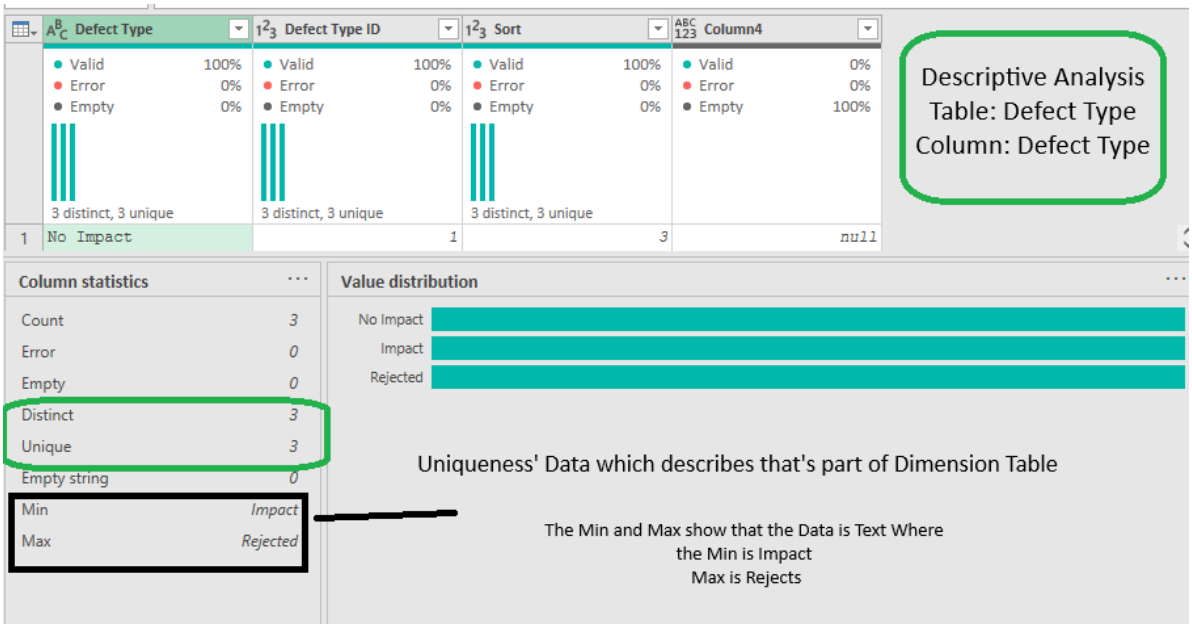
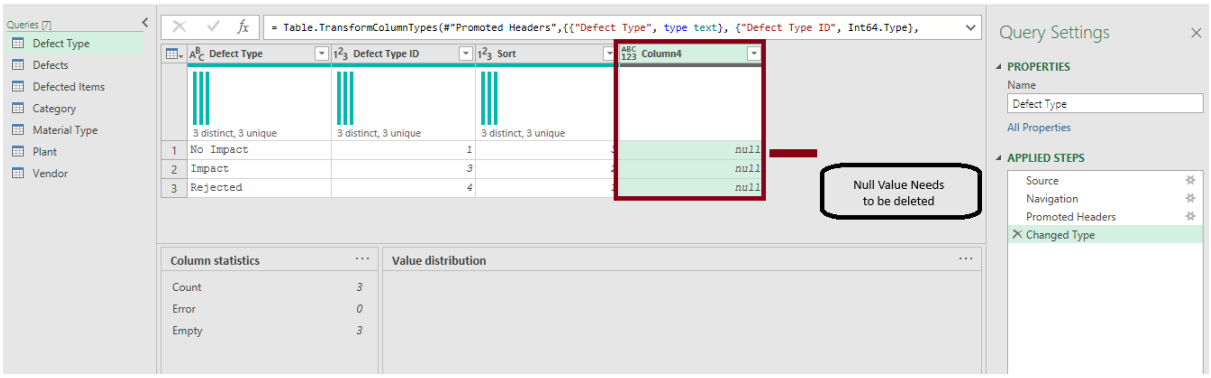
