Final Report

Shujun Zhang

June 15, 2019

Data

The `gapminder` data set contains a collection of 6 variables measured between 1800 and 2015 on the world.It's describing life expentency depending on factors like GDP, Region, population etc.

Some of the variables in the `gapminder` data set are:

- **life** - life expectancy
- **income** - gdp per capita
- **year** - the range from 1800 to 2015
- **county** - countries in the world
- **region** - 6 regions includes all contries
- **population** - census data collected about every 10 years

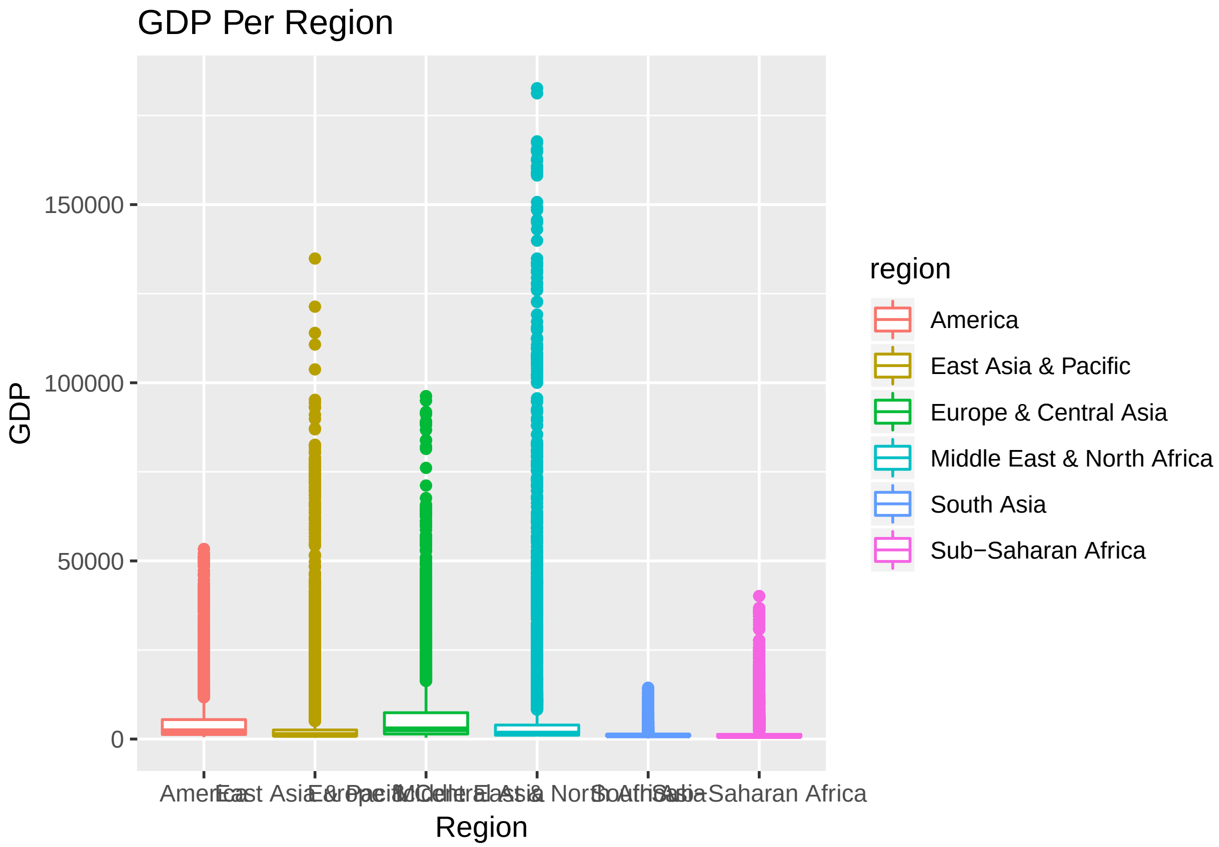
Questions to Answer

1. Is there a big difference in terms of GPD per 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?

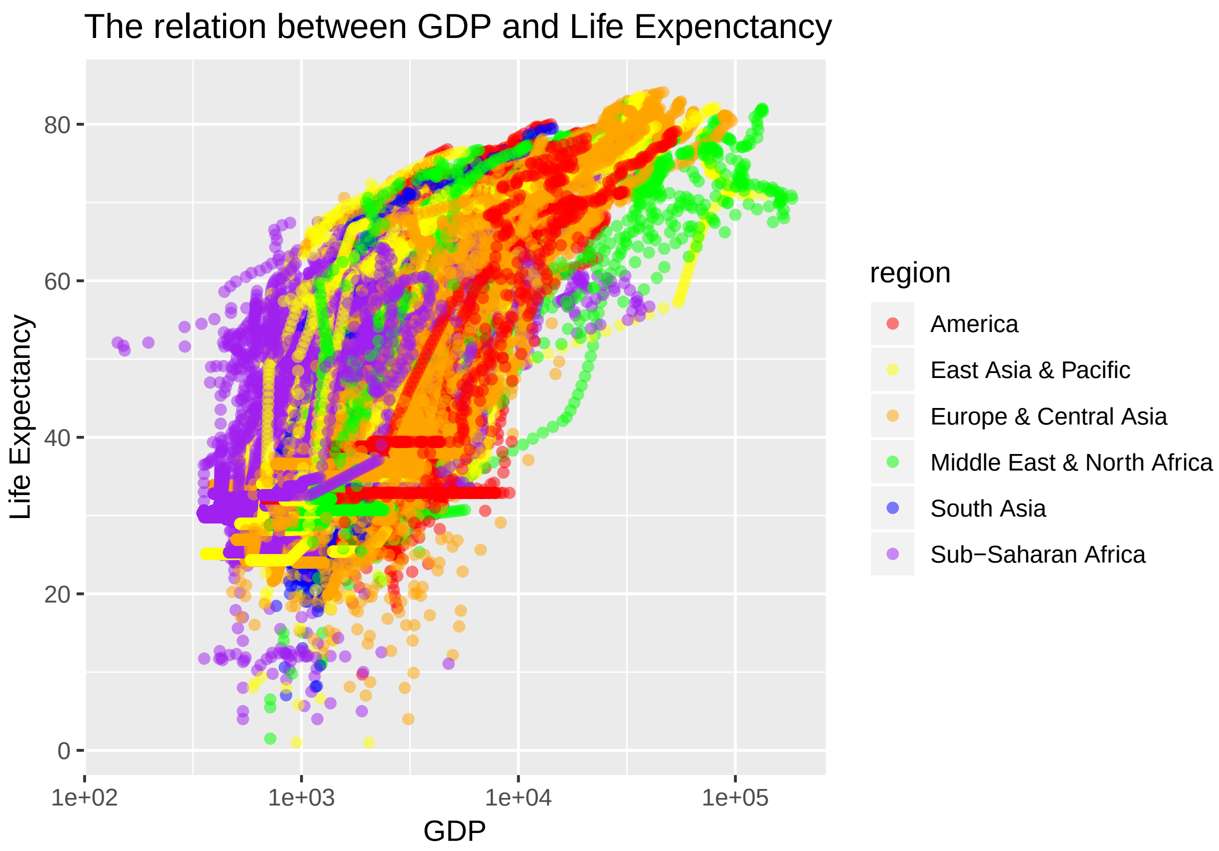
Data narrative summary

1. There are **41284** observations in the data.
2. Thee are **6** variables in the data.
3. Type of variables: "Country" is **factor**, "Year" is **integer**, "life" is **numeric**, "population" is **factor**, "income" is **integer** and "region" is **factor**.
4. How disperse is the data: Range of "Year" is **1800, 2015**. Range of "life" is **1, 84.1**. Range of "income" is **142, 182668**.
5. Data wrangling: The avaerage life expectancy in year 2015 is **71.7634831**.
6. Preprocessing steps: Bascially what I did was filling in the missing population data by the most recent non-missing values using 'fill' method. I also removed the rows with empty income values, and convert the type of population from factor to numeric type for later processing. Please check the in-line code for details.

