EYouth X DEPI Tech Challenge

Data Cleaning and Modeling Report

Presented to: EYouth

Date of Submission: 28/2/2025

Submitted by: Mohammed Yasir Abduljawad

Contents

Objectives of Data Cleaning......3

Data Sources......3

Cleaning steps......4

- Template......4
- Category table......4
- Defect table......5
- Defect type table......6
- Material table......7
- Plant Table......8
- Vendor Table......9
- Defect Report Table:......10

Data Model......11

- Objective of Data Model......11
- Model......11

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 supplier quality monitoring systems and consists of records related to **defect rates, vendor, plant, and material details.**

The dataset is provided in **Google sheets** and exported in **Excel format (Supplier Quality Analysis Dataset.xlsx)** to import it to the Power Query easily, structured into multiple fields capturing essential supplier quality details needed for analysis.

Cleaning steps

Template

The steps will be in this template for each table:

Table name:

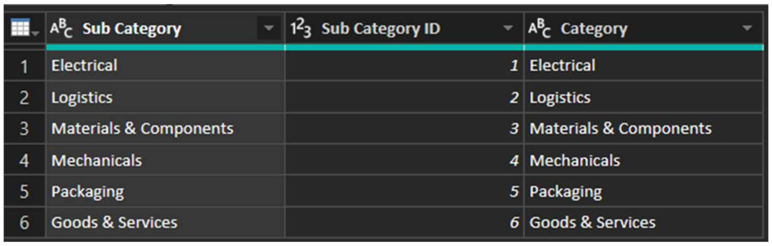
Notes:

Changes made:

Data type changes:

Transformed all sheets except description sheet.

Category table



	Sub Category	Sub Category ID	Category
1	Electrical	1	Electrical
2	Logistics	2	Logistics
3	Materials & Components	3	Materials & Components
4	Mechanicals	4	Mechanicals
5	Packaging	5	Packaging
6	Goods & Services	6	Goods & Services

Figure 1: Category Table Before

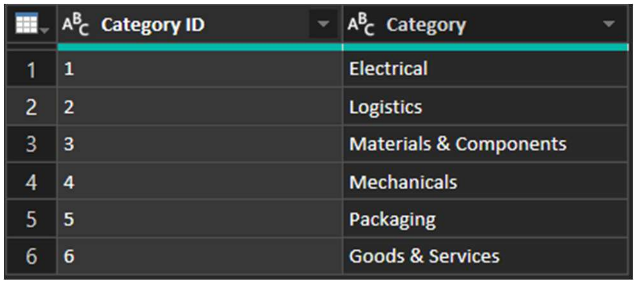
There are two columns **Category, Sub-Category, and Sub-Category ID.** The category columns and sub-category columns are the same. There are no missing data, duplicates or errors.

Changes made:

- Sub-Category column is removed
- Sub-Category ID column header is changed to Category ID

Data type changes:

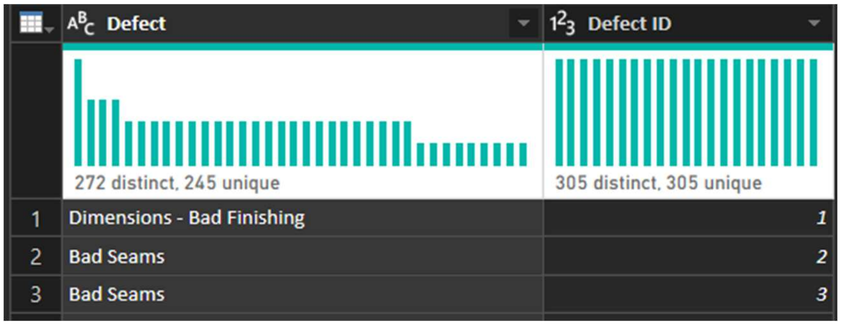
- **Category ID column** is changed from **Whole number** to **Text**



	Category ID	Category
1	1	Electrical
2	2	Logistics
3	3	Materials & Components
4	4	Mechanicals
5	5	Packaging
6	6	Goods & Services

