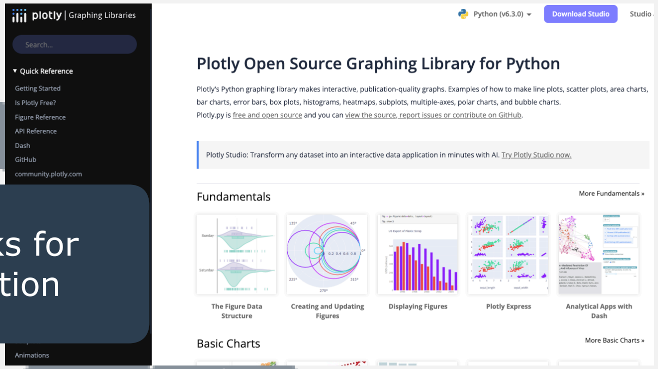
# Gen AI Data Scientist with Python Cheat Sheet

If you want to become a Generative AI Data Scientist with Python, then join our Gen AI Bootcamp for Data Scientists.



plotly | Graphing Libraries

Search...

▾ Quick Reference
- Getting Started
- Is Plotly Free?
- Figure Reference
- API Reference
- Dash
- GitHub
- community.plotly.com
- Animations

Python (v6.3.0) ▾ | Download Studio | Studio

## Plotly Open Source Graphing Library for Python

Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.
Plotly.py is free and open source and you can view the source, report issues or contribute on GitHub.

Plotly Studio: Transform any dataset into an interactive data application in minutes with AI. Try Plotly Studio now.

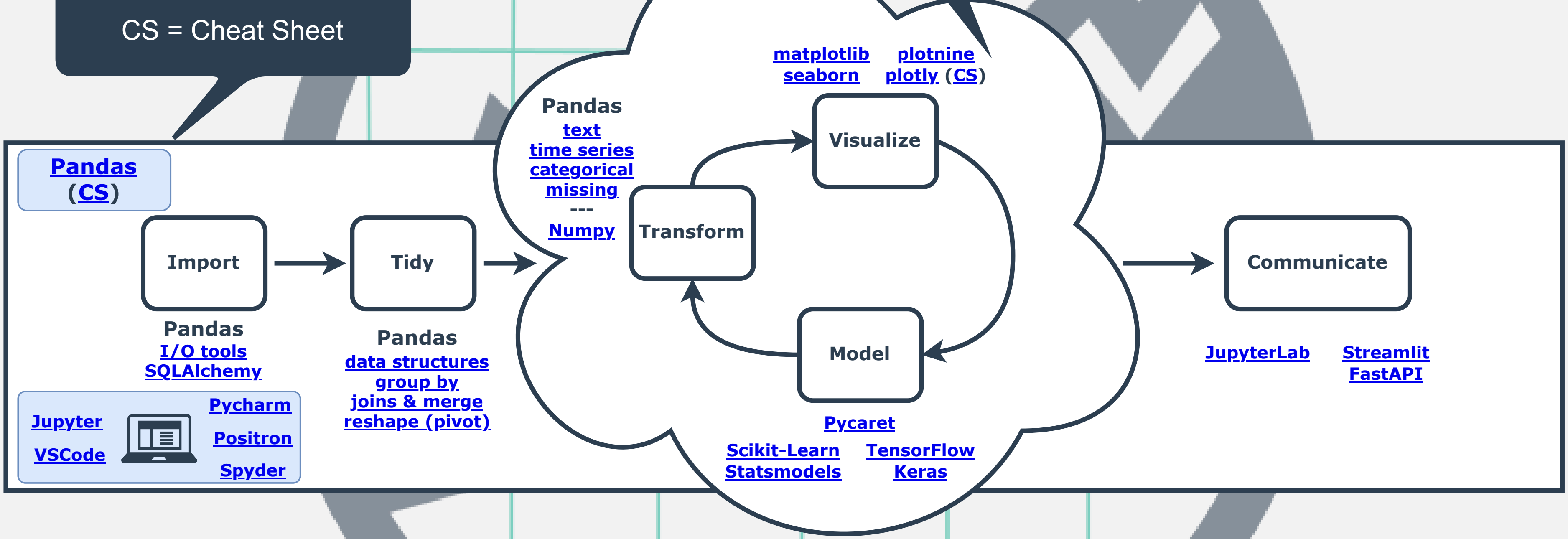
Fundamentals — More Fundamentals »

