uc3m | Universidad Carlos III de Madrid

Master Degree in Big Data Analytics

2023-2024

Technological fundamentals in the big data world

Web Services and Data Retrieval

Gonzalo España-Heredia Llanza (100365421)

Fabio Scielzo Ortiz (100374708)

Teacher in charge: Jesús Carretero Perez

AVOID PLAGIARISM

The University uses the **Turnitin Feedback Studio** for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

CONTENTS

LIST OF FIGURES

1. EXTRACTING DATA FROM EUROSTAT AND IDESCAT

1.1. Introduction

1.1.1. Project Objectives

The primary objective of this project is to extract useful information from web pages. This task is executed via web services and web scraping techniques, leveraging Python's robust programming ecosystem. Specifically, the **requests** library is used for HTTP communication, while **BeautifulSoup** is employed for parsing HTML documents, thereby efficiently extracting data from web pages.

1.1.2. Webs

The following sections discuss the specific web sources selected for this project, namely Eurostat and idescat.cat. Each source has been chosen based on the richness of data available and relevance to the project's objectives.

Eurostat

Eurostat, the statistical office of the European Union, is a critical source of high-quality statistical information about Europe, which contains a great deal of data across various sectors and industries within the EU. Through its web service system, we have been able to extract relevant statistical information.

idescat.cat

Idescat.cat is the official statistics website of Catalonia, providing a comprehensive database of statistical resources about the region. It encompasses a wide range of subjects, including demographics, economics, environment, society, and many more, representing a crucial source of information for regional analysis. In this project we have obtained valuable data from idescat.cat, using both web services and web scraping techniques.

1.2. Extracting information from Eurostat

The code between the lines 13-69 of the file **Ws1.py** allows us to get the balance data for the Consumer confidence indicator for the EU countries as well as for the whole EU in a aggregate way. The code build automatically dictionaries with the most interesting information that is allocate in the provided URL, associated to Eurostat. Later, this information has been used for building insightful and informative plots, that will be displayed in the following sections.

1.2.1. Monthly Time Series of the indicator for the last five years (2017-2021)

The next plots represents the monthly time series of the selected indicator for the last five years (2017-2022), for an specific selection of the EU countries available in the extracted data from Eurostat, and also for the EU as an aggregation.

