
# Project Report: Retail Sales Analysis

**Mukand Krishna**

**7/12/23**

## Abstract:

This report presents an analysis of retail sales data with the objective of gaining insights and making predictions regarding customer behavior and sales trends. The dataset used for this analysis consists of sales records from a supermarket. The report explores various aspects of the data, including data cleaning, exploratory data analysis, and classification tasks to predict customer type. Different models such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machines (SVM) are employed for the classification tasks. The accuracy of each model is evaluated, and a comparison is made to determine the best-performing model.

## 1. Introduction:

The introduction section provides a brief overview of the project, stating the purpose and scope of the analysis. It introduces the dataset used and highlights the importance of retail sales analysis for understanding customer behavior, improving sales strategies, and enhancing business performance.

## 2. Objectives:

The objectives section outlines the specific goals of the analysis. It includes the tasks performed, such as data cleaning, exploratory data analysis, customer analysis, product analysis, payment channel analysis, temporal analysis, clustering analysis, and classification tasks. The objectives section helps readers understand the purpose of each analysis and what insights are sought.

### i. Data Cleaning:

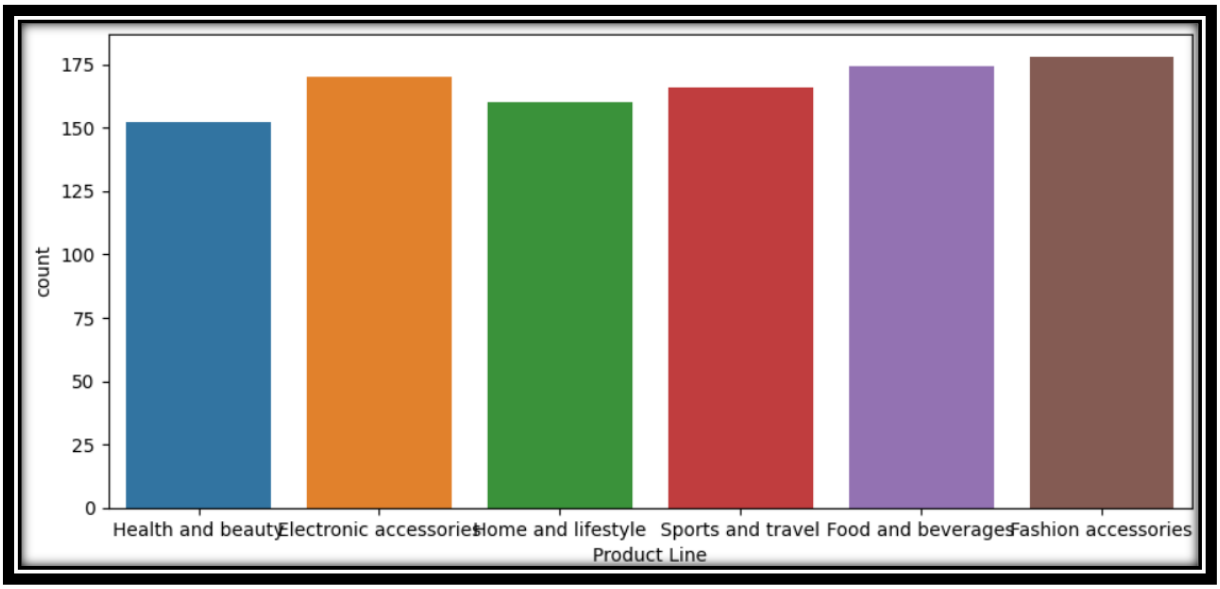
**IQR method** involves calculating the IQR, which is the range between the 25th percentile (Q1) and the 75th percentile (Q3) of the variable. Any data point that falls below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR is considered an outlier and is removed from the dataset.

**Z-score Method** measures how many standard deviations a data point is away from the mean of the variable. Z-score is calculated for each data point. Data points with a Z-score greater than a specified threshold (typically 2 or 3) are considered outliers and are removed from the dataset.

