Wholesale Customer Purchase Behavior

Author: Sreenivasulu Somu
Bellevue University – DSC550 Data Mining
Professor Brett Werner

Introduction

Objective

The project uses wholesale customer dataset, to understand customer purchasing behavior and segment customers based on their buying patterns. The problem at hand is to identify distinct customer groups and predict the sales channel for customers based on their purchase history.

By segmenting customers based on their spending patterns and predicting the sales channel, businesses can tailor marketing strategies, allocate inventory more effectively, and enhance customer retention. Understanding which products drive customer behavior supports more efficient resource allocation and revenue growth.

Introduction and Problem Statement

Understanding customer segments and predicting sales channels is important for optimizing marketing strategies, inventory management, and overall business operations. This analysis can lead to more targeted marketing campaigns, improved customer service, and better resource allocation. To gain stakeholder buy-in, we can emphasize the potential for increased revenue through personalized marketing, reduced costs through optimized inventory management, and improved customer satisfaction through tailored services.

Data Source

The data used in this analysis comes from a CSV file named "Wholesale customers data.csv", which contains information about customer purchases across various product categories.

Summary of Milestone 1-3

Data Origin
The analysis uses the “Wholesale customers data.csv” dataset that includes spending amounts for six key product categories (Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen) as well as categorical attributes like Channel and Region. The dataset is typical of public datasets used in analytics courses and initial exploratory projects.

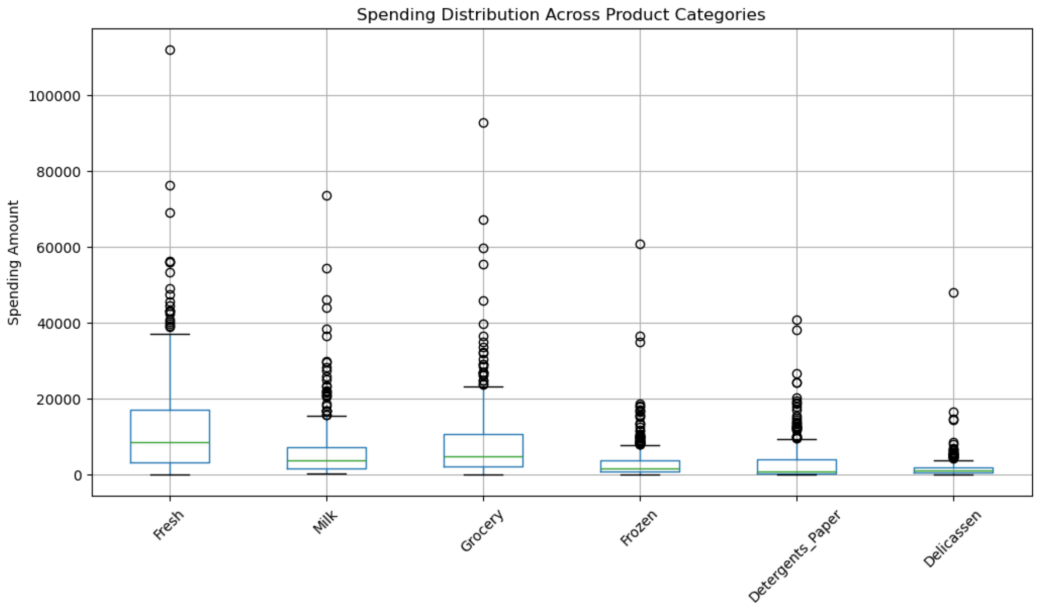
Jupyter Notebook Overview
The Jupyter Notebook in the previous milestones provide an end-to-end project workflow, starting with data ingestion, visualization, and EDA, followed by data transformation, feature engineering, and finally,

model building using both unsupervised (K-Means clustering) and supervised (Random Forest classification) methods. Every step is accompanied by visual outputs and code commentary that detail the reasoning behind each transformation and modeling decision.

Exploratory Data Analysis (EDA)

