null

Dimensionality reduction techniques

This chapter focuses on dimensionality reduction techniques such as principal component analysis (PCA) and Multiple correspondence analysis (MCA).

Read the “human” data

```
human <- read.csv("human_data") # Read the data from my local file
dim(human)
```

```
## [1] 155   9
```

```
str(human)
```

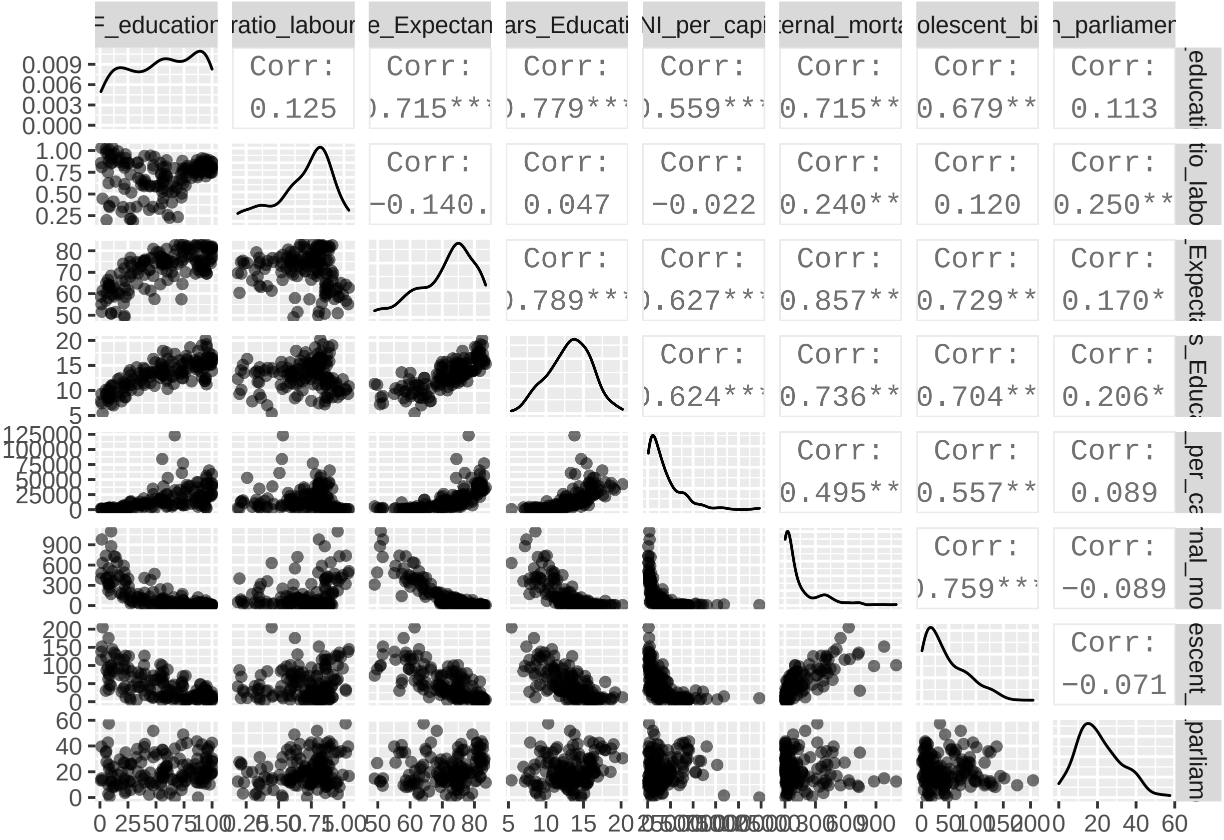
```
## 'data.frame':    155 obs. of  9 variables:
##  $ X                : chr  "Norway" "Australia" "Switzerland" "Denmark" ...
##  $ F_education      : num  97.4 94.3 95 95.5 87.7 96.3 80.5 95.1 100 95 ...
##  $ ratio_labour     : num  0.891 0.819 0.825 0.884 0.829 ...
##  $ Life_Expectancy  : num  81.6 82.4 83 80.2 81.6 80.9 80.9 79.1 82 81.8 ...
##  $ Years_Education  : num  17.5 20.2 15.8 18.7 17.9 16.5 18.6 16.5 15.9 19.2 ...
##  $ GNI_per_capita   : int  64992 42261 56431 44025 45435 43919 39568 52947 42155 32689 ...
##  $ Maternal_mortality: int  4 6 6 5 6 7 9 28 11 8 ...
##  $ Adolescent_birth  : num  7.8 12.1 1.9 5.1 6.2 3.8 8.2 31 14.5 25.3 ...
##  $ In_parliament     : num  39.6 30.5 28.5 38 36.9 36.9 19.9 19.4 28.2 31.4 ...
```

