Final Report

Tianzuo Huang

6/16/2019

```
# Question 1: Which region's income has the highest correlation to its life expectancy?
# Question 2: How does the life expectancy change year by year for China and US? Is there any abnormal
# Question 3: How many clusters is optimal for this dataset? What are the descriptive statistics for ea
```

```
str(gapminder)
```

```
## 'data.frame':    41284 obs. of  6 variables:
##  $ Country   : Factor w/ 197 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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: Factor w/ 15260 levels "","1,005,328,574",..: 7490 1 1 1 1 1 1 1 1 1 ...
##  $ income    : int  603 603 603 603 603 603 603 603 603 603 ...
##  $ region    : Factor w/ 6 levels "America","East Asia & Pacific",..: 5 5 5 5 5 5 5 5 5 5 ...
```

```
# The population data needs to be cleaned by removing the commas.
gapminder$population <- as.numeric(gsub(",", "", gapminder$population))
summary(gapminder)
```

```
##                 Country           Year           life
##  Afghanistan        :  216   Min.   :1800   Min.   : 1.00
##  Albania            :  216   1st Qu.:1854   1st Qu.:31.00
##  Algeria            :  216   Median :1908   Median :35.12
##  Angola             :  216   Mean   :1907   Mean   :42.88
##  Antigua and Barbuda:  216   3rd Qu.:1962   3rd Qu.:55.60
##  Argentina          :  216   Max.   :2015   Max.   :84.10
##  (Other)            :39988
##    population            income                                region
##  Min.   :1.548e+03   Min.   :   142   America                   : 7961
##  1st Qu.:5.335e+05   1st Qu.:   883   East Asia & Pacific       : 6256
##  Median :3.358e+06   Median :  1450   Europe & Central Asia     :10468
##  Mean   :2.119e+07   Mean   :  4571   Middle East & North Africa: 4309
##  3rd Qu.:1.078e+07   3rd Qu.:  3483   South Asia                : 1728
##  Max.   :1.376e+09   Max.   :182668   Sub-Saharan Africa        :10562
##  NA's   :25817       NA's   :2341
```

```
head(gapminder)
```

```
##       Country Year     life population income     region
## 1 Afghanistan 1800 28.21100    3280000    603 South Asia
## 2 Afghanistan 1801 28.20075         NA    603 South Asia
## 3 Afghanistan 1802 28.19051         NA    603 South Asia
## 4 Afghanistan 1803 28.18026         NA    603 South Asia
## 5 Afghanistan 1804 28.17001         NA    603 South Asia
## 6 Afghanistan 1805 28.15977         NA    603 South Asia
```

```
# Noticed some NAs in column population. How about other columns?
colSums(is.na(gapminder))
```

```
##    Country       Year       life population     income     region
##          0          0          0      25817       2341          0
```

```r
