Fundamentals of R

Block 3 - Practical Visualizations

Henrique Sposito and Livio Muller-Silva

2022-10-14

R Markdown

Markdown is a simple formatting syntax for authoring HTML, PDF, and Word documents.

Creating an R Markdown document is just like an R script, you just have to click the new document button and select R Markdowm from the options.

You can embed an R code chunk like this:

In the case above, we are just adjusting the setup for the document and loading some packages for our R Markdown document.

This is the best resource for information on R Markdown!

Some Basics:

Section headers work with #:

First-level header

Second-level header

Third-level header

For changing text styles use *:

Italics

Bold

Italics and bold

For inserting R code click on the **C** buttom above or use Cmd + Option + I on MAC (for Windows: Ctrl + Alt + I).

```r
as.character("R Markdown is awesome")
```

```
## [1] "R Markdown is awesome"
```

Code chunks can be evaluated (run code?), included (should it be displayed in knitted document?), and much more. rmarkdown, as a tidyverse package, also has a cheat sheet!

When you click the **Knit** button a document in HTML or PDF can be generated that includes both content as well as the output of any embedded R code chunks within the document.

Lastly, R Markdown can be further used to create presentations in R (as the ones we use in class, see the xaringan package) or even to write your Master's thesis (check out iheidown).

Visualizations

Mind the GAP

Do you remember the gapminder package and data?

```r
gapminder <- gapminder::gapminder # create an object
summary(gapminder) # summary data
```

The ggplot2 canvas

Before we start, the holy grail duide of visualizations books, using the ggplot2 package, is open source and was written in R Markdown.

Do not forget that ggplot2 is a tidyverse package, therefore, there is a cheat sheet for it!

In ggplot2, we use the '+' operator instead of the '%>%'. Remember the '%>%' takes the result of a previous operation and uses as the first argument to the following operation (for more details see here). The '+', instead, adds layers to a ggplot2 plot (for more information see here)

For example, without layers a ggplot2 plot looks like this:

```r
gapminder %>%
  ggplot(aes(x = year,
             y = lifeExp))
```

This is just an empty canvas waiting to be filled with your art!

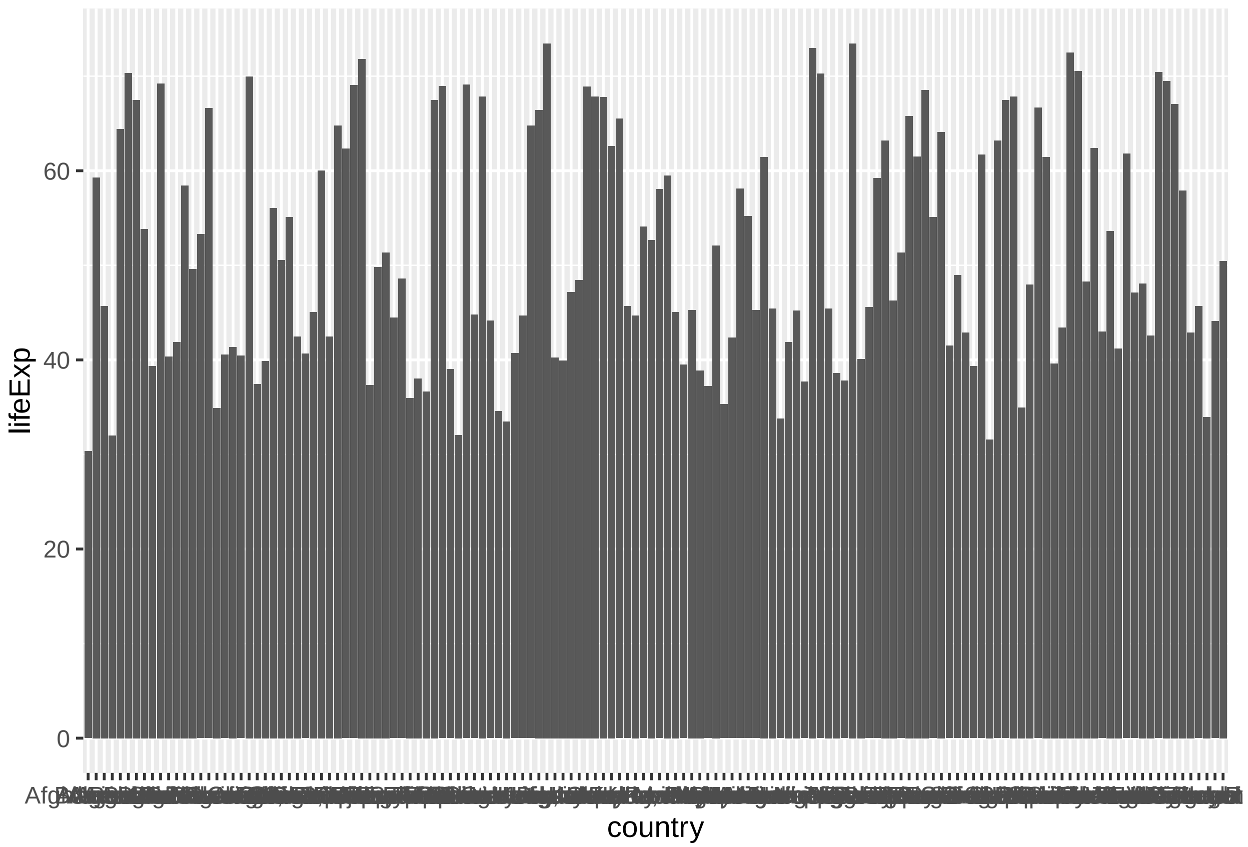
Life expectancy from 1957 to 2007 across continents (Bar plots)

