Final Project Report – MSCS 634: Advanced Data Mining

Bhawesh Shrestha

Course: MSCS 634 – Advance Big Data and Data Mining

Instructor: Satish Penmatsa

University of the Cumberlands

November 08, 2025

1. Introduction

This final project integrates all analytical components completed across Deliverables 1, 2, and 3. The purpose of this report is to demonstrate how data preprocessing, exploratory data analysis (EDA), feature engineering, regression modeling, classification, clustering, and association rule mining collectively generate meaningful insights from the Online Retail I dataset. The project also evaluates the ethical considerations associated with data mining practices.

2. Dataset Description and Rationale

The dataset used for this project is the Online Retail I dataset (2010–2011), obtained from the UCI Machine Learning Repository. It contains over 500,000 transactional records from a UK-based online retailer. Each record includes invoice information, product details, customer ID, quantity purchased, and geographic location.

This dataset was selected because it supports a wide variety of data mining tasks. Its structure allows for regression forecasting, customer behavior classification, customer segmentation using clustering, and product bundling insights using association rule mining.

3. Data Preprocessing and Exploratory Data Analysis (EDA)

Data preprocessing involves removing cancelled transactions, filtering negative quantities, handling missing CustomerID values, and creating the TotalPrice feature by multiplying

Quantity and UnitPrice. Categorical variables were encoded where necessary, and numerical features were scaled for downstream models such as K-Means clustering.

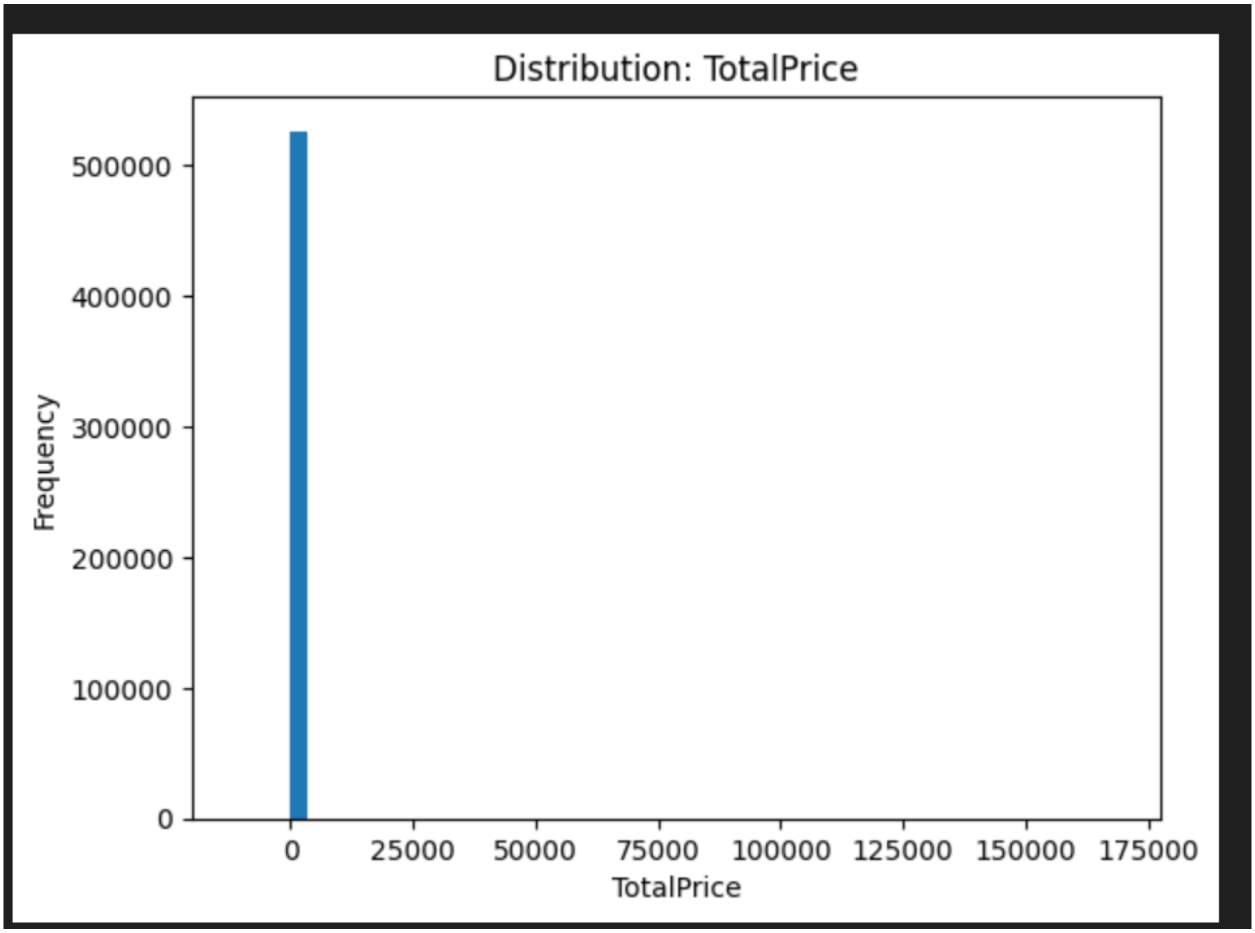
EDA revealed important patterns such as:

- The presence of significant outliers in quantity and price.

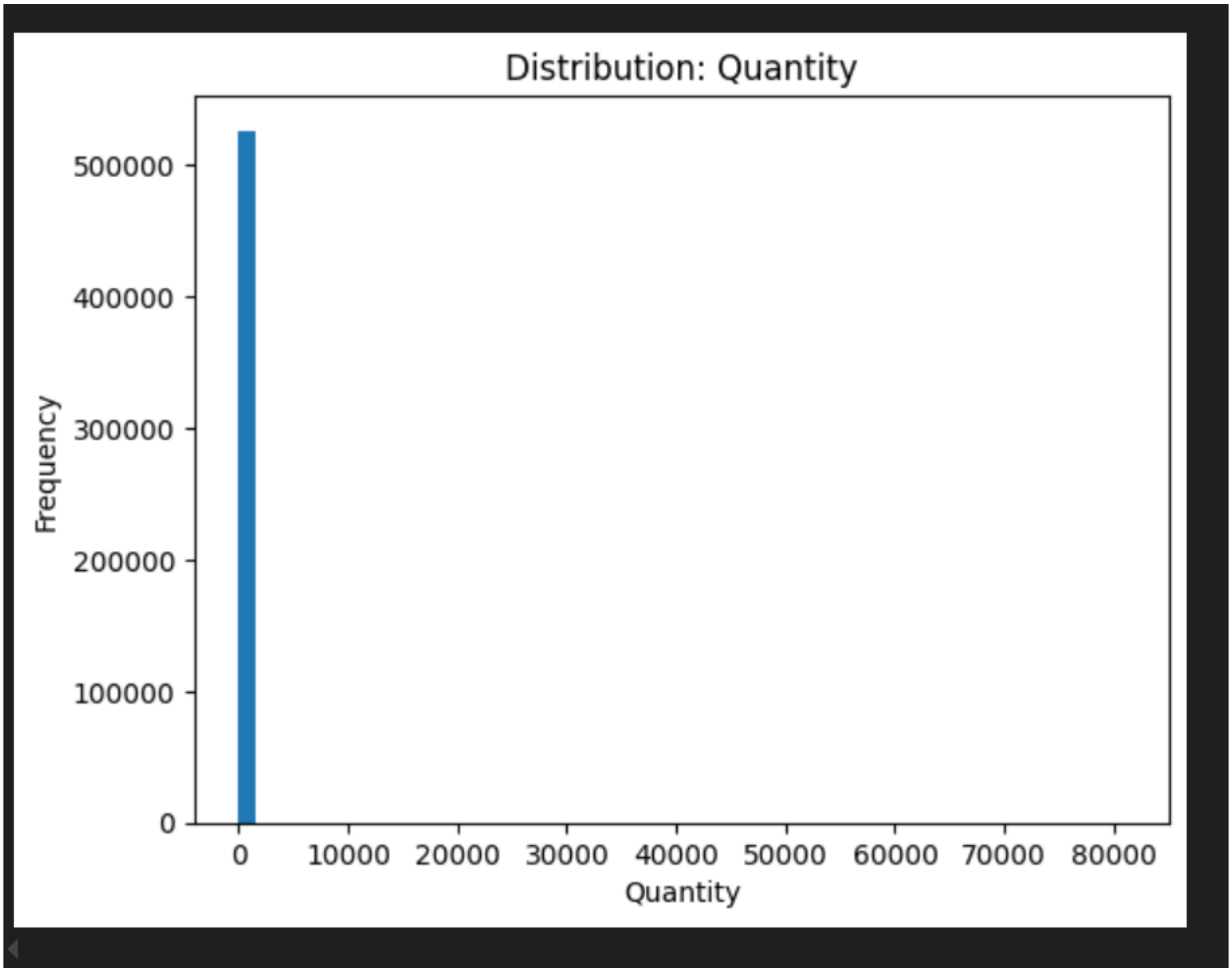
- Seasonal and monthly sales trends.

