Principles of Data Science

Packages used | Installation and Loading

```
if(!require("factoextra")) install.packages("factoextra");
library(factoextra)


if(!require("olsrr")) install.packages("olsrr");
library(olsrr)

# Devtools needed for installing ggbiplot
if(!require("devtools")) install.packages("devtools");
library(devtools)
# if(!require("ggbiplot")) install_github("vqv/ggbiplot");
library(ggbiplot)

if(!require("GGally")) install.packages("GGally");
library(GGally)

if(!require("ggplot2")) install.packages("ggplot2");
library(ggplot2)

if(!require("reshape2")) install.packages("reshape2");
library(reshape2)

if(!require("dplyr")) install.packages("dplyr");
library(dplyr)
```

Dataset

The dataset used in this portfolio consists of an `id` number, `age` (years), `weight` (kg), `height` (cm), and `gender`, for 80 physically active humans, together with seven body girth measurements (cm): `shoulder`, `hip`, `thigh`, `bicep`, `knee`, `ankle`, and `wrist`. There are no missing values in the dataset. The Aim of this Portfolio is to investigate using these body girth measurements in order to predict `weight` and `height`. Throughout this Portfolio, we will every opportunity to investigate the effect of gender on whatever is being investigated.

```
body_sample <- read.csv("body_sample.csv")
head(body_sample)
```

```
##   id age weight height gender shoulder   hip thigh bicep knee ankle wrist
## 1  1  21   65.6  174.0      m    106.2  93.5  51.5  32.5 34.5  23.5  16.5
## 2  5  22   78.8  187.2      m    107.5  98.5  55.4  32.0 37.7  24.4  18.0
## 3 14  26   74.6  176.0      m    113.0  98.0  59.1  35.6 35.8  21.5  16.6
```

```
## 4 17  30   93.8  192.7      m    112.2 105.0  65.8  37.0 40.9  24.2  17.8
## 5 18  22   70.0  171.5      m    120.0  90.1  54.1  31.2 36.4  22.0  17.1
## 6 21  22   78.8  176.0      m    116.0  98.0  57.5  32.0 37.5  21.0  17.3

summary(body_sample)

##        id             age           weight           height
##  Min.   :  1.0   Min.   :18.00   Min.   : 42.00   Min.   :151.1
##  1st Qu.:125.5   1st Qu.:22.00   1st Qu.: 56.58   1st Qu.:162.6
##  Median :244.5   Median :28.50   Median : 68.85   Median :170.2
##  Mean   :252.2   Mean   :29.91   Mean   : 68.98   Mean   :170.0
##  3rd Qu.:385.8   3rd Qu.:35.50   3rd Qu.: 76.65   3rd Qu.:176.0
##  Max.   :505.0   Max.   :53.00   Max.   :116.40   Max.   :192.7
##     gender             shoulder           hip             thigh
##  Length:80          Min.   : 87.00   Min.   : 78.80   Min.   :46.30
##  Class :character   1st Qu.: 99.95   1st Qu.: 91.15   1st Qu.:52.48
##  Mode  :character   Median :105.95   Median : 94.95   Median :55.00
##                     Mean   :107.74   Mean   : 96.76   Mean   :56.85
##                     3rd Qu.:116.35   3rd Qu.:101.67   3rd Qu.:60.20
##                     Max.   :134.80   Max.   :128.30   Max.   :75.70
##      bicep            knee           ankle           wrist
##  Min.   :23.20   Min.   :29.00   Min.   :17.90   Min.   :13.20
##  1st Qu.:27.40   1st Qu.:34.08   1st Qu.:21.00   1st Qu.:14.80
##  Median :30.60   Median :35.40   Median :21.75   Median :15.90
##  Mean   :30.87   Mean   :36.05   Mean   :21.95   Mean   :15.94
##  3rd Qu.:34.25   3rd Qu.:37.70   3rd Qu.:23.02   3rd Qu.:17.02
##  Max.   :40.30   Max.   :49.00   Max.   :27.00   Max.   :19.20
```

Task 1 — Multivariate Statistical Analysis

Firstly, we will sort the rows of the dataset, first by gender and then weight (within gender).

```
sorted_body_sample <- body_sample %>%
  arrange(gender, weight)
head(sorted_body_sample)

##    id age weight height gender shoulder  hip thigh bicep knee ankle wrist
## 1 261  29   42.0  153.4      f     88.7 80.9  48.8  24.0 30.8  17.9  13.2
## 2 381  20   43.2  160.0      f     92.7 78.8  46.3  23.2 31.6  18.6  13.8
## 3 282  21   45.8  152.0      f     87.0 86.0  51.0  24.5 31.5  20.0  14.0
## 4 266  19   47.8  157.0      f     90.1 88.5  54.0  24.6 29.0  19.0  13.2
## 5 505  33   48.6  160.7      f     91.9 86.9  51.8  27.4 34.4  21.2  15.5
## 6 385  28   48.8  160.0      f     92.0 86.0  53.0  24.1 30.0  20.4  14.4
```

(1). Using R to carry out Principal Component Analysis (PCA) using only the seven body girth measurements.

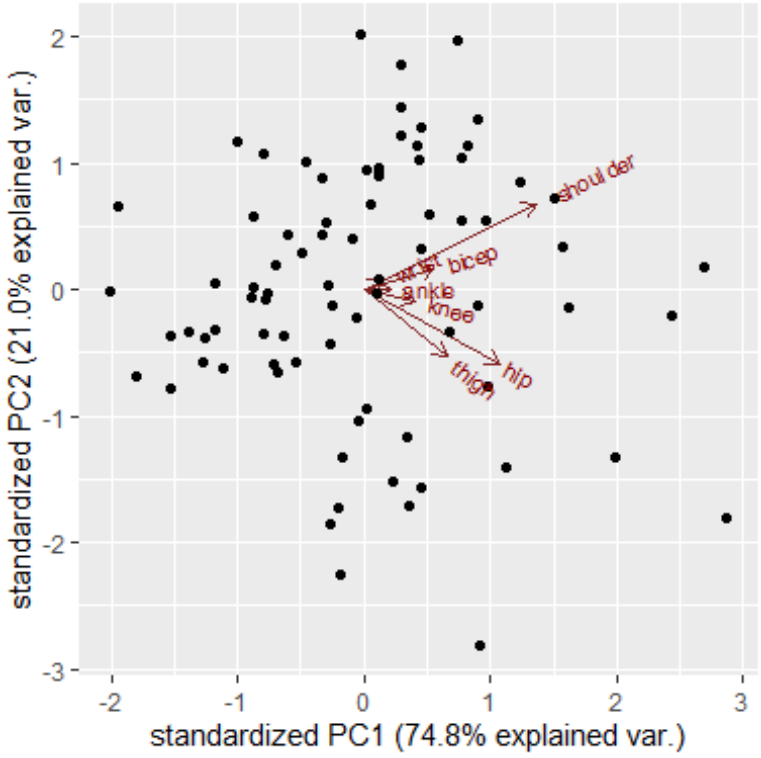
```
body_sample.pca <- prcomp(sorted_body_sample[, 6:12])
summary(body_sample.pca)
```

```
## Importance of components:
##                             PC1    PC2     PC3     PC4     PC5     PC6
PC7
## Standard deviation     13.814 7.3134 2.04141 1.77320 1.50545 0.98278
0.50347
## Proportion of Variance  0.748 0.2097 0.01634 0.01233 0.00888 0.00379
0.00099
## Cumulative Proportion   0.748 0.9577 0.97401 0.98634 0.99522 0.99901
1.00000
```

Including Plots

screeplot, biplot, and loadings plot (loadings variable is included in the biplot)

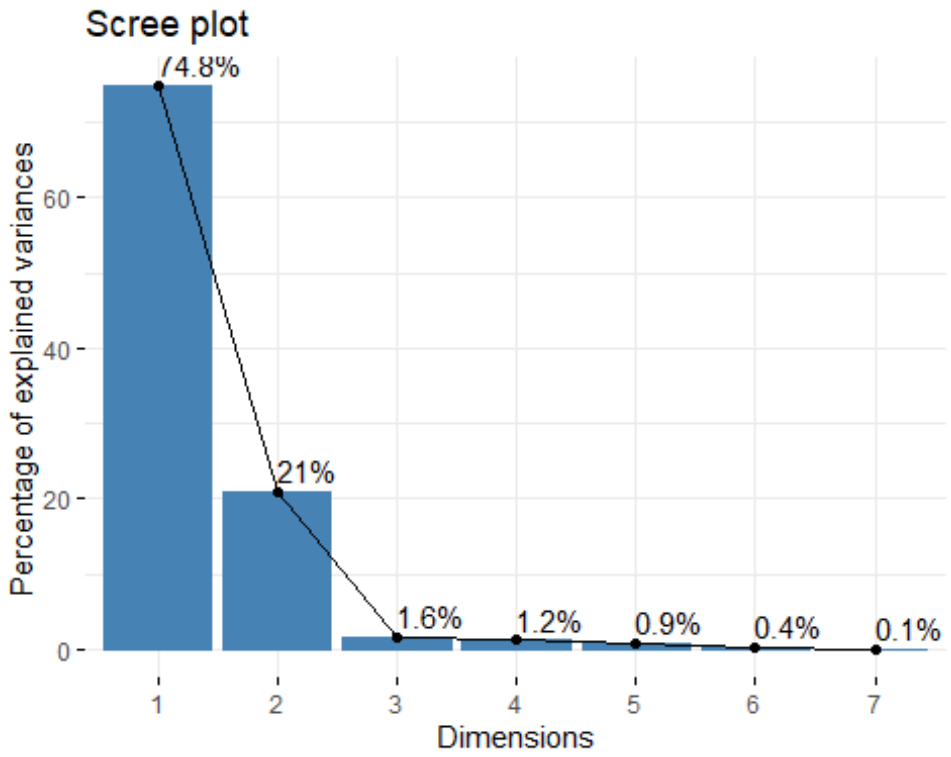
```
body_sample.pca %>%
  ggbiplot()
```



The selected seven body girth measurements have resulted to 7 principal components, that is, PC1 - PC7 each explaining the percentage of variation in the dataset. standardized PC1 explained 74.8% of total variance, PC2 explained 21% hence PC1 and PC2 can explain 95.8% of the variance.

`shoulder` and `hip` variables all contributes to PC1 the arrows indicated in the biplot shows these variables moving the samples to the right of the plot past 1 vertically.

