Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Happiness Report-Principal Component Analyses

Alecsandru Anca-Giorgiana

Group:1077

Discipline: Data Analyses

TABLE OF CONTENTS

1. Theoretial framework…………………………………………………...3
2. Objective…………………………………………………………………6
3. Implementing PCA and interpreting results……………………………..8
4. Conclusions…………………………………………………………….15
5. References……………………………………………………………...16

.1 Theoretical framework for World Happiness Report

Detailed information about each of the predictors that will be used in this paper:

"**1. GDP per capita** is in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank in December 2015. See the appendix for more details. GDP data for 2015 are not yet available, so we extend the GDP time series from 2014 to 2015 using country-specific forecasts of real GDP growth from the OECD Economic Outlook No. 98 (Edition 2015/2) and World Bank's Global Economic Prospects (December 2014 release), after adjustment for population growth. The equation uses the natural log of GDP per capita, since that form fits the data significantly better than does GDP per capita. The statistics of GDP per capita (variable name gdp) in purchasing power parity (PPP) at constant 2011 international dollar prices are from the August 10, 2016 release of the World Development Indicators (WDI). The GDP figures for Taiwan are from the Penn World Table 7.1. Syria and Argentina are missing the GDP numbers in the WDI release but were present in earlier releases. We use the numbers from the earlier release, after adjusting their levels by a factor of 1.17 to take into account changes in the implied prices when switching from the PPP 2005 prices used in the earlier release to the PPP 2011 prices used in the latest release. The factor of 1.17 is the average ratio derived by dividing the US GDP per capita under the 2011 prices with their counterparts under the 2005 prices. The same 1.17 is used to adjust the Taiwanese numbers, which are originally PPP dollars at 2005 constant prices. GDP per capita in 2016 are not yet available as of September 2016. We extend the GDP-per-capita time series from 2015 to 2016 using country specific forecasts of real GDP growth in 2016 first from the OECD Economic Outlook No. 99 (Edition 2016/1) and then, if missing, forecasts from World Bank's Global Economic Prospects (Last Updated: 01/06/2016). The GDP growth forecasts are adjusted for population growth with the subtraction of 2014-15 population growth as the projected 2015-16 growth.

2. The time series of **healthy life expectancy** at birth are constructed based on data from the World Health Organization (WHO) and the World Development Indicators (WDI). WHO publishes the data on healthy life expectancy for the year 2012. The time series of life expectancies, with no adjustment for health, are available in WDI. We adopt the following strategy to construct the time series of healthy life expectancy at birth: first we generate the ratios of healthy life expectancy to life expectancy in 2012 for countries with both data. We then apply the country-specific ratios to other years to generate the healthy life expectancy data. See the appendix for more details.

3. **Social support** (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the Gallup World Poll (GWP) question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

4. **Freedom** to make life choices is the national average of binary responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

5. **Generosity** is the residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.

6. **Perceptions of corruption** are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

7. **Positive affec**t is defined as the average of previous-day affect measures for happiness, laughter and enjoyment for GWP waves 3-7 (years 2008 to 2012, and some in 2013). It is

defined as the average of laughter and enjoyment for other waves where the happiness question was not asked.

8. **Negative affect** is defined as the average of previous-day affect measures for worry, sadness and anger for all waves. See the appendix for more details. " (“THE DISTRIBUTION OF WORLD HAPPINESS” by JOHN F. HELLIWELL, HAIFANG HUANG AND SHUN WANG at (http://worldhappiness.report/wp-content/uploads/sites/2/2016/03/HR-V1Ch2_web.pdf))

Objective

With a large number of variables, the dispersion matrix may be too large to study and interpret properly. There would be too many pairwise correlations between the variables to consider. Graphical display of data may also not be of particular help incase the data set is very large. With 12 variables, for example, there will be more than 200 three-dimensional scatterplots to be studied!

To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component.

What are principal components ?

A principal component is a normalized linear combination of the original predictors in a data set. In image above, *PC1* and *PC2* are the principal components. Let's say we have a set of predictors as $X^1, X^2...,Xp$

The principal component can be written as:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... +\Phi p^1 Xp$$

where,

- Z^1 is first principal component
- Φp^1 is the loading vector comprising of loadings ($\Phi^1, \Phi^2..$) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z^1) along which data varies the most. It results in a line in *p*

dimensional space which is closest to the *n* observations. Closeness is measured using average squared euclidean distance.

- $X^1..Xp$ are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equals to one.