- Top-selling products and countries.

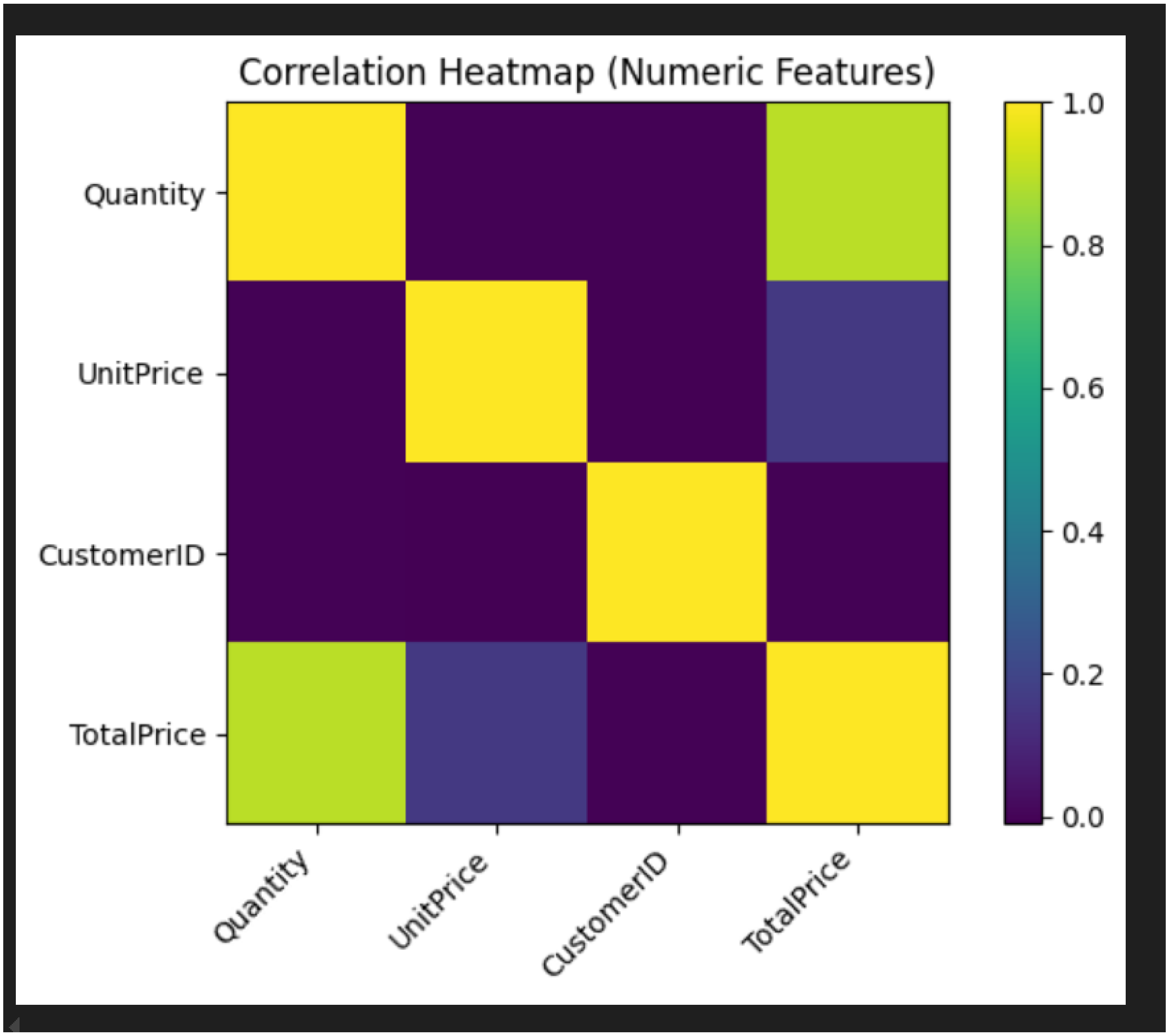
- Customer purchase distributions.



