P.E.S. COLLEGE OF ENGINEERING
MANDYA, 571401
(An Autonomous Institution under VTU, Belgaum)

A Report On

"SUPERSTORE SALES ANALYSYS USING R"

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of
Dr. Deepika Bidri
Asst Professor, Dept of CS&E
P.E.S.C.E, Mandya

Submitted by

MAHADEVASWAMY M R[USN:4PS22CS198]

LIKITH D[USN:4PS22CS199]

HEMANTH M U[USN:4PS22CS201]

PRAVEEN KUMAR R[USN:4PS23CS412]

YASHWANTH GOWDA N R[USN:4PS23CS415]

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
P.E.S. COLLEGE OF ENGINEERING, MANDYA-571401
2024-2025

SL NO	STUDENT NAME AND USN	MARKS OBTAINED
1	MAHADEVASWAMY M R[USN:4PS22CS198]	
2	LIKITH D[USN:4PS22CS199]	
3	HEMANTH M U[USN:4PS22CS201]	
4	PRAVEEN KUMAR R[USN:4PS23CS412]	
5	YASHWANTH GOWDA N R[USN:4PS23CS415]	

List of Figures

CHAPTER 1

1. Introduction:

1.1 Overview of the business problem:

In the highly competitive retail sector, businesses generate large volumes of transactional data from customer purchases, product sales, discounts, and regional performance. However, this raw data often goes underutilized, leaving valuable insights hidden. Retailers struggle to identify which product categories yield the most profit, which regions underperform, or how pricing decisions—like applying discounts—affect profitability. Without actionable insights, making strategic business decisions becomes guesswork rather than a data-driven process.

1.2 Importance of data analysis in retail:

Data analytics plays a vital role in transforming raw business data into meaningful information. Through techniques like Exploratory Data Analysis (EDA), predictive modeling, and clustering, businesses can:

- Track key performance indicators (KPIs) such as total sales, total profit, and profit margins
- Understand the effectiveness of discount strategies
- Identify which product categories or regions are high or low performers
- Group similar customer behaviors or sales patterns for better targeting
- In short, analytics enables retail companies to optimize decision-making, improve customer satisfaction, and increase overall profitability.

1.3 Dataset: Sample Superstore (from Tableau or Kaggle):

This project uses the "Sample - Superstore" dataset, which is widely known in the data analytics and visualization community. It is originally provided with Tableau software and is also available on Kaggle. The dataset simulates real-world retail transactions, including details such as:

- Sales, Profit, and Discount values
- Regional and product-level information (Region, Category, Sub-Category)
- Transaction and shipping details

It provides a comprehensive view of a retail operation, making it ideal for performing both business analysis and applying machine learning techniques like regression and clustering.

2. Objectives:

The primary objective of this project is to apply data analytics techniques using R programming to analyze and extract insights from the Superstore Sales dataset. The project focuses on solving real-world retail business problems by answering key questions such as:

2.1 Sales and Profit Analysis

- Identify how sales and profit vary across different regions, categories, and sub-categories.
- Determine which regions and product types contribute the most to profit.
- Visualize and summarize key performance indicators.

2.2 Discount Impact

- Analyze the relationship between discounts and profitability.
- Understand whether offering discounts helps or hurts the business.
- Use regression techniques to quantify the impact.

2.3 Predictive Modeling with Linear Regression

- Build a linear regression model to predict Profit based on Sales and Discount.
- Use the model to understand which variables significantly influence profitability.
- Interpret regression coefficients and model fit using statistical measures.

2.4 Customer Segmentation using K-Means Clustering

- Apply K-means clustering to group similar transactions based on numerical features like Sales, Profit, and Discount.
- Identify patterns among clusters to assist in customer or transaction segmentation.
- Visualize clusters using dimensionality reduction.

2.5 Data Visualization

- Use ggplot2 and factoextra libraries in R to build insightful charts and graphs.
- Communicate trends, outliers, and groupings clearly to stakeholders.

3. Tools and Technologies

To effectively analyze and visualize the Superstore dataset, this project utilizes the R programming language and several powerful open-source libraries. These tools support the entire data analytics workflow—from data cleaning to advanced statistical modeling and visualization.

3.1 Programming Language

R: A language designed specifically for statistical computing and data visualization. R is well-suited for exploratory data analysis, regression modeling, and clustering tasks.

3.2 Libraries and Packages

- Tidyverse: A collection of R packages that provides tools for data manipulation, cleaning, and visualization:
- dplyr: For data wrangling and transformation
- readr: For reading and writing datasets (CSV, TXT, etc.)

