Verzani Problem Set

Next are considered the problems from Verzani's book on page 31

Problem 4.1
A student evaluation of a teacher is on a 1-5 Leichert scale. Suppose the answers to the first 3 questions are given in this table

Student	Question 1	Question 2	Question 3
1	3	5	1
2	3	2	3
3	3	5	1
4	4	5	1
5	3	2	1
6	4	2	3
7	3	5	1
8	4	5	1
9	3	4	1
10	4	2	1

Enter in the data for questions 1, 2 and 3

```
> q1 <- factor(c(3, 3, 3, 4, 3, 4, 3, 4, 3, 4), levels = c(1, 2, 3, 4, 5))
> q2 <- factor(c(5, 2, 5, 5, 2, 2, 5, 5, 4, 2), levels = c(1, 2, 3, 4, 5))
> q3 <- factor(c(1, 3, 1, 1, 1, 3, 1, 1, 1, 1), levels = c(1, 2, 3, 4, 5))
> eval <- data.frame(q1, q2, q3)
```

1. Make a table of the results of questions 1, 2 and 3 separately

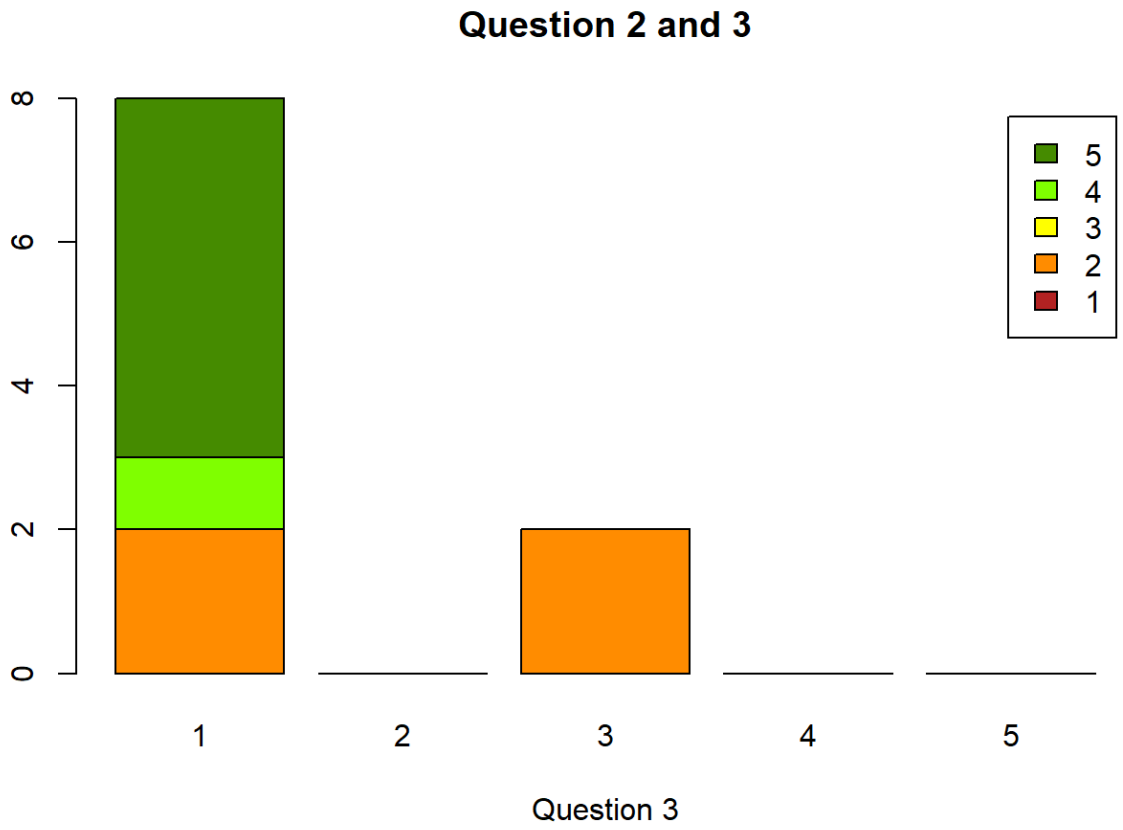
```
> table(q1)
q1
1 2 3 4 5
0 0 6 4 0
> table(q2)
q2
1 2 3 4 5
0 4 0 1 5
> table(q3)
q3
1 2 3 4 5
8 0 2 0 0
```

2. Make a contingency table of questions 1 and 2

```
> table(q1, q2)
   q2
q1  1 2 3 4 5
  1 0 0 0 0 0
  2 0 0 0 0 0
  3 0 2 0 1 3
  4 0 2 0 0 2
  5 0 0 0 0 0
```

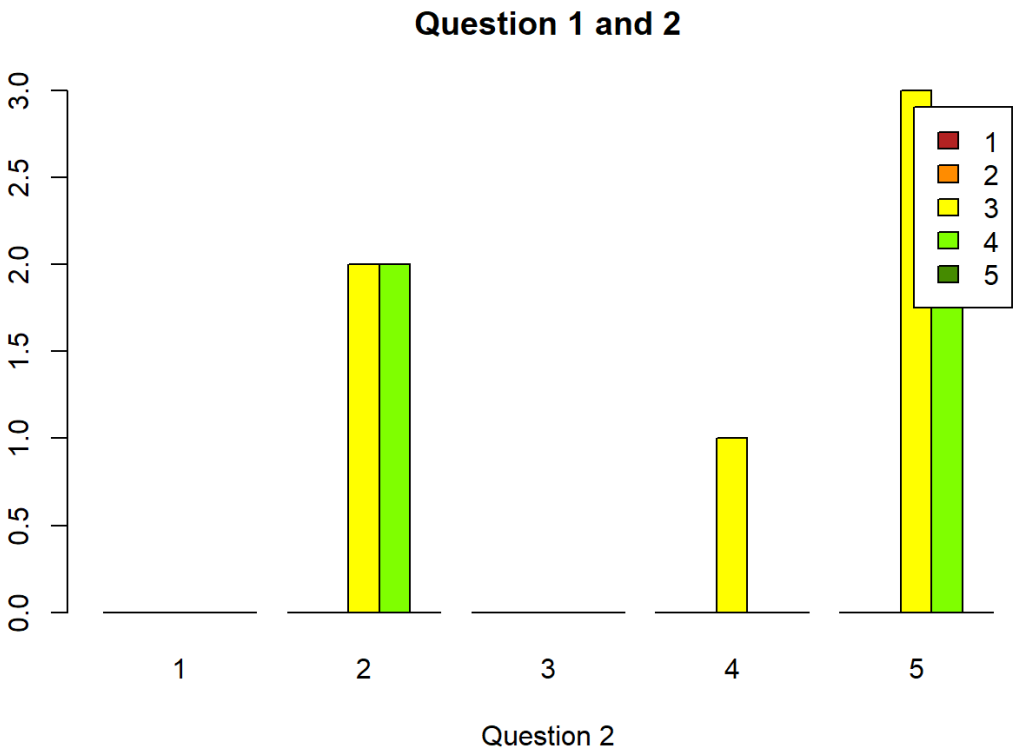
3. Make a stacked barplot of questions 2 and 3

```
> barplot(table(q2, q3),
+         main = "Question 2 and 3",
+         xlab = "Question 3",
+         col = c("firebrick", "darkorange", "yellow", "chartreuse", "chartreuse4"),
+         legend.text = c(1, 2, 3, 4, 5))
```

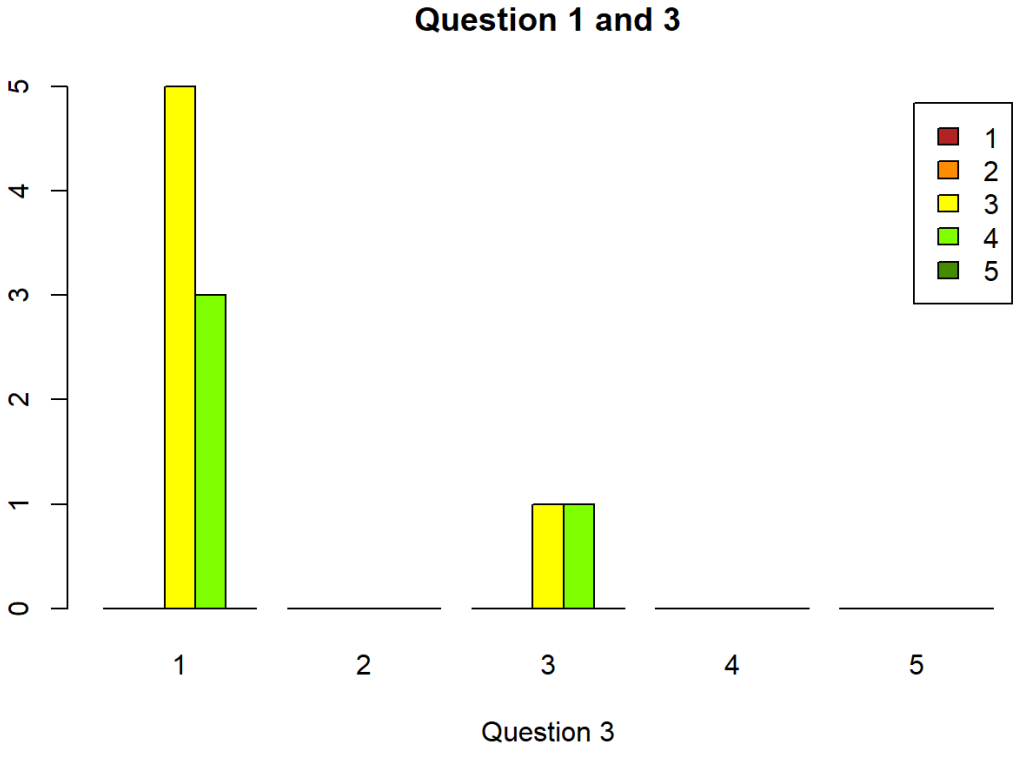


4. Make a side-by-side barplot of all 3 question

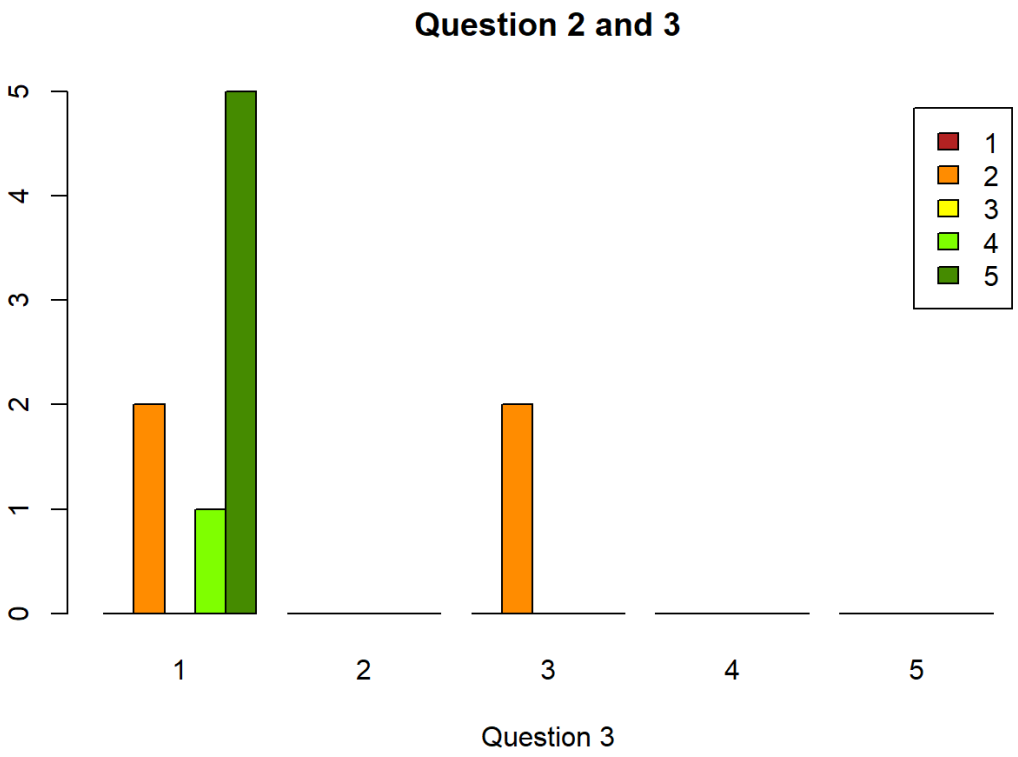
```
> barplot(table(q1, q2),
+         beside = TRUE,
+         main = "Question 1 and 2",
+         xlab = "Question 2",
+         col = c("firebrick", "darkorange", "yellow", "chartreuse", "chartreuse4"),
+         legend.text = c(1, 2, 3, 4, 5))
```




```
> barplot(table(q1, q3),
+         beside = TRUE,
+         main = "Question 1 and 3",
+         xlab = "Question 3",
+         col = c("firebrick", "darkorange", "yellow", "chartreuse", "chartreuse4"),
+         legend.text = c(1, 2, 3, 4, 5))
```



```
> barplot(table(q2, q3),
+         beside = TRUE,
+         main = "Question 2 and 3",
+         xlab = "Question 3",
+         col = c("firebrick", "darkorange", "yellow", "chartreuse", "chartreuse4"),
+         legend.text = c(1, 2, 3, 4, 5))
```



