

GapMinder_Analysis

Haneef Meeran
6/16/2019

Table of Contents

[LINK TO OUTPUT FILE] (<http://rpubs.com/haneefmrn/EDAGapminder>)

Question

To measure the Income and Life expectance rate in different Geographic locations

Data Description

Country describes the primary variable describes the location of the life expectancy measured

Region describes the continent related to the country

Year - Provided for each country from 1800 to 2015

Life - Provided for each country by year

Population - Data available for each starting decade of each country

Income - Continuous data series for each country by year

Income - Gross domestic product per person adjusted for differences in purchasing power (GDP/capita, PPP\$ inflation adjusted)

Life - The average number of years a newborn child would live if current mortality pattern were to stay the same

population - Population for the given country in the given year (Source: [GapMinder](#))

Data Preparation

Set working directory

Import the Raw data

Adding Required libraries

Summary of Data

Explore Data Analysis

Set working directory and Import raw data

```
setwd("~/Desktop/EXPL DATA ANALYSIS/EDA REPORT")
```

```
gapminder <- read.csv("gapminder.csv")
```

Required Libraries

```
library(dplyr)
```

```
library(ggplot2)
```

```
library(matrixStats)
```

```
library(scales)
```

Summary of data

#Summary of Gapminder

```
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
##      :25817  Min.   : 142  America      : 7961
## 121000 :    6  1st Qu.: 883  East Asia & Pacific : 6256
## 14092  :    6  Median : 1450  Europe & Central Asia :10468
## 1432000:    6  Mean    : 4571  Middle East & North Africa: 4309
## 229000 :    6  3rd Qu.: 3483  South Asia      : 1728
## 2574000:    6  Max.    :182668  Sub-Saharan Africa :10562
## (Other):15437  NA's    :2341
```

```
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 ...
## $ region       : Factor w/ 6 levels "America", "East Asia & Pacific",...: 5 5 5 5 5 5 5 5 5 ...
```

Here, we study that Country, population, and income are Factors. Life and income are of num data type while year is of int datatype. We notice that population is a number but due to commas, it is turned a factor. hence we clean the data by removing the commas and creating another field as numeric. ### Study the scope

#Number of Unique Countries

```
length(unique(gapminder$Country))
```

```
## [1] 197
```

#Cleaning of population field. Removing commas and reloading as numeric in a different dataset without NAs.

```
gapminder$population1 <- as.numeric(gsub(",", "", gapminder$population))
```

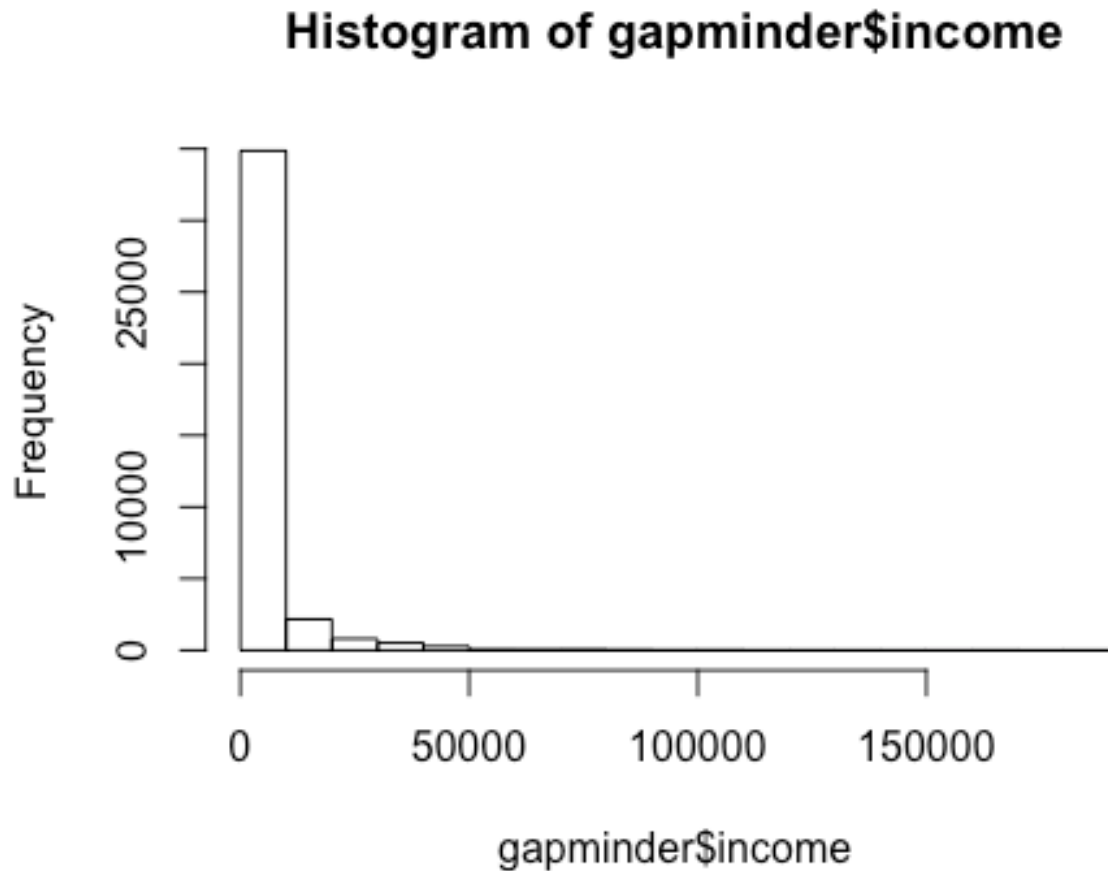
```
gapminder1 <- gapminder %>% filter(!is.na(gapminder$population1))
```

Looking at the summary, now we know there are 197 countries spread across 6 regions, each having 216 entries for years from 1800 to 2015. The life expectancy ranges from 1 to 84%

Analysis

Income

```
hist(gapminder$income)
```



Looking at the histogram, a high frequency of *income* lay in the range below 50,000.

Specifically in 10,000 and reduces as we move towards 50,000.

```
summary(unique(gapminder$income))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's  
##      142    2978    6651   12397   14744  182668         1
```

```
weightedMedian(gapminder$income, w = gapminder$region, na.rm = TRUE)
```

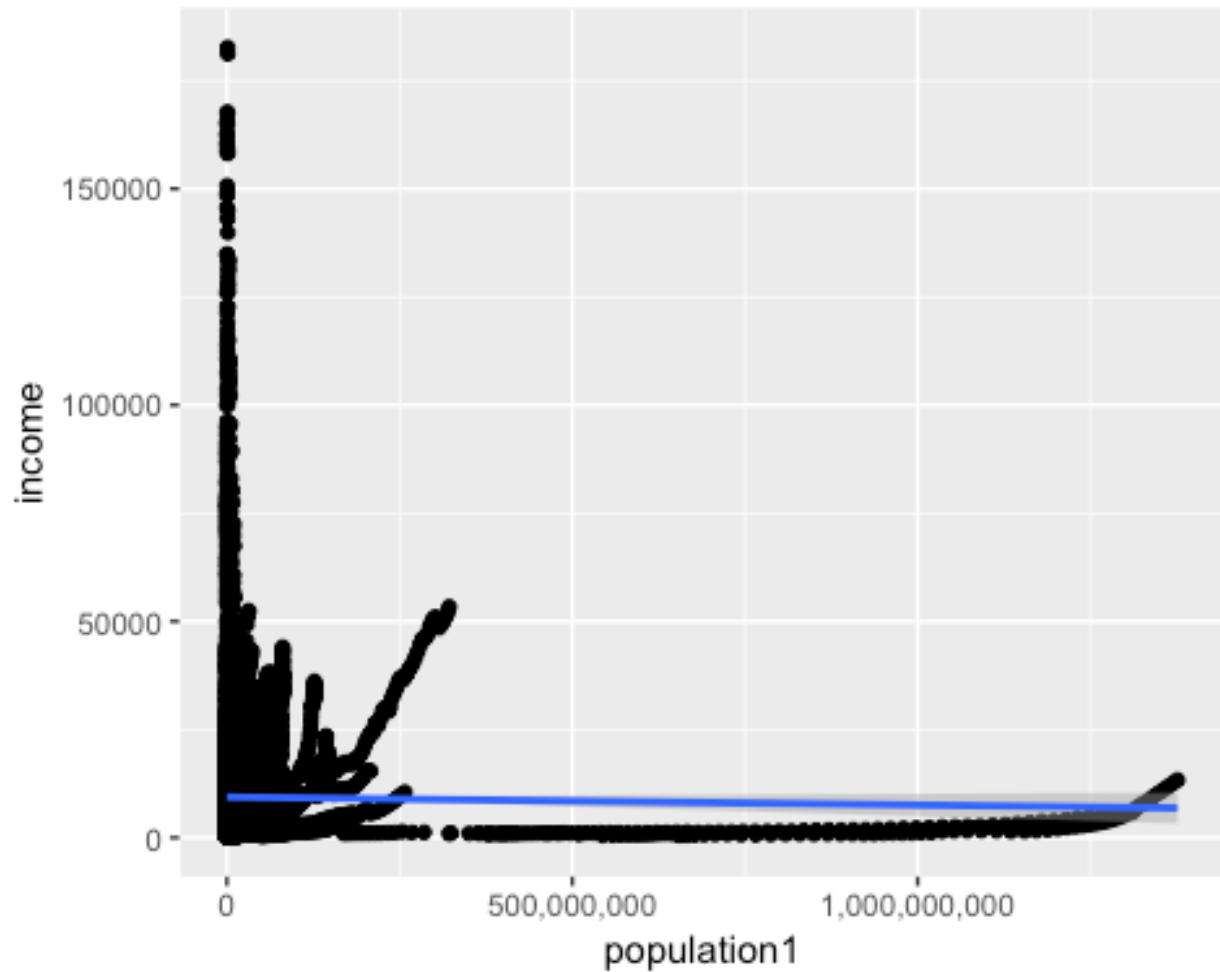
```
## [1] 1182
```

```
weightedMean(gapminder$income, w = gapminder$region, na.rm = TRUE)
```

```
## [1] 3900.356
```

```
d <- ggplot(gapminder1, aes(x = population1, y = income))
```

```
d + geom_point() + geom_smooth(method = "lm") + scale_x_continuous(labels = scales::comma)
```