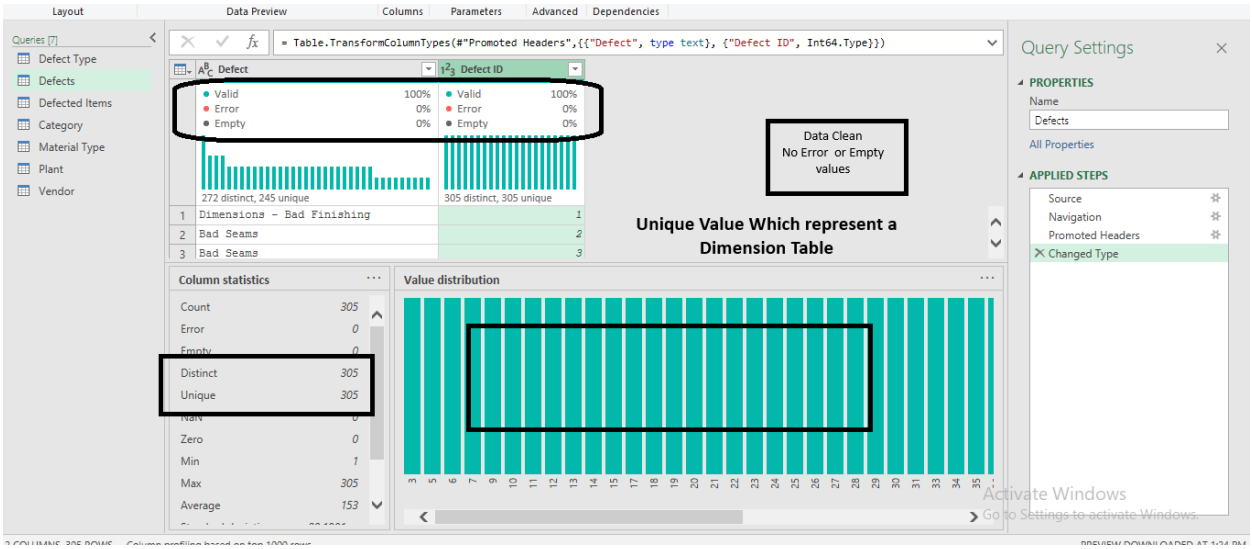
Data Cleaning and Preprocessing Steps (Power Query)

