Data Cleaning Project for Bicycle Manufacturing Company

Ons Chaabene

Introduction

This project involves working with data from a bicycle manufacturer and retailer. The objective is to clean and prepare various datasets related to products, sales, and resellers. The data is provided in an Excel sheet, and Power Query Editor was used for processing.

Loading Data and Promoting Headers

- Imported the **ProductData** Excel sheet into Power BI.
- The first row contained irrelevant data (gibberish) and was removed.
- Promoted the first actual row of data as column headers.
- Verified that the dataset contains 295 rows after cleaning.

Renaming and Reordering Columns

- Renamed **ProductVariantNameDescription** to **Product Name** for simplicity.
- Renamed **Category1** and **Category2** to **Subcategory** and **Category** to improve clarity.
- Reordered **Category** and **Subcategory** to reflect a logical hierarchy.
- Moved **Standard Cost** and **List Price** to the far right, ensuring all text columns appear first, followed by numerical columns. The dataset is now structured in a clearer and more logical manner.

Changing Data Types and Sorting

- Ensured data is in the correct format to optimize analysis and memory usage.
- Changed **Standard Cost** and **List Price** to a **Fixed Decimal Number** format, which is more suitable for currency values.
- Converted the **Product Key** column to a **Whole Number** format to maintain consistency with other numerical identifiers.

- Sorted the data first by the **Category** column in ascending order, then by the **List Price** column in descending order. By sorting data by multiple columns I have better insights.
- Ensured that each product is unique by removing rows with duplicate **ProductKey** values. This step helped maintain consistency.

Identifying and Correcting Issues in the Color Column

An inconsistency was observed in the **Color column** of our dataset, which appeared to be caused by typographical errors. The dataset is expected to contain exactly 10 distinct color categories, including a category for products without a color. However, upon inspecting the column, I identified several issues:

- The **Color column** contained 12 distinct values instead of the expected 10.
- One of the unique values appeared only once, suggesting a possible data entry error.

To address these issues, I took the following steps to clean the data:

- Removed duplicates from the **Color column** to identify any incorrect entries.
- Corrected misspellings:
  - Replaced **multi** with **Multi**.
  - Replaced **yeelow** with **Yellow**.
- Deleted the previous step of duplicate removal after making the corrections.
- Filtered the **Color column**:
  - Selected only the rows where the color was **Multi**, resulting in a subset of 8 rows.

After the cleaning process, the updated distribution of the **Color column** is shown in the figure below:

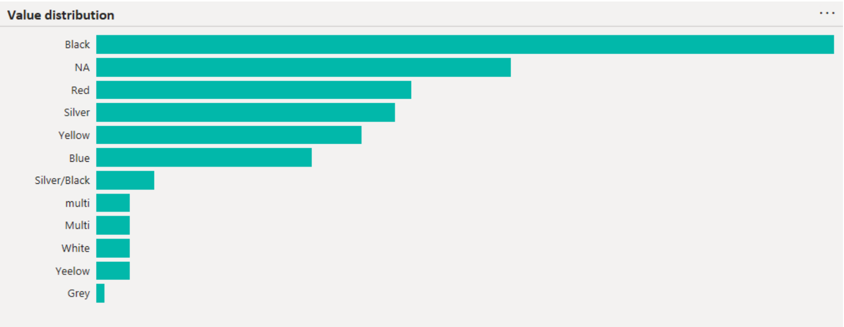


Figure 1: Distribution of the Color column before cleaning

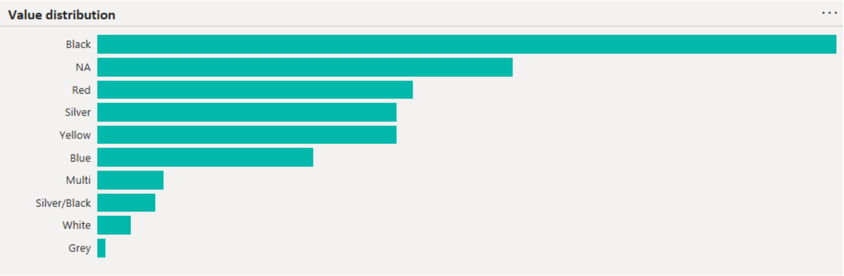
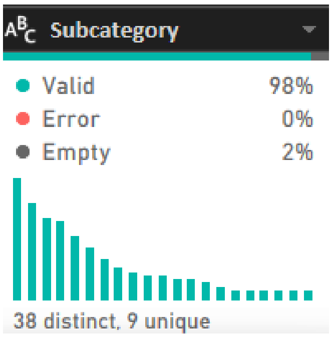


Figure 2: Distribution of the Color column after cleaning

Handling Missing Values in Subcategory Column

The **subcategory column** contains missing values. The following steps were taken to address this:

- Identified missing values in the **subcategory column**.
- Filtered the data to display rows with blank values in the **subcategory column**.
- Replaced blank values with **null** to standardize the missing data.
- Sorted the **ProductKey** column in ascending order to ensure proper organization.
- Filled down the missing values in the **subcategory column** using a fill transformation, ensuring that each missing value is replaced by the most recent non-missing value in the column using **FillDown**.

