Segmenting Consumers of Bath Soap

CASE STUDY
CRISA

Created by:

Gowtham Katari
Sinchan Sinha

CONTENT

Understanding the Business Landscape

CRISA, a respected market research firm, is on a mission. Traditionally, companies grouped customers based on their age, gender, or income, the typical demographics. But today, CRISA aims to go a step further. They want to understand how people *actually shop* — their preferences, behaviours, and what truly drives their purchase decisions.

This case study focuses specifically on the **Bath Soaps** category, a common but competitive market. CRISA has collected a rich dataset covering details like how often customers buy soaps, how loyal they are to certain brands, their sensitivity to price, and what influences them discounts, packaging, or quality.

The Goal?
Create smarter customer segments based on actual **Purchase Behaviour** rather than just demographic labels.

The Problem: Beyond Demographics

While demographics offer a basic snapshot, they don't always tell the full story. Two people of the same age and income can have completely different shopping habits. That's why CRISA wants to shift toward **Behavioural Segmentation**, which dives into the WHY and HOW of purchases:

- **Who buys frequently?**
- **Who always sticks to one brand?**
- **Who is always hunting for discounts?**

Understanding these patterns will allow companies to design precise marketing strategies, improve customer loyalty, and make smarter use of their promotional budgets.

Diving into the Data: Cleaning and Prepping

Before any smart analysis could begin, we had to ensure the data was clean and usable like clearing the canvas before painting a masterpiece.

Step 1: Missing Values

We scanned the dataset for missing or blank values. A few entries lacked details like socioeconomic class or purchase frequency. These were either filled in using average values (for numbers) or the most common category (for text) or removed if they were too incomplete.

Step 2: Outlier Check

Next, we checked for any unusual values for example, someone who bought soap 100 times a month (which is highly unlikely). These were flagged and carefully reviewed. If they were found to be errors, they were corrected or excluded to maintain realistic insights.

Step 3: Normalizing Key Variables

The dataset had features like:

- **Price Sensitivity**: how likely a person is to be affected by soap prices.
- **Brand Loyalty:** how often they stick to the same brand.
- **Purchase Frequency:** how many times they bought soaps over a period.

Since these features had different ranges (some from 1–5, others from 0–100), we scaled them down to a common level to avoid bias during analysis.

Step 4: Making Categorical Data Usable

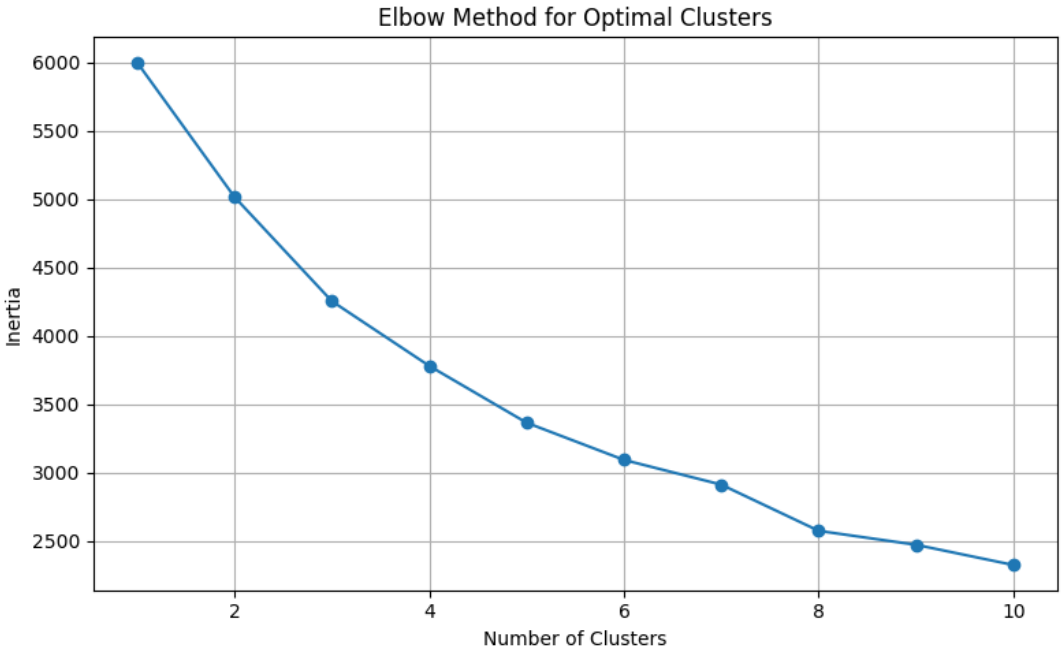
Data like **Gender (male/female)**, **Socioeconomic Class (Upper, Middle, Lower)**, or **Region (North, south, etc.)** were text-based. To allow meaningful analysis, we converted them into numbers using techniques that preserve their meaning but make them digestible for machine learning algorithms.

