

# 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 collage features several cheat sheets: **Array Sorting Algorithms** with a table of time complexities (Best, Average, Worst, Spt, Wo) for algorithms like Bubble Sort, Selection Sort, Insertion Sort, Merge Sort, Quick Sort, and Heap Sort. **Unique Query Features** lists various SQL and NoSQL query capabilities. **and Visualization** covers data tracking and visualization techniques. **Neural Network Graphs** shows different network architectures. **ard Tools** lists tools like Power BI, Tableau, and Alteryx. **Extensions** details Spark and R integrations. **IN Operator**, **BETWEEN Operator**, and **LIKE Operator** sections provide SQL query examples. **Map and Write to CSV** and **Map and Write to Excel** sections show data export methods. **Map and Write to SQL Query or Database Table** shows database integration. A central graphic reads 'Data Science Cheat Sheet' with a starburst effect.

## Cheat Sheets from KDnuggets

This row displays four cheat sheet thumbnails: **Scikit-learn for Machine Learning**, **Linux for Data Science**, **Git for Data Science**, and another **Scikit-learn for Machine Learning** cheat sheet.

Scikit-learn for Machine Learning

Linux for Data Science

Git for Data Science

# Table of Contents

1. [SQL](#)
2. [Web Scraping](#)
3. [Statistics, Probability, & Math](#)
4. [Data Analytics](#)
5. [Business Intelligence](#)
6. [Big Data](#)
7. [Data Structures & Algorithms](#)
8. [Machine Learning](#)
9. [Deep Learning](#)
10. [Natural Language Processing](#)
11. [Data Engineering](#)
12. [Web Frameworks](#)

**Bonus:** [VIP Cheat Sheet](#)



# SQL

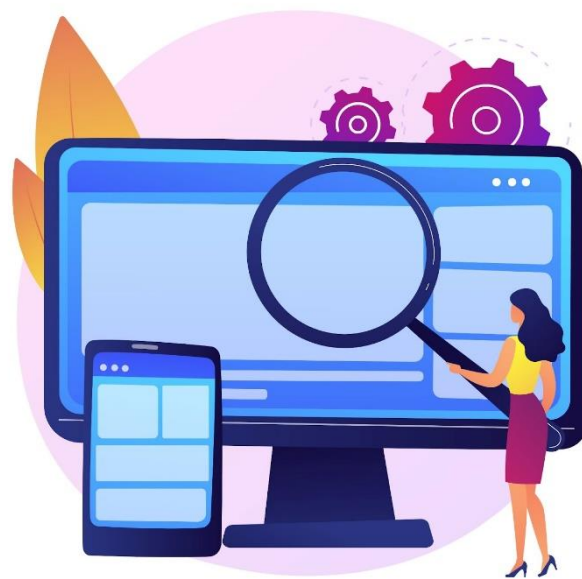
Majority of technical interviews and assessment tests include some type of SQL questions so, it is better to prepare for the interview using the collection of SQL cheat sheets. These cheat sheets will also help you get better at creating and managing databases. It will also help you understand complex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



- [SQL Basics](#)
- [SQL Expert](#)
- [SQL Window Functions Cheat Sheet](#)
- [SQL Joins Cheat Sheet](#)
- [SQL – Data Analysis](#)
- [PostgreSQL](#)
- [SQL for the Job Interview](#)

# Web Scraping

Web Scraping is an essential part of data science, as it is used for gathering data, market research, and maintaining data pipelines. BeautifulSoup 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.

