w203_Lab3_group5_draft1: Joanna wang, Douglas(Zeliang) Xu

Introduction

In this report, we will analyze the crime statistics for a selection of counties in North Carolina and purpose to find out some main factors of crime rate. After the initial analysis on the data, we will purpose our research question.We will then provide several policies or suggestions based on our research findings.

Part1: Looking at the data

This part is to look at the dataset in general to grow our understand of the data:

```
library(car)
```

```
## Loading required package: carData
```

```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
data <- read.csv(file="crime_v2.csv", header=TRUE, sep=",",na.strings=c("`","","NA"))
objects(data)
```

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

Here we have the summary of all the veriables in the data set, to spot any anormaly.

```
summary(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        
##  Min.   :0.06838   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746  
##  1st Qu.:0.34541   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231  
##  Median :0.45283   Median :0.4234   Median : 9.100   Median :0.001485  
##  Mean   :0.55128   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702  
##  3rd Qu.:0.58886   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877  
##  Max.   :2.12121   Max.   :0.6000   Max.   :20.700   Max.   :0.009054  
##  NA's   :6         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
```

We noticed that there are 6 NAs in all the variables. To took a closer look at the last few rows of the date set to verify the NA entires.

```
tail(data,7)
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 91    197   87 0.0141928 0.207595 1.18293 0.360825  12.23 0.00118573
## 92     NA   NA        NA       NA       NA       NA     NA         NA
## 93     NA   NA        NA       NA       NA       NA     NA         NA
## 94     NA   NA        NA       NA       NA       NA     NA         NA
## 95     NA   NA        NA       NA       NA       NA     NA         NA
## 96     NA   NA        NA       NA       NA       NA     NA         NA
## 97     NA   NA        NA       NA       NA       NA     NA         NA
##     density    taxpc west central urban pctmin80    wcon     wtuc    wtrd
## 91 0.889881 25.95258    1       0     0  5.46081 314.166 341.8803 182.802
## 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
##        wfir     wser   wmfg   wfed   wsta   wloc        mix    pctymle
## 91 348.1432 212.8205 322.92 391.72 385.65 306.85 0.06756757 0.07419893
## 92       NA       NA     NA     NA     NA     NA         NA         NA
## 93       NA       NA     NA     NA     NA     NA         NA         NA
## 94       NA       NA     NA     NA     NA     NA         NA         NA
## 95       NA       NA     NA     NA     NA     NA         NA         NA
## 96       NA       NA     NA     NA     NA     NA         NA         NA
## 97       NA       NA     NA     NA     NA     NA         NA         NA
```

Some of the things we see in the dataset and have lead us to some decisions: 1. 6 NAs for all columns. We will remove those entries. 2. Year is alwasy 87. We will take out the year column, because it does not help us with our crime rate analysis 3. “prbarr” max > 1. Probability should not be greater than 1. 4. “prbconv” strange characters and blank spaces; also the probability is bigger than 1 5. taxpc, what is the unit, what does it mean? Outlier at 119. Is the unit %? 6. pctmin80 data is too old 7. 15-23: different industry avg. wages 8. 24 mix: ratio of face-to-face crime 9. percentage young male (what is the age:15-24)

With the import method modified, we are able to address the missing values and the special character error in the $prbconv$ data column

Data Cleaning

1. remove NA

```
data <- na.omit(data)
```

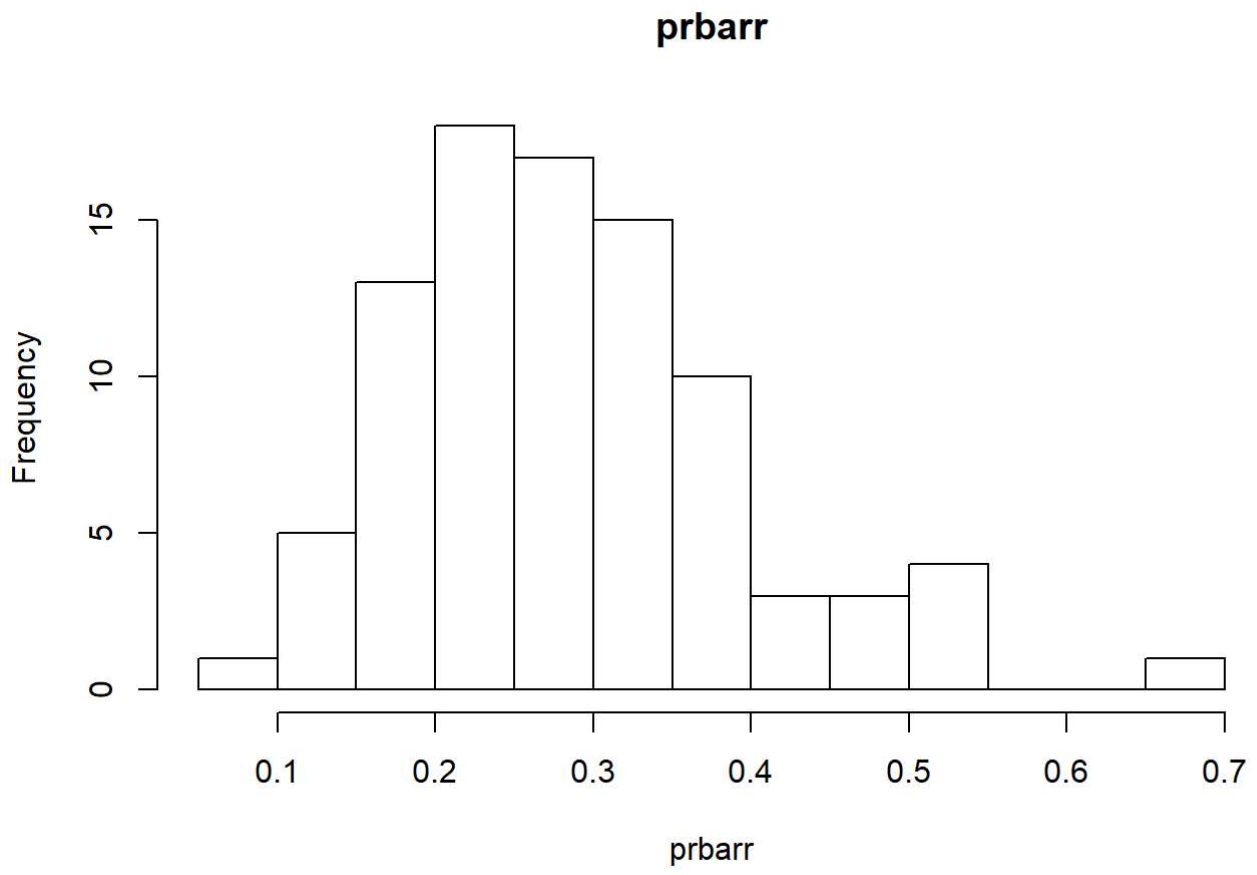
2. take out year column as it does not mean anything in our analysis

```
data_clean <- subset(data,select=-c(year))
objects(data_clean)
```

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

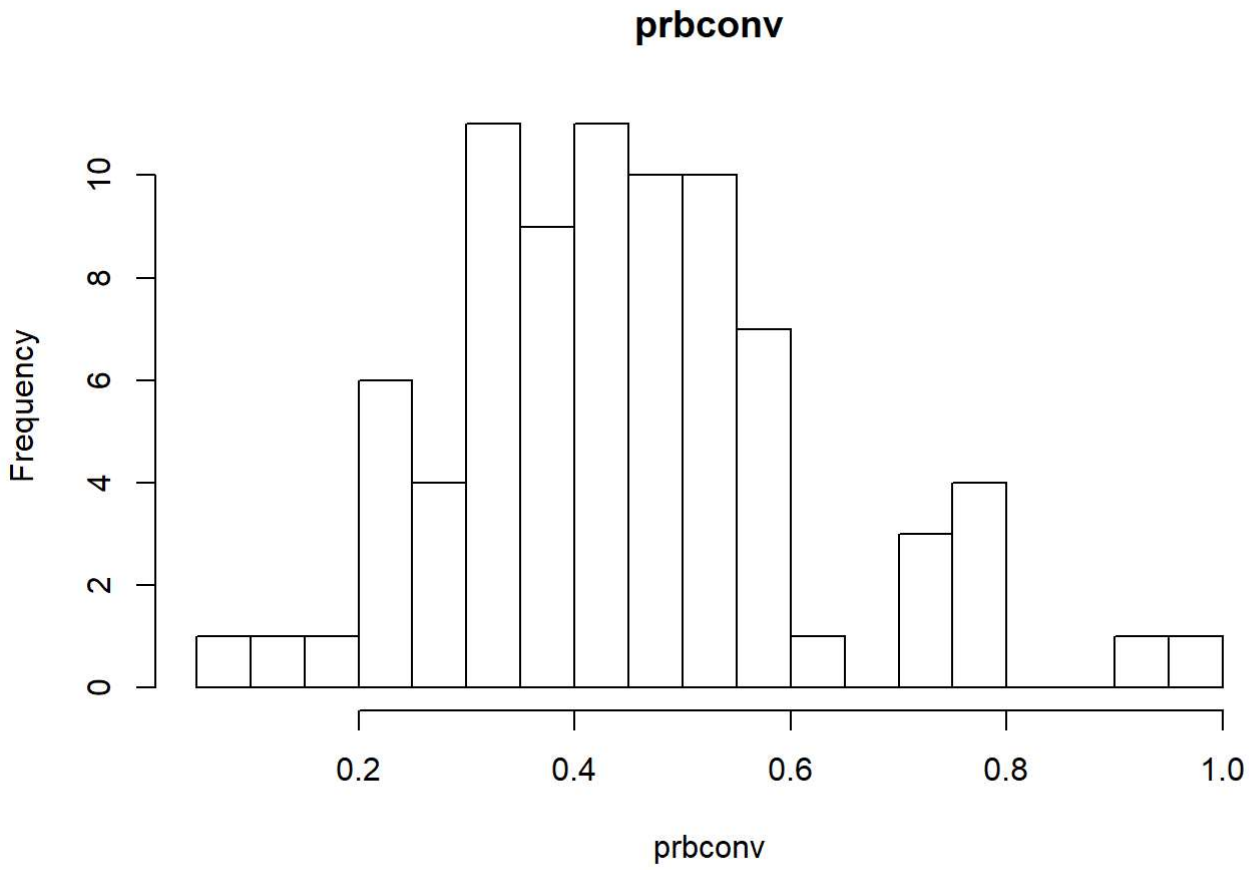
3. “prbarr” max > 1. Becuase this variable is supposed to represent the probability of arrest, the max should never exceed 1.

```
data_clean <- subset(data_clean,data_clean$prbarr <1)
hist(data_clean$prbarr, breaks=20,  main="prbarr", xlab="prbarr")
```



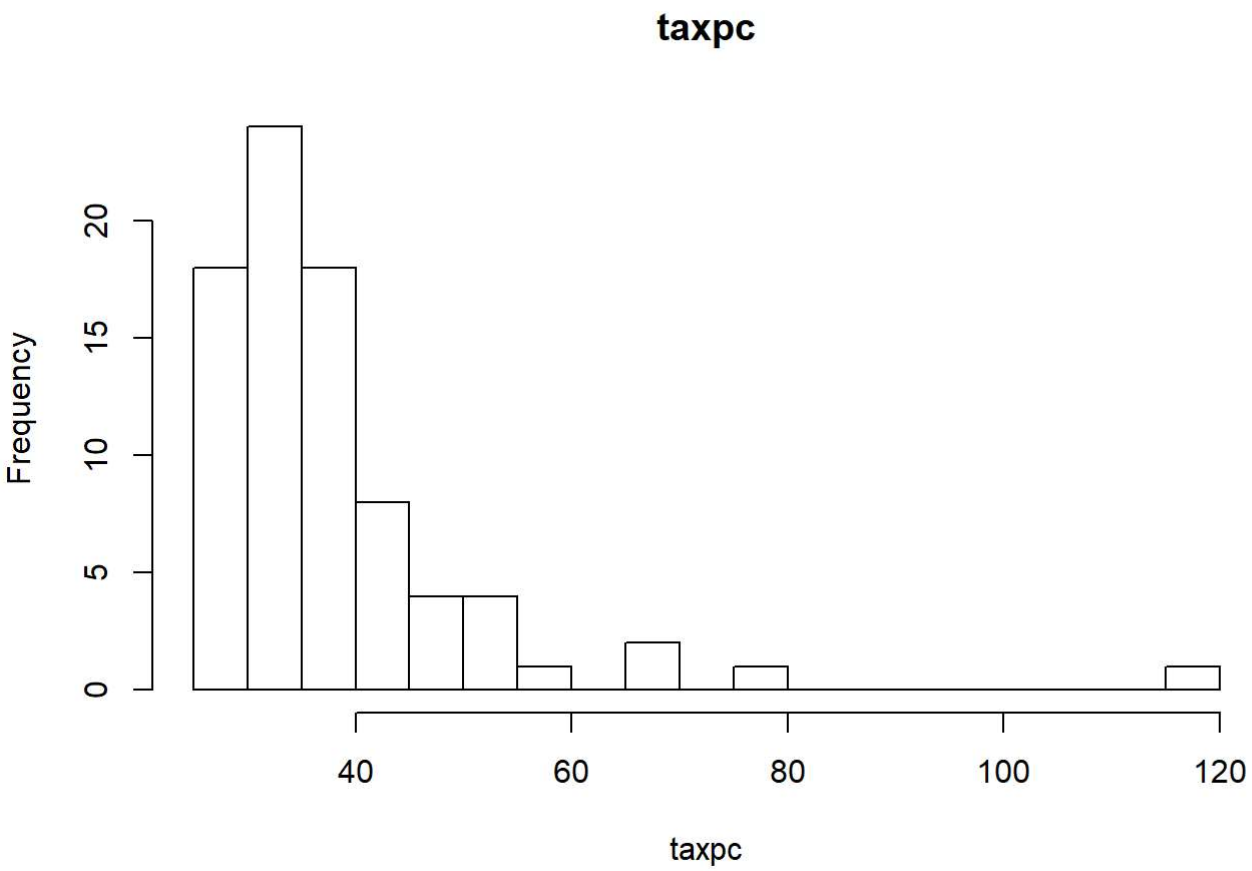
4. “prbconv” column contains strange characters and blank spaces; Also becuase this is the probability, it should not contain entries that are bigger than 1
   blank space and strange characters are taken care of in data import

```
data_clean <- subset(data_clean,data_clean$prbconv <1)
hist(data_clean$prbconv, breaks=20, main="prbconv", xlab="prbconv")
```



