Exploratory Data Analysis Project- Gapminder Dataset

Anjali Rawat

February 18, 2019

EDA Assignment Summary

In this exploratory data analysis assignment I have used the Gapminder Dataset for analysis. After starting with a brief data inspection and exploratory plotting, I set up three specific questions, shown in section 1 below, for my analysis in this project. Thre reset of the presentation and anlysis of the project is as follows.

I first present a systematic data description which includes description of each variables, total number of records and missing values etc. In Data exploration section I have first presented regionwise (subcontinents) discriptive statistics and visulization (boxplots, histograms, and density plots for different) of the different variables present in the data. The analysis results/interpretations are shown below.

1. The analysis of the Life Expectancy vs Per Capita Income plot did not show any conclusive trend, however, plotting percapita income on log scale shows a nearly linear trend with life expectancy with quite a bit of spread. This is ecpected that increase in percapita income will increase life expectancy.
2. The life expectancy trends over the years show a very low and flat expectancy until early parts of the 20th century in all the continents (may be because of lack of health care) and then shown increasing treands in all continents. The increasing trends are set early in high percapita regions, like, US and Europe, and late in poorer regions like middle east/north africa and Asia pacific. There trends are obvious and are in line with the expectations.
3. Similar analysis for United States shows a clearer tresnd for life expectancy whrere above certain level (~$8000) it increases sharply and the slows down beyond certain limit as income cann’t control life beyond certain limit.Similarly, life expectancy in US started increasing very early (before 1900) and is showing a contineous increase over time probably due to improvement in health care, which is also strongly related to income as well.

1. Questions

Q1) What are the descriptive statistics of Life, Income and Population for six Regions

Q2) What is the trend in Life Expectancy (i.e. life) with Per Capita Income (i.e. Income) for the entire dataset.

Q3) What is the trend in Life Expectancy (i.e. life),Population and Income over the years for different regions

Q4) Repeat Q2 and Q3 analysis for United Staes.

2. Data description

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
gapminder <- read_csv("C:/Harrisburg University of Science and Technology/Anly 506-90- O-2018Lat
e Fall - Exploratory Data Analytics/EDA Assignment +Code Portfolio/EDAAssignment_CodePortfolio_A
njaliRawat/gapminder.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Year = col_double(),
##   life = col_double(),
##   population = col_number(),
##   income = col_double(),
##   region = col_character()
## )
```

The gapminder dataset is a comma-separated value (.CSV) file with 41284 records.The data frame has six features/variables

i. Country (a categorical variable), a factor with 197 levels

ii. region (a categorical variable), a factor with 6 levels represent different subcontinental regions in the world

iii. `Year (a categorical variable), a factor with 216 levels`

iv. life (a continuous variable)- denote Life Expectancy

v. population(a discrete variable)

vi. income (a continuous variable)- represents Per Capita Income

```
nrow(gapminder)
```

```
## [1] 41284
```

```
ncol(gapminder)
```

```
## [1] 6
```

```
str(gapminder)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 41284 obs. of  6 variables:
##  $ Country   : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year      : num  1800 1801 1802 1803 1804 ...
##  $ 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    : num  603 603 603 603 603 603 603 603 603 603 ...
##  $ region    : chr  "South Asia" "South Asia" "South Asia" "South Asia" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Country = col_character(),
##   ..   Year = col_double(),
##   ..   life = col_double(),
##   ..   population = col_number(),
##   ..   income = col_double(),
##   ..   region = col_character()
##   .. )
```

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##   Country      Year  life population income region
##   <chr>       <dbl> <dbl>      <dbl>  <dbl> <chr>
## 1 Afghanistan  1800  28.2    3280000    603 South Asia
## 2 Afghanistan  1801  28.2         NA    603 South Asia
## 3 Afghanistan  1802  28.2         NA    603 South Asia
## 4 Afghanistan  1803  28.2         NA    603 South Asia
## 5 Afghanistan  1804  28.2         NA    603 South Asia
## 6 Afghanistan  1805  28.2         NA    603 South Asia
```

```
tail(gapminder)
```

```
## # A tibble: 6 x 6
##   Country  Year  life population income region
##   <chr>   <dbl> <dbl>      <dbl>  <dbl> <chr>
## 1 Åland    1992  80.8      24834     NA Europe & Central Asia
## 2 Åland    1993  81.8      24950     NA Europe & Central Asia
## 3 Åland    1994  80.6      25066     NA Europe & Central Asia
## 4 Åland    1995  79.9      25183     NA Europe & Central Asia
## 5 Åland    1996  80        25301     NA Europe & Central Asia
## 6 Åland    1997  80.1      25419     NA Europe & Central Asia
```

3 Data exploration

3.1 Exploratory regional data analysis

3.1.1 Data summary statistics

a) Mean and Median of life expectancy Region wise

```
aggregate(life ~ region, gapminder, median)
```

```
##                       region     life
## 1                    America 35.37370
## 2         East Asia & Pacific 34.00000
## 3       Europe & Central Asia 41.74110
## 4 Middle East & North Africa 32.30000
## 5                 South Asia 32.64700
## 6          Sub-Saharan Africa 32.40338
```

```
aggregate(life ~ region, gapminder, mean)
```

```
##                       region     life
## 1                    America 44.54065
## 2         East Asia & Pacific 41.76041
## 3       Europe & Central Asia 48.79419
## 4 Middle East & North Africa 41.55366
## 5                 South Asia 37.42302
## 6          Sub-Saharan Africa 37.88242
```

b) Five Num Summary Life Expectancy Region wise

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
SouthAsia=filter(gapminder, region=="South Asia")
fivenum(SouthAsia$life)
```

```
## [1]  7.045063 27.345157 32.647000 44.374110 79.500000
```

```
fivenum(SouthAsia$population)
```

```
## [1]      42378     993915   15661841   69328216 1311050527
```

```
fivenum(SouthAsia$income)
```

```
## [1]   603.0   840.0  1020.0  1286.5 14408.0
```

```
EastAsiaPacific = filter(gapminder,region=="East Asia & Pacific")
fivenum(EastAsiaPacific$life)
```

```
## [1]  1.00000 28.94900 34.00000 57.49076 83.50000
```

```
fivenum(EastAsiaPacific$population)
```

```
## [1]       1548     115956    1241248   18196783 1376048943
```

```
fivenum(EastAsiaPacific$income)
```

```
## [1]    363.0    873.0   1153.5   2557.5 134864.0
```

```
EuropeCentralAsia=filter(gapminder,region=="Europe & Central Asia")
fivenum(EuropeCentralAsia$life)
```

```
## [1]  3.98908 35.60000 41.74110 66.60000 84.10000
```

```
fivenum(EuropeCentralAsia$population)
```

```
## [1]      9584   1960597   4750396  10592125 148435811
```

```
fivenum(EuropeCentralAsia$income)
```

```
## [1]   393  1427  2735  7387 96245
```

```
MiddleEastNorthAfrica = filter(gapminder,region=="Middle East & North Africa")
fivenum(MiddleEastNorthAfrica$life)
```

```
## [1]  1.50000 30.71320 32.30000 53.70766 82.40000
```

```
fivenum(MiddleEastNorthAfrica$population)
```

```
## [1]     2788   548618  2615753  9499387 91508084
```

```
fivenum(MiddleEastNorthAfrica$income)
```

```
## [1]    715.0   1082.0   1537.5   3939.0 182668.0
```

```
America =filter(gapminder,region=="America")
fivenum(America$life)
```

```
## [1]  9.690052 32.124000 35.373700 61.048460 81.700000
```

```
fivenum(America$population)
```

```
## [1]      9899.0    236909.5   2144973.0   8510081.5 321773631.0
```

```
fivenum(America$income)
```

```
## [1]   529.0  1277.5  2214.5  5461.5 53354.0
```

```
SubSaharanAfrica = filter(gapminder,region=="Sub-Saharan Africa")
fivenum(SubSaharanAfrica$life)
```