To ensure the accuracy and reliability of the dataset, several steps were undertaken using Power Query to clean and preprocess the data effectively. These steps ensured that the dataset was well-structured for further analysis and provided reliable insights.

Data Exploration





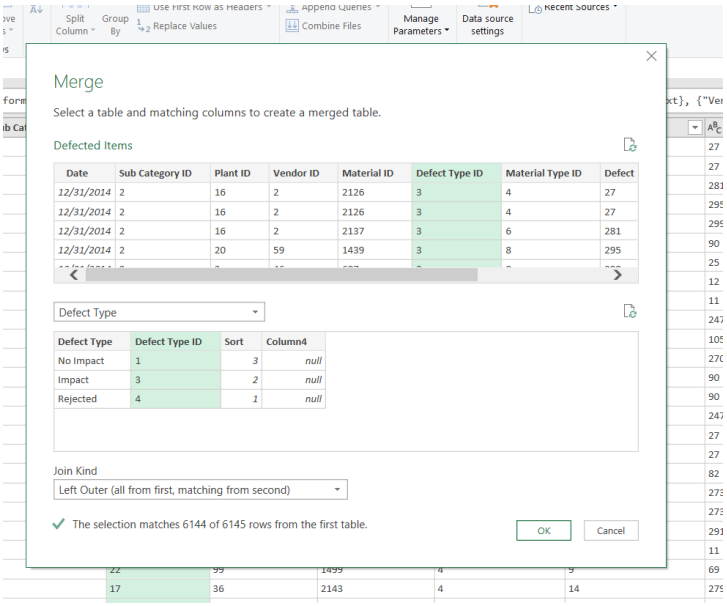


Data Cleaning

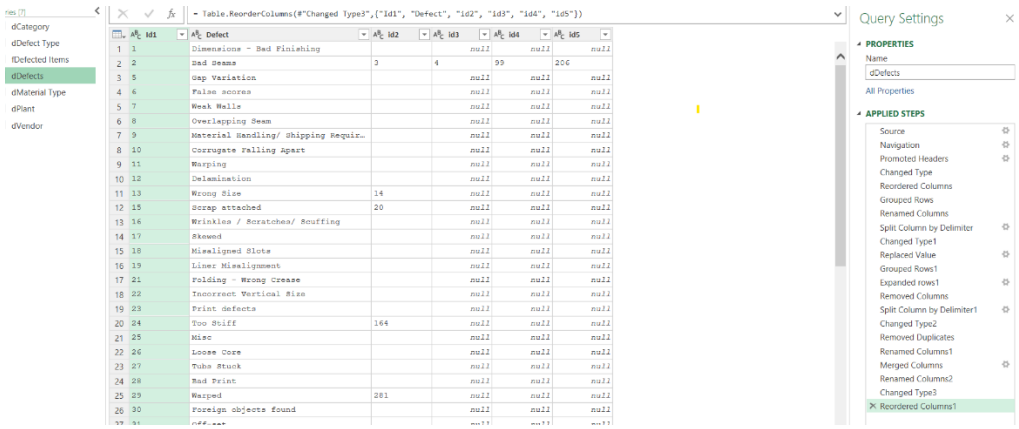
1. Merging Tables

The "Defected Items" sheet was merged with relevant dimension tables (e.g., Vendor, Plant, Material Type, Defect Type, Category) to enrich the dataset and provide additional context for each defect record.

- **Merging the sheets**:
  Using Power Query, the necessary sheets were imported. The `Merge Queries` function was used to combine the "Defected Items" sheet with other dimension tables by matching on key fields (`Vendor ID`, `Plant ID`, `etc`). This allowed the inclusion of additional attributes such as vendor performance and defect classifications.



- **Group By to Unify IDs**:
  In cases where the dimension sheets contained multiple IDs for the same item (e.g., vendors or materials), the `Group By` feature in Power Query was applied. For instance, the Vendor sheet was grouped by `Vendor Name`, and the `Vendor ID` was aggregated to ensure a unified ID for each vendor. The same method was applied to `Defects` to unify identifiers across the dataset.

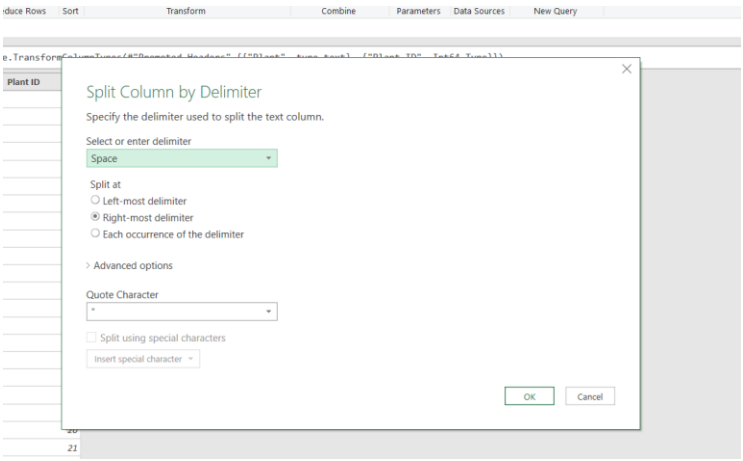


- **Distinct vs. Unique Check**:
  After merging, a check was conducted to ensure that key columns like `Vendor ID`, `Material ID`, and `Defect ID` contained only unique values. This was done by using the `Remove Duplicates` feature in Power Query to eliminate any unintended duplicates, ensuring that each record in these fields was unique.

