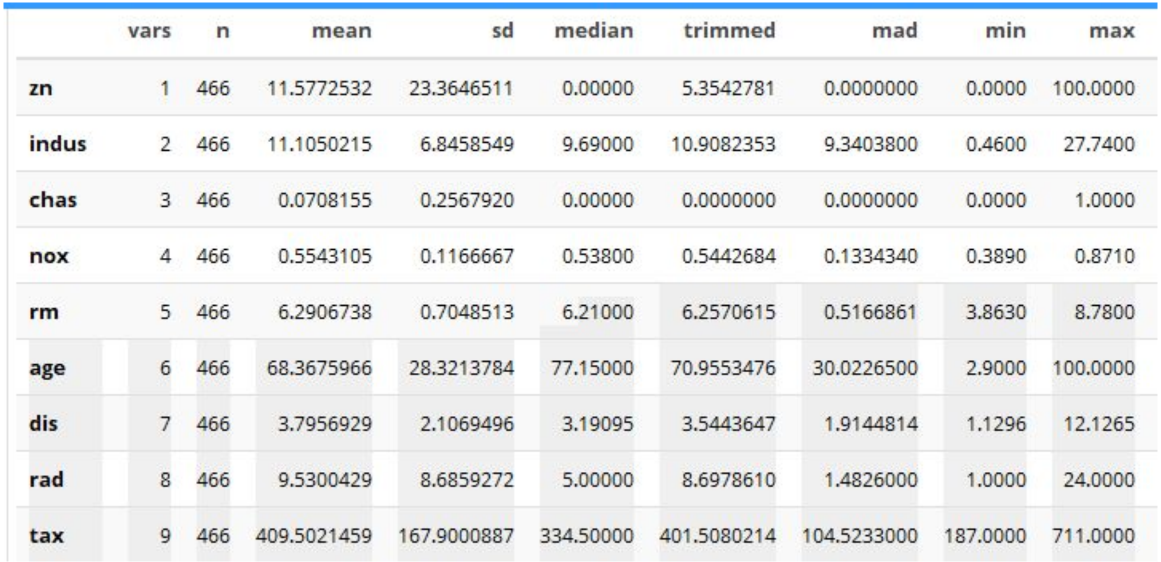
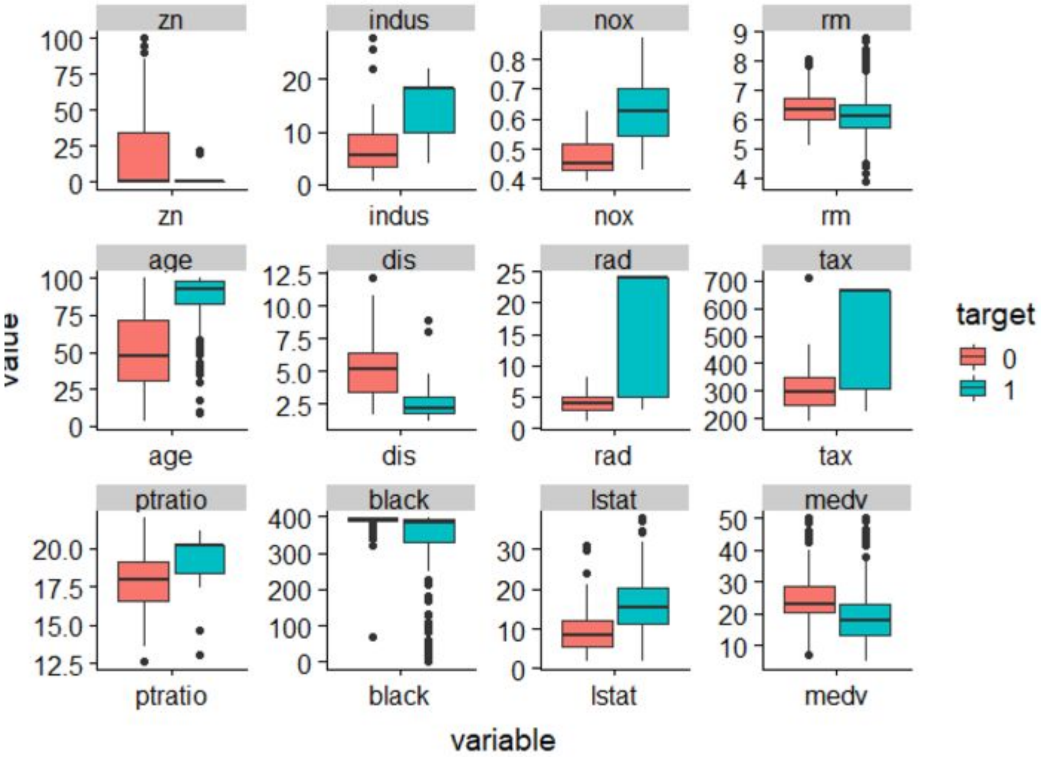
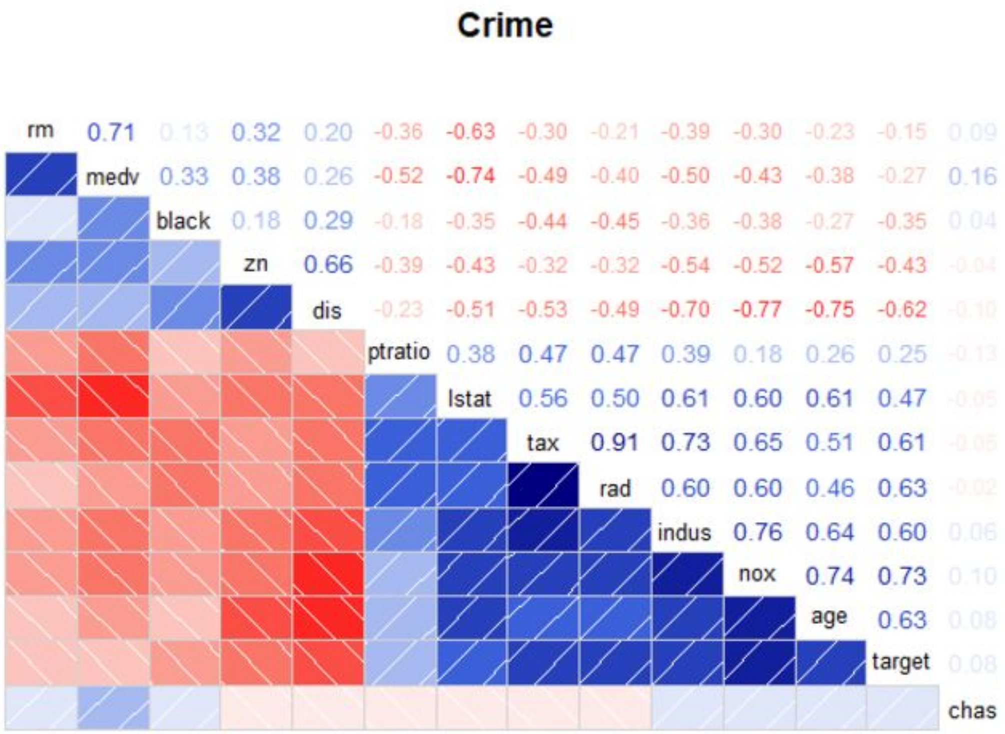
Data Exploration

