# **40 Python Data Science & ML Practice Questions**

**Dataset:** [AI Assistant Usage in Student Life (Synthetic)](https://www.kaggle.com/datasets/ayeshasal89/ai-assistant-usage-in-student-life-synthetic)

**Part A: Basic EDA**

1. Load the dataset and display the first 5 rows.  
    [W3Schools – Pandas Intro](https://www.w3schools.com/python/pandas/pandas_intro.asp)
2. Check the dataset shape (rows, columns).  
    [W3Schools – Pandas DataFrame](https://www.w3schools.com/python/pandas/pandas_dataframes.asp)
3. Display column names and their data types.  
    [GeeksforGeeks – Data Types in Pandas](https://www.geeksforgeeks.org/python-pandas-dataframe-dtypes/)
4. Check for missing values in each column.  
    [Programiz – Handling Missing Data](https://www.programiz.com/python-programming/pandas/missing-data)
5. Show summary statistics for SessionLengthMin and TotalPrompts.  
    [W3Schools – Pandas Describe](https://www.w3schools.com/python/pandas/ref_df_describe.asp)
6. Find the number of unique values in StudentLevel, Discipline, and TaskType.  
    [W3Schools – Pandas Unique](https://www.w3schools.com/python/pandas/ref_series_unique.asp)
7. Which TaskType is the most common?  
    [GeeksforGeeks – value\_counts](https://www.geeksforgeeks.org/python-pandas-series-value_counts-method/)
8. Calculate the average SessionLengthMin for each StudentLevel.  
    [W3Schools – Pandas GroupBy](https://www.w3schools.com/python/pandas/pandas_groupby.asp)

## **Part B: Visualization**

1. Plot a histogram of SessionLengthMin.  
    [W3Schools – Matplotlib Hist](https://www.w3schools.com/python/matplotlib_histograms.asp)
2. Create a bar chart of session counts by StudentLevel.  
    [W3Schools – Matplotlib Bar](https://www.w3schools.com/python/matplotlib_bars.asp)
3. Make a countplot of TaskType using Seaborn.  
    [GeeksforGeeks – Seaborn Countplot](https://www.geeksforgeeks.org/countplot-using-seaborn-in-python/)
4. Plot a boxplot of SessionLengthMin grouped by StudentLevel.  
    [GeeksforGeeks – Boxplot in Seaborn](https://www.geeksforgeeks.org/box-plot-seaborn-python/)
5. Create a pie chart showing proportions of FinalOutcome.  
    [W3Schools – Matplotlib Pie](https://www.w3schools.com/python/matplotlib_pie_charts.asp)
6. Draw a scatterplot of SessionLengthMin vs. TotalPrompts.  
    [W3Schools – Matplotlib Scatter](https://www.w3schools.com/python/matplotlib_scatter.asp)
7. Plot a line chart of average AI\_AssistanceLevel over time (SessionDate).  
    [W3Schools – Matplotlib Line](https://www.w3schools.com/python/matplotlib_line.asp)
8. Create a heatmap of correlations among numeric features.  
    [GeeksforGeeks – Heatmap](https://www.geeksforgeeks.org/heatmaps-in-python/)

**Part C: GroupBy & Aggregations**

1. Find the average SessionLengthMin for each TaskType.  
    [W3Schools – Pandas GroupBy](https://www.w3schools.com/python/pandas/pandas_groupby.asp)
2. Which Discipline had the most sessions?  
    [GeeksforGeeks – Pandas value\_counts](https://www.geeksforgeeks.org/python-pandas-series-value_counts-method/)
3. Compare average AI\_AssistanceLevel across StudentLevel.  
    [TutorialsPoint – GroupBy](https://www.tutorialspoint.com/python_pandas/python_pandas_groupby.htm)
4. Find the most common FinalOutcome for Graduate students.  
    [GeeksforGeeks – Pandas Filtering](https://www.geeksforgeeks.org/python-pandas-dataframe-filter/)
5. Calculate the median SessionLengthMin for each FinalOutcome.  
    [W3Schools – Pandas Median](https://www.w3schools.com/python/pandas/ref_df_median.asp)

