Gapminder data exploration

Krishna Chaitanya Kundety

June 16, 2019

Contents

1 Dataset **1**

2 Missing records **2**
2.1 Are there data samples with NAN values? 3

3 Data visualization **4**
3.1 Population growth over time 4
3.2 How is the current life expectancy disributed? 5
3.3 How is the current world population distributed? 7
3.4 Is there a relationship between life expectancy and oher variables? 9

4 Kmeans clustering **12**
4.1 Can we predict region (using kmeans) if year, population, income and life are given? 12

1 Dataset

The dataset is gapminder. Below is the structure.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    41284 obs. of  6 variables:
##  $ Country   : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year      : int  1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 ...
##  $ life      : num  28.2 28.2 28.2 28.2 28.2 ...
##  $ population: num  3280000 NA NA NA NA NA NA NA NA NA ...
##  $ income    : int  603 603 603 603 603 603 603 603 603 603 ...
##  $ region    : chr  "South Asia" "South Asia" "South Asia" "South Asia" ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 6
##   .. ..$ Country   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Year      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ life      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ population: list()
##   .. .. ..- attr(*, "class")= chr  "collector_number" "collector"
##   .. ..$ income    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ region    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

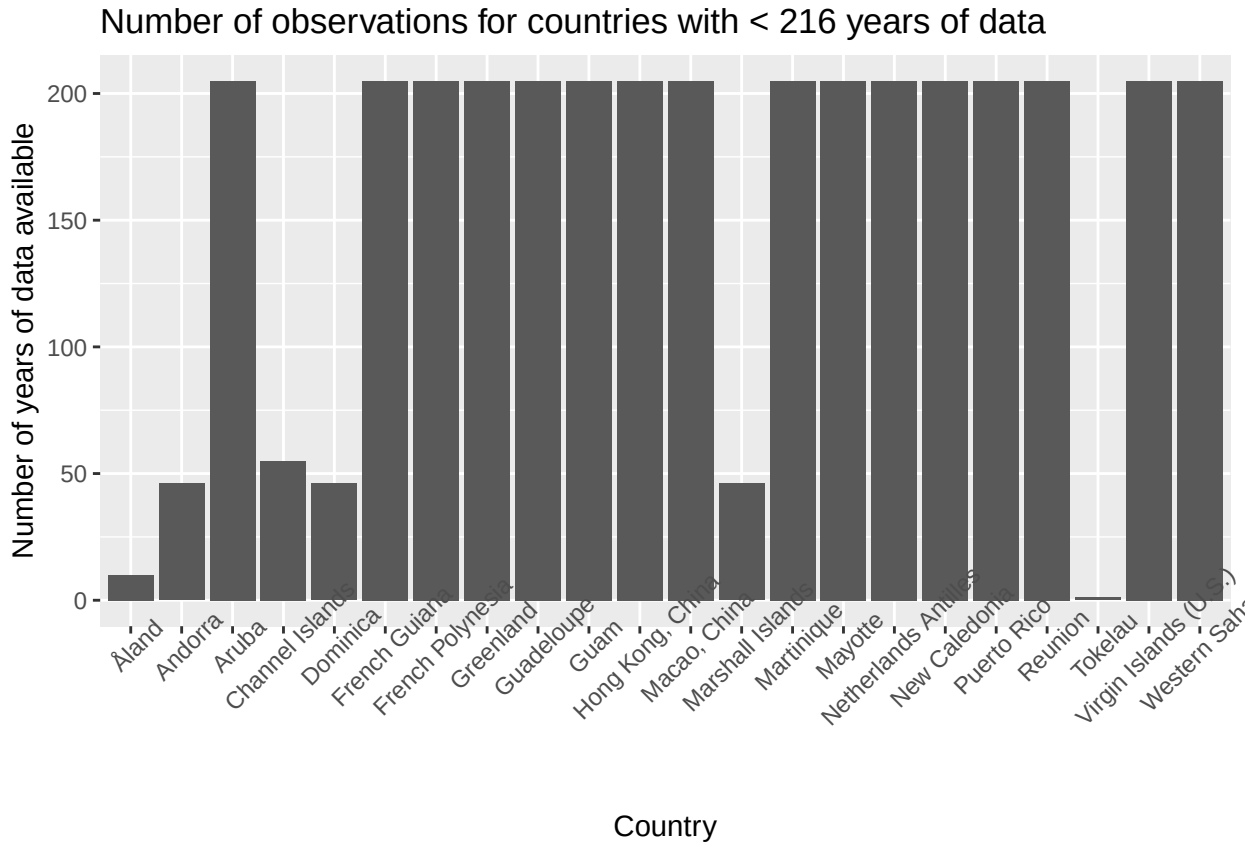
The dataset contains 41284 observations and 6 variables.

The dataset contains life expectancy, income and population of various countries over the past years. Data is provided on 197 countries, from 6 regions around the world. The time range spans over 216 years, from 1800 to 2015.

2 Missing records

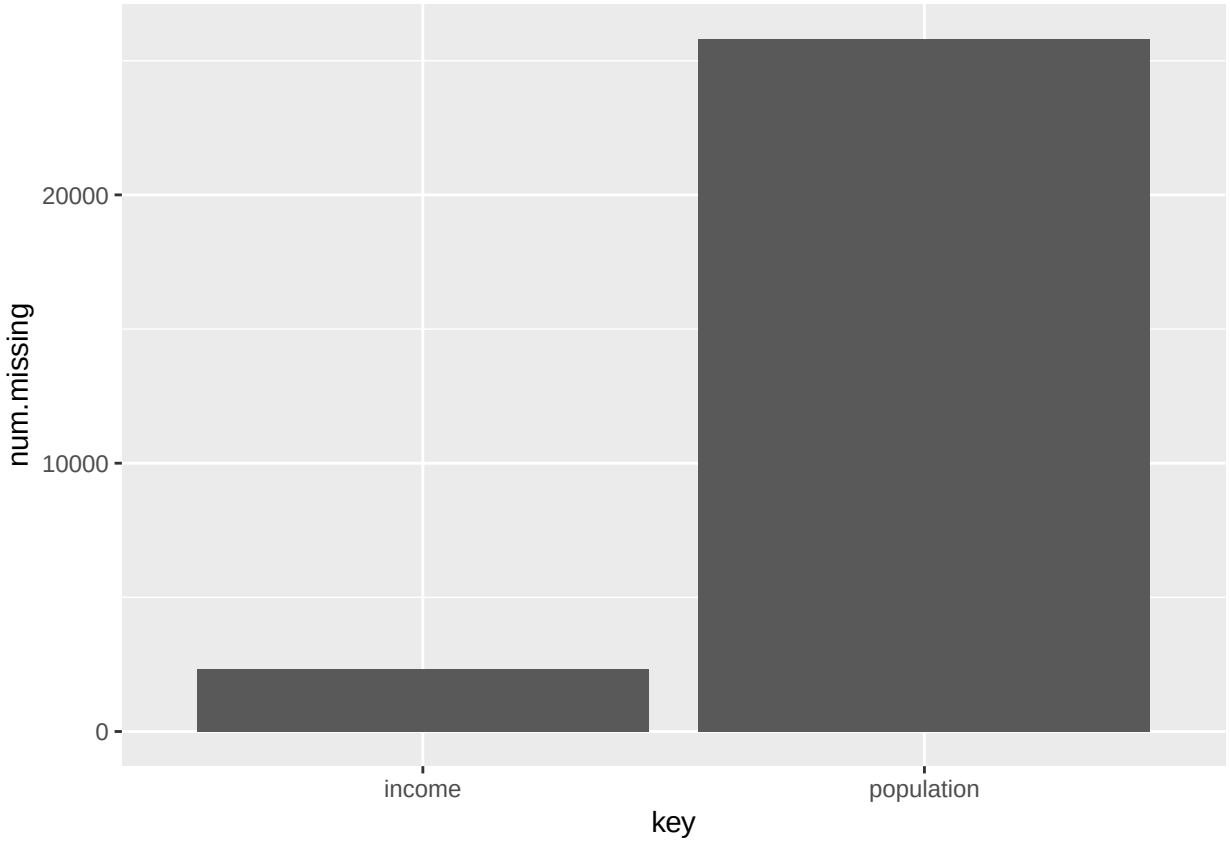
If we assume the dataset to be complete, i.e., each country has 216 observations (one observation for every year), there should be 42552 observations. But this is not the case, as we have seen that there are 41284 observations. This means there are some missing information.

Let us see the countries for which there are less than 216 years in the records.



So 22 countries have less than 216 of data. This can be due to several reasons. The data may not have been recorded for some years, or that the country was founded less than 216 years ago.