To create bar plot in ggplot2 we use the `geom_bar()` function.

What are bar plots good for?

Would a bar plot be a good choice to plot life expectancy by country in 1957?

```r
gapminder %>% # select data
  filter(year == 1957) %>% # filter for years of interest
  ggplot(aes(x = country, # map country in the x axis
             y = lifeExp)) + # map average life expectancy in the y axis
  geom_bar(stat = "identity") # add bars
```



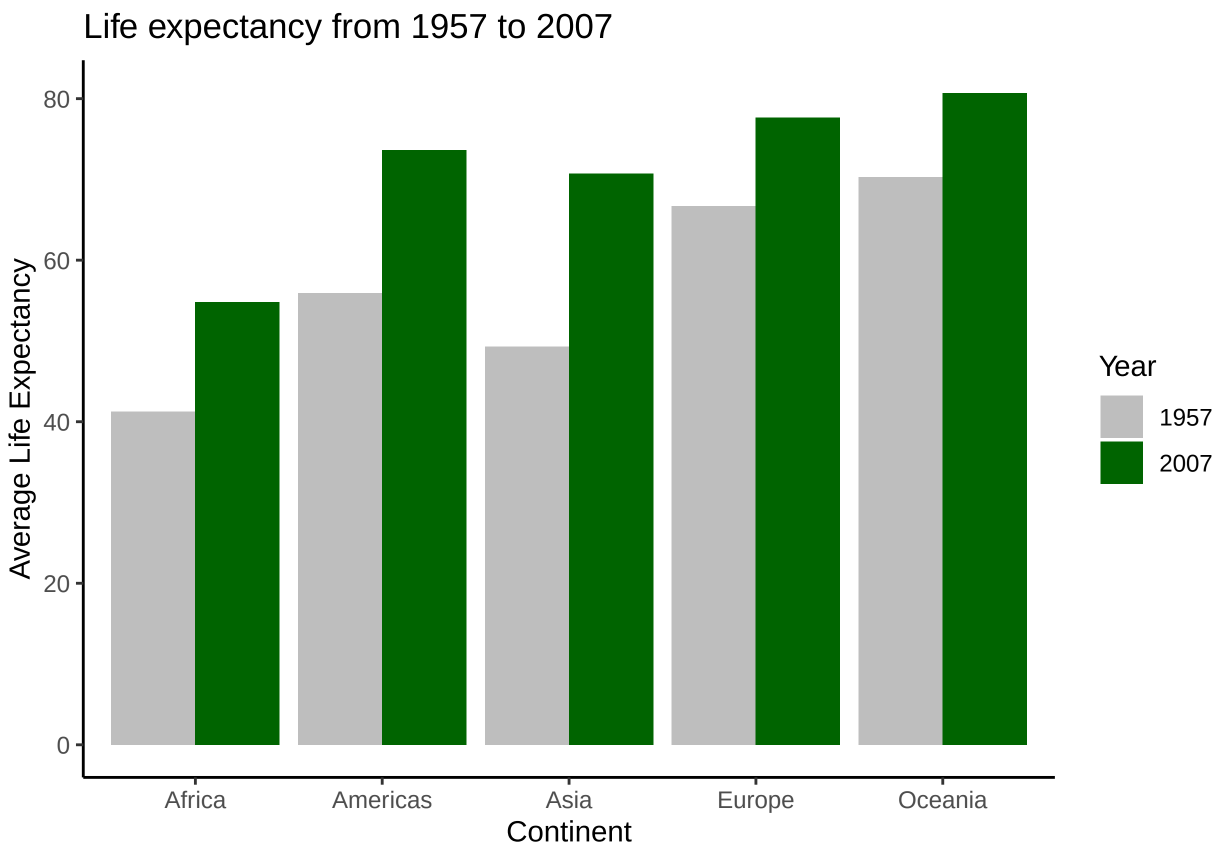
```
# stat = "identity" tells R that the values are provided and need not be counted!
```