A graphical overview of the data and summary of the variables

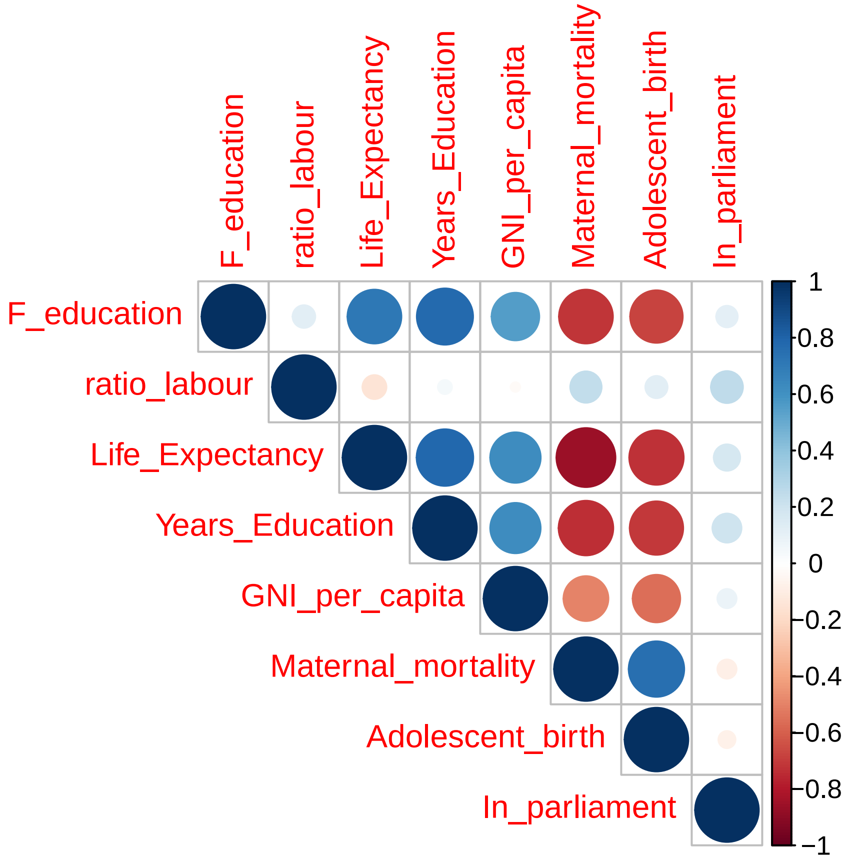
```
library(GGally) # Access the GGally library
library(dplyr) # Access the dplyr library
library(corrplot) # Access the corrplot library

# Remove the "Country" column
human_new <- dplyr::select(human, -X)

# visualize the 'human' variables
ggpairs(human_new, mapping = aes(alpha = 0.3))
```



```
# compute the correlation matrix and visualize it with corrplot
cor(human_new)%>%
corrplot(type = "upper")
```



Interpretations:

- There seems to be a positive correlation between Proportion of females with at least secondary education **(F__education)** with Life expectancy at birth **(Life__expectancy)**, and with Expected years of schooling **(Years__Education)**.
- There seems to be a slightly positive correlation between Proportion of females with at least secondary education **(F__education)** with Gross National Income (GNI) per Capita **(GNI__per__capita)**.
- There seems to be a positive correlation between Life expectancy at birth **(Life__expectancy)** and Expected years of schooling **(Years__education)**. And a slightly one between **Life__expectancy** and **GNI__per__capita**.
- There seems to be a slightly positive correlation one between **Years__education** and **GNI__per__capita**.
- A slightly positive correlation between Maternal Mortality Ratio **(Maternity__mortality)** and Adolescent Birth Rate **(Adolescent__birth)**.
- A negative correlation is observed between **Years__Education** with **Maternity__mortality** and **Adolescent__birth**. Between **Life__expectancy** with **Maternity__mortality** and **Adolescent__birth**. Between **F__education** with **Maternity__mortality** and **Adolescent__birth**.
- **Years__Education** and Percentage of female representatives in parliament **(In__parliament)** variables seem to be normally distributed.
- F__labour / M__labour **(ratio__labour)** and **Life__Expectancy** seem to have more negative values; it is reflected in the left tail.
- **GNI__per__capita**, **Maternal__mortality**, and **Adolescent__birth** seem to have more positive values; it is reflected in the right tail.

```
summary(human_new) # summary of variables
```

```
##   F_education      ratio_labour    Life_Expectancy Years_Education
##  Min.   : 0.90   Min.   :0.1857   Min.   :49.00   Min.   : 5.40
##  1st Qu.: 27.15   1st Qu.:0.5984   1st Qu.:66.30   1st Qu.:11.25  
##  Median : 56.60   Median :0.7535   Median :74.20   Median :13.50  
##  Mean   : 55.37   Mean   :0.7074   Mean   :71.65   Mean   :13.18  
##  3rd Qu.: 85.15   3rd Qu.:0.8535   3rd Qu.:77.25   3rd Qu.:15.20  
##  Max.   :100.00   Max.   :1.0380   Max.   :83.50   Max.   :20.20  
##  GNI_per_capita   Maternal_mortality Adolescent_birth In_parliament  
##  Min.   :   581   Min.   :   1.0     Min.   :  0.60   Min.   : 0.00  
##  1st Qu.:  4198   1st Qu.:  11.5     1st Qu.: 12.65   1st Qu.:12.40  
##  Median : 12040   Median :  49.0     Median : 33.60   Median :19.30  
##  Mean   : 17628   Mean   : 149.1     Mean   : 47.16   Mean   :20.91  
##  3rd Qu.: 24512   3rd Qu.: 190.0     3rd Qu.: 71.95   3rd Qu.:27.95  
##  Max.   :123124   Max.   :1100.0     Max.   :204.80   Max.   :57.50
```

Interpretations: On average,

- The proportion of females with at least secondary education is about 55.37.
- The ratio between F_education and M_education is about 0.71.
- Life expectancy at birth is approximately 72 years.
- The expected years of schooling is approximately 13 years.
- The Gross National Income (GNI) per Capita is about 17628.
- The Maternal Mortality Ratio is about 149.1.
- The Adolescent Birth Rate is about 47.2.
- The Percentage of female representatives in parliament is about 20.91.