- ggplot2: For creating professional-quality plots and visualizations tibble, purrr, stringr, etc., for clean and efficient data handling
- scales: Used for formatting numerical values in plots (e.g., currency, percentages).
- factoextra: A visualization package designed to make it easier to interpret and visualize the output of multivariate data analyses like clustering. Used in this project to visualize K-means clustering results.

3.3 Other Tools:

RStudio (IDE): Used for writing and executing R scripts efficiently.

CSV File: The input dataset, "Sample - Superstore.csv," was analyzed directly in R using read_csv() from readr.

These tools provide a complete, flexible, and scalable environment for conducting professional data analysis and generating actionable insights in a retail business context.

CHAPTER 2

4. Dataset Description

The dataset used in this project is titled “Sample - Superstore” and simulates real-world retail operations of a US-based store. It contains sales transaction data that includes customer details, order details, shipping information, product information, and financial metrics like Sales, Profit, and Discount.

Key Attributes in the Dataset:

Column Name	Description
Order ID	Unique identifier for each order
Order Date	Date when the order was placed
Ship Date	Date when the order was shipped
Category	Product category (e.g., Furniture, Office Supplies)
Sub-Category	More specific product groupings
Region	Geographic region (West, East, South, Central)
Sales	Total sales amount for the item
Profit	Profit earned from the item
Discount	Discount percentage applied

This dataset contains both categorical and numerical features, making it ideal for statistical analysis, visualization, and machine learning tasks such as regression and clustering.

5. Data Preprocessing:

To ensure data consistency and accuracy, the dataset underwent several preprocessing steps before analysis:

5.1 Cleaning and Formatting:

- Renamed columns to remove spaces and special characters using make.names() to ensure compatibility in R.
- Converted categorical variables like Category, Sub-Category, and Region into factor types.

5.2 Handling Missing and Extreme Values:

- Used drop_na() to remove any rows with missing numeric data required for clustering.
- Removed extreme outliers using percentile filtering (e.g., top 1% of Sales and bottom 1% of Profit) before clustering.

5.3 Feature Selection and Scaling:

- Selected relevant numerical features: Sales, Profit, and Discount for clustering.
- Standardized these features using scale() to ensure fair treatment in K-means clustering.

6. Exploratory Data Analysis (EDA):

EDA was performed using ggplot2 to visualize sales and profit trends across various business dimensions.

6.1 Sales by Region

- A bar chart was created to compare total sales across the four regions: West, East, Central, and South.
- Insight: The West region had the highest sales.

6.2 Profit by Category

- Displayed profit contribution from each product category: Furniture, Office Supplies, and Technology.
- Insight: Technology yielded the highest profit margin.

6.3 Profit by Sub-Category

- A horizontal bar chart was used to analyze sub-category-level profitability.
- Insight: Some sub-categories (like Chairs and Tables) showed negative profit despite high sales.

6.4 Sales vs. Profit Scatter Plot

- A scatter plot was created to observe the relationship between Sales and Profit, color-coded by Category.
- Insight: High sales do not always lead to high profit—especially when discounts are involved.

These visualizations helped uncover key performance patterns and areas needing strategic improvement in sales and pricing.

7. Linear Regression Analysis

Linear regression was used to understand how Sales and Discount impact Profit.

7.1 Model Definition:

Profit ~ Sales + Discount

This model estimates how much profit is affected by sales volume and the amount of discount applied.

7.2 Key Findings from Model Summary:

- Sales has a positive relationship with Profit, but the impact is relatively small.
- Discount has a strong negative impact on Profit, meaning higher discounts significantly reduce profitability.
- The intercept represents the baseline profit when both sales and discount are zero.
- R-squared value shows how well the model fits the data (e.g., 0.29 means the model explains 29% of the variance in profit).

7.3 Visualizations:

- Profit vs Sales: Shows a slightly increasing trend.
- Profit vs Discount: Shows a clear decreasing trend with higher discount values.
- These regression insights confirm that aggressive discounting strategies may harm overall profitability and should be applied more strategically.

8. K-Means Clustering

K-Means clustering was used to group similar sales transactions based on three numerical features:

Sales ,Profit and Discount

8.1 Steps Taken:

- Preprocessed the data by removing outliers and scaling features.
- Applied K-Means clustering with 3 clusters using kmeans() in R.
- Visualized the clusters using factoextra::fviz_cluster().