```
## [1]  4.00000 30.43680 32.40338 46.00000 79.64600
```

```
fivenum(SubSaharanAfrica$population)
```

```
## [1]      8219    821457   3019768   8719290 182201962
```

```
fivenum(SubSaharanAfrica$income)
```

```
## [1]   142   596   827  1323 40143
```

3.1.2. Raw Data Visualization

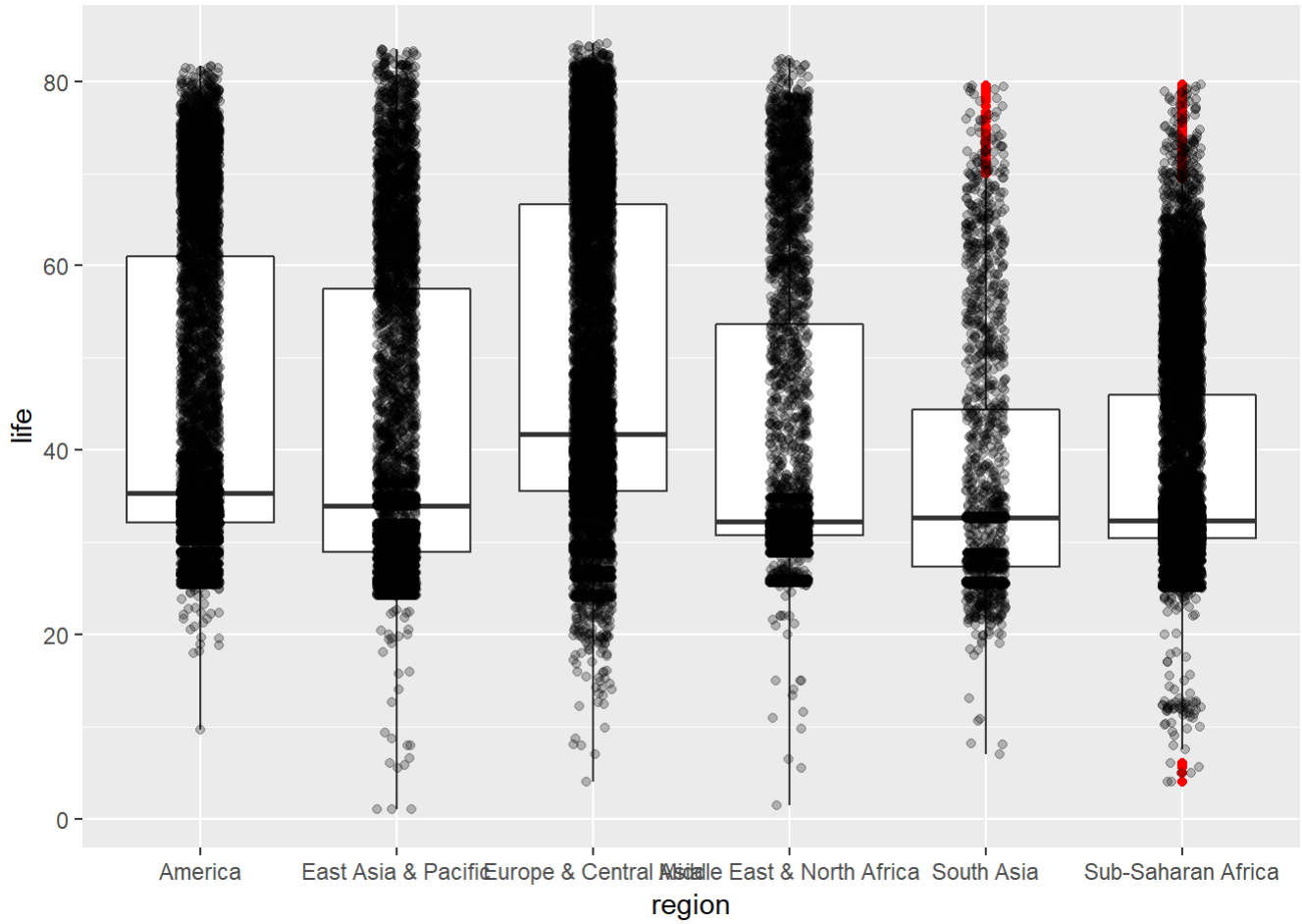
a) Boxplots

Life expectancy comparision in different region

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
ggplot(gapminder, aes(x = region, y = life)) +  geom_boxplot(outlier.colour = "red") +  geom_jit
ter(position = position_jitter(width = 0.1, height = 0), alpha = 1/4)
```

