PCA- Python to R code

Rafia Amin Raka

2025-08-13

```
# install.packages("corrr")
# install.packages("ggcorrplot")
# install.packages("FactoMineR")
# install.packages("factoextra")
# install.packages("broom")

library(ggplot2)
library(corrr)
library(ggcorrplot)
library(FactoMineR)
library(factoextra)
library(broom)
data <- data.frame( Height = c(170, 165, 180, 175, 160, 172, 168, 177, 162, 158),
  Weight = c(65, 59, 75, 68, 55, 70, 62, 74, 58, 54),
  Age = c(30, 25, 35, 28, 22, 32, 27, 33, 24, 21),
  Gender = c(1, 0, 1, 1, 0, 1, 0, 1, 0, 0)
)
print(data)
```

```
##    Height Weight Age Gender
## 1     170     65  30      1
## 2     165     59  25      0
## 3     180     75  35      1
## 4     175     68  28      1
## 5     160     55  22      0
## 6     172     70  32      1
## 7     168     62  27      0
## 8     177     74  33      1
## 9     162     58  24      0
## 10    158     54  21      0
```

```
numerical_data <- data[, c("Height", "Weight", "Age", "Gender")]
head(numerical_data)
```

```
##   Height Weight Age Gender
## 1    170     65  30      1
## 2    165     59  25      0
## 3    180     75  35      1
## 4    175     68  28      1
## 5    160     55  22      0
## 6    172     70  32      1
```

```
data_normalized <- scale(numerical_data)
head(data_normalized)
```

```
##         Height     Weight         Age     Gender
## [1,]  0.1747456  0.1315587  0.48297827  0.9486833
## [2,] -0.4973530 -0.6577935 -0.56697449 -0.9486833
## [3,]  1.5189428  1.4471457  1.53293102  0.9486833
## [4,]  0.8468442  0.5262348  0.06299717  0.9486833
## [5,] -1.1694516 -1.1840283 -1.19694614 -0.9486833
## [6,]  0.4435851  0.7893522  0.90295937  0.9486833
```

```
data.pca <- princomp(data_normalized)
summary(data.pca)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3      Comp.4
## Standard deviation     1.8399723 0.39702020 0.21771746 0.097345178
## Proportion of Variance 0.9404161 0.04378473 0.01316691 0.002632245
## Cumulative Proportion  0.9404161 0.98420084 0.99736775 1.000000000
```

Comment: The principal component analysis indicates that the largest portion of the variance (94.04%) in the dataset is primarily associated with Gender (loading 0.9487), followed by a moderate contribution from Age (loading 0.4830), while Height (0.1747) and Weight (0.1316) contribute much less. Including Age and Gender together captures over 98% of the total variance (cumulative proportion 0.9842), suggesting that most of the variability in the data is explained by these two variables. The remaining variables, Height and Weight, add very little additional information, indicating that dimensionality reduction could focus on Gender and Age without substantial loss of information.

```
data.pca$loadings[, 1:2]
```

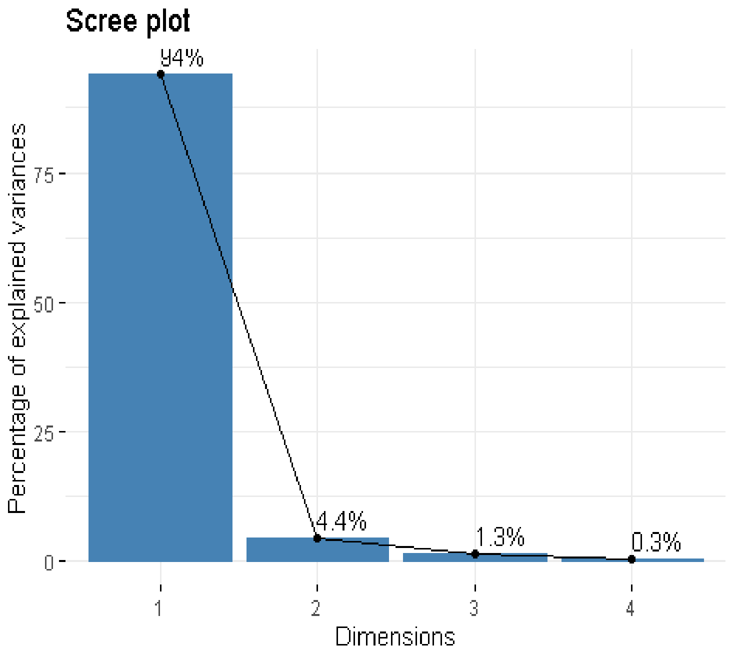
```
##           Comp.1     Comp.2
## Height 0.5045359  0.3018646
## Weight 0.5113665  0.2267456
## Age    0.5037985  0.3013030
## Gender 0.4797238 -0.8756030
```

