**Big Data Frameworks CSE3120**

Lab – 2 Data Cleaning Experiment

**Name:** Naveen Nidadavolu

**Roll No:** 22MIA1049

**Aim:** To clean two CSV files by performing the following tasks:

1. Remove any unnecessary whitespace around the data.
2. Replace missing, null, or n/a values with a default value (N/A).
3. Remove duplicate rows (if a row is repeated in the file, only the first occurrence will be kept).
4. Save the cleaned data into new output files.

**Algorithm**

1. **Input**: Accept two CSV files as input.
2. **Open Files**: Read each file line by line.
3. **Process Header**: Retain the first line (header) as is and write it to the output file.
4. **Process Data Rows**:
5. For each row (after the header), clean it by:
   1. Removing leading and trailing whitespace from each column.
   2. Replacing any "null", "n/a", or empty values with "N/A".
6. If the cleaned row has not been seen before (using a HashSet to track unique rows), write it to the output file.
7. **Output**: Write the cleaned data to new CSV files.
8. **End**: Save and close the files after processing.

**Program:**

A screen shot of a computer program

Description automatically generated A screen shot of a computer code

Description automatically generated

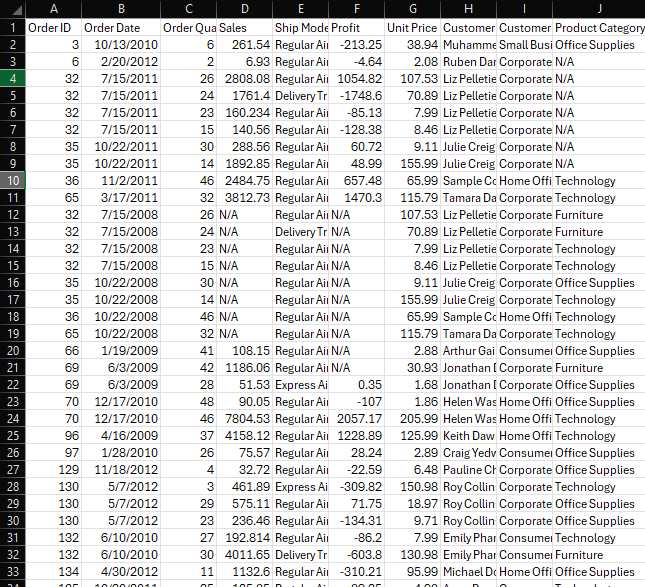
**Output:**

A black screen with white text

Description automatically generated

**Cleaned Table1:**

A screen shot of a computer screen

Description automatically generated

**Cleaned Table2:**

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

**Result:**

Given datasets successfully cleaned using java.