Data 621 Homework 3

Layla Quinones

10/24/2021

Libraries

```
library(tidyverse)
library(ggplot2)
library(VIM)
library(GGally)
library(caret)
```

EDA

```
# Load data
# Training
rawTrain <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-training

#Testing data
rawTest <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-evaluati
```

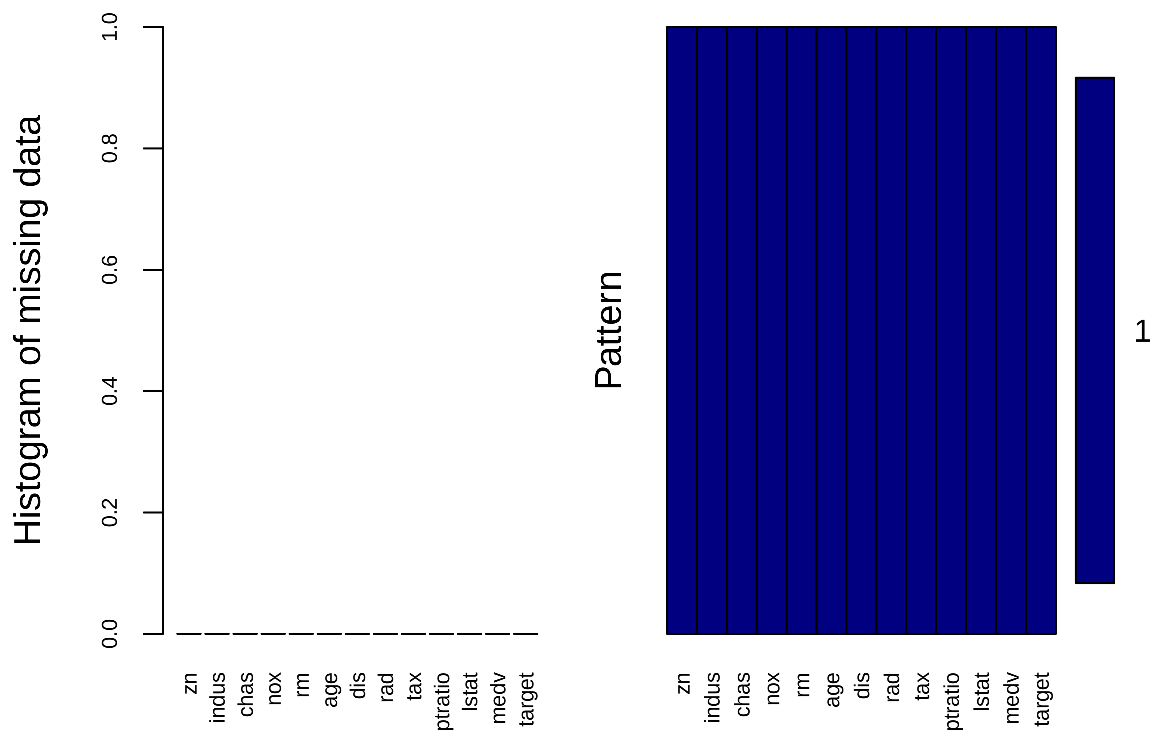
```
# check to see if we need to clean the data
# gives us a sense of what each predictor is
glimpse(rawTrain)
```

```
## Rows: 466
## Columns: 13
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 100, 20...
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5.19, ...
## $ chas    <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ nox     <dbl> 0.605, 0.871, 0.740, 0.428, 0.488, 0.520, 0.693, 0.693, 0.5...
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519, 6.3...
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38.1, 19...
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896, 1.6...
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5, 5, 2...
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330, 398,...
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, 16.4,...
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5.68, 9...
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20.9, 2...
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1,...
```

```r
# All varaibles are numeric
# categorical variables
# chas

#dicrete
#rad, zn, tax

#all others are continuous
```

No Missing Values

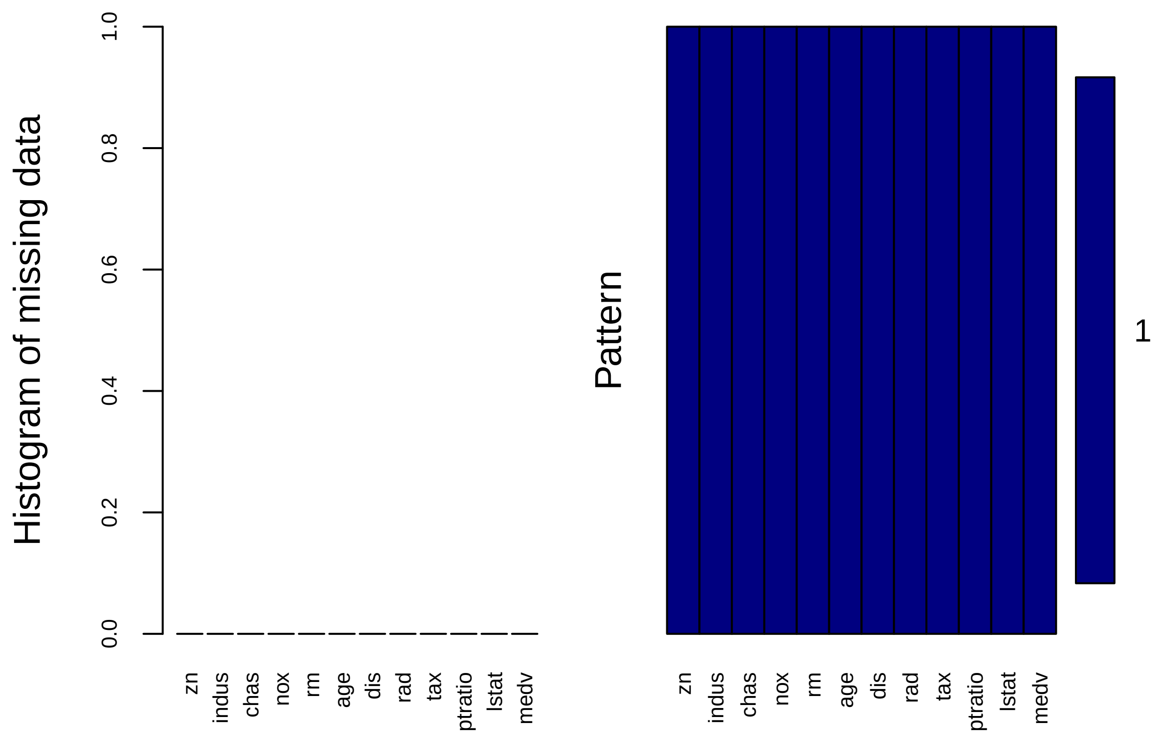
```r
#plot missing values using VIM package
aggr(rawTrain , col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(rawTrain), cex.axis=
```




```
##
##  Variables sorted by number of missings:
##  Variable Count
##        zn     0
##     indus     0
##      chas     0
##       nox     0
##        rm     0
##       age     0
```

```
##       dis     0
##       rad     0
##       tax     0
##   ptratio     0
##     lstat     0
##      medv     0
##    target     0
```

