Evolution of Productivity in European Countries from 1995 to 2021

Data Preparation and Visualization

Master's Degree in Data Science and Engineering - FEUP

November 2022

Group 2

Farzam Salimi (up201007922)

Luis Henriques (up202204386)

Rojan Aslani (up202204382)

Professor: José Luís Moura Borges

---

1. Introduction

Productivity is a measure of economic performance that compares the amount of goods and services produced (output) with the amount of input used to produce those goods and services. With growth in productivity, an economy can increasingly produce and consume more goods and services for the same amount of work. It is an important factor to individuals, business leaders, and analysts (such as policymakers and government statisticians) [1].

Data visualization is a field that enhances the capacity for users to easily understand the behavior of a certain dataset or data stored in a database. A good visualization enables decision-makers to make decisions in a fast paced and competitive environment, and also the general public to easily and quickly understand the state of a certain domain, such as politics, sustainable goals, and so on, or even insights of a company. Moreover, a visualization design enables us to use our perception instead of our cognition, saving our cognitive energy to tasks where we need it more [2].

There are many theories behind ways to maximize the human's capacity to use perception, for example, although we tend to focus on things that are different, if there's many differences, we won't know where to look. This is a very complex topic because the human brain also is, but when understood theoretically and practically, it becomes a powerful tool for data visualization.

In this work, individual European country's productivity from 1995 to 2021 were visualized using R programming language, based on *ggplot2* package. Computer-based visualization systems provide representations of datasets that can help people carry out tasks more effectively and augment their capabilities [2].

Traditionally, graphs representing productivity are plotted with bar charts, line graphs, or maps [3, 4, 5], as demonstrated in Figure 1. Most of these visualization designs compare productivity between different countries in a given year, or 2 different years. In Figure 1 (d) the design does include the trend of each country over several years, but only a limited number of countries are demonstrated, and the high number of lines and shapes makes the graph heavy and hard to analyze at first glance, besides having to remember the color/label attributed to each country.

