**Setting up the stage**

Everybody seems to be starring at plots outlining the spread of the pandemic these days. One thing that caught my interest is how relative small differences in design choices can influence the message that a visual displays.

I use my tidycovid19 package to pull and plot the data.

# remotes::install\_github("joachim-gassen/tidycovid19")

suppressPackageStartupMessages({

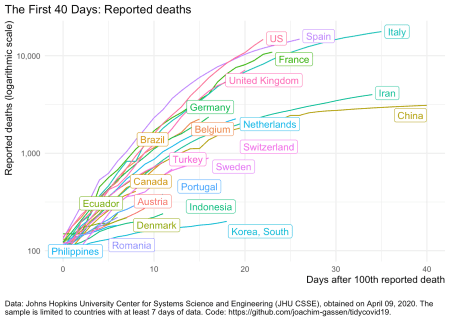
library(tidyverse)

library(tidycovid19)

})

covid19\_dta <- download\_merged\_data(silent = TRUE, cached = TRUE)

plot\_covid19\_spread(covid19\_dta)



The plot\_covid19\_spread() function allows customizing the plot along multiple dimensions. By default it plots, using data from CSSE team of Johns Hopkins University, the cumulative development of reported deaths over country by ‘event time’ starting with the day of the 100th death, using a log scale for the y-axis.

Factors that you can customize are

* the data that you want to plot (deaths, confirmed cases, recovered cases or ‘active cases’ meaning the difference of confirmed and recovered cases),
* the starting point of the X axis
* how many days you want to plot,
* how many days of data the countries need to have to be included,
* whether you want to use a log or linear scale for the Y axis,
* whether you want the data to be scaled by population or not,
* whether you want to look at cumulative cases or at daily changes of cases,
* if you want to look at daily changes, the period over which to average these changes,
* governmental interventions that you want to highlight in the graph and
* countries that you want to highlight

Assuming three different sensible variants for the numeric choices and leaving out the highlighting option, these choice generate a whooping 2,592 degrees of freedom. Many graphs to choose the one that communicates the data most fairly. In the following, I play the devils advocate. I try to produce one graph where the situation looks very bad and one graph where things look as bright as possible. Doing so, I focus on large European countries and the U.S.

**The negative view**

We will focus on deaths mostly as these are less likely to be influenced by variation in testing protocols across countries. To impose a negative view we use a linear scale as this clearly communicates the exponential growth component. A first attempt:

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 100, min\_by\_ctry\_obs = 7,

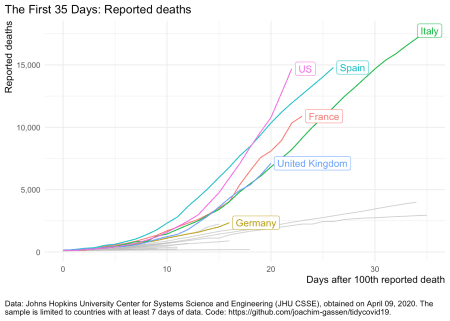
edate\_cutoff = 35, per\_capita = FALSE, log\_scale = FALSE,

cumulative = TRUE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



Now this looks as if all countries are roughly on the same trajectory. Let’s assume that we want to single out the U.S.

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 500, min\_by\_ctry\_obs = 7,

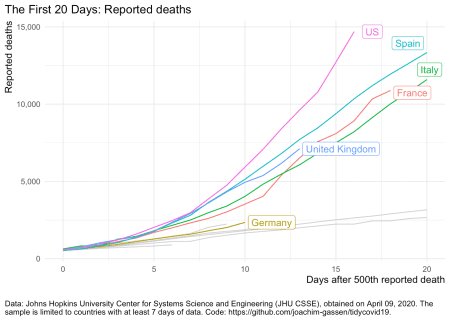
edate\_cutoff = 20, per\_capita = FALSE, log\_scale = FALSE,

cumulative = TRUE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



By moving the X axis starting point we focus the viewers attention on the the period where deaths were growing faster in the U.S. compared to other regions of the world. Can we also make Spain look bad?

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 1, min\_by\_ctry\_obs = 7,

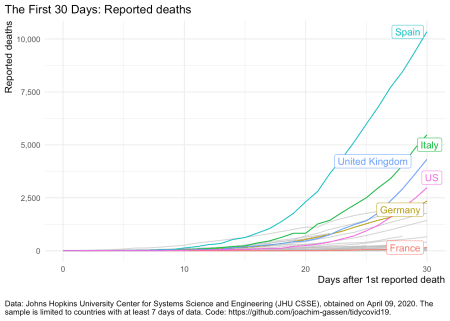
edate\_cutoff = 30, per\_capita = FALSE, log\_scale = FALSE,

cumulative = TRUE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



Now this looks horrible for Spain. BTW: I used a cheap trick in the two last graphs. Guess what: I cut off the X axis. See for yourself:

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 1, min\_by\_ctry\_obs = 7,

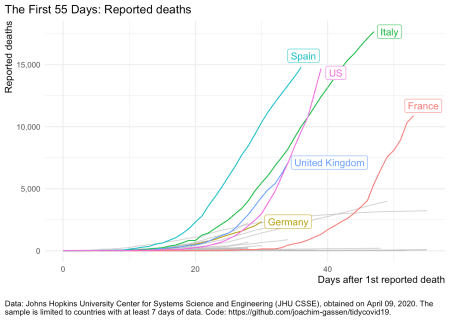
edate\_cutoff = 55, per\_capita = FALSE, log\_scale = FALSE,

cumulative = TRUE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



You see, by shifting around the X axis you can more or less focus on certain regions of the curves where a given country has had, relatively speaking, the fastest growth rate.