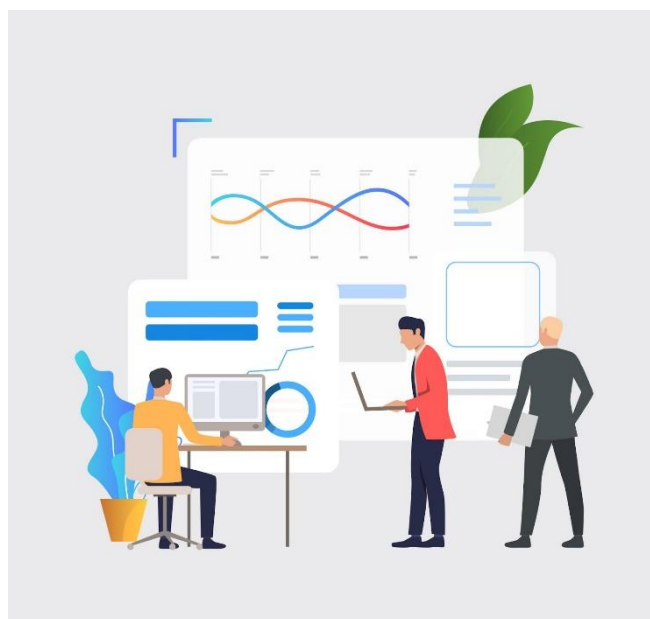


- **Web Scraping with Python**
- **Web scraping with R**
- **Beautiful Soup**
- **Selenium**
- **Scrapy**
- **XPath**
- **HTML Scraping**



# 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.

- **Pandas for Data Science**
- **Pandas: Data Wrangling**
- **Data Visualization**
- **NumPy**
- **Matplotlib**
- **Seaborn**
- **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.

- [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

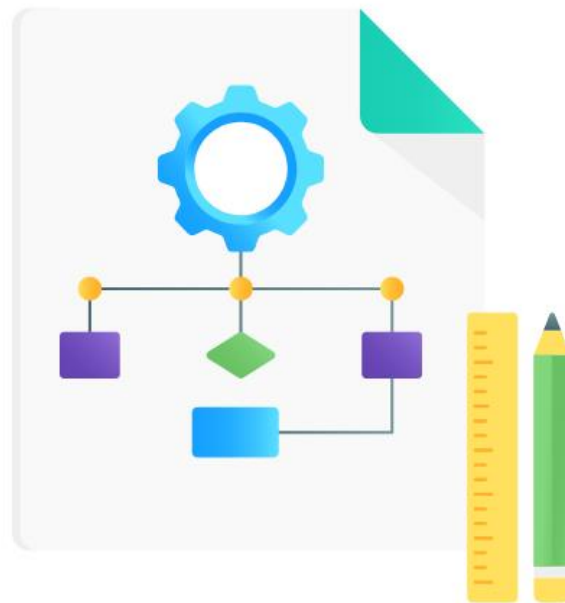
By 2025, it is estimated that 463 exabytes of data will be created each day globally - [weforum.org](http://weforum.org). With that, major data companies are looking for data engineers and data scientists to work on big data solutions. This collection of cheat sheets can give you an introduction to the 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 for interviews](#)
- [Princeton: Algorithms and Data Structures](#)
- [Essential of Data Structures and Algorithms](#)
- [An Executable Data Structures](#)

# 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.

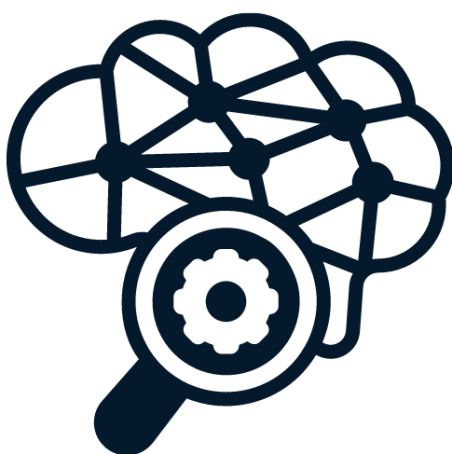


- **Supervised learning**
- **Statistics & Mathematics for Machine Learning**
- **Unsupervised learning**
- **Scikit-Learn: Python Machine Learning**
- **Scikit-Learn: Machine Learning Algorithm Selection**
- **Machine Learning Models**
- **Time Series with R**
- **Machine Learning tips and tricks**
- **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](#)
- [ChatGPT](#)