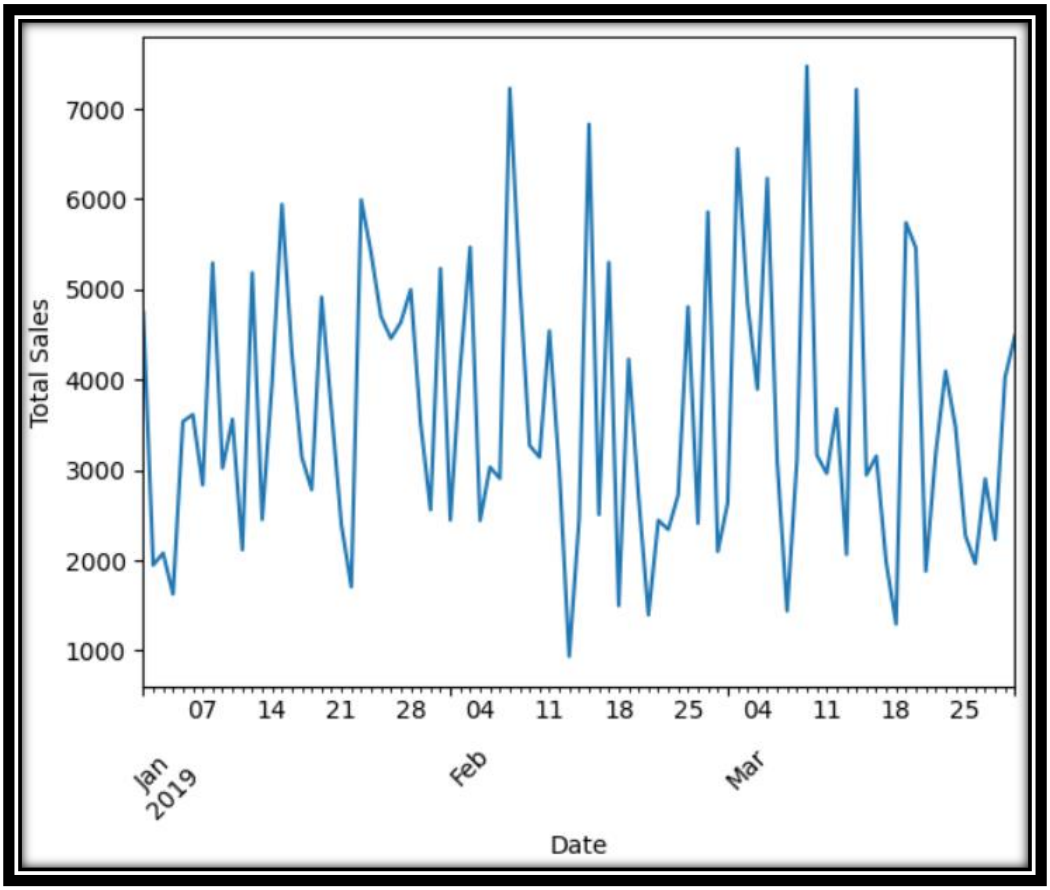
```
17 # Removing Outliers: Using Z-score method:
18
19 from scipy import stats
20 z_scores = stats.zscore(sales['Total'])
21 sales = sales[(np.abs(z_scores) < 3)]
22
23
24 # Using IQR method:
25
26 Q1 = sales['Total'].quantile(0.25)
27 Q3 = sales['Total'].quantile(0.75)
28 IQR = Q3 - Q1
29 sales = sales[~((sales['Total'] < (Q1 - 1.5 * IQR)) | (sales['Total'] > (Q3 + 1.5 * IQR)))]
30
```

