Ngoc Ha

ST 557 - HW 1

In [1]:

```
library(readr)
```

Problem 1

(1a,b)

In [2]:

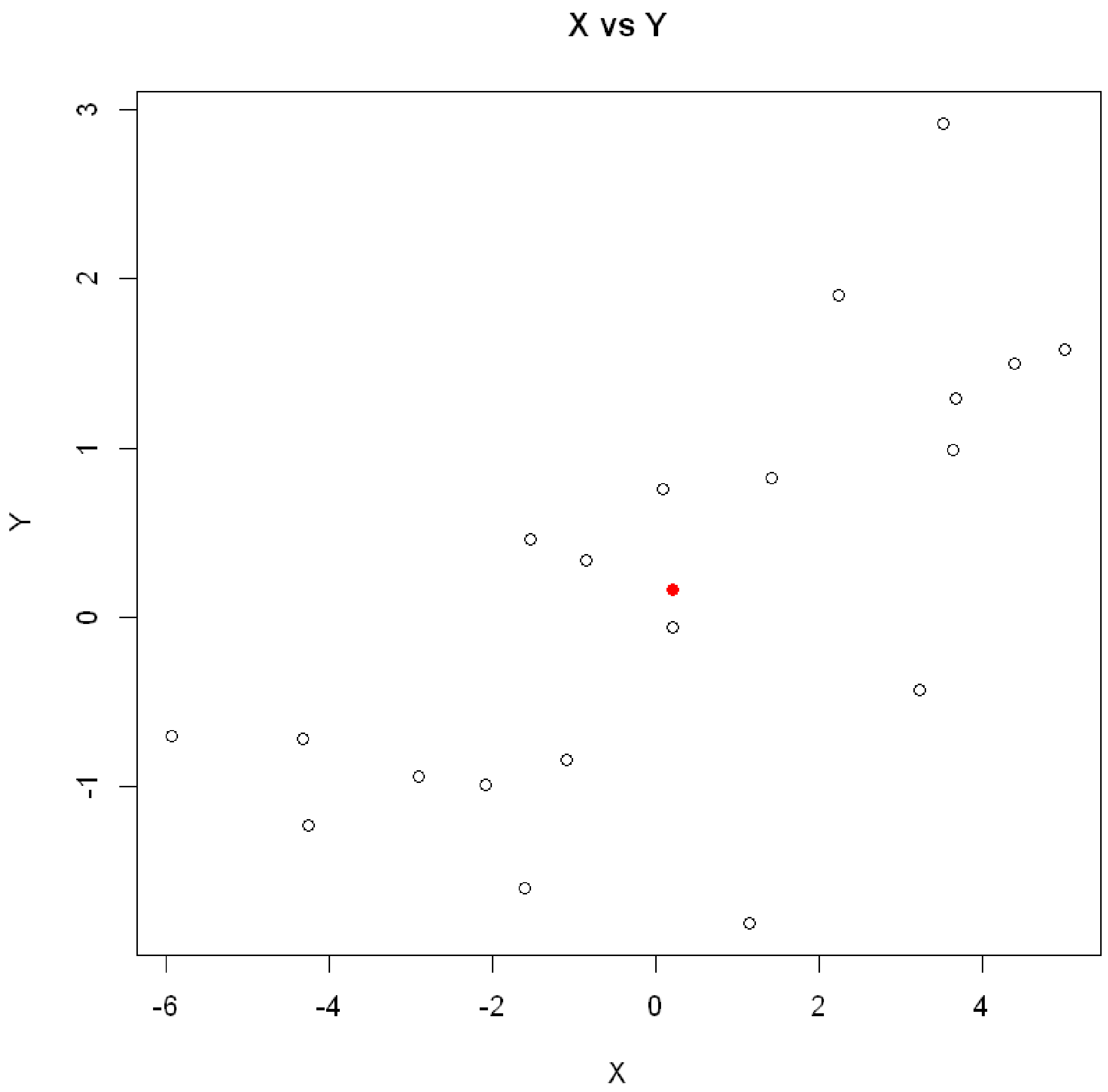
```
data1 <- read_csv('HW1-1.csv', col_types = cols(
  X = col_double(),
  Y = col_double()
))
head(data1)
```