2.1 Are there data samples with NAN values?

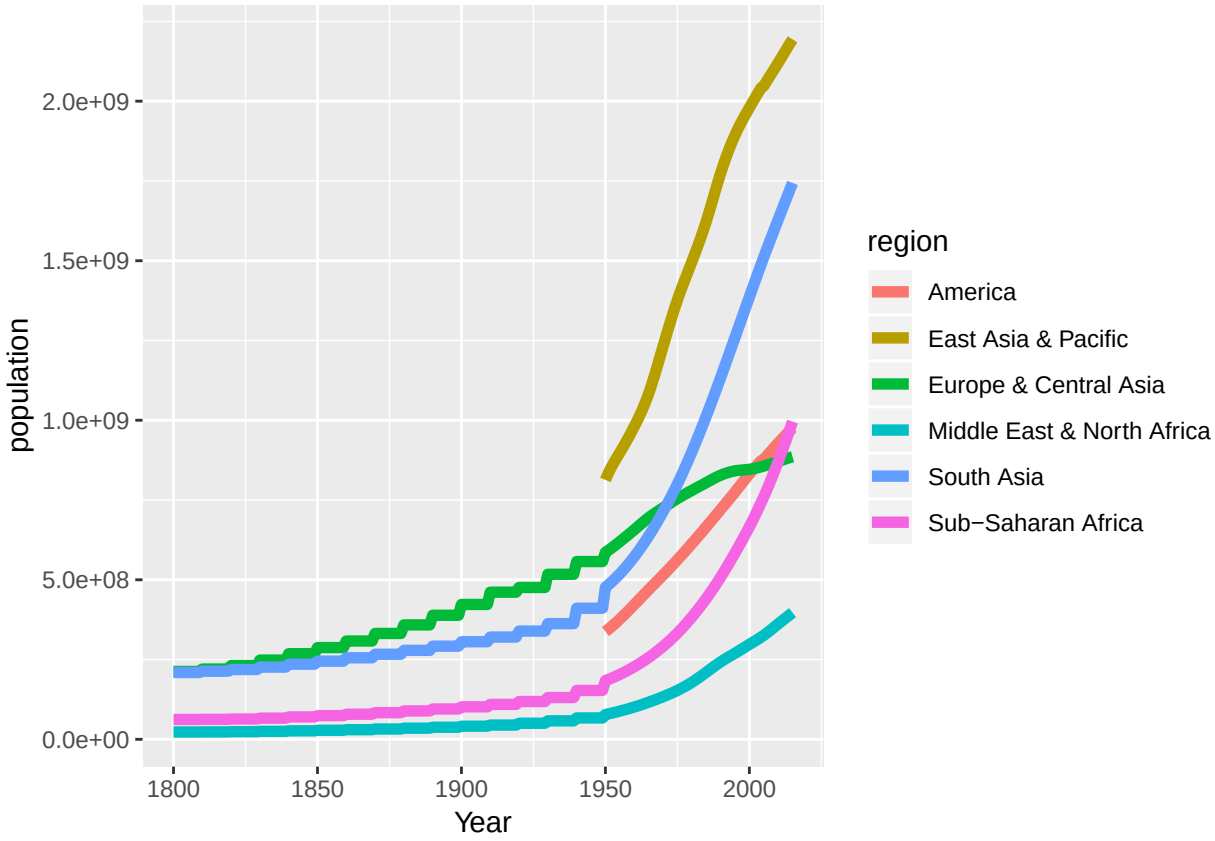


We can see there are many missing values in population column. This may be because the population census is conducted every few years. In this case, the population can be considered the same as the previous years.

Let us assume that the population does not change significantly between two censuses of the country. Then we can backfill the missing population values.

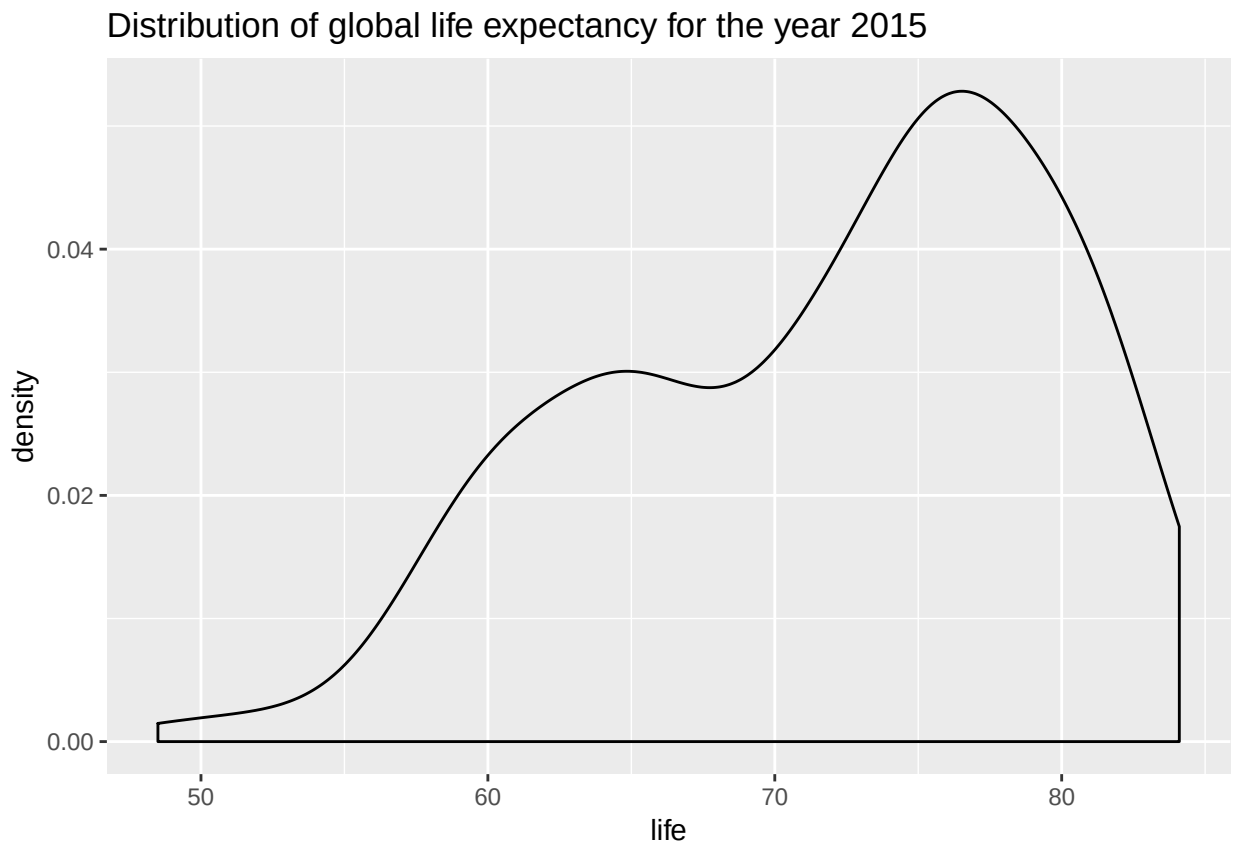
3 Data visualization

3.1 Population growth over time

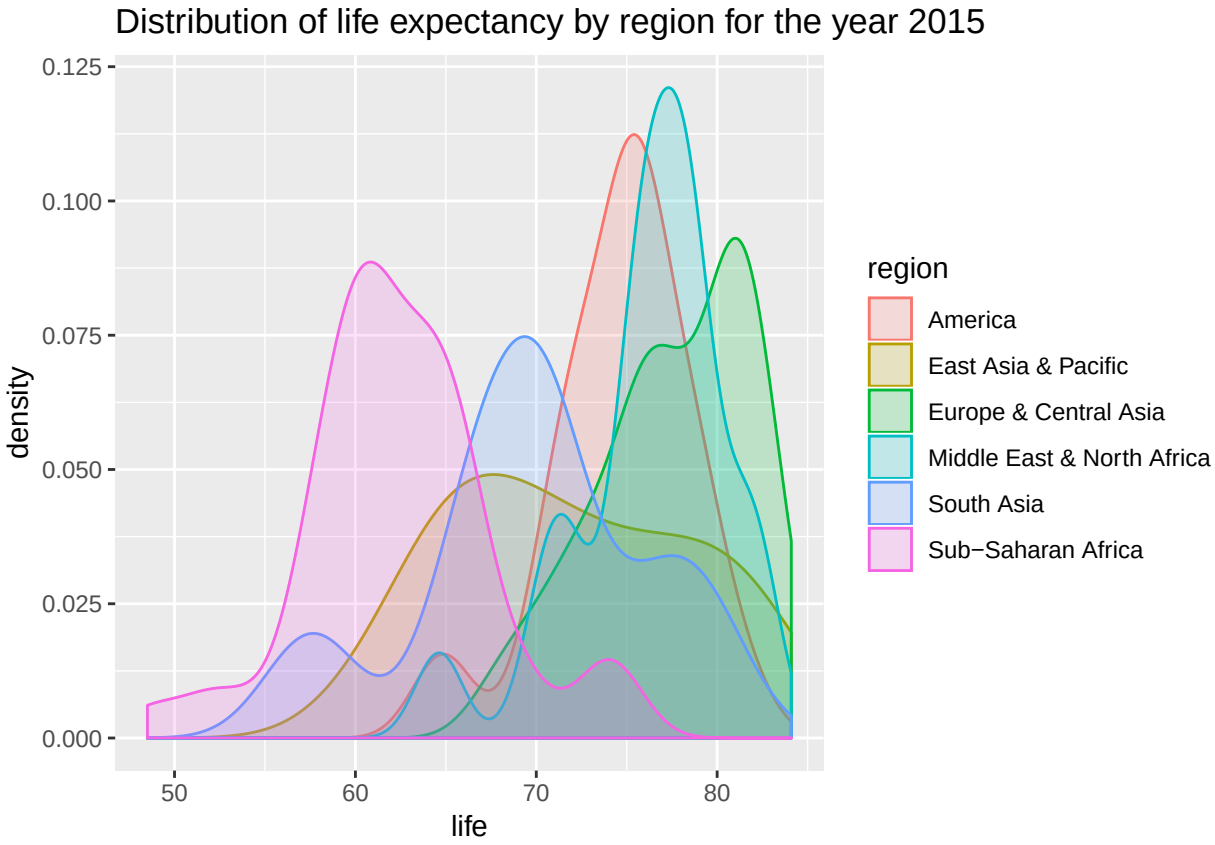


East Asis-Pacific has the highest population in all of recorded history. South Asia has had the highest growth in population. Since 1950, South Asia's population growth rate has been equal to that of East Asia-Pacific.