Therefore,

First principal component is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component. No other component can have variability higher than first principal component.

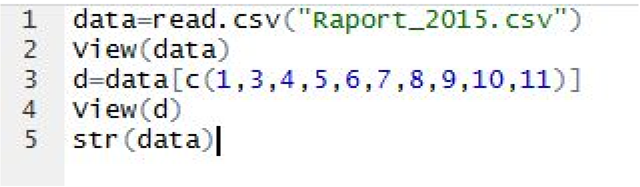
The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.

Similarly, we can compute the second principal component also.

Second principal component (Z^2) is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with Z^1. In other words, the correlation between first and second component should is zero. It can be represented as:

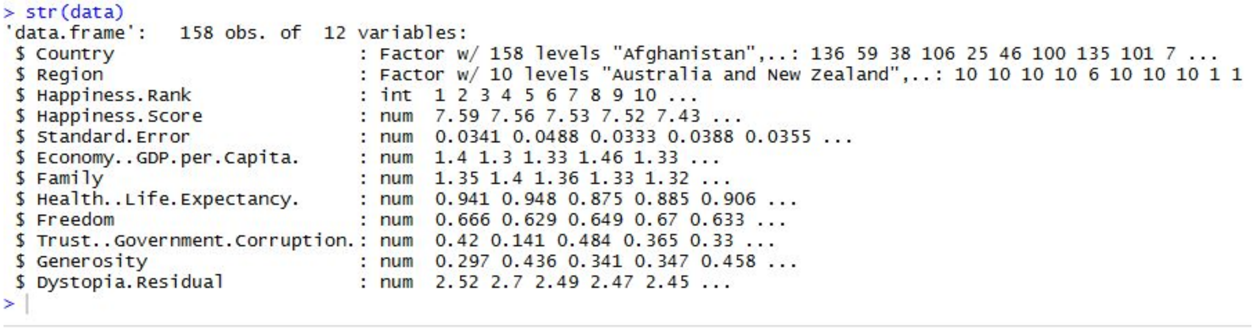
$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi p2Xp$$

IMPLEMENTING PCA AND INTERPRETING



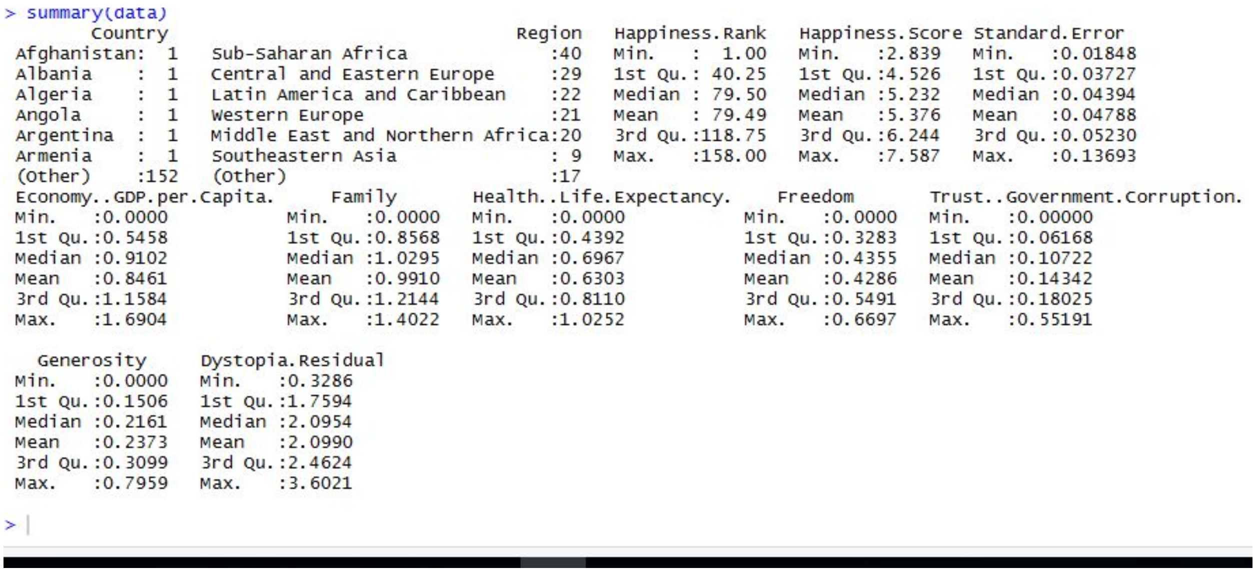
```
data=read.csv("Raport_2015.csv")
View(data)
d=data[c(1,3,4,5,6,7,8,9,10,11)]
View(d)
str(data)
```

