**Deliverable 2**

**Introduction**: This report presents a detailed analysis of the "Supermarket Sales" dataset using statistical methods, data visualization, hypothesis testing, and regression modelling. The objective is to derive insights regarding customer behaviour, product performance, and transaction characteristics while also predicting key variables using machine learning techniques.

**Summary of the Data:** The dataset includes 1,000 observations and 17 attributes, each representing individual sales transactions. These attributes span customer details (such as gender and customer type), transactional information (including unit price, quantity, tax, total, and gross income), and contextual metadata (branch, city, product category, payment method, date, time, and customer rating). There are no missing entries, making the dataset suitable for immediate analysis.

**Data Sampling**: For the development and testing of prediction models, the data was divided into training and testing subsets, with 75% allocated for training and 25% for testing. This split helps ensure robust model performance and accurate generalization.

**Data Visualization and Exploratory Analysis:** Several graphical methods were used to explore the data:

* Customer Distribution: Nearly equal distribution between "Member" and "Normal" customer types.
* Gender Analysis: Both genders are fairly balanced across customer types.
* Product Line Sales: Box plots revealed that "Health and Beauty" and "Home and Lifestyle" categories yielded higher total sales.
* Correlation Heatmap: Strong positive correlations were found among Total, Tax 5%, and Gross Income, reflecting the dependency between these financial variables.

**Variable Relationships**Scatter plots and correlation matrices highlight the following:

* A strong linear association between Total and Gross Income.
* A moderate positive relationship between Rating and Total, hinting that higher spenders may give better reviews.

These insights were supported through Pearson correlation values and regression analysis.

**Hypothesis Testing**

A two-sample **t-test** was conducted to assess whether there was a significant difference in average customer ratings between **Male** and **Female** customers.

* **Null Hypothesis (H₀)**: There is no significant difference in mean rating between genders.
* **Alternative Hypothesis (H₁)**: There is a significant difference in mean rating between genders.

With a **p-value > 0.05**, we **fail to reject H₀**, indicating no significant difference in average ratings across genders.

**Predictive Analysis: Linear Regression**  
A linear regression model was used to forecast Total sales using features such as Gross Income, Unit Price, Quantity, Tax 5%, and Rating. Results revealed:

* A high R² value, indicating strong model performance and explanatory power.
* A scatter plot comparing actual and predicted values displayed tight clustering around the ideal line, demonstrating prediction accuracy.

**Conclusion:** This analysis provided valuable insights into sales dynamics, customer behaviour, and feature relationships. Gender was not a significant factor in customer ratings. Furthermore, the regression model proved effective in estimating total transaction values, validating the dataset’s potential for predictive analytics. These outcomes can support decision-making in areas such as inventory planning, customer segmentation, and marketing strategies.