Delivery Standardization - MSBA Capstone 2025

Code

AUTHOR
Kleyton R. Polzonoff

PUBLISHED
April 18, 2025

Important Note To display the code, click the **"Code"** button in the body of the document or click the **</> Code** button at the top right, then select **"Show All Code."**

1. Business Problem Statement and Objectives

The client, a major beverage supplier, needs a structured system to optimize logistics between its own fleet of Red Trucks and alternative delivery methods (ARTM), which include partner trucks and third-party carriers known as White Trucks. Red Trucks enhance customer relationships and contribute to revenue, while ARTM offers flexibility but limits interaction and control.

To ensure high-quality service and cost efficiency, I will establish clear fleet allocation guidelines based on customer profiles, transaction data, addresses, and delivery costs. This approach will determine the optimal truck type for each customer using a well-defined annual volume threshold. Additionally, customer segmentation will identify shared characteristics, enabling more strategic and data-driven decision-making.

Based on these insights, I will provide actionable recommendations to optimize fleet allocation and enhance operational efficiency.

2. Analytical Approach and Deliveries

The analysis will be conducted separately for two customer groups:

- **All Customers** – The broader customer base, including those who purchase various product types.
- **Local Market Partners Buying Fountain Only** – Customers who purchase only fountain drinks, excluding CO2, cans, or bottles.

To address logistics challenges and transform decisions into data-driven solutions, the approach will combine predictive models with clustering techniques, using both supervised and unsupervised learning methods to build a structured and efficient logistics framework.

Supervised Learning

Supervised learning techniques will be applied to determine whether each customer should be served by Red Trucks (our own fleet) or White Trucks (ARTM), based on a defined set of criteria.

Refining the Fleet Allocation Strategy: Initial Assumptions

As the dataset does not provide predefined fleet allocation criteria, I will establish initial reference points to guide this analysis:

- Annual Volume Threshold: Customers receiving 400 cases and/or gallons per year will initially be assigned to the Red Truck fleet, while those below this threshold will be served by White Trucks.

Unsupervised Learning - Customer Segmentation

A clustering analysis will identify customer groups with similar consumption patterns, refining fleet allocation and enhancing decision-making rules.

Cost Impact Analysis

Considering that the delivery cost of white trucks is five times lower than that of red trucks, different logistics scenarios will be analyzed to compare the costs and strategic impacts of the current approach with the final recommended strategy.

Recommendations

Based on the analysis, data-driven recommendations will be provided to optimize fleet allocation, with a focus on improving service quality, cost efficiency, and strategic decision-making. This approach ensures that each customer receives the most suitable delivery method.

Description of the Data

This project will use four data files provided by the company:

1. customer_profile.csv
2. transactional_data.csv
3. customer_address_and_zip_mapping.csv
4. delivery_cost_data.xlsx

- The customer profile data includes information on all customers they deliver to. This file contains each customer's unique ID along with various categorical variables that describe their location, industry, and delivery preferences.
- The transactional data contains all transactions from all customers with the ordered and delivered amount of product measured in cases and gallons.
- The customer address file only contains two columns – zip code and full address. This can be used in tandem with the customer profile data.
- The delivery cost data maps the cost of delivering a product based on different criteria. This will be used with the transaction data to find the cost of each transaction.

3 Exploratory Data Analysis (EDA) - Part I

This section analyzes the provided data to identify solutions, with a focus on completeness, consistency, and potential issues. Data transformations may include the creation of new variables to improve model

accuracy. Given the large number of variables, the most relevant ones will be prioritized to ensure clarity, while less relevant analyses will be excluded to avoid information overload.

3.1 Loading and Cleaning Datasets

▶ Code

Missing data assessments and any substitutions or modifications will be carried out and will be included in the provided R Markdown file. However, some of these actions will not be displayed in this report to avoid content overload.

Profile Dataset - Cleaning and Adjustments

- The number of unique CUSTOMER_NUMBER in profile_data is greater than in transactional data. This will be addressed later before merging the datasets.
- There are no duplicates or missing values for CUSTOMER_NUMBER.
- Date variables were adjusted to the proper format.
- Logical variables were converted to integers, where 0 represents false and 1 represents true.
- Special characters and extra spaces were removed in factor variables.
- Missing values in the PRIMARY_GROUP_NUMBER field were replaced with zero.
- The CHAIN_MEMBER variable was created to indicate whether the outlet belongs to a chain (has a PRIMARY_GROUP_NUMBER). A value of 1 represents a member, and 0 represents a non-member.

Customer Address Dataset - Cleaning and Adjustments

- The address was split into new columns for each component.
- The dataset does not contain customers' actual addresses but will be used for data aggregation to support customer segmentation. It includes 145 rows with identical geographic coordinates; however, no ZIP codes are duplicated.

Transactional Dataset - Cleaning and Adjustments

- 11,131 null values in the ORDER_TYPE column were replaced with "OTHER."
- The DAYS_AFTER column was added to track the number of days since the transaction, up to February 2, 2025.
- 483 rows with zero values in ORDERED, LOADED, and DELIVERED CASES and GALLONS will be removed from the dataset.
- Negative values in DELIVERED_CASES and DELIVERED_GALLONS have been moved to new columns (RETURNED_CASES and RETURNED_GALLONS), and the original columns were set to zero.
- 30,965 transactions are related to order and/or load but do not have delivery or return data. These will be classified as "order_load" in the DLV_TYPE column.

3.2 Combined Dataset Driven by Transactions

During the exploration, combining all available data was identified as the most effective approach for subsequent analyses. Two files were created: one preserving individual transactions and another compiling information by customer. Both will be used in the exploratory data analysis.

The profile data contains exactly 1801 unique ZIP codes, which were merged with the same number of unique ZIP codes from the customer address dataset. It is important to note that some ZIP codes share the same geographic coordinates, reducing reliability in those cases.

As previously mentioned, the number of unique customer numbers in the profile data (now referred to as full data) is greater than in the transactions dataset. Only customers present in the transactions dataset were included in the merged data.

▶ Code

Below are the first 5 rows and 6 columns of the combined dataset.

▶ Code

```
  CUSTOMER_NUMBER PRIMARY_GROUP_NUMBER FREQUENT_ORDER_TYPE FIRST_DELIVERY_DATE
1       501556470                 376       MYCOKE LEGACY          2024-01-02
2       501556470                 376       MYCOKE LEGACY          2024-01-02
3       501556470                 376       MYCOKE LEGACY          2024-01-02
4       501556470                 376       MYCOKE LEGACY          2024-01-02
5       501556470                 376       MYCOKE LEGACY          2024-01-02
  ON_BOARDING_DATE COLD_DRINK_CHANNEL
1       2023-08-28             DINING
2       2023-08-28             DINING
3       2023-08-28             DINING
4       2023-08-28             DINING
5       2023-08-28             DINING
```

The variable LOCAL_FOUNT_ONLY will be created to identify whether the transaction's customer belongs to the "Local Market Partners Buying Fountain Only" group—customers who purchase only fountain drinks, excluding CO2, cans, or bottles. It will be assigned a value of 1 if the customer belongs to this group and 0 otherwise.

▶ Code

The code below will create a table for an initial overview of the customer types.

▶ Code

Local Market Partners Fountain Only (LFO) - Delivery Quantities Overview

LFO	customers	pct_cust	transactions	pct_trans	qtd_cas	qtd_gal	pct_gal	total_qtd	pct_qtd
0	28,961	95.5	1,013,652	97	26,434,079	9,086,878	94.1	35,520,957	98.4
1	1,359	4.5	31,405	3	0	573,314	5.9	573,314	1.6
Total	30,320	100.0	1,045,057	100	26,434,079	9,660,192	100.0	36,094,271	100.0

▶ Code

Only 4.5% of customers are Local Market Partners who do not purchase CO2 and buy only fountain drinks (LFO = 1), accounting for 3% of transactions. They consumed 5.9% of delivered gallons but represent just 1.9% of the total volume (cases + gallons).

This small group of 1,359 customers includes 83 transactions with positive ordered cases. The last order was placed on December 19, 2024, which would allow for some case deliveries to appear in transactions. Since this didn't occur, these customers will be classified as part of the LFO group, as they consume fountain drinks (gallons), despite ordering cases.

3.3 Combined Dataset Driven by Outlets

The information from the combined transaction dataset (`full_data`) will now be merged by customer and named `full_data_customer`. The goal is to create a unique list of customers who have made transactions. This file will contain a large number of columns and will be used for further analysis.

▶ Code

3.4 Estimated Delivery Costs

The delivery costs will reflect estimated volumes, as they were provided based on the median price within volume ranges and by type of COLD_DRINK_CHANNEL.

▶ Code

The necessary variables will be created to calculate the delivery costs for cases and gallons for the years 2023 and 2024 by customer.

▶ Code

The table below presents the information that constitutes the calculation of the delivery cost per customer.

▶ Code

Show 5 entries Search:

	CUSTOMER_NUMBER	COLD_DRINK_CHANNEL	QTD_DLV_CA_2023	QT
1	501556470	DINING	0	
2	501363456	DINING	0	
3	600075150	DINING	66	
4	500823056	DINING	318	
5	600082383	PUBLIC SECTOR	822	

Showing 1 to 5 of 30,320 entries Previous 1 2 3 4 5 ... 6,064 Next

All costs are being calculated correctly. At this moment, percentage variations for the number of operations, demands, and costs have not been generated because not all customers have a history for 2023 and 2024, which prevents such calculations. However, methods to quantify the growth of each customer will be identified later.

3.5 Target Variables: Initial Assumptions

As initially explained, we will establish classifications related to the target variables to create an initial reference point.

3.5.1 - Demand Threshold and Fleet Assingment

The average annual consumption per customer will be calculated and customers will be classified based on whether they exceed the threshold of 400 units (cases plus gallons).

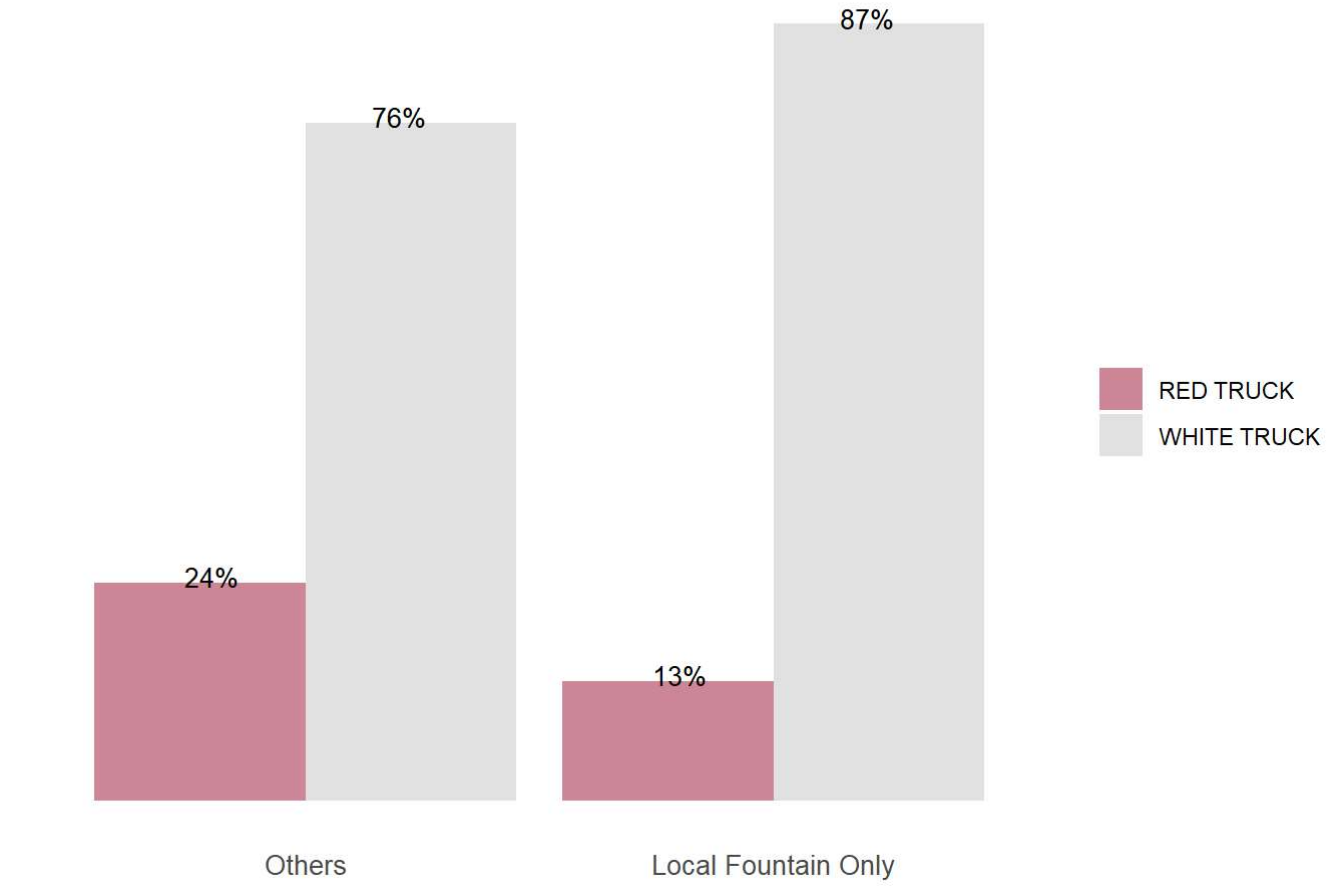
▶ Code

Threshold Reach	Customer Count	Percentage (%)
0	23081	76.1
1	7239	23.9

About 23,081 (76%) of all customers did not reach the threshold of 400 gallons on average per year, while the remaining 7,239 did.

Customers who exceed 400 units annually will be assigned to Red Trucks, while the remaining customers will be allocated to White Trucks.

▶ Code

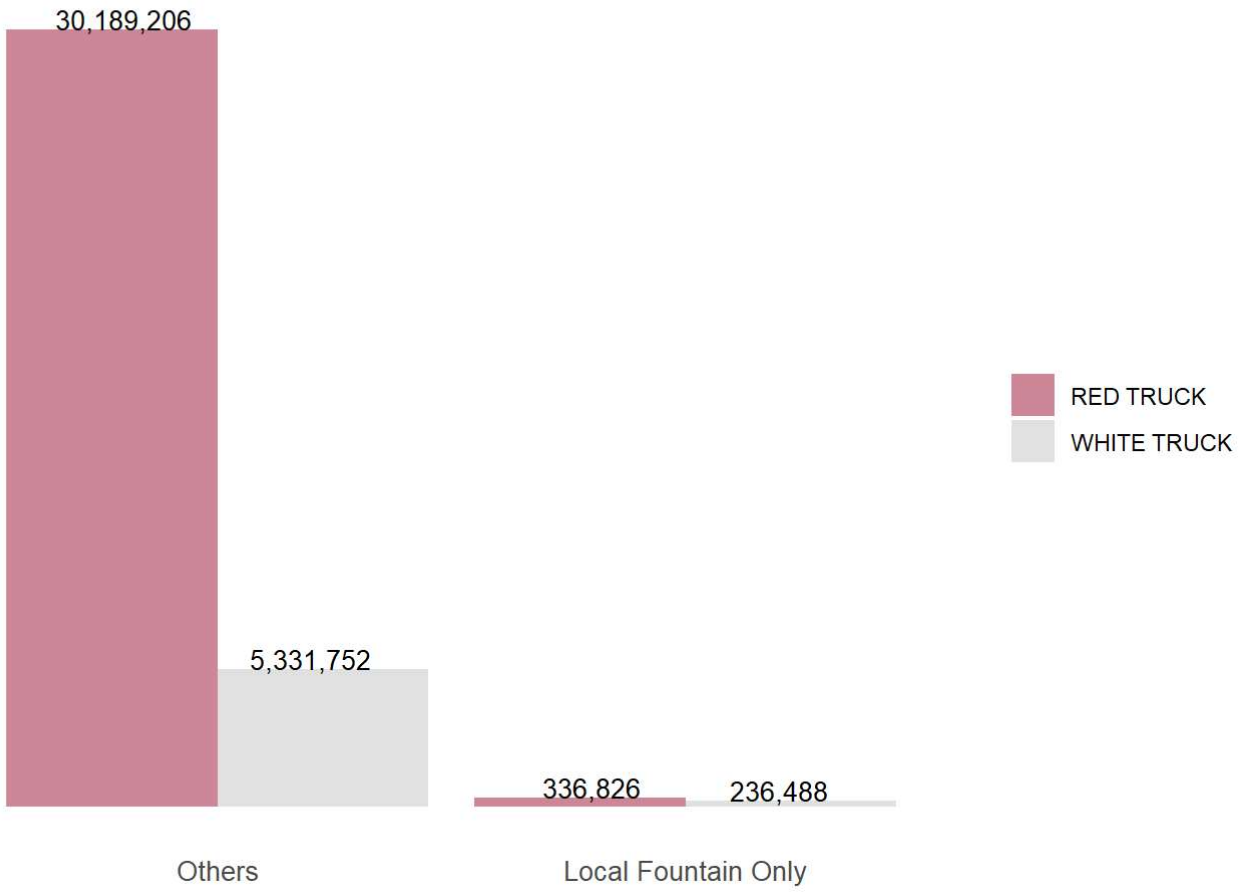


▶ Code

According to these criteria, 13% of Local Fountain Only customers would be assigned to RED TRUCK. Among the other customers, 24% would receive deliveries via RED TRUCK.

▶ Code

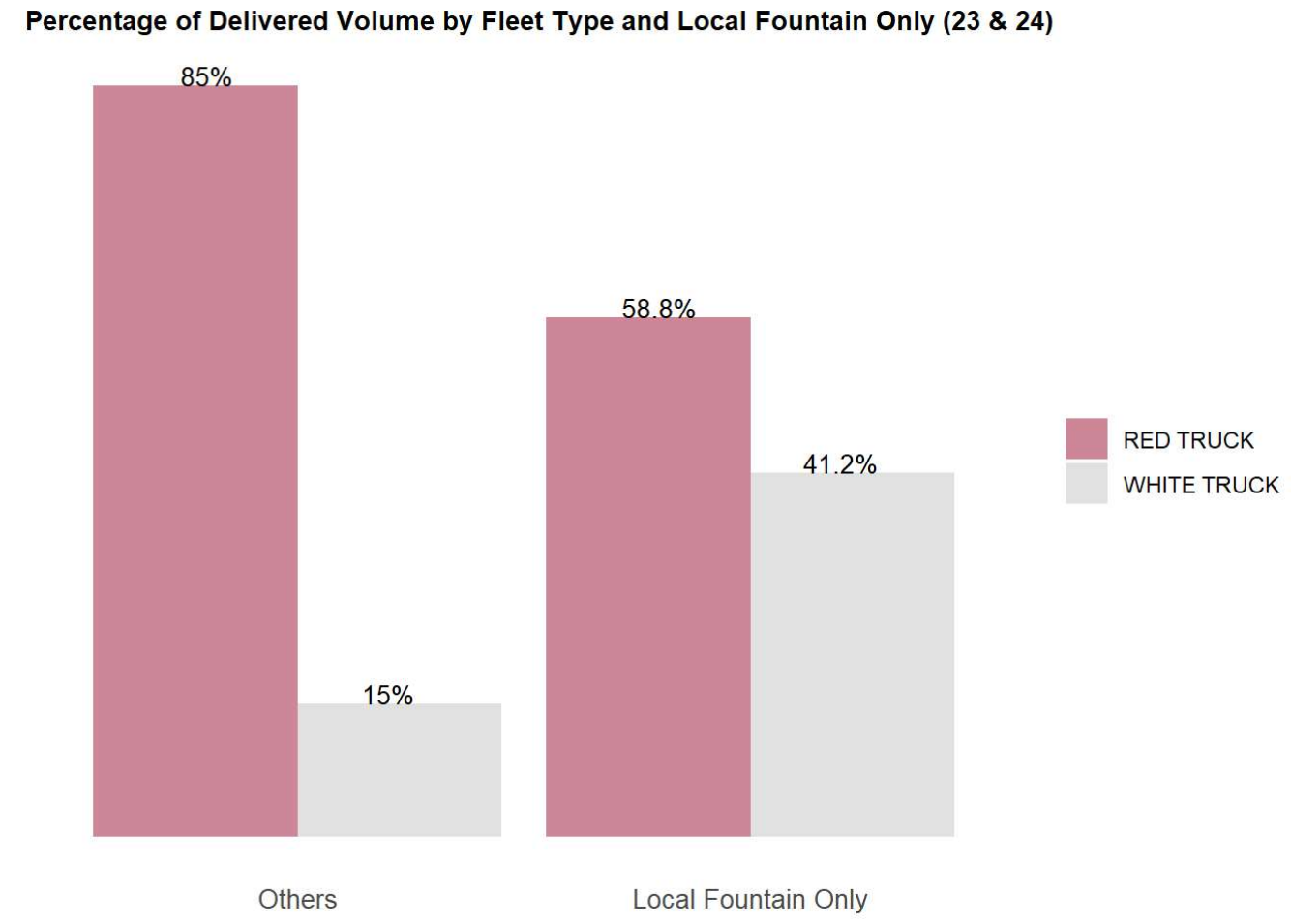
Total Delivered Volume by Fleet Type and Local Fountain Only (23 & 24)



▶ Code

The vast majority of the volume would be delivered by RED TRUCK (85% of the total), with the remaining portion delivered by WHITE TRUCK (15%).

▶ Code



▶ Code

Considering the customer groups independently, nearly 59% of the volume delivered to local partners purchasing fountain only would be transported by RED TRUCKS, while for the remaining customers, almost 85% of the volume would be delivered by RED TRUCKS.

3.6 - Questions and Considerations on Missing Data and Unknown Classes

After the first portion of the EDA, there is a better understanding of the data, but not all questions have been answered. These will continue to be explored in the next section, though some may remain unresolved due to the nature of the questions. The following questions have been identified:

- Based on the available data, what would be a robust statistical approach to calculate the customer growth rate? A simplistic approach was initially used, relying on the average as a reference to visualize the data. However, a more validated method could certainly be applied.
- What is the average load capacity of a Red Truck compared to a White Truck?
- Adding an ID for individual account executives to the customer profile data could be valuable. Is the quality of the account executive a confounding variable when looking at high growth rate customers?
- Does the company set a delivery deadline in days or hours?

4. Exploratory Data Analysis (EDA) - Part II

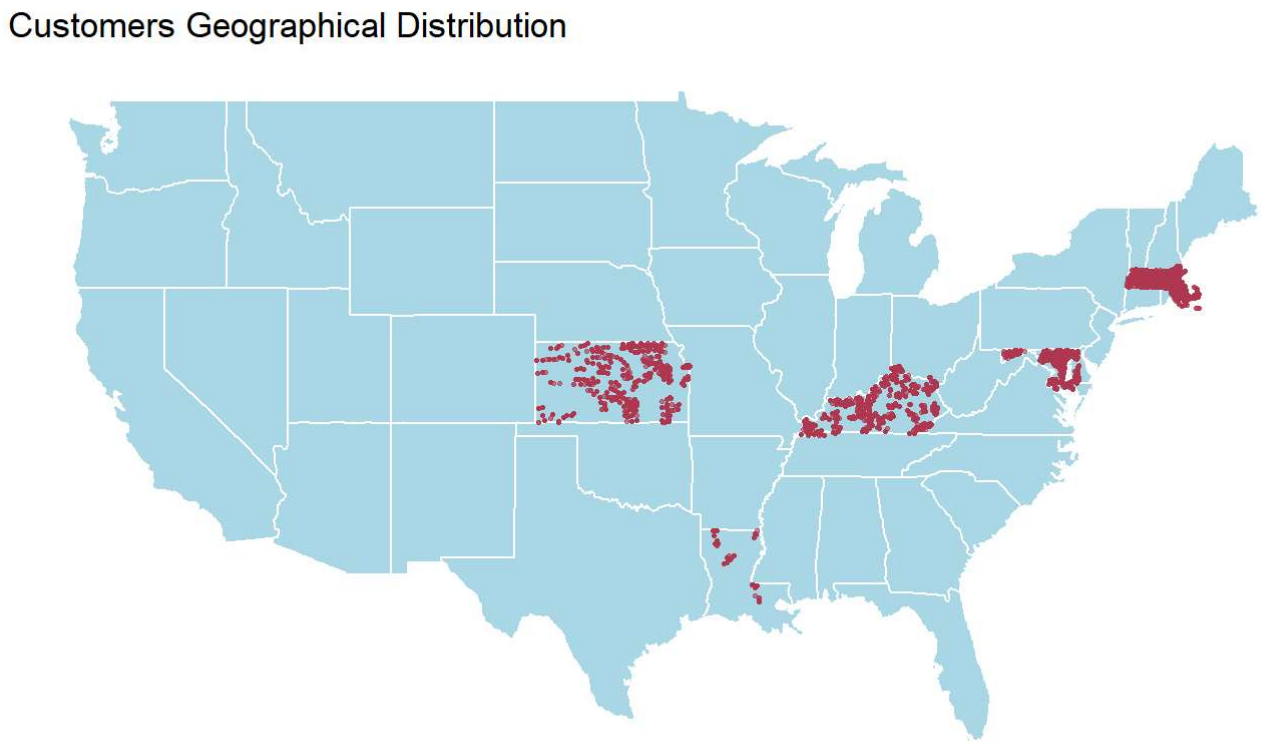
After completing the initial analysis and building the datasets, focusing on the set objectives, we will explore more detailed information about the customers.

4.1 Customers overview

Geographical Distribution of Customers

Although the location data is not real, below you can observe its distribution.

▶ Code



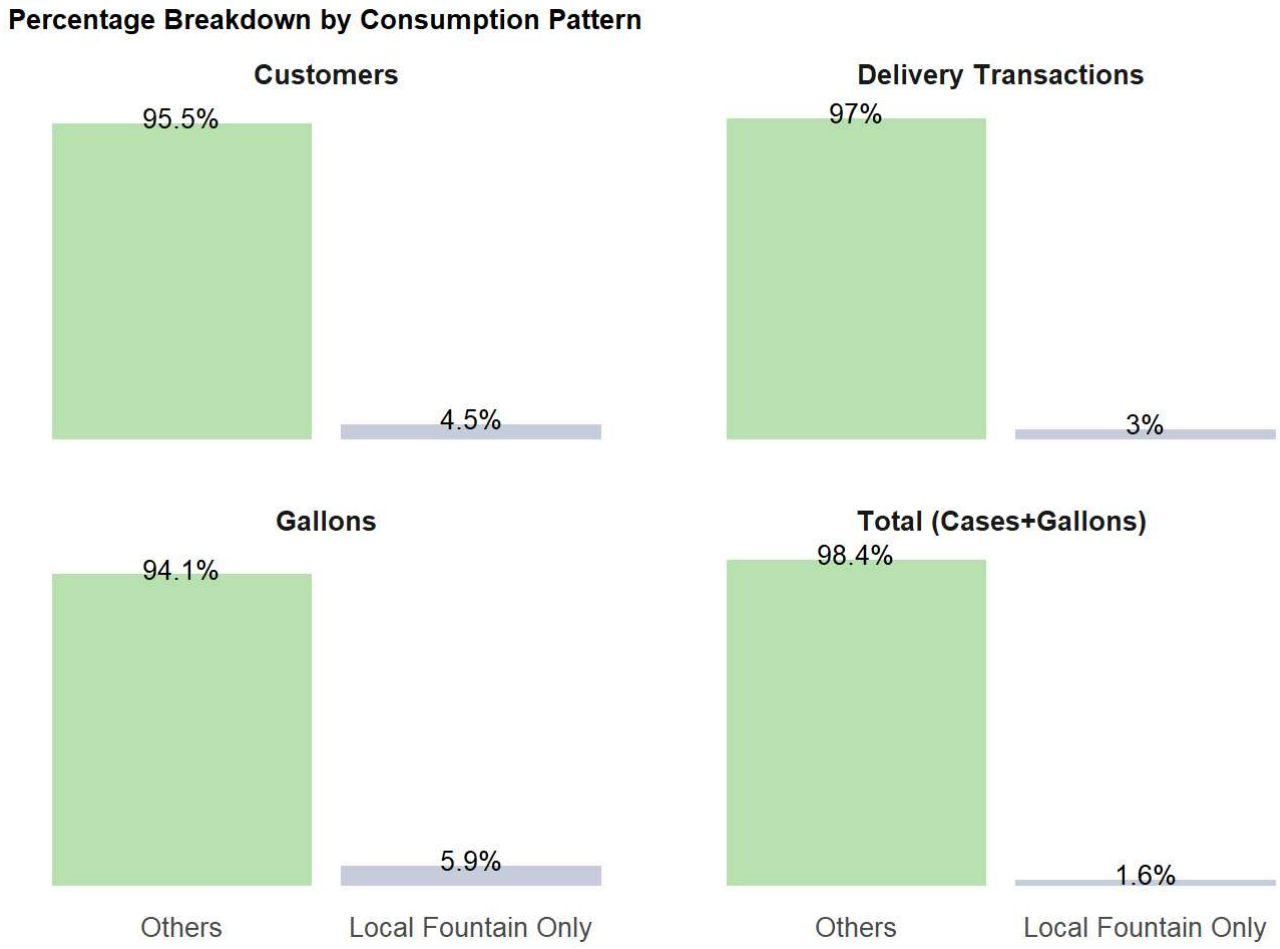
▶ Code

After removing customers who did not make any transactions in 2023 and 2024, there are **30,320 unique customers** who made transactions during these years.
Of these, **18,061 are unique outlets**, while **12,259 belong to 1,020 different chains** that have transacted with the company.

All of their delivery transactions represented a total cost of approximately **$67,907,394**, with an average of **$55.8 per delivery transaction** and **$1.88 per case or gallon delivered**.

4.2 Local Market Partners (Fountain Only)

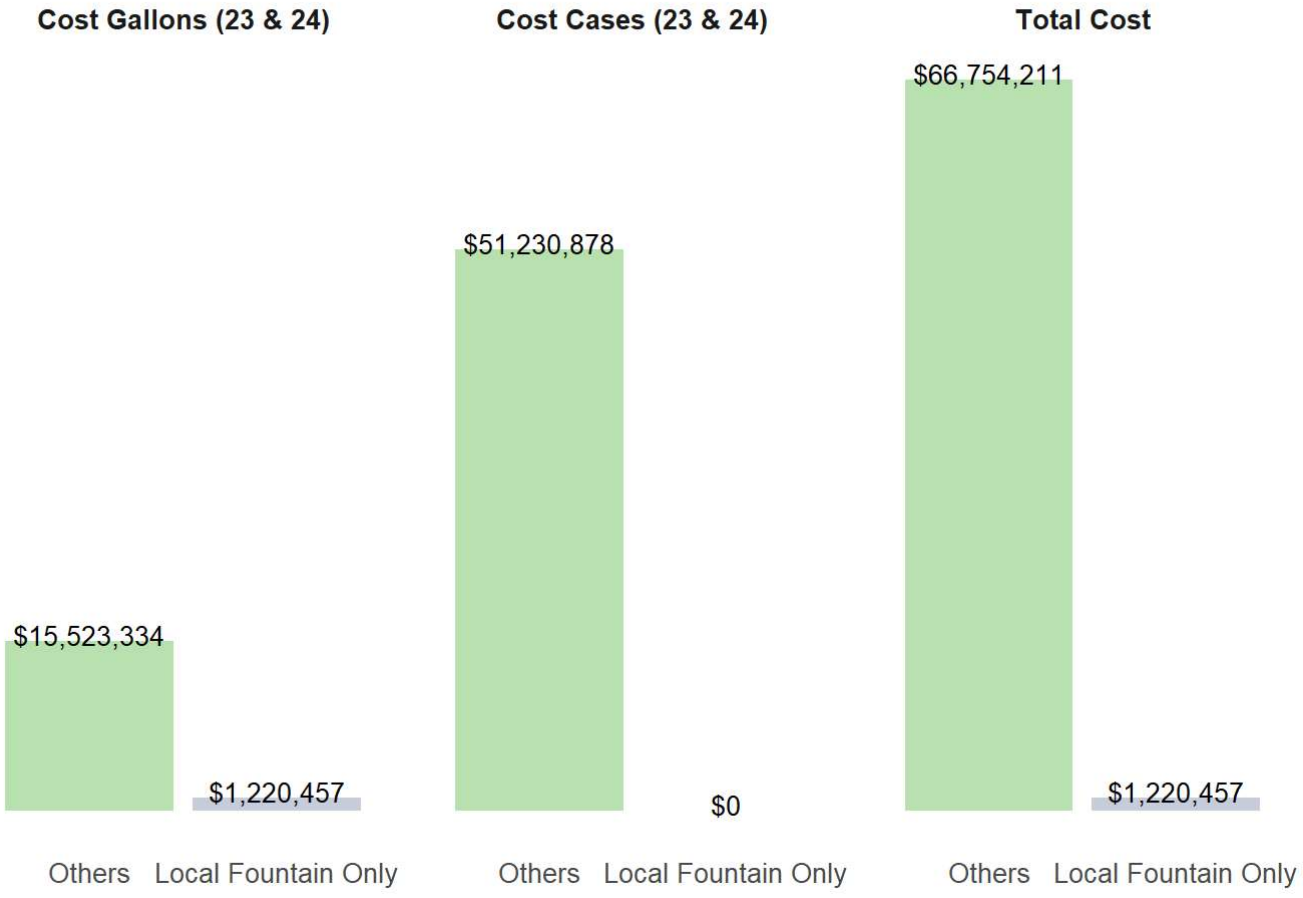
▶ Code



Local market partners who purchase only fountain drinks (Gallons) account for 4.5% of the customers and represent 6% of the company's gallons demand. Their delivery transaction volume is low, contributing only 3%, and the volume delivered accounts for just 1.6% of the total negotiated volume.

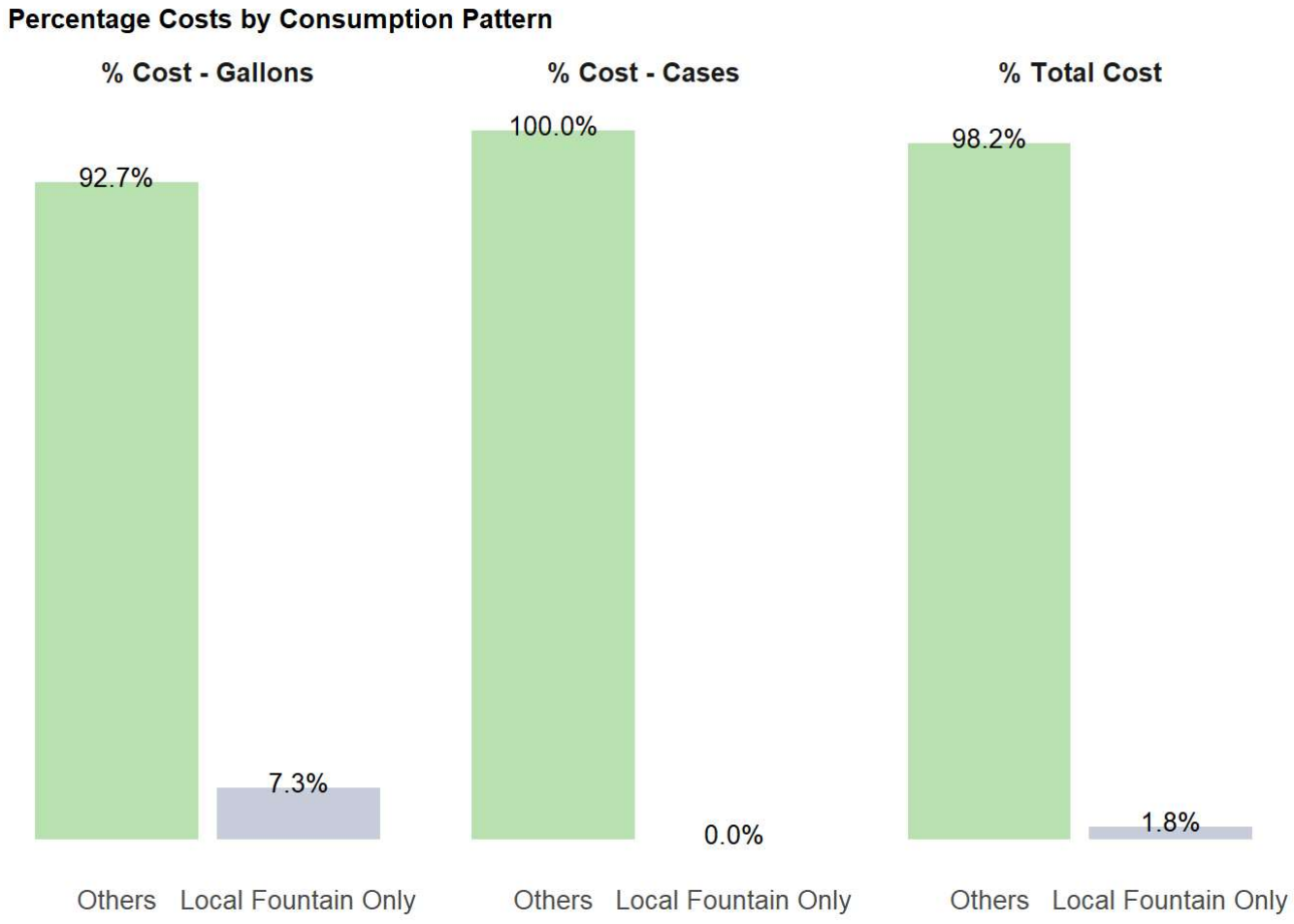
▶ Code

Total Costs by Consumption Pattern



In the years 2023 and 2024, the total delivery cost was 67.9 million, of which only 1.2 million was allocated to local market partners.