X	Y
-1.54	0.46
-4.25	-1.23
-0.85	0.34
-2.90	-0.94
-1.09	-0.84
-5.92	-0.70

In [3]:

```
sampMean1 = c(mean(data1$X), mean(data1$Y))
```

In [4]:
```
plot(data1$X, data1$Y, main = 'X vs Y', xlab = 'X', ylab = 'Y')
points(sampMean1[1], sampMean1[2], pch = 16, col =2)
```



(1c) Covariance matrix

In [5]:
```
covar1 <- var(data1)
covar1
```

	X	Y
X	10.140227	2.852078
Y	2.852078	1.668133

(1d) Eigendecomposition

In [6]:

```
eigendecomp1 <- eigen(covar1)
eigendecomp1
```

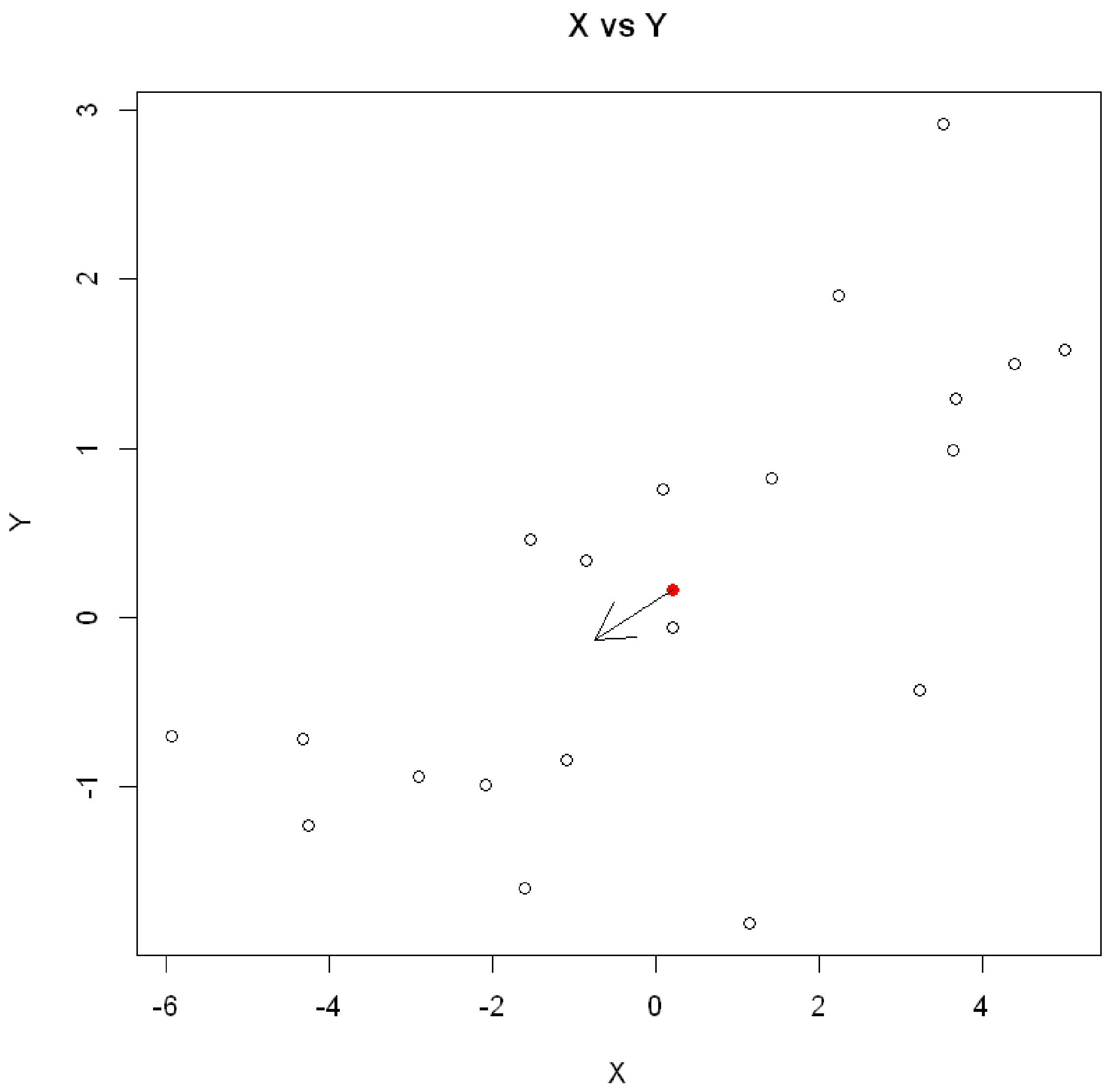
```
eigen() decomposition
$values
[1] 11.0108859  0.7974741

$vectors
           [,1]       [,2]
[1,] -0.9564274  0.2919702
[2,] -0.2919702 -0.9564274
```

