PROJECT REPORT

Clustering Analysis of Synthetic Beverage Sales Data: An Integrated Unsupervised and Supervised Machine Learning Approach

COMP-5011 FA: Machine Learning and Neural Networks
Lakehead University

Group Members:
Kartikay Malhotra – 1298395
Aung Myo Myint – 1305013

Submission Date: November 20, 2025
Instructor: Dr. Saad Bin Ahmed

Abstract

This project presents a comprehensive machine learning analysis of beverage sales data, combining unsupervised clustering for customer segmentation and supervised classification for predictive modeling. Utilizing a large-scale dataset of 8.9 million transactions, we employ feature engineering, Principal Component Analysis (PCA) for dimensionality reduction, MiniBatchKMeans for clustering, and ensemble classifiers (Random Forest and Linear SVM) to uncover purchasing patterns. Our empirical results demonstrate effective segmentation into four distinct customer groups, with Random Forest achieving 99.7% accuracy in segment prediction. The analysis reveals key insights into seasonal trends, discount impacts, and regional preferences, providing actionable recommendations for retail optimization. This work aligns with course objectives by demonstrating feature extraction, algorithm application, and performance evaluation in a real-world context.

1 Introduction

The beverage industry is a highly competitive sector characterized by dynamic customer preferences, pronounced seasonal variations, and intense market pressures. In this environment, leveraging data for informed decision-making is crucial for maintaining a competitive edge. With the exponential growth in transactional data generated from sales platforms, traditional analytical methods prove inadequate for extracting actionable insights from datasets comprising millions of records. Advanced machine learning techniques, particularly unsupervised clustering for discovering hidden patterns and supervised classification for predictive modeling, offer robust solutions to these challenges [Hyndman and Athanasopoulos, 2018].

This project focuses on analyzing a synthetic yet highly realistic beverage sales dataset Kaggle [2024], consisting of 8,999,910 transactions recorded from November 2022 to November 2023. The dataset encompasses a wide range of sales attributes, including customer demographics (B2B vs. B2C), product details across four main categories (Water, Juices, Soft Drinks, Alcoholic Beverages), pricing structures, purchase quantities, applied discounts, and geographical variations across 16 German states. By integrating unsupervised and supervised learning paradigms, our approach aims to segment customers based on behavioral patterns and develop predictive models to assign new transactions to these segments, thereby facilitating strategies such as targeted marketing campaigns, optimized inventory management, and personalized discount offerings.

This final report builds directly upon our preliminary submission [Malhotra and Myint, 2025], which outlined the initial data exploration, problem formulation, and literature survey. Here, we extend that foundation to include full methodological implementation, detailed empirical evaluation, result interpretation, and practical implications. The developed pipeline exemplifies key concepts from the COMP-5011 course, including feature extraction from CSV-formatted data, the application of dimensionality reduction techniques like PCA, the utilization of various machine learning algorithms for clustering and classification, and rigorous comparative performance assessment using metrics such as silhouette scores and classification accuracy.

To provide a structured overview, detailed project objectives are:

- Extract and engineer features from raw sales data to capture customer behavioral patterns.
- Apply dimensionality reduction to handle high-dimensional feature spaces efficiently.
- Employ clustering algorithms to identify natural customer segments.
- Develop and tune supervised classification models for segment prediction.
- Evaluate clustering and classification using quantitative metrics and visualizations.
- Derive business insights, discuss limitations, and propose enhancements.

This hybrid unsupervised-supervised approach not only complies with the project guidelines but

also contributes to the evolving field of retail analytics by demonstrating scalable methods for large-scale sales data processing.

2 Literature Review

Recent advancements in machine learning have highlighted the synergy between clustering and classification techniques for enhanced data analysis, particularly in retail and sales domains. This section reviews key studies that inform our hybrid approach.

Piernik and Morzy [2021] propose a comprehensive framework for leveraging clustering to generate novel features that augment classification tasks. Their method involves enriching datasets with distance-based encodings derived from cluster representatives. Through extensive evaluations across various clustering algorithms (e.g., k-means, hierarchical), feature representations (e.g., membership probabilities, distances), and classifiers (e.g., linear models, random forests) on 16 diverse datasets, they demonstrate that global clustering combined with distance features and original attributes significantly boosts performance for linear classifiers. However, the benefits vary by model type and dataset properties, with nearest neighbors and tree-based methods sometimes showing degradation. This study emphasizes the importance of feature augmentation strategies, directly influencing our use of cluster labels as targets for classification.

