ANLY 506 - Final Project

Antra Chowdhury

6/15/2019

Aim of the study

1) Identify the countries with highest population
2) How has the population affected the income of the country?
3) What is the impact of population on life expectancy?

Data Summary

The given dataset "gapminder.csv" has 41284 observations and 6 variables. "Country" is a factor variable with 197 levels."Year" is an integer variable ranging from 1800 to 2015. "life" is a numeric variable rangine from 1 to 84.10 and has an median value of 35.12. "population" is a factor variable with 15260 levels. "income" is a integer variable ranging from 142 to 182668 and a median value of 1450. "region" is a factor variable with 6 levels.

Data wrangling

There are about 25817 NULL values for population and 2341 NULL values for income. To solve for missing values for population, data from every 10 years will be used for the analysis. The observations with NULL income don't seem to be limited to a particular country or region and so it can be excluded from the analysis later.

Aggregation

By analyzing the population for different countries, it is observed that top 3 countries in terms of population is China, India and USA.

Plots

Using time series plots, we can look at the change of population as the top 3 most populated countries in the dataset. For China (Fig. 1), it looks like the population starting accelerating around 1950. On the other hand, India (Fig. 3) had a gradual increase in population till 1900 and thereafter the population started accelerating. For USA (Fig. 5), the growth in population seems to be somewhat linear.

The income in China (Fig. 2) seems to have been flat till 1990 despite the increase in population from 1950 and it was close to 10,000 in 2010. In India, the income (Fig. 4) started increasing in 1970s and it peaked to 4500 in 2010. In USA, the income increased gradually and it peaked at 50,000 in 2010.

Figure 7 shows that the average life expectancy for United States is much higher than that of India or China.

