Application of Retail Analytics for Strategic Promotion in the Vietnamese Coffee Market

THE CASE OF HIGHLANDS COFFEE

Group 02

GROUP 02 MEMBERS

Ho Song Tin

Team Leader

Bui Thi Ngoc Chau

Member

Bui Thi Hong Thi

Member

Thai Anh Thu

Member

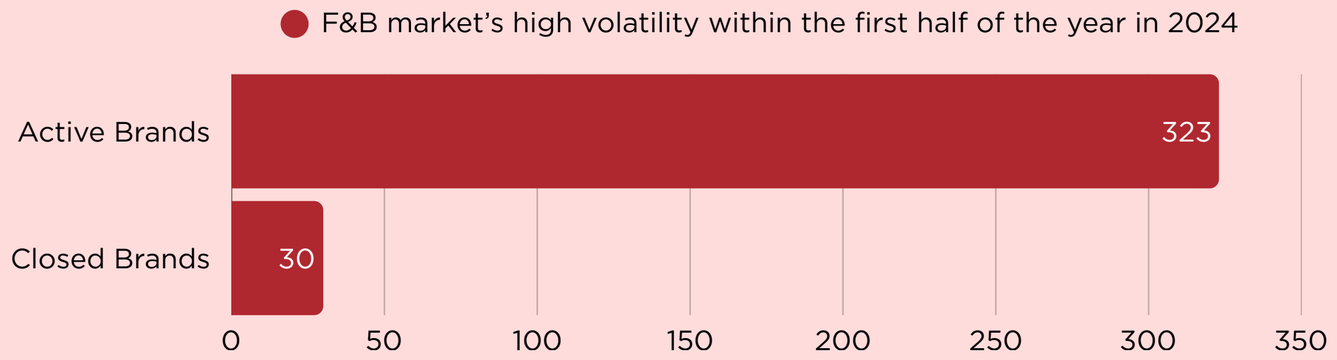
Le Huyen Tran

Member

TABLE OF CONTENTS

PROJECT OVEVRVIEW

PROJECT OVEVRVIEW



The Vietnamese F&B sector, especially the coffee market, is among the most competitive in Southeast Asia

To survive, Highlands Coffee should:

- Promotions that target the right customer segments, Strengthen loyalty
- Prevent churn
- Encourage higher-margin purchases without relying on across-the-board price cuts.

RELATED STUDIES

No	Study	Focus/Method	Key Findings
1	Yang et al. (2023)	Churn prediction in food delivery using ANN, SVM, XGBoost	Predictive models help reduce churn cost and strengthen loyalty.
2	AlShourbaji et al. (2023)	Enhanced Gradient Boosting with SVM	Outperformed traditional models in churn prediction accuracy.
3	Putri et al. (2024)	Customer segmentation via K-Means + RFM	Identified 6 customer types, enabling personalized marketing and lower costs.
4	Avcı et al. (2024)	Churn detection in F&B using supervised & unsupervised ML	Combined models accurately predict churn and guide prevention actions.
5	Gautam & Kumar (2022)	Sustainable marketing via K-Means clustering	Formed 5 income–spending groups to tailor marketing strategies.
6	Juhasz (2025)	RFM + K-Means++ in F&B marketing	Identified 5 customer groups for efficient resource allocation.
7	Berahmana et al. (2020)	RFM with K-Means, K-Medoids, DBSCAN	Helped identify potential customers and improve CRM decisions.

RELATED STUDIES

No	Study	Focus/Method	Key Findings
8	Ramkumar et al. (2025)	RFM + K-Means for loyalty management	Segmented “Best”, “At-risk”, and “Loyal” customers to maximize revenue.
9	Balo et al. (2025)	K-Means + BI for F&B marketing	Classified customers into core, lost, and new groups for digital strategy.
10	Czarniecka-Skubina et al. (2021)	Coffee consumer segmentation in Poland	3 segments: Neutral, Ad hoc, Non-specific; behavior-based offers needed.
11	Yang et al. (2023)	Customer retention via ML models	Retention proven more cost-effective than acquisition.
12	Zhao et al. (2025)	Coffee chain site selection using Random Forest	90% accuracy in store location prediction via geospatial data.
13	Phan (2020)	Restaurant behavior prediction via EDA + ML	Random Forest most accurate; Logistic Regression most interpretable.
14	AlShamsi (2022)	RFM with K-Means, K-Medoids, DBSCAN	Helped identify potential customers and improve CRM decisions.

BUSINESS OBJECTIVES

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

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?

OBJECTS AND SCOPES

Objects

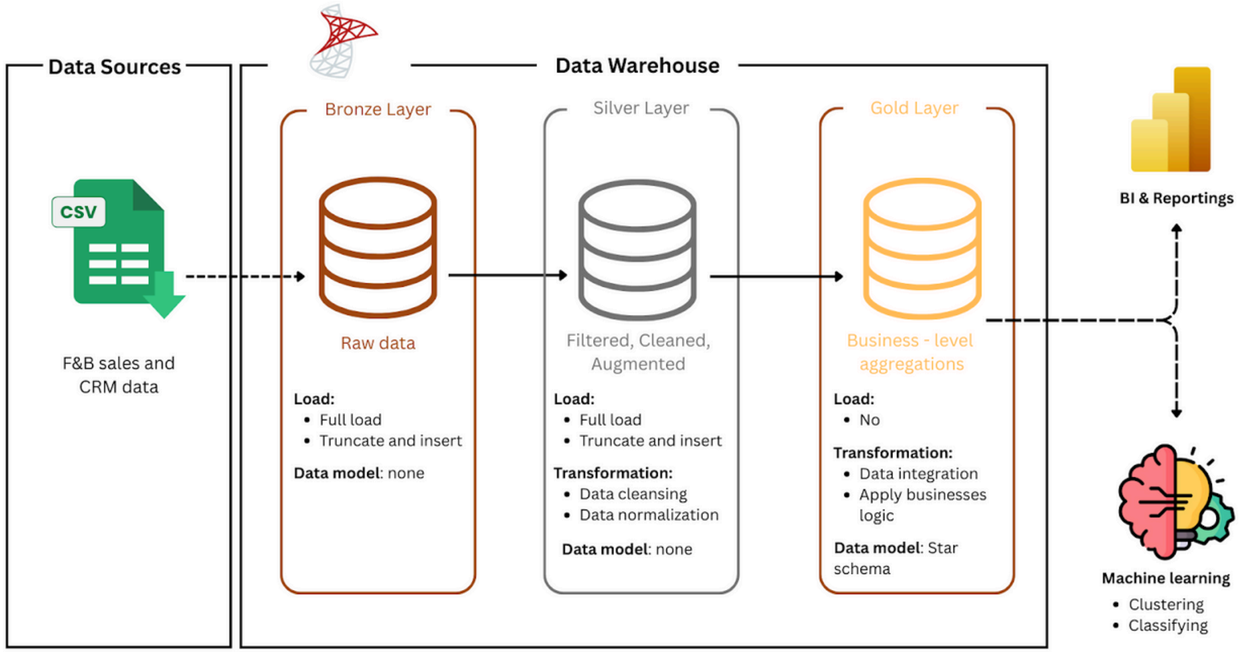
Highland Coffee stores and customers as well as other coffee chains operating in major cities

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



TOOLS AND PROGRAMMING LANGUAGES

SQL server

Excel

Power BI

Python

Google Colab

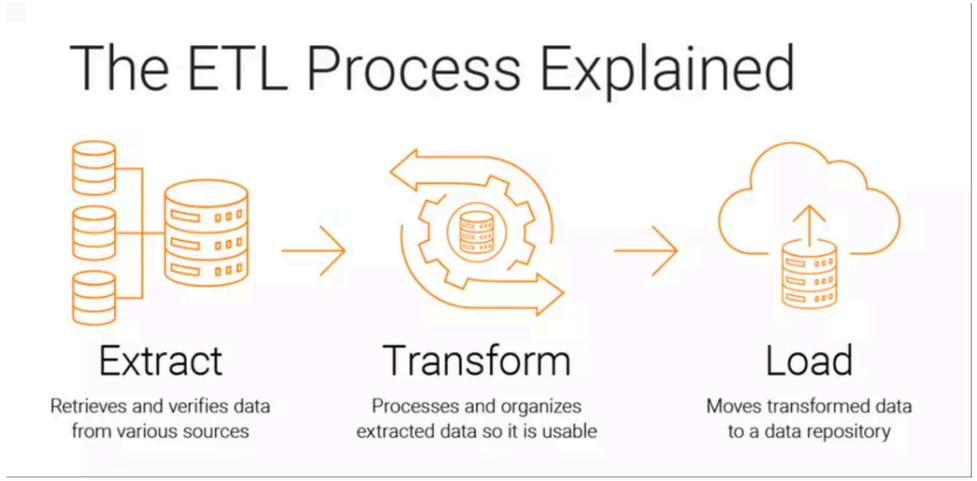
Github

CHAPTER 1. THEORETICAL BACKGROUND

1.1. F&B (FOOD AND BEVERAGE)

- Refers to all **activities** related to **producing**, **serving, and managing food and drinks**.
- Covers a wide range: **restaurants**, **cafés**, **hotels**, **airlines**, **catering services**.
- Success **depends** on **efficient operations**, inventory control, **customer service**, and **data analysis** for trend prediction and performance optimization.

1.2. ETL (EXTRACT – TRANSFORM – LOAD)



A key data integration process used to collect, clean, and load data into a central repository (e.g., data warehouse).

- **Extract:** Collect data from various sources (databases, APIs, files).
- **Transform:** Clean, filter, and format data for consistency and quality.
- **Load:** Transfer processed data to a warehouse/lake for analysis.

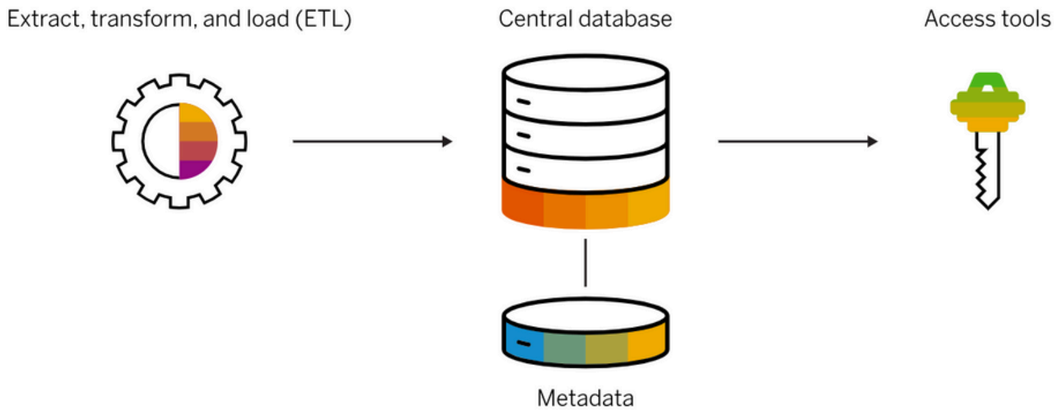
1.3. MEDALLION ARCHITECTURE

- Refers to all **activities** related to **producing**, **serving, and managing food and drinks**.
- Covers a wide range: **restaurants**, **cafés**, **hotels**, **airlines**, **catering services**.
- Success **depends** on **efficient operations**, inventory control, **customer service**, and **data analysis** for trend prediction and performance optimization.

