Project Report

Title: Super Mart Grocery Sales – Retail Analytics Dataset

1. Introduction

The project investigates an imaginary grocery sales dataset for the purpose of practicing **exploratory data analysis (EDA)**, **data visualization**, and **predictive modelling.** The dataset mimics orders received on a grocery delivery platform in Tamil Nadu, India. It gives insights on customer behaviour, distribution of sales, profitability, and time-series trends.

The goal is to analyse the dataset, extract business insights, and train machine learning models to forecast sales or profit.

2. Dataset Description

The data contains **9,994 records** and **11 variables**, with the following:

Order ID – A unique id for every order

Customer Name – Customer's name

Category & Sub-Category – Purchased product type

City & State – Location of the customer

Order Date – Date of purchase

Region – Geographic region (North, South, East, West)

Sales – Total sales value for the order

Discount – Discount offered

Profit – Profit made

Other engineered features: Month, Year, and Month Number.

3. Methodology

Step 1: Data Preprocessing

Dealt with missing values and duplicates

Changed **Order Date** to datetime type

Created day, month, and year for time-series analysis

Label encoded categorical variables

Normalized numerical features

Step 2: Exploratory Data Analysis (EDA)

Category Analysis: Egg, Meat & Fish category had the largest proportion (~15%) of overall sales.

Monthly Trends: Sales were on an upward trend throughout months.

Yearly Analysis: 2017 and 2018 combined accounted for over 50% of total sales.

Regional Sales: Sales in the South region were a key contributor.

Top Cities: Vellore, Krishnagar, and Ooty were among the top city contributors to sales.

Profitability: Discounts had a huge impact on profit, with deep discounts lowering overall profit percentages.

Step 3: Machine Learning Model

Attributes: Category, Sub-Category, City, Region, State, Month, Discount, Profit

Target: Sales

Model Employed: Linear Regression

Model Evaluation Metrics:

Mean Squared Error (MSE): ~1758.26

R^2 Score: ~0.82 (good model fit)

4. Results & Visualizations

Category-wise Distribution of Sales: Identified which categories generate maximum revenue.

Sales Trend Over Time: Uncovered seasonal patterns in demand.

Correlation Heatmap: Demonstrated correlations between sales, discount, and profit.

Actual vs Predicted Sales: There was a good linear fit in the regression model.

Top 5 Cities by Sales: Were crucial geographic insights for making decisions.

5. Conclusion

The **linear regression model** was good with an R^2 of 0.82, signifying valid predictive ability.

Sales growth trend indicates solid business strategies in subsequent years.

Some **categories (Egg, Meat & Fish)** and **cities** are responsible for most of the revenue, indicating the scope for focused marketing opportunities.

Discounts must be managed carefully as they directly cut into profitability.---

6. Recommendations & Next Steps

1. Advanced Modelling: Utilize Random Forest, XGBoost, or Ensemble models for improved accuracy.

2. Feature Engineering: Add customer behaviour features like purchase frequency and basket size.

3. Dashboard Deployment: Develop an interactive BI dashboard (e.g., in Power BI or Tableau) for real-time tracking.

4. Business Strategy:

Prioritize high-performing cities and categories.

Maximize discount strategy to balance profitability and sales growth.

Grow in underperforming areas with tactical campaigns.

7. References

Dataset Source: [Super Mart Grocery Sales – Retail Analytics Dataset] (https://drive.google.com/file/d/1Vx-Ibn11HKofkJasjMZFyigemSu7TOeB/view?usp=sharing)

GitHub Repository: [Supermarket Sales Data Analysis] (https://github.com/sushantag9/Supermarket-Sales-Data-Analysis)

Final Deliverable:

The project illustrates how EDA, visualization, and predictive modelling can yield actionable insights in retail analytics. The conclusions can assist companies to optimize sales, effectively manage discounts, and plan expansion.