The dataset contains 13 variables related to housing, transportation, the environment, education, and crime. For this data exploration, and the data has a target variable with a 1 representing an above average crime rate and a 0 representing below average crime rate and The target has an approximate mean of 0.5 and standard deviation of 0.5. In addition, there is no missing data point in this data set. Further more, from the box plot some variable like rad and tax has large affect to crime rate where higher value tend to lead a crime rate above median while variable like rm tend to have less effect to the target because the distribution are similar between group of high crime rate and low crime rate. It also Indicates that these variable may has less predict power. On the other hand the correlation plot. Shows that there are correlations between variables and rad and tax has highest correlation which is 0.91



	vars	n	mean	sd	median	trimmed	mad	min	max
zn	1	466	11.5772532	23.3646511	0.00000	5.3542781	0.0000000	0.0000	100.0000
indus	2	466	11.1050215	6.8458549	9.69000	10.9082353	9.3403800	0.4600	27.7400
chas	3	466	0.0708155	0.2567920	0.00000	0.0000000	0.0000000	0.0000	1.0000
nox	4	466	0.5543105	0.1166667	0.53800	0.5442684	0.1334340	0.3890	0.8710
rm	5	466	6.2906738	0.7048513	6.21000	6.2570615	0.5166861	3.8630	8.7800
age	6	466	68.3675966	28.3213784	77.15000	70.9553476	30.0226500	2.9000	100.0000
dis	7	466	3.7956929	2.1069496	3.19095	3.5443647	1.9144814	1.1296	12.1265
rad	8	466	9.5300429	8.6859272	5.00000	8.6978610	1.4826000	1.0000	24.0000
tax	9	466	409.5021459	167.9000887	334.50000	401.5080214	104.5233000	187.0000	711.0000





Data Preparation

We applied some transformations to the data. First of all, for variable with a skewed distribution like age we applied log transformation to the variable and. In addition, for variable with high variance like quadratic transformation

Build Models

First we build a model use all variables using original data set and from the result, we see only nine variables are statistically significant AIC for the model is 214.15. Then we build a model with all variables from original data set and transformed variables. The second model has a AIC equal to 213.33 and residual deviance equal to 177.33. Because the model applies transformation which reduced the variance, the residual deviance is smaller than before.

Then we build the third model using Backwards selection against all variables including transformed variables to find the best model and there are ten variables in this model finally. The model's AIC is 261.91 which is higher than previous two models but residual deviance is higher. *Finally we use* glmulti package to build another model against all variables including transformed variables. The model has seven variables and including two transformed variables. The model has a AIC equal to 230 but much lower residual deviance which is only 43

Model Selection

All models have similar AIC but the third model has a much lower residual deviance. So I think it should be more useful despite the fact the the AIC of this model is the highest among all models. Thus I decide to choose the last model.

