Exercises: Preston curve—ggplot2

YSC2210 - DAVis with R

Michael T. Gastner

Preston curve

Introduction

In a classic paper, Preston (1975) discussed scatter plots of life expectancy versus national income per capita (see figure 1), where each point represents one country. The term 'Preston curve' has since then become a synonym for curves fitted to similar data, usually with the per-capita gross domestic product (GDP), instead of national income, as x-value. Preston (1975) and many others have used untransformed x-values and y-values. For a different take on plotting the data, the Swedish foundation Gapminder (2016) uses a logarithmic scale for income (figure 2). A logarithmic scale makes sense because most economic indicators are right-skewed. I recommend to adopt Gapminder's approach.

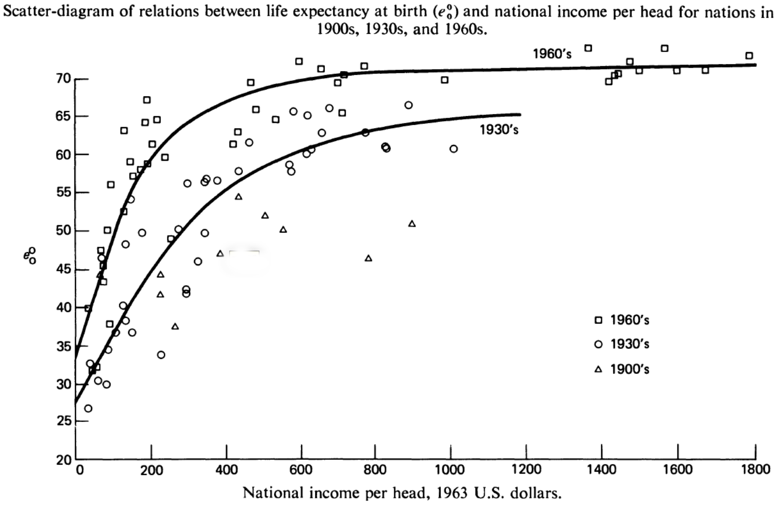
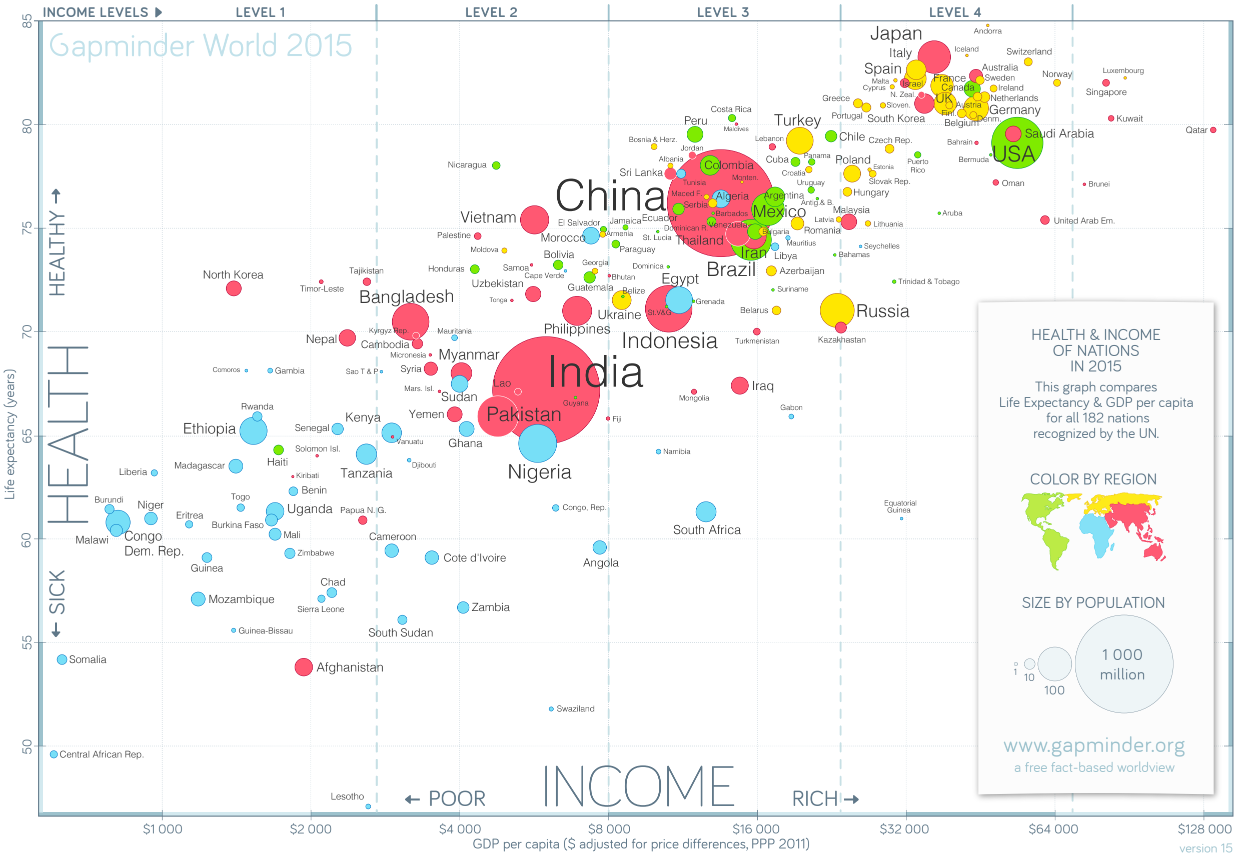


