Fundamentals of R

Block 3 - Practical Visualizations

Henrique Sposito and Livio Muller-Silva

2022-10-14

R Markdown: another way to store code

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.

Markdown allows you to mix chunks of code (in light grey) with text and export a document with your code, text, and plots.

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** button above or use Cmd + Option + I on MAC (for Windows: Ctrl + Alt + I).

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

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

Code chunks can be evaluated (should code be run?), included (should the code 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 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

Setting up the Gapminder data

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

Before we start, the ggplot2 book is a great source for you to learn the details of visualizations in R (and the book was written using an R Markdown).

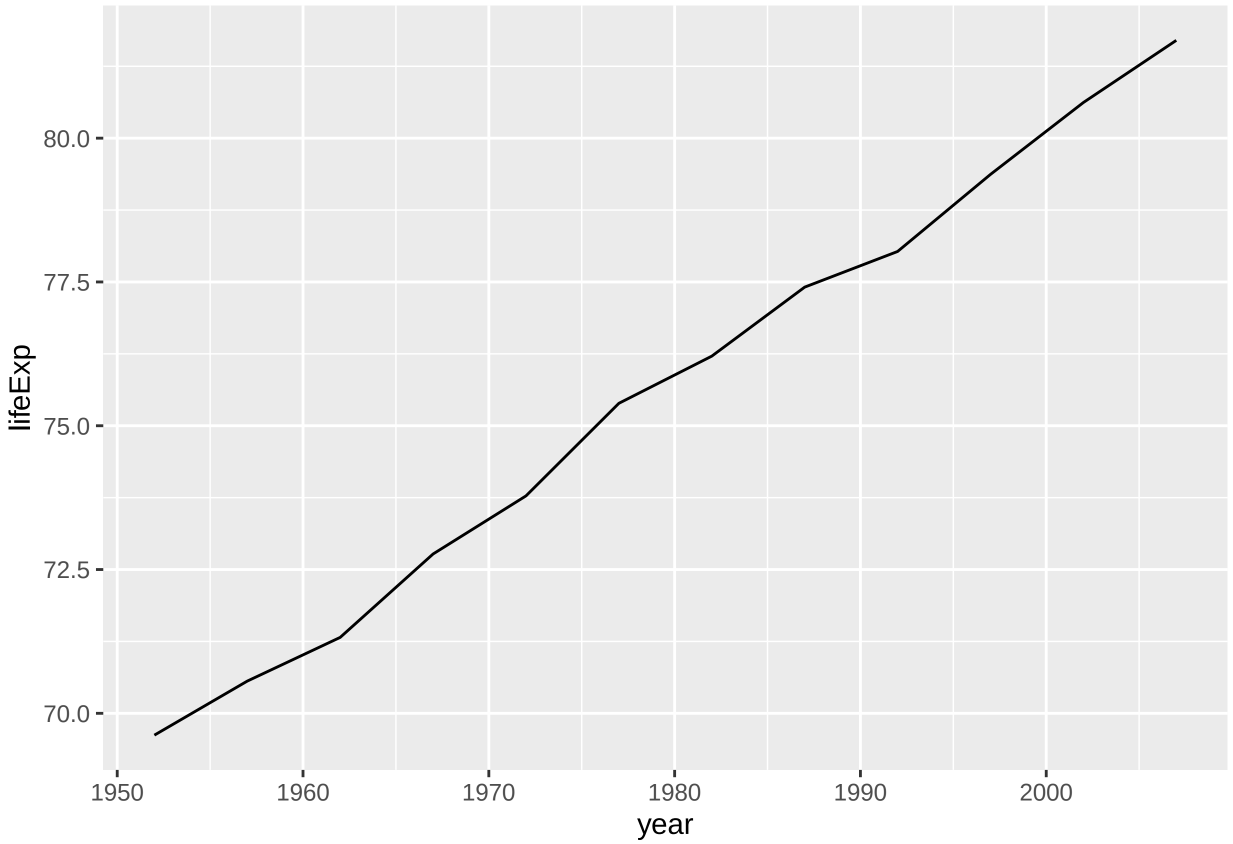
Line plots: The evolution of life expectancy

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

What are line plots good for?

