Final EDA Assignment

Pooja Chopra

16 June 2019

R Markdown

In this EDA project, I have used the Gapminder Dataset for analysis. First We will be following the epicycle analysis in R which we have learned during this course. We will start with setting up questions.

Setting QUESTIONS

Q1 - How per capita income in SouthAsia region has trended over the Years? Which country per capita income increases the most in south asia region? Q2 - What was the percent growth in world GDP per capita in 2015 as compared to 1800? Q3 - Does life expectancy increase with increase in income? Is there enough evidence that high income does not lead to high life expectancy for a country/region?

Let us start with analysis, results and their interpretations:

Importing Required Libraries

Importing raw data

Data Description

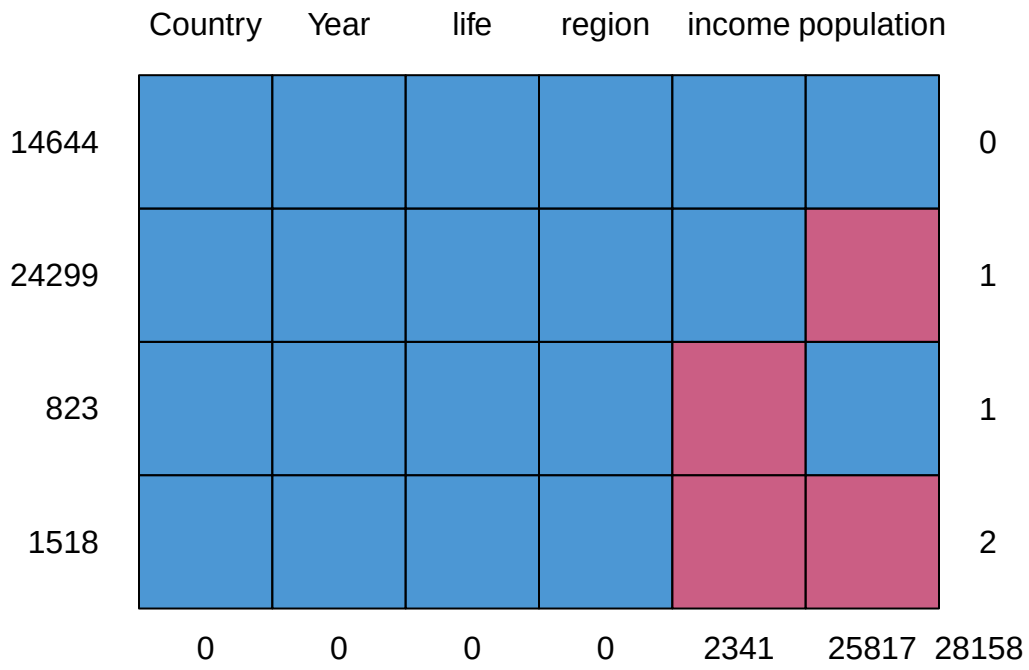
Region/Country describes the continent and country the Life expectancy is measured Population, Income, and Life expectance is provided for each country from years 1800 to 2015

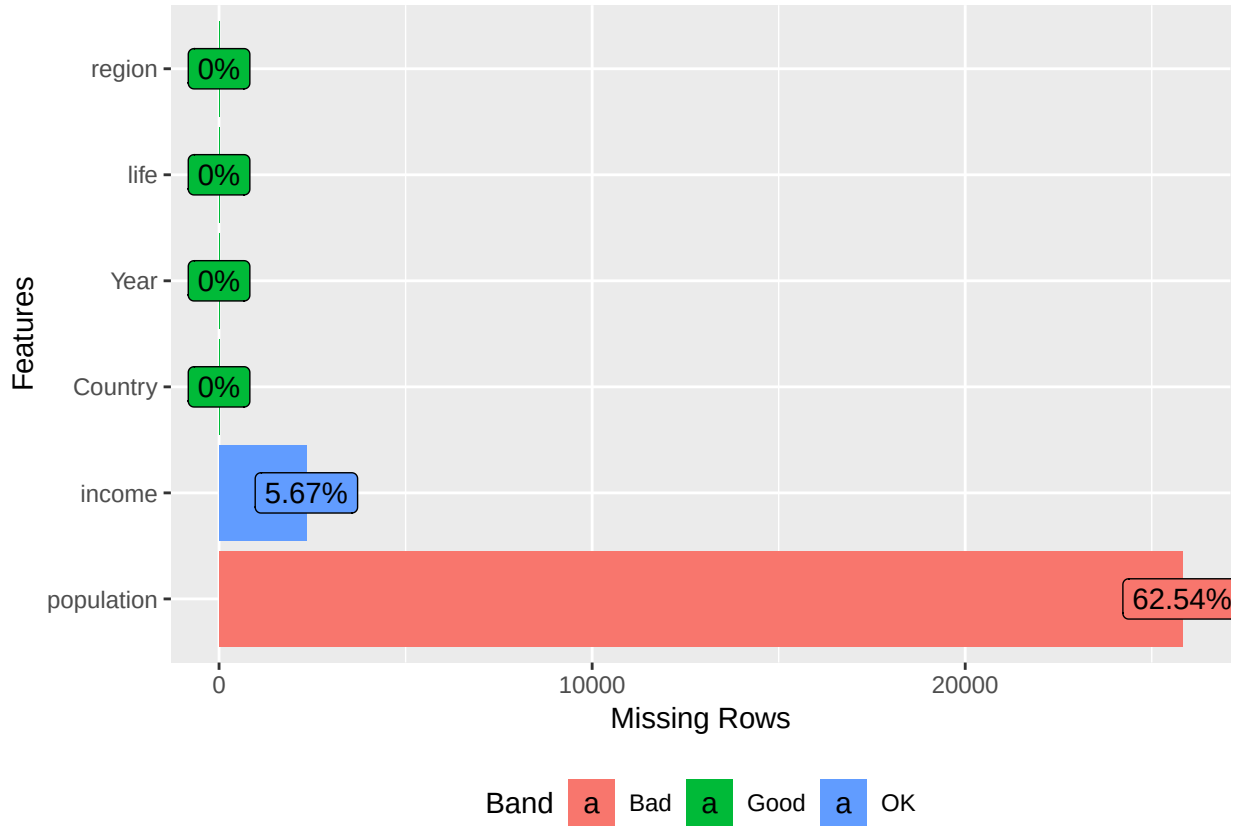
Gapminder dataset have 41284 observation with 6 variables. Type of Variable Country is categorical variable with 197 Levels,Region is categorical variable with 6 Levels,Year,Life and Income are continous variable and Population are descrete variable

Dispersion of Data

In our dataset, data has been gathered for over 215 years starting from 1800 up till 2015. It includes the population of 197 countries spread across 6 regions changed over time. The life expectacy ranges from 1 to 83%.