This is clearly not a good plotting choice (too much information)…

Instead, let's use bars to plot the average difference in life expectancy from 1957 to 2007 across continents.

```
gapminder %>% # select data
  filter(year == 1957 | year == 2007) %>% # filter for years of interest
  group_by(continent, year) %>% # group by year and country
  summarise(Avg_life_expectancy = mean(lifeExp)) %>% # get the means by the groups
  ggplot(aes(x = continent, # map continent in the x axis
             y = Avg_life_expectancy, # map average life expectancy in the y axis
             fill = as.factor(year))) + # fill mapping by year
  geom_bar(stat = "identity", # adding bars to plot according to the fill mapping
           position = "dodge") + # defining the position of stat in bars
  labs(title = "Life expectancy from 1957 to 2007", # add a title
       x = "Continent", # add a label for x axis
       y = "Average Life Expectancy", # add a label for y axis
       fill = "Year") + # sub legend for fill
  scale_fill_manual(values = c("gray", "darkgreen")) + # manually set colors
  theme_classic() # add a theme
```

```
## ‘summarise()‘ has grouped output by ’continent’. You can override using the
## ‘.groups‘ argument.
```



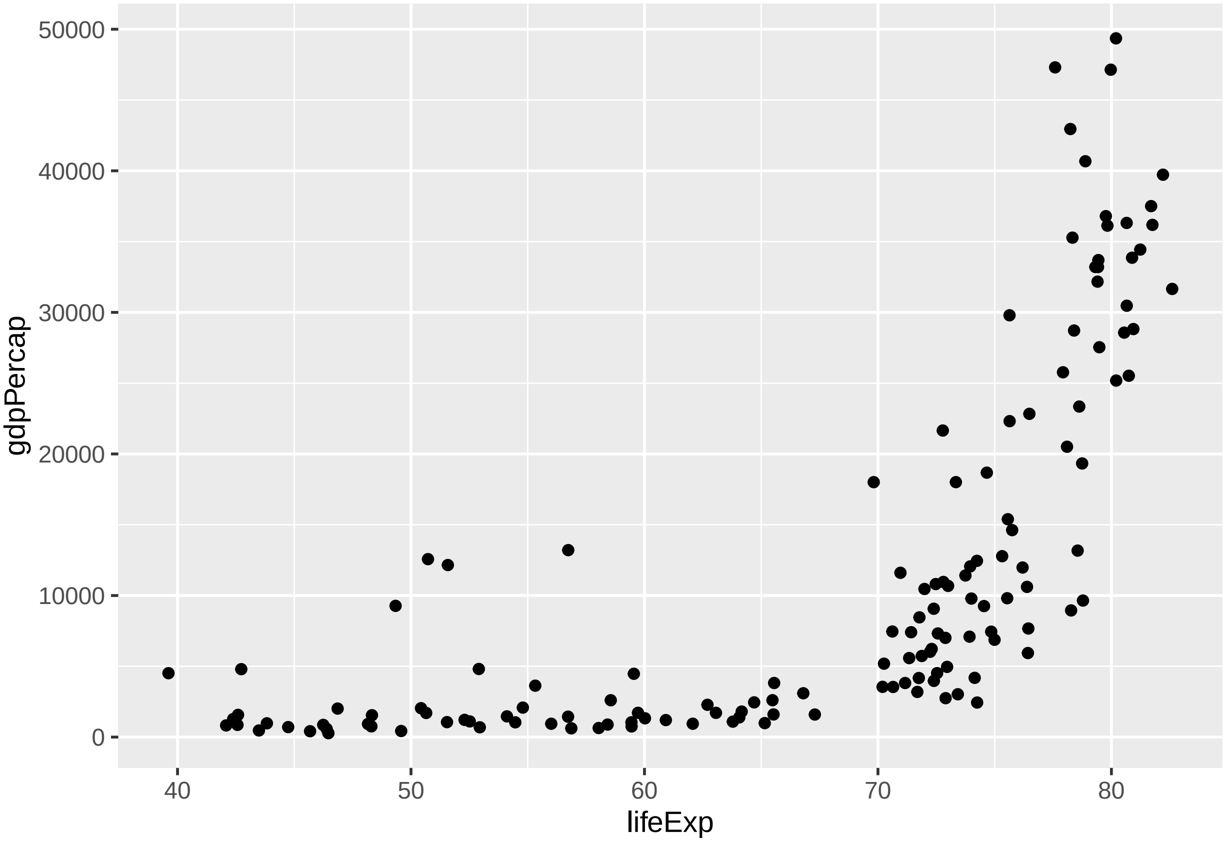
Population, life expectancy and GDP (Scatter plots)