▶ Code

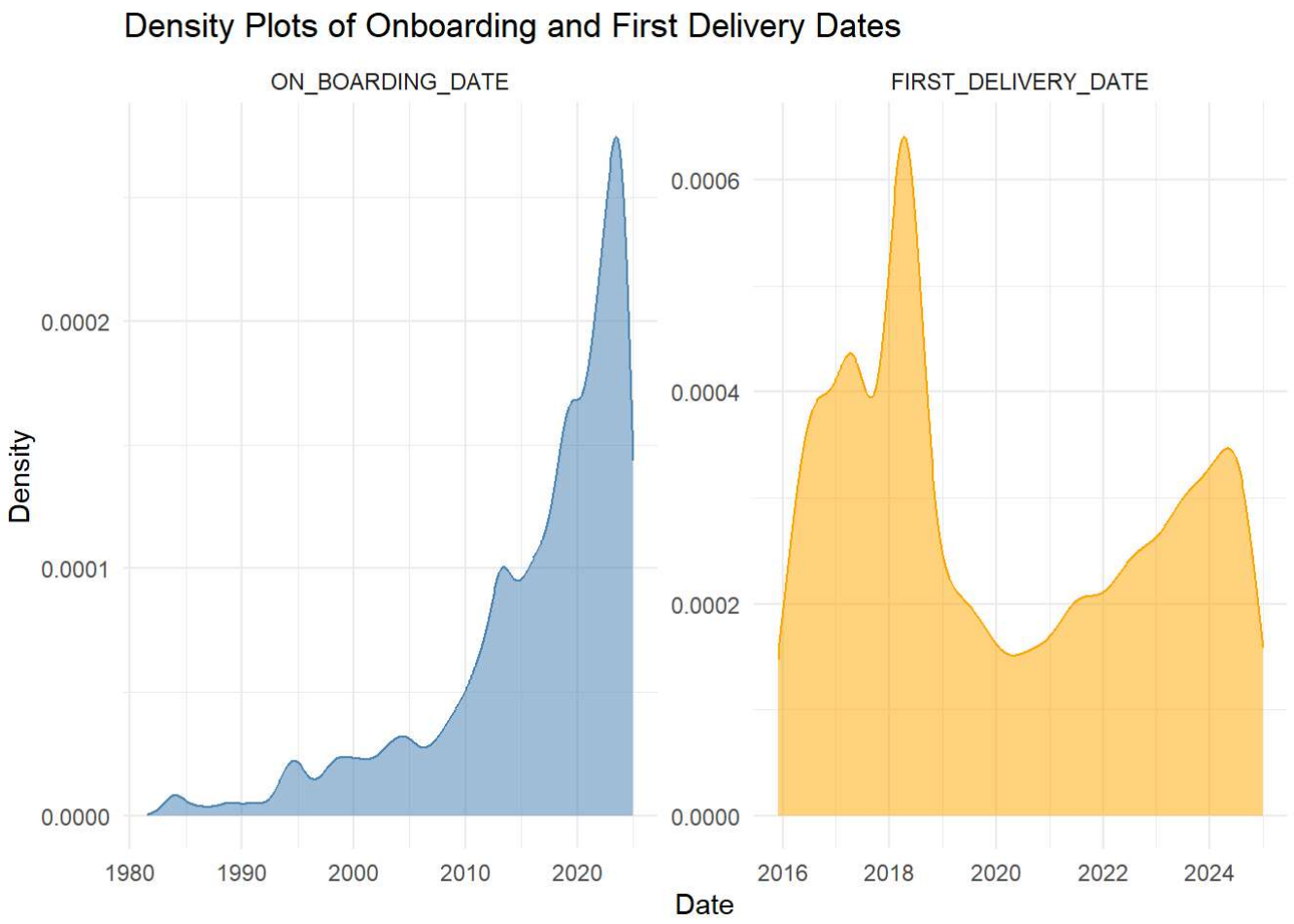


Thus, in 2023 and 2024, the local partners who consume only fountain accounted for 1.8% of the total delivery costs. When we look at their share specifically in gallon deliveries, their participation rises to 7.3%.

4.3 Customers History

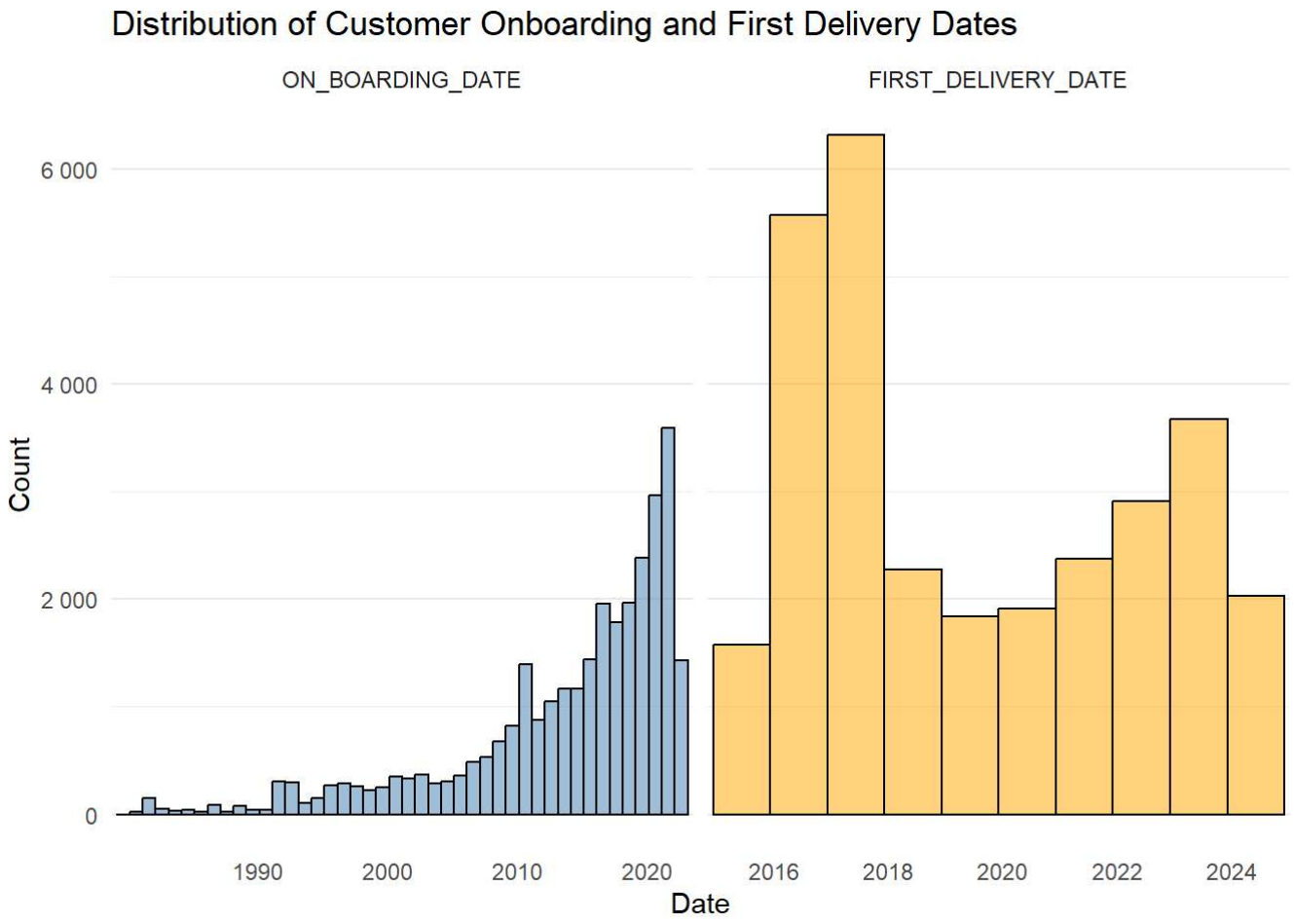
Below is the chart showing the density of customers in relation to the start of their partnership and their first delivery.

▶ Code



The vast majority of customers started to appear after 2010. The figures for the first deliveries show that, since 2016, at least 2,000 customers have received their first delivery each year. There were peaks in 2016 and 2017. In 2024, there was a decrease in the number of customers receiving their first delivery compared to 2023.

▶ Code



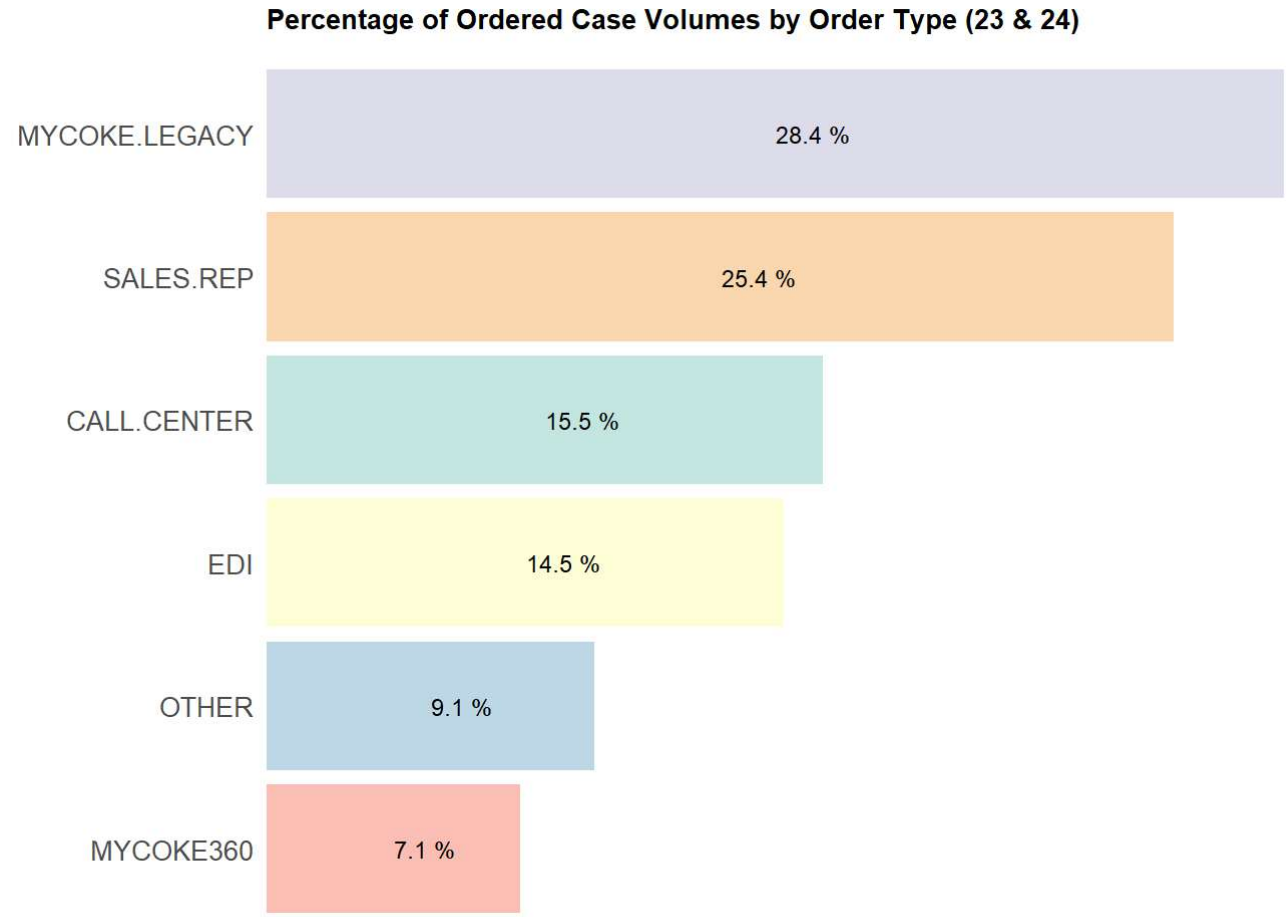
4.4 Order Types

The way orders are placed and by whom is important for understanding customer growth potential. Most customer profiles are associated with sales representatives (65.7%). Other methods follow with 17.6%, and MyCoke 360 accounts for nearly 8%, despite only being launched in Summer 2024 to replace MyCoke Legacy.

However, when analyzing actual transactions from 2023 and 2024, the distribution of order types differs significantly from the customer profiles. For example, sales representatives were responsible for only 27.5% of the orders, not 65.7% as listed in the profiles. Therefore, the analysis will be based on actual transactions rather than profile data.

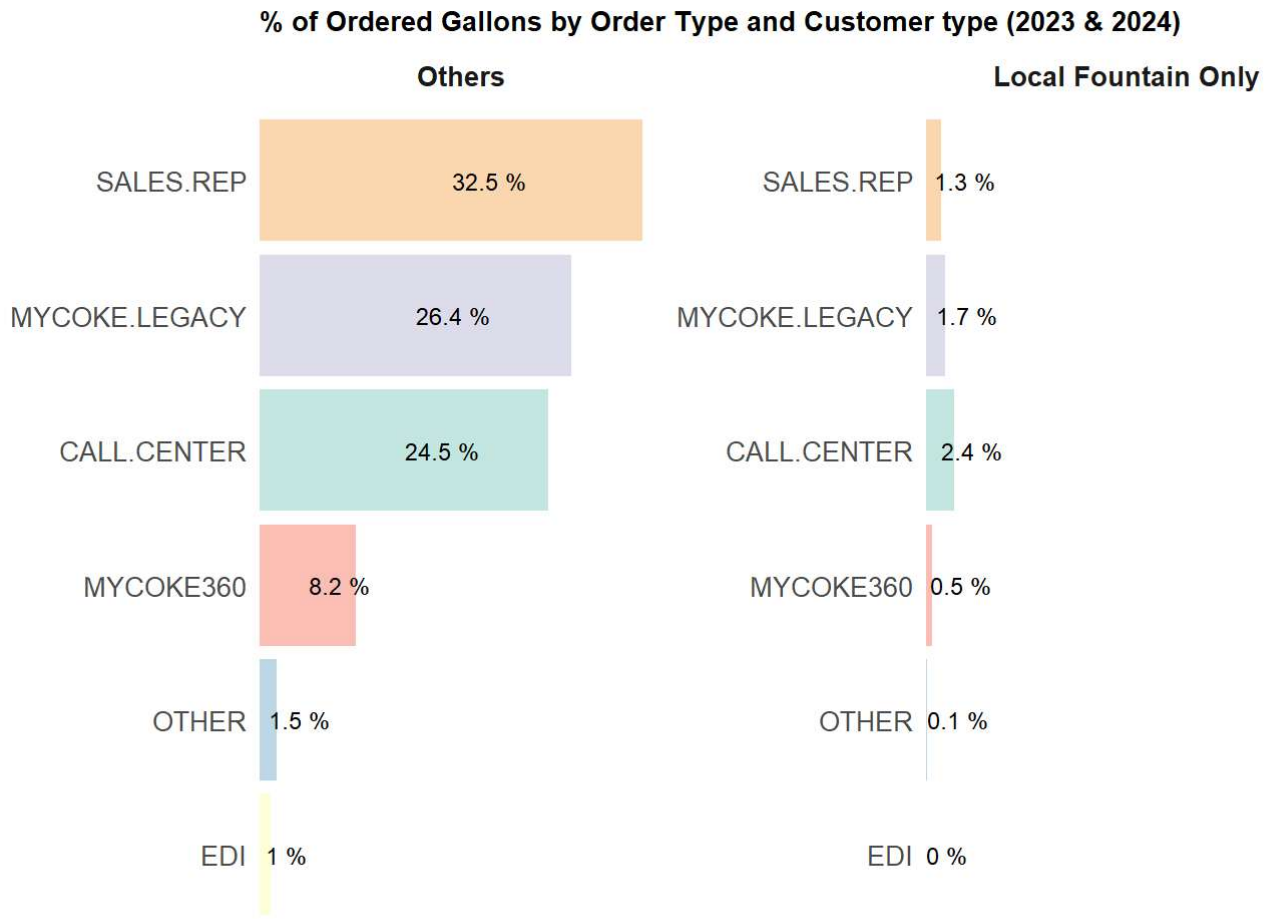
Below are the percentages cases ordered in 2023 and 2024 by order type for each transaction placed in 2023 and 2024.

▶ Code



In the 2023 and 2024 ordered cases transactions, it's clear that the majority of operations were carried out through digital channels, specifically MyCoke Legacy and MyCoke 360, accounting for 35.5%. This was followed by sales representatives with 25.4%, and call centers with 15.5%. MyCoke 360, which was recently launched, makes up 7.1% of the transactions.

▶ Code

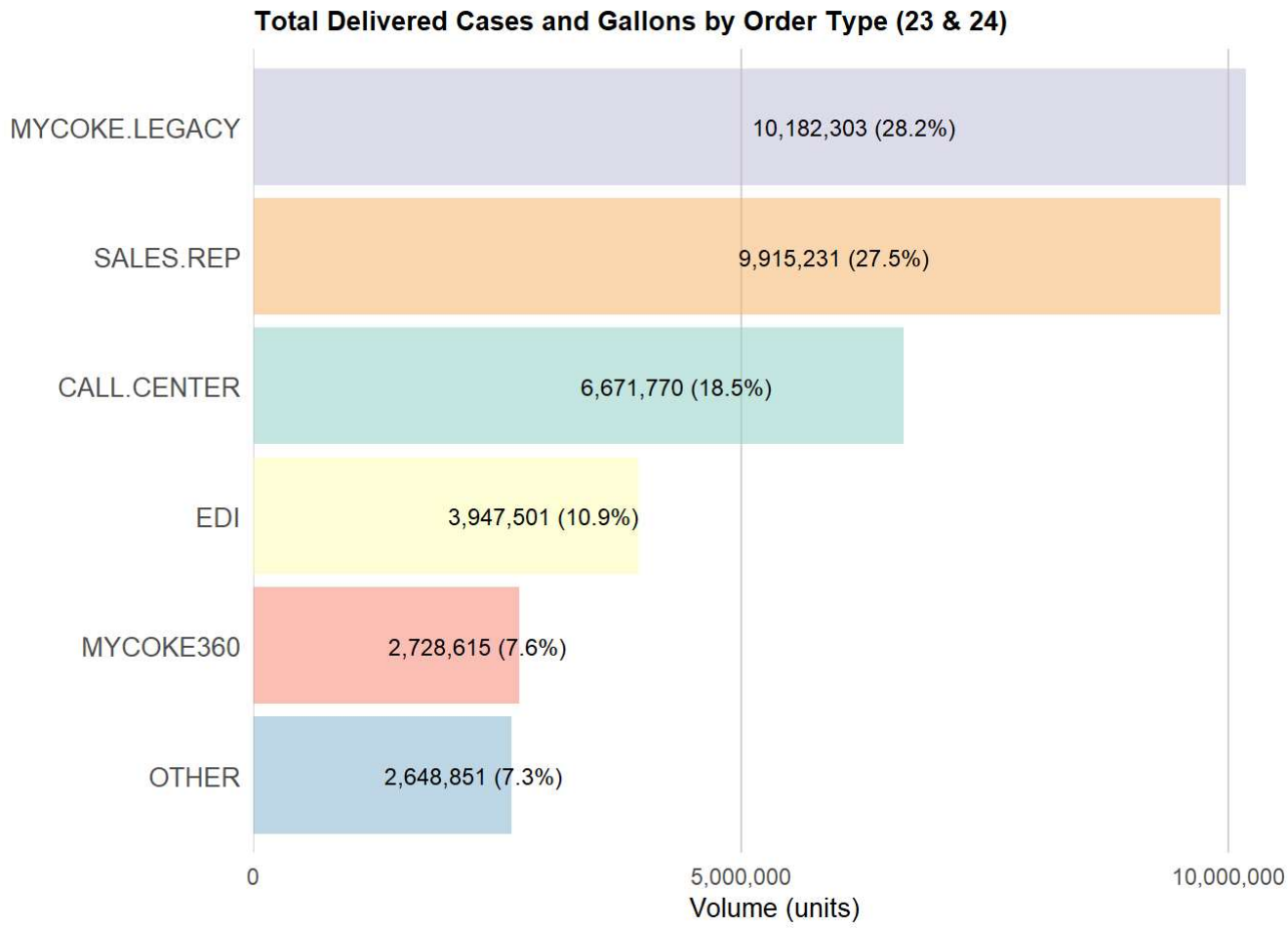


For gallon orders, only a very small fraction (less than 6%) is represented by Local Market Partners that order Fountain Only. For these customers, the majority of their orders are placed via the call center (2.4%), followed by digital channels (2.2%), and finally sales reps (1.3%).

For the remaining customers, digital channels represent 34.6% (MyCoke360 + Legacy), sales reps 32.5%, and call centers 24.5%.

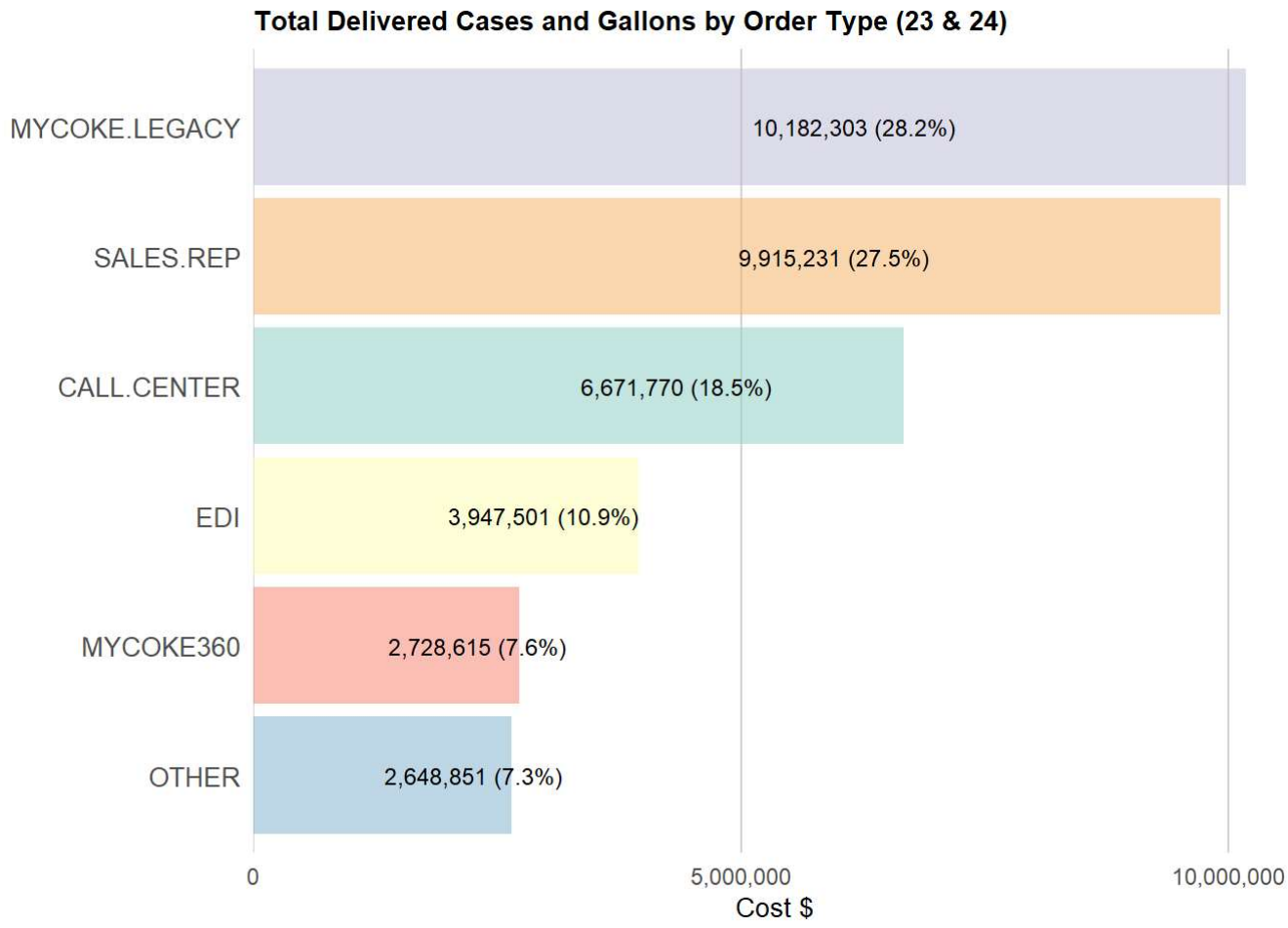
It can be said that digital channels are the most used, accounting for approximately 35% of the total volume of cases and gallons for all customers. Sales reps have a smaller proportional share for case orders but carry more weight for gallon orders.

▶ Code



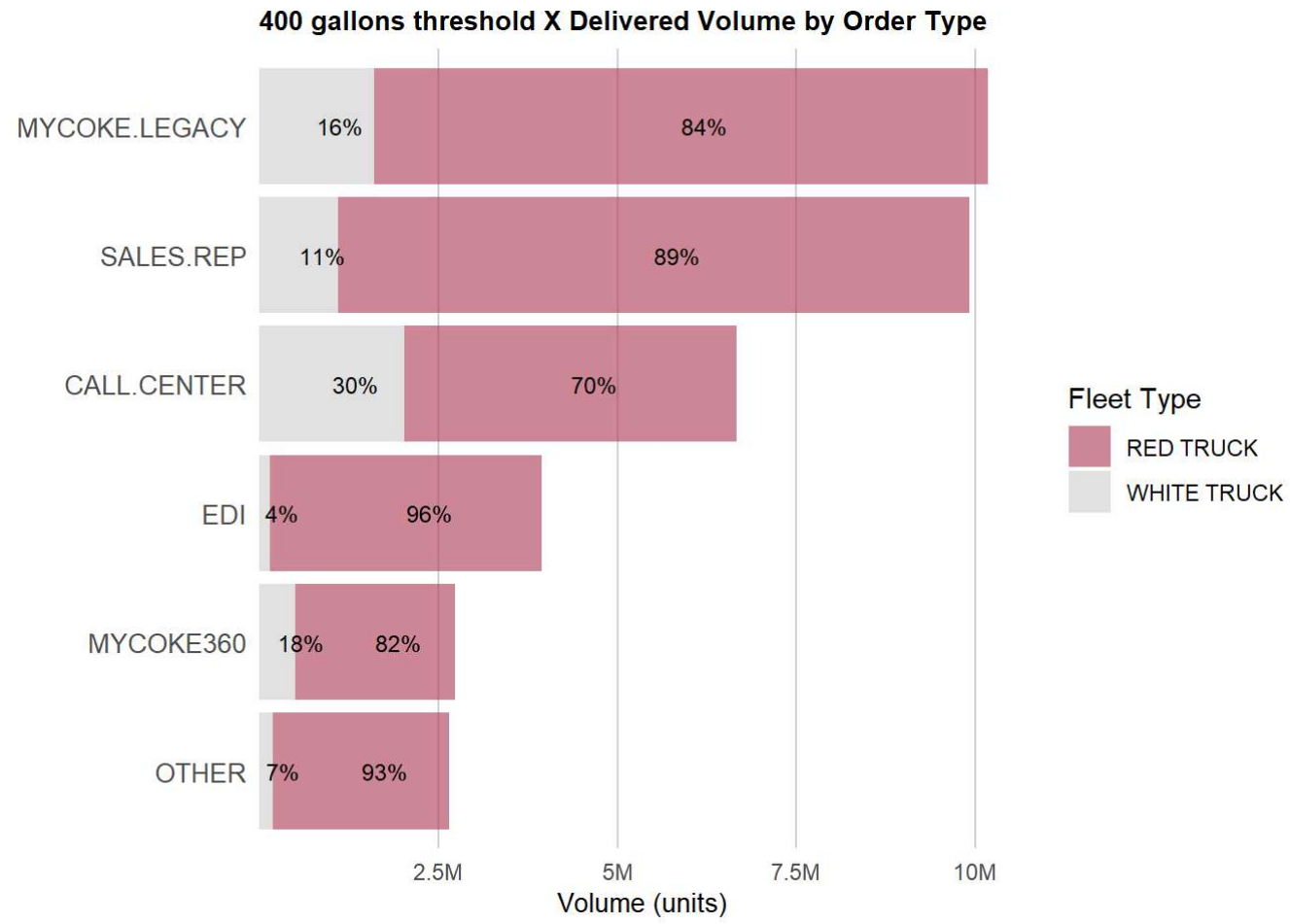
In line with the previous points, digital channels account for nearly 36% of the total volume delivered in 2023 and 2024, followed by sales reps at 27.5% and call centers at 18.5%.

▶ Code



Digital channels account for the majority of the costs, representing 40% of the total delivered cost. Notably, call center costs are slightly higher than sales rep costs, suggesting that their smaller volumes are inflating the costs.

▶ Code



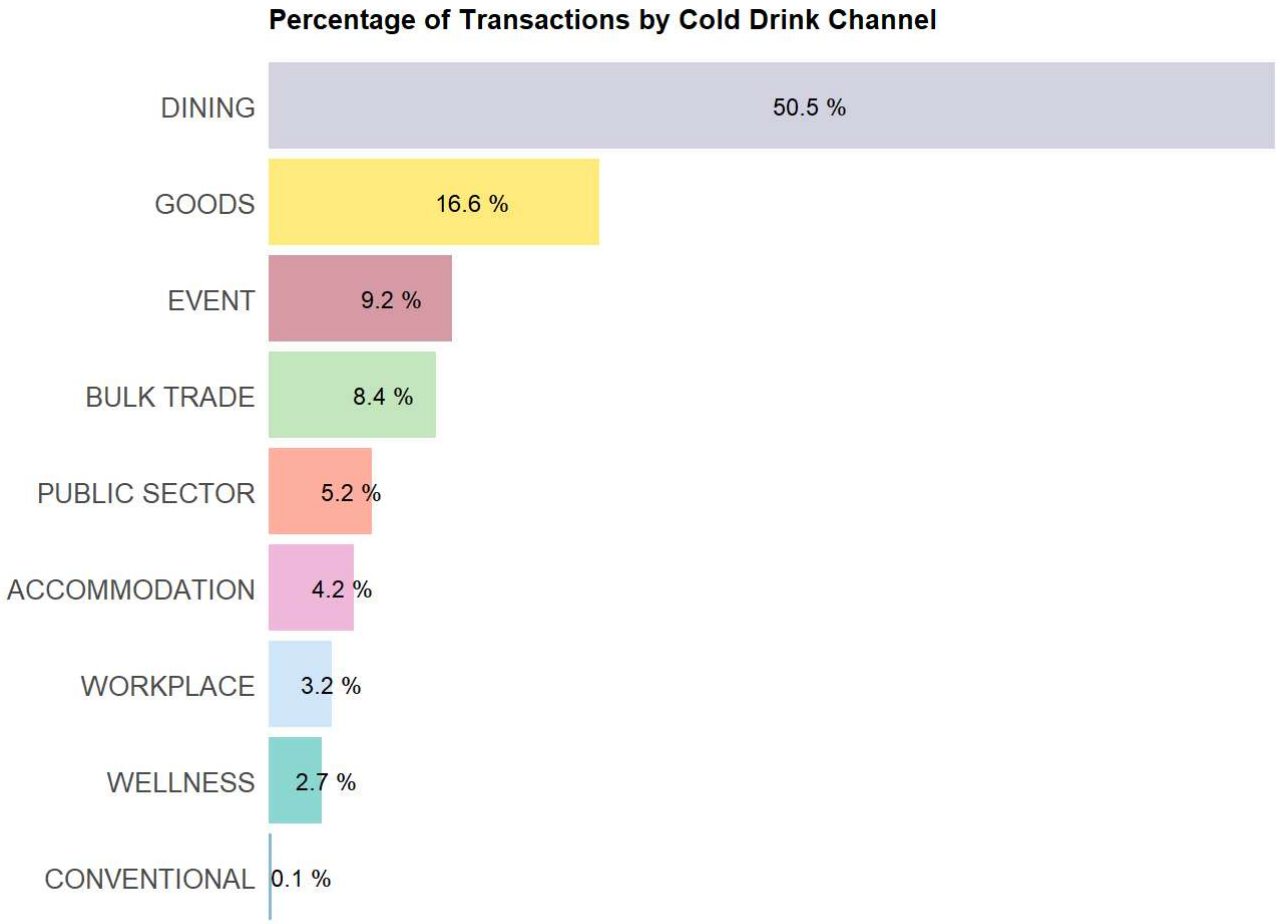
Sales Rep had the highest internal percentage of customers (62%) who would be served by red trucks if the 400-gallon threshold were applied. On the other hand, Call Center showed the highest percentage of customers who would be served by white trucks.

4.5 Channel Types

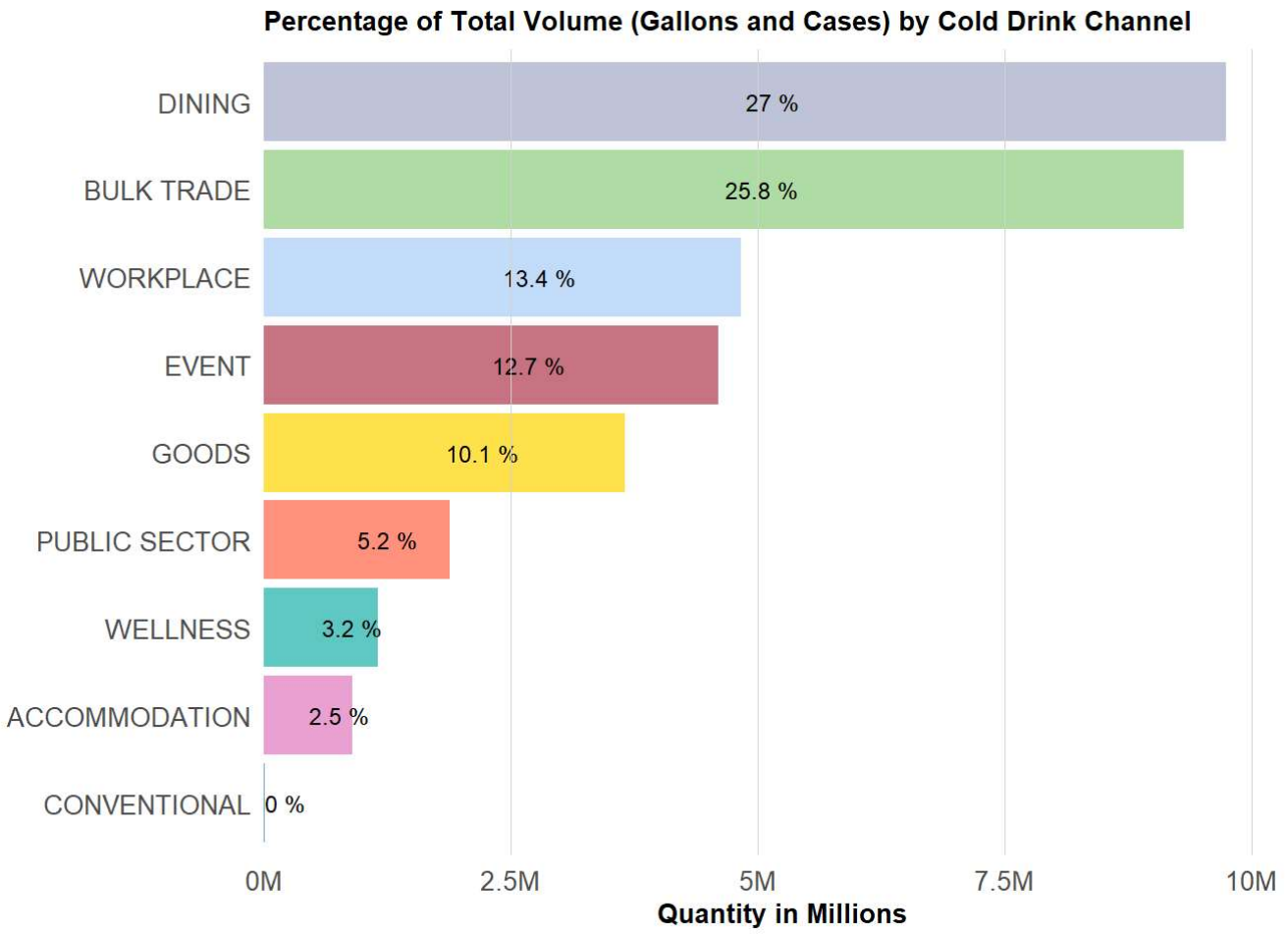
More than 50% of transactions were made through the DINING channel, followed by GOODS (16.6%), EVENTS (9.2%), and BULK TRADE (8.4%). The remaining channels each represent less than 5% of the total.

Transactions for Local Partners Fountain Only are almost entirely concentrated in DINING, with 2.7% of transactions compared to 47.8% for other channels.

▶ Code

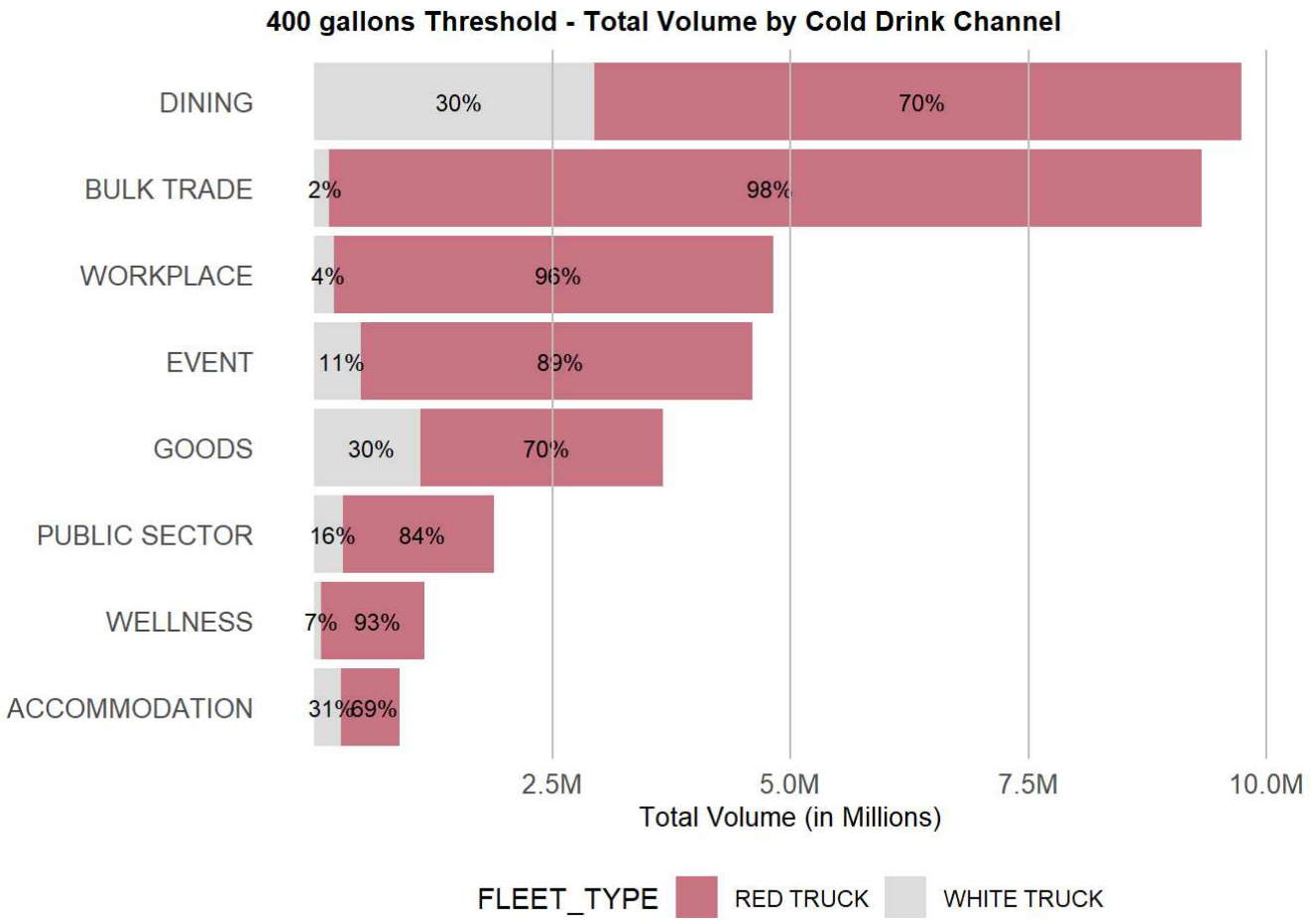

Percentage of Transactions by Cold Drink Channel
DINING
50.5 %
GOODS
16.6 %
EVENT
9.2 %
BULK TRADE
8.4 %
PUBLIC SECTOR
5.2 %
ACCOMMODATION
4.2 %
WORKPLACE
3.2 %
WELLNESS
2.7 %
CONVENTIONAL
0.1 %

▶ Code



Dining was the segment with the highest total consumption, accounting for 27% of the total, followed by Bulk Trade with 25.8% and Workplace with 13.4%. The following section analyzes the information separately by packaging type (cases and gallons) and customer type.