Segmenting Customers: K-Means Clustering

Now came the exciting part: **Clustering.**

We used a technique called **K-Means Clustering**, a method that groups similar customers together. But the big question was how many groups should we create?

To answer that, we used the **Elbow Method**. Think of it as testing how tightly knit different groupings are. We plotted the results and looked for the "elbow" point where adding more groups didn't lead to much improvement. The ideal number turned out to be **4 Distinct Clusters**.



Elbow Method Graph in the report demonstrates a systematic approach to selecting the right number of customer segments (clusters). It gives stakeholders confidence that the number of clusters chosen for analysis is based on solid mathematical reasoning, rather than arbitrary selection. This graph ensures that the segmentation is both meaningful and data-driven, making it easier to justify the business decisions (such as targeted campaigns or product offerings) based on the final number of customer segments.

Meet the Clusters: Understanding Each Group

Once clustered, we explored each segment's personality like creating customer personas.

Cluster 1: The Loyal Regulars

- **Traits**: High brand loyalty, moderate price sensitivity, regular purchase frequency.
- **Behaviour**: Stick to their favourite soap brand, not too swayed by discounts.
- **Marketing Tip**: Develop loyalty programs or offer premium subscriptions.

Cluster 2: The Bargain Hunters

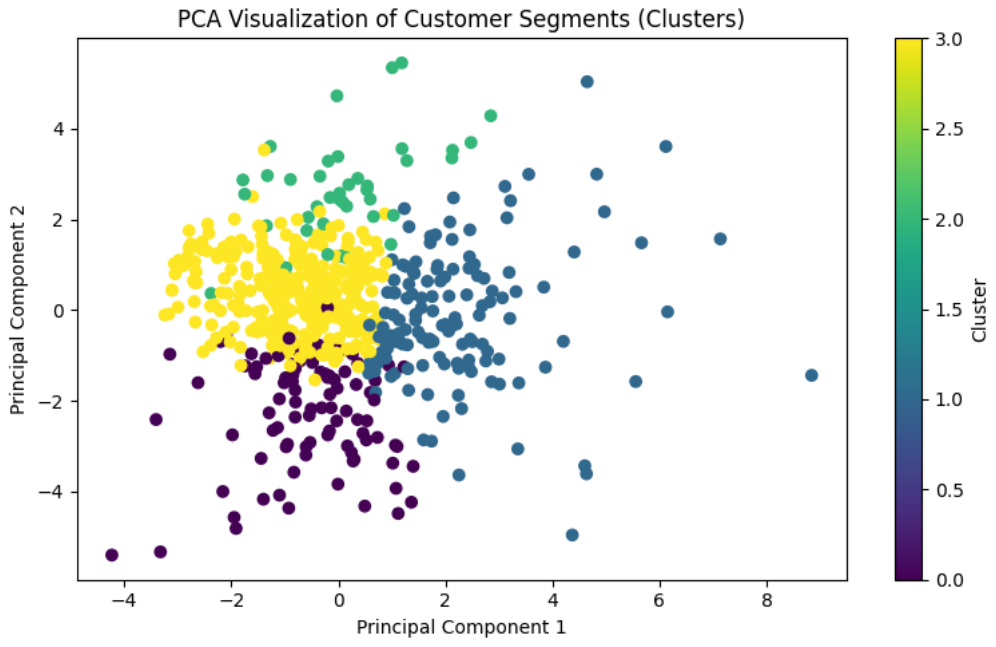
- **Traits**: High price sensitivity, low brand loyalty, buy when there's a deal.
- **Behaviour**: Always on the lookout for offers, switch brands for savings.
- **Marketing Tip**: Target with promotional discounts and limited time offers.

Cluster 3: The Occasional Explorers

- **Traits**: Low purchase frequency, low brand loyalty, curious about new products.
- **Behaviour**: Buy occasionally, often try different brands or new products.
- **Marketing Tip**: Attract them with product innovations, free samples, or bundles.

Cluster 4: The High-Volume Loyalists

- **Traits**: High purchase frequency, high brand loyalty, low price sensitivity.
- **Behaviour**: Buy often and always choose the same brand, regardless of price.
- **Marketing Tip**: Provide exclusive access to products and invite them for brand ambassador programs.

