Final_report

Shipeng Sun

June 14, 2019

Data

For this report, we have dataset gapminder which contains a collection of records measured from 1800 to 2015. There are 6 variables that we will focus on. We will do analysis on life expectancy which maybe effected by GDP, region, country and year.

Variables in the gapminder data set are:

- **life** - life expectancy
- **income** - gdp per capita
- **year** - measured year from 1800 to 2015
- **county** - countries in the world included
- **region** - regions includes all contries
- **population** - census data collected about every 10 years

Questions to Answer

1. What's the relationship between GPD and region?
2. What's a potential relationship between a country's GDP (income) and life expectancy?
3. For example in region America, how does it look like regarding to Life Expectancy Per Region in recent years?

Data narrative summary

1. There are **41284** observations in the dataset.
2. There are **6** variables in the dataset. Number of missing value of the dataset for each variables are **0, 0, 0,** 2.5817×10^4**, 2341, 0**
3. Type of variables:

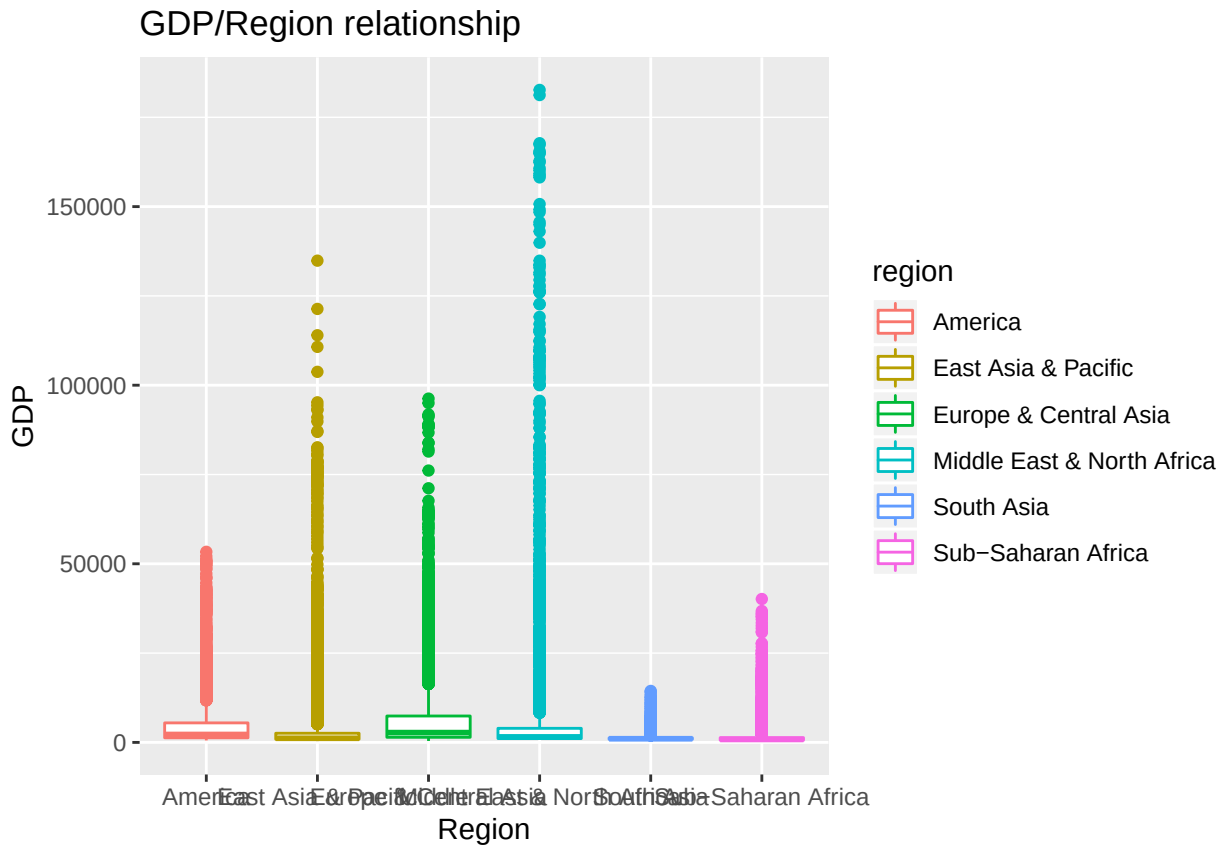
- "Country" is **character**.
- "Year" is **numeric**.
- "life" is **numeric**.
- "population" is **numeric**.
- "income" is **numeric**.
- "region" is **character**.

