HW5 - Make Data Move

05/10/2020

# 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’.

## Warning: replacing previous import 'vctrs::data\_frame' by 'tibble::data\_frame'  
## when loading 'dplyr'

## ── Attaching packages ───────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.0  
## ✓ tidyr 1.1.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.2

## Warning: package 'tibble' was built under R version 4.0.2

## ── Conflicts ──────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Warning: package 'gganimate' was built under R version 4.0.2

## Warning: package 'gapminder' was built under R version 4.0.2

## Warning: package 'gifski' was built under R version 4.0.2

## Warning: package 'av' was built under R version 4.0.2

## 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.

unique(gapminder$year) #checking

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

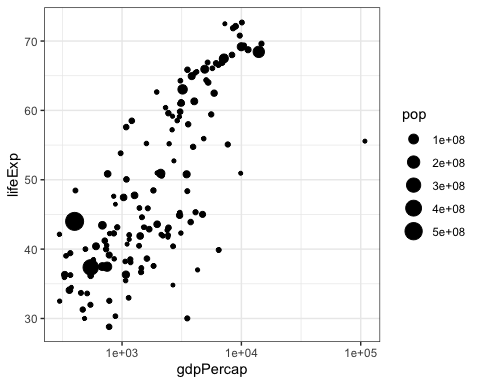
head(gapminder)

## # 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.

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.

theme\_set(theme\_bw()) # set theme to white background for better visibility  
  
ggplot(subset(gapminder, year == 1952), aes(gdpPercap, lifeExp, size = pop)) +  
 geom\_point() +  
 scale\_x\_log10()

 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?

Applying a log\_10 scale to the x-axis results in that for every 1 value of y, the x-axis increases by a factor of 10. The transofrmation is useful if there are large differences between x-values as it compresses the scale. If we do not use a log10 scale for the x-axis, all other datapoints (besides the one outlier) get squished together and are very hard to differentiate. There is a huge difference between 100.000, 10.000 and 1.000 on a normal scale, but on a log\_10 scale, this difference (purely visually) becomes smaller, and we can more easily differentiate the differences between the lower-GDP countries, while on a non-log\_10 scale, it would be very difficult to see the difference. Thus, putting the data on a log\_10 scale makes it less cluttered.

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

We can target the richest country in 1952 by filtering by year set to 1952, and arrange the data, here in ascending order with the largest values appearing last in the dataset, followingly using tail() to target the last value, i.e. the country with the largest GDP.

#Q2  
  
df <- gapminder  
#filtering by year == 1952, and arrangering in ascending (smallest to largest) value in the GDP column  
df\_ordered <- filter(df, df$year == 1952) %>%  
 arrange(gdpPercap)  
  
#Using tail to take the last row in the ordered (ascending in GDP value) dataframe, to get the richest country in 1952 - very last row is the richest country  
tail(df\_ordered, 1)

## # A tibble: 1 x 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Kuwait Asia 1952 55.6 160000 108382.

The richest country in 1952 is Kuwait

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

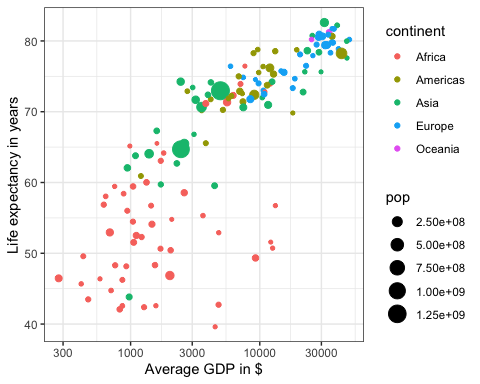
The black bubbles are a bit hard to read, the comparison would be easier with a bit more visual differentiation.

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

Differentiate continents by adding “color=continent” to aes() function of ggplot. Thereby we tell R that you want to add colour to the aesthetics and these should depend on the continent. R searches the dataset and assigns a colour to each unique continent and all values that also share that continent value.

Fixing x and y labels by adding: ylab(“Life expectancy in years”) xlab(“Average GDP in $”) #assuming GDP is in dollars

ggplot(subset(gapminder, year == 2007), aes(gdpPercap, lifeExp, size = pop, color=continent)) +  
 geom\_point() +  
 scale\_x\_log10() +  
 ylab("Life expectancy in years")+ #Setting y axis label  
 xlab("Average GDP in $") #Setting x axis label - assuming GDP is in dollars



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

We’ll apply a filter by year (2007) and then a arrange the data in an ascending manner, and taking the bottom 5 values (the 5 largest values).

The 5 richest countries are: 1. Norway 2. Kuwait 3. Singapore 4. United States 5. Ireland

#Q4  
  
#filtering by year == 2007, and arrangering in ascending (smallest to largest) value in the GDP column  
df\_ordered <- filter(df, df$year == 2007) %>%  
 arrange(gdpPercap)  
  
#Using tail to take the last 5 rows in the ordered (ascending in GDP value) dataframe, to get the richest country in 1952 - very last row is the richest country  
tail(df\_ordered, 5)

## # A tibble: 5 x 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Ireland Europe 2007 78.9 4109086 40676.  
## 2 United States Americas 2007 78.2 301139947 42952.  
## 3 Singapore Asia 2007 80.0 4553009 47143.  
## 4 Kuwait Asia 2007 77.6 2505559 47307.  
## 5 Norway Europe 2007 80.2 4627926 49357.