▶ Code



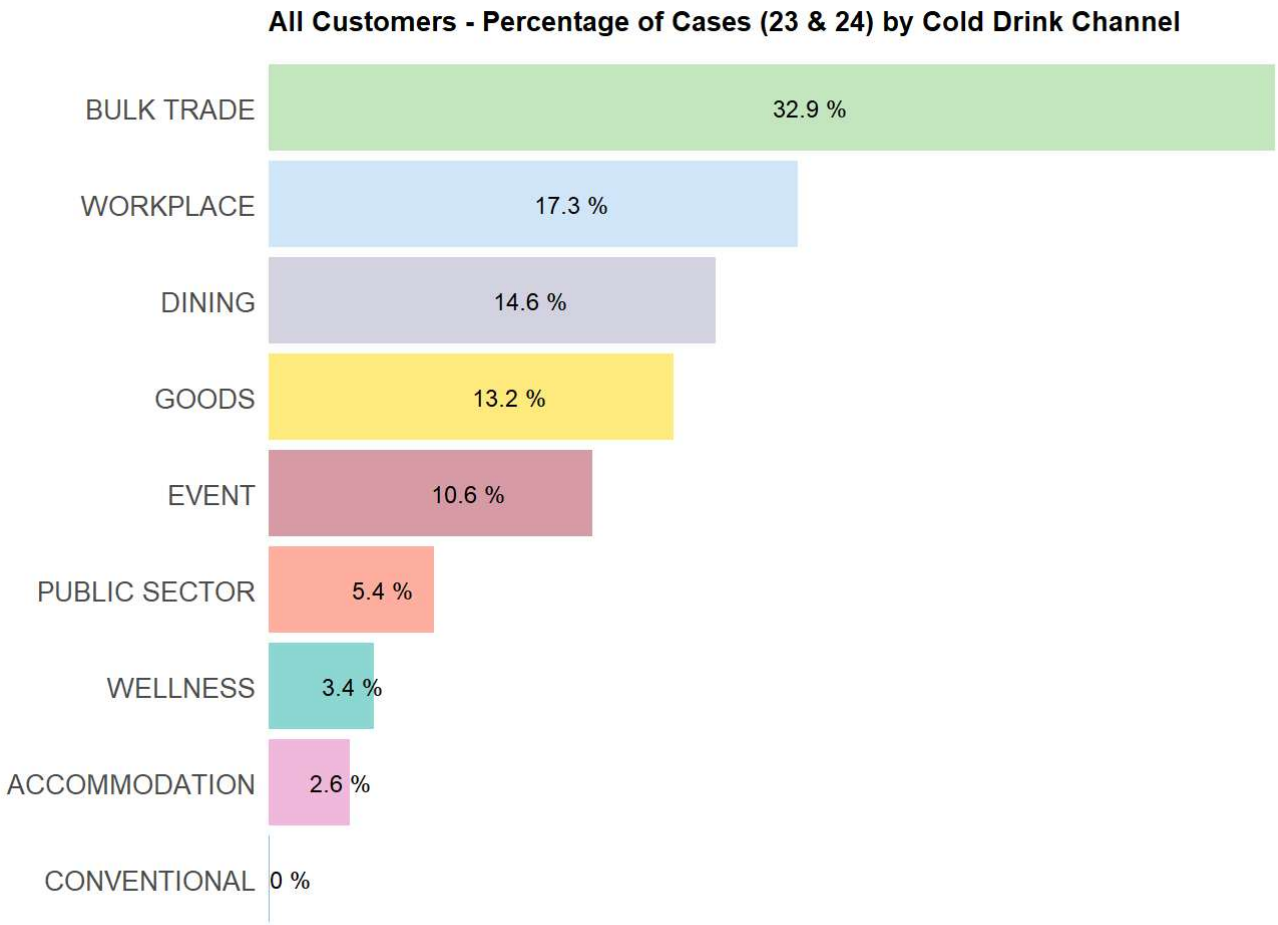
▶ Code

Above are the percentage representations of the volume that would be served by red and white trucks for the 400-gallon threshold. The majority of the volumes would be delivered by red trucks. The "CONVENTIONAL" segment was not displayed due to its extremely low volume, which would overlap with the labels. In this segment, the proportion is 47% for white trucks and 53% for red trucks.

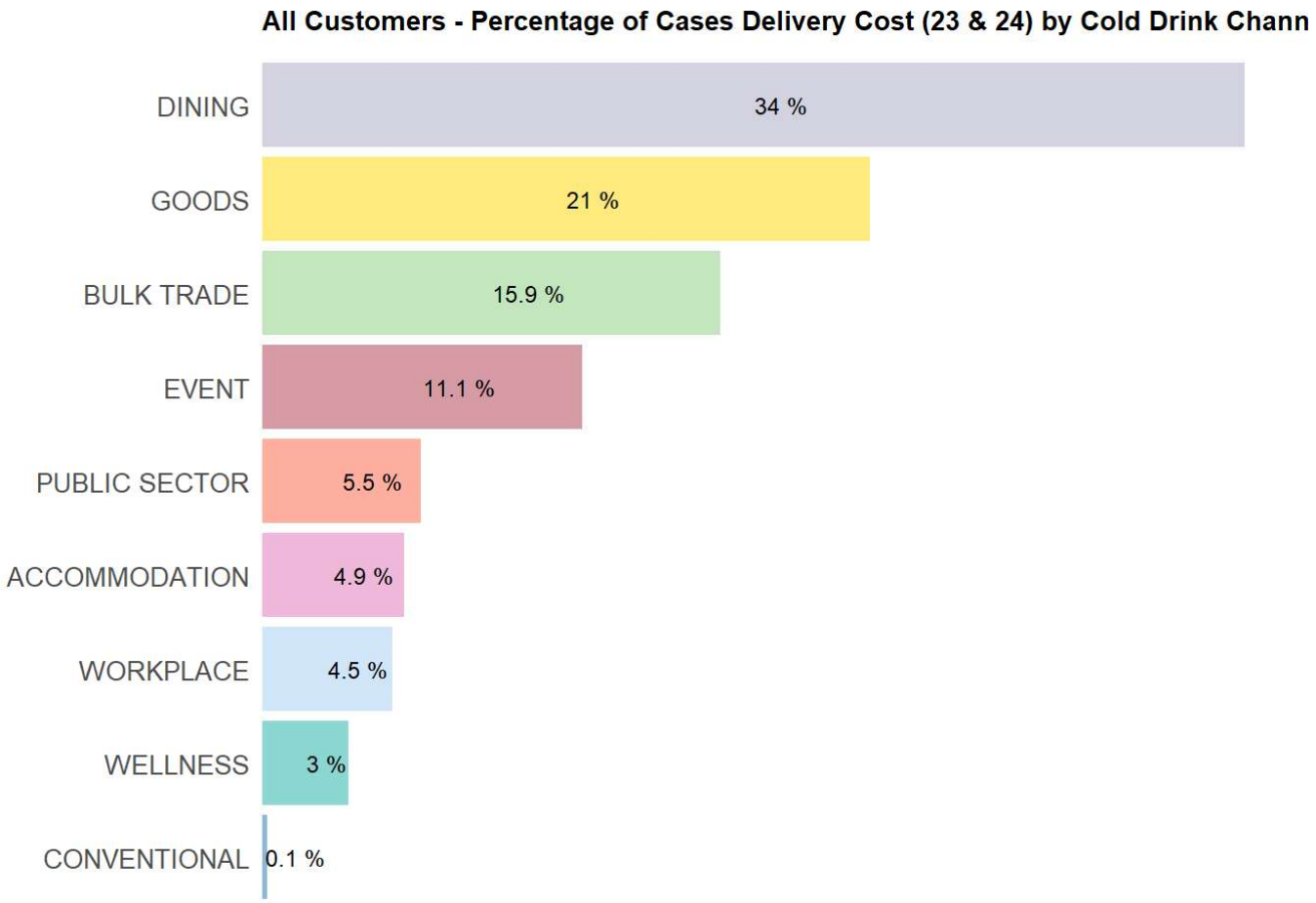
4.5.1 Cold Drink Channel - Delivered Cases for All Customers

Below are the percentages of cases delivered in 2023 and 2024 for all customers by cold drink channel.

▶ Code


All Customers - Percentage of Cases (23 & 24) by Cold Drink Channel
BULK TRADE
32.9 %
WORKPLACE
17.3 %
DINING
14.6 %
GOODS
13.2 %
EVENT
10.6 %
PUBLIC SECTOR
5.4 %
WELLNESS
3.4 %
ACCOMMODATION
2.6 %
CONVENTIONAL
0 %

▶ Code



The main segment receiving cases (bottles, cans, etc.) was Bulk Trade with 33%, followed by Workplace with 17%, and Dining with 14.6%. On the other hand, the segment that presented the highest delivery costs for cases was Dining, accounting for 34% of the cost in 2023 and 2024, followed by Goods at 21%, and Bulk Trade at 16%.

The tables below aim to provide detailed information within this group regarding the number of customers, costs, quartile divisions, and other relevant factors.

▶ Code

CASES (23 & 24) - Deliveries by Cold Drink Channel - All Customers

Channel	T.Cases	Cases %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cust $
BULK TRADE	8,687,959	32.9	8,127,990	1,278	5.3	6,798	1,239	0.94
WORKPLACE	4,567,596	17.3	2,299,625	712	2.9	6,415	164	0.50
DINING	3,859,778	14.6	17,429,159	10,929	45.2	353	82	4.52
GOODS	3,494,064	13.2	10,780,042	5,542	22.9	630	205	3.09
EVENT	2,796,241	10.6	5,677,840	2,785	11.5	1,004	230	2.03
PUBLIC SECTOR	1,422,915	5.4	2,805,085	1,411	5.8	1,008	244	1.97
WELLNESS	903,700	3.4	1,529,502	340	1.4	2,658	532	1.69

CASES (23 & 24) - Deliveries by Cold Drink Channel - All Customers

Channel	T.Cases	Cases %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cust $
ACCOMMODATION	695,490	2.6	2,507,698	1,150	4.8	605	326	3.61
CONVENTIONAL	6,337	0.0	73,937	53	0.2	120	64	11.67

▶ Code

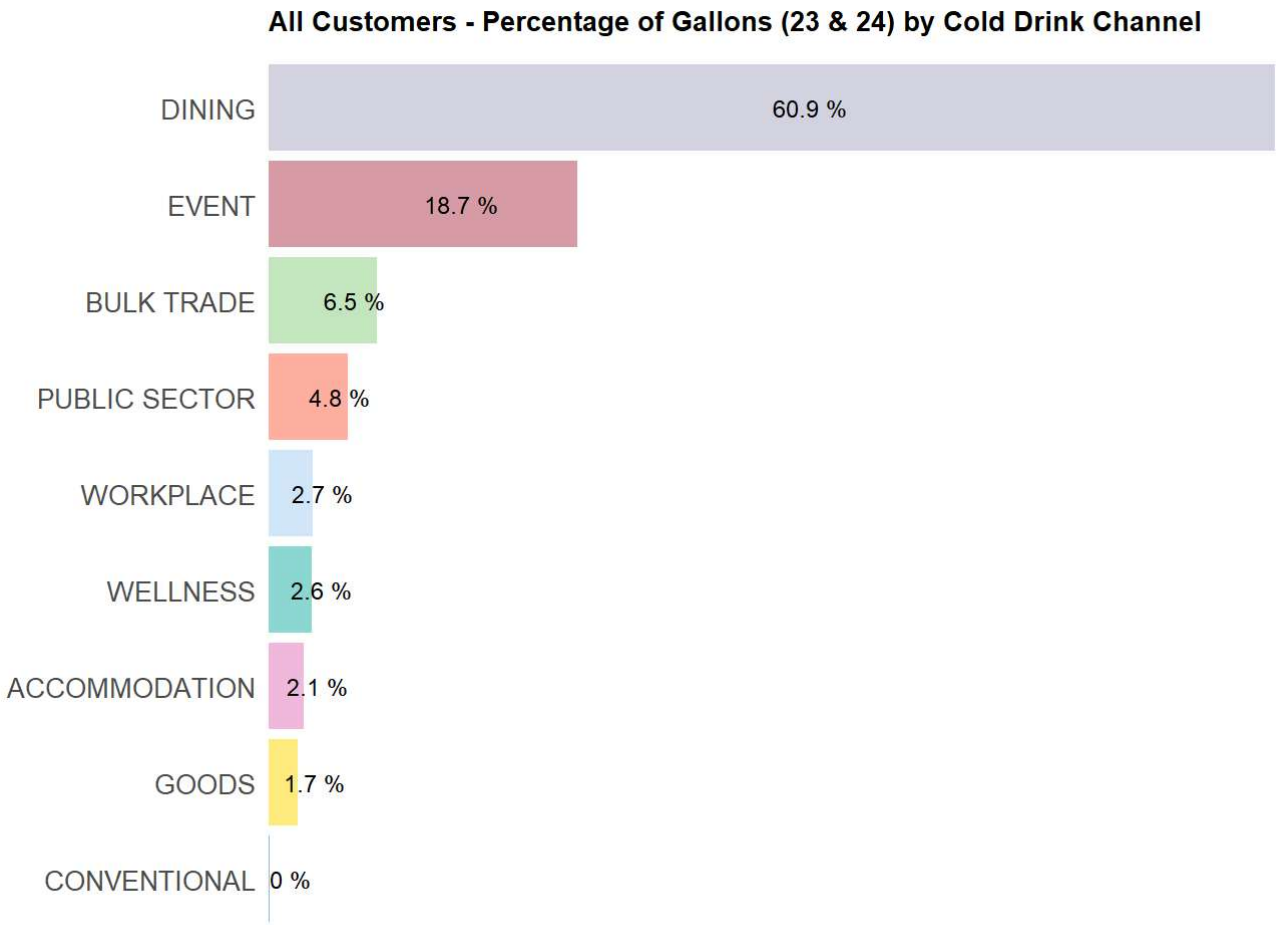
CASES (23 & 24) - Quartile Analysis by Cold Drink Channel - A

Channel	Avg.Qtd.Cust	Median.Qtd.Cust	1Quart.Qtd	2Quart.Qtd	3Quart.Qtd	1Quart.Vol	1Q.Vol%
BULK TRADE	6,798	1,239	384	1,239	4,162	53,070	0.6
WORKPLACE	6,415	164	40	164	546	2,622	0.1
WELLNESS	2,658	532	101	532	2,400	3,224	0.4
PUBLIC SECTOR	1,008	244	68	244	760	10,971	0.8
EVENT	1,004	230	56	230	753	17,174	0.6
GOODS	630	205	98	205	466	68,717	2.0
ACCOMMODATION	605	326	99	326	668	12,725	1.8
DINING	353	82	16	82	318	17,073	0.4
CONVENTIONAL	120	64	26	64	138	228	3.6

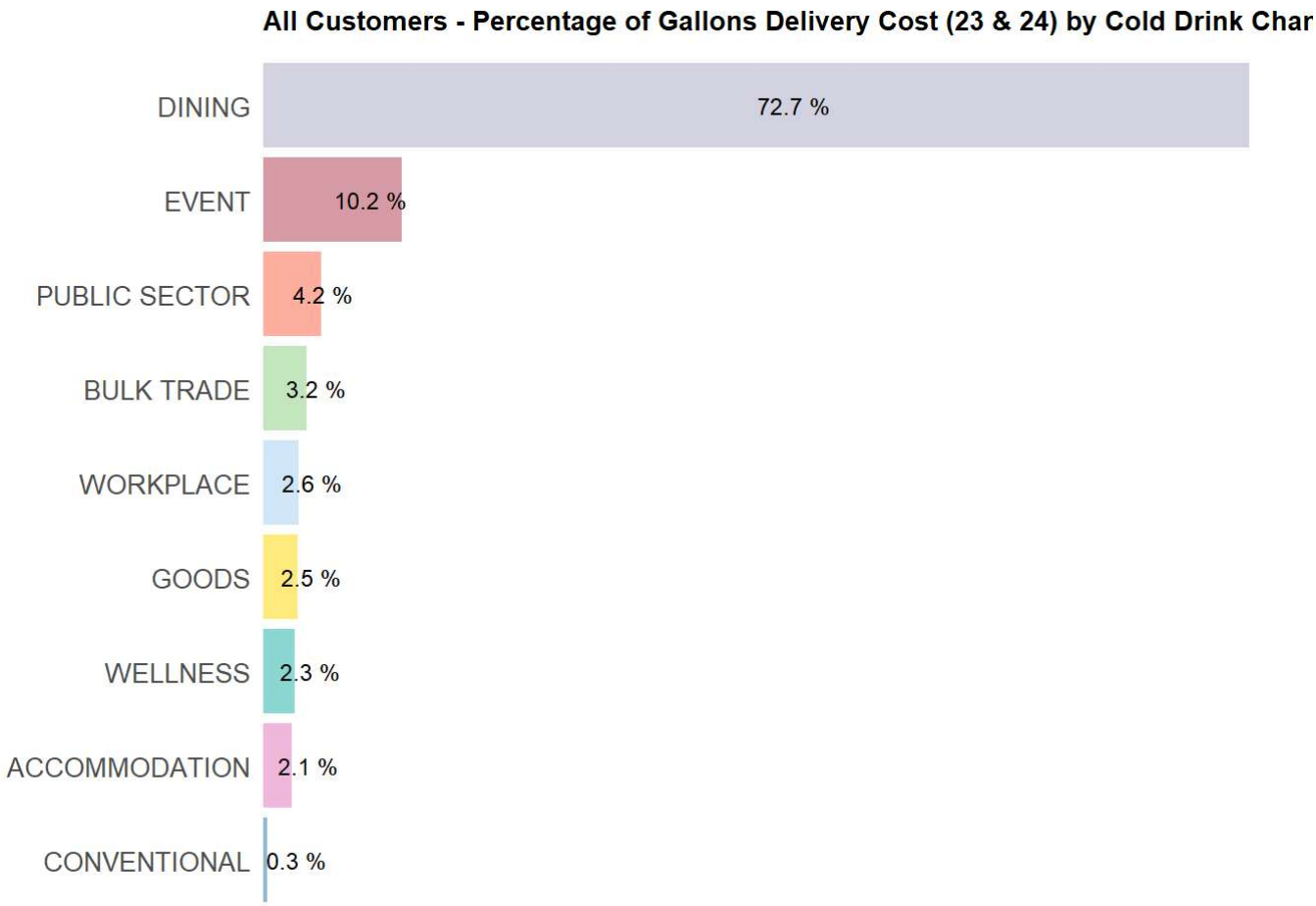
The tables above can be used for different analyses, which will not be discussed here. It is worth highlighting that the bulk trade sector has a high number of outliers, which cause its annual volume average to be very high, while the median is about 5 times lower. This impact can also be observed in the delivery costs.

4.5.2 Cold Drink Channel - Delivered Gallons for All Customers

▶ Code


All Customers - Percentage of Gallons (23 & 24) by Cold Drink Channel
DINING
60.9 %
EVENT
18.7 %
BULK TRADE
6.5 %
PUBLIC SECTOR
4.8 %
WORKPLACE
2.7 %
WELLNESS
2.6 %
ACCOMMODATION
2.1 %
GOODS
1.7 %
CONVENTIONAL
0 %

▶ Code



For gallons, the dining segment is the most representative, accounting for 61% of the volume delivered in 2023 and 2024, and 73% of the cost of gallons. The second segment is events, with 18.7% (10% of the cost), followed by bulk trade with 6.5% (3% of the cost).

The tables below aim to provide detailed information within this group regarding the number of customers, costs, quartile divisions, and other relevant factors.

▶ Code

GALLONS (23 & 24) - Deliveries by Cold Drink Channel - All Customers

Channel	T.Gallons	Gallons %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cus $
DINING	5,881,701	60.9	12,164,673	11,267	71.3	522	235.0	2.07
EVENT	1,802,976	18.7	1,711,570	1,473	9.3	1,224	287.5	0.95
BULK TRADE	631,817	6.5	532,474	464	2.9	1,362	347.5	0.84
PUBLIC SECTOR	460,586	4.8	711,092	635	4.0	725	210.0	1.54
WORKPLACE	258,877	2.7	436,789	682	4.3	380	177.5	1.69
WELLNESS	252,103	2.6	380,152	311	2.0	811	460.0	1.51
ACCOMMODATION	202,090	2.1	346,403	465	2.9	435	150.0	1.71

GALLONS (23 & 24) - Deliveries by Cold Drink Channel - All Customers

Channel	T.Gallons	Gallons %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cus $
GOODS	165,540	1.7	417,493	490	3.1	338	182.5	2.52
CONVENTIONAL	4,502	0.0	43,144	15	0.1	300	135.0	9.58

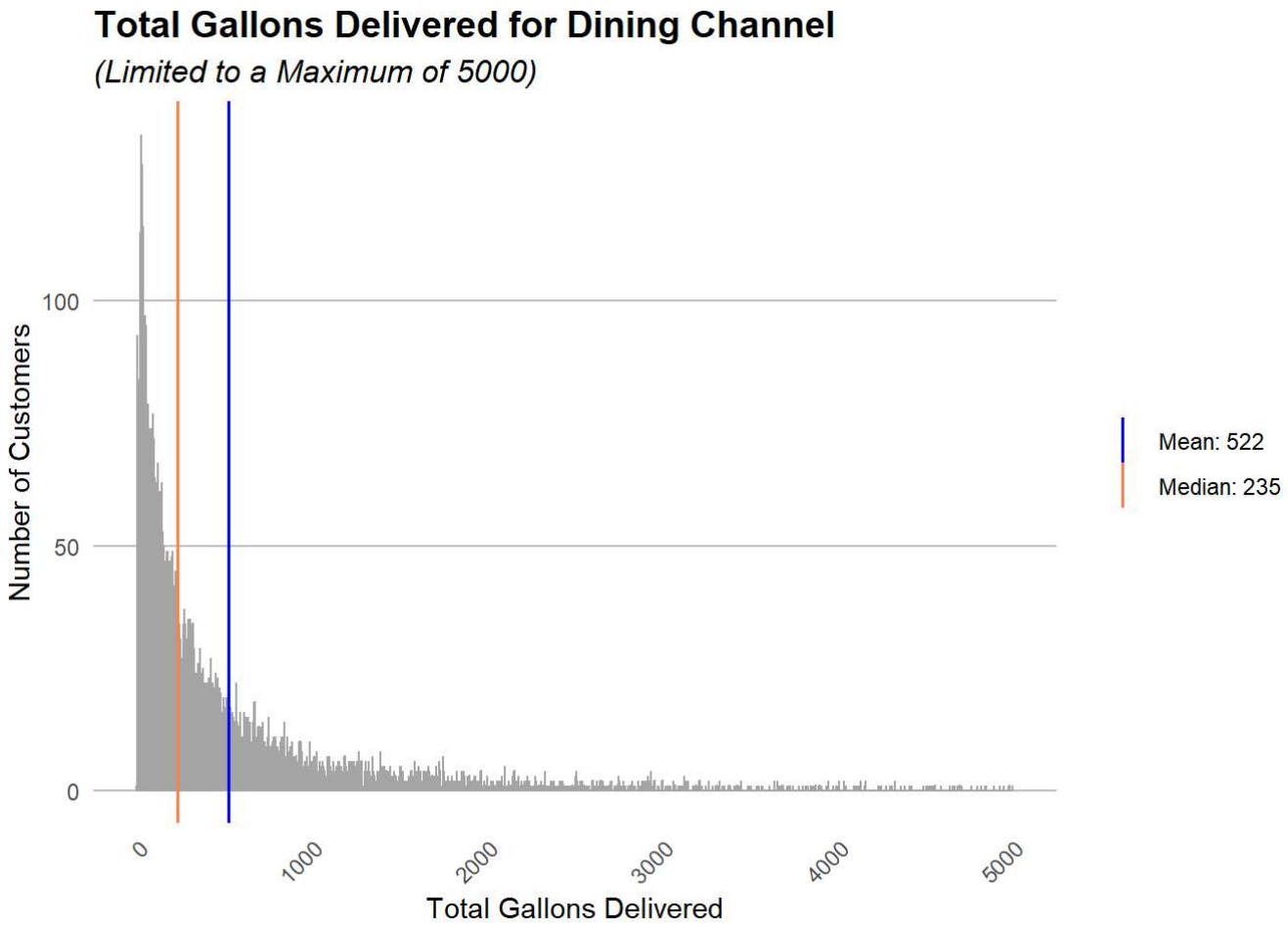
▶ Code

GALLONS (23 & 24) - Quartile Analysis by Cold Drink Channel -

Channel	Avg.Qtd.Cust	Median.Qtd.Cust	1Quart.Qtd	2Quart.Qtd	3Quart.Qtd	1Quart.Vol	1Q.Vol%
BULK TRADE	1,362	348	114	348	1,057	4,870	0.8
EVENT	1,224	288	110	288	800	20,007	1.1
WELLNESS	811	460	158	460	919	5,927	2.4
PUBLIC SECTOR	725	210	79	210	554	6,269	1.4
DINING	522	235	88	235	585	121,015	2.1
ACCOMMODATION	435	150	55	150	518	3,231	1.6
WORKPLACE	380	177	85	177	370	8,121	3.1
GOODS	338	182	82	182	359	5,141	3.1
CONVENTIONAL	300	135	105	135	182	390	8.7

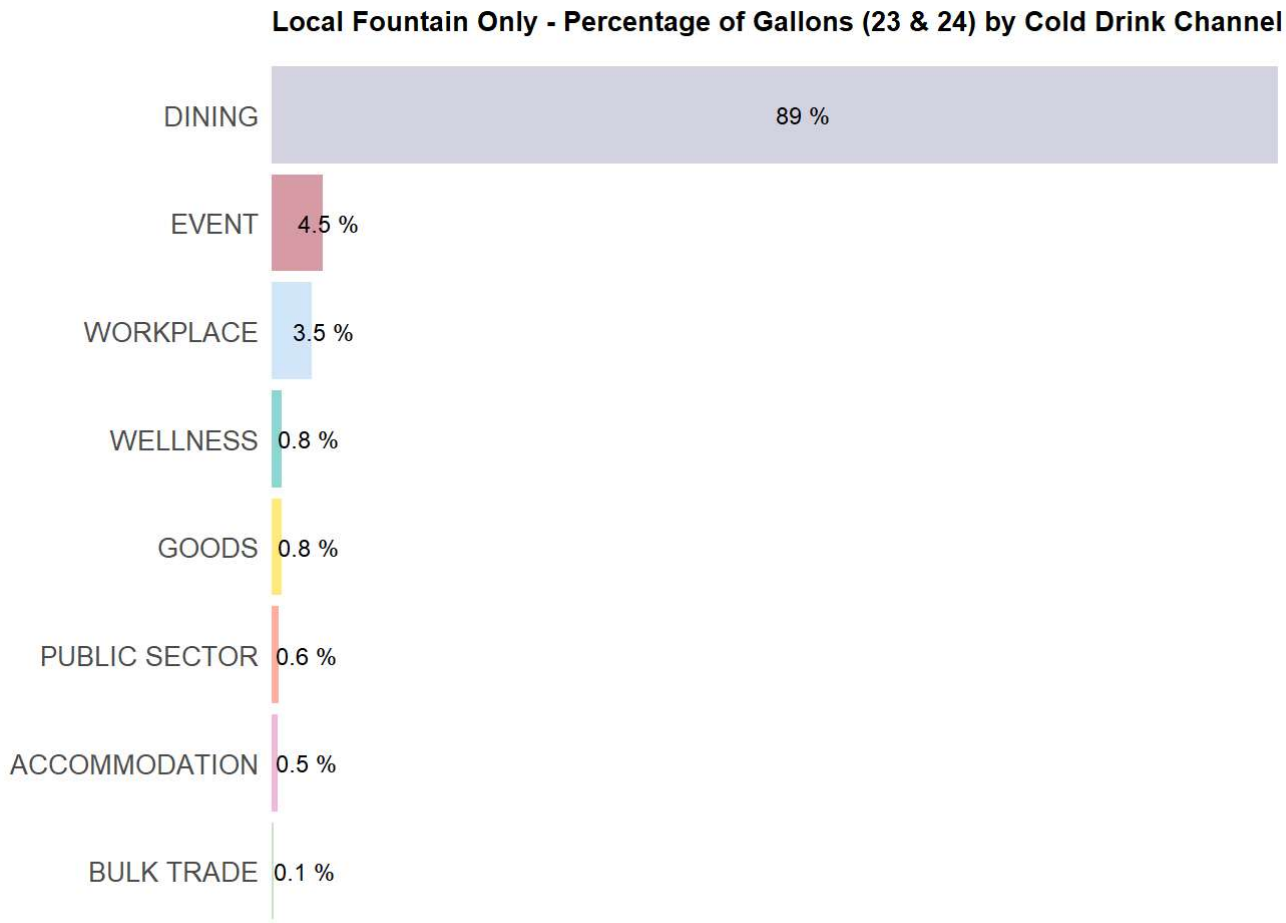
The tables above can be used for different analyses, which will not be discussed here. It is worth highlighting that the dining segment has an average consumption of 522 gallons and a median of 235, resulting in a smaller cost difference when compared to the impact of cases for the bulk trade sector.

▶ Code



4.5.3 Cold Drink Channel - Delivered Gallons for Local Market Partners Fountain Only

▶ Code



▶ Code

Among the local drink-only customers, nearly 90% of the demand is represented by the dining segment, followed by event at 4.5% and workplace at 3.5%. Costs followed nearly the same proportions and were therefore not displayed.

The tables below aim to provide detailed information within this group regarding the number of customers, costs, quartile divisions, and other relevant factors.

▶ Code

GALLONS (23 & 24) - Deliveries by Cold Drink Channel - Local Fountain Only

Channel	T.Gallons	Gallons %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cust $
DINING	510,335	89.0	1,091,574	1,150	84.6	444	177.5	2.14
EVENT	25,774	4.5	50,622	68	5.0	379	121.2	1.96
WORKPLACE	20,029	3.5	30,279	63	4.6	318	116.6	1.51
WELLNESS	4,855	0.8	9,603	9	0.7	539	230.0	1.98
GOODS	4,759	0.8	15,484	25	1.8	190	106.7	3.25
PUBLIC SECTOR	3,655	0.6	12,938	26	1.9	141	81.2	3.54
ACCOMMODATION	3,135	0.5	8,055	13	1.0	241	102.5	2.57

GALLONS (23 & 24) - Deliveries by Cold Drink Channel - Local Fountain Only

Channel	T.Gallons	Gallons %	T.Cost $	N.Cust	P.Cust %	Avg.Qtd.Cust	Median.Qtd.Cust	Avg.Cost.Cust $
BULK TRADE	772	0.1	1,902	5	0.4	154	125.0	2.46

▶ Code

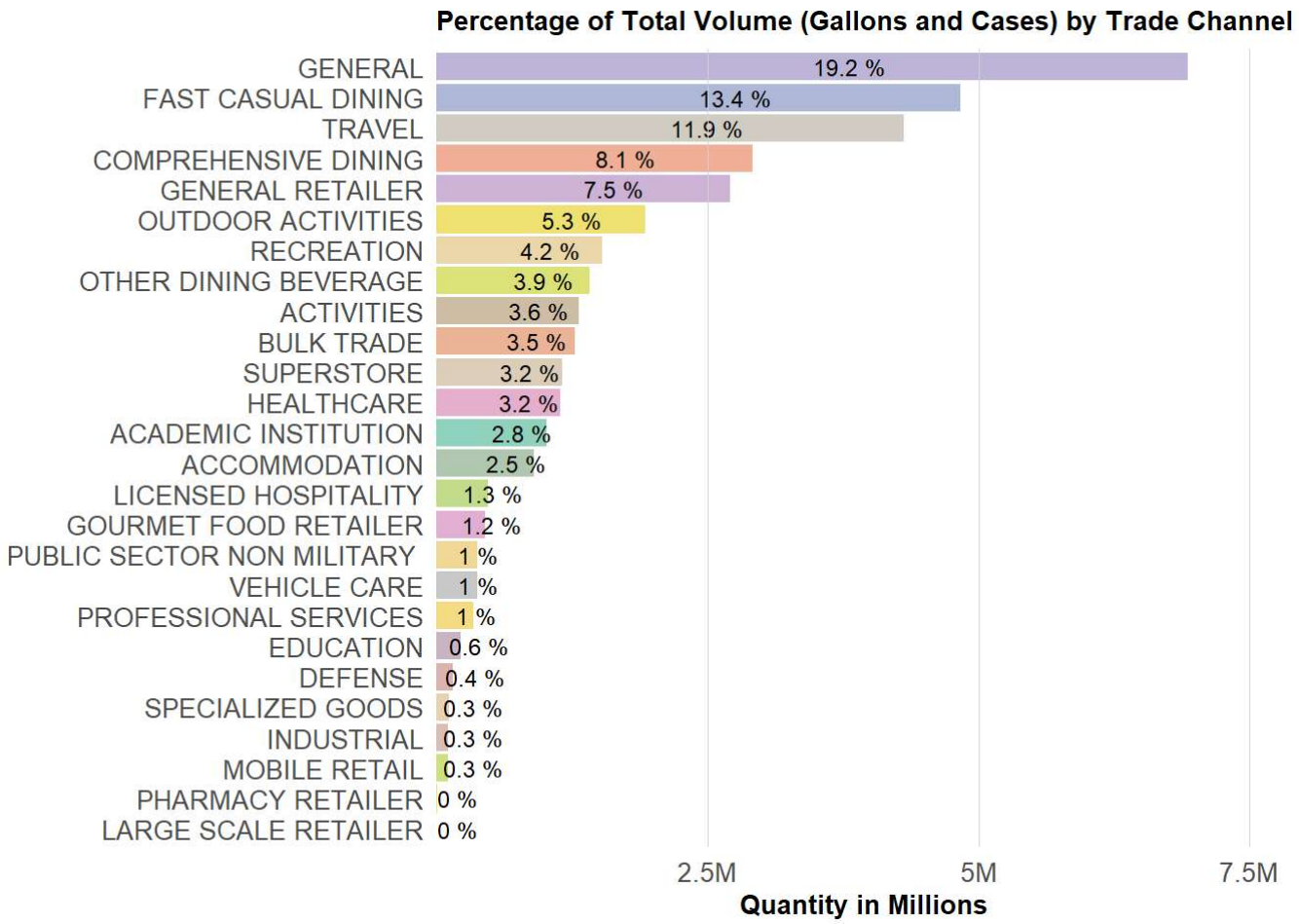
GALLONS (23 & 24) - Quartile Analysis by Cold Drink Channel - Lo

Channel	Avg.Qtd.Cust	Median.Qtd.Cust	1Quart.Qtd	2Quart.Qtd	3Quart.Qtd	1Quart.Vol	1Q.Vol%
WELLNESS	539	230	100	230	998	142	2.9
DINING	444	178	58	178	481	8,472	1.7
EVENT	379	121	39	121	466	358	1.4
WORKPLACE	318	117	55	117	232	449	2.2
ACCOMMODATION	241	102	85	102	323	255	8.1
GOODS	190	107	30	107	222	152	3.2
BULK TRADE	154	125	82	125	235	87	11.3
PUBLIC SECTOR	141	81	38	81	195	198	5.4

The tables above can be used for different analyses, which will not be discussed here. It is worth highlighting that among the local market partners (fountain only), the average consumption was 444 gallons and the median was 177, resulting in an average cost of $2.14 per gallon, which is nearly half of the cost per gallon for customers, which is $3.98.

4.6 Trade Channel

▶ Code



Among the trade channels, Fast Casual Dining (19%), Comprehensive Dining (13.4%), and Travel (12%) rank among the top five in terms of total volume demand. These are also the only segments that individually represent more than 10% of the total volume.

4.7 Sub Trade Channel

The sub trade channel consists of 48 classes, so we decided to create a table for reference and queries.