# Check which column has NA
gapminder <- gapminder[!is.na(gapminder$income), ]
# Only remove NAs in column income because NAs in population can be explained - census data collected a
length(unique(gapminder$Country))
```

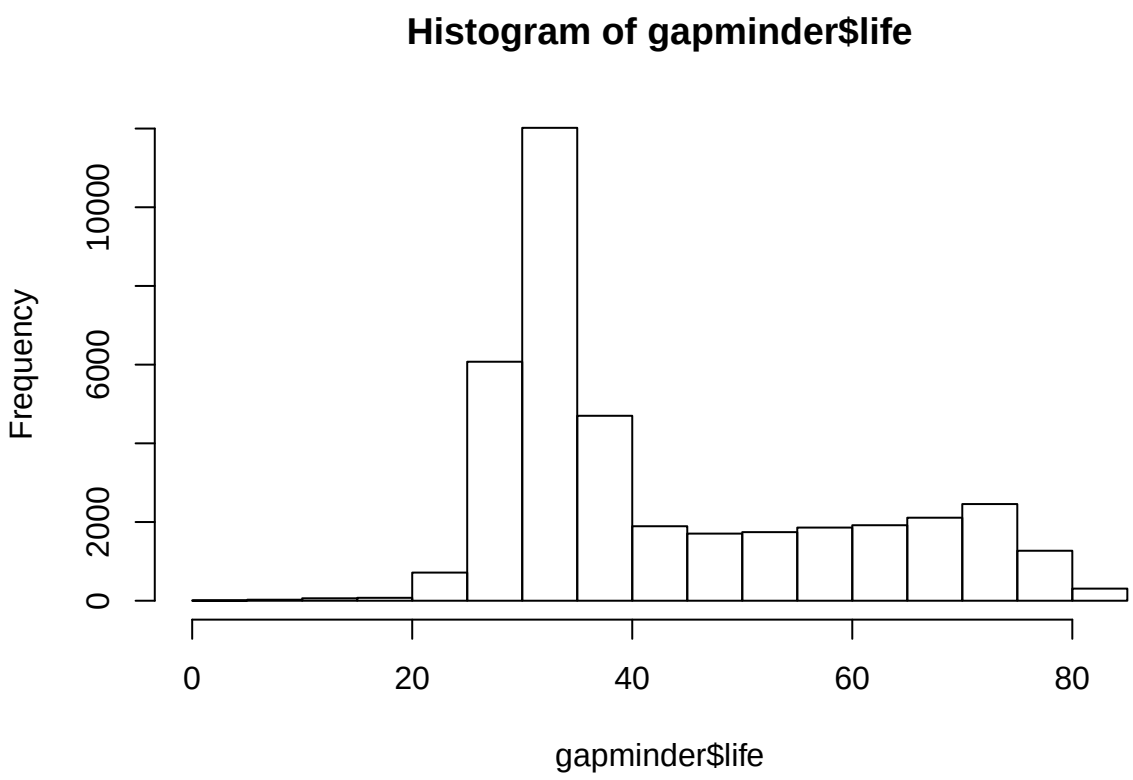
```
## [1] 183
```

We know that there are 183 unique countries and 6 regions in this dataset. The date range is between 1800 and 2015.

```r
# Check the observations for each state
table(gapminder$region)
```

```
##
##                    America        East Asia & Pacific
##                       6936                       5640
##      Europe & Central Asia Middle East & North Africa
##                      10383                       4104
##                 South Asia         Sub-Saharan Africa
##                       1728                      10152
```

```r
# Check the overall life expectancy and the median life expectancy divided by the region
hist(gapminder$life)
```

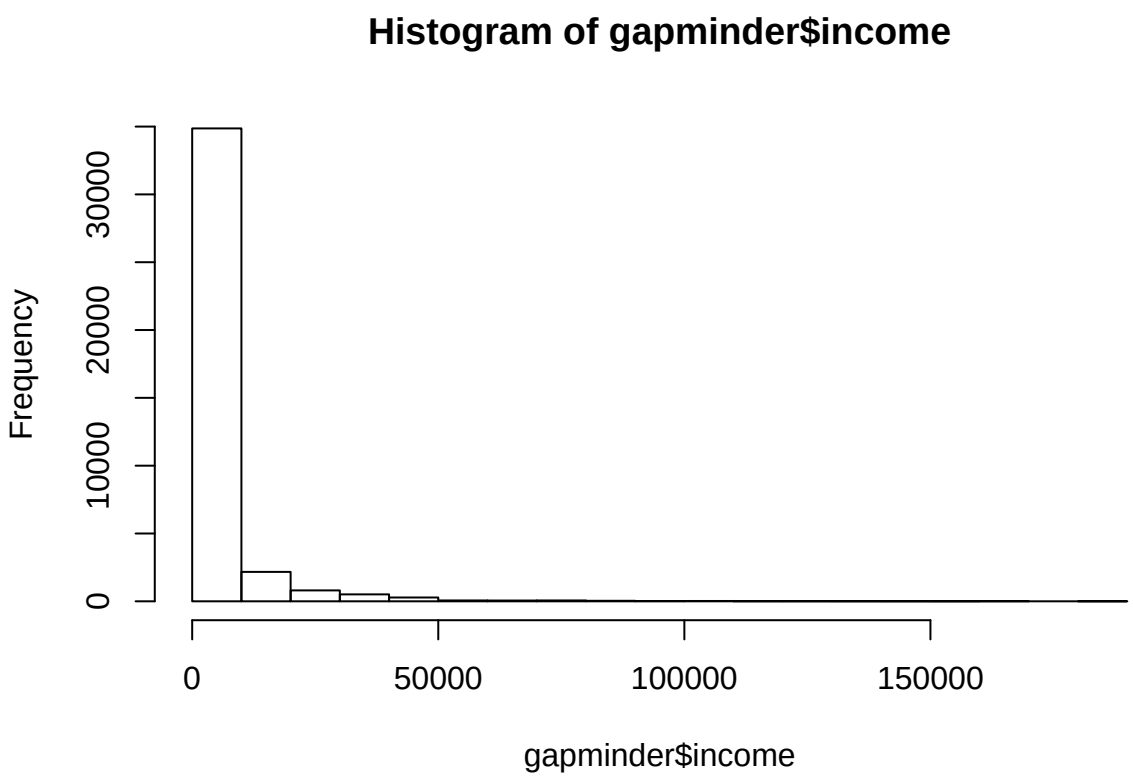



```r
gapminder_life <- aggregate(life ~ region, gapminder, median)
print(gapminder_life)
```

```
##                       region     life
## 1                    America 35.50000
## 2         East Asia & Pacific 34.05000
```

```
## 3      Europe & Central Asia 41.58855
## 4 Middle East & North Africa 32.30000
## 5                 South Asia 32.64700
## 6         Sub-Saharan Africa 32.30000
```