Figure 1: Diagram from Preston (1975).

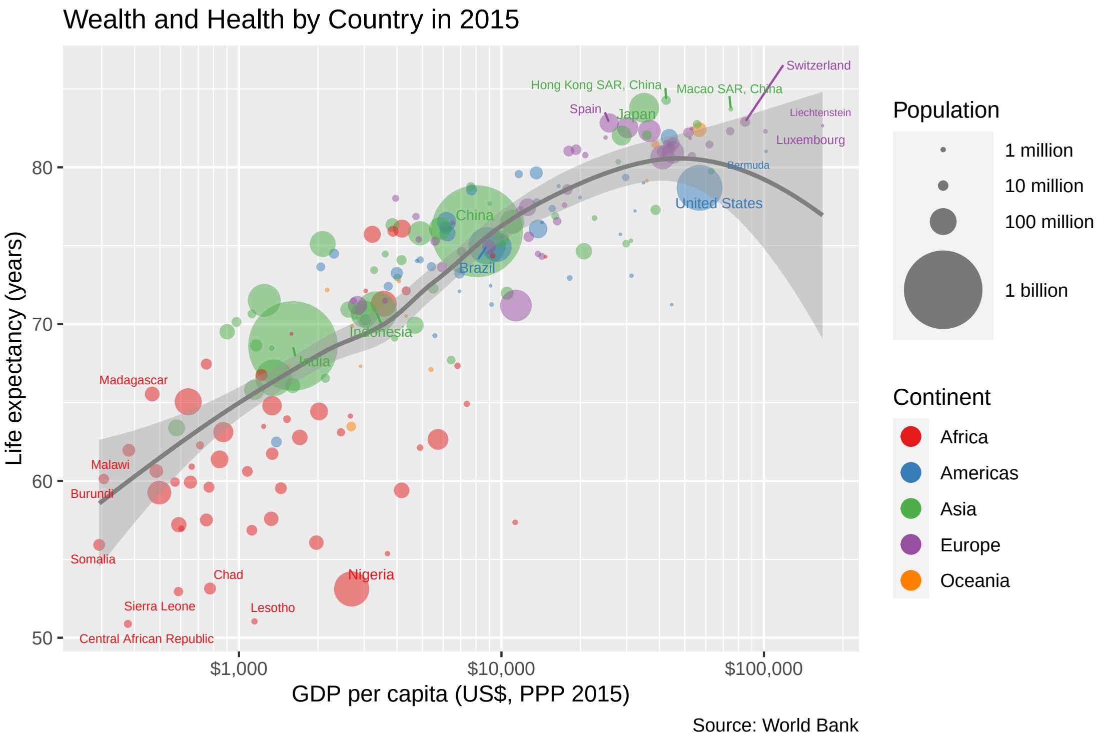


DATA SOURCES—INCOME: World Bank's GDP per capita, PPP (2011 international $). Income of Syria & Cuba are Gapminder estimates. X-axis uses log-scale to make a doubling income show same distance on all levels. POPULATION: Data from UN Population Division. LIFE EXPECTANCY: IHME GBD-2015, as of Oct 2016.
ANIMATING GRAPH: Go to www.gapminder.org/tools to see how this graph changed historically and compare 500 other indicators. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy, share, modify, integrate and even sell them, as long as you mention: "Based on a free chart from www.gapminder.org".

Figure 2: Diagram from Gapminder (2016).

Objectives

We practise our **ggplot2** skills by making a plot that is comparable to the plot of life expectancy as a function of GDP by the Gapminder Foundation (figure 3).


