DATA SCIENCE INTERVIEW QUESTIONS FOR FRESHERS

Habib Shaikh
AI Expert

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is Data Science?

- Data Science is a multidisciplinary field that blends various techniques to analyze large data sets and derive actionable insights.

Key Components

- **Data Collection**: Gathering raw data from various sources.
- **Data Cleaning**: Removing errors and inconsistencies for accuracy.
- **Data Storage**: Warehousing and structuring data for accessibility.
- **Analysis Methods**: Applying statistical, mathematical, and machine learning algorithms.
- **Visualization**: Presenting findings through charts and graphs for better understanding.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Define the terms KPI, lift, model fitting, robustness, and DOE.

- **KPI (Key Performance Indicator)**: A measure of how effectively a business is achieving its objectives.
- **Lift**: A metric used to compare the performance of a target model against a random choice model. Lift quantifies how well the model predicts compared to no model at all.
- **Model Fitting**: Refers to how accurately a model corresponds to the given data observations.
- **Robustness**: The ability of a system to handle variability or noise effectively.
- **DOE** (Design of Experiments): A structured approach to investigating and explaining the variation of information under assumed conditions by reflecting variables.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is the difference between data analytics and data science?

- Data science involves transforming data using advanced analysis methods to extract insights, which can then be applied in various business contexts. In contrast, data analytics focuses on examining existing data to validate hypotheses and support decision-making.
- Data science is forward-looking, involving predictive modeling and innovations for future problem-solving, while data analytics is more focused on understanding past trends for immediate business decisions. Data science encompasses a wider scope of techniques, tools, and methodologies, while data analytics typically deals with more specific, concentrated issues.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What are some sampling techniques and the advantages of sampling?

- Due to the challenges of analyzing large datasets in their entirety, sampling allows for the selection of representative data points for analysis. Two main categories of sampling techniques are:
- **Probability Sampling**: Techniques like cluster sampling, simple random sampling, and stratified sampling.
- **Non-Probability Sampling**: Methods such as quota sampling, convenience sampling, and snowball sampling.
- Sampling ensures that analysis remains manageable while still representing the whole dataset.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

List the conditions for overfitting and underfitting.

- **Overfitting**: Occurs when a model works well on training data but fails on new data. This happens due to low bias and high variance, often seen in decision trees.
- **Underfitting**: Happens when a model is too simple to capture the underlying data relationships, leading to poor performance even on training data. It results from high bias and low variance, typical of linear regression.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Differentiate between long and wide format data.

- **Long Format Data**: Each row represents one data point per subject, with multiple rows for each subject. Common in R analysis and data logging.
- **Wide Format Data**: Multiple observations per subject are stored in separate columns. This format is typically used for repeated measures ANOVA in statistical packages.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What are Eigenvectors and Eigenvalues?

- **Eigenvectors** are unit vectors whose magnitude equals one and are used in eigen decomposition of matrices.
- **Eigenvalues** are coefficients applied to these vectors, altering their magnitude. These concepts are essential in machine learning techniques like PCA, where they help identify significant patterns in data.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What do high and low p-values mean?

- A p-value indicates the likelihood that the observed results are due to chance.
- **Low p-value (≤ 0.05)**: Suggests that the null hypothesis can be rejected, indicating the result is statistically significant.
- **High p-value (≥ 0.05)**: Indicates the null hypothesis remains valid, suggesting the observed data is likely due to random variation.
- **p-value of 0.05**: Implies a borderline result where the hypothesis could go either way.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

When is resampling done?

- Resampling is employed to assess the stability and accuracy of a model, typically by training the model on various subsets of the data. It helps quantify uncertainties, ensuring the model can handle diverse patterns in the data and is validated through different random selections.

What is imbalanced data?

- Imbalanced data refers to a dataset where certain categories or classes are underrepresented, leading to biases in model predictions and accuracy issues.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Are there differences between expected value and mean value?

- While both represent central tendencies, the expected value pertains to random variables, whereas the mean is related to probability distributions. The expected value is often used in stochastic processes, while the mean is a general statistic for averaged data.

What is Survivorship Bias?

- Survivorship bias is the error made when focusing on successful subjects and ignoring those that failed, leading to false conclusions based on incomplete data. This can distort analyses and create misleading interpretations.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is Gradient and Gradient Descent?

- **Gradient**: A vector indicating how much the output of a function changes relative to changes in its input. It represents the slope of the function.
- **Gradient Descent**: A technique used to minimize a function by iteratively adjusting the input in the direction of the steepest decrease, often applied to minimize loss functions in machine learning.

Define confounding variables.

- Confounding variables are extraneous factors that affect both the independent and dependent variables, causing spurious relationships that can distort the conclusions drawn from data analysis.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Explain the bias-variance trade-off.

- **Bias**: Error introduced by oversimplifying the model. Low-bias models like decision trees are complex, while high-bias models like linear regression are simpler.
- **Variance**: Error due to model complexity, where overly complex models may overfit and perform poorly on unseen data.
- The trade-off suggests that as model complexity increases, bias decreases, but variance may increase, leading to overfitting. The goal is to find an optimal balance for model accuracy.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Define the confusion matrix.

- A confusion matrix is a 2x2 matrix used in classification tasks to evaluate a model's performance. It shows the number of correct and incorrect predictions, broken down into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). From these, metrics like accuracy, precision, recall, and F-score are derived.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is logistic regression? Provide an example of its use.

- Logistic regression is a technique used to model binary outcomes using a linear combination of predictor variables. For example, predicting election outcomes based on factors like campaign spending and political history.

What is Linear Regression and its drawbacks?

- Linear regression models the relationship between a dependent variable and independent variables. Drawbacks include assumptions of linearity, inability to model binary outcomes, and vulnerability to overfitting.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is Random Forest and how does it work?

- Random Forest is an ensemble learning technique that builds multiple decision trees and combines their outputs to improve classification accuracy. Each tree is trained on random subsets of data, and predictions are made based on the majority vote across all trees.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

Calculate the chance of seeing a shooting star within an hour given a 0.2 probability every 15 minutes.

- With a 0.2 chance of seeing a shooting star in 15 minutes, the probability of not seeing any in 15 minutes is 0.8. Over an hour (four 15-minute intervals), the chance of seeing no stars is $0.8^4 \approx 0.40$. Thus, the chance of seeing at least one star is 1 - 0.40 = 0.60 or 60%.

DATA SCIENCE

Habib Shaikh
AI Expert

INTERVIEW QUESTIONS FOR FRESHERS

What is deep learning and its difference from machine learning?

- Deep learning is a subset of machine learning that uses layered neural networks to process and learn from data. Unlike traditional machine learning, which uses simpler models, deep learning simulates the human brain's structure for higher accuracy and feature extraction.