# Software Houses in Pakistan - SQL Analysis Questions

## Data Cleaning & Preprocessing

1. Total rows in dataset (audit).

2. Identify and replace fake nulls like 'N/A', 'Not Available', '-' with NULL.

3. Replace errors in contact_no (e.g., #ERROR!) with NULL.

4. Standardize column names and ensure proper formatting.

5. Handle actual NULL values:

   - Update missing services with 'Unknown'.

   - Remove rows with NULL city (no Pakistani office).

6. Check and count actual duplicates.

7. Decide on approach to soft duplicates (company name with slight variations).

8. Standardize city names where applicable (e.g., khi -> Karachi).

## Basic Data Analysis

1. How many companies are there per city?

2. How many companies have complete contact info?

3. What are the top 5 services offered?

4. Which cities are underserved?

   (Cities with the least number of software houses)

5. How many companies have missing or unknown company_name or services?

## Advanced Data Analysis

1. Which companies are present in multiple cities?

   - Companies listed with offices in more than one city.

2. Which companies offer both software development & digital marketing?

   - Using keyword-based LIKE conditions.

3. Compare % of total companies per city.

   - Using CTEs to calculate relative city-wise distribution.

4. Create a "completeness score" per company:

   - 1 point each for:

# Software Houses in Pakistan - SQL Analysis Questions

- Non-null contact

- Valid city

- At least 1 service

- Unique name.