Further, when we look at the summary range, we find that the mean and the median lies in the 2nd quadrant with minimum value of 142 and maximum of 182668.

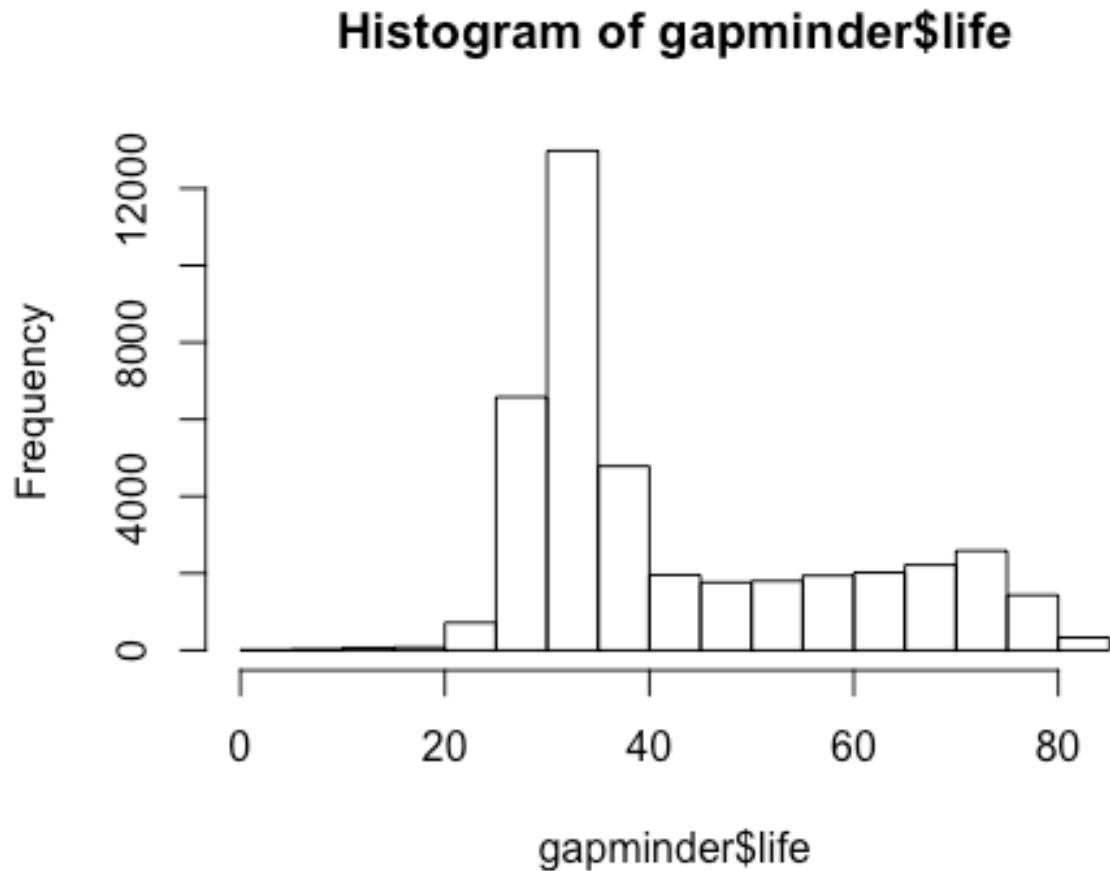
However, things change drastically, when we look at the weighted median and weighted mean based on the *region*. the Median reduces from **6651** to **1457**, while the Mean reduces from **12397** to ****4631**. *We can say that the income* ranges are different in different regions.*

From the plot above (population vs income), we can also find that as the population increases, the income hardly increases for most, while is linear for some. However, if the population is lower always, the income increases exponentially.

###Life Expectancy

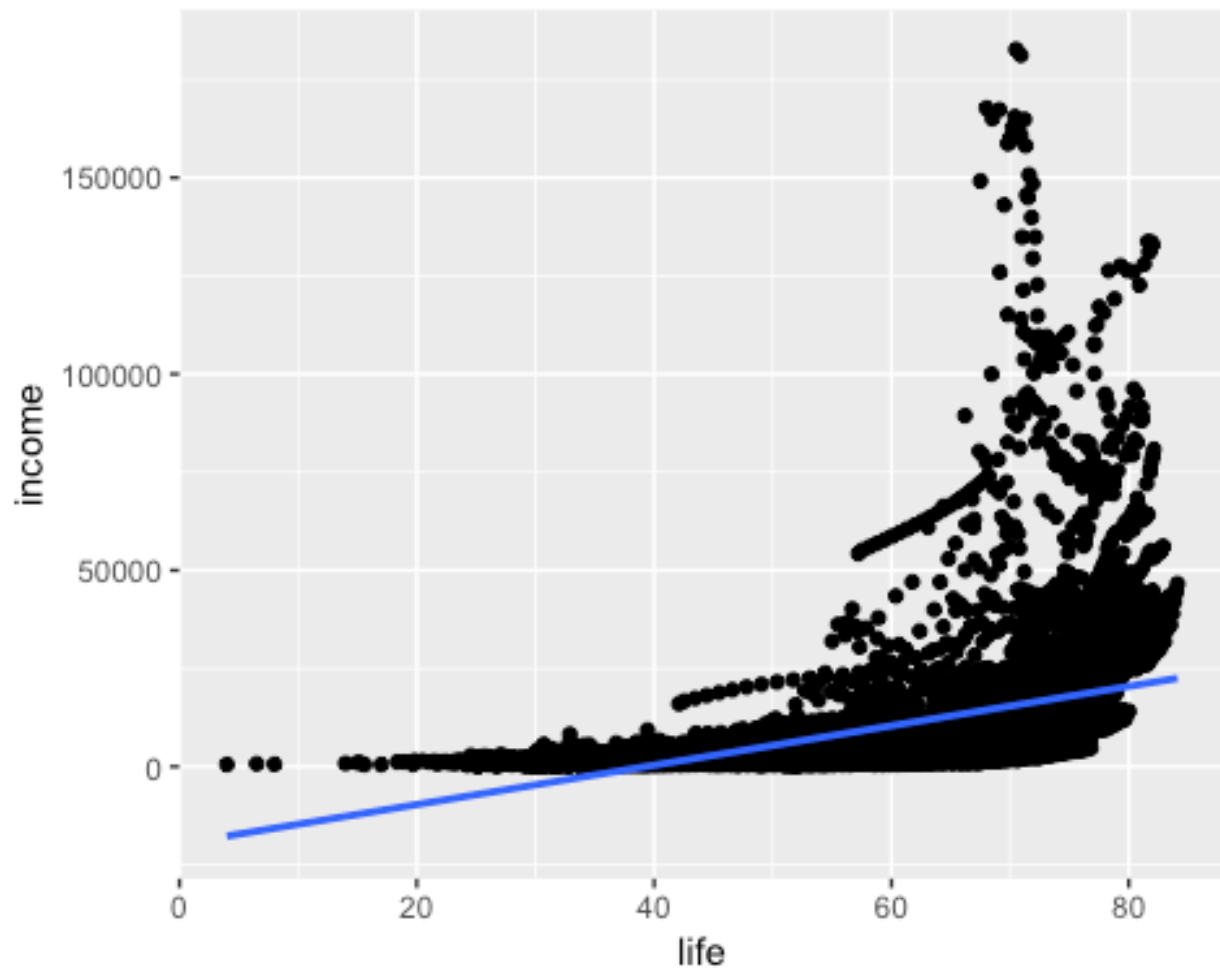
Let's look at the frequencies of the life expectancy and income ranges for the complete data

```
hist(gapminder$life)
```



The Histogram shows a very high life expectancy starting at 20, the highest being 20 to 40 and an average income of about 2000 as we observe towards 80.

```
summary(gapminder$life)
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.00  31.00   35.12   42.88  55.60   84.10
weightedMedian(gapminder$life, w = gapminder$region, na.rm = TRUE)
## [1] 33.86702
weightedMean(gapminder$life, w = gapminder$region, na.rm = TRUE)
## [1] 41.45024
e <- ggplot(gapminder1, aes(x = life, y = income))
e + geom_point() + geom_smooth(method = "lm") + scale_x_continuous(labels = scales::comma)
```



There seems to be a high frequency of *life expectancy* in the range of **25-35 years**. In the next graph, we can see that the life expectancy grows regardless of the income growth. However, as the life expectancy grows past 60, there is a big growth in income.