```
# Check the overall income and the median income divided by the region
hist(gapminder$income)
```




```
gapminder_income <- aggregate(income ~ region, gapminder, median)
print(gapminder_income)
```

```
##                      region income
## 1                   America 2214.5
## 2        East Asia & Pacific 1153.5
## 3      Europe & Central Asia 2735.0
## 4 Middle East & North Africa 1537.5
## 5                 South Asia 1020.0
## 6         Sub-Saharan Africa  827.0
```

```
gapminder_corr <- na.omit(gapminder[, c(2, 3, 4, 5)])
corrmatrix = cor(gapminder_corr, method = "pearson")
corrplot(corrmatrix, method = "number")
```


```
# We didn't see a strong correlation between overall income and life expectancy. Is that still the trut

income_America <- gapminder$income[gapminder$region == "America"]
life_America <- gapminder$life[gapminder$region == "America"]
cor(income_America, life_America)
```

```
## [1] 0.7068546
```

```
income_East_Asia <- gapminder$income[gapminder$region == "East Asia & Pacific"]
life_East_Asia <- gapminder$life[gapminder$region == "East Asia & Pacific"]
cor(income_East_Asia, life_East_Asia)
```

```
## [1] 0.5467629
```

```
income_Europe <- gapminder$income[gapminder$region == "Europe & Central Asia"]
life_Europe  <- gapminder$life[gapminder$region == "Europe & Central Asia"]
