DATA 621HOMEWORK\_3: LOGISTIC REGRESSION

Warner Alexis

2024-11-01

## Objective

The objective is to build a **binary logistic regression model** on the training data set to predict whether the neighborhood will be at risk for high crime levels. Below is a short description of the variables of interest in the data set:

zn: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable) - indus: proportion of non-retail business acres per suburb (predictor variable) - chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable) - nox: nitrogen oxides concentration (parts per 10 million) (predictor variable) - rm: average number of rooms per dwelling (predictor variable) - age: proportion of owner-occupied units built prior to 1940 (predictor variable) - dis: weighted mean of distances to five Boston employment centers (predictor variable) - rad: index of accessibility to radial highways (predictor variable) - tax: full-value property-tax rate per $10,000 (predictor variable) - ptratio: pupil-teacher ratio by town (predictor variable) - lstat: lower status of the population (percent) (predictor variable) - medv: median value of owner-occupied homes in $1000s (predictor variable) - target: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

## Data Exploration

# Loading libraries   
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(corrplot)

## corrplot 0.94 loaded

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(ggplot2)  
library(kableExtra)

##   
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':  
##   
## group\_rows

## Including Plots

You can also embed plots, for example:

## 'data.frame': 466 obs. of 13 variables:  
## $ zn : num 0 0 0 30 0 0 0 0 0 80 ...  
## $ indus : num 19.58 19.58 18.1 4.93 2.46 ...  
## $ chas : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ nox : num 0.605 0.871 0.74 0.428 0.488 0.52 0.693 0.693 0.515 0.392 ...  
## $ rm : num 7.93 5.4 6.49 6.39 7.16 ...  
## $ age : num 96.2 100 100 7.8 92.2 71.3 100 100 38.1 19.1 ...  
## $ dis : num 2.05 1.32 1.98 7.04 2.7 ...  
## $ rad : int 5 5 24 6 3 5 24 24 5 1 ...  
## $ tax : int 403 403 666 300 193 384 666 666 224 315 ...  
## $ ptratio: num 14.7 14.7 20.2 16.6 17.8 20.9 20.2 20.2 20.2 16.4 ...  
## $ lstat : num 3.7 26.82 18.85 5.19 4.82 ...  
## $ medv : num 50 13.4 15.4 23.7 37.9 26.5 5 7 22.2 20.9 ...  
## $ target : int 1 1 1 0 0 0 1 1 0 0 ...

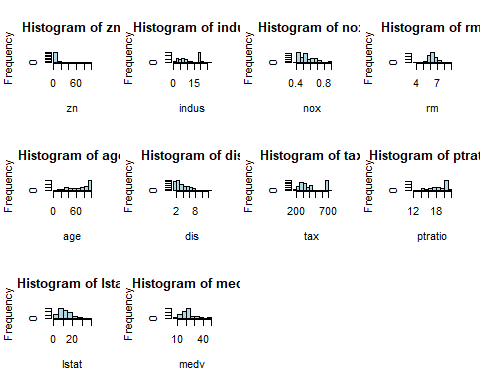
Data Summary

|  | vars | n | mean | sd | median | trimmed | mad |
| --- | --- | --- | --- | --- | --- | --- | --- |
| zn | 1 | 466 | 11.5772532 | 23.3646511 | 0.00000 | 5.3542781 | 0.0000000 |
| indus | 2 | 466 | 11.1050215 | 6.8458549 | 9.69000 | 10.9082353 | 9.3403800 |
| chas | 3 | 466 | 0.0708155 | 0.2567920 | 0.00000 | 0.0000000 | 0.0000000 |
| nox | 4 | 466 | 0.5543105 | 0.1166667 | 0.53800 | 0.5442684 | 0.1334340 |
| rm | 5 | 466 | 6.2906738 | 0.7048513 | 6.21000 | 6.2570615 | 0.5166861 |
| age | 6 | 466 | 68.3675966 | 28.3213784 | 77.15000 | 70.9553476 | 30.0226500 |
| dis | 7 | 466 | 3.7956929 | 2.1069496 | 3.19095 | 3.5443647 | 1.9144814 |
| rad | 8 | 466 | 9.5300429 | 8.6859272 | 5.00000 | 8.6978610 | 1.4826000 |
| tax | 9 | 466 | 409.5021459 | 167.9000887 | 334.50000 | 401.5080214 | 104.5233000 |
| ptratio | 10 | 466 | 18.3984979 | 2.1968447 | 18.90000 | 18.5970588 | 1.9273800 |
| lstat | 11 | 466 | 12.6314592 | 7.1018907 | 11.35000 | 11.8809626 | 7.0720020 |
| medv | 12 | 466 | 22.5892704 | 9.2396814 | 21.20000 | 21.6304813 | 6.0045300 |
| target | 13 | 466 | 0.4914163 | 0.5004636 | 0.00000 | 0.4893048 | 0.0000000 |

Data Summary (Cont)

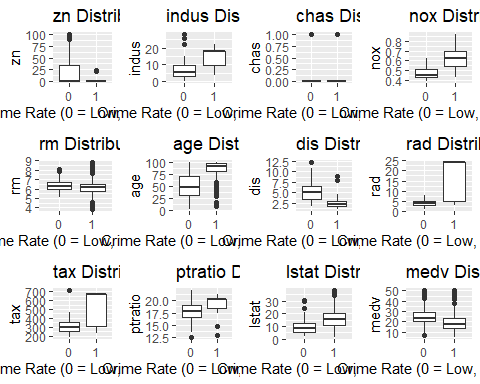
|  | min | max | range | skew | kurtosis | se |
| --- | --- | --- | --- | --- | --- | --- |
| zn | 0.0000 | 100.0000 | 100.0000 | 2.1768152 | 3.8135765 | 1.0823466 |
| indus | 0.4600 | 27.7400 | 27.2800 | 0.2885450 | -1.2432132 | 0.3171281 |
| chas | 0.0000 | 1.0000 | 1.0000 | 3.3354899 | 9.1451313 | 0.0118957 |
| nox | 0.3890 | 0.8710 | 0.4820 | 0.7463281 | -0.0357736 | 0.0054045 |
| rm | 3.8630 | 8.7800 | 4.9170 | 0.4793202 | 1.5424378 | 0.0326516 |
| age | 2.9000 | 100.0000 | 97.1000 | -0.5777075 | -1.0098814 | 1.3119625 |
| dis | 1.1296 | 12.1265 | 10.9969 | 0.9988926 | 0.4719679 | 0.0976026 |
| rad | 1.0000 | 24.0000 | 23.0000 | 1.0102788 | -0.8619110 | 0.4023678 |
| tax | 187.0000 | 711.0000 | 524.0000 | 0.6593136 | -1.1480456 | 7.7778214 |
| ptratio | 12.6000 | 22.0000 | 9.4000 | -0.7542681 | -0.4003627 | 0.1017669 |
| lstat | 1.7300 | 37.9700 | 36.2400 | 0.9055864 | 0.5033688 | 0.3289887 |
| medv | 5.0000 | 50.0000 | 45.0000 | 1.0766920 | 1.3737825 | 0.4280200 |
| target | 0.0000 | 1.0000 | 1.0000 | 0.0342293 | -2.0031131 | 0.0231835 |

