UNIVERSIDAD
POLITÉCNICA
DE MADRID

Data Visualization practical work

European Innovation Scoreboard

Lorenzo Framba: lorenzo.framba@alumnos.upm.es
Ostap Kharysh: ostap.kharysh@alumnos.upm.es
Federico Rodigari: federico.rodigari@alumnos.upm.es

Department of Computer Science
Master EIT Digital in Data Science

January 23, 2021

1 Problem characterization in the application domain

The overall goal of this project is to correctly visualize some deep level questions over a set of data. The method for visualization it's over the R package Shiny.

We decided to provide the answers, through the use of a 2019 open dataset called the *European innovation scoreboard*, which provides a comparative analysis of innovation performance in EU countries, other European countries, and regional neighbours. Thought the tables, it is possible to observe all the relative strengths and weaknesses of national innovation systems. Therefore, is a tool to help countries identify areas they need to address.

2 Data and task abstraction

With the data provided we must answer three questions in order to create a visualization of the data. *What*? *Why*? and *How*?

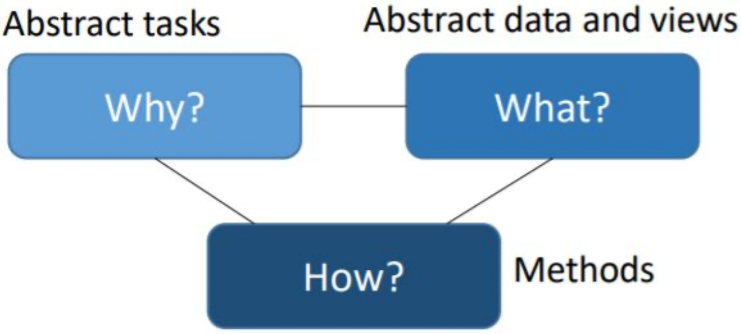


Figure 1: main visualization principles

2.1 Data abstraction – *What*?

With this question we want to describe in depth the chosen dataset and its data types for all the variables.

The dataset is composed by 34 different sheets of different data, and each sheet describes different parameters regarding the innovation.

With our first question we are using 29 rows and 2 columns: 1 quantitative and 1 categorical. Along with the dataset we will use additional information geospatial data with 39 items each consisting of sum of Polygons and 1 categorical value.

The sixth sheet is used to perform the visualization of the question two, which is composed by 37 rows. Each of them correspond to 28 quantitative columns representing the innovation metrics. In order to reach the goals of the question, an additional column is manually added, representing the different region of Europe. This variable is named *"Region"* and is converted to a factor with 4 levels.

The third question will be answered by using the third sheet, which has 37 different observations per 8 years of research. Each Observation is tracking 11 different quantitative attributes, which describe an innovation index parameter per each country. In addition to these columns, an additional categorical attribute was created. This variable is called *"Year"* and it classifies the observation to the corresponding year. This will allow us to analyze trends over time.
We also have another categorical variable, *"Country"*, which is used to classify the observation per each country of reference.

During the import of all these sheets, the rows with NA values were omitted and the names of the columns renamed to have a better understanding of the variables.

2.2 Task abstractions – *Why*?

To properly answer this question, we must make a differentiation among the three visualizations of the data that are going to be produced. We also need to specify what are the various actions and the targets to obtain to answer the question "why?", in each visualization task.

1) *What's the distribution of Research Development expenditure score applied in business sector in the different areas of the Europe?*

We want to understand how aligned are the European countries in terms of financing their RD activities in business sectors and to identify clusters and distribution to evaluate overall situation in Europe.

2) *How do the different European innovation performance metrics correlate with the number of new doctorate students in the different regions?*

We want to understand which are the factors affected by the number of new doctorate students. We look here if the number of the students is in some way correlated with the other innovation metrics taking always into account the overall score provided from the *"Summary Innovation Index"*. We are also interested in understanding if there is a consistent difference in performance between the different European regions.

