Available families Income and how it is spent in goods and services, through the influence of the inflation rate

Miguel Veloso, up202202463
Paulo Portela, up202200871
Tomás Rodrigues, up202202467

Abstract

This work aims at creating one, or more, non-trivial carefully designed plots, using the library ggplot2 in Rstudio. The main goal of this project is to answer the following questions: "The evolution of the purchasing power of families through the analysis of their available income and the inflation rate" and "How the ratios of the families income changes in goods and services, in periods of crisis ?".

So as to do the analysis of the data, the information for the respective questions was collected from the portuguese website "Pordata", from where two datasets were collected, an integration of datasets was carried out in order to have a more complete "final" dataset with more information that would allow us to answer the first question in a more complete and concise way. Regarding the second plot, the information was also collected from "Pordata", only one dataset this time, and the rates were calculated and these same rates were used in the visualisation and in the analysis and interpretation of the results.

As our final output, it's expected to have a line chart (time series) - that shows how quantitative values have changed over time for different categorical items - with two geom lines with a shadow between the lines that would indicate whether the families gained purchasing power or not, while showing the values for the inflation during the years, highlighting sensitive economical periods, more specifically the financial and COVID-19 crisis. The second one it's expected to be a slope graph - that shows how quantitative values have changed over two points in time for different category items - to see the rations of the families income spent in the periods of the economical crisis (2007 - 2013 and COVID-19).

Introduction

This visualisation design is part of the course of Data Visualisation and Preparation from FEUP, Faculty of Engineering - University of Porto, and aims at formulating a question (or a set of questions) that can be answered with one or more datasets of our choice and produce visualisations designs that allow the discovering of the answer to the questions, by using **ggplot2** for this task.

This project follows the Crisp-DM methodology where the raw data was transformed into useful data that can tell us more about a subject and allow us to make visualisations to analyse and, possibly, make predictions of future outcomes.

The raw data (original datasets) was previously downloaded from Pordata website and explored using data science tools (r language and rstudio). The datasets are developed by Pordata, which is a database of certified statistics of Portugal, its municipalities and Europe, that addresses various themes of the society.

Using the data science tools mentioned above, after the data analysing and understanding, the data preparation was carried out, filtered the variables of interest, checked for null/missing values and errors, for the dates of study, 1996 to 2021 for the first visualisation and the 2007 to 2020, with only a few periods, for the second visualisation (for the crisis periods).
After the data manipulation processing was finished it was possible to notice the creation of new columns ("Yearlydiff" which is the difference in the income value when compared with the value from last year and the "var" which is the variation of the income value when compared with the last year), for the first visualisation, that added value for our research and aim. For the second visualisation, the information provided was enough to proceed with the study, analysis and to answer the question previously made. For both scenarios, the main goals were to plot trends and intervals over time, to see the evolution/ increase and on the other hand the decrease / regression for variables in study.
As it will be possible to verify further on this report, the choice of the first visualisation fell on a **Line Chart** and second on a **Slope Graph**, both recommended to show how quantitative values have changed over time / two points for different categorical / category items, respectively.

Methodology

The methodology used in the development of the project consists of the **CRoss Industry Standard Process for Data Mining (CRISP-DM)** methodology that is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organise the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

