MATH 4310 **Name:** Solution

Homework 10

Problem. We consider a data set extracted from the 2017-2018 US National Health and Nutrition Examination Survey consisting of 230 participants aged between 20-25 years. For each participant, data were collected about body measures to estimate the prevalence of overweight and obesity. Data in stored in bodydata.csv

a) Draw a scatterplot of the data in pairs using the R command pairs as in the notes.
b) Compute the correlation of the data matrix. Note: some data row have missing data. You will need to use this version of the command: cor(bodydata,use = "complete.obs")
c) Plot the correlation matrix using numbers as well as circles to display the size of correlation coefficients.
d) Apply hierarchical clustering as in the lectures on the correlation plot using 2 clusters.
e) Compute the p-values on the correlation matrix.
f) Analyze the results: which variables are strongly correlated (correlation coefficient > 0.7) to each other? Which variables are not statistically correlated (use alpha = 0.05)?

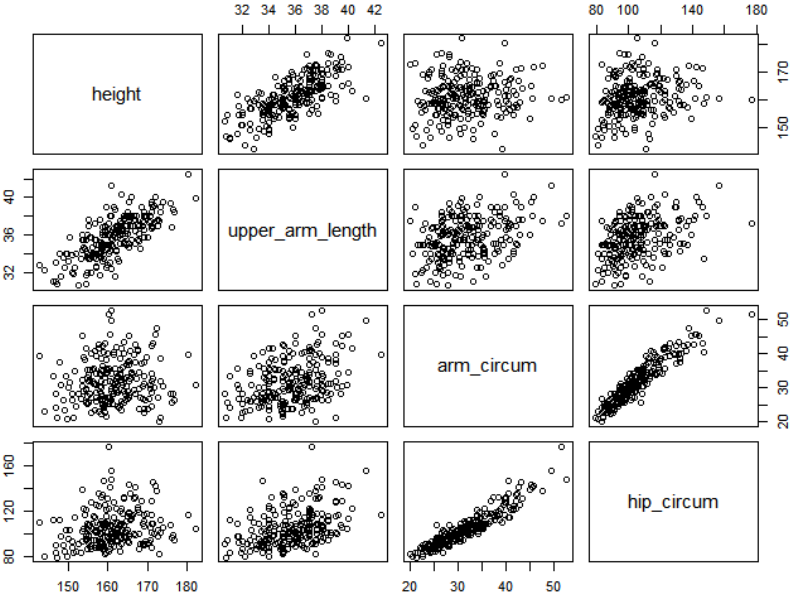
SOLUTION

```
> bodydata <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/bodydata.csv")
> head(bodydata)
  height upper_arm_length arm_circum hip_circum
1  158.4             36.0       26.5      101.1
2  164.7             38.1       30.5       97.4
3  156.9             34.0       28.5      101.7
4  158.1             35.0       22.2       88.7
5  158.2             35.0       32.0      100.3
6  162.0             34.4       32.7       99.3
> dim(bodydata)
[1] 230   4
```

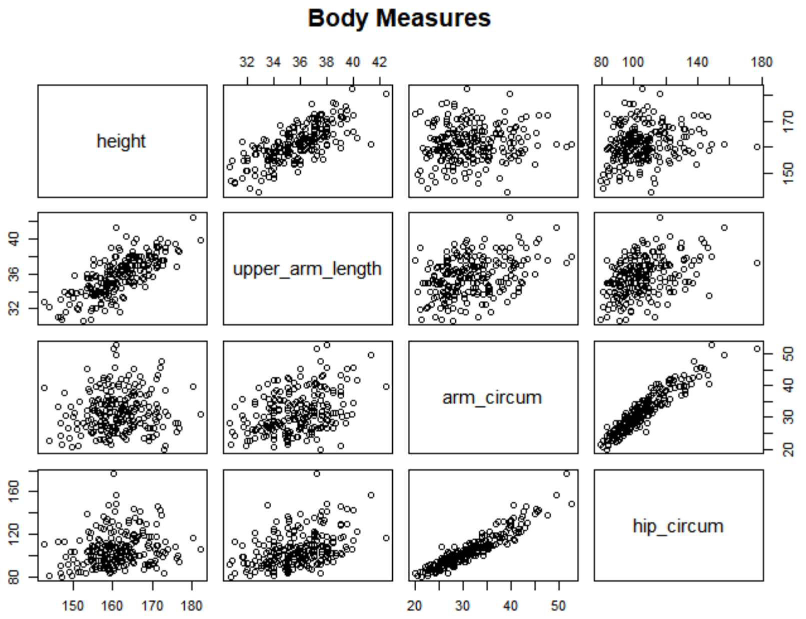
a)

```
> pairs(bodydata[c("height","hip_circum")])
```

```
> pairs(bodydata[c("height","upper_arm_length","arm_circum","hip_circum")])
```



```
>plot(bodydata, main = "Body Measures")
```