cor(income_Europe, life_Europe)
```

```
## [1] 0.7296936
```

```
income_Middle_East <- gapminder$income[gapminder$region == "Middle East & North Africa"]
life_Middle_East <- gapminder$life[gapminder$region == "Middle East & North Africa"]
cor(income_Middle_East, life_Middle_East)
```

```
## [1] 0.5754666
```

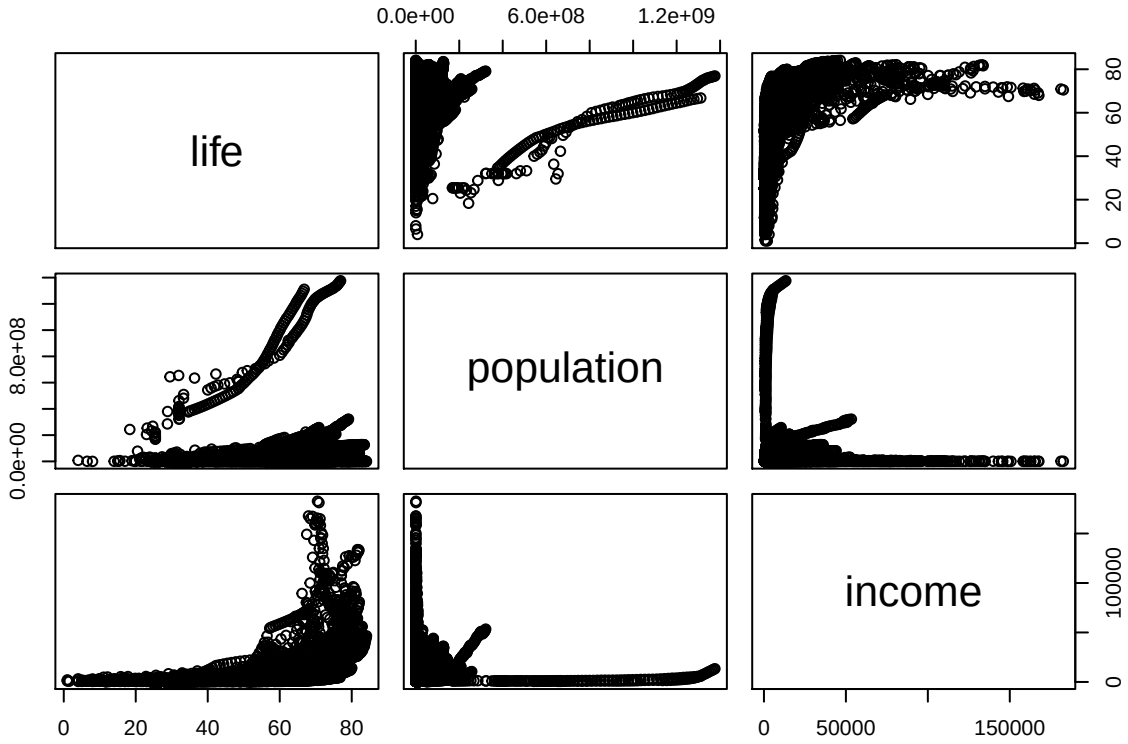
```
income_South_Asia <- gapminder$income[gapminder$region == "South Asia"]
life_South_Asia <- gapminder$life[gapminder$region == "South Asia"]
cor(income_South_Asia, life_South_Asia)
```

```
## [1] 0.6729226
```

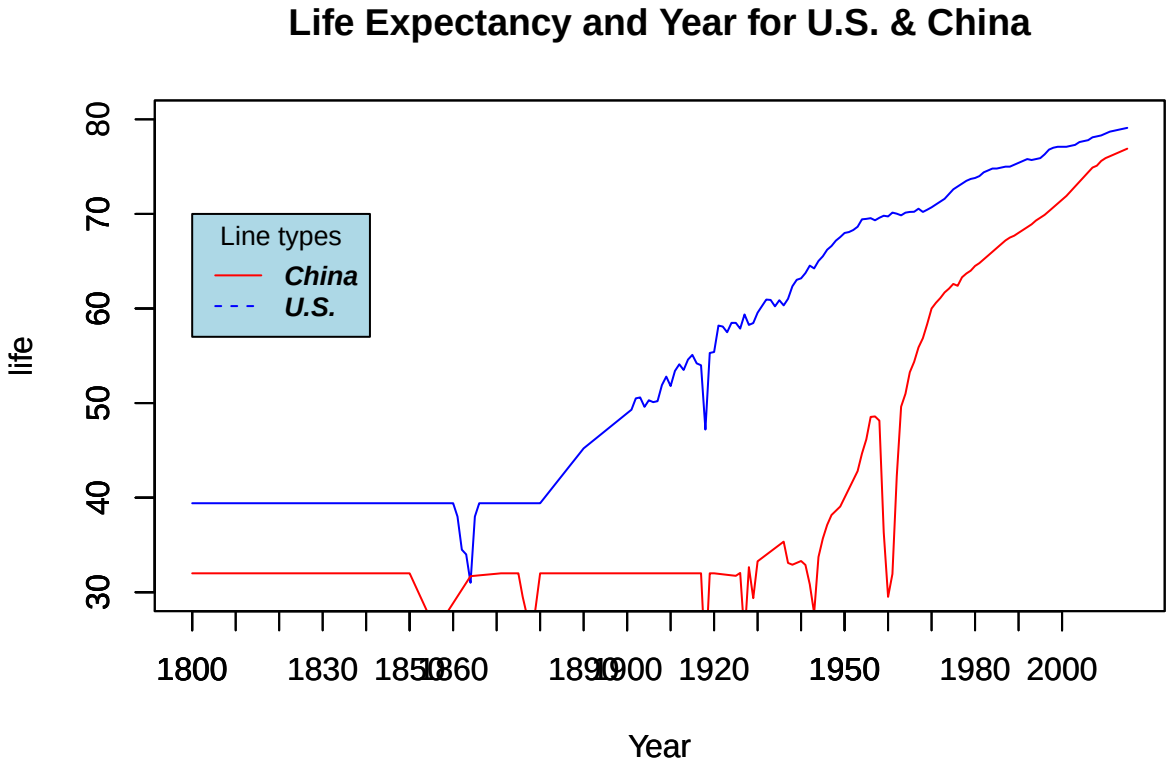
```
income_Africa <- gapminder$income[gapminder$region == "Sub-Saharan Africa"]
life_Africa <- gapminder$life[gapminder$region == "Sub-Saharan Africa"]
cor(income_Africa, life_Africa)
```

```
## [1] 0.4886287
```

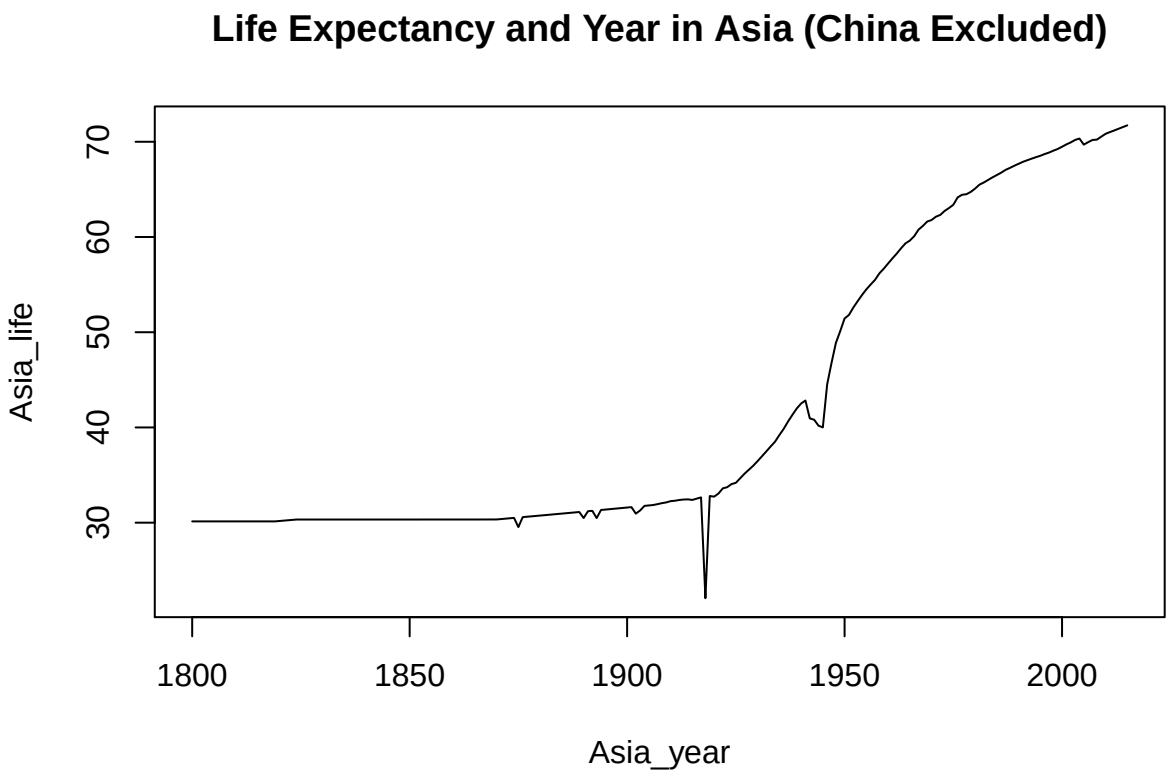
```
plot(gapminder[, 3:5])
```




```
## Didn't get too much valuable information. I need to pick up a specific region or country
plot(life ~ Year, gapminder, subset = Country == "United States",type = "l", col = "Blue", main = "Life
axis(1, seq(1800, 2000, 10))
axis(2, seq(30, 70, 10))
par(new=TRUE)
plot(life ~ Year, gapminder, subset = Country == "China" ,type = "l", col = "Red", ylim = c(30, 80))
axis(1, seq(1800, 2000, 10))
axis(2, seq(30, 70, 10))
legend(1800, 70, legend=c("China", "U.S."),