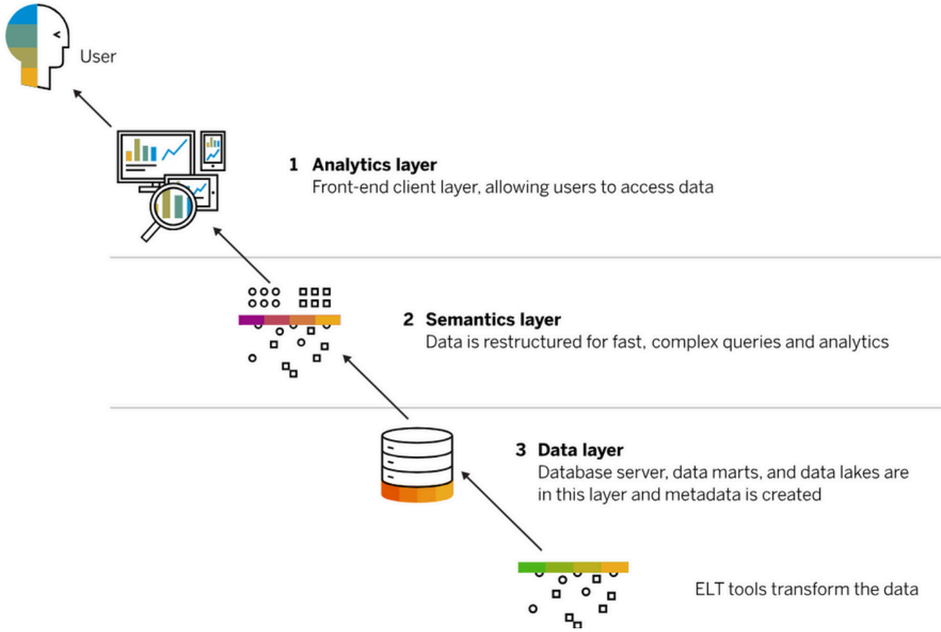
1.4. DATA WAREHOUSE

1.4. DATA WAREHOUSE

KEY COMPONENTS OF A DATA WAREHOUSE



DATA WAREHOUSE ARCHITECTURE



1.5. K-PROTOTYPES ALGORITHM

Combines **K-Means** (for numerical data) and **K-Modes** (for categorical data).

Advantage: Efficient for large, mixed datasets in real-world applications.

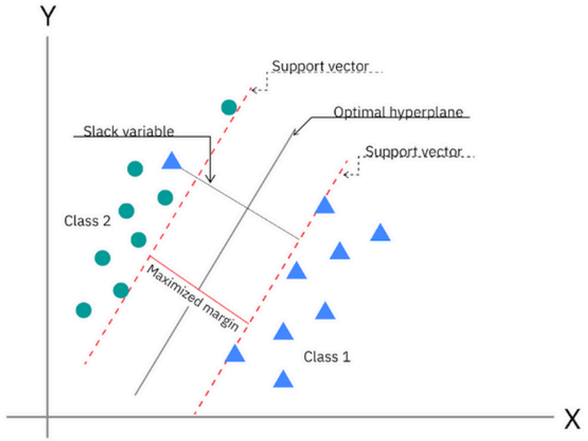
1.6. LOGISTIC REGRESSION

A supervised learning algorithm for binary classification (e.g., churn vs. retain).

Strengths:

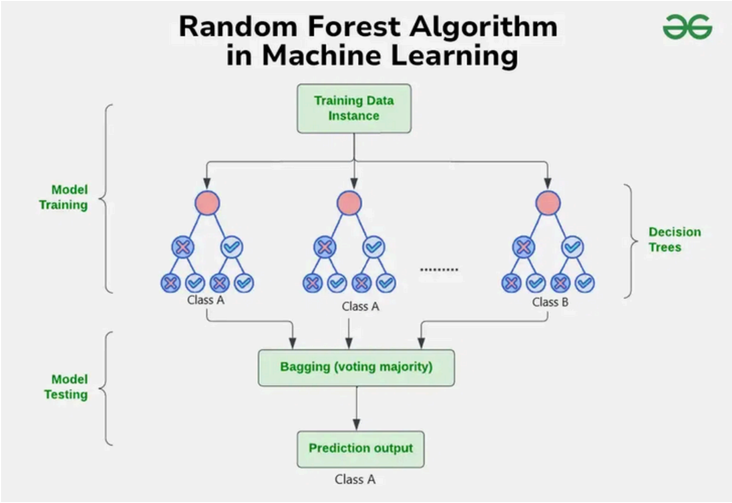
- Simple and interpretable
- Fast to train and easy to implement
- Works well as a baseline model for classification tasks

1.7. SUPPORT VECTOR MACHINE (SVM)



Purpose: Supervised algorithm mainly used for classification.
Key Idea: Finds the best hyperplane that separates data into classes with the maximum margin.
Support Vectors: Data points closest to the hyperplane; determine its position and direction.
Kernel Trick: Maps data to higher dimensions to handle non-linear separation.

1.8. RANDOM FOREST



Type: Ensemble learning algorithm for classification & regression.
Concept: Builds multiple decision trees on random subsets of data and features (bootstrap sampling).
Advantages:

- Reduces overfitting & variance
- Improves accuracy & stability

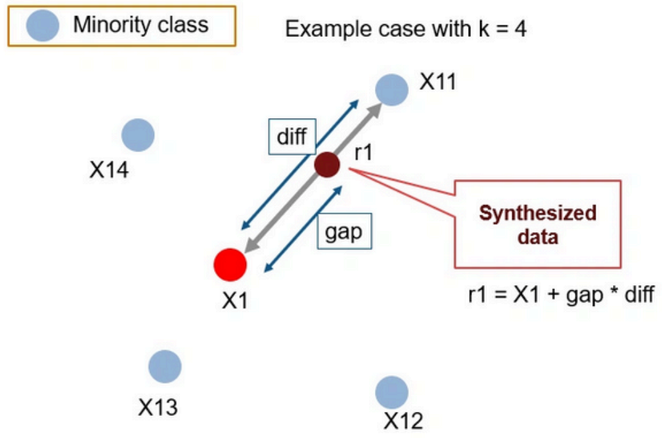
1.9. XGBOOST (EXTREME GRADIENT BOOSTING)

Concept: Sequentially builds an ensemble of weak learners (decision trees) that correct previous errors.

Advantages:

- High accuracy
- Handles non-linear patterns
- Prevents overfitting (regularization)
- Fast training via parallel computation

1.10. SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)



Goal: Solve class imbalance in datasets.

Method: Creates synthetic samples for minority class instead of duplicating.

Benefit: Improves model performance on minority class without losing information.

CHAPTER 2. DATA PREPARATION

This chapter explains data preparation and model setup. It introduces nine datasets from six major Vietnamese cities (2017–2019), describes data cleaning and transformation, performs exploratory analysis to reveal key patterns, and concludes with an aggregated data model (ERD) used for analyzing customer behavior and brand performance.

2.1. DATA COLLECTION AND DESCRIPTION

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.

Dataset Overview

- 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
- Demographics & Behavior 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

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

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	6%	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

Column	Description	Data Type	Null	Role
ID	Unique identifier for each respondent.	Identifier	0%	Identifier
Year	Year of data collection.	Numeric	0%	Contextual
City	The city where the respondent resides.	Categorical	0%	Demographic
Brand	The brand being evaluated.	Categorical	0%	Contextual
Spontaneous	The brand that comes first to the respondent's mind	Categorical	5.835%	Perceptual
Awareness	The brand that the respondent is aware of.	Categorical	15%	Perceptual
Trial	Whether the respondent has ever tried the brand.	Categorical	3.640%	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

Column	Description	Data Type	Null Percentage	Role
Comprehension	How well the respondent	Categorical	6.460%	Perceptual
BrandLikability	The consumer's level of affection or	Categorical	8.612%	Behavioral
Weekly	Indicates if the	Categorical	82%	Behavioral
Daily	Indicates if the respondent uses	Categorical	90%	Behavioral
Fre#Visit	Number of visits made to the brand	Numeric	7.402%	Behavioral
PPA	Price Per Average,	Numeric	8.109%	Analytical
Spending	Total amount of money spent on the	Numeric	8.109%	Behavioral
Segmentation	Customer segment label (e.g., Seg.02 -	Categorical	8.109%	Analytical
NPS#P3M	Net Promoter Score	Numeric	7.097%	Analytical
NPS#P3M#Group	NPS classification group: Promoter, Passive, or Detractor.	Categorical	7.097%	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

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

Description: The Competitor 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

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

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	22%	Behavioral
Visit#Dayofweek	Number of visits made on that particular day of the week.	Numeric	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

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	6%	Behavioral
Visit#Dayofweek	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

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
NeedstateGroup	Broader category grouping similar need states.	Categorical	0%	Analytical

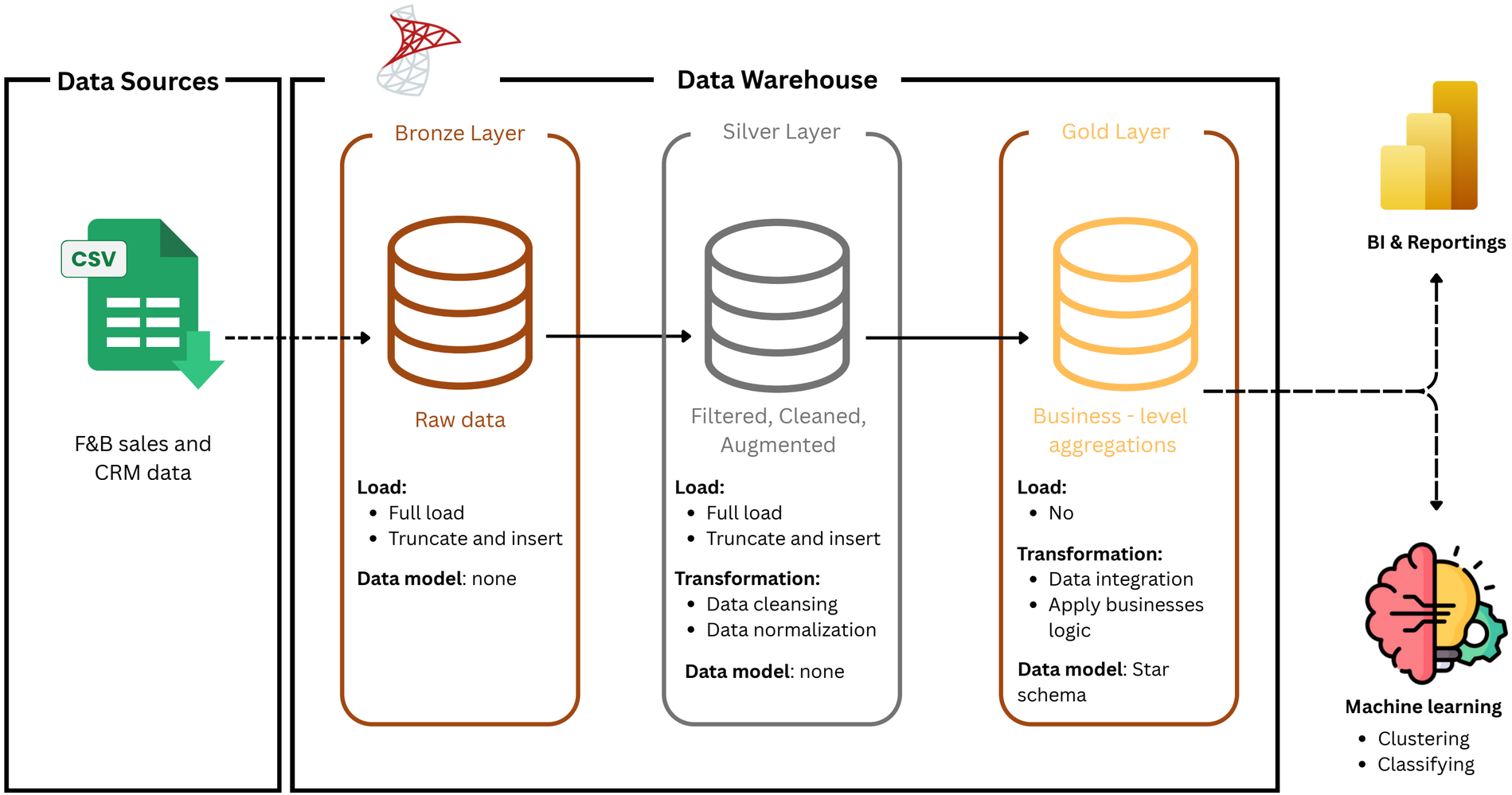
DATASET 9. DEMOGRAPHICS & BEHAVIOR (SA#VAR)

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

Column	Description	Data Type	Null	Role
ID	Unique identifier for	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	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	Categorical	0%	Perceptual
Gender	Gender of the respondent.	Categorical	0%	Demographic

Column	Description	Data Type	Null	Role
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/Derived

DATA WAREHOUSE ARCHITECTURE



2.2 DATA CLEANING

2017Segmentation Case

- Parsed key fields (SegmentCode, CustomerType) from composite text into a Segmentation_Lookup table.
- Corrected typos (“Indepentdent” → “Independent”).
- Scaled Spending and PPA values by 1,000 to reflect true amounts.