8.2 Interpretation of Clusters:

- Cluster 1: High sales with medium profit, low discount.
- Cluster 2: Low sales and profit, possibly high discount—could be loss-making transactions.
- Cluster 3: Medium sales and profit, balanced discount strategy.

Clustering provided useful segmentation of transaction types, which can help tailor pricing and promotion strategies.

9. Conclusion

This project successfully analyzed the Sample Superstore dataset using R, combining exploratory data analysis, predictive modeling, and unsupervised learning.

Key Takeaways:

- The Technology category and the West region were the top contributors to sales and profit.
- Certain Sub-Categories (like Tables and Bookcases) incurred losses, possibly due to poor discount strategies.
- Linear regression revealed that discounts significantly reduce profit, suggesting the need for discount control.
- K-means clustering helped group sales transactions and uncover hidden patterns that traditional charts cannot easily identify.
- These insights can guide better pricing strategies, category management, and regional planning in a real retail environment.

10. Future Scope

The current project provides strong foundational insights, but it can be expanded further in multiple directions:

Potential Enhancements:

Advanced Models: Use decision trees or XGBoost to predict profit with more variables.

Customer Segmentation: Include customer ID and demographics for targeted marketing analysis.

Time Series Analysis: Analyze trends over time for forecasting demand.

Dashboard Integration: Deploy this as an interactive Shiny web app for business users.

Profitability Optimization: Run simulations to test the effect of adjusting discount levels on profit.

These next steps can further enhance decision-making capabilities and drive retail performance.

