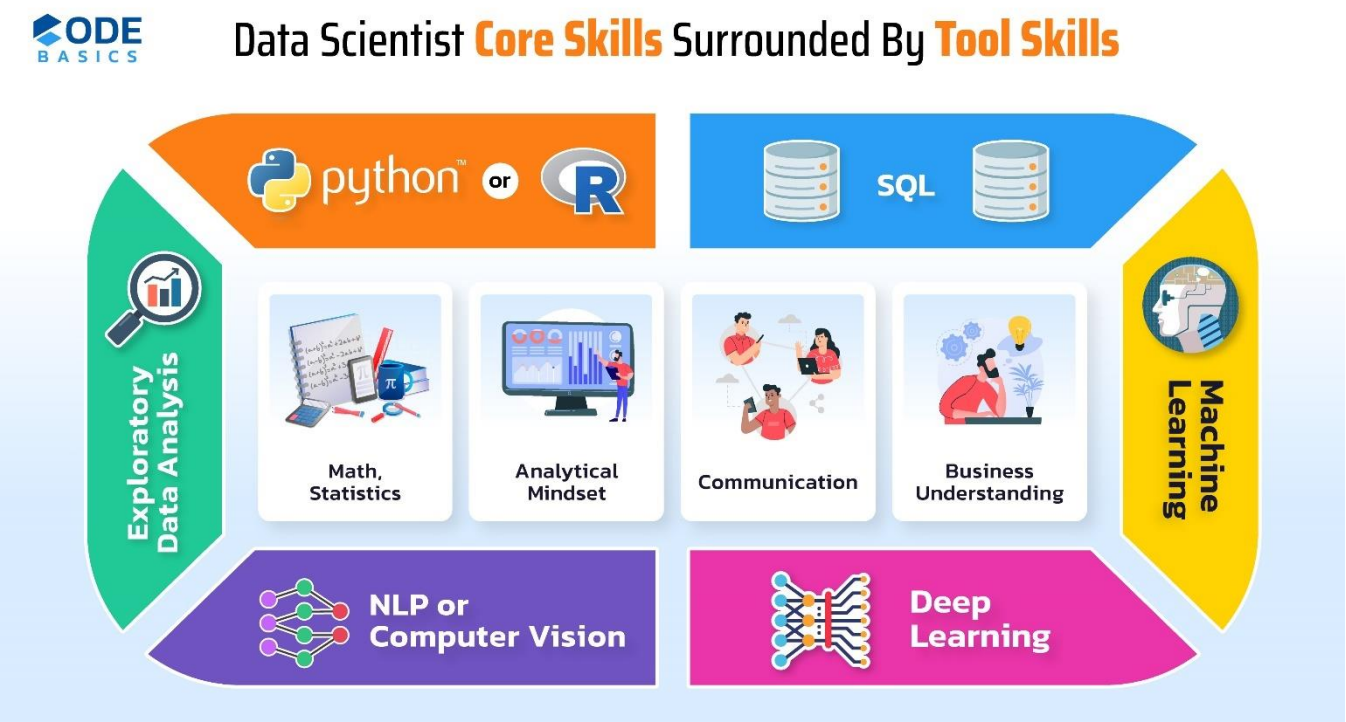
Data Science Roadmap for Beginners

Following is the roadmap to learn **Data Science** skills for a total beginner (**no coding** or **computer science background** needed). It includes FREE learning resources for technical skills (or tool skills) and soft (or core) skills 🏄

Total Duration: **6 Months**

3 hours in Tool Skills + **1 hour** in Core Skills = **4 hours** study Every Day



Week 0: Do Proper Research and protect yourself from SCAMS.

Unfortunately, a lot of systematic scams are happening in ed tech, especially in the data field where aspirants are provided with false promises like a 100% job guarantee or trapped into "Masterclasses" which are nothing but sales pitches to upsell their low-grade courses at exorbitant prices. You need to do complete research about the market and mentors before starting your journey. Providing you the links to a few posts that we have made in this regard which will support your research.

Even though these posts are **NOT** sufficient, do your additional research.