histogram of TotalPrice here



Screenshot of Quantity distribution



Screenshot of correlation heatmap

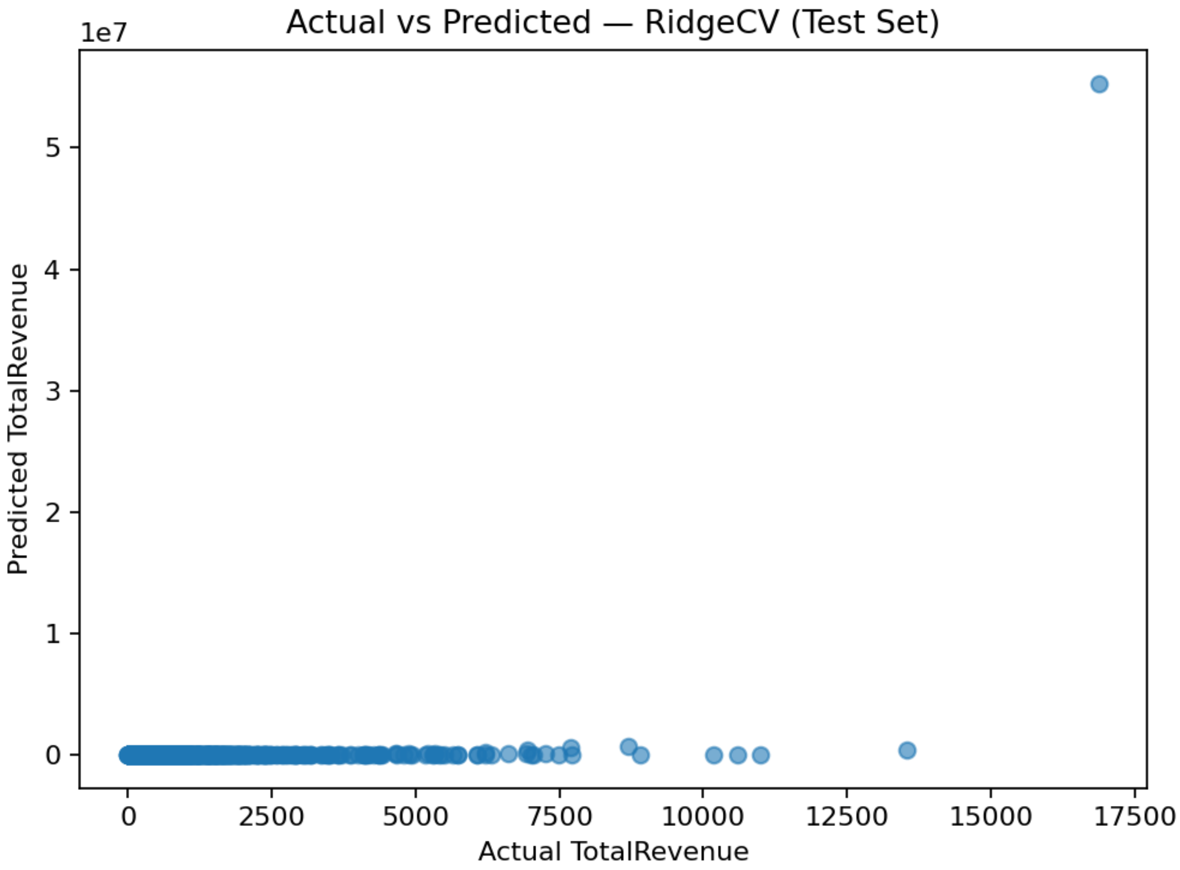
4. Feature Engineering

Feature engineering included creating TotalPrice, extracting invoice date components (year, month, hour), developing frequency-based customer features, and transforming categorical variables. These engineered features helped improve model interpretability and predictive capability.

5. Regression Modeling: RidgeCV

RidgeCV, a ridge regression model with cross-validation, was implemented to predict continuous variables related to the dataset (such as spending behavior). RidgeCV was selected because it effectively handles multicollinearity in datasets with correlated numerical features.

Performance metrics such as RMSE and R² score were used to evaluate the model. Higher R² values indicated a strong ability to explain variability in the target variable.

