DSE6111: Week One Exercises

Adeline Casali

2023-10-25

# Problem 8

1. Read in the dataset

college <- read.csv("Week\_One/Data/college.csv")

1. Create a row.names column with the name of each university and remove the column where the university names were originally stored

rownames(college) <- college[, 1]  
head(college)

## X Private Apps Accept  
## Abilene Christian University Abilene Christian University Yes 1660 1232  
## Adelphi University Adelphi University Yes 2186 1924  
## Adrian College Adrian College Yes 1428 1097  
## Agnes Scott College Agnes Scott College Yes 417 349  
## Alaska Pacific University Alaska Pacific University Yes 193 146  
## Albertson College Albertson College Yes 587 479  
## Enroll Top10perc Top25perc F.Undergrad P.Undergrad  
## Abilene Christian University 721 23 52 2885 537  
## Adelphi University 512 16 29 2683 1227  
## Adrian College 336 22 50 1036 99  
## Agnes Scott College 137 60 89 510 63  
## Alaska Pacific University 55 16 44 249 869  
## Albertson College 158 38 62 678 41  
## Outstate Room.Board Books Personal PhD Terminal  
## Abilene Christian University 7440 3300 450 2200 70 78  
## Adelphi University 12280 6450 750 1500 29 30  
## Adrian College 11250 3750 400 1165 53 66  
## Agnes Scott College 12960 5450 450 875 92 97  
## Alaska Pacific University 7560 4120 800 1500 76 72  
## Albertson College 13500 3335 500 675 67 73  
## S.F.Ratio perc.alumni Expend Grad.Rate  
## Abilene Christian University 18.1 12 7041 60  
## Adelphi University 12.2 16 10527 56  
## Adrian College 12.9 30 8735 54  
## Agnes Scott College 7.7 37 19016 59  
## Alaska Pacific University 11.9 2 10922 15  
## Albertson College 9.4 11 9727 55

college <- college[,-1]  
head(college)

## Private Apps Accept Enroll Top10perc Top25perc  
## Abilene Christian University Yes 1660 1232 721 23 52  
## Adelphi University Yes 2186 1924 512 16 29  
## Adrian College Yes 1428 1097 336 22 50  
## Agnes Scott College Yes 417 349 137 60 89  
## Alaska Pacific University Yes 193 146 55 16 44  
## Albertson College Yes 587 479 158 38 62  
## F.Undergrad P.Undergrad Outstate Room.Board Books  
## Abilene Christian University 2885 537 7440 3300 450  
## Adelphi University 2683 1227 12280 6450 750  
## Adrian College 1036 99 11250 3750 400  
## Agnes Scott College 510 63 12960 5450 450  
## Alaska Pacific University 249 869 7560 4120 800  
## Albertson College 678 41 13500 3335 500  
## Personal PhD Terminal S.F.Ratio perc.alumni Expend  
## Abilene Christian University 2200 70 78 18.1 12 7041  
## Adelphi University 1500 29 30 12.2 16 10527  
## Adrian College 1165 53 66 12.9 30 8735  
## Agnes Scott College 875 92 97 7.7 37 19016  
## Alaska Pacific University 1500 76 72 11.9 2 10922  
## Albertson College 675 67 73 9.4 11 9727  
## Grad.Rate  
## Abilene Christian University 60  
## Adelphi University 56  
## Adrian College 54  
## Agnes Scott College 59  
## Alaska Pacific University 15  
## Albertson College 55

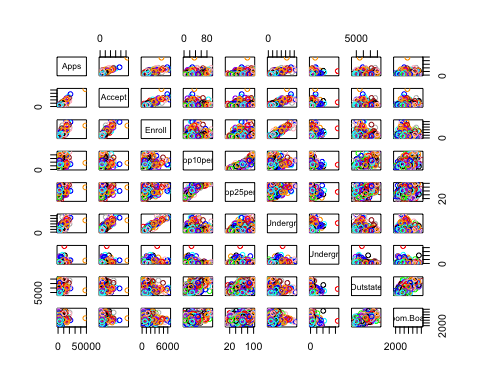
1. Numerical summary of the variables in the dataset

summary(college)

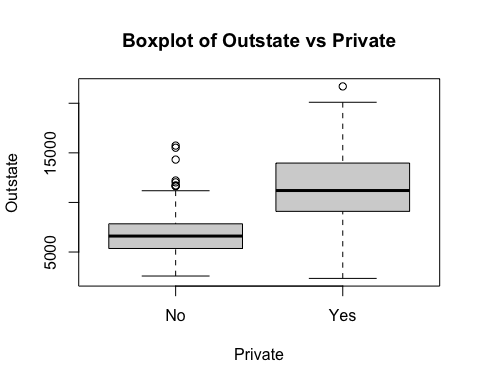
## Private Apps Accept Enroll   
## Length:777 Min. : 81 Min. : 72 Min. : 35   
## Class :character 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242   
## Mode :character Median : 1558 Median : 1110 Median : 434   
## Mean : 3002 Mean : 2019 Mean : 780   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902   
## Max. :48094 Max. :26330 Max. :6392   
## Top10perc Top25perc F.Undergrad P.Undergrad   
## Min. : 1.00 Min. : 9.0 Min. : 139 Min. : 1.0   
## 1st Qu.:15.00 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0   
## Median :23.00 Median : 54.0 Median : 1707 Median : 353.0   
## Mean :27.56 Mean : 55.8 Mean : 3700 Mean : 855.3   
## 3rd Qu.:35.00 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0   
## Max. :96.00 Max. :100.0 Max. :31643 Max. :21836.0   
## Outstate Room.Board Books Personal   
## Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250   
## 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850   
## Median : 9990 Median :4200 Median : 500.0 Median :1200   
## Mean :10441 Mean :4358 Mean : 549.4 Mean :1341   
## 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700   
## Max. :21700 Max. :8124 Max. :2340.0 Max. :6800   
## PhD Terminal S.F.Ratio perc.alumni   
## Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00   
## 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00   
## Median : 75.00 Median : 82.0 Median :13.60 Median :21.00   
## Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74   
## 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00   
## Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00   
## Expend Grad.Rate   
## Min. : 3186 Min. : 10.00   
## 1st Qu.: 6751 1st Qu.: 53.00   
## Median : 8377 Median : 65.00   
## Mean : 9660 Mean : 65.46   
## 3rd Qu.:10830 3rd Qu.: 78.00   
## Max. :56233 Max. :118.00

