December 2, 2019

Happiness of countries around the world

-A study of common characteristics countries famous of happiness and the factors that matter

STAT456 report

Abstract:

This report aims to study common characteristics and influential factors of happiness of countries around the world by investigating the 6 well-being measures provided by World Happiness Report. These variables cover the economic development, political situation, people's physical and mental healthy and precious things that people pursue such as freedom, generosity and support from others. The first question to be answered is what those common characteristic are of countries famous of happiness. The relationship between variables are investigated by data visualization using different kinds of plots. Then, since the WHR data are subjective answers to the nation poll, I wonder what the objective variables are that influence people's opinions and feelings. Data from World Development Indicators are collected and canonical correlation analysis is conducted to dig the reason behind. What's more, to illustrate the following results more succinctly, principal component analysis is conducted to reduce the dimensionality of dataset. The first principal components measures the overall development and people's feeling of happiness of a countries and the second one measures how much freedom and generosity can make up for happiness given other variables.Finally, focusing on the WHR data, clustering analysis is conducted to further investigate the characteristics of happiness of countries. Hierarchical clustering and scree plot are used to give suggestion of number of clusters. Considering the emerging influence of information technology, K-means clustering analysis is conducted on two datasets before and after 2012. Freedom and generosity can make up for the deficiency on economic and medical development of a countries. Since information technology is rapidly developed, countries with no advantage on this aspect suffer lower happiness.

Introduction:

In this analysis, I will use data from World Happiness Report and World Development Indicators. I compile them because I want to use data from World Development Indicators as supplement to study the factors that influence happiness of a country. WDI's data contains 1431 series of data covering 6 topics which are poverty and inequality, people, environment, economy, states and markets and global links. Although the dataset is very big, there are a lot of series with deficient entries. Therefore I chose to select series that are as complete as possible and related with World Happiness Report dataset.

World Happiness Report presents the global data on national happiness and reviews evidence that the quality of people's lives can be reliably assessed by a variety of subjective well-being measures. And World Development Indicators is a compilation of relevant, high-quality, and internationally comparable statistics about global development. The list of all variables and their definitions are described in Table1 .

Table 1: Variable definitions

Variables	Definition
Ladder	Happiness score or subjective well-being;
Log GDP per capita	Log GDP per person
Healthy life expectancy at birth	Life expectancy; based on the data extracted from World Health Organization's Global Health Observatory data repository.
Social support	The national average of the binary responses (either 0 or 1) to the 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?"
Freedom to make life choices	The national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
Generosity	The residual of regressing national average of response to the GWP question"Have you donated money to a charity in the past month" on GDP capita.
Perception of corruption	The national average of the survey response to two questions in the GWP:"Is corruption widespread throughout the government or not?" And "Is corruption widespread within businesses or not?"

Variables	Definition
Access to electricity, rural	Access to electricity, rural is the percentage of rural population with access to electricity
Adolescent fertility rate	Adolescent fertility rate (births per 1,000 women ages 15-19)
Agriculture land(% of land area)	Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures.
Birth rate (per 1,000 people)	Crude birth rate indicates the number of live births per 1,000 midyear population.
Cereal yield (kg per hectare)	Cereal yield, measured as kilograms per hectare of harvested land, includes wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains.
Female to male labor force participation ratio	Ratio of female to male labor force participation rate is calculated by dividing female labor force participation rate by male labor force participation rate and multiplying by 100.
Maternal mortality rate	Maternal mortality ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births.
Population total	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
Population growth, (annual %)	Annual population growth rate. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Profit tax (% of commercial profits)	Profit tax is the amount of taxes on profits paid by the business.
Total tax rate (% of commercial profits)	Total tax rate measures the amount of taxes and mandatory contributions payable by businesses after accounting for allowable deductions and exemptions as a share of commercial profits.
School enrollment secondary	Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown.
Suicide mortality rate(per 100,000 population)	Suicide mortality rate is the number of suicide deaths in a year per 100,000 population. Crude suicide rate (not age-adjusted).