(1e)

In [7]:

```
plot(data1$X, data1$Y, main = 'X vs Y', xlab = 'X', ylab = 'Y')
points(sampMean1[1], sampMean1[2], pch = 16, col =2)
arrows(sampMean1[1], sampMean1[2], sampMean1[1]+eigendecomp1$vectors[,1][1], sampMean1[2]+eigendecomp1$vectors[,1][2])
```



The eigenvector with the largest eigenvalue is the direction along which the data set has the maximum variance.

Problem 2

(2a,b)

In [8]:

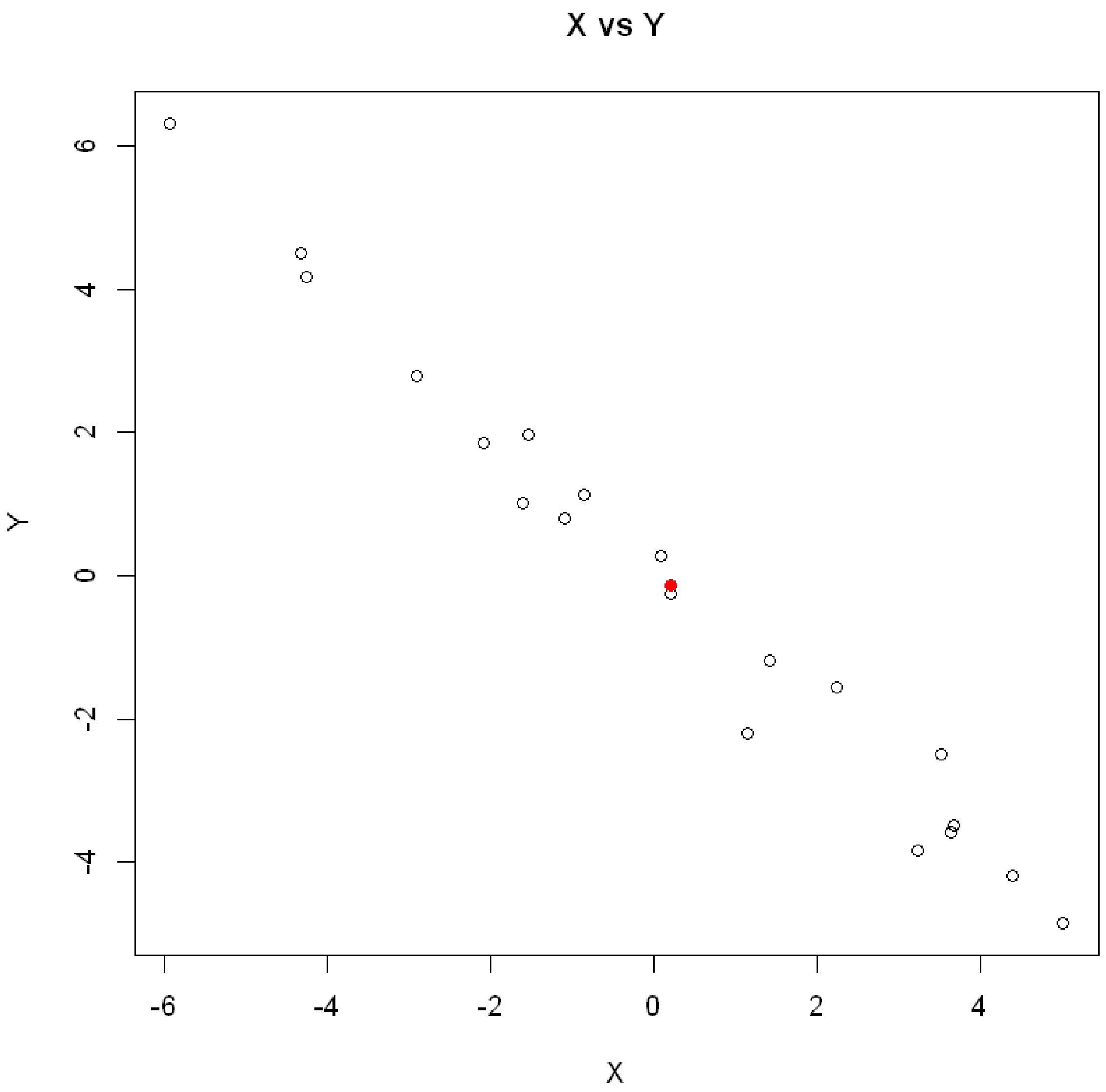
```
data2 <- read_csv('HW1-2.csv', col_types = cols(
  X = col_double(),
  Y = col_double()
))
head(data2)
```