2. Handling Missing Values

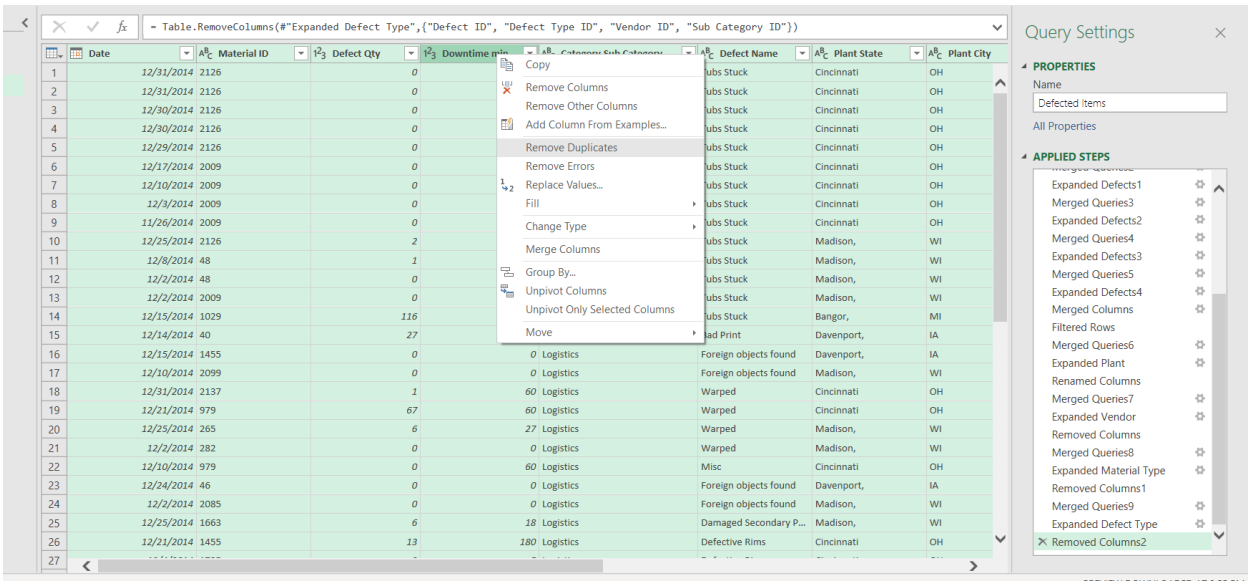
In columns such as `Downtime mins` and `Defected Qty`, zero values were observed instead of nulls. These zero values represented records where no defects were found or reported. These zeros were left unchanged, as they indicated valid data representing defect-free records and ensured data integrity.

3. Splitting the "Plant" Column



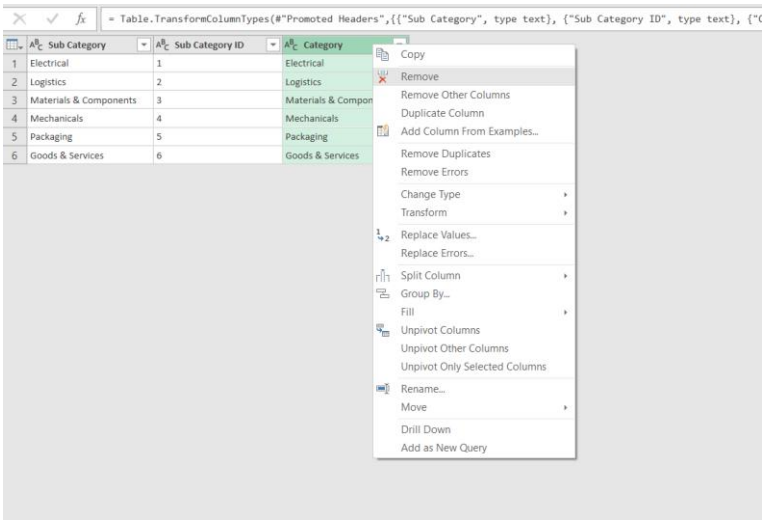
The `Plant Name` column contained both city and state information in the format "City, State" To facilitate location-based analysis, this column was split into two separate columns: `Plant City` and `Plant State`, using the right most space as a delimiter. This allowed for more granular analysis based on the geographic location of each plant.

4. Removing Duplicate Records



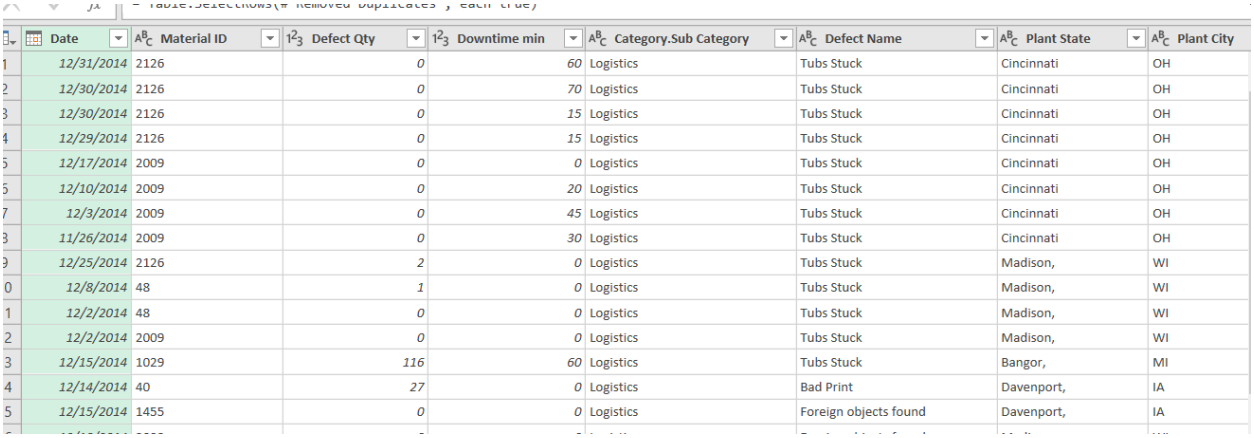
Duplicate rows in the dataset were identified and removed to ensure that each defect record was unique. The `Remove Duplicates` feature was applied across key fields such as `Defect ID` to eliminate redundant records and prevent any skewing of analysis results.