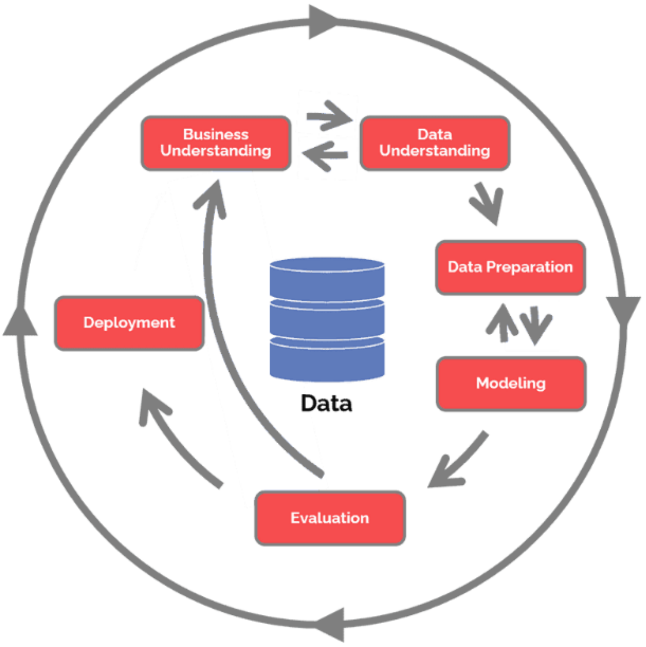


Figure1 - Crisp Tasks

1) Business Understanding

As we all know, inflation is at a high in the year of 2022, and this problem/subject sparked our interest in analysing its evolution throughout the years while comparing it with the available income of the families, and how the ratios of their gross available income are spent in certain goods and services, and how they change.
Thus, our focus was on macroeconomics variables to study the **research questions** which are: "The evolution of the purchasing power of families through the analysis of their available income and the inflation rate" and "How the ratios of the families income changes in goods and services, in periods of crisis ?". The objective was to understand/analyse if throughout the years the families were gaining or losing purchasing power, and how economical cycles changed the way families apply their available income in the various different goods and services.
In order to **assess the situation**, the variables of the study have been carefully checked to see if they were the most appropriate ones. In fact, the macroeconomics variables made the approach of this project slightly harder because it was not possible to know with certainty if the variables were the right ones to be used, especially the gross available income by families.
In terms of the resources available, several datasets from different resources/sites were analysed and viewed, having the choice - of the 3 datasets used - ended up falling on the Pordata website, due to the wide range of topics covered and how easy it was to search for the variables looked for.
In terms of **data mining goals**, the aim was studying the evolution of both variables (Inflation rate and annual income for the families) through the visualisation of **geom lines**, with a **shadow** in between them with an appropriate colour that would indicate whether the families gained purchasing power or not. In addition a **slope graph was raised** to show which goods or services were more sensitive to changes in the ratios when the purchasing power increases or decreases.
According to the knowledge in economics, a predefined notion was taken into account, and confirmed, that in periods of crises the families lose purchasing power, and when that happens a higher ratio of their income is spent in first necessity goods, and the ratio of luxurious goods would decrease, and that is exactly what we wanted to test with these two graphs.

In summary this phase focuses on understanding the project objectives and requirements, convert this knowledge into a data mining problem / questions and move on to the next phase, the data understanding represented below.

2) Data understanding

During the datasets analysis process, in "**Pordata**", and the high exposure of multiple topics within multiple areas, made the choice of the main topic of this project hard to make.
The choice fell on macroeconomics variables, which were the inflation rate, the gross income available by families, and the gross family income spent per type of goods/services.
Therefore, the chosen datasets are the following:

First Visualisation

➢ **Inflation Rate (Growth Rate - Consumer Price Index): total and individual consumption by purpose**

○ https://www.pordata.pt/en/portugal/inflation+rate+(growth+rate+++consumer+price+index)+total+and+individual+consumption+by+purpose-2315 ;

➢ **Gross disposable income of households (2016)**

○ https://www.pordata.pt/en/portugal/gross+disposable+income+of+households+(2016)-2407 ;

Second Visualisation

➢ **Household final consumption in economic territory: total and by goods and services type (2016)**

○ https://www.pordata.pt/en/portugal/household+final+consumption+in+economic+territory+total+and+by+goods+and+services+type+(2016)-2416 ;

The data provided by the website was either numerical and discrete or numerical and continuous, and it was imported through an excel file provided by the website. It was possible to verify/ observe the following for the variables under study:

➢ **Inflation Rate -** the inflation rate had already been calculated, numerical continuous, throughout the years and for the same reason no data manipulation was necessary to perform;

➢ **Gross available income per family** - the gross available income was in numerical value and for that reason more columns were created to be possible to obtain the variation of the gross income, which is a continuous numerical value;

➢ **Income spent per category of goods/services** - the income spent had a discrete numerical value which was converted to a continuous numerical value;

➢ **Ratios of the family income spent per goods and services** - the ratio had to be calculated, because it's only available the numerical discrete values for that, what was converted to a percentage of the occupation in their families income;

In the end, a **data quality** verification was proceeded, based on the following dimensions:

➢ Accurate - measured the degree to which data reflects whats is measuring and the relationship between the data and the real word;

➢ Complete - measured the degree to which we have recorded all relevant properties;

➢ Consistent - measured the agreement between data (also between the different datasets selected);

➢ Timeliness - measured if the data is up to date;

➢ Interpretable - measured how easily the data can be understood;

➢ Trusted - measured how trustable the data are to the users;

In brief, after an initial data collection and proceeding with activities to get familiar with the data, the data quality was identified to discover first insights into the data and detect interesting subsets and variables. After this phase, the next phase is the data preparation addressed then.