## ii. Exploratory Data Analysis:

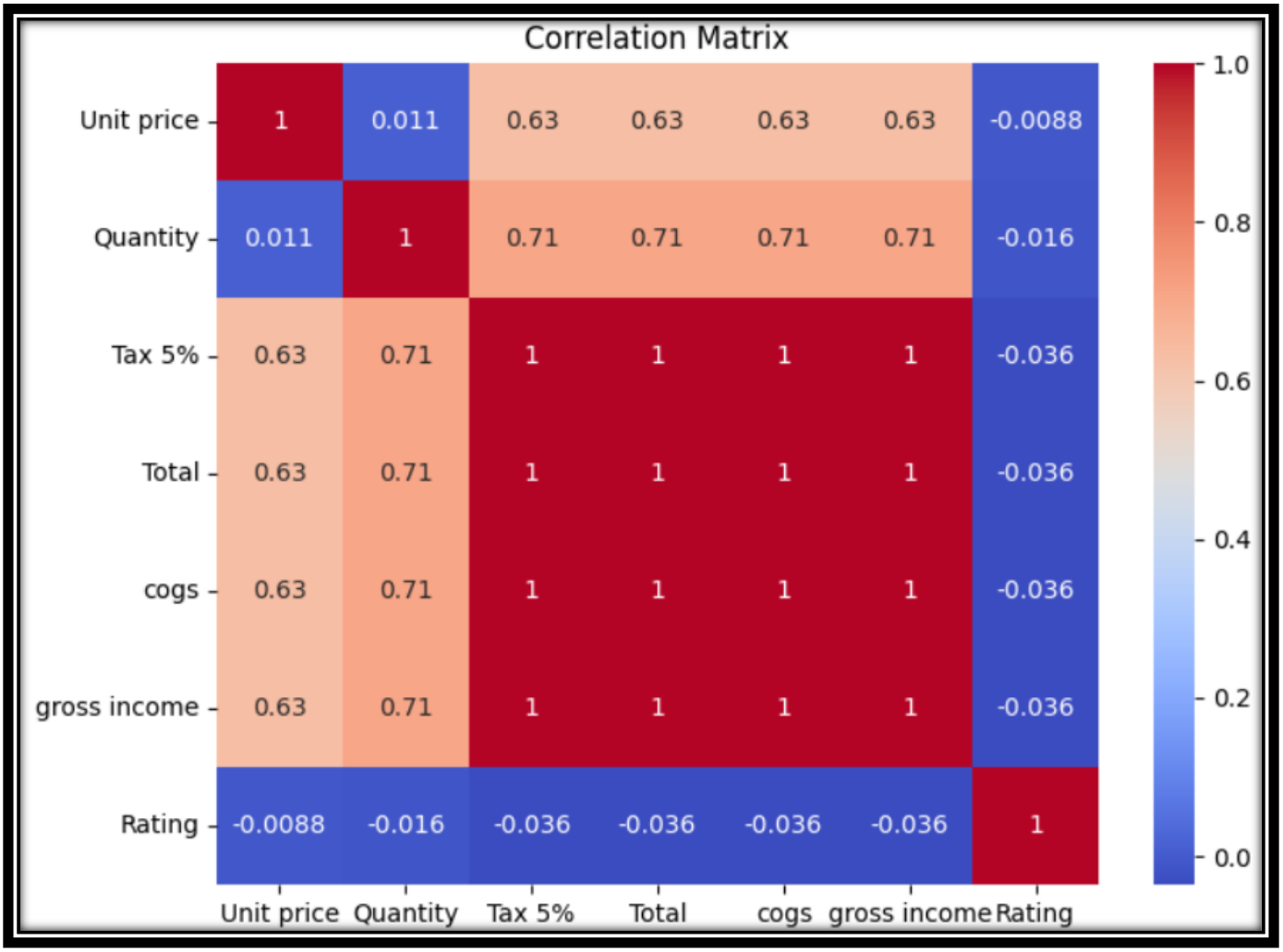
### Bar graph of Number of Products



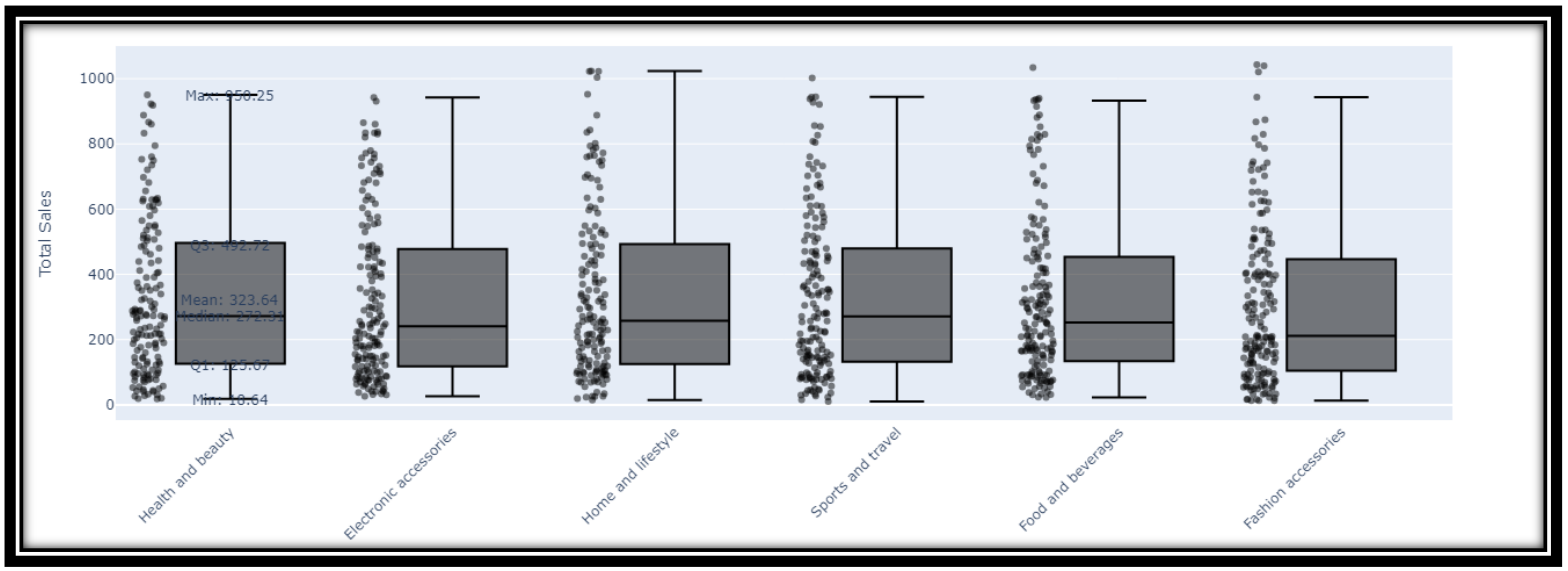

### Line graph of Total sales over Time



## Determining how one variable changes in relation to another variable



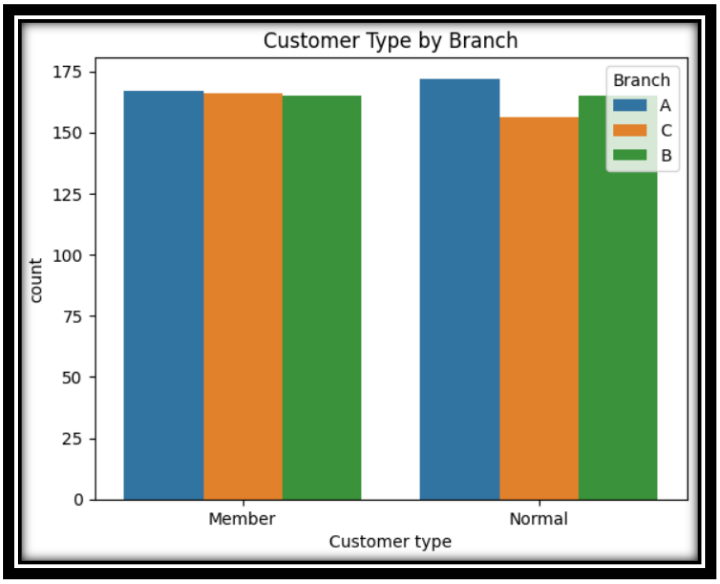

