EDA Project

Tzu Yun Kan

6/15/2019

Explore and uncover insights of gpaminder.csv, and represent the insights by different data visulization

Dataset:

- life - life expectancy
- income - gdp per capita
- year - from 1800 to 2015
- country - **197** unique countries
- region - **6** unique regions
- population - census data collected about every 10 years

Describing data

- Number of observations of the dataset: **41284** rows
- Number of variable of the dataset: **6** variables
- Number of missing value of the dataset for each variables are **0, 0, 0, 0, 2341, 0**
- Types of variables:

1. Country is **factor**
2. Year is **integer**
3. life is **numeric**
4. population is **factor**
5. income is **integer**
6. region is **factor**

Preprocessing: - Imputing the missing data of the population variable by most recent non-missing value, and convert the populaiton variable into numeric type

- Removing rest of the missing value from the dataset
- population is now **numeric**
- Number of missing value of the dataset for each variable are now **0, 0, 0, 0, 0, 0**

How disperse is the data (e.g price), range of variables (e.g. years)

- Standard diviation of the Year is **62.3547089** years
- Standard diviation of the income is 1.0096544×10^4 dollars
- Standard diviation of the life is **16.1453431** years
- Standard diviation of the population is 6.1273623×10^7 number of people

Analysis

Questions:

1. Which 5 countries that have longest life expectency from 1970 to 2015 in average? How does the distribution of life expectency like for those countries?

