**4. Implementation and Results**

4.1 Implementation

4.1.1. Data preparation

The data preparation started with loading, cleaning, and enhancing the dataset to ensure it was accurate and ready for analysis. These steps were executed systematically to prepare the data for building reliable models.

**Data Loading**

The retail transaction data was taken from two Excel files, covering the years 2009 to 2011. Both files were combined into one dataset to create a full record of transactions. Customer survey responses were imported from a CSV file. This survey data added valuable details about customer preferences and behaviors.

**What to Add**:

* **Table**: Show a summary of the combined dataset, including the number of rows, columns, and key features like transaction date, quantity, and price.
* **Image**: Display the first few rows of the dataset to show its structure.

**Data Cleaning**

The dataset had issues that needed fixing. Missing entries in the CustomerID column were removed because they made it impossible to track customer behaviors. Duplicate rows were identified and deleted to avoid counting the same transaction more than once. Transactions with negative values for quantity or price were also removed, as these were likely errors or returns. Columns that were not relevant to the project were dropped to make the data easier to work with.

**What to Add**:

* **Chart**: A bar graph showing the number of rows removed for each issue (missing values, duplicates, negative values).
* **Image**: Example of a problematic row and an explanation of how it was corrected.

**Feature Engineering**

New columns were added to the dataset to extract more useful information. The total value of each transaction (Total\_Amount) was calculated by multiplying quantity and price. Features like the time of day, day of the week, and month were extracted from transaction dates. Another column indicated whether the transaction occurred on a weekend.

Customer behaviors were analyzed more deeply using specific metrics:

* **Recency**: How recently a customer made a purchase.
* **Frequency**: How often a customer made purchases.
* **Monetary Value**: How much money the customer spent overall.

For each customer, additional metrics were calculated, such as their average spending, the largest and smallest transaction amounts, and how much their spending varied. The average time between purchases and how consistently customers made purchases were also calculated.

**What to Add**:

* **Table**: Highlight examples of the new features created, such as total amount, weekend indicator, and shopping frequency.
* **Chart**: A histogram showing shopping behavior by day of the week or time of day.
* **Table**: Show a summary of customer metrics like average spending, recency, and frequency.

**Outlier Handling**

Some transactions had values that were much higher or lower than usual. These extreme numbers were adjusted to prevent them from distorting the results. Setting reasonable limits for these values ensured the analysis and models were not skewed.

**What to Add**:

* **Chart**: A boxplot showing transaction amounts before and after addressing extreme values.
* **Image**: A visual example of how extreme values were identified and fixed.

The cleaned and enhanced dataset provided a reliable foundation for further analysis and model building. Adding tables and charts to explain each step will help make the process clearer and easier to understand.