To create scatter plots in ggplot2 we use the `geom_point()` function.

What are scatter plots good for in this case?

Let's plot population and GDP per capita, in 2007!

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = lifeExp,
             y = gdpPercap)) +
  geom_point() # add points to plot
```

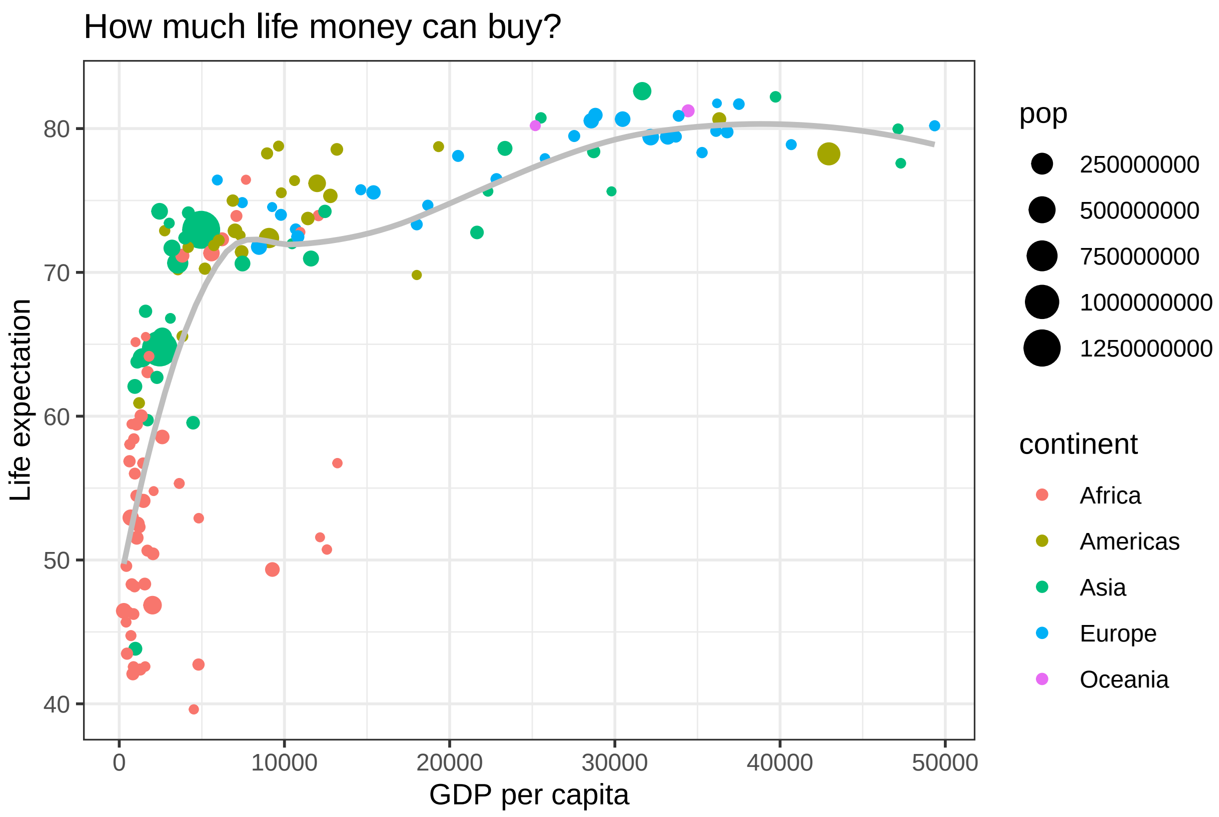


Is this plot informative? How could we improve this?

What if we focus on population, life expectation and GDP?

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point(aes(size = pop, color = continent)) + # color all points blue
  geom_smooth(se = FALSE, color = "gray") + # add a smoothed line
  labs(title = "How much life money can buy?",
       x = "GDP per capita",
       y = "Life expectation",
       caption = "Source: Gapminder") + # add caption
  theme_bw() # add theme
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula = ’y ~ x’
```



