Predictive Customer Segmentation and Classification in Supermarket Sales

INTRODUCTION:

In this machine learning project, predictive data analytics will be carried out utilizing a historical sales dataset from a grocery company. In order to perform this, I intend to categorize customers into these segments using a variety of classification models.

Dataset Link : https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

DATA DESCRIPTION:

Column Name	Description
Invoice ID	A computer generated sales slip invoice identification number
Branch	The branch of the supercenter where the sale took place
City	The location of the supercenter
Customer type	Either Member or Normal. Members are customers who have a member card, and Normal customers are those who do not have a member card
Gender	The gender of the customer
Product line	The general item categorization group. There are six product lines: Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, and Sports and travel
Unit price	The price of each product in $
Quantity	The number of products purchased by the customer
Tax	The 5% tax fee for the customer buying
Total	The total price including tax
Date	The date of purchase
Time	The purchase time
Payment	The payment method used by the customer for purchase. There are three payment methods: Cash, Credit card, and Ewallet

COGS	The cost of goods sold
Gross margin percentage	The gross margin percentage
Gross income	The gross income
Rating	The customer stratification rating on their overall shopping experience

Data Cleaning:

The initial step involves examining the dataset for the presence of null values. The isnull() method is applied to the dataset, and the sum() function is used to count the number of null values in each column. The output indicates that there are no null values in any of the columns, as all counts are zero.

To ensure that each attribute is represented in the correct format, the datatypes of each column are inspected using the dtypes attribute. This step helps identify any inconsistencies or potential issues with the data types.

The absence of null values eliminates the need for imputation or removal, and the correct datatypes ensure that each attribute is appropriately represented. This cleaned dataset is now ready for further exploration, analysis, and modeling, providing a reliable basis for deriving meaningful insights.

Data Pre-processing:

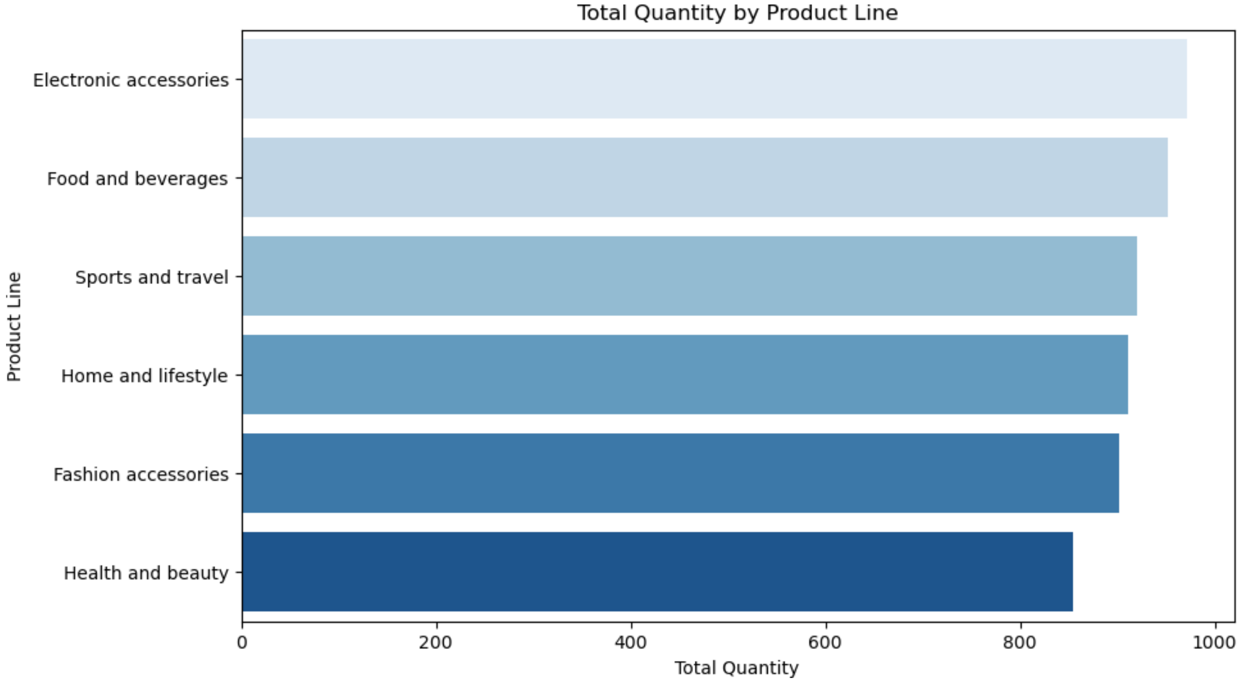
The first step involves converting the 'Date' column to a datetime format. This is a fundamental preprocessing step that enables the dataset to interpret and utilize temporal information effectively. The pandas to_datetime function is employed for this purpose.

Following the conversion of the 'Date' column, two additional features, 'DayOfWeek' and 'Month,' are extracted to provide more granular temporal information. The 'DayOfWeek' feature is created to capture the day of the week, while the 'Month' feature extracts the numeric representation of the month.

These transformations lay the groundwork for further exploratory data analysis (EDA) and modeling, allowing for more insightful and context-rich interpretations of the dataset.