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

Fig. 1 − China population over time

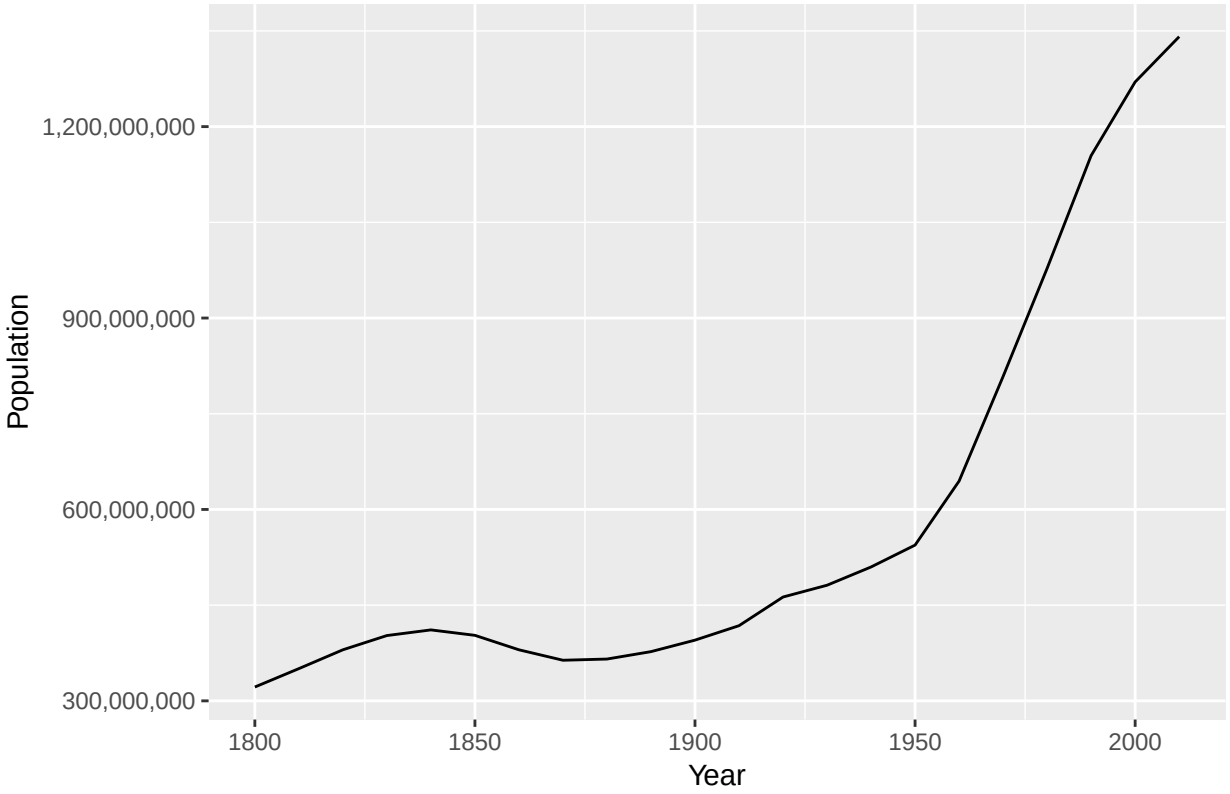


Fig. 2 − China income over time

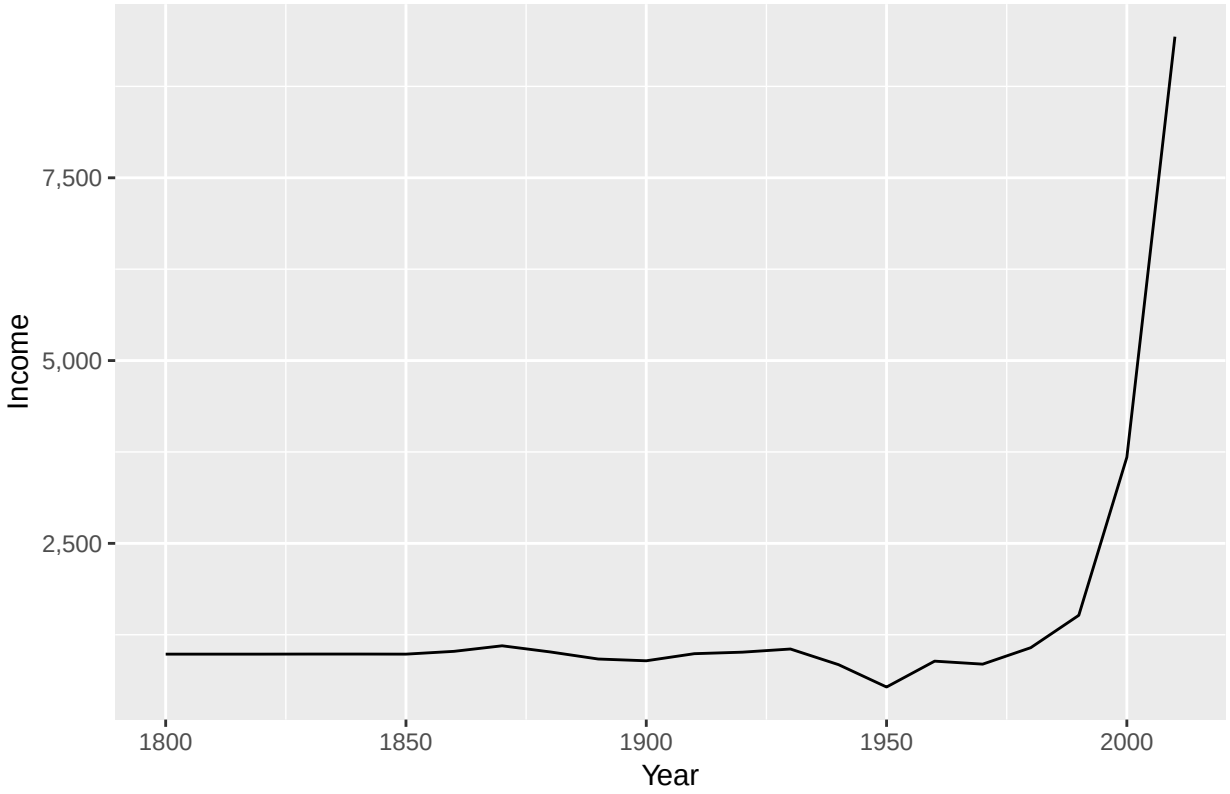


