# What makes a happy country?

*Jingwen Zhang (17-746-487)*

*04/01/2019*

## 1 Introduction

In *Anna Karenina*, Leo Tolstoy wrote, "all happy families are alike, each unhappy family is unhappy in its own way". His beginning word is quoted many times in different ways and nowadays we even have so-called "Anna Karenina principle". When applying this vulnerability of happiness to countries, is it true that happy countries share a common set of reasons which contribute to happiness, while there is any number of ways leading a country to be unhappy?

To answer this question, the World Happiness Index offers us the first glance. The logic behind the ranking of happiness is that we seem to believe there is the happiest country in the world, as Finland is for 2019 since it scores highest in the latest World Happiness Report published by the United Nations in 2019, happiest countries should be alike since they score top in mostly all attributes for the calculation of happiness.

However, this assumption ignores that combinations of attributes and the importance of each attribute can be different. By analyzing the dataset offered in this report, which covers 156 countries from 2008 to 2018 including 26 variables, data will show their patterns and tell us the similarity or dissimilarity among happy countries or unhappy countries. To find the importance of variables and similarities between countries, the most commonly used unsupervised machine learning methods for these purposes are principal component analysis and clustering because it helps find the most important dimensions and cluster

similar countries on these dimensions. Based on these methods, the paper is to find out *what are the most important reasons for a country to be happy*?

# 2 Data

The data are downloaded directly from the official website of "World Happiness Report", this project is a survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be, it is updated annually since 2012 (World Happiness Report, 2019).

The original dataset contains in total 1704 observations and 26 variables from 2008 to 2018 for each country. Since this essay focuses only on the latest data, the subset of the dataset for the year 2018 is selected, dropping all sparse or blank columns and descriptive variables such as residuals, standard deviations of several main indicators. The variable called "life ladder" is excluded because it might predetermine patterns of the analysis here, since its value is very similar to the final happiness score, and the analysis will compare our results with the final score in the end. After the basic selection based on the description of indicators (Helliwell, Huang and Wang, 2019: 1), 11 numeric variables are included, which are shown in Table 1[1], they are log GDP per capita, social support, healthy life expectancy at birth, freedom to make life choices, generosity, perceptions of corruption, positive affect, negative affect, confidence in national government, GINI index (average 2000-16), GINI of household income. [2]

In this analysis, we can not omit cases with missing values because each of them represents a country, instead, they are replaced by the median for each column.[3] A duplicated dataset only includes numeric variables is created because categorical variables can not be processed

[1]Before standardization of data

[2]Positive affect is defined as the average of three positive affect measures in GWP: happiness, laugh and enjoyment, negative affect is defined as the average of three negative affect measures in GWP. They are worry, sadness and anger

[3]There are in total 136 observation in the subest of the original dataset for 2018

by the 2 unsupervised machine learning methods selected in this paper. Since indicators have different scales, and methods are sensitive to this, I standardized all of them in the duplicated numeric dataset.

Table 1: Descriptive statistics

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| GDP | 136 | 9.26 | 1.15 | 6.54 | 8.54 | 10.13 | 11.45 |
| Support | 136 | 0.81 | 0.12 | 0.48 | 0.74 | 0.91 | 0.98 |
| HLE | 136 | 64.72 | 6.63 | 48.20 | 59.47 | 68.93 | 76.80 |
| Freedom | 136 | 0.78 | 0.12 | 0.37 | 0.72 | 0.88 | 0.97 |
| Generosity | 136 | −0.03 | 0.15 | −0.34 | −0.14 | 0.05 | 0.50 |
| Corruption | 136 | 0.73 | 0.18 | 0.10 | 0.69 | 0.85 | 0.95 |
| Positive | 136 | 0.71 | 0.11 | 0.42 | 0.64 | 0.79 | 0.88 |
| Negative | 136 | 0.29 | 0.09 | 0.09 | 0.22 | 0.36 | 0.54 |
| Government | 136 | 0.49 | 0.19 | 0.08 | 0.35 | 0.62 | 0.99 |
| GINIaverage | 136 | 0.38 | 0.08 | 0.21 | 0.33 | 0.43 | 0.63 |
| GINIhousehold | 136 | 0.46 | 0.13 | 0.20 | 0.37 | 0.55 | 0.79 |

# 3 Methodology

In my analysis, principal component analysis (PCA) will be used first and k-means clustering will be applied based on dimensions derived from PCA. PCA inspects patterns among variables and the latter one is for patterns among cases (e.g. countries), but the research question requires a linkage between variables and countries so a combined method between PCA and k-means cluster is used.