4. How disperse is the data:

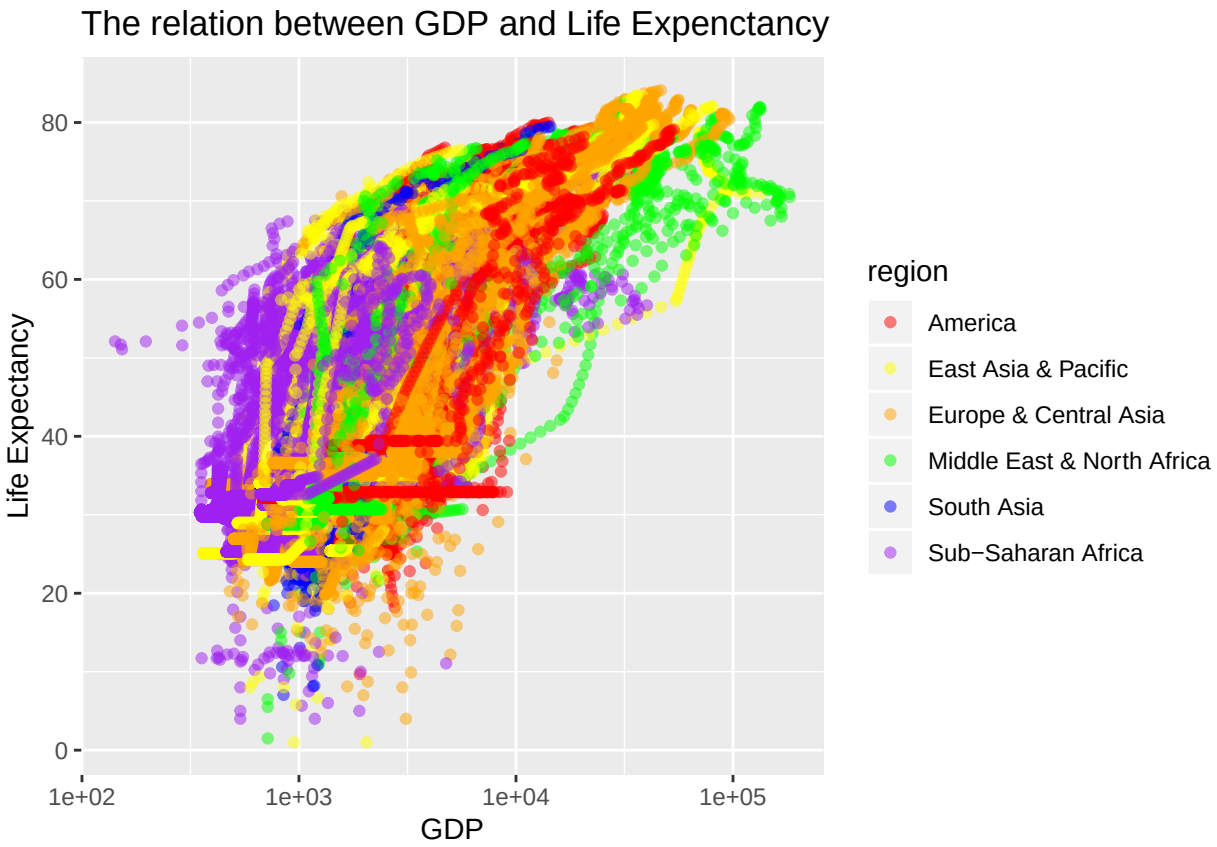
- Range of "Year" is **1800, 2015**.
- Range of "life" is **1, 84.1**.
- Range of "income" is **142,** 1.82668×10^5.

5. Data wrangling: The average life expectancy in year 2015 is **71.7634831**.
6. Preprocessing steps: To deal with missing data, I filled the missing population data by most recent not null values because of data continuity. Besides, rows with empty income values are removed, and type of population is converted from factor to numeric type for later processing. Below shows details.