3.2 How is the current life expectancy disributed?

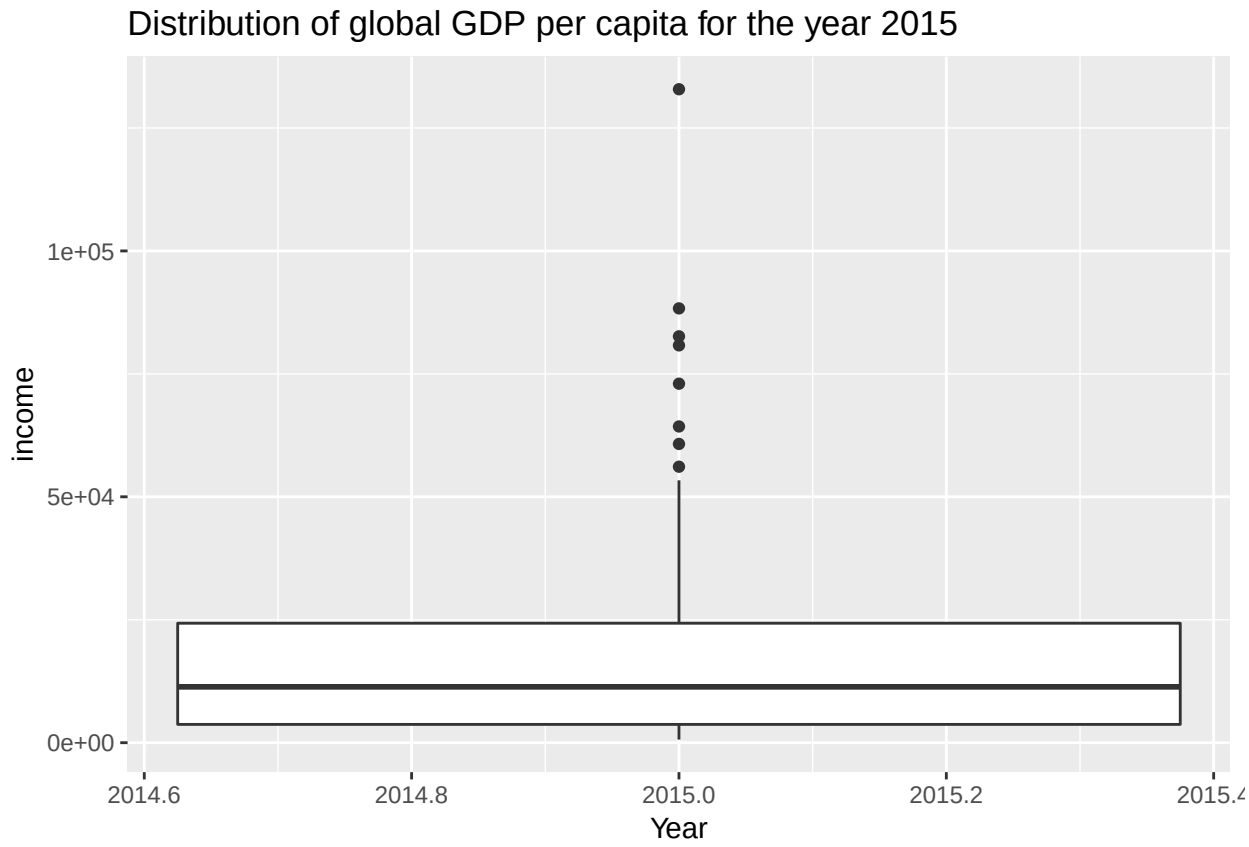


Most countries in the world have a life expectancy of around 75.

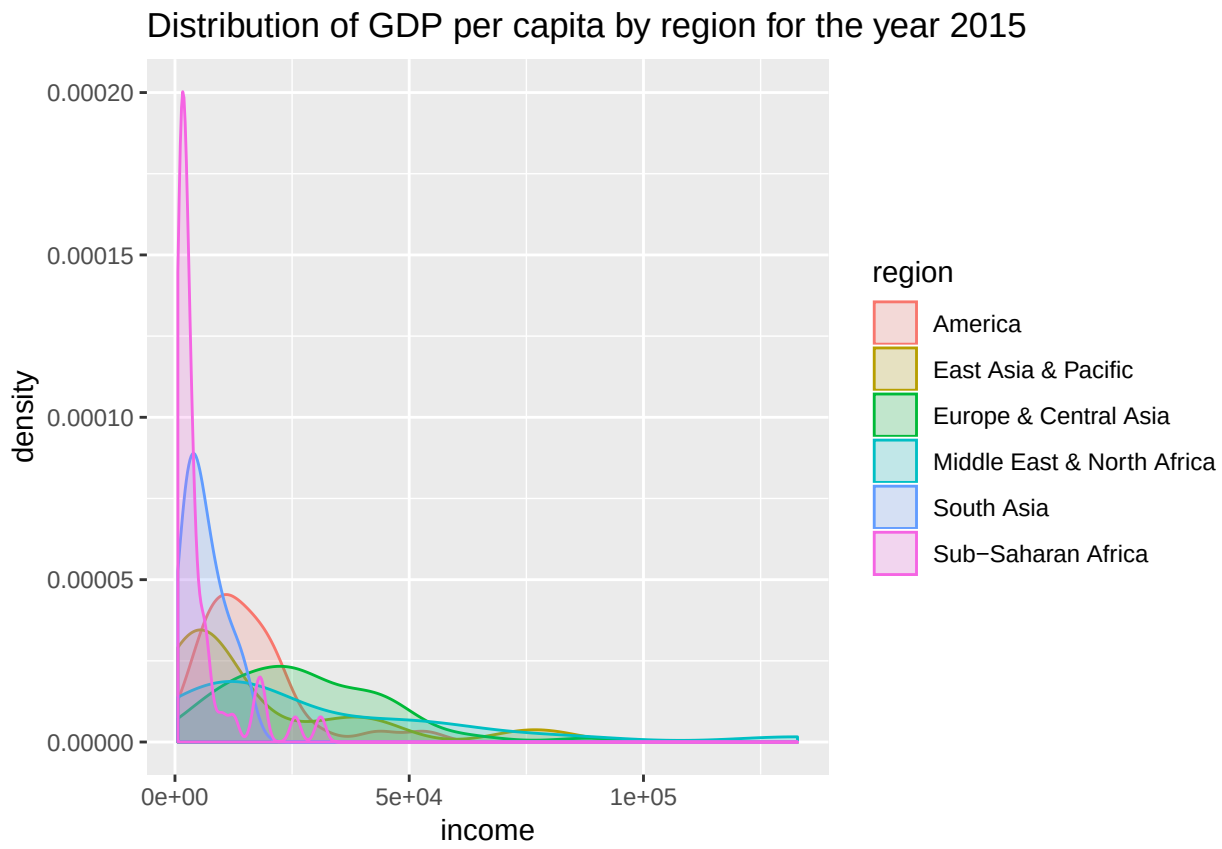


- The American subcontinent has the greatest spread of life expectancy.
- Middle East & North Africa region has the highest mean life expectancy
- Sub-Saharan Africa has the lowest mean life expectancy