Missing values are common in real dataset and it's always a difficult problem we need to handle with before conduct further analytic methods. For the WHR dataset, I decide to separate it into two part regarding to timeline. The first part is from 2005 to 2011 and the second part is from 2012 to 2018. The reason why I choose 2012 to be the cutoff is that happiness of life satisfaction has increase between 1991 and 2011, suddenly declined after 2012 among adults and adolescents. Therefore, I mainly focus on data after 2012 and when I fill in missing values, I use the average of that variable's value from 2012 to 2018. After I

take the average, if there are still missing values, I will simply drop them since I don't have other source of information. Similarly, for the WDI dataset, I fill in the missing values with the average of that variable's value from 2016-2019. After I fill in missing values in both two datasets, I merge them to customize a dataset for this analysis.

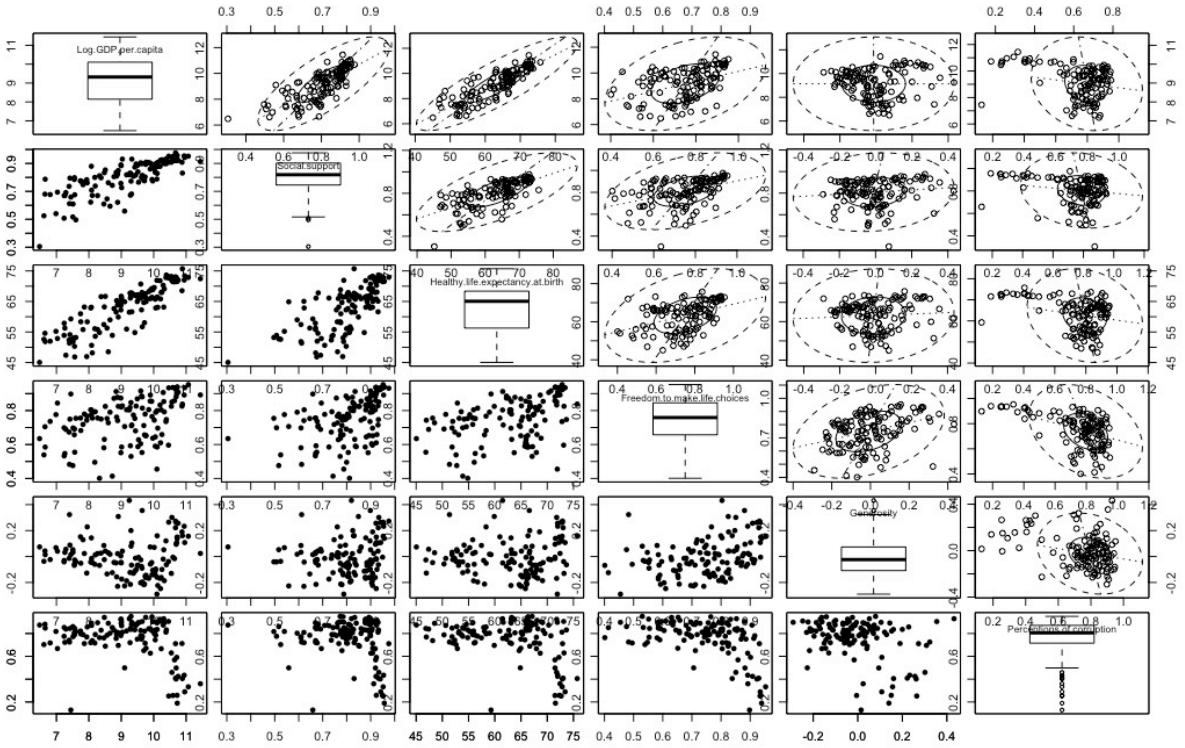
Goals:

1. **Investigate the relationship between variables and detect outliers.**
   - How the variables in WHR dataset are related with each other?
   - Is there any outliers in WHR dataset?
   - What are the common characteristics of TOP 3 countries in recent 5 years?
2. **Describe the datasets with canonical correlation analysis**
   - How to choose WDI variables that are correlated with well-being measures?
   - To what extent are those subjective well-being measures related with world development indicators?
3. **Cluster the countries according to the results of world happiness survey**
   - Is there any clustering patterns among all the countries?
   - Is there any difference between the clustering of 2005-2011 and that of 2012-2018?
   - How to illustrate the results with high dimensions?

Main results:

Goal1: Investigate the relationship between variables and detect outliers.