Problem 4.2

In the library MASS there is a data frame UScereal which contains information about popular breakfast cereals. Attach the data set as follows

```
> library(MASS)
> attach(UScereal)
> names(UScereal)
 [1] "mfr"       "calories"  "protein"   "fat"       "sodium"    "fibre"
 [7] "carbo"     "sugars"    "shelf"     "potassium" "vitamins"
```

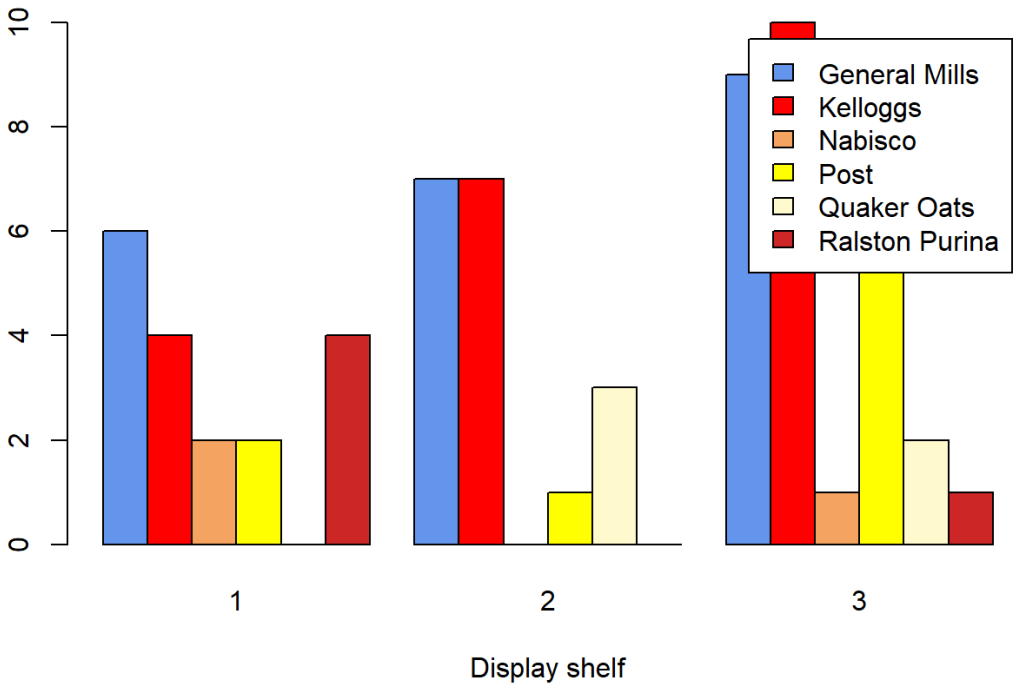
Now, investigate the following relationships, and make comments on what you see. You can use tables, barplots, scatterplots etc. to do your investigation.

1. The relationship between manufacturer and shelf

```
> table(mfr, shelf)
   shelf
mfr  1  2  3
  G  6  7  9
  K  4  7 10
  N  2  0  1
  P  2  1  6
  Q  0  3  2
```

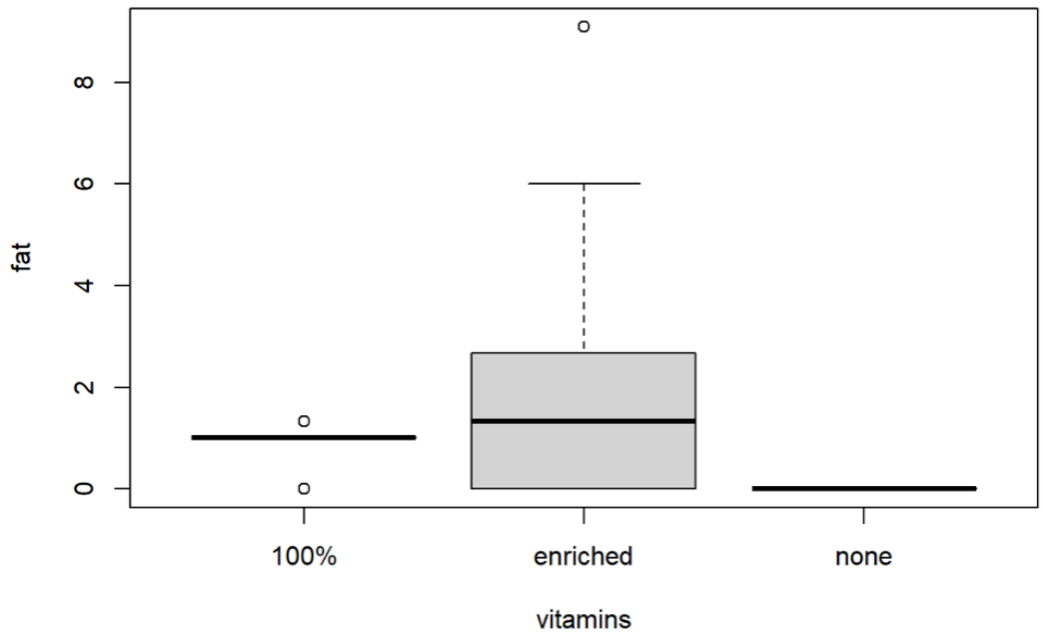
R 4 0 1

```
> barplot(table(mfr, shelf),
+         beside = TRUE,
+              legend.text = c("General Mills", "Kelloggs", "Nabisco", "Post", "Quaker Oats",
"Ralston Purina"),
+                   col = c("cornflowerblue", "red", "sandybrown", "yellow1", "lemonchiffon",
"firebrick3"),
+         xlab = "Display shelf")
```

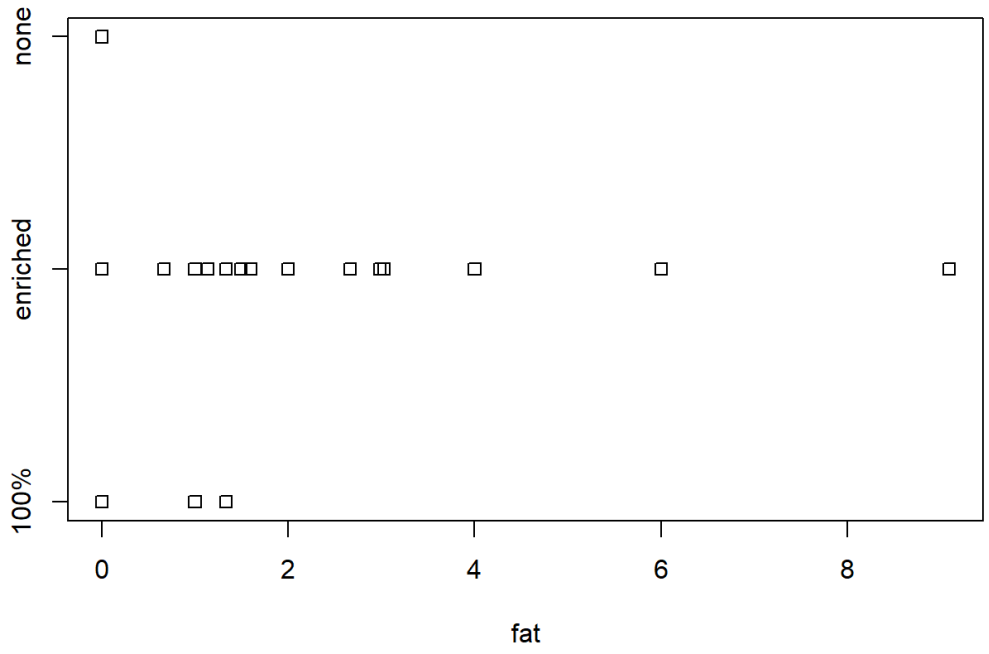


2. The relationship between fat and vitamins

```
> boxplot(fat ~ vitamins)
```

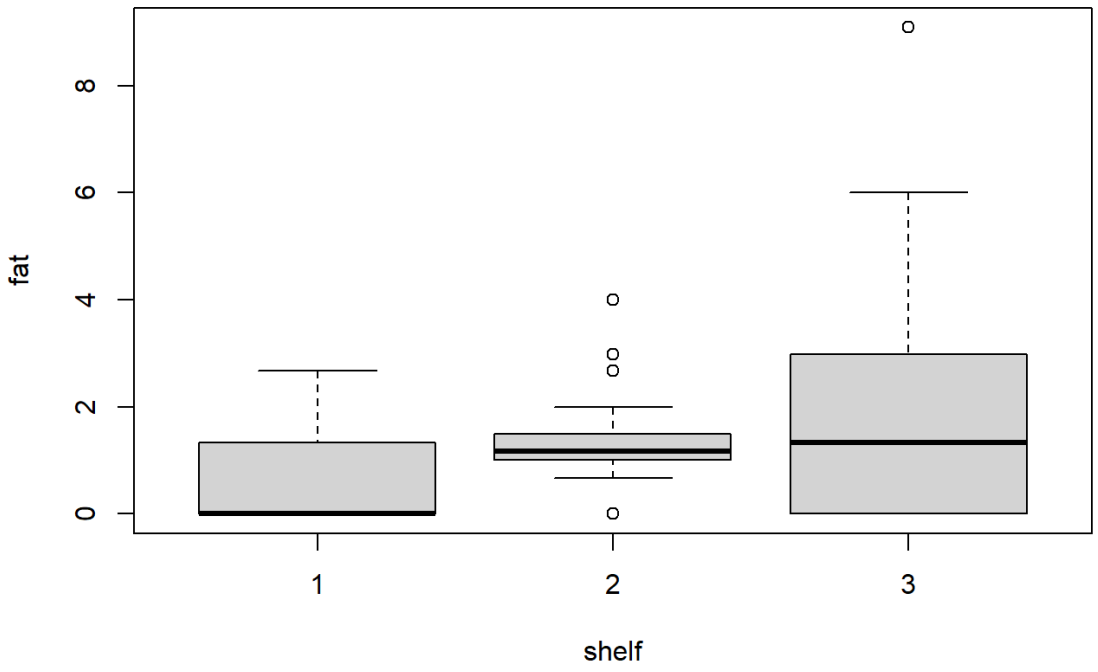


```
> stripchart(fat ~ vitamins)
```



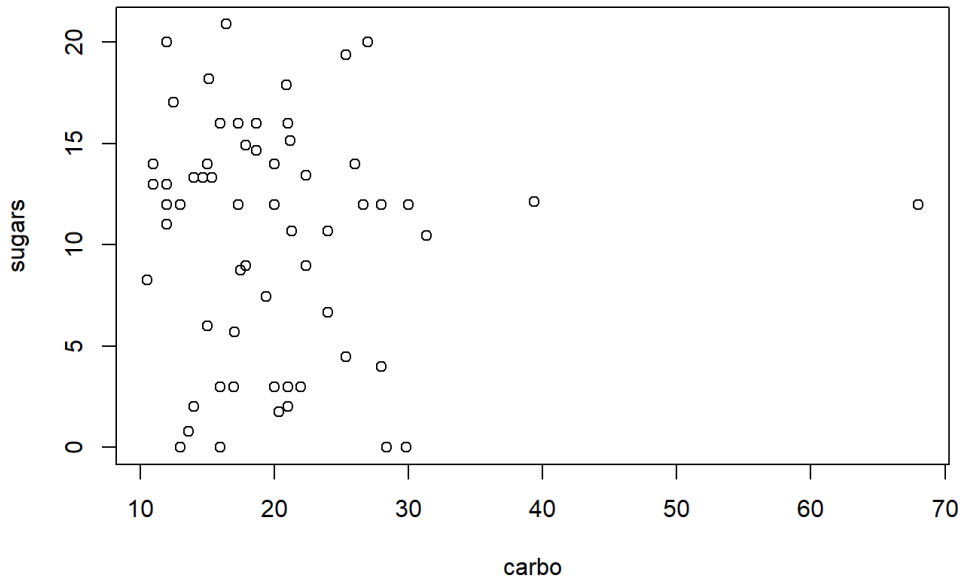
3. The relationship between fat and shelf

```
> boxplot(fat ~ shelf)
```



