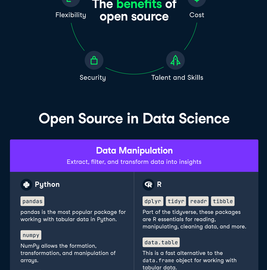
datacamp

Our Guide to Open Source in Data Science

The open-source revolution has transformed software and is driving a new age of data fluency. What are the benefits of open-source software? And what are the most used open-source data science tools?



Open Source in Data Science

Data Manipulation

Extract, filter, and transform data into insights

Python

`pandas`

pandas is the most popular package for working with tabular data in Python.

`numpy`

Numfy allows the formation, transformation, and manipulation of arrays.

`scipy`

Scifty contains a set of tools for statistics, linear algebra, data processing, and more.

R

`dplyr` `tidyr` `readr` `tibble`

Part of the tidyverse, these packages are R essentials for reading, manipulating, cleaning data, and more.

`data.table`

This is a fast alternative to the `data.frame` object for working with tabular data.

`xts`

xts is one of the most popular packages for working with time-series data in R.

Get Started

Courses

- → pandas foundations
- → Working with geospatial data in Python
- → Exploratory data analysis in Python

Tracks

- → Data Manipulation with Python (4 courses)

Courses

- → Introduction to the Tidyverse
- → Data Manipulation with dplyr
- → Exploratory Data Analysis in R

Tracks

- → Data Manipulation with R (5 courses)

Use Cases

Automate legacy Excel workflows Analyze sales data
Analyze traffic rates for city planning

Data Visualization

Visualize your data and make your insights easily interpretable

Python

`matplotlib`

Create and customize various types of data visualizations in Python.

`seaborn`

Built on top of Matplotlib, it allows for the creation of aesthetic plots in Python with few lines of code.

`bokeh` `plotly`

Create and publish interactive data visualizations and widgets in a web-page.

`folium`

Easily create and customize visualizations for geospatial data in Python.

R

`ggplot2`

Create and customize a range of data visualizations with the most popular data visualization package in R.

`leaflet`

Easily visualize geospatial data in R with robust styling capabilities.

`rbokeh` `plotly`

Create and publish interactive data visualizations and widgets in a web-page.

Get Started

Courses

- → Introduction to Data Visualization with Matplotlib
- → Introduction to Data Visualization with Seaborn
- → Interactive Data Visualization with Bokeh

Tracks

- → Data Visualization with Python (5 courses)

Courses

- → Introduction to Data Visualization with ggplot2
- → Interactive Data Visualization with rbokeh
- → Interactive Data Visualization with plotly in R

Tracks

- → Data Visualization with R (3 courses)

Use Cases

Create presentation-ready plots in three lines of code Build interactive dashboards on web-pages Visualize Covid-19 cases across the world

Machine Learning

Make predictions with your data, and automate business processes

Python

`scikit-learn`

The most popular end-to-end library for machine learning across any programming language.

`xgboost` `catboost` `lightbm`

Easily apply one of the most popular techniques for machine learning on tabular data in Python.

`tensorflow`

An end-to-end deep learning framework for building, evaluating, and deploying deep learning models.

`keras`

Built on top of TensorFlow, Keras simplifies building, evaluating, and deploying deep learning models.

`pytorch`

Widely used in research, PyTorch provides a toolset for building and deploying deep learning models.

R

`tidymodels`

Similar to the tidyverse, it's a collection of R packages designed for end-to-end machine learning in R.

`xgboost`

Easily apply one of the most popular techniques for machine learning on tabular data in R.

`metrics`

One of the most popular R packages for evaluating the performance of a range of machine learning predictions.

`rpart`

rpart is a popular package for working with tree-based models in R.

Get Started

Courses

- → Supervised Learning with scikit-learn
- → Unsupervised Learning in Python
- → Introduction to Deep Learning with Keras
- → Introduction to Deep Learning with PyTorch

Tracks

- → Machine Learning Scientist with Python (23 courses)

Courses

- → Supervised learning in R: Classification
- → Supervised learning in R: Regression
- → Unsupervised Learning in R
- → Machine Learning for Marketing Analytics in R
- → Tree-based models in R

Tracks

- → Machine Learning Scientist with R (15 courses)

Use Cases

Create presentation-ready plots in three lines of code Build interactive dashboards on web-pages Visualize Covid-19 cases across the world

Reporting and Communicating Data

Communicate data insights, and easily share data analysis

Python

`dash`

Develop and customize interactive dashboards that can be rendered and shared directly in a browser.

`jupyter notebooks`

Jupyter Notebooks allow creating and sharing documents containing live code, visualizations, and text.

R

`R Markdown`

RMarkdown notebooks allow creating and sharing documents containing live code, visualizations, and text.

`shiny` `shinydashboards`

One of the most popular tools in data science for developing, customizing, and deploying interactive dashboards that can be rendered in a browser.

Get Started

Projects

- → Comparing Search Interest with Google Trends
- → Exploring the evolution of lego
- → Bad passwords and the NIST guidelines
- → Analyzing TV Data

Courses

- → Building Web Applications with Shiny in R
- → Building Dashboards with shinydashboard
- → Reporting with R Markdown
- → Building Dashboards with flexdashboard

Tracks

- → Shiny Fundamentals (4 courses)

Use Cases

Live track department OKRs with a web-based dashboard Share machine learning results with business stakeholders Onboard new hires on data processes

Natural Language Processing

Analyze and extract insights from text data

Python

`gensim`

Python library with efficient tools for topic modeling, document comparison, topic identification, and more.

`spacy`

Perform a range of NLP tasks from tokenization, part-of-speech tagging, text classification, and more.

`nltk`

Python library containing tools for text data preprocessing, classification, sentiment analysis, and more.

R

`tidytext`

Perform a range of NLP tasks from stopword removal, tokenization, sentiment analysis, and more.

`topicmodels`

Perform a range of functions aimed at identifying and summarizing text, and categorizing documents.

`stringr`

Manipulate text data with tools for string detection, string subsetting, joining and splitting strings, and more.

Get Started

Courses

- → Introduction to Natural Language Processing in Python
- → Sentiment Analysis in Python
- → Analyzing Social Media Data in Python

Tracks

- → Natural Language Processing in Python (5 courses)

Courses

- → Introduction to Natural Language Processing in R
- → Introduction to text analysis in R
- → Intermediate Regular Expressions in R

Tracks

- → Text Mining with R (4 courses)

Use Cases

Categorize documents based on topic Analyze social media data Perform sentiment analysis on customer support tickets

Dive Deeper

Learn more about more than 75 open-source data science packages across 10 categories from data cleaning to data engineering, how the most advanced data teams are using open source, and how to get started with data upskilling.

Get the White Paper →

clear as data