Figure 2: Category Table After

Defect table



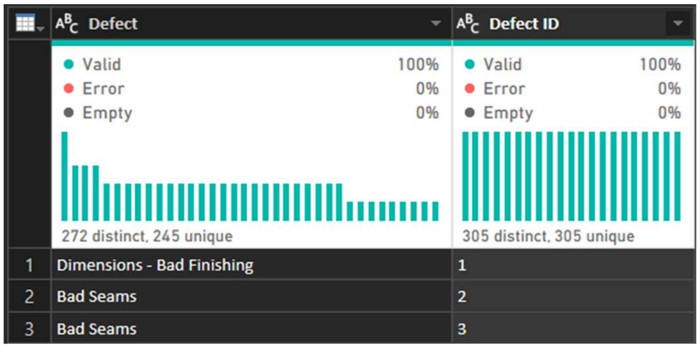
	Defect	Defect ID
	272 distinct, 245 unique	305 distinct, 305 unique
1	Dimensions - Bad Finishing	1
2	Bad Seams	2
3	Bad Seams	3

Figure 3: Defect Table Before

- The defect column has **305 defects with 272 distinct and 245 unique defects**. This indicates that there is a problem with duplicates, deleting them will drop **17%** of the records which will lead to serious problems so we could clarify that instead of deleting them. Also there are a lot of data that is quite similar like **“Wrong Size” and “Wrong Size”, “Warped” and “Warping”** which cannot be deleted for the same reason. Although deleting is not the only solution available but all solutions would be impractical in such situations.

Data type changes:

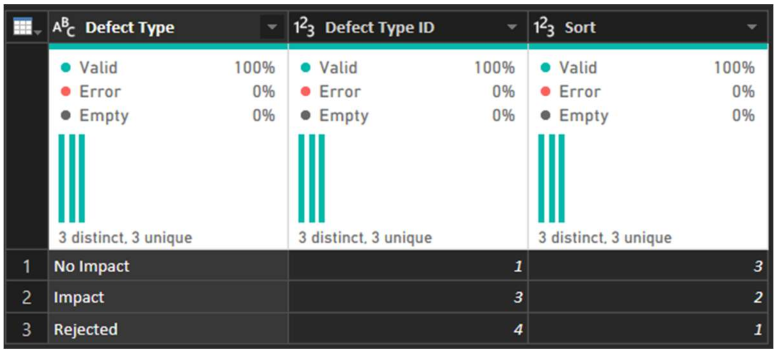
- **Defect ID column** is changed from **Whole number** to **Text**



	Defect	Defect ID
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	272 distinct, 245 unique	305 distinct, 305 unique
1	Dimensions - Bad Finishing	1
2	Bad Seams	2
3	Bad Seams	3

Figure 4: Defect Table After

Defect type table



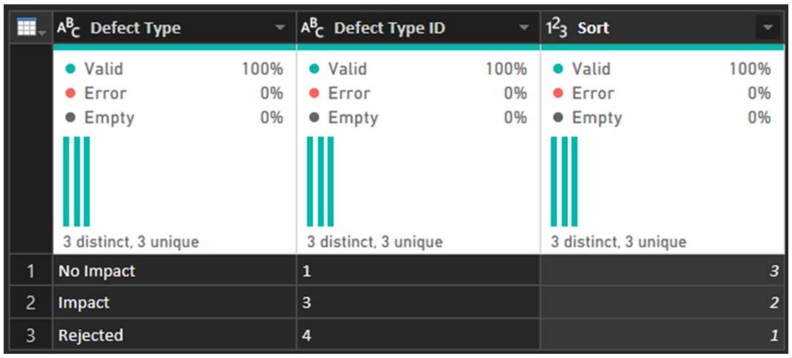
	Defect Type	Defect Type ID	Sort
	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique
1	No Impact	1	3
2	Impact	3	2
3	Rejected	4	1

Figure 5: Defect type Table Before

There are no missing data, duplicates or errors. The IDs only might seem odd but no problem with that.

Data Changes:

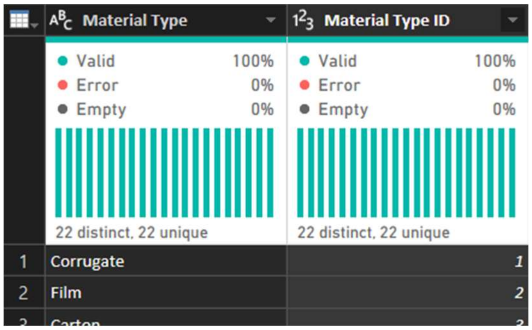
- **Defect Type ID column** is changed from **Whole number** to **Text**



	Defect Type	Defect Type ID	Sort
	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique	● Valid 100% ● Error 0% ● Empty 0% 3 distinct, 3 unique
1	No Impact	1	3
2	Impact	3	2
3	Rejected	4	1

