# Problem Understanding & Framing

***Business Problem Statement***

The company aims to better understand its customer base in order to design targeted marketing strategies, improve customer experience, and optimize promotional spending. Currently, customers are treated as a relatively homogeneous group despite exhibiting differences in demographics, purchasing behavior, loyalty status, and promotion usage.

The key business challenge is the lack of data-driven customer segmentation that can reveal distinct customer groups with similar behaviors and preferences. Without clear segments, marketing campaigns risk being inefficient, generic, and misaligned with customer needs-leading to suboptimal customer satisfaction and return on investment.

The goal of this project is to identify meaningful and actionable customer segments based on demographic and behavioral data.

***Data Science Problem Framing***

From a data science perspective, the objective is to group customers into distinct clusters such that:

- Customers within the same cluster are similar to each other, and
- Customers in different clusters are meaningfully different.

There is no predefined target label (e.g., customer type or churn flag), so the task does not involve prediction but rather structure discovery within the data.

***Type of Machine Learning Task***

- Task Type: Unsupervised Learning – Clustering
- Candidate Algorithms:
  - K-Means
  - Hierarchical Clustering
  - DBSCAN (for density-based structure and noise detection)

*This is not:*

- Classification (no labeled outcome)
- Regression (no continuous target to predict)
- Recommendation (no user-item interaction modeling)
- Anomaly detection (though outliers may be analyzed as a by-product)

***Success Metrics (Data Science Metrics)***

Since clustering lacks ground truth labels, model performance will be evaluated using internal validation metrics, including:

- Silhouette Score (primary metric)
  - Measures how well each data point fits within its cluster compared to other clusters
- Elbow Method (Within-Cluster Sum of Squares, WCSS)
  - Used to guide optimal number of clusters for K-Means
- Davies–Bouldin Index (secondary)
  - Lower values indicate better cluster separation

These metrics assess cluster cohesion and separation, not predictive accuracy.

### ***Business KPIs and Impact***

The resulting customer segments will be evaluated based on their business usefulness, including:

- Improved targeting of promotions (e.g., high-frequency vs low-frequency buyers)
- Potential cost savings from reduced blanket promotions
- Increased marketing uplift through personalized offers
- Better understanding of high-value vs at-risk customer segments
- Insights into loyalty behavior and satisfaction drivers

Success from a business standpoint is defined by whether the clusters are:

- Interpretable
- Actionable
- Aligned with business decisions (marketing, loyalty programs, product strategy)

### ***Objective***

To apply unsupervised clustering techniques to customer demographic and behavioral data in order to uncover distinct, interpretable customer segments that can support data-driven marketing and customer strategy decisions.

## Data Collection & Understanding

***Dataset Source & Description***

This project uses a customer-level transactional and demographic dataset sourced from Kaggle, a public data science platform that hosts real-world and simulated datasets commonly used for analytics and machine learning applications. The dataset is representative of a retail/e-commerce business context and is suitable for customer analytics and segmentation tasks.

The dataset contains one row per customer, capturing a combination of demographic attributes, purchasing behavior, loyalty indicators, and satisfaction metrics.

*Data*: Customer Purchases Behaviour Dataset
*Source*: https://www.kaggle.com/datasets/sanyamgoyal401/customer-purchases-behaviour-dataset
*Granularity*: Customer-level
*Observational Unit*: Individual customer
*Target Variable*: None (unsupervised learning)

This dataset is appropriate for customer segmentation as it integrates multiple dimensions of customer information, including:

- Demographics: age, gender, region, education
- Economic capacity: income
- Behavioral signals: purchase frequency, purchase amount, promotion usage
- Relationship indicators: loyalty status, satisfaction score

The combination of demographic, behavioral, and economic variables enables the identification of meaningful customer groups based on similarity, aligning well with real-world customer relationship management and marketing use cases.