Segmentation_Lookup

SegmentCode	SegmentName	SpendingRange
Seg.01	Mass	<VND 25K
Seg.02	Mass Asp	VND 25K - VND 59K
Seg.03	Premium	VND 60K - VND 99K
Seg.04	Super Premium	VND 100K+

2.2 DATA CLEANING

BrandHealth

- Standardized brand names (e.g., KOI cafe → KOI Cafe, Street Coffee).
- Unified SegmentCode based on the lookup table.
- Converted behavioral variables (Awareness, Trial, Recall, etc.) into binary indicators (1 = Yes, 0 = No).
- Labeled missing values (“N/A”, “Did not answer”).
- Calculated PPA = Spending / Visit Frequency for valid entries only.

Other Tables

- Standardized city names using CityID.
- Assigned “N/A” to blank fields (DayPart, VisitFreq, Needstates).
- Fixed minor formatting issues (e.g., extra characters).

Respondent (SA#Var)

- Removed redundant fields (MPI#Mean, AgeGroup#2, etc.).
- Excluded records with missing demographic info.
- Standardized brand names and age group labels (“18–24”).
- Retained missing data as “N/A” for MPI and BUMO_Previous.

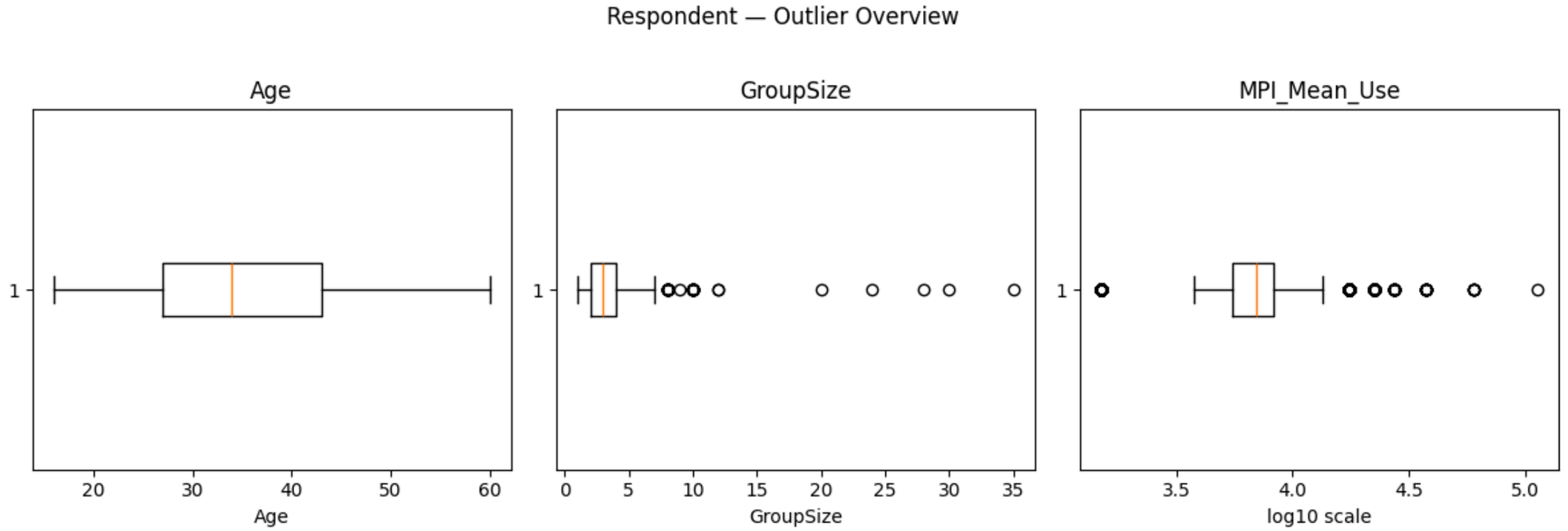
2.3 DATA UNDERSTANDING

Respondent

Variable	mean	median	std	min	max
RespondentKey	58.690	58.690	338.832	10	11737
RespondentID	44.293.075	4.339.430	26.737.144	891.000	8.637.540
CityID	286	20	162	10	60
GroupSize	329	30	133	10	350
Age	3.522	340	1.082	160	600
Year	201.799	20.180	78	20.170	20.190
MPI_Mean_Use	733.874	69.990	466.848	14.990	1.124.990

2.3 DATA UNDERSTANDING

Respondent



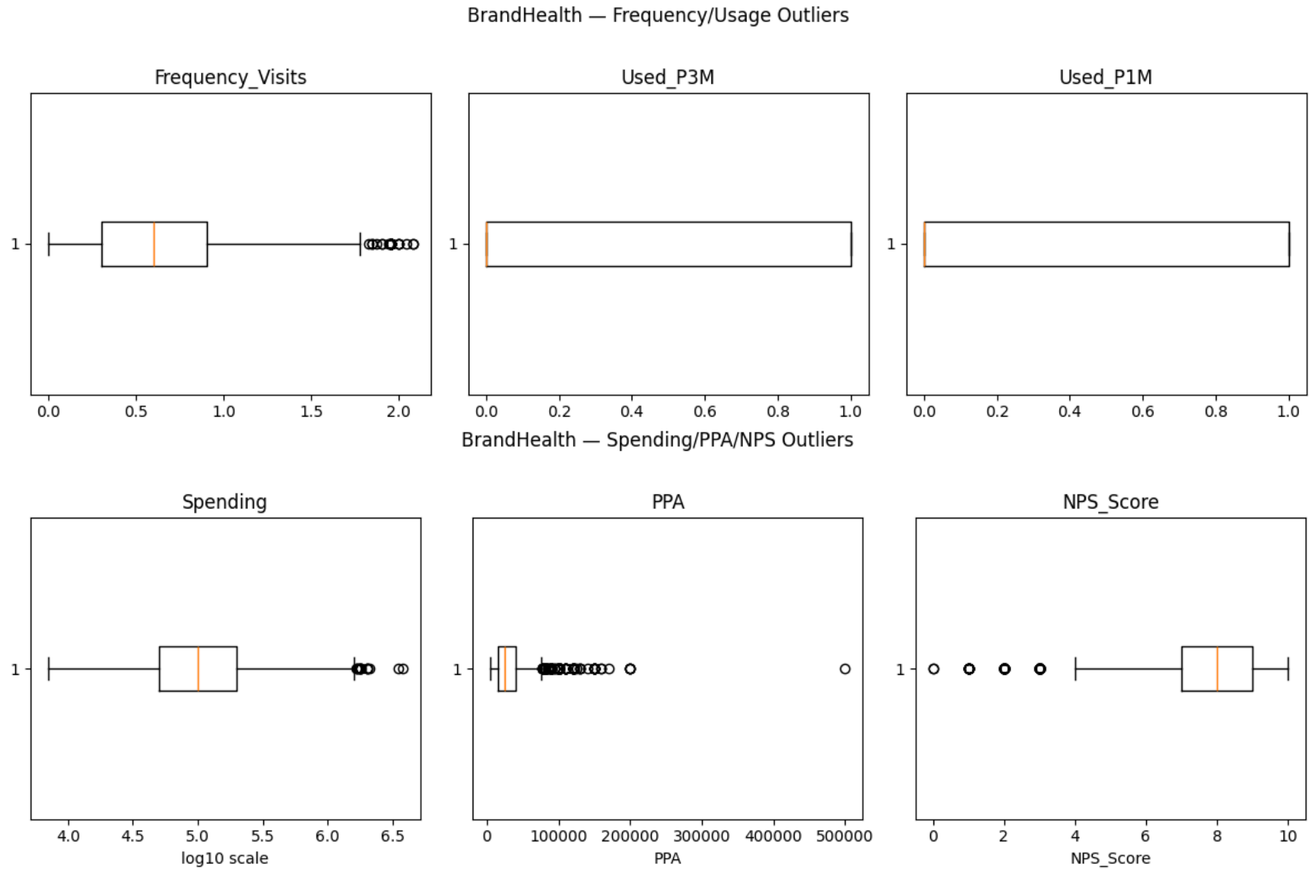
2.3 DATA UNDERSTANDING

BrandHealth Data

Variable	mean	median	std	min	max
RespondentID	47.827.787	4.437.200	26.814.183	891.000	8.637.540
Year	20.180	20.180	78	20.170	20.190
CityID	263	20	160	10	60
Is_Spontaneous_Aware	42	0	49	0	10
Is_Aware	99	10	4	0	10
Is_Trial	64	10	48	0	10
Has_Brand_Likability	14	0	35	0	10
Used_P3M	39	0	49	0	10
Used_P1M	26	0	44	0	10
Frequency_Visits	303	0	694	0	1.200
PPA	2.982.456	250.000	1.907.469	50.000	5.000.000
Spending	15.501.471	1.000.000	17.398.636	70.000	37.500.000
NPS_Score	797	80	135	0	100

2.3 DATA UNDERSTANDING

BrandHealth Data



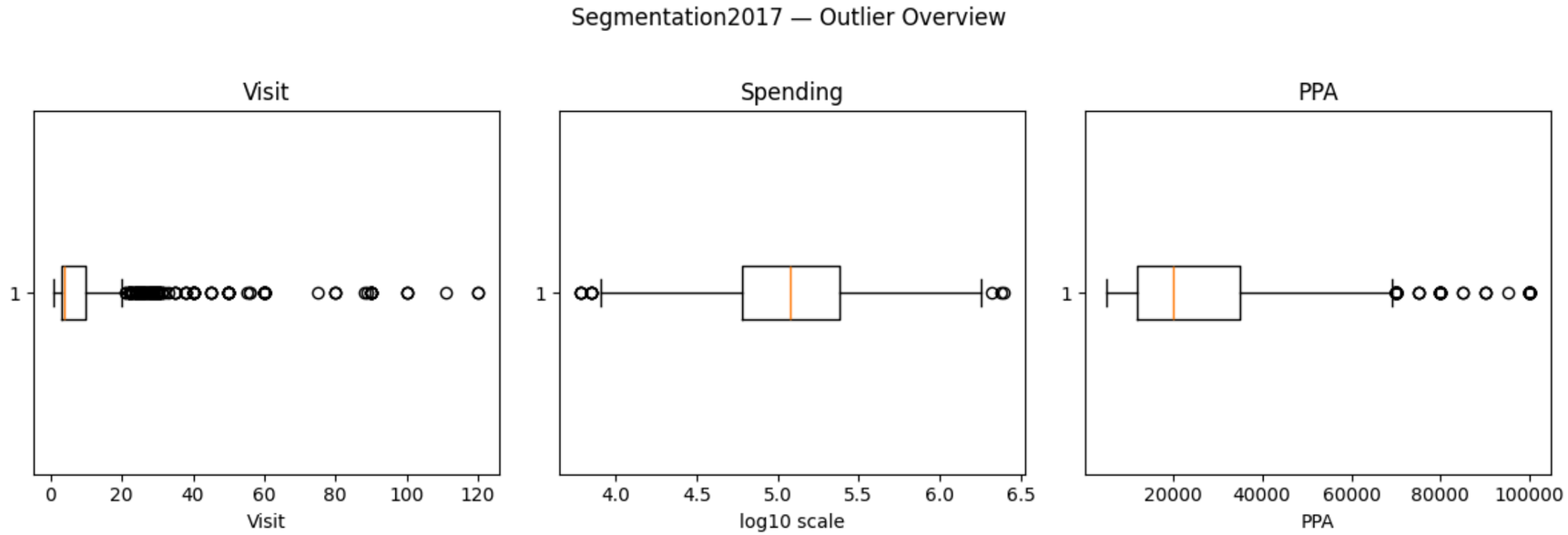
2.3 DATA UNDERSTANDING

Segmentation2017 Data

Variable	mean	median	std	min	max
2017Segmentation CaseKey	247.250	247.250	142.735	10	49.440
RespondentID	12.474.673	1.276.020.000	1.420.000	891.000	1.424.790
Visit	929	40	1.125	10	1.200
Spending	18.521.278	1.200.000	20.838.354	60.000	25.000.000
PPA	2.612.561	200.000	1.745.401	50.000	1.000.000