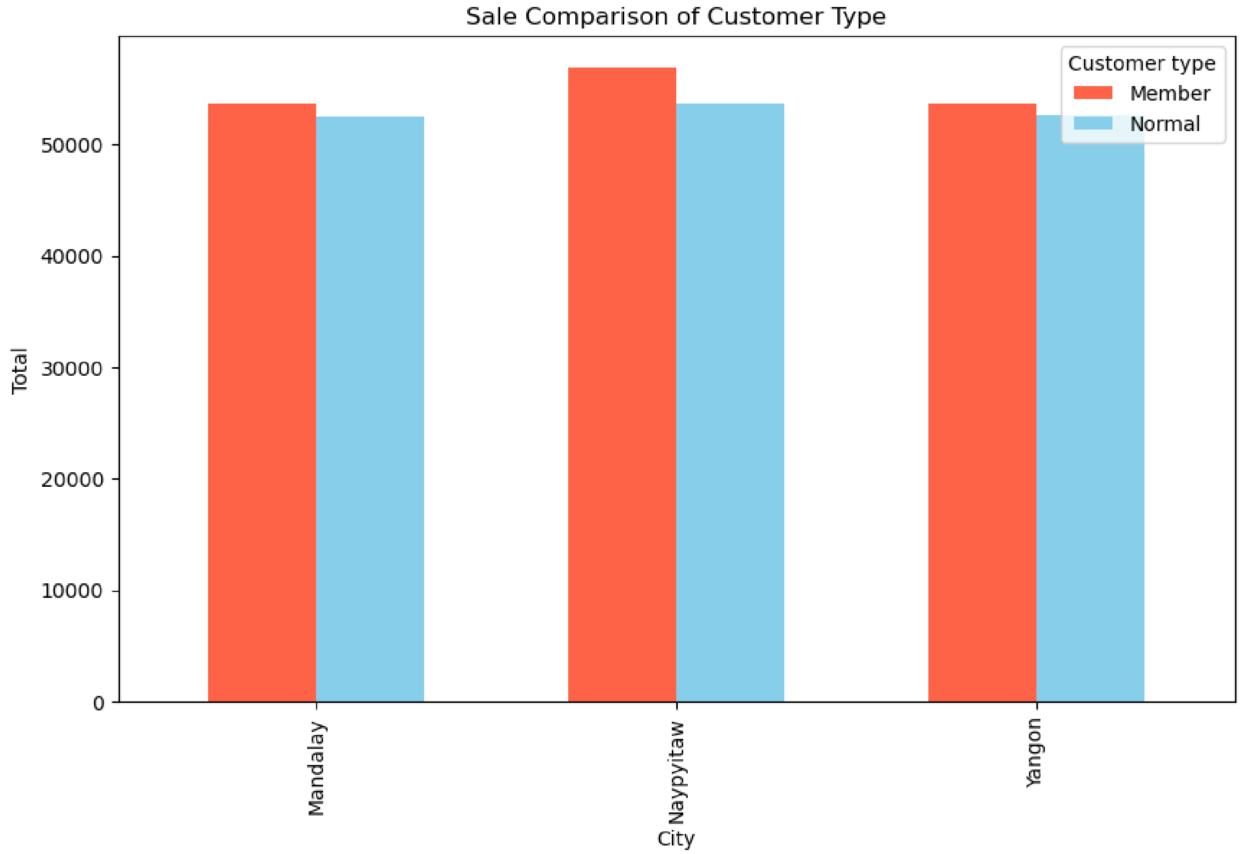
Data Analysis:

Analysis of the number of Quantity sold in each product line:



This data analysis shows that Electronic accessories products are the most popular, followed by Food and beverages, Sports and travel,HOme and Lifestyle, Fashion accessories, and Health and beauty. The top three product lines account for over 60% of total sales and businesses should focus on these product lines in order to maximize profits. The bottom three product lines account for less than 40% of total sales, but businesses should still focus on these product lines if they can be targeted to specific niche markets.

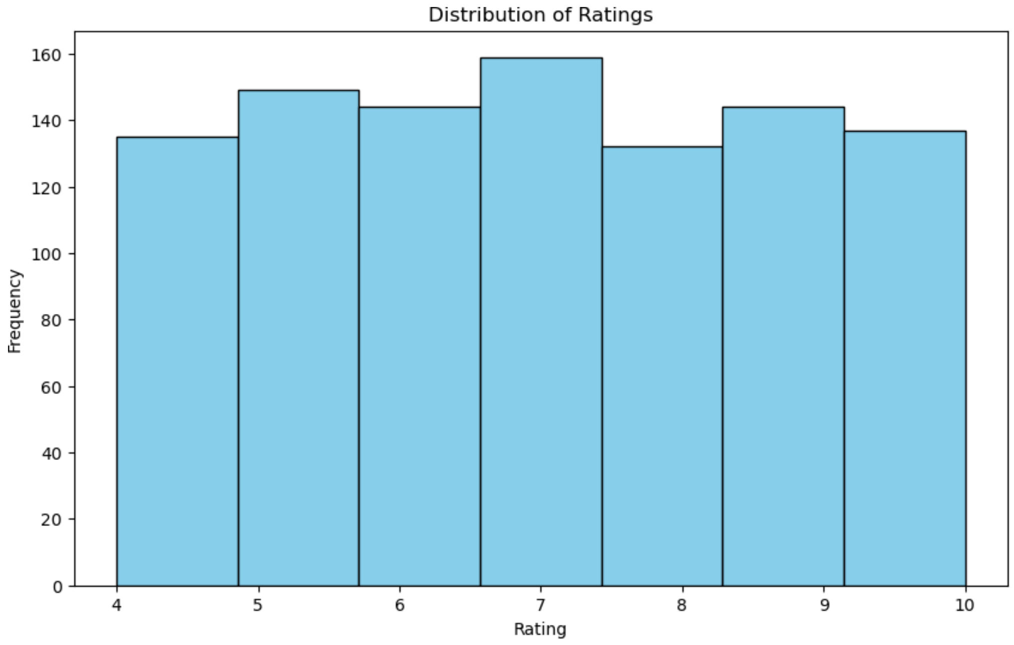
Comparing the sale of each city for each customer type



The bar chart shows the total sales for each city, broken down by customer type. Member customers are the largest customer segment in all three cities, but they make up a larger proportion of sales in Naypyitaw than in Yangon and Mandalay. This suggests that Member customers may be more likely to shop in Naypyitaw than Normal customers.