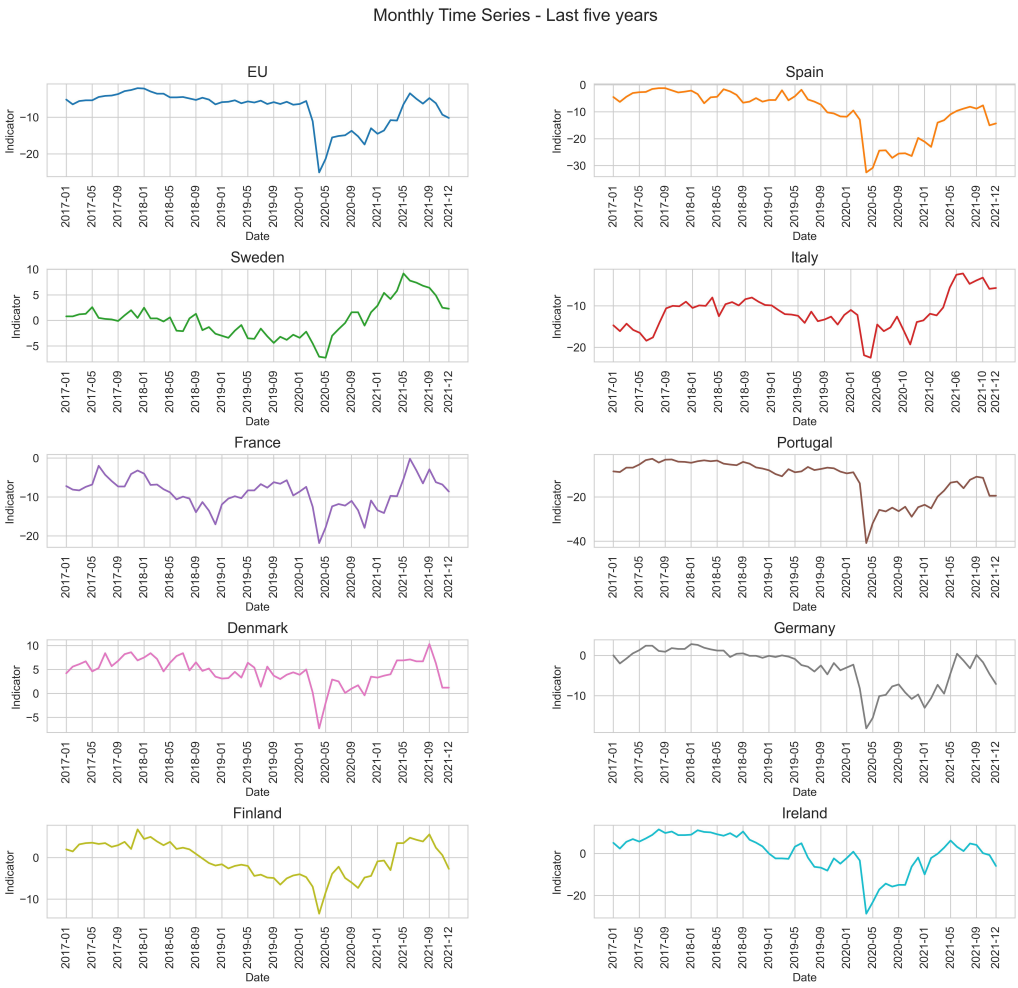

Fig. 1.1. Monthly Time Series - Last five years (2017-2021)

1.2.2. Barplot of the indicator for the last five years (2017-2021) and the months of 2022

These are barplots of the selected indicator for the last five years (2017-2021) and the months of 2022, for a specific selection of EU countries that appeared in the data retrieval form Eurostat, and for the whole EU as an aggregation, as we did before.