Population

```
median(gapminder1$population1)
```

```
## [1] 3358089
```

```
mean(gapminder1$population1)
```

```
## [1] 21187964
```

```
weightedMedian(gapminder1$population1, w = gapminder1$Country, na.rm = TRUE)
```

```
## [1] 3474657
```

```
weightedMean(gapminder1$population1, w = gapminder1$Country, na.rm = TRUE)
```

```
## [1] 19078834
```

The weighted mean is much lower than that of the actual mean when Countries are considered. This means that there is a big difference between the population of some countries.

```
p <- ggplot(gapminder1, aes(reorder(region, population1, FUN=function(x) mean(log10(x))),
population1))
```

```
p <- p + scale_y_log10()
```

```
p + geom_boxplot(outlier.colour="red") + geom_jitter(alpha=1/2)
```

Working with Country data

Comparison between India and China

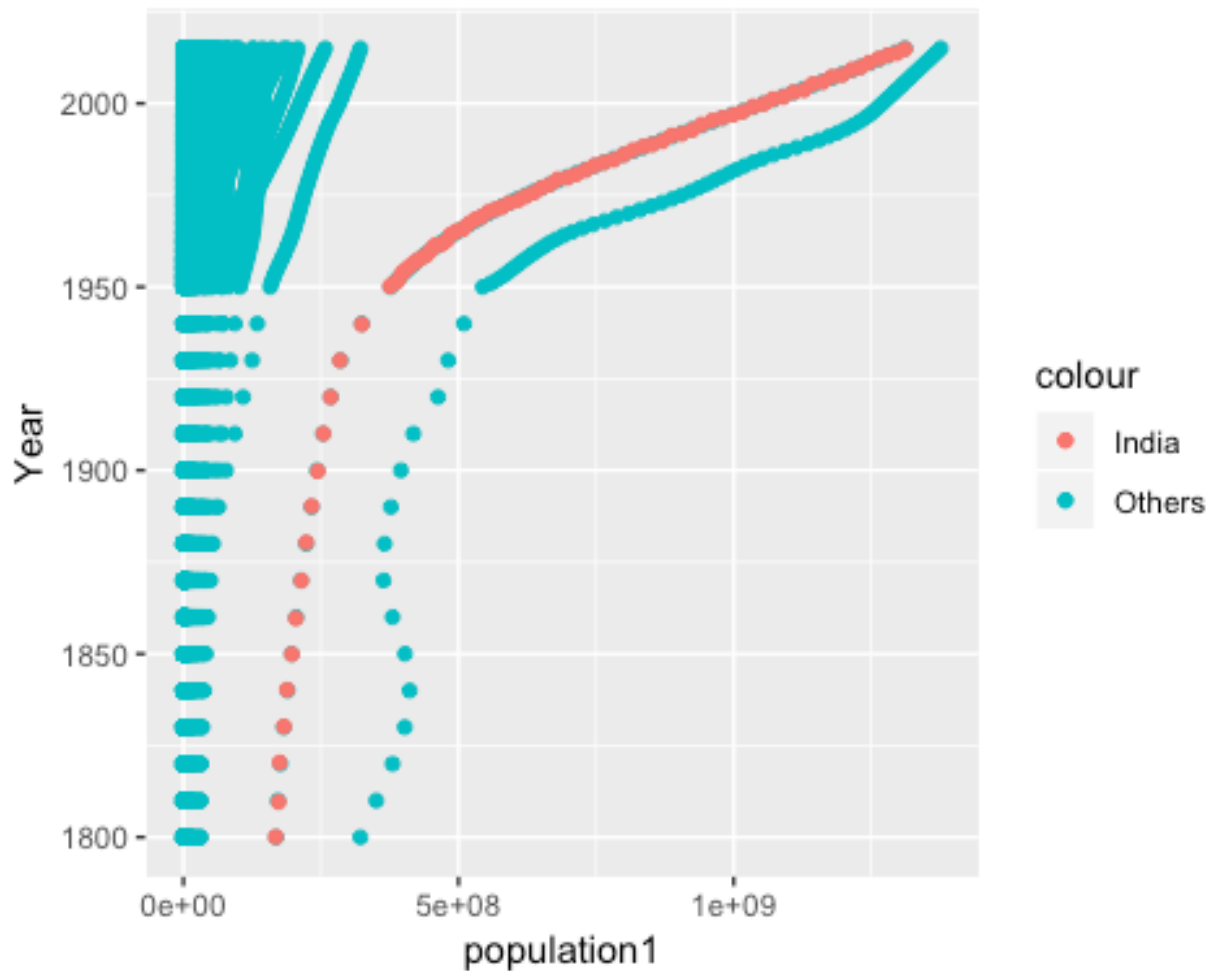
Creating subset of India Country

```
population_india <- gapminder1 %>% filter(Country=='India')
```

Creating subset of India Country

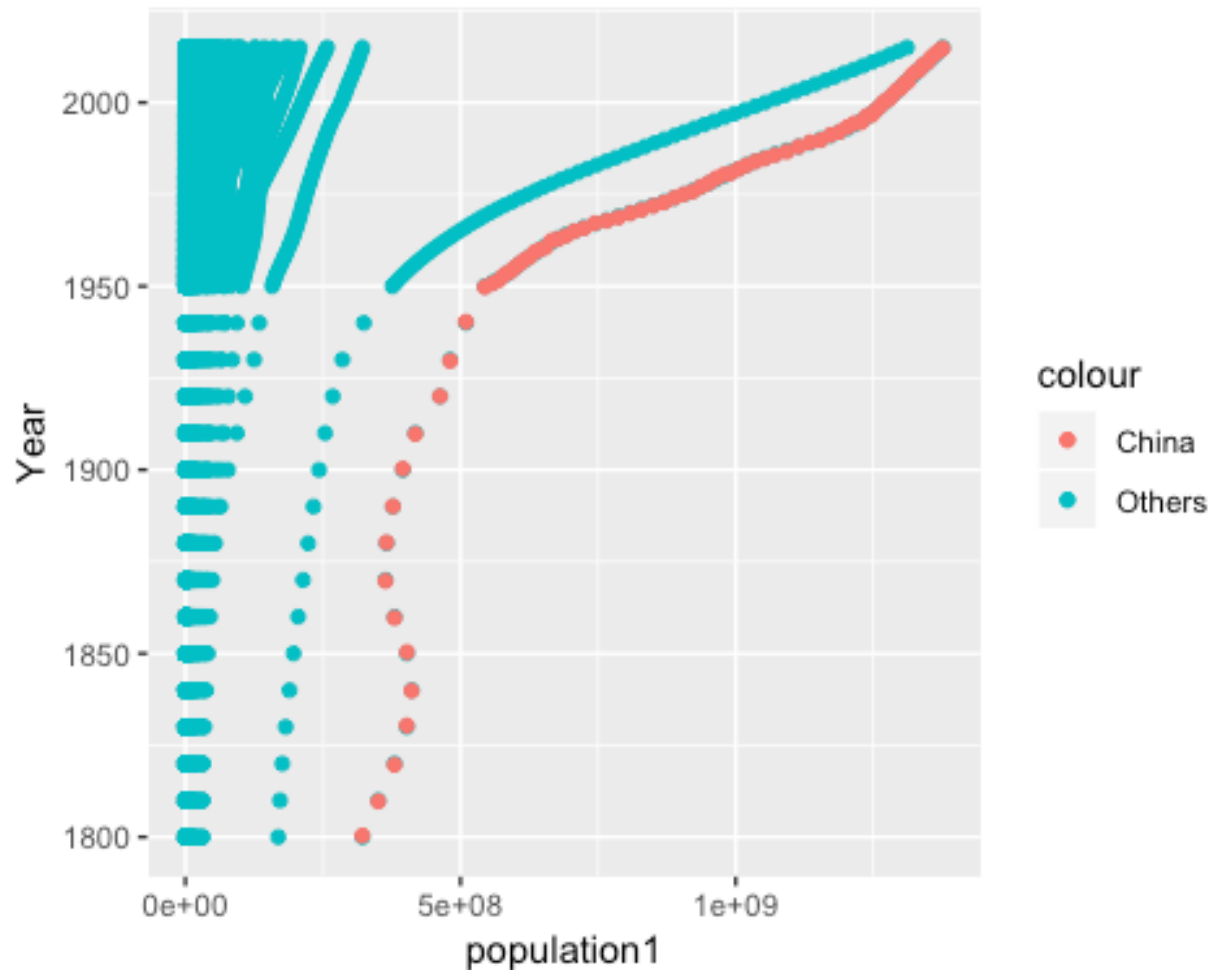
```
population_china <- gapminder1 %>% filter(Country=='China')
```

```
ggplot(data= population_india, mapping  
= aes(y=Year, x=population1, color="India")) +geom_point(data  
= gapminder1, aes(color="Others")) + geom_jitter()
```



Population observations between India and China on the rise starting from 1800, lets perform further analysis.

```
ggplot(data= population_china, mapping = aes(y=Year, x=population1, color="China")) +geom_point(data  
= gapminder1, aes(color="Others")) + geom_jitter()
```



The population in India and China has risen exponentially after 1950 as compared to other countries.

