Lab 3 DRAFT

Siddhartha Jakkamreddy, Neha Kumar, Brian Musisi

11/14/2018

Introduction

The motivation for this analysis is to determine the factors that lead to crime rate in North Carolina counties in 1980. We are assuming the role of data scientists for a political campaign around the same era within North Carolina to determine methods that can be employed to reduce the crime rate. Note, that this requires our analysis to look for causal variables so we can provide concrete and actionable resolutions.

Initial EDA

```
crime = read.csv("crime_v2.csv", header = TRUE)
str(crime)
```

```
## '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 ...
```

Here, we see something interesting. prbconv is a factor due to a rogue ‘ character being added to the bottom of the file. When we view the dataframe, we actually see that this rogue tick mark has also introduced 6 null values in the file. We take steps to remove these records from the file to clean our dataset. We will remove the duplicate value for county 193

```
crime <- crime[!is.na(as.numeric(as.character(crime$prbconv))),]
```

```
## Warning in `[.data.frame`(crime, !
## is.na(as.numeric(as.character(crime$prbconv))), : NAs introduced by
## coercion
```

```
crime$prbconv <- as.numeric(as.character(crime$prbconv))
crime = crime[!duplicated(crime), ]
str(crime)
```

```
## 'data.frame':    90 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 : num  0.528 1.481 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 ...
```

```
summary(crime)
```

```
##      county           year        crmrte             prbarr       
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277  
##  1st Qu.: 51.5   1st Qu.:87   1st Qu.:0.020604   1st Qu.:0.20495  
##  Median :103.0   Median :87   Median :0.030002   Median :0.27146  
##  Mean   :100.6   Mean   :87   Mean   :0.033510   Mean   :0.29524  
##  3rd Qu.:150.5   3rd Qu.:87   3rd Qu.:0.040249   3rd Qu.:0.34487  
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091  
##     prbconv           prbpris           avgsen           polpc          
##  Min.   :0.06838   Min.   :0.1500   Min.   : 5.380   Min.   :0.0007459  
##  1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375   1st Qu.:0.0012378  
##  Median :0.45170   Median :0.4222   Median : 9.110   Median :0.0014897  
##  Mean   :0.55086   Mean   :0.4106   Mean   : 9.689   Mean   :0.0017080  
##  3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465   3rd Qu.:0.0018856  
##  Max.   :2.12121   Max.   :0.6000   Max.   :20.700   Max.   :0.0090543  
##     density            taxpc             west           central      
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.54718   1st Qu.: 30.73   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.97925   Median : 34.92   Median :0.0000   Median :0.0000  
##  Mean   :1.43567   Mean   : 38.16   Mean   :0.2444   Mean   :0.3778  
##  3rd Qu.:1.56926   3rd Qu.: 41.01   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##      urban            pctmin80           wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8   1st Qu.:374.3
##  Median :0.00000   Median :24.852   Median :281.2   Median :404.8
##  Mean   :0.08889   Mean   :25.713   Mean   :285.4   Mean   :410.9
##  3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0   3rd Qu.:440.7
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3   1st Qu.:288.6
##  Median :203.0   Median :317.1   Median : 253.1   Median :321.1
##  Mean   :210.9   Mean   :321.6   Mean   : 275.3   Mean   :336.0
##  3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6   3rd Qu.:359.9
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2   1st Qu.:0.08060
##  Median :448.9   Median :358.4   Median :307.6   Median :0.10095
##  Mean   :442.6   Mean   :357.7   Mean   :312.3   Mean   :0.12905
##  3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8   3rd Qu.:0.15206
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07437
##  Median :0.07770
##  Mean   :0.08403
##  3rd Qu.:0.08352
##  Max.   :0.24871
```

Something we notice here is that both the probability of arrest and probability of conviction have at least 1 record that is over 1. Thinking through this further, this is not impossible. Multiple people can participate in a crime together, leading to multiple arrests and/or convictions per criminal offense and so this anomaly may be a product of the operationalization of the particular variables. Thus, we will not discard this variable.

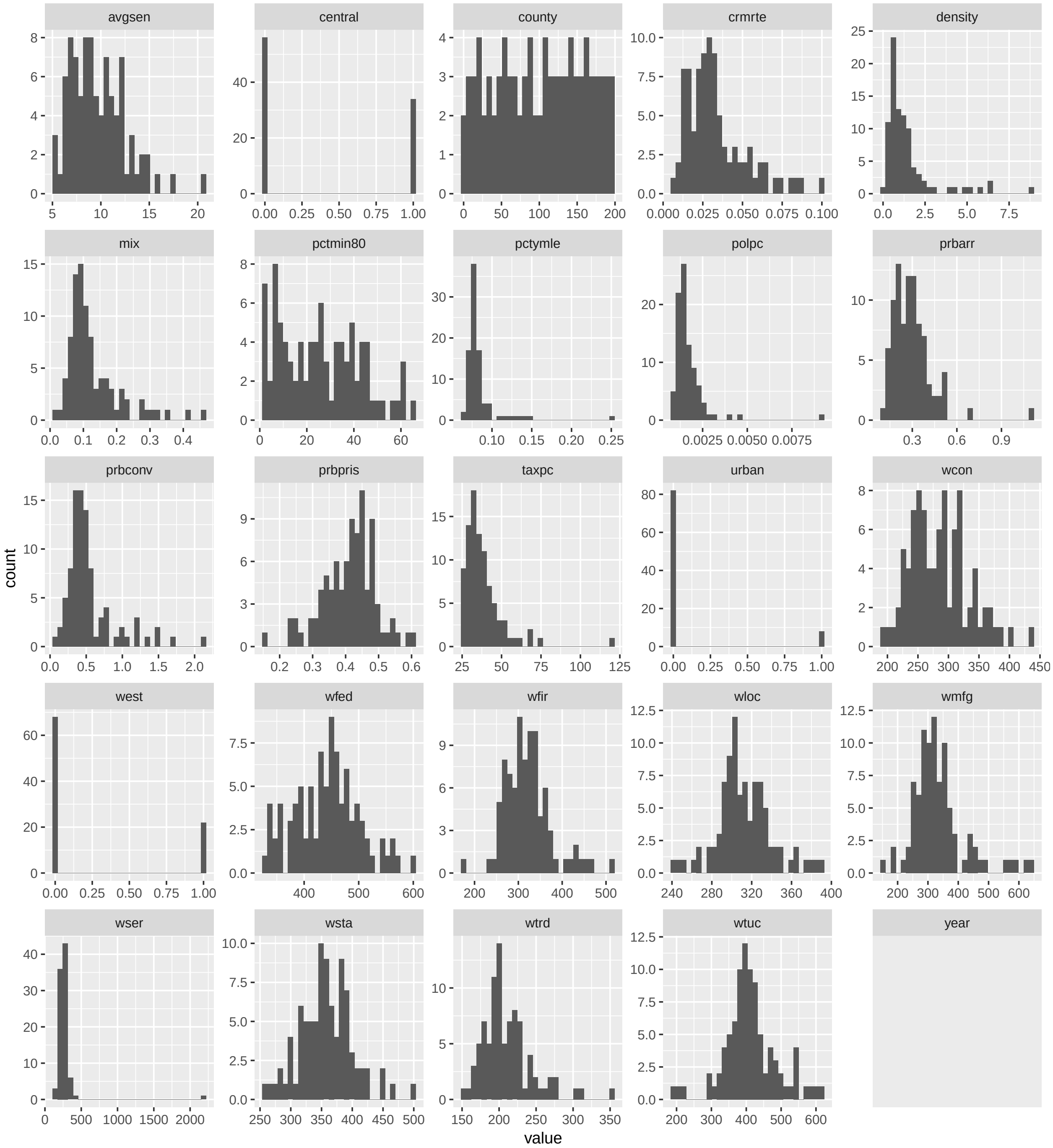
Our data file is now clean and ready for further analysis.

Model Building Process

```r
crime %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Computation failed in `stat_bin()`:
## `binwidth` must be positive
```

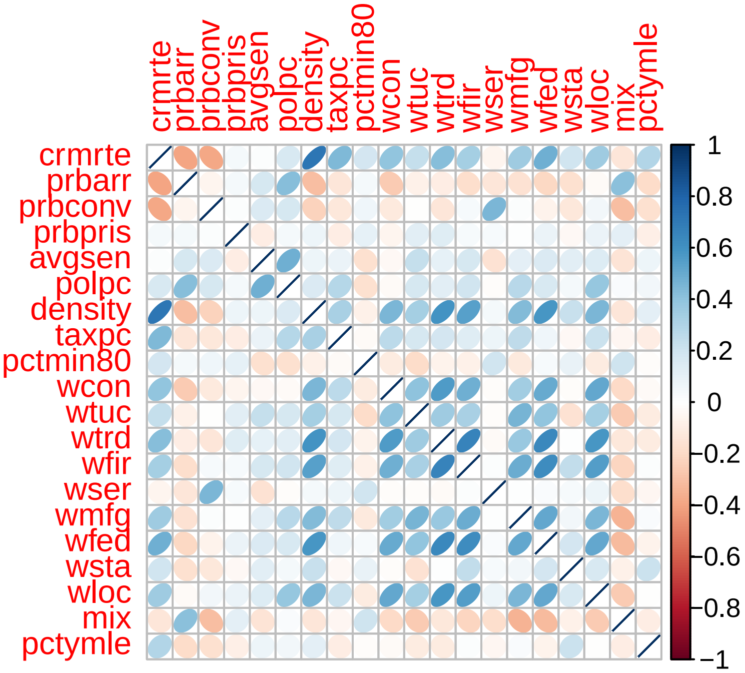


From the above, the distribution of most of the numeric variables presented are approximately normal, with the exception of prbconv, pctymyle and polpc that stand out as non-normal. We now take a look at a scatterplot matrix to examine the bivariate relationships embedded within this dataset.

Looking at the variables, we have a few initial thoughts. There is suspected collinearity between wage variables and tax revenue per capita. We also note that higher tax per capita allows counties to spend more money on police forces, possibly having an effect on the police per capita. Lastly, density could have a confounding effect on per capita variables. Assume we have 2 counties, ceteris paribis, differing only in population density. This means that the per capita measurements for the more densely populated county will be lower than the more sparsely populated county.

We create a correlation matrix to get a high level overview of the correlations between variables.

```
corr_crime = cor(crime[, c(-1,-2,-11,-12,-13)])  ### removing the urban, west and central variables due
corrplot(corr_crime, method = "ellipse")
```

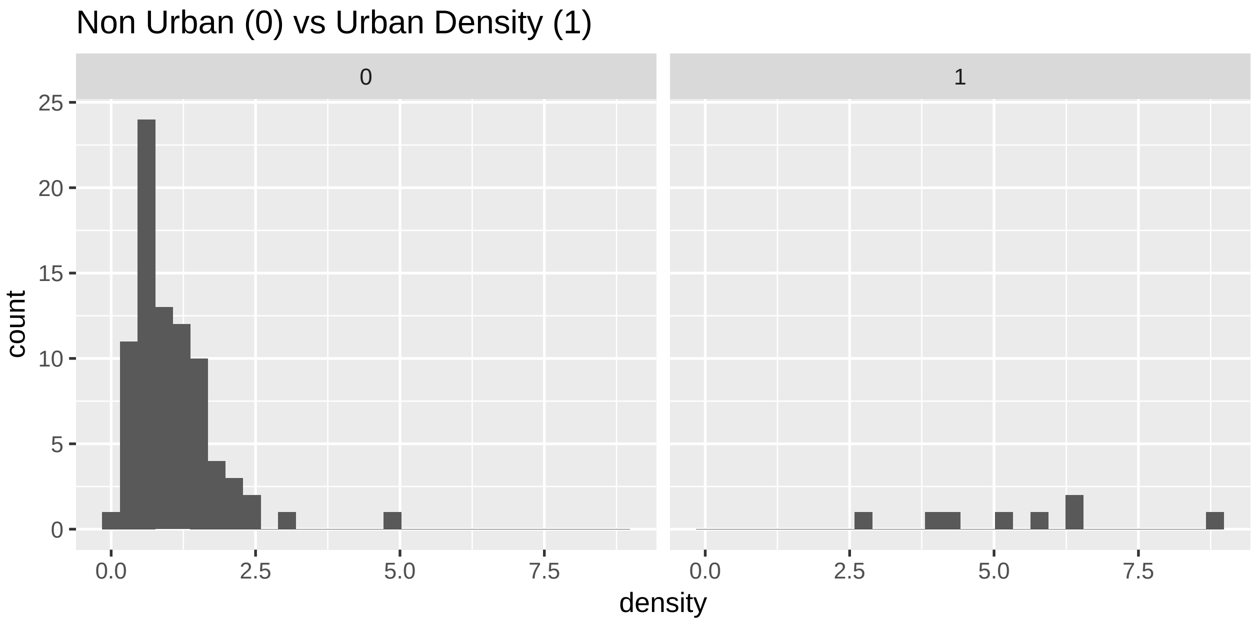


Examining density

The first relationships we decide to investigate further are the density variables against the binary urban, west and central variables. From the correlation matrix and our intuition, we expect that urban areas are more dense.

