Pulse Analysis

Daniel Eduardo López*

15/10/2022

1. Goal

In the present report, the relationship among the **pulse at rest**, **pulse in activity**, **height**, **weight**, and **body mass index** of a group of **male** and **female** patients was assessed.

The question to answer was: **How do these variables relate among each other and with patients' gender?**

2. Analytic Approach

Unsupervised learning techniques such as Principal Component Analysis (PCA) and Clustering Analysis were applied to infer relationships among the variables in study.

3. Data Requirements

Accordingly, patients' data about the variables pulse at rest (restPulse), pulse in activity (actiPulse), height (heig), weight (weig) and Body Mass Index (bmi) was needed.

4. Data Collection

The data was provided by the **University of Zaragoza** in the course of the **Master's in Quantitative Biotechnology** 2020-2021.

5. Data Understanding & Preparation

First, data was loaded into R:

```
load("pulse.RData")
summary(pulse)
```

```
##    restPulse        actiPulse   runActi  smoke       gender        heig
##  Min.   : 48.00   Min.   : 50   no :57   no :64   female:35   Min.   :154.9
##  1st Qu.: 64.00   1st Qu.: 68   yes:35   yes:28   male  :57   1st Qu.:167.6
##  Median : 71.00   Median : 76                                 Median :175.3
##  Mean   : 72.87   Mean   : 80                                 Mean   :174.5
##  3rd Qu.: 80.00   3rd Qu.: 85                                 3rd Qu.:182.9
##  Max.   :100.00   Max.   :140                                 Max.   :190.5
##       weig           acti         bmi
##  Min.   :43.09   low   : 9   Min.   :16.73
##  1st Qu.:56.70   medium:61   1st Qu.:20.00
##  Median :65.77   high  :21   Median :21.42
##  Mean   :65.84   NA's  : 1   Mean   :21.50
```

*https://github.com/DanielEduardoLopez, https://www.linkedin.com/in/daniel-eduardo-lopez/

```
##  3rd Qu.:70.53               3rd Qu.:22.83
##  Max.   :97.52               Max.   :29.16
```

Then, the variables of interest were retrieved from the data to construct a new dataset:

```
restPulse = pulse$restPulse
actiPulse = pulse$actiPulse
heig = pulse$heig
weig = pulse$weig
bmi = pulse$bmi

pulse.gender = pulse$gender
pulse.act = pulse$acti
pulse.data = cbind(restPulse, actiPulse, heig, weig, bmi)
head(pulse.data)
```

```
##      restPulse actiPulse   heig     weig      bmi
## [1,]        48        54 172.72 68.03880 22.80717
## [2,]        54        56 175.26 65.77084 21.41252
## [3,]        54        50 175.26 72.57472 23.62761
## [4,]        58        70 182.88 65.77084 19.66532
## [5,]        58        58 167.64 61.23492 21.78933
## [6,]        58        56 170.18 56.69900 19.57755
```

So, the new dataset only includes the variables of interest: restPulse, actiPulse, heig , weig and bmi.

6. Modeling & Evaluation

6.1. Principal Component Analysis

As a first approach, a Principal Component Analysis (PCA) was performed in order to explore the data.

```
pr.out<-prcomp(pulse.data, scale = TRUE)
summary(pr.out)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5494 1.2014 0.8799 0.61582 0.05074
## Proportion of Variance 0.4801 0.2887 0.1548 0.07585 0.00051
## Cumulative Proportion  0.4801 0.7688 0.9236 0.99949 1.00000
```

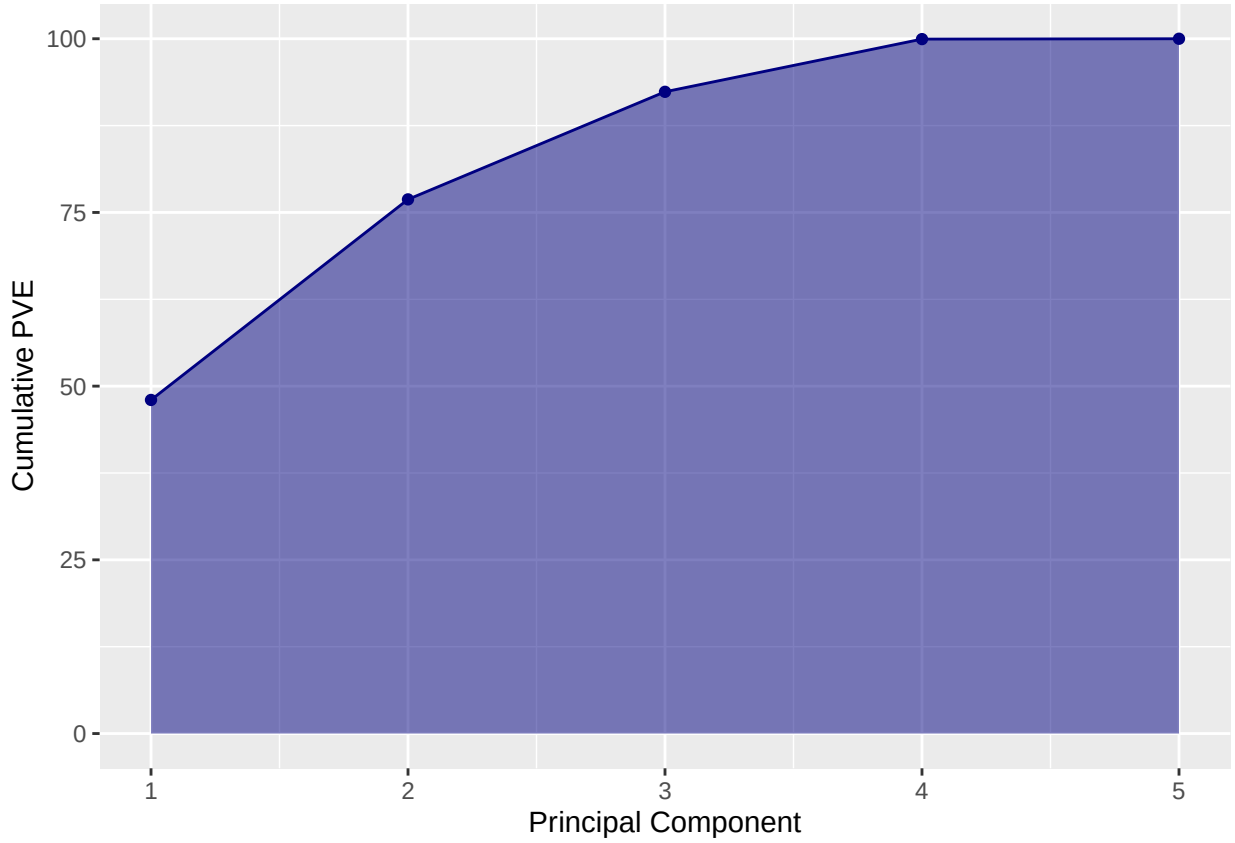
```
head(pr.out$x)
```

```
##             PC1       PC2         PC3          PC4         PC5
## [1,] -1.4672321 -2.172084  0.78194257 -0.602176969 -0.04783121
## [2,] -0.9788028 -1.970037  0.13556884 -0.240570263 -0.03841581
## [3,] -1.9184022 -1.704913  0.86692645 -0.042972391 -0.04348441
## [4,] -0.6642151 -1.335880 -1.11308024 -0.439496473 -0.05095412
## [5,] -0.2490984 -1.863336  0.82211456 -0.144913041 -0.03731935
## [6,]  0.2906541 -2.325297 -0.03744096  0.007706298 -0.01317994
```