5. taxpc: There seems to be some anomaly that is very far apart from other data points, but we have no evidence to say whether the data has anomaly or not. It could be an error or it could just be that the county has high tax per capita

```
hist(data_clean$taxpc, breaks=20, main="taxpc", xlab="taxpc")
```



6. over-paid service industry: we found that the maximum value of the avg. weekly wage of service industry is above 2000, which is way more than the other industries. Based on the background knowledge, we don't believe there should be significant difference between the service industry and other industries in terms of compensation difference, and the data point above 2000 should be an error in the data, and we decided to remove them

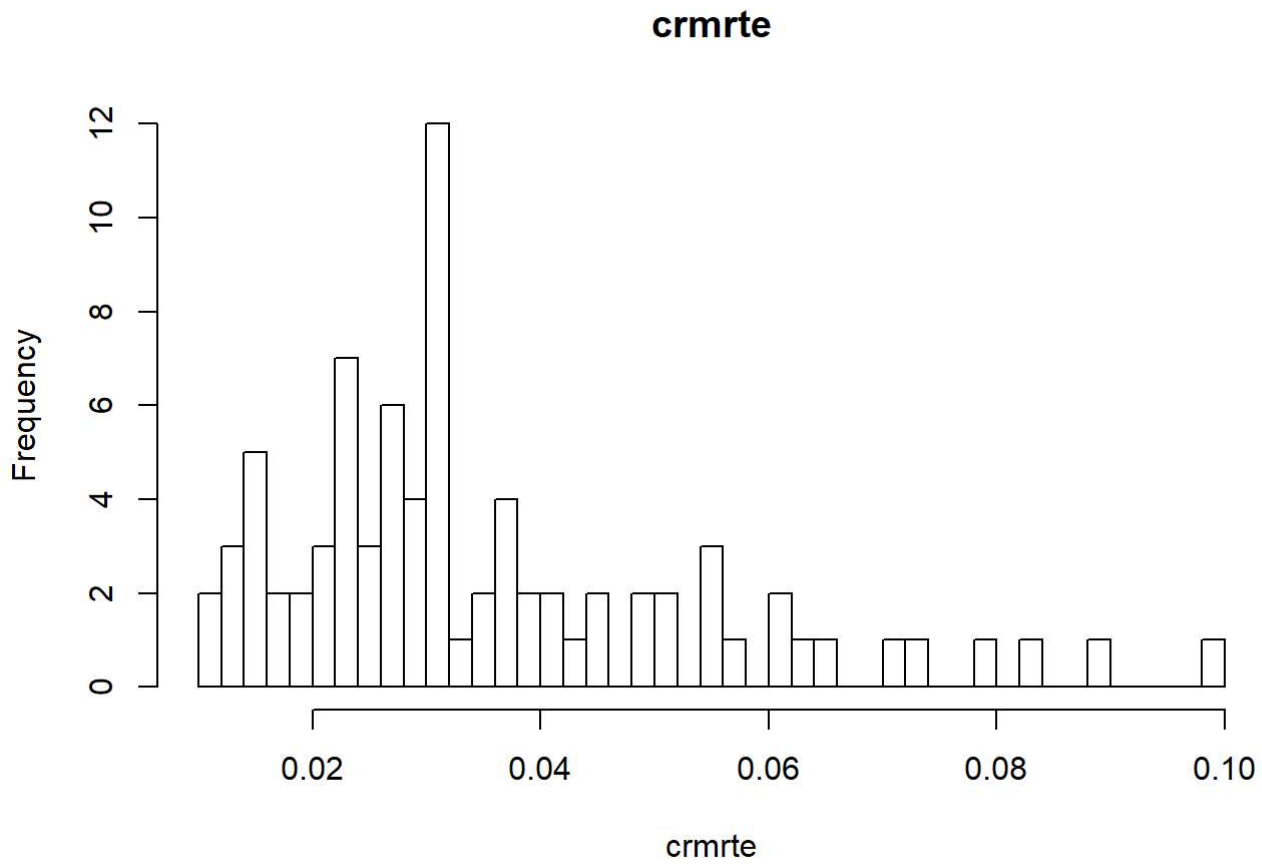
```
data_clean <- subset(data_clean, data_clean$wser<1000)
summary(data_clean$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   230.3   253.6   255.2   278.1   391.3
```

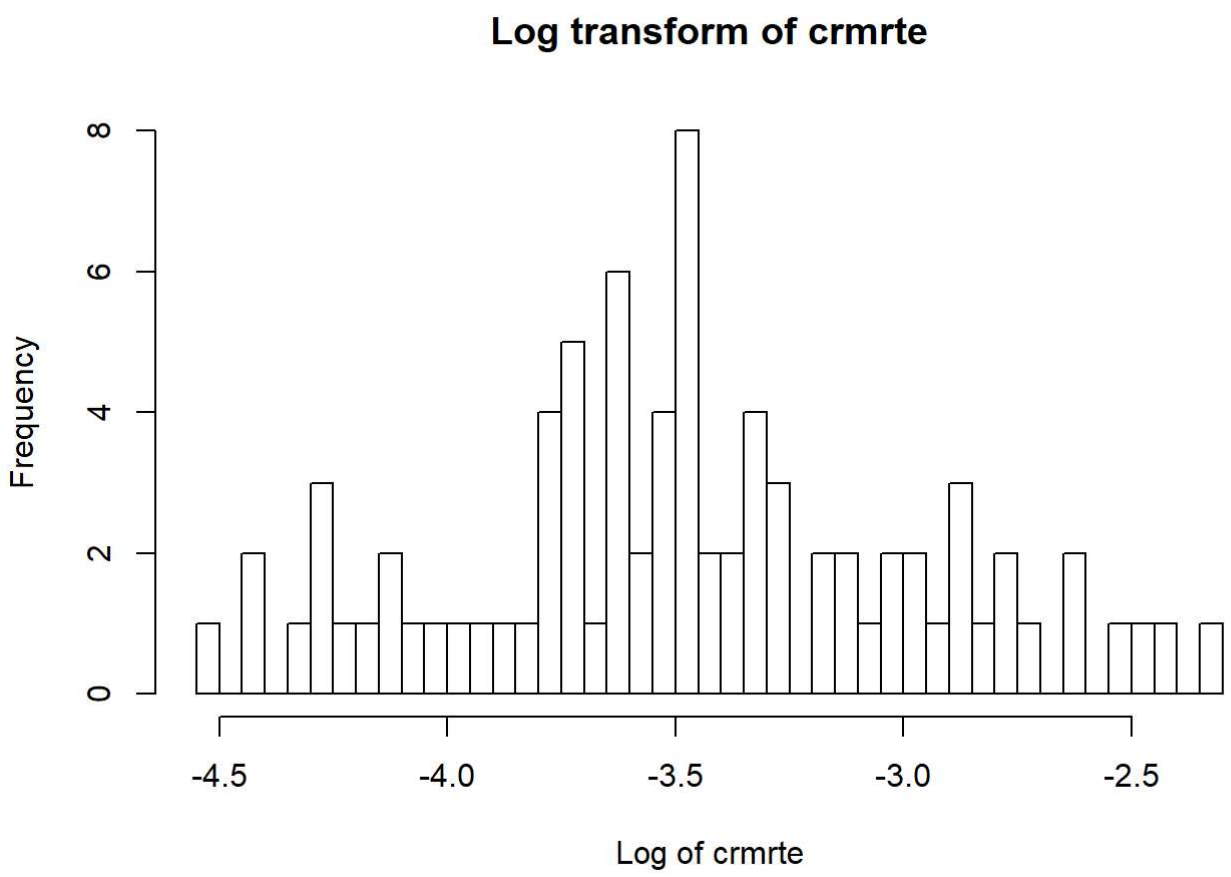
EDA

1. As the first part of the data analysis, we are taking a look at the distributions of the dependent variable:

```
hist(data_clean$crmrte,breaks=50, main="crmrte", xlab="crmrte")
```




```
hist(log(data_clean$crmrte),breaks=50, main="Log transform of crmrte", xlab="Log of crmrte")
```

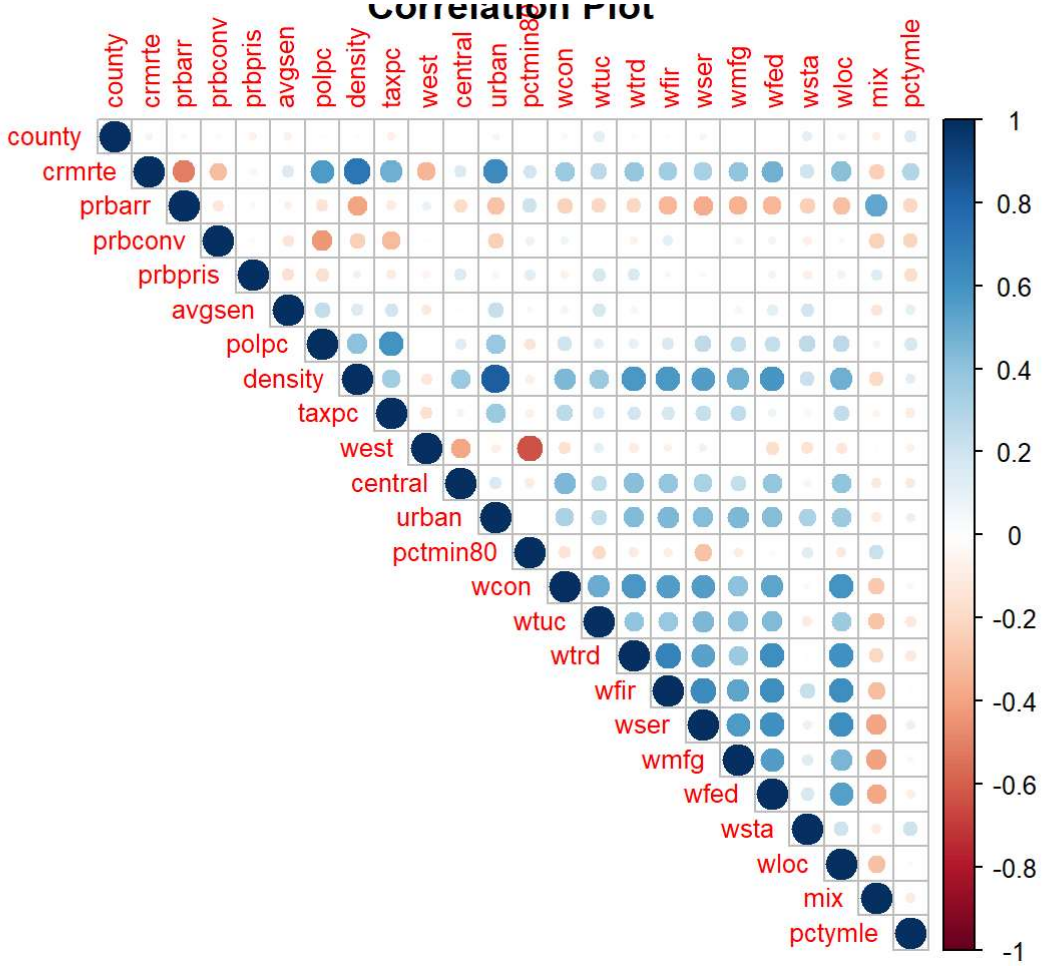


2. As the second part of the EDA, we take an initial look at the simple correlation between each variable inside the dataframe to help us understand the general correlations betwen each varible and the dependent variable to help identify the explanatory variables of interest

```
# correlation plot
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(data_clean[,sapply(data_clean,is.numeric)]),is.corr=T, method = "circle", type='upp
er',main = "Correlation Plot", tl.cex=0.8)
```



Research Question

After the initial data cleaning and data analysis, we have defined our research question to be: how to reduce crime rate within North Carolina, and especially what are the most effective measures in reducing crime rate.

The question is asked to help the political campaign to propose effective strategies to reduce crime rate, especially in the area that might have been neglected before. From the simple correlation plot, we found that variable mix has strong positive correlation with $prbarr$. It can make sense because face-to-face crimes are more severe and usually have more police concentration and resources that lead to higher probability of arrest. However, it might not be very easy for non face-to-face crime, which is still the majority of crimes that happened, but actually seem to have pretty low arrest rate. This conclusion can be obtained by the fact that crime rate ($crmrte$) is negatively correlated with probability of arrest($prbarr$) while percentage of face-to-face crime (mix) is positively correlated with probability of arrest ($prbarr$). Therefore, how to increase $prbarr$ for non face-to-face crime is really the strategy we need to focus our energy on.

