Multivariate Data Analysis
Prof. Dr. Christina Andersson

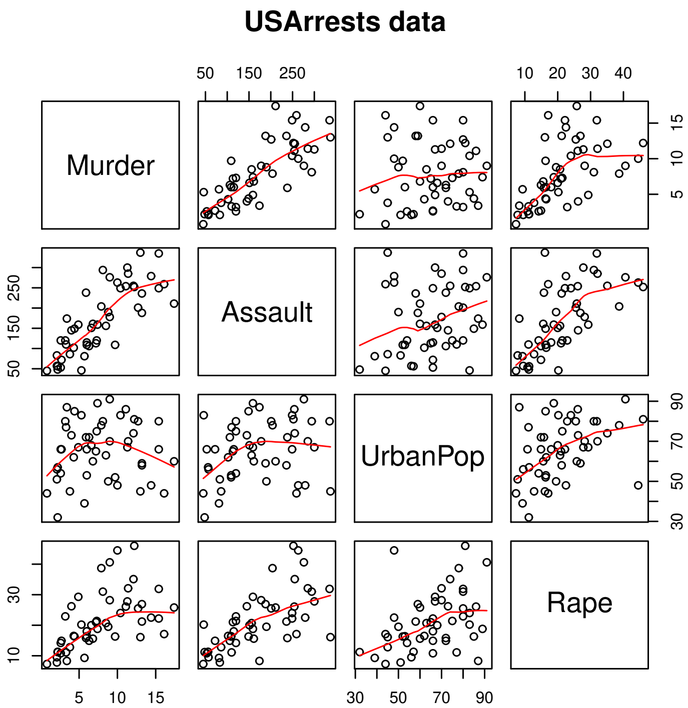
R Solution Exercise Sheet 1: Principal Component Analysis

Computer Problems:

1. (a) head(USArrests)

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

(b)
```
require(graphics)
pairs(USArrests, panel = panel.smooth, main = "USArrests data")
```



(c)
```
> pca1 <- prcomp(USArrests,center = TRUE,scale. = TRUE)
```

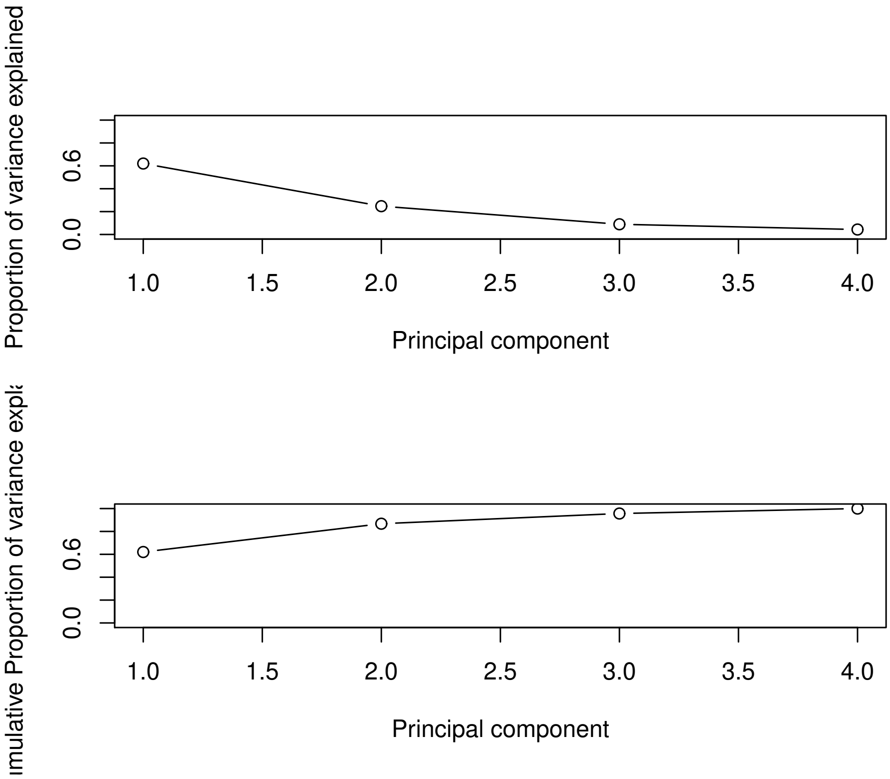
```
> pca1
Standard deviations:
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation:
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

(d) 4, since we have 4 original variables.

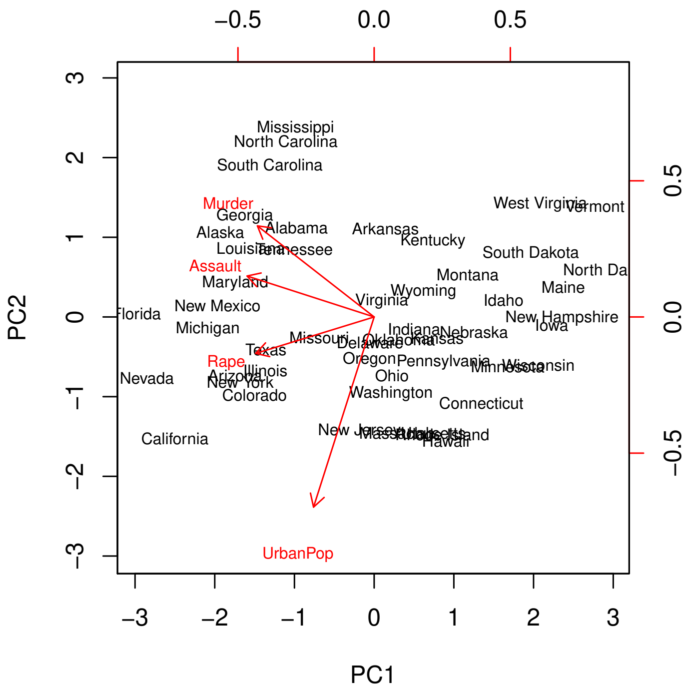
(e)
```
> summary(pca1)
Importance of components:
                                              PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