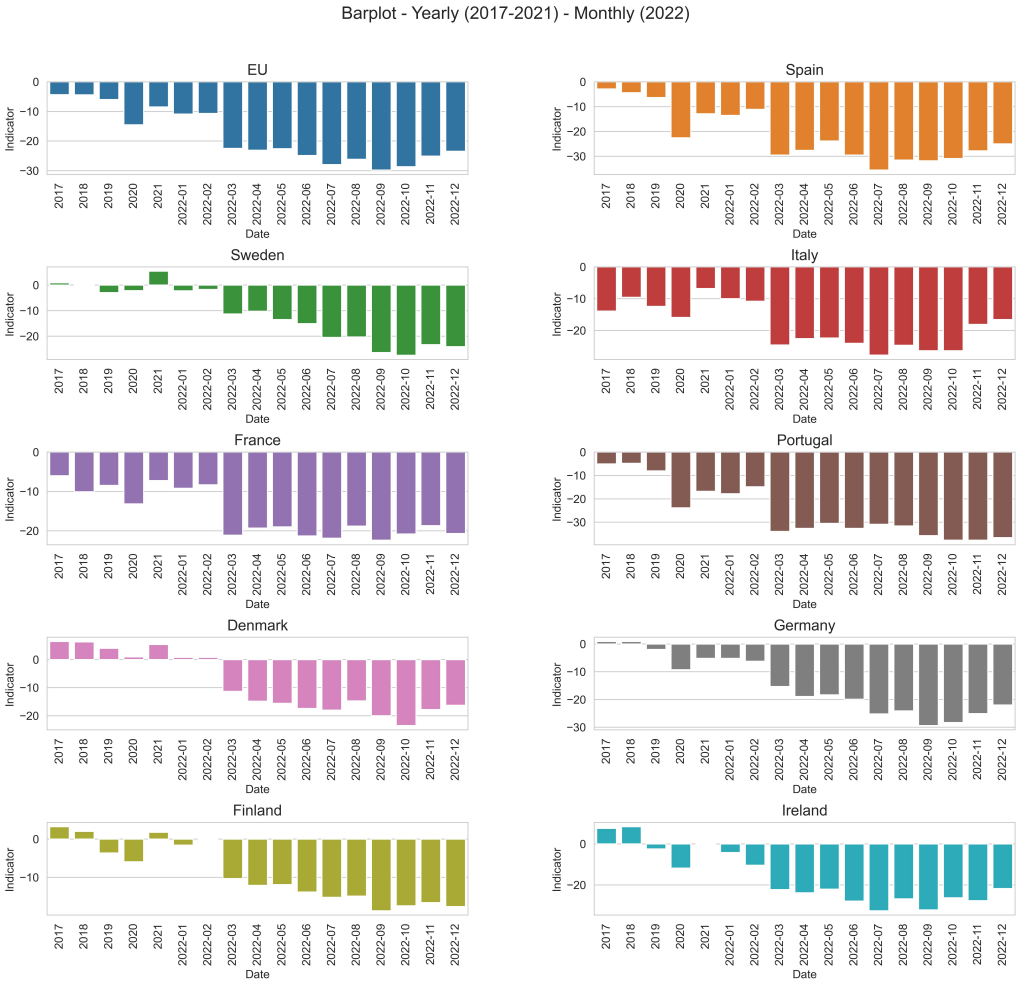

Fig. 1.2. Barplot - Yearly (2017-2021) - Monthly (2022)

1.3. Extracting information from Idescat

1.3.1. Statistical information about Maria's born in Catalonia during the last five years (2018-2022)

The code between the lines 14-49 of the file **Ws2.py** allows us to obtain interesting statistical information about Maria's born in Catalonia during the last five years (2018-2022). The code build automatically dictionaries with the most interesting information that is found in the provided URL, associated to Idescat. After that, this valuable information has been used to generate really informative plots.

The following set of plots contains statistical information about the child's born in Catalonia in the last five years (2018-2022) whose name is Maria.

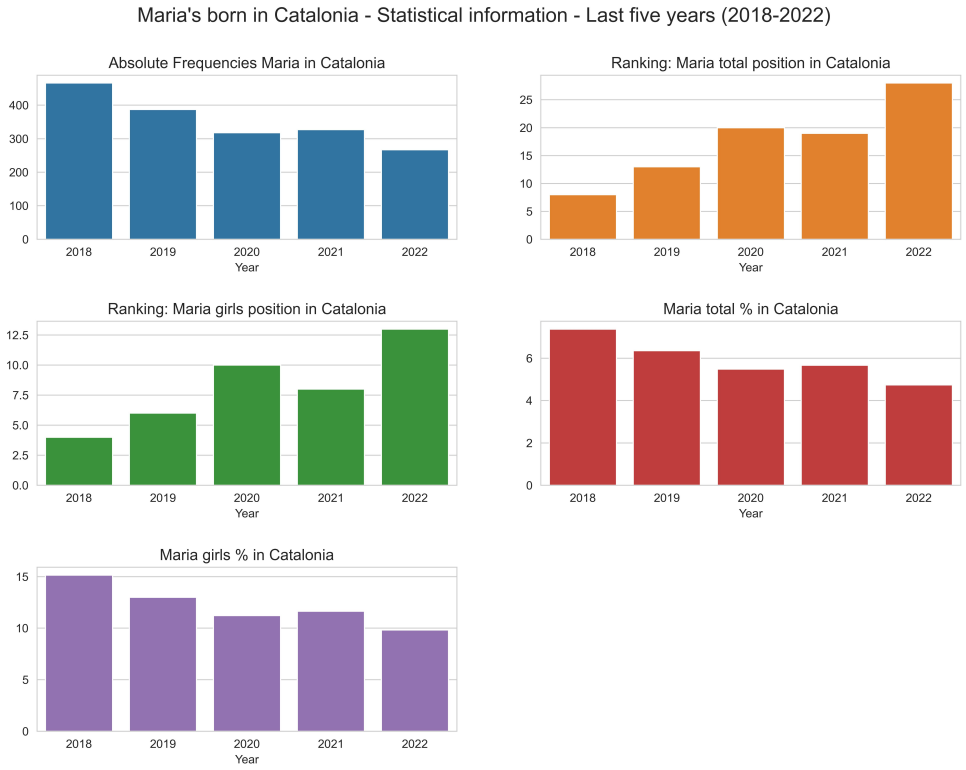

Fig. 1.3. Maria's born in Catalonia - Statistical information - Last five years (2018-2022)