3) *"How does an innovation parameter change over time, given a chosen location?"*

We want to understand if is possible to find knowledge over the years. We are interested in analysing all the trends per each country, per each innovation parameter, to see if there was an overall improvement of the country in that scenario. We also think it's important to compare each selected country, so it shows how countries are performing one vs another. This will be helpful to find trends, outliers, distribution and compare values altogether.

3 Interaction and visual encoding – How?

- **First question**
  With our approach, we want to explore the quantitative characteristics in countries. As a part of the analysis we use geoJSON from Geojson Maps. In order to receive a more appealing view we decided to remove the overseas territories belonging to European countries as those will have no affect on the result of the visualization. It is important to mention that there some countries which are missing the values. We decided to keep them (mentioning "No data provided") to have a complete representation of Europe. However, speaking about Russia, there is no value assigned to it too, and due to the large space taken when visualizing its spacial data, we decided to remove it completely as it does not provide any important information to contribute to answering our question. The interaction with the answer plot consists of ability of user to select the country and see the value assigned to it, with color identification which country is selected.

- **Second question**
  In order to answer the question in an effective way, we considered a bubble chart to be the best solution. This graph has one axe(Y) fixed on the number of new doctorate students and on the other axe(X) an innovation metric that can be chosen from the user. When displayed, the size of the bubble is given by the overall summary innovation index that allows the user to understand if the variable selected are relevant in the overall performance of the nation. Thanks to the distinction made by EuroVoc the nations were splitted in four factors based on their geographical location: "Western", "Eastern", "Northern" and "Southern". This division is visually provided from the colors that were selected with the help of ColorBrewew. The last functionality of this visualization is the printing of a line that estimates the trend of the correlation. This make possible to the user to get an instantaneous impression about the correlation where present.

- **Third question**
  In this case we want to represent the trend of all the innovation parameters in the countries of Europe across the years. This could give an idea if the countries are making efforts in increasing the chosen value and see a clear difference across the years, and it will compare the results per each selected country. Two continuous attributes are going to be represented, value of the chosen parameter(Y) and time (X). It is easy to see patterns since its visual channel, length, is the more effective and expressive. Regarding interaction with the end user, it can be selected the countries to be compared, among all the Europeans' ones, and the parameter among a list of all the relevant metrics to value how innovative is a country, to check for improvements and trends.

4 Implementation

The library to represent the Line Plot was *ggplot2* and to represent the map was *Leaflet*, which is an open source library for mapping applications. In addition to this, for the second question, the library ggtips was used to enhance the ggplots with eye-pleasing tooltips. In order to give interaction, the library *shiny* was implemented. The representation is composed by two parts, the **ui** and the **server**. For the interaction with the graph, different ways are provided, such as selectors(for single and multi choice), radio buttons (for a single choice) and the mouse click.

4.1 How to execute

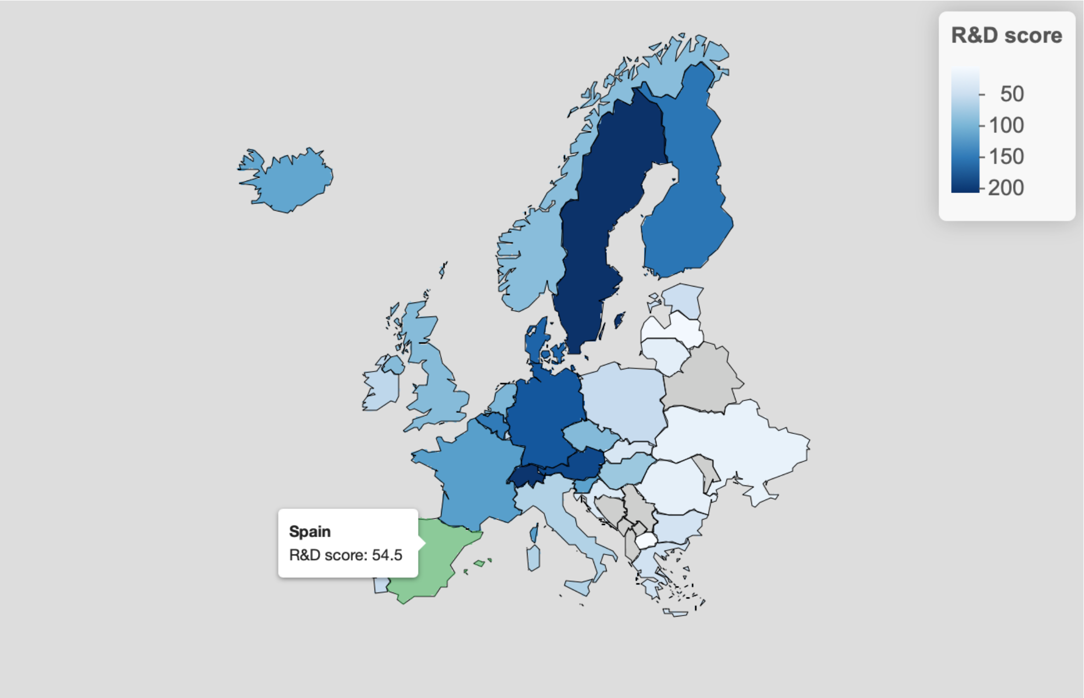
The implementations are loaded to apps.io and could be accessed via the link below each visualization. For question 2, make sure to use Chrome in order to visualize the mouseover function that displays the name of the nations, because for some reasons, Safari doesn't seem to support at its fullest capabilities.