3.3 How is the current world population distributed?

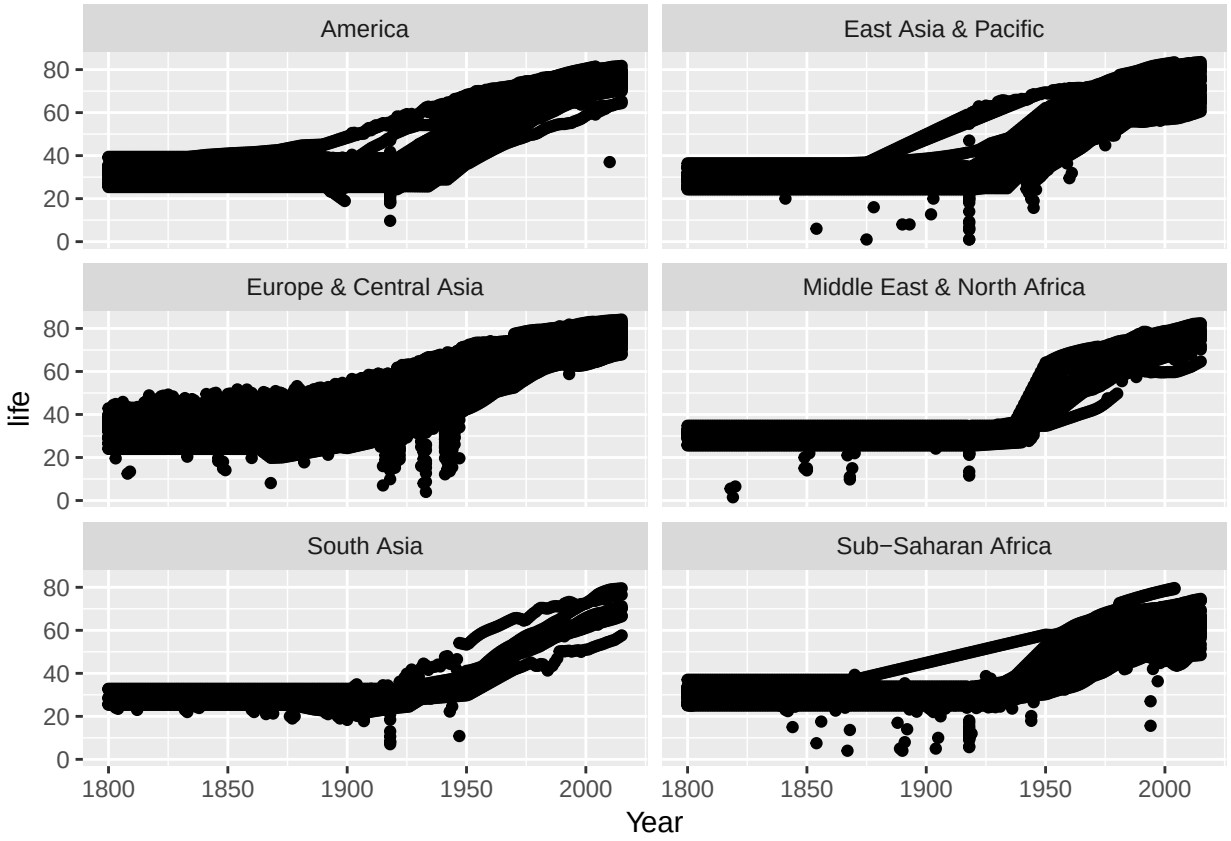


GDP per capita is skewed and has many outliers.



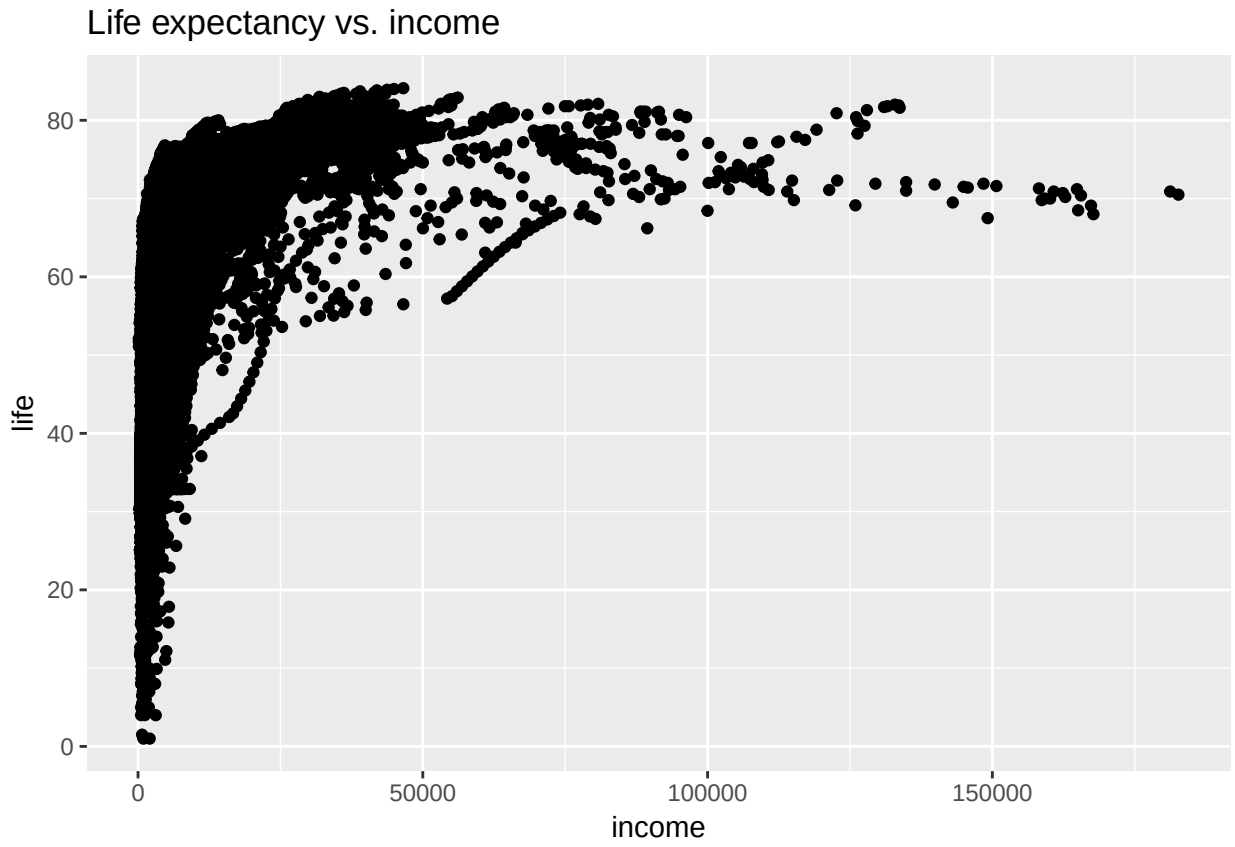
Except for The Americas and Europe & Central Asia, we see vvery similar characteristics.

3.4 Is there a relationship between life expectancy and oher variables?

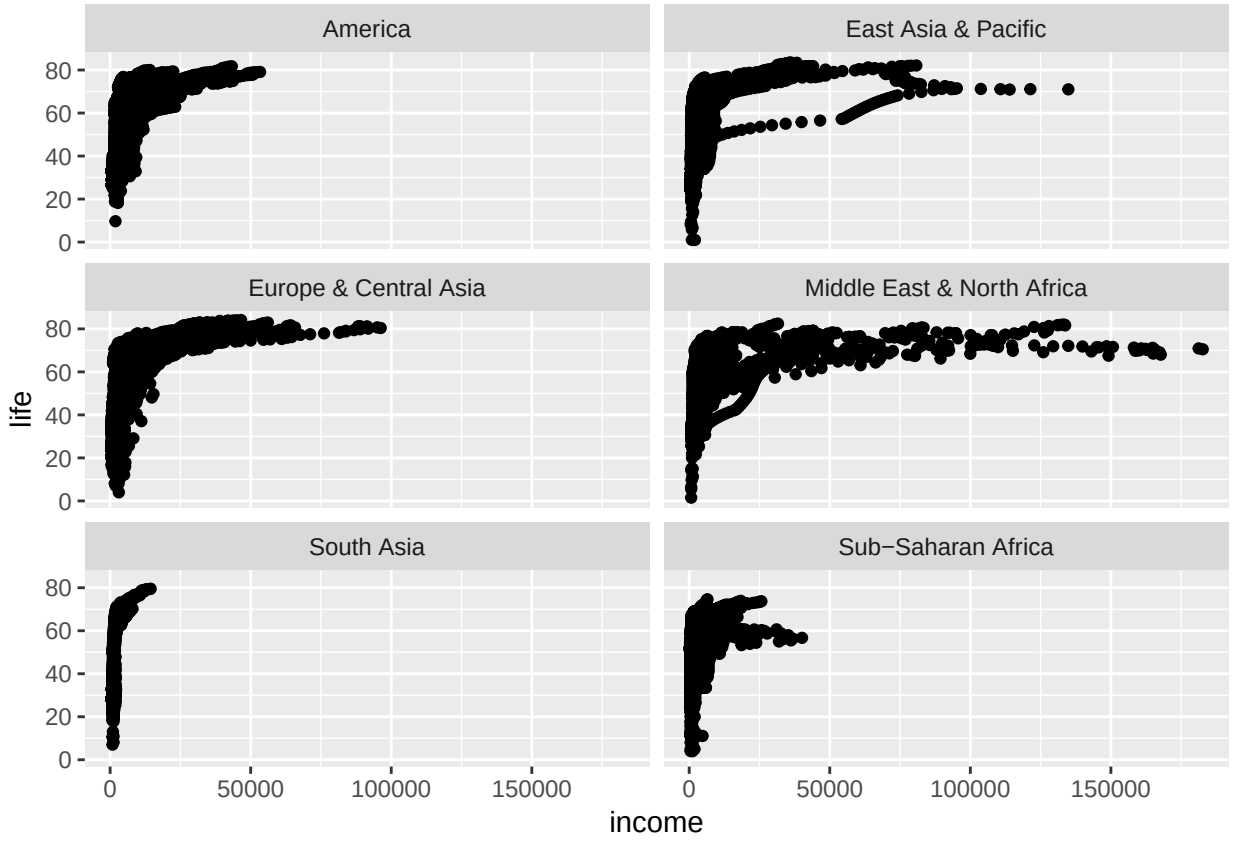


The life expecancy has drastically increased in all regions since 1950.