Fig. 3 − India population over time

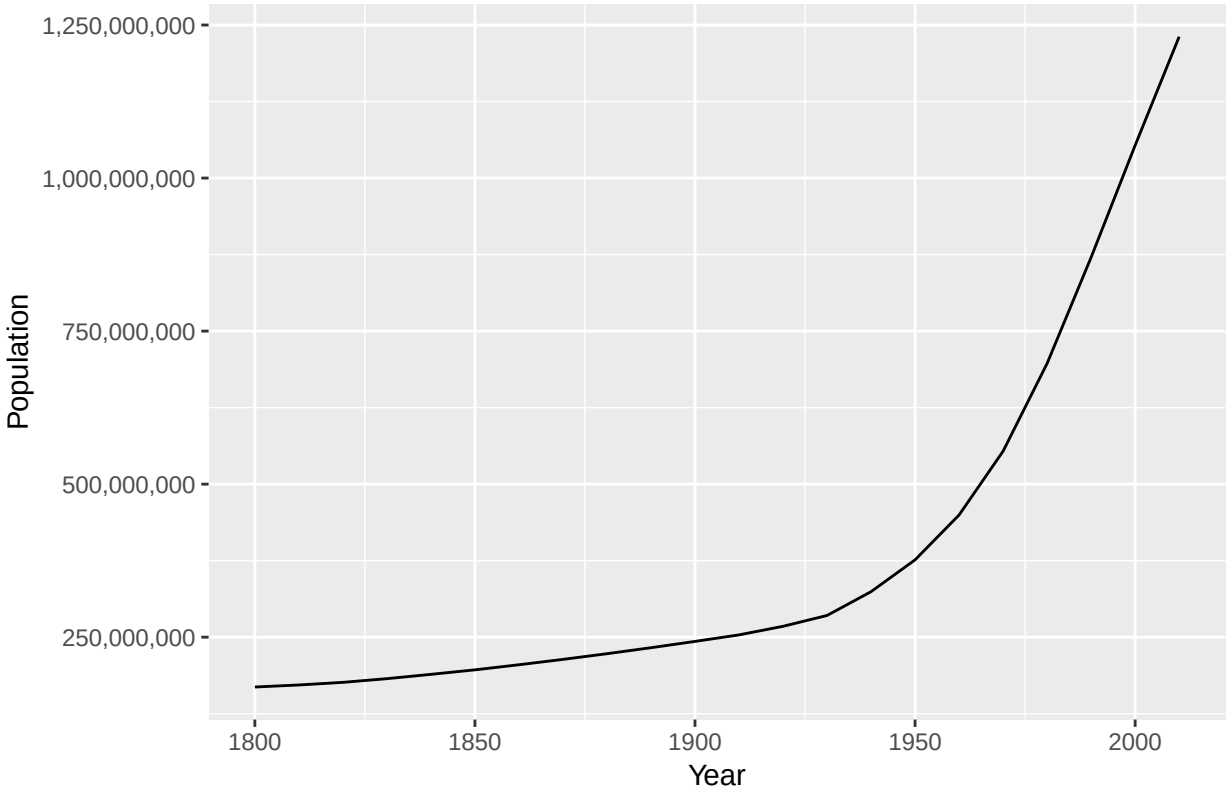


Fig. 4 − India income over time