Data Wrangling



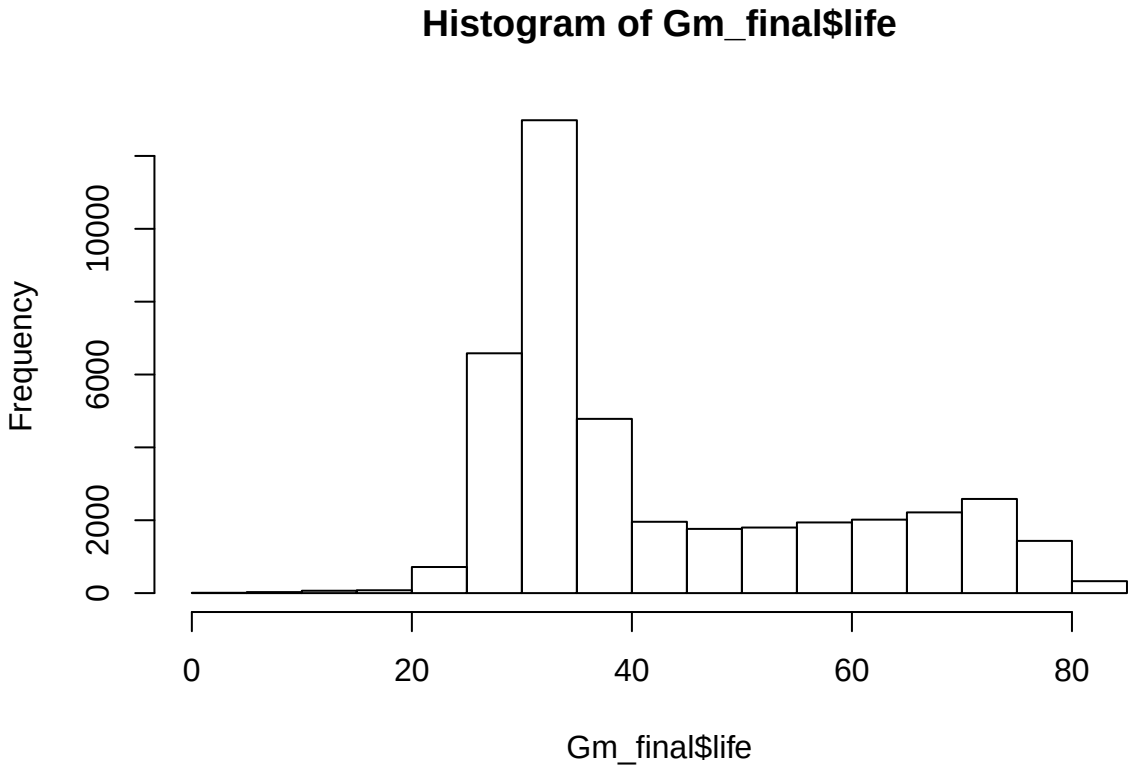


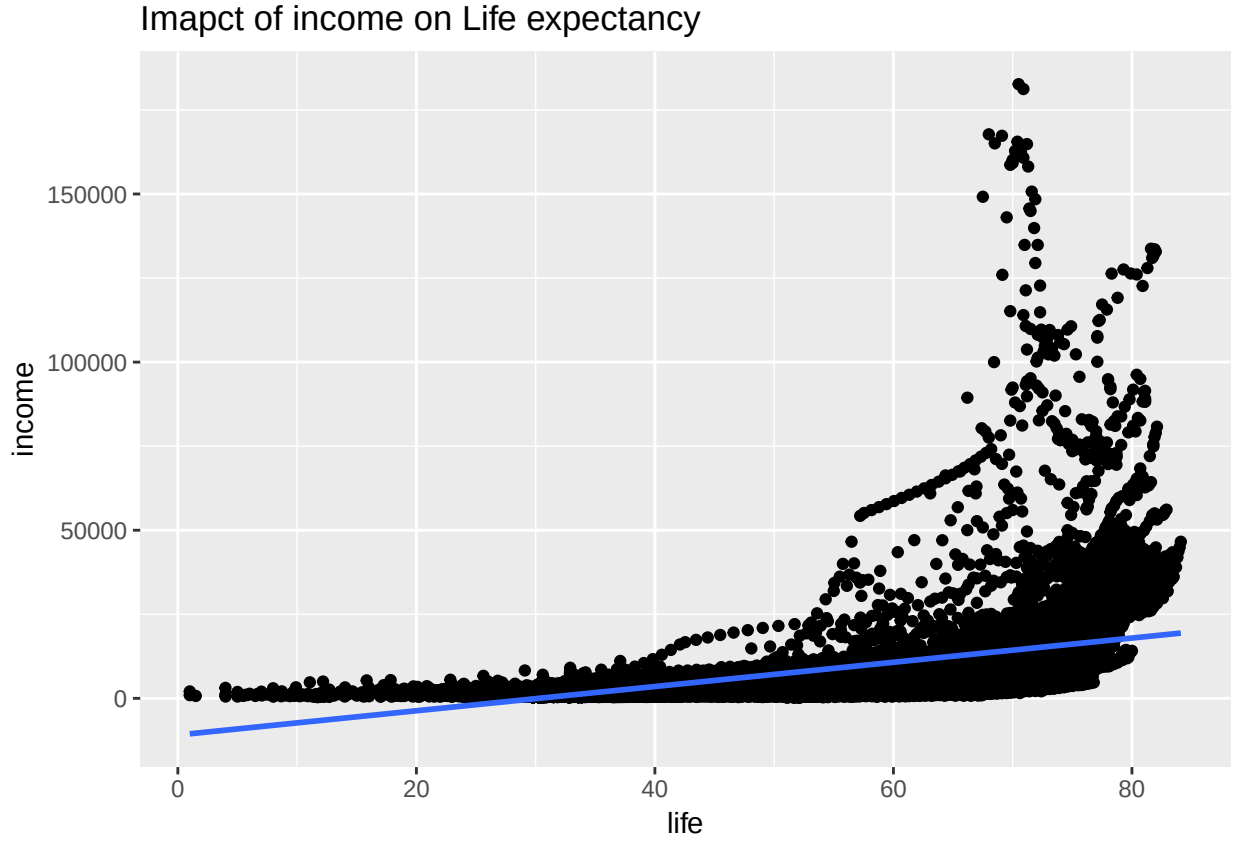
In our data set we have 25817(62.54 %) missing values in population and 2341(5.67%) missing values in Income. In our dataset population is calculated every ten years, we have filled missing values same as calcualted at the start of the decade. In our dataset Missing values of income are for multipe countries and it is observed that Values were never calculated from the start so it is not possible to handle missing value of Income hence we are eliminating observations where we have NA income.