3) Data Preparation

The group wanted to create something that reflected the current situation in Portugal and since the beginning that the inflation rate was an option. After checking a few statistical websites it seemed a good idea to relate the inflation rate with the families yield (yearly) to have an idea about the evolution of the families purchasing power through the years.

In terms of data preparation, there were some things that had to be done. Then, it was decided to do a second visualisation to find out the evolution of the goods consumption by the families in 4 different periods (2 pré crisis and 2 post crisis).

Due to the good quality of data, the data cleaning was a short task, no missing values were found.

Regarding the first visualisation, the two datasets had several columns that were not useful, so the first task was the "**feature selection**". By the end of the features analysis only two features were chosen from each dataset - from the inflation rate "*Year"* and "*Inflation rate"* were selected and from the family yield the "*Year"* and "*Total yield"*.

At this stage, all data from the datasets were filtered, but the unit measure of each other (meaning inflation rate and total yield) was different, one was using percentage and the other integer, so the new task was to use the same unit measure for both and the salaries variation could work well enough. To have the salary variation a new column with the salary difference from each year was created (only the year of 1995 does not have a difference, it is the starting year). Then from that yearly salary difference a new column with the variation was created using the following formula:

$$\frac{second\ year\ salary - first\ year\ salary}{first\ year\ difference} * 100$$

It was all set to merge the two datasets and they were filtered by > Year 1995.

Regarding the second visualisation there was only one dataset with also cleaned data, so, once again this task was short. However, the "**feature selection**" here was a little bit more time consuming than the first visualisation, there were several goods with the respective quantity of consumption and it was necessary a deep analysis of each feature to understand which ones should be selected (which ones were more relevant). After the selection was also necessary to change the unit measure to be more understandable to the final user - in this case the user would have a better visualisation with percentual numbers. It was mentioned before that the idea of this visualisation was to compare 4 periods of time, so the last task was to filter only the years of 2007, 2013, 2019, 2020.

4) Modeling and 5) Evaluation

Modelling is the phase where several modelling techniques were selected and applied and their parameters were calibrated to optimal values and then, the model /or models that appear to have a high quality from a data analysis perspective were the ones selected. In the **Evaluation** part were considered if the selected models / graphs, respond in a clearly and correctly way to the research questions initially defined.
Both phases were developed at the same time, so they were put together in this report.

Below it´s possible to see what was done and which parameters / characteristics were taken into account in the creation and development of both **visualisations**.

First Plot - Explanation - Line chart

The decision to make a **Line chart** for the first plot was intuitive because it is the easiest way to compare the behaviour of two variables within a time series.

For the first plot, the base foundation is two **geom_lines**, one for the inflation rate and another one for the variation of the available gross income of families. Both lines have a distinct colour that helps identify each one of them. A vibrant red for the inflation rate, because red usually suggests something negative, and it had to be vibrant so as to distinguish from the red of the shadow. On the other hand, for the variation of the available income by families the chosen colour was a vibrant green. The approach was similar to the one mentioned before. Green colour because it is usually perceived as something positive, and vibrant because it needed to stand out from the shadow green which is between the lines. Another point worth mentioning is that a **thickness** to both **geom_lines** was added so that they could stand out a bit more from the rest.

All this evolution stretches from **1996 to 2021**, the worthy dates and intended period to study was about these years.

