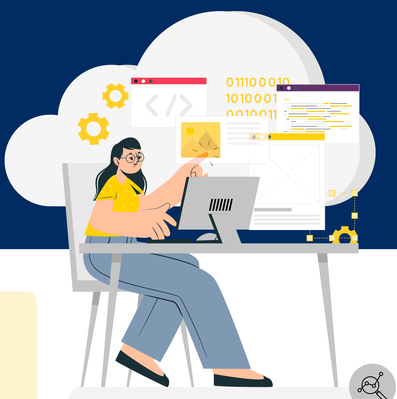
datos.gob.es

OPEN DATA ANALYSIS WITH OPEN SOURCE TOOLS (Part 1)



WHAT IS IT?

Data analysis is a systematic process that allows **transforming raw information into valuable knowledge**, providing:

- Informed decision-making.
- Identification of patterns and trends.
- In-depth understanding of complex data sets.

FIRST STEPS IN THE COMPREHENSIVE DATA ANALYSIS PROCESS

Data cleansing
Data conversion
Data analysis

1. Data preparation and debugging

Objetive: to transform raw data into clean, structured sets.

Examples of tools

OpenRefine (COMPLETELY FREE)

Data cleansing and transformation. It offers an intuitive graphical interface that is cross-platform (requires Java).
Detectability:

- Duplicated
- Incomplete data
- Structural inconsistencies

Talend Open Studio (FREE IN BASIC VERSION)

ETL (Extract, Transform, Load) tool.
(!) Requires intermediate programming skills.
Permitted:

- Component programming
- Integrating data from multiple sources

2. Data conversion

Objetive: to adapt the format of the data to facilitate its analysis.

Examples of tools

Mr Data Converter (COMPLETELY FREE)

- Conversion between CSV, Excel, JSON, HTM, XLM formats
- Simple web interface
- No installation required

Tabula (COMPLETELY FREE)

- Extraction of tables from PDF
- Convert documents into reusable formats
- Useful for reports and official documentation

Pandoc (COMPLETELY FREE)

- Universal document conversion
- Supports more than 20 different formats
- Powerful command line

3. Data analysis

Objetive: to explore, process and gain insights from datasets.

User-friendly analysis software

WEKA (COMPLETELY FREE)

- Machine learning and data mining
- Graphical interface
- Integration with scikitlearn, R and Deeplearning
- Ideal for beginners in machine learning

KNIME (FREE IN BASIC VERSION)

- Visual data analysis
- Workflows by connecting nodes
- Extensive library of components

ORANGE (COMPLETELY FREE)

- Paradigm drag and drop (drag and drop)
- Interactive visualisations
- Accessible statistical analysis

Development environments

Jupyter Notebook (COMPLETELY FREE)

- Executable documents
- Combination of code, visualisations and narrative
- Supports multiple languages
- Ideal for reproducibility

RStudio (FREE IN BASIC VERSION)

- Complete R language environment
- Console, editor and visualisation integration
- Advanced statistical tools

Programming languages for data analysis

R (COMPLETELY FREE)

Especialist in statistics

- Powerful for statistical analysis and visualisation
- Featured libraries:
  - » **Tidyverse**
  - » **ggplot2**

Python (COMPLETELY FREE)

A versatile language

- Recommendation: use Anaconda for environment management
- The main ones libraries for analysis are:
  - » **Pandas** (data manipulation)
  - » **NumPy** (numerical calculation)
  - » **scikit-learn** (machine learning)
  - » **Matplotlib** (visualisation)

Emerging tools

Streamlit (COMPLETELY FREE)

- Rapid creation of data-driven web applications
- Requires Python only
- Rapid prototyping of dashboards.
  - » In this practical exercise we use it to create a public data chat

Polars (COMPLETELY FREE)

- High performance
- Optimised pandas alternative
- Parallel processing

Apache Spark (COMPLETELY FREE)

- Distributed processing
- For Big data.
- APIs in Python, R and Scala

Recommendation
Start with simple tools such as Jupyter and Python, and gradually explore more advanced options as your analysis needs change.

Find out here the benefits and steps of Exploratory Data Analysis (EDA).

GOBIERNO DE ESPAÑA · MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PUBLICA · SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es · iniciativa aporta · datos.gob.es