IDS 2020. Problem Set 1

Muhammad Zahidul Islam Miaji

Last updated at 2021-03-17 19:15:55

```
#install.packages("gapminder")
library(gapminder)
library(tidyverse)

gapminder07 <- filter(gapminder, year %in% 2007)
```

Note: in any graphs, try to provide meaningful labels whenever possible.

Question 1

What is this data set about? (hint: Try ?gapminder to learn more about the data set) What are we doing when we write `gapminder07 <- filter(gapminder, year %in% 2007)`?

?gapminder it's a data set on life expectancy, GDP per capita, and population by country

```
gapminder07 <- filter(gapminder, year %in% 2007)
gapminder07
```

```
## # A tibble: 142 x 6
##    country     continent  year lifeExp       pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
##  1 Afghanistan Asia       2007    43.8  31889923      975.
##  2 Albania     Europe     2007    76.4   3600523     5937.
##  3 Algeria     Africa     2007    72.3  33333216     6223.
##  4 Angola      Africa     2007    42.7  12420476     4797.
##  5 Argentina   Americas   2007    75.3  40301927    12779.
##  6 Australia   Oceania    2007    81.2  20434176    34435.
##  7 Austria     Europe     2007    79.8   8199783    36126.
##  8 Bahrain     Asia       2007    75.6    708573    29796.
##  9 Bangladesh  Asia       2007    64.1 150448339     1391.
## 10 Belgium     Europe     2007    79.4  10392226    33693.
## # ... with 132 more rows
```

it means we are specify the data set by the year of 2007. now it will show the all countries data on lifeExp, pop and gdpPercap of year 2007.

Question 2

How many different countries are in the data? (hint: try glimpse, head, str, summary, or any other commands to get some sense about the data)

```
glimpse(gapminder)
```

```
## Rows: 1,704
## Columns: 6
## $ country   <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Afgha...
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 199...
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 4...
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372,...
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.113...
```

```
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 163
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

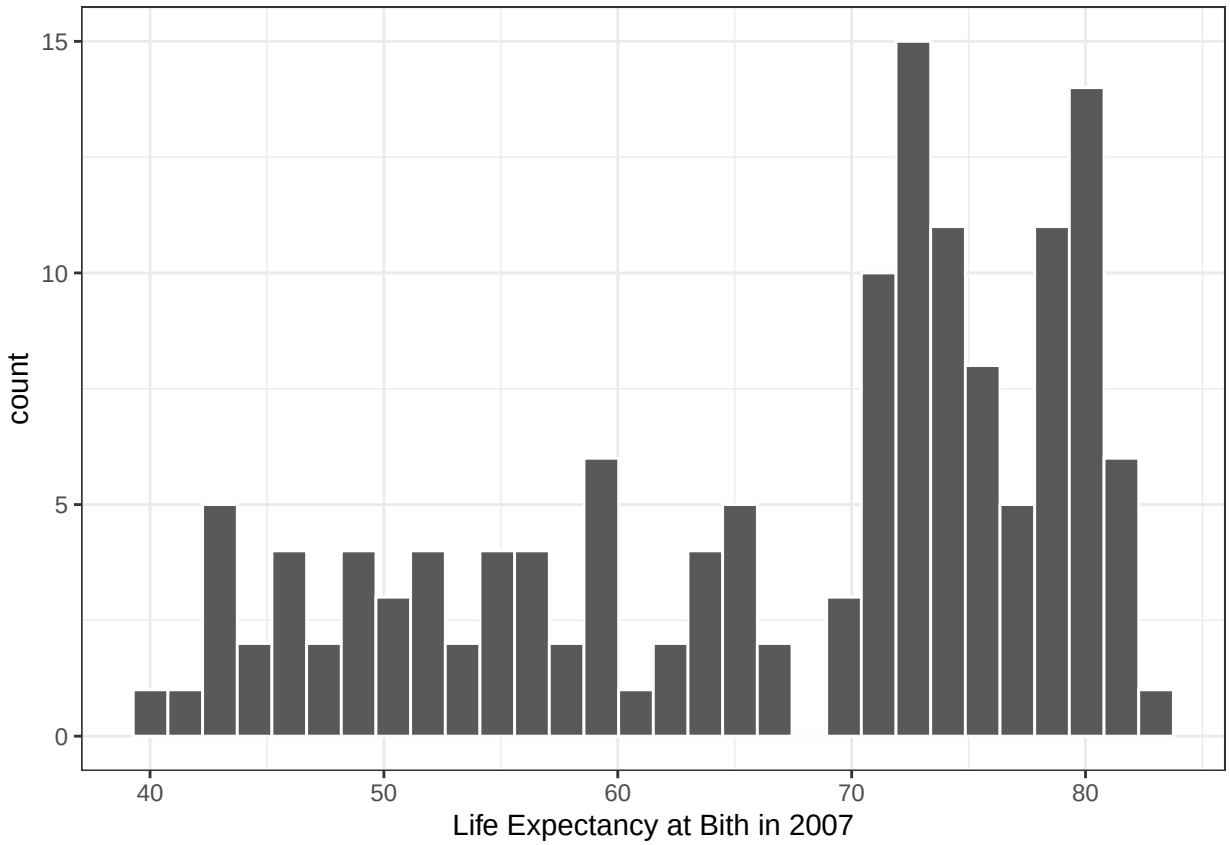
Total 142 different countries exist in the data level

Question 3

What is the distribution of the life expectancy at birth in 2007? (Note: here on out, we will be using the `gapminder07` data for the rest of the exercise. Hint: Try drawing a histogram)

```
histo007 <- ggplot(data = gapminder07) +
  geom_histogram(aes(x = lifeExp),
                 color = "white") +
  scale_x_continuous("Life Expectancy at Bith in 2007") +
  theme_bw()
histo007
```

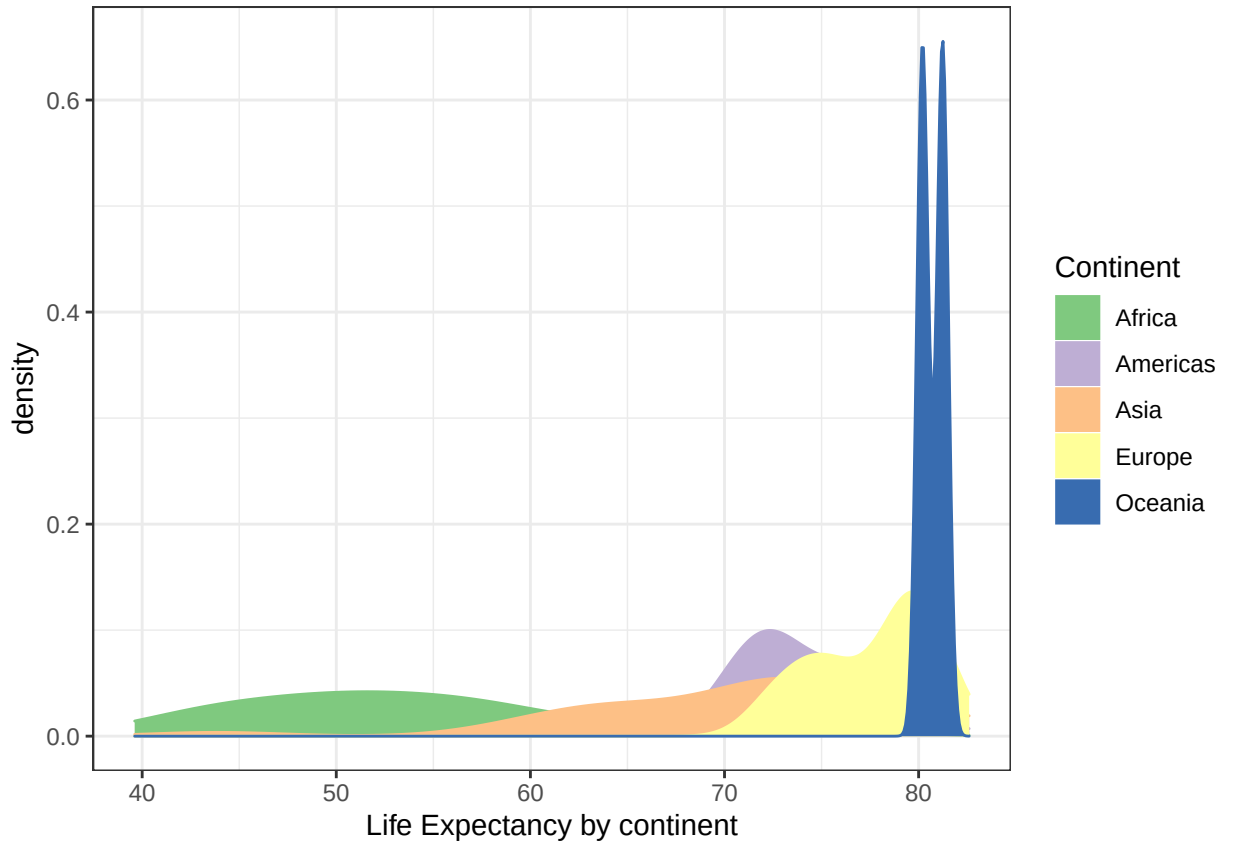
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 4

Now try depicting the same information using a density plot. This time, try also to color the distribution by continent (hint: use aesthetic `fill` to specify the coloring for the continents.)

