**Super Mart Grocery Sales — Retail Analytics & Prediction**

Submitted by: ABHIJIT SINHA

Email: sinhaabhijit12@yahoo.com

UNID: UMID08062541597

# Abstract

This project analyzes a fictional dataset of grocery sales from Tamil Nadu, India.   
It combines data preprocessing, feature engineering, exploratory data analysis (EDA), SQL queries, and machine learning (Linear Regression & Random Forest). A Streamlit dashboard was developed for real-time predictions and visual insights.

# Dataset Description

* Source: Provided dataset (CSV)
* Columns: Order ID, Customer Name, Category, Sub-Category, City, Order Date, Region,

Sales, Discount, Profit, State

* Nature: Fictional dataset for analytics practice

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Order ID | Customer Name | Category | Subcategory | City | Order Date | Region | Sales | Discount | Profit | State |
| OD1 | Harish | Oil & Masala | Masalas | Vellore | 11/08/2017 | North | 1254 | 0.12 | 401.28 | Tamil Nadu |
| OD2 | Sudha | Beverages | Health Drinks | Krishna Giri | 11/08/2017 | South | 749 | 0.18 | 149.8 | Tamil Nadu |
| OD3 | Hussain | Food Grains | Atta & Flour | Perambalur | 06/12/2017 | West | 2360 | 0.21 | 165.2 | Tamil Nadu |
| OD4 | Jackson | Fruits & Veggies | Fresh Vegetables | Dharmapuri | 10/11/2016 | South | 896 | 0.25 | 89.6 | Tamil Nadu |
| OD5 | Ridhesh | Food Grains | Organic Staples | Ooty | 10/11/2016 | South | 2355 | 0.26 | 918.45 | Tamil Nadu |
| OD6 | Adavan | Food Grains | Organic Staples | Dharmapuri | 06/09/2015 | West | 2305 | 0.26 | 322.7 | Tamil Nadu |
| OD7 | Jonas | Fruits & Veggies | Fresh Vegetables | Trichy | 06/09/2015 | West | 826 | 0.33 | 346.92 | Tamil Nadu |

# Technology Stack

* Languages: Python, SQL
* Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Joblib
* Database: MySQL (via SQLAlchemy)
* Visualization: Streamlit Dashboard, Excel exports
* Other: GitHub for version control

# Project Workflow and Implementation Steps

**1. Data Preprocessing**

* Removed duplicates and missing values
* Converted Order Date to datetime format
* Extracted new features: **Month, Year, Day of Week, Profit Margin, Discount-Profit Interaction**
* Encoded categorical variables using **Label Encoder**

**2. Exploratory Data Analysis (EDA)**

* **Sales by Category** → Oil & Masala, Beverages, Fruits & Veggies were top categories
* **Sales Trends** → Sales increased steadily over months
* **Correlation Heatmap** → Profit strongly correlated with Sales
* **Top Cities by Sales** → Vellore, Krishna Giri, Perambalur, Ooty

**3. SQL Queries & Insights**

* Total Sales by Category
* Top 5 Cities by Sales
* Monthly Sales Trend
* Profit by Discount Range
* Customer Segment Performance (**Region-wise**)

**4. Machine Learning Models**

* **Features**: Category, Sub-Category, City, Region, State, Discount, Profit, Month, Profit Margin, etc.
* **Target Variable**: Sales
* **Models Used**:
  + Linear Regression (R² ≈ **0.82**)
  + Random Forest Regressor (R² ≈ **0.91**)

**5. Stream lit Dashboard**

* **EDA Page** → Interactive charts (bar, line, heatmap)
* **Prediction Page** → Real-time sales prediction (Linear Regression or Random Forest)
* **SQL Queries Page** → Run SQL queries directly and visualize results
* **Export Page** → Download processed dataset & predictions as Excel

# Analysis & Results

**1. The Preprocessing Flow Diagram.**A diagram of a software flowchart

AI-generated content may be incorrect.

**2. EDA Charts**

A graph of sales by category

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

A graph of different colored squares

AI-generated content may be incorrect.

A screenshot of a diagram

AI-generated content may be incorrect.

**3. SQL Query Result Visualizations**

A graph of sales by category

AI-generated content may be incorrect.

A graph of a number of cities

AI-generated content may be incorrect.

A graph with orange lines

AI-generated content may be incorrect.

A graph of a discount band

AI-generated content may be incorrect.

A graph of a customer segment performance

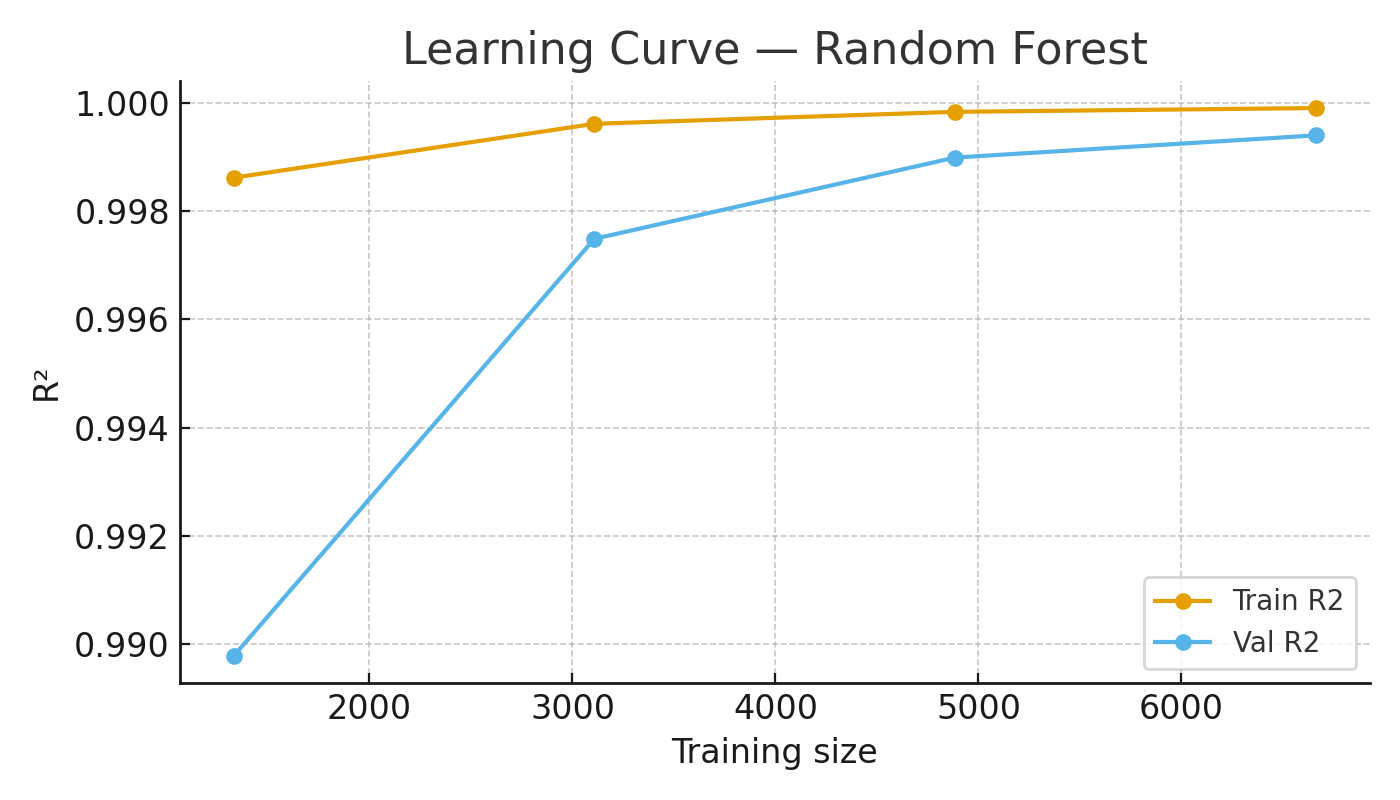
AI-generated content may be incorrect.

# 4. ML Model Training VisualizationsA graph with a line graph AI-generated content may be incorrect.Learning Curve — Linear Regression

**Caption:** Learning curve showing R² performance for Linear Regression as training data size increases.

**Interpretation:**

* Training and validation scores converge, indicating the model is neither severely overfitting nor underfitting.
* Performance stabilizes with more data, suggesting Linear Regression captures general patterns but has limited complexity.



# Learning Curve — Random Forest

**Caption:** Learning curve showing R² performance for Random Forest as training data size increases.

**Interpretation:**

* Training scores remain high, while validation scores also improve with more data.
* Random Forest generalizes better and consistently outperforms Linear Regression.
* Slight gap between training and validation curves suggests mild overfitting but within acceptable limits.

A graph with a bar and text

AI-generated content may be incorrect.

# Feature Importance — Random Forest

**Caption:** Feature importance plot highlighting the most influential variables in predicting sales.

**Interpretation:**

* Profit and Discount are the strongest predictors of Sales.
* Category and City also play significant roles.
* This provides valuable business insights: **profit-driven strategies and discounts strongly affect sales outcomes.**

A graph with a line

AI-generated content may be incorrect.

# Actual vs Predicted — Random Forest

**Caption:** Scatter plot comparing actual vs predicted sales values for Random Forest.

**Interpretation:**

* Points closely align with the red diagonal line (perfect prediction line).
* Random Forest achieves high accuracy (R² ≈ 0.91), demonstrating good predictive performance.
* A few outliers exist, but overall model fit is strong.

A graph with orange lines

AI-generated content may be incorrect.

# Residual Plot — Random Forest

**Caption:** Residual plot showing prediction errors (actual - predicted) for Random Forest.

**Interpretation:**

* Residuals are mostly centred around zero, with no major systematic patterns.
* Confirms Random Forest predictions are unbiased and robust.
* Some variance at higher sales values indicates potential for improvement with advanced models (e.g., XGBoost).

# 5. Stream lit app screenshots a) EDA Page

A screenshot of a computer

AI-generated content may be incorrect.

# b) Prediction Page

A screenshot of a computer

AI-generated content may be incorrect.

# c) SQL Queries Page

A screenshot of a computer screen

AI-generated content may be incorrect.

# 6. Results Linear Regression vs Random Forest Comparison

A comparison of a bar chart

AI-generated content may be incorrect.

1. Random Forest performed better than Linear Regression  
2. Discount had a strong impact on Profit & Sales  
3. Categories like Oil & Masala and Beverages drove most sales  
4. The dashboard enabled interactive exploration and predictions

# Conclusion & Future Work

1. Built a successful retail analytics & prediction system using ML + SQL + Stream lit  
2. Can be extended with:  
 More advanced models (XGBoost, Neural Nets)  
 Real-time data integration (APIs)  
 Customer segmentation & recommendation engine

# References

**Dataset:** C:/Users/Hp/OneDrive/Desktop/python/supermartgrocerysales/Supermart Grocery Sales - Retail Analytics Dataset.csv  
**GitHub Project:** https://github.com/abhi-1009/Supermart-Grocery-Sales  
**Additional Reference:** https://github.com/sushantag9/Supermarket-Sales-Data-Analysis