Stats101C_HW1

Takao Oba

9/30/2022

Name: Takao Oba

Question 1

Upload HW1Train.csv posted on bruinlearn week 2.

(a)Fit the following 5 models. 1) log(x), 2) poly(x,2) 3) poly(x,3) 4) poly(x,4) and 5) poly(x,5). Call them Model 1 through 5 respectively. List the MSE_training R-squared and R-Squared Adjusted for each of the five models.

Reading in the "HW1Train.csv" file

```
setwd("/Users/takaooba/Downloads/")
train <- read.csv("HW1Train.csv")
train <- train[,c(2,3)]
head(train)
```

```
##     x         y
## 1 1.0  83.79707
## 2 1.5  60.01943
## 3 2.0  46.13691
## 4 2.5 178.04065
## 5 3.0  57.10827
## 6 3.5   1.20082
```

Fitting the models

```
# log(x)
model1 <- lm(y ~ log(x), data = train)
# poly(x,2)
model2 <- lm(y ~ poly(x,2), data = train)
# poly(x,3)
model3 <- lm(y ~ poly(x,3), data = train)
# poly(x,4)
model4 <- lm(y ~ poly(x,4), data = train)
# poly(x,5)
model5 <- lm(y ~ poly(x,5), data = train)
```

MSE_training, R-squared, and R-squared Adjusted

```
# model 1
# ANOVA table
anova(model1)$Mean[2]
```

```
## [1] 3572.03
```

```
# Summary Table
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ log(x), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -108.91  -44.69  -10.23   49.00  167.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.166     21.720   0.836    0.405
## log(x)        73.394      6.964  10.539   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.77 on 99 degrees of freedom
## Multiple R-squared:  0.5287, Adjusted R-squared:  0.524
## F-statistic: 111.1 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
# model 2
# ANOVA table
anova(model2)$Mean[2]
```

```
## [1] 3687.86
```

```
# Summary Table
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.534  -46.386   -8.103   43.404  158.662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      6.043  39.441  < 2e-16 ***
```

```
## poly(x, 2)1  601.855     60.728   9.911  < 2e-16 ***
## poly(x, 2)2 -163.561     60.728  -2.693  0.00832 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.73 on 98 degrees of freedom
## Multiple R-squared:  0.5184, Adjusted R-squared:  0.5085
## F-statistic: 52.74 on 2 and 98 DF,  p-value: 2.837e-16
```

```
# model 3
# ANOVA table
anova(model3)$Mean[2]
```

```
## [1] 3628.061
```

```
# Summary Table
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.434  -43.913   -8.613   47.679  169.432
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.993  39.765  < 2e-16 ***
## poly(x, 3)1  601.855     60.233   9.992  < 2e-16 ***
## poly(x, 3)2 -163.561     60.233  -2.715  0.00784 **
## poly(x, 3)3   97.408     60.233   1.617  0.10909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.23 on 97 degrees of freedom
## Multiple R-squared:  0.531,  Adjusted R-squared:  0.5165
## F-statistic: 36.61 on 3 and 97 DF,  p-value: 6.55e-16
```

```
# model 4
# ANOVA table
anova(model4)$Mean[2]
```

```
## [1] 3572.341
```

```
# Summary Table
summary(model4)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 4), data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.758  -39.885   -5.983   46.950  159.819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.947  40.073   <2e-16 ***
## poly(x, 4)1  601.855     59.769  10.070   <2e-16 ***
## poly(x, 4)2 -163.561     59.769  -2.737   0.0074 **
## poly(x, 4)3   97.408     59.769   1.630   0.1064
## poly(x, 4)4  -94.748     59.769  -1.585   0.1162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.77 on 96 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.5239
## F-statistic: 28.51 on 4 and 96 DF,  p-value: 1.286e-15
```

```
# model 5
# ANOVA table
anova(model5)$Mean[2]
```

```
## [1] 3609.788
```

```
# Summary Table
summary(model5)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 5), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.403  -39.569   -6.393   47.391  159.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.978  39.865  < 2e-16 ***
## poly(x, 5)1  601.855     60.082  10.017  < 2e-16 ***
## poly(x, 5)2 -163.561     60.082  -2.722  0.00771 **
## poly(x, 5)3   97.408     60.082   1.621  0.10827
## poly(x, 5)4  -94.748     60.082  -1.577  0.11812
## poly(x, 5)5   -3.854     60.082  -0.064  0.94898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.08 on 95 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.5189
## F-statistic: 22.58 on 5 and 95 DF,  p-value: 7.376e-15
```

The MSE for model 1 is 3572.03, the R-squared is 0.5287, and the adjusted R-squared is 0.524

The MSE for model 2 is 3687.86, the R-squared is 0.5184, and the adjusted R-squared is 0.5085

The MSE for model 3 is 3628.061, the R-squared is 0.531, and the adjusted R-squared is 0.5165

The MSE for model 4 is 3572.341, the R-squared is 0.543, and the adjusted R-squared is 0.5239

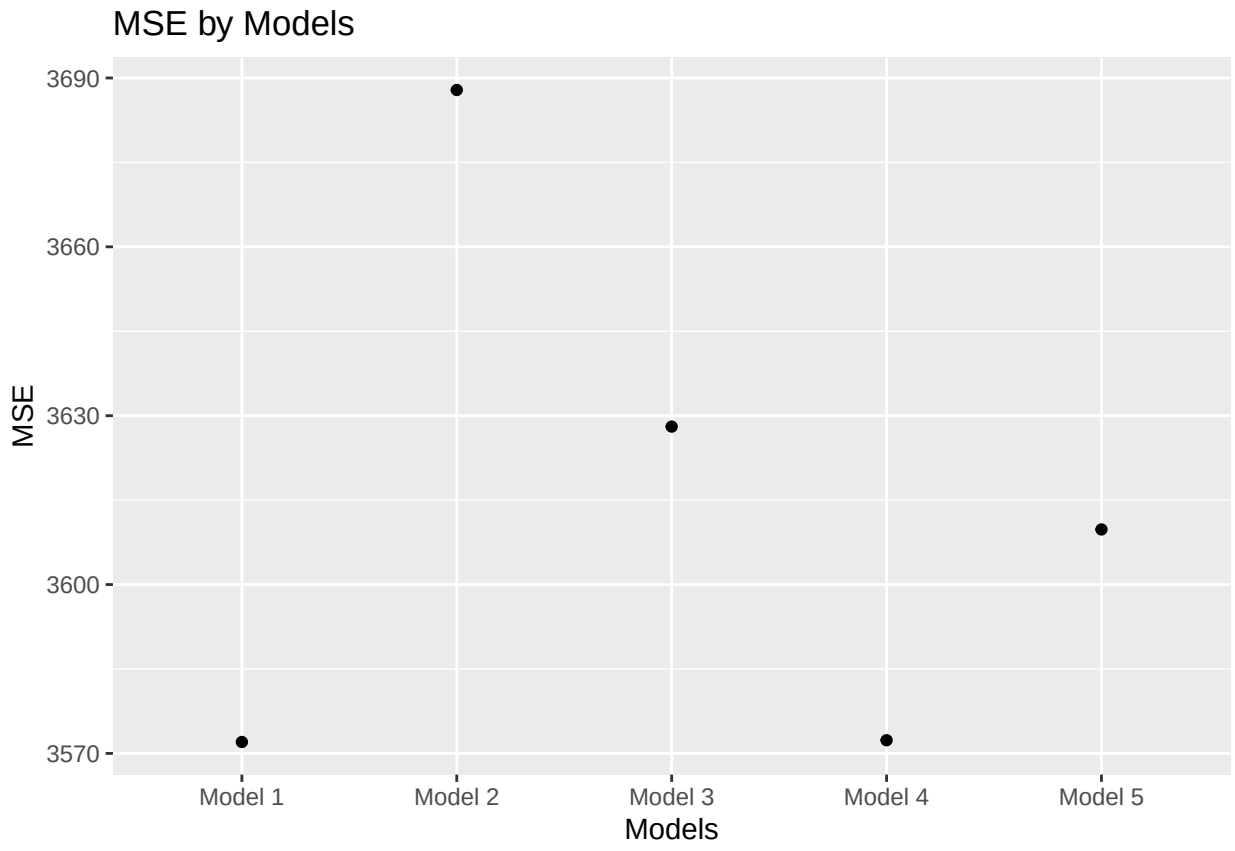
The MSE for model 5 is 3609.788, the R-squared is 0.543, and the adjusted R-squared is 0.5189

(b) Create three ggplots where the x axis is the model and the y axis is either MSE or Rsquared or R-squared adjusted.