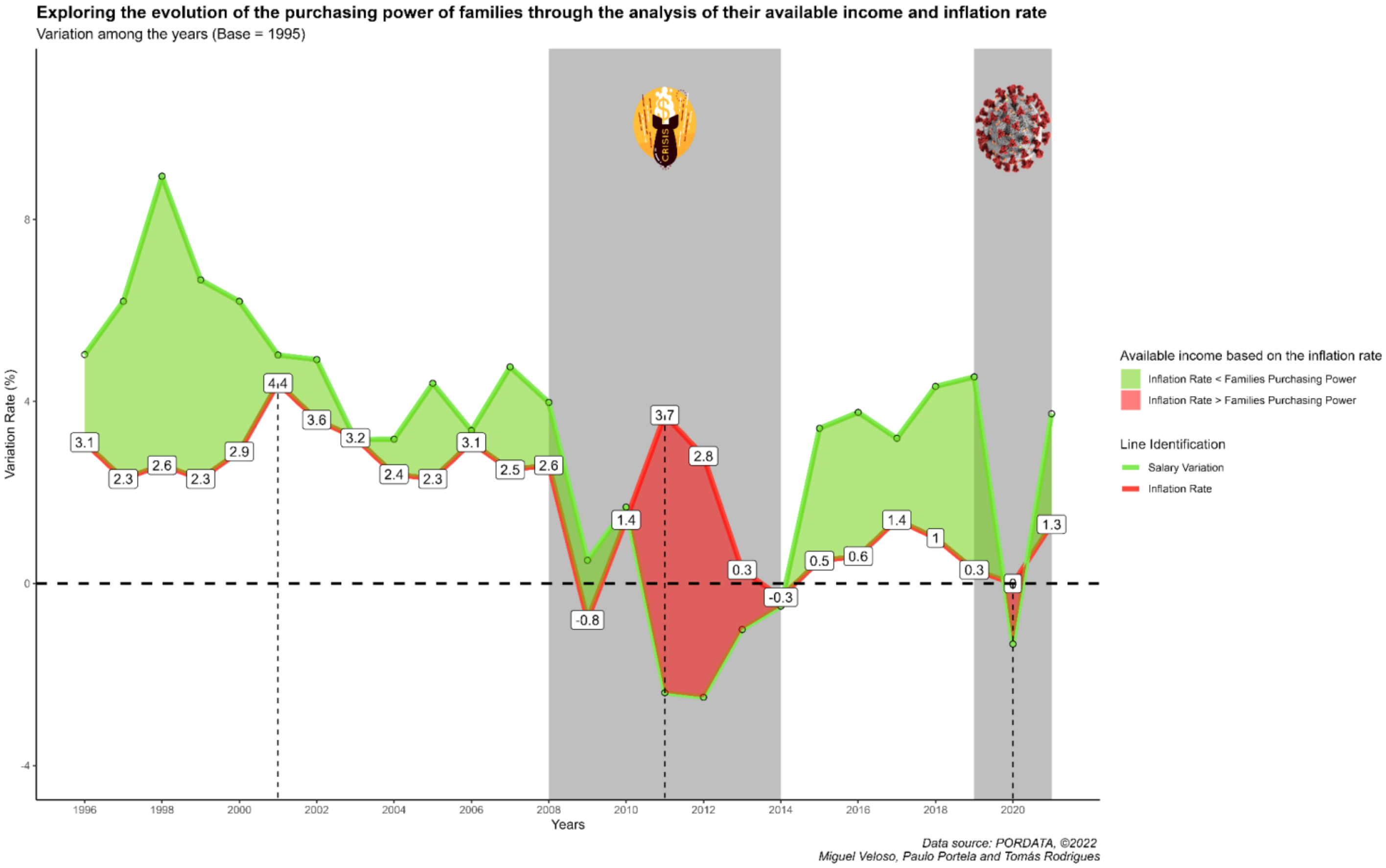
Another important aspect of the graph is the **shadow** in **between** both **lines**, since it´s possible to get immediate perception whether the families gained purchasing power or not. The **shadow** is basically the difference between the variation of the available income by families and the inflation rate. If the difference is positive, it means that the families gained purchasing power, if the difference is negative it means that the families lost purchasing power. Through the colour selection the viewer is able to tell right away whether the families gained purchasing power or lost, since **green** is usually perceived as something **positive** and **red** because it is usually perceived as something **negative**.

Something to outline is the **grey shadow** added in the plot, which outlines two periods of economical crisis, the global crisis of 2008 and the covid crises in 2020, which coincide with two periods where there was loss in purchasing power. These grey shadows went along with two suggesting **pictures of both crises**, so that the user with less overall culture could have a notion of what the grey shadows suggest.

In our plot, a was also added continuous numerical **labels** to the **inflation** in each year, since the study of the inflation was the main drive to create this plot, and will be worth displaying its evolution through the years.