11. Results:

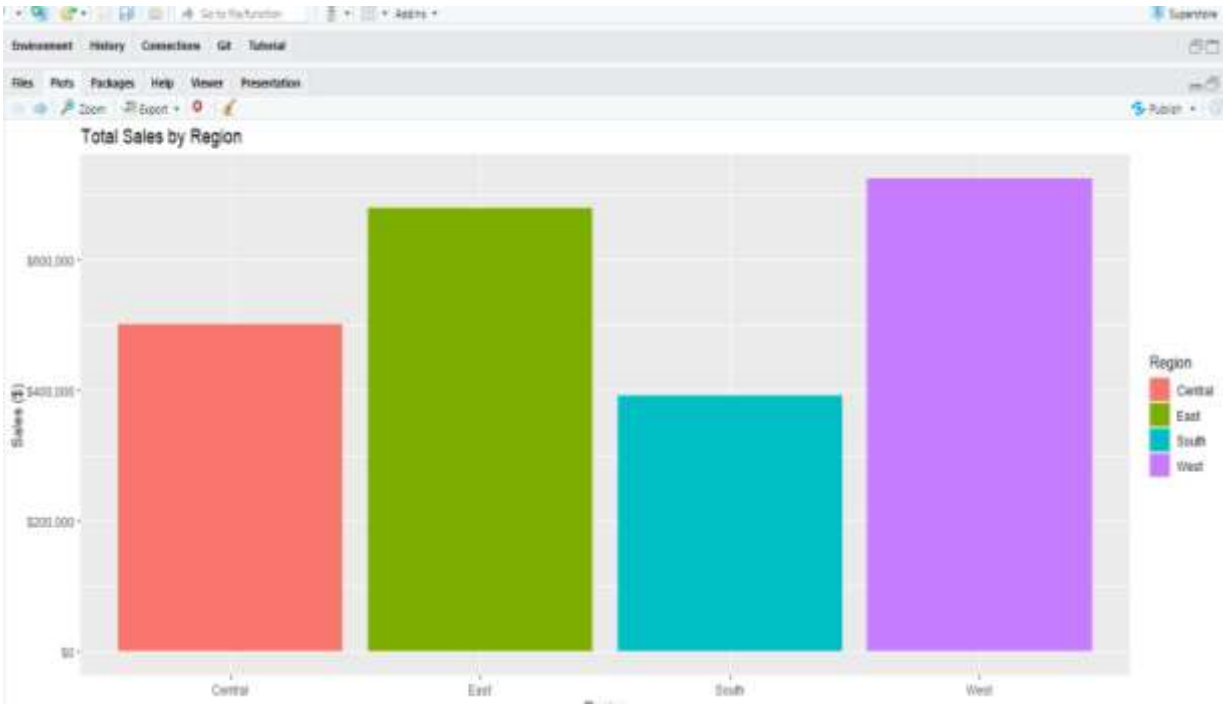

Figure 1: Total Sales by Region

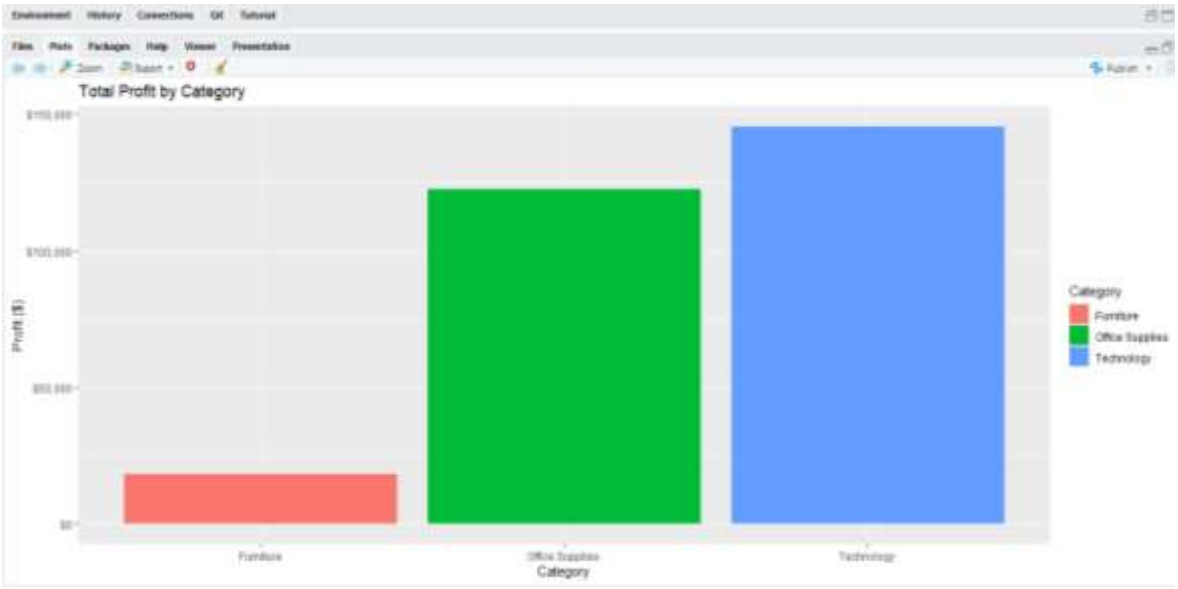

Figure 2: Profit by Product Category

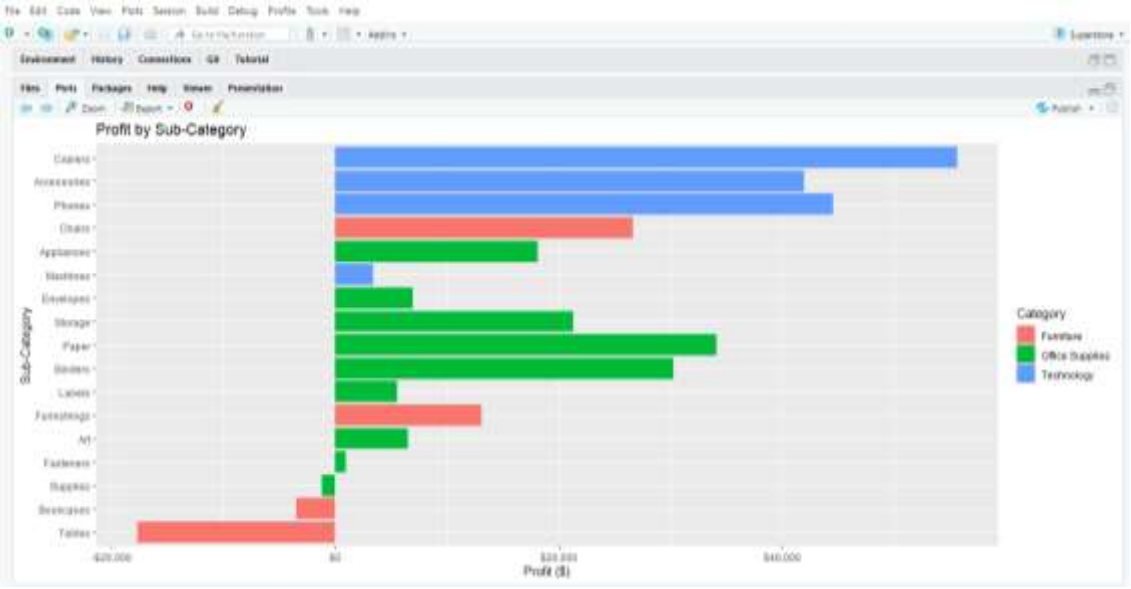