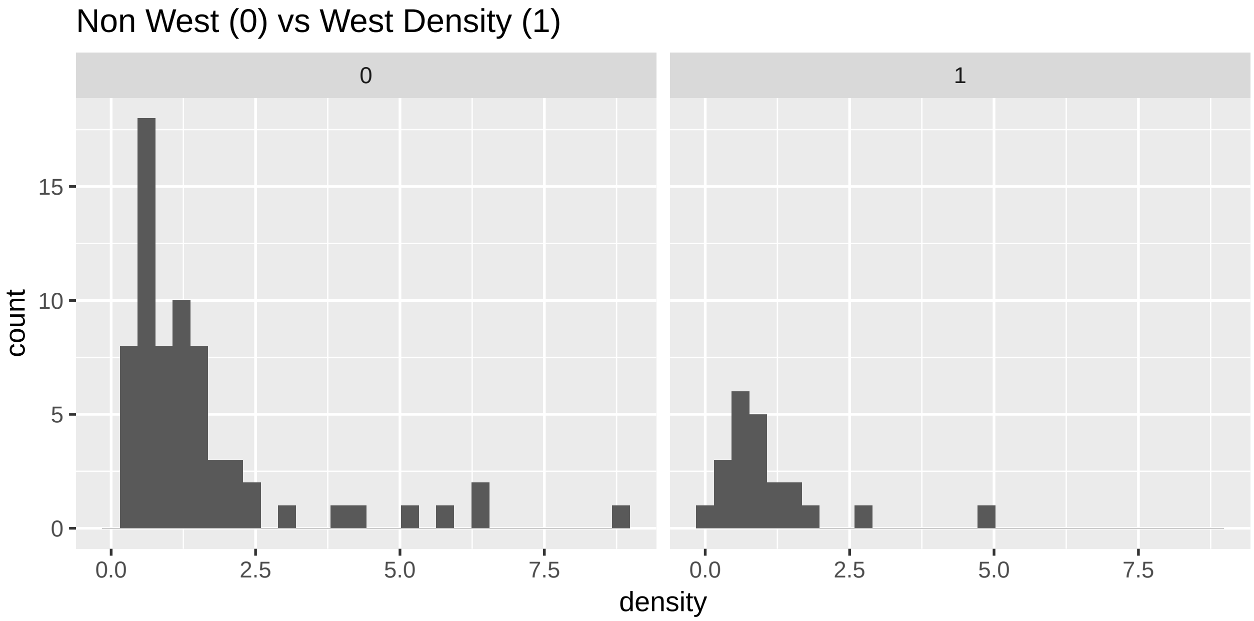
```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ urban) +
    geom_histogram() +
    ggtitle("Non Urban (0) vs Urban Density (1)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



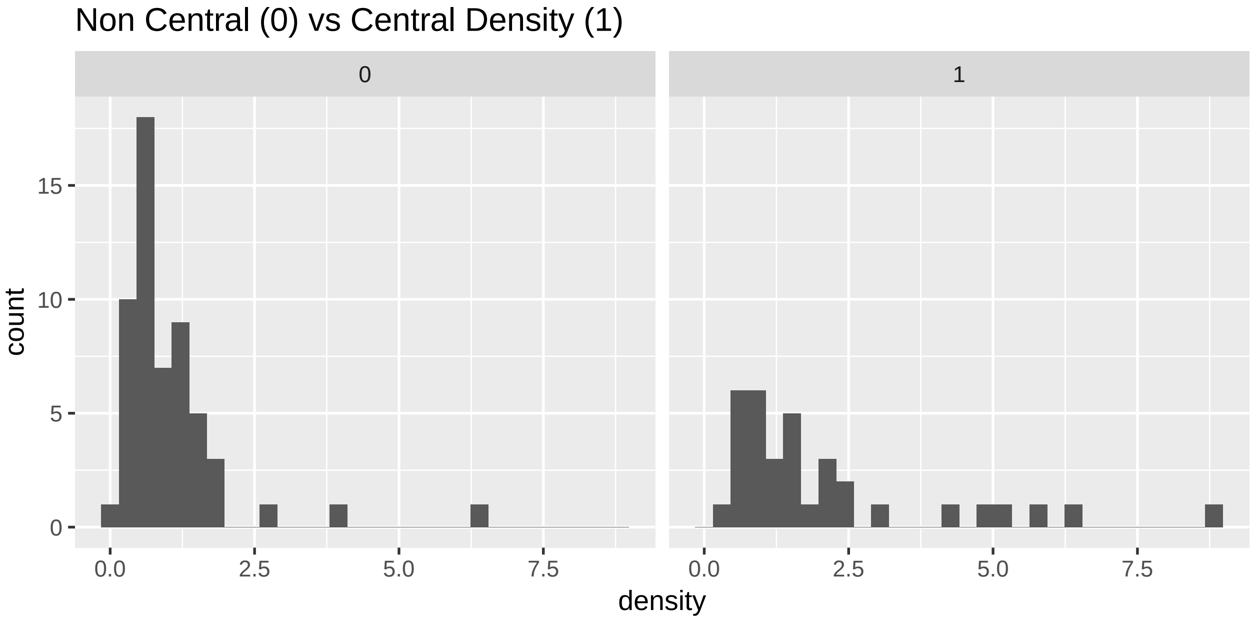
```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ west) +
    geom_histogram() +
    ggtitle("Non West (0) vs West Density (1)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ central) +
    geom_histogram() +
    ggtitle("Non Central (0) vs Central Density (1)")
```

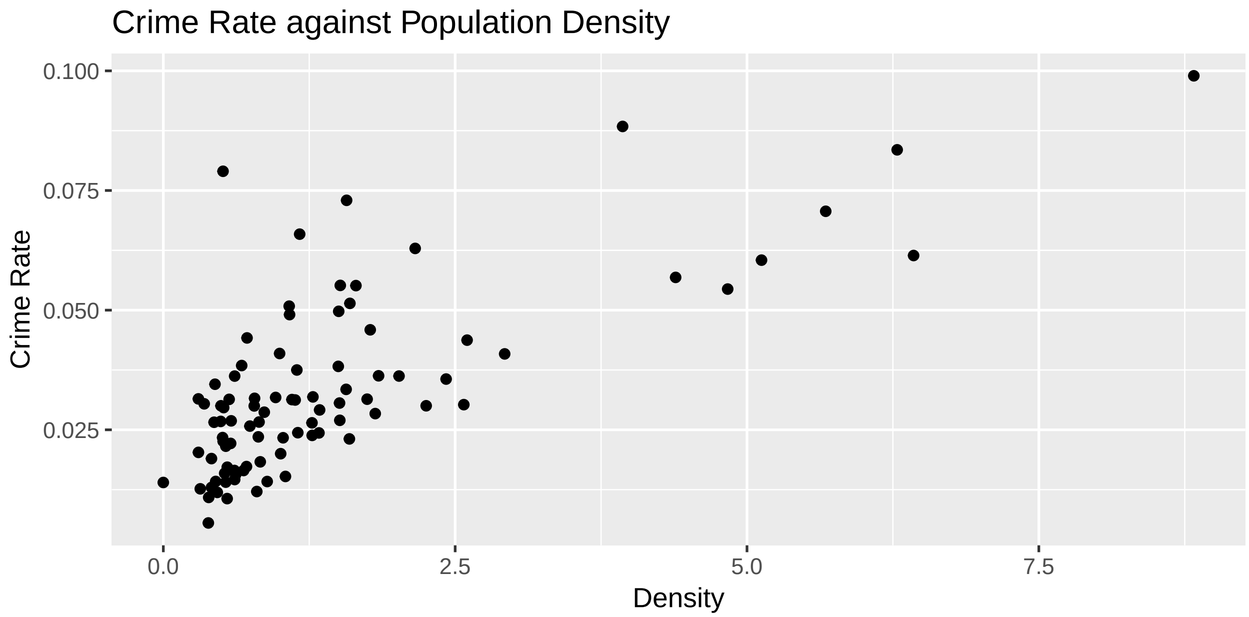
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Urban areas tend to have higher densities, as expected. Thus, the density and urban variables are collinear. The other 2 variables don't show as strong of a relationship with density, and from the scatterplot matrix are not particularly correlated with crime rate. Thus, these variables will likely not be included in our first model specification.

Taking a closer look at crime rate against population density, this does look like a promising variable to include in the first specification.

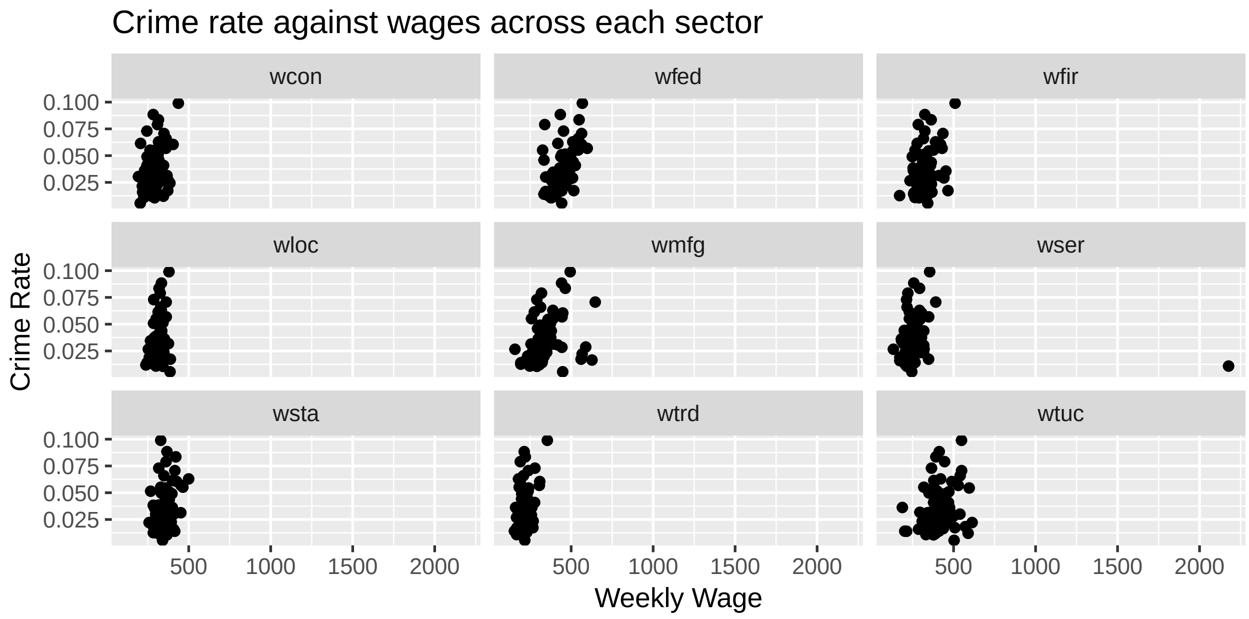
```r
ggplot(crime, aes(x=density, y =crmrte)) +
  geom_point() +
  ggtitle("Crime Rate against Population Density") +
  xlab("Density") +
  ylab("Crime Rate")
```



Examining Income-related variables

Next, we notice from the scatterplot matrix that each wage variable seems to be highly correlated with each other, and there is some positive correlation with the crime rate. We investigate these closer to see any opportunities for transformation.

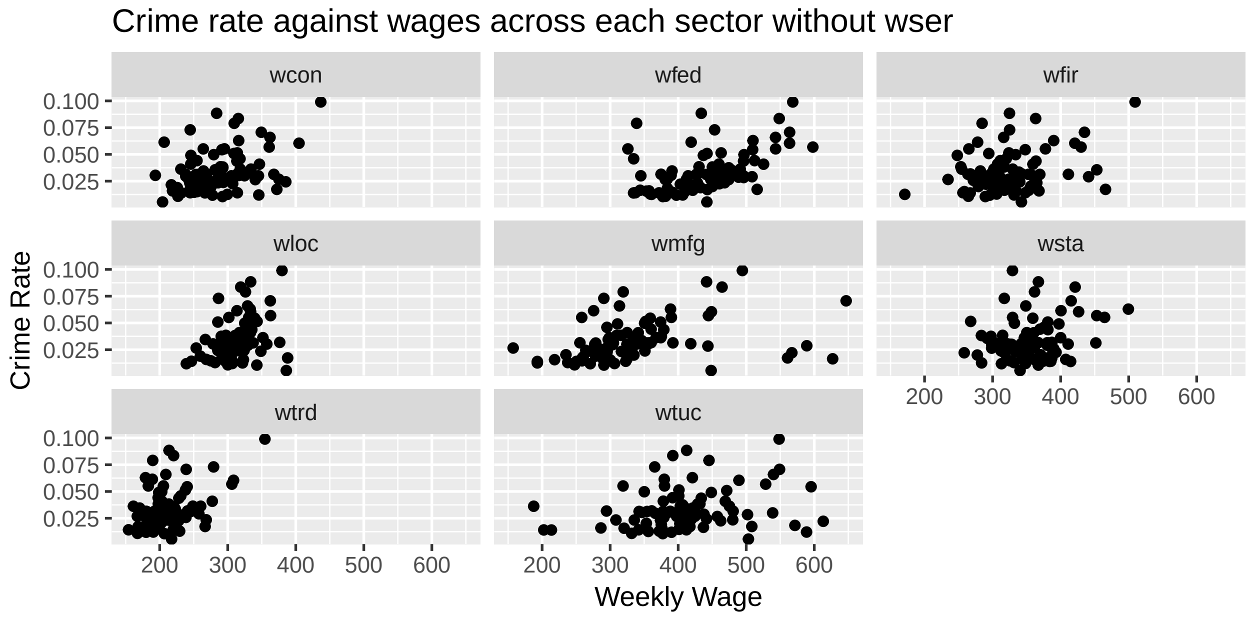
```r
crime_wage <- crime %>%
  select(crmrte, wcon, wtuc, wtrd, wfir, wmfg, wfed, wser, wsta, wloc) %>%
  gather(sector, wkly_wage, -crmrte)
ggplot(crime_wage, aes(x=wkly_wage, y=crmrte)) +
  facet_wrap(~sector) +
  geom_point() +
  ggtitle("Crime rate against wages across each sector") +
  xlab("Weekly Wage") +
  ylab("Crime Rate")
```



```
#+ theme(strip.text.x = element_text(margin = margin(2,0,2,0, "cm")))
```

Here, we see that the outlier in wser affects the x-axis is distorting the x axes for the other facets. We run the same graph without wser for a clearer visual of the other wage variables.

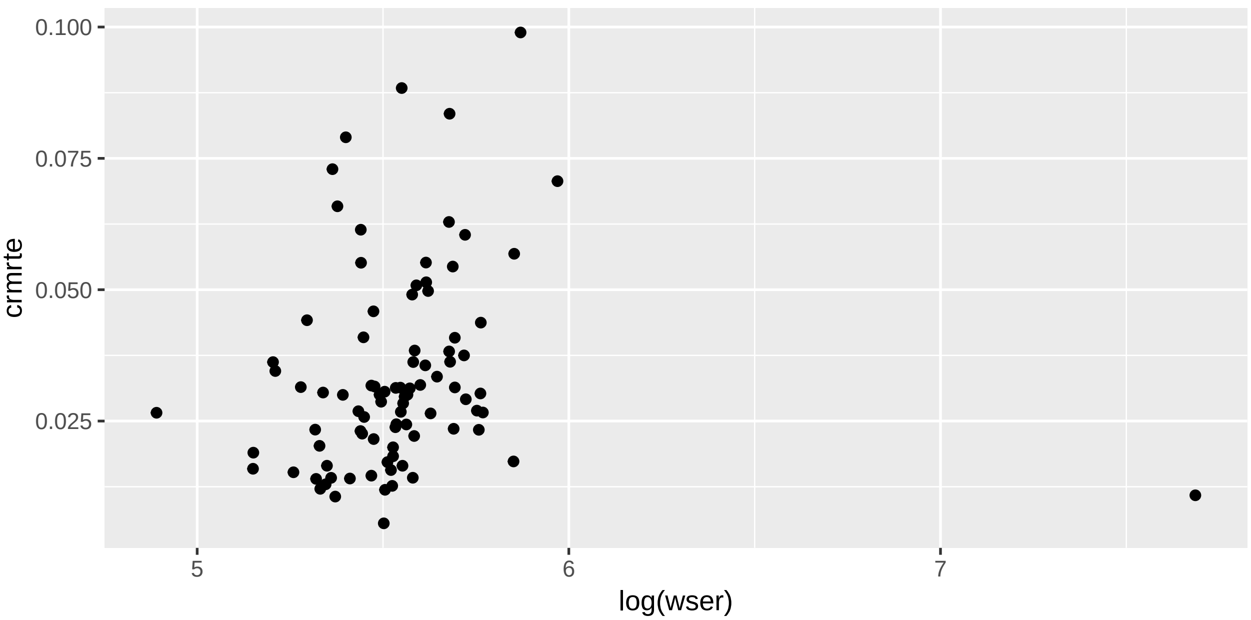
```
crime_wage <- crime %>%
  select(crmrte, wcon, wtuc, wtrd, wfir, wmfg, wfed, wsta, wloc) %>%
  gather(sector, wkly_wage, -crmrte)
ggplot(crime_wage, aes(x=wkly_wage, y=crmrte)) +
  facet_wrap(~sector) +
  geom_point() +
  ggtitle("Crime rate against wages across each sector without wser") +
  xlab("Weekly Wage") +
  ylab("Crime Rate")
```



```
#+ theme(strip.text.x = element_text(margin = margin(2,0,2,0, "cm")))
```

Many of the wage variables have a slight positive relationship with crime rate. The variables wtrd, wmfg, wfed, and (to a lesser degree) wsta and wloc seem to have a good amount of correlation with crime rate. As shown below, wser seems to have a very slight correlation with crime rate (log used because of the outlier)

```
ggplot(crime, aes(x=log(wser), y=crmrte)) + geom_point()
```



These relationships explain the phenomenon that more burglaries / thefts / kidnappings are likely to be targeted on wealthier victims, so areas with higher incomes would end up having higher crime rates.

The one point that stands out is the outlier in wser (service industry wage) when plotted against the crime rate. This reflects a county that has a substantially high wage for service workers, and has a lower crime rate. This is likely an area that has been highly gentrified and is predominantly populated my members of in service roles (with fewer individuals who are in the other, lower paying industries.) We don't believe that removing this outlier is valid, as this county could have another attribute that is worth investigating as a case study of a "successful" community with a low crime rate, and would be of high interest to our political campaign.

Taking a closer look at this outlier:

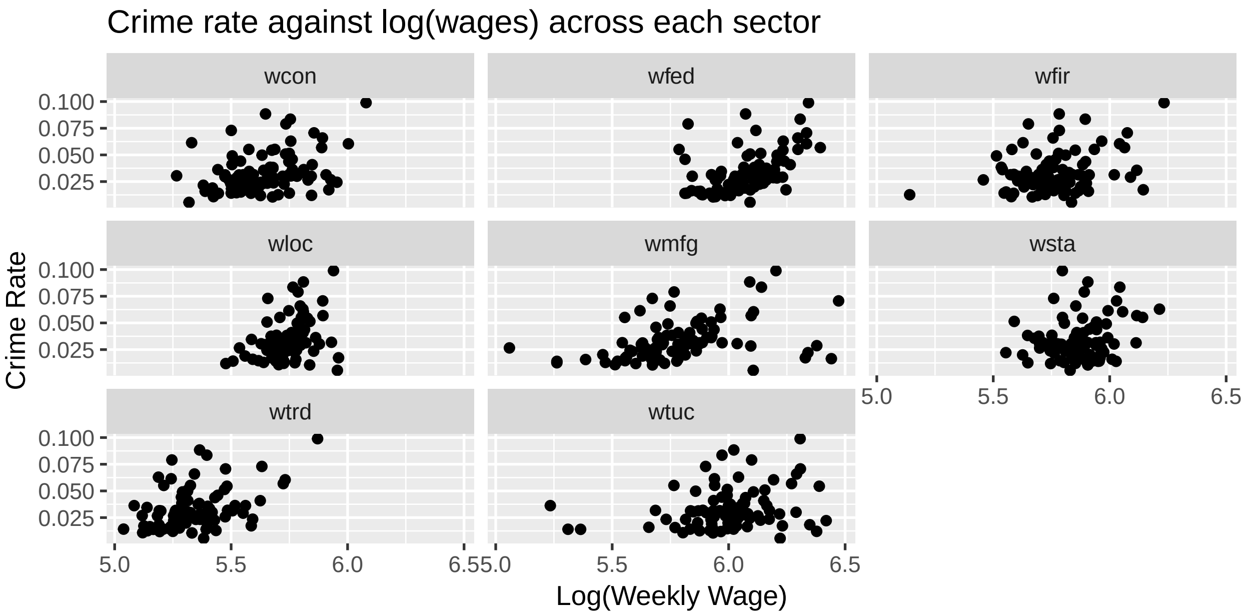
```
crime_outlier <- crime %>% filter(wser>2000)
head(crime_outlier)
```

```
##   county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 1    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##     density    taxpc west central urban pctmin80     wcon    wtuc     wtrd
## 1 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565 167.3726
##       wfir     wser   wmfg   wfed   wsta   wloc        mix    pctymle
## 1 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944 0.07008217
```

Interestingly, the tax revenue per capita is lower than what we would expect for such a supposedly affluent county. Further investigation is required here to better understand the exact job market of this population. It is likely that there are only 1 or 2 members of this county who have high paying jobs, driving up wser. In this dataset, we would hope to see a percentage breakdown of workers in each sector. This would allow us to weight each wage parameter accordingly and provide context as to how much of an influence we would expect a sector's wage to have on its crime rate.

Across all the plots, we see that points are clustered along the lower end of the x axis. Therefore, we take the log of the wage parameters.

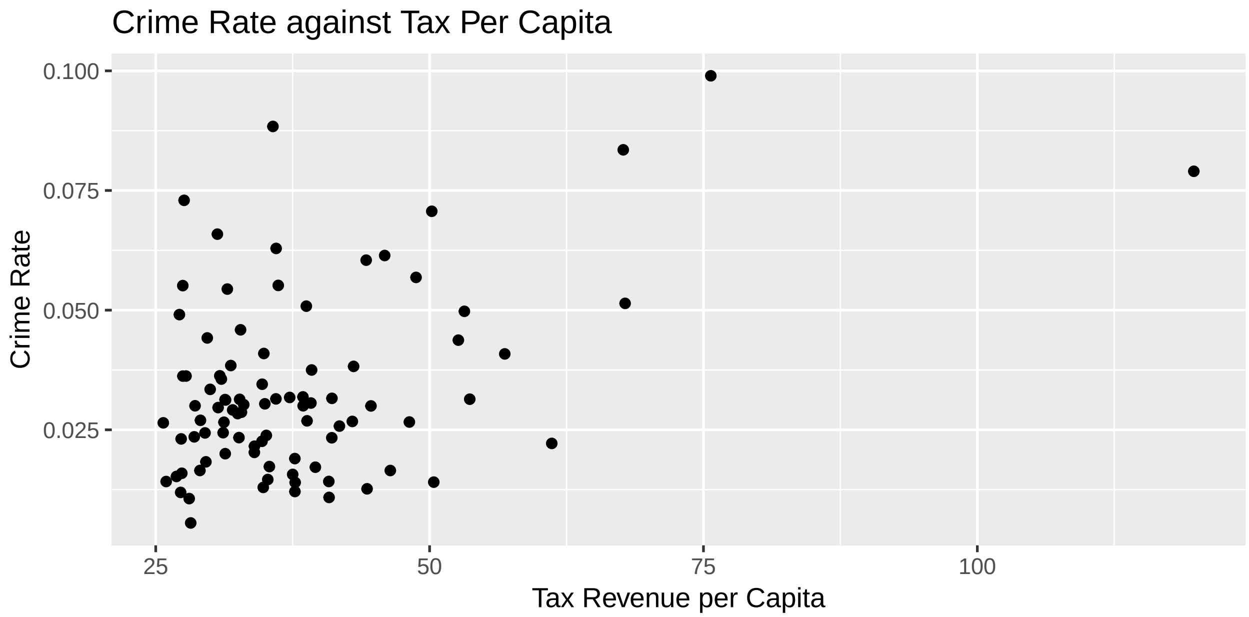
```r
ggplot(crime_wage, aes(x=log(wkly_wage), y=crmrte)) +
  facet_wrap(~sector) +
  geom_point() +
  ggtitle("Crime rate against log(wages) across each sector") +
  xlab("Log(Weekly Wage)") +
  ylab("Crime Rate")
```



This distributes the data much more effectively. Note that even though the wage parameters may not make it into the first specification, in the later specifications we will eventually be factoring these variables in. At that time, we will be taking the log of the wages as shown above.

As mentioned above, we note that taxpc is related to the wages of workers in each county, as higher taxes are applied to individuals with a higher income. We notice this in the correlation matrix at the top of this report, where taxpc shows at least a light blue relationship with each of the wage variables individually (this is expected as we anticipate that taxpc is more tightly correlated with a linear combination of these variables). From the correlation matrix, we expect to see the same positive relationship between taxpc and crmrte. Taking a closer look at this relationship, we can verify that this relationship holds.

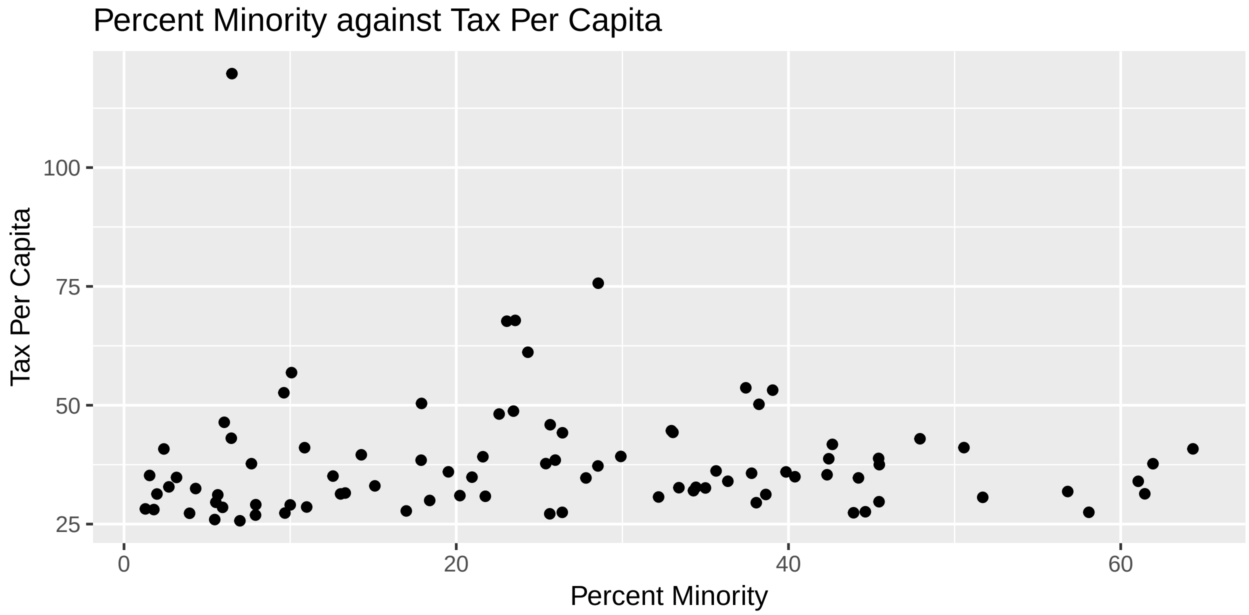
```r
ggplot(crime, aes(x=taxpc, y =crmrte)) +
  geom_point() +
  ggtitle("Crime Rate against Tax Per Capita") +
  xlab("Tax Revenue per Capita") +
  ylab("Crime Rate")
```



Investigating demographic variables

We notice a variable pctmin80, which is the percentage of minority groups in the population. We predict that neighborhoods with higher percent minorities had lower tax revenue per capita, as socio-economic barriers often forced minority groups to take lower paying roles, and racism factors often implied that minority groups would be paid less for the same jobs as white coworkers.

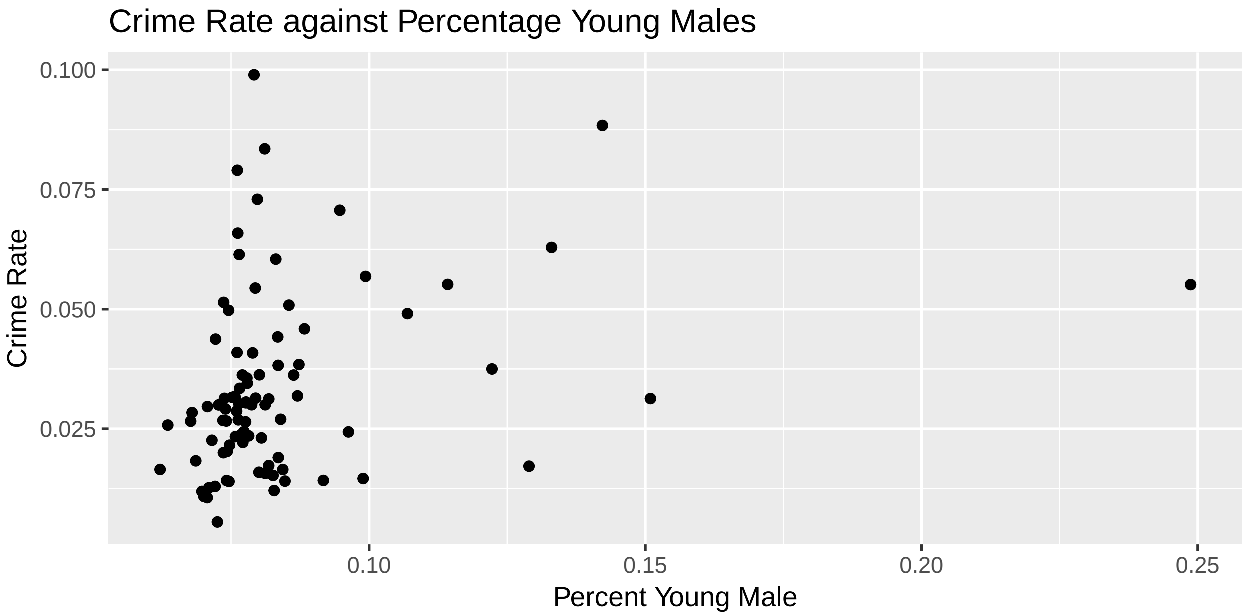
```r
ggplot(crime, aes(x=pctmin80, y =taxpc)) +
  geom_point() +
  ggtitle("Percent Minority against Tax Per Capita") +
  xlab("Percent Minority") +
  ylab("Tax Per Capita")
```



Contrary to the above discussion, Tax Revenue per Capita does not seem to be related to the percentage of minorities in a population.

Now looking at percent young male. This variable is of interest as the perpetrators of crime are often thought to come from this demographic group.

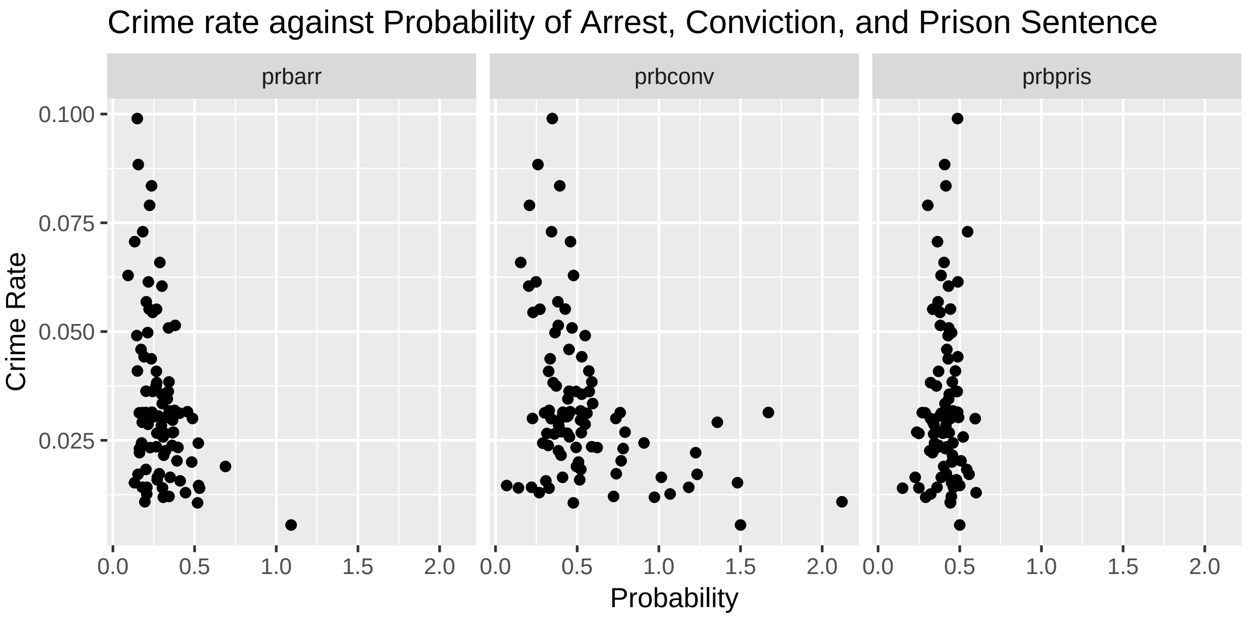
```
ggplot(crime, aes(x = pctymle, y = crmrte)) +
  geom_point() +
  ggtitle("Crime Rate against Percentage Young Males") +
  xlab("Percent Young Male") +
  ylab("Crime Rate")
```



The influence of fear - probability of punishment

Now looking at the probabilities associated with arrest, conviction and prison sentence. These 3 probabilites all illustrate the likelihood of being punished for a crime. Therefore, we expect that only one of these parameters is necessary to include in our model to avoid any confounding effects.

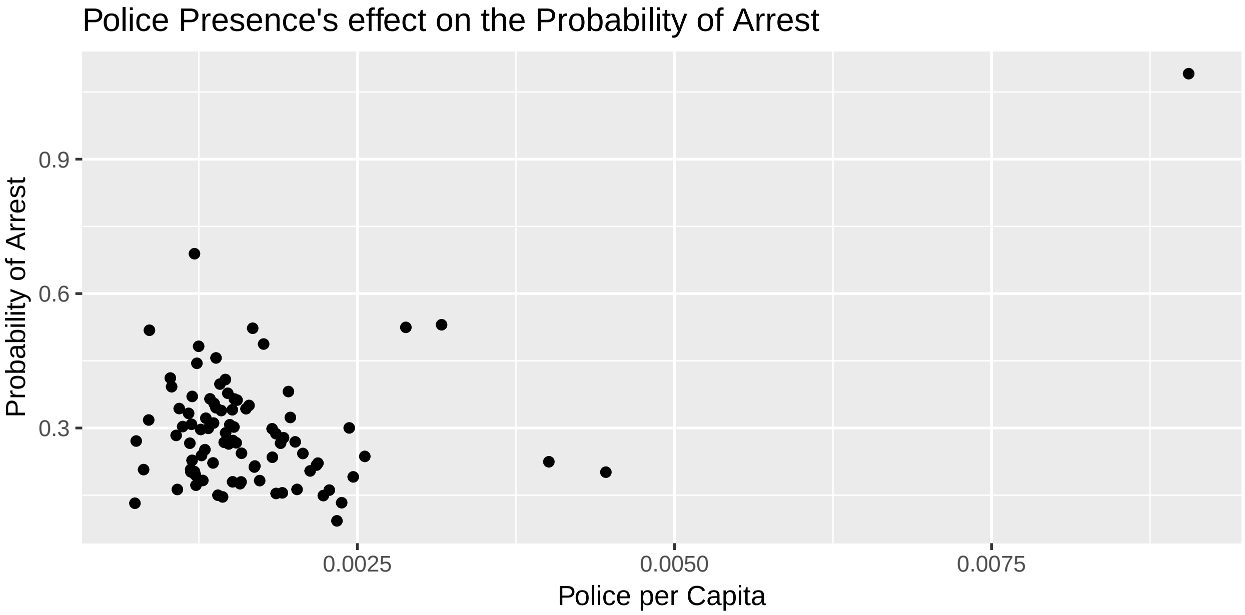
```
crime_prob_punishment <- crime %>%
  select(crmrte, prbarr, prbconv, prbpris) %>%
  gather(punishment, probability, -crmrte)
ggplot(crime_prob_punishment, aes(x=probability, y=crmrte)) +
  facet_wrap(~punishment) +
  geom_point() +
  ggtitle("Crime rate against Probability of Arrest, Conviction, and Prison Sentence") +
  xlab("Probability") +
  ylab("Crime Rate")
```



As suggested by the correlation matrix and the analysis above, there is a negative relationship between the probability of arrest and conviction with crime rate.

From intuition, police presence can either be positively related with crime (more police are needed in more crime active areas) or they can be negatively related (higher plice presence serves as a deterrant of crime). In the former case, police presence is an outcome variable of crime, and in the latter case, crime is the outcome variable.

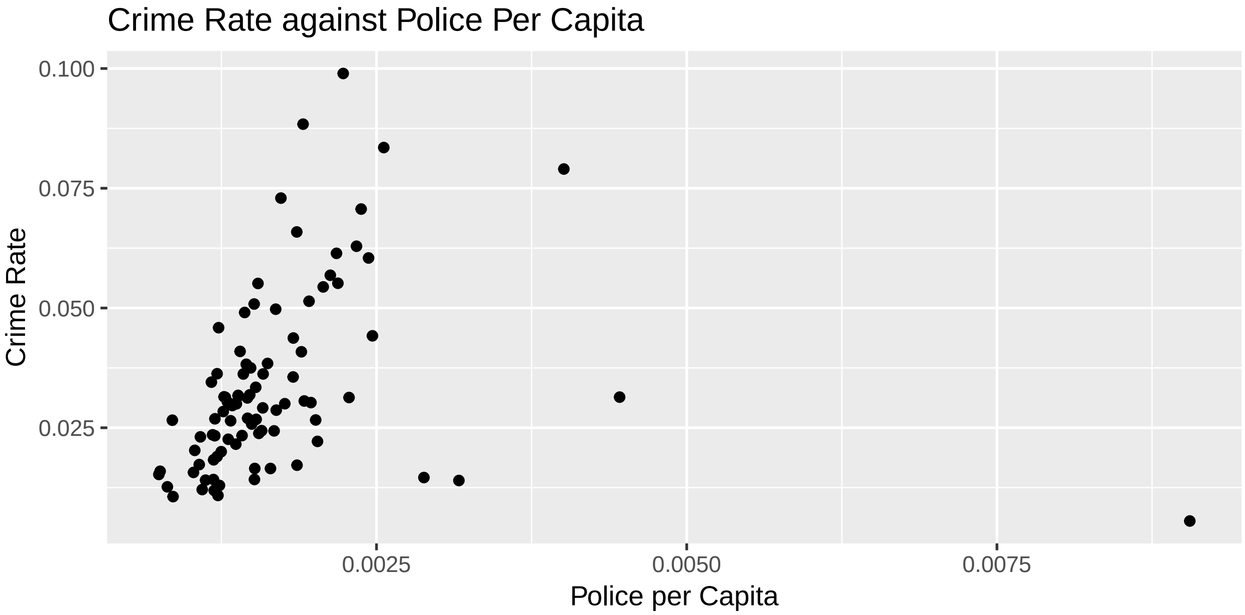
```r
ggplot(crime, aes(x = polpc, y = prbarr)) +
  geom_point() +
  ggtitle("Police Presence's effect on the Probability of Arrest") +
  xlab("Police per Capita") +
  ylab("Probability of Arrest")
```



There is one outlier where a very high police per capita leads to a high probability of arrest. This likely could be a neighborhood where crime is espcially high, so police are typically stationed here. This is not grounds to remove the outlier from our analysis, though we do note that, aside from this point, there doesn't seem to be a relationship between the police per capita and probability of arrest.

We question if more police officers could be a deterrent of crime from occuring in the first place. Thus, let us plot police per capita against the crime rate.

```
ggplot(crime, aes(x = polpc, y = crmrte)) +
  geom_point() +
  ggtitle("Crime Rate against Police Per Capita") +
  xlab("Police per Capita") +
  ylab("Crime Rate")
```



From above, the more police there are the more crime there is, with the exception of the outlier at the bottom right, reflecting an area with high police per capita and low crime. This is the same point as the outlier of the previous graph. Therefore, this one county has a High number of Police with a high likelihood of arreset, and a low crime rate. This likely is an area that is actively cracking down on crime. Overall though, we cannot make any conclusions on whether police presence reduces crime, or whether crime increases police presence. Therefore, this variable will likely not be included as part of our first model specification.

```
crime_outlier2 <- crime %>% filter(polpc>0.0075)
head(crime_outlier2)
```

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

Bringing our analysis together

```
model1 <- lm(crmrte ~ taxpc + density + pctymle + prbarr, data = crime)
model2 <- lm(crmrte ~ taxpc + density + pctymle + prbarr + prbconv + polpc + pctmin80, data = crime)
```

```
model3 <- lm(crmrte ~ taxpc + density + pctymle + prbarr + prbconv +
               prbpris + avgsen + polpc + pctmin80 + log(wcon) +
               log(wtuc) + log(wtrd) + log(wfir) + log(wser) + log(wmfg) +
               log(wfed) + log(wsta) + log(wloc) + mix, data = crime)
```

The Regression Table

We generate a regression table displaying the 3 models side by side

```
stargazer(model1,model2,model3
          , type ="latex"
          , column.labels   = c("Specification 1", "Specification 2 ", "Specification 3", "Specificatio
          , report = "vc*p", title = "Model Summaries"
          , keep.stat = c("rsq","adj.rsq")
          ,font.size = "small"
          ,single.row = TRUE
          , omit.table.layout = "n",
          add.lines=list(c("AIC", round(AIC(model1),1), round(AIC(model2),1), round(AIC(model3),1)),
          c("BIC", round(BIC(model1),1), round(BIC(model2),1), round(BIC(model3),1))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 30, 2018 - 20:39:10

```
# Below is a version of the stargazer table that has heteroskedastic-robust standard errors. This follo
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))

stargazer(model1,model2,model3
          , type = "latex"
          , column.labels   = c("Specification 1", "Specification 2 ", "Specification 3", "Specificatio
          , title = "Model Summaries"
          , omit.stat="f"
          , keep.stat = c("rsq","adj.rsq")
          , se = list(se.model1,se.model2,se.model3)
          , star.cutoffs = c(0.05,0.01,0.001)
          , font.size = "small"
          , single.row = TRUE
          , omit.table.layout = "n"
          , add.lines=list(c("AIC", round(AIC(model1),1), round(AIC(model2),1), round(AIC(model3),1)),
          c("BIC", round(BIC(model1),1), round(BIC(model2),1), round(BIC(model3),1))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 30, 2018 - 20:39:10

It is clear from the results that the second model has the best balance between parsimony and explaining the variation in the outcome variable. However it is interesting to note that both the AIC and BIC point are lower for the third specification that contains nearly all the variables than for the second specification that seems to be more parsimonious and hence could have been expected to be a better model. This also applies for the adjusted R squared value.

Table 1: Model Summaries

	Dependent variable:		
		crmrte	
	Specification 1	Specification 2	Specification 3
	(1)	(2)	(3)
taxpc	0.0004^{***}	0.0002^{**}	0.0002^{**}
	p = 0.0002	p = 0.020	p = 0.023
density	0.007^{***}	0.005^{***}	0.005^{***}
	p = 0.000	p = 0.000	p = 0.00000
pctymle	0.179^{***}	0.088^{**}	0.122^{***}
	p = 0.002	p = 0.034	p = 0.008
prbarr	-0.020^{**}	-0.056^{***}	-0.051^{***}
	p = 0.035	p = 0.00000	p = 0.00001
prbconv		-0.019^{***}	-0.017^{***}
		p = 0.000	p = 0.00001
prbpris			0.0001
			p = 0.996
avgsen			−0.0004
			p = 0.312
polpc		6.539^{***}	6.820^{***}
		p = 0.00001	p = 0.00003
pctmin80		0.0004^{***}	0.0004^{***}
		p = 0.000	p = 0.000
log(wcon)			0.005
			p = 0.546
log(wtuc)			0.004
			p = 0.492
log(wtrd)			0.007
			p = 0.500
log(wfir)			−0.009
			p = 0.261
log(wser)			-0.006^{*}
			p = 0.096
log(wmfg)			−0.001
			p = 0.783
log(wfed)			0.011
			p = 0.298
log(wsta)			−0.008
			p = 0.378
log(wloc)			0.004
			p = 0.789
mix			−0.019
			p = 0.189
Constant	−0.0005	0.018^{***}	−0.017
	p = 0.947	p = 0.009	p = 0.860
AIC	-544.9	-598.1	-590.3
BIC	-529.9	-575.6	-537.8
R^2	0.659	0.823	0.852
Adjusted R^2	0.643	0.808	0.812

Table 2: Model Summaries

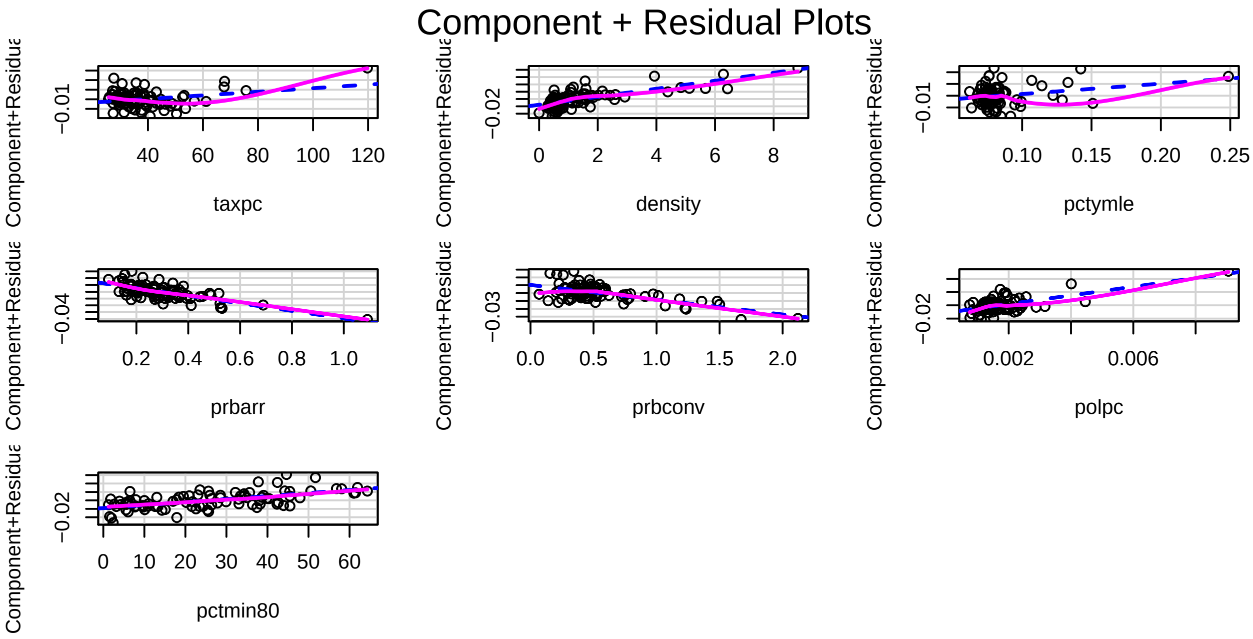
	Dependent variable:		
		crmrte	
	Specification 1	Specification 2	Specification 3
	(1)	(2)	(3)
taxpc	0.0004 (0.0003)	0.0002 (0.0002)	0.0002 (0.0003)
density	0.007*** (0.001)	0.005*** (0.001)	0.005** (0.002)
pctymle	0.179** (0.069)	0.088* (0.041)	0.122** (0.043)
prbarr	−0.020** (0.007)	−0.056*** (0.013)	−0.051*** (0.014)
prbconv		−0.019*** (0.004)	−0.017** (0.006)
prbpris			0.0001 (0.013)
avgsen			−0.0004 (0.0005)
polpc		6.539** (2.152)	6.820* (2.719)
pctmin80		0.0004*** (0.0001)	0.0004*** (0.0001)
log(wcon)			0.005 (0.009)
log(wtuc)			0.004 (0.008)
log(wtrd)			0.007 (0.016)
log(wfir)			−0.009 (0.012)
log(wser)			−0.006 (0.018)
log(wmfg)			−0.001 (0.007)
log(wfed)			0.011 (0.014)
log(wsta)			−0.008 (0.012)
log(wloc)			0.004 (0.025)
mix			−0.019 (0.022)
Constant	−0.0005 (0.012)	0.018* (0.009)	−0.017 (0.149)
AIC	-544.9	-598.1	-590.3
BIC	-529.9	-575.6	-537.8
R^2	0.659	0.823	0.852
Adjusted R^2	0.643	0.808	0.812

Add discussion of practical vs statistical significance

Checking for CLM assumptions

(1) Linear model assumption As we are not restricting the error term, we don't have to worry about the linear model assumption. Another way to look at this is by using the crPlots to check for non-linearlity.

```
crPlots(model2)
```



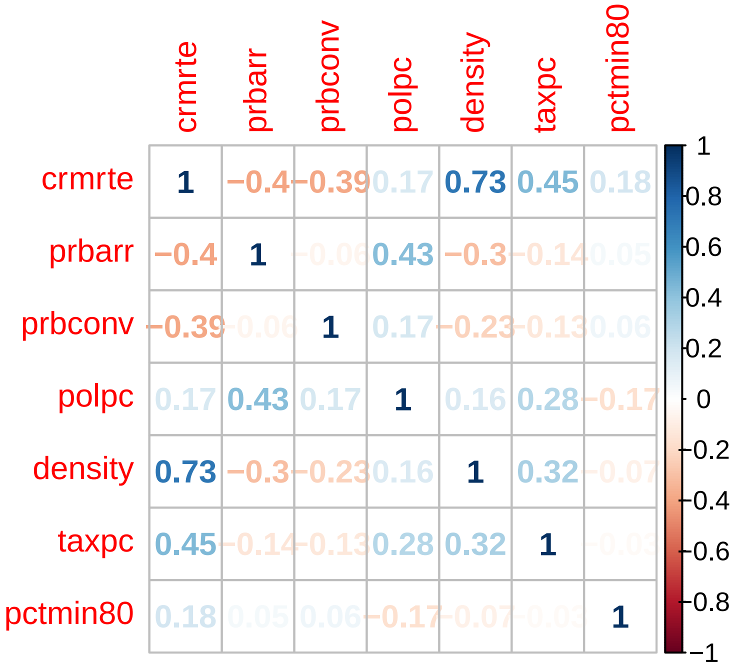
We see that it is linear for most of the independent variables except taxpc.

(2) Random Sampling We have 91 of the 100 counties in Carolina. As of 2016, we have 80 rural counties, we expect that number to be higher in 1980s, so we can assume that we have a random sample.

(3) Multicolinearlity

We know that there isn't perfect multi colinearlity as R would through an error saying it has encountered singularity. We can look at the correlation matrix to understand if there is a multi colinearity.

```
corr_mod2 <- cor(crime[,c(3,4,5,8,9,10,14)])
corrplot(corr_mod2, method =c("number"))
```



From the above correlation matrix we gather that none of the variables have a correlation value of more than 0.43. We can assume that we have no multicolinearlity.

Another way to look at this, is by looking at the Variance Inflation factors.

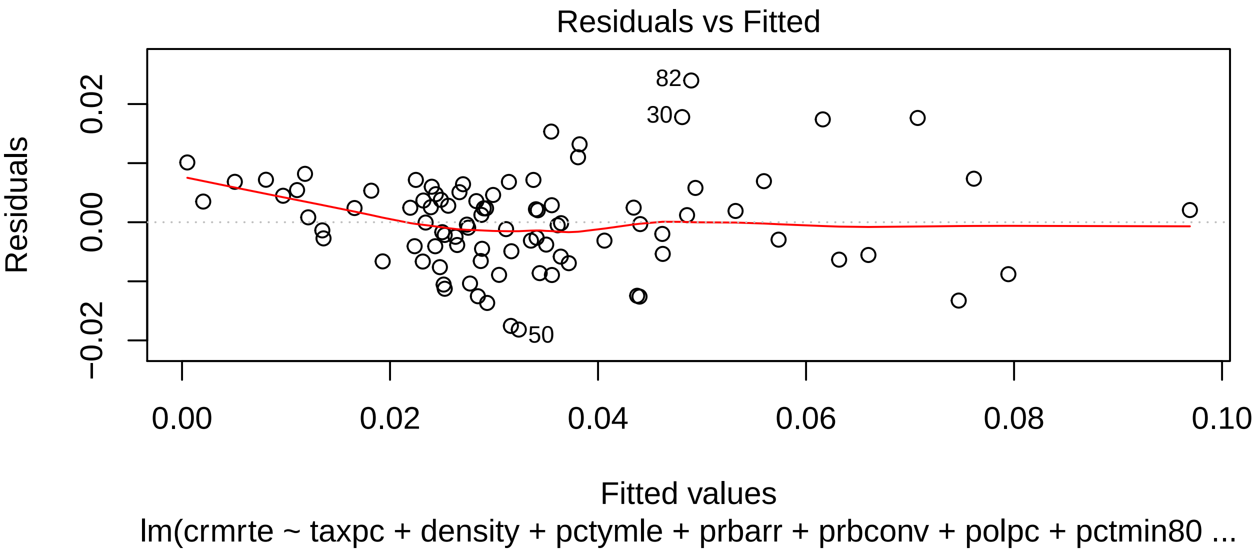
```
vif(model2)
```

```
##    taxpc  density  pctymle   prbarr  prbconv    polpc pctmin80
## 1.381327 1.444747 1.190653 1.969074 1.382107 2.045186 1.081045
```

We notice that none of the values are over 4, so we are safe to say that there is no multicolinearity.

(4) Zero Conditional Mean Lets start looking at the diagnostic plots to talk about the rest of the assumptions.

```
plot(model2, which = 1)
```

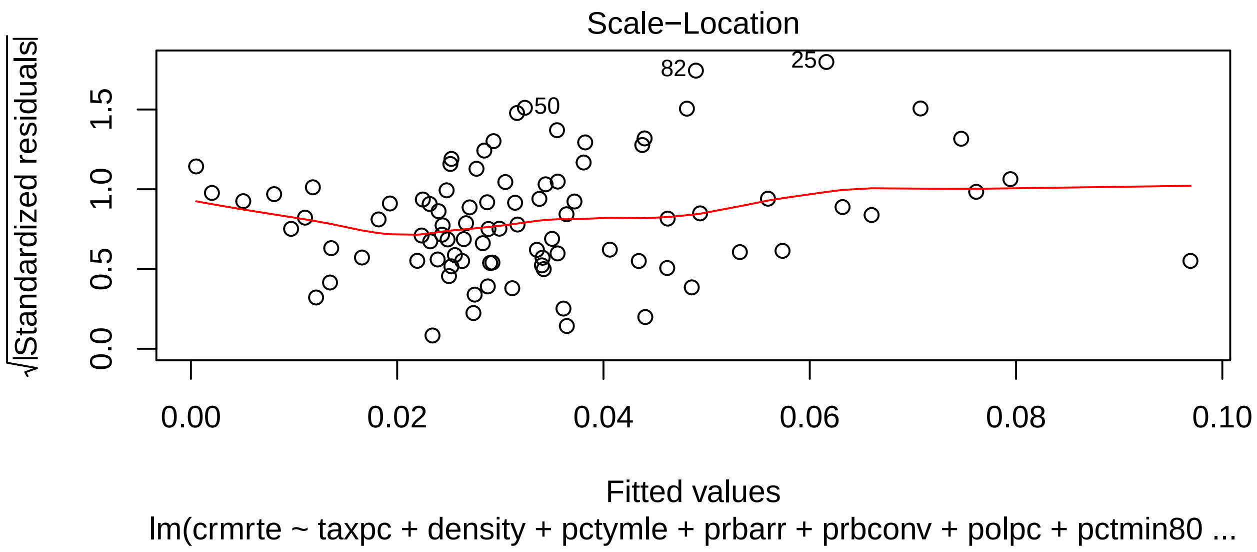


From the residual vs the fitted plot, we see that we have violated the assumptions of zero conditional mean, as values on the left hand side of the plot appear to be higher than those on the right. This means that the coefficients are biased. As we will be discussing in the omitted variables section, there are quite a few variables but none of them seem to be highly correlated to the independent variables that we are using in this model. So, we can assume exogenity.

This will enable use to provide causal inferences from the analysis.

(5) Homoskedasticity

```
plot(model2, which = 3)
```



Here we examine the spread of the residuals from the residuals vs fitted plot as well as the straightness of the mean values of the scale-location plot. From the residuals vs fitted plot, the variance of residuals seems to be fairly even across the fitted values. Looking at the scale-location plot, there is a slight dip in the plot in the center left region of the graph, however this could be attributed to the higher density of points at this region. In either case, we should consider using at the standard errors robust to heteroskedasticity as we cannot make a confident assumption of homoskedasticity

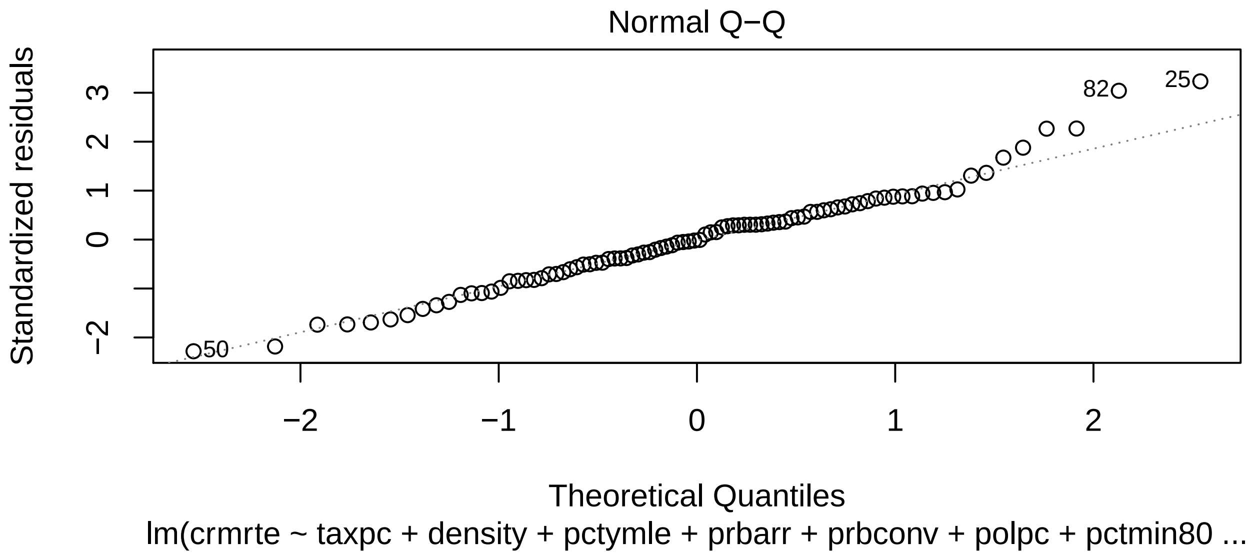
```
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.7604e-02  8.7368e-03  2.0149 0.0471871 *
## taxpc        1.8688e-04  2.4456e-04  0.7641 0.4469804
## density      5.4684e-03  1.3856e-03  3.9465 0.0001666 ***
## pctymle      8.8281e-02  4.0983e-02  2.1541 0.0341696 *
## prbarr      -5.5680e-02  1.2609e-02 -4.4161 3.043e-05 ***
## prbconv     -1.8846e-02  3.9457e-03 -4.7764 7.711e-06 ***
## polpc        6.5391e+00  2.1524e+00  3.0380 0.0031938 **
## pctmin80     3.5614e-04  6.4558e-05  5.5166 3.928e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

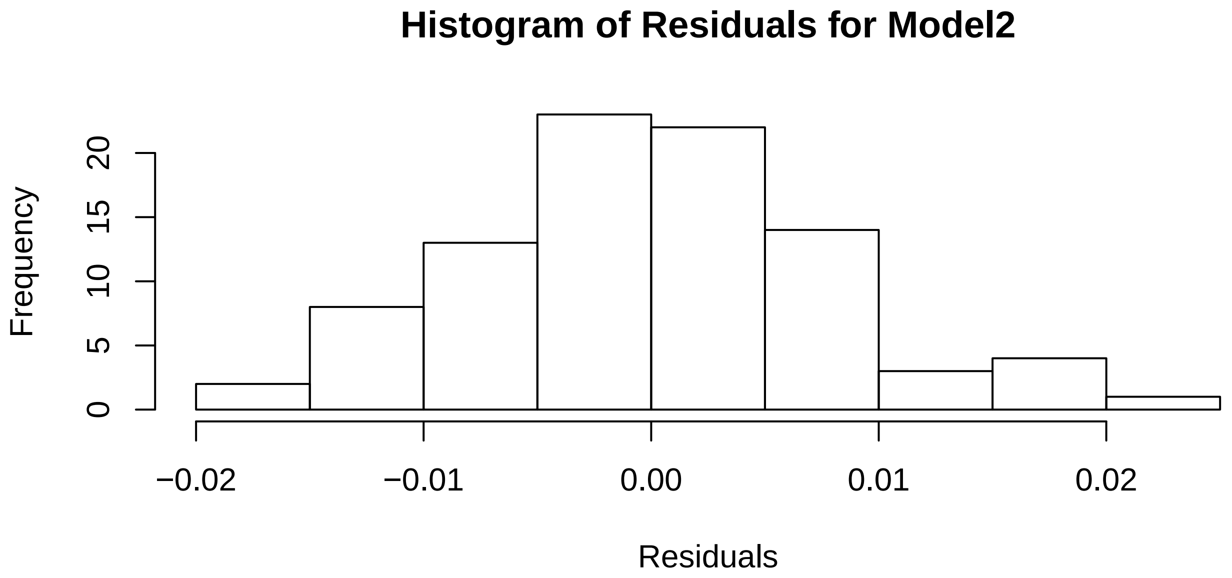
From the above, density, the probability of punishment, and percent minority seem to be the strongest indicators for crime rate. This is followed by the police per capita (which either may be a result of crime rate or a deterrent of crime, so we cannot confidently use it as a predictor variable), and the percentage of young males.

(6) Normality of Errors

```r
plot(model2, which = 2)
```



```r
hist(model2$residuals, main = "Histogram of Residuals for Model2", xlab = "Residuals")
```



When we look at the Q-Q plot, we notice that except for a few points most of the data lies close to the line which ensures that we have a normal error distribution. Additionally, by plotting the residuals above, we see that this distribution looks fairly normal.

The scale-location plot shows the heteroskedastic as we see that there are outliers within the data set. Another way for checking for heteroskedasticity is by looking at the variance. That can be checked as shown below

```r
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 6.888983, Df = 1, p = 0.0086729
```

The above results indeed supports our inference that it is heteroskedastic as the p value is less than 0.05

Also, we see observation 25 falls outside the Cook's distance line of 1. So, this could be an influencial outliers that we should be looking at. The main challenge with this observation is that tax per capita is way more than what we see with all the other observations.

Another way to look at the outliers

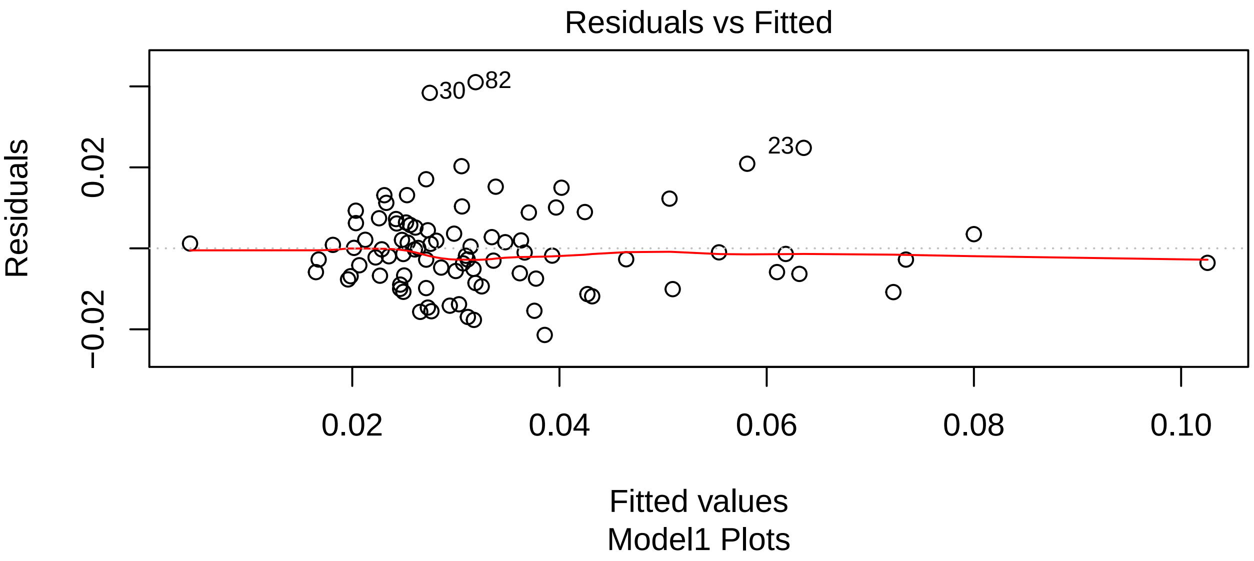
```
outlierTest(model2)
```

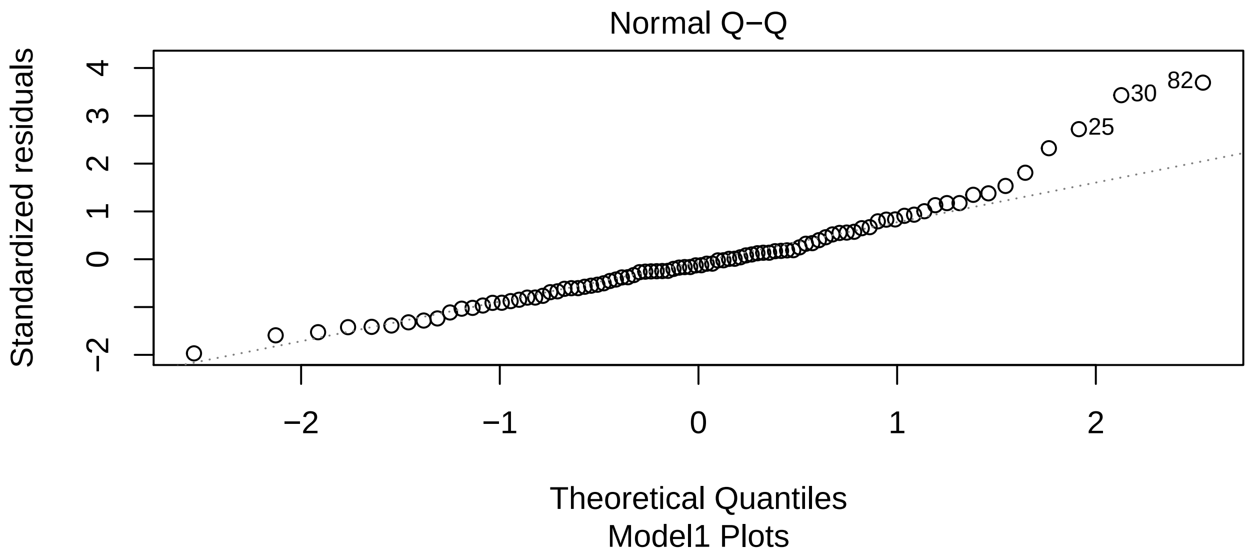
```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 25 3.439746         0.00092279     0.083051
```

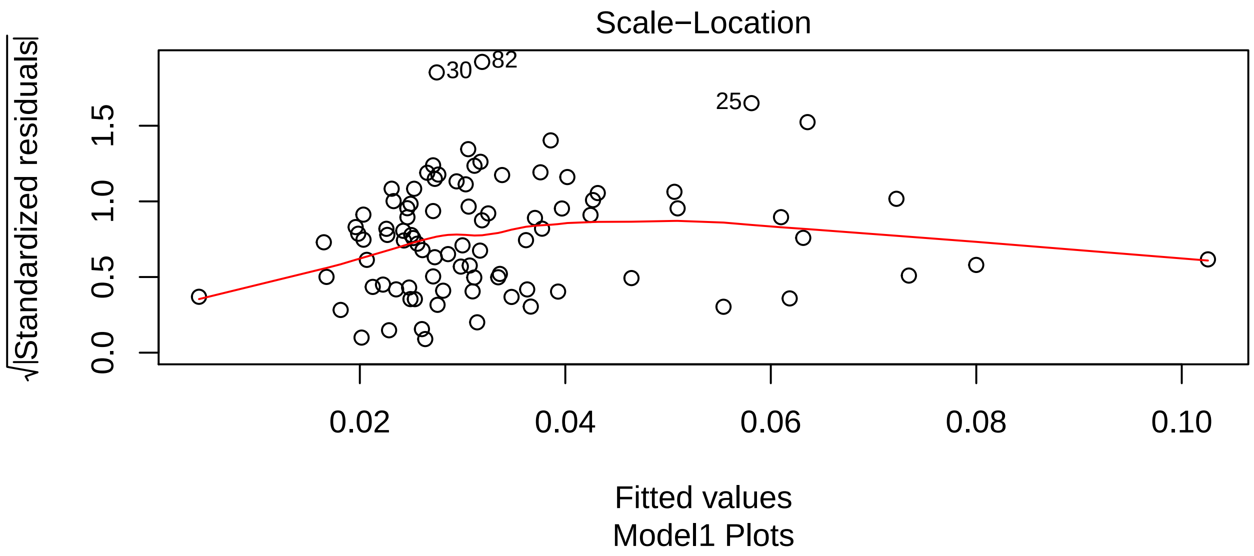
Similar to what we saw in the plot, we see that observation 25 is the most extreme value, but the P value is pretty colse to 0.05, but the earlier plot shows this is an outlier..