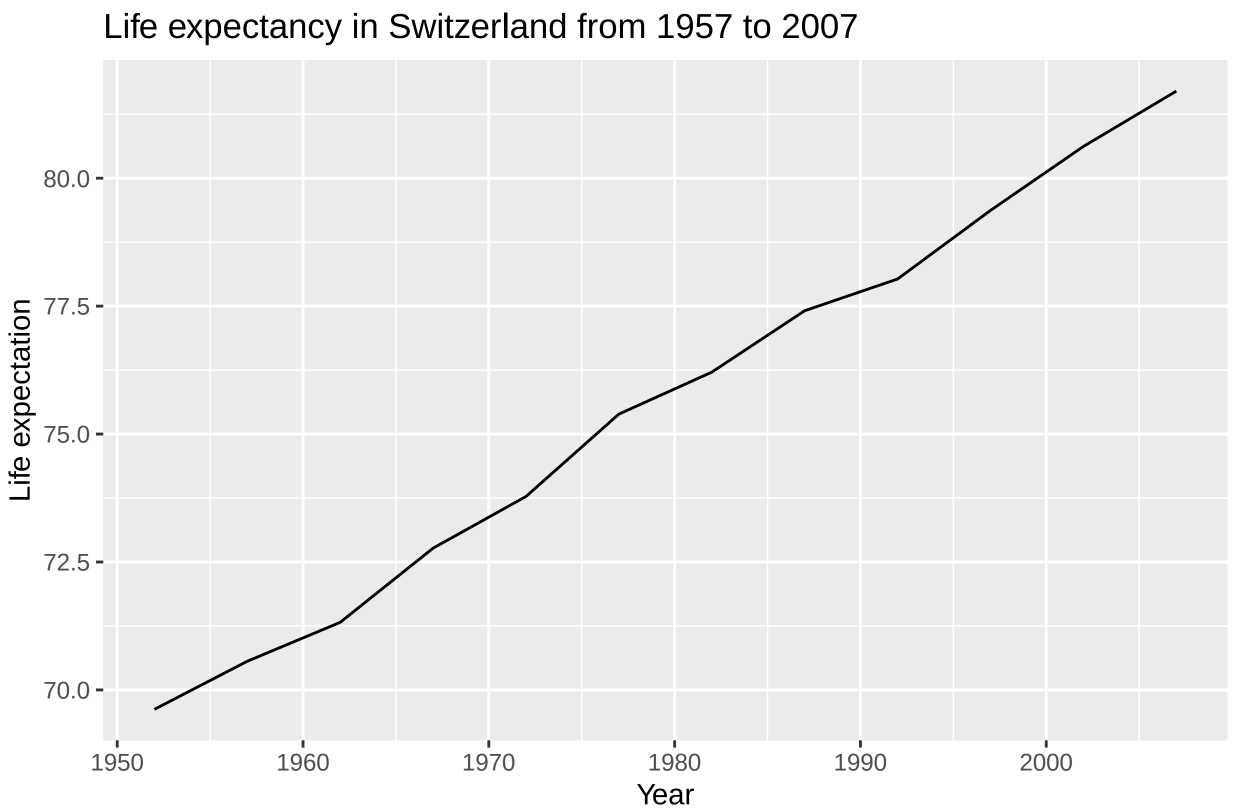
Let's plot life expectancy in time, for Switzerland!

```r
gapminder %>%
  filter(country == "Switzerland") %>% # filter for Switzerland
  ggplot(aes(x = year, y = lifeExp)) + # adds a first ggplot2 layer
  geom_line() # add a line
```



How can we improve this?

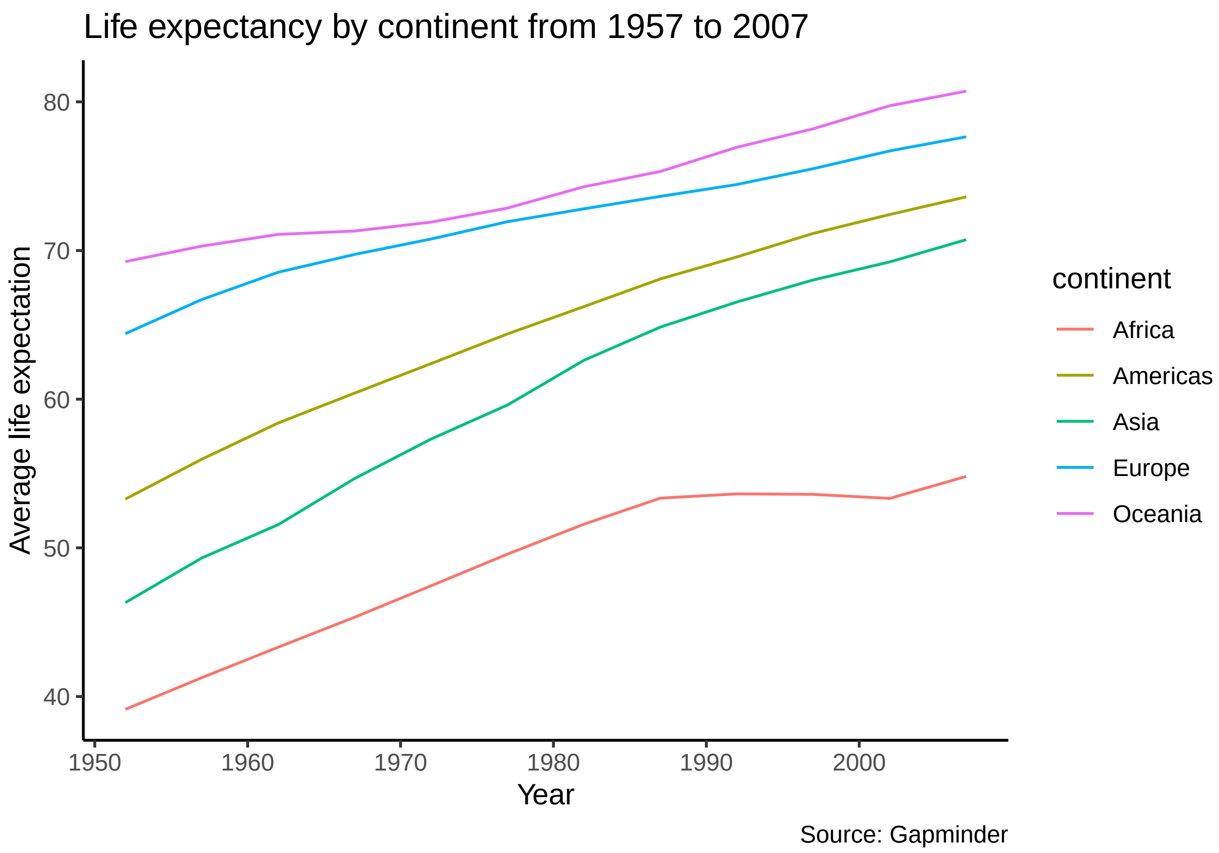
```
gapminder %>%
  filter(country == "Switzerland") %>% # filter for Switzerland
  ggplot(aes(x = year, y = lifeExp)) + # adds a first ggplot2 layer
  geom_line() + # add a line
  labs(title = "Life expectancy in Switzerland from 1957 to 2007", # add title
       x = "Year",  # add label for x axis
       y = "Life expectation", # add label for y axis
       caption = "Source: Gapminder") # add caption
```