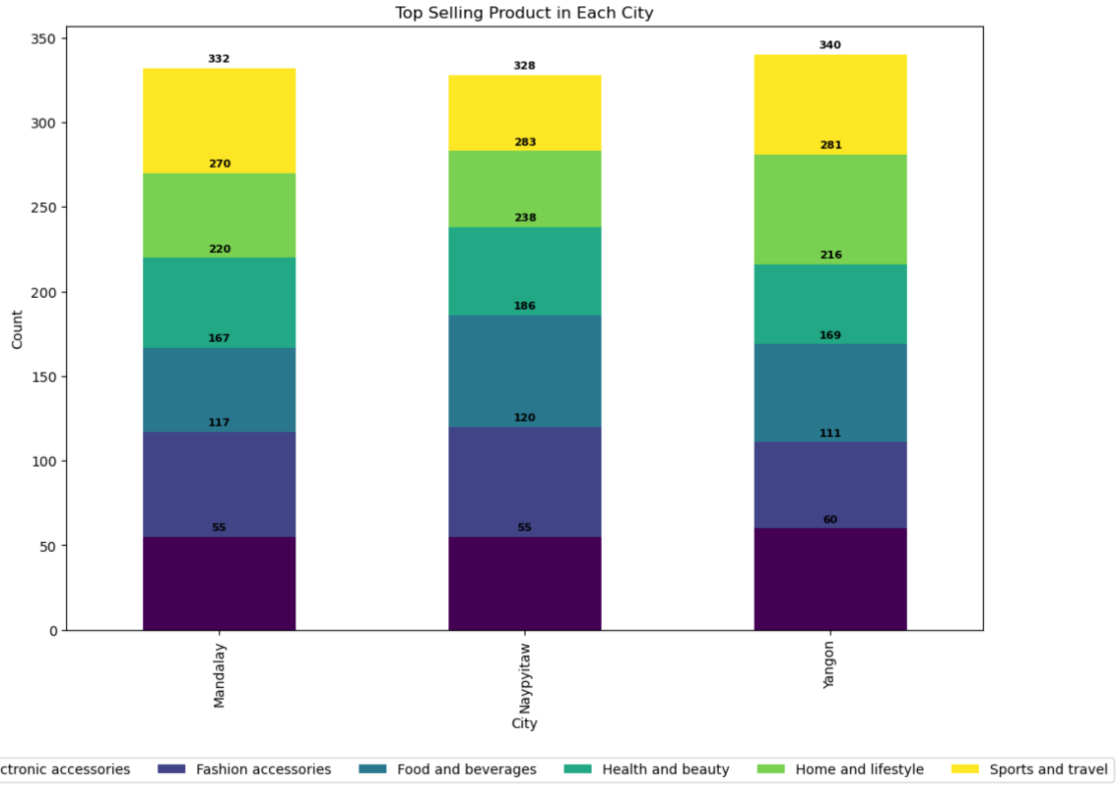
Normal customers make up a little smaller proportion of sales overall in all the cities than the Members.

Comparing the distribution of rating:



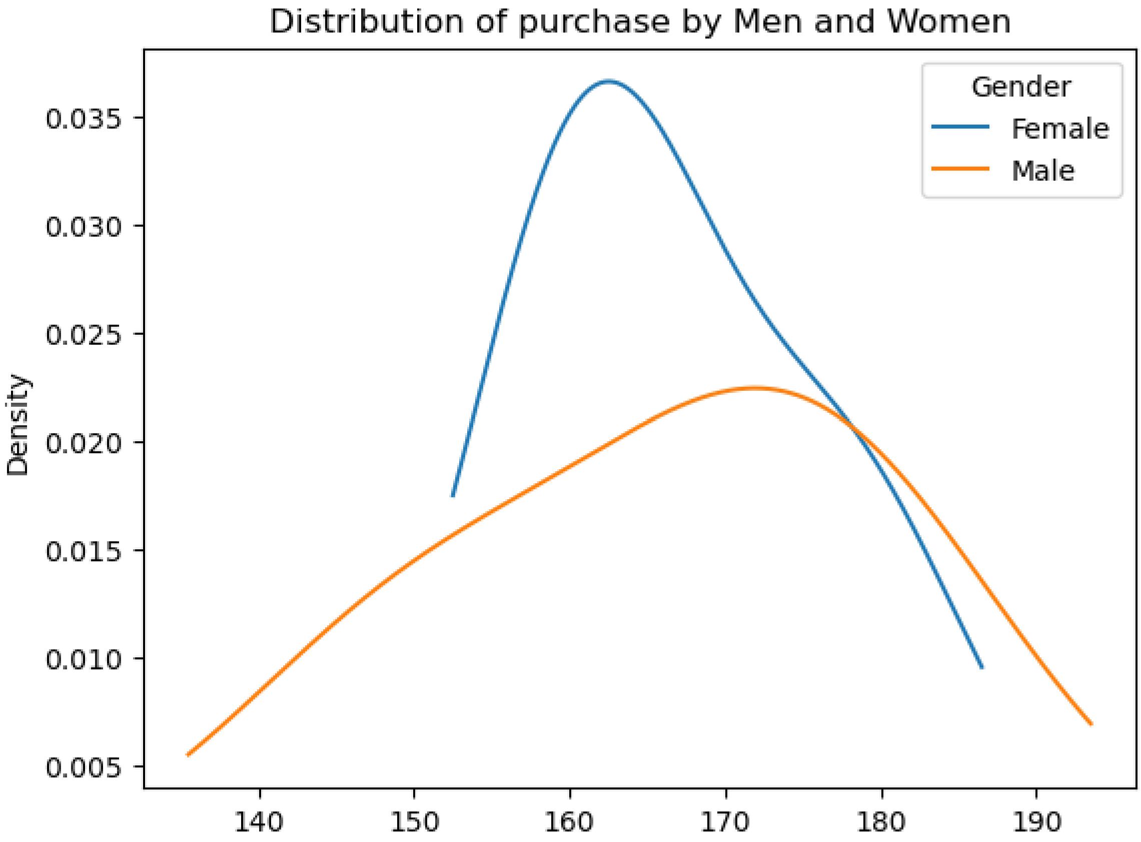
The distribution of customer ratings shows that the majority of customers are satisfied with their purchases.

Analyzing the top selling product in each city



The stacked bar plot shows that Electronic accessories are the top selling product in all three cities followed by Sports and Travel. Businesses should focus on promoting Electronic accessories in all three cities, and tailor their marketing and sales strategies to the specific product preferences of different cities.

Analyzing the distribution of men and Women based on Costs of goods sold:



MODELS:

Model-1: Predicting Total Sales Amount using Linear Regression

The focus of Model 1 is to predict the total sales amount utilizing a Linear Regression approach. Linear Regression is a foundational machine learning algorithm that establishes a linear relationship between the dependent variable (total sales) and a set of independent variables (features). This report provides an overview of the methodology, training process, and evaluation results for Model 1.