India and China after 1950

```
gapminder1 %>% #Filtering data for India
```

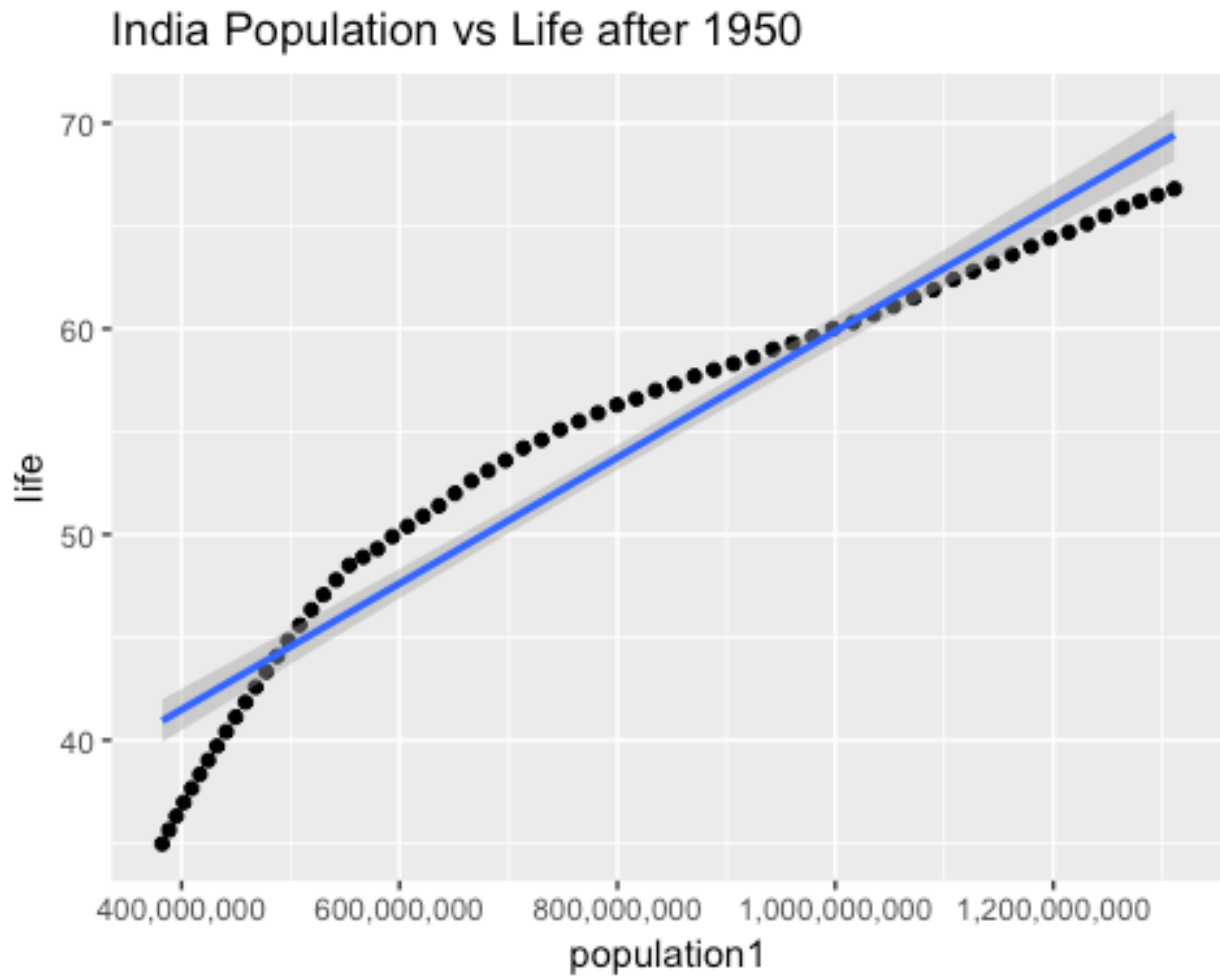
```
filter(Country=='India') %>%
```

```
summarize('LifeExpected (Mean) in India'=mean(population1),
'Population (Standard Deviation) in India'=sd(population1))
## LifeExpected (Mean) in India Population (Standard Deviation) in India
## 1 675758566 343006329
```

```
#Filtering data for India after 1950
```

```
pop_ind_after_1950 <- population_india %>% filter(Year > '1950')
```

```
b <- ggplot(pop_ind_after_1950, aes(x = population1, y = life))
b + geom_point()+geom_smooth(method = "lm") + scale_x_continuous(labels
= scales::comma) + labs(title="India Population vs Life after 1950")
```

The graph above shows the Indian population vs Life, as we can see the age and population are increasing almost linearly, the population appears to be slightly high in age group 45 to 55 and steady increase, followed by a slight decrease after 65.

```
c <- ggplot(pop_ind_after_1950, aes(x = population1, y = income))

c + geom_point() + geom_smooth(method = "lm") + scale_x_continuous(labels
= scales::comma) + labs(title="India Population vs Income after 1950")
```



The population vs Income graph is interesting, a large amount of the population has an income between 1000 to 3000.

```
gapminder1 %>%filter(Country=='China') %>%#Filtering data for India

summarize('LifeExpected (Mean) in China'=mean(population1),
'Population (Standard Deviation) in China'=sd(population1))

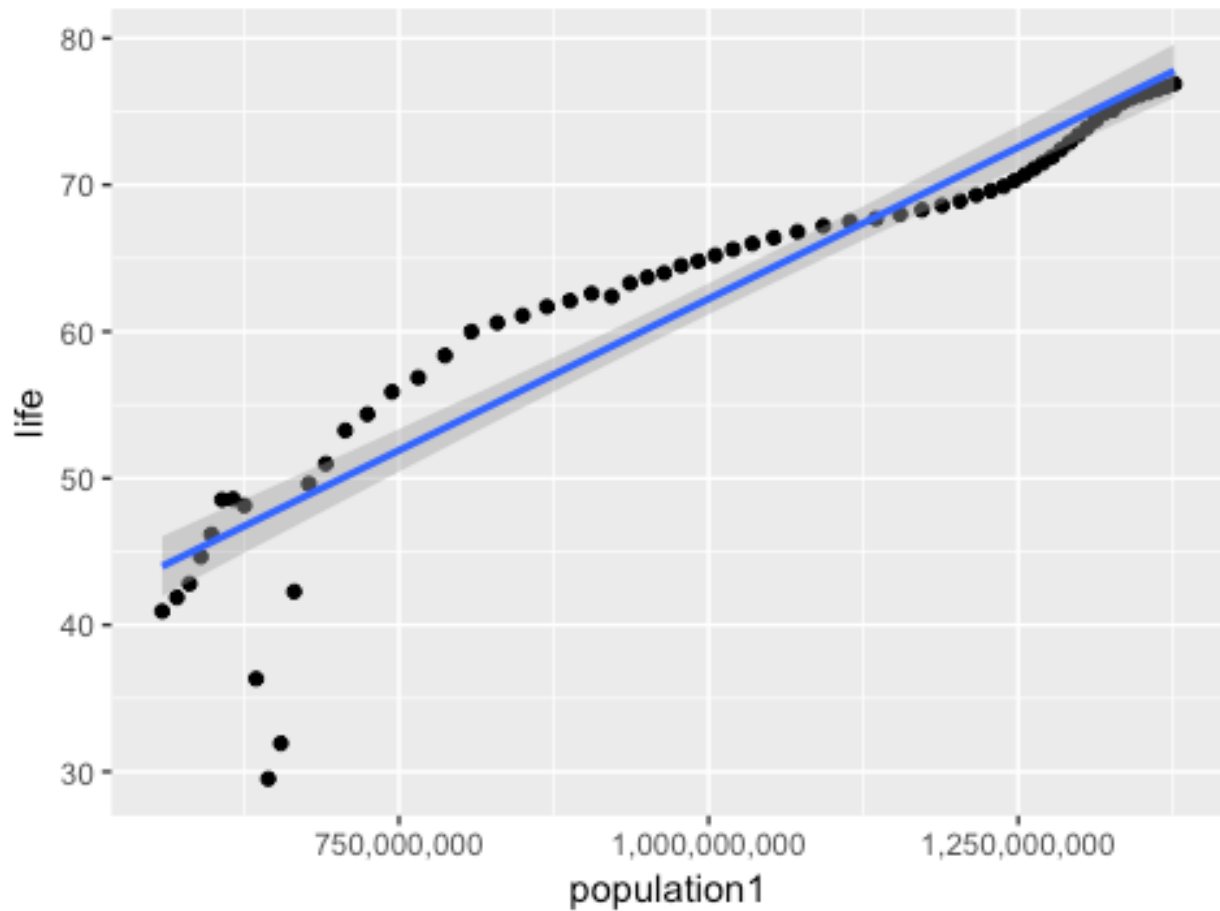
##   LifeExpected (Mean) in China Population (Standard Deviation) in China
## 1                886011895                340405516

#Filtering data for India after 1950

pop_chn_after_1950 <- population_china %>% filter(Year > '1950')

b <- ggplot(pop_chn_after_1950, aes(x = population1, y = life))
b + geom_point()+geom_smooth(method = "lm") + scale_x_continuous(labels
= scales::comma) + labs(title="China Population vs Life after 1950")
```

