Final Project - ANLY 506

Harsha Uppala

6/16/2019

Project Questions

1) Which countries have the highest income?
2) How has the income changed over time?
3) Is income correlated with life expectancy?

Description of data

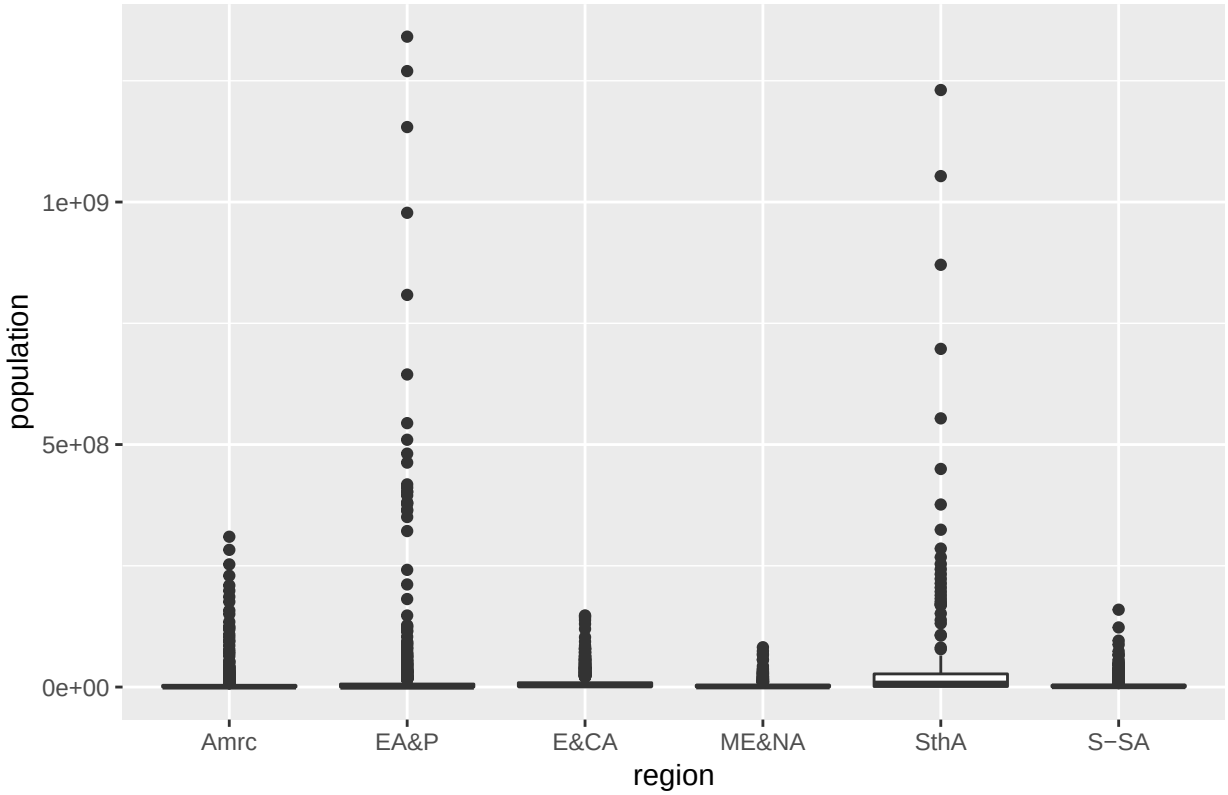
The dataset for this final project has 41284 rows and 6 variables. "Region", "Population" and "Country" are character variables."Year" and "Income" are interger variables. "Life" is a numeric variable and "Population" is a character variable. Therefore, population variable needs to be changed to a numeric variable.

Data modification

Since the population census was conducted every 10 years, there is data missing for population in about 25817 rows. Since, this data was never collected the null values can be dropped from the data set. The boxplot (Fig 1) shows the population distribution in various regions. The median life expectancy is 35 years and the median income value is 1419.5

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

Fig. 1 − Population Distribution by region



Highest Income Countries

After looking at the income in various countries, it is found that the top 3 countries in terms of income are United Arab Emirates (income of 182668), Qatar (income of 158673) and Brunei (income of 121356).

