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. We will then purpose several policies that can potentially 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
```

```
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,11)
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 87    191   87 0.0458895 0.172257 0.450000 0.421053   9.59 0.00122733
## 88    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
## 90    195   87 0.0313973 0.201397 1.670520 0.470588  13.02 0.00445923
## 91    197   87 0.0141928 0.207595 1.182930 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
## 87 1.7725632 32.74533    0       0     0 34.42830 318.0599 400.8570
## 88 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
## 90 1.7459893 53.66693    0       0     0 37.43110 315.1641 377.9356
## 91 0.8898810 25.95258    1       0     0  5.46081 314.1660 341.8803
## 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
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 87 230.9888 320.0345 238.4958 295.26 334.55 375.45 327.62 0.08616445
## 88 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.11050157
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.11050157
## 90 246.0614 411.4330 296.8684 392.27 480.79 303.11 337.28 0.15612382
## 91 182.8020 348.1432 212.8205 322.92 391.72 385.65 306.85 0.06756757
## 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
##       pctymle
## 87 0.08828809
## 88 0.07819394
## 89 0.07819394
## 90 0.07945071
## 91 0.07419893
## 92         NA
## 93         NA
## 94         NA
## 95         NA
## 96         NA
## 97         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)
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
##     prbconv           prbpris           avgsen           polpc
##  Min.   :0.06838   Min.   :0.1500   Min.   : 5.380   Min.   :0.0007459
##  1st Qu.:0.34541   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.0012308
##  Median :0.45283   Median :0.4234   Median : 9.100   Median :0.0014853
##  Mean   :0.55128   Mean   :0.4108   Mean   : 9.647   Mean   :0.0017022
##  3rd Qu.:0.58886   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.0018768
##  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.:0.54741   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
##      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
##       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
##       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.:0.08073
##  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
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
```

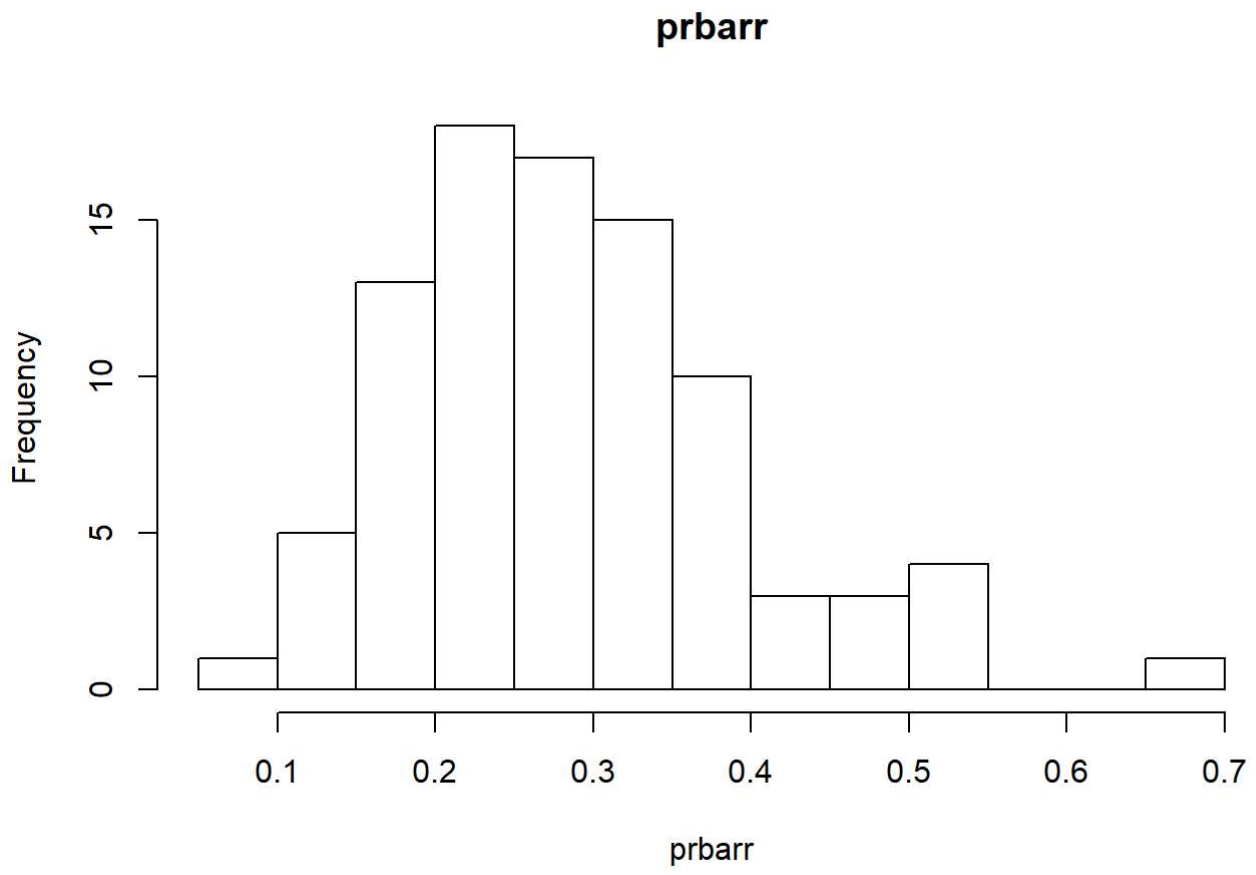
2. take out year column as it does not mean anything in our analysis

