Project Documentation:

Handling Missing Data in Walmart Sales

---

1. Project Title

Handling Null Values in Walmart Sales Data.

2. Project Objective

The main goal of this project was to learn how to deal with missing information (called "null values") in a real-world sales dataset from Walmart. The focus was on using practical and logical methods to fill in these missing values, considering what each type of data represents.

3. Dataset Overview

The project used a realistic Walmart sales dataset. Before fixing any missing data, the first step was to understand the dataset's overall structure.

- Source: Simulated real-world Walmart dataset
- Columns: Order ID, Customer ID, Customer Name, Order Date, City, Region, Category, Quantity, Sales, Profit.
- Rows: 100

4. Missing Values Summary

Column	Missing Values	% of Total	Action Taken
Order ID	7	7%	Unknown
Customer ID	6	6%	Unknown
Customer Name	3	3%	Unknown
Order Date	15	15%	Manual, based on Month / region
City	12	12%	Unknown
Region	4	4%	Manual by Global location

Category	7	7%	Manual, frequency + region
Quantity	7	7%	Manual, Date frequency + Category + Region
Sales	11	11%	Average by Category / Region
Profit	12	12%	Filled by 0.0

5. Null Value Handling Techniques Used

Different types of columns needed different approaches to handle missing values:

1) Identifier Columns (e.g., Order ID, Customer ID, Customer Name):

If these unique identification numbers or names were missing, they were simply filled with the word "**Unknown**." This was done because these are unique labels and can't be guessed or calculated from other data.

2) Date Columns (e.g., Order Date):

Rather than using random fill or forward fill, I used a smart, realistic strategy combining:

- The Year/month distribution
- Regional patterns
- Data proportions

This approach ensures that filled values reflect real-world business logic and to avoid misleading future analysis.

3) Categorical Columns (e.g., City, Region, Category):

- **City:** Missing city names were filled with "**Unknown**" because cities are often unique and don't easily follow patterns for guessing.
- **Region:** Missing region names were filled in by looking at the associated city. For example, if a city belongs to a specific region, that region was assigned. If no city-to-region match was found, the row was removed.

- **Category:** Instead of randomly imputing missing Category values, I performed both frequency analysis and region-category relationship analysis. Based on this, I manually filled null values to preserve realistic patterns in the data.

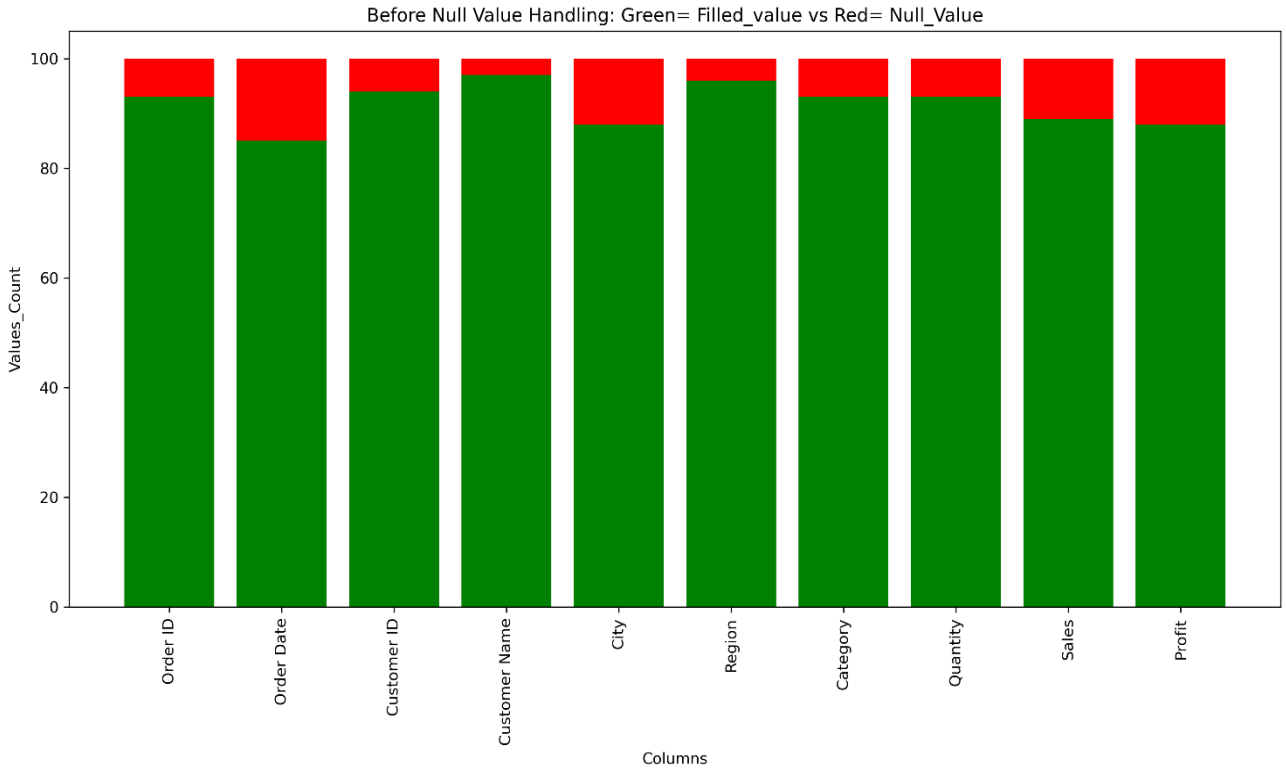
4) Numerical Columns (e.g., Quantity, Sales, Profit):