Preprocessing

We used the function outliers only on the numeric columns which are Life ,Population . we find few outliers in out data which can be ignored as it will not impact our analysis. We have taken log for population column for further analysis.

Data Exploration

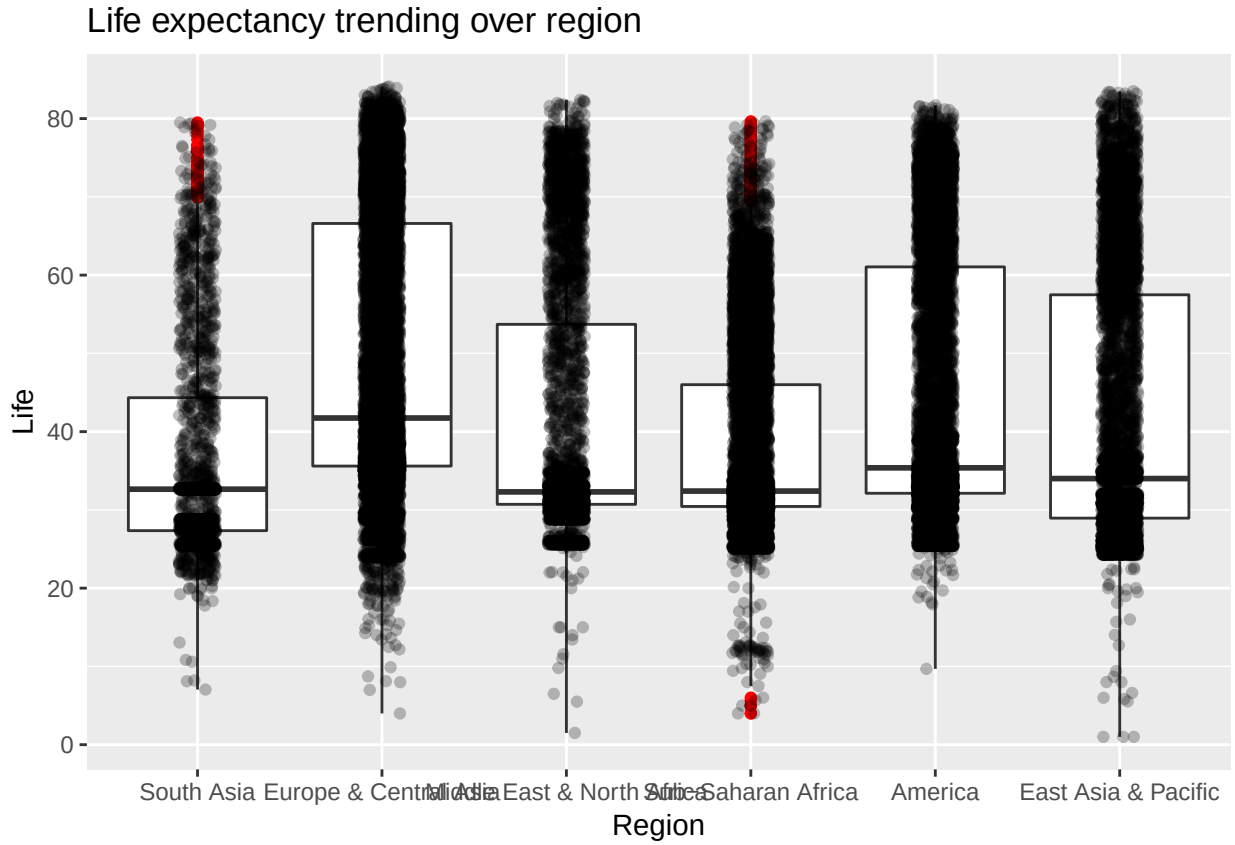




Looking at the histogram , Life expectancy trends to increase after 20 till 35. In other plot we can conclude that life expectancy is increasing regardless of growth in income but EXpectacy level grows above 60 when the income levels increase.

Regionwise ANalysis

```
## # A tibble: 6 x 3
##   region                        mean median
##   <fct>                        <dbl>  <dbl>
## 1 South Asia                    37.4   32.6
## 2 Europe & Central Asia         48.8   41.7
## 3 Middle East & North Africa    41.6   32.3
## 4 Sub-Saharan Africa            37.9   32.4
## 5 America                       44.5   35.4
## 6 East Asia & Pacific           41.8   34
```