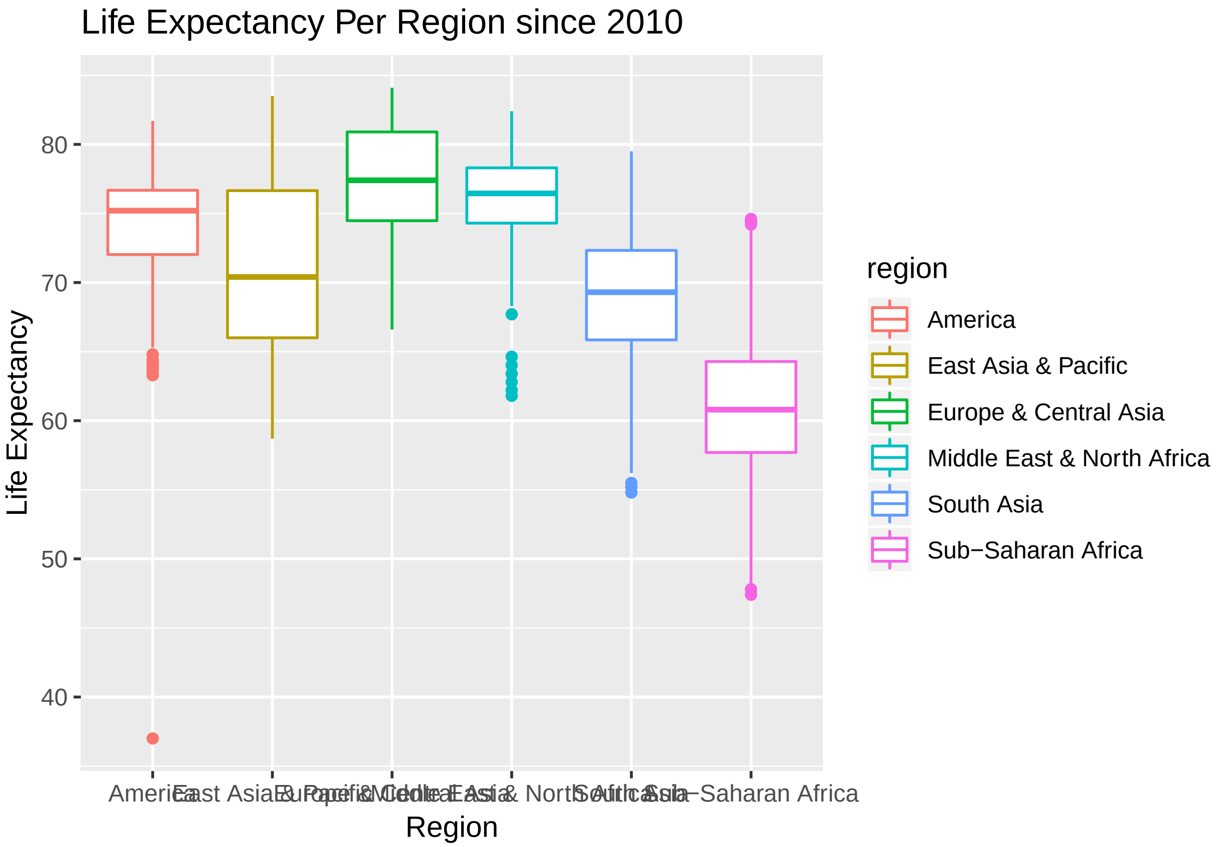
Exploratory Plots



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



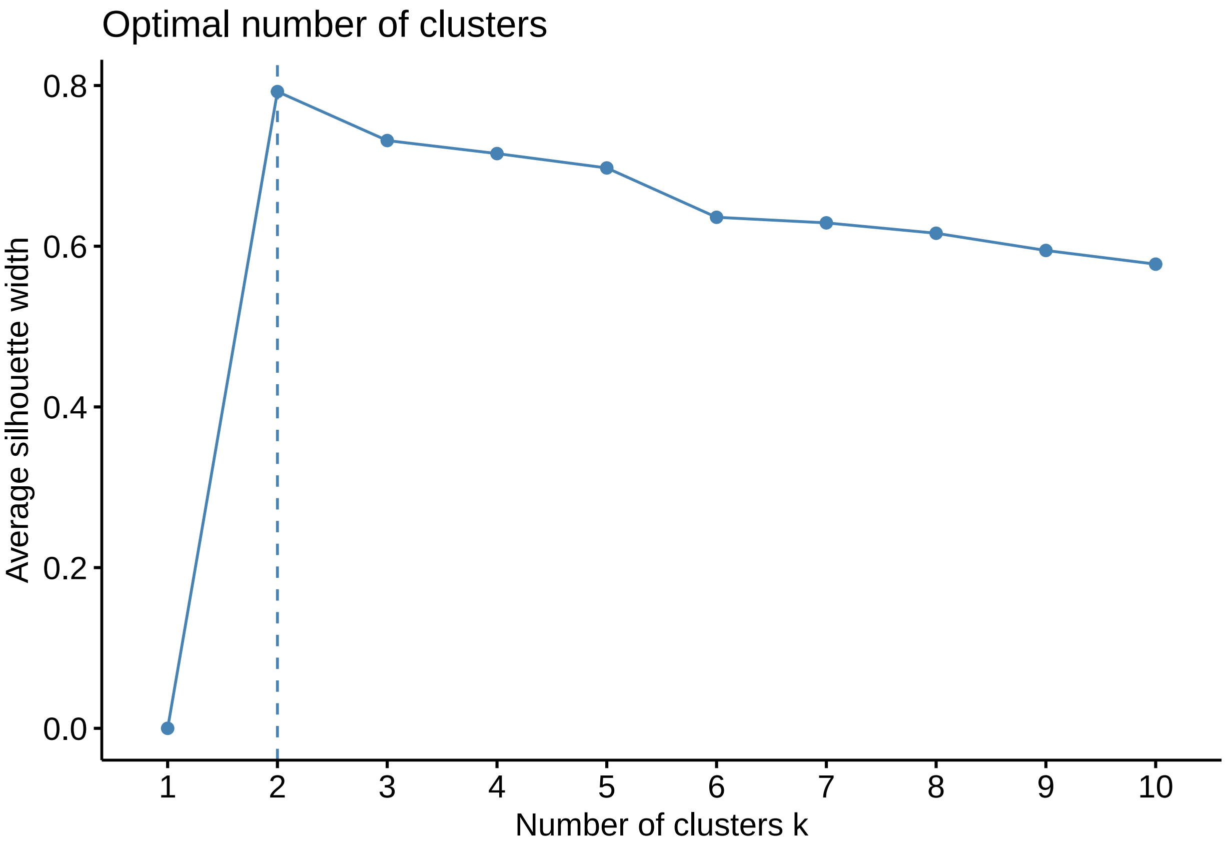
The above **Fig. 2** is a scatter plot which shows the relation between GDP and Life Expencantancy.

