Final Report - ANLY 506-51- B-2019/Late Spring - Exploratory Data Analytics

Rajib Kar

16 Jun 2019

Introduction

This paper is to showcase multiple exploratory data analysis results on the demographic data of the countries across the world, by applying different techniques learned in the class of ANLY 506-51-B-2019/ Late Spring - Exploratory Data Analytics. The following exercise will be undertaken:

- **Studying the changes in economic and health conditions of all global regions for every 50 years from year 1800 to 2000. Two metrics will be used: 1) GDP Per Capita income of each region, and 2) Life Expectancy of each region.**
- **Studying relationship between income and life expectancy across all regions.**
- **Forming clusters using K-Means clustering on the basis of GDP per capita and Life expectancy of each countries in region America for the years 1800, 1900 and 2000, and made the comparisons.**

Description of the data

The "Gapminder" dataset represents the demographic data of several countries across different global regions since year 1800. It consists of **41284 observations** and **6 variables**: **Country, Year, life, population, income**, and **region** with data types **character, double, double, double, double**, and **character** respectively.

The variables represent the following demographic metrics respectively: **life** (Life Expectancy), **income** (GDP Per Capita), **Year** (Year from 1800 to 2015), **County** (Name of the countries: there are total 197 countries), **region** (Name of the global regions: there are total 6 regions), **population** (Census data collected about every 10 years).

Pre-processing and Exploring the data

The dataset doesn't have population data from year 1800 to 1949 for two countries: **Taiwan** and **Netherlands Antilles**. **Taiwan** doesn't even have population data for year 2014 and 2015. Otherwise in the dataset each remaining country has census data for every 10th

year starting from year 1800 till 1950, and for every year from year 1951 to 2015. So for simplicity observations are dropped for **Taiwan** and **Netherlands Antilles**. For remaining **195** countries, from year 1800 to 1950 for each year where the population data is missing, data from the most recent previous census year will be used for imputation i.e census data from year 1800 will be used to impute the missing data of population of years between 1801 and 1809, census data from year 1810 will be used to impute the missing data of population of years between 1811 and 1819, and so on so forth.

Similarly the dataset has **0 observations** with missing income values for **14 Countries** and **5 regions**. However since this paper involves with analysis limited only to certain years i.e. 1800, 1850, 1900, 1950, and 2000 the missing values of income impacts only **53 observations**. Hence all observations with missing income value is dropped.

Insights

In order to get an insight of how the global regions have been changing in terms of their economy the metric **GDP per capita income** is considered. A region with better economy is considered to have a better mean income. Box-plots for GDP per capita income of each Global region for every 50 years from year 1800 to 2000 are drawn to visualise this. But as shown in Fig 1. the chart is not readable due to wide spread of income across different regions, from regions where economy is pretty strong to regions where economy is pretty weak.

To avoid this, the same chart is redrawn using Log10 scale for GDP per capita income as shown in Fig2. From this new chart it is clearly visible that, overall, all the regions are moving towards a higher GDP per capita income over the two centuries, but at a much faster rate during the 19th century. However region like South Asia and Sub-Sahara Africa are lagging far behind compared to the rest. Again between South Asia and Sub-Sahara Africa it is noticeable that, only one time, during year 1950, Sub-Sahara Africa region surpassed the South Asia region in terms of median GDP per capita income. Another observation from this chart is that the income-gap between poor and riches are gradually widening over the period of time. Also the rapid economical growth of Middle East and North Africa region, mostly riding on the oil boom between year 1950 and 2000, is quite easily visible.

Another parameter of prosperity of a region is its health condition which is directly linked with the Life expectancy of its people, i.e. the developed countries have higher median life expectancy than that of the developing countries. Fig 3. shows box plots of Life expectancy of global regions for every 50 years from year 1800 to 2000. It is obvious from this chart that life expectancy of the regions were almost stagnant till 19th century, after which due to the breakthroughs in medical science, it improves dramatically from year 1950 to 2000. Another interesting observation between year 1950 and 2000 is that the developed regions are not only pushing their median life expectancy upwards, they are also being able to reduce the IQR (inter quartile range) of their life expectancy metric, mostly by making better social, economical, and medical benefits available towards their people.