## **Part D: Feature Engineering & Encoding**

1. Convert SessionDate into Year, Month, and Day columns.  
    [W3Schools – Pandas Dates](https://www.w3schools.com/python/pandas/pandas_datetime.asp)
2. Encode StudentLevel using Label Encoding.  
    [GeeksforGeeks – Label Encoding](https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/)
3. Apply One-Hot Encoding to TaskType.  
    [GeeksforGeeks – One Hot Encoding](https://www.geeksforgeeks.org/ml-one-hot-encoding/)
4. Create a new feature: PromptsPerMinute = TotalPrompts / SessionLengthMin.  
    [W3Schools – Pandas New Column](https://www.w3schools.com/python/pandas/pandas_dataframes_newcolumn.asp)
5. Bin SessionLengthMin into categories: Short, Medium, Long.  
    [GeeksforGeeks – Pandas Cut](https://www.geeksforgeeks.org/pandas-cut-method-in-python/)

## **Part E: Machine Learning (Classification Models)**

1. Predict FinalOutcome using a **Decision Tree Classifier**.  
    [W3Schools – Decision Tree](https://www.w3schools.com/python/python_ml_decision_tree.asp)
2. Predict UsedAgain using **Logistic Regression**.  
    [W3Schools – Logistic Regression](https://www.w3schools.com/python/python_ml_logistic_regression.asp)
3. Train a **Random Forest Classifier** to predict FinalOutcome.  
     [GeeksforGeeks – Random Forest](https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/)
4. Use **KNN (K-Nearest Neighbors)** to classify UsedAgain.  
    [W3Schools – KNN](https://www.w3schools.com/python/python_ml_knn.asp)
5. Train a **Naive Bayes Classifier** to predict FinalOutcome.  
    [GeeksforGeeks – Naive Bayes](https://www.geeksforgeeks.org/naive-bayes-classifiers/)
6. Build a **Gradient Boosting Classifier** for UsedAgain.  
    [GeeksforGeeks – Gradient Boosting](https://www.geeksforgeeks.org/ml-gradient-boosting/)
7. Apply an **XGBoost Classifier** to predict FinalOutcome.  
    [GeeksforGeeks – XGBoost](https://www.geeksforgeeks.org/xgboost-for-regression/)
8. Split the dataset into 80% training and 20% testing sets.  
    [W3Schools – Train/Test Split](https://www.w3schools.com/python/python_ml_train_test.asp)
9. Evaluate one model using accuracy and a confusion matrix.  
    [W3Schools – Confusion Matrix](https://www.w3schools.com/python/python_ml_confusion_matrix.asp)
10. Generate a **classification report** (precision, recall, f1-score).  
     [GeeksforGeeks – Classification Report](https://www.geeksforgeeks.org/classification-report-in-machine-learning/)

## **Part F: Model Evaluation & Hyperparameter Tuning**

1. Perform cross-validation for Logistic Regression.  
    [GeeksforGeeks – Cross Validation](https://www.geeksforgeeks.org/cross-validation-machine-learning/)
2. Use GridSearchCV to tune hyperparameters of a Decision Tree.  
    [GeeksforGeeks – GridSearchCV](https://www.geeksforgeeks.org/ml-hyperparameter-tuning-using-gridsearchcv/)
3. Tune a Random Forest Classifier (n\_estimators, max\_depth).  
    [GeeksforGeeks – Random Forest Hyperparameters](https://www.geeksforgeeks.org/random-forest-hyperparameters/)
4. Compare Logistic Regression, Decision Tree, Random Forest, Naive Bayes, KNN, Gradient Boosting, and XGBoost for predicting UsedAgain.  
     [W3Schools – Compare Models](https://www.w3schools.com/python/python_ml_scatter.asp) (basic), + [GeeksforGeeks – Compare ML Models](https://www.geeksforgeeks.org/how-to-compare-machine-learning-models/)