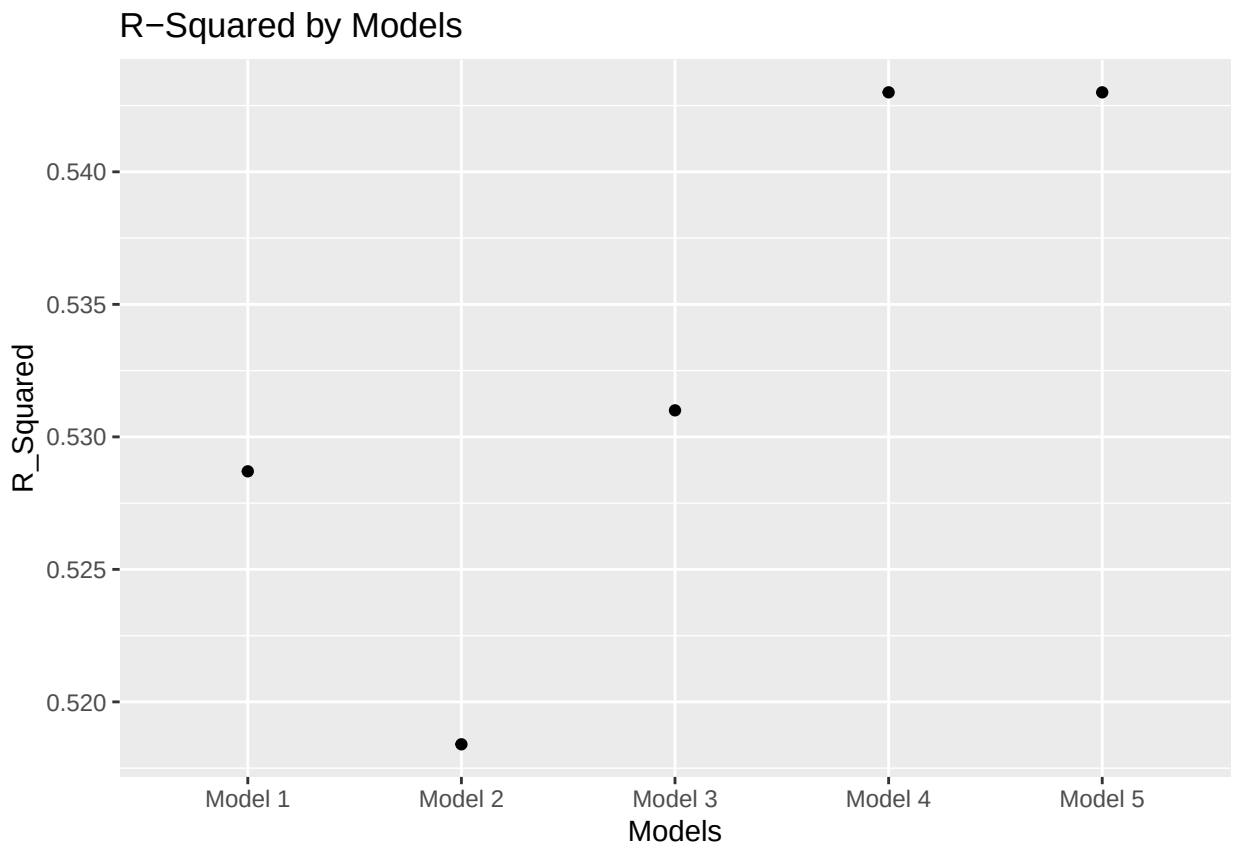
```
df1 <- data.frame(Models <- c("Model 1", "Model 2", "Model 3", "Model 4", "Model 5"), MSE <- c(3572.03,

colnames(df1) <- c("Models", "MSE", "R-Squared", "Adjusted R-Squared")

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

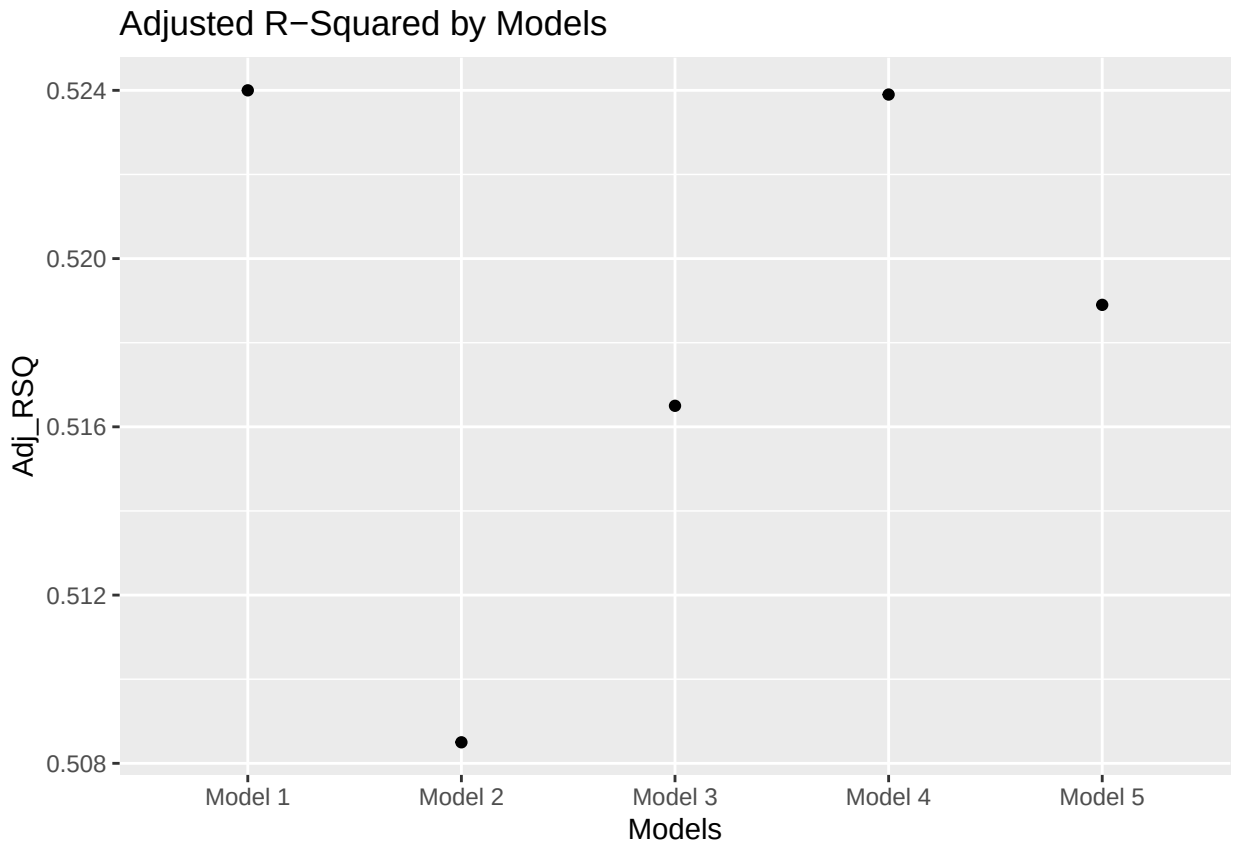
```
ggplot(data = df1, aes(Models, MSE)) + geom_point() + ggtitle("MSE by Models")
```




```
ggplot(data = df1, aes(Models, R_Squared)) + geom_point() + ggtitle("R-Squared by Models")
```




```
ggplot(data = df1, aes(Models, Adj_RSQ)) + geom_point() + ggtitle("Adjusted R-Squared by Models")
```



(c) Based on MSE_training plot, which model would you choose as the best fit?

Based on the MSE training plot, we will choose model 1 or model 4 since they have the smallest MSE

(d) Based on the R-Squared adjusted plots which model would you choose as the best fit?

Based on the R-Squared Adjusted Plot, we will choose model 1 or 4 since they have the highest R-Squared

(e) Report the partial slopes for each model.

