PROJECT: CLEANING OF A MESSY DATA FOR ANALYSIS

INTRODUCTION:

PROBLEM STATEMENT: In this project, I'll be working with a messy dataset. As an analyst my job is to clean and organize the data using Microsoft Excel so it's ready for analysis.

From the dataset I noticed there's inconsistent entries, spacing issues, missing or incorrect values, and more.

OBJECTIVE: Main objective of this project is to clean and prepare the messy data given for analysis using different functions and techniques from text functions to filtering, sorting, and logical formulas.

PROJECT QUESTIONS TO CARRYOUT:

1. Autofit Columns and Rows.

2. Identify and Remove Duplicates.

3. Trim Extra Spaces.

4. Eliminate Blank Cells.

5. Convert Data into Table.

6. Use Find and Replace to correct errors.

7. Validate data to be sure it is thoroughly clean.

TOOL USED: MS EXCEL

DATA TRANSFORMATION PROCESSES:

The following cleaning steps were taken:

- **Copy & Paste the data:** After importing my messy data I copied the data to another sheet to work on so as to maintain the originality and authenticity of the given messy data.
- Click on the first cell --- ctrl A, it highlight all the data --- ctrl C --- ctrl V
- **Auto fits rows and columns:** I autofit the rows and columns by manually place the cursor on rows and columns and double tap on it.
- **Check for duplicate records:** All datasets were reviewed for duplicate entries. A total of 3 duplicates were identified and removed using the ID column.
- **Trimming extra spaces:** Extra spaces was trimmed off using the Trim function in excel.

- **Date Formatting:** The Date column contained both date and time. It was reformatted to short date for consistency.
- **Eliminate blank cells:** Empty columns were identified and removed. Blank cells in the Region column was removed.
- **Convert data into Table format:** The data was converted into Table format using Ctrl T function and a beautiful design was selected.
- **Using Find & Replace:** Using Ctrl H function, transformation was done the following column;
  Region: "Asgard" was replaced with "East"
  Price per Unit: "inf" values were replaced with 0 to avoid calculation errors.
  Ratings: Spelling errors like "Excelent" were corrected to "Excellent."
  Non-rating words such as:
  "worthy" → "Good"
  "leader" → "Excellent"
  "spy" → "Poor"
  "mischief" → "Average"

- **VALIDATION:** Data validation was done and it's properly checked that the data is thoroughly cleaned. All columns were restricted as follows:

- Date: 01/31/2021 – 4/30/2023
- ID: 1 – 29
- Quantity: 0 – 85
- Price per Unit: 0 – 162
- The dataset is now clean, well-structured, and ready for further analysis.

- **RECCOMENDATIONS:** The dataset is now cleaned, well-structured, and ready for analysis.