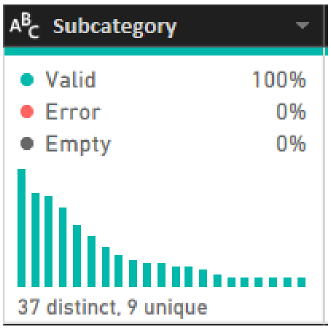


Subcategory

● Valid 98%
● Error 0%
● Empty 2%

38 distinct, 9 unique

Figure 3: **Column Subcategory Data Quality Before Handling Missing Values**



Subcategory

● Valid 100%
● Error 0%
● Empty 0%

37 distinct, 9 unique

Figure 4: **Column Subcategory Data Quality After Handling Missing Values**

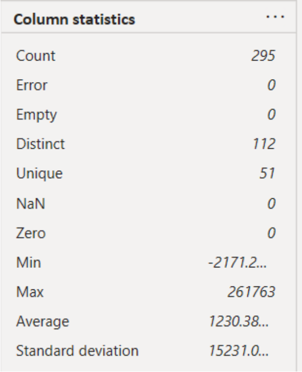
After performing these operations, the **subcategory column** now contains 37 distinct values, with 100% validity.

Handling Issues in the Standard Cost Column

The **Standard Cost** column in the product catalog showed anomalies upon inspection. The minimum value was negative, and the maximum value was unusually high. To address these issues, the following steps were taken:

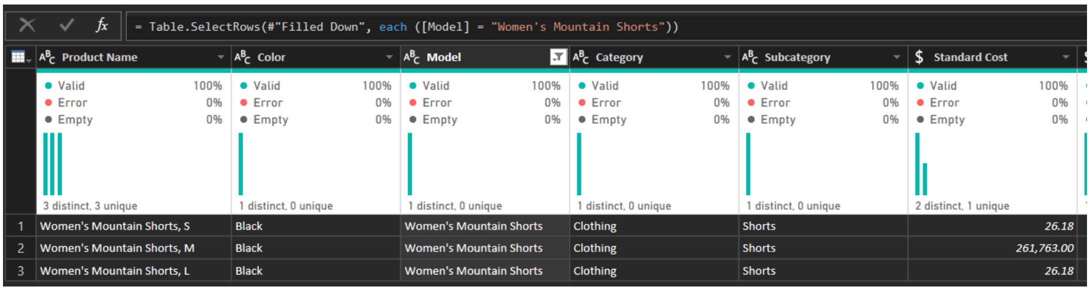
- Identified the outliers by reviewing the column profile, noting a negative minimum value and an unusually high maximum.
- Sorted the values to organize the data in ascending order, allowing easier identification of discrepancies.
- Replaced negative values using the **Replace Values** transformation to convert them into positive numbers, ensuring consistency in the dataset.
- Reviewed specific entries, such as the ”Women’s Mountain Shorts” model, where a missing decimal point was found in one of the values. The outlier was corrected by replacing it with the accurate value.

The following figures illustrate the identified issues and the steps taken to resolve them:



Column statistics	
Count	295
Error	0
Empty	0
Distinct	112
Unique	51
NaN	0
Zero	0
Min	-2171.2...
Max	261763
Average	1230.38...
Standard deviation	15231.0...

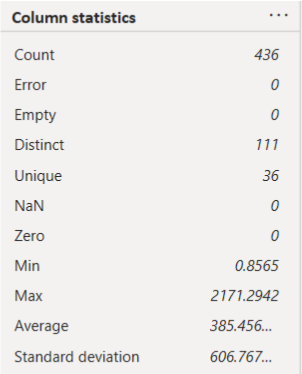
Figure 5: **Incorrect values in the Standard Cost column.**



```
= Table.SelectRows(#"Filled Down", each ([Model] = "Women's Mountain Shorts"))
```

	Product Name	Color	Model	Category	Subcategory	Standard Cost
	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%
	3 distinct, 3 unique	1 distinct, 0 unique	1 distinct, 0 unique	1 distinct, 0 unique	1 distinct, 0 unique	2 distinct, 1 unique
1	Women's Mountain Shorts, S	Black	Women's Mountain Shorts	Clothing	Shorts	26.18
2	Women's Mountain Shorts, M	Black	Women's Mountain Shorts	Clothing	Shorts	261,763.00
3	Women's Mountain Shorts, L	Black	Women's Mountain Shorts	Clothing	Shorts	26.18

Figure 6: Outlier value in Standard Cost column



Column statistics	
Count	436
Error	0
Empty	0
Distinct	111
Unique	36
NaN	0
Zero	0
Min	0.8565
Max	2171.2942
Average	385.456...
Standard deviation	606.767...