Methodology:

The dataset underwent preprocessing steps to ensure its suitability for training a linear regression model. This included creating a copy of the original data (**md1Data**), selecting relevant features, and transforming categorical variables into numerical format through one-hot encoding.

Careful consideration was given to feature selection to identify variables that could significantly impact total sales. The selected features include 'Branch,' 'Customer type,' 'Product line,' 'Unit price,' 'DayOfWeek,' 'Month,' and 'Quantity.'

The dataset was split into training and testing sets using the train_test_split function from scikit-learn. The Linear Regression model was trained on the training set using the LinearRegression class. The training process involved fitting the model to the training data, allowing it to learn the relationships between the selected features and the total sales amount.

Evaluation:

The model's performance was assessed using the R-squared metric, which measures the proportion of the variance in the dependent variable (total sales) that is predictable from the independent variables. The calculated R-squared value was found to be 0.8911, indicating a high level of predictive accuracy.

Results:

The R-squared value of 0.8911 suggests that approximately 89.11% of the variance in total sales can be explained by the selected features in the model. A higher R-squared value signifies a better fit of the model to the data.

Conclusion:

Model 1 has demonstrated strong predictive capabilities, as evidenced by the high R-squared value. The inclusion of relevant features, appropriate preprocessing, and the use of the Linear Regression algorithm have contributed to the success of the model. Further exploration and refinement of features, hyperparameter tuning, and model validation can be considered for potential improvements. The high R-squared value instills confidence in the model's ability to predict total sales accurately, making it a valuable asset for sales forecasting.

Model-2: Predicting Rating using RandomForestClassifier

Methodology:

The dataset underwent preprocessing to enhance its suitability for a RandomForestClassifier model. Key steps included additional feature engineering, dropping irrelevant columns ('Invoice ID', 'Date', 'Time', 'Gender', 'Customer type'), and converting the 'Rating' variable into categories ('low,' 'medium,' 'high'). New features, 'DayOfWeek' and 'Month,' were created.
'Branch,' 'Product line,' 'Payment,' 'Unit price,' 'Quantity,' 'Tax 5%,' 'Total,' 'DayOfWeek,' and 'Month' were selected as features for the model.
And then, The dataset was divided into training and testing sets using the train_test_split function from scikit-learn.

Evaluation:

A RandomForestClassifier model was constructed using a scikit-learn Pipeline. The model included a preprocessor with SimpleImputer for numerical features and OneHotEncoder for categorical features. StandardScaler was applied to scale the features. The model was trained on the training set.
GridSearchCV was employed to find the optimal hyperparameters for the RandomForestClassifier. Parameters tuned included the number of estimators, maximum depth, minimum samples split, and minimum samples leaf.

Results:

The best model obtained from the grid search was used to make predictions on the test set. The accuracy of the model on the testing set was evaluated using accuracy_score and a classification report, providing precision, recall, and F1-score for each category.

Conclusion:

The RandomForestClassifier demonstrated 52% accuracy on the test set. The inclusion of relevant features, appropriate preprocessing, and hyperparameter tuning contributed to the accuracy of the model. Further exploration, including refining features or trying alternative models, could be considered for potential improvements. The classification report provides detailed insights into the model's performance for different categories.

Model-3: Customer Segmentation using K-Means Clustering

The primary objective of Model 3 is to perform customer segmentation through K-Means Clustering, shedding light on distinct customer behavior patterns. Unlike regression-based models, K-Means Clustering is an unsupervised learning technique that categorizes customers into groups based on their purchasing behavior.

Methodology:

The initial steps involved data preprocessing, including the removal of irrelevant columns ('Invoice ID' and 'Date'), statistical summary checks, identification of categorical variables, and ordinal encoding. Ordinal encoding was employed to convert categorical variables into numerical format, facilitating the subsequent analysis.

Exploratory Data Analysis (EDA) was conducted to visualize the correlation matrix and relationships within the data. The heatmap and pairplot generated during EDA offered insights into the interplay between different features, aiding in better understanding the dataset.

The K-Means Clustering process began with feature scaling using StandardScaler. The Elbow method was employed to determine the optimal number of clusters (k) by analyzing inertia values. Subsequently, K-Means clustering with three clusters was performed. Constant and highly correlated features were identified and removed to enhance the clustering process.

