Data Wrangling Report

1. Data Collection and Understanding

The dataset consists of supermarket sales data, containing information such as transaction details, tax, total amount, ratings, and locations where sales occurred. The goal is to clean, analyze, and derive business insights from this data to enhance decision-making.

2. Data Cleaning

- **Missing Values**: A thorough check was conducted, and no missing values were found. This ensured data completeness and accuracy for analysis.
- **Data Types**: Each column's data type was validated. Numerical fields like Total Sales, Tax, and Quantity were confirmed to be in appropriate numerical formats, while categorical fields like City and Ratings were checked for consistency.
- **Duplicates**: The dataset was examined for duplicate records, which were removed to avoid biased analysis.
- **Formatting Standardization**: The dataset contained date-related fields that were standardized into consistent formats to enable better time-based analysis.
- **Categorical Encoding**: City names (Yangon, Naypyitaw, Mandalay) were encoded for numerical representation to allow for effective correlation analysis.
- **Scaling and Normalization**: Since numerical variables varied in scale, normalization techniques were applied where needed to maintain consistency in the dataset.

3. Feature Engineering

- **Extracting Time-Based Features**: Columns for Year, Month, and Day were derived from timestamps to facilitate time-based trend analysis.
- **Creating Revenue-Related Features**: A new feature for total revenue per transaction was computed for better business understanding.
- **Encoding Categorical Variables**: Location-based data was transformed into a numerical format to enable its inclusion in machine learning models.
- **Generating Aggregated Metrics**: Grouped total revenue per city and per time period to allow insights into trends over time and across locations.

4. Outlier Detection and Treatment

- **Boxplot and Statistical Analysis**: A combination of boxplots and statistical measures (IQR method) was used to detect outliers in numerical columns like total revenue and quantity sold.

- **Z-Score Analysis**: Applied to determine extreme values that might skew analysis results.
- **Decision on Outliers**: Outliers in quantity and revenue were analyzed, and some were retained if they provided business value, while extreme anomalies were adjusted to prevent distortions in insights.

5. Correlation Analysis

- **Key Observations from the Heatmap**:
    - **Strong Positive Correlation** between **Quantity Sold and Total Revenue (0.7)**, indicating that the more units sold, the higher the revenue.
    - **Strong Positive Correlation between Tax and Total Revenue (0.98)**, as expected since tax is applied as a percentage of total sales.
    - **Weak Correlation between Location (Yangon, Naypyitaw, Mandalay) and Total Revenue**, suggesting that geographical differences alone do not drive sales volume.
    - **Very Weak Correlation between Ratings and Total Sales**, implying that customer feedback does not have an immediate impact on revenue.

6. Data Visualization

- **Heatmaps**: Used to examine feature correlations and identify variables that influence total revenue.
- **Time-Series Graphs**: Plotted monthly and daily sales trends to understand seasonal variations in purchases.
- **Bar Charts**: Visualized total revenue contributions by city to determine if specific locations had a significant impact on overall sales.
- **Box Plots**: Used to detect anomalies and variations in numerical data such as sales and ratings.