HW #03

Ran Zhang

1/30/2019

Question 1a

Load the dataset as data1

```
data1 <- USJudgeRatings
```

Look at the data

```
head(data1,2)
```

```
##                CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## AARONSON,L.H.   5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
## ALEXANDER,J.M.  6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
```

```
dim(data1)
```

```
## [1] 43 12
```

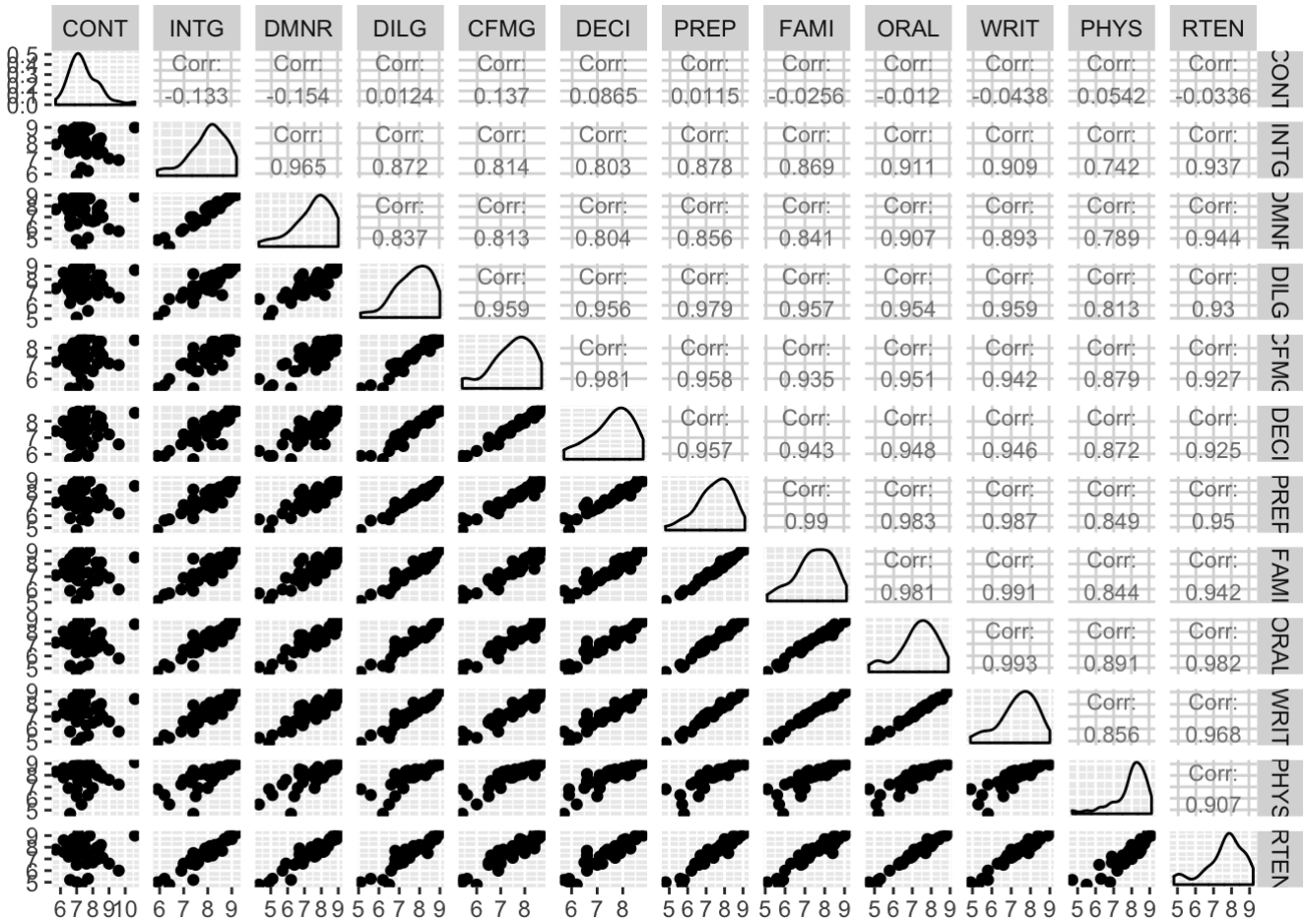
Load the library of ggplot2 and GGally

```
library(ggplot2)
library(GGally)
library(corrplot)
```