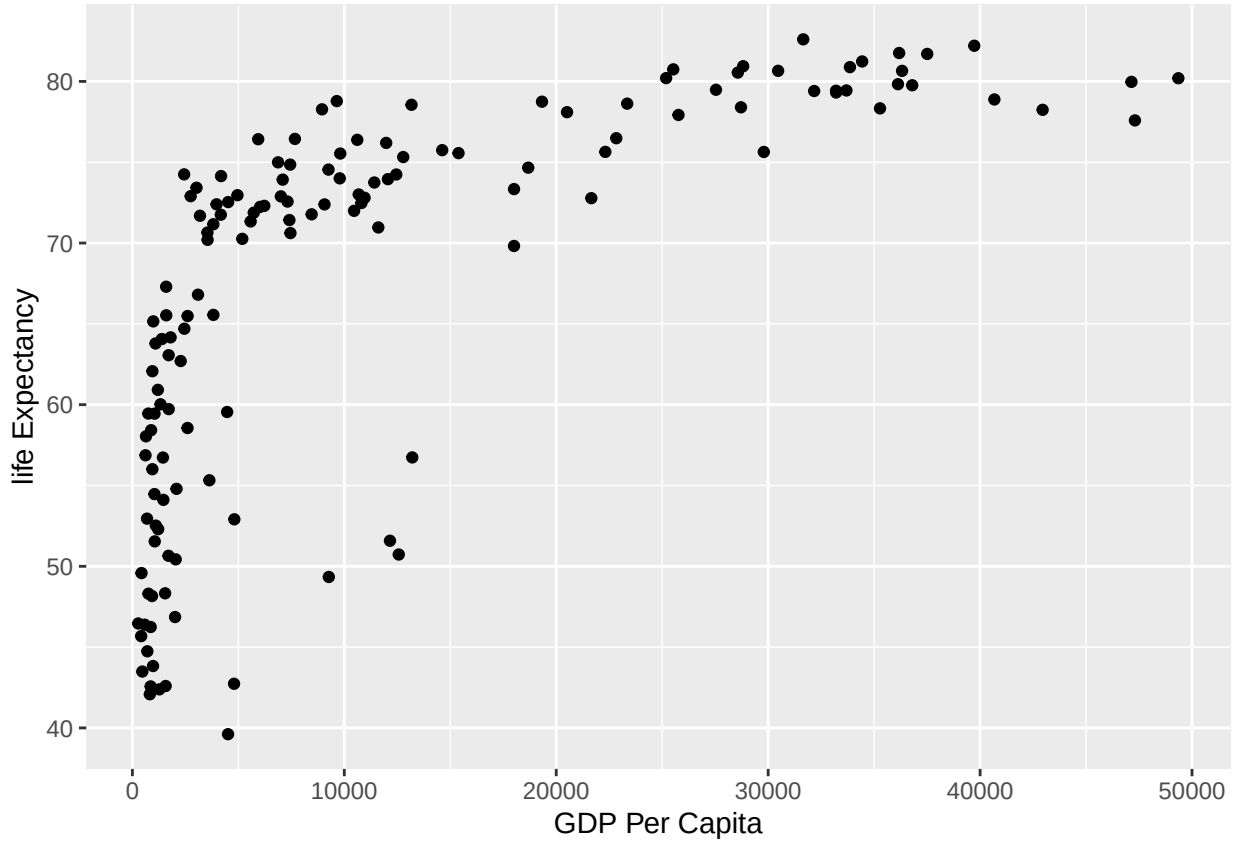
```
ggplot(data = gapminder07) +
  geom_density(aes(x = lifeExp, color = continent, fill = continent))+
  scale_color_brewer("Continent", palette = "Accent") +
    scale_fill_brewer("Continent", palette = "Accent") +
  scale_x_continuous("Life Expectancy by continent") +
  theme_bw()
```



Question 5

