UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT

DATA ANALYTICS IN BUSINESS

TOPIC: APPLICATION OF RETAIL ANALYTICS FOR STRATEGIC PROMOTION IN THE VIETNAMESE COFFEE MARKET: THE CASE OF HIGHLANDS COFFEE

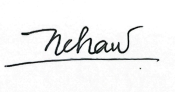
Group: 02

Instructor:

Assoc. Prof. Ho Trung Thanh, Ph.D.

Ho Chi Minh City, September, 2025

Members of Group 02

No.	**Full name**	**Student ID**	**Point / 10 (Individual Contribution) – Chấm điểm chính xác theo đóng góp của mỗi cá nhân**	**Signature**
1	Bùi Thị Ngọc Châu	K224111443	100%	
2	Bùi Thị Hồng Thi	K224111463	100%	
3	Thái Anh Thư	K224111468	100%	
4	Hồ Song Tín	K224111469	100%	
5	Lê Huyền Trân	K224111472	100%	

Acknowledgments

We would like to sincerely thank the University of Economics and Law for the course *"Business Data Analyst"*. This course has provided us with valuable knowledge and practical insights into how to collect data and plan business insights from data. The university's learning environment and resources have greatly supported us during the completion of this project.

We are especially grateful to *Assoc. Prof.* Hồ Trung Thành for his dedicated guidance and support throughout the course. Her helpful feedback and practical exercises helped us connect theory with real applications. We truly appreciate her time, effort, and encouragement.

Completing this project has been a very meaningful experience for us. While we have tried our best to ensure its quality, we know there is always room for improvement, and we welcome any suggestions for future development.

Once again, we sincerely thank everyone who has been part of this journey.

Ho Chi Minh City, October 19th, 2025

Group 02

Commitment

We hereby confirm that the project titled ***"Application Of Retail Analytics For Strategic Promotion In The Vietnamese Coffee Market: The Case Of Highlands Coffee"*** is the result of our team's own research and development. The project was done under the guidance of ***Assoc. Prof.*** **Hồ Trung Thành** at the University of Economics and Law, Vietnam National University – Ho Chi Minh City.

All content in this report comes from our original work in research, system design, and application development. Any ideas, data, or information from other sources have been properly cited and referenced according to academic rules and research standards.

We also confirm that this report has not been submitted for any other course, and it does not include any plagiarized material. The findings, analysis, and conclusions are the result of our team's work, and we take full responsibility for the accuracy and truthfulness of this report.

Ho Chi Minh City, October 19th, 2025

Group 02

List of Abbreviations

No	Abbreviation	Full Meaning
1	F&B	Food and Beverage
2	ETL	Extract, Transform, Load
3	API	Application Programming Interface
4	SQL	Structured Query Language
5	ERP	Enterprise Resource Planning
6	CRM	Customer Relationship Management
7	XML	eXtensible Markup Language
8	JSON	JavaScript Object Notation
9	OLAP	Online Analytical Processing
10	OLTP	Online Transaction Processing
11	KPI	Key Performance Indicator
12	SVM	Support Vector Machine
13	SMOTE	Synthetic Minority Over-sampling Technique
14	XGBoost	eXtreme Gradient Boosting
15	CLV	Customer Lifetime Value
16	EDA	Exploratory Data Analysis
17	DBSCAN	Density-Based Spatial Clustering of Applications with Noise
18	PSO	Particle Swarm Optimization
19	EGBM	Enhanced Gradient Boosting Machine
20	ANN	Artificial Neural Network
21	POS	Point of Sale
22	PPA	Price Per Action

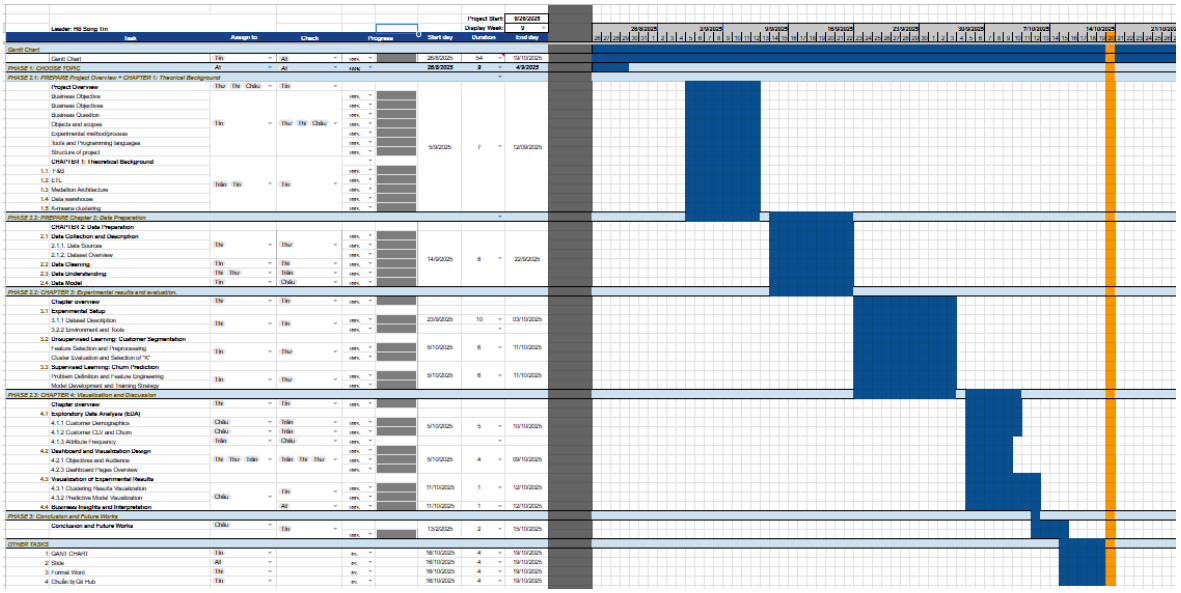
23	NPS	Net Promoter Score
24	MPI	Market Potential Index
25	KPI	Key Performance Indicator
26	BI	Business Intelligence
27	AWS	Amazon Web Services
28	API	Application Programming Interface
29	ML	Machine Learning
30	DBMS	Database Management System
31	EGBM	Enhanced Gradient Boosting Machine
32	RBF	Radial Basis Function
33	JSON	JavaScript Object Notation
34	CLV	Customer Lifetime Value
35	CRM	Customer Relationship Management

Table of Content

List of Tables

List of Figures

GANTT CHART



Project Overview

The F&B sector, particularly the Vietnamese coffee market, is one of the most dynamic and competitive in Southeast Asia. It brings together long-standing domestic brands such as Trung Nguyên and Phúc Long, fast-growing chains like Highlands Coffee, and international players such as Starbucks. Alongside these names, countless independent shops also compete for customer attention. By 2024, Vietnam had an estimated 323,000 active food and beverage (F&B) outlets, but more than 30,000 businesses closed within the first half of the year alone. This high turnover illustrates how volatile the market is, where even well-known brands must constantly adapt to survive. To stay competitive, Highlands Coffee and similar chains need a more precise approach: promotions that target the right customer segments, strengthen loyalty, prevent churn, and encourage higher-margin purchases without relying on across-the-board price cuts.

According to Miguel Alves Gomes et al. (2024): “Customer segmentation is an unsupervised-learning process and utilizes different clustering approaches which have the goal to separate aforementioned customer data based on similarity,". With the meaning of customer clustering, Ann Højbjerg Clarke et al. (2024) once said that “Segmentation builds on and holds implications for marketing strategy, the differentiation of marketing activities, and the allocation of resources. Consequently, segmentation is essential for firms to enhance their competitive edge in navigating a more complex marketing landscape.” Rahma Wati Br Sembiring Berahmanaa et al. pointed that “The main goal of the company is to strengthen the relationship between one customer with another customer to get a significant profit in the market competition.”

(1) Yang et al. (2023) explored consumer behaviour patterns and applied churn prediction models in the food delivery industry. Their study highlighted that consumer retention is often more cost-effective than acquisition, and they used a six-stage churn analysis framework incorporating machine learning techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), AdaBoost, Gradient Boosting, XGBoost, and Deep Neural Networks. The results demonstrated that proactive churn management,

supported by accurate predictive models, can significantly reduce costs and improve customer loyalty.

(2) AlShourbaji et al. (2023) introduced an Enhanced Gradient Boosting Machine (EGBM) model that integrates Support Vector Machines with a Radial Basis Function kernel as base learners. By combining boosting techniques with a modified Particle Swarm Optimization (PSO) algorithm, their model achieved higher predictive performance than traditional Gradient Boosting and SVM approaches. Evaluations across multiple open-source datasets showed that EGBM outperformed state-of-the-art models, making it a robust method for churn prediction in the telecommunications sector.

(3) Putri et al. (2024) investigated retail marketing strategy optimization through customer segmentation using K-Means clustering combined with Recency, Frequency, and Monetary (RFM) analysis. Their findings identified six distinct customer segments—VIP, Loyal, Potential Loyalists, New, At-Risk, and Dormant Customers. By tailoring strategies to each group, such as exclusive offers for VIP customers and reactivation programs for dormant ones, the study demonstrated improved efficiency in targeted promotions, enhanced customer loyalty, and reduced marketing costs. The integration of AI-based segmentation thus proved highly effective in supporting precision retail strategies.

(4) The research proposed by Deniz Altay Avcı et al. (2024) focused on detecting and predicting customer behavior through user classification, particularly customer churn, in the food and beverage industry using machine learning. The authors applied a combined approach of supervised machine such as Support Vector Machines (SVM), Random Forest, Logistic Regression, and XGBoost alongside with unsupervised machine learning techniques such as Support Vector Machines (SVM), Random Forest, Logistic Regression, and XGBoost to detect customer clusters based on features related to the RFM (Recency-Frequency-Monetary) model. The study shows a successfully developed machine learning-based algorithm to detect and predict customer churn with high

accuracy, provides valuable information about the urgency of customer churn probability to take specific actions to prevent future churn.

(5) Nidhi Gautam and Nitin Kumar (2022) posed a study using data mining techniques for customer segmentation using K-means clustering in order to develop sustainable marketing strategies. The authors' approach to addressing the research problem includes key steps such as data collection and preprocessing, exploratory data analysis, and visualization with tools such as R Studio, Weka, and MS Excel. The study also uses Pair-plot, Joint-plot, Box-plot, and correlation matrix to gain further insights into the data. Their study pointed out the result of five main groups of customers based on annual income and spending habits which are low income and high spending habits, low income and low spending habits, medium income & medium spending habits, high income & high spending habits, high income & low spending habits, and thereby proposing sustainable marketing strategies for each segment.

(6) In the work of Jacint Juhasz (2025), they discovered the benefit of customer segmentation approach combining Recency-Frequency-Monetary (RFM) analysis with the K-Means++ clustering algorithm to improve marketing efficiency and customer retention in the increasingly competitive food and beverage (F&B) industry. The process uses data to calculate the three main RFM metrics: Recency (R), Frequency (F), and Monetary Value (M) and applies not only K-Means++ clustering algorithm but also the Elbow method to identify five separate customer groups which are champion, loyal customers, promising, hibernating customers, and lost customers. The study succeeded in identifying five clear customer segments, giving F&B providers practical insights so that the company can allocate resources more effectively and adjust customer management strategies.

(7) Rahma Wati Br Sembiring Berahmana et al. (2020) disclosed the potential customers and the best customers in marketing activities to strengthen customer relationships and achieve significant profit in the competitive market by using RFM Model with K-Means, K-Medoids, and DBSCAN Methods. This problem is solved through customer

segmentation by combining the RFM model (Recency, Frequency, Monetary) with three different clustering algorithms: K-Means, K-Medoids, and DBSCAN and customer relationship management (CRM) to identify five customer segmentations. The results can be used by companies to identify potential customers and understand their characteristics, so they can provide the best service based on each customer's needs and improve customer management.

(8) Govindaraj Ramkumar et al. (2025) investigated the importance of combining the RFM model with K-Means clustering on customer segmentation to maintain customer loyalty and understanding customer needs for a sustained long-term growth and profitability. The process was carried out through data preparation, implementing RFM analysis, grouping customers into clusters using the k-means technique, and cluster evaluation. Through this, customer clusters were identified and described, such as Best-customers, At-risk, Almost Lost, and core loyal segments. By segmenting customers based on purchasing behavior, companies could adjust strategies to meet diverse customer needs, maximize revenue, build loyalty, improve customer relationships, and strengthen organizational resilience.

(9) In another scientific paper investigated the customer segmentation to design targeted and data-driven marketing strategies in the food and beverage industry using the K-means algorithm by Maulana Rumi Irwan Balo et al. (2025), the authors combined machine learning with business intelligence (BI) and behavioral segmentation to improve customer understanding and enhance digital marketing effectiveness. The approach includes using the LRFM model to analyze customer behavior, and determining the optimal number of clusters and cluster characteristics. The clusters which were identified were core customers, lost customers, and new customers. Thereby, planning for each segmentation with the related marketing strategies.

(10) Czarniecka-Skubina et al. (2021) examined consumer coffee choices in Poland and showed that purchasing habits are strongly influenced by socio-demographic characteristics, preparation methods, and service formats. Through clustering analysis,

they identified three distinct segments—"Neutral coffee drinkers," "Ad hoc coffee drinkers," and "Non-specific coffee drinkers." This finding illustrates the challenge of designing one-size-fits-all promotions, reinforcing the importance of behavioral segmentation to deliver targeted offers.

(11) Yang et al. (2023) explored consumer behaviour patterns and churn dynamics in the food delivery industry. Their study emphasized that customer retention is more cost-effective than acquisition, and they applied predictive models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and XGBoost within a multi-stage churn analysis framework. Results confirmed that proactive churn management supported by accurate machine learning models can mitigate attrition risk and strengthen customer loyalty.

(12) From a spatial perspective, Zhao et al. (2025) investigated the challenge of retail site selection for coffee chains in Shanghai. They argued that traditional metrics such as population density and transport accessibility were insufficient to capture retail potential in dynamic urban spaces. By integrating geospatial data with a Random Forest model, their study achieved over 90% prediction accuracy in identifying suitable store locations for Luckin Coffee and Starbucks, highlighting the effectiveness of location intelligence combined with machine learning.

(13) In the Vietnamese context, Phan (2020) analyzed the difficulties faced by new entrants in the F&B market in Ho Chi Minh City. The research found that packaging design, nutritional transparency, and brand consistency directly shaped consumers' purchase intention, emphasizing that building a trustworthy and coherent brand image is a prerequisite for long-term competitiveness.

(14) AlShamsi (2022) addressed the problem of understanding restaurant customer behaviour through data mining. After conducting exploratory data analysis (EDA), the study trained and compared three models—Decision Trees, Logistic Regression, and Random Forest—to predict consumer choices. The results demonstrated that Random Forest yielded higher accuracy, while Logistic Regression provided interpretability and

Decision Trees offered intuitive visualizations, underscoring the value of model comparison in predictive analytics.

Therefore, with the topic *"Application of Retail analytics for strategic promotion in the Vietnamese coffee market: The case of Highlands coffee"*, we will focus on segmenting customers using RFM combined with clustering methods using K-Means, to identify distinct customer groups, developing machine learning models including Logistic Regression, Random Forest, and XGBoost to predict churn and promotion responsiveness, and designing targeted promotion strategies tailored to each segment. This approach, which replaces mass discounting with data-driven personalized offers, is expected to reduce costs, improve ROI, and strengthen customer loyalty.

Business Objective

In this project, our business objectives are shown below:

- Build a robust data warehouse to store reusable data asset that can support ongoing strategic decision-making for Highlands Coffee

- Identify robust customer segmentations based on demographics, visit behavior and brand perception which integrates churn predictions.

- Assess the competitive landscape of Highlands Coffee in the Vietnamese market relative to its competitors using customer perception, geospatial analysis and brand funnel data.

Business Question

Based on our business objectives, we proposed five business questions:

1. What are the demographic and behavioural characteristics of each customer segmentation?

2. What are the typical behavioural patterns of high loyal customer versus high churn risk ones within customer segments?

3. How does Highlands Coffee sustain its leading position in Vietnam's competitive coffee chain market?

4. How can customer segmentation and churn prediction insights be applied to optimize Highlands Coffee's retention and loyalty strategies?

5. How can Highlands define promotional strategies and in-store experience to counter the specific competitive pressures in oversaturated markets?

Mapping business objectives and business questions

Table 1. Mapping business objectives and business question (Source: Authors)

Business Objectives	Business Question
Build a robust data warehouse to store reusable data asset that can support ongoing strategic decision-making for Highlands Coffee	1. What are the demographic and behavioural characteristics of each customer segmentation? Which of them have the highest CLV?
	2. What are the typical behavioural patterns of high loyal customer versus high churn risk ones within customer segments?
	3. How does Highlands Coffee sustain its leading position in Vietnam's competitive coffee chain market?
	4. How can customer segmentation and churn prediction insights be applied to optimize Highlands Coffee's retention and loyalty strategies?
	5. How can Highlands define promotional strategies and in-store

	experience to counter the specific competitive pressures in oversaturated markets?
Identify robust customer segmentations based on demographics, visit behavior and brand perception which integrates churn predictions.	1. What are the demographic and behavioural characteristics of each customer segmentation? Which of them have the highest CLV?
	2. What are the typical behavioural patterns of high loyal customer versus high churn risk ones within customer segments?
Assess the competitive landscape of Highlands Coffee in the Vietnamese market relative to its competitors using customer perception, geospatial analysis and brand funnel data.	3. How does Highlands Coffee sustain its leading position in Vietnam's competitive coffee chain market?
	4. How can customer segmentation and churn prediction insights be applied to optimize Highlands Coffee's retention and loyalty strategies?
	5. How can Highlands define promotional strategies and in-store experience to counter the specific competitive pressures in oversaturated markets?

Objects and scopes

- **Objects:** Highland Coffee stores and customers as well as other coffee chains operating in major cities

- **Scopes**

+ *Time scope:* The study period is from 2017 to 2019

+ *Space scope;* The research is conducted in major Vietnamese cities including: Can Tho, Ho Chi Minh, Da Nang, Hai Phong, Nha Trang, Ha Noi

Experimental method/process

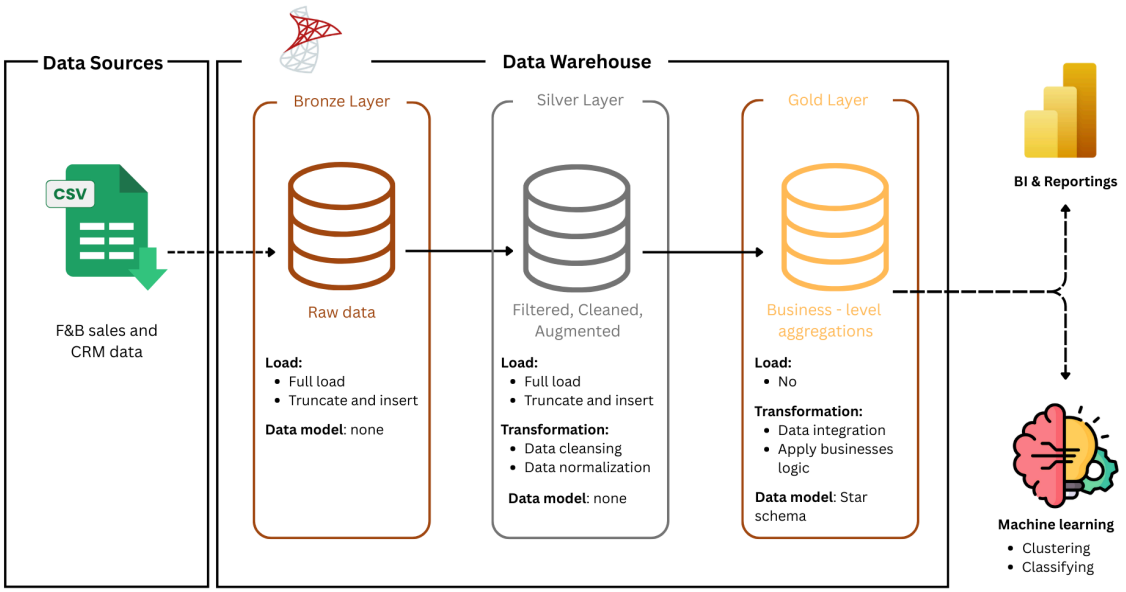


Figure 1. Experimental Process (Source: Authors)

In order to address the research goals, this study employed a data warehouse built on the Medallion architecture, consisting of three layers: Bronze, Silver, and Gold. The raw data from multiple sources, specifically csv files are integrated into the Bronze layer. Then, data is cleaned, normalized, transformed and loaded into the Silver layer after examining constraints such as business rules and consistency. Finally, the Gold layer will be

structured into a star schema with dimensions and facts, preparing for business-level aggregations for advanced analytical tasks with interactive dashboards using Power BI.

Customer segmentations are built to define distinct group attributes based on demographics, behavioural patterns and brand perception. This study applies clustering techniques such as Kmeans and hierarchical clustering to standardize features. Then, Segments are evaluated to interpret and find business insights. Furthermore, to enrich segmentation analysis, the CLV of each segment is evaluated and ranked for long-term revenue potential

The other important task is examining which customers have the high churn risks and identify which factors contribute to this problem. To achieve this goal, supervised learning techniques are trained and evaluated to find the best method to predict customer churn.

Tools and Programming languages

Table 2. Tool Programming languages (Source: Authors)

Tools	**Purposes**
SQL server	Data storage and extraction.
Excel	For quick checks
Power BI	Making dashboards and reports
Python	Machine learning models implementation
Google colab	Environment to run and train models
Github	Version control

Structure of project

Project Overview

This section introduces the research problems, objectives, and scope of the project. Besides, it reviews the relevant literature and theoretical foundations underlying the study. In addition, it outlines the business context of Highland Coffee in Vietnam and experimental method, tools and data sources.

Chapter 1: Theoretical Background

This chapter describes relevant methods, concepts underlying the study.

Chapter 2: Data preparation

This chapter details the data sources and preprocessing methods. It describes the construction of the data warehouse using the Medallion architecture, as well as the processes of data cleaning, integration, transformation, and feature engineering.

Chapter 3: Experimental results and evaluation

This chapter presents the results of customer segmentation and churn analysis. It evaluates the performance of clustering techniques and supervised learning algorithms, as well as their implications for identifying high-value and high-risk customer groups.

Chapter 4: Visualization and discussion

This chapter visualizes the analytical outcomes through interactive dashboards, allowing a clearer interpretation of results.

Conclusion and Future Work

This section summarizes the key findings of the study, highlights its contributions to both theory and practice, and acknowledges its limitations.

References

This section lists all academic, professional, and data sources cited throughout the project, following an appropriate referencing style (e.g., APA, Harvard).

Appendix

The appendix provides supplementary materials, including technical details, extended tables, figures, model parameters, and code snippets that support but are not central to the main text.

Chapter 1. Theoretical Background

1.1. F&B

F&B (Food and Beverage) refers to the sector that covers all aspects of producing, serving, and managing food and drinks for consumers. It is one of the largest and most dynamic segments within the hospitality industry, ranging from small independent restaurants and cafés to multinational chains, luxury hotels, airlines, and large-scale catering services. Success in this industry often depends on effective management of operations, inventory, customer service, and data analysis to understand trends and optimize performance.

1.2. ETL

ETL, which stands for Extract, Transform, and Load, is a fundamental process for integrating data from multiple sources into a centralized repository, typically a data warehouse. It enables organizations to clean, organize, and structure raw data, making it ready for storage, analysis, and machine learning applications. ETL supports business intelligence initiatives by helping generate reports and dashboards, reduce operational inefficiencies, and predict business outcomes.

The ETL process operates in three key stages, typically executed in a recurring cycle to ensure data is kept up-to-date:

a. Extract

During extraction, data is collected from various source systems and temporarily stored in a staging area. Sources can include structured databases, unstructured files, APIs, or web services. The main goal of extraction is to gather data without altering its format, enabling it to be further processed in the next stage.

Types of data sources can include:

- Structured: SQL databases, ERPs, CRMs
- Semi-structured: JSON, XML

- Unstructured: Emails, web pages, flat files

b. Transform

Data extracted in the previous phase is often raw and inconsistent. During transformation, the data is cleaned, aggregated, and formatted according to business rules. This is a crucial step because it ensures that the data meets the quality standards required for accurate analysis.

Common transformations include:

- Data Filtering: Removing irrelevant, duplicate, or incorrect records.
- Data Sorting: Organizing data into a specific order to facilitate analysis.
- Data Aggregation: Summarizing or consolidating data to extract meaningful insights, such as calculating average sales or total revenue.

c. Load

After the data has been cleaned and transformed, the final step is loading, where the processed data is transferred into a target system such as a data warehouse, data lake, or other storage platforms. This stage makes the data accessible for analysis, reporting, and business intelligence purposes.

Depending on the use case, there are two types of loading methods:

- Full Load: The entire dataset is loaded into the target system, typically used when initially populating the data warehouse.
- Incremental Load: Only new or modified data is loaded, which is more efficient for regular updates and ensures that the target system remains current without unnecessary duplication.

1.3. Medallion Architecture

A Medallion Architecture is a data design framework commonly applied in lakehouse systems to logically structure data across multiple layers. Its primary objective is to progressively refine and enhance the quality, consistency, and usability of data as it

moves through the stages, typically referred to as Bronze, Silver, and Gold layers. For this reason, it is sometimes described as a "multi-hop" architecture, since data passes through several transformations before reaching its final curated form.

Although originally associated with lakehouses, the underlying principles of Medallion Architecture are not limited to them. In fact, traditional warehouses already follow a similar structure, often through staging, integration, and presentation layers, which serve functions comparable to Bronze, Silver, and Gold. By aligning with Medallion principles, organizations can strengthen both lakehouse and warehouse architectures to ensure more reliable analytics and decision-making.

a. Bronze layer (raw data)

The Bronze layer serves as the landing zone for raw data collected from external source systems. At this stage, data is stored in its original structure, often accompanied by additional metadata such as load timestamps or process identifiers. The primary goal of this layer is to capture data quickly, preserve historical archives, and maintain lineage and auditability.

b. Silver layer (cleaned and conformed data)

The Silver layer refines the Bronze data by cleansing, conforming, and standardizing it into an enterprise-wide view of key business entities. This layer focuses on reducing duplication and ensuring consistency across multiple sources, enabling self-service analytics, ad-hoc reporting, and machine learning.

c. Gold layer (curated business-level tables)

The Gold layer contains curated, business-ready data optimized for reporting and decision-making. Here, data is organized in project-specific, de-normalized structures designed for fast querying with minimal joins. This layer also applies the final business rules and quality checks, making the data consumption-ready for dashboards, KPIs, and advanced analytics.

1.4. Data warehouse

Data warehousing refers to the process of gathering, integrating, storing, and managing data from multiple sources within a central repository. The primary purpose of a data warehouse is to support decision-making by ensuring that data is clean, consistent, and readily accessible. It is designed to provide rapid data retrieval, even when handling extremely large datasets, enabling organizations to generate insights and make informed business decisions.

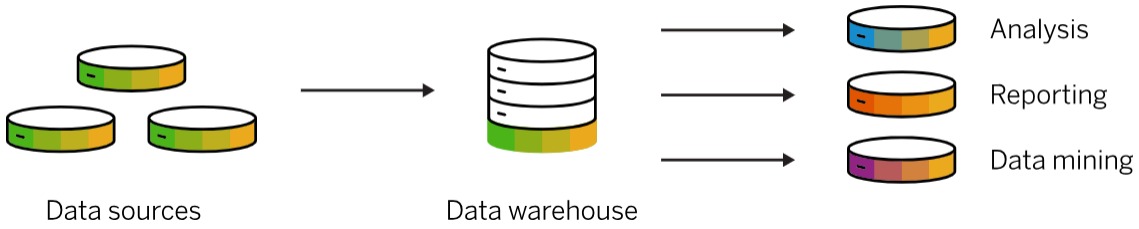


Figure 1.1. Data warehouse Architecture

a. Key Components of a Data Warehouse

A typical data warehouse consists of four main components: a central database, data integration tools, metadata, and access tools. These components are designed for high performance, allowing fast data retrieval and real-time analysis.

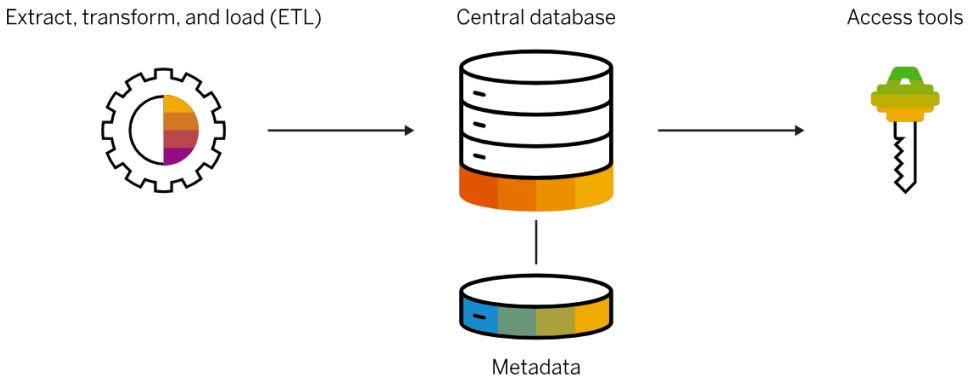


Figure 1.2. The Component Of A Data Warehouse

- **Central database:** forms the foundation of the warehouse. Traditionally, it is a relational database hosted on-premises or in the cloud.

- **Data integration:** involves extracting data from source systems and transforming it to align with the warehouse schema. This is achieved through ETL, real-time replication, bulk-loading, and data quality or enrichment processes, ensuring the data is ready for rapid analysis.

- **Metadata:** provides information about the data itself, including its source, structure, usage, and values. Business metadata adds context for users, while technical metadata describes how to access and manage the data within the warehouse.

- **Data warehouse access tools:** enable users to interact with the warehouse. These include query and reporting tools, OLAP systems, data mining platforms, and application development tools, allowing business users to generate insights and conduct multidimensional analyses efficiently.

b. Data warehouse architecture

Traditionally, data warehouses were structured in layers that mirrored the flow of business data.

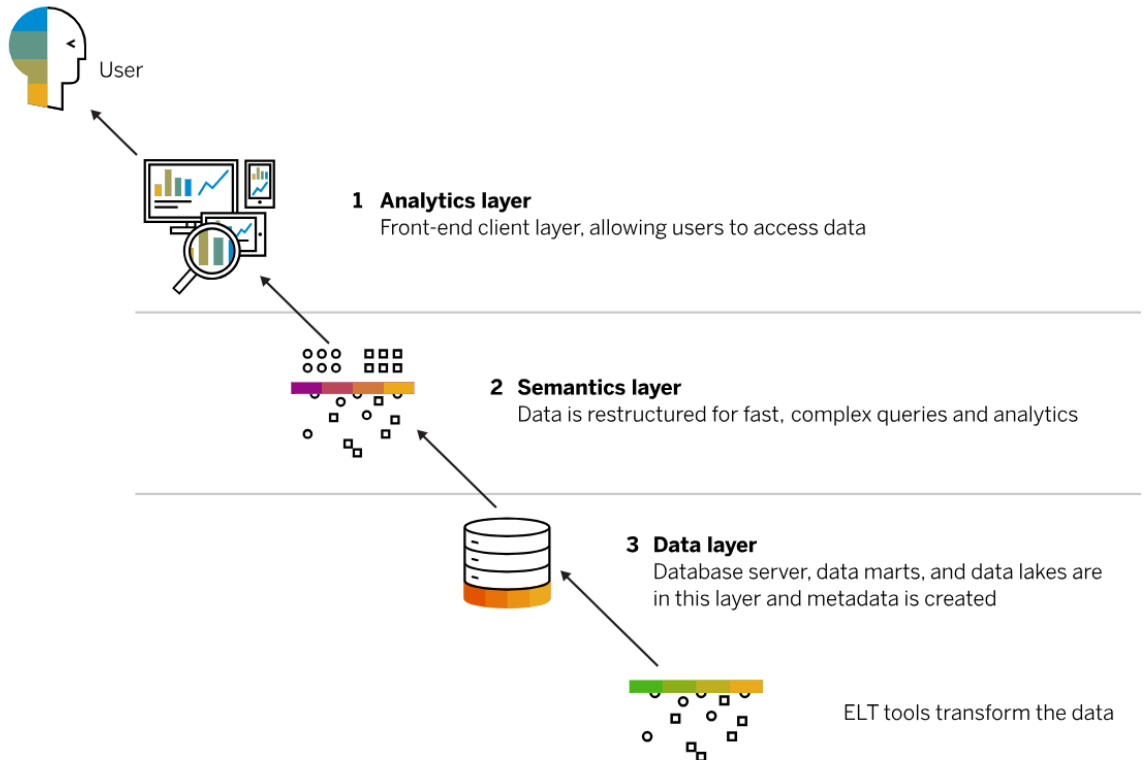


Figure 1.3. Diagram Of Data Warehouse Architecture

- **Data layer:** forms the bottom tier, where data is extracted from source systems, transformed, and loaded using ETL tools. This layer includes the central database, data marts, and sometimes data lakes. Metadata is also created here, and data integration tools, such as data virtualization, help combine and aggregate information seamlessly.

- **Semantics layer**: or middle tier, uses OLAP and OLTP servers to restructure the data for fast, complex queries and analytics. This enables multidimensional analysis and efficient processing of large datasets.

- **Analytics layer**: or top tier, is the front-end interface for users. It includes dashboards, reporting tools, KPI monitoring, data mining platforms, and sandboxes for exploration or developing new data models. This layer allows users to interact directly with data, reducing their reliance on IT teams while supporting informed decision-making.

1.5. K-prototypes

One of the conventional clustering methods commonly used in clustering techniques and efficiently used for large data is the K-Means algorithm. However, its method is not good and suitable for data that contains categorical variables.

To overcome this limitation, Huang (1998) proposed an algorithm called the K-Prototypes algorithm, which was developed as an extension designed to handle mixed-type data, which includes both numerical and categorical variables. This method combines the strengths of K-Means (for numerical data) and K-Modes (for categorical data), providing a unified framework for clustering mixed datasets efficiently.

In the K-Prototypes approach, each cluster is represented by a prototype object - a representative vector that summarizes the characteristics of all data points within that cluster. Specifically:

- For numerical attributes, the prototype value is computed as the mean of the corresponding attribute values across all objects in the cluster.

- For categorical attributes, the prototype value is determined by the most frequent category (mode) among the objects in the cluster.

The K-Prototypes algorithm is a partition-based clustering method that utilizes an objective function (E) and represents each cluster by a prototype object.

Input:

- Initial dataset *X*
- Number of cluster *k*

Output:

- *k* prototype objects such that the objective function E reaches its minimum value.

Algorithm Steps:

Step 1: Initialize *k* prototype objects for dataset *X*, each serving as the representative center of a cluster.

Step 2: Assign each object in *X* to the cluster whose prototype is closest to it. After assignment, update the prototype object of each cluster accordingly.

Step 3: Once all objects have been assigned, check the similarity of each object to all cluster prototypes. If an object is found to be more similar to a different prototype than to the one in its current cluster, reassign the object to the new cluster and update the prototypes of both affected clusters.

Step 4: Repeat Step 3 until no object changes its cluster membership after a full iteration.

1.6. Logistic Regression

Logistic Regression is a fundamental supervised learning algorithm used for binary classification tasks. Despite its name, it's a classification model, not a regression one. Its primary goal is to predict the probability that an instance belongs to a particular class.

The model works by passing input features through a logistic function, also known as the sigmoid function. This S-shaped function takes any real-valued number and maps it to a value between 0 and 1. This output is interpreted as a probability.

$$f(x) = \frac{1}{1+e^{-x}}$$

where:

- $f(x)$ is the output of the sigmoid function.

- e is Euler's number: a mathematical constant ≈ 2.71828.
- x is the input to the sigmoid function.

A decision threshold (typically 0.5) is then applied to this probability. If the calculated probability is greater than the threshold, the instance is classified as belonging to the positive class; otherwise, it's assigned to the negative class.

Logistic Regression is widely used because it's computationally inexpensive, easy to implement, and its results are highly interpretable. It's an excellent baseline model for binary classification problems.

1.7. SVM

A Support Vector Machine (SVM) is a supervised learning algorithm used for both classification and regression, but it is most often applied to classification problems. The main goal of SVM is to find the best hyperplane that separates data points of different classes in a high-dimensional space.

SVM works by finding the hyperplane that divides the data with the largest margin, which is the distance between the hyperplane and the nearest data points from each class. These nearest points are called support vectors, and they play an important role in defining the position and direction of the hyperplane. A larger margin usually means the model will perform better on new, unseen data.

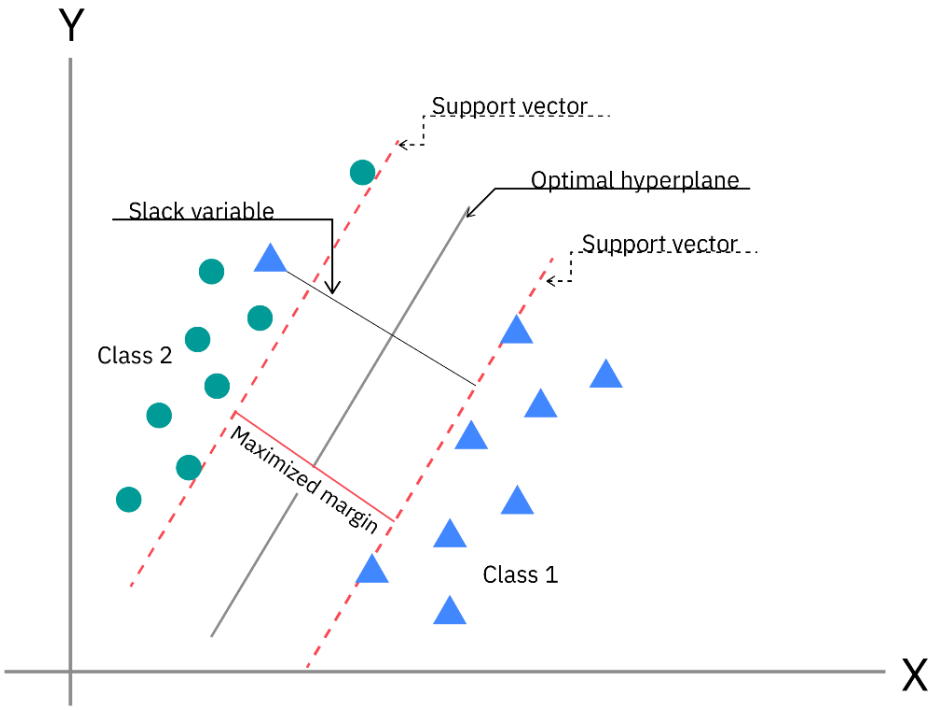


Figure 1.4. Support Vector Machine (SVM) Model

When the data cannot be separated by a straight line, SVM uses the **kernel trick**. This technique maps the data into a higher-dimensional space where it becomes easier to separate the classes. This allows SVM to handle complex, non-linear patterns efficiently without explicitly computing the higher-dimensional coordinates.

Support Vector Machines operate through a structured process that includes data preparation, model construction, and parameter optimization. The following steps outline how an SVM classifier is typically developed and refined.

Step 1: Data Preparation

Similar to other machine learning models, the dataset should first be divided into training and testing subsets. Before this stage, an exploratory data analysis (EDA) is usually conducted to identify missing values, outliers, and potential data imbalances. Although EDA is not a strict requirement for building an SVM, it remains an essential step to ensure data quality and improve model reliability.

Step 2: Model Training and Evaluation

An SVM classifier can be developed by importing the appropriate module from a machine learning library such as scikit-learn. The model is trained on the training dataset and then applied to the test dataset to generate predictions. Its performance is assessed by comparing the predicted outcomes with the actual values. Evaluation metrics such as accuracy, precision, recall, and F1-score are commonly adopted to measure effectiveness.

Step 3: Hyperparameter Tuning

The performance of the SVM model can be enhanced through hyperparameter optimization. Methods such as grid search and cross-validation are employed to determine the optimal combination of parameters, including kernel type, regularization parameter (C), and gamma value. Proper tuning enables the model to achieve a balance between complexity and generalization, ensuring stable and reliable predictions.

1.8. Random Forest

Random Forest is an ensemble learning method used for both classification and regression. It builds multiple decision trees using random subsets of the training data and features, promoting diversity and reducing overfitting. The model aggregates predictions from all trees, with the majority vote used for classification and the average for regression.

Working principle

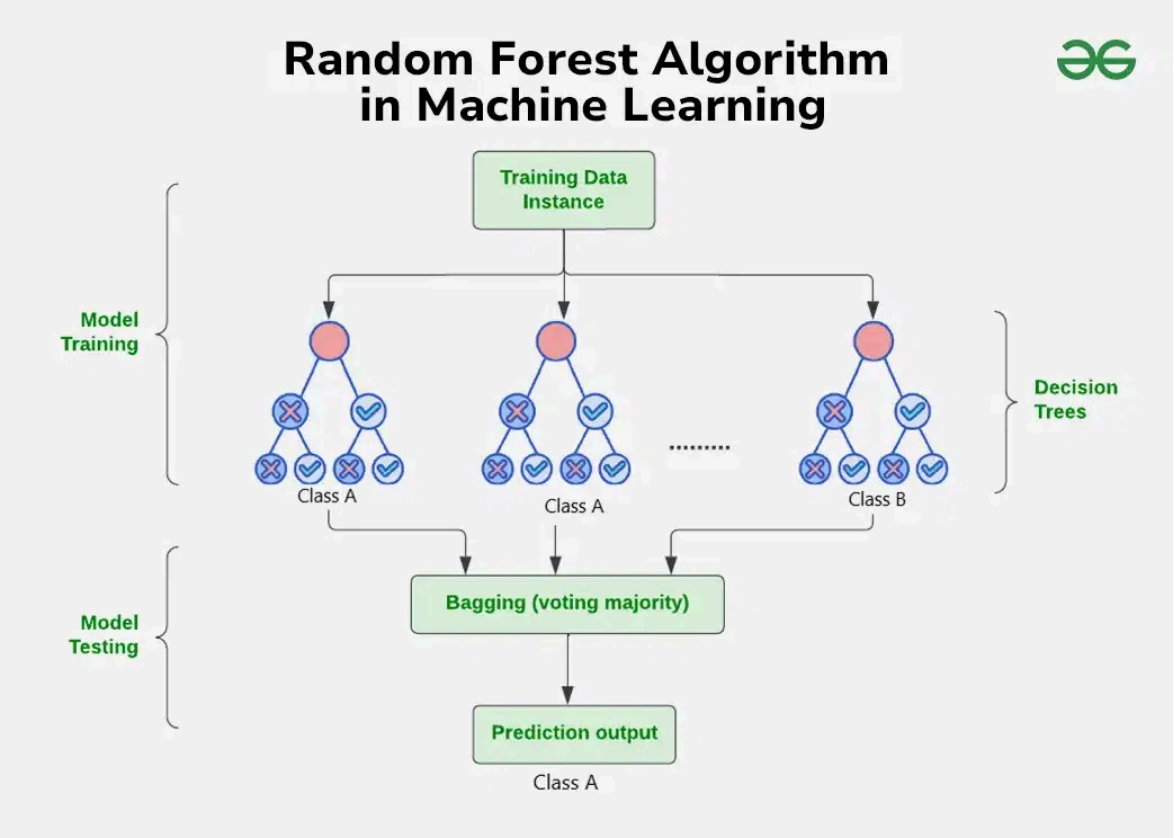


Figure 1.5. Diagram of how the Random Forest algorithm works

Random Forest algorithms have three main hyperparameters that need to be set before training: node size, the number of trees, and the number of features sampled. From there, Random Forest classification can be used to solve regression or classification problems.

Initially, Random Forest applies the bootstrap sampling method to create multiple subsets of data from the original training set. This process is carried out by random sampling with replacement, meaning some data points may appear multiple times in the same subset, while others may not be selected at all. Each subset is the same size as the original training set and is used to build an individual decision tree.

During the construction of each decision tree, the algorithm does not consider the entire feature set but instead randomly selects a small subset of features at each split. This approach helps introduce variability between the trees, reduces variance, and improves the overall performance of the model.

When making predictions on new data, the model aggregates the results from all decision trees. For classification tasks, the predicted label is the one that appears the most, according to the majority voting principle. For regression tasks, the final predicted value is calculated as the mean or median of the results from the trees. Thanks to this method, Random Forest can overcome the limitation of a single decision tree, improving the accuracy and stability of the model.

1.9. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a powerful machine learning algorithm that belongs to the family of ensemble learning methods. It is widely used for supervised learning tasks, including both regression and classification problems.

The main idea behind XGBoost is to build a strong predictive model by combining the outputs of multiple weak learners, typically decision trees, through an iterative boosting process. Each new tree is added to the model sequentially, with the goal of improving upon the mistakes made by the previous trees. During training, XGBoost employs a gradient descent optimization technique to minimize a specified loss function, ensuring that each iteration moves the model closer to optimal predictions.

Key features of XGBoost include its ability to model complex nonlinear relationships, the use of regularization to prevent overfitting, and parallel computation for enhanced training efficiency on large datasets.

Algorithm Steps:

Step 1: Start with a base learner. Begin with a simple model, often a decision tree that predicts a constant value (e.g., the mean of the target variable in regression tasks).

Step 2: Calculate the errors. Measure the difference between the predicted and actual target values to identify the errors that need correction.

Step 3: Train the next tree. Build a new decision tree that focuses on predicting these residuals, effectively learning the aspects the previous model got wrong.

Step 4: Repeat the process. Continue adding new trees, each one refining the predictions made so far, until a predefined stopping condition (such as a maximum number of trees or minimal improvement) is reached.

Step 5: Combine the predictions. The final model output is obtained by summing the weighted predictions of all trees in the ensemble, producing a robust and accurate predictive result.

1.10. SMOTE

Resampling is one of the most widely used strategies for handling imbalanced datasets. Generally, there are two main approaches: i) Undersampling ii) Oversampling. In most cases, oversampling is preferred over undersampling techniques. Among these, oversampling is often preferred because undersampling may remove potentially valuable data points, leading to information loss.

Among various oversampling techniques, SMOTE is one of the most popular and effective methods. It addresses the problem of imbalanced datasets by generating synthetic samples for the minority class, thereby improving the model's ability to learn from underrepresented data. SMOTE, short for Synthetic Minority Oversampling Technique, is a machine learning method designed to address the problem of imbalanced datasets, where one class (the minority class) has significantly fewer samples than the other. Such imbalance can cause models to perform poorly on the minority class, as they tend to favor the majority class during training.

The core idea of SMOTE is to artificially generate new samples for the minority class rather than simply duplicating existing ones. By creating synthetic data points, it helps the model better learn the characteristics of the minority class, improving classification performance.

Working Procedure

Step 1: Define oversampling ratio (N). Define the number of synthetic samples to be created, typically adjusted so that both classes become approximately balanced (e.g., 1:1 ratio).

Step 2: Select a minority instance. Randomly choose a sample from the minority class.

Step 3: Find nearest neighbors. Identify k nearest neighbors (commonly $k = 5$) for the selected instance using a distance metric such as Euclidean distance.

Step 4: Create synthetic samples. Randomly select N neighbors from the k nearest ones. For each neighbor, compute the difference between their feature vectors, multiply it by a random number between *(0, 1]*, and add the result to the original feature vector.

This generates a new synthetic data point that lies between the original and the selected neighbor in the feature space.

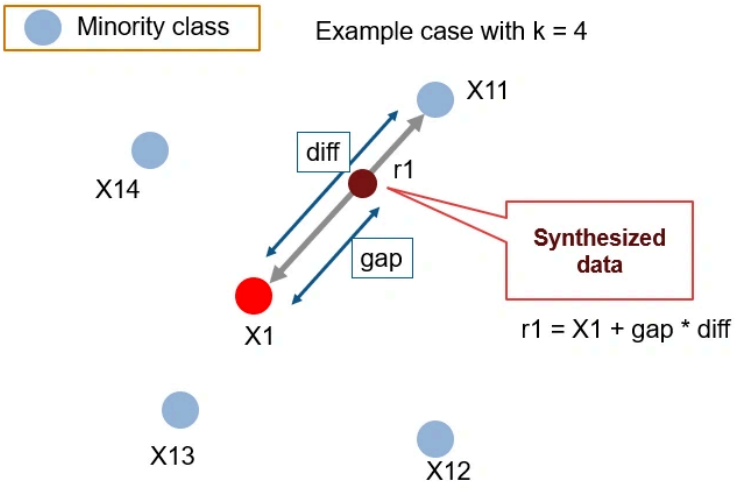


Figure 1.6. Diagram Of How The Smote Algorithm Works

Chapter 2. Data Preparation

This chapter describes prepping the data to be analyzed and establishing the data model to build upon in subsequent research, and it opens by introducing the data sets and the nine representative data sets which filtered to cover information from six large Vietnamese cities across 2017 and 2019. The chapter outline shows cleaning and transformation steps executed to ensure data quality, followed by exploratory analysis to shed light on prevailing characteristics and distributions, and concludes by combining data sets in an aggregated data model, represented in an Entity-Relationship Diagram (ERD), upon which to base an analytically constructed exploration of customer behavior, brand and competitive performance, and competitive dynamics.

2.1. Data Collection and Description

2.1.1. Data Sources

Our data collection is data from surveys in six big cities in Vietnam including Ha Noi, Ho Chi Minh city, Da Nang, Nha Trang, Hai Phong, Can Tho from 2017 to 2019.

2.1.2. Dataset Overview

The 9 datasets in our project which are the 2017 Segmentation Data dataset, Brand Image dataset. Brand Health dataset, Companion dataset, Competitor Database dataset, Day of Week dataset, Day Part dataset, Need State by Day & Daypart dataset and Demographics & Behavior dataset. Those were collected by survey from 2017 to 2019 in six cities in Vietnam. We will describe the meaning of each variable, datatype, role and then identify the initial issues such as duplicates or null values in each dataset.

Dataset 1. 2017 Segmentation Data (2017Segmentation3685case)

Description: This 2017Segmentation3685case dataset contains customer segmentation information based on their visit and spending behavior in 2017.

Data shape: (4944,6). This dataset contains 4944 rows and 6 columns which are ID, Segmentation, Visit, Spending, Brand, and PPA.

Initial issues identified: This dataset contains no duplicates and no null values.

Table 2.1. Dataset 2017 Segmentation Data (Sources: Authors)

Column	Description	Data Types	Null Percentage	Role
ID	Unique identifier for each customer.	Identifier	0%	Identifier
Segmentation	Customer segment label, e.g., Seg.02 - Mass Asp (VND 25K - VND 59K).	Categorical	0%	Analytical (Customer grouping)
Visit	Number of visits made by the customer during the observation period.	Numeric	0%	Behavioral (Frequency)
Spending	Total amount of money spent (in thousand VND) by the customer (e.g., 120 = 120,000 VND).	Numeric	0%	Behavioral (Monetary Value)
Brand	Type of brand chosen: **Independent** (standalone shops), **Chain** (branded chains), or **Street** (informal vendors).	Categorical	0%	Demographic
PPA	Price Per Average, calculated as Spending / Visit.	Numeric	0%	Analytical/ Derived

Dataset 2. Brand Image (Brand_Image)

Description: This Brand_Image dataset contains data on consumers' brand awareness, attribute perceptions, and brand image associations across different cities and years.

Data shape: (643072,6). This dataset contains 643702 rows and 6 columns which are ID, Year, City, Awareness, Attribute, and Brand Image.

Initial issues identified: There are null values in the Awareness variable. In addition, values of Awareness and BrandImage are the same. To address this, one of the variables will be removed.

Table 2.2. Dataset Brand Image (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier
Year	Year of data collection.	Numeric	0%	Contextual (Time Dimension)
City	The city where the respondent resides.	Categorical	0%	Demographic (Geographic Info)
Awareness	The brand that the respondent is aware of.	Categorical	0.06%	Perceptual (Brand Awareness)

Attribute	How the respondent perceives the brand.	Categorical	0%	Perceptual (Brand Perception)
BrandImage	The brand that the respondent associates with a particular image.	Categorical	0%	Perceptual

Dataset 3. Brand Health (Brandhealth)

Description: The Brandhealth dataset contains detailed survey responses measuring brand awareness, usage, perception, and customer segmentation across various coffee brands.

Data Shape: (74419, 20). This dataset contains 74419 rows and 19 columns which are ID, Year, City, Brand, Spontaneous, Awareness, Trial, P3M, P1M, Comprehension, Brand_Likability, Weekly, Daily, Fre#visit, PPA, Spending, Segmentation, NPS#P3M, NPS#P3M#, Group, Spending_use.

Initial issues identified: This dataset contains some null values in variable Spontaneous, Trial, P3M, P1M, Comprehension, Brand_Likability, Weekly, Daily, Fre#visit, PPA, Spending, Segmentation, NPS#P3M, NPS#P3M#, Group, Spending_use but no duplicates found. Additionally, there are inconsistencies in brand names such as Street vs. Street / Half street coffee and multiple versions of the same brand exist. Moreover, qualitative variables Awareness, Trial, Spontaneous, Brand_Likability, P1M, P3M, and Comprehension are not standardized.

Table 2.3. Dataset Brand Health (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role

ID	Unique identifier for each respondent.	Identifier	0%	Identifier
Year	Year of data collection.	Numeric	0%	Contextual (Time Dimension)
City	The city where the respondent resides.	Categorical	0%	Demographic (Geographic Info)
Brand	The brand being evaluated.	Categorical	0%	Contextual
Spontaneous	The brand that comes first to the respondent's mind (unaided awareness).	Categorical	58.35%	Perceptual
Awareness	The brand that the respondent is aware of.	Categorical	0.15%	Perceptual
Trial	Whether the respondent has ever tried the brand.	Categorical	36.40%	Behavioral
P3M	Whether the respondent used the brand in the past 3 months.	Binary	61%	Behavioral

P1M	Whether the respondent used the brand in the past 1 month.	Binary	74%	Perceptual
Comprehension	How well the respondent understands or knows about the brand. (Know it well, Do not know it at all, Maybe do not know it, Know a little, Know it very well)	Categorical	64.60%	Perceptual
BrandLikability	The consumer's level of affection or preference toward the brand.	Categorical	86.12%	Behavioral
Weekly	Indicates if the respondent uses the brand weekly.	Categorical	82%	Behavioral
Daily	Indicates if the respondent uses the brand daily.	Categorical	90%	Behavioral
Fre#Visit	Number of visits made to the brand by the respondent.	Numeric	74.02%	Behavioral

PPA	Price Per Average, calculated as Spending / Visit.	Numeric	81.09%	Analytical
Spending	Total amount of money spent on the brand (in thousand VND).	Numeric	81.09%	Behavioral
Segmentation	Customer segment label (e.g., Seg.02 - Mass Asp (VND 25K - VND 59K)).	Categorical	81.09%	Analytical
NPS#P3M	Net Promoter Score over the past 3 months.	Numeric	70.97%	Analytical
NPS#P3M#Group	NPS classification group: **Promoter**, **Passive**, or **Detractor**.	Categorical	70.97%	Analytical

Dataset 4. Companion (Companion)

Description: This Companion dataset contains information on the typical companion type (e.g., friends, family, alone) customers have when visiting coffee shops.

Data Shape: (11746, 4). This dataset contains 11746 rows and 4 columns which are ID, City, Companion#Group and Year respectively.

Initial issues identified: This dataset contains no null values and no duplicates.

Table 2.4. Dataset Companion (Sources: Authors)

Column	Description	Data Type	Null percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier
City	The city where the respondent resides.	Categorical	0%	Demographic
Companion#group	The usual type of companion the respondent has when visiting a coffee shop.	Categorical	0%	Behavioral
Year	Year of data collection.	Numeric	0%	Contextual

Dataset 5. Competitor Database (Competitordatabase_xlnm#_FilterDatabase)

Description: The Competitordatabase_xlnm#_FilterDatabase dataset contains the number of physical store locations for each coffee brand by city and year, used to assess market presence and competitive density.

Data shape: (234, 5). This dataset contains 234 rows and 5 columns which are No#, Brand, City, Year. StoreCount.

Initial issues identified: This dataset contains no duplicates and no null missing data.

Table 2.5. Dataset Competitor Database (Sources: Authors)

Column	Description	Data Type	Null percentage	Role

No#	Row number or entry index.	Numeric	0%	Index/Identifier
Brand	Name of the coffee brand.	Categorical	0%	Contextual
City	City where the brand's store(s) are located.	Categorical	0%	Demographic
Year	Year in which the store count was recorded.	Numeric	0%	Contextual
StoreCount	Number of stores the brand operated in that city during the given year.	Numeric	0%	Analytical

Dataset 6. Day of Week (Dayofweek)

Description: This Dayofweek dataset contains data on which days of the week consumers typically visit coffee shops, including visit frequency and weekday/weekend classification.

Data Shape: (39095, 6). This dataset contains 39095 rows and 6 columns which are ID, City, Dayofweek, Visit#Dayofweek, Year, Weekday#end.

Initial issues identified: This dataset contains no duplicates but has some null values in variables Dayofweek and variable Visit#Dayofweek.

Table 2.6. Dataset Day of Week (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier

City	City where the respondent resides or visited the coffee shop.	Categorical	0%	Demographic
Dayofweek	Specific day of the week when the visit occurred.	Categorical	0.22%	Behavioral
Visit#Dayofweek	Number of visits made on that particular day of the week.	Numeric	0.138%	Behavioral
Year	Year of data collection.	Numeric	0%	Contextual
Weakday#end	Classification of the day as Weekdays or Weekend.	Categorical	0%	Behavioral

Dataset 7. Day Part (Daypart)

Description: The Daypart dataset contains data on customer visit frequency by time of day, helping to identify peak hours for coffee shop visits.

Data Shape: (11761, 5). This dataset contains 11761 rows and 5 columns which are ID. City, Daypart, Visit#Dayofweek, Year.

Initial issues identified: This dataset contains no duplicates but there are some null percentages in two variables Daypart and Visit#Dayofweek.

Table 2.7. Dataset Day Part (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier

City	City where the respondent resides or visited the coffee shop.	Categorical	0%	Demographic
Daypart	Time range during the day when the visit occurred.	Categorical	0.06%	Behavioral
Visit#Day ofweek	Number of visits made during that specific time range.	Numeric	4%	Behavioral
Year	Year of data collection.	Numeric	0%	Contextual

Dataset 8. Need State by Day & Daypart (NeedstateDayDaypart)

Description: The NeedstateDayDaypart dataset captures consumer need states linked to time of day or day-level behaviors, providing insights into why customers visit coffee shops at specific times.

Data Shape: (75251, 6). This dataset contains 75251 rows and 6 columns which are ID, City, Year, Needstates, Day#Daypart, NeedstateGroup.

Initial issues identified: This dataset contains no duplicates and no missing data.

Table 2.8. Dataset Need State by Day & Daypart (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier
City	City where the respondent resides or visited the coffee shop.	Categorical	0%	Demographic

Year	Year of data collection.	Numeric	0%	Contextual
Needstates	Specific reason or motivation for visiting the coffee shop.	Categorical	0%	Perceptual/ Behavioral
Day#Daypart	Time context for the needed state.	Categorical	0%	Behavioral
Needstate Group	Broader category grouping similar need states.	Categorical	0%	Analytical

Dataset 9. Demographics & Behavior (SA#var)

Description: The SA#var dataset contains detailed demographic and behavioral profiling of coffee shop visitors, including income, age, gender, occupation, and brand preferences.

Data Shape: (11761, 19). This dataset contains 11761 rows and 19 columns which are ID, City, Group_size, Age, MPI#Mean, TOM, BUMO, BUMO_Previous, MostFavourite, Gender, MPI#detail, Age#group, Age#Group#2, MPI, MPI#2, Occupation, Occupation#group, Year, MPI_Mean_Use

Initial issues identified: This dataset contains no duplicates but there are some null values in variables Group_size, Age, MPI#Mean, BUMO, BUMO_Previous, MPI#detail, Age#group, Age#Group#2, MPI, MPI#2.

Table 2.9. Dataset Demographics and Behavior (Sources: Authors)

Column	Description	Data Type	Null Percentage	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier

City	City where the respondent resides or visited the coffee shop.	Categorical	0%	Demographic
Group_size	Number of people in the respondent's visit group.	Numeric	<1%	Behavioral
Age	Age of the respondent.	Numeric	<1%	Demographic
MPI#Mean	Monthly personal income (numerical average, e.g., 5499 = 5.499 million VND).	Numeric	32%	Analytical/Derived
TOM	Top-of-mind coffee brand mentioned by the respondent.	Categorical	0%	Perceptual
BUMO	Brand used most often.	Categorical	0%	Behavioral
BUMO_Previous	Brand used most often previously (if any).	Categorical	48%	Behavioral
MostFavourite	The brand the respondent considers their favorite.	Categorical	0%	Perceptual
Gender	Gender of the respondent.	Categorical	0%	Demographic
MPI#detail	Income range in text format.	Categorical	31%	Demographic

Age#group	Age group category.	Categorical	<1%	Demographic
Age#Group #2	Alternative age group label.	Categorical	<1%	Demographic
MPI	Income category.	Categorica;	32%	Demographic
MPI#2	Grouped income tier.	Categorical	32%	Demographic
Occupation	Respondent's occupation.	Categorical	0%	Demographic
Occupation #group	Broad occupation group.	Categorical	0%	Demographic
Year	Year of data collection.	Numeric	0%	Contextual
MPI_Mean _Use	Same as MPI#Mean; likely used for processing or reporting.	Numeric	32%	Analytical/D erived

2.2. Data Cleaning

To prepare the raw survey data for analysis, a structured, multi-stage data cleaning and transformation process was conducted. The main objective of this procedure was to resolve inconsistencies, address missing information, correct known errors, and restructure fields in order to create a reliable and coherent dataset. Then, the cleaned version of each dataset is loaded into the Silver layer of the data warehouse, which would serve as the foundation for subsequent analysis steps.

2017Segmentation3685Case

This dataset was originally stored as a composite text field, which limited its usability for quantitative analysis. To make this information analytically viable, the following procedures were implemented:

- Key variables such as *SegmentCode* and *CustomerType* were parsed and separated from the composite segmentation string to a lookup table for spending range.
- Identified typographical errors were corrected. Particularly, "Independent" in the *BrandType* field was standardized to the correct form “Independent.”
- Abbreviated figures for *Spending* and *PPA (Price Per Action)* were multiplied by 1,000 to represent their actual values in full numerical terms.

Segmentation_Lookup

This table is created to standardize the classification of customers depending on their spending range which have the description as below

Table 2.10. Structure of Segmentation Lookup (Source:Authors)

SegmentCode	SegmentName	SpendingRange	MinSpending	MaxSpending
Seg.01	Mass	<VND 25K	0	24999
Seg.02	Mass Asp	VND 25K - VND 59K	25000	59000
Seg.03	Premium	VND 60K - VND 99K	60000	99000
Seg.04	Super Premium	VND 100K+	100000	

BrandHealth

The BrandHealth dataset required significant harmonization due to the diversity of brand-related inputs and respondent feedback. The following key steps were undertaken:

Firstly, To ensure the consistent ,, comparison across responses, variations in spelling, abbreviations, and classifications were mapped into standardized categories as bellowed:

Table 2.11. Coffee shop names standardisation (Source: Authors)

Non-standardised	**Standardised**
Street, Street / Half street coffee (including carts)	Street Coffee
Indepedent Cafe, Independent Cafe	Independent Cafe
Other 1,Other 2,Other 3,Other Branded Cafe Chain	Other
KOI cafe	KOI Cafe
Runam cafe	Runam Cafe
Đen Đá	Đen Đá Coffee

The **S**egmentCode was standardized in the same manner as the *2017SegmentationCase*, ensuring consistent classification of respondents by referencing the *Segmentation_LookUp* table, even when raw data did not explicitly contain a segment code. Behavioral variables - such as *awareness, trial, spontaneous recall, brand likability, P3M, and P1M*-were transformed into binary indicators (1 = Yes, 0 = No) to support clearer statistical analysis and modeling. For key metrics (NPS_Score, Frequency_Visits, Spending), missing values were explicitly labeled as “N/A,” distinguishing them from legitimate zero values. The *Comprehension* field was marked as “Did not answer” when responses were blank.

Finally, The PPA metric was calculated as the ratio of Spending to Visit Frequency. This calculation was only performed when both variables contained valid values and visit frequency was non-zero, thus preventing computational errors.

Respondent (SA#Var)

The *Respondent* table contained key demographic and behavioral variables, necessitating systematic cleaning to ensure validity and comparability across records. The following steps were undertaken:

- Redundant fields such as *MPI#Mean*, *MPI#Group*, *MPI#2*, and *AgeGroup#2* were eliminated to streamline the dataset.

- Records missing essential demographic details (*Age*, *Group_size*) were excluded to preserve dataset robustness.

- Brand mentions in variables such as *TopOfMind*, *BrandUseMostOften*, and *MostFavourite* were standardized using consistent naming conventions aligned with the *BrandHealth* dataset.

- Age categories were reformatted into consistent labels (e.g., *"18–24"* instead of *"18–24 y.o."*) to improve readability and uniformity.

- For fields such as *MPI (Monthly Personal Income)* and *BUMO_Previous (Brand Used Most Often – Previous)*, missing values were retained but explicitly marked as *"N/A"*. This approach preserved data completeness while acknowledging inherent limitations.

BrandImage, Competitor, DayOfWeek, DayPart, NeedstateDayDaypart

Other tables also underwent cleaning procedures to enhance uniformity across analytical inputs.

- All city names were converted into standardized numerical identifiers (CityID) using a master lookup table. This improved efficiency and reduced the risk of mismatches in multi-table analyses.

- Missing entries for attributes such as DayPart, VisitFreq, and Needstates were assigned a default value of "N/A." This ensured that no fields remained blank and minimized potential sources of error in subsequent processing.

- Minor formatting adjustments were applied where necessary, such as removing extraneous leading characters in the WeekdayEnd field, thereby ensuring consistent textual representation.

2.3. Data Understanding

This section summarizes the main characteristics of the cleaned datasets, focusing on descriptive statistics, frequency distributions, and outlier detection. The goal is to validate data quality and provide an overview of customer demographics, brand health indicators, and segmentation structure before proceeding to exploratory analysis and modeling.

2.3.1. Respondent Data

Descriptive statistics show that the dataset covers 11,737 respondents across six cities from 2017 to 2019. The average age is 35 years (median = 34, min = 16, max = 60), with most customers falling into the 25–39 age group. Typical group size is 2–4 people (mean = 3.3), although some exceptional cases report up to 30–35 people per visit. Self-reported income (MPI_Mean_Use) has a mean of 7.3k (median ~7k) but varies widely from 1.5k to 112k, indicating substantial heterogeneity in spending potential.

Table 2.12. Descriptive statistics (numeric variables, Respondent dataset) (Source: Authors)

Variable	mean	median	std	min	max
RespondentKey	5869.0	5869.0	3388.32	1.0	11737
RespondentID	442930.75	433943.0	267371.44	89100.0	863754.0
CityID	2.86	2.0	1.62	1.0	6.0
GroupSize	3.29	3.0	1.33	1.0	35.0
Age	35.22	34.0	10.82	16.0	60.0
Year	2017.99	2018.0	0.78	2017.0	2019.0

MPI_Mean_Use	7338.74	6999.0	4668.48	1499.0	112499.0

Frequency distributions highlight a relatively balanced gender split (56% female, 44% male). In terms of brand recall, customers most often mentioned Street Coffee (20%) and Independent Café (15%), with Highlands Coffee ranked fourth (14%). However, when asked about their favorite brand, Highlands rises to 17%, higher than its "most often used" rate (13%). This suggests strong brand affinity that does not fully translate into habitual use. Moreover, nearly 48% of respondents reported having no brand used most often in the past, implying low loyalty and high brand-switching potential. Occupational groups are diverse: blue-collar workers (30%), white-collar professionals (24%), and non-working groups (25%), confirming a heterogeneous customer base.

Outlier analysis identifies two notable patterns. First, about 10% of respondents have MPI_Mean_Use values outside the IQR range (13k–12.3k), including extreme cases exceeding 100k, representing a small but high-value group. Second, only 0.5% reported group sizes above 7, which can be considered plausible outliers tied to events or large gatherings. Age and survey year show no meaningful outliers.

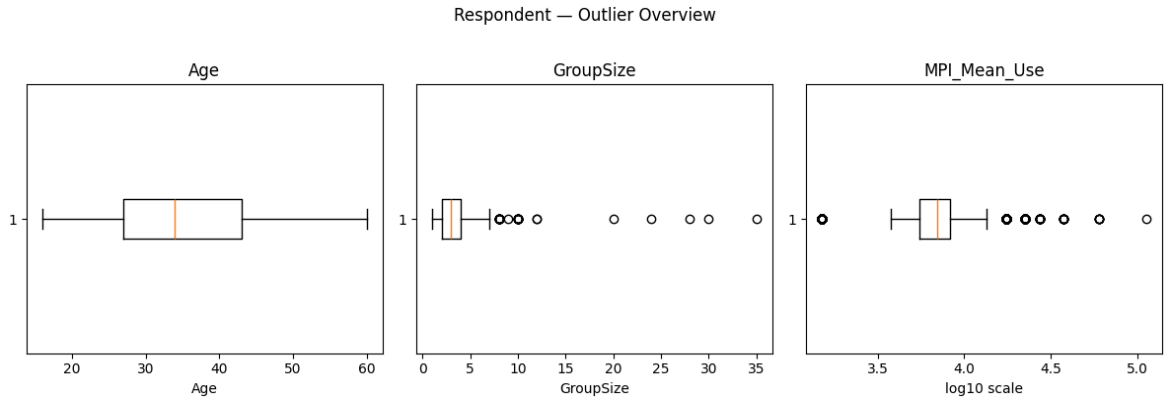


Figure 2.1. Boxplot of outliers (Age, GroupSize, MPI_Mean_Use) (Source: Authors)

The Respondent dataset provides a clear picture of customer demographics and basic behavior. Highlands Coffee is more often favored than regularly used, pointing to a gap between brand love and brand use. Additionally, the presence of high-income/high-spending outliers and occasional large groups, while rare, could be strategically important for premium offerings or event-based marketing.

2.3.2. BrandHealth Data

Descriptive statistics indicate that the dataset contains 74,419 records collected from 2017–2019 across six cities. Awareness levels are almost universal (~100%), but only 42% demonstrate spontaneous awareness. Trial rates are relatively high at 64%, while recent usage drops to 39% in the last three months and 26% in the last month. Visit frequency is skewed, with many reporting 0 visits, while a small group reported over 100. Median spending is VND 100k per visit, but extreme values reach VND 3.75m. The average PPA is ~VND 30k, though outliers go as high as 500k. The average NPS score is 8, reflecting moderate satisfaction.

Table 2.13. Descriptive statistics (numeric variables, BrandHealth dataset) (Source: Authors)

Variable	mean	median	std	min	max
BrandHealthKey	37210.0	37210.0	21483.06	1.0	74419
RespondentID	478277.87	443720.0	268141.83	89100.0	863754.0
Year	2018.0	2018.0	0.78	2017.0	2019.0
CityID	2.63	2.0	1.60	1.0	6.0
Is_Spontaneous_Aware	0.42	0.0	0.49	0.0	1.0
Is_Aware	0.99	1.0	0.04	0.0	1.0
Is_Trial	0.64	1.0	0.48	0.0	1.0
Has_Brand_Likability	0.14	0.0	0.35	0.0	1.0

Used_P3M	0.39	0.0	0.49	0.0	1.0
Used_P1M	0.26	0.0	0.44	0.0	1.0
Frequency_Visits	3.03	0.0	6.94	0.0	120.0
PPA	29824.56	25000.0	19074.69	5000.0	500000.0
Spending	155014.71	100000.0	173986.36	7000.0	3750000.0
NPS_Score	7.97	8.0	1.35	0.0	10.0

Frequency distributions show that Street Coffee (14%) and Trung Nguyên (13%) lead brand mentions, while Highlands Coffee (11%) ranks behind them. Most customers are not assigned to a segment (81%), but Segment 04 accounts for 10% and other segments less than 5%. Regarding NPS, most did not answer (71%), while among valid responses, 15% are Passive, 10% Promoters, and 3% Detractors. Brand likability is low, with only 14% explicitly stating likability.

Outlier analysis reveals three notable patterns:

- 12% of respondents report >7.5 visits, representing heavy users
- 6% spend above VND 425k per visit, likely business or group purchases
- 2% show unusually high PPA (>VND 77.5k), indicating premium consumption

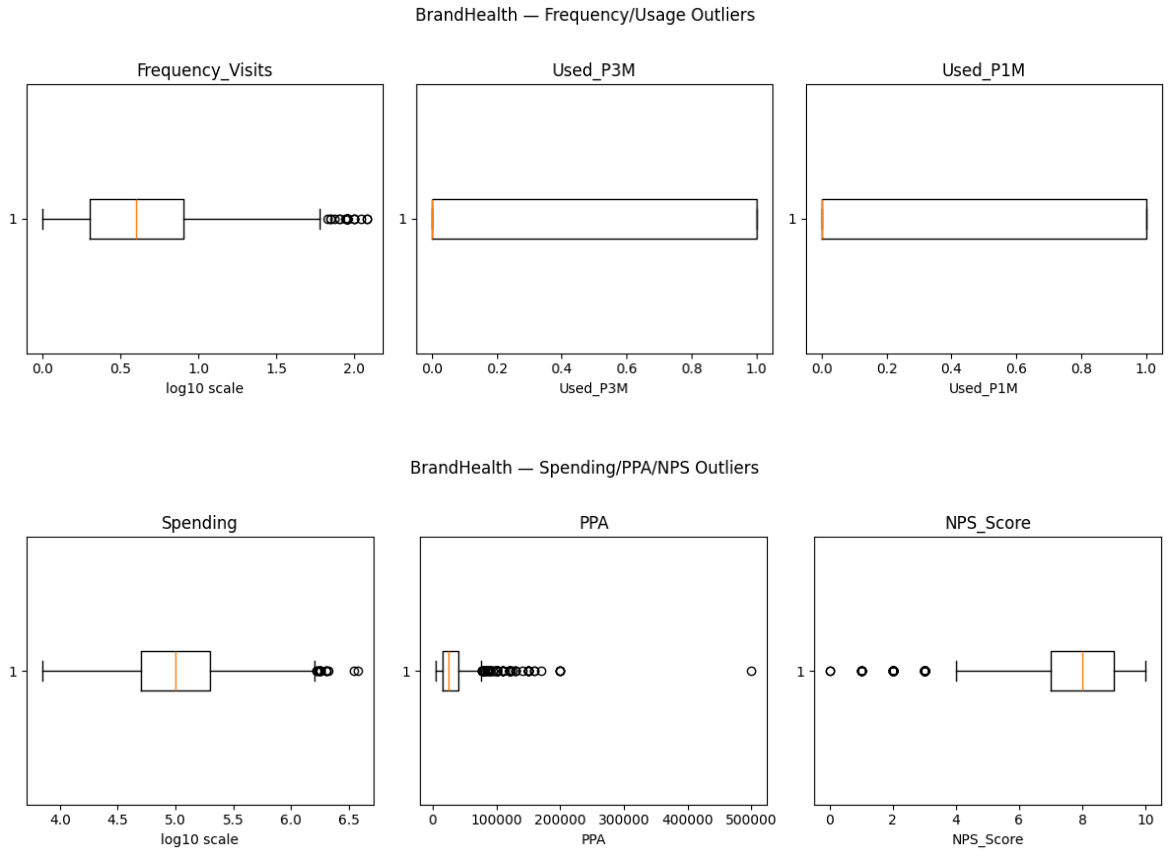


Figure 2.2. Outlier boxplots (Frequency_Visits, Spending, PPA, NPS_Score) (Source: Authors)

BrandHealth data confirms strong awareness but lower levels of recent usage and weak brand advocacy. Highlands Coffee holds a solid position but trails behind local competitors in top-of-mind mentions. Heavy users and premium spenders are important niche groups for retention strategies.

2.3.3. Segmentation2017 Data

Descriptive statistics show that the dataset includes 4,944 cases. The average number of visits is 9.3, though the median is only 4, with outliers up to 120. Average spending is VND 185k, but the median is VND 120k, with extreme values up to 2.5m. The average PPA is ~VND 26k, with most values between 10k and 30k.

Table 2.14. Descriptive statistics (numeric variables, Segmentation2017 dataset) (Source: Authors)

Variable	mean	median	std	min	max
2017SegmentationCaseKey	2472.50	2472.50	1427.35	1.0	4944.0
RespondentID	124746.73	127602000.0	14200.00	89100.0	142479.0
Visit	9.29	4.0	11.25	1.0	120.0
Spending	185212.78	120000.0	208383.54	6000.0	2500000.0
PPA	26125.61	20000.0	17454.01	5000.0	100000.0

Frequency distributions indicate that the market is dominated by Seg.01 (Mass, 52%) and Seg.02 (Mass Asp, 43%), while Premium (4%) and Super Premium (1%) make up only a small fraction. Over half of customers spend less than VND 25k per visit, while less than 1% belong to the 100k+ "Super Premium" category. Independent cafés (44%) are the most common brand type, followed by street cafés (31%) and chains (25%). Visit frequency is concentrated at 2–4 visits, with a niche of heavy users reporting 20–30+ visits.

Outlier analysis shows that 13% of respondents report unusually high visit frequencies (>20), identifying highly engaged customers. Around 6% spend more than VND 510k per visit, and 3% report PPA above 70k, representing premium or group consumption.

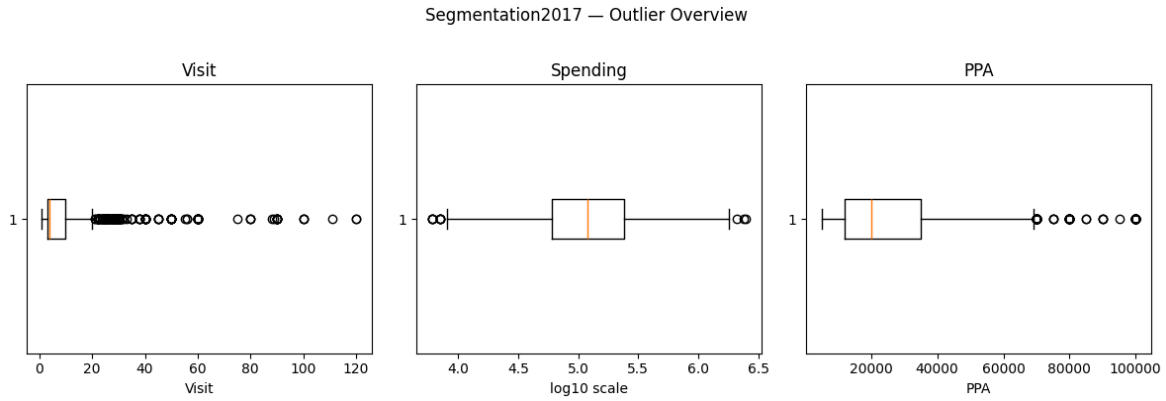


Figure 2.3. Outlier boxplots (Visit, Spending, PPA) (Source: Authors)

The segmentation structure is heavily skewed towards the mass and mass aspirational markets, consistent with Highlands Coffee's positioning in an affordable segment. However, the small premium and super-premium groups, along with heavy spenders, could represent opportunities for targeted strategies.

2.3.4. Brand Image (Brand_Image)

Descriptive analytics show that the Brand_Image has a large sample with 486.938 respondents with a median around 445.000, ranging from 89.100 to 863.754. This Brand_Image dataset was collected between 2017 and 2019 with an average year around 2018 and covered in six big cities with City_ID range from 1 to 6. The BrandImageKey variable spans a wide numerical range from 1 to 643.072 with a high standard deviation 185.639, which reflects the records are widely spread and highly diverse. Moreover, it has both mean and median at 321,536.5, showing a balanced distribution around the center.

Table 2.15. Descriptive statistics (numeric variables, Brand_Image dataset) (Source: Authors)

Variable	mean	median	std	min	max
BrandImageKey	321536.50	321536.5	185639.04	1.0	643072.0
RespondentID	486938.05	444693.0	272601.96	89100.0	863754.0

Year	2018.11	2018.0	0.79	2017.0	2019.0
CityID	2.47	2.0	1.54	1.0	6.0

2.3.5. Companion (Companion)

The Companion dataset with its descriptive statistics records information from an average of 465.471 respondents with median is 439.750, ranging from 89.100 to 863.754. The variable CompanionKey spasm from 1 to 20.739 with both mean and median at 10.370 and a relatively high standard variation around 5.987, which indicates a wide spread of values across different companion records.

Table 2.16. Descriptive statistics (numeric variables, Companion dataset) (Source: Authors)

Variable	mean	median	std	min	max
CompanionKey	10370.0	10370.0	5986.98	1.0	20739.0
RespondentID	465470.81	439750.0	272266.07	89100.0	863754.0

2.3.6. Day of Week (Dayofweek)

The descriptive statistics illustrates the Dayofweek dataset was responded from around 446.773 individuals with a median is 425.233, ranging from 89.100 to 863.754. This dataset was collected from 2017 to 2019, mostly in 2018. Respondents come from six big cities, with an average CityID of 3.57 and a median of 3. The variable VisitFreq shows that people typically visit the coffee shop about 4 times (median =4) while some reported up to 36 visits. The Dayofweek variable ranges from 1 to 31.536 with both mean and median at 15.768 and a relatively high standard deviation of about 9.104, showing a very wide spread of values.

Table 2.17. Descriptive statistics (numeric variables, Dayofweek dataset) (Source: Authors)

Variable	mean	median	std	min	max

DayOfWeekKey	15768.50	15768.5	9103.80	1.0	31536.0
RespondentID	446773.15	425233.0	269119.95	89100.0	863754.0
CityID	3.57	3.0	1.41	2.0	6.0
VisitFreq	3.82	4.0	1.99	1.0	36.0
Year	2018.00	2018.0	0.78	2017.0	2019.0

2.3.7. Day Part (Daypart)

The descriptive analytics of Daypart dataset contains responses from about 454.000 people with a median of 436.161, ranging from 89.100 to 863.754. This dataset was collected between 2017 and 2019 with most responses in 2018. Responses come from six cities with CityID ranges from 1 to 6. In addition, respondents reported visiting the coffee shop about 7 times with a mean of 6.95 a median of 4, but some reported 60 visits, which shows a big difference in visiting frequency. The DayPartKey variable ranges from 1 to 19.189 with both mean and median at 9.595 and a relatively high standard deviation of about 5.540, reflecting a wide range of coded values for different parts of the day.

Table 2.18. Descriptive statistics (numeric variables, Daypart dataset) (Source: Authors)

Variable	mean	median	std	min	max
DayPartKey	9595.0	9595.0	5539.53	1.0	19189.0
RespondentID	454165.15	436161.0	275768.32	89100.0	863754.0
CityID	2.75	2.0	1.62	1.0	6.0
VisitFreq	6.95	4.0	7.72	1.0	60.0
Year	2018.02	2018.0	0.81	2017.0	2019.0

2.3.8. Need State by Day & Daypart (NeedstateDaypart)

The descriptive statistics of NeedstateDaypart dataset shows that the dataset contains answers from around 632.000 people with a median of 788.931, ranging from 89.100 to 864.754. The dataset was collected from 2017 to 2019 and most responses were in 2019. The responses come from six big cities with CityID variables ranging from 1 to 6. The variable NeedsstateDayDayPartKey has values ranging from 1 to over 75.000 with both mean and median at 37.626 and a relatively high standard deviation of around 21.723, which show a very large set of coded categories related to customer needs and times of day.

Table 2.19. Descriptive statistics (numeric variables, NeedstateDaypart dataset) (Source: Authors)

Variable	mean	median	std	min	max
NeedstateDayDaypartKey	37626.0	37626.0	21723.24	1.0	75251.0
RespondentID	631898.21	788931.0	257218.33	89100.0	863754.0
CityID	2.66	2.0	1.62	1.0	6.0
Year	2018.52	2019.0	0.74	2017.0	2019.0

2.4 Data model

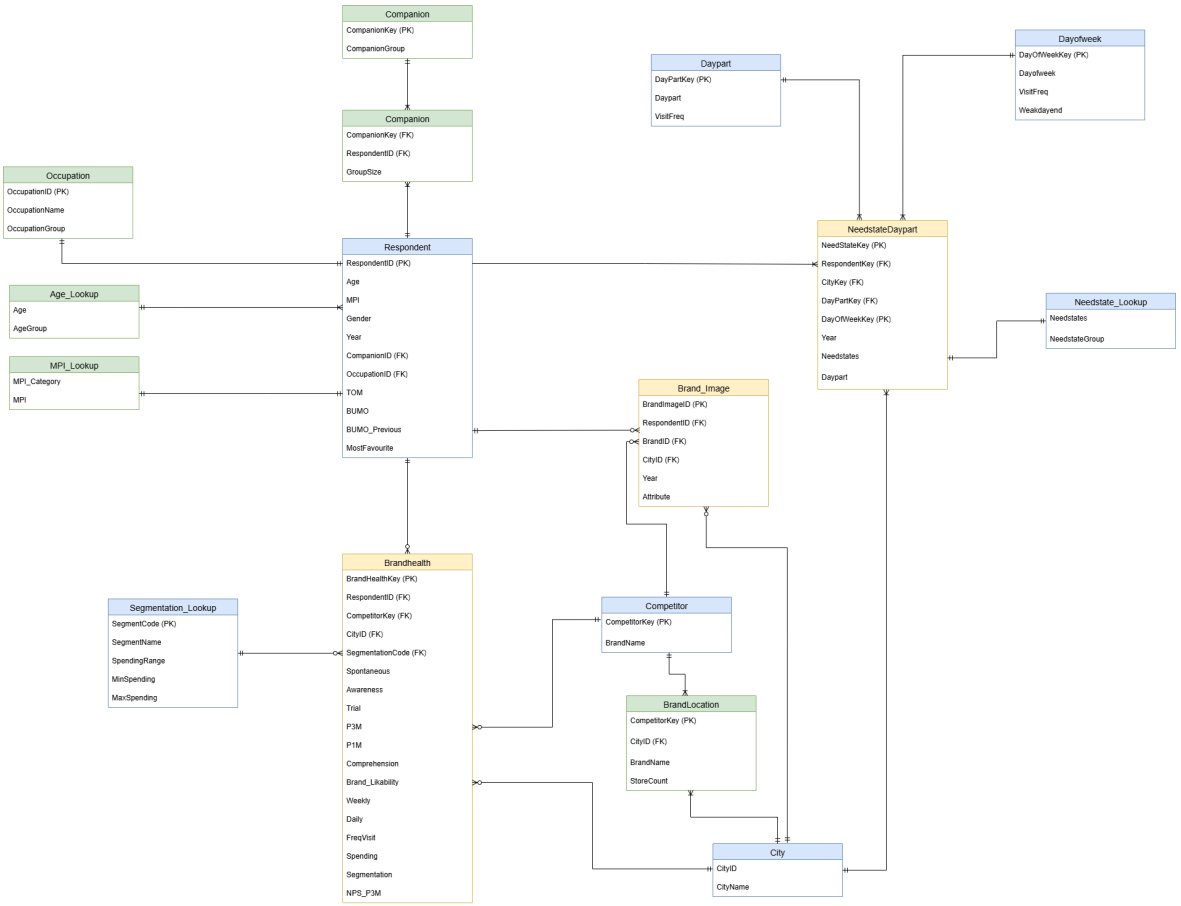


Figure 2.4. Entity-Relationship Diagram of Survey Data (Source: Authors)

After cleaning, the Entity-Relationship Diagram (ERD) is created to provide a complete image of the combined survey structure, interconnecting demographic, behavioural, brand, and contextual dimensions. It ensures consistency in the nine datasets and also supports respondent and aggregate analysis.

Core Entities

- **Respondent:** Central table containing demographic (Age, Gender, Occupation, MPI), behavioural (Top-of-Mind brand, Brand Used Most Often, Previous Brand, Most Favourite), and contextual attributes (City, Year, GroupSize). It is also a central table referring to nearly all other entities.

- **BrandHealth:** It captures brand-related performance indicators like awareness, trial, usage frequency, spend, segmentation, and NPS. It can be attached to Respondent, City, Brand, Competitor, and Segmentation.

- **Brand_Image:** Contains brand characteristic mappings, perceptions, and knowledge. Corresponds to Respondent, Brand, City, and Year.

- **NeedstateDaypart:** Stores consumers' needs at specific days and dayparts. Corresponds to Respondent, City, DayOfWeek, DayPart, and Needstate_Lookup.

Lookup and Dimension Tables

- **Segmentation_Lookup:** Standardizes customer segments (Mass, Mass Asp, Premium, Super Premium) based on spending range, to enrich BrandHealth data.

- **Occupation:** Classifies occupation into standard categories.

- **Age_Lookup:** Provides consistent age category assignment.

- **MPI_Lookup:** Categorizes monthly personal income (MPI) into pre-specified categories.

- **Needstate_Lookup:** Bins more detailed need states into wider categories.

Contextual Entities

- **Companion:** Information on visit companions (CompanionGroup, GroupSize) of Respondent.

- **DayOfWeek:** Weekday/Weekend visit-store patterns.

- **DayPart:** Tracks visit frequency across different periods of day.

- **Competitor:** Contains information regarding competing brands.

- **BrandLocation:** Retail location coverage by Brand and City, including numbers of stores to estimate market penetration.

- **City:** Provides standardized city names and codes to provide geographic consistency.

Relationships

- Respondent entity is core, connecting demographic data to brand comprehensions, health indicators, companions, and visit behavior.
- BrandHealth and Brand_Image are associated with Respondent and Brand, allowing cross-sectional brand performance to be analyzed.
- Segmentation_Lookup, Occupation, Age_Lookup, and MPI_Lookup enrich Respondent and BrandHealth.
- NeedstateDaypart incorporates multiple dimensions (Respondent, City, DayOfWeek, DayPart, Needstate) to account for visit motives to coffee shops.
- Competitor and BrandLocation provide an interface between city-level and brand-level marketplace research.

Chapter 3. Experimental results and evaluation

This chapter presents the experimental settings, parameter configuration, and evaluation of two analytical tasks: customer segmentation and churn prediction. The segmentation experiment applies the K-Prototypes algorithm to identify distinct customer groups while the churn prediction experiment employs classification models to forecast customer attrition. Each task uses slightly different datasets and preprocessing workflow, with classification model validation performed k-fold cross-validation strategies. The results collectively provide descriptive and predictive insights to support Highlands Coffee's targeted marketing and customer retention initiatives.

3.1 Customer Segmentation

The objective of applying clustering is to explore and identify distinct customer groups across the entire dataset. By applying the K-Prototypes clustering algorithm, this analysis aims to uncover patterns in demographics, spending, visit behavior, and motivation shared among different market segments. This segmentation forms the foundation for more targeted marketing strategies.

3.1.1 Dataset Description

The final dataset used for clustering is a consolidated table in which each row represents a unique customer profile. It summarizes individual demographic characteristics and behavioral patterns collected across three survey years (2017, 2018, and 2019).

RespondentID	Spending	Age	GroupSize	VisitFreq_dayofweek	Gender	Occupation	NeedstateGroup	CompanionGroup	DayOfWeek	DayPart	SpendingRange	CityID
344108	260000	30	4	6.666666667	Male	Unskilled Labor (worker	Drinking beverages	Friends	Friday	Before 9 AM	100000+	3
344465	175000	18	5	2.333333333	Female	Pupil / Student	Relaxing & entertainr	Friends	Saturday	5 PM - before 9 PM	100000+	3
344532	36000	28	3	1.5	Female	Skilled Labor (tailor	Relaxing & entertainr	Family	Saturday	5 PM - before 9 PM	25000 - 59000	3
344533	80000	28	4	1.5	Male	Broker/ Service provider with no employee	Socializing	Colleagues / Business partner	Saturday	5 PM - before 9 PM	25000 - 59000	3
344534	60000	21	3	4	Female	Unskilled Labor (worker	Relaxing & entertainr	Family	Saturday	5 PM - before 9 PM	60000 - 99000	3
344625	224000	55	7	4	Female	Housewife	Drinking beverages	Friends	Friday	Before 9 AM	100000+	3
344634	300000	24	4	7.714285714	Male	Semi-skilled labor (salesperson	Socializing	Boyfriend / Girlfriend	Friday	11 AM - before 2 PM	100000+	3
344635	100000	46	3	4	Female	Skilled Labor (tailor	Socializing	Family	Monday	Before 9 AM	0 - 25000	3
344636	64000	26	4	2	Female	Semi-skilled labor (salesperson	Socializing	Colleagues / Business partner	Sunday	9 AM - before 11 AM	60000 - 99000	3
344638	20000	34	2	2	Female	Housewife	Socializing	Family	Sunday	5 PM - before 9 PM	0 - 25000	3
344639	300000	25	2	3	Female	Lecturer / Teacher	Drinking beverages	Friends	Saturday	5 PM - before 9 PM	100000+	3
344641	72000	32	2	4	Female	Officer - Staff level	Drinking beverages	Family	Saturday	9 AM - before 11 AM	0 - 25000	3
344642	56000	36	2	4	Female	Small Business (small shop owner	Drinking beverages	Boyfriend / Girlfriend	Saturday	Before 9 AM	0 - 25000	3

Figure 3.1. Aggregated data for Clustering (Source: Authors)

The dataset contains both numerical and categorical attributes. Each customer, identified by a unique RespondentID, occupies exactly one row in the final dataset.

Table 3.1. Data dictionary of the aggregated data for clustering (Source: Authors)

Category	Column	Meaning
1.Demographic Factors	Gender	Respondent's gender
	Age	Customer's age in years
	Occupation	Job type or work status
	CityID	Identifier for respondent's city
2.Spending Behavior	Spending	Total amount spent by the customer
	SpendingRange	Spending tier
3. Visit Patterns	GroupSize	Number of people in the customer's group
	VisitFreq_dayofweek	Average number of visits per week
	DayOfWeek	Typical day of visit

	DayPart	Typical time of visit
4. Behavioral & Contextual Factors	NeedstateGroup	Primary motivation or consumption need
	CompanionGroup	Typical companions

3.1.2 Data Preprocessing

Variable selections

Multiple data sources were integrated to construct the final dataset, including *Respondent*, *BrandHealth*, *Segmentation*, *NeedStateDayPart*, *Companion*, *DayOfWeek*, and *DayPart*. After resolving minor inconsistencies such as encoding and naming errors, the related tables were merged using common identifiers which is *RespondentID* to create unified yearly datasets. Irrelevant or highly incomplete columns were then removed, data formats were standardized across years and the yearly datasets were concatenated into a single comprehensive table for subsequent aggregation and clustering analysis.

Aggregation by Customer

Since each respondent could appear multiple times across surveys, a group-by aggregation was applied to consolidate all records of the same customer into one entry:

- **Mean:** *GroupSize, VisitFreq_dayofweek*
- **Max:** *Spending, Age.*
- **Mode:** for other categorical variables.

This aggregation process ensured that the final dataset accurately reflects each customer's overall characteristics while removing redundancy and noise from repeated entries, thus forming a solid base for clustering analysis.

3.1.3 Parameter Setting and Experimental Design

The K-Prototypes algorithm was chosen for its capability to handle mixed-type data by combining the strengths of K-Means for numerical features and K-Modes for categorical ones.

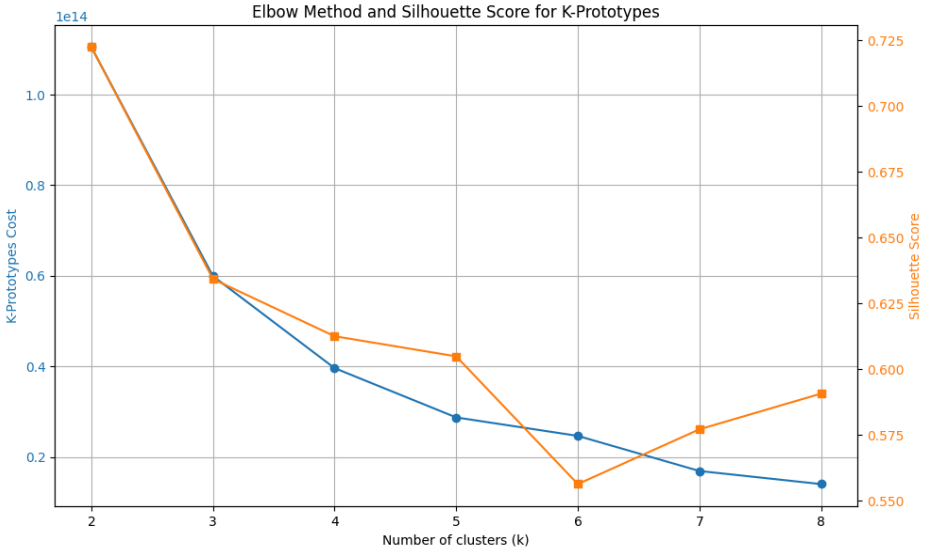


Figure 3.2. Elbow Method and Silhouette by number of clusters (Source: Authors)

The number of clusters (k) was tested from 2 to 10 using the Elbow Method and Silhouette Score. Specifically, compare the K-Prototypes cost (blue line, left axis) and Silhouette score (orange line, right axis) across a number of clusters. As k increases, the clustering cost decreases sharply until around $k = 5$, after which the rate of improvement slows significantly, forming a visible "elbow" point. This indicates that beyond five clusters, additional divisions yield diminishing returns in reducing within-cluster dissimilarity. Meanwhile, the Silhouette score reaches relatively high values between $k = 4$ and $k = 5$, peaking near $k = 3$ but remaining stable around $k = 5$, suggesting that both cohesion and separation between clusters are reasonably balanced in this range.

Considering both metrics together, $k = 5$ provides an optimal trade-off between model simplicity and clustering quality, supporting the choice of five customer segments for subsequent analysis.

3.2 Customer Churn Prediction (Classification)

The objective of the classification experiment is to develop predictive models capable of identifying customers who are likely to churn, specifically to stop engaging with or purchasing Highland coffee by applying and comparing multiple machine learning algorithms: *Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM)*. After finding the best model, the feature importance finding is implemented to find features that heavily affect customer decision to churn.

3.2.1 Dataset Description

RespondentID	Gender	MPI	OccupationGroup	CompanionGroup	Attribute	DayOfWeek_mode	DayPart_mode	GroupSize	Age	Spending	VisitFreq_mean	is_churn
89100	Female	VND 4.5m	Self Employed - Small	Boyfriend / Girlfriend	Ambiance &	Friday	11 AM - before 2 PI	4	39	203333.33	2.4	0
89101	Female	VND 4.5m	None Working	Family	Ambiance &	Saturday	5 PM - before 9 PM	4	33	35000	1.5	0
89102	Male	VND 4.5m	None Working	Friends	Ambiance &	Friday	11 AM - before 2 PI	4	17	300000	2	0
89613	Male	VND 4.5m	White Collar	Colleagues / Business	Ambiance &	Saturday	5 PM - before 9 PM	3	55	96000	4	1
89616	Male	VND 4.5m	Self Employed - Small	Friends	Brand Perce	Friday	9 AM - before 11 AI	2	60	94666.667	5	1
89618	Female	VND 4.5m	None Working	Boyfriend / Girlfriend	Ambiance &	Saturday	2 PM - before 5 PM	2	19	44500	1.5	0
89963	Female	VND 4.5m	None Working	Alone	Ambiance &	Friday	2 PM - before 5 PM	4	49	360000	4.285714286	0
89965	Female	VND 4.5m	Blue Collar	Friends	Ambiance &	Monday	9 AM - before 11 AI	4	39	240000	2.666666667	1

Figure 3.3. Aggregated data for Classification (Source: Authors)

The final dataset used for the churn classification task, was constructed to predict whether a customer is likely to churn. Each row represents a single customer, uniquely identified by RespondentID, and contains a comprehensive profile combining demographic, behavioral, and perceptual information.

Table 3.2. Data dictionary of aggregated data for Classification (Source: Authors)

Category	Column	Meaning
1. Demographic Factors	Age	Customer's age in years
	Gender	Respondent's gender
	OccupationGroup	Customer's job type or occupational group
	MPI	Monthly personal income (income level indicator)

2. Behavioral Factors	GroupSize	Number of people in the customer's group during café visits
	CompanionGroup	Typical companions when visiting
	DayOfWeek_mode	The day of the week the customer most frequently visits
	DayPart_mode	The most common time of day the customer visits
	VisitFreq_mean	Average visit frequency, representing how often the customer visits
3. Financial Factor	Spending	Total amount of money spent by the customer
4. Brand Perception	Attribute	Key brand attribute associated with the customer's perception (generalized from survey responses)
5. Target Variable	is_churn	Indicates whether the customer has churned (1 = churned, 0 = active)

3.2.3 Data Preprocessing

Target Variable Creation

From the BrandHealth table, only records related to Highlands Coffee were retained. Based on the field Used_P3M (indicating whether the customer had used the brand in the past three months), a new binary variable *is_churn* was generated:

- is_churn = 1: The customer has churned
- is_churn = 0: The customer remains active.

Variable selections

The *Respondents* table, containing demographic information, was merged with the using RespondentID as the common identifier. Behavioral data from *Companion, DayOfWeek, DayPart*, and *BrandImage* tables were then processed separately before integration. For *DayOfWeek* and *DayPart*, visit data were aggregated by respondents to obtain the mode (most frequent visit time) and mean (average visit frequency). The *BrandImage* attributes were standardized by mapping detailed perceptions to broader categories using an AttributeMapping reference table. All processed sources were then combined into a unified dataset.

Data Imputation

To fix the issue of missing value, a multi-step imputation strategy was employed. The clustered data was utilized to perform filling null values effectively.

- Missing values in *Spending* and *VisitFreq_mean* were filled using the *Spending* and *VisitFreq_mean* of other respondents in the same cluster
- Any remaining numerical missing values were replaced with the mean of the respective columns.
- For categorical variables including *MPI* and *DayOfWeek_mode*, missing entries were imputed using the mode of the corresponding cluster, as defined in the earlier clustering analysis.

Feature engineering

To maintain consistency and prevent data leakage, all preprocessing and feature transformation steps were encapsulated in a unified machine learning pipeline.

- **Cross-validation target encoding:** Categorical features were encoded using TargetEncoder, which replaces each category with the mean churn rate (*is_churn*) for that category across 5 fold of splitted data. This method captures the statistical relationship

between categorical levels and the churn outcome, improving model interpretability and performance while preventing data leakage when applying cross-validation.

- **Scaling:** Numerical features were standardized using StandardScaler to ensure a mean of 0 and standard deviation of 1. This normalization step is crucial for sensitivity to feature scales like SVM.

- **5-Fold Cross-Validation:** The dataset was split into five subsets; in each iteration, four folds were used for training and one for testing.

- **Handling Data Imbalance:** Since the dataset exhibited class when imbalance non-churn customers (62.9%) significantly outnumber churners (37.1%) , the SMOTE (Synthetic Minority Over-sampling Technique) was applied to generate synthetic examples of the minority class. Importantly, SMOTE was used only within the training folds during cross-validation, also to prevent data leakage and ensure fair evaluation on unseen test data.

3.2.4 Experimental Design

Stage 1: Initial Models Comparison

In the first stage, all models were trained using default parameter settings to provide a fair comparison of their out-of-the-box performance. The following configurations were used:

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- Support Vector Machine

Table 3.3. Model Performance - Default Model (Source: Authors)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	66.34%	89.53%	10.61%	18.96%	59.45%

Random Forest	67.67%	61.46%	34.98%	44.55%	68.25%
XGBoost	68.44%	64.82%	33.01%	43.7%	70.48%
SVC	66.16%	89.43%	10.08%	18.11%	60.43%

The table illustrates the average 5-fold cross-validation performance of four classification models with default parameters. XGBoost achieved the best overall performance with the highest accuracy (68.44%) and ROC-AUC (70.48%), indicating stronger discriminative ability compared to other models. Its F1-score (43.7%) also surpassed that of Logistic Regression and SVM, reflecting a better balance between precision and recall, though recall remains relatively modest (33.01%), suggesting that the model still misses a notable portion of actual churners. The Random Forest model produced comparable accuracy (67.67%) and a similar F1-score (44.55%), but slightly lower AUC (68.25%), implying that while it identifies churners somewhat effectively, it lacks the finer discriminative power of XGBoost. In contrast, Logistic Regression and SVM (RBF) yielded high precision (~89%) but extremely low recall (~10%), leading to poor F1-scores (<20%). This indicates that these linear or margin-based models tend to classify most customers as non-churners, capturing few true churn cases, probably due to data imbalance and non-linear patterns that they cannot fully model.

Table 3.4. Model Performance With SMOTE (Source: Authors)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	60.37%	45.64%	35.09%	39.65%	59.40%
Random Forest	66.69%	58.18%	36.85%	45.09%	66.91%

XGBoost	67.93%	62.16%	34.98%	44.74%	70.14%
SVC	60.89%	47.27%	46.05%	46.59%	61.36%

After applying SMOTE to balance the class distribution, the overall performance metrics demonstrate notable changes, particularly in recall and F1-score, indicating that the oversampling technique effectively improved the models' ability to detect more churners. Among the tested models, XGBoost again achieved the best overall performance, with the highest accuracy (67.93%) and ROC-AUC (70.14%). Its precision (62.16%) and F1-score (44.74%) remained competitive, showing that it maintained strong discriminative power even after oversampling. However, recall (~35%) suggests that further improvements may still be needed to enhance sensitivity to churn cases. The Random Forest model achieved similar performance, with slightly lower AUC (66.91%) and F1-score (45.09%), reflecting solid but less stable recall (37%). It remains a reliable ensemble method but underperforms XGBoost in distinguishing borderline churners. Interestingly, SVM demonstrated a substantial improvement compared to the non-SMOTE scenario. Its recall rose sharply from 10% to 46%, while F1-score increased to 46.59%, the highest among all models. This suggests that balancing the dataset helped the SVM better capture minority-class patterns, although its overall AUC (61.36%) and accuracy (60.86%) remained lower than the tree-based models. Logistic Regression, despite gaining higher recall (35% vs. 10% previously), still produced the lowest accuracy (60.37%) and moderate F1-score (39.65%), implying limited capability in modeling the complex, non-linear relationships underlying churn behavior.

With the application of SMOTE, models really show significant improvement of classification results. Although improved model sensitivity, particularly for SVM. but XGBoost remained the most consistent and well-balanced performer in terms of overall predictive power (accuracy and AUC).

Stage 2: Hyper Parameter Tuning for XGboost

To optimize XGBoost performance, GridSearchCV was employed to systematically search for the most effective combination of hyperparameters. This method evaluates all parameter combinations within a predefined grid using cross-validation, ensuring the model achieves an optimal balance between bias and variance. After searching, GridSearchCV identified the following optimal parameter configuration for XGBoost:

Table 3.5. Optimal Parameter for XGBoost (Source: Authors)

Parameter	Value	Description
n_estimators	100	Number of boosting rounds or trees to be built sequentially.
max_depth	3	Maximum depth of each decision tree.
learning_rate	0.1	Step size shrinkage that controls how much each tree contributes to the final model.
subsample	0.8	Fraction of the training samples randomly selected to grow each tree.
colsample_bytree	0.8	Fraction of features randomly selected for each tree.

Table 3.6. Model Performance Interpretation (Before vs. After Fine-Tuning) (Source: Authors)

Metric	Before Fine-Tuning	After Fine-Tuning	Change
Accuracy	67.93%	66.12 %	−1.81 %
Precision	62.16 %	55.49 %	−6.67 %

Recall	34.98 %	44.70 %	+9.72 %
F1-score	44.74 %	49.49 %	+4.75 %
ROC-AUC	70.14%	68.72 %	−1.42 %

After fine-tuning and applying SMOTE, the XGBoost model exhibited a clear improvement in recall and F1-score, demonstrating enhanced sensitivity to churn cases and better balance between precision and recall. Specifically, recall increased by 9.72%, indicating that the model captured a higher proportion of actual churners, thereby reducing false negatives which is a critical improvement for churn prediction tasks. Although precision decreased by 6.67%, this reduction is a typical trade-off when recall improves, reflecting a more inclusive detection of potential churners at the cost of a few additional false positives. Similarly, accuracy dropped marginally (−1.81%) and ROC-AUC decreased slightly (−1.42%), but these changes remain within acceptable bounds, suggesting that the model's discriminative ability between churn and non-churn classes remained stable.

Overall, in the context of churn prediction for the F&B industry, this trade-off is both acceptable and strategically valuable. Identifying as many potential churners as possible enables proactive retention actions. Therefore, the fine-tuned XGBoost model delivers a more balanced and business-aligned performance, favoring improved recall and F1-score, a desirable outcome where minimizing customer loss is prioritized over reducing false alarms.

Chapter 4. Visualization and Discussion

This chapter represents the visualization of the data from the Vietnamese coffee brands insights into interactive dashboards. This chapter mainly focuses on interpreting the raw data into meaningful insights which support business decisions and give the views of customer behaviors, brand positions and performance. Moreover, it shows the analysis of customer churn, identifying differences in characteristics from customers. Overall, this chapter visualizes to offer a comprehensive view of the market performance and customer analysis for Highland Coffee, therefore providing the brand with strategic planning and business development.

4.1. Exploratory Data Analysis (EDA)

4.1.1. Customer demographics

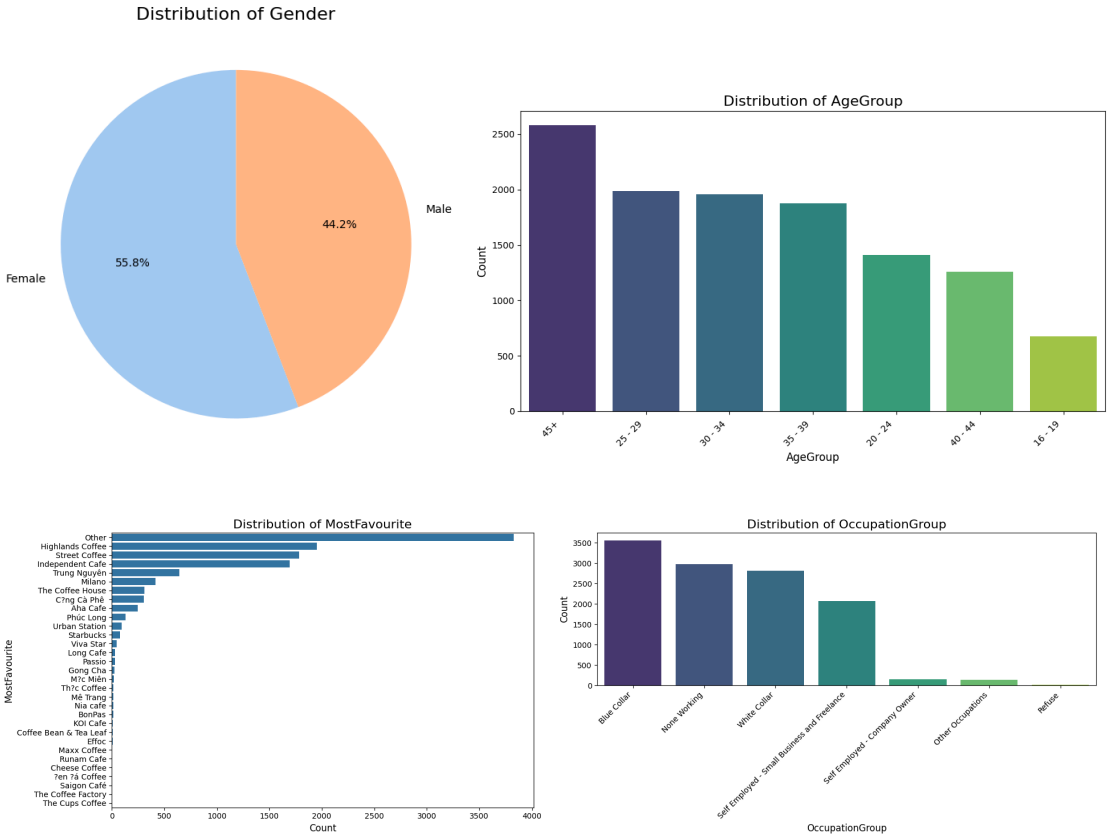
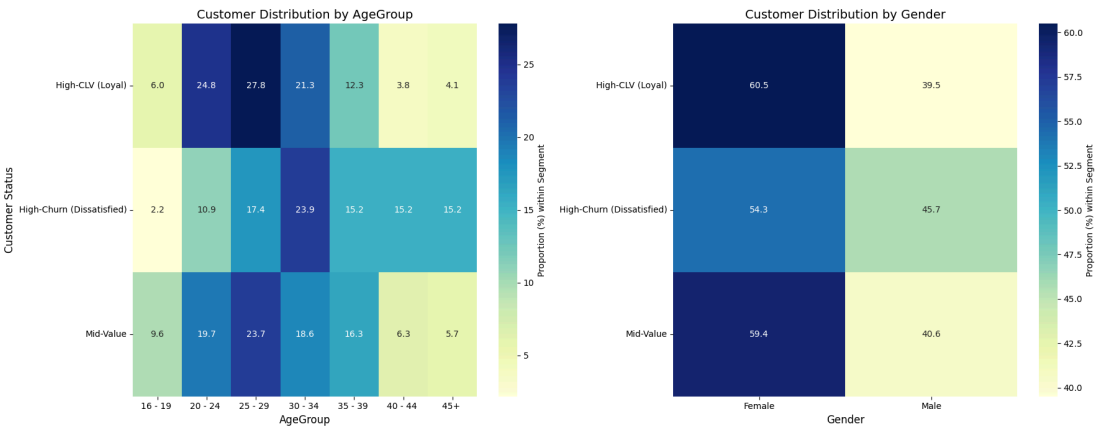


Figure 4.1. Charts about distribution of Highlands's customer demographics (Source: Authors)

Based on this demographic analysis, we can paint a clear picture of the typical Highlands Coffee customer. The brand holds a strong market position, with 16.64% of all respondents citing it as their favorite brand, making it a leader in customer preference. This success is built on a broad demographic appeal that reflects the general market's composition. The customer base is primarily composed of young to middle-aged adults, with the largest segments falling into the 25-29 (22.38%) and 30-34 (20.06%) age groups. The customer base also shows a slight female majority, with 55.79% of respondents being female. In terms of occupation, the customer base is diverse, with the most common groups being Blue Collar (30.38%), None Working (25.42%), and White Collar (23.94%). This wide appeal across age, gender, and occupation is a key factor in the brand's success.

→ Highlands Coffee has successfully established itself as a brand with broad market reach. Its ability to attract customers from diverse age groups and occupations has been a key factor in its market-leading position. The brand's customer profile aligns with the general market trends, indicating its appeal is not limited to a single niche. This broad demographic appeal is a significant strategic advantage, providing the brand with a stable and diverse customer base.

4.1.2. Customer CLV and Churn



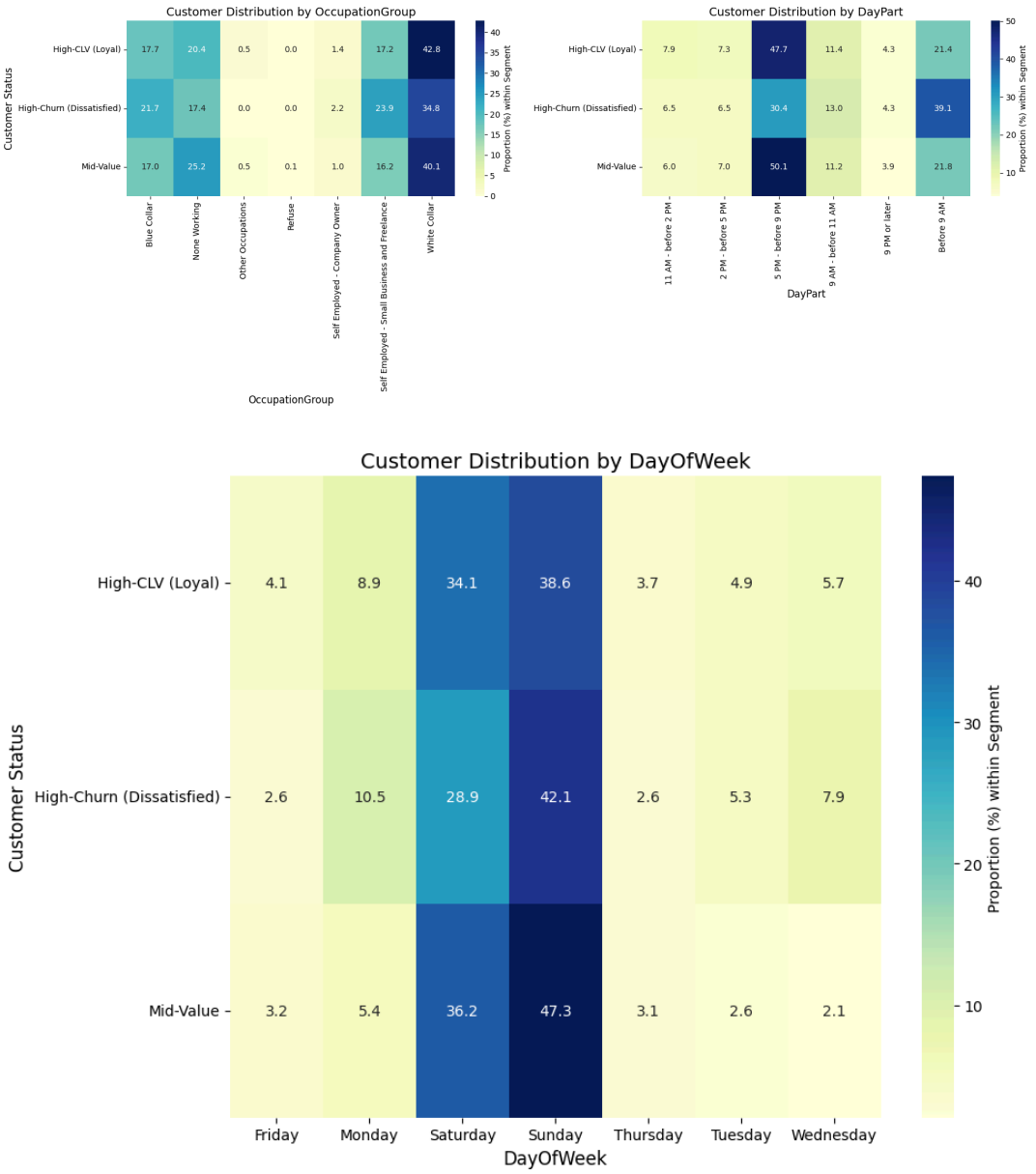


Figure 4.2. Percentage of high-CLV and high-churn customers (Source: Authors)

Based on the detailed data analysis of Highlands Coffee's customer base, we can draw significant insights into the key differences between high-value and high-risk customers.

Demographic profile (AgeGroup, Gender and Occupation)

- **Age:** The analysis clearly shows that High-CLV (Loyal) customers are younger, with the highest concentration in the 25–29 age group (27.79%), closely followed by the 20–24 group (24.80%). This highlights young professionals as the brand's financial

backbone. In stark contrast, the High-Churn (Dissatisfied) group is older, with its largest proportion in the 30–34 age group (23.91%), and a significantly higher representation in the 45+ group (15.22%)—nearly four times the rate of the loyal segment.

- **Gender:** While Female customers dominate the loyal segment (60.49%), Male customers (45.65%) are proportionally much higher in the High-Churn group than in the High-CLV group (39.51%), suggesting men who frequent Highlands are more prone to dissatisfaction and defection.

- **Occupation:** The loyal segment is heavily characterized by the White Collar group (42.78%). However, the High-Churn group shows significant risk associated with Self Employed/Freelance (23.91%) and Blue Collar (21.74%) professions, indicating that customers with less structured work environments or manual labor roles are far more likely to be dissatisfied with the brand experience.

Visit Behavior (DayOfWeek & DayPart)

- **Time of Visit (DayPart):** Customer loyalty is strongly correlated with evening usage. High-CLV customers show a dominant preference for the Evening (5 PM - before 9 PM) time slot (47.70%), reinforcing the role of Highlands as a social/leisure destination. The key behavioral differentiator is the morning rush: High-Churn customers have a disproportionately high tendency to visit Before 9 AM (39.13%), nearly double the rate of High-CLV customers (21.41%). This strongly suggests that dissatisfaction is rooted in convenience and transactional speed during the morning commute.

- **Day of Visit (DayOfWeek):** Both loyal and churning customers are primarily weekend users, with the highest usage on Sunday (38.62% for High-CLV, 42.11% for High-Churn). However, the High-Churn group shows a higher tendency to visit on Mondays (10.53%) compared to the High-CLV group (8.94%). Coupled with the morning preference, this indicates that the High-Churn segment's usage is likely linked to necessity or daily routine during weekdays, not leisure.

These findings generate two distinct customer profiles:

=> For High-CLV customers: These are young, White collar females (20–29) who treat Highlands as a destination for socializing and leisure, primarily visiting on weekend evenings. Their loyalty is built on the in-store experience and ambience.

=> For high-churn customers: This group is typically older (30s+), Self employed/Blue collar men who use Highlands for a quick, convenient transaction, with a peak during the early morning commute. Their high risk of defection is primarily driven by the brand's failure to meet the demands of fast, convenient service.

4.1.3. Attribute frequency

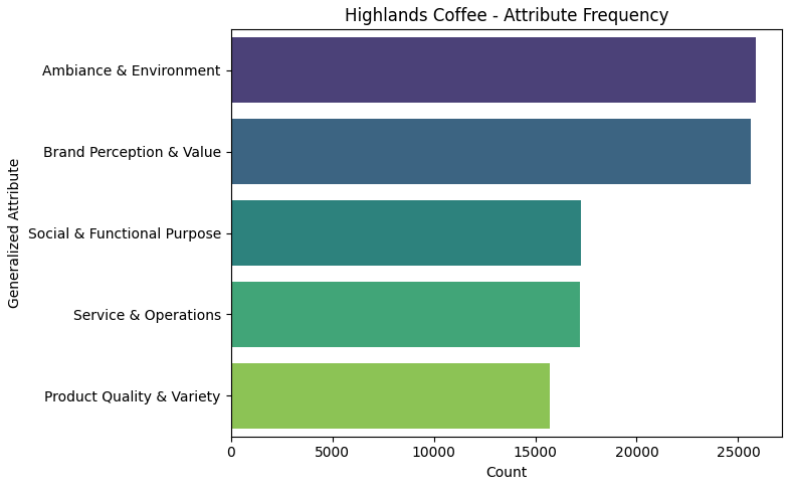


Figure 4.3. Count of attribute frequency that customers remember Highlands (Source: Authors)

For Highlands Coffee, generalized attribute groups stand out most in what customers recall:

- **Ambiance & Environment and Brand Perception & Value** dominate, far surpassing the other categories. This indicates that customers primarily remember Highlands Coffee for its store atmosphere, which may include its design, comfort, or convenient locations. In addition, Highlands' brand value is very strong and clearly recognized by customers, potentially related to its prestige and a premium yet accessible brand image.

- **Social & Functional Purpose and Service & Operations** are mid-tier. The Highlands is seen as a place suitable for study, work or socialising, but these aspects are less defining than its atmosphere and brand image.

- **Product Quality & Variety** ranks lowest. This is quite surprising for a major coffee chain. It may imply that while Highlands' beverages are acceptable, they are not the most prominent factor that makes customers remember this brand.

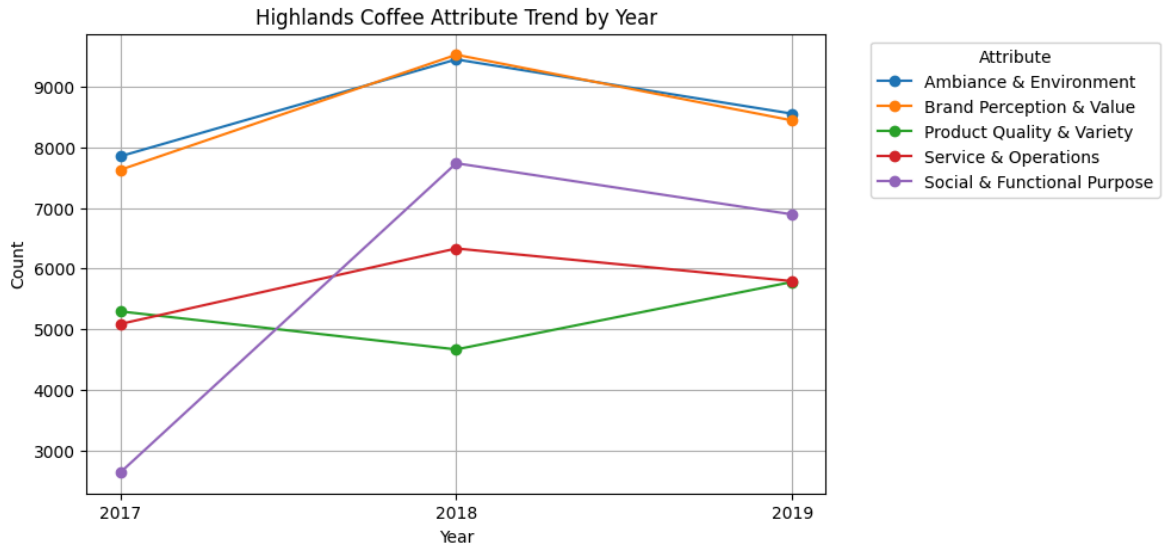


Figure 4.4. Chart of Highlands Coffee attribute trend by year (Source: Authors)

The "Highlands Coffee Attribute Trend by Year" chart shows the change in customer perception of the brand from 2017 to 2019.

- **Ambiance & Environment** and **Brand Perception & Value** consistently remain the top two factors, peaking in 2018 and staying at a high level in 2019. This confirms that Highlands' strategy of creating a comfortable, familiar space and building a strong brand value has been successful and effective over time.

- **Social & Functional Purpose** saw a significant increase from 2017 to 2018, showing that Highlands was increasingly seen as an ideal place for meeting friends, working, or studying. However, this factor saw a slight decline in 2019.

- **Service & Operations** also trended upward from 2017 to 2018, reflecting improvements in service quality. However, this factor also saw a slight decrease in 2019.

- **Product Quality & Variety** showed a continuous decline from 2017 to 2018, with a slight recovery in 2019, but it remains the attribute with the lowest number of mentions. This indicates that product quality is not the most prominent factor that customers remember about Highlands and even showed a downward trend during the 2017-2018 period.

Overall, the trend shows that Highlands Coffee's brand strength lies firmly in its ambiance and brand value, while product quality remains its weakest point despite slight improvements.

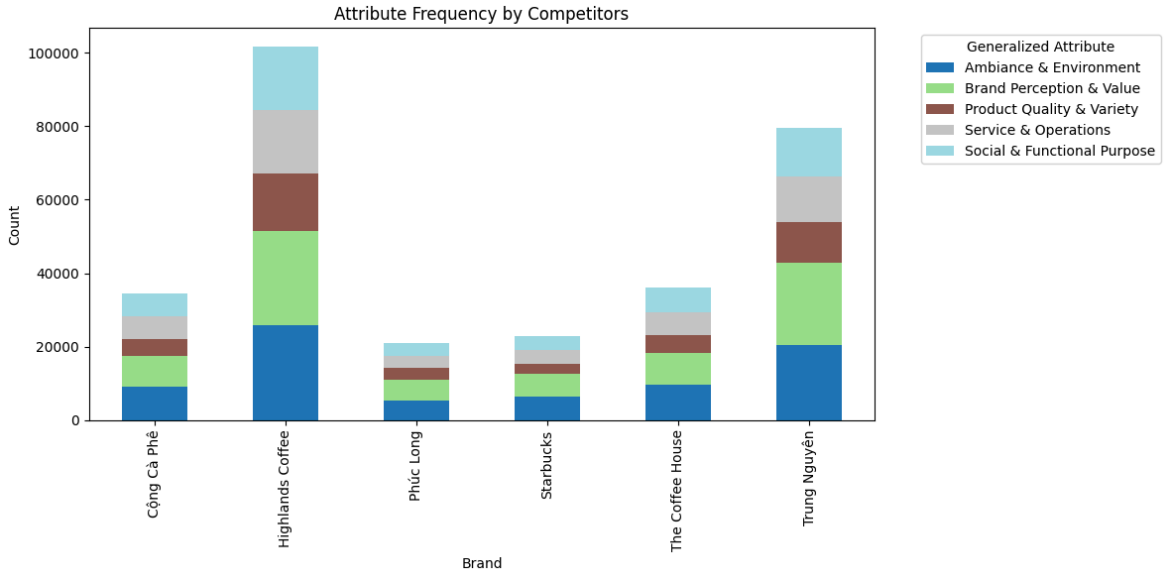


Figure 4.5. Chart of attribute frequency by competitors (Source: Authors)

When compared with its competitors, most coffee chains share a strong association with these two attributes. Atmosphere and brand value are not exclusive to Highlands or Starbucks; rather, they represent the core factors that customers prioritize when recalling coffee brands.

- **Highlands Coffee** has the highest total number of attribute mentions, indicating that it is one of the most prominent and discussed brands.

- **Trung Nguyên** ranks second, with a more balanced presence across Ambiance, Brand Value, and Social Purpose, reflecting its role as both a meeting place and a cultural symbol of Vietnamese coffee.

- **Cộng Cà Phê** is smaller in scale but distinctively remembered for its nostalgic, unique Ambiance, though less so for products or operations.

- **Phúc Long** has a good balance across all attributes, suggesting that customers perceive the brand in a more holistic way.

- **Starbucks** is strongly tied to Brand Perception & Value, reflecting its global brand power rather than local product or space experiences.

- **The Coffee House** still has a noticeably higher proportion of the Product Quality & Variety attribute, making it a distinct highlight for their brand.

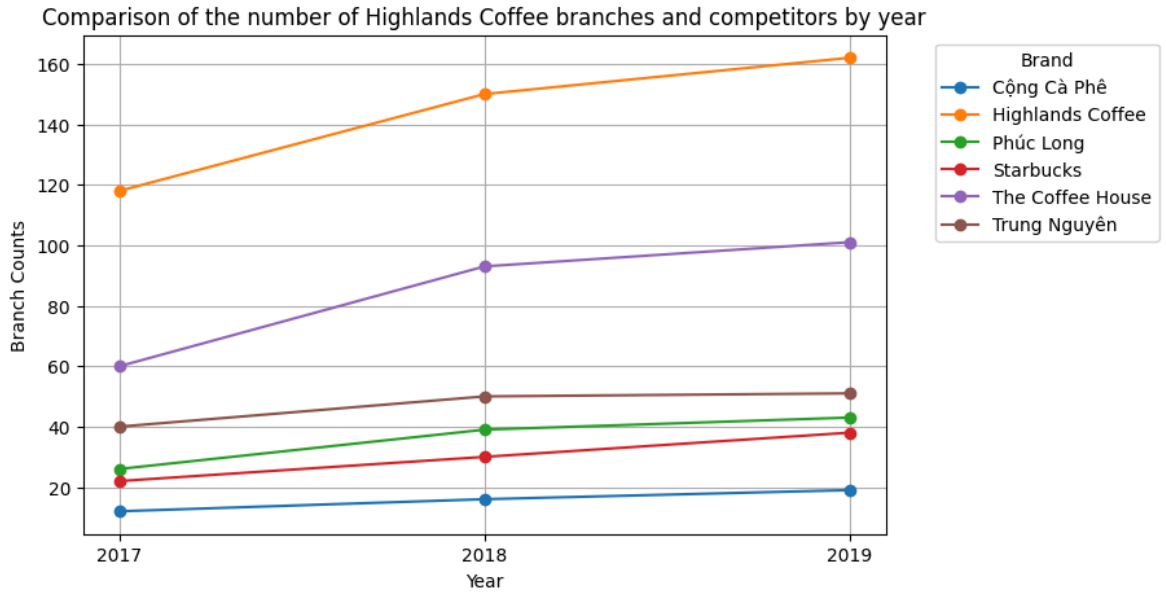


Figure 4.6. Chart about comparison of the number of Highlands coffee branches and competitors by year (Source: Authors)

The chart illustrates the growth in the number of Highlands Coffee branches compared to its competitors from 2017 to 2019. Highlands Coffee consistently maintains the highest number of branches, expanding rapidly from 118 in 2017 to over 160 in 2019. The Coffee House shows steady growth, reaching just above 100 branches by 2019, making it the closest competitor. Trung Nguyên, Phúc Long, and Starbucks exhibit slower growth, while Cộng Cà Phê remains the smallest player with minimal expansion during the period.

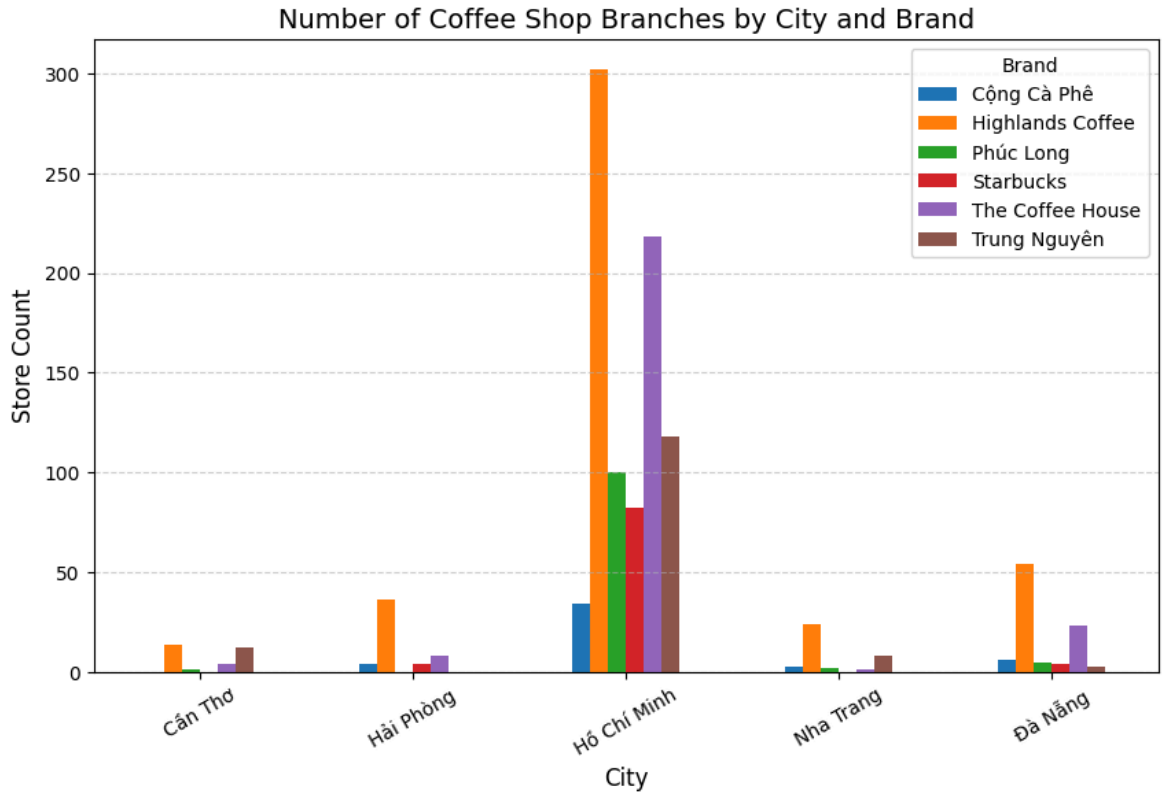


Figure 4.7. Count of number of coffee shop branches by city and brand (Source: Authors)

Highlands Coffee dominates across all cities, with a particularly strong presence in Ho Chi Minh City. The Coffee House is concentrated mainly in Ho Chi Minh City, with limited presence elsewhere. Phúc Long and Starbucks are also notable in Ho Chi Minh City but remain much smaller in scale compared to Highlands Coffee and The Coffee House. Cộng Cà Phê operates on a modest scale yet maintains a presence in most cities. Trung Nguyên, meanwhile, shows a relatively balanced distribution but still has its strongest foothold in Ho Chi Minh City. Overall, Highlands Coffee holds an absolute advantage in the number of stores across all cities, especially in Ho Chi Minh City.

4.2. Dashboard and Visualization design

BrandHealth Overview

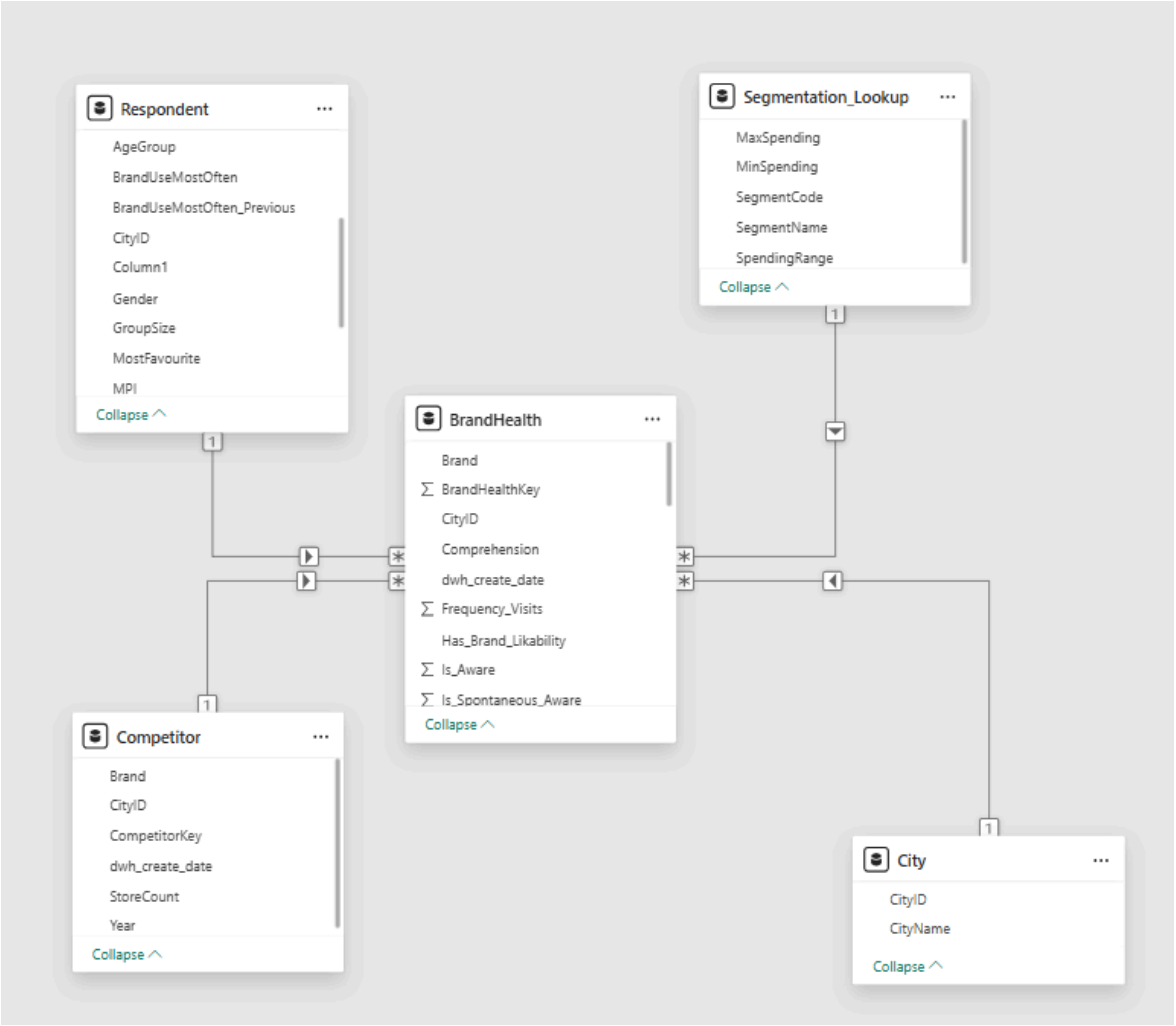


Figure 4.8. Data model of BrandHealth (Source: Authors)

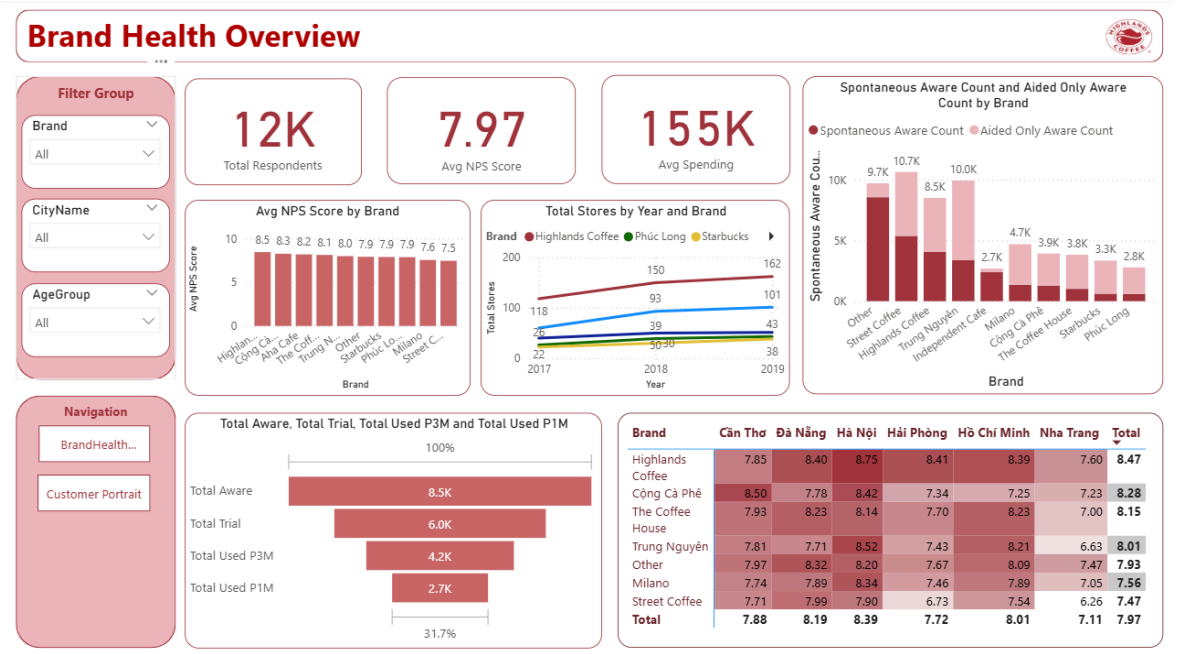


Figure 4.9. Brand Health Overview(Source: Authors)

This report page is designed to provide a panoramic view of the competitive landscape within the chain coffee market. The information focuses on core brand health metrics and market scale, helping to establish the company's current position relative to key competitors.

This report provides a comprehensive market overview through key performance indicators. The industry-wide average Net Promoter Score (NPS) stands at 7.97, indicating a relatively healthy level of customer satisfaction. Concurrently, the average spending is recorded at 155K VND. Delving into individual brand performance reveals distinct differences in customer loyalty. Highlands Coffee and Cong Ca Phe are the current leaders in NPS, with scores of 8.5 and 8.3, respectively.

Between 2017 and 2018, both Highlands Coffee and The Coffee House experienced rapid expansion. Highlands Coffee increased its number of stores from 118 to 150, while The Coffee House grew from 60 to 93. However, from 2018 to 2019, this growth pace slowed down, with Highlands Coffee reaching 162 stores and The Coffee House 101. This parallel slowdown indicates a strategic transition from rapid expansion to a more sustained and stable growth phase.

Among the top coffee brands, Street Coffee ranks highest in total awareness with approximately 10.7K, followed by Trung Nguyen with 10K, and Highlands Coffee with 8.5K. Milano, Cộng Cà Phê, and The Coffee House follow with awareness levels ranging from 4.7K to 3.3K. Notably, Highlands Coffee stands out for its high spontaneous awareness, suggesting that consumers can recall the brand easily without prompts—an indicator of strong top-of-mind recognition and brand familiarity compared to its competitors.

The funnel for Highlands Coffee shows a clear drop across the customer journey. Out of 8.5K people aware of the brand, about 6.0K have tried it, but only 4.2K used it in the past three months and 2.7K in the past month. This means while the brand achieves good awareness and trial rates, the retention and frequency of use remain limited. The gap between trial and recent usage suggests that Highlands Coffee may need to strengthen customer loyalty and encourage repeat visits.

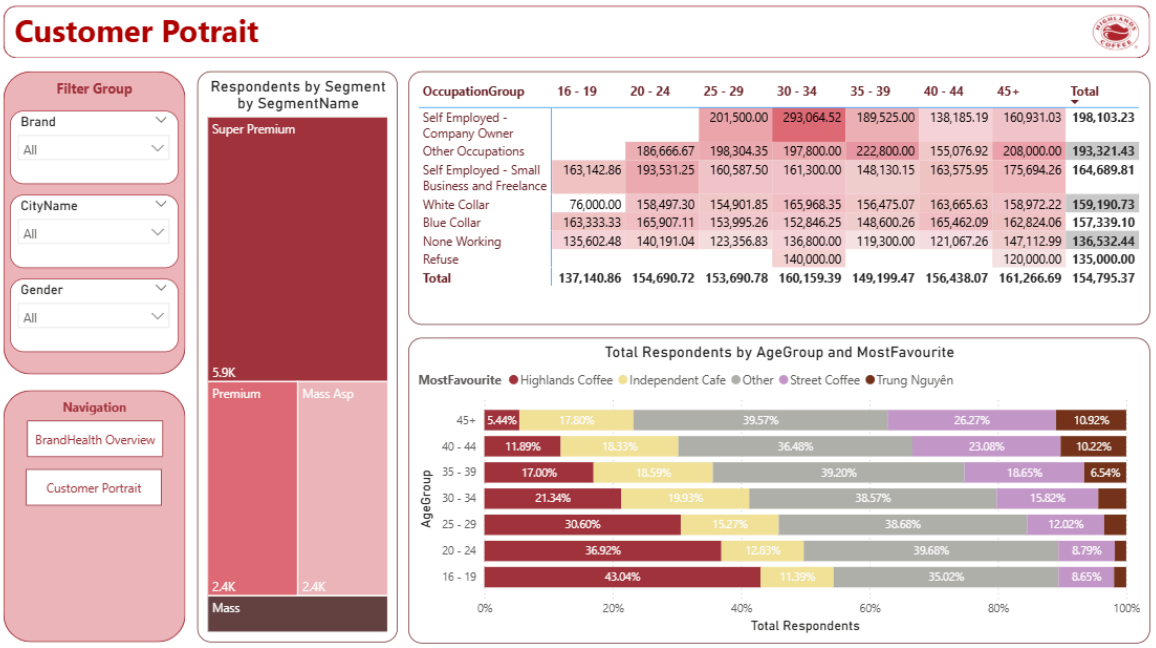


Figure 4.10. Customer Portrait (Source: Authors)

The Customer Portrait dashboard provides an overview of respondents' segmentation, occupation, and favorite coffee brands across age groups. Most respondents fall into the

Super Premium segment (5.9K), followed by Premium (2.4K) and Mass Aspiring (2.4K), indicating that the survey sample is dominated by higher-income consumers.

The Self-Employed Company Owner group records the highest average spending (198K), with a clear peak at 293K in the 30–34 age range. Other Occupations (193K) and Small Business/Freelance workers (165K) follow, showing consistent mid-to-high spending. White-Collar employees (159K) and Blue-Collar workers (157K) contribute moderate levels, while Non-Working and Refuse groups remain the lowest. Overall, spending is strongest among entrepreneurial and professional groups aged 25–39, indicating that this segment holds the greatest purchasing power within the coffee market.

The chart shows brand preference across different age groups. Highlands Coffee stands out as the most favored brand among younger consumers, with 43.0% preference in the 16–19 group and 36.9% in the 20–24 group. This indicates strong appeal among Gen Z and young adults, likely due to the brand's modern image and accessibility. As age increases, preference for Highlands declines, while Street Coffee and Trung Nguyen gain traction. Trung Nguyen holds a notable share among older consumers aged 35+, peaking at 10.9% in the 45+ group, reflecting loyalty toward traditional Vietnamese coffee culture. Overall, brand preference shifts clearly with age — from modern chains like Highlands Coffee among younger consumers to traditional and local-focused brands among older generations.

Brand Image

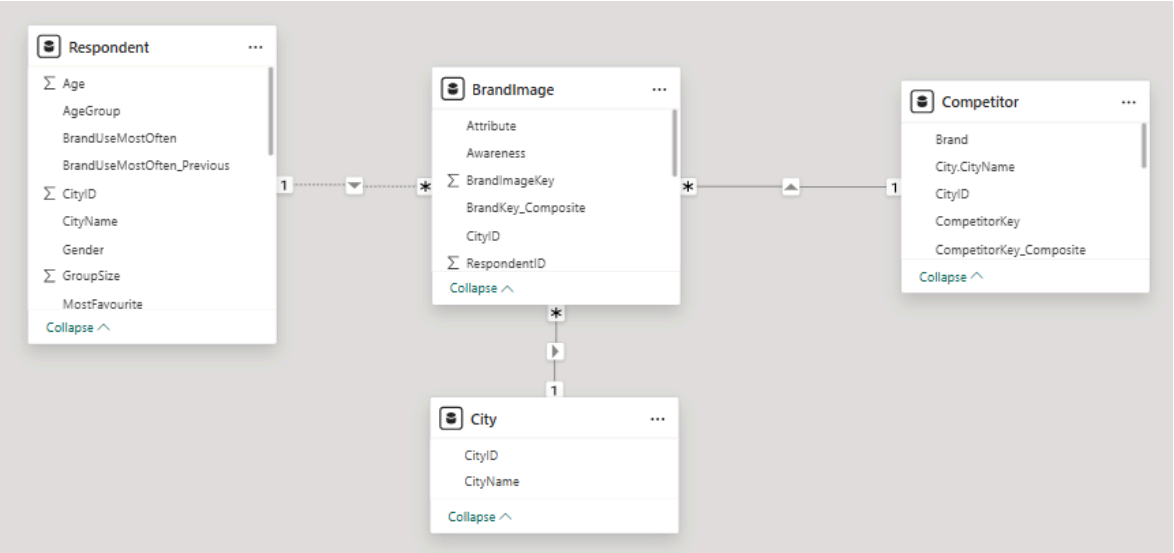


Figure 4.11. Data Model of Brand Image (Source: Authors)

The figure Data Model of Brand Image shows the relationship between the Fact Table Brand Image with its dimensional Table including Competitors, City and Respondent. The fact table BrandImage has many to one relationships with Respondent, City and Competitors.

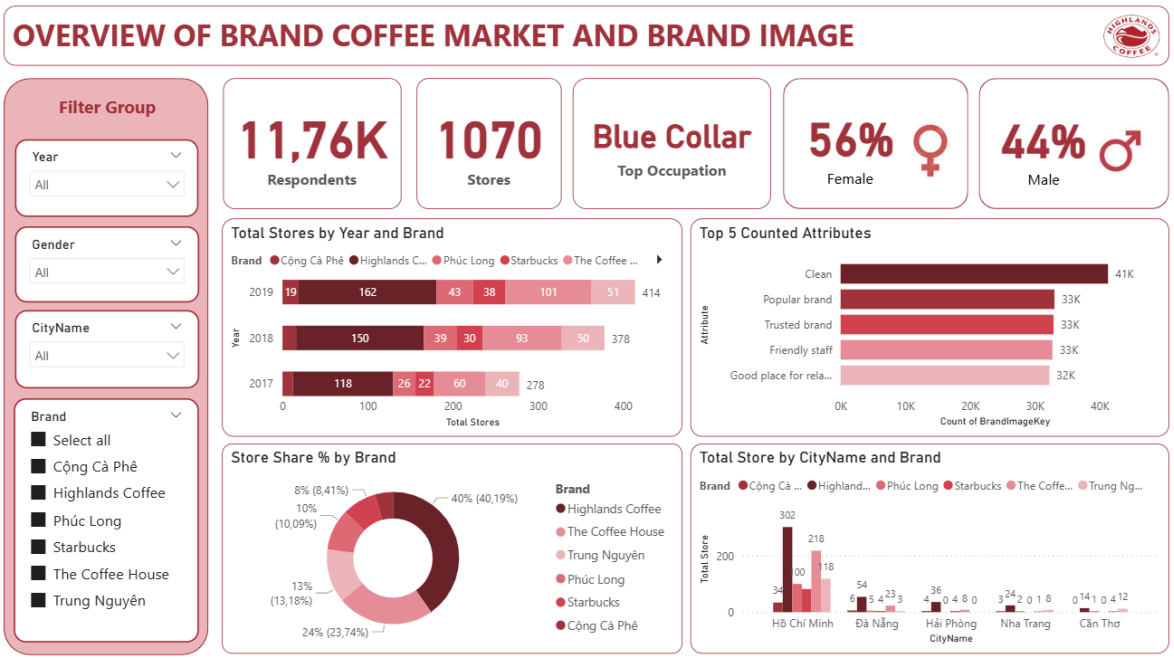


Figure 4.12. Overview of Coffee Store Market and Brand Image (Source: Authors)

The Overview of Coffee Store Market and Brand Image dashboard illustrates the overall market and brand perception of six major coffee brands in Vietnam, including Highlands Coffee, Phúc Long, Cộng Cà Phê, Starbucks, The Coffee House, and Trung Nguyên Coffee. It visualizes key factors: the total respondents, the total count number of stores, and the characteristics of respondents including gender and occupation. A total of 11,760 people participated in the survey, with 56% were female and 44% were male. Most of the respondents belong to the Blue Collar group.

The total number of coffee shops from these six brands has increased from 2017 to 2019 (278 stores in 2017 and 414 stores in 2019). Among them, Highlands Coffee has the highest number of stores opened over all three years, with 44 new stores (115 stores in 2017, 150 in 2018, and 162 in 2019). Meanwhile, Cộng Cà Phê had the fewest new openings with only 7 more stores (12 stores in 2017, 16 stores in 2018, and 19 stores in 2019). The Coffee House ranked second in terms of growth with 41 new stores (60 stores in 2017, 93 stores in 2018, and 101 stores in 2019). The remaining brands are Trung Nguyên with 11 stores, Starbucks with 16 stores, and Phúc Long with 17 stores. The year 2018 recorded the highest increase of store openings across all brands, showing the spurge of F&B market in Vietnam.

Among major brand attributes were mentioned, and the five most frequent ones are “Clean”, “Popular brand”, “Trusted brand”, “Friendly staff” and “Good place for relaxing”. This indicates that these brands are generally perceived as providing good service quality, having clean spaces, and being well-known and trusted by consumers.

In terms of market share, Highlands Coffee holds the largest proportion of stores with over 40%, followed by The Coffee House with 24%. Trung Nguyên, Phúc Long, and Starbucks rank at 3rd, 4th, and 5th with 13%, 10%, and 8% respectively. However, Cộng Cà Phê has the smallest share at around 4.38%. This shows that Highlands is currently the leading brand in terms of brand positioning.

Moreover, Hồ Chí Minh City has the highest number of coffee stores with a total of 772 stores from the six brands, followed by Đà Nẵng, Hải Phòng, Nha Trang and Cần Thơ. Highlands Coffee also has the largest number of stores in each city, and Hồ Chí Minh City is the city which has the most Highlands Coffee stores with 302 stores. This indicates that Hồ Chí Minh City is the main hub for coffee chains due to its high population density and consumption demand.

Based on the above insights, the following business objectives for Highlands Coffee are proposed. The first future implementation is the expansion in the Tier-2 cities. Hồ Chí Minh City has the highest number of coffee stores and the fiercest competition, indicating a saturated market. In contrast, cities such as Đà Nẵng, Nha Trang and Hải Phòng have fewer stores but they still show many chances to increase the expansion as competition is lower, rental costs are cheaper, and tourism demand is high. The second implementation is brand positioning. From the dashboard, consumers mainly associate the brands with "Clean", "Popular brand", "Trusted brand", "Friendly staff" and "Good place for relaxing". Therefore, Highlands should continue to maintain the attributes "trusted" and "clean" as they are highly valued by customers. However, current coffee brands show limited emphasis on "premium quality" or "unique taste", indicating a gap in differentiation, which Highlands can position itself as a premium or uniquely styled brand to attract customers seeking distinctive experiences.