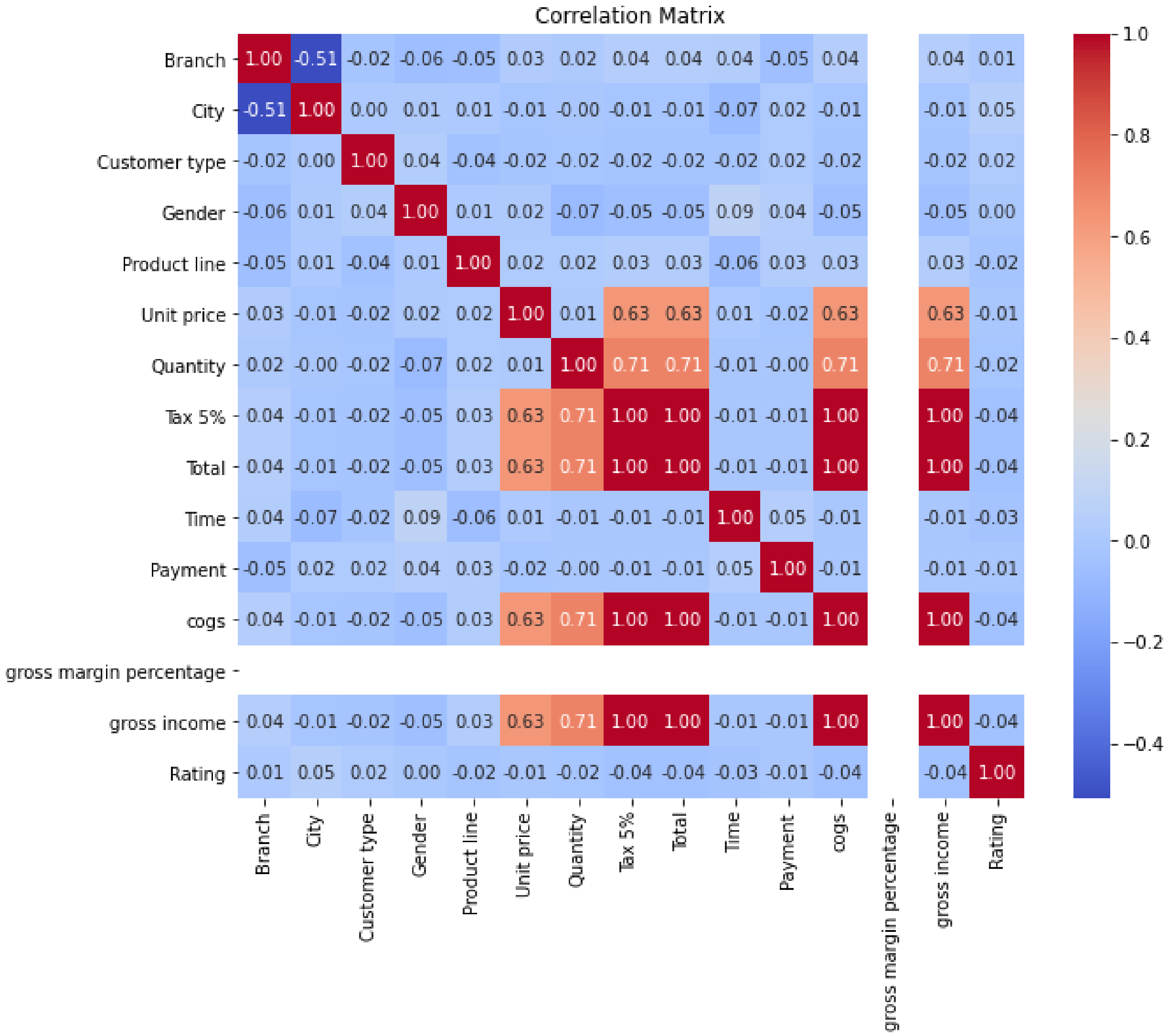
Evaluation of the clustering results involved calculating Silhouette Score, Davies Bouldin Score, and Calinski-Harabasz Score. These metrics provided quantitative assessments of the quality and coherence of the identified clusters.

Results:

The following pairplot visualizations provide a comprehensive understanding of the clustered data, showcasing patterns and relationships within each cluster. Each point in the plot represents a customer, color-coded by the assigned cluster. These plots are instrumental in visually discerning the characteristics of each cluster and understanding the variations in purchasing behavior among customer segments.

Pairplot Visualization 1: Correlation Matrix

The correlation matrix heatmap visually represents the correlation coefficients between different features. This aids in understanding the relationships and dependencies within the dataset.



	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Time	Payment	cogs	gross margin percentage	gross income	Rating
Branch	1.00	-0.51	-0.02	-0.06	-0.05	0.03	0.02	0.04	0.04	0.04	-0.05	0.04		0.04	0.01
City	-0.51	1.00	0.00	0.01	0.01	-0.01	-0.00	-0.01	-0.01	-0.07	0.02	-0.01		-0.01	0.05
Customer type	-0.02	0.00	1.00	0.04	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02	0.02	-0.02		-0.02	0.02
Gender	-0.06	0.01	0.04	1.00	0.01	0.02	-0.07	-0.05	-0.05	0.09	0.04	-0.05		-0.05	0.00
Product line	-0.05	0.01	-0.04	0.01	1.00	0.02	0.02	0.03	0.03	-0.06	0.03	0.03		0.03	-0.02
Unit price	0.03	-0.01	-0.02	0.02	0.02	1.00	0.01	0.63	0.63	0.01	-0.02	0.63		0.63	-0.01
Quantity	0.02	-0.00	-0.02	-0.07	0.02	0.01	1.00	0.71	0.71	-0.01	-0.00	0.71		0.71	-0.02
Tax 5%	0.04	-0.01	-0.02	-0.05	0.03	0.63	0.71	1.00	1.00	-0.01	-0.01	1.00		1.00	-0.04
Total	0.04	-0.01	-0.02	-0.05	0.03	0.63	0.71	1.00	1.00	-0.01	-0.01	1.00		1.00	-0.04
Time	0.04	-0.07	-0.02	0.09	-0.06	0.01	-0.01	-0.01	-0.01	1.00	0.05	-0.01		-0.01	-0.03
Payment	-0.05	0.02	0.02	0.04	0.03	-0.02	-0.00	-0.01	-0.01	0.05	1.00	-0.01		-0.01	-0.01
cogs	0.04	-0.01	-0.02	-0.05	0.03	0.63	0.71	1.00	1.00	-0.01	-0.01	1.00		1.00	-0.04
gross margin percentage															
gross income	0.04	-0.01	-0.02	-0.05	0.03	0.63	0.71	1.00	1.00	-0.01	-0.01	1.00		1.00	-0.04
Rating	0.01	0.05	0.02	0.00	-0.02	-0.01	-0.02	-0.04	-0.04	-0.03	-0.01	-0.04		-0.04	1.00

Correlation Matrix Insights:

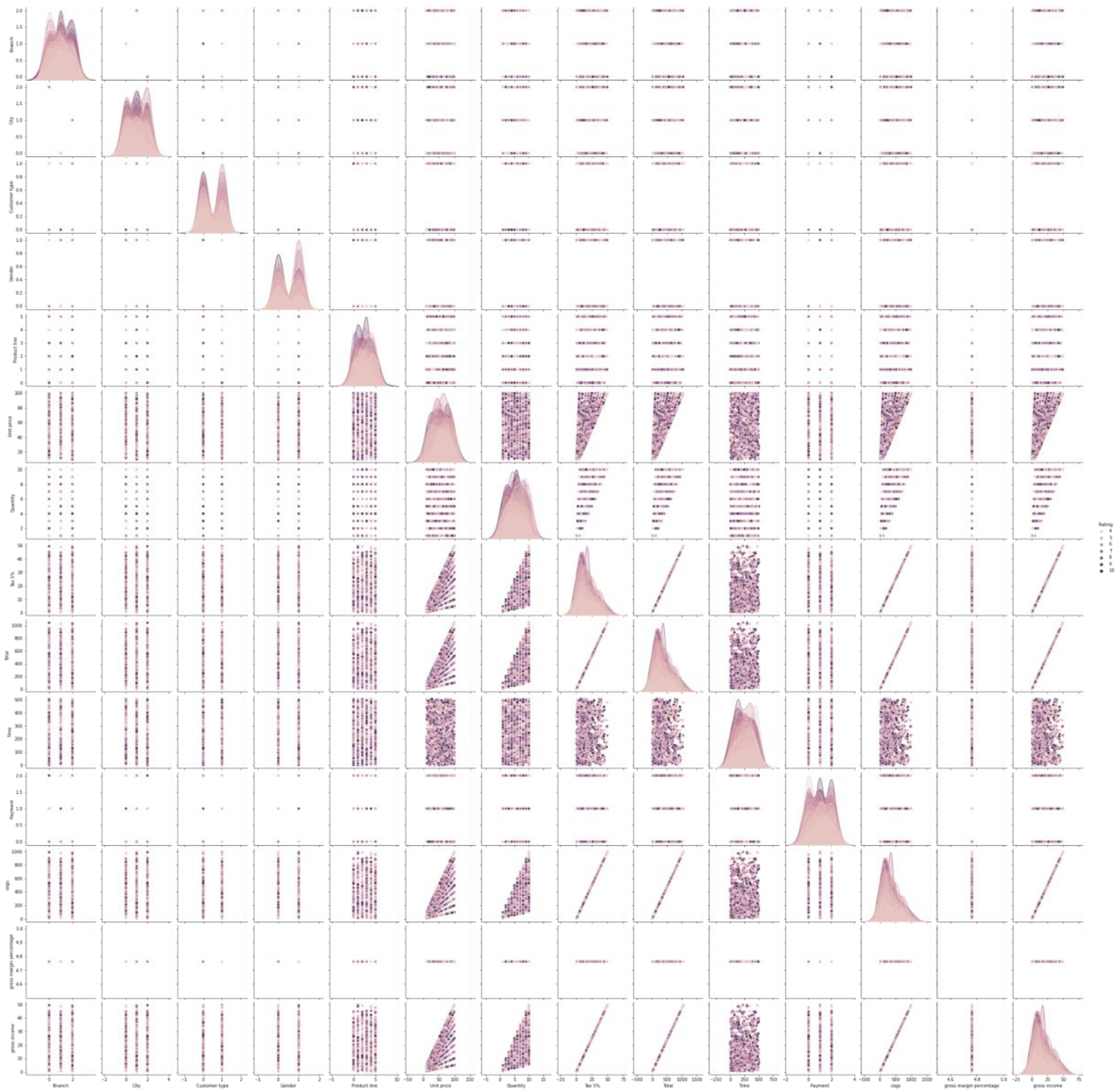
The correlation matrix visually represents the correlation coefficients between different features in the dataset. Here are additional insights into specific high correlations:

- High Correlation (Correlation Coefficient = 1):
    - 'COGS,' 'Total,' and 'Tax 5%' exhibit a perfect correlation of 1. This implies a linear relationship where changes in one variable are completely reflected in the other.
- Moderate to High Correlation:
    - 'Quantity' demonstrates a correlation of 0.71 with 'COGS,' 'Total,' and 'Tax 5%.' This indicates a substantial positive linear relationship, suggesting that an increase in quantity sold is associated with an increase in these financial metrics.
    - 'Unit Price' has a correlation of 0.63 with 'COGS,' 'Total,' and 'Tax 5%.' This suggests a moderate positive linear relationship, indicating that changes in unit price are associated with corresponding changes in these financial indicators.
- Low Correlation:
    - The remaining features have correlations of less than 0.05 with 'COGS,' 'Total,' and 'Tax 5%,' indicating a weak linear relationship.

These correlation insights provide a deeper understanding of the relationships between key financial metrics and individual product characteristics. They can guide further analysis and decision-making processes related to pricing, quantity optimization, and overall sales strategy.

Pairplot Visualization 2: Relationships in the Data

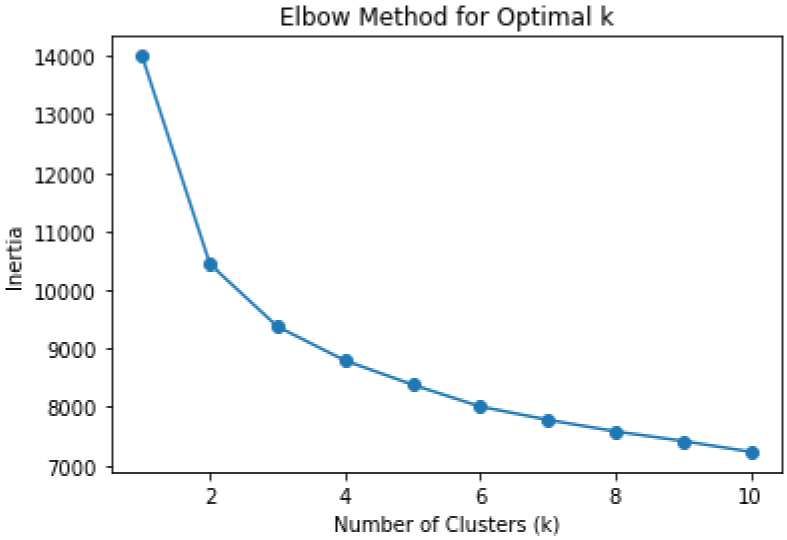
The pairplot provides scatterplots for each pair of features, facilitating the visualization of relationships between variables. The hue parameter is set to 'Rating,' allowing for differentiation based on customer ratings.



Elbow Plot:

The Elbow Method is employed to determine the optimal number of clusters (k) for the K-Means Clustering algorithm. The inertia, which represents the sum of squared distances between data points and their assigned cluster center, is calculated for different values of k. The point where the inertia begins to decrease at a slower rate is considered the "elbow" and is indicative of the optimal number of clusters.

