Big Data: Visualization
Practical Work

December, 9th 2018

Javier de la Rúa Martínez

M. Hamza Malik

Indice

1. Problem characterization in the application domain

When it comes to monitoring the progress of a country, there are many factors easy to obtain that are involved. The **problem** is the difficulty of considering all the controlling factors at a time and extracting valuable information and conclusions from them. So, it is important to have a visualization tool that gives a good picture of the progress of a country on the basis of major factors (e.g., economy, infant mortality, population growth, life expectancy, pollution, etc.) for **every user** who wants to explore how these factors have evolved throughout the years and find possible relationships between them. This **tool** will provide different ways of exploring the data, including interactive options like filtering, aggregation and sorting, helping the user look for associations between these factors such as how one factor might affect the others or how it has evolved during the years.

In a high level of abstraction, the tasks of the common user of this tool are located in the categories of **discovery** and **enjoyment**. In order to provide the best tool that helps this user to get a deep insight of the data set, we have estimated what kind of questions are the most probable and interesting to be asked. Among these **questions**, and taking into account the dataset explained below, we have focused on the following ones:

- ❏ How has the infant mortality rate evolved in Europe from 2000 to 2014?
- ❏ Is there any pattern between the infant mortality rate, CO2 emission level and population of a country? Does larger population mean higher level of pollution or infant mortality rate?
- ❏ Is there any relationship between the GDP expenditure of a country and its infant mortality rate or life expectancy? What about considering the amount of GDP invested in healthcare?
- ❏ Have there been cuts in the CO2 emissions from 2000 to 2014? What countries have polluted the most throughout the years?
- ❏ How has varied the GDP expenditure of each country during the years? Is there any tendency or feature? What about the percentage of GDP invested in healthcare?
- ❏ How has population evolved over years in Europe from 2000 to 2014 in each country?
- ❏ Is there any correlation between the factors of progress of a country?

The data set is focused on the **European countries** in order to compare them on the basis of different factors. More specifically, the data set is an aggregation of six different data sets all of which share the evolution of any factor for each country and throughout the years from 2000 to 2014. These **factors** are: CO2 emissions, GDP expenditure, life expectancy, infant mortality, population and size. As explained in the next section, some of these factors are represented by more than one variable.

2. Data and task abstractions

At this point, an abstraction of the data and tasks out of the application domain is required in order to solve the problems described in the previous step. The following sections describe the data types, data semantics and tasks abstractions that best fits the questions.

2.1 Data abstraction

2.1.1 Data types

The final data set is the result of combining six different data sets that provide information of different factors of the European countries from 2000 to 2014. By doing this combination, the result is a **time-varying** data set in the form of a **multidimensional table** with two keys: country and year. The data set is composed of 420 items and 14 attributes, 2 keys and 12 values as a whole. A brief explanation of each attribute can be found below.

The main two attributes of the data set are **country** and **year**. These variables are key attributes, or independent variables, so the rest of attributes will be dependent values of these variables. At the same time, country is a categorical attribute while year is a temporal, sequential and quantitative ordered attribute.

The types of each attribute of the data set can be found in the following table.

ATTRIBUTE	TYPE
Country name	Key attribute. Categorical. Character
Country code	Value attribute. Categorical. Character
Year	Key attribute. Ordered. Sequential. Numeric.
CO2 emission	Value attribute. Continuous. Numeric
CO2 emission per capita	Value attribute. Continuous. Numeric
CO2 emission level	Value attribute. Categorical. Character
GDP per capita	Value attribute. Continuous. Numeric
% of GDP for Healthcare	Value attribute. Continuous. Numeric
Life expectancy	Value attribute. Continuous. Numeric
Infant mortality rate	Value attribute. Continuous. Numeric

Infant deaths	Value attribute. Continuous. Numeric
Population	Value attribute. Continuous. Numeric
Area	Value attribute. Continuous. Numeric
Size	Value attribute. Categorical. Character

In the questions described above, we are dealing with various data types i-e items and attributes in order to answer above questions.

- In the first question, we are dealing with one quantitative variable i-e infant mortality. We will take into consideration all the 420 item sets containing these attributes.
- In the second question, there are three quantitative variables which are used i-e CO2 emission level, infant mortality and population.
- In the third question, we are dealing with five quantitative variables i-e, infant mortality, life expectancy, GDP per capita, population and health GDP..
- In the fourth question, we have Population which is quantitative variable.
- In fifth question, we have two main attributes i-e, country, year and one quantitative variable i-e CO2 emission which represents pollution level in a particular country.
- In the last question, we have one quantitative variable i-e GDP.

2.1.2 Data semantics

Despite the fact that almost all the attributes of the data set are easily understandable and the user can infer the meaning from the name of the variables, it is important to ensure that the real-world meaning is clear in order to interpret properly the dataset and future explorations.

The following list explains briefly each attribute of the data set:

- **Country**: The name that represents the real instance of the country.
- **Country code**: 3-letter code that represents the real instance of the country.
- **Year**: Value that represents a year from 2000 to 2014.
- **CO2 emission**: CO2 emissions measured by thousand metric tons (kt).
- **CO2 emission per capita**: CO2 emissions measured by thousand metric tons(kt) per capita (pc)
- **CO2 emission level**: Level of CO2 emissions (kt). There are five different levels: Low, Mid-Low, Middle, Mid-High and High.
- **GDP expenditure per capita**: Value that indicates the GDP expenditure per capita of a country.
- **% of GDP for healthcare**: Percentage of the GDP per capita of a country invested in healthcare.
- **Life expectancy**: Age that represents an estimation of the life expectancy of a country.