1. Scatterplot matrix of the first ten numeric columns of the data

pairs(college[, 1:10][, sapply(college[, 1:10], is.numeric)], col = c("red", "blue", "green", "orange", "purple", "pink", "cyan", "brown", "black", "gray"))

 iii. Boxplots of Outstate vs Private colleges

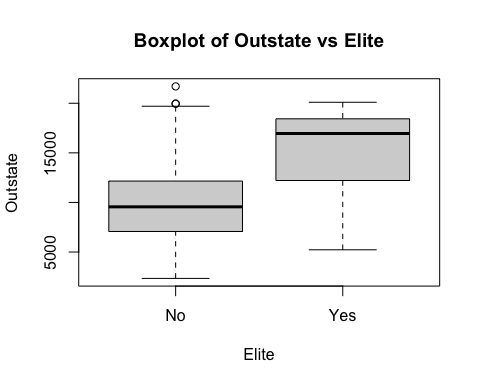
boxplot(Outstate ~ Private, data = college, xlab = "Private", ylab = "Outstate", main = "Boxplot of Outstate vs Private")

 iv. Creation of the Elite variable and boxplot depicting outstate vs elite colleges

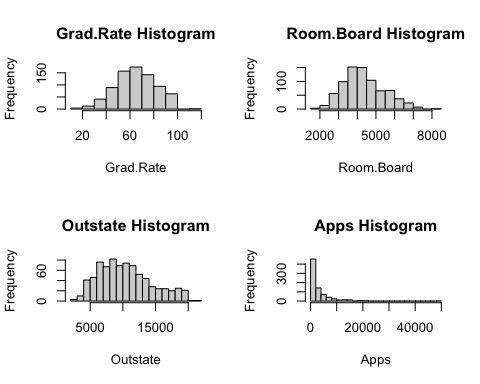
Elite <- rep("No", nrow(college))  
Elite[college$Top10perc > 50] <- "Yes"  
Elite <- as.factor(Elite)  
college <- data.frame(college, Elite)  
summary(college)

## Private Apps Accept Enroll   
## Length:777 Min. : 81 Min. : 72 Min. : 35   
## Class :character 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242   
## Mode :character Median : 1558 Median : 1110 Median : 434   
## Mean : 3002 Mean : 2019 Mean : 780   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902   
## Max. :48094 Max. :26330 Max. :6392   
## Top10perc Top25perc F.Undergrad P.Undergrad   
## Min. : 1.00 Min. : 9.0 Min. : 139 Min. : 1.0   
## 1st Qu.:15.00 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0   
## Median :23.00 Median : 54.0 Median : 1707 Median : 353.0   
## Mean :27.56 Mean : 55.8 Mean : 3700 Mean : 855.3   
## 3rd Qu.:35.00 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0   
## Max. :96.00 Max. :100.0 Max. :31643 Max. :21836.0   
## Outstate Room.Board Books Personal   
## Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250   
## 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850   
## Median : 9990 Median :4200 Median : 500.0 Median :1200   
## Mean :10441 Mean :4358 Mean : 549.4 Mean :1341   
## 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700   
## Max. :21700 Max. :8124 Max. :2340.0 Max. :6800   
## PhD Terminal S.F.Ratio perc.alumni   
## Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00   
## 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00   
## Median : 75.00 Median : 82.0 Median :13.60 Median :21.00   
## Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74   
## 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00   
## Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00   
## Expend Grad.Rate Elite   
## Min. : 3186 Min. : 10.00 No :699   
## 1st Qu.: 6751 1st Qu.: 53.00 Yes: 78   
## Median : 8377 Median : 65.00   
## Mean : 9660 Mean : 65.46   
## 3rd Qu.:10830 3rd Qu.: 78.00   
## Max. :56233 Max. :118.00

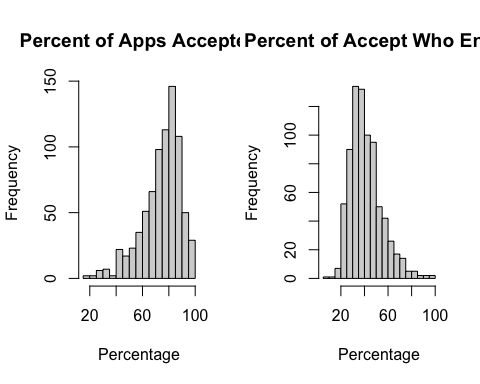
boxplot(Outstate ~ Elite, data = college, xlab = "Elite", ylab = "Outstate", main = "Boxplot of Outstate vs Elite")

 v. Histograms of graduation rate, room and board, out of state tuition, and number of applications

par(mfrow = c(2, 2))  
hist(college$Grad.Rate, breaks = 10, main = "Grad.Rate Histogram", xlab = "Grad.Rate")  
hist(college$Room.Board, breaks = 20, main = "Room.Board Histogram", xlab = "Room.Board")  
hist(college$Outstate, breaks = 15, main = "Outstate Histogram", xlab = "Outstate")  
hist(college$Apps, breaks = 30, main = "Apps Histogram", xlab = "Apps")

 vi. I created two new columns; the first calculated the percentage of applicants who were accepted, and the second calculated the percentage of those accepted who enrolled. I then produced a histogram of each new variable and discovered that most colleges have a high admittance rate (around 80%), but the majority of colleges have less than 40% of accepted students actually enroll.

