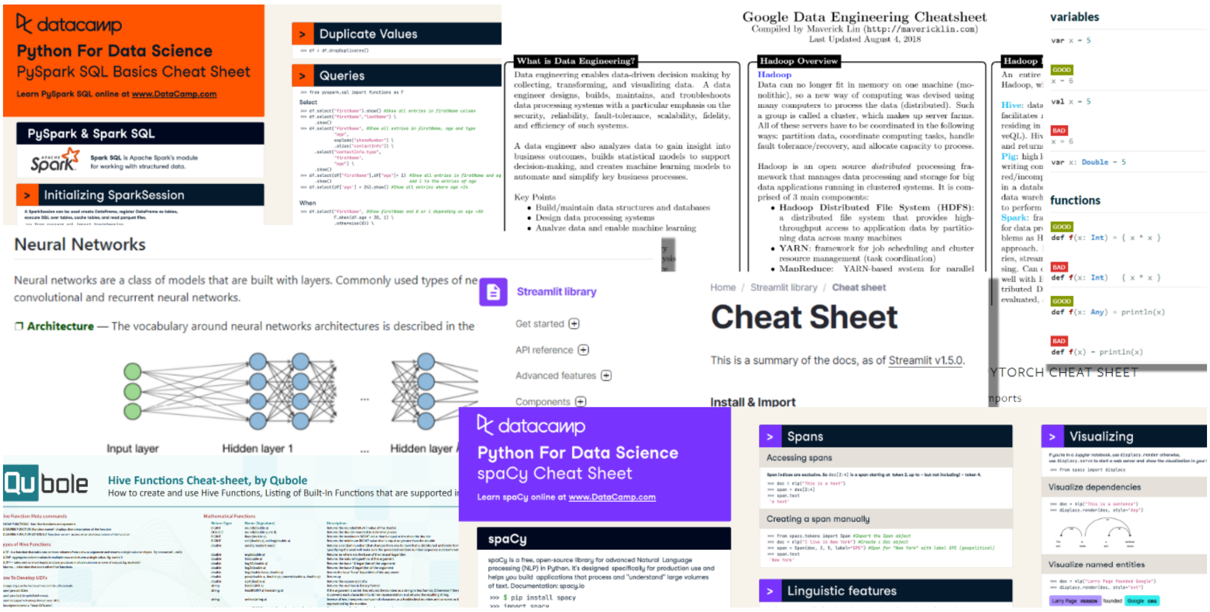
The Complete Collection of Data Science Cheat Sheets

Abid Ali Awan

Data Scientist & Writer at KDnuggets

A collection of cheat sheets that will help you prepare for a technical interview, assessment tests, class presentation, and help you revise core data science concepts.



The cheat sheets can help us revise the concepts of statistics, programming language syntax, data analytics tools, and machine learning frameworks. It can also help you ace technical interviews and assessment tests. Jupyter Notebook is the essential cheat sheet that everyone should learn. It contains shortcuts, tricks, and functions for running a Python notebook.

I use cheat sheets to prepare for technical interviews, as tech recruiters want to assess the subject matter expertise. Searching for the cheat sheet that works for you can take hours as most of them are not easy to comprehend. The collection has 12 subcategories that include easy-to-follow and summarized sheet cheats to revise all the concepts of data sciences.

Table of Contents

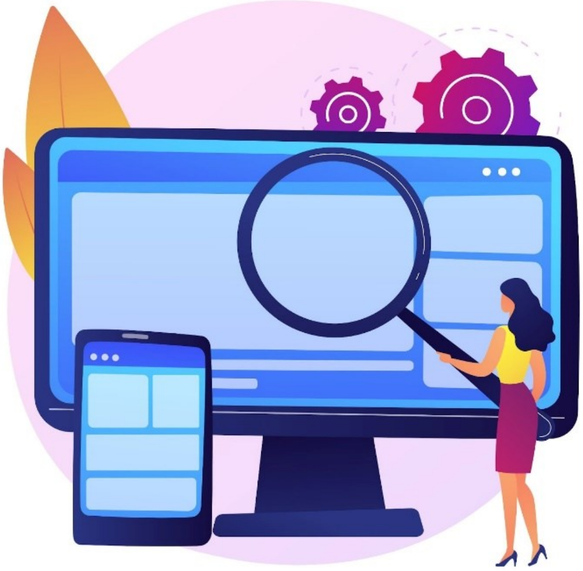
SQL

Majority of technical interviews and assessment tests include some type of SQL questions so, it is better to prepare for the interview using the collection of SQL cheat sheets. These cheat sheets will also help you get better at creating and managing databases. It will also help you understand complex SQL queries.

