MOST IMPORTANT LIBRARIES TO BE USED IN DATA SCIENCE WITH THEIR SPECIFIC FUNCTION

Data Science libraries are crucial because they provide prebuilt tools and functionalities that streamline complex tasks like data analysis, machine learning and visualization.

Sr.No	Library Name	Function	Key Features
1	Pandas	Handles structured data with Data Frames, offering powerful data wrangling and transformation tools.	Data Frames and Series; data cleaning; merging/joining; handling missing data; time series support.
2	NumPy	Supports high-performance mathematical operations with multi-dimensional arrays.	Fast numerical operations; multi-dimensional arrays; broadcasting; linear algebra tools.
3	Dask	Scales data processing to larger datasets, working efficiently on distributed systems.	Parallel and distributed computation; works with datasets larger than memory; Pandas-like API.
4	Vaex	Optimized for handling large datasets efficiently with fast computations.	High performance, lazy computation, memory efficiency, efficient joints and visualization
5	Polars	Alternative to Pandas, providing high-speed data processing using Rust-based architecture.	Speed and performance, lazy evaluation, expressive API and big data handling
Data Visualization			
6	Matplotlib	Provides fundamental plotting functions for graphs and charts.	Low-level control; supports all basic plot types; publication-quality visuals.
7	Seaborn	Extends Matplotlib with statistical visualizations and enhanced styling	High-level abstraction over Matplotlib; attractive statistical plots; great integration with Pandas.

8	Plotly	Generates interactive plots, dashboards, and visualizations.	Interactive, web-based charts; 3D plots; supports Dash for dashboards.
9	Bokeh	Supports interactive web-based visualizations.	Real-time interactivity; dashboard-ready; supports streaming data.
10	ggplot (for Python)	inspired by the R ggplot2 package, simplifies complex plots creation.	Layered Approach, Aesthetic approach, Faceting, Custom themes, Scalability, Extensions, Annotations and labels.
Machine Learning			
11	Scikit-learn	Essential machine learning library for classification, regression, and clustering.	Consistent API for models; preprocessing tools; model evaluation; grid search.
12	TensorFlow	Deep learning framework for scalable neural networks and AI applications.	Graph-based deep learning; scalable for production; TensorBoard for visualization.
13	Py Torch	Popular framework for deep learning with dynamic computation graphs	Dynamic computation graphs; intuitive debugging; strong support for research.
14	XGBoost	Optimized gradient boosting algorithm for structured data analysis.	High-speed gradient boosting; handles missing values; built-in regularization.
15	LightGBM	A fast, efficient gradient boosting library for machine learning tasks.	Fast Training Speed, Lower Memory Usage, Parallel & Distributed Learning, and Handling Large-Scale Data:

Natural Language Processing (NLP)			
16	NLTK	Provides tools for text processing, tokenization, and linguistic analysis.	Corpus & Lexical Resources, Text Processing, Classification & Parsing, Named Entity Recognition (NER), Integration with Other Libraries, Visualization & Analysis.
17	spaCy	Designed for efficient NLP processing with pre-trained models.	Pre-trained Models, Tokenization, Named Entity Recognition (NER), Dependency Parsing.
18	Transformers (Hugging Face)	Powerful library for leveraging state-of-the-art deep learning NLP models.	Pre-trained Models, Unified API, Text Processing, Parallel Processing.
Data Storage & Retrieval			
19	SQLAlchemy	Python's SQL toolkit for database management.	Type Safety, Performance Optimization, Flexibility, Integration, Modern Design.
20	PyMongo	Interface for interacting with MongoDB NoSQL databases.	Efficient Database Operations, Flexible Schema, Aggregation Framework, and security features.
Data Engineering & Pipeline Management			

21	Apache Airflow	Workflow automation tool for data pipeline management.	Integration with Cloud & Big Data, Extensibility, Workflow Automation, and Scalability
22	Luigi	Assists with building complex pipelines of tasks.	Task Dependency Management, Failure Handling & Retry Mechanism, and Integration with Big Data Tools
23	Great Expectations	Enables automated validation of data quality.	Data Validation, Automated Testing, Customizable Expectations, and Integration with Data Pipelines
Statistics & Mathematics			
24	SciPy	Houses advanced mathematical functions, optimization, and statistics.	Advanced Mathematical Functions: Provides modules for optimization, integration, interpolation, linear algebra, and statistics, High-Performance Computing.
25	Statsmodels	Provides statistical models like regression, hypothesis testing, and time-series analysis.	Regression Models, Time Series Analysis, Statistical Tests and Nonparametric Methods