# An Exploratory Data Analysis of Globel Peace Index (GPI): 2015-2019

Genghua Chen

December 9, 2019

## Contents

## *Introduction*

*Parasite* is the first Korean film to win the Palme d'Or, the highest prize awarded at the Cannes Film Festival. The movie received widespread critical acclaim and was chosen by Time magazine as one of the top ten films of 2019. The whole story begins with an impoverished family living in a semi-basement. Because of an accidental opportunity, four members of this family all find a well-paid job at a wealthy family through deceiving. However, this seemingly good storyline turned sharply at the rest of half of the movie, and eventually, it becomes a thriller and ended in tragedy. Even though many movie critics said economic inequality and patriarchy are obvious themes of the movie, yet to dig deeper, the whole movie is centered around one core word – happiness.

Before getting involved in this research, we will first introduce ten necessary column name's explanations to help readers better understand what we are going to do next. First, there are four major coulumn names that are easy to understand: Country, Year, Happiness_Rank, and Happiness_Score. Second, six factors in the dataset will directly affect the happiness index, which are Economy, which is GDP per Capita, Family, Health, which is life expectancy, Freedom, Trust, which is Government Corruption, and Generosity. You are going to hear these ten column names repeatedly in the research, guaranteed, so make sure you are familiar with them before continuing.
In the following report, we have raised three main questions, namely: Is there any country that has notable ranks or scores change from 2015 to 2017? How does the six key factors affected people's happiness score for 2015 to 2019? and Is the 'sub-saharan Africa' was the poorest region in 2015? It is gratifying that at the end of the study, every question has a certain answer. Meanwhile, throughout the exploration process, we also confronted numerous obstacles like every beginner, but fortunately, we eventually solved all these problems. Some can be seen intuitively in the code, but more questions are only those who have experienced the entire process can experience it personally. In order to ensure a detailed and accurate explanation of the process of exploring the questions and answers, we decided to intersperse the explanations with the preceding or following questions of each plot.

## *Data Cleaning*

Before we start our exploration, we need to clean the data first. We used five datasets for this report, so we want to combine all five data frames into one dataframe to make our exploration easier. Fist, we have to find out whether we can combine them or not. We found that these five data frmes have many similar variables, but those variables have different names. Then, we need to unify the variable names for all five data frames, so we can start to combine them. We used snake case to separate the longer variable names to make our code look nicely.

We encountered a problem on how to combine those five data frames. Initially, we want to use some joins to combine them into one data frame, however it is not what we are expected. After some discussion, we figured out that we merge them together, so we checked the documentation for merge(). Finally, we obtained a new data fram called df after some modification.

Next, we want to make sure that our new data fram is clean. One of our team member realized that some of the country names were not unified. For example, Hongkong is called "Hong Kong S.A.R., China" in one of the data frame, Taiwan is called "Taiwan Province of China" in one of the data fram, and Trinidad and Tobago is called "Trinidad & Tobago" in two of the data fram. Then, we unified those names with a same country name. We also checked if each country have five rows, because we suppose to have five years of data for each country. Initially, We want to delete the data for countries that have less than five rows. However, we think it is not necessary, because it will not affect the result of our exploration.

When we were doing our exploration, we can not do our plot with variable "Trust". One of our team member spend an hour to solve this problem. Finally, we found the data type for Trust is character, because there is a string in the observation called "N/A". We googled this problem, then we successfully convert the data type of the variable "Trust" from character into double.

## *Questions and Findings*

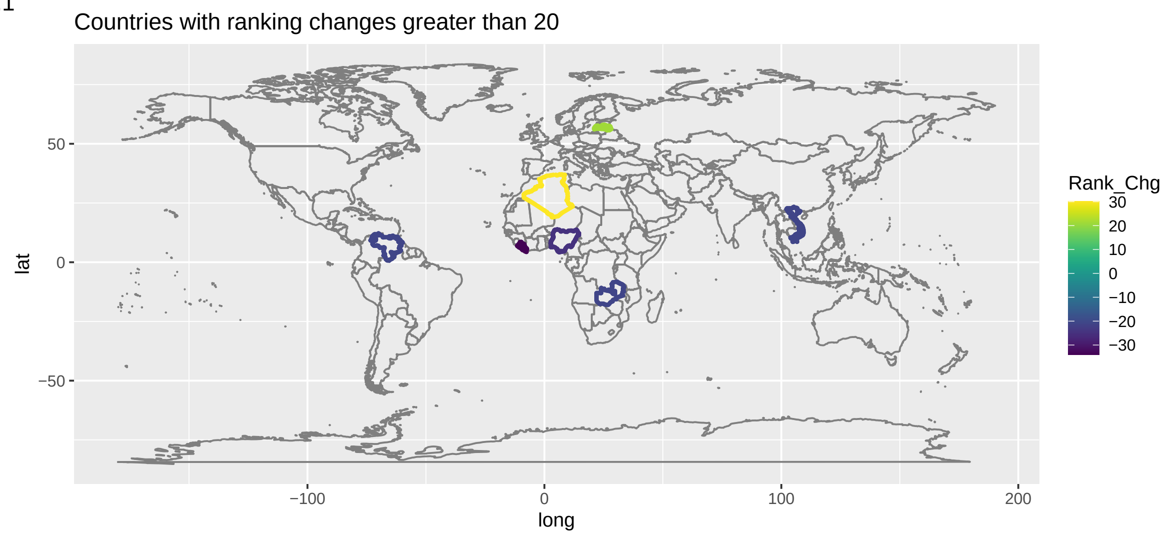
### *Question 1*

### *Is there any country that has notable ranks or scores change from 2015 to 2017?*

My question was not very clear at the beginning of the research, but I remember that there was a mention of happiness_score or happiness_rank change in "Inspiration", so from the very beginning, my thought line went in this direction. The initial idea popped up is to create a world map and use geom_point to locate a big colorful point at the country which has the greatest happiness rank change or score change. It sounds familiar, doesn't it? This is a question we have done at least three times this semester. It is about New York airport and the special day of June 13, 2013.

First, I limited the time between 2015 and 2016, and then "spread" the happiness_Rank value into two columns which are 2015 and 2016. In this case, I could find the difference values between 2015 and 2016, and then I mutated a new column called "Rank_Chg", which is column 2015 minus column 2016, and the new column is exactly the happiness change of two years. I used "arrange" to organize the column, but later on I found it is unnecessary. In the end, in order not to generate a global map as a clown's face, I would like to filter out the country that has less than 20 happiness rank change. In the short code in the middle, I used left_join to join the world map dataset into my first data, which called "diff". In the last piece of code, I utilized the question from June 13, and after many attempts, I made Figure 1.1, the world map that looks good.

Figure 1.1



**Features:**

1. The darker the color, the higher the boost of happiness rank, the visa versa.

2. The countries with large ranking changes are concentrated in Africa, South America and South Asia.

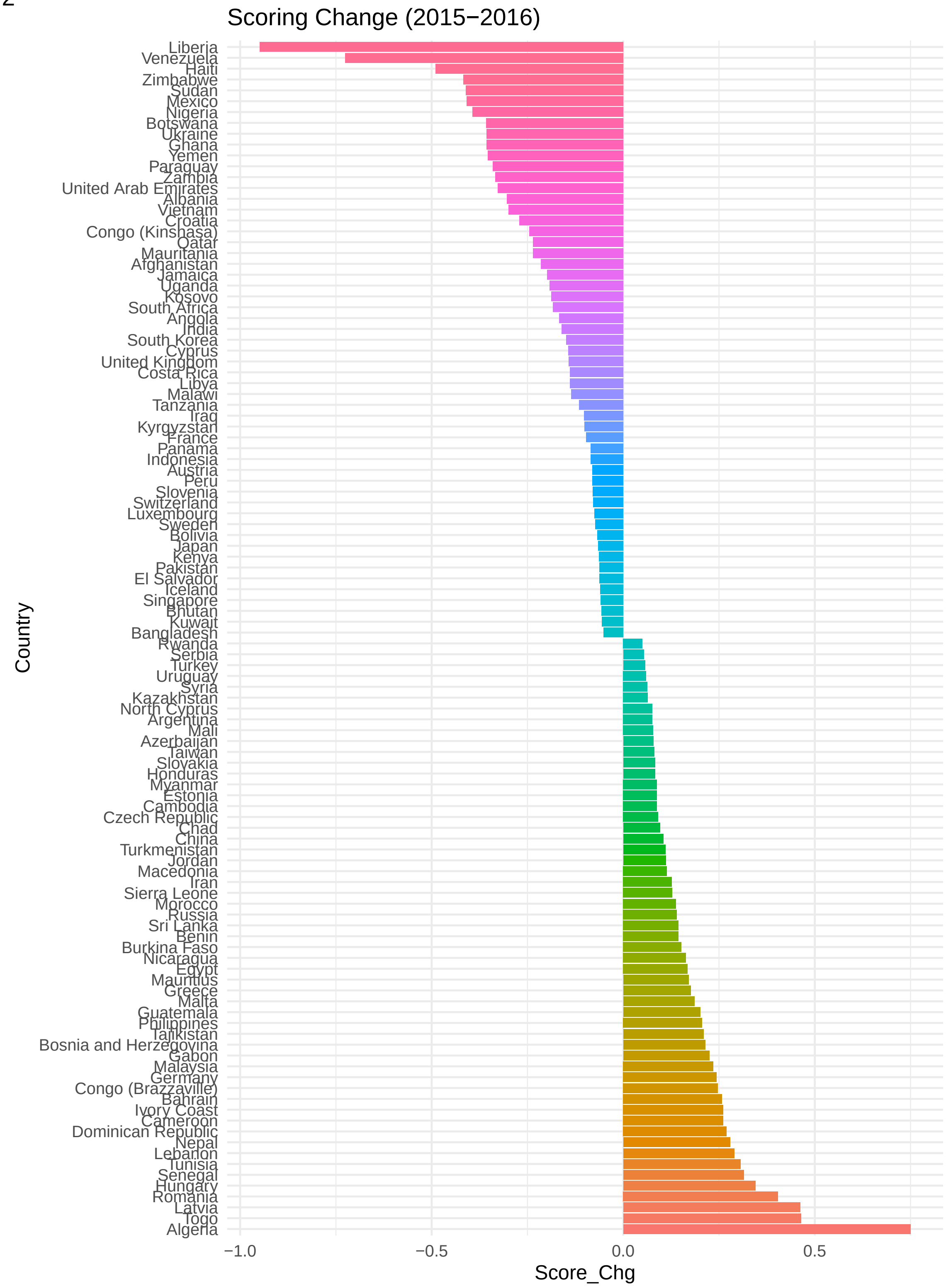
3. The plot is not intuitive, for example, we cannot see the name of country.