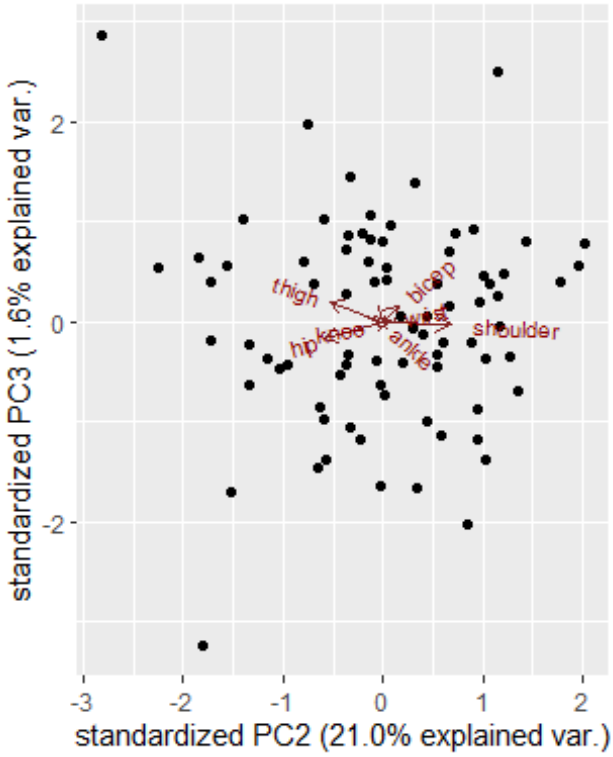
```
# Visualize eigenvalues/variances
fviz_screeplot(body_sample.pca, addlabels = TRUE)
```



From the scree plot it is evident that PC1 contributes 74.8% followed by PC2 which contributes 21%.

a biplot using PC2 and PC3 as the axes

```
body_sample.pca %>%
  ggbiplot(choices = 2:3)
```



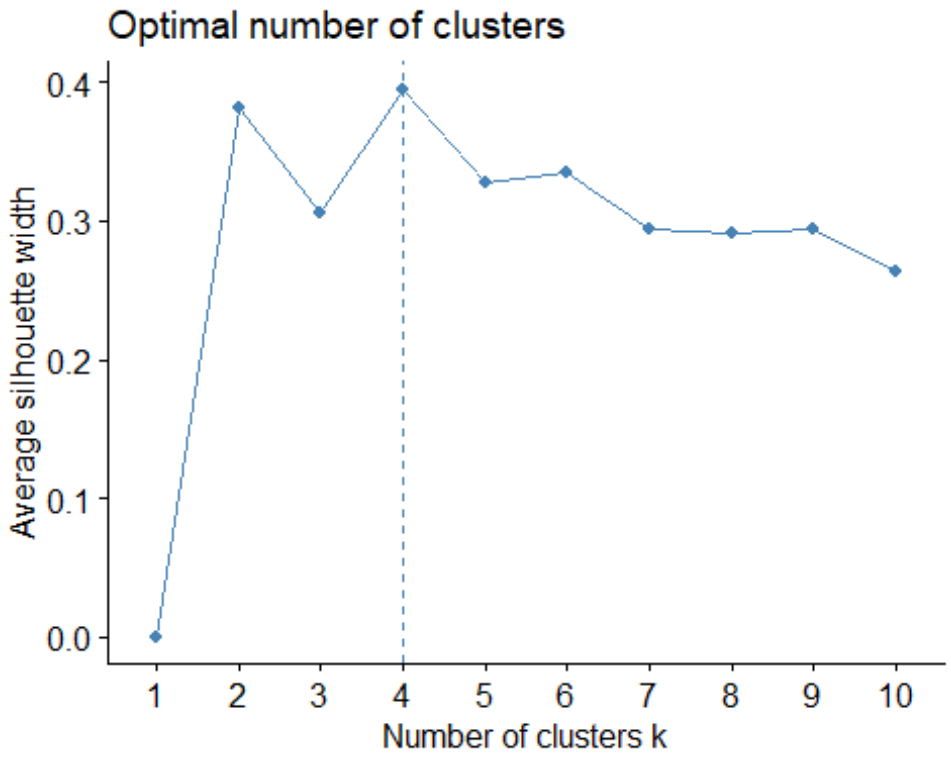
`bicep`, `shoulder`, `ankle` and `wrist` contribute to PC2, with higher values of these variables moving the samples to the right of PC2 vs PC3 biplot. `hip` is contributing to PC3

From both plots, we can see that the `shoulder` variable is contributing to both PC1 and PC2, the `hip` variable is contributing to PC1 and PC3. other variables have smaller contribution to PCs

(2). Carrying out Cluster Analysis on the dataset using hierarchical clustering on the seven body girth measurements.

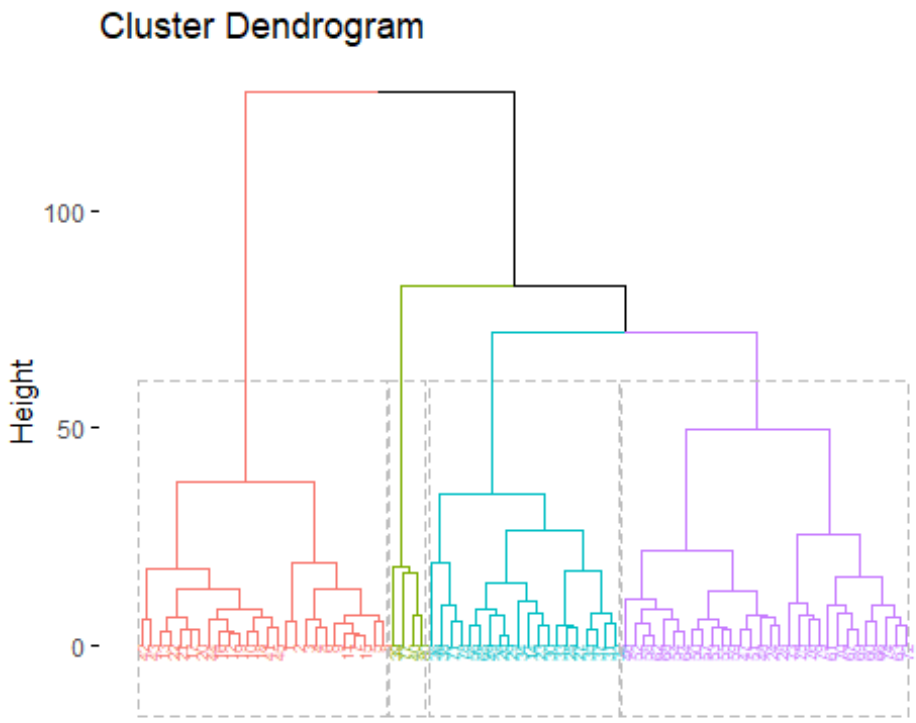
Before performing the cluster analysis, we need to find the optimal number of clusters that will be used to cut the tree.

```
# Determine the optimal number of clusters and plot
sorted_body_sample[,6:12] %>%
  fviz_nbclust(kmeans, method = "silhouette")
```



4 is the optimal number of cluster, we will use it to compute hierarchical clustering and cut tree.Since we are performing based on the seven body girth measurements, we will not scale the data because all measurements are of one unit of measurement.

```
# Compute hierarchical clustering on seven body girth measurements
body_sample.hcut <- hcut(sorted_body_sample[,6:12], k=4)

# Visualize the dendogram
fviz_dend(
  body_sample.hcut,
  rect = T,
  cex = .4
)
```

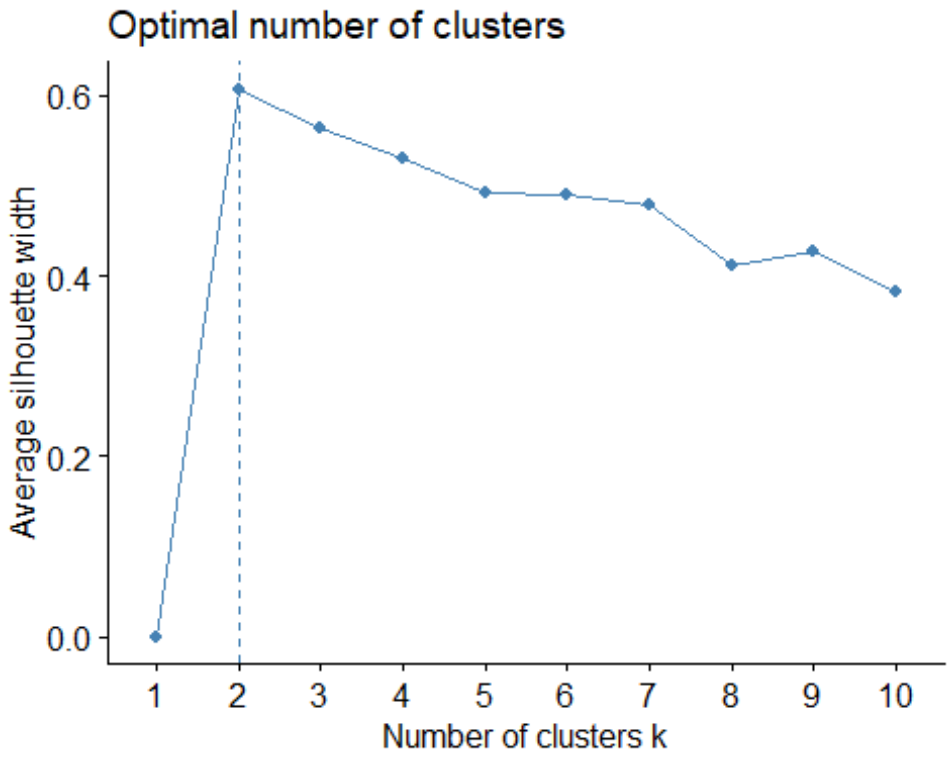


`hcut()` function computes hierarchical clustering and cut the tree into specified clusters (2 by default). Here we used 4 clusters, the rounded rectangles are the yielded clusters, by default `hcut()` used the *eucliden* distance metric.

