Data Cleaning and Analysis of Retail Sales Dataset for Insight Generation

AYOMIDE MAKANJUOLA

OUTLINE

1. **Introduction**
   1.1. Background of the project
   1.2. Purpose and scope
   1.3. Dataset description

2. **Data Cleaning Process**
   2.1 Auto Fit Columns and Rows:
   2.2 Removing Duplicates
   2.3 Eliminate Blank Cells
   2.4 Use find and replace to correct Errors
   2.5 Using the Appropriate Data Types
   2.6 Convert Data To Table
   2.7 Validate data to be sure it is thoroughly clean.

3. **Conclusion**

1. Introduction

As part of my Data Analysis Course at Axia Africa, I carried out a project to clean, standardize, and analyze a structured retail sales dataset containing 31 transaction records. The primary objective was to improve its accuracy, completeness, and consistency, making it a reliable foundation for analysis.

The raw dataset exhibited common data quality challenges such as inconsistent name formatting, spelling errors in categorical fields, missing or invalid values, and irregular numeric entries. By applying systematic data cleaning techniques, I transformed the dataset into a high-integrity resource suitable for deeper analysis, trend identification, and actionable recommendations.

1.2 Purpose and Scope

This project was completed as part of the requirements for the 4-month Data Analysis training with Axia Africa. It provided a hands-on opportunity to apply critical data preparation skills, from detecting and resolving inconsistencies to structuring the data for meaningful insights. The ultimate goal was to demonstrate my ability to work with real-world dirty data and prepare it for business decision-making.

1.3 Dataset Description

The dataset contains entries of shopping records, consisting of the Date of Sale, the ID and Name of the Customer, the Product bought, the quantity bought and the price per unit of the Product bought. The dataset has 31 rows and 8 columns:

- Date: This Contains information for the date of the purchase.
- ID: This Contains unique Identifier for each customer.
- Name: This Contains information on the Name of the customer.
- Region: This Contains the location for the sale.
- Rating: This Contains the information on the customer experience.
- Product: This Contains the Product information.
- Quantity: This Contains the Information in the Quantity of the Product Bought.
- Price Per Unit: This Contains the information on the unit price of the product.

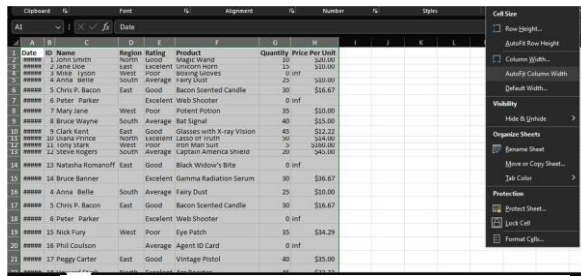|	Date	ID	Name	Region	Rating	Product	Quantity	Price Per Unit
1								
2	#####	1	John Smith	North	Good	Magic Wand	10	$20.00
3	#####	2	Jane Doe	East	Excelent	Unicorn Horn	15	$10.00
4	#####	3	Mike Tyson	West	Poor	Boxing Gloves	0	inf
5	#####	4	Anna Belle	South	Average	Fairy Dust	25	$10.00
6	#####	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
7	#####	6	Peter Parker		Excelent	Web Shooter	0	inf
8	#####	7	Mary Jane	West	Poor	Potent Potion	35	$10.00
9	#####	8	Bruce Wayne	South	Average	Bat Signal	40	$15.00
10	#####	9	Clark Kent	East	Good	Glasses with X-ray Vision	45	$12.22
11	#####	10	Diana Prince	North	Excelent	Lasso of Truth	50	$14.00
12	#####	11	Tony Stark	West	Poor	Iron Man Suit	5	$160.00
13	#####	12	Steve Rogers	South	Average	Captain America Shield	20	$45.00
14	#####	13	Natasha Romanoff	East	Good	Black Widow's Bite	0	inf
15	#####	14	Bruce Banner		Excelent	Gamma Radiation Serum	30	$36.67
16	#####	4	Anna Belle	South	Average	Fairy Dust	25	$10.00
17	#####	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
18	#####	6	Peter Parker		Excelent	Web Shooter	0	inf
19	#####	15	Nick Fury	West	Poor	Eye Patch	35	$34.29
20	#####	16	Phil Coulson		Average	Agent ID Card	0	inf
21	#####	17	Peggy Carter	East	Good	Vintage Pistol	40	$35.00