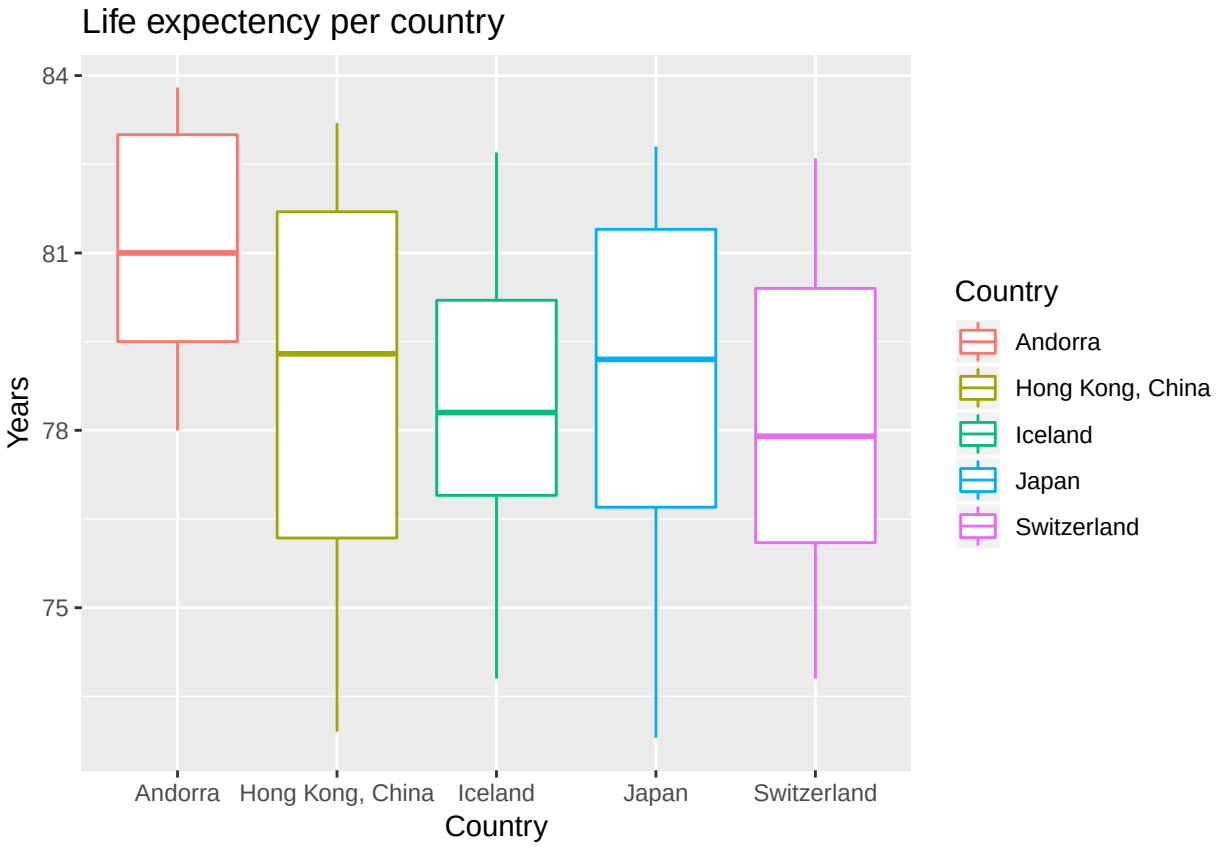


Figure 1 is a boxplot which shows the life expectency of Andorra, Hong Kong, Iceland, Japan, and Switzerland since 1970s. Besides Andorra, Hong Kong's life expectency is higher in average and also more sparse than other countries.