- **Infant mortality**: Rate that represents the number of deaths per 1000 live infants.
- **Infant deaths**: Number of infant deaths.
- **Population**: Number of inhabitants of a country.
- **Area**: Surface of a country represented by km2.
- **Size**: Level that represents the size of a country based on the area of the country. There are four levels: Small, Normal, Large and Very large.

2.2 Tasks abstractions

Since the demands of the common user is to explore the factors of each European country from 2000 to 2014, all of them take part in the categories of **discovery** and **enjoyment**. The next sections explain the decisions taken in order to answer the potential questions.

2.2.1 Stacked bar chart

Question. How has the infant mortality rate evolved in Europe from 2000 to 2014?

In order to solve the first question, we have decided to use a stacked bar chart (or stream graph by steps).

- **Why.** The objective is to provide an easy way to show the change in infant mortality rate per country over the years.

  Actions: By using this chart, the user should be able to explore the infant mortality rates per country, or as a whole, from 2000 to 2014 in order to find new trends or satisfy his/her curiosity about the evolution of this factor. This plot offers a suitable way of summarizing the records per country and year, and making easy to lookup known values. At the same time, this plot provides a whole visualization (part-to-whole relationship) of the total of infant mortality in Europe per year.

  Targets: In this case, the focus of the task remains in looking up for values and possible existent trends in the evolution of infant mortality over the years.

- **What.** A stacked bar chart (or stream graph by steps) to represent the infant mortality rates (quantitative value) of the multidimensional table based on the country and year (categorical key attributes)

- **How.** To build this chart we establish one bar per year which have one section per each country. The thickness of each section correspond to the value of infant mortality based on both key attributes. Also, we provide the possibility to filter or select the desired countries and years to be plotted, as explained in the next section, in order to focus the search and query on more specific items.

2.2.2 Tree map

Question. Is there any pattern between infant mortality rate, CO2 emission level and population of a country? Does larger population mean higher level of pollution or infant mortality rate?

We have decided to answer this question by using a tree map.

- **Why.** The objective is to provide an easy way to discover how the change in population over years is impacting infant mortality and pollution level.

  Actions: By using this chart, the user can discover correlations between the levels of pollution with the infant mortality rate, as well as with the population of a country. In addition, this chart provides an easy way to lookup specific countries to check their pollution levels based on the groups formed by the tree map. The user can also make comparisons between countries.

  Targets: In this case, the focus of the task remains in exploring the classification of each country regarding pollution rates and compare this level to the population and infant mortality rate over the years. Additionally, the user can find outliers based on the colors of the squares.

- **What.** A treemap to represent infant mortality, with change in pollution and population

- **How.** To build this chart we represent each country with one square which size indicates its population and the color indicates the infant mortality rate. User can see the pattern for a particular selected year. In addition, the chart is divided into four main groups and each group represents one of the CO2 emission levels. Each one of these groups contains the countries which have the corresponding CO2 emission level.

2.2.3 Parallel coordinates

Question. Is there any relationship between the GDP expenditure of a country and its infant mortality rate or life expectancy? What about considering the amount of GDP invested in healthcare?

We are answering this question by using a parallel coordinates plot.

- **Why:** The goal is to facilitate a visual way to find association between all GDPs, infant mortality, life expectancy and population of a country.

  Actions: The user can select the year in order to see the association between two GDPs, life expectancy, infant mortality and population of a country in that particular year.

  Targets: The principal target is to look for possible trends or associations between infant mortality, life expectancy and the variation of GDP per capita and % of GDP invested in

healthcare. Also, this associations have to be able to be related at the same time to the population of each country.

- **What.** A parallel coordinates chart that explains the association between these quantitative variables (GDP per capita, % of GDP for healthcare, infant mortality and life expectancy) with respect to main variables i-e country and year.

- **How:** To create this chart we establish one line per country in a specific year selected by the user. The color of each line represents the population of inhabitants in that particular country and the factors life expectancy, GDP per capita, % of GDP for healthcare and infant mortality are represented by dimensions (columns) of the parallel coordinate plots. The user can see how changes in the colors (population) influence the other dimensions. Also, the reordering of the columns facilitates the exploration of relationships between two dimensions (segments of the lines) as new information is presented with different orders.

2.2.4 Heat map

Question. Have there been cuts in the CO2 emissions from 2000 to 2014? What countries have polluted the most over the years?

We have decided to use heat map in order to answer this question.

- **Why.** The purpose is to explain the change in pollution levels of different European countries over years.

  Actions: By using this graph, the user should be able to interpret the change in CO2 emission level over the selected years and countries. This map gives a perfect representation of the pollution for all the European countries.

  Targets: Our focus remains in representing the changes in CO2 emissions over years for different countries of Europe in order to find outliers or discover trends in specific years.

- **What.** A heat map representing the CO2 emissions for different countries of Europe and years.

- **How:** To implement this chart we establish tiles as the combination of a year and country. The color of that tile represents the CO2 emission level. In this way, the user can discover changes in CO2 emission levels by observing variations in the color of tiles per column or row.

2.2.5 Stream graph

Question. How has varied the GDP expenditure of each country during the years? Is there any tendency or feature? What about considering the amount of GDP invested in healthcare?

We have decided to answer the first part of this question by using a stream graph.

- **Why.** The objective is to explore the change in GDP per capita of each country over years.

  Actions: By using this graph, the user should be able to visualize changes in GDP for different european countries from 2000 to 2014. Additionally, the user can select a particular set of countries and years in order to find trends.