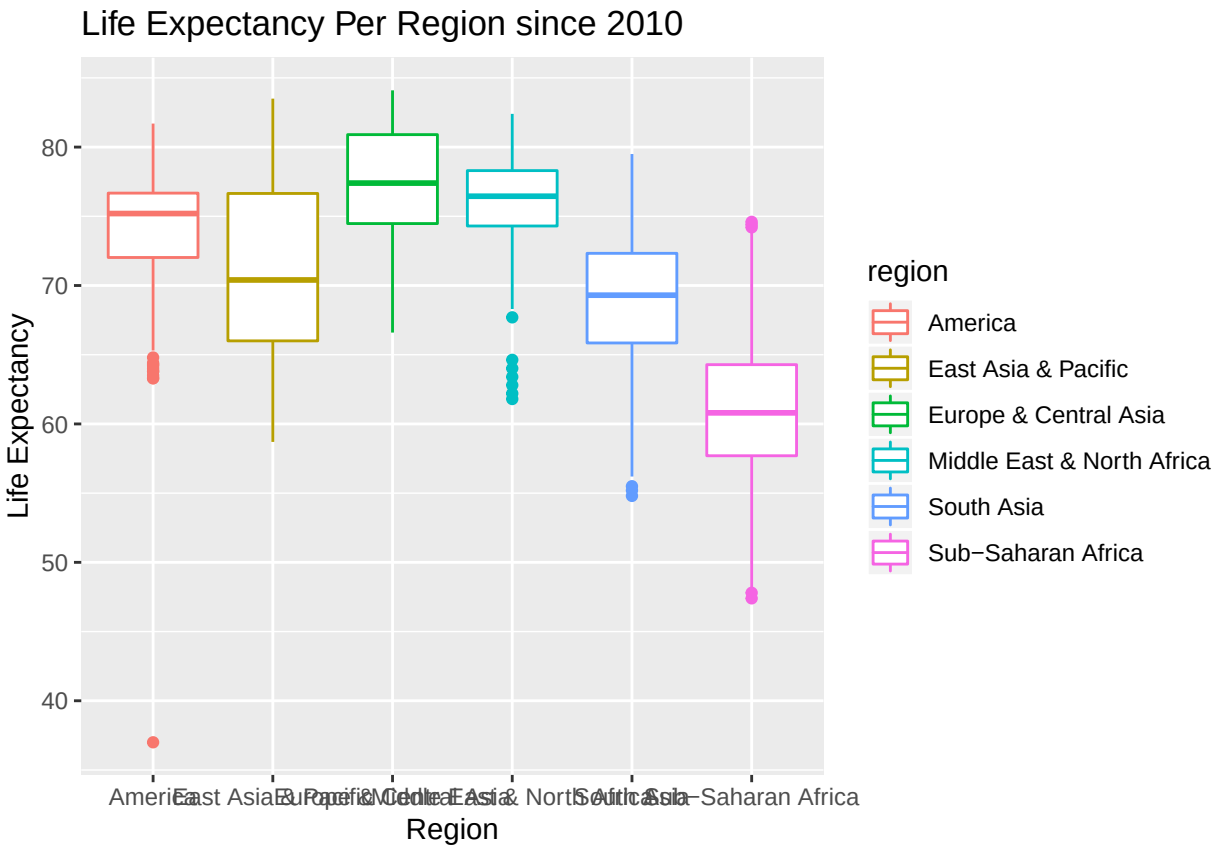
Exploratory Plots



The above **Fig. 1** is a boxplot which shows the total GPD (income) per Region.