5. Removing the Duplicate "Category" Column



This redundant column was removed to avoid confusion and streamline the dataset, ensuring that only one instance of the `Category` column remained.

6. Data Type Conversion



	Date	Material ID	Defect Qty	Downtime min	Category.Sub Category	Defect Name	Plant State	Plant City
1	12/31/2014	2126	0	60	Logistics	Tubs Stuck	Cincinnati	OH
2	12/30/2014	2126	0	70	Logistics	Tubs Stuck	Cincinnati	OH
3	12/30/2014	2126	0	15	Logistics	Tubs Stuck	Cincinnati	OH
4	12/29/2014	2126	0	15	Logistics	Tubs Stuck	Cincinnati	OH
5	12/17/2014	2009	0	0	Logistics	Tubs Stuck	Cincinnati	OH
6	12/10/2014	2009	0	20	Logistics	Tubs Stuck	Cincinnati	OH
7	12/3/2014	2009	0	45	Logistics	Tubs Stuck	Cincinnati	OH
8	11/26/2014	2009	0	30	Logistics	Tubs Stuck	Cincinnati	OH
9	12/25/2014	2126	2	0	Logistics	Tubs Stuck	Madison,	WI
10	12/8/2014	48	1	0	Logistics	Tubs Stuck	Madison,	WI
11	12/2/2014	48	0	0	Logistics	Tubs Stuck	Madison,	WI
12	12/2/2014	2009	0	0	Logistics	Tubs Stuck	Madison,	WI
13	12/15/2014	1029	116	60	Logistics	Tubs Stuck	Bangor,	MI
14	12/14/2014	40	27	0	Logistics	Bad Print	Davenport,	IA
15	12/15/2014	1455	0	0	Logistics	Foreign objects found	Davenport,	IA

To ensure data consistency and support effective analysis, several data type conversions were performed:

- **Date Column**:
  The `Defect Date` column was converted to the `Date/Time` format to enable accurate time-based analysis, such as tracking defects over time.
- **Numeric Columns**:
  Columns such as `Defected Qty` and `Downtime mins` were converted to appropriate numeric types (e.g., `Whole Number` or `Decimal Number`) to ensure consistency and facilitate calculations.
- **Categorical Columns**:
  Categorical columns such as `Defect Type`, `Material Type`, and `Vendor Name` were converted to the `Text` data type to enable grouping, filtering, and category-based analysis.

7. Standardizing Column Names

Inconsistencies were found in the column names, such as spaces and varying capitalization. To resolve this, column names were standardized using Power Query's `Trim` and `Clean` functions to remove leading/trailing spaces and non-printable characters. Column names were then renamed to follow a consistent naming convention, improving readability and ensuring uniformity across the dataset.

8. Capitalization and Spell Check

- **Capitalization**:
  In columns such as `Vendor Name`, `Plant City`, `Plant State`, `Material Type`, and `Defect Type`, the first letter of each word was capitalized to improve consistency and readability across the dataset. This transformation was applied using Power Query's `Capitalize Each Word` function.
- **Spell Check on Defects Column**:
  A spell check was conducted on the `Defects` column to ensure that defect descriptions were correctly spelled. Common spelling errors were identified and corrected using the `Replace Values` function. This step helped maintain consistency in defect types and minimized the risk of misinterpretation during analysis.