- https://bit.ly/4at9Jaw
- https://bit.ly/477IOOs
- https://bit.ly/3GPD7dp

Week 1 and 2: Python

- **Topics**
  - Variables, Numbers, Strings
  - Lists, Dictionaries, Sets, Tuples
  - If condition, for loop
  - Functions, Lambda Functions
  - Modules (pip install)
  - Read, Write files
  - Exception handling
  - Classes, Objects
- **Learning Resources**
  - Track A (Free)
    - Free Python Tutorials on YouTube (first 16 videos) - https://bit.ly/3X6CCC7
    - Codebasics python HINDI tutorials - https://bit.ly/3vmXrgw
  - Track B (Affordable Fees)
    - Python course: https://codebasics.io/courses/python-for-beginner-and-intermediate-learners

- **LinkedIn - Core Skill**

  ☐ Create a professional-looking LinkedIn profile.

  - Have a clear profile picture and banner image.
  - Add tags such as: Open to work etc.

  ☐ Use this LinkedIn Checklist to create a profile: Click here.

- **Motivation**
  - Physics to Data Scientist Transition -> https://bit.ly/47cA8GU

- **Assignment**

  (Use the assignment tracker: Click here)

  ☐ Track A: Finish all these exercises: https://bit.ly/3k1mof5
  ☐ Track B: Finish exercises and quizzes for relevant topics
  ☐ Create a professional-looking LinkedIn profile.

Week 3: Numpy, Pandas, Data Visualization 📊

- **Tech Skills**
  - **Numpy**
    - numpy YouTube playlist: https://bit.ly/3GTppa8
  - **Pandas, Matplotlib, Seaborn**
    - Go through chapter 3 in this course (entire chapter is free): https://codebasics.io/courses/math-and-statistics-for-data-science

- **Core/Soft Skills**
  - **Linkedin**
    - Start following prominent data science influencers.
      - Daliana Liu: https://www.linkedin.com/in/dalianaliu/
      - Nitin Aggarwal: https://www.linkedin.com/in/ntnaggarwal/
      - Steve Nouri: https://www.linkedin.com/in/stevenouri/
      - Dhaval Patel: https://www.linkedin.com/in/dhavalsays/

        - Increase engagement.
            - Start commenting meaningfully on data science and career-related posts.
            - Helps network with others working in the industry build connections.
            - Learning and brainstorming opportunity.
        - Remember ***online presence is a new form of resume***
    - **Business Fundamentals - Soft Skill**
        - Learn business concepts from ThinkSchool and other YT Case Studies
        - Example: How Amul beat competition: https://youtu.be/nnwqtZiYMxQ
    - **Discord**
        - Start asking questions and get help from the community. This post shows how to ask questions the right way: https://bit.ly/3I70EbI
        - Join codebasics discord server: https://discord.gg/r42Kbuk
- **Assignment**

    ☐ Write meaningful comments on at least **10 data science related LinkedIn posts**
    ☐ Note down your key learnings from **3 case studies** on ThinkSchool and share them with your friend.

Week 4, 5, 6, 7: Statistics and Math for Data Science 👨‍🎓

- **Math and Statistics for Data Science**
    - Topics to Learn
        - Basics: Descriptive vs inferential statistics, continuous vs discrete data, nominal vs ordinal data
        - Basic plots: Histograms, pie charts, bar charts, scatter plot etc.
        - Measures of central tendency: mean, median, mode
        - Measures of dispersion: variance, standard deviation
        - Probability basics
        - Distributions: Normal distribution
        - Correlation and covariance
        - Central limit theorem
        - Hypothesis testing: p value, confidence interval, type 1 vs type 2 error, Z test, t test, ANOVA

- Learning Resources
  - Track A (Free)
    - Learn the above topics from this excellent Khan academy course on statistics and probability.
    - Course link: https://www.khanacademy.org/math/statistics-probability
    - While doing khan academy course, when you have doubts, use statquest YouTube channel: https://www.youtube.com/@statquest
    - Use this free YouTube playlist: https://bit.ly/3QrSXis
  - Track B (Affordable Fees)
    - Khan academy course doesn't have python coding and it is generic. To learn using Python and specifics of applying statistics to data science check this course: https://codebasics.io/courses/math-statistics-for-data-professionals

- **Motivation**
  - Petroleum engineer to data scientist: https://bit.ly/3REsqiL

- **Assignment**
  - ☑ Finish all exercises in this playlist: https://bit.ly/3QrSXis
  - ☑ Finish all exercises in Khan academy course.

