

Data Science

Module 1: Python Fundamentals

- Python installation
- Variables & data types
- Lists, tuples, dictionaries, sets
- Operators
- Conditional statements
- Loops (for, while)
- Functions
- Modules & packages
- File handling

Practice:

- 30–40 Python coding exercises
- Simple automation scripts

Module 2: Python for Data Analysis

Libraries:

- NumPy
- Pandas

Topics:

- NumPy arrays
- Mathematical operations
- Pandas Series & DataFrame
- Data loading (CSV, Excel, JSON)
- Data filtering
- Data aggregation
- GroupBy operations

Practice:

- Analyze small datasets

Module 3: Statistics & Mathematics

- Mean, median, mode
- Variance & standard deviation
- Probability basics
- Normal distribution
- Hypothesis testing
- Correlation & covariance
- Sampling methods

Use libraries:

- SciPy
- Statsmodels

Module 4: Data Visualization

Use libraries:

- Matplotlib
- Seaborn

Topics:

- Line charts
- Bar charts
- Histograms
- Scatter plots
- Box plots
- Heatmaps
- Data storytelling

Tools:

- Tableau
- Power BI

Module 5 : SQL for Data Science

Learn databases using:

- MySQL
- PostgreSQL

Topics:

- Database basics
- SELECT queries
- Filtering data
- GROUP BY
- JOINS
- Subqueries
- Window functions
- Data cleaning with SQL

Module 6 : Exploratory Data Analysis (EDA)

Topics :

- Data cleaning
- Handling missing values
- Outlier detection
- Feature engineering
- Data transformation
- Correlation analysis

Tools :

- Pandas
- Seaborn
- Matplotlib

Practice:

- Full EDA on multiple datasets.

Module 7 : Machine Learning Fundamentals

Topics :

- What is machine learning
- Supervised vs unsupervised learning
- Train-test split
- Model training
- Model evaluation

Algorithms:

- Linear regression
- Logistic regression
- Decision trees
- KNN

Metrics:

- Accuracy
- Precision
- Recall
- F1 score

Module 8 : Advanced Machine Learning

Algorithms:

- Random Forest
- Gradient Boosting
- XGBoost
- Support Vector Machine

Techniques:

- Cross validation
- Hyperparameter tuning
- Grid search
- Feature selection
- Dimensionality reduction (PCA)

Libraries:

- XGBoost

Module 9 : Deep Learning

Use:

- TensorFlow
- PyTorch

Topics:

- Neural networks
- Activation functions
- Backpropagation
- CNN (image tasks)
- RNN (sequence tasks)
- LSTM

- Applications:
- Image classification
- Text analysis

Module 10 : Natural Language Processing

Libraries:

- NLTK
- spaCy

Topics:

- Text preprocessing
- Tokenization
- Stopword removal
- Sentiment analysis
- Topic modeling

- Applications:
- Chatbots
- Text classification

Module 11 : Big Data & Deployment

Big Data tools:

- Apache Hadoop
- Apache Spark

Deployment:

Model APIs

Cloud deployment

Version control:

- GitHub

Module 12 : Advanced Machine Learning

Tasks:

- Build portfolio
- Publish projects
- Write case studies
- Prepare resume
- Practice interviews

- Create projects on:
 - GitHub
 - Kaggle

Example platform:

Kaggle

20 Real Data Science Projects That Impress Recruiters

Business Projects

1. Customer churn prediction
2. Sales forecasting system
3. Product recommendation system
4. Customer segmentation using clustering
5. Marketing campaign optimization

Finance Projects

6. Credit card fraud detection
7. Loan approval prediction
8. Stock price prediction
9. Cryptocurrency trend analysis

Healthcare Projects

10. Disease prediction system
11. Diabetes prediction model
12. Medical image classification Datasets from World Health Organization

Entertainment Projects

13. Movie recommendation system
14. Netflix content analysis

Company example:

- Netflix

Transportation Projects

15. Ride demand prediction for Uber

Social Media Projects

16. Twitter sentiment analysis
17. Fake news detection

HR Analytics

18. Employee attrition prediction

Advanced AI Projects

19. Chatbot using NLP
20. Image classification using CNN

Top 100 Data Science Interview Questions

Python for Data Science (1–20)

Practice using Python.

1. What is Python?
2. Why is Python popular for Data Science?
3. What are Python data types?
4. Difference between list and tuple.
5. What is dictionary?
6. What is list comprehension?
7. What is lambda function?
8. What is a generator?
9. What is iterator?
10. What is exception handling?
11. What is virtual environment?
12. What is pip?
13. What is file handling?
14. What is JSON?
15. What is API?
16. What is multithreading?
17. What is multiprocessing?
18. What is GIL?
19. What is object-oriented programming?
20. What is a decorator?