# Plot the distribution   
  
numeric\_vars <- c("zn", "indus", "nox", "rm", "age", "dis", "tax", "ptratio", "lstat", "medv")  
par(mfrow=c(3,4))  
for (var in numeric\_vars) {  
 hist(crime\_data[[var]], main = paste('Histogram of', var), xlab= var, col= 'lightblue')  
}



The box plots show the distribution of various predictor variables for areas with low (0) and high (1) crime rates, offering insights into how each variable relates to crime. In areas with low crime rates, the zn variable (proportion of residential land) has a higher median and wider range, suggesting more residential zoning in lower-crime areas. In contrast, high-crime areas tend to have higher values of indus (non-retail business acreage), nox (nitrogen oxide concentration), age (older properties), rad (proximity to highways), and tax (property taxes). This indicates that industrial zones, pollution, older buildings, highway access, and higher taxes may correlate with higher crime. Additionally, rm (average rooms per dwelling) and medv (median property value) have higher medians in low-crime areas, suggesting that larger residences and higher property values are associated with lower crime rates. The dis variable (distance to employment centers) has a lower median for high-crime areas, implying that areas closer to job hubs may have more crime. Furthermore, the ptratio (pupil-teacher ratio) and lstat (percentage of lower socioeconomic status individuals) are slightly higher in high-crime areas, hinting at possible links between socioeconomic factors, educational resources, and crime. Overall, these plots reveal that lower crime areas tend to have more residential zoning, larger homes, and higher property values, while high-crime areas are associated with industrial zones, pollution, older properties, highway access, and socioeconomic challenges.

# Load necessary libraries  
library(ggplot2)  
  
  
# Define a list of all predictor variables  
predictor\_vars <- c("zn", "indus", "chas", "nox", "rm", "age", "dis",   
 "rad", "tax", "ptratio", "lstat", "medv")  
  
# Create a list to store each plot  
plots <- list()  
  
# Generate boxplots for each predictor variable against the target  
for (var in predictor\_vars) {  
 p <- ggplot(crime\_data, aes(x = as.factor(target), y = .data[[var]])) +  
 geom\_boxplot() +  
 labs(x = "Crime Rate (0 = Low, 1 = High)", y = var) +  
 ggtitle(paste(var, "Distribution by Crime Rate"))  
 plots[[var]] <- p # Add each plot to the list  
}  
  
# Arrange plots in a 3x4 grid  
grid.arrange(grobs = plots, ncol = 4, nrow = 3)



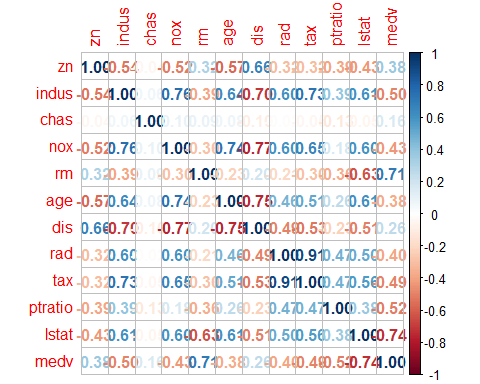
Strong positive correlations are observed between tax and rad (0.91), as well as between indus and nox (0.76). These suggest that areas with high tax rates tend to have better highway accessibility (rad), and industrial zones (indus) correlate with higher nitrogen oxide levels (nox).

Negative correlations are seen between dis and variables like nox (-0.77) and rad (-0.71), indicating that neighborhoods farther from employment centers tend to have lower nitrogen oxide levels and less highway access.

The correlations between each predictor variable and the target variable (whether the crime rate is above or below the median) reveal significant patterns in neighborhood characteristics associated with crime. A strong positive correlation with crime rate is seen for nox (nitrogen oxide concentration, 0.73), indus (proportion of industrial land, 0.60), age (proportion of older homes, 0.63), and rad (access to radial highways, 0.63). These values suggest that neighborhoods with higher pollution levels, more industrial areas, older housing stock, and better highway accessibility tend to have higher crime rates.

Conversely, there are negative correlations for variables such as zn (proportion of large residential lots, -0.43), dis (distance from employment centers, -0.62), and medv (median home value, -0.27). This implies that neighborhoods with larger residential lots, greater distances from employment centers, and higher home values generally experience lower crime rates. Moderate positive correlations are found for tax (property tax rate, 0.61) and lstat (lower-status population percentage, 0.47), indicating that areas with higher tax rates and a higher proportion of lower-status residents also tend to have increased crime rates. Other variables like chas (proximity to the Charles River, 0.08) and rm (average number of rooms per dwelling, -0.15) show weak correlations, suggesting a limited relationship with crime rate. These correlation values help identify influential features that may be predictive of crime levels in neighborhoods.

# Correlation matrix and correlation with target  
corr\_matrix <- cor(crime\_data[, -which(names(crime\_data) == "target")])  
corrplot(corr\_matrix, method = "number")



# Correlation of each variable with target variable  
cor\_with\_target <- sapply(crime\_data[, -which(names(crime\_data) == "target")], function(x) cor(x, crime\_data$target))  
print(cor\_with\_target)

## zn indus chas nox rm age   
## -0.43168176 0.60485074 0.08004187 0.72610622 -0.15255334 0.63010625   
## dis rad tax ptratio lstat medv   
## -0.61867312 0.62810492 0.61111331 0.25084892 0.46912702 -0.27055071

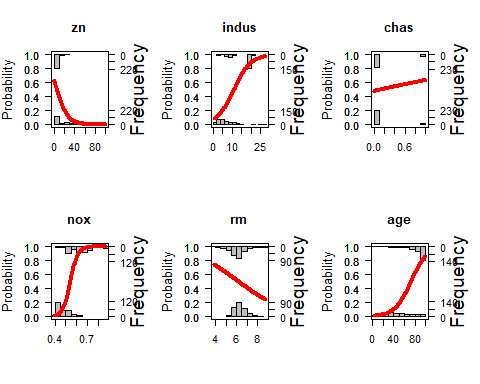
This set of binary plots illustrates the relationship between several predictor variables and the probability of high crime rates in neighborhoods. Each panel represents a different predictor variable and its influence on the likelihood of high crime, with a logistic curve (in red) showing the estimated probability. The variables dis (distance to employment centers) and medv (median home value) both show a negative relationship with crime probability, suggesting that neighborhoods further from employment centers and with higher home values are less likely to have high crime rates. Conversely, rad (access to radial highways), tax (property tax rate), ptratio (pupil-teacher ratio), and lstat (lower status of the population) are positively associated with crime probability. This implies that neighborhoods with better highway access, higher property taxes, lower socioeconomic status, and fewer educational resources are more prone to higher crime rates. Overall, these plots provide a visual summary of how each predictor variable impacts the likelihood of high crime rates, highlighting socioeconomic and accessibility factors as key contributors.

library(popbio)