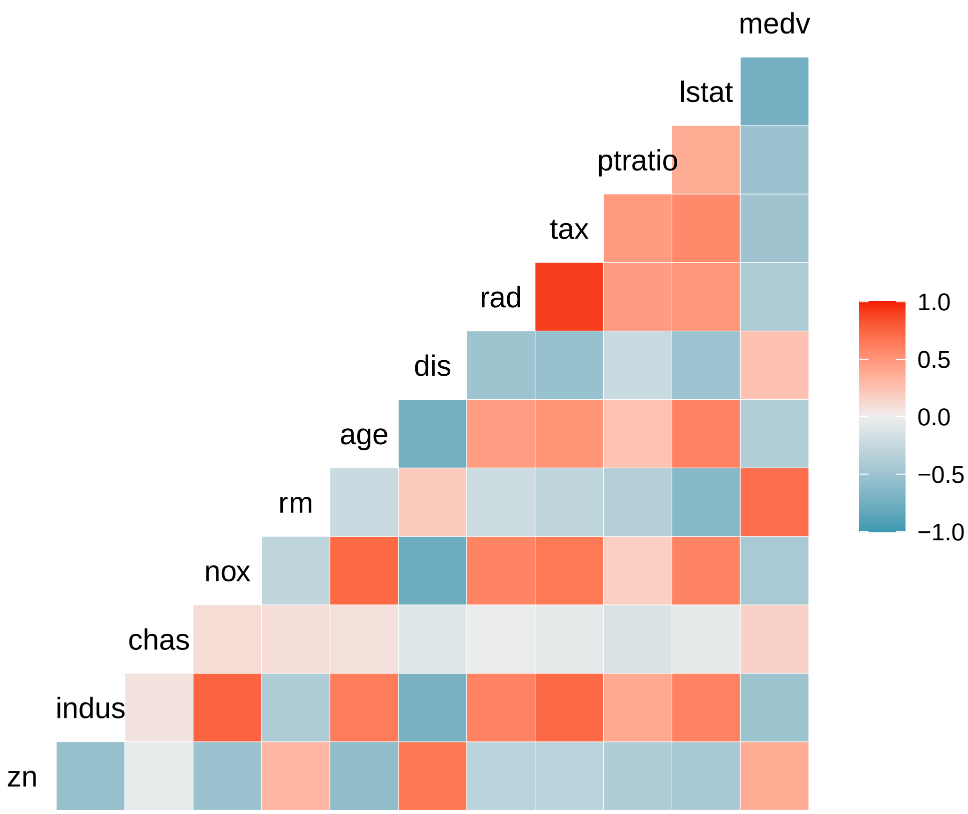
```r
#plot missing values using VIM package
aggr(rawTest , col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(rawTrain), cex.axis=.
```