```
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")
```

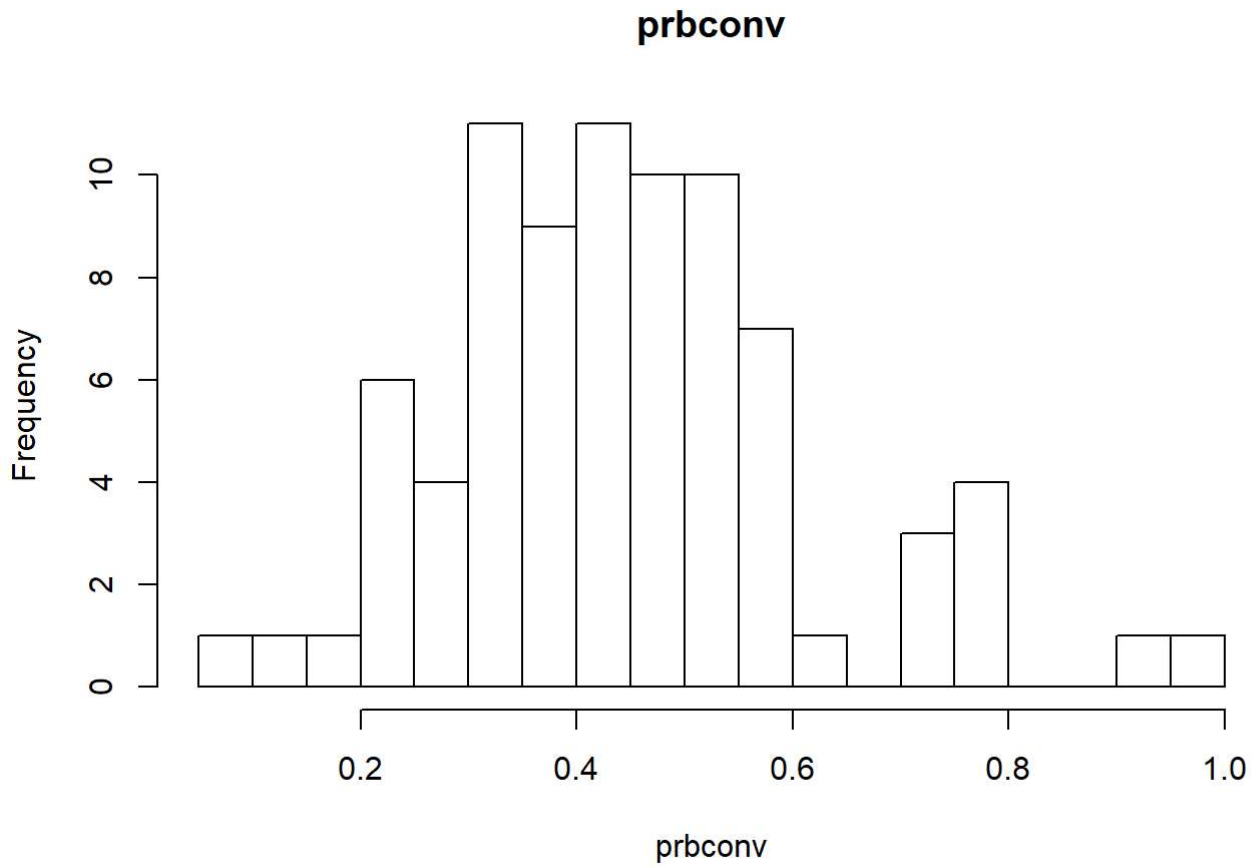


4.

“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

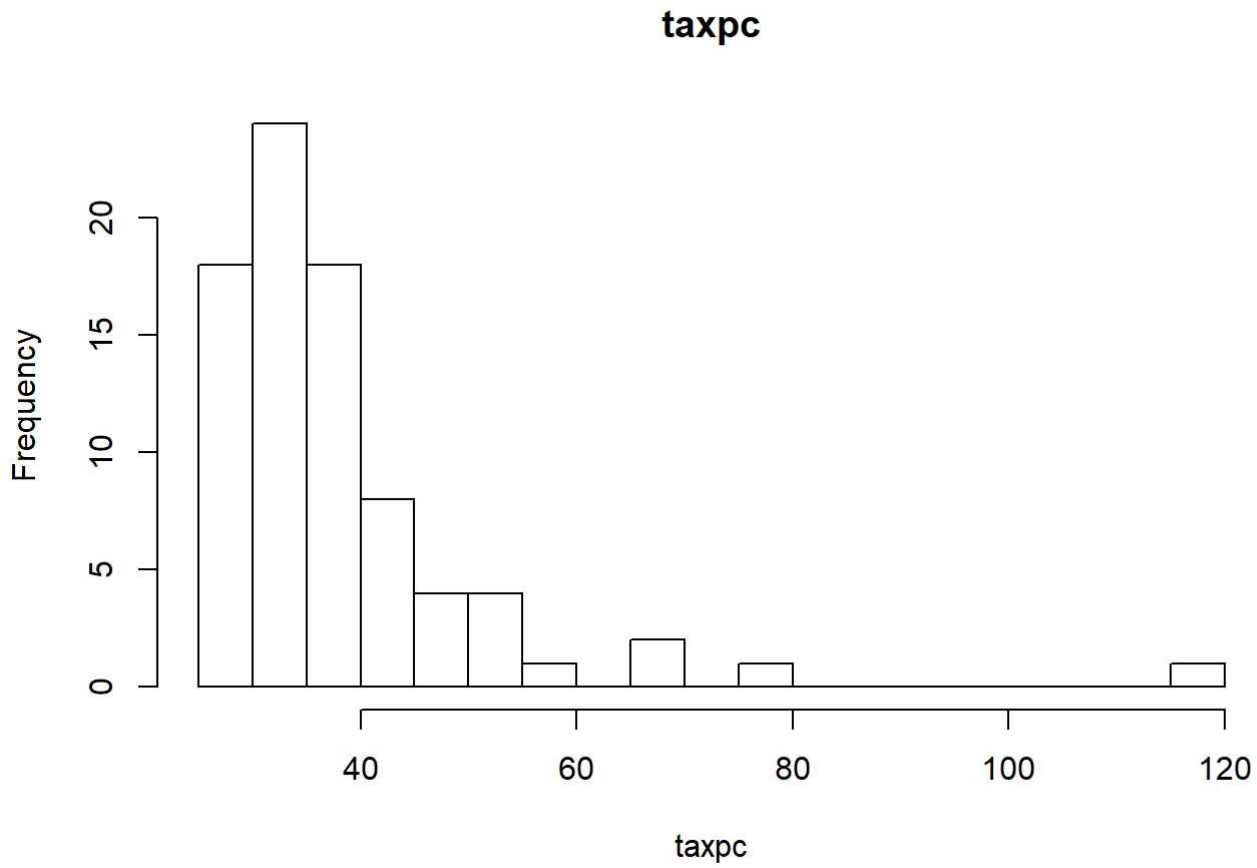
```
data_clean <- subset(data_clean,data_clean$prbconv <1)
hist(data_clean$prbconv, breaks=20, main="prbconv", xlab="prbconv")
```



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

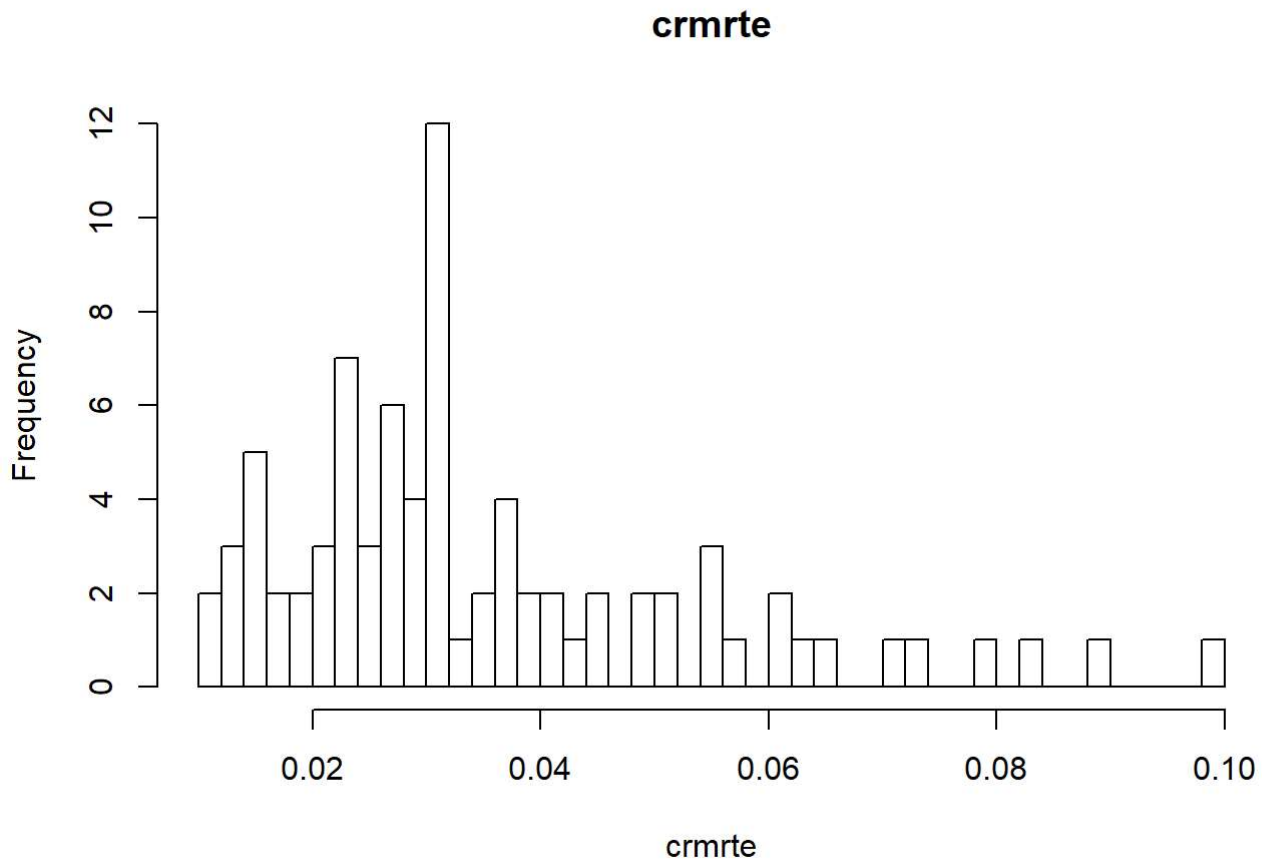
```
data_clean <- subset(data_clean, data_clean$wser<1000)
summary(data_clean$wser)
```

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

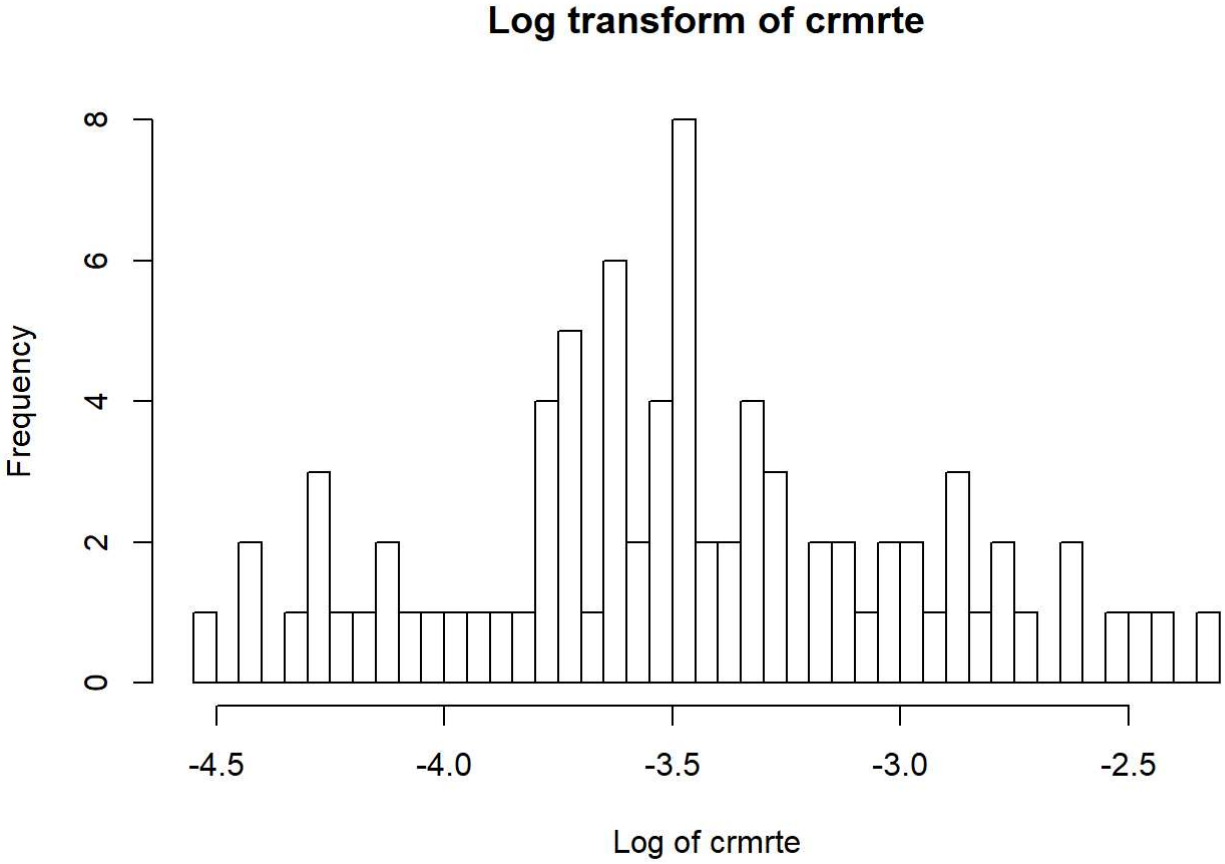
EDA

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")
```

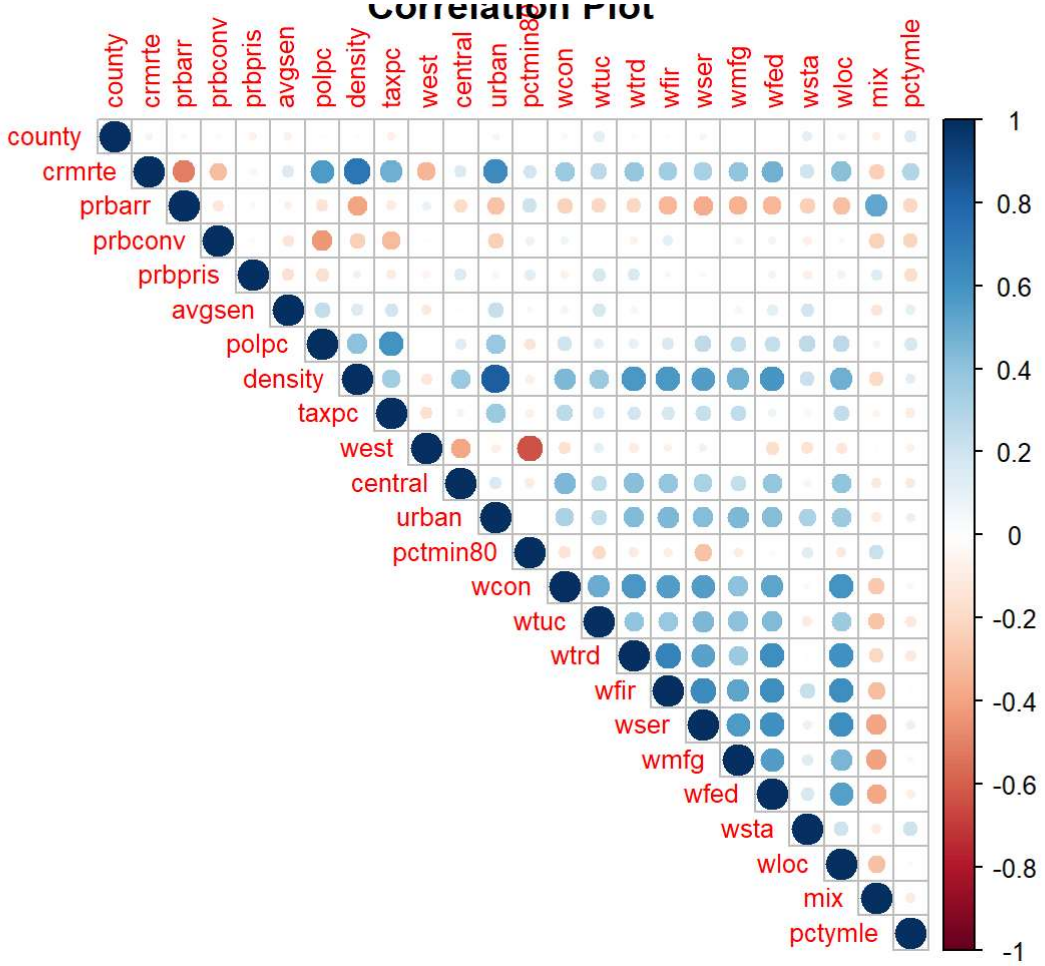


2. As the second part of the EDA, we take an initial look at the simple correlation between each variable inside the dataframe to help us understand the general correlations betwen each varible and the dependent variable to help identify the explanatory variables of interest

```
# 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 that might have been neglected before.

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.

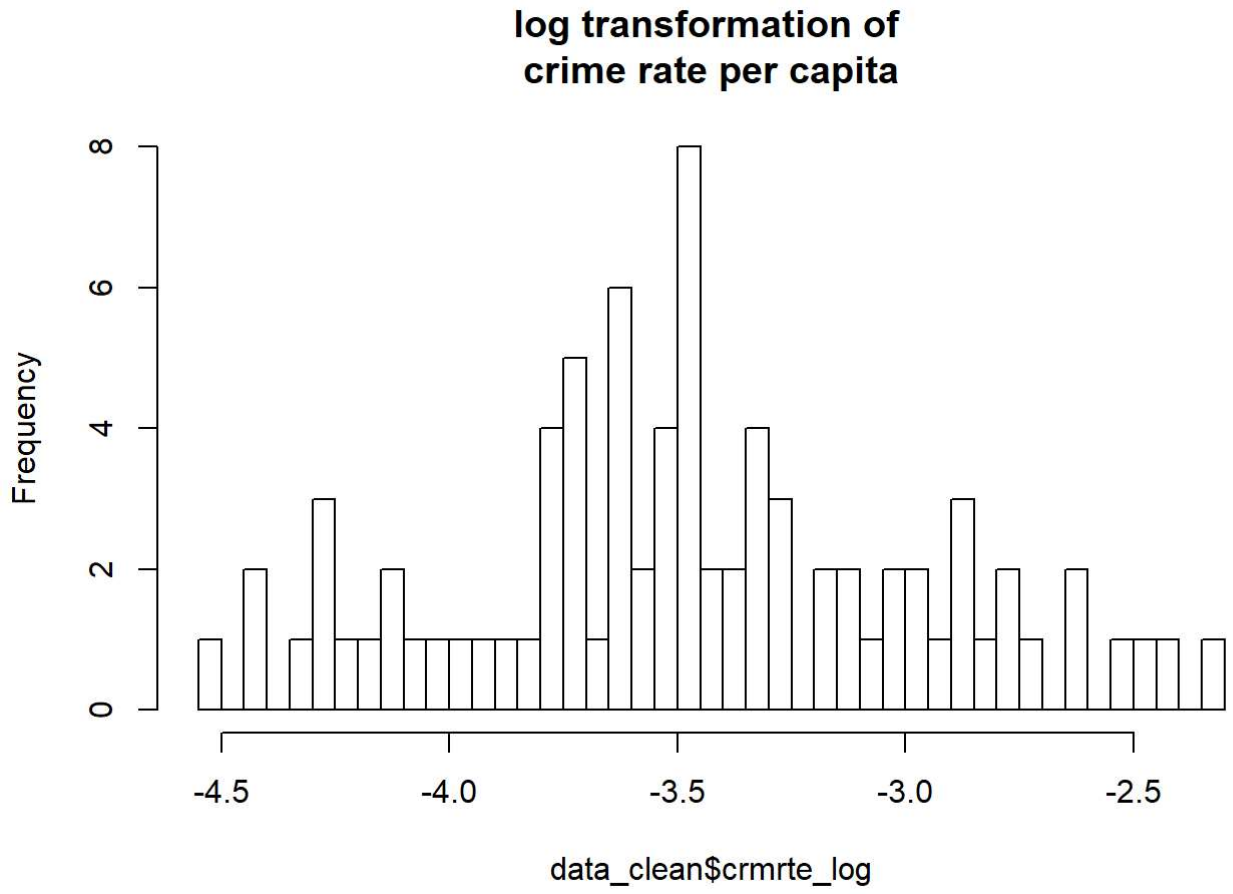
Fitting model 1

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")
```

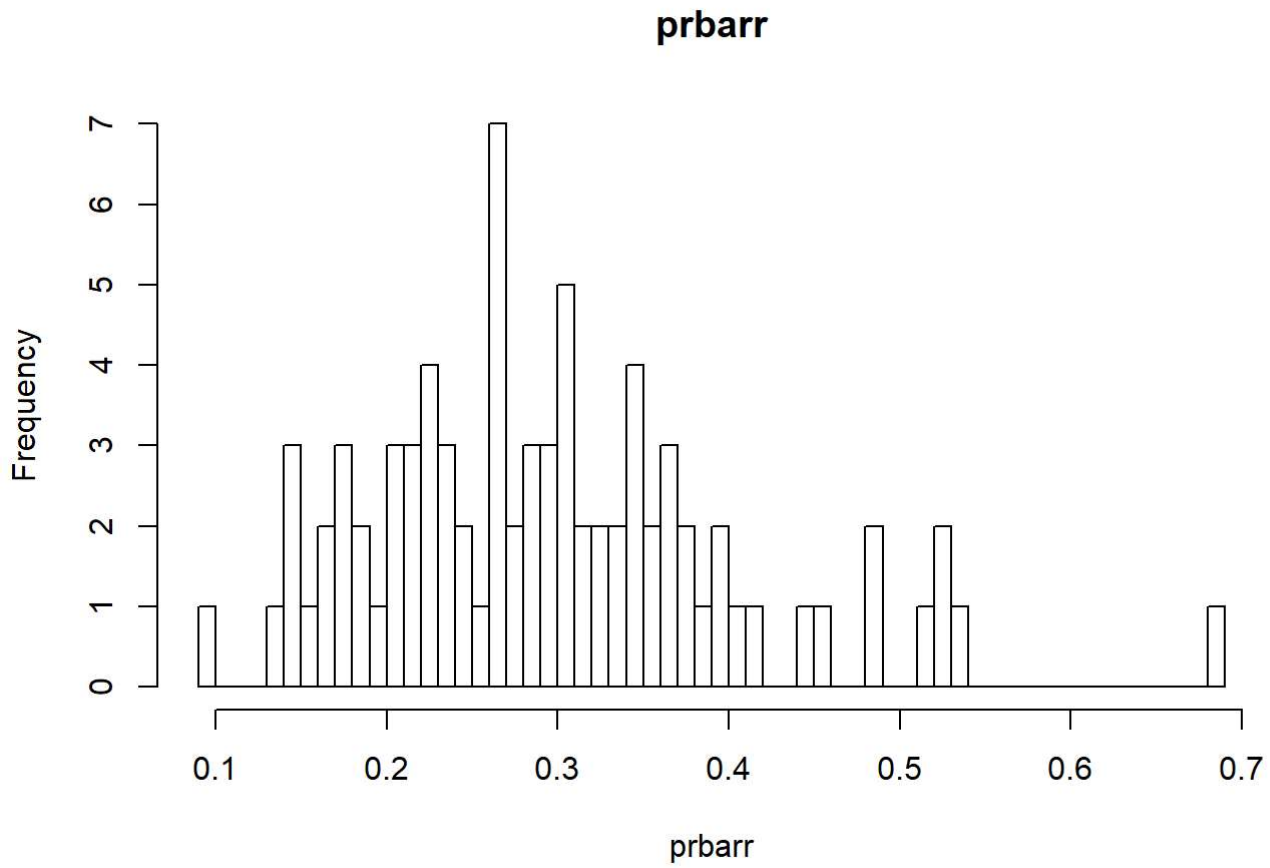


Looking at the other variables, and the simple correlation plot in initial EDA, we are proposing the explanatory variables that we believe contribute to crime rate:

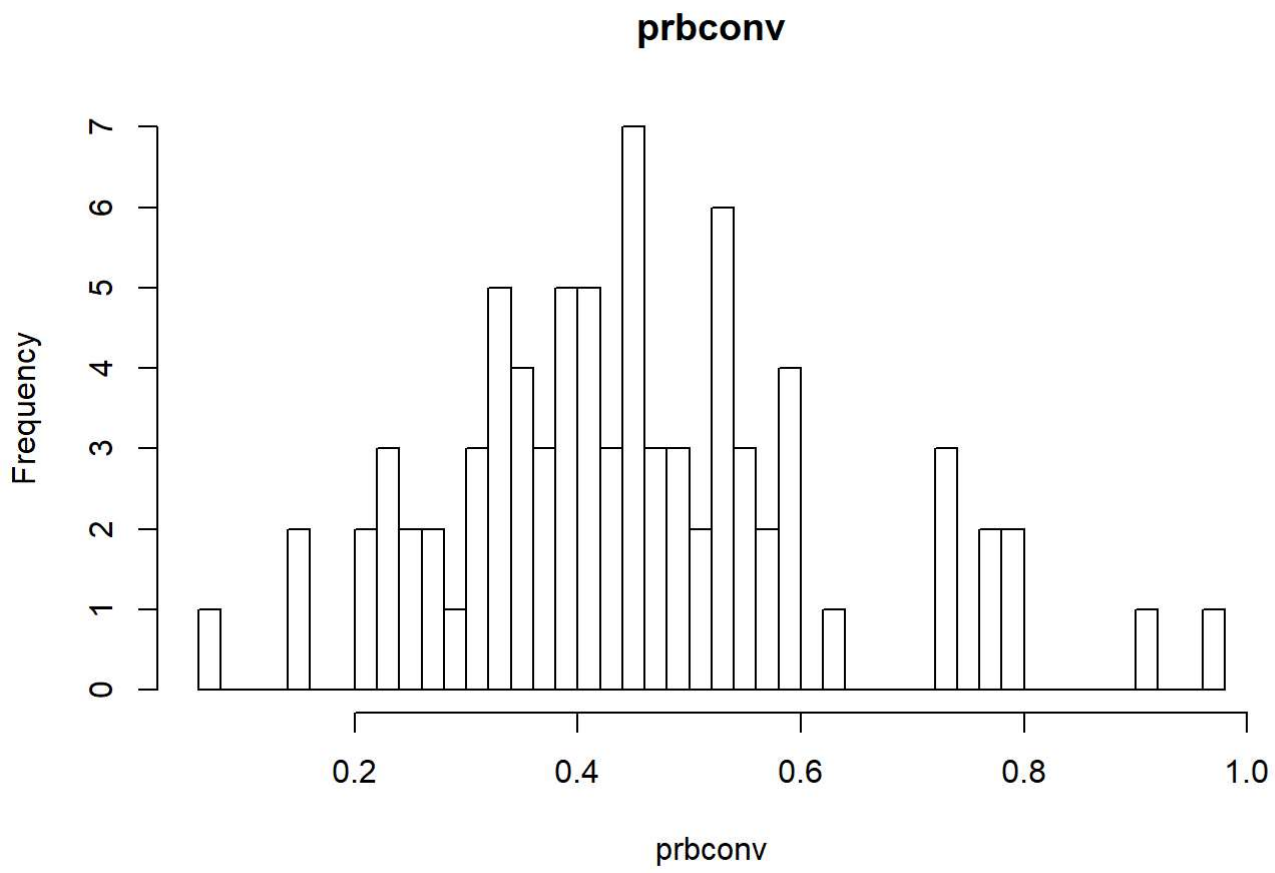
1. prbarr: the probability of arrest should be a direct contributing factor to crime rate. In other words, if people who have potential to commit a crime believe the chance of them getting arrested is small, then it might encourage them to commit a crime
2. prbconv: after getting arrested, getting suspects 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:

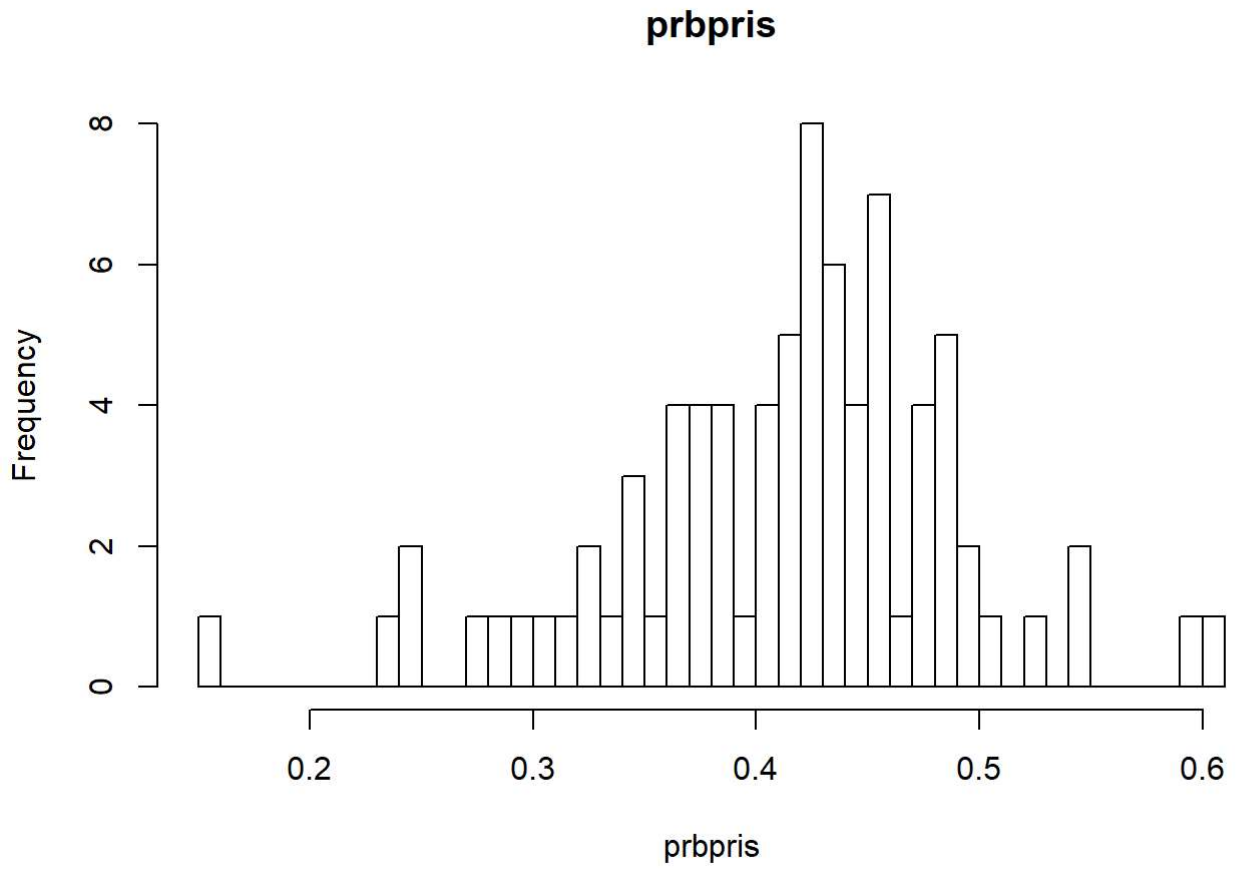
```
# prbarr
hist(data_clean$prbarr,breaks=50,main="prbarr", xlab="prbarr")
```



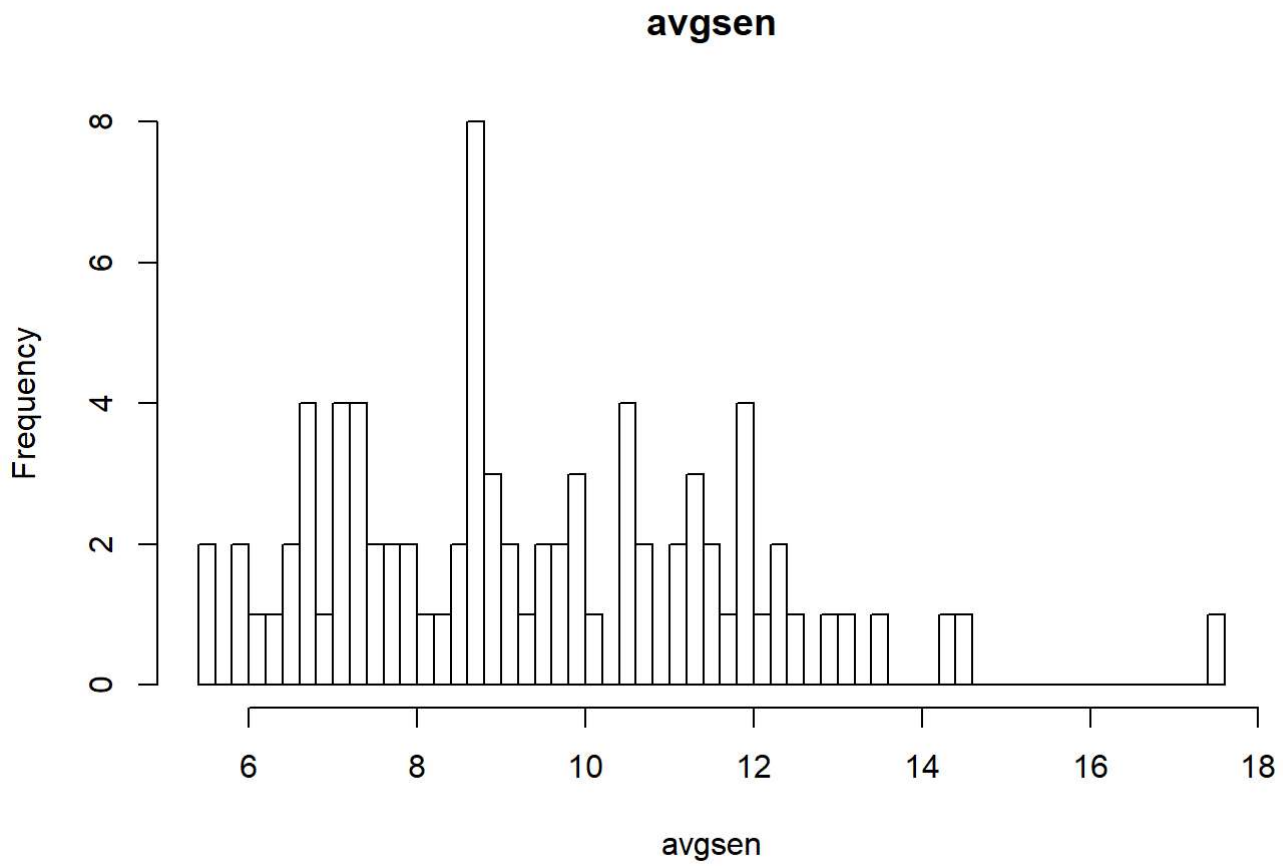

```
#prbconv
hist(data_clean$prbconv,breaks=50,main="prbconv", xlab="prbconv")
```




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



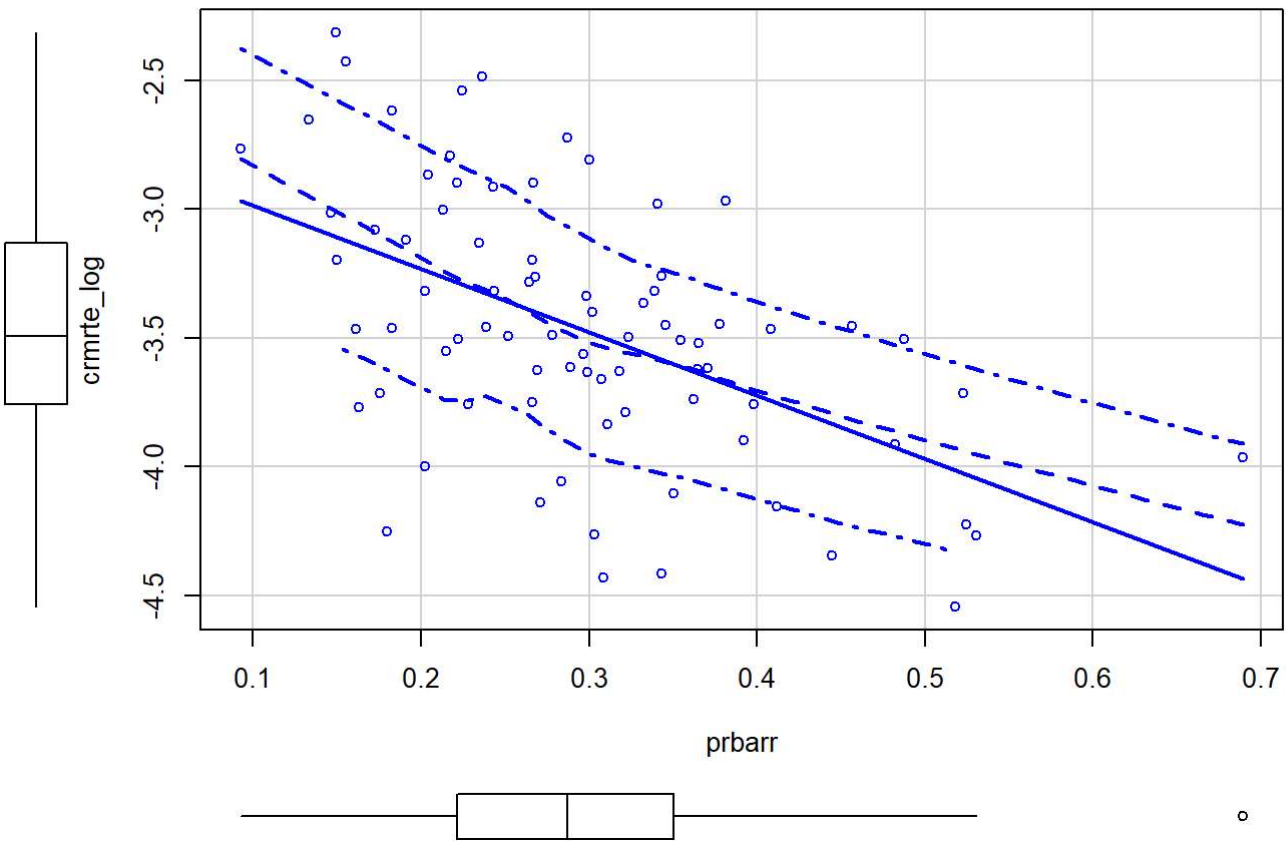

```
#avgsen
hist(data_clean$avgsen,breaks=50,main="avgsen", xlab="avgsen")
```



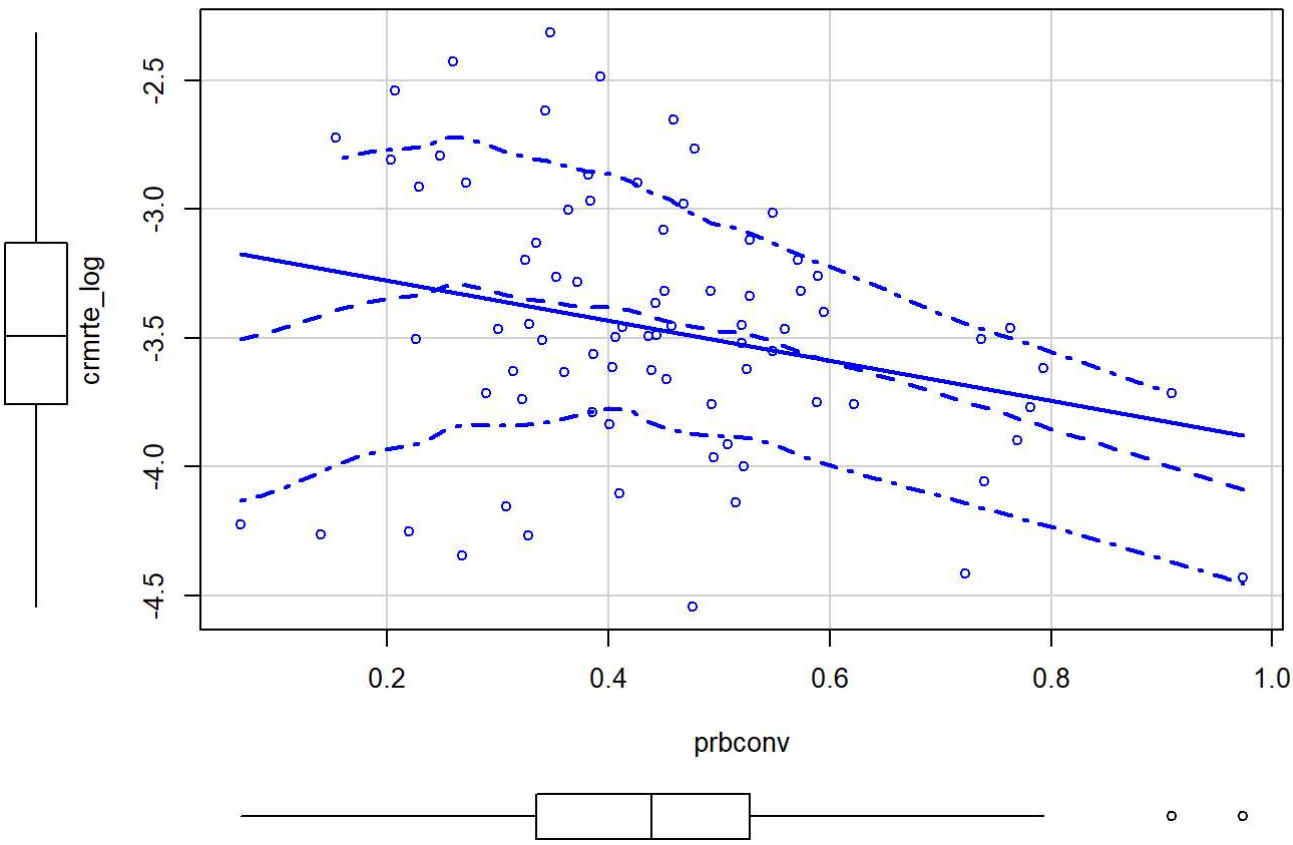
The above histograms of each explanatory variable all seem rather normally distributed, with some anomalies implied in $prbarr$ and $avgsen$. However, we don't think there is enough evidence to make a decision on whether the data is anomaly or not, and we will use the data for model building.

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

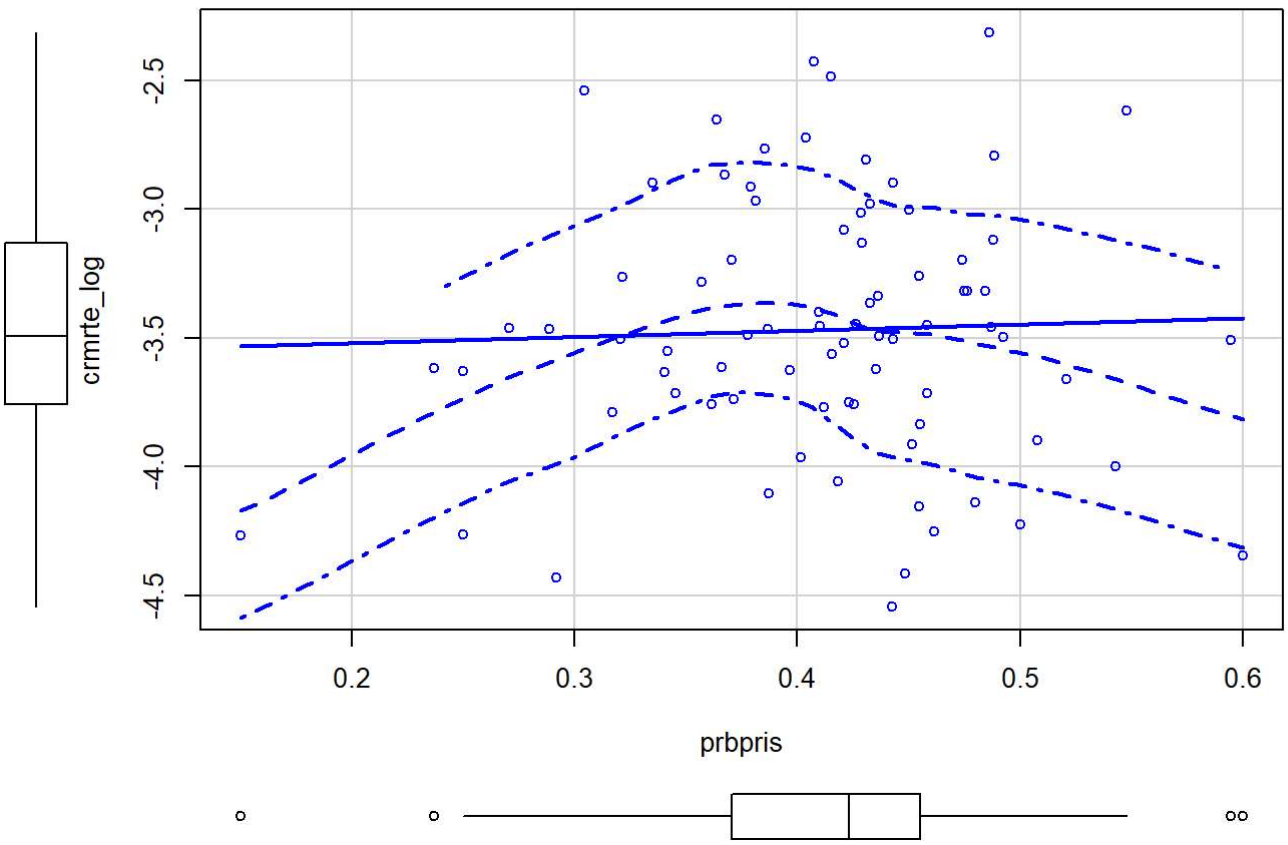
```
scatterplot(data_clean$prbarr,data_clean$crmrte_log, main="prbarr vs. crmrte_log", xlab="prbarr"
, ylab="crmrte_log")
```



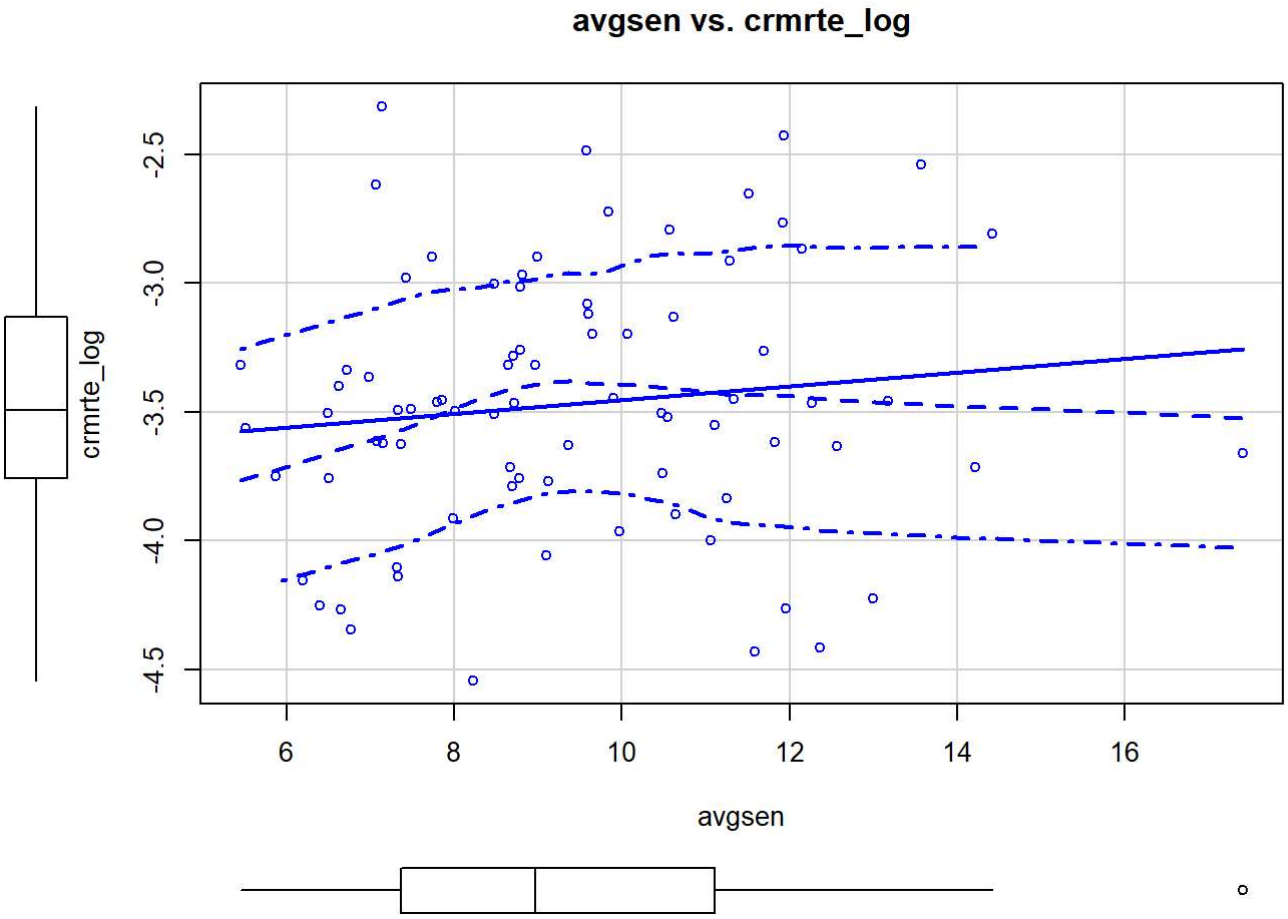

```
scatterplot(data_clean$prbconv,data_clean$crmrte_log, main="prbconv vs. crmrte_log", xlab="prbco
nv", ylab="crmrte_log")
```




```
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")
```



From the above scatterplots, we believe that the explanatory variables are displaying relatively linear relationship with dependent variable except for $prbpris$, which seems to be more in parabolic shape. Therefore, we have decided to create a new variable with square of probability of prison to better reveal the linearity.

```
data_clean$prbpris_sq <- data_clean$prbpris^2
```

_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 $

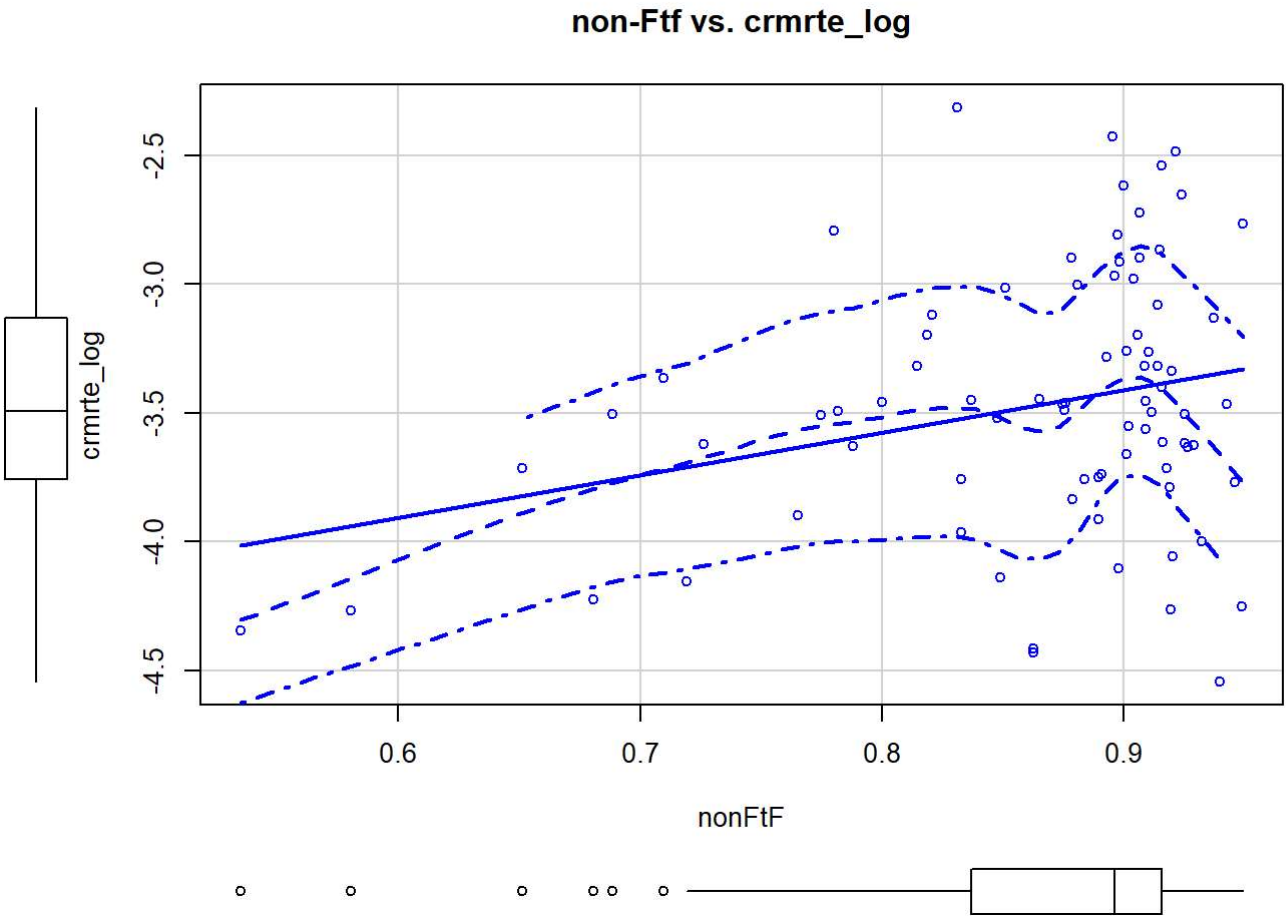
```
model1 <- lm(crmrte_log ~ prbarr + prbconv + avgsen + prbpris, data=data_clean)
model1
```

```
##
## 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
```

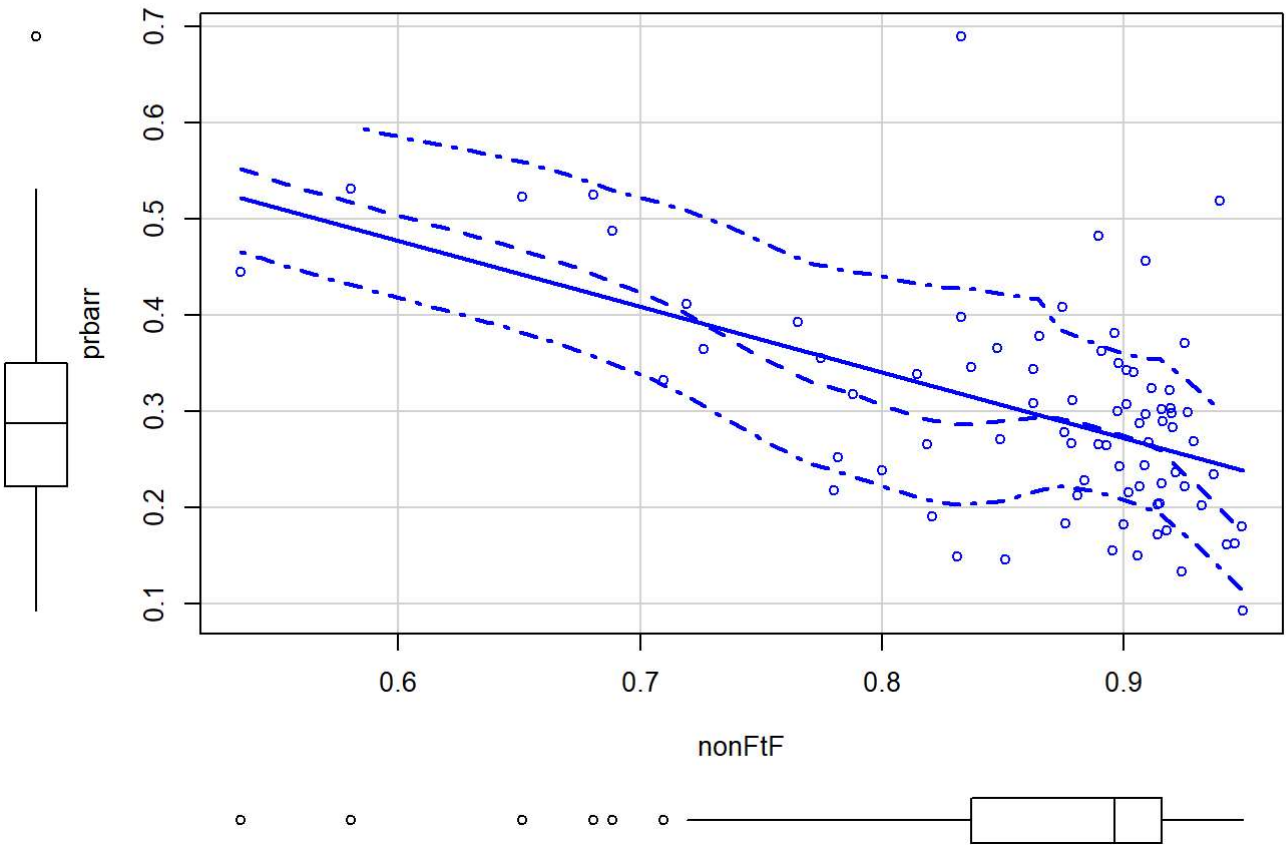
Fitting model 2

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:

```
data_clean$nonFtF <- 1-data_clean$mix
scatterplot(data_clean$nonFtF,data_clean$crmrte_log, main="non-Ftf vs. crmrte_log", xlab="nonFt
F", ylab="crmrte_log")
```




```
scatterplot(data_clean$nonFtF,data_clean$prbarr, main="non-Ftf vs. prbarr", xlab="nonFtF", ylab=
"prbarr")
```



1. mix: The above two scatterplot 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: 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
```

_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. For the purpose of building the model, we added two additional variables: $nonFtF$ and $polpcMult$. $nonFtf$ is the transformation of mix. It represent the non face to face offense. And $polpcMult$ is the transformation of polpc. We need to multiply polpc by 100 becuase the numbers are too small. Without the multiplication, polpc skewed the model significantly. After the data transformation, the model is as below:
crmrte_log = $_0 + _1 * prbarr + _2 * prbconv + _3 * avgsen + _4 * prbpris + _5 * nonFtf + _6 * taxpc + _7 * urban + _8 * pctymle + _9 * polpcMutl + u $

```
data_clean$nonFtf <- (1- data_clean$mix)
data_clean$polpcMult <- data_clean$polpc * 100
model2 <- lm(crmrte_log ~ prbarr + prbconv + avgsen + prbpris_sq + nonFtf + taxpc + urban + pcty
mle + polpcMult, data=data_clean)
model2
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + avgsen + prbpris_sq +
##     nonFtf + taxpc + urban + pctymle + polpcMult, data = data_clean)
##
## Coefficients:
## (Intercept)       prbarr      prbconv       avgsen   prbpris_sq
##    -4.33062     -1.53728     -0.29128     -0.01616      0.60003
##      nonFtf        taxpc        urban      pctymle    polpcMult
##     0.73421      0.00358      0.42856      3.06260      2.62029
```

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

Fitting model 3

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

```
model3 <- lm(crmrte_log ~ prbarr+prbconv+prbpris_sq+avgsen+polpcMult+density+taxpc+west+central+
urban+pctmin80+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+nonFtf+pctymle, data=data_clean)
model3
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + prbpris_sq + avgsen +
##     polpcMult + density + taxpc + west + central + urban + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     nonFtf + pctymle, data = data_clean)
##
## Coefficients:
## (Intercept)       prbarr      prbconv   prbpris_sq       avgsen
##  -4.5967576   -1.7058163   -0.2850715    0.0050278   -0.0218338
##   polpcMult      density        taxpc         west      central
##   2.7118838    0.1146256    0.0032761   -0.1629276   -0.1255685
##       urban     pctmin80         wcon         wtuc         wtrd
##  -0.0493747    0.0090044    0.0004000    0.0003957    0.0008749
##        wfir         wser         wmfg         wfed         wsta
##  -0.0023738   -0.0012955   -0.0002377    0.0022246   -0.0017153
##        wloc       nonFtf      pctymle
##   0.0019754    0.5863199    3.6248234
```

Here is the latex table in a PDF document:

	Dependent variable:		
	crmrte_log		
	(1)	(2)	(3)
prbarr	-2.626***	-1.537***	-1.706***
	(0.423)	(0.465)	(0.305)
prbconv	-0.962***	-0.291	-0.285
	(0.268)	(0.276)	(0.182)
avgsen	0.009	-0.016	-0.022*
	(0.020)	(0.018)	(0.012)
prbpris	0.100		
	(0.589)		
prbpris_sq		0.600	0.005
		(0.692)	(0.454)
nonFtf		0.734	0.586
		(0.599)	(0.418)
taxpc		0.004	0.003
		(0.004)	(0.003)
west			-0.163
			(0.115)
central			-0.126
			(0.077)
urban		0.429***	-0.049
		(0.152)	(0.163)
pctmin80			0.009***
			(0.003)
wcon			0.0004
			(0.001)
wtuc			0.0004
			(0.0004)
wtrd			0.001
			(0.001)
wfir			-0.002***
			(0.001)
wser			-0.001
			(0.001)
wmfg			-0.0002
			(0.0004)
wfed			0.002***
			(0.001)
wsta			-0.002**
			(0.001)
wloc			0.002
			(0.001)
pctymle		3.063	3.625***
		(1.868)	(1.305)
polpcMult		2.620**	2.712***
		(1.030)	(0.769)
density			0.115***

			(0.035)
Constant	-2.385***	-4.331***	-4.597***
	(0.402)	(0.741)	(0.568)
Observations	81	81	81
R^2	0.390	0.573	0.870
Adjusted R^2	0.358	0.519	0.820
Residual Std. Error	0.405 (df = 76)	0.351 (df = 71)	0.214 (df = 58)
F Statistic	12.156*** (df = 4; 76)	10.589*** (df = 9; 71)	17.585*** (df = 22; 58)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

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.

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.