Cluster the people and body girth measurements separately

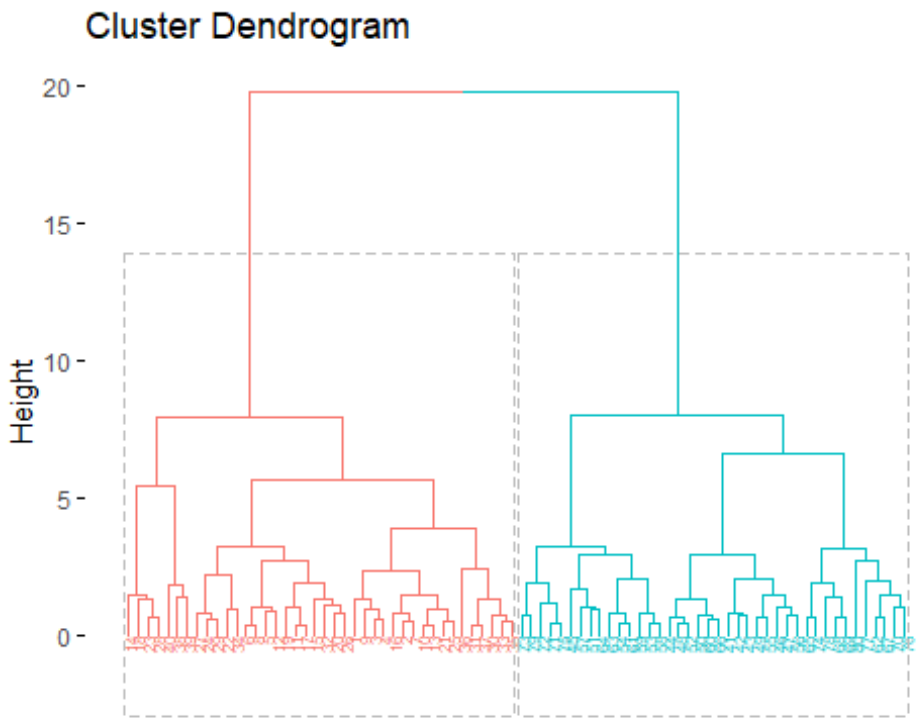
We have already created cluster for body girth measurements, below we will cluster the *people*. We will first check for the optimal clusters

```
# Determine the optimal number of clusters and plot
sorted_body_sample[,1:5] %>%
  # Encode gender with an integer 1 for male and 2 for females
  mutate(gender_int=ifelse(gender=="m",1,2))%>%
  # remove the text gender columns
  select(-gender)%>%
  # Change gender column to factor
  fviz_nbclust(kmeans, method = "silhouette")
```



2 is the optimal number of clusters.

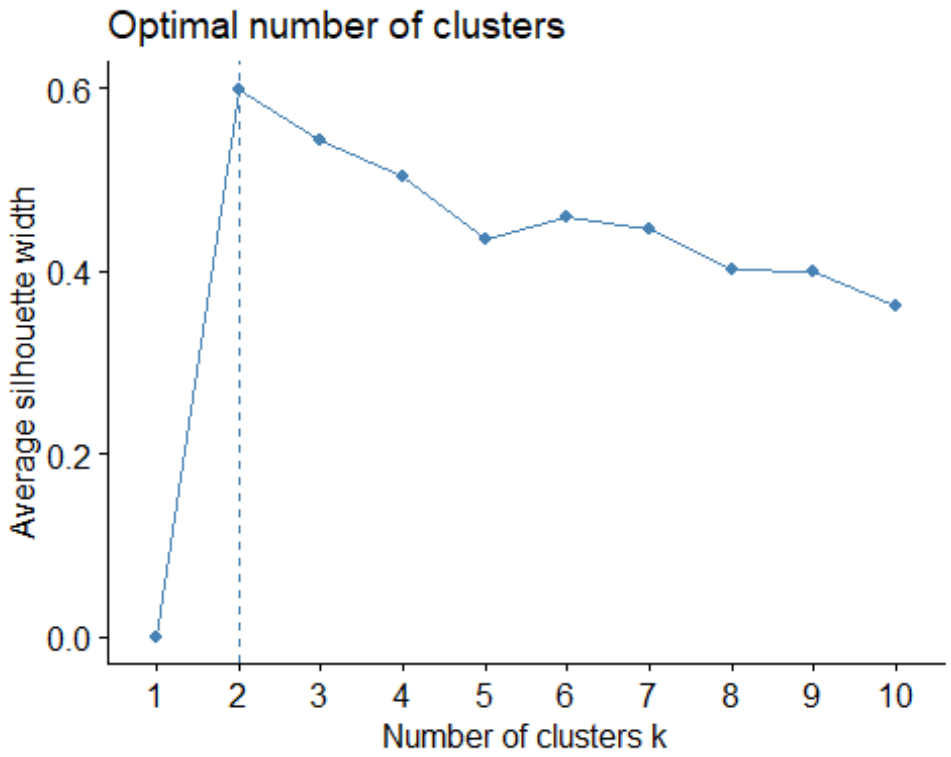
```
# Compute hierarchical clustering on seven body girth measurements
body_sample.hcut <- # Determine the optimal number of clusters and plot
sorted_body_sample[,1:5] %>%
  # Encode gender with an integer 1 for male and 2 for females
  mutate(gender_int=ifelse(gender=="m",1,2))%>%
  # remove the text gender columns
  select(-gender)%>%
  hcut(k=2, stand=T)

# Visualize the dendogram
fviz_dend(
  body_sample.hcut,
  rect = T,
  cex = .4
)
```



This has resulted to 2 clusters, this may be based on a persons gender, or age group with a 20 distance of merge.

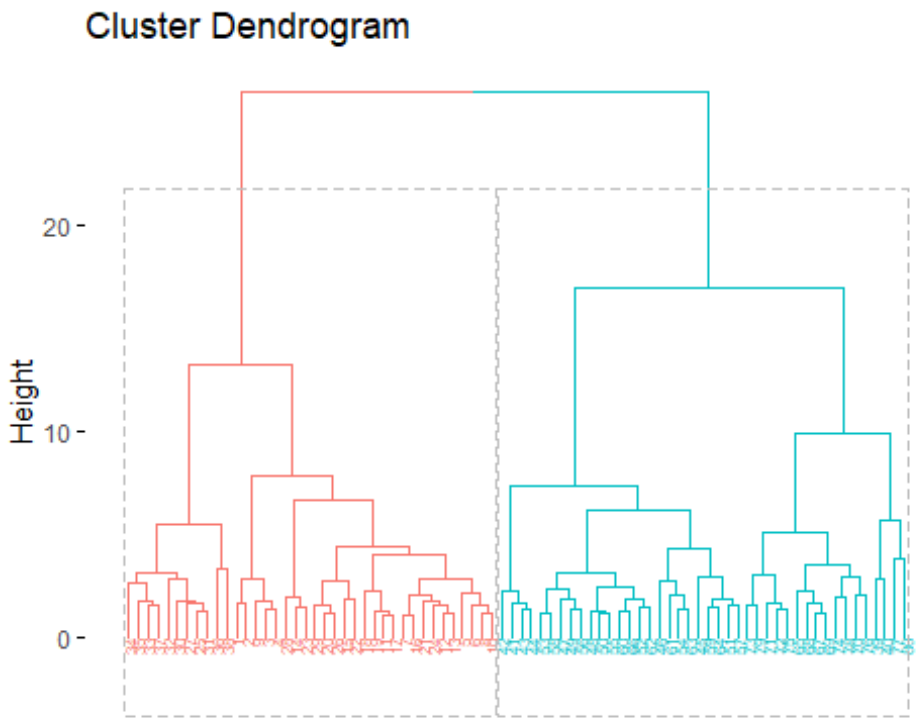
Cluster analysis including age, weight, height

```
# Determine the number of clusters
sorted_body_sample %>%
  # Encode gender with an integer 1 for male and 2 for females
  mutate(gender_int=ifelse(gender=="m",1,2))%>%
  # remove the text gender columns
  select(-gender)%>%
  # Change gender column to factor
  fviz_nbclust(kmeans, method = "silhouette")
```



Using all variables results to 2 clusters, we will use the optimal clusters to plot the hierarchical clustering.

```
# Compute hierarchical clustering on seven body girth measurements
body_sample.hcut <- sorted_body_sample %>%
  # Encode gender with an integer 1 for male and 2 for females
  mutate(gender_int=ifelse(gender=="m",1,2))%>%
  # remove the text gender columns
  select(-gender)%>%
  hcut(k=2, stand=T)

# Visualize the dendogram
fviz_dend(
  body_sample.hcut,
  rect = T,
  cex = .4
)
```



From the dendogram, we can see the groupings generated with 25 distance of split. From all the clustering, it is evident that with seven body girth measurement gives 4 clustering the distance of merge for the seven body girth measurements was 150.

(3) assess the methods applied and insights gained

The PCA, Principal Component Analysis assisted in visualizing the variations present in body girth measurements, from PCA the 1 dimensions contributed over 74%, the variables resulting to this where `shoulder`, `hip` and `thigh` contributing the highest variations.

Clustering helped in identifying the groupings available in the dataset. This was the findings:

- 4 groups: with the seven body girth
- 2 groups: the people
- 2 groups using all the variables

With all these, it is evident that there can be a hidden pattern in the dataset that is formed with as a result of the variables. there are 4 different groups that have different characteristics formed by the seven body girth measurements.