Data plots

Now, we can look at how the income in the top 3 countries has changes over time. UAE, Qatar and Brunei have all seen sharp increase and decrease in the income of the country. However, even at the lowest point, the income in these countries is about 35 times the median income of the dataset.For the most part, life expectancy is positive correlated with income for all these countries.

Fig. 2 − UAE income over time

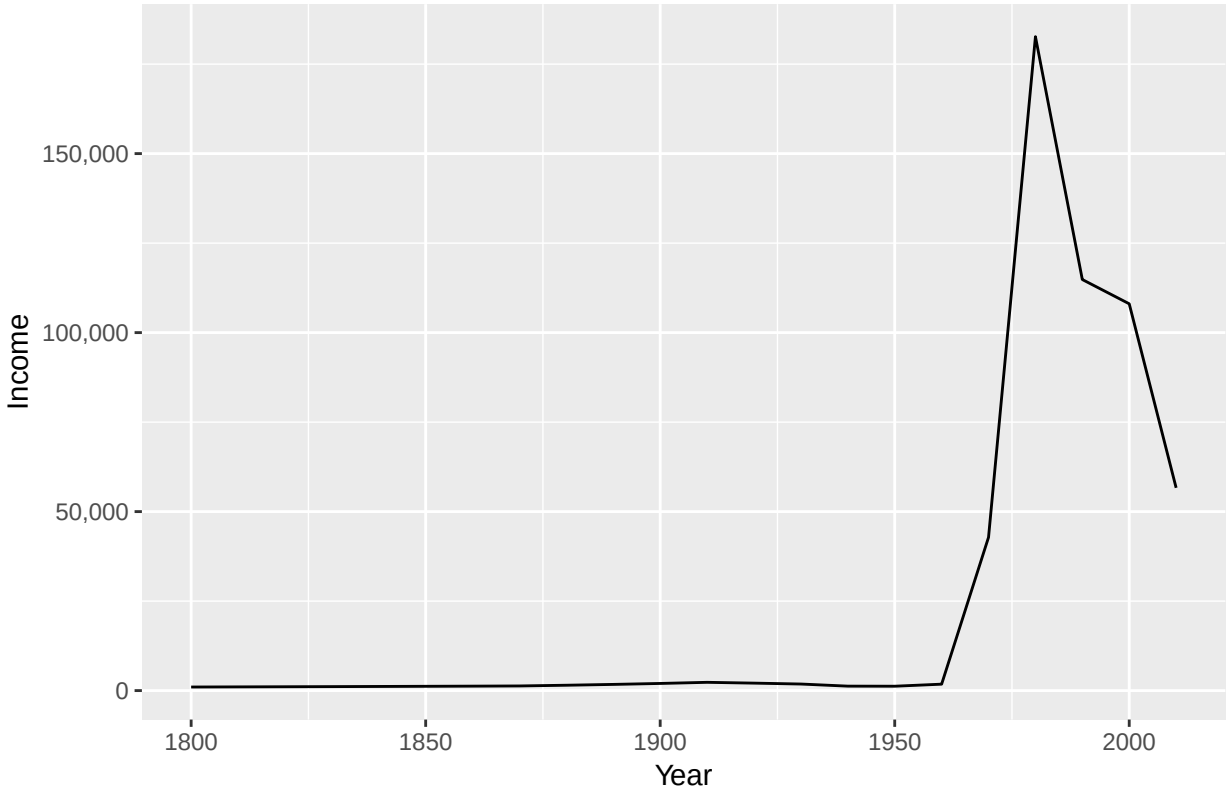