```
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ log(x), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -108.91  -44.69  -10.23   49.00  167.06
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.166     21.720   0.836    0.405
## log(x)        73.394      6.964  10.539   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.77 on 99 degrees of freedom
## Multiple R-squared:  0.5287, Adjusted R-squared:  0.524
## F-statistic: 111.1 on 1 and 99 DF,  p-value: < 2.2e-16
```

The partial slope for Model 1 can be found through the summary function. b_1 = 73.394

```
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.534  -46.386   -8.103   43.404  158.662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      6.043  39.441  < 2e-16 ***
## poly(x, 2)1  601.855     60.728   9.911  < 2e-16 ***
## poly(x, 2)2 -163.561     60.728  -2.693  0.00832 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.73 on 98 degrees of freedom
## Multiple R-squared:  0.5184, Adjusted R-squared:  0.5085
## F-statistic: 52.74 on 2 and 98 DF,  p-value: 2.837e-16
```

The partial slopes for Model 2 can be found through the summary function. b_1 = 601.855, b_2 = -163.561

```
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.434  -43.913   -8.613   47.679  169.432
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.993  39.765  < 2e-16 ***
## poly(x, 3)1  601.855     60.233   9.992  < 2e-16 ***
## poly(x, 3)2 -163.561     60.233  -2.715  0.00784 **
```

```
## poly(x, 3)3   97.408     60.233   1.617  0.10909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.23 on 97 degrees of freedom
## Multiple R-squared:  0.531,  Adjusted R-squared:  0.5165
## F-statistic: 36.61 on 3 and 97 DF,  p-value: 6.55e-16
```

The partial slopes for Model 3 can be found through the summary function. b_1 = 601.855, b_2 = -163.561, b_3 = 97.408

```
summary(model4)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 4), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.758  -39.885   -5.983   46.950  159.819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.947  40.073   <2e-16 ***
## poly(x, 4)1  601.855     59.769  10.070   <2e-16 ***
## poly(x, 4)2 -163.561     59.769  -2.737   0.0074 **
## poly(x, 4)3   97.408     59.769   1.630   0.1064
## poly(x, 4)4  -94.748     59.769  -1.585   0.1162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.77 on 96 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.5239
## F-statistic: 28.51 on 4 and 96 DF,  p-value: 1.286e-15
```

The partial slopes for Model 4 can be found through the summary function b_1 = 601.855, b_2 = -163.561, b_3 = 97.408, b_4 = -94.748

```
summary(model5)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 5), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.403  -39.569   -6.393   47.391  159.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.326      5.978  39.865  < 2e-16 ***
## poly(x, 5)1  601.855     60.082  10.017  < 2e-16 ***
```

```
## poly(x, 5)2 -163.561     60.082  -2.722  0.00771 **
## poly(x, 5)3   97.408     60.082   1.621  0.10827
## poly(x, 5)4  -94.748     60.082  -1.577  0.11812
## poly(x, 5)5   -3.854     60.082  -0.064  0.94898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.08 on 95 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.5189
## F-statistic: 22.58 on 5 and 95 DF,  p-value: 7.376e-15
```

The partial slopes for Model 5 can be found through the summary function b_1 = 601.855, b_2 = -163.561, b_3 = 97.408, b_4 = -94.748, b_5 = -3.854

(f) Now upload HW1Test.csv posted on bruinlearn week 2. Use the five models you got from part (a) to predict the y-values for x values in the testing data. Then, use the generated y-values (predicted values) along with the testing y values to compute the MSE_testing for each of the five models. (Hint: use the predict() command to get the predicted y values for the vector x).

```
testVal <- read.csv("/Users/takaooba/Downloads/HW1Test.csv")
testVal <- testVal[,c(2,3)]
predict1 <- predict(model1, newdata = testVal)
predict2 <- predict(model2, newdata = testVal)
predict3 <- predict(model3, newdata = testVal)
predict4 <- predict(model4, newdata = testVal)
predict5 <- predict(model5, newdata = testVal)
MSE_1 <- mean((testVal[,2] - predict.lm(model1, testVal))^2)
MSE_2 <- mean((testVal[,2] - predict.lm(model2, testVal))^2)
MSE_3 <- mean((testVal[,2] - predict.lm(model3, testVal))^2)
MSE_4 <- mean((testVal[,2] - predict.lm(model4, testVal))^2)
MSE_5 <- mean((testVal[,2] - predict.lm(model5, testVal))^2)
cat("Solved the MSE for each model through the formula given in the lectures. This is biased since it is
```

```
## Solved the MSE for each model through the formula given in the lectures. This is biased since it is
```

```
cat("The MSE for model 1 utilizing (1/n)*.... is ", as.character(MSE_1), "\n")
```

```
## The MSE for model 1 utilizing (1/n)*.... is  3756.98878669711
```

```
cat("The MSE for model 2 utilizing (1/n)*.... is ", as.character(MSE_2), "\n")
```

```
## The MSE for model 2 utilizing (1/n)*.... is  3731.96440929781
```

```
cat("The MSE for model 3 utilizing (1/n)*.... is ", as.character(MSE_3), "\n")
```

```
## The MSE for model 3 utilizing (1/n)*.... is  3827.5639456516
```

```
cat("The MSE for model 4 utilizing (1/n)*.... is ", as.character(MSE_4), "\n")
```

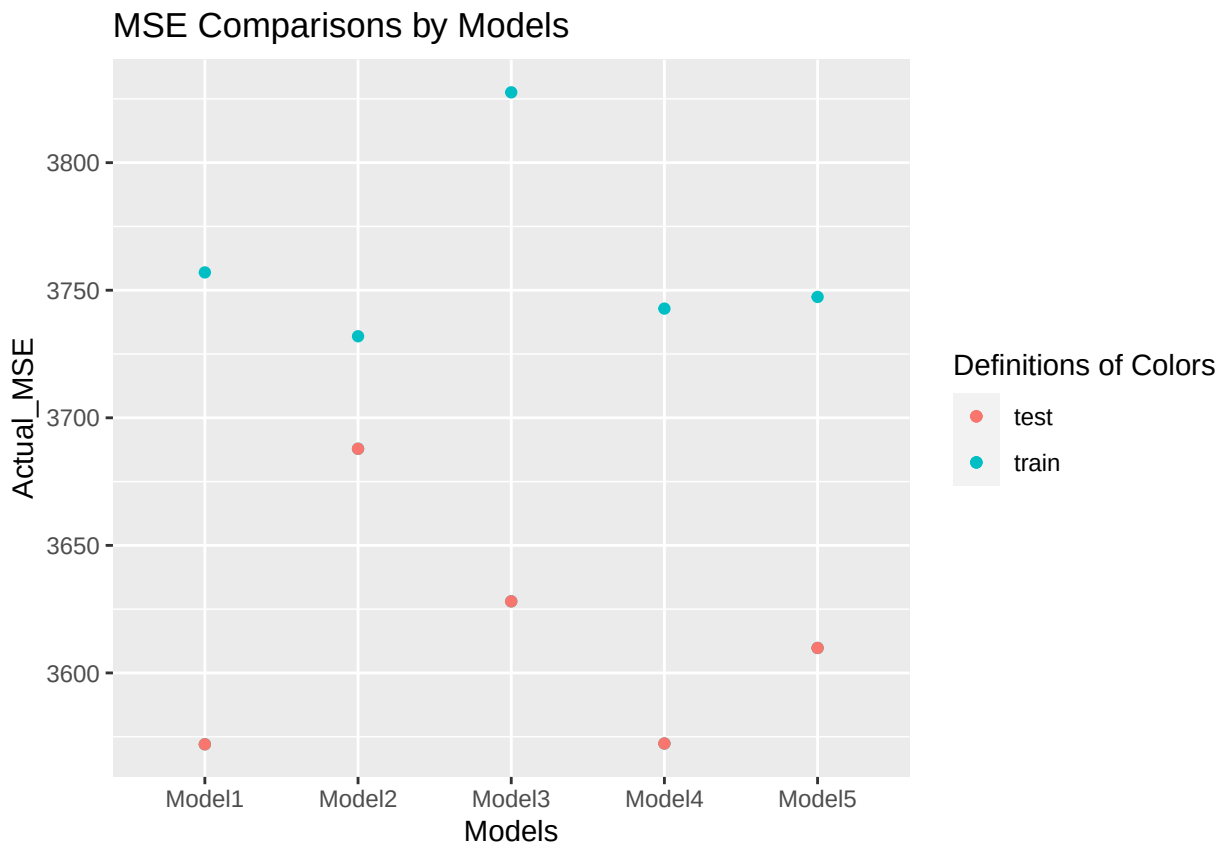
```
## The MSE for model 4 utilizing (1/n)*.... is  3742.80522856432
```

```
cat("The MSE for model 5 utilizing (1/n)*.... is ", as.character(MSE_5), "\n")
```

```
## The MSE for model 5 utilizing (1/n)*.... is  3747.37425167557
```

(g) Write a sentence or two describing how the MSE_testing and MSE_training compared to each other for each model. Now based on your answers to part b and part c and your MSEs, which of the five models you think is the true model used to create such data? (preferable to create a ggplot using the 5 models vs the testing and training MSEs. "Use two different colors to distinguish testing MSEs from Training MSEs"

```
ActualMSE <- c(3572.03, 3687.86, 3628.061, 3572.341, 3609.788)
PredictedMSE <- c(3756.98878669711, 3731.96440929781,
                  3827.5639456516, 3742.80522856432, 3747.37425167557)

df2 <- data.frame(Models = c("Model1","Model2", "Model3", "Model4", "Model5"),
                  Actual_MSE = ActualMSE, Predicted_MSE = PredictedMSE)

ggplot(data = df2, aes(Models, Actual_MSE, color = "train")) + geom_point() +
  geom_point(data = df2,  aes(Models, Predicted_MSE))+ geom_point(aes(color = "test")) +
  ggtitle("MSE Comparisons by Models") + labs(color = "Definitions of Colors")
```



Question 2

Check out the Heart Data file heart.csv on the bruinlearn under Week 2. The Heart data set contain a binary outcome HD for 303 patients who presented with chest pain. An outcome value of Yes indicates the presence of heart disease based on an angiographic test, while No means no heart disease. There are 13 predictors including Age, Sex, Chol (a cholesterol measurement), Thal (Thallium stress test) and other heart and lung function measurements. Some of these variables are categorical and some are numerical.

(a) Write four questions that could be answered with these data. Two of your questions should be questions that require estimating the parameters, and two should be prediction questions. Indicate which question is a 'parameter estimation' question and which is a 'prediction' question. (The Statistical Learning textbook classifies these as 'inference' vs. 'prediction' questions.)