- **SQL for beginners**
- **SQL Expert**
- **SQL – Data Analysis**
- **PostgreSQL**

Web Scraping

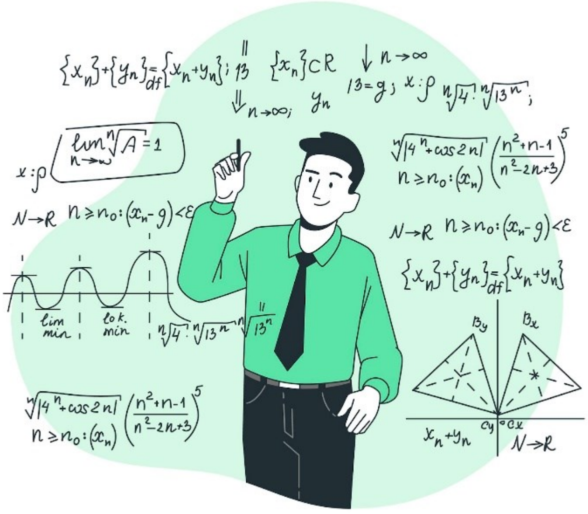
Web Scraping is an essential part of data science, as it is used for gathering data, market research, and maintaining data pipelines. Beautiful Soup is a popular library for parsing HTML/Java scripts and converting them into human-readable dataframe. The section consists of tools that are used to parse scripts in Python and R.



- **Web Scraping with Python**
- **Web scraping with R**
- **Beautiful Soup**
- **Selenium**
- **Scrapy**
- **XPath**
- **HTML Scraping**

Statistics, Probability, & Math

Artificial intelligence, data analytics, and modern research depend on statistics. It is the backbone of our modern society, so if you want to review old concepts or learn new complex ideas then check out a collection of statistical cheat sheets.



- **Probability**
- **William Chen's Probability Cheatsheet 2.0**
- **Stanford: Algebra and Calculus**
- **Statistics, Probability & Math**
- **MIT: Statistics**
- **Stanford: Statistics**
- **Calculus for Machine Learning**
- **Linear algebra for deep learning**
- **SciPy: Linear Algebra in Python**

Data Analytics

Data analytics is used for making business decisions, marketing campaigns, scientific research, and designing unique data products. Entire IT industry depends on it. This category is further divided into three subcategories: **Python**, **R**, **Julia**. All of these languages are popular among data scientists and data analysts.

Python

The list contains the most used Python packages from data ingestion, manipulation, and visualization. Numpy and Pandas are the most popular tools among the data community for performing scientific calculation and data augmentation.

- **Python For Data Science Cheat Sheet For Beginners**
- **Pandas for Data Science**
- **Pandas: Data Wrangling**
- **NumPy**
- **Matplotlib**
- **Python Seaborn**
- **Data Visualization: Bokeh**
- **Importing Data**
- **PySpark**

R

R is quite famous among statisticians and data analytics professionals. It is recommended to learn syntax and functions of famous Packages such as Tidyverse. The Tidyverse contains a complete data science solution from importing data to creating visually simulating data reports.

- **Python with R and reticulate**
- **Tidyverse For Beginners**
- **Data visualization with ggplot2**
- **Data transformation with dplyr**
- **Data tidying with tidyr**
- **Data import with readr, readxl, and googlesheets4**
- **Apply functions with purrr**
- **Factors with forcats**
- **Dates and times with lubridate**
- **Dynamic documents with rmarkdown**
- **Advanced R**
- **The data.table R Package**
- **xts Cheat Sheet: Time Series in R**
- **cartography**

Julia

Julia is an emerging language, and, in my opinion, it is the future of data science. The list contains a quick introduction of Julia syntax, data wrangling, and data visualization.

- **Fast Track to Julia**
- **Data Wrangling with DataFrames.jl**
- **Plots.jl**
- **MATLAB Vs. Python Vs. Julia**
- **Pluto.jl**
- **Make.jl Examples**

Business Intelligence

No code applications for Business Intelligence are becoming industry standards. These applications can help you create data analytical reports, dashboards, and immersive visualization. These tools are helping businesses make data-driven decisions. The most popular tools are MS Excel, Power BI, and Tableau.

