Dimensionality Reduction: PCA

# Dimensionality Reduction: PCA & MCA

Created human file/data from the given source and wrangled it enough for further usage. Reading the wrangled data and confirming the dimensionalities match with the given one.

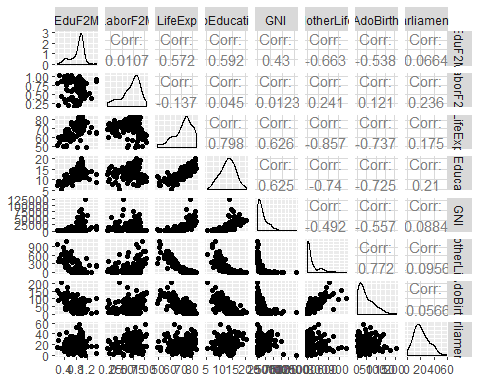
dim(human\_data)

## [1] 155 8

head(human\_data)

## EduF2M LaborF2M LifeExp ExpEducation GNI MotherLife  
## Afghanistan 0.1979866 0.1987421 60.4 9.3 1885 400  
## Albania 0.9306030 0.6854962 77.8 11.8 9943 21  
## Algeria 0.8612903 0.2105263 74.8 14.0 13054 89  
## Arab States 0.7289916 0.3081009 70.6 12.0 15722 155  
## Argentina 0.9774306 0.6333333 76.3 17.9 22050 69  
## Armenia 0.9894737 0.7465565 74.7 12.3 8124 29  
## AdoBirth parliamentF  
## Afghanistan 86.8 27.6  
## Albania 15.3 20.7  
## Algeria 10.0 25.7  
## Arab States 45.4 14.0  
## Argentina 54.4 36.8  
## Armenia 27.1 10.7

ggpairs(human\_data)



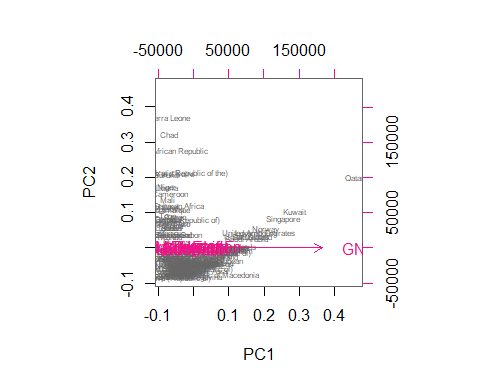
summary(human\_data)

## EduF2M LaborF2M LifeExp ExpEducation   
## Min. :0.1717 Min. :0.1857 Min. :49.00 Min. : 5.40   
## 1st Qu.:0.7264 1st Qu.:0.5965 1st Qu.:67.25 1st Qu.:11.25   
## Median :0.9359 Median :0.7523 Median :74.00 Median :13.50   
## Mean :0.8524 Mean :0.7026 Mean :71.72 Mean :13.16   
## 3rd Qu.:0.9957 3rd Qu.:0.8454 3rd Qu.:77.15 3rd Qu.:15.20   
## Max. :1.4967 Max. :1.0380 Max. :83.50 Max. :20.20   
## GNI MotherLife AdoBirth parliamentF   
## Min. : 581 Min. : 1.0 Min. : 0.60 Min. : 0.00   
## 1st Qu.: 4737 1st Qu.: 11.5 1st Qu.: 12.65 1st Qu.:12.50   
## Median : 12122 Median : 50.0 Median : 32.80 Median :19.30   
## Mean : 17717 Mean : 148.0 Mean : 46.58 Mean :21.02   
## 3rd Qu.: 24512 3rd Qu.: 186.5 3rd Qu.: 71.65 3rd Qu.:27.65   
## Max. :123124 Max. :1100.0 Max. :204.80 Max. :57.50

The results show that we have 155 rows and 8 columns now. We have mapped the data with countries as rows for better understanding.