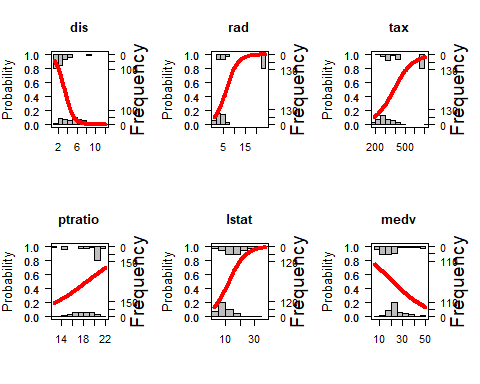
##   
## Attaching package: 'popbio'

## The following object is masked from 'package:caret':  
##   
## sensitivity

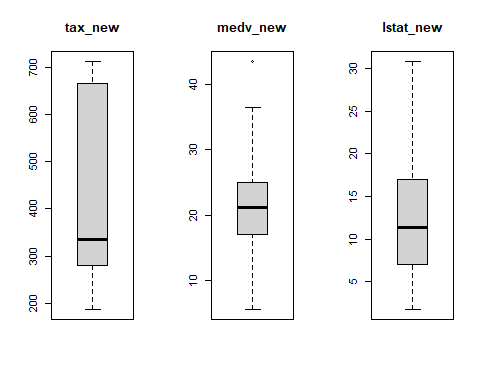
# Relationship are not linear   
x<-crime\_data[,]  
  
par(mfrow=c(2,3))  
logi.hist.plot(x$zn,x$target,logi.mod = 1, type="p", boxp=FALSE,col="gray", mainlabel = "zn")  
logi.hist.plot(x$indus, x$target,logi.mod = 1, type="hist",boxp=FALSE,col="gray", mainlabel = "indus")  
logi.hist.plot(x$chas,x$target,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "chas")  
logi.hist.plot(x$nox, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "nox")  
logi.hist.plot(x$rm,x$target,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "rm")  
logi.hist.plot(x$age, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "age")



logi.hist.plot(x$dis,x$target,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "dis")  
logi.hist.plot(x$rad, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "rad")  
logi.hist.plot(x$tax,x$target,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "tax")  
logi.hist.plot(x$ptratio, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "ptratio")  
#logi.hist.plot(x$black,x$target,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "black")  
logi.hist.plot(x$lstat, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "lstat")  
logi.hist.plot(x$medv, x$target,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "medv")



# remove Outliers   
treat\_outliers <- function(x) {  
 qnt <- quantile(x, probs=c(.25, .75), na.rm = T)  
 caps <- quantile(x, probs=c(.05, .95), na.rm = T)  
 H <- 1.5 \* IQR(x, na.rm = T)  
 x[x < (qnt[1] - H)] <- caps[1]  
 x[x > (qnt[2] + H)] <- caps[2]  
   
 return(x)  
}  
  
crime\_data\_train<-crime\_data  
train\_test <- read.csv('crime-evaluation-data\_modified.csv') # load evaulation/test set   
train\_test\_mod<-train\_test # chamging name to test set   
  
crime\_data\_train$tax\_new <- treat\_outliers(crime\_data$tax) # remove outliers for tax  
crime\_data\_train$medv\_new <- treat\_outliers(crime\_data$medv) # remove outliers for medv  
crime\_data\_train$lstat\_new <- treat\_outliers(crime\_data$lstat)# # remove outliers for lsatt  
  
# do the same for test set   
train\_test\_mod$tax\_new <- treat\_outliers(train\_test$tax) # remove outliers for tax  
train\_test\_mod$medv\_new <- treat\_outliers(train\_test$medv)# remove outliers for medv  
train\_test\_mod$lstat\_new <- treat\_outliers(train\_test$lstat) # remove outliers for lsatt  
  
# plot the newly made variables with no outliers   
par(mfrow=c(1,3))  
# Box plot to distributetions and mean values   
boxplot(crime\_data\_train$tax\_new,main="tax\_new")   
boxplot(crime\_data\_train$medv\_new,main="medv\_new")  
boxplot(crime\_data\_train$lstat\_new,main="lstat\_new")

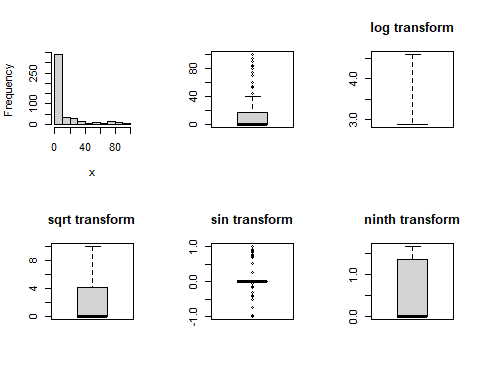


These plots show a series of data transformations applied to an original variable (shown in the top left histogram) to explore its effect on the distribution. Each transformation can impact the spread and symmetry of the data, potentially helping to improve model performance or meet assumptions of normality.

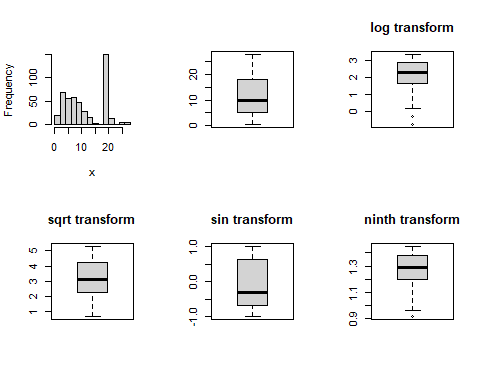
1. **Original Data (Top Left)**: Shows the frequency distribution of the raw data, which appears skewed to the right.
2. **Log Transform**: Compresses the data range, especially for larger values, helping to reduce skewness. This can be useful when data spans multiple orders of magnitude.
3. **Square Root Transform**: Similar to the log transform, but less aggressive. It can moderate right-skewed data while preserving data order.
4. **Sine Transform**: Applies the sine function to the data. This is less common for normalization, but may be useful for cyclic or periodic data.
5. **Ninth Root Transform**: A specific type of root transformation that reduces the influence of large values, though less drastically than the square root.

Each transformation changes the data distribution, which can help in reducing skewness, stabilizing variance, or meeting other modeling requirements.

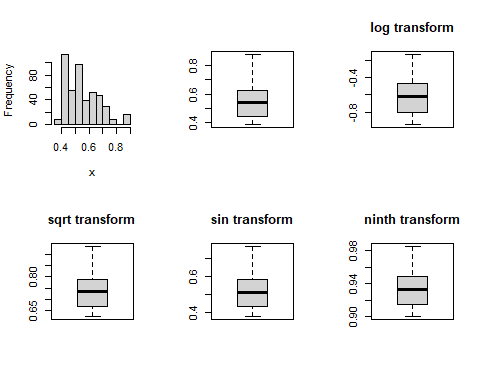
#Checking how different transformation works on this Data set  
  
crime\_data\_train$rm\_new<-sin(crime\_data$rm)  
  
crime\_data\_train$dis\_new<-sin(crime\_data$dis)  
  
train\_test\_mod$rm\_new<-sin(train\_test$rm)  
  
train\_test\_mod$dis\_new<-sin(train\_test$dis)  
  
  
par(mfrow=c(1,2))  
  