Fig. 3 − UAE life expectancy vs income

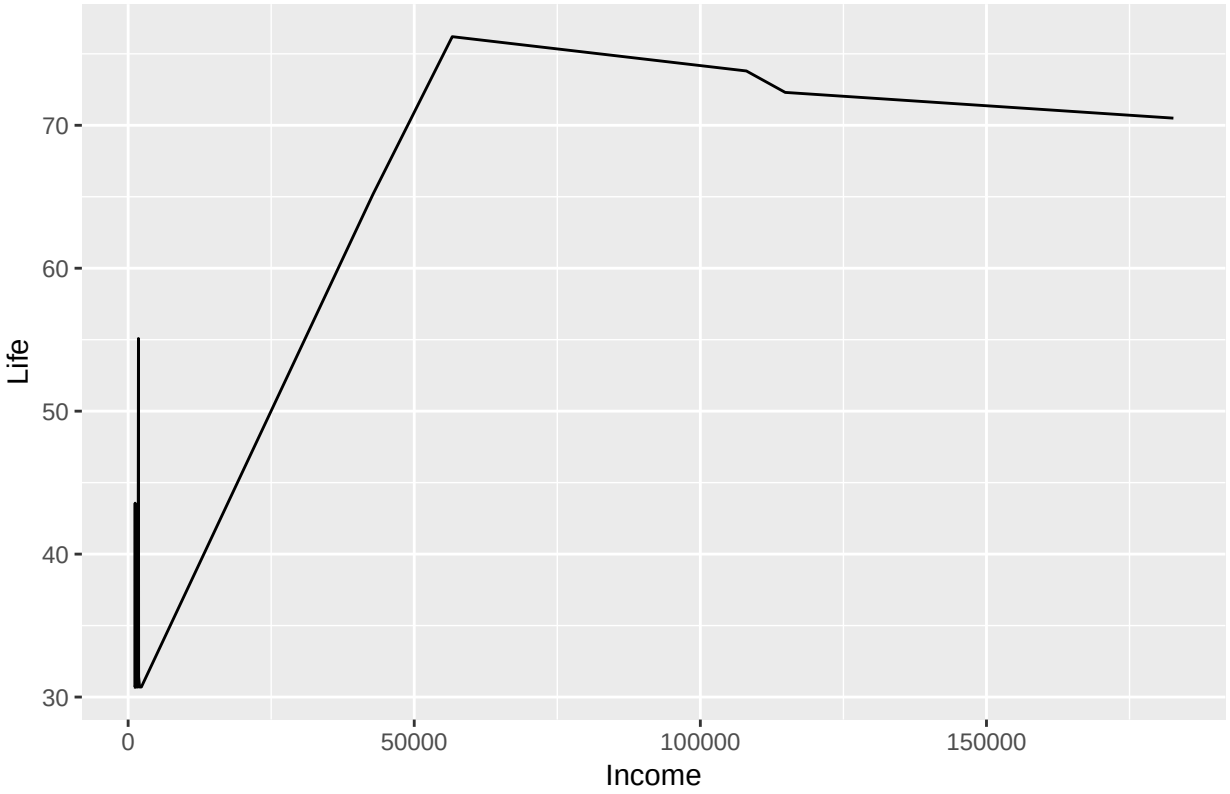


Fig. 4 − Qatar income over time

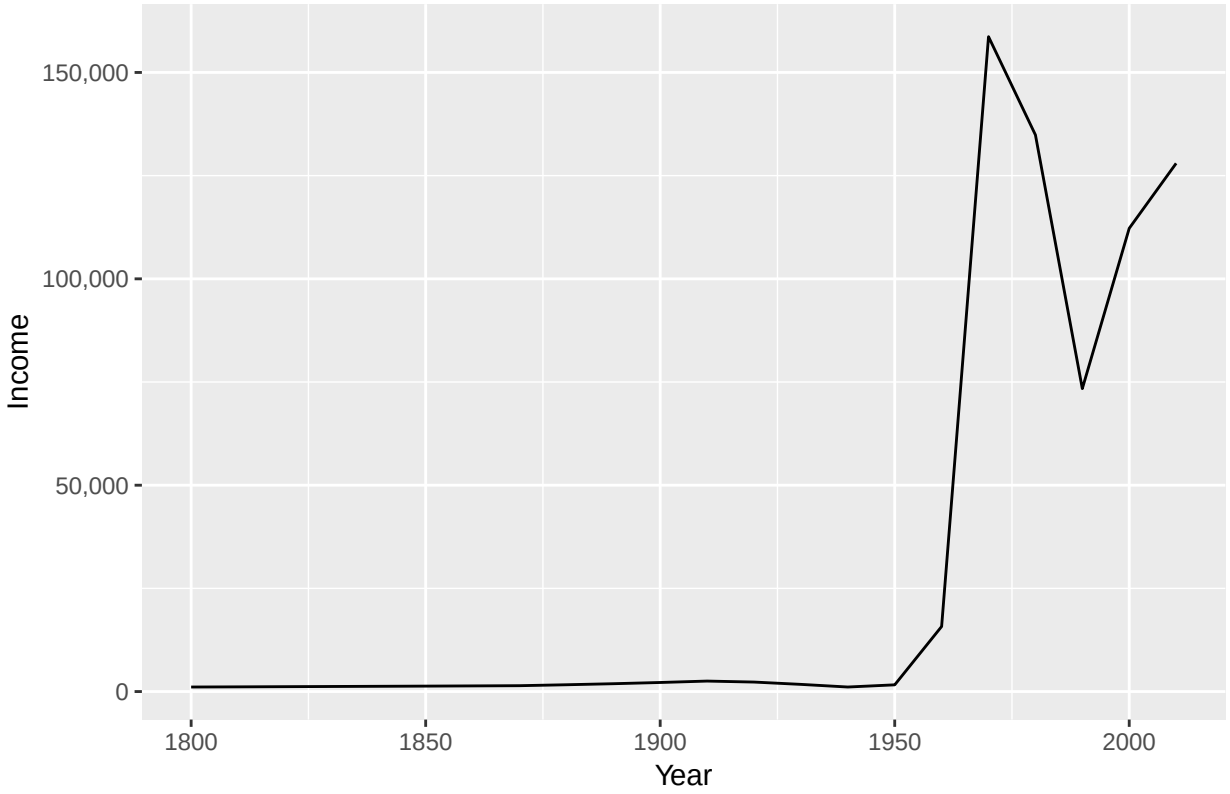