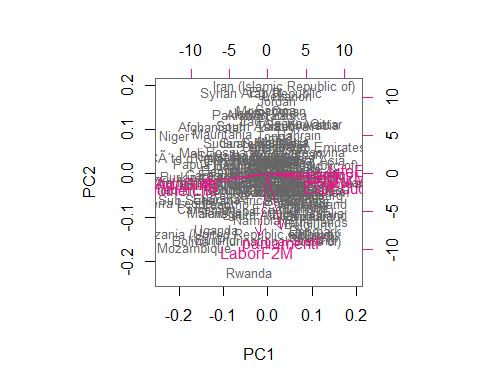
The data consists of 8 features which are representing Women empowerment in different countries regarding education, work, Expected life and average life with representation in the parliament.

Then performed Prinicple Component Analysis (PCA) without normalizing the data. It is difficult to understand the resutls in graphical format as the features are making no sense.



For a better understanding, we have to standardize the data.

human\_std <- scale(human\_data)  
  
human\_pca <- prcomp(human\_std)  
  
biplot(human\_pca, choices = 1:2, cex=c(0.8,1), col=c("grey40", "deeppink2"))

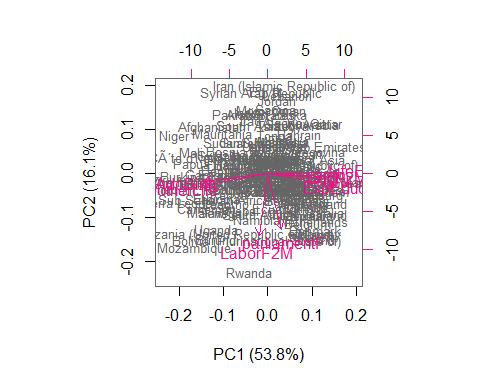


s <- summary(human\_pca)  
s

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.0755 1.1342 0.88683 0.77944 0.66236 0.51776  
## Proportion of Variance 0.5385 0.1608 0.09831 0.07594 0.05484 0.03351  
## Cumulative Proportion 0.5385 0.6993 0.79758 0.87352 0.92836 0.96187  
## PC7 PC8  
## Standard deviation 0.45219 0.31709  
## Proportion of Variance 0.02556 0.01257  
## Cumulative Proportion 0.98743 1.00000

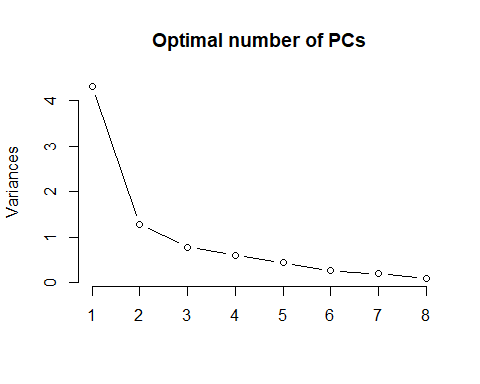
Now the results are creating some understanding as we cans ee that Female to male education ratio, Average life expectancy, Gross National Income and Expected Education is strongly correlated towards to Pricniple Component 1. The results depict that Adolesent birth and Mother’s life are correlated to Principle Component 2. two interesting features Female representation in Parliament and Labor ration with male is uncorrelated to Principle Component 2 and note correlated with Principle Component 1 either.

We can see from the plot given below the percentage of first 2 principle components and the correlation of different features with them.



If we look at the sumamries given above and the graph showing below, we can see that the optimal number of Principle Components which are more significant to data are two. Screeplot is showing the exponential change in the variance when the Principle Components went from 1 to 2.

screeplot(human\_pca, npcs = 8, type = "lines", main = "Optimal number of PCs")



Since we are studying component analysis in this week and PCA works for quantitative date or numerical values but if our data is qualitative or categorical, we will use Multiple Correspondence Analysis (MCA). MCA takes multiple categorical variables and seeks to identify associations between levels of those variables. We are using “tea” dataset here which has 300 X 36 dimensions. Just for the understanding of the concepts, we are using only 6 features here.

