Final Project Report

Phi Dang (732080)

12 April, 2021

1 Introduction

Happiness and human well-being have been such a universal concern as they produce a more impartial and balance motivation to a healthy economy. GDP is often considered a key factor to raise happiness. However, many happy countries are generating sustainable development based on trust, freedom, and life expectancy as well. On March 20th, the United Nations proclaimed the International Day of Happiness to emphasize the happiness as the ultimate goal in the lives of human beings. Since then, happiness become an indispensable part of every policy decisions, especially restriction and lockdown policies in the current virus situation.

First publised in 2012, the World Happiness Report measures the state of happiness through specific indicators. The report has gained attentions from governments and many global organizations and become effective assessment criteria of a nation's development. As an data analysis enthusiast in Finland, I am curious about underlying components inducing the fact that Finland is the happiest country in the world. In this report, the lastest version of data published in 2019 is utilized to analyze the relationships among attributes.

2 Research questions

This project aims to propose detail answers for the following questions:
1. How different attributes correlated to each other?
2. How could the variance in the data be performed with fewer components? What does this new performance tell?
3. Is it possible to have an unsupervised clustering method for this data?

3 Univariate Data Analysis

The dataset consists of attributes as follows:
- Rank: Rank of the country based on the Happiness Score
- Country: Country or Region
- Score: A metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."

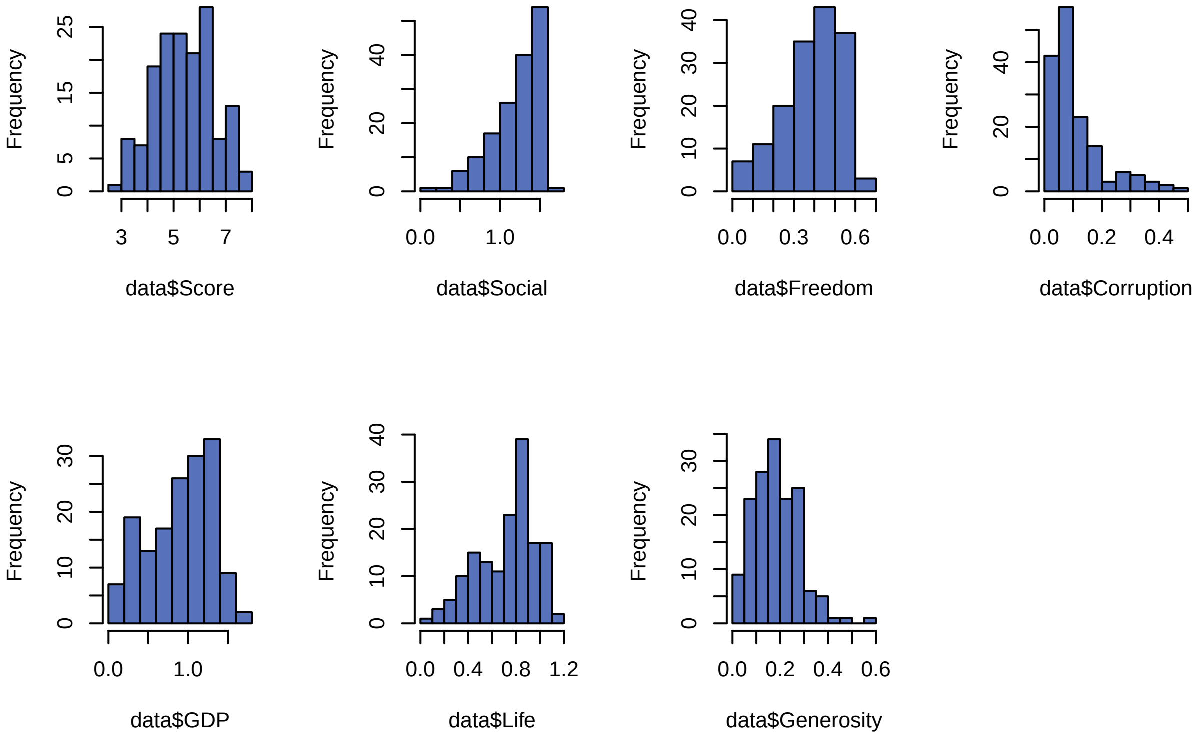
The following attributes contributes to the calculation of the Happiness Score. Note that it might be unreliable to build a Happiness prediction model based on these attributes.
- GDP
- Social: Social supports
- Life: Healthy life expectancy
- Freedom: Freedom to make life choices
- Generosity

- Corruption: The perception of corruption. The higher the value is, the less corrupt that the country perceives.