```
heart <- read.csv("/Users/takaooba/Downloads/Heart.csv")
heart <- heart[,-1]
head(heart)
```

```
##   Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca
```

```
## 1  63   1      typical    145  233   1       2   150     0     2.3     3  0
## 2  67   1 asymptomatic    160  286   0       2   108     1     1.5     2  3
## 3  67   1 asymptomatic    120  229   0       2   129     1     2.6     2  2
## 4  37   1   nonanginal    130  250   0       0   187     0     3.5     3  0
## 5  41   0   nontypical    130  204   0       2   172     0     1.4     1  0
## 6  56   1   nontypical    120  236   0       0   178     0     0.8     1  0
##         Thal AHD
## 1      fixed  No
## 2     normal Yes
## 3 reversable Yes
## 4     normal  No
## 5     normal  No
## 6     normal  No
```

In the Statistical Learning textbook, the author expands on the difference between inference and predictions. For inference, utilizing a set of data, the researcher intends to "infer" how the output generates as a function of the data. In comparison,for prediction, the researcher will be given a new measurement and will aim to use the current data to predict or choose the correct outcome.

Parameter Estimation Questions

1. What quantitative predictors are associated with experiencing heart disease?
2. Are men more likely to experience heart disease?

Prediction Question

1. Out of the new batch of patients, who will experience heart disease?
2. Given someone is a 50 year old man with a cholesterol of 260, is he safe from a heart disease?

(b) Choose one of your questions and answer it. It doesn't have to be a good answer (so you don't have to find the best model or justify the model), but you do have to give the answer to your question and explain how you answered it.

Are men more likely to experience heart disease?

To approach this problem, we will assess the questions through the null hypothesis of "There is no significant difference between the observed and the expected value"

We will look at the predictor "Sex" which has the vales 0 and 1.

```
unique(heart[,2])
```

```
## [1] 1 0
```

We will conduct a chi test goodness of fit to assess this question.

```
heartTemp <- heart[,c(2,14)]
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## 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
```

```
heartTemp %>% group_by(AHD) %>% count(Sex)
```

```
## # A tibble: 4 x 3
## # Groups:   AHD [2]
##   AHD     Sex     n
##   <chr> <int> <int>
## 1 No        0    72
## 2 No        1    92
## 3 Yes       0    25
## 4 Yes       1   114
```

```
# Conducting chi squared test (GOF)
testingVal <- c(72,92,25,114)
chisq.test(testingVal, p = c(1/4,1/4,1/4,1/4))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  testingVal
## X-squared = 56.987, df = 3, p-value = 2.587e-12
```

We have that the test statistic is 56.987. Further, we have that the p-value is 2.587 e-12 < 0.05. Thus, we have that the test statistic is significant. Thus, there is a significant difference between the observed and the expected value.

Question 3 The objective of a study was to find out whether there was any relationship between level of education (high school, four-year college, and graduate work) and attitude toward pre-screening for breast cancer (i.e., going for mammograms). A sample size of 512 were selected for this study. Complete the anova table. Find the numerical values of A,B, C, D, E, F, G, and H

Attached below is the answers to the respective slots from A-H with works shown

Q3) The objective of a study was to find out whether there was any relationship between level of education (high school, four-year college, and graduate work) and attitude toward pre-screening for breast cancer (i.e., going for mammograms). A sample size of 512 were selected for this study. Complete the anova table. Find the numerical values of A,B, C, D, E, F, G, and H.

	Sum of Squares	degrees of freedom	Mean Square	F	P-Value
Between group	A	B	E	G	H
Within group	51912.079	C	F		
Total	53727.147	D			

$A = SS_{Total} - SS_{within}$

$= 53727.147 - 51912.079$

$= \boxed{1815.068}$

$a = 3$

$n = 512$

$B = a - 1 = 3 - 1 = \boxed{2}$

$C = (n-1) - (a-1) = (512-1) - (3-1) = 511 - 2 = \boxed{509}$

$D = n - 1 = 512 - 1 = \boxed{511}$

$E = SS_{between} / df_{between}$

$= \frac{SS_{between}}{a-1} = \frac{1815.068}{2} = \boxed{907.534}$

$F = \frac{SS_{within}}{Df_{within}} = \frac{SS_{within}}{(n-1)-(a-1)} = \frac{51912.079}{509} = \boxed{101.988}$

$G = MS_{between} / MS_{within} = \frac{907.534}{101.988} = \boxed{8.898}$

$H = 1 - pf(5.2, 2, 509) = \boxed{0.005813}$

$0.005813 < 0.05$

↳ Significant!

Question 4

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   0.3.5
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'purrr' was built under R version 4.1.2
```

```
## Warning: package 'stringr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Introduce read_csv
auto <- read_csv("/Users/takaooba/Downloads/Auto.csv")
```

```
## Rows: 397 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): horsepower, name
## dbl (7): mpg, cylinders, displacement, weight, acceleration, year, origin
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(auto)
```

```
## [1] 397   9
```

```r
# Determine if there is any NAs
apply(X = is.na(auto), MARGIN = 2, FUN = sum)
```

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

```r
# However we introduce "?"
auto[auto$horsepower == "?", ]
```

```
## # A tibble: 5 x 9
##     mpg cylinders displacement horsepower weight acceleration  year origin name
##   <dbl>     <dbl>        <dbl> <chr>       <dbl>        <dbl> <dbl>  <dbl> <chr>
## 1  25           4           98 ?            2046         19      71      1 ford~
## 2  21           6          200 ?            2875         17      74      1 ford~
## 3  40.9         4           85 ?            1835         17.3    80      2 rena~
## 4  23.6         4          140 ?            2905         14.3    80      1 ford~
## 5  34.5         4          100 ?            2320         15.8    81      2 rena~
```

```r
auto$horsepower <- as.numeric(auto$horsepower)
```

```
## Warning: NAs introduced by coercion
```

```r
auto <- na.omit(auto)

# Have successfully omitted the five NA values
dim(auto)
```

```
## [1] 392   9
```

```r
# Notice that the dimension was originally 397 by 9 and now is 392 by 9

auto[auto$horsepower == "?", ]
```

```
## # A tibble: 0 x 9
## # ... with 9 variables: mpg <dbl>, cylinders <dbl>, displacement <dbl>,
## #   horsepower <dbl>, weight <dbl>, acceleration <dbl>, year <dbl>,
## #   origin <dbl>, name <chr>
```

```r
# Notice there are no "?"
```

(a) Which of the predictors are quantitative, and which are qualitative?

```r
sapply(auto, class)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
##         year       origin         name
##    "numeric"    "numeric"  "character"
```

Numeric does not guarantee quantitative, so we will further assess through the head function

```
head(auto)
```

```
## # A tibble: 6 x 9
##     mpg cylinders displacement horsepower weight acceleration  year origin name
##   <dbl>     <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <dbl>  <dbl> <chr>
## 1    18         8          307        130   3504         12      70      1 chev~
## 2    15         8          350        165   3693         11.5    70      1 buic~
## 3    18         8          318        150   3436         11      70      1 plym~
## 4    16         8          304        150   3433         12      70      1 amc ~
## 5    17         8          302        140   3449         10.5    70      1 ford~
## 6    15         8          429        198   4341         10      70      1 ford~
```

Qualitative predictors are "year", "origin", "name"

Quantitative predictors are : "mpg" "cylinders", "displacement", "horsepower", "weight", "acceleration"

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
autoInter0 <- auto[,c(1,2,3,4,5,6)]
apply(autoInter0, MARGIN = 2, range)
```

```
##       mpg cylinders displacement horsepower weight acceleration
## [1,]  9.0         3           68         46   1613          8.0
## [2,] 46.6         8          455        230   5140         24.8
```

```
# range(auto$mpg)
# range(auto$cylinders)
# range(auto$displacement)
# range(auto$horsepower)
# range(auto$weight)
# range(auto$acceleration)
```

mpg: Range is 9.0 and 46.6

cylinders: Range is 3 and 8

displacement: Range is 68 and 455

horsepower: Range is 46 and 230

weight: Range is 1613 and 5140

acceleration: Range is 8.0 and 24.8

(c) What is the mean and standard deviation of each quantitative predictor?