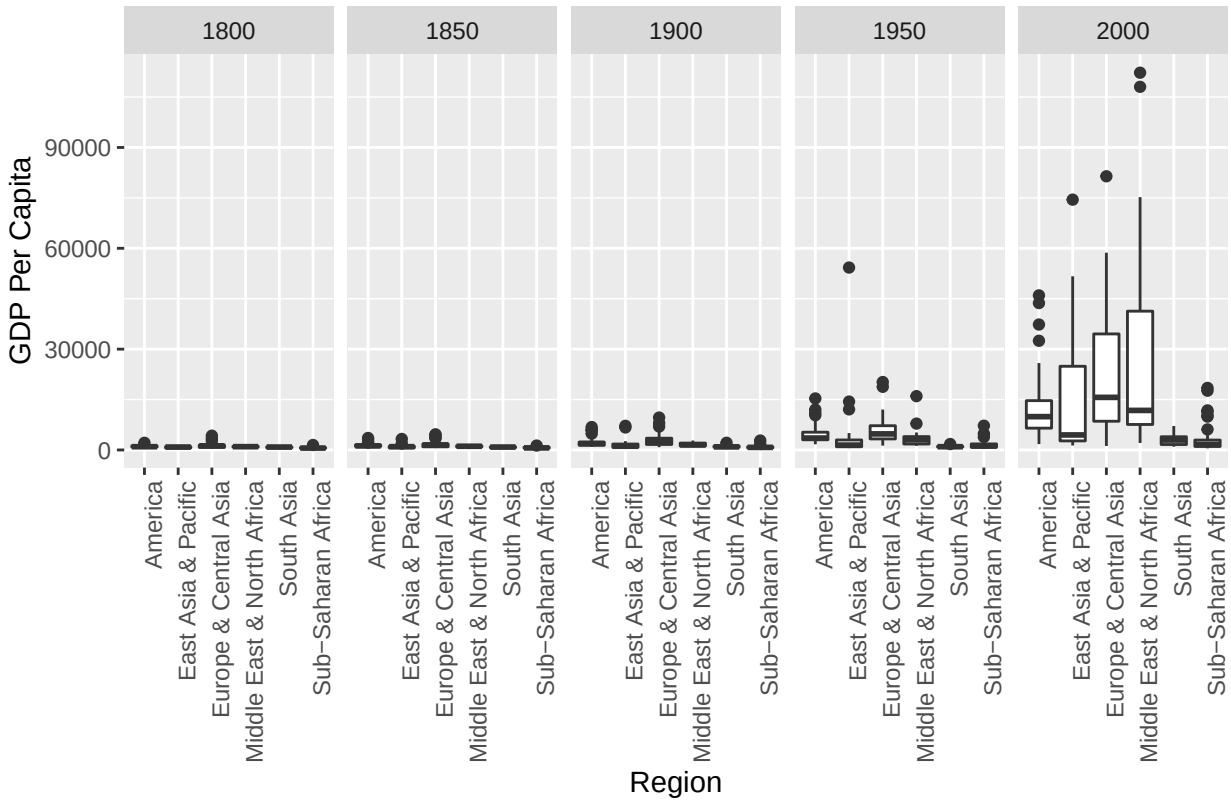

Fig 1: Box Plot of GDP Per Capita of Regions

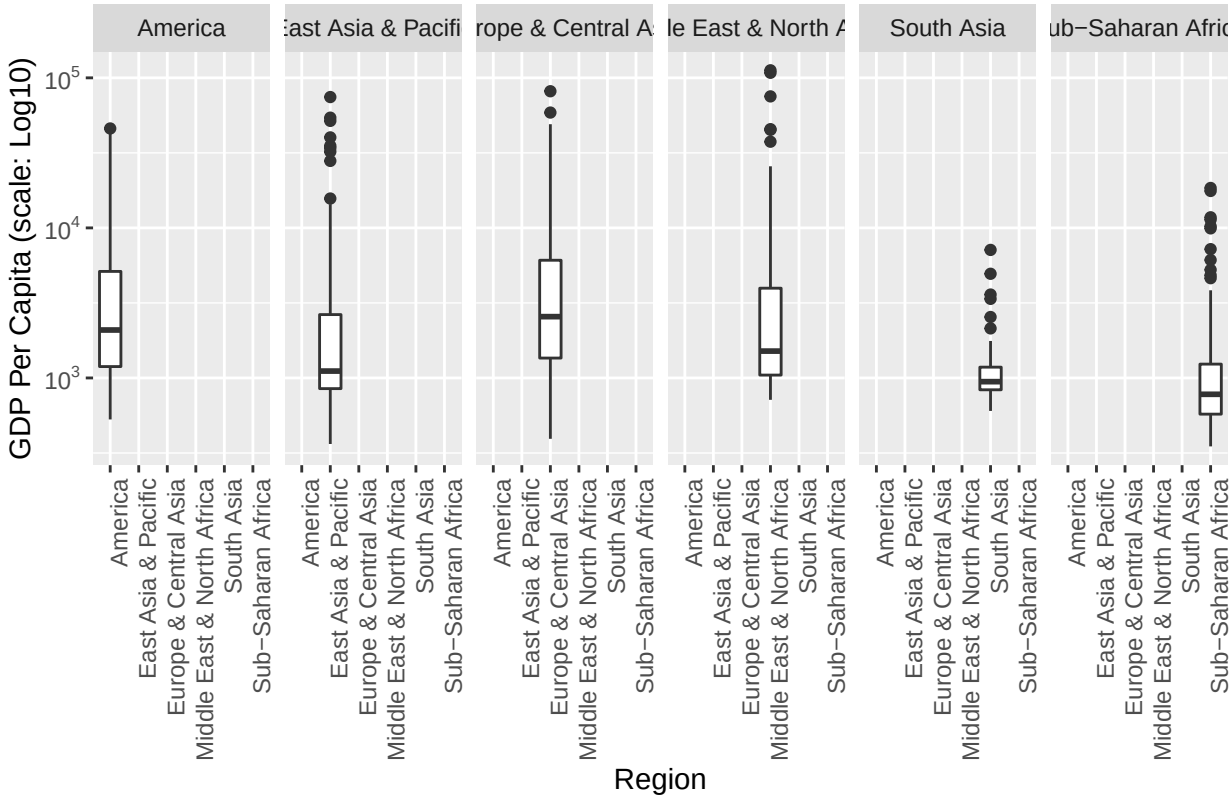

Fig 2: Box Plot of GDP Per Capita (scale:Log10) of Regions

Since like income of regions in Fig 2., the life expectancy of regions in Fig 3. has a similar upward trend over the years while comparing between developed and developing regions, a scatter plot is attempted as shown in Fig 4. to assess the behavior of income against life expectancy for different regions across every 50 years between year 1800 and 2000.The scatter plot clearly shows a positive correlation between income and life expectancy, i.e. higher the income better is the life expectancy.

