|  |
| --- |
| itle: “HW5 - Make Data Move” |
| uthor: “” |
| ate: “05/10/2020” |
| utput: html\_document |

# Explore the recent global developments with R

Today, you will load a filtered gapminder dataset - with a subset of data on global development from 1952 - 2007 in increments of 5 years - to capture the period between the Second World War and the Global Financial Crisis.

**Your task: Explore the data and visualise it in both static and animated ways, providing answers and solutions to 7 questions/tasks below.**

## Get the necessary packages

First, start with installing the relevant packages ‘tidyverse’, ‘gganimate’, and ‘gapminder’.

## Look at the data

First, see which specific years are actually represented in the dataset and what variables are being recorded for each country. Note that when you run the cell below, Rmarkdown will give you two results - one for each line - that you can flip between.

# Inspect data  
unique(gapminder$year)

## [1] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2002 2007

head(gapminder)

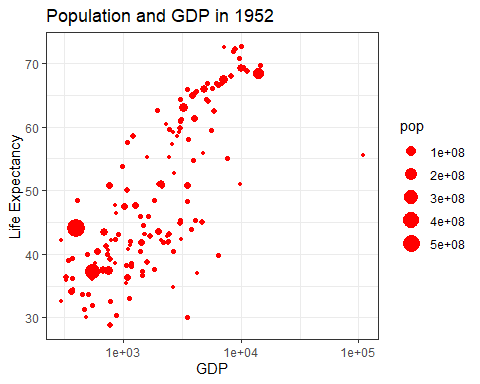
## # A tibble: 6 x 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.8 8425333 779.  
## 2 Afghanistan Asia 1957 30.3 9240934 821.  
## 3 Afghanistan Asia 1962 32.0 10267083 853.  
## 4 Afghanistan Asia 1967 34.0 11537966 836.  
## 5 Afghanistan Asia 1972 36.1 13079460 740.  
## 6 Afghanistan Asia 1977 38.4 14880372 786.

# I load the data in as a df to make it easier to inspect  
gapminder <- as.data.frame(gapminder)

The dataset contains information on each country in the sampled year, its continent, life expectancy, population, and GDP per capita.

Let’s plot all the countries in 1952.

# set theme to white background for better visibility  
theme\_set(theme\_bw())   
  
  
### make ggplot object  
y1952 <- ggplot(subset(gapminder, year == 1952), aes(gdpPercap, lifeExp, size = pop)) +  
 geom\_point(color = "red") +  
 labs(title = "Population and GDP in 1952", x = "GDP", y = "Life Expectancy") + scale\_x\_log10()  
  
# plot it  
y1952

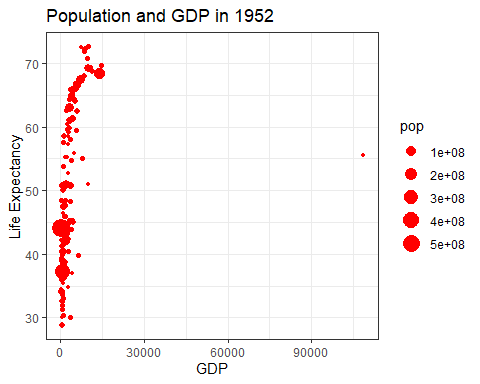


We see an interesting spread with an outlier to the right. Answer the following questions, please:

## Q1. Why does it make sense to have a log10 scale on x axis?

The difference in GDP per capitan from the richest country (the outliere) in 1952 to the rest of the countries is so huge, that the differences between GDP in the rest of the countries will get ‘drowned’ completely. This becomes apparent in the plot in the in the following chunk. Over time, the change in GDP per capitan will increase in a tremendous way. The log10 scale will facilitate the visibility this spread.

### make ggplot object without log transforming the x-axis  
y1952\_minuslog <- ggplot(subset(gapminder, year == 1952), aes(gdpPercap, lifeExp, size = pop)) +  
 geom\_point(color = "red") +  
 labs(title = "Population and GDP in 1952", x = "GDP", y = "Life Expectancy")  
  
# plot it  
y1952\_minuslog



## Q2. What country is the richest in 1952 (far right on x axis)?

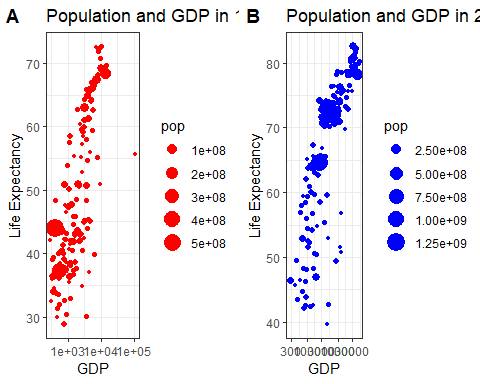
# I use pipes, filter and subset to isolate the richest country  
gapminder %>% filter(year == 1952) %>% subset(gdpPercap == max(gdpPercap))