```
apply(autoInter0, MARGIN = 2, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
```

```
apply(autoInter0, MARGIN = 2, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
```

Mean of mpg: 23.44592

SD of mpg: 7.805007

Mean of cylinders: 5.471939

SD of cylinders: 1.705783

Mean of displacement: 194.412

SD of displacement: 104.644

Mean of horsepower: 104.4694

SD of of horsepower: 38.49116

Mean of weight: 2977.584

SD of weight: 849.4026

Mean of acceleration: 15.54133

SD of acceleration: 2.758864

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
auto1 <- auto[-c(10:85), ]
head(auto1)
```

```
## # A tibble: 6 x 9
##     mpg cylinders displacement horsepower weight acceleration  year origin name
##   <dbl>     <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <dbl>  <dbl> <chr>
## 1    18         8          307        130   3504         12      70      1 chev~
## 2    15         8          350        165   3693         11.5    70      1 buic~
## 3    18         8          318        150   3436         11      70      1 plym~
## 4    16         8          304        150   3433         12      70      1 amc ~
## 5    17         8          302        140   3449         10.5    70      1 ford~
## 6    15         8          429        198   4341         10      70      1 ford~
```

```
autoInter <- auto1[,c(1,2,3,4,5,6)]

apply(autoInter, MARGIN = 2, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
```

```
apply(autoInter, MARGIN = 2, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
```

```
apply(autoInter, MARGIN = 2, range)
```

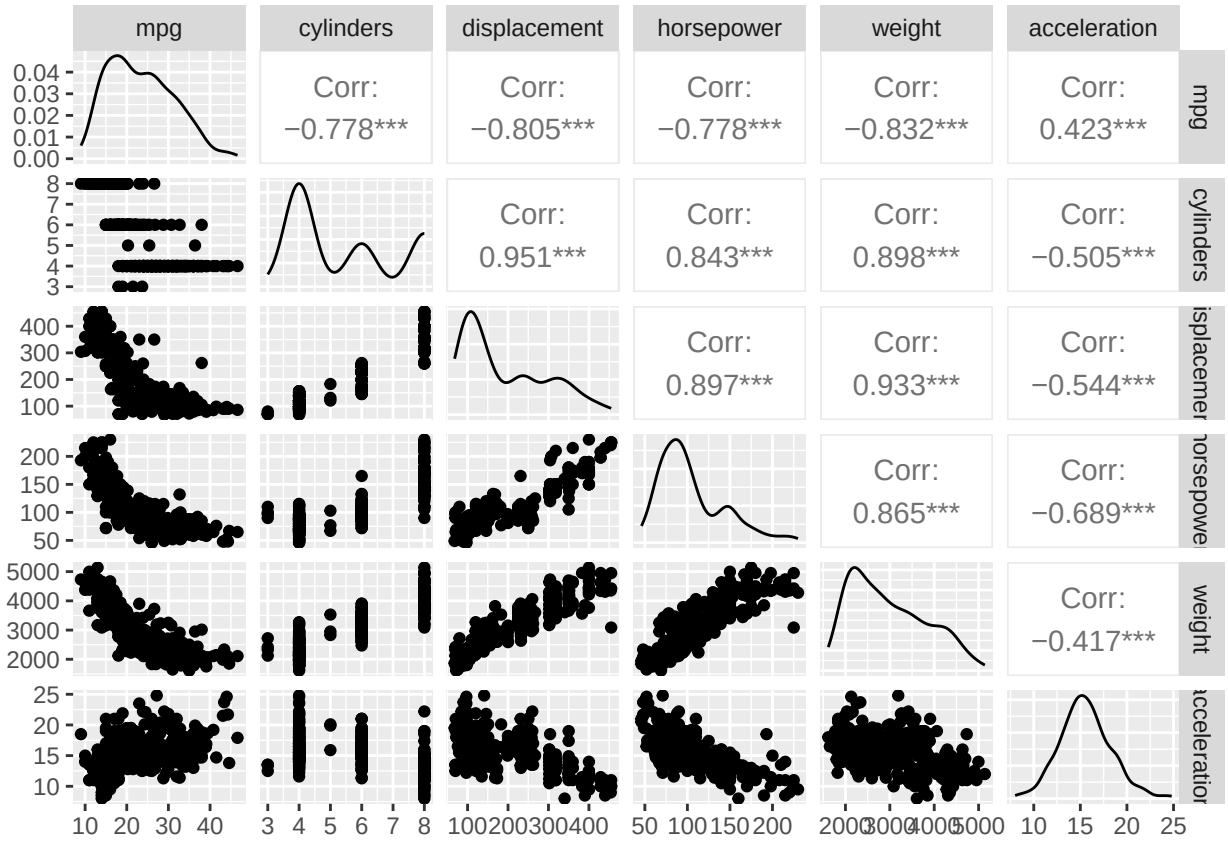
```
##       mpg cylinders displacement horsepower weight acceleration
## [1,] 11.0         3           68         46   1649          8.5
## [2,] 46.6         8          455        230   4997         24.8
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
#install.packages("GGally")
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
autoQuantitative <- auto[,c(1,2,3,4,5,6)]
ggpairs(autoQuantitative)
```




```
autoCategorical <- auto[,c(7,8,9)]
```

There were some interesting findings through the plot above. There is a high correlation between horsepower and cylinder, horsepower and displacement and a negative relationship between horsepower and mpg. This makes sense as more cylinders generally indicates more power. For example, a ferrari 812 GTS is a v12 indicating it has 12 cylinders and makes 788 hp. On the contrary, a toyota prius is a 4 cylinder car making 121 hp. Additionally, the negative relationship between horsepower and mpg makes sense because for example, a prius does around 50 mpg where the ferrari 812 GTS does around 12 mpg. As a car lover, it is amazing to visually see these relationships amongst the various components of a car.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

```
cor(autoQuantitative)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
```

```
##              acceleration
## mpg             0.4233285
## cylinders      -0.5046834
## displacement   -0.5438005
## horsepower     -0.6891955
## weight         -0.4168392
## acceleration    1.0000000
```

From the investigation that we have generated in the previous part, it suffice to say that the "weight" variable is best in predicting mpg because they have the highest correlation. Then, the next best predictor will be "displacement" with the second highest correlation. Further, considering the numeric variables, we will conclude that the order of correlation will be the order of "horsepower", "cylinders" and "acceleration" will be the worst predictor with the lowest correlation.

We will assess this through an anova test

```
anova(lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration, data = auto))
```

```
## Analysis of Variance Table
##
## Response: mpg
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## cylinders      1 14403.1 14403.1 798.5086 < 2.2e-16 ***
## displacement   1  1073.3  1073.3  59.5063 1.050e-13 ***
## horsepower     1   403.4   403.4  22.3650 3.166e-06 ***
## weight         1   975.7   975.7  54.0943 1.156e-12 ***
## acceleration   1     1.0     1.0   0.0536    0.8171
## Residuals    386  6962.5    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice how acceleration is not significant and the p-value is 0.8171.

Reassessing the model with the significant predictors, we have that

```
anova(lm(mpg ~ cylinders + displacement + horsepower + weight, data = auto))
```

```
## Analysis of Variance Table
##
## Response: mpg
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## cylinders      1 14403.1 14403.1 800.466 < 2.2e-16 ***
## displacement   1  1073.3  1073.3  59.652 9.795e-14 ***
## horsepower     1   403.4   403.4  22.420 3.079e-06 ***
## weight         1   975.7   975.7  54.227 1.085e-12 ***
## Residuals    387  6963.4    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we have that the predictors "cylinders" "displacement" "horsepower" and "weight" are all significant.