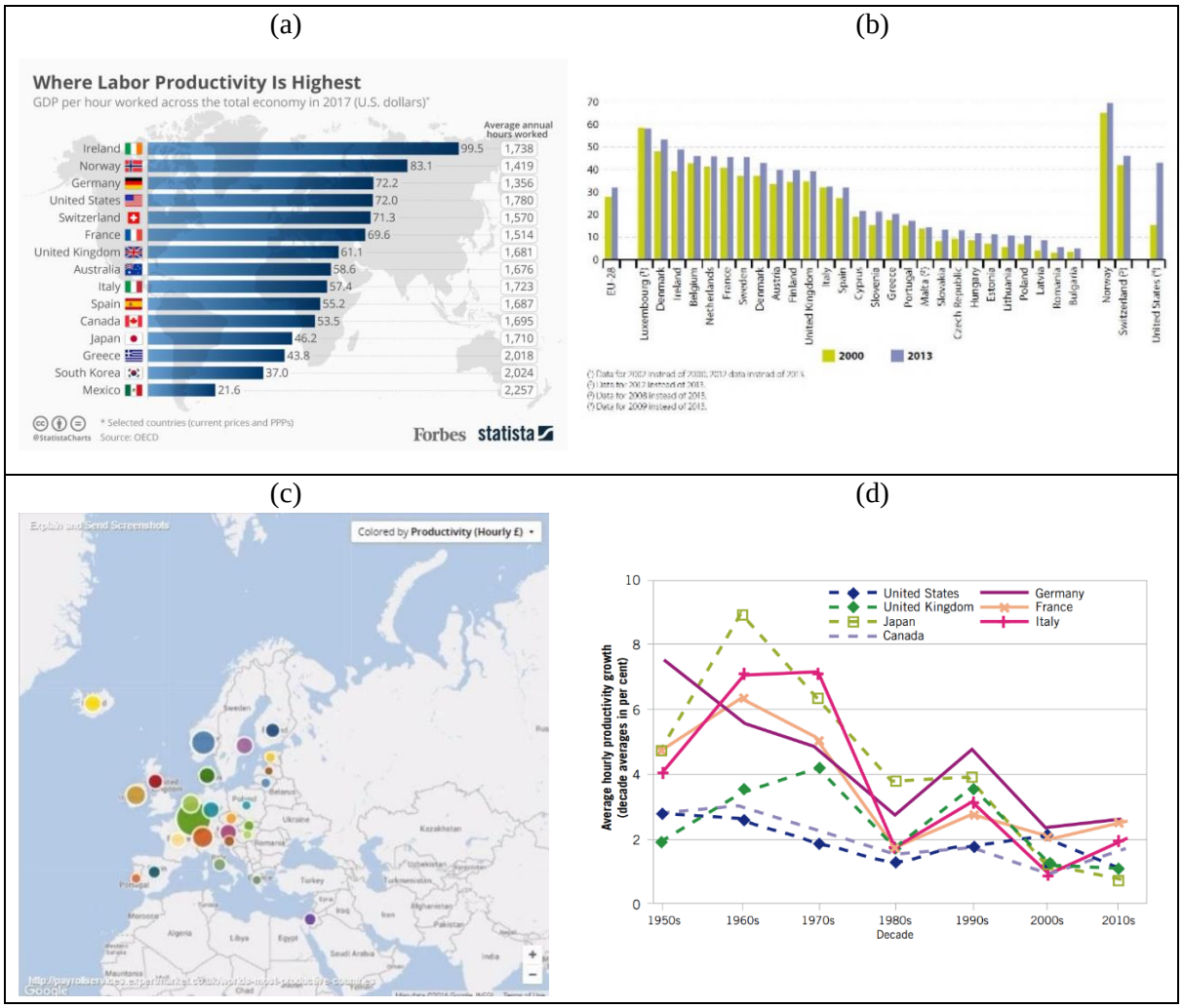

Figure 1. Examples of the state of the art – visual demonstration of productivity of different countries. (a) labor productivity in 2017 [6]. (b) labor productivity in 2000 and 2017 [4]. (c) Analysis of GDP data (year not specified) [5]. (d)Productivity growth from 1950 to 2010 in a few countries [3].

With that in mind and based on the challenge proposed in the Data Preparation and Visualization course, the authors aimed to design an innovative, attractive, engaging, complete and simple visualization, while providing detailed information about the data.

Hence, this project was focused on creating a visualization design that showcased the evolution of the productivity in European countries from 1995 to 2021. The challenge would be to include all the countries, years, and productivities available in the chosen dataset, and overcome the limitations of visualizations in the state of the art of this domain.

2. Materials and Methods

To answer the proposed question, a dataset was obtained from PORDATA website [7]. This dataset contains data about productivity of per working hour in euros (€) from 1995 to 2021 (756 values of productivity for 28 countries in a time interval of 27 years). The dataset is considered accurate, trustable, and complete, besides the few missing values. Although the year is 2022, the dataset is timely because only in the next year productivity can be measured for the previous year. The dataset has few attributes (unrepeated), so there're no inconsistency issues, also given the nature of the variables.

The used dataset includes three variables with distinct types:

- Country name: Qualitative - nominal variable
- Productivity (€/hour): Quantitative - continuous - ratio variable
- Year: Quantitative – discrete – interval variable

The raw dataset is presented in a table where rows represent the years, columns represent countries, and the intersection cell includes the productivity of a given country in a given year. Moreover, the table itself includes metadata, including the dictionary of some symbols present in the dataset.