college$Percent\_Apps\_Accepted <- (college$Accept / college$Apps) \* 100  
college$Percent\_Accept\_Enroll <- (college$Enroll / college$Accept) \* 100  
par(mfrow = c(1, 2))  
hist(college$Percent\_Apps\_Accepted, breaks = 20, main = "Percent of Apps Accepted", xlab = "Percentage")  
hist(college$Percent\_Accept\_Enroll, breaks = 20, main = "Percent of Accept Who Enroll", xlab = "Percentage")



# Problem 9

Read in dataset and check for and remove NA values

auto <- read.csv("Week\_One/Data/Auto.csv", na.strings = "?")  
colMeans(is.na(auto))

## mpg cylinders displacement horsepower weight acceleration   
## 0.00000000 0.00000000 0.00000000 0.01259446 0.00000000 0.00000000   
## year origin name   
## 0.00000000 0.00000000 0.00000000

auto <- na.omit(auto)  
colMeans(is.na(auto))

## mpg cylinders displacement horsepower weight acceleration   
## 0 0 0 0 0 0   
## year origin name   
## 0 0 0

1. From looking at the data, the only column that is clearly qualitative is name. However, cylinders, year, and origin are also qualitative predictors as they have a set number of categorical variables within them. The remaining variables (mpg, displacement, horsepower, weight, and acceleration) are all quantitative variables.

head(auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

1. Calculating the range of each quantitative predictor

cat("The range of mpg is", range(auto$mpg), "\n")

## The range of mpg is 9 46.6

cat("The range of displacement is", range(auto$displacement), "\n")

## The range of displacement is 68 455

cat("The range of horsepower is", range(auto$horsepower), "\n")

## The range of horsepower is 46 230

cat("The range of weight is", range(auto$weight), "\n")

## The range of weight is 1613 5140

cat("The range of acceleration is", range(auto$acceleration), "\n")

## The range of acceleration is 8 24.8

1. Calculating the mean and standard deviation of each quantitative predictor

cat("The mean of mpg is", mean(auto$mpg), "and the standard deviation is", sd(auto$mpg), "\n")

## The mean of mpg is 23.44592 and the standard deviation is 7.805007

cat("The mean of displacement is", mean(auto$displacement), "and the standard deviation is", sd(auto$displacement), "\n")

## The mean of displacement is 194.412 and the standard deviation is 104.644

cat("The mean of horsepower is", mean(auto$horsepower), "and the standard deviation is", sd(auto$horsepower), "\n")

## The mean of horsepower is 104.4694 and the standard deviation is 38.49116

cat("The mean of weight is", mean(auto$weight), "and the standard deviation is", sd(auto$weight), "\n")

## The mean of weight is 2977.584 and the standard deviation is 849.4026

cat("The mean of acceleration is", mean(auto$acceleration), "and the standard deviation is", sd(auto$acceleration), "\n")

## The mean of acceleration is 15.54133 and the standard deviation is 2.758864

1. Removing the 10th through 85th observations and re-calculating the range, mean, and standard deviation of each qualitative predictor.

auto\_subset <- auto[-(10:85), ]  
cat("The mean of mpg is", mean(auto\_subset$mpg), ", the standard deviation is", sd(auto\_subset$mpg), ", and the range is", range(auto\_subset$mpg), "\n")

## The mean of mpg is 24.40443 , the standard deviation is 7.867283 , and the range is 11 46.6

cat("The mean of displacement is", mean(auto\_subset$displacement), ", the standard deviation is", sd(auto\_subset$displacement), ", and the range is", range(auto\_subset$displacement), "\n")

## The mean of displacement is 187.2405 , the standard deviation is 99.67837 , and the range is 68 455

cat("The mean of horsepower is", mean(auto\_subset$horsepower), ", the standard deviation is", sd(auto\_subset$horsepower), ", and the range is", range(auto\_subset$horsepower), "\n")

## The mean of horsepower is 100.7215 , the standard deviation is 35.70885 , and the range is 46 230

cat("The mean of weight is", mean(auto\_subset$weight), ", the standard deviation is", sd(auto\_subset$weight), ", and the range is", range(auto\_subset$weight), "\n")

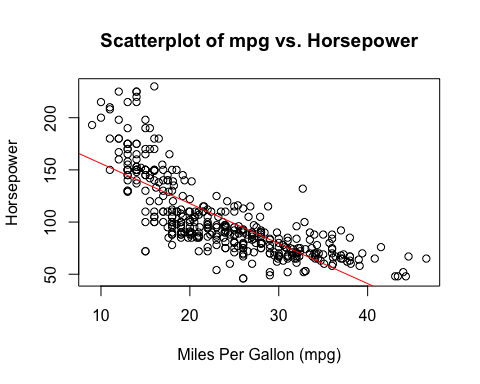
## The mean of weight is 2935.972 , the standard deviation is 811.3002 , and the range is 1649 4997

cat("The mean of acceleration is", mean(auto\_subset$acceleration), ", the standard deviation is", sd(auto\_subset$acceleration), ", and the range is", range(auto\_subset$acceleration), "\n")

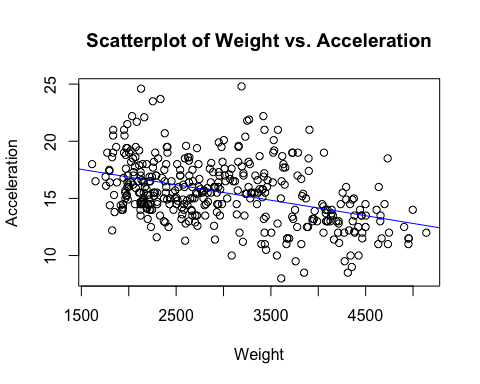
## The mean of acceleration is 15.7269 , the standard deviation is 2.693721 , and the range is 8.5 24.8

