EDA Crime Lab

Eduarda Espindola, Laura Pintos, Payman Roghani and Pri Nonis

20/11/2018

Crime Lab

Eduarda, Laura, Payman, Pri

EDA

```
setwd("/Users/eduardaespindola/Documents/Mestrado/W203 - Stats/Lab3/w203-lab3")
crime_data <- read.csv("crime_v2.csv")
```

```
head(crime_data)
```

```
##   county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
## 1      1   87 0.0356036 0.298270 0.527595997 0.436170   6.71 0.00182786
## 2      3   87 0.0152532 0.132029 1.481480002 0.450000   6.35 0.00074588
## 3      5   87 0.0129603 0.444444 0.267856985 0.600000   6.76 0.00123431
## 4      7   87 0.0267532 0.364760 0.525424004 0.435484   7.14 0.00152994
## 5      9   87 0.0106232 0.518219 0.476563007 0.442623   8.22 0.00086018
## 6     11   87 0.0146067 0.524664 0.068376102 0.500000  13.00 0.00288203
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 1 2.4226327 30.99368    0       1     0 20.21870 281.4259 408.7245
## 2 1.0463320 26.89208    0       1     0  7.91632 255.1020 376.2542
## 3 0.4127659 34.81605    1       0     0  3.16053 226.9470 372.2084
## 4 0.4915572 42.94759    0       1     0 47.91610 375.2345 397.6901
## 5 0.5469484 28.05474    1       0     0  1.79619 292.3077 377.3126
## 6 0.6113361 35.22974    1       0     0  1.54070 250.4006 401.3378
##       wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 1 221.2701 453.1722 274.1775 334.54 477.58 292.09 311.91 0.08016878
## 2 196.0101 258.5650 192.3077 300.38 409.83 362.96 301.47 0.03022670
## 3 229.3209 305.9441 209.6972 237.65 358.98 331.53 281.37 0.46511629
## 4 191.1720 281.0651 256.7214 281.80 412.15 328.27 299.03 0.27362204
## 5 206.8215 289.3125 215.1933 290.89 377.35 367.23 342.82 0.06008584
## 6 187.8255 258.5650 237.1507 258.60 391.48 325.71 275.22 0.31952664
##      pctymle
## 1 0.07787097
## 2 0.08260694
## 3 0.07211538
## 4 0.07353726
## 5 0.07069755
## 6 0.09891920
```

```
str(crime_data)
```

```
## '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",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ 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(crime_data)
```

```
##      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
```

Understanding the meaning of some of the variables, we are able to do some cross checks, and make sure all the data makes sense:

1. County (county): It is the county identifier, and as for the problem statement, we should have only one entry (one row) per county:

```
crime_data[which(is.na(crime_data$county)),]
```

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 92     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 93     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 94     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 95     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 96     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 97     NA   NA     NA     NA       `      NA     NA    NA      NA    NA
##    west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta
## 92   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 93   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 94   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 95   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 96   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 97   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
##    wloc mix pctymle
## 92   NA  NA      NA
## 93   NA  NA      NA
## 94   NA  NA      NA
## 95   NA  NA      NA
## 96   NA  NA      NA
## 97   NA  NA      NA
```

We have no data in these 6 rows, so for the purpose of our analysis, we can get it out

```
crime_data<-crime_data[which(!is.na(crime_data$county)),]
```

Now, we must finally check for duplicate values:

```
crime_data[duplicated(crime_data),]
```

```
##    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
```

We have seen that we have two entries for county 193. The data structure we have should be one row for one county, which is why we are going to discard the extra entry for county 193

```
crime_data<-unique(crime_data)
```

If we check again for duplicates, it shows us none:

```
crime_data[duplicated(crime_data),]
```

```
##  [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
## <0 rows> (or 0-length row.names)
```

2. Year (year): we have that all the observations come from the year of 1987, therefore, we should just check if there are other years on this dataset

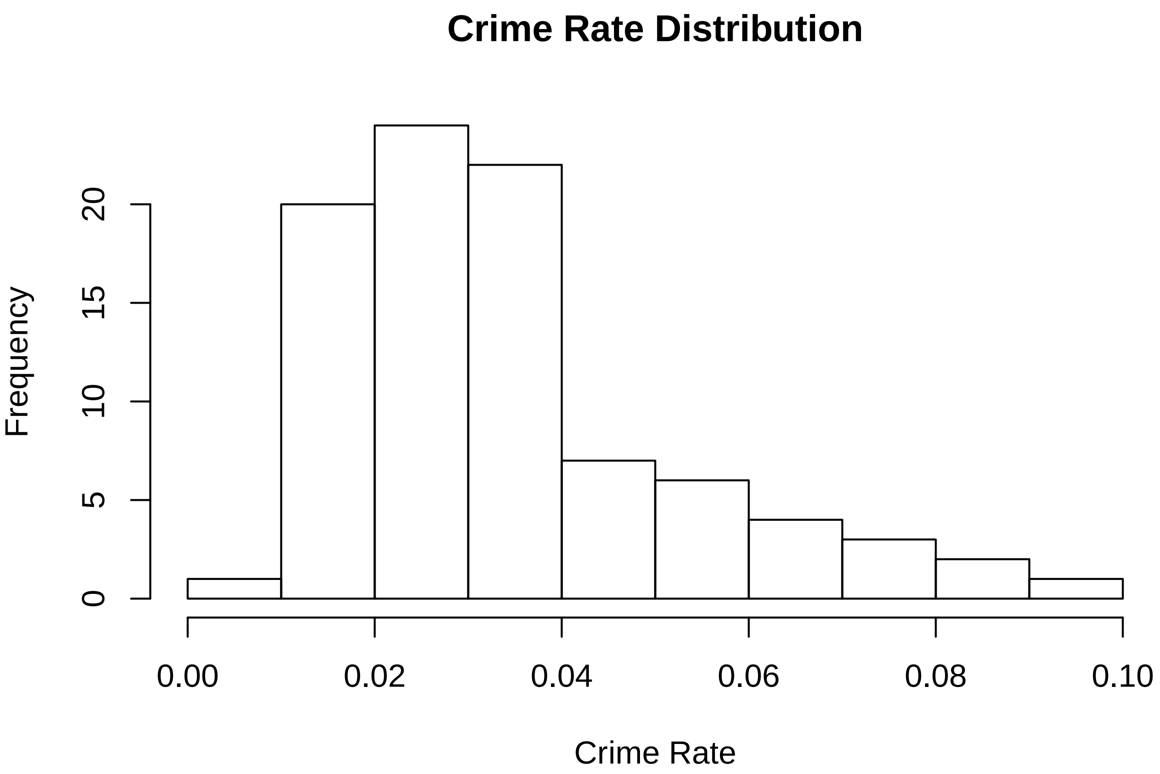
```
summary(crime_data$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      87      87      87      87      87      87
```

And there we have it, only observations for 1987.

3. Crime Rate (crmrte): It is calculated as ratio of number of reported crimes to the total population of the county. Theoretically, we could have values ranging from zero (no crimes commited in that county in 1987) to infinity (so many crimes committed that the ratio goes to infinity), however both these cases are extremes that don't make any logical sense. So we should check the distribution of this variable to try and spot weird observations:

```
hist(x = crime_data$crmrte, main = "Crime Rate Distribution", xlab = "Crime Rate", ylab = "Frequency")
```



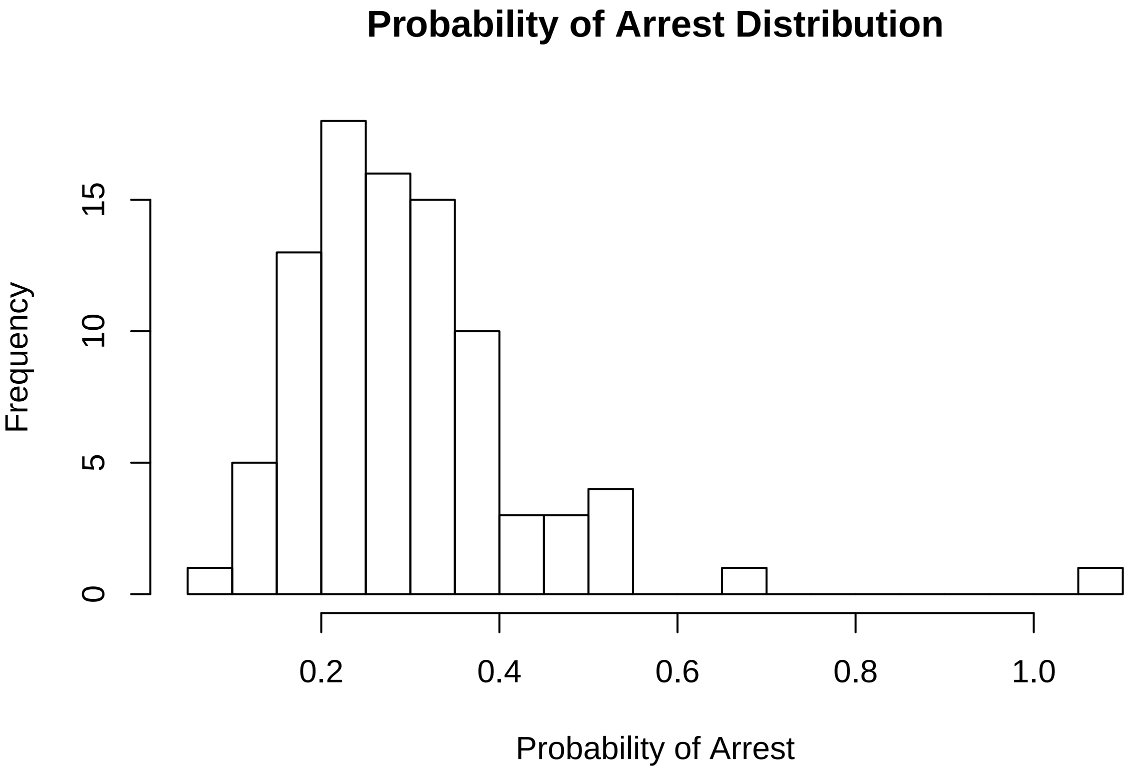
```
summary(crime_data$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```

There is nothing abnormal with the data, so it is safe to proceed.

4. Probability of arres (prbarr): The probability of arrest is proxied by the ratio of arrests to offenses.

```
hist(x = crime_data$prbarr, breaks=20, main = "Probability of Arrest Distribution", xlab = "Probability
```



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

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

Probabilities should not be over 100%, so we should take a closer look at the observations where the probability of arrest were higher than 1

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

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

For county 115, another thing jumps to the eye, the probability of conviction (prbpris, proxied by the ratio of convictions to arrests), is also higher than 1. Probabilities should range from 0 to 1, however, these anomalies might be due to the way those variables were proxied: probability of arrest is proxied by the ratio of arrests to offenses and the probability of conviction, by the ratio of convictions to arrests. They are not actual probabilities. One may argue that it makes no sense to have more arrests than offenses, or more convictions than arrests, however, we are looking at snapshot of 1987, and arrests made in that year might be referring both to offenses mad in 1987 and previously, which could explain the ration being over than one. The same line of thought applies for the probability of conviction variable: the convictions made in 1987 might be referring both to arrests made in 1987 and previously. For those reasons, we choose not to discard this observation.

