Homework 3

Miles Tweed

Problem 1

1.1

```r
# These empty vectors will be filled with the values calculated in the looping construct.
# They will be used to create a matrix object in later exercises.
mean_vec <- vector()
sd_vec <- vector()

# The first for loop iterates over every column
for (i in 1:length(cystfibr)) {
  # the mean for the i'th column is calculated and printed
  cf_mean <- sum(cystfibr[i])/length(cystfibr[,1])
  print(paste('The mean for column', names(cystfibr)[i], 'is: ', cf_mean))
  # This vector will contained the squared differences between eachvalue and the mean
  x = vector()
  # This for loop iterates over each value in the i'th column
  for (j in 1:length(cystfibr[,i])) {
    # the squared difference from the mean is calculated for each value and stored in x
    diff <- (cystfibr[j,i]-cf_mean)^2
    x = append(x, diff)
  }
  # The sample standard deviation is the square root of the sum of the values in x
  # divided by the length minus one. The standard deviation is printed.
  cf_sd <- sqrt(sum(x)/(length(cystfibr[,i])-1))
  print(paste('The standard deviation for column', names(cystfibr)[i], 'is: ', cf_sd))

  # The mean and sd for each colum is stored in these variables to be used later.
  mean_vec <- append(mean_vec, cf_mean)
  sd_vec <- append(sd_vec, cf_sd)
  }
```

```
## [1] "The mean for column age is:  14.48"
## [1] "The standard deviation for column age is:  5.05898540552682"
## [1] "The mean for column sex is:  0.44"
## [1] "The standard deviation for column sex is:  0.506622805119022"
## [1] "The mean for column height is:  152.8"
## [1] "The standard deviation for column height is:  21.5"
## [1] "The mean for column weight is:  38.404"
## [1] "The standard deviation for column weight is:  17.8981256001851"
## [1] "The mean for column bmp is:  78.28"
## [1] "The standard deviation for column bmp is:  12.0052766176655"
## [1] "The mean for column fev1 is:  34.72"
## [1] "The standard deviation for column fev1 is:  11.1971722620788"
```

```
## [1] "The mean for column rv is:  255.2"
## [1] "The standard deviation for column rv is:  86.0169556928555"
## [1] "The mean for column frc is:  155.4"
## [1] "The standard deviation for column frc is:  43.7187983976382"
## [1] "The mean for column tlc is:  114"
## [1] "The standard deviation for column tlc is:  16.9681073389658"
## [1] "The mean for column pemax is:  109.12"
## [1] "The standard deviation for column pemax is:  33.4369057579595"
```

1.2

```r
# A matrix with dimensions [10,2] is created using the
# values stored in the looping construct.
cf_matrix <- matrix(c(mean_vec, sd_vec), nrow = 10, ncol = 2)
# The row names are the variable names from the original data frame
rownames(cf_matrix) <- names(cystfibr)
# The column names are added
colnames(cf_matrix) <- c('mean', 'sd')
cf_matrix
```

```
##           mean         sd
## age     14.480  5.0589854
## sex      0.440  0.5066228
## height 152.800 21.5000000
## weight  38.404 17.8981256
## bmp     78.280 12.0052766
## fev1    34.720 11.1971723
## rv     255.200 86.0169557
## frc    155.400 43.7187984
## tlc    114.000 16.9681073
## pemax  109.120 33.4369058
```

1.3

a)

The first 10 observations are:

```r
head(cystfibr, n = 10L)
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 1    7   0    109   13.1  68   32 258 183 137    95
## 2    7   1    112   12.9  65   19 449 245 134    85
## 3    8   0    124   14.1  64   22 441 268 147   100
## 4    8   1    125   16.2  67   41 234 146 124    85
## 5    8   0    127   21.5  93   52 202 131 104    95
## 6    9   0    130   17.5  68   44 308 155 118    80
## 7   11   1    139   30.7  89   28 305 179 119    65
## 8   12   1    150   28.4  69   18 369 198 103   110
## 9   12   0    146   25.1  67   24 312 194 128    70
## 10  13   1    155   31.5  68   23 413 225 136    95
```

b)