***Feature Overview***

| Feature Type | Variables |
|---|---|
| Demographic | age, gender, education, region |
| Economic | income |
| Behavioral | purchase_frequency, purchase_amount, promotion_usage |
| Relationship / Attitudinal | loyalty_status, satisfaction_score |
| Product Preference | product_category |

***Expected Data Quality Considerations***

- Missing values:
  - Possible in income, education, satisfaction score
- Outliers:
  - Income and purchase_amount may contain extreme values
- Categorical imbalance:
  - Some regions or product categories may be underrepresented
- Scaling sensitivity:
  - Numeric variables vary significantly in magnitude (e.g., age vs income)

These considerations are critical because distance-based clustering algorithms are sensitive to scale and noise.

***Data Dictionary***

| Variable | Type | Description | Units / Values |
|---|---|---|---|
| Age | Integer | Customer age | Years |
| Gender | Categorical | Customer gender | {Male, Female} |
| Income | Numeric | Annual income | Currency units |
| Education | Categorical | Highest education attained | {HighSchool, Bachelor, Masters, PhD} |
| Region | Categorical | Customer location | {North, South, East, West} |
| loyalty_status | Categorical | Loyalty program tier | {Regular, Silver, Gold, Platinum} |
| purchase_frequency | Categorical | Frequency of purchases | {rare, occasional, frequent} |
| purchase_amount | Numeric | Total purchase amount | Currency units |
| product_category | Categorical | Primary product category | {Books, Clothing, Food, Electronics, …} |
| promotion_usage | Binary | Promotion usage indicator | {0 = No, 1 = Yes} |
| satisfaction_score | Integer | Customer satisfaction rating | Scale: 1–10 |

# EDA + Feature Engineering Report

***Objective***

To prepare a clean, well-encoded, well-scaled feature matrix suitable for distance-based clustering, while producing interpretable EDA and explainability artifacts.

***EDA + Feature Engineering Report***

1) Cleaning decisions (with justification)

- Duplicates: drop duplicate id rows (customer should be unique).
- Missing values:
  - Numeric (age, income, purchase_amount, satisfaction_score): impute using median (robust to outliers).
  - Categorical (gender, education, region, loyalty_status, purchase_frequency, product_category): impute using mode (most common).
- Outliers (income, purchase_amount): cap using IQR winsorization (keeps all customers while limiting extreme values that distort distance).

2) Feature engineering (with justification)

- Encoding
  - Ordinal encoding for ordered categories:
    - loyalty_status: Regular < Silver < Gold < Platinum
    - purchase_frequency: rare < occasional < frequent
    - education: HighSchool < Bachelor < Masters < PhD
  - One-hot encoding for nominal categories: gender, region, product_category
- Scaling
  - Use RobustScaler for numeric features to reduce outlier sensitivity.
- Binning
  - age_band (optional) to capture lifecycle segments without assuming linearity.
- Domain-derived features
  - promo_intensity = promotion_usage * purchase_amount (proxy for “value impacted by promos”)
  - spend_per_purchase_proxy = purchase_amount / freq_multiplier (approximation of avg ticket / engagement)
  - value_index = z(income) + z(purchase_amount) (simple composite “capacity + value”)

3) Applied EDA (clustering readiness)

- Distributions: histograms + boxplots (spot skew/outliers).
- Relationships: pairwise correlations among numeric features.
- Clustering tendency: compute Hopkins statistic (values close to 1 suggest clusterable structure; ~0.5 suggests randomness).

4) Feature selection (at least one)

- Filter method:
  - remove near-zero variance features (VarianceThreshold)
  - remove highly correlated numeric features (keep one of each pair above threshold)
- (Optional) PCA-based selection: inspect explained variance and loadings.

5) Dimensionality reduction

- PCA for compact representation used for clustering (or for visualization).
- t-SNE / UMAP for 2D visualization of cluster structure (not for training).

