**✅ 1. What are the key differences between INNER JOIN and OUTER JOIN in SQL?**

* **INNER JOIN** returns only the matching rows from both tables.
* **OUTER JOIN** (LEFT, RIGHT, or FULL) returns all rows from one or both tables, filling in NULLs where there's no match.

**✅ 2. How do you handle missing data in a dataset?**

* **Identify missing values** using .isnull() or .info()
* **Techniques**:
  + Imputation (mean/median/mode)
  + Drop rows/columns (if too many missing)
  + Use ML models for advanced imputation
  + Add indicators (e.g., is\_missing flag)

**✅ 3. What is the difference between variance and standard deviation?**

* **Variance**: Measures the average squared deviation from the mean.
* **Standard Deviation**: Square root of variance; gives dispersion in original units.

**✅ 4. Explain the concept of normalization in databases.**

Normalization is the process of organizing data to reduce redundancy and improve integrity. It involves breaking down tables into smaller ones and linking them using keys (1NF, 2NF, 3NF...).

**✅ 5. What is the role of a primary key in a relational database?**

A **primary key** uniquely identifies each record in a table. It enforces **uniqueness** and **non-nullability**, ensuring data integrity.

**✅ 6. How would you detect outliers in a dataset?**

* **Visual Methods**: Box plots, scatter plots
* **Statistical**:
  + Z-score (>3 or <-3)
  + IQR method (outside 1.5×IQR)
  + Isolation Forest or DBSCAN (advanced)

**✅ 7. What is data wrangling and why is it important?**

Data wrangling is the process of cleaning and transforming raw data into a usable format. It's essential because real-world data is messy and needs preparation before analysis.

**✅ 8. Describe a situation where you used data to solve a business problem.**

In my YouTube Trending Video Analytics project, I analyzed trending videos across countries. Using SQL and Python, I uncovered which regions prefer which content types — helping identify target content for different markets.

**✅ 9. What is the difference between a clustered and non-clustered index?**

* **Clustered Index**: Sorts and stores the data rows in the table based on the index.
* **Non-Clustered Index**: Maintains a separate structure from the table and points to the actual data rows.

**✅ 10. Explain the difference between supervised and unsupervised learning.**

* **Supervised Learning**: Uses labeled data (e.g., regression, classification).
* **Unsupervised Learning**: Finds patterns in unlabeled data (e.g., clustering, dimensionality reduction).

**✅ 11. What is the purpose of the GROUP BY clause in SQL?**

GROUP BY is used to **aggregate data** based on one or more columns. It groups rows that share a value, allowing you to apply functions like COUNT(), SUM(), or AVG() on each group.

**✅ 12. How do you handle duplicate data entries in a dataset?**

* Identify duplicates using df.duplicated()
* Remove them using df.drop\_duplicates()
* You can keep the **first** or **last** occurrence or drop all duplicates depending on business logic.

**✅ 13. What is a pivot table and how have you used it?**

A pivot table summarizes data by grouping and aggregating it. In Excel/Power BI, I used pivot tables to quickly analyze sales by **region, category, or month**, helping management spot trends.

**✅ 14. Explain the differences between a bar chart and a histogram.**

* **Bar Chart**: Displays categorical data using bars.
* **Histogram**: Displays frequency distribution of numerical (continuous) data.

**✅ 15. How do you optimize a slow SQL query?**

* Use **indexes** on filtered or joined columns.
* Avoid SELECT \* — only fetch needed columns.
* Use **LIMIT**, proper **JOINs**, and **WHERE** filters.
* Check for unnecessary subqueries or large intermediate datasets.

**✅ 16. What are common KPIs used in business analysis?**

* Sales Revenue
* Conversion Rate
* Customer Retention Rate
* Average Order Value
* Net Promoter Score (NPS)
* Return on Investment (ROI)

**✅ 17. What is A/B testing and how is it used in data analysis?**

A/B testing compares two versions (A and B) of a variable (e.g., a web page) to see which performs better based on a **key metric** like click-through rate. It uses hypothesis testing to make data-driven decisions.

**✅ 18. How do you ensure data accuracy and integrity in a project?**

* Validate sources and schema
* Use constraints like **primary keys**, **foreign keys**, **NOT NULL**
* Perform regular audits and sanity checks
* Maintain version control and clear data pipelines

**✅ 19. What is a correlation matrix and how do you interpret it?**

A correlation matrix shows **pairwise correlation coefficients** (between -1 and 1) for variables. Values close to ±1 indicate strong relationships; values near 0 suggest weak or no correlation.

**✅ 20. What is the difference between correlation and causation?**

* **Correlation**: A relationship exists between variables (e.g., X ↑ as Y ↑).
* **Causation**: One variable **causes** a change in another. Correlation does **not** imply causation without controlled testing.