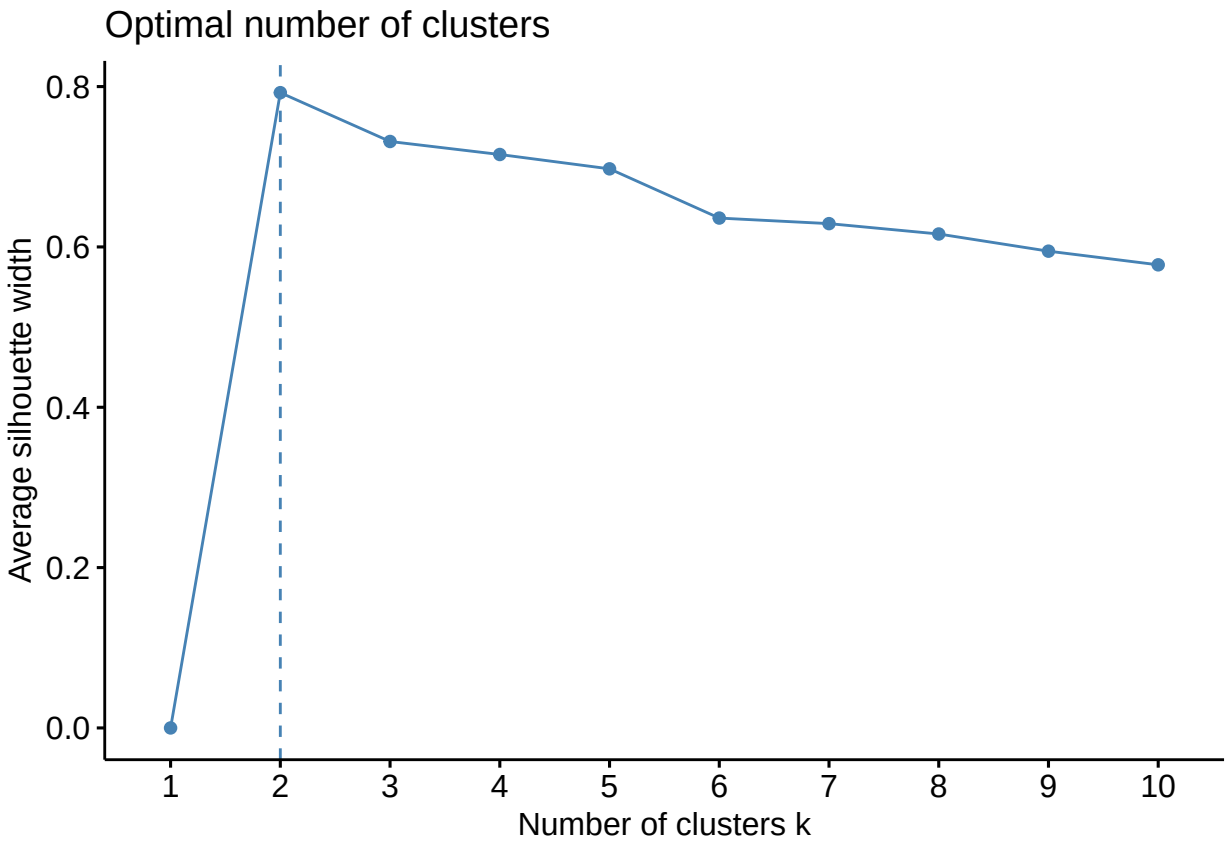
The above **Fig. 2** is a scatter plot which shows the relation between GDP and Life Expenctancy.