2.3 DATA UNDERSTANDING

Segmentation2017 Data



2.3 DATA UNDERSTANDING

Brand Image

Variable	mean	median	std	min	max
BrandImageKey	32.153.650	3.215.365	18.563.904	10	6.430.720
RespondentID	48.693.805	4.446.930	27.260.196	891.000	8.637.540
Year	201.811	20.180	79	20.170	20.190
CityID	247	20	154	10	60

2.3 DATA UNDERSTANDING

Companion

Variable	mean	median	std	min	max
CompanionKey	103.700	103.700	598.698	10	207.390
RespondentID	46.547.081	4.397.500	27.226.607	891.000	8.637.540

2.3 DATA UNDERSTANDING

Day Part

Variable	mean	median	std	min	max
DayPartKey	95.950	95.950	553.953	10	191.890
RespondentID	45.416.515	4.361.610	27.576.832	891.000	8.637.540
CityID	275	20	162	10	60
VisitFreq	695	40	772	10	600
Year	201.802	20.180	81	20.170	20.190

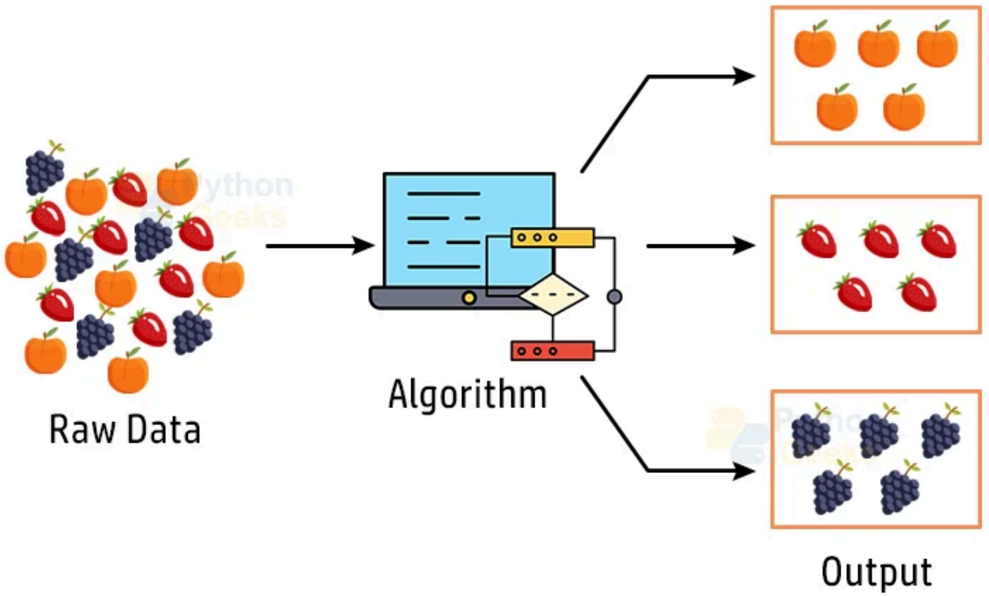
2.3 DATA UNDERSTANDING

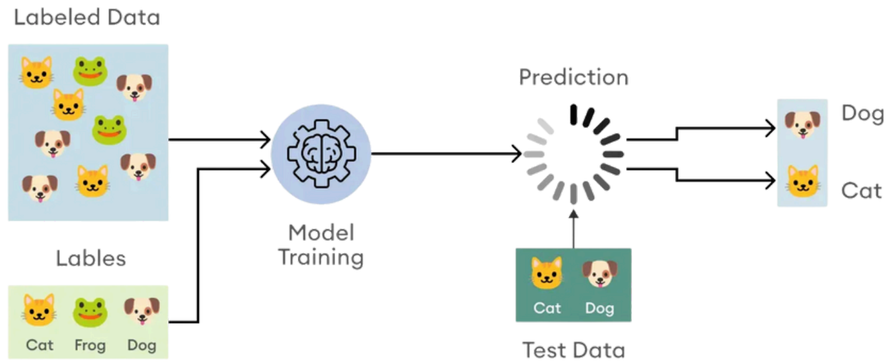
Need State by Day & Daypart

Variable	mean	median	std	min	max
NeedstateDayDaypartKey	376.260	376.260	2.172.324	10	752.510
RespondentID	63.189.821	7.889.310	25.721.833	891.000	8.637.540
CityID	266	20	162	10	60
Year	201.852	20.190	74	20.170	20.190

2.4 DATA MODEL

CHAPTER 3 EXPERIMENTAL RESULTS AND EVALUATION

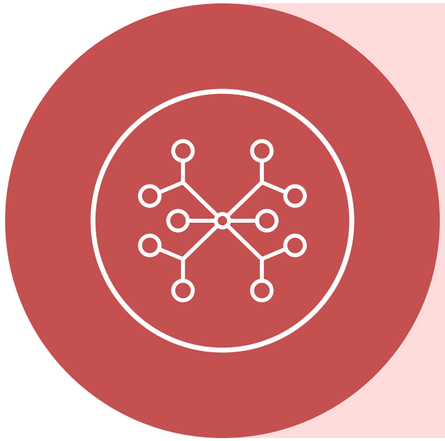




Clustering

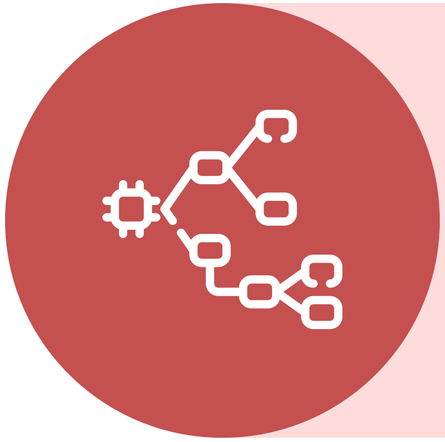
Classification

OBJECTIVES



Clustering

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.



Classification

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.

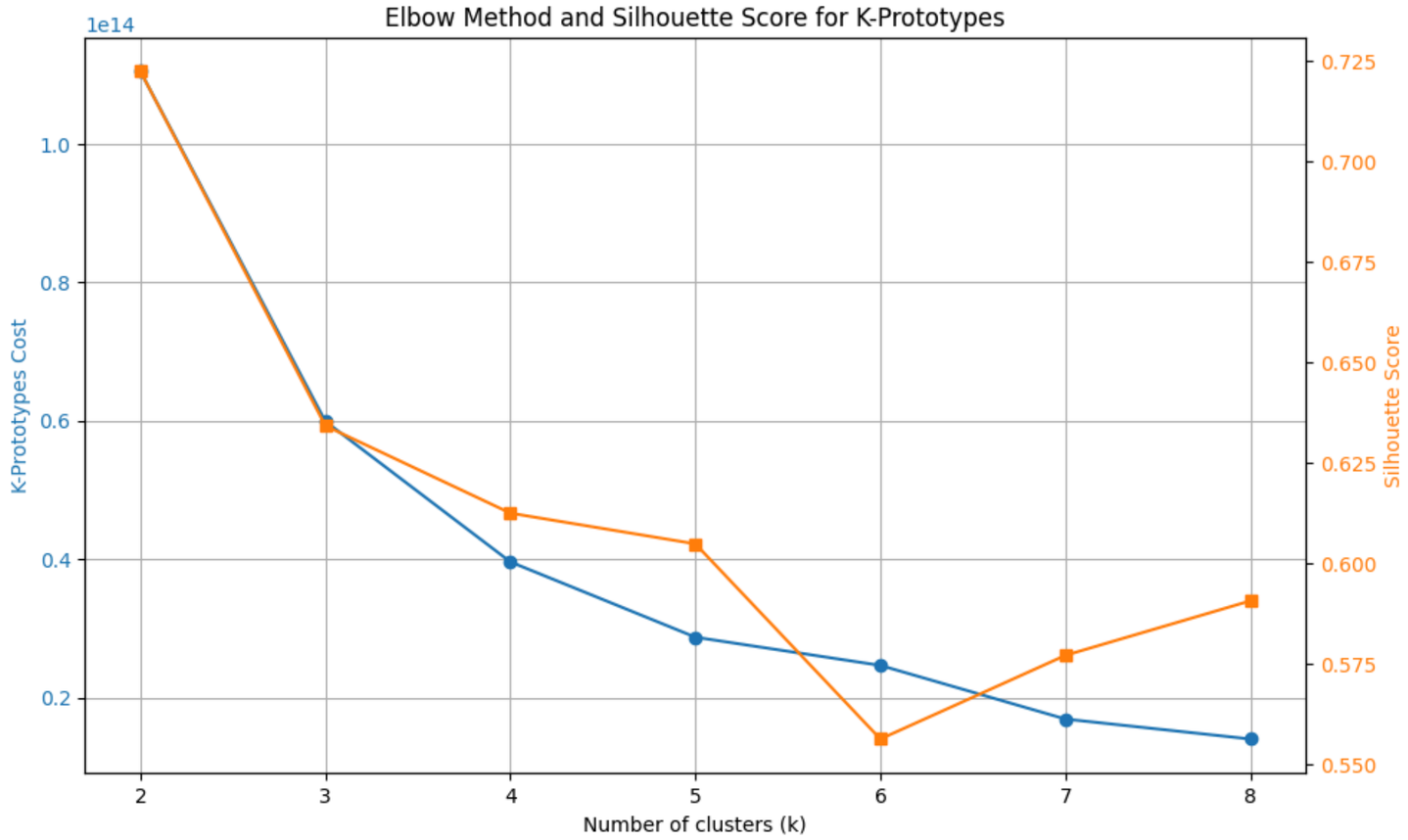
3.1 CUSTOMER SEGMENTATION

Dataset Description For Clustering

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 CUSTOMER SEGMENTATION

Parameter Setting and Experimental Design



3.1 CUSTOMER CHURN PREDICTION

Dataset Description For Churn Prediction

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 CUSTOMER CHURN PREDICTION

Experimental Design - Stage 1

Model Performance - Default Model

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	4,37	70,48
SVC	66,16	89,43	10,08	18,11	60,43

3.2 CUSTOMER CHURN PREDICTION

Experimental Design - Stage 1

Model Performance - SMOTE

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	60,37	45,64	35,09	39,65	59,4
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

3.2 CUSTOMER CHURN PREDICTION

Stage 2: Hyper Parameter Tuning for XGboost

Optimal Hyperparameters

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	1	Step size shrinkage that controls how much each tree contributes to the final model.
subsample	8	Fraction of the training samples randomly selected to grow each tree.
colsample_bytree	8	Fraction of features randomly selected for each tree.