# look at the summaries and structure of the data  
dim(tea\_time)

## [1] 300 6

names(tea\_time)

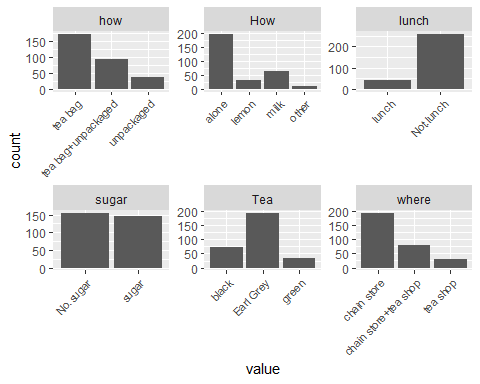
## [1] "Tea" "How" "how" "sugar" "where" "lunch"

summary(tea\_time)

## Tea How how sugar   
## black : 74 alone:195 tea bag :170 No.sugar:155   
## Earl Grey:193 lemon: 33 tea bag+unpackaged: 94 sugar :145   
## green : 33 milk : 63 unpackaged : 36   
## other: 9   
## where lunch   
## chain store :192 lunch : 44   
## chain store+tea shop: 78 Not.lunch:256   
## tea shop : 30   
##

str(tea\_time)

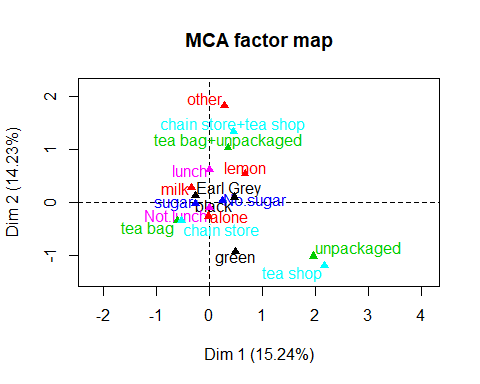
## 'data.frame': 300 obs. of 6 variables:  
## $ Tea : Factor w/ 3 levels "black","Earl Grey",..: 1 1 2 2 2 2 2 1 2 1 ...  
## $ How : Factor w/ 4 levels "alone","lemon",..: 1 3 1 1 1 1 1 3 3 1 ...  
## $ how : Factor w/ 3 levels "tea bag","tea bag+unpackaged",..: 1 1 1 1 1 1 1 1 2 2 ...  
## $ sugar: Factor w/ 2 levels "No.sugar","sugar": 2 1 1 2 1 1 1 1 1 1 ...  
## $ where: Factor w/ 3 levels "chain store",..: 1 1 1 1 1 1 1 1 2 2 ...  
## $ lunch: Factor w/ 2 levels "lunch","Not.lunch": 2 2 2 2 2 2 2 2 2 2 ...