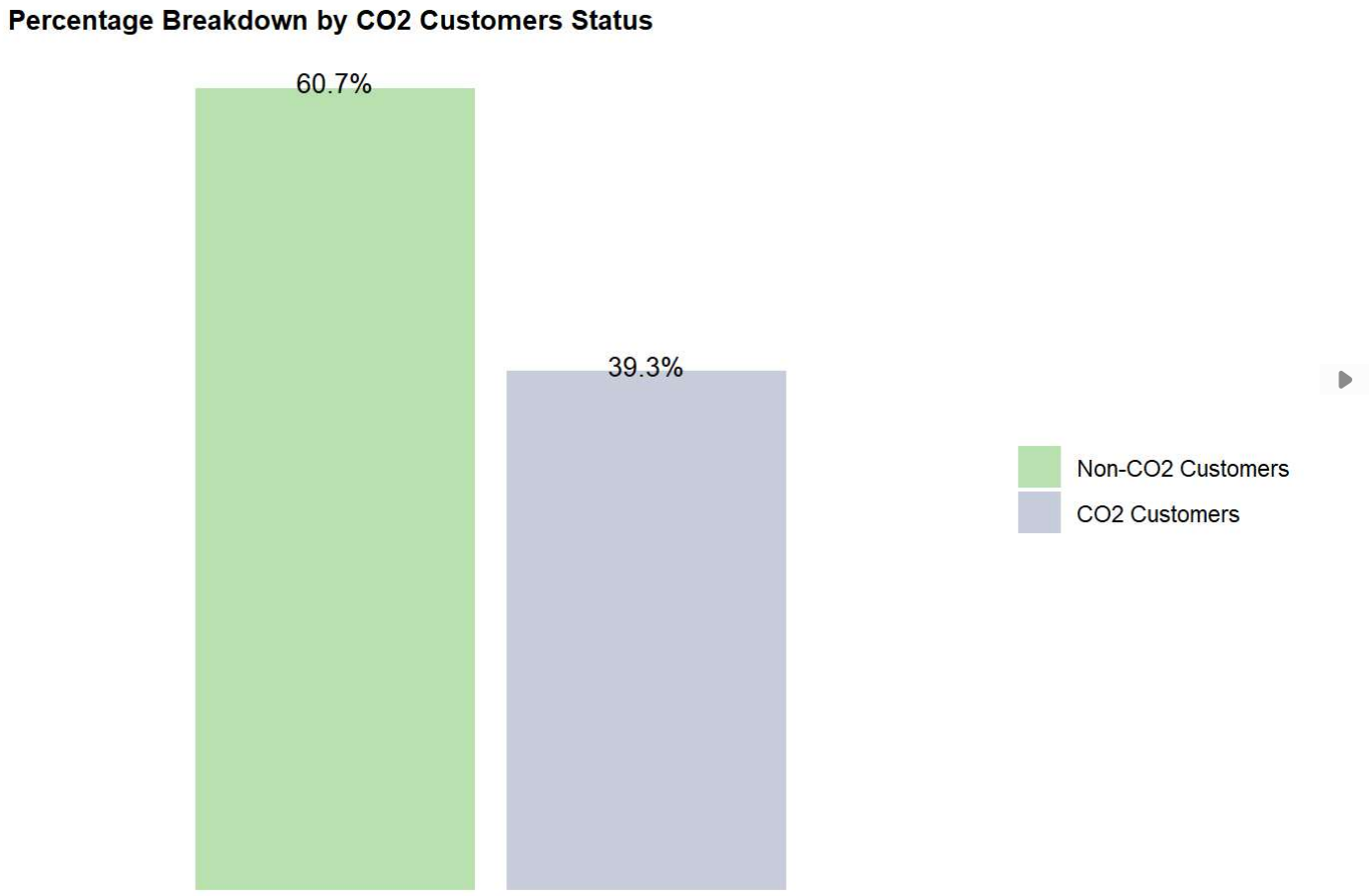
▶ Code

Search:

	SUB_TRADE_CHANNEL	Count
1	ASIAN FAST FOOD	945
2	BOOKS OFFICE	81
3	BULK BEVERAGE RETAIL	1
4	BULK TRADE	104
5	BURGER FAST FOOD	304

4.8 CO2 Customers

▶ Code



Around 61% of customers do not consume CO2, including all local market partners. However, we still find that the percentage of customers consuming CO2 is relatively high, at nearly 39%.

4.9 Transactions by Cases

▶ Code

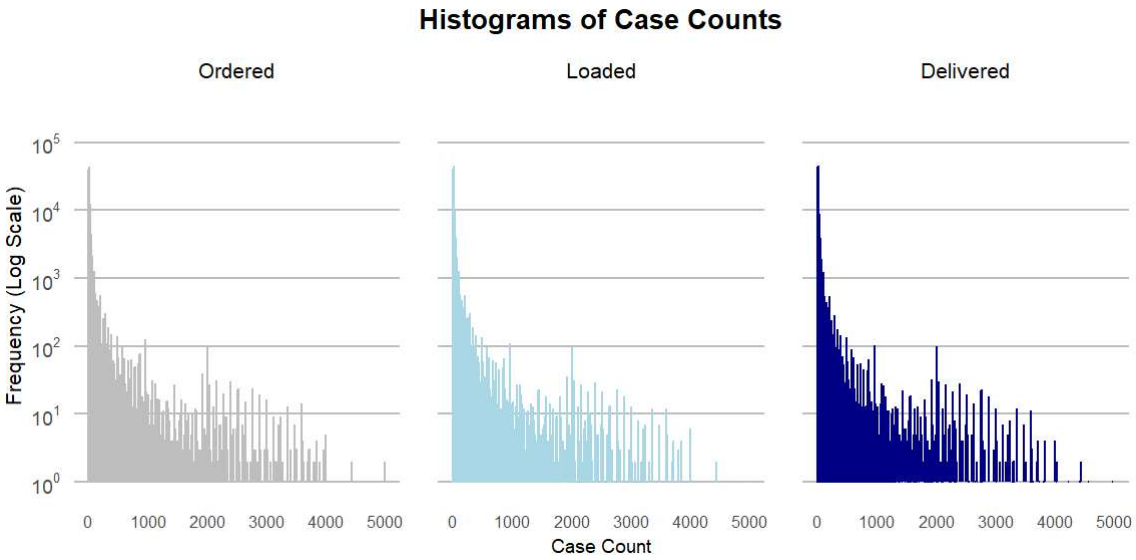
CASES - Statistics by transactions greater than 0

type	min	median	max	sum_qtd	num_trans	avg_qtd_by_trans
ORDERED_CASES	0.0898	11.5	8479	28,074,470	772,877	36
LOADED_CASES	0.0898	11.0	8171	27,103,098	770,624	35
DELIVERED_CASES	0.0001	11.0	8069	26,434,079	750,872	35
RETURNED_CASES	0.0390	8.0	3132	156,165	2,582	60

Considering all case transactions, we created the table above to generate some key metrics. The values for ORDERED CASES, LOADED CASES, and DELIVERED CASES are similar, as expected. There are records with quantities less than 1 unit, and the maximum values exceed 8,000 cases, with the average per transaction being approximately 35 cases.

The number of transactions for RETURNED CASES is much smaller, but there was a return of 3,132 cases. The average number of cases per transaction is 60.

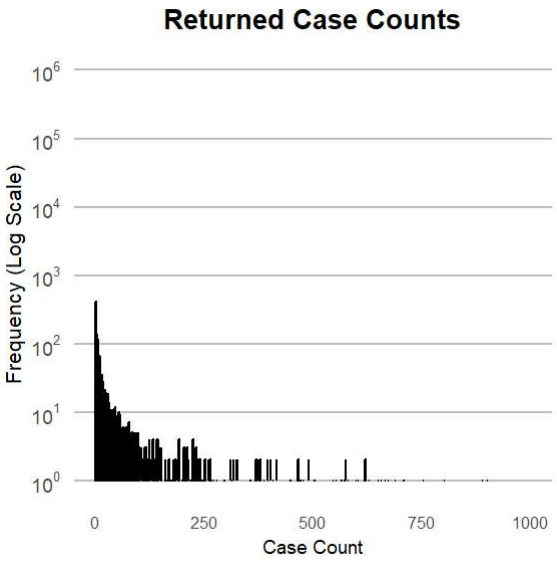
▶ Code



Above, we have the histogram of transactions related to case counts. We have limited the visualization to 5000 cases and applied a logarithmic scale for better interpretation. It is noticeable that the number of transactions decreases near 1900 cases and then increases again around 2000. This could potentially correlate with the larger clients.

Below is the histogram of returned cases, where it is evident that the number of transactions is relatively low, with quantities generally not exceeding 250 cases. There are some transactions exceeding 1,000 cases, but they are rare. These were excluded to make the chart more interpretable.

▶ Code



4.10 Transactions by Gallons

▶ Code

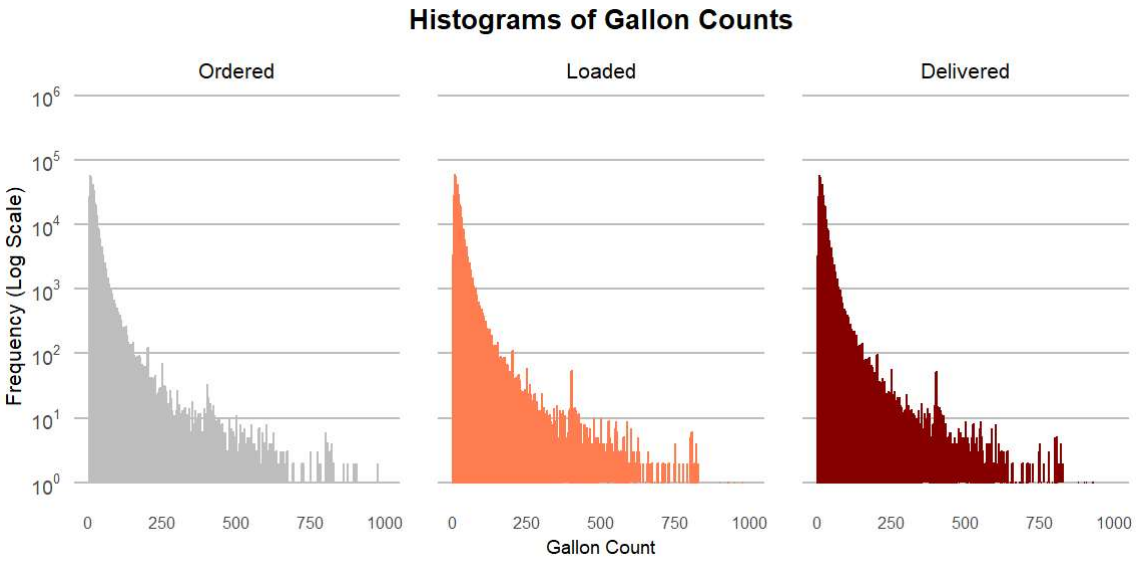
GALLONS - Statistics by transactions greater than 0						
type	**min**	**median**	**max**	**sum_qtd**	**num_trans**	**avg_qtd_by_trans**
ORDERED_GALLONS	0.0898	15.0	2562	10,323,336	482,518	21
LOADED_GALLONS	0.0898	15.0	2562	10,042,299	479,599	21
DELIVERED_GALLONS	0.0159	15.0	2292	9,660,192	464,231	21
RETURNED_GALLONS	0.0156	7.5	1792	32,513	1,760	18

The values for ORDERED GALLONS, LOADED GALLONS, and DELIVERED GALLONS are similar, as expected. There are records with quantities less than 1 unit, and the maximum values exceed 2,200 gallons, with the average per transaction being approximately 21 gallons.

The number of gallon transactions is significantly lower than that of cases, at about 60%.

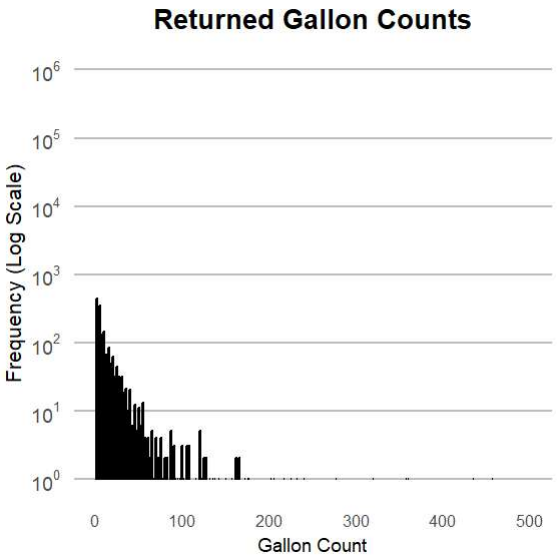
The number of transactions for RETURNED GALLONS is much smaller, but there was a return of 1,792 gallons. The average number of gallons per transaction is 18.

▶ Code



We limited the histograms of gallon counts per transaction to 1000 for better visualization. There are only a few operations that exceed this limit. The vast majority of transactions do not exceed 500 gallons.

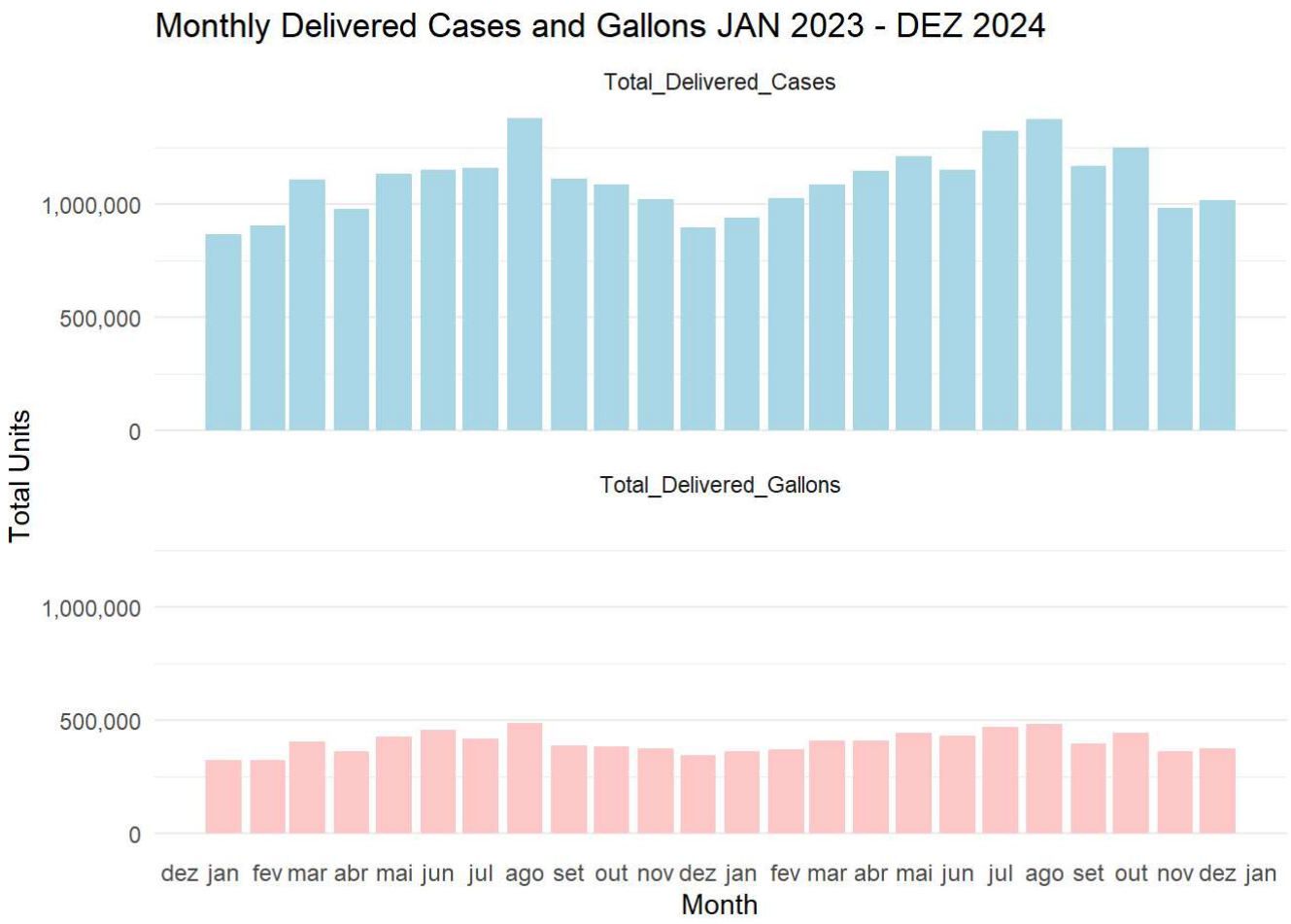
▶ Code



The number of returned gallon transactions is much lower compared to cases. Overall, these transactions do not exceed 100 gallons.

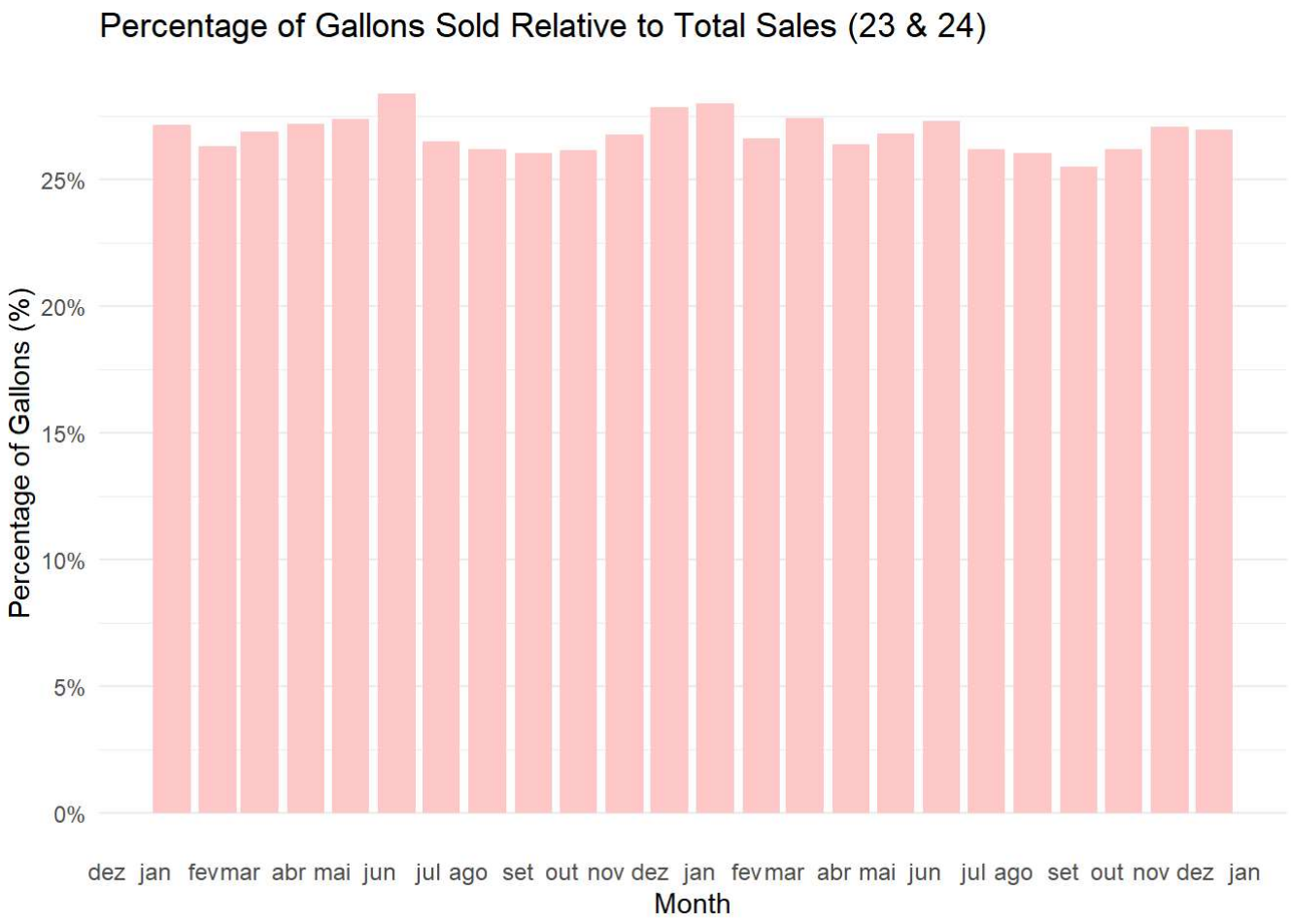
4.11 Transaction Dates Overview

▶ Code



The seasonal effect, related to lower temperatures (OCT-MAR), is more pronounced for the number of delivered cases than for gallons. Additionally, this chart highlights the significant difference in consumption between the two, as both quantities are represented on the same scale.

▶ Code



The sale of gallons over the months remains between 20% and 25% of the total volume.

4.12 Retailer Consumption Quantities

▶ Code

```
Number of Retailers: 1021
```

▶ Code

```
Number of Outlets/Stores: 30320
```

Of the 30,320 stores, many belong to the same chains, with 1,020 networks represented in the dataset. (PRIMARY_GROUP_NUMBER = 0 represents the single stores.)

▶ Code

CASES - Statistics by deliveries greater than 0				
customer_type	**qtd_cases_dlv_23**	**qtd_cases_dlv_24**	**total_qtd_cases_dlv**	**perc_total_qtd_cases**
Retailer Group	10,099,875	10,770,367	20,870,242	79
Single Store	2,684,696	2,879,141	5,563,837	21
Total	12,784,571	13,649,508	26,434,079	100

Considering cases, 80% of the volume went to stores that belong to larger groups.

▶ Code

GALLONS - Statistics by deliveries greater than 0				
customer_type	**qtd_gallons_dlv_23**	**qtd_gallons_dlv_24**	**total_qtd_cases_dlv**	**perc_total_qtd_cases**
Retailer Group	4,565,535	4,565,535	4,565,535	47
Single Store	5,094,657	5,094,657	5,094,657	53
Total	9,660,192	9,660,192	9,660,192	100

As for gallons, the distribution is similar, with 53% going to single stores and 47% to retailer groups, indicating that local stores have a greater share in gallon consumption compared to cases.

▶ Code

TOTAL - Combined Deliveries Quantities for Cases and Gallons				
customer_type	**qtd_cas_gal_23**	**qtd_cas_gal_24**	**total_qtd_cas_gal**	**perc_total_qtd**
Retailer Group	14,665,410	15,335,902	25,435,777	70
Single Store	7,779,354	7,973,798	10,658,494	30
Total	22,444,764	23,309,700	36,094,271	100

The table below helps to better explore the data presented above.

▶ Code

Show 10 entries Search:

Quantity Delivered

PGN	cas_qtd_dlv23	gal_qtd_dlv23	cas_qtd_dlv24	gal_qtd_dlv24	tot
0	2,684,696	2,461,570	2,879,141	2,633,088	5
2487	677,295	668	706,428	888	
8407	501,901	1,996	547,129	4,944	
330	464,824	0	505,619	267	
404	503,598	0	459,051	0	
481	420,259	0	494,011	73	
7353	359,366	1,565	386,132	0	

PGN	cas_qtd_dlv23	gal_qtd_dlv23	cas_qtd_dlv24	gal_qtd_dlv24	tot
464	260,542	0	308,498	0	
374	270,470	130	293,298	375	
408	22,033	35	494,448	42	

Showing 1 to 10 of 1,021 entries Previous 1 2 3 4 5 ... 103 Next

▶ Code

5. Feature Engineering

Considering all the previous analyses, the goal now is to complement the information that can enhance the robustness of the modeling process. Several feature engineering techniques were attempted, but only the most relevant ones will be described.

5.1 Census Data

The data used for updating the location information comes from the U.S. Census Bureau, specifically the American Community Survey (ACS), which annually adjusts its results based on the most recent data. For 2023, the ACS data was retrieved, which is adjusted using the 2020 Census data. However, data for 2024 was not yet available at the time of retrieval.

The decision to use coordinates for store locations, even when there are multiple instances of identical coordinates across different ZIP codes, was made due to the challenges encountered when retrieving Census data based on ZIP codes. Different stores or customers within the same ZIP code can share coordinates, particularly in areas like shopping centers with multiple businesses.

Below are the descriptions of the import data:

▶ Code

Show 10 entries Search:

List of Census Variables and Descriptions

	variable	description
1	MED_HH_INC	Median household income
2	GINI_IDX	Gini index of income inequality
3	PER_CAP_INC	Per capita income

4	MED_HOME_VAL	Median home value
5	POV_POP	Population below poverty
6	INC_LVL_1	Income less than $10,000
7	INC_LVL_2	$10,000 to $14,999
8	INC_LVL_3	$15,000 to $19,999
9	INC_LVL_4	$20,000 to $24,999
10	INC_LVL_5	$25,000 to $29,999

Showing 1 to 10 of 37 entries Previous 1 2 3 4 Next

▶ Code

During the modeling process, it became clear that the absence of 2024 data limited the analysis. In addition, correlations between the census variables and, in particular, customer demand volumes were very low. Because of this, these variables were not explored further in the document. The goal is for this initial process to serve as a foundation for future analyses.

5.2 RFM Score

The RFM (Recency, Frequency, Monetary) analysis segments customers based on purchasing behavior, providing insights into consumption patterns. Adapting this model to analyze customer orders helps assess both the frequency and volume of purchases.

5.2.1 Frequency - Days Between Orders

To adapt the RFM analysis by considering purchase periods and quantities ordered, the analysis will focus on customer orders. Before calculating the number of days between orders (frequency), the total number of orders per customer will be determined, considering only those with a quantity of gallons or cases greater than 0.

▶ Code

There are 135 customers who do not have order transactions greater than zero in the dataset; for these customers, I will consider the number of delivery transactions as orders.

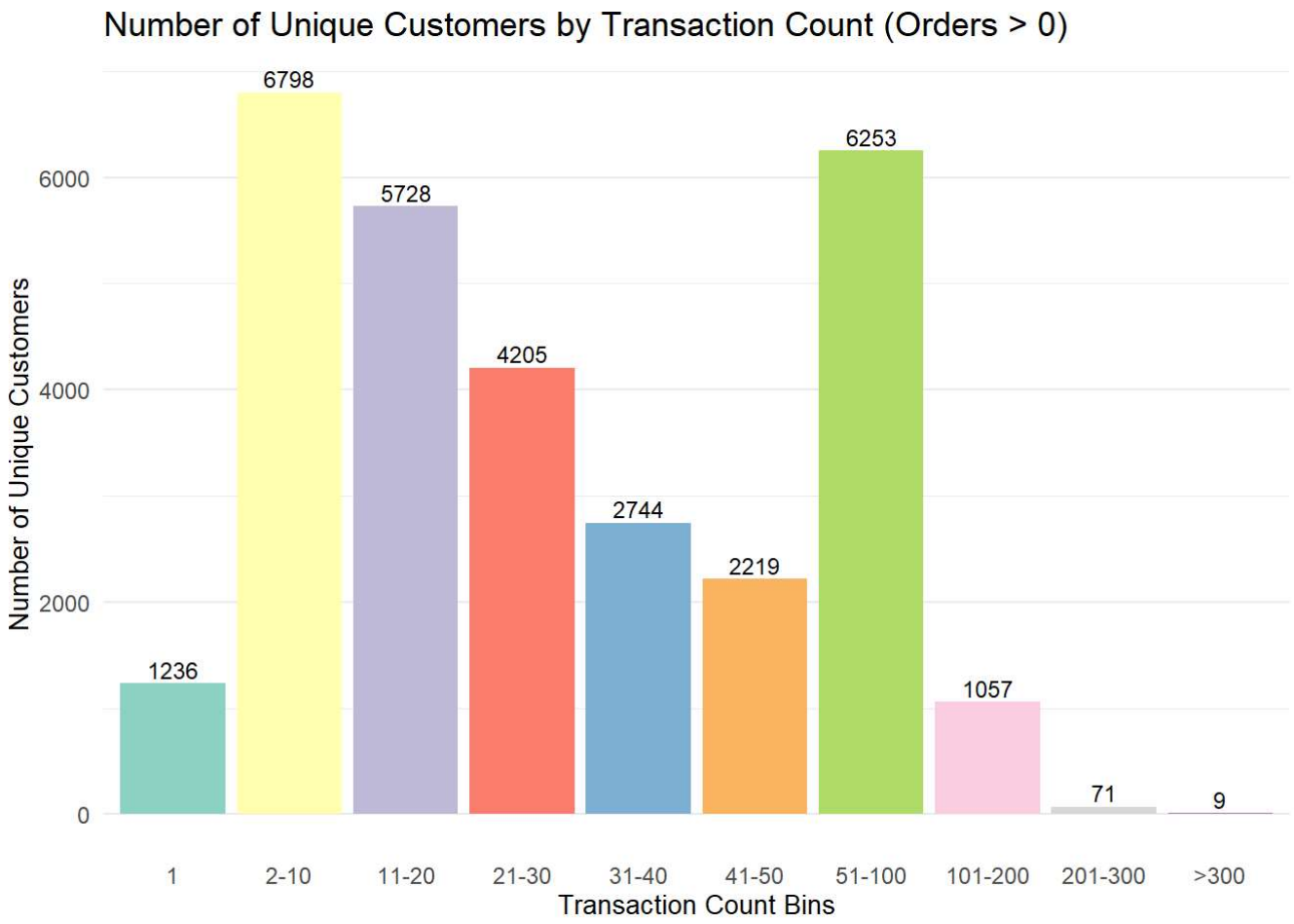
▶ Code

Considering all the order transactions recorded in 2023 and 2024, each unique customer has a minimum of 1 transaction and a maximum of 392 transactions.

To better understand the consumption profile of each customer, below we will visualize the number of customers in transaction bins where the orders of cases or gallons were greater than 0. For the 135 unique

customers who did not have order transactions but received volume, we considered these operations as orders.

▶ Code

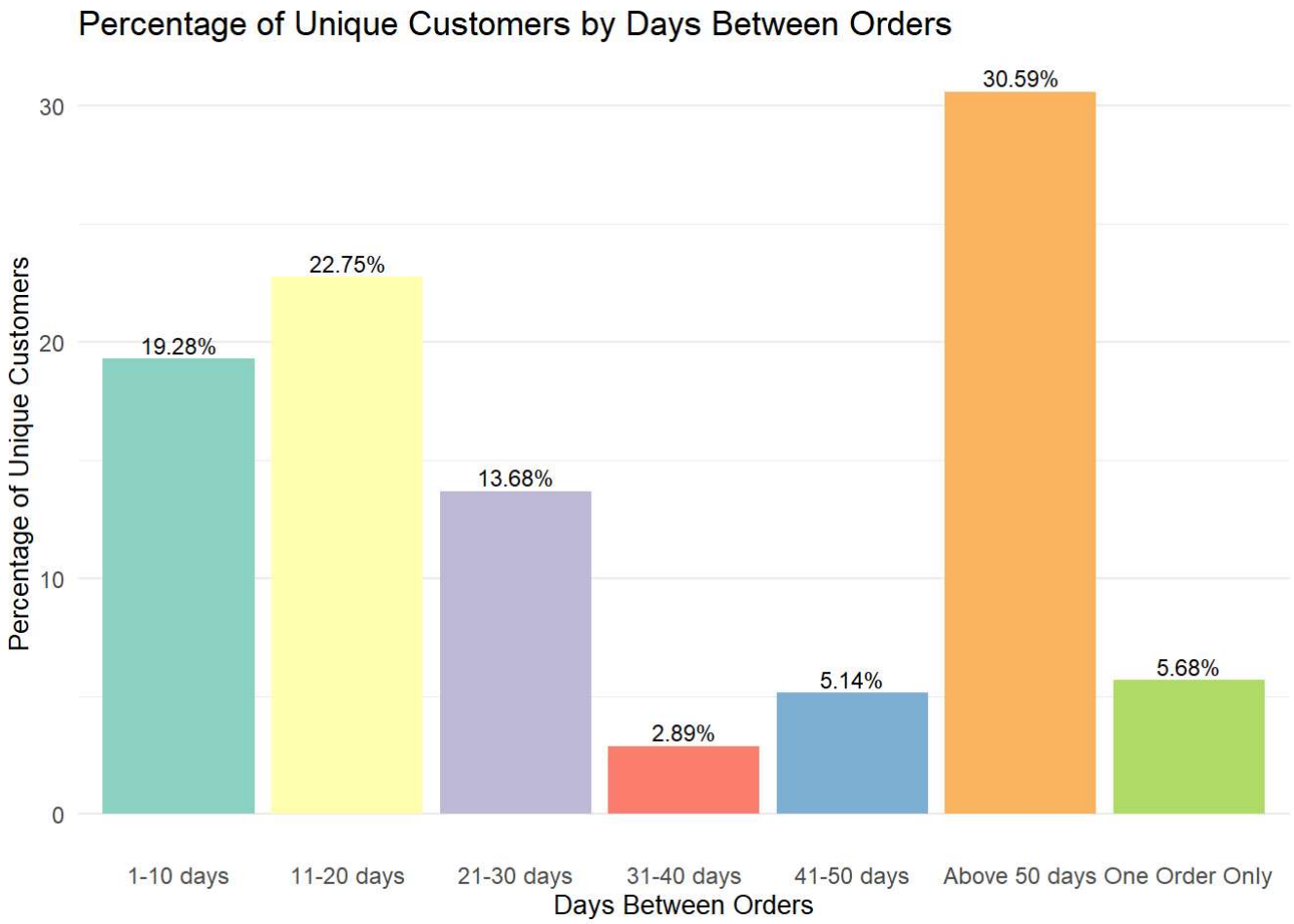


▶ Code

The histogram shows that 1,218 customers have only one order transaction, making it impossible to calculate the days between orders. Additionally, 6,798 customers have between 2 and 10 orders. To ensure more reliable figures, we will consider only customers with at least 11 orders for this indicator. As a result, all customers with fewer transactions will be assigned a value of 731 days between orders, indicating low order frequency over a two-year range.

▶ Code

▶ Code



▶ Code

Around 20% of customers had an average order interval of up to 10 days, while 44% showed an average interval of more than 30 days.Approximately 5% of customers placed only one order, making it impossible to calculate the number of days between orders.

5.2.2 Recency - Time Since Last Order

To calculate recency, I will consider the number of days between the date of the last order and 01-01-2025.

▶ Code

There are 5,754 transaction rows where assigning the last transaction date based on orders is not possible. For these, the date of the last delivery operation will be used as the reference date. The last two transactions, referring to return transactions, will be excluded.

▶ Code

5.2.3 Total Quantity Ordered

As there is no access to the prices charged, and considering that they likely vary among customer types and demanded volumes, the focus will be on the quantities demanded instead of monetary values. This approach aligns with the current objective of customer segmentation.

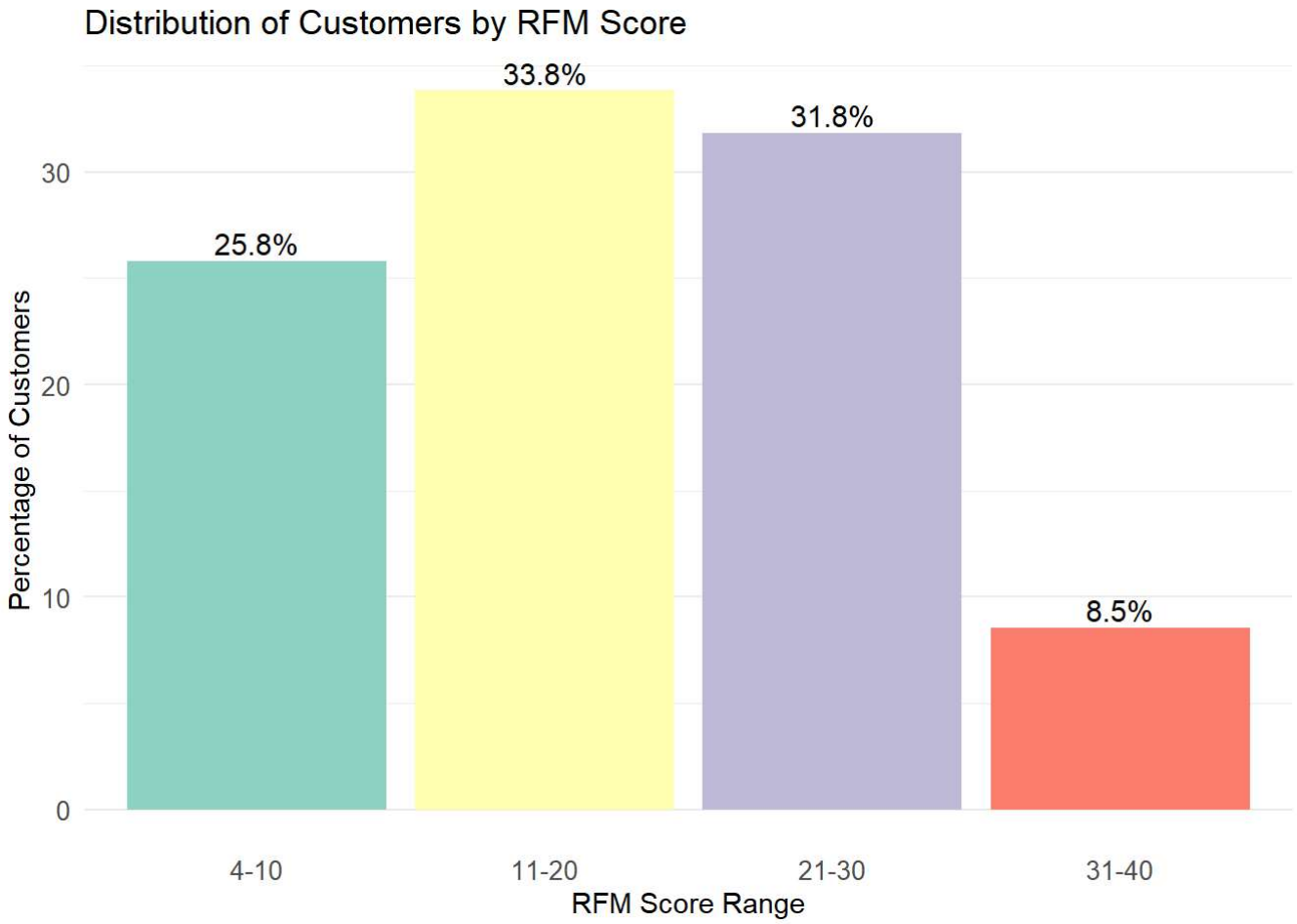
▶ Code

5.2.4 Adapted RFM Score

Scores were assigned to classes based on the distribution of the created variables. The total score, combined with its relative weight, formed the *RFM_SCORE*, which served as an additional variable for customer analysis and segmentation.

To define these scores, the quantitative distribution of each variable was used, especially considering the wide range observed in some of them. Each variable received a score from 1 to 10. In the case of frequency, two separate variables were created, and weight was given not only to the number of orders but also to the interval between them. As a result, the total score ranged from 4 to 40.