       col=c("red", "blue"), lty=1:2, cex=0.8,
       title="Line types", text.font=4, bg='lightblue')
```



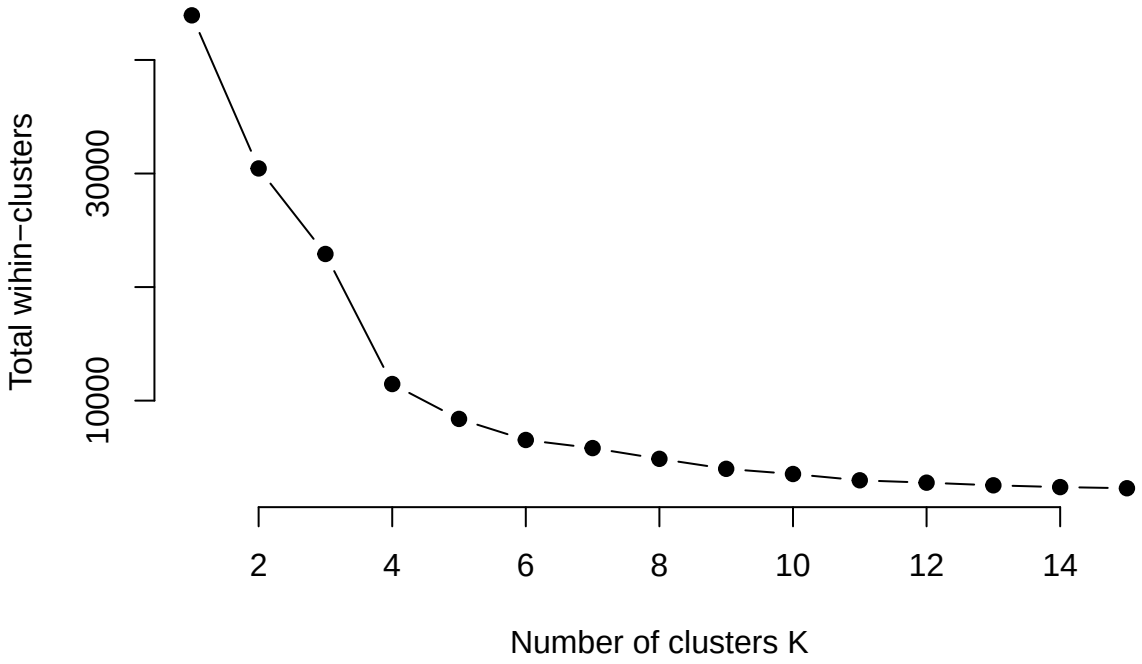

```
# We noticed that China experienced a big drop in life expectancy around 1960. Did this situation happe
gapminder_asia <- gapminder[gapminder$region == "East Asia & Pacific" & !gapminder$Country == "China",]
Asia_year <- c(min(gapminder_asia$Year):max(gapminder_asia$Year))
Asia_life = Asia_year
for (i in c(1:length(Asia_year))){
  tmp = gapminder_asia[gapminder_asia$Year == Asia_year[i],3]
  Asia_life[i]= mean(tmp)
}
plot(Asia_year, Asia_life, type ="l", main = "Life Expectancy and Year in Asia (China Excluded)")
```



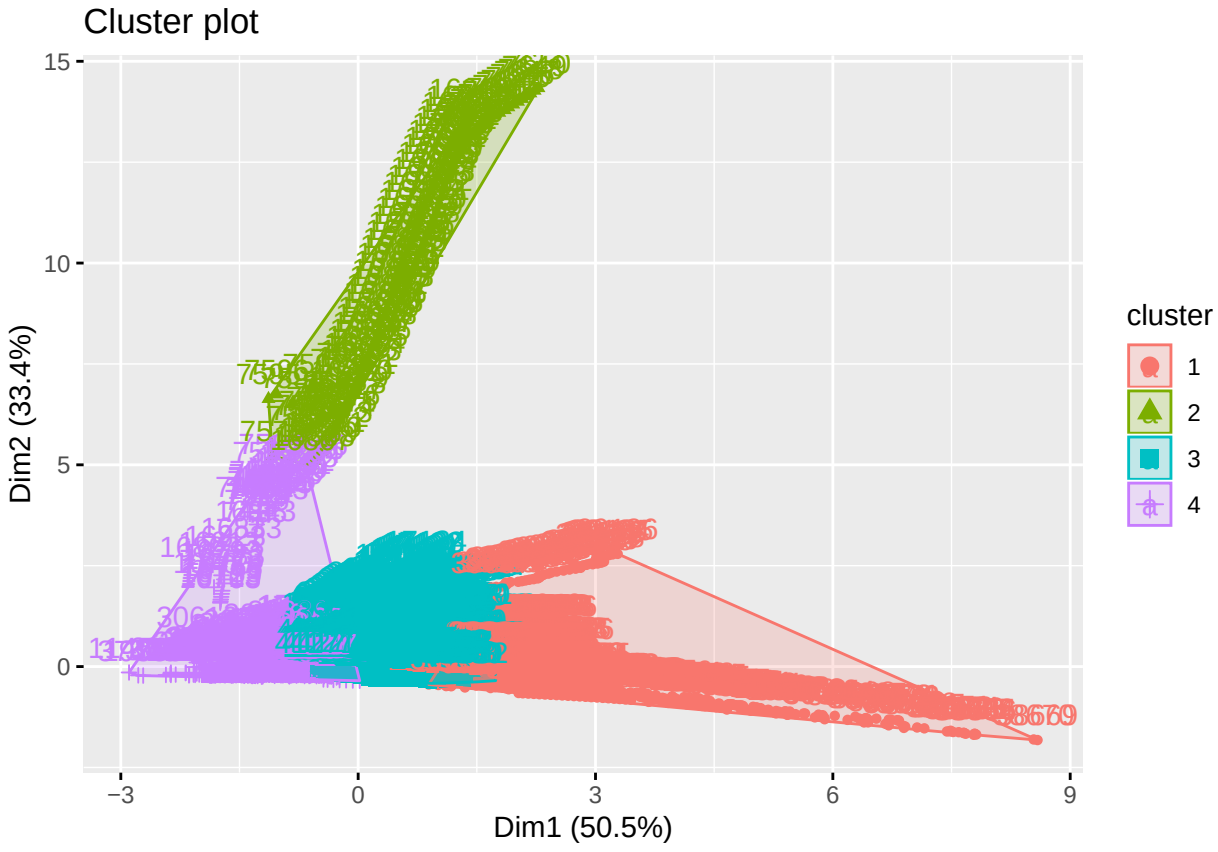

```
gapminder_cluster <- na.omit(gapminder[, c(3, 4, 5)])
gapminder_cluster_scale <- scale(gapminder_cluster)
wss <- function(k){
  kmeans(gapminder_cluster_scale, k, nstart = 10)$tot.withinss
}
k.values <- 1:15
wss_values <- map_dbl(k.values, wss)
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 732200)
```

```
plot(k.values, wss_values, type = "b", pch = 19, frame = FALSE, xlab = "Number of clusters K", ylab = "
```




```
# With the ELbow method, it suggests 4 as the number of optimal clusters. We can perform the final analy
final <- kmeans(gapminder_cluster_scale, 4, nstart = 25)
fviz_cluster(final, data = gapminder_cluster_scale)