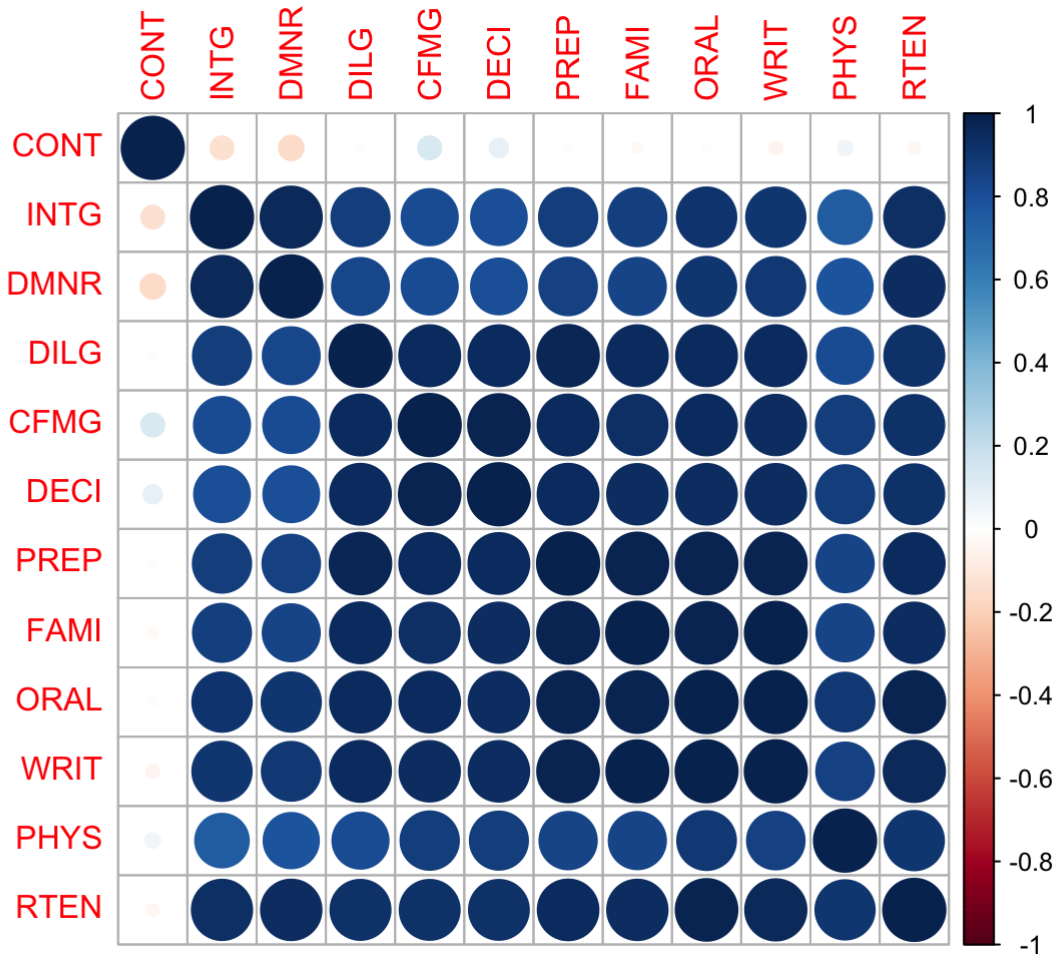
```
## corrplot 0.84 loaded
```

Examine the data by pairs-plot

```
ggpairs(data1[,1:12],upper = list(continuous = wrap("cor", size = 3)))
```




```
corrplot(cor(data1[,1:12]))
```



Comment: All these variables are correlated to each other except the variable “Cont”.

