Data Science and Big Data Analytics: 50 High-Probability Questions

Unit I: Introduction (CO1)

1. **What is Data Science?** Data Science is the process of extracting insights from data using statistics, machine learning, and visualization.
2. **Why is Data Science important?** It helps make data-driven decisions in business, healthcare, and other fields by analyzing large datasets.
3. **What is Big Data?** Big Data refers to large, complex datasets that traditional tools cannot process easily.
4. **What are the 5 V's of Big Data?** Volume (size), Velocity (speed), Variety (types), Veracity (accuracy), Value (usefulness).
5. **What is data explosion?** Data explosion is the rapid increase in data generation due to digital technologies like social media and IoT.
6. **Name two applications of Data Science.** Fraud detection in banking and personalized recommendations in e-commerce.
7. **What is Business Intelligence?** Business Intelligence uses data analysis to support business decisions, often with dashboards and reports.
8. **How is Data Science different from Business Intelligence?** Data Science builds predictive models, while Business Intelligence focuses on historical data analysis.
9. **What is the Data Science Life Cycle?** It includes problem definition, data collection, preparation, analysis, modeling, and deployment.
10. **What is Data Wrangling?** Data Wrangling is cleaning and preparing raw data for analysis.
11. **What is Data Cleaning?** Data Cleaning removes errors, duplicates, or missing values from a dataset.
12. **What is Data Transformation?** Data Transformation converts data into a suitable format, like scaling or encoding.
13. **What is Data Integration?** Data Integration combines data from different sources into a single dataset.
14. **What are the types of data?** Structured (tables), unstructured (text, images), and semi-structured (JSON, XML).
15. **What is Data Discretization?** Data Discretization converts continuous data into discrete categories, like age ranges.

Unit II: Statistical Inference (CO2)

1. **Why is statistics important in Data Science?** Statistics helps analyze data, find patterns, and make predictions.
2. **What is the Mean?** Mean is the average of a dataset, calculated by summing values and dividing by count.
3. **What is the Median?** Median is the middle value in a sorted dataset.
4. **What is the Mode?** Mode is the most frequent value in a dataset.
5. **What is the Mid-range?** Mid-range is the average of the maximum and minimum values in a dataset.
6. **What is the Range?** Range is the difference between the maximum and minimum values.
7. **What is Variance?** Variance measures how spread out data points are from the mean.
8. **What is Standard Deviation?** Standard Deviation is the square root of variance, showing data spread.
9. **What is Bayes Theorem?** Bayes Theorem calculates the probability of an event based on prior knowledge.
10. **What is Hypothesis Testing?** Hypothesis Testing checks if a claim about data is true using statistical methods.
11. **What is Pearson Correlation?** Pearson Correlation measures the strength of a linear relationship between two variables.
12. **What is a t-test?** A t-test compares means of two groups to check if they are significantly different.
13. **What is a Chi-Square Test?** A Chi-Square Test checks if there is a relationship between categorical variables.

Unit III: Big Data Analytics Life Cycle (CO3)

1. **What is Big Data Analytics?** Big Data Analytics processes large datasets to find patterns and insights.
2. **Name two sources of Big Data.** Social media and IoT devices.
3. **What is the Data Analytic Life Cycle?** It includes Discovery, Data Preparation, Model Planning, Model Building, Communication, and Operationalize.
4. **What happens in the Discovery phase?** Identify business problems, goals, and data sources.
5. **What is Data Preparation?** Clean and format data for analysis.
6. **What is Model Planning?** Choose algorithms and techniques for analysis.
7. **What is Model Building?** Create and test predictive or analytical models.
8. **What is the Communication phase?** Present findings to stakeholders using reports or visualizations.
9. **What does Operationalize mean?** Deploy models into production for real-world use.

Unit IV: Predictive Big Data Analytics with Python (CO4, CO2)

1. **What is Python used for in Data Science?** Python is used for data analysis, visualization, and building machine learning models.
2. **Name two Python libraries for Data Science.** Pandas and NumPy.
3. **What is Data Preprocessing?** Data Preprocessing prepares data by removing duplicates, handling missing values, and transforming data.
4. **What are the types of analytics?** Predictive (forecasting), Descriptive (summarizing), Prescriptive (recommending actions).
5. **What is the Apriori Algorithm?** Apriori finds frequent itemsets in data for association rule mining, like market basket analysis.
6. **What is Linear Regression?** Linear Regression predicts a continuous output based on input variables.
7. **What is Scikit-learn?** Scikit-learn is a Python library for machine learning tasks like regression and classification.

Unit V: Big Data Analytics and Model Evaluation (CO4, CO2)

1. **What is K-Means Clustering?** K-Means groups data into K clusters based on similarity.
2. **What is a Confusion Matrix?** A Confusion Matrix shows true vs. predicted classifications to evaluate a model.
3. **What is the AUC-ROC Curve?** AUC-ROC measures a classifier's performance by plotting true positive vs. false positive rates.
4. **What is Text Preprocessing?** Text Preprocessing cleans text data, like removing punctuation or converting to lowercase.

Unit VI: Data Visualization and Hadoop (CO4)

1. **What is Data Visualization?** Data Visualization presents data using charts, graphs, or plots to communicate insights.
2. **What is Hadoop?** Hadoop is a framework for storing and processing large datasets across distributed systems.