From the the summary, we can undersand PC1 explains 62

(f)
```
pcaCharts <- function(x) {
    x.var <- x$sdev ^ 2
    x.pvar <- x.var/sum(x.var)
    print("proportions of variance:")
    print(x.pvar)

    par(mfrow=c(2,1))
    plot(x.pvar,xlab="Principal component",
ylab="Proportion of variance explained", ylim=c(0,1), type="b")
    plot(cumsum(x.pvar),xlab="Principal component",
 ylab="Cumulative Proportion of variance explained", ylim=c(0,1),
 type="b")
    par(mfrow=c(1,1))
}

pcaCharts(pca1)
```



(g) `> biplot(pca1,scale=0, cex=.7)`



For each of 50 stats in the USA, the data set contains the number of arrests per 100,000 residents for each three crimes: Assault, Murder and Rape. Also urbanpop represents percent of the population in each state living in urban areas. The plot shows the first two principal component scores and the loading verctors in a singple biplot display.

(h) >pca1

```
Standard deviations:
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation:
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

From the plot as wells from the above loadings what we can understand is, first loading vector places approximately equal weight on Assault, Murder and Rape, with much less weight on urbanpop. Hence this component roughly corresponds to a measure of overall rates of serious crimes.

The second loading vector places most of it weight on Urbanpop and much less weight on the other 3 features. Hence, this component roughly corresponds to the level of urbanization of the state. Overall, we see that the crime-related varaibales are located close to each other, and that the urbanpop variable is far from other three. This indicates hat the crime related variables are correlated with each other-States with high murder rates tend to had high assault and rape rates. Urabnpop variable is less correlated with the other three.

2. (a) Iris is a data frame with 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

(b) Iris is a data frame with 150 cases (rows).