5. Probability of Conviction (prbconv): As we have seen previously, the probability of conviction is proxied by the ratio of convictions to arrests.

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

```
##                       ` 0.068376102 0.140350997 0.154451996 0.203724995 
##           0           0           1           1           1           1 
## 0.207830995    0.220339 0.226361006 0.229589999 0.248275995 0.259833008 
##           1           1           1           1           1           1 
## 0.267856985 0.271946996  0.28947401 0.300577998 0.308411002 0.314606994 
##           1           1           1           1           1           1 
## 0.322580993 0.325300992 0.327868998 0.328664005 0.334701002 0.340490997 
##           1           1           1           1           1           1 
## 0.343023002 0.347799987 0.352941006  0.36015299 0.364353001 0.371879011 
##           1           1           1           1           1           1 
##    0.381908 0.384236008 0.385495991 0.386925995 0.393413007    0.401198 
##           1           1           1           1           1           1 
## 0.403780013 0.406780005 0.410596013 0.412698001 0.426777989 0.436441004 
##           1           1           1           1           1           1 
## 0.438960999 0.443114012 0.443681002 0.449999988 0.450567007 0.452829987 
##           1           1           1           1           1           1 
## 0.457210004 0.459215999 0.468531013 0.476563007 0.477732986 0.492940009 
##           1           1           1           1           1           1 
## 0.493438005 0.495575011  0.50819701 0.515464008 0.520606995 0.520709991 
##           1           1           1           1           1           1 
## 0.522387981 0.525424004 0.527595997 0.528302014 0.548494995 0.549019992 
##           1           1           1           1           1           1 
## 0.559822977 0.571429014 0.573943973 0.588859022 0.589905024 0.595077991 
##           1           1           1           1           1           1 
##  0.62251699 0.722972989 0.736908972 0.739394009 0.763333023 0.769231021 
##           1           1           1           1           1           1 
## 0.781608999 0.793232977 0.909090996 0.972972989 1.015380025 1.068969965 
##           1           1           1           1           1           1 
## 1.182929993 1.225610018 1.234380007 1.358139992 1.481480002         1.5 
##           1           1           1           1           1           1 
## 1.670519948 2.121210098 
##           1           1
```

The probability of conviction has some weird values, such one that is empty and another one that is '. We should take a look at those observations

```r
crime_data$prbconv
```

```
##  [1] 0.527595997 1.481480002 0.267856985 0.525424004 0.476563007
##  [6] 0.068376102 0.520606995 0.769231021 0.436441004 1.225610018
## [11] 0.334701002 0.403780013 0.406780005 0.352941006 0.515464008
## [16] 0.325300992 0.385495991 0.972972989 0.452829987 0.450567007
## [21] 0.763333023 0.371879011 0.259833008 0.140350997 0.207830995
## [26] 0.736908972 0.62251699  0.493438005 0.459215999 0.154451996
## [31] 0.248275995 0.739394009 0.229589999 0.528302014 0.308411002
## [36] 0.203724995 0.457210004 0.549019992 0.548494995 0.386925995
## [41] 0.589905024 0.573943973 0.595077991 1.234380007 0.571429014
## [46] 0.384236008 0.364353001 0.781608999 0.522387981 0.220339   
## [51] 1.5         0.793232977 0.347799987 0.226361006 0.438960999
## [56] 1.358139992 0.393413007 0.495575011 0.271946996 0.477732986
## [61] 1.068969965 0.28947401  0.412698001 0.314606994 0.340490997
## [66] 0.426777989 1.015380025 0.36015299  0.520709991 0.559822977
## [71] 0.443681002 0.492940009 0.50819701  0.401198    0.468531013
## [76] 0.322580993 0.722972989 0.909090996 0.327868998 0.410596013
## [81] 0.328664005 0.343023002 0.381908    2.121210098 0.443114012
## [86] 0.300577998 0.449999988 0.588859022 1.670519948 1.182929993
## 92 Levels:  ` 0.068376102 0.140350997 0.154451996 ... 2.121210098
```

```
crime_data[crime_data$prbconv == '' | crime_data$prbconv=='`',]
```

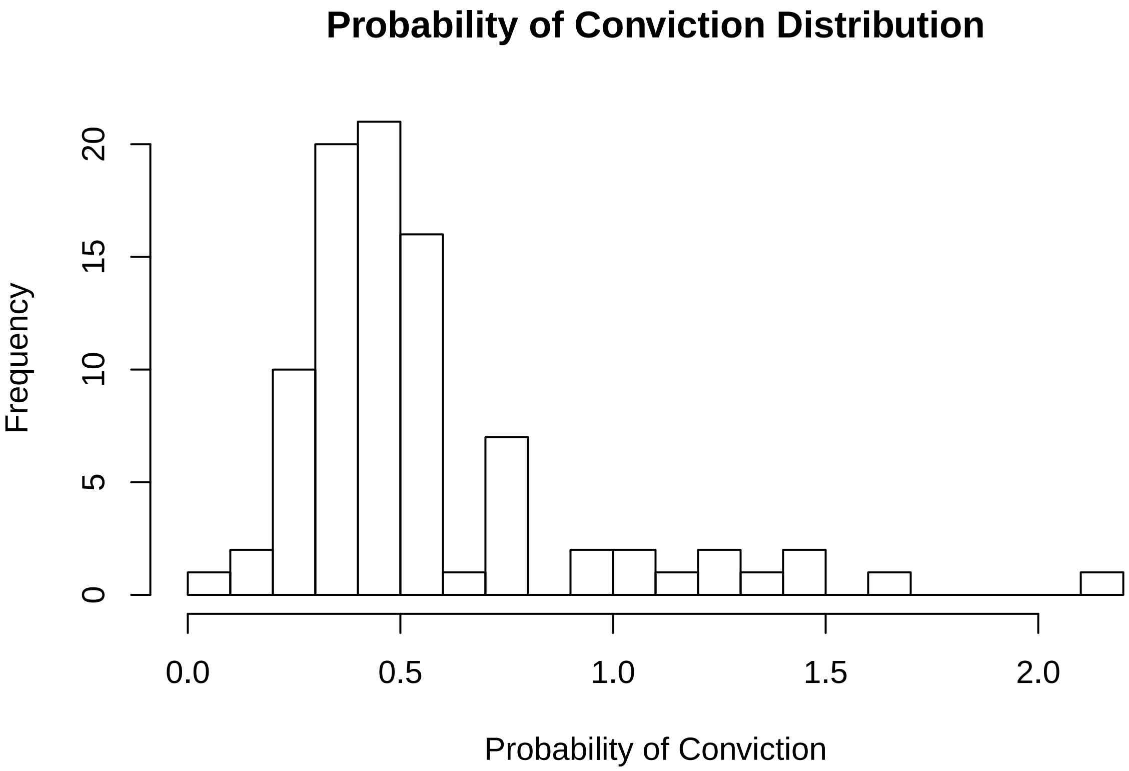
```
##  [1] county   year     crmrte   prbarr   prbconv  prbpris  avgsen
##  [8] polpc    density  taxpc    west     central  urban    pctmin80
## [15] wcon     wtuc     wtrd     wfir     wser     wmfg     wfed
## [22] wsta     wloc     mix      pctymle
## <0 rows> (or 0-length row.names)
```

The observations with these weird values have already been discarded on previous analysis, however, they still show up as factors, since they were first loaded like that. One way we could go is transforming that variable into a numeric one

```
crime_data$prbconv<-as.numeric(as.character(crime_data$prbconv))
```

Now we can perform the usual analysis:

```
hist(x = crime_data$prbconv, breaks=20, main = "Probability of Conviction Distribution", xlab = "Probab
```



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

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

Again, we see observations in which the probability of conviction is higher than 1, which shouldn't happen, if they were in fact probabilities. However, as we previously stated, by the method they were proxied, values above 1 are possible. But, nonetheless, we must analyze those cases in more detail.

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

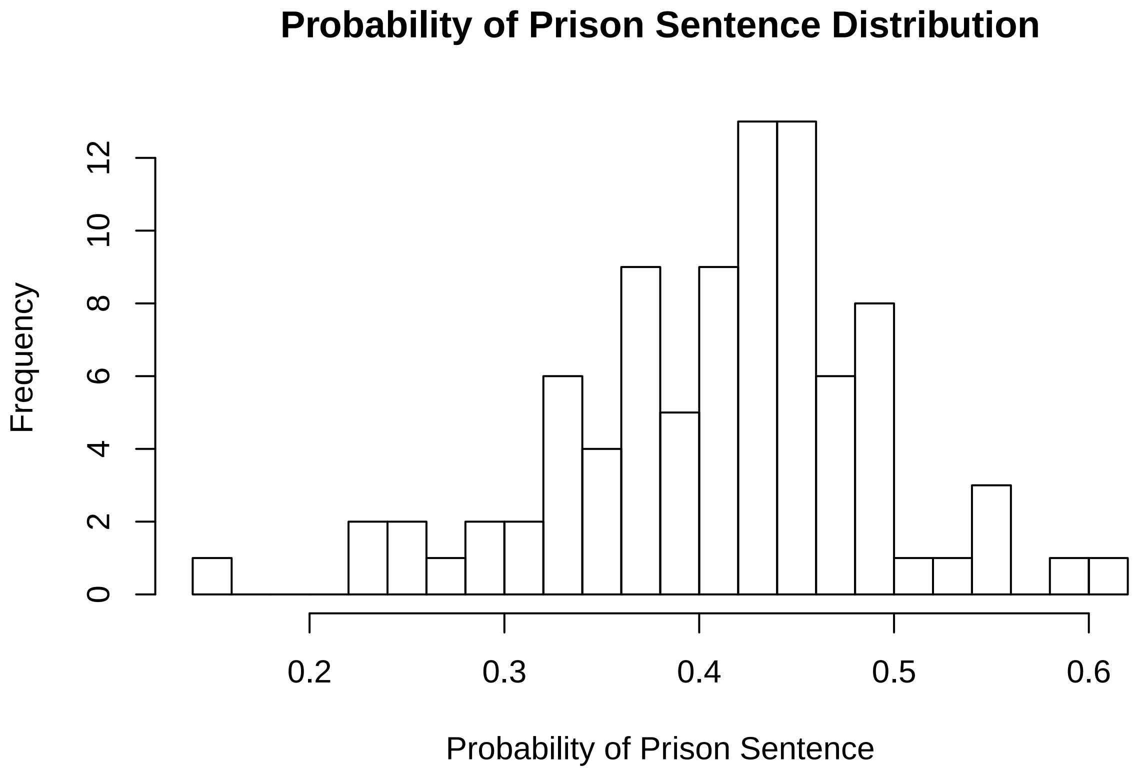
```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 2       3   87 0.0152532 0.132029 1.48148 0.450000   6.35 0.00074588
## 10     19   87 0.0221567 0.162860 1.22561 0.333333  10.34 0.00202425
## 44     99   87 0.0171865 0.153846 1.23438 0.556962  14.75 0.00185912
## 51    115   87 0.0055332 1.090910 1.50000 0.500000  20.70 0.00905433
## 56    127   87 0.0291496 0.179616 1.35814 0.335616  15.99 0.00158289
## 61    137   87 0.0126662 0.207143 1.06897 0.322581   6.18 0.00081426
## 67    149   87 0.0164987 0.271967 1.01538 0.227273  14.62 0.00151871
## 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.00122210
## 90    195   87 0.0313973 0.201397 1.67052 0.470588  13.02 0.00445923
## 91    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
## 90 1.7459893 53.66693    0       0     0 37.43110 315.1641 377.9356
## 91 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
## 90 246.0614 411.4330  296.8684 392.27 480.79 303.11 337.28 0.15612382
## 91 182.8020 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
## 90 0.07945071
## 91 0.07419893
```

Those observations fall into the same issue we have seen for the probability of arrest variable. By the way they were proxied, the ratio of convictions to arrests in 1987 doesn't necessarily matches convictions in 1987 referring to arrests only made in 1987. There might be some convictions made in 1987 referring to arrests made in previous years in the mix, which is why we decide to keep those observations, as the same effect migh also be present in the observations where the probability of conviction was below 1.