In order to explain at least 90% of variance from the data, a suitable number of principal components was selected according to the criterion of the Cumulative Percentage of Variance Explained (PVE), plotted in the following figure:

```
pve <-100*pr.out$sdev^2/sum (pr.out$sdev^2)

library(ggplot2)
```

```
pve.df = as.data.frame(cbind(1:length(pve),cumsum(pve)))
ggplot(data = pve.df, aes(x = pve.df[1:length(pve),1], y =pve.df[1:length(pve),2])) + geom_point(color=
```



Thus, in view of the former figure, 3 principal components are required to explain at least 90% of variance from the data, effectively reducing the dimension of the dataset from 6 variables to only 3.

To interpret the resulting components, it is useful to check the coefficients of each component according to the variables:

```
pr.out$rotation
```

```
##                 PC1       PC2          PC3         PC4          PC5
## restPulse  0.3201933 0.6233876 -0.007349452  0.71330851 -0.001073276
## actiPulse  0.2948346 0.6406360 -0.145878518 -0.69370784  0.012122048
## heig      -0.4890108 0.1550802 -0.708308317  0.07596045 -0.478902834
## weig      -0.6055892 0.2855428 -0.018705245  0.02321703  0.742184681
## bmi       -0.4524146 0.3088509  0.690373005 -0.06042667 -0.468685267
```

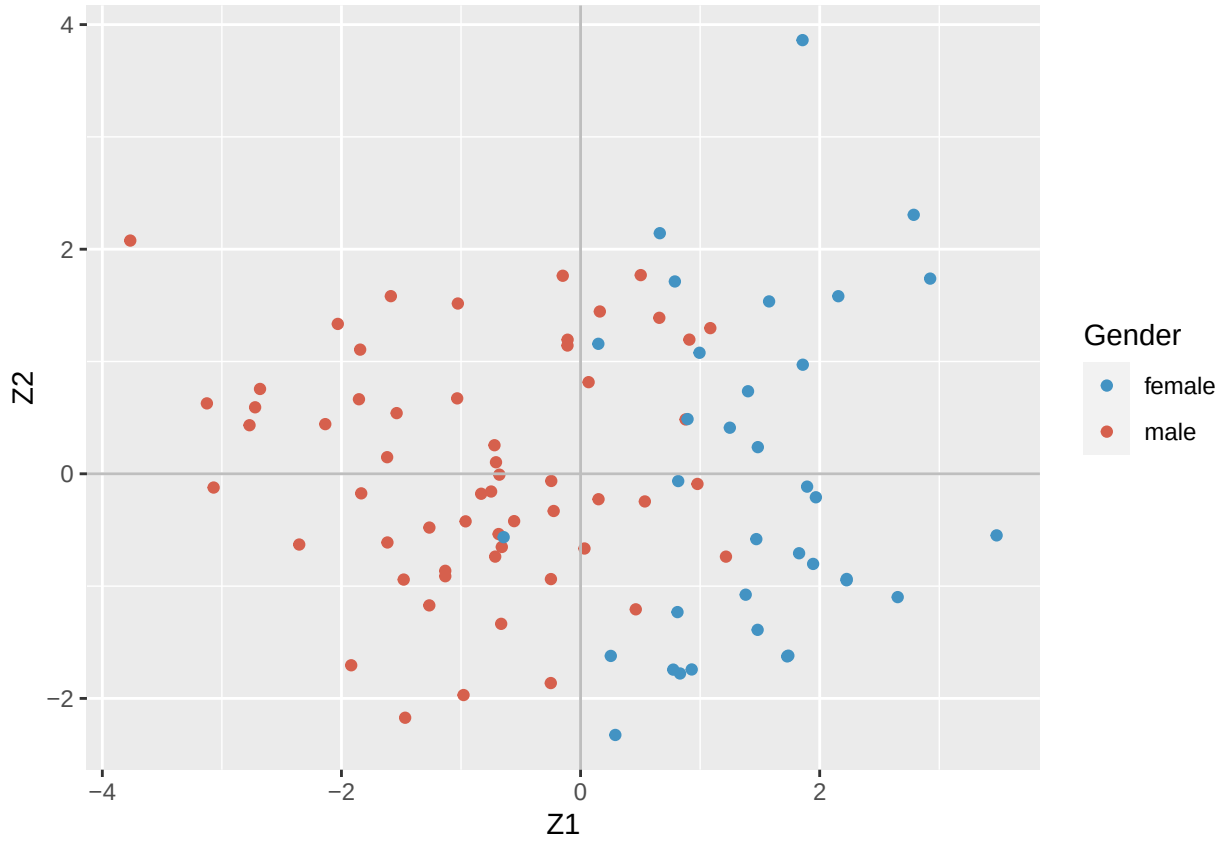
In this context, we can see that the PC1 is positively correlated to the variables restPulse (0.3201933) and actiPulse (0.2948346) and negatively correlated to the variables heig (-0.4890108), weig (-0.6055892) and bmi (-0.4524146). On the other hand, PC2 is positively correlated with all the variables and PC3 is negatively correlated with all the variables except bmi.

Thus, we may infer that the PC1 will be the most useful to look for relationships within the data in the analysis below.