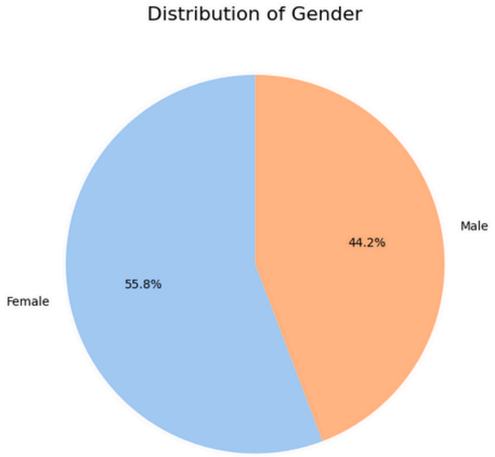
Before and After Optimizing

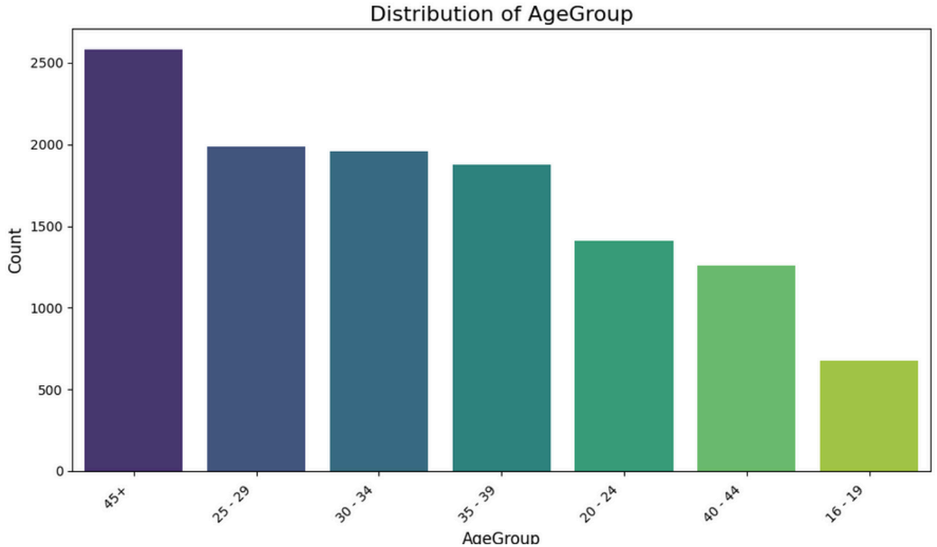
Metric	Before Fine-Tuning	After Fine-Tuning	Change
Accuracy	67,93	66,12	–1,81
Precision	62,16	55,49	–6,67
Recall	34,98	44,70	9,72
F1-score	44,74	49,49	4,75
ROC-AUC	70,14	68,72	–1.42

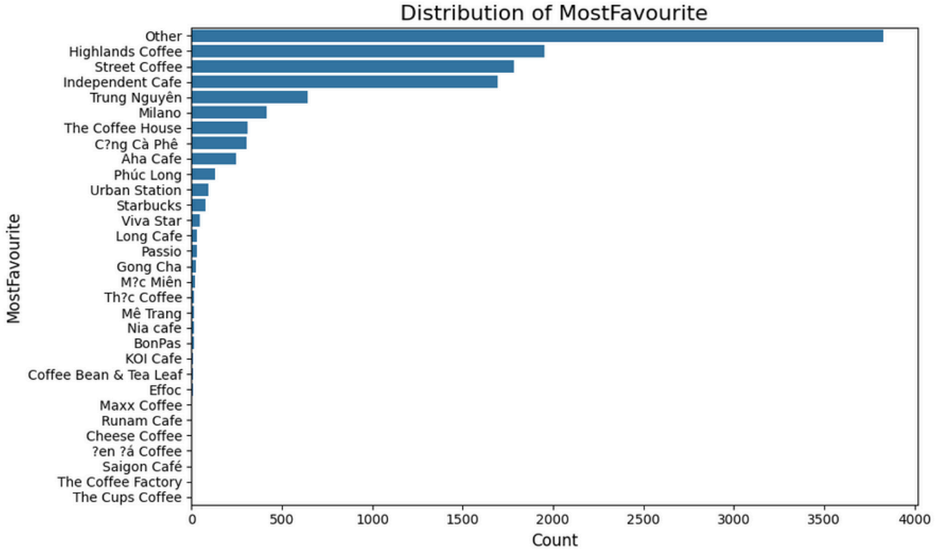
CHAPTER 4.
DASHBOARD AND VISUALIZATION

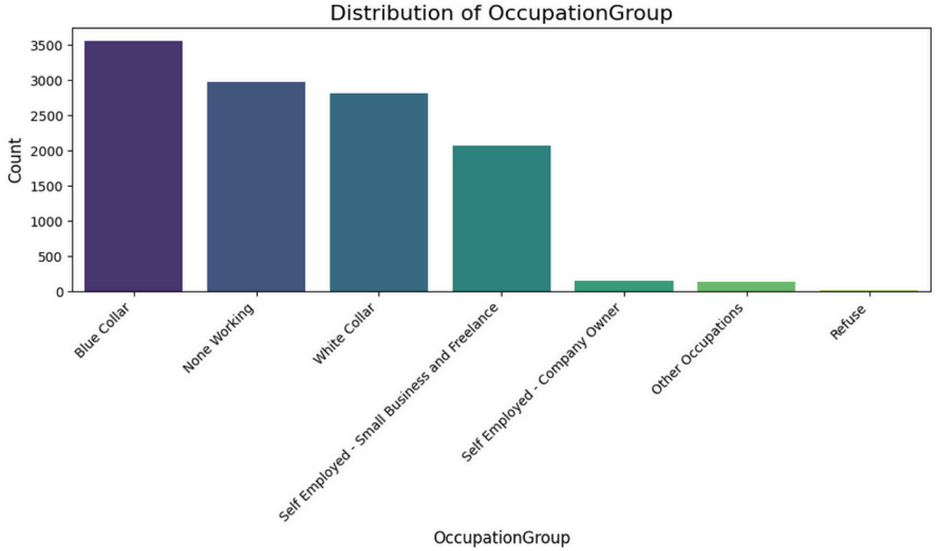
4.1. EXPLORATORY DATA ANALYSIS (EDA)