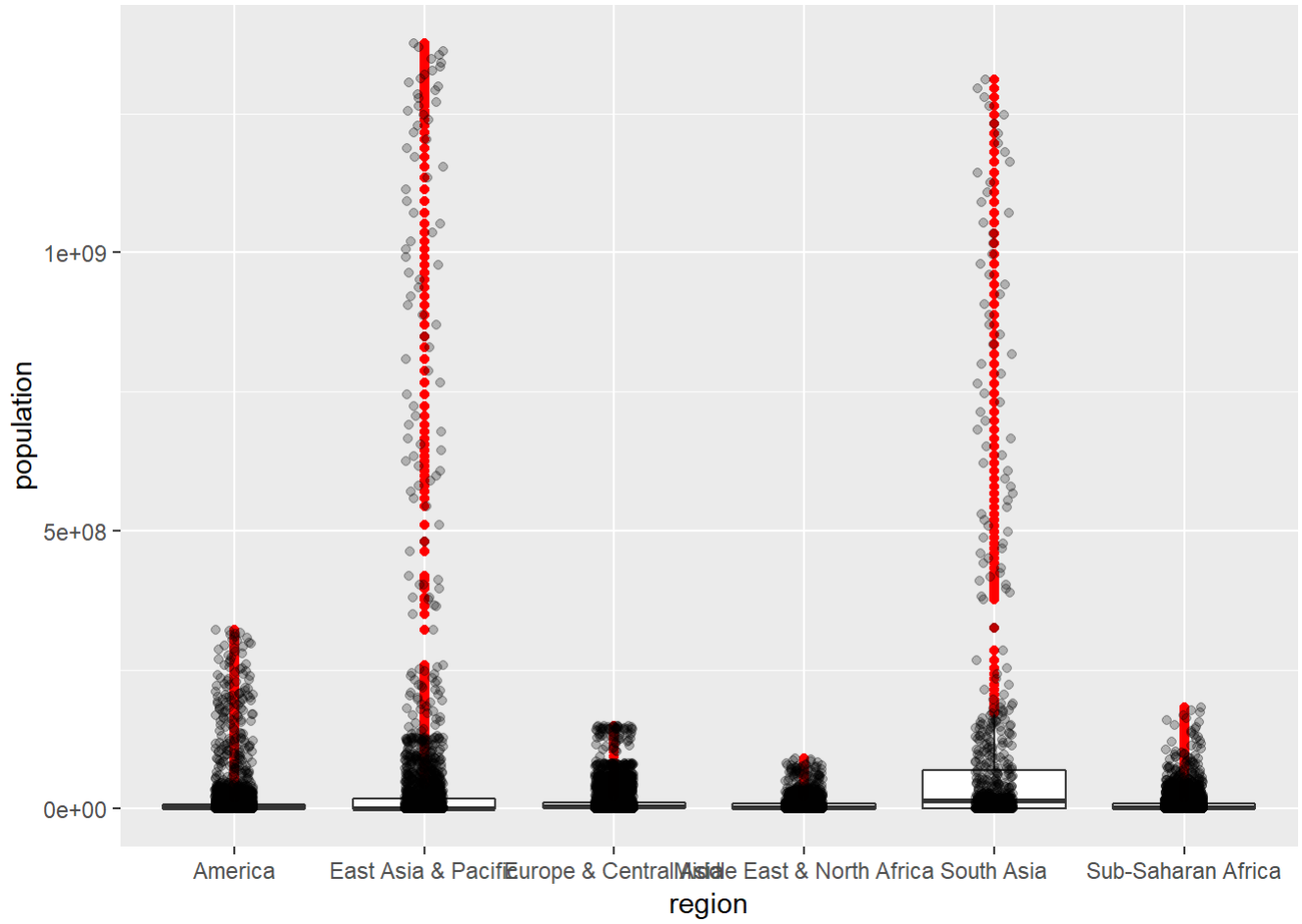


Population comparision in different region

```
ggplot(gapminder, aes(x = region, y = population)) +  geom_boxplot(outlier.colour = "red") +  ge
om_jitter(position = position_jitter(width = 0.1, height = 0), alpha = 1/4)
```

```
## Warning: Removed 25817 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 25817 rows containing missing values (geom_point).
```

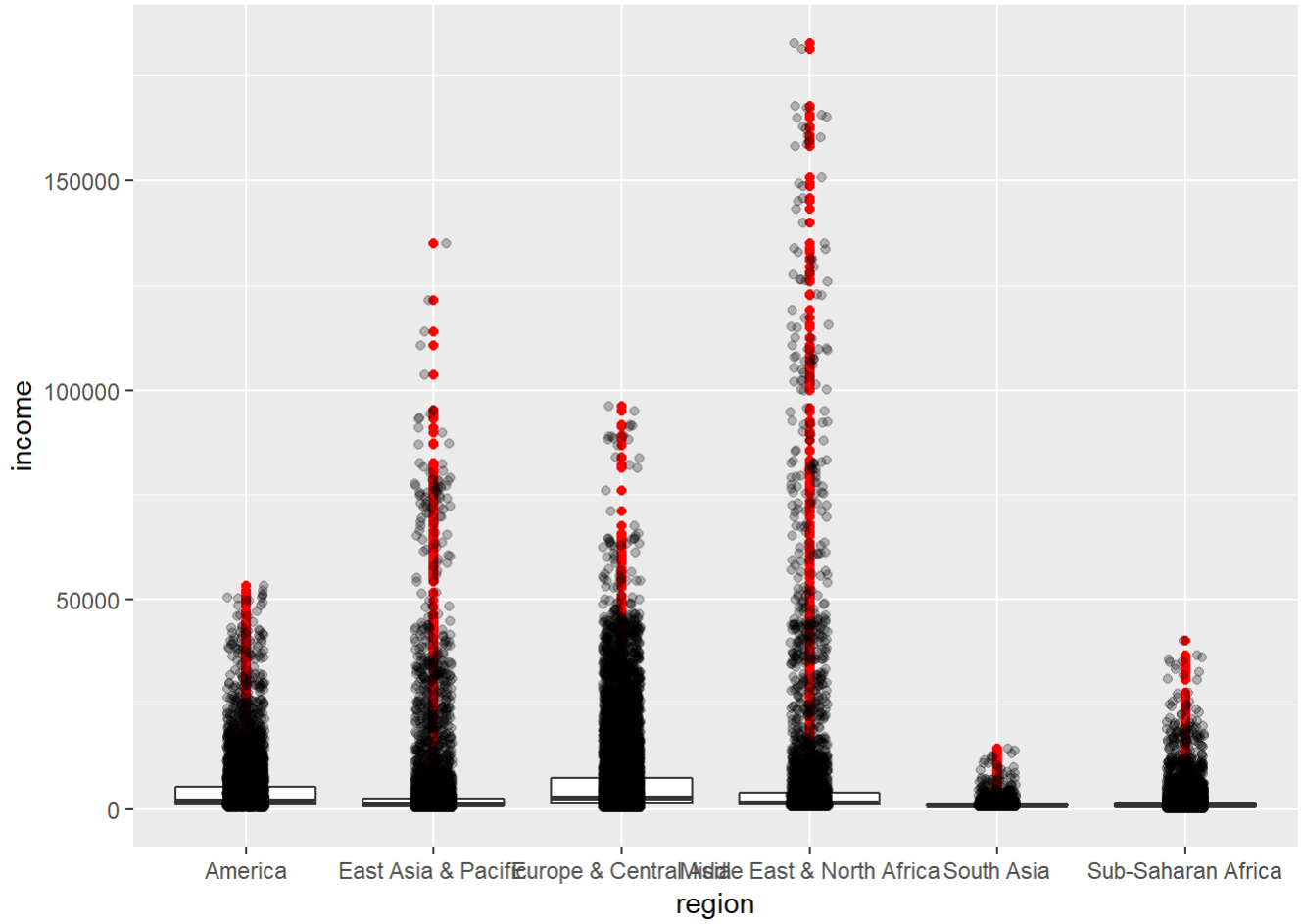


Income comparision in different region

```
ggplot(gapminder, aes(x = region, y = income)) +  geom_boxplot(outlier.colour = "red") +  geom_j
itter(position = position_jitter(width = 0.1, height = 0), alpha = 1/4)
```

```
## Warning: Removed 2341 rows containing non-finite values (stat_boxplot).
```

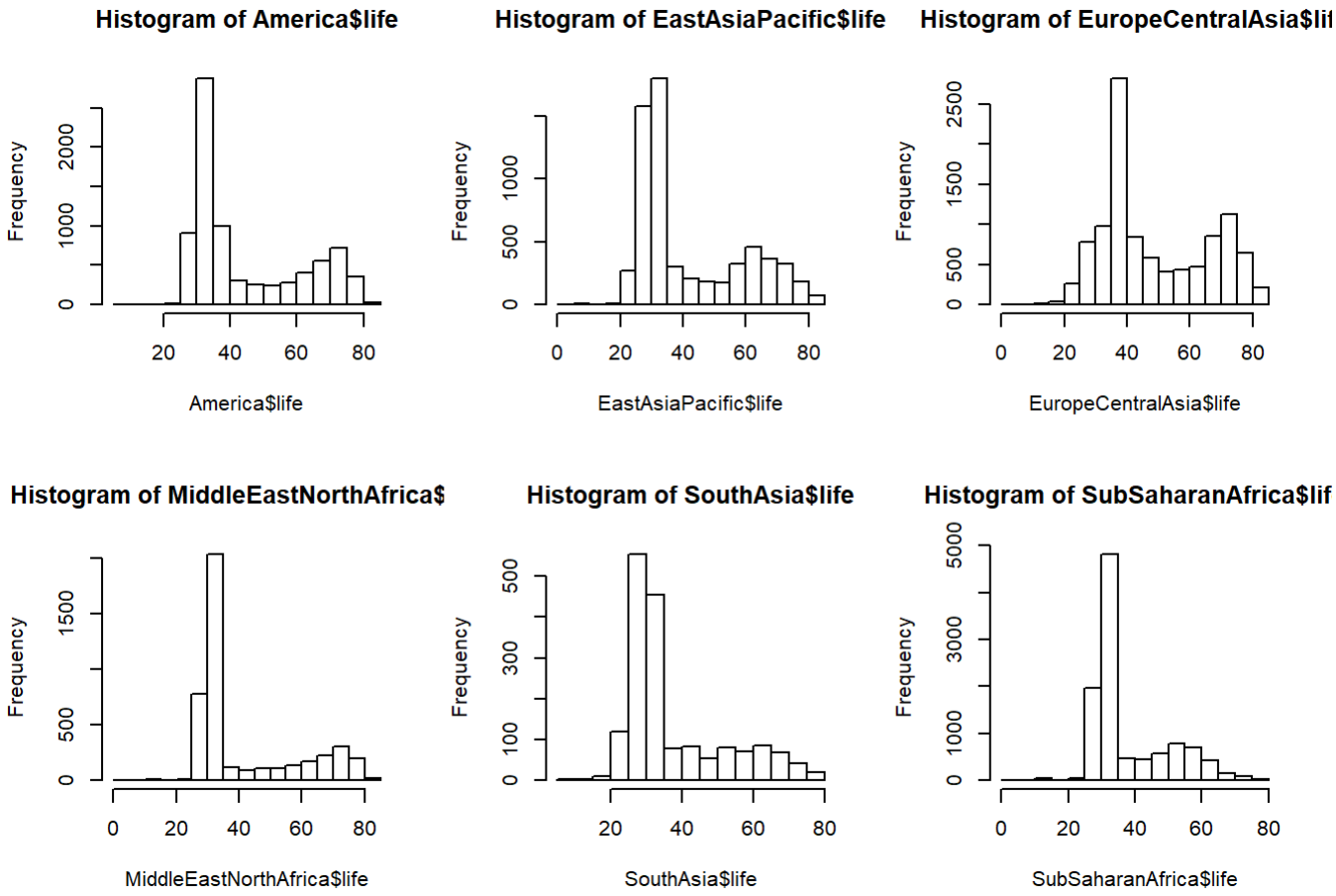
```
## Warning: Removed 2341 rows containing missing values (geom_point).
```