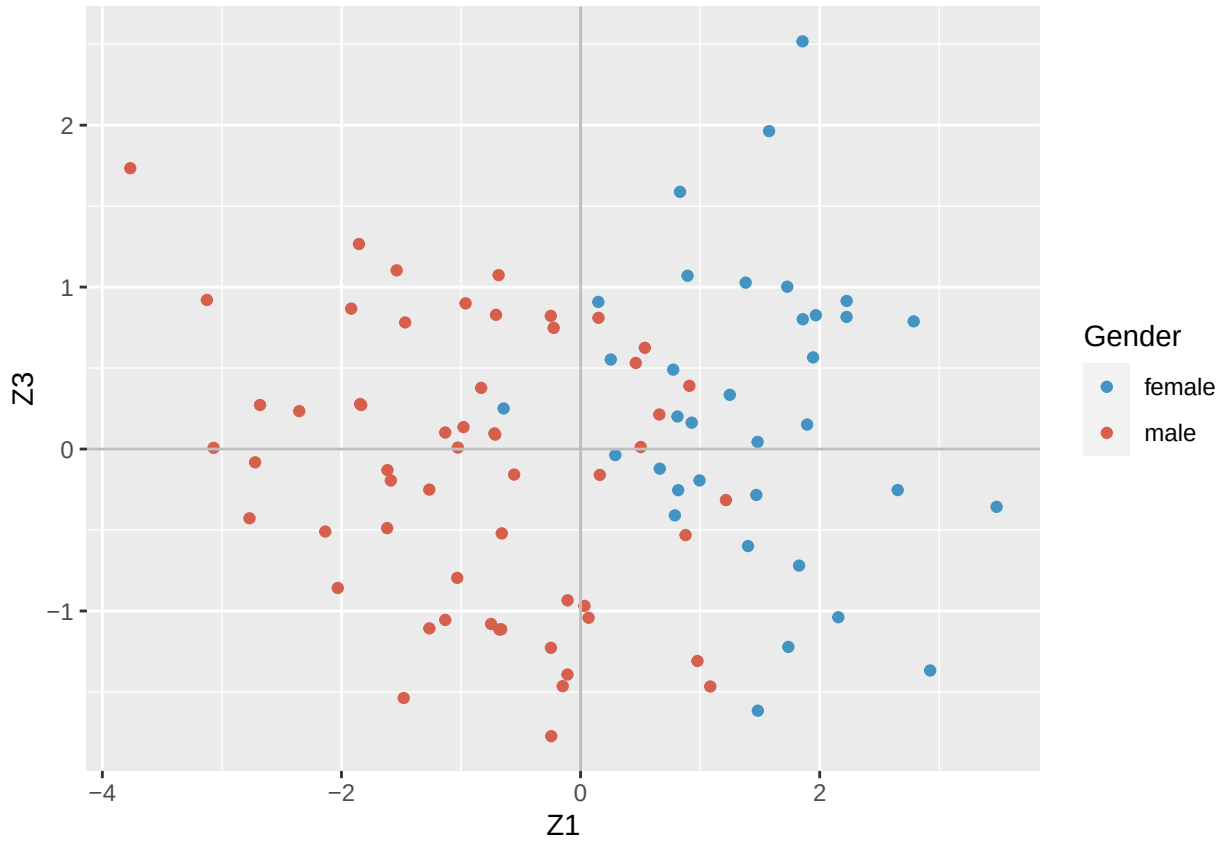
Furthermore, as we have reduced the dimensions from 6 to 3, we may interpret that, indeed, the variables are somewhat correlated.

Then, the observations were represented graphically in the new reduced space, differentiating between the gender of the patients in the plots.

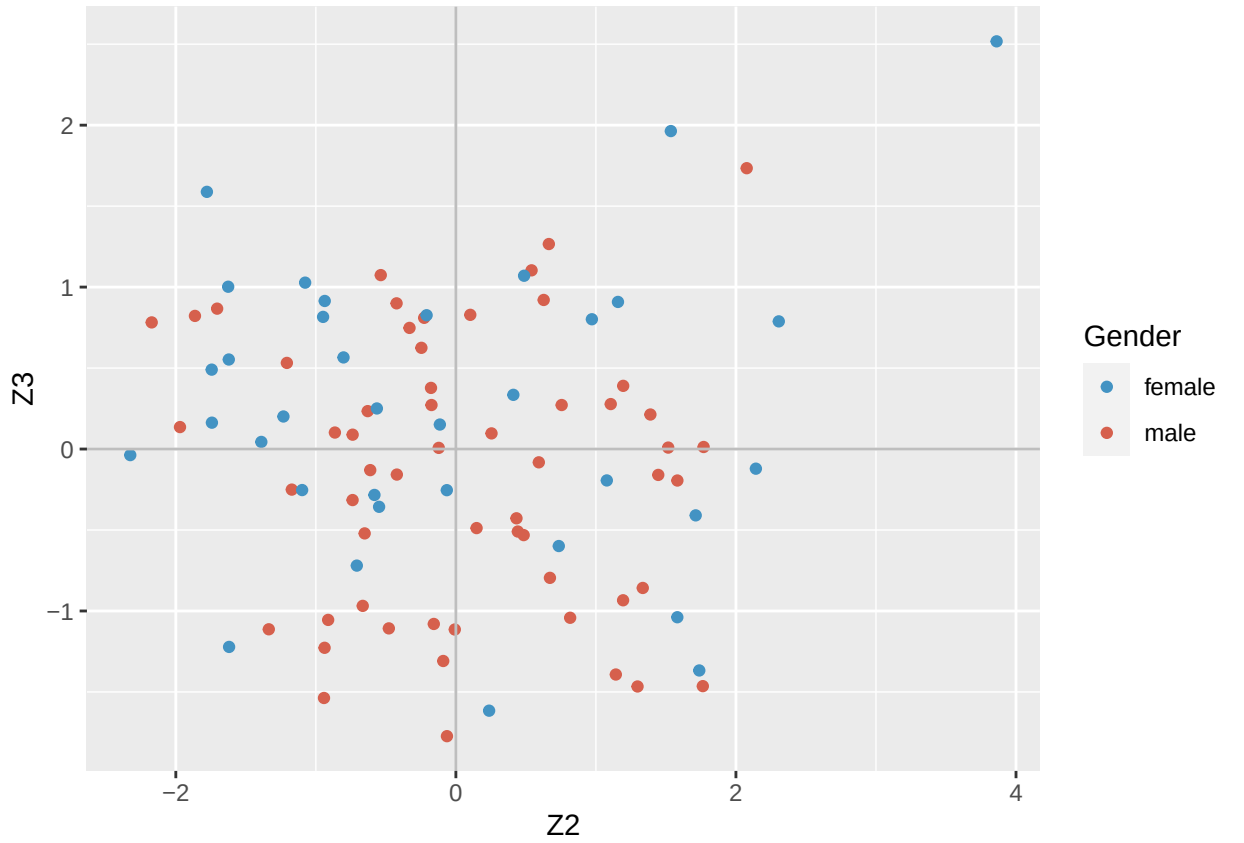
```
# Converting datasets into data frames for plotting
po.df = as.data.frame(pr.out$x[,1:3])
pr.df = as.data.frame(pr.out$rotation[,1:3])

# Set of colors for female and male patients
colors <- c("female" = "#4393C3", "male" = "#D6604D")

# Score Plot of Z1 vs Z2
ggplot(data = po.df) + geom_point(aes(x = PC1, y = PC2, color  = pulse.gender)) + labs(x = "Z1", y = "Z
```




```
# Score Plot of Z1 vs Z3
ggplot(data = po.df) + geom_point(aes(x = PC1, y = PC3, color  = pulse.gender)) + labs(x = "Z1", y = "Z
```




```
# Score Plot of Z2 vs Z3
ggplot(data = po.df) + geom_point(aes(x = PC2, y = PC3, color  = pulse.gender)) + labs(x = "Z2", y = "Z
```



From the plots of Z1 vs Z2 and Z1 vs Z3, it is possible to clearly distinguish that the principal component 1 is able to capture the relationship between the gender of the patients and the data. In particular, the PC1 is positively correlated with the female gender and negatively correlated with the male gender, thus generating two distinct groups.