The visualization's design was done following a structured process, **The Four Level Nested Model** [8]. This framework allows addressing different concerns separately. In this method, an output from the upper nest is being used as an input for the lower nest. As demonstrated in Figure 2, the four layers consist of:

1. **Domain problem and data characterization**: concentrate on the domain of target users and their interest, their data, and their question. A clear understanding of the needs of the audience must be developed, keeping in mind that the audiences can be different groups with different needs and goals in mind.
2. **Operation and data type abstraction**: abstraction of detailed domain questions and data in a domain-specific form. This form is the one decided in the top level and will be abstracted into a generic representation.
3. **Visual encoding and interaction design**: decision making about details and techniques to create a visual representation of the data, according to the previous levels. There can be different approaches to deal with this level, keeping in mind the two main concerns – encoding and interaction.
4. **Algorithm design**: create an algorithm to carry out the visual encoding and interaction designs automatically.

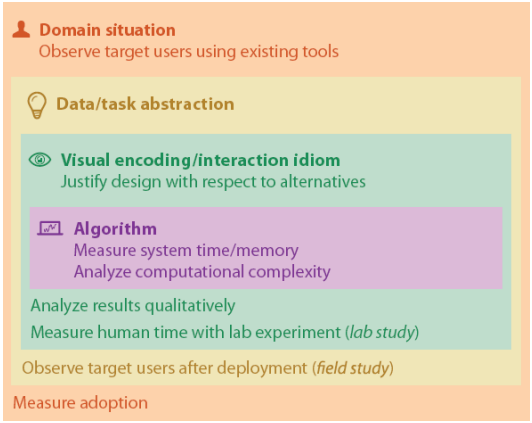

Figure 2. The four level nested layers of visualization design [2].

That being said, and as in other project development processes, it's important to validate the output that is created with the intended audience (sample) for the visualization. To validate the visualization an agile methodology was applied where a minimum viable product (MVP) was developed as soon as possible, and changes were implemented and validated again.

2.1 Domain Situation

Having the task of visualization of the evolution of productivity in European countries, one of the audiences is the general public who is interested in this type of information – someone who is interested in such topics and wants to be updated about new insights of several topics, like macroeconomy. Therefore, it should be kept in mind that the visualization should be comprehensive to the general perception.

Another target audience for this design is decision makers, like investors, shareholders, managers and politicians. For example, venture capital firms may assess start-ups to invest in based on the productivity of the countries where they are based in. More productive countries can be more appealing locations to invest since they can be more assured about the return of their investment. Hence, the visualization must feel reliable, trustworthy, and easily assessed, with a wide range of countries, so that these users can make confident decisions based on it.

2.2 Data/Task Abstraction

As mentioned earlier, the data included information about the productivity of some European countries from 1995 to 2021, which is calculated through the division of Gross Domestic Product (GDP) of the country by the number of working hours of all labor. Besides this data, three additional columns were provided, related to some cumulative measures. As these three columns of data are at this point irrelevant to this project, they were eliminated from the dataset.

At an initial phase, exploratory data analysis was done using traditional charts. An apparent observation was the missing values of productivity of some countries in some years. Removing instances containing missing values is not an option in this project, and a meaningful method to replace these values by prediction was required (explained in section 2.4). The first glance at the data showed an increasing trend related to the productivity during the years. Since 1995, most of the countries have had an increase in their productivity, and what is more interesting is that these countries almost increased their productivity every year. This trend helps in the decision-making process about what method to use for replacing missing value in a realistic way.

2.3 Visual Encoding/Interaction Idiom

In this project several abstraction methods were considered and discussed, each of them having their own pros and cons in showing the data. For example, one of the primary options was a visual encoding using a line chart. In terms of the encoding, line charts would be an excellent option, since they can easily show the evolution of the productivity between different years. These types of charts would be near ideal if we wanted to represent just one or two countries or even a discretized version of the countries based on a specific criteria, like regions.

After considering different options and going through a knowledge development process that included reading several technical books and articles, a heatmap visualization technique was chosen, where the data values for productivity were mapped into colors with different brightnesses. This method highlights

trends and makes visual comparisons easier for the user. In this heatmap conceptualization phase, the productivities of the countries could be easily compared with each other, or for the same country, the evolution throughout the years can be studied, or even for the same year compare all the countries, which can lead to further conclusions/questions.