Question 5

This exercise involves the Boston housing data set

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

How many rows are in this data set? How many columns? What do the rows and columns represent?

```
library (MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
##
## Attaching package: 'MASS'
```

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

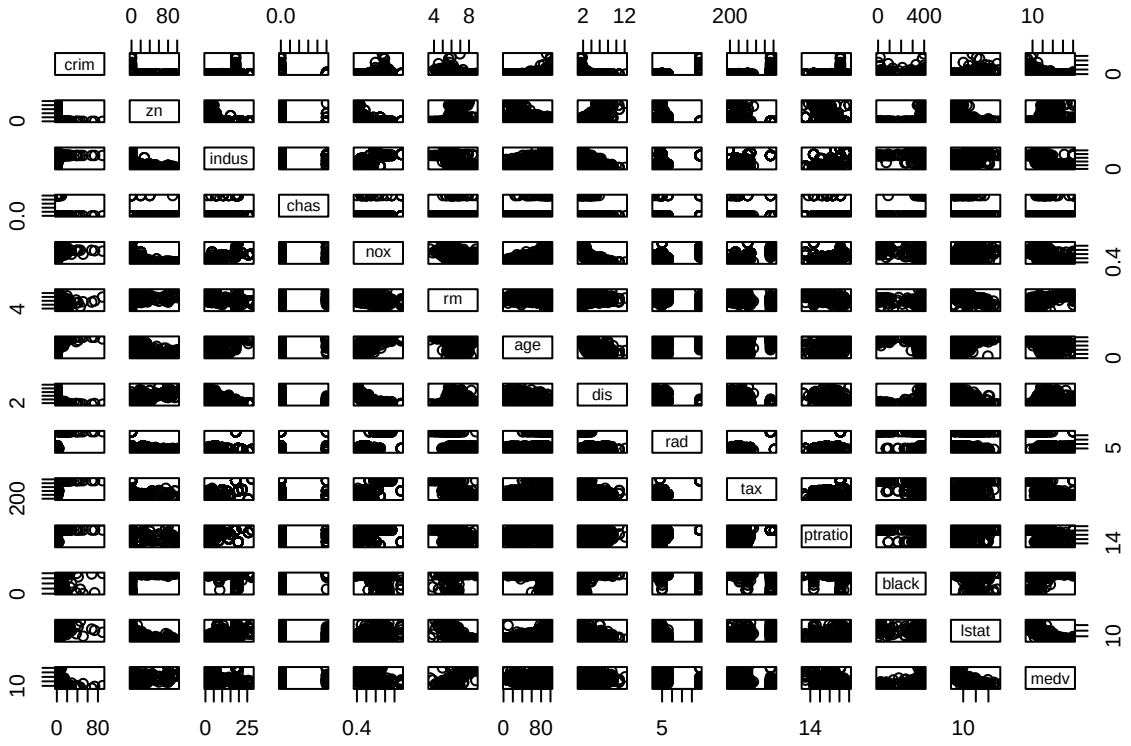
```
# Boston
?Boston
dim(Boston)
```

```
## [1] 506  14
```

There are 506 rows and 14 columns Each row represents the housing observations and the column represents the various predictors.

b Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings

```
Boston$chas <- as.numeric(Boston$chas)
Boston$rad <- as.numeric(Boston$rad)
pairs(Boston)
```



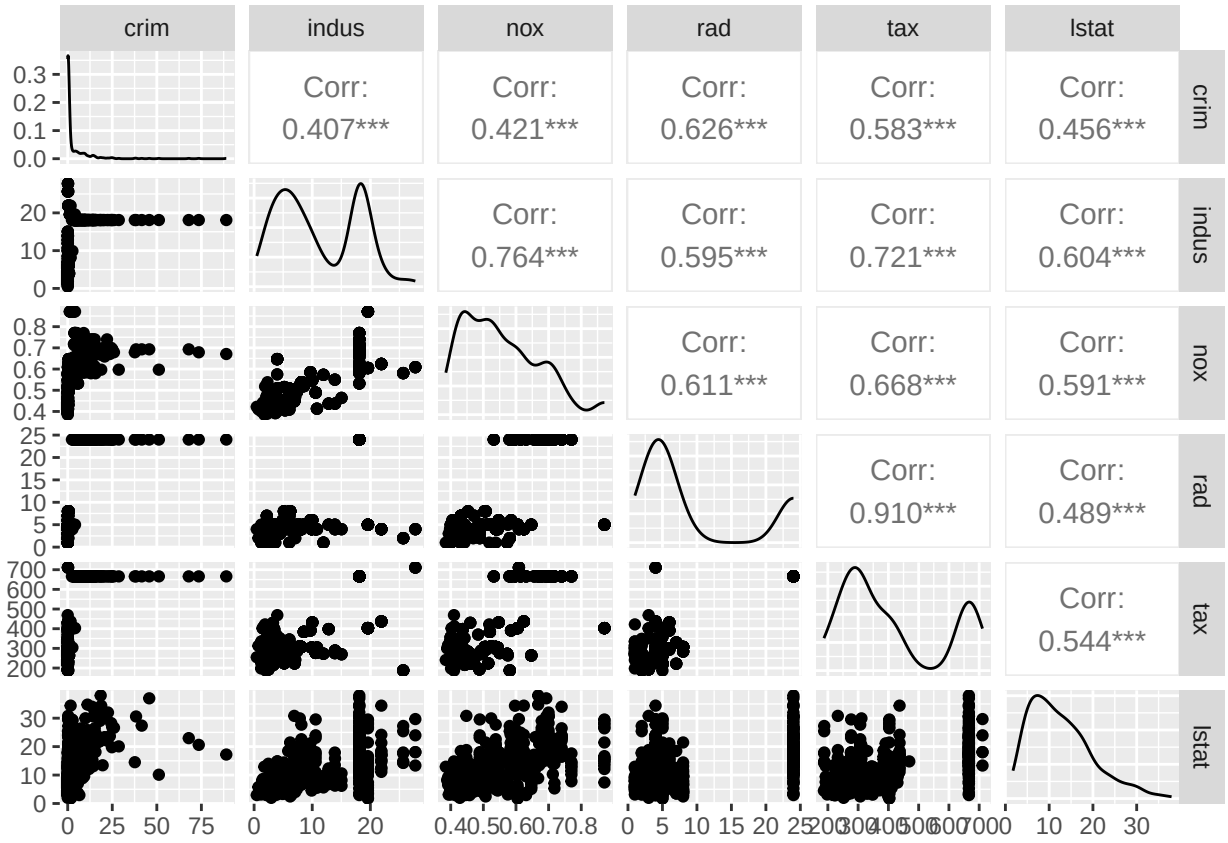

```
cor(Boston)
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
```

```
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801 -0.1773833
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##               black      lstat       medv
## crim    -0.38506394  0.4556215 -0.3883046
## zn       0.17552032 -0.4129946  0.3604453
## indus   -0.35697654  0.6037997 -0.4837252
## chas     0.04878848 -0.0539293  0.1752602
## nox     -0.38005064  0.5908789 -0.4273208
## rm       0.12806864 -0.6138083  0.6953599
## age     -0.27353398  0.6023385 -0.3769546
## dis      0.29151167 -0.4969958  0.2499287
## rad     -0.44441282  0.4886763 -0.3816262
## tax     -0.44180801  0.5439934 -0.4685359
## ptratio -0.17738330  0.3740443 -0.5077867
## black    1.00000000 -0.3660869  0.3334608
## lstat   -0.36608690  1.0000000 -0.7376627
## medv     0.33346082 -0.7376627  1.0000000
```

If we were to go directly into making the pairwise scatter plot, although we can see the general picture, it is difficult to assess the details. We will take the predictors with the highest correlation with capita crime rate.

```
bostonInter <- Boston[,c(1,3,5,9, 10, 13)]
ggpairs(bostonInter)
```

