EDA Final Project

Ankur Gupta

6/12/2019

Narrative summary of the data

The data set contains observations on the population and income of people of different continets. On further analysis, we can identify the relation of regions and income on the life expectancy.

Following are the variables in the data set:

Life = life expectancy (type - Numeric) Income = gdp per capita (type - Integer) Year = The dataset contains range from 1800 to 2015. (type - Integer) Country = Contries of the world. (type - Factor) Region = 6 continents. (type - Factor) Population - census data collected about every 10 years. (type - Factor)

A close look at the dataset tells us that the income census was done every 10 years till 1950 and after 1950, the population census is done every year.

There are `6` variables in the data set. There are `4208` of observations in the data set.

Checking for missing values

The original data set has `**2341**` missing values.

I have modified the data set so that it contains the observations evey 10 years. This modified data has `**240**` missing values in Income.

From here on, I will be using the modified data set. As a first step, I have removed the missing values for further abalysis

```
## ï..Country       Year       life population     income     region
##           0          0          0          0        240          0
```

```
##    Country       Year       life population     income     region
##           0          0          0          0       2341          0
```

From this, we can see that there are 2341 values missing in the income column.

There are values that might seem missing from the Population column as well, but on closer observation, we can deduce that before 1950, the population census was done every 10 years. It was only after 1950 that the census was done annually.

Questions to answer

By analyzing this data set, we can derive many conclusions. I will focus on answer the following questions:

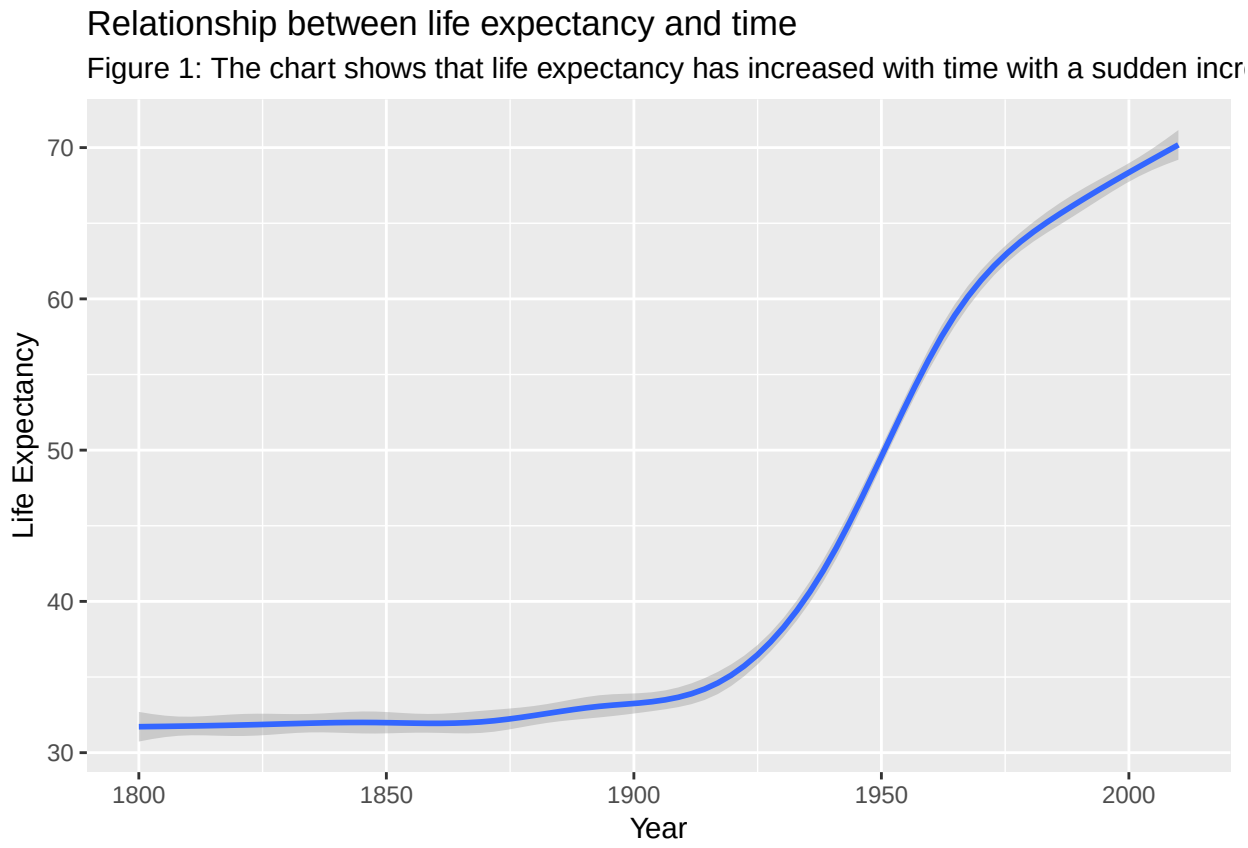
1. What is the mean life expectancy for each continent?
2. Is there a time period with sudden increase in life expectancy for various regions?
3. Is there a relation between the and life expectancy for regions?