If you want to run the code manually, you could download it from the GitHub repository and run them from the directory as the ordinary apps. But, mind that for the question 2 the installation of "ggtips" is required. To install ggtips, use devtools: devtools::install_github("Roche/ggtips")

5 Results

5.1 Idiom 1: Choropleth map

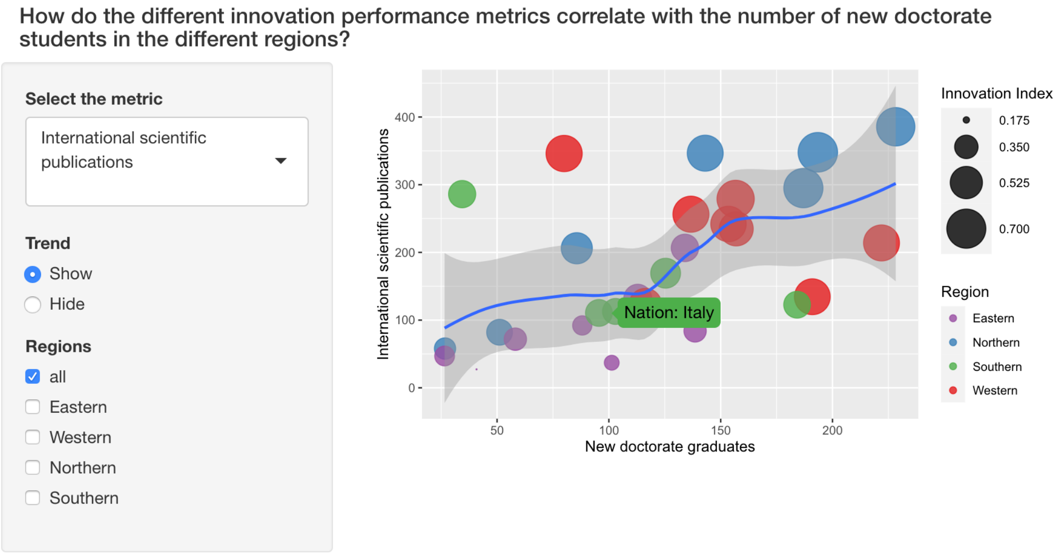
Idiom 1: Choropleth map	
Question:	• What's the distribution of Research & Development expenditure score applied in business sector in the different areas of the Europe?
Data:	• Table of quantitative data • Geographic Geometry data (Polygons)
Encode:	• Spacial representation: us of the geometrical representation of countries' boundaries • Color: sequential colormap of blues
Task:	• Find clusters and distribution among certain regions of Europe to infer a summary
Scale:	• Dataset items: 58, geoJSON items: 78 • Quantitative attributes



Link to open the implementation

5.2 Idiom 2: Bubble Plot

Idiom 3: Bubble Plot	
Question:	• How do the different European innovation performance metrics correlate with the number of new doctorate students in the different regions?
Data:	• Table: three quantitative attributes and one categorical as factor.
Encode:	• Spacial representation: Bubble Plot with size as channel • Color: hue to divide the categorical values from each other.
Task:	• Find trends, outliers, distribution, correlations and differences between regions
Scale:	• Dataset items: 28 observations of 31 variables • 28 Quantitative and 3 categorical attributes



How do the different innovation performance metrics correlate with the number of new doctorate students in the different regions?