Figure 3: Profit by Sub-Category

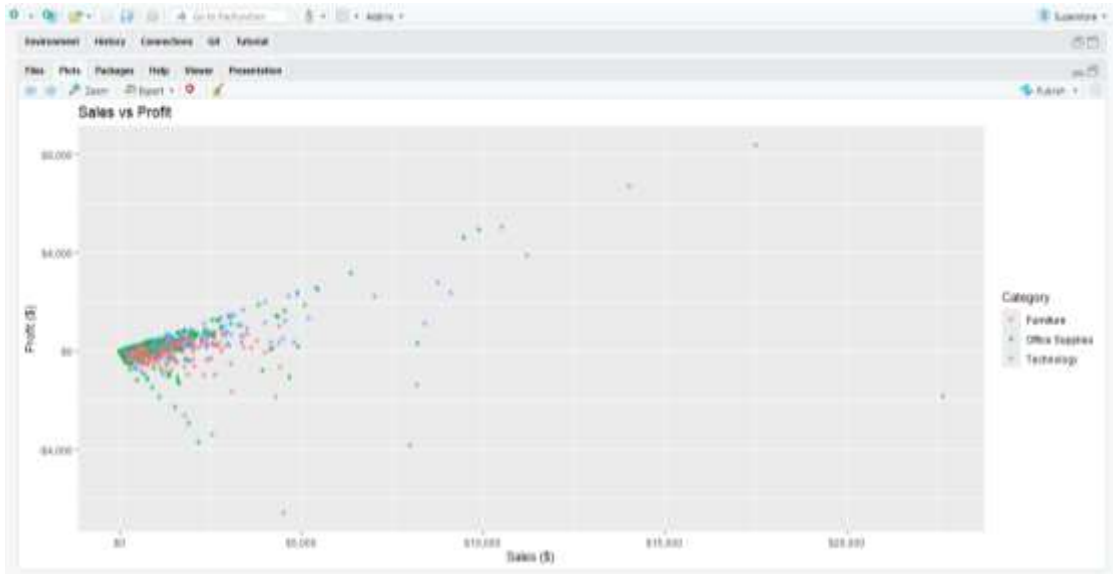

Figure 4: Sales vs Profit

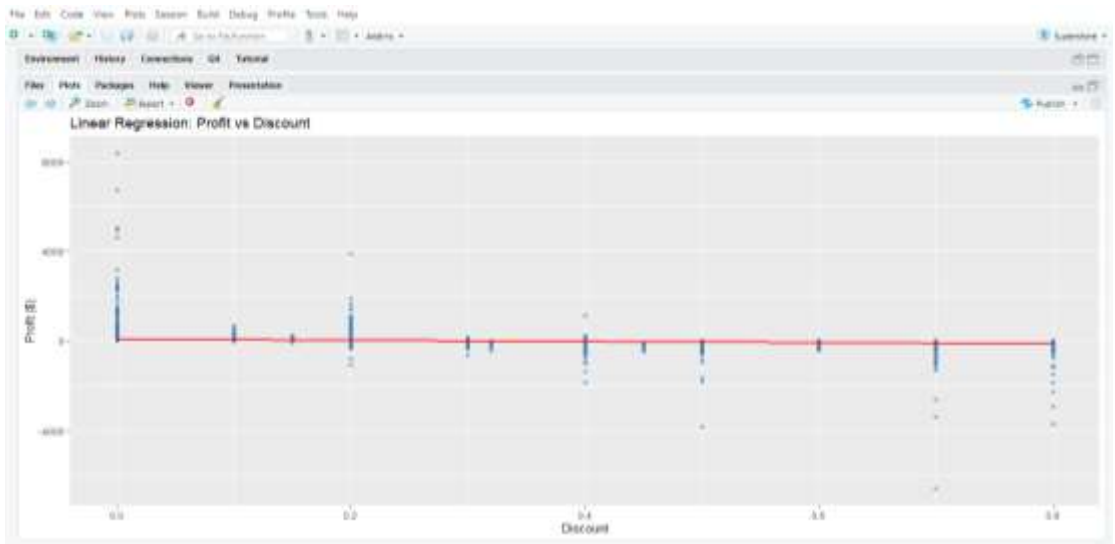

Figure 5: Linear Regression – Profit vs Discount