3.4.1 Has income affected the life expectancy?

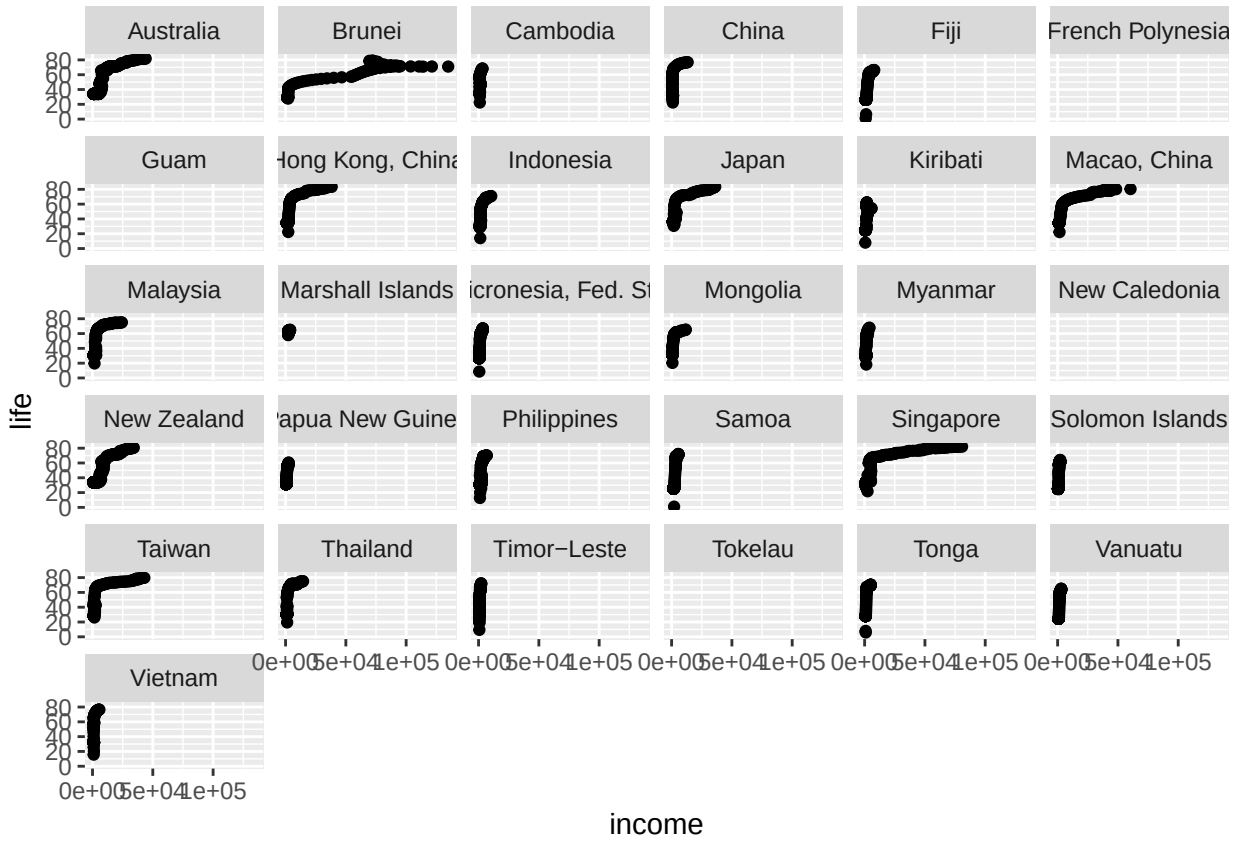


In all the cases, The highest life expectancies are found in high GDP per capita countries. The relationship is non-linear, increase in life expectancy dampening around 80 years.



We can compare life expectancy vs. GDP per capita in each region. We see the same non-linear relationship in all regions. South Asia and Sub-Saharan Africa have achieved some high life expectancy without he need for very high GDP per capita. It is evident from this plot that the op echelon of GDP per capita are all in Middle East & North Africa region. In the East Asia & Pacific region, one or more countries have a lower life expectancy despite having above average GDP per capita.

Let us zoom into East Asia & Pacific region to find out which country has lower life expectancy.



Brunei has comparatively lower life expectancy growth with GDP per capita compared to other countries.

4 Kmeans clustering

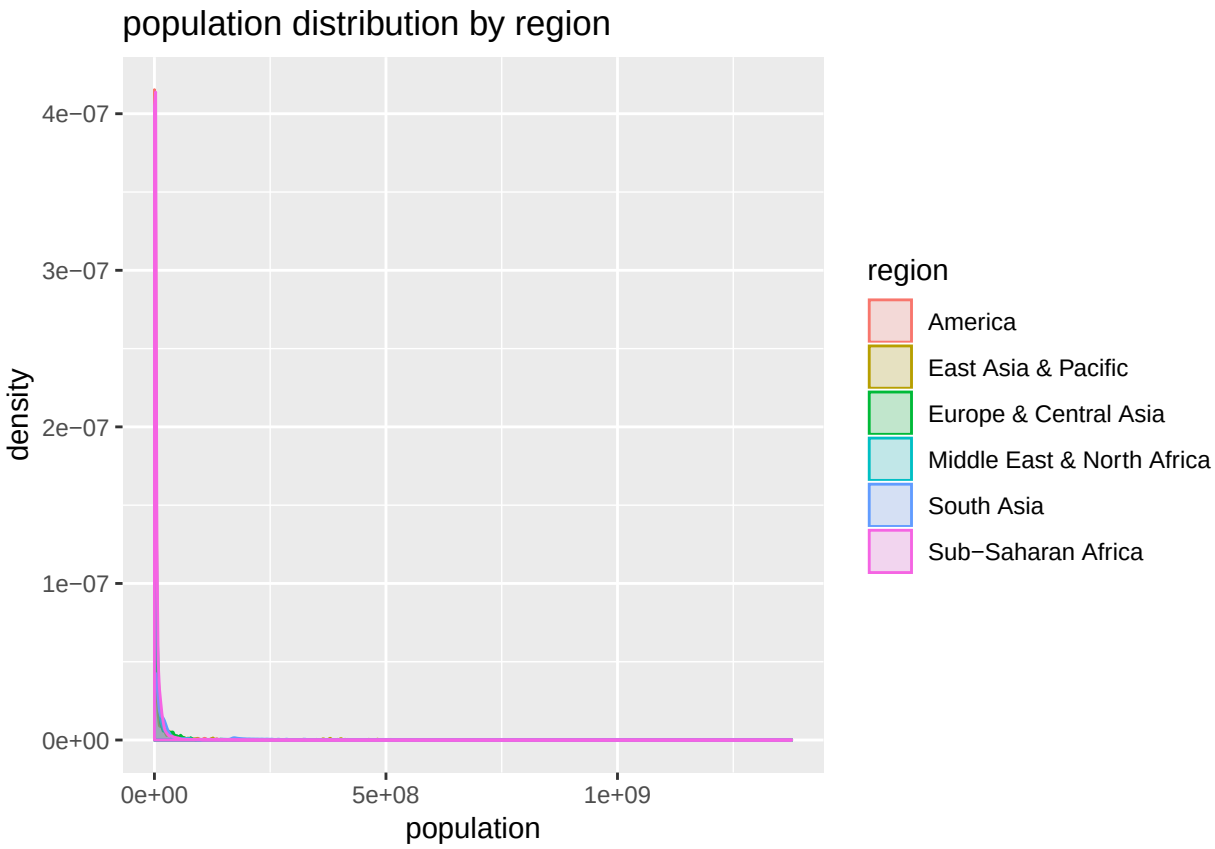
4.1 Can we predict region (using kmeans) if year, population, income and life are given?