# Data Engineering

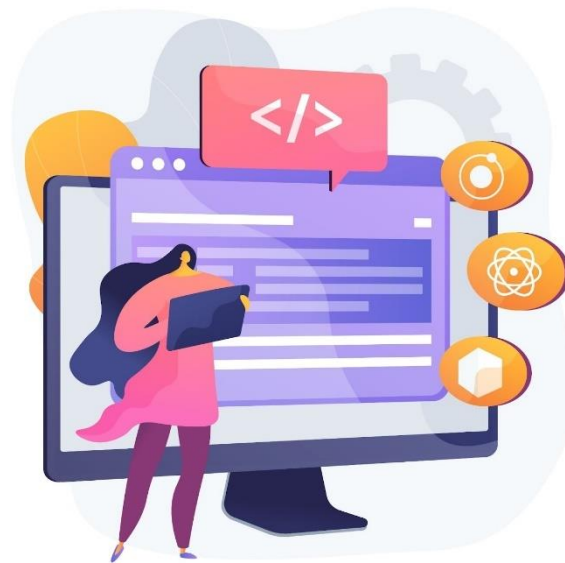
The data engineer's job requirement includes proficiency in SQL, Extract-Transform-Load (ETL) operations, creating & managing databases, automating data pipelines, and processing big data. The data engineer jobs are in demand, and companies want to hire the best engineer for creating and managing fully 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.

Data Science Cheatsheet  
Compiled by Haverick Lin (http://havericklin.com)  
Last updated: April 10, 2019

**What is Data Science?**

Multi-disciplinary field that brings together concepts from computer science, mathematics, learning, and data analysis to understand and extract insights from the overwhelming amounts of data.

Two paradigms of data research:

- Hypothesis Driven:** Given a problem, what kind of data do we need to help solve it?
- Data Driven:** Given some data, what interesting problems can be solved with it?

The heart of data science is to answer all questions. All ways to answer about the world:

- What can we learn from the data?
- What action can we take now on that information? In what we are looking for?

**Support Vector Machines**

Separate data between two classes by maximizing the margin between the hyperplanes and minimizing the slack of data points of any class. (Relies on the idea of soft margin)

**Support Vector Classifiers** - account for outliers through the regularization parameter  $C$ , which penalizes misclassifications in the margin by a factor of  $C > 0$ .

**Kernel Functions** - solve nonlinear problems by computing the similarity between points  $a, b$  and mapping the data to a higher dimension. Common functions:

- Polynomial:  $(a \cdot b)^\gamma$
- Radial:  $e^{-\gamma \|a-b\|^2}$ , where smaller  $\gamma \rightarrow$  smoother boundaries

**Hinge Loss** -  $\max(0, 1 - \gamma(w^T x_i - b))$ , where  $w$  is the margin width,  $b$  is the offset bias, and classes are labeled  $\pm 1$ . Acts as the cost function for SVM. Note, even a correct prediction inside the margin gives loss  $> 0$ .

**Probability Overview**

Probability theory provides a framework for reasoning about the likelihood of events.

**Sample Space**  $\Omega$ : set of possible outcomes of an experiment. e.g. tossing a die:  $\Omega = \{1, 2, 3, 4, 5, 6\}$

**Event**  $E$ : set of outcomes of an experiment, e.g. event that roll is 4 or the event that sum of 2 rolls is 7.

**Probability of an Outcome** or  $P(\omega)$ : number that satisfies 2 properties:

- For each outcome  $\omega$ ,  $0 \leq P(\omega) \leq 1$
- $\sum_{\omega} P(\omega) = 1$

**Probability of an Event**  $E$ : sum of the probabilities of the outcomes of the experiment:  $P(E) = \sum_{\omega \in E} P(\omega)$

**Descriptive Statistics**

Provides a way of capturing a given data set or sample. There are two main types: central tendency and variability measures.

**Central Tendency**

- Arithmetic Mean:** Useful to characterize symmetric distributions without outliers.  $\mu = \frac{1}{N} \sum x_i$
- Geometric Mean:** Useful for averaging ratios. Always less than arithmetic mean.  $\mu = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N}$
- Median:** Does not take extreme values as much. Useful to avoid distribution or data with outliers.
- Mode:** Most frequent element in a dataset.

**Variability**

- Standard Deviation:** Measures the typical difference between the individual datapoints and the mean.  $\sigma = \sqrt{\frac{1}{N} \sum (x_i - \mu)^2}$

**Neural Network**

Separates data through different hidden layers and relies on nonlinear functions to reach an output.