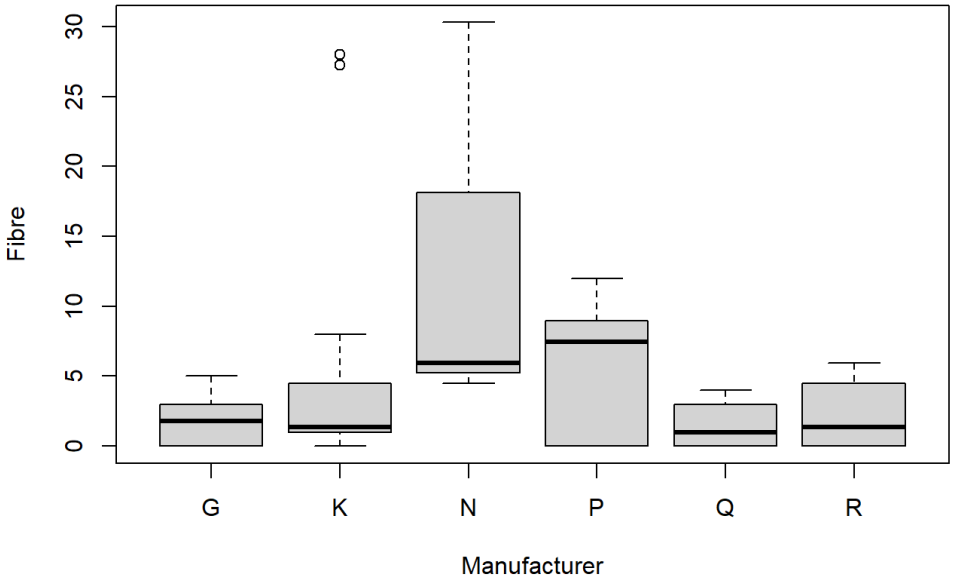
4. The relationship between carbohydrates and sugars

```
> cor(carbo, sugars)
[1] -0.04082599
> plot(carbo, sugars)
```



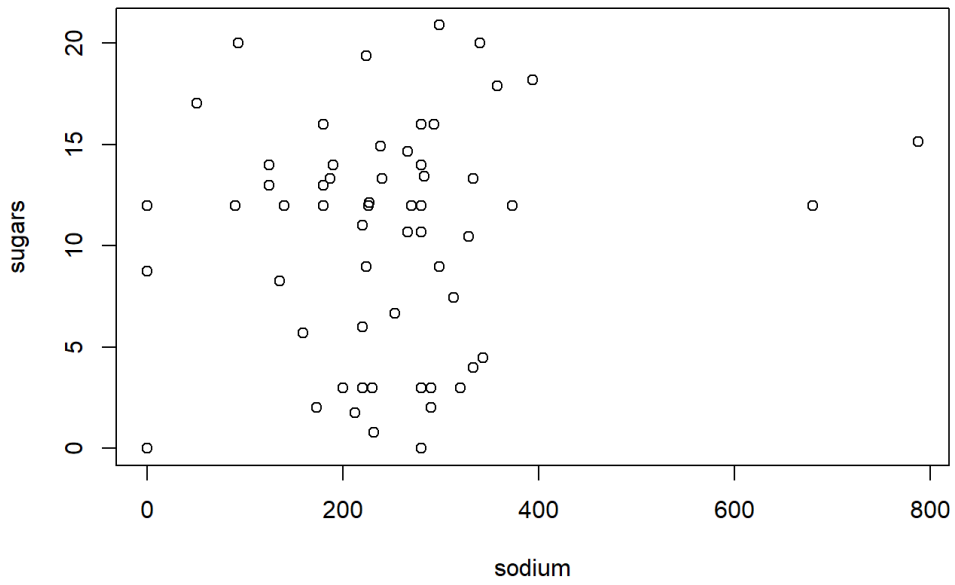
5. The relationship between fibre and manufacturer

```
> boxplot(fibre ~ mfr, xlab = "Manufacturer", ylab = "Fibre")
```



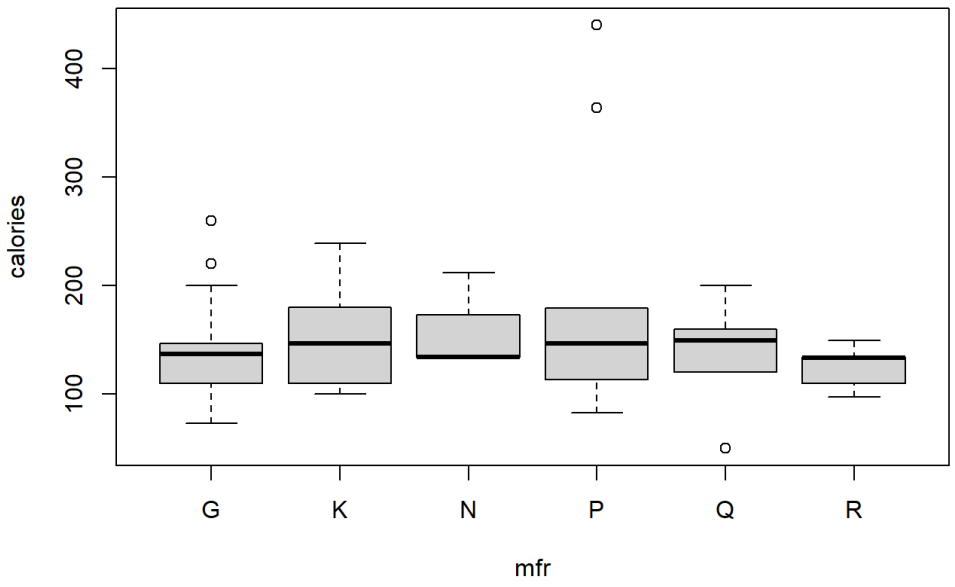
6. The relationship between sodium and sugars

```
> cor(sodium, sugars)
[1] 0.2112437
> plot(sodium, sugars)
```

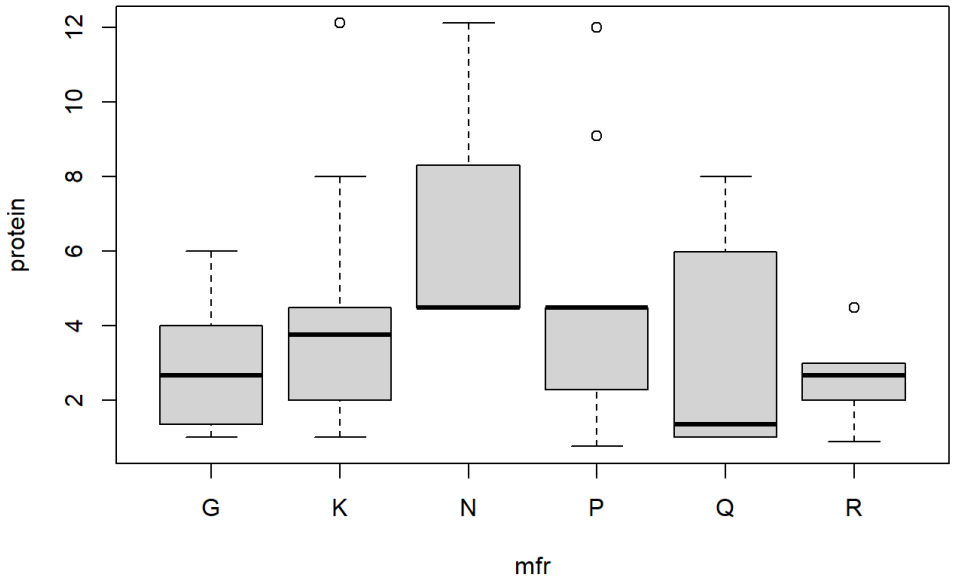


Are there other relationships you can predict and investigate?