  Targets: In this case, the focus of the task remains in looking up for values and possible existent trends in the evolution of GDP over the years.

- **What.** A stream graph by cardinals is used in the case that takes GDP, years and a country as input.

- **How:** To implement this chart we establish a layer (geometry shape) of a particular color per country along the axis. The shape of these layers are due to the GDP value. Stream curves go thicker if there summary of GDP for the corresponding year increase.

2.2.6 Bubble chart

Question. How has varied the GDP expenditure of each country during the years? Is there any tendency or feature? What about considering the amount of GDP invested in healthcare?

We have decided to answer the second part of this question by using a bubble chart,

- **Why.** The objective is to explore the changes in the % of GDP for healthcare per country and year and, at the same time, facilitate the search of relationships between these values and GDP per capita.

  Actions: By using this graph user should be able to visualize changes in the % of GDP for healthcare and the GDP per capita over the selected years for the selected european countries.

  Targets: In this case, the target is to explore GDP per capita and % of GDP for healthcare for different European countries over years looking for trends or extreme cases.

- **What.** A bubble chart will be used in this case which represent the two quantitative variables, GDP per capita and % of GDP for healthcare, along the countries and years.

- **How:** To build this chart we establish a bubble per country and year. All the bubbles regarding the same country are stacked in the same column and their order is given by the year. Each bubble is assigned two visual channels: the color of the bubble represents the GDP per capita and size of bubble represents the % of GDP for healthcare. The user can interpret the variations of the two types of GDP by comparing the color and size of the bubbles.

2.2.7 Bar chart

Question. How has population evolved in European countries from 2000 to 2014?

We are answering this question by using histogram.

- **Why.** In this case, the objective is to provide a simple way to visualize the evolution of population of one country at a time, over years.

  Actions: By using this graph user should be able to visualize rapidly any change in the population of the selected country over the selected years.

  Targets: The target is mainly focused on providing a general overview of one attribute (population) for the user to lookup extreme values or see its distribution.

- **What.** A histogram will be used in this case to represent the distribution of the population of one country over the years.

- **How:** To implement this chart we establish one bar per year which height indicates the number of population of a country. The user can filter a particular country to see the evolution of the population of that country and selected the desired years subject to exploration.

2.2.8 Principal Component Analysis & Correlation matrix

Question. Is there any correlation between the factors of progress of a country?

In order to solve the last question, we have decided to use a principal component analysis and a correlation matrix.

- **Why.** The objective is to provide an easy way to represent the possible correlations between the factors of progress of the countries.

  Actions: By using this algorithm and the correlation matrix, the user should be able to check easily the correlations between the factors by checking the angle of the arrows in the PCA chart or the size of the ellipsis in the correlation matrix.

  Targets: In this case, the focus of the task remains in explore or browse possible correlation between all the factors.

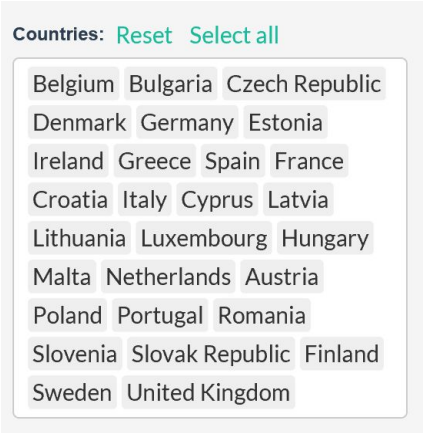
- **What.** An implementation of the Principal Component Analysis algorithm and a correlation matrix to show the correlations.

- **How.** To build this chart we use a correlation matrix in which each cell correspond to a relationship between two factors. These cells contains a ellipsis which size represent the degree of correlation.

3. Interaction and visual encoding

After having defined the data and tasks abstractions we can get into the interactivity and visual encoding. In order to increase the offerings of each chart and make it easier for the user to get into the data and to discover new trends and features, every chart is subject to options that the user can change for pleasure. In fact, there are two main types of interaction: through options in the sidebar of the app and directly with the plot. As we will discuss in the next sections, not all the charts are directly interactive but all of them are subject to options in the sidebar panel. In each section the specific options regarding the chart are explained.

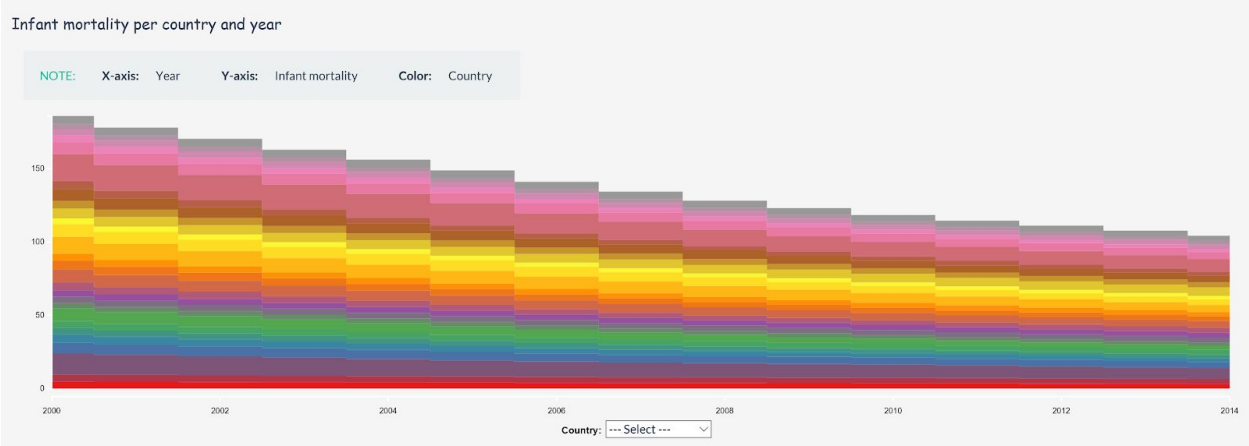
Regarding the general options, the user can filter and reorder the countries used by the charts to represent the data in order to facilitate comparisons and queries. The following figures show the options about countries.



