# INF552: Data Visualization

Celine Hajjar

December 2020

## Introduction

The Covid-19 pandemic has generated humongous amounts of data. For this project, I chose to visualize a dataset provided by Our World in Data (**link**). I chose to use the Tableau software for the visualization, for two main reasons: I have already acquired the coding skills in javascript (d3 and vegalite) during the practical sessions, and I created a customized visualization in the last session, merging d3 and VegaLite. Thus, doing the same for my project would be redundant, and learning to use Tableau would be interesting. Second, using a high-level interface takes less time and leaves room for more creative functionalities than low-level coding. The final dashboard is available in the submitted file, and at this **link**.

## 1 Dataset Overview

The dataset is collected from a variety of sources (ECDC[1], Johns Hopkins University, the UN,...). A detailed description of the different variables can be found **here**. Some pre-processing and data cleaning was necessary, and was done using the Pandas Python package.

## 2 Visualizations

I made several types of visualizations: some are classic, like maps and line plots we see in every dashboard, and some are a bit more insightful, like scatter plots showcasing the relation between several attributes, and finally, some are aesthetically pleasing.

### 2.1 Maps

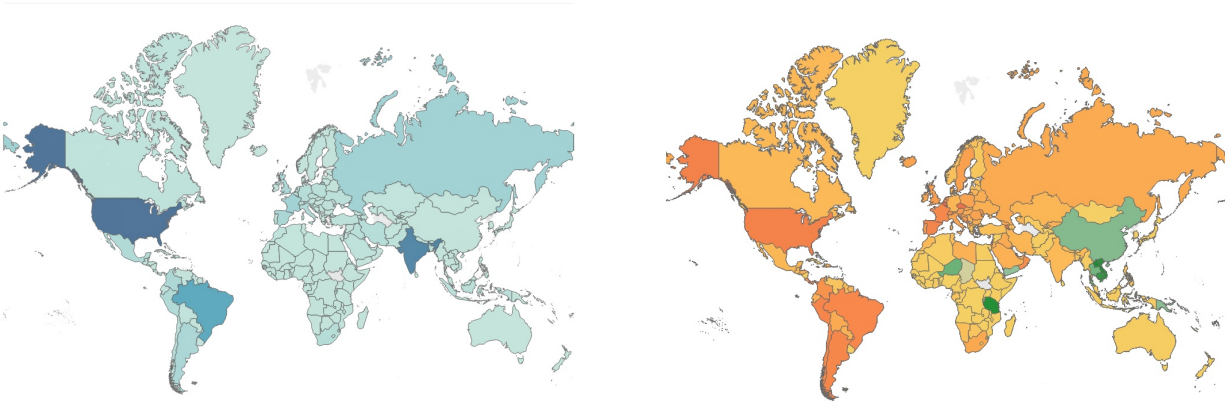
The main map I show in my dashboard represents the temporal evolution of the total number of cases per million inhabitants in every country.

**Color Choices** The color legend chosen by default was the blue-green gradient shown in figure 1a. However, this palette fails to depict the contrast between the number of cases. It is difficult to tell when a country first makes the shift from not being affected at all or having no confirmed cases to having cases, because the difference in saturation between the various shades of blue isn't striking. That is why I created a customized palette, with colors diverging from green to yellow to red. Consequently, we can clearly make out the degree of the severity of the disease across countries. Humans are more likely to perceive green as positive, yellow as cautionary, and red as dangerous. Both gradients are shown in figure 1.

### 2.2 Line Plots

Another basic visualization is the line plot of the number of new cases recorded with time. As seen in session 3 of the course, it wouldn't make sense to show all the countries at once. The graph would be

[1] European Centre for Disease Prevention and Control



(a) Ambiguous Gradient

(b) Better Gradient

Figure 1: Two Examples of Gradients

incomprehensible. Instead, I chose to group countries by continents, so we end up with way less elements. Also, I represent the lines as part of a whole, which in this case would be the total new cases in the world (figure not shown in this report, please refer to the dashboard).

## 2.3 Sunburst Plot

Another interesting temporal visualization is a sunburst line graph representing the Case Fatality Rate (CFR) variation with time in all countries.

$$CFR = \frac{Total\ Deaths}{Total\ Cases}$$

This type of graph is not provided by the Tableau interface, so I had to make some manipulations in the data. I duplicated the data source, so every country is represented by 2 points (start and end of line), and I added calculated attributes such as the radial angle and the $(x, y)$ coordinates to place the countries in a circle. Details about this method can be found **here**. Moreover, I had to manually add an index column sorted by continent such that countries in the same continent are shown next to each other.