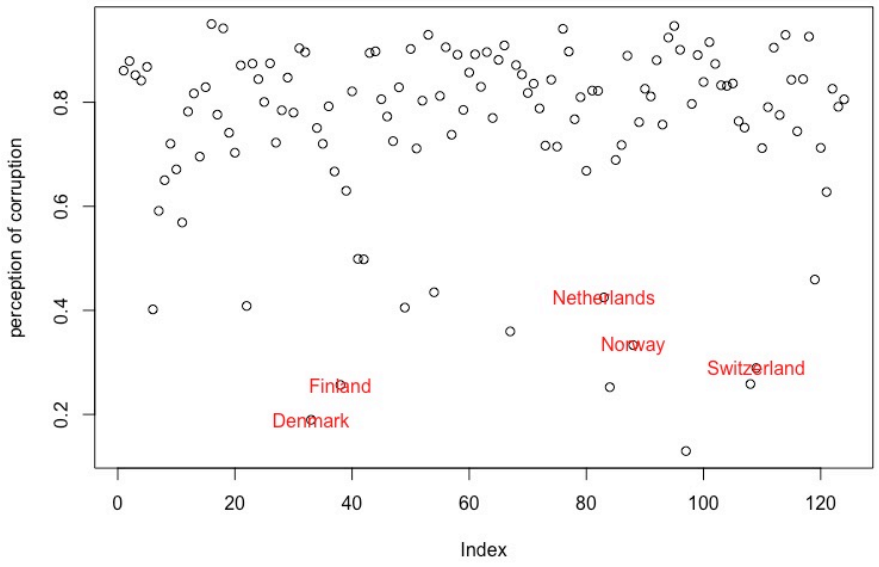
Figure 1. Scatterplot matrix



From Figure 1, we can see that the first three variables, Log GDP per capita, social support and healthy life expectancy of birth are highly correlated. The reason is probably that country with high GDP per capita have more advanced medical system and technology and their people enjoy better social welfare and education resources. Therefore, their life expectancy of birth is longer. And they are more likely to help others and family ties are tighter, thus leading to more social support. Relationship between them inspire me to conduct principal component analysis among them which is very helpful in results illustration.

The other three variables don't show apparent correlation which can be checked by chi-plot between each pairs among them. And there are many outliers in variable--perception of corruption. By looking up for them, I find that all of those TOP 5 countries with highest happiness score are outliers with apparently lower perception of corruption.

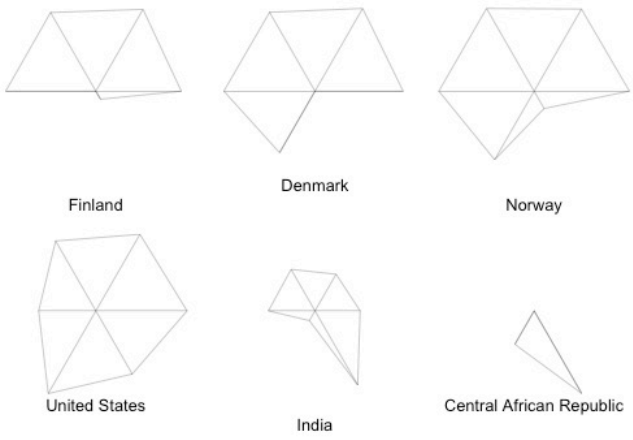
Figure 2 : scatterplot of perception of corruption



The TOP 5 happiest countries in recent 5 years are Denmark, Finland, Netherland, Norway and Switzerland. As depicted in the above scatterplot, they all have very low perception of corruption compared with the majority which is a common characteristic of them.

Figure 3: star plots of TOP 3 happiest countries and representative

Before analyzing those star plots, we need to know that star plot use relative range while computing. Therefore, the plot above compares TOP 3 happiest



countries with representative developed, developing and poorest countries. It shows that countries with higher Log GDP per capita, social support, life expectancy of birth and more freedom to make life choices. It's worthy to mention that Finland has very low generosity which is even lower that of India and Central African Republic. The reason may be that gap of wealth is small and rare people needs donation, thus leading to less generosity. Comparing TOP 3 countries with others, we can also see that they enjoy lower perception of corruption. Speaking of United States, the only disadvantage is corruption in government and business.

Goal 2: Describe the datasets with canonical correlation analysis

First, I select variables from WDI dataset, I choose those highly correlated with at least one of well-being measures and check its sufficient in the meantime. Finally, I choose eleven variables which are listed below in Table 2. These variables cover people, environment, economy, states and markets. Since data about poverty and inequality are very deficient, I didn't choose them in order to keep sample size not too small.

