Data Wrangling Report

Overview

This report outlines the data wrangling process applied to the supermarket sales dataset, detailing the steps taken to clean, transform, and prepare the data for analysis.

1. Handling Missing Values

- Checked for missing values in all columns.
  - Recalculated `Tax 5%` using formula: `(Unit price × Quantity) × 0.05`
  - Recalculated `Total` as `(Unit price × Quantity) + Tax 5%`

2. Correcting Data Types

- **Date:** Converted to `datetime` format for accurate time-based analysis.
- **Time:** Standardized to a 24-hour format and removed AM/PM inconsistencies.
- **Unit Price:** Removed currency symbols and converted to a numeric type.

3. Handling Duplicates

- Checked for duplicate rows and removed them to maintain data integrity.

4. Cleaning Categorical Data

- **Customer Type:** Fixed typos (e.g., 'Memberr' → 'Member').
- **City Information:** Extracted correct city names from binary indicators.

5. Feature Engineering

- **City**: Created from binary columns `Yangon`, `Naypyitaw`, `Mandalay`
- **Time of Day**: Categorized into Morning (5AM-12PM), Afternoon (12PM-5PM), Evening (5PM-9PM), Night (9PM-5AM)
- **Day of Week**: Extracted from `Date` column to identify busy days
- **Weekend Indicator**: Created `Is Weekend?` column to identify weekend sales.
- **Revenue per Item**: Computed as `Total / Quantity` to measure per-unit revenue.

6. Handling Outliers

- Boxplots were used to detect outliers in numerical columns.
- Outliers in the Rating column were removed using IQR-based filtering.

7. Data Visulaization

- Made some insightful visuals that could help in decision making (More in Business Insight Report)

Conclusion

The dataset is now structured, cleaned, and ready for further analysis and visualization. The preprocessing steps improved data quality, ensured consistency, and eliminated errors that could affect business insights.