A Brief Overview of the Dataset

2. Data Cleaning Process

Microsoft Excel was used for the data cleaning and initial exploration stages. Several issues within the dataset were identified and rectified, as summarized below:
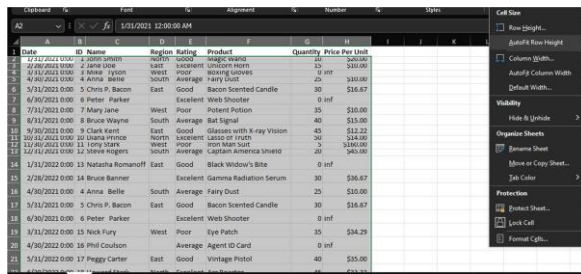
2.1 Auto Fit Columns and Rows:

Due to the disparities in the data, and some Columns not showing properly due to the column length, I ensured all columns and rows were resized to fit their contents for better readability.
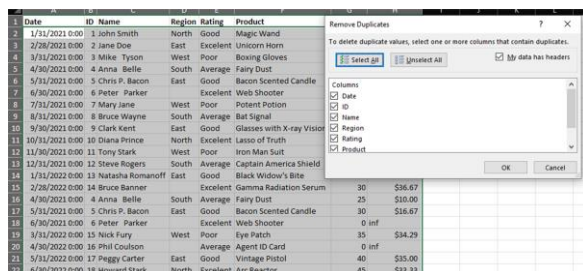

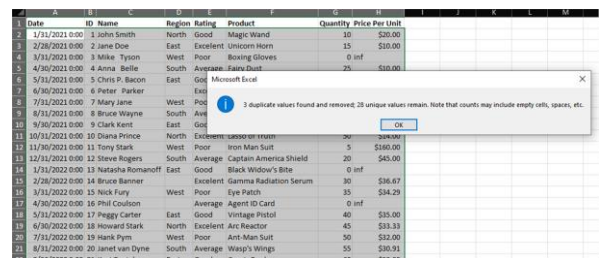
Before Autofit Column

After Autofit column

Before Autofit Row

After Autofit Rows

2.2 Removing Duplicates

With further steps to improving data quality, duplicate values were removed, 3 duplicates were removed in total and 28 unique values were remaining
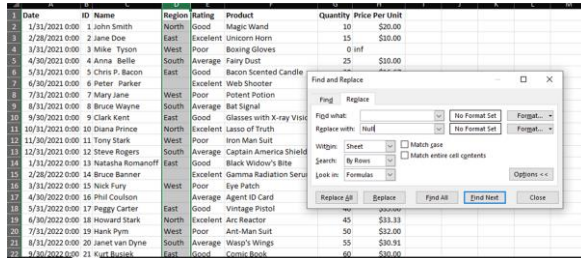
Before Removing Duplicates
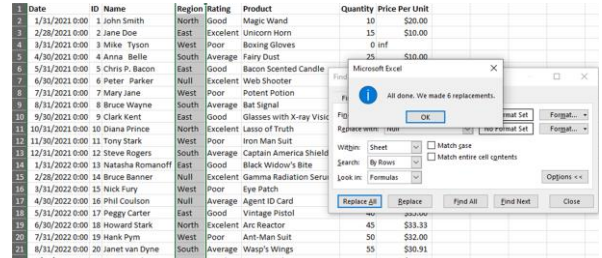
After Removing Duplicates

2.3 Eliminate Blank Cells

To continue with the data cleaning, blank cells needed to be fixed. A total of 6 blank cells were found at the Region Column, and after checking for the trend in the data and its consistencies, Null values were better replaced in the empty spaces than guessing or removing the entire row.
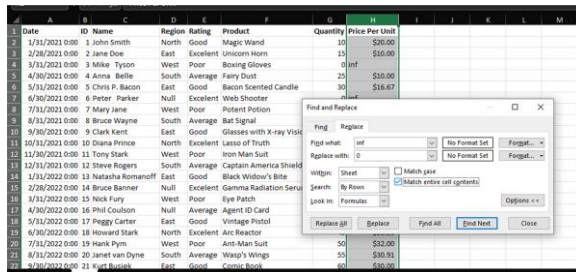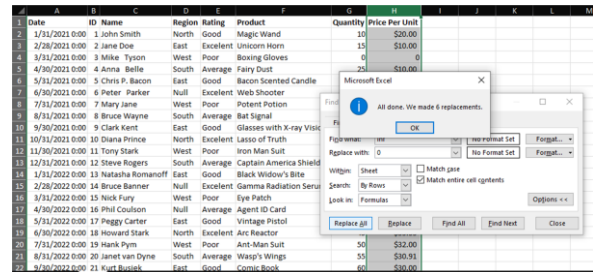