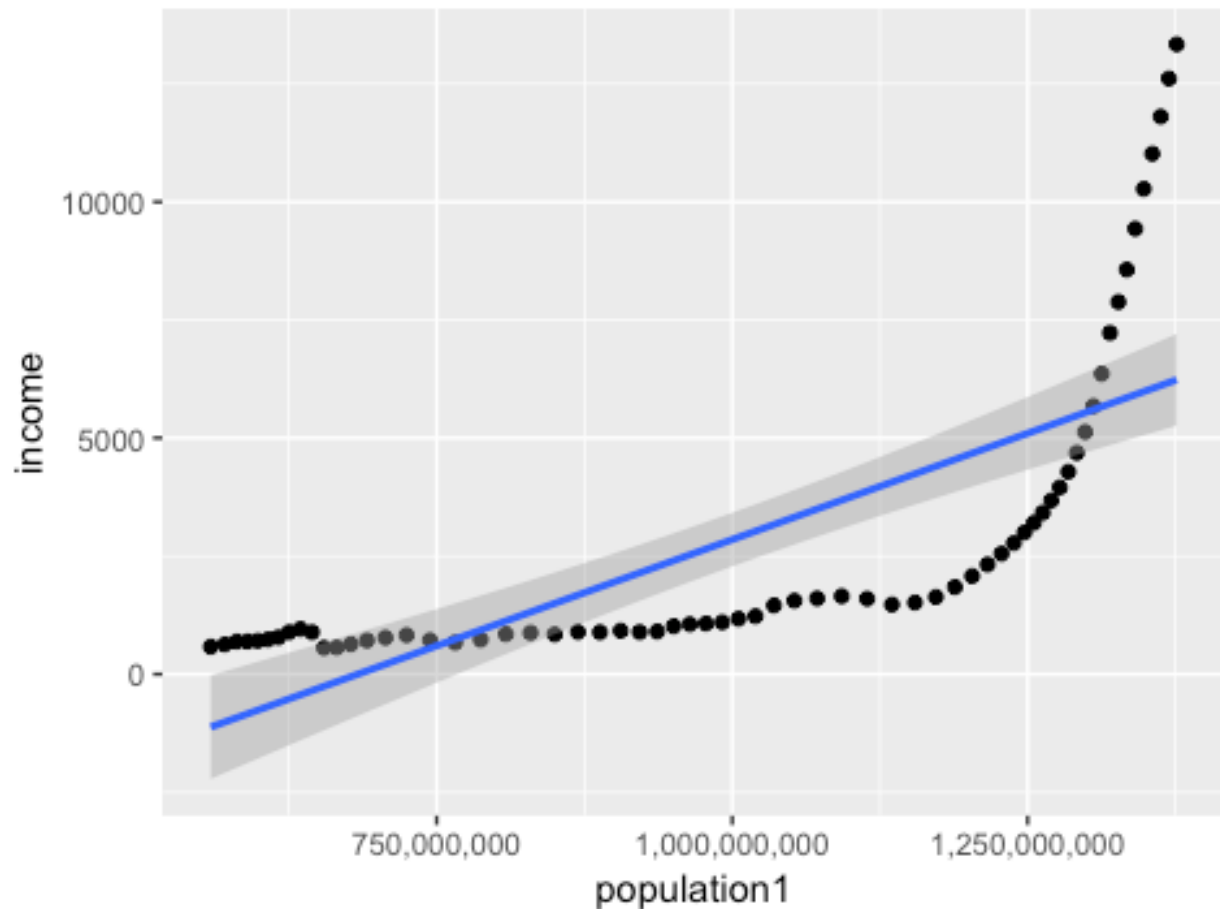
China Population vs Life after 1950



The population of China has a high number of age group between 50 to 60.

```
c <- ggplot(pop_chn_after_1950, aes(x = population1, y = income))  
  
c + geom_point()+geom_smooth(method = "lm") + scale_x_continuous(labels  
= scales::comma) + labs(title="China Population vs Income after 1950")
```

China Population vs Income after 1950



We can see here the income rose for China and India after population crossed 1.1 billion. However, the rise in life expectancy reduced when the population crossed 600 million and flattened between 55 to 65 years.

Working with 2015 data

#Creating a dataset with 2015 data

```
gapminder_2015 <- gapminder %>%
  filter(Year==2015)
```

Overview of 2015 data

```
summary(gapminder_2015)
```

```
##           Country      Year      life
## Afghanistan      : 1  Min.   :2015  Min.   :48.50
## Albania          : 1  1st Qu.:2015  1st Qu.:65.35
## Algeria          : 1  Median :2015  Median :73.50
## Andorra          : 1  Mean    :2015  Mean    :71.76
## Angola           : 1  3rd Qu.:2015  3rd Qu.:77.97
## Antigua and Barbuda: 1  Max.   :2015  Max.   :84.10
## (Other)          :172
##      population  income      region
##           : 1  Min.   : 624  America      :31
## 1,376,048,943: 1  1st Qu.: 3715  East Asia & Pacific :25
## 10,349,803   : 1  Median :11360  Europe & Central Asia :48
## 10,954,617   : 1  Mean    :17717  Middle East & North Africa:19
```