b) Histograms

Life Expectancy by region

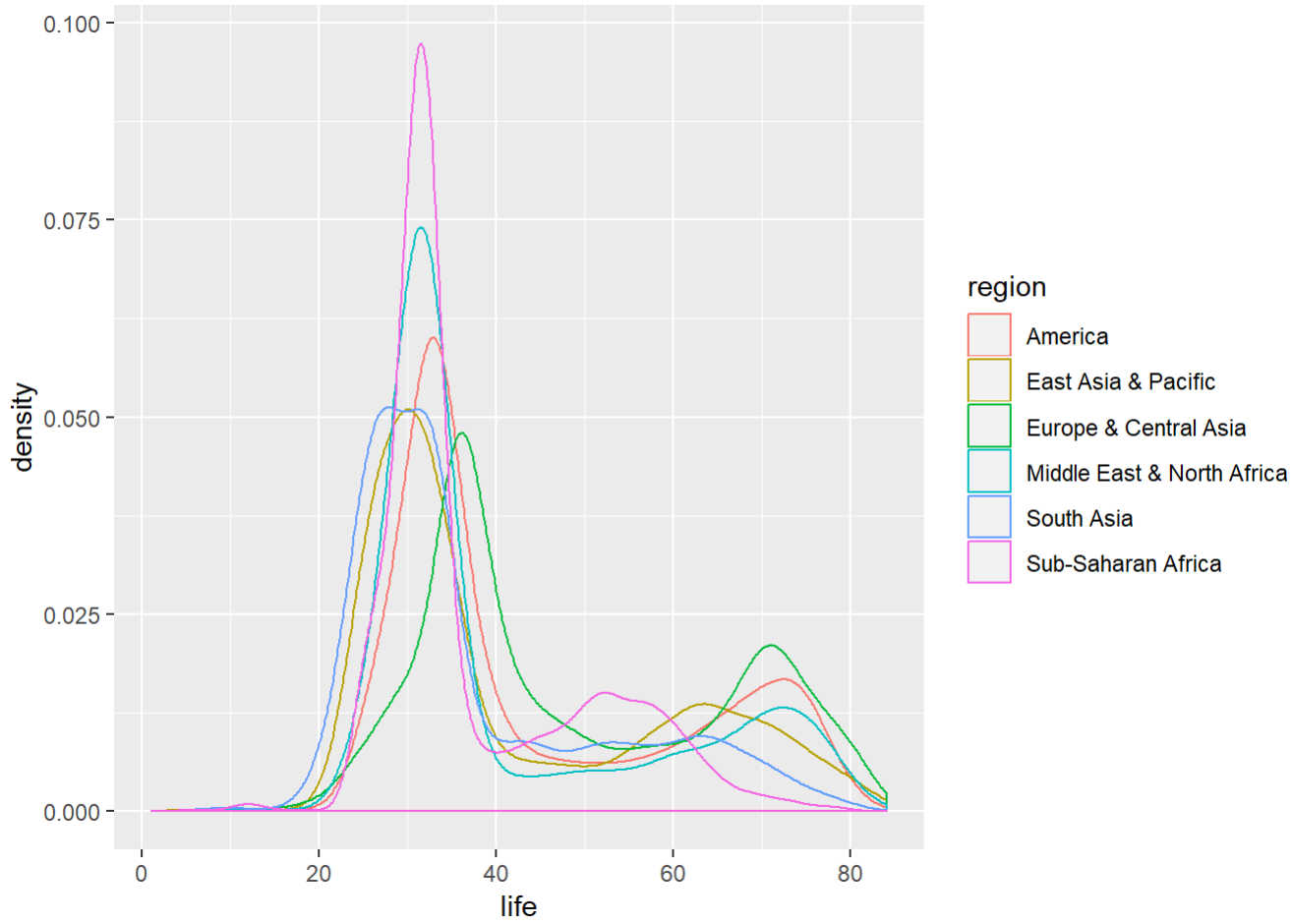
```
library(ggplot2)
par(mfrow=c(2,3))
hist(America$life)
hist(EastAsiaPacific$life)
hist(EuropeCentralAsia$life)
hist(MiddleEastNorthAfrica$life)
hist(SouthAsia$life)
hist(SubSaharanAfrica$life)
```