Customer demographics



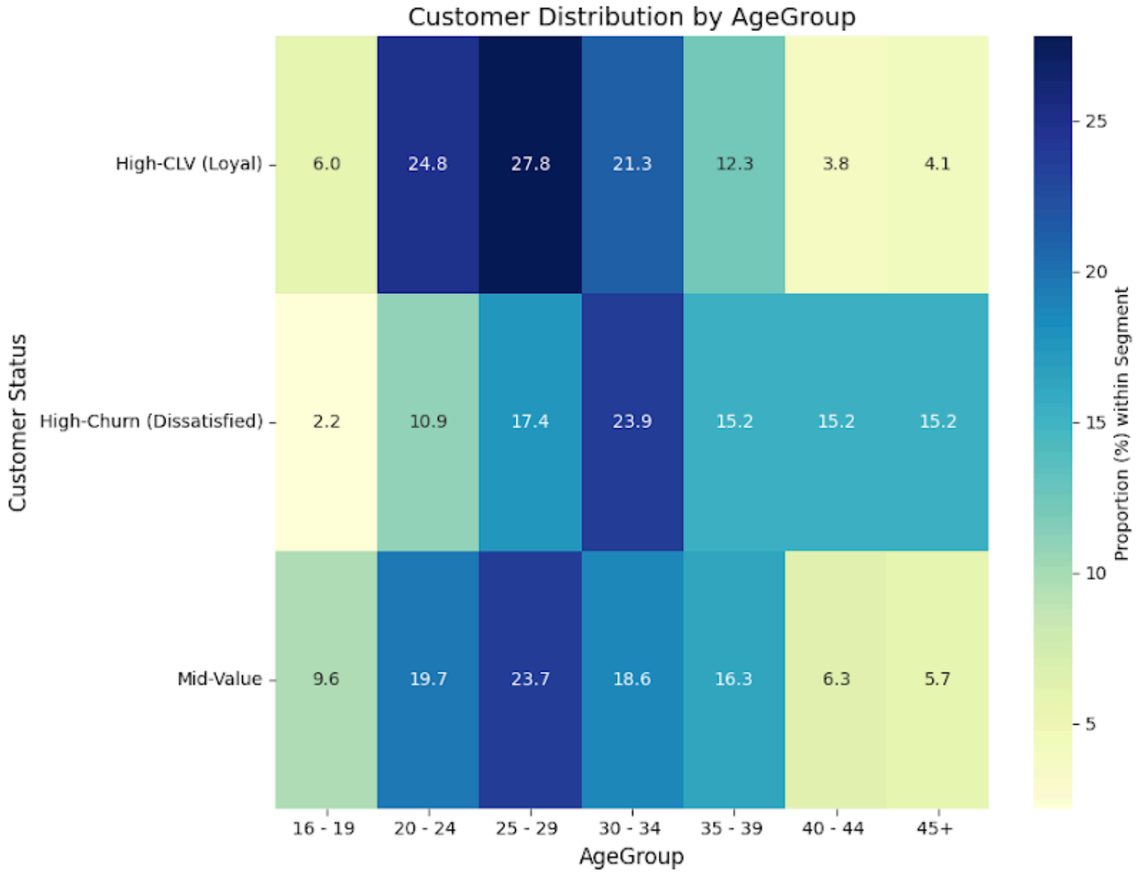


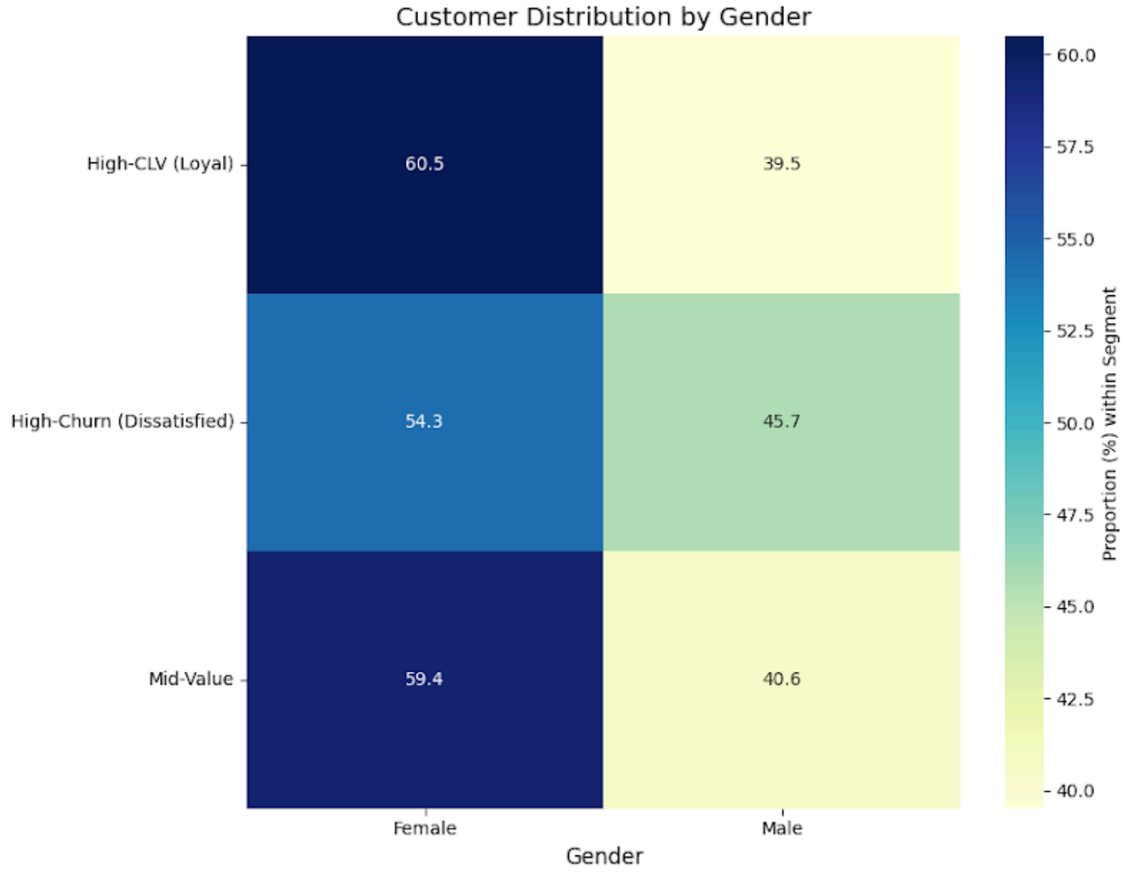




4.1. EXPLORATORY DATA ANALYSIS (EDA)

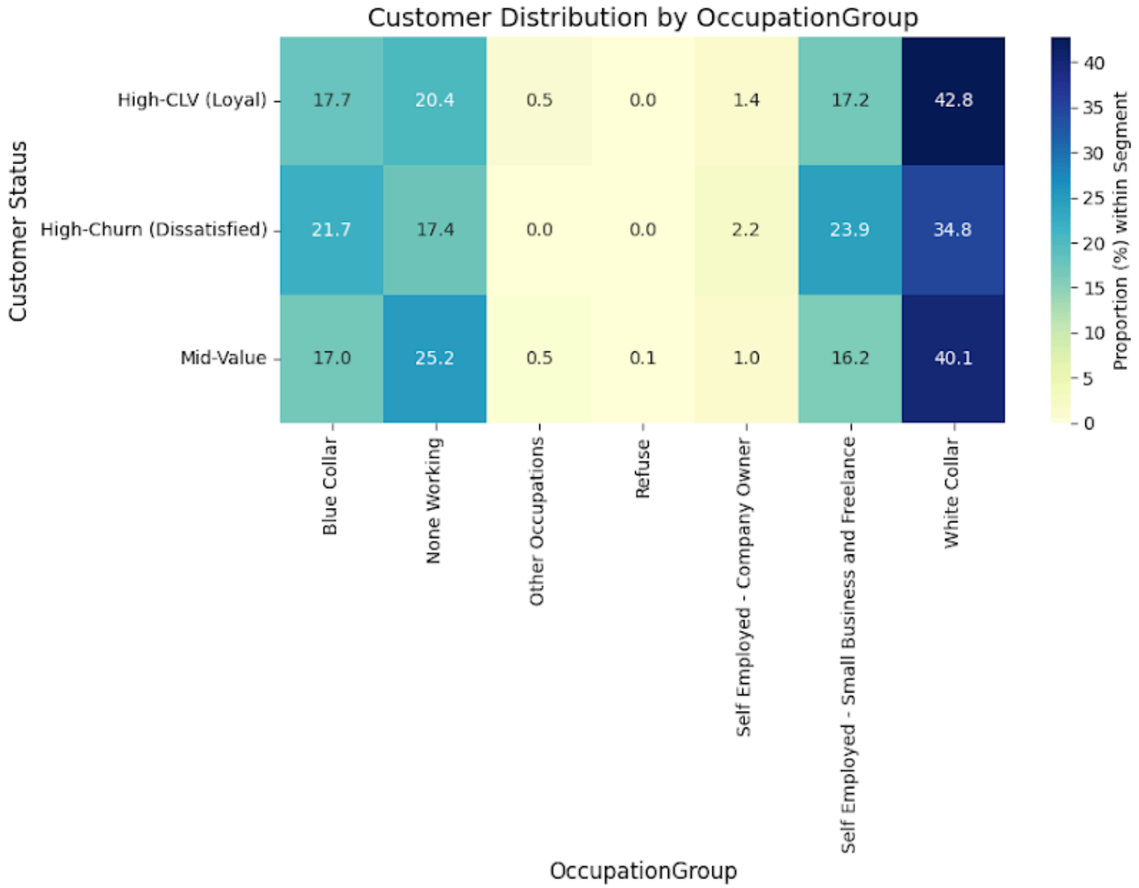
Customer CLV and Churn

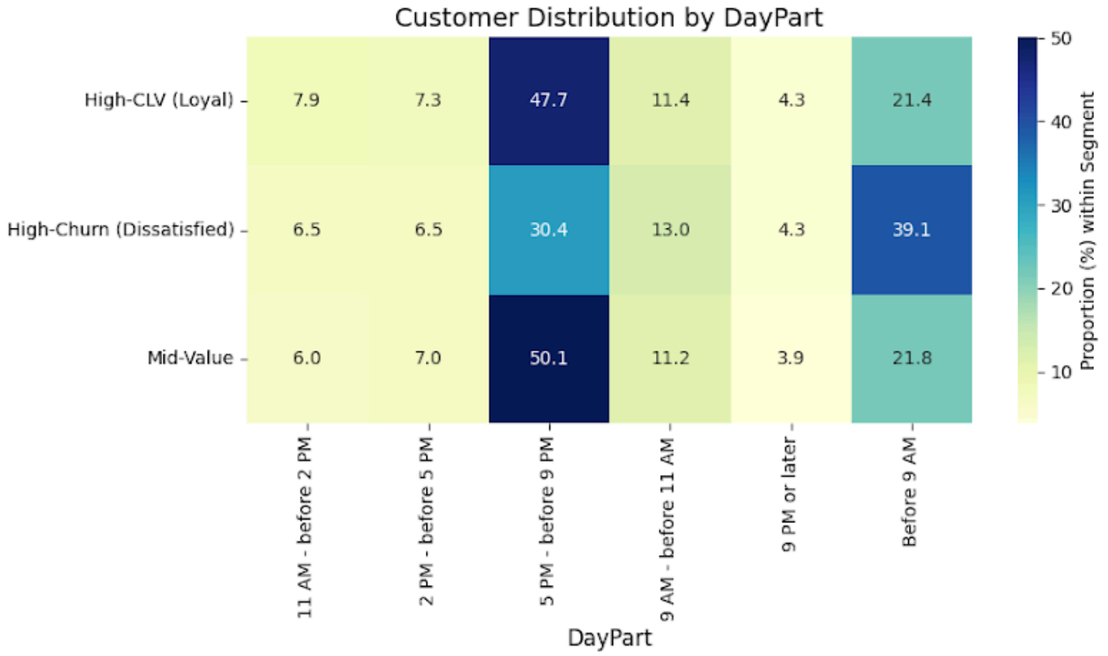




4.1. EXPLORATORY DATA ANALYSIS (EDA)

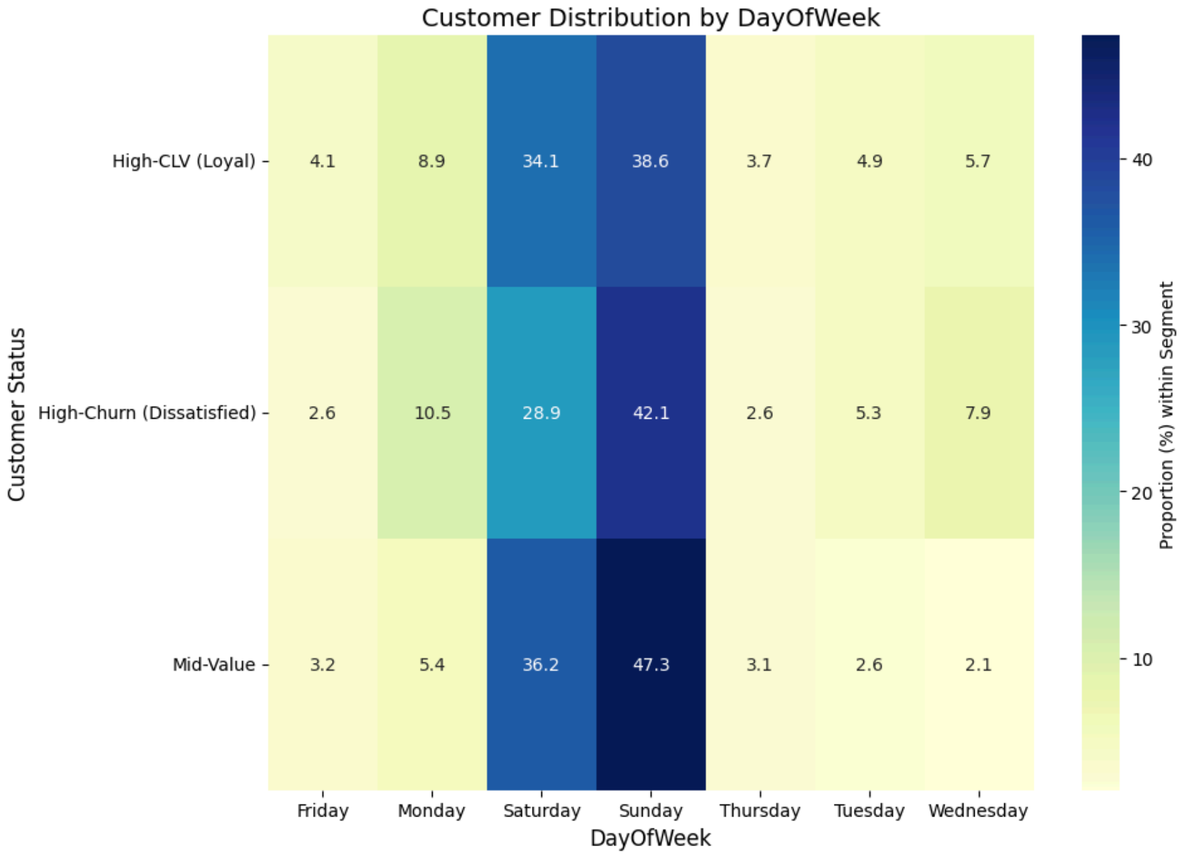
Customer CLV and Churn





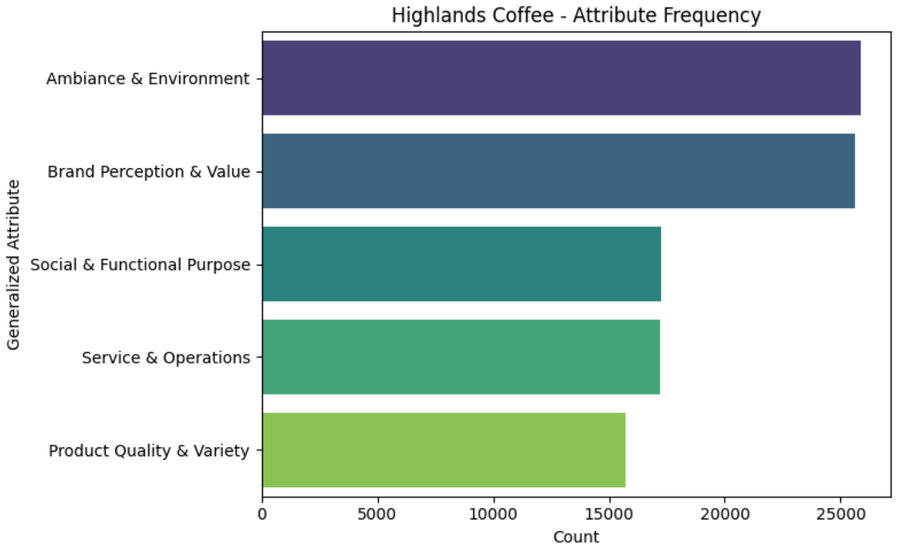
4.1. EXPLORATORY DATA ANALYSIS (EDA)