Figure 6: Defect type Table After

Material table



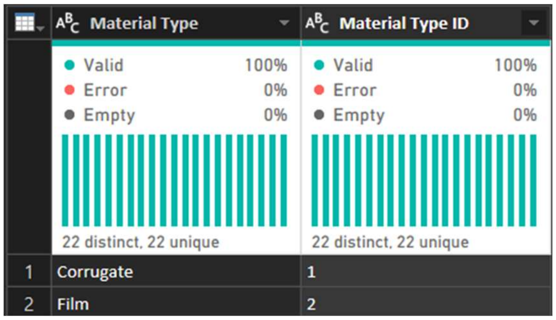
	Material Type	Material Type ID
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	22 distinct, 22 unique	22 distinct, 22 unique
1	Corrugate	1
2	Film	2
3	Carton	3

Figure 7: Material Table Before

There are no missing data, duplicates or errors.

Data changes:

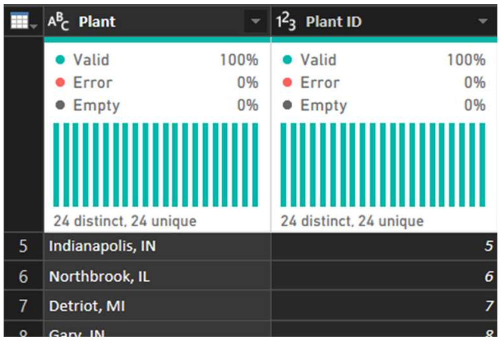
- **Material Type ID column** is changed from **Whole number** to **Text**



	Material Type	Material Type ID
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	22 distinct, 22 unique	22 distinct, 22 unique
1	Corrugate	1
2	Film	2

Figure 8: Material Table After

Plant Table



	Plant	Plant ID
	● Valid 100% ● Error 0% ● Empty 0% 24 distinct, 24 unique	● Valid 100% ● Error 0% ● Empty 0% 24 distinct, 24 unique
5	Indianapolis, IN	5
6	Northbrook, IL	6
7	Detriot, MI	7
8	Gary, IN	8

Figure 9: Plant Table Before

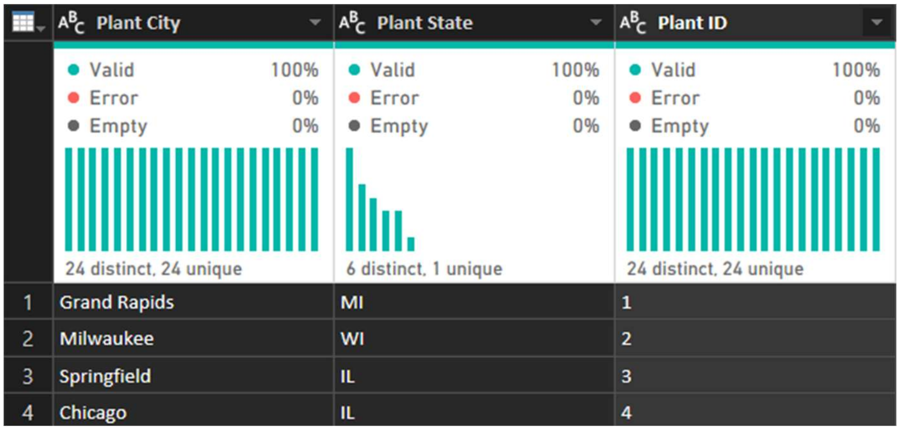
There are no missing data, duplicates or errors. It will be better to separate state abbreviations from the cities which can be used later to apply a hierarchy.

Changes made:

- Replaced **Cincinnati OH** with **Cincinnati, OH**.
- Split Plant column to two Plant.1 and Plant.2 using delimiters ",".
- Trimmed Plant.2 column
- Rename columns plant.1 = Plant City and plant.2 = Plant State
- Replaced **Wi** with **WI**

Data changes:

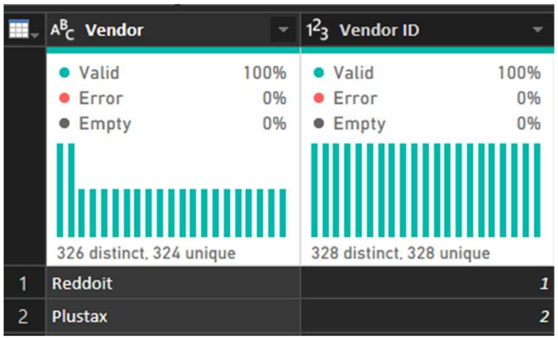
- **Material Type ID column** is changed from **Whole number** to **Text**.