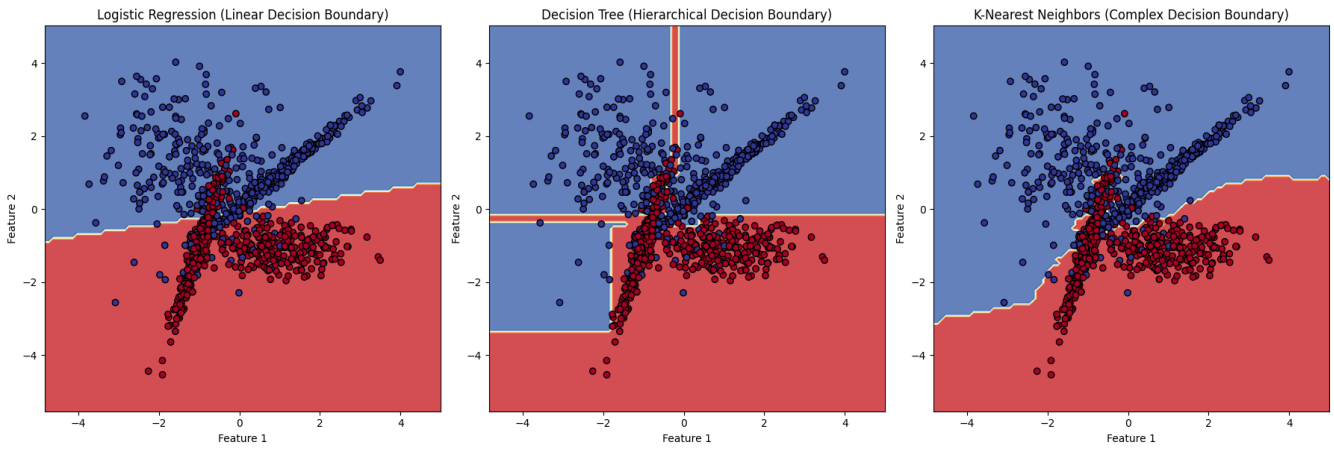


This visualization shows how different customer segments (e.g., price-sensitive, brand-loyal, frequent buyers) are distributed in relation to each other. If the clusters are clearly separated, it confirms that the clustering model has successfully grouped customers with similar traits. If the clusters overlap, it indicates that the model might need adjustments in the number of clusters or other clustering parameters.

Building a Classification Model

Once we understood these segments, the next challenge was predicting where new customers might fit. To do this, we built a Classification Model (a smart system that could look at a new shopper's data and place them into the right group).
We experimented with various models like:

- **Logistic Regression** (A Statistical Model),
- **Decision Trees** (like flowcharts to make decisions), and
- **K-Nearest Neighbours (KNN)** (which groups based on similar behaviours).



Three subplots are created side by side to compare how each model behaves with the same dataset. Each plot represents the decision boundary of a different model.

- **Left Plot**: Logistic Regression (shows a straight line as decision boundary).
- **Middle Plot**: Decision Tree (shows segmented regions).
- **Right Plot**: KNN (shows complex, curved decision boundaries).