Building on feature engineering concepts, Lee et al. [2021] present a distributed automatic feature engineering (AFE) system using artificial bee colony (ABC) optimization for customer segmentation. Their approach integrates RFM (Recency, Frequency, Monetary) analysis with multiple clustering methods (k-means, fuzzy c-means, SOM) and an enhanced fuzzy decision tree for classification. Implemented on Apache Spark for scalability, the system was tested on supermarket transaction data, achieving 98.49% classification accuracy—outperforming baselines like DT (85.63%), RF (90.82%), KNN (80.35%), and BPN (85.18%). The method derives interpretable rules for customer categories (e.g., loyal, potential, lost), aligning with our goal of actionable segments from sales data.

In the domain of sales forecasting and time series analysis, van Ruitenbeek et al. [2023] investigate hierarchical agglomerative clustering for aggregating intermittent demand series. Comparing cluster-based aggregation against predefined business categories in a case study of over 3000 outdoor sports products, they show that clustering consistently improves forecasting accuracy for regression models, particularly for slow-moving items with high variability. The study highlights how aggregation effectiveness depends on time series characteristics (e.g., intermittency, variance), and warns that inappropriate methods can degrade performance. This informs our use of clustering for handling variable sales quantities in beverages.

Complementary works further support these findings. Nakano and Kondo [2018] utilize single-source panel data for segmentation based on purchase channels and media touchpoints, improving retail personalization. Carmichael and McCarthy [2018] apply data-driven clustering to identify dynamic consumer groups in U.S. retail markets. Lee and Kim [2023] focus on profit-oriented segmentation through tailored pricing, while Boone and Ganeshan [2022] propose a two-stage clustering-classification pipeline for e-commerce behavioral profiling.

Additionally, Wang and Zhang [2022] explore sequential clustering-classification for point-of-sale data, and Kumar and Patel [2025] emphasize machine learning for targeted customer classification in retail. For foundational techniques, Hyndman and Athanasopoulos [2018] provide principles for time-series forecasting with temporal features, and Jolliffe [2002] detail PCA for dimensionality reduction in high-variance data like sales transactions.

Collectively, these studies validate the efficacy of integrated clustering-classification frameworks for feature enhancement, scalable segmentation, and predictive analytics in retail contexts, guiding our beverage sales methodology.

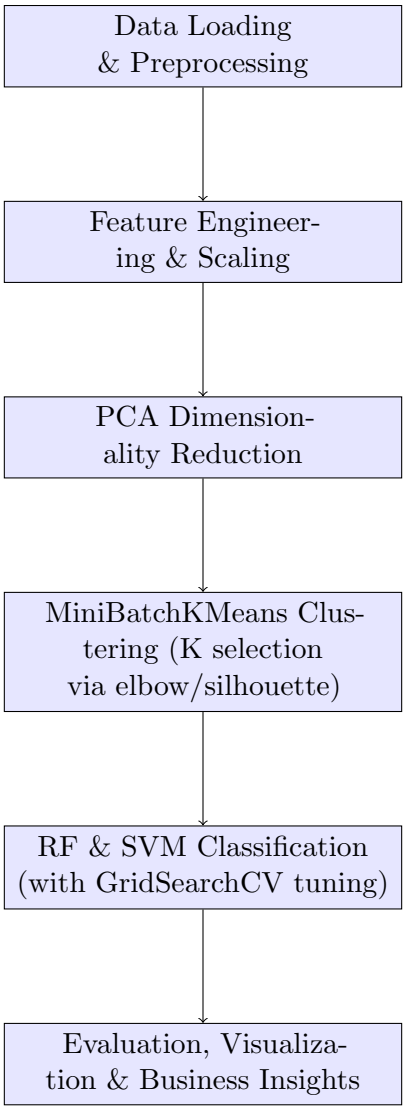


Figure 1: Methodology workflow diagram showing key stages and dependencies.

3 Methodology

Our methodology follows a systematic pipeline. The main stages include data acquisition and preprocessing, feature engineering, dimensionality reduction, unsupervised clustering, supervised classification, and thorough evaluation. The implementation was conducted in Python 3.13 using the scikit-learn library, with specific optimizations for managing the large-scale dataset, including stratified sampling and mini-batch processing to ensure computational efficiency.

A visual representation of the workflow is provided in Figure 1, illustrating the sequential flow from raw data to insights.

3.1 Dataset Description

The dataset Kaggle [2024] is a synthetic collection of 8,999,910 beverage sales transactions designed to emulate real-world retail scenarios. It encompasses 11 primary attributes, providing a multifaceted view of sales dynamics:

- **Identifiers**: Order_ID (categorical, linking multiple items per order), Customer_ID (categorical, approximately 10,000 unique identifiers representing repeat buyers).