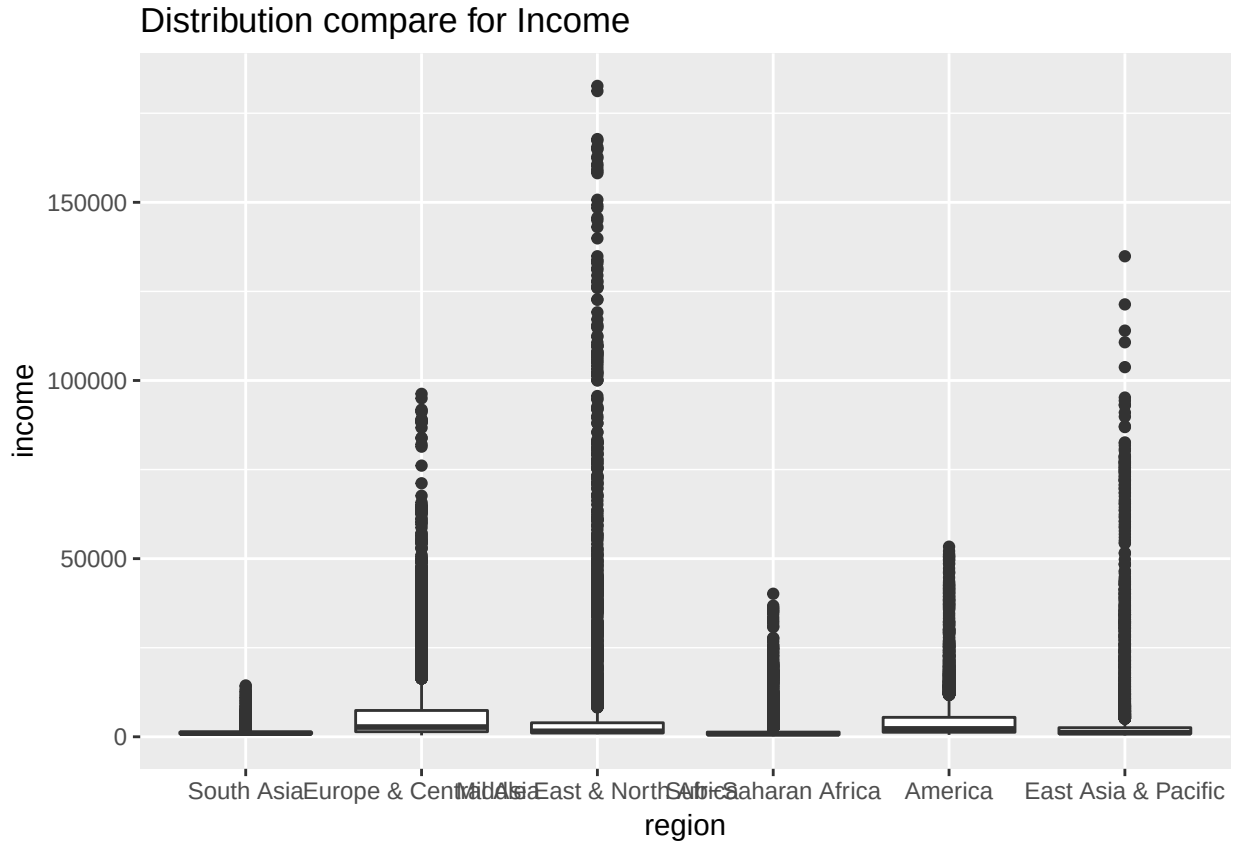


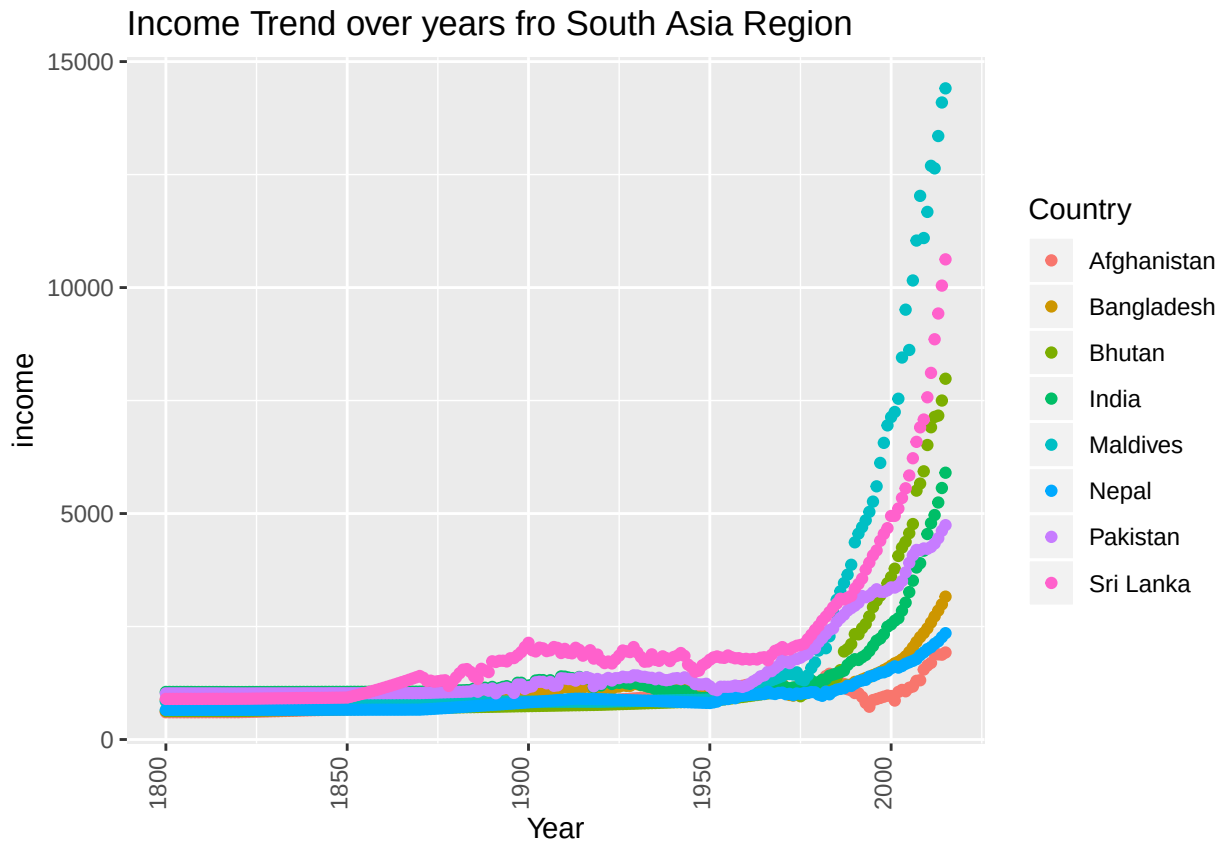
Looking at the Life expectacy trending over region Graph, Europe region has high level of Life Expectacy as compared to our Regions , America and East Asia pacific is also high levl of life expectact.