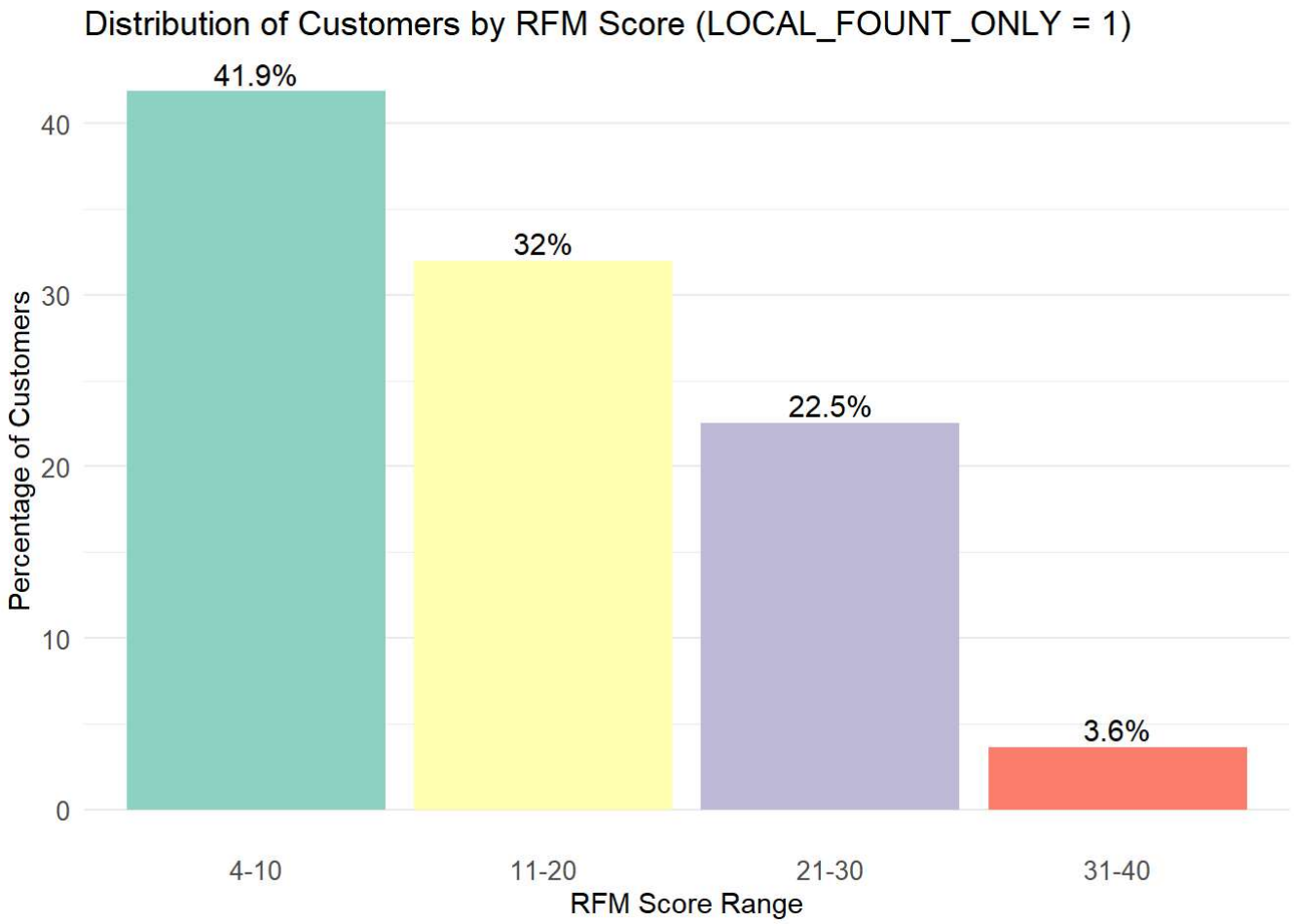
▶ Code



▶ Code

The adapted RFM Score is a method developed to condense various pieces of information related to store consumption. It was observed that 60% of stores have a score up to 20 (the median), 32% have scores between 21-30, and 8.5% have scores above 30. This suggests that only a small percentage of stores exhibit high consumption patterns.

▶ Code



▶ Code

For customers who are local partners and consume only fountain drinks, it is clear that their consumption patterns are even lower. Nearly 74% of them have scores up to 20, and among the remaining customers, less than 3.6% have scores above 30.

5.3 Customer Demand and Growth

5.3.1 Low Demand Customers

It is known that a few customers exhibit very high consumption volumes, causing the average to be skewed above the median. The table below explores metrics related to customers whose demand falls below the first quartile.

▶ Code

Customers Analysis by Cold Drink Channel

Cold Drink Channel	Total Cust.	Avg. Vol Cust.	Median Vol Cust.	1st Quartile Qtd	Cust. Below 1st Quart	Vol % Below 1st Quart
ACCOMMODATION	1235	727	376	122	310	2.0
BULK TRADE	1320	7060	1420	444	330	0.7
CONVENTIONAL	57	190	99	39	15	3.1

Customers Analysis by Cold Drink Channel						
Cold Drink Channel	**Total Cust.**	**Avg. Vol Cust.**	**Median Vol Cust.**	**1st Quartile Qtd**	**Cust. Below 1st Quart**	**Vol % Below 1st Quart**
DINING	15400	633	283	98	3860	1.8
EVENT	3074	1496	329	91	771	0.7
GOODS	5826	628	209	104	1465	2.1
PUBLIC SECTOR	1736	1085	283	94	435	1.1
WELLNESS	479	2413	625	182	119	0.9
WORKPLACE	1193	4046	200	87	303	0.3

For customers with total consumption volumes in 2023 and 2024 below the first quartile, the sums represent very low percentages, ranging from 0.3% to 3.1% of the total for each segment. In the dining segment, for example, 25% of customers showed demand below the first quartile.

Some of these customers have been identified as having high growth potential, as their demand growth is above average. This happens because any increase in demand from these low-volume customers results in higher growth percentages.

The low RFM scores also indicate that these customers have low recency, frequency, and total volume of purchases. Therefore, a flag, **LOW_DEMAND_CUST**, will be created, where a value of 1 will indicate low-consumption customers. With this flag, a white truck will be assigned to these customers, regardless of their growth indices.

Below are the cut volumes by segment:

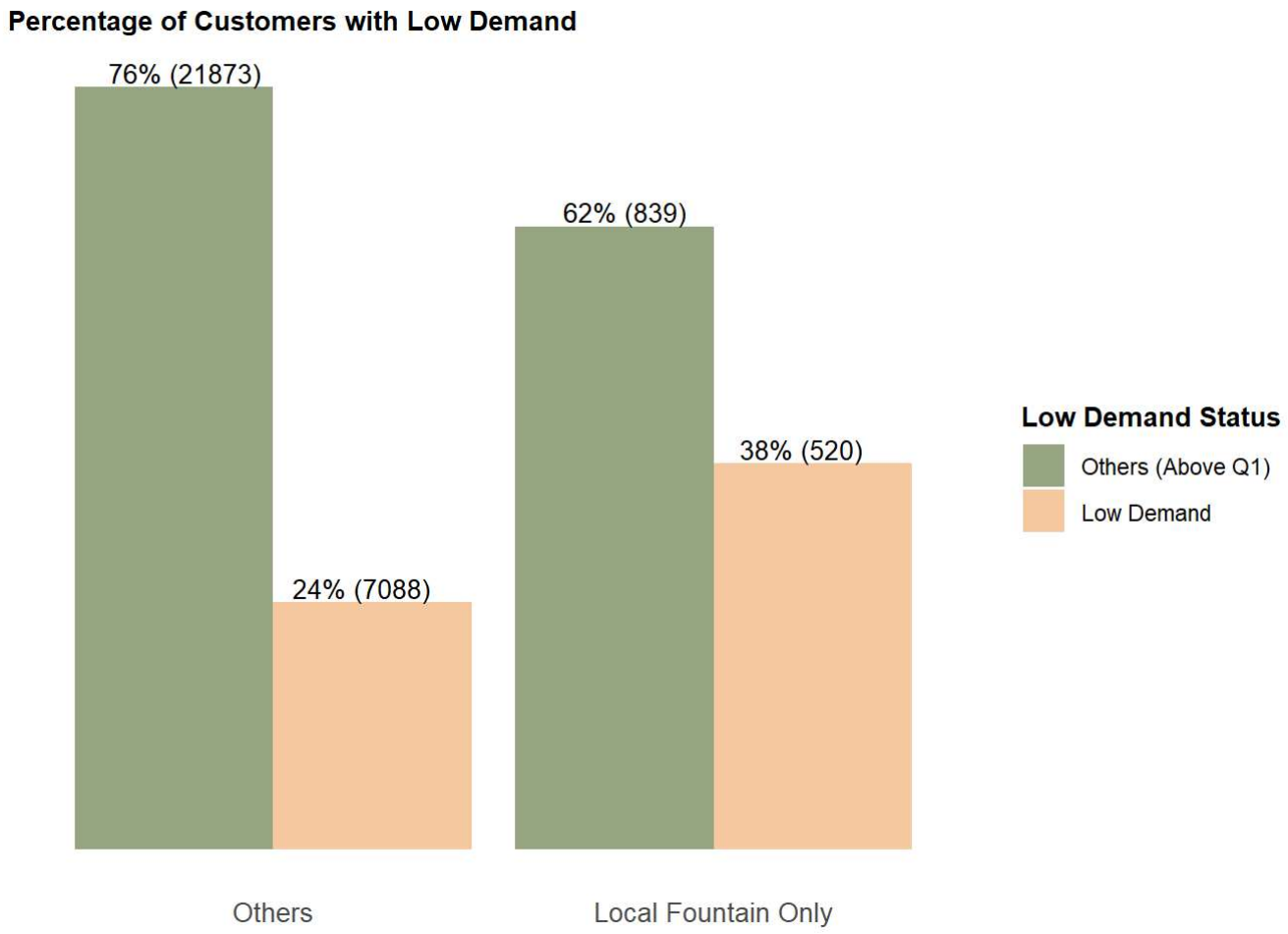
▶ Code

```
ACCOMMODATION     BULK TRADE   CONVENTIONAL         DINING          EVENT
          122            444             39            98             91
       GOODS PUBLIC SECTOR       WELLNESS     WORKPLACE
          104             94            182            87
```

▶ Code

In the plot below, the numbers represent the percentages and the number of customers who received this flag.

▶ Code



5.3.2 Demand Variation between all stores

To measure demand growth patterns across our customer base (January 2023 - December 2024):

1. **Data Preparation:** Combined monthly case and gallon deliveries for each customer into total monthly volumes.
2. **Eligibility:** Required ≥6 months with positive orders for reliable analysis. Customers with <6 ordering months were classified as having no growth potential (6,026 customers).
3. **Growth Calculation:**
    - Split each qualifying customer's order history into two equal time periods
    - For odd numbers of months, divided the middle month equally between periods
    - Calculated growth rate as: (Second Period Total - First Period Total) / First Period Total
4. **Classification:** Customers with growth rates exceeding the average positive growth rate were categorized as high growth potential (HIGH_GROW_POT = 1), while all others received a standard classification (HIGH_GROW_POT = 0).

▶ Code

```
Calculated mean of positive DEMAND_VARIATION:  0.2843618
```

Considering all customers, there was an average demand growth variation of 28%. However, 6,026 customers were excluded from the analysis as their growth could not be calculated due to having fewer than 6 periods of orders. For these customers, it was assumed that they have no growth potential.

Below, the number of customers whose growth exceeded the average, regardless of the segment.

▶ Code

▶ Code

Approximately 9% of customers (123) identified as local market partners who purchase fountain-only products show high growth potential according to the established criteria. For other customers, about 12% (3450) are classified as having high growth potential.

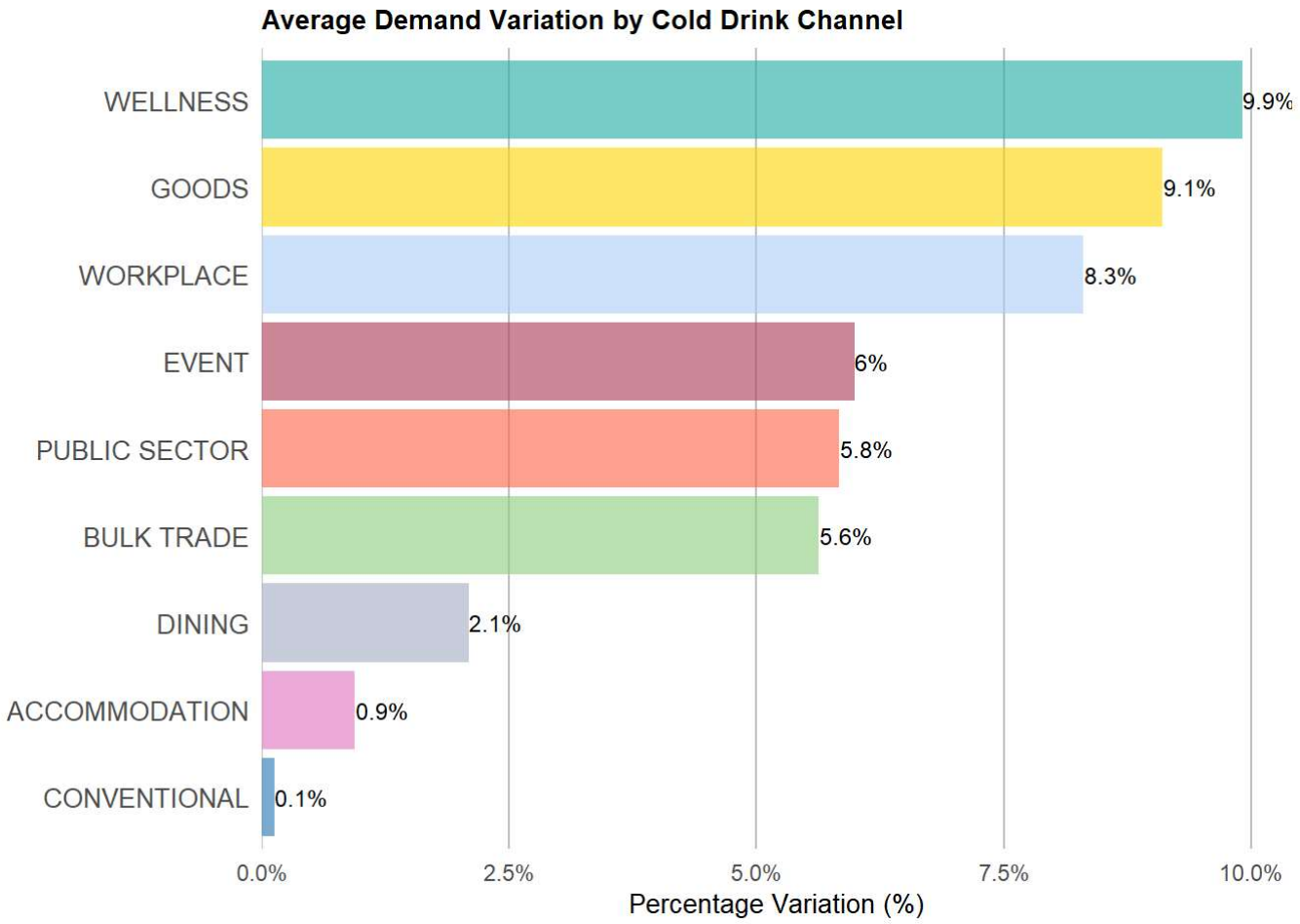
Customers with high volumes are somewhat penalized by this criterion, as significant demand growth is more difficult to achieve. However, their substantial volume already places them as strategic partners, making them essential for close monitoring and prioritized deliveries via red trucks. For these customers, lower distribution costs allow for more competitive pricing, supporting the long-term sustainability of the partnership.

5.3.3 Demand Variation by Cold Drink Channel

Each customer's growth potential was considered within their respective segment. Following the same criteria as before, only customers whose demand variation exceeded the segment average were classified as

high potential. Below is the calculated demand variation for each Cold Drink Channel during the period.

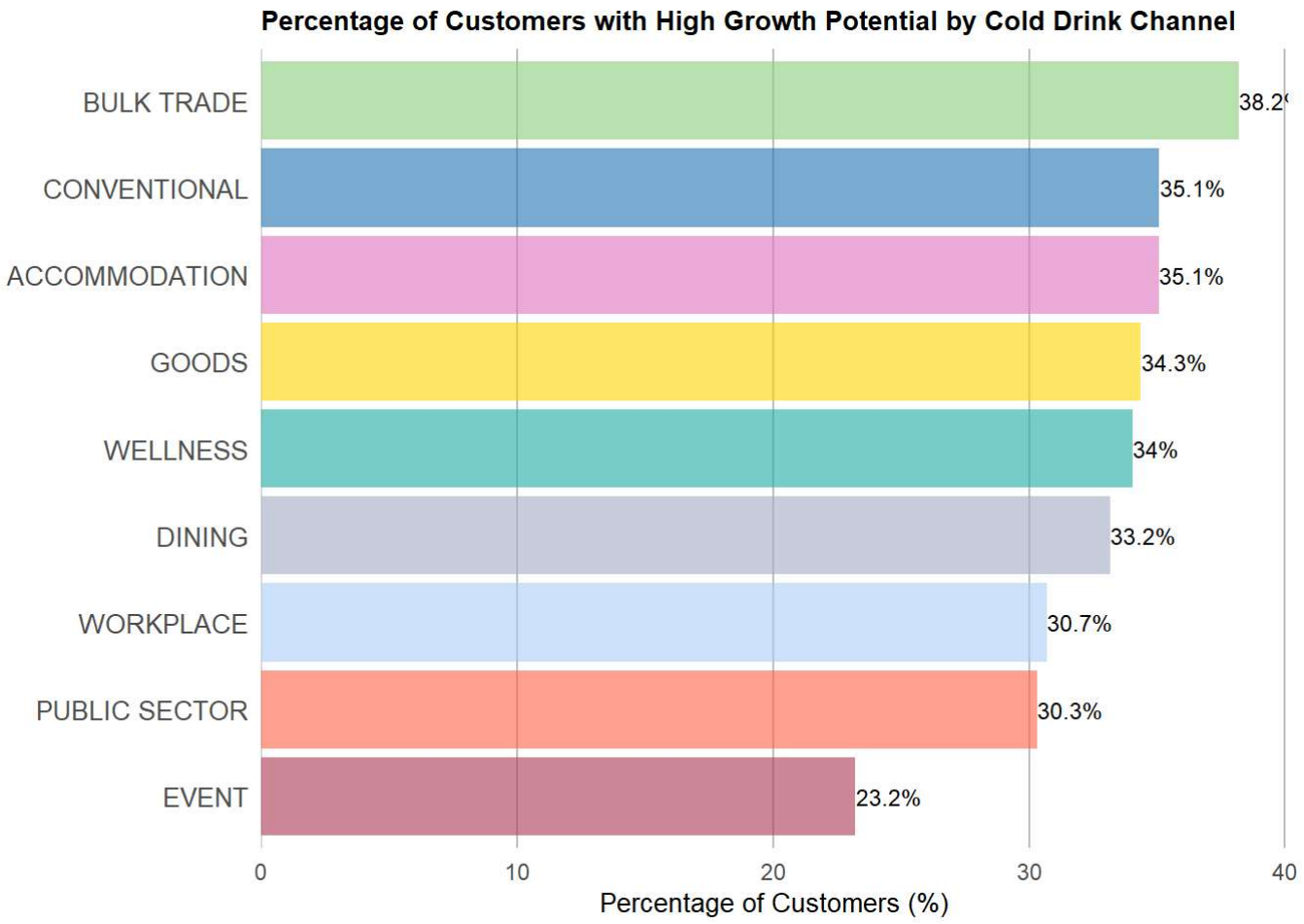
▶ Code



Dining and bulk trade are the most important channels, with customers increasing their demand by 2.1% and 5.6%, respectively, on average.

Wellness experienced the highest variation at almost 10%, but it accounts for only 3.2% of the total volume sold. Goods had the second-highest variation, at 9%, and represents 10% of the total volume.

▶ Code



The majority of segments showed more than 30% of stores with growth above the average for their group. Only the 'Events' segment presented a lower percentage, close to 23%. These customers will be classified as high-growth in their respective segments.

The number of customers with a variation higher than the average for each cold drink channel significantly expands the high-potential customer base. Even when simulating the number of customers with 100% growth above the average, the base was still elevated. Therefore, this criterion will need further analysis before potentially being considered.

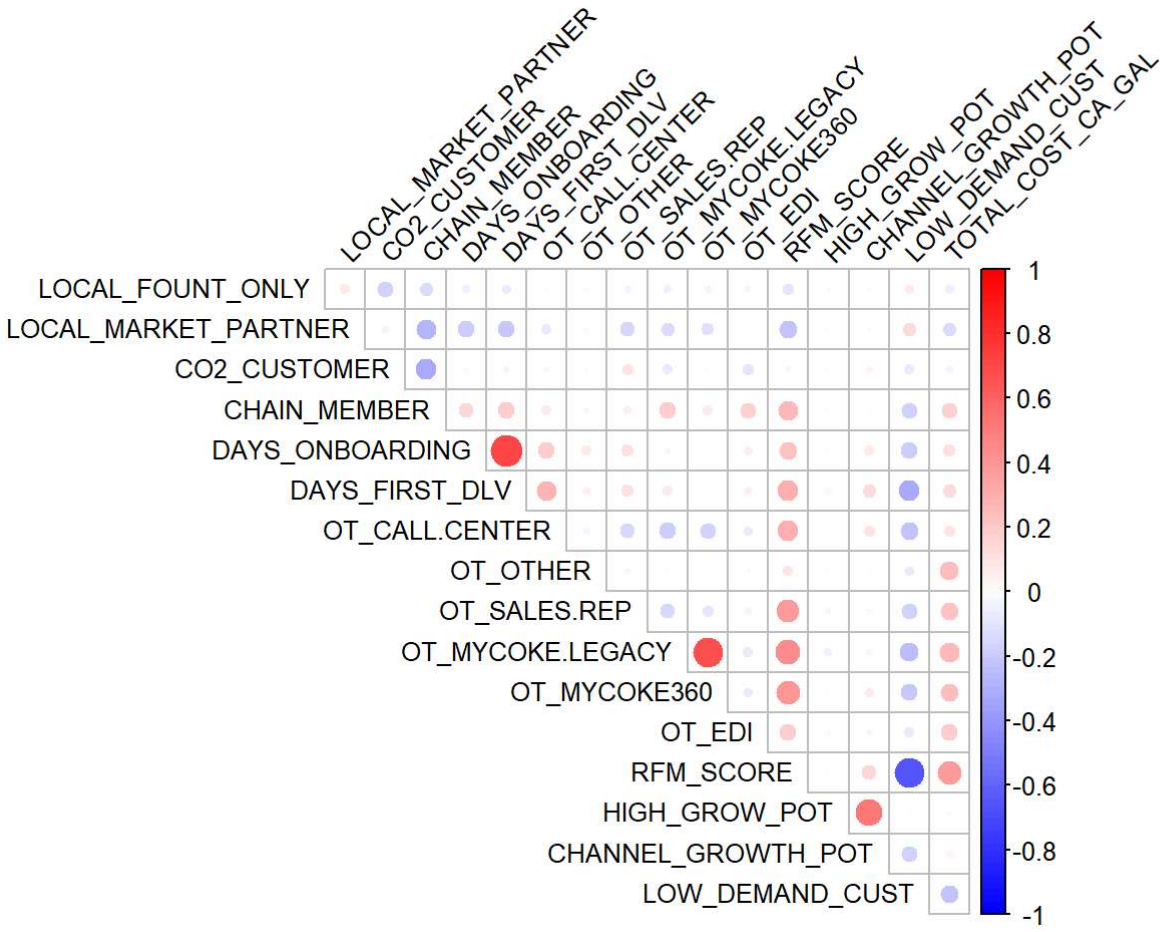
6. Correlations

Customer Features X RFM_SCORE

Seeking to understand how the variables correlate, based on our understanding of the dataset and with the goal of obtaining clear information without multicollinearity, we chose to select numeric variables and display only the most significant correlations (disregarding the range between -0.2 and 0.2).

▶ Code

▶ Code



The strongest correlations were observed between Days Onboarding and Days After First Delivery (0.7) and between the order types MyCoke Legacy and MyCoke 360 (0.66). Both relationships make sense: customers who onboarded earlier tend to have older orders, except for cases where a new store belongs to an established chain. Similarly, customers who previously used the legacy channel transitioned to the newer 360 platform.

There is a correlation of 0.53 between overall customer growth and growth within the Cold Drink Channel, suggesting that expansion trends align across segments. The RFM Score also correlates with various variables that were not directly considered in its calculation, with correlations ranging from 0.44 to 0.27.

Among the negative correlations, the most notable is between RFM Score and Low Demand Customer (-0.65), indicating that lower RFM scores effectively capture low-demand customers.

Census X Total Ordered

All the correlations between the 2023 updated census data showed very low correlations, close to zero, in relation to the customers' consumption patterns.

For this reason, these variables will be excluded, along with others no longer required, to streamline `full_data_customer`. However, the process will be retained in this document, as the company may obtain different results when applying real locations.

7. Customer Segmentation

Since all customers in the original dataset are served by red trucks, there is no prior information on the characteristics of those who would be served by white trucks. The only available reference is the average annual consumption threshold of 400 gallons or cases.

To address this, customers were segmented based on their most relevant characteristics within the available scope, including variables created during the analysis.

Variables selected represent store-level traits or consumption behavior, with geographic and census data excluded.

The variables selected are listed below:

Customer Type & Relationship:
These variables represent customers' relationship with the company and their type:
- LOCAL_FOUNT_ONLY: Customers who only consume fountain drinks.
- LOCAL_MARKET_PARTNER: Local market partners.
- CO2_CUSTOMER: Customers who are CO2 consumers.
- CHAIN_MEMBER: Customers who are part of a chain.

Time-Related Metrics:
Time-related metrics track customers' activity and engagement over time:
- DAYS_ONBOARDING: Number of days since onboarding.
- DAYS_FIRST_DLV: Number of days since the first delivery.
- DAYS_AF_LAST_ORD: Number of days after the last order.
- AVG_DAYS_BET_ORD: Average number of days between orders.

Order & Sales Behavior:
These variables represent customer behaviors in terms of orders and sales:
- NUM_ORDERS: Total number of orders.
- TOTAL_ORDERED: Total volume of orders.
- RFM_SCORE: Adapted Recency, Frequency, Monetary score.
- TOTAL_COST_CA_GAL: Total cost in deliveries for 2023 and 2024.

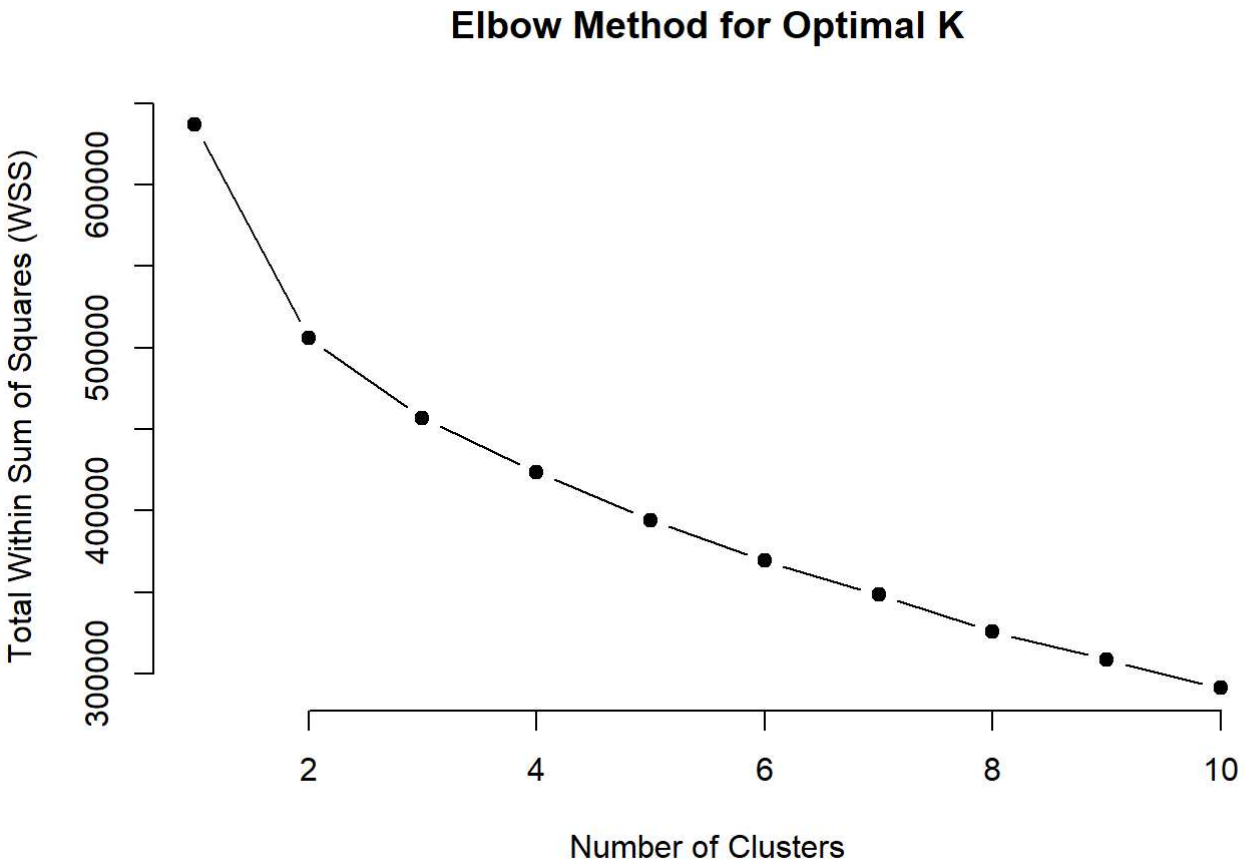
Order Channels:
This category contains data on the various channels through which customers make their transactions:
- OT_CALL.CENTER: Transactions via call center.
- OT_OTHER: Transactions made through other means (emails, trade fairs, etc.).
- OT_SALES.REP: Transactions via sales representatives.
- OT_MYCOKE: Transactions via MyCoke (legacy platform).
- OT_EDI: Transactions via electronic direct ordering (EDI).

Growth & Demand Potential:
These flags indicate customers' growth and demand potential:
- HIGH_GROW_POT: Flag for customers with above-average growth potential across all segments.

- CHANNEL_GROWTH_POT: Flag for customers with above-average growth within their segment.
- LOW_DEMAND_CUST: Flag for customers with low demand (below the 1st quartile) by segment.

Three variables have a wide range of values with extreme outliers. For these variables—**NUM_ORDERS**, **TOTAL_ORDERED**, and **TOTAL_COST_CA_GAL**—we will apply a logarithmic transformation.

▶ Code



After testing different compositions to calculate the silhouette score and ARI score—varying the number of clusters from 2 to 4, using multiple distance metrics (Euclidean, Manhattan), and applying different algorithms (Hartigan-Wong, MacQueen, Lloyd)—the most relevant metrics are presented below.

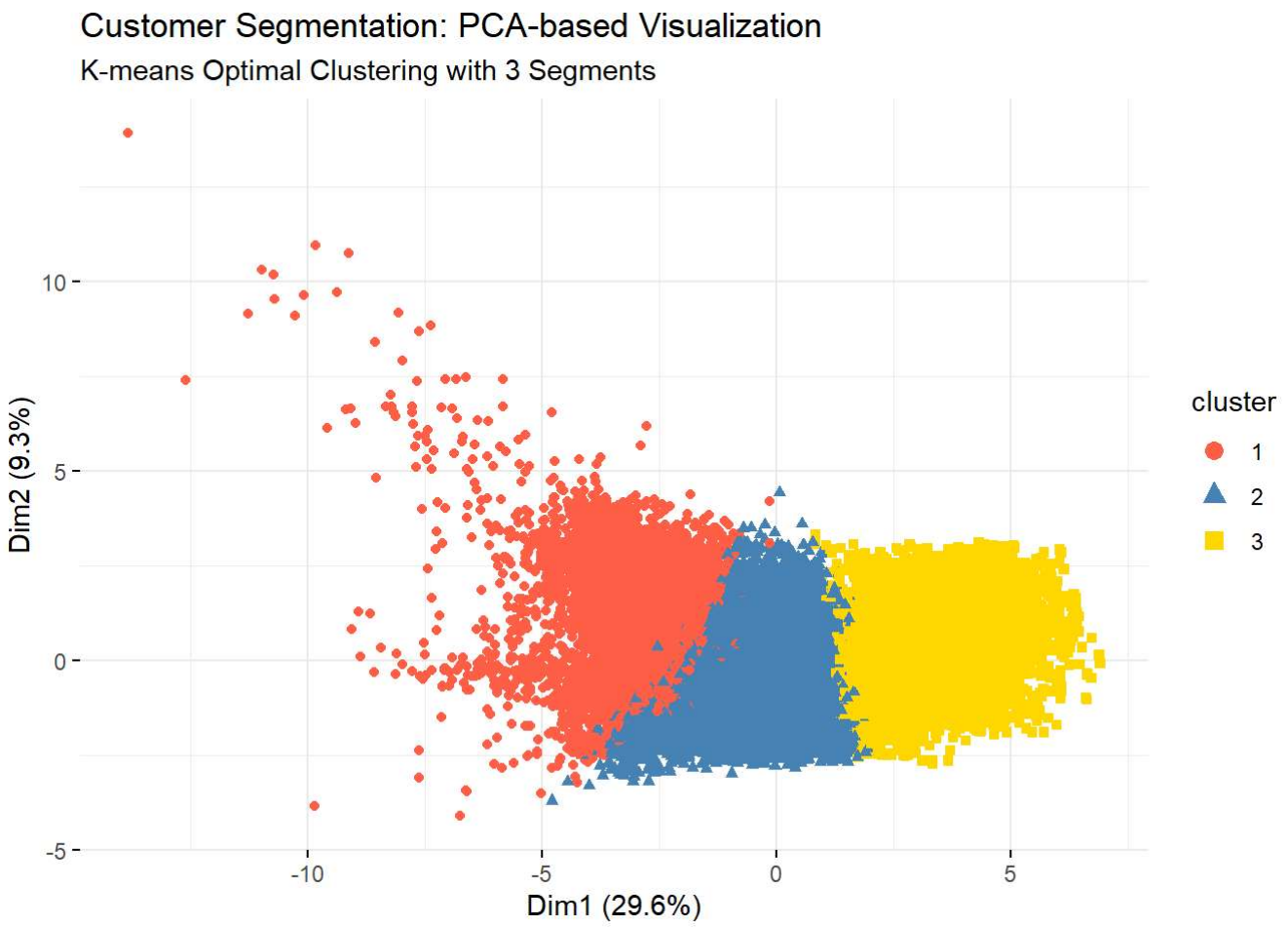
▶ Code

Parameter	Silhouette Score	Adjusted Rand Index (ARI)
Euclidean, 2 Clusters	0.210	-0.054
Euclidean, 3 Clusters	0.180	0.043
Euclidean, 4 Clusters	0.176	0.048

Given the results, the combination "Euclidean, 3 Clusters" was selected, using `centers = 3, nstart = 25`, and the "Hartigan-Wong" algorithm (default), as it demonstrated the best performance among the tested options. Still, the separation between clusters remains marginal and relatively weak.

Below is the visualization of the clusters based on the two principal components.

▶ Code



The customer segmentation will be discussed later, including the interpretation of each cluster.

7.1 Clusters and principal components

Given the visualization of the clusters through their principal components, the decision was made to further explore the characteristics of the two main components, as they account for 39% of the total variability.

▶ Code

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
LOCAL_FOUNT_ONLY	-0.05	0.00	0.00	0.03	-0.18	0.29	0.73	-0.24	0.00
LOCAL_MARKET_PARTNER	-0.12	0.02	0.22	-0.14	-0.13	0.35	0.18	0.41	0.29
CO2_CUSTOMER	0.00	-0.02	0.24	-0.36	0.52	-0.10	-0.25	0.08	0.10
CHAIN_MEMBER	0.14	0.00	-0.30	0.30	-0.36	-0.18	-0.25	-0.11	-0.23
DAYS_ONBOARDING	0.20	-0.43	-0.17	0.25	0.26	0.00	0.11	0.10	0.19

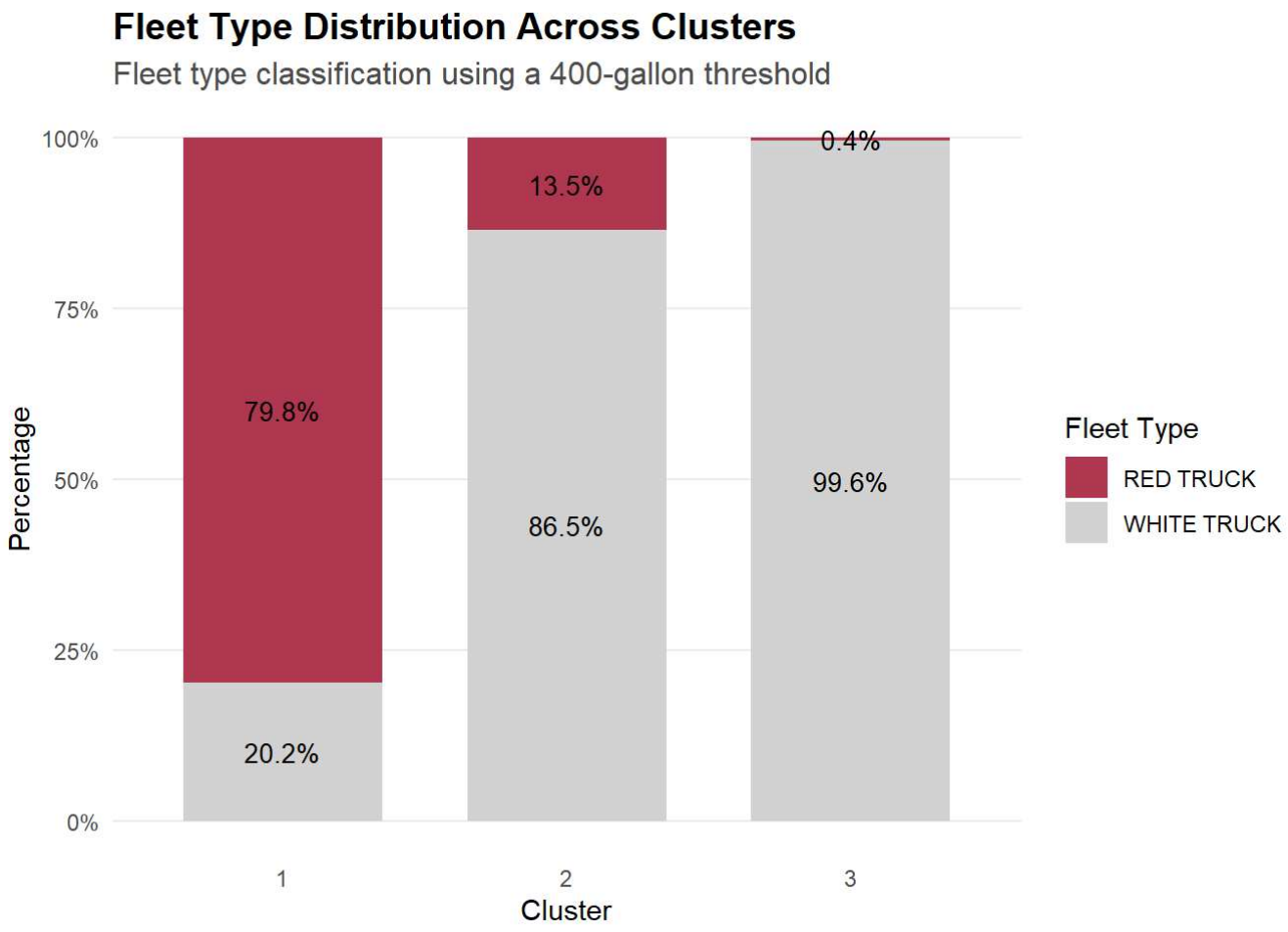
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
DAYS_FIRST_DLV	0.22	-0.43	-0.14	0.26	0.26	0.01	0.11	0.10	0.21
DAYS_AF_LAST_ORD	-0.22	-0.10	-0.23	0.22	0.19	-0.17	0.11	0.04	0.06
AVG_DAYS_BET_ORD	-0.32	0.07	-0.12	0.04	0.02	-0.12	0.02	0.00	-0.03
OT_CALL.CENTER	0.12	-0.34	0.00	-0.11	-0.04	0.48	-0.22	-0.20	-0.37
OT_OTHER	0.04	-0.01	-0.05	-0.06	0.05	-0.19	0.27	0.67	-0.62
OT_SALES.REP	0.14	-0.03	-0.10	-0.42	-0.01	-0.52	0.31	-0.26	0.12
OT_MYCOKE.LEGACY	0.19	0.42	0.03	0.39	0.20	0.06	0.04	0.02	0.07
OT_MYCOKE360	0.16	0.44	0.16	0.30	0.21	0.05	0.02	0.01	0.04
OT_EDI	0.07	0.00	-0.19	-0.05	-0.47	-0.05	-0.19	0.39	0.48
NUM_ORDERS	0.34	0.12	-0.10	-0.09	-0.01	-0.10	0.10	-0.06	0.01
TOTAL_ORDERED	0.37	0.07	-0.03	-0.11	-0.02	-0.01	0.05	0.05	0.00
RFM_SCORE	0.37	0.11	0.02	-0.14	-0.07	0.04	0.00	-0.05	-0.02
HIGH_GROW_POT	0.02	-0.20	0.53	0.24	-0.22	-0.29	0.01	0.00	-0.04
CHANNEL_GROWTH_POT	0.09	-0.21	0.55	0.19	-0.15	-0.20	0.03	-0.05	0.00
LOW_DEMAND_CUST	-0.30	0.05	-0.08	0.10	-0.02	-0.15	0.01	-0.07	-0.01
TOTAL_COST_CA_GAL	0.36	0.03	-0.02	-0.04	-0.06	0.03	0.00	0.06	-0.02
Variance_Explained	0.30	0.09	0.07	0.07	0.06	0.06	0.05	0.05	0.05
Cumulative_Variance	0.30	0.39	0.46	0.54	0.60	0.66	0.71	0.76	0.80

Principal Component 1 has the highest weight from the variables RFM_SCORE, NUM_ORDERS, TOTAL_ORDERED, and TOTAL_COST_CA_GAL, representing 30% of the variance.
Principal Component 2 adds another 9% of variance, with the highest weight from the OT_MYCOKE variables.

