Data Wrangling I

Q1: What is data wrangling?
A: Data wrangling is the process of cleaning, structuring, and enriching raw data into a desired format for better decision making.

Q2: Which Python libraries are commonly used for data wrangling?
A: Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

Q3: How do you check for missing values in a DataFrame?
A: Using `isnull()` and `sum()` functions in Pandas. Example: `df.isnull().sum()`

Q4: What is data normalization?
A: It's the process of scaling numerical data into a specific range, usually [0,1] using methods like MinMaxScaler.

Q5: How do you convert categorical variables to numeric ones in Python?
A: Using techniques like Label Encoding (`LabelEncoder`) or One Hot Encoding (`get_dummies`).

Q1: What is the first step before performing any operation in Python?
A: Import all the required Python libraries like `pandas`, `numpy`, etc.

Q2: Name any two libraries commonly used in data wrangling.
A: `pandas`, `numpy`.

Q3: How can you find open source datasets for practice?
A: Websites like Kaggle (https://www.kaggle.com), UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php).

Q4: After downloading a dataset, how do you load it in Python?
A: Using `pandas.read_csv('filename.csv')`.

Q5: What function is used to check for missing values in a dataset?
A: `isnull()` function.

Q6: What does the `describe()` function do in Pandas?
A: It provides basic statistical details like mean, median, min, max, etc., of numeric columns.

Q7: What function is used to check the size (dimensions) of a DataFrame?
A: `shape` property. (`df.shape`)

Q8: How can you see the data types of each column?
A: Using `dtypes` attribute. (`df.dtypes`)

Q9: Why is it important to check data types in a dataset?
A: To ensure the data is correctly interpreted during analysis and modeling.

Q10: How can you convert a column's data type?
A: Using `.astype()` function. (Example: `df['column'] = df['column'].astype(int)`)

Q11: What are categorical variables?
A: Variables that take on a limited number of categories or distinct groups (e.g., Gender: Male/Female).

Q12: Why do we need to convert categorical variables into numerical form?
A: Machine learning models work better with numerical values.

Q13: Which method is used to turn categorical variables into numerical values?
A: Label Encoding or One-Hot Encoding.

Q14: Name a function in Pandas used for one-hot encoding.
A: `get_dummies()`. Example: `pd.get_dummies(df['column'])`

Q15: What is data normalization?
A: Scaling numeric data into a specific range (like 0 to 1) for better model performance.

Q16: What is the purpose of checking missing values?
A: To clean the data — missing values can cause wrong results if not handled properly.

Q17: Give an example of a numeric variable and a categorical variable.
A: Numeric: Age (23, 45, 30),
Categorical: Gender (Male, Female).

Q18: Which Pandas function gives a quick summary of all columns including non-numeric?
A: `info()` function. (`df.info()`)

Q19: What can you do if there are too many missing values in a column?
A: Either drop the column or fill missing values using mean/median/mode.

Q20: What does `head()` function do in Pandas?
A: It shows the first 5 rows of the DataFrame for a quick look at the data.

Data Wrangling II (Academic Performance Dataset)

Q6: How do you handle missing values?
A: Techniques include filling with mean/median/mode, forward/backward fill, or dropping the rows/columns.

Q7: What are outliers and how can you detect them?
A: Outliers are extreme values. Detection methods include Boxplots, IQR method, and Z-score.

Q8: Why do we apply transformations on variables?
A: To normalize distributions, stabilize variance, or linearize relationships for better modelling

Question	**Answer**
Q1: What is missing data?	Missing data occurs when no data value is stored for a variable in an observation.
Q2: How can we handle missing values?	Techniques like mean/median imputation, deletion, or prediction models can be used.
Q3: What are outliers?	Outliers are extreme values that differ significantly from other observations.
Q4: Name two methods to detect outliers.	Boxplot visualization and the Interquartile Range (IQR) method.
Q5: Why do we transform variables?	To normalize data, reduce skewness, stabilize variance, or linearize relationships.
Q6: What is a log transformation?	A transformation that applies the natural logarithm to data values, often used to reduce right skewness.
Q7: What is IQR?	IQR (Interquartile Range) is the difference between the 75th percentile (Q3) and 25th percentile (Q1).

Q8: Why is normal distribution important?	Many statistical techniques assume data to be normally distributed for valid results.

Descriptive Statistics

Q9: What are measures of central tendency?
A: Mean, median, and mode.

Q10: What is standard deviation?
A: It measures the dispersion of a dataset relative to its mean.

Q11: How do you calculate group-wise statistics in Pandas?
A: Using `groupby()` function. Example: `df.groupby('age_group')['income'].mean()`

Data Analytics I (Linear Regression)

Q12: What is the objective of linear regression?
A: To model the relationship between a dependent variable and one or more independent variables.

Q13: What dataset is used in the Boston Housing problem?
A: Boston Housing Dataset from Kaggle.

Q14: What is the formula for a simple linear regression line?
A: y=mx+cy = mx + cy=mx+c where m is the slope and c is the intercept.

Data Analytics II (Logistic Regression)

Q15: When do you use logistic regression?
A: When the target variable is categorical (usually binary).

Q16: What is a confusion matrix?
A: It shows the counts of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

Q17: How is Accuracy calculated?
A: Accuracy=TP+TNTP+FP+TN+FNAccuracy = \frac{TP + TN}{TP + FP + TN + FN}Accuracy=TP+FP+TN+FNTP+TN

Data Analytics III (Naive Bayes)

Q18: What is Naïve Bayes classifier?
A: It’s a probabilistic classifier based on Bayes' Theorem with the assumption of independence among predictors.