c) Density plots

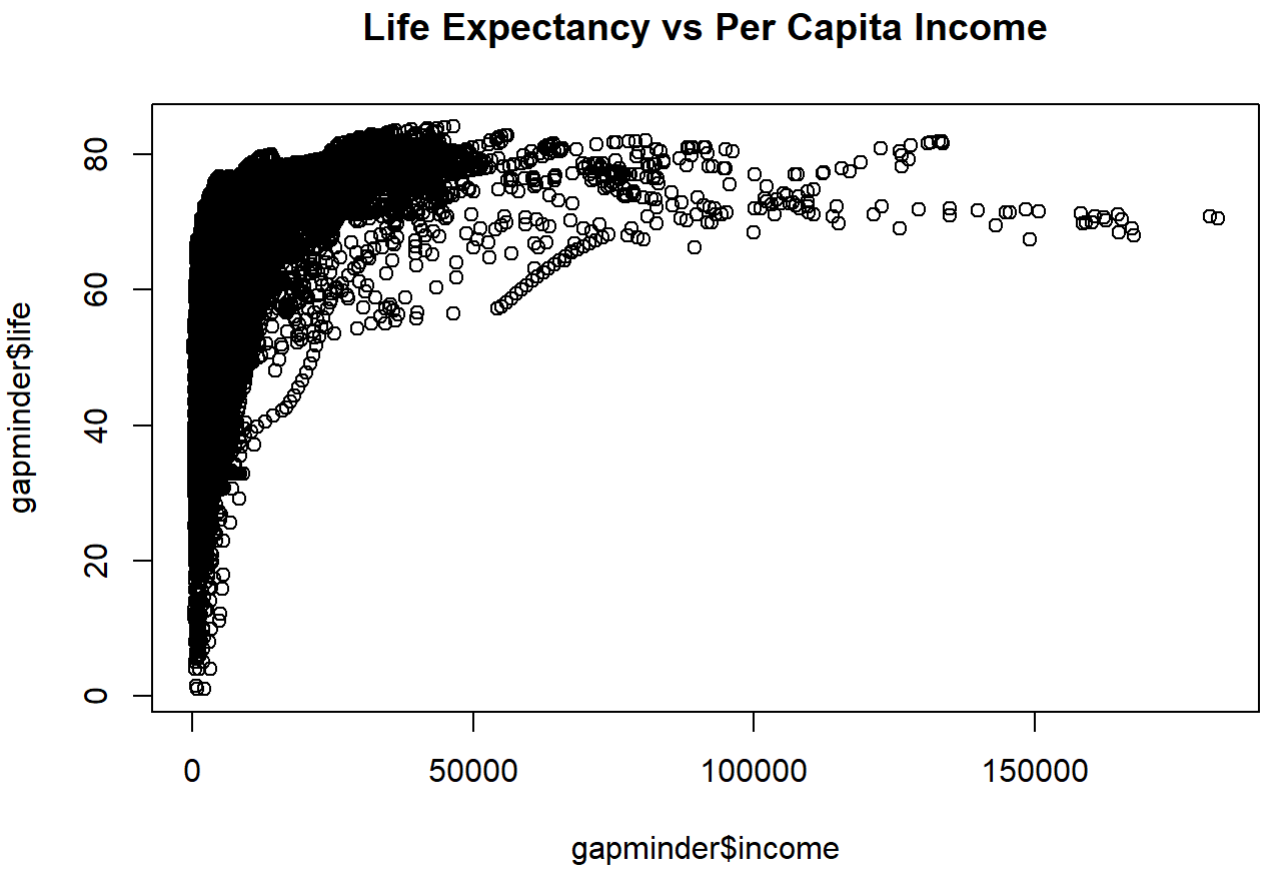
Density Plot of life expectancy by region

```
ggplot(gapminder,aes(life))+geom_density(aes(color=region))
```

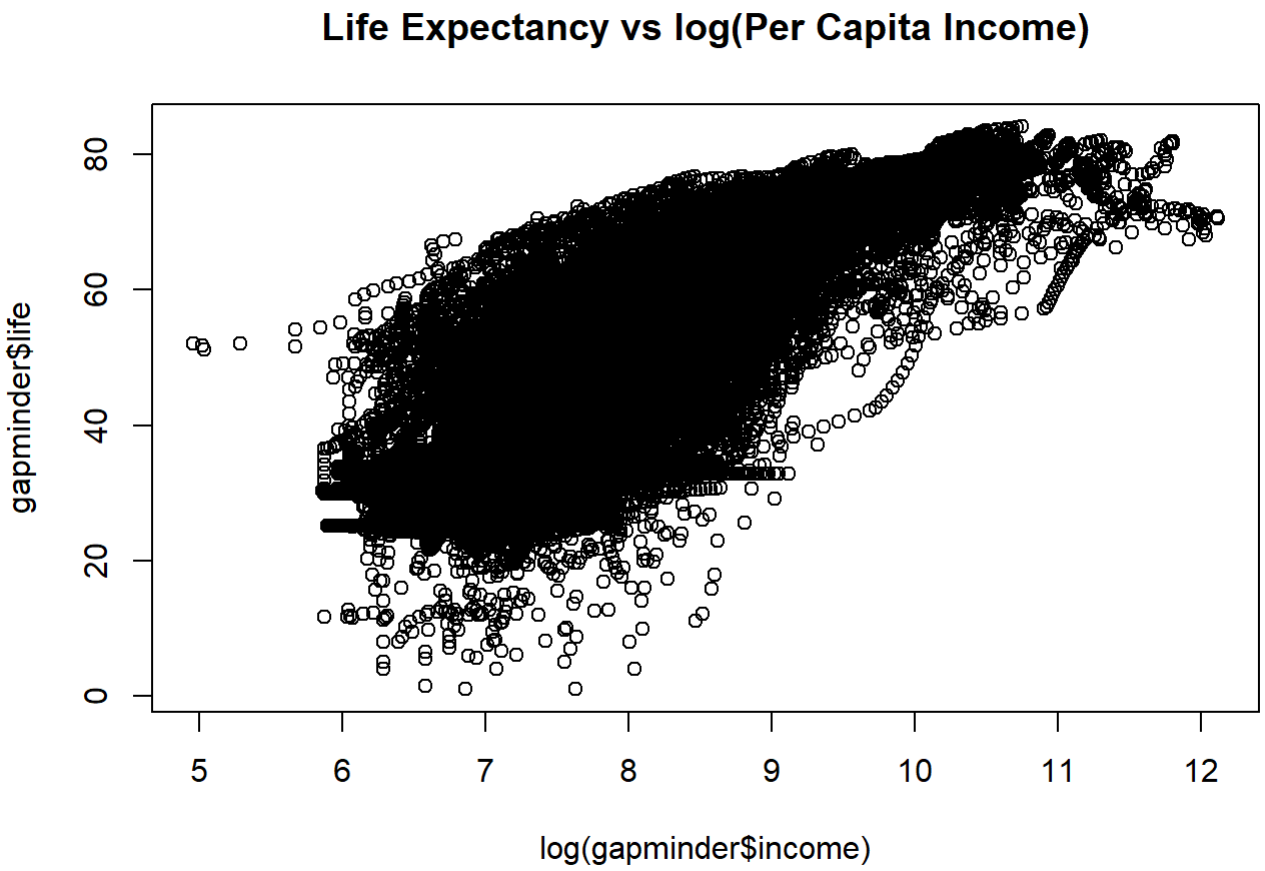


3.1.3. Analysis of Life Expectancy vs Per capita income for different subcontinents

```
plot(gapminder$income, gapminder$life, main = "Life Expectancy vs Per Capita Income" )
```