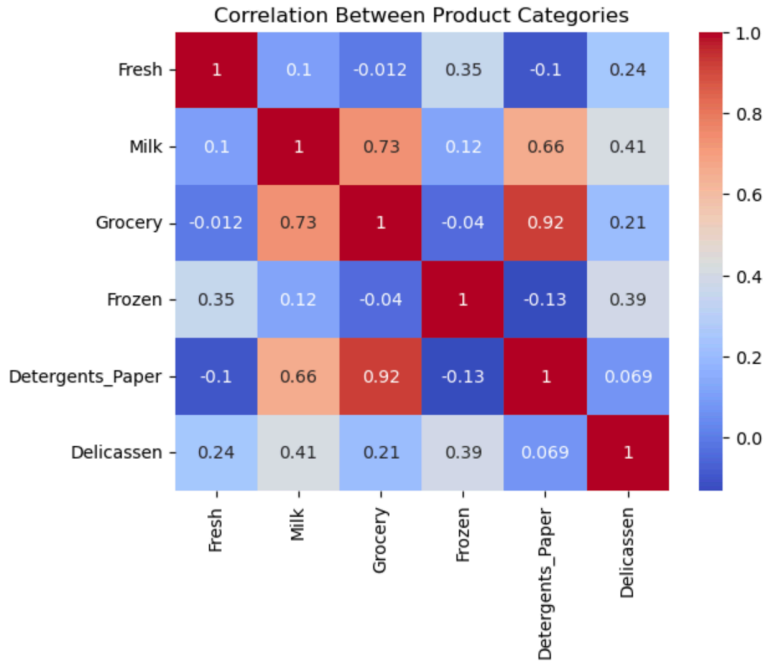
Visualizations and Their Insights:

- **Box Plots**



A box plot for spending across the product categories shows the distribution and spread of values. This helps in quickly identifying outliers and understanding the variability in each product category's sales1.

- **Correlation Heatmap**

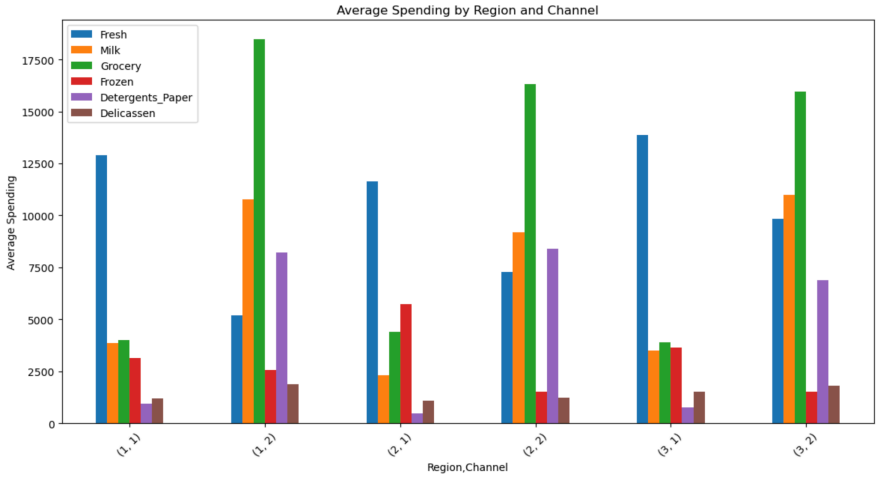


The correlation heatmap shown above is for the six spending categories, uses a "coolwarm" palette. It clearly illustrates how different product categories relate to each other—revealing, for instance, which items tend to have similar spending patterns, which is critical for dimensionality reduction or feature selection strategies.

- **PCA Scatter Plot**

After reducing the multi-dimensional spending data to two principal components, a scatter plot colored by K-Means clusters reveals three distinct clusters. This visualization provides a visual confirmation that natural customer segments exist and each cluster representing a group with similar purchasing behaviors.

- **Grouped Bar Chart**



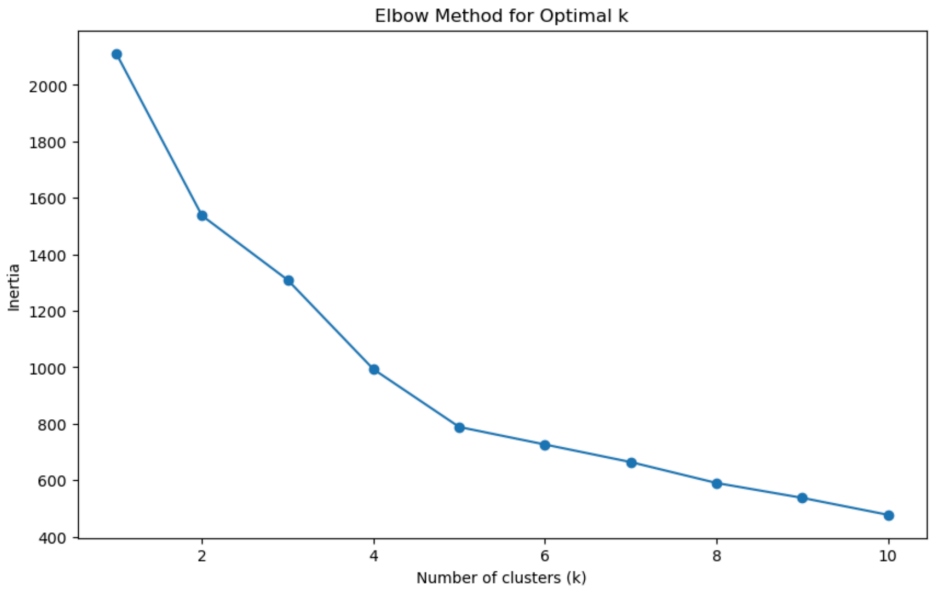
A bar chart above shows an average spending by Region and Channel offers insight into regional or channel-specific behaviors and tendencies, aiding the understanding of market segmentation further.