The next sections describe the interactivity and visual encoding used for each task.

3.1 Stacked bar chart

Question. How has the infant mortality rate evolved in Europe from 2000 to 2014?



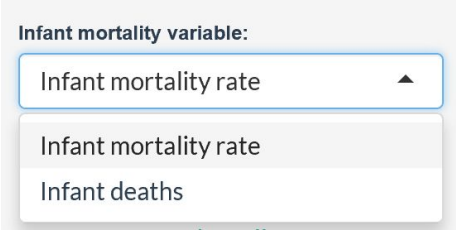
To answer the first question related to infant mortality per country and year, we decided to choose a stacked bar chart and build it based on the following design choices:

- **<u>Interaction</u>.** We decided to build this chart due to its suitable visual representation of values based on two key attributes, one of them referring to a temporal attribute.

  As explained at the beginning of this section, we opted to make it responsive to the selection of countries and years in order to increase the possibilities of exploration and **comparison** of different specific countries and years.

  In addition, this chart is directly interactive in two ways, achieving the same result. It is possible to select a country in the combobox under de bars in order to automatically highlight the sections of each bar corresponding to that country. Also, this effect can be achieved through a hover action with the mouse over any section of any bar. The highlighting is very useful to find a country in the chart and to see its evolution without having to change the selection of countries in the chart options.

  Lastly, it is also possible to choose which attribute to use in representation of infant mortality being able to choose between the rate of infant mortality per 1000 lives and the number of infant deaths.



- **<u>Visual encoding</u>.** In this chart there are two containment marks related to area and following the grouping criteria: small rectangles that represent each item of the table and indirectly a big rectangle, or bar, grouping the items based on the temporal attribute.

  Regarding the channels, we can differentiate between identity and magnitude channels.

  Identity channels. We are using a qualitative or categorical colormap to represent each value of the first key attribute, the country. We have tried to use a rainbow colormap to show the most variable colors to make the perception of each country easier. However, since the number of countries is 28 which exceeds considerably the maximum number of discriminable colors, we have taken the approach of reducing the number of sections by including the filtering options to allow the user focused on already known countries subject to exploration. In addition, we are using the bars as a way to represent the sum of values based on the temporal attribute and sequential ordering of these bars to represent the evolution from left to right.

  Magnitude channels. In this chart we are using the size of each rectangle as a magnitude channel to visually represent its value.

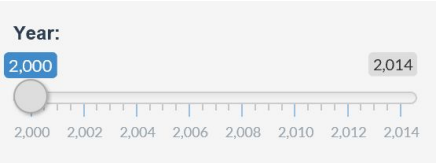
3.2 Tree map

Question. Is there any pattern between infant mortality rate, CO2 emission level and population of a country? Does larger population mean higher level of pollution or infant mortality rate?

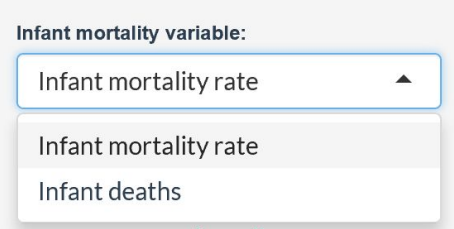
To answer the second question related to the relationship between infant mortality, CO2 emissions and population, we decided to use a treemap based on the following design choices:

- **<u>Interaction</u>.** We decided to build this chart due to its suitable visual representation of grouping quantitative values based on two categorical attributes: CO2 emission level and country.

  For this chart we also opted to make it responsive to the selection of countries in order to increase the possibilities of exploration and **comparison.** In this case, the treemap is built on top of a selected year that the user can modify at any time in the options panel.



Lastly, it is also possible to choose which attribute to use in representation of infant mortality being able to choose between the rate of infant mortality per 1000 lives and the number of infant deaths.



- **Visual encoding.** In this chart there are two containment marks related to area and following the grouping criteria: small rectangles that represent each item of the table previously filtered by the temporal attribute, and big rectangles grouping these items based on another categorical variable, CO2 emission level. By doing this, we manage to display hierarchical data by nesting those rectangles.