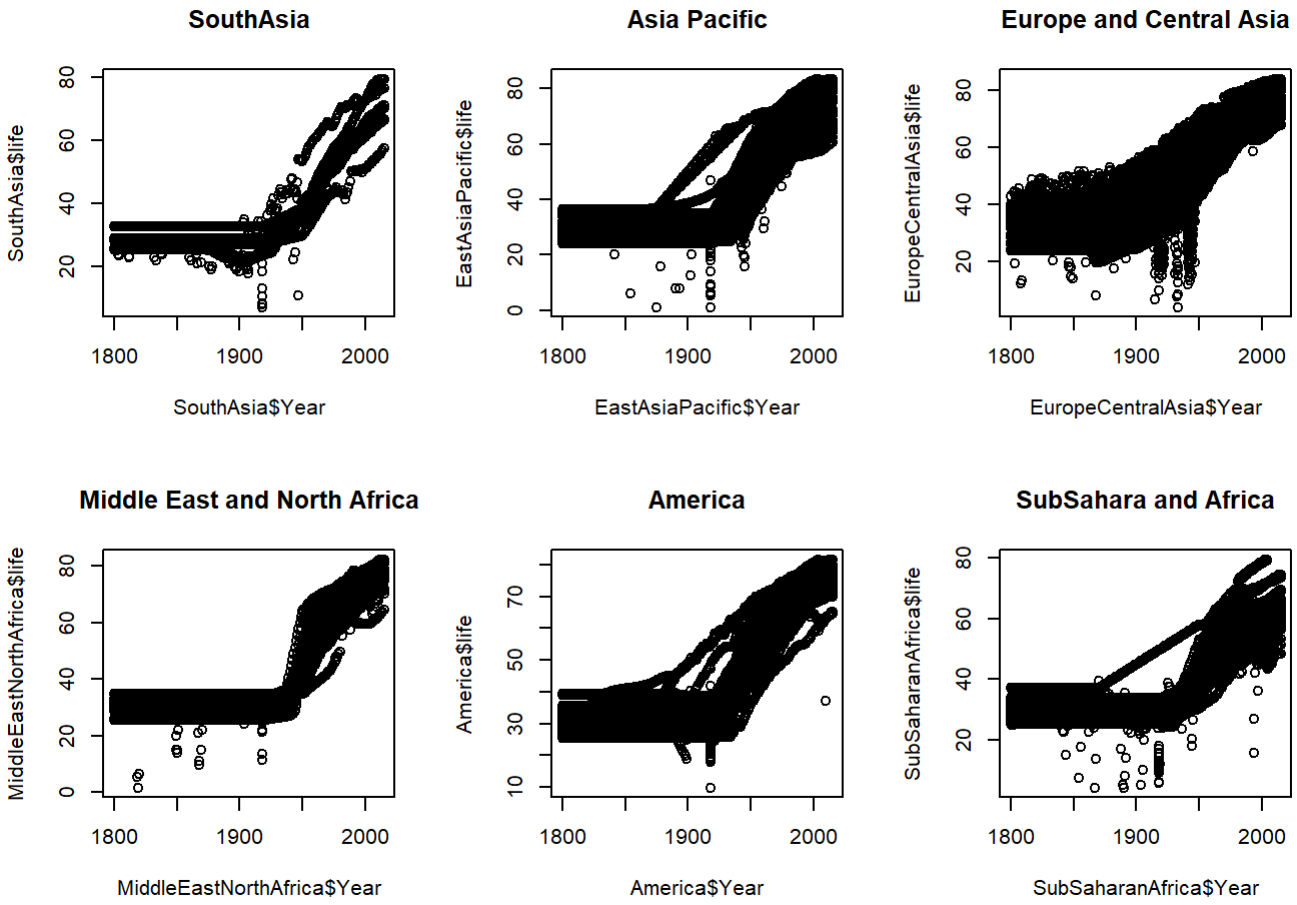



```
plot(log(gapminder$income), gapminder$life, main = "Life Expectancy vs log(Per Capita Income)")
```



3.1.4 Life expectancy trends over the years for different regions/subcontinents

```
par(mfrow=c(2,3))
plot(SouthAsia$Year, SouthAsia$life, main = "SouthAsia")
plot(EastAsiaPacific$Year, EastAsiaPacific$life, main = "Asia Pacific")
plot(EuropeCentralAsia$Year, EuropeCentralAsia$life, main = "Europe and Central Asia")
plot(MiddleEastNorthAfrica$Year, MiddleEastNorthAfrica$life, main = "Middle East and North Afric
a")
plot(America$Year, America$life, main = "America")
plot(SubSaharanAfrica$Year, SubSaharanAfrica$life, main = "SubSahara and Africa")
```



3.2 Exploratory data analysis for a single country (United States)

```
gapminderUS=filter(gapminder,Country=="United States")
```

Data summary and statistics

Five-number summaries for different variables for the US

```
fivenum(gapminderUS$life)
```

```
## [1] 31.000 39.410 50.550 69.938 79.100
```

```
fivenum(gapminderUS$population)
```

```
## [1]   6801854 170796378 218963561 266275528 321773631
```

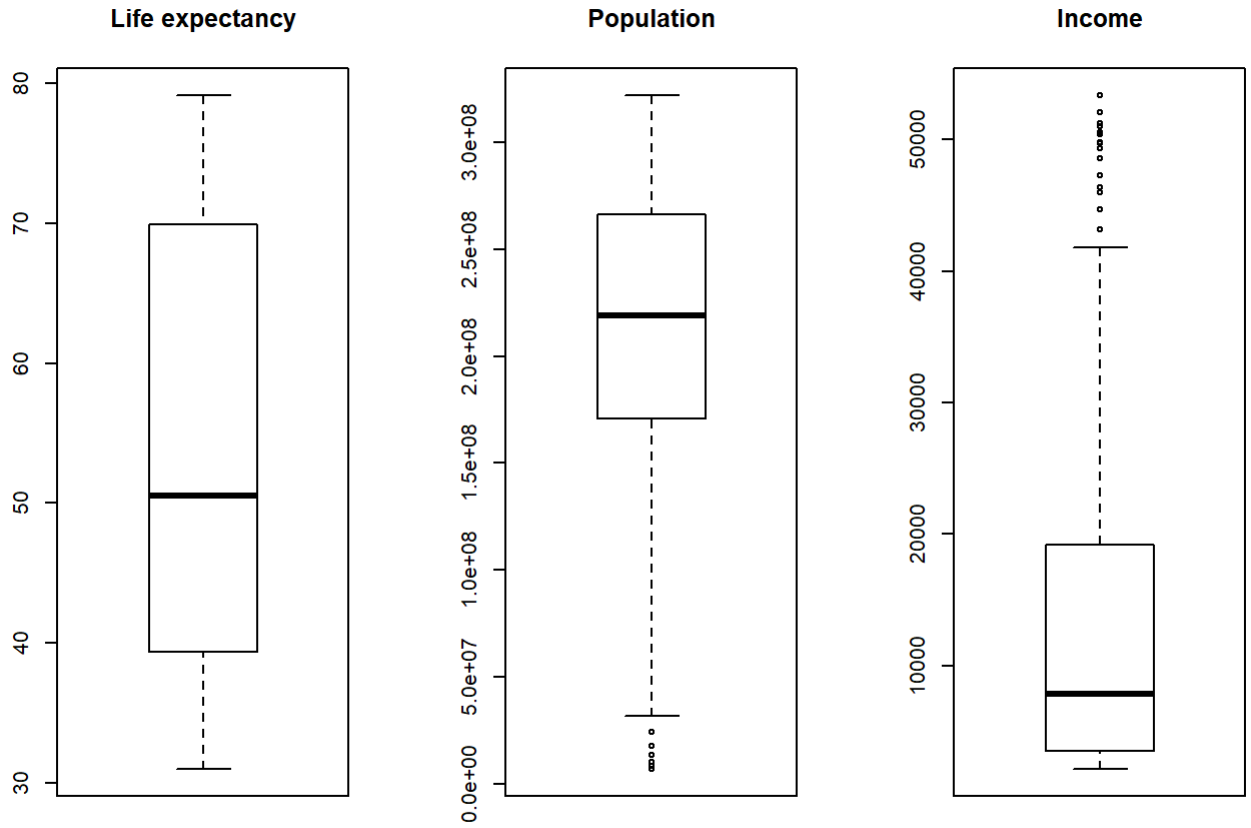
```
fivenum(gapminderUS$income)
```

```
## [1]  2115.0  3505.0  7875.0 19231.5 53354.0
```