Therefore, if we recall that PC1 is positively correlated with the variables restPulse (0.3201933) and actiPulse (0.2948346) and negatively correlated to the variables heig (-0.4890108), weig (-0.6055892) and bmi (-0.4524146); we may infer that the female patients tend to exhibit a higher restPulse and actiPulse but smaller values of height, weight and bmi; which is reasonable as women tend to have lowers values of height, weight and bmi in comparison to men.

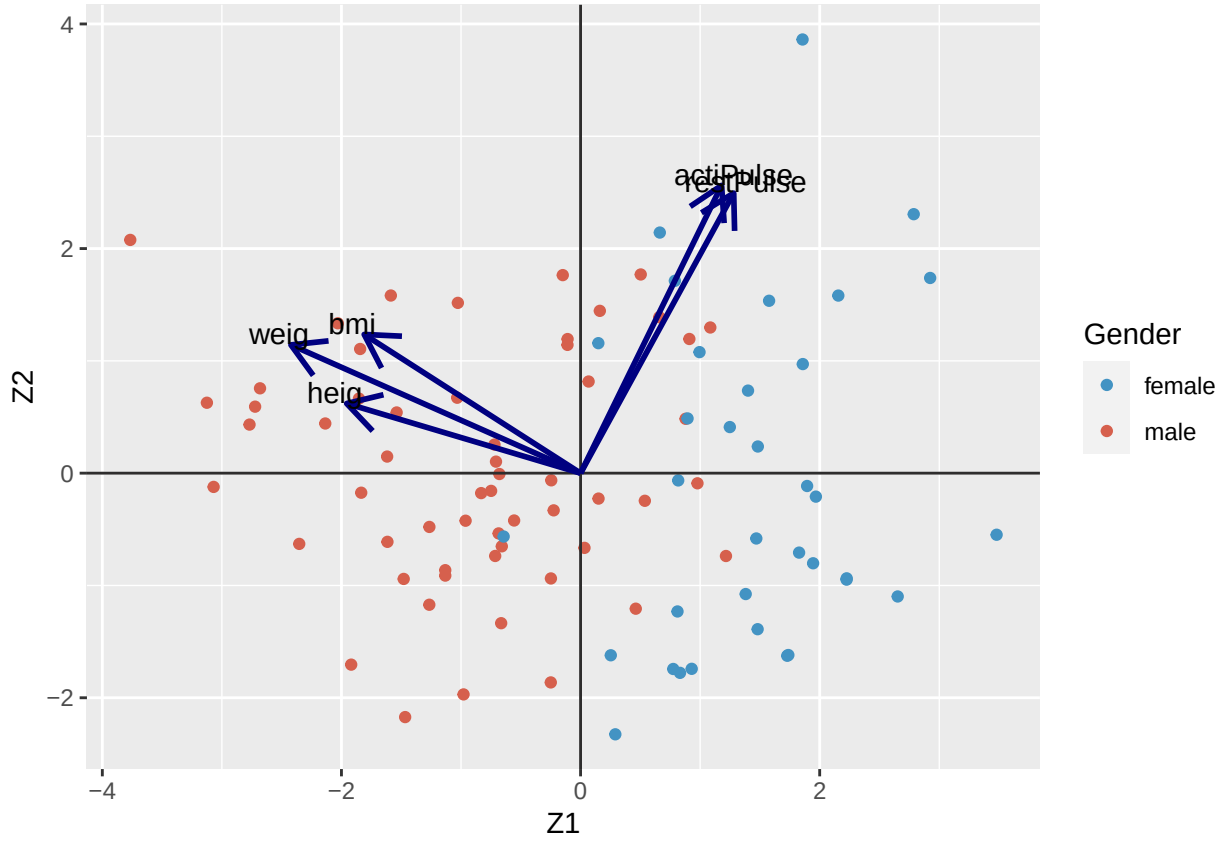
In this sense, we can also interpret that male patients tend to have a smaller restPulse and actiPulse but higher values of height, weight and bmi which, again, is reasonable according to the reasoning exposed above.

On the other hand, the plot of Z2 vs Z3 fails in generating a clear pattern between the observations and the gender of the patients.

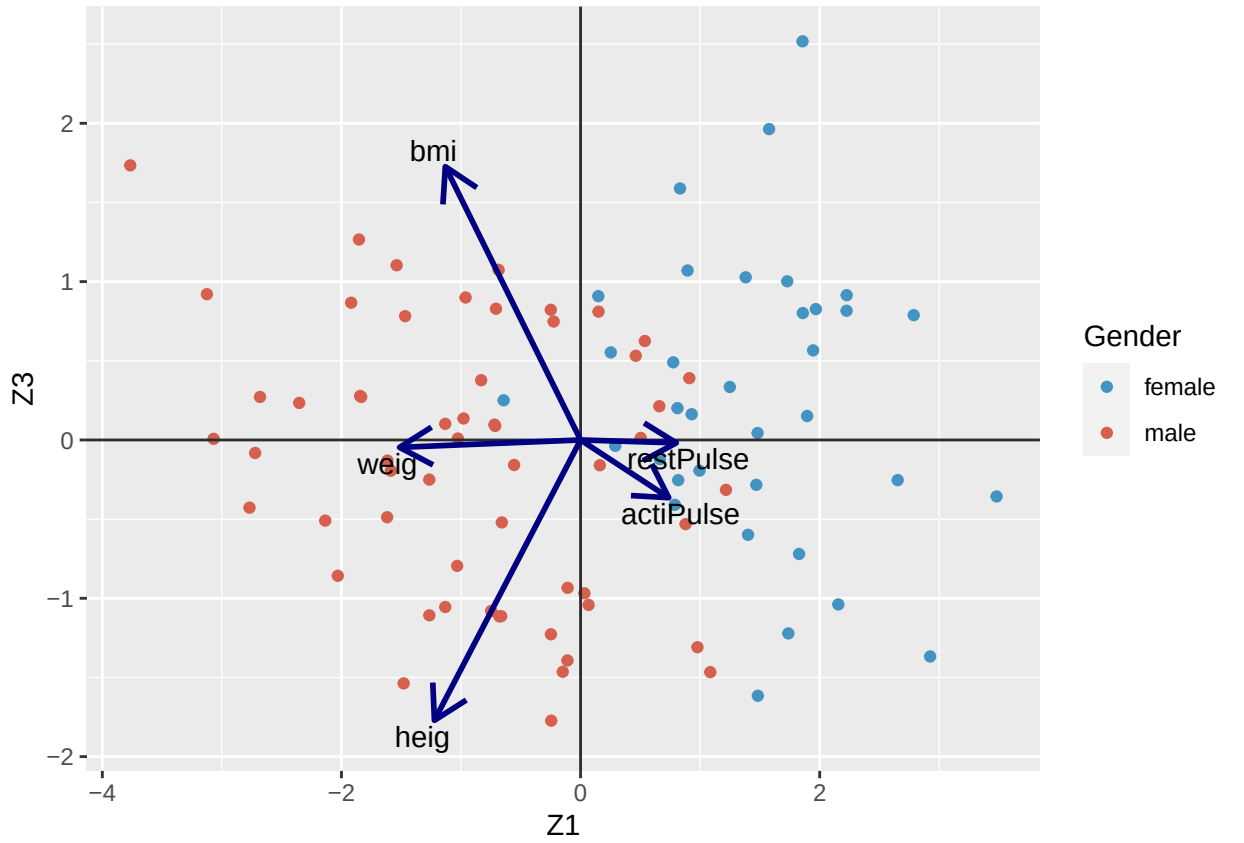
This interpretation is coherent if we take a look to the biplots of Z1 vs Z2 and Z1 vs Z3:

```
# Function for adjusting the position of the labels in the biplots
labeladj = function(v){
  counter = 1
  for (i in v){
  if (i >= 0){
    v[counter] = i + 0.1
  }
  if (i < 0){
    v[counter] = i - 0.1
  }
```

```
  counter = counter + 1
  }
  return(v)
}

# Z1 vs Z2
scale = 4
ggplot() + geom_point(data = po.df, aes(x = PC1, y = PC2, color  = pulse.gender)) + labs(x = "Z1", y =
```




```
# Z1 vs Z3
scale = 2.5
ggplot() + geom_point(data = po.df, aes(x = PC1, y = PC3, color  = pulse.gender)) + labs(x = "Z1", y =
```



