Retail Customer Behavioral Data Dictionary

MegaMart Retail - Customer Analytics Dataset

File: `retail_customer_data.csv`
Observations: 3,000 active customers
Time Period: 12 months (2023-2024)
Business Context: National retail chain - general merchandise, apparel, electronics, home goods
Data Type: Unsupervised learning (NO predefined customer segments or labels)

Variable Descriptions

Identification Variables

Variable	Type	Description	Values
`customer_id`	String	Unique customer identifier	CUST_0001 to CUST_3000

Purchase Behavior Metrics

Variable	Type	Scale	Description
`monthly_transactions`	Float	0.2-18.0	Average number of purchases per month
`avg_basket_size`	Float	1.0-35.0	Average number of items per transaction
`total_spend`	Float	50—9,000	Total spending over 12-month period
`recency_days`	Integer	1-90	Days since last purchase (lower = more recent)

Engagement Metrics

Variable	Type	Scale	Description
`avg_session_duration`	Float	2.0-75.0	Average website/app session duration (minutes)
`email_open_rate`	Float	0.00-0.95	Proportion of marketing emails opened
`product_views_per_visit`	Float	3.0-65.0	Average number of products viewed per session

Loyalty and Satisfaction Indicators

Variable	Type	Scale	Description
`return_rate`	Float	0.00-0.50	Proportion of purchases returned/exchanged
`customer_tenure_months`	Integer	1-48	Months since first purchase (customer age)

Data Quality Notes

Completeness

- **Missing Data**: No missing values in the dataset (100% complete data)
- **Data Quality**: All 3,000 observations have complete information across all 10 variables
- **Outliers**: Some extreme values present (e.g., very high spenders, very frequent shoppers) - these are legitimate customer behaviors, not data errors

Variable Distributions

- **Monthly Transactions**: Right-skewed distribution (most customers shop 1-5 times/month, some shop 10+ times)
- **Total Spend**: Highly right-skewed (mean = $2,250$, $median$ =1,280, range: 50—9,000)
- **Session Duration**: Bimodal distribution (quick shoppers vs browsers)
- **Email Engagement**: Varies widely (some customers never open emails, others are highly engaged)

- **Return Rate**: Right-skewed (most customers have low return rates <15%, some have high rates >30%)

Expected Cluster Patterns

Note for Instructors: The dataset contains **5 underlying customer behavioral patterns**, though students are not told this. Expected natural segments include:

1. **High-Value Loyalists**: High spend, frequent transactions, low return rate, high engagement
2. **Bargain Hunters**: Moderate frequency, small baskets, price-sensitive behavior
3. **Window Shoppers**: High browsing (session duration, product views), low conversion (spending)
4. **Occasional Splurgers**: Infrequent but high-value purchases, moderate engagement
5. **New Explorers**: Recent customers (low tenure), moderate behavior across metrics

Students may discover 3-6 clusters depending on their methodology. Award full credit for well-justified solutions.

Analytical Considerations

Clustering Suitability

- **Sample Size**: N=3,000 is excellent for clustering (sufficient for detecting patterns)
- **Variable Types**: All 9 behavioral variables are continuous/numeric (ideal for distance-based clustering)
- **Measurement Scales**: Variables measured on different scales (dollars, counts, rates) - **standardization is critical**
- **Expected Separation**: Clear behavioral differences exist between segments

Required Preprocessing

1. **Remove Identification Column**: Exclude customer_id from clustering analysis (not a behavioral variable)
2. **Standardization**: Apply StandardScaler to all 9 behavioral variables before clustering
    - Variables have vastly different scales (total_spend in thousands, return_rate 0-0.5)

    - Without standardization, total_spend would dominate distance calculations
    - Use StandardScaler (mean=0, std=1) not MinMaxScaler for clustering
3. **Outlier Handling**: Outliers are legitimate customer behaviors - do not remove, but be aware they may influence cluster centroids