Table 2: canonical analysis WDI variables

WHR variables	Correlated WDI variables
Social support	1. Suicide mortality rate

WHR variables	Correlated WDI variables
Healthy life expectancy at birth	1. Birth rate 2. Maternal mortality rate 3. Adolescent fertility rate
Freedom to make life choices	1. Population total 2. Population growth 3. Ratio of female to male labor force participation rate 4. Agriculture land 5. Cereal yield 6. Access to electricity rural
Generosity	1. Log GDP per capita 2. School enrollment secondary
Perception of corruption	1. Profit tax

For social support, lower suicide rate may imply better mental health of people and tighter bind among people which sequentially implies that one feel more support from people around him/her. For healthy life expectancy at birth, those three variables are good criterion to evaluate the medical development of a country. For freedom to make life choices, variables about population and environment are good criterion of resources. More resources people have, more likely they are to make life choices instead of worrying about basic life condition. Ratio of female and male labor force participation rate is a gender statistic. The reason I select is that I think freedom of people of a country may partially conditioning on freedom of female of a country. Additionally, women's power in workplace can be a reasonable indicator of female's freedom. For generosity, I think people with better economic condition and higher education level may be more empathic and more willing to donate. For corruption perception in government and business, high tax rate means more income for people work in government and business which sequentially means more probable to corrupt and more profit.

Canonical correlation analysis is conducted between WHR variables and WDI variables. The canonical correlations are listed in Table 3.

Table 3: Canonical correlations

Canonical correlations					
0.5386697	0.3528146	0.3106746	0.2543119	0.2300860	0.1806128

We can see that the WDI variables are correlated with those well-being variables. But there still are something that cannot be explained by indicators which are more about people's feelings and personal opinions. These are very difficult to be observed or captured by explicit indicators. However, happiness is a subjective feeling itself which is also difficult to observed. That's why this national poll is necessary and important.

U are canonical variables for WHR data and V are those for WDI data which are listed in Table 4 and 5 below.

Table 4: Canonical variables U

U	"Log.GDP.per.capita"	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perception of corruption
U1	-10.386319	-5.898396	5.765494	40.67487	23.69495	-44.82417
U2	-7.630059	-5.094501	5.481091	31.86383	18.76211	-34.81389
U3	-10.301967	-6.068009	6.039311	40.59533	23.80467	-44.83613

U is mainly a contrast between freedom to make life choices and generosity on one hand, and perception of corruption on the other. Corruption implies press and bureaucratism which sequentially means that people may have little freedom and power. Therefore, U is best interpreted as a measure of more freedom and generosity and less corruption.

Table 5: Canonical variables V

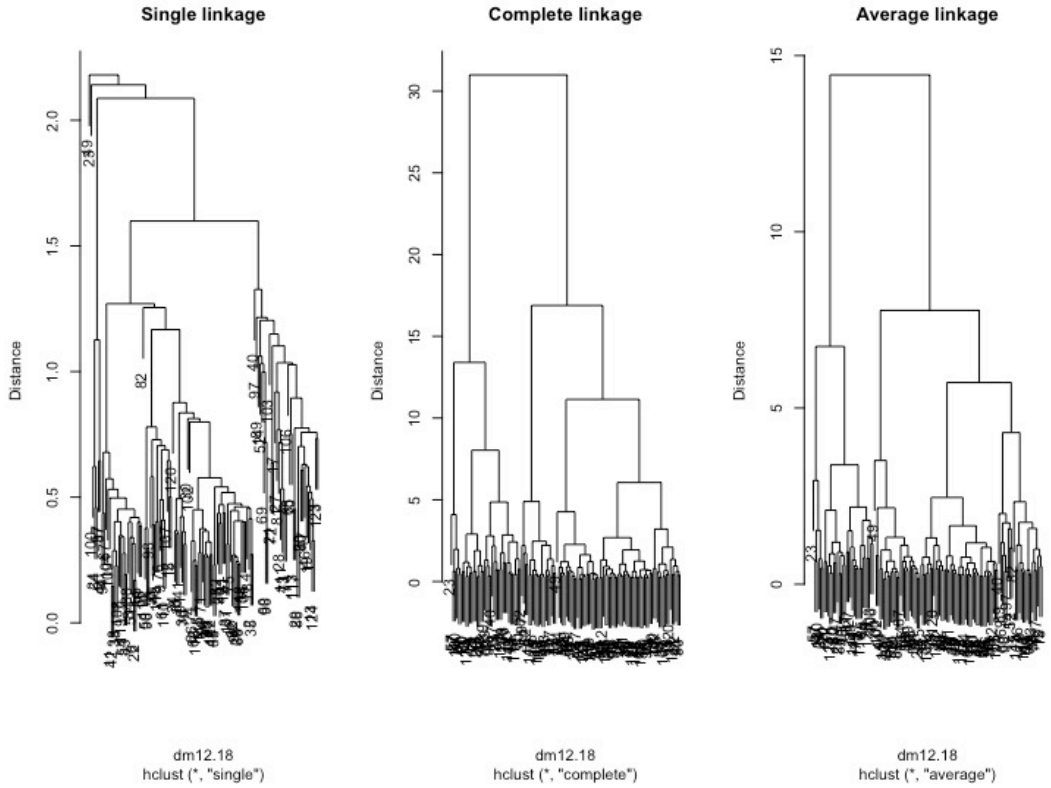
V	Access to electricity, rural	Adolescent fertility rate	Agriculture land	Birth rate	Cereal yield	Maternal mortality rate
V1	-327.9027	52.649158	221.40630	-42.341639	-59.600590	-27.43823
V2	-285.0150	7.287065	181.54985	-8.839524	-6.582801	-17.15795
V3	-187.8282	-15.745452	66.32685	-77.073183	-18.762942	-106.92968