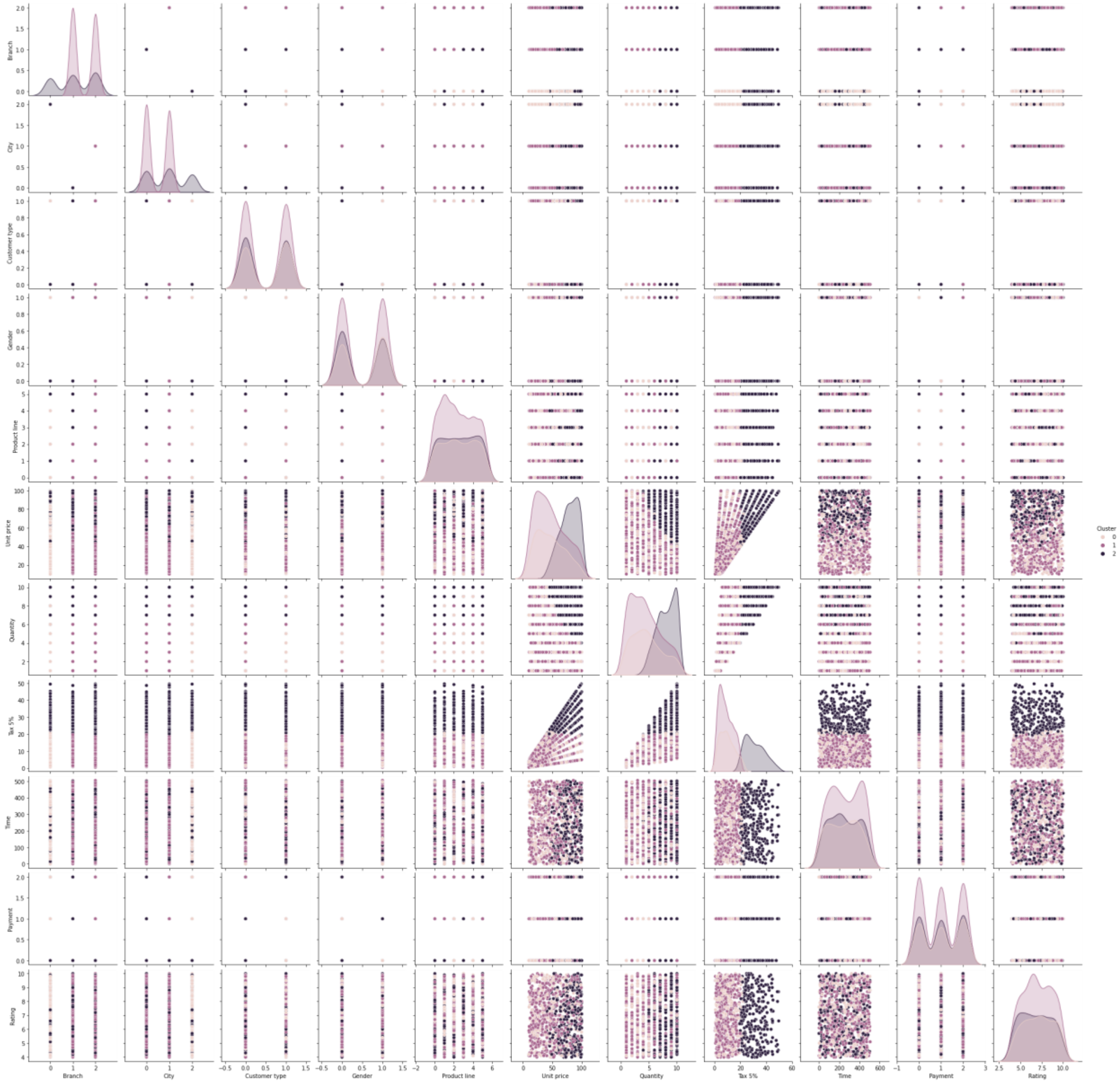
The elbow plot aids in identifying the optimal number of clusters by visually inspecting the rate of inertia reduction across different k values. Here's the code used to generate the elbow plot:



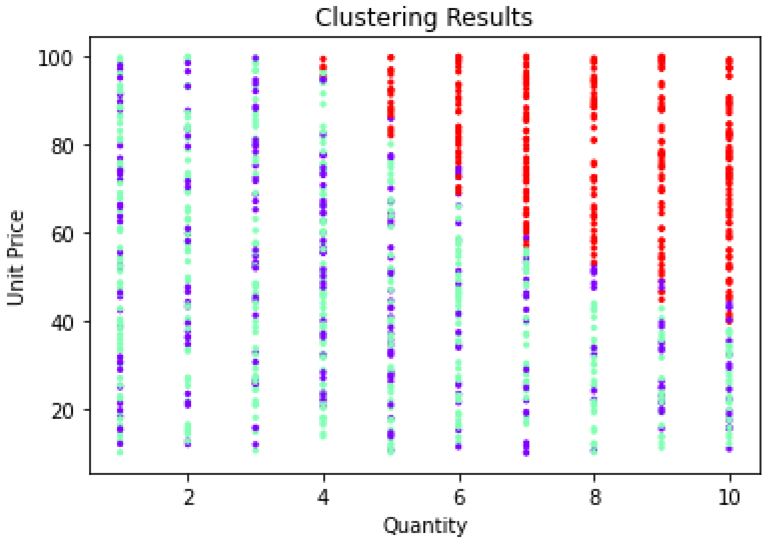
In this specific case, the "elbow" in the plot occurs at k=3, suggesting that three clusters are appropriate for segmenting the customers based on their purchasing behavior. This informed the decision to set the number of clusters to 3 in the subsequent K-Means Clustering process.

Pairplot Visualization 3: Relationships in Clustered Data

After clustering, a pairplot is generated to visualize relationships within the clustered data. Each point is color-coded by the assigned cluster, enabling a clear understanding of patterns in each segment.



Pairplot Visualization 4: Clustering Results



In this Pairplot Visualization, I explored the results of the K-Means clustering applied to the supermarket sales dataset. Each point on the plot represents a data point, colored according to the assigned cluster. The Pairplot is a grid of scatterplots, where each pair of features is visualized against each other, providing a comprehensive view of the relationships within the clustered data.

This visualization aids in discerning patterns, trends, and potential separations between clusters based on different feature combinations. The distinct color-coding of clusters allows for a quick and intuitive understanding of how the clustering algorithm has grouped the data points in the multidimensional feature space.

Key elements to observe include:

- Scatterplots: Scatterplots reveal how pairs of features relate to each other within each cluster. Different clusters may exhibit unique patterns, helping to identify characteristics that distinguish one cluster from another.
- Diagonal Distribution: Along the diagonal, histograms display the distribution of each feature, providing insights into the overall distribution of values within clusters.