The Figure Data Structure | Creating and Updating Figures | Displaying Figures | Plotly Express | Analytical Apps with Dash

Basic Charts — More Basic Charts »

Click the links for Documentation

CS = Cheat Sheet



**Import** — Pandas: I/O tools, SQLAlchemy

Jupyter, VSCode, Pycharm, Positron, Spyder

**Tidy** — Pandas: data structures, group by, joins & merge, reshape (pivot)

**Transform** — Pandas: text, time series, categorical, missing --- Numpy

**Visualize** — matplotlib, seaborn, plotnine, plotly (CS)

**Model** — Pycaret, Scikit-Learn, Statsmodels, TensorFlow, Keras

**Communicate** — JupyterLab, Streamlit, FastAPI

## Important Resources

- **NEW:** **Awesome Generative AI Data Scientist GitHub**
- **NEW:** **AI Data Science Team (AI / ML Agents)**
- **Anaconda Distribution:** https://www.anaconda.com/download/
- **Python Documentation:** https://docs.python.org/
- **Python Standard Library:** https://docs.python.org/3/library

*Click here to join the* ***Generative AI Data Scientist Bootcamp***

| | | |
|---|---|---|
| OpenAI | LangChain | LangGraph |
| learn | Streamlit | FastMCP |

Business Science University
university.business-science.io

version: 4.0

# Data Science & ML

Special Topics

## Time Series Forecasting

- Nixtla - TimeGPT, StatsForecast, MLForecast, NeuralForecast, Hierarchical Forecast
- sktime - Scikit-Learn Extension for Time Series
- statsmodels - Time Series Analysis
- GluonTS - MXNet/Gluon Deep Learning for Time Series

## Time Series Analysis

- PyTimetk - Time series analysis in python
- TSFresh - Time Series Feature Engineering
- tslearn - Time Series Features
- Pandas Time Series
- Arrow - Human-Friendly Time

## EDA

- pandas-profiling, SweetViz, lux

## Web

- beautifulsoup - Extract data from HTML
- requests-html - HTML Parsing
- scrappy - Web crawling

## MS Office, Google & PDF

- XlsxWriter - Create Excel Workbooks
- pyexcel -Read/Write Excel
- xlwings - Call python from Excel
- python-docx - Word Documents
- python-pptx - PowerPoint Documents
- pdfminer - Text extraction from PDF
- textract - Extract text from any document
- PyPDF2 - Create PDF documents
- gspread - Google Sheets

## Text Analysis & NLP

- NLTK - Text Tokenization & Modeling
- spaCy - NLP using Cython for Speed
- fuzzywuzzy - Fuzzy String Matching

## Recommendation Systems

- Annoy - Approximate Nearest Neighbors
- LightFM - Popular recommendation algo's.

## Apps & APIs

- Streamlit - User-Friendly Web App Framework
- FastAPI - Web framework for building APIs in Python
- Flask - Web App Development

## MLOps

- Pycaret MLFlow Integration
- MLFlow - Machine Learning Lifecycle, Tracking, Deployment
- MetaFlow - Scalable AWS Jobs for Data Scientists

## Cloud (AWS, Azure, GCP)

- boto3 (AWS) - AWS Python SDK
- Google Cloud - GCP Python SDK
- Azure - Azure Python SDK

## ETL & Automations

- Airflow - Workflow Scheduling & Monitoring
- Luigi - Batch Job Tool, Scheduling, Monitoring
- Prefect - Open-source Orchestration Software