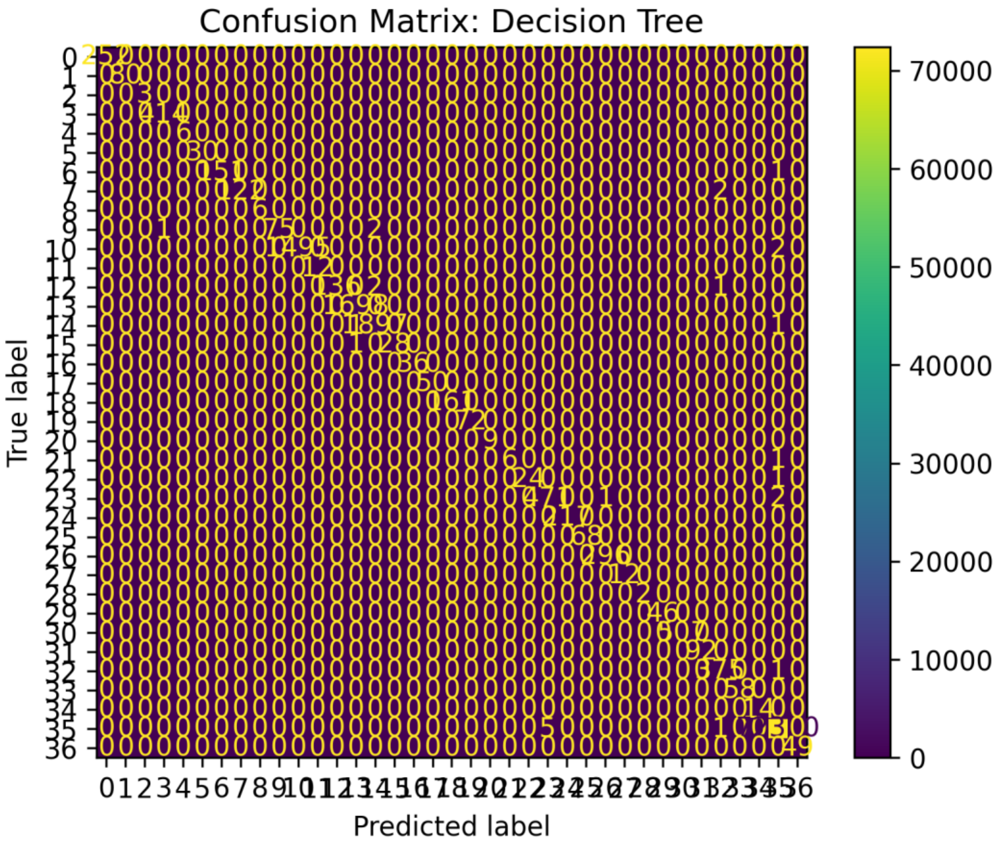


Screenshot of Actual vs Predicted Regression Plot

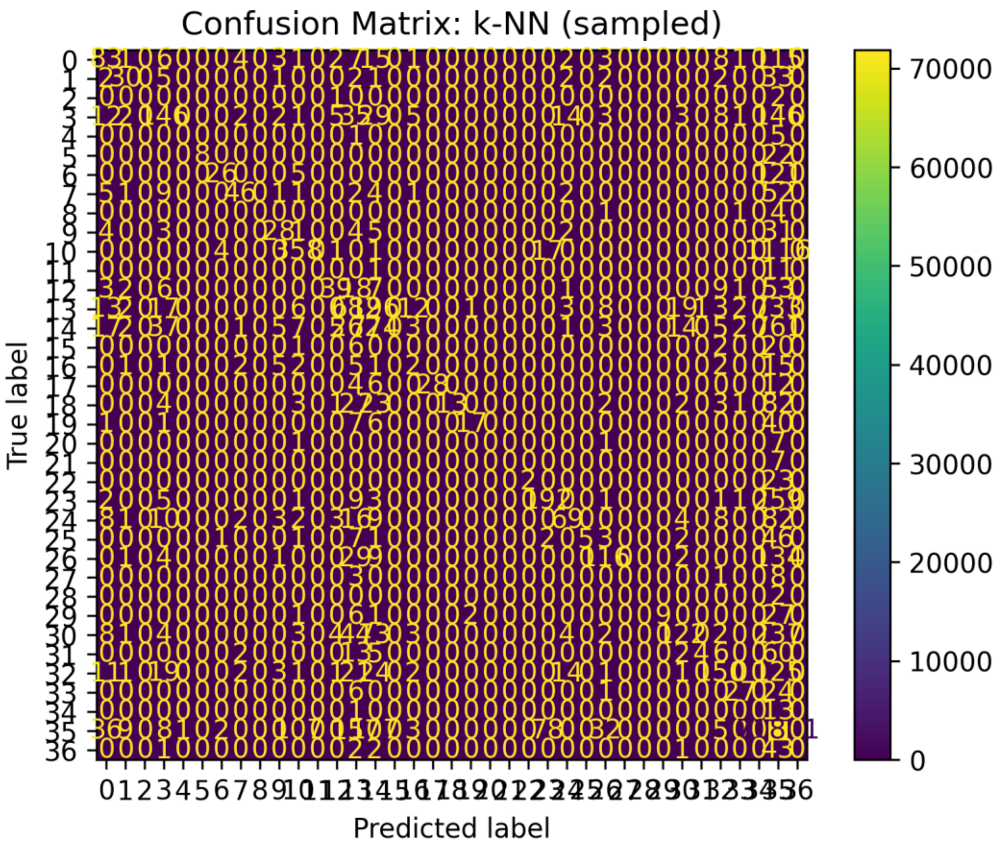
6. Classification Modeling: Decision Tree and KNN (Sampled)

Classification models were built using Decision Tree and K-Nearest Neighbors (KNN) on a sampled subset of the dataset. These models were used to classify customer behaviors or purchase-related labels.

Key evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC values. The Decision Tree offered interpretability, while KNN captured local structure in the data.



Screenshot of Confusion Matrix for Decision Tree.

