Sales Insight Prediction:

Forecasting Future Sales Using Data Mining & Data Warehousing

Course: ICT 333 1.5 – Data Mining and Data Warehousing

Name: K. A. N. N. Kodikara

Index No: AS2021907

Date: May 09, 2025

1. Introduction

Sales prediction is a critical aspect of business strategy for retail and e-commerce organizations. It helps optimize inventory, allocate resources effectively, and improve decision-making.

This project leverages data mining techniques to analyze historical sales data from the Superstore dataset, aiming to uncover insights into sales trends, customer behavior, and product performance. The project also builds predictive models to forecast future sales, enabling actionable recommendations for business growth.

2. Dataset Description

The dataset used is the **Sample—Superstore dataset** from Kaggle, which contains retail transaction records from an e-commerce store.

- **Rows:** 9,994
- **Columns:** 21
- **Key Features:**
  - Order Date, Ship Date
  - Category, Sub-Category
  - Sales, Profit, Quantity, Discount
  - Customer ID, Region, State

The dataset was loaded into Python and cleaned by converting date columns to DateTime format, removing irrelevant columns, and creating additional features such as shipping duration and profit margin.

3. Methodology

Data Preprocessing

- Converted Order Date and Ship Date to datetime.
- Created Ship Duration (days between order and ship date).
- Calculated Profit Margin = Profit / Sales.
- Handled missing values (none found in critical fields).
- Exported cleaned data to CSV for dashboard use.

Exploratory Data Analysis

- Analyzed **sales trends over time** → monthly/quarterly sales patterns.
- Identified **top-selling categories and sub-categories**.
- Plotted **sales by region** to find high-performing areas.

- Visualized **discount vs. profit** → found negative correlation (higher discounts reduce profits).

Customer Segmentation

Used **K-Means clustering** on RFM (Recency, Frequency, Monetary) features:

- Recency: Days since last purchase
- Frequency: Number of orders
- Monetary: Total sales per customer
  → Identified 3 distinct customer segments (VIP, mid-value, low-value).

Predictive Modeling

Built models to predict Sales:

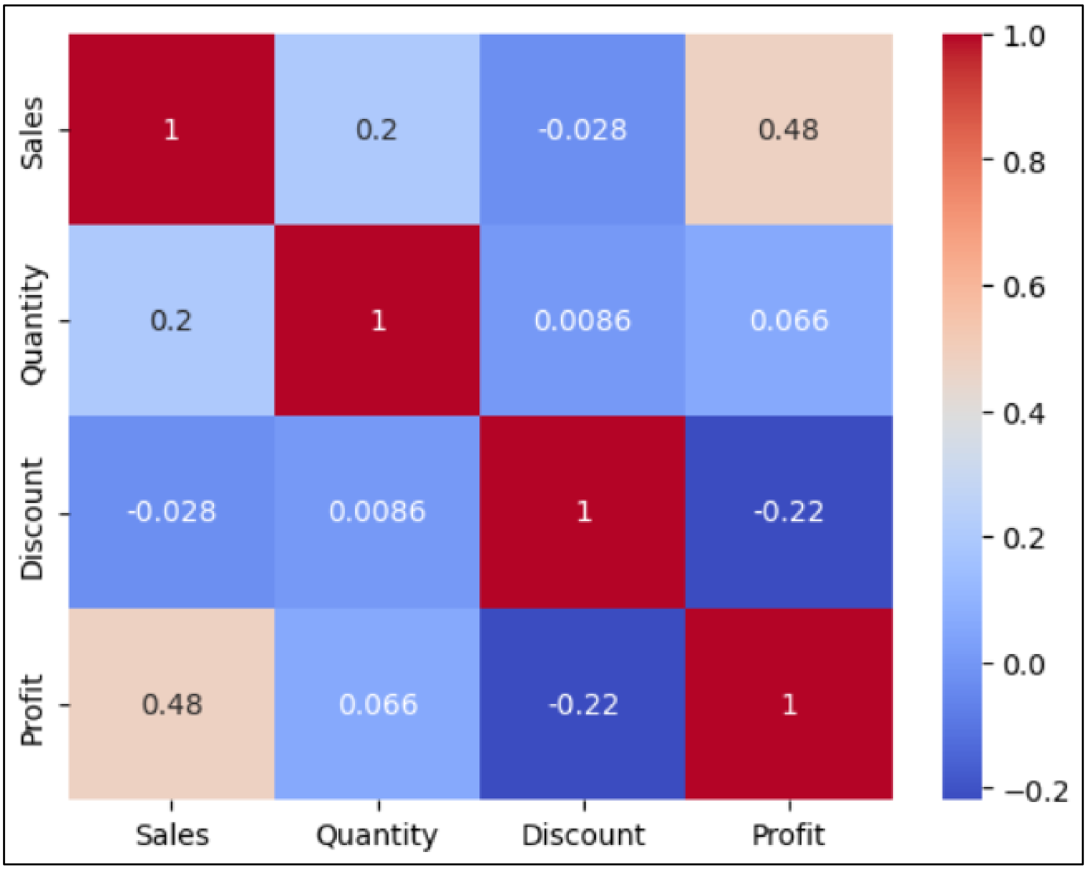
- Linear Regression
- Random Forest Regression