## Box Plot of Total Sales by Product line



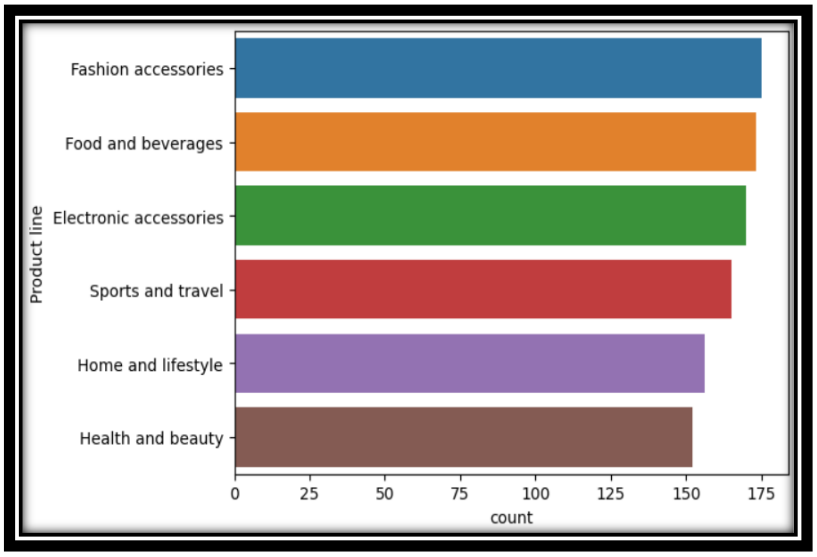
## 3. Results and Analysis:

The results and analysis section present the findings obtained from the analysis. It includes visualizations and key observations from each analysis, such as distribution analysis, customer behavior patterns, product performance, payment channel preferences, temporal trends, and customer segmentation. The section highlights important trends, relationships, and patterns discovered in the data.

