A1

Question	Answer
What does `pd.read_csv()` do?	It reads a CSV file into a pandas DataFrame.
What is the purpose of `df.isnull().sum()`?	It identifies the number of missing values in each column.
What does `df.describe(include='all')` show?	It provides descriptive statistics for both numerical and categorical features.
Why use `astype('category')`?	It converts a column to categorical datatype to save memory and improve model performance.
What is One-Hot Encoding?	It converts categorical variables into a set of binary (0/1) columns.
What is `drop_first=True` in `pd.get_dummies`?	It drops the first category to avoid multicollinearity (dummy variable trap).
What does `df.dtypes` return?	It returns the data types of all columns in the DataFrame.
What is the role of Seaborn and Matplotlib imports here?	They are imported for visualization but not yet used in the given code.
What is multicollinearity?	It occurs when independent variables in a model are highly correlated. It can mislead model predictions.
If a typo like `'2urvived'` exists, what will happen?	It will raise a **KeyError** because that column does not exist unless it's intentionally named so.
What does `df.shape` tell you?	It gives the tuple (number of rows, number of columns) of the dataset.
How will you handle missing values if you find them?	Options: drop missing rows, fill with mean/median/mode, or use predictive imputation.

A2

Question	Answer
What does `fillna()` do?	It replaces missing values with specified values (here, mean of the column).
Why use `np.nan`?	It represents missing or undefined numerical data.
What is the Z-Score method for outlier detection?	A data point is considered an outlier if its Z-Score is greater than 3 or less than -3.
What does `abs(zscore()) > 3` mean?	Identifies absolute values of Z-Score greater than 3, marking potential outliers.
What is IQR?	Interquartile Range (Q3 - Q1); it captures the middle 50% of the data.
How are outliers treated here?	Using capping: outliers are replaced with nearest acceptable boundary (lower/upper bound).
Why apply a Log Transformation?	To reduce right-skewness and normalize data.
What is skewness?	It measures asymmetry of the data distribution. 0 = perfect symmetry.
What happens if data is highly skewed?	It can distort statistical analysis and affect machine learning models.
Why add 1 in `np.log(df['Math_Score']+1)`?	To avoid `log(0)` error since log(0) is undefined.
What is Winsorizing?	Replacing extreme values (outliers) with nearest acceptable values to limit influence.
Difference between Z-Score and IQR outlier detection?	Z-Score is based on standard deviation; IQR is based on percentiles.

A3

Question	Answer
What is the Iris dataset?	A famous dataset with measurements of 150 iris flowers (3 species × 50 samples each) with 4 features.
What does `load_iris(as_frame=True)` do?	Loads Iris dataset as a pandas DataFrame instead of numpy arrays.
Why create a 'species' column?	To map numeric labels (0,1,2) into readable species names ('setosa', etc.).
What is the purpose of `.groupby()`?	It groups the dataset based on a key ('species' here) to perform aggregation operations like mean, std, etc.
What does `.agg(['mean', 'median', 'min', 'max', 'std'])` do?	Calculates multiple statistical metrics at once for each group.
Why use `.apply(list)`?	To convert the grouped values into lists for more detailed inspection or plotting.
What is `df.describe()` used for?	It provides summary statistics (count, mean, std, min, 25%, 50%, 75%, max) for numerical columns.
What is `.unique()` used for here?	To list all unique species names for iteration.
Difference between `.groupby().agg()` and `.describe()`?	`.agg()` allows manual control over specific aggregations; `.describe()` gives a standard set of statistics.
Why use percentiles in `describe(percentiles=[.25, .5, .75])`?	To get statistical insight at 25th, 50th (median), and 75th percentiles for better understanding data distribution.

A4

Question	Answer
What is the Boston Housing dataset?	Dataset containing information about housing prices in Boston suburbs. Target: MEDV (median value).
Why drop 'MEDV' from x?	'MEDV' is the **output/target**; x should only have **input features**.
What is train_test_split?	It splits the data into training and testing sets to evaluate model performance properly.
What is Linear Regression?	A basic ML algorithm modeling **linear relationship** between independent variables and a dependent variable.
What is Mean Squared Error (MSE)?	Average of squared differences between predicted and actual values; lower is better.
What is R^2 Score?	It indicates proportion of variance explained by the model (1 = perfect, 0 = poor fit).
Why random_state=42?	To make results **consistent and reproducible**.
What does scatter plot show?	Visual comparison between **actual prices and predicted prices**. Ideally, points lie close to the diagonal line.
What are good MSE and R^2 values?	Lower MSE and R^2 close to 1 are considered good.
Why use Linear Regression here?	Problem is **regression (continuous target)** and initial assumption is linear relationship between features and MEDV.

