

Data Science Foundations

Data Science · Answer Key · 20 Questions

1. Which programming language is most commonly cited as the primary language for data science and machine learning tasks?

- A) Java
- B) Python**
- C) HTML
- D) Assembly

2. What is the primary purpose of a 'train-test split' in machine learning?

- A) To format the data for Excel
- B) To evaluate model performance on unseen data**
- C) To increase the size of the dataset
- D) To delete outliers from the database

3. In a confusion matrix, what does a 'False Positive' represent?

- A) Correctly identifying a positive case
- B) Failing to identify a positive case
- C) Predicting a positive result when the actual result is negative**
- D) Predicting a negative result when the actual result is positive

4. Which library is widely used in Python for data manipulation and analysis, providing structures like DataFrames?

- A) Pandas**
- B) PyQt
- C) Tkinter
- D) Requests

5. What is the term for a subset of artificial intelligence that involves training algorithms to learn patterns from data?

- A) Cloud Computing
- B) Machine Learning**
- C) Quantum Physics
- D) Blockchain

6. What is the main goal of 'Data Cleaning' in the data science pipeline?

- A) To encrypt the data for security
- B) To visualize the data in charts
- C) To fix or remove incorrect, corrupted, or duplicate data**
- D) To compress the data for faster storage

7. Which statistical measure represents the most frequently occurring value in a dataset?

- A) Mean
- B) Median
- C) Mode**
- D) Range

8. What does SQL stand for in the context of database management?

- A) Standard Query Language
- B) Structured Query Language**
- C) Simple Quality Language
- D) System Query Logic

9. In supervised learning, what kind of data is required to train a model?

- A) Unlabeled data
- B) Labeled data**
- C) Random noise
- D) Deleted files

10. Which visualization tool is best suited for showing the distribution of a single continuous variable?

- A) Histogram**
- B) Pie chart
- C) Line chart
- D) Network graph

11. What is the 'Overfitting' phenomenon in machine learning?

- A) When a model performs poorly on training data
- B) When a model learns the noise in training data too well, failing to generalize**
- C) When a model is too simple to capture the pattern
- D) When a model uses too much RAM

12. Which of the following is considered a 'categorical' variable?

- A) Height in centimeters
- B) Blood type (A, B, AB, O)**
- C) Time taken to run a race
- D) Annual income

13. What is 'Data Mining'?

- A) Manually typing data into a spreadsheet
- B) The process of discovering patterns in large data sets**
- C) The physical extraction of minerals
- D) Hardware installation for servers

14. Which Python library is the standard for creating static, animated, and interactive visualizations?

- A) Matplotlib**
- B) TensorFlow
- C) PyTorch
- D) Flask

15. What is the 'null hypothesis' in statistical testing?

- A) The assumption that there is no effect or no difference**
- B) The assumption that the data is corrupted
- C) The initial guess of the researcher
- D) The final conclusion of an experiment

16. What type of chart is best for showing the relationship between two continuous variables?

- A) Bar chart
- B) Scatter plot**
- C) Histogram
- D) Area chart

17. In data science, what is a 'Feature'?

- A) The final output of a model
- B) An input variable used in making predictions**
- C) The name of the data scientist
- D) The computer hardware used

18. What is 'Big Data' typically characterized by?

- A) Volume, Velocity, and Variety**
- B) Speed, Memory, and Pixels
- C) Color, Weight, and Sound
- D) Price, Size, and Name

19. Which algorithm is a popular choice for classification tasks and is based on the idea of drawing a boundary between classes?

- A) Logistic Regression**
- B) Linear Regression
- C) K-Means Clustering
- D) Principal Component Analysis

20. What is the purpose of 'Normalization' in data preprocessing?

- A) To scale numerical values to a common range**
- B) To translate text into another language
- C) To convert all data to binary
- D) To increase the data file size