X	Y
-1.54	1.96
-4.25	4.17
-0.85	1.13
-2.90	2.79
-1.09	0.80
-5.92	6.31

In [9]:

```
sampMean2 = c(mean(data2$X), mean(data2$Y))
```

In [10]:
```r
plot(data2$X, data2$Y, main = 'X vs Y', xlab = 'X', ylab = 'Y')
points(sampMean2[1], sampMean2[2], pch = 16, col =2)
```



(2c) Covariance matrix

In [11]:
```r
covar2 <- var(data2)
covar2
```

	X	Y
X	10.140227	-9.983968
Y	-9.983968	10.046978

(2d) Eigendecomposition

In [12]:

```
eigendecomp2 <- eigen(covar2)
eigendecomp2
```

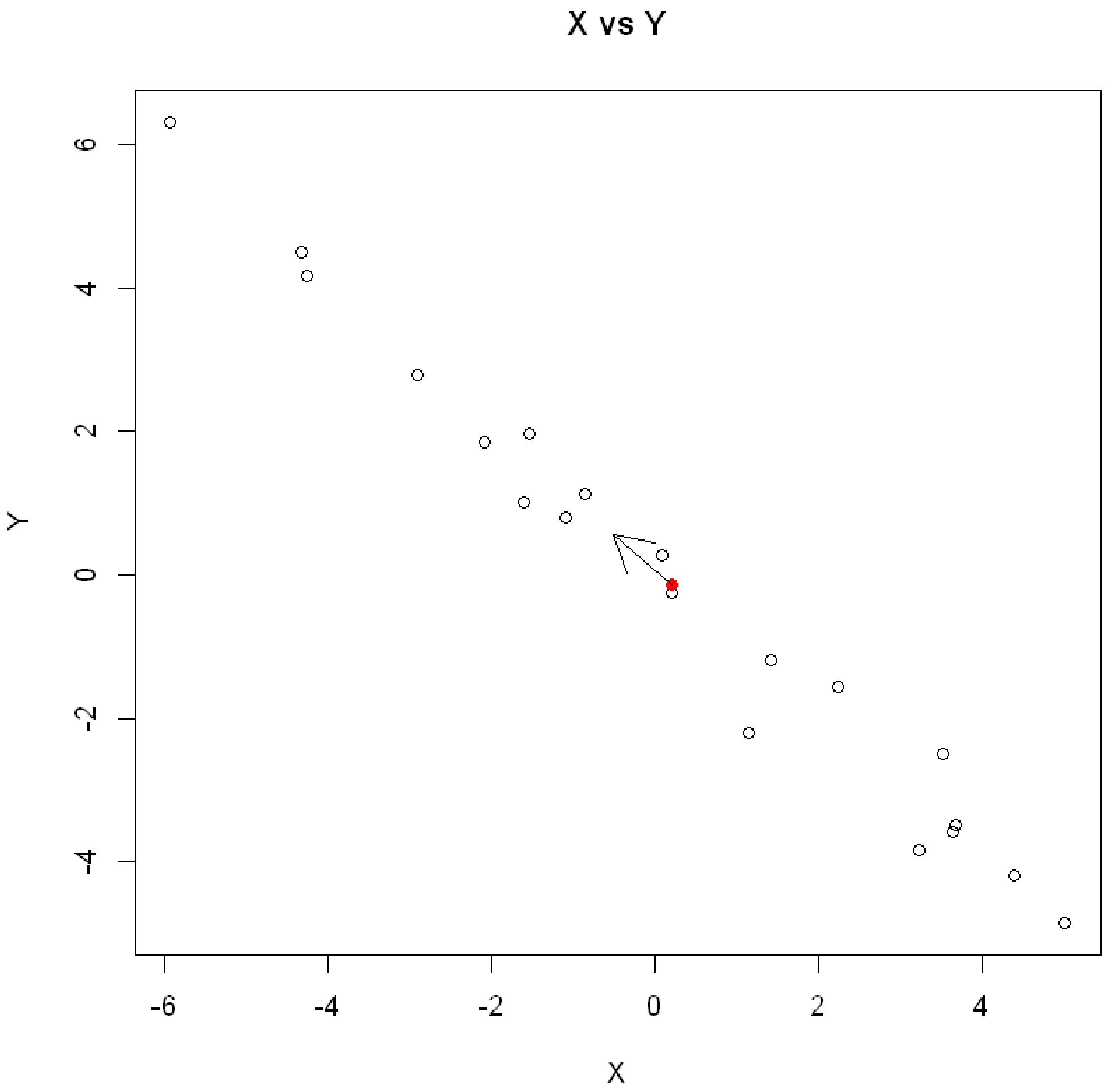
```
eigen() decomposition
$values
[1] 20.0776798  0.1095252

$vectors
           [,1]       [,2]
[1,] -0.7087559 -0.7054538
[2,]  0.7054538 -0.7087559
```

(2e)

In [13]:

```
plot(data2$X, data2$Y, main = 'X vs Y', xlab = 'X', ylab = 'Y')
points(sampMean2[1], sampMean2[2], pch = 16, col =2)
arrows(sampMean2[1], sampMean2[2], sampMean2[1] + eigendecomp2$vectors[,1][1],
sampMean2[2] + eigendecomp2$vectors[,1][2])