## Machine Learning

- Scikit-Learn - ML in Python
- H2O - Scalable & AutoML
- TPOT - TPOT Automated ML Tool
- PyCaret - PyCaret Low Code ML
- Dask ML - Scalable ML with Dask
- ML Packages: XGBoost, LightGBM, CatBoost

## Feature Engineering

- Sklearn Data Transformations
- sklearn-pandas - Sklearn Extension for Pandas
- Featuretools - Automated Feature Engineering
- category_encoders - Categorical Encoding
- imbalanced-learn - Resampling for Imbalanced
- fancyimpute - Extended imputation strategies

## Deep Learning

- TensorFlow & Keras
- PyTorch
- MXNet, Gluon, & GluonTS
- OpenAI Gym - Reenforcement Learning

## Image & Comp Vision

- OpenCV - Open Source Computer Vision
- Scikit Image - Image Processing
- Pillow - Python Imaging Library

## Speed & Scale

- Polars - Rust Speed Up
- Dask (CS) - Parallel Pandas & Scikit Learn
- RAPIDS (CS)- GPU Accelerated Pandas
- PySpark - Spark Clusters

## Coming from R?

- R-to-Pandas Comparison
- siuba & plydata - dplyr/tidyr ports
- datatable - data.table port
- plotnine - ggplot2 port

Business Science University
university.business-science.io

# AI Python Stack

**Generative Artificial Intelligence (AI) & LLMs**

## AI LLM Frameworks

Frameworks for Large Language Models (LLMs)

- LangChain - Application development framework for apps powered by LLMs with many integrations, tools, and community extensions, support for a broad spectrum of Agents and more.
- LangGraph - Build DAG Graphs to combine multiple LLMs and Agents.
- LLamaIndex - An alternative to LangChain that focuses on RAG (Retrieval Augmented Generation) and Vector Indexing and Retrieval.

## Vector Databases

Used to store text embedding and similarity search

- ChromaDB - Open source vector DB
- FAISS - Facebook AI Similarity Search
- Pinecone - Scalable, cloud-based vector DB
- Milvus - Scalable cloud-based vector DB
- Zilliz - Fully managed cloud built on Milvus

More vector databases

## Embedding Models

Text Embeddings Models:

- OpenAI Embedding
- Hugging Face Transformers Library

More Embedding Models

## LLM Models & APIs

LLM Inference Models, APIs, and SDKs:

- Hugging Face Models - Massive library of open-source models for data science, machine learning, and AI.
- OpenAI Python SDK and API - A software development kit for interfacing with OpenAI API.
- Anthropic Claude SDK and API - A software development kit for interfacing with Anthropic API and Claude models
- Meta Llama Models - Open source LLMs by Facebook / Meta
- Ollama - Run open-source LLMs such as Llama 2 and 3 locally
- Groq - Blazing fast inference
- Mistral AI - Open source and commercial LLM models

LangChain LLM Models
More LangChain LLM Integrations

## Document Loaders

LangChain Native Document Loaders:

- PDF
- CSV
- HTML
- JSON
- Cloud

All Document Loaders

## Text Splitters

LangChain Native Document Transformers:

- Character Splitter
- Recursive Character Splitter
- HTML Splitters

More Splitters

**Now using LangChain v1**

## Agents

Building agents with Tools:

- Agents (LangChain)
- Deep Agents
- Swarm Agents
- Supervisor Agents

## Tools

Integrations and Community Tools:

- Pandas DataFrame
- SQL Databases
- Spark SQL
- MCP Toolbox for Databases

LangChain Tools
More Tool Integrations

## MCP Servers

Building agents with Tools:

- LangChain MCP Adapters
- FastMCP

## Extra AI Concepts

- LangChain Messages
- LangChain Structured Output
- LangChain Short-Term Memory
- LangChain Streaming
- LangChain Middleware

*Click here to join the* ***Generative AI Data Scientist Bootcamp***

Business Science University
university.business-science.io