Output:

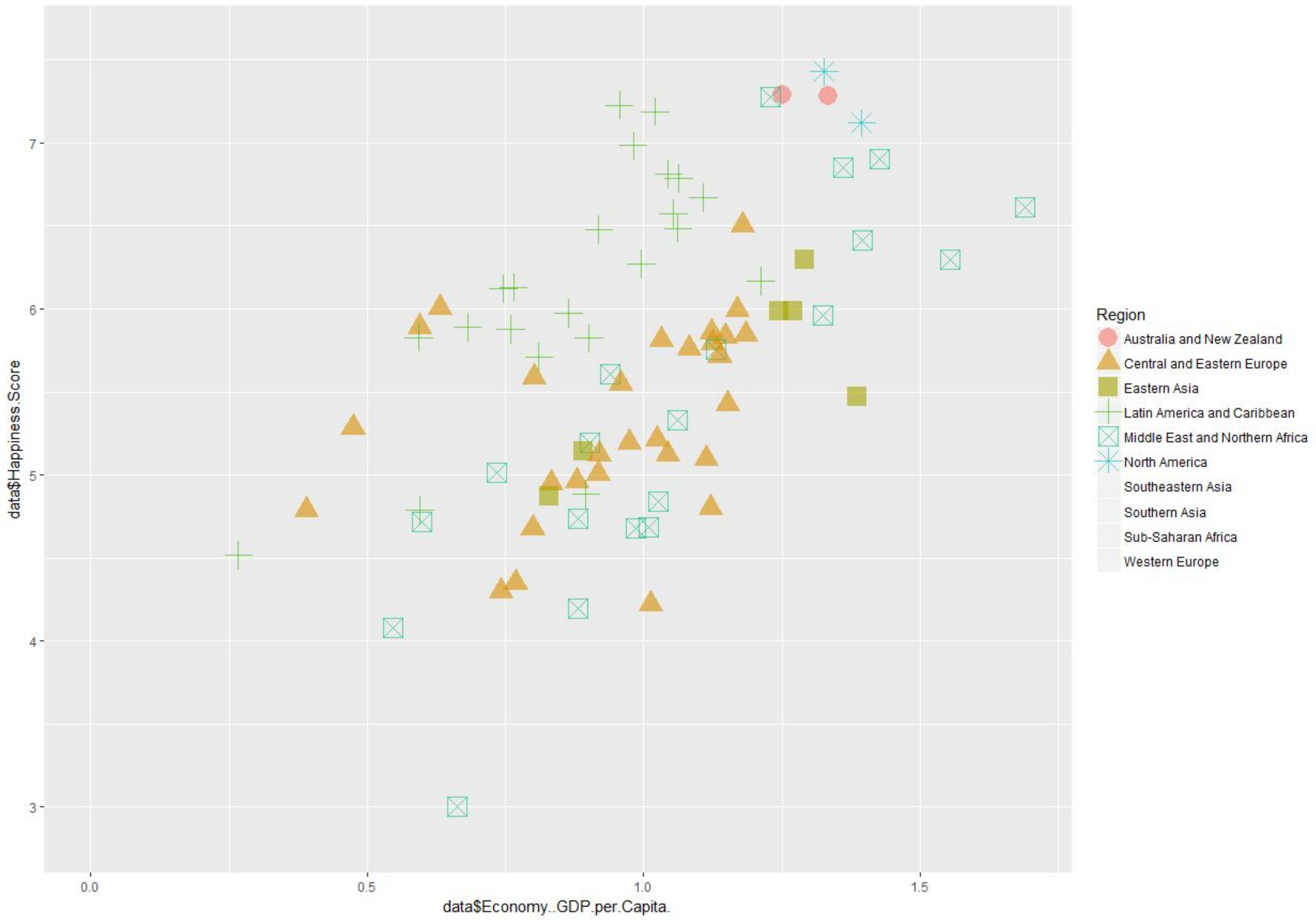


```
> str(data)
'data.frame':   158 obs. of  12 variables:
 $ Country                      : Factor w/ 158 levels "Afghanistan",..: 136 59 38 106 25 46 100 135 101 7 ...
 $ Region                       : Factor w/ 10 levels "Australia and New Zealand",..: 10 10 10 10 6 10 10 10 1 1
 $ Happiness.Rank               : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Happiness.Score              : num  7.59 7.56 7.53 7.52 7.43 ...
 $ Standard.Error               : num  0.0341 0.0488 0.0333 0.0388 0.0355 ...
 $ Economy..GDP.per.Capita.     : num  1.4 1.3 1.33 1.46 1.33 ...
 $ Family                       : num  1.35 1.4 1.36 1.33 1.32 ...
 $ Health..Life.Expectancy.     : num  0.941 0.948 0.875 0.885 0.906 ...
 $ Freedom                      : num  0.666 0.629 0.649 0.67 0.633 ...
 $ Trust..Government.Corruption.: num  0.42 0.141 0.484 0.365 0.33 ...
 $ Generosity                   : num  0.297 0.436 0.341 0.347 0.458 ...
 $ Dystopia.Residual            : num  2.52 2.7 2.49 2.47 2.45 ...
> 
```