Hence, as demonstrated in Table 1, productivity was represented by color, where lighter/brighter colors are associated with lower productivity, and higher productivity levels with darker colors. The horizontal axis was dedicated to years variable since it's more usual for the users to read the evolution during the years horizontally, like in a time series. The vertical axis was dedicated to names of countries. One of the reasons for it being the fact that the countries are represented by text, and the vertical axis allows an easier reading of the axis text. On the other hand, because some countries have a relatively big name, it's best if we place that nominal variable on the y-axis; this is also applicable to when a horizontal bar plot is used instead of the standard vertical one.

Table 1. Data names and types and the visual encoding used to represent them in the visual design

Variable	Variable type	Visual encoding
Country name	Qualitative – Nominal	Position (y-axis)
Productivity (€/hour)	Quantitative continuous – ratio	Sequential color scale
Year	Quantitative discrete – interval	Position (x-axis)

Color, position, dimensions, lightning, and contrast are characteristics that play an important role in visual perception. Each of these characteristics was carefully assessed and chosen.

2.4 Algorithm

This level is influenced by all the design choices, which is a detailed and comprehensive process letting the computer do the design task to reach our intended visualization. In the previous section was chosen a heatmap as our encoding method, so the goal is to efficiently encode the data into this type of visual representation. The compilation of the code was done in RStudio, using the R programming language [9].

Initially, the data went through a pre-processing phase to improve its quality. The steps include:

- **Data format:** The data was transformed from wide to long format, increasing the number of rows and decreasing the number of columns. The 1st column was dedicated to the year number, the 2nd column was dedicated to country name (translated to English – the original dataset is in Portuguese), and the 3rd column was filled with the productivity of each country in each year.

- **Replacing missing values:** This step was done using k-near neighbor (kNN) [10] algorithm. The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then these 'k' samples are used to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset. In this work, we did the calculations with k = 2.

- **Ordering:** The countries were ordered in a decreasing order according to the country with the highest productivity (maximum value throughout all years).

Next, the data was plotted. Within the process of plotting the data, variables were attributed according to Table 1 (productivity was used as *fill* of the *geom_tile()* function). Reversed *Mako* colorblind-safe sequential color scale was used to represent the productivity. Economist theme was used but adjusted to fit the authors' vision. Labels were chosen carefully to best describe the data. The x and y axis labels (year and country name, respectively) were removed from the design, as the title and the data make these variables obvious to the user, leaving only their values on the axis lines.

3. Results and Discussion

The outcome of this work is a visualization designed to present the evolution of productivity in European countries through the years. This visualization is presented in Figure 3.

This type of visualization (heatmap) is interesting to be used for this case because it allows encoding the three variables in the same graph in an easily perceivable way, giving obvious visual cues to the reader. In comparison to other techniques of visualization (bar plots, geographical maps), heatmap is a promising choice because it allows the demonstration of a big set of data, without being too heavy on the reader's eye. Also, it avoids the misleading effect of areas, using geographical maps.

Human vision is limited and is optimized to detect contrast [11]. The chosen color palette (mako) has limited number of colors, with high variance in the brightness levels, which are associated to higher and lower values in human eyes, by nature. This color palette saves the user from confusions and having to use their memory to remember what each color represents, ameliorating the visual perception. Moreover, according to color psychology, the emphasize on blue (background and in the plot) which is a peaceful and calming color, can trigger various emotions such as trust and reliability to the visualization [11].

The typeface was also chosen based on the emotions and associations the user does. With that in mind, and because we intend to pass a trustworthy and professional image, the chosen typeface was *Merriweather*, a serif typeface. *PT Sans* was also considered as a promising typeface because of its minimal and modern look, while having a more creative vibe than serif typeface. Ultimately, in order to emphasize on the feeling of trust and the other upper mentioned emotions, and in line with the colors chosen, a serif typeface was used [11].