Source: Gapminder

Do you think that life expectancy increased for across all continents in time?

```
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)) + # map
  geom_line(aes(color = continent)) + #here we are mapping color by continent at the geom_level
  labs(title = "Life expectancy by continent from 1957 to 2007", # add title
       x = "Year", # add
       y = "Average life expectation", # add lable for y axis
       caption = "Source: Gapminder") + # add caption
  theme_classic() # add theme
```



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

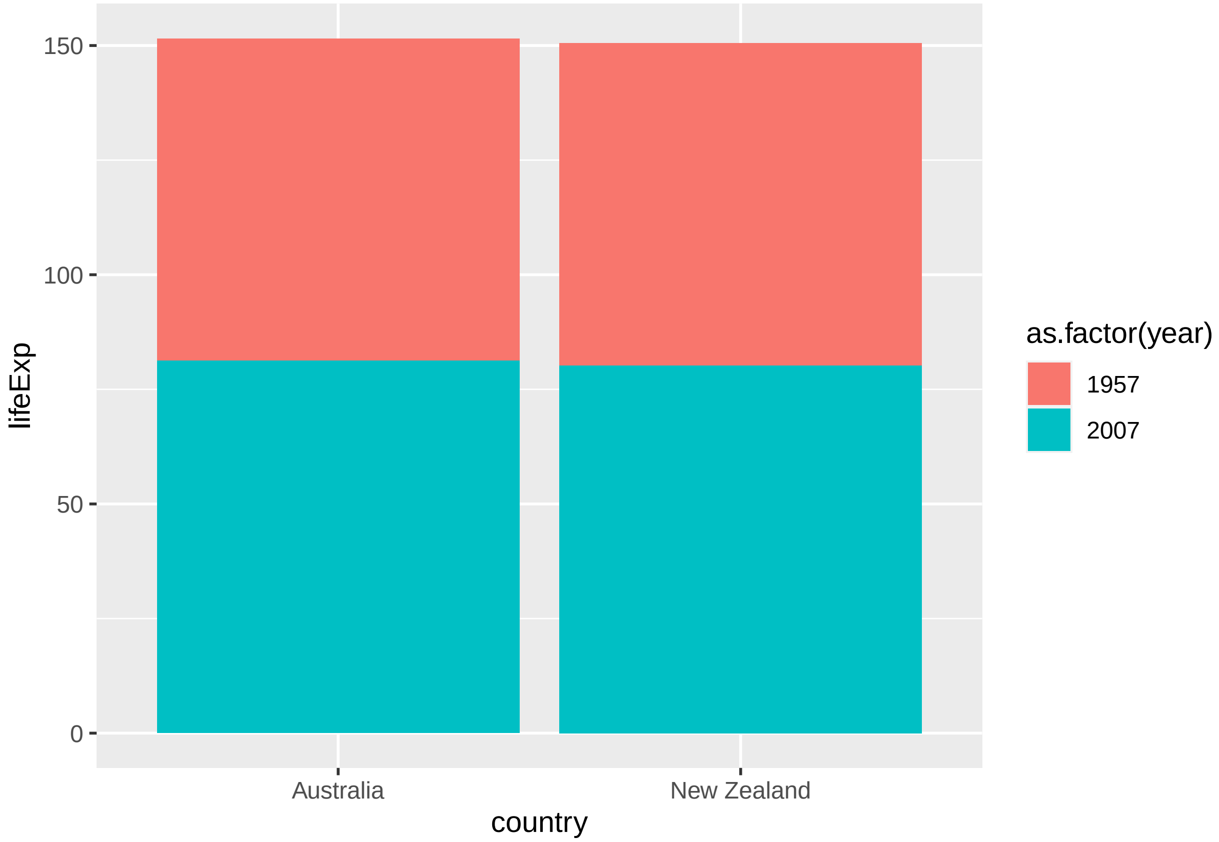
Bar plots: Life expectancy from 1957 to 2007 across continents

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

What are bar plots good for?