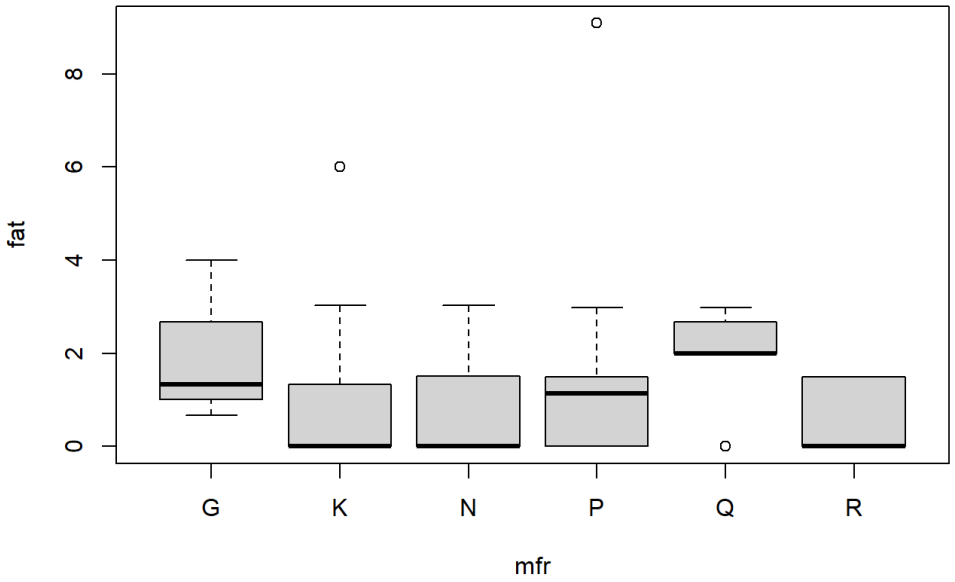
```
> boxplot(calories ~ mfr)
```



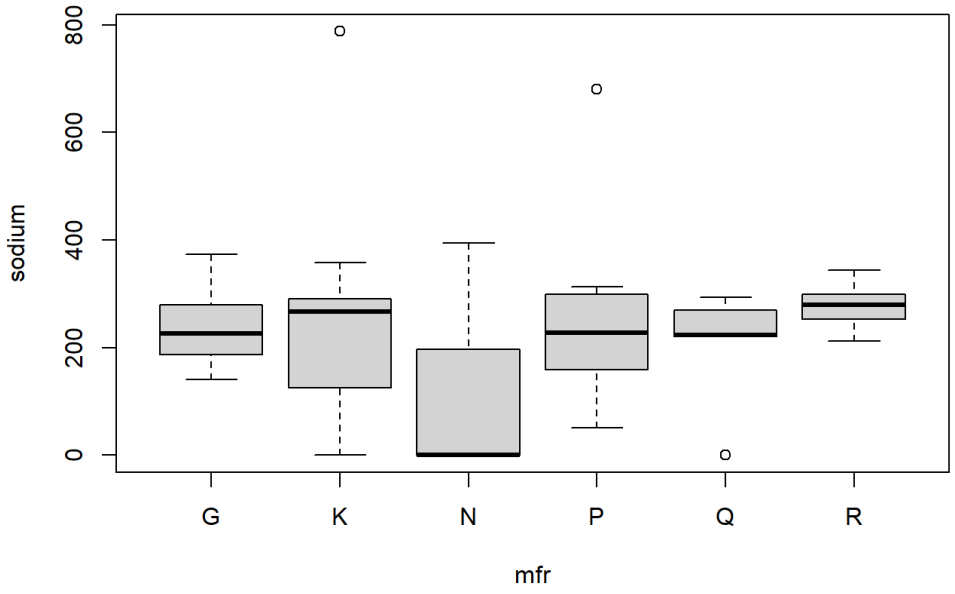
```
> boxplot(protein ~ mfr)
```



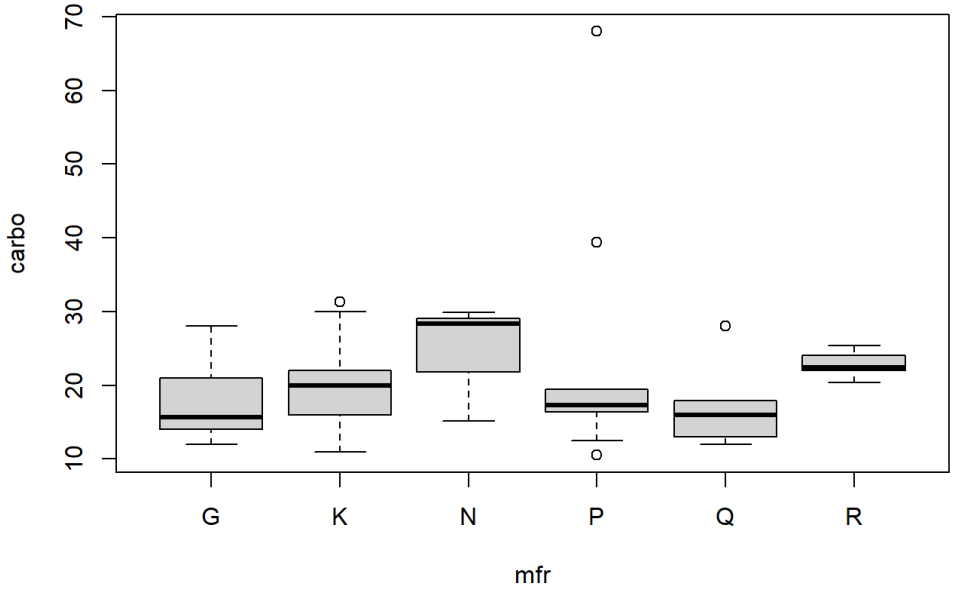

```
> boxplot(fat ~ mfr)
```



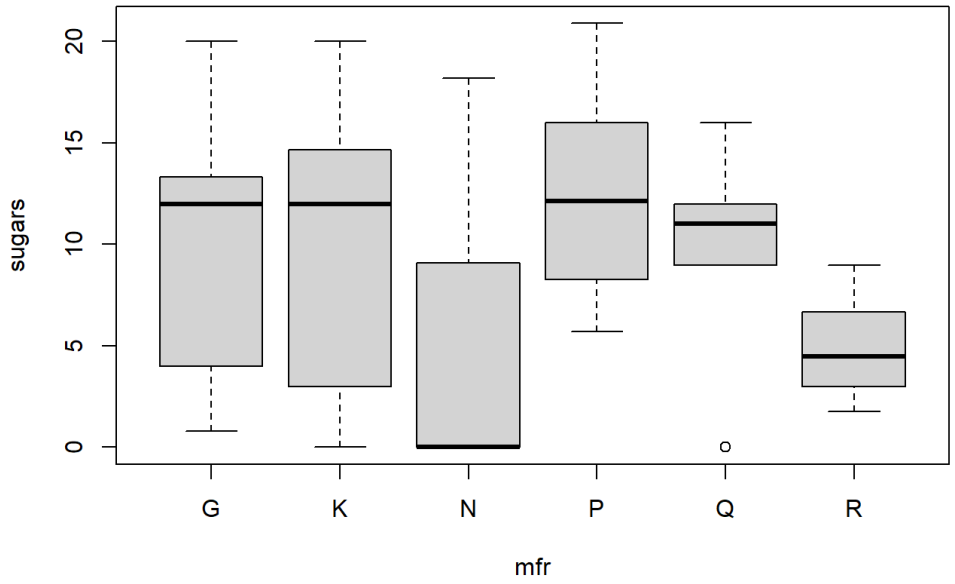
```
> boxplot(sodium ~ mfr)
```



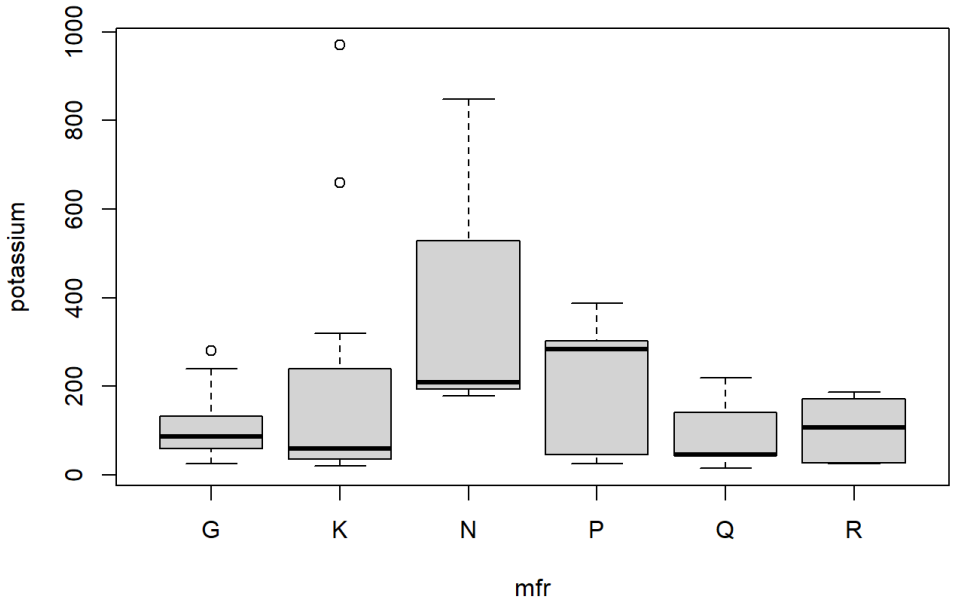

```
> boxplot(carbo ~ mfr)
```



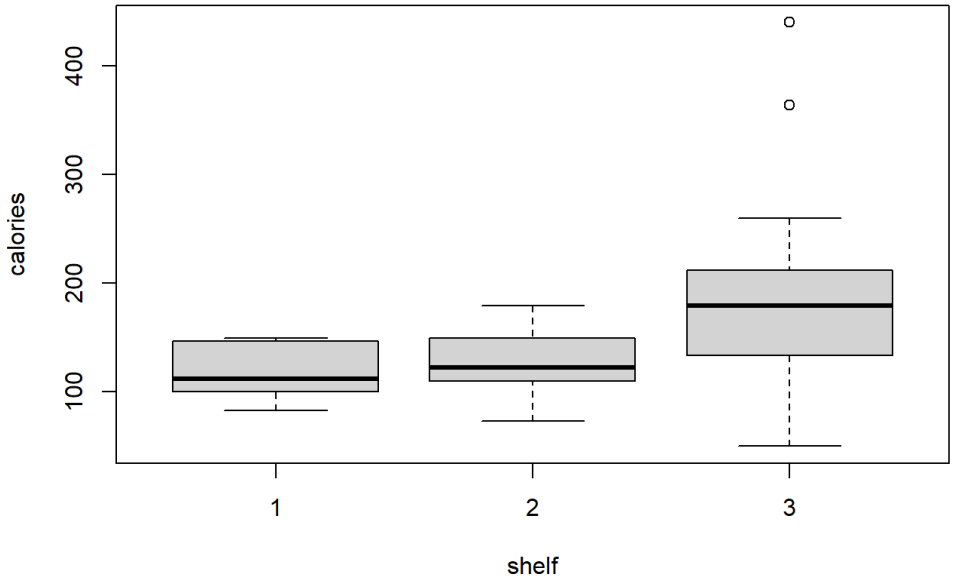
```
> boxplot(sugars ~ mfr)
```



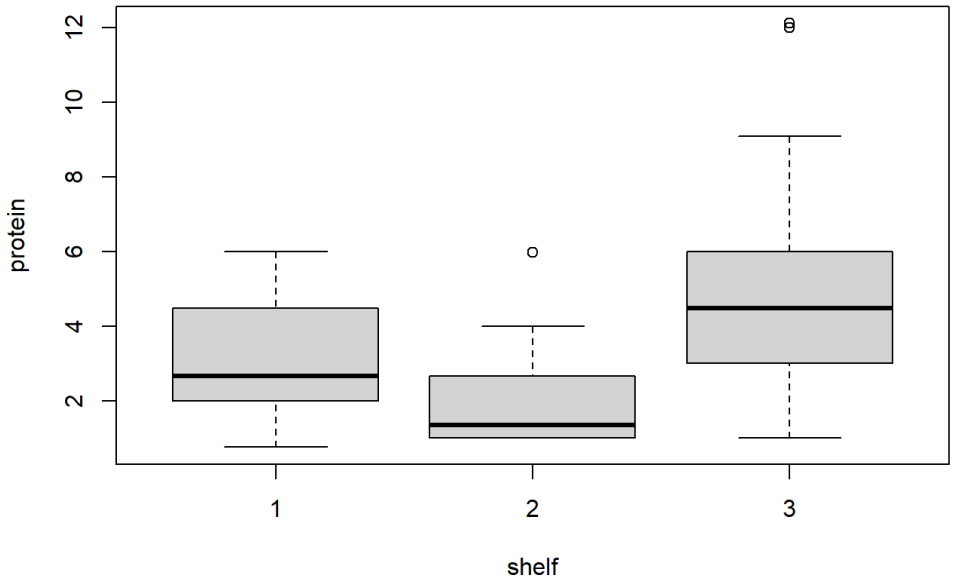

```
> boxplot(potassium ~ mfr)
```



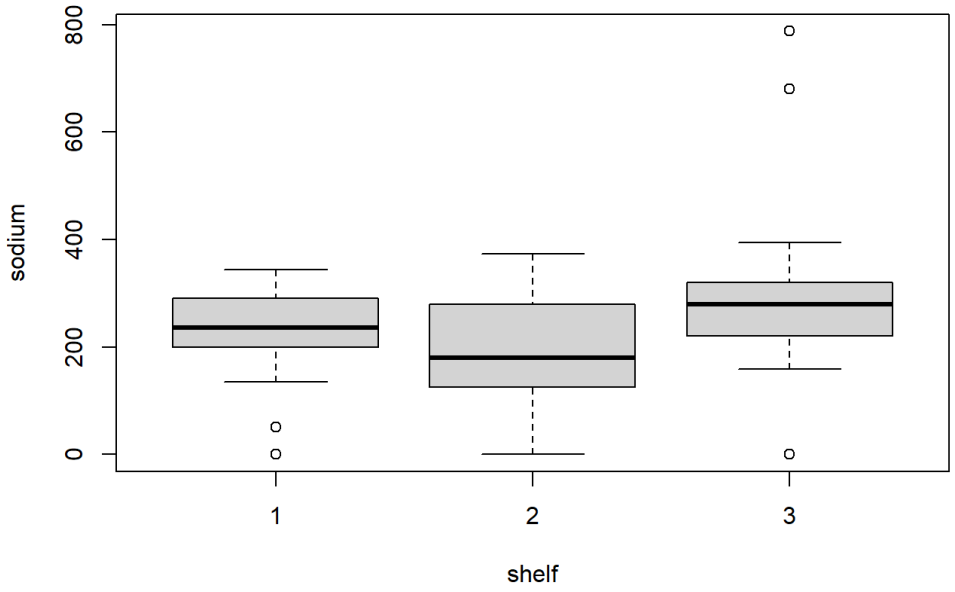
```
> boxplot(calories ~ shelf)
```



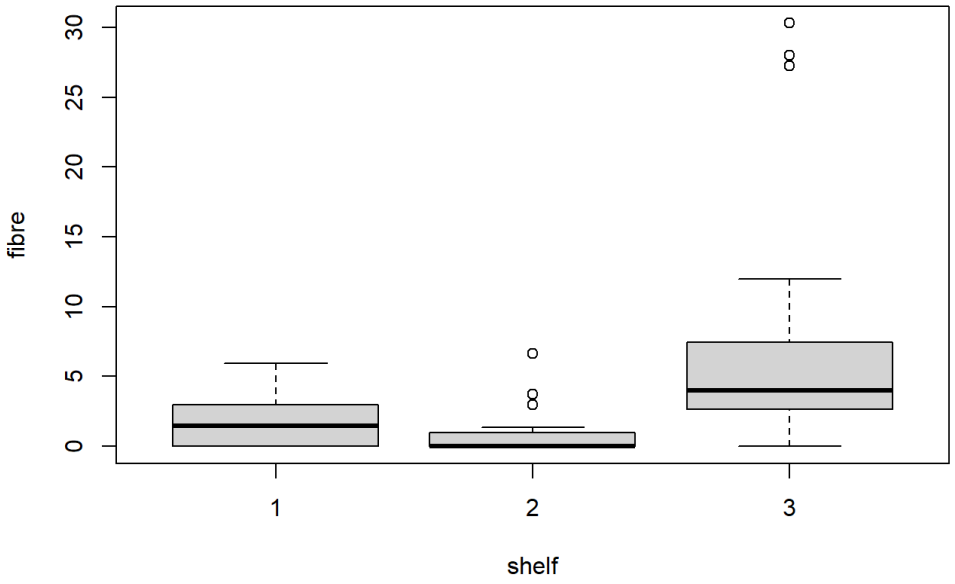

```
> boxplot(protein ~ shelf)
```



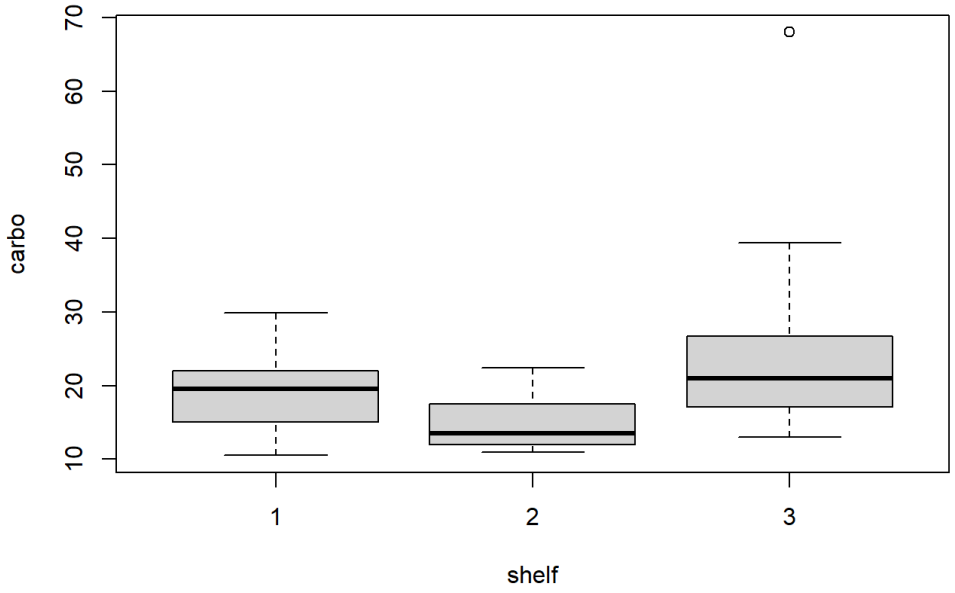
```
> boxplot(sodium ~ shelf)
```



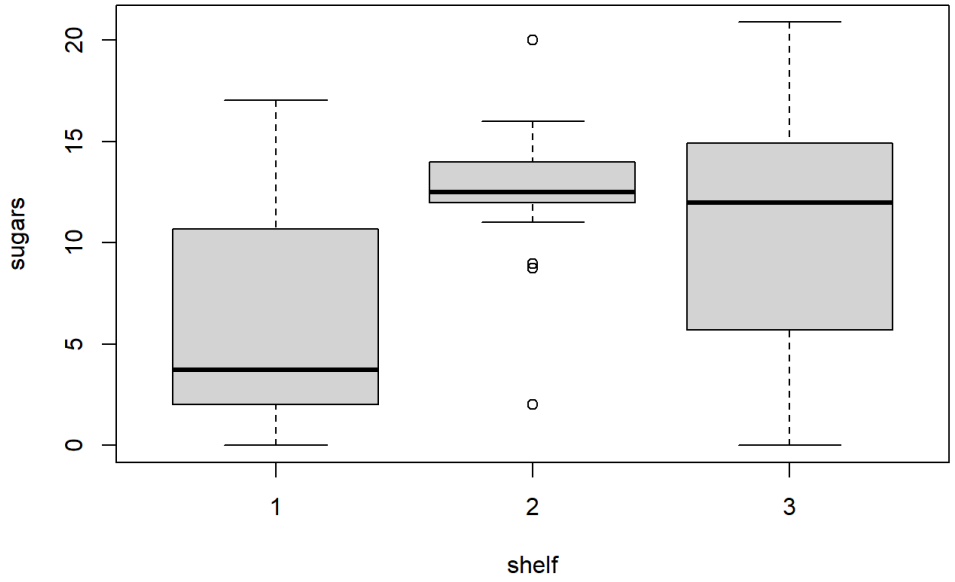

```
> boxplot(fibre ~ shelf)
```



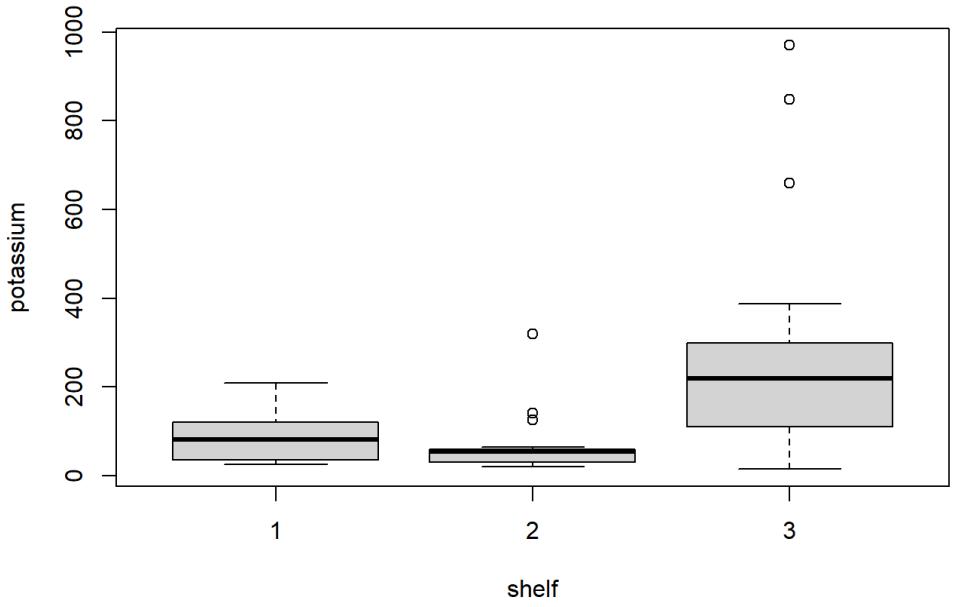