2. How does these 5 countries life expectency trend like from 1800 to 2015

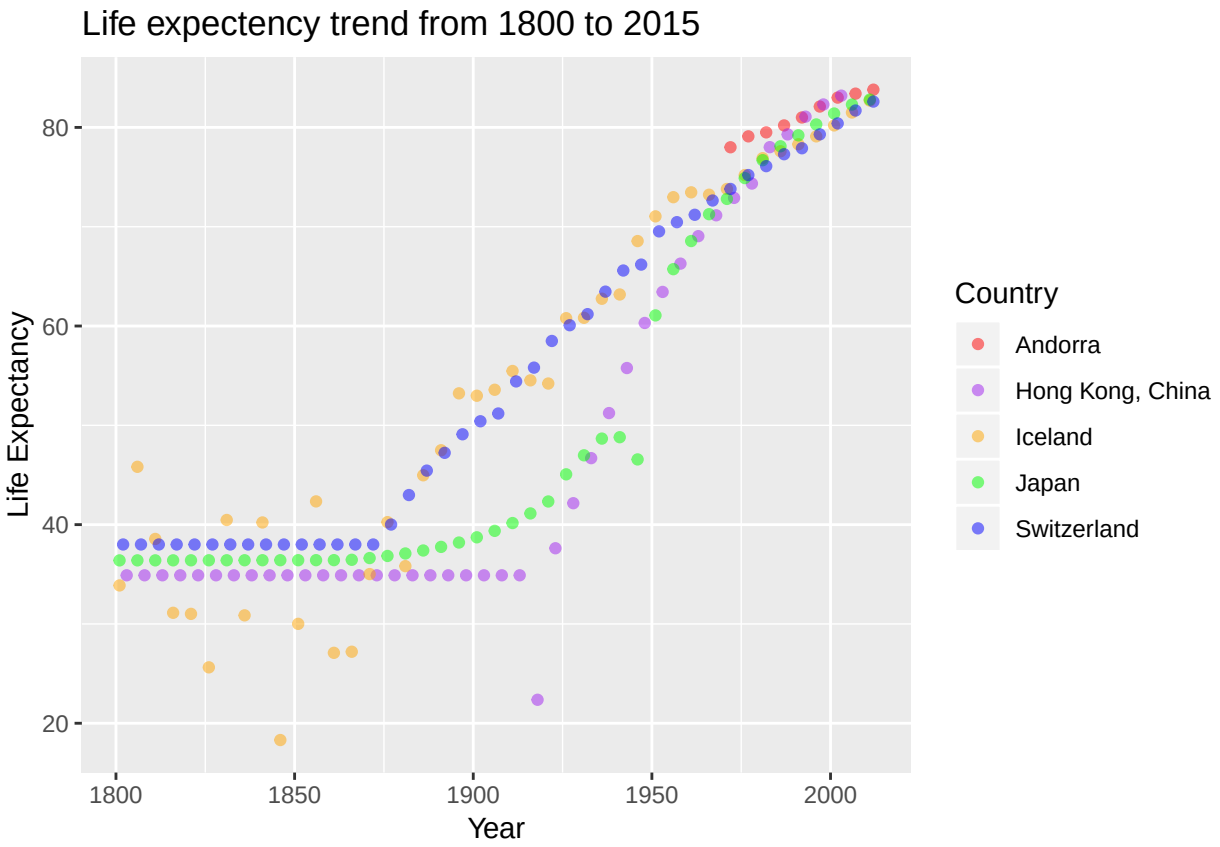


Figure 2 is a scatter plot which shows the life expectency trend of Andorra, Hong Kong, Iceland, Japan, and Switzerland since 1800. From 1875 to 1950 is the time period that life expectency increased the most.

