IVS Final Project

Umberto Pasinetti

27 May 2022

1 Before the Plot

1.1 Dataset

The dataset in use is referred to the ***2015 World happiness report (WHR)***. It is composed of 158 rows and 12 variables, and the data are categorical and numerical. I found this dataset searching between the ones available in Kaggle. To find a suitable one, I've filtered the datasets available specifying a maximum file-dimension of 70Kb, and then, ordering the results obtained by usability.

1.2 Approach followed

I decided to realize a big visualization in which a lot of small scatter-plots are located close to each other. My **goal** was to compare some different variables to show which are the ones that affect the most the countries score (and consequently, the positioning in the overall happiness scale).

The representation offers the user **two evident messages.** The first one regards the variables that affect the most the position of the countries in the happiness scale. On one hand, the paths followed by the points in the plots related to the *GDB per Capita, Family* and *Healthy Life Expectancy* show that a high positive correlation with the *Score* exits. On the other, the plot regarding the *Freedom* has a low positive correlation, while the ones on *Corruption Perception* and *Generosity* demonstrate that these last two variables do not significantly affect the y-axis variable. The second message the user can perceive is the average positioning of each continent (deducible from the zone of the plot in which the biggest part of the countries belonging to a specific continent are located). They exist some exception, but in the majority of the cases, starting from the top and reaching the bottom, the order is: Oceania, America, Europe, Asia and Africa.

2 During the plot

2.1 Tools

Once decided to use the 2015 dataset, the **first step** consisted in arranging the data to make them ready for the representation. I removed two columns (*Standard.Error* and *Dystopia.Residual*) to have an even number of variables to plot, and renamed the columns to both have easier labels to recall and labels already correctly named.
Then, I exploited the data contained in the *"Region"* variable to insert the different countries in

some small sub-matrices using the methods *filter()* (from the *dplyr* library) and *cbind()*. And after, I've joined all these sub-matrices with *rbind()* creating a new table ready for my plots.

During the **second step,** I realized the visualizations starting from the particular and expanding to the general. Here, after coding the first plots skeleton, I've improved its general appearance thanks to the libraries *"RColorBrewer"* and *"hrbrthemes"*. They allowed me to choose the palette and to insert the dark theme. I've also manually modified some other theme parameters to make the representation lighter. Then, I've realized all the other plots simply changing the variable on the x-axis, and especially, I've located the plots close to each other using the method *ggarrange()* from the *"ggpubr"* library.

2.2 Problems

The **main problem** I faced has been the dataset choice and the data manipulation[1]. In my opinion it would have been more interesting to base my project on the 2022 (or, at least on the 2021) statistics. Unfortunately, even having found these dataset versions on the WHR official site, I was not able to manipulate the data offered to obtain a division by continent of the countries.
Indeed, the 2015 version of the dataset was more manageable thanks to an additional column called *"Region"*. This allowed me to group the countries to assign them their relative continent inside a new *"Continent"* column.
Furthermore, after adding this variable, I tried to merge the 2022 to the 2015 matrix, to obtain a new 2022 table with the *"Region"* column. In the end, I realized this was only partially a good idea: even if I should have done a bit of data cleaning once the merging of the two previous tables was done, it wouldn't have been worth it, because the 2022 data would have been mutilated. The graph would have been an ad-hoc graph in which some countries of the 2022 report wouldn't have been present because of their manual removal.

During the 80% of the time spent on this project I worked on the 2022 dataset. As it's possible to see in *Figure 1* and in *Figure 2* (see document last page), the final result was almost the same: the only difference was the presence of the *"Rank Position"* variable (instead of the *"Continent"* one) related to the color dimension.
I was very frustrated because I knew that the 2022 dataset was better but without finding a way to insert the "Continent" column. And I knew also that this column was crucial: associate the color to the *"Rank Position"* was redundant (see *Figure 2*)[2], while associating it to a *"Continent"* variable, would have represented an opportunity to smartly use this other plot dimension.

3 After the Plot

3.1 Critique

I've decided to use the scatter-plot because this kind of graph is one of the best in plotting distributions and, if present, highlight eventual correlations.
In the general plot, three perceptual tasks are involved. The first one requires the user to decode the position in the common scale, and is related to every small plot x and y-axis. Then the user has to understand the position in a non aligned-scale, observing the big visualization x and

[1] Initially, I was planning to base my project on a total different dataset related to chocolate bars ratings. Its problem was the too big amount of categorical data, and, in my opinion, the scarcity of visualizations it was able to offer.

[2] There was already the *"Score"* variable; basically, having a high score means being located in a high position in the rank

y-axis. In the end, the third one regards the meaning of the color. They cover respectively the 1st, the 2nd and the 6th position in the perceptual task ranking.

