

Strong answers stay concrete. Refer to projects you built. Share numbers. Explain trade offs. Show how your model improved revenue, reduced risk, or saved time. Structured preparation makes a difference. If you want guided practice, explore [Data Science Training in Chennai](#) to strengthen your technical base and project experience before attending interviews.

List of Data Science Interview Questions for Freshers

1. What is data science?
2. What is the difference between supervised and unsupervised learning in data science?
3. Define survivorship bias.
4. What is a random forest model?
5. Define Recommender Systems
6. Define feature vectors.
7. What is logistic regression?
8. Explain a linear model or linear regression, along with its limitations.
9. What is the purpose of A/B testing?
10. Define the law of large numbers
11. Define confounding variables

Data Science Technical Interview Questions and Answers for Freshers

Data science technical interviews for freshers test your understanding of core concepts and your ability to apply them. Interviewers focus on fundamentals in statistics, Python, SQL, and machine learning. You must explain ideas clearly and solve small practical problems. Below are common technical questions with direct answers to help you prepare.

1. What is data science?

Data Science is the collection of tools, algorithms, scientific methods, and principles used to explore the hidden patterns from structured and unstructured raw data. These insights help business people make better decisions according to trending business growth.

2. What is the difference between supervised and unsupervised learning in data science?

In Supervised learning, the input data is labeled and uses a training data set. Supervised learning is used for prediction and allows classification and regression.

In unsupervised learning, the input data is unlabeled, and the input data is set for performing the analysis. It enables classification, density estimation, and dimension reduction.

3. Define survivorship bias.

Survivorship bias is a logical error that focuses on aspects that support a survival process by overlooking them because they lack prominence. It leads to wrong conclusions in many ways.

4. What is a random forest model?

A random forest is developed on several decision trees and if we split the data into various packages and make a decision tree in all groups of data, then the random forest puts trees together for fetching useful insights.

5. Define Recommender Systems

The recommender systems are used to predict the user rate for a particular product according to their preferences. It happens in two different areas, such as collaborative filtering and content-based filtering.

6. Define feature vectors.

A feature vector is an n-dimensional of numerical features that means an object. It is implemented in the machine learning process to represent symbolic or numeric characteristics and objects mathematically, as it is easy to analyze.

**Take your Knowledge
Test Report**

Check your Score



7. What is logistic regression?

Logistic regression is the technique used to forecast the binary outcome from a linear combination of predictor variables.

8. Explain a linear model or linear regression, along with its limitations.

Linear regression is the method that uses the largest square by connecting a line through plotted data points. This line is positioned to minimize the distance of all data points; this distance is called “residues” or “errors.”.

The assumption of linearity of the errors can't be used for counting outcomes and binary outcomes, and overfitting problems are the drawbacks of linear models or linear regression.

9. What is the purpose of A/B testing?

A/B testing is the testing of statistical hypotheses for randomized experiments with two variables, such as A and B. The goal of A/B testing is to detect changes to a web and maximize the outcome of a particular strategy.

10. Define the law of large numbers

The law of large numbers is a theorem that defines the result of performing the same experiment every time. It forms as per the frequency-style thinking and it denotes the sample mean, variance, and sample standard deviation that converge to give the estimation.

11. Define confounding variables

Confounding variables are extraneous variables in a statistical model that combine directly or indirectly with dependent and independent variables. The estimate will fail to account for the confounding factor.

Check out our [Machine Learning Training in Chennai](#) to learn algorithms, build models, and gain hands-on project experience.

List of Data Science Interview Questions for Experienced

12. What are the steps to follow for creating a decision tree?
13. What is selection bias and what are the types of selection biases?
14. What are the steps involved in developing a random forest model?
15. Explain collaborative filtering and content-based filtering.
16. Explain univariate, bivariate, and multivariate.
17. How do you avoid overfitting the model?
18. What are the feature selection methods applied for selecting the right variables?
19. What are the steps to maintain a deployed model?
20. What does the p-value indicate?
21. Define Cross-Validation
22. Describe star schema
23. How often must an algorithm be updated?
24. What are Eigenvalue and Eigenvector?

Data Science Technical Interview Questions and Answers for Experienced

Data science interviews for experienced professionals focus on depth, scale, and impact. Employers expect you to design systems, improve model performance, and connect analytics to business results. You must show project ownership, clear reasoning, and measurable outcomes. Below are common interview questions with structured answers.

12. What are the steps to follow for creating a decision tree?

Step 1: Take up the entire data set as input

Step 2: Calculate the entropy of predictor attributes and target variable

Step 3: Calculate the gained information of all attributes

Step 4: Choose the attribute that has the highest value gain as the root node

Step 5: Repeat the process until the decision node of all branches is finalized.

13. What is selection bias and what are the types of selection biases?

Selection bias is the type of error that occurs by individuals, groups, or analyzed data when it is done without achieving proper randomization.