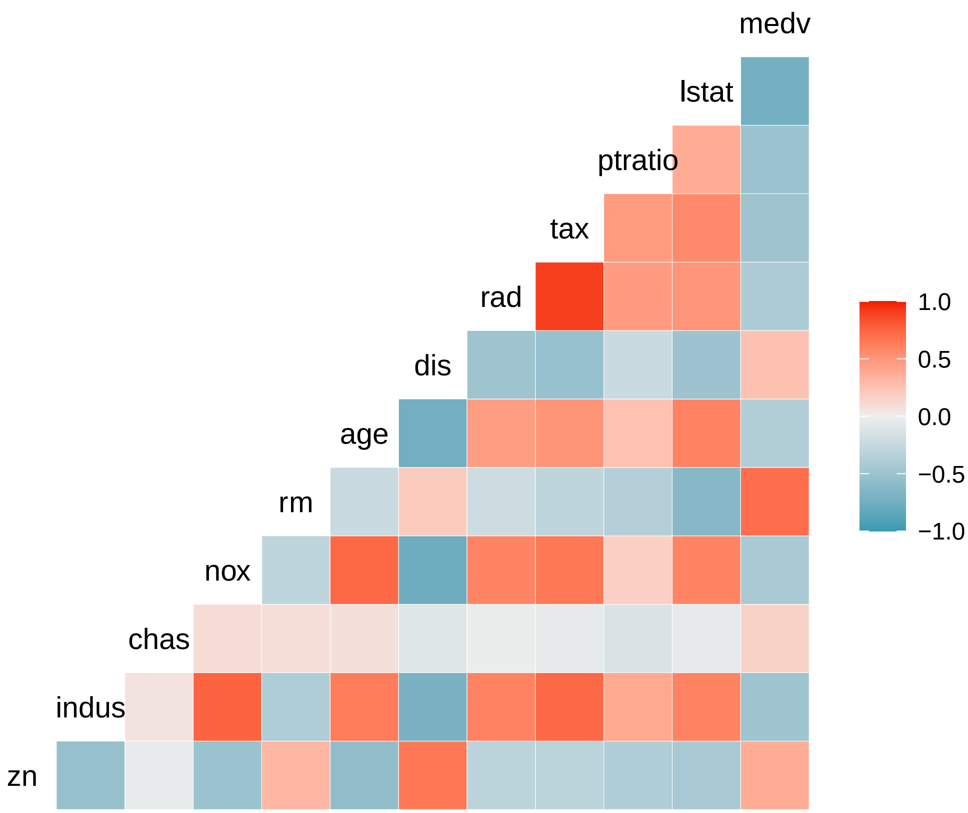

```
##
##  Variables sorted by number of missings:
##  Variable Count
##        zn     0
##     indus     0
##      chas     0
##       nox     0
##        rm     0
##       age     0
##       dis     0
##       rad     0
##       tax     0
##   ptratio     0
##     lstat     0
##      medv     0
```

Correlation

```
#correlation matrix for predictors
ggcorr(rawTrain%>% select(zn:medv))
```




```
#Idetify highly correlated variables
ggcorr(rawTrain%>% select(zn:medv))
```




```r
#Lets look at some highly correlated variables and drop them
findCorrelation(cor(rawTrain%>% select(zn:medv)),
                cutoff = 0.75,
                verbose = TRUE,
                names = TRUE)
```

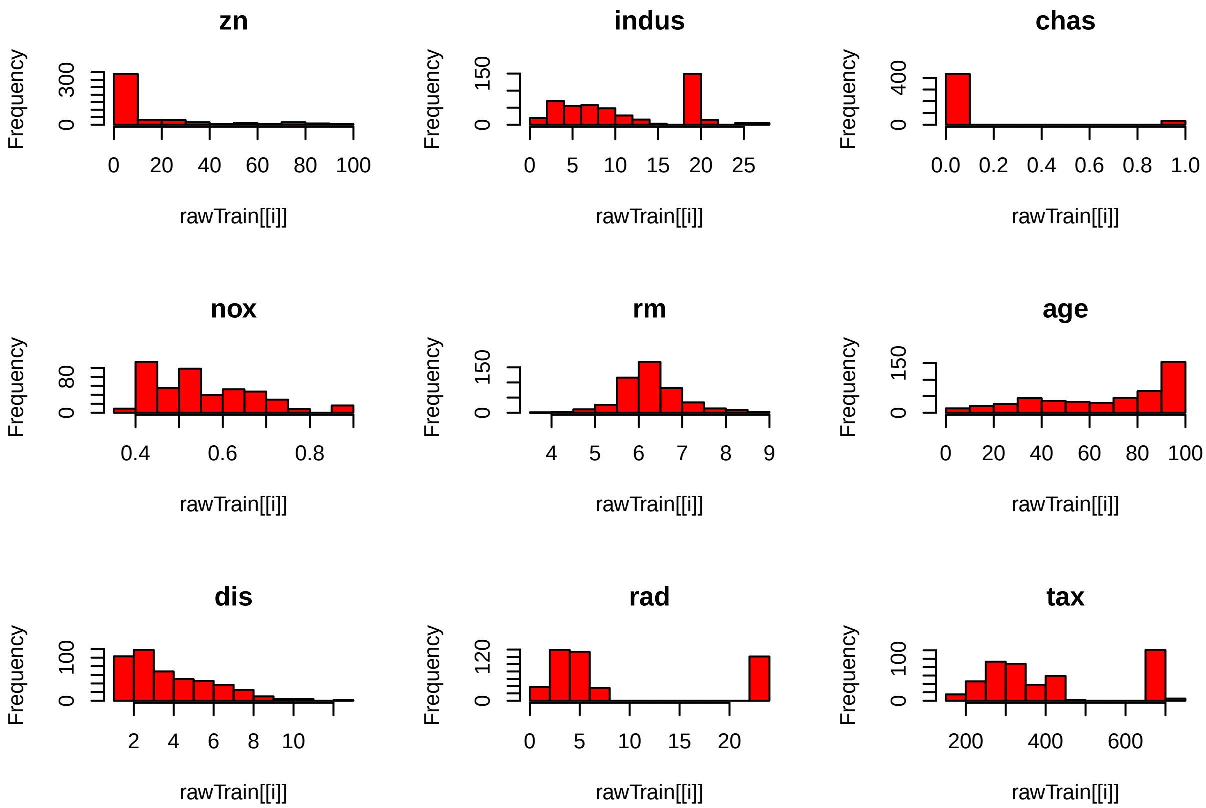
```
## Compare row 2  and column  4 with corr  0.76
##   Means:  0.539 vs 0.416 so flagging column 2
## Compare row 4  and column  7 with corr  0.769
##   Means:  0.487 vs 0.395 so flagging column 4
## Compare row 9  and column  8 with corr  0.906
##   Means:  0.46 vs 0.377 so flagging column 9
## Compare row 6  and column  7 with corr  0.751
##   Means:  0.417 vs 0.357 so flagging column 6
## All correlations <= 0.75

## [1] "indus" "nox"   "tax"   "age"
```

```r
# There are 4 highly correlated variables
# I will drop the highest one which is tax which seems to be the most highly correlated
#tax and rad are 0.9 correlated lets look at their relationship to the predictor to see which one to dr
```