In view of the two previous biplots, it is possible to interpret that, indeed, the observations from the female patients are associated with higher values for restPulse and actiPulse, while being also associated with lower values for weigth, height and bmi. On the contrary, the observations from the male patients are associated with lower values for restPulse and actiPulse, while being also associated with higher values for weigth, height and bmi.

With these relationships in mind, it would be possible, for instance, to design a further test of hypothesis to compare the means of the observations from male and female patients in one or several of the variables assessed in this analysis.

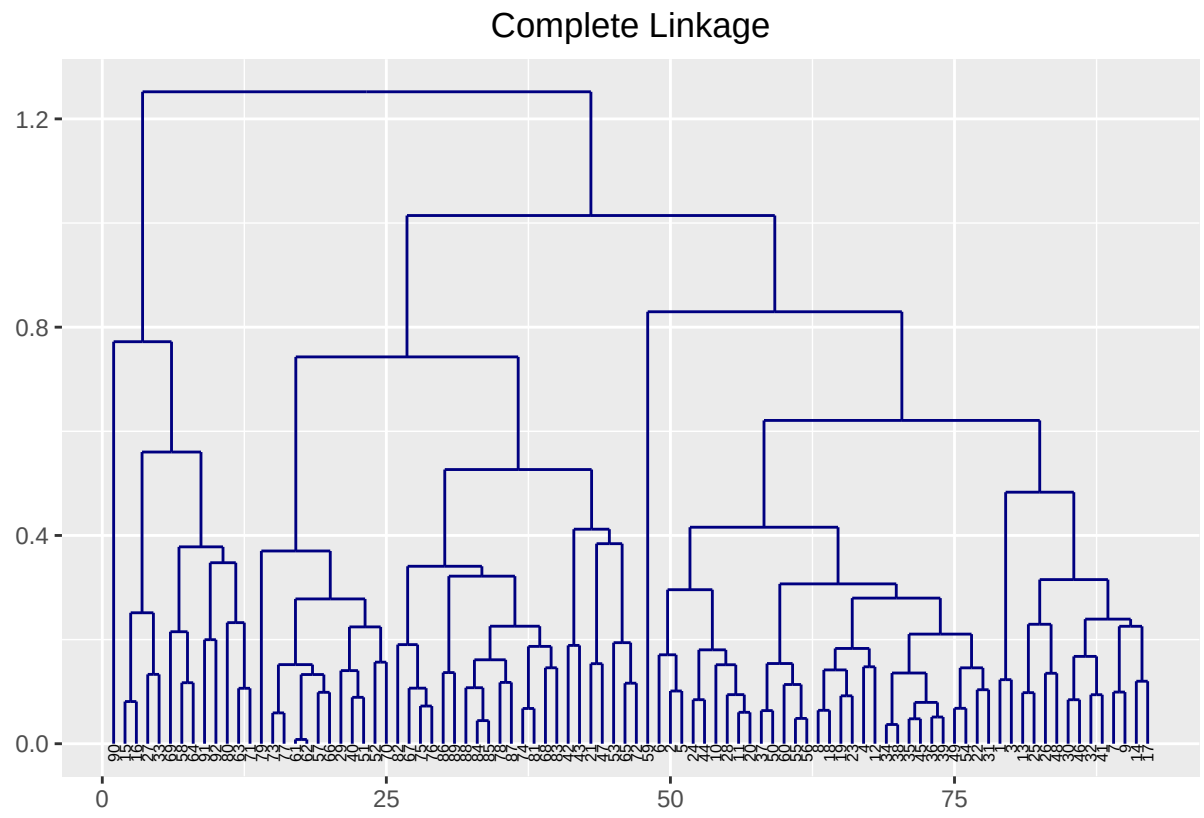
Thus, we may conclude that the current PCA analysis was successful in to capture some important relationships within the observations and the gender in the data, which shows the great potential that this statistical technique have as an exploratory tool for datasets with many variables.