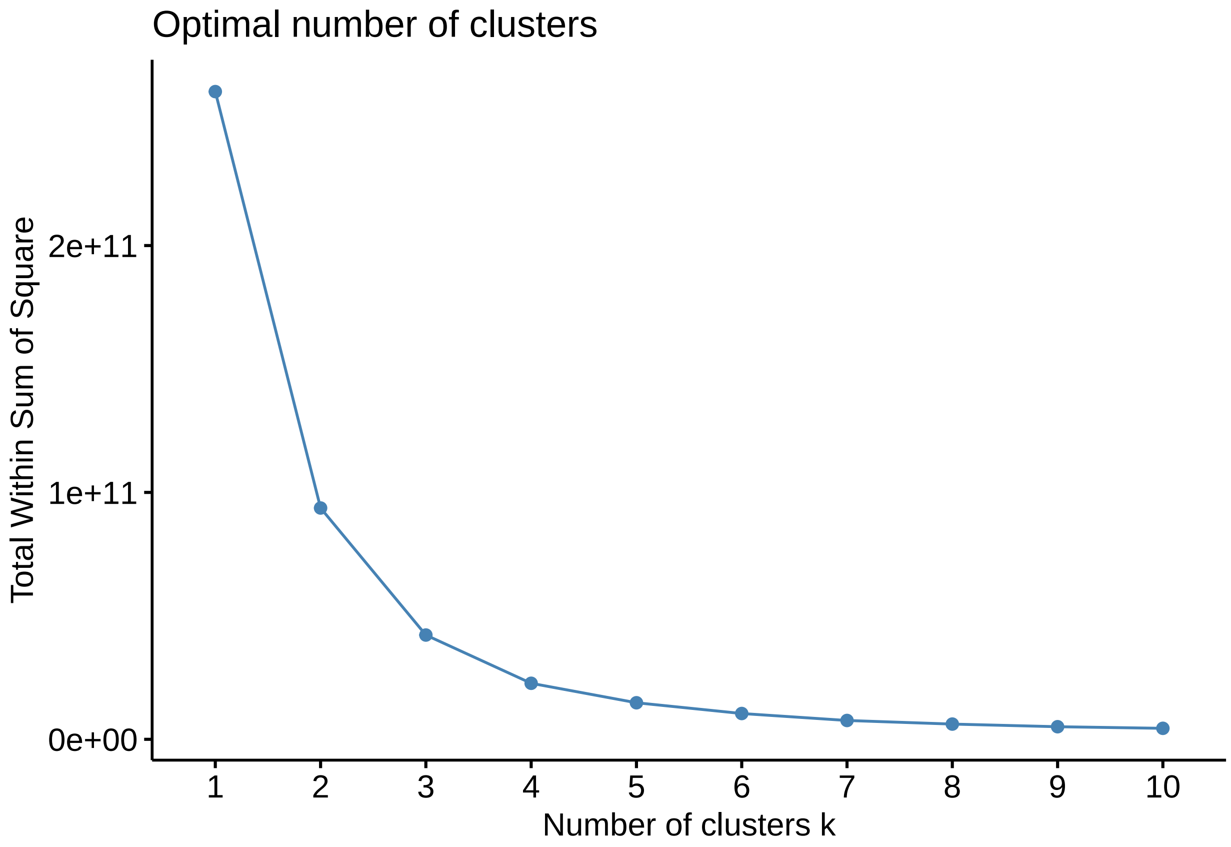


The above **Fig. 3** is a boxplot which shows the Life Expectancy Per Region in the recent years.