As we can observe we have 158 observation of 12 variable from which 10 are numeric and 2 are factor variables.



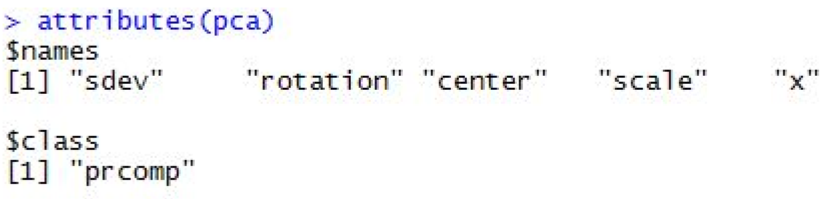
```
> summary(data)
        Country                              Region    Happiness.Rank   Happiness.Score Standard.Error
 Afghanistan:  1   Sub-Saharan Africa             :40   Min.   :  1.00   Min.   :2.839   Min.   :0.01848
 Albania    :  1   Central and Eastern Europe     :29   1st Qu.: 40.25   1st Qu.:4.526   1st Qu.:0.03727
 Algeria    :  1   Latin America and Caribbean    :22   Median : 79.50   Median :5.232   Median :0.04394
 Angola     :  1   Western Europe                 :21   Mean   : 79.49   Mean   :5.376   Mean   :0.04788
 Argentina  :  1   Middle East and Northern Africa:20   3rd Qu.:118.75   3rd Qu.:6.244   3rd Qu.:0.05230
 Armenia    :  1   Southeastern Asia              : 9   Max.   :158.00   Max.   :7.587   Max.   :0.13693
 (Other)    :152   (Other)                        :17
 Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.    Freedom       Trust..Government.Corruption.
 Min.   :0.0000           Min.   :0.0000   Min.   :0.0000           Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.5458           1st Qu.:0.8568   1st Qu.:0.4392           1st Qu.:0.3283   1st Qu.:0.06168
 Median :0.9102           Median :1.0295   Median :0.6967           Median :0.4355   Median :0.10722
 Mean   :0.8461           Mean   :0.9910   Mean   :0.6303           Mean   :0.4286   Mean   :0.14342
 3rd Qu.:1.1584           3rd Qu.:1.2144   3rd Qu.:0.8110           3rd Qu.:0.5491   3rd Qu.:0.18025
 Max.   :1.6904           Max.   :1.4022   Max.   :1.0252           Max.   :0.6697   Max.   :0.55191

   Generosity     Dystopia.Residual
 Min.   :0.0000   Min.   :0.3286
 1st Qu.:0.1506   1st Qu.:1.7594
 Median :0.2161   Median :2.0954
 Mean   :0.2373   Mean   :2.0990
 3rd Qu.:0.3099   3rd Qu.:2.4624
 Max.   :0.7959   Max.   :3.6021

> |
```



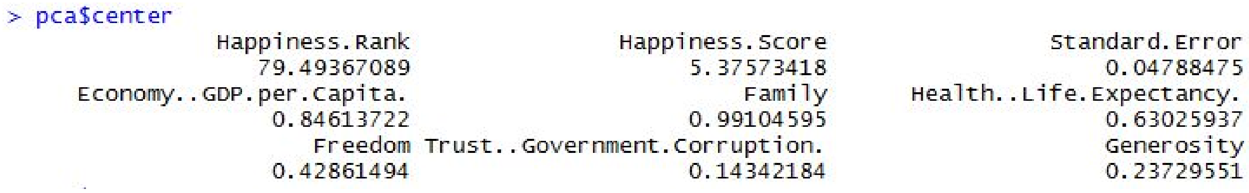
After plotting the data based on the regions we can observe the clusters formed.