What is the relationship between GDP per capita and life expectancy? (Hint: try to draw out a scatter plot using points)

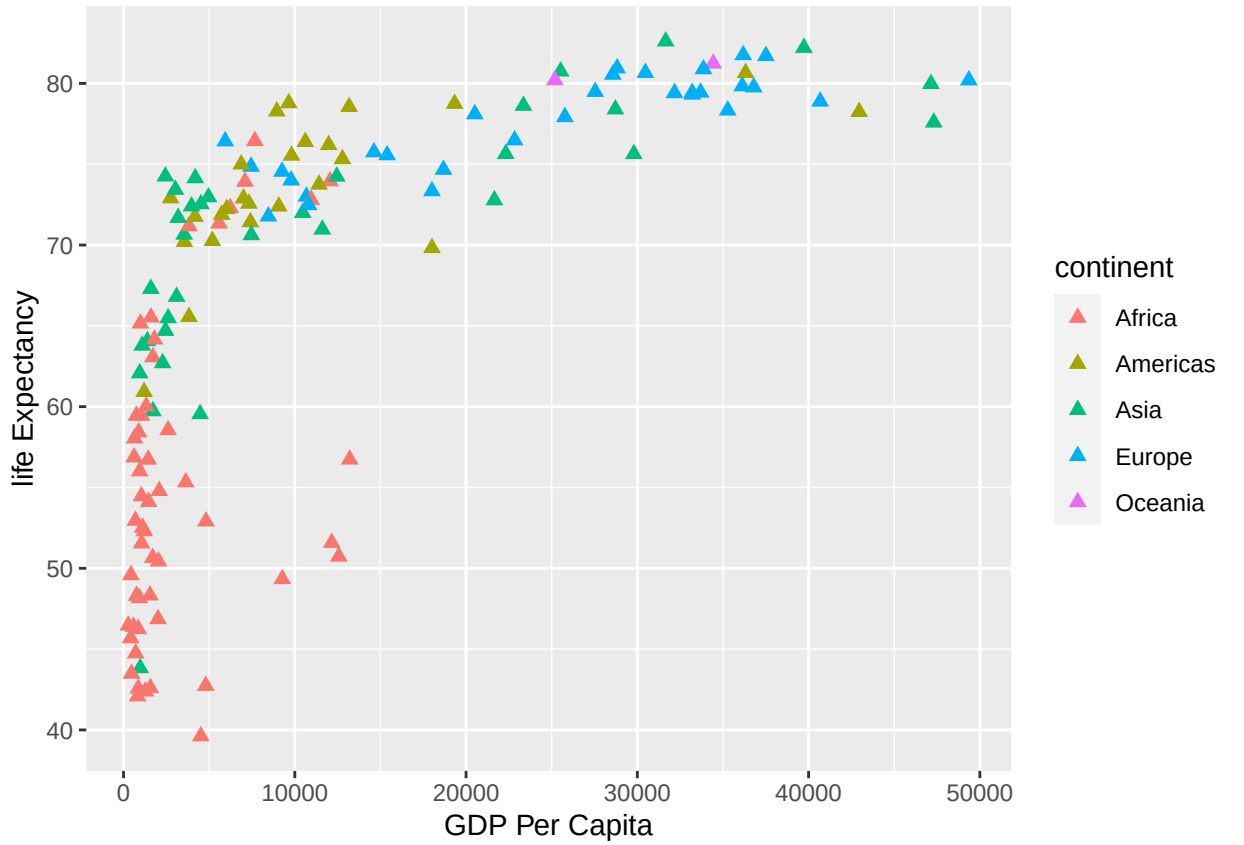
```
ggplot(data = gapminder07) +
  geom_point(aes(x = gdpPercap, y = lifeExp)) +
  xlab("GDP Per Capita") + ylab("life Expectancy")
```



Question 6

Next, try to plot the same information, this time with the points colored by continents and varying in their size by population (hint: try using `color` and `size` in the `aes`)