V is mainly a contrast between agriculture land on one hand and access to electricity(rural population) on the other. Therefore, it can be interpreted as a measure of resources.

Goal 3: Cluster the countries according to the results of world happiness

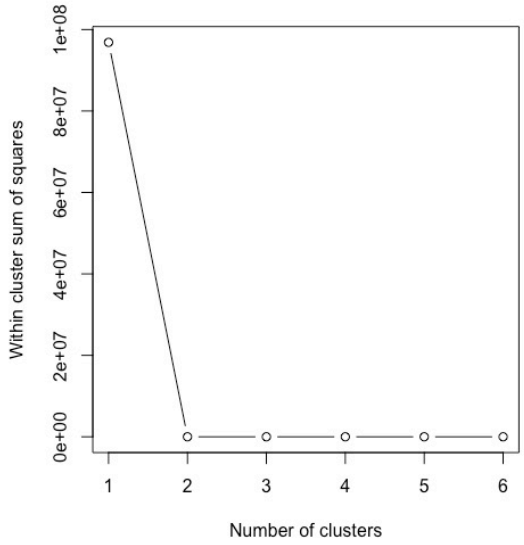
I'm interested in whether there are any cluster pattern of those countries. I use World Happiness Report data to conduct the analysis. In order to choose a reasonable K value in K-means clustering technique, I first use hierarchical clustering analysis to get a rough suggestion about the number of clusters. Meanwhile, I also draw scree plot to get a suggestion about the value of K.

Figure 4: hierarchical clustering analysis



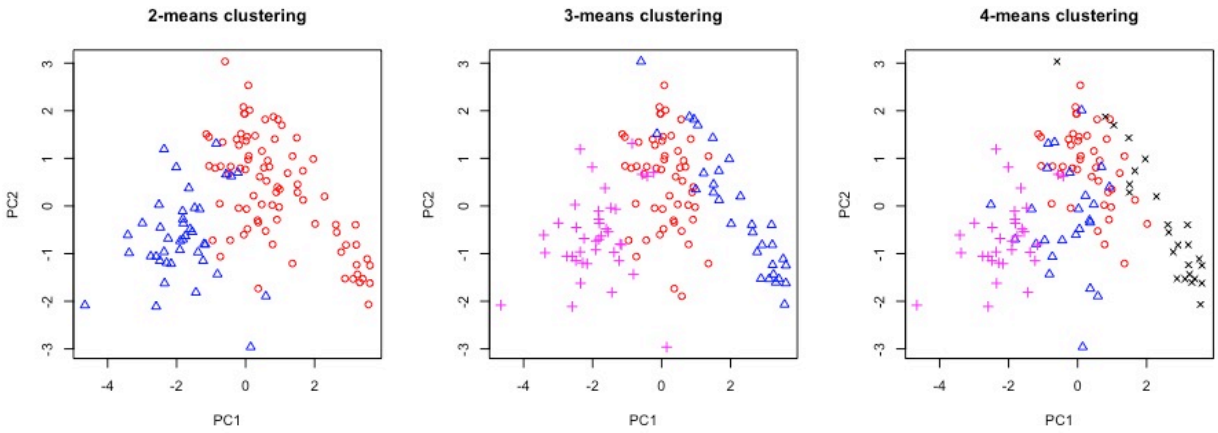
Both Single linkage suggests 4 clusters. Complete linkage suggests 3 or 4 clusters and average linkage suggests 2 clusters. The scree plot suggests K equals 2. Therefore, I will try k=2, 3 and 4 in K-means clustering technique.