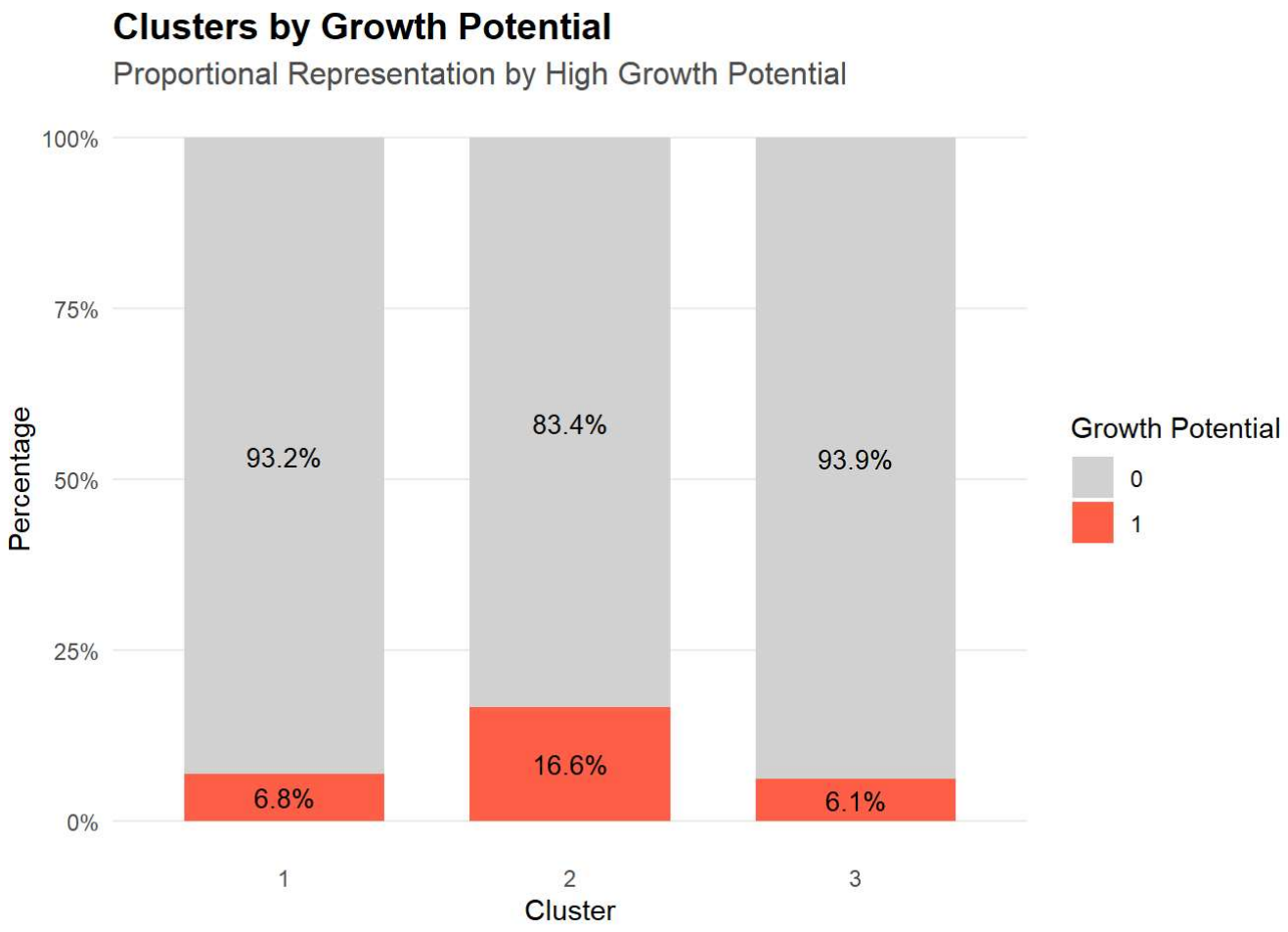
7.2 Clusters Features

The clusters will be characterized based on their relationships with other variables.

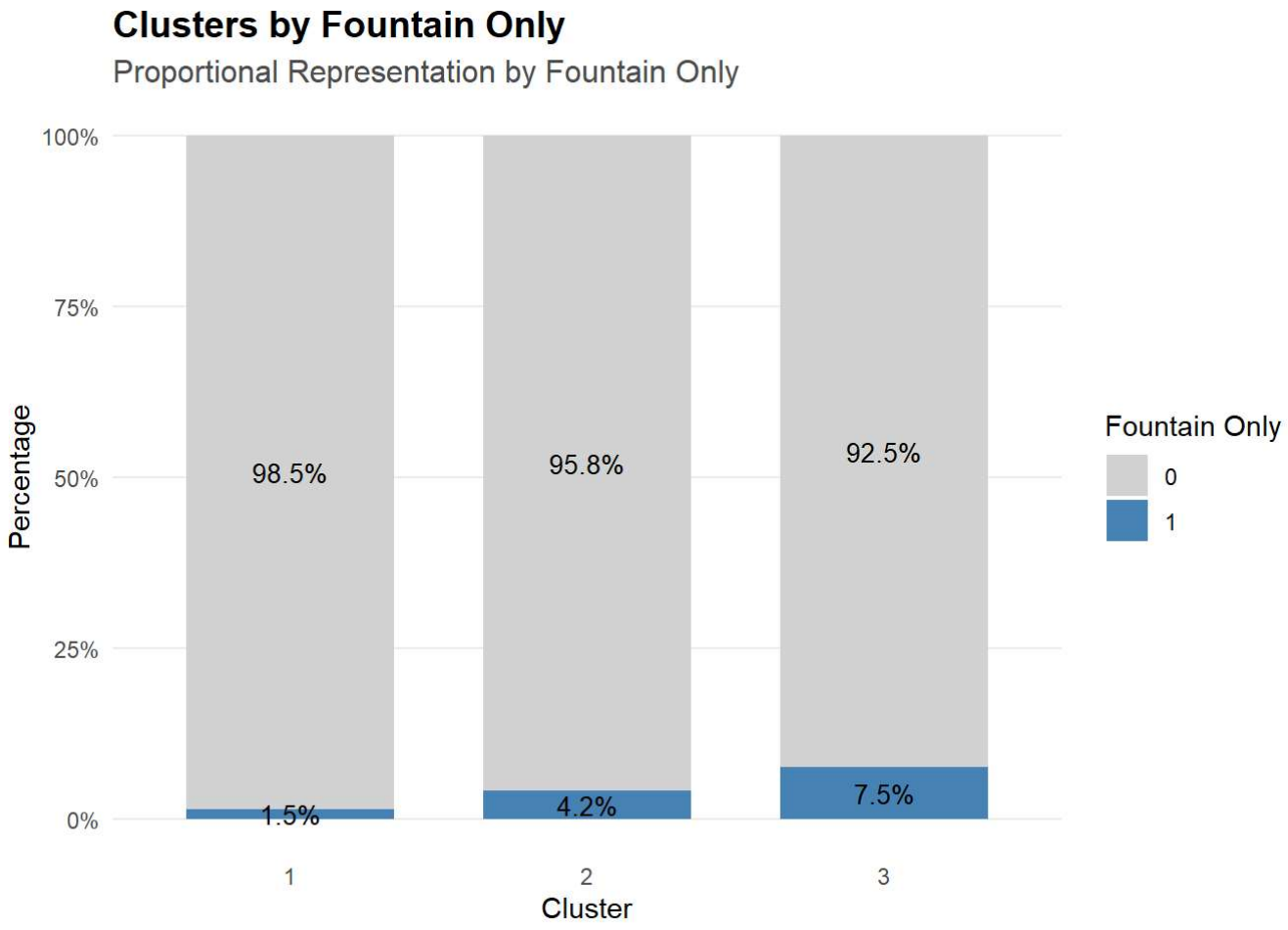
▶ Code


Fleet Type Distribution Across Clusters
Fleet type classification using a 400-gallon threshold
100%
75%
50%
25%
0%
Percentage
79.8%
20.2%
13.5%
86.5%
0.4%
99.6%
1
2
3
Cluster
Fleet Type
RED TRUCK
WHITE TRUCK

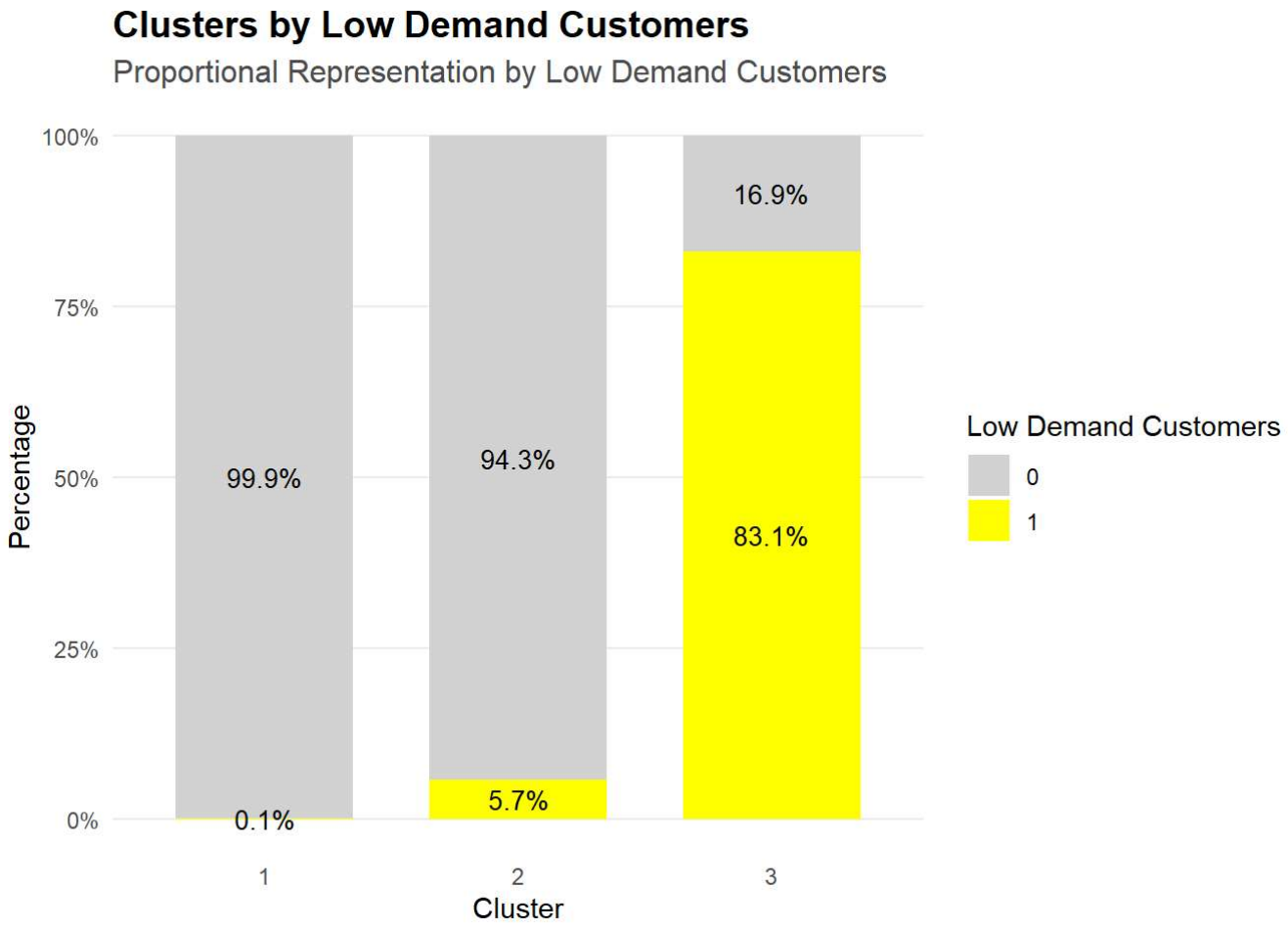
▶ Code


Clusters by Growth Potential
Proportional Representation by High Growth Potential
100%
75%
50%
25%
0%
Percentage
93.2%
83.4%
93.9%
6.8%
16.6%
6.1%
1
2
3
Cluster
Growth Potential
0
1

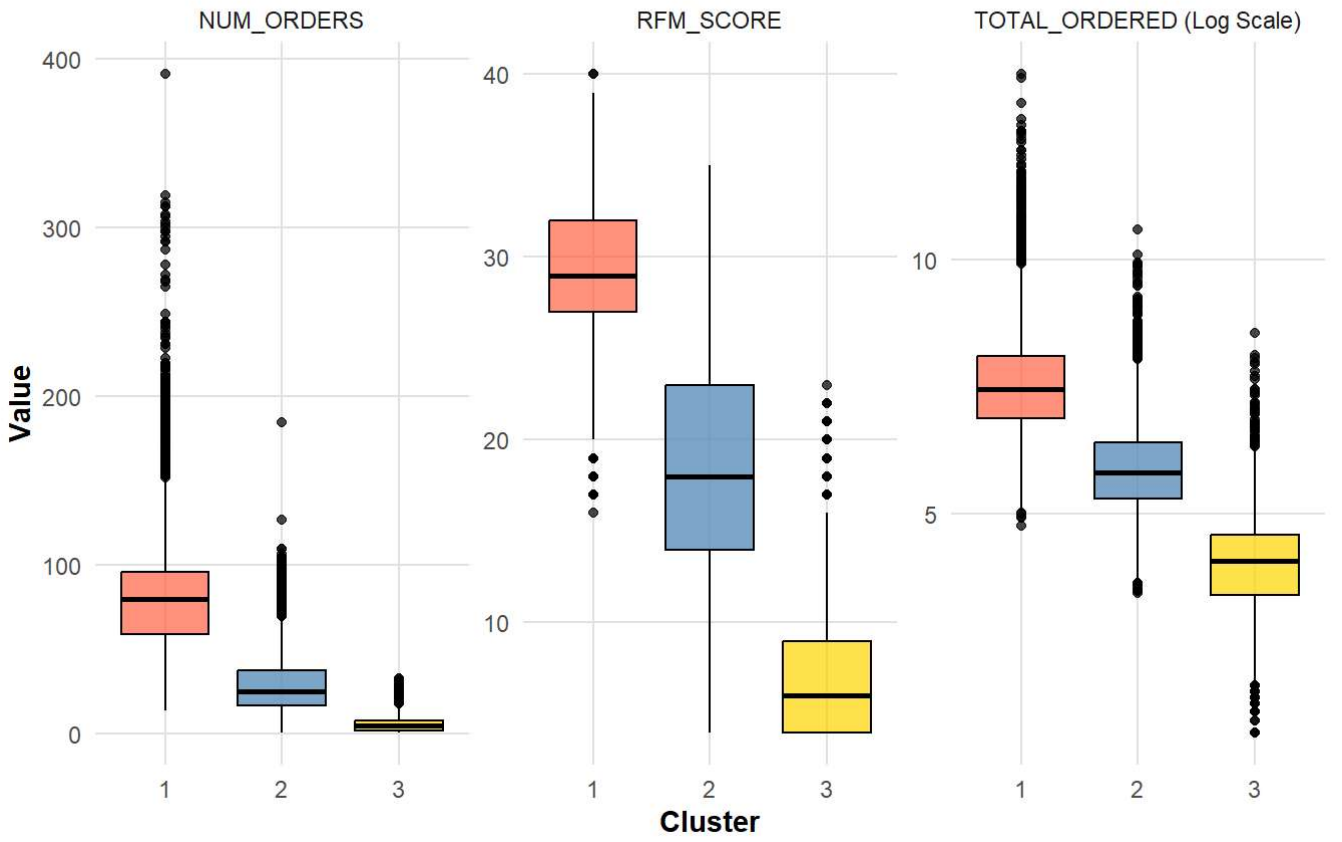
▶ Code


Clusters by Fountain Only
Proportional Representation by Fountain Only
100%
75%
50%
25%
0%
Percentage
98.5%
95.8%
92.5%
1.5%
4.2%
7.5%
1
2
3
Cluster
Fountain Only
0
1

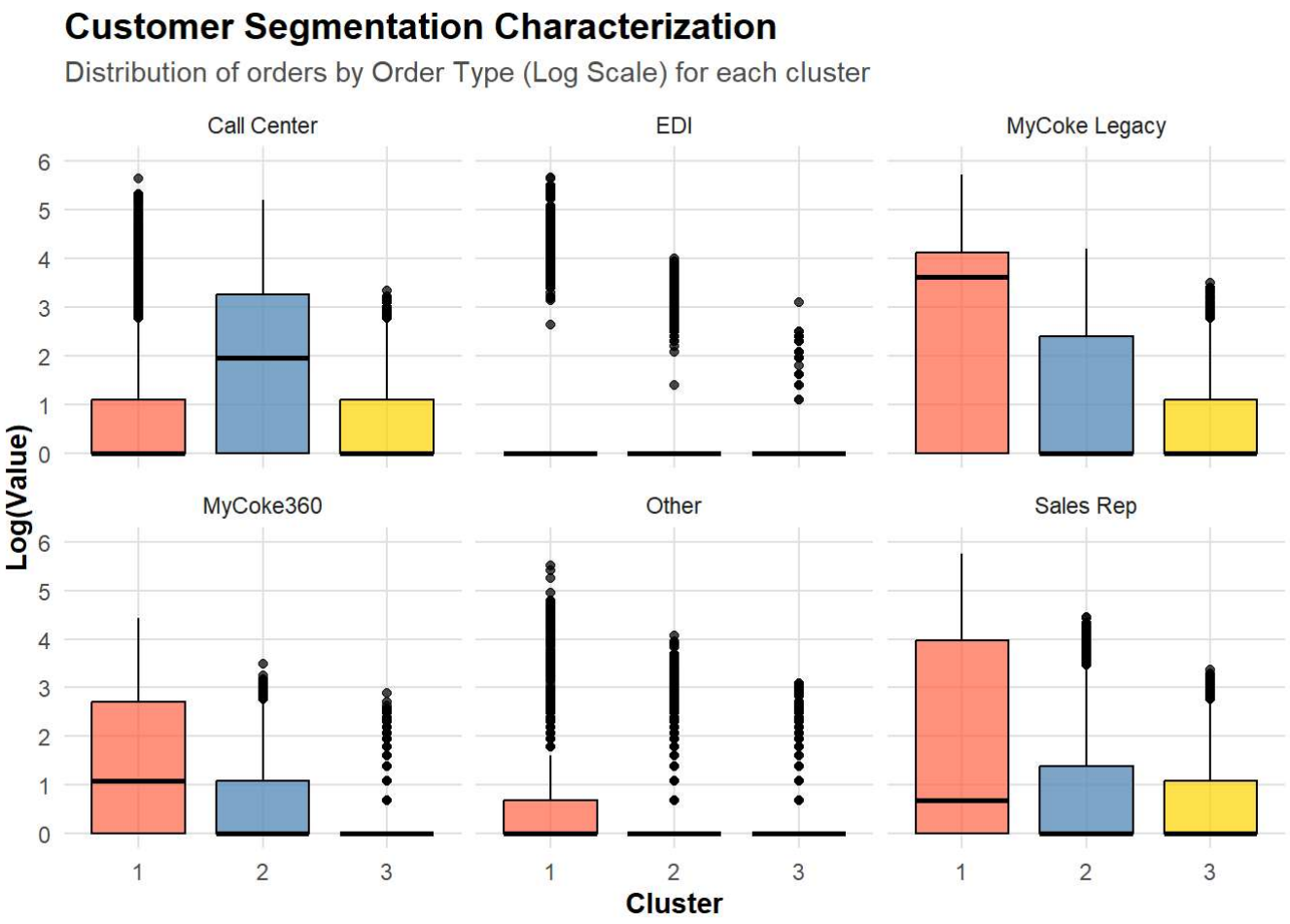
▶ Code


Clusters by Low Demand Customers
Proportional Representation by Low Demand Customers
100%
75%
50%
25%
0%
Percentage
99.9%
0.1%
94.3%
5.7%
16.9%
83.1%
1
2
3
Cluster
Low Demand Customers
0
1

▶ Code


Customer Segmentation Characterization
Distribution of RFM Score, Number of Orders, and Log-Transformed Total Ordered for each clus
NUM_ORDERS
RFM_SCORE
TOTAL_ORDERED (Log Scale)
Value
Cluster
0
100
200
300
400
10
20
30
40
5
10
1
2
3

▶ Code



Cluster 1 (Red): High Demand Customers

- Composition: Approximately 80% of customers receive deliveries via red trucks (based on the benchmark threshold of 400 gallons on average per year).
- Growth: Around 7% of customers exhibit high growth potential.
- Local Fountain Only: Only 1.5% of customers are local fountain-only.
- Average RFM: The average RFM score for this cluster is 29, the highest among the three clusters.
- Average Number of Orders: The average number of orders per customer was 81 in 2023 and 2024, with many outliers showing significantly higher order volumes.
- Total Ordered Volume: The average total ordered volume per customer in 2023 and 2024 is 4,638 gallons. This cluster has the highest number of outliers with elevated volumes, which skews the average. The volume representing the median is 1,707 gallons. - Volume Share: This cluster represents 76% of the total volume consumed in 2023 and 2024.
- It has the highest number of orders through digital channels and is the cluster most served by sales representatives.

Cluster 2 (Blue): Intermediate Customers with Growth Potential

- Composition: Approximately 87% of customers receive deliveries via white trucks (based on the benchmark threshold of 400 gallons on average per year).
- Growth: This cluster has the highest percentage of customers with high growth potential, at 16.6%.
- Local Fountain Only: Around 4.2% of customers are local fountain-only.
- Average RFM: The average RFM score for this group is 18.7, the second highest among the clusters.
- Average Number of Orders: The average number of orders per customer was 30 in 2023 and 2024.

- Total Ordered Volume: The average total ordered volume per customer in 2023 and 2024 is 525 gallons. The median volume is 331 gallons.
- Volume Share: This cluster represents approximately 22% of the total volume consumed in 2023 and 2024.
- It is the cluster with the highest average number of orders placed via the call center. It has fewer orders through digital channels compared to Cluster 2, but more than Cluster 1. The number of orders through sales representatives is similar to Cluster 1

Cluster 3 (Yellow): Less Active Customers with Low Order Volume
- Composition: Only 0.4% of customers receive deliveries via red trucks (based on the benchmark threshold of 400 gallons on average per year).
- Growth: Approximately 6% of customers exhibit high growth potential.
- Local Fountain Only: This cluster has the highest percentage of local fountain-only customers, at 7.5%.
- Average RFM: The average RFM score is 7, indicating these are the least active customers.
- Average Number of Orders: The average number of orders per customer was 5.5 in 2023 and 2024.
- Total Ordered Volume: The average total ordered volume per customer in 2023 and 2024 is around 80 gallons, while the median is 57 gallons, indicating a large number of customers with smaller volumes.
- Volume Share: This cluster represents only 1.7% of the total volume consumed in 2023 and 2024.
- The cluster shows orders concentrated through call centers, digital channels, and sales representatives, although in smaller absolute quantities compared to the other clusters.

8. Classification Models for Explaining Clusters

To better understand the variables influencing cluster composition and facilitate future predictions without the need for re-clustering, two classification models will be used: decision trees and multinomial logistic regression. These models will help identify the key characteristics that drive cluster formation.

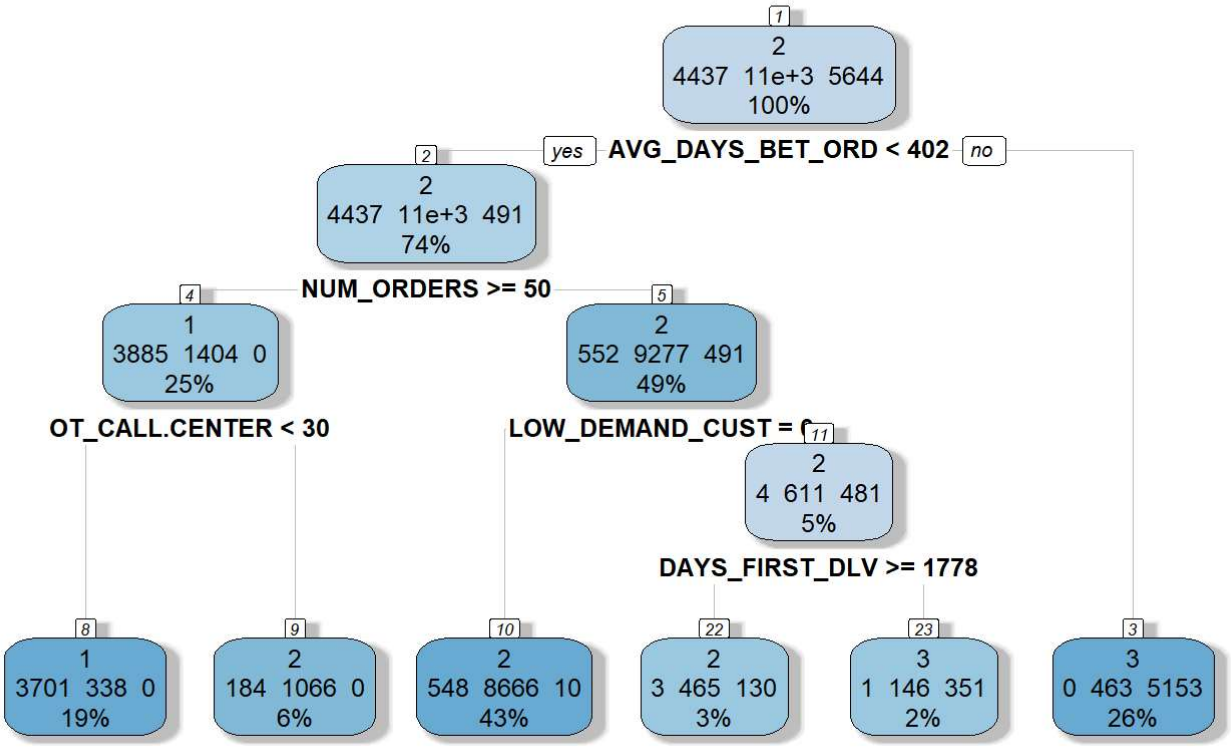
By applying these models to new data, cluster assignments can be predicted, streamlining the analysis process and eliminating the need to recreate the clusters whenever new data is introduced.

8.1 Decision Tree

The selected variables will be analyzed to explain the clusters using a decision tree, with the dataset split into training and test sets, applying 20-fold cross-validation.

▶ Code

Decision Tree: Explaining Customer Clusters



Below are the prediction performance metrics:

▶ Code

```
--- Decision Tree Model Performance ---
```

▶ Code

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1568  170    0
         2  332 4336   72
         3    1  270 2346

Overall Statistics

               Accuracy : 0.9071
                 95% CI : (0.9009, 0.913)
    No Information Rate : 0.5251
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8477
```

```
 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8248   0.9079   0.9702
Specificity            0.9764   0.9065   0.9594
Pos Pred Value         0.9022   0.9148   0.8964
Neg Pred Value         0.9547   0.8990   0.9889
Prevalence             0.2090   0.5251   0.2659
Detection Rate         0.1724   0.4767   0.2579
Detection Prevalence   0.1911   0.5212   0.2877
Balanced Accuracy      0.9006   0.9072   0.9648
```

▶ Code

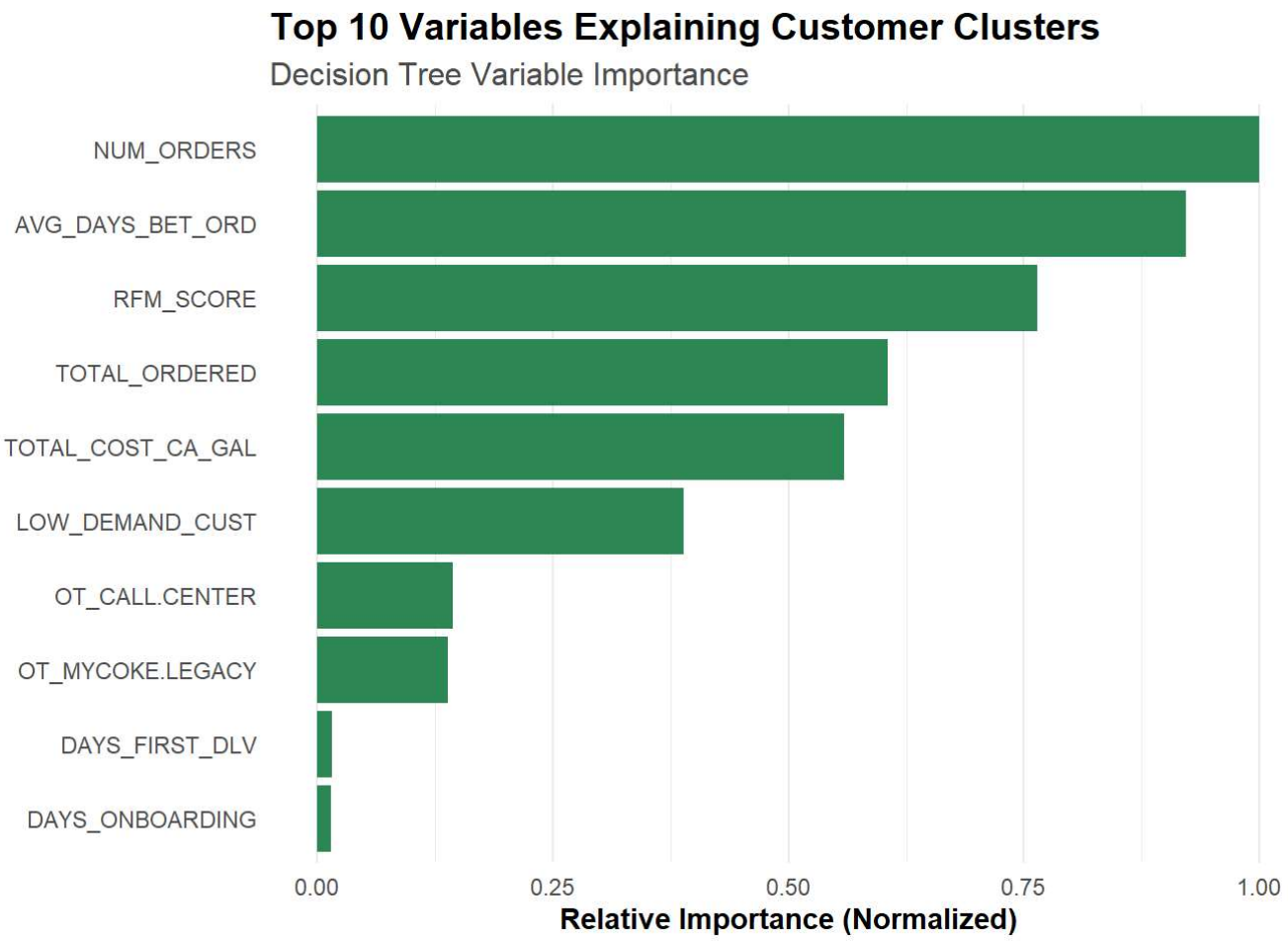
Decision Tree Accuracy Comparison (Train vs Test)

Dataset	Accuracy
Train	0.91
Test	0.91

The model has an accuracy of 91% on both the train and test sets, demonstrating strong performance across all clusters. In Cluster 1: High Demand Customers, precision is 90.2% and recall is 82.5%. For Cluster 2: Intermediate Customers with Growth Potential, precision is 91.5% and recall is 90.8%. For Cluster 3: Less Active Customers with Low Order Volume, precision is 89.6% and recall is 97.0%.

Overall, the model performs well across all clusters, with strong precision and recall values for Cluster 1 and Cluster 3, and solid performance in Cluster 2. The accuracy comparison between the train and test sets is identical at 91%, indicating good generalization.

▶ Code



The most important variables in the model were the number of orders per customer, average days between orders, RFM score, total ordered volume, total cost, and the low demand customers flag.

8.2 Multinomial Logistic Regression

The influence of the selected variables on customer clusters will be explored using multinomial logistic regression to predict the probabilities of new customers belonging to each of the established clusters. This method is well-suited for modeling the relationship between the predictors and the probabilities of customers being assigned to one of the three clusters, helping to assess the likelihood of a customer belonging to each specific group based on their characteristics.

Variable standardization and Elastic Net regularization will be used in the model development process.

▶ Code

```
glmnet

21225 samples
   21 predictor
    3 classes: '1', '2', '3'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19102, 19103, 19102, 19103, 19103, 19102, ...
```

```
Resampling results across tuning parameters:

  lambda  Accuracy   Kappa
  0.100   0.8948407  0.8217432
  0.325   0.7363958  0.5030291
  0.550   0.5250412  0.0000000
  0.775   0.5250412  0.0000000
  1.000   0.5250412  0.0000000

Tuning parameter 'alpha' was held constant at a value of 0.5
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were alpha = 0.5 and lambda = 0.1.
```

▶ Code

```
 Accuracy     Kappa
0.8931281 0.8189136
```

▶ Code

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1 1346    2    0
         2  555 4575  216
         3    0  199 2202

Overall Statistics

               Accuracy : 0.8931
                 95% CI : (0.8866, 0.8994)
    No Information Rate : 0.5251
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8189

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7080   0.9579   0.9107
Specificity            0.9997   0.8215   0.9702
Pos Pred Value         0.9985   0.8558   0.9171
Neg Pred Value         0.9284   0.9464   0.9677
Prevalence             0.2090   0.5251   0.2659
Detection Rate         0.1480   0.5030   0.2421
Detection Prevalence   0.1482   0.5878   0.2640
Balanced Accuracy      0.8539   0.8897   0.9404
```

▶ Code

Multinomial Logistic Regression Accuracy Comparison (Train vs Test)

Dataset	Accuracy
Train	0.89
Test	0.89

The model achieved an accuracy of 89.3% on the test set, reflecting strong performance. In Cluster 1 (Red): High Demand Customers, recall is 70.8% and precision is 99.8%. For Cluster 2 (Blue): Intermediate Customers with Growth Potential, recall is 95.8% and precision is 85.6%. Finally, Cluster 3 (Yellow): Less Active Customers with Low Order Volume shows recall of 91 % and precision of 91.7%. Overall, the model performs well, with Cluster 2 showing the highest recall and Cluster 1 having the strongest precision.

The relatively low recall in Cluster 1 (Red) (70.8%) suggests that the model may not always correctly identify customers in this group, leading to false negatives.

▶ Code

The model indicates that:

For **Cluster 1**, the key variables included the number of orders, RFM score, order type (MyCoke Legacy), order type (MyCoke 360), order type (using sales representatives), and chain member.

For **Cluster 2**, the most significant variables were the average number of days between orders, low demand customers, order type (call center), order type (MyCoke Legacy), and channel growth potential.

For **Cluster 3**, the most important variables were the average number of days between orders, low demand customers, RFM score, and days since the first delivery.

The models created to predict clusters for new customers performed well and provide insights that clearly help in understanding the characteristics influencing the clusters. Therefore, we can proceed with the final analysis for fleet assignment.