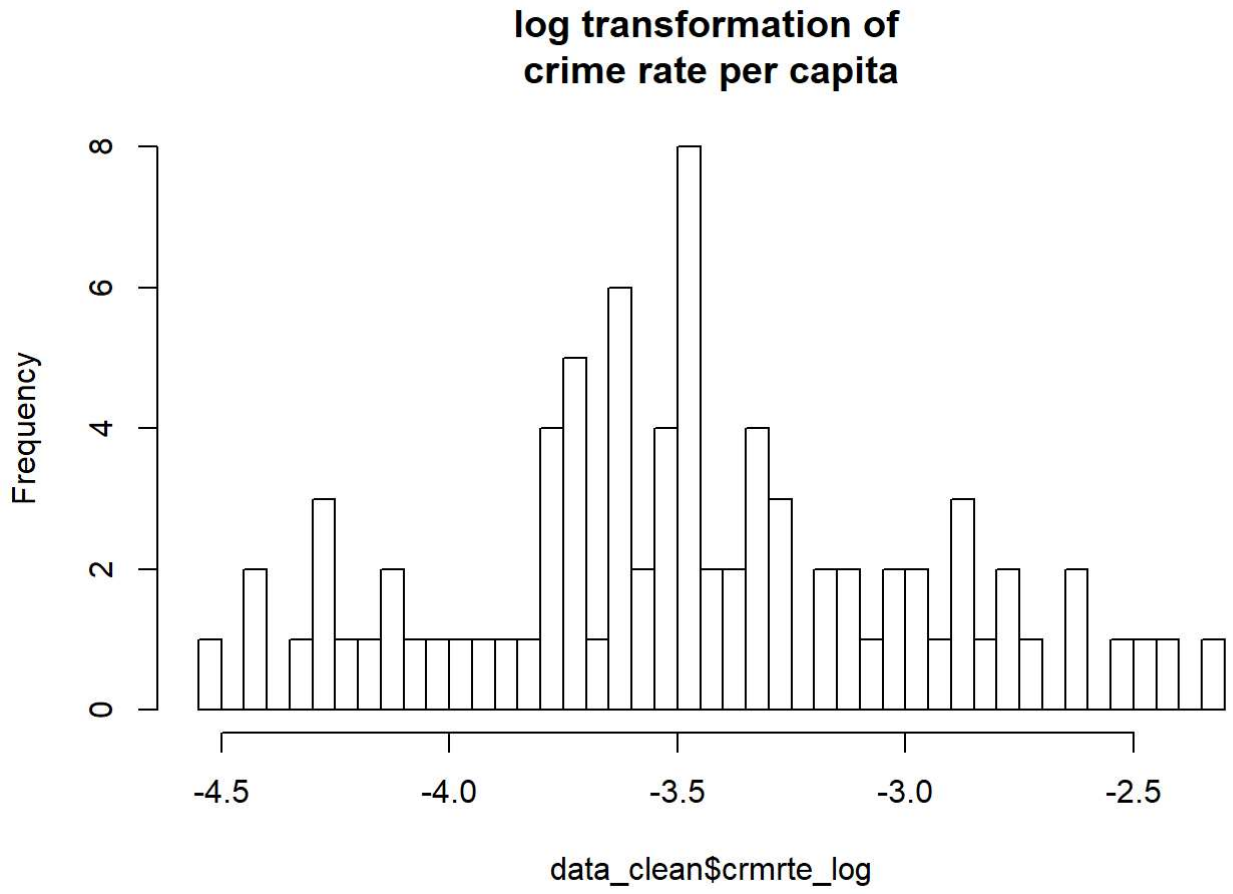
Fitting model 1

model 1: $crmrte = _0 + _1 * prbarr + _2 * prbconv + _3 * avgsen + _4 * prbpris + u $

Based on the understanding of what each parameter stands for, we have chosen crmrte as the only dependent variable to use for our analysis. The variable is a direct indicator of average crime commited to North Carolina counties. To start with, we took a look at the distribution of the dataset.

From the analysis of the EDA, the distribution of crmrte does not look particularly normal, but the log transformed crmrte looks much more normally distributed. To reduce the standard error in the model building process, we decided that in our model fitting, we are going to use the log transformed $crmrte$ as our dependent variable. That being said, we are creating another variable that indicates the log transformed $crmrte$

```
data_clean$crmrte_log <- log(data_clean$crmrte)
hist(data_clean$crmrte_log, breaks=50, main="log transformation of
crime rate per capita")
```

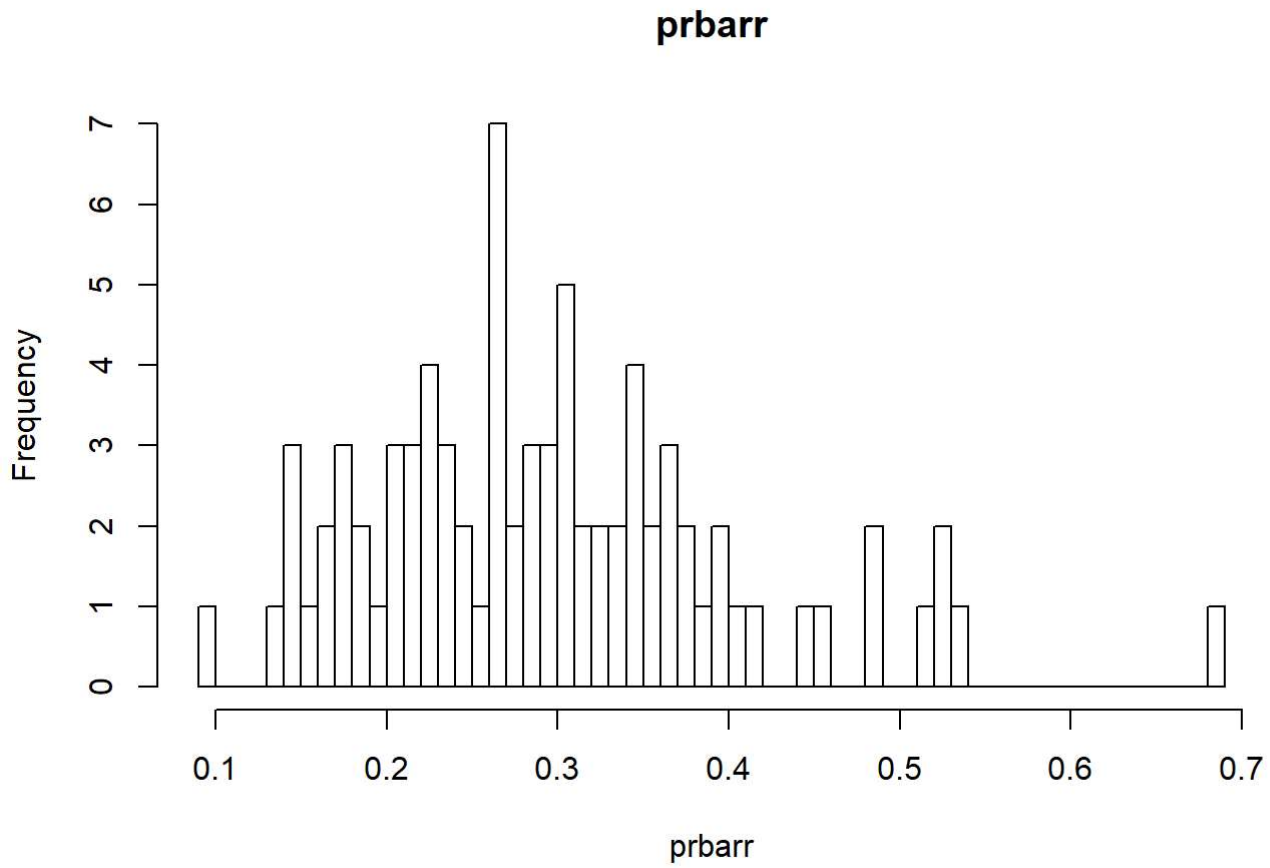


Looking at the other variables, and the simple correlation plot in initial EDA, we are proposing the explanatory variables that we believe contribute to crime rate:

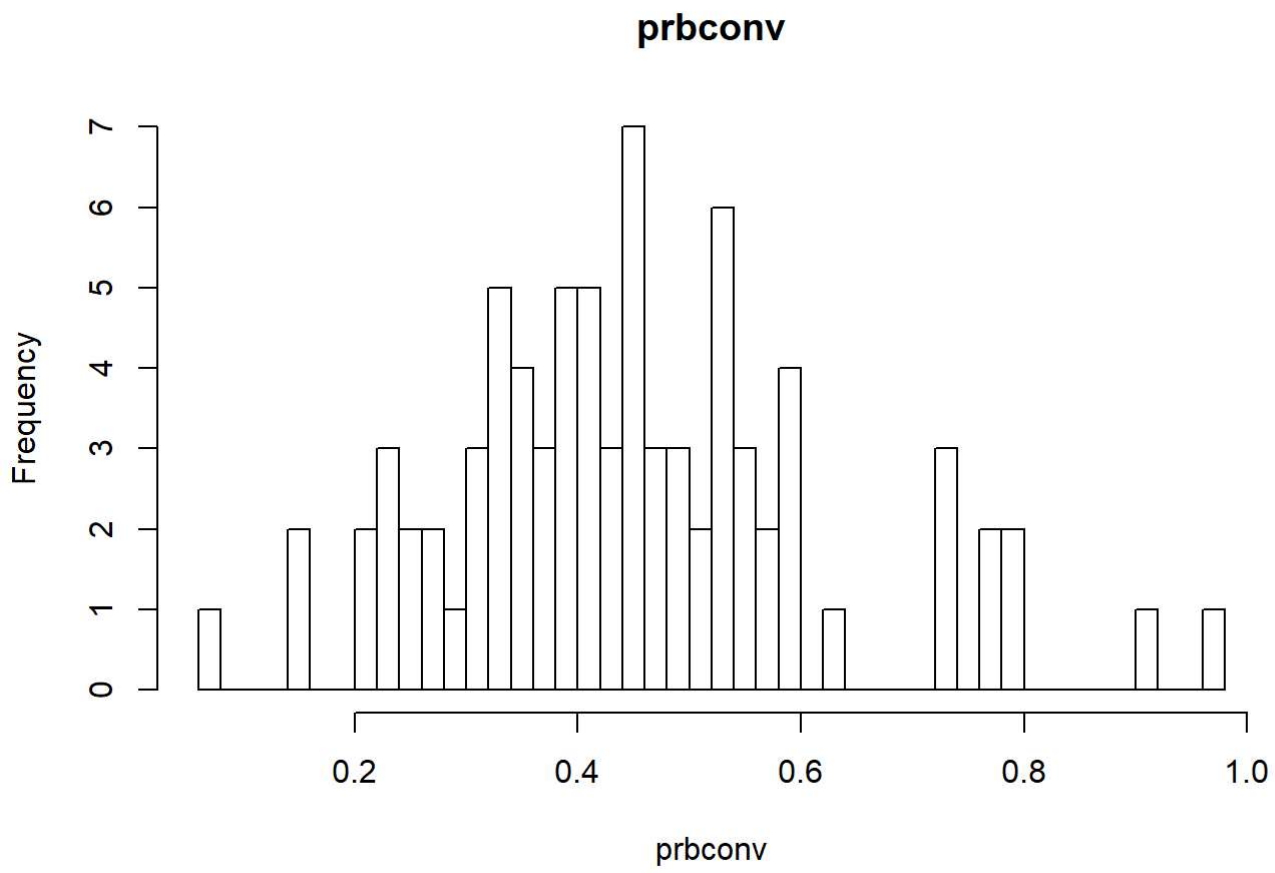
1. prbarr: the probability of arrest should be a direct contributing factor to crime rate. In other words, if people who have potential to commit a crime believe the chance of them getting arrested is small, then it might encourage them to commit a crime
2. prbconv: after getting arrested, getting suspects convicted are the only way to let them take the punishment they deserve.
3. prbpris
4. avgsen: both probability of prison sentence reflect the severity of the punishment, which should directly impact the crime rate

With the above being proposed, we decided to take a look at the distribution of each explanatory variable:

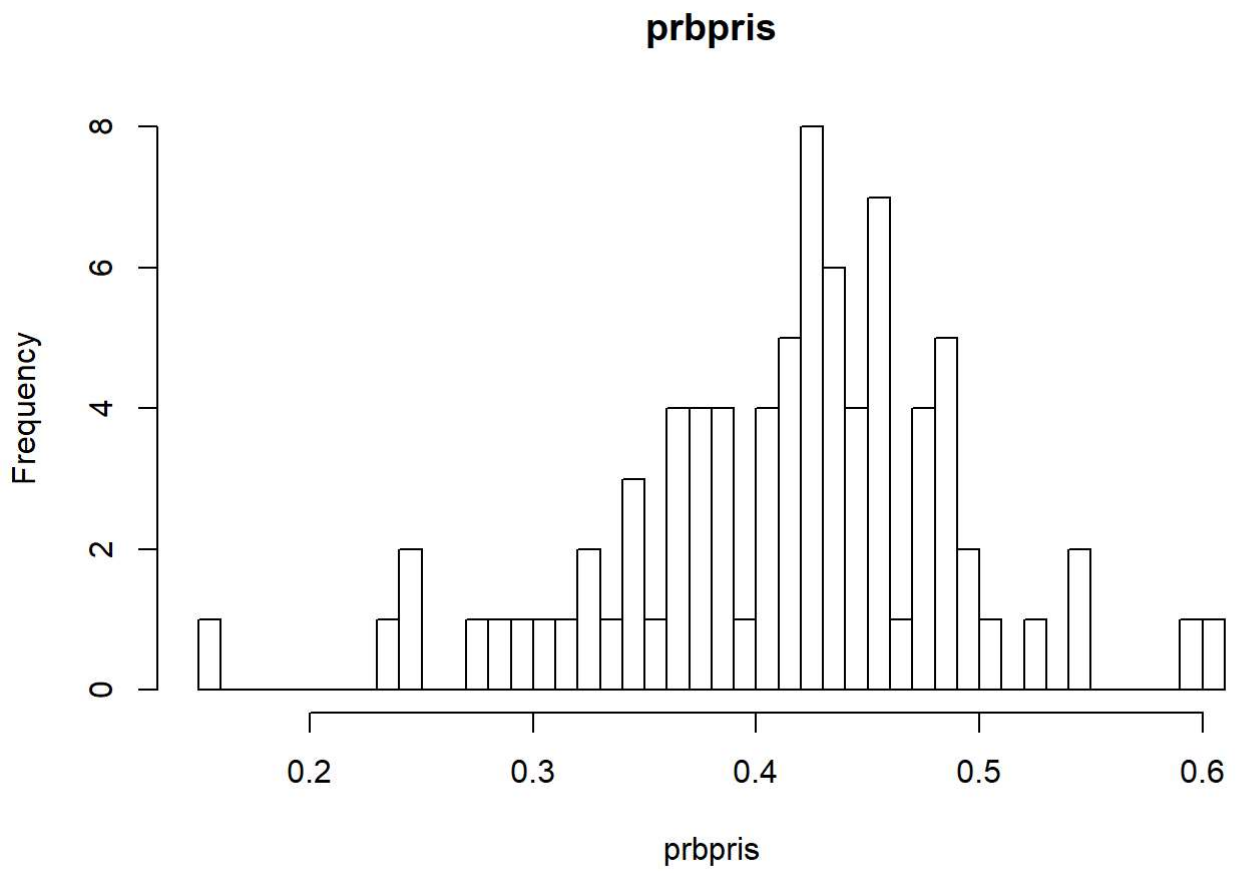
```
# prbarr
hist(data_clean$prbarr,breaks=50,main="prbarr", xlab="prbarr")
```