Q19: Why is it called 'Naive'?
A: Because it assumes that the features are independent, which is rarely true in real life.

Text Analytics

Q20: What is Tokenization?
A: Breaking down text into individual words or tokens.

Q21: What is POS tagging?
A: Part-Of-Speech tagging identifies whether a word is a noun, verb, adjective, etc.

Q22: What is Stemming and Lemmatization?
A:

- Stemming cuts words to their base form.
- Lemmatization uses a dictionary to get the correct base form.

Q23: What is TF-IDF?
A: Term Frequency-Inverse Document Frequency measures how important a word is in a document relative to a corpus.

Data Visualization I and II

Q24: What is the Titanic dataset?
A: A dataset containing information about passengers aboard the Titanic, including survival status.

Q25: Which Python libraries are used for visualization?
A: Matplotlib, Seaborn, and Plotly.

Q26: How do you plot a histogram for 'fare' in Titanic dataset?
A: Using `sns.histplot(data=titanic, x='fare')`

Q27: How do you create a boxplot for 'age' with respect to 'sex' and 'survived'?
A: Using `sns.boxplot(x='sex', y='age', hue='survived', data=titanic)`

Data Visualization III (Iris Dataset)

Q28: What are the features in Iris dataset?
A: Sepal length, sepal width, petal length, and petal width (all numeric).

Q29: What plots can be used to identify outliers?
A: Boxplots and scatterplots.

Big Data Analytics – JAVA/SCALA

Q30: What is Hadoop MapReduce?
A: It's a programming model for processing large datasets with a distributed algorithm.

Q31: What is HDFS?
A: Hadoop Distributed File System - a scalable and fault-tolerant storage system.

Q32: What is Apache Spark?
A: A fast, in-memory distributed computing framework.

Mini Projects/Case Studies

Q33: What is the objective of the GINA case study?
A: To analyze and discover business problems, plan models, and find key insights.

Q34: What is a recommendation system?
A: A system that suggests items (like movies) to users based on their preferences.

Q35: How can you classify tweets into positive and negative?
A: Using Natural Language Processing (NLP) techniques and a classification algorithm like Logistic Regression or Naïve Bayes.

📚 Data Wrangling I

Q1: What is the purpose of data preprocessing?
A: To clean and prepare data for analysis and modeling.

Q2: Name any two functions to get initial statistics of a dataset.
A: `describe()`, `info()`.

Q3: What does `shape` function in Pandas return?
A: Number of rows and columns (Rows, Columns).

📚 Data Wrangling II (Academic Performance Dataset)

Q4: How do you fill missing values with the mean in Pandas?
A: `df.fillna(df.mean())`

Q5: What is the Interquartile Range (IQR)?
A: The difference between the 75th percentile and 25th percentile (Q3 - Q1).

Q6: Why is scaling important in data pre-processing?
A: To bring all features to the same scale for better model performance.

📚 Descriptive Statistics

Q7: Define mode.
A: The most frequent value in a dataset.

Q8: What does variance measure?
A: It measures the spread of data points from the mean.

Q9: Which function in Pandas gives the percentile of data?
A: `quantile()`

📚 Data Analytics I (Linear Regression)

Q10: What is the dependent variable?
A: The variable we are trying to predict.

Q11: What does a low p-value indicate in regression?
A: Strong evidence against the null hypothesis (feature is significant).

Q12: What is R-squared value?
A: It shows how well the model explains the variability of the output.

📚 Data Analytics II (Logistic Regression)

Q13: What type of output does logistic regression produce?
A: Probability values (between 0 and 1), later classified into classes.

Q14: Define Precision.
A: Precision = TP / (TP + FP)

Q15: What is Recall?
A: Recall = TP / (TP + FN)

📚 Data Analytics III (Naive Bayes)

Q16: State Bayes Theorem formula.
A:

P(A|B)=P(B|A)×P(A)P(B)P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}P(A|B)=P(B)P(B|A)×P(A)

Q17: In Naïve Bayes, why is independence assumption made?
A: To simplify calculations.

📚 Text Analytics

Q18: Name any two stop words.
A: "the", "is".

Q19: What is the main goal of Lemmatization?
A: To reduce a word to its dictionary form.

Q20: What is the role of TF in TF-IDF?
A: Measures how frequently a term occurs in a document.

📚 Data Visualization I and II

Q21: What is a histogram used for?
A: To show the distribution of a numerical variable.

Q22: What does a boxplot show?
A: Minimum, 1st quartile, median, 3rd quartile, maximum, and outliers.

Q23: Which Seaborn function is used for boxplot?
A: `sns.boxplot()`

📚 Data Visualization III (Iris Dataset)

Q24: How many classes are there in the Iris dataset?
A: Three (Setosa, Versicolor, Virginica).

Q25: What are the types of features in the Iris dataset?
A: Numeric.

📚 Big Data Analytics – JAVA/SCALA

Q26: What is MapReduce?
A: A programming model to process big data across multiple nodes.

Q27: Name two components of Hadoop.
A: HDFS and YARN.

Q28: What is Apache Pig?
A: A high-level platform for creating MapReduce programs using a scripting language.

Q29: Name a library in Spark for Machine Learning.
A: MLlib.

📚 Mini Projects/Case Studies

Q30: What is the aim of a recommendation system?
A: To suggest items to users based on preferences or behavior.

Q31: Which algorithm is often used for text classification?
A: Naïve Bayes.

Q32: What does 'label encoding' mean?
A: Converting categorical labels into numeric form.

Q33: What is the use of HBase in Hadoop ecosystem?
A: Real-time read/write access to large datasets.