1. Plot 1: Scatterplot of mpg vs Horsepower - there seems to be a negative linear relationship between the mpg and horsepower of cars in the auto dataset Plot 2: Scatterplot of weight vs Acceleration - there is a slight negative linear relationship between the weight and acceleration predictors Plot 3: Barplot of Cylinders vs Acceleration - the plot shows an inverted u pattern Plot 4: Plot of Year vs Displacement - the plot shows a decrease in displacement over car years (negative linear relationship)

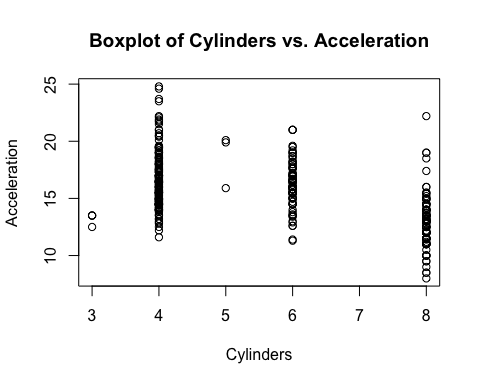
plot(auto$mpg, auto$horsepower, xlab = "Miles Per Gallon (mpg)", ylab = "Horsepower", main = "Scatterplot of mpg vs. Horsepower")  
abline(lm(auto$horsepower ~ auto$mpg), col = "red")



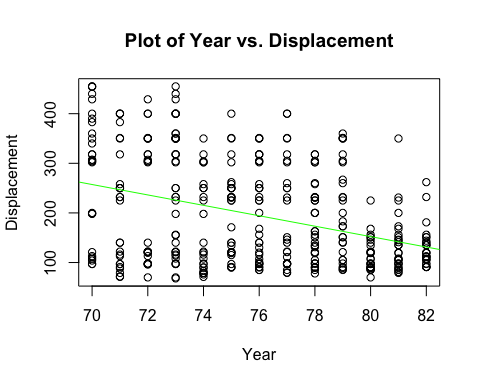
plot(auto$weight, auto$acceleration, xlab = "Weight", ylab = "Acceleration", main = "Scatterplot of Weight vs. Acceleration")  
abline(lm(auto$acceleration ~ auto$weight), col = "blue")



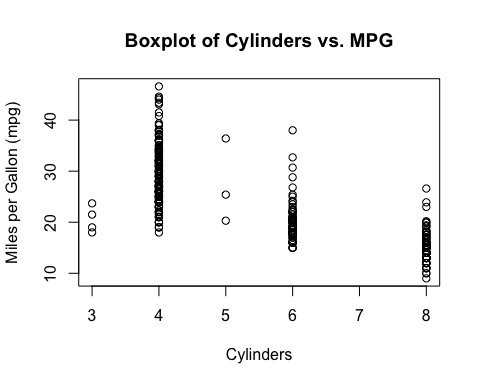
plot(auto$cylinders, auto$acceleration, xlab = "Cylinders", ylab = "Acceleration", main = "Boxplot of Cylinders vs. Acceleration")



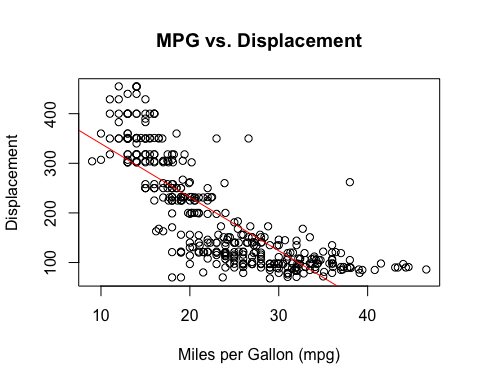
plot(auto$year, auto$displacement, xlab = "Year", ylab = "Displacement", main = "Plot of Year vs. Displacement")  
abline(lm(auto$displacement ~ auto$year), col = "green")

 f) According to the plots produced below, there looks to be a negative linear relationship between displacement, horsepower, and weight and mpg. Additionally, there looks to be a weak positive linear relationship between acceleration and mpg as well as a stronger positive linear relationship between year and origin and mpg.

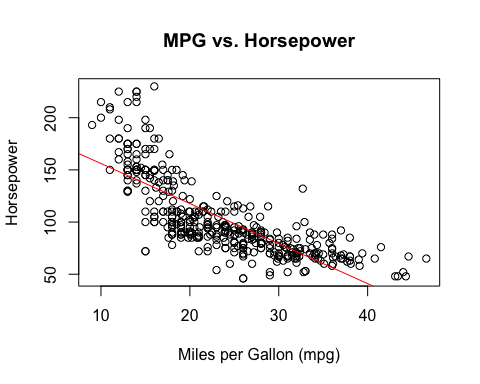
# Plot 1: Boxplot of Cylinders vs. MPG  
plot(auto$cylinders, auto$mpg, xlab = "Cylinders", ylab = "Miles per Gallon (mpg)", main = "Boxplot of Cylinders vs. MPG")



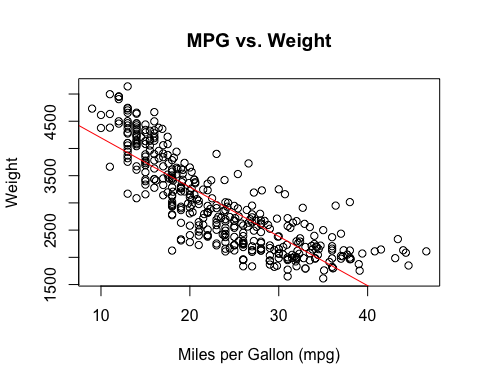
# Plot 2: MPG vs. Displacement  
plot(auto$mpg, auto$displacement, xlab = "Miles per Gallon (mpg)", ylab = "Displacement", main = "MPG vs. Displacement")  
abline(lm(auto$displacement ~ auto$mpg), col = "red")