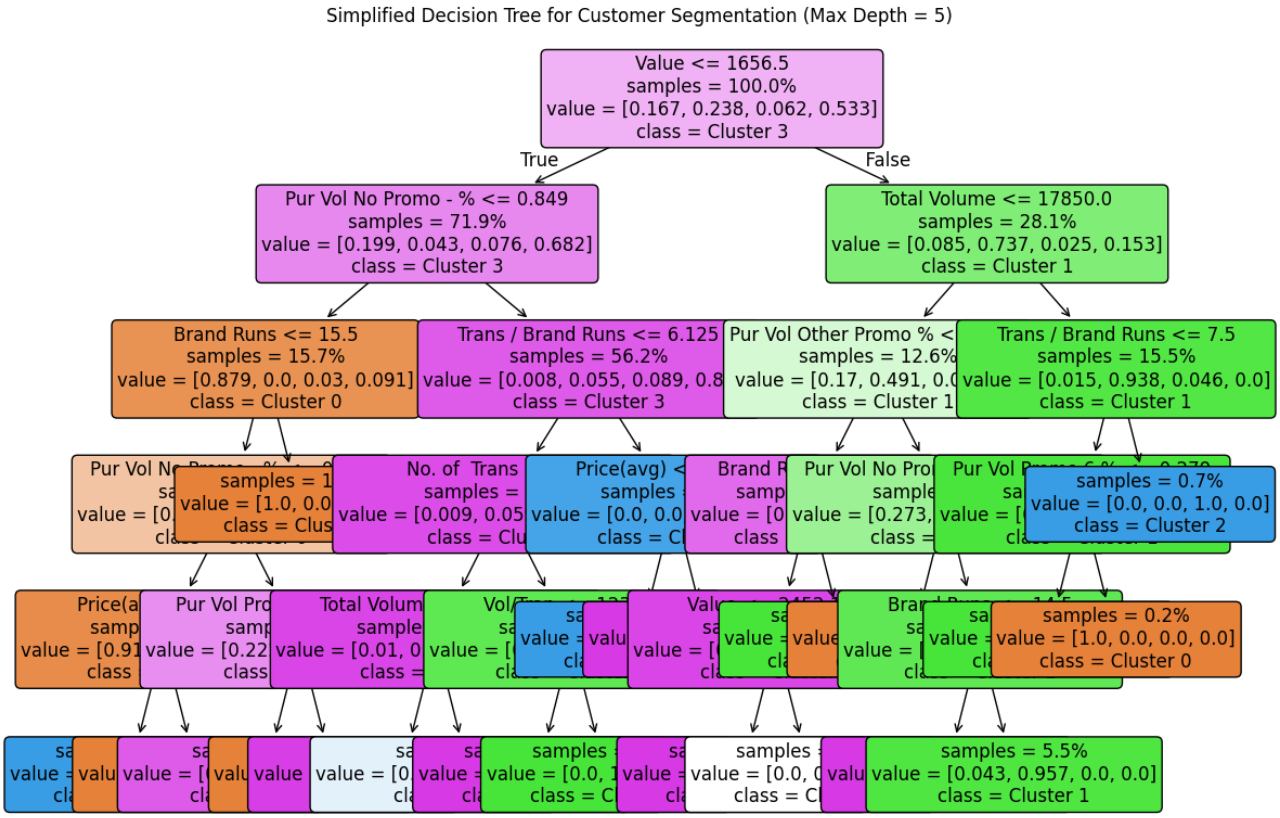
In comparing the three models, Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN), it becomes clear that each model has its strengths and limitations based on the nature of the data and the problem at hand:

Logistic Regression works well when the relationship between the features and target variable is linear. It is a simple and interpretable model, providing a clear decision boundary. However, it struggles with more complex, nonlinear relationships and may not perform well when the data isn't linearly separable.

On the other hand, **Decision Trees** are more flexible and can capture nonlinear relationships in the data. They divide the feature space into distinct regions, making them highly interpretable and useful when the data has clear decision thresholds. While they are prone to overfitting if not properly controlled (e.g., by pruning the tree), they provide a good balance between complexity and interpretability.

K-Nearest Neighbors (KNN) offers the most flexibility by adapting to the local structure of the data. It can model highly complex and irregular decision boundaries, which makes it useful for datasets with intricate patterns. However, KNN can be computationally expensive and sensitive to noise, requiring careful selection of the number of neighbours (K) and proper data preprocessing.

After testing accuracy and performance, **Decision Tree** provided the most understandable and effective results. This model is now ready to classify any future customer, enabling CRISA to scale this segmentation work.



Business Recommendations: Turning Insight into Action

With these insights, CRISA can help businesses move from a one-size-fits-all approach to precision marketing.

For the Brand Loyalists (Clusters 1 & 4):

- Offer **Exclusive Memberships**, early access to new products.
- Send **Personalized Messages** that celebrate their loyalty.
- Bundle products or introduce loyalty rewards.

For the Price-Sensitive Buyers (Cluster 2):

- Focus on **Price Promotions**, discounts, and value packs.
- Highlight **Cost-Saving Benefits** in advertisements.
- Use platforms where deals are frequently promoted (like flyers, coupon apps).

For the Occasional Explorers (Cluster 3):

- Send **Product Recommendations** based on new or trending items.
- Offer **Trial Packs** or **mini sizes**.
- Encourage them through engaging content and storytelling.

Final Thoughts: A New Era of Smart Marketing

Thanks to this data-driven approach, CRISA can now help companies:

- **Increase customer retention** by giving each group what they value most.
- **Improve marketing ROI** by targeting the right offers to the right people.
- **Boost product innovation** by understanding what different customer types of care about.

In the age of personalized experiences, this kind of segmentation is no longer optional, it's essential. CRISA's journey with this bath soap data is just the beginning. With these tools and insights, they're ready to transform how brands connect with their customers one cluster at a time.