Running the Principle Component Analyses using the prcomp() function in R Studio we can observe the attributes returned by it.

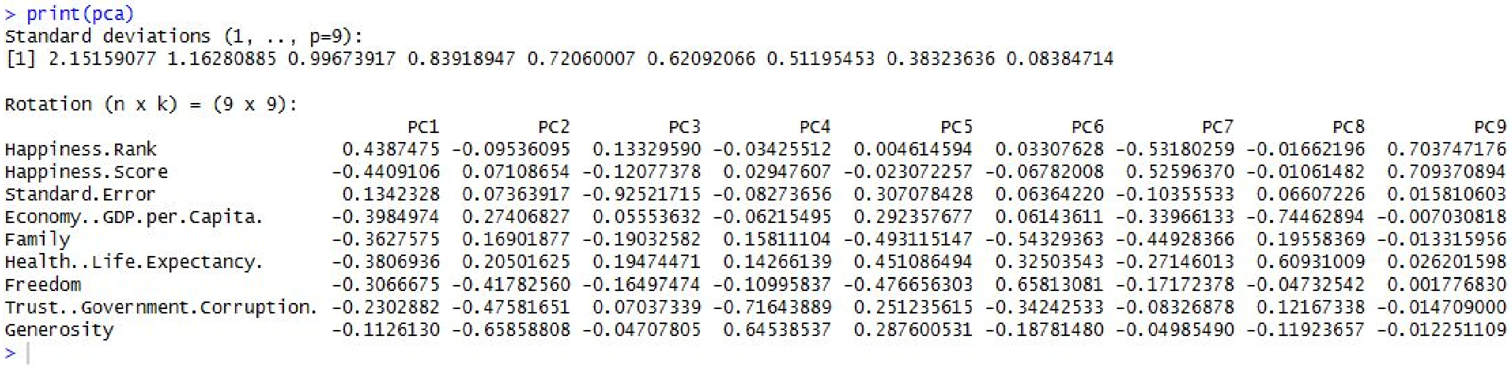
```
pca=prcomp(d[,-1],center=TRUE,scale.=TRUE)
attributes(pca)
```



```
> attributes(pca)
$names
[1] "sdev"     "rotation" "center"   "scale"    "x"

$class
[1] "prcomp"
```



```
> pca$center
                Happiness.Rank              Happiness.Score               Standard.Error
                   79.49367089                   5.37573418                   0.04788475
      Economy..GDP.per.Capita.                       Family     Health..Life.Expectancy.
                    0.84613722                   0.99104595                   0.63025937
                       Freedom Trust..Government.Corruption.                  Generosity
                    0.42861494                   0.14342184                   0.23729551
```



```
> print(pca)
Standard deviations (1, .., p=9):
[1] 2.15159077 1.16280885 0.99673917 0.83918947 0.72060007 0.62092066 0.51195453 0.38323636 0.08384714

Rotation (n x k) = (9 x 9):
                                     PC1         PC2         PC3         PC4          PC5         PC6         PC7         PC8          PC9
Happiness.Rank                 0.4387475 -0.09536095  0.13329590 -0.03425512  0.004614594  0.03307628 -0.53180259 -0.01662196  0.703747176
Happiness.Score               -0.4409106  0.07108654 -0.12077378  0.02947607 -0.023072257 -0.06782008  0.52596370 -0.01061482  0.709370894
Standard.Error                 0.1342328  0.07363917 -0.92521715 -0.08273656  0.307078428  0.06364220 -0.10355533  0.06607226  0.015810603
Economy..GDP.per.Capita.      -0.3984974  0.27406827  0.05553632 -0.06215495  0.292357677  0.06143611 -0.33966133 -0.74462894 -0.007030818
Family                        -0.3627575  0.16901877 -0.19032582  0.15811104 -0.493115147 -0.54329363 -0.44928366  0.19558369 -0.013315956
Health..Life.Expectancy.      -0.3806936  0.20501625  0.19474471  0.14266139  0.451086494  0.32503543 -0.27146013  0.60931009  0.026201598
Freedom                       -0.3066675 -0.41782560 -0.16497474 -0.10995837 -0.476656303  0.65813081 -0.17172378 -0.04732542  0.001776830
Trust..Government.Corruption. -0.2302882 -0.47581651  0.07037339 -0.71643889  0.251235615 -0.34242533 -0.08326878  0.12167338 -0.014709000
Generosity                    -0.1126130 -0.65858808 -0.04707805  0.64538537  0.287600531 -0.18781480 -0.04985490 -0.11923657 -0.012251109
> |
```

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, which of these numbers are large in magnitude, the farthest from zero in either positive or negative direction. Which numbers we consider to be large or small is of course is a subjective decision. You need to determine at what level the correlation value will be of importance. **Here a correlation value above 0.5 is deemed important.** We will now interpret the principal component results with respect to the value that we have deemed significant.