Perform principal component analysis (PCA) – Non standarized data

```
# perform principal component analysis
pca_human_new <- prcomp(human_new)
summary(pca_human_new)
```

```
## Importance of components:
##                              PC1      PC2   PC3   PC4   PC5   PC6   PC7    PC8
## Standard deviation     1.854e+04 186.1920 25.97 19.25 11.42 3.723 1.431 0.1649
## Proportion of Variance 9.999e-01   0.0001  0.00  0.00  0.00 0.000 0.000 0.0000
## Cumulative Proportion  9.999e-01   1.0000  1.00  1.00  1.00 1.000 1.000 1.0000
```

```
# draw a biplot of the principal component
biplot(pca_human_new, choices = 1:2, col = c("blue", "red"))
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```

Interpretations: With no-standardized data, we can not really capture the variability captured by the principal components. This is because PCA is sensitive to the relative scaling of the original features and assumes that features with larger variance are more important than features with smaller variance. That

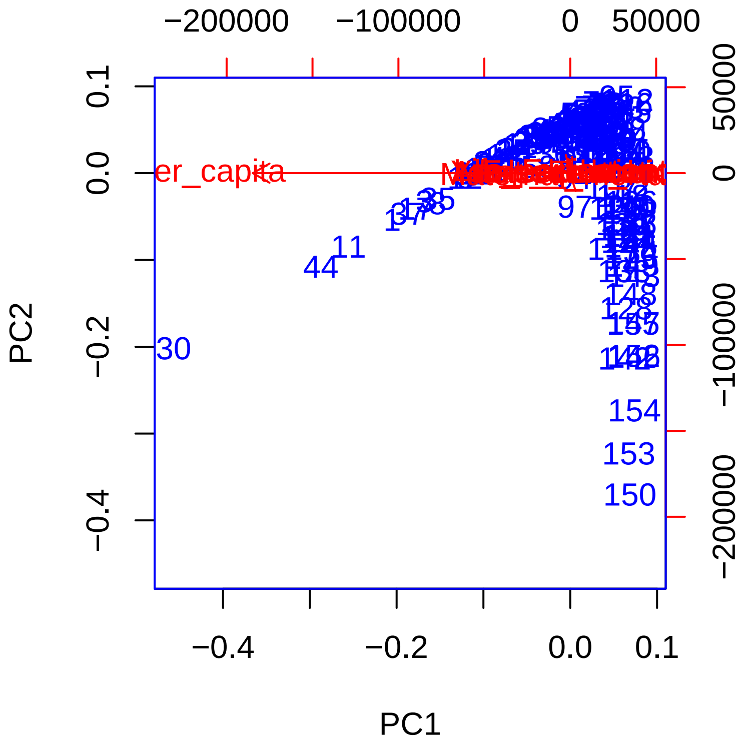


Figure 1: CPA – Non standarized data

is the reason why probably the GNI_per_capita has a larger arrow. Also, 99% of variation explained by 1 PCA component.

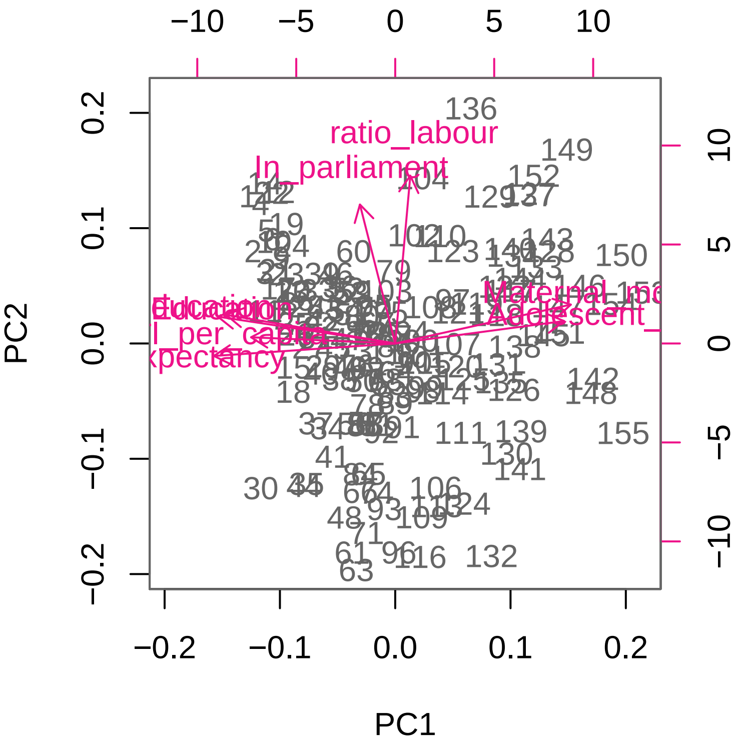
Perform principal component analysis (PCA) – Standarized data

```r
# standardize the variables
human_std <- scale(human_new)

# perform principal component analysis
pca_human_std <- prcomp(human_std)
summary(pca_human_std)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1194 1.1478 0.89070 0.73763 0.55201 0.48552 0.44894
## Proportion of Variance 0.5615 0.1647 0.09917 0.06801 0.03809 0.02947 0.02519
## Cumulative Proportion  0.5615 0.7261 0.82532 0.89333 0.93142 0.96089 0.98608
##                            PC8
## Standard deviation     0.33372
## Proportion of Variance 0.01392
## Cumulative Proportion  1.00000
```

```r
# draw a biplot of the principal component
biplot(pca_human_std, choices = 1:2, col = c("grey40", "deeppink2"))
```



Interpretations: