Project Dissertation Report on

ANALYSIS ON THE PRESCRIPTIVE PACKAGING FRAMEWORKS IN THE INDIAN DAIRY SECTOR

A CLUSTER-BASED ARTIFICIAL INTELLIGENCE APPROACH

Submitted By

Abhinav Choudhary

Roll No.: 24/DMBA/006

Under the Guidance of

Dr. Apoorva Jain

Delhi School of Management

DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road, Delhi - 110042

May 2026

CERTIFICATE

This is to certify that the project report on the topic - “ANALYSIS ON THE PRESCRIPTIVE PACKAGING FRAMEWORKS IN THE INDIAN DAIRY SECTOR” submitted by Abhinav Choudhary, Roll No. 24/DMBA/006, as part of my major research project for the Master of Business Administration course is original and not disclosed anywhere, and has not been copied or summarized. All the citations have been clearly mentioned and given credits.

Date: _____

Place: Delhi

_____

Signature of the Internal Guide

Delhi School of Management

Delhi Technological University

DECLARATION

I Abhinav Choudhary, Roll No. 24/DMBA/006, M.B.A (DMBA) student of Delhi School of Management, Delhi Technological University, declare that the topic - “ANALYSIS ON THE PRESCRIPTIVE PACKAGING FRAMEWORKS IN THE INDIAN DAIRY SECTOR” is part of my major project report, under the mentorship of Dr. Apoorva Jain, is original. All the citations (books, journal articles, reports, datasets, online resources, etc.) have clearly been mentioned in the Reference section of the project. I declare that the academic integrity has been maintained throughout the process, from approval from the mentor, to the publication of the major project report. To the best of my knowledge and belief, this report contains no matter which has previously been published or written by others except where due reference is made to this in the text.

I also certify that this report in its entirety or in part has not been presented to any other university or institution for any degree, diploma, degree with honors, fellowship or other qualification.

Throughout the report, I have maintained the plagiarism threshold. I have read the consequences and I will be viable for punishment if found guilty against any academic dishonesty.

Date: _____

Place: New Delhi

_____

Abhinav Choudhary

Roll No.: 24/DMBA/006

Master of Business Administration

Delhi School of Management, DTU

ACKNOWLEDGEMENT

The successful completion of the Major Research Project will not be completed with my institution's contribution in providing the right education. I would like to thank my institution, Delhi School of Management, Delhi Technological University for creating the perfect support. I would also like to thank all my teachers and professors for guiding me throughout the journey.

I would also like to thank my mentor, Dr. Apoorva Jain, for her constant support in the process of making this report. Without her guidance, this would not have been possible. From approval of the project to project publication, guiding through every step of the way, her feedback helped me in making course corrections in creating this report.

Gratitude to the report on which this is based upon - Wang, Wang and Cho (2022) — whose conjoint-analysis study of yogurt packaging in Sustainability provided the foundation against which the present analysis was concluded.

I would also like to thank the Kaggle community for bringing in extensive reports. Because of them, thorough analysis was possible.

Finally, heartfelt thanks are extended to family members, classmates and friends, whose encouragement and criticism helped in the creation of the report.

Abhinav Choudhary

Roll No.: 24/DMBA/006

EXECUTIVE SUMMARY

India is the largest dairy producing market, which produces about a quarter of global supply. It contributes nearly 5% to the national economy, directly affecting the farmers and the small dairy owner. This sector is one of the largest women workforce sectors, having nearly 70% of the total workforce contributing to women. The sector assumes a compounded annual growth (CAGR) of about 6.8% a year.

The Indian dairy sector is an example of operational efficiency, having milk collected from farmers and rural dairy owners. The industry is largely dominated by the cooperatives, having Amul and Mother Dairy as one of the largest milk processing units in the world. The country produces about 239.30 million tonnes of milk, with production growing year on year. The White Revolution with the introduction of the project “Operation Flood”, an idea of Dr. Verghese Kurien, India transformed itself from milk-deficient nation to the largest milk producing nation in the world

The industry faces several problems in its way. The industry faces the problems of low cold storage, which is not effective in keeping the total national produce cold. Milk is a perishable item having 5-7 days with a temperature of $0^{o}C$ to $4^{o}C$. Other dairy products also have similar on-shelf conversion. Due to this, cold-storage becomes a necessity rather than an option. 20-25% of the total produce is estimated to be wasted due to the inefficiency in the cold storage in India. Dairy products are highly susceptible to bacteria and fermentation, with bacteria forming instantly when kept at room temperature.

Packaging becomes an important tool in the daily milk and dairy production. It is important to keep dairy products away from the bacteria and away from direct sunlight. Apart from the utility use, the packaging describes the product effectively. It can be used to generate an emotion in the minds of the

people. With the growing urban population and the emergence of tier-2 cities, packaging holds all the more importance.

The Indian dairy segment, due to its bulk nature, contains homogeneous packaging, ignoring the consumer construct in its entirety. The large scale restaurant, preparing thousands of dishes with various dairy products would require a different packaging than a normal consumer. The large scale nature of the industry ignores the end packaging. This is why this research has been done to ensure that the packaging remains in parallel with the efficient operational machinery of the industry, with recommendation and the implementation of the packaging segmentation based on the previous studies done on the packaging materials.

The Wang, Wang & Cho (2022) is taken as the base of the research. The study concluded the utility behaviours of the users on the basis of packaging in the Chinese industry. The study looked at various features which concluded to the emotional and tasteful factors in addition to the product. Four features to doubt during the conjoint analysis were- shape (39.017 percent relative importance), graphics (31.330 percent), label text (15.495 percent) and colour (14.157 percent).

During our study, we used elbow analysis to get the optimum number of clusters for which we would have to segment the total database. The database used was taken from the Kaggle community, named Dairy Good Sales dataset. Various clusters were formed using K Means clustering. Four behaviourally distinct consumer-transaction clusters were identified. Cluster C0, the Everyday Mainstream segment, accounted for 38.2 percent of transactions and was characterised by mid-range volume, low unit prices and a balanced channel mix dominated by Mother Dairy and Amul. Cluster C1, the Institutional and Bulk segment, accounted for 19.0 percent of transactions and was characterised by

very high quantities sold, the highest average revenue per transaction, and a tilt toward retail and wholesale channels. Cluster C2, the Premium Fresh Urban segment, accounted for 31.5 percent of transactions and was characterised by the highest average unit price, low to medium volumes and short shelf life. Cluster C3, the Shelf-Stable Specialty segment, accounted for 11.2 percent of transactions and was dominated by ghee and cheese, with mean shelf life approaching one hundred days.

Perspective analysis done on the basis of earlier studies mentioned in the research to conclude the packaging type used along various clusters. The urban premium sector could use a gabblet of box with a preference in the shape as stated in the previous research having minimalist labels with bold content, optimised for quick commerce like Zepto and Blinkit. The traditional consumers could buy traditional pouches which are efficient in stackability and cost efficiency with warm colour such as Red and gold having concrete graphics, JBL top boxes could respond at a high cost which would be ineffective for the traditional everyday use consumers. Bulk buyers use squared shift containers which would be very useful for stacking up and bulk purchases. Lastly the products with high shelf life can be contained in tin boxes with warm tone graphics to ensure authenticity and richness.

The GABDO model (Goals, Analysis, Build or Buy, Deploy, Optimise) and the TCPR readiness model (Time, Cost, Performance, Requirements) were applied to effectively provide a complete plan for the industry to roll out to end consumers.

The study bridges the gap of descriptive and predictive contribution of the packaging material in the large Indian Dairy organisation. The limitation of the study remains that the previous study was done on a Chinese consumer sample; although Indian context adjustments were added in the study to give more accurate results.

TABLE OF CONTENTS

TITLE	PAGE NO.
Certificate	**i**
Declaration	**ii**
Acknowledgement	**iv**
Executive Summary	**v**
List of Tables	**ix**
List of Figures	
Chapter 1: Introduction	**1**
1.1 Background of the Indian Dairy Sector	1
1.2 Problem Statement	3
1.2.1 Demand-side manifestation	3
1.2.2 Supply-side manifestation	4
1.2.3 Formal problem statement	5
1.3 Objectives of the Study	5
1.4 Scope of the Study	6
1.4.1 Inclusions	6
1.4.2 Exclusions	7
1.5 Significance of the Study	8
1.6 Structure of the Report	8
Chapter 2: Literature Review	**10**
2.1 Overview	10
2.2 Packaging Design and Consumer Purchase Behaviour	10
2.3 Conjoint Analysis and the Wang, Wang and Cho (2022) Study	11
2.3.1 The Conjoint Analysis Method	11
2.3.2 The Wang, Wang and Cho (2022) study	11
2.3.3 Critical considerations for the Indian context	11

TITLE	PAGE NO.
2.4 Unsupervised Learning and Consumer Segmentation	12
2.5 The Indian Dairy Market	12
2.6 Research Gaps	12
Chapter 3: Research Methodology	**14**
3.1 Research Design	14
3.2 Data Sources	14
3.2.1 Primary analytical dataset	14
3.2.2 Behavioural utility benchmark	16
3.3 Analytical Workflow	17
3.4 Algorithmic Specifications	18
3.4.1 K-Means clustering	18
3.4.2 Principal component analysis	18
3.5 Tools and Software	18
3.6 Evaluation Criteria	19
3.7 Ethical Considerations	19
3.8 Limitations of the Methodology	19
Chapter 4: Introduction to the Case	**21**
4.1 The Analytical Setting	21
4.2 The Kaggle Dairy Goods Sales Dataset	21
4.3 The Wang, Wang and Cho (2022) Utility Matrix	22
4.4 Operating Environment	23
Chapter 5: Data Collection and Analysis	**25**
5.1 Data Collection	25
5.2 Data Preparation	25
5.3 Descriptive Statistics	26
5.4 Optimal-k Determination	28
5.5 K-Means Clustering Results	29
Chapter 6: Findings, Discussion and Recommendations	**31**
6.1 Cluster Visualisation	31

TITLE	PAGE NO.
6.2 Cluster Profiles	33
6.2.1 Cluster C0 — Everyday Mainstream segment	33
6.2.2 Cluster C1 — Institutional and Bulk segment	34
6.2.3 Cluster C2 — Premium Fresh Urban segment	35
6.2.4 Cluster C3 — Shelf-Stable Specialty segment	36
6.3 The Prescriptive Packaging Framework	37
6.3.1 Recommendation for Cluster C0 — Everyday Mainstream	37
6.3.2 Recommendation for Cluster C1 — Institutional and Bulk	38
6.3.3 Recommendation for Cluster C2 — Premium Fresh Urban	38
6.3.4 Recommendation for Cluster C3 — Shelf-Stable Specialty	39
6.4 Strategic Implementation — The GABDO Model	39
6.5 Organisational Readiness — The TCPR Model	40
Chapter 7: Conclusion	**42**
References	**45**
Annexure	**46**
Annexure A: Variable Inventory of the Kaggle Dairy Goods Sales Dataset	46
Annexure B: Indicative Python Pseudocode for the K-Means Clustering Workflow	46
Annexure C: Glossary of Selected Terms	47

LIST OF TABLES

Table No.	Caption
Table 4.1	Description of Significant Attribute Levels in Yogurt Packaging (after Wang, Wang and Cho, 2022)
Table 4.2	Relative Importance of Packaging Attributes (Wang, Wang and Cho, 2022)
Table 5.1	Variable Inventory and Treatment in the Kaggle Dairy Goods Sales Dataset
Table 5.2	Elbow and Silhouette Diagnostics for Candidate Values of k
Table 6.1	Profile of Cluster C0 — Everyday Mainstream Segment
Table 6.2	Profile of Cluster C1 — Institutional and Bulk Segment
Table 6.3	Profile of Cluster C2 — Premium Fresh Urban Segment
Table 6.4	Profile of Cluster C3 — Shelf-Stable Specialty Segment
Table 6.5	Prescriptive Packaging Framework by Cluster
Table 6.6	GABDO Implementation Roadmap for the Proposed Framework
Table 6.7	TCPR Readiness Assessment for an Indian Dairy Firm

LIST OF FIGURES

Figure No.	Caption
Figure 4.1	Conjoint-Derived Packaging Attribute Importance (Wang, Wang and Cho, 2022)
Figure 5.1	Product Distribution in the Kaggle Dairy Goods Sales Dataset
Figure 5.2	Optimal-k Diagnostics: Elbow Method and Silhouette Analysis
Figure 6.1	K-Means Clusters Projected onto the PCA-Reduced Feature Space
Figure 6.2	Cluster Size Distribution and Average Revenue and Quantity Sold by Cluster
Figure 6.3	Sales Channel Mix by Cluster

CHAPTER 1
INTRODUCTION

India is the largest producer of milk, contributing to a quarter of the global milk production. It amounts to nearly 5% of the overall GDP contribution. The dairy industries are highly integrated into the rural environment, effectively helping the small scale farmers and rural Dairy owners. India has a very large network of operational milk acquisition and distribution channels, having cooperatives such as Amul and Mother Dairy at the centre of the affairs. The dairy industry supports millions of households and enables small rural economies to be integrated into the large urban formal markets.

India was not a milk sufficient country during the time of partition. It was because of the contribution of Dr Varghese Kurian, the operation flood was launched. Operation flood was also called the White revolution, over 700 cities were integrated during the 1970s to distribute milk and dairy products to every household resulting in fare prices to consumers and an increase in the overall milk production.

The Indian Dairy segment is now entering into a different phase. The industries are constantly evolving with the current times, protein is now being added to the milk and other dairy products to increase the sales in the urban premium market. Various milk based products such as cheese and Greek yogurt are made in-house by various factories, instead of importing from countries like the United States and Mexico.

The market is expanding at the rate of 6.8% compounded annual growth rate (CAGR), with various new distribution channels being introduced in the market. The companies are now moving from traditional distribution channels to now handling door to door sales. Companies like Country Delight directly supply their milk to the end consumer rather than going through various large and small scale distributors. This cuts off the middleman but introduces a new challenge in

the form of distribution and requires a new strategy to be implemented to facilitate the new economy. The rise of quick commerce is also one of the challenges that the traditional large supply industries are looking to tackle. The quick commerce model involves the introduction of having dark stores in multiple locations, facilitating a quick delivery to the end consumers. The consumer segment in this type of economy is restricted to mostly urban and premium users. There's also a rise in the tier-2, experimenting with new alternative choices given by brands. The customers are now seeking healthy and low fat alternatives rather than choosing the traditional dairy products. Protein is also one of the new trends in the urban premium health-conscious industry.

The adoption of the Quick commerce industry like Zepto and Blinkit, also brings out new operational challenges. The customer base that the quick commerce industry is adopting is focused primarily to the tier-1 and tier-2 consumers with different demands than the rural tier-3 area. The industry also demands a package that can be stacked up differently than the kirana store packaging.

Thus, the study takes in account of the scientific and analytical studies on the consumer mindset on the package and takes in account the Indian conditions to come up with conclusions regarding the packaging material used in the industry as segments. A study by Wang, Wang and Cho (2022), published in Sustainability, looked at the yogurt packaging preferences using conjoint analysis. The paper concluded that the Chinese consumer preferences can be categorised into four factors, namely shape, graphics, colour, and label text. The research was concluded on the good audience of 311 respondents. The shape was most preferred among the consumers with nearly 39% importance, followed by graphics (~31%), label text (~15%) and colour (~14%). The respondants gave the importance to Gable-top in shape, concrete graphics in graphics, simple text in label and cool colours in colour. This was obtained by getting a utility value against each of the preferences, for example the consumer gave a utility value to concrete graphics of +0.041 and abstract graphics of -0.026.

positive utility was given to determine that the consumer preferred the type while the negative utility gave the impression that the consumer described the type.

To get an idea of the Indian consumers, the Kaggle dataset was used named Dairy goods sales dataset. It contains 4325 transactions across 15 different Indian states, having 10 product categories. It was expanded to over 11 brands with three sales channels. This was during the time 2019-2022, covering the rise of quick commerce and the modern online transaction channels.

Today the dairy industry follows the descriptive analysis where they take the previous data and find the patterns to come up with conclusions. They do not follow the prescriptive analysis where they take the previous research data to come up with future predictions.

The present research project proposes that the unsupervised clustering combined with the previous research data along with the implementation model such as GABDO and TCPR will help in the designing of most logistically efficient packaging which will also be beneficial in enhancing the preferences among consumers.

1.2 Problem Statement

Our current dairy industry poses a problem of having homogeneous packaging across different verticals, providing a one-size-fits-all solution. With the emergence of new segments and industries, the consumer preferences of packaging along with the logistic inefficiency that the Dairy sector is facing are huge problems that the research work is trying to fix. There is a mismatch between the demand side and the supply side where the need for the packaging is different but the supply of the packaging is the same across all the demand channels.

1.2.1 Demand-side manifestation

The Indian Dairy consumer is a mix of different segments across the country rather than a homogeneous mix. The country is divided into three different tiers,

with different customer segments from having a premium consumer to a rural high volume consumer. The urban consumer, typically living in tier-1 cities like Delhi, Bangalore and Mumbai, have different needs and prefer different packaging and different options of dairy products. They even purchase through a different channel than a high volume traditional consumer living in tier-2 cities. The characteristics of the urban consumer will be that they are willing to pay higher prices, and what consume different premium and aesthetically pleasing products with high nutritional value having cleaner and visually appealing packaging. The traditional high value consumer lives in rural areas in Uttar Pradesh, Rajasthan and Himachal Pradesh, buys daily products when larger quantities, they prefer cooperative brands such as Amul and Sudha for the daily transactions as the companies possess a traditional authenticity and deep cultural impact. The consumer packaging used here would be different than the urban packaging having cows and traditional imagery ensuring trustworthiness and quality. There is also a bulk buyer who does not care about the colour and the text of the dairy product packaging but rather is more focused on the shape of the packaging and how efficiently they can be stacked on top of each other to get more units in a single shipment.

1.2.2 Supply-side manifestation

Since India is the largest producer of the milk product, the company contains a large amount of transactional data and the customer profiles. This posses a huge opportunity for the companies to use predictive and prescriptive analysis to come up with the most efficient and appealing packaging material but the company uses traditional data to do descriptive analysis (sales dashboard and regional comparisons) and performs supervised predictive task to come up with conclusion (demand forecasting and churn prediction). While this may give sense of the customer and what they need, it fails to predict the future trends and emerging sectors to invest in. This is where prescriptive analysis requiring unsupervised learning is of utmost importance. Unsupervised learning gives a larger picture of the emerging trends to be focused upon to sustain in the near

future. Furthermore, we have also observed that the unsupervised learning observation is not integrated with the packaging design or the consumer utility analysis. This gives a structural disconnect between the transactional data analysis and the correlation in the packaging strategy used in the market. This is clearly evident as the Indian Dairy industry really uses heterogeneous packaging for different segments of consumers.

1.2.3 Formal problem statement

The study tries to find how supervised learning techniques combined with utility scores of the conjoint base analysis regarding consumer preferences of packaging be applied to the Indian Dairy sector looking over the constraint and the new demand created by different segments and the introduction of different industries. This will help in reducing the marketing inefficiencies, create a more precise consumer targeting, look over the cold storage conditions of the Indian Dairy sector and finally take an account of the cultural and economic diversities of the Indian consumers.

1.3 Objectives of the Study

The objective of the study is to determine five determinants:

1. To determine and divide the Indian consumers segments into distinct clusters based on K Means clustering analysis. The dataset used from Kaggle, named Dairy Goods Sales Dataset as source of the overall transaction data. The data contains over 4000 datapoints factored over 11 columns giving a clear picture of the consumers.
2. To utilise the clusters derived from the Indian consumers dataset and use those clusters with findings from the study by Wang, Wang and Cho (2022). The study used conjoint analysis to give importance scores across four attributes - Shape, Graphics, Label Text and Color.

3. To give out a definitive prescripted strategy by transforming descriptive findings into recommended actions. This will be done by giving out clear shape, graphic, label text and colour to each segment on the cluster findings
4. Correlating the predictive findings with respect to the Indian cultural and economic environment. The previous research study focuses on Chinese perception by the consumer and it lacks Indian cold storage problems, rise of quick commerce, the emergence of urban premium markets and the traditional bulk purchases of the rural economy.
5. Define a clear framework of task to be done in order to implement the packaging by using the strategic frameworks of artificial intelligence implementation like goals, analysis, build, deploy and optimise (GABDO) and time, cost, performance and requirements (TCPR).

1.4 Scope of the Study

1.4.1 Inclusions

build, deploy and optimise (GABDO) and time, cost, performance and requirements (TCPR).

The analytical territory of the research has the following inclusions:

- Industry scope: it focuses on the Indian dairy sector only, particularly yogurt, paneer, cheese, ice cream, buttermilk, curd, lassi, ghee and packaged milk.
- Geographic scope: the research covers over 15 states in the dataset, which includes Delhi Chandigarh, Karnataka Telangana Tamil Nadu, Uttar Pradesh, Rajasthan, Haryana, Madhya Pradesh, Bihar, Jharkhand, West Bengal Gujarat, Maharashtra and Kerala.
- Data Scope: the study analyses over 4000 transactional data over Kaggle community dataset, the data set is linked with the production, sales, distribution channel, geographical location, brand and lastly the pricing of the Dairy product. Secondary data sources include the

research study conducted by Wang, Wang & Cho to predict the packaging using the utility scores in the study.

- Methodological Scope: the research uses unsupervised learning algorithms, in particular K-Means Clustering, principal component analysis (PCA) for visual diagnostic, elbow analysis and lastly silhouette analysis. Strategic evaluation tools for the implementation of artificial intelligence in business like GABDO and TCPR.
- Design-attribute scope: the packaging analysis is only focused upon the four design attributes that were a result of the previous research study, which include:
    - Shape
    - Graphics
    - Colour
    - Label text

1.4.2 Exclusions

The following are not part of the research project and are not under the scope of analysis made under this project:

- The project does not conduct Indian ko joint experiment but rather realise on the previous research project results to find best packaging for different segments
- The materials of the packaging, the chemistry like:
    - Polymer composition
    - Barrier-layer engineering
    - Chemical migration analysis
    - Environmental life-cycle assessment

  are excluded from the research.
- The frameworks used to implement the unsupervised learning predictions gives out a broader level picture of what needs to be done by the business and does not cover the detailed company level modelling.

- The studies are only related to dairy products and do not cover any other fast moving consumer goods (FMCG) like biscuits, snacks, cold drinks, soaps and other material.
- The study does not include the supervised predictive model like:
    - Demand forecasting
    - Churn prediction

1.5 Significance of the Study

The study offers a link between the two areas: transactional consumer segmentation and the packaging goods material. The consumer segmentation extracted from the data sets are directly applied to the previous study to determine the packaging preferences. The study will also help in the proper implementation of the above predictive analysis via known frameworks. This study also helps in finding the usefulness of using unsupervised learning algorithms and conjoint analysis. This study helps in implementation of the findings without having to spend a lot of money conducting expensive consumer surveys. This study removes the barrier of managerial studies like descriptive analysis to find results and predictive and predictive analysis used academically to find future trends.

1.6 Structure of the Report

The studies follow the guidelines given by the Delhi School of Management, Delhi Technological University and are broken down into different chapters. The report has been broken down into five chapters. Chapter 2 highlights the academic and literature review on the packaging design, conjoint analysis, unsupervised learning and Indian Dairy market. It also analyzes the research gap in the study. Chapter 3 explains the research methodology which includes

the data sources, workflow, algorithm specifications and evaluation criteria. Chapter 4 presents the context of the case which comprises The Kaggle Dairy Goods Sales Dataset, the Wang, Wang and Cho (2022) utility matrix, Descriptive statistics and exploratory analysis. In chapter 5 we are doing the core analytical work. Cluster identification, cluster profiling, utility mapping and the development of prescriptive packaging framework is done in this chapter. The application of the GABDO and TCPR framework is also implemented in this chapter. Chapter 6 concludes the discussion by analysing the key findings, managerial implications, academic contribution, study limitations and future research directions.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

The study connects the four streams - packaging design and the consumer purchase behaviour, conjoint analysis methodology, unsupervised learning mechanisms and the K-Means clustering and Indian Dairy Market consumer data. The study takes in references from the major research study conducted conjoint analysis on Chinese consumers on the basis of Yogurt packaging done by the Wang, Wang and Cho (2022). In this research, unsupervised learning techniques have been adopted to find clusters in the Kaggle community named dairy goods sales dataset. The study finds conclusive packaging along the clusters and the method by which the packaging materials could be implemented easily. The study finally concludes by stating the research gaps and future study scopes.

2.2 Packaging Design and Consumer Purchase Behaviour

Packaging is not nearly a container but rather a taste and psychological consumer decision enhancer. Several empirical research has been made so that these specific observations could be made to prove this conclusion. The research paper Rebollar, Lidón, Gil-Pérez and Marín (2019) states that the images on the text of the packaging material influences customers expectations of sweetness and affect the buying behaviour. Sousa, Carvalho and Pereira (2020) shows that consumers' expectations and their judgment is directly affected by the color and shape of the designs on the packaging material. The Wang, Wang & Cho (2022) directly find the consumer preference of the packaging material by doing conjoint analysis on various attributes such as shape and color. Thus, it can be concluded that the product packaging features

such as shape, color, label and graphics directly affects the users buying perceptions by influencing the cognitive behaviours.

2.3 Conjoint Analysis and the Wang, Wang and Cho (2022) Study

2.3.1 The Conjoint Analysis Method

Conjoint analysis was coined by Lucy and Tukey (1964). It gave a value of preference of attributes among multiple sets of attributes. It works by identifying the key features of the product, assigning variations to each attribute, combining these attributes into a single package asking consumers to read their preferred options. Then the options are rated along the statistical models to get the exact value of each specific feature.

2.3.2 The Wang, Wang and Cho (2022) study

The study applied conjoint analysis to you a good packaging using a sample size of 311 Chinese consumers. The author finds out that the four attributes - shape, size, color and label text, having relative importance scores with each attribute having the most preferred combination such as concrete graphics for graphics attribute, cool colour for colour attribute, gable-top shape for shape and lastly simple label text.

2.3.3 Critical considerations for the Indian context

The elements could not be directly applied without considering the Indian context and the Indian Dairy consumer base. The industry faces a large problem of cold-storage, India being a largely warm country has to implement a more safe packaging while keeping the marketing consumer preferences. Moreover with the rise of different platforms such as quick commerce the packaging needs to change according to the preferences of the urban premium needs. Furthermore, the packaging such as gable top prove more expensive than the traditional Dairy pouches and can affect the rural low per capita income households.

2.4 Unsupervised Learning and Consumer Segmentation

The study uses unsupervised learning algorithms to find correlation and prescriptive analysis between the two studies. The study uses the K-Means clustering algorithm which divides the data set into k mutually exclusive feature clusters. These clusters have a common centroid which is the mean of the observations of the data points inside the cluster. The number of clusters is decided by the elbow method which examines the sum of squares as a function of k and identifies the point at which rate of decrease or the elbow supply decreases. The study also uses silhouette coefficient which evaluates how observed clusters fit closer to the relative next cluster.

2.5 The Indian Dairy Market

The Indian Dairy market is slightly different from the Chinese values observed in the research paper. The compounded annual growth rate for the Indian Dairy sector is about 6.8%. India possesses a different risk having cold chain weakness resulting in waste of 20-25% loss for perishable goods like milk, cream, yogurt and cheese. The market is constantly evolving with the onset of quick commerce and the rise of the healthy market industry. The importance of protein and whey protein have been significant in the minds of the urban consumer. They now resemble the most European and American counterparts. While India is seeing this growth, the rural sector is still in its rudimentary state having a large traditional segment with different cultural orientation. There are other observable facts that the restaurants now want different packaging then the consumers get. All of these poses are a huge challenge in the segment and require a thorough predictive and prescriptive analysis to curb the demands.

The study provides a comprehensive literature over the marketing and the proper implementation of the recommendation there are few gaps that the future research holds:

First, the conjoint analysis derived from the yogurt packaging by the Chinese researchers has taken into account the Chinese factors. The conjoint analysis method is present, but it is not applied to the Indian data. Output needs to be mapped with the Indian data to provide a more comprehensive user base study of the Indian output taking into account the Indian challenges.

Second, the unsupervised clustering observations are not applied on the final consumers side of the industry. Algorithms like this have already been implemented on the supply side of the chain such as farmers productivity rather than consumer facing problems like the packaging design.

Third, the Indian Dairy packaging is largely descriptive analysis and focuses on more bulk supply and one size fits all models.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Research Design

The study focuses on qualitative analysis of the transactional data among the Indian consumers and matching them with the secondary behavioural data conducted by various researchers. It is an exploratory approach because the nature of the consumers segment and the number of Total transactions in the Indian Dairy market are not pre specified. Secondly the study does a prescriptive analysis on the segments mapped by the clustering and map set along the packaging design recommendation by the conjoint analysis done by the researchers Wang, Wang and Cho (2022).

That study uses two approaches - it uses conjoint analysis to find the consumer value packaging features and it uses unsupervised learning algorithm K-Means clustering to group the transactional patterns.

3.2 Data Sources

3.2.1 Primary analytical dataset

The primary analytical dataset is the Kaggle Dairy Goods Sales Dataset, a publicly available dataset of dairy transactions contributed to the Kaggle data-science platform. The dataset has 4,325 transactional records and 23 fields, which include Location, Total Land Area, Number of Cows, Farm Size, Product Identifier, Product Name, Brand, Quantity in litres or kilograms, Price per Unit, Total Value, Shelf Life in days, Storage Condition, Production Date, Expiration Date, Quantity Sold, Price per Unit Sold, Approximate Total Revenue in Indian rupees, Customer Location, Sales Channel, Quantity in Stock, Minimum Stock Threshold and Reorder Quantity.

Sample JSON:

```
[
  {
    "Location": "Telangana",
    "Total Land Area (acres)": 310.84,
    "Number of Cows": 96,
    "Farm Size": "Medium",
    "Date": "2022-02-17",
    "Product ID": 5,
    "Product Name": "Ice Cream",
    "Brand": "Dodla Dairy",
    "Quantity (liters/kg)": 222.4,
    "Price per Unit": 85.72,
    "Total Value": 19064.128,
    "Shelf Life (days)": 25,
    "Storage Condition": "Frozen",
    "Production Date": "2021-12-27",
    "Expiration Date": "2022-01-21",
    "Quantity Sold (liters/kg)": 7,
    "Price per Unit (sold)": 82.24,
    "Approx. Total Revenue(INR)": 575.68,
    "Customer Location": "Madhya Pradesh",
    "Sales Channel": "Wholesale",
    "Quantity in Stock (liters/kg)": 215,
    "Minimum Stock Threshold (liters/kg)": 19.55,
    "Reorder Quantity (liters/kg)": 64.03
  },
  {
    "Location": "Uttar Pradesh",
    "Total Land Area (acres)": 19.19,
    "Number of Cows": 44,
```

```
    "Farm Size": "Large",
    "Date": "2021-12-01",
    "Product ID": 1,
    "Product Name": "Milk",
    "Brand": "Amul",
    "Quantity (liters/kg)": 687.48,
    "Price per Unit": 42.61,
    "Total Value": 29293.5228,
    "Shelf Life (days)": 22,
    "Storage Condition": "Tetra Pack",
    "Production Date": "2021-10-03",
    "Expiration Date": "2021-10-25",
    "Quantity Sold (liters/kg)": 558,
    "Price per Unit (sold)": 39.24,
    "Approx. Total Revenue(INR)": 21895.92,
    "Customer Location": "Kerala",
    "Sales Channel": "Wholesale",
    "Quantity in Stock (liters/kg)": 129,
    "Minimum Stock Threshold (liters/kg)": 43.17,
    "Reorder Quantity (liters/kg)": 181.1
  }
]
```


3.2.2 Behavioural utility benchmark

The behavioural utility benchmark is the conjoint-derived utility matrix reported by Wang, Wang and Cho (2022) in the journal Sustainability. The study applied conjoint analysis to you a good packaging using a sample size of 311 Chinese consumers. The author finds out that the four attributes - shape, size, color and label text, having relative importance scores with each attribute having the most

preferred combination such as concrete graphics for graphics attribute, cool colour for colour attribute, gable-top shape for shape and lastly simple label text.

3.3 Analytical Workflow

The analytical workflow contains the following workflow:

1. Data preparation: the raw data step is transformed by handling the missing values, type inconsistency and data outliers. The date fields are converted into standard formats, categorical values are converted into numerical values and the continuous variables are standardized using z score normalisation for further clustering of data.
2. Optimal cluster determination: two methods are used in order to compute the optimal K clusters which balances the marginal reduction within cluster sum of squares against the interpretability of the result cluster structure while keeping the right coefficient silhouette coefficient. The methods used were elbow method and silhouette coefficient.
3. K-Means clustering: the data set is broken down into K clusters, the centroids are recorded and each transaction is labeled accordingly.
4. Cluster Profiling: each cluster is profiled against columns like the location, dominant products, dominant brand, dominant sales channel, average quantity sold, average revenue, average per price unit and average self life. The profiling is then changed into personas for further analysis.
5. Utility mapping and framework construction: Each cluster is mapped against the Wang, Wang and Cho (2022) by processing it against the Indian context considering price sensitivity, cold-chain robustness and quick-commerce handling. The output is then given in the form of prescriptive analysis with cluster-specific recommendations on shape, colour, graphic style and label text. Models like GABDO and TCPR frameworks are applied to evaluate the strategic implementation and the organisational readiness.

3.4 Algorithmic Specifications

3.4.1 K-Means clustering

K-Means clustering algorithm which divides the data set into k mutually exclusive feature clusters. These clusters have a common centroid which is the mean of the observations of the data points inside the cluster. The number of clusters is decided by the elbow method which examines the sum of squares as a function of k and identifies the point at which rate of decrease or the elbow supply decreases. The study also uses silhouette coefficient which evaluates how observed clusters fit closer to the relative next cluster.

The feature set used for clustering has five variables: Quantity Sold in litres or kilograms, Price per Unit Sold in Indian rupees, Approximate Total Revenue in Indian rupees, Shelf Life in days and an ordinal encoding of Sales Channel in which Online is coded as 2, Retail as 1 and Wholesale as 0. All variables are standardized to ensure that no variable stands out as a dominant variable in the K means clustering.

3.4.2 Principal component analysis

Principal Component Analysis (PCA) was applied to standardize the dataset onto two-dimensional visualization of clusters. The original data set contains 5 dimensions on which the clusters were formed, PCA reduces it to two dimensional components while preserving variance as much as possible. It is used only for the visualisation purposes as the K means clustering is still performed along the five dimensional standardized set rather than a two dimensional structure.

3.5 Tools and Software

Python was used to analyse the data set, having software libraries to support the unsupervised learning. For the data preprocessing, pandas libraries was

used, scikit-learn library was used for the K means clustering principle component analysis, Silhouette coefficient and lastly matplotlib library was used to demonstrate the graphs and clustering patterns. Initial data was visualized under the Microsoft Excel software.

3.6 Evaluation Criteria

The evaluation criteria was analyse using the cluster quality which was determined by silhouette coefficient and elbow method by using within cluster sum of squares (WCSS). High silhouette coefficient define better cluster separation and well defined consumers segments, WSCC determines how tightly data points are grouped together. Data was also evaluated on the basis of how they perform under the categorization of segments using other data points.

3.7 Ethical Considerations

The Kaggle dataset is publicly available and it does not contain any personal information of any user. It was taken up as a secondary data set and no human level interaction has been made during the research project.

3.8 Limitations of the Methodology

There are multiple limitations to the study, the first being the Kaggle data set was used as its secondary data set and no primary collected data was taken during the whole research process. These processes are at risk of high generalization. The research also does not consider the Indian sample but takes in the Chinese consumer sample and price to evolve within the Indian cultural adjustments. Even though that Indian context has been considered, the transferability issue cannot be fully eliminated. K means clustering generalizes the clusters and does not represent each individual value but rather a group of

cluster profiles with comparable density. The study recommends that the future study holds more Indian context with conjoint analysis which gives a more detailed and accurate result.

CHAPTER 4
INTRODUCTION TO THE CASE

4.1 The Analytical Setting

The study is intended for the Indian Dairy Sector companies and takes in the marketing perspective to rely on different consumer segments. The study takes in account that the companies use standardized packaging to supply different consumer segments and does not account the consumer choices and behaviours. It introduces segment specific packaging that will be based on the prescriptive analysis over the dataset. The dataset used is not company specific but taken in the broader picture of the whole industry across multiple states. This gives a generalised strategy that could easily be applied across the nation. It also analyzes the strategic implementation of the recommendations over different frameworks. The study can be taken by dairy companies to optimize consumer segmentation, packaging personalization, premium positioning and better alignment with regional consumer behaviour.

4.2 The Kaggle Dairy Goods Sales Dataset

Kaggle Dairy Goods Sales Dataset. This is a public dataset of Indian Dairy transactions. The dataset comprises 4,325 transactions across fifteen states, ten product categories, eleven brands and three sales channels. Each record consists of 23 fields grouped in five thematic groups: farm characteristics (Total Land Area, Number of Cows, Farm Size), product characteristics (Product Name, Brand, Quantity, Shelf Life, Storage Condition), pricing characteristics (Price per Unit, Total Value), sales characteristics (Quantity Sold, Price per Unit Sold, Approximate Total Revenue, Customer Location, Sales Channel) and inventory characteristics (Quantity in Stock, Minimum Stock Threshold, Reorder Quantity).

The list of variables and the treatment for each are listed in table 5.1 in chapter 5. The distribution of products in the dataset is shown in Figure 5.1. The distribution of products in the dataset is almost balanced in the ten product categories. The maximum number of records are of curd, lassi, paneer and yogurt.

4.3 The Wang, Wang and Cho (2022) Utility Matrix

The Wang, Wang and Cho (2022) conjoint study offers the behavioral utility matrix that supports the prescriptive part of this analysis. Table 4.1 shows the significant attribute levels in yogurt packaging from the study, and Table 4.2 displays the relative importance values for each attribute. Figure 4.1 illustrates the relative importance of the four attributes.

Attribute	Level Description	Level No.
Graphic	Mainly abstract	1
	Mainly concrete (photographic)	2
	Mainly text	3
Packaging Colour	Cool (blue, green, etc.)	1
	Warm (red, yellow, etc.)	2
Packaging Shape	Square-bag	1
	Bowl	2
	Bottle	3
	Gable-top	4
Label Text	Complex	1
	Simple	2

Table 4.1: Description of significant attribute levels in yogurt packaging (after Wang, Wang and Cho, 2022)

Source: Adapted from Wang, Wang and Cho (2022), Sustainability, 14(6), 3463.

Packaging Attribute	Relative Importance (%)	Globally Preferred Level
Packaging Shape	39.017	Gable-top
Graphic	31.330	Concrete / photographic
Label Text	15.495	Simple
Packaging Colour	14.157	Cool

Table 4.2: Relative importance of yogurt packaging attributes (Wang, Wang and Cho, 2022)

Source: Adapted from Table 4 of Wang, Wang and Cho (2022).

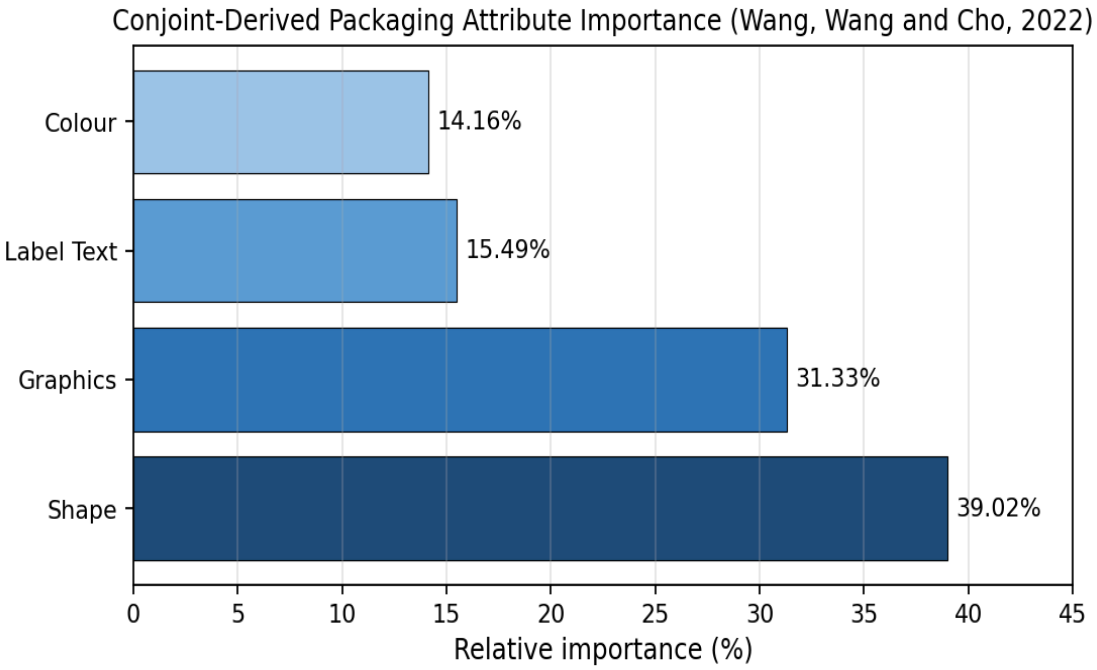

Figure 4.1: Conjoint-derived packaging attribute importance (Wang, Wang and Cho, 2022)

Source: Own creation based on Wang, Wang and Cho (2022).

4.4 Operating Environment

Presently the Indian Dairy market has four features:

- Consumers have two types, one who are premium and urban segment who buy their products from online shopping and realize quick commerce

for their Dairy needs. Secondly there are traditional segments which rely on retail and wholesale stores.

- India's cold chain storage is a sign of weakness as it is predicted that about 20-25% of the overall milk production is wasted.
- Quick commerce platform stores perishable dairy products at high temperature and handling at the last mile delivery becomes extremely important.
- Lastly nutritional labelling is now becoming popular with the emergence of health based consumers.

CHAPTER 5
DATA COLLECTION AND ANALYSIS

5.1 Data Collection

The main data for analysis in this study was taken from the Kaggle data-science platform, with the title “Dairy Goods Sales Dataset”. The data set is released under a permissive license for non-commercial academic use and contains no personally identifiable information. The data set was downloaded as a comma separated value format and loaded into a python analytical environment using the pandas library.

The behavioral utility data used in the prescriptive layer of the analysis were extracted from the open-access journal article by Wang, Wang and Cho (2022) published in Sustainability, which is an MDPI-indexed journal with a documented peer-review process.

5.2 Data Preparation

The data set was evaluated on the basis of four parameters. First the data set was analyse for missing values in the 5 attributes - quantity sold, price per unit sold, approximate total revenue, shelf life and sales channel. It was evaluated across all 4325 records. Next the date given in the data set was transformed into a standard date. Third categorical values were transformed for descriptive and predictive analysis. Lastly the five categorical attributes were processor under K means clustering. All the data set was standardized to ensure that the data does not give abrupt results.

Variable	Type	Range / Categories	Treatment
Location	Categorical	15 Indian states	Descriptive
Product Name	Categorical	10 product categories	Descriptive
Brand	Categorical	11 brands	Descriptive
Sales Channel	Categorical (ordinal)	Online / Retail / Wholesale	Ordinal-encoded
Quantity Sold (l/kg)	Continuous	1 – 960; mean 248.1	Standardised
Price per Unit Sold (₹)	Continuous	Variable; mean ~₹55	Standardised
Approx. Total Revenue (₹)	Continuous	12.5 – 89,108; mean 13,580	Standardised
Shelf Life (days)	Continuous	Variable; mean ~28	Standardised

Table 5.1: Variable inventory and treatment in the Kaggle Dairy Goods Sales Dataset

Source: Own analysis of the Kaggle Dairy Goods Sales Dataset (n = 4,325).

5.3 Descriptive Statistics

The dataset is approximately balanced across the ten product categories. Curd accounts for 479 records (11.1 percent), lassi for 447 records (10.3 percent), paneer for 441 records (10.2 percent), yogurt for 437 records (10.1 percent), buttermilk for 435 records (10.1 percent), butter for 431 records (10.0 percent), milk for 429 records (9.9 percent), ice cream for 423 records (9.8 percent), ghee for 402 records (9.3 percent) and cheese for 401 records (9.3 percent). Figure 5.1 presents the product distribution graphically.

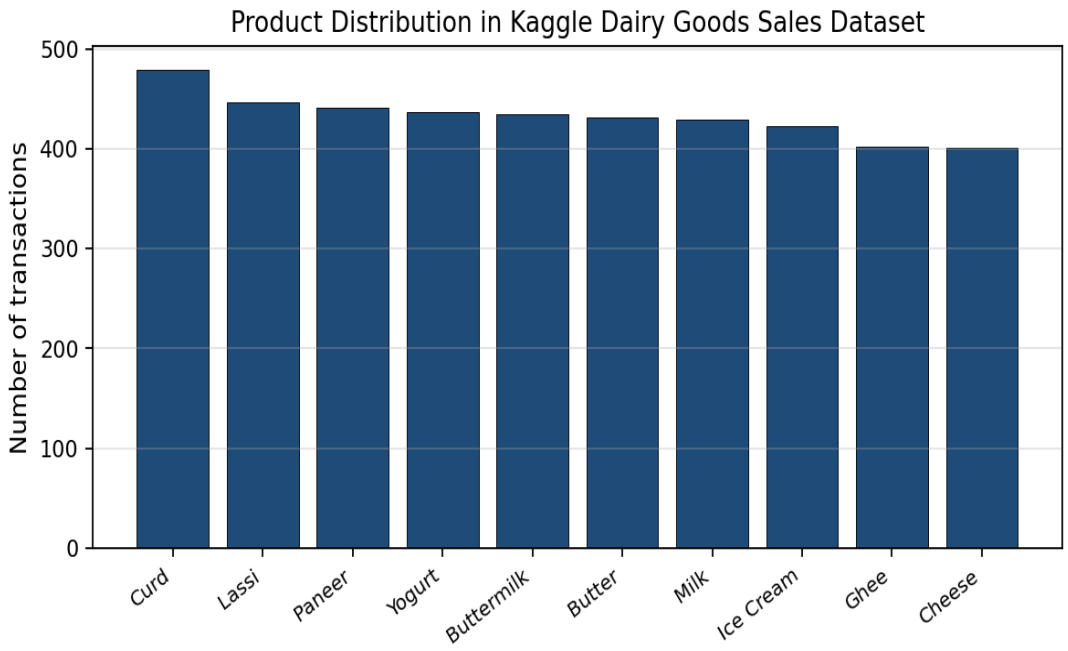

Figure 5.1: Product distribution in the Kaggle Dairy Goods Sales Dataset

Source: Own analysis using matplotlib (n = 4,325).

The brand-level distribution is also represented as:

- Amul 1,053 records
- Mother Dairy 1,010 records
- Raj 685 records
- Sudha 648 records

And so on

The sales channels distribution are:

- Retail - 1,478 records, 34.2% of the total distribution.
- Wholesale - 1,476 records, 34.1% of the dataset.
- Online sales - 1,371 records, 31.7% overall.

The state-wise distribution is highly concentrated in Delhi and Chandigarh:

- Delhi - 525 records (highest contribution)
- Chandigarh - 519 records.

- The remaining thirteen states each contribute between 223 and 276 records, indicating a comparatively even distribution among them.

The continuous variables show substantial dispersion across the dataset.

- Quantity Sold ranges from 1 to 960 units.
- The mean Quantity Sold is 248.1 units with a standard deviation of 217.0 units.
- Approximate Total Revenue ranges from ₹12.54 to ₹89,108.90.
- The mean Approximate Total Revenue is ₹13,580.27.
- The standard deviation of Approximate Total Revenue is ₹14,617.01.

5.4 Optimal-k Determination

The number of clusters taken for Elbow method and the Silhouette coefficient having K ranging from two to seven.

k	Inertia (Within-Cluster SS)	Silhouette Coefficient
2	15,926.4	0.2865
3	13,092.1	0.3008
4	**10,815.9**	**0.2479**
5	9,303.0	0.2513
6	8,291.6	0.2506
7	7,497.7	0.2473

Table 5.2: Elbow and Silhouette diagnostics for candidate values of k

Source: Own analysis using scikit-learn KMeans on the standardised feature matrix (random_state = 42, n_init = 10).

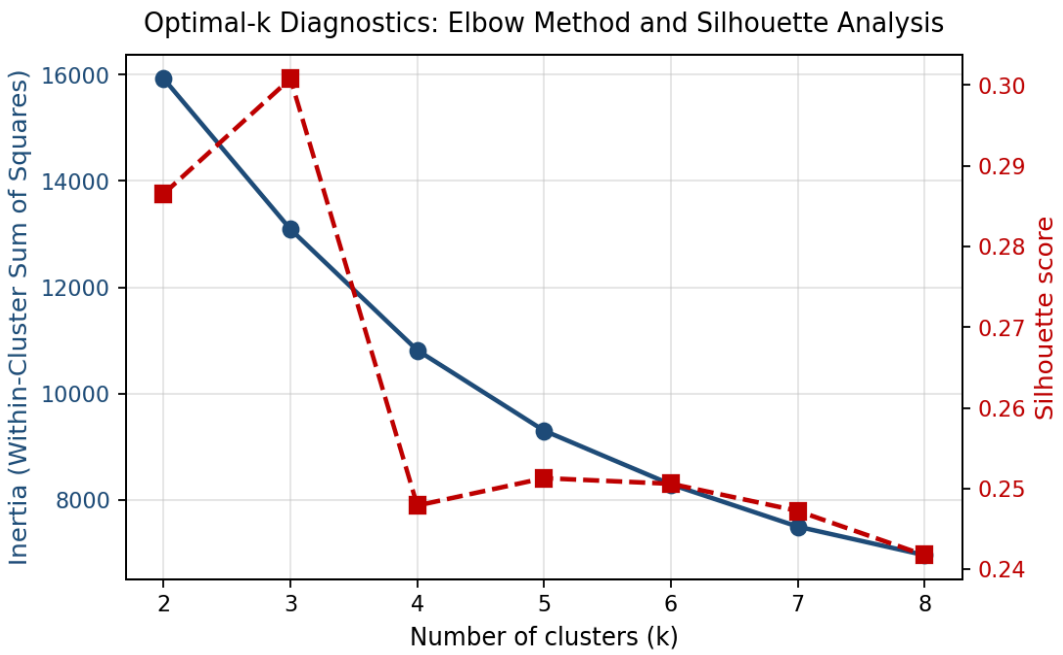

Figure 5.2: Optimal-k diagnostics — Elbow method and Silhouette analysis

Source: Own creation using matplotlib.

The inertia curve shows a clear inflection point between k = 3 and k = 4. Beyond k = 4, the marginal reduction in inertia begins to taper off. The Silhouette coefficient reaches its highest value at k = 3 (0.3008). The Silhouette coefficient decreases slightly at k = 4 (0.2479). A small recovery in the Silhouette coefficient is observed at k = 5 (0.2513). Despite the slightly higher Silhouette value at k = 3, k = 4 was selected for the analysis. The k=4 value was selected as it better categorized the clusters on the basis of cluster profiling. The shift from k = 3 to k = 4 results in a 17.4% reduction in within-cluster sum of squares. The four-cluster solution is managerially interpretable and aligns with practical consumer segmentation. It was selected to attain a more marketing interpretable stance.

5.5 K-Means Clustering Results

The K-Means clustering was applied along 4 clusters to identify the cluster formed and to further analyse. The 5 attributes were standardized to prevent any abrupt values. The model underwent 0 random initialisations, which means

the algorithm was executed twenty times with different starting centroids to reduce the risk of poor local solutions and improve clustering reliability. A random seed value of 42 was used to ensure reproducibility of results, allowing the same clustering output to be generated again under identical conditions. The complete dataset of 4325 transactions were transformed into 4 clusters. The original data set contains 5 dimensions on which the clusters were formed, PCA was used to reduce it to two dimensional components while preserving variance as much as possible.

- **Cluster C0** (Everyday Mainstream): 1,654 transactions (38.2 percent of the dataset).
- **Cluster C1** (Institutional and Bulk): 822 transactions (19.0 percent of the dataset).
- **Cluster C2** (Premium Fresh Urban): 1,364 transactions (31.5 percent of the dataset).
- **Cluster C3** (Shelf-Stable Specialty): 485 transactions (11.2 percent of the dataset).

Chapter 6 presents the full profile of each cluster (average quantity sold, average revenue per transaction, average price per unit sold, average shelf life, dominant products, dominant brands and dominant channels and dominant states) along with the prescriptive packaging framework.

CHAPTER 6
FINDINGS, DISCUSSION AND RECOMMENDATIONS

6.1 Cluster Visualisation

The four k-means clusters are represented along two dimensional space using principal component analysis are shown in the figure. The two leading components have the most variance in the features phase and provide visual diagnostic of the cluster structure. The figure also indicates that the 4 clusters are separated in different parts of the features space having very little overlap providing separability as predicted by the silhouette coefficient.

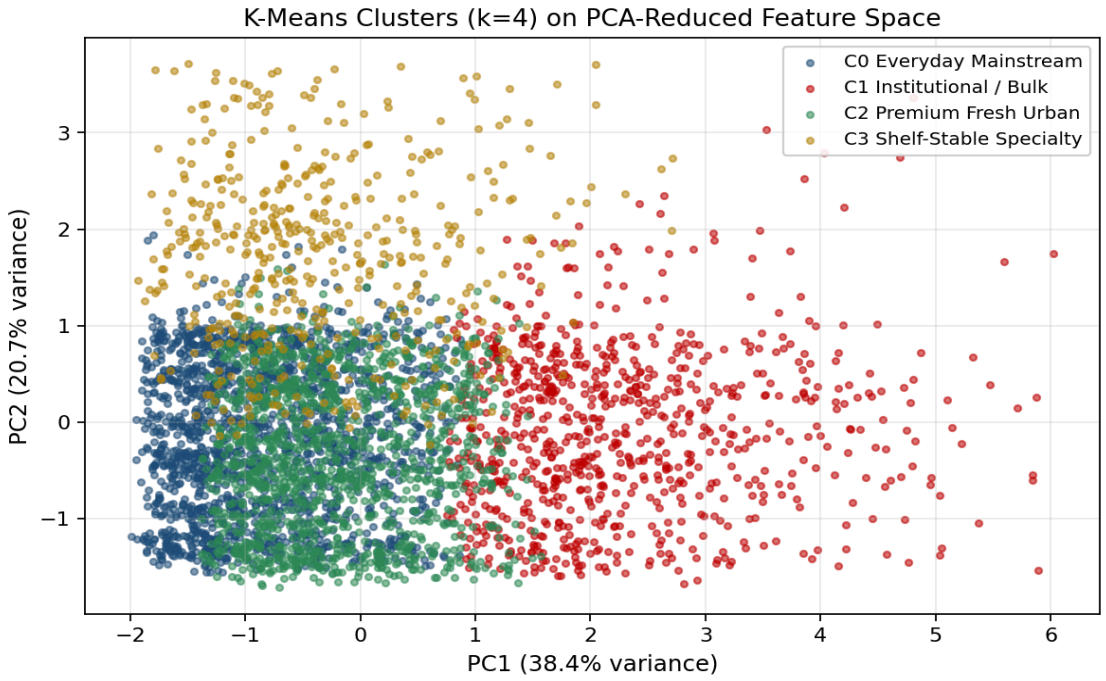

Figure 6.1: K-Means clusters (k = 4) projected onto the PCA-reduced feature space

Source: Own creation using scikit-learn and matplotlib (n = 4,325).

The next bar graph is complementary to the cluster structure. The size of the cluster on the left panel and the comparative profile of average revenue and

average quantity sold on the right panel is shown in the figure. From the chat we can see that the C0 is the largest in terms of the number of transactions while C1 is the largest segment by far in terms of the revenue per transaction and the physical volume per transaction.

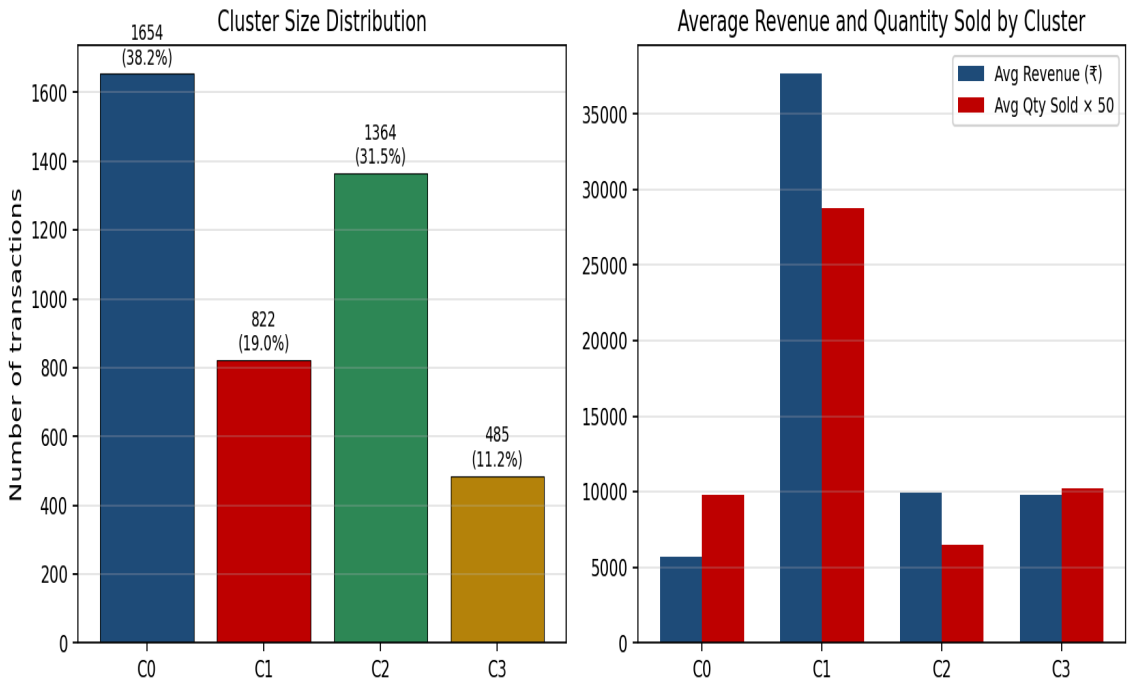

Figure 6.2: Cluster size distribution and average revenue and quantity sold by cluster

Source: Own creation using matplotlib (n = 4,325).

Figure 6.3 shows the sales channel mix by each cluster. It shows that the cluster C3 has the highest wholesale share, while the other cluster shows equal or nearly equal wholesale share. This shows that the modern Indian Dairy market is interconnected in terms of the sales channel rather than being isolated. For example earlier products were categorised and based on that different channels were adopted like premium products through details store, bulk product through wholesale.

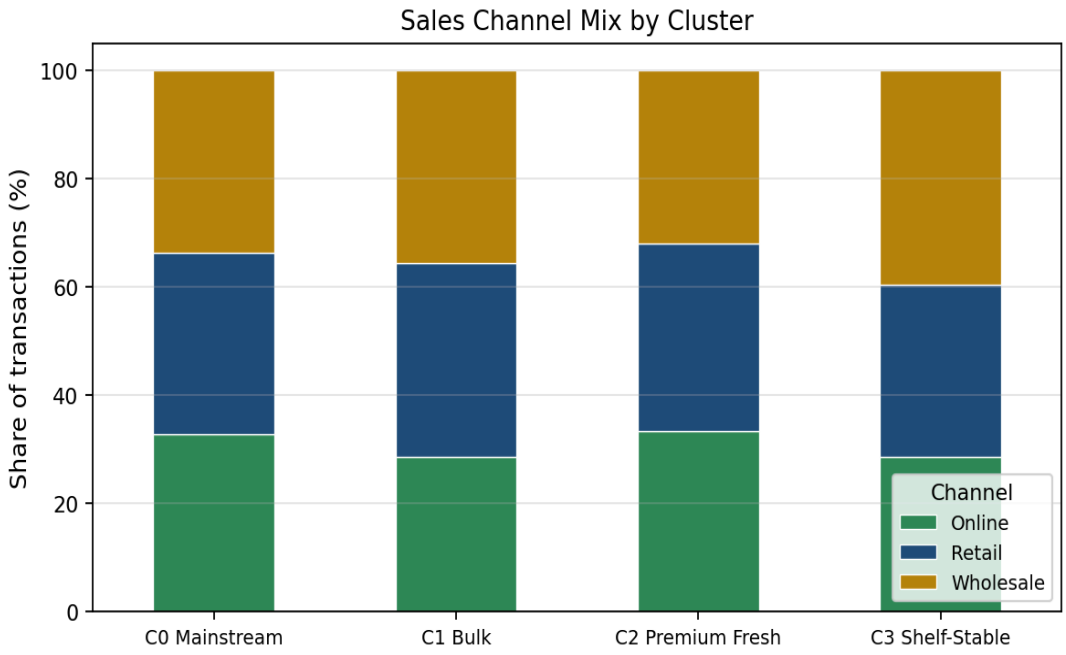

Figure 6.3: Sales channel mix by cluster

Source: Own analysis using matplotlib (n = 4,325).

6.2 Cluster Profiles

6.2.1 Cluster C0 — Everyday Mainstream segment

The largest segment in terms of the number of transactions is Cluster C0 which consists of 1,654 transactions or 38.2 percent of the dataset. This segment is characterized by a mid-range quantity sold (mean 195.6 units), low to mid-range revenue per transaction (mean ₹5,691), low unit prices (mean ₹30.75), short shelf life (mean 19.2 days) and a near balanced channel mix, with all three sales channels contributing roughly one third of transactions each. The product mix is dominated by curd, yogurt and ice cream while the brand mix is dominated by Mother Dairy and Amul, each with about 399 records. The geographic distribution is led by Delhi (206 records) and Chandigarh (190 records), with secondary representation in Rajasthan, Gujarat and Telangana. This is the mainstream Indian dairy transaction: a regular-volume, regular-price purchase of fresh dairy from a legacy national brand.

Cluster name	Everyday Mainstream
Transactions	1,654 (38.2% of dataset)
Average quantity sold	195.6 units
Average revenue per transaction	₹5,690.8
Average price per unit	₹30.75
Average shelf life	19.2 days
Dominant channels	Wholesale (33.6%), Retail (33.6%), Online (32.8%)
Dominant products	Curd, Yogurt, Ice Cream
Dominant brands	Mother Dairy, Amul, Sudha
Top geographies	Delhi, Chandigarh, Rajasthan, Gujarat, Telangana

Table 6.1: Profile of Cluster C0 — Everyday Mainstream segment

Source: Own analysis using scikit-learn KMeans on the Kaggle Dairy Goods Sales Dataset.

6.2.2 Cluster C1 — Institutional and Bulk segment

The segment with the highest economic intensity per transaction is cluster C1, which comprises 822 transactions, or 19.0 percent of the dataset. The mean quantity sold is 576.1 units, almost thrice the dataset mean and the mean revenue per transaction is ₹37,657 – 2.7 times the dataset mean. The average unit price of ₹67.68 is the second highest in the dataset and indicates that the segment is not just buying in bulk, but is buying mid-premium products in bulk. The channel mix is tilted towards retail (35.9 percent) and wholesale (35.5 percent) which combine to account for 71.4 percent of transactions. The product mix is led by curd, milk and lassi and the brand mix is led by Amul, Mother Dairy and the regional cooperative Raj. The geographic distribution is more scattered than that of Cluster C0 with the highest shares in Chandigarh and Delhi followed by Tamil Nadu, Maharashtra and Uttar Pradesh. This is the institutional and bulk transaction, a high volume, mid-premium purchase, typically made through retail or wholesale channels.

Cluster name	Institutional and Bulk
Transactions	822 (19.0% of dataset)
Average quantity sold	576.1 units
Average revenue per transaction	₹37,657.1
Average price per unit	₹67.68
Average shelf life	24.0 days
Dominant channels	Retail (35.9%), Wholesale (35.5%), Online (28.6%)
Dominant products	Curd, Milk, Lassi
Dominant brands	Amul, Mother Dairy, Raj
Top geographies	Chandigarh, Delhi, Tamil Nadu, Maharashtra, Uttar Pradesh

Table 6.2: Profile of Cluster C1 — Institutional and Bulk segment

Source: Own analysis using scikit-learn KMeans on the Kaggle Dairy Goods Sales Dataset.

6.2.3 Cluster C2 — Premium Fresh Urban segment

The C2 cluster has 1,364 transactions, accounting for 31.5 percent of the dataset. The average quantity sold of 129.5 units is the lowest in the data set and the average unit price of ₹76.97 is the highest. The average revenue per transaction at ₹9,974 suggests a low volume, high price purchasing pattern consistent with premium-fresh consumption. The segment is buying perishables as evidenced by the low average shelf life of 19.1 days. The mix of channels is roughly even. The product mix is led by buttermilk, butter and curd and the brand mix is dominated by Mother Dairy, Amul and Raj. The geographic distribution centres on Delhi (167 records) and Chandigarh (166 records), with secondary representation in Uttar Pradesh, Karnataka and Haryana. This is the premium urban transaction: a low-volume, high-price purchase of fresh dairy by an urban household that values quality over volume.

Cluster name	Premium Fresh Urban
Transactions	1,364 (31.5% of dataset)
Average quantity sold	129.5 units
Average revenue per transaction	₹9,974.2
Average price per unit	₹76.97

Average shelf life	19.1 days
Dominant channels	Retail (34.8%), Online (33.3%), Wholesale (32.0%)
Dominant products	Buttermilk, Butter, Curd
Dominant brands	Mother Dairy, Amul, Raj
Top geographies	Delhi, Chandigarh, Uttar Pradesh, Karnataka, Haryana

Table 6.3: Profile of Cluster C2 — Premium Fresh Urban segment

Source: Own analysis using scikit-learn KMeans on the Kaggle Dairy Goods Sales Dataset.

6.2.4 Cluster C3 — Shelf-Stable Specialty segment

Cluster C3 is the smallest cluster, accounting for 485 transactions or 11.2 percent of the dataset, and is the most distinctive by product profile. The mean shelf life of 99.8 days is roughly five times the dataset mean, reflecting the dominance of ghee (343 records) and cheese (142 records) within the segment. The mean quantity sold of 204.7 units is mid-range, and the mean revenue per transaction of ₹9,821 is comparable to that of Cluster C2 despite a lower unit price (₹52.44). The channel mix tilts toward wholesale (39.6 percent). The brand mix is led by Amul, Mother Dairy and Sudha. The geographic distribution is led by Delhi and Chandigarh, with secondary representation in Karnataka, Uttar Pradesh and Gujarat. This is the shelf-stable specialty transaction: a long-shelf-life, mid-volume purchase of value-added dairy concentrates, often routed through wholesale channels for downstream institutional or retail distribution.

Cluster name	Shelf-Stable Specialty
Transactions	485 (11.2% of dataset)
Average quantity sold	204.7 units
Average revenue per transaction	₹9,821.0
Average price per unit	₹52.44
Average shelf life	99.8 days
Dominant channels	Wholesale (39.6%), Retail (31.8%), Online (28.7%)
Dominant products	Ghee, Cheese
Dominant brands	Amul, Mother Dairy, Sudha

Top geographies	Delhi, Chandigarh, Karnataka, Uttar Pradesh, Gujarat

6.3 The Prescriptive Packaging Framework

The prescriptive packaging framework is constructed by mapping each cluster onto the Wang, Wang and Cho (2022) utility matrix, moderated by Indian-context variables. The four moderating variables are authenticity cues, price sensitivity, cold-chain robustness and quick-commerce handling. The recommendations are summarised in Table 6.5 and elaborated in the narrative that follows.

Cluster	Shape	Graphic	Colour	Label
C0 Everyday Mainstream	Laminated pouch / rigid cylindrical tub	Concrete, traditional purity cues	Balanced cool–warm	Simple with brand legacy
C1 Institutional / Bulk	Square / palletisation-optimised	Functional, minimal	Neutral / warm	Oversized batch and expiry
C2 Premium Fresh Urban	Gable-top (reinforced)	Concrete, premium aesthetic	Cool palette	Simple, protein/nutrition forward
C3 Shelf-Stable Specialty	Rigid jar / tin-look canister	Concrete, traditional authenticity	Warm (gold, red)	Heritage-led, claim-rich

Table 6.5: Prescriptive packaging framework by cluster

Source: Own analysis, integrating Wang, Wang and Cho (2022) utility scores with Indian-context moderating variables.

6.3.1 Recommendation for Cluster C0 — Everyday Mainstream

For the Everyday Mainstream segment I think the best packaging is the laminated pouch for dairy and the rigid cylindrical polypropylene tub for fresh fermented dairy. I do not think the gable-top shape is an idea even though Wang, Wang and Cho said it was the best in 2022. This is because the gable-top shape is a lot more expensive than the laminated pouch. The Everyday Mainstream segment is very sensitive to price. The average price of a unit is ₹30.75 which is not enough to cover the extra cost.

The graphics on the packaging should be simple. Show traditional images of purity like cows, pastures or churning vessels. This is because the market is very worried about adulteration and these images show that the product is authentic.

The colour of the packaging should be a mix of warm colours so it is easy to recognise the brand like the blue of Amul or the green of Mother Dairy.

The label on the packaging should be easy to read and have the brand name prominently displayed along with the information that the government requires.

6.3.2 Recommendation for Cluster C1 — Institutional and Bulk

For the Institutional and Bulk segment what is most important is that the product is useful not that it looks good. The best shape for this segment is a square or rectangle because it is easier to stack on pallets. This is what Wang, Wang and Cho said is important for the shape. For this segment it is more about making it easy to transport than making it look nice. The graphics on the product should be simple and not fancy because the people who buy these products are not looking at them to put on a shelf they are looking to use them. The colour of the product should be neutral or warm. The label on the product should have batch numbers and expiry dates so that the people who buy them know exactly what they are getting and when it will expire. This is very important for the people who buy in bulk because they need to know that the products are safe. This segment is where the global scores are very different from what's best, for India.

6.3.3 Recommendation for Cluster C2 — Premium Fresh Urban

For the Premium Fresh Urban segment the best advice is to use a shape that's like a gable-top. This is what Wang, Wang and Cho found out in 2022. The gable-top shape is good because it looks nice and it can handle the handling that happens when people get their deliveries quickly.

The pictures on the package should be photos that look fancy and show that the food is fresh and good quality. Do not use the pictures that people are used to. The colours should be cool like what Wang, Wang and Cho said.

The label should be simple. Have information about protein and nutrition on the front. This is what people who live in cities and care about what they eat want to see. The Premium Fresh Urban segment likes this kind of thing. Wang, Wang and Cho also said that the gable-top shape and cool colours are an idea. The simple label with nutrition information is a choice, for the Premium Fresh Urban segment.

6.3.4 Recommendation for Cluster C3 — Shelf-Stable Specialty

For the Shelf-Stable Specialty segment the advice is to make the product look special because it has a lot of ghee and cheese. The best shape for the container is a jar or a canister that looks like a tin, which keeps everything fresh and looks expensive. The pictures on the container should be real. Show traditional things like making ghee or old patterns. The color should be warm like gold, red or deep amber, which means the product is rich and has a lot of fat. This is different from what Wang, Wang and Cho do because in India people think warm colors mean quality ghee. The label on the container should talk about where the product comes from and how it is made and it should be proud of its history and the way it is prepared.

6.4 Strategic Implementation — The GABDO Model

The GABDO model — Goals, Analysis, Build or Buy, Deploy, Optimise — is applied to translate the prescriptive packaging framework into an organisational implementation plan. Table 6.6 summarises the implementation roadmap.

GABDO Phase	Implementation focus and expected output
Goals	Set the operational target as a 15% increase in value-added product share through segment-specific packaging within 18 months. Define success as a reduction in cost-per-acquisition in Premium Fresh Urban and a reduction in cold-chain spoilage in Institutional and Bulk, rather than as model accuracy.
Analysis	Conduct a data audit integrating internal CRM data with the transactional dataset used in this study. Reconcile data silos between marketing, logistics and product teams. Validate that customer location, shelf life and sales channel are populated for at least 95% of recent transactions before model deployment.
Build or Buy	Larger entities such as Amul and Mother Dairy can build internal MLOps pipelines. Mid-sized firms such as Dodla Dairy and regional cooperatives are likely to obtain better economics from partnering with specialised AI vendors, particularly for the initial proof of concept.
Deploy	Adopt a pilot-first rollout. Deploy the reinforced gable-top first in Premium Fresh Urban geographies (Delhi, Karnataka). Validate uplift in on-shelf conversion before extending to other segments. Roll out the Institutional and Bulk recommendations next, since cold-chain spoilage reduction provides a clear measurable benefit.
Optimise	Establish a quarterly retraining cadence for the clustering model. As consumer tastes evolve — particularly toward plant-based hybrid dairy — the cluster boundaries will shift, and the prescriptive framework must shift with them.

Table 6.6: GABDO implementation roadmap for the proposed framework

Source: Own analysis.

6.5 Organisational Readiness — The TCPR Model

The TCPR model — Time, Cost, Performance, Requirements — is applied to assess organisational readiness for the implementation of the proposed framework. Table 6.7 summarises the readiness assessment.

TCPR Dimension	Readiness assessment
Time	A proof-of-concept implementation using the Kaggle dataset and internal sales data is achievable in 8 to 12 weeks. A production-grade implementation, including integration with internal data warehouses and the deployment of the first pilot packaging line, is achievable in approximately six months for a mid-sized firm.
Cost	Cost drivers include data-engineering effort, model-development effort and the incremental material cost of the prescribed packaging configurations. The gable-top configuration recommended for Premium Fresh Urban carries a material cost premium of approximately 15 to 25 percent over the laminated pouch baseline.

TCPR Dimension	Readiness assessment
Performance	Performance is measured through market resonance — that is, through statistically significant uplift in sales within the target cluster after deployment of the prescribed packaging. The recommended evaluation horizon is a minimum of two quarters per pilot, to absorb seasonal variation.
Requirements	Data readiness is the single most important precondition. Specifically, the customer location, shelf life and sales channel fields must be populated and accurate at the transaction level. Without this foundation, the clustering algorithm yields uninterpretable centroids.

Table 6.7: TCPR readiness assessment for an Indian dairy firm

Source: Own analysis.

CHAPTER 7
CONCLUSION

The Indian dairy industry is going under transformation. The market which was earlier volume dependent is now going under change with the increase of different segments of consumers having different preferences and behaviours. Customer differentiation, premiumisation urbanization and value added dairy products are on the rise due to the onset of new premium and urban consumers. The value added Dairy segment is growing at a compound annual growth rate of 6.8% year on year. Due to the new segment being introduced the consumer groups have now been divided into urban consumers, premium consumers, traditional consumers and institutional buyers. Each demands a different product strategy but the traditional bulk volume industry is still dependent on legacy cooperatives, and mid size Dairy firms produce generic packaging to all the consumer segments.

There are several other factors which are to be considered for the product packaging. Resources should be spent on making packaging designs that should help in the overall longevity of the dairy product. Premium segment should also be tapped having differentiation, aesthetic that influences purchasing decisions of the consumers and the perceived quality that the branding and the packaging represents. Packaging should also be restructured to consider the cold storage problems in India. The study addresses these gaps and provides a long term predictive analysis and finds implementation based framework strategy. The study also found the linkage between the machine learning based consumer segmentation and the conjoint analysis based packaging preferences.

Machine learning algorithms such as K-Means clustering is used in the Kaggle dataset of 4325 dairy transaction records collected over 15 States from 2019 to 2022.

During the study, four consumer segments were identified. Cluster one represents the everyday mainstream segment having 38.2% of all the transactions. It is the largest cluster in the data set and is characterized by routine consumption, affordability and mass buying patterns. Institutional and bulk segments represent 19% of the total transactions. These are associated with bulk buying, institutions wholesaler and large volume purchases. A new category segment called premium fresh urban segment having 31.5% of the total transaction and characterized by urban, quality conscious premium oriented customers which are strongly affected by aesthetic and the packaging quality. Lastly shelf stable specialities segment, representing 11.2% transactions are dominated by speciality dairy products, long shelf life, ghee and cheese.

These cluster profiles are then integrated with conjoint analysis done by Wang, Wang and Cho (2022), to find the perfect shape, colour, graphics and label complexity. This was done while keeping in mind the Indian context of dairy products. It was observed that the premium urban consumers match the global preferences and they do want similar product design and packaging materials such as concrete graphics, cool colours, gable-top and simple text. They were willing to pay higher for premium products. Institutional buyers, being the highest revenue per transaction paid importance to the functionality and the operational efficiency of the packaging. Traditional consumers still relied on plastic packaging with cows as graphics to ensure reliability. Shelf-Stable speciality products required warm colour signaling to ensure the richness, intelligence, tradition and product depth.

Finally the study tested implementation feasibility using the GABDO and TCPR framework. While GABDO is used for strategic planning, TCPR is used for accessing organisation readiness for change.

A timeline was also recorded by the study. It was concluded that the mid size Indian Dairy could implement the framework in 6 months of time. Data readiness is highly important in this scenario, with time stretching if the data is not of the perfect quality.

The study provides Indian Dairy firms structural packaging decision framework, segment specific packaging recommendation and practical method without expensive consumer primary data.

The limitations of the study were focused on the secondary data usage. Chinese research was used to find correlation between the Indian Dairy sector and consumer preferences, which eliminate the cultural preferences of Indian consumers. There were few limitations of K-means clustering as well regarding the cluster shaped centroid placement and predetermined cluster numbers. The research was very industry specific rather than being company specific.

REFERENCES

TechSci Research. (2026). *Yogurt market — Size, share, trends and forecast 2030*. Retrieved from https://www.techsciresearch.com/report/global-yogurt-market/3126.html

Velasco, C., Woods, A. T., Petit, O., Cheok, A. D., & Spence, C. (2016). Crossmodal correspondences between taste and shape, and their implications for product packaging: A review. *Food Quality and Preference*, *52*, 17–26.

Wang, E. S. T. (2013). The influence of visual packaging design on perceived food product quality, value, and brand preference. *International Journal of Retail and Distribution Management*, *41*(10), 805–816.

Wang, F., Wang, H., & Cho, J. H. (2022). Consumer preference for yogurt packaging design using conjoint analysis. *Sustainability*, *14*(6), 3463.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston, MA: Kluwer Academic Publishers.

Zhou, X., & Xu, Y. (2019). Conjoint analysis of consumer preferences for dress design. *International Journal of Clothing Science and Technology*, *32*(1), 73–84.

ANNEXURE

Annexure A: Variable inventory of the Kaggle Dairy Goods Sales Dataset

The following twenty-three fields constitute the schema of the Kaggle Dairy Goods Sales Dataset used as the primary analytical corpus of this study: Location, Total Land Area (acres), Number of Cows, Farm Size, Date, Product ID, Product Name, Brand, Quantity (litres/kg), Price per Unit, Total Value, Shelf Life (days), Storage Condition, Production Date, Expiration Date, Quantity Sold (litres/kg), Price per Unit Sold, Approximate Total Revenue (INR), Customer Location, Sales Channel, Quantity in Stock (litres/kg), Minimum Stock Threshold (litres/kg) and Reorder Quantity (litres/kg). Five of these fields — Quantity Sold, Price per Unit Sold, Approximate Total Revenue, Shelf Life and Sales Channel — constitute the feature set used in the K-Means clustering reported in Chapter 5.

Annexure B: Indicative Python pseudocode for the K-Means clustering workflow

```
# Step 1: Ingest and prepare the data

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette_score

df = pd.read_csv('dairy_dataset.csv')

feat = df[['Quantity Sold (liters/kg)', 'Price per Unit (sold)',

           'Approx. Total Revenue(INR)', 'Shelf Life (days)']].copy()

feat['ChannelOrd']                        =                        df['Sales
Channel'].map({'Online':2,'Retail':1,'Wholesale':0})

X = StandardScaler().fit_transform(feat)

# Step 2: Choose k via Elbow and Silhouette

for k in range(2, 8):

    km = KMeans(n_clusters=k, n_init=10, random_state=42).fit(X)
```

```
    print(k, km.inertia_, silhouette_score(X, km.labels_))
# Step 3: Fit final model
km = KMeans(n_clusters=4, n_init=20, random_state=42).fit(X)
df['Cluster'] = km.labels_
```

Annexure C: Glossary of selected terms

- **Conjoint analysis.** A multivariate statistical method that decomposes a product into its constituent attributes and quantifies the part-worth utility consumers attach to each attribute level.
- **Cluster centroid.** The mean position of all observations assigned to a cluster in the relevant feature space; it represents the prototypical observation of the cluster.
- **Silhouette coefficient.** A measure of how similar an observation is to its own cluster compared to the next-closest cluster; values range from −1 to +1, with higher values indicating better-defined clusters.
- **Elbow method.** A heuristic for determining the optimal number of clusters by plotting the within-cluster sum of squares against k and identifying the point at which marginal reductions taper off.
- **Gable-top.** A folding-carton packaging shape, common to liquid dairy and juice products, in which the top of the package is folded into a triangular peak.
- **Quick-commerce.** A retail-delivery model in which consumer goods, including perishable dairy, are delivered to the consumer within ten to thirty minutes of ordering, typically through dark-store networks operated by platforms such as Blinkit, Zepto and Instamart.