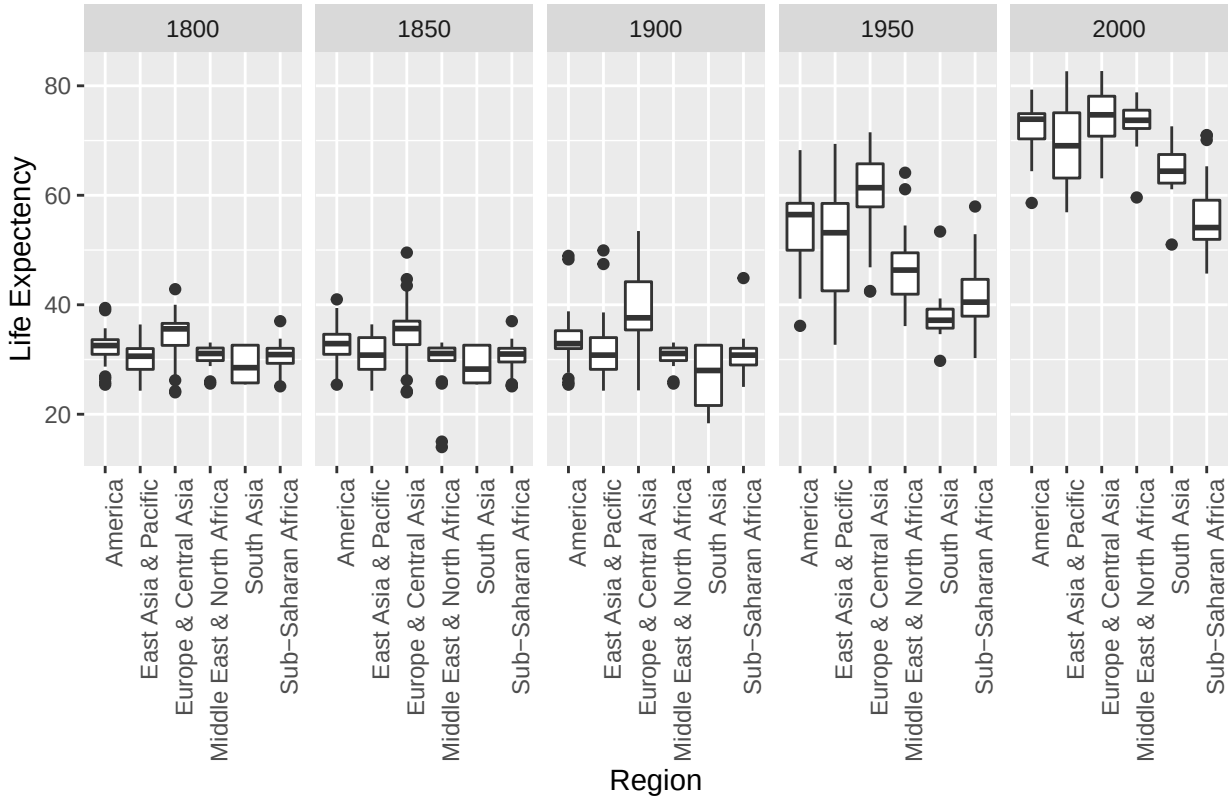

Fig 3: Box Plot of Life Expectancy of Regions

From Fig 4. it is apparent that the more the dots of the scatter plot are moving towards the right hand top corner of the chart over the period of years the more prosperous the corresponding regions are becoming. From this chart, as expected, it is clearly visible that regions like America, Middle East and North Africa, Europe and Central Asia, East Asia and Pacific are prospering at a much faster rate than that of the region of South Asia and Sub-Saharan Africa region.