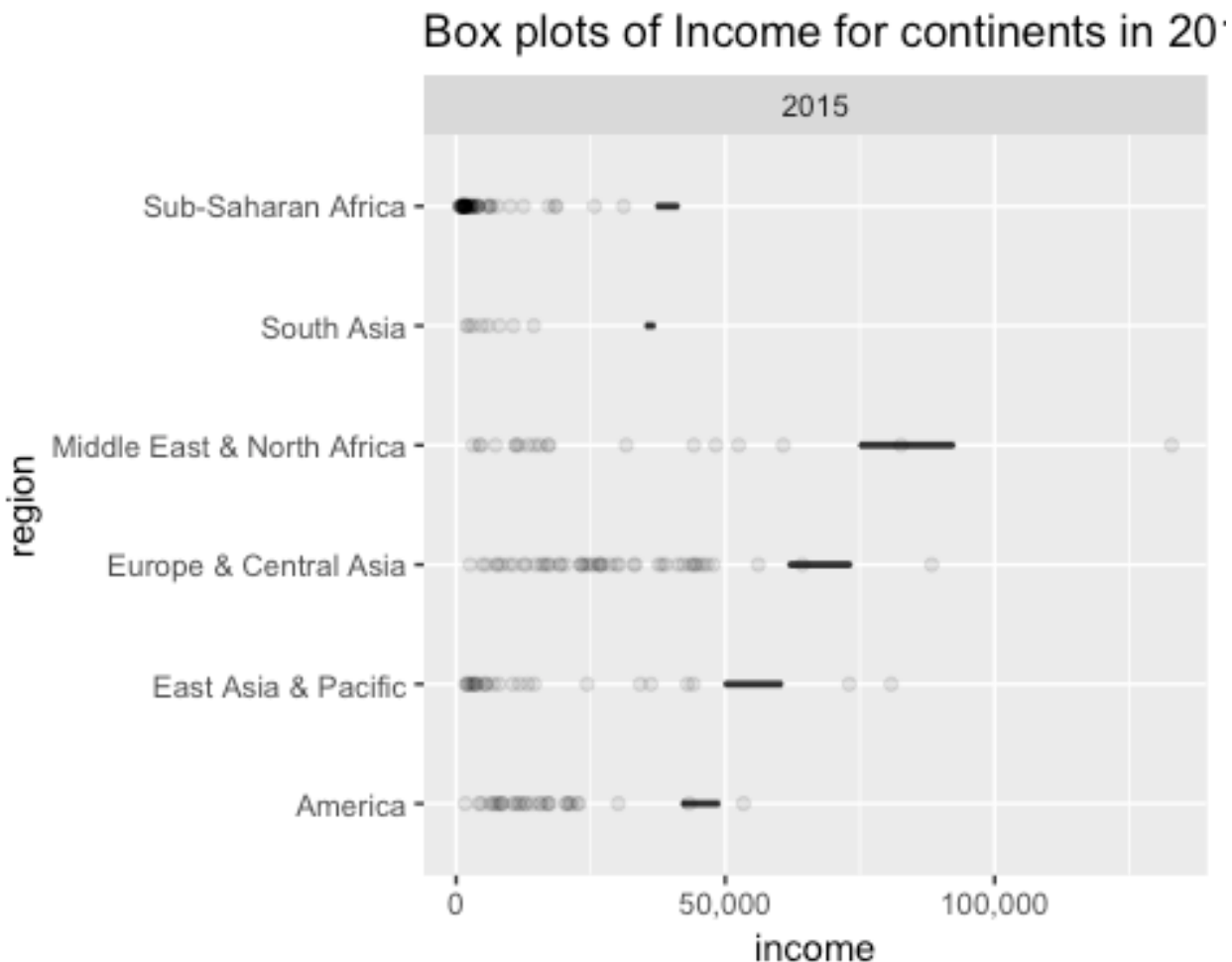
```
## 100699395 : 1 3rd Qu.: 24290 South Asia : 8
## 104460 : 1 Max. :132877 Sub-Saharan Africa :47
## (Other) :172
## population1
## Min. :5.299e+04
## 1st Qu.:2.235e+06
## Median :8.545e+06
## Mean :4.050e+07
## 3rd Qu.:2.851e+07
## Max. :1.376e+09
## NA's :1
```

Let us take a case for a single year (2015) and see at the income range and life expectancy comparison with respect to regions.

Income in 2015

#Plotting 2015 data with income vs region

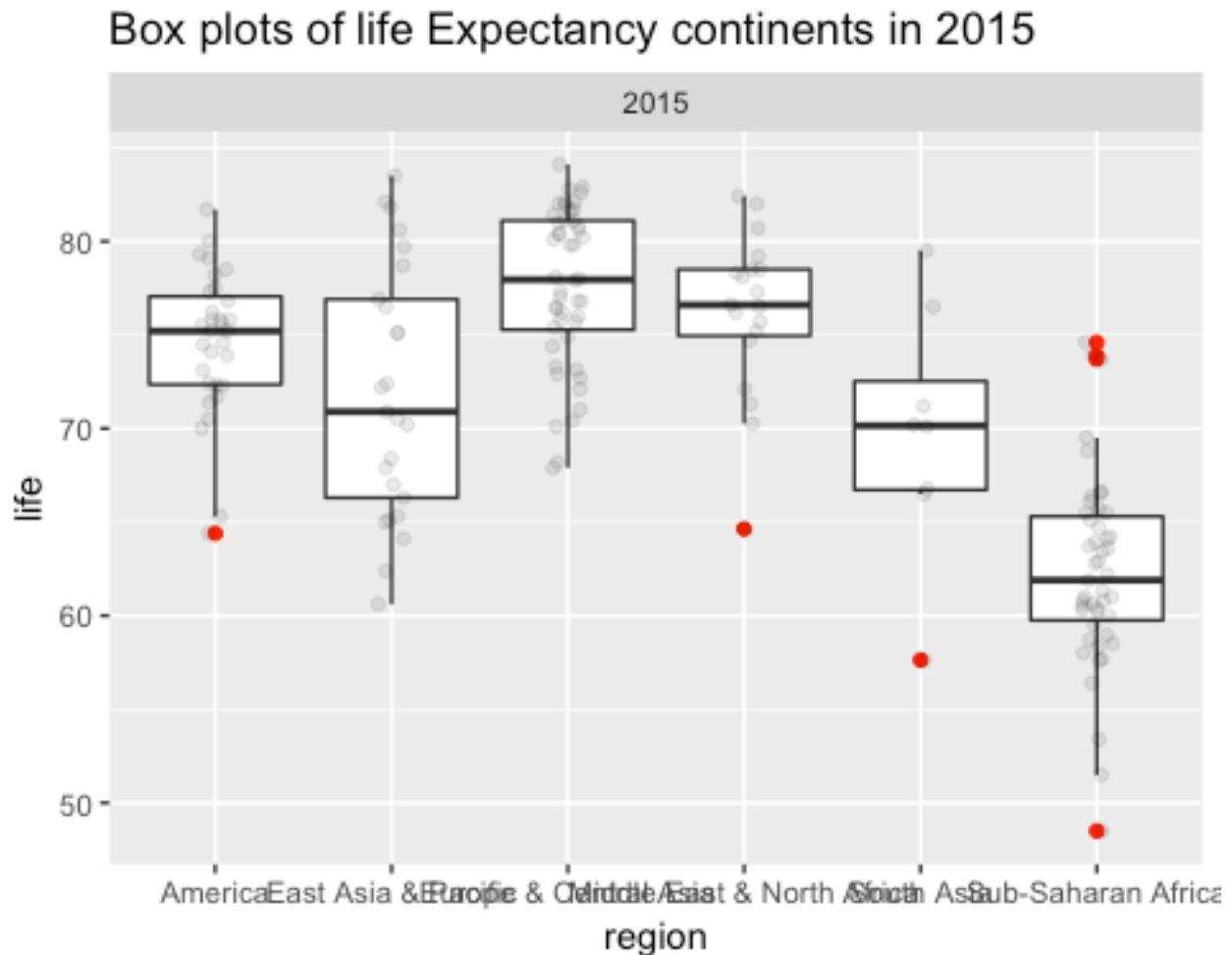
```
ggplot(gapminder_2015, aes(x = income, y = region)) + facet_wrap(~Year) +
geom_boxplot(outlier.colour = 'red') +
ggtitle('Box plots of Income for continents in 2015 across all countries') +
geom_jitter(position = position_jitter(width = 0.09, height = 0), alpha
= 1/10) +scale_x_continuous(labels = scales::comma)
```



Here we see that incomes for South Asia, Sub-Saharan Africa, and America is lower than 50000, while other regions go above this range and thus affecting the weighted mean and median.

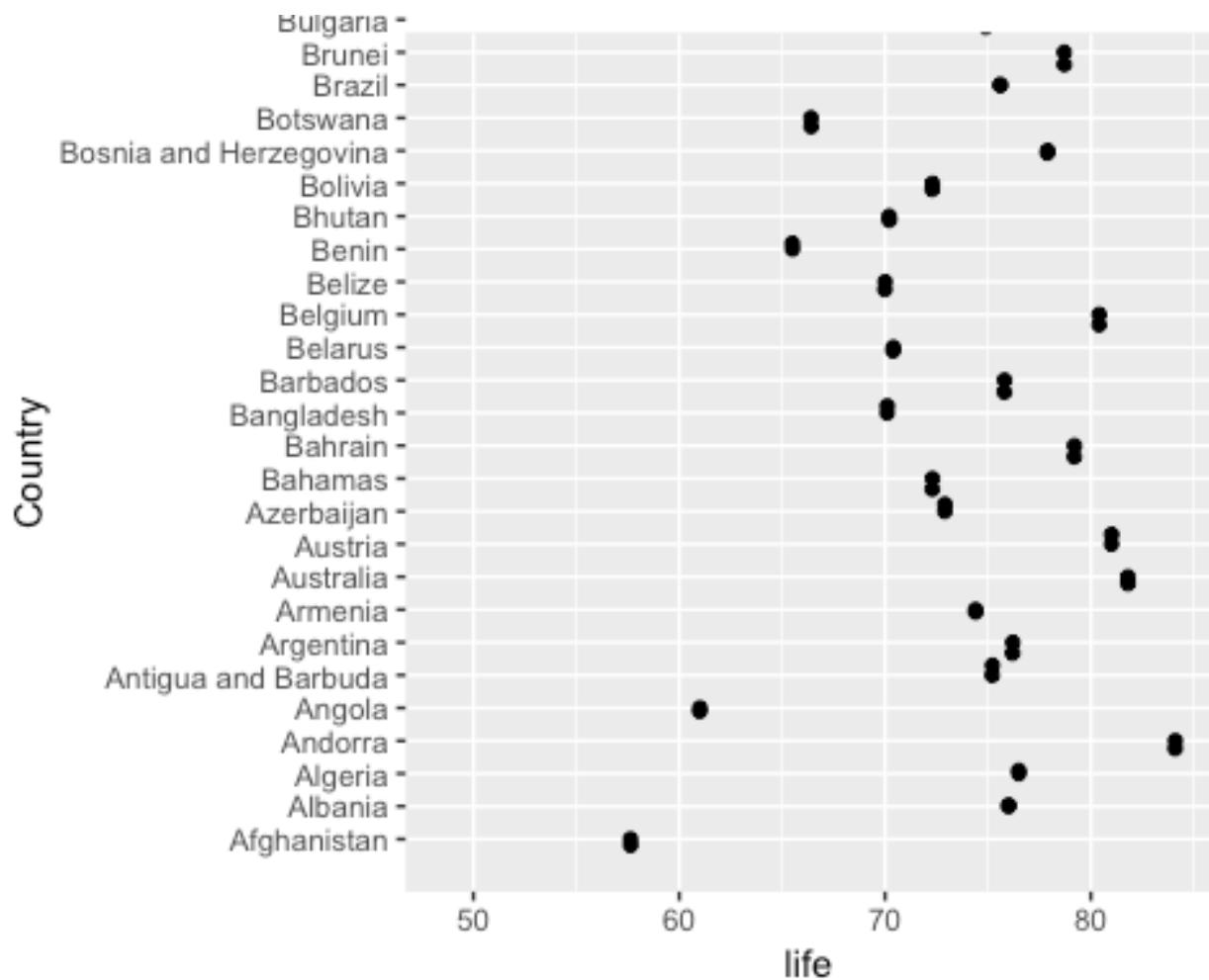
Life Expectancy in 2015

```
# Regions with Life expectancy mean
ggplot(gapminder_2015, aes(x = region, y = life)) + facet_wrap(~Year) +
geom_boxplot(outlier.colour = 'red') + # extreme values marked red
ggtitle('Box plots of life Expectancy continents in 2015') +
geom_jitter(position = position_jitter(width = 0.09, height = 0), alpha = 1/10)
```