6. Probability of Prison Sentence (prbpris): The probability of prison sentence is proxied by the convictions resulting in a prison sentence to total convictions. In that case, unlike the other two previous variables

we analyzed, the ratio is calculated in the same set of convictions: how many of such set of convictions resulted in a prison sentence. Therefore, for this variable, we should have the values ranging from 0 to a maximum of 1.

```
hist(x = crime_data$prbpris, breaks=20, main = "Probability of Prison Sentence Distribution", xlab = "Pr
```



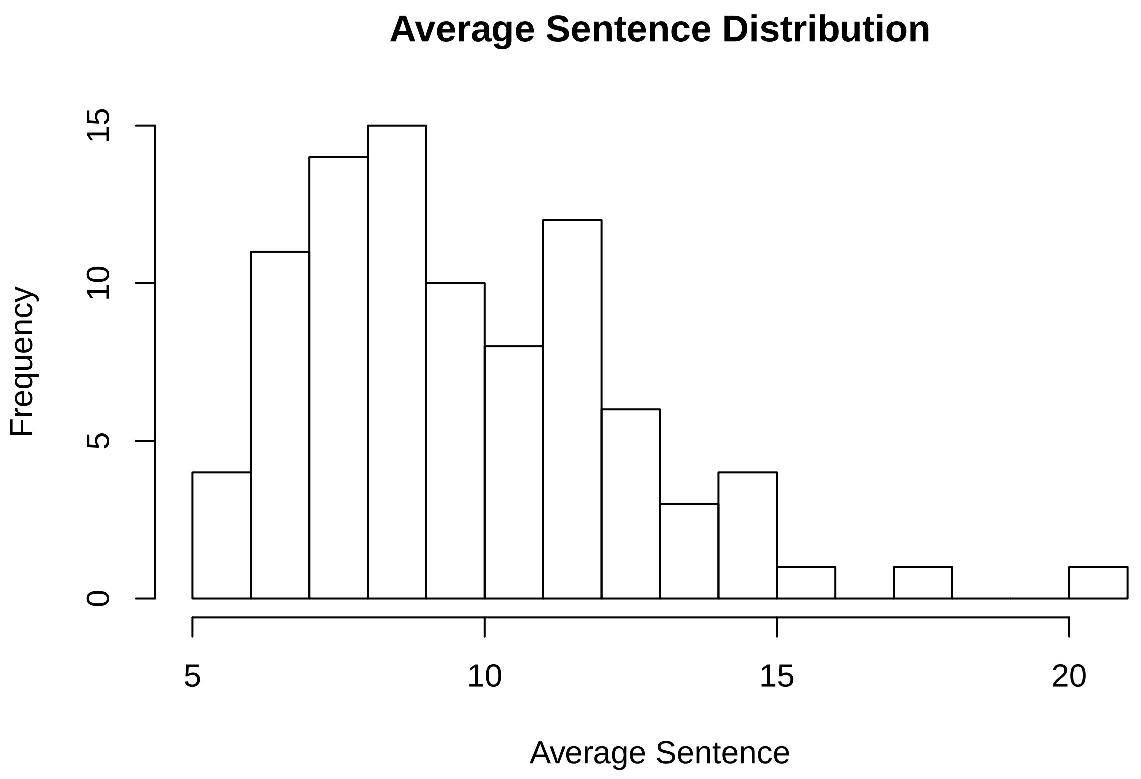
```
summary(crime_data$prbpris)
```

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

The variable behaves as we expected, and we can move on to analyzing other variables.

7. Average Sentence, days (avgsen): The average sentence time in days. This variable doesn't have a theoretical limit, it only shouldn't be negative. So we just need to be wary of outliers and understand if the values are actually true or some sort of measurement mistake.

```
hist(x = crime_data$avgsen, breaks=20, main = "Average Sentence Distribution", xlab = "Average Sentence
```



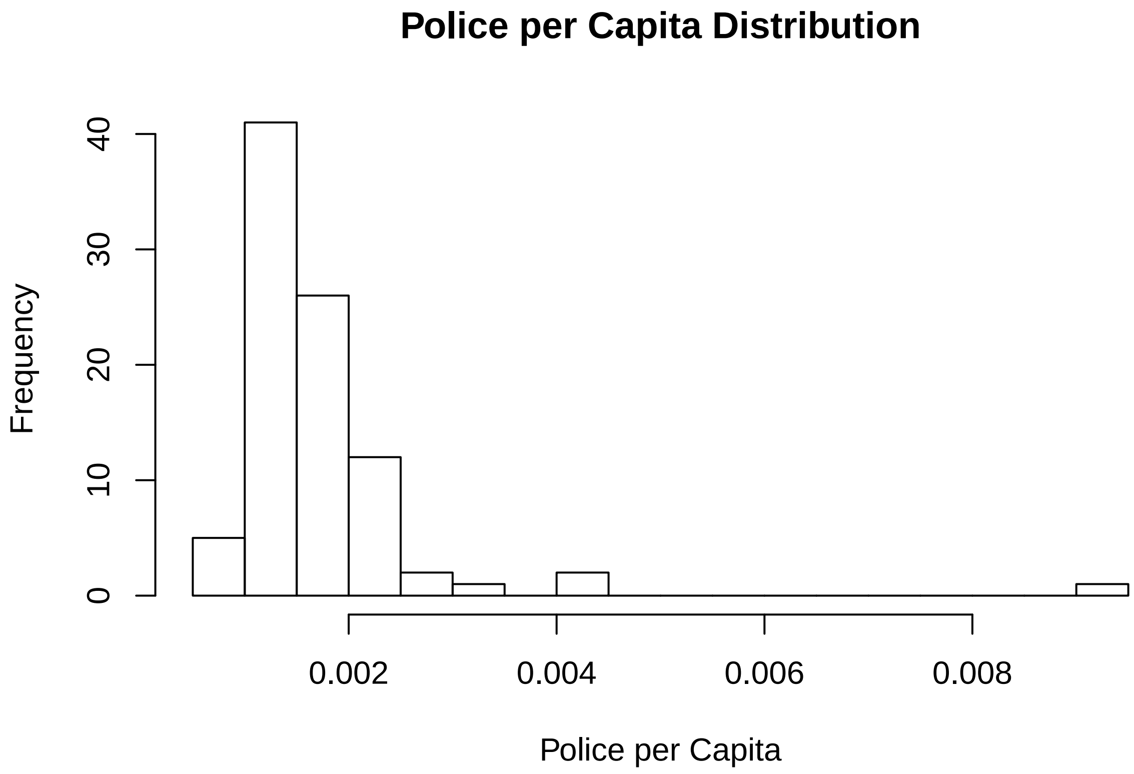
```
summary(crime_data$avgsen)
```

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

The variable behaves as we expected, and we can move on to analyzing other variables.

8. Police per Capita (polpc): The ratio of the number of police officers to the total population of the county. The values must be in the range from 0 (no cops in the county) to 1 (everyone in the county is a cop).

```
hist(x = crime_data$polpc, breaks=20, main = "Police per Capita Distribution", xlab = "Police per Capita
```



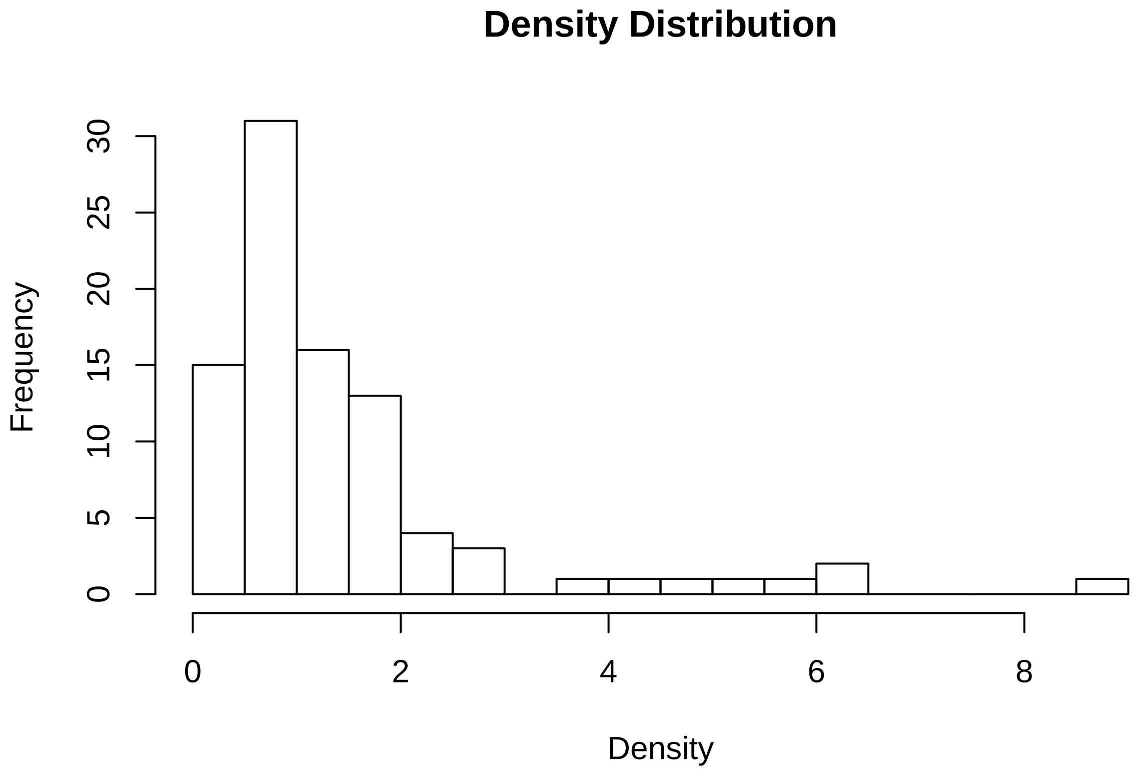
```
summary(crime_data$polpc)
```

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

The variable behaves as we expected, and we can move on to analyzing other variables.