Fig. 5 − Qatar life expectancy vs income

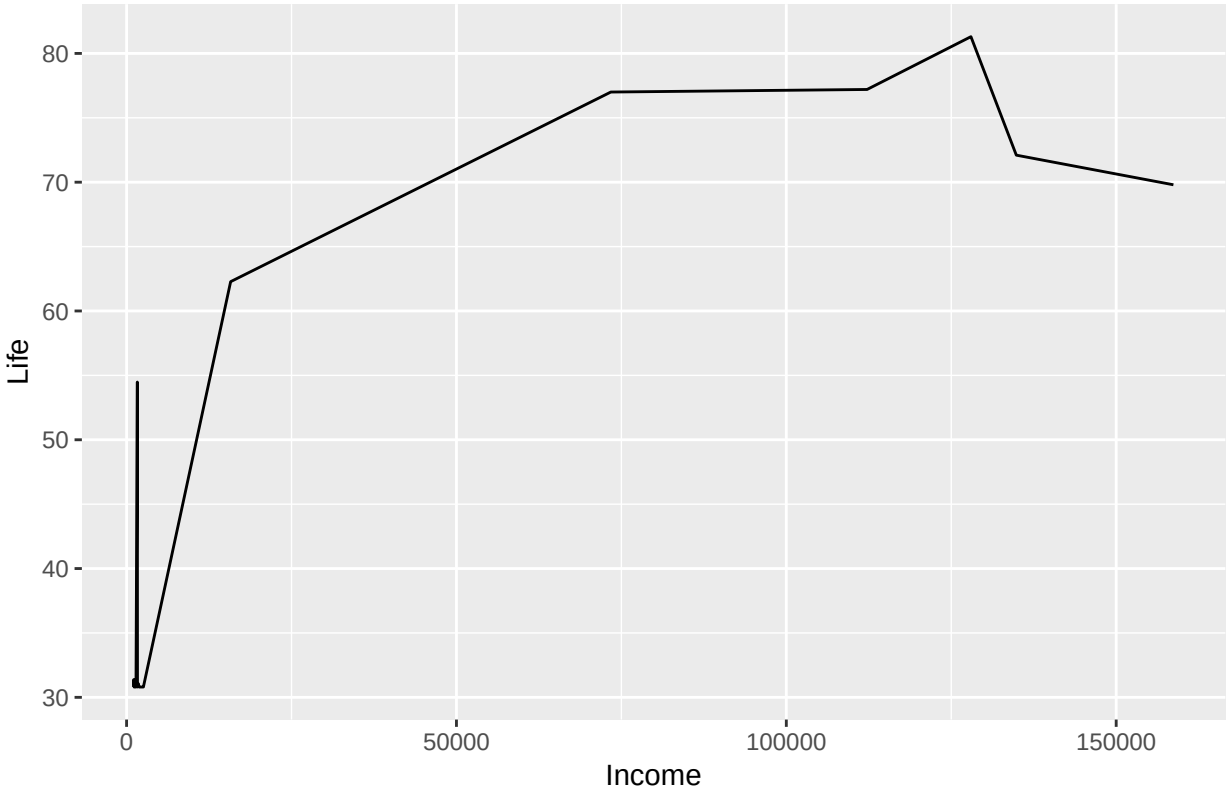
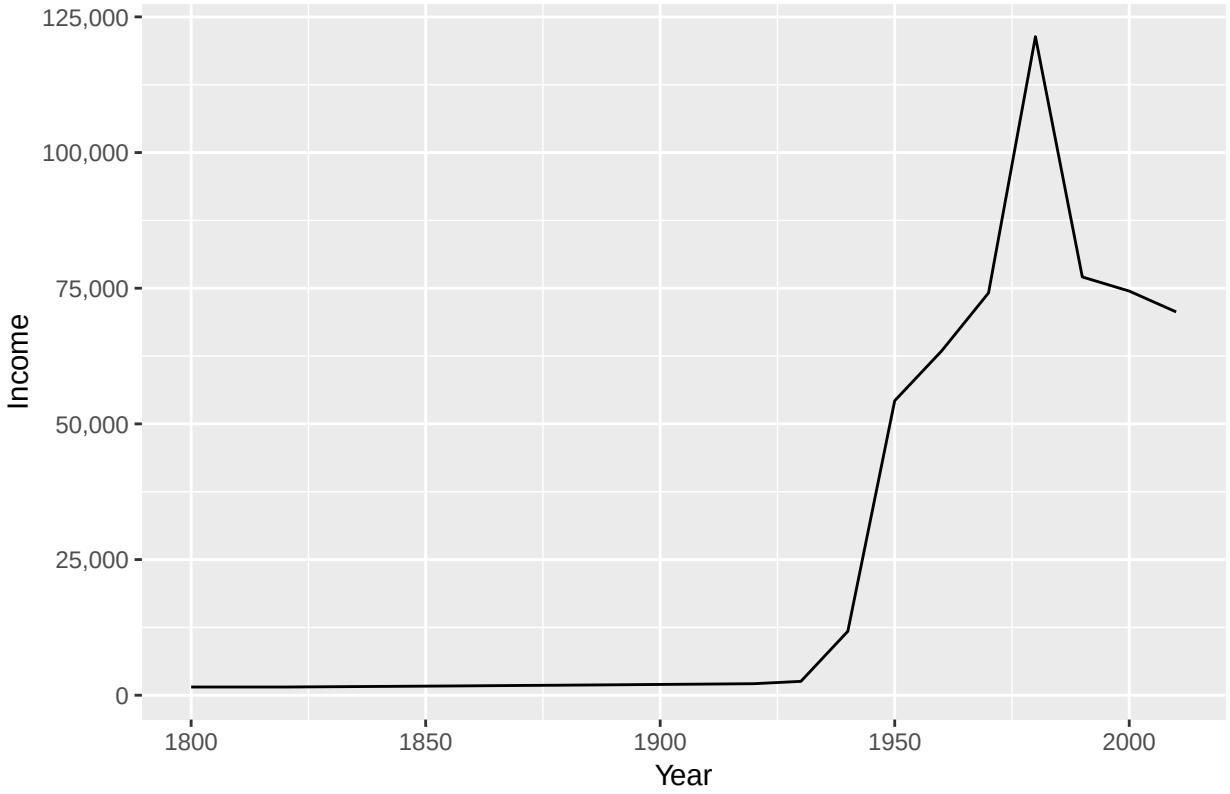
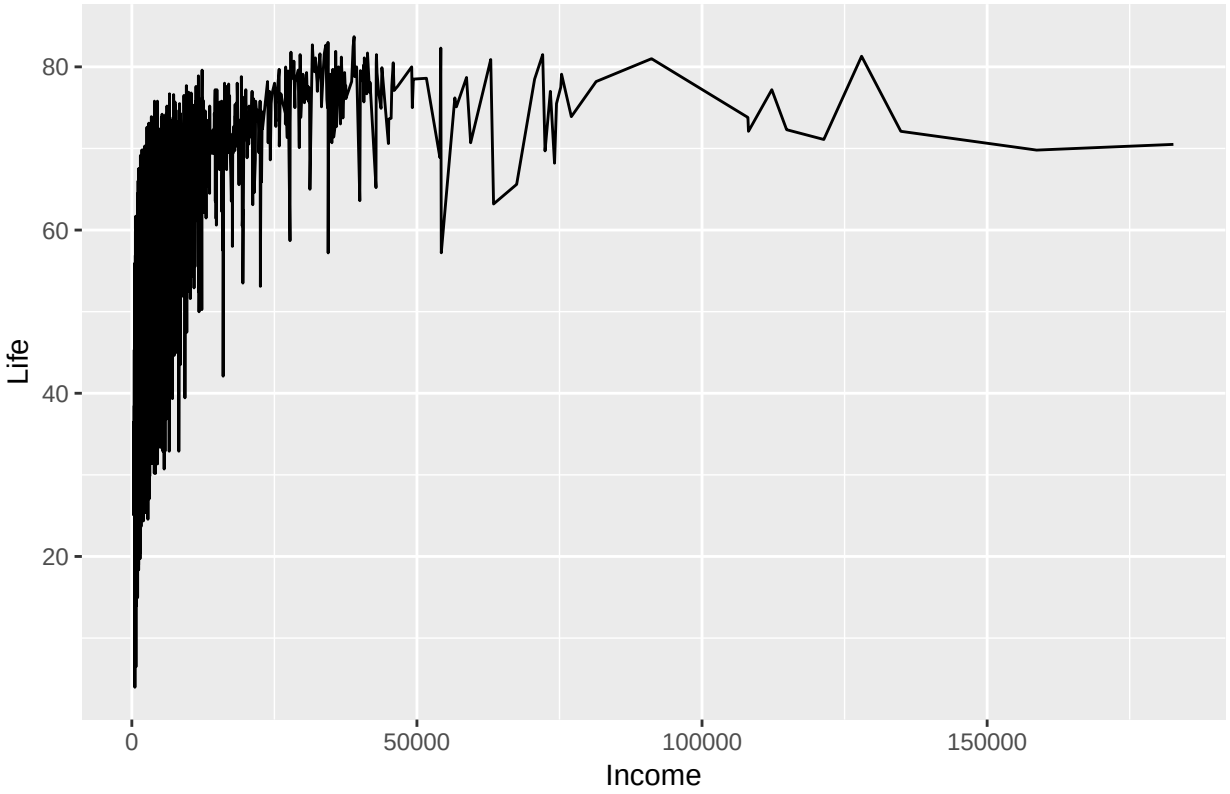