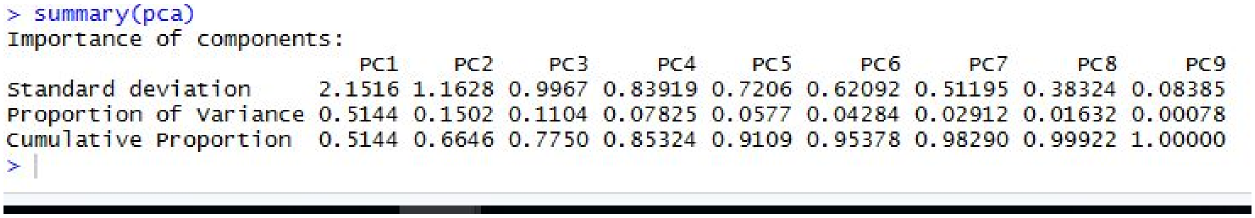
First Principal Component Analysis - PCA1

The first principal component is strongly negatively correlated with seven of the original variables. The first principal component decreases with decreasing GDP,, Health, Freedom,, Generosity and Trust in the Government. This suggests that these seven criteria vary together.

Second Principal Component Analysis - PCA2

The second principal component decreases with only three of the values, decreasing Freedom, Trust in the government and Generosity.

Each Principle Component is a normalised linear combination of all the original variables listed(Happiness Score, GDP/Capita, Family, Health,Freedom, Generosity, Government Corruption).