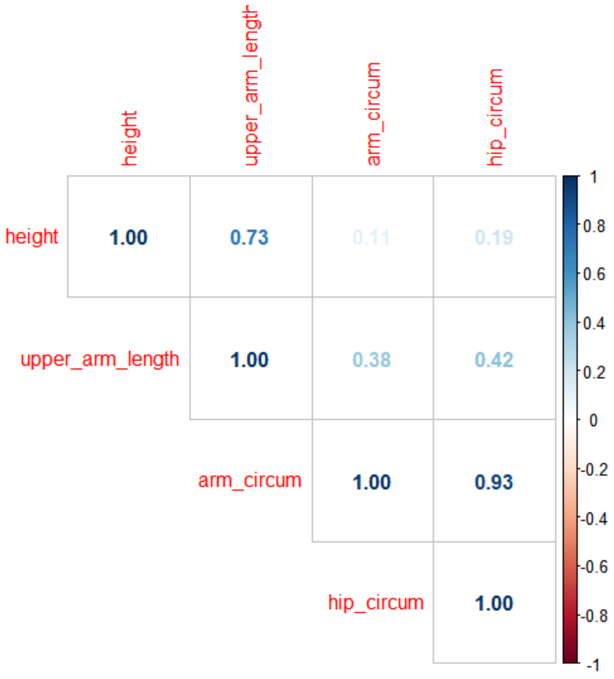


b)

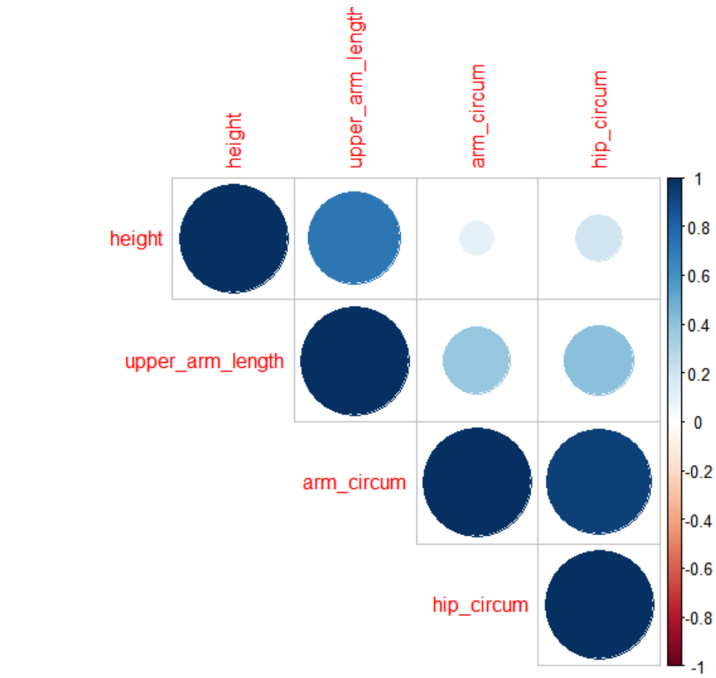
```
> cor(bodydata,use = "complete.obs")
                    height upper_arm_length arm_circum hip_circum
height           1.0000000        0.7259228  0.1061935  0.1942494
upper_arm_length 0.7259228        1.0000000  0.3843140  0.4187889
arm_circum       0.1061935        0.3843140  1.0000000  0.9332575
hip_circum       0.1942494        0.4187889  0.9332575  1.0000000
```

c)

```
> corrplot(cor(bodydata,use = "complete.obs"),method = "number",type = "upper")
```