Exploratory Data Analysis

Before answering the questions, I am doing an Exploratory Data Analysis on the data set to identify trends and relationships. This would help us make sense of the data in a better way.

The following Figure 1: tells us about the evolution of life over the years. We can see that there is a sudden spike between 1925 and 1950. This is a thing that should be analyzed in detail.

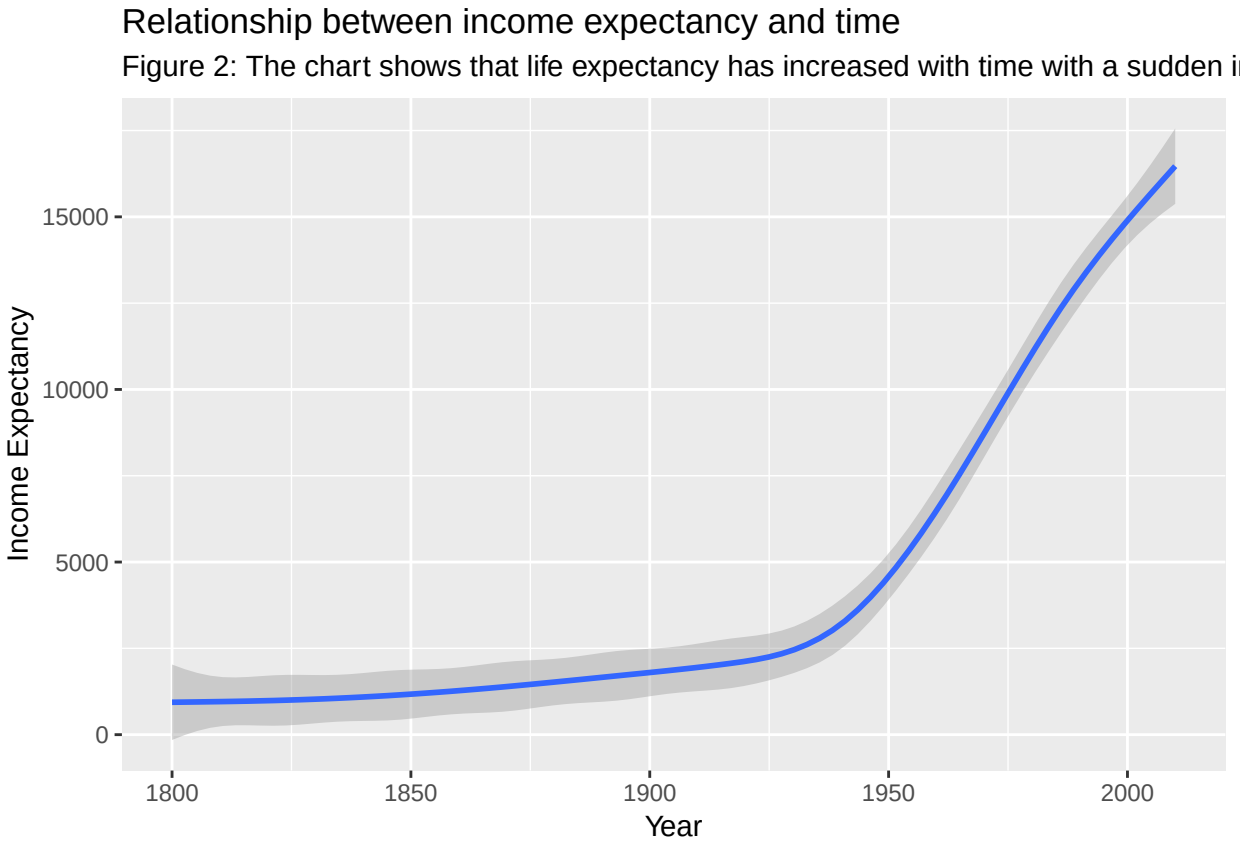
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



For further analysis, I am seeking to identify the relation between income and time to see if it also follows the same trend.