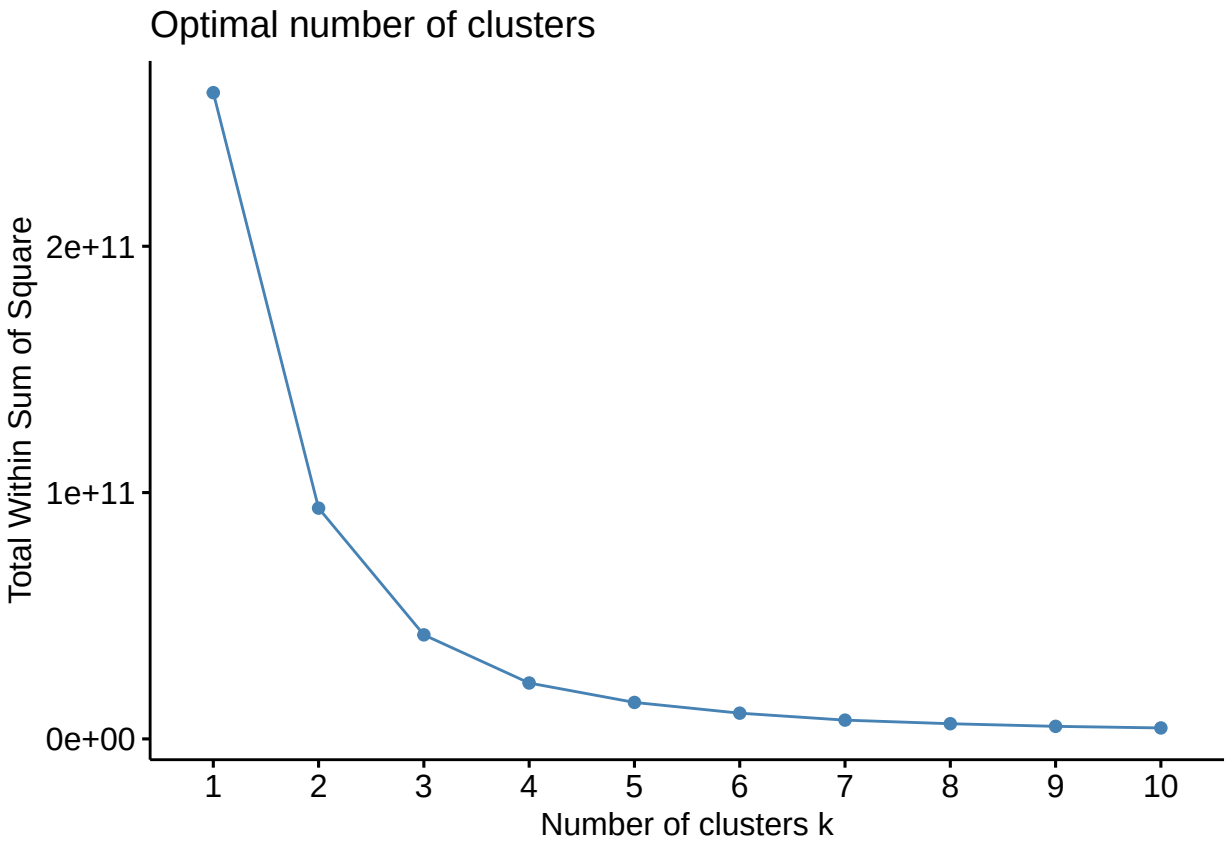


The above **Fig. 3** is a boxplot which shows the Life Expectancy Per Region in the recent years.