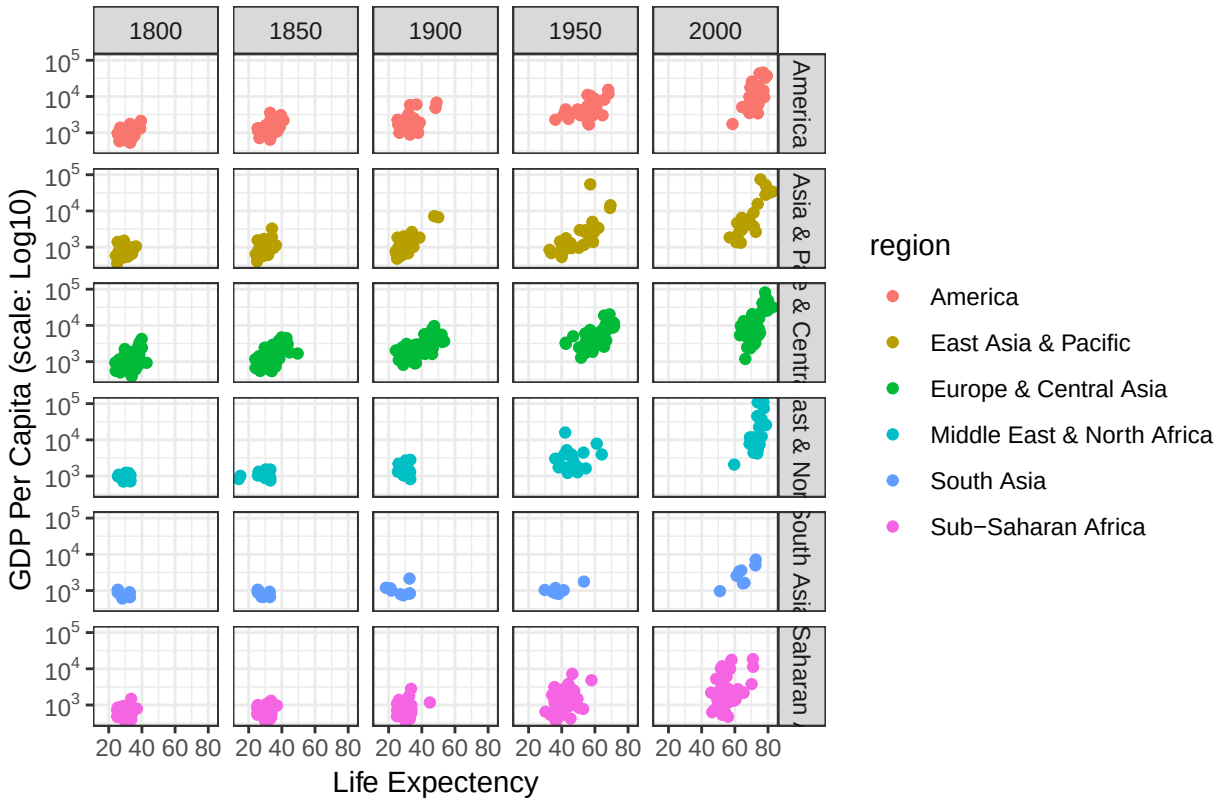

Fig 4. Income vs Life Expectancy of Regions

Finally an attempt is made to form clusters using K-Means clustering technique on the basis of GDP per capita income and Life expectancy of the countries falling under the region America, for the year 1800, 1900 and 2000 respectively. Clustering optimization technique suggests that the optimum number of clusters i.e. k value should be 3, 2 and 3 for year 1800, 1900 and 2000 respectively as shown in Fig 5.

The clusters are formed for year 1800, 1900, and 2000 as shown in Fig 6, Fig 7, and Fig 8 respectively. It is visible from these clustering charts that USA and Canada remain in the best cluster where both income and life expectency both are above average. While USA remain at top, Canada gradually reduces the gap over the years. On the otherhand countries like Uruguay and Argentina keep pace with the other countries in terms of economic growth, but gradually falling short interms of life expectancy, especially between year 1900 and 2000.