income Study

```
## # A tibble: 6 x 5
##   region                      mean median   min    max
##   <fct>                      <dbl>  <dbl> <int>  <int>
## 1 South Asia                1385.   1020    603  14408
```

```
## 2 Europe & Central Asia      6829.  2735    393  96245
## 3 Middle East & North Africa 8064.  1538.   715 182668
## 4 Sub-Saharan Africa         1464.   827    142  40143
## 5 America                    4595.  2214.   529  53354
## 6 East Asia & Pacific        4412.  1154.   363 134864
```





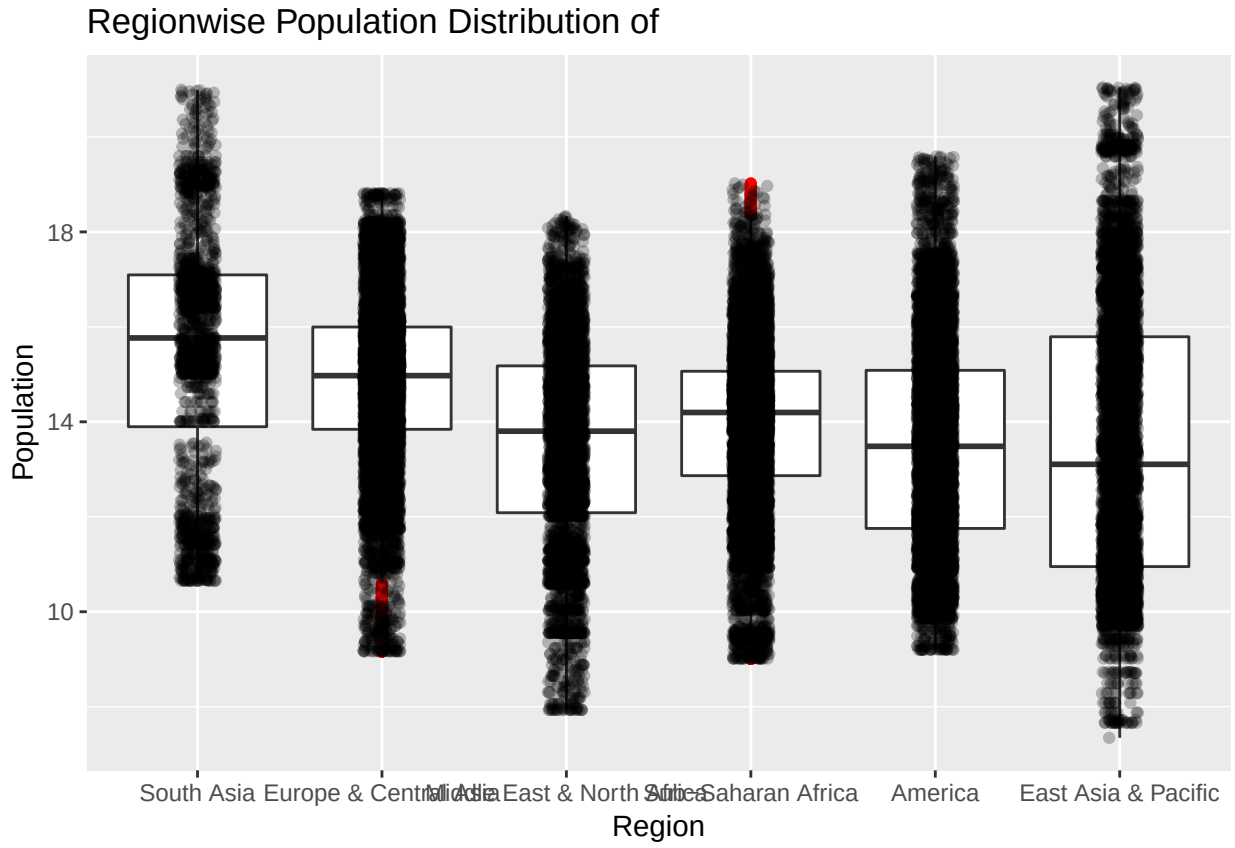
"The Percent growth (or decline) in GDP per capita in 2015 for the world was 18.62%". From distribution compare for income Plot we observed that , Europe income is substantially higher than all regions. If we look into Income and Region data for south asia region we observed that Per capita income is tend to increase over the Year and Among all countries Maldives per capita increases the most in south Asia region till 2014.

Population

```
## # A tibble: 6 x 3
##   region                        max   min
##   <fct>                       <dbl> <dbl>
## 1 South Asia                   21.0 10.7
## 2 Europe & Central Asia        18.8  9.17
## 3 Middle East & North Africa   18.3  7.93
## 4 Sub-Saharan Africa           19.0  9.01
## 5 America                      19.6  9.20
## 6 East Asia & Pacific          21.0  7.34
```