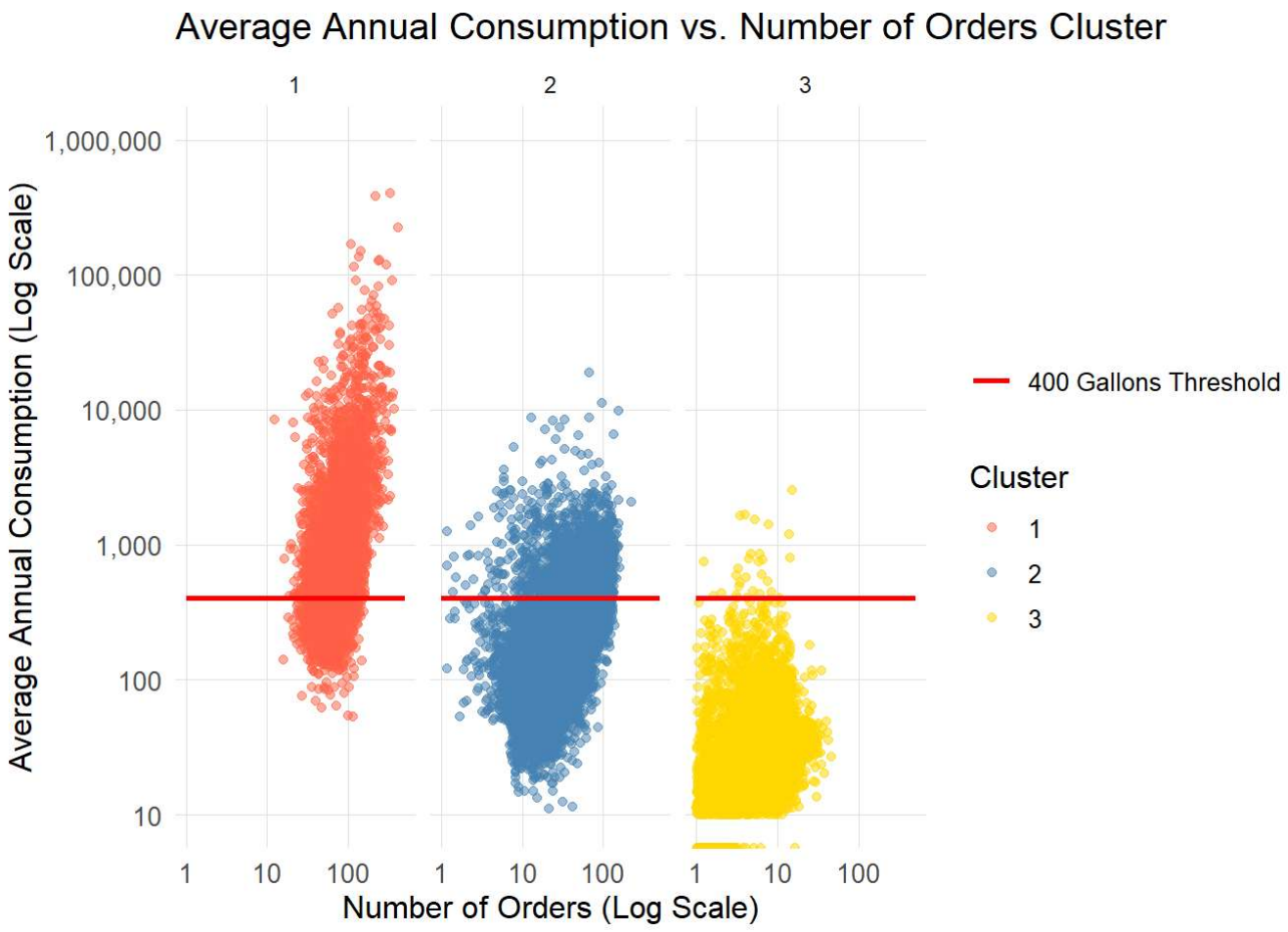
9. Data driven fleet assingment

Based on all the previous analyses, it is concluded that the fleet type designated for clients should be defined by considering different criteria, not just the average annual volume demand.

The main criteria shaping this approach include the similarities among clients represented by the clusters, the analysis of volume distributions by cold drink channel segment, and the growth potential of the clients.

Before proceeding, the relationship between the 400 gallons annual threshold for each cluster will be analyzed.

▶ Code



Regarding the 400-gallon benchmark for defining clients to be served by red trucks, it is possible to note that:

- Cluster 1: This cluster mainly selects clients with higher demand volumes or a larger number of orders. A smaller portion of these clients would fall below the 400-gallon threshold, with some still close to a minimum of 100 gallons.
- Cluster 2: This cluster has large number of clients above and below the threshold, so it requires further refinement.
- Cluster 3: The vast majority of clients fall below the threshold. However, the few clients above it tend to place a small number of orders per year.

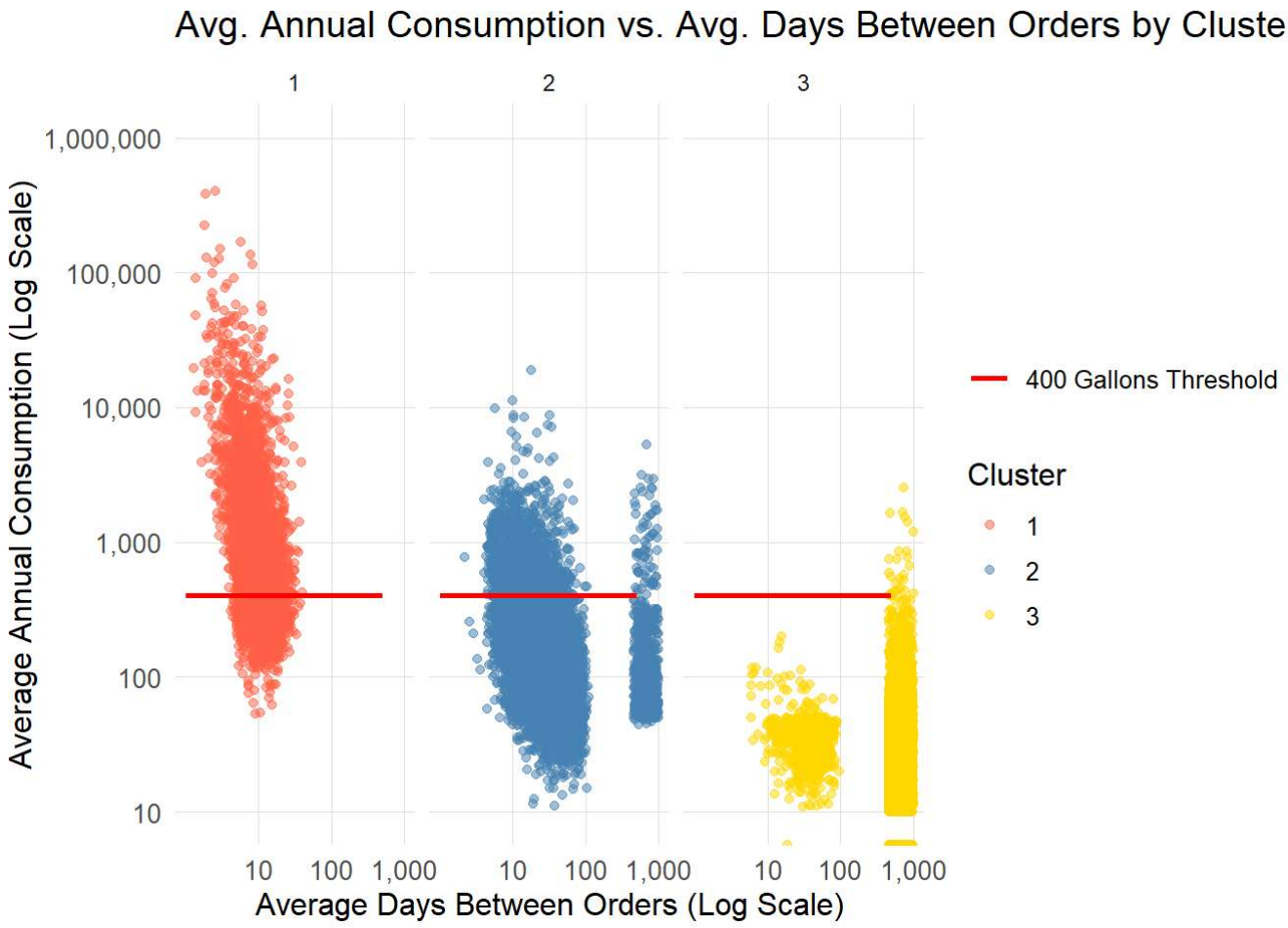
9.1 Cluster 2 Analysis for Fleet Assignment

Cluster 2 comprises just over half of all clients, making it difficult to define clear criteria for fleet designation.

The multinomial regression model indicated that the variable “Average Days Between Orders” (AVG_DAYS_BET_ORD) was the most important, while in the decision tree model, it was the second most important variable.

Therefore, below is the plot showing the relationship between the average annual consumption of each client and their average days between orders.

▶ Code



▶ Code

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Lower_Bound	3	11	14	17	20	24	28	33	40	52
Upper_Bound	11	14	17	20	24	28	33	40	52	731

▶ Code

	Statistic	Value
1	Min.	3
2	1st Qu.	16
3	Median	24
4	Mean	56
5	3rd Qu.	37
6	Max.	731

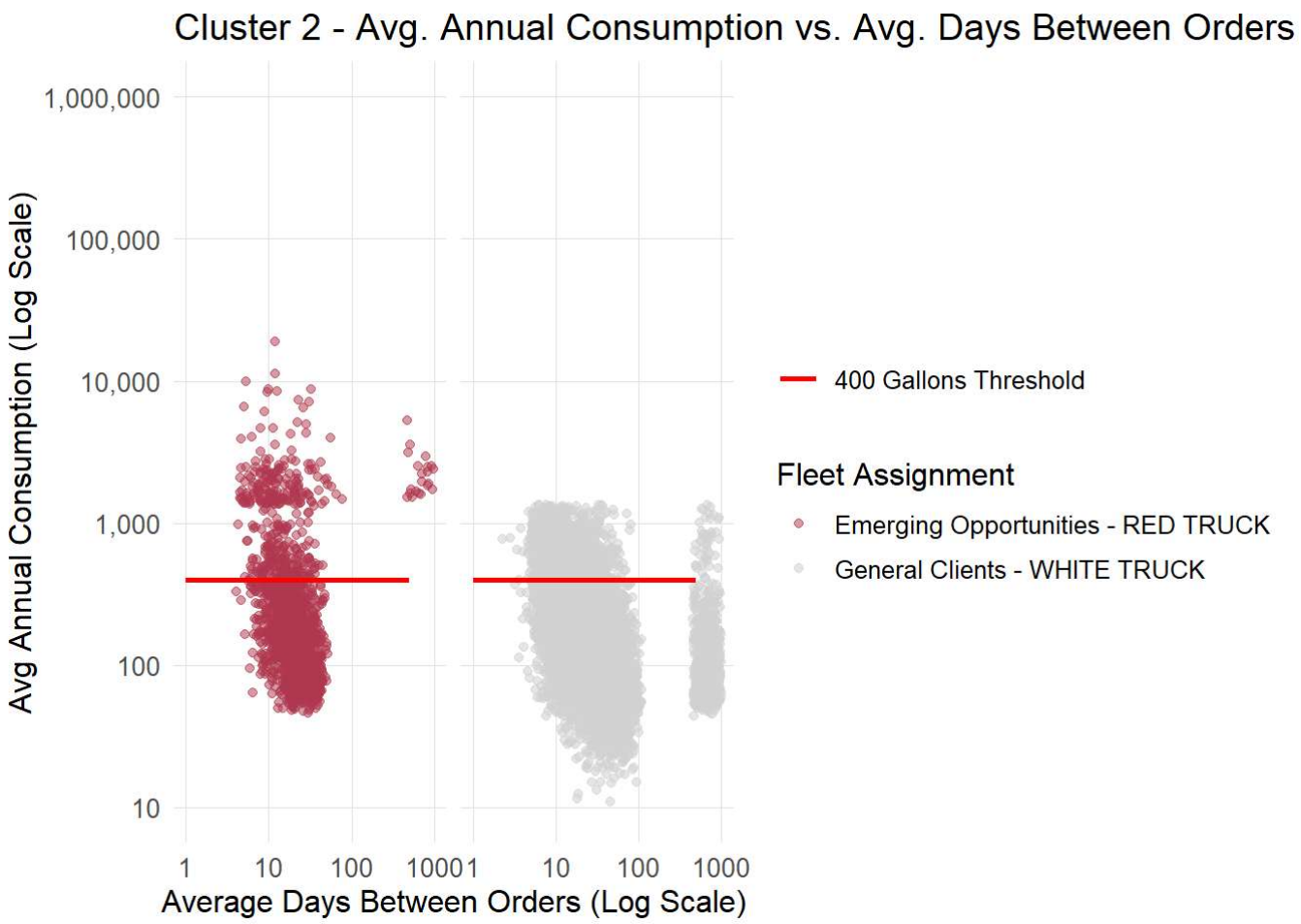
When filtering the average days between orders for Cluster 2, it is observed that 60 percent of customers have an average of 33 days or fewer between orders. The group's average is 56.4 days, with a median of 24 days.

Building upon the previously calculated variables, low demand customers and high growth potential customers, additional criteria relevant to the business will be introduced to better segment customers within Cluster 2.

These new criteria include an average annual consumption greater than 1,349 gallons and an average of 52 or fewer days between orders. The first threshold was chosen because it represents the point at which delivery costs are minimized. The second threshold was selected due to its significant influence on the clustering model, and because customers with high growth potential (excluding low demand customers) and an average time between orders of 33 days or fewer—representing nearly two-thirds of customers—are believed to have the potential to order more frequently, thus reducing the order interval.

As a result, in the plot below, customers who are not low demand, show high growth potential, or have an average annual consumption greater than 1,349 gallons and an average of 33 or fewer days between orders will be classified as Emerging Opportunities and assigned to the red truck category.

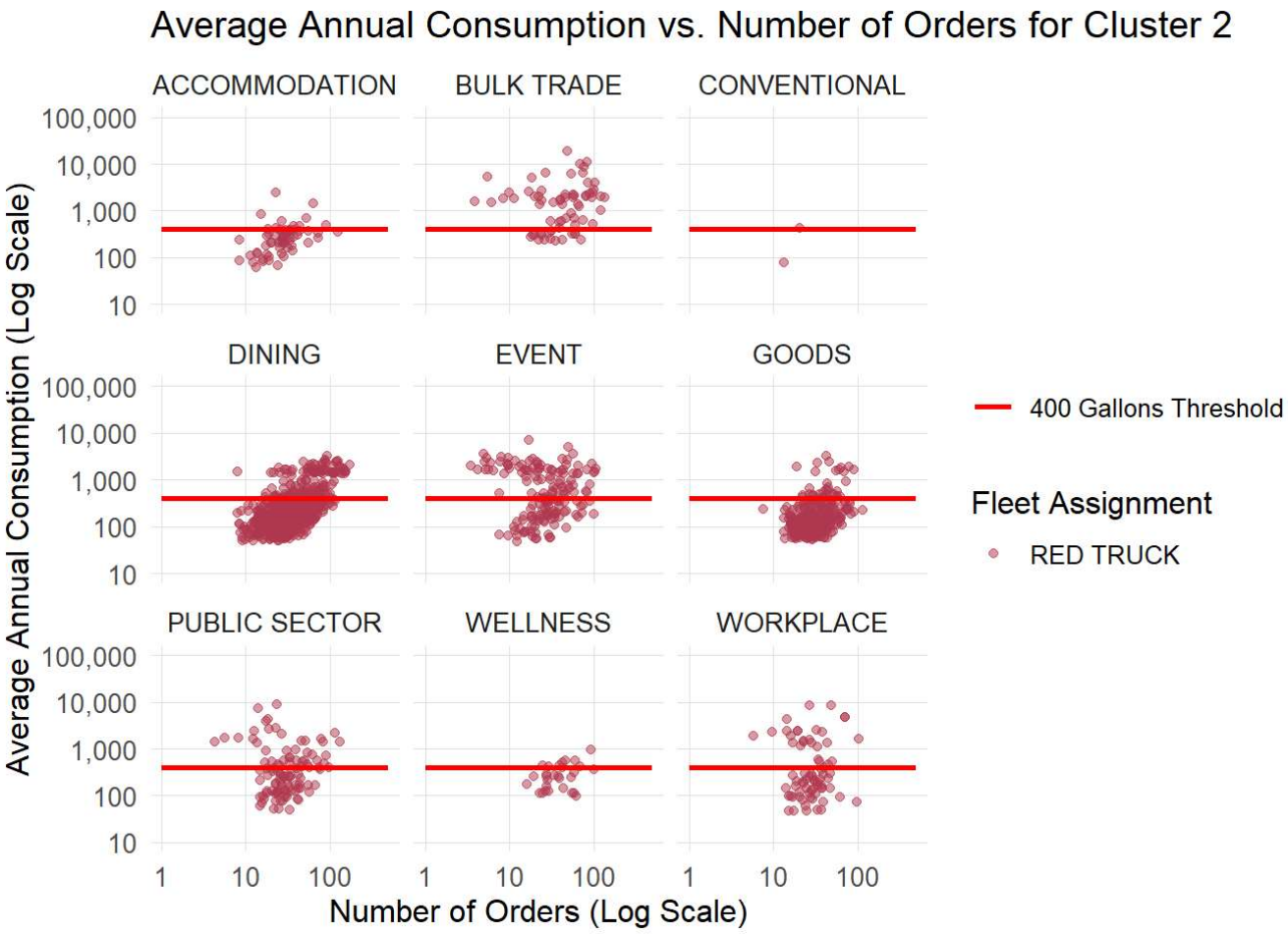
▶ Code



▶ Code

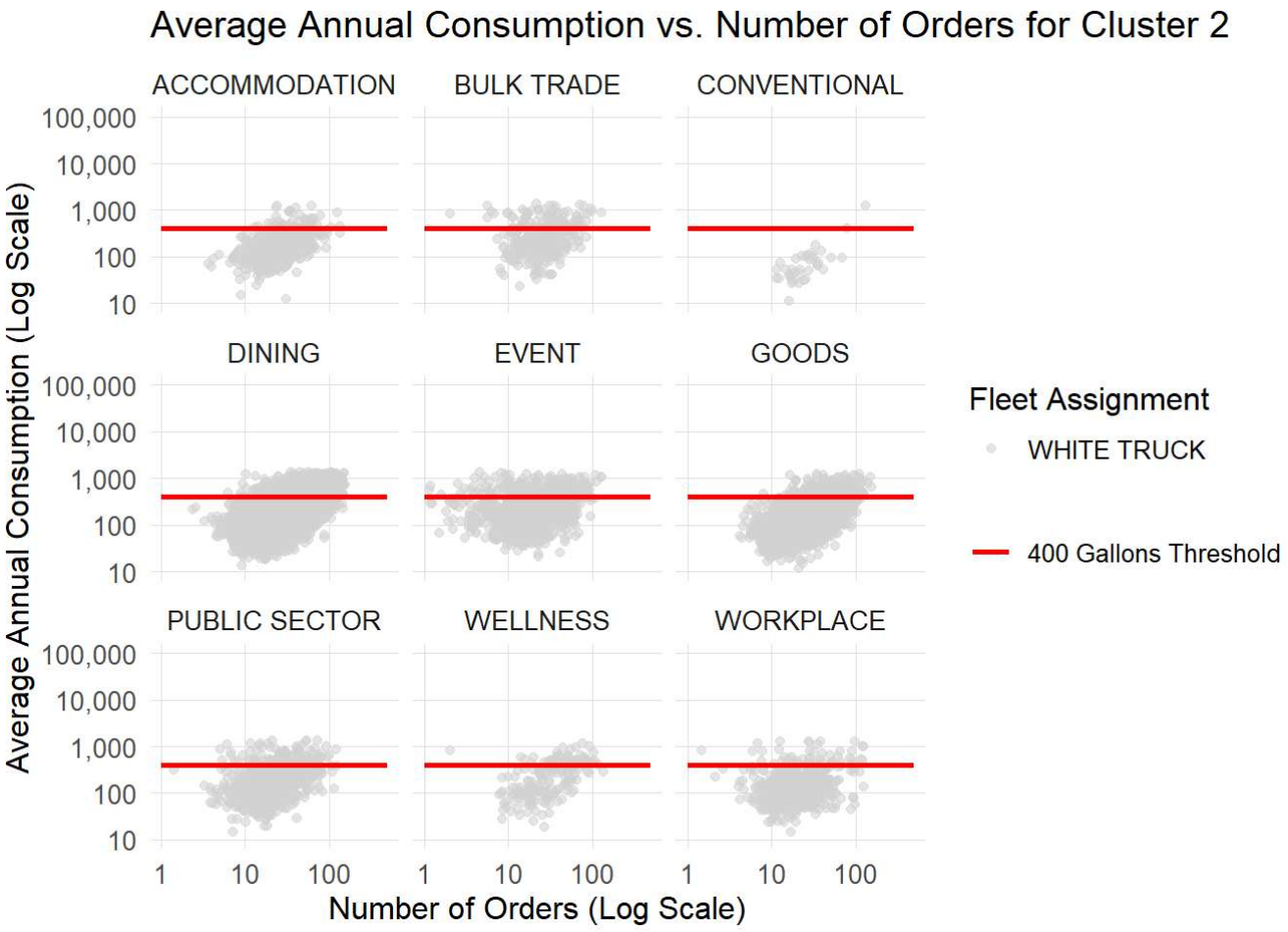
Below, the impact of fleet assignment on each cold drink channel will be explored using the previous criteria for Cluster 2, and its relation to average annual consumption and the number of orders will be analyzed.

▶ Code



In an effort to explore growth opportunities, almost all sectors would have a considerable number of clients with a volume of less than 400 gallons but using red trucks.

▶ Code



On the other hand, the criteria naturally assign white trucks to a large number of clients with an average annual volume of less than 400 gallons in each segment, while still ensuring that high-volume clients are served by red trucks.

Also, the previous graphs represent an opportunity for the company to develop targeted strategies for each segment.

9.2 Fleet Assignment Criteria

Based on the analysis, the recommended fleet assignment will be determined by the following criteria:

1. Customers with an average annual consumption greater than **1349 gallons** will be assigned to **RED TRUCKS**.
2. **Low-demand customers** (identified by `LOW_DEMAND_CUST == 1`) will be assigned to **WHITE TRUCKS**.
3. All customers in **Cluster 1** will be assigned to **RED TRUCKS**, according to the previous rules.
4. All customers in **Cluster 3** will be assigned to **WHITE TRUCKS**, after applying the previous rules.
5. **Customers in Cluster 2** will be assigned to **RED TRUCKS** if they meet at least one of the following conditions:
    - They are classified as **high growth potential** (`HIGH_GROW_POT == 1`).

- Their **average days between orders** are less than or equal to 33 (`AVG_DAYS_BET_ORD <= 33`).

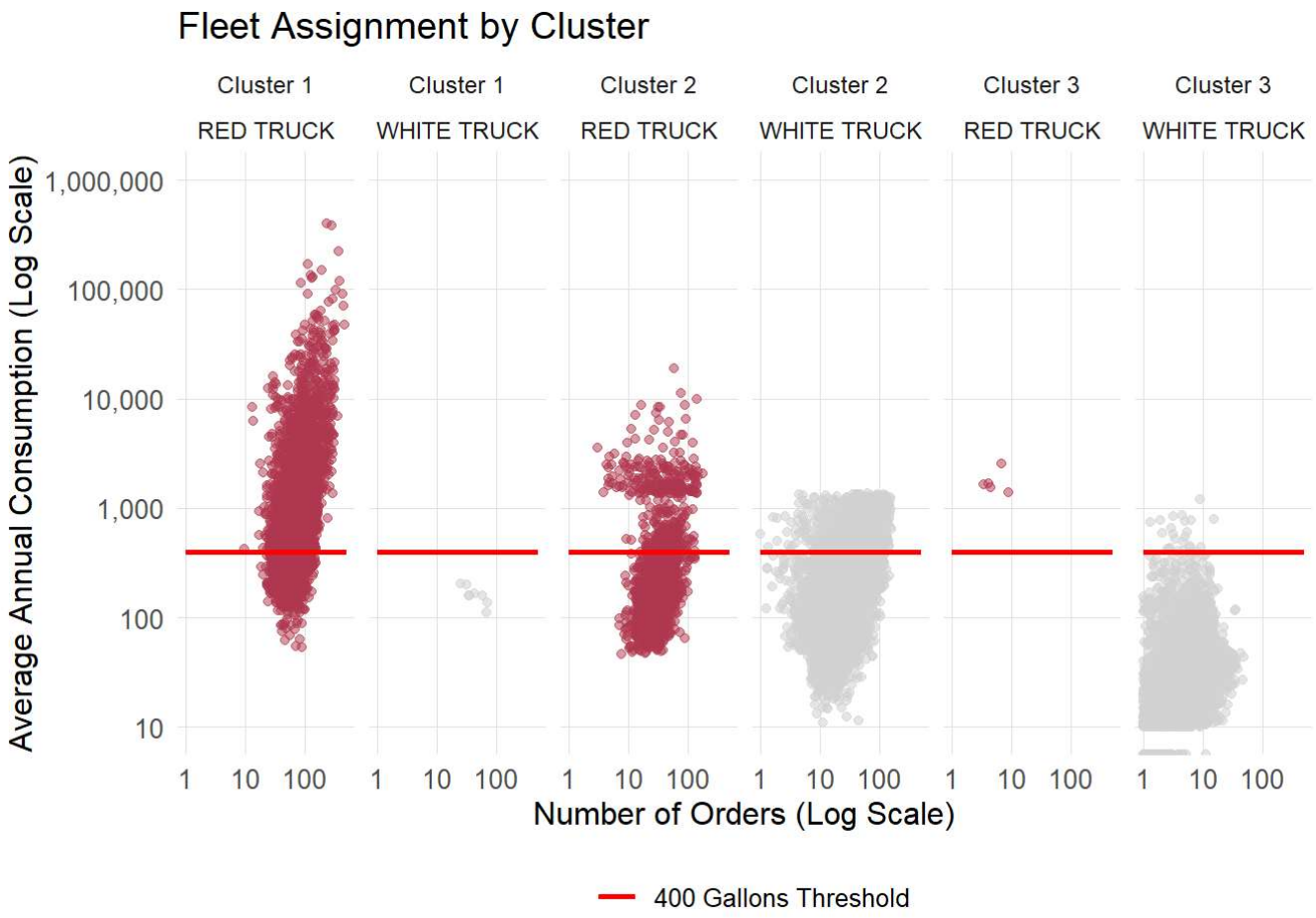
6. The remaining customers in **Cluster 2** will be assigned to **WHITE TRUCKS**.

7. Any customers who do not meet any of these criteria will remain unclassified (`NA`).

▶ Code

Below are the representations of the clusters and the designated fleet.

▶ Code



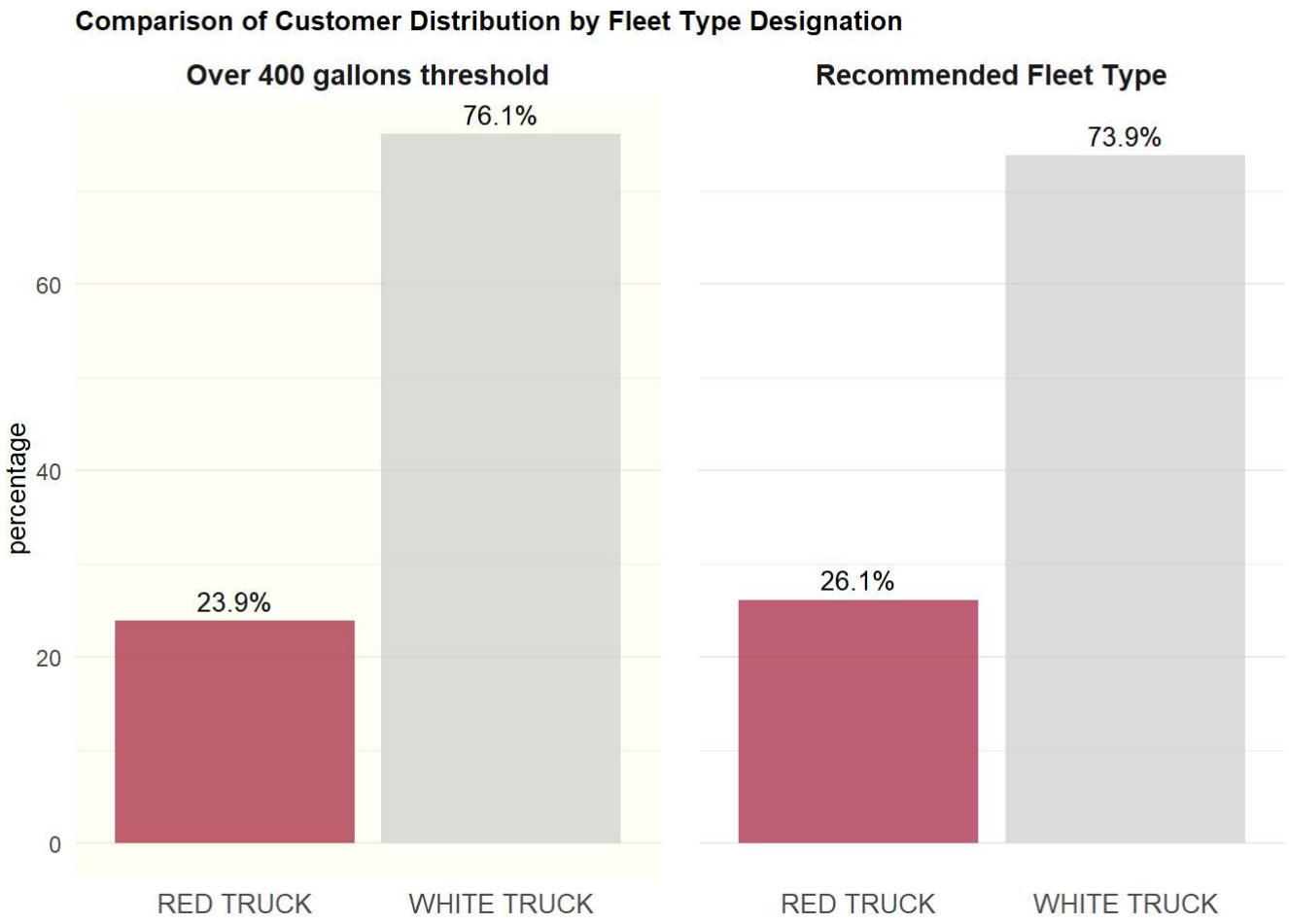
▶ Code

The new criteria established labels for all customers. A total of 7,926 customers were assigned to “Red Truck”, while 22,394 customers were assigned to “White Truck”.

The annual average consumption criterion of 400 gallons would have assigned 7,239 customers to be served by “Red Truck” and 23,081 customers to be served by “White Trucks”.

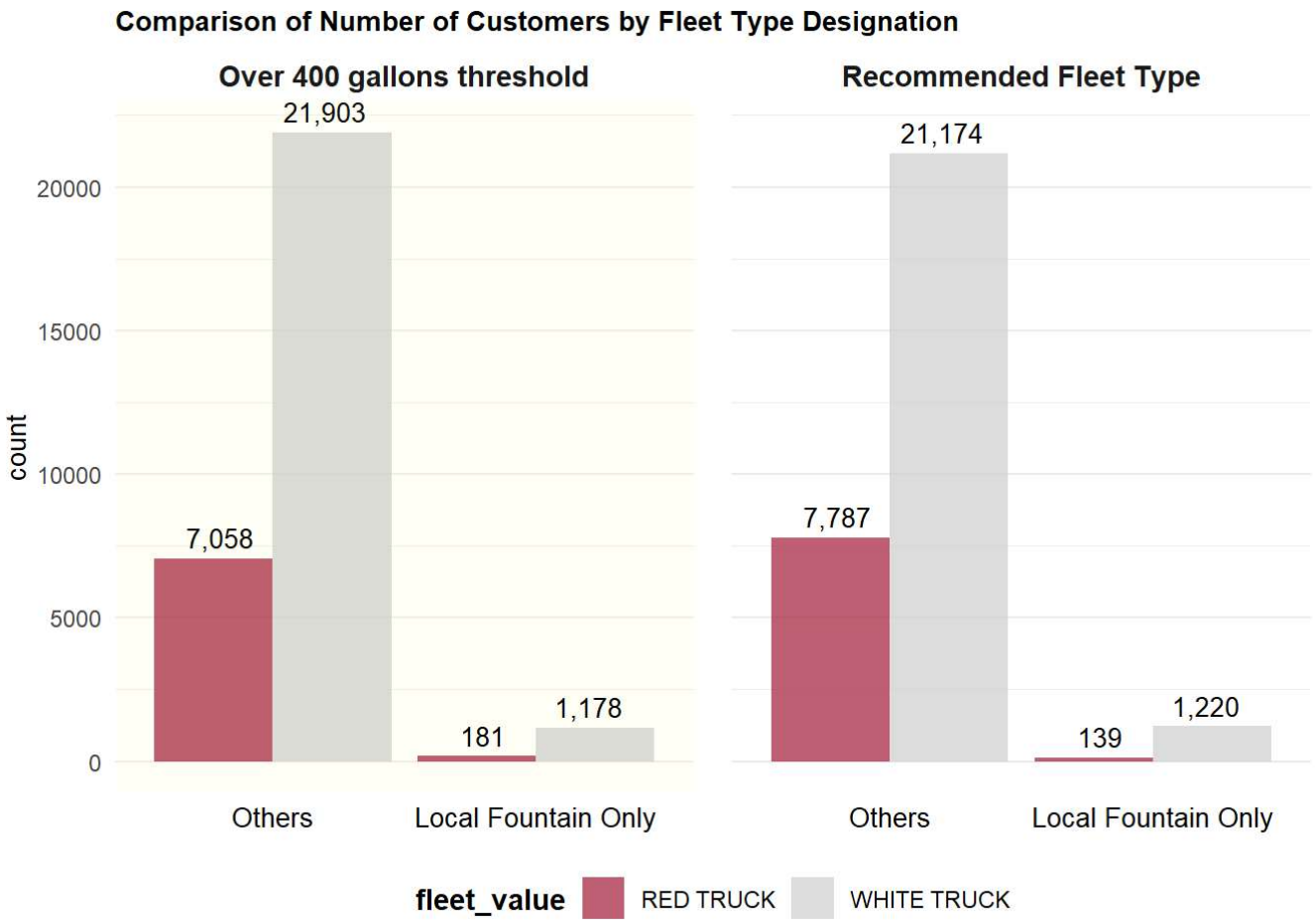
Therefore, 687 clients who were previously served by white trucks and who present higher growth potential will now be served by red trucks.

▶ Code



According to the criteria, 26% of customers would be served by red trucks and 74% by white trucks.

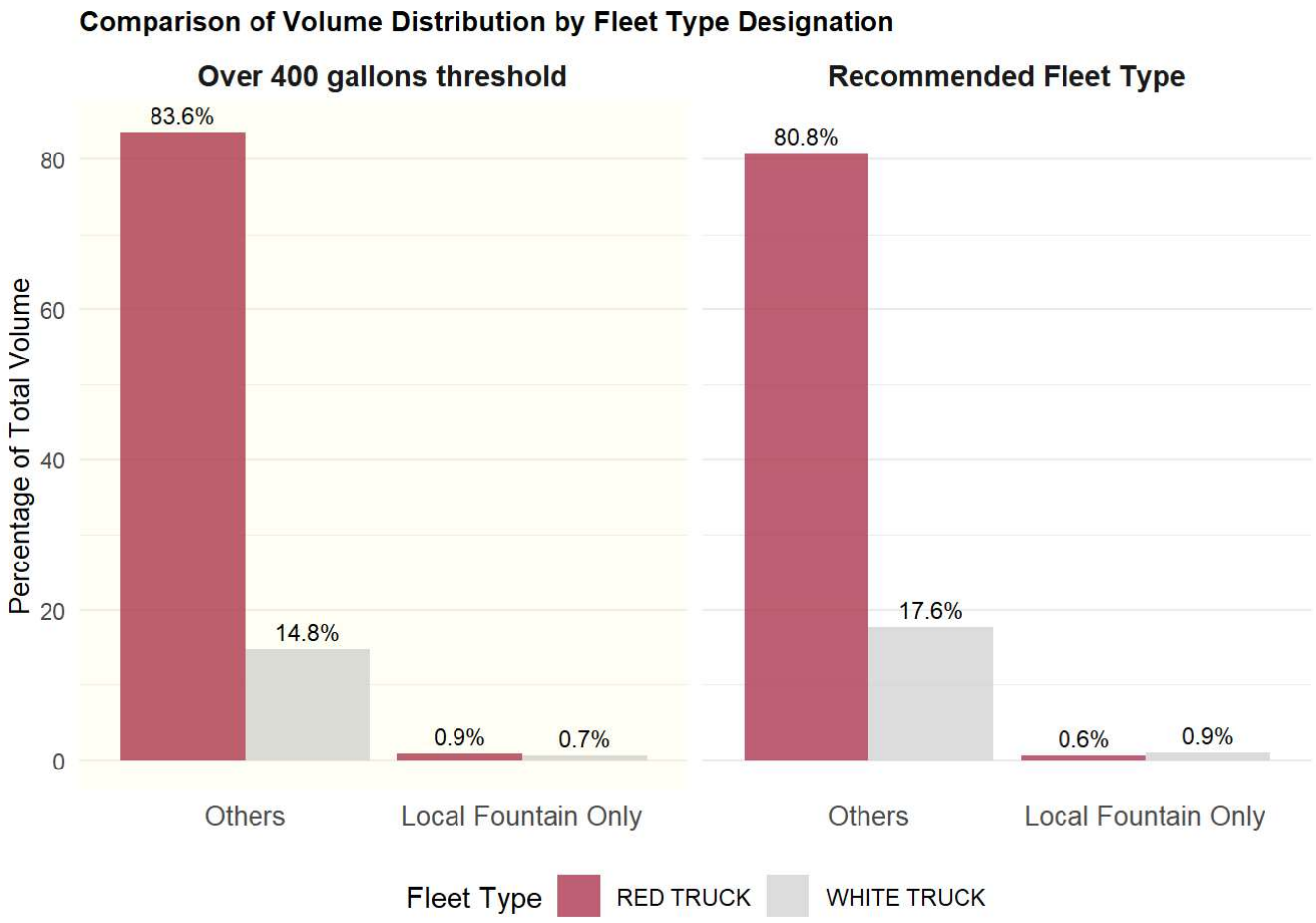
▶ Code



Considering only the "Others" group (customers who order from multiple sources), our recommendation would result in 729 additional stores being served by red trucks, compared to the 400-gallon threshold—an increase of 10.3%.