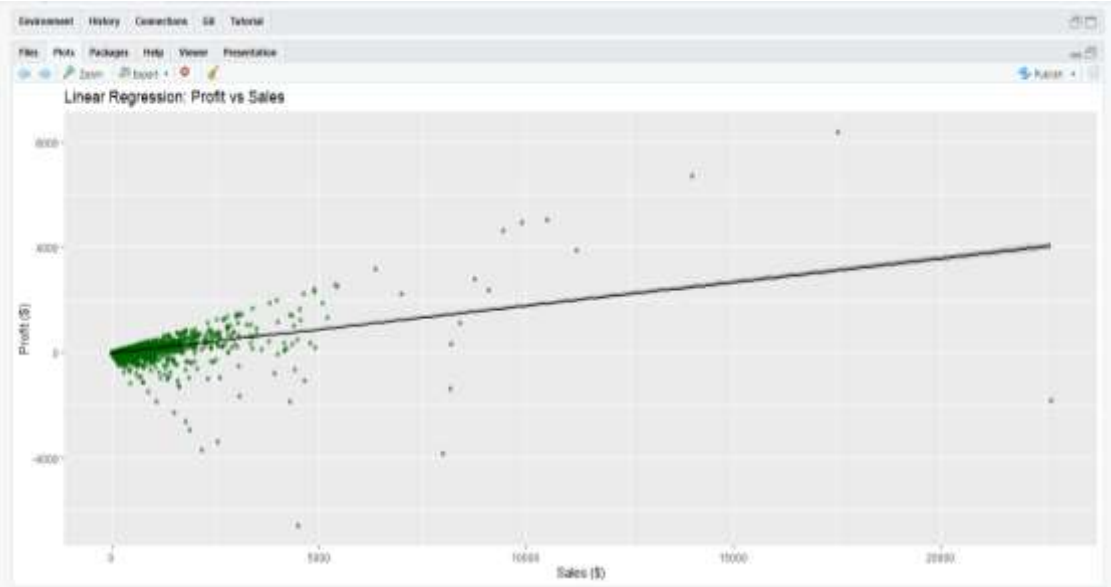

Figure 6: Linear Regression – Profit vs Sales

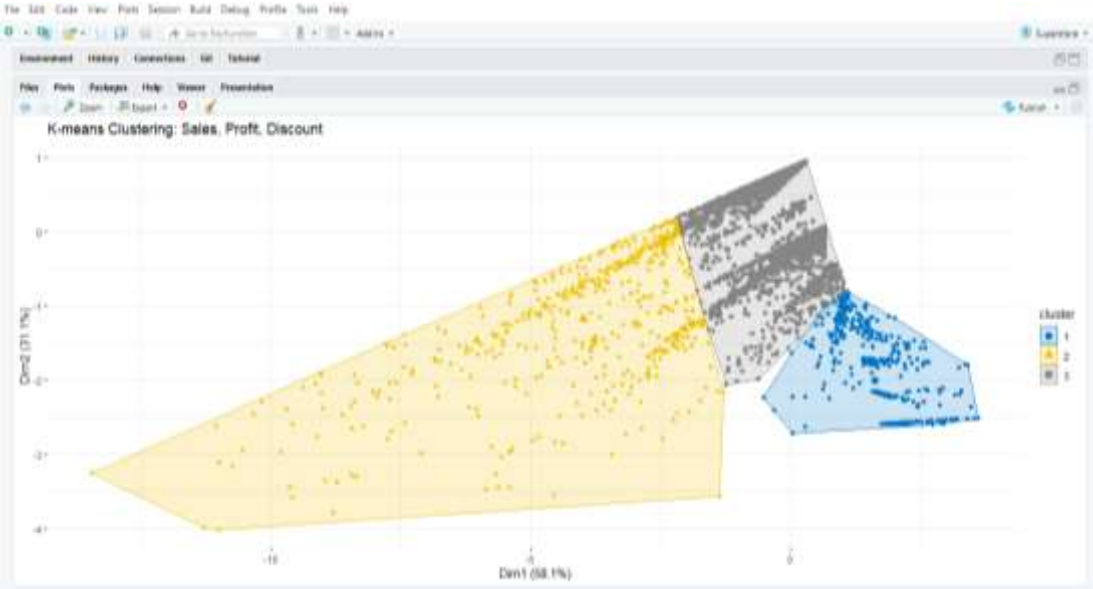

Figure 7: K-Means Cluster Plot (Sales, Profit, Discount)