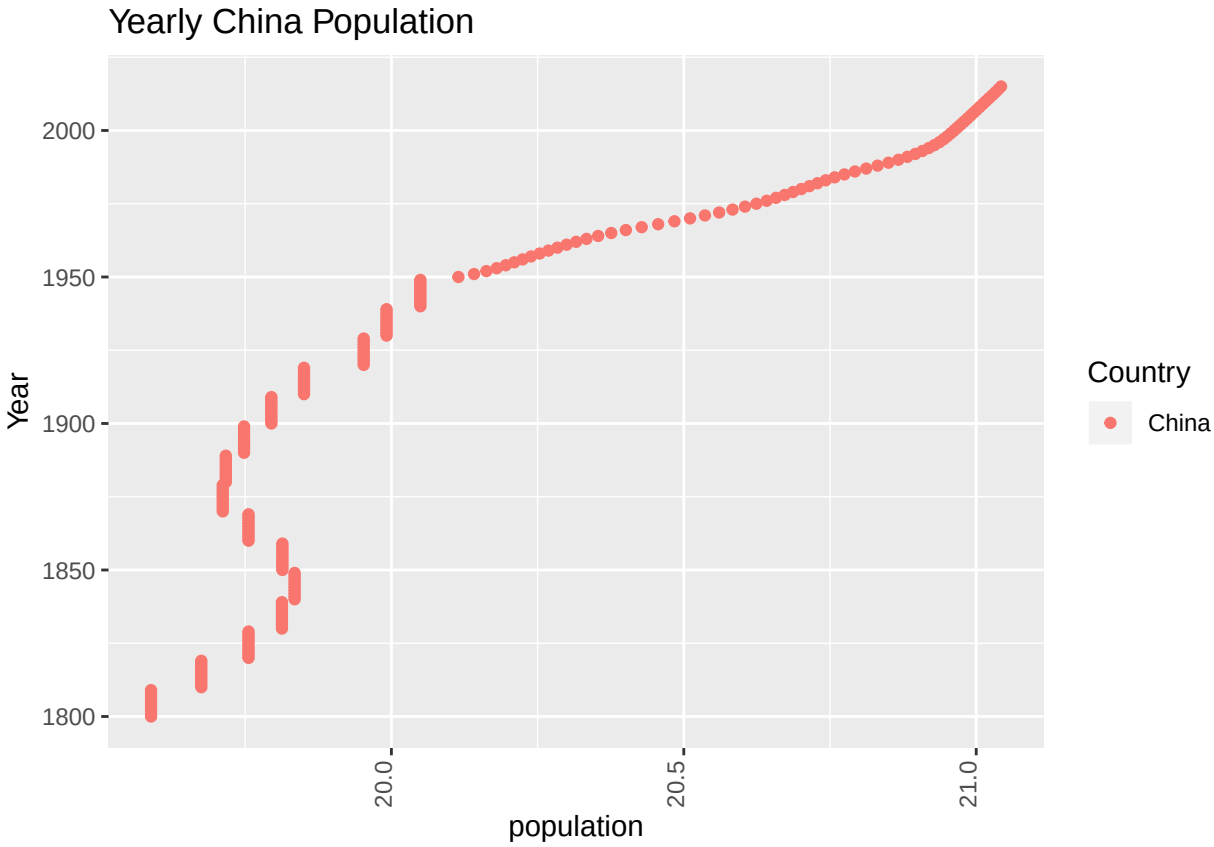
```
## # A tibble: 6 x 2
##   region                     Total_population
##   <fct>                                 <dbl>
## 1 South Asia                            26819.
## 2 Europe & Central Asia                155695.
## 3 Middle East & North Africa            58338.
## 4 Sub-Saharan Africa                   147470.
## 5 America                              107625.
## 6 East Asia & Pacific                   84407.
```

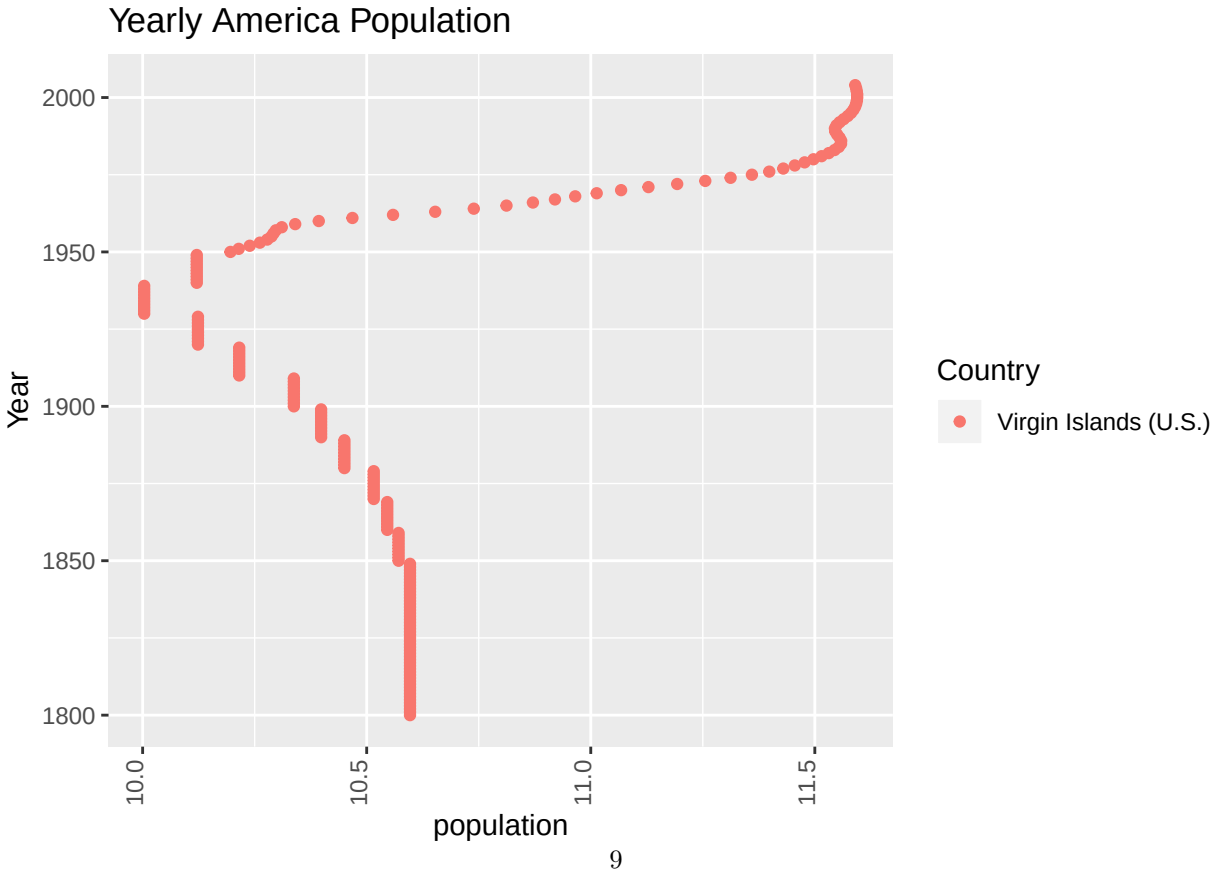
```
## # A tibble: 6 x 3
##   region                       Total_population   Income
##   <fct>                                   <dbl>    <int>
## 1 South Asia                             26819.  2393238
## 2 Europe & Central Asia                 155695. 70901008
## 3 Middle East & North Africa             58338. 33096223
## 4 Sub-Saharan Africa                    147470. 14860160
## 5 America                               107625. 31867866
## 6 East Asia & Pacific                    84407. 24881983
```



With this analysis we observed that population of Asia pacific region is higher than other regions. If we compare Regionwise population with income Europe & Central Asia has highest population with highest income we can say that increase in population increase income.