- **Data Science for Business Leaders**
- **PowerBI**
- **PowerBI: DAX**
- **Tableau**
- **MS Excel**
- **Business Intelligence**

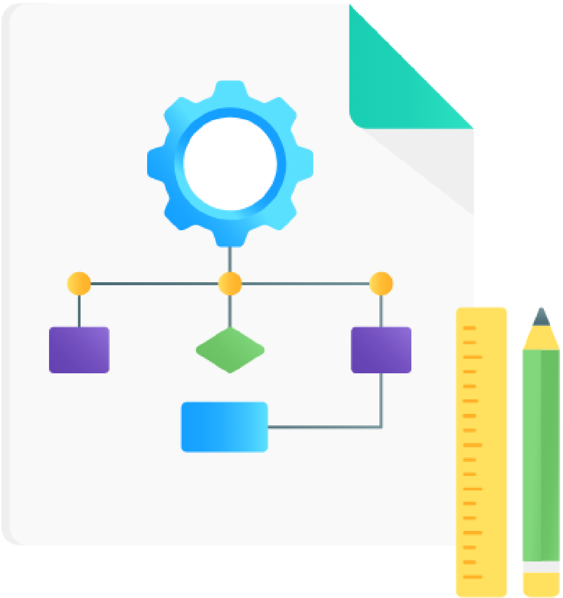
Big Data

By 2025, it is estimated that 463 exabytes of data will be created each day globally - weforum.org. With that, major data companies are looking for data engineers and data scientists to work on big data solutions. This collection of cheat sheets can give you an introduction to the essential big data tools.

- **Hadoop**
- **Scala**
- **Spark**
- **Hive Functions**
- **Spark with sparklyr**

Data Structures & Algorithms

The most common technical interview questions are about data structures and algorithms. If you are a software engineer or data scientist then you must know common data structure operations, search & sorting algorithms, and data structure types. The list was created to help you understand complex sorting functions and algorithms.



- **Big-O Complexity Chart**
- **Common Data Structure Operations / Array Sorting Algorithms**
- **Data Structures**
- **Princeton: Algorithms and Data Structures**
- **Essential of Data Structures and Algorithms**

Machine Learning

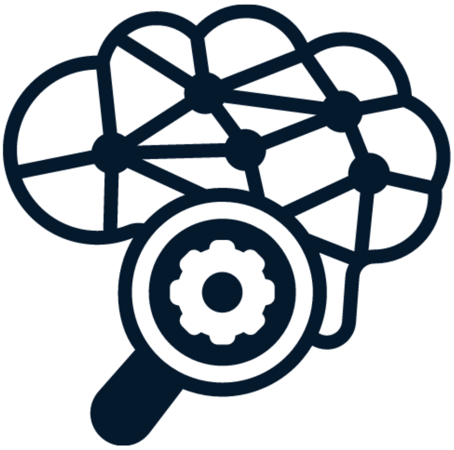
This is the most in-demand cheat sheet among the data community. Whenever I have a machine learning or deep learning interview, I spend a couple of hours revising all of the key concepts of machine learning and model architecture. Sometimes hiring managers won't have the technical knowledge, so they will also use cheat sheets for preparations. The collection consists of machine learning frameworks, algorithms and neural network architectures cheat sheets.

- **Supervised learning**
- **Unsupervised learning**
- **Scikit-Learn: Python Machine Learning**
- **Scikit-Learn: Machine Learning Algorithm Selection**
- **Machine Learning Algorithm**
- **Time Series with R**
- **Machine Learning tips and tricks**
- **Caret: Modeling and machine learning in R**
- **Machine Learning Modeling with R**

Bonus: Get a free machine learning crash course by subscribing to Machine Learning Mastery **here**. The crash course includes free eBooks, code-based content, and a gift “ML Performance Improvement Cheat Sheet”.

Deep Learning

Modern machine learning applications run on deep neural networks and every data-related job expects you to have some knowledge about deep learning or Advance AI technologies. The deep learning models are driving modern technologies such as computer vision, automatic speech recognition, natural language processing, medical research, and self-driving cars. The list below contains information about deep learning frameworks (Pytorch/Keras/Tensorflow), model architectures, graph neural networks, and data processing techniques.