Evaluating the model

We split the training data set in 20/80 and run the model against two data set. The result confusion matrices and accuracy are showing below. The model return a high 80% accuracy for both data set which is good.

```
##            PredictedValue
## ActualValue FALSE TRUE
##           0   138   37
##           1    12  176

## [1] 0.873

##            PredictedValue
## ActualValue FALSE TRUE
##           0    48   14
##           1     1   40

## [1] 0.894
```

Finally we run the model on the given evaluation dataset and the model predicts that there are 12 observations below the median crime rate, and about 28 above the median crime rate.

```
## predict12
##  0  1
## 12 28
```

Appendix

```
library(psych) library(readr) library(kableExtra) library(ggiraph) library(cowplot) library(reshape2)
library(corrgram) library(gridExtra) library(usdm) library(mice) library(pROC) library(reshape2) library(caTools)
library(caret) library(ROCR)

crime_train<-read_csv(“crime-training-data.csv”)

crime_eval<-read_csv(“crime-evaluation-data.csv”)

train <- describe(crime_train)

train$na_count <- sapply(crime_train, function(y)

sum(length(which(is.na(y)))))

kable(train, “html”, escape = F) %>%

kable_styling(“striped”, full_width = T) %>%

column_spec(1, bold = T) %>%

scroll_box(width = “100%”, height = “700px”)

long <- melt(crime_train, id.vars= “target”)%>%

filter(variable != “chas”) %>%

mutate(target = as.factor(target))

ggplot(data = long, aes(x = variable, y = value)) + geom_boxplot(aes(fill = target)) + facet_wrap( ~ variable,
scales = “free”)

crime_hist <- crime_train

crime_hist %>% keep(is.numeric) %>%
gather() %>%
ggplot(aes(value)) + facet_wrap(~ key, scales = “free”) + geom_histogram(bins = 35)

ggplot(crime_train, aes(crime_train$medv ,target)) + geom_point() + geom_smooth(method = “glm”,
method.args = list(family = “binomial”), se = FALSE)

kable(cor(drop_na(crime_train))[,14], “html”, escape = F) %>% kable_styling(“striped”, full_width = F) %>%
column_spec(1, bold = T) %>% scroll_box(height = “500px”)

corrgram(drop_na(crime_train), order=TRUE, upper.panel=panel.cor, main=“Moneyball”)
```

```
library(Amelia)

missmap(crime_eval, main = “Missing values vs observed”)

transform_crime <- crime_train

transform_crime$logage<−log(transformcrime$age)

transform_crime$loglstat<−log(transformcrime$lstat)

  transform_crime$quadzn<−transformcrime$zn^2

transform_crime$quadrad<−transformcrime$rad^2

crime_eval1 <- crime_eval

crime_eval1$logage<−log(crime_val$age)

crime_eval1loglstat<−log(crime_val$lstat)

crime_eval1$quadzn<−crime_val$zn^2

crime_eval1$quadrad<−crimeeval$rad^2

model1 <- glm(target ~., family = “binomial”, data=crime_train) summary(model1)

Model2 <- glm(target ~., family = “binomial”, data=transform_crime) summary(model3)

Model3 <- glm(target ~ indus + nox + rm + age + dis + tax + ptratio + black +medv + logage, family = “binomial”,
data=transform_crime) summary(model5)

library(rJava)

library(glmulti)

glmulti.lm.out <- glmulti(crime_train$target ~., data = crime_train, level = 1, # No interaction considered
method = “h”, # Exhaustive approach crit = “aic”, # AIC as criteria confsetsize = 5, # Keep 5 best models plotty
= F, report = F, # No plot or interim reports fitfunction = “lm”) # lm function

modelglmulti <- glm(transform_crime$target ~ nox + age + rad + ptratio + medv + logage + quadrad, data =
transform_crime)

summary(modelglmulti)

splitdata <- transform_crime

split <- sample.split(splitdata, SplitRatio = 0.8) split training <- subset(splitdata, split == “TRUE”) testing <-
subset(splitdata, split == “FALSE”)
```

```
modelglmulti3 <- glm(training$target ~ 1 + nox + age + rad + ptratio + medv, family=“binomial”, data = training)
res <- predict(modelglmulti3, newdata=training, type=“response”)

ROCRPred = prediction(res, training$target)

ROCRPref <- performance(ROCRPred, “tpr”,“fpr”)

plot(ROCRPref, colorize=TRUE, print.cutoffs.at=seq(0.1,by=0.1))

(table(ActualValue=training$target, PredictedValue=res>0.3)) round((149+167)/(149+167+9+37),3)

res <- predict(modelglmulti3, newdata=testing, type=“response”) (table(ActualValue=testing$target, PredictedValue=res>0.3)) round((42+51)/(42+51+9+2),3)

predict1 <- predict(modelglmulti, newdata=crime_eval1, type=“response”) predict12 <- ifelse(predict1 > 0.3, 1, 0) table(predict12) summary(predict1)

predict2 <- predict(model1, newdata=crime_eval1, type=“response”) summary(predict2) predict11 <- ifelse(predict2 > 0.5, 1, 0) table(predict11)
```