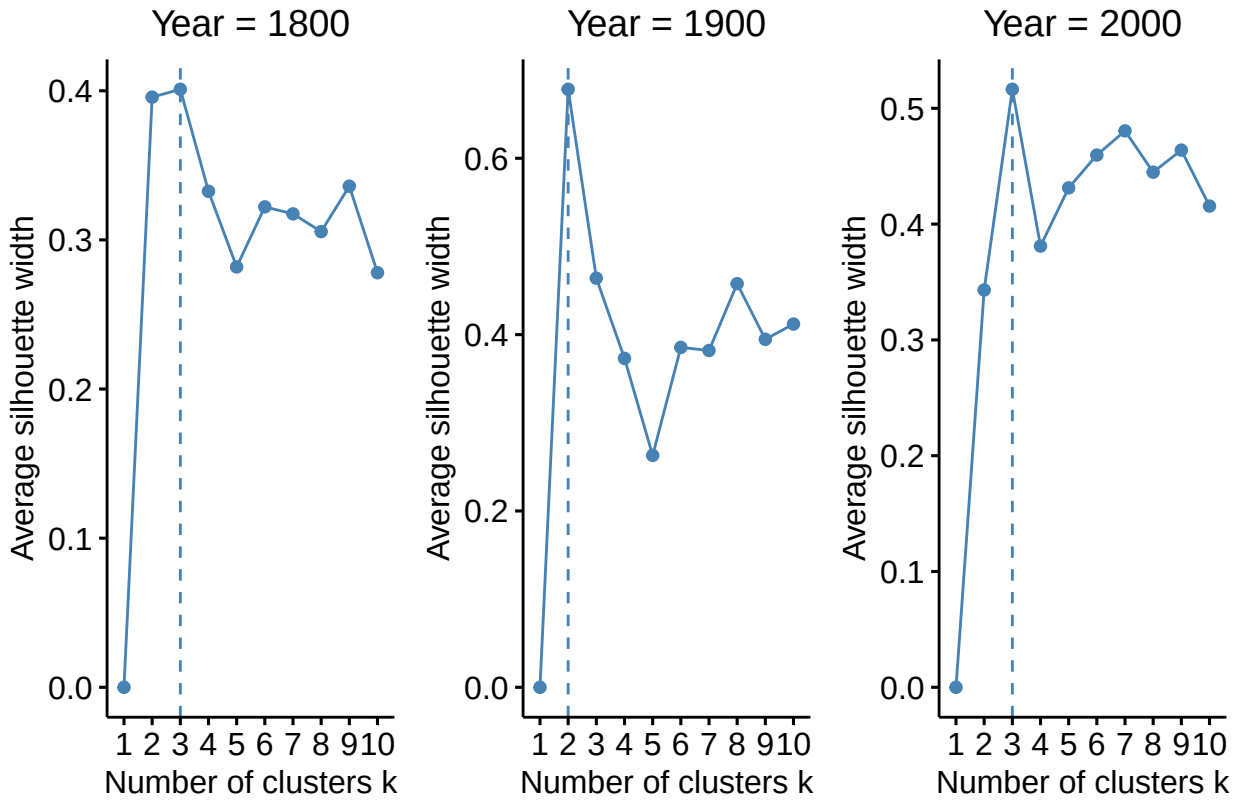

Fig 5. Optimization of k (Region = America)