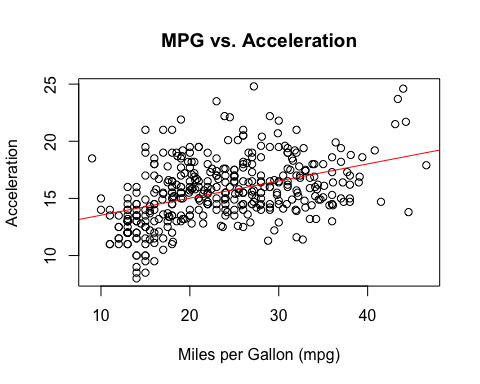
# Plot 3: MPG vs. Horsepower  
plot(auto$mpg, auto$horsepower, xlab = "Miles per Gallon (mpg)", ylab = "Horsepower", main = "MPG vs. Horsepower")  
abline(lm(auto$horsepower ~ auto$mpg), col = "red")



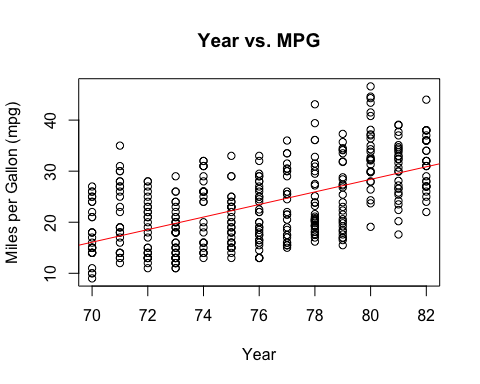
# Plot 4: MPG vs. Weight  
plot(auto$mpg, auto$weight, xlab = "Miles per Gallon (mpg)", ylab = "Weight", main = "MPG vs. Weight")  
abline(lm(auto$weight ~ auto$mpg), col = "red")



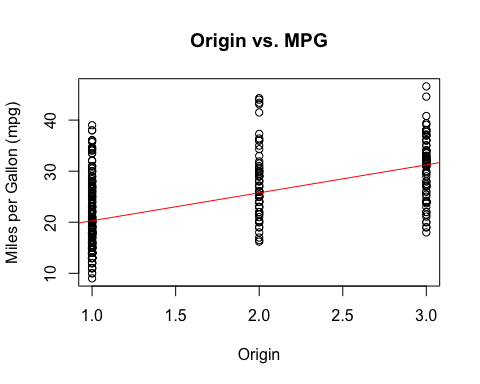
# Plot 5: MPG vs. Acceleration  
plot(auto$mpg, auto$acceleration, xlab = "Miles per Gallon (mpg)", ylab = "Acceleration", main = "MPG vs. Acceleration")  
abline(lm(auto$acceleration ~ auto$mpg), col = "red")



# Plot 6: Year vs. MPG  
plot(auto$year, auto$mpg, xlab = "Year", ylab = "Miles per Gallon (mpg)", main = "Year vs. MPG")  
abline(lm(auto$mpg ~ auto$year), col = "red")



# Plot 7: Origin vs. MPG  
plot(auto$origin, auto$mpg, xlab = "Origin", ylab = "Miles per Gallon (mpg)", main = "Origin vs. MPG")  
abline(lm(auto$mpg ~ auto$origin), col = "red")



# Problem 10

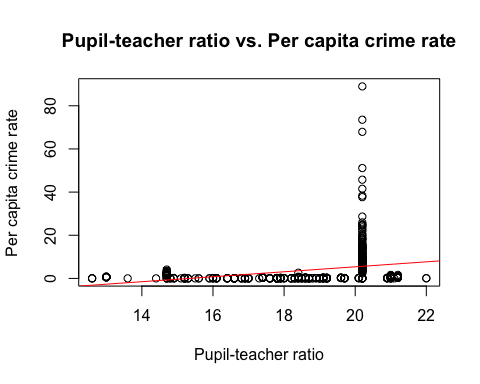
1. Load dataset and check its structure and components. It has 506 rows and 14 columns. Each row represents a suburb in the Boston area and the columns represent different housing values pertaining to that suburb - for example, per capita crime rate, the median value of homes, and the tax rate, among others.

library(MASS)  
?Boston  
str(Boston)

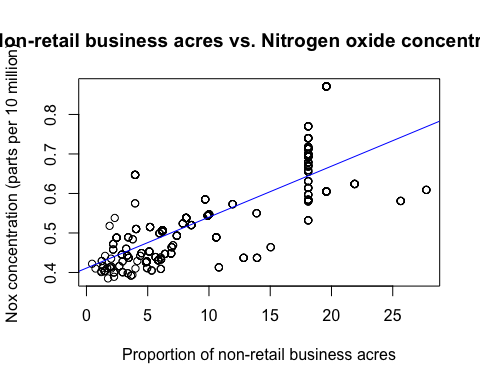
## 'data.frame': 506 obs. of 14 variables:  
## $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ chas : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ rm : num 6.58 6.42 7.18 7 7.15 ...  
## $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ rad : int 1 2 2 3 3 3 5 5 5 5 ...  
## $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  
## $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ black : num 397 397 393 395 397 ...  
## $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  
## $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

1. There seems to be a weak positive linear relationship between the pupil-teacher ratio and the crime rate per capita, with the highest crime rates occurring around a 20:1 ratio. Additionally, there is a strong positive linear correlation between the proportion of non-retail business acres and the nitrogen oxide concentration. Finally, there is a weak negative linear correlation between the proportion of units built prior to 1940 and the median value of homes.

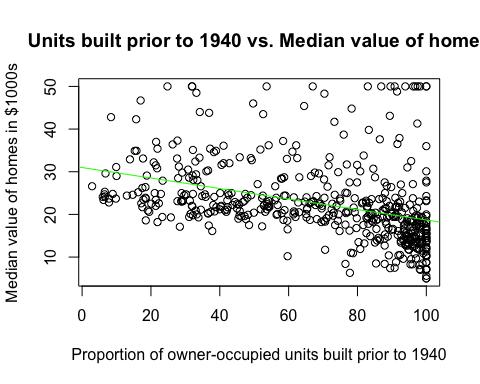
plot(Boston$ptratio, Boston$crim, xlab = "Pupil-teacher ratio", ylab = "Per capita crime rate", main = "Pupil-teacher ratio vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$ptratio), col = "red")



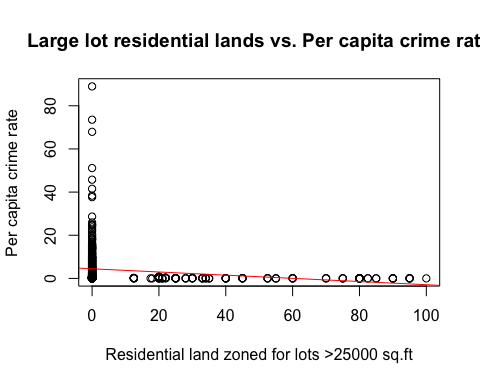
plot(Boston$indus, Boston$nox, xlab = "Proportion of non-retail business acres", ylab = "Nox concentration (parts per 10 million)", main = "Non-retail business acres vs. Nitrogen oxide concentration")  
abline(lm(Boston$nox ~ Boston$indus), col = "blue")



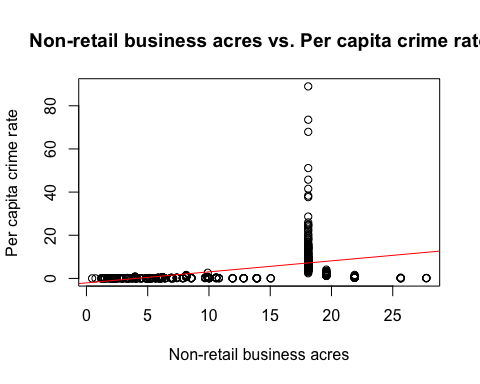
plot(Boston$age, Boston$medv, xlab = "Proportion of owner-occupied units built prior to 1940", ylab = "Median value of homes in $1000s", main = "Units built prior to 1940 vs. Median value of homes")  
abline(lm(Boston$medv ~ Boston$age), col = "green")

 c) There are weak positive linear relationships between the number of non-retail business acres, the proportion of units built prior to 1940, accessibility to highways, property tax rate, pupil teacher ratio, and lower status of the population when compared with the crime rate per capita. There are weak negative linear relationships between the number of large residential lands, the number of rooms per dwelling, distance to employment centers, proportion of blacks, and median value of homes when compared to the crime rate per capita. Additionally, the towns with the highest crime rates appeared to be off of the Charles River.

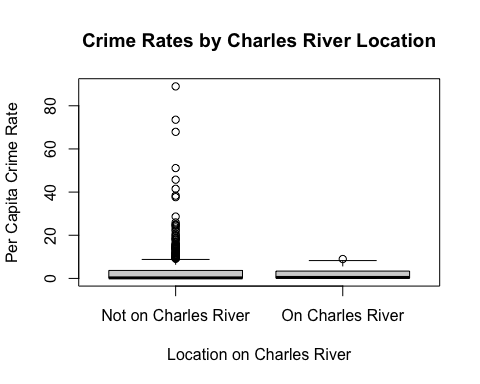
# Plot 1: Large lot residential lands  
plot(Boston$zn, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Residential land zoned for lots >25000 sq.ft",  
 main = "Large lot residential lands vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$zn), col = "red")



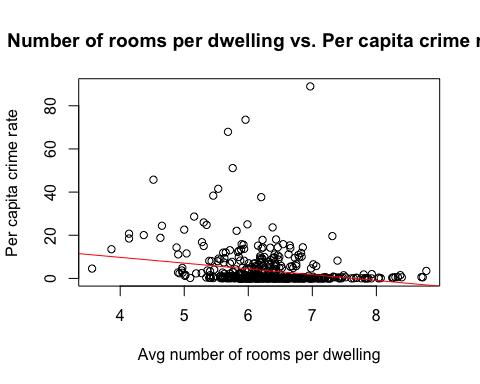
# Plot 2: Non-retail business acres  
plot(Boston$indus, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Non-retail business acres",  
 main = "Non-retail business acres vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$indus), col = "red")



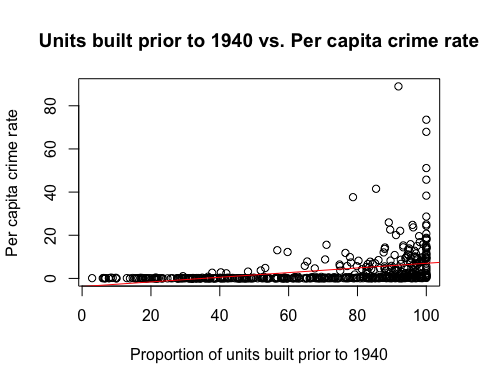
# Boxplot for Crime Rates by Charles River Location  
boxplot(crim ~ chas, data = Boston, names = c("Not on Charles River", "On Charles River"),  
 ylab = "Per Capita Crime Rate",  
 xlab = "Location on Charles River",  
 main = "Crime Rates by Charles River Location")