9. Density (density): People per square mile. This variable should be above zero. Other than that, we should only take a deeper look at outliers.

```
hist(x = crime_data$density, breaks=20, main = "Density Distribution", xlab = "Density", ylab = "Frequen
```



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

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

There is a strangely small value for the minimum density, so we should take a deeper look:

```
crime_data[crime_data$density<0.0001,]
```

```
##    county year    crmrte    prbarr  prbconv prbpris avgsen      polpc
## 79    173   87 0.0139937 0.530435 0.327869    0.15   6.64 0.00316379
##        density    taxpc west central urban pctmin80    wcon     wtuc
## 79 2.03422e-05 37.72702    1       0     0  25.3914 231.696 213.6752
##        wtrd     wfir     wser   wmfg   wfed    wsta    wloc       mix
## 79 175.1604 267.094 204.3792 193.01 334.44 414.68 304.32 0.4197531
##       pctymle
## 79 0.07462687
```

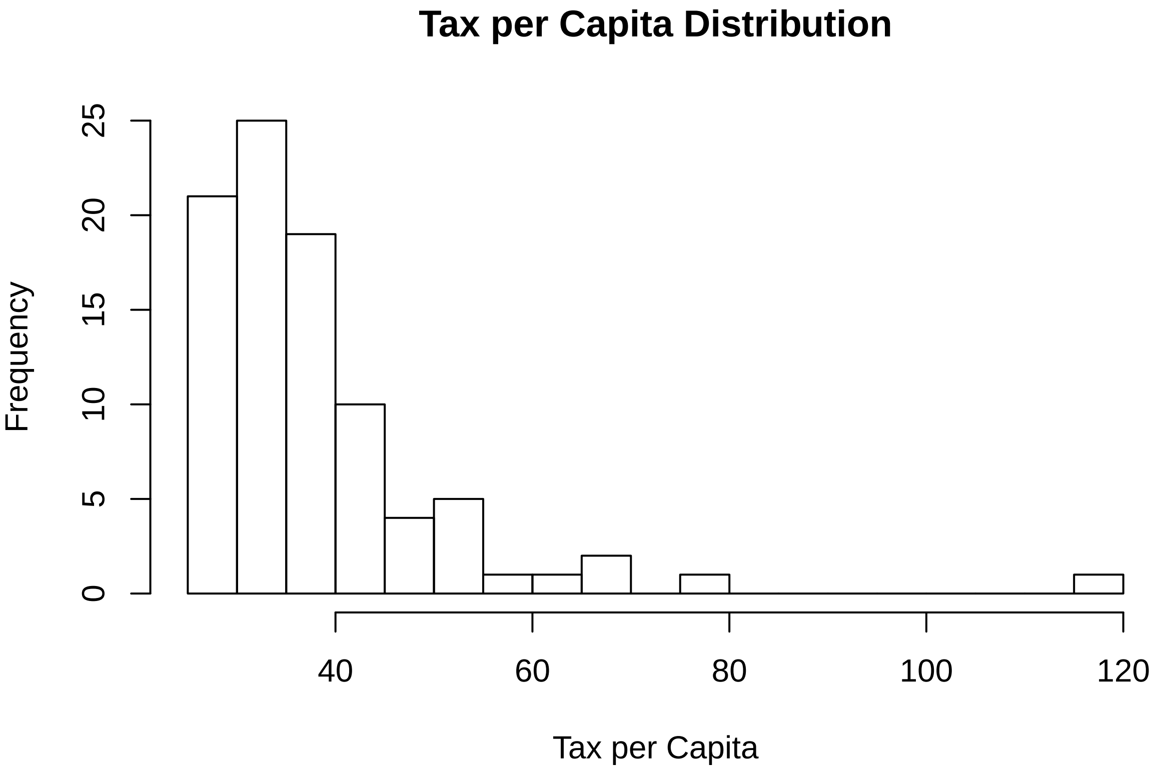
Searching for the FIPS code of this county (173), we see that it is Swain County. That is clearly an arithmetic error, and the true density value is 0.02. So we must correct it

```
crime_data$density[crime_data$density<0.0001]<- crime_data$density[crime_data$density<0.0001]*1000
crime_data[crime_data$county==173,]
```

```
##    county year    crmrte    prbarr  prbconv prbpris avgsen      polpc
## 79    173   87 0.0139937 0.530435 0.327869    0.15   6.64 0.00316379
##      density    taxpc west central urban pctmin80    wcon     wtuc
## 79 0.0203422 37.72702    1       0     0  25.3914 231.696 213.6752
##        wtrd     wfir     wser   wmfg   wfed    wsta    wloc       mix
## 79 175.1604 267.094 204.3792 193.01 334.44 414.68 304.32 0.4197531
##       pctymle
## 79 0.07462687
```

10.Tax Revenue per Capita (taxpc): This variable should be above zero. Other than that, we should only take a deeper look at outliers.

```
hist(x = crime_data$taxpc, breaks=20, main = "Tax per Capita Distribution", xlab = "Tax per Capita", yla
```



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

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

The observation in which tax per capita is almost 120 catches the eye, and so we should take a deeper look at that one.

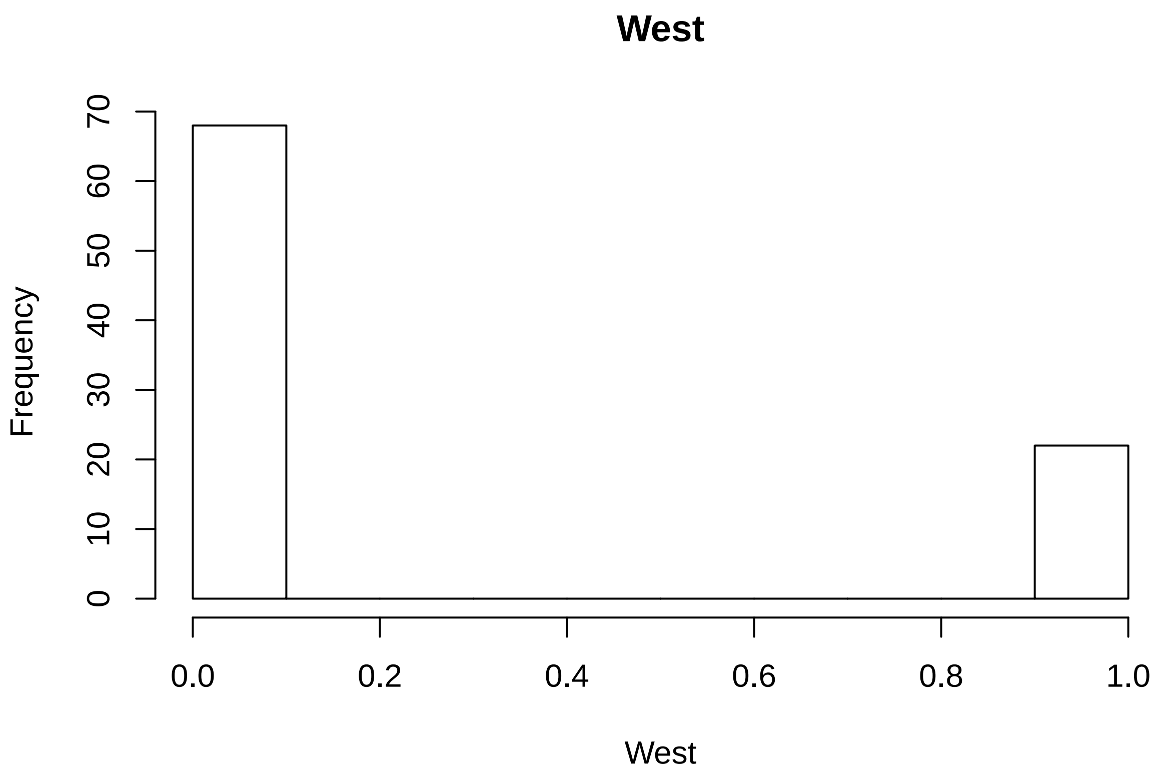
```
crime_data[crime_data$taxpc>100,]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 25     55   87 0.0790163 0.224628 0.207831 0.304348  13.57 0.00400962
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 25 0.5115089 119.7615    0       0     0  6.49622 309.5238 445.2762
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 25 189.7436 284.5933 221.3903 319.21 338.91 361.68 326.08 0.08437271
##       pctymle
## 25 0.07613807
```

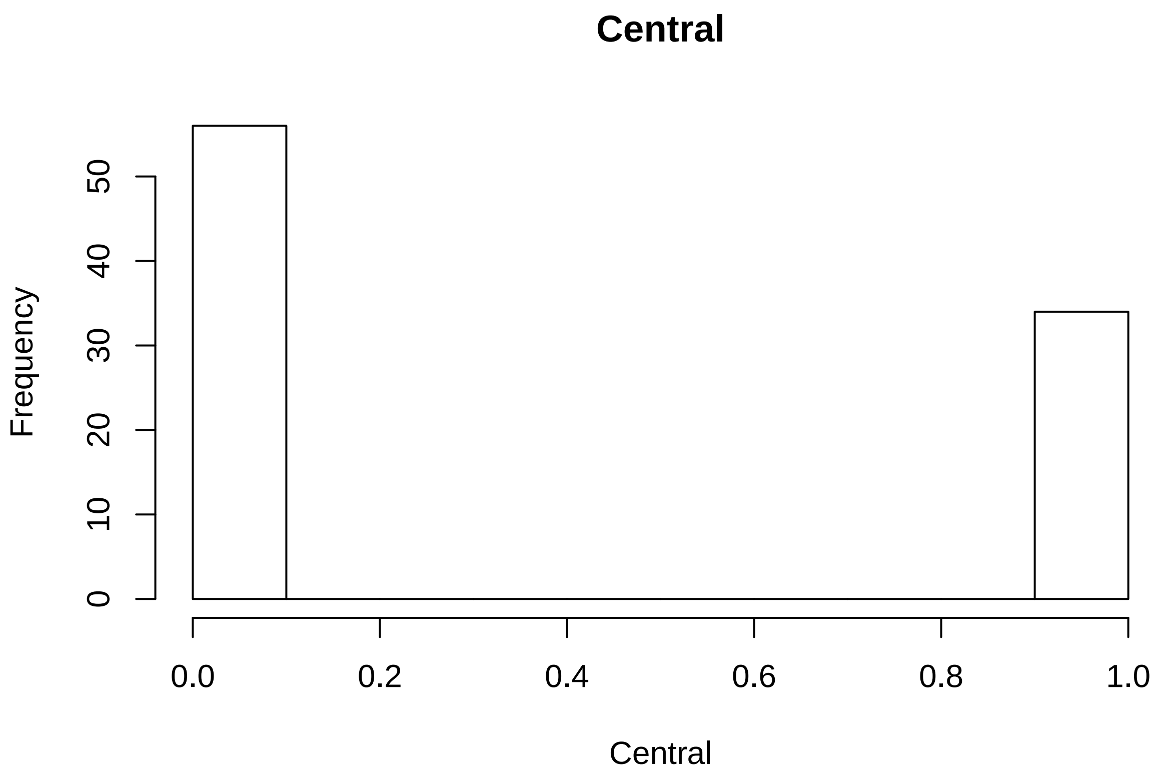
The other variables seem to be ok, so, it is safe to keep these observation.