The next graph is extremely unfair as I am also changing the data that I present. I now focus on the daily change of active cases, meaning the difference of confirmed and recovered cases.

plot\_covid19\_spread(

covid19\_dta, type = "active", min\_cases = 1000, min\_by\_ctry\_obs = 20,

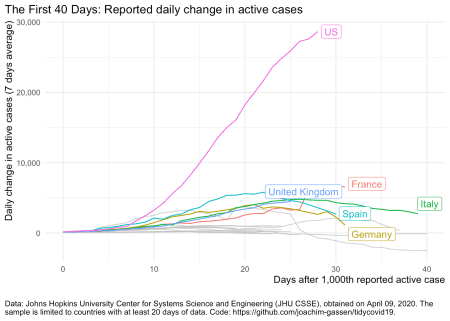
edate\_cutoff = 40, per\_capita = FALSE, log\_scale = FALSE,

cumulative = FALSE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



It communicates the impression that the situation in the U.S. is dwarfing the situation in Europe which clearly is not the case. Time to become less pessimistic and switch the focus on the more positive side of things.

**The positive view**

To make things look better, let’s first switch to the log scale for Y axis. We return to focus on deaths.

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 100, min\_by\_ctry\_obs = 7,

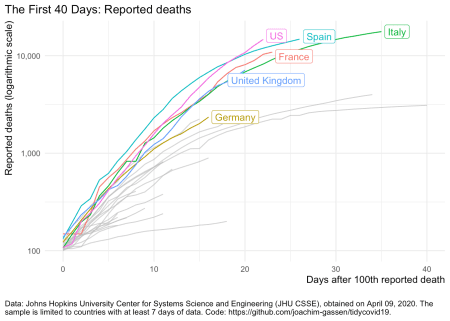
edate\_cutoff = 40, per\_capita = FALSE, log\_scale = TRUE,

cumulative = TRUE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



The log scale enables us to see the declining growth rates. This is good news that is far less obvious without the log scaled Y axis. Can we make this change in growth more pronounced? Yes. Let’s look at daily new cases.

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 500, min\_by\_ctry\_obs = 7,

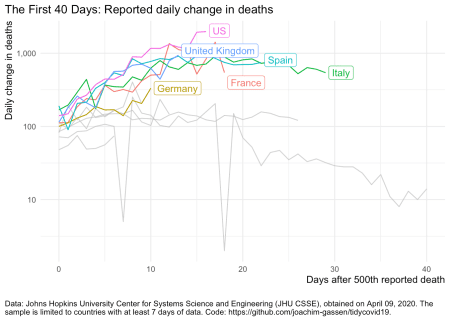
edate\_cutoff = 40, per\_capita = FALSE, log\_scale = TRUE,

cumulative = FALSE, change\_ave = 1,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



Uuh uuh. This looks messy. Daily changes are noisy, also because of data collection reasons. Let’s smooth this over seven days.

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 500, min\_by\_ctry\_obs = 7,

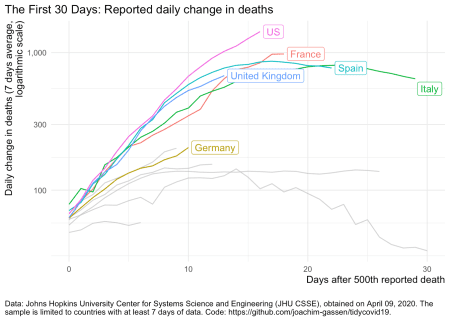
edate\_cutoff = 30, per\_capita = FALSE, log\_scale = TRUE,

cumulative = FALSE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



Ah. This looks better. Now you can see that Italy and Spain have “flattened the curve”. It is easy to miss-understand this graph in a way that fewer people are affected by the pandemic. This is not the case. There are still many people dying and confirmed and active cases are mostly on the rise but there is a trend in Italy and Spain that fewer people die every day. This certainly is good news. For the U.S. however, unfortunatley, things do not look that good yet.

For the last graph, let’s assume that we would like to put less emphasis on the differences across countries and more on the overall trend.

plot\_covid19\_spread(

covid19\_dta, type = "deaths", min\_cases = 100, min\_by\_ctry\_obs = 7,

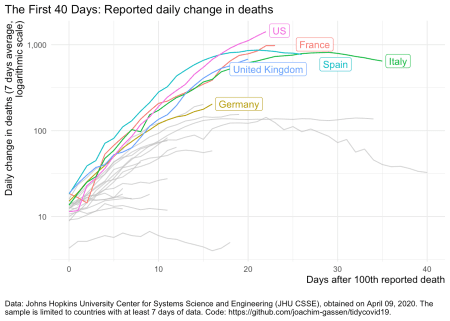
edate\_cutoff = 40, per\_capita = FALSE, log\_scale = TRUE,

cumulative = FALSE, change\_ave = 7,

highlight = c("FRA", "DEU", "ITA", "ESP", "GBR", "USA"),

intervention = NULL

)



**Summing up**

Even using the same data, graphs can communicate very differently. Getting a ‘neutral’ message to the reader is far from trivial. One also has to factor in the level of experience that a reader has, e.g., with different axis scales. Visualizations without guidance can be particularly misleading.