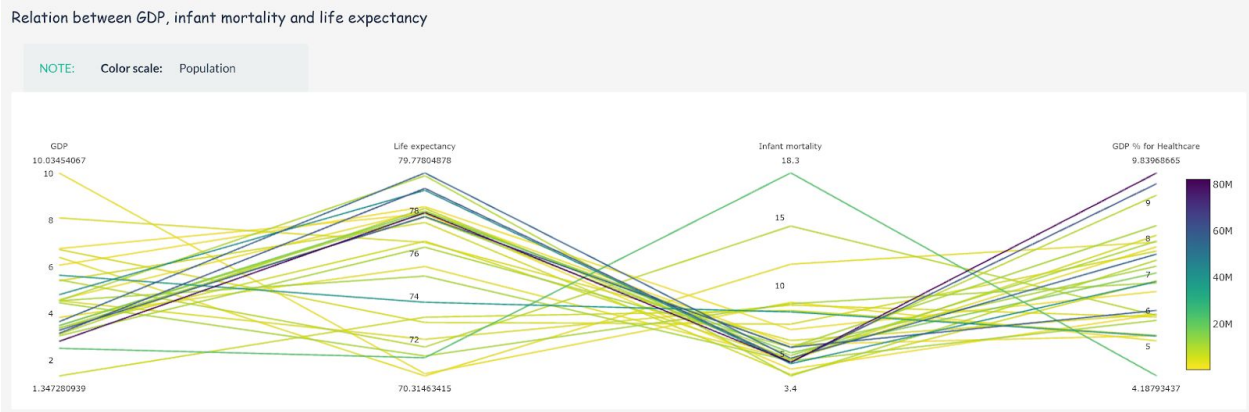
  Regarding the channels, we can differentiate between identity and magnitude channels.

  Identity channels. We are using rectangles to represent both the items subjected to the first categorical variable (country) and the groups formed based on the second categorical variable (CO2 emission level). In addition, we use ordering of grouping rectangles, positioning the biggest ones on the left side of the treemap.

  Magnitude channels. In this chart we are using two magnitude channels: the size of each rectangle as a way to represent the population of each country and a sequential colormap to represent the infant mortality factor.

3.3 Parallel coordinates

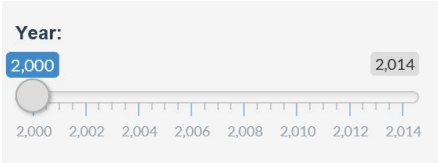
Question. Is there any relationship between the GDP expenditure of a country and its infant mortality rate or life expectancy? What about considering the amount of GDP invested in healthcare?



To answer the third question related to the relationship between infant mortality, GDP and life expectancy, we decided to use a parallel coordinates chart based on the following design choices:

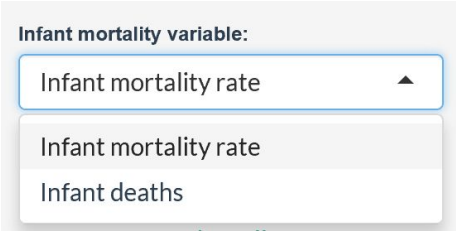
- **<u>Interaction</u>.** We decided to build this chart due to its scalability, providing an effective visual representation multiple attributes.

  Regarding the interaction through options, we also decided to make it responsive to the selection of countries in order to increase the possibilities of exploration and **comparison.** This chart is built on top of a selected year that the user can modify at any time in the options panel.



  Both options, countries and a specific year, lets us overcome the problem of limited number of attributes that can be represented without losing effectiveness (dozens).

  In addition, it is also possible to choose which attribute to use in representation of infant mortality being able to choose between the rate of infant mortality per 1000 lives and the number of infant deaths.



  Lastly, one of the things that make this chart very useful for this question is the possibility of **reordering** the axes to help the users with the exploration and discovery of new patterns since the changes in the layout can provide new information.

- **<u>Visual encoding</u>.** In this chart, connection marks with the form of a line are used to represent the relationships between the attributes as segments between the parallel axis. Due to its parallel

layout, a horizontal spatial position is used to separate the axes and a vertical spatial position is used to separate the different segments connecting the attributes.

Regarding the channels, we can differentiate between identity and magnitude channels.

Identity channels. We are using lines to represent each item of the table based on the country and each segment of the line represent the connection between the consecutive attributes at both ends. In addition, we are using vertical axis to represent different attributes, in this case GDP, infant mortality and life expectancy.

Magnitude channels. We are using a sequential colormap as a magnitude channel to represent the population of each country.

3.4 Heat map

Question. Have there been cuts in the CO2 emissions from 2000 to 2014? What countries have polluted the most over the years?

To answer the fourth question related to the evolution of CO2 emissions in the European countries from 2000 to 2014, we decided to use a heat map based on the following design choices:

- **Interaction.** We decided to use this chart due to its effective visual representation of a quantitative variable (CO2 emissions) based on two categorical key attributes (country and year) and the scalability of these categorical variables.

  For this chart, we also opted to make it responsive to the selection of countries and years in order to increase the possibilities of exploration and **comparison** of different specific countries and years. The following figure shows the years selection options.

- **Visual encoding.** In this chart there is a principal mark that represents the items of the table based on the two categorical attributes (country and year) and the value of a quantitative attribute (CO2

emission), with the form of a square or cell which form a general view of a table with all the items.

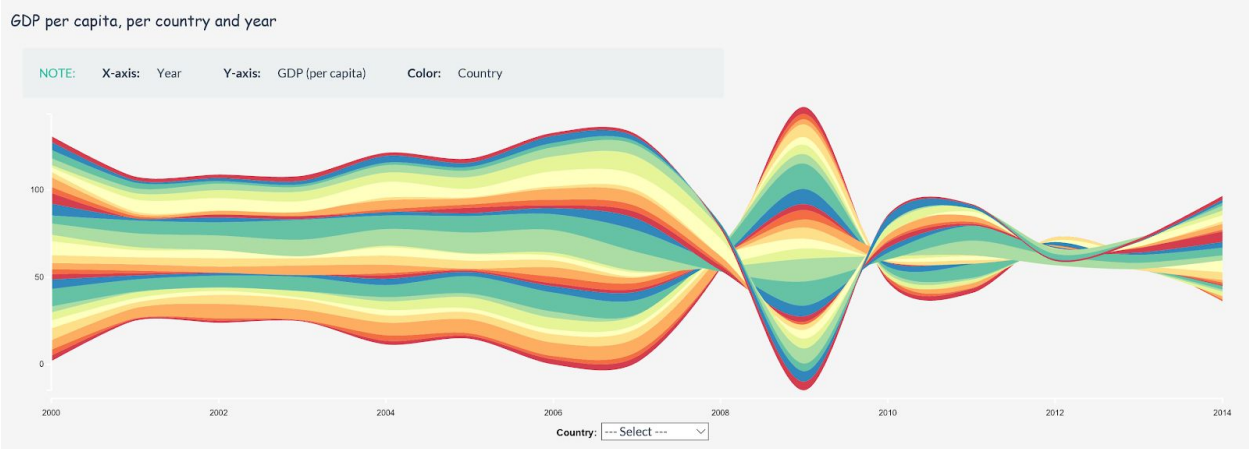
Regarding the channels, we can differentiate between identity and magnitude channels.

