# Contents

# 1. Project Overview

## 1.1 Overview

This project aims to analyze Superstore sales data to uncover insights related to customer behavior, product performance, and business trends. The goal is to generate actionable insights using data visualization, customer segmentation, and a personalized product recommendation system.

## 1.2 Scope

The project includes:

- Sales data analysis over four years
- Customer segmentation via K-Means clustering
- A product recommendation engine based on purchasing patterns
- Interactive dashboard and GUI for visualization and recommendations

## 1.3 Key Deliverables

- Exploratory data analysis (EDA) of the Superstore Dataset
- Power BI dashboard for sales and customer insights
- K-Means clustering model for customer segmentation
- Cluster-based recommendation system
- GUI-based recommendation interface
- Complete project documentation

## 1.4 Stakeholders

- Business managers and decision-makers
- Marketing and sales teams
- Customer support teams
- Data science and analytics teams
- End-users interacting with the recommendation system

# 2. Dataset Overview

## 2.1 Data Description

The Superstore dataset contains over 10,000 records across 21 columns. It includes transactional data on orders, customers, shipping, and product sales across four years. This dataset is ideal for time-series, customer behavior, and sales pattern analysis, and the Kaggle link for the data : Superstore Dataset - Kaggle .

## 2.2 Data Dictionary

| | | |
|---|---|---|
| Row ID | Unique identifier for each record | Integer |
| Order ID | Unique identifier for each customer order | String |
| Order Date | Date when the order was placed | DateTime |
| Ship Date | Date when the order was shipped | DateTime |
| Ship Mode | Shipping method selected by customer | String |
| Customer ID | Unique identifier for each customer | String |
| Customer Name | Name of the customer | String |
| Segment | Business segment the customer belongs to | String |
| Country | Country of the customer | String |
| City | City of the customer | String |
| State | State or province of the customer | String |
| Postal Code | Postal code of the customer's address | String |
| Region | Geographical region classification | String |
| Product ID | Unique identifier for each product | String |
| Category | Main product category | String |
| Sub-Category | Specific product sub-category | String |
| Product Name | Name of the product | String |
| Sales | Revenue generated from the sale | Float |
| Quantity | Quantity of the product | Integer |
| Discount | Discount provided | Float |
| Profit | Profit/lose generated from the sale | Float |

## 2.3 Data Quality Assessment

- Missing values identified in 'State' column (11 entries)
- One duplicate record removed
- Date formats standardized
- Text fields normalized to lowercase

# 3. Methodology

## 3.1 Data Preprocessing

- Removed whitespace, converted text to lowercase
- Transformed date columns(‘Order Date’ and ’Ship Date’) to datetime format
- Imputed missing 'State' with appropriate values
- Removed duplicates and irrelevant columns (e.g., Row ID)

## 3.2 Exploratory Data Analysis

Comprehensive visualizations were created to understand patterns in the data like:

- Sales distribution and trend analysis
- Top-selling products and profit contribution
- Regional and category-wise performance
- Customer segment-product heatmaps

## 3.3 Dashboard Development

- Created interactive Power BI dashboards for sales, profits, and customer insights

## 3.4 Customer Segmentation

K-means clustering was employed to segment customers based on their purchasing behavior :

- Calculated median sales per customer
- Normalized sales data using MinMaxScaler
- Used Elbow Method to determine optimal clusters (k=4)
- Applied K-Means (n_clusters=4, random_state=42)
- Achieved silhouette score of 0.69

## 3.5 Product Recommendation System

A recommendation engine was developed based on the clustering results:

- Identified top-selling products per cluster
- Recommended top 10 products per customer cluster
- Used collaborative filtering for personalized suggestions

# 4. Implementation

## 4.1 Technology Stack

The project was implemented using the following technologies:

- **Programming:** Python 3.x
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Visualization:** Power BI

## 4.2 Machine Learning Pipeline

The machine learning workflow was structured as follows:

1. Data ingestion and preprocessing : Reading, Cleaning and transforming the raw data
2. Feature engineering (customer-level metrics)
3. K-Means clustering model training
4. Cluster evaluation using silhouette score
5. Product recommendation generation : Creating product recommendations based on cluster-specific popularity

# 5. Model Evaluation

## 5.1 Clustering Performance

The K-means clustering model was evaluated using multiple metrics:

- **Elbow Method:** Determined that 4 clusters provide the optimal balance between complexity and explanatory power
- **Silhouette Score:** Achieved a score of 0.69, indicating well-formed and distinct clusters
- **Silhouette Plot:** Visual validation confirmed the quality of the clustering with minimal overlap
- **Cluster Distribution Analysis:** Examined the size and characteristics of each cluster

## 5.2 Recommendation Accuracy

The recommendation system's effectiveness was assessed through:

- Analysis of recommendation relevance within each cluster
- Examination of product diversity in recommendations
- Verification that recommended products align with cluster purchasing patterns

# 6. **Results and Insights**

## 6.1 Customer Segments Analysis

The clustering analysis revealed four distinct customer segments:

1. **Economy Buyers (Cluster 0):**
    - 638 customers
    - median sales ( $47)
    - Focus on essential, lower-priced items
2. **Premium Shopper (Cluster 1):**
    - 1 customer
    - median sales ( $1,920)
    - Purchases high-value items in significant quantities
    - Potential VIP customer requiring special attention
3. **Mid-Market Consumers (Cluster 2):**
    - 150 customers
    - median sales ( $147)
    - Balanced purchasing behavior across various product categories
    - Solid revenue contributors with potential for up-selling
4. **High-Value Customers (Cluster 3):**
    - 4 customers
    - median sales ( $720)
    - Prefer premium products but purchase less frequently than Cluster 1
    - Important segment for targeted premium offerings

## 6.2 Product Recommendations

The recommendation system generates tailored product suggestions based on cluster membership:

- **Economy Buyers:** Popular, affordable products
- **Premium Shoppers:** High-end, complementary items
- **Mid-Market:** Value-focused and diverse offerings
- **High-Value:** Premium and exclusive recommendations

# 7. **Conclusion**

The Superstore project successfully demonstrates how sales data can be transformed into actionable insights using machine learning and data visualization. The system segments customers into meaningful clusters and delivers personalized product recommendations. With an intuitive dashboard and interactive interface, business users can easily access and apply these insights to drive growth, improve customer experience, and enhance operational efficiency.