Wealth and Health by Country in 2015

Source: World Bank

Figure 3: Data from Gapminder (2016) plotted with **ggplot2**.

Tasks

(1) Download https://michaelgastner.com/DAVisR_data/life_quality.csv.[1] This CSV file contains the columns:

- `country_name`
- `country_code`: standardised 3-letter code (ISO 3166-1 alpha-3)
- `gdp_per_capita` (US$, PPP 2015)
- `life_expectancy` (in years)
- `pop`: population
- `continent`

These numbers are not exactly the same as those used by Gapminder (2016). Thus, please do not worry if your final plot does not look identical.

Import the CSV as a tibble.

(2) Make a bubble chart with **ggplot2** where:

- the x-coordinate is the GDP per capita.

[1] These data are based on information available from the World Bank (accessed on 14 February 2022).
- GDP per capita (US$, PPP 2015): https://data.worldbank.org/indicator/NY.GDP.PCAP.KD
- Life expectancy at birth, total (years): https://data.worldbank.org/indicator/SP.DYN.LE00.IN
- Population: https://data.worldbank.org/indicator/SP.POP.TOTL

- the y-coordinate is the life expectancy.
- the colour indicates the continent.
- the size of the bubble indicates the population.

Change the axis labels and give the plot a title. Give credit to the World Bank as data source in the form of a caption. Make the bubbles semitransparent. (An improvement compared to Gapminder!)

Do not worry about the scales for the coordinates and the bubble areas yet. We will fix them shortly.

(3) Change the x-coordinates to a logarithmic scale. Change the minor breaks as shown in figure 3, where they appear as thin white lines. See section 10.1.5 in Wickham *et al.* (2021) for related examples. Change the default tick mark labels ('1e+03', '1e+04', '1e+05') to reader-friendlier labels ('$1,000', '$10,000', '$100,000'). Why do you think **ggplot2** puts axis ticks, by default, at integer powers of 10 in contrast to Gapminder's tick positions (e.g. '$8,000', '$16,000', '$32,000')?

(4) Use `scale_size_area()` so that the areas of the bubbles in the legend represent populations of 1 million, 10 million, 100 million and 1 billion. Change the numbers in the legends from '1e+06', '1e+07', '1e+08', '1e+09' to the reader-friendlier strings '1 million', '10 million', '100 million' and '1 billion'. Increase `max_size` so that the bubble areas are approximately the same as in the Gapminder figure.

(5) Change the colour scale to the ColorBrewer palette 'Set1'. See section 11.3.1 in Wickham *et al.* (2021) for related examples. In my opinion, Set1 provides clearer contrasts than **ggplot2**'s default colours. These are not the same colours as in the Gapminder figure, but let us not worry about it.

(6) Semi-transparent colours are great at dealing with overplotting in the bubble plot. However, they are not optimal for the legend, where we would like to see clear contrasts between the colours. Override the alpha aesthetic in the legend. Also increase the sizes of the circles in the colour legend so that the colours are easier to read. See section 11.3.6 in Wickham *et al.* (2021) for related examples.

(7) Fit a single LOESS curve to all data points. Use the countries' population sizes as weighting variable. Use a neutral colour for the curve to indicate that the curve is not specific to any continent. Do not show the geom for the LOESS curve in the legend.

(8) Using `ggrepel::geom_text_repel()`, add the country names as labels to:

- the 5 countries with the highest GDP.
- the 5 countries with the lowest GDP.
- the 5 countries with the longest life expectancy.
- the 5 countries with the shortest life expectancy.
- the 5 most populous countries.

(Some countries are in more than one of these categories.)

Make the text colour equal to the continent's colour. Choose the font sizes and text positions so that the labels are easily legible.

(9) Feel free to make more adjustments if you think they improve the quality of the plot. Then adjust the figure dimensions in the knitted file with the code chunk options `fig.width`, `fig.height` and `out.width`. Labels should be clearly legible without appearing disproportionately large. Figure 3 shows my attempt.

(10) Write a few sentences about the data. What does the plot reveal about the data? If you refer to specific countries, make sure to add the corresponding labels in the plot if necessary.

References

Gapminder (2016). Updated Gapminder World Poster 2015! URL: https://www.gapminder.org/downloads/updated-gapminder-world-poster-2015/. Accessed on 2020-11-26.

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies*, **29**(2), 231–248.

Wickham, H., Navarro, D., and Pedersen, T. L. (2021). *ggplot2: elegant graphics for data analysis*. Springer, 3rd edition.