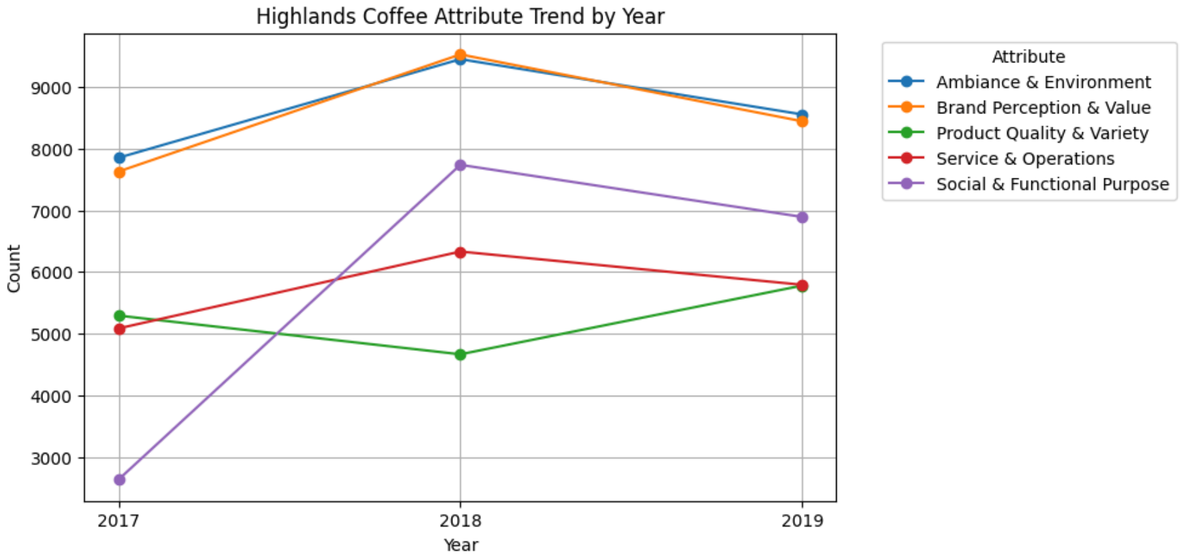
Customer CLV and Churn

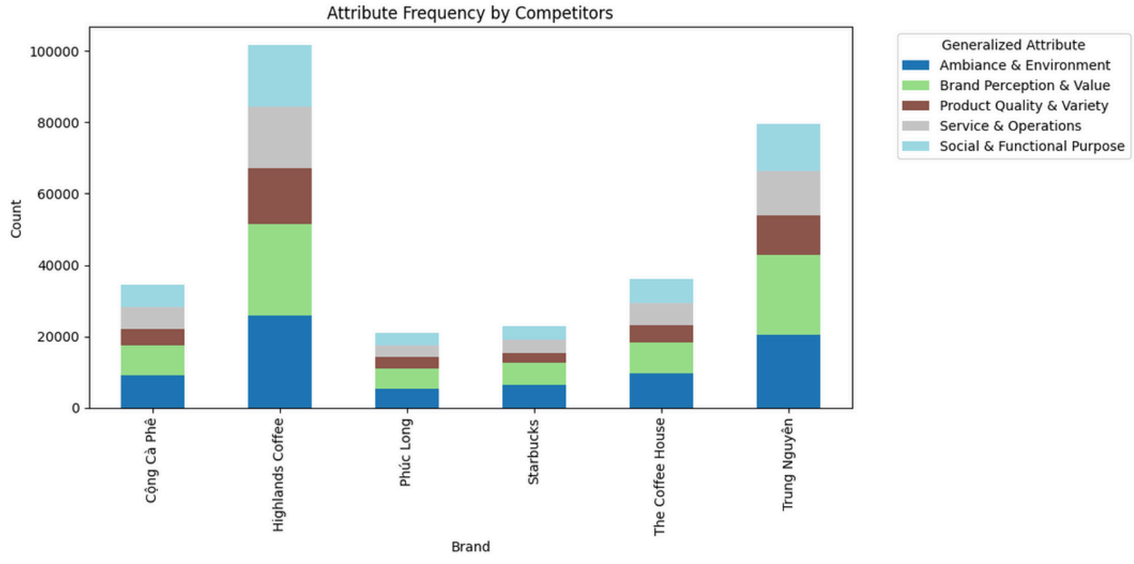


4.1. EXPLORATORY DATA ANALYSIS (EDA)

Attribute frequency

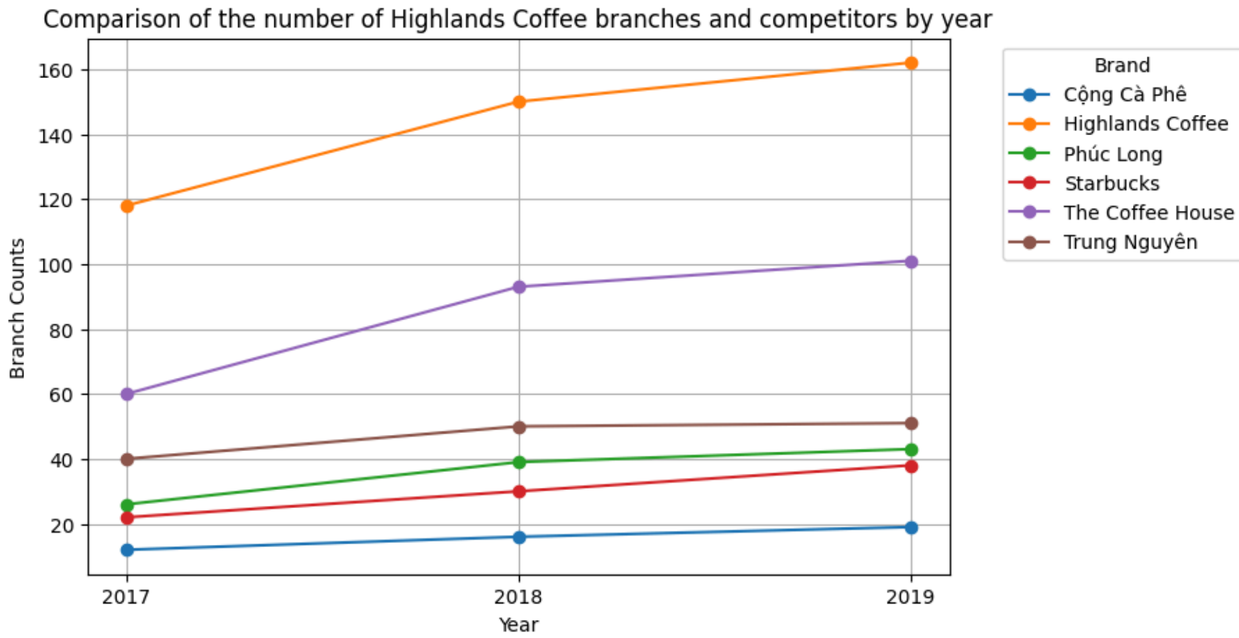


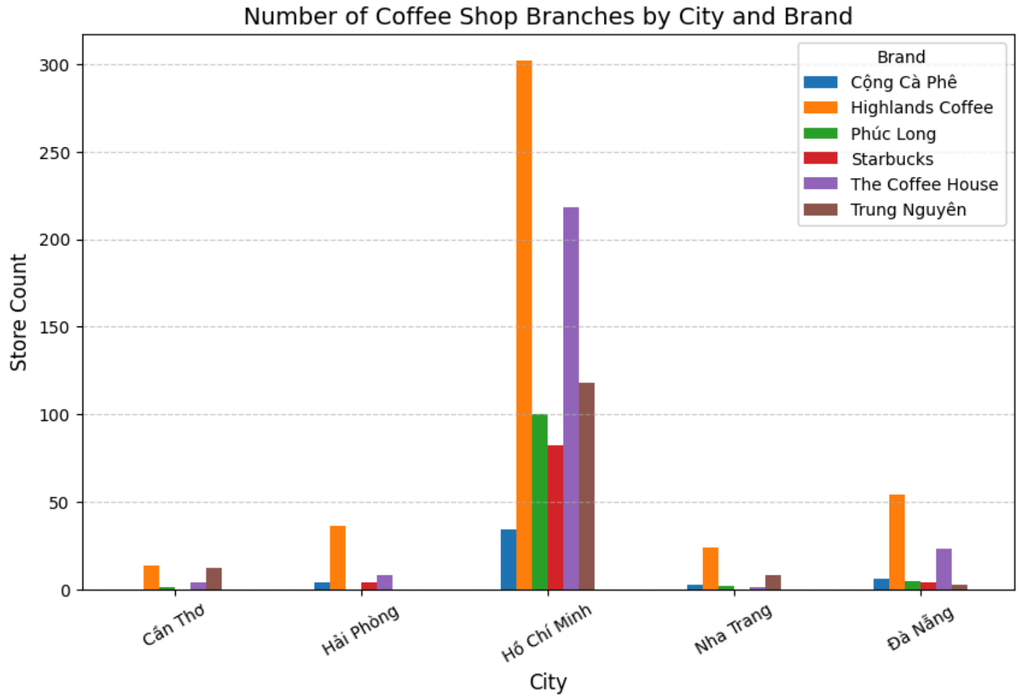




4.1. EXPLORATORY DATA ANALYSIS (EDA)

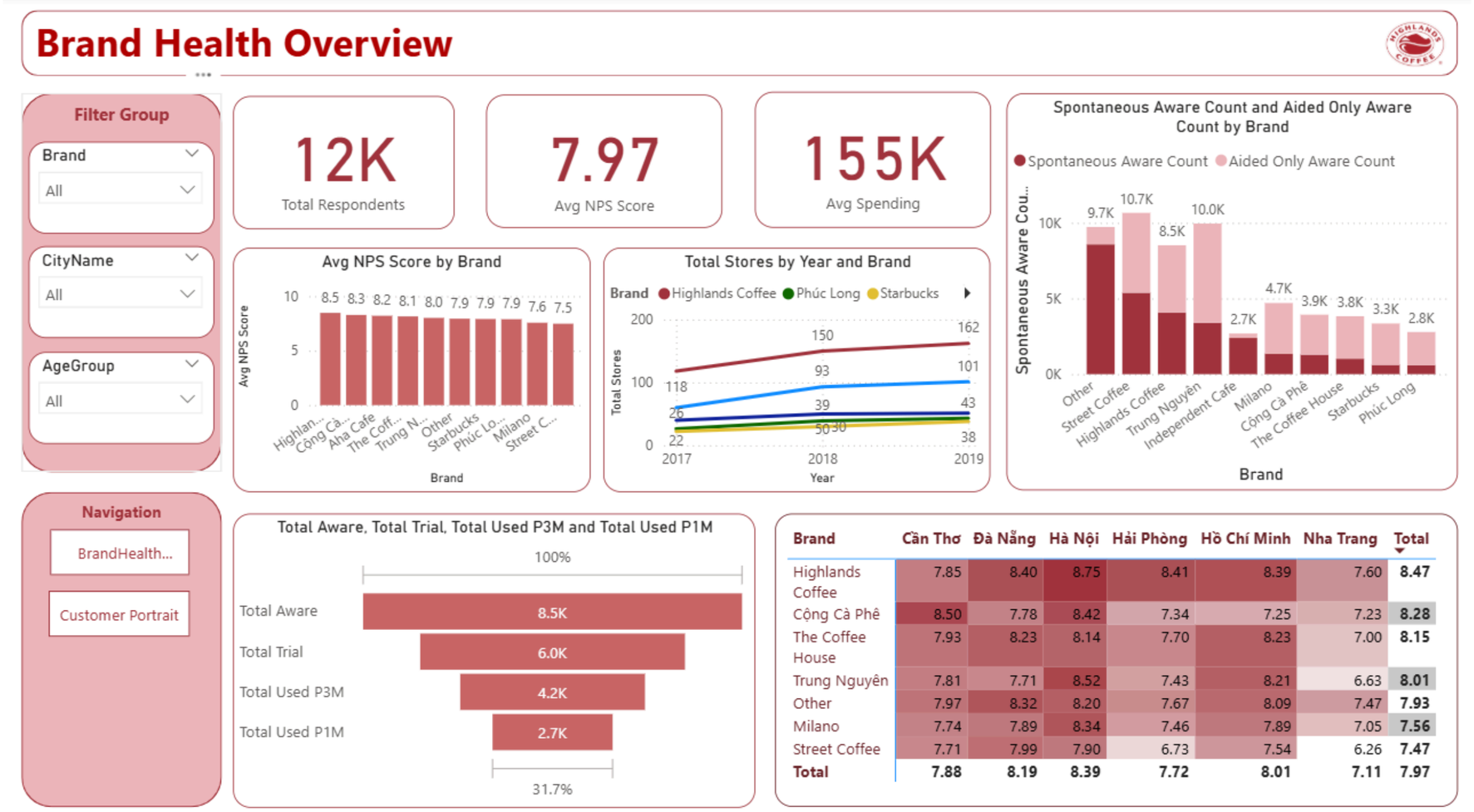
Brand Counts





4.1. DASHBOARD AND VISUALIZATION

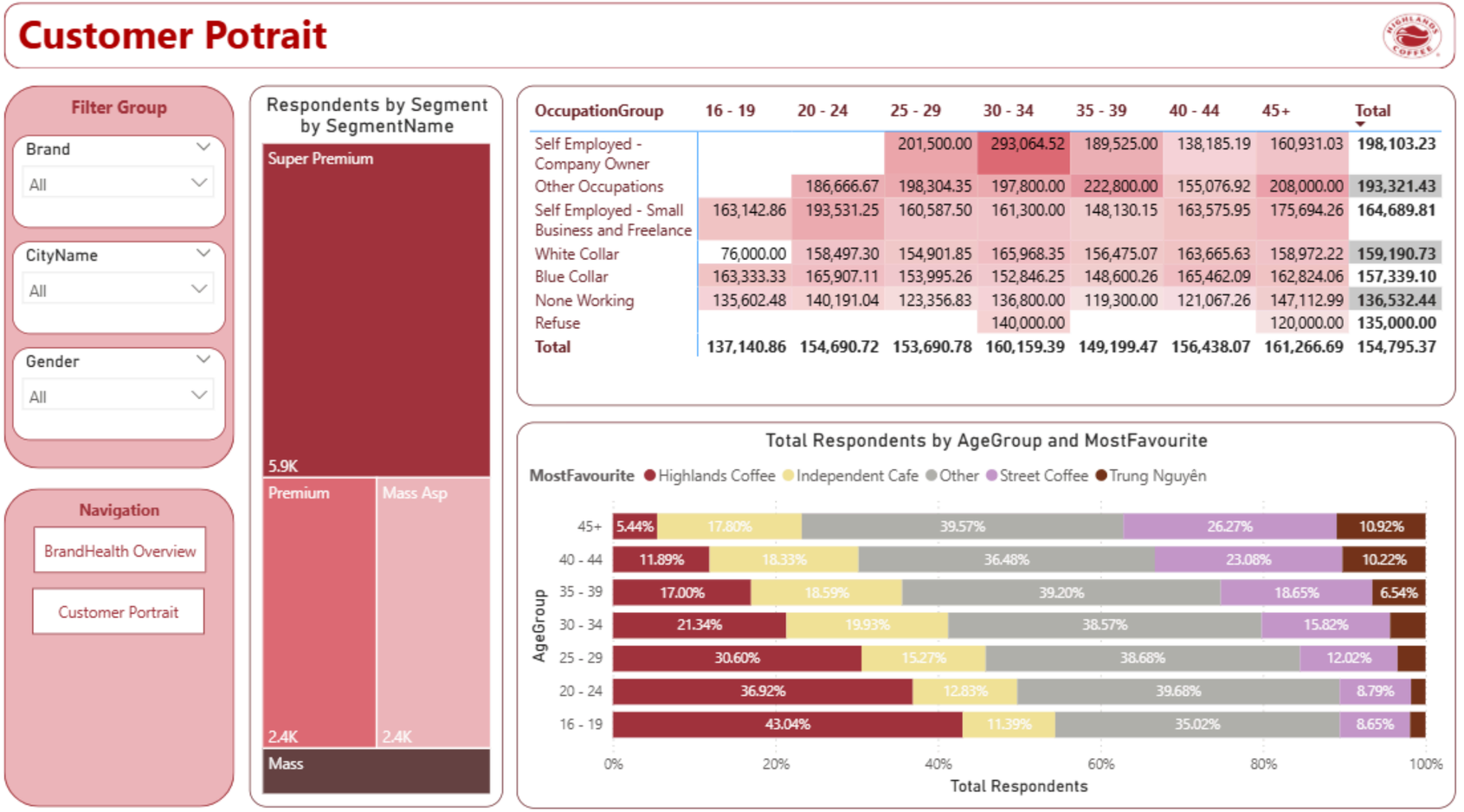
BrandHealth Overview



Brand	Cần Thơ	Đà Nẵng	Hà Nội	Hải Phòng	Hồ Chí Minh	Nha Trang	Total
Highlands Coffee	7.85	8.40	8.75	8.41	8.39	7.60	**8.47**
Cộng Cà Phê	8.50	7.78	8.42	7.34	7.25	7.23	**8.28**
The Coffee House	7.93	8.23	8.14	7.70	8.23	7.00	**8.15**
Trung Nguyên	7.81	7.71	8.52	7.43	8.21	6.63	**8.01**
Other	7.97	8.32	8.20	7.67	8.09	7.47	**7.93**
Milano	7.74	7.89	8.34	7.46	7.89	7.05	**7.56**
Street Coffee	7.71	7.99	7.90	6.73	7.54	6.26	**7.47**
Total	**7.88**	**8.19**	**8.39**	**7.72**	**8.01**	**7.11**	**7.97**

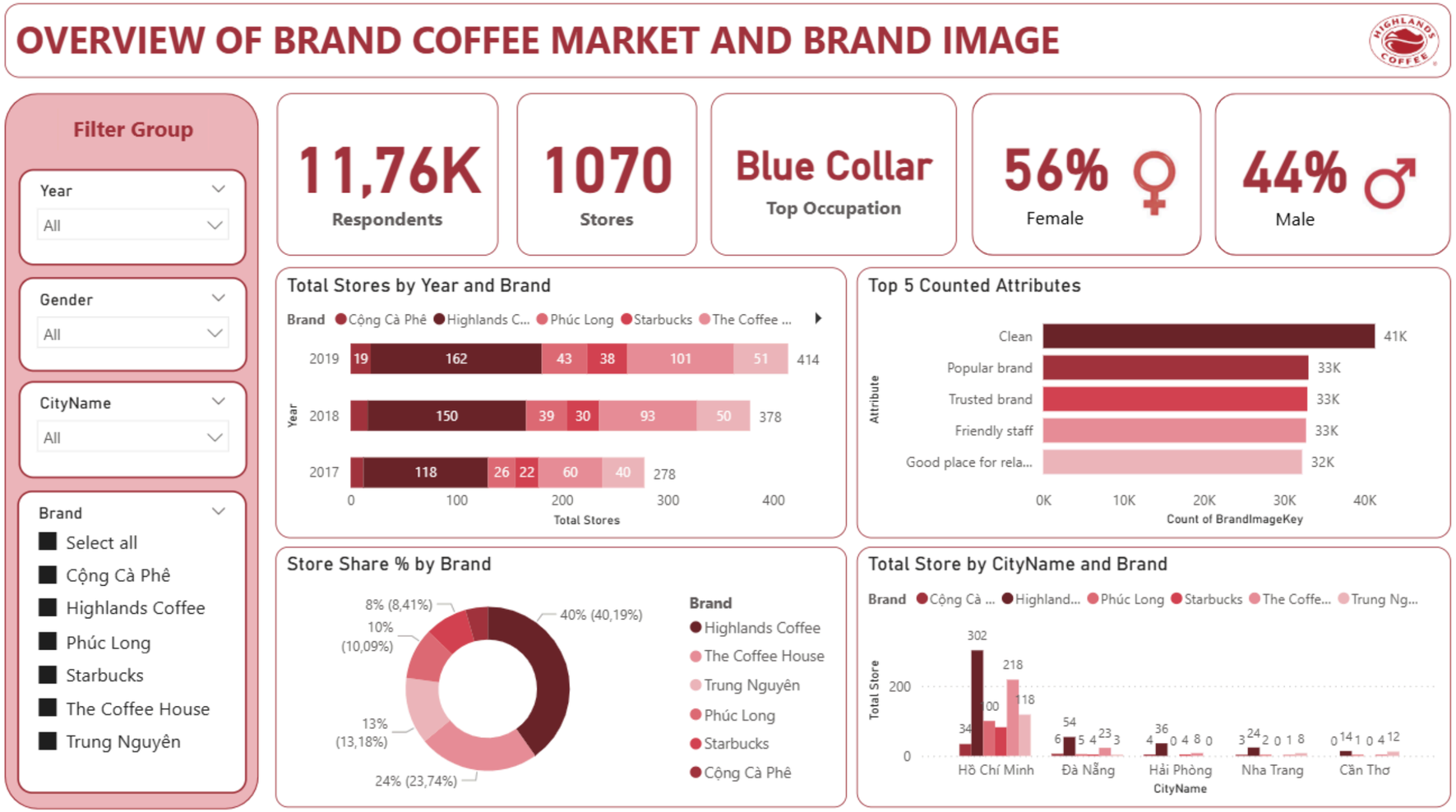
4.1. DASHBOARD AND VISUALIZATION

Customer Potrait



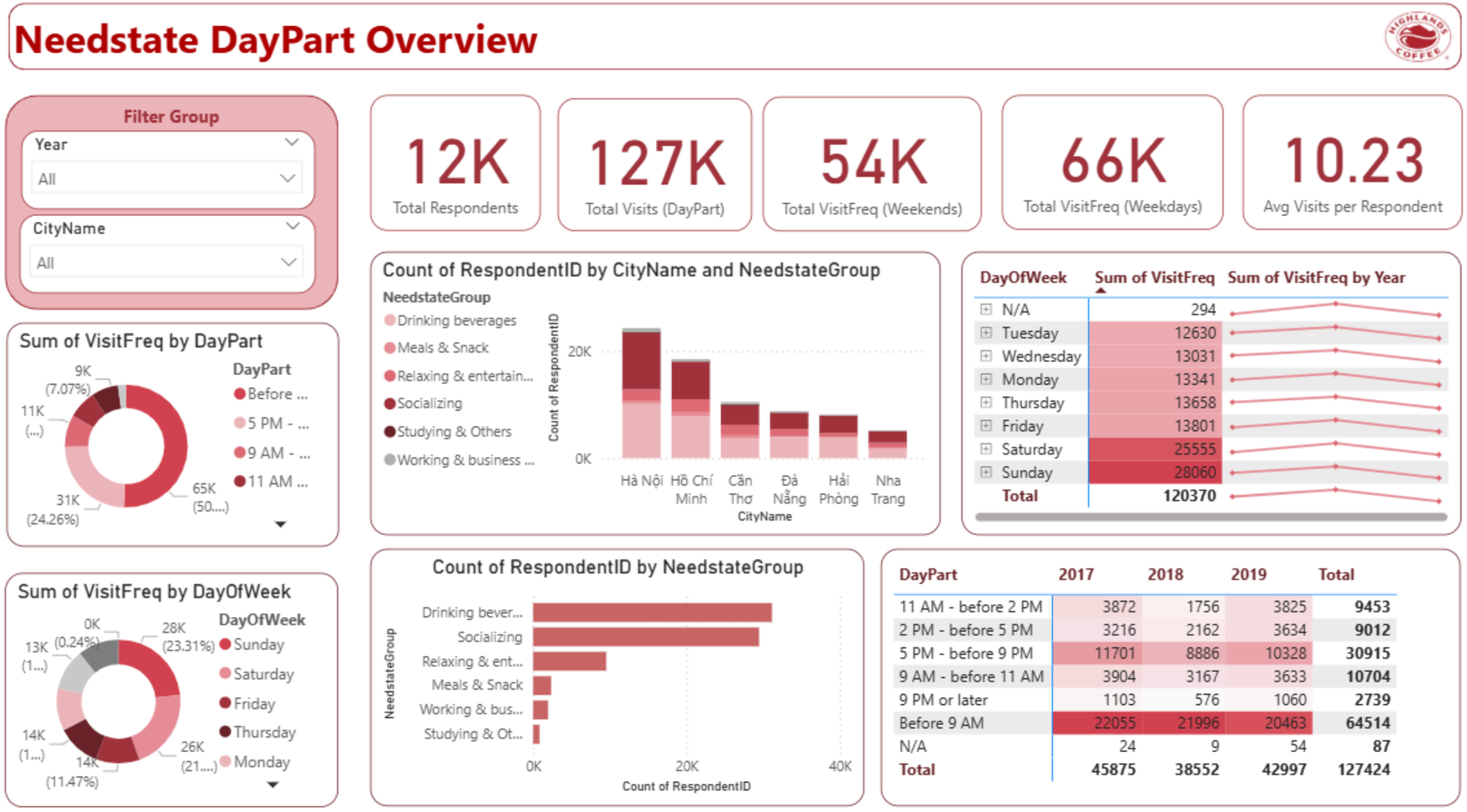
4.1. DASHBOARD AND VISUALIZATION

Brand Image



4.1. DASHBOARD AND VISUALIZATION

NeedstateDayPart Overview



BUSINESS INSIGHT & STRATEGY

Business Insight

- **Market position:** Highlands Coffee leads Vietnam's coffee chain market but faces low customer retention.
- **Customer profile:** Core customers are female, white-collar, aged 20–34
- **Brand image:** "clean, friendly, trustworthy"
- **Customer segmentation:** 5 cluster
- **Market development:** HCMC is saturated
- **Churn drivers:** Brand perception is the strongest churn determinants