Question 1b

Check the sd of each of variable

```
apply(data1, 2, sd)
```

```
##      CONT      INTG      DMNR      DILG      CFMG      DECI      PREP
## 0.9408768 0.7701447 1.1437054 0.9008978 0.8601102 0.8029362 0.9533702
##      FAMI      ORAL      WRIT      PHYS      RTEN
## 0.9489868 1.0100437 0.9611328 0.9395753 1.1009711
```

Standardize the variables

```
std.data1 <- scale(data1)
apply(std.data1,2,sd)
```

```
## CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
##    1    1    1    1    1    1    1    1    1    1    1    1
```

Perform the PCA

```
data1.pca <- prcomp(std.data1)
summary(data1.pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     3.1833 1.05078 0.57698 0.50383 0.29061 0.19310
## Proportion of Variance 0.8445 0.09201 0.02774 0.02115 0.00704 0.00311
## Cumulative Proportion  0.8445 0.93647 0.96421 0.98537 0.99240 0.99551
##                           PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.14030 0.12416 0.08851 0.07491 0.05708 0.04539
## Proportion of Variance 0.00164 0.00128 0.00065 0.00047 0.00027 0.00017
## Cumulative Proportion  0.99715 0.99844 0.99909 0.99956 0.99983 1.00000
```

The loadings for the first two PCs which covered 90% more

```
round(data1.pca$rotation[,1:2],3)
```

```
##         PC1    PC2
## CONT  0.003 -0.933
## INTG -0.289  0.182
## DMNR -0.287  0.198
## DILG -0.304 -0.036
## CFMG -0.303 -0.168
## DECI -0.302 -0.128
## PREP -0.309 -0.032
## FAMI -0.307  0.001
## ORAL -0.313  0.004
## WRIT -0.311  0.031
## PHYS -0.281 -0.089
## RTEN -0.310  0.039
```

Question 1c

For PC1, is roughly the averarage of 13 variables except the first variable; For PC2, represents the difference between the X1, X4, X5, X6, X7, X11 and the rest of the variables. The result match the result what I find in part (a). As the the first variable X1 has nearly no correlation with other variables, and all other variables have a relative tight same correlation with each other.