The productivity measure in this dataset is not only dependent on the labor itself (hours worked) but also on the GDP of the country in each year. This can influence the results greatly. As an example, Ireland seems to have evolved more than all other presented countries in this time interval, which might be associated to the fact that the country decreased tax percentages in certain domains and many multinational companies moved their headquarters to this country, increasing the GDP of the country even though the labor might not necessarily be working more productively.

It is also interesting to note that Greece was having an increase in productivity, reaching its maximum productivity at around 2008 (crisis) and then slowly declining from then on. These observations raise further questions regarding the association of productivity with other factors.

One weaknesses of the proposed visual representation is not being able to read the exact value of each cell, since it is represented by a color scale. However, keeping in mind the objective of the work (finding trends and distinguishing the most productive countries) and the intended audience, this representation clearly visualizes the intended objectives. One iteration of the visualization was with black lines separation every cell in the heatmap, but because that would confuse the user on what was the objective of this visualization (compare every cell vs productivity evolution), the gridlines were removed.

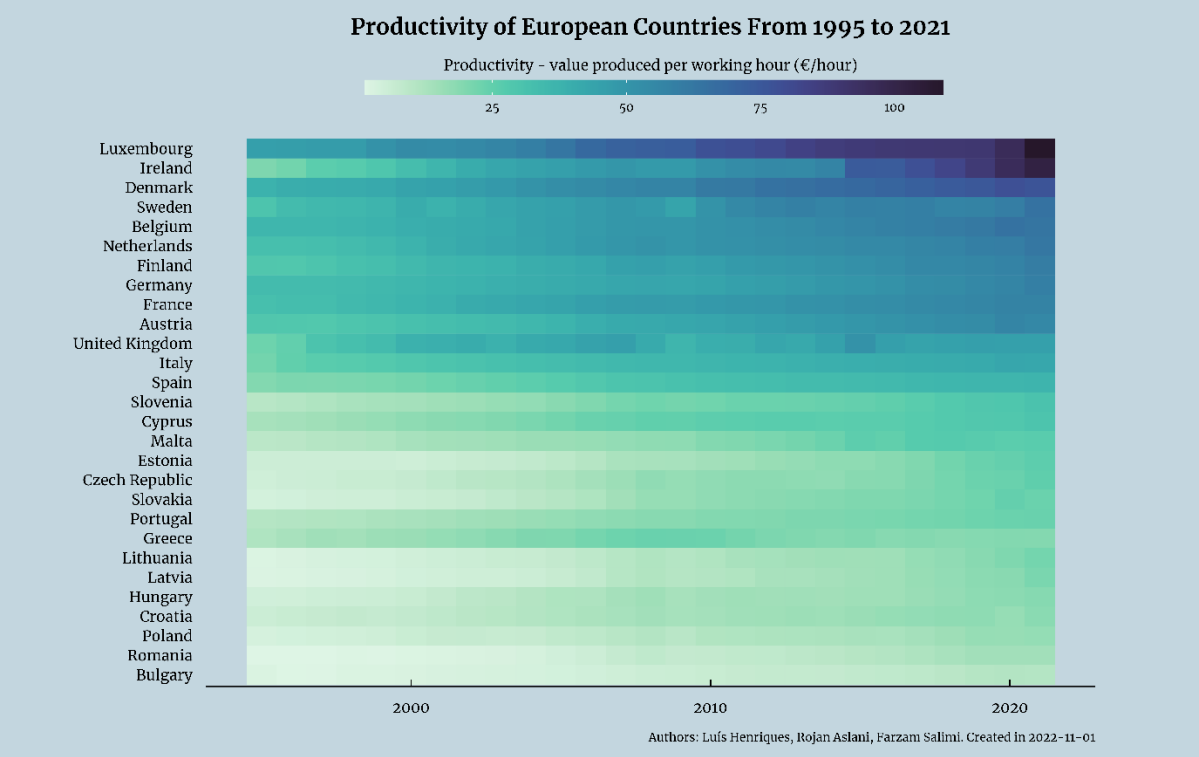

Figure 3. Proposed visualization to demonstrate the evolution of productivity of European countries from 1995 to 2021. Productivity (€/hour) is presented using Mako color scale, with brighter colors representing low productivity. Countries were ordered according to the maximum value of productivity in the latest year. Data source: PORDATA available at [7].

Regarding the validation process, the design presented in Figure 3 was shared with 15 unbiased users. The users were asked to identify the task of the visualization and share their insights. All users identified the task in line with the authors' objective. Users were also asked to give an estimate of the productivity (€/hour) in a certain year for a certain country. Even though on average they did not find this task easy, they were all able to determine an approximate value to the real one.

Overall, although a heatmap can make it trickier to analyze the data in detail, in general the users were able to identify key factors and interpret the main goal. To finalize, the users were asked if they associate brighter colors with higher or lower productivity. Contrary to the authors' belief, most users associated bright colors to lower productivity, hence we used the presented (reversed) color palette.