```
> boxplot(carbo ~ shelf)
```



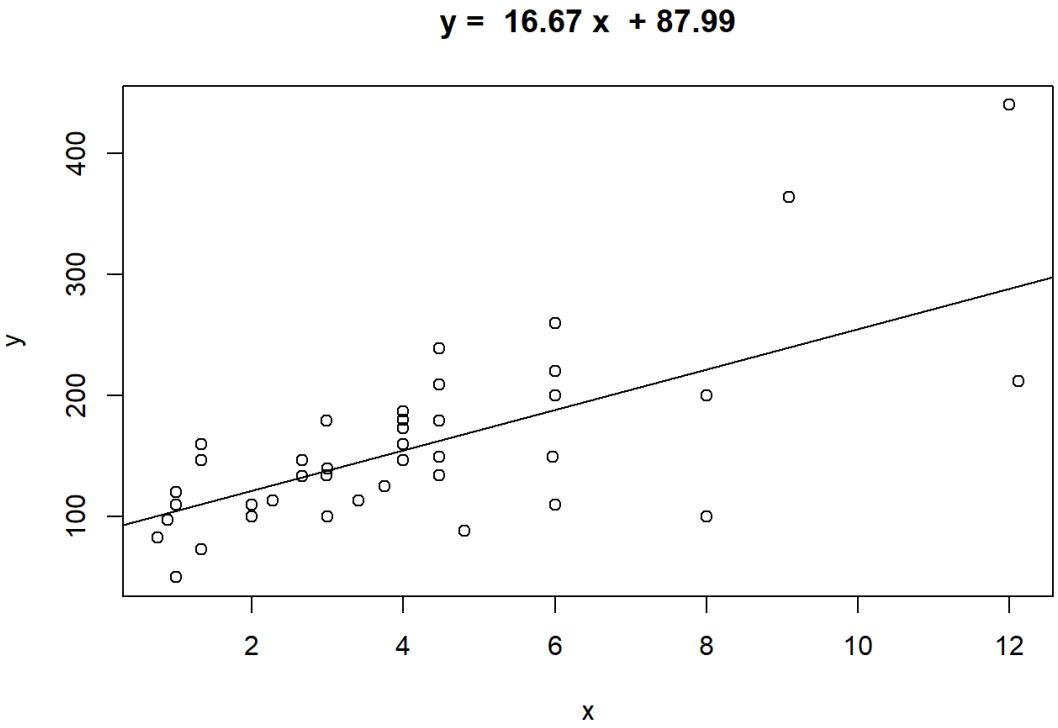

```
> boxplot(sugars ~ shelf)
```



```
> boxplot(potassium ~ shelf)
```



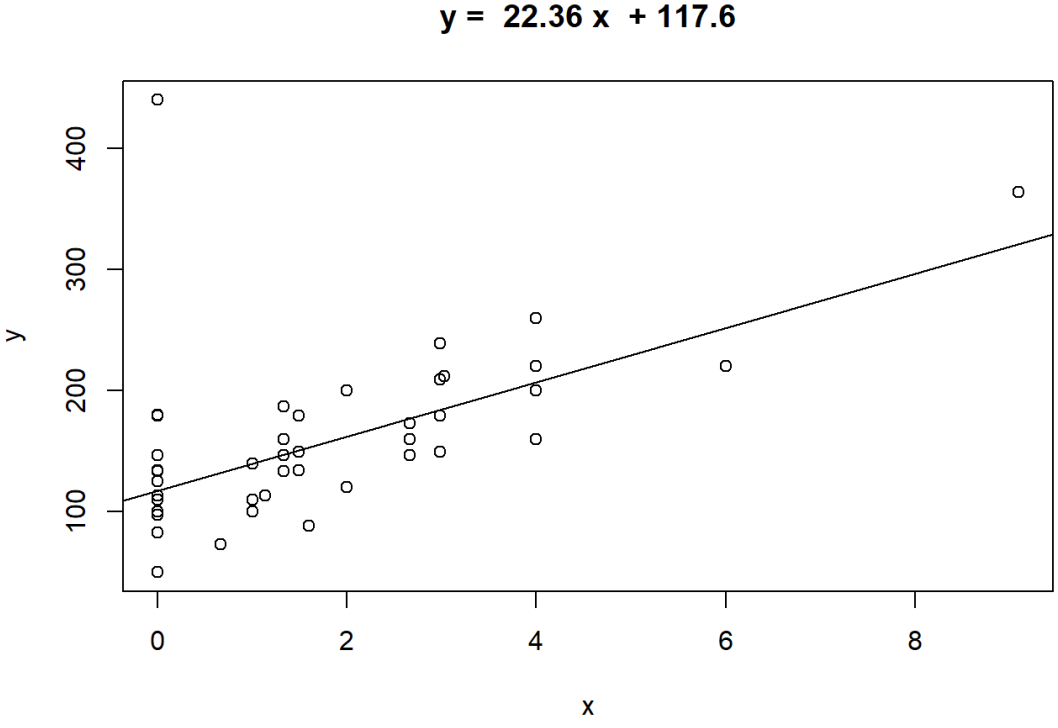

```
> library(UsingR)

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':

    cancer
The following object is masked from 'UScereal':

    fat

> attach(UScereal)
The following object is masked from package:UsingR:

    fat
The following objects are masked from UScereal (pos = 10):

    calories, carbo, fat, fibre, mfr, potassium, protein, shelf,
    sodium, sugars, vitamins

> cor(protein, calories)
[1] 0.7060105
> simple.lm(protein, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
87.99 16.67

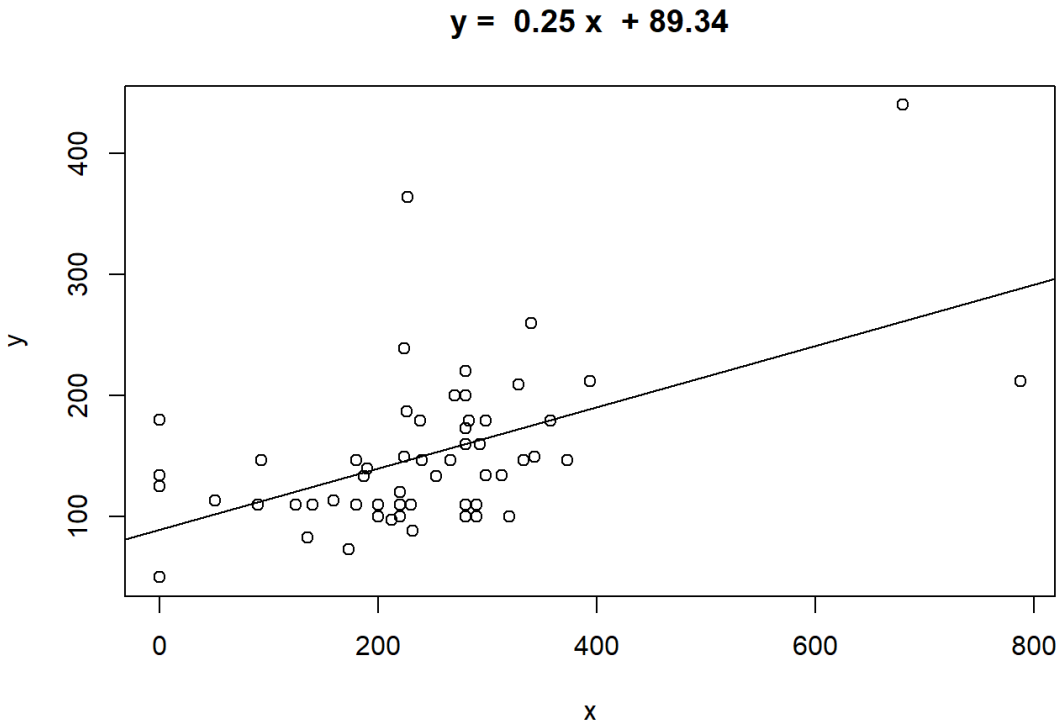
```
> cor(fat, calories)
[1] 0.5901757
> simple.lm(fat, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
117.60 22.36