A5

Question	Answer
What is Logistic Regression?	A supervised ML algorithm used for **binary classification** (predicting classes 0 or 1).
Why do we scale features before Logistic Regression?	Because Logistic Regression is sensitive to feature magnitudes; scaling improves performance.
What is the purpose of train_test_split?	To **evaluate the model** on unseen data.
What is a confusion matrix?	A table showing TP, FP, TN, FN — helps visualize classification performance.
Define True Positive (TP) and False Positive (FP).	TP: Correctly predicted positive. FP: Incorrectly predicted positive.
How is Accuracy calculated?	(TP + TN) / Total Predictions.
What is Precision?	Of all positive predictions, how many were actually positive (TP/(TP+FP)).
What is Recall?	Of all actual positives, how many were captured (TP/(TP+FN)).
Why is error rate important?	It tells us how often the model is wrong (should be minimized).
Why Logistic Regression, not Linear Regression here?	Target is **categorical**, not continuous — so **Logistic Regression** fits better.

A6

Question	Answer
What is Naive Bayes?	A **probabilistic classification algorithm** based on **Bayes' theorem** assuming **independence** between features.
What is Gaussian Naive Bayes?	A version of Naive Bayes where **features are assumed to follow a Normal (Gaussian) distribution**.
Why use Naive Bayes?	It is **simple, fast, performs well** even with small datasets, and works well when feature independence is a reasonable assumption.
Why do we split the dataset?	To **train** on one part and **test** on another — avoids overfitting and ensures generalization.
What is a confusion matrix?	A table showing TP, FP, TN, FN across all classes to evaluate prediction performance.
What is macro average precision/recall?	**Average precision/recall across all classes**, treating all classes equally (important in multi-class problems).
What does the heatmap represent?	Visual representation of Confusion Matrix, showing how well model classified each class.
What is accuracy?	**(Correct Predictions) / (Total Predictions)** — overall model performance.
What is error rate?	**1 - Accuracy** — represents misclassification rate.
Why is Gaussian Naive Bayes suitable for Iris Dataset?	Because Iris features (like petal lengths) **follow a roughly normal distribution**, and features are **mostly independent**.

A7

Question	Answer
What is tokenization?	Splitting text into small parts (tokens) like words or phrases.
What is POS tagging?	Assigning each token a grammatical category (Noun, Verb, etc.).
What are stopwords?	Very common words (like "the", "is") that usually add little meaning and are removed.
Difference between stemming and lemmatization?	Stemming cuts words crudely (may not be a real word), while lemmatization gives valid words using vocabulary and grammar.
What is TF-IDF?	A statistical measure that shows how important a word is to a document relative to the whole corpus.
Why lowercase text before processing?	To avoid treating 'Natural' and 'natural' as two different tokens.
Why remove punctuation?	Punctuation usually doesn't carry useful meaning for NLP tasks.
What is the use of the TF-IDF matrix?	It helps in feeding text data into machine learning models numerically.
What happens if you don't remove stopwords?	Stopwords may dominate and bias the model without adding meaningful information.
Why use WordNet Lemmatizer?	Because it understands grammar and gives **correct dictionary forms** of words.

A8

Question	Answer
What is a Countplot?	A bar plot that shows the counts of observations in a categorical variable.
Why use hue in Countplot or Histplot?	To add a second categorical variable (like survived/died) for better comparison.
What does KDE stand for?	Kernel Density Estimation: A smooth curve showing probability distribution.
What is a heatmap in Seaborn?	A graphical representation of data where individual values are represented as colors.
Why only select numeric data for correlation?	Correlation requires numeric values; strings/categories can't be correlated.
What is correlation?	A measure of relationship between two variables (positive, negative, or no correlation).
What is the Titanic dataset?	A dataset containing demographic and survival info about Titanic passengers.
What does dropna() do?	Removes rows with missing (NaN) values to avoid errors during analysis.
Why plot Fare distribution?	To understand ticket pricing spread and detect outliers.
What is the importance of using bins in histplot?	It controls the granularity/smoothness of the histogram (more bins = finer granularity).