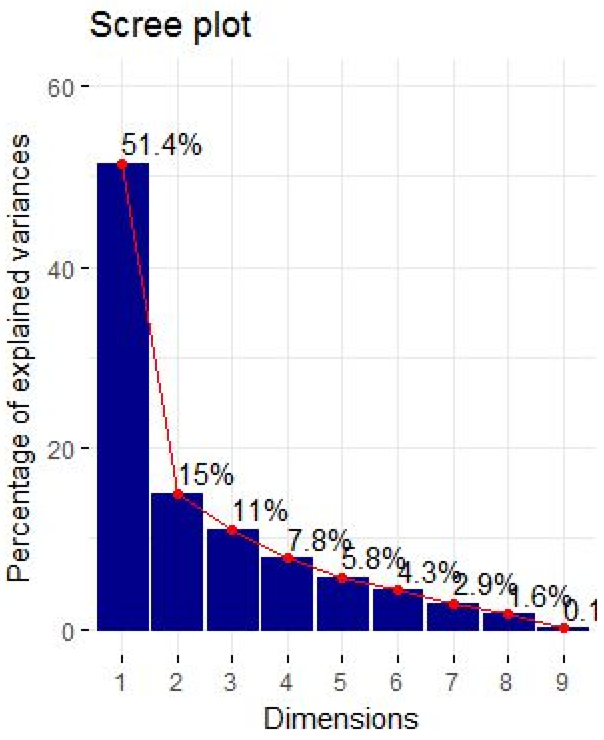


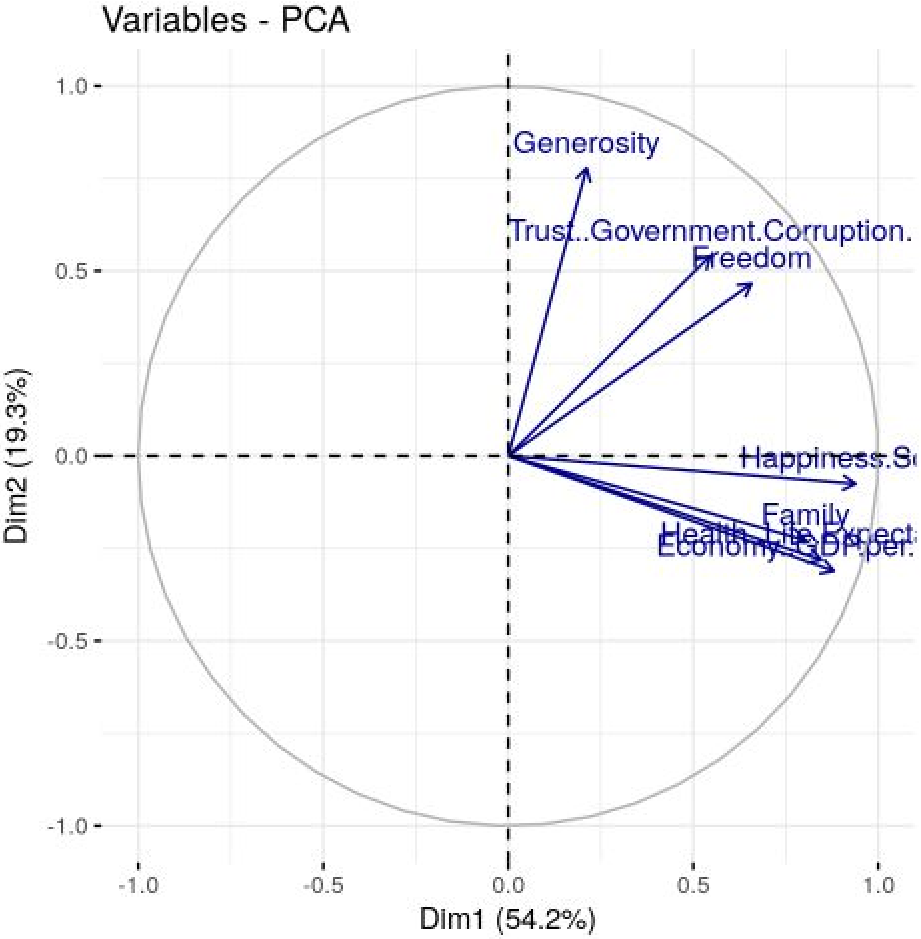
```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7     PC8     PC9
Standard deviation     2.1516 1.1628 0.9967 0.83919 0.7206 0.62092 0.51195 0.38324 0.08385
Proportion of Variance 0.5144 0.1502 0.1104 0.07825 0.0577 0.04284 0.02912 0.01632 0.00078
Cumulative Proportion  0.5144 0.6646 0.7750 0.85324 0.9109 0.95378 0.98290 0.99922 1.00000
> |
```

We can observe that only the first Principle Component explains about 51% of the variation.

Looking at the cumulative proportion we can observe that the first 4 principle components explain almost 85% of the variance in tha dataset.

The same thing can be observed from the Scree Plot below.


Scree plot
Percentage of explained variances
60
40
20
0
51.4%
15%
11%
7.8%
5.8%
4.3%
2.9%
1.6%
0.1
1
2
3
4
5
6
7
8
9
Dimensions



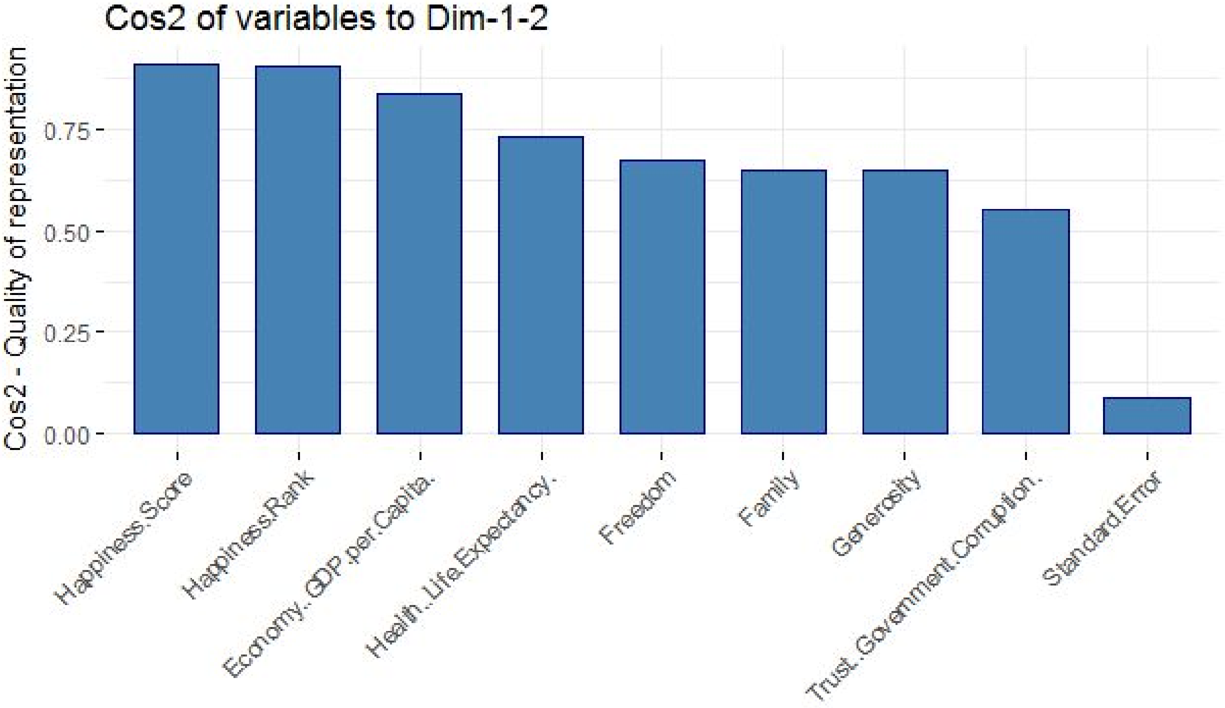
We see that for instance happiness score and family, health and economy are highly correlated. Trust in the government and freedom are also correlated.