Data Preparation

- **Cleaning and Transformation**
  - The notebook begins with a cursory look at the dataset, displaying a few rows and checking for missing values. No missing data was found.
  - To handle skewed distributions in several spending columns, a log transformation (via np.log1p) is applied, improving the normality of distributions for the later modeling tasks.
  - Two new features are created:
    - **Total_Spending:** The sum of spending across all product categories.
    - **Milk_Grocery_Ratio:** A ratio of spending on Milk to Grocery, offering insight into purchasing patterns.
  - The categorical features (Channel and Region) are converted into dummy variables to prevent multicollinearity. Finally, the numerical features are scaled using a StandardScaler to ensure that all variables contribute equally during model building.

Model Building and Evaluation

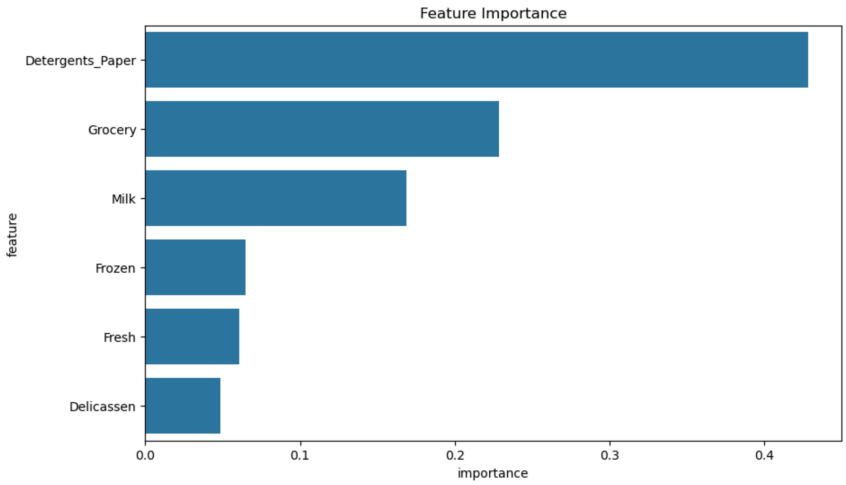
- **K-Means Clustering**
  - **Clustering Process**



After reducing dimensionality with PCA, the model applies K-Means clustering. The elbow method (plotting inertias for k values from 1 to 10) helps decide an optimal k, which is identified as 3.

- **Evaluation**
  The silhouette score of approximately 0.54 indicates that the clusters have reasonable cohesion and separation, thereby supporting the segmentation result.

- **Random Forest Classification**
  - **Model Setup**
    Using all purchasing features (after dropping raw categorical columns) as predictors and Channel as the target, a Random Forest Classifier is trained.
  - **Performance Metrics**
    The model achieves a high accuracy of about 91% on the test set. The classification report further details that many classes are predicted with high precision and recall.
  - **Feature Importance**



A feature importance bar plot is generated, which ranks the spending categories

showing which ones most influence the model's decision regarding customer channel. This provides insight into product-specific impacts on channel determination.

Conclusions

Interpretation of Analysis and Model Readiness.

- The clustering confirms that there are distinct customer segments based on purchasing behavior, while the Random Forest model accurately predicts the sales channel from these patterns. In essence, spending patterns in product categories are strong indicators of overall customer behavior and channel preferences.
- Although the Random Forest model shows strong performance metrics, it is based on historical data from a limited dataset. Deployment would require additional validation—such as cross-validation, testing on unseen data, and sensitivity analyses—to ensure robustness in a production environment.

Recommendations and Future Opportunities

Recommendations

- **Targeted Marketing:** Use segmentation insights to craft customized promotions and campaigns.
- **Inventory and Product Strategy:** Leverage feature importance insights to focus on high-impact product categories tailored for respective customer segments.
- **Further Validation:** Conduct additional tests and hyperparameter tuning on the Random Forest model before considering deployment.

Challenges and Future Opportunities

- **Scalability:** Ensure model performance remains robust as more customer data becomes available over time.
- **Seasonality and External Factors:** Incorporate time-series elements or external market data to capture seasonal trends or regional variations.
- **Integration with Other Data:** Consider merging demographic or geo-spatial data to further refine customer segmentation and enhance predictive capabilities.
- **Overfitting Concerns:** Monitor for overfitting given the dataset's size; regularization techniques and continuous model evaluation can help mitigate this risk.