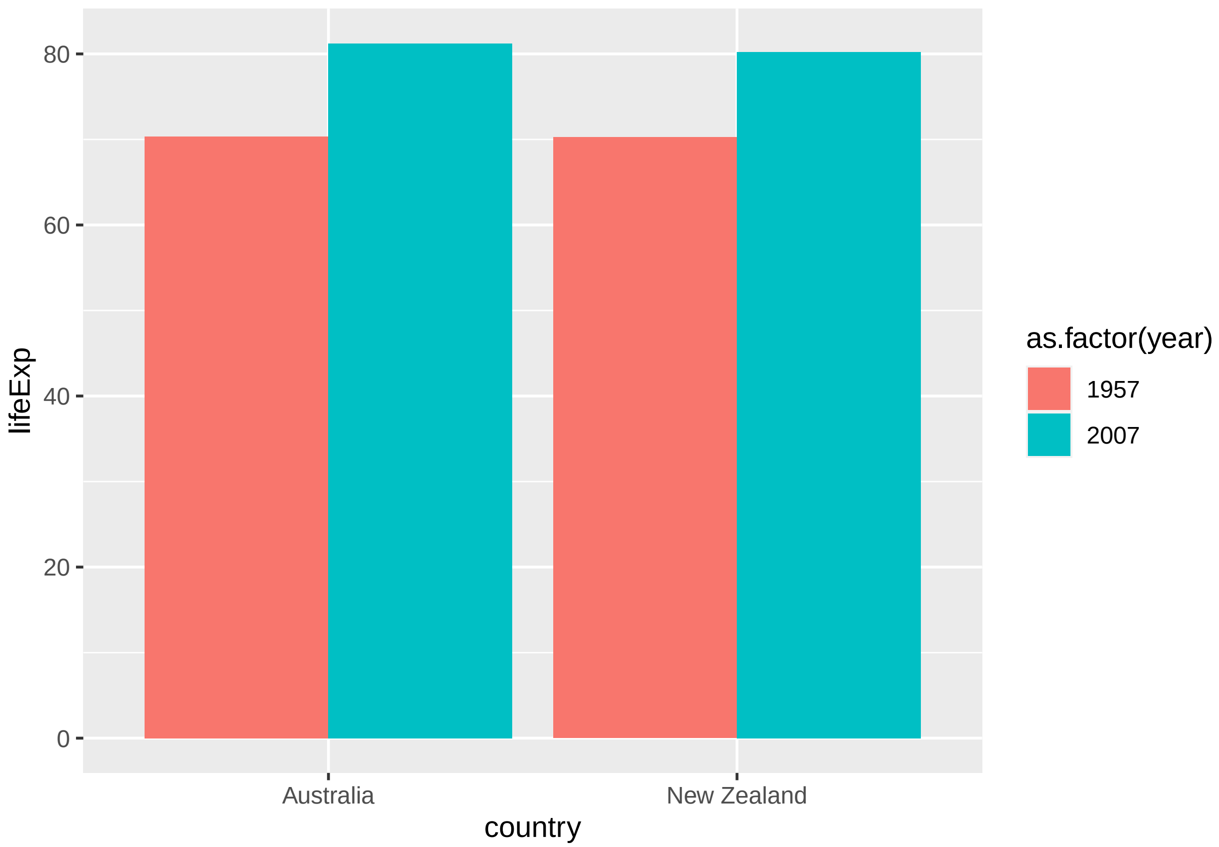
Let's start with a simple bar plot, one continent at two different points in time.

```
gapminder %>% # select data
  filter(year == 1957 | year == 2007,
         continent == "Oceania") %>% # filter for years of interest and Oceania
  ggplot(aes(x = country, # country in the x axis
             y = lifeExp, # map average life expectancy in the y axis
             fill = as.factor(year))) +
  # map fill by year (as factor), fill means the color filling the geom...
  geom_col()
```