Comment: The loadings for the first two principal components show how each variable contributes to the components. For Comp.1, Height (0.5045), Weight (0.5114), Age (0.5038), and Gender (0.4797) all have relatively high and similar positive loadings. This indicates that Comp.1 represents a combination of all variables, capturing the overall variation in the dataset.

For Comp.2, Gender (-0.8756) has a strong negative loading, while Height (0.3019), Weight (0.2267), and Age (0.3013) have smaller positive loadings. This suggests that Comp.2 primarily contrasts Gender against the other variables, highlighting variation associated with Gender that is independent of Height, Weight, and Age.

In practical terms, Comp.1 can be interpreted as a general size/age factor, while Comp.2 reflects Gender-related differences in the dataset. Together, these two components capture most of the meaningful variability, allowing for dimensionality reduction while retaining key patterns.

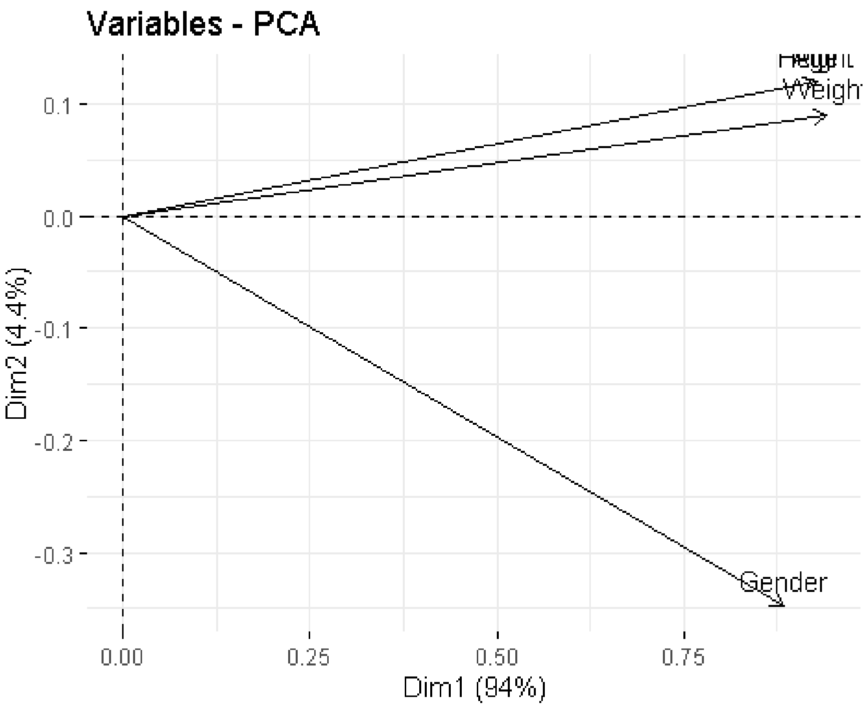
```
fviz_eig(data.pca, addlabels = TRUE)
```



Comment: The plot shows eigen values in down curve, from highest to lowest.The first two components are considered to be the most significance since they contains the most variance , 1st component 94% and 2nd component 4.4%