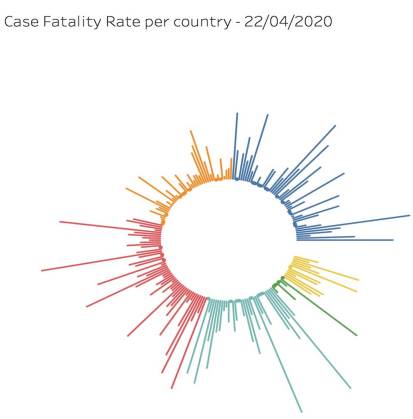



Figure 2: CFR in April


Figure 3: CFR in November

The major impression we get from this animation is the notable reduction in the fatality of the disease. In figure 2, the "sunburst" is intense, which clearly depicts the first wave that hit most countries. However, as at-risk groups became more cautious, and medical staff more prepared, the death rate went down (figure 3), even in countries that witnessed a second, more powerful surge in cases (like European countries, shown in red).

Another major observation is the orange line that stands out from the rest in figure 3. This corresponds to the CFR in Yemen, a country whose healthcare system, already devastated by five years of war, collapsed due to the pandemic[2].

In the animation, we can also visualize the evolution of the number of affected countries with time. As the number of sunrays increases, we know that more countries have entered the Covid mortality race.
A video of the animation is available in the submitted file (name: video_CFR.mp4).

## 2.4 Multi-variate Visualizations

The dataset has several interesting attributes. In order to exploit them effectively in my project, I was inspired by the methods and principles mentioned in the course material, with regards to multivariate visualizations.

### 2.4.1 Relation between age and death rate

In one part of my dashboard, I visualize a treemap that represents for each country the relation between the death rate and the percentage of the population aged 65 and more.

I chose to encode the age with the size of the rectangles, and the death prevalence with color. The reason for that is the fact that the death rate is the important variable I'm trying to show, and color is more dominant than size, especially with the gradient I chose (black-white). Results are shown in figure 4. We notice that Japan has the highest percentage of old people, but a relatively low death rate. The opposite is true for Peru.

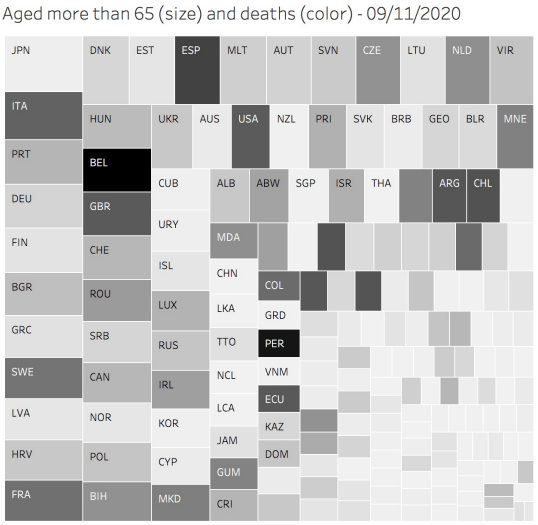



Figure 4: Deaths per Million (color) vs Old Population (size)

[2]Looi, Mun-Keat. 'Covid-19: Deaths in Yemen Are Five Times Global Average as Healthcare Collapses'. BMJ, July 2020, p. m2997. DOI.org (Crossref), doi:10.1136/bmj.m2997.

### 2.4.2 Testing vs GDP

To make it simple, I chose to make a scatterplot showing the number of tests per thousand with respect to the GDP per capita. Unsurprisingly, richer countries tend to have more tests (figure 5).

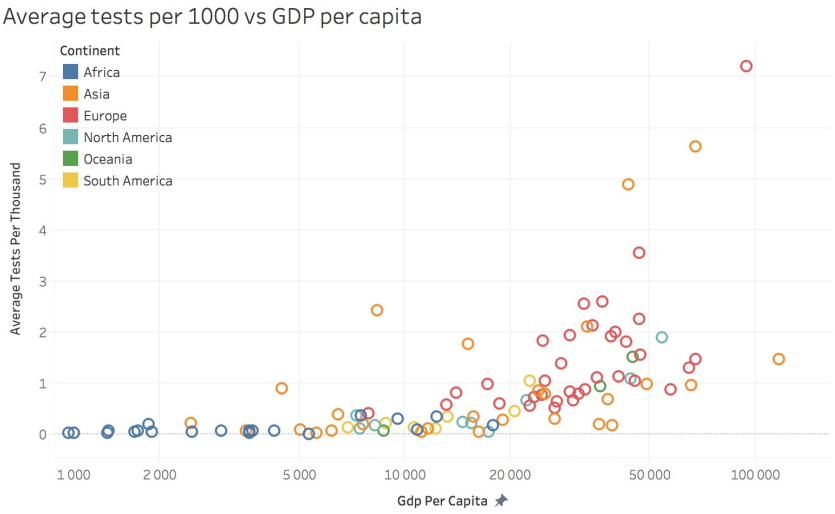



Figure 5: Countries with higher GDP per capita do more tests

### 2.4.3 Parallel Coordinates Plot

In order to plot 3 variables, I used a parallel coordinates plot. The three variables I took for each country are: the **average stringency index**, which is a composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, the **total number of tests per thousand**, and the maximum **CFR** recorded. As seen in figures 7 and 6, it is more logical to separate countries per continent. After the separation, I was able to notice a few features. For example, South Africa and Moroco, the topmost lines in the Africa group, have high stringency, high number of tests, and low CFR. They represent the ideal scenario.