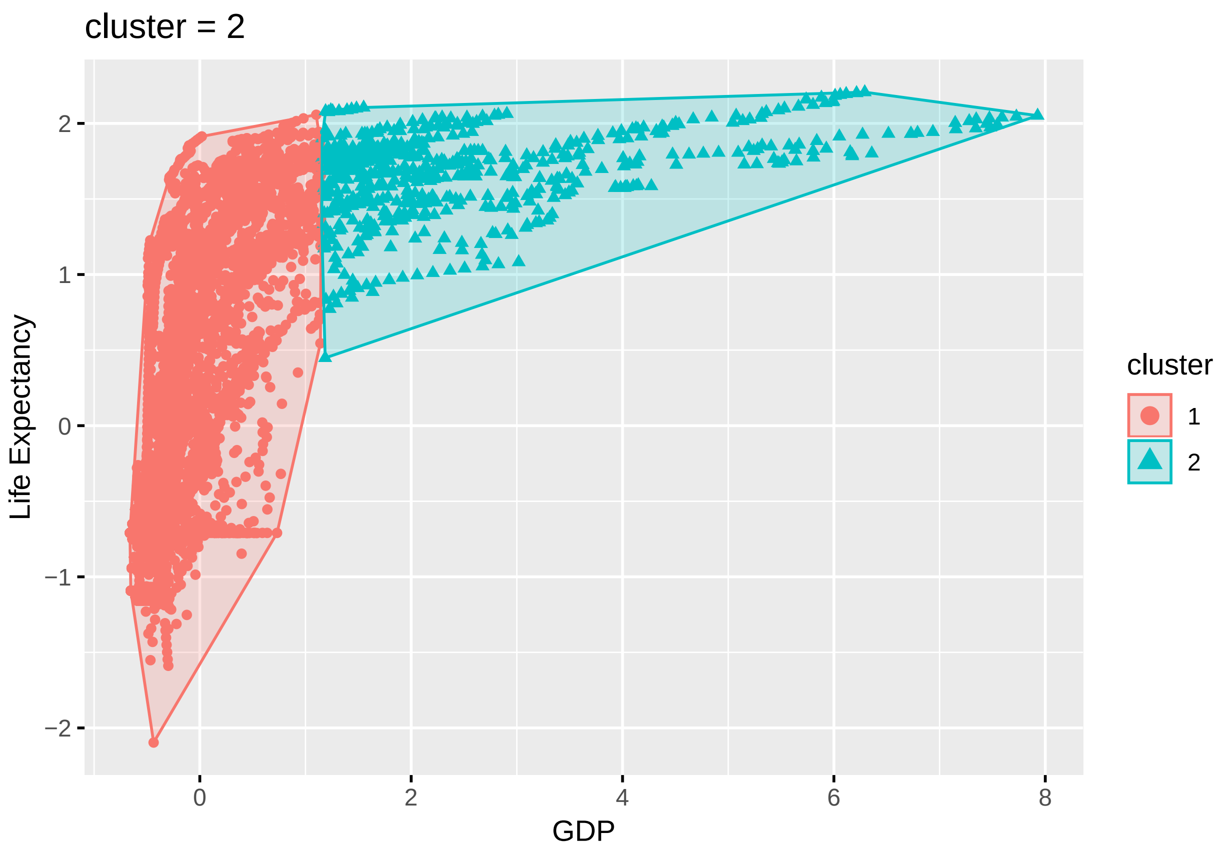
Clustering Analysis

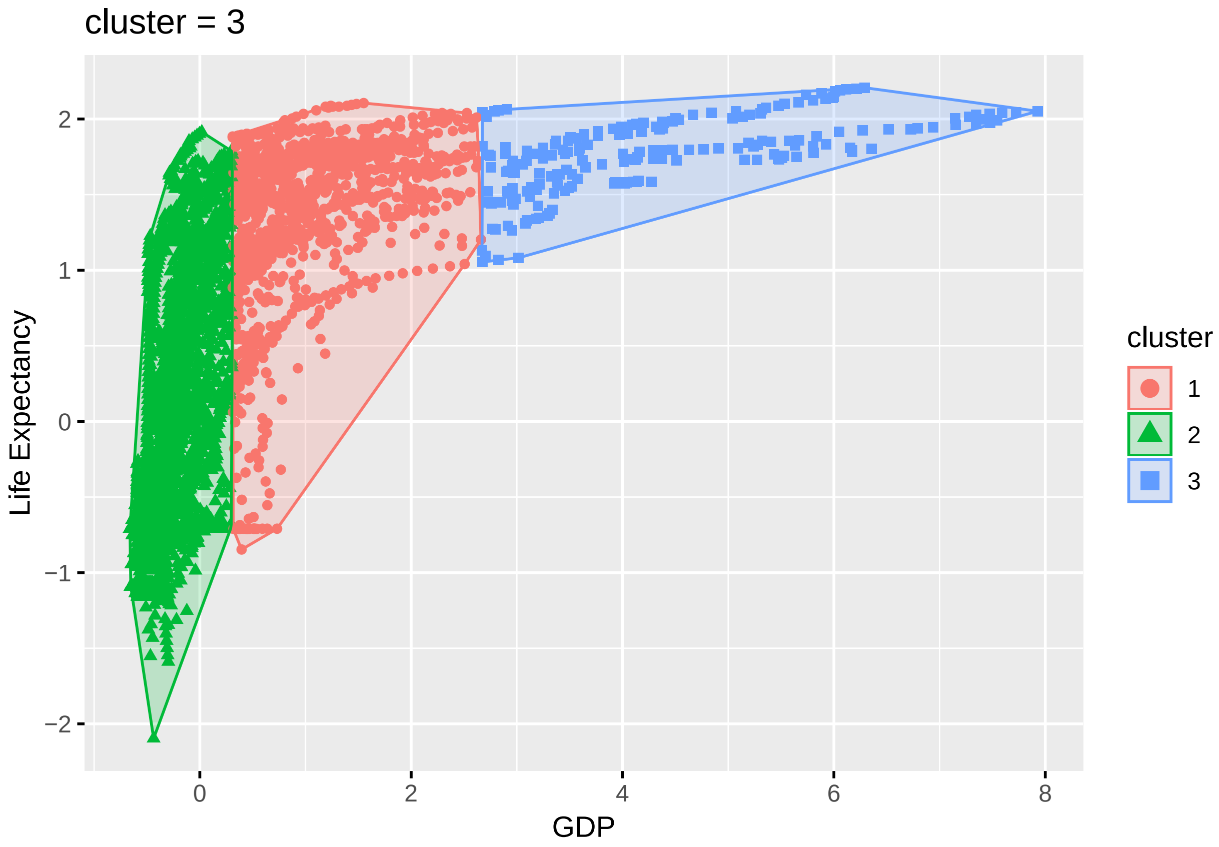
Note: I only consider data in region America to save running time.


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
2
1
0
−1
−2
0
2
4
6
8
GDP
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 could see differences between regions, especially for region Africa.
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.

References:

1. Phillips, N. D. (2016). Yarrr! The pirate's guide to R.
2. Peng, R. D. & Matsui, E. (2018). The Art of Data Science: A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC.
3. Grolemund, G., & Wichham, H. (2018). R for Data Science.