```




```
# Visualize the results by using fviz_cluster
gapminder_cluster %>%
```

```
  mutate(Cluster = final$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")
```

```
## # A tibble: 4 x 4
##   Cluster  life population income
##     <int> <dbl>      <dbl>  <dbl>
## 1       1  76.6  22437516. 46697.
## 2       2  59.5 938949293.  2567.
## 3       3  66.2  17215045.  8162.
## 4       4  39.7   9096633.  1665.
```

```
# We can extract the clusters and add to our initial data to do some descriptive statistics at the clus
```

With the ELbow method, it suggests 4 as the number of optimal clusters.

From the statistics above, we can see that the 2nd cluster has the largest popution. The 4th cluster has the highest income.

```
### In Europe & Central Asia and America, the correlation between income and life expectancy is higher
### For the second question, since we didn't see a similar drop for other countries in East Asia & Paci
```

```
# Quick R: Cluster Analysis
# https://www.statmethods.net/advstats/cluster.html
# UC Business Analytics R Programming Guide: Dealing with Missing Values
# http://uc-r.github.io/missing_values
# UC Business Analytics R Programming Guide: K-means Cluster Analysis
# https://uc-r.github.io/kmeans_clustering
# Great Chinese Famine
# https://en.wikipedia.org/wiki/Great_Chinese_Famine
```