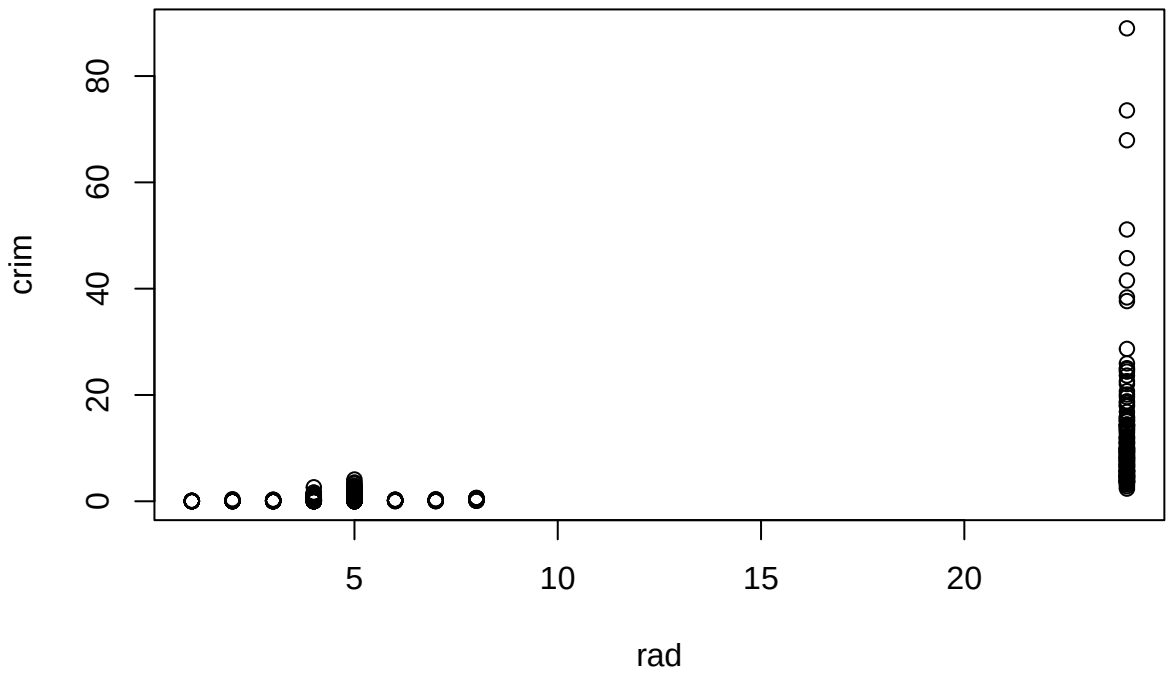


From above, we can see the scatter plot for some of the predictors as well as the correlations. Considering the "crim" or per capita crime rate by town to be the outcome, we have that the predictors "rad" "tax" and "lstat" are the best predictors for "crim".

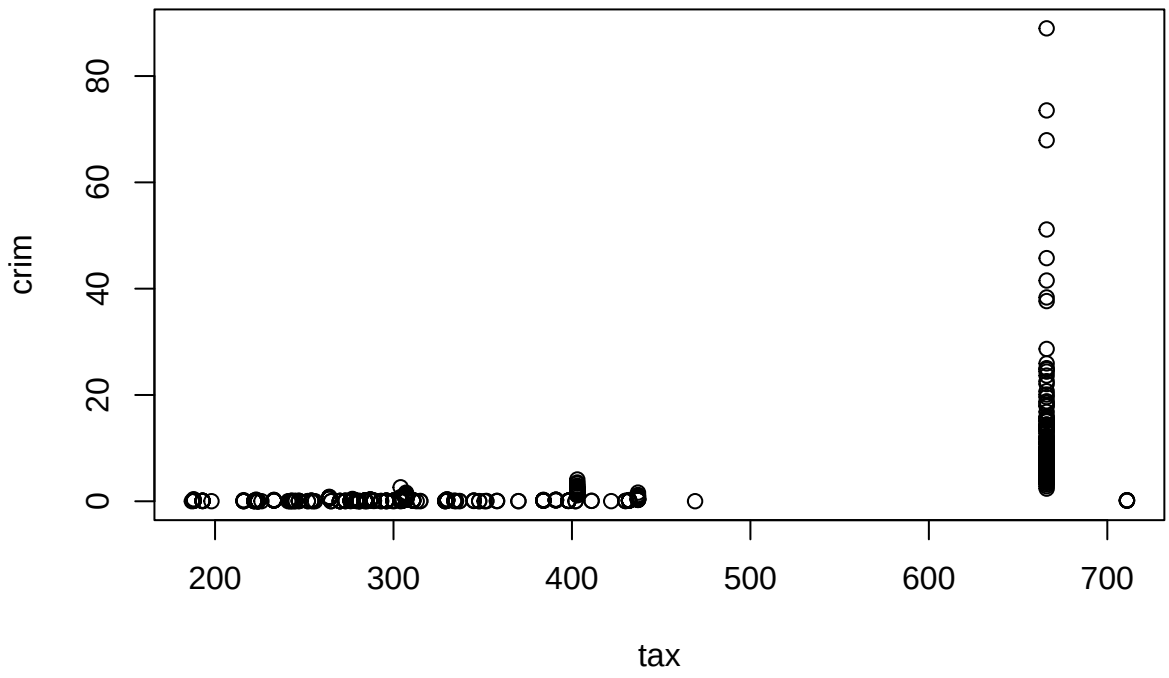
(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

The predictors that appear to be associated with per capita crime rate are "rad" "tax" and "lstat" through looking at the correlation with "crim"

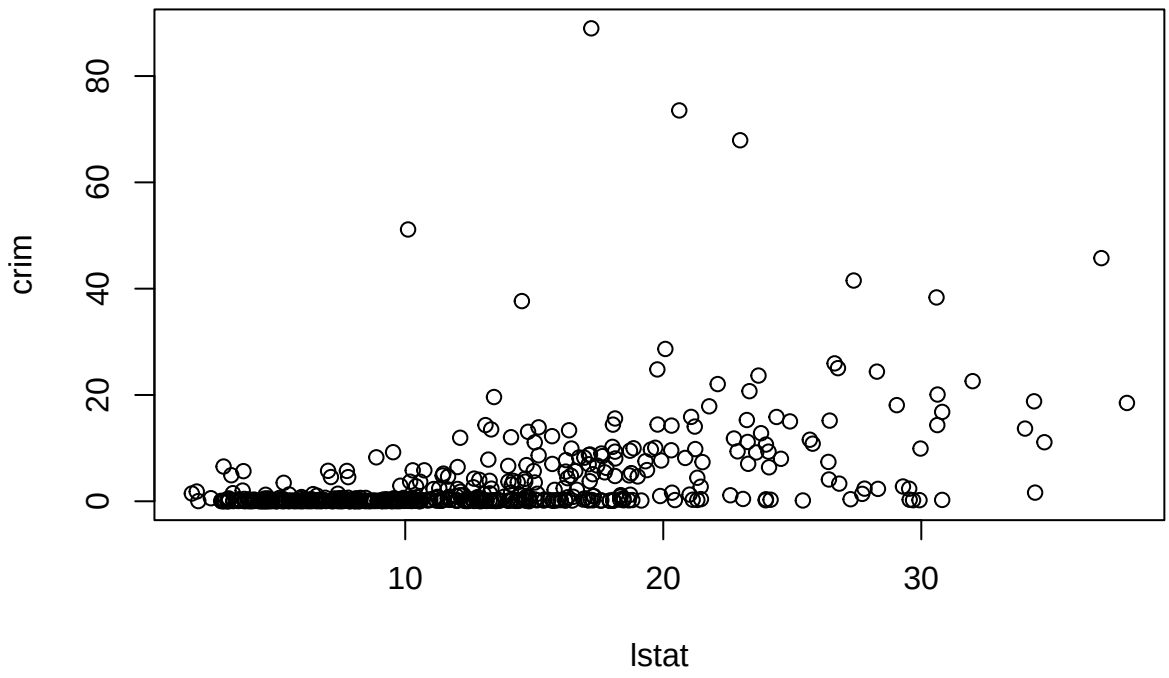
```
plot(crim ~rad, data = Boston)
```




```
plot(crim ~ tax, data = Boston)
```




```
plot(crim ~ lstat, data= Boston)
```

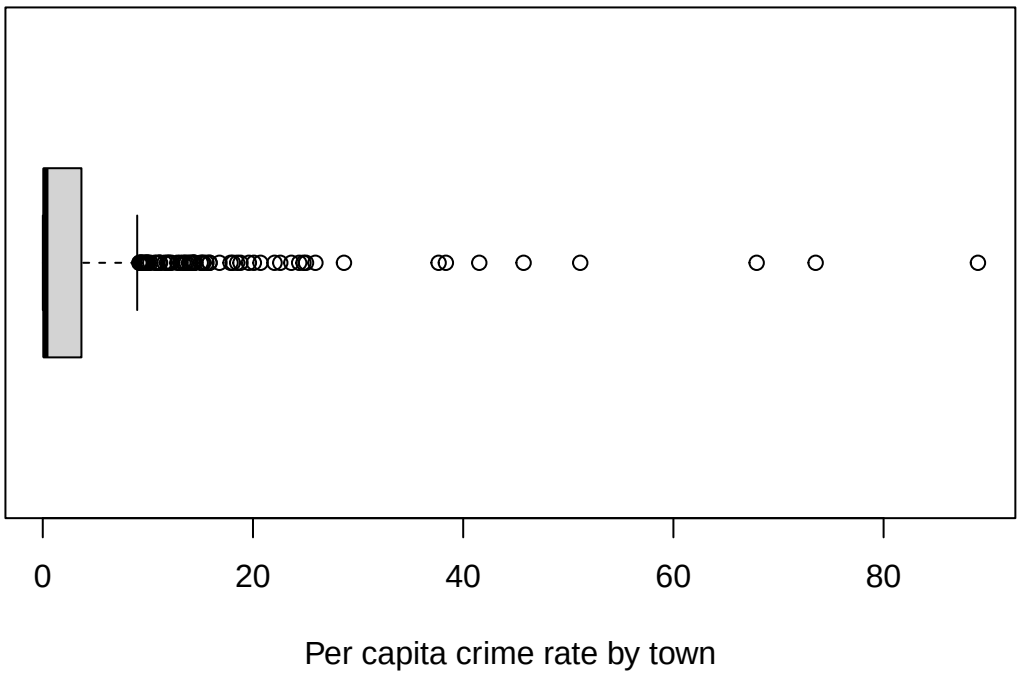


We assess these three predictors through the correlation plot and we do in fact see that these predictors are associated with per capita crime rate.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