#boxplot(city\_crime\_train\_mod$rm\_new,main="rm\_new")  
#boxplot(city\_crime\_train\_mod$dis\_new,main="dis\_new")  
show\_charts <- function(x, ...) {  
   
 par(mfrow=c(2,3))  
   
 xlabel <- unlist(str\_split(deparse(substitute(x)), pattern = "\\$"))[2]  
 ylabel <- unlist(str\_split(deparse(substitute(y)), pattern = "\\$"))[2]  
   
 hist(x,main=xlabel)  
 boxplot(x,main=xlabel)  
   
 y<-log(x)  
 boxplot(y,main='log transform')  
 y<-sqrt(x)  
 boxplot(y,main='sqrt transform')  
 y<-sin(x)  
 boxplot(y,main='sin transform')  
 y<-(x)^(1/9)  
 boxplot(y,main='ninth transform')  
   
}  
  
library(stringr)  
  
attach(crime\_data\_train)  
  
show\_charts(zn)



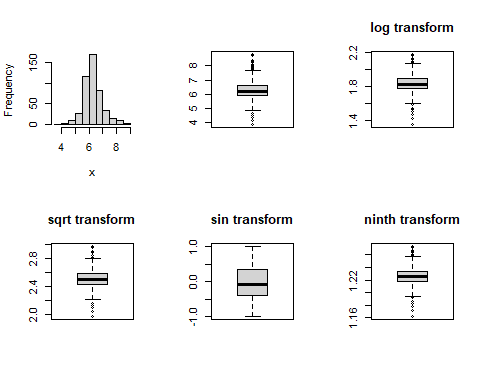
show\_charts(indus)



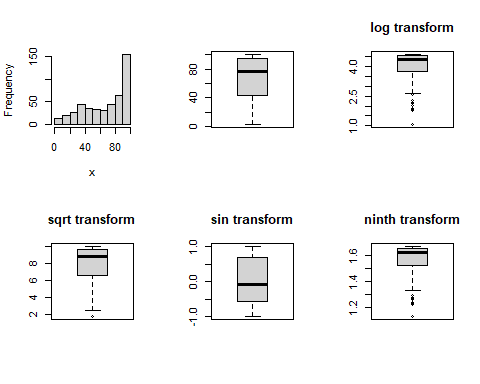
show\_charts(nox)



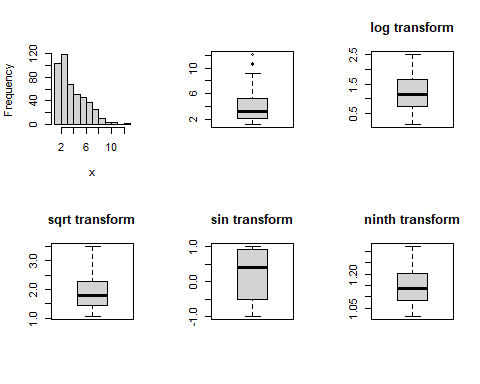
show\_charts(rm)



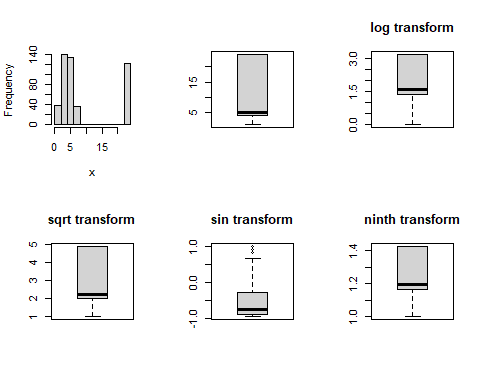
show\_charts(age)



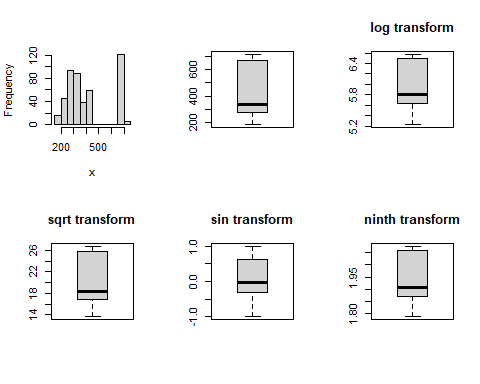
show\_charts(dis)



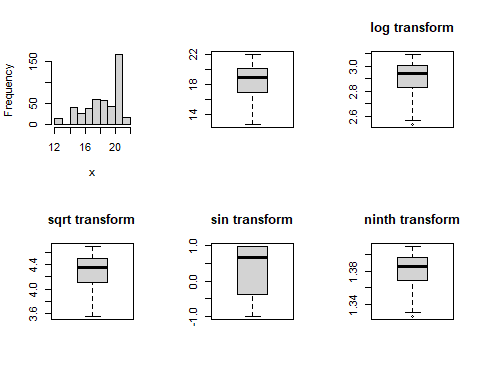
show\_charts(rad)



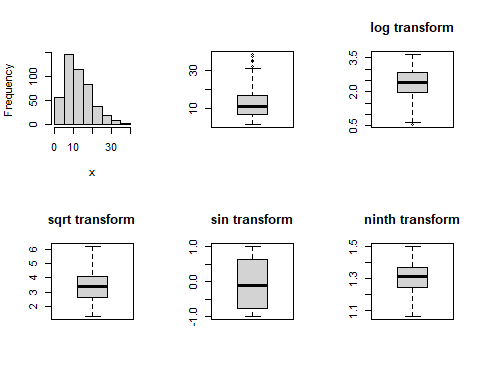
show\_charts(tax)



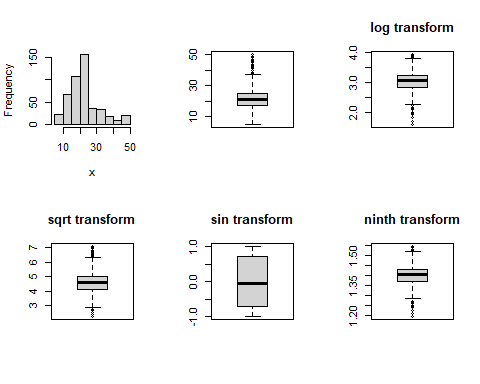
show\_charts(ptratio)



show\_charts(lstat)



show\_charts(medv)



We create **Ratio and Interations** combining variables such as tax with ptratio (as a measure of affordability) or rm and age (to capture older, larger homes) can create meaningful new variables. These interaction terms can capture unique relationships between predictors and crime rate. We normalize variables with skewed distributions. The log or square root transformations can help normalize the data, making it more suitable for logistic regression

crime\_data <- crime\_data\_train  
# Creating Categories   
crime\_data$lstat\_log <- log(crime\_data$lstat +1)  
crime\_data$tax\_log <- log(crime\_data$tax + 1)  
  
train\_test$lstat\_log <- log(train\_test$lstat +1)  
train\_test$tax\_log <- log(train\_test$tax + 1)  
# Creating new ratio variables  
crime\_data$tax\_ptratio\_ratio <- crime\_data$tax / crime\_data$ptratio  
crime\_data$rm\_age\_interaction <- crime\_data$rm \* crime\_data$age  
  
train\_test$tax\_ptratio\_ratio <- train\_test$tax / train\_test$ptratio  
train\_test$rm\_age\_interaction <- train\_test$rm \* train\_test$age  
  