6.2. Clustering Analysis

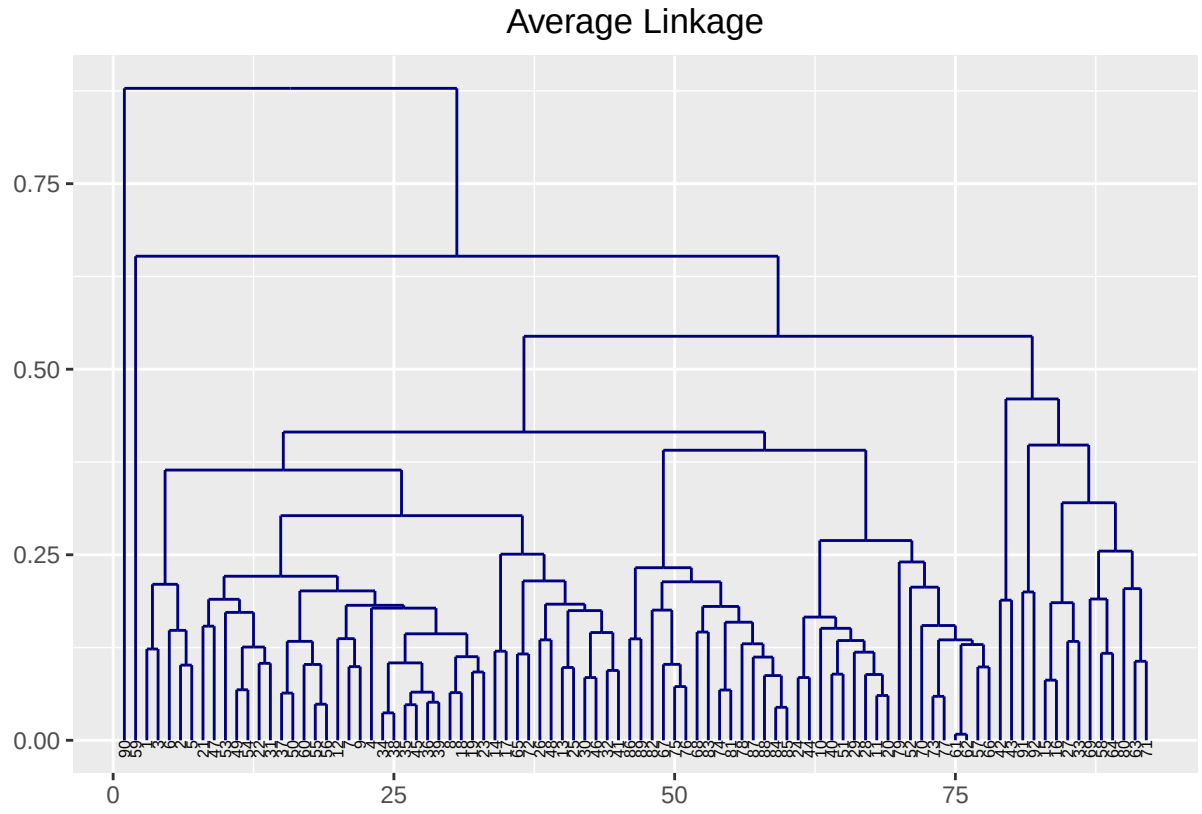
Afterwards, a cluster analysis was performed to the data. To do so, first, a hierarchical cluster analysis was applied in order to select the most suitable number of clusters.

```
# Data scaling
sd.data<-scale(pulse.data, center = FALSE , scale = TRUE)

# Computation of distances matrix
data.dist<-dist(sd.data)

# Hierarchical Clustering with the Complete Linkage Method
clustcomp <- hclust(data.dist, method = "complete")
```

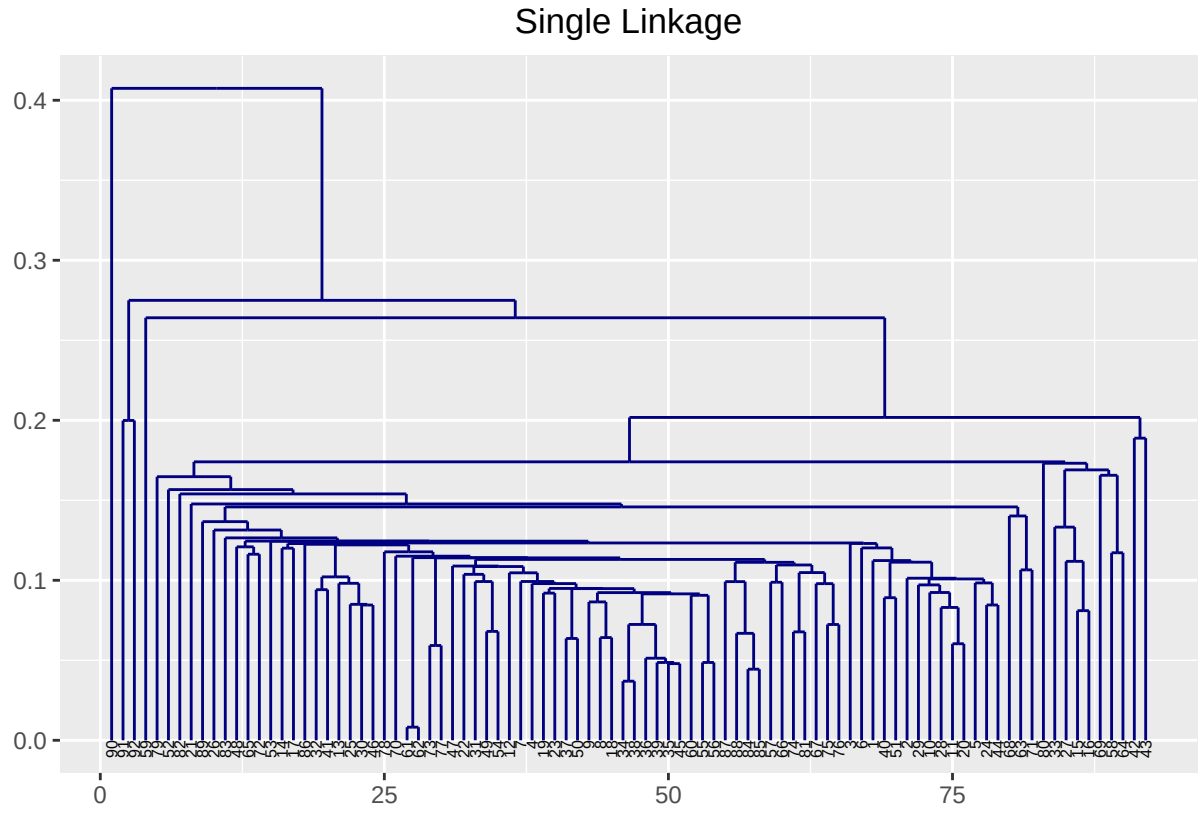
```
library("ggdendro")

# Dendrogram plot with Complete Linkage
dendcomp <- as.dendrogram(clustcomp)
dendcomp_data <- dendro_data(dendcomp, type = "rectangle")
ggplot(dendcomp_data$segments) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), color = "navyblue")+
  geom_text(data = dendcomp_data$labels, aes(x, y, label = label),
            hjust = 1, angle = 90, size = 2) + labs(title="Complete Linkage", x="", y="" ) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Hierarchical Clustering with the Average Linkage Method
clustav <- hclust(data.dist, method = "average")
dendav <- as.dendrogram(clustav)
dendav_data <- dendro_data(dendav, type = "rectangle")

# Dendrogram plot with Average Linkage
ggplot(dendav_data$segments) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), color = "navyblue")+
  geom_text(data = dendav_data$labels, aes(x, y, label = label),
            hjust = 1, angle = 90, size = 2) + labs(title="Average Linkage", x="", y="" ) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Hierarchical Clustering with the Single Linkage Method
clustsin <- hclust(data.dist, method = "single")
dendsin <- as.dendrogram(clustsin)
dendsin_data <- dendro_data(dendsin, type = "rectangle")

# Dendrogram plot with Single Linkage
ggplot(dendsin_data$segments) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), color = "navyblue")+
  geom_text(data = dendsin_data$labels, aes(x, y, label = label),
            hjust = 1, angle = 90, size = 2) + labs(title="Single Linkage", x="", y="" ) +
  theme(plot.title = element_text(hjust = 0.5))
```



From the three dendrograms plotted above, the ones that generates the most distinguishable clusters were the ones constructed with the Complete Linkage and the Average Linkage methods, yielding about 8 and 7 subgroups, respectively. On the other hand, in the Single Linkage method it is not possible to easily differentiate the clusters so it will be discarded from further analysis.