## country continent year lifeExp pop gdpPercap  
## 72 Kuwait Asia 1952 55.565 160000 108382.4

The richest Country is Kuwait

You can generate a similar plot for 2007 and compare the differences

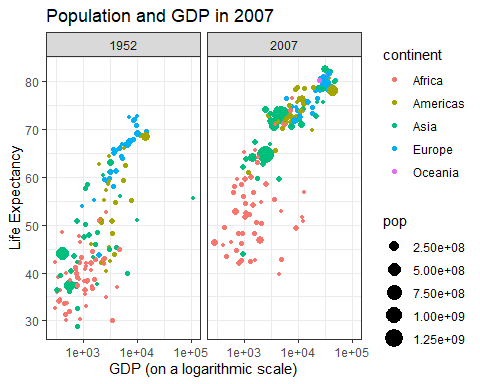
y2007 <- ggplot(subset(gapminder, year == 2007), aes(gdpPercap, lifeExp, size = pop, color = "blue")) +  
 geom\_point(color = "blue") +  
 scale\_x\_log10() +  
 labs(title = "Population and GDP in 2007", x = "GDP", y = "Life Expectancy")  
  
# plot them next to each other  
cowplot::plot\_grid(y1952, y2007, labels = "AUTO")



The bubbles are a bit hard to read, the comparison would be easier with a bit more visual differentiation.

## Q3. Can you differentiate the continents by color and fix the axis labels?

# I color by continent   
y1952\_2007 <- ggplot(subset(gapminder, year %in% c(1952, 2007)), aes(gdpPercap, lifeExp, size = pop, color = continent)) +  
 geom\_point() +  
 scale\_x\_log10() +  
 labs(title = "Population and GDP in 2007", x = "GDP (on a logarithmic scale)", y = "Life Expectancy") +  
 facet\_wrap(~ year)  
  
y1952\_2007



## Q4. What are the five richest countries in the world in 2007?

gapminder %>% subset(year == 2007) %>% arrange(desc(gdpPercap)) %>% head(5) %>% select(country, gdpPercap)

## country gdpPercap  
## 1 Norway 49357.19  
## 2 Kuwait 47306.99  
## 3 Singapore 47143.18  
## 4 United States 42951.65  
## 5 Ireland 40676.00

Norway, Kuwait, Singapore, US, and Ireland

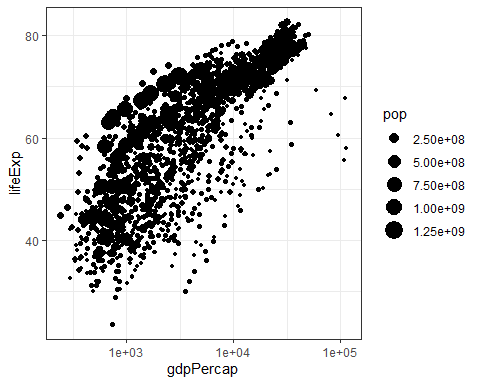
## Make it move!

The comparison would be easier if we had the two graphs together, animated. We have a lovely tool in R to do this: the gganimate package. And there are two ways of animating the gapminder ggplot.

### Option 1: Animate using transition\_states()

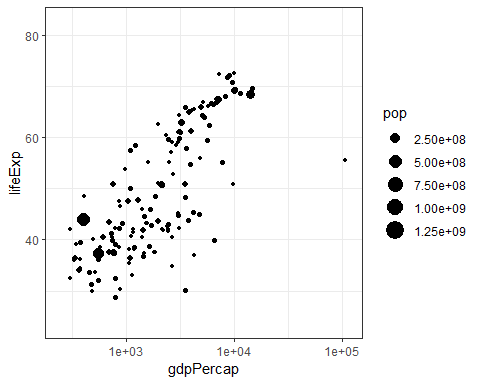
The first step is to create the object-to-be-animated

anim <- ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop)) +  
 geom\_point() +  
 scale\_x\_log10()  
  
anim # plot



This plot collates all the points across time. The next step is to split it into years and animate it. This may take some time, depending on the processing power of your computer (and other things you are asking it to do). Beware that the animation might appear in the ‘Viewer’ pane, not in this rmd preview. You need to knit the document to get the viz inside an html file.

