SQL Data Engineering Project: Restaurant Data Cleaning

Project Goal

Practice cleaning messy data and checking data quality using SQL.

Setup

Create a database and two raw tables:

```
CREATE DATABASE restaurant_data;

CREATE TABLE raw_orders (
    order_id VARCHAR(50),
    customer_email VARCHAR(100),
    order_date VARCHAR(50),
    item_name VARCHAR(100),
    quantity VARCHAR(10),
    price VARCHAR(10)
);

CREATE TABLE raw_customers (
    email VARCHAR(100),
    name VARCHAR(100),
    phone VARCHAR(50),
    signup_date VARCHAR(50)
);
```

Import the provided CSV files into these tables.

Task 1: Clean Customer Data

Create a clean table:

```
CREATE TABLE clean_customers (
    customer_id INT PRIMARY KEY AUTO_INCREMENT,
    email VARCHAR(100),
    first_name VARCHAR(50),
    last_name VARCHAR(50),
    phone VARCHAR(20),
    signup_date DATE
);
```

Write an INSERT query that: - Removes duplicate emails (keep the first one) - Splits "name" into first_name and last_name - Converts signup_date from text to DATE - Removes spaces and dashes from phone numbers

---

Task 2: Clean Order Data

Create clean tables:

```
CREATE TABLE clean_orders (
    order_id INT PRIMARY KEY,
    customer_email VARCHAR(100),
    order_date DATE
);

CREATE TABLE clean_order_items (
    item_id INT PRIMARY KEY AUTO_INCREMENT,
    order_id INT,
    item_name VARCHAR(100),
    quantity INT,
    price DECIMAL(6,2)
);
```

Write INSERT queries that: - Convert order_id from text to INT - Convert order_date from text to DATE - Convert quantity from text to INT - Convert price from text to DECIMAL - Handle any NULL or empty values

---

Task 3: Find Data Quality Issues

Write queries to find these problems in the raw tables:

1. Find duplicate emails in raw_customers
2. Find orders where customer_email doesn't exist in raw_customers
3. Find orders with NULL, empty, or zero prices
4. Find orders with negative or zero quantities
5. Find dates that can't be converted (like "20/31/2024" or "invalid")
6. Find phone numbers that don't have 10 digits

Save each query with a comment explaining what issue it finds.

Task 4: Simple Analysis

Using your clean tables, write queries to answer:

1. How many customers are in the clean table?
2. How many total orders were placed?
3. What is the total revenue (sum of price × quantity)?
4. Which item was ordered most frequently?

What You'll Learn

✅ Converting data types (VARCHAR → INT, DATE, DECIMAL)
✅ String manipulation (splitting names, cleaning phone numbers)
✅ Handling duplicates
✅ Data validation queries
✅ Working with messy real-world data