- **Quantity:** For missing product quantities, a more detailed approach was used. This involved looking at the product's category, region, and date to find similar sales records and then manually deciding on a reasonable quantity. In a few cases, if there wasn't enough information to make a good guess, the rows were removed.
- **Sales:** Missing sales figures were filled by calculating the average sales for similar groups of products or transactions. Sometimes, manual decisions were made if a clear logical value could be assigned.
- **Profit:** Missing profit values were imputed with 0.0 to represent no profit or unknown.

6) Visual Aids (e.g., Before, After)

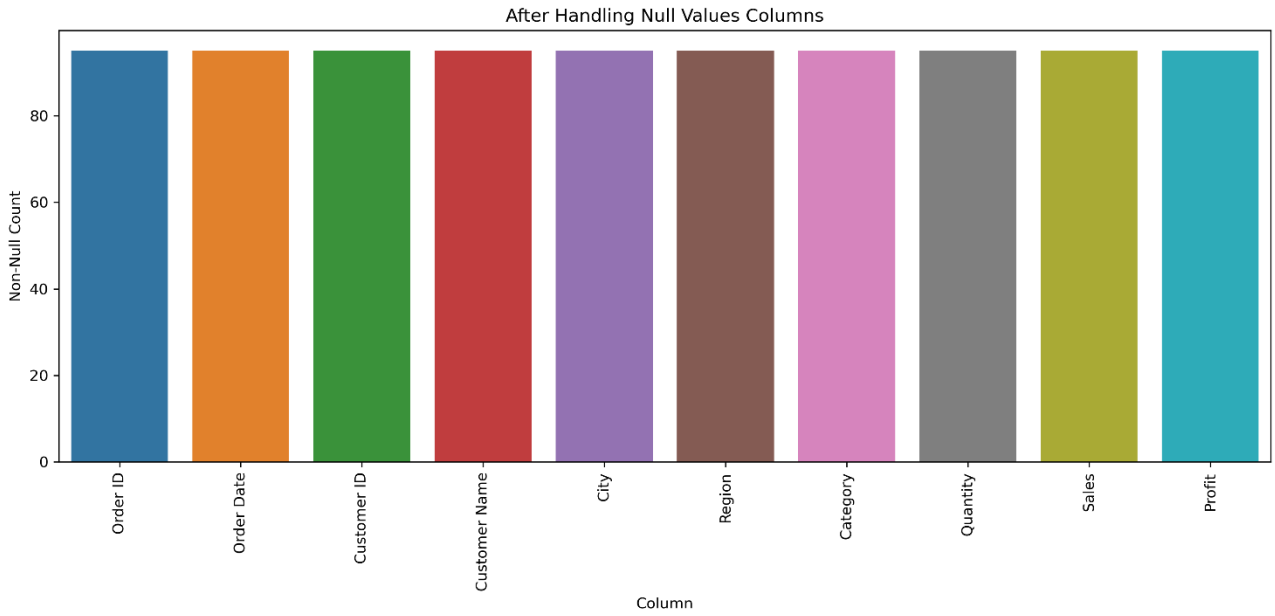
1) **Before:**

Columns = 10, Rows = 100, Total missing values = 84 – Bar plot of Missing values and Filled values per column using Matplotlib library.



2) **After:**

Columns = 10, Rows = 95, Total missing values = 0 – Bar plot of Filled values per column using Seaborn library.



7) Key Learnings

This project showed that handling missing data isn't a one-size-fits-all task. It's important to:

- **Understand your data:** Know what each column means.
- **Think logically:** Use real-world knowledge to decide how to fill in missing values.
- **Be flexible:** Different types of data (identifiers, dates, categories, numbers) need different strategies.

By applying these thoughtful methods, the dataset became cleaner and more reliable for future analysis, ensuring that any insights gained are based on more complete and accurate information.

8) Tools Used

- Jupyter Notebook
- Python 3.x
- Numpy
- Pandas
- Matplotlib
- Seaborn