In order to evaluate the goodness of clusterings generated, it is possible to use the Dunn's Index; which is the ratio between the minimum inter-cluster distances to the maximum intra-cluster diameter. The diameter of a cluster is the distance between its two furthermost points. In order to have well separated and compact clusters, it is necessary to aim for a higher Dunn's index.

```
library(clValid)

# Dunn's Index for the Hierarchical Clustering with Complete Linkage method
clustc <- cutree(clustcomp, 8)
dunn(data.dist, clustc)
```

```
## [1] 0.1642313
```

```
# Dunn's Index for the Hierarchical Clustering with Average Linkage method
clusta <- cutree(clustav, 7)
dunn(data.dist, clusta)
```

```
## [1] 0.1548116
```

As shown by the Dunn function, the Dunn's Index is higher for the Complete Linkage method.

In this context, as the Complete Linkage method exhibited the highest Dunn's Index value, the clusters are more distinguishable and it is the linkage technique most used in cluster analysis, this method was selected as the most suitable for the next analysis.

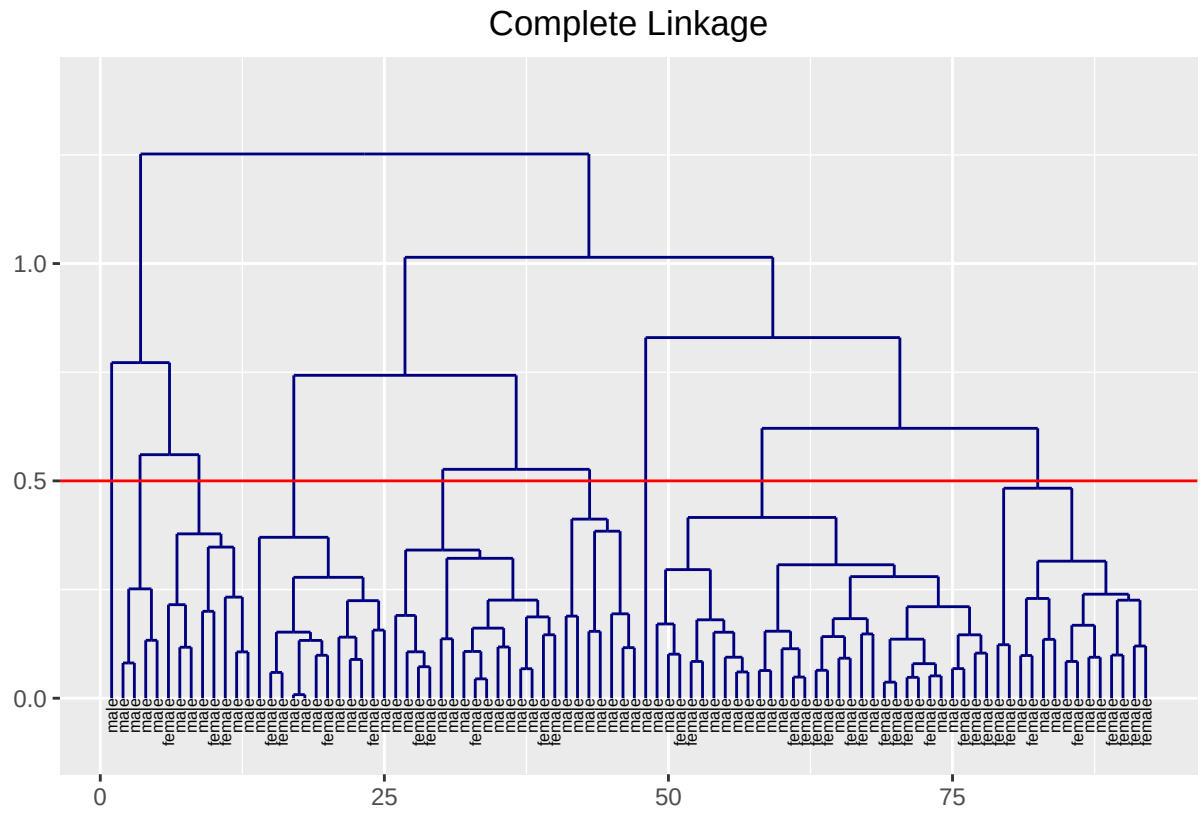
So, by building a table with the results using the Complete Linkage method and 8 clusters, the following characterization according to the gender of each patient was obtained:

```
hc.clusters<-cutree(clustcomp,8)
table(hc.clusters, pulse.gender)
```

```
##             pulse.gender
## hc.clusters female male
##           1      0   14
##           2      7   23
##           3      3    1
##           4      6   16
##           5     11    1
##           6      7    1
##           7      0    1
##           8      1    0
```

And the following labeled dendrogram according to the gender of each patient was generated:

```
ggplot(dendcomp_data$segments) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), color = "navyblue")+
  geom_text(data = dendcomp_data$labels, aes(x, y, label = pulse.gender),
            hjust = 1, angle = 90, size = 2) + labs(title="Complete Linkage", x="", y="" ) +
  theme(plot.title = element_text(hjust = 0.5)) + geom_hline(aes(yintercept = 0.5), color = "red") +
  ylim(-0.10,1.4)
```



After selecting the number of clusters from the Hierarchical Cluster Analysis, the K-means Cluster Analysis was applied with 8 clusters, using a set.seed(2) and nstart = 20.

```
set.seed(2)
km.out<-kmeans(sd.data, 8, nstart = 20)
km.out
```

```
## K-means clustering with 8 clusters of sizes 10, 25, 10, 15, 13, 6, 7, 6
##
## Cluster means:
##   restPulse actiPulse      heig      weig       bmi
## 1 1.0622034 0.9387777 0.9376252 0.7628548 0.8670426
## 2 0.9177869 0.8765168 1.0200784 1.0295835 0.9906554
## 3 0.8732473 0.8098782 0.9358186 0.8216920 0.9378195
## 4 1.1823255 1.0733521 0.9890531 0.9707012 0.9927161
## 5 0.9343985 0.9269916 1.0483837 1.2537371 1.1407690
## 6 0.7558245 0.6850321 0.9996518 1.0369776 1.0353341
## 7 0.9293557 1.2855209 0.9884681 0.8936686 0.9116823
## 8 1.2102190 1.4734271 0.9587022 0.8983382 0.9824497
##
## Clustering vector:
##  [1] 6 6 6 2 6 3 5 6 2 3 3 2 5 6 7 7 5 2 2 3 2 2 2 3 5 5 7 3 3 5 2 5 7 2 2 2 2 2
## [39] 2 3 2 7 5 3 2 5 2 5 2 2 3 1 2 2 2 2 1 7 5 2 1 1 7 8 5 1 1 4 8 1 4 5 1 4 4 4
## [77] 1 4 1 8 4 4 4 4 4 4 4 4 4 4 8 8 8
##
## Within cluster sum of squares by cluster:
## [1] 0.15848585 0.45562984 0.11666951 0.28078437 0.36017792 0.09443633 0.23840939
## [8] 0.31973796
##  (between_SS / total_SS =  78.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Then, the gender of the patients was examined by cluster.