Figure 5: Scree plot



To illustrate the results, I conduct principal component analysis and choose the first two PC as coordinates. Results of K-means are shown below.

Figure 6: 2 and 3-means clustering results



Comparing the results of three kinds of clustering, I think 2 or 3 clusters are reasonable. Before I interpret the plots, the meaning of PCs should be clarified. The first PC is

$$PC1 = 0.498 * \text{Log GDP per capita} + 0.466 * \text{Social support} + 0.492 * \text{Life expectancy} + 0.397 * \text{Freedom} + 0.168 * \text{Generosity} - 0.328 * \text{Corruption}$$

The second PC is :

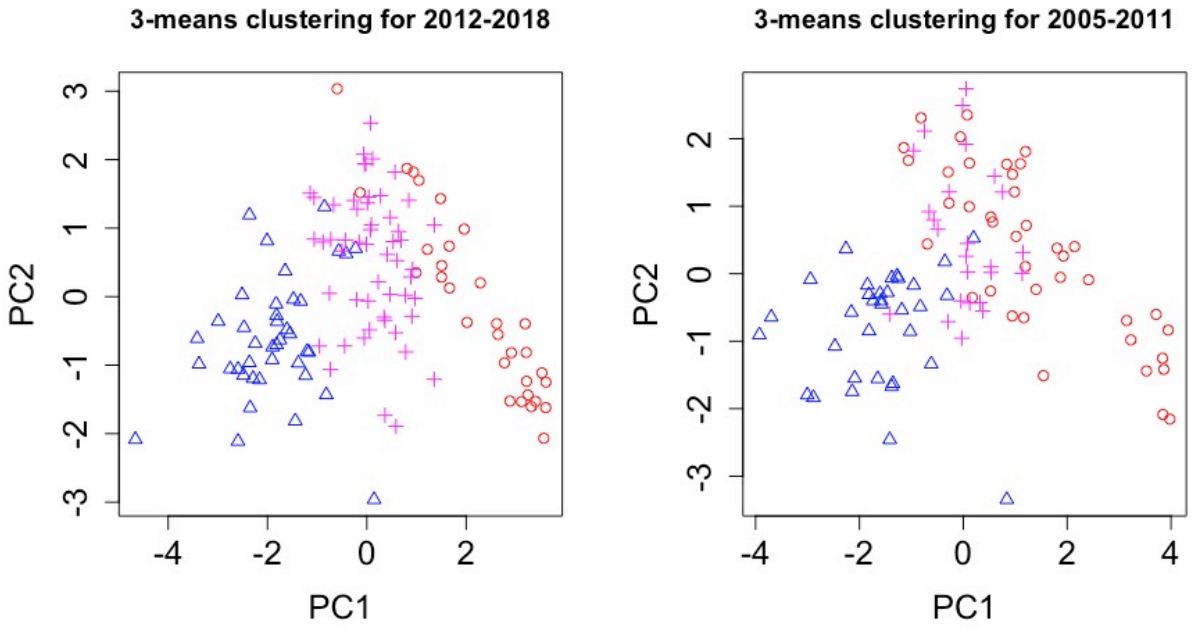
$$PC1 = 0.310 * \text{Log GDP per capita} + 0.254 * \text{Social support} + 0.259 * \text{Life expectancy} - 0.355 * \text{Freedom} - 0.665 * \text{Generosity} + 0.451 * \text{Corruption}$$

PC_1 stands for the overall well-being measures of a country. A country with higher GDP per capita, more social support and generosity among people, longer healthy life expectancy and more freedom to make life choices will gain higher PC_1. PC_2 stands for how much power the government have and how free people can be. A country may enjoy powerful and developed economy but people don't have that much freedom and aren't willing to donate. This kind of country will gain higher PC_2.