Libraries used in Data Science:

- NumPy
- Pandas

Statistics & Mathematics (21–40)

Statistics is a core part of Data Science.

21. What is mean?
22. What is median?
23. What is mode?
24. What is variance?

25. What is standard deviation?
26. What is probability?
27. What is correlation?
28. What is covariance?
29. What is hypothesis testing?
30. What is p-value?
31. What is confidence interval?
32. What is normal distribution?
33. What is sampling?
34. What is central limit theorem?
35. What is A/B testing?
36. What is Bayesian statistics?
37. What is regression analysis?
38. What is overfitting?
39. What is underfitting?
40. What is bias vs variance?

Machine Learning (41–70)

Using libraries like Scikit-learn.

Basics

41. What is machine learning?
42. Types of machine learning.
43. Supervised learning.
44. Unsupervised learning.
45. Reinforcement learning.

Algorithms

46. Linear regression
47. Logistic regression
48. Decision tree
49. Random forest
50. K-nearest neighbors
51. Naive Bayes
52. Support vector machine

Clustering

53. K-means clustering
54. Hierarchical clustering

Model Evaluation

55. Accuracy
56. Precision
57. Recall
58. F1 score
59. Confusion matrix

Feature Engineering

60. Feature scaling
61. Standardization
62. Normalization

63. Handling missing values
64. Encoding categorical variables

Advanced ML

65. Cross validation
66. Grid search
67. Hyperparameter tuning
68. Feature selection
69. Dimensionality reduction
70. Principal Component Analysis (PCA)

Deep Learning (71–85)

Libraries like:

- TensorFlow
- PyTorch

71. What is deep learning?
72. What is neural network?
73. What is a neuron?
74. What is activation function?
75. What is ReLU?
76. What is sigmoid?
77. What is softmax?
78. What is backpropagation?
79. What is gradient descent?
80. What is CNN?
81. What is RNN?
82. What is LSTM?
83. What is dropout?
84. What is batch normalization?
85. What is transfer learning?

Data Processing & Visualization (86–100)

Visualization libraries:

- Matplotlib
- Seaborn

86. What is data cleaning?
87. How to handle missing values?
88. What is exploratory data analysis (EDA)?
89. What is feature engineering?
90. What is outlier detection?
91. What is data normalization?
92. What is data standardization?
93. What is data pipeline?
94. What is big data?
95. What is Hadoop?
96. What is Spark?
97. What is model deployment?
98. What is A/B testing in data science?

99. What is MLOps?

100. What is model monitoring?

50 Datasets for Data Science Practice

Datasets are available on Kaggle and UCI Machine Learning Repository.

Business / Marketing

1. E-commerce sales dataset
2. Customer segmentation dataset
3. Retail store dataset
4. Walmart sales dataset
5. Marketing campaign dataset
6. Customer lifetime value dataset
7. Customer churn dataset
8. Product recommendation dataset
9. Sales forecasting dataset
10. Retail inventory dataset

Media / Entertainment

11. Netflix movies dataset
 12. IMDB movies dataset
 13. Spotify songs dataset
 14. YouTube trending dataset
 15. Movie recommendation dataset
- Companies include Netflix, Spotify, and YouTube.

Finance

16. Stock price dataset
17. Cryptocurrency dataset
18. Credit card fraud dataset
19. Loan prediction dataset
20. Bank marketing dataset

Healthcare

21. COVID-19 dataset
 22. Diabetes dataset
 23. Heart disease dataset
 24. Medical insurance dataset
 25. Cancer dataset
- Data sources include the World Health Organization.

HR Analytics

26. Employee attrition dataset
27. HR analytics dataset
28. Salary dataset
29. Job satisfaction dataset
30. Workforce diversity dataset

Transportation

31. Uber trips dataset
 32. Taxi trips dataset
 33. Airline delays dataset
 34. Traffic accidents dataset
 35. Flight dataset
- Company example: Uber

General Data Science

36. Global temperature dataset
37. Population dataset
38. Housing prices dataset
39. Crime dataset
40. Education dataset

Technology / Internet

41. Mobile usage dataset
42. App store dataset
43. Social media dataset
44. Website traffic dataset
45. Online reviews dataset

Gaming

46. Video game sales dataset
47. Steam games dataset
48. Game ratings dataset
49. Esports dataset
50. Gaming revenue dataset