NeedstateDayPart Overview

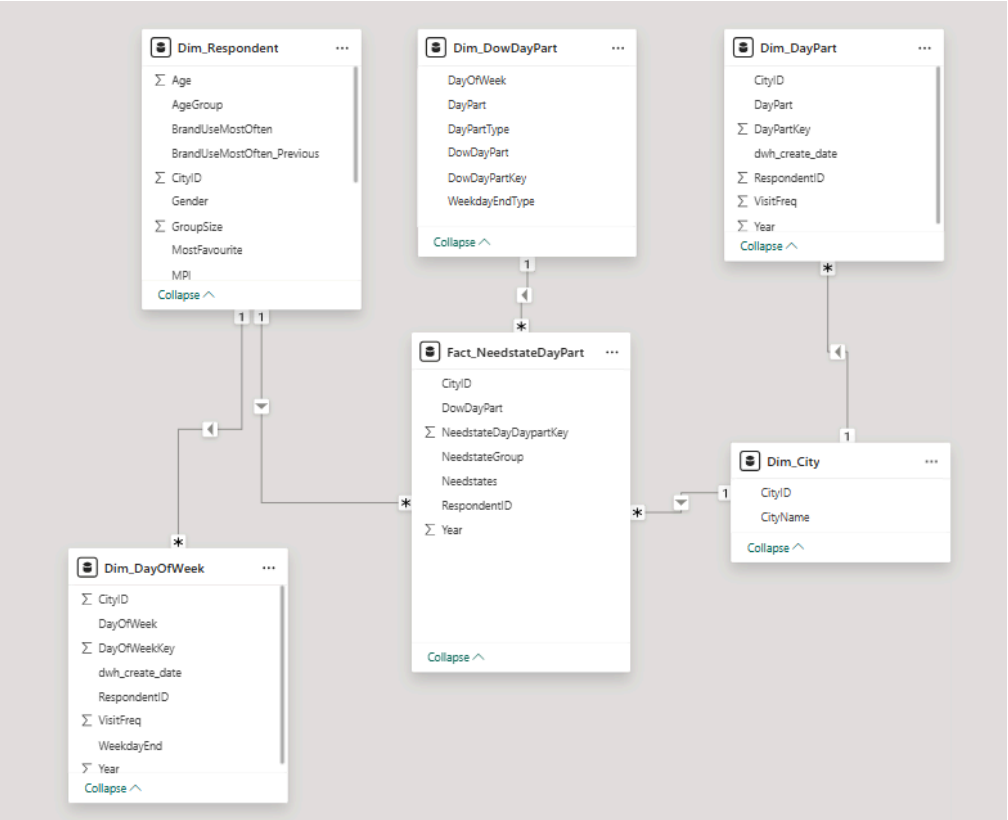


Figure 4.13. Data Model of NeedstateDayPart (Source: Authors)

The data model consists of one fact table, Fact_NeedstateDayPart, and five dimension tables: Dim_Respondent, Dim_DowDayPart, Dim_DayPart, Dim_DayOfWeek, and Dim_City. The relationships between these tables form a star schema, enabling analysis of respondent behavior and need states across various time periods, days, and cities.

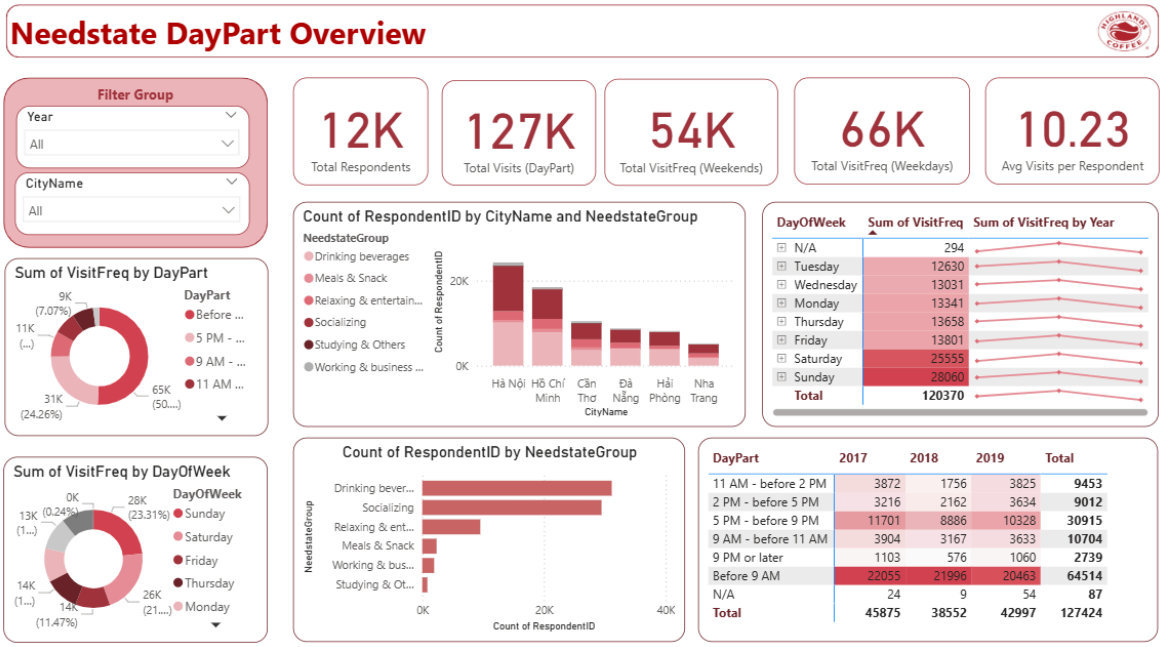


Figure 4.14. Needstate DayPart Overview Dashboard (Source: Authors)

The Needstate DayPart Overview Dashboard illustrates customer visit behavior at Highlands Coffee. Over *12K respondents* recorded approximately *127K visits by time of day,* including *66K visits on weekdays and 54K visits on weekends* - averaging *10.23 visits per person*, indicating strong engagement with the brand.

Most visits occurred *before 9 AM* (around 50%), highlighting a strong morning peak. This suggests Highlands Coffee is a preferred destination for early relaxation, social gatherings, or a light breakfast. The *5 PM - before 9 PM* period ranked second, indicating a notable demand for evening coffee as well.

By day of week, *Sundays (28%)* and *Saturdays (26%)* saw the highest traffic, aligning with leisure-time behavior, while most weekdays like Tuesday and Wednesday had fewer visits, mostly work or study-related. Visits increased from 2017 to 2018 but declined in 2019, mainly due to an overall drop in total visitors that year.

In terms of *Needstate Groups*, the main motivation for visits was *"Drinking Beverages"* followed closely by *"Socializing"* This demonstrates that customers view Highlands Coffee as a place for both a go-to place for their coffee fix and a space for social

connection. Other needs such as *"Studying & Working"* or *"Meals & Snack"* represent a smaller share, reinforcing the brand's image as a relaxed social space than a workspace.

Geographically, *Hanoi* and *Ho Chi Minh City* recorded the highest participation, underscoring the brand's strength in major urban markets, while *Can Tho, Da Nang,* and *Hai Phong* showed lower engagement but potential for growth.

Overall, Highlands Coffee stands out as a popular destination in the morning and on weekends, serving primarily social and beverage-related needs in Vietnam's key cities - insights that can guide time-based and location-driven marketing strategies to enhance customer experience and business performance.

Conclusion

These visualized dashboards provide the answer for the key business question *"How does Highlands Coffee sustain its leading position in Vietnam's competitive coffee chain market?"* by highlighting that Highland's success is driven by a combination of broad market coverage, strong brand awareness and deep customer engagement across different demographic groups.

The findings from BrandHealth reveal that Highland maintains one of the highest Net Promoter Score (8.5) among all competitors alongside with a significant store opening from 118 stores in 2017 to 162 stores in 2019, demonstrating its ability to strengthen customer satisfaction and royalty while scaling operations.

The BrandImage dashboard further confirms Highland's strong market presence when it holds more than 40% of total market share across six major brands and is recognized most frequently with positive brand attributes such as "Clean", "Trusted brand brand" and "Popular".

Moreover, insights from Customer Portrait Dashboard illustrates Highland's broad demographic reach, particularly among GenZ and young adults aged 16-24, who show the highest brand preference. This suggests that the brand's modern and accessible position resonates with the young segment while still maintaining appeal across higher-income groups. The Needstastes dashboard also demonstrates this by showing

strong engagement in morning and weekend visits, reflecting the Highlands's maintaining customer's perception strategy as a preferred destination for socialising and beverage enjoyment rather than a workplace.

Collectively, these visual insights emphasize that Highland Coffee's leading position is not dependent on the store volume, but rather on its multi-dimensional brand strategy by integrating reach and reputation and relevance. However, the findings also disclose the untapped opportunities for future growth in Tier-2 cities such as Đà Nẵng, Nha Trang, Hải Phòng, where the competitive levels are lower but the demand consumption is high from foreigner visitors. Additionally, Highland can further differentiate itself through product innovation and customized marketing to sustain long-term customer loyalty.

In conclusion, these dashboards bring a data-driven foundation for understanding Highland Coffee's market dominance and offer actionable business plans in the Vietnamese coffee market.

4.3. Visualization of experimental results

4.3.1. Clustering results visualization

Table 4.1. Characteristics of each cluster (Source: Authors)

Cluster ID	Size (%)	Avg Spending	Avg Age	AvgVisit Freq (Weekly)	Avg Group Size	Most Frequent Gender	Most Frequent Motivation	Most Frequent DayPart	Most Frequent Occupation	Characteristic
0	11.8	429,960 VND	34.60	4.30	3.36	Male	Drinking beverages	5 PM - before 9 PM	Officer - Staff level	High-value core customers, spending above average, visiting regularly.
1	54.1	84,729 VND	34.39	2.95	3.32	Female	Drinking beverages	5 PM - before 9 PM	Unskilled Labor (worker)	The majority segment, low value.
2	0.9	1,484,660 VND	33.49	4.96	3.40	Female	Drinking beverages	5 PM - before 9	Officer - Staff level	Super VIP customers, highest

								PM		spending, visiting quite regularly.
3	28.4	240,033 VND	34.43	3.87	3.35	Male	Drinking beverages	5 PM - before 9 PM	Unskilled Labor (worker	Spending at a medium-low level, visiting quite regularly.
4	4.9	752,463 VND	34.00	5.16	3.29	Male	Drinking beverages	5 PM - before 9 PM	Officer - Staff level	High-level spending, visiting frequently.

The clustering analysis (k=5) results show a quite clear distribution of customers through percentage and average values of the features. In terms of volume, over half of the customers are concentrated in Cluster 1 (54.1%) and Cluster 3 (28.4%), while high-value clusters (Clusters 0, 2, 4) account for a total of only about 17.6%.

Cluster 0: The stable core customer

Cluster 0 has stable spending characteristics with an average spending over the average (429,960 VND) and a quite regular visit frequency (4.30 times/week). This group accounts for 11.8% and is primarily Male in the Officer - staff level occupation group. This group is stable, spends well, but has not reached maximum engagement.

Cluster 2: The super VIP customer

Cluster 2 shows the characteristics of this customer group with the most outstanding average spending (1.48 million VND) and a high visit frequency (4.96 times/week). This cluster, although accounting for only 0.9% of the total customers, is the largest potential source of profit contribution. The demographics of this group are primarily female in the Officer - staff level occupation group.

Cluster 1: The low-value majority customer

Cluster 1 is the largest group (54.1%) but has the lowest average spending (84,729 VND) and low frequency (2.95 times/week). The demographics of this group are primarily female in the Unskilled labor (worker) occupation group.

Cluster 3: The revenue base customer

Cluster 3 maintains a quite regular visit frequency (3.87 times/week), but the average spending is only at a medium-low level (240,033 VND). This group accounts for 28.4% and primarily consists of males in the Unskilled labor (worker) occupation group.

Cluster 4: The loyal customer

Cluster 4's distinguishing characteristic is its highest visit frequency (5.16 times/week), coupled with high average spending (752,463 VND). This group accounts for 4.9% and is primarily male in the Officer - staff level occupation group. These are active customers who have a strong engagement with the brand.

Distribution of Spending by Cluster

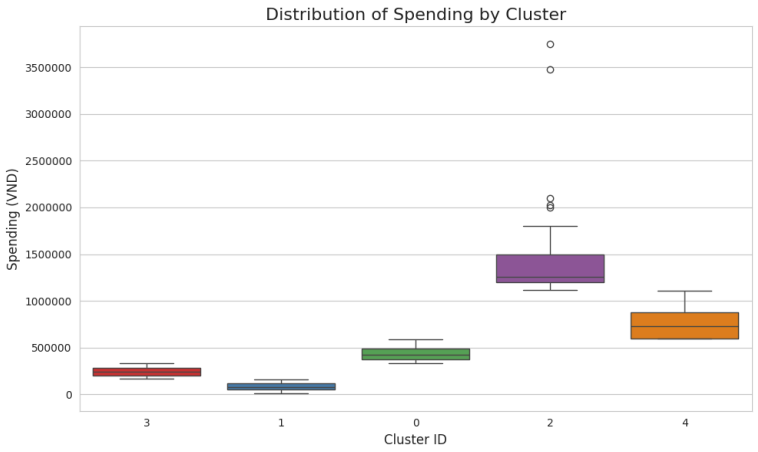


Figure 4.15. Distribution Of Spending By Cluster (Source: Authors)

Boxplot for Average Spending by cluster reveals an extremely significant disparity in actual spending levels (VND) among the segments. Cluster 2 stands out completely with the highest average spending (1,484,660 VND). The presence of outliers in this cluster confirms the existence of "Super Spenders" who far exceed even the already high average of this group, reinforcing its "Super VIP" status. Cluster 4 (752,463 VND) and cluster 0 (429,960 VND) continue to maintain high, stable spending levels, forming the premium customer base for the brand. Conversely, cluster 3 sits at a medium-low level (240,033

VND), but the biggest challenge lies with Cluster 1, which accounts for over 50% of customers yet has the lowest Average Spending (84,729 VND). This huge disparity underscores the imperative to address the profit challenge presented by the majority low-value segment.

Distribution of Age by Cluster

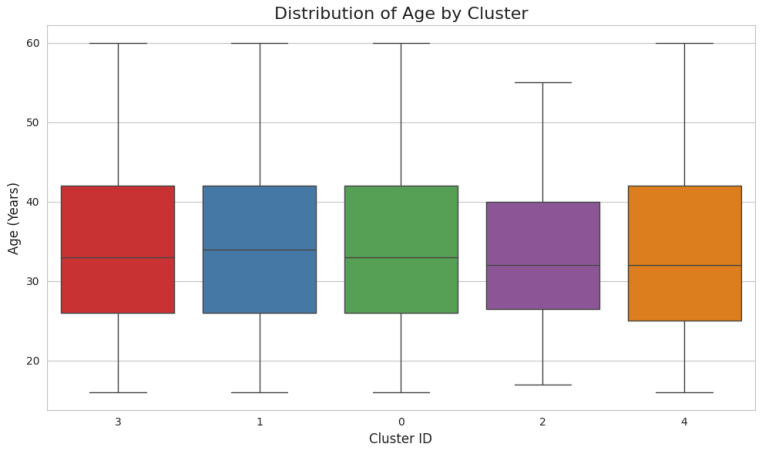


Figure 4.16. Distribution Of Age By Cluster (Source: Authors)

Boxplot of Age by cluster shows a high consistency in the median age across the entire customer base. The median value of all clusters is concentrated within the 33 to 35 age range, confirming that age is not a core differentiating factor in this clustering model. However, there are subtle differences: cluster 2 and cluster 4 tend to have a slightly lower median age compared to the remaining clusters. This implies that the most active and highest financially valuable customer segments tend to be slightly younger than the overall average.

Distribution of Occupation by Cluster

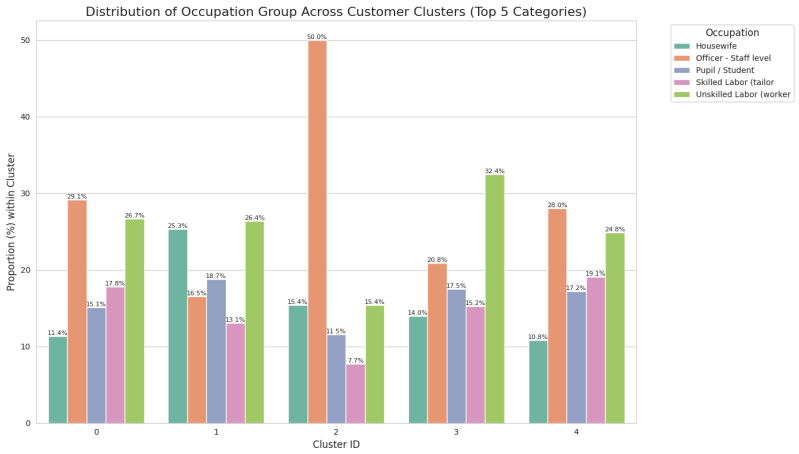


Figure 4.17. Distribution of occupation by cluster (Source: Authors)

The heatmap Distribution of Occupation group by cluster reveals a clear polarization in demographic profiles, confirming this is the strongest clustering differentiator in terms of financial value. The chart confirms that the highest-value clusters (cluster 0, cluster 2, and cluster 4) are defined by the Officer-staff level group, affirming that this is the core source of the brand's financial and frequency contribution. Conversely, the two largest and lowest-value clusters (cluster 1 and cluster 3) are heavily concentrated in the Unskilled labor (worker) group. This division is visual and sharp, indicating that Highlands Coffee serves two distinct primary target groups with completely different spending capacities and needs.

Distribution of Visit frequency by Cluster

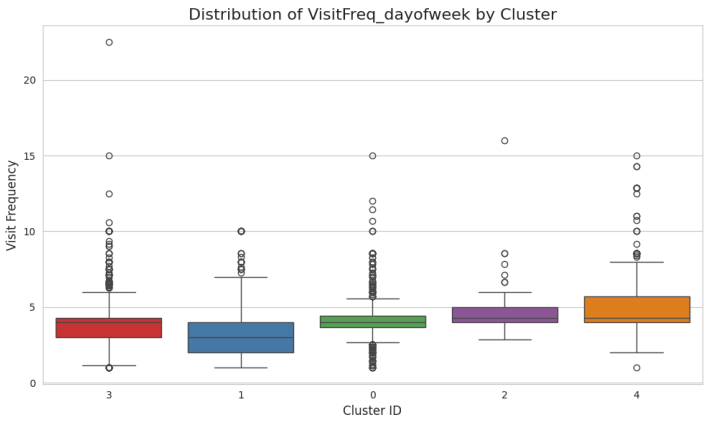


Figure 4.18. Distribution of visit frequency by cluster (Source: Authors)

Boxplot of Visit frequency confirms that engagement levels are clearly stratified. High-value clusters (cluster 4 and cluster 2) exhibit the highest level of engagement, with cluster 4 reaching the highest frequency (5.16 times/week) and cluster 2 maintaining a very high frequency (4.96 times/week). The presence of outliers in the upper portion of the Boxplot for these clusters, particularly cluster 4, highlights the existence of "Super Users" whose visit frequency far exceeds the already high average of their segment. Conversely, visit frequency decreases proportionally with customer value: cluster 0 maintains an above-average frequency (4.30 times/week), cluster 3 is at a medium-low level (3.87 times/week), and cluster 1 (the largest, lowest-value group) has the lowest frequency (2.95 times/week). This clear opposition confirms that visit frequency is a key factor in defining and separating customer profiles, linking tightly with their potential revenue contribution.

Distribution of Needstate group (Motivation) by Cluster

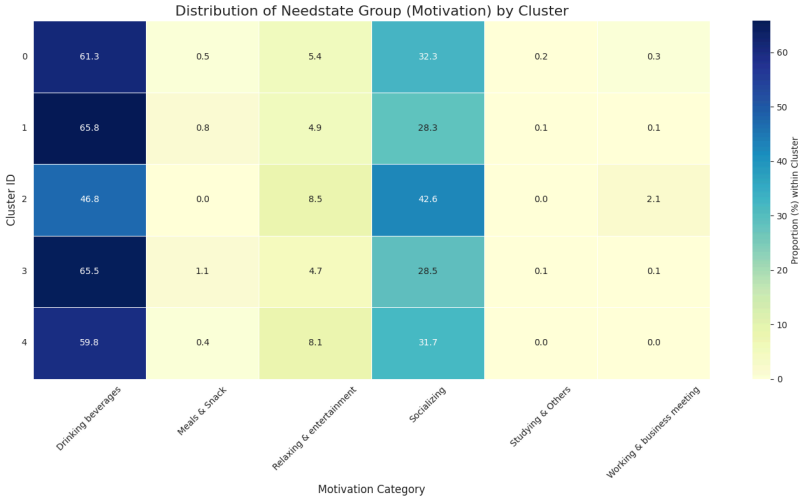


Figure 4.19. Distribution of needstate by cluster (Source: Authors)

Distribution of the Needstate group (Motivation/Need) by cluster reveals a high consistency in the primary purpose of visiting but a subtle difference in secondary needs. The main motivation for the entire customer base, regardless of the cluster, is "Drinking

beverages." This confirms that coffee is a basic consumption product and not an effective clustering differentiator. However, the difference emerges in secondary needs: high-value clusters (clusters 0, 2, 4) show a higher proportion of secondary motivations like "Socializing" or "Working/Business," reinforcing their profile as the Officer-level group utilizing the coffee shop as a work or meeting space. Conversely, low-value clusters (clusters 1, 3) tend to focus more on basic needs. This distinction confirms that high-value clusters do not only purchase beverages but also the accompanying space and services, while the majority segment primarily fulfills only basic product needs.

Distribution of Daypart by Cluster

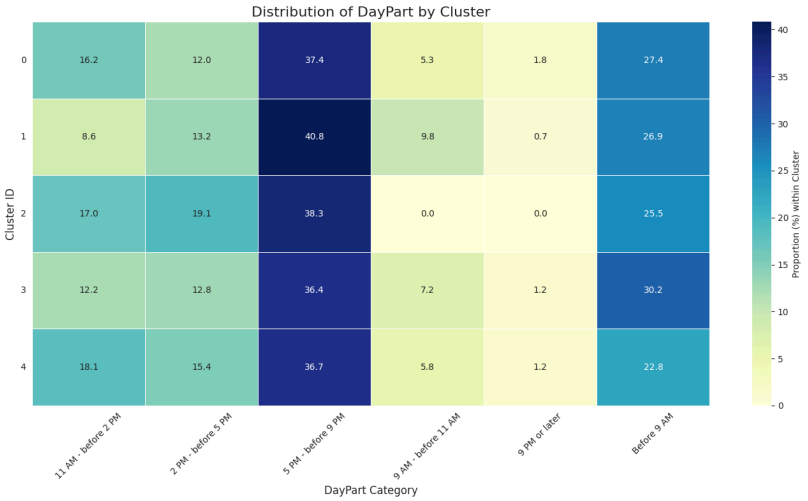


Figure 4.20. Distribution of daypart by cluster (Source: Authors)

Distribution of Daypart by cluster shows a high consistency in the primary visiting time but a clear difference in secondary habits. "5 PM - before 9 PM" (Evening) is the most popular visiting time for all five customer clusters, confirming this is a common peak hour linked to the needs for socializing or relaxation after work for the majority of Highlands Coffee customers. However, the differentiation occurs in secondary habits: the low-value clusters (Cluster 1 and Cluster 3) show a significantly higher proportion of visits occurring "Before 9 AM" (Early Morning) compared to the high-value clusters (clusters 0, 2, 4). This disparity indicates that Cluster 1 and Cluster 3 tend to use the brand for convenience and quick transactions, often during their commute. Conversely,

the high-value, Officer-level clusters (clusters 0, 2, 4) primarily use Highlands as a destination for socializing or relaxation in the evening. This highlights that the operating strategy must clearly distinguish between the high-speed service experience required in the morning and the leisure/relaxation experience offered in the evening.

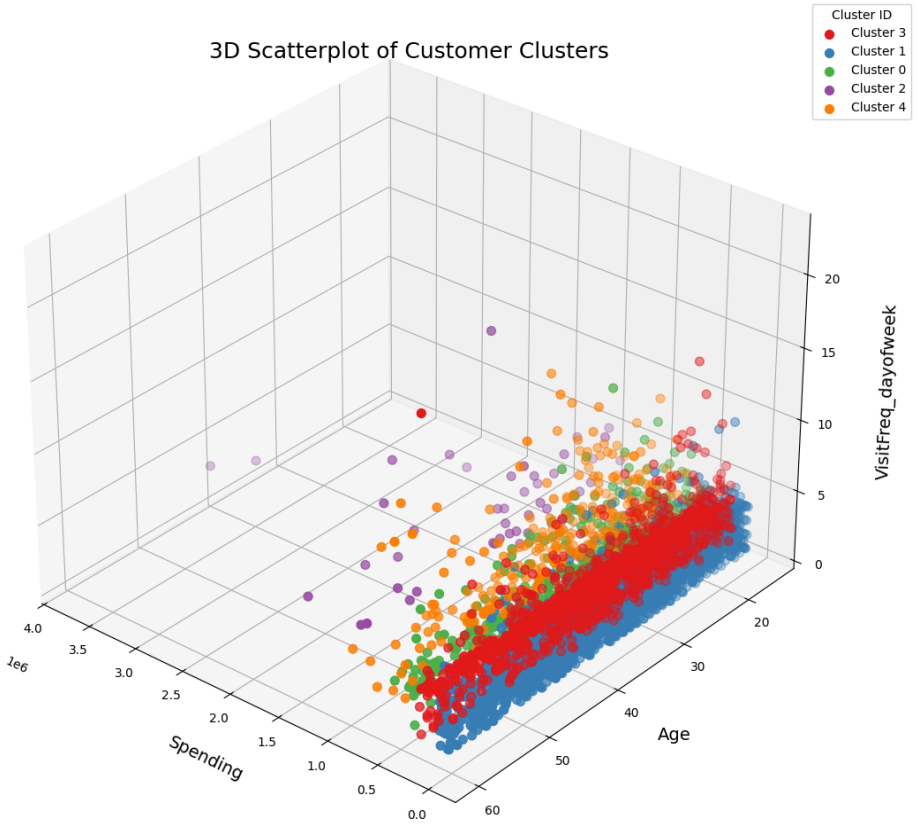


Figure 4.21. Scatter Plot for customer clusters (Source: Authors)

The 3D Scatter Plot illustrates the separation of customer clusters in a three-dimensional space defined by Spending (X-axis), Age (Y-axis), and Visit frequency (Z-axis).