For 2-means case, countries clustered into the 1st group are those with higher PC_1 and the 2nd group with lower PC_1. They are not apparently separated in PC_2's direction. We can see that the first group symboled by blue triangle mainly locates at the upper right side and the second group symboled by red circle mainly locates at the lower left side. If we say that group 1 stands for happy countries and group2 stands for unhappy countries. Countries with higher PC_2 and medium PC_1 can be clustered into group1 and those with modestly high PC_1 score are clustered into group2. This implies that freedom to make life choices and people's generous quality are very important criterion for people's happiness.

For 3-means case, I conduct 3-means clustering analysis twice on dataset before 2012 and after 2012. And then I compare the clusters of them.

Figure 7: Comparing 3-means clustering results for data before and after 2012



Countries in the 1st group symboled by red circle are those with high PC_1 scores and widespread PC_2 scores. Those in the 2nd group are with medium PC_1 scores but high PC_2 scores. Those in the 3rd group are with both low PC_1 and PC_2 scores. If we say that group1,2 and 2 are happier, happy and unhappy countries. The results implies basically the same conclusion with 2-means case that country with more freedom for their people to make life choices and more generous atmosphere can make up for their disadvantage in economic development.

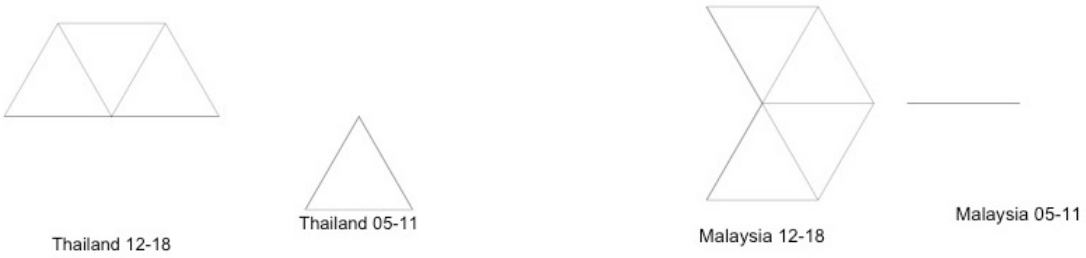
Comparing the results of data from 2005 to 2011 and data from 2012 to 2018, we can see that some countries' cluster are changed. Those are listed in Table 6. And I draw star plot of them and try to find some reason behind this change.

Table 6: Countries changed different cluster

country	Cluster 2005-2011	Cluster 2012-2018
Thailand	3	2
Malaysia	3	2
Nicaragua	1	2
Mauritius	1	2

The following star plot show the well-being measure values of Thailand and Malaysia before and after 2012 in Figure 8.

Figure 8: Comparing star plots of Thailand and Malaysia



Comparing the two star plot for Thailand, Thailand has grown in economic and medical development. In the meantime, people think there are less corruption and they are not as generous as past. These leads that people enjoy more happiness than before. Similar for Malaysia, it generally developed in economy and medical. People enjoy more freedom and social support. There are lower perception of corruption in their government. These lead to a happier Malaysia.

Conclusion:

In this project, I studied the common characteristics and influential factors of happiness of a country. Through data visualization, I see that Log GDP per capita, social support and healthy life expectancy at birth are highly correlated. Log GDP per capita can a foundation of the other two variables. Stronger economic power implies better medical development level in both physical and mental aspects. What's more, the TOP 5 happiest countries share the same characteristic which is low perception of corruption. It a natural idea that open and transparent government can provide people with more sense of security and power. However, some countries with less generosity are still famous of happiness due to small gap of wealth probably. Then, canonical correlation analysis is conducted to discover some explicit observable variables to explain those subjective well-being measures. Finally, clustering analysis is conducted to investigate cluster pattern among countries. I conclude that freedom and generosity can make up for less developed economy and medical system. Countries are separated into 2 and 3 groups which capture different features. Additionally, k-means clustering analysis is conducted on both dataset before and after 2012. 2012 is a turning point of the emerging development of information technology. Therefore, I compare the clustering results and find Thailand and Malaysia become happier because of their development in economy and other aspects.