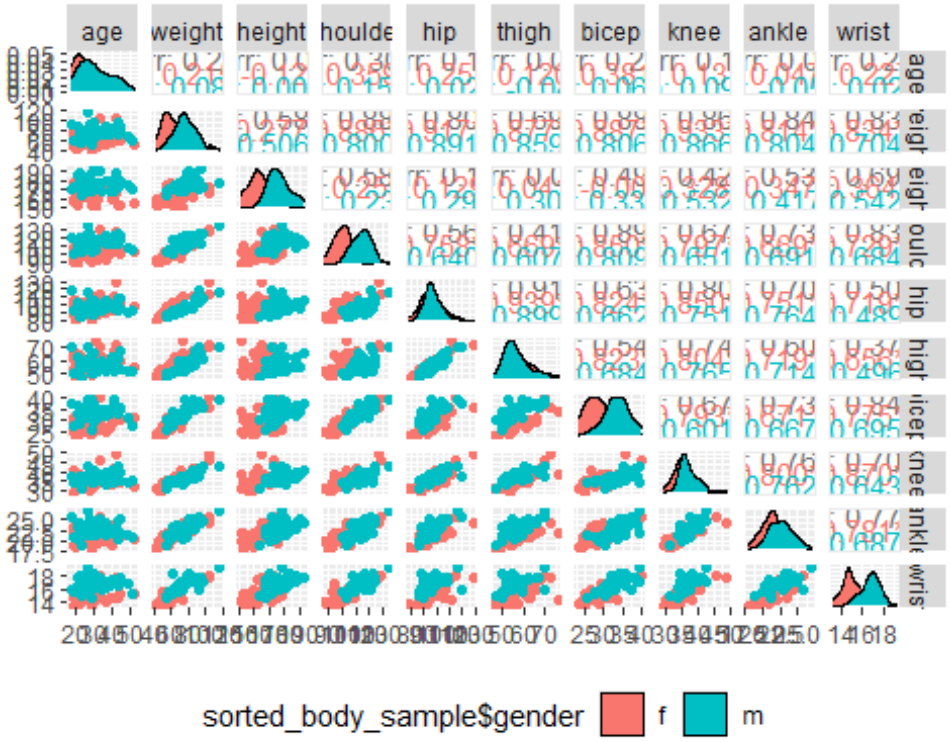
Task 2— Exploratory Data Analysis and Linear Models

The aim is to use the seven body girth measurements to predict body weight and body height using linear models (regression) in R. We are primarily concerned with critically

assessing any linear models proposed, and with *model selection* (which predictors to include in any final linear models recommended).

(2) Using R to build a *scatter matrix* using `ggpairs()`.

```
sorted_body_sample %>%
  select(-id, -gender) %>%
  ggpairs(
    aes(colour=sorted_body_sample$gender),
    progress = FALSE,
    legend=1
  ) +
  theme(legend.position = "bottom")
```



From the scatter matrix above we can see that:

1. There is a strong correlation between:
- `thigh` and `hip`: 0.912
- `wrist` and `ankle`: 0.775
- `wrist` and `knee`: 0.7
- `wrist` and `bisep`: 0.842
- `wrist` and `shoulder`: 0.836
- `wrist` and `weight`: 0.83
- `ankle` and `weight`: 0.843 among others

2. Week Correlations is witnessed between:
- `height` and `age`

- hip and height
- thigh and height

Looking at the hip and shoulder relationship, we can see a moderate positive correlation of 0.566 but female have a stronger positive correlation of hip-shoulder of 0.758 as compared to males whose correlation is 0.640. It is not sufficiently enough to state that hip and shoulder can sufficient to identify gender.

(2) single-predictor linear model "best" predicts body weight and which single-predictor linear model "best" predicts body height

```
# Create model to predict weight using body girth measurements
lr_weight <- lm(
  weight ~ shoulder + hip + thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)

# Create model to predict height using body girth measurements
lr_height <- lm(
  height ~ shoulder + hip + thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)

# Check the summaries of the models
summary(lr_weight)

##
## Call:
## lm(formula = weight ~ shoulder + hip + thigh + bicep + knee +
##     ankle + wrist, data = sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3325  -2.1335   0.4357   2.3820   8.4240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -108.1439     6.7898 -15.927  < 2e-16 ***
## shoulder       0.4534     0.0983   4.613 1.69e-05 ***
## hip            0.3642     0.1332   2.734  0.00787 **
## thigh          0.1746     0.1845   0.946  0.34712
## bicep          0.5550     0.2636   2.105  0.03877 *
## knee           0.8059     0.2650   3.042  0.00328 **
## ankle          0.6913     0.4189   1.650  0.10325
## wrist          1.3642     0.7180   1.900  0.06145 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.426 on 72 degrees of freedom
```

```
## Multiple R-squared:  0.9536, Adjusted R-squared:  0.9491
## F-statistic: 211.6 on 7 and 72 DF,  p-value: < 2.2e-16

summary(lr_height)

##
## Call:
## lm(formula = height ~ shoulder + hip + thigh + bicep + knee +
##     ankle + wrist, data = sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0455  -3.3239  -0.3182   4.5360  17.4880
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.7777    12.9689   7.385 2.14e-10 ***
## shoulder      0.1748     0.1878   0.931  0.35487
## hip          -0.1229     0.2544  -0.483  0.63043
## thigh        -0.4467     0.3524  -1.268  0.20897
## bicep        -0.4852     0.5036  -0.964  0.33850
## knee          0.4277     0.5061   0.845  0.40085
## ankle         1.2181     0.8001   1.522  0.13228
## wrist         4.1070     1.3715   2.995  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.545 on 72 degrees of freedom
## Multiple R-squared:  0.5653, Adjusted R-squared:  0.523
## F-statistic: 13.38 on 7 and 72 DF,  p-value: 6.288e-11
```

To select single-predictor linear model that best predicts body weight and best predict linear model we will perform an all possible regression that involves all subset regression test for all the seven body girth measurements and select one single predictor that has the largest R^2 and also small Mean Squared Error.

Single predictorlinear model which best predicts body weight.

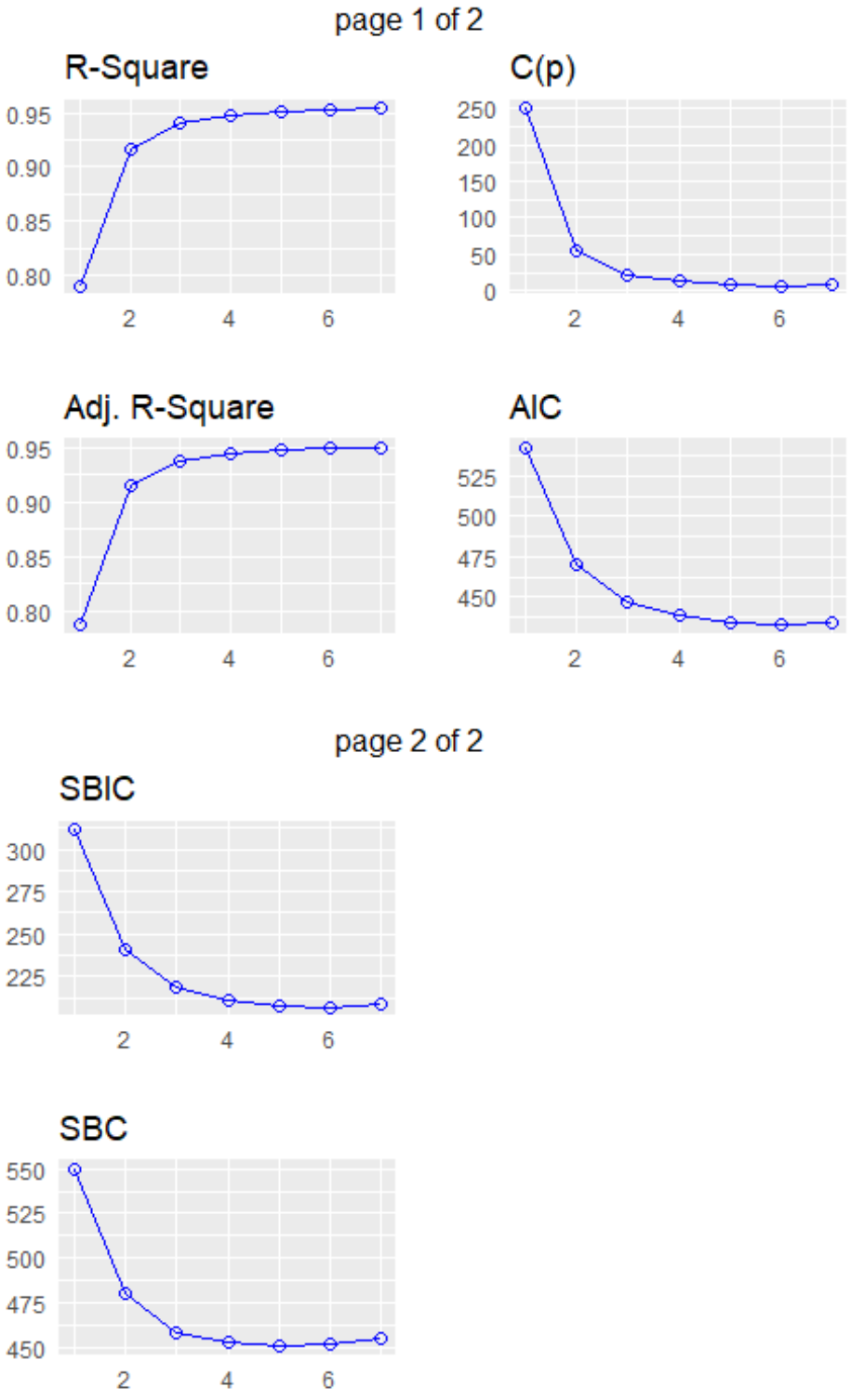
The model used is `lr_weight`

```
best_predictor_weight <- lr_weight %>%
  ols_step_best_subset()
best_predictor_weight

##                 Best Subsets Regression
## ---------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------
##      1         bicep
##      2         shoulder hip
##      3         shoulder hip wrist
```

```
##      4         shoulder hip bicep knee
##      5         shoulder hip bicep knee ankle
##      6         shoulder hip bicep knee ankle wrist
##      7         shoulder hip thigh bicep knee ankle wrist
## ----------------------------------------------------------
##
##                                                        Subsets Regression
Summary
## ------------------------------------------------------------------------
----------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)        AIC
SBIC        SBC         MSEP         FPE       HSP       APC
## ------------------------------------------------------------------------
----------------------------------------------------------
##   1        0.7902      0.7875      0.7754    249.8809    542.4300
311.2432    549.5761    3923.9363    50.2750    0.6370    0.2206
##   2        0.9166      0.9144      0.9086     55.5893    470.6571
241.0381    480.1852    1580.9183    20.4989    0.2600    0.0899
##   3        0.9399      0.9376      0.9312     21.2691    446.3466
218.1378    458.2567    1153.0029    15.1279    0.1921    0.0664
##   4        0.9469      0.9441      0.9343     12.4403    438.4734
211.1455    452.7656    1032.9085    13.7111    0.1744    0.0602
##   5        0.9512      0.9479      0.9391      7.7691    433.7227
207.3564    450.3969     962.3277    12.9221    0.1647    0.0567
##   6        0.9531      0.9492      0.9392      6.8956    432.6297
206.9450    451.6860     938.6913    12.7489    0.1628    0.0559
##   7        0.9536      0.9491      0.9375      8.0000    433.6407
208.3637    455.0790     940.2166    12.9139    0.1654    0.0567
## ------------------------------------------------------------------------
----------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

The single-predict linear model that best predict body weight will have `bicep` as the independent predictor. The plot below shows how fit criterion was done.

```
plot(best_predictor_weight)
```



Single predictorlinear model which best predicts body heigth

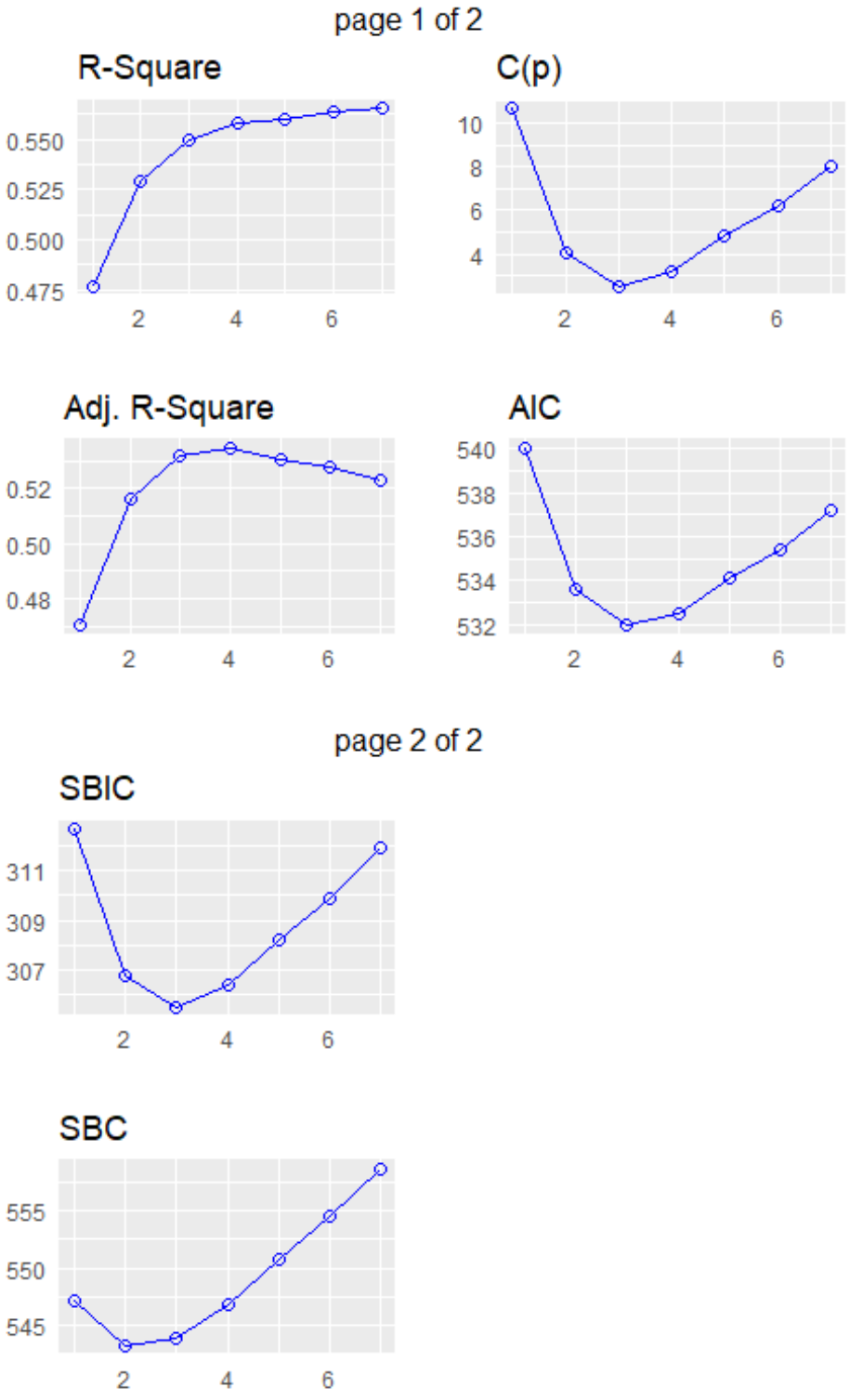
The model used is `lr_height`

```
best_predictor_height <- lr_height %>%
  ols_step_best_subset()
best_predictor_height
```

```
##                 Best Subsets Regression
## ---------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------
##      1         wrist
##      2         thigh wrist
##      3         thigh ankle wrist
##      4         thigh knee ankle wrist
##      5         shoulder thigh bicep ankle wrist
##      6         shoulder thigh bicep knee ankle wrist
##      7         shoulder hip thigh bicep knee ankle wrist
## ---------------------------------------------------------
##
##                                                    Subsets Regression
Summary
## ------------------------------------------------------------------------
----------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC
SBC         MSEP         FPE       HSP       APC
## ------------------------------------------------------------------------
----------------------------------------------------------
##   1        0.4769      0.4702       0.448    10.6407    539.9910
312.6425    547.1371    3806.1107    48.7654    0.6179    0.5499
##   2        0.5288      0.5165      0.4884     4.0529    533.6404
306.7587    543.1685    3473.9673    45.0450    0.5713    0.5080
##   3        0.5499      0.5321      0.5034     2.5582    531.9758
305.5189    543.8860    3362.6703    44.1196    0.5602    0.4975
##   4        0.5581      0.5346      0.4833     3.1889    532.4930
306.3761    546.7852    3345.5236    44.4093    0.5648    0.5008
##   5        0.5603      0.5306      0.4809     4.8215    534.0904
308.2237    550.7646    3374.3278    45.3103    0.5774    0.5109
##   6        0.5639      0.5281      0.4538     6.2335    535.4418
309.8937    554.4980    3393.5667    46.0899    0.5886    0.5197
##   7        0.5653      0.5230      0.4313     8.0000    537.1828
311.9057    558.6210    3430.2402    47.1144    0.6033    0.5313
## ------------------------------------------------------------------------
----------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

The single-predict linear model that best predict body weight will have `wrist` as the independent predictor. The plot below shows how fit criterion was done.

```
plot(best_predictor_height)
```



(3) AIC

The acronym AIC stands for Akaike information criterion, this is a regression metric used for comparing how well several regression models approximate a target function (or fit). The best model explains the highest variation using very few features.

Considering only the seven body girth measurements, we will perform the `ols_step_subest` which selects the subset of predictors that do best fit wile having largest R^2 or smallest mean squared error, we will use it to determine what two-predictor and four-predictor linear models to recommend to predict body weight and body height.

```
# Create a multitarget linear model with seven body girth measurements
lr_weight_model <- lm(
  weight ~ shoulder + hip + thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)
lr_height_model <- lm(
  height ~ shoulder + hip + thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)
# print summaries of the models
summary(lr_weight_model)
```

```
##
## Call:
## lm(formula = weight ~ shoulder + hip + thigh + bicep + knee +
##     ankle + wrist, data = sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3325  -2.1335   0.4357   2.3820   8.4240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -108.1439     6.7898 -15.927  < 2e-16 ***
## shoulder       0.4534     0.0983   4.613 1.69e-05 ***
## hip            0.3642     0.1332   2.734  0.00787 **
## thigh          0.1746     0.1845   0.946  0.34712
## bicep          0.5550     0.2636   2.105  0.03877 *
## knee           0.8059     0.2650   3.042  0.00328 **
## ankle          0.6913     0.4189   1.650  0.10325
## wrist          1.3642     0.7180   1.900  0.06145 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.426 on 72 degrees of freedom
## Multiple R-squared:  0.9536, Adjusted R-squared:  0.9491
## F-statistic: 211.6 on 7 and 72 DF,  p-value: < 2.2e-16
```

```
summary(lr_height_model)
```

```
##
## Call:
## lm(formula = height ~ shoulder + hip + thigh + bicep + knee +
##     ankle + wrist, data = sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0455  -3.3239  -0.3182   4.5360  17.4880
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.7777    12.9689   7.385 2.14e-10 ***
## shoulder      0.1748     0.1878   0.931  0.35487
## hip          -0.1229     0.2544  -0.483  0.63043
## thigh        -0.4467     0.3524  -1.268  0.20897
## bicep        -0.4852     0.5036  -0.964  0.33850
## knee          0.4277     0.5061   0.845  0.40085
## ankle         1.2181     0.8001   1.522  0.13228
## wrist         4.1070     1.3715   2.995  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.545 on 72 degrees of freedom
## Multiple R-squared:  0.5653, Adjusted R-squared:  0.523
## F-statistic: 13.38 on 7 and 72 DF,  p-value: 6.288e-11
```

```
lr_weight_model %>%
  ols_step_both_aic(
    details = T
  )
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1 . shoulder
## 2 . hip
## 3 . thigh
## 4 . bicep
## 5 . knee
## 6 . ankle
## 7 . wrist
##
##  Step 0: AIC = 665.3484
##  weight ~ 1
##
##
## Variables Entered/Removed:
##
```

```
##                             Enter New Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## bicep         1    542.430    14407.481    3825.807    0.790        0.787
## shoulder      1    547.534    14155.454    4077.834    0.776        0.773
## knee          1    559.053    13523.894    4709.394    0.742        0.738
## ankle         1    567.930    12971.236    5262.052    0.711        0.708
## wrist         1    571.693    12717.843    5515.445    0.698        0.694
## hip           1    583.006    11880.057    6353.231    0.652        0.647
## thigh         1    613.726     8905.735    9327.553    0.488        0.482
## ----------------------------------------------------------------------
##
## - bicep added
##
##
##  Step 1 : AIC = 542.43
##  weight ~ bicep
##
##                             Enter New Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## knee          1    472.876    16669.128    1564.160    0.914        0.912
## hip           1    492.641    16230.766    2002.522    0.890        0.887
## ankle         1    506.691    15846.312    2386.976    0.869        0.866
## thigh         1    514.466    15602.678    2630.610    0.856        0.852
## shoulder      1    529.436    15061.342    3171.946    0.826        0.822
## wrist         1    533.868    14880.645    3352.643    0.816        0.811
## ----------------------------------------------------------------------
##
## - knee added
##
##
##  Step 2 : AIC = 472.8764
##  weight ~ bicep + knee
##
##                          Remove Existing Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## knee          1    542.430    14407.481    3825.807    0.790        0.787
## bicep         1    559.053    13523.894    4709.394    0.742        0.738
## ----------------------------------------------------------------------
##
##                             Enter New Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## shoulder      1    460.054    16933.670    1299.618    0.929        0.926
```

```
## ankle          1    462.824    16887.878    1345.410    0.926        0.923
## hip            1    462.902    16886.569    1346.719    0.926        0.923
## thigh          1    472.432    16716.205    1517.083    0.917        0.914
## wrist          1    473.178    16701.985    1531.303    0.916        0.913
## ---------------------------------------------------------------------------
##
## - shoulder added
##
##
##  Step 3 : AIC = 460.0541
##  weight ~ bicep + knee + shoulder
##
##                          Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## shoulder       1    472.876    16669.128    1564.160    0.914        0.912
## bicep          1    478.856    16547.742    1685.546    0.908        0.905
## knee           1    529.436    15061.342    3171.946    0.826        0.822
## ---------------------------------------------------------------------------
##
##                             Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## hip            1    438.473    17265.447     967.841    0.947        0.944
## thigh          1    446.405    17164.580    1068.708    0.941        0.938
## ankle          1    452.553    17079.211    1154.077    0.937        0.933
## wrist          1    461.995    16934.629    1298.659    0.929        0.925
## ---------------------------------------------------------------------------
##
## - hip added
##
##
##  Step 4 : AIC = 438.4734
##  weight ~ bicep + knee + shoulder + hip
##
##                          Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## bicep          1    451.310    17068.228    1165.060    0.936        0.934
## hip            1    460.054    16933.670    1299.618    0.929        0.926
## knee           1    462.342    16895.959    1337.329    0.927        0.924
## shoulder       1    462.902    16886.569    1346.719    0.926        0.923
## ---------------------------------------------------------------------------
##
##                             Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS      R-Sq     Adj. R-Sq
```

```
## ------------------------------------------------------------------------
## ankle         1    433.723    17343.767    889.521    0.951        0.948
## wrist         1    433.906    17341.723    891.565    0.951        0.948
## thigh         1    440.366    17266.745    966.543    0.947        0.943
## ------------------------------------------------------------------------
##
## - ankle added
##
##
##  Step 5 : AIC = 433.7227
##  weight ~ bicep + knee + shoulder + hip + ankle
##
##                         Remove Existing Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## ankle         1    438.473    17265.447     967.841    0.947        0.944
## bicep         1    445.747    17173.326    1059.962    0.942        0.939
## knee          1    451.291    17097.262    1136.026    0.938        0.934
## hip           1    452.553    17079.211    1154.077    0.937        0.933
## shoulder      1    454.521    17050.459    1182.829    0.935        0.932
## ------------------------------------------------------------------------
##
##                           Enter New Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## wrist         1    432.630    17377.501    855.787    0.953        0.949
## thigh         1    435.554    17345.638    887.650    0.951        0.947
## ------------------------------------------------------------------------
##
## - wrist added
##
##
##  Step 6 : AIC = 432.6297
##  weight ~ bicep + knee + shoulder + hip + ankle + wrist
##
##                         Remove Existing Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## wrist         1    433.723    17343.767     889.521    0.951        0.948
## ankle         1    433.906    17341.723     891.565    0.951        0.948
## bicep         1    438.435    17289.802     943.486    0.948        0.945
## knee          1    443.320    17230.391    1002.897    0.945        0.941
## shoulder      1    451.287    17125.374    1107.914    0.939        0.935
## hip           1    454.154    17084.946    1148.342    0.937        0.933
## ------------------------------------------------------------------------
##
##                           Enter New Variables
```

```
## ---------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------
## thigh         1    433.641    17388.016    845.272    0.954        0.949
## ---------------------------------------------------------------------
##
##
## No more variables to be added or removed.
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.976       RMSE                 3.424
## R-Squared                0.953       Coef. Var            4.964
## Adj. R-Squared           0.949       MSE                 11.723
## Pred R-Squared           0.939       MAE                  2.561
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                  Sum of
##                 Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    17377.501         6       2896.250    247.055    0.0000
## Residual        855.787        73         11.723
## Total         18233.288        79
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
--------------------
##       model        Beta    Std. Error    Std. Beta       t        Sig
lower      upper
## ----------------------------------------------------------------------------
--------------------
## (Intercept)    -106.645         6.598                -16.164    0.000
-119.794    -93.496
##       bicep       0.657         0.240        0.186      2.735    0.008
0.178      1.136
##        knee       0.887         0.250        0.194      3.542    0.001
0.388      1.386
##    shoulder       0.415         0.090        0.288      4.638    0.000
0.237      0.593
##         hip       0.456         0.091        0.257      4.996    0.000
0.274      0.638
```

```
##       ankle       0.728          0.417         0.086       1.747     0.085
-0.103       1.559
##       wrist       1.162          0.685         0.104       1.696     0.094
-0.203       2.527
## ----------------------------------------------------------------------------
--------------------

##
##
##                                  Stepwise Summary
## ----------------------------------------------------------------------------
------
## Variable     Method       AIC         RSS         Sum Sq       R-Sq
Adj. R-Sq
## ----------------------------------------------------------------------------
------
## bicep       addition    542.430    3825.807    14407.481    0.79017
0.78748
## knee        addition    472.876    1564.160    16669.128    0.91421
0.91199
## shoulder    addition    460.054    1299.618    16933.670    0.92872
0.92591
## hip         addition    438.473     967.841    17265.447    0.94692
0.94409
## ankle       addition    433.723     889.521    17343.767    0.95121
0.94792
## wrist       addition    432.630     855.787    17377.501    0.95306
0.94921
## ----------------------------------------------------------------------------
------
```

- The two- predictors for weight are:`bicep` + `knee` with an AIC of 472.8764
- The four-predictors for weight are: `bicep` + `knee` + `shoulder` + `hip` with an AIC of 438.4734

```
lr_height_model %>%
  ols_step_both_aic(details = T)

## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1 . shoulder
## 2 . hip
## 3 . thigh
## 4 . bicep
## 5 . knee
## 6 . ankle
## 7 . wrist
##
```

```
##  Step 0: AIC = 589.8319
##  height ~ 1
##
##
## Variables Entered/Removed:
##
##                            Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## wrist         1    539.991    3383.384    3710.928    0.477        0.470
## shoulder      1    558.707    2405.231    4689.081    0.339        0.331
## ankle         1    564.638    2044.395    5049.917    0.288        0.279
## bicep         1    569.014    1760.450    5333.862    0.248        0.239
## knee          1    575.945    1277.751    5816.561    0.180        0.170
## hip           1    589.127     235.857    6858.455    0.033        0.021
## thigh         1    591.643      16.776    7077.536    0.002       -0.010
## ---------------------------------------------------------------------------
##
## - wrist added
##
##
##  Step 1 : AIC = 539.991
##  height ~ wrist
##
##                            Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## thigh         1    533.640    3751.209    3343.102    0.529        0.517
## hip           1    536.077    3647.818    3446.494    0.514        0.502
## bicep         1    538.284    3551.438    3542.874    0.501        0.488
## knee          1    540.738    3441.057    3653.255    0.485        0.472
## shoulder      1    541.980    3383.897    3710.415    0.477        0.463
## ankle         1    541.990    3383.417    3710.895    0.477        0.463
## ---------------------------------------------------------------------------
##
## - thigh added
##
##
##  Step 2 : AIC = 533.6404
##  height ~ wrist + thigh
##
##                         Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## thigh         1    539.991    3383.384    3710.928    0.477        0.470
## wrist         1    591.643      16.776    7077.536    0.002       -0.010
## ---------------------------------------------------------------------------
```

```
##
##                           Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## ankle         1    531.976    3900.893    3193.419    0.550        0.532
## knee          1    533.421    3842.663    3251.649    0.542        0.524
## bicep         1    535.102    3773.647    3320.665    0.532        0.513
## shoulder      1    535.149    3771.690    3322.622    0.532        0.513
## hip           1    535.496    3757.219    3337.093    0.530        0.511
## ---------------------------------------------------------------------------
##
## - ankle added
##
##
##  Step 3 : AIC = 531.9758
##  height ~ wrist + thigh + ankle
##
##                         Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## ankle         1    533.640    3751.209    3343.102    0.529        0.517
## thigh         1    541.990    3383.417    3710.895    0.477        0.463
## wrist         1    552.085    2884.334    4209.977    0.407        0.391
## ---------------------------------------------------------------------------
##
##                           Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## knee          1    532.493    3959.538    3134.774    0.558        0.535
## bicep         1    533.422    3922.941    3171.371    0.553        0.529
## shoulder      1    533.835    3906.492    3187.820    0.551        0.527
## hip           1    533.971    3901.076    3193.236    0.550        0.526
## ---------------------------------------------------------------------------
##
##
## No more variables to be added or removed.
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.742       RMSE                 6.482
## R-Squared               0.550       Coef. Var            3.814
## Adj. R-Squared          0.532       MSE                 42.019
## Pred R-Squared          0.503       MAE                  5.014
## ----------------------------------------------------------------
```

```
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression    3900.893         3       1300.298    30.946    0.0000
## Residual      3193.419        76         42.019
## Total         7094.312        79
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ------------------------------------------------------------------------
----------------
##       model       Beta    Std. Error    Std. Beta      t        Sig
lower      upper
## ------------------------------------------------------------------------
----------------
## (Intercept)    101.727         9.598                 10.599    0.000
82.611    120.844
##       wrist      4.242         0.862        0.609     4.919    0.000
2.524      5.959
##       thigh     -0.540         0.154       -0.344    -3.509    0.001    -
0.847     -0.234
##       ankle      1.428         0.756        0.271     1.887    0.063    -
0.079      2.934
## ------------------------------------------------------------------------
----------------

##
##
##                                  Stepwise Summary
## ------------------------------------------------------------------------
-----
## Variable     Method       AIC        RSS        Sum Sq      R-Sq      Adj.
R-Sq
## ------------------------------------------------------------------------
-----
## wrist       addition    539.991    3710.928    3383.384    0.47692
0.47021
## thigh       addition    533.640    3343.102    3751.209    0.52876
0.51652
## ankle       addition    531.976    3193.419    3900.893    0.54986
0.53209
## ------------------------------------------------------------------------
-----
```

- The two- predictors for height are: `wrist` + `thigh` with an AIC of 533.6404
- The four-predictors for height are: **No more variables were to be added or removed.** hence stopped at 3.

Compare these models with the best linear models using body girth measurements from the legs and arms only, i.e., not including `shoulder` and `hip`.

We will create more models covering only 5 features, without including `shoulder` and `hip`.

```
# New weight model
lr_weight_new <- lm(
  weight ~ thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)
# New height model
lr_height_new <- lm(
  height ~ thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)

## Summaries
summary(lr_weight_new)

##
## Call:
## lm(formula = weight ~ thigh + bicep + knee + ankle + wrist, data =
sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4481  -2.8752   0.3574   2.8384  12.0780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90.1145     7.3625 -12.240  < 2e-16 ***
## thigh         0.2907     0.1423   2.044  0.04456 *
## bicep         1.4907     0.2310   6.453 1.01e-08 ***
## knee          1.3396     0.3018   4.438 3.11e-05 ***
## ankle         1.3121     0.4907   2.674  0.00922 **
## wrist         1.2199     0.8656   1.409  0.16290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.146 on 74 degrees of freedom
## Multiple R-squared:  0.9302, Adjusted R-squared:  0.9255
## F-statistic: 197.3 on 5 and 74 DF,  p-value: < 2.2e-16

summary(lr_height_new)
```

```
##
## Call:
## lm(formula = height ~ thigh + bicep + knee + ankle + wrist, data =
sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8796  -3.9194  -0.3675   4.2039  16.5367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99.4063    11.5382   8.615 8.87e-13 ***
## thigh        -0.6459     0.2229  -2.897  0.00495 **
## bicep        -0.1801     0.3620  -0.498  0.62028
## knee          0.4994     0.4730   1.056  0.29457
## ankle         1.2903     0.7690   1.678  0.09757 .
## wrist         4.1726     1.3565   3.076  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.498 on 74 degrees of freedom
## Multiple R-squared:  0.5596, Adjusted R-squared:  0.5298
## F-statistic: 18.81 on 5 and 74 DF,  p-value: 5.103e-12
```

weight predictors using body girth measurements from legs and arms only

```
lr_weight_new %>%
  ols_step_both_aic(
    details = T
  )

## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1 . thigh
## 2 . bicep
## 3 . knee
## 4 . ankle
## 5 . wrist
##
##  Step 0: AIC = 665.3484
##  weight ~ 1
##
##
## Variables Entered/Removed:
##
##                              Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
```

```
## ---------------------------------------------------------------------
## bicep         1    542.430    14407.481    3825.807    0.790        0.787
## knee          1    559.053    13523.894    4709.394    0.742        0.738
## ankle         1    567.930    12971.236    5262.052    0.711        0.708
## wrist         1    571.693    12717.843    5515.445    0.698        0.694
## thigh         1    613.726     8905.735    9327.553    0.488        0.482
## ---------------------------------------------------------------------
##
## - bicep added
##
##
##  Step 1 : AIC = 542.43
##  weight ~ bicep
##
##                             Enter New Variables
## ---------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------
## knee          1    472.876    16669.128    1564.160    0.914        0.912
## ankle         1    506.691    15846.312    2386.976    0.869        0.866
## thigh         1    514.466    15602.678    2630.610    0.856        0.852
## wrist         1    533.868    14880.645    3352.643    0.816        0.811
## ---------------------------------------------------------------------
##
## - knee added
##
##
##  Step 2 : AIC = 472.8764
##  weight ~ bicep + knee
##
##                          Remove Existing Variables
## ---------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------
## knee          1    542.430    14407.481    3825.807    0.790        0.787
## bicep         1    559.053    13523.894    4709.394    0.742        0.738
## ---------------------------------------------------------------------
##
##                             Enter New Variables
## ---------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------
## ankle         1    462.824    16887.878    1345.410    0.926        0.923
## thigh         1    472.432    16716.205    1517.083    0.917        0.914
## wrist         1    473.178    16701.985    1531.303    0.916        0.913
## ---------------------------------------------------------------------
##
## - ankle added
##
##
```

```
##  Step 3 : AIC = 462.8244
##  weight ~ bicep + knee + ankle
##
##                          Remove Existing Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## ankle         1    472.876    16669.128    1564.160    0.914        0.912
## knee          1    506.691    15846.312    2386.976    0.869        0.866
## bicep         1    530.315    15026.319    3206.969    0.824        0.820
## ----------------------------------------------------------------------
##
##                             Enter New Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## thigh         1    462.462    16927.019    1306.269    0.928        0.925
## wrist         1    464.735    16889.376    1343.912    0.926        0.922
## ----------------------------------------------------------------------
##
## - thigh added
##
##
##  Step 4 : AIC = 462.4624
##  weight ~ bicep + knee + ankle + thigh
##
##                          Remove Existing Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## thigh         1    462.824    16887.878    1345.410    0.926        0.923
## ankle         1    472.432    16716.205    1517.083    0.917        0.914
## knee          1    489.497    16355.467    1877.821    0.897        0.893
## bicep         1    530.504    15098.098    3135.190    0.828        0.821
## ----------------------------------------------------------------------
##
##                             Enter New Variables
## ----------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------
## wrist         1    462.343    16961.168    1272.120    0.930        0.926
## ----------------------------------------------------------------------
##
## - wrist added
##
##
##  Step 5 : AIC = 462.3433
##  weight ~ bicep + knee + ankle + thigh + wrist
##
##                          Remove Existing Variables
```

```
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## wrist         1    462.462    16927.019    1306.269    0.928         0.925
## thigh         1    464.735    16889.376    1343.912    0.926         0.922
## ankle         1    467.722    16838.247    1395.041    0.923         0.919
## knee          1    479.222    16622.585    1610.703    0.912         0.907
## bicep         1    496.058    16245.297    1987.991    0.891         0.885
## ---------------------------------------------------------------------------
##
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.964       RMSE                 4.146
## R-Squared                0.930       Coef. Var            6.011
## Adj. R-Squared           0.926       MSE                 17.191
## Pred R-Squared           0.914       MAE                  3.164
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ---------------------------------------------------------------------
##                  Sum of
##                 Squares        DF    Mean Square       F          Sig.
## ---------------------------------------------------------------------
## Regression    16961.168         5       3392.234    197.328    0.0000
## Residual       1272.120        74         17.191
## Total         18233.288        79
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
------------------
##       model       Beta    Std. Error    Std. Beta       t        Sig
lower      upper
## ----------------------------------------------------------------------------
------------------
## (Intercept)    -90.114         7.362                -12.240    0.000    -
104.784    -75.444
##       bicep      1.491         0.231        0.421      6.453    0.000
1.030      1.951
##        knee      1.340         0.302        0.294      4.438    0.000
0.738      1.941
##       ankle      1.312         0.491        0.156      2.674    0.009
```

```
0.334      2.290
##       thigh      0.291         0.142        0.115      2.044    0.045
0.007      0.574
##       wrist      1.220         0.866        0.109      1.409    0.163
-0.505      2.945
## ----------------------------------------------------------------------
-------------------

##
##
##                                  Stepwise Summary
## ----------------------------------------------------------------------
------
## Variable     Method       AIC        RSS        Sum Sq       R-Sq
Adj. R-Sq
## ----------------------------------------------------------------------
------
## bicep       addition    542.430    3825.807    14407.481    0.79017
0.78748
## knee        addition    472.876    1564.160    16669.128    0.91421
0.91199
## ankle       addition    462.824    1345.410    16887.878    0.92621
0.92330
## thigh       addition    462.462    1306.269    16927.019    0.92836
0.92454
## wrist       addition    462.343    1272.120    16961.168    0.93023
0.92552
## ----------------------------------------------------------------------
------
```

- The two- predictors for weight are: `bicep + knee` with an AIC of 472.8764
- The four-predictors for weight are: `bicep + knee + ankle + thigh` with an AIC of 462.4624

height predictors using body girth measurements from legs and arms only

```
lr_height_new %>%
  ols_step_both_aic(
    details = T
  )

## Stepwise Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1 . thigh
## 2 . bicep
## 3 . knee
## 4 . ankle
## 5 . wrist
```

```
##
##  Step 0: AIC = 589.8319
##  height ~ 1
##
##
## Variables Entered/Removed:
##
##                            Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## wrist         1    539.991    3383.384    3710.928    0.477        0.470
## ankle         1    564.638    2044.395    5049.917    0.288        0.279
## bicep         1    569.014    1760.450    5333.862    0.248        0.239
## knee          1    575.945    1277.751    5816.561    0.180        0.170
## thigh         1    591.643      16.776    7077.536    0.002       -0.010
## ---------------------------------------------------------------------------
##
## - wrist added
##
##
##  Step 1 : AIC = 539.991
##  height ~ wrist
##
##                            Enter New Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## thigh         1    533.640    3751.209    3343.102    0.529        0.517
## bicep         1    538.284    3551.438    3542.874    0.501        0.488
## knee          1    540.738    3441.057    3653.255    0.485        0.472
## ankle         1    541.990    3383.417    3710.895    0.477        0.463
## ---------------------------------------------------------------------------
##
## - thigh added
##
##
##  Step 2 : AIC = 533.6404
##  height ~ wrist + thigh
##
##                         Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## thigh         1    539.991    3383.384    3710.928    0.477        0.470
## wrist         1    591.643      16.776    7077.536    0.002       -0.010
## ---------------------------------------------------------------------------
##
##                            Enter New Variables
## ---------------------------------------------------------------------------
```

```
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## ankle         1    531.976    3900.893    3193.419    0.550        0.532
## knee          1    533.421    3842.663    3251.649    0.542        0.524
## bicep         1    535.102    3773.647    3320.665    0.532        0.513
## ------------------------------------------------------------------------
##
## - ankle added
##
##
##  Step 3 : AIC = 531.9758
##  height ~ wrist + thigh + ankle
##
##                        Remove Existing Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## ankle         1    533.640    3751.209    3343.102    0.529        0.517
## thigh         1    541.990    3383.417    3710.895    0.477        0.463
## wrist         1    552.085    2884.334    4209.977    0.407        0.391
## ------------------------------------------------------------------------
##
##                           Enter New Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## knee          1    532.493    3959.538    3134.774    0.558        0.535
## bicep         1    533.422    3922.941    3171.371    0.553        0.529
## ------------------------------------------------------------------------
##
##
## No more variables to be added or removed.
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.742       RMSE                 6.482
## R-Squared               0.550       Coef. Var            3.814
## Adj. R-Squared          0.532       MSE                 42.019
## Pred R-Squared          0.503       MAE                  5.014
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
```

```
##                Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    3900.893         3       1300.298    30.946    0.0000
## Residual      3193.419        76         42.019
## Total         7094.312        79
## ----------------------------------------------------------------------
##
##                                    Parameter Estimates
## ------------------------------------------------------------------------------
----------------
##       model       Beta    Std. Error    Std. Beta      t        Sig
lower      upper
## ------------------------------------------------------------------------------
----------------
## (Intercept)    101.727         9.598                 10.599    0.000
82.611    120.844
##       wrist      4.242         0.862        0.609     4.919    0.000
2.524      5.959
##       thigh     -0.540         0.154       -0.344    -3.509    0.001    -
0.847     -0.234
##       ankle      1.428         0.756        0.271     1.887    0.063    -
0.079      2.934
## ------------------------------------------------------------------------------
----------------

##
##
##                                  Stepwise Summary
## ------------------------------------------------------------------------------
-----
## Variable     Method       AIC        RSS        Sum Sq      R-Sq      Adj.
R-Sq
## ------------------------------------------------------------------------------
-----
## wrist       addition    539.991    3710.928    3383.384    0.47692
0.47021
## thigh       addition    533.640    3343.102    3751.209    0.52876
0.51652
## ankle       addition    531.976    3193.419    3900.893    0.54986
0.53209
## ------------------------------------------------------------------------------
-----
```

- The two- predictors for height are: `wrist + thigh` with an AIC of 533.6404
- The four-predictors for height are: **No more variables were to be added or removed.** hence stopped at 3.

Comparing best models using only body girth measurements as predictors with the linear best models including any of the available predictors.

```
# Using all predictors
lr_weight_all <- lm(
  weight ~ .,
  data = sorted_body_sample
)
# Predict height using all predictors
lr_height_all <- lm(
  height ~ .,
  data = sorted_body_sample
)
```

Check the AIC of weight using all predictors

```
AIC(lr_weight_all)
```

```
## [1] 409.6935
```

Check the AIC of height using all predictors

```
AIC(lr_height_all)
```

```
## [1] 511.717
```

Summarising the results from best linear models in a small table.

Model Name	Model Description	AIC
`lr_weight_model`	The two- predictors for weight are:bicep + knee	472.8764
`lr_weight_model`	The four-predictors for weight are: bicep + knee + shoulder + hip	438.4734
`lr_height_model`	The two- predictors for height are: wrist + thigh	533.6404
`lr_height_model`	No more variables were to be added or removed. hence stopped at stepwise 3.	
`lr_weight_new`	using body girth measurements from the legs and arms only, i.e., not including shoulder and hip. The two-predictors for weight are: bicep + knee	472.8764
`lr_weight_new`	using body girth measurements from the legs and arms only, i.e., not including shoulder and hip. The four-predictors for weight are: bicep + knee + ankle + thigh	462.4624
`lr_height_new`	using body girth measurements from the legs and arms only, i.e., not including shoulder and hip. The two-predictors for height are: wrist + thigh	533.6404
`lr_height_new`	using body girth measurements from the legs and arms only, i.e., not including shoulder and hip.The four-predictors for height are: No more variables were to be added or removed. hence stopped at 3.	

`lr_weight_all`	Use all predictors	409.6935
`lr_height_all`	Use all predictors	511.717

From the model summaries, using all the predictors results to a model with lower Akaike Information Criterion hence better models to predict weight and height.

(4) Linear model using `shoulder` and `hip` to predict weight (model A) and the linear model using the other five body firth measurements to predict weight (model B).

```
# model A uses shoulder and hip
model.A <- lm(
  weight ~ shoulder+hip,
  data = sorted_body_sample
)

# model B users other 5 body girth measurements
model.B <- lm(
  weight ~ thigh + bicep + knee + ankle + wrist,
  data = sorted_body_sample
)
```

Comparing the residuals from these two models for each individual in the dataset

In order to compare the residuals of these two models, we will create a dataframe that will assist in comparison, then check the summary using `summary()` function.

```
# Create model residuals
models.resids <- data.frame(
  model_A_residuals=residuals(model.A),
  model_B_residuals=residuals(model.B)
)

# Check the summary of the created dataframe to compare residuals
summary(models.resids)
```

```
##  model_A_residuals  model_B_residuals
##  Min.   :-12.0025   Min.   :-11.4481
##  1st Qu.: -2.9455   1st Qu.: -2.8752
##  Median :  0.3886   Median :  0.3574
##  Mean   :  0.0000   Mean   :  0.0000
##  3rd Qu.:  2.5889   3rd Qu.:  2.8384
##  Max.   : 14.1628   Max.   : 12.0780
```

From summary statistics, we can see that there is a small negligable change between the models. We will check the summary of the models to see how each model is performing

```
summary(model.A)
```

```
##
## Call:
```

```
## lm(formula = weight ~ shoulder + hip, data = sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0025  -2.9455   0.3886   2.5889  14.1628
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -106.05369    6.13474  -17.29   <2e-16 ***
## shoulder       0.90054    0.05759   15.64   <2e-16 ***
## hip            0.80619    0.07087   11.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.445 on 77 degrees of freedom
## Multiple R-squared:  0.9166, Adjusted R-squared:  0.9144
## F-statistic: 422.9 on 2 and 77 DF,  p-value: < 2.2e-16
```

```
summary(model.B)
```

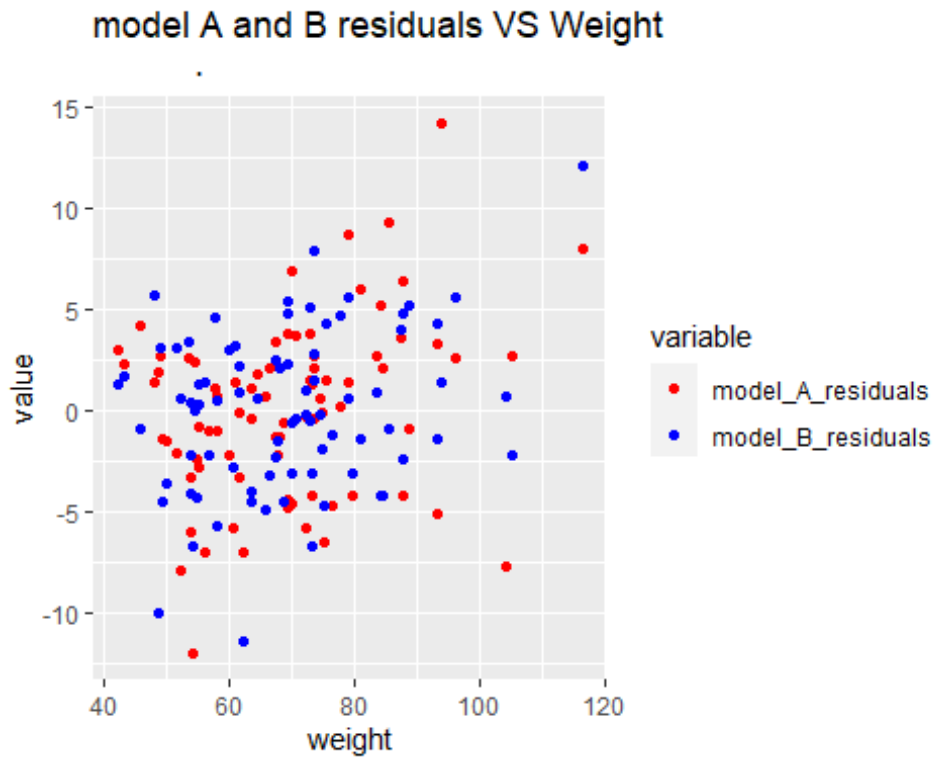
```
##
## Call:
## lm(formula = weight ~ thigh + bicep + knee + ankle + wrist, data =
sorted_body_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4481  -2.8752   0.3574   2.8384  12.0780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90.1145     7.3625 -12.240  < 2e-16 ***
## thigh         0.2907     0.1423   2.044  0.04456 *
## bicep         1.4907     0.2310   6.453 1.01e-08 ***
## knee          1.3396     0.3018   4.438 3.11e-05 ***
## ankle         1.3121     0.4907   2.674  0.00922 **
## wrist         1.2199     0.8656   1.409  0.16290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.146 on 74 degrees of freedom
## Multiple R-squared:  0.9302, Adjusted R-squared:  0.9255
## F-statistic: 197.3 on 5 and 74 DF,  p-value: < 2.2e-16
```

Model B has a higher Multiple R^2 than model A, both models have same p-values.

customised plot where the residuals from the two models is given on the vertical axis, body weight is given on the horizontal axis, and residuals from each individual are somehow linked in the plot.

Since all are numerics, a scatterplot will be the most effective plot to plot here. We will create a dataframe that contains linked points of residuals, and weight as they appear in individuals

```
# Create a dataframe contating weight and models residuals
models_weight_resids <- data.frame(
  weight=sorted_body_sample$weight,
  model_A_residuals = residuals(model.A),
  model_B_residuals = residuals(model.B)
)
# develop a customized plot
models_weight_resids %>%
  melt(id.vars="weight") %>%
  ggplot(aes(weight, value, colour=variable))+
  geom_point()+
  scale_colour_manual(values=c("red", "blue")) +
  ggtitle("model A and B residuals VS Weight
          .")
```



From this plot we can see that model B residuals are majorly higher than model A residuals in every weight except for first minimum and last maximum weights.

(5) Assessing conclusions from fitting linear models and drawing comparisons with the results from PCA and Cluster Analysis

Using all predictors has yielded better models this was seen from the AIC generated by different models tested, therefore the future medical organizers need to collect accurate measurements that can be used to predict `weight` and `height`. From this portfolio generally Supervised Machine Learning (fitting linear models) gives better results than PCA and Cluster Analysis. The Cluster Analysis was harder to determine the prevailing clusters that may be formed by seven body girth measurements. Clustering assisted in determining the hidden pattern in the dataset. With all these, it is evident that there can be a hidden pattern in the dataset that is formed with as a result of the variables. there are 4 different groups that have different characteristics formed by the seven body girth measurements

The PCA, Principal Component Analysis assisted in visualizing the variations present in body girth measurements, from PCA the 1 dimensions contributed over 74%, the variables resulting to this where `shoulder`, `hip` and `thigh` contributing the highest variations.

Clustering helped in identifying the groupings available in the dataset. This was the findings:

- 4 groups: with the seven body girth
- 2 groups: the people
- 2 groups using all the variables