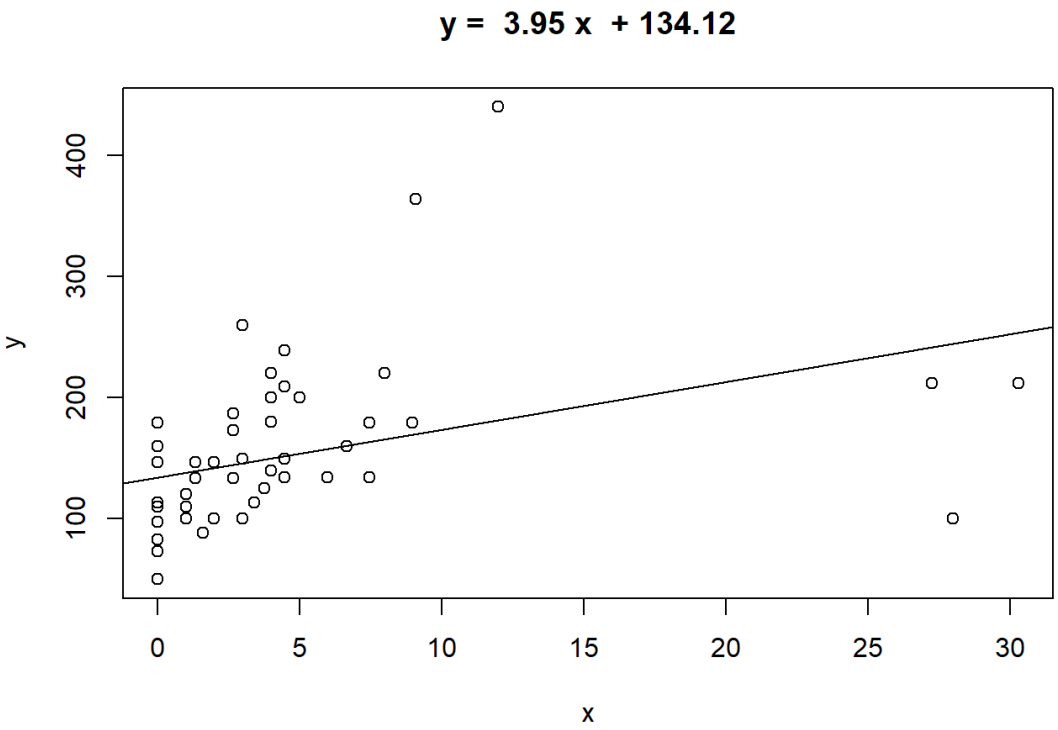
```
> cor(sodium, calories)
[1] 0.5286552
> simple.lm(sodium, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
89.3352 0.2526

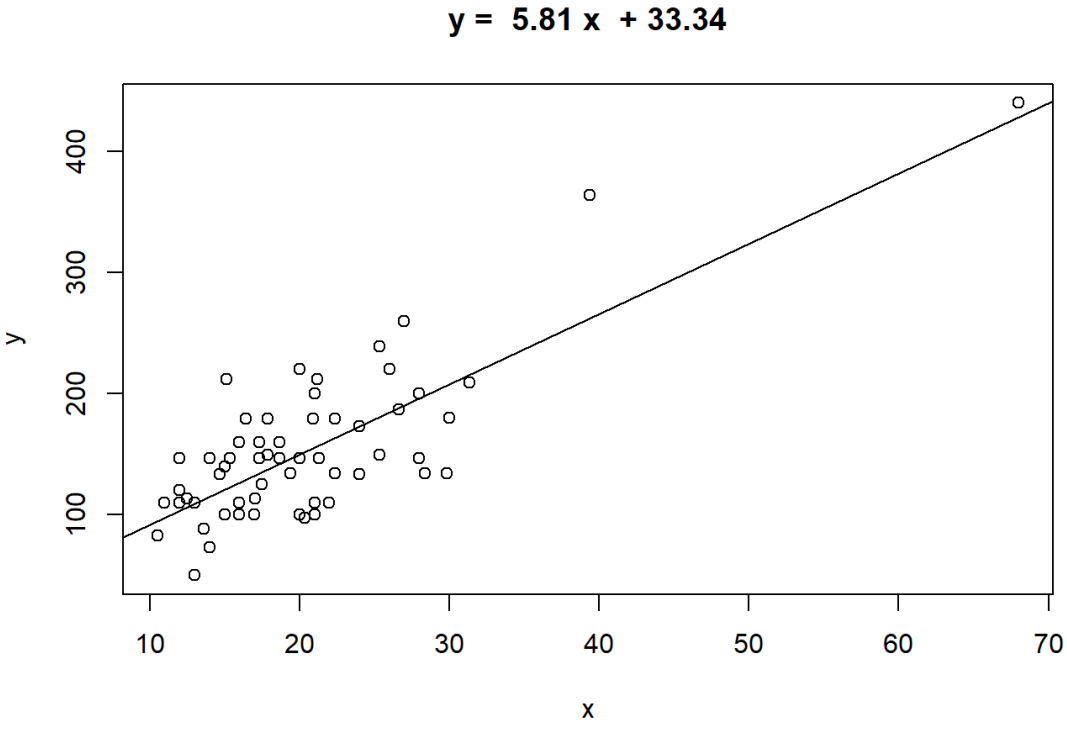
```
> cor(fibre, calories)
[1] 0.3882179
> simple.lm(fibre, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
134.12 3.95

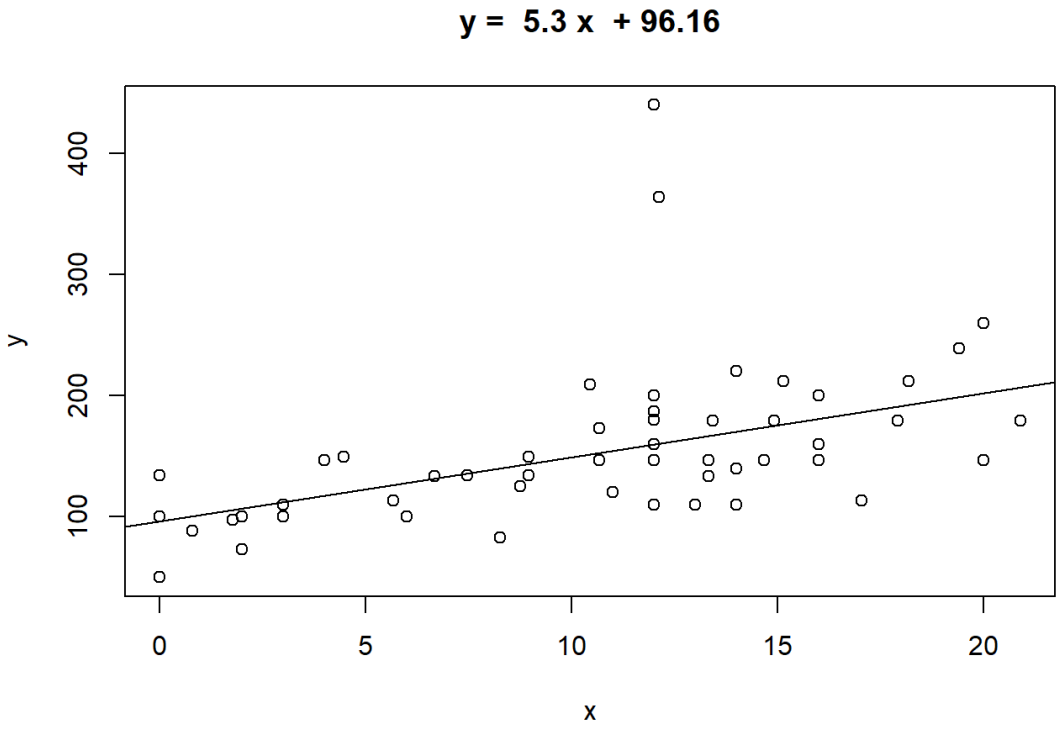
```
> cor(carbo, calories)
[1] 0.7887227
> simple.lm(carbo, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
33.340 5.813

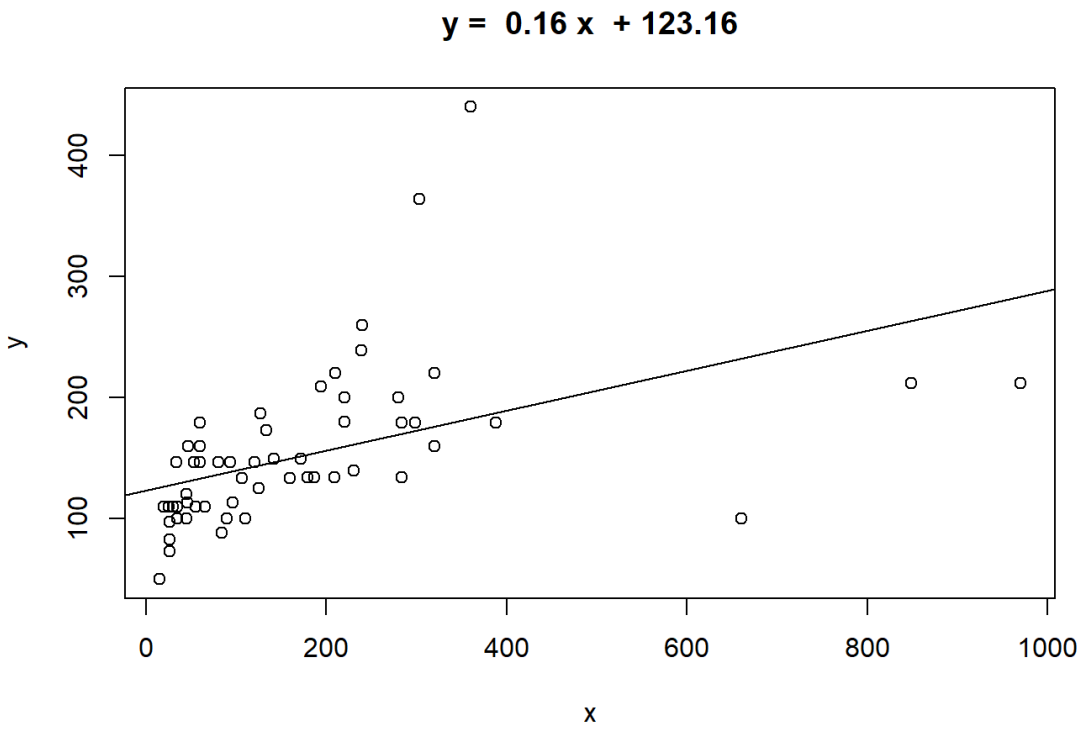
```
> cor(sugars, calories)
[1] 0.4952942
> simple.lm(sugars, calories)
```



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
96.164 5.298

```
> cor(potassium, calories)
[1] 0.4765955
> simple.lm(potassium, calories)
```




```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    123.156        0.165
```

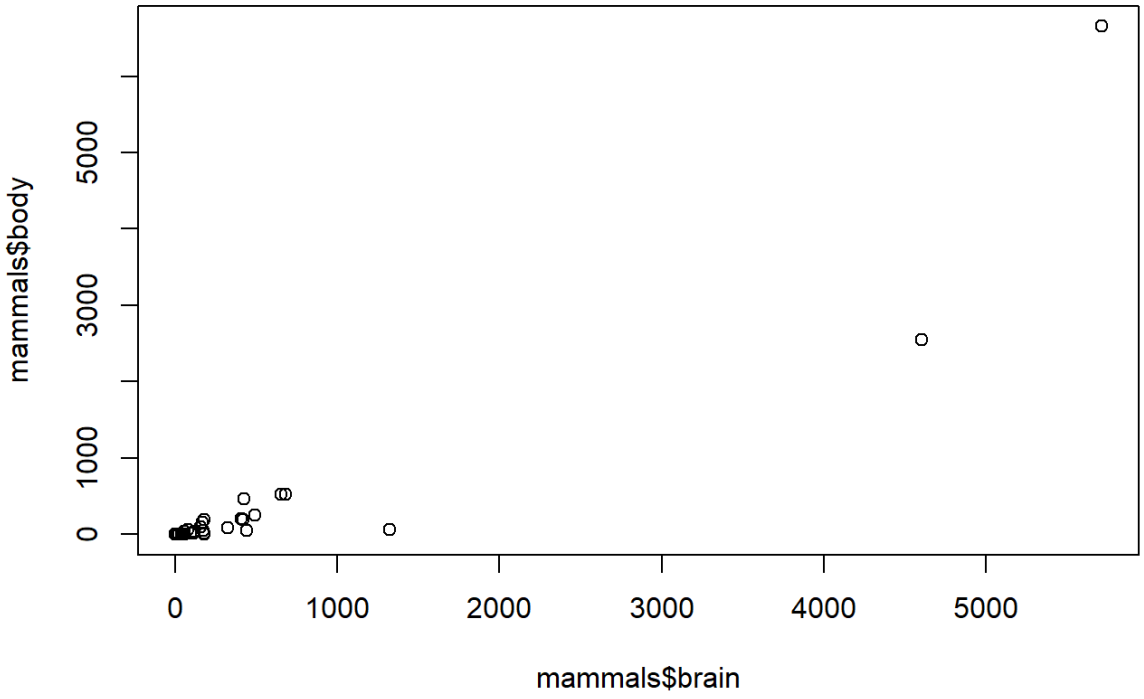
Problem 4.3

The built-in data set mammals contains data on body weight versus brain weight. Use the cor to find the Pearson and Spearman correlation coefficients. Are they similar?