Country


Yearly China Population
2000
1950
1900
1850
1800
Year
20.0
20.5
21.0
population
Country
China

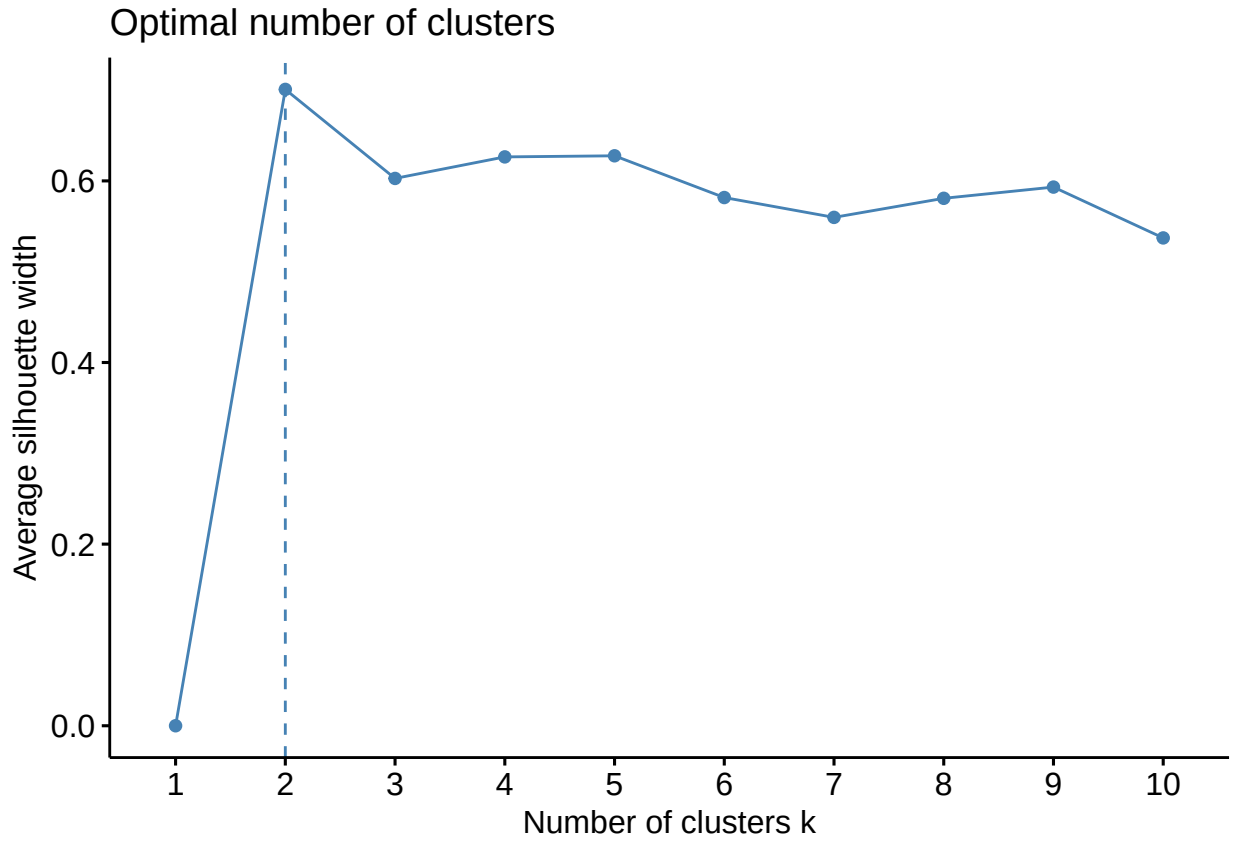

Yearly America Population
2000
1950
1900
1850
1800
Year
10.0
10.5
11.0
11.5
population
Country
Virgin Islands (U.S.)

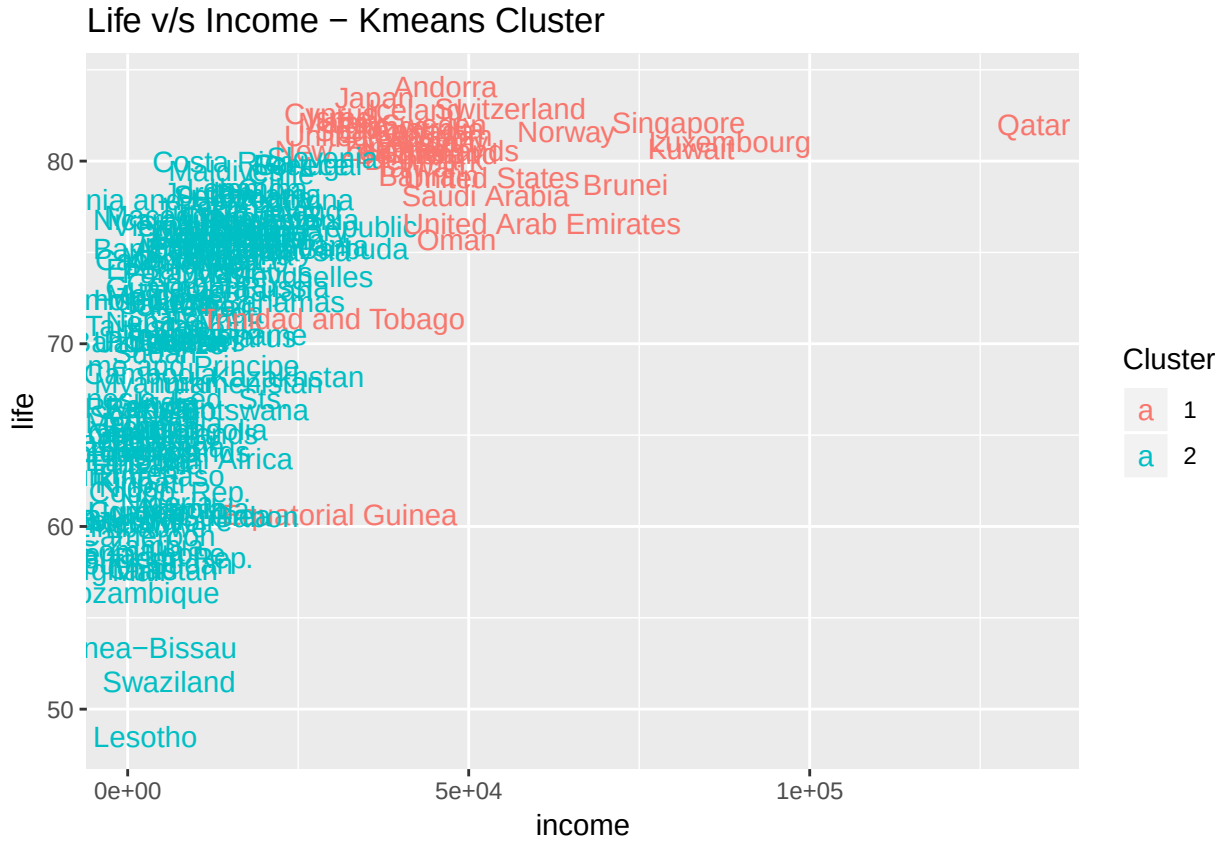
Here we can observe that population for both countries China and Virginia island increase after 1950's

