STT 481 Homework 1

Derien Weatherspoon

2023-01-24

Contents

Question 3a:

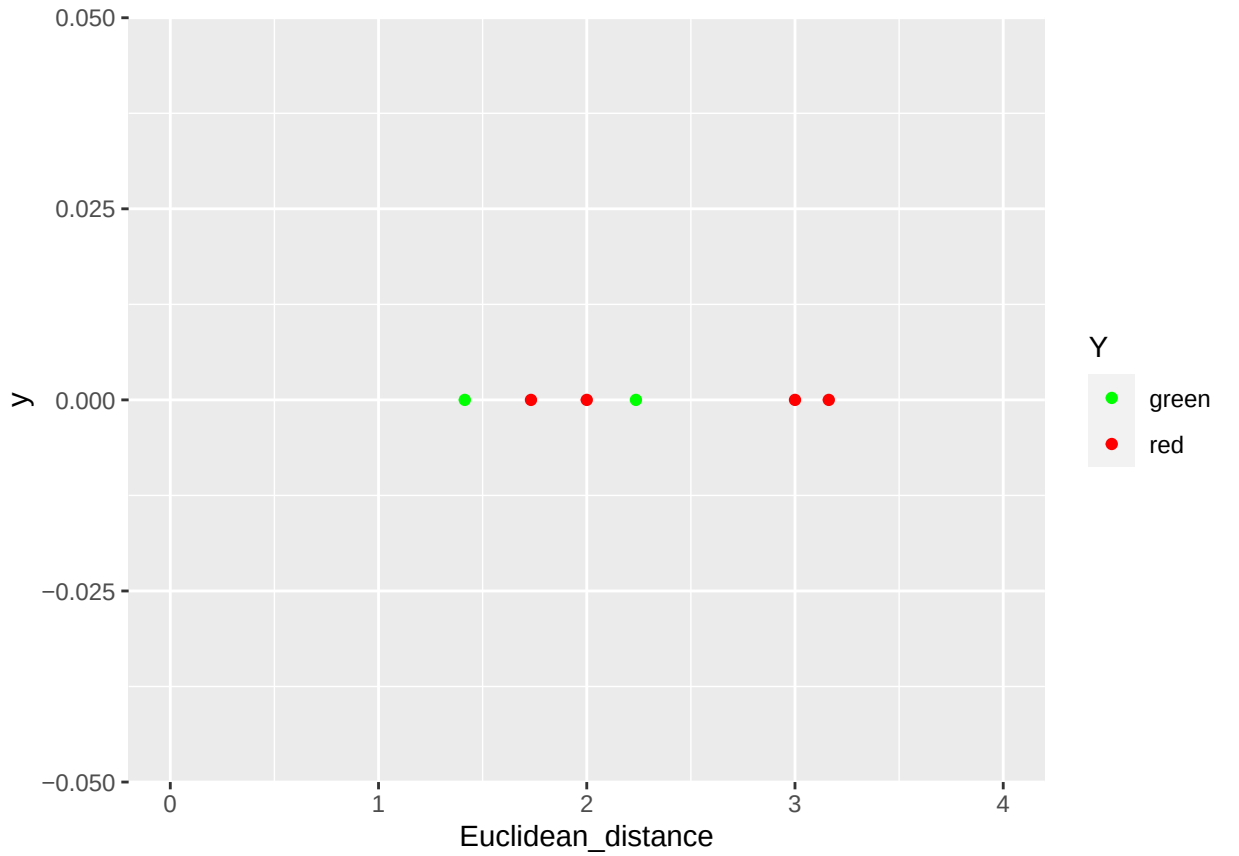
The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when X1 = 1, X2 = 0, X3 = 1 using K-nearest neighbors. Compute the Euclidean distance between each observation and the test point, (X1, X2, X3) = (1, 0, 1).

```
euclidean_df <- data.frame(observation=c(1,2,3,4,5,6),
                           X1 = c(0,2,0,0,1,1),
                           X2 = c(3,0,1,1,0,1),
                           X3 = c(0,0,3,2,1,1),
                           Y = c("red", "red", "red", "green", "green", "red"),
                           Euclidean_distance = c(sqrt(3*3), sqrt(2*2),sqrt(1*1+3*3), sqrt(1*1+2*2),sqrt
euclidean_df
```

```
##   observation X1 X2 X3     Y Euclidean_distance
## 1           1  0  3  0   red           3.000000
## 2           2  2  0  0   red           2.000000
## 3           3  0  1  3   red           3.162278
## 4           4  0  1  2 green           2.236068
## 5           5  1  0  1 green           1.414214
## 6           6  1  1  1   red           1.732051
```

Question 3b: What is our prediction with K=1? Why?

```
ggplot(euclidean_df, aes(x = Euclidean_distance, y = 0))+
  geom_point(aes(colour = Y))+
  scale_color_manual(values = c( "green", "red"))+
  xlim(0,4)
```




```
#The nearest neighbor to the testing point (0, 0, 0) is Obs 5, (-1, 0, 1) with
#euclidean distance of about 1.41. Obs 5 was Green, so we can predict (K = 1)
#that the test point is Green too.
```

Question 3c:

What is our prediction with K = 3? Why? c) The answer here is red, as most of the points on the plot made is red.

Question 3d:

If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

d) A highly nonlinear Bayes boundary suggest that there is lesser of an advantage to generalizing more because of high variance, so it is best for K to be small.

Question 3e:

Use R to find predictions in (b) and (c). USE KNN Slides in NOTES

Question 4a,b:

```
college <- read.csv("College.csv")

rownames(college) <- college[, 1]
#View(college)

college <- college[, -1]
#View(college)
```

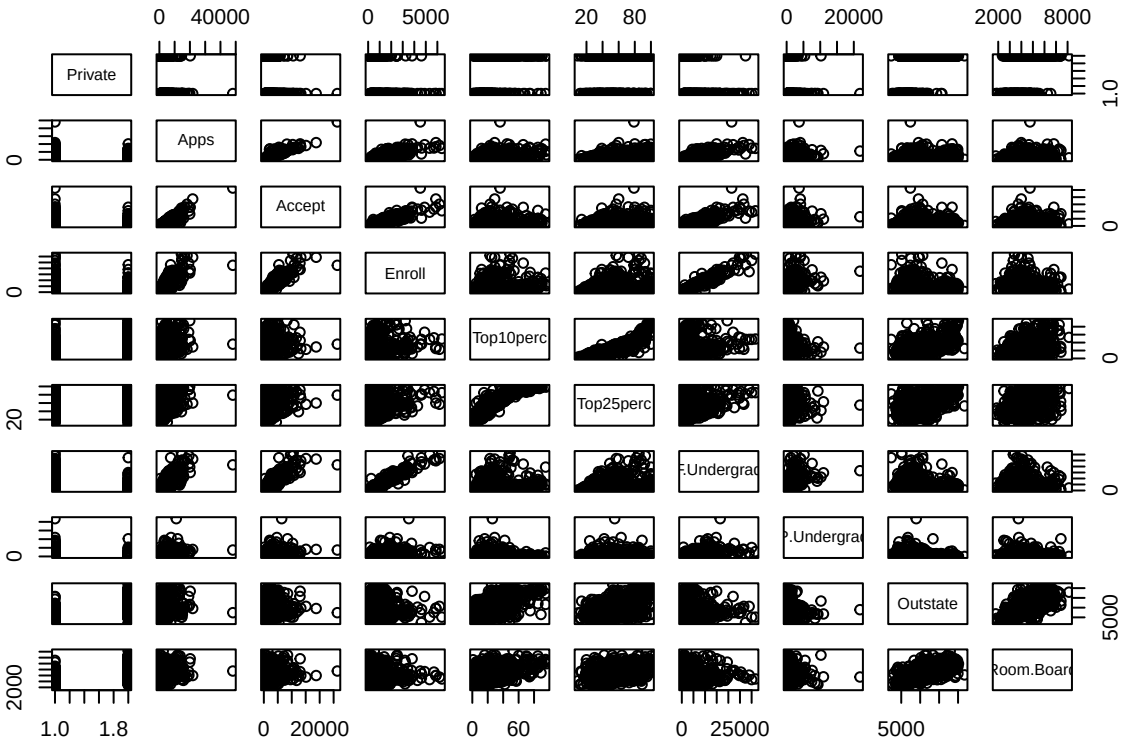
Question 4c (i, ii, and iii):

```
summary(college)
```

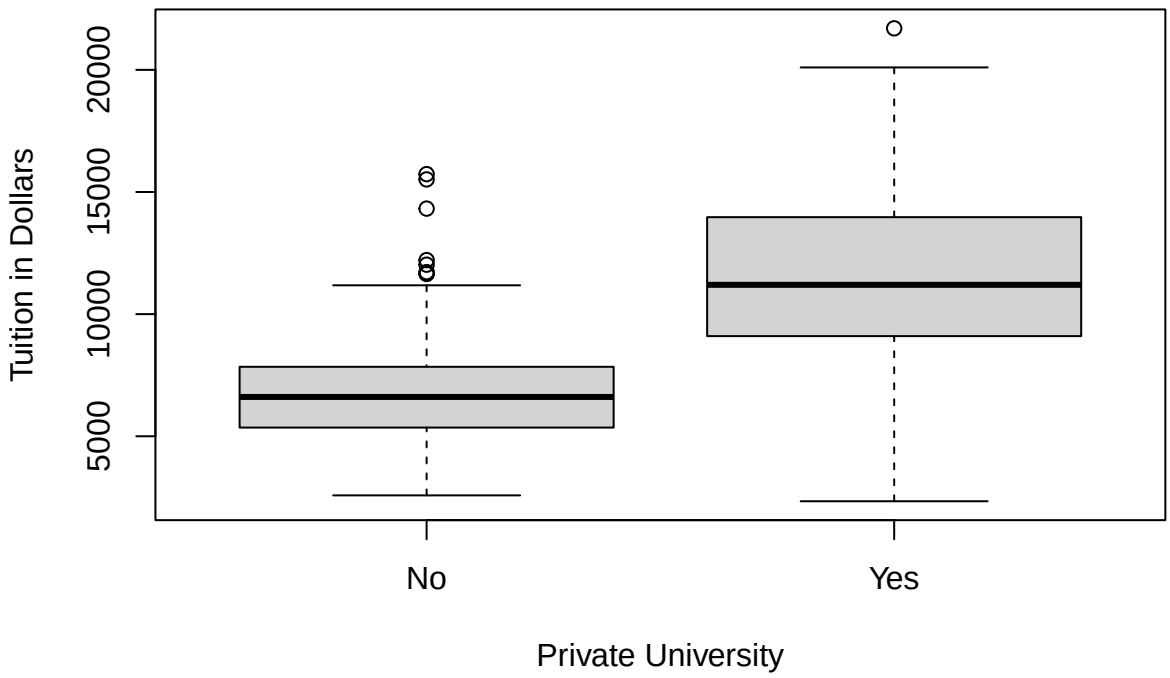
```
##    Private                Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 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
```

```
#Use the pairs() function to produce a scatterplot matrix of the first ten
#columns or variables of the
#data. Recall that you can reference the first ten columns of a matrix A using
#A[,1:10]. We change Private with as.factor, else the pairs function won't work
college$Private <- as.factor(college$Private)
pairs(college[,1:10])
```




```
#use the plot() function to produce side-by-side boxplots of Outstate versus Private
plot(college$Private, college$Outstate,
     xlab = "Private University", ylab = "Tuition in Dollars")
```



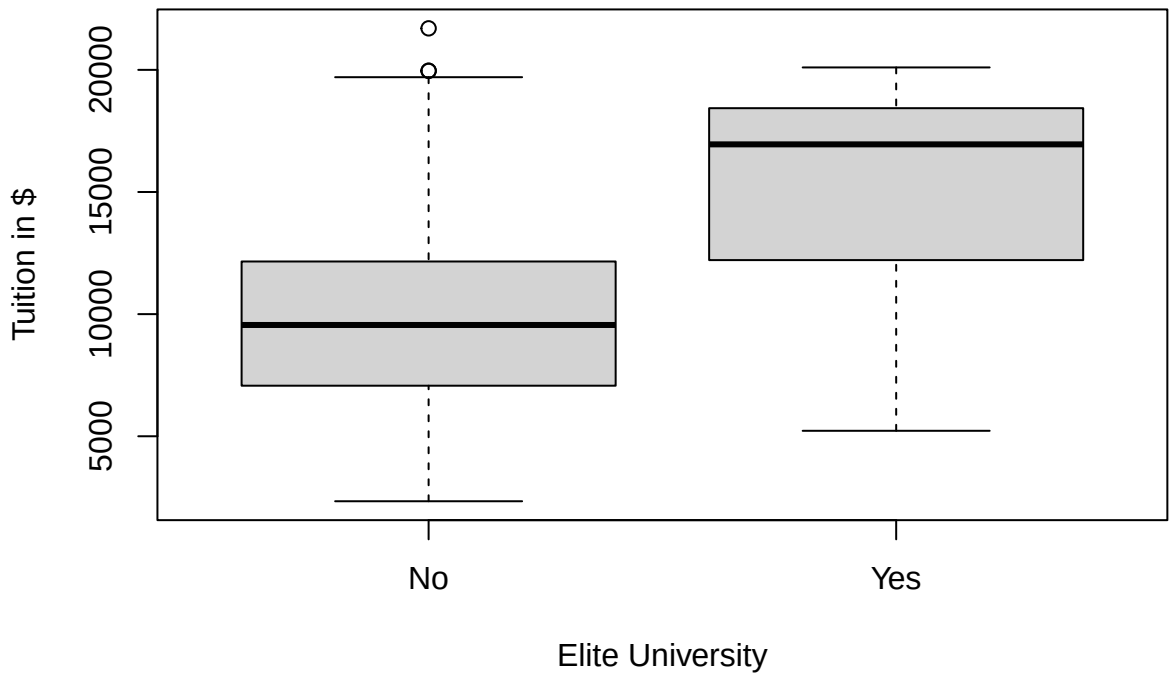
Question 4c (iv):

Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. (Code is given in textbook)

```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
#Use the summary() function to see how many elite universities there are.
#Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.
summary(Elite)
```

```
##  No Yes
## 699  78
```

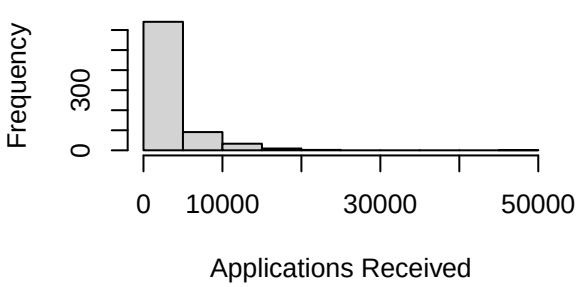
```
plot(college$Elite, college$Outstate,
     xlab = "Elite University", ylab = "Tuition in $")
```

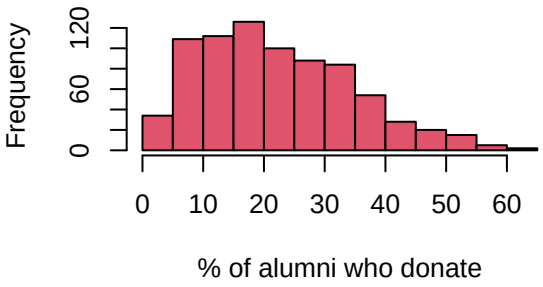


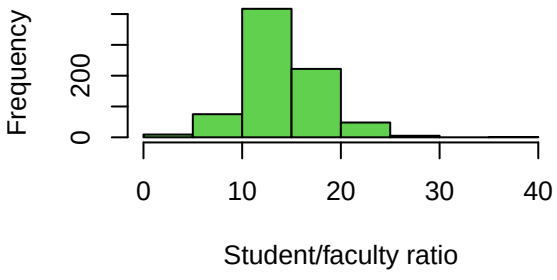
Question 4c (v):

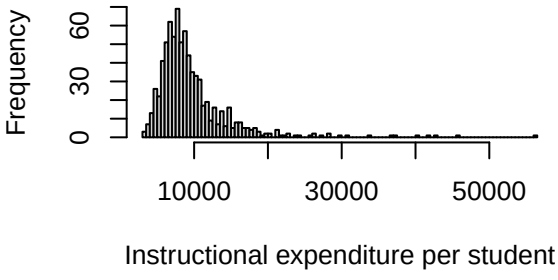
Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
par(mfrow=c(2,2))
hist(college$Apps, xlab = "Applications Received", main = "")
hist(college$perc.alumni, col=2, xlab = "% of alumni who donate", main = "")
hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/faculty ratio", main = "")
hist(college$Expend, breaks=100, xlab = "Instructional expenditure per student", main = "")
```









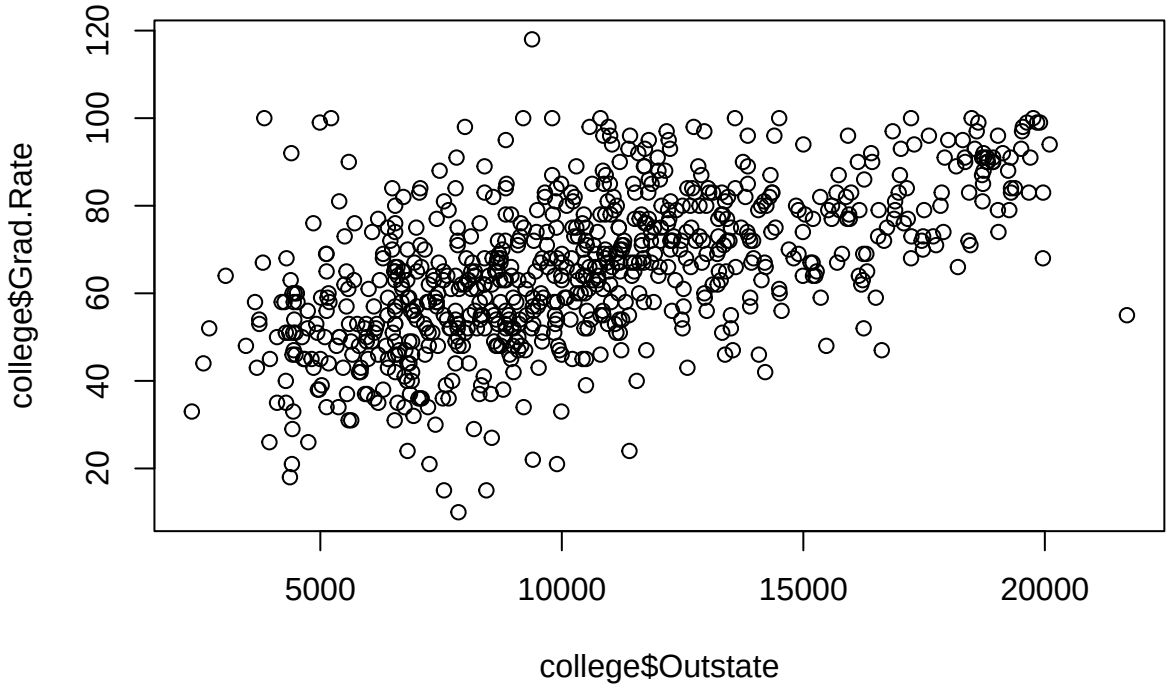
Question 4c (vi):

Continue exploring the data, and provide a brief summary of what you discover.

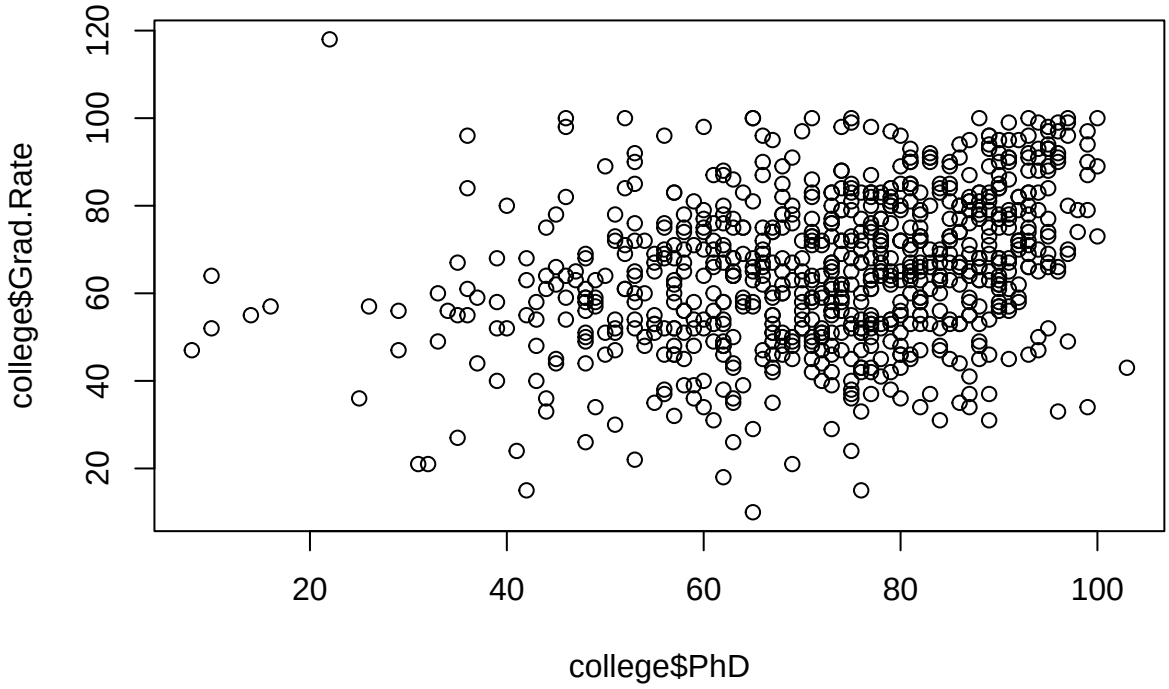
```
#university with the max number of students in top 10% of their class
row.names(college)[which.max(college$Top10perc)]  ##MIT
```

```
## [1] "Massachusetts Institute of Technology"
```

```
#looking at out of state graduation rates and graduation rates with the number of PhD faculty
plot(college$Outstate, college$Grad.Rate)
```



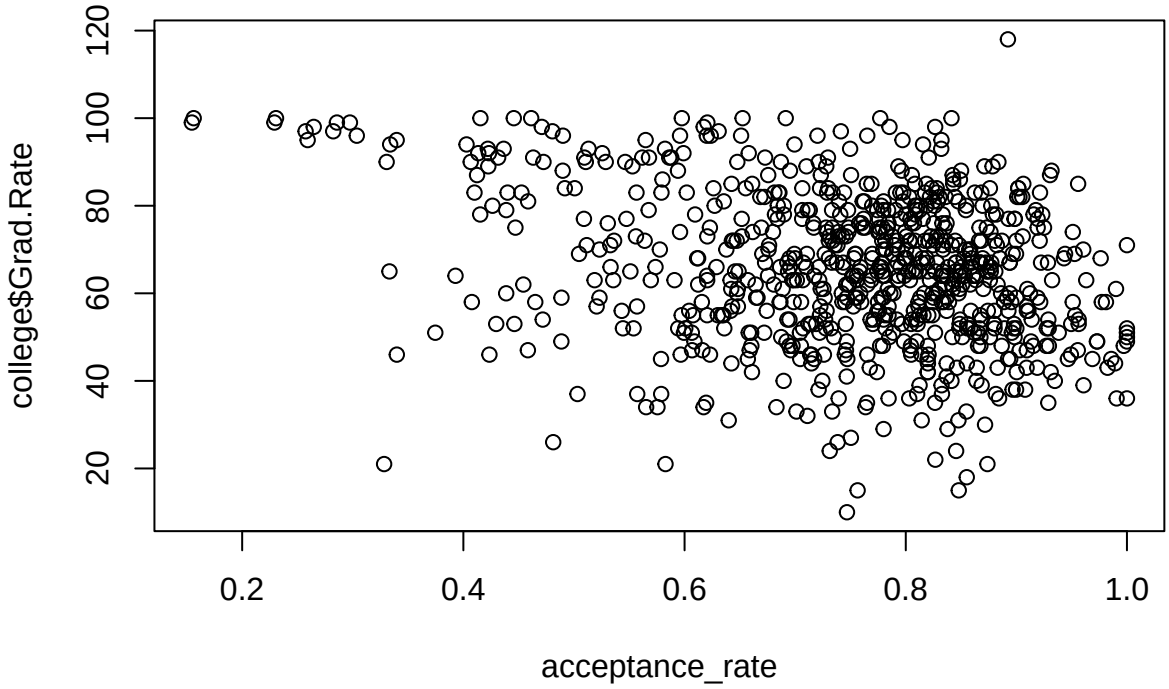

```
plot(college$PhD, college$Grad.Rate)
```




```
#which university has the highest accceptance rate
acceptance_rate <- college$Accept / college$Apps
row.names(college)[which.max(acceptance_rate)] # Emporia State University
```

```
## [1] "Emporia State University"
```

```
#do colleges with high acceptance rate have better graduation rates?
plot(acceptance_rate, college$Grad.Rate)
```



Question 5a:

Which of the predictors are quantitative, and which are qualitative?

```
auto <- na.omit(Auto)
dim(auto)
```

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

```
#summary(auto)
#glimpse(auto)
#the origin row doesn't look right, so after researching the data set I found out
#that the integers correspond to places on Earth. usa = 1, EU = 2, JP = 3
auto$origin.factored <- factor(auto$origin, labels = c("usa", "eu", "jp"))
with(auto,table(origin.factored, origin))
```

```
##                origin
## origin.factored   1   2   3
##             usa 245   0   0
##             eu    0  68   0
##             jp    0   0  79
```

```
#So the qualitative variables here are origin, origin.factored, and name.
#Everything else is quantitative.
```

Question 5b:

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

```
#First we must group together qualitative predictors
qualitative_columns <- which(names(auto)
      %in% c("name", "origin", "origin.factored"))
qualitative_columns
```

```
## [1]  8  9 10
```

```
#use range on columns NOT qualitative
sapply(auto[,-qualitative_columns],range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
```

Question 5c:

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

```
sapply(auto[,-qualitative_columns],mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
sapply(auto[,-qualitative_columns],sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

Question 5d:

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

```
sapply(auto[-seq(10,85), -qualitative_columns], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
##         year
##    77.145570
```

```
sapply(auto[-seq(10,85), -qualitative_columns], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

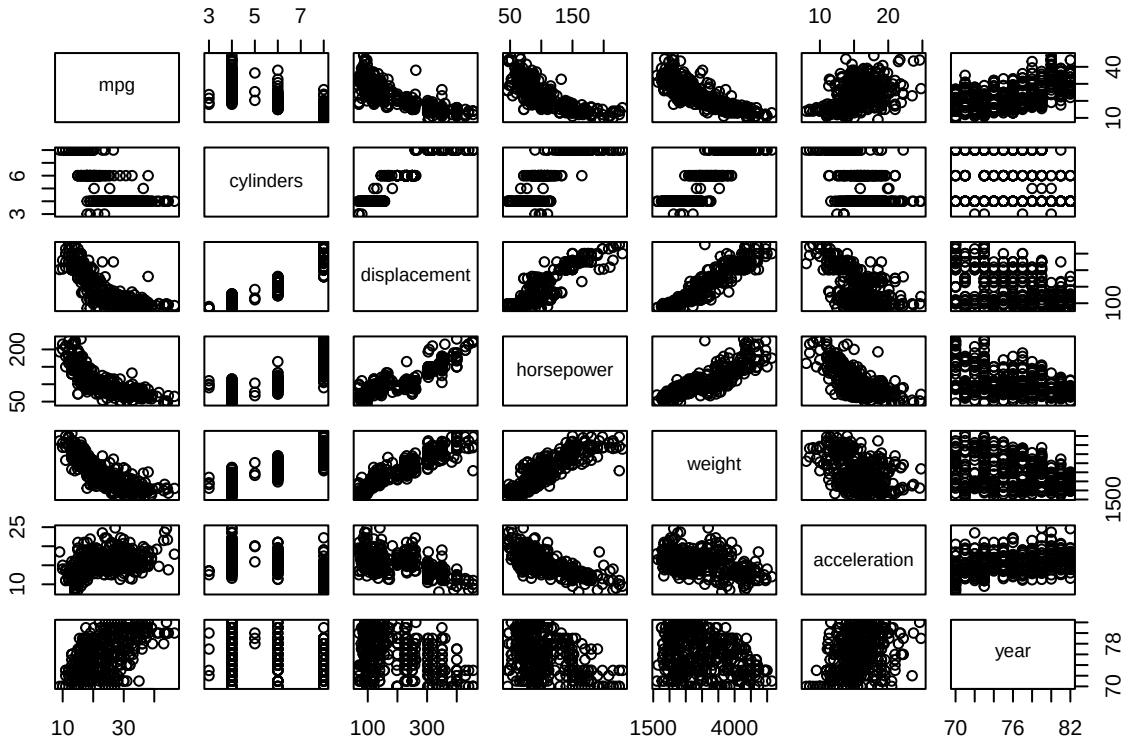
```
sapply(auto[-seq(10,85), -qualitative_columns], range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
```

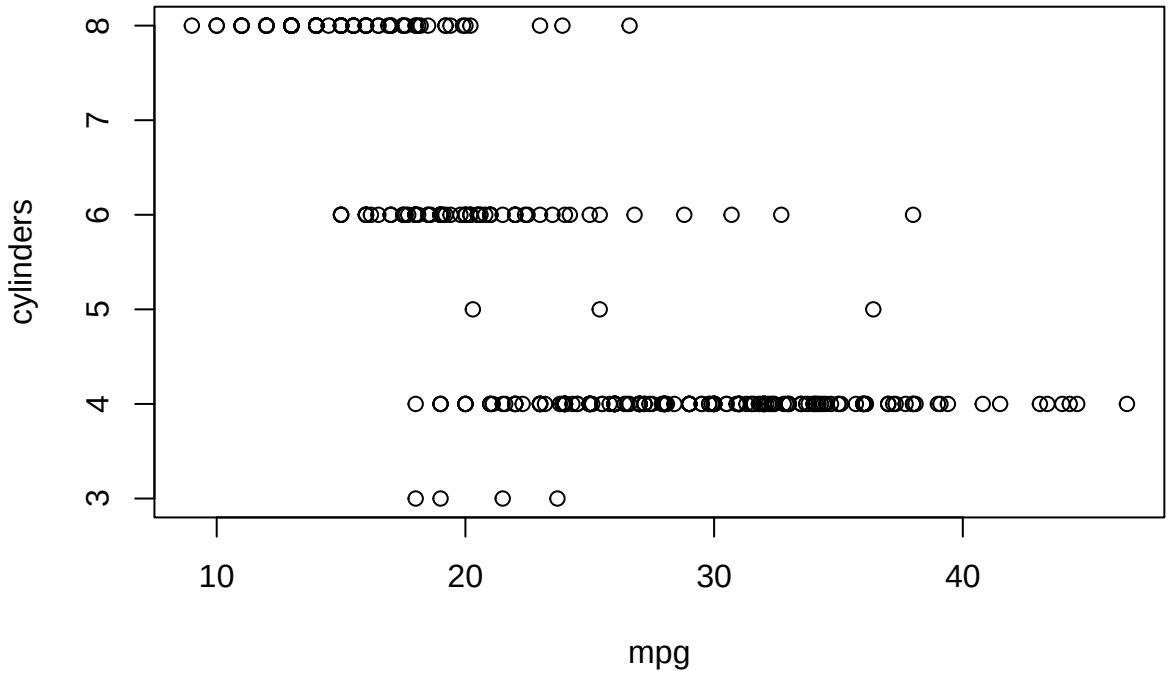
Question 5e:

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.

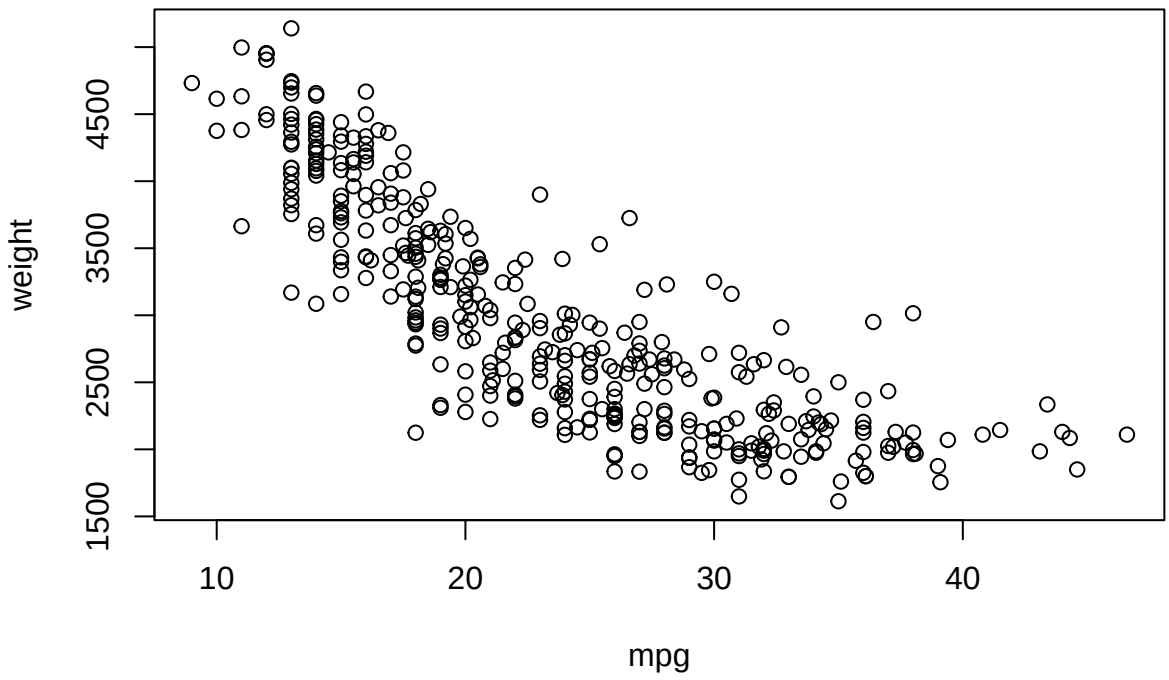
```
pairs(auto[, -qualitative_columns]) #need to remove qualitative columns so the scatter plot
```



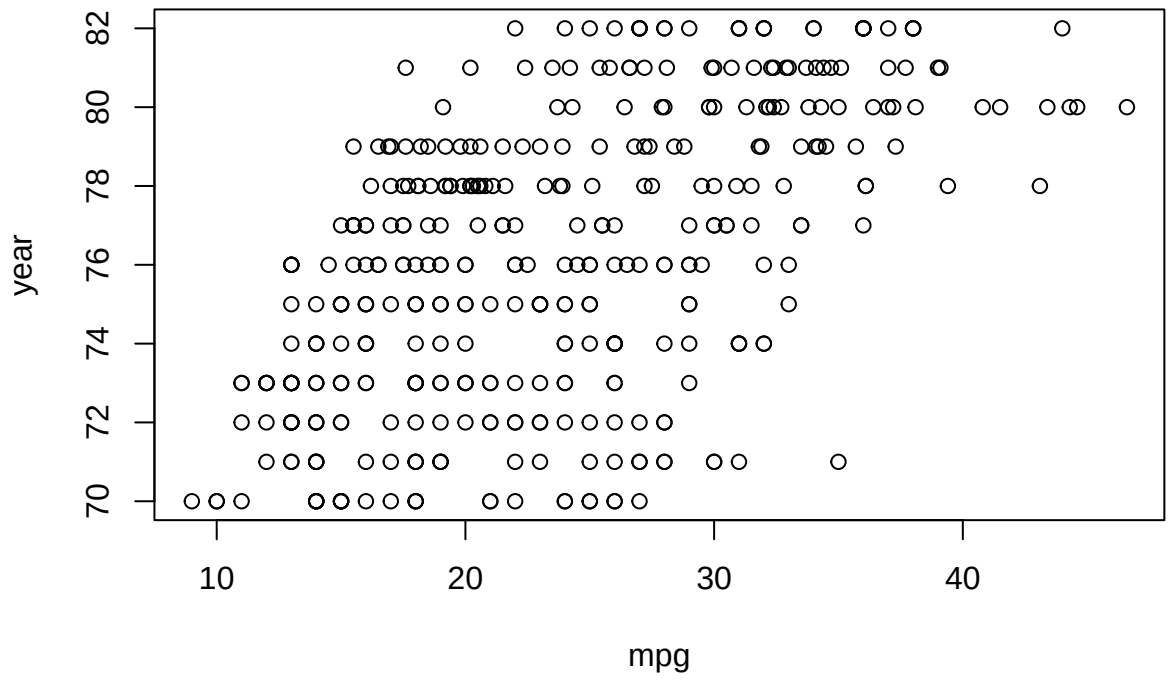
```
#matrix doesn't look erroneous

#looking at the matrix enlarged, it looks like there are some interesting relations
#between mpg ~ cylinders, mpg ~ weight, mpg ~ year
with(auto, plot(mpg, cylinders)) #the more cylinders a car has, the less mpg it gets
```



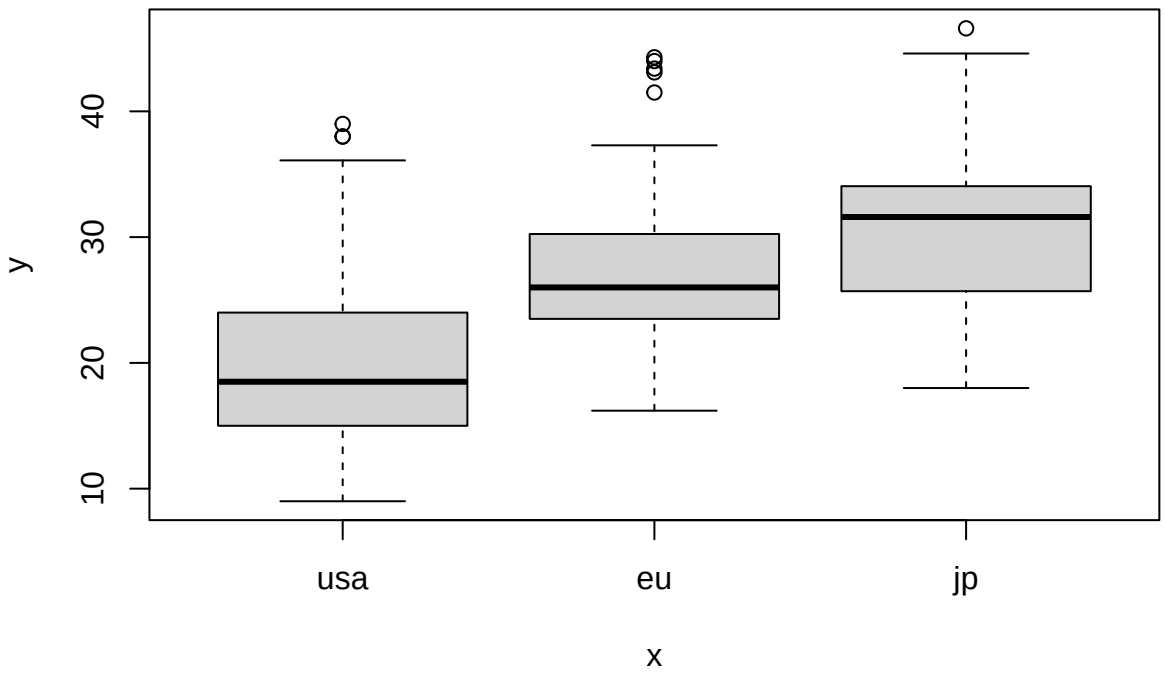

```
with(auto, plot(mpg, weight)) #the heavier a car is, the less mpg it gets, which
```




```
#makes sense intuitively (trucks)
with(auto, plot(mpg, year)) #cars start to gain efficiency over time
```




```
with(auto, plot(origin.factored,mpg),ylab = "miles per gallon")
```




```
#boxplot of mpg in different regions on Earth
```

Question 5f:

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.

f) Using the plots we did in part e, it seems that all predictors are somewhat correlated with mpg in some way. The predictor "name" has few observations, so even if we wanted to use this, it would not be wise. This would cause the data to be overfitted.

Question 6a:

Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get B^0 = 50,B^1 = 20,B^2 = 0.07,B^3 = 35,B^4 = 0.01,B^5 = -10.

a) The correct option here is (iii). If IQ and GPA are fixed values, then males make more on average than females, if it is provided that the GPA is high enough.

This due to the least squared line being yhat = 50 + 20GPA + 0.07IQ + 35$Gender$ + 0.01GPA × IQ - 10*GPA × Gender. For males and females respectively, the equations become:

yhat = 50 + 20*GPA + 0.07*IQ + 0.01*GPA × IQ (males) yhat = 85 + 10*GPA +0.07*IQ + 0.01*GPA × IQ (females)

Question 6b:

Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

b) For this question, we can simply plug in the given values into the equations seen above. yhat = 85 + 40 + 7.7 + 4.4 = 137.1. This gives us a starting salary of 137,100 dollars.

Question 6c:

True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

c) The answer here is FALSE. This is false, as we would need test the hypothesis of B^4 = 0. Find the p-value associated with t or f stat. Furthermore, this depends on the standard error of the beta.

Question 7a:

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response.I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. Y = Beta0 +Beta1X +Beta2X2 +Beta3X3 + epsilon.

Suppose that the true relationship between X and Y is linear, i.e. Y = Beta0 + Beta1X + epsilon. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

a) The trained RSS will be lower than the (untrained) linear regression. The trained RSS would conform to the data better, i.e., it would make a tighter fit with th data with a wider error (epsilon).

Question 7b:

Answer (a) using test rather than training RSS.

b) Unlike in part a, the test RSS will be higher than the linear regression. The trained RSS will have more error than the linear regression, due to being overfitted.

Question 7c:

Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer

c) The trained RSS will be lower for either cubic regression and polynomial regression. The flexible model follow points more closely and reduce the trained RSS.

Question 7d:

Answer (c) using test rather than training RSS.

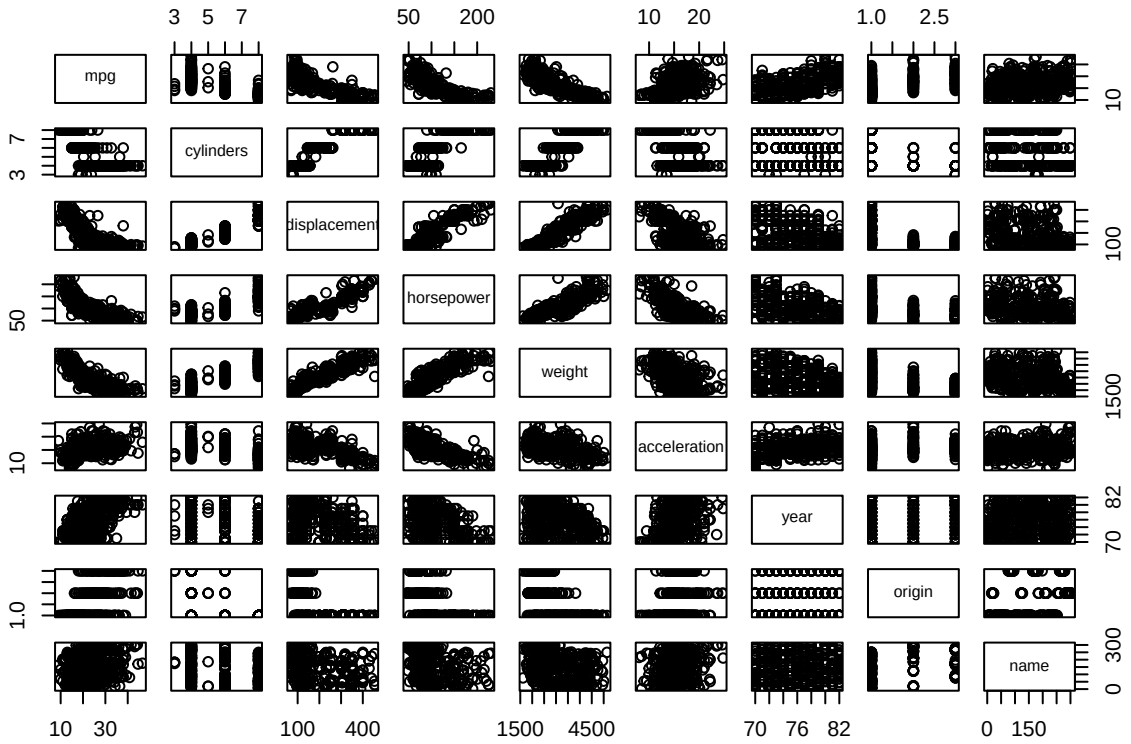
d) We do not know the underlying distribution here. We cannot even say that it is far from linear, because we do not know. We can say that if it is far from linear, then a flexible model will do better here. And if it is close to linear, the model may be over-fitting and result in higher test RSS.

Question 8a and 8b:

This question involves the use of multiple linear regression on the Auto data set.

a) Produce a scatter plot matrix which includes all of the variables in the data set.
b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
pairs(Auto) #From ISLR
```




```
cor(subset(Auto, select = -name))
```

```
##                     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
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance: -Is there a relationship between the predictors and the response? -Which predictors appear to have a statistically significant relationship to the response? -What does the coefficient for the `year` variable suggest?

```
lm.fitQ8 <-  lm(mpg ~ . - name, data = Auto)
summary(lm.fitQ8)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i) There is a relationship between predictors and the response. The F-stat is far from 1, as well as having a very small p-value. This means the evidence is against the null hypothesis.

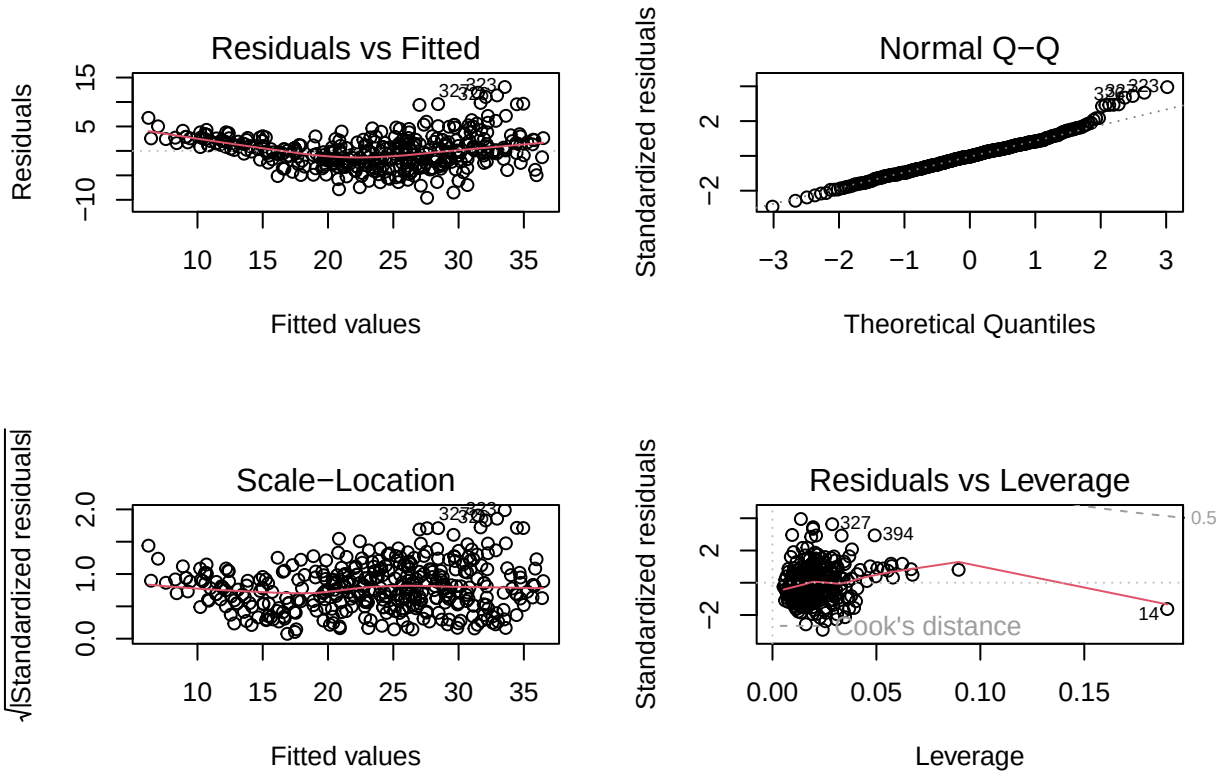
ii) If we look at the p-values that are associated with each predictor's t-stat, then it is simple to tell that displacement, year, origin, and weight are statistically significant. Cylinders, acceleration and horsepower to not have the same significance.

iii) The coefficient for year says that for every 1 year, mpg increases by the coefficient listed. Cars become more fuel efficient by that much each year.

Question 8d:

Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow = c(2, 2))
plot(lm.fitQ8)
```




```
#Based on the residual vs fitted values, there is some non linearity in the data.
#Based on the residual vs leverage plot, there are some outliers present.
#There is also high leverage for obs greater than 5% (0.05)
```

Question 8e

Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
lm.fitQ8_2 <-  lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto)
summary(lm.fitQ8_2)
```

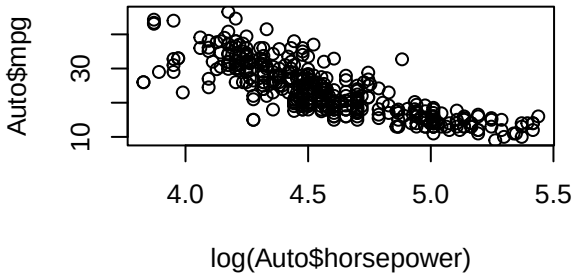
```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders                   7.606e-01  7.669e-01   0.992    0.322
## displacement               -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                     -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement     -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight         2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

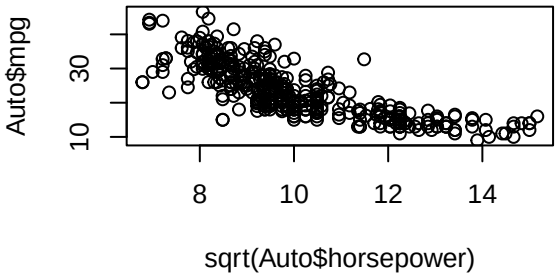
```
#Displacement and weight has significant interaction. Cylinders and displacement
#not have a significant interaction.
```

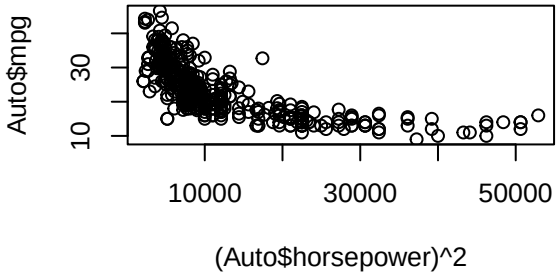
Question 8f:

Try a few different transformations of the variables, such as log(X), sqrt(X), X2. Comment on your findings.

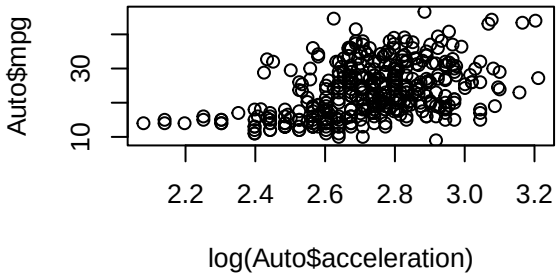
```
par(mfrow = c(2, 2))
plot(log(Auto$horsepower), Auto$mpg)
plot(sqrt(Auto$horsepower), Auto$mpg)
plot((Auto$horsepower)^2, Auto$mpg)
par(mfrow = c(2, 2))
```

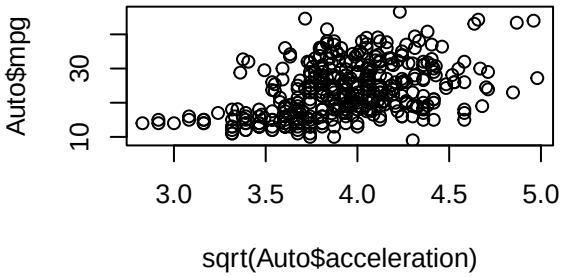


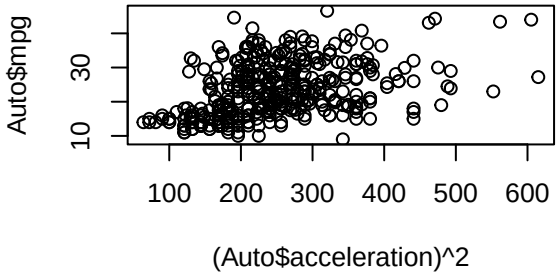




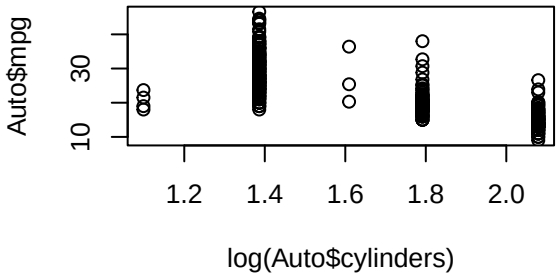

```
plot(log(Auto$acceleration), Auto$mpg)
plot(sqrt(Auto$acceleration), Auto$mpg)
plot((Auto$acceleration)^2, Auto$mpg)
par(mfrow = c(2, 2))
```

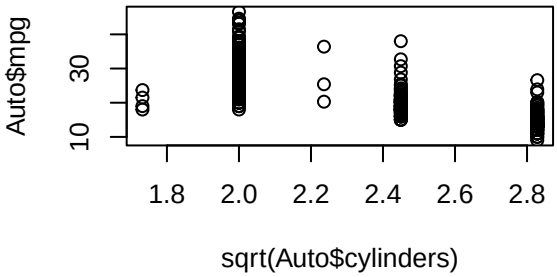


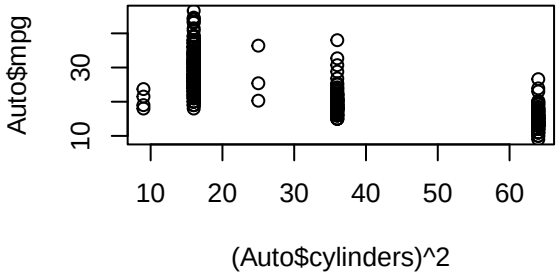





```
plot(log(Auto$cylinders), Auto$mpg)
plot(sqrt(Auto$cylinders), Auto$mpg)
plot((Auto$cylinders)^2, Auto$mpg)
#Can't use weight, non-numeric

#Horsepower, acceleration, cylinders do not show a relationship with mpg.
#Horsepower is the closest to being linear though
#Residual plots indicate a few outliers.
```







Question 9a:

Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
attach(Carseats)
lm.fitQ9 <- lm(Sales ~ Price + Urban + US)
summary(lm.fitQ9)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Question 9b:

Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

b) From the summary above: The effect of increase of 1 dollar to 'price' equates to a decrease in around 54 dollars in 'sales', if all everything else in the model remains fixed. There is no relationship between 'sales' and 'urban', this is seen by looking at the p-value. Lastly, 'US' shows that sales, on average in the US stores are around 1200 dollars more than a store not based in the US, again assuming all other predictors are fixed.

Question 9c:

Write out the model in equation form, being careful to handle the qualitative variables properly.

c) Sales = 13.043469 - 0.054459(Price) - 0.021916(UrbanYes) + 1.200573(USYes) Important to note that UrbanYes = 1 for Urban, and 0 means not Urban. Same goes for USYes, 1 = US, 0 = non US stores.

Question 9d:

For which of the predictors can you reject the null hypothesis H0 :Betaj =0?

d) Price and US. Both variables have an effect on Sales.

Question 9e:

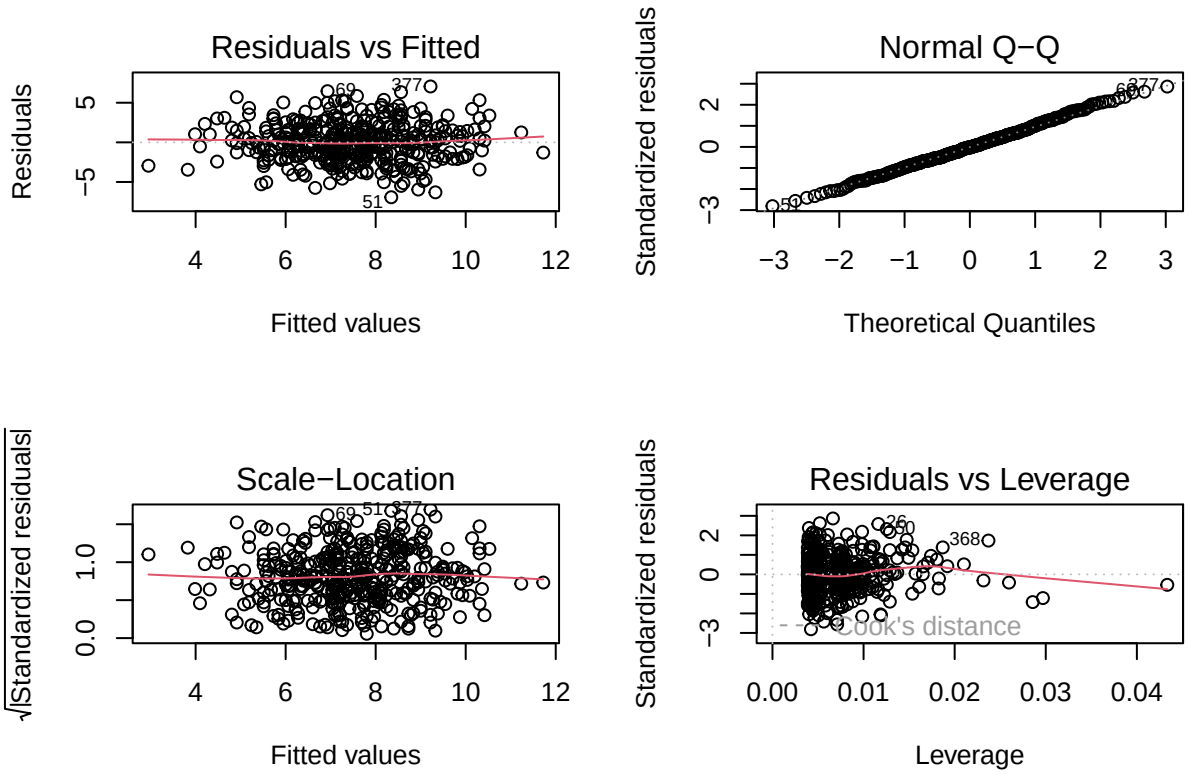
On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
lm.fitQ9_2 <- lm(Sales~Price + US)
summary(lm.fitQ9_2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
```

```
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm.fitQ9_2)
```



Question 9f: How well do the models in (a) and (e) fit the data?

f) Both models have very similar R^2 values. So, around 24% of the variability can be explained by the model. This means, they are not very good fits.

Question 9g:

Using the model from (e), obtain 95% confidence interval for the coefficients.

```
confint(lm.fitQ9_2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

Question 9h:

Is there evidence of outliers or high leverage observations in the model from (e)?

h) Looking at the residual vs fitted plot above from (e), there seems to be some linearity in the data. Looking at the residual vs leverage plot from (e), there are also some outliers. This is seen by looking at the data points outside the range of -2 and 2. Also, there is high leverage for observations more than 10%.

Question 10a:

Perform the following commands in R: The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100) / 10
y <- 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

The linear model is: Y = 2 + 2X_1 + 0.3X_2 + epsilon epsilon ~ N(0,1) r.v. The regression coefficients are 2, 2 and 0.3 respectively.

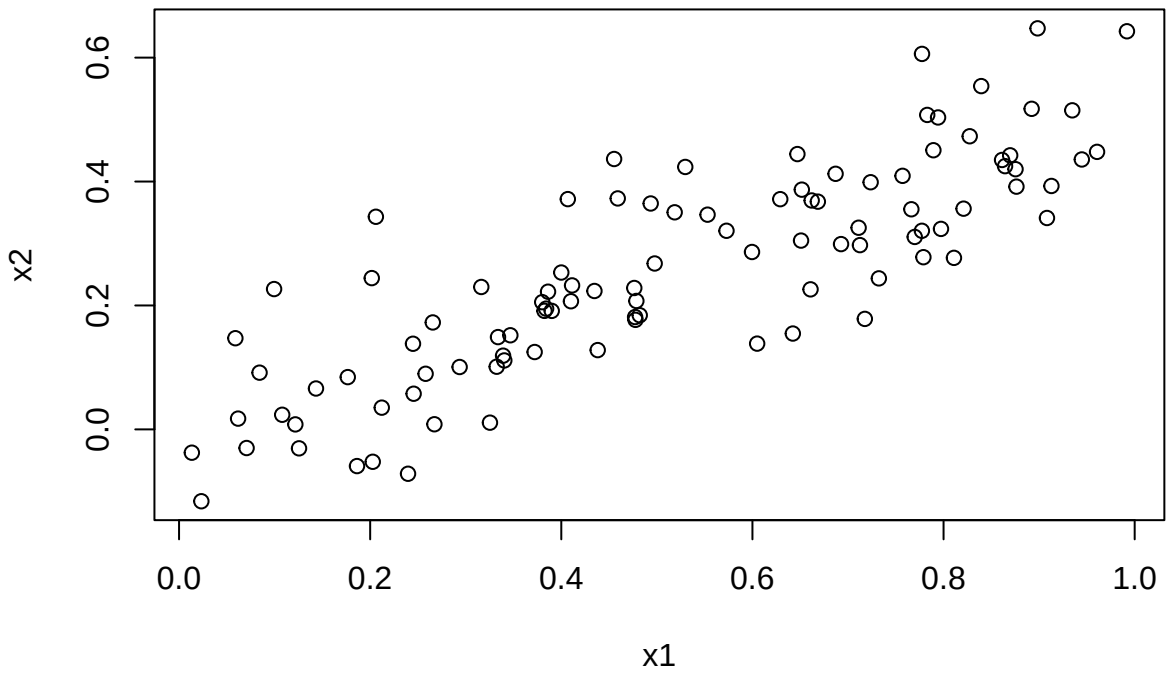
Question 10b:

What is the correlation between "x1" and "x2" ? Create a scatterplot displaying the relationship between the variables.

```
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
plot(x1,x2)
```




```
#Strong positive correlation.
```

Question 10c:

Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are Beta^0, Beta^1, and Beta^2? How do these relate to the true Beta0, Beta1, and Beta2? Can you reject the null hypothesis H0 : Beta1 = 0? How about the null hypothesis H0 : Beta2 = 0?

```
lm.fitQ10 <- lm(y ~ x1 + x2)
summary(lm.fitQ10)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

Beta^0, Beta^1, and Beta^2 respectively have the coefficients: 2.1304996, 1.43955554, and 1.0096742. Beta^0 is the only one close to Beta0. The p-value is less than 0.05, we reject the null hypothesis for Beta1, we cannot do this for Beta2, the p-value is greater than 0.05.

Question 10d:

Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 :Beta1 = 0?

```
lm.fitQ10_2 <- lm(y~x1)
summary(lm.fitQ10_2)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

The coefficient x1 in this model is highly significant. Its p-value is very low, in this case, we can reject the null hypothesis.

Question 10e:

Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 :Beta1 = 0?

```
lm.fitQ10_3 <- lm(y~x2)
summary(lm.fitQ10_3)
```

```
##
```

```
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

Same thing as in part d, the coefficient x2 is very different from before. This is highly significant, as the p-value is very low. Again, reject the null.

Question 10f:

Do the results obtained in (c)-(e) contradict each other? Explain your answer.

f) The results do NOT contradict each other. Predictors x1 and x2 are highly correlated, this means collinearity. It is hard to determine which predictor is separately associated with the response variable. Collinearity reduces accuracy in the regression coefficients as well.

Question 10g:

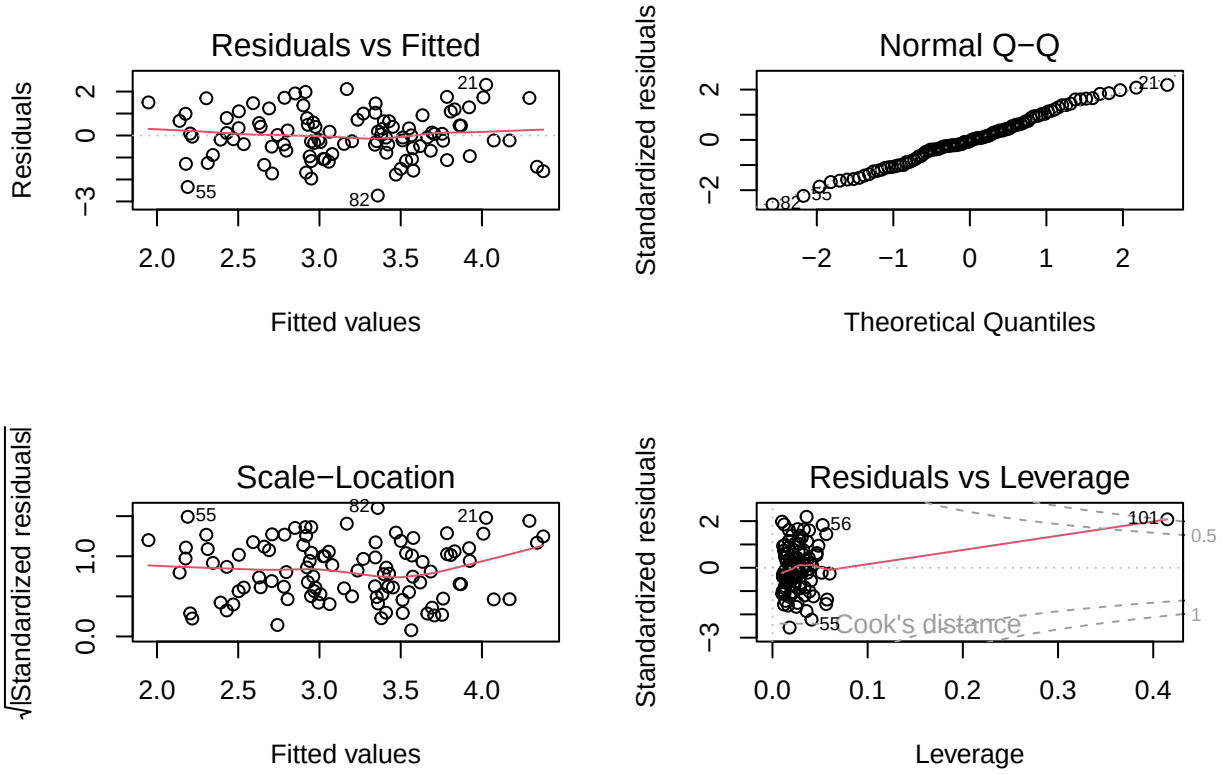
Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
lm.fitQ10_4 <- lm(y ~ x1 + x2)
lm.fitQ10_5 <- lm(y ~ x1)
lm.fitQ10_6 <- lm(y ~ x2)
summary(lm.fitQ10_4)
```

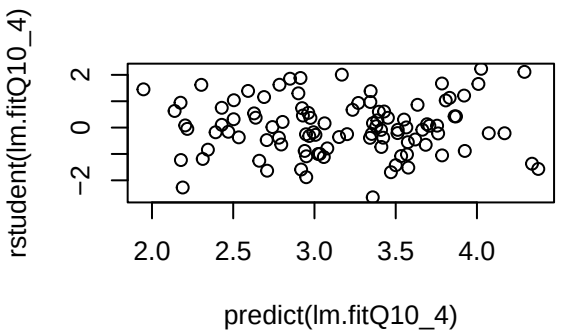
```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
par(mfrow=c(2,2))
plot(lm.fitQ10_4)
```



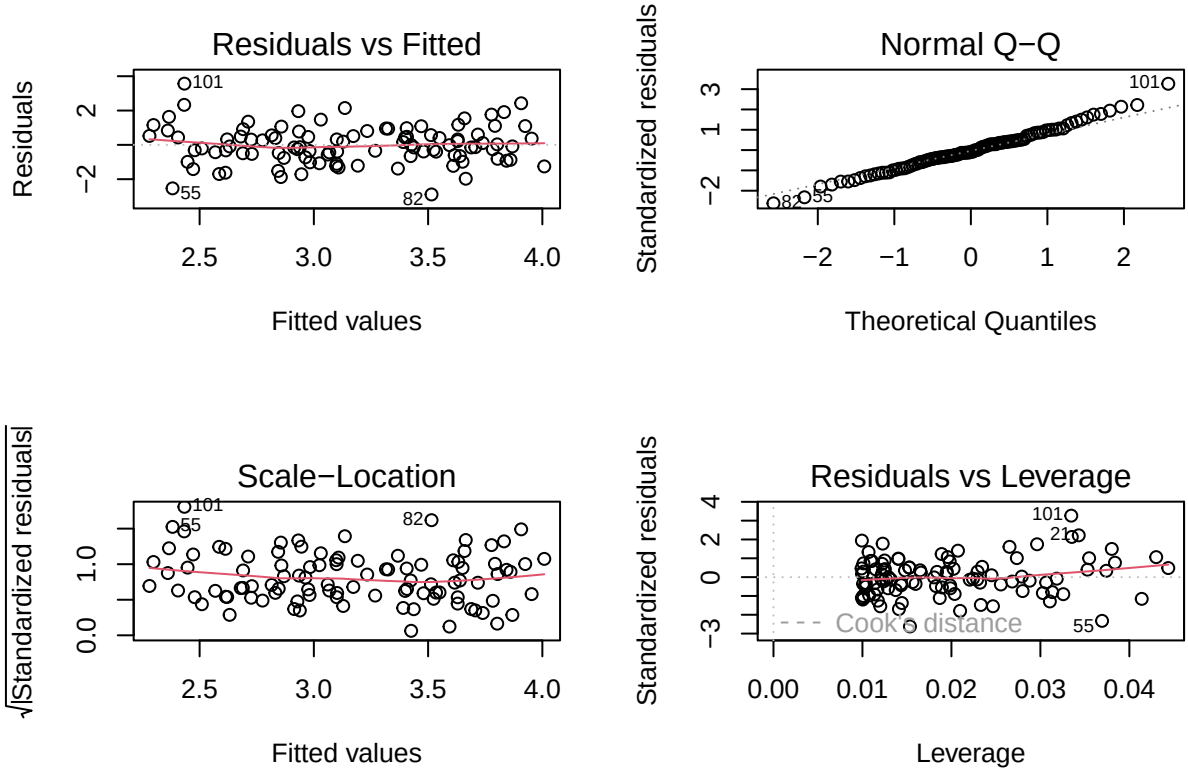

```
plot(predict(lm.fitQ10_4), rstudent(lm.fitQ10_4))
```



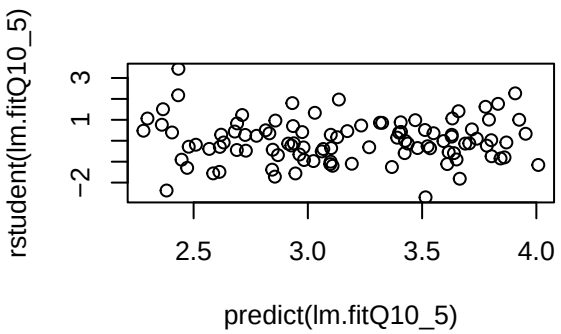

```
summary(lm.fitQ10_5)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```
par(mfrow=c(2,2))
plot(lm.fitQ10_5)
```



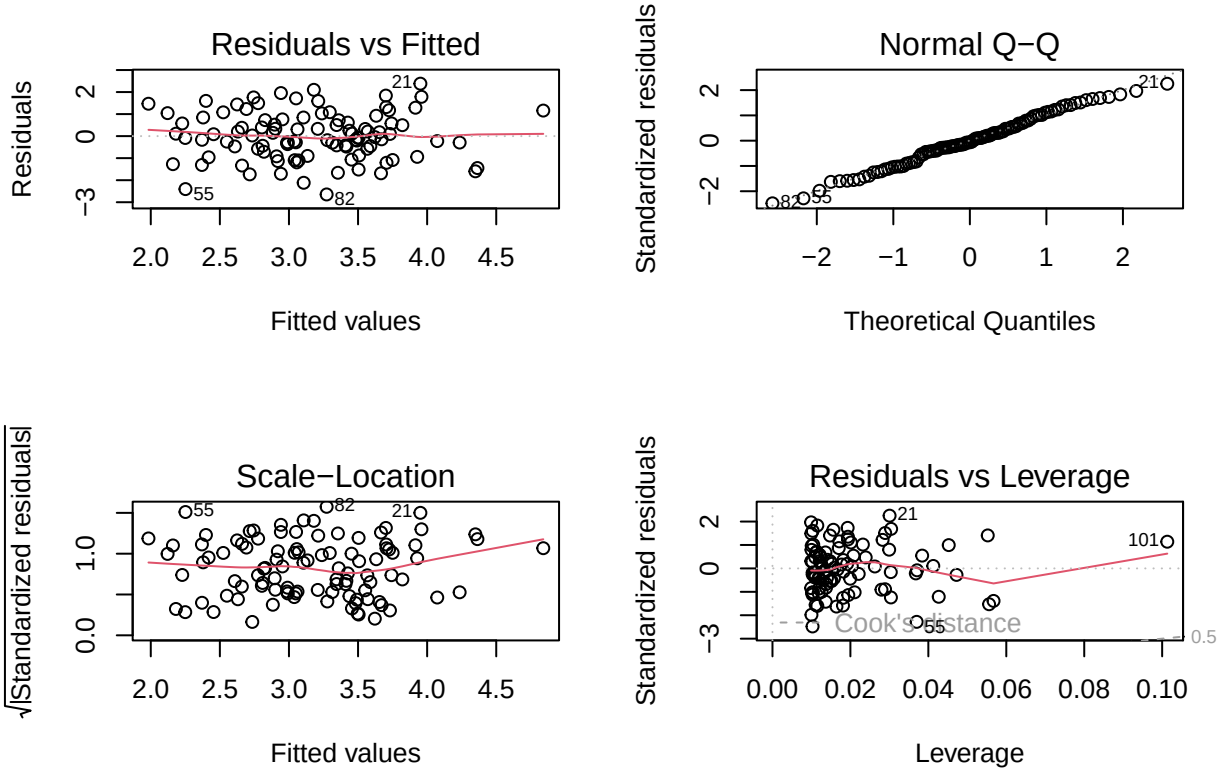

```
plot(predict(lm.fitQ10_5), rstudent(lm.fitQ10_5))
```



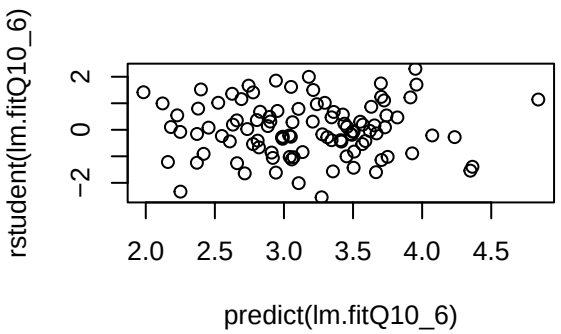

```
summary(lm.fitQ10_6)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```
par(mfrow=c(2,2))
plot(lm.fitQ10_6)
```




```
plot(predict(lm.fitQ10_6), rstudent(lm.fitQ10_6))
```



In the first model, x1 is not significant, in the second model it is. x2 is significant in both models.

In the first model, both x1 and x2, the 3rd model, x2, the last point is high leverage point. Though, in the second model, x1, the last point is not high leverage.

In the first model, x1 and x2, and third model, x2, the last point is not an outlier. In the second model, x1, the last point is an outlier, the point is outside abs(3).

Question 11a:

Use KNN method with K = 10 to predict the prices.