It is also referred to as the effect that the population is not analyzed to ensure a perfect result. It occurs when people volunteer to study analytics.

The four types of selection bias are sampling bias, time interval bias, data bias, and attrition bias.

- Unpredictable sample selection leads to systemic errors known as **sampling bias**.
- **A time interval** is a trial that terminates at an extreme value.
- **The data bias** is the conclusion.
- **Attrition** is the loss of participants.

14. What are the steps involved in developing a random forest model?

Step 1: Randomly select 'k' features from the total number of 'm' features when $k \ll m$

Step 2: Calculate the D node among 'k' features using the best-split point

Step 3: Split the nodes into sub-nodes using the best split

Step 4: Repeat the steps until leaf nodes are formed

Step 5: Develop a forest by repeating the above steps one to 'n' times to generate 'n' number of trees.

15. Explain collaborative filtering and content-based filtering.

Collaborative Filtering: The marketer recommends products based on the similar interests of users.

Ex: *When the user checks something on Amazon, it shows "Users who bought this also bought..." with recommendations.*

Content-based Filtering: The app shows the recommendation of the same properties as per the user's interests.

Ex: *Spotify used to recommend music according to the latest listening of its users.*

16. Explain univariate, bivariate, and multivariate.

Univariate: Univariate data contains only one variable and the purpose of this analysis is to define the data to extract the pattern within it.

This pattern will include median, mean, mode, range or dispersion, minimum, and maximum.

Ex: Height report of students

Bivariate: Bivariate data includes two different variables.

This kind of data analysis deals with causes and relationships and it is used to determine the relationship between two given variables.

The relationship is visible for users to make better decisions.

Ex: the analysis of temperature and the sales of ice cream.

Multivariate: Multivariate data includes three or more variables and it is categorized as per the total number of dependent variables.

It will be studied by fetching the conclusions through mean, median, minimum, maximum, and dispersion or range.

Ex: Analysis of the price attributes of a house.

17. How do you avoid overfitting the model?

Overfitting is the model that sets up every small amount of data and avoids the bigger picture. The following methods are used to avoid overfitting:

- Keep the sample model that takes fewer variables into account and removes the noise of training data
- Utilize cross-validation techniques such as k-fold cross-validation
- Apply regularization techniques such as LASSO that penalize the models that have the possibility of overfitting

**Take your Knowledge
Test Report**

Check your Score



18. What are the feature selection methods applied for selecting the right variables?

There are two main methods filter methods and wrapper methods.

- Filter methods include linear discrimination analysis, ANOVA, and Chi-Square.
- The wrapper methods include FSelection, Backward Selection, and Recursive Feature Elimination.

The Wrapper Methods requires high-end computer systems and it is labor-intensive to perform the analysis.

19. What are the steps to maintain a deployed model?

Step 1: Monitor: Continuous monitoring of all models leads to determining performance accuracy.

Step 2: Evaluate: Evaluation metrics of the current model should be calculated to determine if any new algorithm is required.

Step 3: Compare: Comparing the new models to each other brings out the performance of the best model.

Step 4: Rebuild: The best-performing model will be rebuilt as per the current state of data.

20. What does the p-value indicate?

If p-value < 0.05: It indicates strong evidence against the null hypothesis, and we can reject them

If the p-value > 0.05: it indicates weak evidence against the null hypothesis, and we can accept them

If the p-value is at 0.05: it means it could be either way, as it is considered marginal.

21. Define Cross-Validation

Cross-validation is one of the model validation techniques used for evaluating how the outcomes of a statistical analysis will generalize to a single data set.

It is mainly applied to backgrounds to estimate the accuracy of the model and test the training phase to limit problems such as overfitting and insight-gaining.

22. Describe star schema

The star schema is a traditional database schema that has a central table.

Satellite tables map IDs to physical descriptions and are connected to the central fact tables using ID fields. Tables are also referred to as lookup tables and are used for real-time applications as they save memory storage.

This star schema includes several layers of summarization to recover the information quickly and accurately.

23. How often must an algorithm be updated?

When the underlying data source changes, the model changes as data flows through the infrastructure, or there is an instance of non-stability, we need to adjust the algorithm.

24. What are Eigenvalue and Eigenvector?

- Eigenvalues are the directions of a particular linear transformation served through compressing, stretching, and flipping.
- Eigenvectors are the understanding of linear transformations used to calculate a correlation or covariance matrix.

Conclusion

A data science interview brings your preparation into the open. You sit down and your thinking becomes visible through your answers. Clear basics in statistics, SQL, Python, and model building matter more than memorised lines. Interviewers pay attention to how you structure a solution and how you support claims with numbers from your work.

Preparation works best when it is steady and practical. Revisit topics where you hesitate. Solve problems without looking at notes. Speak your answers out loud. Before you attend an interview, go through the [Data Science](#)

[Syllabus](#) once more and confirm you have covered each core area with confidence.