11. West (west) / 12. Central (central) / 13. Urban (urban): Binary variables that indicate if the county is on West North Carolina, Central North Carolina or in SMSA. All of them should be either 0 or 1 for each observation.

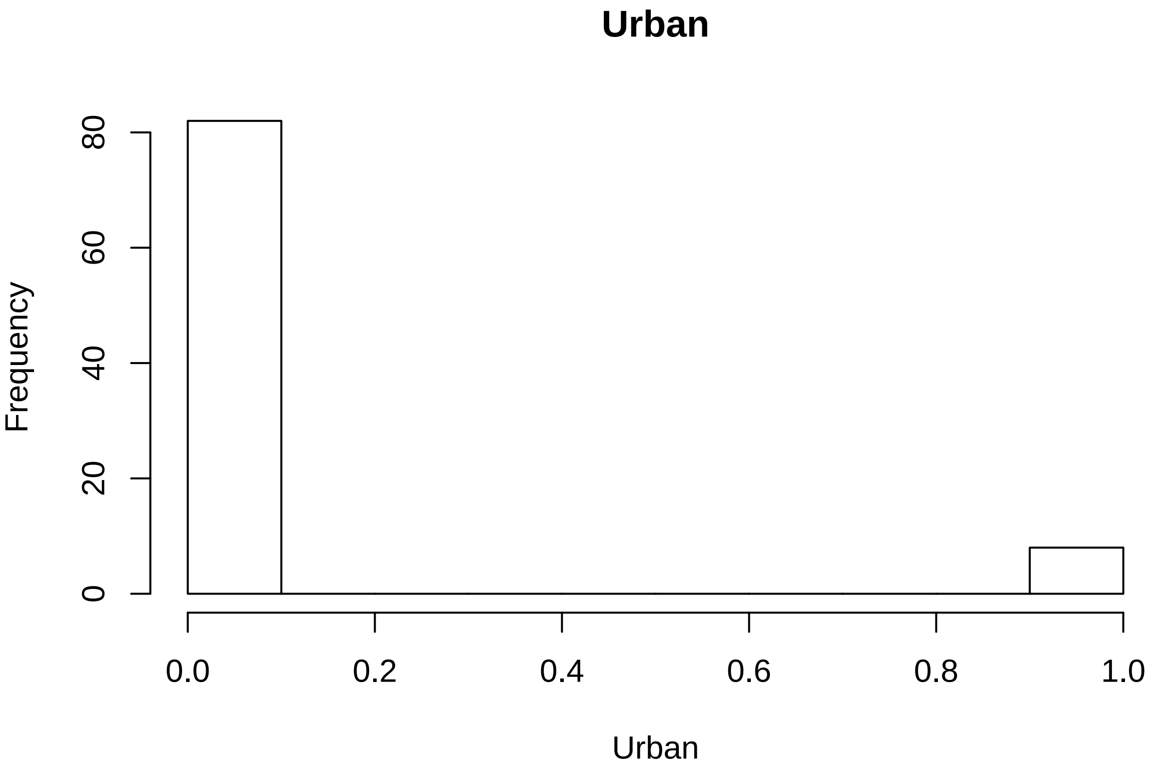
```
hist(x=crime_data$west, main = "West", xlab= "West", ylab= "Frequency")
```



```
hist(x=crime_data$central, main = "Central", xlab= "Central", ylab= "Frequency")
```



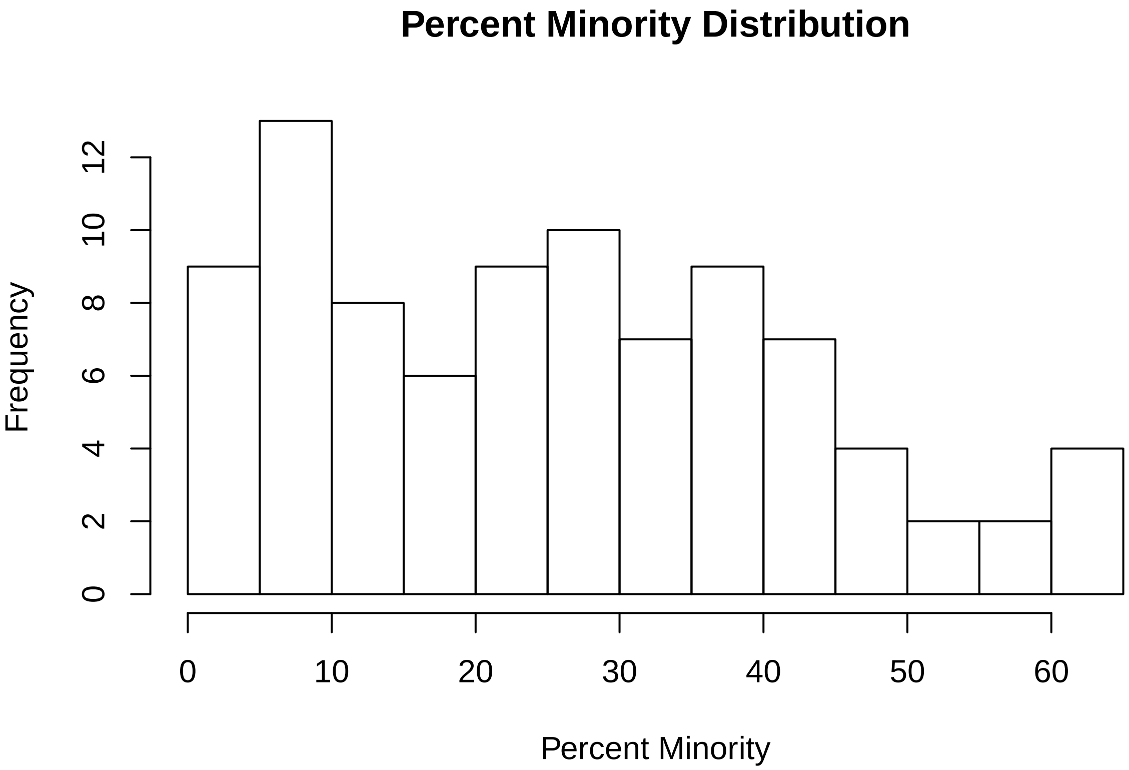
```
hist(x=crime_data$urban, main = "Urban", xlab= "Urban", ylab= "Frequency")
```



The variables behave as we expected, and we can move on to analyzing other variables.

14. Percent Minority, 1980 (pctmin80): Percentage of population within minority groups in the year of 1980. It should be between 0 and 1, because it represents the fraction of the population that is within minority groups

```
hist(x = crime_data$pctmin80, breaks=20, main = "Percent Minority Distribution", xlab = "Percent Minori
```



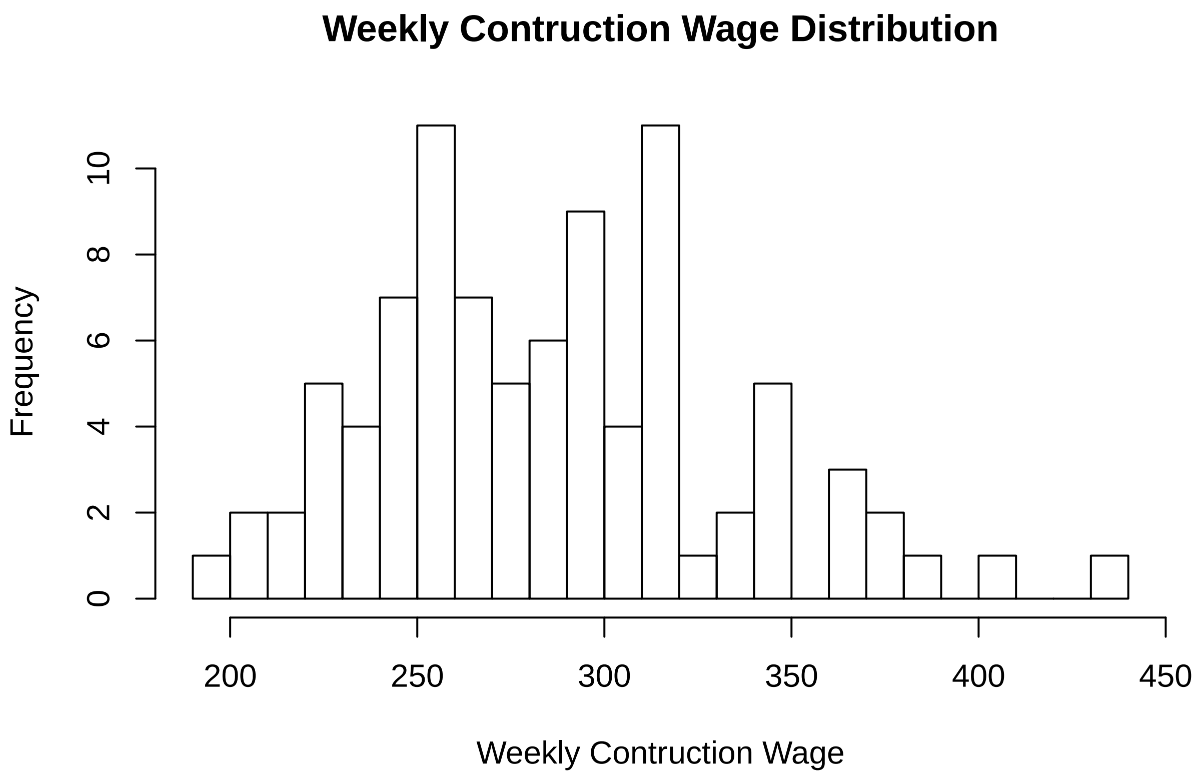
```
summary(crime_data$pctmin80)
```

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

The variable behaves as we expected, and we can move on to analyzing other variables.