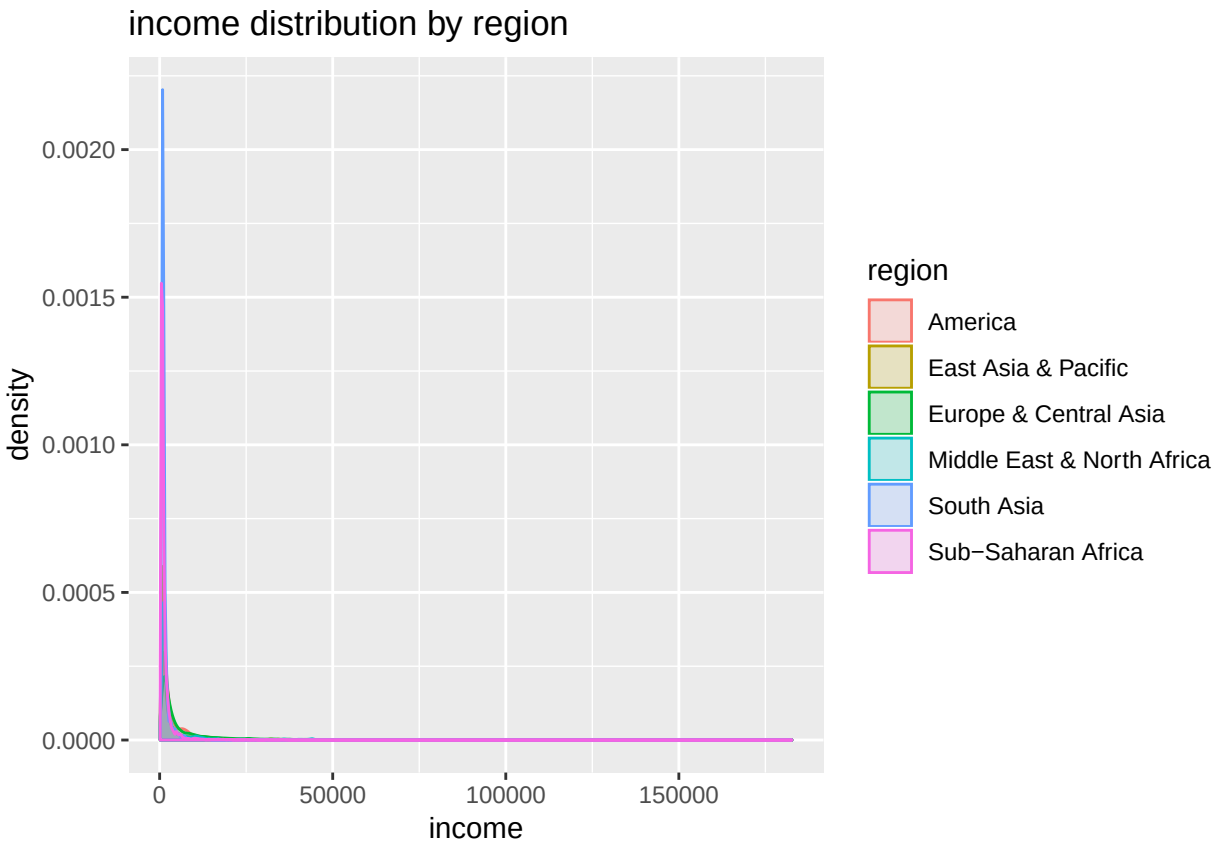
Another way to frame the same question is:

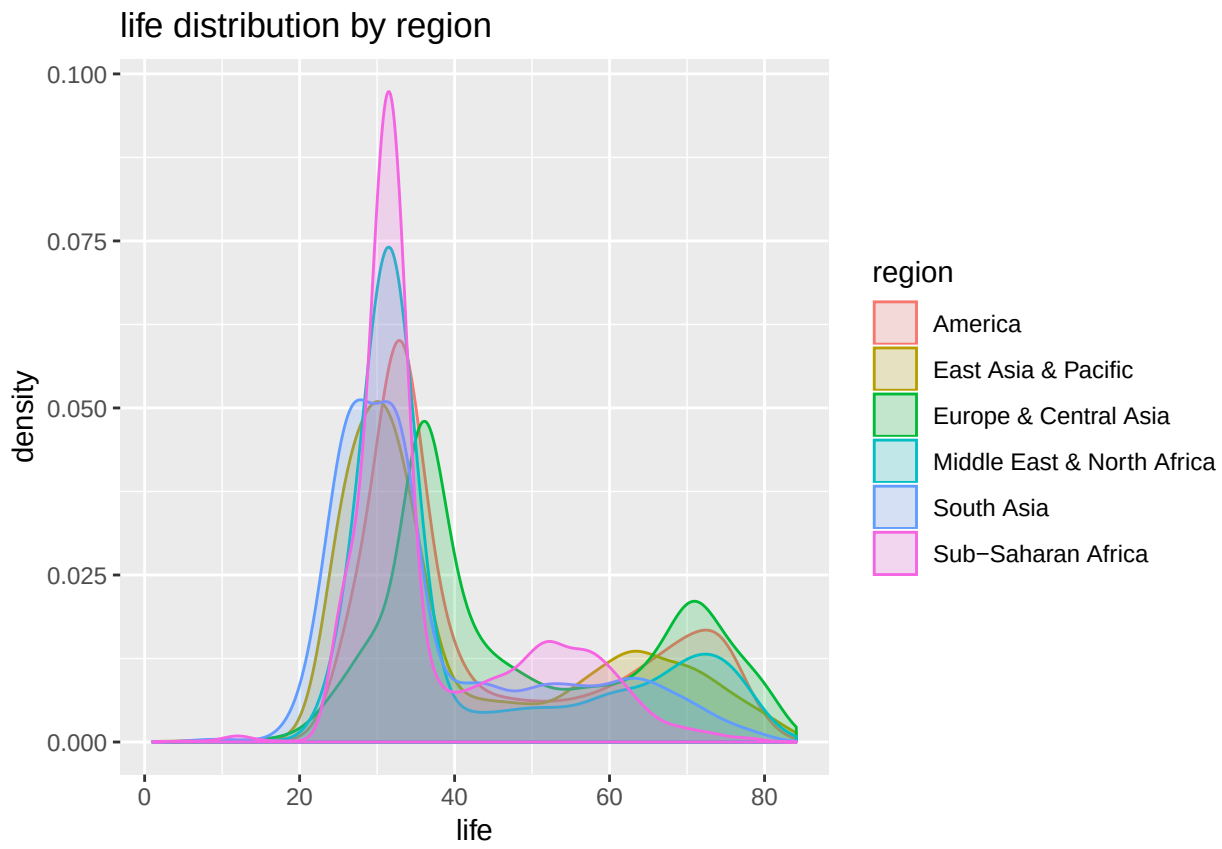
If the input variables are year, population, income and life, number of centers = 6, can kmeans accurately cluster the data into regions?

For this, We can perform some preliminary visualizations to see if there is any significant distinction between regions' data points.

First, let us see the distribution of each variable grouped by region. We have seen this in earlier section, but it was only for 2015. This time we need to see the same plot for all years.

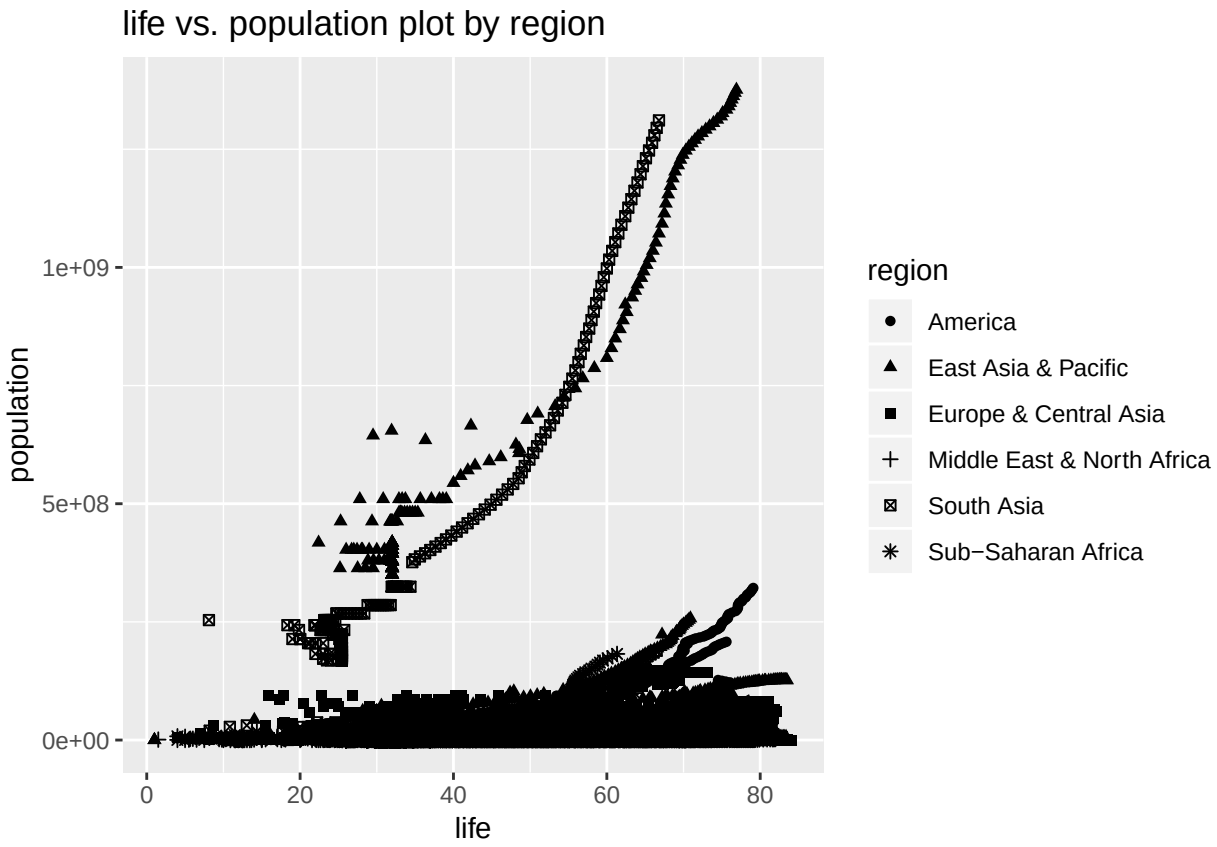

population distribution by region
density
4e−07
3e−07
2e−07
1e−07
0e+00
0e+00
5e+08
1e+09
population
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa

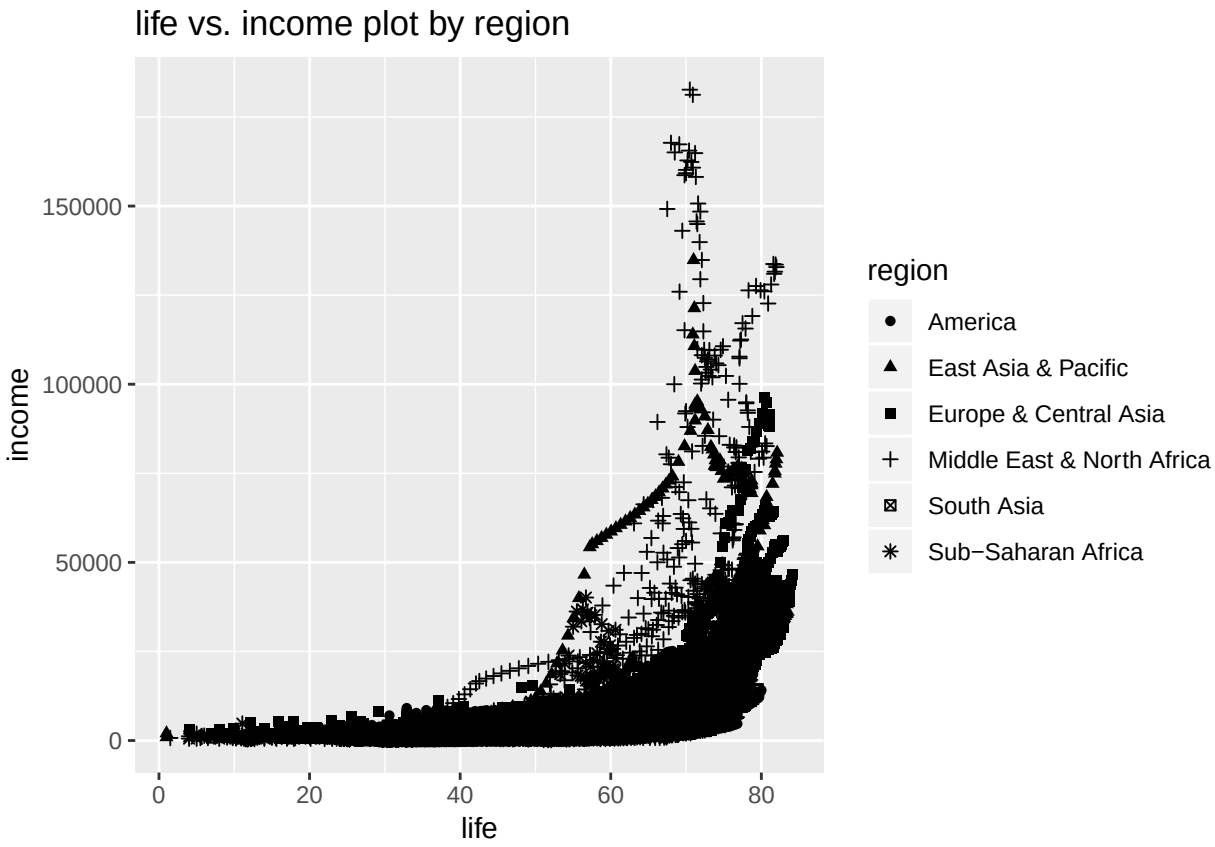

income distribution by region
density
0.0020
0.0015
0.0010
0.0005
0.0000
0
50000
100000
150000
income
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa

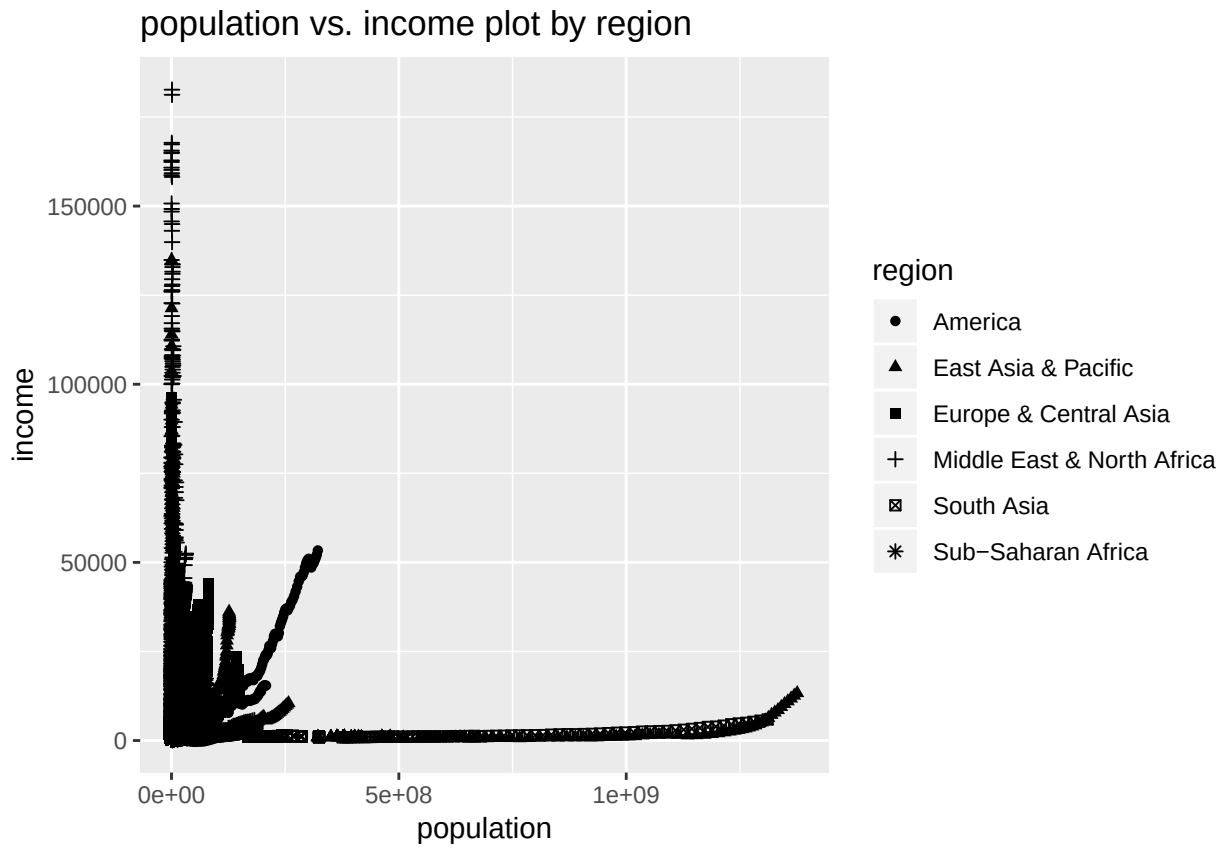


We can see from above 3 plots that there is no difference in the distribution between regions. This means that clustering the variables will not give any meaningful results.

It is, however, possible that the combined will produce meaningful clusters. Below, 3 scatter plots are shown, one for each pair combination of the numeric variables.

