Oral Questions and Answers for DSBDA Topics

---

1. What is Data Science? Why is it needed?

- Data Science is the study of data to extract meaningful insights for business or research purposes. It combines domain expertise, programming skills, and knowledge of mathematics and statistics.
- It is needed because of the massive amount of data generated daily that needs analysis for decision-making.

2. What is Big Data? Explain the 5 V's of Big Data.

- Big Data refers to datasets that are so large or complex that traditional data processing software can't manage them.
- 5 V's: Volume, Velocity, Variety, Veracity, and Value.

3. Applications of Data Science?

- Healthcare, Banking, E-commerce, Fraud Detection, Recommendation Systems, Image and Speech Recognition.

4. What is Data Explosion?

- Rapid growth of data due to social media, IoT devices, transactions, etc.

5. How are Data Science and Information Science related?

- Information Science focuses on organizing and accessing information; Data Science focuses on analyzing and predicting from information.

6. Business Intelligence vs Data Science?

- Business Intelligence: Deals with descriptive analytics (what happened?).
- Data Science: Deals with predictive and prescriptive analytics (what will happen and how can we make it happen?).

7. What are the phases of the Data Science Life Cycle?

- Data Collection, Data Preparation, Model Planning, Model Building, Communication of Results, Operationalization.

8. What are different Data Types?

- Structured, Semi-structured, Unstructured data.

9. What is Data Wrangling and why is it needed?

- The process of cleaning and unifying messy and complex data sets for easy access and analysis.

10. Methods of Data Wrangling?

- Data Cleaning, Integration, Reduction, Transformation, Discretization.

11. Why are statistics important in Data Science and Big Data Analytics?

- Statistics help to analyze data patterns, model relationships, and make predictions.

12. Define Measures of Central Tendency.

- Mean: Average
- Median: Middle value
- Mode: Most frequent value
- Mid-range: (Minimum + Maximum)/2

13. Define Measures of Dispersion.

- Range: Difference between maximum and minimum values.
- Variance: Average squared deviation from mean.
- Mean Deviation: Average of absolute deviations.
- Standard Deviation: Square root of variance.

14. What is Bayes Theorem?

- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

15. What is Hypothesis and Hypothesis Testing?

- Hypothesis: A statement to be tested.
- Hypothesis Testing: Process to determine if there is enough evidence to support a particular belief.

16. Explain Pearson Correlation.

- Measures the linear relationship between two variables (ranges from -1 to 1).

17. What is Sample Hypothesis Testing?

- Testing assumptions about a population parameter based on sample data.

18. What is Chi-Square Test?

- A statistical test used to determine if a significant relationship exists between categorical variables.

19. What is a t-test?

- A test used to compare the means of two groups.

20. What are sources of Big Data?

- Social Media, Sensors, Internet Transactions, Logs, Mobile Apps.

21. Phases of Data Analytic Lifecycle?

- Discovery, Data Preparation, Model Planning, Model Building, Communicate Results, Operationalize.

22. Essential Python Libraries for Data Science?

- NumPy, pandas, matplotlib, scikit-learn, seaborn.

23. What are Analytics Types?

- Predictive, Descriptive, Prescriptive Analytics.

24. What is Association Rule Mining?

- Discovering interesting relations between variables in large datasets. Algorithms: Apriori, FP-growth.

25. Explain Linear and Logistic Regression.

- Linear Regression: Predicts continuous output.
- Logistic Regression: Predicts categorical output (binary/multiclass).

26. What are Naïve Bayes and Decision Trees?

- Naïve Bayes: Classification technique based on Bayes theorem.
- Decision Trees: Tree-like model for decision making and classification.

27. What is Clustering?

- Grouping similar data points together (unsupervised learning).

- Algorithms: K-Means, Hierarchical Clustering.

28. What is Time-Series Analysis?

- Analyzing data points collected over time to forecast future trends.

29. Basics of Text Analysis?

- Preprocessing text (tokenization, removing stopwords), Bag of Words, TF-IDF, Topic Modeling.

30. Need for Social Network Analysis?

- To study relationships and interactions in social structures.

31. Metrics for Evaluating Classifier Performance?

- Accuracy, Precision, Recall, F1-Score, ROC-AUC.

32. What is Holdout Method and Random Subsampling?

- Holdout: Splitting data into training and testing.
- Random Subsampling: Repeated random splits for evaluation.

33. What is Parameter Tuning and Optimization?

- Finding the best parameters for a model to improve performance.

34. Common Model Evaluation Tools?

- Confusion Matrix, ROC Curve, AUC, Elbow Plot.

35. Challenges of Big Data Visualization?

- Scalability, Interactivity, Real-time Rendering.

36. Types and Techniques of Data Visualization?

- Line Plot, Scatter Plot, Histogram, Density Plot, Box Plot.

37. Tools for Data Visualization?

- Tableau, Power BI, matplotlib, seaborn.

38. Hadoop Ecosystem Overview?

- Components: HDFS, MapReduce, Pig, Hive, HBase, Spark.

39. What is MapReduce?

- Programming model for processing large data sets with a distributed algorithm.

40. What is Hive and Pig?

- Hive: Data warehouse software to manage large datasets.
- Pig: Platform for analyzing large datasets with a high-level scripting language.

End of Questions