6) Explainability (SHAP/LIME equivalent for clustering)

Since clustering has no labels, we explain clusters by:

1. Fit clusters (later step) → get cluster_label
2. Train a surrogate classifier (e.g., RandomForest) to predict cluster labels from features
3. Use SHAP on the surrogate model to identify what features most separate clusters

# Model Implementation and Evaluation

***Objective***

The objective of this step is to implement and compare multiple unsupervised clustering algorithms in order to identify meaningful customer segments from the engineered feature matrix (X_preprocessed). The comparison considers not only quantitative clustering metrics, but also scalability, stability, coverage, and business interpretability, which are critical for real-world customer segmentation.

***Models Considered***

Three clustering algorithms were evaluated:

1. K-Means Clustering

K-Means is a centroid-based algorithm that partitions the dataset into a predefined number of clusters by minimizing within-cluster variance. It is computationally efficient and scales well to large, high-dimensional datasets, making it a common industry choice for customer segmentation.

2. Hierarchical (Agglomerative) Clustering

Agglomerative clustering is a bottom-up approach that recursively merges observations based on a linkage criterion. While it provides interpretability in how clusters are formed, it has quadratic memory complexity, which limits its applicability to large datasets.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based algorithm that identifies dense regions as clusters and labels sparse observations as noise. It does not require the number of clusters to be specified in advance and is well suited for discovering irregularly shaped clusters and outliers.

***Evaluation Metrics***

Because clustering is an unsupervised task with no ground-truth labels, internal validation metrics were used:

- Silhouette Score (Primary Metric):
  Measures how similar observations are within their cluster compared to other clusters. Higher values indicate better separation.
- Davies-Bouldin Index:
  Measures average similarity between clusters; lower values indicate better separation.
- Calinski-Harabasz Index:
  Measures the ratio of between-cluster dispersion to within-cluster dispersion; higher values indicate better defined clusters.

Silhouette score was used as the primary selection metric, with secondary metrics used to support interpretation.

***Model Implementation***

1. *K-Means Clustering*

K-Means was trained on the full preprocessed dataset.

The number of clusters (k) was swept from 2 to 10.

Model selection was guided by:

- The Elbow Method (within-cluster sum of squares)
- Silhouette scores across values of k

The best performing configuration occurred at k = 2, which yielded the highest silhouette score among K-Means configurations.

2. *Hierarchical Clustering and Computational Constraints*

Hierarchical clustering was initially attempted on the full feature matrix. However, due to its O($n^2$) memory complexity, it resulted in memory allocation errors when applied to the full, high-dimensional dataset.

To address this limitation while maintaining methodological completeness:

- The data was first reduced to 10 principal components using PCA
- A random subsample of up to 3,000 observations was used for clustering

Linkage methods (ward, complete, average) and cluster counts (2–10) were evaluated. The best hierarchical configuration used Ward linkage with k = 2.

Despite this mitigation, hierarchical clustering was not selected as the final model due to scalability limitations and weaker cluster separation.

3. *DBSCAN Clustering*

DBSCAN was applied to the full preprocessed dataset.

Hyperparameters were tuned as follows:

- A k-distance plot was used to guide the choice of eps
- Multiple combinations of eps and min_samples were evaluated

DBSCAN identified 5 dense clusters with a high silhouette score (0.69). However, it labeled a very large proportion of customers (99,937 observations) as noise.

***Model Comparison***

The best performing configuration from each model family is summarized below:

| Model | Configuration | Clusters | Silhouette | Davies-Bouldin | Calinski-Harabasz | Noise Points |
|---|---|---|---|---|---|---|
| DBSCAN | eps = 0.3, min_samples = 10 | 5 | 0.69 | 0.37 | 624.33 | 99,937 |
| Hierarchical | Ward linkage, k = 2 | 2 | 0.25 | 1.51 | 767.23 | 0 |
| K-Means | k = 2 | 2 | 0.21 | 1.70 | 20,554.09 | 0 |

***Interpretation of Results***

Although DBSCAN achieved the highest silhouette score, this metric was computed only on non-noise observations and reflects the quality of a small dense subset rather than the full population. Excluding the majority of customers as noise significantly limits DBSCAN’s usefulness for business-oriented segmentation.