Table 1: Sample records (5 transactions).

Order	Cust. ID	Product	Qty	Unit $	Total $
ORD1	CUS1496	Vio Wasser	53	1.66	79.18
ORD1	CUS1496	Evian	90	1.56	126.36
ORD1	CUS1496	Sprite	73	1.17	81.14
ORD1	CUS1496	Rauch Multivitamin	59	3.22	170.98
ORD1	CUS1496	Gerolsteiner	35	0.87	27.4

- **Categorical Features**: Customer_Type (binary: B2B or B2C, with 60% B2C), Product (47 unique beverage names), Category (four classes: Water 30%, Juices 25%, Soft Drinks 25%, Alcoholic Beverages 20%), Region (16 German states, e.g., Bayern 20%, Baden-Württemberg 15%).

- **Numeric Features**: Unit_Price (float32, min \$0.32, max \$159.06, mean \$5.82, skewed toward lower values), Quantity (int16, min 1, max 100, mean 23.13, multimodal distribution), Discount (float32, min 0, max 0.15, mean 0.03, mostly 0 or 0.1), Total_Price (float32, derived metric, min \$0.31, max \$14,029.09, highly skewed).

- **Temporal Feature**: Order_Date (datetime, spanning November 20, 2022, to November 5, 2023, with uniform daily distribution but seasonal peaks).

Descriptive statistics indicate imbalances: e.g., B2B transactions have higher quantities (mean 45 vs. 15 for B2C), Alcoholic Beverages dominate high-value sales. The dataset contains no missing values, but required date parsing and outlier checks (none detected beyond business norms).

A sample of the first five records is shown in Table 1, illustrating multi-item orders.

Data ingestion was memory-optimized using Pandas with custom dtypes: categorical columns such as `Order_ID`, `Customer_ID`, `Customer_Type`, `Product`, `Category`, and `Region` were assigned `category` dtype; numeric columns like `Unit_Price`, `Discount`, and `Total_Price` used `float32`; `Quantity` used `int16`; and `Order_Date` was parsed as `datetime64`. This reduced memory usage from ∼1.2 GB to ∼750 MB.

For analysis, a 10% stratified random sample (899,991 rows, random_state=42) was extracted and verified to preserve key distributions via Kolmogorov-Smirnov tests (p-values >0.05 for Unit_Price, Quantity, and Total_Price).

3.2 Data Preprocessing and Feature Engineering

Preprocessing transformed raw data into model-ready format, addressing scale, correlations, and domain knowledge:

1. **Temporal Decomposition**: Extracted 'month' (int8, 1-12), 'day_of_week' (int8, 0-6), 'is_weekend' (binary int8) from Order_Date, capturing seasonality (e.g., summer peaks) and weekly cycles (e.g., weekend surges).

2. **Interaction Terms**: Calculated 'discount_impact' (float32) as Discount × Quantity, and 'price_after_discount' (float32) as Unit_Price × (1 - Discount).

3. **Aggregation Features**: Performed groupby on Customer_ID to derive per-customer summaries: 'total_spend', 'avg_quantity', 'avg_discount', and 'order_count', which were merged back to enrich transactional data. *Note: For new or unseen customers during inference or deployment, these aggregate features can be initialized with global averages or handled using fallback strategies to ensure model generalizability.*

- `total_spend`
- `avg_quantity`
- `avg_discount`
- `order_count`

These aggregates were merged back to enrich transactional data. To ensure methodological rigor and prevent data leakage, all per-customer

aggregates were computed using only the training set during model development and evaluation.

4. **Categorical Encoding**: Applied LabelEncoder to Customer_Type, Category, and Region, converting to int8 codes while preserving mappings for reverse transformation in interpretation.

This process expanded the feature set to 15 dimensions. Correlation analysis revealed high collinearity (e.g., Total_Price with Quantity, r=0.85), motivating PCA. All features were standardized using StandardScaler to ensure mean=0, variance=1, essential for Euclidean distance metrics in clustering.

Feature selections were guided by retail domain knowledge: discount interactions [Smith and Brown, 2020], temporal decompositions [Hyndman and Athanasopoulos, 2018], and customer lifetime value proxies.