15. Weekly Wage, Contruction (wcon) / 16. Weekly Wage, Transportation, Utilities and Community (wtuc) / 17. Weekly Wage, Wholesale and Retail Trade (wtrd) / 18. Weekly Wage, Financial, Insurance and Real Estate (wfir) / 19. Weekly Wage, Service Industry (wser) / 20. Weekly Wage, Manufacturing (wmfg) / 21. Weekly Wage, Federal Employees (wfed) / 22. Weekly Wage, State Employees (wsta) / 23. Weekly Wage, Local Government Employees (wloc): All of these variables refer to the average weekly wage in different sectors of the economy. We should check for outliers, and if they do happen, investigate them more deeply.

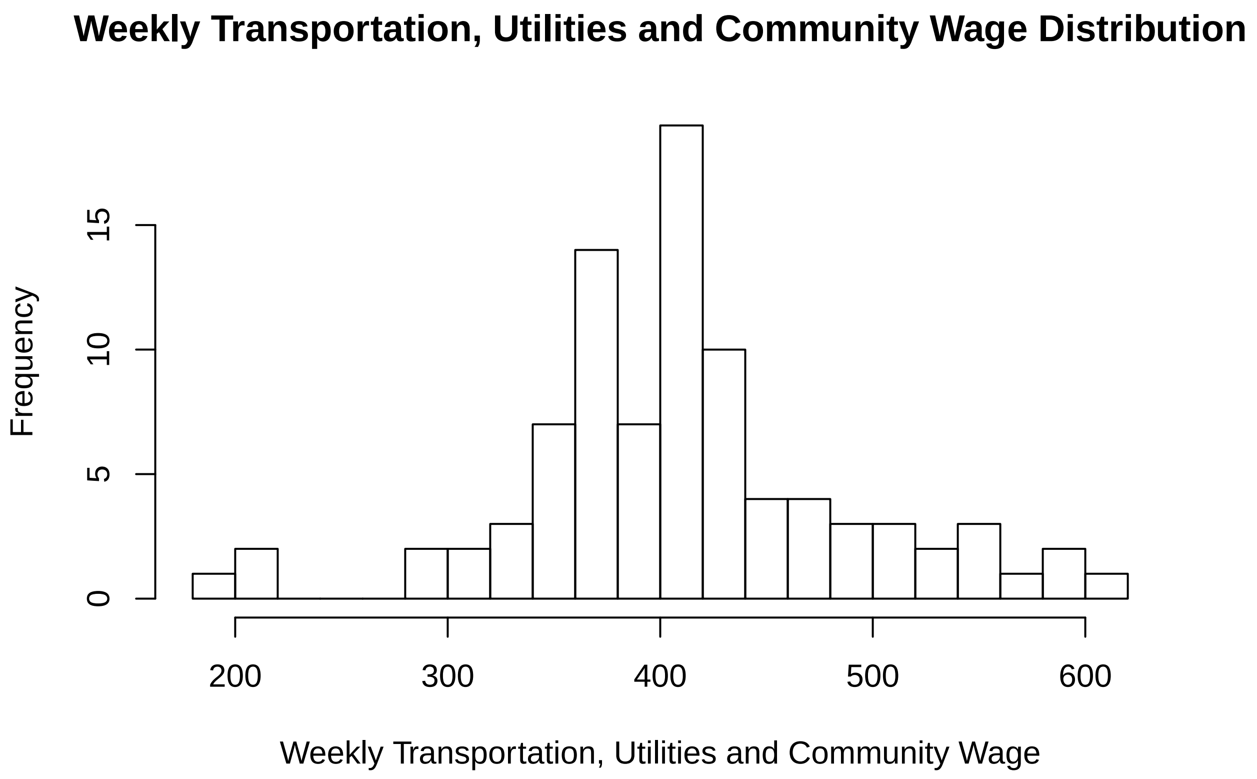
```
hist(x = crime_data$wcon, breaks=20, main = "Weekly Contruction Wage Distribution", xlab = "Weekly Cont
```



```
summary(crime_data$wcon)
```

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

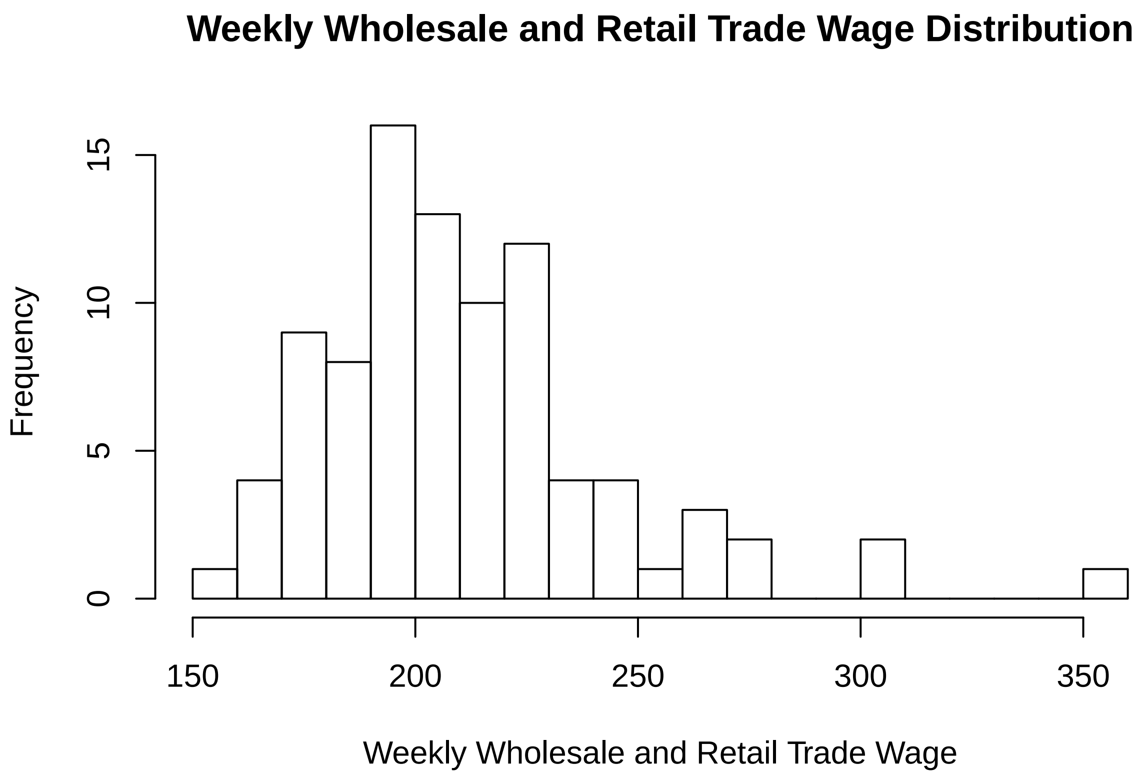
```
hist(x = crime_data$wtuc, breaks=20, main = "Weekly Transportation, Utilities and Community Wage Distri
```



```
summary(crime_data$wtuc)
```

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

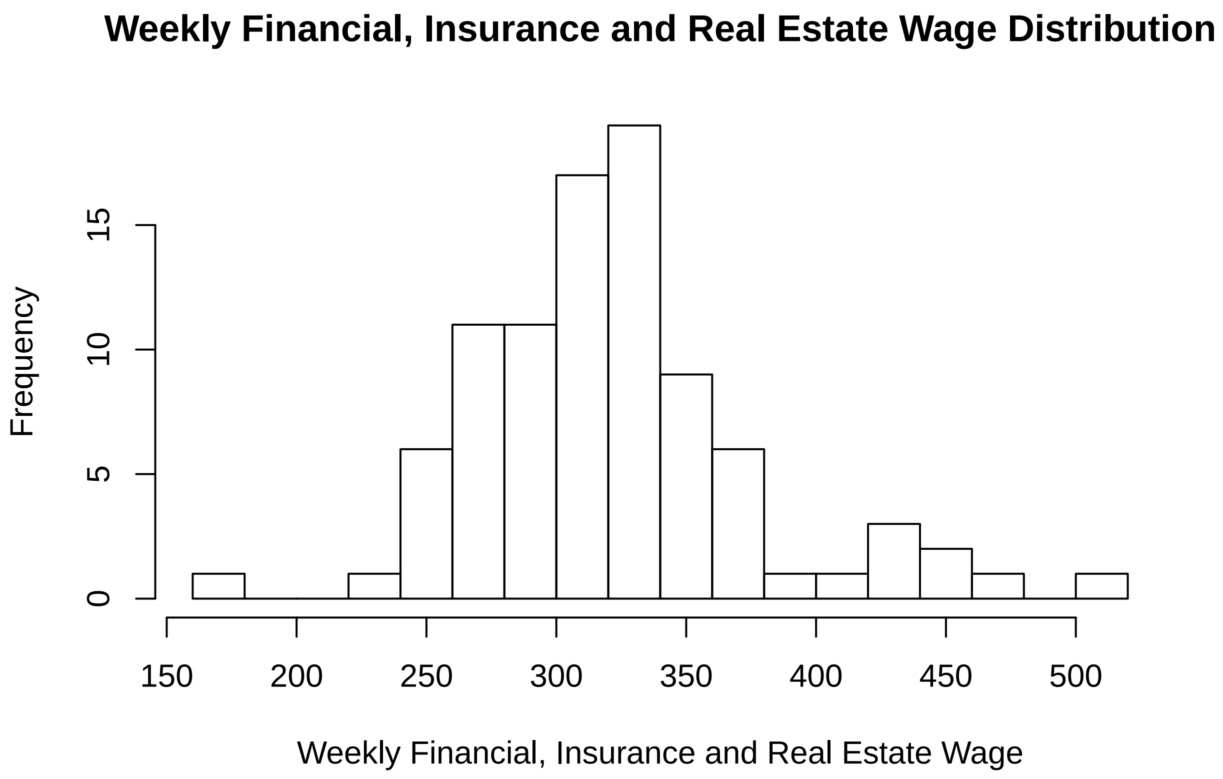
```
hist(x = crime_data$wtrd, breaks=20, main = "Weekly Wholesale and Retail Trade Wage Distribution", xlab
```



```
summary(crime_data$wtrd)
```