By scrutinizing this Pairplot, I gained valuable visual insights into the efficacy of the clustering algorithm and potential patterns within the identified clusters. Patterns, overlaps, or separations in the scatterplots offer valuable information about the relationships between different features and how these relationships contribute to the clustering results.

Cluster Analysis Conclusion:

The K-Means clustering algorithm has effectively grouped customers into three distinct clusters, each shedding light on unique purchasing behaviors and preferences. Let's delve into each cluster individually to understand their characteristics and implications:

1. Cluster 0 - Balanced Buyers (259 clients):

- **Description:** This cluster represents clients showcasing a balanced purchasing pattern, with moderate quantities of products across various price points.

- **Insight:** These customers exhibit versatility in their buying choices, presenting an opportunity for targeted marketing strategies that cater to a broad spectrum of products and price ranges.

- *Average Quantity:* Moderate (mean = 4.67)
- *Unit Price:* Moderate
- *Purchasing Pattern:* Balanced quantities across different price points.
- *Insight:* This group represents clients with a versatile buying pattern, showing a balance between the quantity of products purchased and their associated unit prices.

2. Cluster 1 - Diverse Shopper Base (454 clients):

- **Description:** Cluster 1 comprises clients with moderate purchasing intensity and a diverse range of unit prices.

- **Insight:** The larger size of this cluster indicates a varied customer base with distinct shopping preferences. Targeted campaigns focusing on a diverse product range could effectively resonate with these customers.

- *Average Quantity:* Moderate (mean = 4.37)
- *Unit Price:* Slightly lower than Cluster 0
- *Purchasing Pattern:* Slightly lower purchasing intensity with a diverse range of unit prices.
- *Insight:* This larger cluster consists of clients with varied shopping preferences, demonstrating a more extensive range of unit prices.

3. Cluster 2 - Focused High-Spenders (287 clients):

- **Description:** This cluster represents clients with a higher purchasing intensity, concentrating on higher-priced items.

- **Insight:** These high-spending customers present an opportunity for premium offerings and exclusive promotions. Tailored marketing strategies could enhance their experience and potentially increase overall revenue.

- *Average Quantity:* Higher (mean = 8.07)
- *Unit Price:* Significantly higher
- *Purchasing Pattern:* Concentrated on higher-priced items.
- *Insight:* This group comprises clients with a focused approach to purchasing, showing a preference for higher-priced items and a more substantial quantity of products.

Key Metrics:

- Silhouette Score: 0.17 (indicating reasonable separation between clusters)

- Davies Bouldin Score: 2.00 (reflecting well-separated and compact clusters)

- Calinski-Harabasz Score: 245.74 (indicating well-defined clusters)

Additional Insights:

- Correlation analysis reveals relationships between different features, contributing to a comprehensive understanding of inter-feature dependencies.

Practical Implications:

- These identified clusters offer actionable insights for targeted marketing campaigns. By tailoring strategies to the specific preferences of each cluster, businesses can enhance customer engagement and satisfaction.

Strategic Considerations:

- Understanding the distinct needs of each cluster enables businesses to formulate personalized marketing approaches, improving overall customer experience and maximizing the effectiveness of sales strategies.

Conclusion

In conclusion, the clustering analysis serves as a valuable tool for customer segmentation, providing a nuanced understanding of customer behaviors. This understanding, coupled with the identified clusters, empowers businesses to craft strategic marketing initiatives that resonate with their diverse customer base, ultimately driving success and profitability.

Model-4: Predicting Quantity using Decision Tree

Model 4 employs a Decision Tree classifier to predict the quantity of products based on the given features. Decision Trees are versatile machine learning models capable of capturing non-linear relationships and hierarchies within the data. This report provides an overview of the model, its methodology, and the evaluation results.

Methodology:
The dataset is divided into the target variable 'Quantity' and the feature set 'X' by dropping the 'Quantity' column. Numerical and categorical columns are identified to guide the preprocessing steps. A preprocessor is constructed using a ColumnTransformer to handle both numerical and categorical features separately. Numerical features undergo standard scaling, while categorical features are one-hot encoded to ensure compatibility with the Decision Tree model. A pipeline is established to streamline the workflow. The pipeline consists of the preprocessor and a Decision Tree classifier. This encapsulates the entire process from feature transformation to model training.

Model Evaluation:
The model's performance is assessed using cross-validation with a 5-fold strategy, utilizing the negative mean absolute error (MAE) as the evaluation metric. The MAE is calculated by negating the mean absolute error scores obtained during cross-validation.

Results:
The model achieved a mean absolute error of approximately 0.244. The MAE of 0.244 indicates the average absolute difference between the predicted and actual quantity values. Lower values of MAE suggest better accuracy in predicting the quantity of items.

Conclusion:
Model 4, utilizing a Decision Tree classifier, demonstrates promising performance in predicting the quantity of items. The incorporation of appropriate preprocessing steps ensures that the

model is well-equipped to handle both numeric and categorical features. The mean absolute error of 0.244 suggests that, on average, the model's predictions are close to the actual quantity values. Further analysis, hyperparameter tuning, and exploration of alternative algorithms could be considered to potentially enhance the model's predictive capabilities. This report provides insights into the methodology, preprocessing, and evaluation of Model 4, offering a foundation for iterative improvements in future modeling efforts.

CONCLUSION:

Models 1 and 4 showcase promising predictive capabilities.

Model 1, employing Linear Regression, achieves high accuracy in predicting total sales, as indicated by the substantial R-squared value. The inclusion of relevant features, effective preprocessing, and the algorithm's choice contribute to its success. Future enhancements may involve feature refinement, hyperparameter tuning, and model validation.

Model 4, utilizing a Decision Tree classifier, performs well in predicting item quantity. The mean absolute error suggests close alignment between predicted and actual values. Appropriate preprocessing accommodates diverse features. Further exploration, hyperparameter tuning, and alternative algorithm investigation could enhance predictive capabilities.

In conclusion, both models demonstrate effectiveness with their respective tasks. Future work may focus on iterative improvements, including refining features and exploring alternative algorithms for enhanced predictive accuracy.