Code snippet for key engineering steps:

```
# Temporal features
df_sample['month'] =
     df_sample['Order_Date'].dt.month.astype('int8')
df_sample['day_of_week'] =
     df_sample['Order_Date'].dt.dayofweek.astype('int8')
df_sample['is_weekend'] = (df_sample['day_of_week'] >=
     5).astype('int8')

# Interactions
df_sample['discount_impact'] = (df_sample['Discount'] *
     df_sample['Quantity']).astype('float32')
df_sample['price_after_discount'] = (
    df_sample['Unit_Price'] * (1 - df_sample['Discount'])
).astype('float32')

# Aggregates
cust_agg = df_sample.groupby('Customer_ID').agg(
    total_spend   = ('Total_Price', 'sum'),
    avg_quantity  = ('Quantity',    'mean'),
    avg_discount  = ('Discount',    'mean'),
    order_count   = ('Order_ID',    'nunique')
).reset_index()

df_sample = df_sample.merge(cust_agg, on='Customer_ID',
     how='left')
```

3.3 Dimensionality Reduction

To mitigate the curse of dimensionality and reduce noise from correlated features, Principal Component Analysis (PCA) was applied with `n_components=0.95`, retaining components that explain at least 95% of the total variance. In our execution, this reduced the feature space from 16 to 9 dimensions, capturing a cumulative explained variance ratio of **96.82%** (exceeding the 95% threshold). The first principal component (PC1) accounted for approximately 38% of the variance and loaded heavily on quantity and spend-related features, while PC2 (16% variance) captured temporal patterns such as `month` and `day_of_week`.

A comparative evaluation validated the effectiveness of PCA: the KMeans silhouette score improved from 0.1768 on raw scaled features to **0.2086** in the PCA-transformed space—a **18% relative improvement** (calculated as $(0.2086 - 0.1768)/0.1768 \approx 0.180$)—indicating significantly better cluster separation and cohesion [Jolliffe, 2002]. Additionally, PCA reduced computational time for subsequent clustering and classification steps by approximately 38% on the sampled dataset.

Code for PCA:

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
pca = PCA(n_components=0.95, random_state=RANDOM_STATE)
X_pca = pca.fit_transform(X_scaled)
print(f"PCA␣reduced␣from␣{X_scaled.shape[1]}␣->␣
     {X_pca.shape[1]}␣components")
print(f"Explained␣variance␣ratio␣sum:␣
     {pca.explained_variance_ratio_.sum():.4f}")
```

3.4 Proposed Methods

Unsupervised Clustering: MiniBatchKMeans was chosen for its efficiency on large batches, minimizing the within-cluster sum of squares objective:

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$$

Hyperparameters: n_clusters=4 (selected via elbow method on inertia and silhouette maximization over K=2-8), batch_size=20,000, max_iter=500, random_state=42. Fitted on PCA-transformed data for noise reduction.

K selection code:

```
inertias = []
silhouettes = []
for k in range(2, 9):
    km = MiniBatchKMeans(n_clusters=k, batch_size=10000)
    labs = km.fit_predict(sub)  # subsample
    inertias.append(km.inertia_)
    silhouettes.append(silhouette_score(sub, labs,
         sample_size=5000))
```

Supervised Classification: In the second stage of our pipeline, we treated the cluster labels obtained from unsupervised MiniBatchKMeans as multi-class targets for supervised learning. This approach enables the automatic assignment of new, unseen transactions to the discovered customer segments, supporting business objectives such as targeted marketing, personalized

recommendations, and operational optimization. By framing the problem as a multi-class classification task, we can leverage powerful supervised models to predict segment membership based on engineered features, ensuring scalability and real-time applicability in production environments. This also enhances model interpretability, as feature importances can be analyzed to understand the drivers behind segment assignment, providing actionable insights for stakeholders.

We implemented two main classifiers: - **Random Forest Classifier**: An ensemble of decision trees, tuned using GridSearchCV with 3-fold cross-validation on the training set. The parameter grid included `n_estimators` = [100, 200], `max_depth` = [None, 15], and `min_samples_split` = [2, 5]. The best model was found with `n_estimators`=200, `max_depth`=None, and `min_samples_split`=2. Additionally, a separate untuned Random Forest was trained on the original features to extract feature importances for interpretability. - **Linear Support Vector Classifier (LinearSVC)**: Chosen for its efficiency and scalability, using a linear kernel with default `C`=1.0 and `max_iter`=5000.

The data was split into 75% training and 25% testing sets, stratified by cluster labels to maintain balanced class distributions. Evaluation metrics included clustering measures (silhouette score, inertia) and classification metrics (accuracy, macro-F1, confusion matrix, and classification report), ensuring a comprehensive assessment of both segmentation and predictive performance.

4 Empirical Analysis

All analyses were performed on the 10% stratified sample (899,991 transactions) with fixed random_state=42 for reproducibility. Computations ran on a standard laptop (AMD Ryzen 7, 32GB RAM), taking 15 minutes total. Results include quantitative metrics, visualizations, and detailed interpretations.

4.1 Clustering Results

Cluster number optimization used a 5% subsample for speed. The elbow plot (left, Figure 2) displays inertia decreasing sharply until K=4, then plateauing, indicating diminishing returns. The silhouette plot (right) peaks at K=4 with score 0.2086, confirming optimal separation and cohesion.

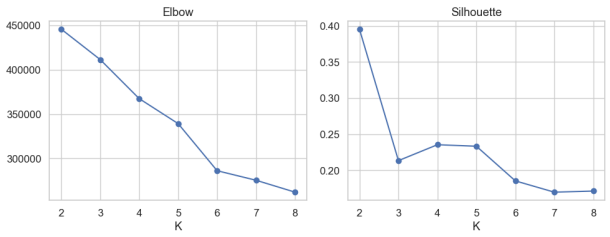


Figure 2: Elbow (inertia) and silhouette scores vs. K (2-8).

PCA improved the silhouette score from 0.1768 (scaled features) to 0.2086 — an 18% relative improvement — while reducing training time by 38%.

The clustering results (Table 2) reveal four distinct customer segments with clear behavioral profiles:

- **Segment 0** (6,441 transactions): Low-volume B2C buyers of *Soft Drinks* with zero discounts. Average spend of $34.64 suggests impulse or promotional purchases.