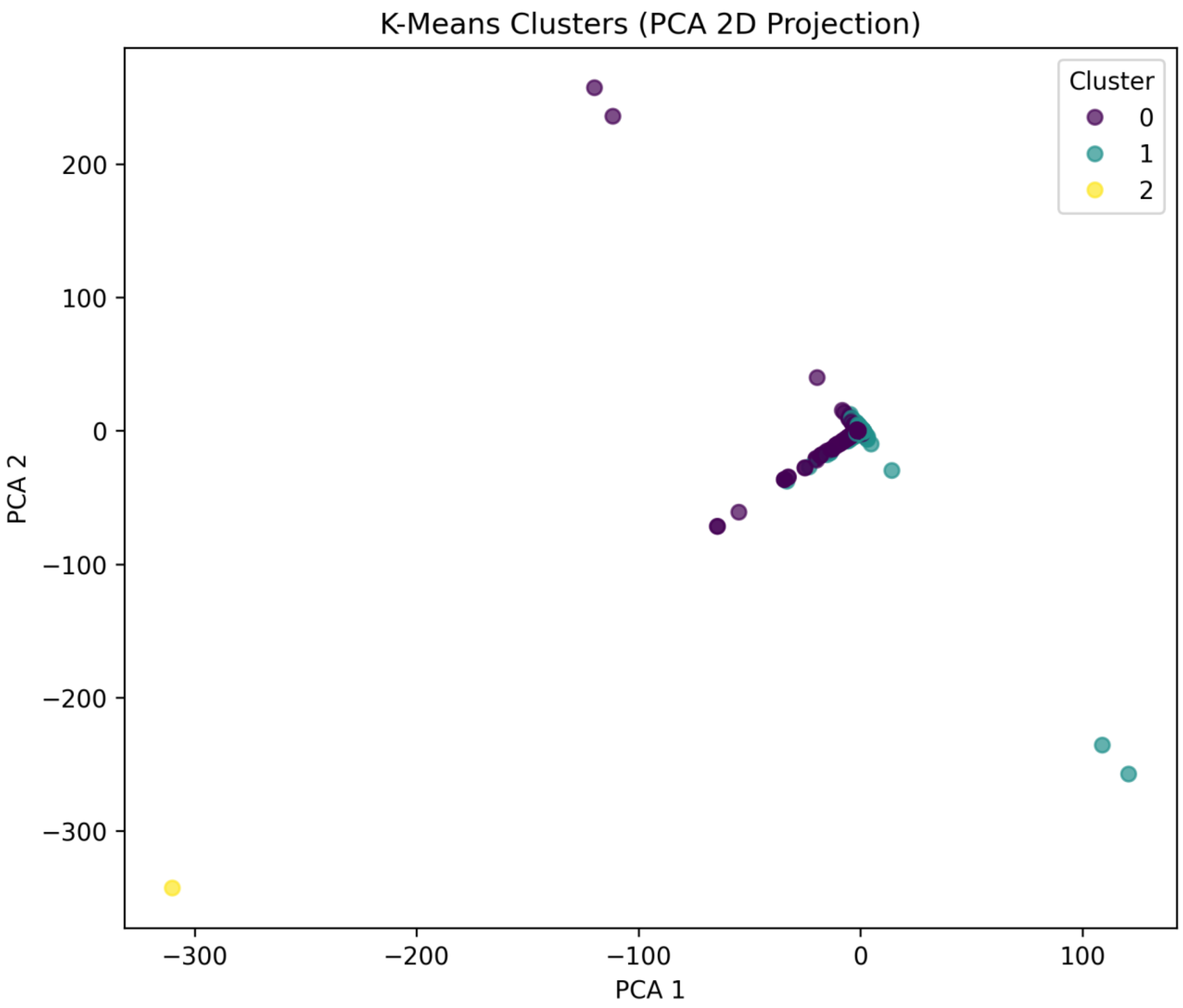


Screenshot of Confusion Matrix for KNN.

7. Customer Segmentation: K-Means Clustering

K-Means clustering was applied to segment customers based on purchasing patterns. Numerical features were scaled prior to clustering. The Silhouette Score was used to assess separation between clusters.