```
ggplot(data = gapminder07) +
  geom_point(aes(x = gdpPercap, y = lifeExp, color = continent), size = 2, shape= 17) +
  xlab("GDP Per Capita") + ylab("life Expectancy")
```



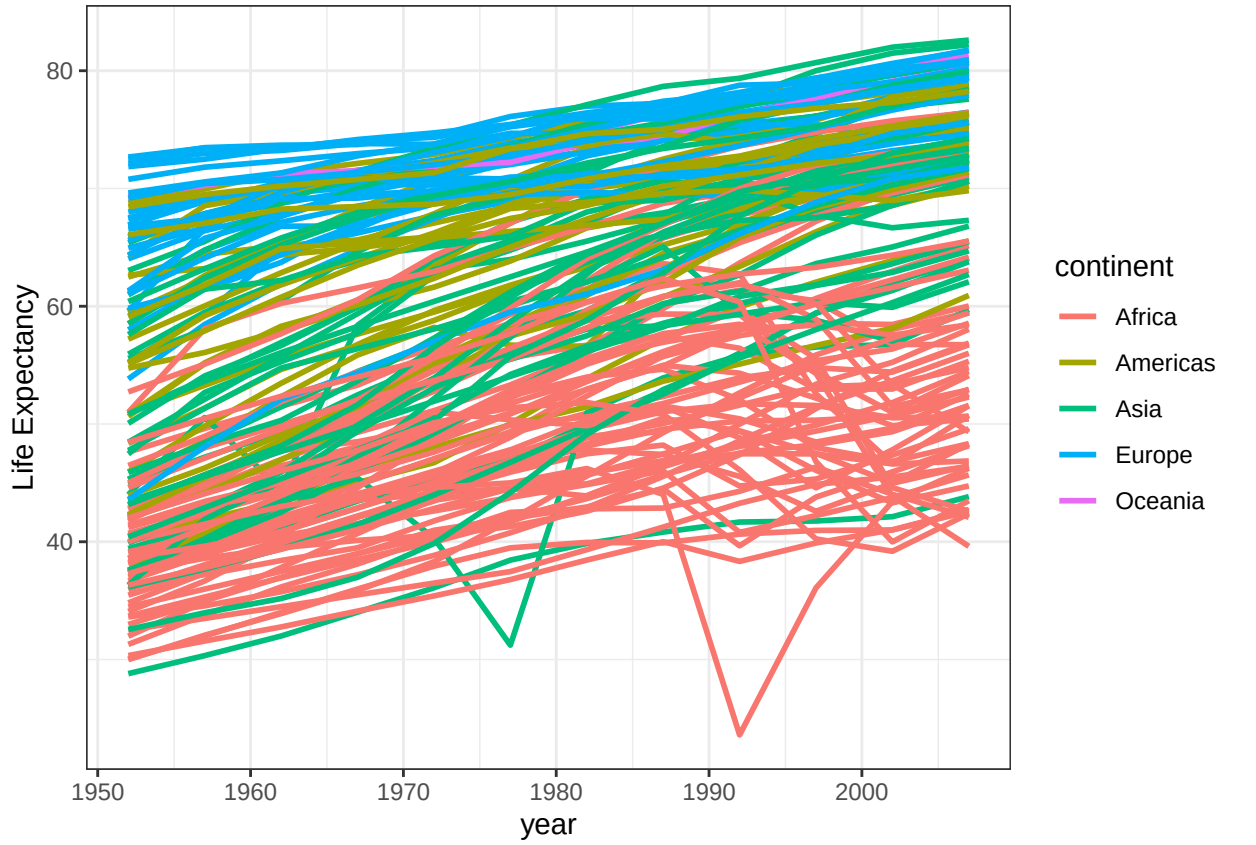
Question 7

This time, use `gapminder` instead of `gapminder07` as your dataset. Using a line graph, plot the overtime change in the life expectency across different countries and color them by continents. (hint: draw a line graph. put the year on the x-axis and the life expectancy on the y-axis. Also specify `group = country` and `color=continent` in the `aes`.)