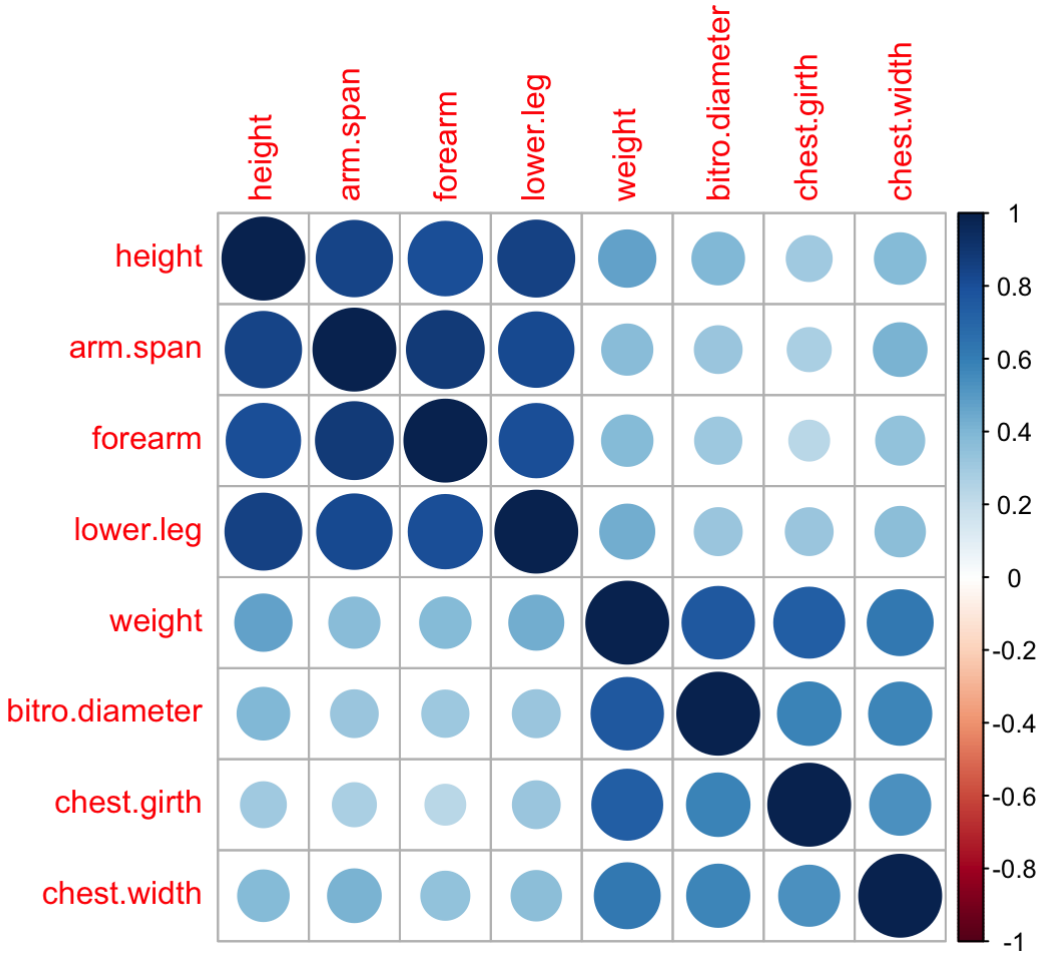
Question 2a

Load the Harmo23.cor data

```
data2 <- Harman23.cor$cov
```

Visualized the matrix

```
library(corrplot)
corrplot(data2)
```



Comment: There are different seperate parts (two mainly groups) of the realtionship of the variables. One group covers the height, arm.span, forearm and lower.leg, these variables have relative strong correlations to each other; and rest of four variables have kindly low correlations to each other.

Question 2b

Compute the eigenvector

```
eig.out2 <- eigen(data2)
str(eig.out2)
```

```
## List of 2
##  $ values : num [1:8] 4.673 1.771 0.481 0.421 0.233 ...
##  $ vectors: num [1:8, 1:8] -0.398 -0.389 -0.376 -0.388 -0.351 ...
##  - attr(*, "class")= chr "eigen"
```

Compute the eigenvalues

```
lam2 <- eig.out2$values
```

Calculate the standard deviation, proportion and cumulative proportion

```
tab2 <- rbind(lam2, lam2/sum(lam2), cumsum(lam2)/sum(lam2))
rownames(tab2) <- c("Standard deivation", "Proportion of variance", "Cumulative proporti
on")
```

Results using eigen

```
round(tab2, 4)
```

```
##                           [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
## Standard deivation     4.6729 1.7710 0.4810 0.4214 0.2332 0.1867 0.1373
## Proportion of variance 0.5841 0.2214 0.0601 0.0527 0.0292 0.0233 0.0172
## Cumulative proportion  0.5841 0.8055 0.8656 0.9183 0.9474 0.9708 0.9879
##                           [,8]
## Standard deivation     0.0965
## Proportion of variance 0.0121
## Cumulative proportion  1.0000
```

Extract the first two PCs

```
eig.out2$vectors[,1:2]
```

```
##            [,1]       [,2]
## [1,] -0.3975776 -0.2797405
## [2,] -0.3893198 -0.3314202
## [3,] -0.3761601 -0.3446045
## [4,] -0.3883899 -0.2970667
## [5,] -0.3506669  0.3942422
## [6,] -0.3119078  0.4007179
## [7,] -0.2855270  0.4359188
## [8,] -0.3102250  0.3144488
```

Question 2c

The PC1 captures the variance of 0.5842 (58.42%) of total variance, the PC2 captures 0.2214 (22.14%) of total variance; These two PC1 and PC2 capture 0.8054 (80.54%) of total variance.

Question 2d

PC1 is roughly the average value of all variables and interpreted as the projection of X onto the direction a1. And PC1 captures the most variability in the original data by any linear combination,

PC2 is roughly the difference between first 4 groups and the last 4 groups and the direction such that it is orthogonal to the PC1 and it captures the most of the remianing variability while being uncorrelated to X1.