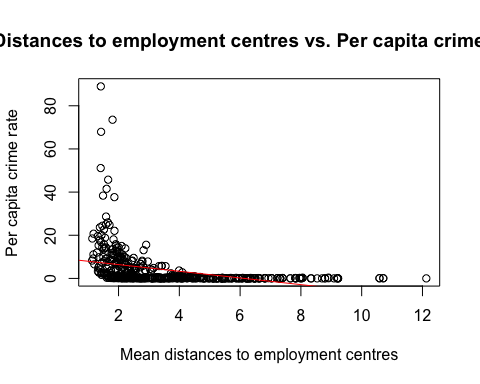
# Plot 3: Number of rooms per dwelling  
plot(Boston$rm, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Avg number of rooms per dwelling",  
 main = "Number of rooms per dwelling vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$rm), col = "red")



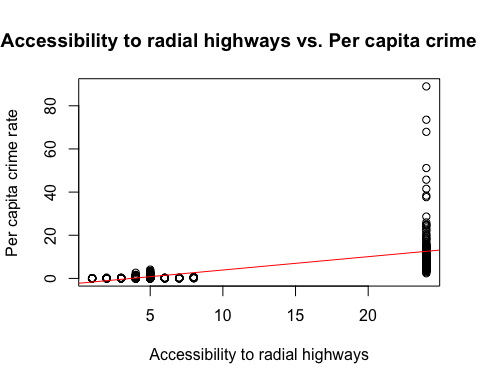
# Plot 4: Proportion of units built prior to 1940  
plot(Boston$age, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Proportion of units built prior to 1940",  
 main = "Units built prior to 1940 vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$age), col = "red")



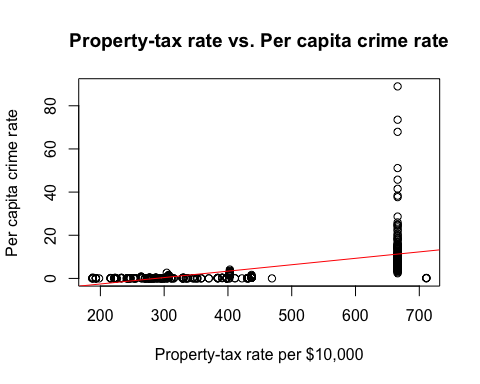
# Plot 5: Mean distances to employment centres  
plot(Boston$dis, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Mean distances to employment centres",  
 main = "Distances to employment centres vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$dis), col = "red")



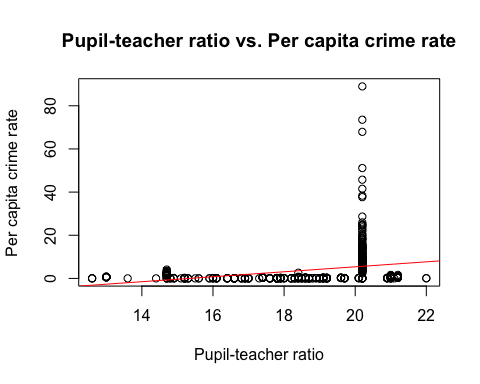
# Plot 6: Accessibility to radial highways  
plot(Boston$rad, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Accessibility to radial highways",  
 main = "Accessibility to radial highways vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$rad), col = "red")



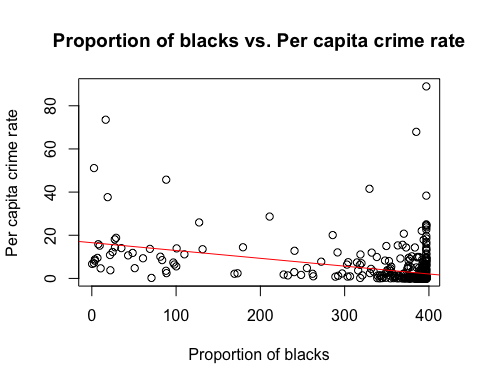
# Plot 7: Property-tax rate per $10,000  
plot(Boston$tax, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Property-tax rate per $10,000",  
 main = "Property-tax rate vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$tax), col = "red")



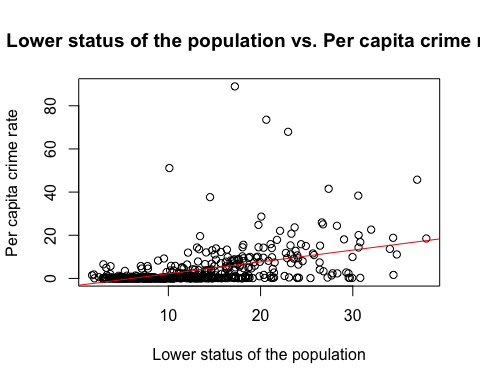
# Plot 8: Pupil-teacher ratio  
plot(Boston$ptratio, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Pupil-teacher ratio",  
 main = "Pupil-teacher ratio vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$ptratio), col = "red")



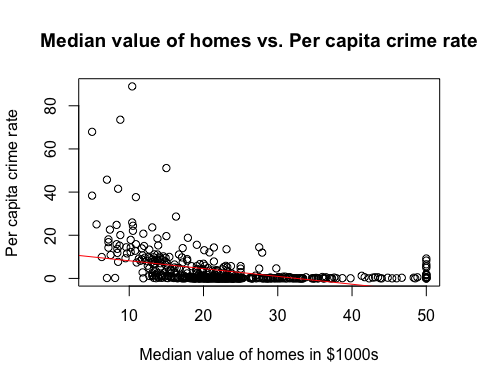
# Plot 9: Proportion of blacks  
plot(Boston$black, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Proportion of blacks",  
 main = "Proportion of blacks vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$black), col = "red")



# Plot 10: Lower status of the population  
plot(Boston$lstat, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Lower status of the population",  
 main = "Lower status of the population vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$lstat), col = "red")



# Plot 11: Median value of homes in $1000s  
plot(Boston$medv, Boston$crim, ylab = "Per capita crime rate",   
 xlab = "Median value of homes in $1000s",  
 main = "Median value of homes vs. Per capita crime rate")  
abline(lm(Boston$crim ~ Boston$medv), col = "red")

 d) There are a wide range of values for each predictor, all listed below. Some of the most notable ranges are the crime rate per capita, which ranges from 0.00632 to 88.9762, the proportion of buildings built prior to 1940 which ranges from 2.9 to 100, and the pupil-teacher ratios which ranges from 12.6 to 22.

