**Data Science Sample Interview Questions – Theory**

1. **What are the key features of Python that make it suitable for data science?**

Python is open-source, easy to learn, has a large collection of libraries like NumPy, Pandas, and integrates well with other tools.

1. **What is the difference between descriptive and inferential statistics?**

Descriptive statistics summarize data; inferential statistics draw conclusions from sample data about a population.

1. **What is linear regression?**

A supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables.

1. **Name any two supervised machine learning algorithms.**

Linear Regression and Decision Trees.

1. **What is the role of EDA in data science?**

EDA helps understand the structure, trends, and anomalies in the data.

1. **What are outliers? How can they be handled?**

Outliers are unusual values in data. They can be handled using removal, capping, or transformation methods.

1. **Define overfitting in a machine learning model.**

Overfitting is when a model performs well on training data but poorly on new, unseen data.

1. **What is feature scaling?**

The process of standardizing or normalizing data features to bring them to a similar scale.

1. **Mention two evaluation metrics for classification problems.**

Accuracy and F1-Score.

1. **What is the function of SQL in data science?**

SQL is used for querying and managing structured data in relational databases.

1. **Explain the use of NumPy in Python.**

NumPy provides support for large, multi-dimensional arrays and matrices, along with mathematical functions.

1. **What is Pandas used for?**

Pandas is a data manipulation library used for handling structured data and performing operations like filtering and aggregation.

1. **What is data preprocessing?**

Data preprocessing is the cleaning and transformation of raw data before analysis.

1. **Name two types of data types in Python.**

Integer and String.

1. **What are missing values? How can they be treated?**

Missing values are absent data entries. They can be handled using imputation or removal.

1. **Define classification in machine learning.**

Classification is predicting the category or class of given data points.

1. **What is logistic regression used for?**

Logistic regression is used for binary and multi-class classification problems.

1. **What is a confusion matrix?**

A confusion matrix is a table used to evaluate the performance of a classification algorithm.

1. **What is a decision tree?**

A decision tree is a flowchart-like model used for classification and regression.

1. **What is entropy in decision trees?**

Entropy measures the disorder or uncertainty in a dataset.

1. **What is information gain?**

Information gain measures the reduction in entropy achieved by splitting the data.

1. **Define k-NN algorithm.**

k-NN (k-nearest neighbors) is a classification algorithm that assigns a class based on the majority class among k closest data points.

1. **What is normalization?**

Normalization is scaling data to a range of [0,1].

1. **What is standardization?**

Standardization scales data to have zero mean and unit variance.

1. **What are supervised learning algorithms?**

Algorithms that learn from labeled data.

1. **What is unsupervised learning**

Learning from unlabeled data to identify patterns or groupings.

1. **Define clustering.**

Clustering is grouping similar data points together without prior labels.

1. **What is K-means algorithm?**

K-means is a clustering algorithm that groups data into k clusters based on feature similarity.

1. **What is PCA?**

Principal Component Analysis reduces the dimensionality of data while retaining most of the variance.

1. **What is a neural network?**

A set of algorithms inspired by the human brain, used to recognize patterns.

1. **What is deep learning?**

A subset of machine learning that uses neural networks with multiple layers.

1. **What is a perceptron?**

A single-layer neural network used for binary classification.

1. **What is backpropagation?**

An algorithm for training neural networks by adjusting weights using gradient descent.

1. **What is overfitting in neural networks?**

When the model learns noise and details from the training data, reducing its generalization.

1. **What is dropout in neural networks?**

A regularization technique that randomly disables neurons during training.

1. **What is a convolutional neural network (CNN)?**

A deep learning model mainly used for image data.

1. **What is a recurrent neural network (RNN)?**

A deep learning model suitable for sequential data like text and time series.

1. **What is an activation function?**

A function that adds non-linearity to neural networks.

1. **Name any two activation functions.**

ReLU and Sigmoid.

1. **What is TensorFlow used for?**

TensorFlow is an open-source library for machine learning and deep learning.

1. **What is PyTorch?**

An open-source machine learning library used for applications such as computer vision and NLP.

1. **What is model evaluation?**

The process of assessing how well a model performs on test data.

1. **What is cross-validation?**

A technique for assessing how the results of a model will generalize to an independent dataset.

1. **What is precision in classification?**

Precision is the ratio of true positives to total predicted positives.

1. **What is recall?**

Recall is the ratio of true positives to all actual positives.

1. **What is F1-score?**

The harmonic mean of precision and recall.

1. **What is ROC curve?**

A graphical plot that shows the diagnostic ability of a classifier.

1. **What is AUC?**

AUC (Area Under Curve) measures the entire two-dimensional area underneath the ROC curve.