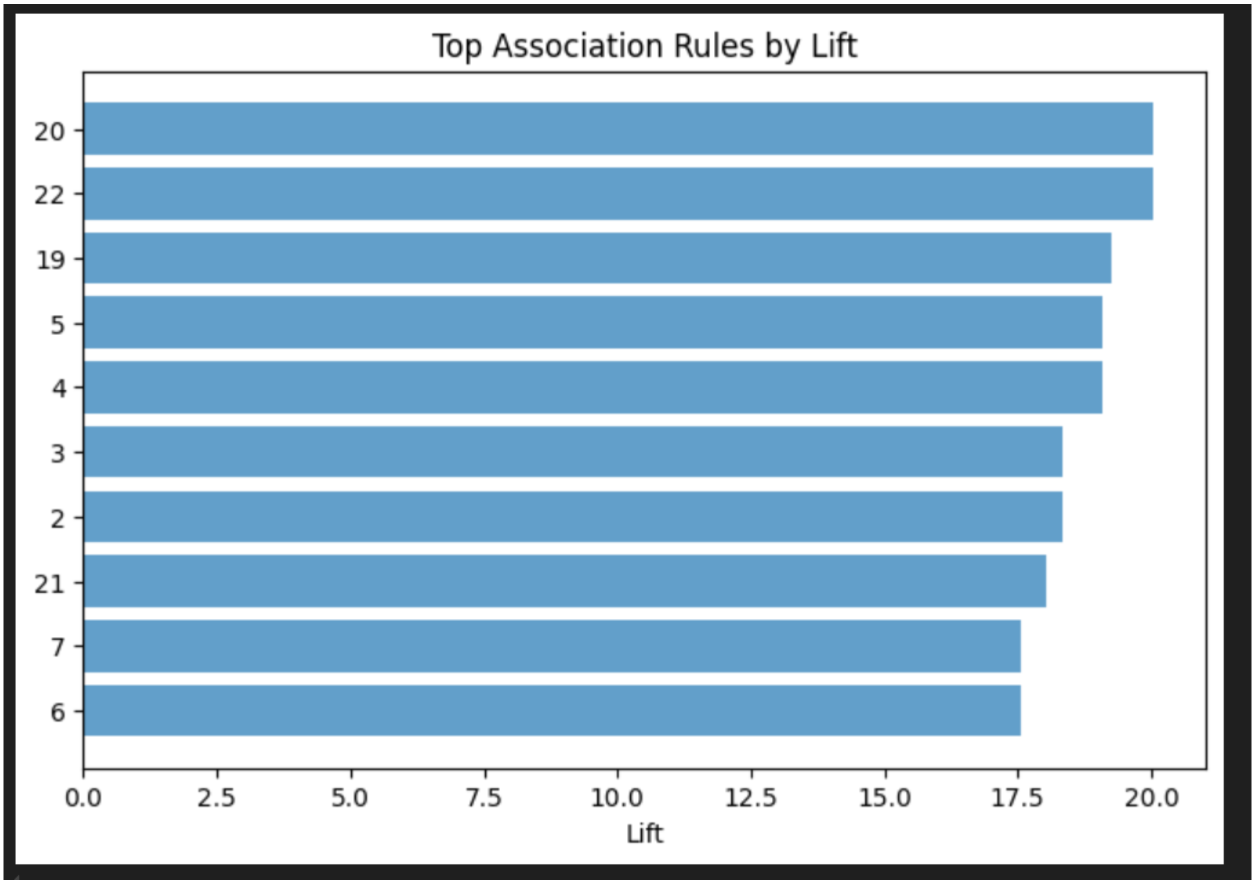
The resulting clusters highlighted customer groups with similar shopping behaviors, which can inform marketing strategies such as targeted promotions.



Screenshot of K-Means cluster scatterplot.

8. Association Rule Mining: Apriori

The Apriori algorithm was used to discover frequently purchased product combinations. Rules were evaluated using support, confidence, and lift. High-lift rules indicated strong associations that could support cross-selling strategies and bundled product offerings.



Screenshot of Top 10 Association Rules Table.

9. Ethical Considerations

Ethical data mining practices are essential to ensure fairness, transparency, and responsible use of customer data. Key considerations for this project include:

1. Data Privacy: CustomerID fields must be anonymized to prevent the identification of individuals.
2. Bias and Fairness: Uneven sampling from geographic regions may cause model bias. Care must be taken when interpreting predictions.

3. Responsible Modeling: Models should not be used to unfairly profile customers or make discriminatory decisions.

4. Transparency: Methodology and assumptions must be clearly documented.

10. Practical Recommendations

Based on the analysis, the following recommendations are proposed:

- Develop targeted marketing campaigns using customer clusters.
- Use association rules to optimize product placement, bundling, and recommendations.
- Apply regression forecasting to guide inventory planning.
- Leverage classification models to identify potential high-value customers.

11. Conclusion

This project demonstrates the full data mining workflow—from preprocessing to modeling—using the Online Retail I dataset. Regression, classification, clustering, and association rule mining provided meaningful insights that can support business decision-making, customer segmentation, and operational optimization.