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

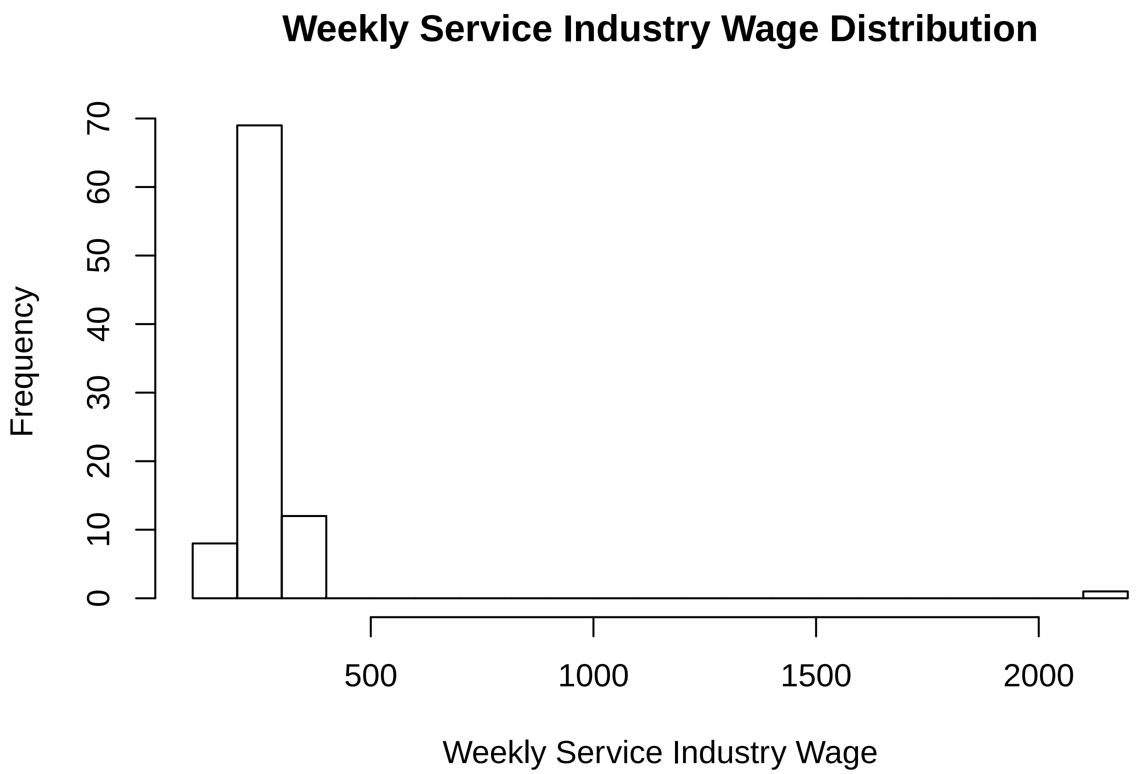
```
hist(x = crime_data$wfir, breaks=20, main = "Weekly Financial, Insurance and Real Estate Wage Distributi
```



```
summary(crime_data$wfir)
```

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

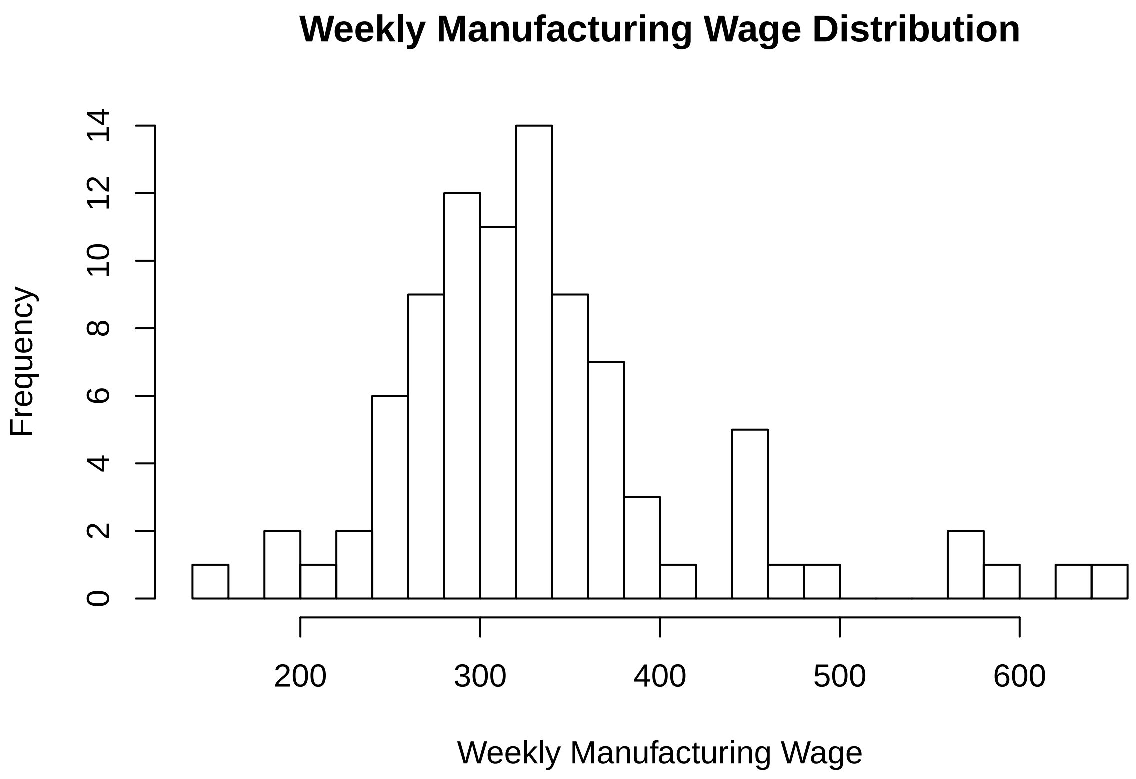
```
hist(x = crime_data$wser, breaks=20, main = "Weekly Service Industry Wage Distribution", xlab = "Weekly
```



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

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

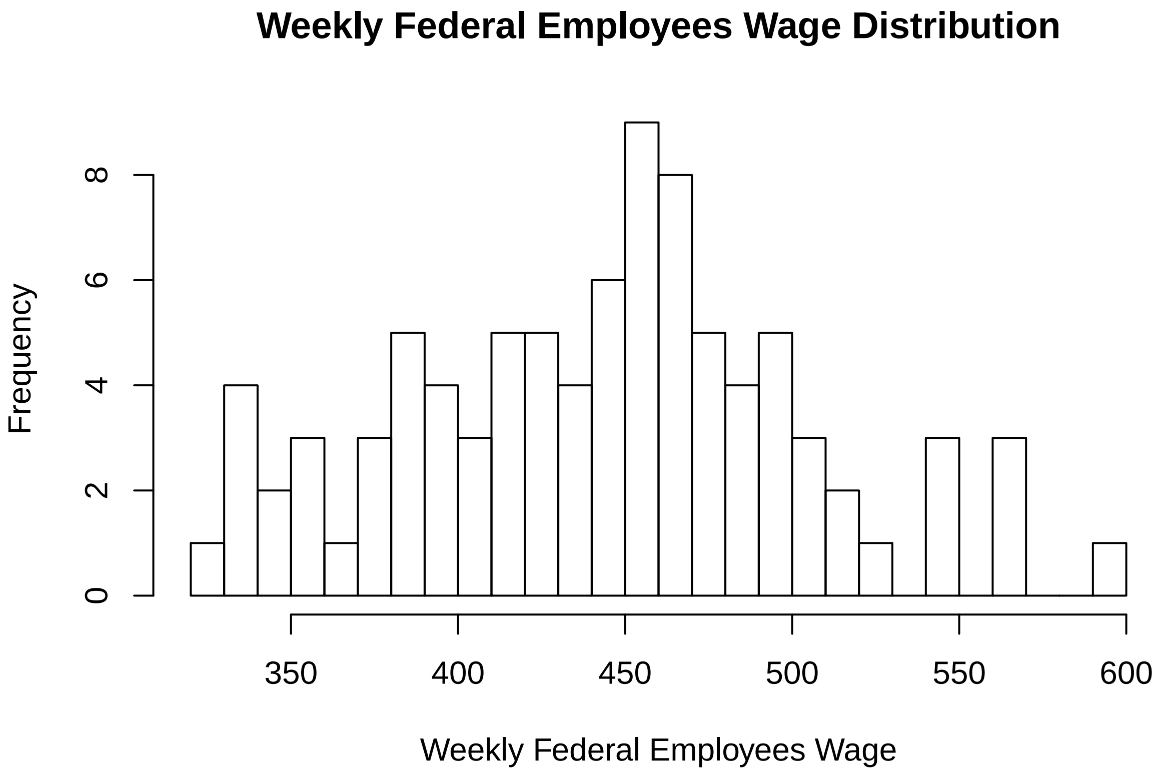
```
hist(x = crime_data$wmfg, breaks=20, main = "Weekly Manufacturing Wage Distribution", xlab = "Weekly Ma
```



```
summary(crime_data$wmfg)
```

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

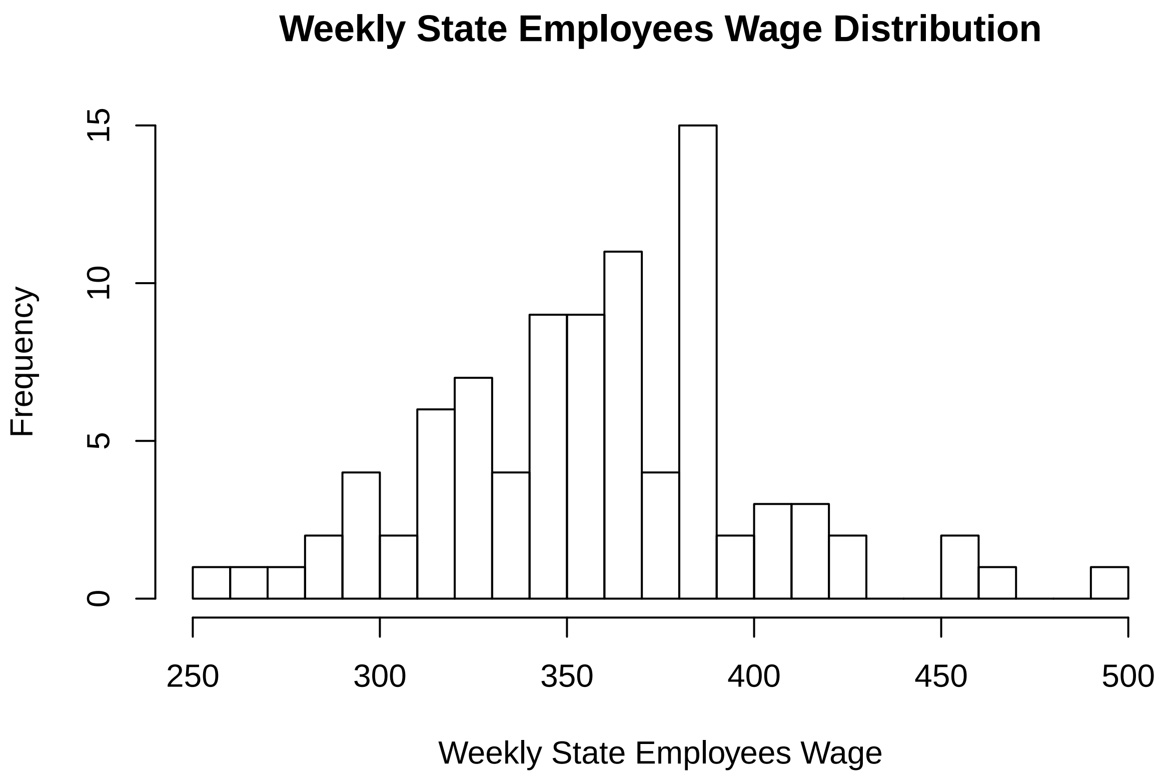
```
hist(x = crime_data$wfed, breaks=20, main = "Weekly Federal Employees Wage Distribution", xlab = "Weekl
```



```
summary(crime_data$wfed)
```