The structure of the representation is the following: in the y-axis of each small-scatter plot, it's always located the *"Score"* variable, while on the x-axis, there is a different variable every time (for instance, *"GDP per capita", "Family" "Healthy Life Expectancy"...*).
Moreover, I decided to add the color dimension to show how the countries are located (according to the previous variables involved) in respect of the continent they belong to. And, the color palette, that goes from red to dark-green, is not chosen without reason. The meaning of these colors (located in the extremes of the plots) is linked with sociocultural meanings and conventions strongly fixed in humans visual communication habits. Their main function is to represent a specific Continent. But, their association is also made in a way in which, very often, moving from the dark-green (*"Oceania"* countries) to the red (*"Africa"* countries) represents also an implicit transaction from countries of continents that have an average good-positioning to others that are more and more low-positioned. Being green a color commonly associated to something positive and red related to the ideas of danger, the situation it's very likely to a contraposition between "Ok continents" and "not so Ok continents".

The representation respects the 5 principles of the graphical excellence. It doesn't hide or over-process the data and it doesn't distort the data[3].
The main visualization doesn't distract the user. I'm aware that taking the decision to use a dark-theme background was a bit risky. Nonetheless, in my opinion, it was challenging, cool and updated to the current technology trends. I'm also very satisfied regarding its contrast with the color palette. It's also possible to say that this representation is dense, showing a lot of information in a small space, and that the way in which the scatter-plots are located helps the user eye to compare the different plots.

I also tried to reduce the Data-ink the most possible. To do so, I removed the plots x and y-axis lines, the y-axis *"Score"* label, always the same in every plot, and modified their grid, making it lighter outlining it.

[3]Even if little, this had happened modifying the dataset obtaining just the records of the 2022 matrix present also in the 2015 table, to add them the *"Continent"* column

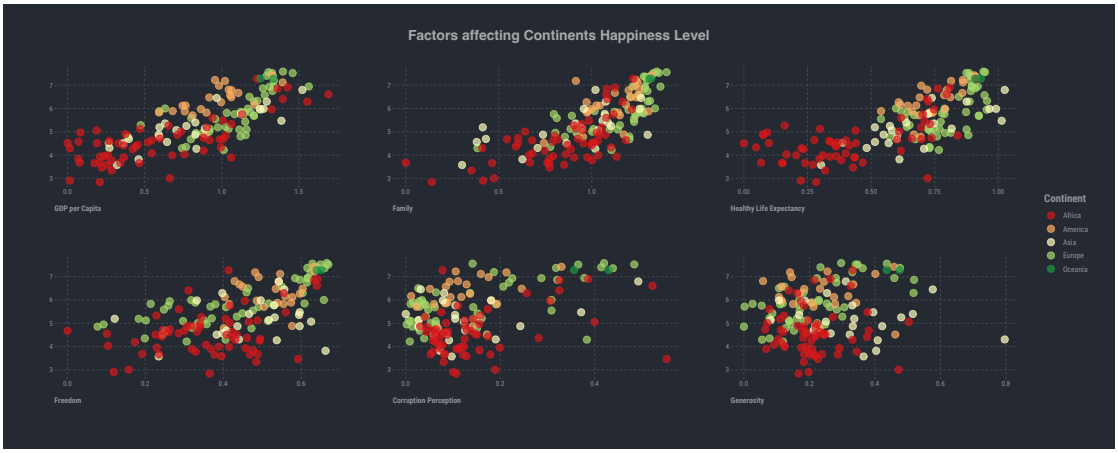

Figure 1: definitive Final Plot

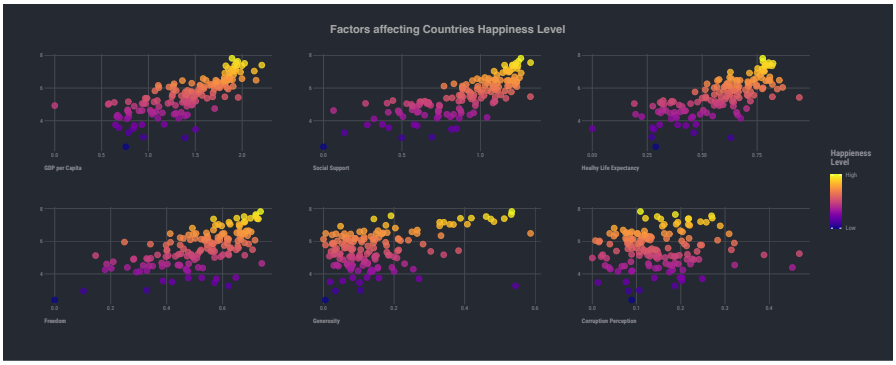

Figure 2: old Final Plot