Evaluated models using **MAE, MSE, R^2** metrics. Random Forest achieved higher accuracy.

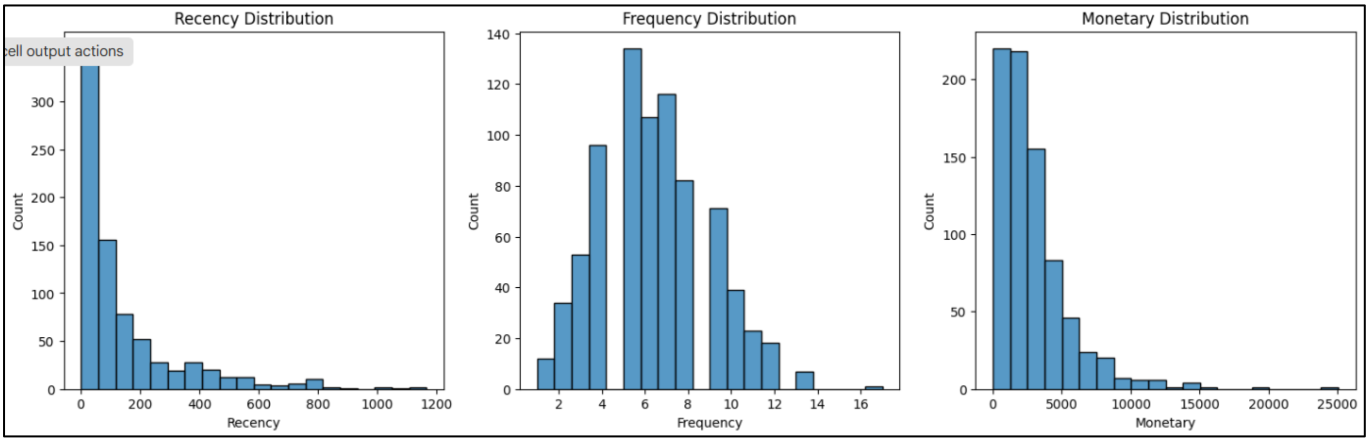
Association Rule Mining

Attempted Apriori algorithm for product associations.
→ No significant frequent itemsets found due to sparse co-purchases.