Figure 2 tells us about the evolution of life over the years. We can see that there is a sudden spike between 1925 and 1950. This is a thing that should be analyzed in detail.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



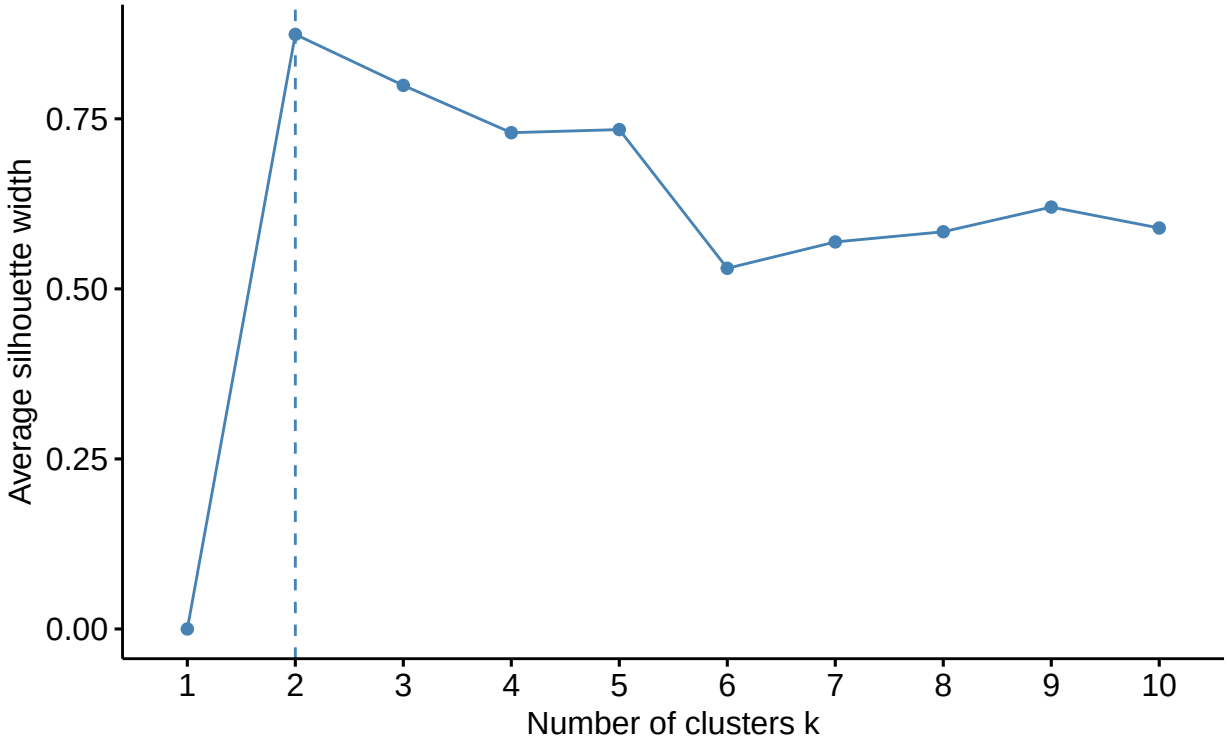
From figure2, we can see that the income curve closely resembles the life expectancy curve. This can point towards the following conclusions: 1. The increase in disposable income has increased the access to medicare 2. Because of the increase in income, people might have more acess to healthcare facilities thus aiding them in living longer.

