# Supermarket Grocery Sales Analysis – Detailed Project Report

## 1. Introduction

The grocery retail industry produces a large volume of transactional data. Analyzing this data can help businesses optimize their operations, increase profitability, and improve customer satisfaction. This project focuses on a dataset from a supermarket and performs in-depth analysis to understand sales patterns, customer behavior, and product performance. Additionally, a regression model is built to predict sales based on various features.

## 2. Objective

The main objectives of this project are:

1. To clean and preprocess the dataset for accurate analysis.
2. To analyze sales trends across different branches, products, and time periods.
3. To visualize customer demographics and payment preferences.
4. To build a machine learning model to predict gross income or total sales.
5. To evaluate model accuracy using suitable metrics.

## 3. Tools and Technologies Used

The following tools were used for this analysis:

**Python**

**Pandas & NumPy** – data manipulation

**Matplotlib & Seaborn** – data visualization

**Scikit-learn** – machine learning model building

**Google Colab** – development environment

## 4. Dataset Overview

The dataset used is named:
Supermarket Grocery Sales - Retail Analytics Dataset.csv

It includes attributes such as:

- Invoice ID
- Branch (A/B/C)
- City
- Customer Type (Member/Normal)
- Gender
- Product Line (e.g., Food and beverages, Health and beauty)
- Unit Price
- Quantity
- Tax
- Total
- Date & Time
- Payment Method
- Gross Income
- Rating

These attributes were used to perform various types of analysis.

---

# 5. Data Preprocessing

### □ Cleaning the Data

Loaded using pandas.read_csv().

Checked using df.head(), df.info(), and df.describe().

Verified and handled **null values** and **data types**.

Converted date and time columns to datetime objects using pd.to_datetime().

Extracted new features: **Hour**, **Day**, and **Month** from the datetime column.

Encoded categorical variables using LabelEncoder for ML model compatibility.

---

# 6. Exploratory Data Analysis (EDA)

### □ A. Sales Performance

**Branch-wise analysis**: Compared sales totals among Branch A, B, and C.

**Product line analysis**: Identified best and worst-performing product categories.

**Monthly sales**: Revealed temporal sales trends.

### □ B. Customer Demographics

Analyzed gender and customer type distributions.

Found whether members or normal customers made higher-value purchases.

### □ C. Payment Method Insights

Evaluated which payment modes were most commonly used (Cash, Credit card, E-wallet).

### □ D. Correlation Analysis

Used a heatmap to find correlation between unit price, quantity, total, and gross income.

---

## 7. Feature Engineering

New columns were created to support analysis and modeling:

**Hour, Day, Month** – extracted from the timestamp.

**Total Purchase** – calculated using Unit price × Quantity.

Categorical columns were encoded for model training.

---

## 8. Machine Learning Model

### □ Objective:

To predict the **Gross Income** or **Total Sale** based on other features using regression.

### □ Model Used:

**Linear Regression** from Scikit-learn

### □ Workflow:

Feature selection: Used columns like Unit Price, Quantity, Tax, etc.

Data split: Used train_test_split() with an 80-20 ratio.

Model fitting: LinearRegression().fit(X_train, y_train)

Prediction and Evaluation:

- $R^2$ Score
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

□ **Model Performance:**

The model gave a good **$R^2$ Score**, showing how well features explained variations in gross income.

MSE and RMSE were used to assess prediction errors.

---

# 9. Key Insights

**Food and Beverages** and **Health and Beauty** were the top-selling product lines.

**Branch B** generated the most revenue.

**Evening hours (15:00–19:00)** had the highest number of purchases.

**Male and Female** customers showed relatively balanced purchasing behavior.

**E-wallets** emerged as a frequently used payment method.

**Linear regression** performed well in predicting gross income based on other numerical features.

---

# 10. Conclusion

The Supermarket Grocery Sales Analysis provided actionable insights into sales trends, customer behavior, and product performance. The integration of machine learning helped predict sales-related metrics, which can help businesses forecast revenue and plan inventory.

---