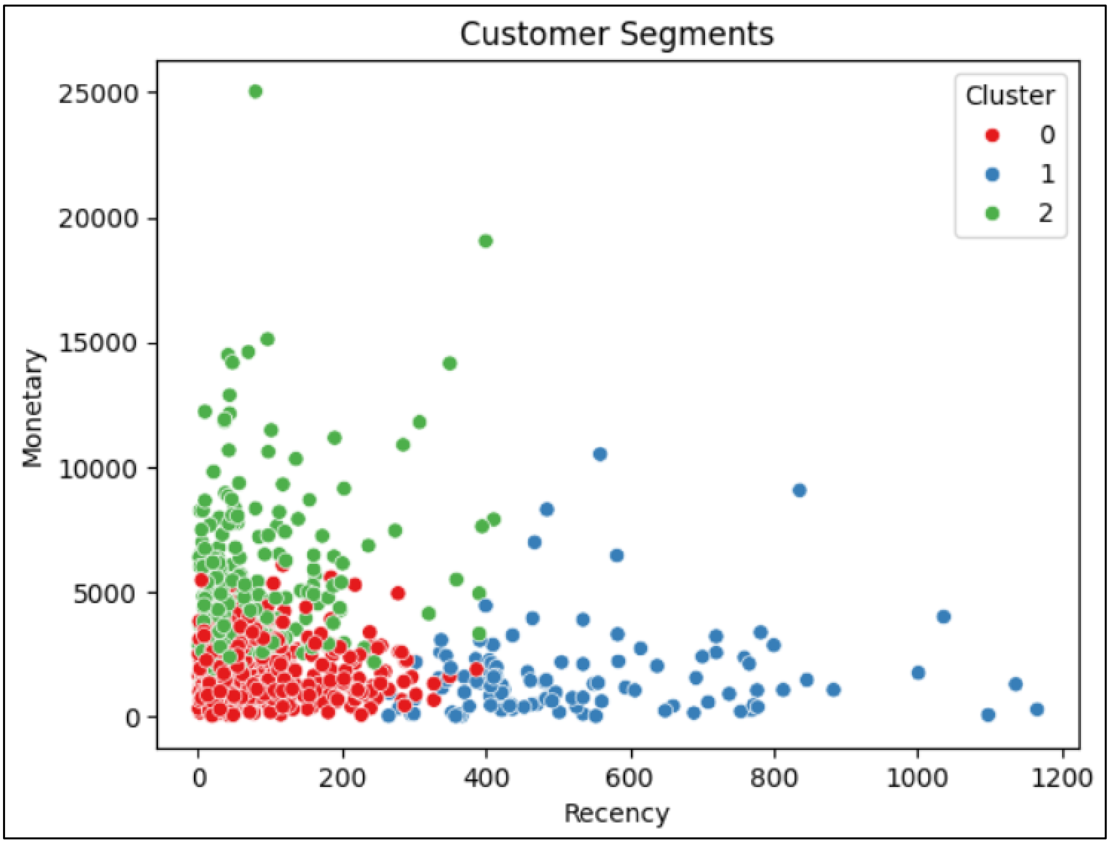
Heatmap of correlations



RFM Features Distribution



Scatterplot



Linear Regression Model

Linear Regression Model

```
[35] from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

     lr_model = LinearRegression()
     lr_model.fit(X_train, y_train)

     y_pred_lr = lr_model.predict(X_test)

     # Evaluation
     print("Linear Regression MAE:", mean_absolute_error(y_test, y_pred_lr))
     print("Linear Regression MSE:", mean_squared_error(y_test, y_pred_lr))
     print("Linear Regression R^2:", r2_score(y_test, y_pred_lr))
```

```
Linear Regression MAE: 218.16059538696962
Linear Regression MSE: 186793.7952128799
Linear Regression R^2: 0.3730058915278802
```

Random Forest Model

Random Forest Model

```
from sklearn.ensemble import RandomForestRegressor

rf_model = RandomForestRegressor(n_estimators=100, random_state=100)
rf_model.fit(X_train, y_train)

y_pred_rf = rf_model.predict(X_test)

# Evaluation
print("Random Forest MAE:", mean_absolute_error(y_test, y_pred_rf))
print("Random Forest MSE:", mean_squared_error(y_test, y_pred_rf))
print("Random Forest R^2:", r2_score(y_test, y_pred_rf))
```

```
Random Forest MAE: 92.53681724193049
Random Forest MSE: 68586.3627758969
Random Forest R^2: 0.7697822599888282
```