Summary

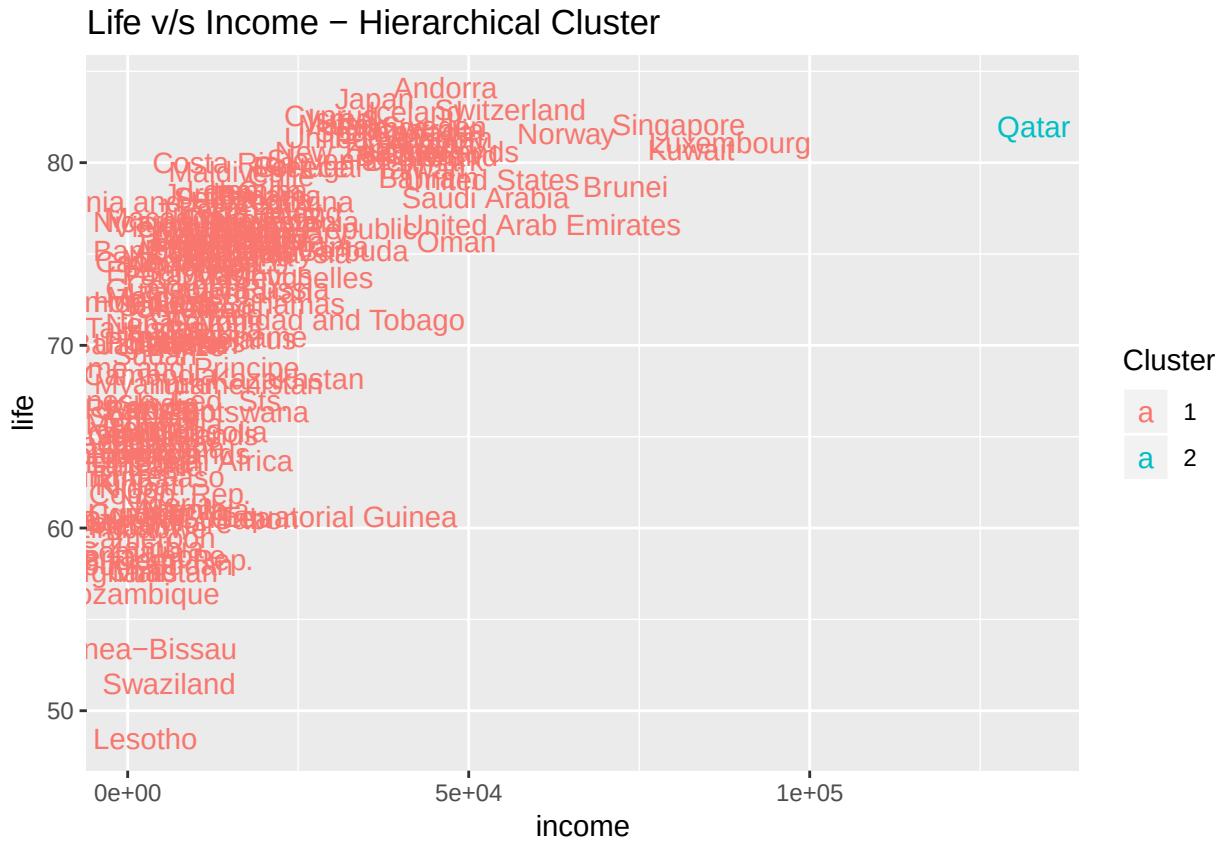
In this Analysis we observed that The mean of Life Expectancy for *Sub-Saharan Africa* region is lower than all other regions, while Europe & Central Asia is highest. *Q1*If we look into Income and Region data for south asia region we observed that Per capita income is tend to increase over the Years and Maldives per capita increases the most in south Asia region till 2015. *Q2* paste0("The Percent growth (or decline) in GDP per capita in 2015 for the world was", round(World_GDP_summarise[2,2]/World_GDP_summarise[1,2], 2),"%") Life expectancy in the Gapminder dataset is higher in the range of 25-35 years. Though it increase regardless of income *Q3* the income levels rises higher,life expectancy grows above 60 .With this analysis of population with region we observed that population of Asia pacific region is higher than other regions. If we compare Regionwise population with income Europe & Central Asia has highest population with highest income we can say that increase in population increase income.

Clustering


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Life v/s Income − Kmeans Cluster
life
80
70
60
50
0e+00
5e+04
1e+05
income
Andorra
Japan
Switzerland
Norway
Singapore
Qatar
Kuwait
Saudi Arabia
Brunei
United Arab Emirates
Oman
Trinidad and Tobago
Equatorial Guinea
Swaziland
Lesotho
Cluster
a 1
a 2

In Kmeans Cluster the countries are more or less divided based on the life expectacy in 2015. The cutoff seems to be near 70 years of life expectancy. This also points to an interesting fact that richer countries don't actually live longer than poorer countries.



We observe results with the cluster being separated long life variable but here the cutoff seems to be slightly lower than 70 years.