Raw data visualization

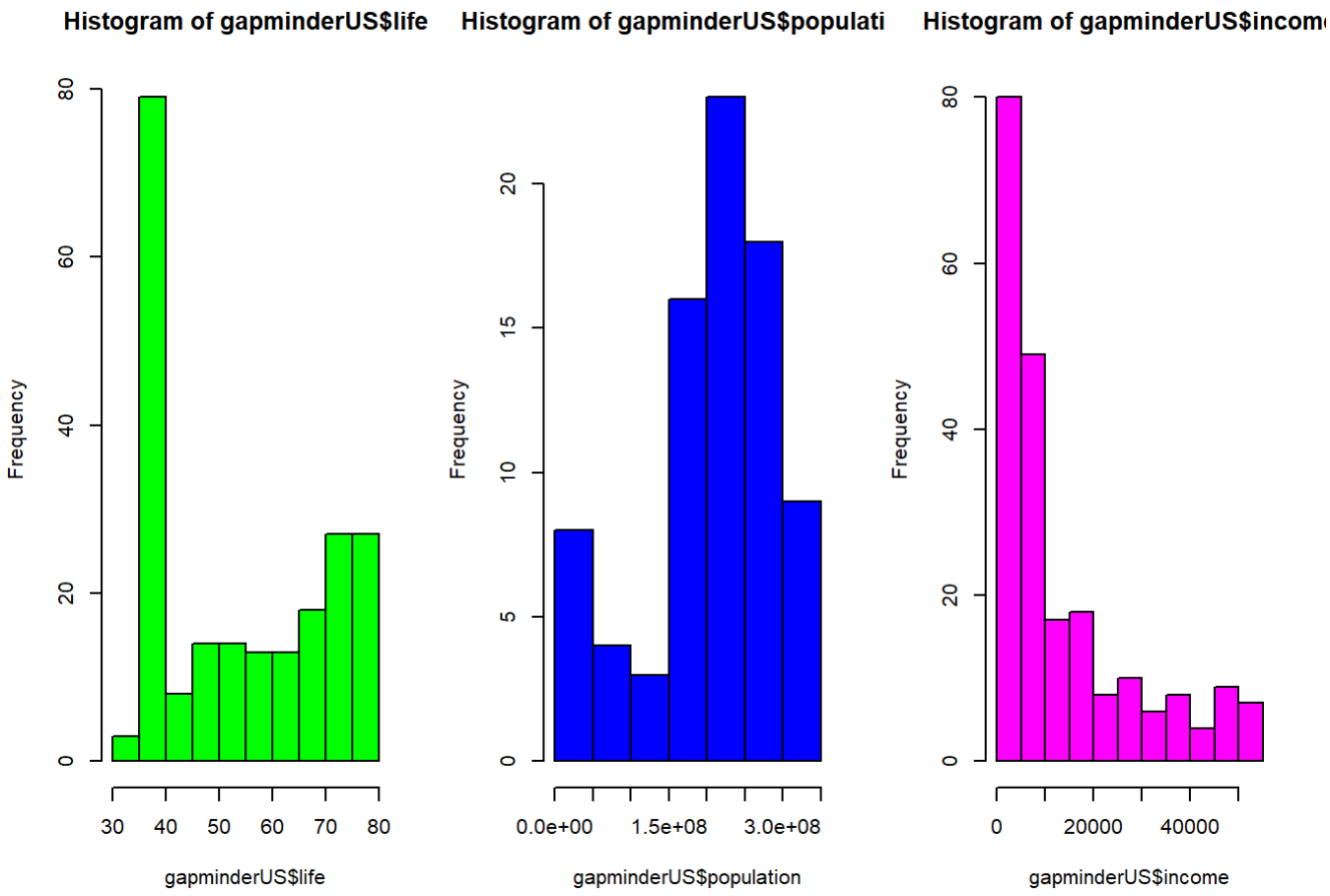
a) Boxplots for different variables for the US

```
par(mfrow=c(1,3))
boxplot(gapminderUS$life, main = "Life expectancy")
boxplot(gapminderUS$population, main = "Population")
boxplot(gapminderUS$income, main = "Income")
```



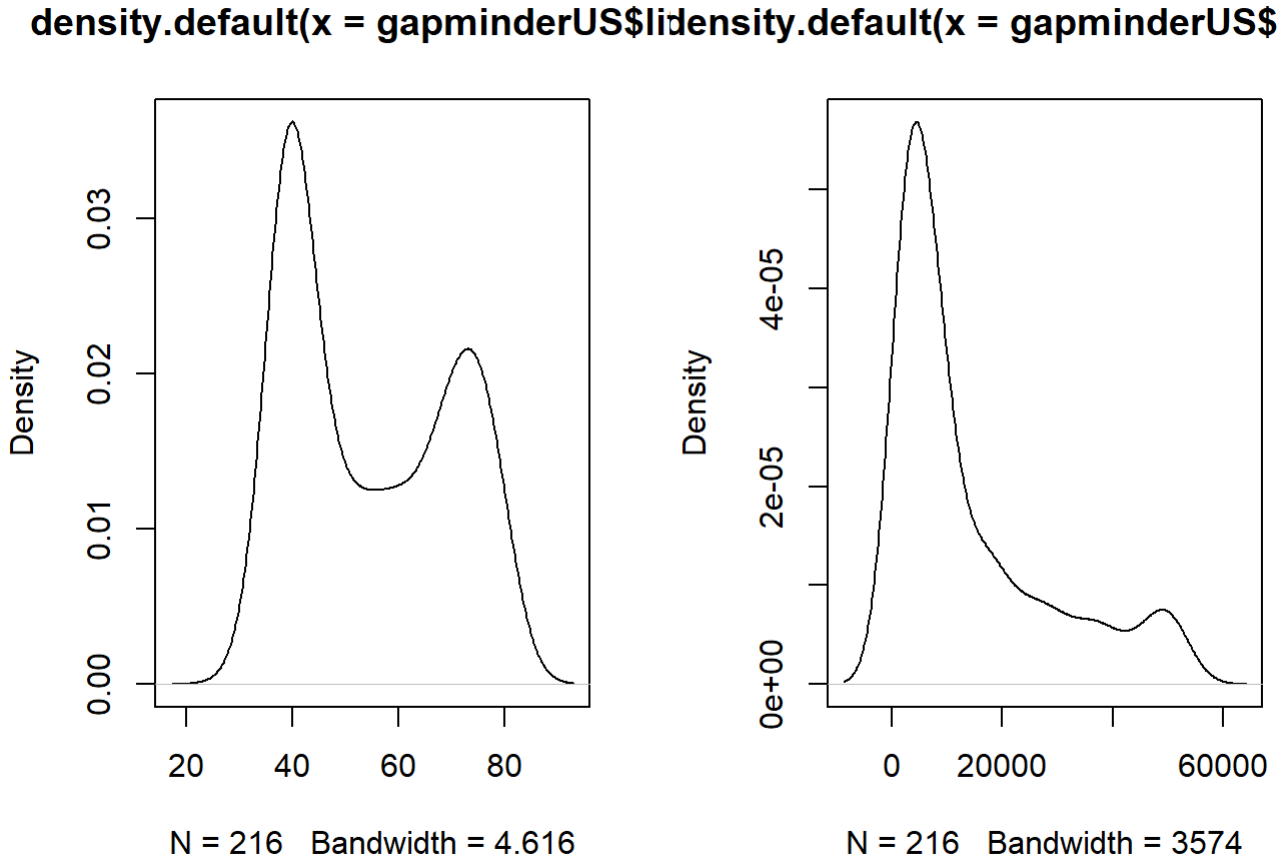
b) Histograms for different variable for the US

```
par(mfrow=c(1,3))
hist(gapminderUS$life,col = "green")
hist(gapminderUS$population,col = "blue")
hist(gapminderUS$income,col="magenta")
```