Hierarchical clustering showed moderate separation but required dimensionality reduction and subsampling, making it unsuitable for large-scale deployment.

K-Means, despite a lower silhouette score, successfully clustered the entire customer base, produced stable and interpretable segments, and scaled efficiently. In practical customer analytics, moderate silhouette scores are common due to overlapping behaviors and heterogeneous customer profiles.

### *Final Model Selection*

K-Means clustering with k = 2 was selected as the final model for downstream analysis.

This decision was based on:

- Full coverage of the customer population
- Computational scalability
- Stability across runs
- Ease of interpretation and profiling
- Suitability for actionable business segmentation

Quantitative metrics were considered alongside business relevance, ensuring that the selected model supports real-world decision-making rather than purely mathematical optimality.

### *Reproducibility*

To ensure reproducibility:

- All model configurations and random seeds were fixed
- Trained models, cluster labels, and evaluation metrics were saved as artifacts
- Diagnostic plots and summary tables were exported for traceability

# Bias & Fairness Analysis

***Objective***

This step evaluates the selected K-Means (k = 2) customer segmentation model from an explainability, fairness, and ethical deployment perspective. Although the task is unsupervised, cluster assignments may influence downstream business actions such as promotion targeting, prioritization, or personalization. Therefore, transparency, bias auditing, and appropriate safeguards are essential.

***Explainability of Model Decisions***

*Challenge in Unsupervised Models*

K-Means clustering does not natively provide feature importance or decision rules. As a result, traditional explainability tools (e.g., coefficients or odds ratios) are not directly applicable.

*Methodology: Surrogate Model with SHAP-style Explainability*

To enable interpretability in an unsupervised setting, a post-hoc explainability approach was applied:

- Final cluster labels were treated as pseudo-targets
- A Random Forest surrogate classifier was trained to predict cluster membership
- SHAP-style feature contributions were analyzed on the surrogate model

This approach is a commonly accepted method for explaining clustering outcomes and allows interpretation without altering the original clustering model.

*Key Drivers Identified*

The clusters were primarily differentiated by:

- Purchase amount
- Purchase frequency
- Loyalty status
- Promotion usage

Economic and demographic attributes (e.g., income, age, gender) played a comparatively secondary role, suggesting that the segmentation is predominantly behavior-driven rather than demographically determined.

***Data & Model Limitations***

*1. Cluster Imbalance*

The final segmentation exhibits an approximate 80/20 split between clusters. While not inherently problematic, this imbalance requires caution to avoid over-favoring the majority cluster in downstream applications.

*2. Data Leakage*

Traditional target leakage is not applicable due to the absence of a labeled outcome variable. However, representational leakage remains a potential risk if features correlated with sensitive attributes (e.g., income) are used directly in downstream decision rules.

*3. Overfitting*

Overfitting risk is limited due to:

- Use of a simple, low-variance model (K-Means)
- Small number of clusters (k = 2)
- Absence of outcome optimization

Nonetheless, clusters may be sensitive to feature scaling and snapshot timing.

*4. Temporal Limitation*

The model represents a static snapshot of customer behavior and does not capture behavioral drift or customer migration between segments over time.

***Bias & Fairness Auditing***

*Sensitive Attribute Audited*

Fairness auditing focused on demographic and economic attributes that may act as sensitive or proxy variables:

- Income band (socioeconomic proxy)
- Age group
- Gender
- Education
- Region

*Fairness Metric and Methodology*

A demographic parity–style audit was conducted by comparing normalized cluster assignment rates across income groups. This approach is appropriate for unsupervised models where no ground truth labels exist.

