



Supclair
Clarity in Every Chain



Capstone Project

Data Preprocessing Notebook

Date of Submission: 1/3/2025

Submitted by: Supclair

Contents

Objectives of Data Cleaning.....	3
Data Sources	3
Cleaning steps.....	4
Product Table	5
Transportation Table	6
Customer Table	6
Supplier Table	7
Report Table	7

Objectives of Data Cleaning

The primary objective of this data cleaning process is to ensure the accuracy, consistency, and reliability of the **Supplier Quality Dataset**. By addressing common data quality issues, this process aims to enhance the validity of subsequent analysis and reporting. The specific goals of data cleaning include:

- **Handling Missing Data:** Identifying and addressing any missing values in key columns to prevent gaps in analysis.
- **Correcting Inconsistencies:** Standardizing naming conventions, data formats, and measurement units to ensure uniformity across records.
- **Removing Duplicates:** Eliminating redundant records to avoid double counting and biased results.
- **Resolving Errors:** Identifying and correcting incorrect values, such as negative defect counts or unrealistic downtime durations.
- **Ensuring Data Integrity:** Verifying relationships between different columns, such as supplier IDs matching the correct supplier names.

By conducting a thorough data cleaning process, the dataset will be optimized for accurate supplier quality analysis, ensuring meaningful insights for decision-making.

Data Sources

The dataset used in this analysis originates from supply chain management systems and consists of records related to lead times, stock levels, shipping details, and supplier performance.

The dataset is provided in CSV format (Supply chain.csv) for seamless import into data analysis tools, structured into multiple fields capturing essential supply chain metrics needed for evaluation and optimization.

Cleaning steps

The steps will be in this template for each table:

Table name:

Notes:

Changes made:

Data type changes:

The flat file “supply_chain_dataset” was transformed to the Power Query.

Our approach was to normalize the given dataset for better overall performance and data integrity. This was done by dividing the flat files into 1-fact table and 4-dimension tables. The fact table is the **Report Table**, and the dimension tables are **Products Table**, **Supplier Table**, **Customer Table**, and **Transportation table**.

To normalize the dataset, we followed these steps:

1. Make a copy of the **Report Table** as **Duplicate**.
2. Choose the **columns needed** for the table and **remove the others**.
3. Remove the **duplicates** as needed to make a **unique dimension table**.
4. Add an **indexed column** if needed.
5. Change the **Data types** if needed.
6. Go back to the **Report Table** and merge the new table with the **intended column** if needed. Choose the **ID column**.
7. Remove the **normalized column** from the **Report Table** leaving just the **ID column**

Example:

Product Table

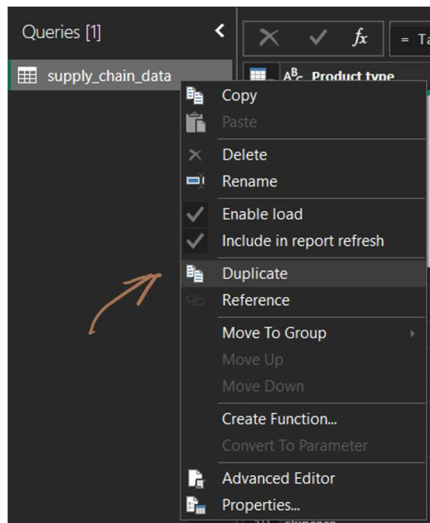


Figure 1: Duplicate Report Table

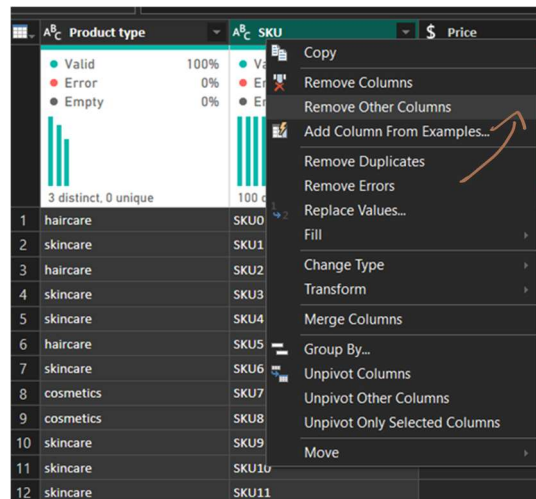


Figure 2: Removing other columns

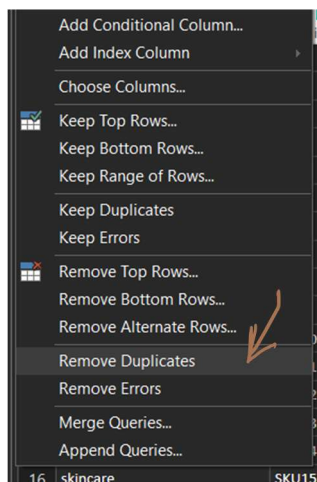


Figure 3: Removing Duplicates

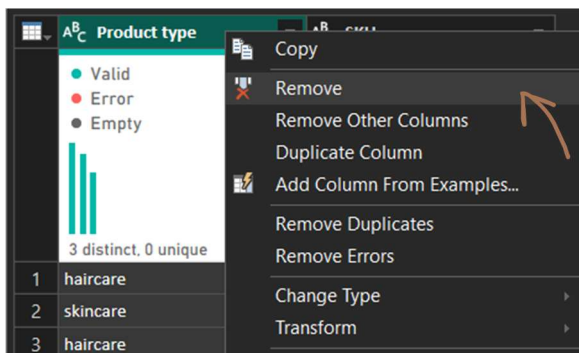


Figure 4: Remove Product Type from Report Table

A screenshot of the final Power BI report. It shows two tables side-by-side. The 'SKU' table has 100 distinct, 100 unique values. The 'Product type' table has 3 distinct, 0 unique values. The data is as follows:

SKU	Product type
1	haircare
2	skincare
3	haircare
4	skincare
5	skincare
6	haircare
7	skincare
8	cosmetics
9	cosmetics
10	skincare
11	skincare
12	skincare
13	skincare
14	skincare
15	skincare
16	skincare
17	skincare
18	skincare
19	skincare
20	skincare
21	skincare
22	skincare
23	skincare
24	skincare
25	skincare
26	skincare
27	skincare
28	skincare
29	skincare
30	skincare
31	skincare
32	skincare
33	skincare
34	skincare
35	skincare
36	skincare
37	skincare
38	skincare
39	skincare
40	skincare
41	skincare
42	skincare
43	skincare
44	skincare
45	skincare
46	skincare
47	skincare
48	skincare
49	skincare
50	skincare
51	skincare
52	skincare
53	skincare
54	skincare
55	skincare
56	skincare
57	skincare
58	skincare
59	skincare
60	skincare
61	skincare
62	skincare
63	skincare
64	skincare
65	skincare
66	skincare
67	skincare
68	skincare
69	skincare
70	skincare
71	skincare
72	skincare
73	skincare
74	skincare
75	skincare
76	skincare
77	skincare
78	skincare
79	skincare
80	skincare
81	skincare
82	skincare
83	skincare
84	skincare
85	skincare
86	skincare
87	skincare
88	skincare
89	skincare
90	skincare
91	skincare
92	skincare
93	skincare
94	skincare
95	skincare
96	skincare
97	skincare
98	skincare
99	skincare
100	skincare

Figure 5: Final Product Table

Instead of creating an indexed column, SKU is used.

There is not any missing data or inconsistency

Transportation Table

	A ^B _C TransportationID	A ^B _C Shipping carriers	A ^B _C Routes	A ^B _C Transportation modes
	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>33 distinct, 33 unique</div>	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>3 distinct, 0 unique</div>	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>3 distinct, 0 unique</div>	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>4 distinct, 0 unique</div>
1	1	Carrier A	Route B	Air
2	2	Carrier B	Route C	Air
3	3	Carrier C	Route B	Air

There is no missing data or inconsistency.

Data Type Changes:

Transportation ID column is changed from Whole number to Text.

Customer Table

	A ^B _C GenderID	A ^B _C Customer demographics
	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>4 distinct, 4 unique</div>	<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>4 distinct, 4 unique</div>
1	1	Non-binary
2	2	Female
3	3	Unknown
4	4	Male

There is no missing data or inconsistency.

Data Type Changes:

Gender ID column is changed from Whole number to Text.

Supplier Table

	Supplier ID	Supplier name	Location
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	25 distinct, 25 unique	5 distinct, 0 unique	5 distinct, 0 unique
1	1	Supplier 1	Bangalore
2	2	Supplier 1	Mumbai

There is no missing data or inconsistency.

Data Type Changes:

Supplier ID column is changed from Whole number to Text.

Report Table

	GenderID	SKU	SupplierID	TransportationID	Price	Availability	NumberofProductsSold	RevenueGenerated	StockLevels
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	4 distinct, 0 unique	100 distinct, 100 unique	25 distinct, 1 unique	33 distinct, 8 unique	100 distinct, 100 unique	63 distinct, 37 unique	96 distinct, 92 unique	100 distinct, 100 unique	65 distinct, 39 unique
1	1	SKU0	13	23	69.81	55	802	8,662.00	
2	2	SKU29	5	23	63.45	3	253	8,318.90	
3	3	SKU20	4	1	96.34	22	320	8,128.03	

The ID columns were added to the report and the normalized columns were removed.

There is no missing data, duplicates or inconsistency.

Data type Changes:

Price column is changed from Decimal number to fixed decimal number (currency).

RevenueGenerated column is changed from Decimal number to fixed decimal number (currency).

ShippingCost column is changed from Decimal number to fixed decimal number (currency).

ManufacturingCost column is changed from Decimal number to fixed decimal number (currency).

Costs column is changed from **Decimal number** to **fixed decimal number** (currency).