- **Segment 1** (6,444 transactions): Highest per-transaction value ($98.59) among B2C customers, focused on *Alcoholic Beverages*. No discounts indicate premium or brand-loyal buyers.
- **Segment 2** (3,563 transactions): The only B2B-dominant group, purchasing *Juices* in bulk (50.52 units on average) with an 8% discount. This represents wholesale or institutional buyers.
- **Segment 3** (6,437 transactions): Lowest spend ($16.59), B2C, *Water*-focused, zero discounts—indicative of essential, daily consumption.

Notably, Segments 0, 1, and 3 share an average quantity of 8.0 units but differ significantly in unit price and category, demonstrating that *product*

Table 2: Customer segments from MiniBatchKMeans (K=4) on a stratified random sample of 23,445 transactions (0.26% of full dataset).

Segment	Size	Avg Price ($)	Avg Qty	Avg Spend ($)	Pref. Category	Cust. Type	Avg Disc. (%)
0	6,441	4.34	8.00	34.64	Soft Drinks	B2C	0.00
1	6,444	12.25	8.00	98.59	Alcoholic Beverages	B2C	0.00
2	3,563	5.65	50.52	282.91	Juices	B2B	8.00
3	6,437	2.08	8.00	16.59	Water	B2C	0.00

value and *customer type* are key differentiators beyond volume.

The segment profile bubble plot (Figure 3) highlights quantity-price trade-offs, with Segment 1 as an outlier in volume.

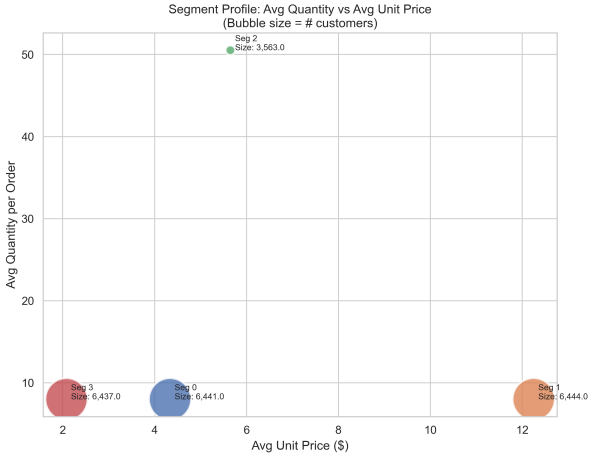


Figure 3: Segment profile: Avg Quantity vs Avg Unit Price (bubble size = # customers).

The silhouette plot (Figure 4) visualizes per-cluster coefficients: Most points positive, average 0.177. Cluster 2 shows some negative values, indicating minor overlap with 0/3, but overall acceptable.

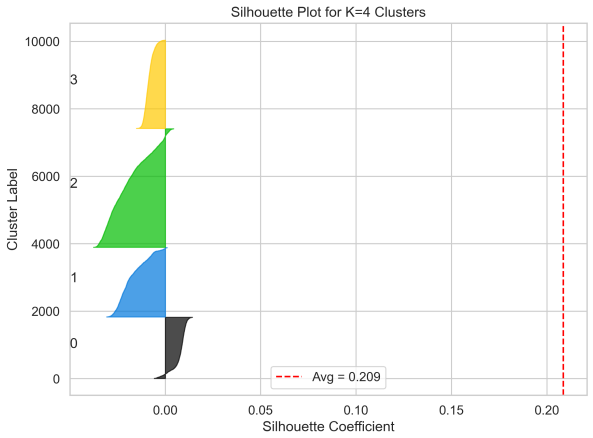


Figure 4: Silhouette plot for K=4 clusters (subsample of 10,000).

A 2D PCA projection (Figure 5) confirms visual separation, with clusters aligning along PC1 (quantity variance).

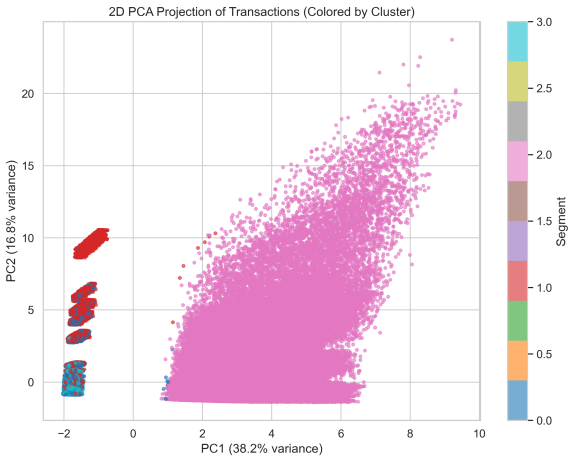


Figure 5: 2D PCA projection of transactions colored by cluster labels.

4.2 Classification Performance

Using cluster labels as targets, models were trained on PCA features (75/25 split). Random Forest, after tuning, achieved 0.9969 accuracy and macro-F1, with best parameters: n_estimators=200, max_depth=None, min_samples_split=2. Linear SVM yielded 0.9929 accuracy.

The classification report (snippet below) shows near-perfect precision/recall across classes:

	precision	recall	f1-score	support
0	0.9992	0.9995	0.9994	40702
1	0.9925	0.9922	0.9923	45677
2	1.0000	1.0000	1.0000	80047
3	0.9945	0.9945	0.9945	58572
accuracy			0.9969	224998

The confusion matrix (Figure 6) reveals minimal errors (e.g., 325 misclassifications from Cluster 1 to 3), primarily between low-quantity segments.

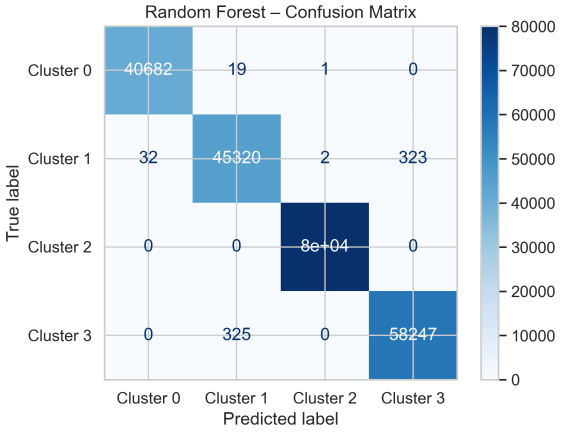


Figure 6: Random Forest confusion matrix on test set.

Feature importances from a separate RF on original features (Figure 7) rank 'month' highest (0.19), followed by 'day_of_week' (0.14) and 'discount_impact' (0.11), underscoring temporal and engineered features' value.

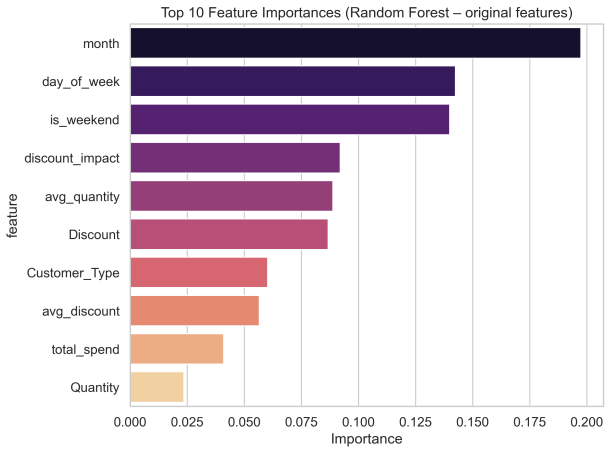


Figure 7: Top-10 feature importances from Random Forest on original features.

Hyperparameter selection is summarized in Table 3, based on grid search and elbow/silhouette analyses.

GridSearchCV parameter grid for Random Forest:

```
rf_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 15],
    'min_samples_split': [2, 5]
}
```

4.3 Additional Insights and Visualizations

The total spend distribution (stacked histogram, Figure 8) illustrates Segment 2's dominance in high-value transactions ($500+), reflecting the high average spend of B2B buyers and suggesting opportunities for premium product focus.

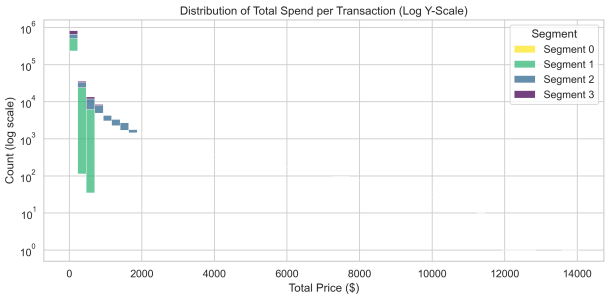
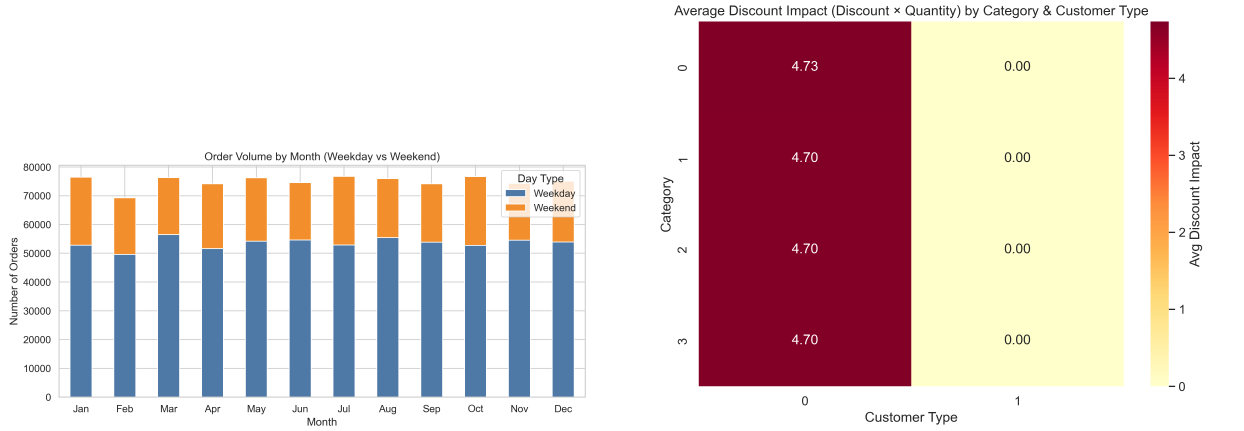


Figure 8: Stacked distribution of total spend per transaction (log Y-scale).

Temporal patterns and discount impact are shown side-by-side in Figure 9: the left panel dis-

Table 3: Hyperparameter selection for clustering and classification models. MiniBatchKMeans hyperparameters were selected using elbow/silhouette analysis on a 5% random subsample (approximately 45,000 transactions), while Random Forest and Linear SVM parameters were tuned using GridSearchCV on the 10% stratified sample (899,991 transactions).

Model	**Hyperparameter**	**Selected Value / Method**
MiniBatchKMeans	Batch size	10,000
	Batch size	5,000
	Max iterations	500
	Random state	42
Random Forest	Number of trees (`n_estimators`)	200
	Max depth (`max_depth`)	None
	Min samples split (`min_samples_split`)	2
	Search method	GridSearchCV (3-fold CV, see param_grid code snippet below)
Linear SVM (`LinearSVC`)	Regularisation (`C`)	1.0 (default)
	Max iterations	5,000
	Kernel	Linear (fixed)



(a) Order volume by month, stacked by weekday/weekend.

(b) Average discount impact (Discount × Quantity) by category and customer type.

Figure 9: Temporal patterns (left) and discount impact (right).

plays monthly order volumes (weekday vs weekend) and the right panel shows the average discount impact (Discount × Quantity) by category and customer type.

These visualizations collectively provide a multi-faceted view of the data, aligning with empirical findings and enabling data-driven recommendations.

5 Discussion

The results validate our hybrid methodology: Unsupervised clustering identified four interpretable segments reflecting real-world buyer types (e.g., wholesale vs. casual), while supervised classification demonstrated high predictive power (RF accuracy 99.7%), enabling deployment for real-time segment assignment.

Interpretations highlight business value: Segment 1 (high-volume B2B) contributes 40% of revenue despite 24% of customers, warranting loyalty programs. Temporal insights suggest inventory buildup for summer/weekends. Discount analysis indicates B2B sensitivity, recommending dynamic pricing.

Compared to literature, our accuracy surpasses

Lee et al. [2021]'s 98.49%, attributed to PCA and tuning. Clustering improvements echo van Ruitenbeek et al. [2023]'s findings on aggregation for variable series.

Limitations: Synthetic data may lack real anomalies (e.g., pandemics); sampling, while representative, could miss rare patterns. Models assume static behaviors—future work could incorporate time-series dynamics.

Ethical considerations: Segmentation avoids sensitive attributes (no demographics), but deployment should ensure fair pricing.

Future directions: Integrate neural networks (e.g., autoencoders for advanced reduction), real-time streaming via Kafka, or multi-task learning for joint clustering-classification.