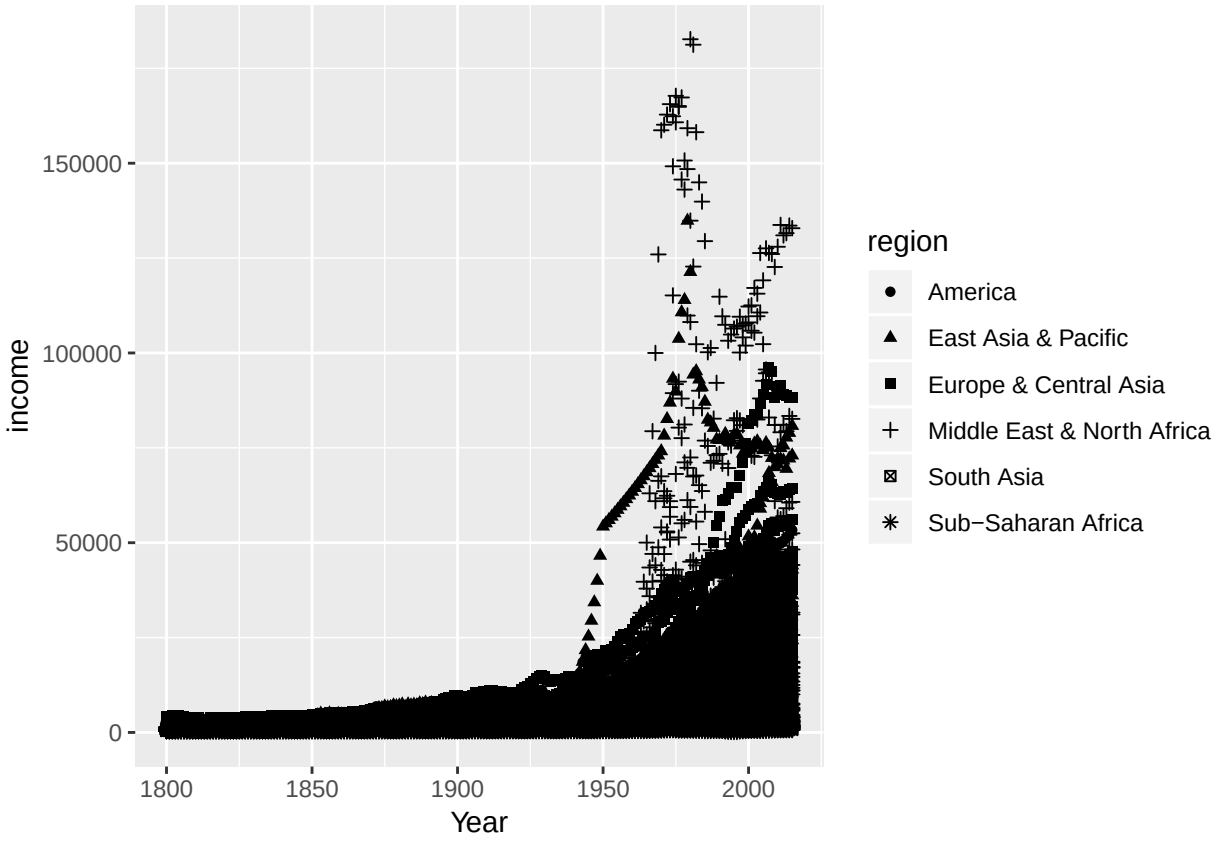

life vs. population plot by region
population
1e+09
5e+08
0e+00
0
20
40
60
80
life
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa

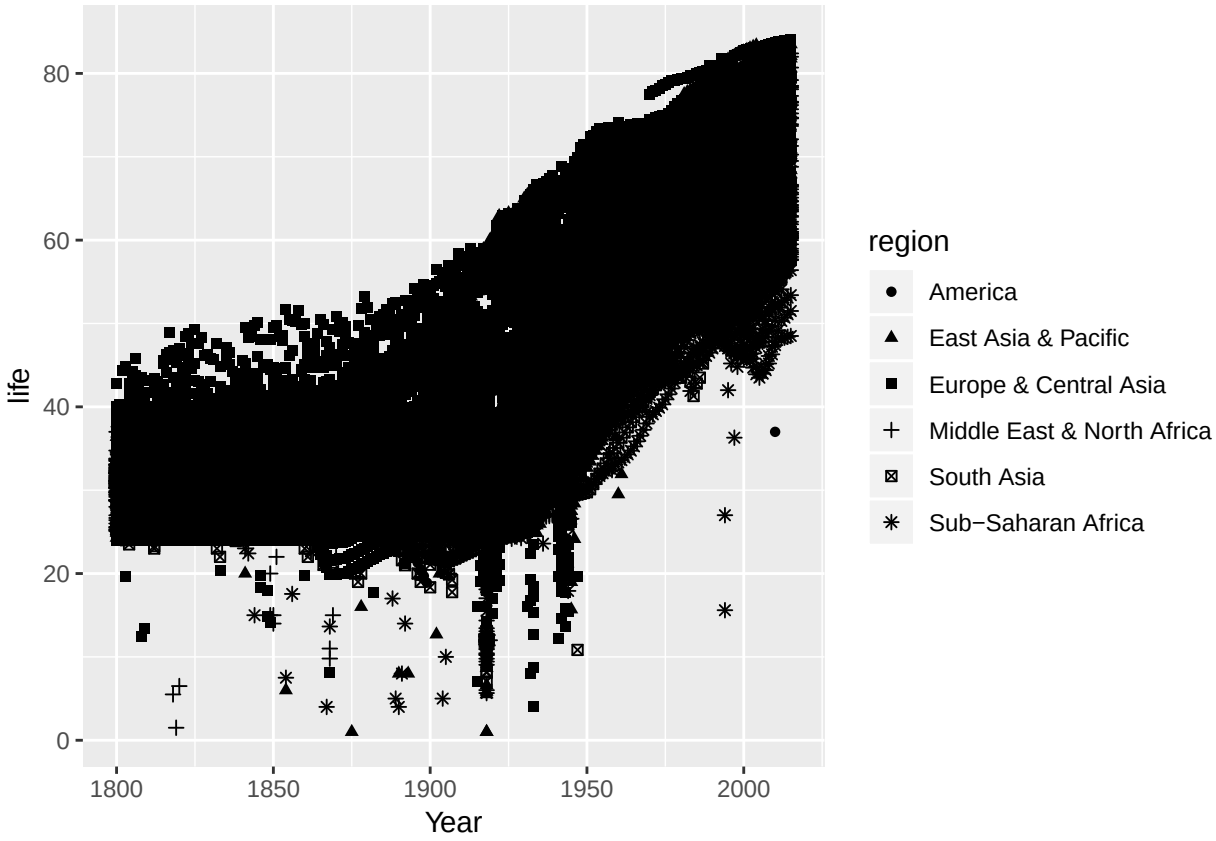

life vs. income plot by region
income
150000
100000
50000
0
0
20
40
60
80
life
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa

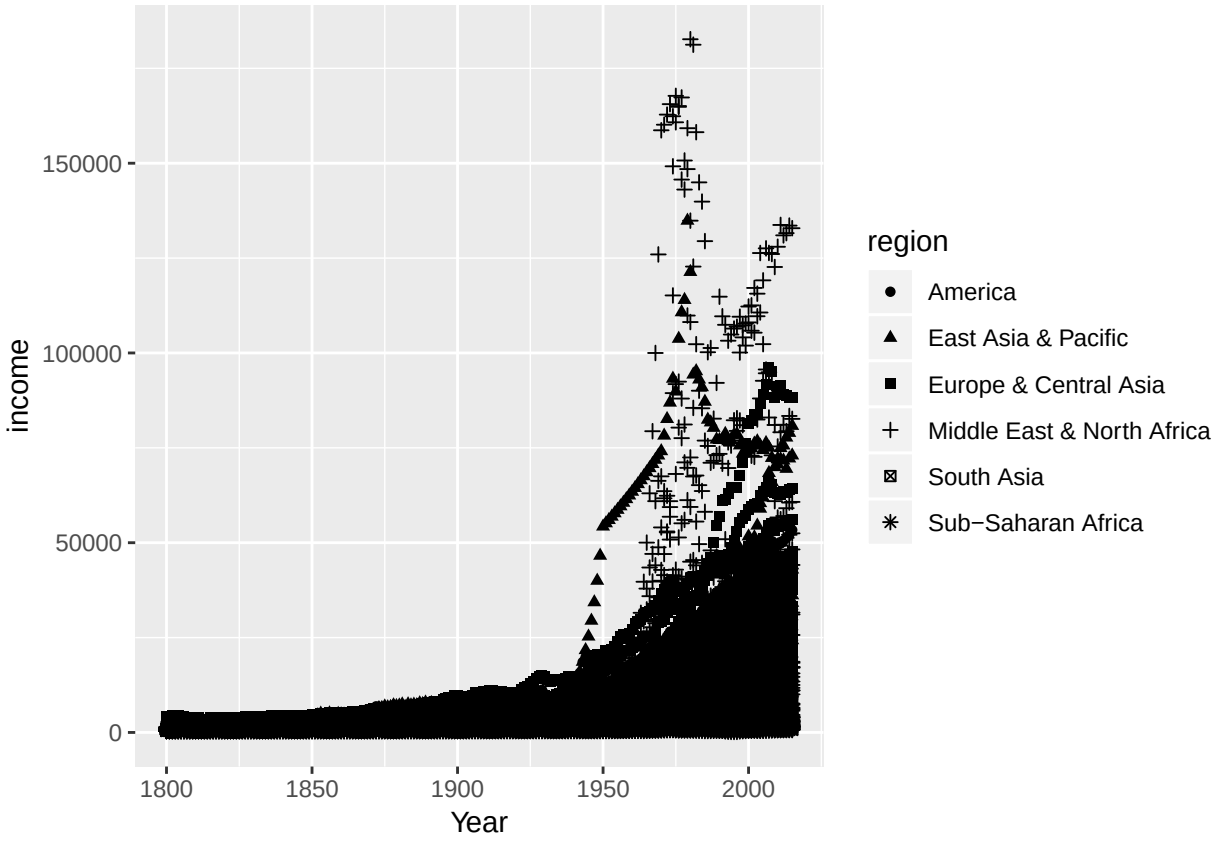


We see that no clear distinction between regions in the 2D space.

It can be We argued that we have disregarded the year variable. For this analysis, Year is better treated as as a categorical variable with 216 levels instead time or continuous variable. But let us look at bivariate plots with year on x axis anyway.


150000
100000
50000
0
income
1800
1850
1900
1950
2000
Year
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa


life
80
60
40
20
0
1800
1850
1900
1950
2000
Year
region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub−Saharan Africa



Once again, there is no spatial distinction for kmeans (or any distance based algorithm) to give meaningful results.

In conclusion, we should not use kmeans to cluster the data into regions.