  
  
# Bucketing age into categories  
crime\_data$age\_group <- cut(crime\_data$age, breaks = c(0, 35, 70, 100), labels = c("Low", "Medium", "High"))  
train\_test$age\_group <- cut(train\_test$age, breaks = c(0, 35, 70, 100), labels = c("Low", "Medium", "High"))  
# Bucketing dis into categories  
crime\_data$dis\_group <- cut(crime\_data$dis, breaks = quantile(crime\_data$dis, probs = seq(0, 1, 0.33)), include.lowest = TRUE)  
train\_test$dis\_group <- cut(train\_test$dis, breaks = quantile(train\_test$dis, probs = seq(0, 1, 0.33)), include.lowest = TRUE)  
## Standardizing selected variables  
crime\_data <- crime\_data %>%  
 mutate(across(c(zn, indus, nox, rm, age, dis, rad, tax, ptratio, lstat, medv), scale))  
  
train\_test <- train\_test %>%  
 mutate(across(c(zn, indus, nox, rm, age, dis, rad, tax, ptratio, lstat, medv), scale))  
  
head(crime\_data)

## zn indus chas nox rm age dis  
## 1 -0.4955029 1.2379723 0 0.4344813 2.3243572 0.9827348 -0.8304864  
## 2 -0.4955029 1.2379723 1 2.7144813 -1.2593775 1.1169090 -1.1742535  
## 3 -0.4955029 1.0217831 0 1.5916242 0.2756981 1.1169090 -0.8625232  
## 4 0.7884880 -0.9020088 0 -1.0826616 0.1451741 -2.1385822 1.5376766  
## 5 -0.4955029 -1.2628111 0 -0.5683758 1.2262532 0.8414987 -0.5197528  
## 6 -0.4955029 -0.3717609 0 -0.2940901 0.6956449 0.1035403 -0.4459494  
## rad tax ptratio lstat medv target tax\_new  
## 1 -0.5215382 -0.03872628 -1.6835500 -1.2576171 2.9666315 1 403  
## 2 -0.5215382 -0.03872628 -1.6835500 1.9978540 -0.9945441 1 403  
## 3 1.6659082 1.52768147 0.8200407 0.8756176 -0.7780864 1 666  
## 4 -0.4064095 -0.65218635 -0.8186732 -1.0478138 0.1202130 0 300  
## 5 -0.7517957 -1.28947011 -0.2724352 -1.0999126 1.6570625 0 193  
## 6 -0.5215382 -0.15188882 1.1386795 -0.6986110 0.4232537 0 384  
## medv\_new lstat\_new rm\_new dis\_new lstat\_log tax\_log tax\_ptratio\_ratio  
## 1 43.4 3.70 0.9971874 0.8892453 1.547563 6.001415 27.41497  
## 2 13.4 26.82 -0.7708569 0.9691109 3.325755 6.001415 27.41497  
## 3 15.4 18.85 0.2004475 0.9180734 2.988204 6.502790 32.97030  
## 4 23.7 5.19 0.1095941 0.6833306 1.822935 5.707110 18.07229  
## 5 43.4 4.82 0.7654978 0.4268374 1.761300 5.267858 10.84270  
## 6 26.5 7.67 0.4775066 0.2816302 2.159869 5.953243 18.37321  
## rm\_age\_interaction age\_group dis\_group  
## 1 762.7698 High [1.13,2.37]  
## 2 540.3000 High [1.13,2.37]  
## 3 648.5000 High [1.13,2.37]  
## 4 49.8654 Low (4.25,9.7]  
## 5 659.6910 High (2.37,4.25]  
## 6 483.4853 High (2.37,4.25]

head(train\_test)

## zn indus chas nox rm age dis  
## 1 -0.3864190 -0.6244111 0 -0.8386288 1.4261128 -0.3764993 0.55698900  
## 2 -0.3864190 -0.4738235 0 -0.1969278 -0.1737603 0.5143080 0.31844789  
## 3 -0.3864190 -0.4738235 0 -0.1969278 0.4124191 0.8911880 0.31504826  
## 4 -0.3864190 -0.4738235 0 -0.1969278 -0.3882520 0.4191362 0.09563010  
## 5 -0.3864190 -0.7806283 0 -0.5596284 -0.5351642 -1.1226456 0.06928292  
## 6 0.7020852 -0.8974393 0 -0.9874290 -0.6952984 -0.1823490 1.62329420  
## rad tax ptratio lstat medv lstat\_log tax\_log  
## 1 -0.8579380 -0.8543307 -0.80987496 -1.15653394 1.46235907 1.615420 5.493061  
## 2 -0.6372466 -0.4877862 1.15345828 -0.34465682 -0.41903856 2.421257 5.730100  
## 3 -0.6372466 -0.4877862 1.15345828 -0.01365074 -0.39623374 2.624669 5.730100  
## 4 -0.6372466 -0.4877862 1.15345828 1.92938102 -0.98915906 3.357245 5.730100  
## 5 -0.5269009 -0.6456823 0.04908333 -0.53882968 -0.09977109 2.279316 5.634790  
## 6 -0.1958637 -0.6174866 0.35585415 0.03196033 -0.36202651 2.649715 5.652489  
## tax\_ptratio\_ratio rm\_age\_interaction age\_group dis\_group  
## 1 13.59551 439.0035 Medium (3.98,9.02]  
## 2 14.61905 515.1120 High (3.98,9.02]  
## 3 14.61905 613.1280 High (3.98,9.02]  
## 4 14.61905 487.9000 High (3.98,9.02]  
## 5 14.53125 242.7750 Medium (2.38,3.98]  
## 6 14.41624 380.0542 Medium (3.98,9.02]

## Model Development

**MODEL 1**

The Analysis of Deviance table provides insights into the contribution of each predictor variable to the logistic regression model predicting crime rates. The results show that indus (proportion of industrial land) and nox (nitrogen oxide concentration) significantly reduce the deviance, with highly significant p-values (< 2.2e-16), indicating that these are strong predictors of crime rate. Adding rad (access to radial highways) also significantly improves the model, reducing deviance by 58.076 with a p-value of 2.522e-14, suggesting that highway accessibility is relevant in predicting higher crime rates. Similarly, tax (property tax rate) further improves the model with a deviance reduction of 10.184 and a significant p-value of 0.001416, indicating a meaningful association with crime. In contrast, variables like age (proportion of older homes) and dis (distance to employment centers) show high p-values and minimal reductions in deviance, implying that they do not add significant explanatory power to the model once other variables are included. Overall, the results suggest that crime rates are closely associated with industrial land, pollution levels, highway access, and property tax rates, while factors such as housing age and proximity to employment centers contribute less to the model.

set.seed(123)  
# Model Development  
m1 <- glm(target ~ indus + nox + age + dis + age + rad + tax, data = crime\_data, family = 'binomial')  
# Summary of the model   
summary(m1)