Question 3a

Read the data of the cancer dataset

```
data3 <- read.table("https://www.stat.ncsu.edu/people/maity/courses/st537-S2019/HW/cance
r.txt", header = TRUE)
```

Examine the standard deviation of each volume

```
data3_sd <- apply(data3[ ,2:11],2,sd)
data3_sd
```

```
## All.cancers        Lung       Colon    Melanoma    F.breast    Pancreas
##   47.536999   14.588088    7.210099    5.390848    8.008908    1.577709
##    Leukemia     Ovarian      Cervix    Prostate
##    2.373649    1.395516    1.510094   18.048835
```

Question 3b

As the different standard deviations we have, we have to standardzied the variables first then do the PCA.
Question 3c Standardize the all variables

```
std.dat3 <- scale(data3[,2:12], center=T, scale = T)
apply(std.dat3,2,sd)
```

```
## All.cancers        Lung       Colon    Melanoma    F.breast    Pancreas
##           1           1           1           1           1           1
##    Leukemia     Ovarian      Cervix    Prostate       Liver
##           1           1           1           1           1
```

Compute the principal components analysis

```
data.pca3 <- prcomp(std.dat3)
summary(data.pca3)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.8872 1.4110 1.2070 1.04549 0.92966 0.8028 0.69483
## Proportion of Variance 0.3238 0.1810 0.1324 0.09937 0.07857 0.0586 0.04389
## Cumulative Proportion  0.3238 0.5048 0.6372 0.73660 0.81517 0.8738 0.91765
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.59711 0.52480 0.48278 0.20192
## Proportion of Variance 0.03241 0.02504 0.02119 0.00371
## Cumulative Proportion  0.95007 0.97511 0.99629 1.00000
```

Question 4a

Load the library of HSAUR3

```
library(HSAUR3)
```

```
## Loading required package: tools
```

Snapshot of the data

```
head(heptathlon,2)
```

```
##                     hurdles highjump  shot run200m longjump javelin
## Joyner-Kersee (USA)   12.69     1.86 15.80   22.56     7.27   45.66
## John (GDR)            12.85     1.80 16.23   23.65     6.71   42.56
##                     run800m score
## Joyner-Kersee (USA)  128.51  7291
## John (GDR)           126.12  6897
```

Extract the first 7 colunms of the data

```
data4 <- heptathlon[,1:7]
```

Transform the latter 3 columns of the data

```
data4_sub <- data4[,c(1,4,7)]
newx <- apply(data4_sub,2,function(x) max(x)-x)
```

Rebuild the new matrix

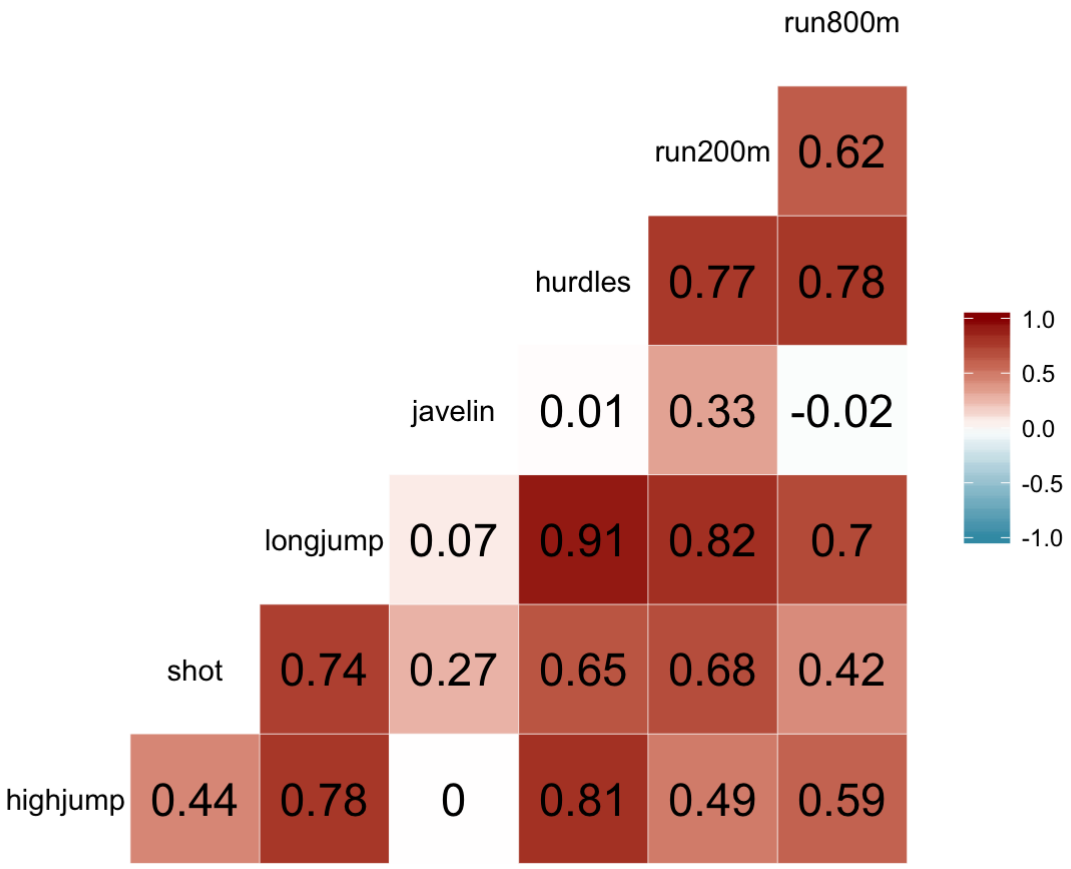
```
data4_new <- cbind(heptathlon[,-c(1,4,7,8)], newx)
```