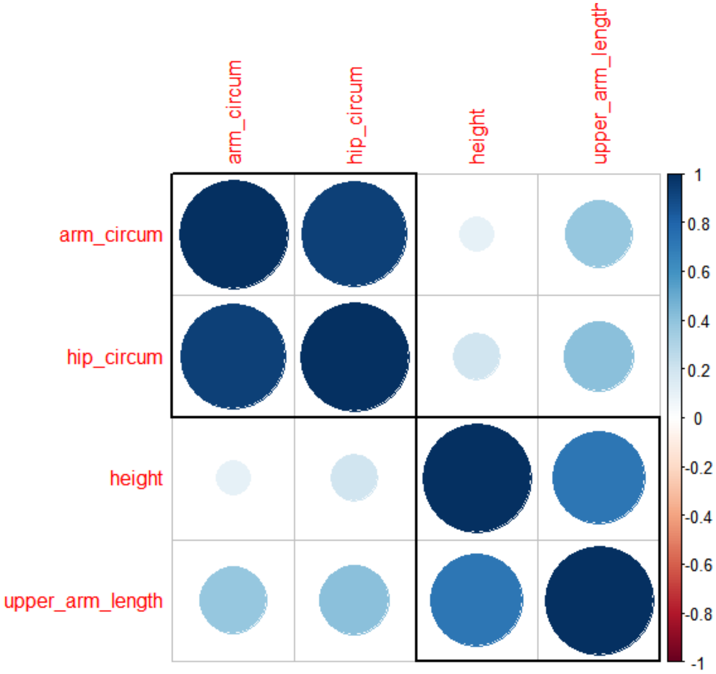


```
> corrplot(cor(bodydata,use = "complete.obs"),method = "circle",type = "upper")
```



d)

```
> corrplot(cor(bodydata,use = "complete.obs"), order = "hclust", addrect = 2)
```



e)

```
> X <- as.matrix(bodydata)
> res <-rcorr(X)
> round(res$P, 3)
                 height upper_arm_length arm_circum hip_circum
height               NA                0      0.176      0.004
upper_arm_length  0.000               NA      0.000      0.000
arm_circum        0.176                0         NA      0.000
hip_circum        0.004                0      0.000         NA
```

f)

CONCLUSION:

- **The variables arm-circum and hip_circum are strongly correlated; so are the variables height and upper_arm_length.**
- **The variables height and arm_circum are not statistically correlated.**

Problem 2. Load the Iris dataset in R

```
library(datasets)
data(iris)
```

It is a data frame with 150 samples (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

a) Generate a statistical summary of the data
b) Apply PCA and LDA analysis to the data and generate a plot to represent the features with respect to 2 dimensions. Make sure to label the axes appropriately and display the different species using different colors or symbols.
c) Concisely discuss the performance of the two methods.

SOLUTION

```
> library(datasets)
> data(iris)
a)
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

b) We first apply PCA and display the PCA coordinates

```
> irisPCA=prcomp(iris[,1:4],scale=TRUE)
> summary(irisPCA)
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
```

Next we apply LDA and display the LDA coordinates

```
> library(MASS)
> irisLDA <-lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
,iris, prior=c(1,1,1)/3)

(alternatively)
> irisLDA <-lda(Species~.,data=iris)
> irisLDA
Call:
lda(Species ~ ., data = iris)
```

```
Prior probabilities of groups:
    setosa versicolor  virginica 
 0.3333333  0.3333333  0.3333333 

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2 
0.9912 0.0088 
```

Next we display the first 2 principal components of the PCA and LDA methods

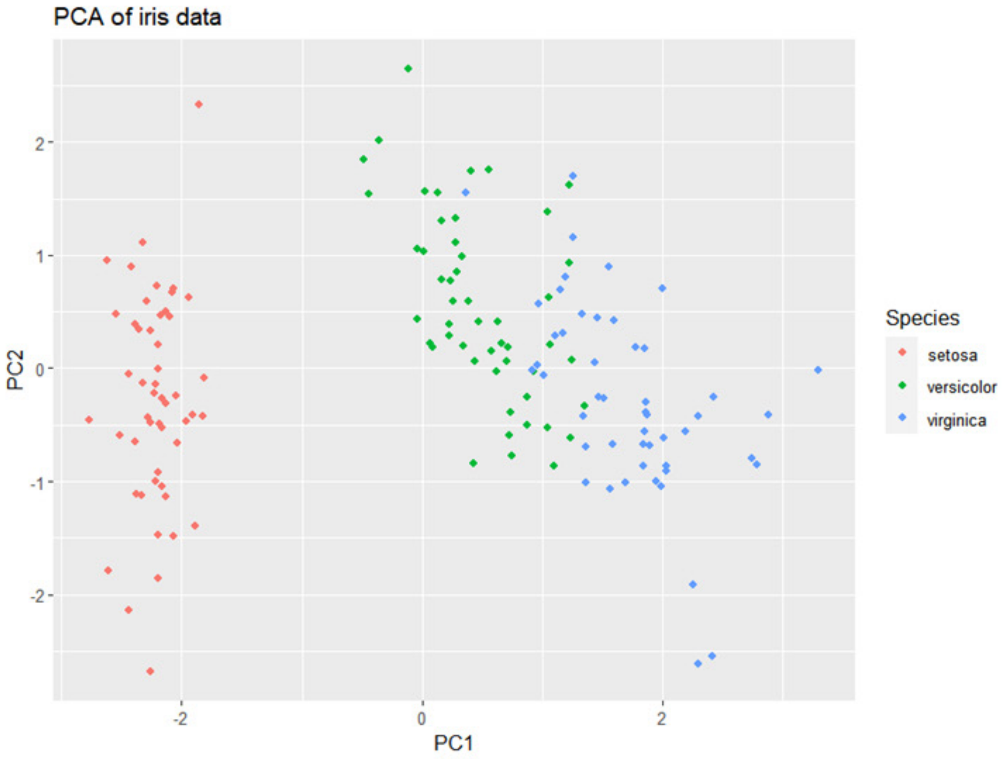
We start with the first two principal components PC1 and PC2 (figure next page)