Let's take a look at the dataset:

```
##   Rank     Country Score   GDP Social  Life Freedom Generosity Corruption
## 1    1     Finland 7.769 1.340  1.587 0.986   0.596      0.153      0.393
## 2    2     Denmark 7.600 1.383  1.573 0.996   0.592      0.252      0.410
## 3    3      Norway 7.554 1.488  1.582 1.028   0.603      0.271      0.341
## 4    4     Iceland 7.494 1.380  1.624 1.026   0.591      0.354      0.118
## 5    5 Netherlands 7.488 1.396  1.522 0.999   0.557      0.322      0.298
## 6    6 Switzerland 7.480 1.452  1.526 1.052   0.572      0.263      0.343
```

To have a closer observation of data's distribution, let's look at the following histograms of all attributes:

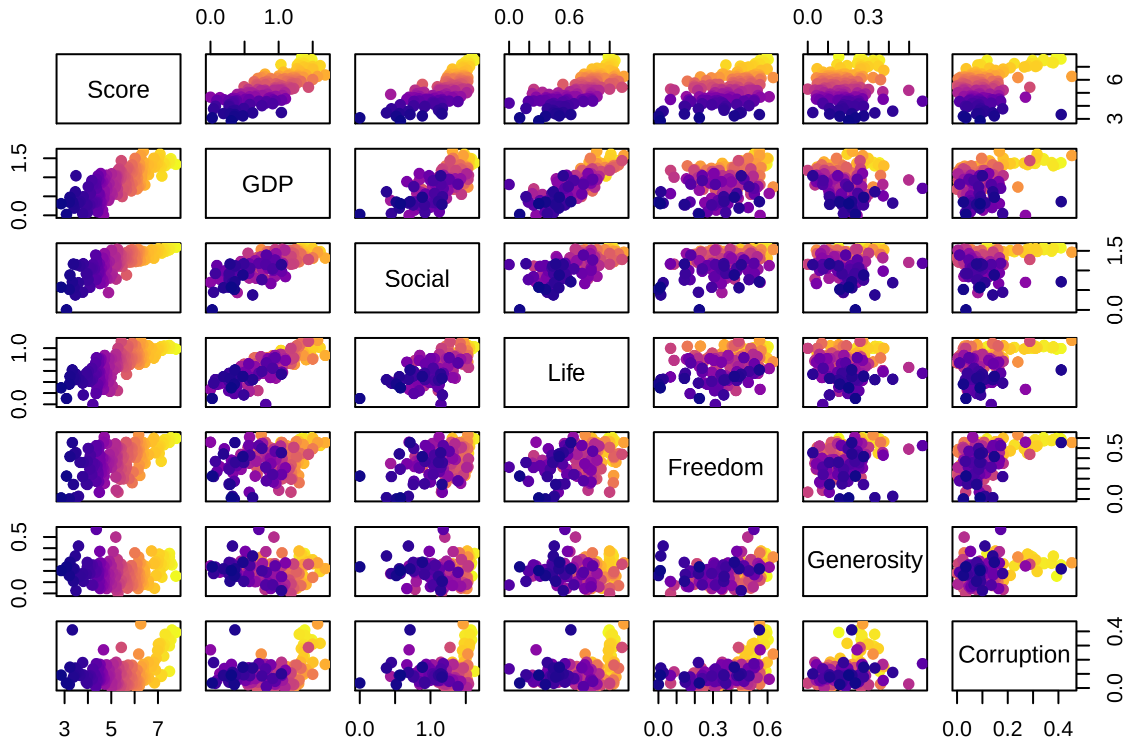


Some useful statistics could be found in the summary of the dataset:

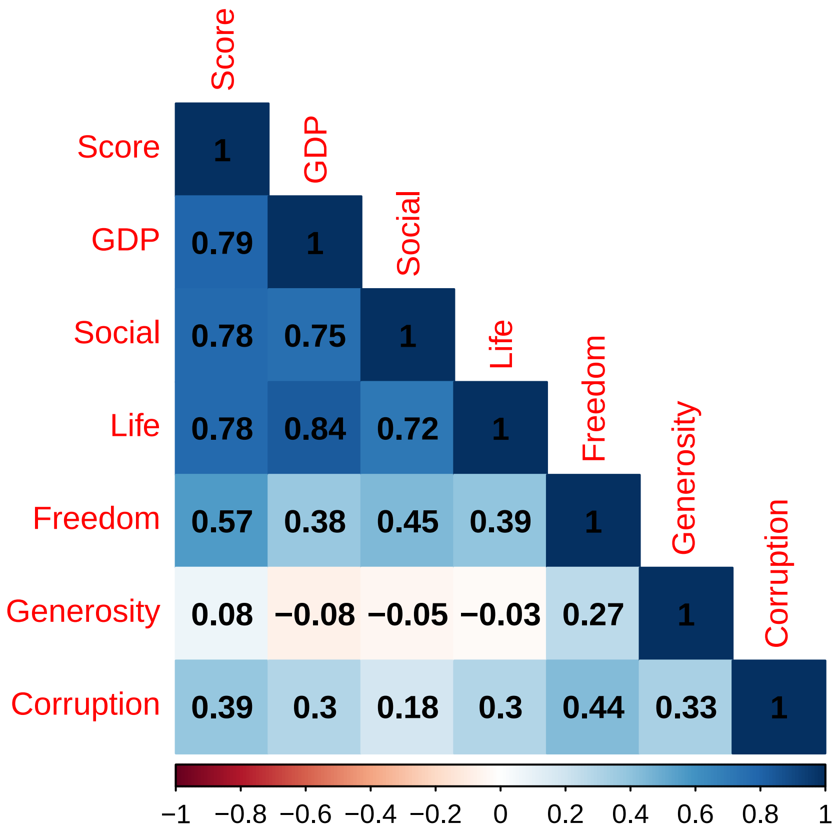
```
##      Score            GDP              Social           Life       
##  Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477  
##  Median :5.380   Median :0.9600   Median :1.272   Median :0.7890  
##  Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252  
##  3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818  
##  Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410  
##     Freedom         Generosity        Corruption    
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.3080   1st Qu.:0.1087   1st Qu.:0.0470  
##  Median :0.4170   Median :0.1775   Median :0.0855  
##  Mean   :0.3926   Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.5072   3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.6310   Max.   :0.5660   Max.   :0.4530
```

Some observations:
- The Score is almost normally distributed. It could be noteworthy that average Happiness Score all over the world is about 5.4.
- Freedom, Life, GDP, and especially Social are left-skewed. This indicates most of the countries succeeded in these criteria.
- Corruption and Generosity are right-skewed. This proposes that not many countries have accomplished these attributes and they might be indicators to make differences in Happiness Score.

4 Bivariate Data Analysis



From the pairplot, we can see that Score seems to have strong relationship with GDP, Social, and Life. Score's relationship with Freedom is slightly weaker, and with Generosity and Corruption is much weaker.



The correlation matrix provides a better insight of the associations among attributes. Besides outstanding correlations of GDP, Social, and Life with the target varable - Score, we can also see that GDP highly correlated to Social, Life. This happens to Social and Life as well, raising a possibility of multicollinearity issue. Fortunately, we perform PCA as a multivaraite analysis method in this project, which expresses the data in fewer component that no longer has multicollinearity.