Strategy

- **Redesign morning experience**: Launch "Grab & Go", drive-thru, and mobile pre-order services, improve service speed
- **Product Innovation & Unique identity:** Develop a "Highlands Signature" drink line, upgrade quality, and enhance packaging.
- **Personalized Loyalty & Retention:** Expand membership tiers with rewards, birthday perks, and product trials. Use Loyalty App data advocates.

CONCLUSION & FUTURE WORK

Conclusion

- Provides a **comprehensive understanding** of **customer segmentation** and **churn behavior of** Highlands Coffee.
- The results highlight clear distinctions between **high-value** and **high-churn customer** segments and demonstrate that **data-driven customer analytics** can effectively support **strategic decisions** in marketing, retention, and personalized service design, enabling Highlands Coffee
- However, the study still has several limitations. It did not conduct cross-model comparisons, and the predictive accuracy of churn models remains moderate. Moreover, the dataset lacked real-time behavioral variables, limiting the ability to capture dynamic customer changes.

Future work

- Expand datasets with real-time, recency-based, and multi-source data to improve model robustness and segmentation accuracy.
- Enhance model comparison by evaluating multiple clustering (K-means, GMM, DBSCAN) and classification algorithms (XGBoost, Random Forest, Neural Networks) to optimize segmentation and churn prediction.
- Integrate sentiment and behavioral analytics from customer feedback and social media to capture loyalty.

THANK YOU