```
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

```
ggplot(data = gapminder)+
  geom_line(aes(x= year, y= lifeExp, group= country, color= continent), lwd=1)+
  xlab("year")+ ylab("Life Expectancy")+
  theme_bw()
```

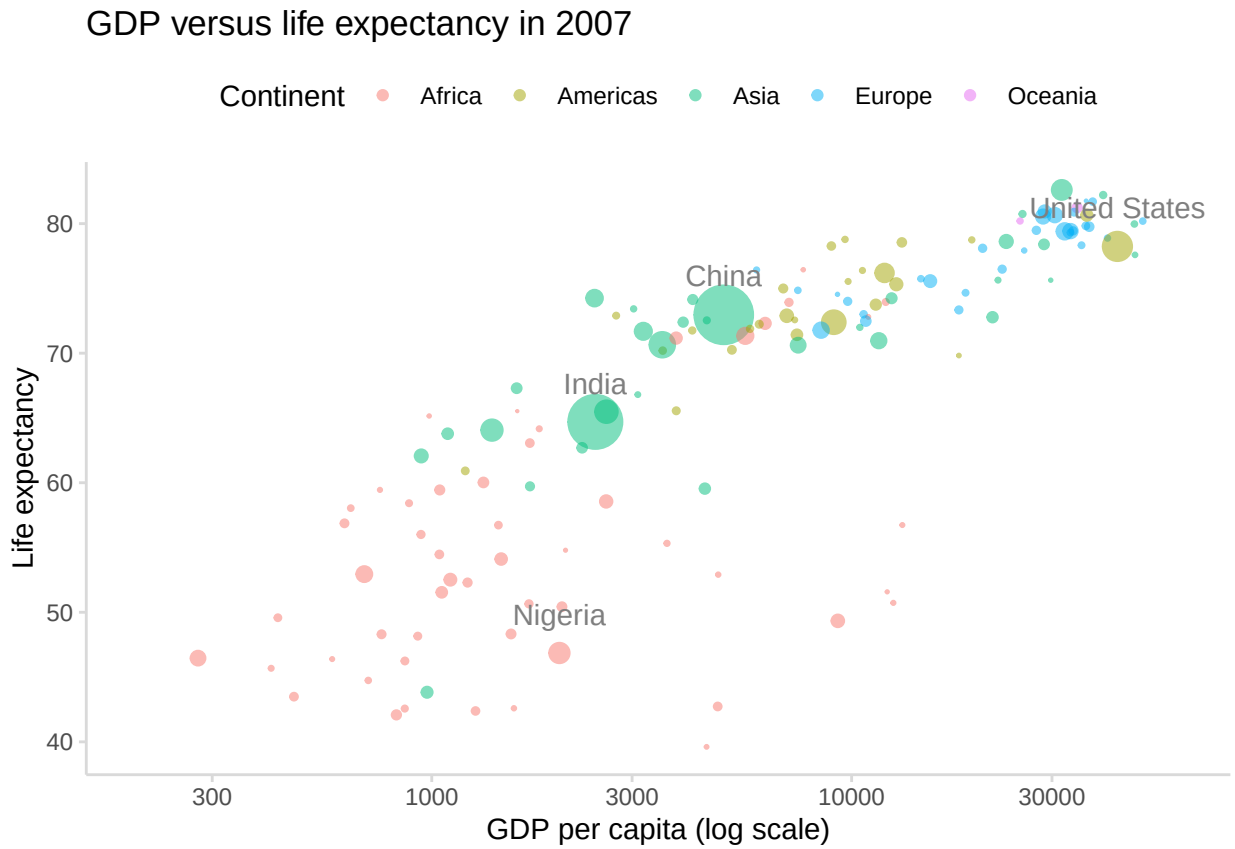


Question 8

Finally, try executing the following code, and write in words what your take away is from the information provided in the resulting graph. (hint: you need to remove the `eval=FALSE` from the code chunk bracket below.)

```
ggplot(gapminder07) +
  # add scatter points
  geom_point(aes(x = gdpPercap, y = lifeExp, color = continent, size = pop),
             alpha = 0.5) +
  # add some text annotations for the very large countries
  geom_text(aes(x = gdpPercap, y = lifeExp + 3, label = country),
            color = "grey50",
            data = filter(gapminder07, pop > 1000000000 | country %in% c("Nigeria", "United States"))) +
  # clean the axes names and breaks
  scale_x_log10(limits = c(200, 60000)) +
  # change labels
  labs(title = "GDP versus life expectancy in 2007",
       x = "GDP per capita (log scale)",
```

```
      y = "Life expectancy",
      size = "Popoulation",
      color = "Continent") +
# change the size scale
scale_size(range = c(0.1, 10),
           # remove size legend
           guide = "none") +
# add a nicer theme
theme_classic() +
# place legend at top and grey axis lines
theme(legend.position = "top",
      axis.line = element_line(color = "grey85"),
      axis.ticks = element_line(color = "grey85"))
```



This code used for creating scatterplot. it is usually shows us relationship between two continuous variables. In the code we try to discover "gapminder07" data, GDP per capital in the x-axis and life expectancy in the y-axis. we labeled continents in various color and marked bubble size with population size. we try to see the USA and Nigeria population and which countries population > 1000000000.

In the graph, we are seeing that USA has higher GDP per capita and higher life expectancy than designated criteria (Nigeria and population > 1000000000). we also get china and india are the largest population size (Bubble size). Finally we can conclude that, USA is the far ahead than china and India but Nigeria has very low life expectancy rate.