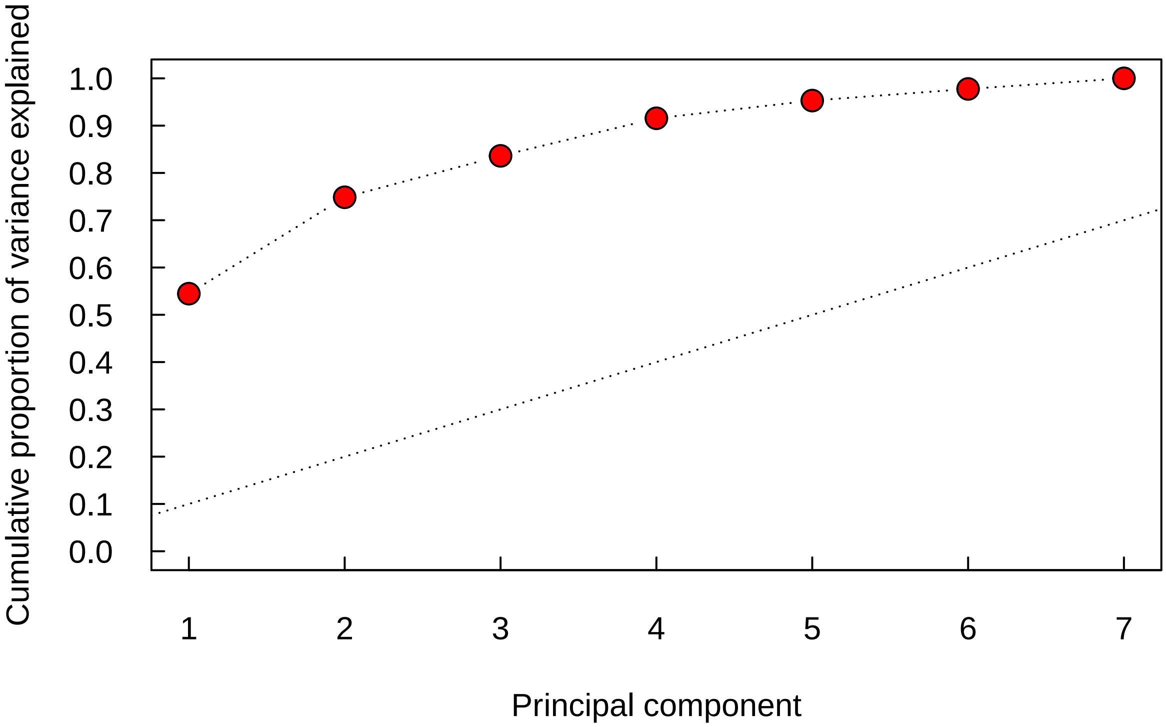
5 Multivariate Data Analysis

a) Principle Component Analysis

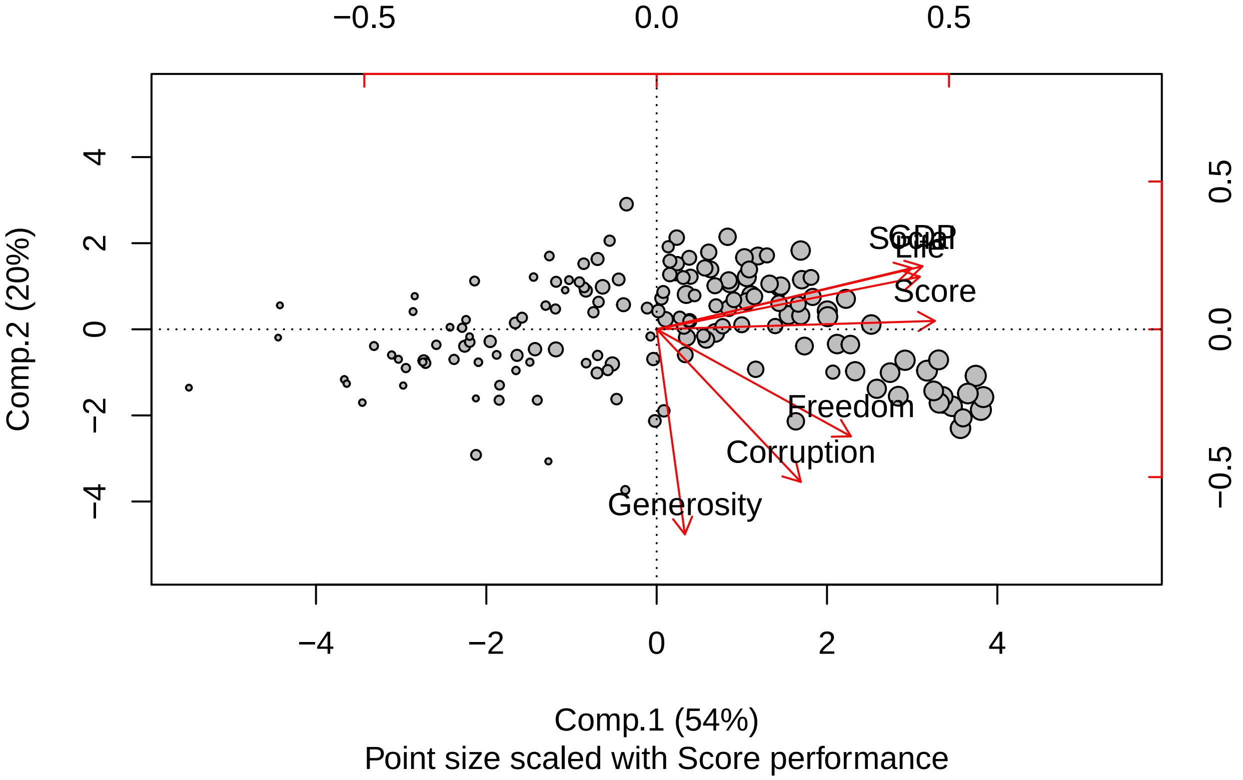
Principal Component Analysis (PCA) looks for few linear combinations of p variables, losing in the process as little information as possible. More precisely, PCA transformation is an orthogonal linear transformation that transforms a p-variate random vector to a new coordinate system such that, the obtained new variables are uncorrelated, and the greatest possible variance lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. (Pauliina, 2021)