## What is next?

Next, While changing a plot to express the data visually, I also want to see the change of happiness score between 2015 and 2016. If the result, which is the scoring change, is similar as the ranking change. Then I might be able to tell that I only need to study the happiness score or happiness level, one of the two.

In the first piece of the code, I filtered out other years, and keep 2015 and 2016. Then since I wanted to know the score change between 2015 and 2016, so I "spread" year to column `2015` and `2016`. Then I "mutate" a new column to contain the value of score change between 2015 and 2016. Also, as I did before, in order to make the plot looks clean and clear, I filter out the country that has less than 0.05 happiness score change. The "gather" part can be omitted, but I did it for the sake of simplicity. In the second piece of code, in order to make plot looks orgnized and clear, I convert column "Country" as factor. In the third piece of code, I created Figure 1.2, a bar plot but with y axis, which is the change of happiness score. I also change the background color of Figure 1.2 to white, and because the x axis contains too many names of county, I omited the lengend and used coord—flip.

Figure 1.2



Indeed, by looking at Figure 1.2, I cannot draw a firm conclusion. But I want to make a bold assumption: I suspect that the source of inspiration given by the professor is related to Venezuela. There are two reasons behind it.

First, It seems like the change of happiness index of *Liberia* is significantly lower than the second-ranked

*Venezuela*. However, if we compare *Venezuela*'s decrease and *Algeria*, the top one happiness score increased country, it seems like the decrease of *Venezuela* is also huge.

Second, there is a sentence from a Venezuelans that I still remember it today. In early 2018, I and my Venezuelan friend were walking in the Chelsea Market in New York at the time. He pointed down a bag of small pistachios lightly and said, “This price is basically a month's salary for the people of our country.” From the calculation of the time point, I think the situation he mentioned is correct, and I suspect the Venezuelan happiness score might be lower next year.

Based on the two reasons, I prefer to continue my analysis conservatively to see the scoring change between 2016-2017.

### Why do I dicided to only focusing on scoring change?

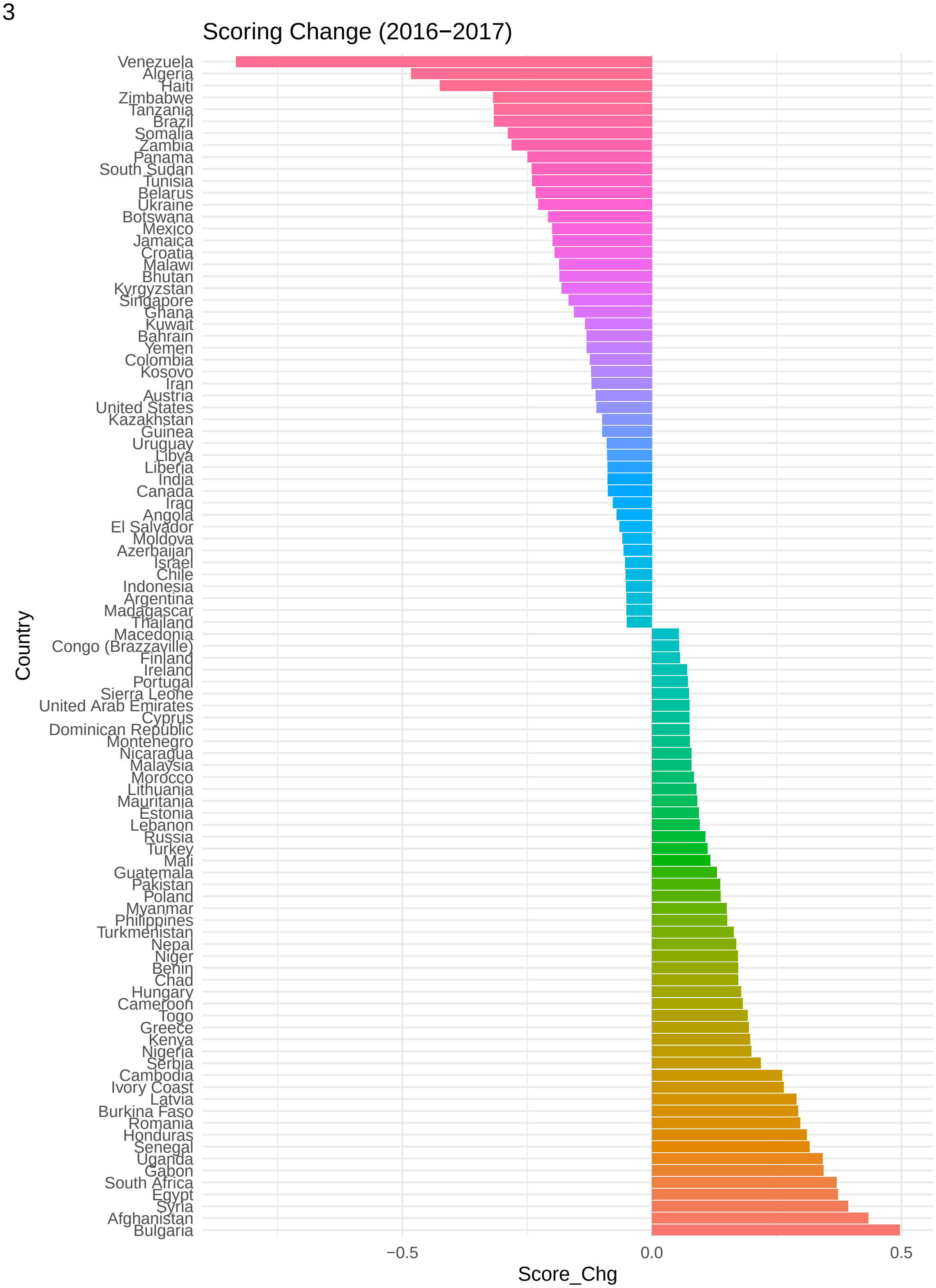
1. The plot is huge and fansy. If providing double number of plots, it will definitely reduce reader interest.

2. Analyze both scoring change and ranking change may lead to similar conclusions.

3. Ranking changes are also affected by other countries' changes, while socring changes are not.

In the following case, I created Figure 1.3. The entire calculation and generation procedure is the same as Figure 1.2, and the only difference is that it refers to 2016-2017. I expect to see more information about *Venezuela*.

Figure 1.3



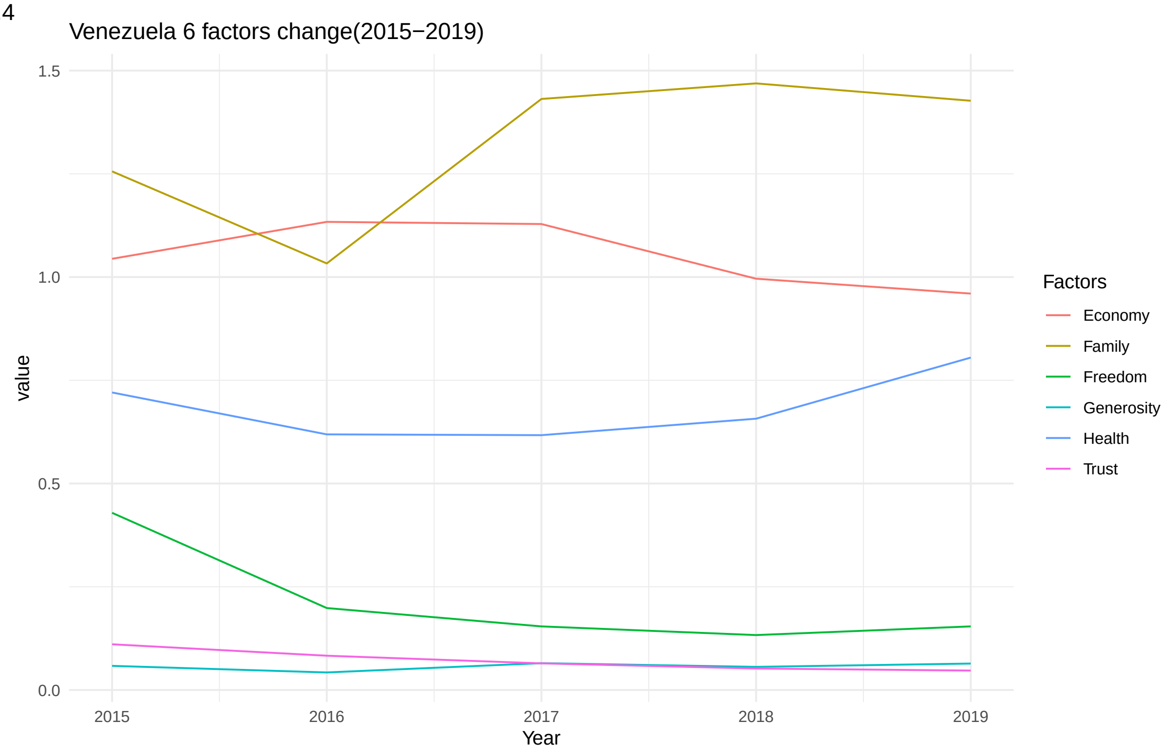
Yes! Compared with Figure 1.2 and other countries in Figure 1.3, we can verify that my conjecture is correct. Between 2015-2016, *Venezuela* lost almost 0.75 points of its happiness score, and between 2016-2017, the decline of *Venezuela*'s happiness index is far ahead of all other countries: -0.834!

Another interesting thing we could find is in the last year, *Algeria* was the country with the highest improvement in the happiness score, but in this report, it is the second-ranked country with the largest decline in the happiness score.

