Q1: What are Python's key data structures, and when would you use a list vs. a dictionary?

A1: List: Ordered collection; use when order matters.

Dict: Key-value pairs; use for fast lookups by keys.

Q2: How would you handle missing data in a Pandas DataFrame?

A2: Use df.dropna() to drop rows, df.fillna(value) to impute.

Detect with df.isnull().sum().

Q3: Write a query to find the second-highest salary from an Employee table.

A3: SELECT MAX(salary) FROM Employee WHERE salary < (SELECT MAX(salary) FROM Employee);

Q4: What's the difference between ROW_NUMBER() and RANK() in SQL?

A4: ROW_NUMBER: Unique sequence.

RANK: Same rank for ties, skips ranks.

Q5: What's the difference between VLOOKUP, INDEX-MATCH, and XLOOKUP?

A5: VLOOKUP: Search right only.

INDEX-MATCH: Flexible, search left/right.

XLOOKUP: Replaces both.

Q6: How would you use a PivotTable to analyze sales by region and product category?

A6: Insert PivotTable, drag Region to rows, Product Category to columns, Sales to values.

Q7: How would you create a dynamic dashboard with filters in Tableau?

A7: Drag filter field to Filters shelf, select 'Show Filter', apply to all relevant sheets.

Q8: When would you use a box plot vs. a violin plot in Seaborn?

A8: Box: Summary stats.

Violin: Shows KDE, better for distribution shape.

Q9: How do you decide whether to use Linear Regression or Decision Trees?

A9: Linear Regression: Linear data.

Decision Trees: Nonlinear, interpretable splits.

Q10: What are precision and recall? When is recall more important?

A10: Precision: TP / (TP+FP).

Recall: TP / (TP+FN).

Recall is key in medical/fraud cases.

Q11: How does K-Means clustering work?

A11: Random centroids -> assign points -> recalculate -> repeat until convergence.

Q12: How do you choose the number of clusters in K-Means?

A12: Elbow method, Silhouette Score, Gap Statistic.

Q13: What is stationarity in time series? Why is it important?

A13: Stationary = constant mean/variance.

Required for ARIMA. Check with ADF test.

Q14: When would you use Prophet over ARIMA?

A14: Prophet: Seasonal, missing data, holidays.

ARIMA: Fine-tuned control.

Q15: What are Type I and Type II errors? Which is worse?

A15: Type I: False positive.

Type II: False negative.

Depends on context (e.g., court vs. health).

Q16: How would you design an A/B test?

A16: Define hypothesis -> Split groups -> Run test -> Analyze with t-test/z-test.

Q17: What are the first steps you take when analyzing a new dataset?

A17: Check shape, nulls, data types -> use describe/info -> plot distributions.

Q18: How do you detect outliers?

A18: IQR method, Z-score, boxplots.

Q19: What is multicollinearity, and how do you detect it?

A19: High feature correlation.

Detect with VIF, correlation matrix.

Q20: Give examples of feature transformation techniques.

A20: Binning, Scaling, Encoding, Date parts extraction.

Q21: What's the difference between batch size, epochs, and iterations?

A21: Batch size: Samples per update.

Epoch: Full pass through data.

Iterations: Batches per epoch.

Q22: When would you use an RNN instead of a CNN?

A22: RNN: Sequential data (text, time series).

CNN: Spatial data (images).

Q23: What's the difference between git pull, git fetch, and git clone?

A23: pull = fetch + merge.

fetch = just download.

clone = copy repo to local.

Q24: How would you containerize a Streamlit app using Docker?

A24: Create Dockerfile -> install deps -> copy files -> CMD streamlit run app.py -> build and run.

Q25: What's the role of S3, EC2, and Lambda in AWS?

A25: S3: Storage.

EC2: Compute.

Lambda: Serverless functions.

Q26: How would you deploy a model on cloud for real-time inference?

A26: Save model -> create API (Flask/FastAPI) -> deploy via EC2, Lambda, or Cloud Run.

Q27: How do you interpret a ROC curve?

A27: TPR vs. FPR plot.

Closer to top-left = better.

AUC closer to 1 = good.

Q28: What's the difference between cross-validation and train/test split?

A28: Split: One-time.

CV: Repeated splits, better generalization.

Q29: How does GridSearchCV work?

A29: Tests all hyperparameter combos with CV.

Returns best based on scoring.

Q30: What is overfitting and how do you prevent it?

A30: High train but low test accuracy.

Fix with CV, regularization, pruning, simpler models.