# Calculate and print the range of each column  
for (col\_name in names(Boston)) {  
 col\_range <- range(Boston[[col\_name]], na.rm = TRUE)  
 cat(paste("Range of", col\_name, ":", col\_range[1], "to", col\_range[2], "\n"))  
}

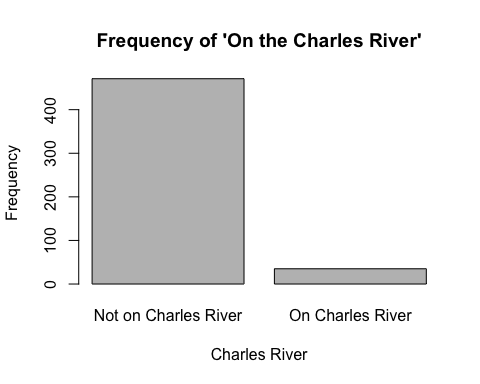
## Range of crim : 0.00632 to 88.9762   
## Range of zn : 0 to 100   
## Range of indus : 0.46 to 27.74   
## Range of chas : 0 to 1   
## Range of nox : 0.385 to 0.871   
## Range of rm : 3.561 to 8.78   
## Range of age : 2.9 to 100   
## Range of dis : 1.1296 to 12.1265   
## Range of rad : 1 to 24   
## Range of tax : 187 to 711   
## Range of ptratio : 12.6 to 22   
## Range of black : 0.32 to 396.9   
## Range of lstat : 1.73 to 37.97   
## Range of medv : 5 to 50

1. 35 of the suburbs in this dataset are located on the Charles River.

cat("Number of towns on the Charles River: ", sum(Boston$chas == 1), "\n")

## Number of towns on the Charles River: 35

barplot(table(Boston$chas),   
 names.arg = c("Not on Charles River", "On Charles River"),  
 xlab = "Charles River",  
 ylab = "Frequency",  
 main = "Frequency of 'On the Charles River'")

 f) The median pupil-teacher ratio for the towns located in this dataset is 19.05.

cat("Median Student-Teacher Ratio: ", median(Boston$ptratio), "\n")

## Median Student-Teacher Ratio: 19.05

1. There are two towns that have the lowest median value of owner-occupied homes, at $5000. They both have relatively high crime rates, levels of industry, levels of nitrogen oxides, accessibility to highways, and tax rates. Additionally, they both have some of the highest proportions of old houses, lower status populations, and blacks in the dataset. They are both not located on the Charles River.

cat("Suburb with the lowest median value of owner-occupied homes (in 1000s):", min(Boston$medv), "\n")

## Suburb with the lowest median value of owner-occupied homes (in 1000s): 5

print(Boston[Boston$medv == 5, ])

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## 399 38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 396.90 30.59  
## 406 67.9208 0 18.1 0 0.693 5.683 100 1.4254 24 666 20.2 384.97 22.98  
## medv  
## 399 5  
## 406 5

1. The suburbs that average more than 8 rooms per dwelling tend to have low crime rates, industrialization, and nitrogen oxides. They tend to have higher proportions of old houses (minus one outlier) and median incomes.

cat("Number of suburbs with more than 7 rooms per dwelling: ", nrow(Boston[Boston$rm > 7, ]), "\n")

## Number of suburbs with more than 7 rooms per dwelling: 64

cat("Number of suburbs with more than 8 rooms per dwelling: ", nrow(Boston[Boston$rm > 8, ]), "\n")

## Number of suburbs with more than 8 rooms per dwelling: 13

print(Boston[Boston$rm > 8, ])

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## 98 0.12083 0 2.89 0 0.4450 8.069 76.0 3.4952 2 276 18.0 396.90 4.21  
## 164 1.51902 0 19.58 1 0.6050 8.375 93.9 2.1620 5 403 14.7 388.45 3.32  
## 205 0.02009 95 2.68 0 0.4161 8.034 31.9 5.1180 4 224 14.7 390.55 2.88  
## 225 0.31533 0 6.20 0 0.5040 8.266 78.3 2.8944 8 307 17.4 385.05 4.14  
## 226 0.52693 0 6.20 0 0.5040 8.725 83.0 2.8944 8 307 17.4 382.00 4.63  
## 227 0.38214 0 6.20 0 0.5040 8.040 86.5 3.2157 8 307 17.4 387.38 3.13  
## 233 0.57529 0 6.20 0 0.5070 8.337 73.3 3.8384 8 307 17.4 385.91 2.47  
## 234 0.33147 0 6.20 0 0.5070 8.247 70.4 3.6519 8 307 17.4 378.95 3.95  
## 254 0.36894 22 5.86 0 0.4310 8.259 8.4 8.9067 7 330 19.1 396.90 3.54  
## 258 0.61154 20 3.97 0 0.6470 8.704 86.9 1.8010 5 264 13.0 389.70 5.12  
## 263 0.52014 20 3.97 0 0.6470 8.398 91.5 2.2885 5 264 13.0 386.86 5.91  
## 268 0.57834 20 3.97 0 0.5750 8.297 67.0 2.4216 5 264 13.0 384.54 7.44  
## 365 3.47428 0 18.10 1 0.7180 8.780 82.9 1.9047 24 666 20.2 354.55 5.29  
## medv  
## 98 38.7  
## 164 50.0  
## 205 50.0  
## 225 44.8  
## 226 50.0  
## 227 37.6  
## 233 41.7  
## 234 48.3  
## 254 42.8  
## 258 50.0  
## 263 48.8  
## 268 50.0  
## 365 21.9