1. **What is data wrangling?**

The process of cleaning and unifying complex data sets.

1. **What is a data pipeline?**

A series of data processing steps.

1. **What is feature engineering?**

The process of using domain knowledge to create features that make machine learning models more effective.

1. **Define one-hot encoding.**

A method of converting categorical data into a binary matrix.

1. **What is label encoding?**

Assigning each unique category in a variable with an integer value.

1. **What is the purpose of using matplotlib in Python?**

For creating visualizations such as line plots, histograms, bar charts, etc.

1. **What is seaborn in Python?**

A Python data visualization library based on matplotlib that provides high-level interface.

1. **What is a boxplot used for?**

To display the distribution of data and identify outliers.

1. **Define correlation.**

A statistical measure that expresses the extent to which two variables are linearly related.

1. **What is multicollinearity?**

When independent variables in a regression model are highly correlated.

1. **How is multicollinearity detected?**

Using Variance Inflation Factor (VIF).

1. **What is the role of the learning rate in training?**

It determines the step size at each iteration while moving toward a minimum of a loss function.

1. **What is the epoch in neural networks?**

One complete forward and backward pass of all training examples.

1. **What is batch size?**

The number of training examples used in one iteration.

1. **What is gradient descent?**

An optimization algorithm used to minimize a loss function.

1. **What is the loss function?**

A function that maps values of one or more variables onto a real number representing the cost.

1. **What is hyperparameter tuning?**

The process of optimizing the parameters that are not directly learned from the training process.

1. **What are big data technologies?**

Technologies used for storing and processing large and complex data sets (e.g., Hadoop, Spark).

1. **What is Hadoop**

An open-source framework for distributed storage and processing of large datasets.

1. **What is MapReduce?**

A programming model for processing large datasets with a distributed algorithm.

1. **What is Apache Spark?**

A fast and general-purpose cluster computing system for big data processing.

1. **What is Hive?**

A data warehouse tool built on Hadoop that provides SQL-like querying.

1. **What is Tableau used for?**

For creating interactive and shareable dashboards and data visualizations.

1. **What is a dashboard in Tableau?**

A collection of several views and visualizations that tell a story about the data.

1. **What is a calculated field in Tableau?**

A new field created by applying a formula to existing fields.

1. **What is ML model deployment?**

Integrating a trained model into a production environment to make real-time predictions.

1. **What is Flask in Python?**

A micro web framework used for deploying machine learning models as APIs.

1. **What is Pickle in Python**

A module to serialize and deserialize Python objects.

1. **What is REST API?**

An application programming interface that follows REST architectural principles for communication.

1. **What is cloud deployment?**

Hosting machine learning models on cloud platforms like AWS, GCP, or Azure.

1. **What are Capstone Projects?**

End-to-end projects that demonstrate understanding and application of data science concepts.

1. **Why are Capstone Projects important?**

They validate your hands-on skills and problem-solving ability in real-world scenarios.

1. **What is a case study in data science?**

A detailed analysis of a real or hypothetical problem solved using data science techniques.

1. **What is version control in projects?**

A system to track changes in code over time (e.g., Git).

1. **What is GitHub?**

A platform for hosting and sharing code repositories with version control.

1. **What is continuous integration?**

The practice of merging code changes frequently to detect issues early.

1. **What is exploratory analysis?**

The process of summarizing the main characteristics of data often using visual methods.

1. **What is model drift?**

The change in model performance over time due to data changes.

1. **What is data labeling?**

The process of identifying raw data and adding labels to provide context.

1. **What is NLP?**

Natural Language Processing is a field of AI that deals with the interaction between computers and human language.

1. **What are stopwords in NLP?**

Common words like "the", "is", "in" that are often removed in text processing.

1. **What is tokenization?**

The process of breaking down text into smaller units such as words or phrases.

1. **What is stemming?**

The process of reducing words to their root form.

1. **What is lemmatization?**

Similar to stemming but results in valid words.

1. **What is TF-IDF?**

A statistical measure used to evaluate the importance of a word in a document relative to a corpus.

1. **What is a word embedding?**

A technique to represent words in dense vector form based on context.

1. **What is a bag-of-words model?**

A representation of text that describes the occurrence of words within a document.

1. **What is ensemble learning?**

Combining multiple models to improve performance.

1. **What is Random Forest?**

An ensemble method using multiple decision trees.

1. **What is boosting?**

A sequential ensemble technique that builds models iteratively to correct the previous ones.

1. **What is XGBoost?**

An efficient and scalable implementation of gradient boosting.

100**. What are the steps in a data science lifecycle?** Problem Definition, Data Collection, Data Cleaning, EDA, Modeling, Evaluation, Deployment, Monitoring.