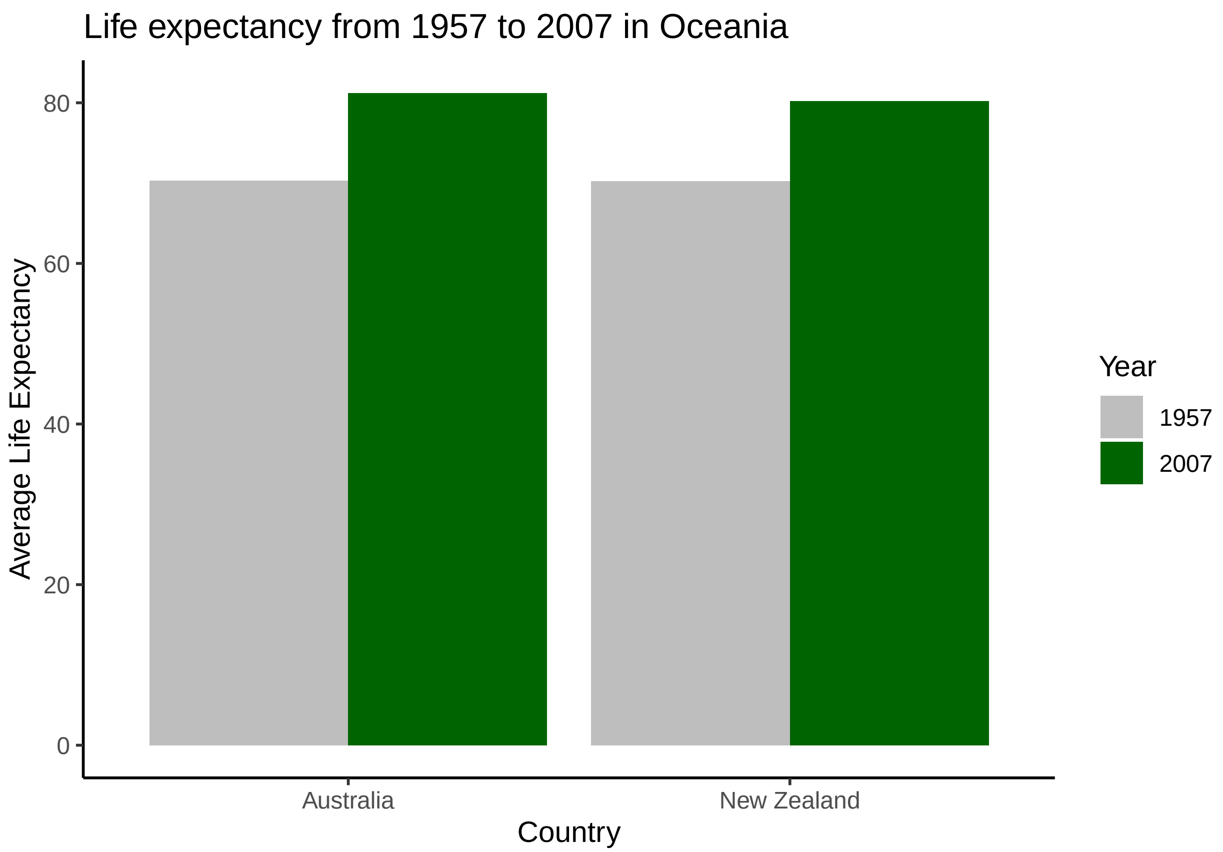
What is the issue with this plot?

```
gapminder %>% # select data
  filter(year == 1957 | year == 2007,
         continent == "Oceania") %>% # filter for years of interest and Oceania
  ggplot(aes(x = country, # country in the x axis
             y = lifeExp, # map average life expectancy in the y axis
             fill = as.factor(year))) +
  # map fill by year (as factor), remember that fill means the color filling the geom
  geom_col(position= "dodge") # separated the columns
```



This looks nicer, but there are several improvements we can still make!

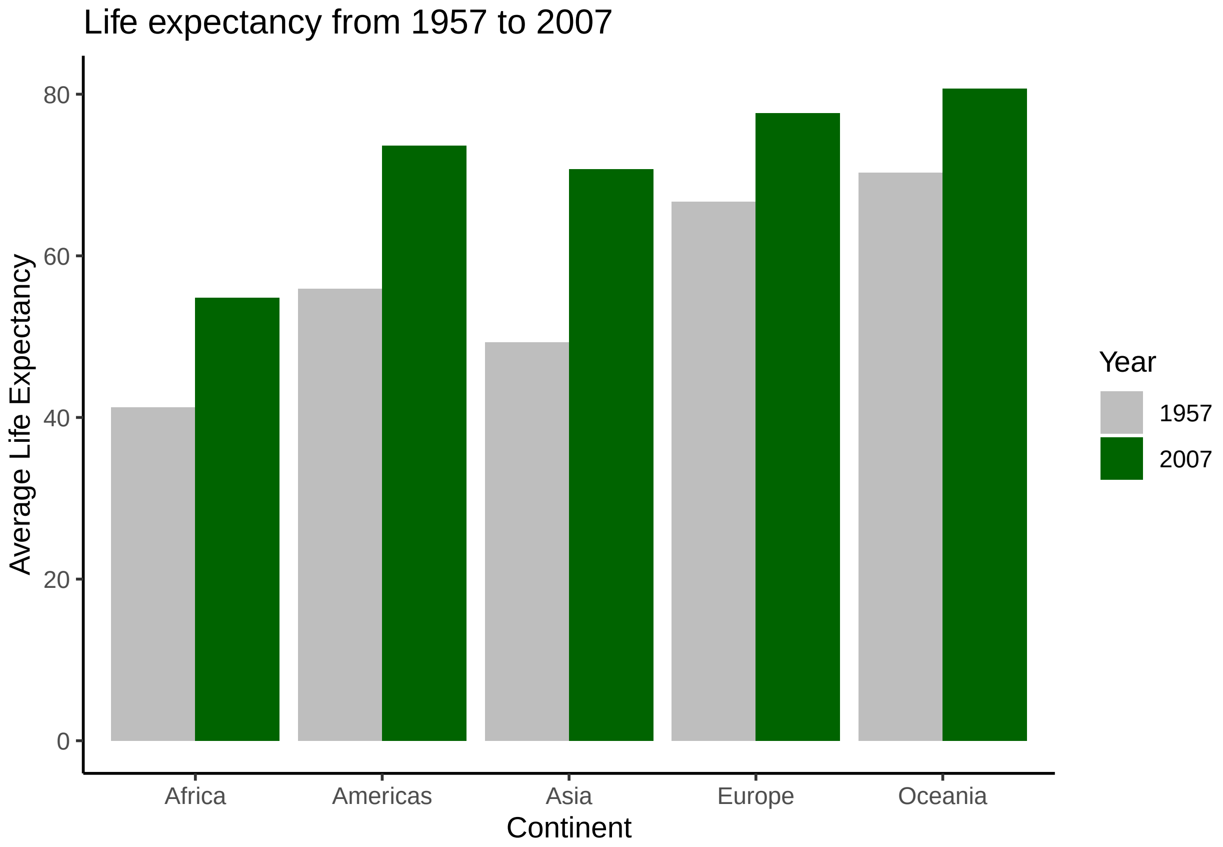
```
gapminder %>% # select data
  filter(year == 1957 | year == 2007,
         continent == "Oceania") %>% # filter for years of interest and Oceania
  ggplot(aes(x = country, # country in the x axis
             y = lifeExp, # map average life expectancy in the y axis
             fill = as.factor(year))) +
  # map fill by year (as factor), remember that fill means the color filling the geom
  geom_col(position= "dodge") + # separated the columns
  labs(title = "Life expectancy from 1957 to 2007 in Oceania", # add a title
       x = "Country", # 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()
```



```
  # adds a theme; theme_classic is a "clean" theme that removes unnecessary stuff
```