Source: Gapminder

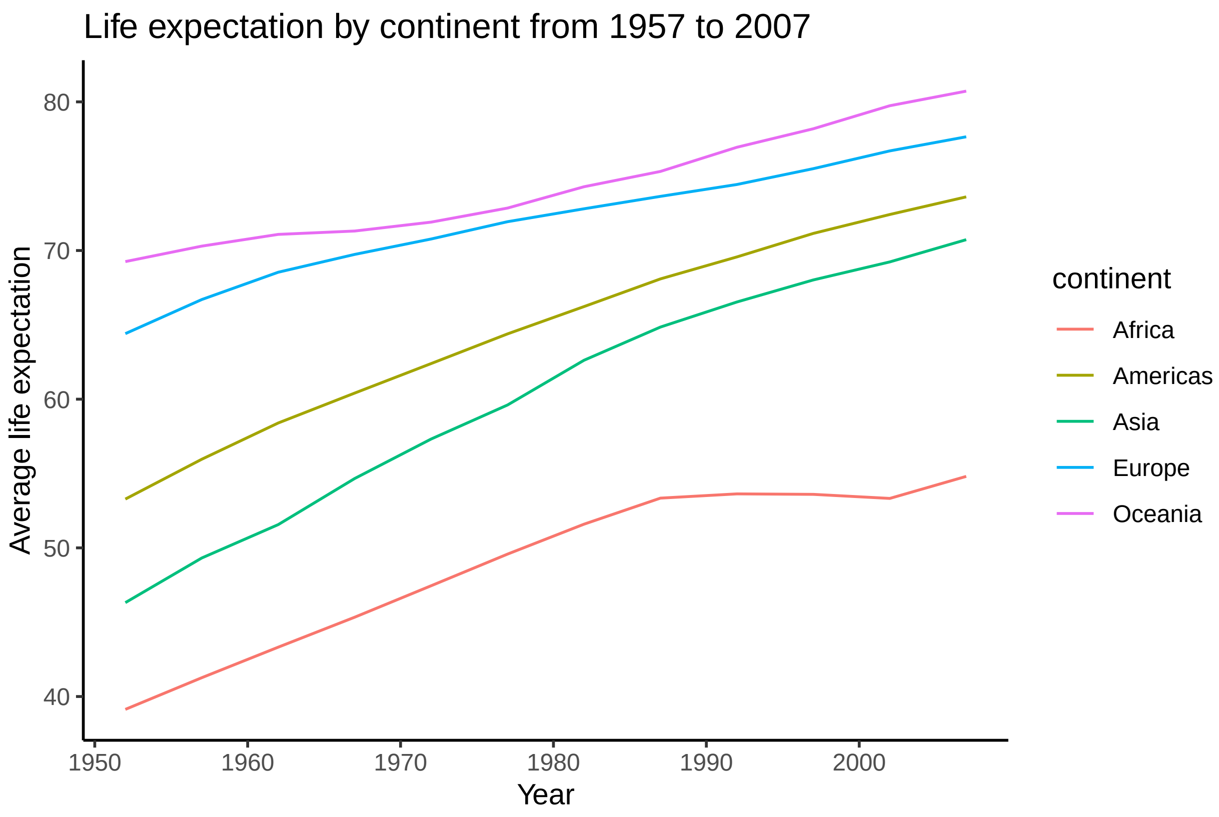
The evolution of life expectancy (Line plots)

To create scatter plots in ggplot2 we use the `geom_line()` function.