Find the correlation matrix of the new matrix

```
cor(data4_new)
```

```
##             highjump      shot   longjump      javelin     hurdles
## highjump 1.000000000 0.4407861 0.78244227  0.002153016 0.811402536
## shot     0.440786140 1.0000000 0.74307300  0.268988837 0.651334688
## longjump 0.782442273 0.7430730 1.00000000  0.067108409 0.912133617
## javelin  0.002153016 0.2689888 0.06710841  1.000000000 0.007762549
## hurdles  0.811402536 0.6513347 0.91213362  0.007762549 1.000000000
## run200m  0.487663685 0.6826704 0.81720530  0.333042722 0.773720543
## run800m  0.591162823 0.4196196 0.69951116 -0.020049088 0.779257110
##            run200m     run800m
## highjump 0.4876637  0.59116282
## shot     0.6826704  0.41961957
## longjump 0.8172053  0.69951116
## javelin  0.3330427 -0.02004909
## hurdles  0.7737205  0.77925711
## run200m  1.0000000  0.61681006
## run800m  0.6168101  1.00000000
```

```
library(GGally)
ggcorr(data4_new, low = "#3B9AB2", mid="#FFFFFF", high="#990000", label = T, label_color
 = "black", label_size = 6, label_round = 2)
```



Comment: Almost of the variables have positive relationships to each other, except the 800m run and the javelin has a really low negative relationship. Javelin has really low realationship to other activities (variables). Jump (long/high) has a strong relationship to the run (200m/800m).

Question 4b

Standardize the new data

```
std.data4_new <- scale(data4_new)
apply(std.data4_new,2,sd)
```

```
## highjump     shot longjump  javelin  hurdles  run200m  run800m
##        1        1        1        1        1        1        1
```

Perform PCA

```
data4.pca <- prcomp(std.data4_new)
summary(data4.pca)
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.1119 1.0928 0.72181 0.67614 0.49524 0.27010
## Proportion of Variance 0.6372 0.1706 0.07443 0.06531 0.03504 0.01042
## Cumulative Proportion  0.6372 0.8078 0.88223 0.94754 0.98258 0.99300
##                            PC7
## Standard deviation     0.2214
## Proportion of Variance 0.0070
## Cumulative Proportion  1.0000
```

Explaination: The standard deviation of the first PC (PC1) is 2.1119, which gives the variance of that is 4.46. While the total variance is 7 and PC1 occupies around 4.5, which means the PC1 explains the 63.72% of the total variance. The second PC (PC2) is the most capture of the remaining variane except the PC1, the standard deviation of that is 1.0928 and the variance of that is 1.23, which explains the 17.06% of the total variance. And the direction of the PC2 is orthornol to the PC1. The cumulative of the PC1 and PC2 is 80.78% of the total variance, so the two components is good enough to explain the most of the variance. We can lower the dimension by using these two components to analyze the question here.

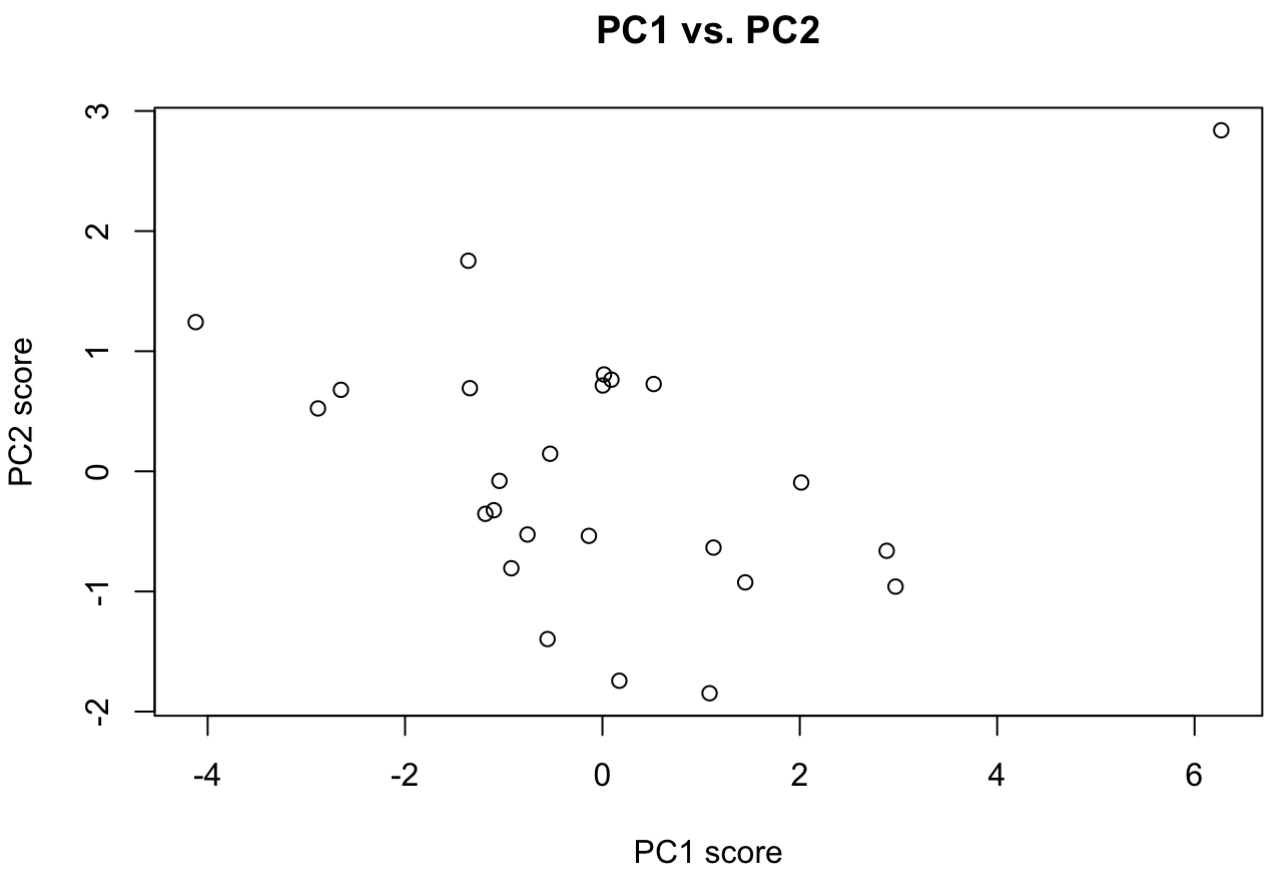
Question 4c

Extract the first two PC1 score and PC2 score of the data

```
data4_PC1 <- data4.pca$x[,1]
data4_PC2 <- data4.pca$x[,2]
```

Plot of pc scores

```
plot(x=data4_PC1,y=data4_PC2,xlab="PC1 score",ylab="PC2 score", main="PC1 vs. PC2")
```



Answer: There is a strong positive correlation pattern, but there is also a point as an outlier of the pattern.