Figure 7: **Standard Cost column After Fixing Errors.**

Trimming, Cleaning, and Formatting Reseller Data

In preparation for analysis, the dataset containing all resellers of the bike company was reviewed and cleaned to ensure consistency and accuracy. The following steps were taken:

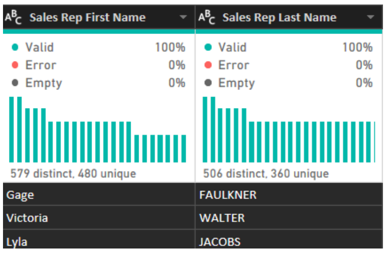
- Opened the **ResellerData** dataset, which contains the information of all resellers.
- Selected all columns for cleaning and applied trimming to remove any leading or trailing spaces.
- Standardized the formatting of the **Business Type** column by capitalizing the first letter of each word, ensuring consistency in naming conventions.
- After cleaning, the **Business Type** column now contains 3 distinct values, reflecting the standardized business types.
- Ensured that the **Reseller** column followed the same cleaning process to remove inconsistencies and ensure uniformity across entries.

Implementing Naming Convention for Reseller Sales Representatives

To distinguish reseller sales representatives I implemented a naming convention. The following steps were performed:

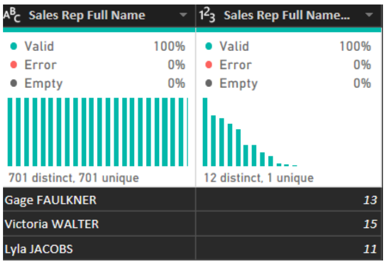
- Verified the length of the longest sales representative name in the dataset.
- Transformed the **Sales Rep Last Name** column to uppercase to ensure consistency and improve readability.
- Combined the **Sales Rep Last Name** and **Sales Rep First Name** columns, using a space delimiter, and created a new column titled **Sales Rep Full Name**.
- Duplicated the **Sales Rep Full Name** column, then extracted the length of each name and renamed the new column as **Name Length**.
- Sorted the data by **Name Length** in descending order, with the longest name having a length of 19 characters.

These steps ensure the names of the reseller sales representatives are formatted consistently and in compliance with the CRM software's limitations.



Sales Rep First Name	Sales Rep Last Name
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
579 distinct, 480 unique	506 distinct, 360 unique
Gage	FAULKNER
Victoria	WALTER
Lyla	JACOBS

Figure 8: **Columns related to names before modification**



Sales Rep Full Name	Sales Rep Full Name...
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
701 distinct, 701 unique	12 distinct, 1 unique
Gage FAULKNER	13
Victoria WALTER	15
Lyla JACOBS	11

Figure 9: **Modified columns related to names**

Sales Data Processing

To ensure consistency in sales-related figures, I applied several transformations to the dataset from **SalesData** Excel sheet.

- Rounded all sale-related figures to one decimal place following company conventions.
- Used Power Query to analyze the column profile for the entire dataset.
- Examined the distribution of **Order Quantity** and **Extended Amount**.
- Applied an absolute value transformation to both columns.
- Rounded the **Unit Price** and **Extended Amount** to one decimal place.
- Fixed inconsistencies in the **Discount** and **Unit Price Discount Percentage** columns.
  - Converted **Discount Amount** to a negative value by multiplying by -1, aligning with company reporting standards.
  - Multiplied **Unit Price Discount Percentage** by 100 to correct its scale.
- Applied a retroactive tax by subtracting 1 from each row in the **Sales Total** column.
- Rounded **Sales Total** and **Sub Total** to one decimal place.
- Confirmed the highest discount percentage in the dataset is 13%.

Seasonality Analysis of Orders and Shipping

To analyze the seasonality of orders and shipping patterns, I performed the following transformations:

- Duplicated the **Order Date** column and extracted the **week of the month** to identify trends.
- Transformed the **Ship Date** column to display the corresponding weekday name.
- Verified that shipment providers often delay orders by one day (Tuesday instead of Monday)
- Assessed which week of the month most orders are received to optimize warehouse staffing.

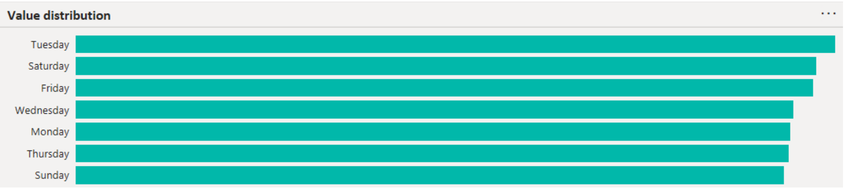


Figure 10: The ship day falls on a Tuesday

Conclusion

Through a systematic approach to data cleaning and preparation, the datasets related to products, sales, and resellers have been refined to ensure accuracy, consistency, and usability. Key steps included renaming and reordering columns for clarity, correcting inconsistencies in categorical data, handling missing values, and addressing anomalies in numerical fields. These transformations have improved data quality, making it more reliable for analysis and reporting. With a well-structured and cleaned dataset, further analysis and visualization can now be conducted efficiently to derive meaningful business insights.