4. Conclusions and Future works

In this work a computer-based visualization model was designed to represent information about productivity of European countries from 1995 to 2021. R programming language and *ggplot2* package (alongside with other packages) were used. The obtained plot correctly matches the designers' conceptual model with users' mental model, based on the validation process implemented throughout the development of the visualization and the implementation of the nested model upper mentioned.

Additional questions were raised about the productivity of the countries and what might be impacting them. In future works it would be interesting to evaluate the effect of factors, such as average salary, population's education, geographical location, on the productivity of countries. To do so, ridgeline plots can be considered as an interesting visualization technique. As an initial step it would be interesting to replace the productivity variable by the mentioned factors to produce similar heatmaps, and compare the resulting visualization with the current design.

5. References

[1] K. Burton, Encyclopedia of Busine$$ and Finance, New York: Macmillan Reference USA, 2001.

[2] T. Munzner, Visualization Analysis and Design, A K Peters/CRC Press, 2014.

[3] E. Ernst, R. Merola e D. Samaan, "Economics of Artificial Intelligence: Implications for the Future of Work," 2019.

[4] E. s. explained, "Labour productivity per hour worked by country 2000 and 2013," 2015. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Labour_productivity_per_hour_worked_by_country_2000_and_2013.jpg.

[5] L. Dishman, "The world's most productive countries also have the shortest workdays," 2016. [Online]. Available: https://www.fastcompany.com/4016006/the-worlds-most-productive-countries-also-have-the-shortest-workdays.

[6] N. McCarthy, "Where Labor Productivity Is Highest [Infographic]," Forbes, 5 2 2019. [Online]. Available: https://www.forbes.com/sites/niallmccarthy/2019/02/05/where-labor-productivity-is-highest-infographic. [Acedido em 25 10 2022].

[7] F. F. M. d. Santos, "Produtividade do trabalho por hora trabalhada (Euro)," PORDATA - Estatisticas sobre Portugal e Europa, 27 05 2022. [Online]. Available: https://www.pordata.pt/europa/produtividade+do+trabalho+por+hora+trabalhada+(euro)-3019. [Acedido em 17 10 2022].

[8] T. Munzner, "A Nested Model of Visualization Design and Validation," *IEEE TVCG*, vol. 15, nº 6, pp. 921-928, 2009.

[9] H. Wickham, "ggplot2," tidyverse, 2022. [Online]. Available: https://ggplot2.tidyverse.org.

[10] L. Targo, Data Mining with R: Learning with Case Studies, Minnesota, USA: Chapman & Hall/CRC, 2011.

[11] L. Holtzschue, Understanding Color: An Introduction for Designers, Wiley, 2011.