Identity channels. We are using squares to represent the items subjected to the categorical key attributes (country and year) and using ordering of the rows based on the temporal key attribute (year). This lets the user not only facilitate the discovery of trends and patterns over the years but also lookup the countries with higher or least levels of pollution.

Magnitude channels. The main magnitude channel used in this chart is a sequential colormap to represent the quantitative variable (CO2 emissions) in each square or item of the table.

3.5 Stream graph

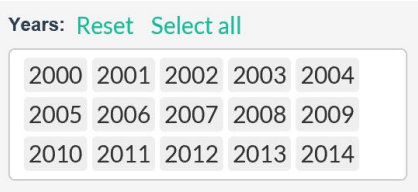
Question. How has varied the GDP expenditure of each country during the years? Is there any tendency or feature? What about considering the amount of GDP invested in healthcare?



To answer the first part of the fifth question related to the evolution of GDP per country and year, we decided to choose a stream graph based on the following design choices:

- **<u>Interaction</u>.** We decided to build this chart due to its complex but expressive representation of values based on two key attributes, one of them referring to an ordered key attribute, and a quantitative attribute. This chart lets us represent the evolution of the quantitative variable, the GDP per capita, over time in a more scalable way than using stacked bars.

  When building this chart, we opted to make it responsive to the selection of countries and years in order to increase the possibilities of exploration and **comparison** of different specific countries and years. The following figure shows the years selection option.



In addition, this chart is directly interactive in two ways, achieving the same result. It is possible to select a country in the combobox under de stream in order to automatically highlight the layer of the stream corresponding to that country. Also, this effect can be achieved through a hover action with the mouse over any layer of the stream. The highlighting is very useful to find a country in the chart and to see its evolution without having to change the selection of countries in the chart options.

- **Visual encoding.** In this chart the main mark refers to the geometry of the layers that represent each item of the table based on the key attribute (country) and the ordered key attribute (year), and takes into account the quantitative attribute (GDP per capita).

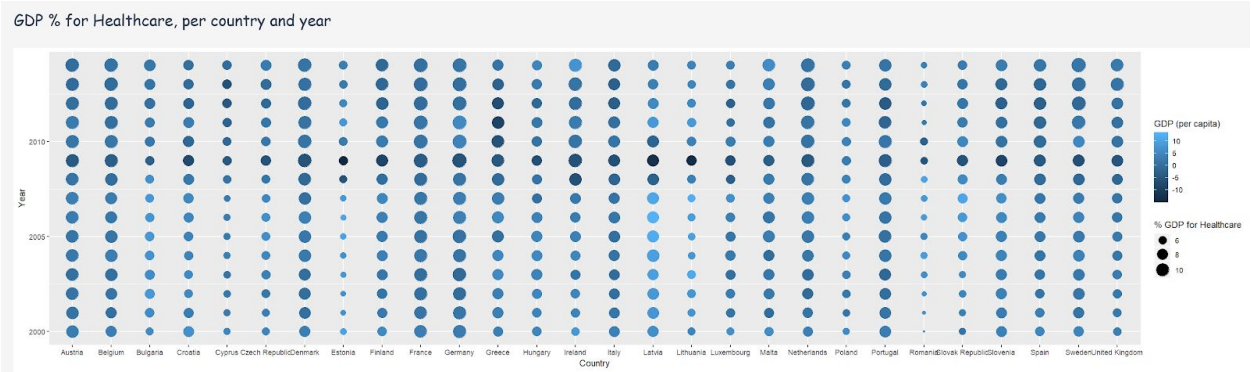
  Regarding the channels, we can differentiate between identity and magnitude channels.

  Identity channels. We are using a qualitative or categorical colormap to represent each value of the first key attribute, the country. We have tried to use a rainbow colormap to show the most variable colors to make the perception of each country easier. However, since the number of countries is 28 which exceeds considerably the maximum number of discriminable colors as happened before with the stacked bar chart, we have taken the same approach of reducing the number of sections by including the filtering options to allow the user focused on already known countries subject to exploration. In addition, we are using the shape of each layer and the whole stream as a way to represent better the tendency or evolution of the GDP based on the temporal attribute ordered from left to right.

  Magnitude channels. In this chart, the height of each layer is used as a magnitude channel to visually represent the value of the quantitative attribute in that item, as well as the height of the whole stream to represent the sum of these values.

3.6 Bubble chart

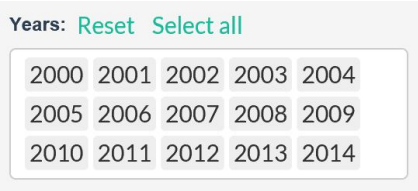
Question. How has varied the GDP expenditure of each country during the years? Is there any tendency or feature? What about considering the amount of GDP invested in healthcare?



To answer the second part of the fifth question related to the evolution of % of GDP for healthcare per country and year, we decided to choose a bubble chart based on the following design choices:

- **<u>Interaction</u>.** We have decided to build this chart due to the clear representation of values based on two key attributes, one of them referring to an ordered key attribute, and two quantitative attributes. This chart lets us represent the value of both quantitative values per item of the table in order to explore the evolution and discover any trend.

  For this chart, we also opted to make it responsive to the selection of countries and years in order to increase the possibilities of exploration and **comparison** of different specific countries and years. The following figure shows the years selection options.



- **<u>Visual encoding</u>.** In this chart there is a principal mark that represents the items of the table based on the two categorical attributes (country and year) and the values of two quantitative attributes (GDP and % of GDP for healthcare) at the same time, using bubbles and different channels over them.

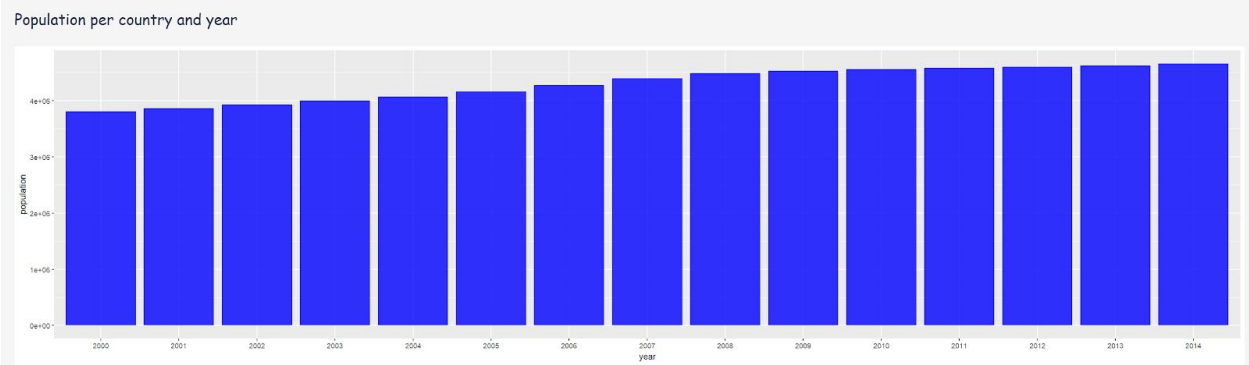
  Regarding the channels, we can differentiate between identity and magnitude channels.

  Identity channels. We are using columns of bubbles to represent the items subjected to the categorical key attribute (country) along the time scale established by the temporal key attribute, using bottom-up ordering of the rows. This lets the user not only facilitate the discovery of trends and patterns over the years but also lookup the countries with higher or least GDP or % of GDP for healthcare.

  Magnitude channels. used in this chart. The first one stands for the use of a sequential colormap over the bubbles to represent the GDP per capita. The second one is the size of each bubble which represents the % of GDP invested in healthcare in that item.

3.7 Bar chart

Question. How has population evolved in European countries from 2000 to 2014 in each country?



To answer the sixth question related to the evolution of the population per country from 2000 to 2014, we decided to choose a bar chart based on the following design choices:

- **<u>Interaction</u>.** We have decided to build this chart due to the simple representation of items based on a categorical key attribute (year) and a quantitative attribute (population). Since we wanted to fix the values to a specific categorical key attribute (country), we apply filtering of the items of the table based on the value selected by the user in the options panel, and we treat the categorical key attribute year as an ordered attribute printed in ascendent order from left to right. Additionally, for this chart we also opted to make it responsive to the selection of years. The following figures show the combobox for selecting the country and the years multiselect.

  By using changing these options we enforce the utility of the bar chart to compare values and lookup specific ones.

- **<u>Visual encoding</u>.** In this chart there is a principal mark that represents the items of the table based on an categorical key attribute (year) and the values of a quantitative attribute (population), with the form of a rectangle or bar on top of the corresponding categorical key attribute (year).

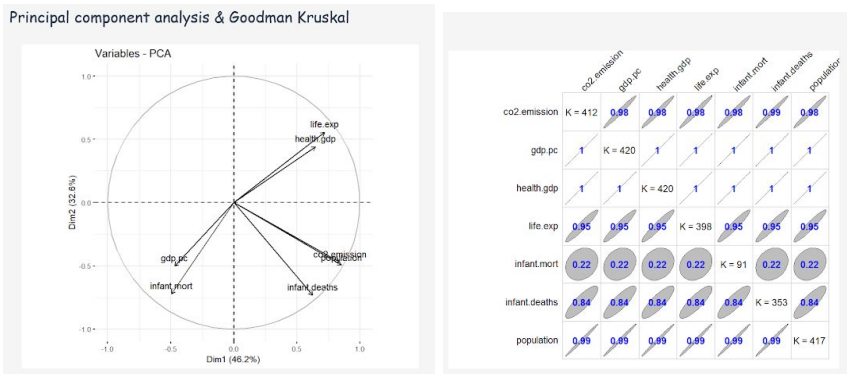
  Regarding the channels, we can differentiate between identity and magnitude channels.

Identity channels. The main identity channel is related to the horizontal spatial position of the bar which indicates the corresponding value of the ordered key attribute (year).

Magnitude channels. The magnitude channel in this chart refers to the height of each rectangle or bar which represents the value of the population of the selected country in the year indicated in the axis.

3.8 Principal Component Analysis & Correlation matrix

Question. Is there any correlation between the factors of progress of a country?



To answer the last question related to the possible existence of correlations between the factor of progress of a country, we decided to implement a Principal Component Analysis and plot the results using a --- and a correlation matrix based on the following design choices:

- **<u>Interaction</u>.** We decided to implement these charts due to the easy interpretation the user can make over the correlation of the different factors of progress of a country.

  When implementing the charts, we opted to make them responsive to the selection of countries and years in order to increase the possibilities of exploration and **comparison** of the values across different sets of countries and years. The following figure shows the years selection option.

- **<u>Visual encoding.</u>** In the first chart, we are using a line mark to represent each vector which, at the same time, show the impact of the corresponding factor in the dataset. In the second plot, the mark is an ellipsis per cell that visually explains the correlation between the corresponding two variables based on the position of the cell (row and column).

  Regarding the channels, we can differentiate between identity and magnitude channels.

Identity channels. In the first plot, the channel to identify each factor is an arrow with the name of the factor indicated in a label in the furthest position to the center of the circle. Regarding the second plot, the position of each cell is what identify to which attributes the value represented inside stands for.

Magnitude channels. In the first plot, the magnitude channel is the angle of the arrow with regard to the center point of the circle. This lets the user to approximately quantify the correlation between factors by comparing the angle and observing the position and proximity of the arrows. Concerning the second plot, the magnitude channel refers to the size of each ellipse which represents the value of correlation between the factors corresponding to the cell.

4. Algorithm implementation

Filter:

We are using selectizeInput feature in UI function by which we can select any number of countries we want to visualize, any number of years or if want to take any value (i-e infant mortality or infant deaths) as input. Inside server function we are using if conditions that observe and filter the values entered from selectizeInput by the user and are used inside the plot function as input. We are also using other action buttons, for example select all or reset. We also observe them inside the server function before passing the input to plot function. The libraries we are using here are **shiny**, **dplyr** and **reshape2**

4.1 Stacked bar chart

In order to plot this graph we are using the ggplot function and the ggplot2 library. ggplot2 is a system for declaratively creating graphics, based on the Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. We are using the geom_bar feature of ggplot in order to plot the stacked bar chart. We are using a categorical colormap for countries in order to give more spatial look.

4.2 Tree map

In order to plot this graph, we are using the tree map plot function and it uses the treemap library. Tree branches are used as rectangles and sub-branches are shown by smaller rectangles. We are using CO2 emission levels and countries as index in the tree map function. It has divided our tree into four branches representing CO2 emission level and we are representing the countries as sub-branches. Population is represented by the size of the sub-branch in the plot and the variation of color represents the different values of infant mortality.

4.3 Parallel coordinates

It uses the plotly function that uses the plotly library. It provides abstractions for doing common things (e.g. mapping data values to fill colors (via color) or creating animations. Inside the server function, we are using the plot function with **parcoords** type in order to plot parallel coordinates plot. We are using its color variable in order to show a visual representation of the population of different countries and we are using dimensions to visualize other quantitative variables i-e GDP per capita, % of GDP for healthcare, life expectancy and infant mortality.

4.4 Heat map

We are using the ggplot function in order to plot this chart that uses the ggplot2 library. Also, we are using the geom_tile utility to represent the pollution values of different european countries over years as color variation of the tiles. Finally, it is important to mention that we are using the square root of CO2 emission per capita to normalize the values so as to better appreciate the color variations among tiles. The countries and the years are the x and y variables of aesthetics, respectively.

4.5 Stream graph

It uses the stream graph function and the library which we are using in this case is streamgraph library (i-e **by using devtools library and installing the hrbrmstr/streamgraph github package**). The size of each individual stream shape is proportional to the values in each category. The axis that a stream graph flows parallel to, is used for the timescale which is years in our case. The value variable which we are using in our stream plot is the GDP per capita. Additionally, we are using a categorical colormap to draw each layer that represent the countries. We are scaling it with respect to years from 2000 to 2014 but this can be configured by the user.

4.6 Bubble chart

We are using the ggplot function in order to draw the bubble plot which uses the ggplot2 library. The Geom_point feature is used in the ggplot function to visualize the bubble plot which is also called scatter plot. In our case, we are using the color channel of each bubble to represent the GDP per capita of the countries and the size channel to represent the % of GDP for healthcare. The countries and the years are the x and y variables of aesthetics, respectively.

4.7 Bar chart

We are using the ggplot function in order to draw the histogram plot which uses the ggplot2 library. The Geom_histogram feature is used in the ggplot function to visualize the histogram. We are using the length of the bars to represent the values of the population of a country. The population and the years are the x and y variables of aesthetics, respectively.

4.8 Principal Component Analysis & Correlation matrix

In order to implement a principal component analysis plot and a correlation matrix, we are using the prcomp and GKtauDataframe functions. We are drawing the principal component analysis plot by using the fviz_pca_var function. Also, we are using the eigenvalue interpretation of our principal component analysis by using the fviz_eig function. The libraries which we are using in order to plot this analysis are GoodmanKruskal, factoMiner, factoextra, corrplot and cluster. Principal component analysis helps to reduce the dimensionality of a data set while retaining as much as possible of the variation present in the

data. Moreover, the correlation matrix which we achieve by using the GoodmanKruskal function gives a very good representation of the correlations between factors. Finally, we have set the value of corrColors variable as blue in correlation matrix function to make the ellipsis more appealing.