Team EidosReport DataViz

breve frase

[Dashboard](https://public.tableau.com/app/profile/antonio.sabbatella/viz/DataVisualization_17046596687520/FinalReport)w | Gennaio 2024

Alex Calabrese  
Kamila Fido  
Cristian Longoni  
Antonio Sabbatella  
Babak Khalilvandian  
Giuseppe Ciulla

Introduction

This report utilizes recent **economic data** to construct insightful visualizations that shed light on crucial aspects of the **Italian economy**. Leveraging multiple statistical sources, it graphs **trends** over time, comparisons across sectors, and **regional disparities** related to household incomes, expenditure patterns, employee compensation, and disposable earnings.

The graphical analysis aims to highlight noteworthy narratives in the data through effective data representation techniques that transform complex quantitative information into **visually intuitive** and easy-to-grasp graphs.

Index

[**Data Sources 2**](#_4lqp25cx7kth)

[**Data Cleaning (Pre-processing) 3**](#_ovawxjoi07ia)

[**Design Choices 5**](#_lzr2si3oshbv)

[Evolution of italian’s wages (Slide 3) 5](#_vs8n9u9uqpuo)

[Expenditure in Italy (Slide 4) 5](#_q0jru6iua1ig)

[Salary in different sectors (Slide 5) 5](#_uc6vxz88cfxq)

[Sectors with the Top 4 Wages (Slide 6) 6](#_kemz7iwlpoe3)

[Disposable Income by regions (Slide 7) 6](#_44fpyur8p3ze)

[Top 5 Regions (Slide 8) 6](#_ewsjbrm1kiu)

[**Challenges 7**](#_vi4iu9y3uz3x)

# 

# Data Sources

The economic situation of Italy and its population can be analyzed from different perspectives, such as the **income**, **expenditure**, and **compensation** of the households and the employees.

In this report, we will use **six sources** of data to explore these aspects and compare them with other regions.

The sources are:

1. [Compensation of employees breakdown by industry](https://esploradati.istat.it/databrowser/#/en/dw/categories/IT1,DATAWAREHOUSE,1.0/UP_ACC_NFAIST/IT1,94_1063_DF_DCCN_ISTITUZ_ANA1_3,1.0): data on the compensation of employees and its components by industry in Italy from 2016 to 2022. It includes wages and salaries, employers’ social contributions, and taxes and subsidies on production and imports.
2. [Average Wage Italy](https://www.statista.com/statistics/416213/average-annual-wages-italy-y-on-y-in-euros): average annual wage in Italy from 2000 to 2022 in euros.
3. [Italian Monetary Value Coeff](https://www.istat.it/it/archivio/286776): tables of coefficients for converting monetary values of different years from 1861 to 2022 into values of a reference year. It allows us to compare the purchasing power of money over time and adjust for inflation.
4. [Italian Household Expenditure by Category](https://esploradati.istat.it/databrowser/#/en/dw/categories/IT1,DATAWAREHOUSE,1.0/UP_ACC_TRIMES/IT1,163_24_DF_DCCN_QNA_4,1.0): distribution of household expenditure in Italy by category in 2020.
5. [Disposable Income of Household By Region](https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,DATAWAREHOUSE,1.0/UP_ACC_TERRIT/IT1,93_1095_DF_DCCN_ISTITUZ_TNA1_1,1.0): This link presents the disposable income of the households in the Italian regions from 2016 to 2022.
6. [Disposable Income Per Person](https://ec.europa.eu/eurostat/databrowser/view/nama_10r_2hhinc__custom_9198844/default/table?lang=en): disposable income per person across European countries in 2022.

# 

# 

# Data Cleaning (Pre-processing)

**Data preprocessing** is a crucial step in the data analysis pipeline. It involves transforming **raw data** into a clean format that is ready for **analysis** and **visualization**. This document outlines common preprocessing techniques used in the preprocessing of the utilized datasets in **Python**, using salary dataset as an example where extracting part of the utilized code.

Key shared steps across the datasets include:

**Data Import**: The raw data is imported from CSV and Excel files. This involves reading the files into dataframes, which are tabular data structures that make data manipulation easier.

* Example: df = pd.read\_csv("annual\_wage\_by\_industry.csv")
* Description: Importing CSV data into dataframe

**Data Selection**: Relevant columns and categories are selected for analysis. This step helps to focus on the data that is important for the specific analysis or model being developed.

* Example: selected\_df\_salaries = df[df["Breakdown by industry (NACE Rev2)"].isin(categories)]
* Description: Selecting subsets of data by industry category

**Data Reshaping**: The data is reshaped for analysis using operations like pivot and melt. Pivoting transforms the data from long format to wide format, making it easier to understand and analyze. Melting does the opposite, transforming data from wide format to long format, which is often necessary for certain types of data processing or modeling.

* Example: salaries = grp\_salaries.pivot(index='Breakdown by industry (NACE Rev2)', columns='TIME', values='Value')
* Description: Pivoting data from long to wide format

**Data Cleaning**: The data is cleaned and processed. This can involve filtering the data, grouping it by certain categories, and calculating derived metrics like averages. Cleaning is crucial to ensure the quality and reliability of the analysis or model.