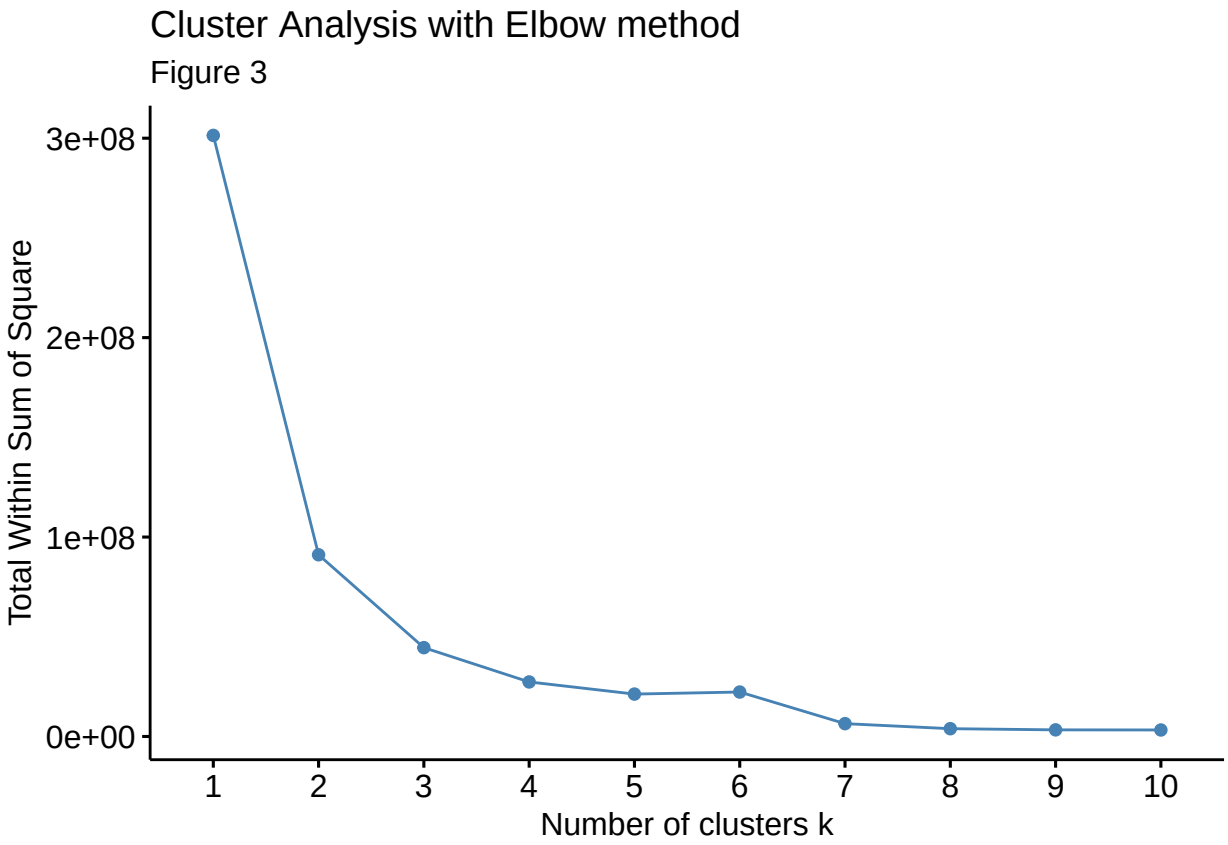
Clustering Analysis

I am only focussing on the analysis of just South Aian region and have sub setted my data.