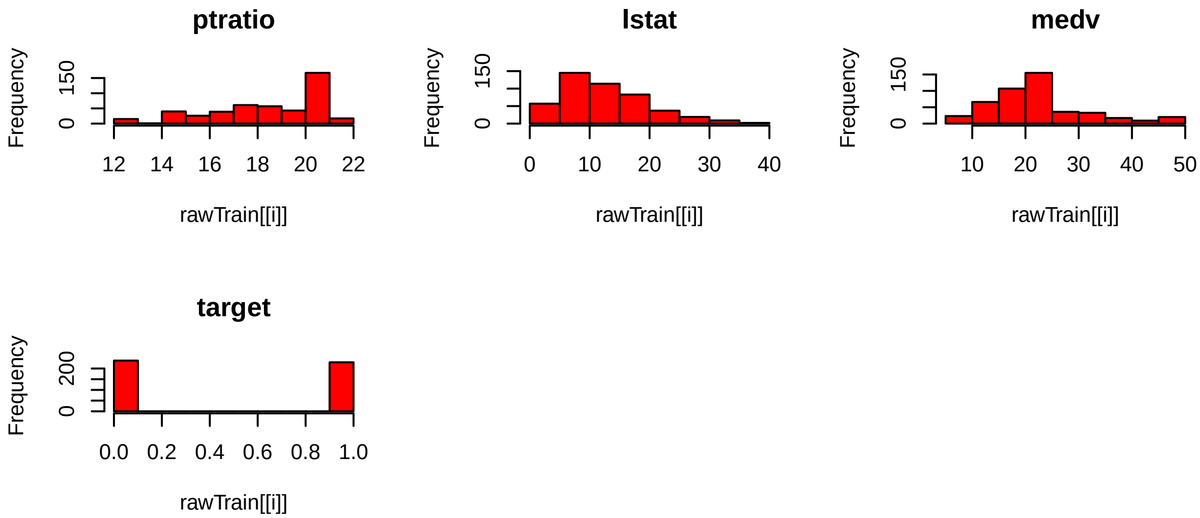
Distribution of Predictors

ADD VARIANCE AND INFLATION FACTORS TO THIS SECTION

```r
par(mfrow = c(3,3))
for(i in 1:ncol(rawTrain)) {#distribution of each variable
  hist(rawTrain[[i]], main = colnames(rawTrain[i]), col = "red")
}
```

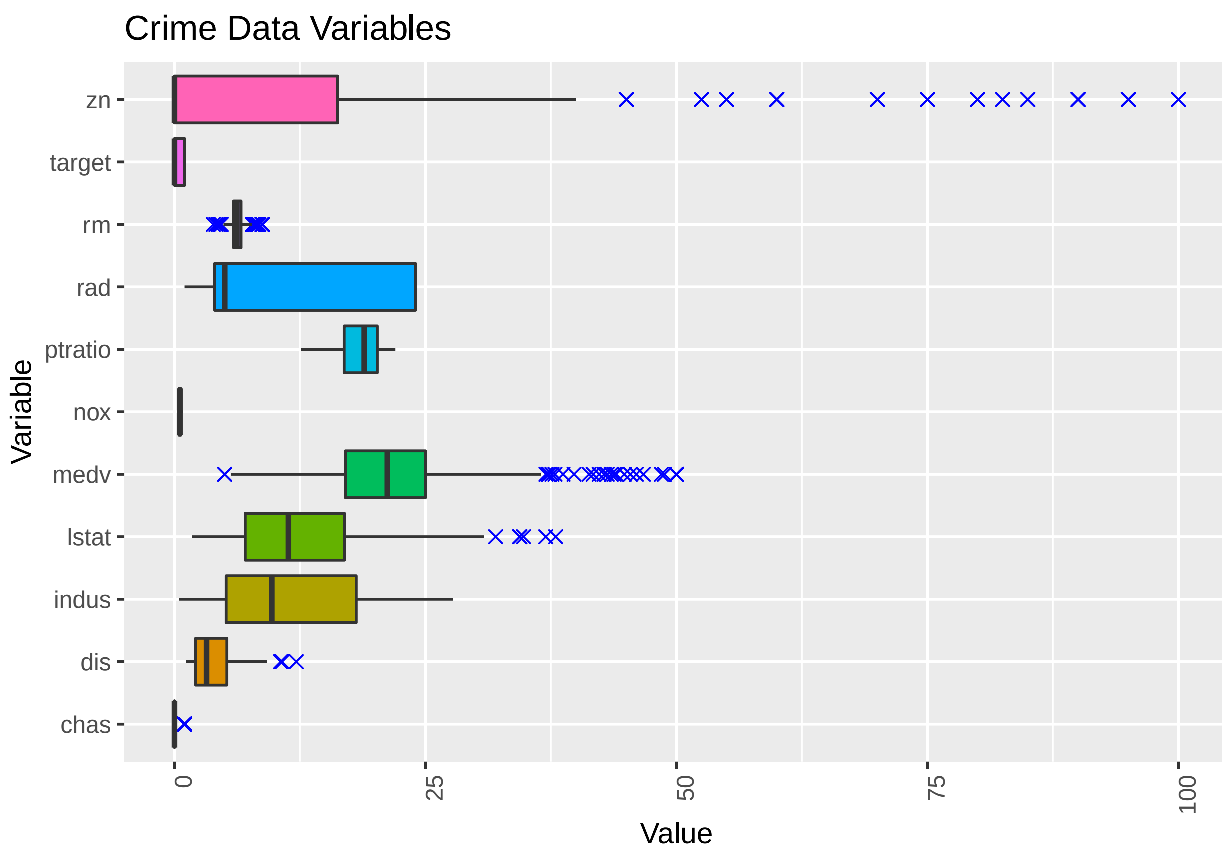



```r
#binomial data
# indus, tax and rad

#all other variables ar skewed excpet RM
```



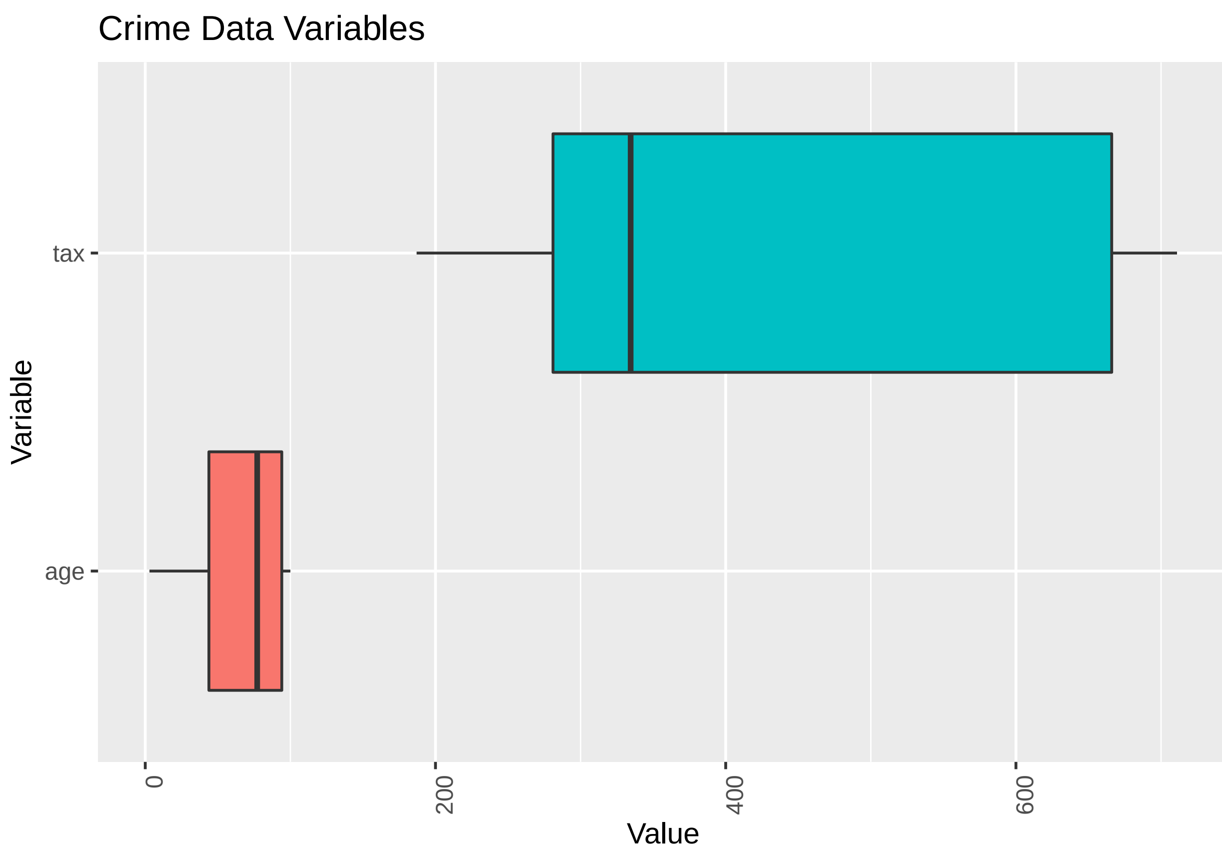
Box Plots

```r
#make long
#tax and age has a much different scale so we are seperating it here
longData <- rawTrain %>%
  select(-tax, -age) %>%
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData, aes(Variable, Value, fill = Variable)) +
  geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Crime Data Variables", y="Value")
```




```
#we can see that zn, medv and lstat has MANY outliers
```

```
#make long
#tax and age has a much different scale so we are seperating it here
longData <- rawTrain %>%
  select(tax, age) %>%
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData, aes(Variable, Value, fill = Variable)) +
  geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Crime Data Variables", y="Value")
```




```
# no outliers for tax and age
```

Model Building

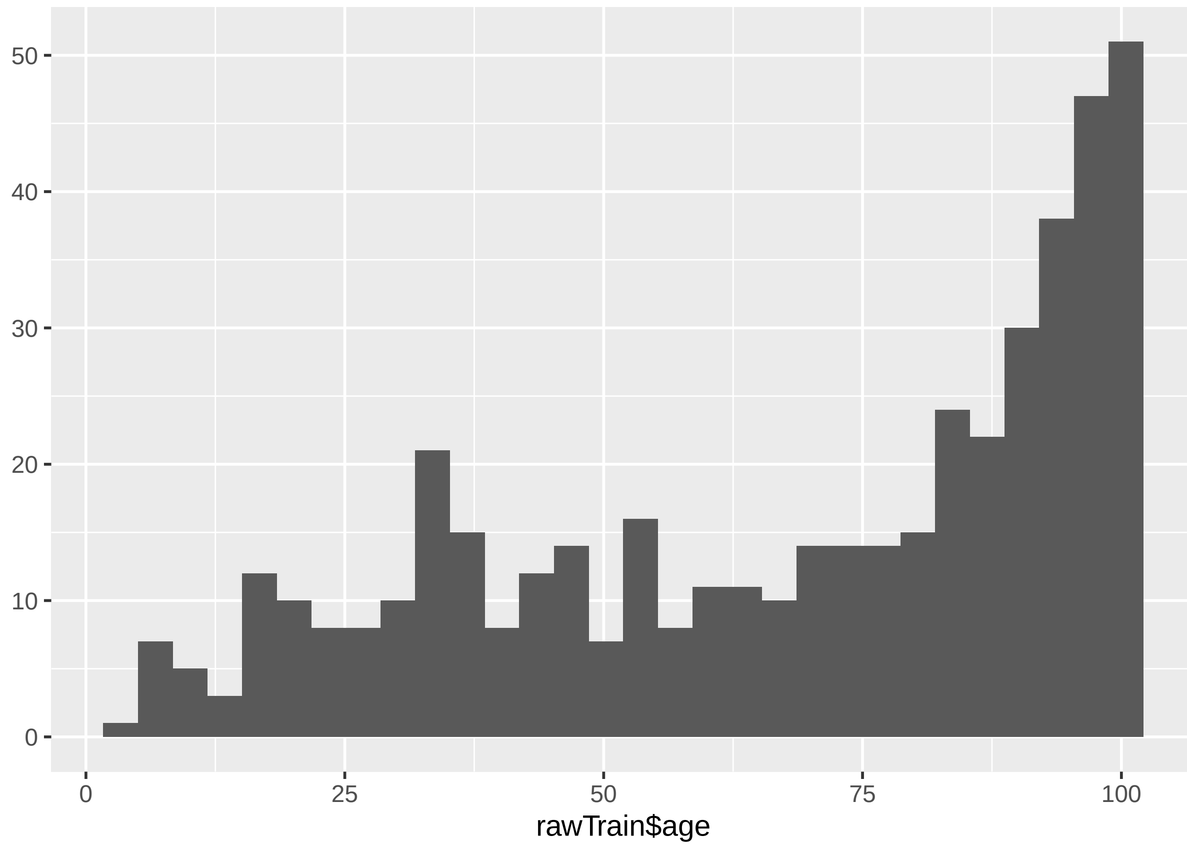
```
#remove Tax due to high correlation with other variables
modelOne <- glm(target ~ zn + indus + chas + nox + rm + age + dis + rad + ptratio + lstat + medv , data

modelOne
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age + dis +
##     rad + ptratio + lstat + medv, family = "binomial", data = rawTrain)
##
## Coefficients:
## (Intercept)           zn        indus         chas          nox           rm
##   -41.17734     -0.07141     -0.11249      1.25335     49.11180     -0.69362
##         age          dis          rad      ptratio        lstat         medv
##     0.03471      0.83505      0.50619      0.38009      0.03387      0.19946
##
## Degrees of Freedom: 465 Total (i.e. Null);  454 Residual
## Null Deviance:       645.9
## Residual Deviance: 196.6     AIC: 220.6
```

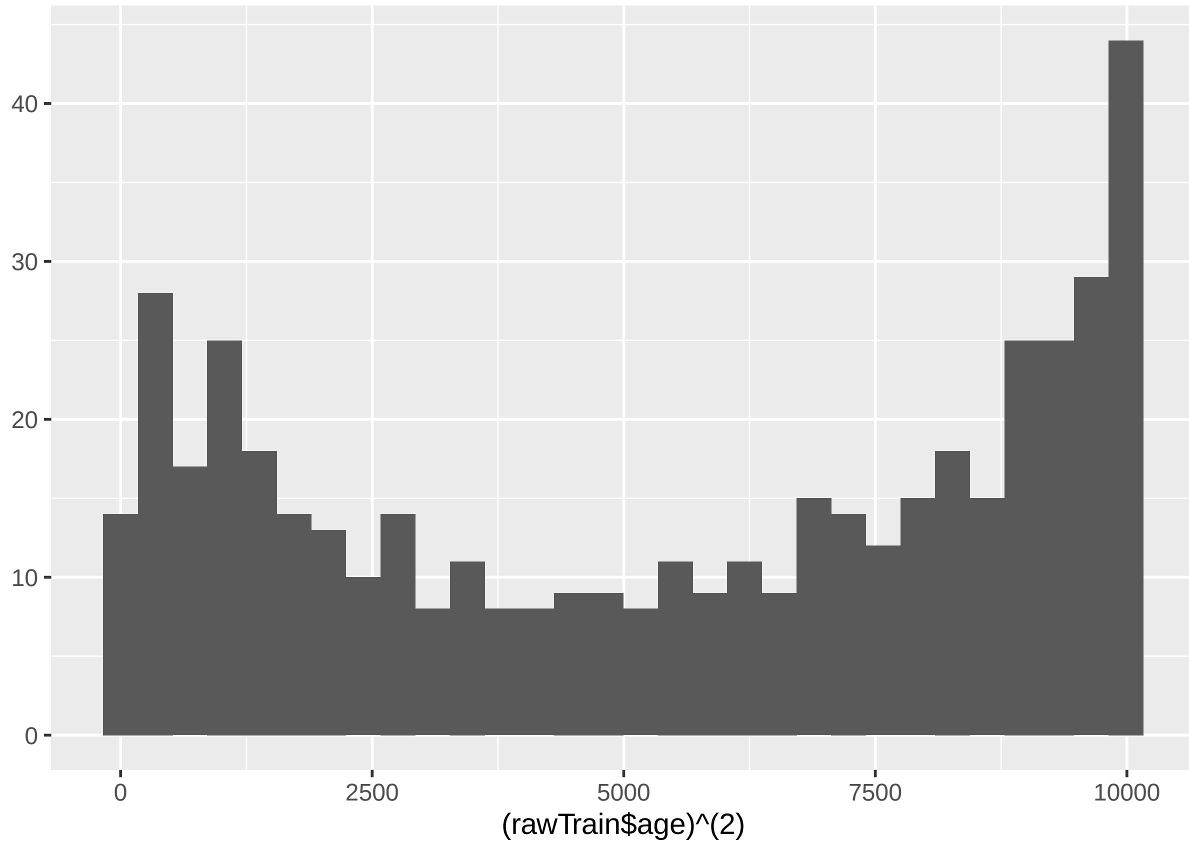
```r
# squared transformation to age and lstat

#age before squared
summary(rawTrain$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.90   43.88   77.15   68.37   94.10  100.00
```

```r
#age before squared
qplot(rawTrain$age)
```



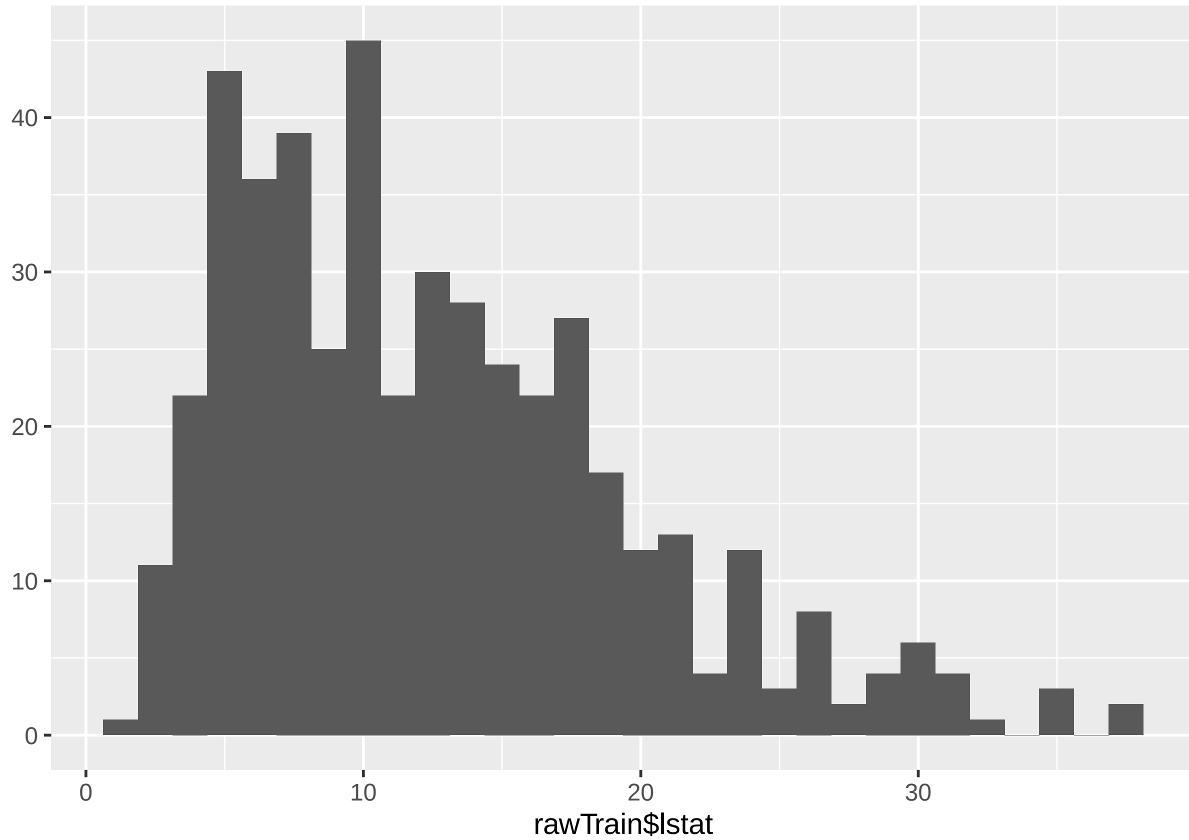

```r
#age after squared
qplot((rawTrain$age)^(2))
```



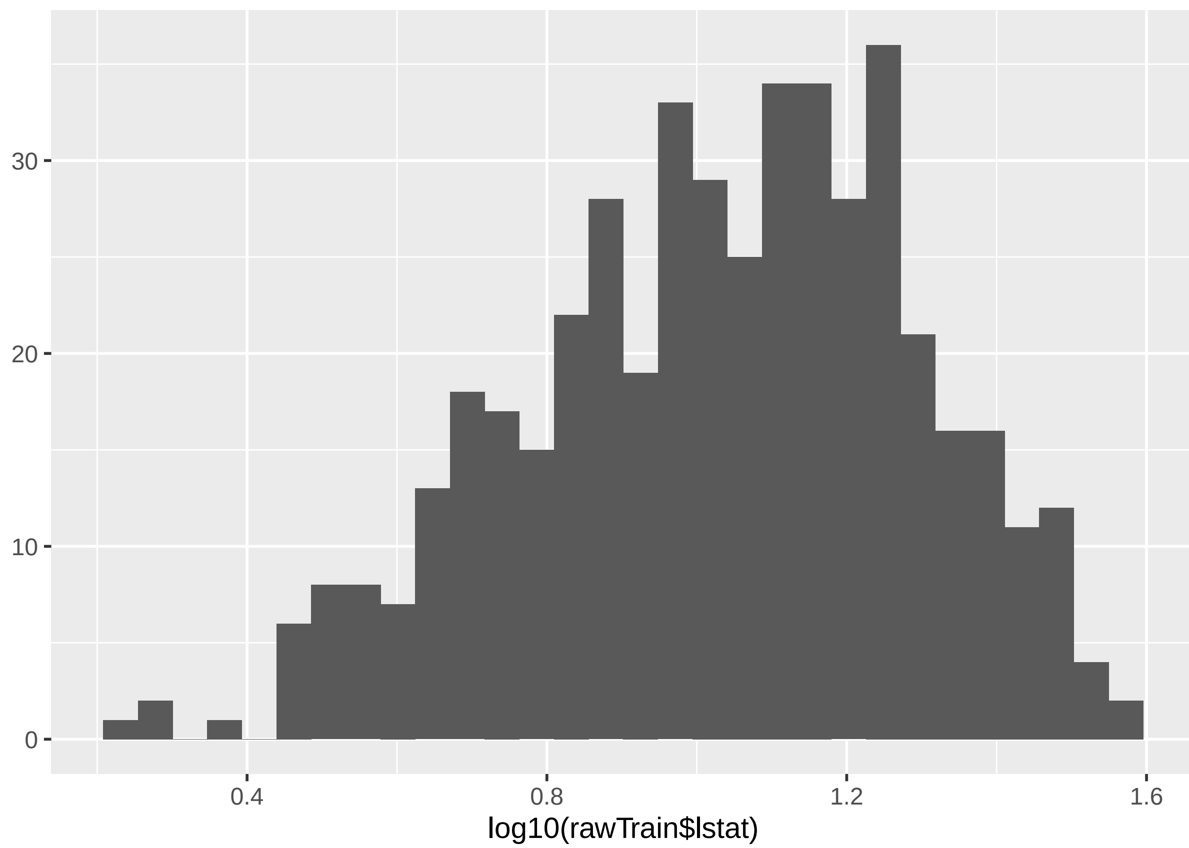

```
#lstat before log
summary(rawTrain$lstat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.730   7.043  11.350  12.631  16.930  37.970
```

```
#lstat before log
qplot(rawTrain$lstat)
```




```
#lstat afterlog
qplot(log10(rawTrain$lstat))
```




```r
#remove Tax squared age and log lstat
modelTwo <- glm(target ~ zn + indus + chas + nox + rm + age^2 + dis + rad + ptratio + log10(lstat) + me

modelTwo
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age^2 +
##     dis + rad + ptratio + log10(lstat) + medv, family = "binomial",
##     data = rawTrain)
##
## Coefficients:
##  (Intercept)           zn        indus         chas          nox
##    -40.12700     -0.06734     -0.10958      1.33475     49.33564
##           rm          age          dis          rad      ptratio
##     -0.89104      0.03896      0.84436      0.51164      0.39222
## log10(lstat)         medv
##     -0.14881      0.19872
##
## Degrees of Freedom: 465 Total (i.e. Null);  454 Residual
## Null Deviance:       645.9
## Residual Deviance: 197   AIC: 221
```

```r
#remove Tax squared age and log lstat - log dis and zn
modelThree <- glm(target ~ log10(zn + 1) + indus chas + nox + rm + age^2 + log10(dis) + rad + ptratio

modelThree
```

```
##
## Call:  glm(formula = target ~ log10(zn + 1) + indus + chas + nox + rm +
##     age^2 + log10(dis) + rad + ptratio + log10(lstat) + medv,
##     family = "binomial", data = rawTrain)
##
## Coefficients:
##   (Intercept)  log10(zn + 1)          indus           chas            nox
##     -46.69939       -1.00777       -0.07120        1.11180       54.23952
##            rm            age     log10(dis)            rad        ptratio
##      -1.01689        0.04484       10.03136        0.55084        0.41541
##  log10(lstat)           medv
##       0.12432        0.23433
##
## Degrees of Freedom: 465 Total (i.e. Null);  454 Residual
## Null Deviance:       645.9
## Residual Deviance: 189.2     AIC: 213.2
```

```
#remove Tax squared age and log lstat - log dis - log zn
modelFour <- glm(target ~ zn + indus + chas + nox + rm + age^2 + log10(dis) + rad + ptratio + log10(lsta

modelFour
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age^2 +
##     log10(dis) + rad + ptratio + log10(lstat) + medv, family = "binomial",
##     data = rawTrain)
##
## Coefficients:
##  (Intercept)            zn         indus          chas           nox
##    -45.71222      -0.04939      -0.06976       1.17911      53.13061
##           rm           age    log10(dis)           rad       ptratio
##     -1.05455       0.04416       9.47828       0.55075       0.44318
## log10(lstat)          medv
##     -0.16944       0.23047
##
## Degrees of Freedom: 465 Total (i.e. Null);  454 Residual
## Null Deviance:       645.9
## Residual Deviance: 189.8     AIC: 213.8
```

NEXT I WANT TO TRY BOX COX TRANSFORMATIONS

WEE NEED QQ PLOTS AND ACCURACY