***Results: Economic and Demographic***

*Income (%)*

| **income_band** | **Cluster 0** | **Cluster 1** |
|---|---:|---:|
| Low | 24.9 | 25 |
| Lower-Mid | 25.6 | 24.9 |
| Upper-Mid | 24.6 | 25.1 |
| High | 24.9 | 25 |

*Cluster composition across income bands is nearly uniform, with each income group contributing approximately 25% to both clusters. No income band shows disproportionate concentration in either cluster.*

*Age (%)*

| age_group | Cluster 0 | Cluster 1 |
|---|---:|---:|
| <25 | 10.3 | 11 |
| 25–34 | 73.8 | 73.2 |
| 35–44 | 15.8 | 15.7 |
| 45–54 | 0.1 | 0.1 |

*Age distributions are highly similar across clusters. Both clusters are dominated by customers aged 25–34, with minimal representation from older age groups. No age group exhibits material over- or under-representation.*

*Gender (%)*

| gender | Cluster 0 | Cluster 1 |
|---|---:|---:|
| Female | 50.1 | 50.1 |
| Male | 49.9 | 49.9 |

*Gender composition is effectively identical across clusters, with both clusters split almost exactly 50/50 between male and female customers.*

*Education (%)*

| education | Cluster 0 | Cluster 1 |
|---|---:|---:|
| Bachelor | 30.5 | 30.2 |
| College | 39.1 | 40.1 |
| HighSchool | 20.1 | 20 |
| Masters | 10.2 | 9.7 |

*Education levels are evenly distributed across clusters, with only negligible percentage differences across Bachelor, College, High School, and Master's categories.*

*Region (%)*

| region | Cluster 0 | Cluster 1 |
|---|---:|---:|
| East | 30.2 | 30 |
| North | 19.9 | 19.9 |
| South | 20.2 | 20.1 |
| West | 29.8 | 30 |

*Geographic distribution is nearly identical across clusters, indicating no regional concentration or geographic bias.*

### *Fairness Metrics: Disparate Impact Perspective*

In the absence of labeled outcomes, traditional fairness metrics such as equalized odds are not applicable. Instead, a disparate impact–style assessment was conducted by examining whether the probability of cluster assignment differs materially across groups.

For all audited demographic and economic attributes, cluster membership probabilities are effectively invariant, resulting in implied disparate impact ratios close to 1.0. This indicates no evidence of adverse impact associated with cluster assignment.

From a disparate impact perspective, the clustering model does not systematically disadvantage or privilege any demographic or socioeconomic group.

### *Results: Behavioral, Relationship / Attitudinal, and Product Preference*

While demographic and economic variables show parity, behavioral attributes exhibit strong differentiation, confirming that the clusters capture meaningful behavioral structure.

*Purchase Frequency (%)*

| purchase_frequency | Cluster 0 | Cluster 1 |
|---|---:|---:|
| frequent | 100 | 0 |
| occasional | 0 | 37.4 |
| rare | 0 | 62.6 |

*Cluster 0 consists entirely of frequent purchasers, while Cluster 1 is composed of occasional and rare purchasers.*

*Purchase Amount (Mean)*

| Cluster | purchase_amount |
|---|---:|
| 0 | 9,616.85 |
| 1 | 9,639.28 |

*Average purchase amounts are similar across clusters, suggesting that engagement frequency, rather than spend per transaction, is the primary differentiator.*

*Promotion Usage (Mean)*

| Cluster | promotion_usage |
|---|---:|
| 0 | 0.3 |
| 1 | 0.3 |

*Promotion usage rates are identical, indicating similar price sensitivity across clusters.*

*Loyalty Status (%)*

| loyalty_status | Cluster 0 | Cluster 1 |
|---|---|---|
| Gold | 9.9 | 9.9 |
| Regular | 59.7 | 60.3 |
| Silver | 30.4 | 29.8 |

*Loyalty tier distributions are nearly identical across clusters, suggesting that loyalty status alone does not drive segmentation.*

*Satisfaction Score (Mean)*

| Cluster | satisfaction_ score |
|---|---|
| 0 | 5.01 |
| 1 | 5.01 |

*Average satisfaction scores are the same across clusters, indicating comparable customer sentiment.*

*Product Category (%)*

| product_category | Cluster 0 | Cluster 1 |
|---|---|---|
| Beauty | 5.1 | 5 |
| Books | 14.9 | 14.9 |
| Clothing | 20.1 | 20 |
| Electronics | 29.6 | 30.1 |
| Food | 14.8 | 14.8 |
| Health | 10.3 | 10.1 |
| Home | 5.2 | 5 |

*Product category distributions are highly similar across clusters, suggesting that segmentation is not driven by category preference, but rather by engagement patterns.*

***Interpretation***

*1. Fairness Interpretation (Economic & Demographic Attributes)*

From a fairness perspective, analysis focused on economic and demographic attributes (income, age, gender, education, and region), as these variables may act as sensitive or proxy attributes in downstream decision-making.