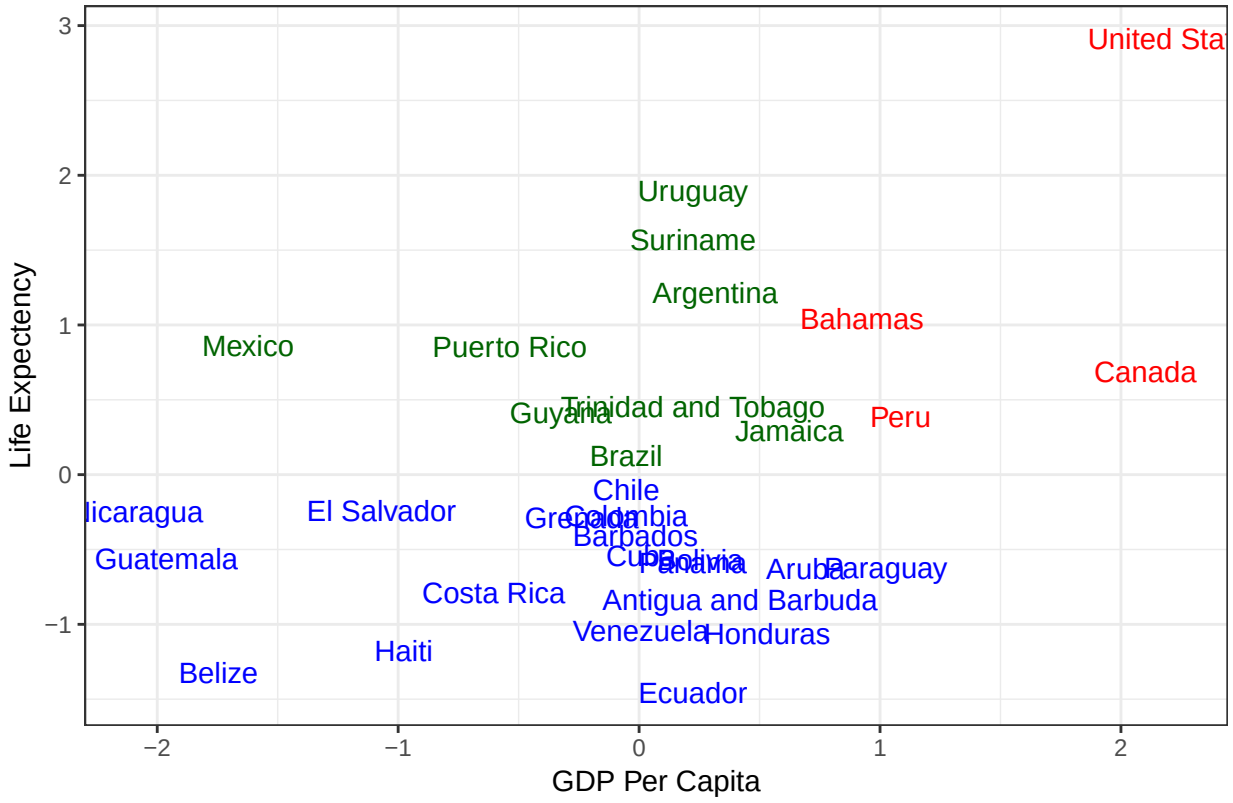

Fig 6: K−Means Clustering (Region = America, Year = 1800, k = 3)

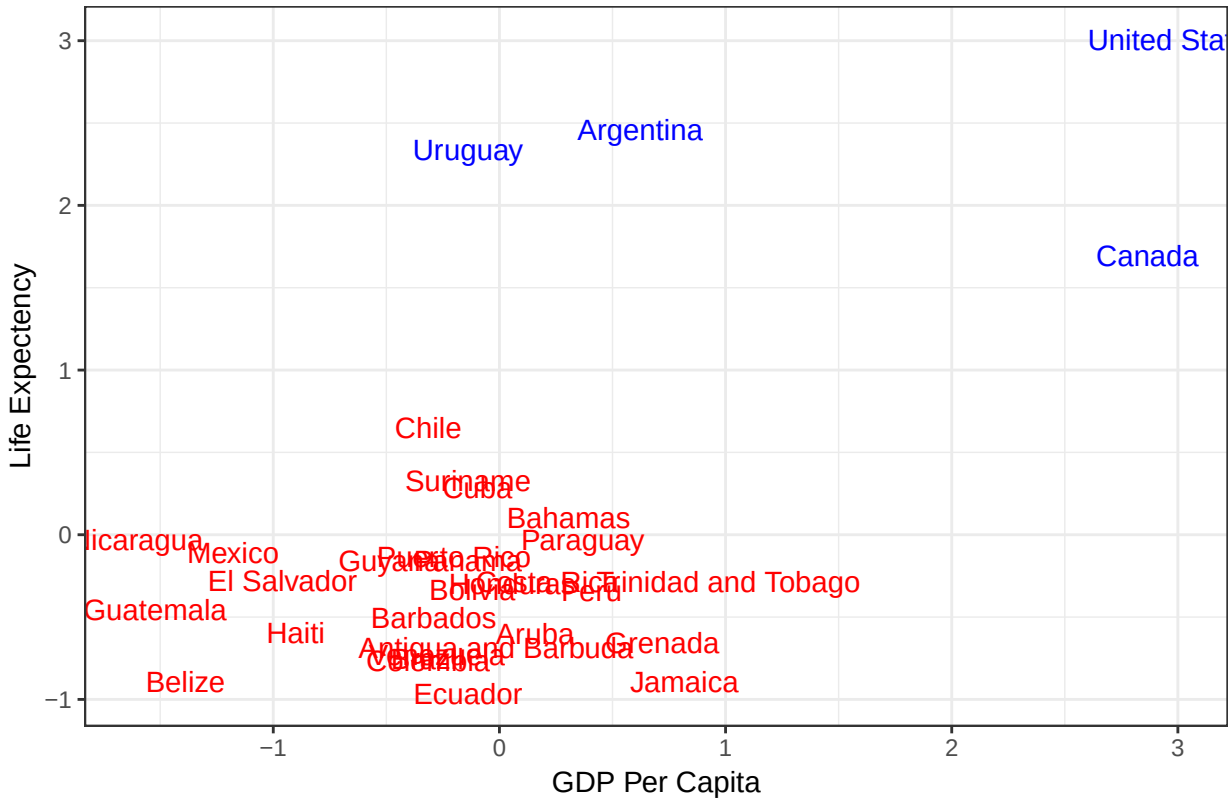

Fig 7: K−Means Clustering (Region = America, Year = 1900, k = 2)