Generally speaking, PCA is used for data reduction because it can identify patterns in quantitative variables we are interested in and reduce dimensions of the data in multiple regression among others. Since there are 11 variables therefore 11 dimensions, if I focus on all of them it might result in "overfitting", and it is also difficult to plot cases on all 11 dimensions in the visualization. Since the analysis focuses on similarity between each country, clustering is a suitable method, it focuses on finding similarity/dissimilarity between cases

and try to put them into different grouped based on their similar characteristics. K-means clustering is chosen among other clustering methods such as the commonly used hierarchical clustering because I prefer clusters to be flexible here, and retain the optimal result. Before moving forward to the interpretation, a brief explanation of how they work is given as following.

PCA: First we retrieve the covariance matrix of all 11 variables. Based on the matrix, we can attain eigenvectors and eigenvalues, they can be understood as the direction and magnitude of our data. PCA is like projecting high-dimensional data on a low-dimensional space. We sort the eigenvalues from on a descending order so that it gives us the components in order of importance, and we multiply them accordingly with the scaled version of the original dataset and retain a new matrix. In this matrix, each component is a combination of the original variables weighted by the eigenvector and it is independent of one another since each eigenvector is also independent of one another. In R these can be accomplished by several functions, for example, this analysis uses PCA function. Since we want to keep as less as possible variables and ensure they explain as much as possible of the total variance, we need to also calculate the proportion of variance explained by each component, for example, if the first 2 components explain in total 60% which we think it is sufficient, then we retain 2 components and drop other components. We will also calculate loadings of each variable in these 2 components, we drop variables in which loadings are smaller than 0.35 for instance, to find the most important variables for the first 2 components. Loadings of those 11 variables in each component can be understood as coefficients in multiple regression.

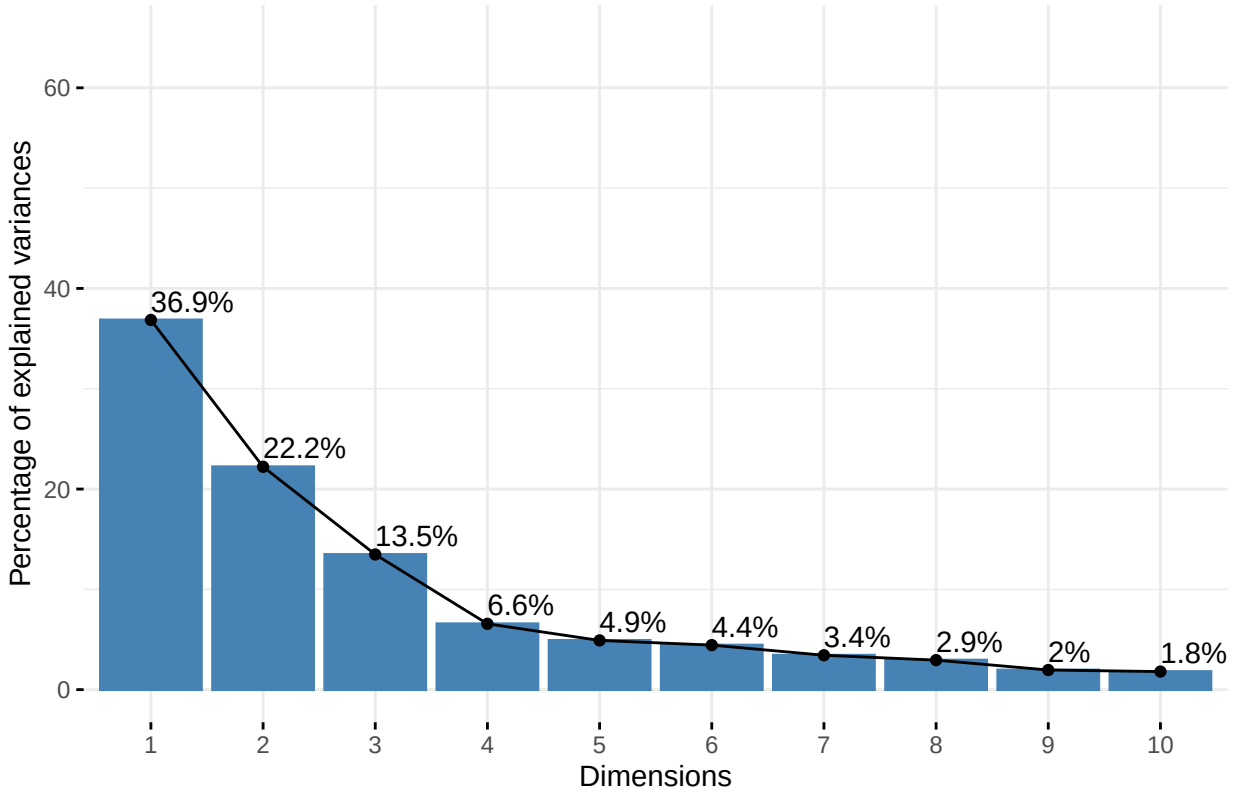
K-means clustering: As one of the clustering methods, k-means literally refers to attaining "means" of the data to find the centroid of clusters and allocates observations to the nearest cluster. The K-means algorithm is an iterative algorithm, given by the number of clusters we required, it begins with randomly selected k centroids and repeats until the convergence is reached when positions of centroids no longer move. Before the process, we need to define how many clusters we would like to retain. By increasing the number of clusters "k", we

preserve more information so precision will increase but we lose the efficiency of the method. There are several ways to determine k, in this essay, I use function "mclustBIC" from the package "mcluster". It works by setting the minimal and maximal number of clusters and it will compare BIC values of different k and show the optimal value of k which has the smallest absolute BIC values among other options.

# 4 Results

As we can see in the scree plot that the first 2 components explained 59.1% of the total variance, which is slightly lower than my threshold 60%. Since the general goal is to reduce the number of components, therefore, only the first 2 components will be retained.

Figure 1: Scree Plot




As shown in Table 2, since I want to include variables in which absolute values of their

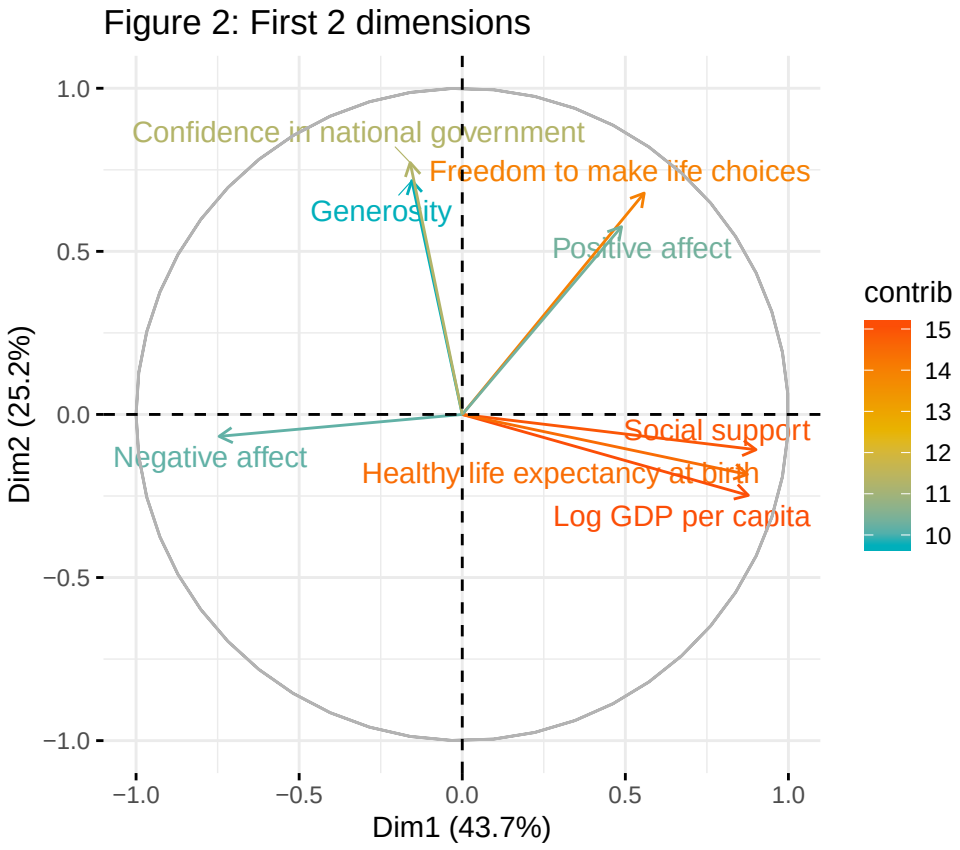
loadings are higher than 0.35 in either one of the components, perception of corruption, GINI index, GINI of household income will be neglected in further interpretation. Based on the reduced version of the dataset, I put important variables on the 2-dimensional graph as shown in Figure 2. The first principal component explained approximately 43.7% of the total variation, which represents mainly variables like log GDP per capita, social support, healthy life expectancy at birth since they have the highest values of loadings in the first component. If a country has higher values in log GDP per capita, social support, healthy life expectancy at birth, it is likely to score higher in the first component. The first component has a emphasis on the socioeconomic conditions in one country.

Table 2: Loadings on the first 2 components

| | First component | Second component |
|---|---|---|
| Log GDP per capita | 0.4373698 | 0.1095311 |
| Social support | 0.4357200 | 0.0408444 |
| Healthy life expectancy at birth | 0.4392845 | 0.0704962 |
| Freedom to make life choices | 0.2455780 | -0.4419593 |
| Generosity | -0.0677013 | -0.3960190 |
| Perceptions of corruption | -0.2170431 | 0.3477810 |
| Positive affect | 0.1708292 | -0.4177775 |
| Negative affect | -0.3716831 | 0.0498791 |
| Confidence in national government | -0.0590853 | -0.4713195 |
| GINI index (World Bank estimate), average 2000-16 | -0.2254361 | -0.1422335 |
| gini of household income reported in Gallup, by wp5-year | -0.3037133 | -0.2980869 |

The second principal component explained a further 25.2% of the total variation and appears to represent confidence in the national government, freedom to make life choices. If a country has higher values in confidence in the national government, freedom to make life choices, it

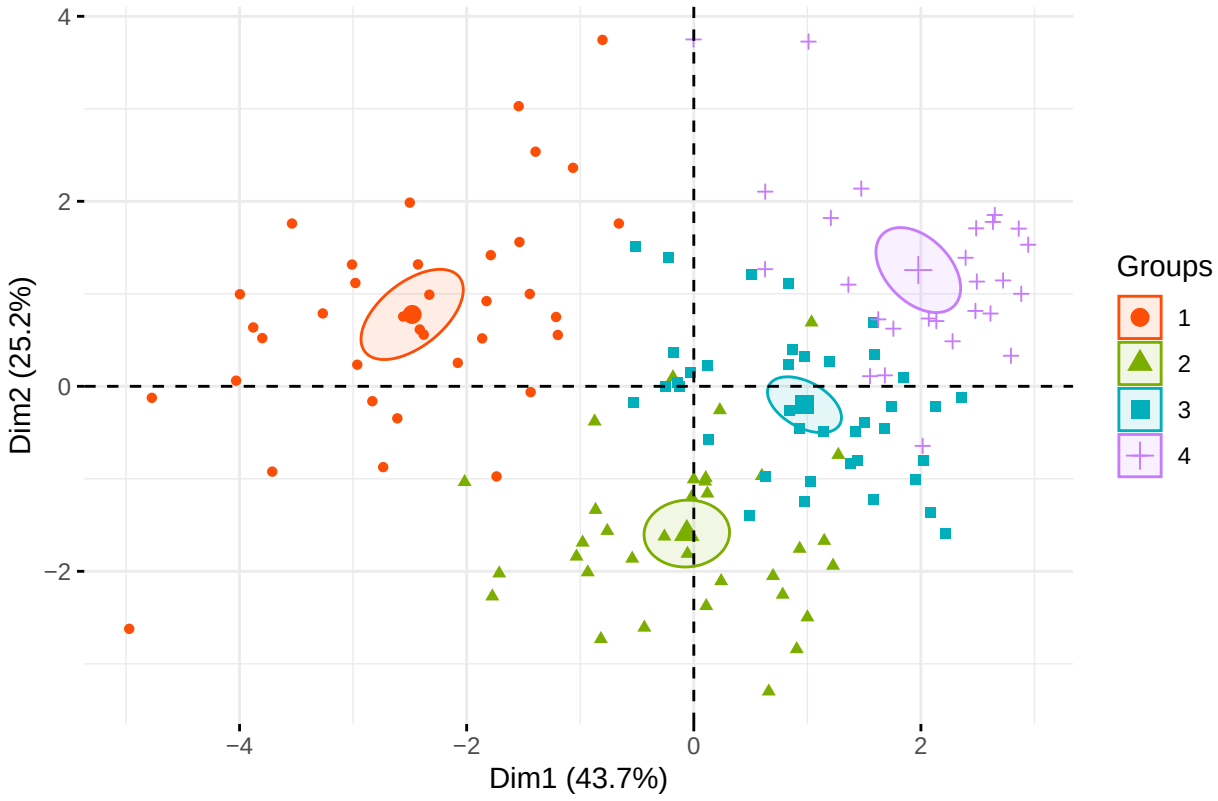
is likely to score higher in the second component. The second component has a focus on psychological feelings about the social values and environment in one country. As the first component is more important than the second one, I assume that data patterns convey that socioeconomic conditions are more important than social values or ideology in our dataset.




This paper aims to find out patterns of relationship between the characteristics of variables and different countries. Clusters of similar countries will be plotted on this 2-dimensional space for comparison. The minimal and maximal number of clusters I require is 2 and 5, it turns out the the best k among others is 4. Therefore, as shown in Figure 4, 4 clusters of countries are placed on the same dimensional space as in Figure 3. The points in different colour inside circles represents the centroid of different clusters in different colours, points with different shapes are countries, their colours correspond to clusters they belong to. I assume cluster 4 represents top-ranking countries in happiness scores since its centroid is

located in very positive side of both dimensions. According to distance of different centroids to the centroid of cluster 4, it should be followed by cluster 3, cluster 2 and cluster 1, the least similar. According to my assumption, I will plot these clusters on the world map to see whether it is in line with our expectations, as shown in Figure 4.

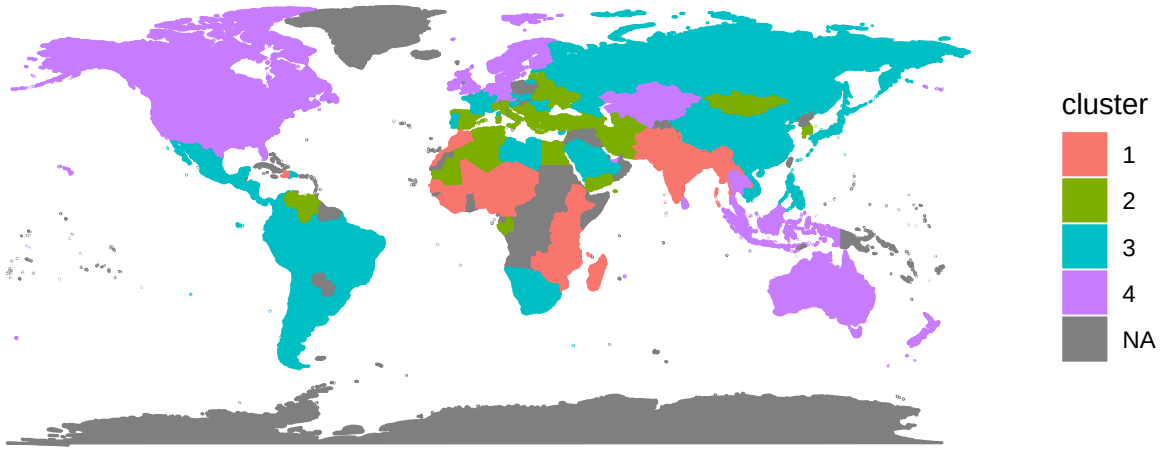
Figure 3: Clusters on the first 2 dimensions




