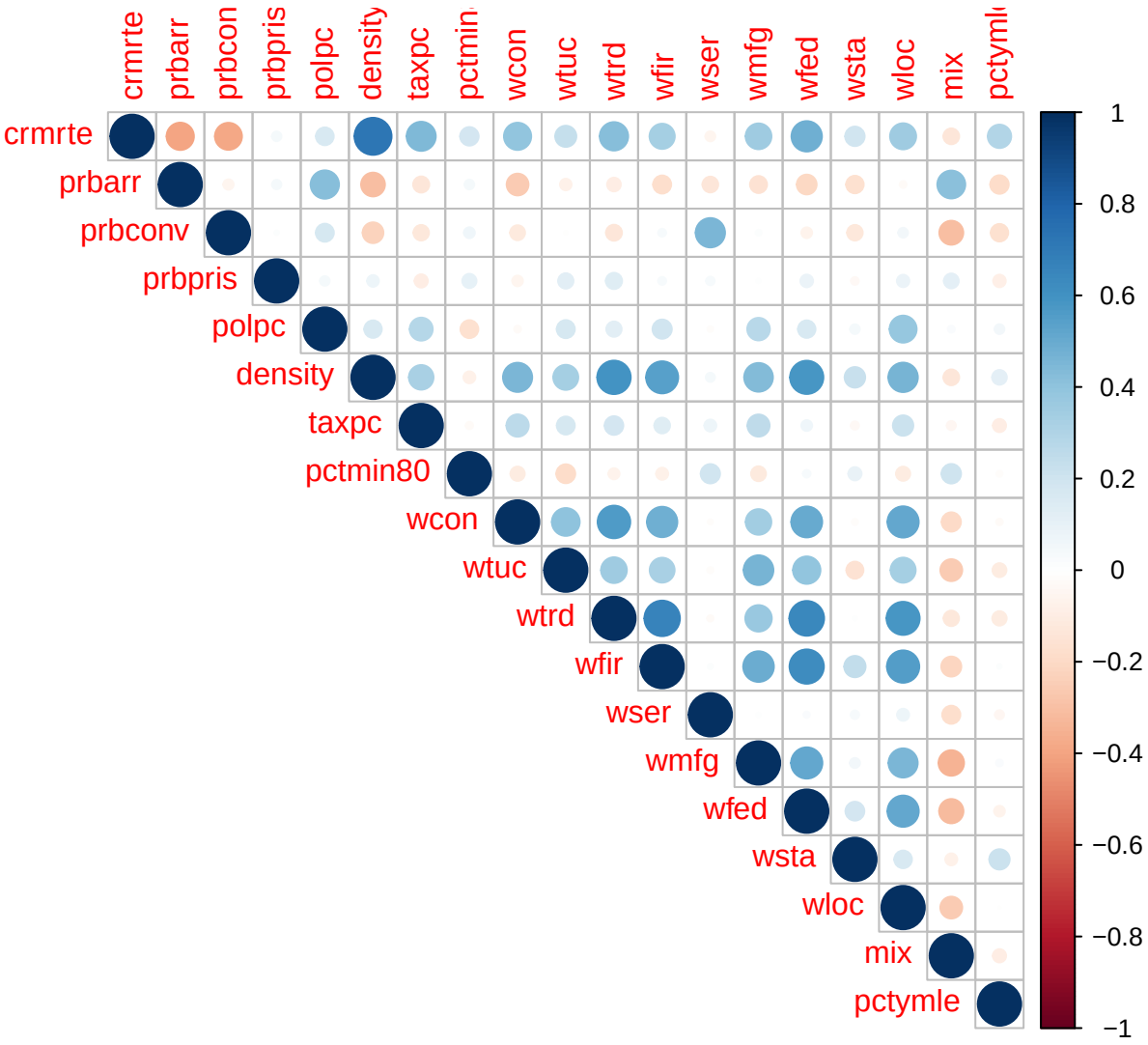
Here is the correlation plot for the non-indicator variables.

Looking at the plot, looks like the following variables correlate with crmrte highly and positively:

1. density - Makes sense since the crime rate tends to increase in more populated areas.
2. taxpc - Somewhat makes sense since crime rate could increase in areas where there are more tax revenues collected.
3. wage variables - Somewhat makes sense since as wages go up, there may be higher likelihood for crime.

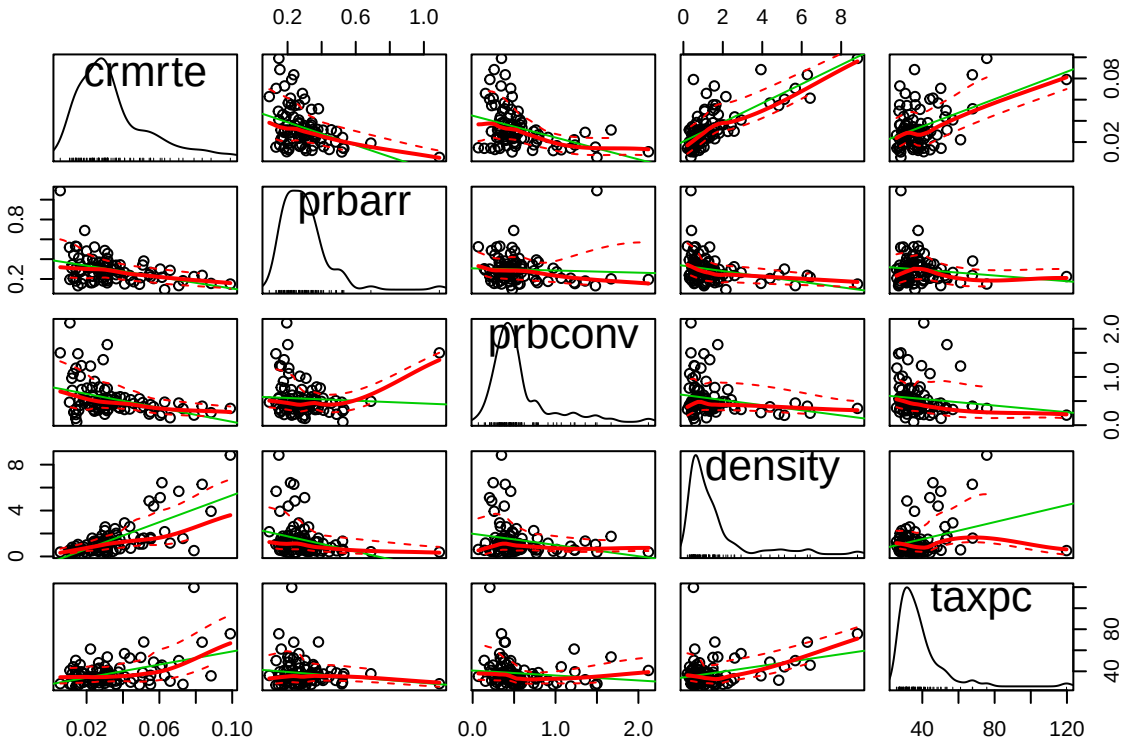
Looks like the following variables correlate with crmrte highly and negatively: 1. prbarr - Makes sense since if probabily of arrests go down, then there are more criminals out on the streets. 2. prbconv - Makes sense since if probability of convictions go down, then there are more (potential) criminal out on the streets.

```r
corrplot(cor(data[ , (names(data) %in% c("crmrte", "prbarr", "prbconv", "prbpris", "polpc", "density", "pctmi
                                         "taxpc", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "w
                                         "mix"))]),
                                          method="circle", type="upper")
```



Let's focus more on the variables that correlate highly with crmrte.

```
scatterplotMatrix(data[ , (names(data) %in% c("crmrte", "prbarr", "prbconv", "density",
                                              "taxpc"))])
```


crmrte
prbarr
prbconv
density
taxpc
0.2
0.6
1.0
0
2
4
6
8
0.02
0.08
0.2
0.8
0.0
1.0
2.0
0
4
8
40
80
120
0.02
0.06
0.10
0.0
1.0
2.0