100 Short Viva Questions with Answers

Data Wrangling

1. **What is data wrangling?** Process of cleaning, structuring, and enriching raw data.
2. **Why is data wrangling important?** Prepares data for analysis, making it usable and accurate.
3. **What is missing data?** Data that is not recorded or unavailable.
4. **How do you detect missing values in Python?** Using `isnull()` and `sum()` functions in pandas.
5. **What is data normalization?** Scaling features to a range like [0,1].
6. **Which Python library is used for data analysis?** Pandas.
7. **How to replace missing values in pandas?** Using `fillna()` function.
8. **What is data type conversion?** Changing the data type of a variable.
9. **What is one-hot encoding?** Converting categorical variables into numeric dummy variables.
10. **Which function in pandas is used to describe statistics?** `describe()`.

Data Preprocessing

11. **What are outliers?** Extreme values differing significantly from others.
12. **How do you detect outliers?** Using box plots or Z-score.
13. **What is data transformation?** Changing the format or structure of data.
14. **Why apply log transformation?** To reduce skewness and normalize distribution.
15. **What is feature scaling?** Standardizing or normalizing data features.
16. **What is the purpose of label encoding?** Converting categorical labels into numeric form.
17. **What is standardization?** Scaling data with mean=0 and variance=1.
18. **Which library provides StandardScaler in Python?** Scikit-learn.
19. **What is skewness?** Measure of asymmetry in data distribution.
20. **Name two ways to handle missing data.** Deletion and Imputation.

Descriptive Statistics

21. **What is mean?** Average value.
22. **What is median?** Middle value of sorted data.
23. **What is mode?** Most frequently occurring value.
24. **What is variance?** Measure of dispersion.
25. **What is standard deviation?** Square root of variance.
26. **What is percentile?** Value below which a given percentage falls.
27. **What is interquartile range (IQR)?** Difference between 75th and 25th percentile.
28. **What is a boxplot?** A graphical representation of data distribution.
29. **How is a histogram different from a bar plot?** Histogram: continuous data, Bar plot: categorical data.

30. **What does a small standard deviation indicate?** Data points are close to the mean.

Data Analytics I (Linear Regression)

31. **What is linear regression?** Predicts a dependent variable using independent variables.
32. **What is the Boston Housing dataset used for?** Predicting house prices.
33. **What does R^2 (R-squared) mean?** Measure of how well data fits a regression model.
34. **Which function fits a linear regression in scikit-learn?** `LinearRegression().fit()`.
35. **What is the formula of simple linear regression?** $Y = aX + b$.
36. **What is multicollinearity?** High correlation between independent variables.
37. **How to measure model error?** Using RMSE (Root Mean Square Error).
38. **What is overfitting?** Model learns noise instead of pattern.
39. **What is underfitting?** Model is too simple to capture pattern.
40. **What is residual?** Difference between observed and predicted values.

Data Analytics II (Logistic Regression)

41. **What is logistic regression?** Classification technique based on probability.
42. **What is the output of logistic regression?** Probability values between 0 and 1.
43. **What is confusion matrix?** Matrix showing TP, FP, TN, FN.
44. **What is True Positive (TP)?** Correct positive prediction.
45. **What is True Negative (TN)?** Correct negative prediction.
46. **What is Precision?** TP / (TP + FP).
47. **What is Recall?** TP / (TP + FN).
48. **What is F1 Score?** Harmonic mean of Precision and Recall.
49. **What is ROC curve?** Graph showing true positive rate vs. false positive rate.
50. **What is AUC?** Area Under the ROC Curve.

Data Analytics III (Naïve Bayes)

51. **What is Naïve Bayes algorithm?** Classification based on Bayes' theorem.
52. **Why is it called "naïve"?** Assumes features are independent.
53. **What are common types of Naïve Bayes classifiers?** Gaussian, Multinomial, Bernoulli.
54. **What type of data does Multinomial Naïve Bayes handle?** Discrete data like text classification.
55. **What is prior probability?** Initial probability before new evidence.
56. **What is posterior probability?** Updated probability after considering evidence.
57. **What is likelihood?** Probability of evidence given hypothesis.
58. **What is conditional probability?** Probability of event A given B has occurred.
59. **Is Naïve Bayes good for small datasets?** Yes.
60. **Can Naïve Bayes be used for text classification?** Yes.

Text Analytics

61. **What is tokenization?** Splitting text into words or phrases.
62. **What is POS tagging?** Part of Speech tagging (noun, verb, etc.).
63. **What are stopwords?** Common words like "is", "the", "an" removed from text.
64. **What is stemming?** Reducing words to their root form.
65. **What is lemmatization?** Reducing words to dictionary form.
66. **What is TF-IDF?** Term Frequency - Inverse Document Frequency.
67. **What is bag-of-words?** Text representation using word counts.
68. **Which library is used for text preprocessing in Python?** NLTK.
69. **What is the difference between stemming and lemmatization?** Stemming cuts off words roughly, lemmatization uses vocabulary.
70. **What is N-gram?** Sequence of n words.

Data Visualization I & II

71. **What is Seaborn?** Python library for data visualization.
72. **How do you plot a histogram in seaborn?** `sns.histplot()`.
73. **How to plot a boxplot in seaborn?** `sns.boxplot()`.
74. **What is the Titanic dataset used for?** Survival prediction.
75. **What does a boxplot show?** Medians, quartiles, outliers.
76. **What is a pairplot?** Plots pairwise relationships between features.
77. **What is KDE plot?** Kernel Density Estimation plot.
78. **What is correlation heatmap?** Matrix showing correlations between variables.
79. **Which function is used for correlation heatmap?** `sns.heatmap()`.
80. **What is a violin plot?** Combination of boxplot and KDE plot.

Big Data Analytics (Hadoop, Spark)

81. **What is Hadoop?** Open-source framework for distributed storage and processing.
82. **What is HDFS?** Hadoop Distributed File System.
83. **What is YARN?** Yet Another Resource Negotiator.
84. **What is MapReduce?** Programming model for processing large datasets.
85. **What is Apache Spark?** Fast, in-memory big data processing engine.
86. **What is the purpose of HBase?** NoSQL database for real-time reads/writes.
87. **What is Pig in Hadoop?** High-level platform for creating MapReduce programs.
88. **What is Hive in Hadoop?** SQL-like interface for querying big data.
89. **What is Mahout?** Machine learning library for big data.
90. **What is Scala?** Programming language used with Apache Spark.

Mini Project Topics

91. **What is a recommendation system?** System suggesting products or services based on user preferences.
92. **Which algorithm can classify tweets?** Naïve Bayes or Logistic Regression.
93. **What is sentiment analysis?** Classifying text into positive, negative, or neutral sentiment.
94. **How can you classify covid vaccination data?** Using pandas groupby and visualization.
95. **Which dataset is used for covid vaccine analytics?** covid_vaccine_statewise.csv.
96. **What is business analytics?** Analyzing data to improve business decisions.
97. **What are the steps in an analytics project?** Define problem ➔ Collect data ➔ Analyze ➔ Report findings.
98. **What is supervised learning?** Training a model on labeled data.
99. **What is unsupervised learning?** Finding patterns in unlabeled data.
100. **What is clustering?** Grouping similar data points together.