# Coffee Dataset – Data Cleaning Summary (SQL-Ready)

## 🗂 Dataset Overview

- Goal: Prepare coffee quality dataset for structured SQL analysis  
- Initial Data Format: CSV with mixed types (text, ranges, inconsistent date formats)  
- Final Output: cleaned\_coffee\_sql\_ready.csv (ready for import into a relational database)

## 🔧 Data Cleaning Steps

### 1. 🧱 Handling Missing Values

- Fields like Farm Name, Mill, Region, and Variety had sparse missing values → filled with 'Unknown'.  
- Rows with missing ICO Number (critical identifier) → dropped entirely to preserve referential integrity.  
  
SQL Example:  
UPDATE coffee\_table SET region = 'Unknown' WHERE region IS NULL;  
DELETE FROM coffee\_table WHERE ico\_number IS NULL;

### 2. 🏔 Splitting Altitude Range

- Altitude values like '1700-1930' split into:  
 - Min\_Altitude: 1700  
 - Max\_Altitude: 1930  
- If only a single value (e.g., '1300'), both Min\_Altitude and Max\_Altitude were set to 1300.  
  
SQL Pseudocode: Use SPLIT or SUBSTRING functions to separate values.

### 3. ⚖️ Converting Bag Weight

- Removed 'kg' suffix from Bag Weight values like '69 kg'.  
- Stored the numeric result as float in new column: Bag\_Weight\_KG.  
  
SQL Example:  
UPDATE coffee\_table SET bag\_weight\_kg = CAST(REPLACE(bag\_weight, 'kg', '') AS FLOAT);

### 4. 🗓 Date Conversion

- Harvest Year: converted to INTEGER.  
- Grading Date: Parsed from formats like 'September 21st, 2023' to proper DATE.  
 - Removed ordinal suffixes (st, nd, etc.) before parsing.  
  
SQL Example: Use STR\_TO\_DATE or equivalent functions.

### 5. 📎 Duplicate Check

- Checked duplicates based on Lot Number and Company.  
- Duplicates were flagged and exported for manual review (in duplicate\_lots.csv), not deleted automatically.  
  
SQL Example:  
SELECT lot\_number, company, COUNT(\*)  
FROM coffee\_table  
GROUP BY lot\_number, company  
HAVING COUNT(\*) > 1;

## 📊 Summary After Cleaning

|  |  |
| --- | --- |
| Metric | Description |
| Records Removed | Rows with missing ICO number |
| Text Replacements | Replaced nulls in 4 text fields with 'Unknown' |
| New Columns Added | Min\_Altitude, Max\_Altitude, Bag\_Weight\_KG |
| Converted Columns | Harvest Year, Grading Date |
| Duplicate Check | Flagged for review, not auto-deleted |