Lastly, let's use the same bar plots 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_col(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
```

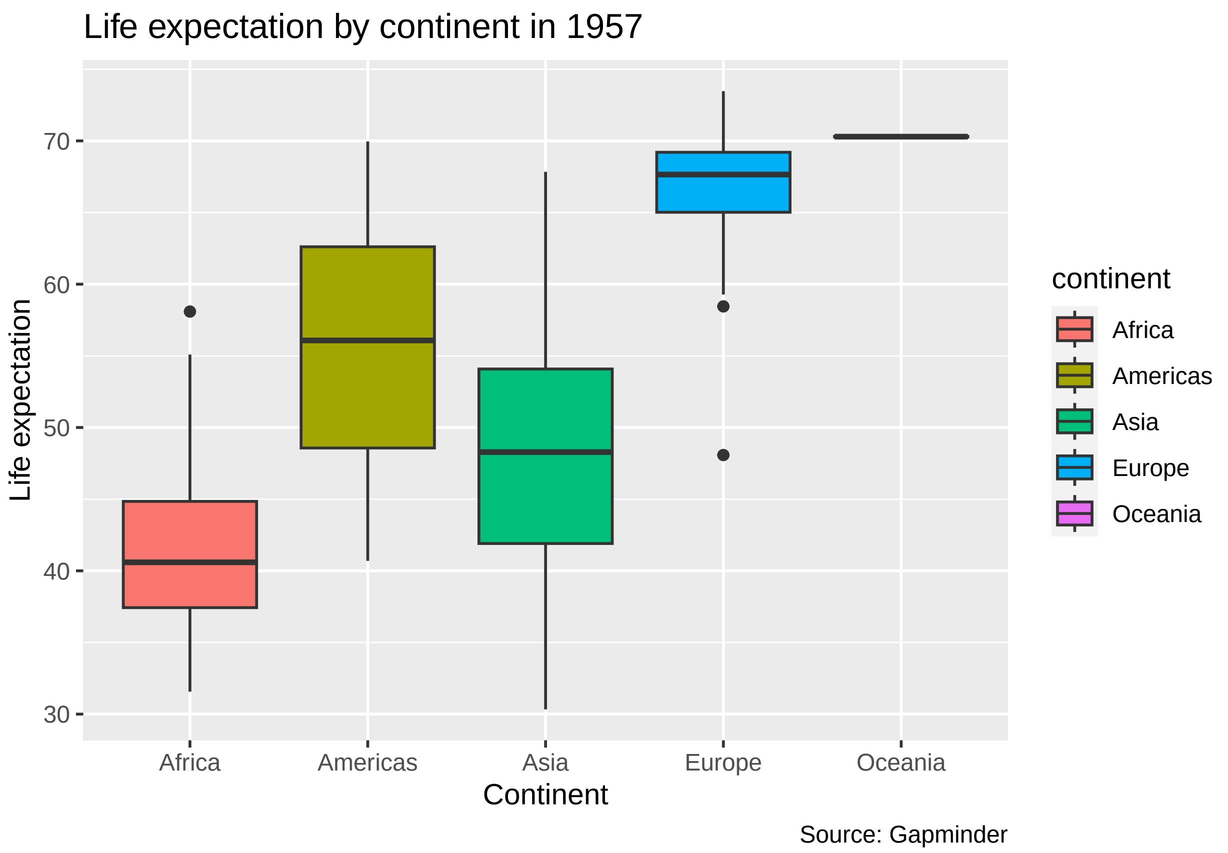


Box plots: Distribution of life expectancies across continent

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

What are box plots good for?

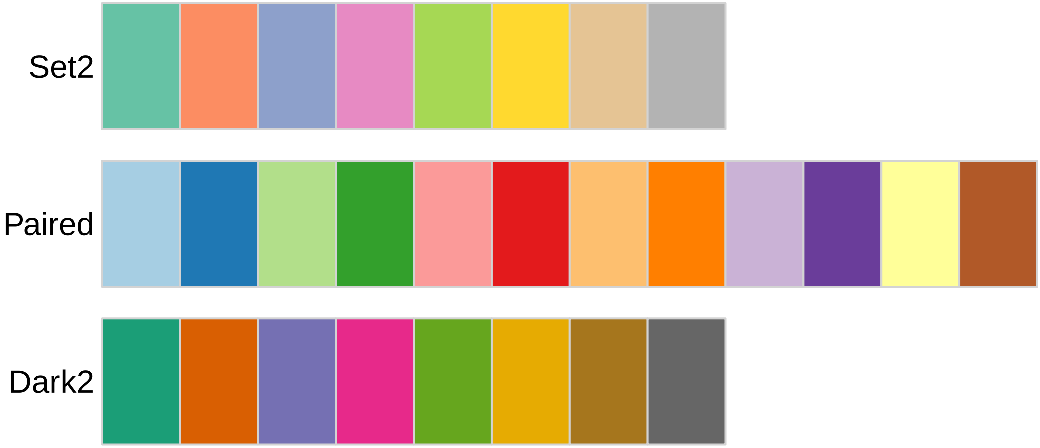
```
gapminder %>% # get data
  filter(year == 1957) %>% # filter by year
