**Data Cleaning Documentation**

|  |  |
| --- | --- |
| **Project Details** | |
| Client | Retail Company |
| **Project** | Retail Company Sales Performance Analysis |
| **Start Data** | 2023-10-26 |
| **End Date** | 2023-11-25 |
| **Email** | [Alaamhassan2001@gmail.com](mailto:Alaamhassan2001@gmail.com) |
| **GitHub** | <https://github.com/alaamhassan> |
| **LinkedIn** | [www.linkedin.com/in/alaamhassan](http://www.linkedin.com/in/alaamhassan) |
| **Portfolio** |  |

Table of Contents

[Metadata 2](#_Toc152597923)

[Changelog 3](#_Toc152597924)

[Cleaning Process 4](#_Toc152597925)

[1. Business Logic (does the data make sense? ) 4](#_Toc152597926)

# Metadata

|  |  |
| --- | --- |
| **Field** | **Description** |
| **Consumer ID** | Unique identifier for each customer. |
| **Order ID** | Unique identifier for each order. |
| **Month** | The month when the order was placed. |
| **Year** | The year when the order was placed. |
| **Total order value** | The total value of the order before any discounts are applied. |
| **Discount** | The amount of discount applied to the order. |
| **Line Value (net discount)** | The value of the order after the discount has been applied. |
| **Line Category** | The category of the product in the order. |
| **Line SKU** | Unique identifier for each product. |
| **Line SKU Production Cost** | The production cost of the product. |

# Changelog

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Change ID** | **Reason** | **Title of Change** | **Description** | **# Values** | **Process**  **Link** | **Version**  **Link** |
| **C1** | Data constraint violation. | Fixing mistyping of Line SKU Production Cost Value. | Change Line SKU Production Cost from 432,000 € to 43.2 €. | 1 | [C1-P](#C1_P) | [C1-V](https://github.com/alaamhassan/RetailCompany_PerformanceAnalysis/tree/44bc2d7178ee4e43471d926f714bceb54e3c5ad2) |
| **C2** |  |  |  |  |  |  |
| **C3** |  |  |  |  |  |  |

# Cleaning Process

## Business Logic (does the data make sense?)

Constrains based on data:

* Total order value >= Discount
* Line Value (net discount) = Total order value – Discount
* Line SKU Production Cost < Line Value (net discount)

Each constrain was checked using **conditional formatting**:

1) Total order value >= Discount

A screenshot of a computer

Description automatically generated

**Output**:

No value violates the constraint.

2) Line Value (net discount) = Total order value – Discount

A screenshot of a computer

Description automatically generated

**Output**:

No value violates the constraint.

A screenshot of a computer

Description automatically generated3) Line SKU Production Cost < Line Value (net discount)

**Output**:  
one value violates the constraint



**Correctness**:

**Steps**:

1. Filter for the ‘FRA5’ Line SKU

A screenshot of a computer

Description automatically generated

**Observations**:

170 records have the category ‘FRA5’. All these records have a Line SKU Production Cost of 43.2 €.

1. Change the Line SKU Production Cost of the first record from 432,000 € to 43.2 €.

A screenshot of a computer

Description automatically generated

## Check for Duplicates

Unique Constraints:

* A Record can’t be duplicated.
* Order ID can’t appear twice.

Check for each constrain using **Remove Duplicates** and **Conditional Formatting**:

1. A screenshot of a computer

   Description automatically generatedA Record can’t be duplicated.

**Output**:

No record violates the constraint.

A screenshot of a computer error

Description automatically generated

1. Order ID can’t appear twice.

Check for duplicates using **conditional formatting**:

A screenshot of a computer

Description automatically generated

**Output**:

Two values violate the constraint.



as the value of the Order ID is N/A, no values will be removed.

## Check for Mistyping

In the ‘Line Category’ column there are two categories with the same name:

* Mini bags
* Mini bag

A screenshot of a computer

Description automatically generatedbut one is plural and the other is singular.

**Correctness:**

As the two ‘Line Category’ have records with the same ‘Line SKU’. Then the two are likely the same.

**steps:**

1. Filter for the ‘Mini bag’ Line Category.
2. A screenshot of a computer

   Description automatically generatedChange ‘Mini bag’ to ‘Mini bags’ using **Find and Replace**.

Output:

A screenshot of a computer error

Description automatically generatedThree values were changes