* Example: df = df[df["Aggregate (millions of euro)"] == "domestic wages and salaries"]
* Description: Filtering to only relevant data

**Data Integration**: Different datasets are merged together. This is often necessary when data is collected from multiple sources or in different formats.

* Example: merged\_df = pd.concat([salaries, df\_avg\_annual\_wage], axis=0)
* Description: Merging wage data and average wage data

**Output**: The final, cleaned, and processed dataframes are outputted. These dataframes are now ready for further analysis or modeling.

* Example: df\_transposed.to\_csv('average\_annual\_wage\_by\_year.csv')
* Description: Output cleaned dataframe to CSV

# Design Choices

## Evolution of italian’s wages (Slide 3)

**Use of line charts**: The use of a line chart allows for a good visual representation of the trend in real and indexed wages over time. It is particularly useful here as the data are continuous and there is a clear relationship between the two variables (time and average wage).

**Use of two lines**: The use of two lines, one for real wages and one for indexed, Deinflationized wages, allows for a direct comparison of the two measures.

**Use of bar graph**: The use of bar graph to compare real and Deinflationized wages is used to represent the data clearly and concisely using a value calculated by the differences between the two values.   
**Use of color**: The use of red and green color in the bar chart represent the danger, anger of less purchasing power.

## Expenditure in Italy (Slide 4)

The main graph is a **treemap** showing expenditure average by category. We decided to use a treemap that uses nested rectangles to show **hierarchical data** and relative proportions.

The size and color of each rectangle represents the value of the category. This graph allows the user to quickly **compare the expenditure categories** and see which ones are more or less significant.

To the right, there’s a **donut chart** labeled “Goods Subdivision” showing non-durable goods in dark blue and semi/durable goods in light blue. We used it to show the percentage or **proportion** of a whole, usually with one or two categories.

The graph details the composition of “Goods”, showing that non-durable goods account for a larger share than semi/durable goods.

## Salary in different sectors (Slide 5)

The graph utilizes an **area chart** design to represent the **salaries in different sectors** over the years. Each sector is represented by a distinct color, and the areas are stacked on top of each other, allowing for a visual **comparison** of the **proportion** of the growth or decline in salaries across sectors over time.

The choice of different colors for each sector may be intended to create a **contrast** and a categorization of the data, making it easier to identify and compare the sectors.

## Sectors with the Top 4 Wages (Slide 6)

The graph employs a **line graph** design to represent the **wage growth** in four different sectors over time. Each sector is represented by a distinct color, making them stand out and easy to compare. The y-axis shows the **percentage increase** in wages, while the x-axis represents years from 2000 to 2022. The choice of a line graph may be intended to show the trends and changes in the wage growth over time, as well as the relationships and differences between the sectors.

The choice of different colors for each sector may be intended to create a contrast and a categorization of the data, making it easier to identify and compare the sectors.

## Disposable Income by regions (Slide 7)

We chose a **map graph** to show the **geographical distribution** of disposable income by regions in Italy. This helps to highlight the **regional disparities** and patterns in the data.

We used a color scale from red to green to indicate the ratio of “Population per Region” to “Salary per Employee” for each year. Red represents a low ratio, meaning low disposable income per capita, while green represents a high ratio, meaning high disposable income per capita.

This makes it easy to **compare the regions** and see the changes over time.

We added a legend to explain the meaning of the color shades and a slider to allow users to select different years from 2001 to 2021. This provides more details and interactivity to the graph.

## Top 5 Regions (Slide 8)

We chose a **graphical representation** showcasing the **disposable income trends** over the years for the **top 5 regions in Italy**.

**Graph on the Left**: It uses distinct colors for each region to clearly differentiate them and make it easy to follow their income trends over time. The Y-axis represents the disposable income, and the X-axis represents time from 2000 to 2020, providing a two-decade overview of financial progression.

**Map on the Right**: It provides a geographical context to where these top 5 regions are located in Italy, with different shades of green indicating their average disposable incomes. This visual aid helps in quickly grasping regional financial standings.

# Challenges

**Data Management**: The first challenge we encountered was related to data handling.

This involved several steps:

* **Data Discovery**: We had to locate and identify relevant data sources that could provide the necessary information for our project.
* **Data Integration**: Once the data was found, the next step was to integrate it from various sources, ensuring that it could interact and function together cohesively.
* **Data Filtering**: After integration, we had to filter the data to remove any irrelevant information, focusing only on the data that was pertinent to our objectives.
* **Data Cleaning**: The final step in this process was data cleaning, which involved checking for and correcting errors, inconsistencies, and inaccuracies in the data to ensure its quality and reliability.

**Visualization Selection**: The second challenge was choosing the appropriate **visualization** and **color scheme** for presenting our data. This was crucial as the right visualization can significantly enhance the comprehension and impact of our data. We had to consider various factors such as the type of data, the intended audience, and the key message we wanted to convey.



**Team Management**: The third challenge was managing a diverse team of six individuals, each with different backgrounds and areas of expertise. This required effective communication, conflict resolution, and leadership skills to ensure that everyone was aligned with the project goals and working together efficiently. It was essential to leverage the unique strengths of each team member while fostering a collaborative and inclusive environment.