### What to do next?

From the data point of view, what happened in *Venezuela*? I want to glance at the six factors that determine the change in happiness scores in Venezuela. In the following Figure 1.4, I met the problem we have mentioned at the very beginning: Trust. I talked with my teammate, and the problem was solved by, actually, Google. Then, I filter all other countries, but keep "Venezuela". Next gathered six factors to one column called "Factors" and all their values to "value".

Figure 1.4



By observing Figure 1.4, my biggest question is: The six factors in Venezuela, except the "economy", have all declined between 2015 to 2016, yet why did the great decline of happiness scores be reflected in the report in 2017 instead of 2016? From this Wikipedia website: https://en.wikipedia.org/wiki/World_Happiness_Report, I found that the annual score report depends on the comprehensive situation of the previous three years. This can clearly explain why the decline in the six factors was mainly concentrated in 2016, while the overall decline in happiness score was reflected in the 2017 report. By the way, from 2015 to 2019, the website says: "The biggest loser is Venezuela, down 2.2 points."

### *Question 2*

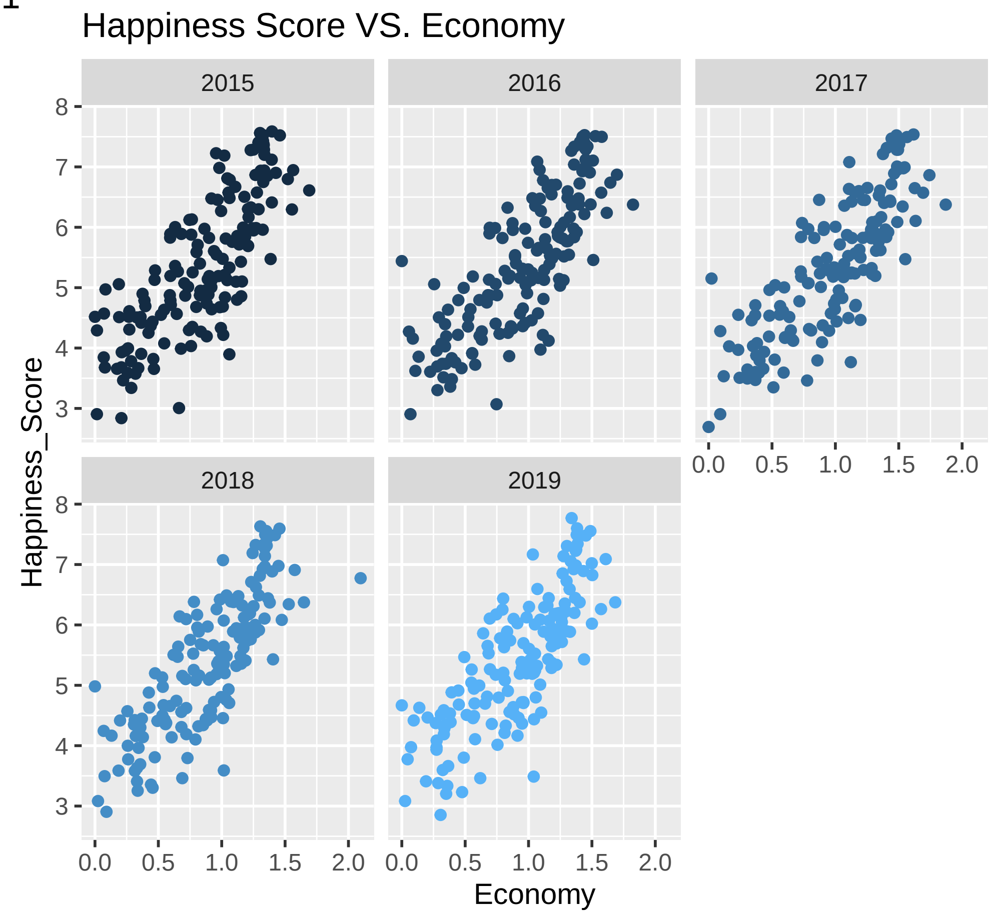
### *How does the six key factors affected people's happiness score for 2015 to 2019?*

When I saw this dataset, this first question came to my mind is what are the variables that will affect people's happiness. As we all know, happiness is an important part of our life. What makes people happy? I found that economic production, social support, life expectancy, freedom, absence of corruption, and generosity will contribute to people's happiness score when I was reading the data content. Then, I decided to explore how does those factors affected people's happiness score. The purpose of this question is to understand how the factors affected people's happiness score. What is their correlation?

### Economy

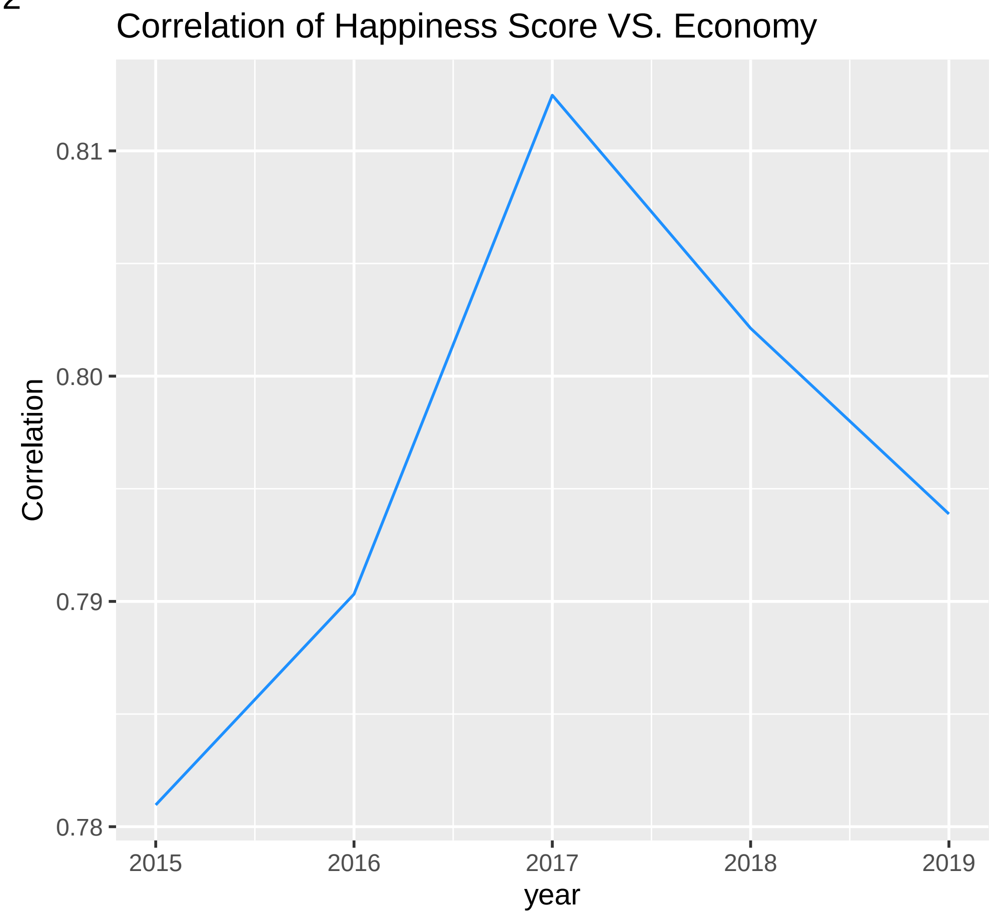
Economy is the first key factor that I'm going to introduce. Obviously, money will make a lot of people happy, because we need money to survive on this planet. Figure 2.1 is a scatter plot for the Happiness Score vs. the GDP per capital for every country in five years. From Figure 2.1, we can see there is a positive trend in all five graphs, as the the economy of those countries increased the happiness score of those countries also increased. In fact, this is true for 2015, to 2019. Since the trend for all five years is very similar, I want to calculate their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.78, 0.79, 0.81, 0.80, 0.79 respectively. Since the happiness score is highly correlated with economy, we can conclude that economy is definitely a factor that will affect the happiness score.

Figure 2.1



Then, I want to check the correlation changed over time for happiness score vs economy. Figure 2.2 is a plot for the correlation change of happiness score vs economy from 2015 to 2019. From figure 2.2, we can see the correlation increased about three percent from 2015 to 2017, then decreased about two percent from 2017 to 2019. Since the correlation change is not that significant, we didn't want to explore it further.