##   
## Call:  
## glm(formula = target ~ indus + nox + age + dis + age + rad +   
## tax, family = "binomial", data = crime\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.94477 -0.26091 -0.02967 0.00597 2.79697   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.5155 0.5973 4.211 2.54e-05 \*\*\*  
## indus -0.3565 0.3164 -1.127 0.25982   
## nox 4.6850 0.8341 5.617 1.95e-08 \*\*\*  
## age 0.6216 0.2740 2.269 0.02328 \*   
## dis 0.5096 0.3294 1.547 0.12188   
## rad 5.3456 1.0992 4.863 1.16e-06 \*\*\*  
## tax -1.3018 0.4356 -2.988 0.00281 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 645.88 on 465 degrees of freedom  
## Residual deviance: 216.88 on 459 degrees of freedom  
## AIC: 230.88  
##   
## Number of Fisher Scoring iterations: 8

# Deviance analyis   
anova(m1, test = 'Chi') # use to ananlyse deviance in all variables

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: target  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 465 645.88   
## indus 1 192.641 464 453.23 < 2.2e-16 \*\*\*  
## nox 1 165.125 463 288.11 < 2.2e-16 \*\*\*  
## age 1 1.421 462 286.69 0.233317   
## dis 1 1.553 461 285.14 0.212682   
## rad 1 58.076 460 227.06 2.522e-14 \*\*\*  
## tax 1 10.184 459 216.88 0.001416 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Odd Ratio   
s <- c("indus" , "nox" , "age" , "dis" , "age" , "rad" , "tax")  
or\_m1 <- exp(coef(m1)[s])  
print(or\_m1)

## indus nox age dis age rad   
## 0.7000880 108.3076344 1.8618580 1.6646041 1.8618580 209.6889640   
## tax   
## 0.2720426

step\_m1 <- step(m1, direction = 'backward')

## Start: AIC=230.88  
## target ~ indus + nox + age + dis + age + rad + tax  
##   
## Df Deviance AIC  
## - indus 1 218.17 230.17  
## <none> 216.88 230.88  
## - dis 1 219.25 231.25  
## - age 1 222.23 234.23  
## - tax 1 227.06 239.06  
## - nox 1 273.04 285.04  
## - rad 1 276.24 288.24  
##   
## Step: AIC=230.17  
## target ~ nox + age + dis + rad + tax  
##   
## Df Deviance AIC  
## <none> 218.17 230.17  
## - dis 1 220.44 230.44  
## - age 1 223.30 233.30  
## - tax 1 234.17 244.17  
## - nox 1 278.97 288.97  
## - rad 1 284.67 294.67

summary(step\_m1)

##   
## Call:  
## glm(formula = target ~ nox + age + dis + rad + tax, family = "binomial",   
## data = crime\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.84242 -0.28115 -0.03076 0.00521 2.74702   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.6043 0.5958 4.371 1.24e-05 \*\*\*  
## nox 4.2210 0.6886 6.130 8.78e-10 \*\*\*  
## age 0.5982 0.2685 2.228 0.025902 \*   
## dis 0.4937 0.3260 1.514 0.129987   
## rad 5.6934 1.0669 5.336 9.48e-08 \*\*\*  
## tax -1.4404 0.4049 -3.557 0.000375 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 645.88 on 465 degrees of freedom  
## Residual deviance: 218.17 on 460 degrees of freedom  
## AIC: 230.17  
##   
## Number of Fisher Scoring iterations: 8

exp(coef(m1)[s])/(1 + exp(coef(m1)[s]))

## indus nox age dis age rad tax   
## 0.4117952 0.9908515 0.6505767 0.6247097 0.6505767 0.9952537 0.2138628

R performs a hypothesis test for each coefficient, that is, versus $β\_j for j=0,1 $ via the Wald test, and print the in the last column. We can thus compare these to the chosen significance level (usually ) to conclude whether or not each of the coefficient is significantly different from 0.

* When X and Y are independent
* When $\_1 > 0 $, The probability That Y = 1 increase with X, and
* When $\_1 < 0 $,the probability that Y = 1 decreases with X

Which means in our context: - When the probability of having crime is independent of the predictors we are comparing,

* When $\_1 > 0 $, the probability of having crime increases with the the predictors, and
* When $\_1 < 0 $,the probability of having crime decreases with the the predictors.

The result shows the probability of a higher crime rate (target = 1) when each predictor variable (indus, nox, age, dis, age, rad, and tax) is evaluated individually with all other predictors set to zero in the logistic regression model.

The probabilities indicate how each predictor alone influences the likelihood of a high crime rate. nox (99.1%) and rad (99.5%) show strong positive associations with crime, suggesting pollution and highway access are linked to higher crime rates. age (65.1%) and dis (62.5%) have moderate associations, indicating older properties and proximity to job centers also relate to crime. indus (41.2%) shows a weaker, moderate association, while tax (21.4%) is inversely related to crime, suggesting that higher-taxed areas tend to have lower crime rates.

This logistic regression model examines the relationship between crime rate and predictors such as nox (pollution), age (property age), dis (distance to employment centers), rad (highway access), and tax (property tax) in crime\_data. Significant predictors (p < 0.05) include nox, age, rad, and tax. Higher nox and rad values are associated with an increased likelihood of higher crime rates, while higher tax reduces this likelihood. The age of properties also slightly increases the likelihood of higher crime. The non-significant dis variable suggests that proximity to job centers is not strongly linked to crime rates in this model. The model reduces deviance from 645.88 to 218.17, indicating that these predictors explain a substantial portion of the variation in crime rates. An AIC of 230.17 suggests a reasonable fit.

**MODEL 2**

# Model 2   
  
crime\_data\_train\_scaled <- crime\_data[,1:22]  
s <- crime\_data[,1:22] %>%  
 mutate(across(where(is.numeric) & !c(target, chas), scale))  
m2 <- glm(target ~ indus + tax\_new + medv\_new + dis\_new + tax\_log , family = 'binomial', data = s)  
summary(m2)

##   
## Call:  
## glm(formula = target ~ indus + tax\_new + medv\_new + dis\_new +   
## tax\_log, family = "binomial", data = s)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9689 -0.6879 -0.3159 0.4436 2.1566   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.10144 0.13830 0.733 0.4633   
## indus 1.02716 0.18819 5.458 4.81e-08 \*\*\*  
## tax\_new -0.30253 0.95254 -0.318 0.7508   
## medv\_new 0.37613 0.16506 2.279 0.0227 \*   
## dis\_new -0.06872 0.14421 -0.477 0.6337   
## tax\_log 1.56350 0.90356 1.730 0.0836 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 645.88 on 465 degrees of freedom  
## Residual deviance: 406.37 on 460 degrees of freedom  
## AIC: 418.37  
##   
## Number of Fisher Scoring iterations: 5

# Odd Ratio   
s2 <- c("indus" , "tax\_new" , "medv\_new" , "dis\_new " , "tax\_log" )  
or\_m2 <- exp(coef(m2)[s2])  
print(or\_m2)

## indus tax\_new medv\_new <NA> tax\_log   
## 2.7931151 0.7389498 1.4566358 NA 4.7755184

step\_m2 <- step(m2, direction = 'backward')