Across all audited groups, cluster membership distributions are highly consistent. Income bands contribute approximately equal proportions to both clusters, gender distributions are identical, and age, education, and regional compositions show no meaningful divergence. These results indicate no observable demographic or socioeconomic bias in cluster assignment.

Accordingly, the model satisfies a demographic parity–style fairness criterion, with implied disparate impact ratios close to 1.0 across all audited groups.

*2. Behavioral, Relationship, and Product Interpretation (Non-Fairness Dimensions)*

Behavioral, relationship, and product variables were analyzed for cluster characterization rather than fairness auditing, as differences along these dimensions are expected and desirable in customer segmentation.

Behavioral Attributes

The clusters exhibit a clear separation in purchase frequency. Cluster 0 consists entirely of frequent purchasers, while Cluster 1 comprises occasional and rare purchasers. In contrast, average purchase amounts and promotion usage rates are nearly identical across clusters, indicating that the segmentation reflects engagement frequency rather than spending level or price sensitivity.

Relationship / Attitudinal Attributes

Loyalty status and satisfaction scores show minimal differentiation between clusters. This suggests that differences in engagement frequency are not driven by dissatisfaction or loyalty tier, and that both clusters maintain comparable relationships with the business.

Product Preference

Product category distributions are almost identical across clusters, indicating that product preference does not drive segmentation. Customers in both clusters engage with similar product categories, reinforcing the conclusion that clusters represent differences in engagement intensity rather than product affinity.

*3. Integrated Interpretation*

Taken together, these results indicate that:

- The clustering model is fair with respect to demographic and economic attributes
- Behavioral differences are the intended and meaningful drivers of segmentation
- Relationship strength and product preferences remain stable across clusters

The resulting segments therefore represent engagement-driven customer groups, which are appropriate and actionable for real-world business use cases such as activation, retention, and lifecycle management.

***Limitations & Residual Risks***

Despite the absence of detected bias in cluster assignment, several limitations remain:

- Proxy Risk - Income may still act as a proxy in downstream decision rules, even if the clustering itself is neutral.
- Usage Risk - Bias can be introduced during application, for example if one cluster is systematically favored with promotions or benefits.
- Static Snapshot - The audit reflects a single time period and does not capture temporal changes in customer behavior.

***Mitigation Strategies***

To ensure ethical deployment, the following safeguards are recommended:

- Purpose Limitation: Use clusters for personalization and insight generation, not as sole criteria for exclusion or disadvantage.

- Policy Constraints: Ensure that benefits or promotions are not allocated exclusively to a single cluster.
- Monitoring: Periodically re-audit cluster distributions across demographic groups as new data becomes available.
- Human Oversight: Maintain managerial review when cluster outputs inform high-impact decisions.

### ***Conclusion***

The fairness audit indicates that the selected K-Means (k = 2) clustering model does not exhibit systematic demographic or socioeconomic bias. Cluster assignments are consistent across income, age, gender, education, and region, satisfying a demographic parity–style fairness criterion.

At the same time, strong differentiation is observed in behavioral engagement, particularly purchase frequency, confirming that the model captures meaningful customer behavior rather than demographic stratification. This demonstrates that ethical AI considerations are applicable even in unsupervised learning contexts and reinforces the importance of evaluating both model impact and model intent.

By integrating explainability, quantitative fairness auditing, and governance-focused mitigation strategies, this project ensures that the resulting customer segmentation is analytically sound, ethically responsible, and suitable for real-world deployment.