Week 8: Exploratory Data Analysis (EDA) 👨‍🎓

- **Exploratory Data Analysis (EDA)**
  - https://www.kaggle.com/code?searchQuery=exploratory+data+analysis
  - Use the above link to search for exploratory data analysis notebooks.
  - Practice EDA using at least 3 datasets.
    - e.g. https://www.kaggle.com/datasets/rishabhkarn/ipl-auction-2023/data

- **Assignment**
  - ☐ Perform EDA (Exploratory data analysis on **at least 2 additional datasets** on Kaggle)

Week 9, 10: SQL 💰

- **Topics**
  - Basics of relational databases.
  - Basic Queries: SELECT, WHERE LIKE, DISTINCT, BETWEEN, GROUP BY, ORDER BY
  - Advanced Queries: CTE, Subqueries, Window Functions
  - Joins: Left, Right, Inner, Full
  - No need to learn database creation, indexes, triggers etc. as those things are rarely used by data scientists.
- **Learning Resources**
  - Track A
    - Khan academy: https://bit.ly/3WFku20
    - https://www.w3schools.com/sql/
    - https://sqlbolt.com/
  - Track B
    - SQL course for data professionals: https://codebasics.io/courses/sql-beginner-to-advanced-for-data-professionals
- **Core/Soft Skills**
  - Presentation skills
    - Death by PowerPoint: https://youtu.be/Iwpi1Lm6dFo
- **Assignment**

☐ Participate in SQL resume project challenge on https://codebasics.io/

- Link: https://codebasics.io/challenge/codebasics-resume-project-challenge/7
- These challenges help you improve technical skills, soft skills and business understanding.

☐ Make a LinkedIn post with a submission of your resume project challenge

- Sample post: https://bit.ly/48Bg5mB
- Codebasics is promoting winning entries to employers. This way you can get interview calls. We do this in two ways:
  - We have a database of employers hiring for data analyst positions. We send first 10 or 20 profiles based on their performance.
  - LinkedIn post by Dhaval (who has more than 100k followers and some of them are HR managers, data analytics senior managers): https://bit.ly/3jnni5c

Week 11, 12, 13, 14, 15: Machine Learning ⚒️

- **Machine Learning: Preprocessing**
  - Handling NA values, outlier treatment, data normalization
  - One hot encoding, label encoding
  - Feature engineering
  - Train test split
  - Cross validation
- **Machine Learning: Model Building**
  - Types of ML: Supervised, Unsupervised
  - Supervised: Regression vs Classification
  - Linear models
    - Linear regression, logistic regression
    - Gradient descent
  - Nonlinear models (tree-based models)
    - Decision tree
    - Random forest
    - XGBoost
  - Model evaluation
    - Regression: Mean Squared Error, Mean Absolute Error, MAPE
    - Classification: Accuracy, Precision-Recall, F1 Score, ROC Curve, Confusion matrix
  - Hyperparameter tunning: GridSearchCV, RandomSearchCV
  - Unsupervised: K means, Hierarchical clustering, Dimensionality reduction (PCA)
- **Learning Resources**
  - YouTube playlist (more than 2 million views): https://bit.ly/3io5qqX
    - First 21 videos
  - Feature engineering playlist: https://bit.ly/3IFa3Yf
- **Core/Soft Skills**
  - **Project Management**
    - Scrum: https://scrumtrainingseries.com/
    - Kanban: https://youtu.be/jf0tlbt9lx0
    - Tools: JIRA, Notion

- **Motivation**
  - How Kaggle helped this person become ML engineer**:** https://bit.ly/3RFVruy
- **Assignment**
  - ☐ Complete all exercises in ML playlist: https://bit.ly/3io5qqX
  - ☐ Work on **2 Kaggle ML notebooks**
  - ☐ Write **2 LinkedIn posts** on whatever you have learnt in ML
  - ☐ Discord: Help people with **at least 10 answers**

Week 16, 17, 18: Machine Learning Projects with Deployment 🎯

- You need to finish **two** end to end ML projects. One on **Regression**, the other on **Classification**
- Regression Project: Bangalore property price prediction
  - YouTube playlist link: https://bit.ly/3ivycWr
  - Project covers following
    - Data cleaning
    - Feature engineering
    - Model building and hyper parameter tuning
    - Write flask server as a web backend
    - Building website for price prediction
    - Deployment to AWS
- Classification Project: Sports celebrity image classification
  - YouTube playlist link: https://bit.ly/3ioaMSU
  - Project covers following
    - Data collection and data cleaning
    - Feature engineering and model training
    - Flask server as a web backend
    - Building website and deployment
- **ATS Resume Preparation**
  - Resumes are dying but not dead yet. Focus more on online presence.
  - Here is the resume tips video along with some templates you can use for your data analyst resume: https://www.youtube.com/watch?v=buQSI8NLOMw
  - Use this checklist to ensure you have the right ATS Resume: Check here.