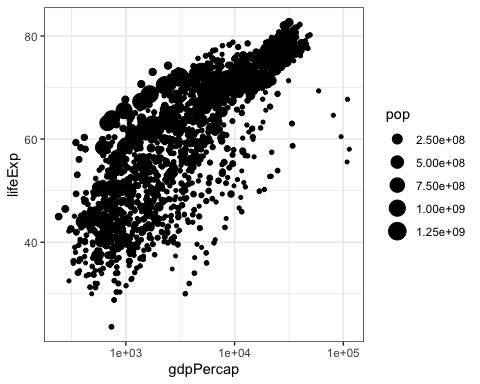
## Make it move!

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()

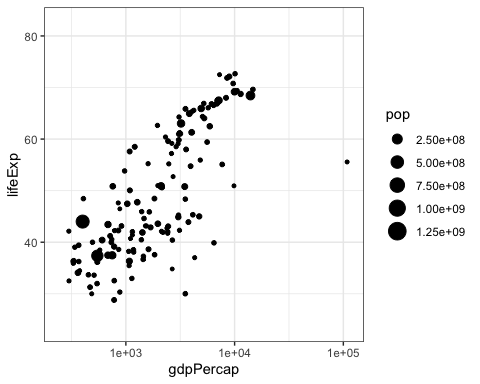
The first step is to create the object-to-be-animated

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



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.

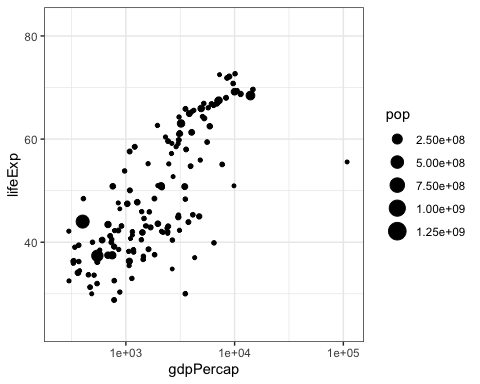
p <- anim + transition\_states(year,   
 transition\_length = 1,  
 state\_length = 1)  
  
p

 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.

animation2 <- p + transition\_time(year)  
  
animation2



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.

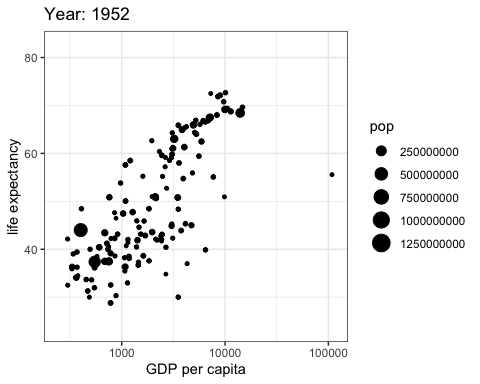
Q5 Can you add a title to one or both of the animations above that will change in sync with the animation? [hint: search labeling for transition\_states() and transition\_time() functions respectively]

This can be done by adding labs(title = ‘Year: {frame\_time}’) to the ggplot object

Q6 Can you made 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]

Removing scientific notation can be done by options(scipen=10000) - setting scipen to a high number removes scientific notation To give the axes more comprehensible names, this can be manually set using the labs function

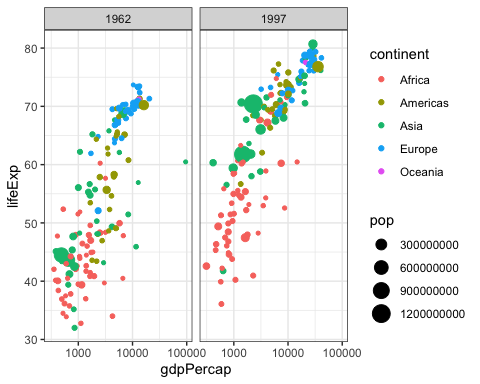
options(scipen=10000) #Setting scipen to a high number removes scientific notation  
  
animation3 <- animation2 + labs(title = 'Year: {frame\_time}', #adding the year variable as title  
 x = 'GDP per capita', #adding explanatory title for x axis  
 y = 'life expectancy') #adding explanatory title for y axis  
  
animation3



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/> ]

Question: how have life expetancy changed fom 1962 to 1997 in different continents?

df <- gapminder # transforming the gapmidner data into a dataframe  
  
df\_1997\_1962 <- filter(df, year == 1962 | year == 1997) #filtering the data to only include years 1997 (my birth year) and 1962 (one year earlier than my parents' birth year)  
  
p97\_62 <- ggplot(df\_1997\_1962, aes(gdpPercap, lifeExp, size = pop, color=continent)) +  
 geom\_point() +  
 scale\_x\_log10()+  
 facet\_wrap(~year)  
p97\_62



We can see from the plots that overall, the life expectancy had increased across different continents (relative to their starting point in 1962) - However, Africa still has relatively lower life expextancy than the rest of the world.