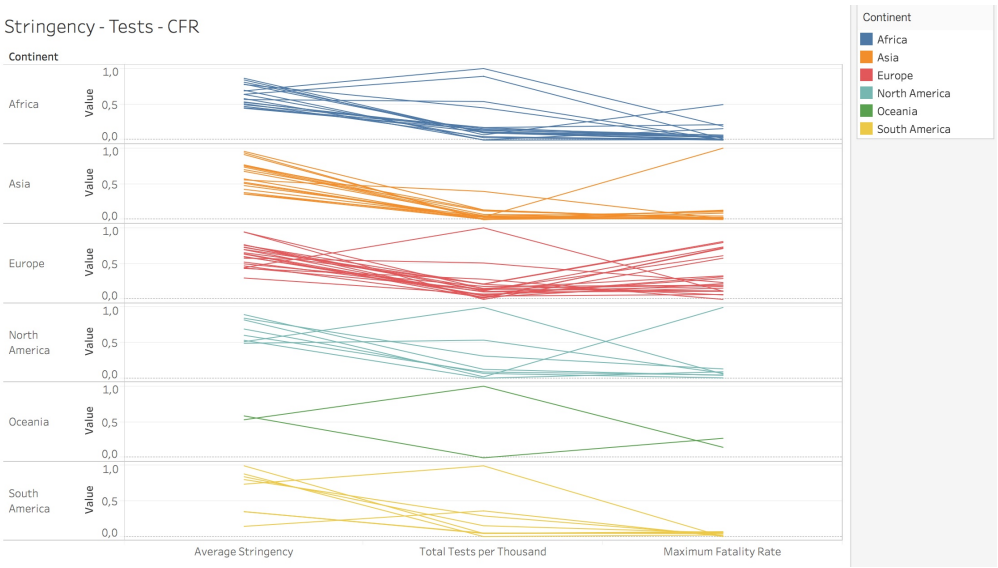



Figure 6: Clear Parallel Coordinates Plot

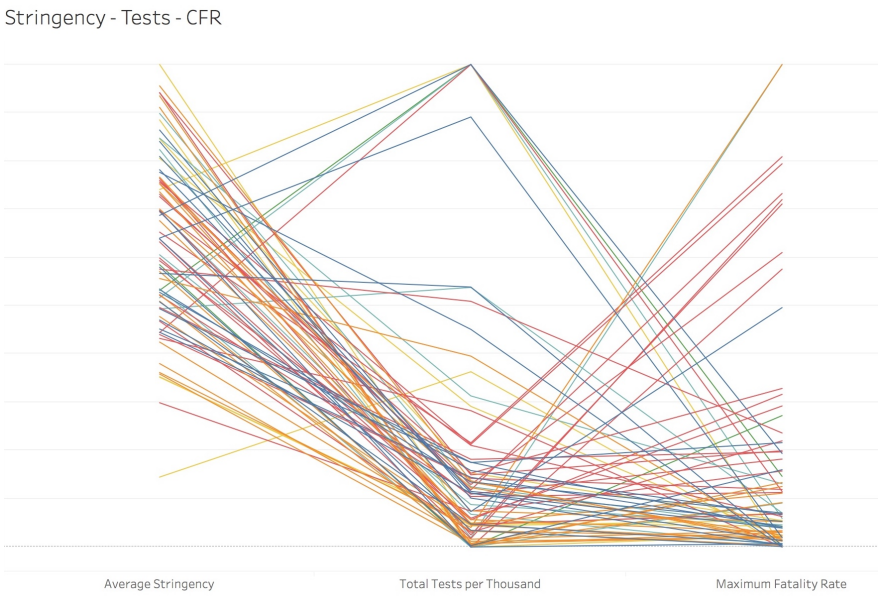



Figure 7: Ambiguous Parallel Coordinates Plot

# 3 Final Dashboard

The final dashboard is composed of three main parts, juxtaposed vertically. The first part, related to the number of cases, is represented by a map showing the total cases and a line plot. The second part consists of death-related visualizations. The first two parts are animated, one has to click on play in the corresponding widget. The last part consists of the analytical graphs that try to make sense of the relation between multiple variables. The dashboard, the corresponding sheets, as well as some extra exploration sheets can be found in the **Covid_INF552.twbx** file in the submitted folder. They are also available at this **link**.