	Plant City	Plant State	Plant ID
	● Valid 100% ● Error 0% ● Empty 0% 24 distinct, 24 unique	● Valid 100% ● Error 0% ● Empty 0% 6 distinct, 1 unique	● Valid 100% ● Error 0% ● Empty 0% 24 distinct, 24 unique
1	Grand Rapids	MI	1
2	Milwaukee	WI	2
3	Springfield	IL	3
4	Chicago	IL	4

Figure 10: Plant Table After

Vendor Table



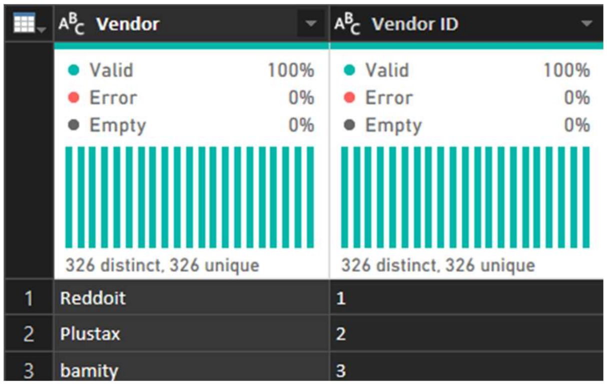
	Vendor	Vendor ID
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	326 distinct, 324 unique	328 distinct, 328 unique
1	Reddoit	1
2	Plustax	2

Figure 11: Vendor Table Before

There are no missing data or errors. But there are two duplicate **roundphase (80 & 113)** and **Quotefix (125 & 144)**. But deleting them may cause problems as they are recorded in the report so we could **clarify** that **instead of deleting** them.

Data changes:

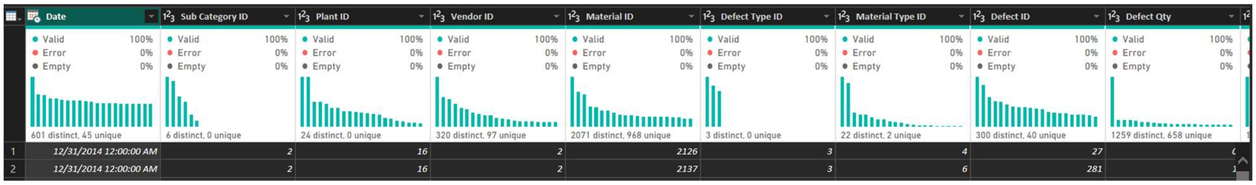
- **Vendor ID column** is changed from **Whole number** to **Text.**



	Vendor	Vendor ID
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	326 distinct, 326 unique	326 distinct, 326 unique
1	Reddoit	1
2	Plustax	2
3	bamity	3

Figure 12: Vendor Table After

Defect Report Table:



	Date	Sub Category ID	Plant ID	Vendor ID	Material ID	Defect Type ID	Material Type ID	Defect ID	Defect Qty
	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%
	601 distinct, 45 unique	6 distinct, 0 unique	24 distinct, 0 unique	320 distinct, 97 unique	2071 distinct, 968 unique	3 distinct, 0 unique	22 distinct, 2 unique	300 distinct, 40 unique	1259 distinct, 658 unique
1	12/31/2014 12:00:00 AM	2	16	2	2126	3	4	27	
2	12/31/2014 12:00:00 AM	2	16	2	2137	3	6	281	1

Figure 13: Defect Report Table Before

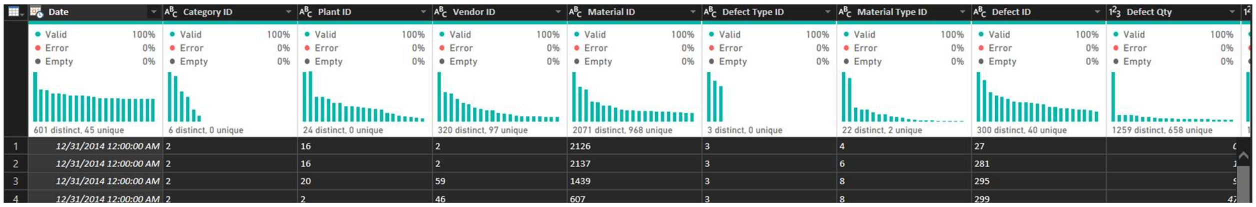
There are no missing data or errors. There is 6145 record, 193 of which are duplicates. The precious similarity in numbers may suggest a problem with the record entry that caused this entry. Anyways this should be sent back to the data owner but let’s assume the presence of the duplicates. After removing them the records will be 5952.