The **title** is in bold and the font size is slightly bigger (14) than the font size in the **subtitle** (12), and that is because it is intended to highlight the main title of the project. At the bottom, for the **caption**, there is the data source along with the name of the authors. The font size was adequate (10) and it´s in italic. For all, the colour black was defined and the family is the same.

The **subtitles on the right** help the readers to understand which line matches the inflation and which one matches the variance of the available income. The other graphic subtitle is for the shadow in between lines.
The **vertical dashed** lines represent some inflation picks, the first line shows the point where inflation was at its high before the 2008 global crisis. The second line shows an inflation pick during the 2008 global crisis, and the last one the inflation during the covid crises. These lines are important for the graph understanding and perception.
In addition, a **horizontal line** representing the <u>value 0</u> was added in order to make it easy, for the readers, to see which values in the graph are above and below a variation rate of 0 percent, in black to be neutral.



Exploring the evolution of the purchasing power of families through the analysis of their available income and inflation rate
Variation among the years (Base = 1995)

Data source: PORDATA, ©2022
Miguel Veloso, Paulo Portela and Tomás Rodrigues

Figure2 - Line Graph for the exploration of the evolution of the purchasing power

Second Plot - Explanation - Slope chart

It was important to use the **Slope graph** due to its **simplicity in showing the ratios** spent in different goods and services, according to the family's Income. It is the most appropriate chart since the nature of the message is to compare values between certain years, in order to identify trends over time. It represents a quick and direct way to **show transitions** and changes over periods, in this case, with our ratios.

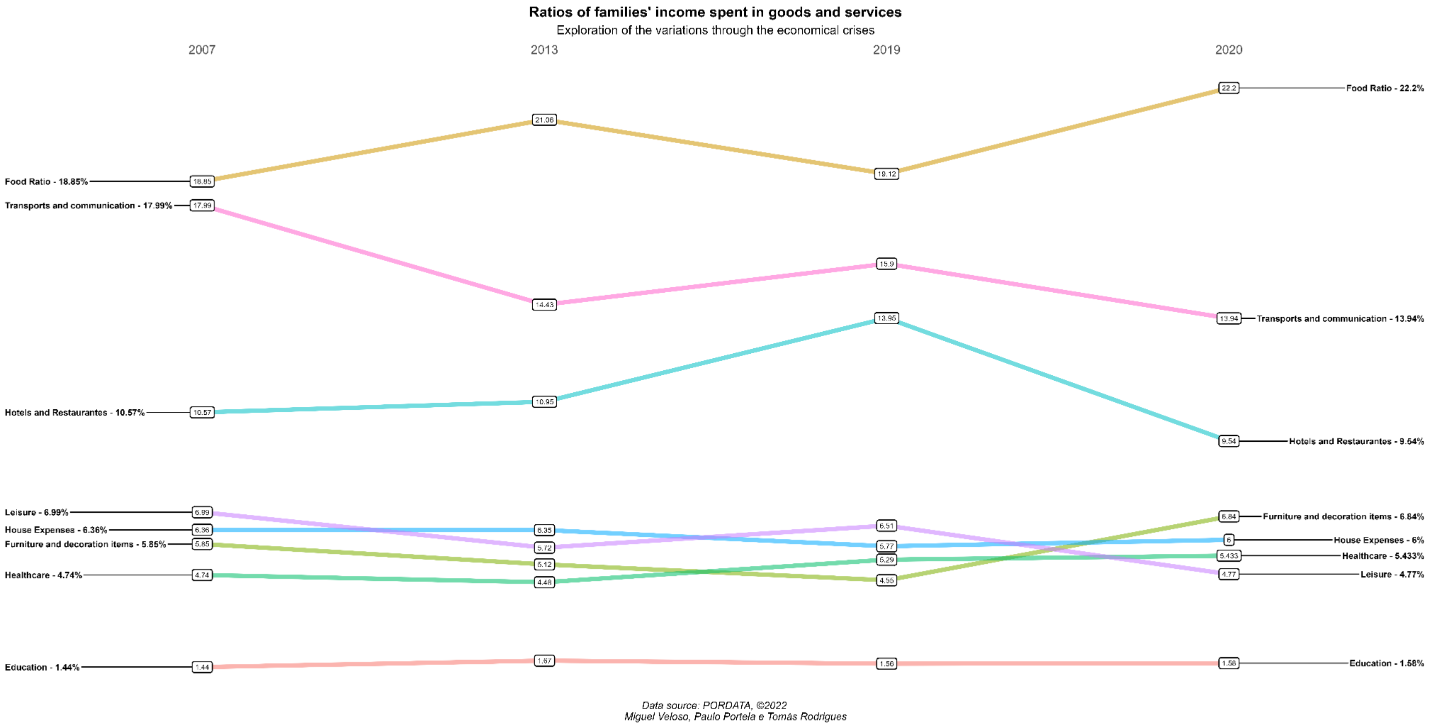
With this graph it is possible to extract a lot of information, from the evolution of the ratios through certain economical periods to the distribution of how families spend their income. What they prioritise, what they spend more money on and so on.

