Completing this ETL project provided an opportunity to build a practical data science pipeline. Throughout the process, we encountered several challenges and gained a better understanding of working with real-world data formats, APIs, and visualizations.

One of the primary challenges we faced was formatting the numerical axes on the graphs in millions of euros. Although the data existed in large numbers, configuring the y-axis to properly reflect this in the visual output required precise formatting techniques and the use of custom tick formatters, which took trial and error to get right. Additionally, implementing comprehensive error handling—especially across multiple file formats and data types—was more difficult than we anticipated, as it required handling a range of potential issues and designing clear, specific feedback for users. Data visualization also posed difficulties in ensuring that the trends are accurately represented in a manner that is clear and easy to read. We had to ensure that the size of the plots, the details within, and the values of the axis were realistic and easy to analyze.

On the other hand, we found working with two different datasets surprisingly easy. Ingesting both a local CSV and a remote Kaggle dataset was pretty simple once we understood the structure. Implementing the file conversion feature between formats like CSV, JSON, and SQLite also turned out to be easier than expected.

This project emphasized the importance of modular, flexible pipelines in data science. A utility like this can be incredibly useful for future projects where quick ingestion, transformation, and export of datasets is required. It allows for scalable, repeatable workflows where data from APIs and local files can be quickly standardized and stored. Whether in academic research or industry applications, building a reusable ETL pipeline is an essential aspect of data science.