# summary of the model  
summary(mca)

##   
## Call:  
## MCA(X = tea\_time, graph = FALSE)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6  
## Variance 0.279 0.261 0.219 0.189 0.177 0.156  
## % of var. 15.238 14.232 11.964 10.333 9.667 8.519  
## Cumulative % of var. 15.238 29.471 41.435 51.768 61.434 69.953  
## Dim.7 Dim.8 Dim.9 Dim.10 Dim.11  
## Variance 0.144 0.141 0.117 0.087 0.062  
## % of var. 7.841 7.705 6.392 4.724 3.385  
## Cumulative % of var. 77.794 85.500 91.891 96.615 100.000  
##   
## Individuals (the 10 first)  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3  
## 1 | -0.298 0.106 0.086 | -0.328 0.137 0.105 | -0.327  
## 2 | -0.237 0.067 0.036 | -0.136 0.024 0.012 | -0.695  
## 3 | -0.369 0.162 0.231 | -0.300 0.115 0.153 | -0.202  
## 4 | -0.530 0.335 0.460 | -0.318 0.129 0.166 | 0.211  
## 5 | -0.369 0.162 0.231 | -0.300 0.115 0.153 | -0.202  
## 6 | -0.369 0.162 0.231 | -0.300 0.115 0.153 | -0.202  
## 7 | -0.369 0.162 0.231 | -0.300 0.115 0.153 | -0.202  
## 8 | -0.237 0.067 0.036 | -0.136 0.024 0.012 | -0.695  
## 9 | 0.143 0.024 0.012 | 0.871 0.969 0.435 | -0.067  
## 10 | 0.476 0.271 0.140 | 0.687 0.604 0.291 | -0.650  
## ctr cos2   
## 1 0.163 0.104 |  
## 2 0.735 0.314 |  
## 3 0.062 0.069 |  
## 4 0.068 0.073 |  
## 5 0.062 0.069 |  
## 6 0.062 0.069 |  
## 7 0.062 0.069 |  
## 8 0.735 0.314 |  
## 9 0.007 0.003 |  
## 10 0.643 0.261 |  
##   
## Categories (the 10 first)  
## Dim.1 ctr cos2 v.test Dim.2 ctr  
## black | 0.473 3.288 0.073 4.677 | 0.094 0.139  
## Earl Grey | -0.264 2.680 0.126 -6.137 | 0.123 0.626  
## green | 0.486 1.547 0.029 2.952 | -0.933 6.111  
## alone | -0.018 0.012 0.001 -0.418 | -0.262 2.841  
## lemon | 0.669 2.938 0.055 4.068 | 0.531 1.979  
## milk | -0.337 1.420 0.030 -3.002 | 0.272 0.990  
## other | 0.288 0.148 0.003 0.876 | 1.820 6.347  
## tea bag | -0.608 12.499 0.483 -12.023 | -0.351 4.459  
## tea bag+unpackaged | 0.350 2.289 0.056 4.088 | 1.024 20.968  
## unpackaged | 1.958 27.432 0.523 12.499 | -1.015 7.898  
## cos2 v.test Dim.3 ctr cos2 v.test   
## black 0.003 0.929 | -1.081 21.888 0.382 -10.692 |  
## Earl Grey 0.027 2.867 | 0.433 9.160 0.338 10.053 |  
## green 0.107 -5.669 | -0.108 0.098 0.001 -0.659 |  
## alone 0.127 -6.164 | -0.113 0.627 0.024 -2.655 |  
## lemon 0.035 3.226 | 1.329 14.771 0.218 8.081 |  
## milk 0.020 2.422 | 0.013 0.003 0.000 0.116 |  
## other 0.102 5.534 | -2.524 14.526 0.197 -7.676 |  
## tea bag 0.161 -6.941 | -0.065 0.183 0.006 -1.287 |  
## tea bag+unpackaged 0.478 11.956 | 0.019 0.009 0.000 0.226 |  
## unpackaged 0.141 -6.482 | 0.257 0.602 0.009 1.640 |  
##   
## Categorical variables (eta2)  
## Dim.1 Dim.2 Dim.3   
## Tea | 0.126 0.108 0.410 |  
## How | 0.076 0.190 0.394 |  
## how | 0.708 0.522 0.010 |  
## sugar | 0.065 0.001 0.336 |  
## where | 0.702 0.681 0.055 |  
## lunch | 0.000 0.064 0.111 |

# visualize MCA  
plot(mca, invisible=c("ind"), habillage = "quali")



The Correspondence Analysis plot shows the cloud of categories of the six variables as projected onto the two principal axes. The results show that tea bag are related to chain store while unpackages have a correlation with tea shop. Green tea has less correlation for the 2nd component which is on vertical axes similarly Earl Grey has less influence against first component which is on horizontal axes.