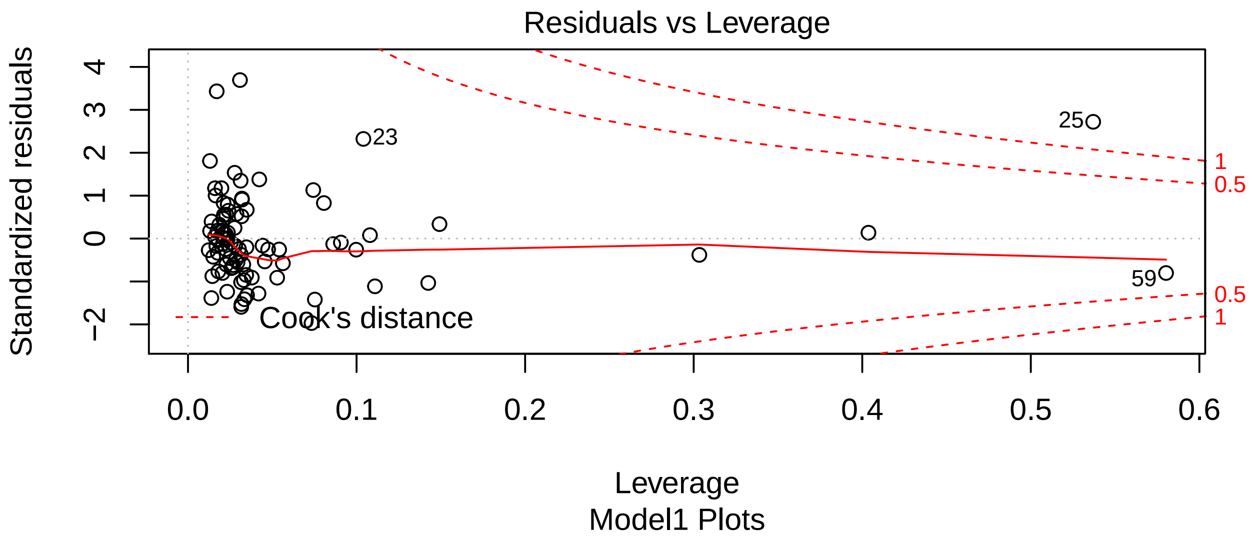
We can look at the diagnostic plots for other models as well:

```
plot(model1, sub.caption  = "Model1 Plots", cex.caption = 1)
```



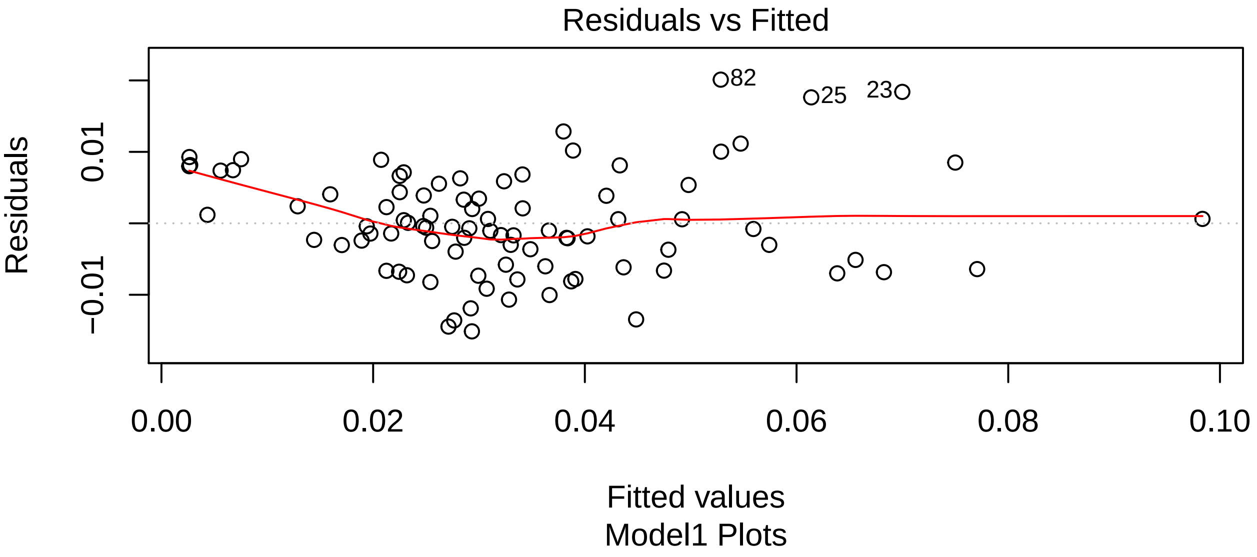

Normal Q−Q
Standardized residuals
−2
0
1
2
3
4
−2
−1
0
1
2
30
82
25
Theoretical Quantiles
Model1 Plots

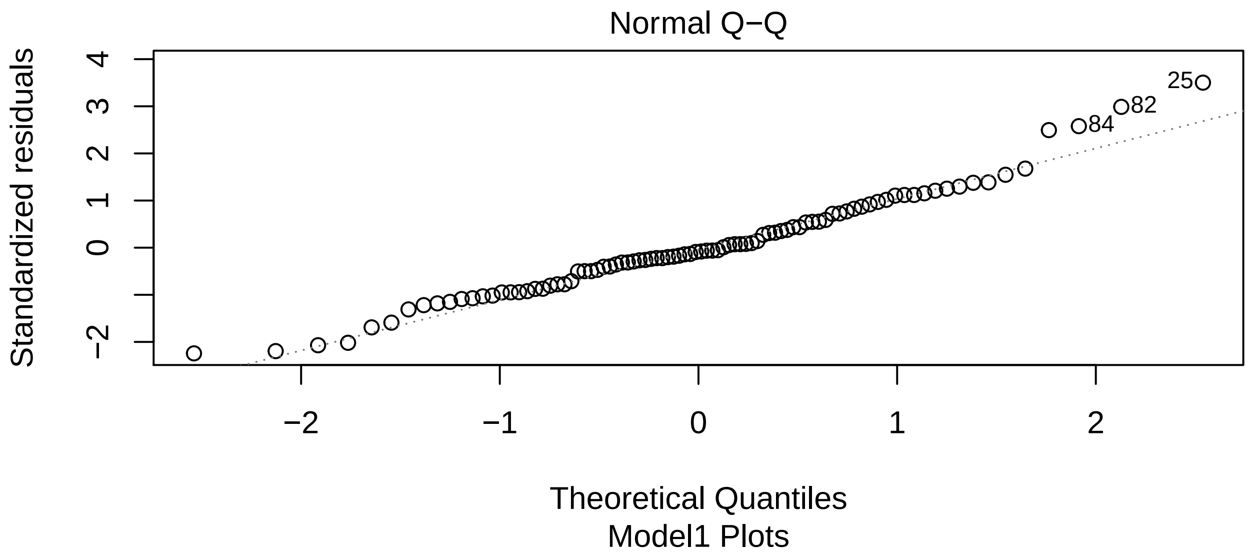

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.02
0.04
0.06
0.08
0.10
30
82
25
Fitted values
Model1 Plots

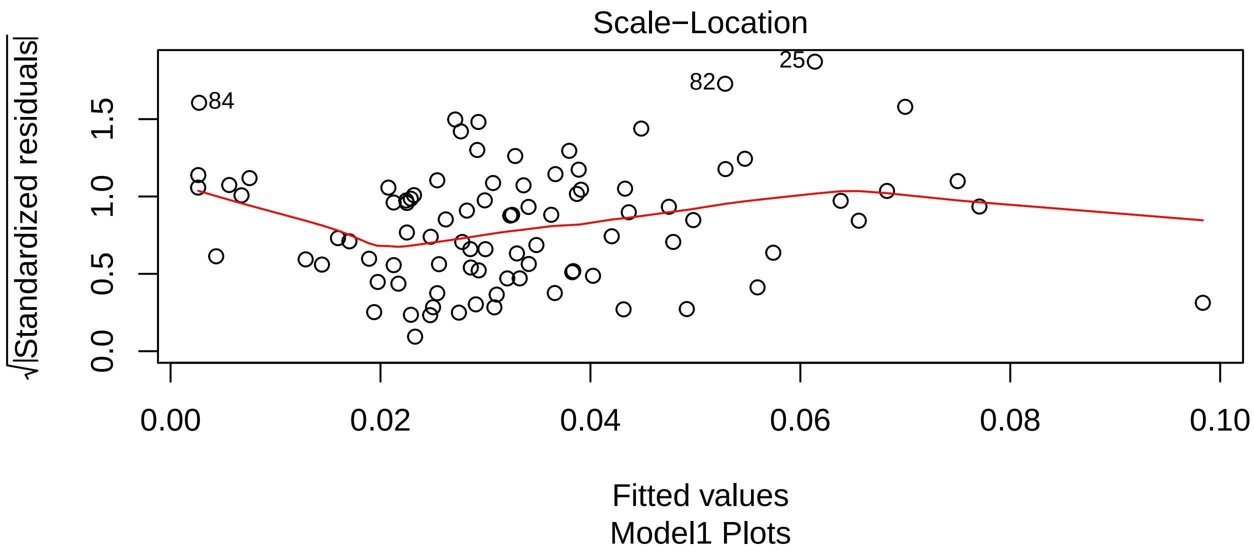


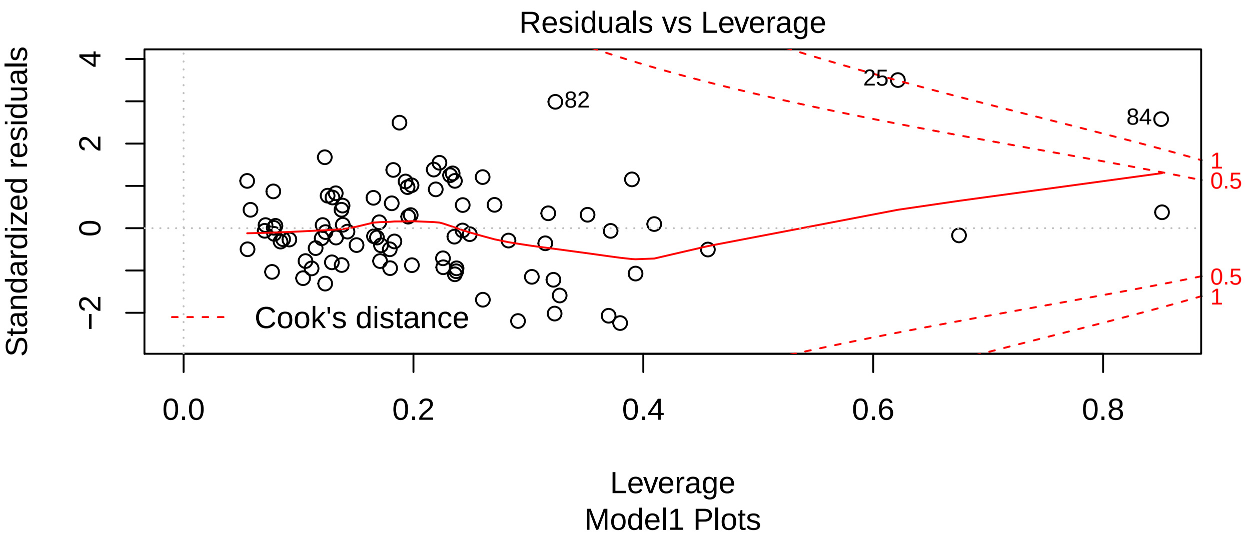
Interestingly, the residuals vs fitted plot for Model1 is the flattest, however it is more heteroskedastic. The residuals seem to have a similar degree of normality as Model specfication 2. Even if the Adjusted R squared is lower for this plot than for Model2, it may be worth considering that Model 1 is a better predictor once we use heteroskedastic-robust standard errors to assess the model.

```r
plot(model3, sub.caption  = "Model1 Plots", cex.caption = 1)
```




Normal Q−Q
Standardized residuals
4
3
2
1
0
−2
25
82
84
−2
−1
0
1
2
Theoretical Quantiles
Model1 Plots


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
84
82
25
0.00
0.02
0.04
0.06
0.08
0.10
Fitted values
Model1 Plots



Comment - Should we add this (needs gvlma package) NK: my only hesitation is that I believe this isn't covered in our course and the instructions tell us to not use anything outside of what we've learned in the course.

Another way to compare all the assumptions is through gvlma function.

```
model2_gvlma <- gvlma(model2)
summary(model2_gvlma)
```

```
## 
## Call:
## lm(formula = crmrte ~ taxpc + density + pctymle + prbarr + prbconv + 
##     polpc + pctmin80, data = crime)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -0.0181722 -0.0052538 -0.0001114  0.0047106  0.0239860 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.760e-02  6.545e-03   2.690  0.00866 ** 
## taxpc        1.869e-04  7.858e-05   2.378  0.01973 *  
## density      5.468e-03  6.925e-04   7.896 1.12e-11 ***
## pctymle      8.828e-02  4.079e-02   2.164  0.03337 *  
## prbarr      -5.568e-02  8.936e-03  -6.231 1.89e-08 ***
## prbconv     -1.885e-02  2.910e-03  -6.477 6.50e-09 ***
## polpc        6.539e+00  1.265e+00   5.168 1.63e-06 ***
## pctmin80     3.561e-04  5.367e-05   6.636 3.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.008271 on 82 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8082 
## F-statistic: 54.59 on 7 and 82 DF,  p-value: < 2.2e-16
## 
## 
```

```
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model2)
##
##                       Value p-value                   Decision
## Global Stat        5.97057 0.20136    Assumptions acceptable.
## Skewness           1.02297 0.31182    Assumptions acceptable.
## Kurtosis           0.59332 0.44114    Assumptions acceptable.
## Link Function      4.33337 0.03737 Assumptions NOT satisfied!
## Heteroscedasticity 0.02091 0.88503    Assumptions acceptable.
```

The Omitted Variables Discussion

There are several omittied variables that would be valuable in conducting this analysis:

1. Severity of crime. Crimes can vary from being petty (jaywalking or parking in a no parking zone) to severe crimes that do warrant arrest, conviction and prison sentences (kidnapping, thefts, sexual violence). Having a parameter that indicates the severity of the crime would help differentiate the varying levels of crime and focus analysis on reducing the likelihood of harsher crimes. The crime severity would be positively correlated with the crime rate and the probability of conviction but negatively correlated with the probability of arrest and the average sentence. This may lead to a negative coefficient because of the higher magnititude of the coefficient for the probability of arrest.

2. Income gap. There are several variables that point to the affluence of a region, but we are interested in seeing the percentage of upper/middle class individuals compared to percentage of lower class. We predict that the difference in these percents would be a better indicator of crime rate. Currently, we only have the wage within each sector (it is unclear whether this wage is a median or a mean or some other aggregated measure). There also could be omitted sectors, and we don't know the relative proportion of individuals in each sector. The size of the income gap may be positively correlated with the crime rate as well as the tax revenue and wage variables for high paying sectors like service while being negatively correlated with the wage variables for low paying sectors like manufacturing. As such the size of the income gap is likely to be have a positive coefficient.