```
km.clusters<-km.out$cluster
table(km.clusters, pulse.gender)
```

```
##            pulse.gender
## km.clusters female male
##           1      9    1
##           2      1   24
##           3      9    1
##           4      6    9
##           5      0   13
##           6      0    6
##           7      4    3
##           8      6    0
```

Afterwards, a comparison of the results from the Hierarchical Clustering and the K-means Cluster Analysis was summarized in the following table:

```
table(km.clusters, hc.clusters)
```

```
##            hc.clusters
## km.clusters  1  2  3  4  5  6  7  8
##           1  0  0  0  1  9  0  0  0
```

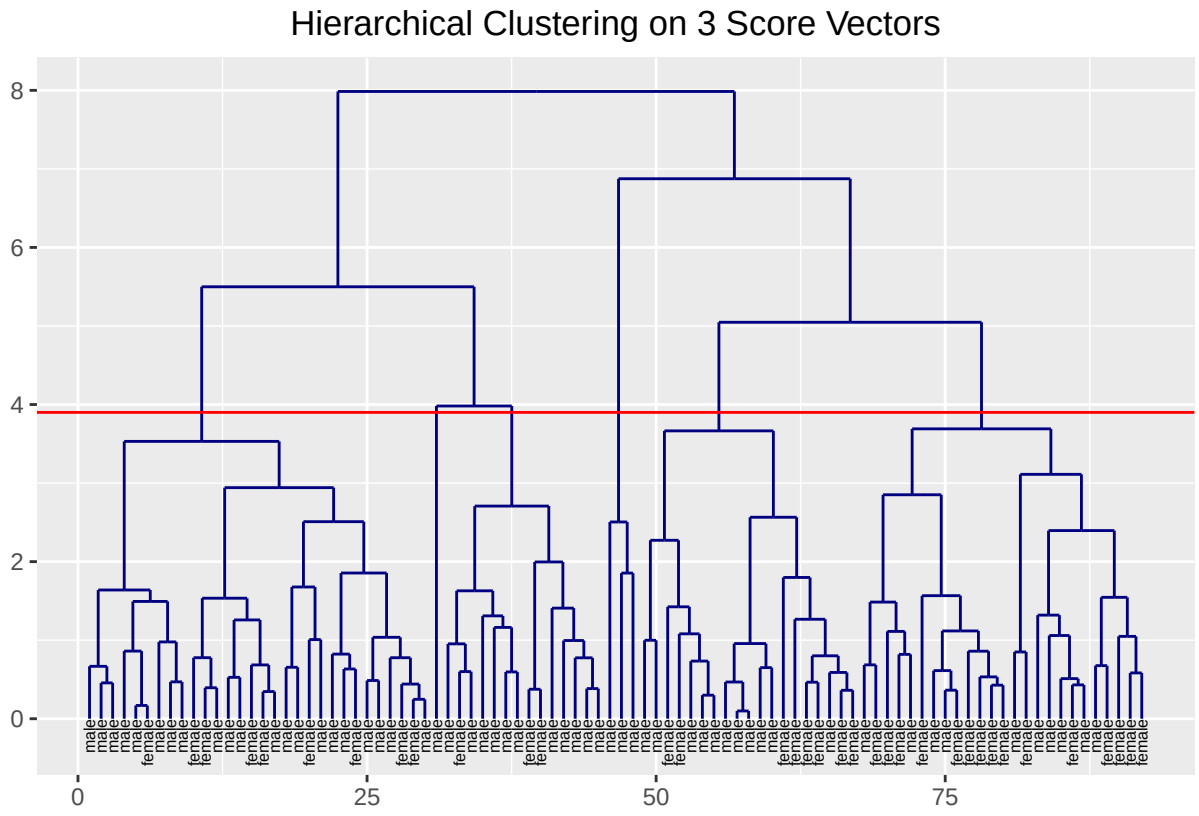
```
##           2  2 20  0  3  0  0  0  0
##           3  0  7  0  0  3  0  0  0
##           4  0  0  0 14  0  1  0  0
##           5  9  0  0  3  0  0  1  0
##           6  3  3  0  0  0  0  0  0
##           7  0  0  4  1  0  2  0  0
##           8  0  0  0  0  0  5  0  1
```

Finally, a Hierarchical Clustering Analysis was performed based on the already calculated principal components.

```
# Hierarchical Clustering Analysis with PC
clustpca <- hclust(dist(pr.out$x[,1:3]), method = "complete")
table(cutree(clustpca,8), pulse.gender)
```

```
##    pulse.gender
##     female male
##   1      1   29
##   2     11    1
##   3      0   14
##   4      6    1
##   5      5    7
##   6      0    1
##   7      9    4
##   8      3    0
```

```
# Dendrogram plot with principal components
dendpca <- as.dendrogram(clustpca)
dendpca_data <- dendro_data(dendpca, type = "rectangle")
ggplot(dendpca_data$segments) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), color = "navyblue")+
  geom_text(data = dendpca_data$labels, aes(x, y, label = pulse.gender),
            hjust = 1, angle = 90, size = 2) + labs(title="Hierarchical Clustering on 3 Score Vectors",
  theme(plot.title = element_text(hjust = 0.5)) + geom_hline(aes(yintercept = 3.9), color = "red") +
  ylim(-0.3,8)
```



From the whole cluster analysis, we can observe that the clusters generated from both the Hierarchical analysis and the K-means analysis somewhat coincide. For instance, from the former table, both methods share 20 elements in their respective cluster 2 and 14 elements in their respective cluster 4.

It is also noteworthy that, indeed, both cluster methods were able to group the gender of the patients in mostly homogeneous subgroups. For example, in the hierarchical clustering, the clusters 1, 2 and 4 are mostly made of male patients; while the clusters 3, 5, 6 and 8 are mostly made of female patients.

Likewise, in the k-means method, the clusters 2,5 and 6 are mostly made of male patients; while the clusters 1, 3 and 8 are mostly made of female patients.

Furthermore, the Hierarchical Clustering Analysis based on the principal components also yielded a similar behavior than the previous clusters, as some clusters groups a majority of male patients whereas others groups a majority of female ones.

Thus, the clustering analysis has proved to be useful in to relate some subgroups of observations in terms of the gender of the patients, which is also consistent with the findings from the PCA.

7. Conclusions

Pulse at rest and pulse in activity are negatively correlated with the weight, height and body mass index in both male and female patients.

On the other hand, female patients had a tendency to exhibit lower values of weight, height and body mass index and higher values of pulse at rest and pulse in activity. On the contrary, male patients exhibited a higher values of weight, height and body mass index and lower values of pulse at rest and pulse in activity.

Finally, the present study showed that most of the patients tend to display a similar behavior in terms of pulse at rest, pulse in activity, weight, height and body mass index according to their gender, forming mostly

distinctive groups in both PCA and clustering analysis.