## Start: AIC=418.37  
## target ~ indus + tax\_new + medv\_new + dis\_new + tax\_log  
##   
## Df Deviance AIC  
## - tax\_new 1 406.47 416.47  
## - dis\_new 1 406.60 416.60  
## <none> 406.37 418.37  
## - tax\_log 1 409.49 419.49  
## - medv\_new 1 411.73 421.73  
## - indus 1 440.15 450.15  
##   
## Step: AIC=416.47  
## target ~ indus + medv\_new + dis\_new + tax\_log  
##   
## Df Deviance AIC  
## - dis\_new 1 406.70 414.70  
## <none> 406.47 416.47  
## - medv\_new 1 411.77 419.77  
## - indus 1 441.23 449.23  
## - tax\_log 1 448.96 456.96  
##   
## Step: AIC=414.7  
## target ~ indus + medv\_new + tax\_log  
##   
## Df Deviance AIC  
## <none> 406.70 414.70  
## - medv\_new 1 411.84 417.84  
## - indus 1 441.94 447.94  
## - tax\_log 1 452.70 458.70

summary(step\_m2)

##   
## Call:  
## glm(formula = target ~ indus + medv\_new + tax\_log, family = "binomial",   
## data = s)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9778 -0.7032 -0.3498 0.4274 2.1530   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.1172 0.1302 0.901 0.3678   
## indus 0.9864 0.1704 5.788 7.12e-09 \*\*\*  
## medv\_new 0.3654 0.1633 2.237 0.0253 \*   
## tax\_log 1.2529 0.1901 6.589 4.42e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 645.88 on 465 degrees of freedom  
## Residual deviance: 406.70 on 462 degrees of freedom  
## AIC: 414.7  
##   
## Number of Fisher Scoring iterations: 5

exp(coef(m2)[s2])/(1 + exp(coef(m2)[s2]))

## indus tax\_new medv\_new <NA> tax\_log   
## 0.7363644 0.4249403 0.5929393 NA 0.8268554

The logistic regression model summary shows the estimated coefficients and significance levels for the predictors indus, medv\_new, and tax\_log, along with the intercept. The coefficient for indus is 0.9864 (p < 0.001), indicating a strong, positive association with the likelihood of high crime levels. Similarly, tax\_log has a high positive coefficient of 1.2529 (p < 0.001), showing a significant influence on increasing crime probability. The variable medv\_new has a smaller positive effect (0.3654, p < 0.05), indicating a weaker but still statistically significant association. The model’s AIC is 414.7, suggesting a balance between goodness of fit and model complexity. Overall, the model significantly reduces deviance from the null (645.88) to residual (406.70), indicating a better fit with these predictors.

**Model 3**

The stepwise selection process using AIC (Akaike Information Criterion) has iteratively removed and added predictors to find the most parsimonious model for predicting the target variable. Starting from an initial model with multiple predictors, the stepwise selection method aimed to minimize the AIC, a measure balancing model fit and complexity. At each step, variables were assessed for their contribution to the model, with the least significant predictors being removed. The process stopped at an AIC of 157.95, suggesting this model best balances fit and simplicity. Key variables that remain in the model include indus, nox, rad, tax, ptratio, lstat, medv, rm\_new, dis\_new, lstat\_log, tax\_log, tax\_ptratio\_ratio, and rm\_age\_interaction. These variables were retained as they contribute meaningful information for predicting the target. Additionally, no further improvements in AIC were possible by adding or removing individual variables, suggesting that this model is near optimal in terms of predictive power relative to complexity.

m3 <- stepAIC(glm(target ~ ., data = s, family = "binomial"), direction = "both")

## Start: AIC=169.21  
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +   
## ptratio + lstat + medv + tax\_new + medv\_new + lstat\_new +   
## rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=169.21  
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +   
## ptratio + lstat + medv + medv\_new + lstat\_new + rm\_new +   
## dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio + rm\_age\_interaction

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - medv\_new 1 127.22 167.22  
## - lstat\_new 1 127.22 167.22  
## - chas 1 127.26 167.26  
## - dis 1 127.43 167.43  
## - medv 1 127.44 167.44  
## - lstat 1 127.66 167.66  
## - age 1 127.97 167.97  
## - rm 1 128.20 168.20  
## - zn 1 128.29 168.29  
## - rm\_age\_interaction 1 128.66 168.66  
## <none> 127.20 169.21  
## - rm\_new 1 130.79 170.79  
## - lstat\_log 1 132.52 172.52  
## - tax\_ptratio\_ratio 1 132.91 172.91  
## - indus 1 134.68 174.68  
## - ptratio 1 137.62 177.62  
## - dis\_new 1 143.56 183.56  
## - tax\_log 1 154.38 194.38  
## - tax 1 156.92 196.92  
## - nox 1 167.70 207.70  
## - rad 1 196.62 236.62  
##   
## Step: AIC=167.21  
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +   
## ptratio + lstat + medv + lstat\_new + rm\_new + dis\_new + lstat\_log +   
## tax\_log + tax\_ptratio\_ratio + rm\_age\_interaction