While trying to do the graph, there were many issues regarding the **colours** of the **lines**. The first chose red and green, red if the line was descending and green if the line was ascending. However, there was a problem regarding the overlaying of the lines, and it was decided to have **different colours** for each line so that the viewer/reader could perceive each line easily, even if there was line **overlaying**.

In the beginning, we tried different ways of having this type of graph, once the **text was overlaying** in the middle of the graph, because the text was identifying each line, for each year. To solve this issue, the text in between the graph was removed, and **added labels** with the values of each line. It ended up only having **text** in the very **beginning** of the graph, and at the very **end**. Again, there was some text overlaying which was fixed through the library ggrepel, more precisely with geom_text_repel. With this library it's possible to see and identify each line matching the corresponding text that identifies them.

To make this graph, four years were chosen (2007, 2013, 2019 and 2020). The year of 2007 it is when there is the outbreak of the global crisis. The year of 2013 is the end of the period of the economical crisis and there is recuperation afterwards. Then the year of 2019 is the year prior to the covid crisis, and it's when everything has recovered from the economic crisis. And finally the year of 2020 was the pandemic of COVID-19 crises which affected the way that people spend their income.

The **title** is in bold and the font size is slightly bigger (14) than the font size in the **subtitle** (12), and that is because it is intended to highlight the main title of the project. The hjust (hjust = 0.5) was used to set both in the centre. At the bottom, for the **caption**, there is the data source along with the name of the authors. The font size was adequate (10) and it´s in italic. For all, the colour black was defined and the family is the same.



Ratios of families' income spent in goods and services

Exploration of the variations through the economical crises

Data source: PORDATA, ©2022
Miguel Veloso, Paulo Portela e Tomás Rodrigues

Figure3 - Slope Graph for the ratios of families income spent in goods and services

Conclusions

When it comes down to what this report wants to answer, in other words, “Exploring the evolution of the purchasing power of families through the analysis of their available income and the inflation rate”, and “Analysing how the the ratios of the family income is spent in goods and services change” there are several findings worth mentioning.

Usually variations of the salaries are positive and are above the variations of the inflation rate, however, during periods of crisis or with economic tension the inflation rates tend to be higher (due to economic reasons/ mechanisms to prevent worst scenarios) and the salary variations are more likely to be close to 0. This indicates that, overall, the families gain purchasing power, with the exception of periods of crisis, where it is evident that they lose purchasing power, since the inflation rate is higher than the Income variations rate.

In terms of goods and services consumption, there can be a negative variation in luxurious services and items like hotels and restaurants in periods of crises, contrasting with first necessity goods like food and house bills which tend to increase. This means that when the families gain purchasing power, they are more likely to spend it on luxurious items, whereas in periods of crises, when families lose purchasing power, families are more likely to spend their income in first necessity goods and decrease the ratio of their income spent in none first necessity goods/services, which matches the first pre conception of economics.

Looking back and ahead, there were some improvements that could be interesting to do. Starting with more detailed data of the goods consumption to find out if there are some goods with unexpected growth in unexpected periods.

Having the information used in the first plot (namely the family yield) for more countries to compare variations and see which ones have (or not) better strategies to deal with crises and even which countries are more affected by tense periods.

To reach these conclusions the methodology chosen was essential and helped in terms of process organisation and management. Having a methodology allowed to set and split tasks and at the same time following a path with a final goal - answering the research questions.

This research was a good way to explore the immense capabilities and tunes of R language and its library ggplot2 - it allows all the members of the group to grow up as users of this tool and its language.