In fact, PCA transformation is highly sensitive for scaling of the variables. One can address this problem by standardizing the variables first. The data can be standardized by subtracting the sample mean $\overline{x}$, and then dividing each variable by the corresponding square root of the sample variance. In R, one can simply use scale() function to preprocess data.

With R's princomp package, we can transform the dataset from seven attributes into fewer components. The plot below shows the cummilative proportion of variance explained by each number of components:



More than 70% of the variance is explained by only two first components. Let's visualize the scores produced by them:



Some interpretation:
- PCA plot illustrates cluster of samples based on their ranks. The right-half region contains bigger datapoints than the left-half region. Inside the right-half region, bigger datapoints are in the lower half.
- Principle component 1, which explains most of the variance (54%), is strongly influenced by Score, Life, GDP, Social. The angles among these attributes are also small, suggesting close correlation between one with each other.
- Principle component 2 is strongly influenced by Generosity.
- Corruption and Generosity are not likely to be correlated to Score as the angles are almost 90 degrees.
- Freedom is also noticeable as biggest points tend to be dragged to this indicator.

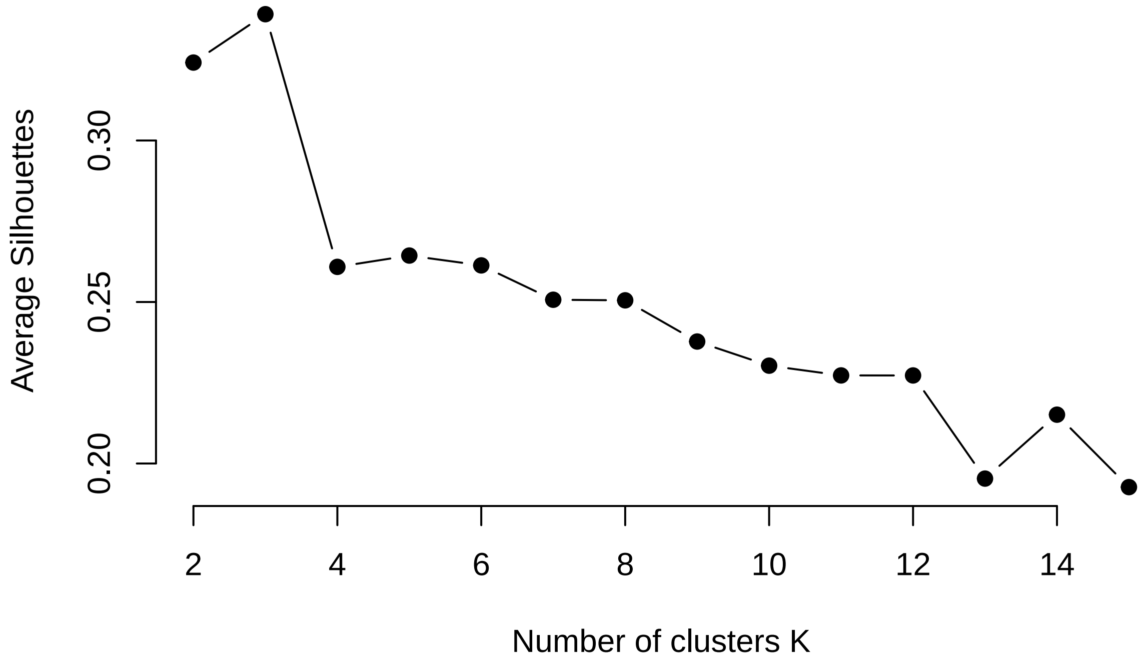
b) K-means clustering

The target of the K-means algorithm is to divide datapoints in K clusters so that the within-cluster sum of squares (withinss) is minimized. Moving Centers Method is utilized as follows (Pauliina, 2021):
1. Choose randomly k data points $c_1, ...c_k$ out of $x_1, ..., x_n$.
2. Define k sets $A_1, ...A_k$ such that $A_t = \{x_i | d(x_i, c_t) \leq d(x_i, c_j), for\ j \neq t\}$.
3. Calculate new centers $c_1, ..., c_k$ of the sets $A_1, ..., A_k$
4. Repeat steps 2 and 3 until convergence.

There are several considerations:
- What distance is the most appropriate?
- How to define center?
- Which k is the best one?

Using R's kmeans package, the algorithm of Hartigan and Wong (1979) is used by default. It exploits Euclidean distances and define centers as the mean of their Votonoi sets.

For determining the best k, the project used Average Silhouette Method. In a nutshell, the average silhouette method assesses the clustering's accuracy. In other words, it establishes how well each object fits into its cluster. A successful clustering is shown by a high average silhouette diameter. For various values of k, the average silhouette formula computes the average silhouette of observations. Over a set of potential values for k, the ideal number of clusters k is the one that maximizes the average silhouette.

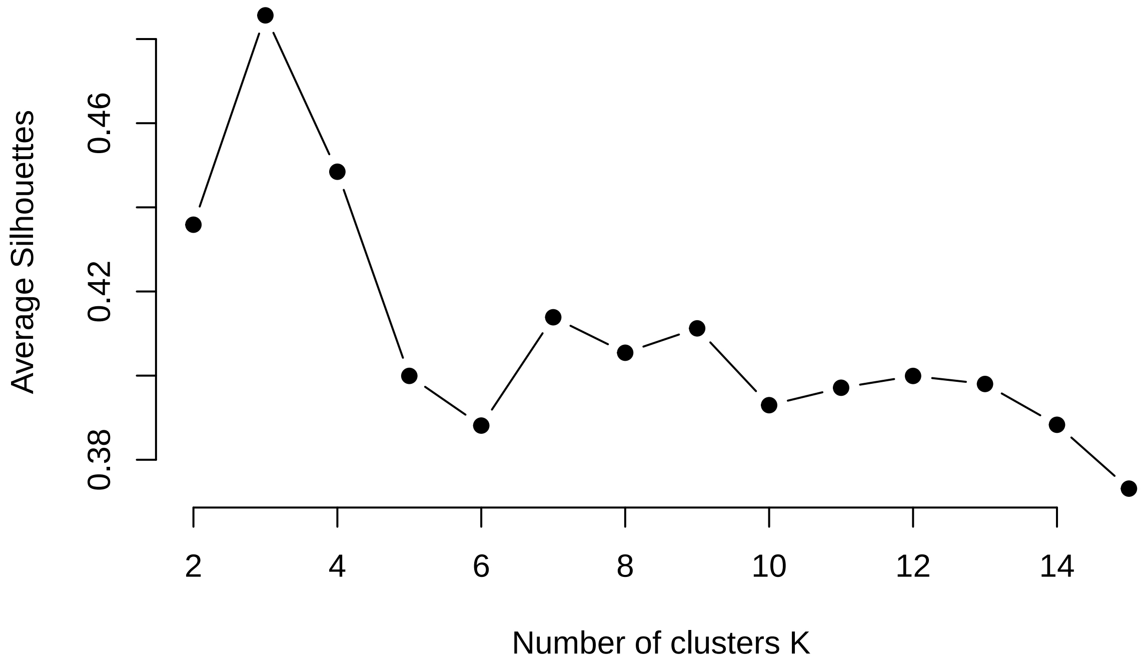


Now we can use K-means clustering with $k = 3$:

```
data.kmeans2 <- kmeans(data.scaled,centers=3,nstart=25)
print(paste('Total variance explained by clustering:',
            round(data.kmeans2$betweenss/data.kmeans2$totss,2)*100,'%'))
```

```
## [1] "Total variance explained by clustering: 55 %"
```