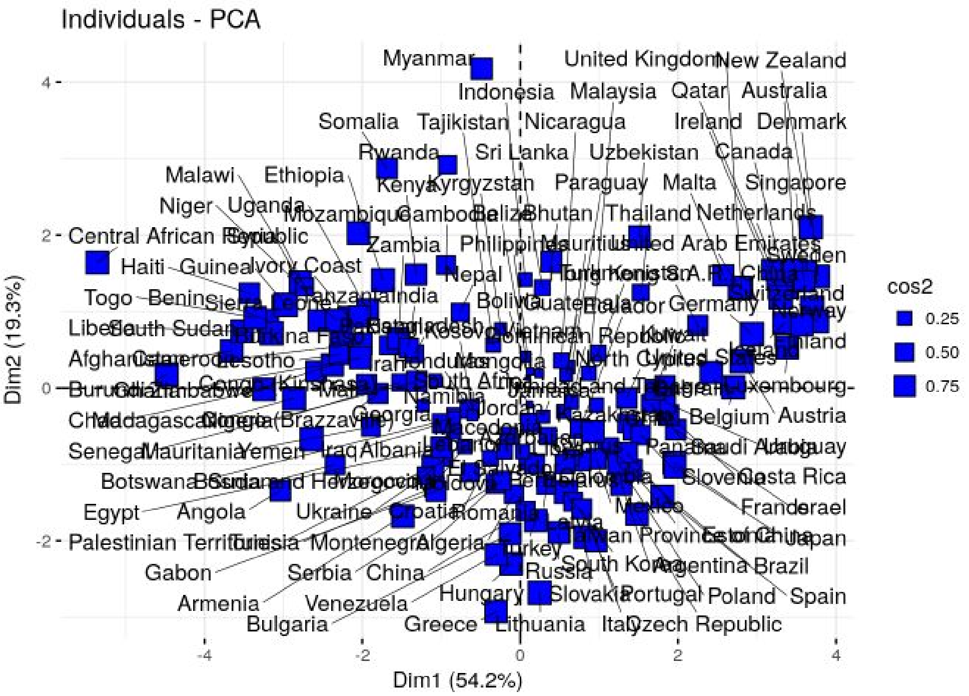
We also see that happiness score,Family, Economy, Health are more correlated with the first dimension whereas freedom, generosity,Trust in the Government are more correlated with the second dimension.

Cos2 shows the quality of representation

```
fviz_pca_ind (pca, pointsize = "cos2", pointshape = 22, fill = "blue", repel = TRUE)
```

Output:





But it’s not very visible

CONCLUSIONS

Running a Principle Component Analyses on this dataset has showed the high correlation between variables like GDP/Capita, Health, Family and the Happiness Score as these variables are represented by the first Principle Component. As these variables decrease the happiness Score also decreases dramatically due to the high level of correlation.

The other variables: Generosity, Trust in the Government and Freedom are also correlated, not in such a strong manner, but they influence the Happiness Score as they are comprised by the second Principle Component.

REFERENCES:

https://onlinecourses.science.psu.edu/stat505/book/export/html/49

https://www.kaggle.com/unsdsn/world-happiness

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/screeplot.html

http://www.gastonsanchez.com/visually-enforced/how-to/2012/06/17/PCA-in-R/