Based on the results of PCA and K-means clusters, although cluster 4 located in very different regions, varying from developed countries to developing countries, have different ethnicities, languages, religions, economic conditions, but most of them are still located in western Europe and North America. Cluster 3 seems more homogenous since it contains mainly developing countries like China, Brazil, Russia, etc. Cluster 2 is mainly located in areas that encountered the Arab Spring or the refugee crisis in recent years. Cluster 1 mainly contains countries with ongoing armed conflicts or less-developed countries. It seems that the happiest countries are very different from each other but the most unhappy countries

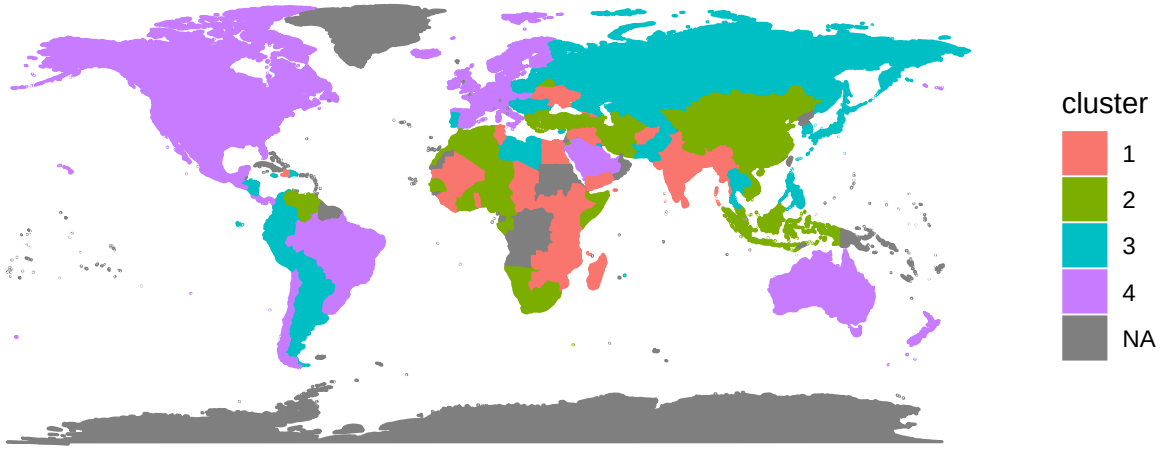
are similar.

Figure 4: Application of K−means Clusters on World Happiness Index
Based on data from World Happiness Report 2019




As a reference, I use the final happiness scores from the report and divide all countries into 4 groups and plot them on the same map for comparison, as shown in Figure 4. Although 4 classes are somewhat rough for displaying scores, this is to be in accordance with 4 clusters generated before.

Figure 5: Clusters on World Happiness Index
Based on data from World Happiness Score 2019




As shown in Figure 5, compared to clusters generated before, it appears that the happiest

countries are more heterogeneous than in the previous map.

Happiness of a country is just as realistic as our lives. No matter how high a country reaches in the second dimension (in Figure 2), as long as it is not at the right side of the baseline "0", it still belongs to an very unhappy cluster e.g. cluster 1. Once a country is at the right side of the base axis of the first dimension, it will belong to a happier cluster if it scores higher in the second dimension. This implies that economics and health are basic conditions for a country to be happy since they have the highest loadings in the first dimension shown in Table 2. Only these have been fulfilled, indicators like freedom, generosity, and confidence in the government can start to help a country be happier.

# 6 Conclusion

Based on the data given by the Happiness Index and the results above, the most important reasons for a country to be happy are good economics and health conditions. There are many ways towards happiness, such as improving freedom to make life choices, generosity, confidence in the government, etc, as listed both in the first and the second dimensions, but if basic socio-economic conditions can not be ensured, it leads directly to unhappiness. Indicators listed in the second dimension can be considered as secondary needs, compared to the primary needs in the first dimension. The logic of happiness is similar to Maslow's hierarchy of needs, the most important reasons behind happiness - good economic and health conditions, they corresponds to the first 2 levels of needs, which are physiological and safety needs. In the future the research can combine theories from Maslow's hierarchy of needs with the happiness index by applying supervised machine learning methods, we can understand happiness scores deeper in the future.