Select the metric

International scientific publications ▾

Trend

● Show

○ Hide

Regions

☑ all

□ Eastern

□ Western

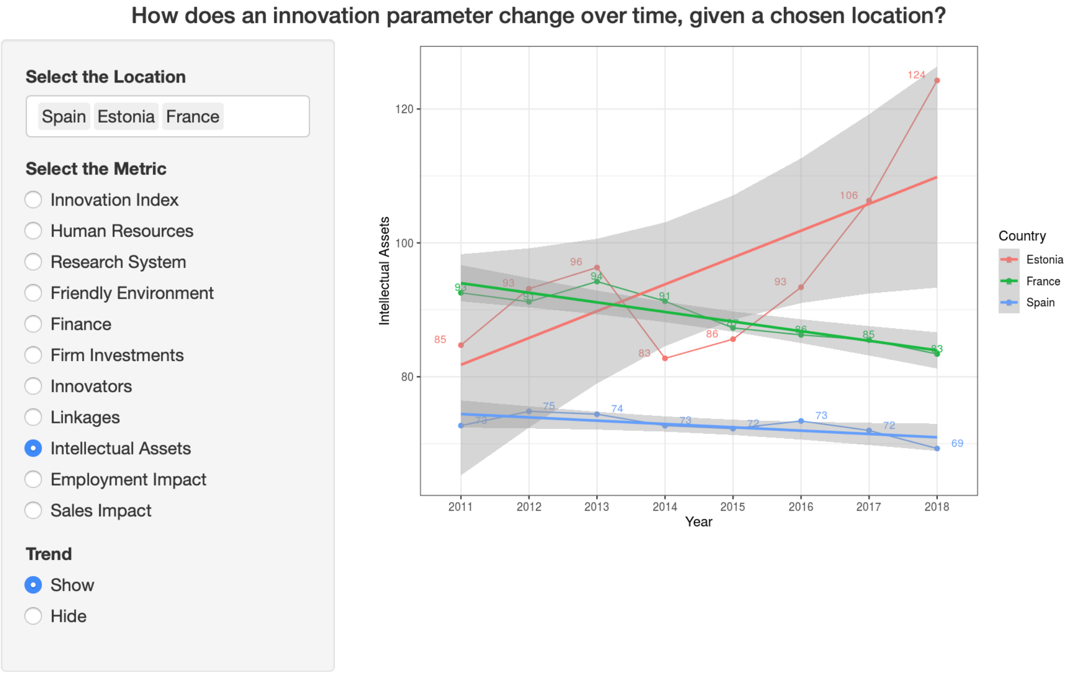
□ Northern

□ Southern

Link to open the implementation

5.3 Idiom 3: Line Plot

Idiom 3: Line Plot	
Question:	• How does an innovation parameter change over time, given a chosen location?
Data:	• Table: two quantitative attributes, one as factor.
Encode:	• Spacial representation: with Line Plot with length as channel • Color: to divide the categorical values from each other.
Task:	• Find trends, outliers, distribution and compare values
Scale:	• Dataset items: 4144 • 11 Quantitative attributes, 2 Categorical attributes



How does an innovation parameter change over time, given a chosen location?

Select the Location

Spain Estonia France

Select the Metric

○ Innovation Index
○ Human Resources
○ Research System
○ Friendly Environment
○ Finance
○ Firm Investments
○ Innovators
○ Linkages
● Intellectual Assets
○ Employment Impact
○ Sales Impact

Trend

● Show
○ Hide

Link to open the implementation