Cluster Analysis with Silhoutte method

Figure 3



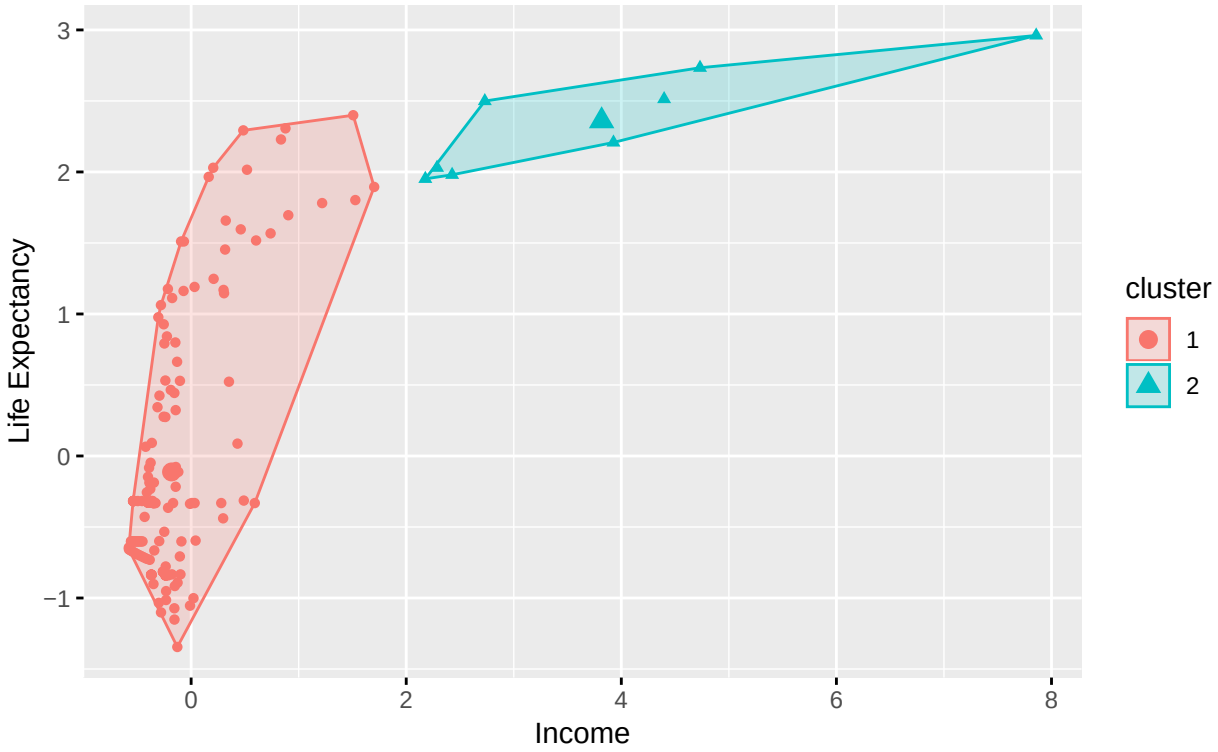


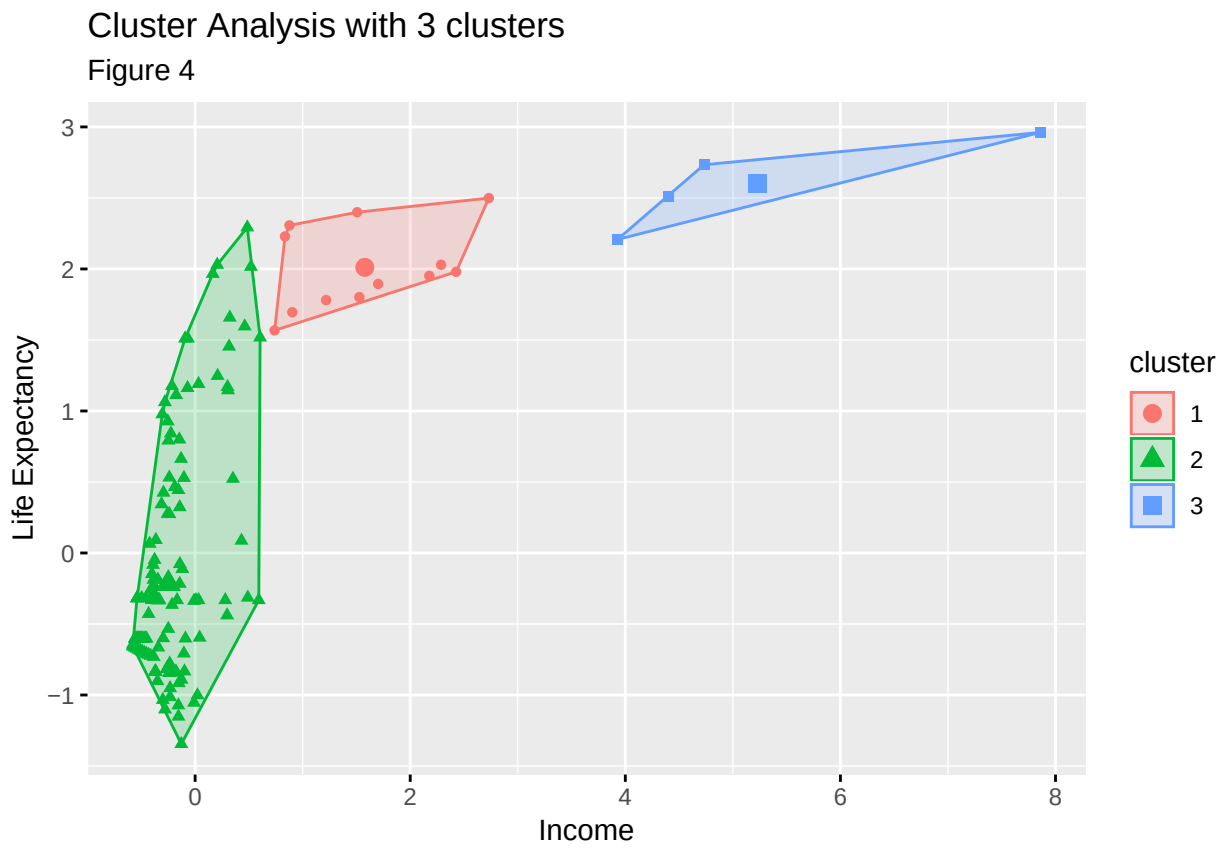
The above two figures tell us about finding the number of clusters using Silhouette Method and Elbow Method.

Once we have found the optimal number of clusters, we can do further analysis.

Cluster Analysis with 2 clusters

Figure 4





Through the above figures, we can seggregate the data into different clusters.