Fig. 6 − Brunei income over time




```
## Warning: Removed 240 rows containing missing values (geom_path).
```

Fig. 7 − Brunei life expectancy vs income



Cluster Analysis Further we can try to uncover any clusters in the data set that might help explain the relationship between life expectancy and income. Using the elbow method, it looks like the optimal number of clusters for this data set is 2. Thus, conducting kmean clustering using 2 centers for life and income variables in the dataset. There seems to be 2 available groups in the given data set. First group is the high income and long life expectancy group. The other group is low income and moderate life expectancy group.

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

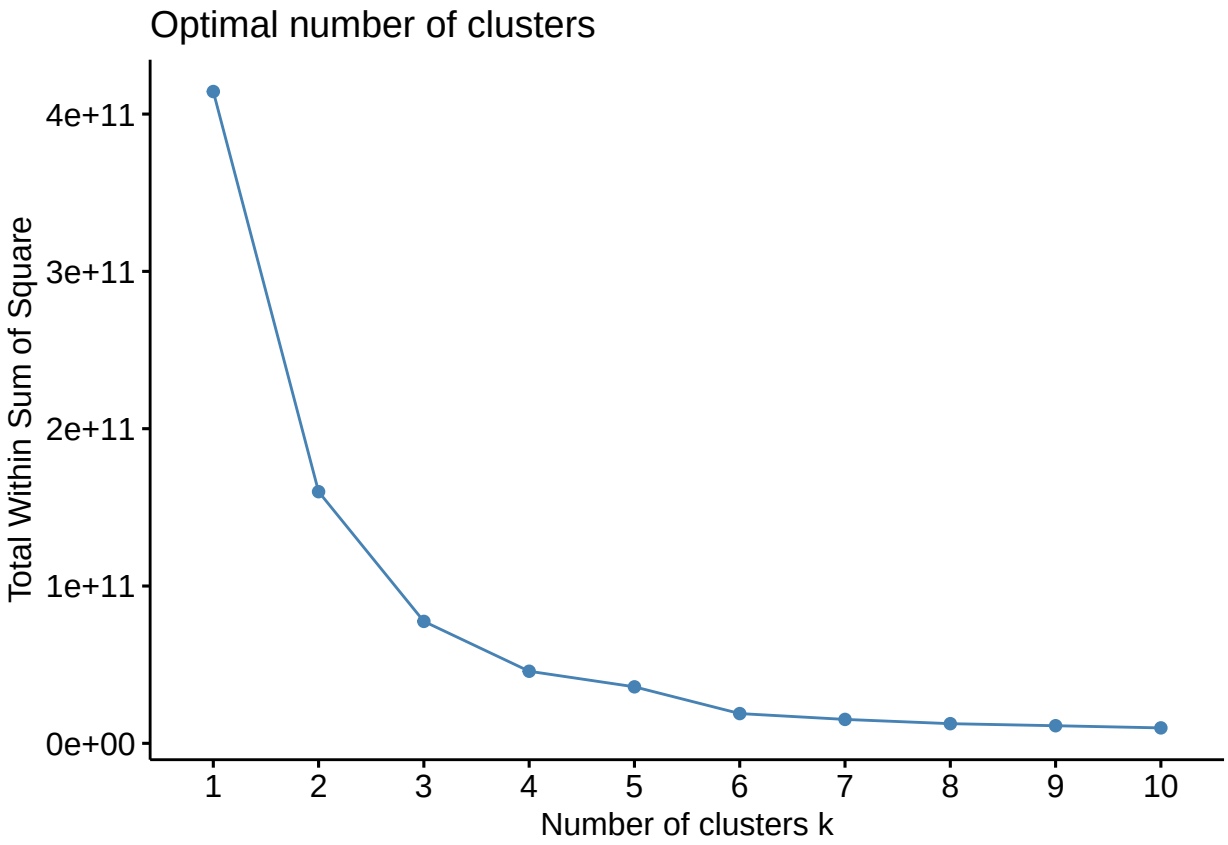
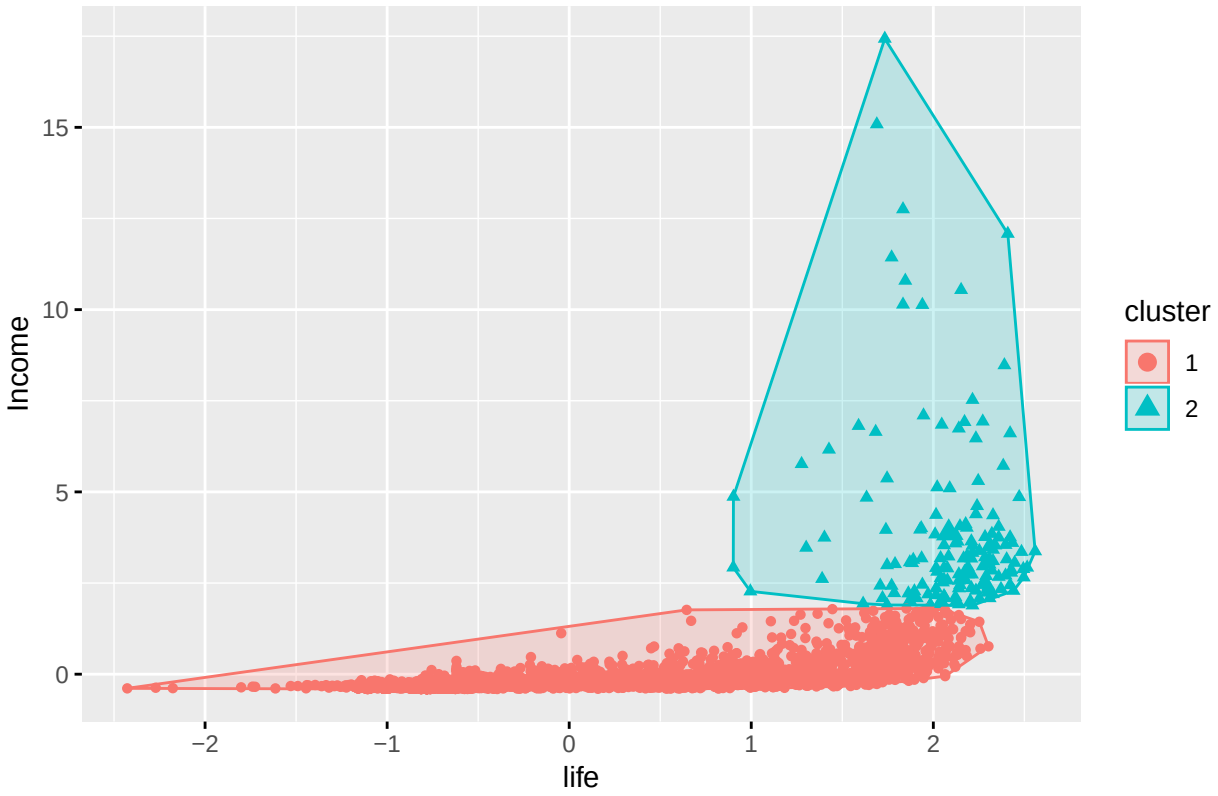

Optimal number of clusters
Total Within Sum of Square
4e+11
3e+11
2e+11
1e+11
0e+00
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Fig. 8 − Cluster − Life and Income



#Conclusion

From the given data set, we were able to observe that the top richest countries in the world have an oil dependent economy and the income for these countries is far greater than the median income , indicating high economic disparity. Further, the income in these countries is volatile and is probably highly correlated with oil prices. Also, the life expectancy in these countries seems to increase with the increase in income.