What are line plots good for in this case?

```
gapminder %>%
  group_by(continent, year) %>% # group by year and country
  summarise(Avg_life_expectancy = mean(lifeExp)) %>%  #  mean life expectancy
  ggplot(aes(x = year, y = Avg_life_expectancy)) +
  geom_line(aes(color = continent)) + # color all points blue
  labs(title = "Life expectation by continent from 1957 to 2007",
       x = "Year",
       y = "Average life expectation",
       caption = "Source: Gapminder") + # add caption
  theme_classic() # add theme
```

```
## ‘summarise()‘ has grouped output by ’continent’. You can override using the
## ‘.groups‘ argument.
```



Source: Gapminder

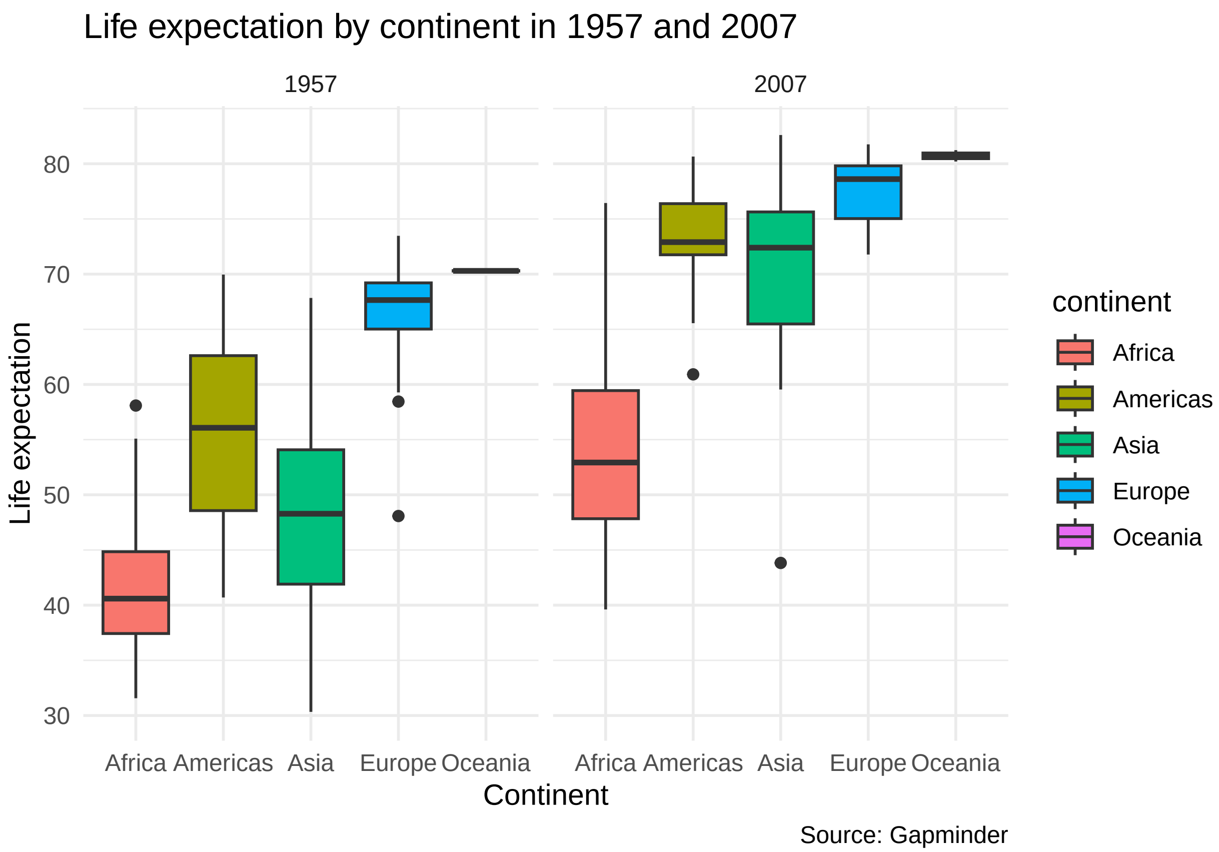
Could you make the same line plots for GDP per capita across continents in time?

Distribution of life expectancies across continent (Box plots)

To create scatter plots in ggplot2 we use the `geom_boxplot()` function.

What are box plots good for in this case?

```
gapminder %>%
  filter(year == 1957 | year == 2007) %>%
  ggplot(aes(x = continent, y = lifeExp, fill = continent)) +
  geom_boxplot() +
  facet_wrap("year") +
  # facet_wrap("continent") +
  labs(title = "Life expectation by continent in 1957 and 2007",
       x = "Continent",
       y = "Life expectation",
       caption = "Source: Gapminder") + # add caption
  theme_minimal()
```



Can you make the same box plot for GDP?