Answering the Questions

1. Does the life expectancy change with regions?

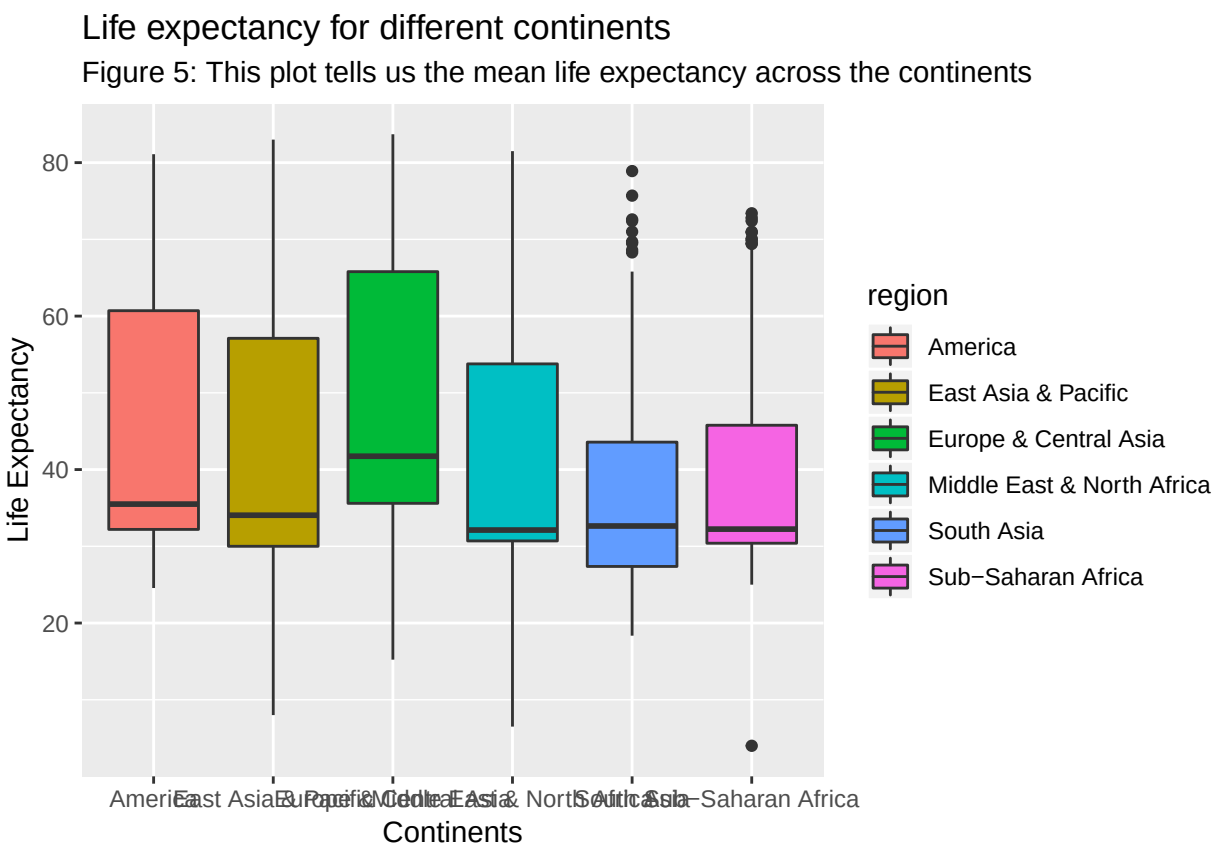


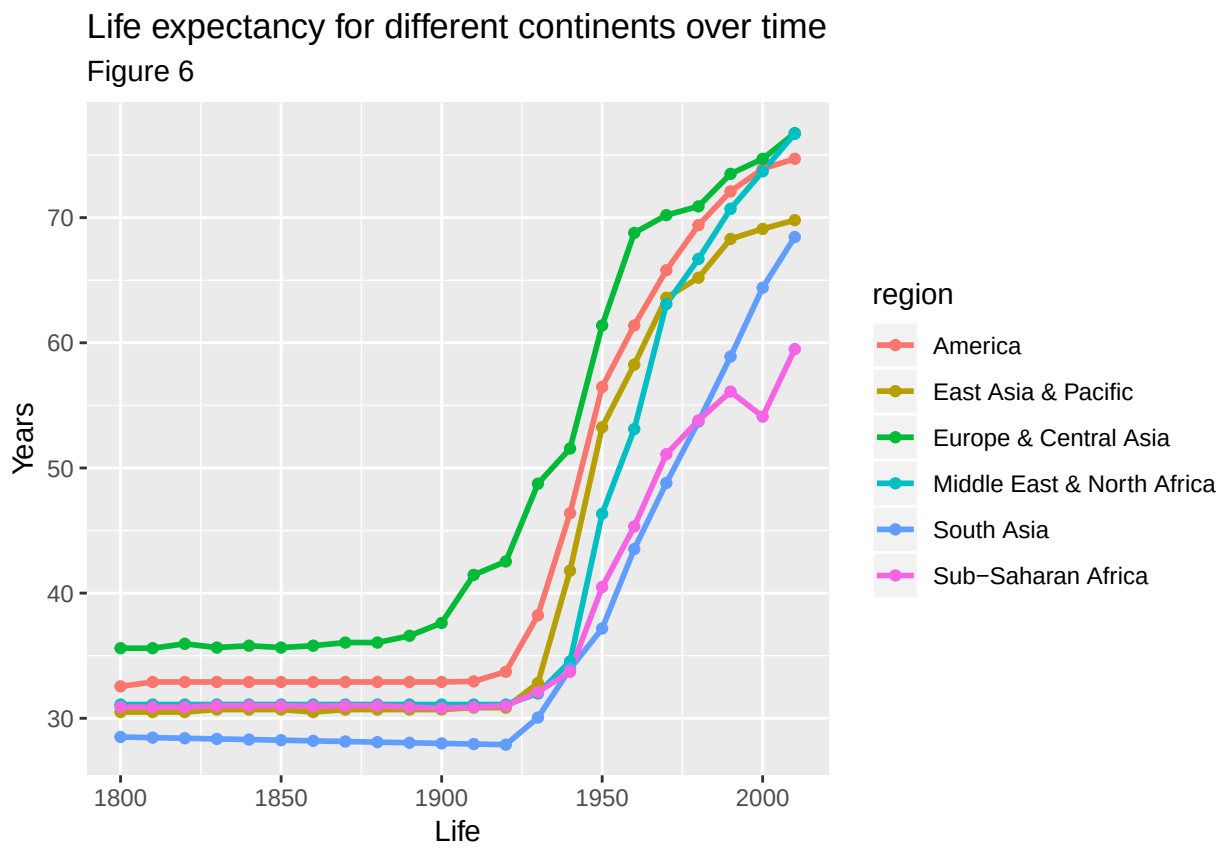
Figure 5 tells us that the median life Expectancy is highest for Europe and Central Asia over the years. This means that the people of this region tend to live longer than peopel from other regions. This can be attributed to following subjective factors:

1. Better access to healthcare
2. Better living conditions and environment

It also tells that lowest life expectanct is from Sub-Saharan Africa region.

Also, quartile range provides more color in the life expectancy. Rather than just focussing on the median, we can see the range of life expectancy in the regions.

2. Is there a time period with sudden increase in life expectancy for various regions?



From the Figure 1, we observed that there is a sudden increase in lif Expectancy. Figure 6 throws more color on to this and tells us the increase in life expectancy in different regions.

We can clearly see that expectancy has increased in all of the regions after 1925 but it has increased the most for Europe and Central Asia between 1900 and 1950.

3. Relationship between life expectancy and GDP

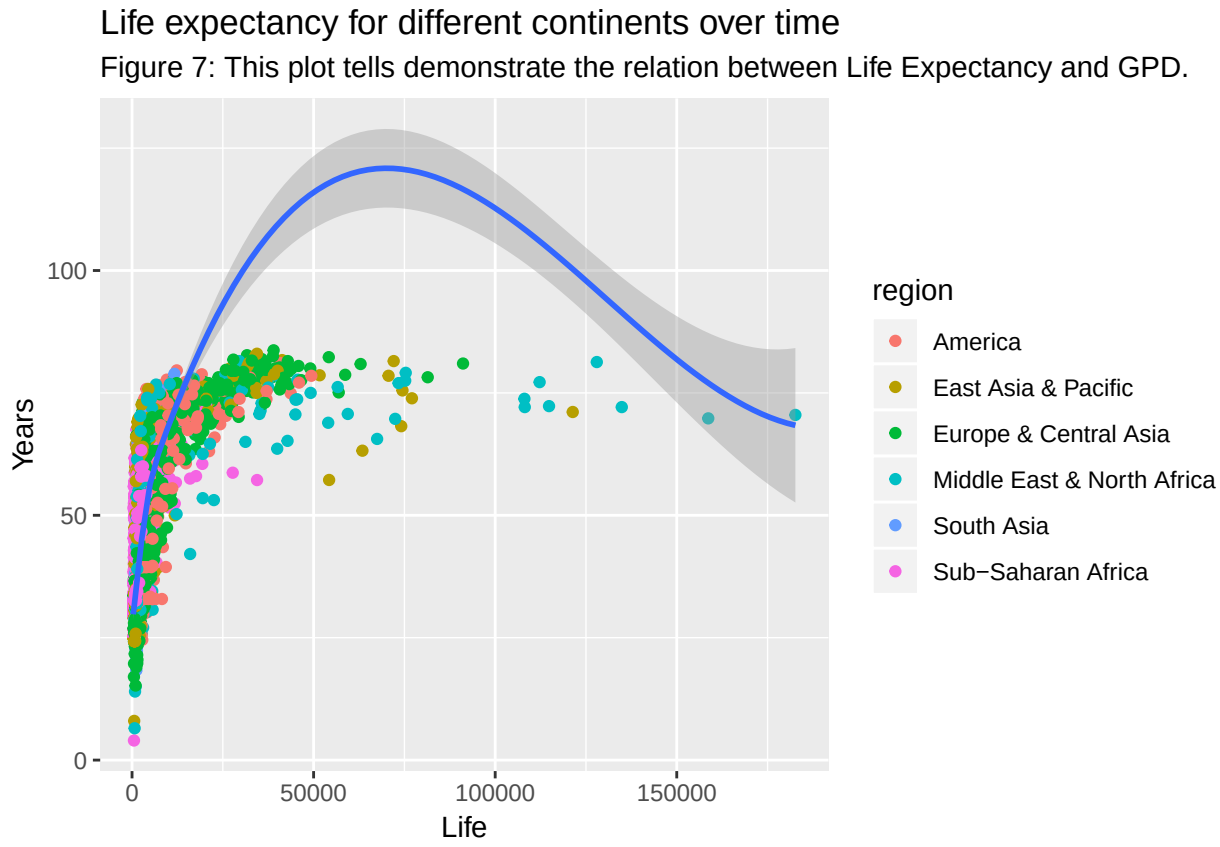


