Project 3

Authors: Sayan Das, James Gao, Jacob Tosh

Political Campaign Objective

> It is important to determine the causes of crime in order to make an impact on our society. As we delved into the logistics of our political campaign, and after conducting some high-level exploratory data analysis on the given dataset of crime statistics, we have decided to explore three key variables from the dataset in particular: **crmrte, polpc, taxpc**. These variables were decided for the following purposes - does an increase in police per capita result in a decrease of crime rate for a given county? Answering this question can lead to certain policy implementations that will ultimately help decrease crime rate. Moreover, how does tax revenue play a role with regards to the aforementioned variables? This question ultimately implores us to examine whether changes in tax revenue in various counties have significant impacts on the crime rate.
>
> Overall, the objective of analyzing these questions is to propose policy suggestions that would help deter the proliferation of crime in these counties, and our research group is confident that variables such as police and tax revenue per capita will have a significant impact and provide us insight on which policies to propose to combat and ultimately decrease our crime rates in society.

Measurements

> The key variables that we are interested in are:
>
> - crmrte
> - polpc
> - taxpc
>
> In addition to these three key variables, the following variables will help supplement our analysis, as mentioned later on in our exploratory data analysis (EDA):
>
> - prbarr
> - prbconv

EDA

> Prior to developing any models, it is important to get a high level understanding of each of our columns of interest from our dataset. To do this, let's take a look at a summary of each of the columns:

In [3]:

```
data <- read.csv('crime_v2.csv')
summary(data$crmrte)
summary(data$polpc)
summary(data$taxpc)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.005533 0.020927 0.029986 0.033400 0.039642 0.098966 
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0007459 0.0012308 0.0014853 0.0017022 0.0018768 0.0090543 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.69   30.66   34.87   38.06   40.95  119.76 
```

> Taking a quick look at the datasets above, everything seems to be reasonable when paying in attention to specifically the Min, Mean, and Max variables. For instance, a .098966 (approximately 10%) crime rate for the maximum in a county seems usual. Let's take a deeper dive into our supplementary variables of analysis (**prbarr**

maximum in a county seems usual. Let's take a deeper dive into our supplementary variables of analysis (**prbarr, prbconv**)

In [4]:

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.09277 0.20568 0.27095 0.29492 0.34438 1.09091 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.06838 0.34541 0.45283 0.55128 0.58886 2.12121 
```

> Now, we see statistics that seem to be nonsensical. For example, the maximum value in our prbarr column is 1.09091, which implies that 1.09 people are arrested for every 1 offense. However, this is simply not possible, for how can there be a greater number of arrests than offenses. Based on similar logic, the maximum valuable of 2.12121 for the prbconv variable does not also make sense; this statistic implies that 2.12 people are convicted for every arrest. Once again, how are more than two people being convicted for each arrest?
>
> Based on these arguments, our group concludes that the maximum value that can occur for both of these variables is 1.0. This is based on the assumption that if more than one person is convicted for an arrest, they are already included in the arrest statistic. This assumption would corroborate our argument that the maximum value can only be 1.0, for of course this implies that the number of convictions cannot exceed the number of arrests. Moreover, due to the double jeopardy clause that is specified in the 4th Amendment of the constitution, an individual cannot be convicted more than once for a single offense.

In [5]:

```
cleanedData = data[data$prbarr <= 1.0 & data$prbconv <= 1.0,]
summary(cleanedData$prbarr)
summary(cleanedData$prbconv)
length(cleanedData$prbarr)
length(data$prbarr)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.09277 0.22154 0.28733 0.29673 0.35035 0.68902 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.06838 0.33470 0.43896 0.44824 0.52760 0.97297 
```

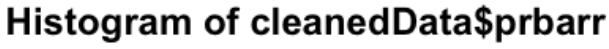
81

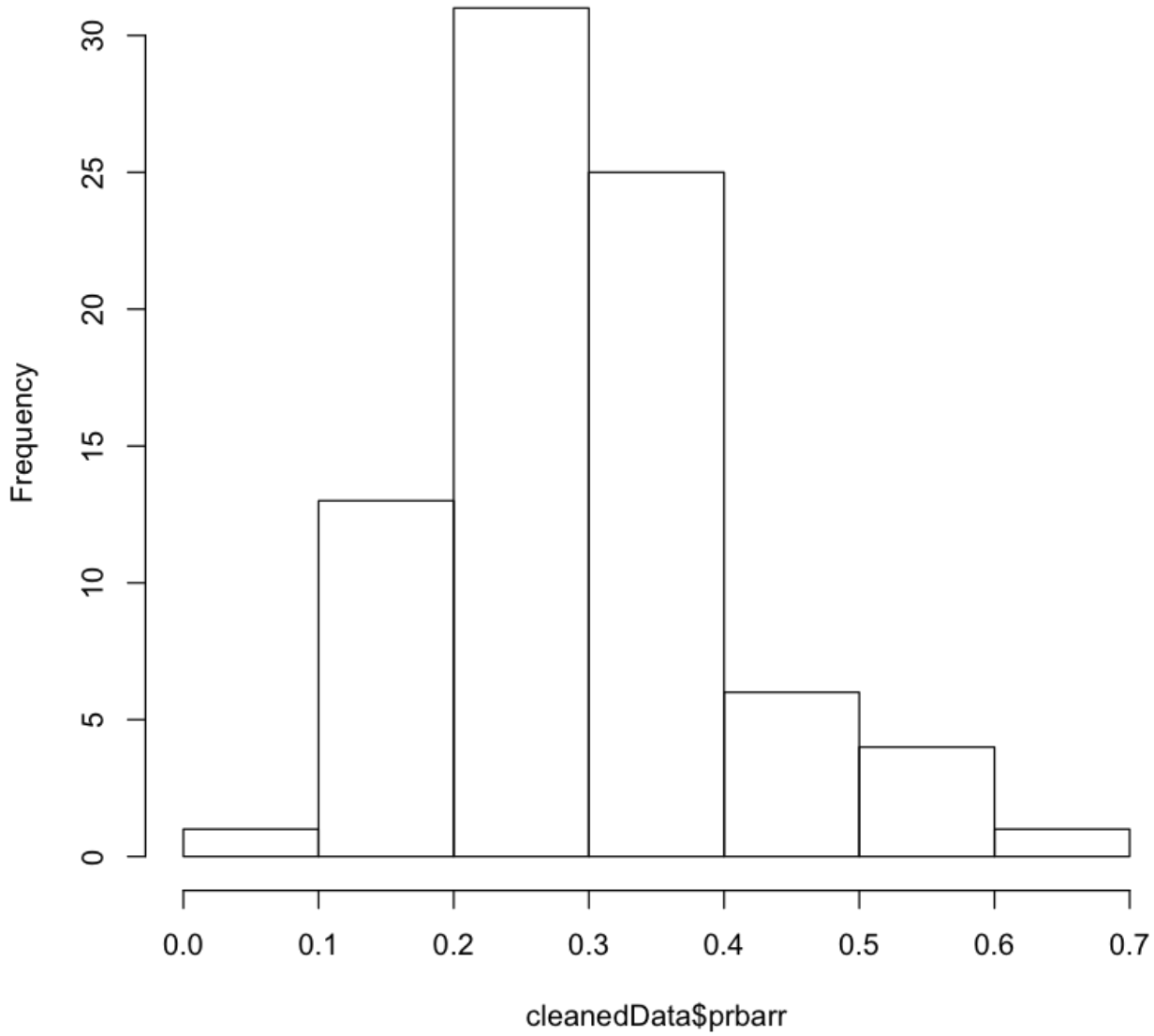
91

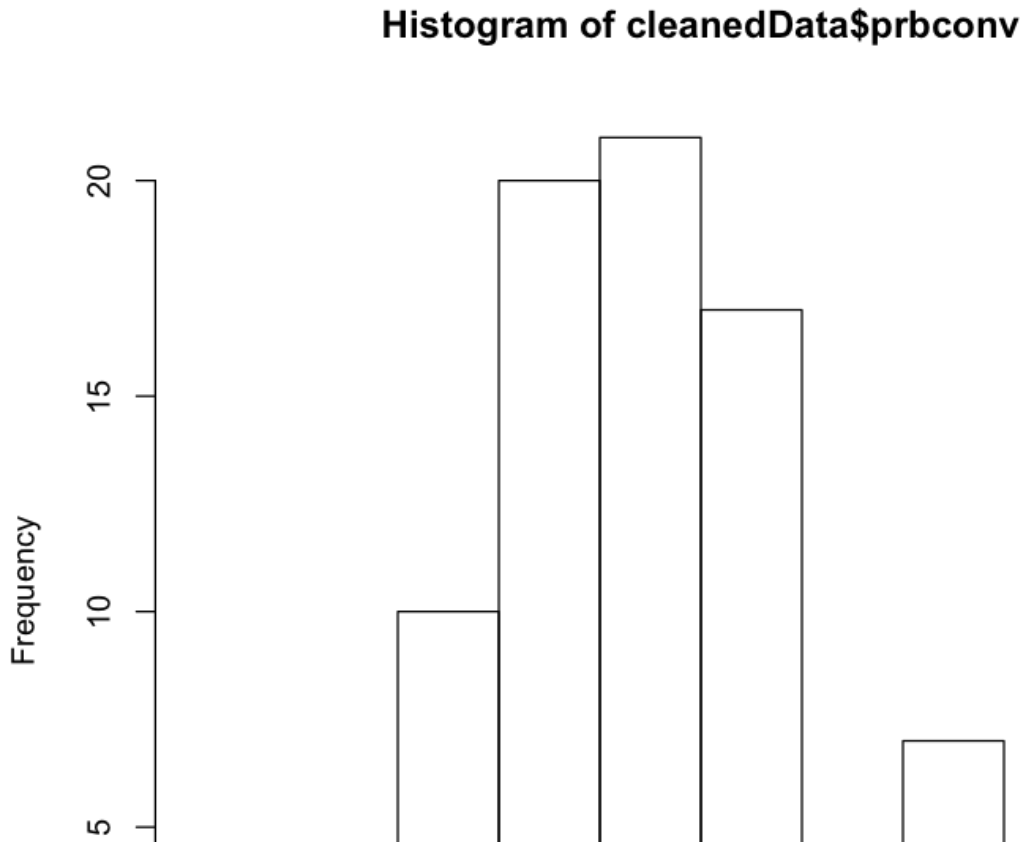
> After cleaning up our data, we see that 10 entries were removed from our original dataset. Now, we can be confident that all of the columns that we using for our EDA are of significant value.
>
> Next, let's take a quick look at the histograms

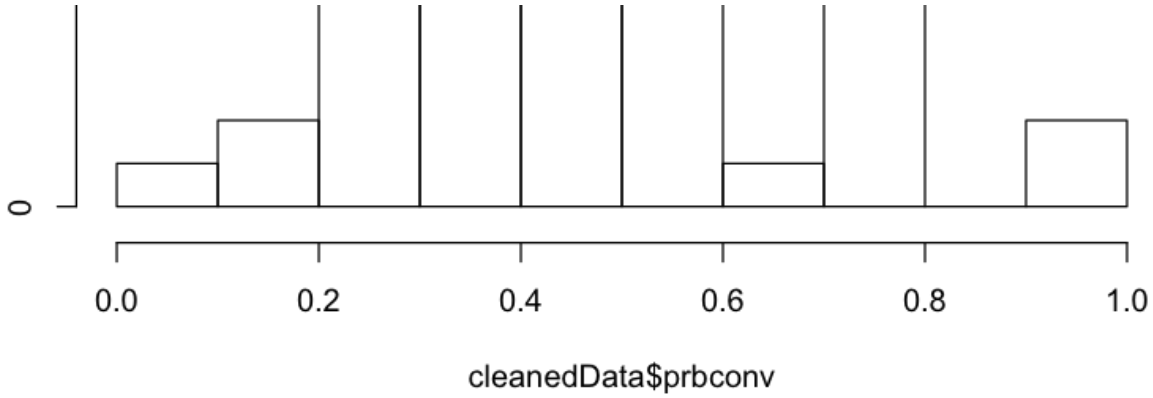
In [6]:

```
hist(cleanedData$prbarr)
hist(cleanedData$prbconv)
```




Frequency
0
5
10
15
20
25
30
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
cleanedData$prbarr

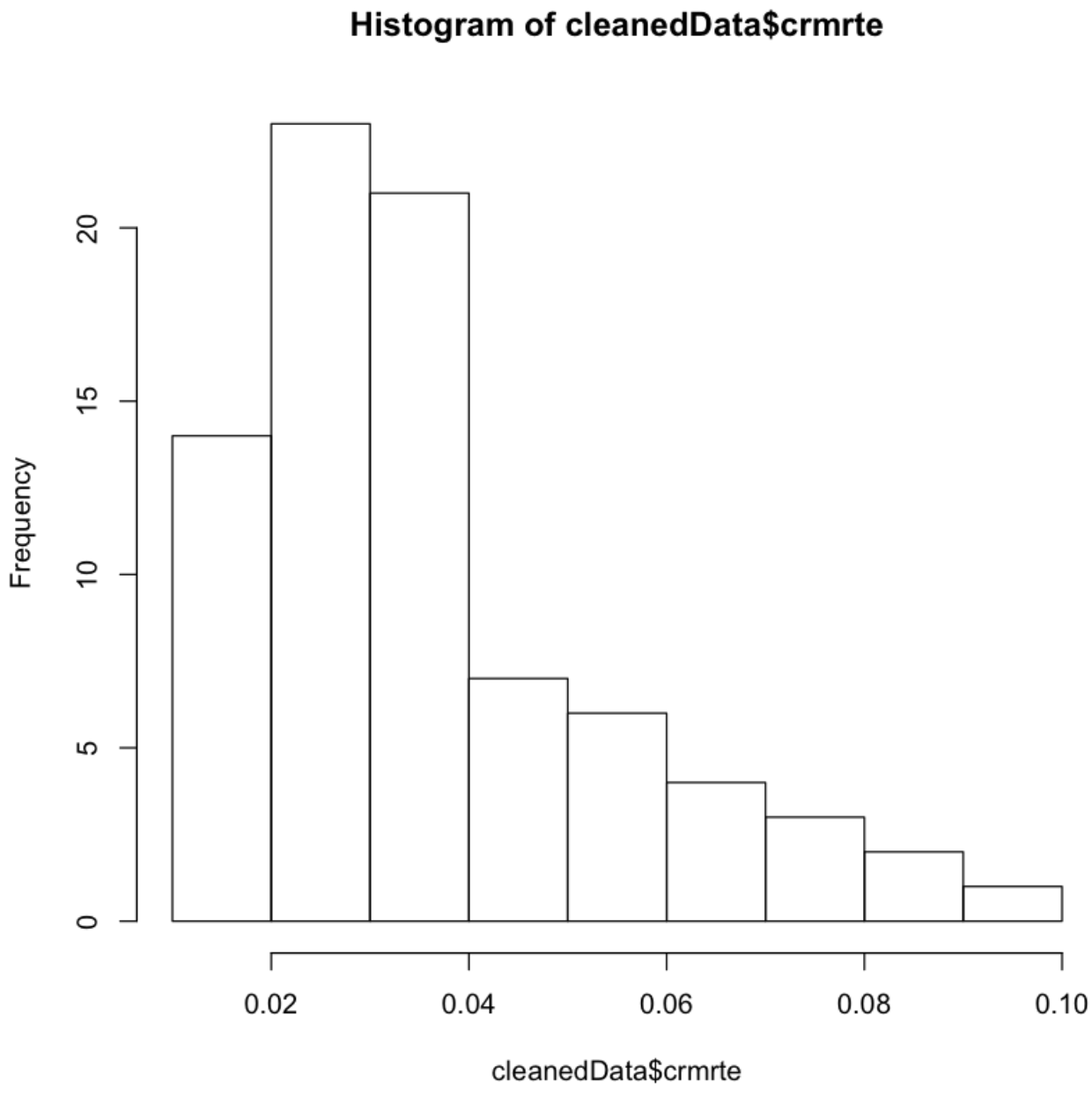

Histogram of cleanedData$prbconv
Frequency
5
10
15
20



> We see that the histograms of both our prbarr and prbconv are normally distributed. Let's take a quick look at our prbcrmrte column as well.

In [7]:

```
hist(cleanedData$crmrte)
```

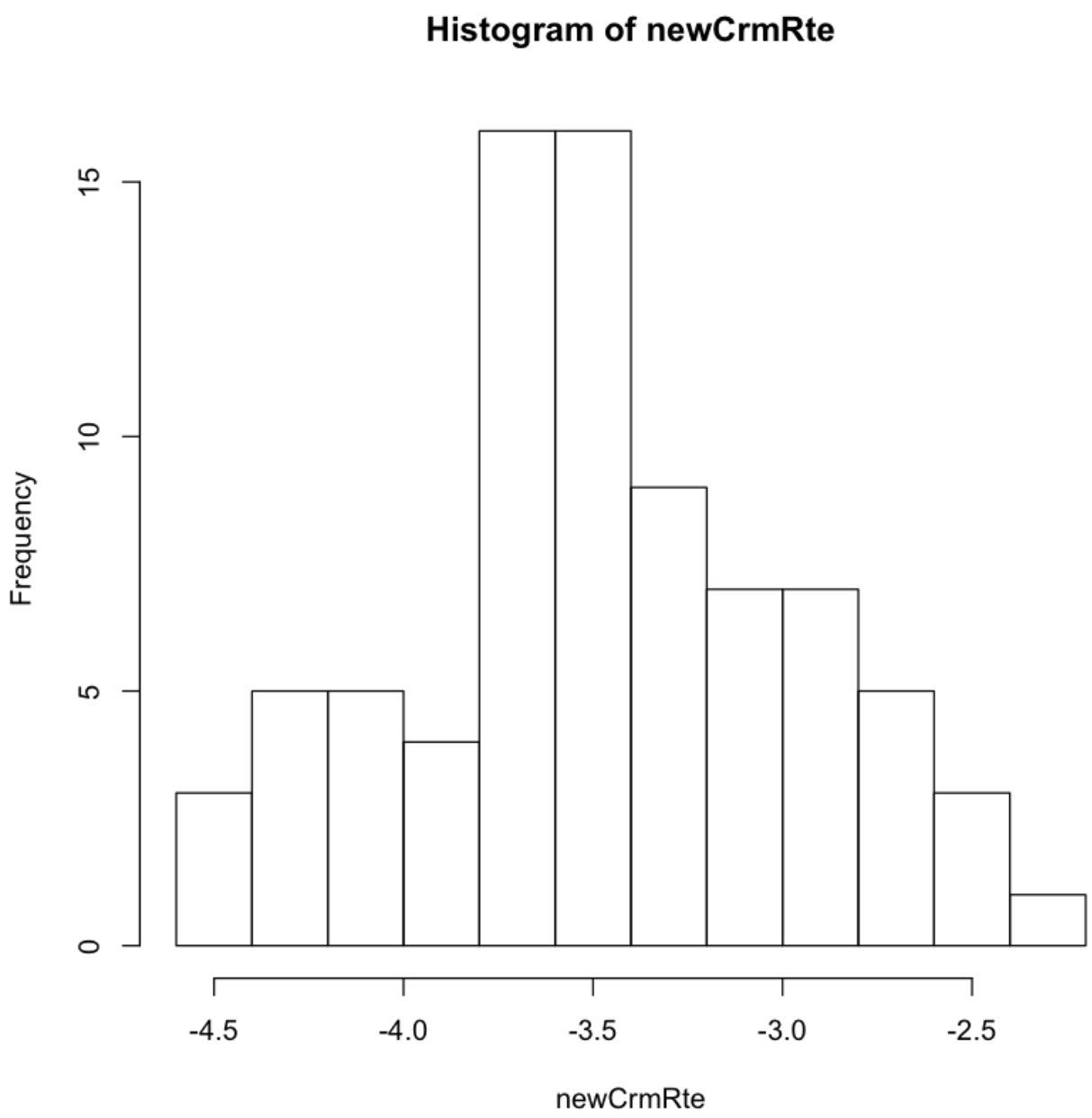


Sidenote on Transformation of our outcome variable of interest: Crimerate

> We see that there is a slight positive skew in our distribution, but this should not be of too much concern for we have a relatively large sample size. Regardless, let's try applying a logarithimic transformation to see if it makes the distribution more normal.

In [8]:

```
newCrmRte = log(cleanedData$crmrte)
hist(newCrmRte)
```



> We see that our distribution appears to be much more normal now. This distribution can be used when proceeding with our EDA. However, there would be an extra step involved of taking the inverse of the logarithm of our newCrmRte at the end of our analysis, so our results are actually interpretable pragmatically. Because of this extra step, and more importantly, because our original distribution for crmrte is not incredibly skewed while also exceeding a sample size of 30 and following Central Limit Theorem, our group has decided to use the original

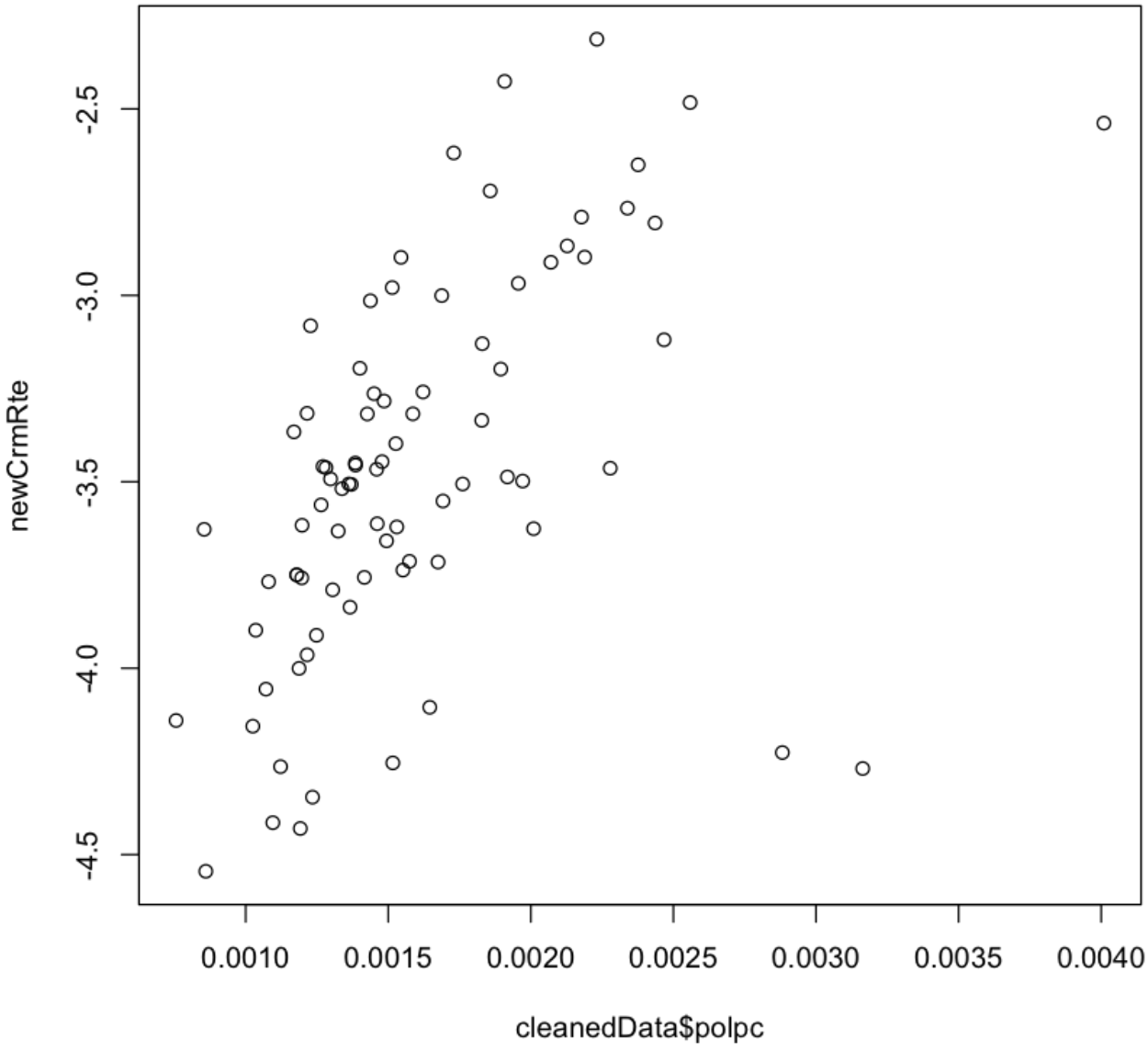
> distribution for the rest of our EDA. This is so we do not have to worry about the interpretation of our results in our final analysis.
>
> In the end, it is most important to be mindful that such transformations can help supplement our analysis when conducting EDA.

Continuing Our Plot Analysis

> Now, let's take a look at the scatterplots of our key columns of interest:

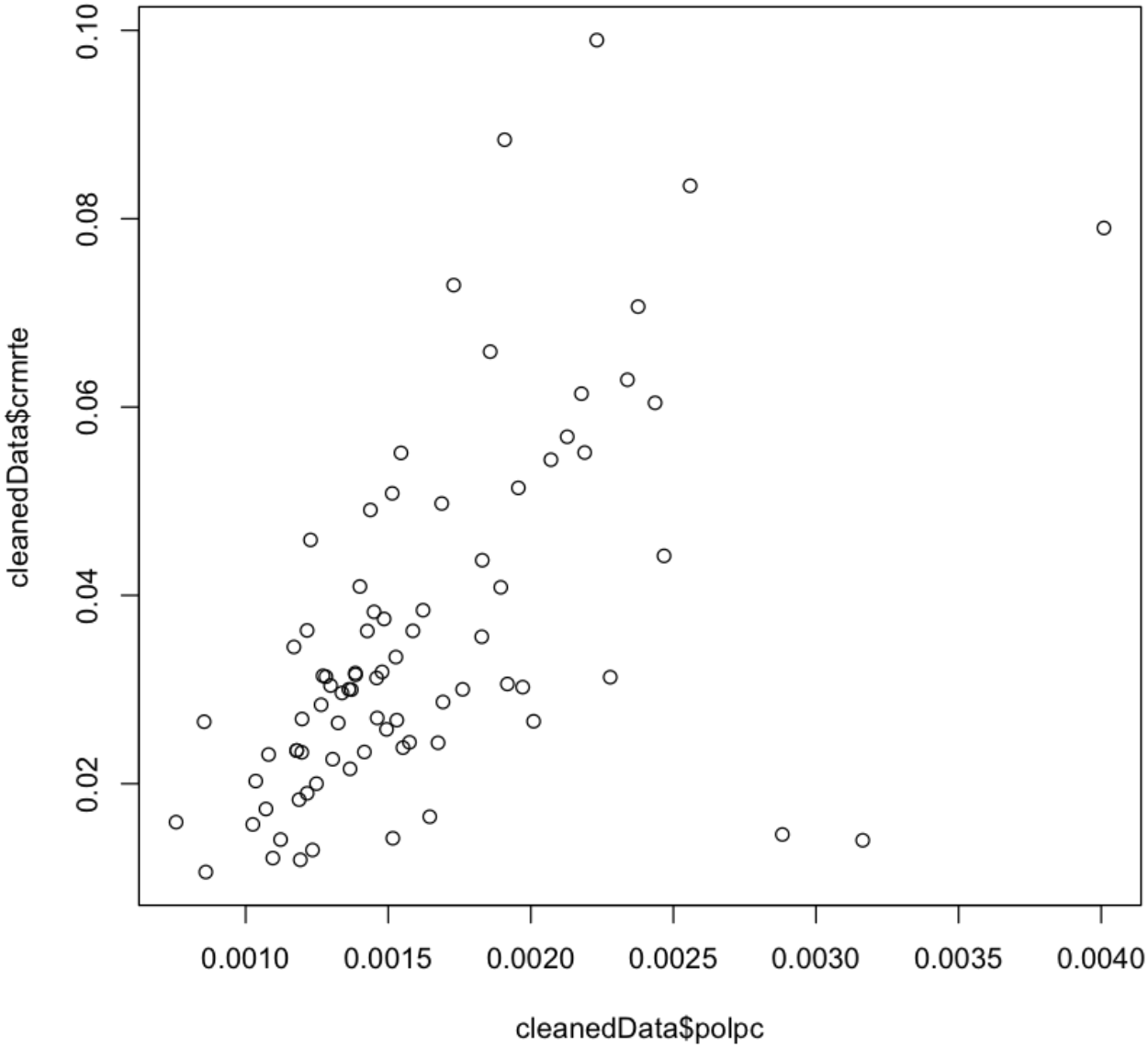
In [9]:

```
plot(cleanedData$polpc, newCrmRte)
```



In [10]:

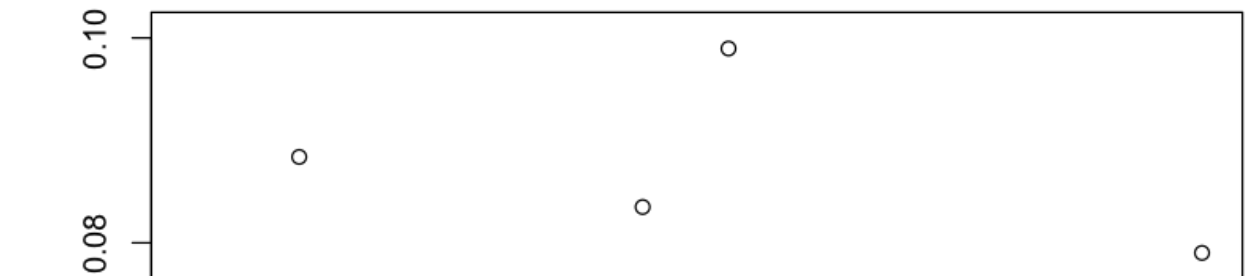
```
plot(cleanedData$polpc, cleanedData$crmrte)
```

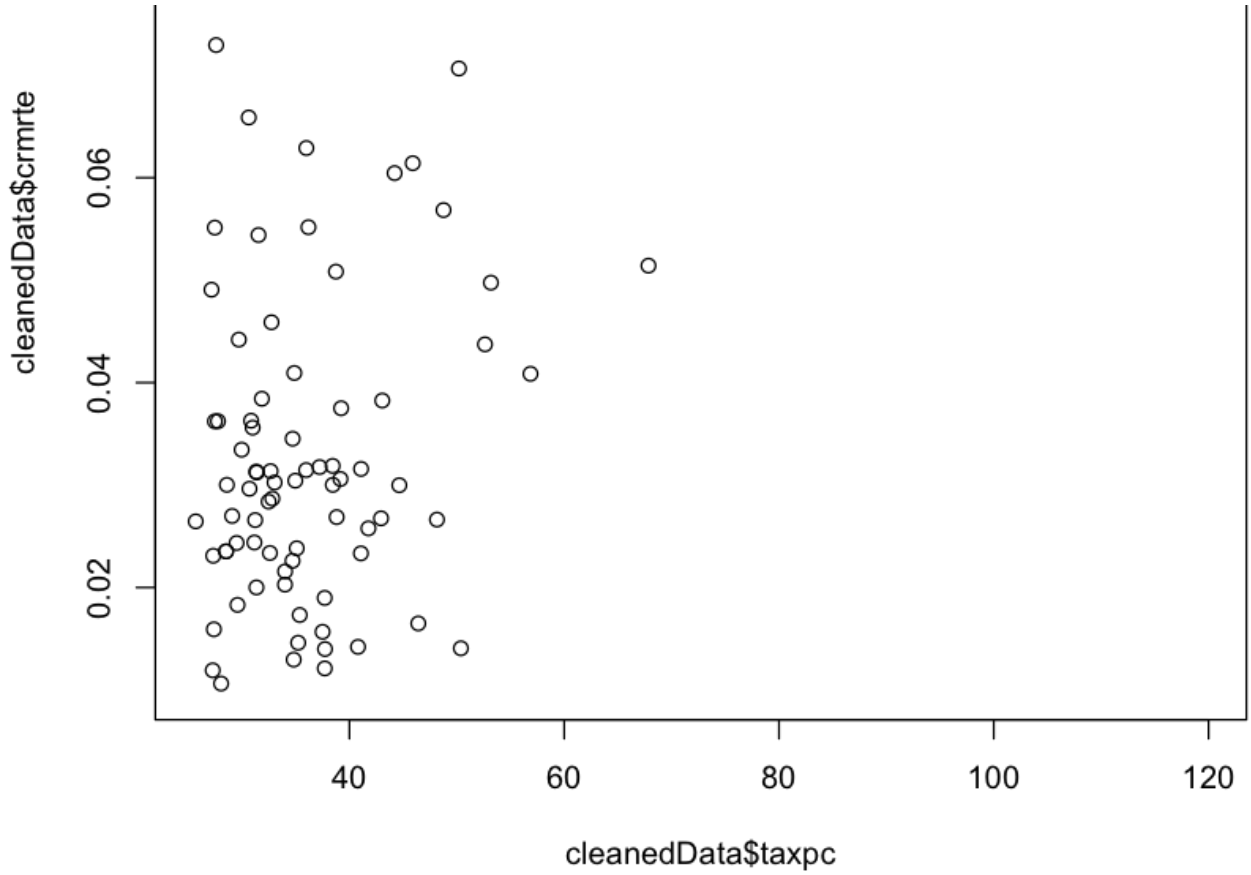


> Based on the above scatterplot, it appears that there is a positive correlation between police per capita and crime rate. This suggests us to explore our other columns of interest as well to see if the correlation is ubiquitous.

In [11]:

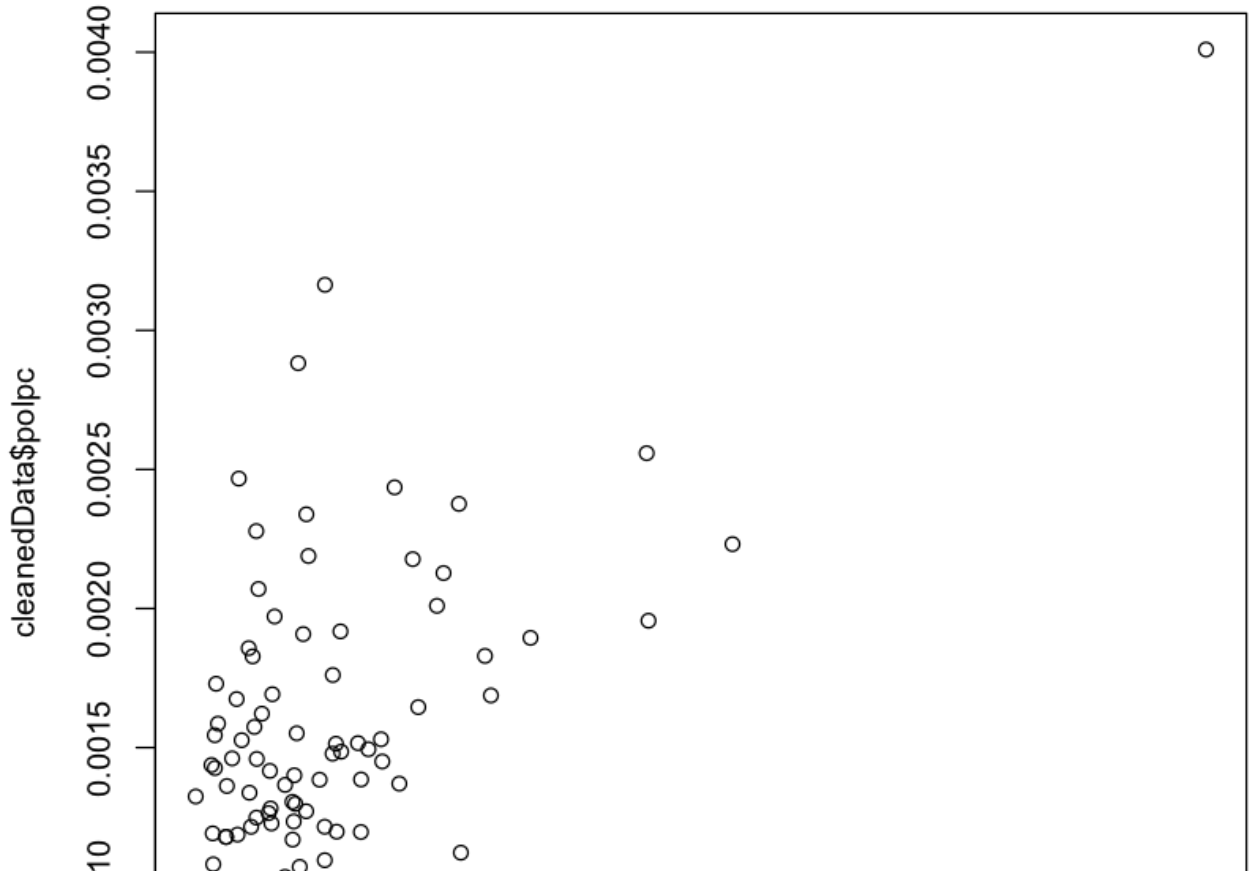
```
plot(cleanedData$taxpc, cleanedData$crmrte)
```

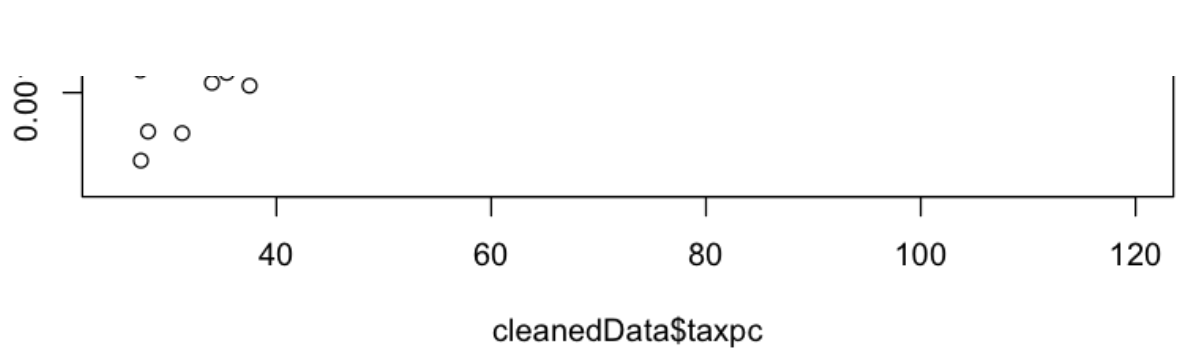




In [12]:

```
plot(cleanedData$taxpc, cleanedData$polpc)
```





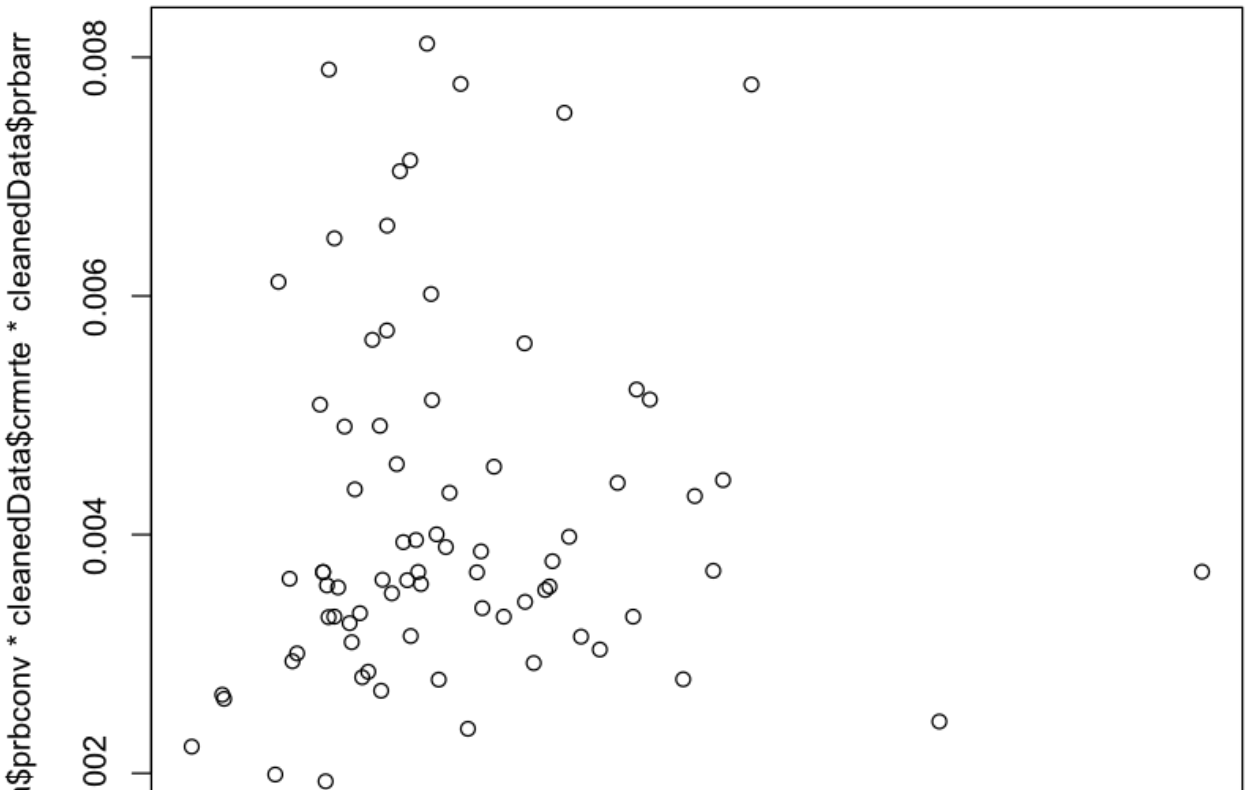
We can see that ultimately all three of our scatterplots have a positive correlation. Namely, in addition to our relationship regarding police per capita and crime rate, it appears that as tax revenue per capita increases, the crime rate increases as well.

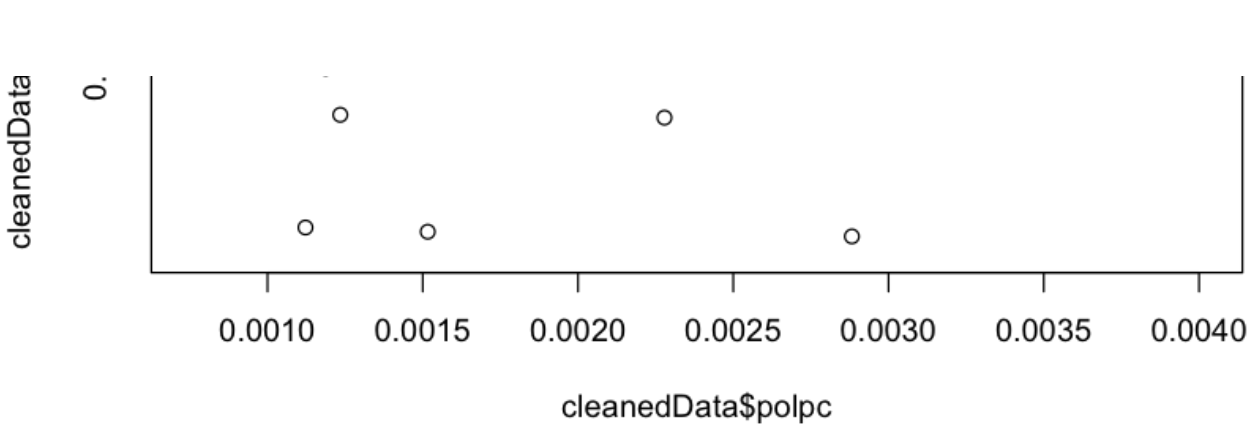
Most importantly, when examining our third scatterplot, we see that there is also a positive relationship between tax revenue and police revenue per capita. This implies that our two covariates have a relationship with each other, which indicates that we must be watchful of multicollinearity when both creating and drawing conclusions from our final model. Ultimately, taxpc and polpc may be quite problematic when conducting our final analysis.

Let's take a look at the ratio of number of convictions per person instead of the ratio of convictions per arrests by multiplying our crime rate with our prbarr and prbconv. This computes the total number of convictions because crime rate is defined as total number of offenses per person, and prbarr is defined as the ratio of arrests per offenses, so multiplying these two will give the ratio of arrests per person. Finally, multiplying this ratio of arrests per person by the ratio of convictions per arrests (prbconv) will give us the ratio of convictions per person, our ultimate goal.

In [13]:

```
plot(cleanedData$polpc, cleanedData$prbconv * cleanedData$crmrte * cleanedData$prbarr)
```





> We see that there is a positive correlation between police per capita and the ratio of convictions per person, and this makes sense, for regions that have a higher number of police officers per person will be able to convict offenders with a higher probability than regions that have less police officers per person.

Model Building

> **Note: To see more detailed analysis on our outcome variable as well as our key explanatory variables of interest, see the EDA above.**
>
> Let's start by building our first model that includes our explanatory variables of key interest: namely, polpc and crmrte. We specifically focus on these two variables based on the positive correlation that appears to occur from our EDA.

In [14]:

```
modelOne = lm(crmrte ~ polpc, data = cleanedData) # builds our first model
```

In [15]:

```
modelOne$coefficients
```

(Intercept)
0.00344610097660025
polpc
19.8219934319961

> Note that the polpc coefficient the crime rate variable increases by 20 each time polpc increases by one. Clearly, there is a positive relationship between crmrte and polpc based on the output of our modelOne.

In [16]:

```
summary(modelOne)$r.square
```

0.317260901062281

> This implies that our model explains 31.7% of the variability in our crime rate variable. Let's see if we can do better by incorporating more covariates. Let's incorporate taxpc to see if it helps increase the accuracy of our results.
>
> Note that we assume that taxpc will have an impact on our outcome variable based on our EDA from above.

In [17]:

```
modelTwo = lm(crmrte ~ taxpc + polpc, data = cleanedData)
```

In [18]:

```
modelTwo$coefficients
```

(Intercept)
-0.00110093457884648
taxpc
0.000318467779532367
polpc
15.1220149034285

> It appears that taxpc has a miniscule if any effect on our crime rate. The extremely small coefficient value for taxpc pales in comparison to the magnitude of the polpc variable. Let's take a look at our R squared for our second model.

In [19]:

```
summary(modelTwo)$r.square
```

0.349665303180333

> We see that our R squared value increased, which is to be expected, for R squared only increases as more covariates are added. It is important to remember that our taxpc variable does not appear to have any impact on determining our crime rate based on the magnitude of the coefficient. Let's introduce even more covariates to demonstrate the robustness of our model. Namely, we introduce:
>
> - Prbarr
>   This variable indicates the probability of being arrested which may or may not be connected with how likely someone is to commit a crime. After all, an extremely low probability of being arrested could incentivize individuals to perform more crimes, as there would be less punishment.
> - Prbconv
>   Similar to the logic specified above, a low probability of being convicted could incentivize individuals to perform more crimes, for there would be a smaller likelihood of being punished if caught.
> - Wmfg
> - Wser
> - Wfir
>   The three wage variables above are introduced to see if wage has any impact on crime rate. They are not critical or involved in our initial research question, but they may end up having an impact on our outcome variable that is unforeseen from our hypothesis.

In [20]:

```
modelThree = lm(crmrte ~ polpc + taxpc + prbarr + prbconv + wmfg + wser + wfir, cleanedData)
```

In [21]:

```
modelThree$coefficients
```

(Intercept)
0.0244629702923924
polpc
10.1266439549304
taxpc
0.000235034256410296
prbarr
-0.0718698512977742

prbconv
-0.0223109701402396
wmfg
2.75077798944281e-05
wser
-4.01326811047764e-05
wfir
5.57666245754976e-05

> Based on the coefficients of modelThree, it appears that none of the covariates that are introduced have a significant impact on our outcome variable of crime rate. The magnitudes for the wage variables in particular are extremely small (e-05 in magnitude), and once again, the only variable that seems to have an impact is our original variable of analysis, polpc.
>
> Let's take a look at the variability in our modelThree.

In [22]:

```
summary(modelThree)$r.square
```

0.58253639872753

> This time, we have incorporated many other variables that may possibly explain the crimerate. Once again, based on the coefficient values, it is apparent that polpc has the most significant impact on crmrte, and none of the other covariates appear to have any impact at all. Of course, our model in this case explains more of the variability in the crime rate, but this is just attributed to the increase in covariates with increases our R squared.

CLM Assumptions

It is important to evaluate our CLM assumptions for our three models that we have calculated. These assumptions are specified below:

- OLS Assumption 1: The linear regression model is “linear in parameters.”
- OLS Assumption 2: There is a random sampling of observations
- OLS Assumption 3: The conditional mean should be zero.
- OLS Assumption 4: There is no multi-collinearity (or perfect collinearity).
- OLS Assumption 5: Spherical errors: There is homoscedasticity and no autocorrelation.
- OLS Assumption 6: Error terms should be normally distributed.

Note that we will complete this analysis in part 3 as specified in the Lab 3 assignment document.

Omitted Variables

> Possible omitted variables that may have an impact on crime rate include the following: unemployment rate, type of crime, per capita income, and percentage of people living in poverty, as well as the actual crime rate, where actual implies crimes that have not been reported.

- **Unemployment Rate**

  > Our group's hypothesis is that an increase in unemployment rate will have a large increase in crime rate. This is attributed to the fact that when individuals are unemployed and/or in poverty, they have a smaller cost to their lifestyle of commiting a crime. This is because their lifestyle is already in a poor situation, so of course, this would ultimately result in a higher probability of comitting a crime.

- **Type of Crime**

> It is important to understand what types of crimes are included in the dataset. Specifically, do the crimes include everything ranging from a simple speeding ticket to a homicide? This is of important consideration because statistics such as tax per capita would have less impact on an individual speeding or not. However, individuals that are taxed more may be more inclined to commit crimes that are of higher degree such as robbing a convenience store, robbing a bank, etc. We hypothesize this because individuals who are already in a poor economic situation may be inclined to commit crimes that will benefit their financial wellbeing. For this reason, it is important to consider the type of crime, for misdemeanours such as speeding are unlikely to be correlated with statistics such as tax per capita. Ultimately, our group hypothesizes that the type of crimes in this dataset incorporate crimes that may in fact be heavily biased towards misdemeanours such as speeding tickets, for of course the frequency of such crimes is more rampant compared to more notorious crimes such as robbing a bank. If the dataset was composed of crimes of higher degree/significance, we hypothesize that variables such as tax per capita would have a more significant impact on the crime rate, in particular, a large and positive impact. This is because a larger tax per capita would hypothetically lead to more individuals who are inclined to commit crimes of higher degree to benefit their financial and overall wellbeing.

- **Per Capita Income**

> Per capita income is a variable of prominent concern, for individuals that have more income on average would be less likely to commit crimes due to having no incentive to from a financial standpoint. Therefore, this omitted variable would have a negative and large bias. The significance of this variable is also corroborated by the idea that many crimes such as theft and robberies hinge upon the objective of getting more income, so more income would ultimately result in a lower frequency of such aforementioned crimes.

- **Percentage of people living in poverty**

> Based on similar logic that is presented in both per capita income and unemployment rate, people who are impoverished would be more likely to commit a crime due to the fact they have more to gain and less to lose. Based on this logic, this omitted variable would result in a postive and large bias, for a higher percentage of impoverished people in a given county would likely result in a higher crime rate.

- **Actual Crime Rate**

> The crime rate that is presented in our dataset is the document crime rate, and this corroborates the idea that as the police per capita increased, the crime rate increased. However, the actual crime rate is an omitted variable in the sense that there is no way of identifying whether the true number of crimes have been actually reported. In fact, regions that have a signficantly smaller number of police per capita may have an even smaller number of reported crimes, for there would be less police officers to report/identify crimes. Moreover, we can assume the direction of the relationship between actual crime rate and the reported crime rate is positive, which leads us to assume that the relationship between police per capita and crime rate is also positively correlated. Ultimately, we can determine whether this omitted variable bias is postitive or negative based on the relationship this has with the reported crime rate and the police per capita variable.

Regression Table

In [26]:

```
install.packages("stargazer")
library(stargazer)
stargazer(modelOne, modelTwo, modelThree, type = "text", report = "vc", title = "Regression Table f
or Analysis of Crime Rates", keep.stat = c("rsq", "n"), omit.table.layout = "n")
```

```
Installing package into '/Users/sdas115/Library/R/3.6/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
 /var/folders/wk/6pjd0d_96dqbhytxj4gzhzpsk8tt5m/T//RtmpU1Tb3Z/downloaded_packages
```

```
Please cite as:

 Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
 R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
Regression Table for Analysis of Crime Rates
===================================
              Dependent variable:
             ---------------------
                      crmrte
               (1)    (2)     (3)
-----------------------------------
taxpc                0.0003  0.0002

prbarr                        -0.072

prbconv                       -0.022

wmfg                         0.00003

wser                         -0.00004

wfir                          0.0001

polpc          19.822 15.122  10.127

Constant       0.003  -0.001  0.024

-----------------------------------
Observations    81     81       81
R2             0.317  0.350   0.583
===================================
```

> Indeed, it is clear that our regression table matches our model analysis from above. This validates our individual model analysis and also easily highlights our key coefficients from our regression analysis.
>
> Note that practical significance will be discussed in part 3 based on the Lab 3 document.

Conclusion

> In conclusion, it appears that based on the data that is provided, polpc is the only significant variable that has an impact on crime rate. The other variables have miniscule if any effect based on our model regression analysis, so they are not of important interest. For this reason, our group highly recommends proposing policies that are specifically related to police per capita. In particular we recommend the following policy:

- Counties that have higher crime rates should have higher police per capita.

  > This policy appears to be obvious at first glance, but it is particularly corroborated by the model regression analysis as well as the scatterplots that manifest the positive correlation that exists between crime rate and police per capita. Indeed, as crime rate increases, police per capita increases as well.

> Logically, this makes sense. Regions that have higher rates of crime need more police officers to patrol and guard the region. Yet, it is safe to say that this apparently evident policy is not properly enacted in our society, for otherwise, our crime rates would not be so proliferate in many of the counties that are present in the dataset. One particular metric stands out in particular: the maximum crime rate for a county in this dataset is 0.09897, which implies that if there is a group of 10 people in the aforementioned county, 1 of them have committed a crime. This is simply ludicrious, and it is imperative that our society enacts policies to deter such effects.
>
> Most importantly, it is crucial to remember the omitted variables that were highlighted in this report. Further analysis of such variables may lead to more insights that will help reduce the crime in counties, which overall, will indubitably make our society a safer place.