```



The eigenvector with the largest eigenvalue is the direction along which the data set has the maximum variance.

Problem 3

In [14]:

```
A = cbind(c(5.125, 3.875, 2.125, -1.125, 0.000), c(3.875, 5.125, -1.125, 2.125
, 0.000), c(2.125, -1.125, 5.125, 3.875, 0.000), c(-1.125, 2.125, 3.875, 5.125
, 0.000), c(0.000, 0.000, 0.000, 0.000, -3.000))
print(A)
```

```
       [,1]   [,2]   [,3]   [,4] [,5]
[1,]  5.125  3.875  2.125 -1.125    0
[2,]  3.875  5.125 -1.125  2.125    0
[3,]  2.125 -1.125  5.125  3.875    0
[4,] -1.125  2.125  3.875  5.125    0
[5,]  0.000  0.000  0.000  0.000   -3
```

(a)

In [15]:

```
eigen(A)
```

```
eigen() decomposition
$values
[1] 10.0  8.0  4.5 -2.0 -3.0

$vectors
     [,1] [,2] [,3] [,4] [,5]
[1,]  0.5 -0.5 -0.5  0.5    0
[2,]  0.5 -0.5  0.5 -0.5    0
[3,]  0.5  0.5 -0.5 -0.5    0
[4,]  0.5  0.5  0.5  0.5    0
[5,]  0.0  0.0  0.0  0.0    1
```

(b)

A is *not* positive definite.