Figure 7: This plot tells demonstrate the relation between Life Expectancy and GPD.

From this Figure 7, we can see the relation between life Expectancy and Income. We see that the life expectancy is closely tied with income and increases with it. But after a point, income does not have much effect on life expectancy.

Relation between life expectancy and regions.

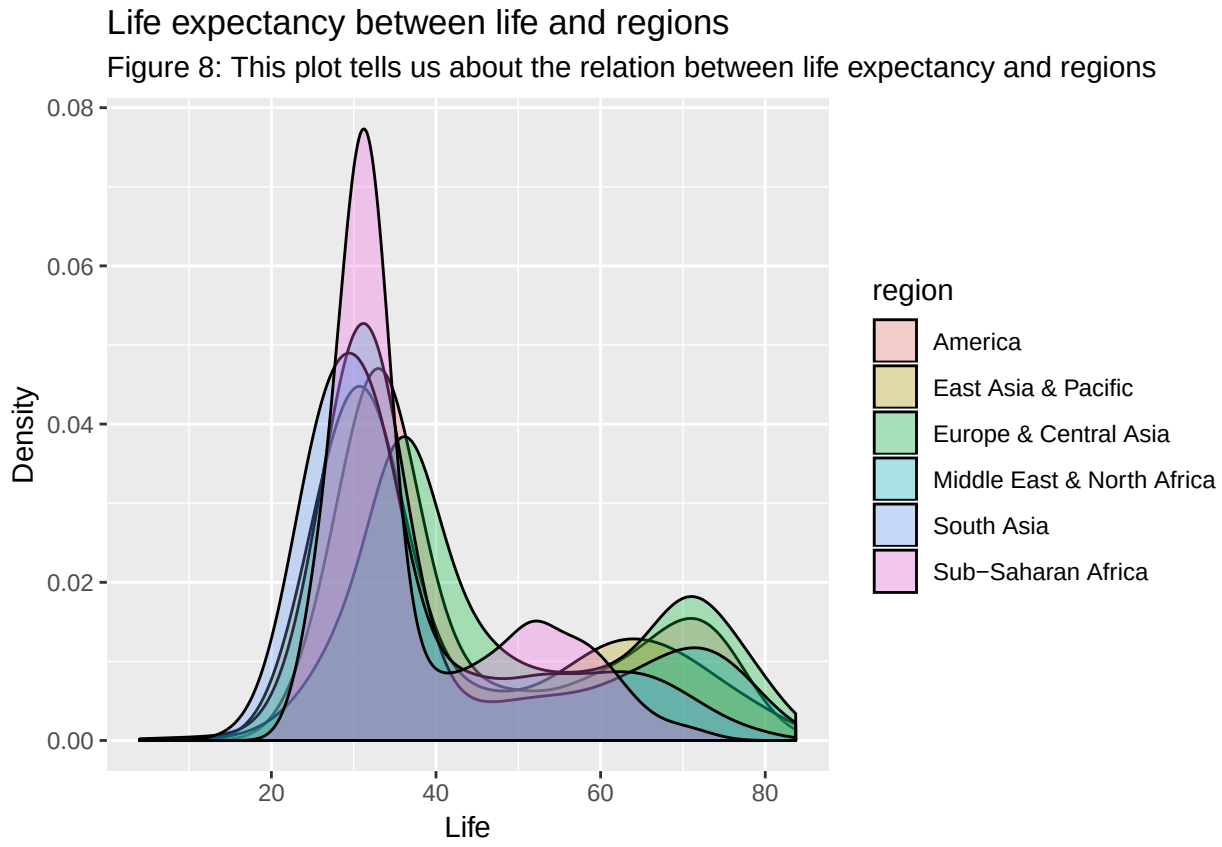


Figure 8: This plot tells us about the relation between life expectancy and regions

In conclusion, we can see that that life expectancy is directly related to income and is also related to years. We can clearly see a sudden spike in life expectancy after 1925 - 1950 for different regions.

This is attributed to the following:

1. Access to medical facilities
2. End of World Wars
3. Increase in dispoable incomes.