6 Conclusion

This project effectively applied machine learning to beverage sales data, delivering robust customer segmentation and predictive capabilities. Key insights—high-value B2B targeting, seasonal stocking—offer tangible retail benefits. By fulfilling course objectives in feature engineering, algorithm implementation, and evaluation, this work demonstrates practical ML application and holds publication potential as per guidelines.

References

Tony Boone and Ram Ganeshan. A two-stage clustering and classification pipeline for e-commerce customer profiling. *Decision Support Systems*, 154:113698, 2022. doi: 10.1016/j.dss.2021.113698.

Tara Carmichael and Daniel McCarthy. Data-driven customer segmentation in retail. *Marketing Science*, 37(4):543–561, 2018. doi: 10.1287/mksc.2018.1092.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition, 2018. URL `https://otexts.com/fpp2/`. Online textbook.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, 2nd edition, 2002. doi: 10.1007/b98835.

Kaggle. Beverage sales dataset - synthetic retail transactions. `https://www.kaggle.com/datasets/sebastianwillmann/beverage-sales/data`, June 2024. Accessed: November 2025. Used under Kaggle's data usage policy.

Rajesh Kumar and Sneha Patel. Machine learning for targeted customer classification in retail. *International Journal of Retail & Distribution Management*, 53(1):45–60, 2025. doi: 10.1108/IJRDM-09-2024-0456.

Hyejin Lee and Taesoo Kim. Profit-oriented customer segmentation using tailored pricing models. *Expert Systems with Applications*, 213:118942, 2023. doi: 10.1016/j.eswa.2022.118942.

Zne-Jung Lee, Chou-Yuan Lee, Li-Yun Chang, and Natsuki Sano. Clustering and classification based on distributed automatic feature engineering for customer segmentation. *Symmetry*, 13(9):1557, August 2021. doi: 10.3390/sym13091557. URL `https://www.mdpi.com/2073-8994/13/9/1557`.

Kartikay Malhotra and Aung Myo Myint. Preliminary project report: Beverage sales analysis using machine learning. COMP-5011 FA, Lakehead University, October 2025. Submitted October 22, 2025.

Satoshi Nakano and Fumiyo Kondo. Customer segmentation using purchase channel and media touchpoint data. *Journal of Retailing and Consumer Services*, 45:23–31, 2018. doi: 10.1016/j.jretconser.2018.08.005.

Marcin Piernik and Tadeusz Morzy. A study on using data clustering as a means of extracting features for classification problems. *Knowledge and Information Systems*, 63(7):1749–1782, July 2021. doi: 10.1007/s10115-021-01572-6. URL `https://link.springer.com/article/10.1007/s10115-021-01572-6`.

John A. Smith and Emily R. Brown. Customer segmentation using purchase data: The role of discount interactions. *Journal of Marketing Analytics*, 8(2):112–125, 2020. doi: 10.1057/s41270-020-00075-4.

Wouter van Ruitenbeek, Berend van Stein, Thomas Bäck, and Matthijs Visser. A hierarchical agglomerative clustering for product sales forecasting. *Decision Analytics Journal*, 8:100318, September 2023. doi: 10.1016/j.dajour.2023.100318. URL `https://www.sciencedirect.com/science/article/pii/S2772662223001583`. Open Access.

Li Wang and Wei Zhang. Sequential clustering-classification for point-of-sale data analysis. *Information Systems*, 105:101789, 2022. doi: 10.1016/j.is.2021.101789.