Changes made:

- Removed duplicate rows.
- Renamed the Sub Category ID to Category ID to fit the change done on the Category Table.

Data Changes:

- **Defect Type ID column** is changed from **Whole number** to **Text**
- **Category ID column** is changed from **Whole number** to **Text**
- **Material Type ID column** is changed from **Whole number** to **Text**
- **Plant ID column** is changed from **Whole number** to **Text**
- **Vendor ID column** is changed from **Whole number** to **Text**
- **Material ID column** is changed from **Whole number** to **Text**
- **Defect ID column** is changed from **Whole number** to **Text**



	Date	Category ID	Plant ID	Vendor ID	Material ID	Defect Type ID	Material Type ID	Defect ID	Defect Qty
	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%
	601 distinct, 45 unique	6 distinct, 0 unique	24 distinct, 0 unique	320 distinct, 97 unique	2071 distinct, 968 unique	3 distinct, 0 unique	22 distinct, 2 unique	300 distinct, 40 unique	1259 distinct, 658 unique
1	12/31/2014 12:00:00 AM	2	16	2	2126	3	4	27	
2	12/31/2014 12:00:00 AM	2	16	2	2137	3	6	281	1
3	12/31/2014 12:00:00 AM	2	20	59	1439	3	8	295	5
4	12/31/2014 12:00:00 AM	2	2	46	607	3	8	299	47

Figure 14: Defect Report Table After

Data Model

Objective of Data Model

The objective of data modeling is to organize the dataset into a clear and structured format for accurate analysis. This involves linking the **defect report table** with **related tables** such as category, vendor, material, plant, and defect to ensure consistency. **ID formats are standardized,** and a calendar table is created to manage dates properly. These steps help in tracking defects, evaluating supplier. With a well-structured **Star Schema**, the data becomes easier to analyze in Power BI, leading to better decision-making.