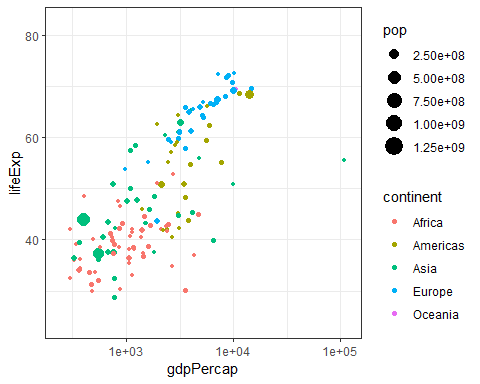
anim + transition\_states(year,   
 transition\_length = 1,  
 state\_length = 1)

 Notice how the animation moves jerkily, ‘jumping’ from one year to the next 12 times in total. This is a bit clunky, which is why it’s good we have another option.

### Option 2 Animate using transition\_time()

This option smoothes the transition between different ‘frames’, because it interpolates and adds transitional years where there are gaps in the timeseries data.

anim2 <- ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, color = continent)) +  
 geom\_point() +  
 scale\_x\_log10() +   
 transition\_time(year)  
  
anim2 # plot

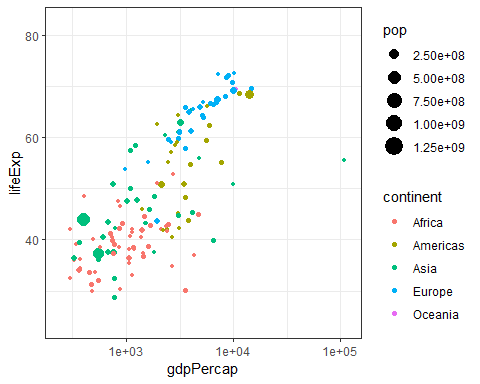


The much smoother movement in Option 2 will be much more noticeable if you add a title to the chart, that will page through the years corresponding to each frame.

## Q5 Can you add a title to one or both of the animations above that will change

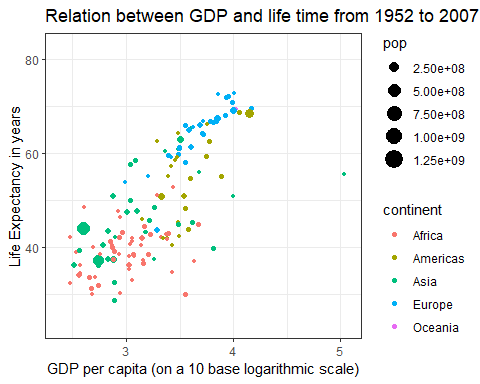
in sync with the animation? [hint: search labeling for transition\_states() and transition\_time() functions respectively]

# I add '{frame\_time} in labs.  
anim2 +  
 labs(title = 'Year: {frame\_time}', x = 'GDP per capita (logarithmic transformed)', y = 'Life Expectancy')  
  
anim2 # plot



## Q6 Can you make the axes’ labels and units more readable? Consider expanding the abreviated lables as well as the scientific notation in the legend and x axis to whole numbers.[hint:search disabling scientific notation]

gapminder$gdpLog <- log10(gapminder$gdpPercap)  
  
anim3 <- ggplot(gapminder, aes(gdpLog, lifeExp, size = pop, color = continent)) +  
 geom\_point() +  
 transition\_time(year) +   
 labs(title = 'Relation between GDP and life time from 1952 to 2007: {frame\_time}', x = 'GDP per capita (on a 10 base logarithmic scale)', y = 'Life Expectancy in years') +  
 scale\_x\_continuous(labels = comma\_format(big.mark = ".",  
 decimal.mark = ","))  
  
anim3 # plot



## Q7 Come up with a question you want to answer using the gapminder data and write it down. Then, create a data visualisation that answers the question and explain how your visualization answers the question.

(Example: you wish to see what was mean life expectancy across the continents in the year you were born versus your parents’ birth years). [hint: if you wish to have more data than is in the filtered gapminder, you can load either the gapminder\_unfiltered dataset and download more at <https://www.gapminder.org/data/> ]

I wish to compare life expectancyies over time across continents in the world. That will enable me to see the development by continent but also compare the development across continent. I use a bar plot to clearly indicate the life expectancies and use facet\_wrap to create a plot for each continent. This reduces confusion compared to plotting everyhting on top of each other.

dataSub <- subset(gapminder) %>% group\_by(continent, year) %>% dplyr::summarize(  
 lifeExpect = mean(lifeExp)  
 )  
  
PlotQ7 <- ggplot(dataSub, aes(year, lifeExpect, fill = continent)) +  
 geom\_bar(stat = "identity", width = 2) + facet\_wrap(~ continent) +   
 labs(title = "Life expectancies by continent from 1952 to 2007", x = "Year", y = "Life Expectancies in Years")  
  
PlotQ7