```
> irisPCA_frame <-data.frame(irisPCA$x,Species=iris$Species)
> ggplot(irisPCA_frame,aes(x=PC1,y=PC2,color=Species))+geom_point()+labs(title="PCA of iris data",x="PC1",y="PC2")
```

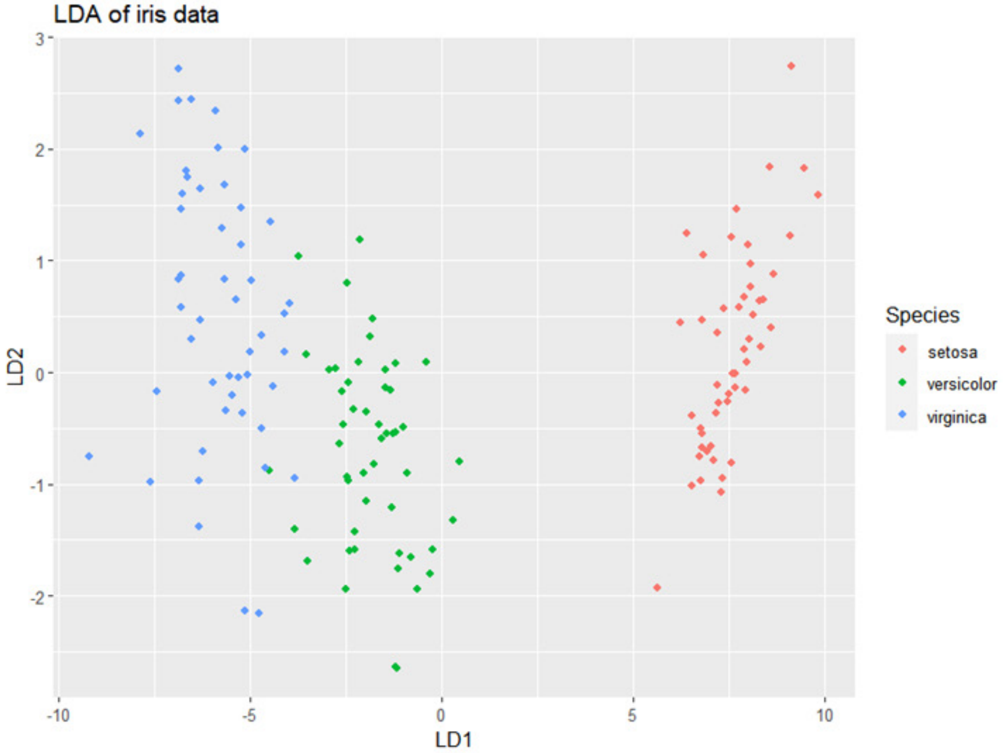
Next we plot the first two LDA components LD1 and LD2. For convenience we plot the PCA and LDA figures side by side in the next page

```
> irisLDAmodel<-predict(irisLDA)
> irisLDA_frame <-data.frame(irisLDAmodel$x,Species=iris$Species)
> ggplot(irisLDA_frame,aes(x=LD1,y=LD2,color=Species))+geom_point()+labs(title="LDA of iris data",x="LD1",y="LD2")
```

d) **The plots show that LDA separates the 3 species almost exactly using the single coordinate LD1. By contrast, The versicolor and virgica species has some overlap in the PCA projection. Thus, the LDA approach results in a more effective separation of the 3 species.**



PCA of iris data



LDA of iris data