Observations #5, 10, and 15 are:

```r
cystfibr[c(5, 10, 15),]
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 5    8   0    127   21.5  93   52 202 131 104    95
## 10  13   1    155   31.5  68   23 413 225 136    95
## 15  16   1    160   35.9  66   31 302 133 101   134
```

c)

The last 10 observations are:

```r
tail(cystfibr, n = 10L)
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 16  17   1    153   34.8  70   29 204 118 120   134
## 17  17   0    174   44.7  70   49 187 104 103   165
## 18  17   1    176   60.1  92   29 188 129 130   120
## 19  17   0    171   42.6  69   38 172 130 103   130
## 20  19   1    156   37.2  72   21 216 119  81    85
## 21  19   0    174   54.6  86   37 184 118 101    85
## 22  20   0    178   64.0  86   34 225 148 135   160
## 23  23   0    180   73.8  97   57 171 108  98   165
## 24  23   0    175   51.1  71   33 224 131 113    95
## 25  23   0    179   71.5  95   52 225 127 101   195
```

d)

The name of the first variable is:

```r
names(cystfibr[1])
```

```
## [1] "age"
```

And it contains the values:

```r
cystfibr[,1]
```

```
##  [1]  7  7  8  8  8  9 11 12 12 13 13 14 14 15 16 17 17 17 17 19 19 20 23 23 23
```

e)

The values #5, 10, and 15 from the first variable are:

```r
cystfibr[c(5,10,15),1]
```

```
## [1]  8 13 16
```

f)

The information for people with above average height is:

```r
cystfibr[cystfibr$height > mean(cystfibr$height),]
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 10  13   1    155   31.5  68   23 413 225 136    95
## 11  13   0    156   39.9  89   39 206 142  95   110
## 12  14   1    153   42.1  90   26 253 191 121    90
## 13  14   0    160   45.6  93   45 174 139 108   100
## 14  15   1    158   51.2  93   45 158 124  90    80
```

```
## 15   16   1    160   35.9  66   31 302 133 101   134
## 16   17   1    153   34.8  70   29 204 118 120   134
## 17   17   0    174   44.7  70   49 187 104 103   165
## 18   17   1    176   60.1  92   29 188 129 130   120
## 19   17   0    171   42.6  69   38 172 130 103   130
## 20   19   1    156   37.2  72   21 216 119  81    85
## 21   19   0    174   54.6  86   37 184 118 101    85
## 22   20   0    178   64.0  86   34 225 148 135   160
## 23   23   0    180   73.8  97   57 171 108  98   165
## 24   23   0    175   51.1  71   33 224 131 113    95
## 25   23   0    179   71.5  95   52 225 127 101   195
```

1.4

```r
head(cystfibr[order(cystfibr$age, decreasing = T),], n = 5L)
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 23  23   0    180   73.8  97   57 171 108  98   165
## 24  23   0    175   51.1  71   33 224 131 113    95
## 25  23   0    179   71.5  95   52 225 127 101   195
## 22  20   0    178   64.0  86   34 225 148 135   160
## 20  19   1    156   37.2  72   21 216 119  81    85
```