4. Results & Insights

- **Sales Trends:**
  - Peak sales during November–December → holiday season.
- **Best-Selling Categories:**
  - Highest sales in **Office Supplies** and **Furniture** categories.
  - Top sub-categories: **Chairs**, **Phones**.
- **Best Regions:**
  - **West** and **East** regions generated most sales.
- **Discount Impact:**
  - Negative correlation between **Discount** and **Profit** → excessive discounts reduce profitability.
- **Customer Segments:**
  - Cluster 0: Recent, high-spending → VIP customers.
  - Cluster 1: Older purchases, medium spending → needs reactivation.
  - Cluster 2: Infrequent, low spending → low-value customers.

5. Dashboard Overview

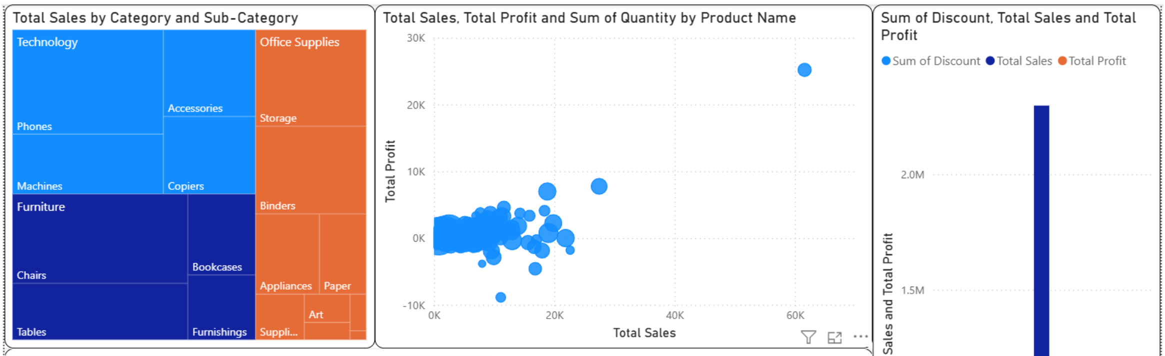
A dashboard was created using Tableau [or Power BI] to visualize:

- Sales trend over time
- Sales by category and region
- Discount vs profit relationship
- Customer segment distribution

The dashboard allows interactive filtering by region, category, and date.