```
#FRESH START ON 11
library(FNN)
library(class)
```

```
##
## Attaching package: 'class'
```

```
## The following objects are masked from 'package:FNN':
##
##     knn, knn.cv
```

```
toyota <- read.csv("toyota.csv")
toyota_index <- data.frame(model = c("RAV4", "Camry"), year = c(2019, 2019),
                     transmission = c("Automatic", "Automatic"),
```

```
                    mileage = c(12345, 50000), fuelType = c("Diesel", "Hybrid"),
                    tax = c(150, 130), mpg = c(25, 50), engineSize = c(2,3))


test_toyota <- rbind(toyota[,-3], toyota_index)
model.x <- model.matrix(~., data = test_toyota)[,-1]
model.x <- scale(model.x)

x.train <- model.x[1:nrow(toyota),]
x.test <- model.x[-(1:nrow(toyota)),]
prediction.out <- knn.reg(train = x.train, test = x.test, y = toyota$price, k = 10)
prediction.out$pred
```

```
## [1] 20893.9 26902.1
```

Question 11b:

Use KNN method with K = 100 to predict the prices.

```
prediction.out_2 <- knn.reg(train = x.train, test = x.test, y = toyota$price, k = 100)
prediction.out_2$pred
```

```
## [1] 25197.33 16842.34
```

Question 11c:

According to bias-variance trade-off, which one is going to give higher prediction bias and lower pre- diction variance? Why?

c) K = 100 has a higher prediction bias because it is a flexible model, which pulls from more of the data. K = 10 will have a lower prediction variance because it has less information from the data.

Question 11d:

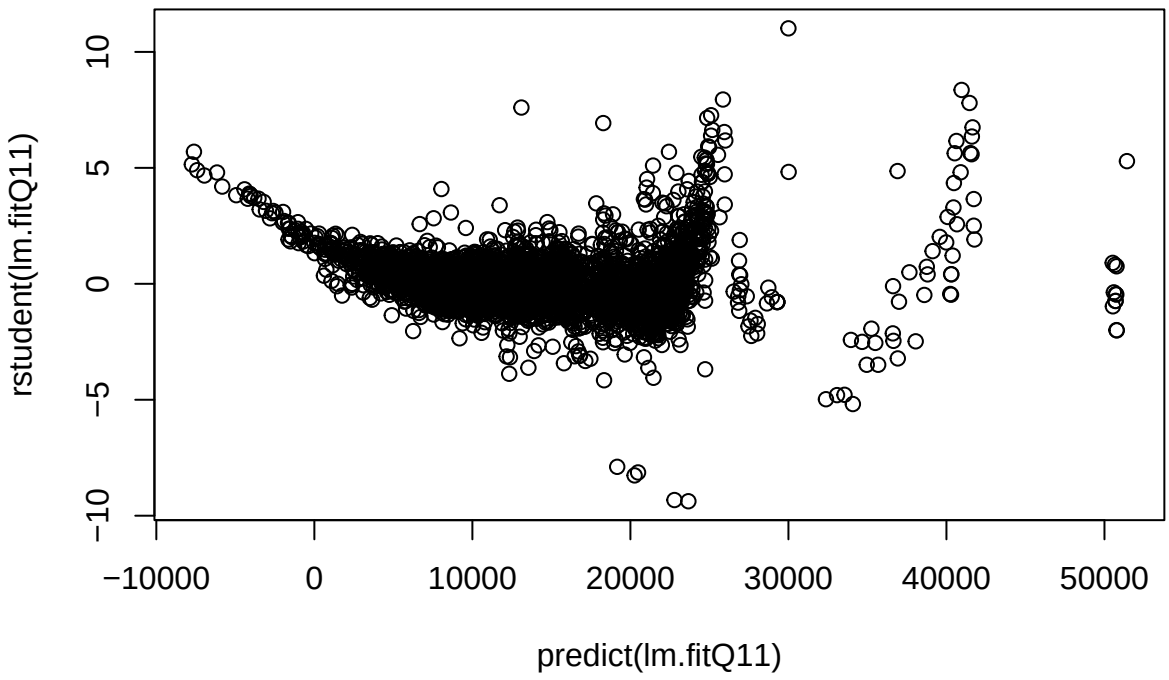
Use a linear regression model to predict the prices.

```
lm.fitQ11 <- lm(data = toyota, price ~.)
summary(lm.fitQ11)
```

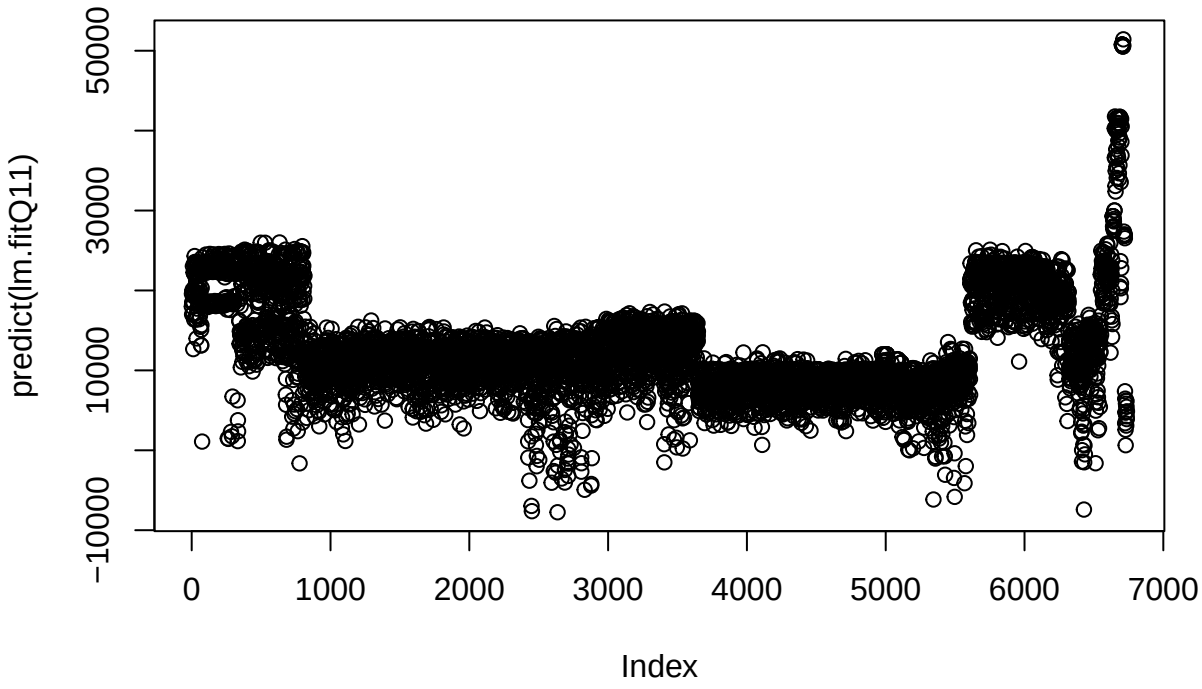
```
##
## Call:
## lm(formula = price ~ ., data = toyota)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15669.1   -826.3   -169.0    602.7  17990.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           -1.566e+06  2.933e+04 -53.411  < 2e-16 ***
## modelAvensis           1.324e+03  1.960e+02   6.755 1.55e-11 ***
## modelAygo             -2.694e+03  1.228e+02 -21.942  < 2e-16 ***
## modelC-HR              5.559e+03  1.052e+02  52.850  < 2e-16 ***
## modelCamry             6.847e+03  5.349e+02  12.800  < 2e-16 ***
## modelCorolla           5.001e+03  1.288e+02  38.831  < 2e-16 ***
## modelGT86              6.185e+03  2.342e+02  26.411  < 2e-16 ***
## modelHilux             8.352e+03  2.934e+02  28.465  < 2e-16 ***
## modelIQ               -3.093e+02  6.144e+02  -0.503    0.615
## modelLandCruiser       2.275e+04  3.780e+02  60.201  < 2e-16 ***
## modelPrius             4.987e+03  1.387e+02  35.950  < 2e-16 ***
## modelPROACEVERSO       1.358e+04  4.768e+02  28.486  < 2e-16 ***
## modelRAV4              4.908e+03  1.577e+02  31.117  < 2e-16 ***
## modelSupra             3.037e+04  5.663e+02  53.620  < 2e-16 ***
## modelUrbanCruiser     -6.856e+00  8.585e+02  -0.008    0.994
## modelVerso             1.204e+03  1.862e+02   6.466 1.08e-10 ***
## modelVerso-S           2.536e+02  9.892e+02   0.256    0.798
## modelYaris            -1.557e+03  8.358e+01 -18.631  < 2e-16 ***
## year                   7.811e+02  1.455e+01  53.696  < 2e-16 ***
## transmissionManual    -1.221e+03  8.463e+01 -14.428  < 2e-16 ***
## transmissionOther      8.927e+02  1.707e+03   0.523    0.601
## transmissionSemi-Auto  7.322e+01  1.382e+02   0.530    0.596
## mileage               -6.233e-02  1.691e-03 -36.851  < 2e-16 ***
## fuelTypeHybrid         3.240e+03  1.590e+02  20.378  < 2e-16 ***
## fuelTypeOther          2.941e+03  2.253e+02  13.052  < 2e-16 ***
## fuelTypePetrol         1.587e+03  1.295e+02  12.261  < 2e-16 ***
## tax                   -3.780e+00  3.637e-01 -10.391  < 2e-16 ***
## mpg                   -9.591e+00  2.081e+00  -4.609 4.12e-06 ***
## engineSize             2.997e+03  1.698e+02  17.648  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1706 on 6709 degrees of freedom
## Multiple R-squared:  0.928,  Adjusted R-squared:  0.9277
## F-statistic:  3089 on 28 and 6709 DF,  p-value: < 2.2e-16
```

```
plot(predict(lm.fitQ11),rstudent(lm.fitQ11))
```




```
#or
plot(predict(lm.fitQ11))
```



Question 11e:

Pick one of the coefficients in the model (d) to interpret. Which predictors are significantly important?

e) The coefficient I wish to you use is mileage. Mileage is highly significant for the data. It is telling us that as we see an increase in price, we see a decrease in mileage on the car by around 6.233. Other significant predictors include fuelType, tax, mpg, engineSize, year, and a few more.

Question 11f:

Can you tell which method preforms better in terms of prediction accuracy? KNN or linear regression model?

f) Linear regression model performs better in terms of prediction accuracy. This is because the KNN model performs good when you have few predictors, and linear model here has a R^2 value of 92%, which is very high.

References

https://www.statology.org/euclidean-distance-in-r/

https://search.r-project.org/CRAN/refmans/ISLR2/html/Auto.html

https://book.stat420.org/transformations.html

https://online.stat.psu.edu/stat462/node/260/

https://www.statology.org/confint-r/

https://stackoverflow.com/questions/25166624/insert-picture-table-in-r-markdown

https://rdrr.io/cran/Rfit/man/rstudent.rfit.html

https://www.statology.org/studentized-residuals-in-r/

https://www.rdocumentation.org/packages/KODAMA/versions/0.0.1/topics/knn.predict

https://www.edureka.co/blog/knn-algorithm-in-r/#Practical%20Implementation%20Of%20KNN%20Algorithm%20In%20R