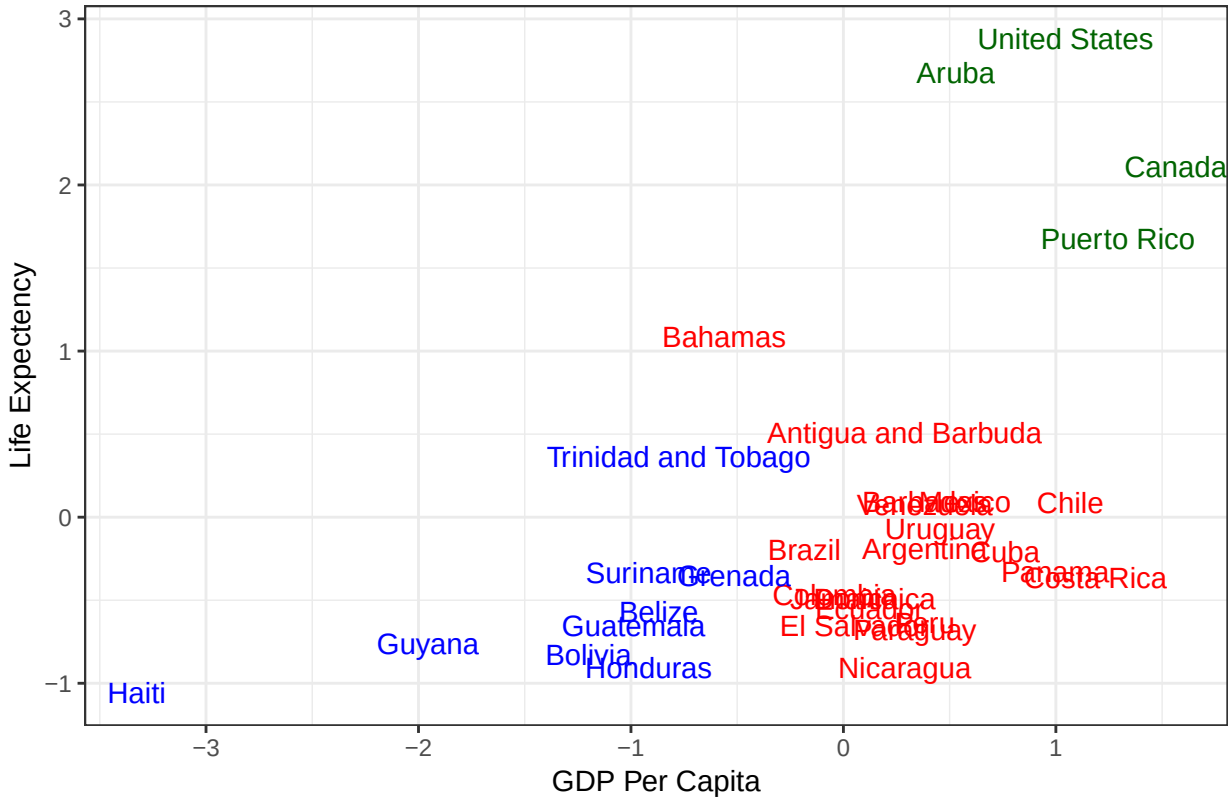

Fig 8: K−Means Clustering (Region = America, Year = 2000, k=3)

Conclusion

Over the period of two centuries between year 1800 and 2000, the development of the region or countries can be easily visualise from the changes of their GDP per capita income and life expectancy. In general higher the income better is the life expectancy. Also over the period of time the gap between wealthy and poors have been widened gradually.

Reference

https://uc-r.github.io/kmeans_clustering)

https://rafalab.github.io/dsbook/

http://r-statistics.co/Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html