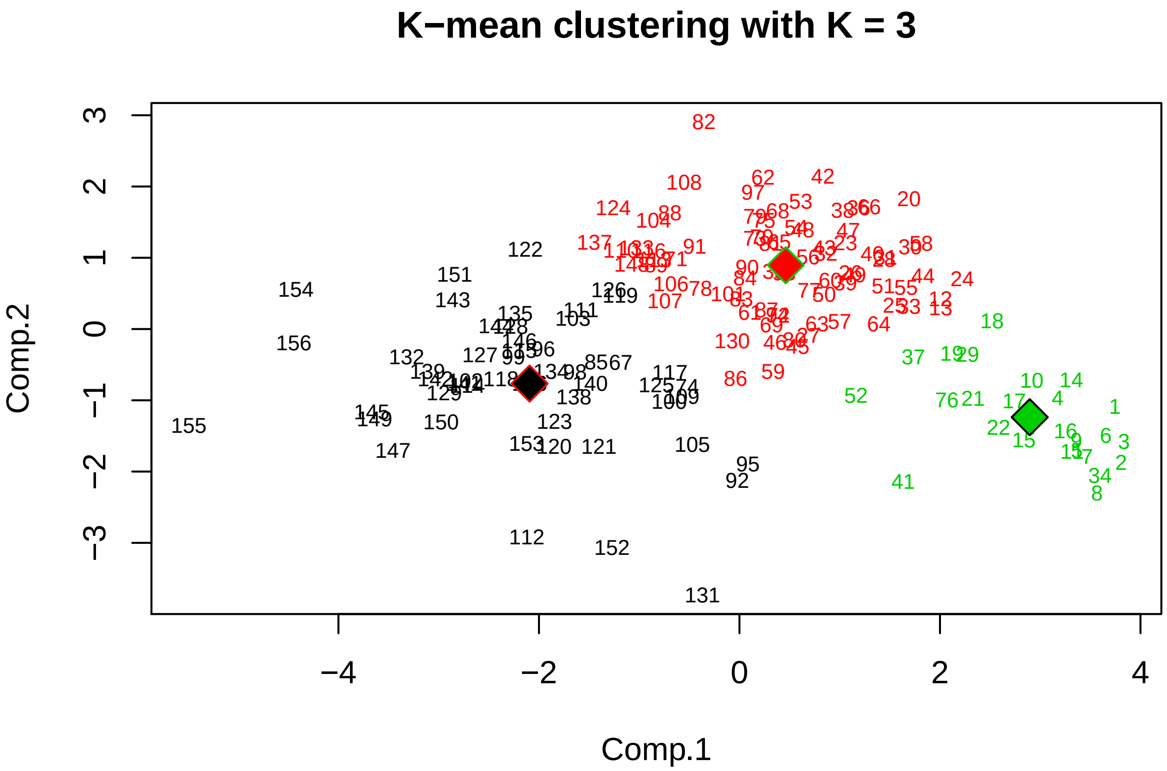
The variance explained by this clustering way is not really high. Bear in mind that PCA-transformed data is available, we can perform K-means clustering on the first two components. Starting with Silhouette once again to find the ideal k:



```
data.kmeans <- kmeans(data.PC1PC2,centers=3,nstart=25)
print(paste('Total variance explained by clustering:',
            round(data.kmeans$betweenss/data.kmeans$totss,2)*100,'%'))
```

```
## [1] "Total variance explained by clustering: 72 %"
```

The total variance explained increase significantly. In two-dimensional case, the results can be easily illustrated with a scatter plot:



We can observe that the red cluster contains most of high-ranking countries in term of Happiness Score. Likewise, the green cluster and the black cluster contain entries with middle and low ranks, respectively.

6 Conclusion

With all above analyses, answers for the research questions will be summarized as follows:
1. According to the bivariate data analysis, Score has strong relationships with GDP, Social, and Life. This goes well with a common perception that money could buy happiness. Having a good social network and healthy life is also important to be happy.
2. The dataset can be performed with two components capturing more 70% of total variance by applying PCA to the standardized data. Once again, GDP, Social, and Life appear to be dominating indicator at this point. Freedom could also be a good measure of happiness.
3. K-means is an reasonable clustering method for this dataset as long as it goes with PCA transformation.

7 Critical Evaluation

Some ecomonists have challenged the notion that a survey can capture subjective well-being. They found that people's evaluation of happiness might be influenced by their country's school system grades tests. This might induce bias in happiness score and lead to misleading resuilts.

Futhermore, some similar work has suggested the possibility that Corruption would be a good predictor while Social would be a bad one. This might be capture if the PCA in this project is analyzed deeper by looking at other components rather than just the first two.