1.5

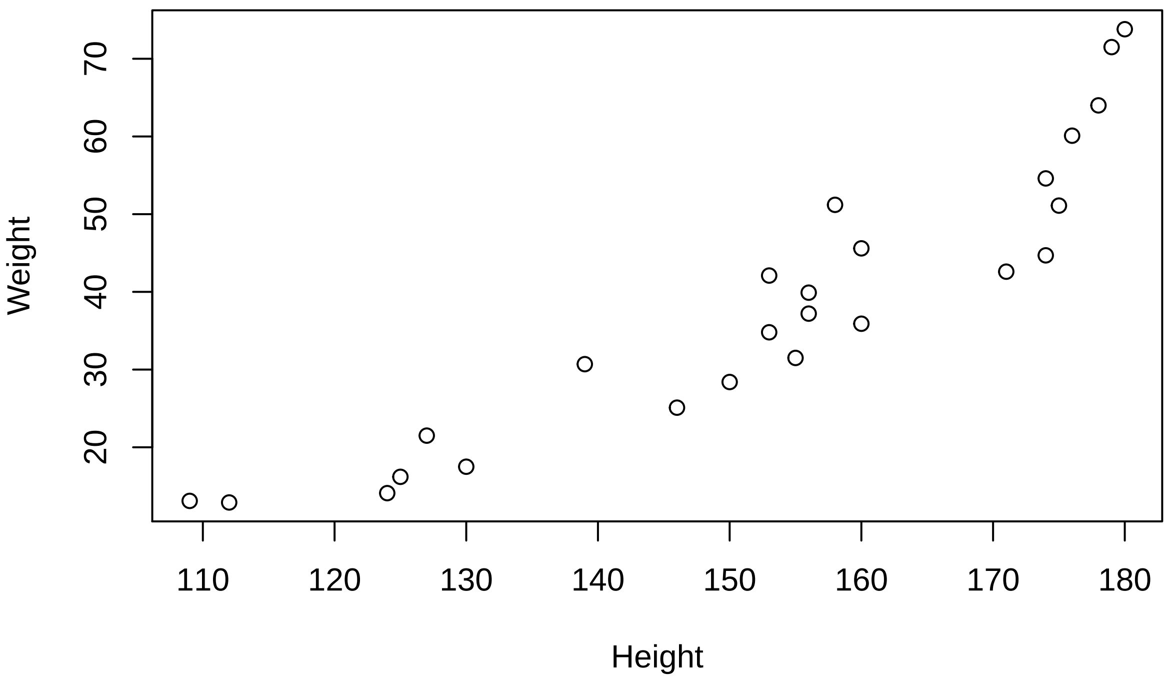
```r
subset(cystfibr, cystfibr$age > mean(cystfibr$age) & cystfibr$height > mean(cystfibr$height))
```

```
##    age sex height weight bmp fev1  rv frc tlc pemax
## 14  15   1    158   51.2  93   45 158 124  90    80
## 15  16   1    160   35.9  66   31 302 133 101   134
## 16  17   1    153   34.8  70   29 204 118 120   134
## 17  17   0    174   44.7  70   49 187 104 103   165
## 18  17   1    176   60.1  92   29 188 129 130   120
## 19  17   0    171   42.6  69   38 172 130 103   130
## 20  19   1    156   37.2  72   21 216 119  81    85
## 21  19   0    174   54.6  86   37 184 118 101    85
## 22  20   0    178   64.0  86   34 225 148 135   160
## 23  23   0    180   73.8  97   57 171 108  98   165
## 24  23   0    175   51.1  71   33 224 131 113    95
## 25  23   0    179   71.5  95   52 225 127 101   195
```

1.6

I plotted height versus weight because I assumed that there might be some correlation. Indeed, the two variables seem to have a positive correlation in that and increase in one corresponds to an increase in the other.

```r
plot(cystfibr$height, cystfibr$weight,xlab = 'Height', ylab = 'Weight')
```



Problem 2

I created a loop that adds 1 to a counter variable for every occurrence of a male with igf1 greater than 400.

```r
# The count variable is initialized with the value of 0
count <- 0
# This iterates over each record in the sex column
for (i in 1:length(juul_clean$sex)) {
  # This condition checksthat the sex is male
  if (juul_clean[i,'sex'] == 1) {
    # This condition checks that the igf1 variable is
    # greater than 400
    if (juul_clean[i,"igf1"] > 400){
      # The counter variable is only increased if both
      # condition are met
        count = count + 1
    }
  }
}
print(paste('The number of males with an insulin-like growth factor greater than 400 is:', count))
```

```
## [1] "The number of males with an insulin-like growth factor greater than 400 is: 144"
```

I used the 'ifelse' function to perform a vectorized conditional operation on the data set. This operation returns a one if both conditions are met and a zero otherwise. Finally, the resulting vector is summed.

```r
sum(ifelse(juul_clean$sex == 1 & juul_clean$igf1 > 400, 1, 0))
```

```
## [1] 144
```

To count the desired values using subsetting I first used direct indexing using conditional statements.

```r
length(juul_clean[juul_clean$sex == 1 & juul_clean$igf1 > 400,'sex'])
```

```
## [1] 144
```

Next, I used the subset function.

```r
length(subset(juul_clean[juul_clean$sex == 1,], igf1 > 400, sex) == 1)
```

```
## [1] 144
```

All queries resulted in similar values.

Problem 3

3.3

a)

The response variable is happiness and the explanatory variable is income.

b)

```r
nH <- c(21/360, 96/850, 143/604, 260/1814)
H <- c(213/360, 506/850, 347/604, 1006/1814)
vH <- c(126/360, 248/850, 114/604, 488/1814)
g_matrix <- matrix(c(nH, H, vH), nrow = 4, ncol = 3)
row.names(g_matrix) <- c('Above Average', 'Average', 'Below Average', 'Total')
colnames(g_matrix) <- c('Not Too Happy', 'Pretty Happy', 'Very Happy')
g_matrix
```

```
##               Not Too Happy Pretty Happy Very Happy
## Above Average    0.05833333    0.5916667  0.3500000
## Average          0.11294118    0.5952941  0.2917647
## Below Average    0.23675497    0.5745033  0.1887417
## Total            0.14332966    0.5545755  0.2690187
```

c)

Looking the the value in the 'Total' row of the 'Very Happy' column in the conditional proportions table above, the total proportion of individuals that report being very happy is about 27%.

3.61

a)

The response variable is assessed value and the explanatory variable is square feet.

b)

The response variable is party preference and the explanatory variable is gender.

c)

The response variable is annual income and the explanatory variable is number of years of education.

d)

The response variable is the number of pounds lost on a diet and the explanatory variable is the type of diet.

3.63

a)

```
lad_yes <- c(621/808, 834/979)
lad_no <- c(187/808, 145/979)
g_matrix <- matrix(c(lad_yes, lad_no), nrow =2, ncol = 2)
row.names(g_matrix) <- c('Male', 'Female')
colnames(g_matrix) <- c('Yes', 'No')
g_matrix
```

```
##              Yes        No
## Male   0.7685644 0.2314356
## Female 0.8518897 0.1481103
```

b)

Looking at the conditional proportion table above, Females are more likely to report believing in a life after death by 8 percentage points. This value is the difference of proportions and is calculated by subtracting the proportion of male/yes responses from the female/yes responses (0.85-0.77=0.08). The ratio of proportions would the the ratio of these two values (0.85/0.77=1.1). These values indicate that there is not a large difference between the responses from males and females. Looking at the difference of proportion, 8 percentage points is not a very sizable increase, but more tellingly, the ratio indicates that the two values are almost identical since it is very close to one.

3.14

a)

The most strongly correlated variables have a correlation close to one or negative one. The value of -0.07 is very close to zero, therefore, 'political ideology' and 'times a week reading a newspaper' are weakly correlated. The correlation value can be thought of as the slop of a line-of-best-fit for the scatter plot of the two variables. A correlation of zero would indicate a horizontal line and an increase in one variable would not translate to and increase in the other. Conversely, a correlation of one would indicate that an increase in one variable would indicate a proportionate increase in the other.

b)

'Religiosity' has a stronger correlation to 'political ideology' because 0.58 is closer to one than -0.07. Using the reasoning outlined above, the slope of the linear correlation for 'religiosity' and 'political ideology' would be steeper than the one for 'times a week reading a newspaper' and 'political ideology'.

3.16

1. c
2. a
3. d
4. b