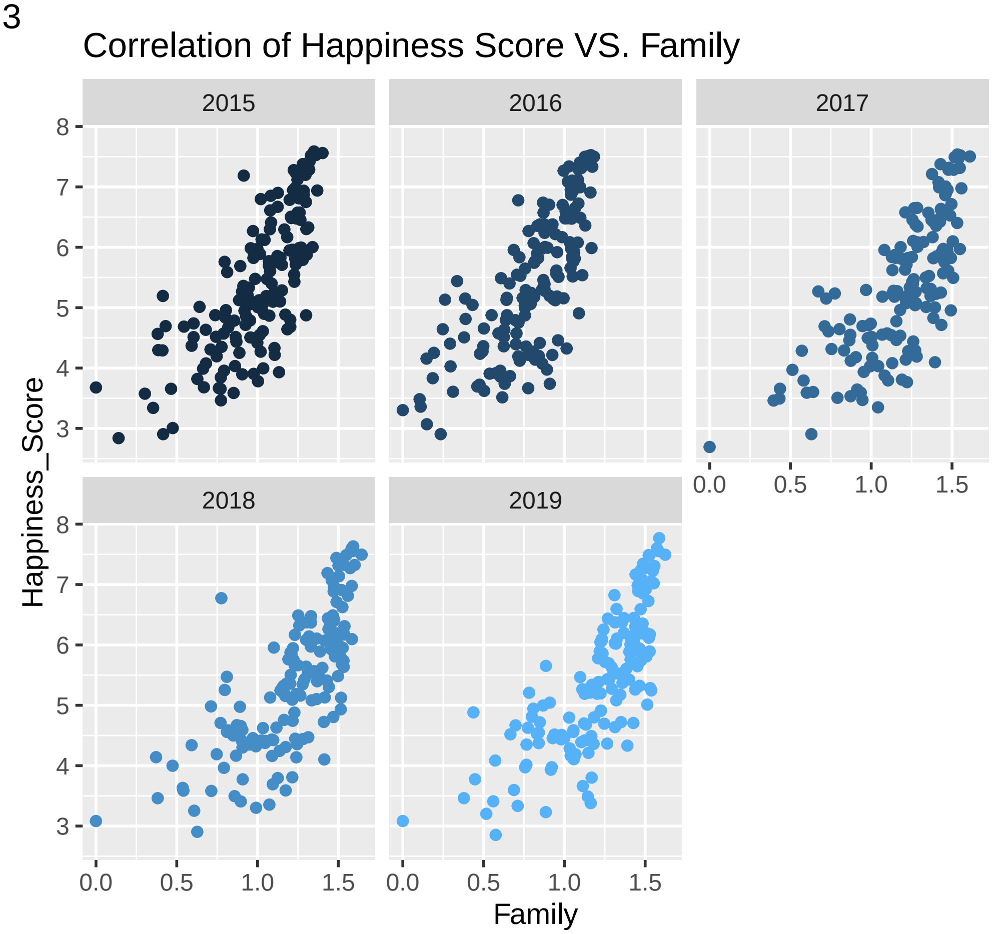
Figure 2.2



**Family**

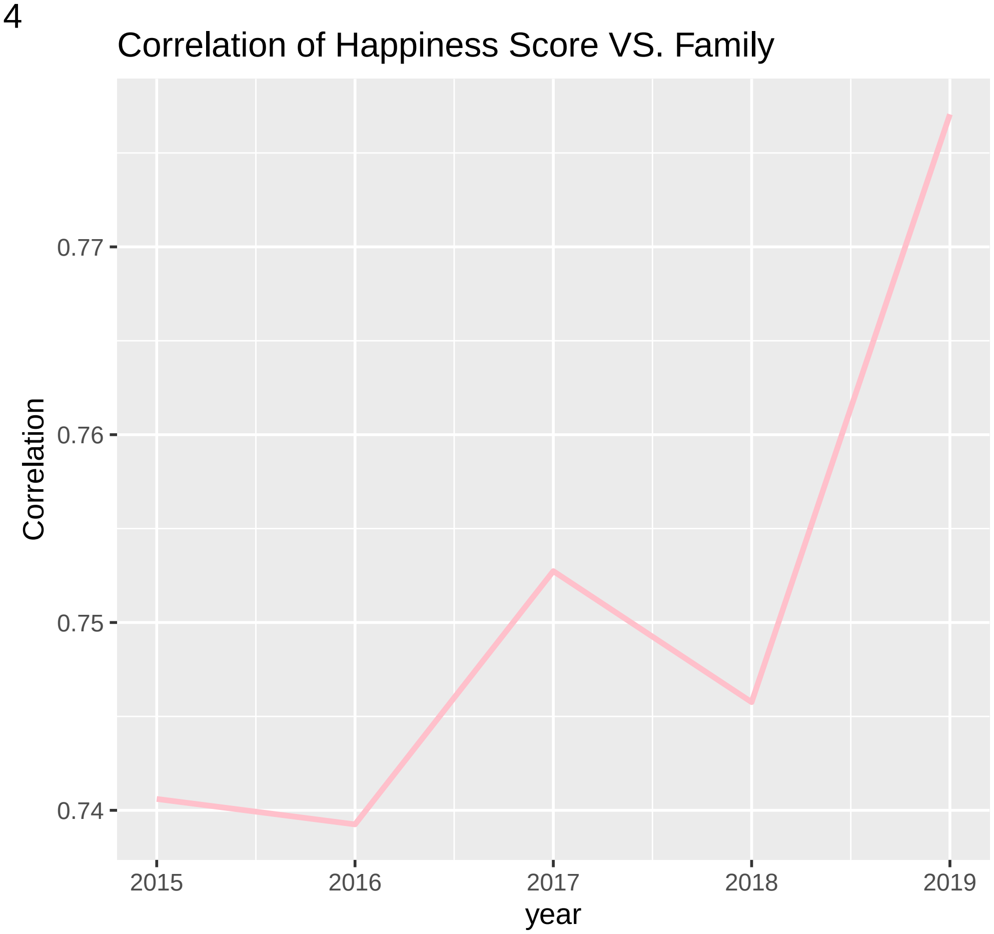
Family is the second factor that I'm going to introduce. We spend a lot of time with our family, of course we had a lot of fun with our family.
We will be happy when we were celebrate holidays with our family members. We will also be happy when we received gifts from our kids or parents.
Figure 2.3 is a scatter plot for the Happiness Score vs. Family contribution for every country in five years. From Figure 2.3, we can see there is a positive trend in all five graphs, as the the family contribution of those countries increased the happiness score of those countries also increased. Actually, this is true for 2015 to 2019. Since the trend for all five years is very similar, I also calculated their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.74, 0.739, 0.75, 0.746, 0.777 respectively. Since the happiness score is highly correlated with the family contribution, then we can conclude that family contribution is absolutely a factor that will affect the happiness score.

Figure 2.3



Then, I will check the correlation changed over time for happiness score vs family. Figure 2.4 is a plot for the correlation change of happiness score vs family from 2015 to 2019. From figure 2.4, we can see the correlation changed very little form 2015 to 2018, then increased three percent from 2018 to 2019. Since the correlation change is not very significant, we stopped our exploration about family contribution.

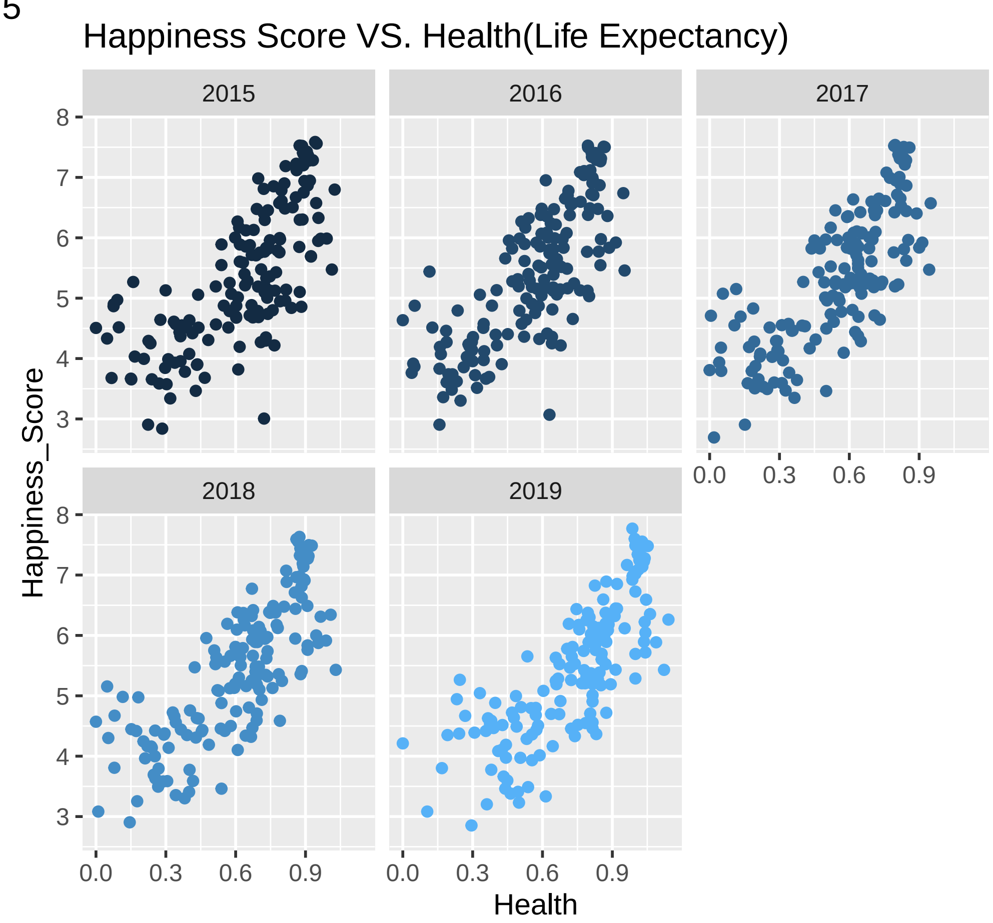
Figure 2.4



**Health**

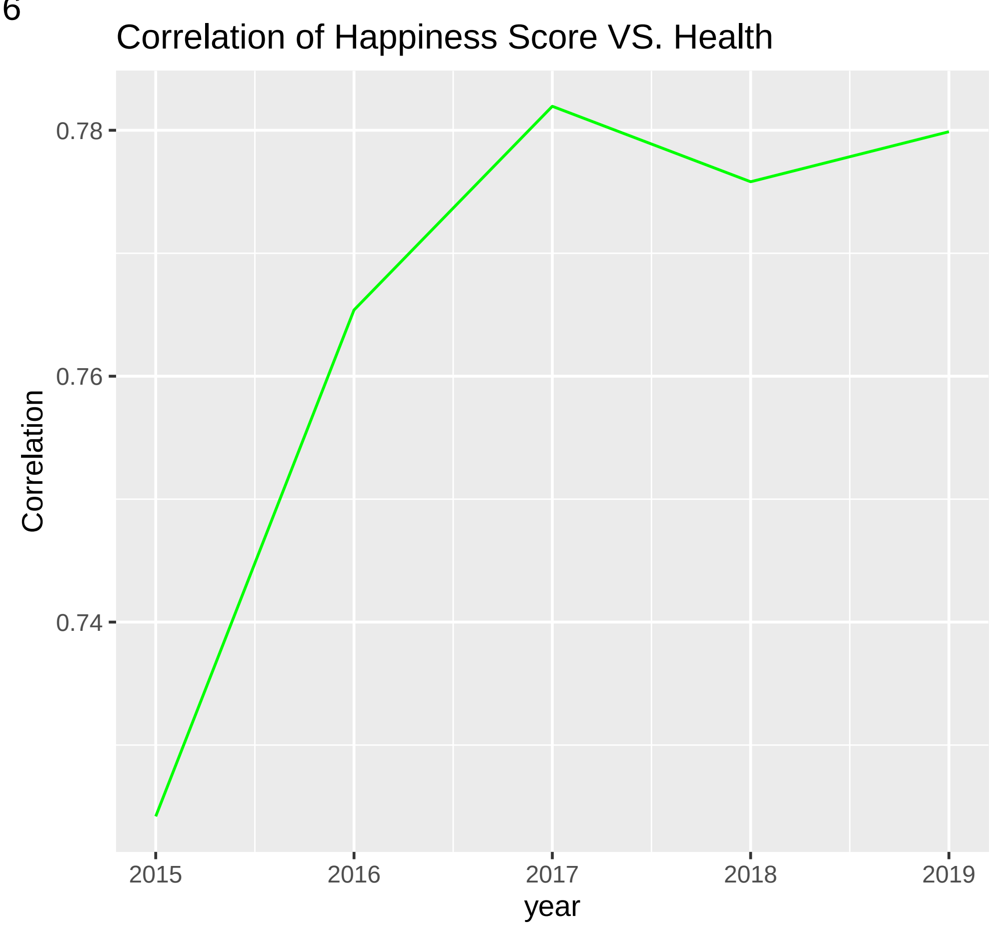
The third factor that I'm going to illustrate is health (or life expectancy). Sickness or bad health condition will definitely affect our happiness. Next, let's see how health is affected people's happiness score. Figure 2.5 is a scatter plot for the Happiness Score vs. Health(Life Expectancy) for every country in five years. From Figure 2.5, we can see there is a positive trend in all five graphs, as the life expectancy of those countries increased the happiness score of those countries also increased. Actually, this is true for 2015 to 2019. Since the trend for all five years is very similar, I want to calculate their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.72, 0.765, 0.78, 0.776, 0.780 respectively. Since the happiness score is highly correlated with life expectancy, we can conclude that life expectancy is absolutely a factor that will affect the happiness score.