- **Deep Learning**
- **PyTorch**
- **Neural Network Architectures**
- **Neural Network Graphs**
- **Neural Network Cells**
- **Neural Network Type with Diagram**
- **Keras: Neural Networks in Python**
- **Deep learning with Keras in R**
- **TensorFlow**

Natural Language Processing

Natural Language Processing (NLP) is used for processing and cleaning text, audio, and image data so we can extract useful information. NLP applications are limitless, as it is used for language translation, transcription, conversation AI, question & answering, generative technology, classification, name entity recognition, and many more. The collection of cheat sheets contains bite-size information about the most famous NLP tools and algorithms.

- **spaCy: Advanced NLP in Python**
- **String manipulation with stringr**
- **Regular Expressions with R**
- **NLP for Beginners**
- **Python & nltk**
- **Advanced NLP**
- **Transformers Documentation**
- **NLP Python Introduction**
- **Gensim**

Data Engineering

The data engineer's job requirement includes proficiency in SQL, Extract-Transform-Load (ETL) operations, creating & managing databases, automating data pipelines, and processing big data. The data engineer jobs are in demand, and companies want to hire the best engineer for creating and managing fully automated data pipelines. The list below contains cheat sheets on the most popular data engineer tools such as Apache Airflow and Kafka.

- **Spark DataFrames in Python**
- **Data Engineering**
- **Data Engineering on Microsoft Azure**
- **Apache Kafka**
- **dbt(data built tool)**
- **AWS Redshift**
- **Apache Airflow**
- **Docker**
- **BigQuery**

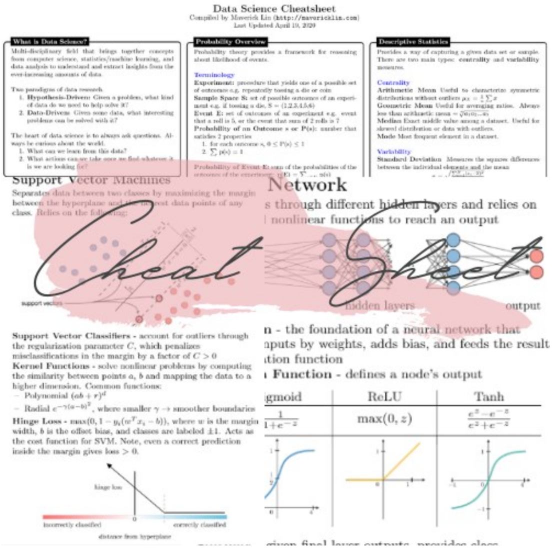
Web Frameworks

Even though this is optional, I have been asked in the past by hiring managers about my experience with end-to-end machine learning applications. They will ask you about Django, Flask, and FastAPI or experience in deploying models to production. It is good practice to learn about web frameworks before a technical interview. The list consists of R-shiny, Plumber, Golem, Streamlit, FastAPI, Flask, and Django web frameworks.

- **Interactive web apps with shiny**
- **Web APIs for R with plumber**
- **Golem with R**
- **Streamlit**
- **FastAPI**
- **Flask**
- **Django**

Bonus: VIP Cheat Sheet

VIP cheat sheets are a data science goldmine that contains bit size information about data science and its core subjects. The cheat sheets include the basic information about data types, algorithms, NLP, machine learning, data analytics, and data processing. If you are preparing for a general data interview, then I will suggest you download any VIP cheat sheet and revise all the core topics on data science and machine learning.



- **Stanford: Super VIP Cheat Sheet**
- **Data Science Cheat Sheet by Aaron Wang**
- **Data Science Cheat Sheet by Maverick Lin**
- **Machine Learning Bites by Rishabh Anand**
- **Machine Learning Interviews**

Conclusion

If you are preparing for an interview or presentation, use these collections of cheat cheats to revise the core concepts of data science. We have covered SQL, Web Scraping, Statistics, Probability, & Math, Data Analytics, Business Intelligence, Big Data, Data Structures & Algorithms, Machine learning, Deep Learning, Natural Language Processing, Data Engineering, Web Frameworks. If you want to ace your next interview, then save this PDF so that you can always come back and prepare for the technical interview.

About Author

Abid Ali Awan (@1abidaliawan) is a certified data scientist professional who loves building and deploying machine learning models. Currently, he is focusing on content creation and writing technical blogs on machine learning and data science. Abid holds a master's degree in Technology Management and a bachelor's degree in Telecommunication Engineering. His vision is to build an AI product using a graph neural network for students struggling with mental illness.