1.3.2. Maria's analysis in Catalonia by city (comarca)

The code between lines 99-132 of the file **Ws2.py** works for obtaining insightful statistical information about Maria's born in Catalonia, differentiating by city (comarca), during the last five years (2018-2022). The code build automatically dictionaries with the most interesting information that is contained in the provided URL, associated with Idescat web service. Then we have created the following plots to expose the absolute frequencies

(counts) of Maria name in child's born in Catalonia, in the last five years, grouping by city (comarca). As relevant information extracted from this analysis, we can see that the Catalonia city (comarca) with more Maria's born in each one of those five years (2018-2022) is Barcelonès.

Fig. 1.4. Absolute Frequencies Maria by Catalonia cities (comarcas) - Last five years (2018-2022)

1.3.3. Gender info for the child's born in Catalonia in the last nine years (2014-2022)

In this case we have used web scraping techniques to extract information from Idescat about the child's born in Catalonia in the last nine years, focusing the analysis in the their gender (boy/girl). The code (lines 203-246 of **Ws2.py**) build a data-frame with the most useful information that we have located in the studied URL from Ideascat. Then, plots have been made to illustrate this information in a fancy way.

These plots show the relative frequencies (%) of the gender of the child's born in Catalonia in the last nine years (2014-2022).

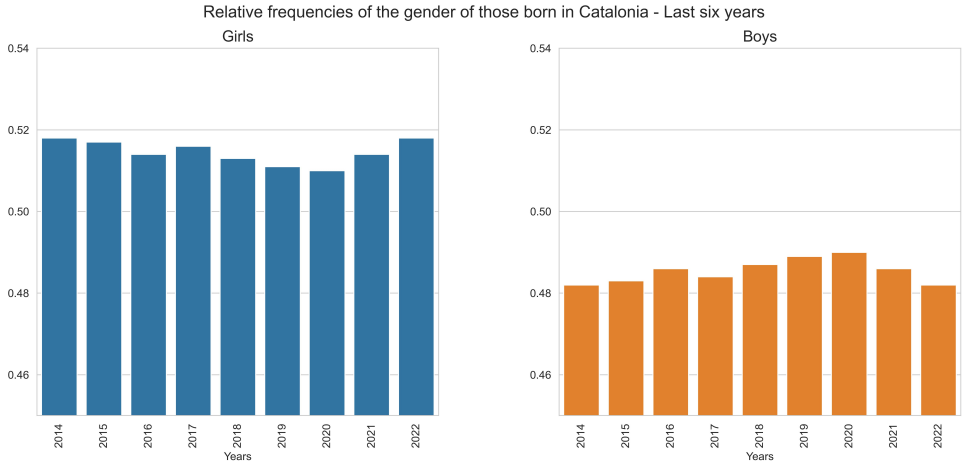

Fig. 1.5. Relative frequencies of the gender of child's born in Catalonia - Last nine years (2014-2022)

1.4. Conclusions

Along this project we have applied both web services and web scraping techniques through Python, so, it has been a great opportunity to learn more about data retrieval from the web, using these two approaches.

A relevant conclusion that we have drawn from the project is that, in some cases, when it is not possible to extract the desired information from a web page, either because it doesn't have a web services system, or because it has one, but the information is not accessible through it, Web scraping can help us solve the problem by allowing us to extract the HTML content of the page, which can be dissected for useful information.

Anyway, the project has been quite fruitful to show us the power of web services and web scraping for extracting data from web sites, which is a pretty important skill for any data scientist.