  ggplot(aes(x = continent, y = lifeExp, fill = continent)) + # map first layer
  geom_boxplot() + # box plot
  labs(title = "Life expectation by continent in 1957", # adds title
       x = "Continent", # adds x axis label
       y = "Life expectation", # adds y axis label
       caption = "Source: Gapminder") # adds caption
```


Life expectation by continent in 1957

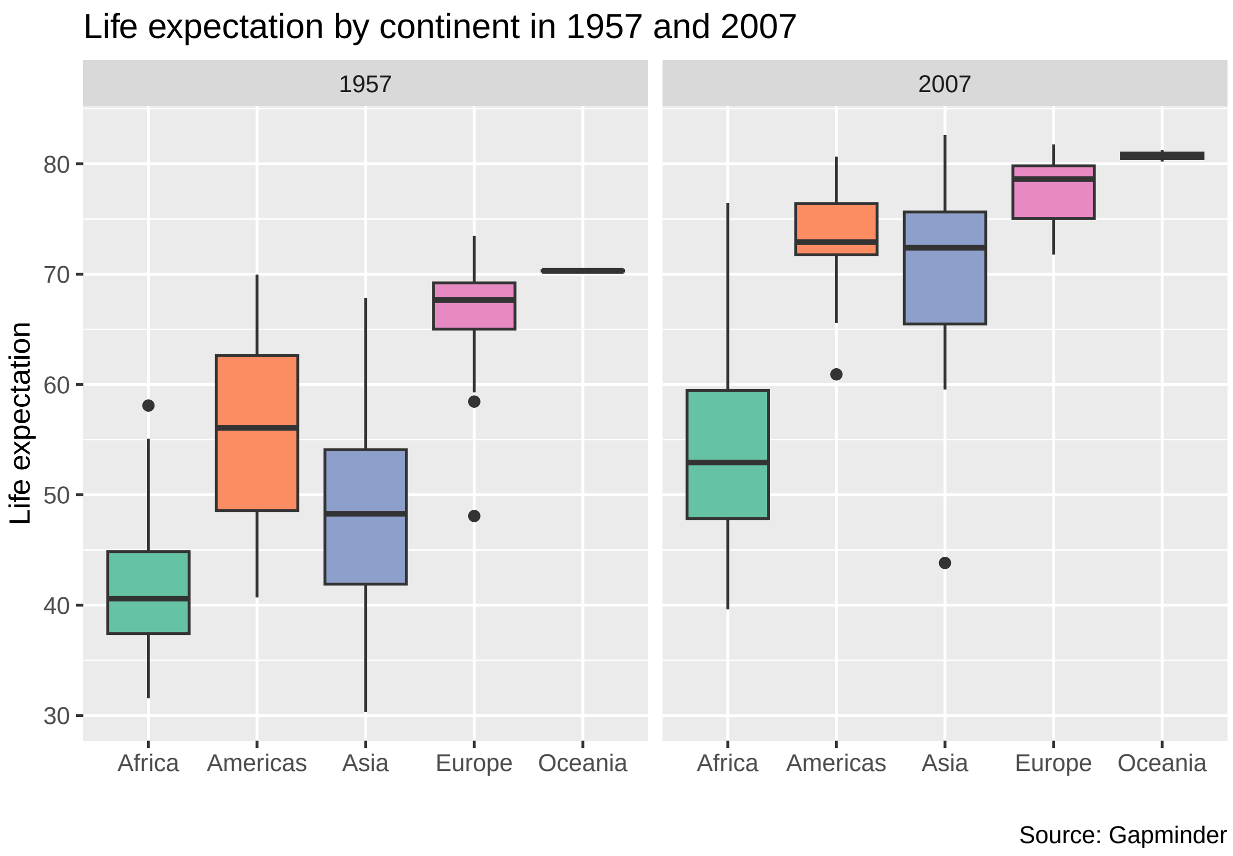
Source: Gapminder

There is a lot of redundant information in this plot, no?

```
library(RColorBrewer)
display.brewer.all(colorblindFriendly = TRUE, type = "qual")
```



```r
# "qual" stands for qualitative!
gapminder %>%
  filter(year == 1957 | year == 2007) %>% # filter
  ggplot(aes(x = continent, y = lifeExp, fill = continent)) + # fill by continent
  geom_boxplot() + # box plot
  facet_wrap(~year) + # separate plots by year; 2 different plots in this case
  labs(title = "Life expectation by continent in 1957 and 2007",
       x = "", # remove labels for x axis
       y = "Life expectation",
       caption = "Source: Gapminder") +
  theme(legend.position="none") +
  scale_fill_brewer(palette = "Set2")
```



```
#scale_fill_brewer() colours the aesthetics fill with the "Set2"
```

Could you make the same box plot for GDP?

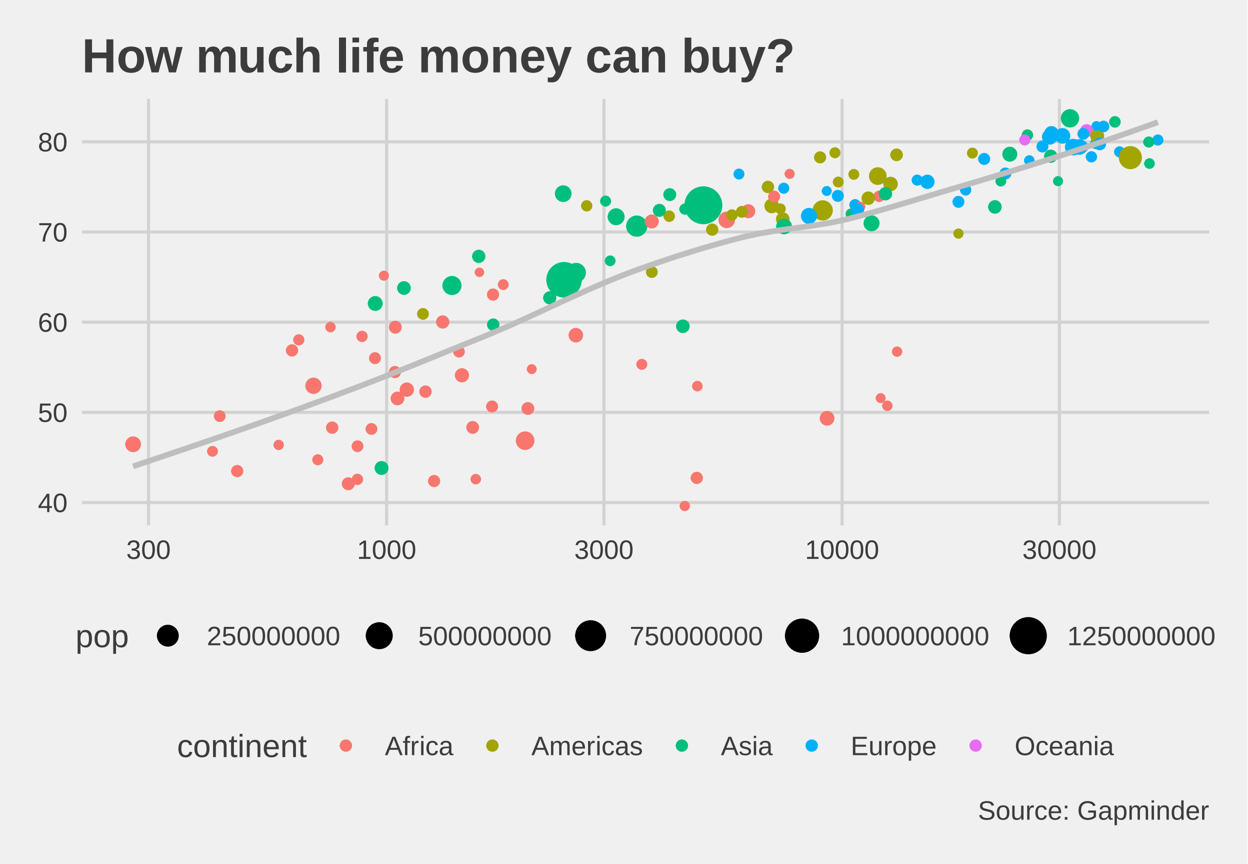
Scatter plots: Population, life expectancy and GDP

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

What are scatter plots good for?

```
library(ggthemes) # This package has several aditional cool themes!
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point(aes(size = pop, color = continent)) +
  # size points by population and color by continent
  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
  scale_x_log10() +
  theme_fivethirtyeight() # add theme from the ggthemes package
```

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


How much life money can buy?
80
70
60
50
40
300
1000
3000
10000
30000
pop
250000000
500000000
750000000
1000000000
1250000000
continent
Africa
Americas
Asia
Europe
Oceania
Source: Gapminder