- **Portfolio Building Resources:**
  You need a portfolio website in 2024. You can build your portfolio by using these free resources.
  - GitHub
    - Upload your projects with code on github and using github.io create a portfolio website
    - Sample portfolio website: http://rajag0pal.github.io/
  - Linktree
    - Helpful to add multiple links in one page.

- **Assignment**
  - In above two projects make following changes

    ☐ Use **FastAPI** instead of **flask**. FastAPI tutorial: https://youtu.be/Wr1JjhTt1Xg

    ☐ **Regression project**: Instead of property prediction, take any other project of your interest from Kaggle for regression

    ☐ **Classification project**: Instead of sports celebrity classification, take any other project of your interest from Kaggle for classification and build end to end solution along with **deployment to AWS or Azure**

    ☐ Add a link of your projects in your resume and LinkedIn.

    (Tag Codebasics, Dhaval Patel and Hemanand Vadivel with the hashtag #dsroadmap24 so we can engage to increase your visibility)

Week 19, 20, 21: Deep Learning 🤖

- **Topics**
  - What is a neural network? Forward propagation, back propagation
  - Building multilayer perceptron
  - Special neural network architectures
    - Convolutional neural network (CNN)
    - Sequence models: RNN, LSTM

- **Learning Resources**
  - Deep Learning playlist (tensorflow): https://bit.ly/3vOZ3zV
  - Deep learning playlist (pytorch): https://bit.ly/3TzDbWp
  - End to end potato disease classification project: https://bit.ly/3QzkVJi

- **Assignment**
  - ☐ Instead of potato plant images use tomato plant images or some other image classification dataset.
  - ☐ Deploy to Azure instead of GCP.
  - ☐ Create a presentation as if you are presenting to stakeholders and upload video presentation on LinkedIn.

Week 22, 23, 24: NLP or Computer Vision

- Many data scientists choose a specialized track which is either NLP or Computer vision. You don't need to learn both.
- **Natural Language Processing (NLP)**
  - Topics
    - Regex
    - Text presentation: Count vectorizer, TF-IDF, BOW, Word2Vec, Embeddings
    - Text classification: Naïve Bayes
    - Fundamentals of Spacy & NLTP library
    - One end to end project
  - Learning Resources
    - NLP YouTube playlist: https://bit.ly/3XnjfEZ
- **Computer Vision (CV)**
  - Topics
    - Basic image processing techniques: Filtering, Edge Detection, Image Scaling, Rotation
    - Library to use: OpenCV
    - Convolutional Neural Networks (CNN) – Already covered in deep learning.
    - Data preprocessing, augmentation – Already covered in deep learning.
- **Assignment**
  - ☐ NLP Track: Complete exercises in this playlist: https://bit.ly/3XnjfEZ

Week 25 onwards.... 😃 😃 😃

- More projects 📽️
- Online brand building through LinkedIn, Kaggle, Discord, Opensource contribution 👥
- Job application and Success 🚀

Tips of effective learning 🔥

- **Spend less time in consuming information, more time in**
    - Digesting
    - Implementing
    - Sharing
- **Group learning**
    - Use **partner-and-group-finder** channel on codebasics discord server for group study and hold each other accountable for the progress of your study plan. Here is the discord server link: https://discord.gg/r42Kbuk

FAQs 🔭

- **Do I need to learn cloud tech (Amazon sagemaker, Azure etc.)?**
    - Big cloud service providers such as AWS, Azure, Google Cloud have their own ML offering such as Amazon Sagemaker in case of AWS. As a fresher it is ok if you are not familiar with these cloud platforms but once you have some experience it is good to have experience and know-how of at least one cloud ML platform.

- **Do I need to learn Gen AI?**
    - Gen AI is a fancy topic and majority of the junior data science positions do not demand gen ai skills. In case you have additional time and If you want to learn a famous framework for building GenAI apps called langchain then here is the playlist: https://bit.ly/3RYpxuw

- **How about BI tool (Power BI or Tableau)**
    - BI tools nowadays are mainly used by BI developers, data analysts etc. Hence it is ok if you don't learn them as a data scientist. Majority of the time whenever data scientists have a need of BI dashboards they will take help of BI or data analyst teams. In small organizations however, sometimes data scientist work on building BI dashboards but in general you should not worry about learning BI tool for a data scientist career