The box plot above describes the Life expectancy of continents in 2015, a large set between 65 to 75 in China, the smallest North Africa but at an older age group between 75 to 80.

```
#List of countries with Life expectancy in 2015 arranged in descending order
life_gap_2015 <- gapminder_2015 %>% group_by(Country) %>% summarize(life) %>% arrange(desc(life))
ggplot(data= life_gap_2015, mapping
= aes(y=Country, x=life))+ geom_point()+ geom_jitter()+ coord_cartesian(ylim=c(0, 25))
```



The plot above shows the age groups (LIFE) for countries.

```
gapminder_2015 %>% # For each continent in 2015
```

```
group_by(region) %>%
```

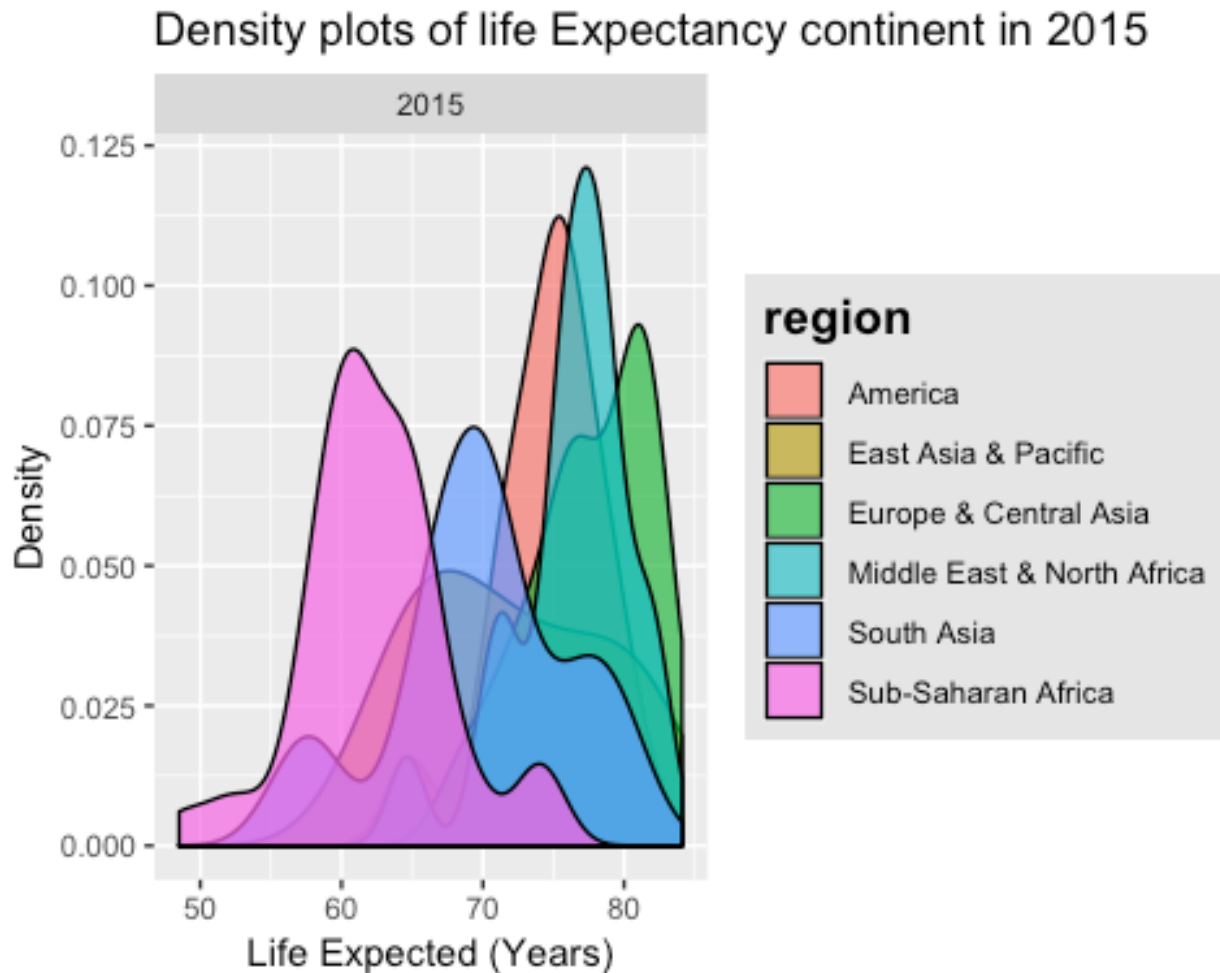
```
summarize('Standard Deviation'=sd(life),
'Inter Quartile Range'=IQR(life),
'Number'=n())
```

```
## # A tibble: 6 x 4
```

##	region	`Standard Deviation`	`Inter Quartile Range`	Number
##	<fct>	<dbl>	<dbl>	<int>
## 1	America	3.85	4.7	31
## 2	East Asia & Pacific	6.75	10.6	25
## 3	Europe & Central Asia	4.20	5.82	48
## 4	Middle East & North Afr...	4.31	3.55	19
## 5	South Asia	6.65	5.8	8
## 6	Sub-Saharan Africa	5.14	5.55	47

```
gapminder_2015 %>%
```

```
ggplot(aes(x = life, fill = region)) + facet_wrap(~Year) + # aes = aesthetics
geom_density(alpha = 0.7) +
ggtitle('Density plots of life Expectancy continent in 2015') +
theme(legend.title = element_text(color = 'Black', size = 14, face = 'bold'),
legend.background = element_rect(fill = 'gray90', size = 0.5, linetype = 'dashed')) +
labs(x='Life Expected (Years)', y='Density')
```



Life Expectancy for America in 2015

Let us focus on a single region and in it a single country. Here, the example considered is the region of **America** and the country sample is **United States**.