```
> cor(mammals$brain, mammals$body)
[1] 0.9341638
> cor(rank(mammals$brain), rank(mammals$body))
[1] 0.9534986
```

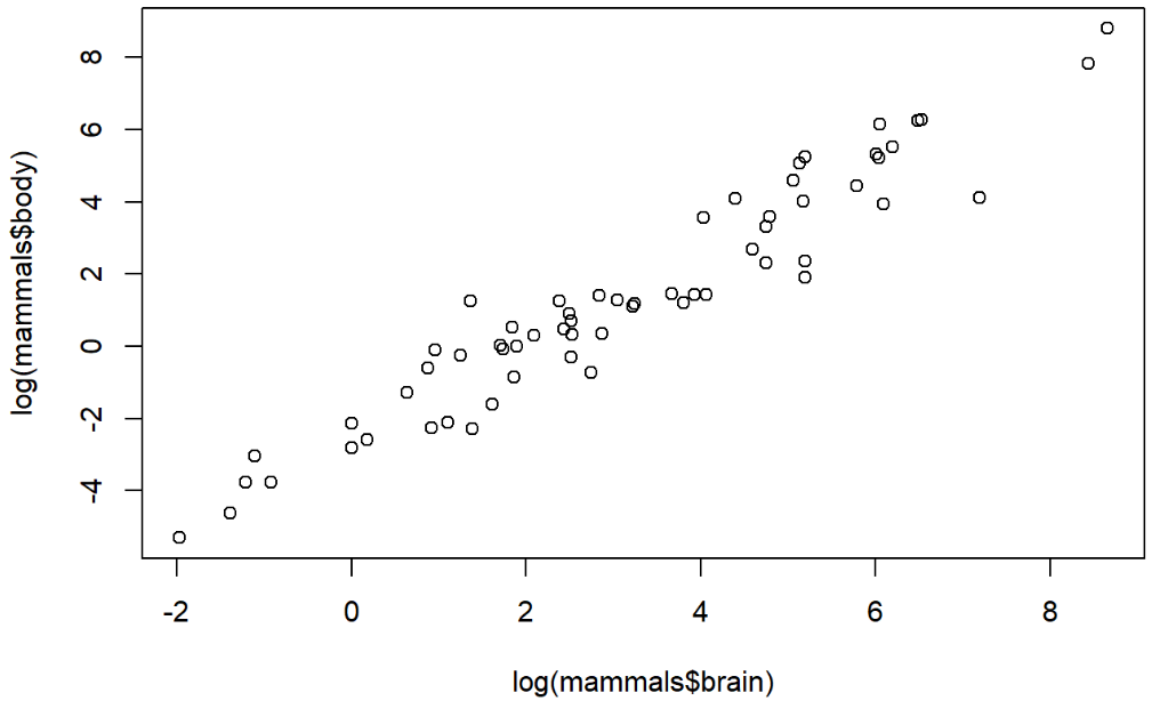
Plot the data using the plot command and see if you expect them to be similar.

```
> plot(mammals$brain, mammals$body)
```



You should be unsatisfied with this plot. Next, plot the logarithm (log) of each variable and see if that makes a difference.

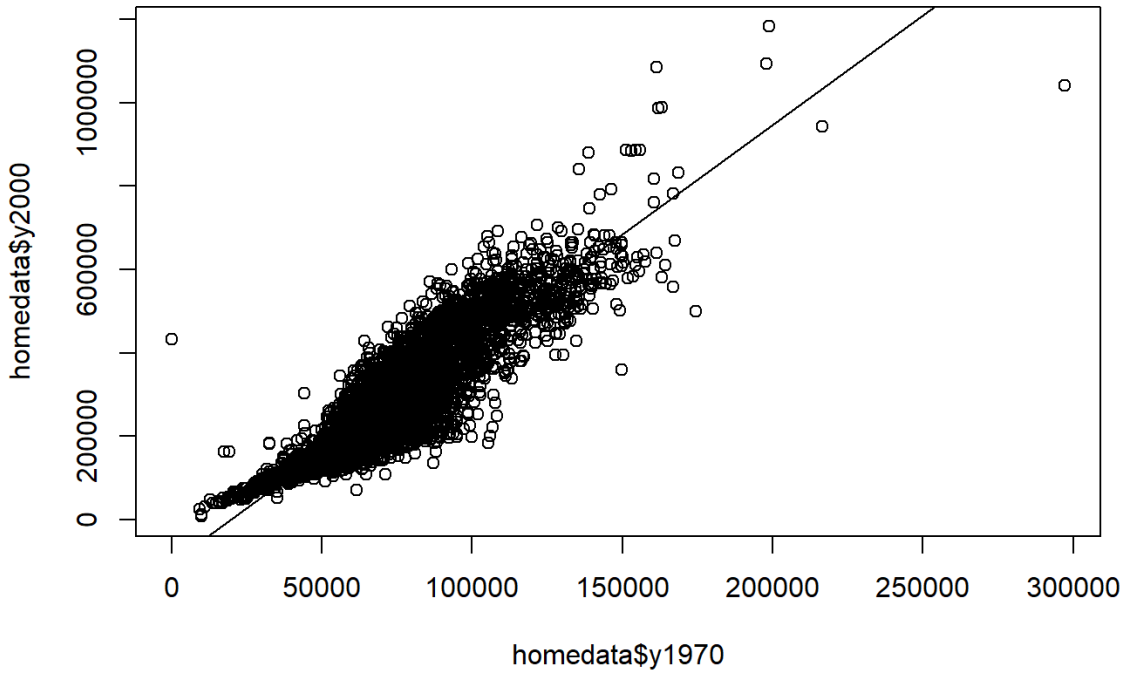
```
> plot(log(mammals$brain), log(mammals$body))
```



Problem 4.4

For the data set on housing prices, homedata, investigate the relationship between old assessed value and new. Use old as the predictor variable. Does the data suggest a linear relationship? Are there any outliers? What may have caused these outliers?

```
> plot(homedata$y2000 ~ homedata$y1970)
> model <- lm(homedata$y2000 ~ homedata$y1970)
> summary(model)

Call:
lm(formula = homedata$y2000 ~ homedata$y1970)

Residuals:
    Min      1Q  Median      3Q     Max
-416665  -36308     809   34372  536605

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.040e+05  2.337e+03  -44.51   <2e-16 ***
homedata$y1970  5.258e+00  3.147e-02  167.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58000 on 6839 degrees of freedom
Multiple R-squared:  0.8032,    Adjusted R-squared:  0.8032
F-statistic: 2.791e+04 on 1 and 6839 DF,  p-value: < 2.2e-16

> abline(model)
```



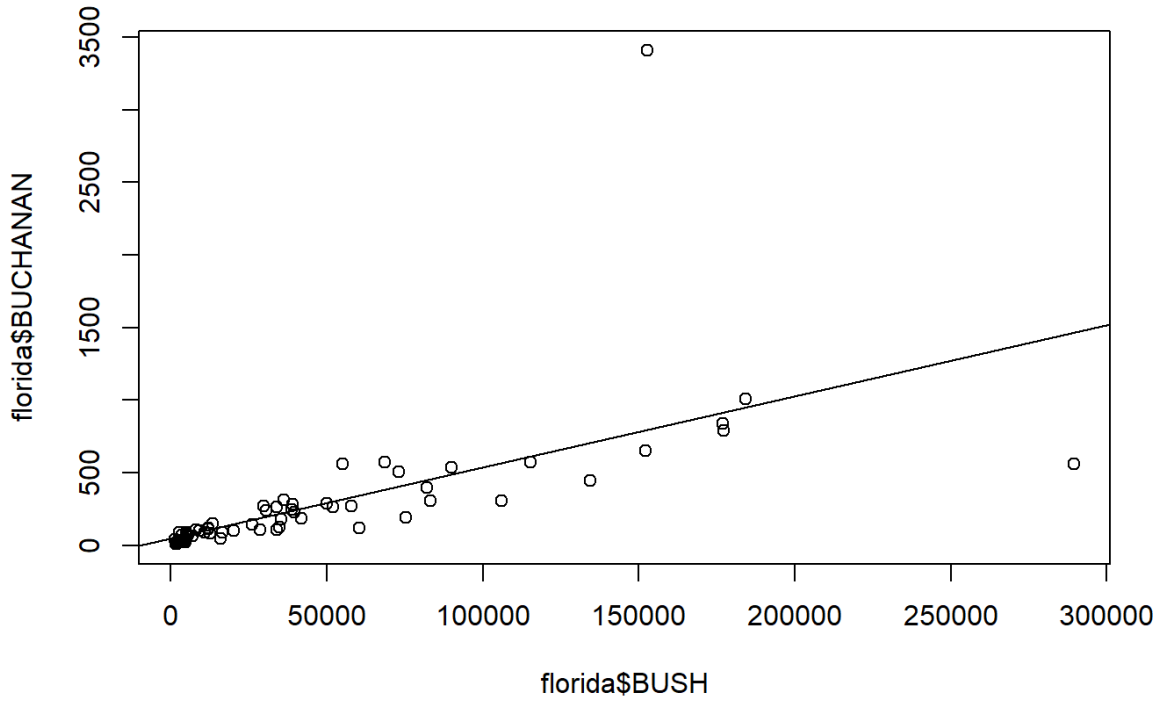
What is the predicted new assessed value for a $75,000 house in 1970.

```
> model$coefficients[1] + model$coefficients[2] * 75000
(Intercept)
   290343.2
```

Problem 4.5

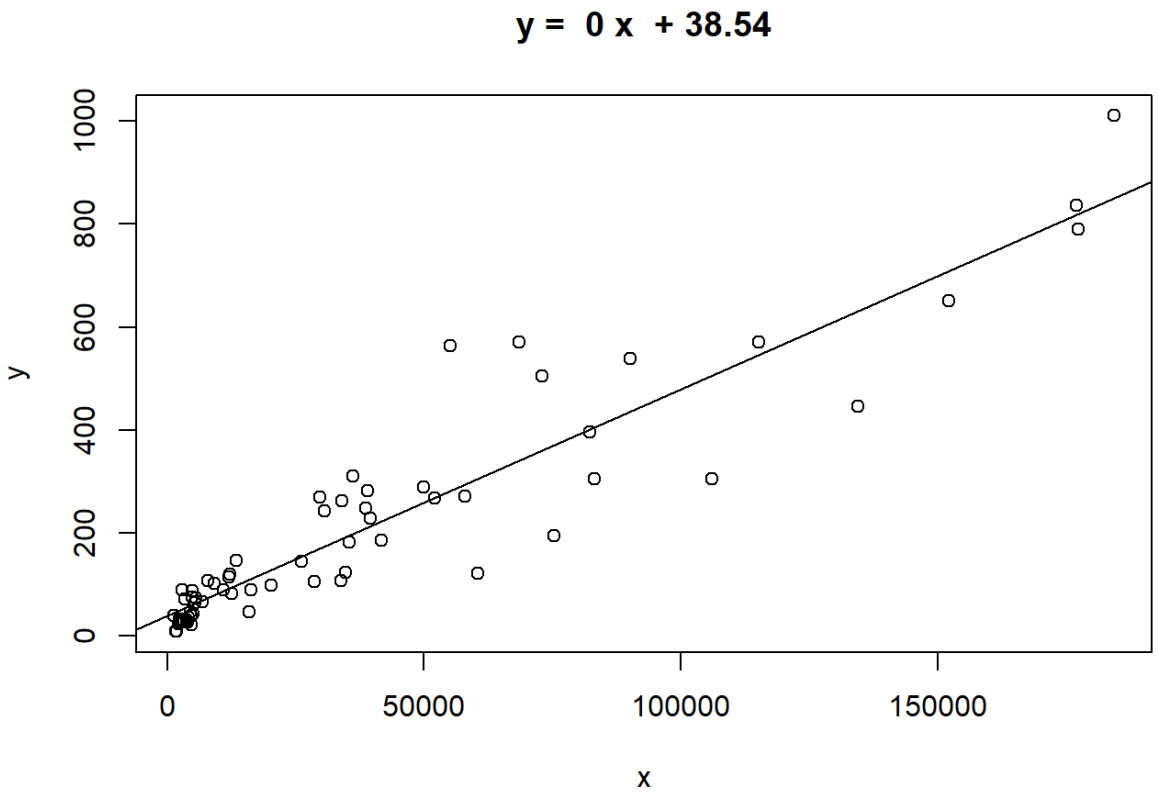
For the florida data set of Bush vs. Buchanan, there is another obvious outlier that indicated Buchanan received fewer votes than expected. If you remove both the outliers, what is the predicted value for the number of votes Buchanan would get in Miami-Dade county based on the number of Bush votes?