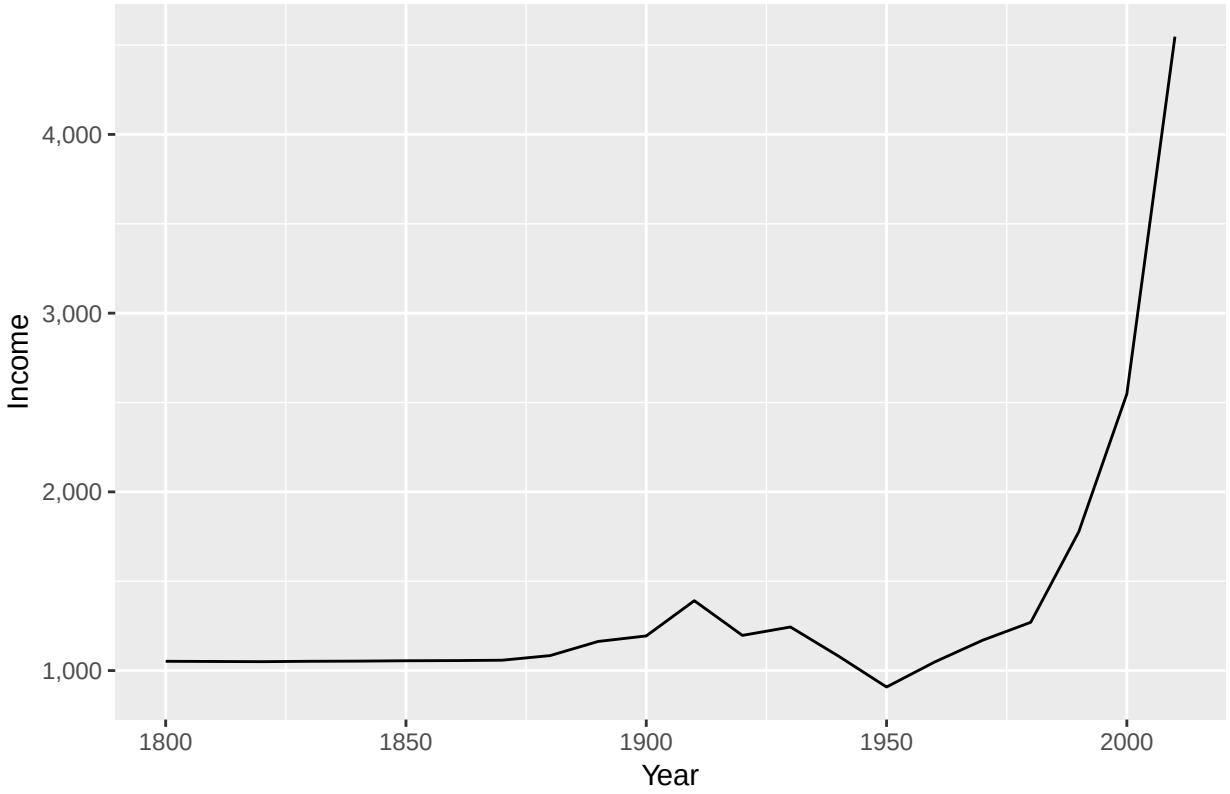


Fig. 5 − USA population over time