3. How does life expectency related to GDP for these five countries?

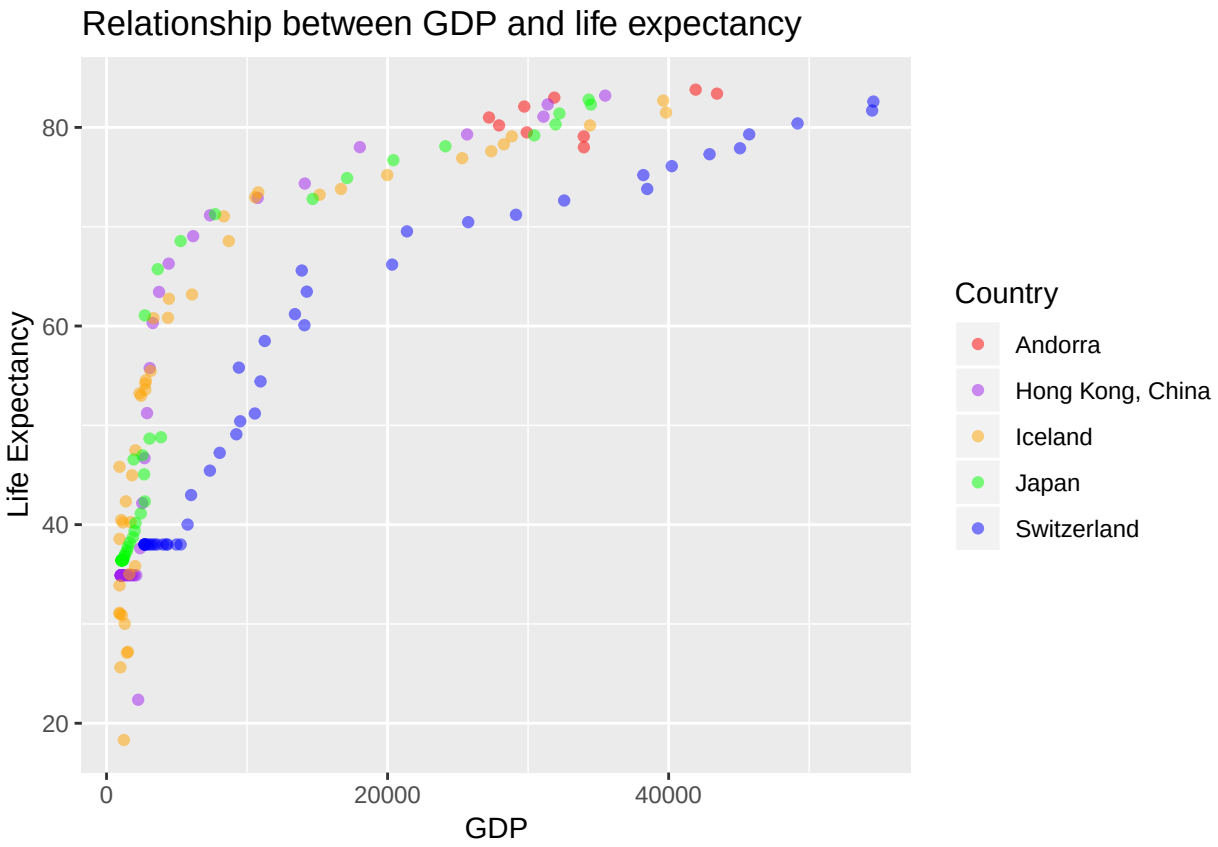
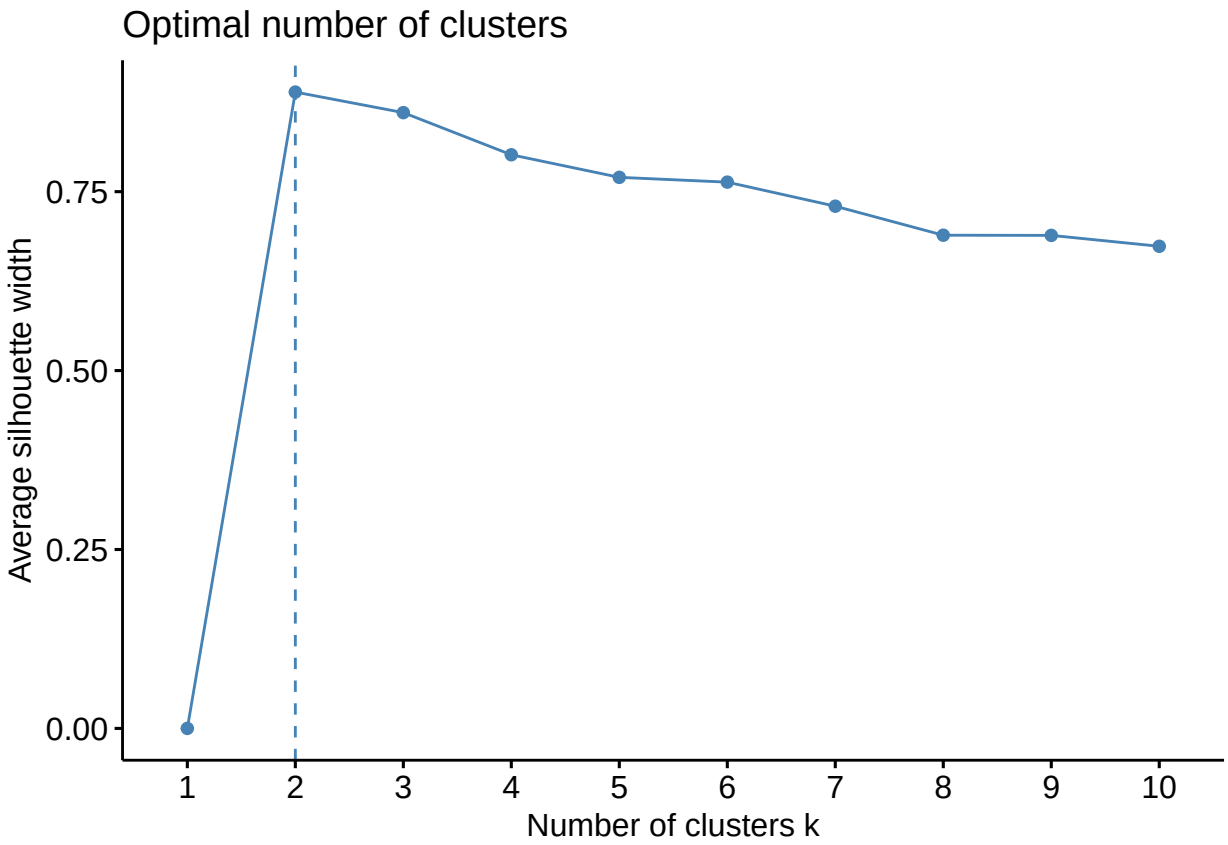
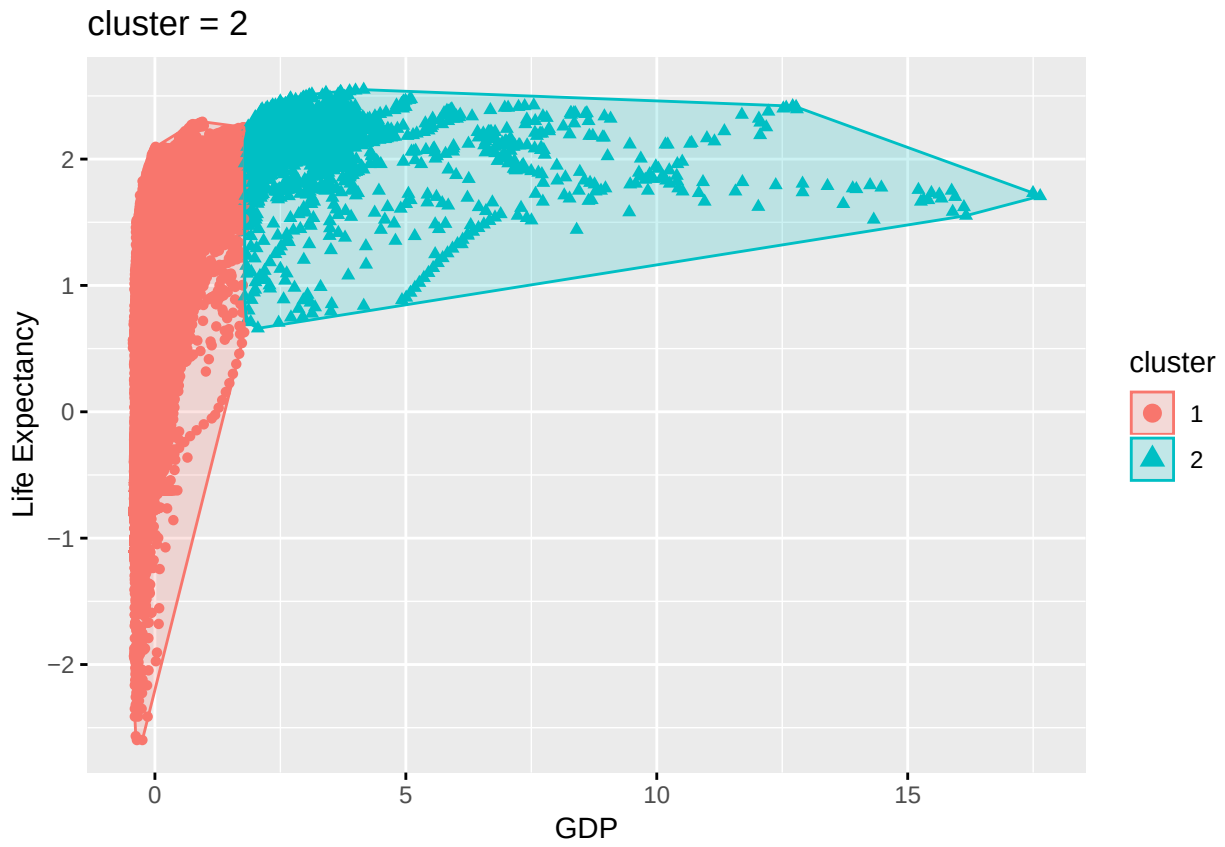


Figure 3 is a scatter plot which shows the life expectency related to GDP of Andorra, Hong Kong, Iceland, Japan, and Switzerland. It indicates higher the GDP a country has, higher the life expectancy it is.

Clustering Analysis



The **Figure 4** is for finding the number of clusters using Silhouette Method.



The **Figire 5** is the visualization of kmeans of 2 clusters. It indicates that high GDP will lead to higher life expectancy not only for the above 5 countries, but also applied for rest of the world. However, when the income comes to a certain level, the increase of life expectancy will slow down.