A9

Question	Answer
What is a boxplot?	A graphical summary of the distribution of a dataset based on five summary statistics: minimum, first quartile, median, third quartile, and maximum.
What does hue mean in seaborn plots?	It adds a third variable by color-coding the groups (in this case, survival status).
Why use boxplot for age and sex?	To easily visualize the distribution of ages across genders and see survival patterns.
What are outliers in a boxplot?	Data points lying outside 1.5 times the IQR above Q3 or below Q1, shown as dots.
Why set figure size manually?	To make the plot readable and professional, especially if there are many categories or a large dataset.
What does the line inside the box represent?	The median (50th percentile) value of the data.
How does hue enhance boxplots?	It allows comparison between sub-categories (here, survival vs. non-survival) within each main category (male/female).
What are whiskers in a boxplot?	Lines extending from the box to the highest and lowest values within 1.5×IQR from the quartiles.
What happens if there are NaN values in 'age'?	Those entries are ignored automatically by seaborn while plotting.

A10

Question	Answer
What is a histogram?	A plot showing the **frequency distribution** of a numerical variable.
Why use histograms in EDA?	To understand data spread, skewness, and modality (one peak, two peaks, etc.).
What does a boxplot show?	Median, quartiles (Q1, Q3), spread (IQR), and outliers of a feature.
Why do we exclude 'species' while plotting histograms and boxplots?	'species' is a **categorical** column, not a numerical feature.
What does tight_layout() do?	Automatically adjusts spacing between subplots for better visualization.
What is the purpose of `edgecolor='black'` in histograms?	Makes the bars stand out clearly, improving readability.
How does seaborn's boxplot differ from matplotlib's boxplot?	Seaborn's boxplot is more **aesthetically pleasing** and easier to customize.
How do outliers appear in a boxplot?	As **individual points** beyond the whiskers.
Why use subplots in this code?	To compare **multiple features side-by-side** in the same figure.

Scala: High-level programming language combining object-oriented and functional programming.

Apache Spark: Open-source big data processing framework for fast, large-scale data analysis.

describe(): Gives summary statistics like mean, median, std for numerical columns.

astype(): Converts a column or data to a different data type.

info(): Displays basic information about dataset like column names, non-null values, and types.

TN (True Negative): Model correctly predicts negative class.

FN (False Negative): Model incorrectly predicts negative for a positive case.

TP (True Positive): Model correctly predicts positive class.

FP (False Positive): Model incorrectly predicts positive for a negative case.

TF-IDF: Term Frequency-Inverse Document Frequency; measures word importance in a document.

Tokenize: Splitting text into smaller units like words or sentences.

Test: Dataset used to evaluate model performance after training.

Train: Dataset used to build and fit the machine learning model.

POS Tagging: Assigning parts of speech (noun, verb, etc.) to each word in a text.

Lemmatization: Reduces words to their base or dictionary form.

Stemming: Cuts words down to their root form by removing suffixes.

Stopwords: Common words (like "the", "is") removed during text preprocessing.

toarray(): Converts sparse matrix (like after vectorization) into a dense NumPy array.

KDE: Kernel Density Estimate; smooths data distribution as a continuous curve.

Histplot: Plots histogram showing frequency distribution of data.

Distplot: (Deprecated) Combined histogram and KDE plot for distribution visualization.

Catplot: Categorical plot that visualizes data across categories (bar, box, strip, etc.).

Shape: Returns dimensions of dataset as (rows, columns).

Bins: Number of intervals in histogram for grouping data.

Hue: Adds color separation based on a categorical variable.

Kind: Specifies plot type in functions like catplot (e.g., 'box', 'violin').

Box Plot: Visualizes data spread using quartiles and highlights outliers.

Average = 'macro' in catplot: Computes metrics independently for each class and takes unweighted mean.

Statistics: Science of collecting, analyzing, and interpreting data.

Logistic Regression: Used for binary classification problems, outputs probability.

Linear Regression: Predicts continuous dependent variable based on independent variables.

Naive Bayes Algorithm: Classification technique based on Bayes' Theorem with independence assumption.

Confusion Matrix: Table showing true vs predicted classifications to evaluate model performance.

R2 Score: Measures goodness of fit; how well predictions approximate real data points.

Mean Squared Error: Average squared difference between predicted and actual values.

Ravel: Flattens multi-dimensional arrays into a 1D array (mainly in NumPy).

Pandas: Library for data manipulation and analysis using DataFrames.

NumPy: Library for numerical operations and array handling.

Matplotlib: Library for creating static, animated, and interactive visualizations.

Sklearn: Machine learning library offering tools for classification, regression, and clustering.

NLTK: Toolkit for working with human language data (text processing, NLP tasks).

Seaborn: Statistical data visualization library based on Matplotlib.