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

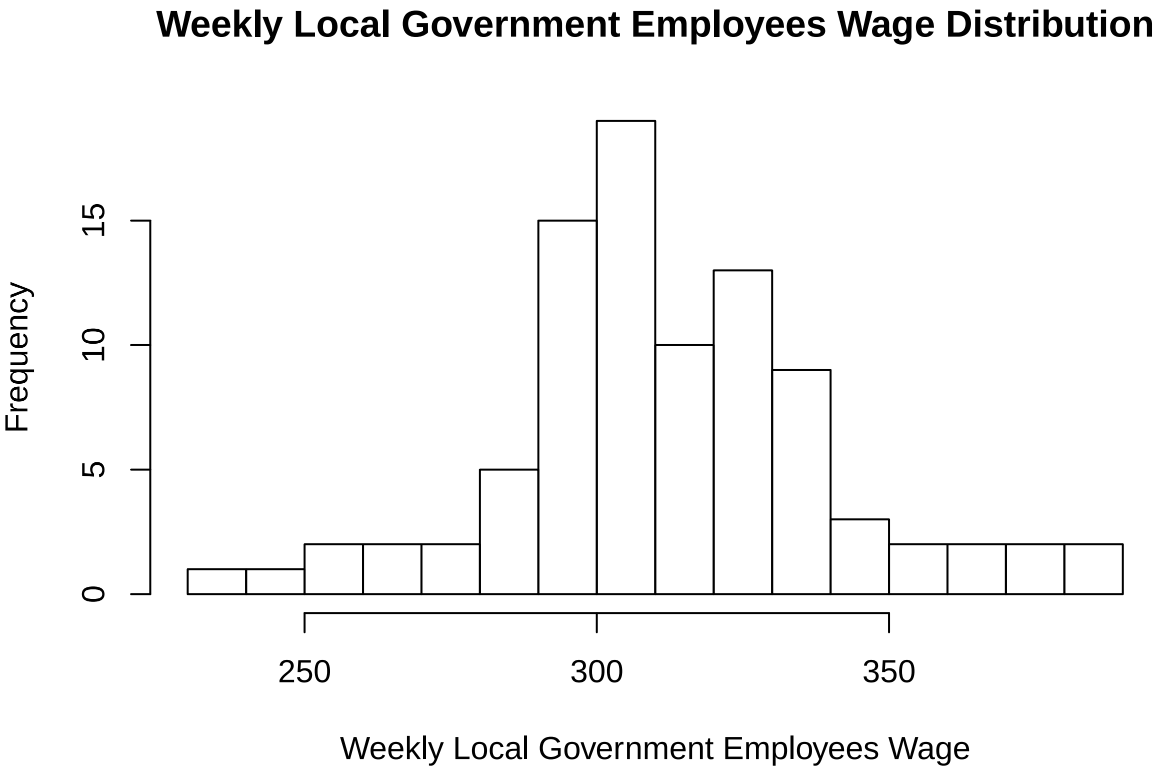
```
hist(x = crime_data$wsta, breaks=20, main = "Weekly State Employees Wage Distribution", xlab = "Weekly
```



```
summary(crime_data$wsta)
```

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

```
hist(x = crime_data$wloc, breaks=20, main = "Weekly Local Government Employees Wage Distribution", xlab
```



```
summary(crime_data$wloc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   239.2   297.2   307.6   312.3   328.8   388.1
```

For the service industry, there is one observation in particular that catches the eye, which is way above the second largest value. For that, we take a deeper look

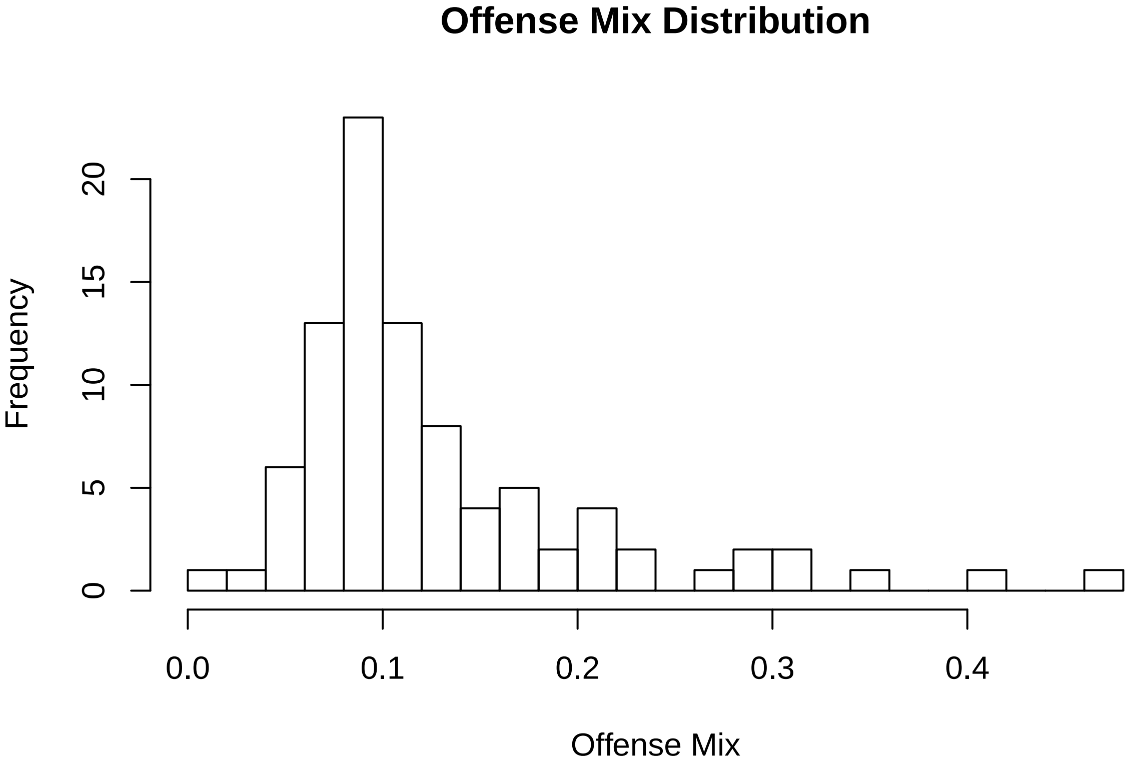
```
crime_data[crime_data$wser>2000,]
```

```
##    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
## 84 0.07008217
```

It is county 185, Warren County. The only sector that has a weekly wage so much higher than for the other counties is the service industry, with all other sectors having a weekly wage very close to the state average. One might wonder if this county is particularly attractive for tourism, or some other sort of services, to explain such a difference. That is not the fact: Warren county is a center of tobacco and cotton plantations,educational later textile mills (https://en.wikipedia.org/wiki/Warren_County,_North_Carolina). It is very likely a dot was misplaced, and the actual value is 217.7068 instead of 2177.068. However, since we cannot atest that with certainty, we will leave the value as it is, and will not discard the observation.

24. Offense mix, face-to-face / other (mix): Represents the ratio of criminal offenses made face-to-face (such as armed robbery) to other types. The values can range within any positive number, however, we should dig deeper in the case of outliers.

```
hist(x = crime_data$mix, breaks=20, main = "Offense Mix Distribution", xlab = "Offense Mix", ylab = "Fre
```



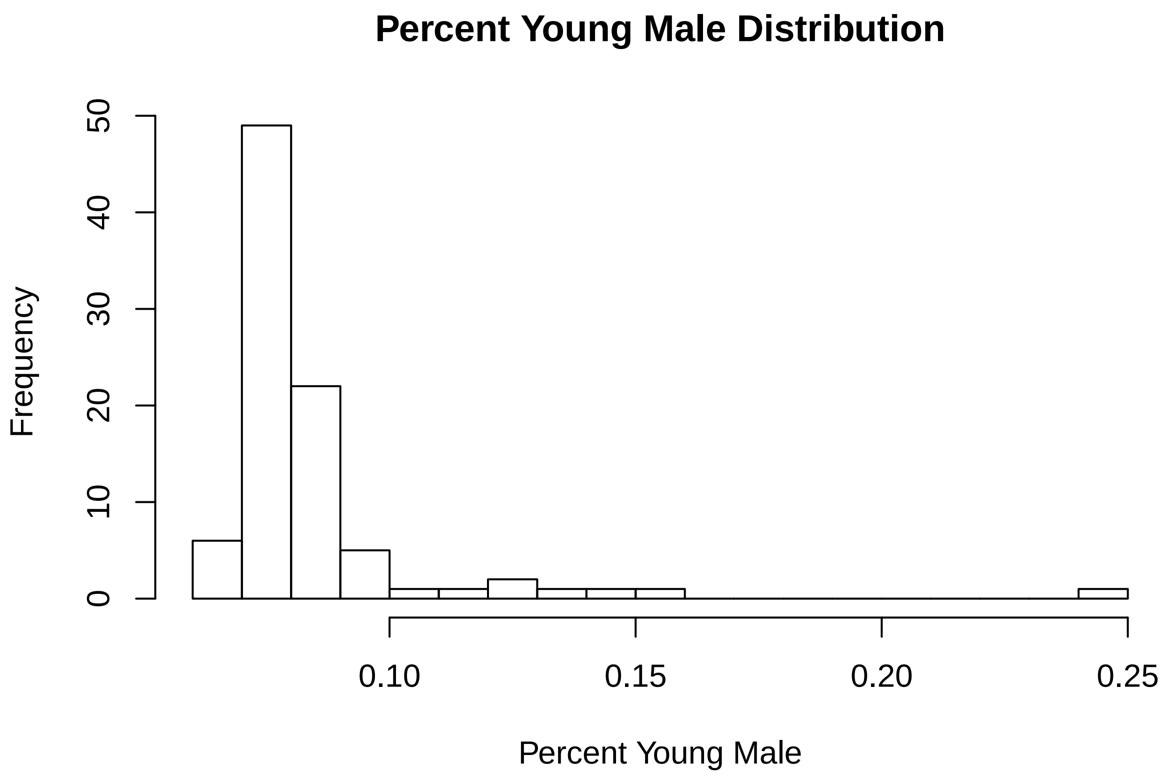
```
summary(crime_data$mix)
```

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

The variable behaves as we expected, and we can move on to analyzing other variables.

25. Percent Young Male (pctymle): Represents the percent of the population composed by males between the age of 15 and 24. Should be a number between 0 and 1.

```
hist(x = crime_data$pctymle, breaks=20, main = "Percent Young Male Distribution", xlab = "Percent Young
```



```
summary(crime_data$pctymle)
```

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

The variable behaves as expected and now we can finally move on to the research question.