Data Wrangling Report

Table of Contents

- **Note**: I have included these steps to make it easier to navigate through each step I performed in the wrangling process.

Introduction

This project is about data cleaning. We want to make sure the data is correct and ready for analysis.

Goals of Data Wrangling

The goal of this wrangling is to have clean data. This helps us do better analysis.

Dataset Description

This dataset has sales information from many branches. It has 16 columns and 1006 rows. The data includes:

- Invoice number
- Branch
- Customer type
- Product type
- Quantity sold
- Price per unit
- Taxes
- Total
- Payment method
- Customer ratings

Data Assessing

Trend Analysis

Each column has values like:

- Yangon
- Naypyitaw
- Mandalay

Quality Assessment

1. Completeness:

- Remove "USD" from the unit price columns.
- Remove "(pm)" and "(-)" from the time columns.
- For empty "Total" values, calculate: Quantity * Unit Price + 5% Tax.
- Fill empty "5% Tax" values.
- Replace "(-)" with the most common value.

2. Validity:

- Change column names to: 'Invoice_ID', 'Customer_type', 'Product_line', 'Unit_price', 'Tax_5%'.
- Change unit price type from object to int.
- Change date type from int to date.

3. Accuracy:

- Change rating (97.0) to (9.7).
- Replace negative values.

4. Consistency:

- Remove duplicate entries.

Data Cleaning

Cleaning Steps Taken

- Completeness: Removed "USD" from unit price columns, removed "(pm)" and "(-)" from time columns, filled empty "Total" values by calculating Quantity * Unit Price + 5% Tax, filled empty "5% Tax" values, replaced "(-)" with the most common value.
- Validity: Changed column names to "Invoice_ID", "Customer_type", "Product_line", "Unit_price", "Tax_5%", changed unit price type from object to int, changed date type from int to date.
- Accuracy: Changed rating (97.0) to (9.7), replaced negative values.
- Consistency: Removed duplicate entries.

Data Assessing After Cleaning

"The data has been cleaned. There are no more duplicates, no empty values, no inconsistent data types, and no out-of-range values

Data Storing

The cleaned data has been saved in a CSV file using Pandas and is now ready for analysis."