**Activation Function** - defines a node's output

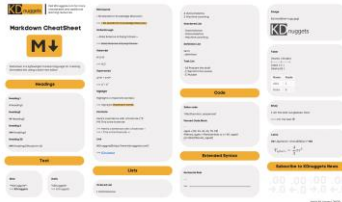
sigmoid	ReLU	Tanh
$\frac{1}{1 + e^{-x}}$	$\max(0, x)$	$\frac{e^x - e^{-x}}{e^x + e^{-x}}$

sigmoid, ReLU, Tanh

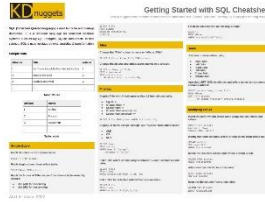
- [Stanford: Super VIP Cheat Sheet](#)
- [Python from Zero to Hero](#)
- [Collection of R Cheat Sheet by posit](#)
- [Data Science Cheat Sheet by Aaron Wang](#)
- [Master NLP](#)
- [NLP Starter Kit](#)
- [Machine Learning Bites by Rishabh Anand](#)
- [Machine Learning Interviews](#)
- [Deep Learning Super VIP](#)



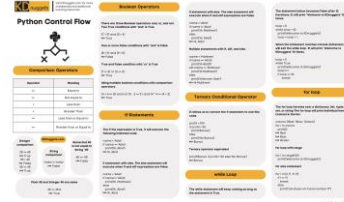
# More Cheat Sheets from KDnuggets



[Markdown](#)



[Getting Started with SQL](#)



[Python Control Flow](#)

## About Author

**Abid Ali Awan** (@labidaliawan) 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.



Visit KDnuggets.com for more cheat sheets and additional learning resources.

### ChatGPT Cheat Sheet

ChatGPT is a large language conversational AI built by OpenAI. It was trained using Reinforcement Learning from Human Feedback similar to GPT-4. ChatGPT understands the prompt and generates detailed responses that can help you with research, coding, and various data science tasks.

- Ideas**
- Dataset Suggestion**
- Suggest Resources**
- AI Testing**
- Career Coaching**
- Coding**

**Unit Test**  
Write a unit test for a function. The test cases are x, should not be null, and only should return numerical value.

**Code Explanation**  
Can you explain what the code is doing? (code snippet)

**Optimize Code**  
Can you improve the time complexity of the code? (code snippet)

**SQL**

**SQL Formatting**  
Format the following SQL code and convert all reserved keywords to uppercase. (code snippet)

**Translate Between DBMS**  
What is the equivalent of PostgresSQL's DATE\_TRUNC for MSSQL?

**Calculate Average**  
Write the SQL code that works for PostgreSQL. I have a table with two columns (date, sales). I would like to calculate an average sales.

**Calculate Runway**  
Write SQL to calculate my runway.

**Spreadsheets**

**Summarize Formula**  
Create a spreadsheet formula to calculate the sum of cells B1 to B20?

**Dummy Data**  
Generate the dummy data for me to use as placeholders in my spreadsheet.

**Tips**  
Give me some tips on how to improve the efficiency of my spreadsheet?

**Date Analysis**

**Generate Data**  
Generate a fake data with 100 rows and 4 columns (id, name, grade, subject).

**Data Cleaning**  
I have a CSV file of customer data. Write Python code for data cleaning.

**Data Exploration**  
I have a dataset of 100 rows and four columns (id, name, grade, subject). Write a code for data visualization and exploration.

**Date Visualization**  
I have a dataset with 100 rows (id, name, grade, subject). Create a regression bar chart of subject vs. grade.

**Machine Learning**

**Train Regression Model**  
You are a data scientist. Write Python code for the 2 feature dataset with columns (brand, speed). Please build a machine learning model that predicts speed.

**Hyperparameter Tuning**  
I have a logistic regression model. Write Python code to tune hyperparameters.

**Influence Data**  
I have a virtualized dataset with target column, 'sales'. I would like to know the top 10 variables that influence my sales?

**Explain the Model**  
I have a model of a logistic model. Write a Python code to explain the output using a series of plots with Shop.

**Research**

**Explain the Concept**  
Explain a concept to an undergraduate on a data science instructor.

**Stakeholders**  
I am an aspiring data science reports to a business stakeholder.

**Summarize the paper**  
Please summarize the paper 'Adding Conditional Control to Text-to-Image Diffusion Models' in simple terms in one paragraph.

**Writing Blog**  
Write an outline for a blog 'Python Lists'.

**Research History**  
Can you research the history of the graph neural network?

**Subscribe to KDnuggets News**



Abid Ali Awan | 2023

## ChatGPT for Data Science Cheat Sheet