```
> dim(iris)
[1] 150   5
```

(c)
```
require(graphics)
pairs(iris, panel = panel.smooth, main = "Iris data")
```

(d)
```
> pca1 <- prcomp(~Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width, data=iris,center = TRUE,scale. = TRUE)
> pca1
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
```

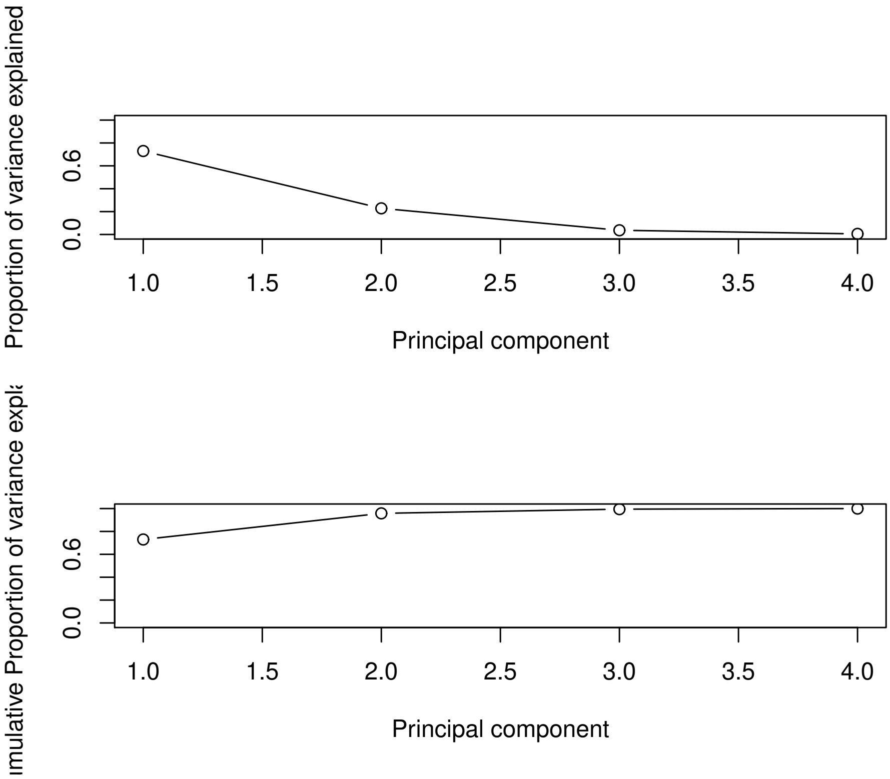
```
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

(e) 4, since we have 4 original variables.

(f)
```
> summary(pca1)
Importance of components:
                                   PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

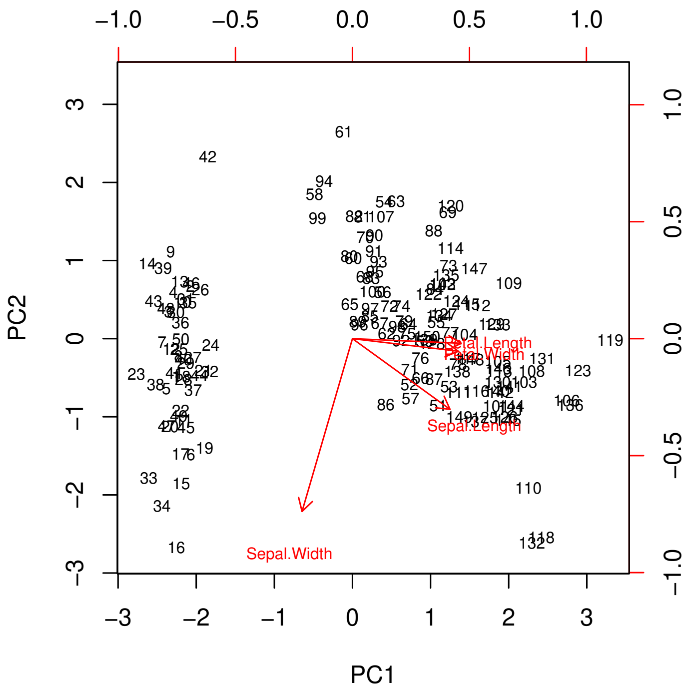
PC1 and PC2 cover 95% of variability in the data.

(g)
```
pcaCharts <- function(x) {
    x.var <- x$sdev ^ 2
    x.pvar <- x.var/sum(x.var)
    print("proportions of variance:")
    print(x.pvar)

    par(mfrow=c(2,1))
    plot(x.pvar,xlab="Principal component",
ylab="Proportion of variance explained", ylim=c(0,1), type="b")
    plot(cumsum(x.pvar),xlab="Principal component",
 ylab="Cumulative Proportion of variance explained", ylim=c(0,1),
 type="b")
    par(mfrow=c(1,1))
}

pcaCharts(pca1)
```



PCA charts also confirms result above (Look for elbow shape).

(h) `> biplot(pca1,scale=0, cex=.7)`




```
>pca1
> pca1
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:
                   PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```