```
> plot(florida$BUCHANAN ~ florida$BUSH)
> abline(lm(florida$BUCHANAN ~ florida$BUSH))
> identify(florida$BUSH, florida$BUCHANAN)
```




```
integer(0)
> florida[13,]
       County   GORE    BUSH BUCHANAN NADER BROWN HAGELIN HARRIS
MCREYNOLDS
13   DADE 328702 289456      561  5355   759     119    88       36
  MOOREHEAD PHILLIPS  Total
13       124       69 625269
```

```
> florida.cleaned <- florida[-c(13, 50), ]
> linearmodel <- simple.lm(florida.cleaned$BUSH, florida.cleaned$BUCHANAN)
```

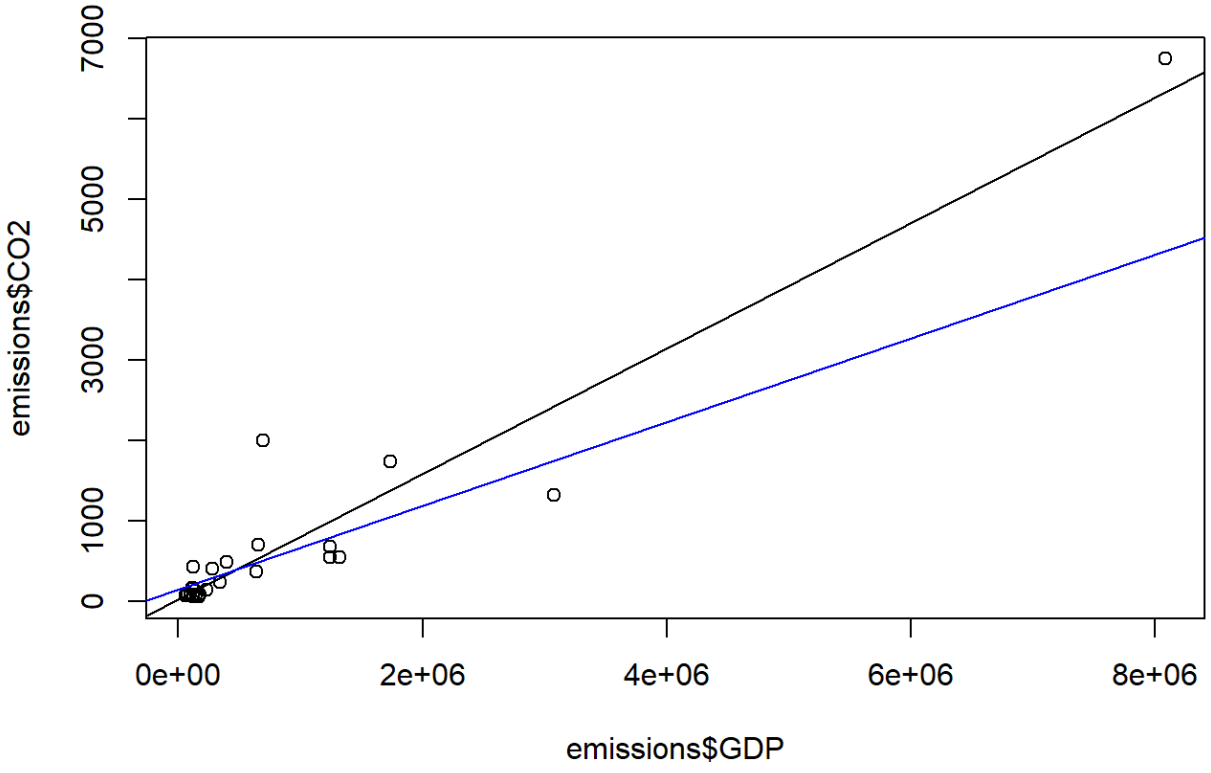



```
> summary(linearmodel)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-200.94  -28.47  -11.06   27.52  281.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.854e+01  1.314e+01   2.934  0.00467 **
x           4.404e-03  2.193e-04  20.077  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.03 on 63 degrees of freedom
Multiple R-squared:  0.8648,    Adjusted R-squared:  0.8627
F-statistic: 403.1 on 1 and 63 DF,  p-value: < 2.2e-16
> linearmodel$coefficients[1] + linearmodel$coefficients[2] * florida[13,"BUSH"]
(Intercept)
     1313.2
```

Problem 4.6

For the data set emissions plot the per-Capita GDP (gross domestic product) as a predictor for the response variable CO_2 emissions. Identify the outlier and find the regression lines with this point, and without this point.

```
> plot(emissions$CO2 ~ emissions$GDP)
> model <- lm(emissions$CO2 ~ emissions$GDP)
> abline(model)
> identify(emissions$GDP, emissions$CO2)
integer(0)
> model.clean <- lm(emissions[-1,]$CO2 ~ emissions[-1,]$GDP)
> abline(model.clean, col = "Blue")
```



Problem 4.7

Attach the data set babies:

```
> attach(babies)
```

This data set contains much information about babies and their mothers for 1236 observations. Find the correlation coefficient (both Pearson and Spearman) between age and weight.

```
> cor(wt, age)
[1] 0.02904064
```

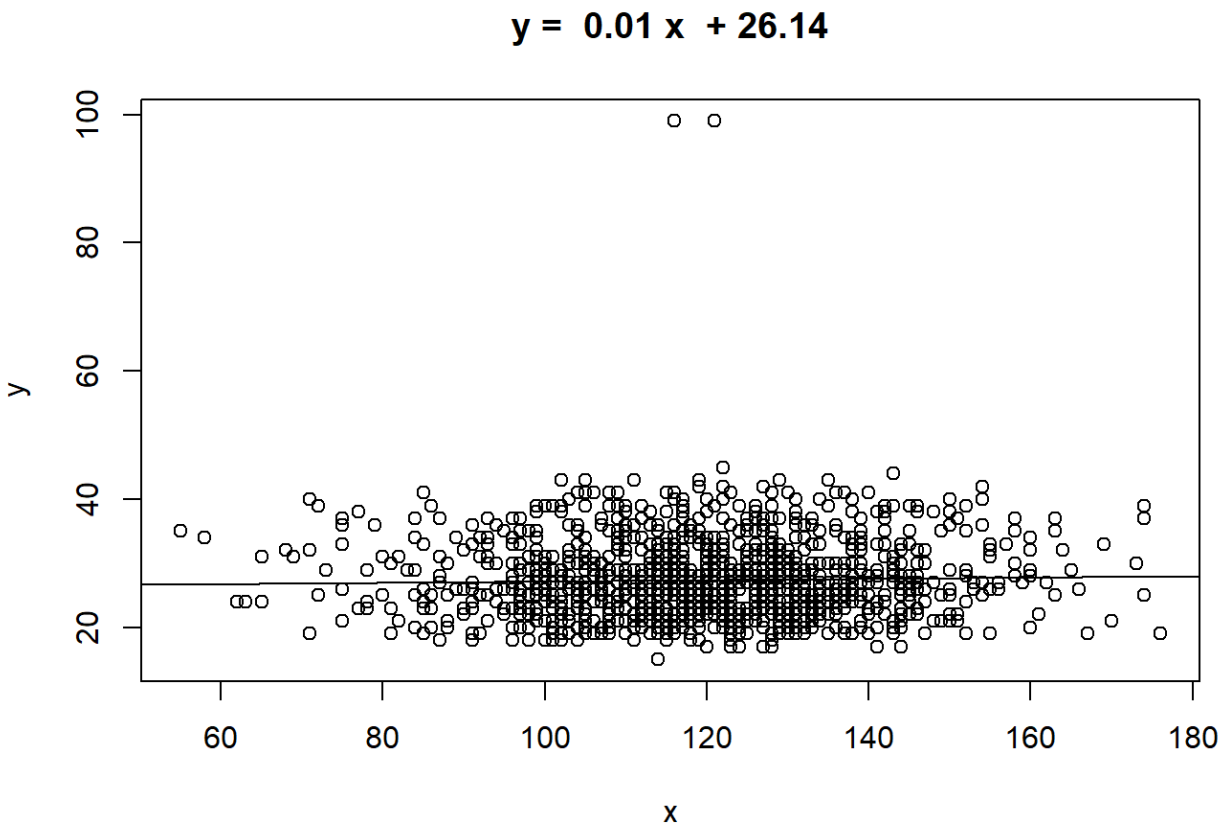
```
> cor(rank(wt), rank(age))
[1] 0.04170028
```

Repeat for the relationship between height and weight.

```
> cor(wt, ht)
[1] 0.1255413
> cor(rank(wt), rank(ht))
[1] 0.214745
```

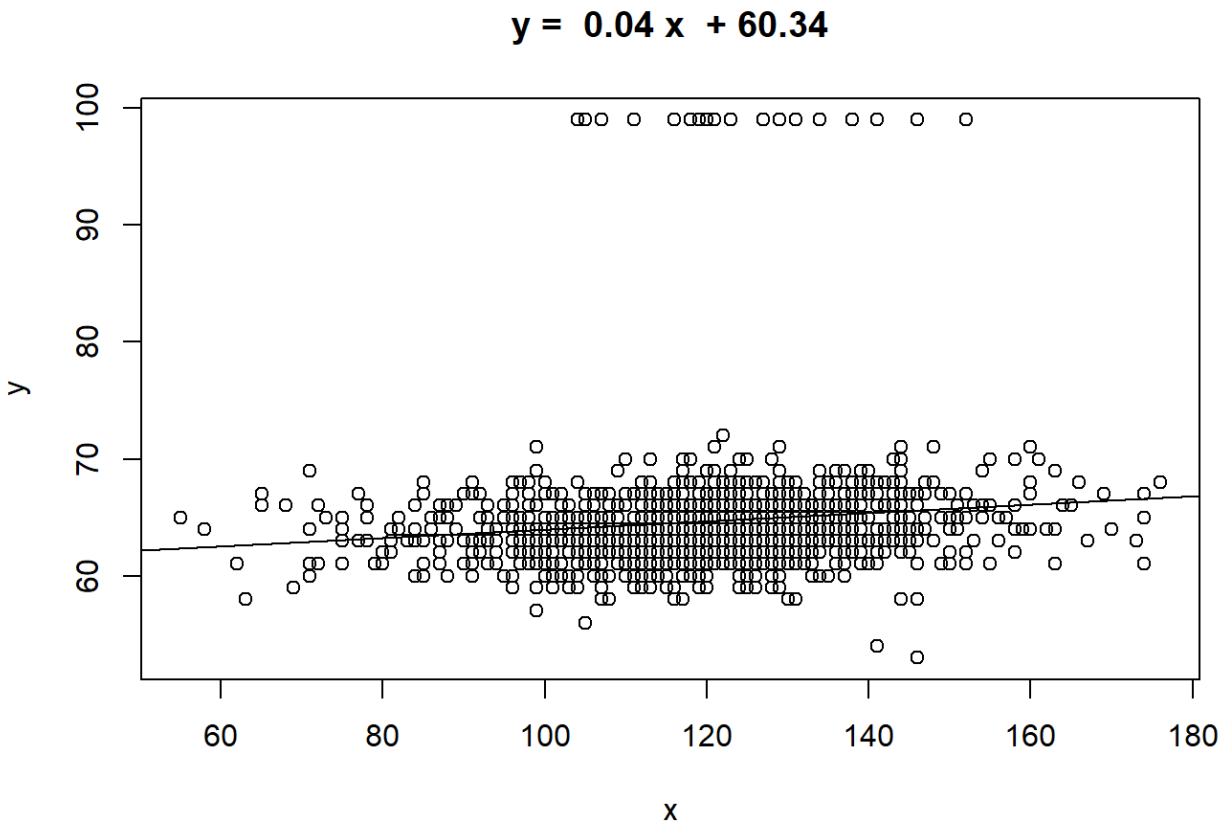
Make scatter plots of each pair and see if your answer makes sense.

```
> simple.lm(wt, age)
```




```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
   26.14182      0.01028
```

> simple.lm(wt, ht)



Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
60.33911 0.03622

Problem 4.8

Find a dataset that is a candidate for linear regression (you need two numeric variables, one a predictor and one a response.) Make a scatterplot with regression line using R.

> data()

Problem 4.9

The built-in data set mtcars contains information about cars from a 1974 Motor Trend issue. Try to answer the following:

1. What are the variable names? (Try names.)

> names(mtcars)
[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"
[11] "carb"

2. What is the maximum mpg

```
> max(mtcars$mpg)
[1] 33.9
```

3. Which car has this?

```
> rownames(mtcars[which.max(mtcars$mpg), ])
[1] "Toyota Corolla"
```

4. What are the first 5 cars listed?

```
> rownames(mtcars)[1:5]
[1] "Mazda RX4"         "Mazda RX4 Wag"     "Datsun 710"
[4] "Hornet 4 Drive"    "Hornet Sportabout"
```

5. What horsepower (hp) does the “Valiant” have?

```
> mtcars["Valiant", "hp"]
[1] 105
```

6. What are all the values for the Mercedes 450slc (Merc 450SLC)?

```
> mtcars["Merc 450SLC", ]
              mpg cyl  disp  hp drat   wt qsec vs am gear carb
Merc 450SLC 15.2   8 275.8 180 3.07 3.78   18  0  0    3    3
```

7. Make a scatterplot of cylinders (cyl) vs. miles per gallon (mpg). Fit a regression line. Is this a good candidate for linear regression?