This chart demonstrates that the cluster separation is primarily shaped by the Spending axis and the Visit frequency axis, confirming these are the two core factors in defining customer value and behavior. Specifically, cluster 2 and cluster 4 are located in the region of high Spending and high Frequency, representing the group with the greatest value and engagement. Conversely, cluster 1 forms a large, dense cloud concentrated in the area of low spending and low frequency.

Most importantly, the Age axis shows high consistency, as the clusters are not significantly dispersed along the Y-axis. This confirms that, although the model has successfully segmented the clusters based on spending capacity and frequency, the age factor is not the primary driver in differentiating the various customer profiles. The 3D chart serves as the final piece of evidence confirming the model's effectiveness in classifying customers based on financial behavior.

Main features of each cluster

Table 4.2. Main features of each cluster (Source: Authors)

Cluster	Cluster name	Key characteristics
0	Stable core customer	Accounts for 11.8%, representing a solid revenue base, with high average spending (429,960 VND) and stable frequency (4.30 times/week).
1	Low-value majority customer	Accounts for 54.1%, being the largest segment but simultaneously having the lowest average spending (84,729 VND) and lowest frequency (2.95 times/week).
2	Super VIP customer	The smallest size (0.9%) but the highest average spending (1,484,660 VND), with a high visit frequency (4.96 times/week).
3	Revenue base customer	Accounts for 28.4%, serving as a large, stable source of customer traffic, with a fairly regular visit frequency (3.87 times/week) but medium-low average spending (240,033 VND).
4	Loyal customer	Accounts for 4.9%, showing the highest level of engagement, with the highest visit frequency (5.16

		times/week), coupled with very high Average Spending (752,463 VND).

Strategy proposal

Cluster 0, the stable core customer (Avg Spending: 429,960 VND), is a solid revenue foundation with regular spending and frequency but needs to be motivated to reach the level of Cluster 4. Firstly, personalized upsell and cross-sell strategy is the key solution, as Verhoef et al. (2009) confirms that suggesting complementary products or upgrading orders through smart recommendations will significantly increase transaction value. Concurrently, Periodic Promotions and Personalized Service are also very important; according to Parasuraman et al. (1988), providing personalized support (e.g., shopping consultation, birthday offers) will help satisfied customers maintain and reinforce loyalty. Finally, integrating them into the Loyalty Program is also an approach suggested by Kotler & Keller (2016), helping to create a sense of "priority" and encouraging this group to increase their spending to achieve higher reward tiers.

Cluster 1, the low-value majority customer (avg spending: 84,729 VND), is the largest group (54.1%) but has the lowest value, representing the biggest profit challenge. Firstly, assessing potential and selective activation is the key solution, as Rust & Huang (2012) proposes evaluating this group's potential and using only low-cost activation campaigns (e.g., email marketing) for the most likely responders. Concurrently, cost management and digital channel enhancement are also very important; shifting transactions to the app or self-service can help reduce operating costs per low-value transaction, thereby optimizing profitability. Finally, product consultation and special discounts are reasonable approaches; Kotler & Keller (2016) acknowledges that attractive short-term offers can draw low-frequency buyers back, especially when they have not encountered quality issues.

Cluster 2, though accounting for only 0.9% of total customers, is the most critical source of revenue and needs dedicated special attention to ensure absolute loyalty and maintain high spending. Firstly, exclusive care and privileges are the key solutions, because

according to Kotler & Keller (2016), the highest-value customer segment should be offered VIP services, such as priority support services, separate new product experience events, or premium personalized gifts to meet maximum expectations. Concurrently, enhancing service quality is also very important; Parasuraman et al. (1988) emphasize that improving service quality and providing personalized support channels will meet this segment's high expectations, strengthening loyalty and reducing dissatisfaction. Finally, introducing new premium products and upsell is an effective way to leverage this group's high purchasing power, as Verhoef et al. (2009) shows that upsell and cross-sell help maximize revenue and maintain purchase interest among VIP customers.

Cluster 3, the revenue base customer (avg spending: 240,033 VND), provides good frequency but lower average spending, requiring an optimization strategy for order value. Firstly, optimizing average order value (AOV) is the key solution, through implementing the bundling/combo promotion strategy, as Rust & Huang (2012) argues that offering attractive product bundles not only increases spending but also boosts marketing effectiveness per transaction. Furthermore, simplifying and enhancing operational experience is also very important; due to their Unskilled labor profile, prioritizing fast, convenient service experience is essential to ensure the purchasing process does not cause delay, thus preventing dissatisfaction leading to defection (Parasuraman et al., 1988). Finally, reactivation and building trust is a reasonable approach to maintain a long-term relationship, helping to foster engagement and encouraging this group to increase their spending in the future.

Cluster 4 exhibits the highest engagement and has very strong average spending, making it an ideal customer group for exploiting purchase frequency. Firstly, leveraging the power of Loyalty is a key solution, through implementing Loyalty Programs featuring Frequency-based rewards to encourage them to maintain high purchasing volume, as Kotler & Keller (2016) suggests. Furthermore, Utilizing Engagement Levels is also very important; Verhoef et al. (2009) points out that turning them into Brand Advocates by offering "refer-a-friend" incentives or allowing them to experience new products before the general public is an effective strategy to maximize CLV. Finally, developing

community and private events for this group is an approach that Parasuraman et al. (1988) suggests to strengthen emotional bonds and maintain long-term attachment.

4.3.2. Predictive model visualization

Table 4.3. Feature importance on predict customer churn (Source: Authors)

Feature	Importance
Attribute	18.518656
Gender	5.733407
OccupationGroup	2.957462
CompanionGroup	2.601267
DayOfWeek_mode	2.351748
DayPart_mode	2.343934
VisitFreq_mean	1.786088
Spending	1.745259
MPI	1.516034
GroupSize	1.489848
Age	1.451023

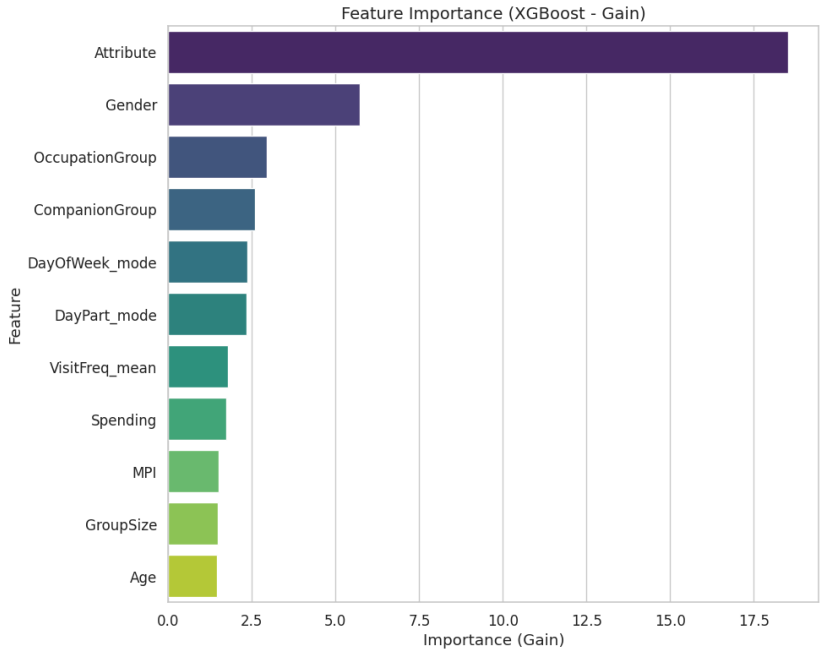


Figure 4.22. Feature importance on predict customer churn (Source: Authors)

The "Feature Importance (XGBoost - Gain)" chart illustrates the relative influence of various factors on the prediction of customer churn at Highlands Coffee. Importance is quantified by Gain, a metric representing the improvement in model performance provided by each feature to the XGBoost algorithm.

The analysis reveals that three features overwhelmingly drive the prediction: Attribute emerges as the most critical factor, with a Gain of nearly 18, suggesting that the customer's core perception of the brand is the primary mechanism of churn. These Attributes include factors like Ambiance & Environment, Service & Operations, and Product Quality & Variety. Following this are Gender (Gain ≈5.5) and OccupationGroup (Gain ≈3.2), which form the second tier of importance and affirm the significant role of core demographic backgrounds. A clear distinction is seen with secondary factors: Behavioral and Time-based features such as CompanionGroup, DayOfWeek_mode, and DayPart_mode show moderate importance (Gain 2.0−3.0), indicating that the *timing* and *social context* of customer interaction hold more relevance than basic metrics. Conversely, variables typically considered vital in marketing, such as Spending and

especially Age, register very low impact (Gain below 1.0), demonstrating that this model is primarily steered by customer Attribute perceptions and Gender, rather than conventional age or expenditure levels.

4.4. Business insights and strategy

4.4.1 Business insights

The comprehensive data analysis shows that Highlands Coffee currently holds a leading position in the Vietnamese coffee chain market, boasting high brand recognition (8.5 NPS) and the largest store network (over 160 branches). However, the chain still faces a major challenge in low customer retention rates: of the 8,500 people aware of the brand, only 2,700 used it in the most recent month. This reflects a critical issue in converting "awareness" into "loyalty," demanding a strategy focused on reinforcing customer experience and retention.

The core customer base of Highlands is primarily composed of females and white-collar employees aged 20–34, who typically visit in the evening and on weekends for relaxation, socializing, or work. In contrast, males and self-employed groups constitute a higher proportion of the churn group, often visiting in the morning for convenience. This indicates that Highlands needs to upgrade the morning exploratory, increasing service speed and expanding "grab & go" or "self-order kiosk" loyalty, to retain these high-frequency but low-loyalty customers.

In terms of brand image, Highlands is recognized for being "clean, friendly, and trustworthy," but is weak in the area of "beverage quality and product differentiation." This is a core vulnerability as competitors like The Coffee House are closing the gap through menu innovation and quality. Therefore, the brand must invest in product R&D, for example, developing signature lines (Highlands Signature), refining seasonal or regional menus, and creating a healthy drink line to expand its appeal to younger, lifestyle-conscious customer segments.

Customer clustering analysis shows that the highest-value groups (Super VIP and Loyal) only account for less than 6% but contribute the majority of revenue. Therefore, Highlands must build separate retention strategies for each segment:

- **VIP Customers:** Develop exclusive care programs (birthday privileges, new product trials, "Highlands Lounge" access) to reinforce loyalty.

- **Loyal Customers:** Implement a frequency-based loyalty program (point accrual/redemption), encouraging combo purchases or size upgrades to increase Average Order Value (AOV).

- **The Low-value Majority Customer:** Apply low-cost activation strategies like app discounts, automated upsells, or off-peak promotions to boost transaction volume.

Brand Health results and market share data suggest that Ho Chi Minh City has reached a saturation point, while secondary cities like Da Nang, Hai Phong, and Nha Trang still hold development potential. Highlands should therefore pivot its expansion toward these Tier 2 cities, where rental costs are lower, tourism is growing, and competition is less intense. This will be a strategic move to optimize costs, diversify the market, and sustain stable growth.

XGBoost model suggests that the most critical factor determining churn is "brand perception/attribute" rather than "age" or "expenditure." Consequently, Highlands needs to invest heavily in emotional experience and service consistency—from staff attitude, speed of service, music, lighting, and aroma—to reinforce its brand image as "friendly, modern, and memorable." Concurrently, it must leverage behavioral data from the Loyalty app to personalize drink recommendations and promotional programs

4.4.2 Strategy

In summary, Highlands Coffee needs to transition from a model of "rapid scale expansion" to "sustainable growth driven by customer value," with three core strategic directions as below

The first direction is Highlands Coffee must focus on redesigning the morning experience to better capitalize on the high-frequency, low-loyalty customer segment. Currently, most customers visiting before 9 AM are males and self-employed workers who tend to buy quickly and leave quickly. However, the brand does not yet have a specialized service system for this group, leading to a high risk of defection. To address this, Highlands can implement a "Grab & Go" model and quick service counters, integrating pre-ordering via the mobile application to reduce wait times. Additionally, the brand should expand its drive-thru service to other branches nationwide so customers can purchase conveniently without parking, while training specialized morning shift staff to optimize performance and service attitude. This will not only help Highlands expand revenue during morning hours but also convert "casual" customers into "regular" users, contributing to increased usage frequency and brand loyalty.

The second one, Highlands Coffee needs to accelerate product innovation and build a unique identity to enhance customer perceived value. In a context where coffee chains in Vietnam are increasingly uniform in terms of ambiance, creating a distinctive product identity is a vital factor. Highlands can develop a "Highlands Signature" beverage line that embodies modern Vietnamese flavors, combining local ingredients with exclusive recipes, and simultaneously implement seasonal or regional menus to evoke curiosity and sense of occasion. Furthermore, the brand should invest in upgrading ingredient quality, standardizing brewing procedures, and improving packaging aesthetics to create a premium, consistent feel. By transforming coffee into a part of the lifestyle—rather than just a drink—Highlands can create a competitive emotional advantage, enhancing recognition and long-term loyalty.

The third is Highlands Coffee needs to build a personalized strategy and reward loyal customers to increase Customer Lifetime Value (CLV) and turn them into brand ambassadors. Although a point accumulation program exists, Highlands has not yet made this program widely accessible to customers. Highlands should promote this program more extensively, implement a multi-tiered membership program (Silver – Gold – Diamond) with privileges such as point accumulation, birthday offers, or new product

trials. Additionally, leveraging behavioral data from the Loyalty app allows the brand to suggest beverages based on personal preferences or propose suitable promotions based on usage patterns, creating a feeling of being valued and understood. Concurrently, Highlands should develop a program about getting rewarded when referring a friend and host community events for close members to strengthen emotional attachment. When customers feel they are part of the brand, they not only purchase more frequently but also willingly share and spread a positive image of Highlands to the community.

Conclusion and Future Works

Conclusion

Through the implementation of the research project "Application of Retail Analytics for Strategic Promotion in the Vietnamese Coffee Market: The Case of Highlands Coffee", this study has achieved several important outcomes.

First, the research established a comprehensive analytical framework integrating Exploratory Data Analysis (EDA), K-prototypes clustering, and XGBoost churn prediction to analyze customer behavior and identify key factors influencing loyalty and defection. This framework enabled the segmentation of Highlands Coffee's customers into distinct behavioral and demographic groups, revealing clear differences between high-value (High-CLV) and high-churn customers in terms of age, visiting time, and consumption motivations.

Second, the results demonstrated the strategic potential of retail analytics in guiding data-driven decision-making for the F&B industry. Specifically, the study proposed three actionable strategic directions for Highlands Coffee: (1) optimizing the morning experience for high-frequency, low-loyalty customers; (2) accelerating product innovation and differentiation to enhance perceived brand value; and (3) expanding personalized loyalty programs to strengthen retention and increase Customer Lifetime Value (CLV).

However, several limitations should be acknowledged:

First, the study did not perform comparative testing across multiple clustering and classification algorithms, which could have strengthened the validity and robustness of the segmentation and churn models.

Second, the predictive models—particularly those using machine learning classifiers—did not reach high accuracy levels, indicating the need for improved data preprocessing, feature engineering, and model optimization.

Third, the dataset lacked real-time variables, such as recent purchase recency or digital interaction metrics, which limited the ability to capture dynamic behavioral shifts and evolving customer preferences.

Overall, this study demonstrates that applying retail analytics provides a powerful foundation for evidence-based marketing and strategic promotion. The case of Highlands Coffee illustrates how transforming behavioral data into actionable insights can enhance customer understanding, strengthen loyalty strategies, and foster sustainable, customer-centric growth within Vietnam's evolving coffee chain market.

Future works

Building upon the limitations identified, several promising directions are suggested for future research.

First, future studies should conduct comparative analyses across various clustering and classification algorithms—such as DBSCAN, Gaussian Mixture Models, and advanced ensemble classifiers—to enhance segmentation robustness and model reliability.

Second, researchers could explore hybrid analytical frameworks that integrate deep learning architectures with traditional machine learning models. Techniques such as autoencoder-based feature extraction or neural clustering may uncover hidden behavioral dimensions, offering a more nuanced representation of customer diversity and purchasing dynamics.

Third, incorporating real-time data from multiple touchpoints—including POS transactions, loyalty applications, and mobile ordering platforms—would allow continuous monitoring of behavioral changes and the development of adaptive churn prediction systems.

Finally, integrating qualitative analytics such as sentiment analysis and text mining from customer feedback, online reviews, and social media discussions would complement quantitative models. This multimodal approach would provide deeper insight into emotional and experiential factors influencing loyalty, ultimately enabling brands like Highlands Coffee to design more human-centered and responsive marketing strategies.

In summary, future research should aim to construct a more dynamic and intelligent retail analytics ecosystem—where predictive, behavioral, and emotional data converge—to support sustainable customer engagement and strategic growth in the Vietnamese F&B industry

References

[1] *Amazon Web Services, Inc. (n.d.). What is ETL? - Extract Transform Load Explained - AWS*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/etl/?nc1=h_ls

[2] Informatica. (n.d.). *What is ETL?* https://www.informatica.com/resources/articles/what-is-etl.html

[3] GeeksforGeeks. (2025c, July 19). *ETL process in data warehouse*. GeeksforGeeks. https://www.geeksforgeeks.org/dbms/etl-process-in-data-warehouse/

[4] *SAP. (n.d.). What is a data warehouse? | Definition, components, architecture | SAP*. (n.d.). SAP. https://www.sap.com/products/data-cloud/datasphere/what-is-a-data-warehouse.html

[5] GeeksforGeeks. (2025d, August 1). *Data warehousing*. GeeksforGeeks. https://www.geeksforgeeks.org/dbms/data-warehousing/

[6] Lucie, & Lucie. (2022, June 20). *Data Warehouse là gì? Tổng quan về kho dữ liệu*. TopDev. https://topdev.vn/blog/data-warehouse-la-gi-tong-quan-ve-kho-du-lieu/

[7] *NVIDIA. (n.d.). What is K Means?* (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/k-means/

[8] GeeksforGeeks. (2025e, August 22). *K means Clustering – Introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/

[9] *Databricks. (n.d.). What is a Medallion Architecture? https://www.databricks.com/glossary/medallion-architecture*

[10] Mssaperla. (n.d.). *What is the medallion lakehouse architecture? - Azure Databricks*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion

[11] Idera. (2025, April 10). *How Medallion Architecture's layers transform data workflows?* ER/Studio. https://erstudio.com/blog/understanding-the-three-layers-of-medallion-architecture/

[12] Banerjee, S. (2020, June 16). *Food and Beverage Service 101: The basics, types, and roles explained*. Food and Beverage Knowledge. https://foodandbeverageknowledge.com/food-and-beverage-service-101-basics-types-and-role-explained/

[13] *Food and beverage services - quick guide*. (n.d.). https://www.tutorialspoint.com/food_and_beverage_services/food_and_beverage_services_quick_guide.htm

[14] Standard Insights. (2025b, June 15). The Vietnam Coffee Industry | Standard Insights. Standard Insights. https://standard-insights.com/insights/the-coffee-industry-in-vietnam/

[15] Company, B. (2025, August 4). *Technology adoption in Vietnam's F&B SMEs: Current gaps, challenges, and strategic solutions - B-Company*. B-Company. https://b-company.jp/technology-adoption-in-vietnams-fb-smes-current-gaps-challenges-and-strategic-solutions/

[16] Yang, P. Y., Chaw, J. K., Cheng, X., Ang, M. C., & Salim, M. H. M. (2023). *Exploring consumer behaviour patterns and modelling churn prediction in the food delivery service industry: A case study*. Journal of Information System and Technology Management, 8(32), 53–68. https://www.researchgate.net/publication/374614083

[17] AlShourbaji, I., Helian, N., Sun, Y., Hussien, A. G., Abualigah, L., & Elnaim, B. (2023). *An efficient churn prediction model using gradient boosting machine and metaheuristic optimization*. Scientific Reports, 13(14441). https://doi.org/10.1038/s41598-023-41093-6

[18] Putri, Y., Aldo, D., & Ilham, W. (2024). *Retail marketing strategy optimization: Customer segmentation with artificial intelligence integration and K-Means clustering*. Sinkron: Jurnal dan Penelitian Teknik Informatika, 8(4). https://doi.org/10.33395/sinkron.v8i4.14000

[19] Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, *21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

[20] Clarke, A. H., Freytag, P. V., & Mora Cortez, R. (2024). *Revisiting the strategic role of market segmentation: Five themes for future research*. *Industrial Marketing Management, 121*, A7–A10. https://doi.org/10.1016/j.indmarman.2024.07.012

[21] Avcı, D. A., Şahin, G., & Kan, M. (2024). Churn detection and user classification via machine learning in the food and beverage sector. *The European Journal of Research and Development*, *4*(4), 1–16. https://doi.org/10.56038/ejrnd.v4i4.552

[22] Gautam, N., & Kumar, N. (2022). Customer segmentation using k-means clustering for developing sustainable marketing strategies. *Business Informatics*, *16*(1), 72–82. https://doi.org/10.17323/2587-814x.2022.1.72.82

[23] Juhasz, J. (2025). Machine Learning-Driven Customer Segmentation: A Behavior-Based Approach for F&B Providers. *SEA - Practical Application of Science*, *XIII*(39), 169–176. https://doi.org/10.70147/s39169176

[24] Brahmana, R. W. S., Mohammed, F. A., & Chairuang, K. (2020). Customer segmentation based on RFM model using K-Means, K-Medoids, and DBSCAN methods. *Lontar Komputer Jurnal Ilmiah Teknologi Informasi*, *11*(1), 32. https://doi.org/10.24843/lkjiti.2020.v11.i01.p04

[25] Ramkumar, G., Bhuvaneswari, J., Venugopal, S., Kumar, S., Ramasamy, C. K., & Karthick, R. (2025). Enhancing customer segmentation: RFM analysis and K-Means clustering implementation. In *CRC Press eBooks* (pp. 70–76). https://doi.org/10.1201/9781003559139-9

[26] Rumi, Maulana & Rakib, Muhammad & Ashdaq, Muhammad. (2025). Customer Segmentation Using the K-Means Algorithm for Marketing Strategy Design: Case Study at the Icon Yasika Makassar. International Journal of Innovative Science and Research Technology. 10. 1041-1047. 10.38124/ijisrt/25jul508.

[27] AlShamsi, A. Y. (2022). *Understanding customer behaviour in restaurants based on data mining prediction technique*. RIT Digital Institutional Repository. https://repository.rit.edu/theses/11210

[28] Czarniecka-Skubina, E., Korzeniowska-Ginter, R., Pielak, M., Sałek, P., Owczarek, T., & Kozak, A. (2021). *Consumer choices and habits related to coffee consumption by Poles*. International Journal of Environmental Research and Public Health, 18(8), 3948. https://doi.org/10.3390/ijerph18083948

[29] Phan, N. T. T. (2020). *A market expansion research for F&B distribution agents in Ho Chi Minh City, Vietnam: Importing Valio Oddly Good® through brand image*. Karelia University of Applied Sciences. https://www.theseus.fi/handle/10024/354618

[30] Yang, P. Y., Chaw, J. K., Cheng, X., Ang, M. C., & Salim, M. H. M. (2023). *Exploring consumer behaviour patterns and modelling churn prediction in the food delivery service industry: A case study*. Journal of Information System and Technology Management, 8(32), 53–68. https://www.researchgate.net/publication/374614083

[31] Zhao, Z., Chen, G., Duan, J., & Xu, Y. (2025). *Site selection analysis and prediction of new retail stores from an urban commercial space perspective: A case study of Luckin Coffee and Starbucks in Shanghai*. ISPRS International Journal of Geo-Information, 14(6), 217. https://doi.org/10.3390/ijgi14060217

[32] Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson Education. https://www.scirp.org/reference/referencespapers?referenceid=3155681

[33] Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). *SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality.* Journal of Retailing.
https://www.researchgate.net/publication/200827786_SERVQUAL_A_Multiple-item_Scale_for_Measuring_Consumer_Perceptions_of_Service_Quality

[34] Risselada, R., Verhoef, P. C., & Bijmolt, T. H. A. (2010). *Dynamic Effects of Social Influence and Direct Marketing on the Adoption of High-Technology Products.* Journal of Marketing. https://journals.sagepub.com/doi/abs/10.1509/jm.11.0592

[35] Rust, R. T., & Huang, M. H. (2012). *Optimizing service productivity*. Journal of Marketing.
https://www.researchgate.net/publication/261970843_Optimizing_Service_Productivity

[36] Verhoef, P. C., Reinartz, W. J., & Krafft, M. (2009). *Customer Engagement as a New Perspective in Customer Management.* Journal of Retailing. https://www.researchgate.net/publication/229821476_Customer_Engagement_as_a_New_Perspective_in_Customer_Management

[37] Z. Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values (1998). Data Mining and Knowledge Discovery. 2(3): 283–304.

[38] Aprilliant, A. (2025, March 5). *The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)*. Towards Data Science. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb/

[39] *RPubs - Phân cụm dữ liệu hỗn hợp bằng K-Prototypes*. (n.d.). https://rpubs.com/namvk/895126

[40] GeeksforGeeks. (2025f, September 1). *Random Forest algorithm in machine learning*.GeeksforGeeks.https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[41] GeeksforGeeks. (2025g, September 5). *XGBoost*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/xgboost/

[42] Tyagi, A. (2025, April 23). *What is XGBoost Algorithm?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[43] Swastik. (2025, April 24). *SMOTE for Imbalanced Classification with Python*. Analytics Vidhya.

https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[44] *Logistic regression: Calculating a probability with the sigmoid function*. (n.d.). Google for Developers. https://developers.google.com/machine-learning/crash-course/logistic-regression/sigmoid-function

[45] Kavlakoglu, E. (2025, October 16). Support Vector machine. *IBM*. https://www.ibm.com/think/topics/support-vector-machine

Appendix

Appendix 1. The source code for this project

https://github.com/hoaloken61998/Highland-Coffee-Analytics.git