# Supermart Grocery Sales

**Project Title:** Supermart Grocery Sales - Retail Analytics Dataset
**Domain:** Data Analytics & Data Science
**Tools Used:** Python, Jupyter Notebook, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

## Objective:

The objective of this project is to analyze grocery sales data collected from a fictional supermarket chain in Tamil Nadu, India. The goal is to derive insights from historical sales, perform exploratory data analysis (EDA), and build a machine learning regression model to predict future sales.

## Dataset Description:

The dataset consists of transaction-level information on customer purchases including:

- Order ID
- Customer Name
- Category & Sub-Category
- City, State, Region
- Order Date
- Sales, Discount, Profit

## Data Preprocessing & Feature Engineering

### 1.Data Cleaning:

- Removed missing values and duplicate entries.
- Converted ‘Order Date’ to datetime format.
- Extracted additional features: Order Day, Order Month, Order Year.

### 2. Categorical Encoding:

Applied Label Encoding to categorical variables such as Category, Sub Category, City, Region, and State.

### 3. Feature Creation:

- Extracted month_no, Month name, and year from order date.
- Created relevant features like Total Sales by Category, Sales Trends by Month and Year, Top Cities by Sales.

### 4. Final Feature Set:

Category, Sub Category, City, Region, State, Order Month, Order Year, Discount, Profit

## Exploratory Data Analysis (EDA)

### 1. Sales by Category:

A bar chart showed that ‘Egg, Meat & Fish’ contributed most to total sales, indicating a strong customer preference.

**2. Monthly Sales Trend:**

Line plot revealed increasing sales trends over the months. Sales tend to peak during certain periods, suggesting promotional success or seasonal demand.

**3. Yearly Sales Distribution:**

Pie chart showed that 2017 and 2018 accounted for over 50% of total sales.

**4. Top Cities by Sales:**

Bar chart indicated the top 5 cities contributing to revenue, helping identify high-performing regions.

**5. Correlation Analysis:**

Heatmap revealed strong correlations between Profit and Sales, and moderate correlation between Discount and Sales.

## Model Building & Evaluation

**1. Train-Test Split:**

Used 80/20 split on features and target (Sales).

**2. Feature Scaling:**

Applied StandardScaler to normalize feature values.

**3. Models Used:**

- Linear Regression
- Random Forest Regressor (for improved accuracy)

## 4. Performance Metrics:

1. **Linear Regression:**

Mean Squared Error (MSE): 212,935.59

R-Squared Value: 0.35

2. **Random Forest Regressor:**

Improved R-Squared observed in testing (suggested, exact value may vary)

## 5. Visualization:

Scatter plot for Actual vs Predicted sales from both models.

## Conclusion:

- Linear Regression provided a baseline model but had limited performance ($R^2 = 0.35$).
- Random Forest showed better predictive capability by capturing non-linear relationships.
- Feature importance analysis suggested Profit and Discount were key drivers of Sales.