We will consider an observation to be in the higher outlier to be considered "particularly high". We can see if an observation is an outlier simply through looking at a boxplot.

```
boxplot(Boston$crim, horizontal = TRUE, xlab = "Per capita crime rate by town")
```



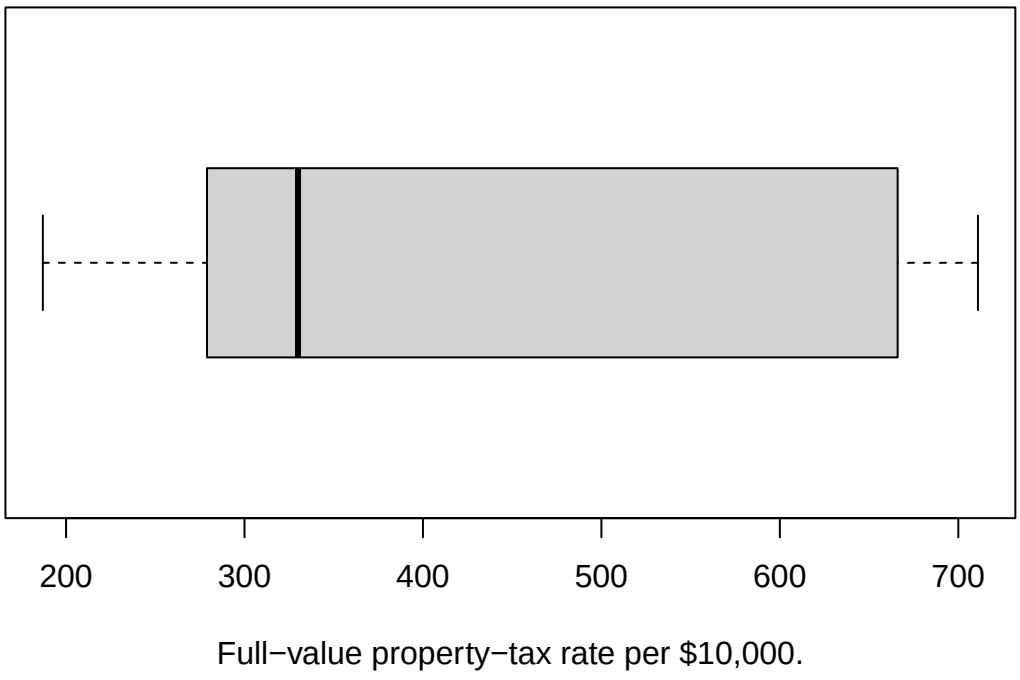
There seems to be some suburbs of Boston to have particularly high crime rates.

```
range(Boston$crim)
```

```
## [1]  0.00632 88.97620
```

The range of crime rate in Boston ranges from 0.00632 to 88.9762

```
boxplot(Boston$tax, horizontal = TRUE, xlab = "Full-value property-tax rate per $10,000.")
```



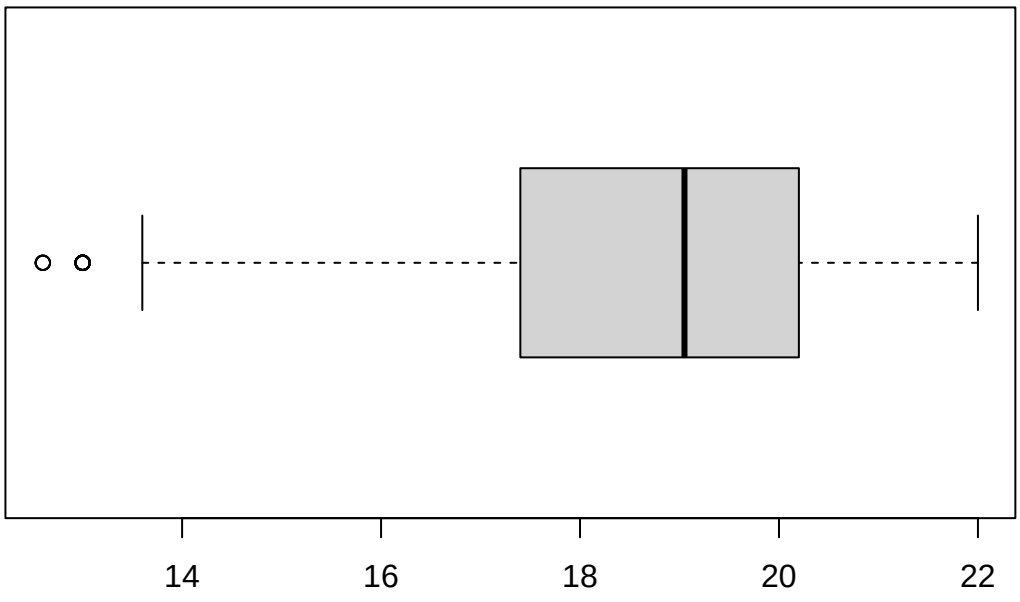
There does not seem to be any suburbs of Boston with high property-tax rate.

```
range(Boston$tax)
```

```
## [1] 187 711
```

The range of the property-tax rate in Boston is 187 and 711

```
boxplot(Boston$ptratio, horizontal = TRUE, xlab = "Pupil-teacher ratio by town.")
```



Pupil−teacher ratio by town.

There does not seem to be any suburbs of Boston with particularly high pupil-teacher ratio.

```
range(Boston$ptratio)
```

```
## [1] 12.6 22.0
```

The range of the pupil-teacher ratio is from 12.6 to 22.0.

(e) How many of the suburbs in this data set bound the Charles river?

To assess the amount of suburbs that bound the Charles river, we will look at the predictor "chas". When looking at the R Documentation for this particular variable, we have that 1 indicates that the observation bounds the river, thus we would simply count the amount of observations in the data frame with "chas" = 1

```
sum(Boston$chas == 1)
```

```
## [1] 35
```

35 suburbs were bounded by the Charles river.

(f) What is the median pupil-teacher ratio among the towns in this data set?

We will focus on the "ptratio" predictor

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

The median pupil-teacher ratio among the towns in this data set is 19.05

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
# The lowest median value of owner occupied homes is given below
min(Boston$medv)
```

```
## [1] 5
```

```
# Then we will look at the indeces where the lowest median value occurs
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
```

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
```

```
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Comparing the values from indeces 399 and 406 to the overall ranges of those predictors, we notice that

crim - The crime rate is particularly higher than other observations for both 399 and 406.

zn - Both 0.0 for observations 399 and 406 which is the minimum as well as the median for all observations.

indus - The proportion of non-retail business acres per town is particularly high for both observations.

chas - Both not bounded by river (majority for other observations as well)

nox - The nitrogen oxides concentration is particularly high for both observations.

rm - The average number of rooms per dwelling is below the 1st quantile for both observations.

age - The proportion of owner occupied units built prior to 1940 are both observations are the maximum values

dis - The weighted mean of distances to five Boston employment centres is below 1st quantile for both observations.

rad - the index of accessibility to radial highways is max for both observations.

tax - The full-value property tax rate per $10000 is at the 3rd quantile for both observations.

ptratio - The pupil-teacher ratio by town is both at the 3rd quantile for both observations.

black - The black predictor is max for observation 399 and is particularly high for observation 406

lstat - The lower status of the population is particularly high for both observations.

Of course, the medv or the median value of owner-occupied homes in $1000s is the lowest for both observations.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling

```
sum(Boston$rm > 7)
```

```
## [1] 64
```

```
sum(Boston$rm > 8)
```

```
## [1] 13
```

There are 64 suburbs that average more than 7 rooms per dwelling

There are 13 suburbs that average more than 8 rooms per dwelling

```
interVar <- subset(Boston, rm > 8)

summary(interVar)
```

```
##       crim                zn             indus              chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat            medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

```
summary(Boston)
```

```
##       crim                 zn               indus             chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
```

```
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

We see that the suburbs with more than 8 rooms are generally nicer suburbs/towns. The crime rate is lower, the percent of lower status is lower, the nitrogen oxides concentration is lower, and a low pupil teacher ratio.