Before Eliminating Blank Cells

After Eliminating Blank Cells

2.4 Use find and replace to correct Errors

In the data, at the price per unit column, the column contained values like 'inf' which doesn't actually depict what should be in the column. The column is better represented with zero as its consistent with the zero on the quality column. After changing the 'inf' to 0, a total of 6 replacements were made and the data looks better for analysis and insights.

Before using find and replace

After using find and replace

2.5 Using the Appropriate Data Types

The Data types of the columns were incorrect, so there was a need to put the individual columns in their correct data types, which proved to be good, as the data became perfect for analysis and insights.

Before changing data types

After changing data types

2.6 Convert Data To Table:

For the closing steps of the data cleaning, the data is converted to Table, so it can be further used to generate insights for perform further calculative or visualization tasks.

Before converting data to table

After converting data to table

2.7 Validate data to be sure it is thoroughly clean.

For the last steps, the Data quantity was validated and was put to effect for other columns.

Validating data

3.0 Conclusion

This project demonstrated the importance of thorough data cleaning in ensuring the accuracy and reliability of datasets used for analysis. By systematically applying Excel cleaning techniques, the dataset was transformed into a well-structured and reliable resource for generating actionable insights.

Date	ID	Name	Region	Rating	Product	Quantity	Price Per Unit
1/31/2021	1.00	John Smith	North	Good	Magic Wand	10.00	$ 20.00
2/28/2021	2.00	Jane Doe	East	Excelent	Unicorn Horn	15.00	$ 10.00
3/31/2021	3.00	Mike Tyson	West	Poor	Boxing Gloves	0.00	$ -
4/30/2021	4.00	Anna Belle	South	Average	Fairy Dust	25.00	$ 10.00
5/31/2021	5.00	Chris P. Bacon	East	Good	Bacon Scented Candle	30.00	$ 16.67
6/30/2021	6.00	Peter Parker	Null	Excelent	Web Shooter	0.00	$ -
7/31/2021	7.00	Mary Jane	West	Poor	Potent Potion	35.00	$ 10.00
8/31/2021	8.00	Bruce Wayne	South	Average	Bat Signal	40.00	$ 15.00
9/30/2021	9.00	Clark Kent	East	Good	Glasses with X-ray Vision	45.00	$ 12.22
10/31/2021	10.00	Diana Prince	North	Excelent	Lasso of Truth	50.00	$ 14.00
11/30/2021	11.00	Tony Stark	West	Poor	Iron Man Suit	5.00	$ 160.00
12/31/2021	12.00	Steve Rogers	South	Average	Captain America Shield	20.00	$ 45.00
1/31/2022	13.00	Natasha Romanoff	East	Good	Black Widow's Bite	0.00	$ -
2/28/2022	14.00	Bruce Banner	Null	Excelent	Gamma Radiation Serum	30.00	$ 36.67
3/31/2022	15.00	Nick Fury	West	Poor	Eye Patch	35.00	$ 34.29
4/30/2022	16.00	Phil Coulson	Null	Average	Agent ID Card	0.00	$ -
5/31/2022	17.00	Peggy Carter	East	Good	Vintage Pistol	40.00	$ 35.00
6/30/2022	18.00	Howard Stark	North	Excelent	Arc Reactor	45.00	$ 33.33
7/31/2022	19.00	Hank Pym	West	Poor	Ant-Man Suit	50.00	$ 32.00
8/31/2022	20.00	Janet van Dyne	South	Average	Wasp's Wings	55.00	$ 30.91
9/30/2022	21.00	Kurt Busiek	East	Good	Comic Book	60.00	$ 30.00
10/31/2022	22.00	George Perez	North	Excelent	Drawing Pad	0.00	$ -

AN OVERVIEW OF THE FINAL DATASET