Model

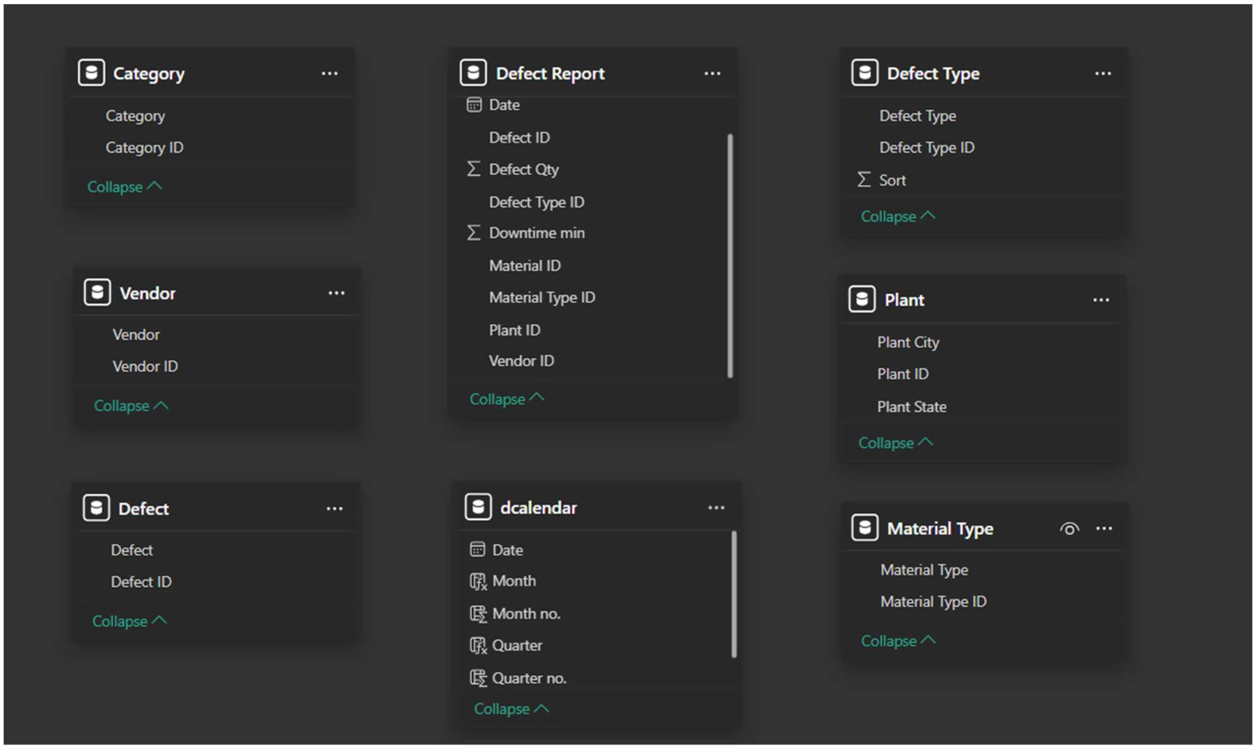


Figure 15: Before Data Modeling

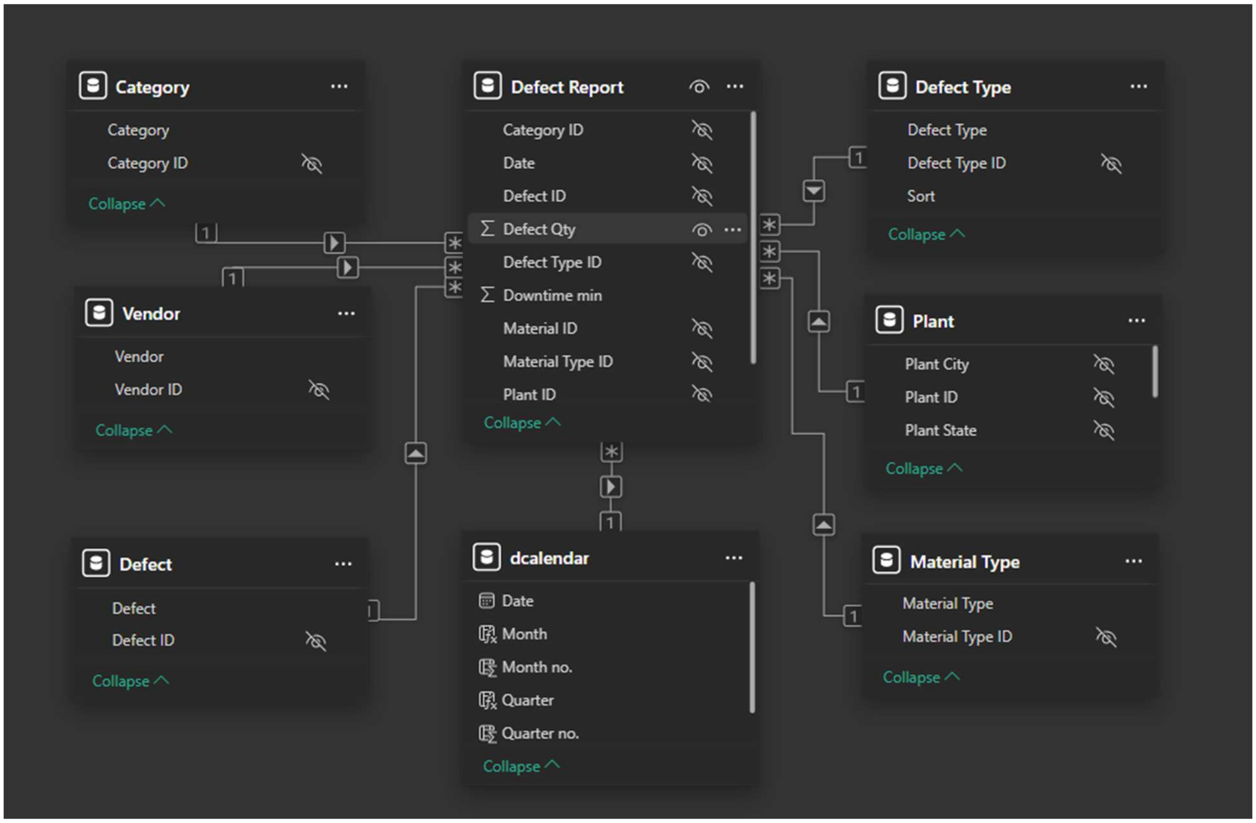


Figure 16: Data Modelled based on Star Schema