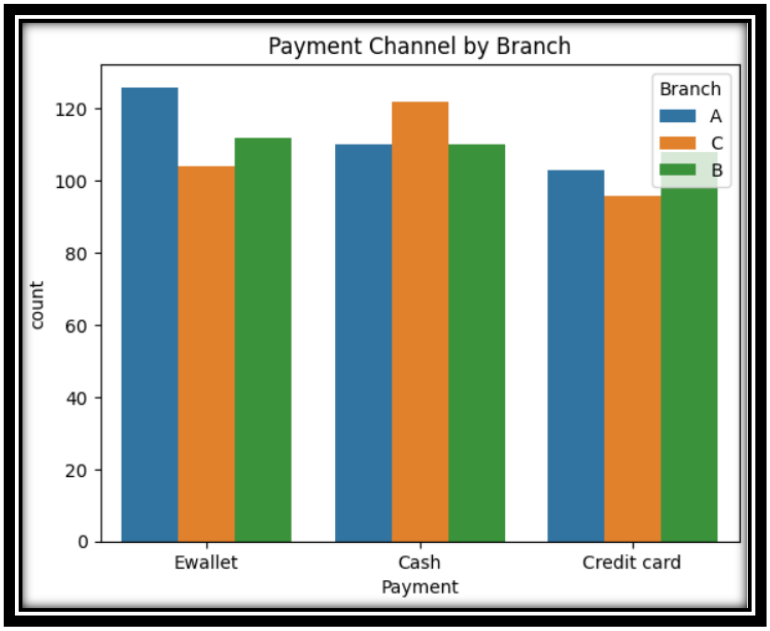
### i. Customer Analysis



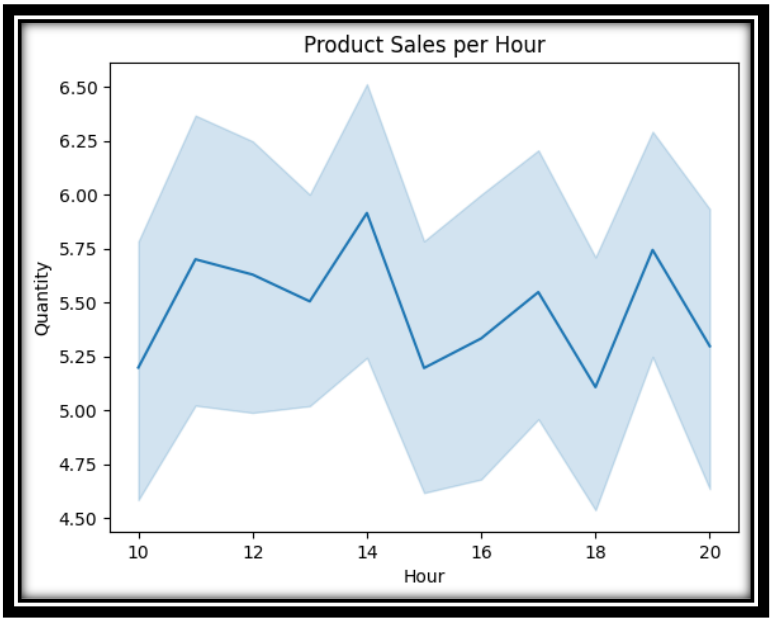

### ii. Product Analysis



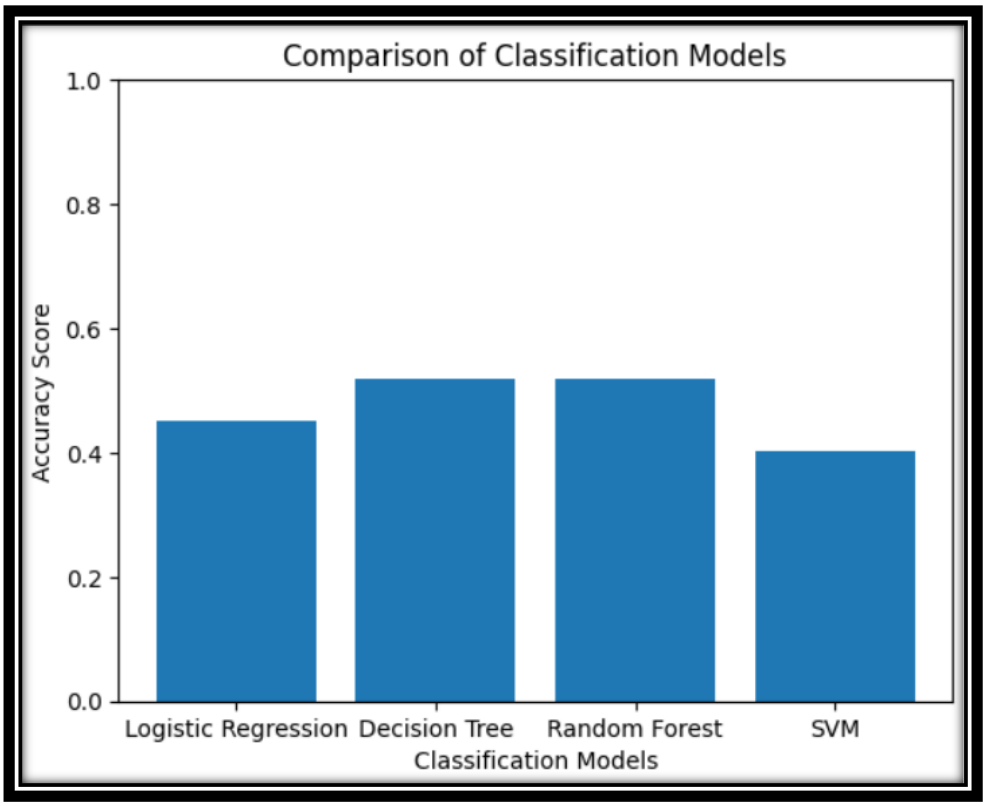

### iii. Payment Channel Analysis



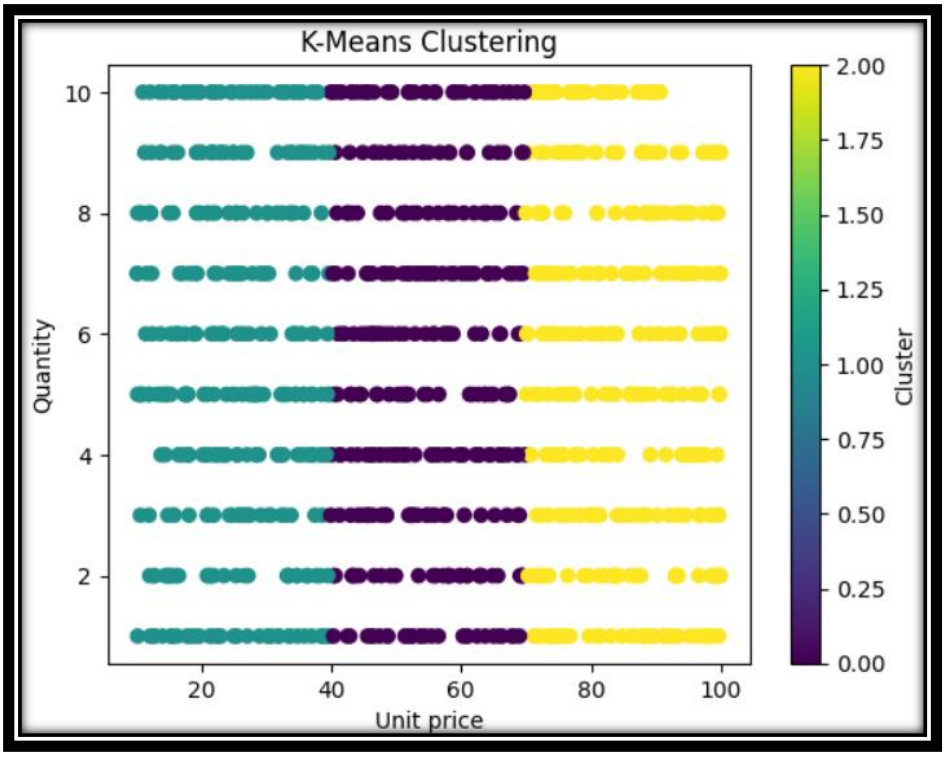

### iv. Temporal Analysis



**Performing classification tasks using different models to predict the Customer type**




**Sales data points are grouped together based on their similarities**



## 4. Power Bi Dashboard:

The dashboard includes interactive visualizations that allow users to explore the data and gain insights.


## 5. Conclusion:

In conclusion, the sales trends, customer behavior, and product performance were examined. The analysis revealed that certain product lines, such as *Fashion Accessories and Health and Beauty*, have higher sales and customer ratings. It also highlights the importance of understanding customer types, with customers who *spend a moderate amount of money and buy a variety of product* being the most common. Discovered peak sales hours and monthly variations. Analyzed payment channels and found that *E-wallet and Cash* were the preferred methods. Project also includes the application of machine learning models for customer type prediction, achieving moderate accuracy rates. The *decision tree* model is the most accurate among the models tested.

To enhance the business based on the findings, several actions can be taken. For instance, allocating resources to popular product lines, optimizing inventory management based on peak sales hours, and further analyzing customer preferences for targeted marketing efforts. Additionally, considering the feedback from low-rated products can lead to improvements in customer satisfaction.