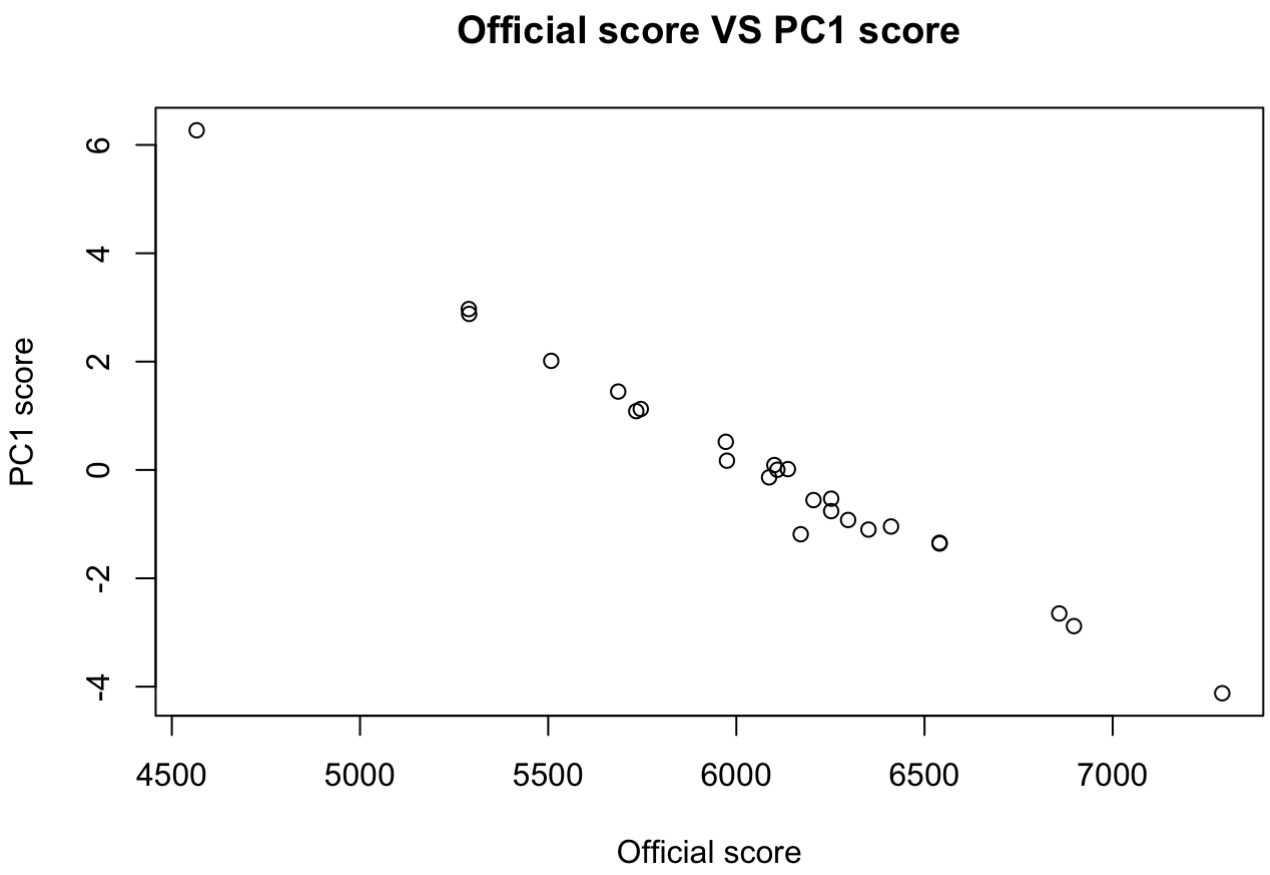
Question 4d

Extract the score colume of the data

```
data4_score <- heptathlon[,8]
```

Plot the official scores versus PC1 scores

```
plot(x=data4_score,y=data4_PC1,xlab = "Official score", ylab = "PC1 score", main="Offici
al score VS PC1 score")
```



Explaination: Yes, as the PC1 score and the official score has a linear relationship, they are aligned with each other.