Since $v^TAv = \lambda v^Tv = \lambda||v||$ and $||v|| > 0$, we have:

$$v^TAv < 0 \iff \lambda < 0 \iff \lambda \in \{-2.0, -3.0\}$$

In [16]:

```
x <- c(0.5, -0.5, -0.5, 0.5, 0.0) # eigenvector with corresponding eigenvalue
 -2.0
t(x)%*%A%*%x
```

```
-2
```

(c)

$x = 4v_1 + 2v_5$, where v_1 is the eigenvector corresponding to the largest eigenvalue λ_1 of A, and v_5 is the eigenvector corresponding to the smallest eigenvalue λ_5 of A.

$$Ax = A(4v_1 + 2v_5) = 4Av_1 + 2Av_5 = 4\lambda_1 v_1 + 2\lambda_5 v_5$$

Problem 4

```
In [17]: iris = read.csv('IrisData.csv')
         head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	1
4.9	3.0	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5.0	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1

(4a) sample mean vector

```
In [18]: sapply(iris[,1:4], mean)
```

Sepal.Length	5.84333333333333
Sepal.Width	3.05733333333333
Petal.Length	3.758
Petal.Width	1.19933333333333

(4b) sample mean vector for each vector

Species 1

```
In [19]: meanVec1 <- sapply(iris[iris$Species == 1,1:4], mean)
         print(meanVec1)
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
       5.006        3.428        1.462        0.246 
```

Species 2

In [20]:

```
meanVec2 <- sapply(iris[iris$Species == 2,1:4], mean)
print(meanVec2)
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
       5.936        2.770        4.260        1.326 
```

Species 3

In [21]:

```
meanVec3 <- sapply(iris[iris$Species == 3,1:4], mean)
print(meanVec3)
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
       6.588        2.974        5.552        2.026 
```

(4c) sample correlation matrix

In [22]:

```
cor(iris[,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Petal Length and *Petal Width* are most highly correlated.

(4d) individual sample correlation matrix

Species 1

In [23]:

```
cor(iris[iris$Species == 1,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	0.7425467	0.2671758	0.2780984
Sepal.Width	0.7425467	1.0000000	0.1777000	0.2327520
Petal.Length	0.2671758	0.1777000	1.0000000	0.3316300
Petal.Width	0.2780984	0.2327520	0.3316300	1.0000000

Species 2

```
In [24]: cor(iris[iris$Species == 2,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	0.5259107	0.7540490	0.5464611
Sepal.Width	0.5259107	1.0000000	0.5605221	0.6639987
Petal.Length	0.7540490	0.5605221	1.0000000	0.7866681
Petal.Width	0.5464611	0.6639987	0.7866681	1.0000000