Figure 2.5



Next, let's check the correlation changed over time for happiness score vs health. Figure 2.6 is a plot for the correlation change of happiness score vs health from 2015 to 2019. From figure 2.6, we can see the correlation increased six percent from 2015 to 2017, then it kept steady from 2017 to 2019. Since the correlation changed significantly from 2015 to 2017, it might be the health or life expectancy is getting more important on people's happiness.

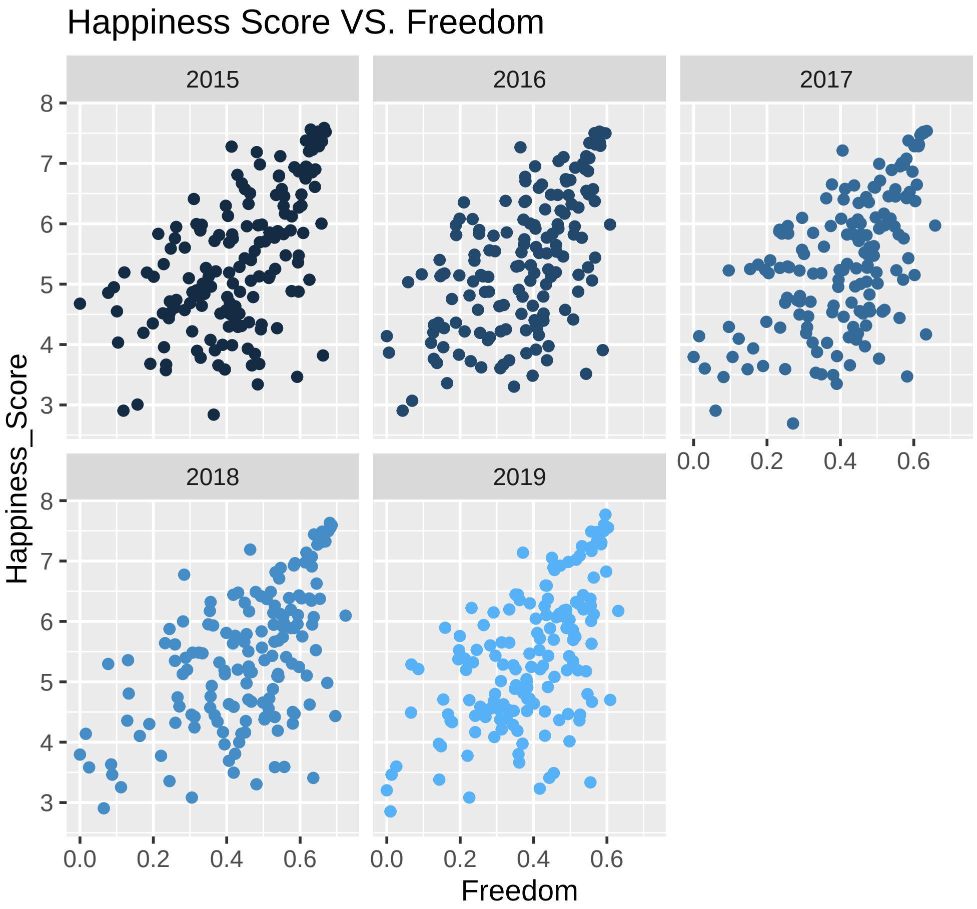
Figure 2.6



**Freedom**

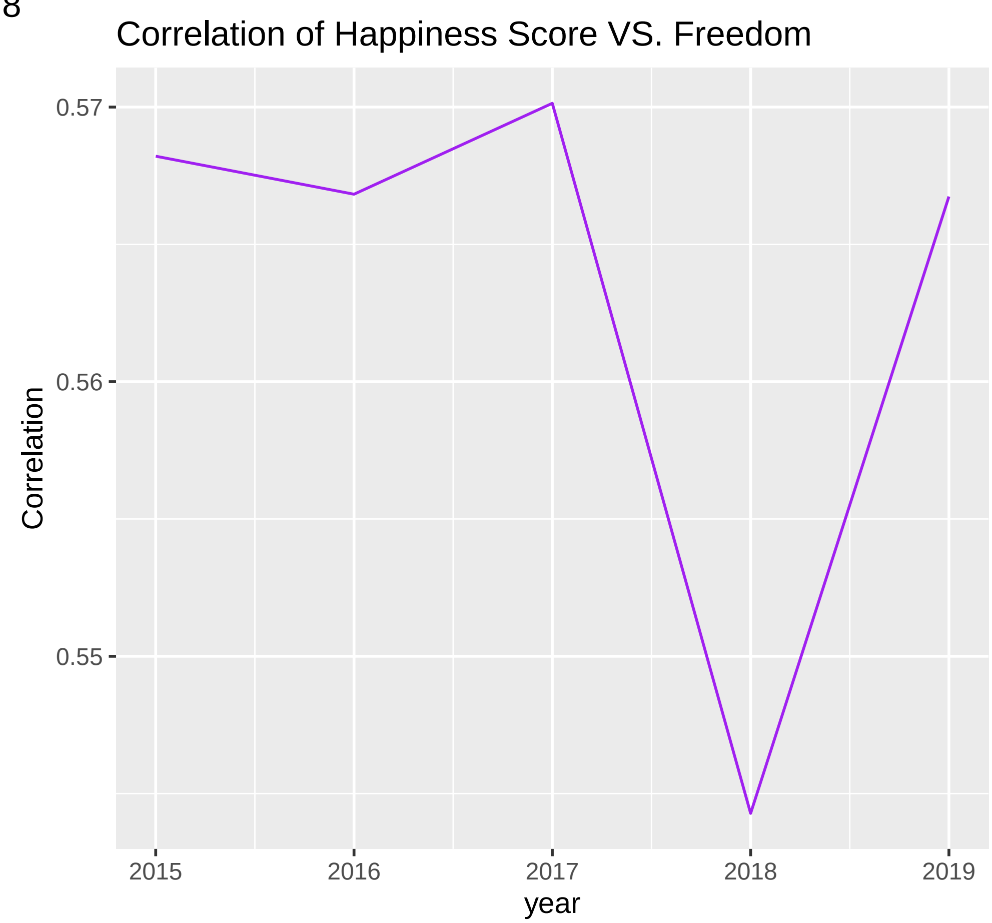
Freedom is the fourth key factor that I'm going to explore. Figure 2.7 is a scatter plot for the Happiness Score vs. Freedom for every country in five years. From Figure 2.7, we can see there is a positive trend in all five graphs, as the freedom of those countries increased the happiness score of those countries also increased. Actually, this is true for 2015 to 2019. Since the trend for all five years is very similar, I also calculated their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.568, 0.567, 0.570, 0.544, 0.567 respectively. Since the correlation between happiness score and freedom is not that bad, we can conclude that Freedom is a factor that will affect the happiness score.

Figure 2.7



Next, let's see the correlation changed over time for happiness score vs freedom. Figure 2.8 is a plot for the correlation change of happiness score vs freedom from 2015 to 2019. From figure 2.8, we can see the correlation kept steady from 2015 to 2017, then it decreased 2.6% from 2017 to 2018 and bounced back from 2018 to 2019. Since the correlation change is not that significant, we didn't explore it further.