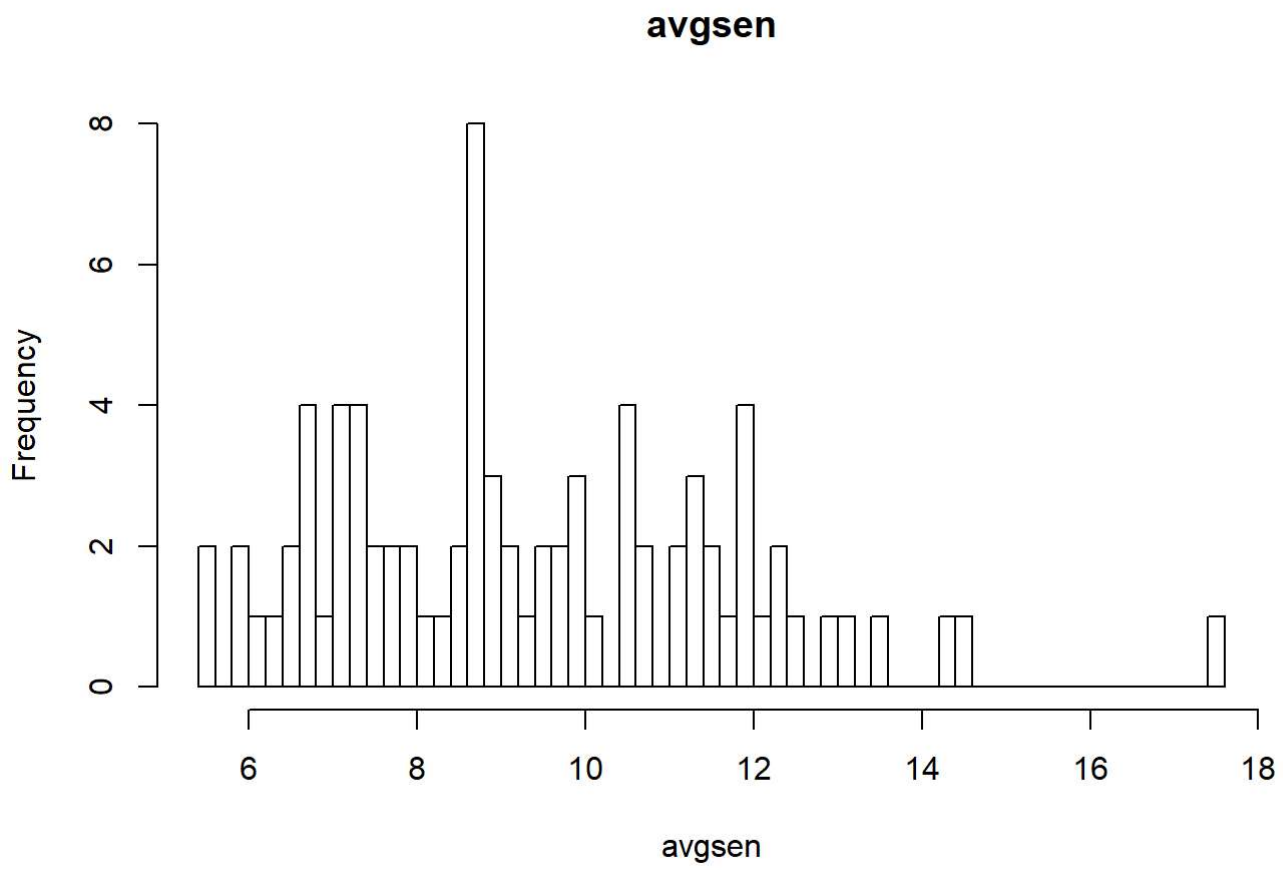

```
#prbconv
hist(data_clean$prbconv,breaks=50,main="prbconv", xlab="prbconv")
```




```
#prbpris
hist(data_clean$prbpris,breaks=50,main="prbpris", xlab="prbpris")
```



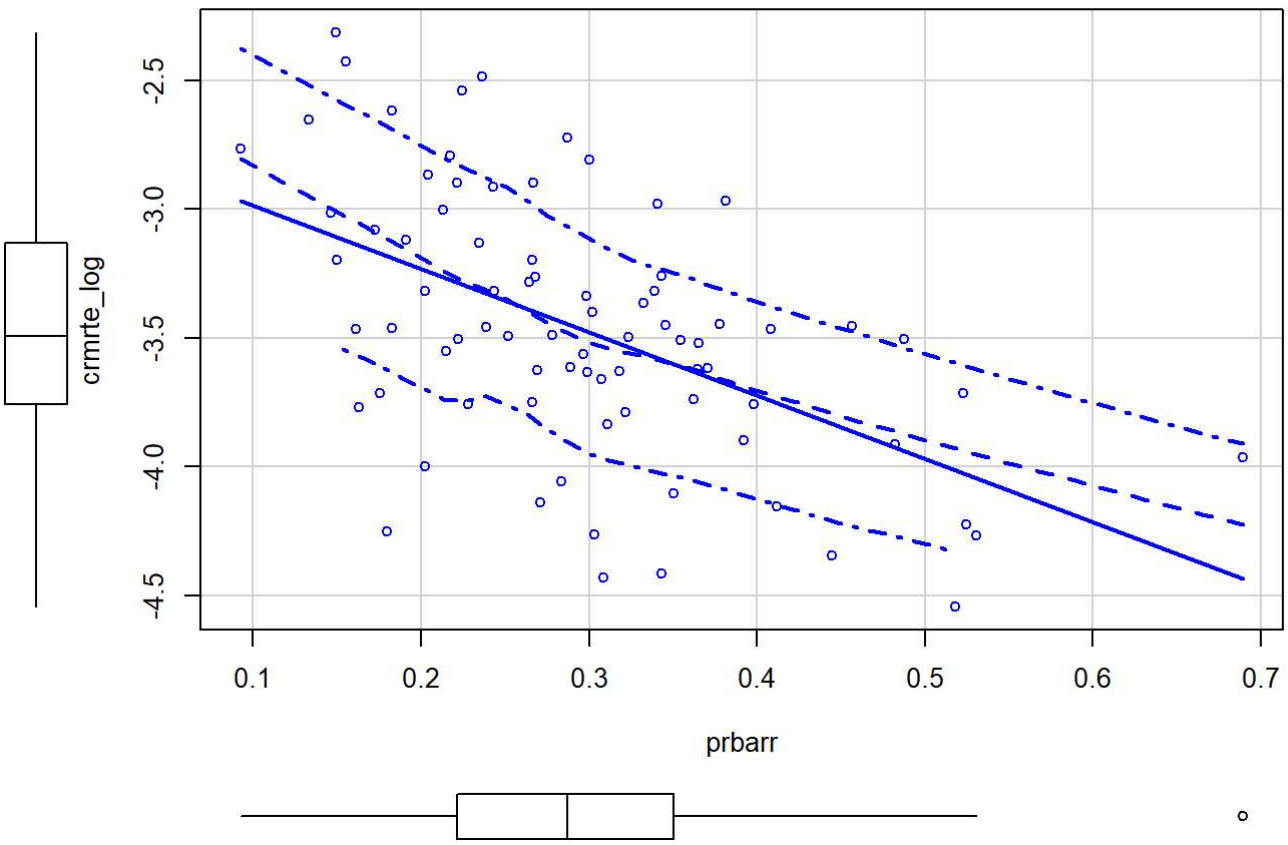

```
#avgsen
hist(data_clean$avgsen,breaks=50,main="avgsen", xlab="avgsen")
```



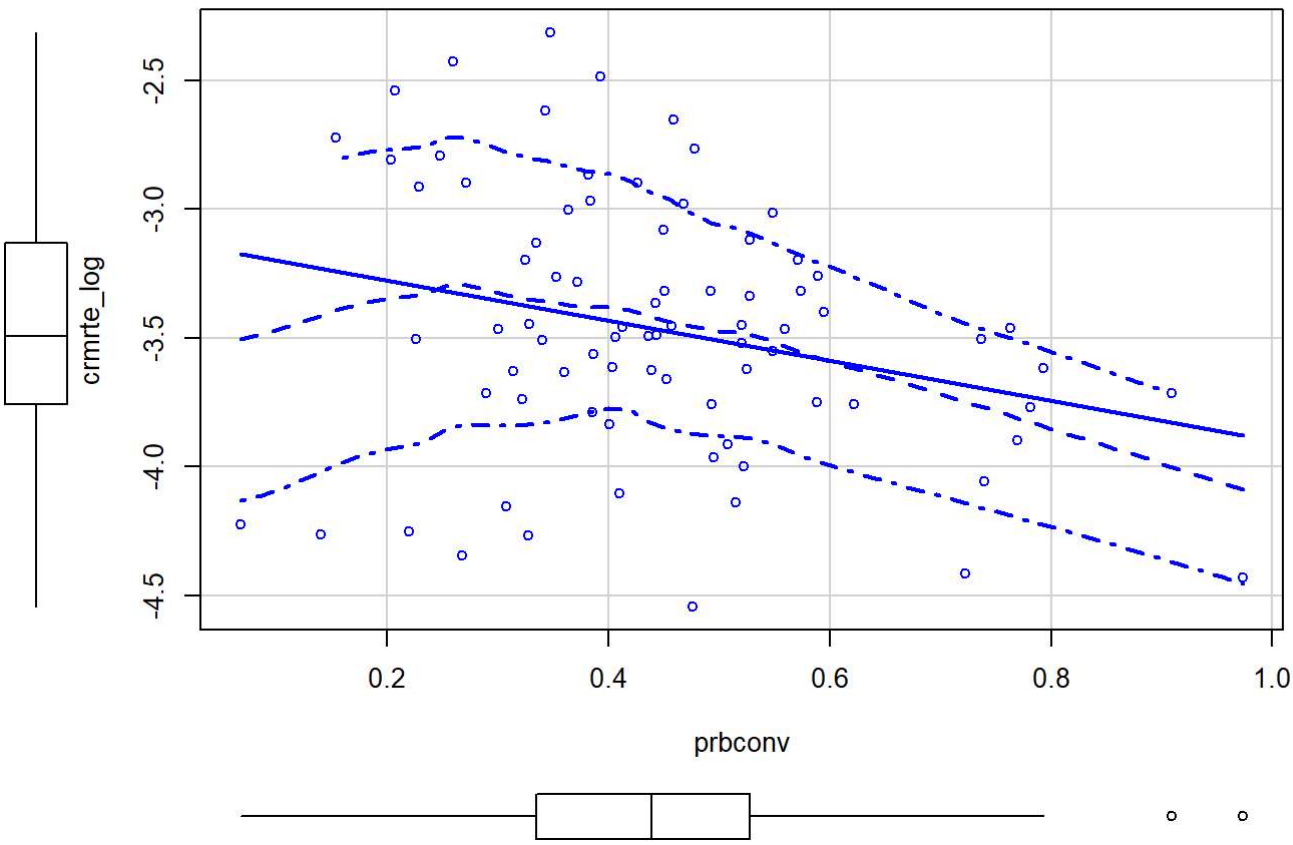
The above histograms of each explanatory variable all seem rather normally distributed, with some anomalies implied in $prbarr$ and $avgsen$. However, we don't think there is enough evidence to make a decision on whether the data is anomaly or not, and we will use the data for model building.

Next, we want to look at the scatterplot between each key explanatory variable and the dependent variable to decide on model specification:

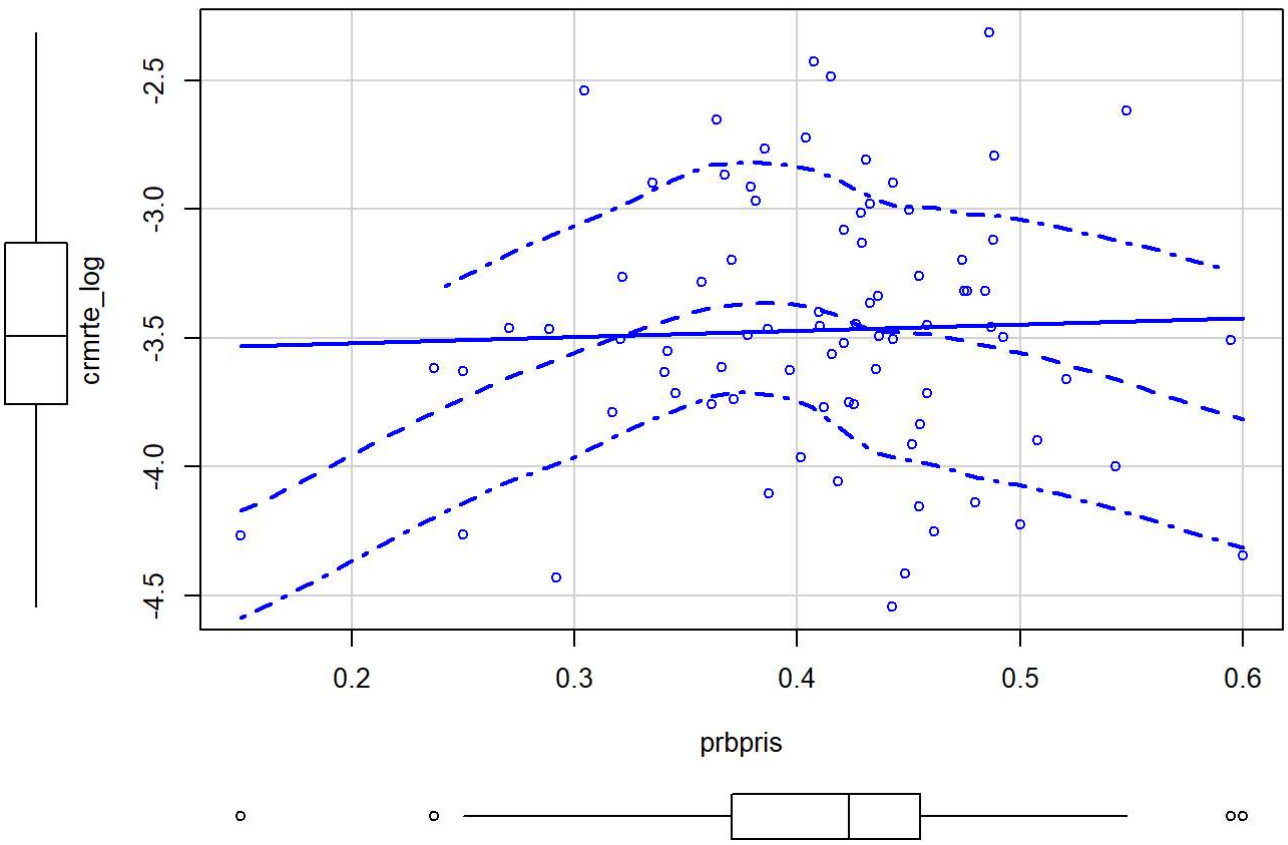
```
scatterplot(data_clean$prbarr,data_clean$crmrte_log, main="prbarr vs. crmrte_log", xlab="prbarr"
, ylab="crmrte_log")
```