Species 3

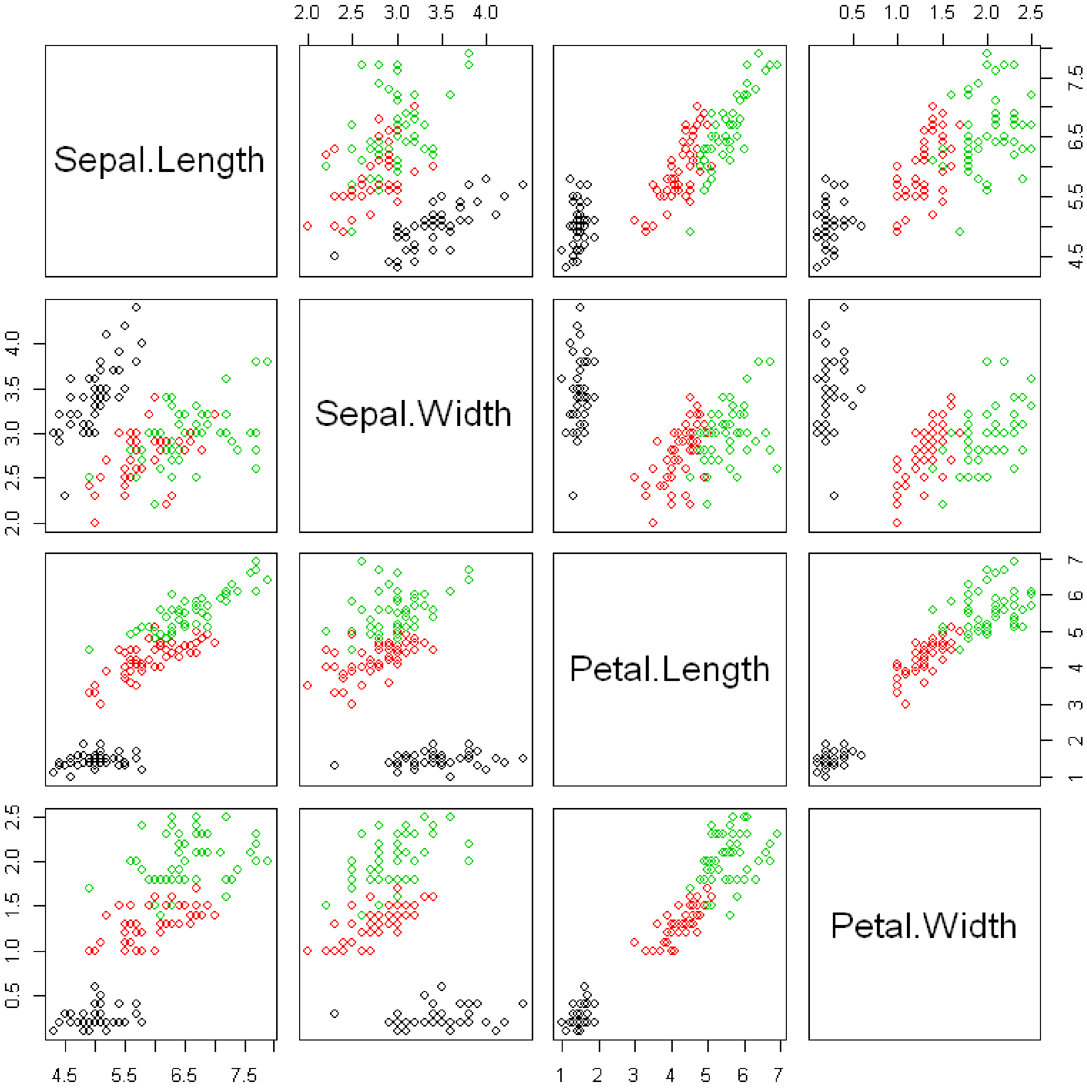
```
In [25]: cor(iris[iris$Species == 3,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	0.4572278	0.8642247	0.2811077
Sepal.Width	0.4572278	1.0000000	0.4010446	0.5377280
Petal.Length	0.8642247	0.4010446	1.0000000	0.3221082
Petal.Width	0.2811077	0.5377280	0.3221082	1.0000000

Petal Length and *Petal Width* are no longer the most highly correlated pair of variables across all species.

(4e)

```
In [26]: pairs(iris[,1:4], col = iris$Species)
```



There are distinct groups of flowers.

I can completely predict Species 1 just by using Petal width as they have distinctly small petal widths; Species 2 & 3 would be a little more tricky if petal width is around 1.5.

I would choose **petal width** and **petal length**, as the groups seem most "separated" when plotting with petal width and length.

Problem 5

Let A be any $n \times p$ matrix, for arbitrary dimensions n and p, and let B be the product matrix $B = A^T A$

(5a) Show that B is symmetric

$B^T = A^T(A^T)^T = A^TA = B \iff B$ is symmetric

(5b) Show that B is postitive semi-definite: $x^TBx \geq 0 \, \forall x \in \mathbb{R}^p$

$x^TBx = x^TA^TAx = (Ax)^TAx = ||Ax||^2 >= 0 \, \forall x \in \mathbb{R}^p$

(5c) $S = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$. Argue that S is positive-definite.

The sample covariance matrix S has the form A^TA (scaled by a constant factor), so it is positive semi-definite.

In [ ]: