**1. Introduction to Data Cleaning**

**Definition:**  
Data cleaning is the process of correcting or removing incorrect, incomplete, or irrelevant data.

**Common data issues:**

* Null values
* Duplicates
* Inconsistent formats (e.g., dates, casing)
* Invalid entries (e.g., negative age)
* Truncated or misspelled strings

**2. Sample Tables Setup**

CREATE TABLE Customers (

CustID INT,

FullName VARCHAR(100),

Email VARCHAR(100),

DOB DATE,

Country VARCHAR(50),

Phone VARCHAR(20)

);

INSERT INTO Customers VALUES

(1, 'John Doe', 'john@example.com', '1980-05-20', 'USA', '1234567890'),

(2, 'sara K', 'SARA@example.COM', NULL, 'usa', '9876543210'),

(3, 'M!ke 77', NULL, '1992-11-30', 'India', '99999'),

(4, 'Linda', 'linda@example.com', '1975-07-25', NULL, NULL),

(5, 'John Doe', 'john@example.com', '1980-05-20', 'USA', '1234567890'); -- duplicate

**3. Identify Dirty Data**

-- Find NULLs

SELECT \* FROM Customers WHERE Email IS NULL OR Country IS NULL;

-- Detect duplicate rows

SELECT FullName, Email, COUNT(\*)

FROM Customers

GROUP BY FullName, Email

HAVING COUNT(\*) > 1;

-- Find invalid characters in names

SELECT \* FROM Customers WHERE FullName LIKE '%[^a-zA-Z ]%';

-- Find inconsistent casing

SELECT DISTINCT Country FROM Customers;

**4. Fixing Nulls, Duplicates, Casing**

-- Update missing countries to 'Unknown'

UPDATE Customers SET Country = 'Unknown' WHERE Country IS NULL;

-- Standardize country names (USA → UPPER)

UPDATE Customers SET Country = UPPER(Country);

-- Remove duplicates

DELETE FROM Customers

WHERE CustID NOT IN (

SELECT MIN(CustID)

FROM Customers

GROUP BY FullName, Email

);

**5. Handling Data Types**

-- Trim and format names

SELECT TRIM(FullName) AS CleanName,

INITCAP(FullName) AS ProperCase

FROM Customers;

-- Replace invalid characters

SELECT REGEXP\_REPLACE(FullName, '[^a-zA-Z ]', '', 'g') AS CleanName

FROM Customers;

**Working with Dates:**

-- Calculate age

SELECT FullName, DOB,

EXTRACT(YEAR FROM CURRENT\_DATE) - EXTRACT(YEAR FROM DOB) AS Age

FROM Customers

WHERE DOB IS NOT NULL;

**Working with NULLs and Defaults:**

-- Replace nulls in phone with 'NA'

SELECT COALESCE(Phone, 'NA') AS PhoneNumber FROM Customers;