Biplot for the Attributes

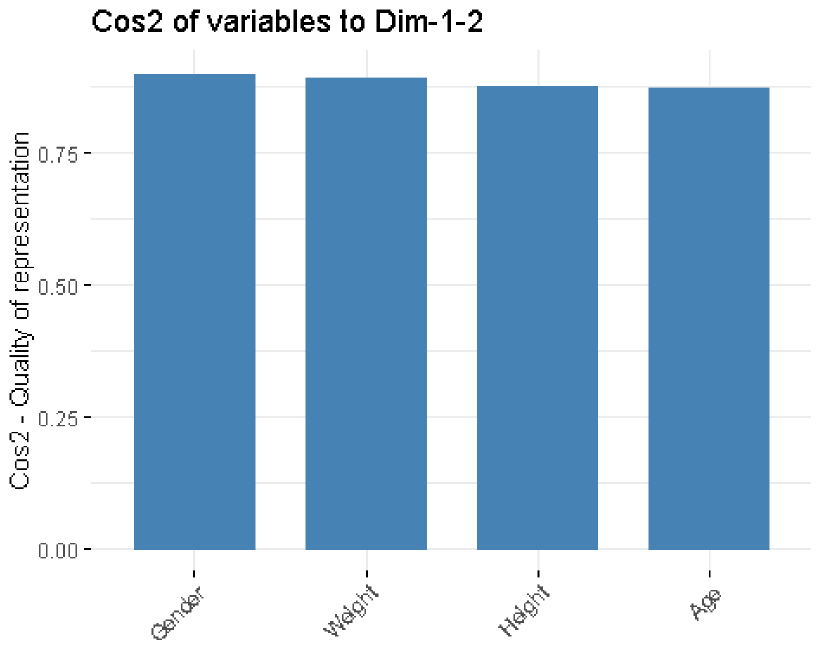
```
fviz_pca_var(data.pca, col.var = "black")
```



Comment: The PCA biplot offers a compelling visual representation of how the variables Height, Weight, and Gender relate to the principal components. The x-axis (Dim1), which explains 94% of the variance, captures the dominant structure in the data, while the y-axis (Dim2) accounts for a modest 4.4%. The arrows for Height and Weight point in nearly the same direction along Dim1, indicating a strong positive correlation between these two variables and their significant contribution to the first principal component. In contrast, the arrow for Gender points in a different direction, suggesting it is less correlated with Height and Weight and contributes more distinctly to Dim2. This separation highlights how PCA can reveal underlying patterns and relationships among variables, making it easier to interpret complex datasets and identify which features drive the most variation.

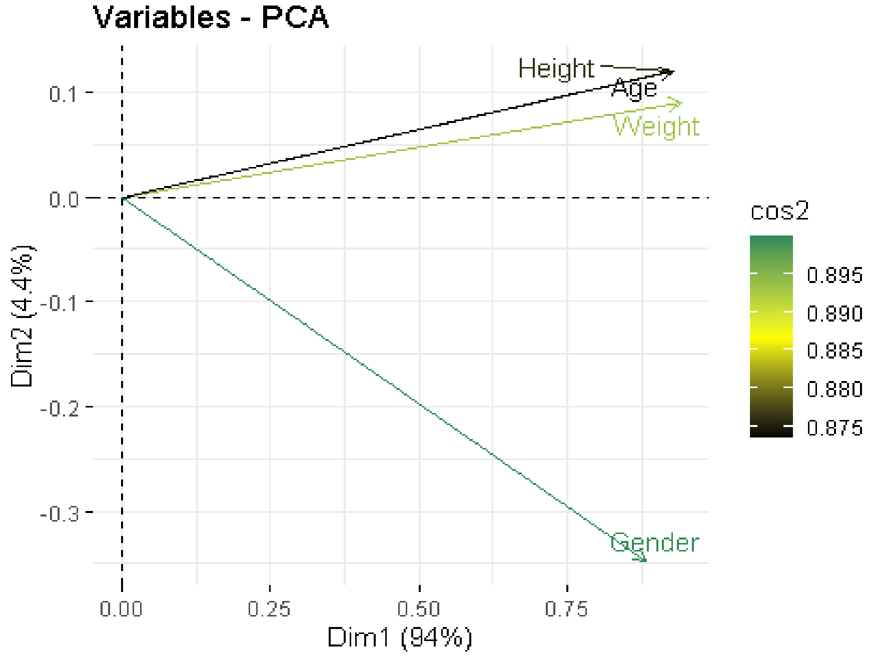
Implementation of PCA

```
fviz_cos2(data.pca, choice = "var", axes = 1:2)
```



Biplot combine with cos2

```
fviz_pca_var(data.pca, col.var = "cos2", gradient.cols = c("black", "yellow", "seagreen"), repel = TRUE)
```



Comment: The PCA biplot effectively illustrates how the variables Height, Age, Weight, and Gender relate to the first two principal components. Here, high cos2 attributes are colored in Green: Gender, Weight. And low cos2 attributes have a darker to black color: Height and Age.