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - lstat\_new 1 127.23 165.23  
## - chas 1 127.27 165.27  
## - dis 1 127.44 165.44  
## - lstat 1 127.67 165.67  
## - age 1 127.97 165.97  
## - rm 1 128.20 166.20  
## - zn 1 128.29 166.29  
## - rm\_age\_interaction 1 128.66 166.66  
## <none> 127.22 167.22  
## - rm\_new 1 131.06 169.06  
## + medv\_new 1 127.20 169.21  
## - lstat\_log 1 132.54 170.54  
## - medv 1 132.58 170.58  
## - tax\_ptratio\_ratio 1 132.98 170.98  
## - indus 1 134.69 172.69  
## - ptratio 1 137.69 175.69  
## - dis\_new 1 143.56 181.56  
## - tax\_log 1 154.62 192.62  
## - tax 1 157.22 195.22  
## - nox 1 167.82 205.82  
## - rad 1 196.63 234.63  
##   
## Step: AIC=165.23  
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +   
## ptratio + lstat + medv + rm\_new + dis\_new + lstat\_log + tax\_log +   
## tax\_ptratio\_ratio + rm\_age\_interaction  
##   
## Df Deviance AIC  
## - chas 1 127.28 163.28  
## - dis 1 127.45 163.45  
## - age 1 127.98 163.98  
## - rm 1 128.21 164.21  
## - zn 1 128.31 164.31  
## - rm\_age\_interaction 1 128.67 164.67  
## <none> 127.23 165.23  
## - rm\_new 1 131.06 167.06  
## + lstat\_new 1 127.22 167.22  
## + medv\_new 1 127.22 167.22  
## - medv 1 132.59 168.59  
## - lstat\_log 1 132.74 168.74  
## - tax\_ptratio\_ratio 1 132.99 168.99  
## - lstat 1 134.15 170.15  
## - indus 1 134.71 170.71  
## - ptratio 1 137.71 173.71  
## - dis\_new 1 143.57 179.57  
## - tax\_log 1 154.67 190.67  
## - tax 1 157.25 193.25  
## - nox 1 167.87 203.87  
## - rad 1 196.80 232.80  
##   
## Step: AIC=163.28  
## target ~ zn + indus + nox + rm + age + dis + rad + tax + ptratio +   
## lstat + medv + rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction  
##   
## Df Deviance AIC  
## - dis 1 127.47 161.47  
## - age 1 128.10 162.10  
## - rm 1 128.26 162.26  
## - zn 1 128.40 162.40  
## - rm\_age\_interaction 1 128.80 162.80  
## <none> 127.28 163.28  
## - rm\_new 1 131.19 165.19  
## + chas 1 127.23 165.23  
## + lstat\_new 1 127.27 165.27  
## + medv\_new 1 127.27 165.27  
## - medv 1 132.61 166.61  
## - lstat\_log 1 132.86 166.86  
## - tax\_ptratio\_ratio 1 133.34 167.34  
## - lstat 1 134.20 168.20  
## - indus 1 135.56 169.56  
## - ptratio 1 138.21 172.21  
## - dis\_new 1 143.60 177.60  
## - tax\_log 1 156.08 190.08  
## - tax 1 157.38 191.38  
## - nox 1 171.81 205.81  
## - rad 1 202.89 236.89  
##   
## Step: AIC=161.47  
## target ~ zn + indus + nox + rm + age + rad + tax + ptratio +   
## lstat + medv + rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction  
##   
## Df Deviance AIC  
## - rm 1 128.27 160.27  
## - zn 1 128.40 160.40  
## - age 1 128.55 160.55  
## - rm\_age\_interaction 1 129.37 161.37  
## <none> 127.47 161.47  
## - rm\_new 1 131.20 163.20  
## + dis 1 127.28 163.28  
## + lstat\_new 1 127.45 163.45  
## + chas 1 127.45 163.45  
## + medv\_new 1 127.46 163.46  
## - lstat\_log 1 132.99 164.99  
## - tax\_ptratio\_ratio 1 133.39 165.39  
## - medv 1 134.21 166.21  
## - lstat 1 134.29 166.29  
## - indus 1 135.58 167.58  
## - ptratio 1 138.29 170.29  
## - dis\_new 1 144.11 176.11  
## - tax\_log 1 156.31 188.31  
## - tax 1 158.95 190.95  
## - nox 1 183.68 215.68  
## - rad 1 203.90 235.90  
##   
## Step: AIC=160.27  
## target ~ zn + indus + nox + age + rad + tax + ptratio + lstat +   
## medv + rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction  
##   
## Df Deviance AIC  
## - zn 1 129.23 159.23  
## - age 1 129.23 159.23  
## - rm\_age\_interaction 1 129.96 159.96  
## <none> 128.27 160.27  
## + rm 1 127.47 161.47  
## - rm\_new 1 131.99 161.99  
## + chas 1 128.23 162.23  
## + dis 1 128.26 162.26  
## + lstat\_new 1 128.26 162.26  
## + medv\_new 1 128.27 162.27  
## - lstat\_log 1 134.35 164.35  
## - tax\_ptratio\_ratio 1 134.62 164.62  
## - lstat 1 135.85 165.85  
## - indus 1 136.47 166.47  
## - medv 1 137.50 167.50  
## - ptratio 1 139.93 169.93  
## - dis\_new 1 144.91 174.91  
## - tax\_log 1 156.79 186.79  
## - tax 1 160.10 190.10  
## - nox 1 184.00 214.00  
## - rad 1 203.93 233.93  
##   
## Step: AIC=159.23  
## target ~ indus + nox + age + rad + tax + ptratio + lstat + medv +   
## rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction  
##   
## Df Deviance AIC  
## - age 1 129.95 157.95  
## - rm\_age\_interaction 1 130.56 158.56  
## <none> 129.23 159.23  
## + zn 1 128.27 160.27  
## - rm\_new 1 132.29 160.29  
## + rm 1 128.40 160.40  
## + chas 1 129.11 161.11  
## + dis 1 129.15 161.15  
## + lstat\_new 1 129.22 161.22  
## + medv\_new 1 129.23 161.23  
## - tax\_ptratio\_ratio 1 135.58 163.58  
## - lstat\_log 1 135.83 163.83  
## - indus 1 136.61 164.61  
## - lstat 1 137.64 165.64  
## - medv 1 138.11 166.11  
## - ptratio 1 140.47 168.47  
## - dis\_new 1 145.71 173.71  
## - tax\_log 1 156.79 184.79  
## - tax 1 160.12 188.12  
## - nox 1 186.28 214.28  
## - rad 1 205.19 233.19  
##   
## Step: AIC=157.95  
## target ~ indus + nox + rad + tax + ptratio + lstat + medv + rm\_new +   
## dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio + rm\_age\_interaction  
##   
## Df Deviance AIC  
## <none> 129.95 157.95  
## - rm\_new 1 133.11 159.11  
## + age 1 129.23 159.23  
## + zn 1 129.23 159.23  
## + rm 1 129.25 159.25  
## + chas 1 129.76 159.76  
## + medv\_new 1 129.94 159.94  
## + lstat\_new 1 129.94 159.94  
## + dis 1 129.95 159.95  
## - lstat\_log 1 135.84 161.84  
## - tax\_ptratio\_ratio 1 136.23 162.23  
## - rm\_age\_interaction 1 136.66 162.66  
## - indus 1 136.82 162.82  
## - lstat 1 137.66 163.66  
## - medv 1 139.91 165.91  
## - ptratio 1 141.16 167.16  
## - dis\_new 1 146.53 172.53  
## - tax\_log 1 157.51 183.51  
## - tax 1 160.75 186.75  
## - nox 1 194.75 220.75  
## - rad 1 206.15 232.15

summary(m3)

##   
## Call:  
## glm(formula = target ~ indus + nox + rad + tax + ptratio + lstat +   
## medv + rm\_new + dis\_new + lstat\_log + tax\_log + tax\_ptratio\_ratio +   
## rm\_age\_interaction, family = "binomial", data = s)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8775 -0.0796 0.0000 0.0306 3.9887   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.2875 0.5667 0.507 0.611972   
## indus 1.3813 0.5496 2.513 0.011956 \*   
## nox 5.7438 0.9670 5.940 2.86e-09 \*\*\*  
## rad 11.2322 2.0838 5.390 7.03e-08 \*\*\*  
## tax -42.7932 9.8462 -4.346 1.39e-05 \*\*\*  
## ptratio 5.0678 1.7689 2.865 0.004171 \*\*   
## lstat 3.4429 1.3126 2.623 0.008719 \*\*   
## medv 2.3714 0.8528 2.781 0.005426 \*\*   
## rm\_new -1.0878 0.6266 -1.736 0.082568 .   
## dis\_new -1.1713 0.3290 -3.560 0.000371 \*\*\*  
## lstat\_log -3.4665 1.5020 -2.308 0.021000 \*   
## tax\_log 20.1628 4.5014 4.479 7.49e-06 \*\*\*  
## tax\_ptratio\_ratio 12.2851 5.7830 2.124 0.033640 \*   
## rm\_age\_interaction 1.1262 0.4535 2.483 0.013014 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 645.88 on 465 degrees of freedom  
## Residual deviance: 129.95 on 452 degrees of freedom  
## AIC: 157.95  
##   
## Number of Fisher Scoring iterations: 9