3. Police bias. Bias among police officers in certain areas may contribute to the crime rate because of spurious arrests and convictions. This may be difficult to measure directly. We would expect police bias to be positively correlated with the probability of arrests and to a lesser extent the probability of convictions. It would also be positively correlated with the crime rate leading to the coefficient being positive.

4. Crime rate in neighbouring counties. Proximity to other areas where crime is high may have an influence on the crime rate in a particular county due to spillovers of activity. This variable may be correlated with other variables like the probability of convictions and probability of arrests as well as the outcome variable, crime rate. We would as a result expect the coefficient of police bias to be positive.

5. Size of the economy. The size of the economy for each county may be a factor. Explanations could be made for crime rate to be higher or lower in a given county depending on other counties. It would be intersting to see how the crime rate varies with the size of the economy (measured by GDP or similar measure). This would likely be positively correlated with the density and tax per capita variables as well as the wage variables and the crime rate variable. The sign of the coefficient for this variable would be expected to be positive

6. Unemployment rate- We have the wage level within each sector, but we don't have the unemployment

rate within each county. An unemployed person has a higher propensity to commit a crime than someone who is working. So a higher unemployment rate in a county would increase the crime rate in the county. We expect that this has a positive bias on the coefficients with a positive slope.

7. Family Composition. Having a variable which defines the degree of cohesiveness or divorces will play an important role in the crime rate. People with a less than healthy childhood has a higher chance of commiting crime than a person who had a normal childhood. The higher the family composition, the lower the crime rate which would imply that we will have a negative coefficient. We expect there to be no correlation between the family composition and the other variables currently in the model. Therefore, it is likely absorbed by the error terms.
8. Poverty Level - In our current model taxpc acts as a proxy for the poverty level, but the challenge with this is that people within or close to the poverty level do not contribute to taxes and there might be outliers with higher income that can skew the data substantially.

Comments on the Modeling Process

Between Specification 1 and 2, we observed an increase in the adjusted R squared value, implying that the robustness of the model did indeed increase with the addition of the prbconv, polpc, and pctmin80 variables. Looking between the second and third specification, all of the variables in the second specification more or less retained their coefficient value, suggesting that the variables in specification 2 are indeed the ones we should be focusing on, hwowever the AIC and BIC scores for the third specification were better than that for the second specification which is an indication that a reasonable amount of the variation is explained by the wage variables that were left out of the second model specification.

There are a number of omitted variables, discussed above, that are suspected to have an impact on the crime rate, or that we suspect are highly correlated with the variables available in the dataset.

Secondly there are confounding factors we must consider. We noticed that crime rate is going up due to police presence, which may seem counterintuitive. However, more police may lead to more reports of crime, so there are a large number of unreported crimes that are potentially being missed here. This leads us to question whether the crmrte variable is really a true representation of how safe a neighborhood is, as, in some areas most crimes can go unreported either due to a lack of trust in law enforcement, or simply because victims do not want to spend the energy to report a crime. We must also consider the possibility that there is a higher police presence in some variables because there is a high crime rate. This would actually suggest that polpc is an effect of crmrate than the other way round. If this is true, then we would actually advise removing the polpc variable from our model specifications.

Also, there are a variety of economic variables that could be highly correlated with another economic parameter that is causal to crmrate. Specifically, we would suggest exploring income gaps and unemployment data in counties, as these have a clearer causal mechanism to crime rate.

Implications for the Political Campaign

In terms of actionable steps for the political campaign, we recommend looking at the outputs of the regression model with a more critical eye. From first glance, it looks like reducing the number of police, lowering taxes, and forcing young males out of counties would be the solution. These are neither advised nor are they ethical in some cases. Instead, we recommend looking at why these explanatory variables are related to crime rate.

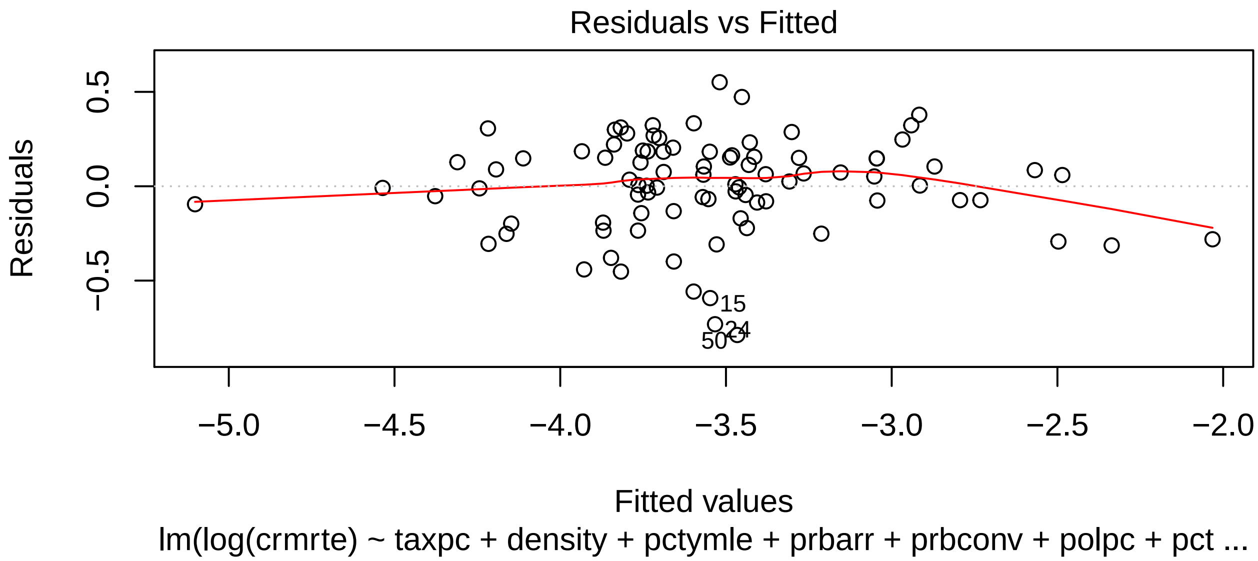
Some actionable steps we can recommend are creating programs to keep young men employed and off the streets, programs to improve the relationship between civilians and police forces, and making the penalties for crimes known as to deter crime from happening in the first place. Results from a future analysis that looks at income gaps explicitly would also inform which groups of individuals require a wage increase or better employement opportunities (if any).

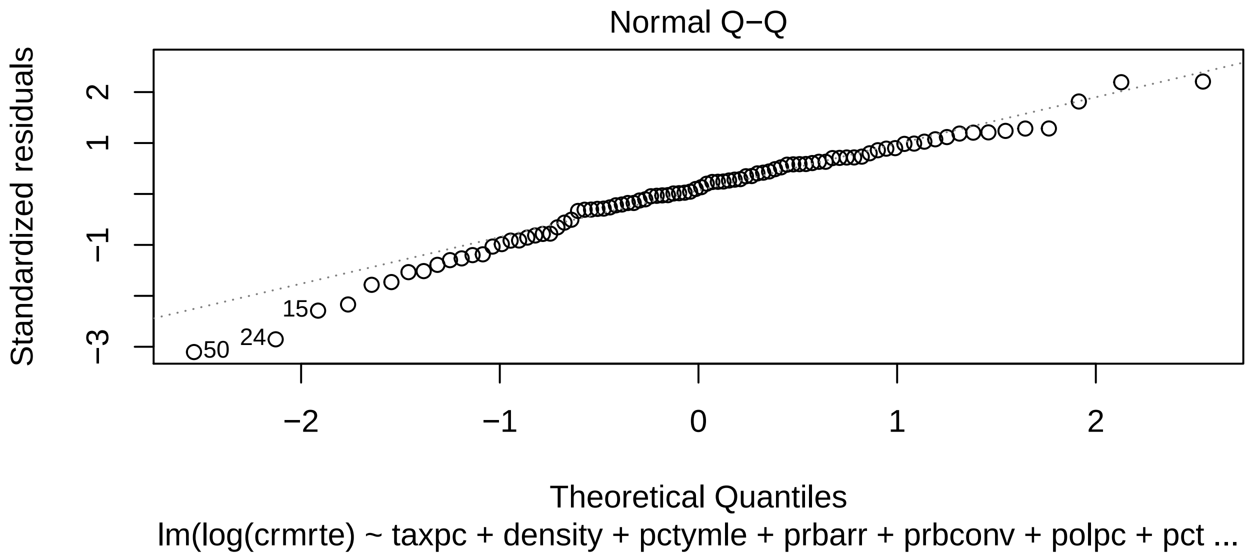
Appendix

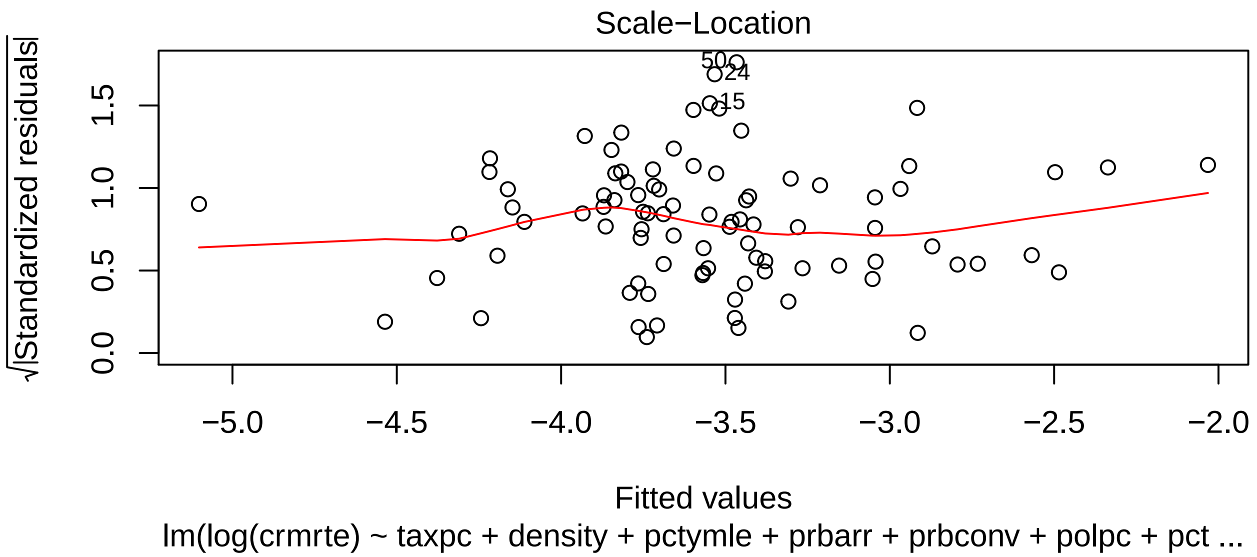
We found the CLM assumptions to reflect well on the models built so far. For all 3 specifications, the QQ plots show residuals are fairly close to normal. The thickness of the band for the Residuals vs Fitted plot for Specification 2 is fairly even, suggesting that there is even variance across all fitted values, suggesting homoskedasticity. In the Scale-Location plot there is a slight dip in the middle left portion of the graph, though this can be attributed to the larger number of points there.

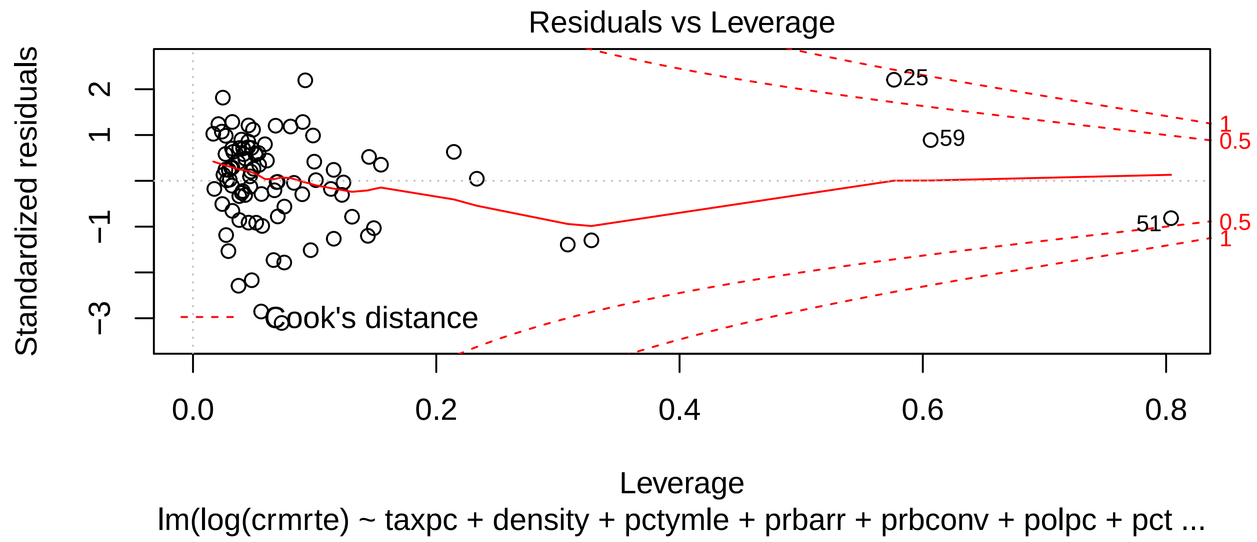
Still, we do see from the histogram of crmrte that this variable is skewed right. We try modifying Specification 2 by taking the log of crime rate and plot graphs to comment on how this holds up as a model for inference using the 6 CLM assumptions:

```r
model4 <- lm(log(crmrte) ~ taxpc + density + pctymle + prbarr + prbconv + polpc + pctmin80, data = crim
plot(model4)
```









The residuals vs fitted plot appears to peak in the center of the graph, so we violate the zero conditional mean condition (CLM 4). The normal QQ plot also deviates from normality more than model2, and the line in the scale-location graph is less flat, indicating a greater degree of heteroskedasticity. Overall, we can feel confident sticking to the untransformed crmrte in our preferred specifications.