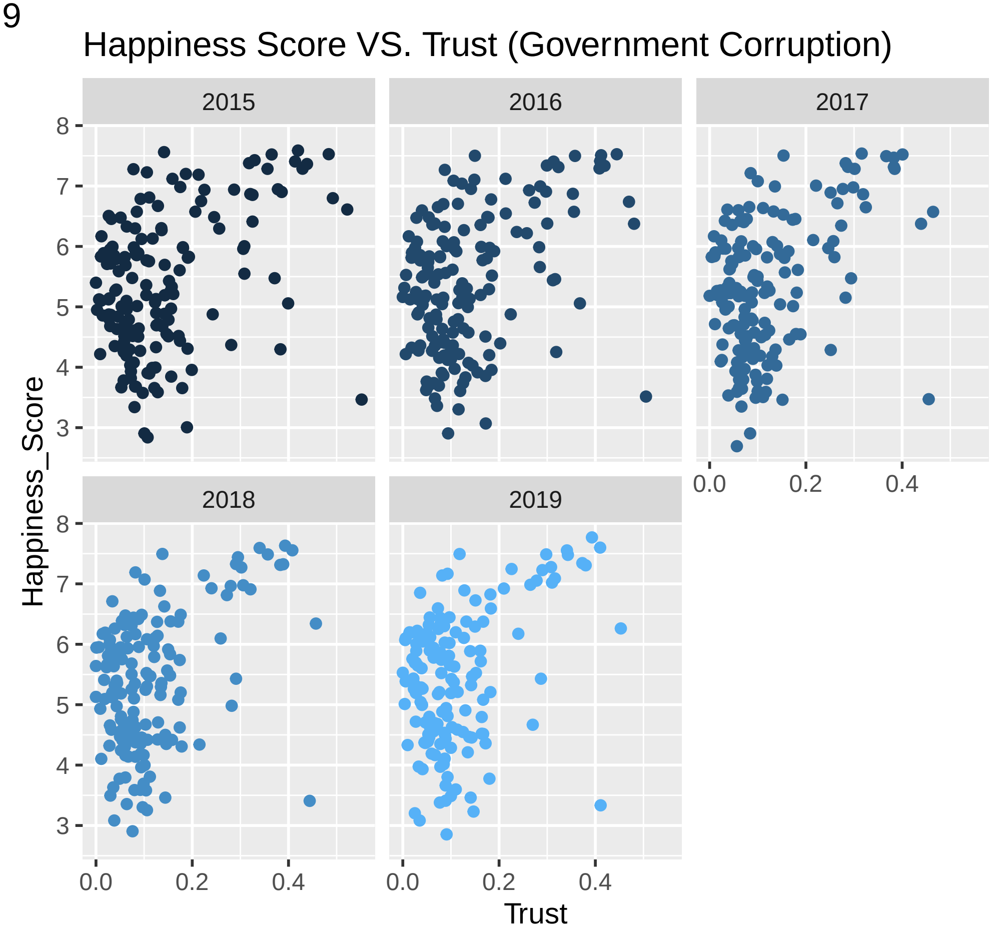
Figure 2.8



**Trust**

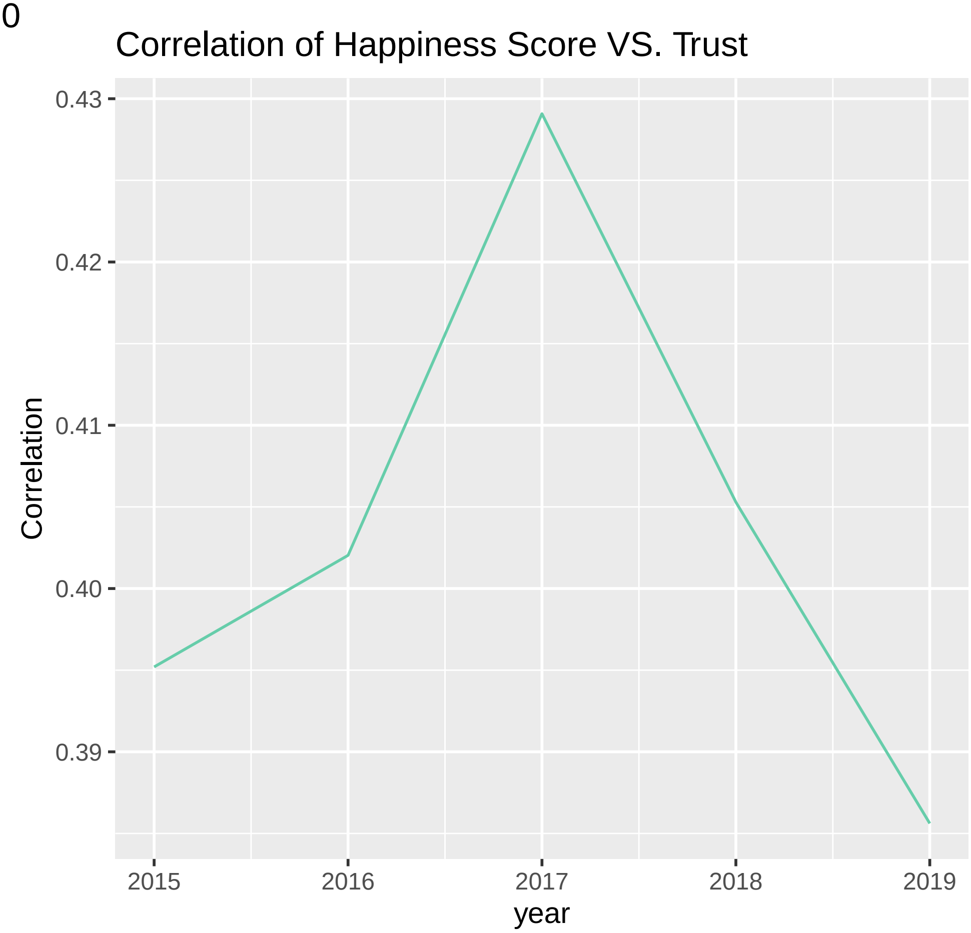
The fifth factor that we are going to observe is the trust of government or government corruption. Figure 2.9 is a scatter plot for the Happiness Score vs. Trust for every country in five years. From figure 2.9, we can barely see that there is a positive trend in all five graphs, as the Trust (Government Corruption) of those countries increased the happiness score of those countries also increased. Since the trend for all five years is very similar, I also calculated their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.395, 0.402, 0.429, 0.405, 0.386 respectively. Since the correlation between happiness score and Trust (Government Corruption) is not that bad, we can conclude that Trust (Government Corruption) is a factor that will affect the happiness score.

Figure 2.9



Then, we observed the correlation changed over time for happiness score vs trust. Figure 2.10 is a plot for the correlation change of happiness score vs trust from 2015 to 2019. From figure 2.10, we can see the correlation increased about 3% from 2015 to 2017, then straightly decreased about 4.3% from 2017 to 2019. Since the correlation decreased significantly from 2017 to 2019, it perhaps the government corruption is getting less important on people's happiness or the government corruption is getting better from 2017 to 2019.

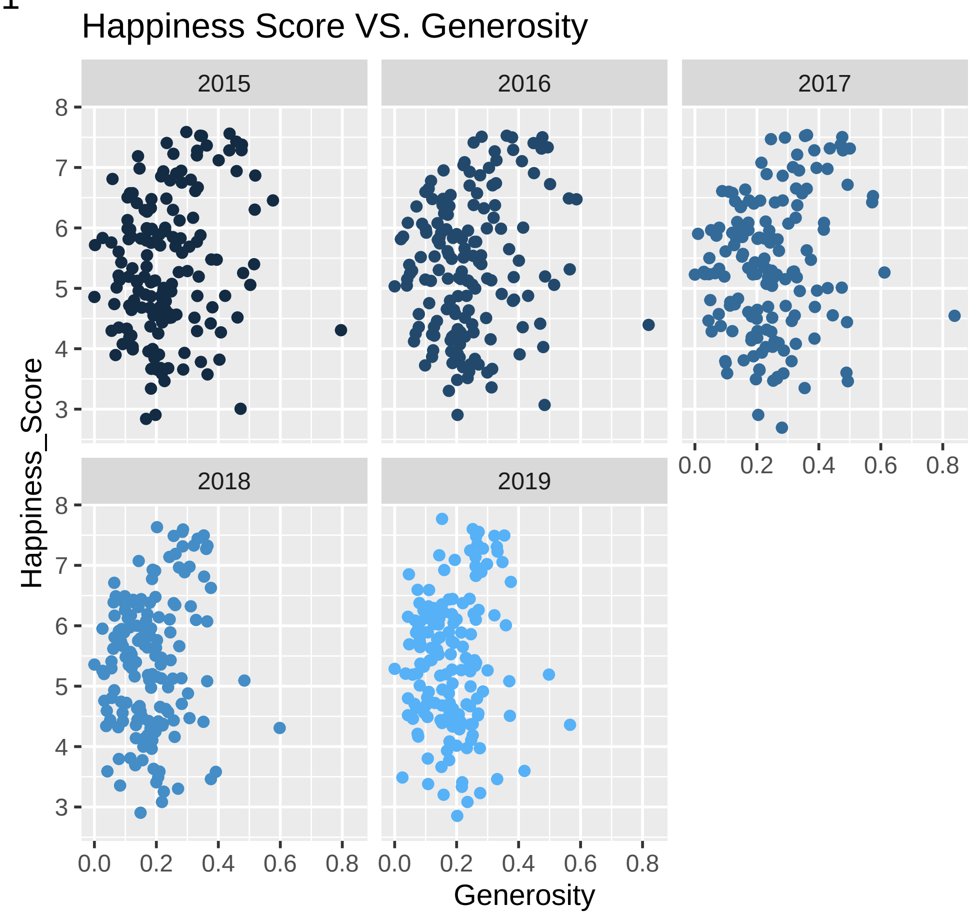
Figure 2.10



**Generosity**

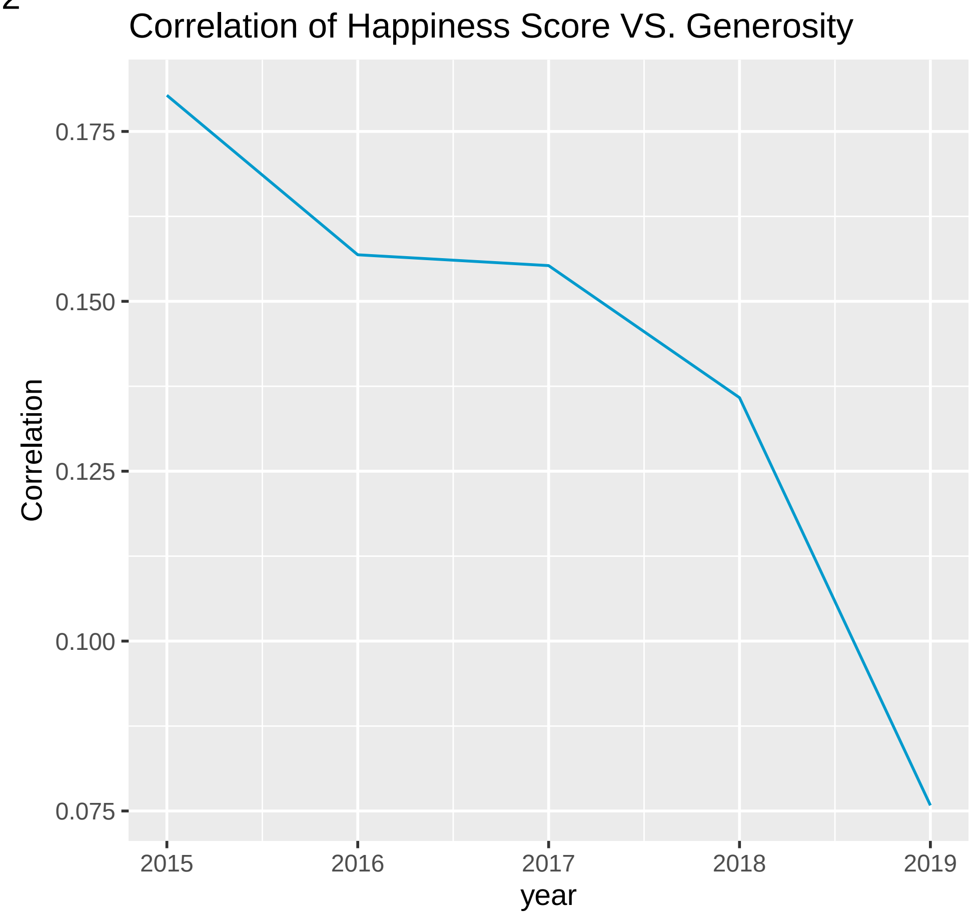
Generosity is the sixth factor that we are going to observe. Figure 2.11 is a scatter plot for the Happiness Score vs. Generosity for every country in five years. From figure 2.11, we can barelly see the trend in all five graphs. Since the trend for all five years is very similar, I want to see the relationship between happiness score and generosity by calculating their correlation. The correlation for 2015, 2016, 2017, 2018 and 2019 are approximately 0.180, 0.157, 0.155, 0.136, 0.076 respectively. However, the correlation between happiness score and generosity is very small, but we still can conclude that generosity is a factor that will slightly affect the happiness score.

Figure 2.11



Next, we want to observe the correlation changed over time for happiness score vs generosity. Figure 2.12 is a plot for the correlation change of happiness score vs generosity from 2015 to 2019. From figure 2.12, we can see the correlation decreased about 10.4 % from 2015 to 2019, the correlation decreased significantly. The reason might be that the generosity is getting less important on people's happiness.

Figure 2.12



As a result, we can conclude that Economy(GDP per Capita), Family, Health(Life Expectancy), Freedom, Generosity, and Trust Government Corruption were all contributing to people's happiness score. However, the Economy will contribute the most, since the correlation is the highest among these six factors. Family and health(life expectancy) were also contributed a lot to the happiness score, but generosity only slightly affected the happiness score, because the correlation for generosity is very small.

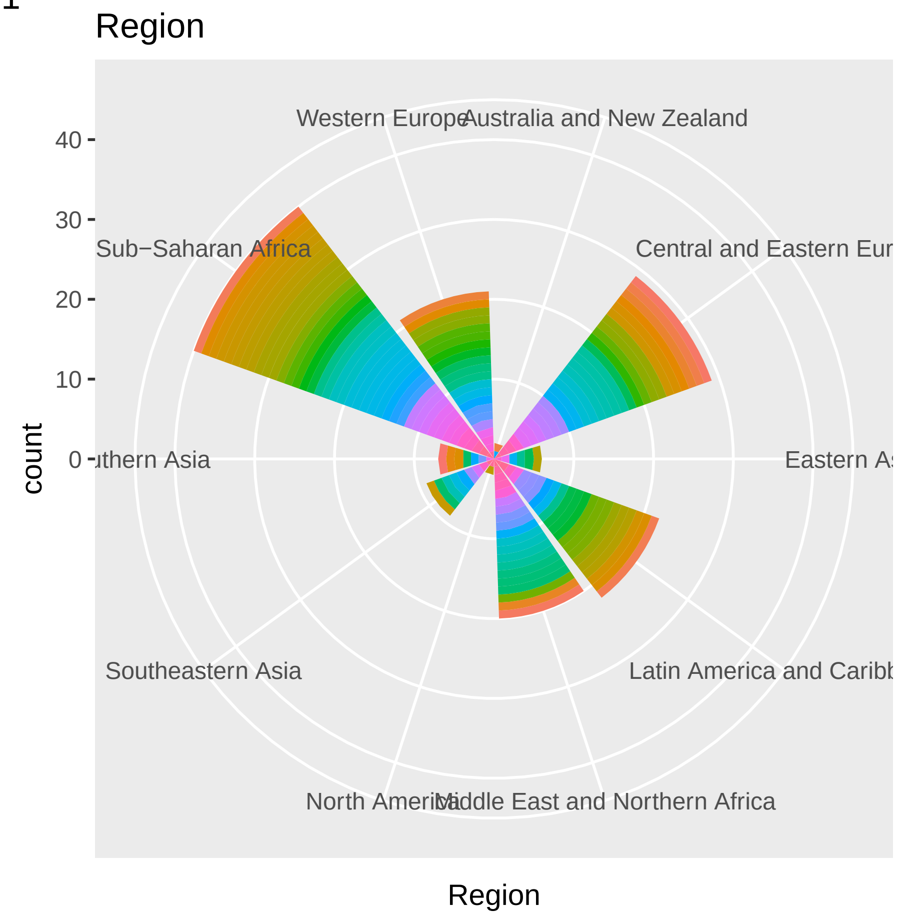
***Question 3***

***Is the 'sub-saharan Africa' was the poorest region in 2015?***

Africa located in the western eastern hemisphere, covers an area of 11.7 million square miles, accounting for 20 percent of the world land area. It is the second largest continent and the second largest population (about 1.2 billion) in the world. Africa hosts a large diversity of ethnicities, cultures and languages. In the late 19th century, European countries colonised almost all of Africa; most present states in Africa emerged from a process of decolonisation in the 20th century. African nations cooperate through the establishment of the African Union, which is headquartered in Addis Ababa.

In this dataset, a total of 10 regions and 158 countries were surveyed, 'Sub-Saharan Africa' accounting for the largest proportion. As we all know, Africa's economic development is lagging behind, the reason behind that is there has no no infrastructure, no education for a long time and there is no such good government policy and the government is corrupt.

Figure 3.1



Now let's look at the question: "Is the 'sub-saharan Africa' was the poorest region in 2015?" The Figure 3.2 that is happiness score and economy, color are present each regions. The purple color represent Africa. In the plot, most of the purple is concentrated in the lower left corner. That means the economy is fall behind and that region is unhappy. You can also see that the pink is from Western Europe, so Western Europe is the happiest region. Actually, I have tried to use bar chart to answer this question, but it is not very clear. So I decided to do it with geom_point. It's more intuitive to see if Africa is happiness. In addition, I use bar chat and found the happiest country. It's Mauritius.

Figure 3.2

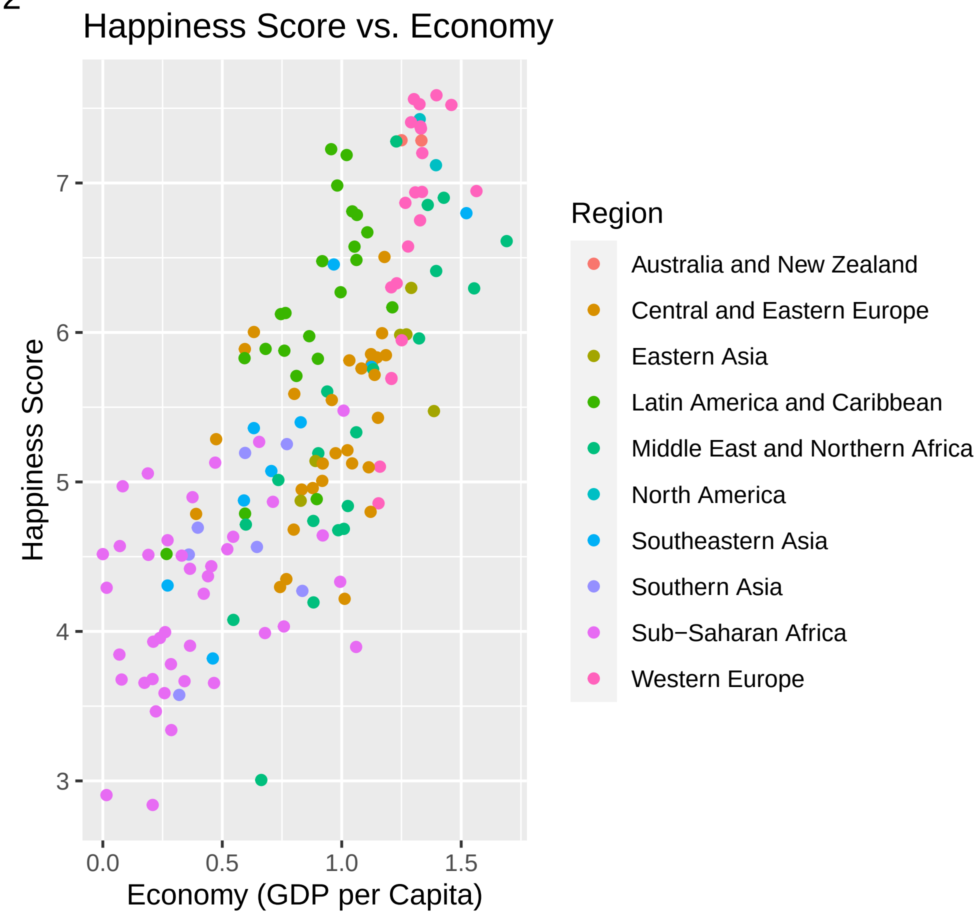
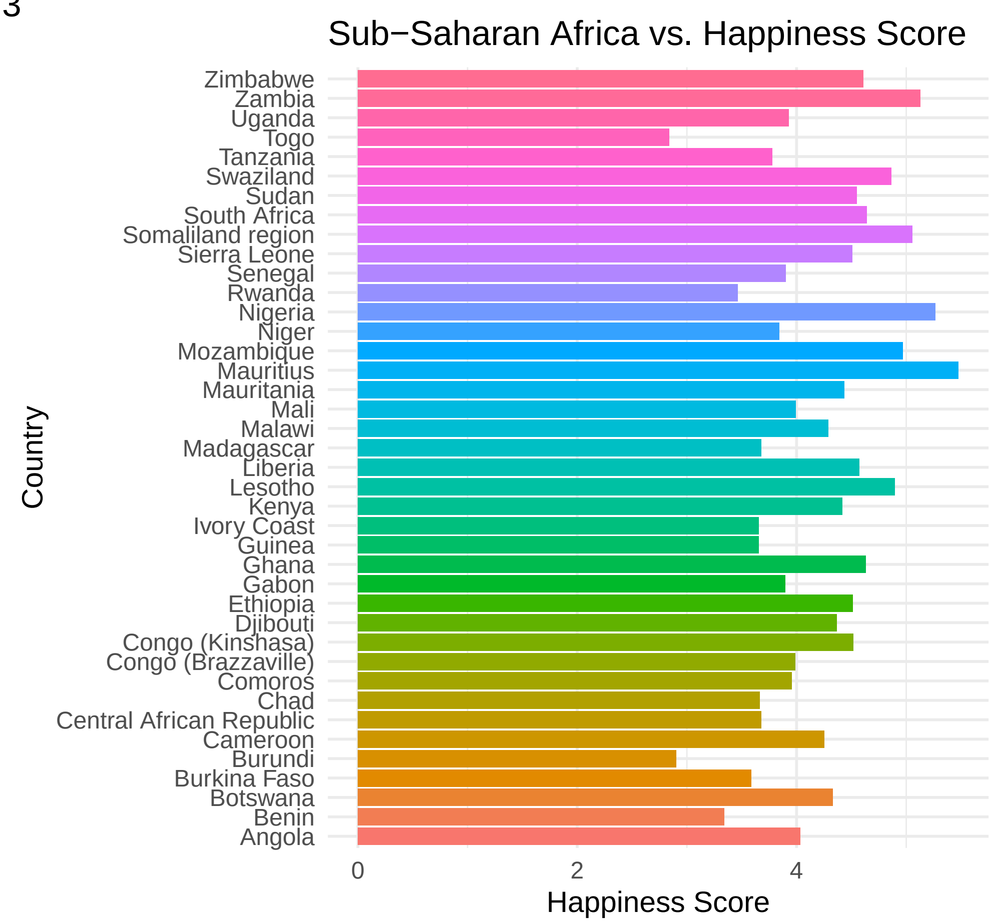

Happiness Score vs. Economy
Happiness Score
7
6
5
4
3
0.0
0.5
1.0
1.5
Economy (GDP per Capita)
Region
Australia and New Zealand
Central and Eastern Europe
Eastern Asia
Latin America and Caribbean
Middle East and Northern Africa
North America
Southeastern Asia
Southern Asia
Sub−Saharan Africa
Western Europe

Figure 3.3

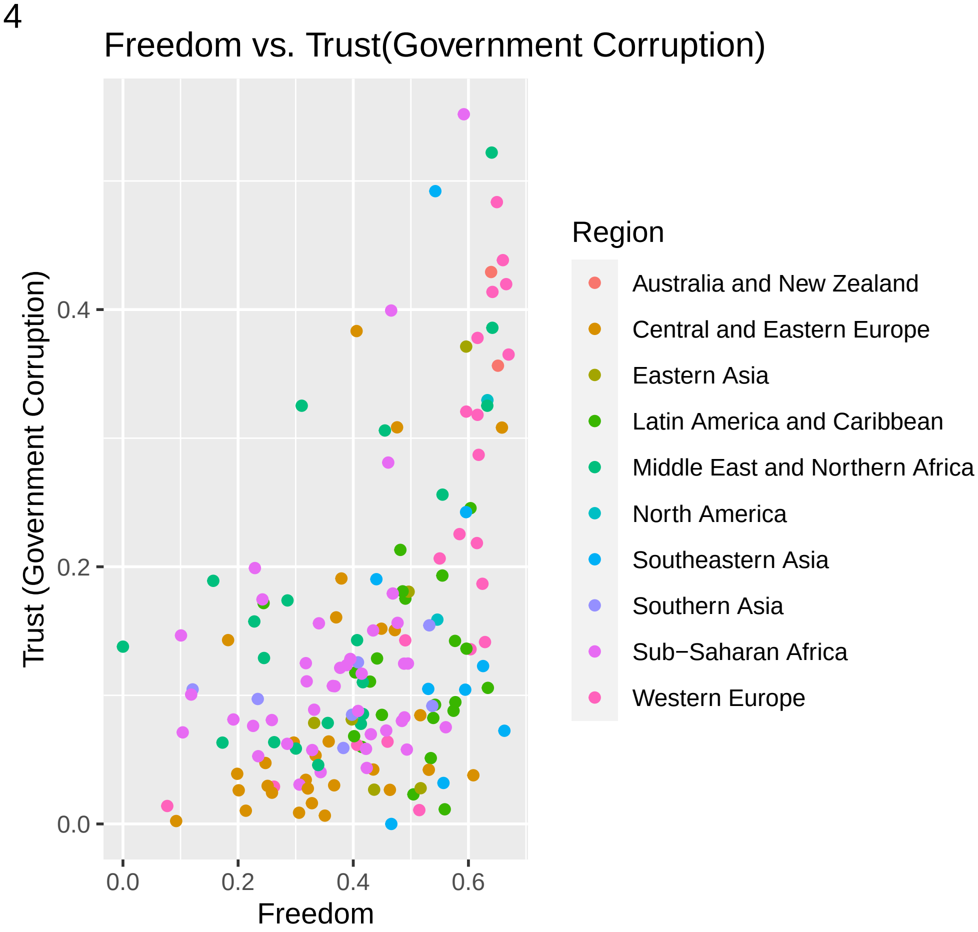


In the 158 countries, more than 40 of them are from Africa, there is a variable called "Middle East and Northern Africa", in this variable including country from middle east, such as Turkey, Qatar. So we are not using that variable, we only focus on "Sub-Saharan Africa". "Sub-Saharan Africa" more than 25 percent in this report, like what I said, most of the countries in Africa are unhappy, and there are a lot of countries in Africa, I wondering if Africa has bring down the happiness score of the whole dataset, then I get the mean. In fact, it's not much lower, because most African countries are still below average. Next, I am interesting to see what can directly affect happiness in Africa besides economy, for example, health, freedom, govenment corruption, etc.

```
## # A tibble: 1 x 1
##   `Happiness Score`
##               <dbl>
## 1              5.38
```

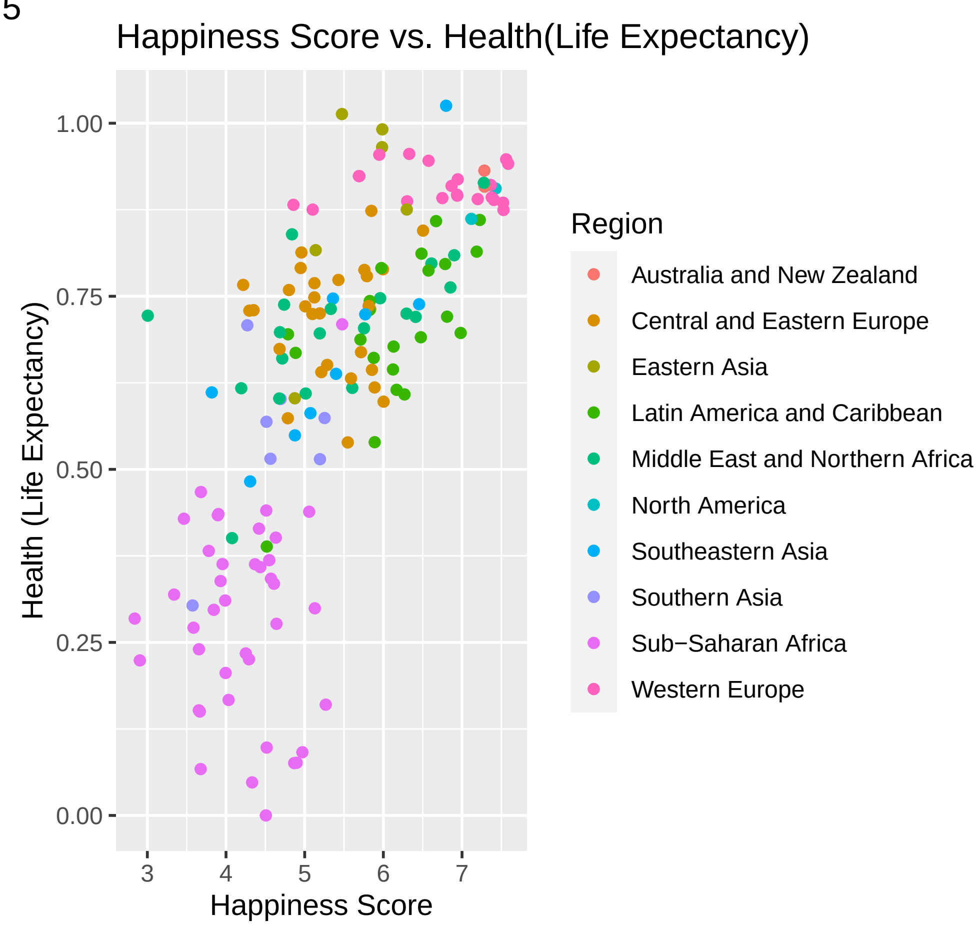
The figure 3.4 shows that the most purple below 0.2, lower than average, meant most African governments are corrupt. Government corruption is also a main factor in Africa's unhappiness. In fact, if the black people in power, the lack of education level led to the unlimited increase of desire, once in power, increased corruption, resulting in the corruption of the country. However, the life of the black people at the bottom has not been improved at all. The people at the bottom who have lost hope, without the traditional education that poor people are immune to poverty, have become lazy, thieves and robbers. For the freedom, some countries have totally indulged their people, making crime everywhere. Most people no longer trust the government.

Figure 3.4



According to the article, the average age in Africa is 19 and Africa's total fertility rate is 4.7, that means an average of 4700 children per 1000 African women. Coupled with the famine in some cities, many people starve to death every year. It is obvious from the plot that only Africa has lowest life expectancy.

Figure 3.5



## *Conclusion*