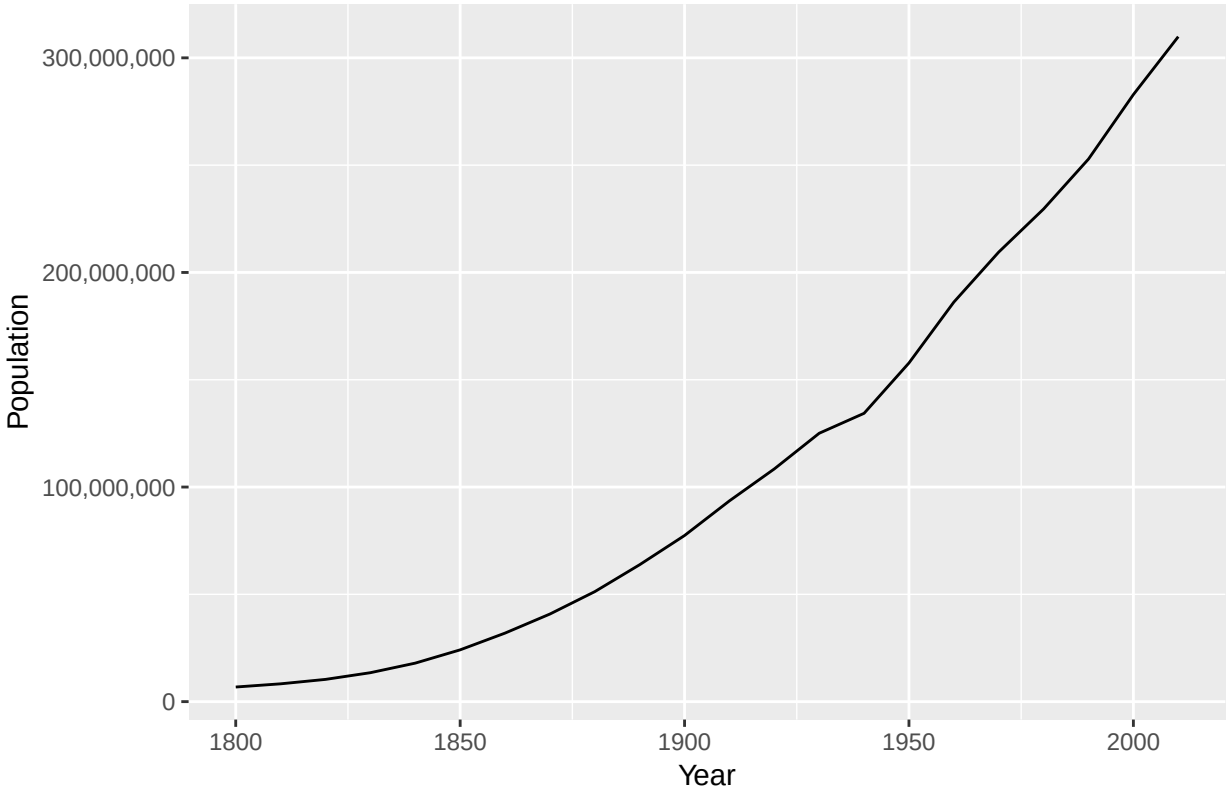
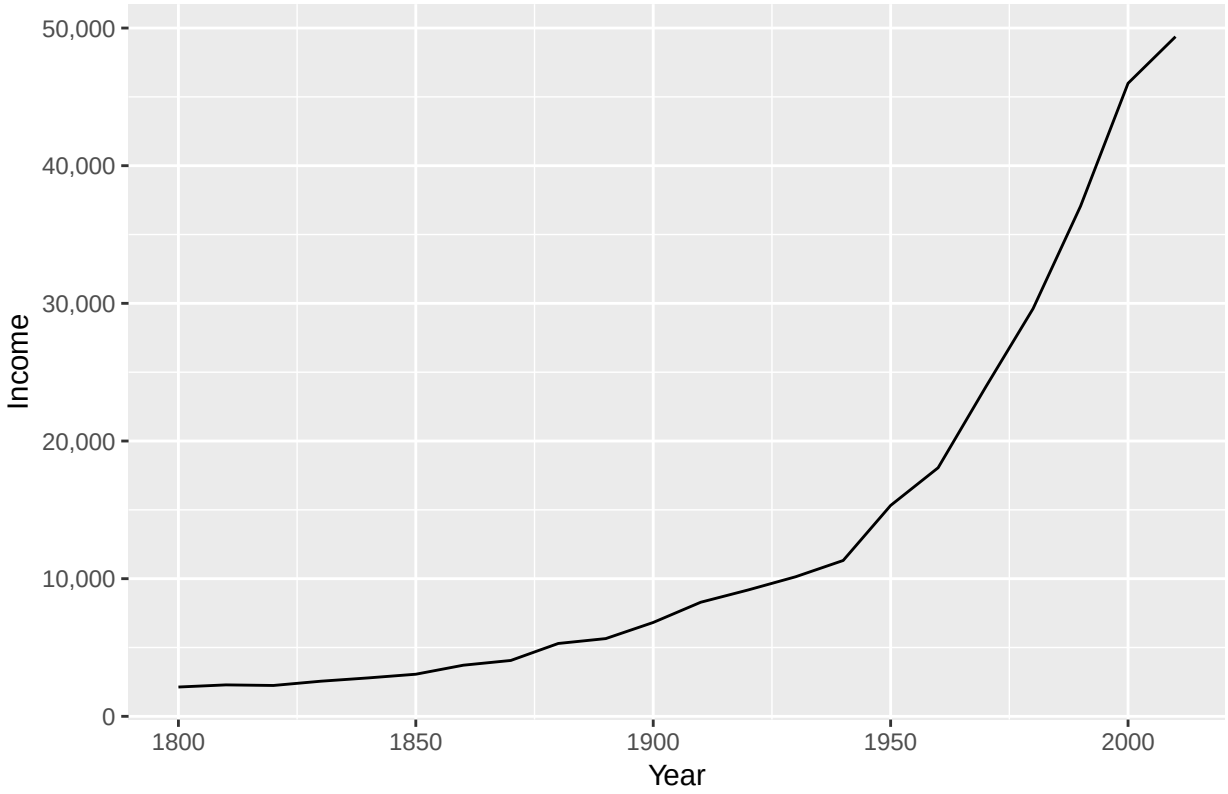
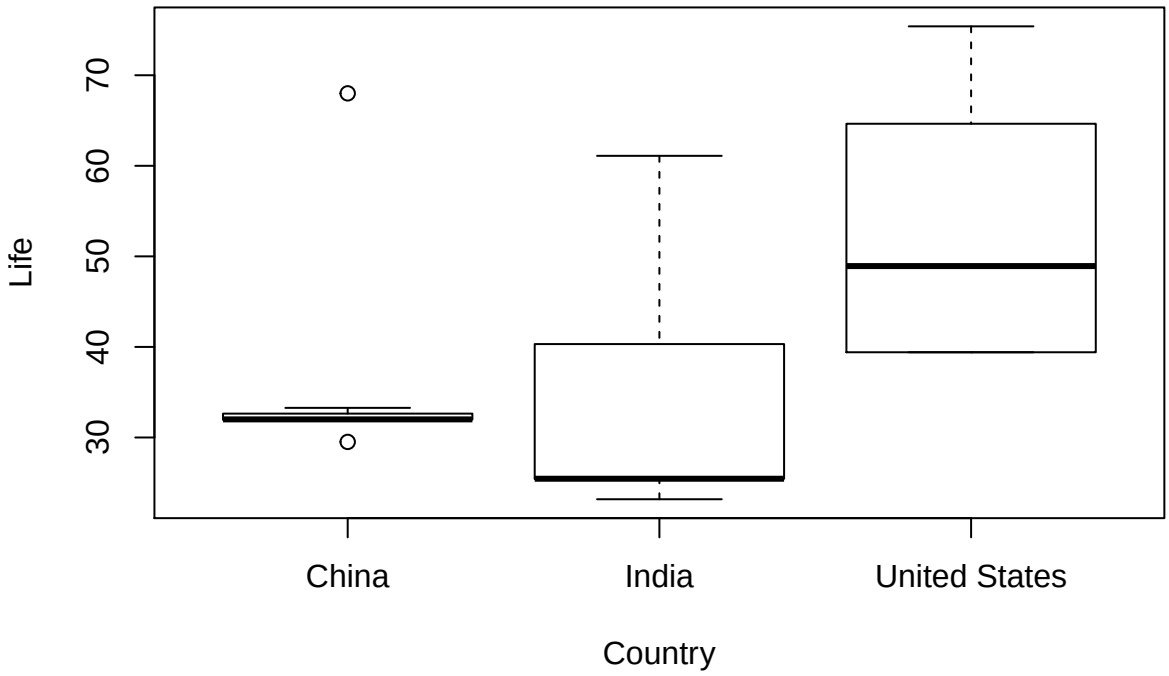