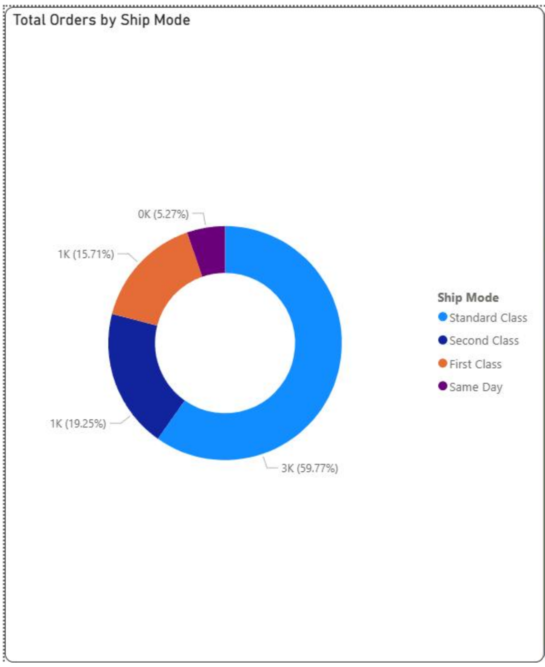
Overview Page

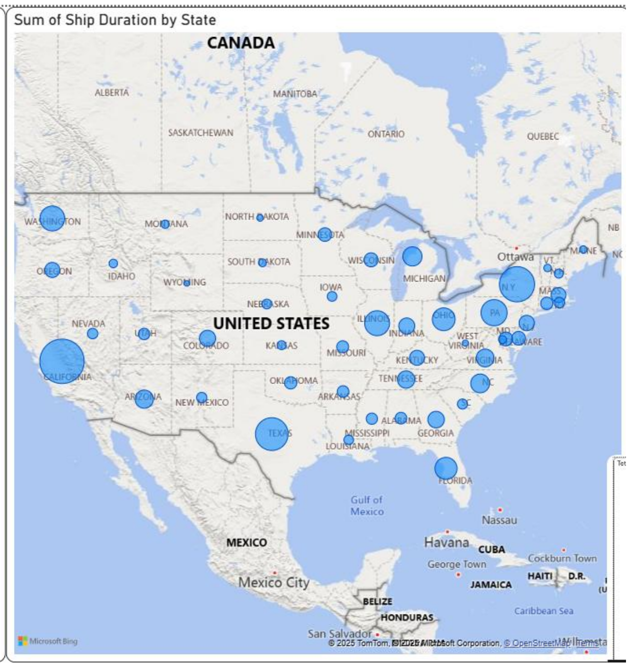
Product Performance



Product Name	Total Sales	Total Profit	Sum of Quantity
"While you Were Out" Message Book, One Form per Page	25.23	10.39	8
#10- 4 1/8" x 9 1/2" Recycled Envelopes	286.67	115.28	37
#10- 4 1/8" x 9 1/2" Security-Tint Envelopes	146.69	64.86	21
#10 Gummed Flap White Envelopes, 100/Box	41.30	16.77	11
#10 Self-Seal White Envelopes	108.68	52.12	10
#10 White Business Envelopes,4 1/8 x 9 1/2	488.90	223.14	32
#10-4 1/8" x 9 1/2" Premium Diagonal Seam Envelopes	176.29	63.75	13
#6 3/4 Gummed Flap White Envelopes	71.28	24.95	9
1.7 Cubic Foot Compact "Cube" Office Refrigerators	2,706.08	578.68	14
1/4 Fold Party Design Invitations & White Envelopes, 24 8-1/2" X 11" Cards, 25 Env./Pack	49.98	22.71	7
12 Colored Short Pencils	18.20	3.02	8
12-1/2 Diameter Round Wall Clock	551.45	91.11	36
14-7/8 x 11 Blue Bar Computer Printout Paper	422.75	172.94	10
Total	**2,297,200.86**	**286,397.02**	**37873**

Shipping Performance





6. Conclusion & Recommendations

This project successfully applied data mining techniques to analyze sales data, segment customers, and build predictive models to forecast sales. The analysis identified key trends and insights to support business decision-making.

Key findings include:

- **Sales trends** peaked during November and December, aligning with the holiday seasons.
- **Office Supplies** and **Furniture** were the best-selling categories, with **Chairs** and **Phones** among the top products.
- The **West** and **East** regions generated the highest sales.
- There is a **negative correlation between Discount and Profit**, indicating that excessive discounts lower profitability.

In predictive modeling:

- **Linear Regression** achieved an R^2 of **0.373**, MAE of **218.16**, and MSE of **186,793.79**, indicating lower predictive accuracy.

- **Random Forest Regression** performed better with an R^2 of **0.770**, MAE of **92.54**, and MSE of **68,586.36**, providing a more reliable model for sales prediction.

Customer segmentation via **K-Means clustering** identified three distinct segments:

1. **High-value customers** – recent, frequent, high spending
2. **Mid-value customers** – moderate recency and spending
3. **Low-value customers** – infrequent, low spending

---

7. References

- Superstore Dataset, Kaggle
- sci-kit-learn, pandas, seaborn libraries, numpy
- Power BI Public documentation

- END -