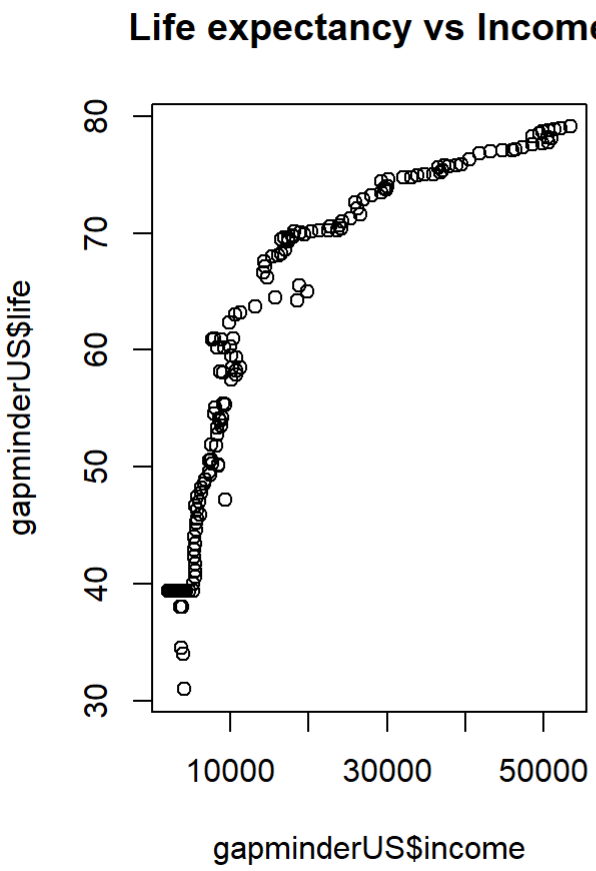
c) Density Plots for different variables for the US

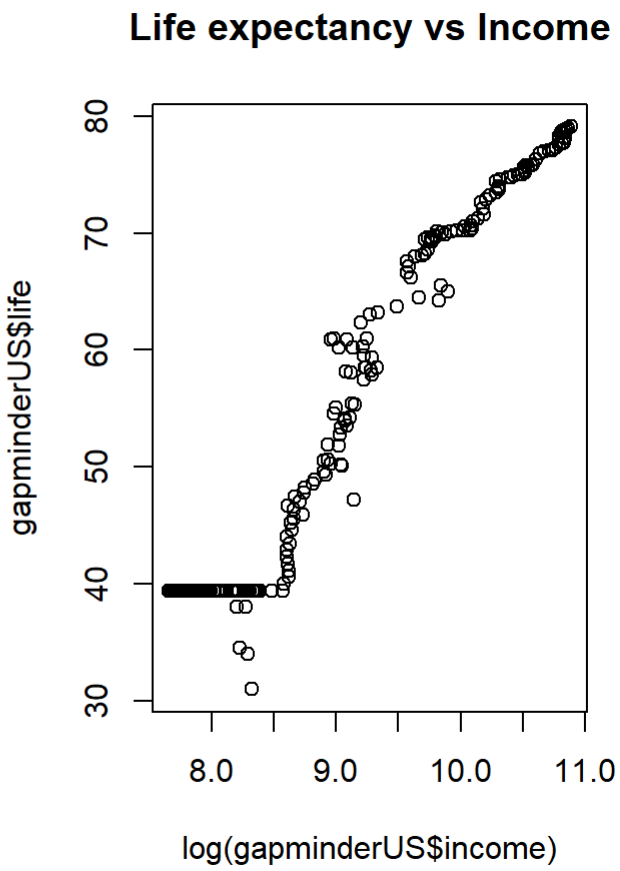
```
par(mfrow=c(1,2))
plot(density(gapminderUS$life))
plot(density(gapminderUS$income))
```



3.2.3. Analysis of ‘Life Expectancy’ vs ‘Per Capita Income’ for the United States

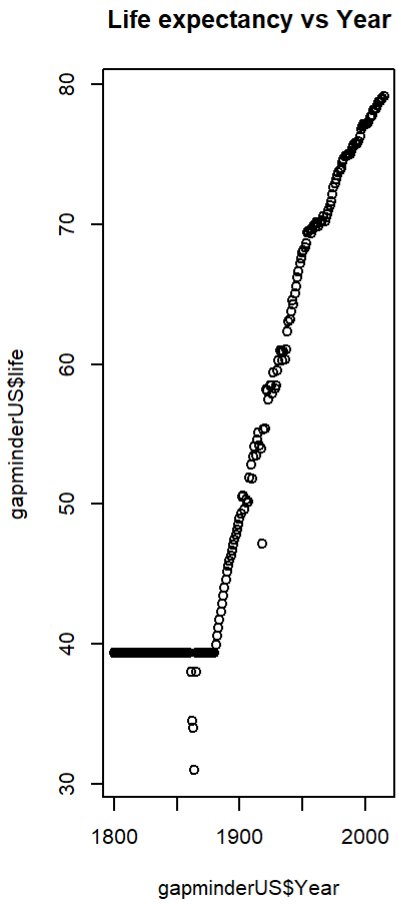
```
par(mfrow=c(1,2))
plot(gapminderUS$income, gapminderUS$life, main = "Life expectancy vs Income")
plot(log(gapminderUS$income), gapminderUS$life, main = "Life expectancy vs Income")
```

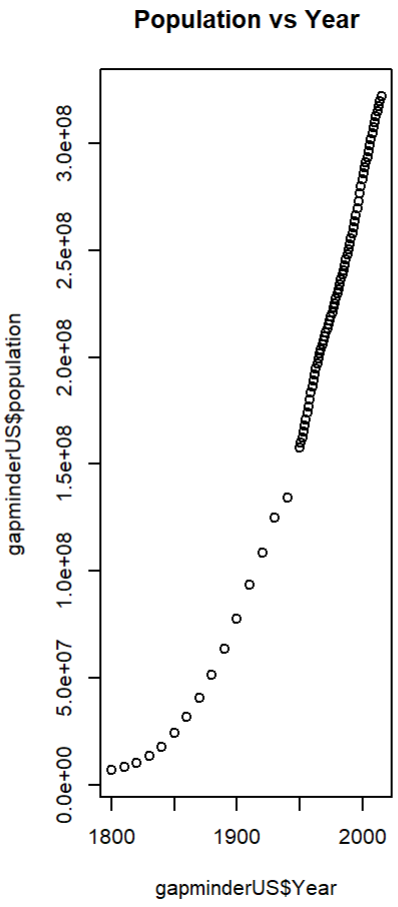


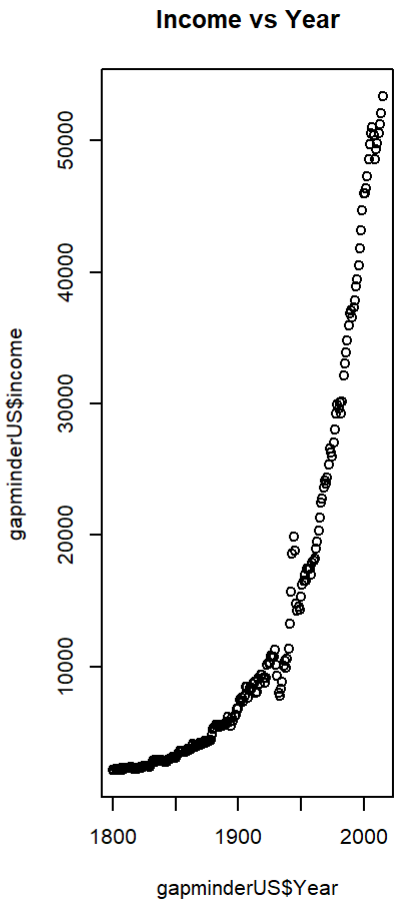


3.2.3. Analysis of ‘Life Expectancy’, ‘Population’ and ‘Per Capita Income’ variation over the years for the United States

```
par(mfrow=c(1,3))
plot(gapminderUS$Year, gapminderUS$life, main = "Life expectancy vs Year")
plot(gapminderUS$Year, gapminderUS$population, main = "Population vs Year")
plot(gapminderUS$Year, gapminderUS$income, main = "Income vs Year")
```


Life expectancy vs Year
gapminderUS$life
gapminderUS$Year
30
40
50
60
70
80
1800
1900
2000


Population vs Year
gapminderUS$population
gapminderUS$Year
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
3.0e+08
1800
1900
2000


Income vs Year
gapminderUS$income
gapminderUS$Year
10000
20000
30000
40000
50000
1800
1900
2000