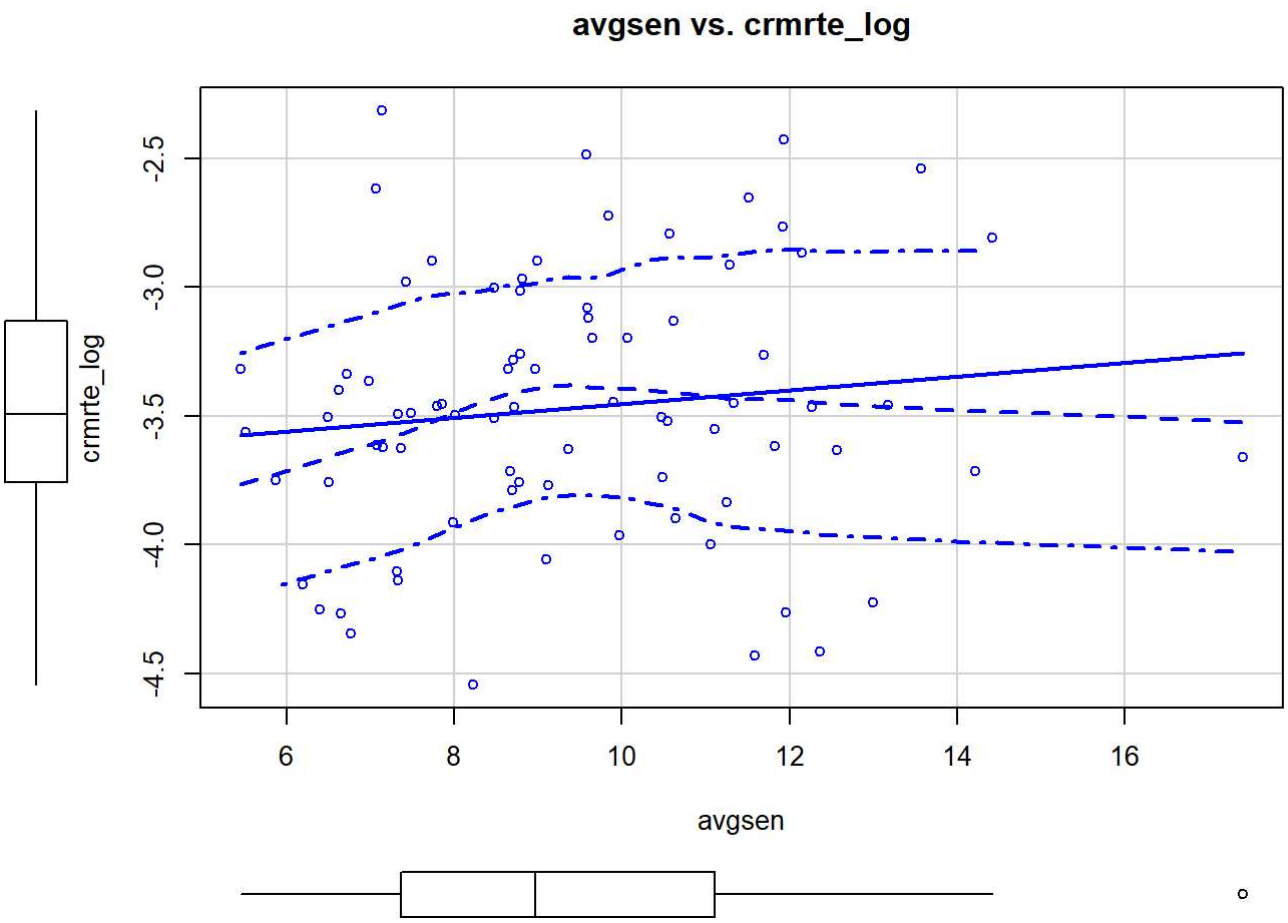

```
scatterplot(data_clean$prbconv,data_clean$crmrte_log, main="prbconv vs. crmrte_log", xlab="prbco
nv", ylab="crmrte_log")
```




```
scatterplot(data_clean$prbpris,data_clean$crmrte_log, main="prbpris vs. crmrte_log", xlab="prbpr
is", ylab="crmrte_log")
```




```
scatterplot(data_clean$avgsen,data_clean$crmrte_log, main="avgsen vs. crmrte_log", xlab="avgsen"
, ylab="crmrte_log")
```



_Fitting model 1__:
With the above explained reason, we are building the model1 with the explanatory variable of key interest as below:
crmrte_log = $_0 + _1 * prbarr + _2 * prbconv + _3 * avgsen + _4 * prbpris + u $

```
model1 <- lm(crmrte_log ~ prbarr + prbconv + avgsen + prbpris, data=data_clean)
model1
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + avgsen + prbpris,
##     data = data_clean)
##
## Coefficients:
## (Intercept)       prbarr      prbconv       avgsen      prbpris
##   -2.384761    -2.626160    -0.961712     0.009033     0.100288
```

Now let's look at whether the model coefficients satisfy the 6 CLM assumptions:

1. linear population model:
   we have not defined the error term, therefore we don't have to talk about the linear population model yet.

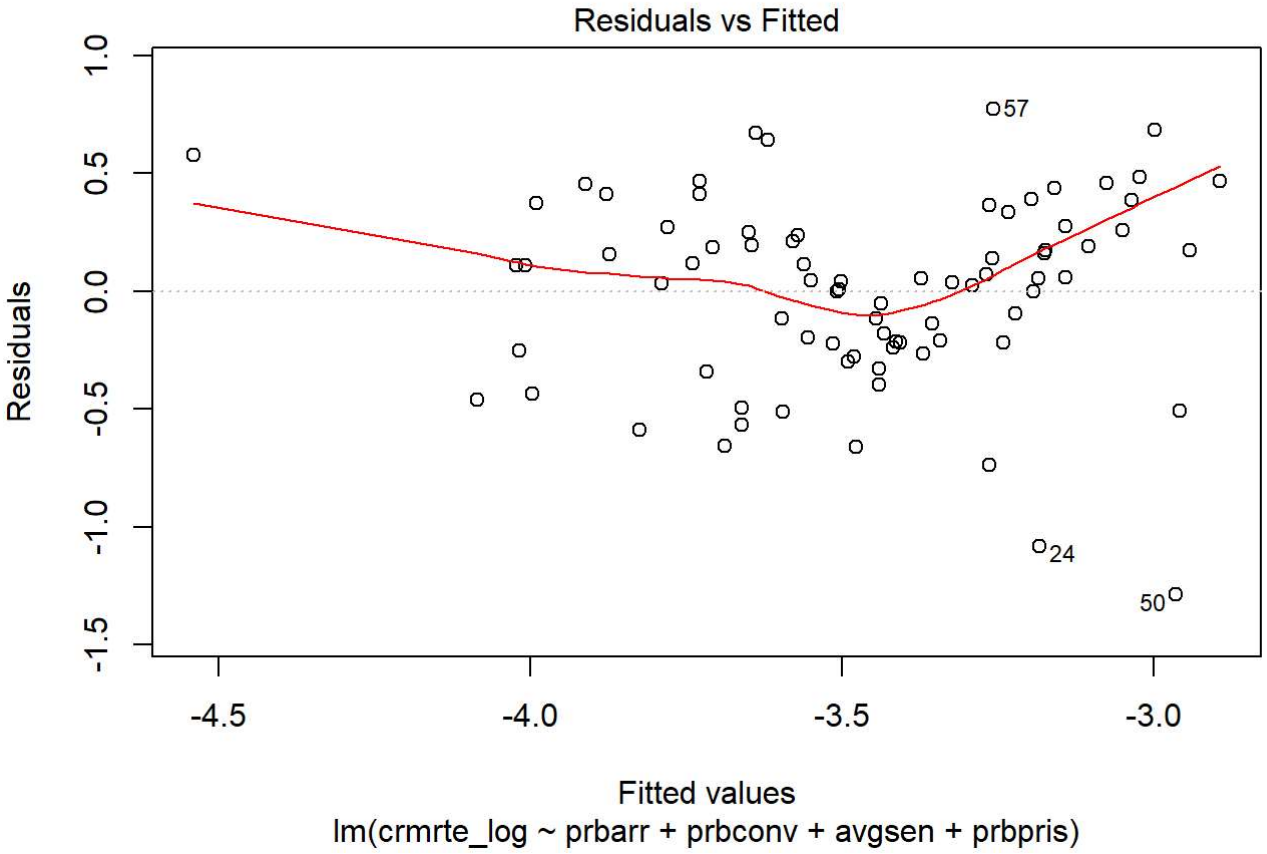
2. Random sampling:
   we do not know how the data is collected. But based on the source of the data, the probability of arrest, conviction and sentence should be random-sampled since they all come from credible sources that only

collect data from actual events Therefore, we should be confident enough to say that the datset have randome sampling.

3. No perfect collinearity
   We are relying on R to alert us when this happens

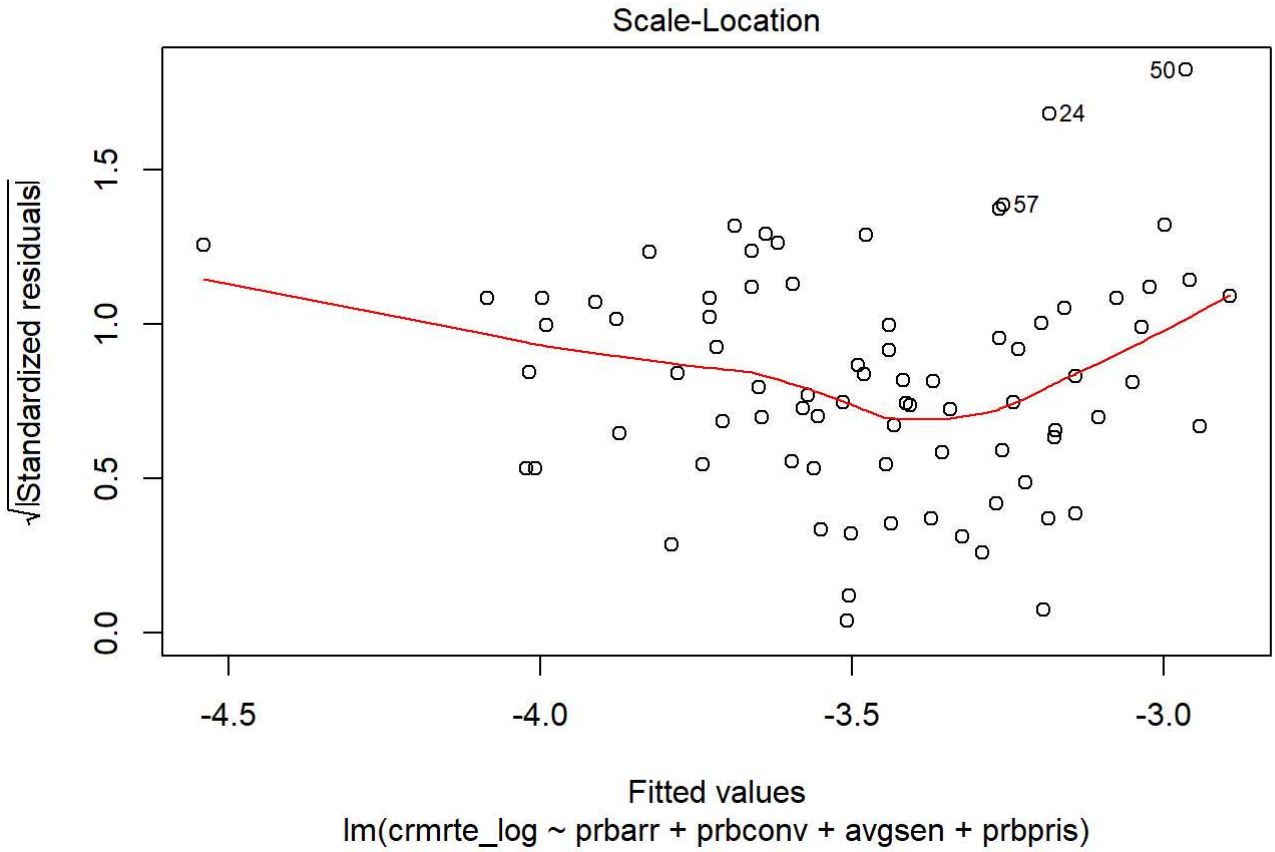
4. Zero-condition mean
   Based on the chart below, it definitely deviates from zero-condition mean. It seems to be rather following a parabolic shape.
   We believe it can be caused by both the face that: a. there are omitted variables b. some parameters might have been misspecified. However, from the scatterplot, we cannot seem to fine strong evidence that the model is misspecified, therefore we are leaning towards the fact that there are other important variables that are omitted from the model

```
plot(model1,which = 1)
```



5. Homoskedasticity
   The below plot is not anywhere close to being flat. It suggests that there are heteroskedasticity in our model, therefore, to look at the errors for the model, we need to rely on the robust standard errors