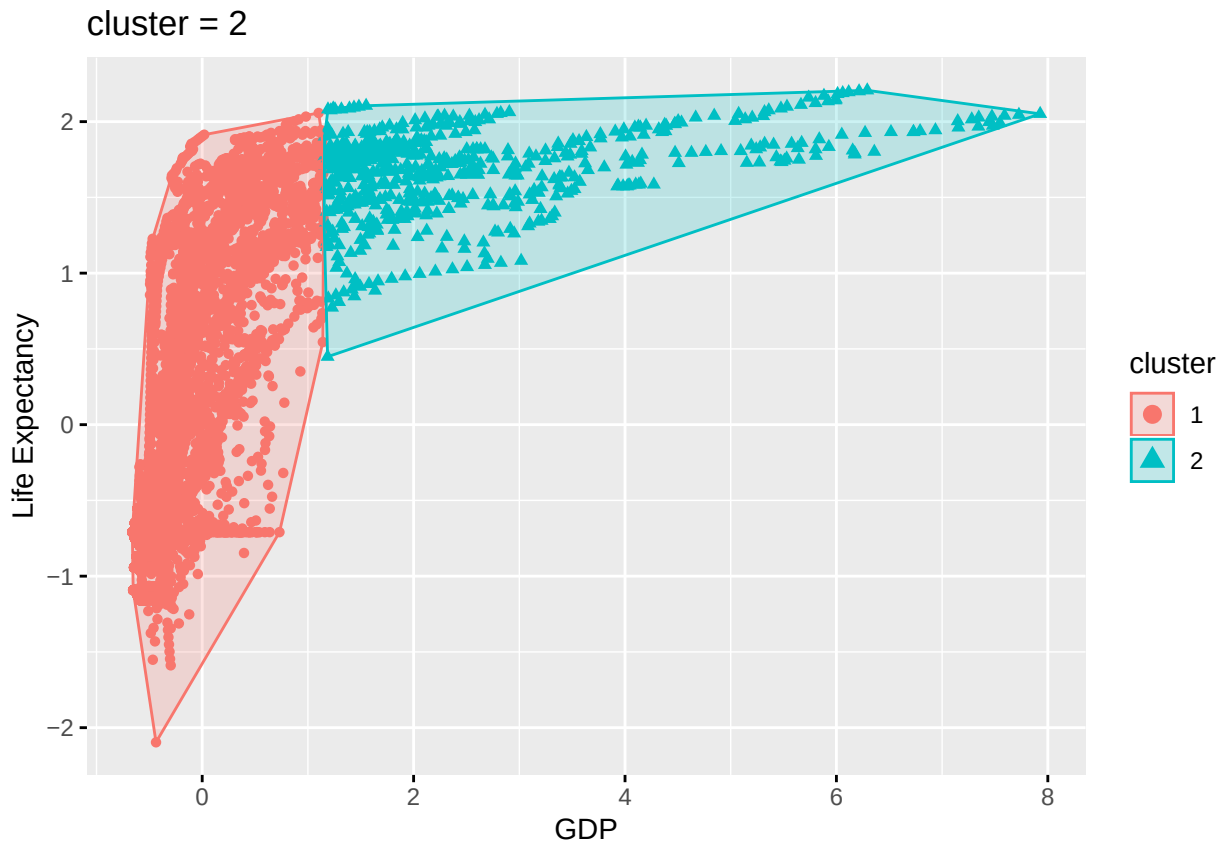
Clustering Analysis

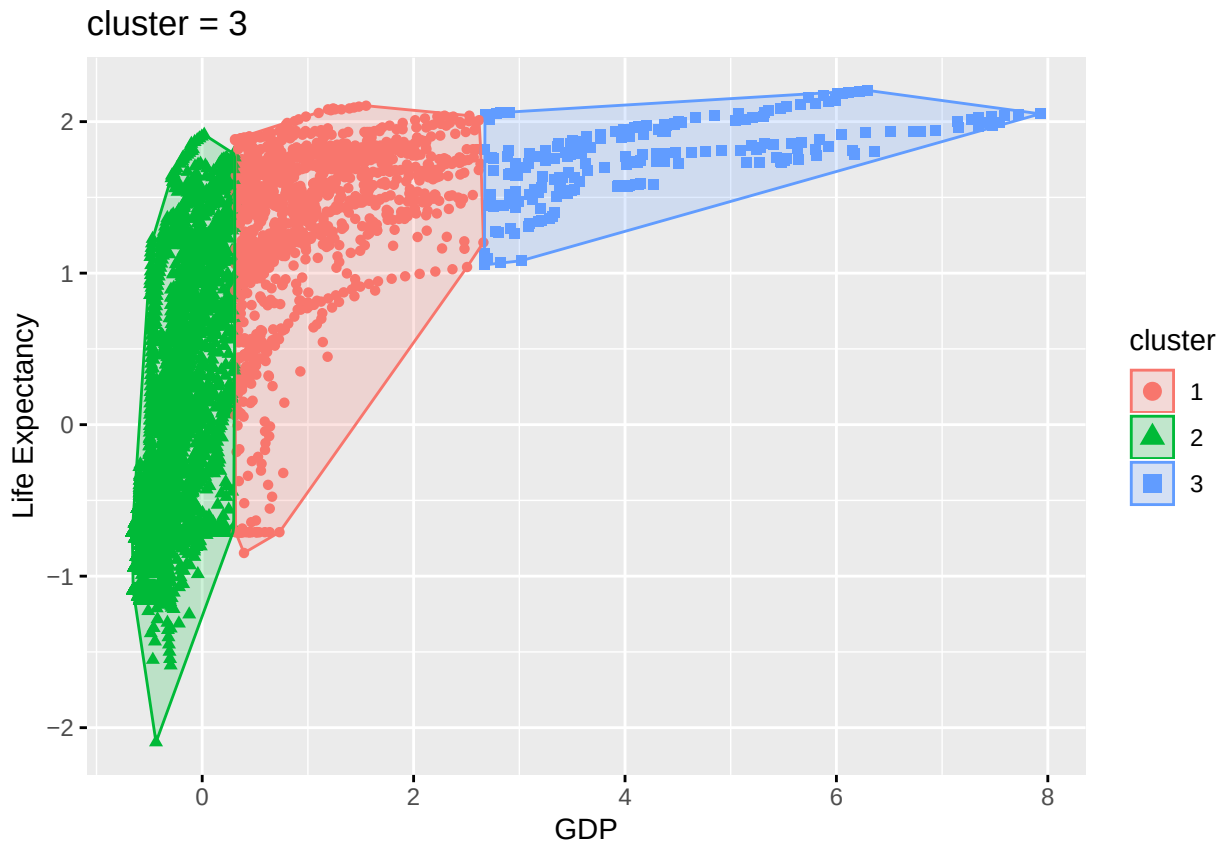
Note: I take America region data as an example here.


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
Number of clusters k



The **Fig. 4** (above two figures) is for finding the number of clusters using Silhouette Method and Elbow Method.


cluster = 2
Life Expectancy
GDP
2
1
0
−1
−2
0
2
4
6
8
cluster
1
2



The **Fig. 5** (above two figures) is the visualization of kmeans of 2 clusters and 3 clusters .

Answers:

1. From **Fig. 1**, we know region has effect on GPD and GPD have big difference between diffrent regions.
2. From **Fig. 2**, there's a positive relationship between a country's GDP (income) and life expectancy.
3. From **Fig. 5**, lower income (GDP) will lead lower life expectancy, but when the income comes to a certain high level, the life expectancy won't increase too much.