Fig. 6 − USA income over time




```
## Warning in Country == list("China", "India", "United States"): longer
## object length is not a multiple of shorter object length
```

Fig. 7 − Life Expectancy Distribution



Cluster Analysis After performing cluster analysis on income and population on the entire data set, it is observed that the data set can divided into 2 clusters, namely low population-high income and high population-low income.

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

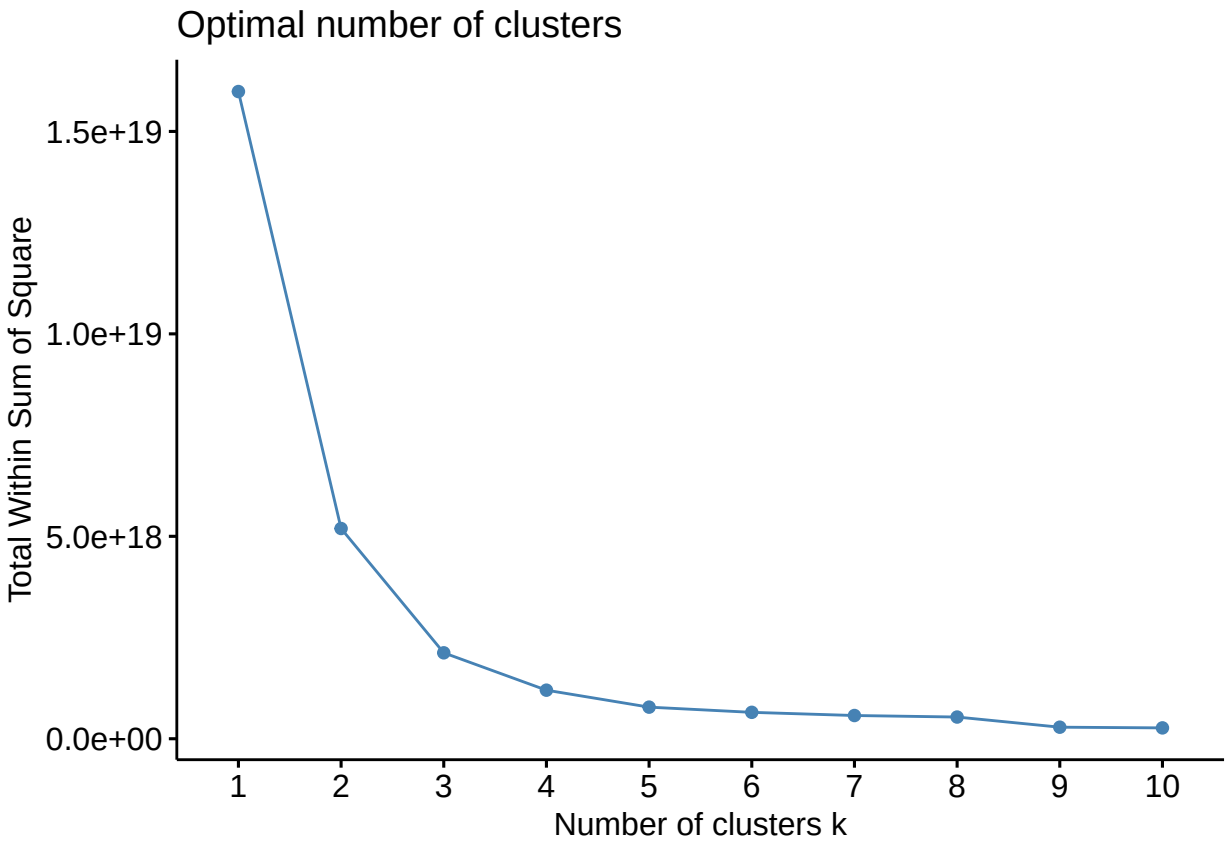
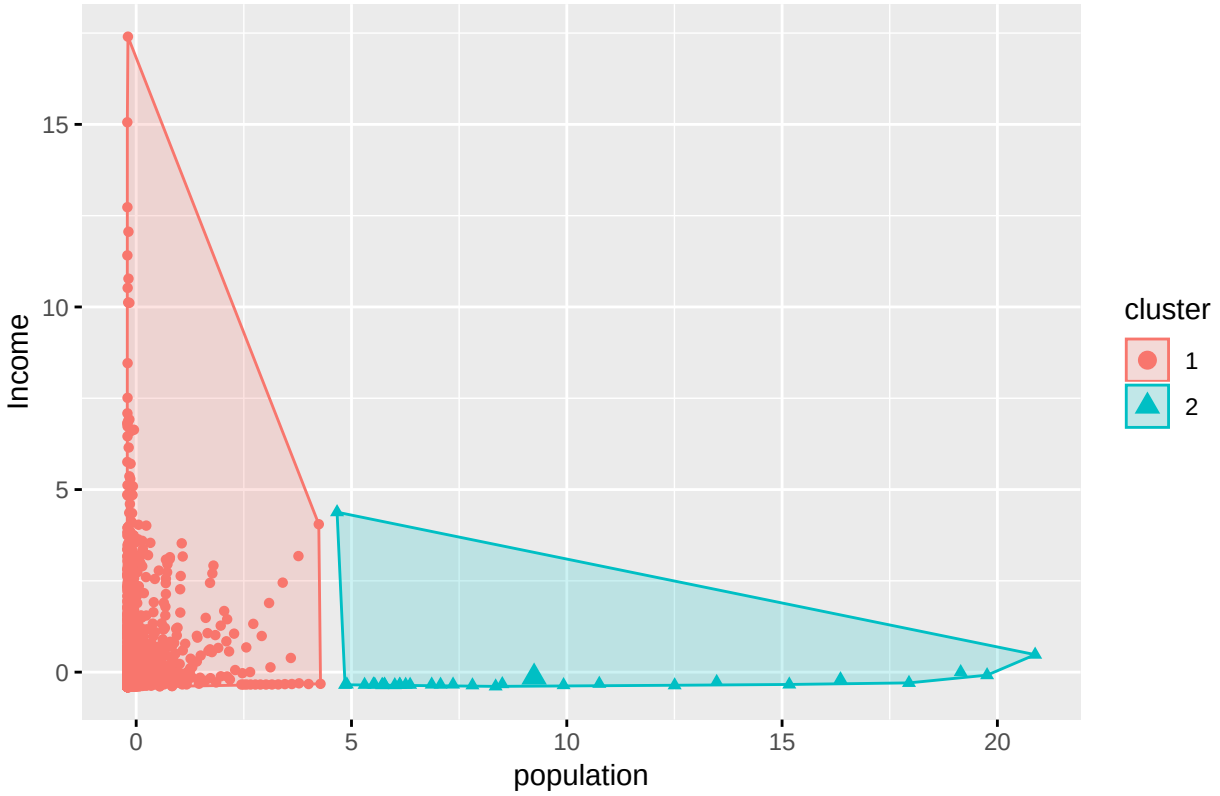

Optimal number of clusters
Total Within Sum of Square
1.5e+19
1.0e+19
5.0e+18
0.0e+00
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Fig. 8 − Cluster with 2 centers



#Summary

Based on the above analysis, it can be concluded that countries with high population tend to have lower income as well as lower life expectancy.