Calculating the standard deviation for America in 2015:

```
gapminder_2015 %>% #compute stats
filter(region=='America') %>%
summarize('LifeExpectancy (Mean) in America for 2015'=mean(life),
'LifeExpectancy (Standard Deviation) in America for 2015'=sd(life))
## LifeExpectancy (Mean) in America for 2015
## 1 74.64516
## LifeExpectancy (Standard Deviation) in America for 2015
## 1 3.853729
```


America and United States in 2015

```
# Creating subset of America region
life_America <- gapminder %>% filter(region=='America')

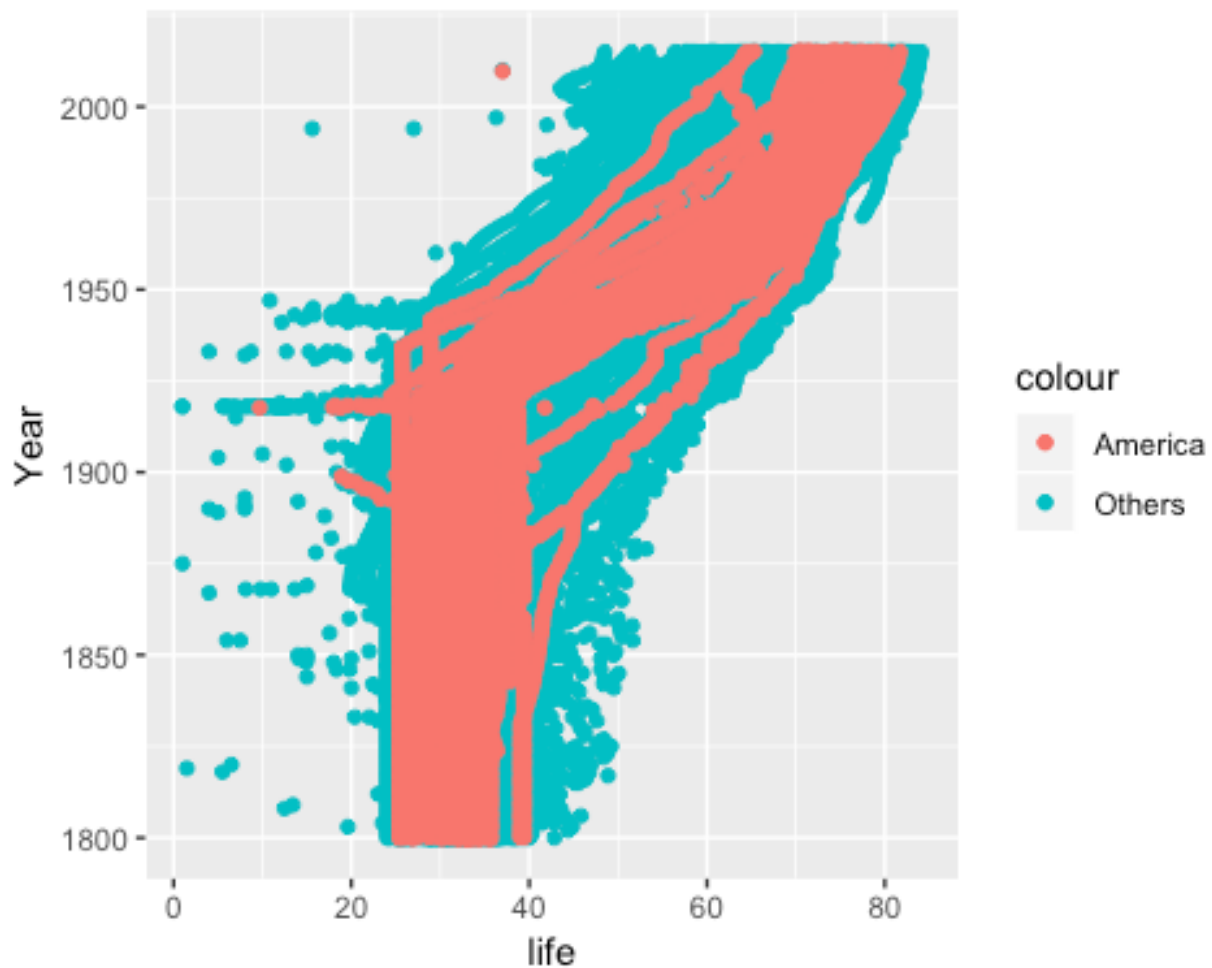
# Creating subset of United States country
life_USA <- gapminder %>% filter(Country=='United States')
```

Plotting Life Expectancy in 2015:

America region vs other region in world

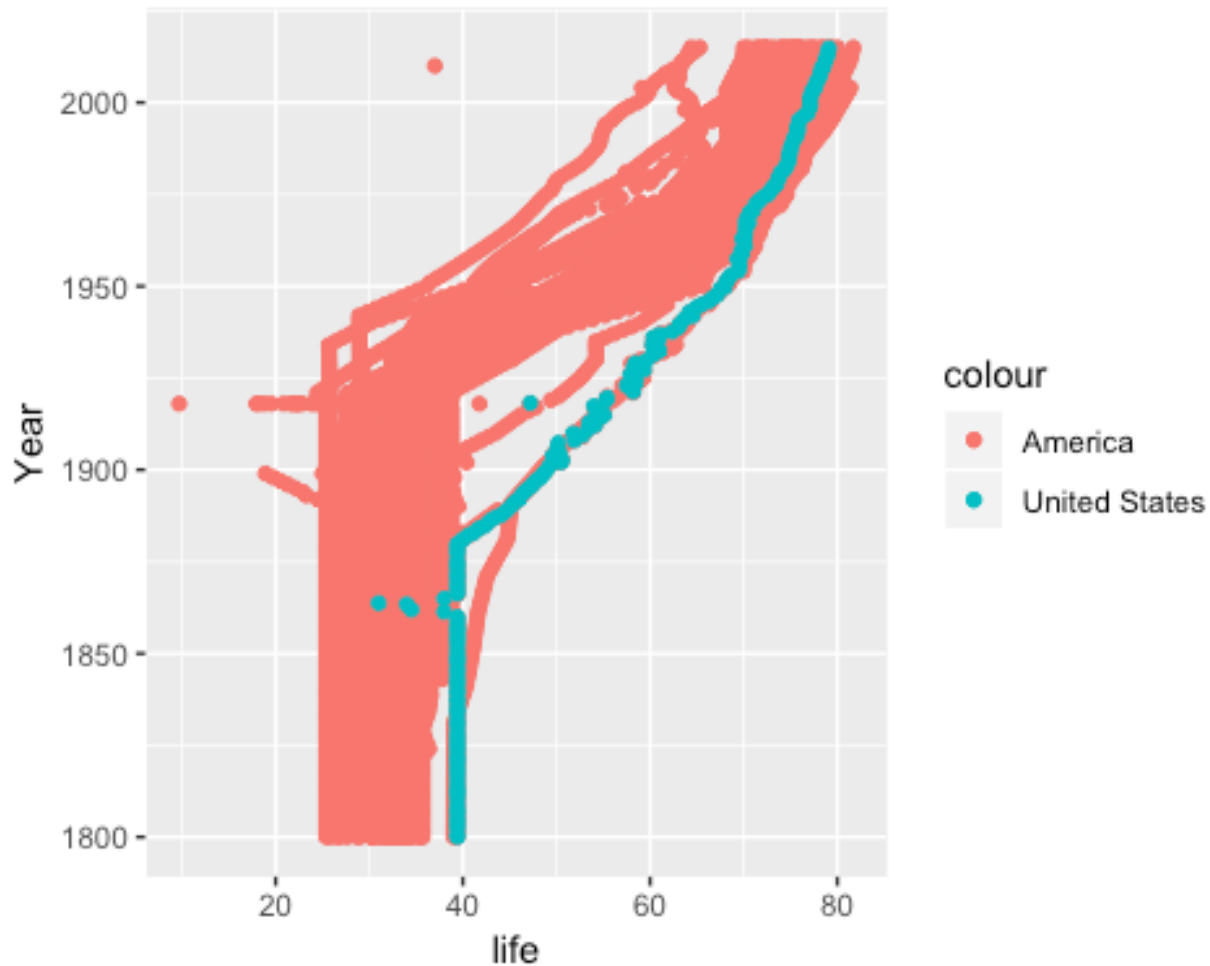
United States country vs other countries in America

```
ggplot(data= life_America, mapping = aes(y=Year, x=life, color="America")) + geom_point(data
= gapminder, aes(color="Others")) + geom_jitter()
```



The graph shows that the Life Expectancy is below average in the early years for America, but since 1950s, America's life expectancy has improved and to be above average than most of the other regions.

```
ggplot(data= life_USA, mapping = aes(y=Year, x=life, color="United States")) + geom_point(data
= life_America, aes(color="America")) + geom_jitter()
```



While comparing United States to America, the Life expectancy of United States is higher than most of other countries in the same region since the beginning.

Summary

Income for some of the regions are higher than compared to other regions. In 2015, average Income Middle East & North Africa has been the highest compared to South Asia where it recorded lowest average income. **Income rises when the population remains lower.**

The average income is 12397 across all region, but if the regions are considered, then the income reduces to 3900, meaning that income in some of the region is much lower than that of other regions.

Life Expectancy We find that life expectancy in the given dataset is higher in the range of 25-35 years. The mean of Life Expectancy in 2015 for Sub-Saharan Africa region is lower than all other regions, while Europe & Central Asia is highest. **There is a relationship between Life Expectancy and income.**

As the life expectancy grows above 60, the income levels rise higher.

When magnified the analysis to American region, the life Expectancy is nearly 75%. Though the average life expectancy of America improves from below average to above average from 1800 to 2015, the life expectancy of United States has been highest than most of the American countries from 1800 to 2015.

Population We see that India and China are higher in population than other countries. **There is a exponential growth in these countries after 1950.** We also saw that population causes a flattening effect on life expectancy (over 600 million) and also extreme population (over 1 billion) witnesses an increased growth in income.