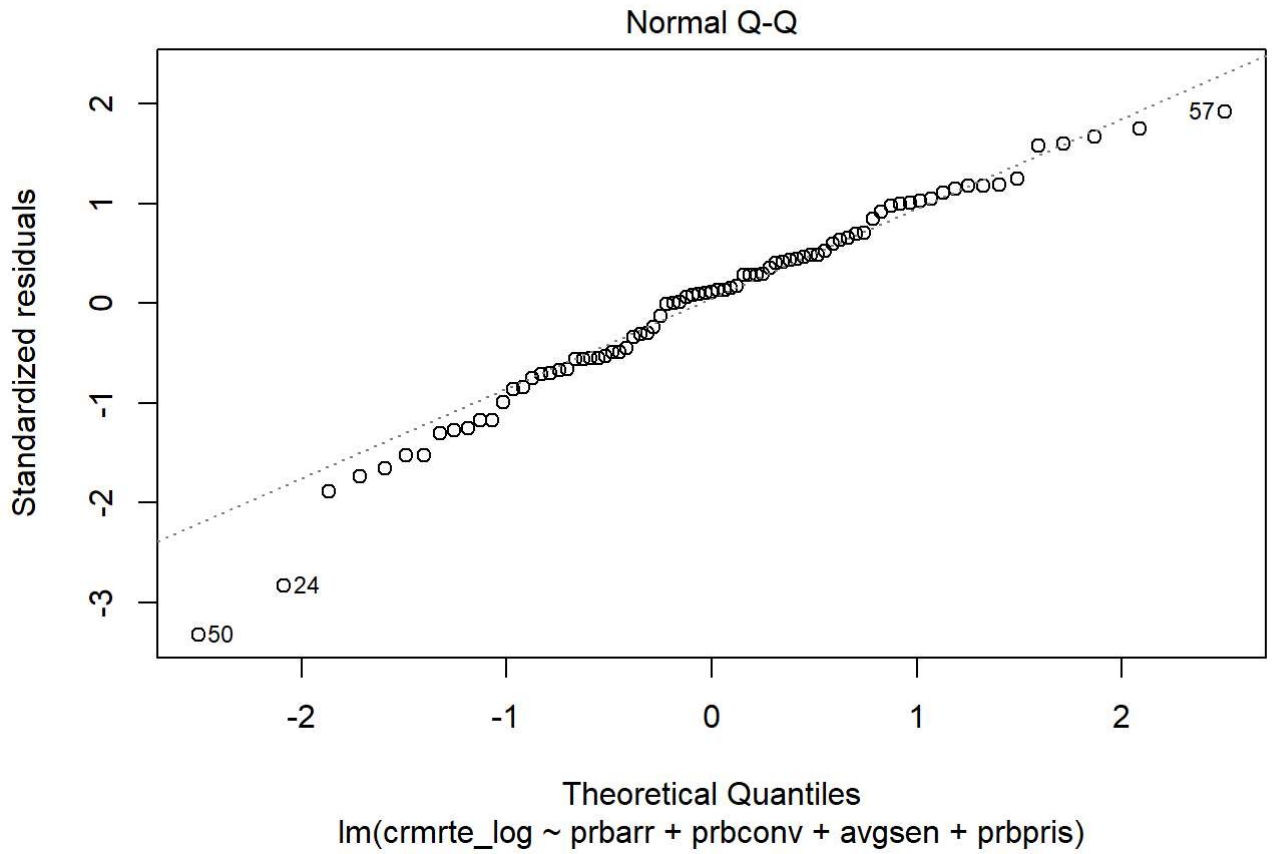
```
plot(model1,which = 3)
```



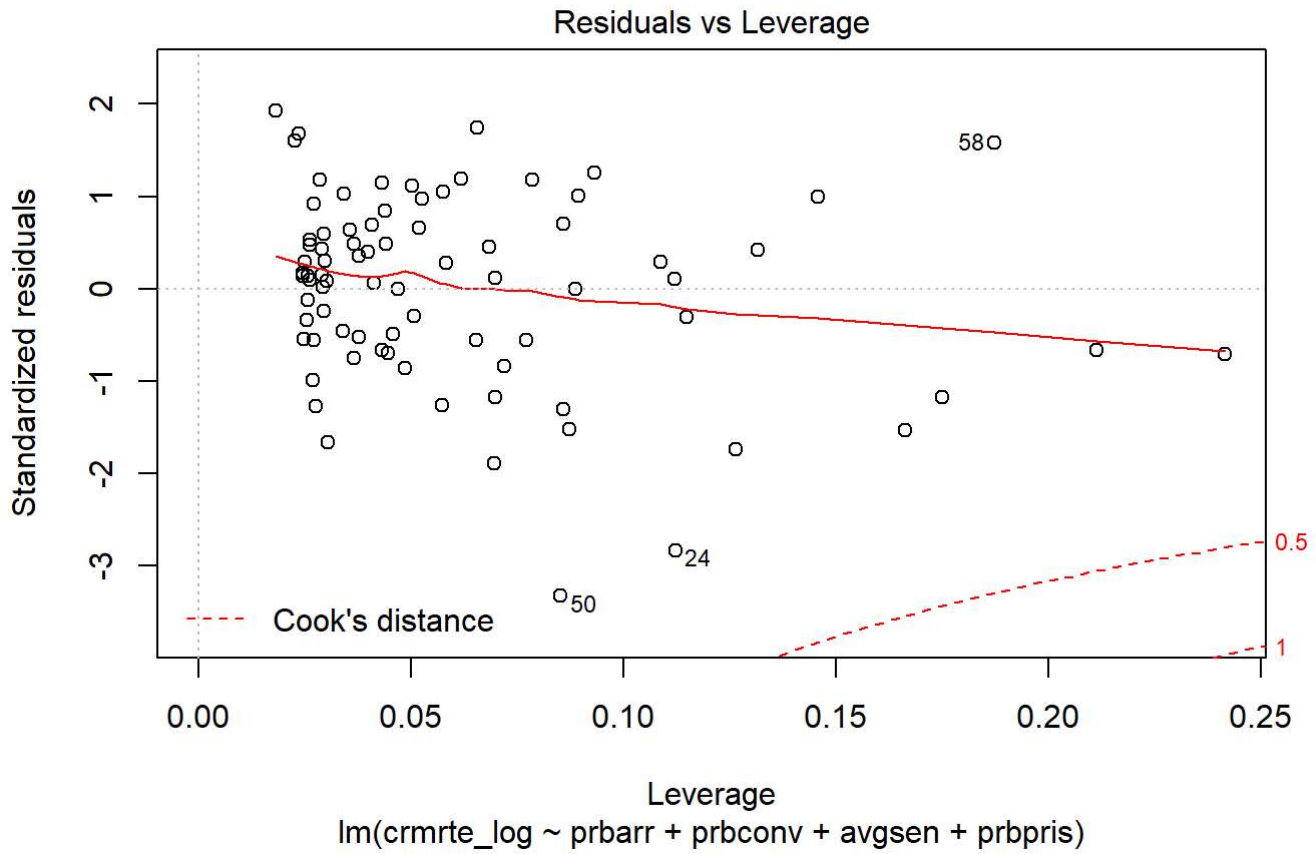
6. Normality of Errors

   The Q-Q plot below suggests that most of the data points follow normal distribution relatively well, with some divergence towards the two ends. Since the two ends have fewer data points and we have close to 100 data points, we can take advantage of central limit theorem.

```
plot(model1,which =2)
```




```
plot(model1,which = 5)
```



The above plot shows that even though there are some points that have high leverage, but since it stays within the 0.5 Cook's distance, there is nothing that we need to worry about.

Next, we are doing a t-test for the coefficients to see whether they have statistical significance

```
coeftest(model1,vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.3847606  0.5030916 -4.7402 9.755e-06 ***
## prbarr      -2.6261600  0.5414970 -4.8498 6.413e-06 ***
## prbconv     -0.9617120  0.3524799 -2.7284  0.007902 **
## avgsen       0.0090325  0.0218154  0.4140  0.680008
## prbpris      0.1002884  0.7078536  0.1417  0.887708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we want to test whether the two explanatory variables that are not statistically significant are jointly significant, we build a model1.1 with just $prbarr$ and $prbconv$

```
model1.1 <- lm(crmrte_log ~ prbarr + prbconv, data=data_clean)
```

To test whether the difference in fit is significant, we use the wald test, which generalizes the usual F-test of overall significance, but allows for a heteroskedasticity-robust covariance matrix

```
waldtest(model1, model1.1, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: crmrte_log ~ prbarr + prbconv + avgsen + prbpris
## Model 2: crmrte_log ~ prbarr + prbconv
##   Res.Df Df      F Pr(>F)
## 1     76
## 2     78 -2 0.1015 0.9036
```

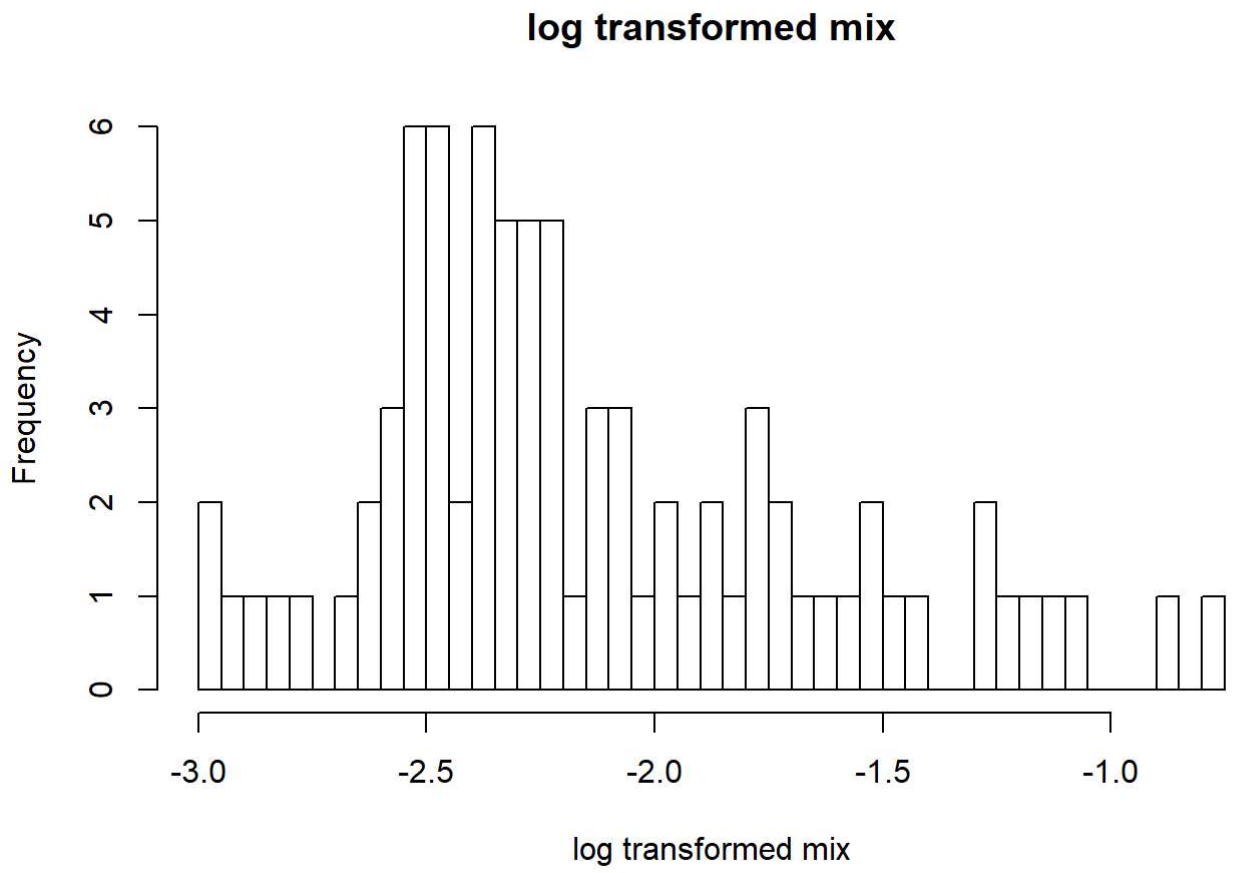
Summary for model 1
1. There are strong evidence that omitted variables exist for model1, and it leading to biases.Therefore, in the second model, we will add more covariates that we believe to be relevant and can help us reduce the omitted variable biases. 2. We need to use robust standard errors when looking at the model standard errors. 3. Only $prbarr$ and $prbconv$ are statistically significant 4. The above results suggest that the variable $avgsen$ and $prbpris$ are not jointly significant.Therefore, we have decided to remove the two variables from our model specification, and thus we are not going to use them for policy suggestions.

Fitting model 2

Based on the analysis from research question section, we do believe that the variable mix has certain influence on crime rate, especially in $prbarr$. The fact that $prbarr$ has strong correlation with mix lead us to think that we should really be focusing on the portion of crime that is non face-to-face, which has relatively low $prbarr$. Therefore, we are adding an extra column that reflects percentage of non face-to-face crime:

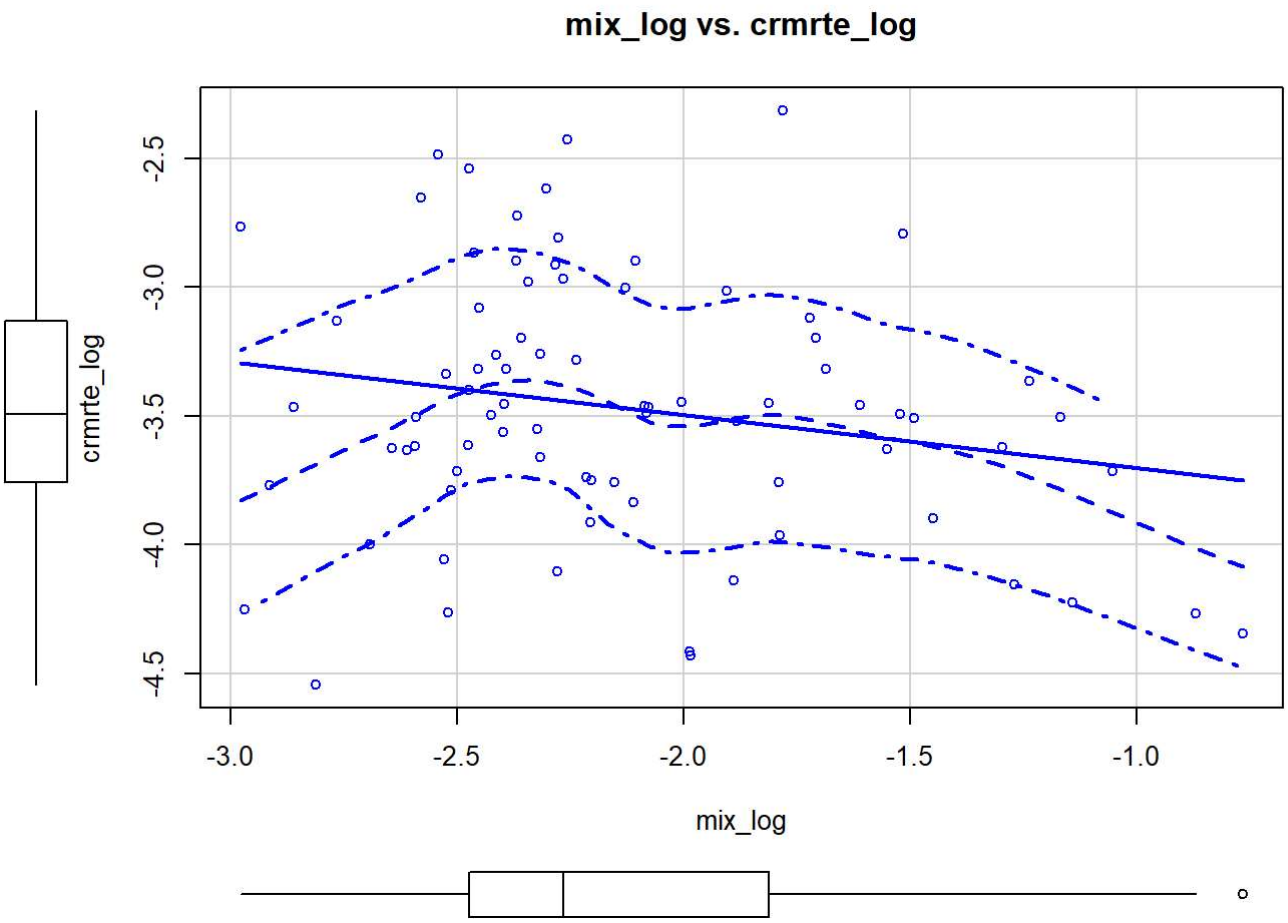
Below are the explanation of what and why we want to introduce those covariates into model2, and what transformation we are applying to them
The original histogram of mix does not seem very normally distributed. Since it is left skewed, we believe that log transformation is helpful:

```
hist(log(data_clean$mix),breaks=50,main="log transformed mix", xlab="log transformed mix")
```



We will create a new variable that has the log transformed mix, and name it mix_log.

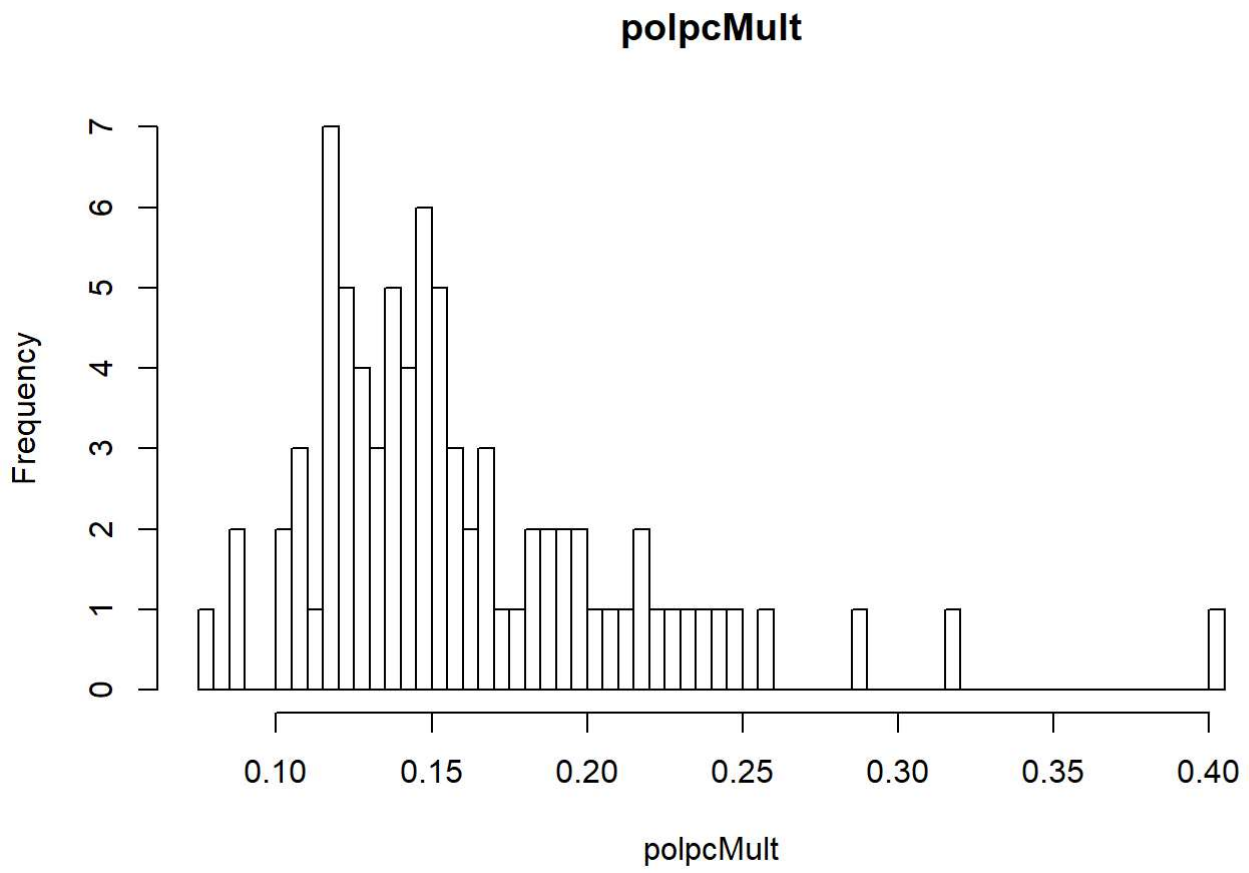
```
data_clean$mix_log <- log(data_clean$mix)
scatterplot(data_clean$mix_log,data_clean$crmrte_log, main="mix_log vs. crmrte_log", xlab="mix_l
og", ylab="crmrte_log")
```



1. mix: The above two scatterplots show enough evidence that variable $nonFtF$ is an important covariate. Logically, we believe that $nonFtF$ should directly impact $prbarr$ since the non face-to-face crime tend to be less severe, and there are not as much evidence that can lead to arrest. With the probability of arrest being low for this kind of crime, people are willing to take the risk to commit those crimes. Reducing the non face-to-face crime is very important in reducing the ocerall $crmrte$. Therefore, we believe that $nonFtF$ is crucial in model building and it will help make our model more accurate and less biased.

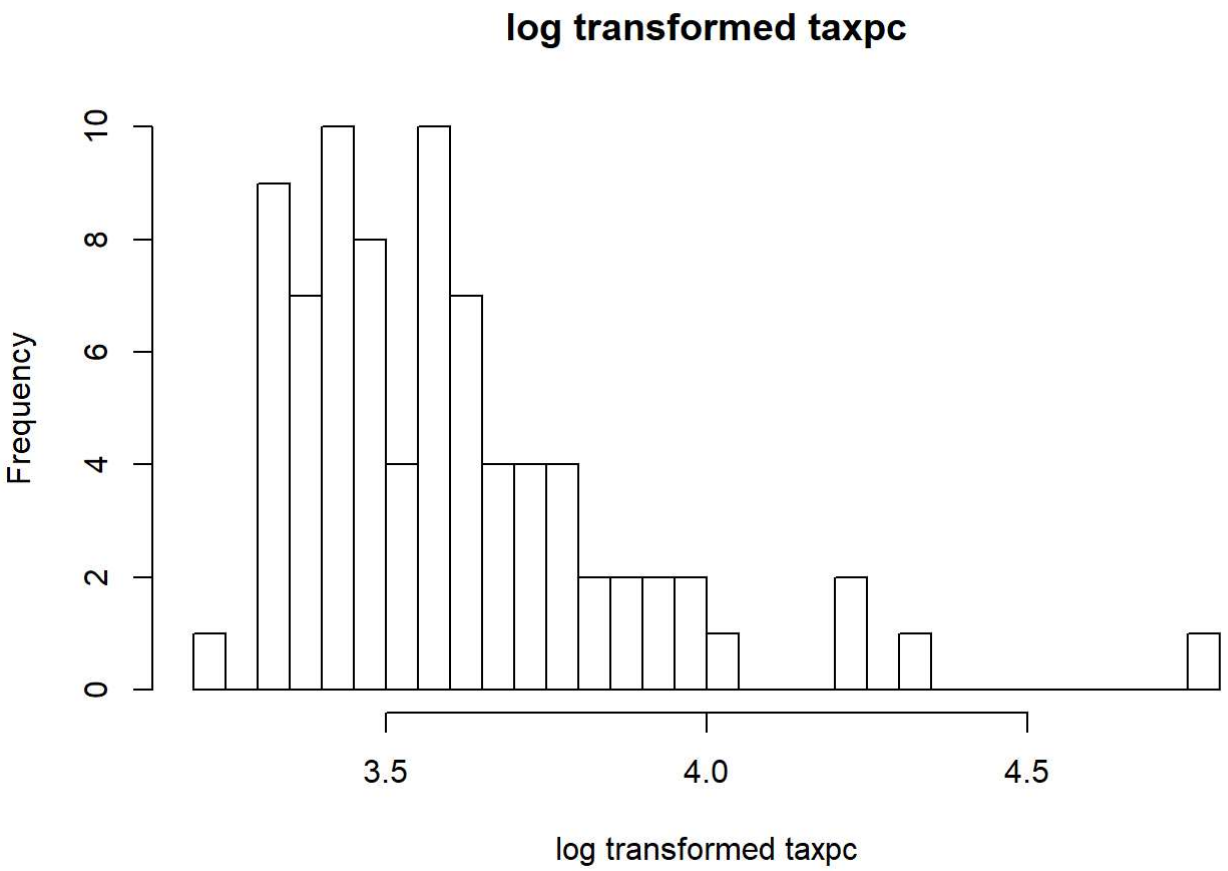
2. polpc: we believe that the police per capita is going to help with our model accuracy. The density or distribution of police forces is very likely to affect the effectiveness of the arrest when crime happens. The magnitude of the $polpc$ is at least 100 times smaller than the other probability variables, and we believe that it might add some imbalance in our model building process, therefore we decided to create another variable that is 100 times of polpc:

```
data_clean$polpcMult <- data_clean$polpc*100
hist(data_clean$polpcMult,breaks=50,main="polpcMult", xlab="polpcMult")
```



3. taxpc: Related to police density and effectiveness of arrest, we believe that the tax income of the local government will affect how much resources they have to implement systematic measures to reduce crime rate. Therefore, we consider $taxpc$ as an important covariate in model2

```
hist(log(data_clean$taxpc),breaks=50,main="log transformed taxpc", xlab="log transformed taxpc")
```

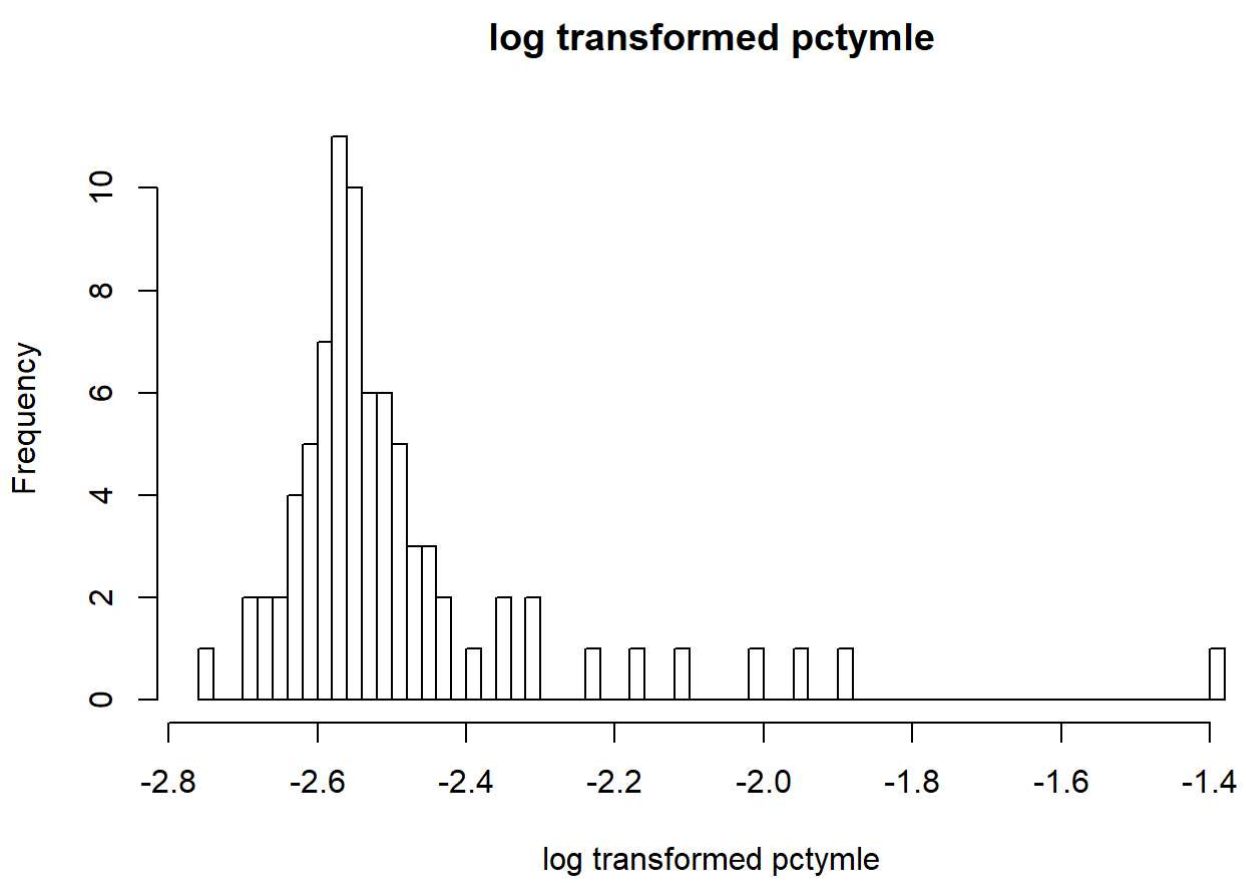


The log transformed taxpc seems to be more normal compared to the original dataset. We will create a new variable with log transformed $taxpc$, and name it $taxpc_log$.

```
data_clean$taxpc_log <- log(data_clean$taxpc)
```

4. urban: we will use $urban$ as indicator variable to see how we compare the data between areas outside the city and within the city
5. pctymle: The percentage of young male seems to have rather big correlations with crime rate. Therefore, we will add this covariate into model2. We want to see whether we can draw some conclusions on young males, maybe increasing education opportunities for young males so that they acquire more skills and knowledge to find a job.
   According to the distribution below, there seems to be one point that has much higher percentage of young males. However, we don't have enough evidence to see whether it is an error in the data or an outlier.

```
hist(log(data_clean$pctymle), breaks=50, main="log transformed pctymle", xlab="log transformed p
ctymle")
```



we will use the log transformed version of the data, and create a new variable $pctymle_log$

```
data_clean$pctymle_log <- log(data_clean$pctymle)
```

Fitting model 2:
With the second model, we are building it on top of model 1 but adding more variables that can have influence on the dependent variable crmte. After the data transformation, the model is as below:
crmrte_log = $_0 + _1 * prbarr + _2 * prbconv + _3 * avgsen + _4 * prbpris + _5 * mix_log + _6 * taxpc_log + _7 * urban + _8 * pctymle_log + _9 * polpcMult + u $

```
model2 <- lm(crmrte_log ~ prbarr + prbconv + avgsen + prbpris + mix_log + taxpc_log + pctymle_lo
g + polpcMult, data=data_clean)

model2
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + avgsen + prbpris +
##     mix_log + taxpc_log + pctymle_log + polpcMult, data = data_clean)
##
## Coefficients:
## (Intercept)       prbarr      prbconv       avgsen      prbpris
##   -3.500277    -2.098403    -0.276904    -0.005076     0.673577
##     mix_log    taxpc_log  pctymle_log    polpcMult
##    0.013473     0.315053     0.406114     2.830542
```

Based on the model, we notice that nonFtf has a siginificant positive effect on crmrte_log. That means the higher the non face to face offense ratio, the higher the crmrte_log. This might be interprated as, non face to face offense tends to be more violent and dangerous than face to face offend, thus get reported more. Taxpc has almost none effect on crmrte. Urban has a relativly siginificant.

We will look again at the residuals and standard errors of the new model, and how is the model fitting satisfy the CLM assumptions.

Next, we are interested in knowing how is model2 fitting into the classical linear model assumptions:
1. model2 has a slightly better residual vs. fitted value plot at the red curve is closer to 0 and also more straight. It indicates that we have smaller biases by introducing more variables into the specification
2. There is one point in the leverage plot which is almost on the Crook's line, but it is still within the acceptable range, and we will leave the point in our analysis. Otherwise there are no other warning signs that we see for model2.

Next, we are doing a t-test for the coefficients to see whether they have statistical significance

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

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -3.5002773  1.0013886 -3.4954 0.0008143 ***
## prbarr      -2.0984027  0.5982449 -3.5076 0.0007832 ***
## prbconv     -0.2769044  0.4370999 -0.6335 0.5284121
## avgsen      -0.0050755  0.0240975 -0.2106 0.8337753
## prbpris      0.6735766  0.9711675  0.6936 0.4901809
## mix_log      0.0134730  0.1364366  0.0987 0.9216116
## taxpc_log    0.3150530  0.2872738  1.0967 0.2764273
## pctymle_log  0.4061144  0.2696352  1.5062 0.1364005
## polpcMult    2.8305418  1.9313685  1.4656 0.1471210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary for model 2
1. $Prbarr$ remains to be statistically significant, but $prbconv$ has a lot of standard error introduced by adding more variables in, meaning that $prbconv$ is not the parameter to add in for a robust predictino model.

2. The addition of the other variables helped us in making the model more accurate. The adjusted r square is higher meaning that the new variables added to the model accuracy and the impact is more than by chance.

```
paste("adjusted R square for model1 is:",summary(model1)$adj.r.squared,"; ", "adjusted R square
 for model2 is:",summary(model2)$adj.r.squared)
```

```
## [1] "adjusted R square for model1 is: 0.358068104042064 ;  adjusted R square for model2 is:
0.46952892519363"
```

3. We want to test whether mix_log, $taxpc_log$, $pctymle_log$ and $polpcMult$ are jointly significant, meaning that whether their coefficients are all zero.

```
linearHypothesis(model2, c("mix_log = 0", "taxpc_log = 0", "pctymle_log = 0", "polpcMult = 0"),
 vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## mix_log = 0
## taxpc_log = 0
## pctymle_log = 0
## polpcMult = 0
##
## Model 1: restricted model
## Model 2: crmrte_log ~ prbarr + prbconv + avgsen + prbpris + mix_log +
##     taxpc_log + pctymle_log + polpcMult
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1     76
## 2     72  4 3.5942 0.00996 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above tell us that the the 4 variables are jointly significant. There is a probability that there is multicollinearity among those four variables and which explains why none of them is individually significant.

Fitting model 3

Bringing in other covariates that we do not believe to be too relavent:

```
model3 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen+polpcMult+density+taxpc_log+west+central
+urban+pctmin80+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+mix_log+pctymle_log, data=data_clea
n)
model3
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + prbpris + avgsen +
##     polpcMult + density + taxpc_log + west + central + urban +
##     pctmin80 + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
##     wsta + wloc + mix_log + pctymle_log, data = data_clean)
##
## Coefficients:
## (Intercept)       prbarr      prbconv      prbpris       avgsen
##  -3.1274148   -1.7126251   -0.2565360    0.0310823   -0.0215137
##   polpcMult      density    taxpc_log         west      central
##   2.8752873    0.1201781    0.1119469   -0.1908106   -0.1408822
##       urban     pctmin80         wcon         wtuc         wtrd
##  -0.0555075    0.0087980    0.0004136    0.0003945    0.0010412
##        wfir         wser         wmfg         wfed         wsta
##  -0.0024401   -0.0015116   -0.0002204    0.0020816   -0.0020028
##        wloc      mix_log  pctymle_log
##   0.0021864   -0.1047079    0.4275211
```

Next, we are interested in knowing how is model3 fitting into the classical linear model assumptions:
1. model3 has a better residual vs. fitted value plot at the red curve is closer to 0 and also more straight. It indicates that we have smaller biases by introducing more variables into the specification
2. There are no other warning signs that we see for model3, except that the introduction of other varialbes have impacted the normality. However, we believe that 81 sample size should be enough of a sample size to not have normality affect our analysis.

Next we want to check whether all variables related to wage are jointly significant. We built another model with all other variables that are not wage related, and want to use the wald test to check for statistical significance.

```
model4 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen+polpcMult+density+taxpc_log+west+central
+urban+pctmin80+mix_log+pctymle_log, data=data_clean)
waldtest(model3, model4, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpcMult +
##     density + taxpc_log + west + central + urban + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     mix_log + pctymle_log
## Model 2: crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpcMult +
##     density + taxpc_log + west + central + urban + pctmin80 +
##     mix_log + pctymle_log
##   Res.Df Df      F   Pr(>F)
## 1     58
## 2     67 -9 3.3034 0.002579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the above result tells us that those variables are jointly significant, and there is probability a great amount of multicollinearity that caused most of them to be none-statistically-significant. Also, wage factor and economics are widely studied factors related to crime, which is out of the scope of our research question. We just want to make

sure that omitting wage factors does not introduce significant amount of biase.

Next we want to use Stargzer to directly compare the 4 models that we builts:

output: pdf_document

==========

Dependent variable:

——— crmrte_log

(1) (2) (3) (4)

——— prbarr -2.626*** -2.098*** -1.713*** -1.583***

(0.423) (0.471) (0.304) (0.357)

prbconv -0.962*** -0.277 -0.257 -0.361
(0.268) (0.290) (0.188) (0.207)

avgsen 0.009 -0.005 -0.022 -0.015
(0.020) (0.018) (0.012) (0.013)

prbpris 0.100 0.674 0.031 0.250
(0.589) (0.566) (0.359) (0.406)

mix_log 0.013 -0.105 -0.113
(0.101) (0.068) (0.073)

taxpc_log 0.315 0.112 -0.037
(0.202) (0.153) (0.159)

west -0.191 -0.282*
(0.113) (0.125)

central -0.141 -0.163
(0.076) (0.084)

urban -0.056 -0.239
(0.165) (0.185)

pctmin80 0.009** 0.008**
(0.003) (0.003)

wcon 0.0004
(0.001)

wtuc 0.0004
(0.0004)

wtrd 0.001
(0.001)

wfir -0.002**
(0.001)

wser -0.002
(0.001)

wmfg -0.0002
(0.0004)

wfed 0.002**
(0.001)

wsta -0.002**
(0.001)

wloc 0.002
(0.001)

pctymle_log 0.406 0.428* 0.100
(0.240) (0.166) (0.176)

polpcMult 2.831** 2.875*** 3.190***
(1.031) (0.704) (0.763)

density 0.120** 0.162***
(0.036) (0.037)

Constant -2.385*** -3.500*** -3.127*** -3.471***
(0.402) (0.867) (0.675) (0.651)

Observations 81 81 81 81
R2 0.390 0.523 0.867 0.784
Adjusted R2 0.358 0.470 0.816 0.742
Residual Std. Error 0.405 (df = 76) 0.368 (df = 72) 0.217 (df = 58) 0.257 (df = 67)
==========
Note: *p<0.05;* ***p<0.01;*** p<0.001

What does the above table tell us:

1. model3 really help us to identify the most cost-effective measures that we should suggest in the policy to reduce crime rate. In the above table comparison, it is easy to see that the wage related variables that we intend to drop off from the analysis due to our research question, do have strong statistical significance yet the coefficients are small enough that we would not consider them to be the most important factors in our analysis. The addition of those variables does increase the adjusted R squre and the residuals plot shows more consistency with the CLM assumptions. However, the strong statistical significance + small coefficients serve as confirmation that those variables have small (close to zero) impact on the crime rate, which is not what we are trying to find in this research.
2. $prbarr$ is one of the most relevant factor in reducing crime rate. This is re-affirmed by comparing between different models. It is highly significant and the coefficient is high enough for us to consider it as an cost-effective measures to suggest policies for reducing crime rate
3. $polpcMult$ police per capita has high statistical significance, but we believe it does not show causality, or at least it does not have practical significance for us to suggest policies. With the background knowledge, we know that the police density will increase due to the high crime rate in a certain area, instead of the other way around.

Based on the comparison above, we want to see if we can have another model with the variables that showed statistical significance and at the same time are relevant to our research question. We are focusing on the measures that are most cost-effective, and therefore, we want to find those explanatory variables that have statistical significance while have larger slopes.

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + density + polpcMult + pctmin80,
##     data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67676 -0.18074  0.02077  0.18124  0.56525
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.921643   0.144419 -27.155  < 2e-16 ***
## prbarr      -1.914598   0.306928  -6.238 2.30e-08 ***
## density      0.121121   0.022371   5.414 6.93e-07 ***
## polpcMult    3.292720   0.617821   5.330 9.74e-07 ***
## pctmin80     0.011936   0.001842   6.479 8.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2693 on 76 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7163
## F-statistic:  51.5 on 4 and 76 DF,  p-value: < 2.2e-16
```

Based on the above analysis, we are achieving relatively high adjusted R-squared value with 4 key variables, which all have statistical significance. Based on the analysis of 6 classical linear model assumptions, we definitely see improvement in terms of zero-condition mean and normality. However, we believe that there should still be omitted variable biases due to some other key aspects missing from the dataset:

Omitted Variable

1. **education level**: One aspect of education level that is measurable would be the years of education $year_educ$.
   The years of education could result in the biase over how much impact the percentage of young males $pctymle$ is correlated with $crmrte_log$. With the addition of education level, $pctymle$ should really become vague as a explanatory variable. In its place, there should be most likely $pctymle_lowEduc$ (percentage of low-education (<6 years) male) that is really more correlated. The bias from this omiitted variabled should be towards zero.

2. **surveillance camera**: Number of surveillance cameras $surCams$ has some impact on the prbability of arrest and the crime rate: Leaving other variables unchanged, we assume that: $crmrte_log = _0 + _1 * prbarr + _2 * surCams + u $ $surCams = _0 + _1 * prbarr + v $

Based on our understanding, γ_1 should be positive, and β_2 should be negative. β_1 is negative. Therefore, the OLS coefficient of $prbarr$ will be scaled away to be more negative, gaining statistical significance.

3. **economic growth**: The annual growth in economic $ecoGrowth$ will lead to more job availability and increase in tax income, and will give more funding to the government to hire police forces.

$crmrte_log = _0 + _1 * taxpc + _2 * ecoGrowth + u $ $ecoGrowth = _0 + _1 * taxpc + v $ Based on our understanding, γ_1 should be positive, if β_2 is positive, and β_1 is positive Therefore, the OLS coefficient of $prbarr$ will be scaled away to be more positive, gaining statistical significance.

4. **arrest rate for minor crimes**: the rate or probability of arrest for minor crimes (vandalism, public drinking, etc.) $prbarr_minor$ is the biggest percentage of overall crime rate, and it should be directly related to $prbarr$ and $crmrte_log$.

$crmrte_log = _0 + _1 * prbarr + _2 * prbarr_minor + u $ $prbarr_minor = _0 + _1 * prbarr + v $ Based on our understanding, γ_1 should be positive, indicating that actually reducing overall crime rate will usually reduce crime rate for minor crime, and vice versa; β_2 should be negative so OMVB=_2_1<0, and β_1 is negative Therefore, the OLS coefficient of $prbarr$ will be scaled away to be more negative, gaining statistical significance.

5. **average age**: the average age is affecting the crime rate because it is usually the group of people under 30 that commit more crime than people over 30: $crmrte_log = _0 + _1 * pctymle + _2 * avgAge + u $ $avgAge = _0 + _1 * pctymle + v $ γ_1 should be negative, and β_2 should be negative so OMVB=_2_1>0, and β_1 is positive Therefore, the OLS coefficient of $pctymle$ will be scaled away to be more positive, gaining statistical significance.

6. **income inequality**: income alone is usually not enough to tell how it can be related to crime rate, we always need extra information like income growth and income inequality to analyze the social stability of the society. It is hard to say exactly what coefficients income inequality is going to affect, but it should be related to tax income, population density, what type of crime that happens, etc.

Suggested Statistical Test:

1. Heavier penalties and arrest rate for minor crimes: the analysis suggest that less severe crimes consist of the most crimes commited, and the arrest rate is low for this type of crime. One good test in this regard is to lift the penalties and arrest rate for minor crimes for about 2 weeks, and see if the overall crime rate over those two weeks go down.

Final conclusion:

Based on our models, we think that to reduce crime rate in general, there are severl approaches and can eventaully evolve into policies:

1. Increase the probability of arrest on non face to face crime. In order to achieve that, we need to furthur categorize non face to face crime and come up with strategies that can deal with each caregory. For example, most non face-to-face crimes are hard to identify and there are fewer evidences left. In order to catch more of those crimes in actions and with evidences, one policy that we will suggest is to install more surveillance cameras, especially in the areas where non face-to-faces crimes are most reported (downtowns, subways, malls, etc.)

2. If the above suggested statistical test has a positive result, we might need to consider increase penalties for minor crimes. Penalties for minor crime do not necessarily mean longer prison time. It could be in other forms like social services.

3. Above are two policies on the measures we can take directly related to crime and arrest. However, in most of our omitted variables, if we have enough information about those variabels, there is a possiblity that we can draw conclusion from other aspects, for example, economically how to reduce income inequality. Collect

more data on the omitted variables. We need further investigation after collecting information on the omitted variables. The models we have in this report all have their biases, so with more information, we can refine our models thus uncover the other key measures to reduce crime rate.