Recommended Analyses

1. **Exploratory Data Analysis**:
    - Distribution plots for each variable
    - Correlation matrix to identify multicollinearity
    - Box plots to detect outliers
    - Scatter plots to visualize relationships
2. **Hierarchical Clustering**:
    - Compare linkage methods: single, complete, average, Ward's
    - Create dendrograms to visualize cluster hierarchy
    - Identify optimal cutting point (look for large vertical gaps)
    - Ward's method typically performs best for customer segmentation
3. **K-Means Clustering**:
    - Apply elbow method (plot inertia vs k for k=2 to 10)
    - Calculate silhouette scores for each k value
    - Compare results with hierarchical clustering
    - Select optimal k balancing statistical metrics and business actionability
4. **Cluster Validation**:
    - Silhouette analysis (coefficients should be >0.3 for reasonable clustering)
    - Cluster profiling (mean values of each variable per cluster)
    - Business interpretation (can you describe each cluster in plain English?)
    - Stability checks (do results make sense?)
5. **Visualization**:
    - PCA projection to 2D for cluster visualization
    - Heatmaps showing cluster profiles across variables
    - Silhouette plots showing cohesion within clusters

Distance Metric Considerations

- **Default**: Euclidean distance is appropriate for this dataset (continuous variables, standardized)
- **Alternatives**: Manhattan distance less sensitive to outliers but not necessary here
- **Inappropriate**: Cosine similarity (better for text/sparse data), Hamming distance (for categorical data)

Business Context Variables

Industry Benchmarks

- **Average Transaction Frequency**: 2-4 times per month (retail industry standard)
- **Email Open Rates**: 15-25% is typical (MegaMart's top segment reaches 70%+)
- **Return Rates**: 10-15% is normal (higher for apparel, lower for electronics)
- **Customer Lifetime Value**: Top 20% of customers typically generate 60-80% of revenue

Retail Segmentation Best Practices

- **Actionability**: Segments must be large enough to justify targeted campaigns (typically >10% of customer base)
- **Distinctiveness**: Clear differences in behavior between segments
- **Stability**: Segments should be relatively stable over time (not change dramatically month-to-month)
- **Profitability**: Consider segment value (high-spend segments warrant more investment)

Marketing Strategy Frameworks

Students should consider segment-specific strategies across:

1. **Product**: Which products to promote to each segment
2. **Price**: Discount sensitivity and promotional strategies
3. **Place**: Channel preferences (online vs in-store)
4. **Promotion**: Email frequency, content style, messaging tone

File Format

- **Encoding**: UTF-8
- **Delimiter**: Comma (,)
- **Missing Values**: None present
- **Header**: First row contains variable names
- **Index**: No row index included (use customer_id as primary key)

Expected Analysis Outcomes

Typical Student Findings

Optimal Number of Clusters: Most students will select k=4 or k=5 based on:

- Elbow method shows inflection around k=4-5
- Silhouette scores peak around k=4-5
- Dendrogram shows 4-5 natural groupings

Common Cluster Interpretations:

- VIP/Premium customers (high value, high frequency)
- Value-seekers (price-sensitive, moderate engagement)
- Browsers (high engagement, low conversion)
- Occasional buyers (infrequent but deliberate purchases)
- New/Emerging customers (recent, still exploring)

Grading Considerations

Full credit should be awarded for:

- Well-justified selection of k (even if different from "true" k=5)
- Clear business interpretation of discovered clusters
- Appropriate use of validation metrics
- Actionable marketing strategies

Common student errors to watch for:

- Forgetting to standardize data (major technical error)
- Choosing k without justification
- Describing clusters only with statistics (no business interpretation)
- Creating too many clusters (k>7) that aren't actionable

- Ignoring silhouette scores indicating poor clustering quality

---

Note: This dataset is synthetically generated for educational purposes. All customer identifiers and behavioral data are fictional but based on realistic retail industry patterns and customer segmentation research. The underlying structure was designed to reward rigorous analysis while allowing flexibility in cluster interpretation.

Instructor Reference: Data Generation Details

Synthetic Data Structure:

- Generated from 5 multivariate normal distributions with distinct means and covariance matrices
- Cluster proportions: VIP (18%), Bargain (28%), Window (32%), Splurger (14%), New (8%)
- Correlations designed to reflect realistic customer behavior:
    - Positive: monthly_transactions and total_spend
    - Positive: session_duration and product_views
    - Negative: return_rate and email_open_rate (satisfied customers engage more)
    - Negative: recency_days and monthly_transactions (frequent shoppers are recent)

Expected Silhouette Scores:

- k=3: ~0.42 (merges some distinct groups)
- k=4: ~0.46 (good separation)
- k=5: ~0.48 (optimal - reflects true structure)
- k=6: ~0.44 (starts over-segmenting)
- k=7+: <0.40 (poor clustering quality)

Validation: Hierarchical clustering with Ward's linkage should recover the 5-cluster structure nearly perfectly (Adjusted Rand Index >0.95). K-means should achieve ARI >0.90 with k=5.