In contrast, within the 'Local Fountain Only' group, the number of customers served by red trucks would decrease by 42, representing a 23.2% reduction.

▶ Code



Although the number of customers served by red trucks has increased, the overall volume transported remains relatively stable.

Within the “Others” customer group, there would be a reduction of approximately 1,038,637 gallons over two years, representing a 3.4% decrease. This volume would now be delivered by white trucks.

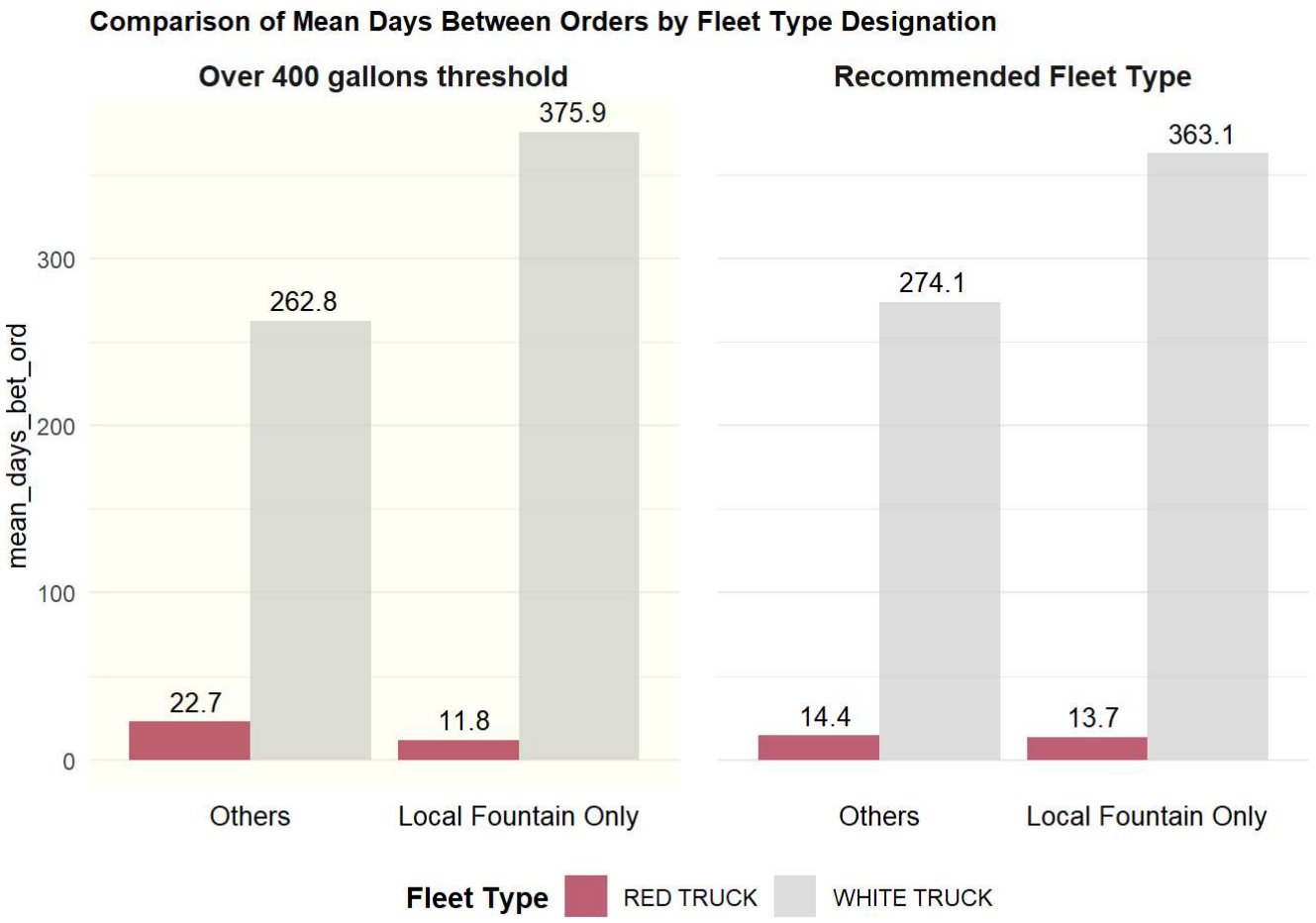
For the “Local Fountain Only” group, the reduction in volume transported by red trucks is around 104,895 gallons over two years a 31% decrease.

Despite the increase in the number of customers served by red trucks, which may lead to higher travel times and costs, the recommendation optimizes the delivery system by allowing red trucks to focus on strategic customers while reducing overall costs through higher-volume deliveries using white trucks.

A geographic distribution analysis of the customer base can be carried out at a later stage. One opportunity that emerges from this recommendation is to encourage customers within the same ZIP code to coordinate delivery dates. This would help consolidate volumes, streamline the delivery process, and further reduce operational costs.

Below is the average number of days between orders for each group.

▶ Code



The red trucks should be optimized to serve the "Others" group, which has an average order interval of 14 days, compared to 23 days under the 400-gallon threshold model. The difference for the "Local Fountain Only" group in relation to the white trucks would be approximately 2 days.

The white trucks, on the other hand, would serve more sporadic customers, with an average interval of over 260 days between orders.

10. Recommendation Impacts

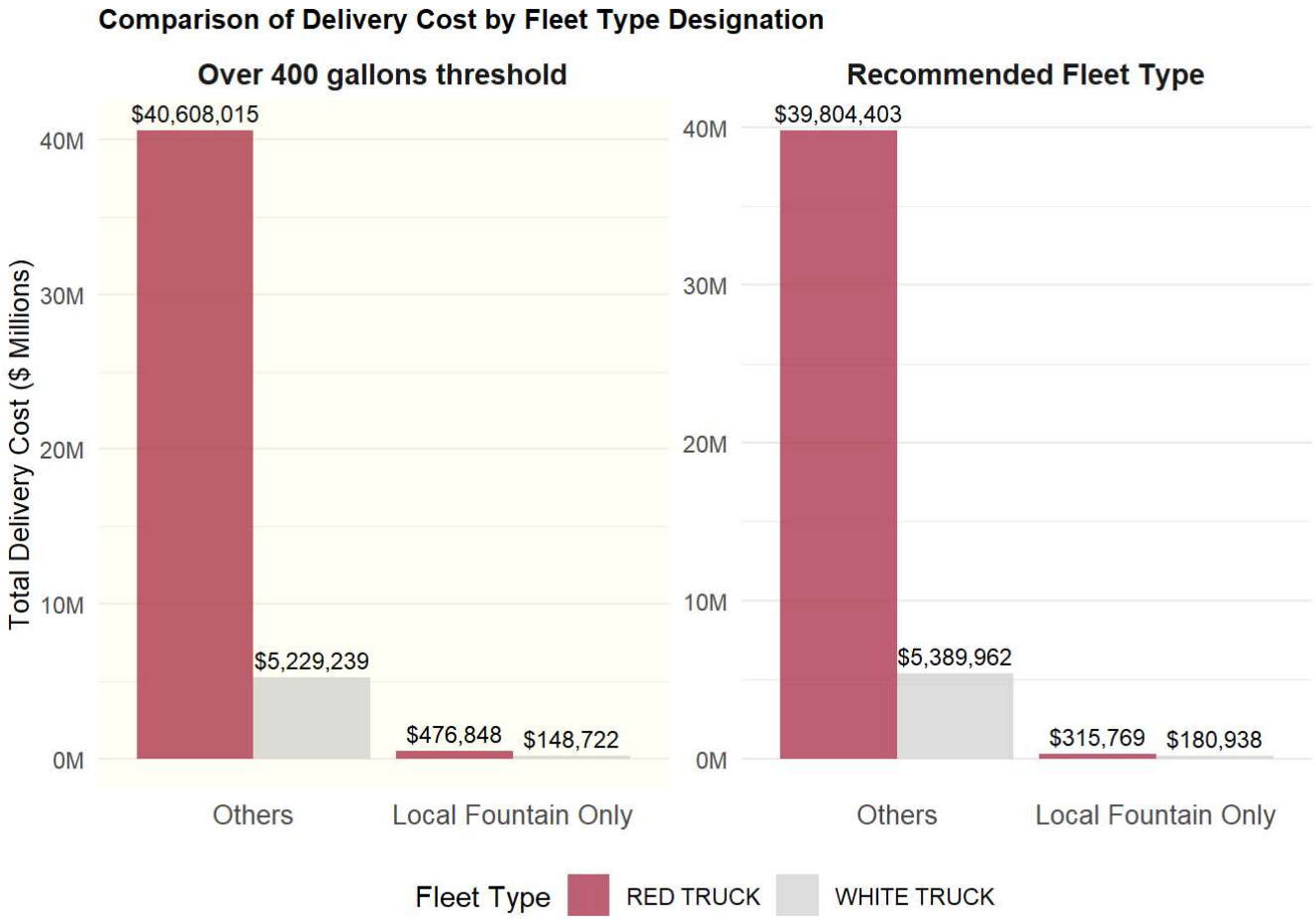
10.1 Impact on Costs

The cost impact of using red trucks is significantly higher compared to white trucks. For OPEX, the delivery cost for red trucks is approximately 700% more than for white trucks when considering only variable costs.

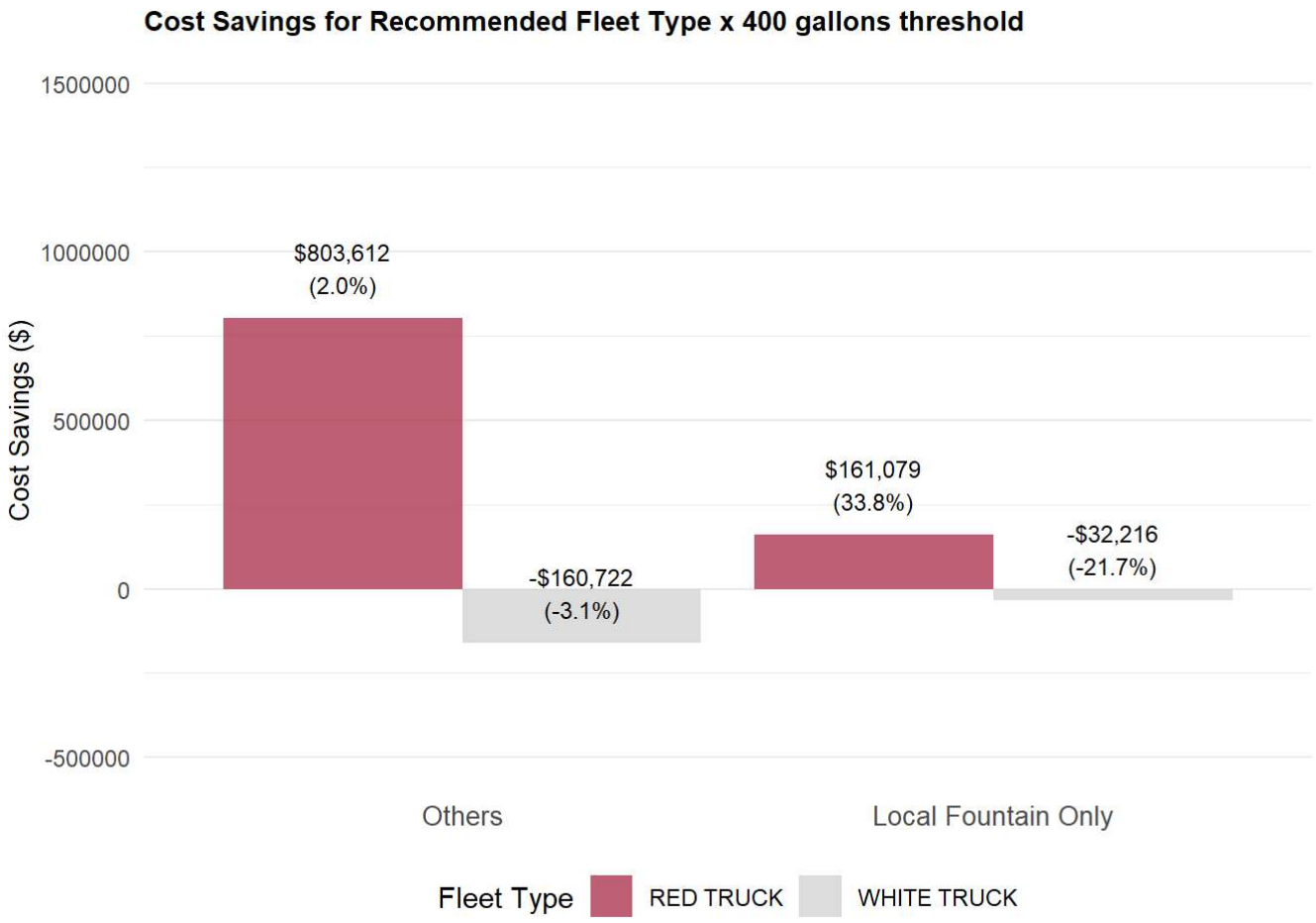
The calculated cost for the total volume delivered to each customer via red trucks is represented in the column `total_cos_ca_gal`. To provide conservative estimates, a 400% difference is assumed, and the red truck cost is divided by 5 to estimate the cost for white trucks, represented by `ARTM_TOTAL_COST`.

Below is the cost comparison.

▶ Code


Comparison of Delivery Cost by Fleet Type Designation
Over 400 gallons threshold
Recommended Fleet Type
Total Delivery Cost ($ Millions)
40M
30M
20M
10M
0M
$40,608,015
$5,229,239
$476,848
$148,722
$39,804,403
$5,389,962
$315,769
$180,938
Others
Local Fountain Only
Fleet Type
RED TRUCK
WHITE TRUCK

▶ Code



Regarding the 400-gallon threshold, over a two-year period the estimated differences would be:

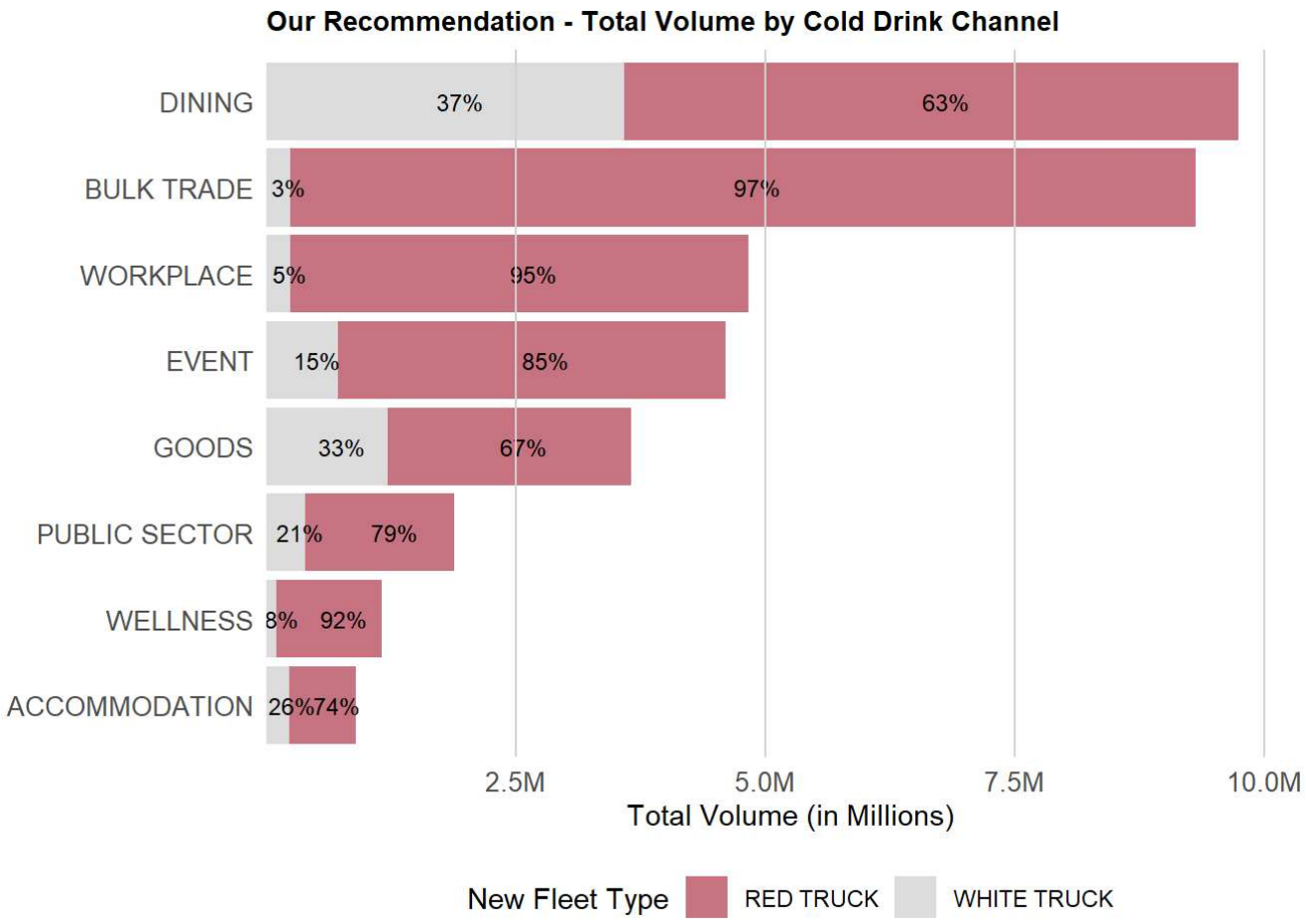
- **Others – Red Truck**: cost reduction of **$803,612** (2%);
- **Others – White Truck**: cost increase of **$160,722** (3%);
- **Local Fountain Only – Red Truck**: cost reduction of **$161,079** (34%);
- **Local Fountain Only – White Truck**: cost increase of **$32,216** (22%).

The total cost using the 400-gallon threshold over two years would be $46,462,823, while the recommendation totals $45,691,071. The net result over these two years would be a total savings of $771,752, representing a 1.7% reduction compared to the original 400-gallon threshold strategy.

These values were calculated based on actual historical delivery volumes. Predicting whether these savings will continue in the future is highly uncertain due to many potential influencing factors—such as economic shifts, customer reactions, competitor strategies, and more. Additionally, the limited historical data (only two years) adds uncertainty to future projections.

10.2 Impact on Fleet Assignment by Cold Drink Channel

▶ Code



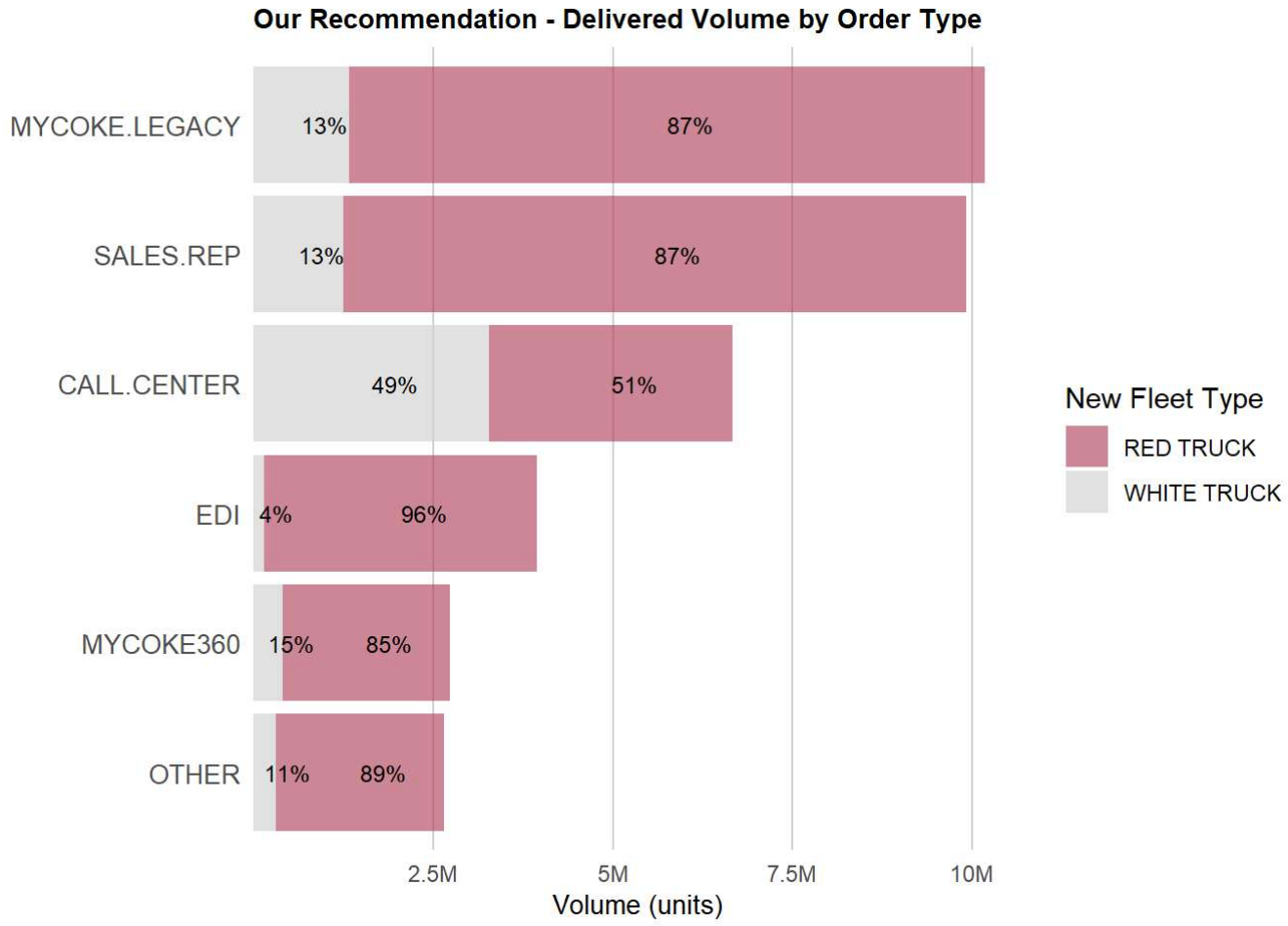
With the recommendation, the dining segment saw a 7% reduction in the number of customers previously served by red trucks, who are now being served by white trucks.

Events and Public Sector experienced a near 5% reduction in customers served by red trucks. The remaining segments saw changes of less than 2%.

The conventional segment was not displayed due to the low volume, but the change in this segment was also less than 2%.

10.3 Impact on Order Types

▶ Code



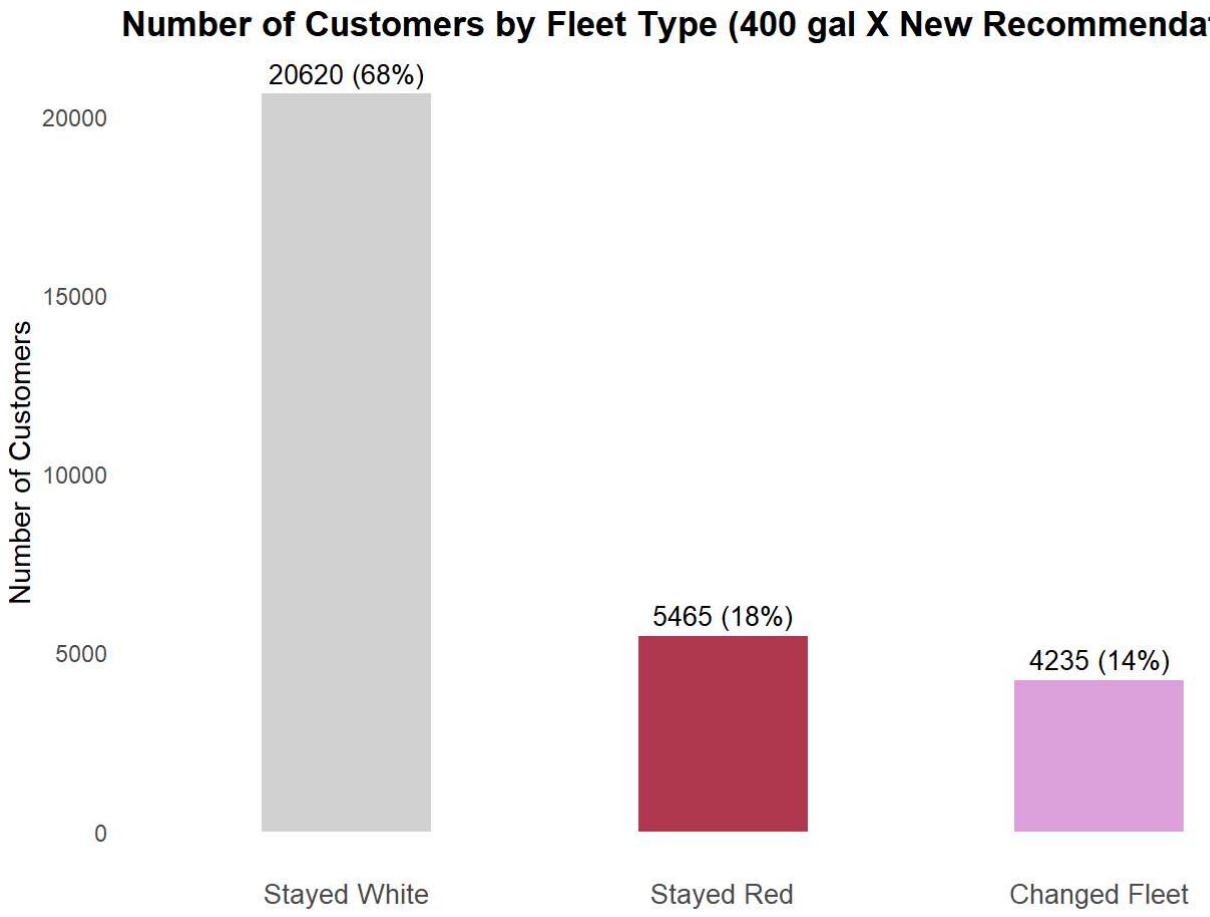
The key takeaway here is that the volume served by sales reps would see only a slight reduction of about 2 percent compared to the 400 gallon threshold. This helps avoid abrupt changes that could potentially harm relationships with customers who have closer contact with our sales team.

The most significant shift however would occur with orders placed through the call center. Approximately 20 percent of the volume that would have been served by red trucks under the 400 gallon threshold would now be served by white trucks. This allows red trucks to be redirected to other types of orders with greater potential to strengthen customer relationships.

10.4 Customers Impacted

All Customers

▶ Code

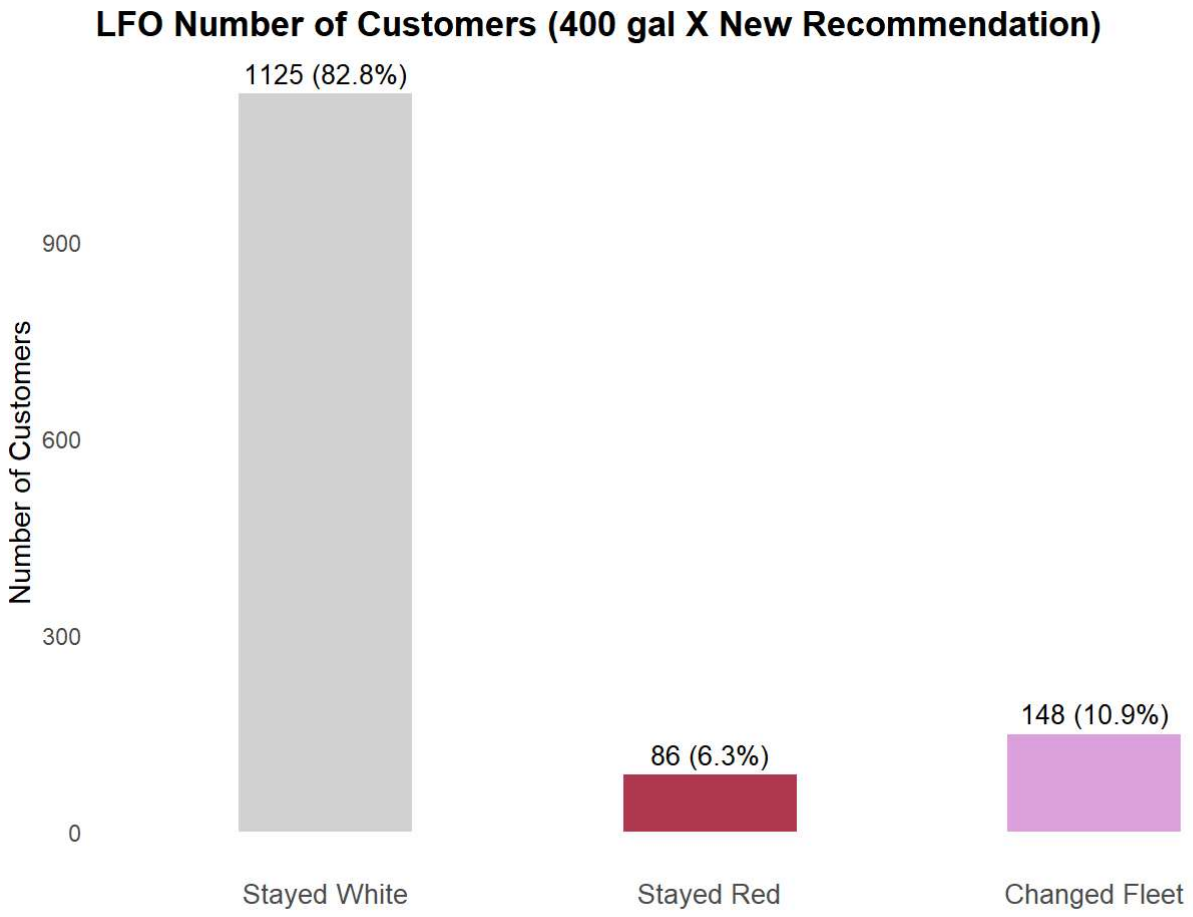


Among all customers, compared to the 400-gallon threshold, 14 percent (4,325) would have their fleet assignment changed, either from red truck to white truck or vice versa.

These 4,235 customers represent 9.3% of the total volume sold in 2023 and 2024. Of these, 2,461 would switch from white trucks to red trucks (20% of the white truck volume), while 1,774 would switch from red trucks to white trucks (7.4% of the red truck volume).

Local Market Partners - Local Fountain Only

▶ Code

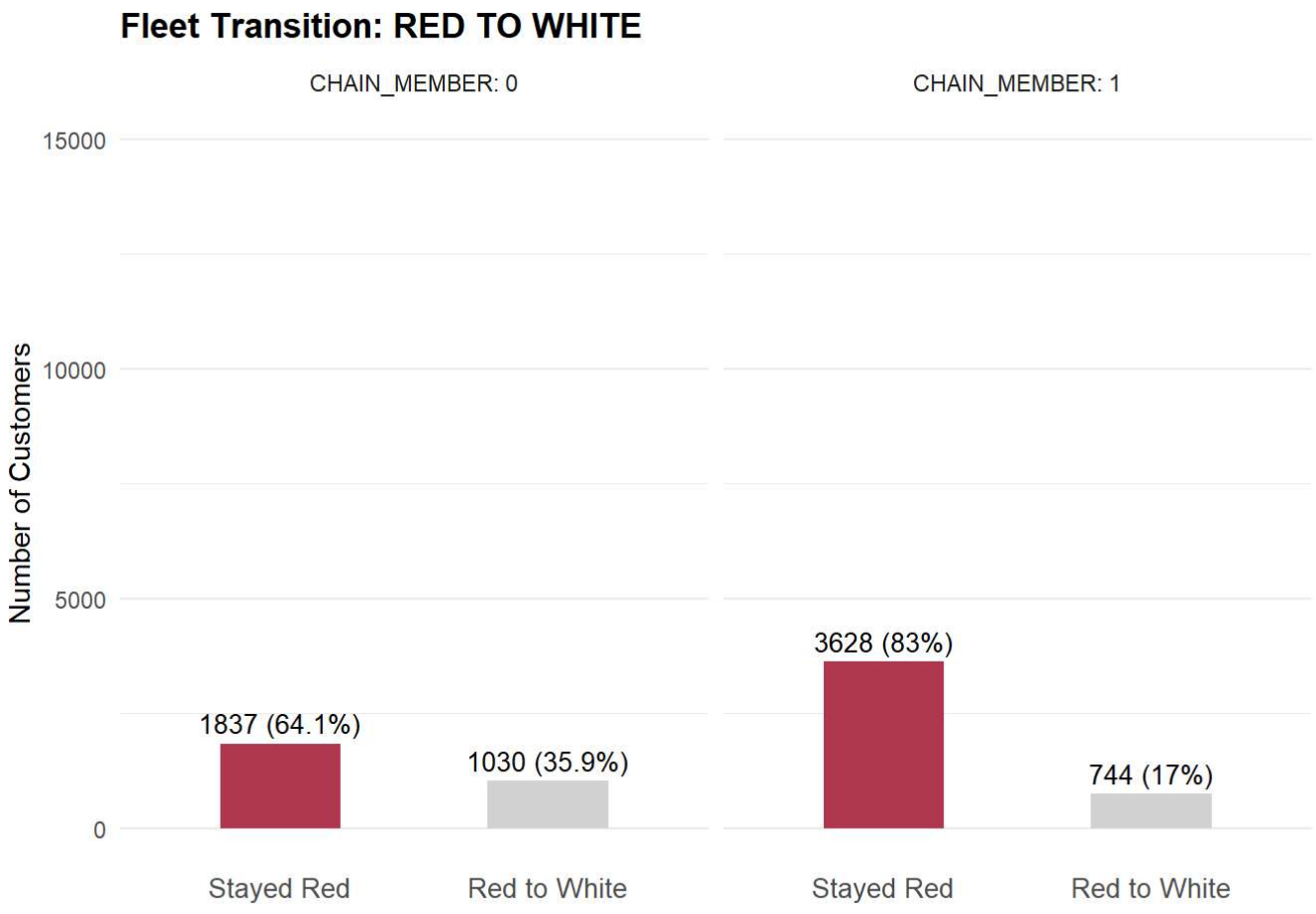


Among local market partners, 148 customers (11%) would switch fleets, making up 25% of the group's total volume. Of these, 95 switched from red trucks to white trucks, which is 52% of red truck customers and 37% of the red truck volume in this group.

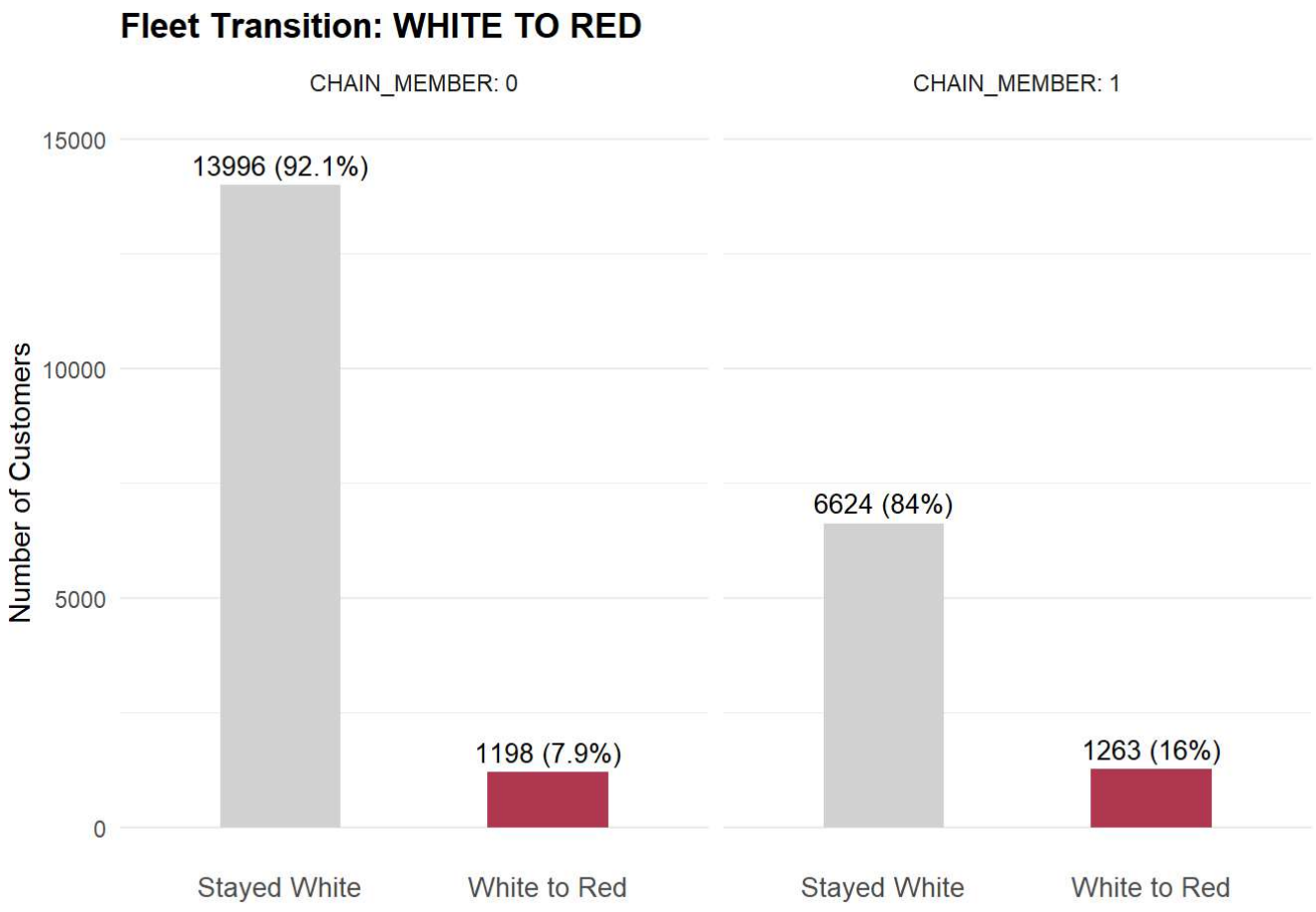
Additionally, 53 customers switched from white trucks to red trucks, representing 4.5% of white truck customers and 9% of the white truck volume in this group.

Impacts on Chain Members

▶ Code


Fleet Transition: RED TO WHITE
CHAIN_MEMBER: 0
CHAIN_MEMBER: 1
Number of Customers
15000
10000
5000
0
1837 (64.1%)
1030 (35.9%)
3628 (83%)
744 (17%)
Stayed Red
Red to White
Stayed Red
Red to White

▶ Code



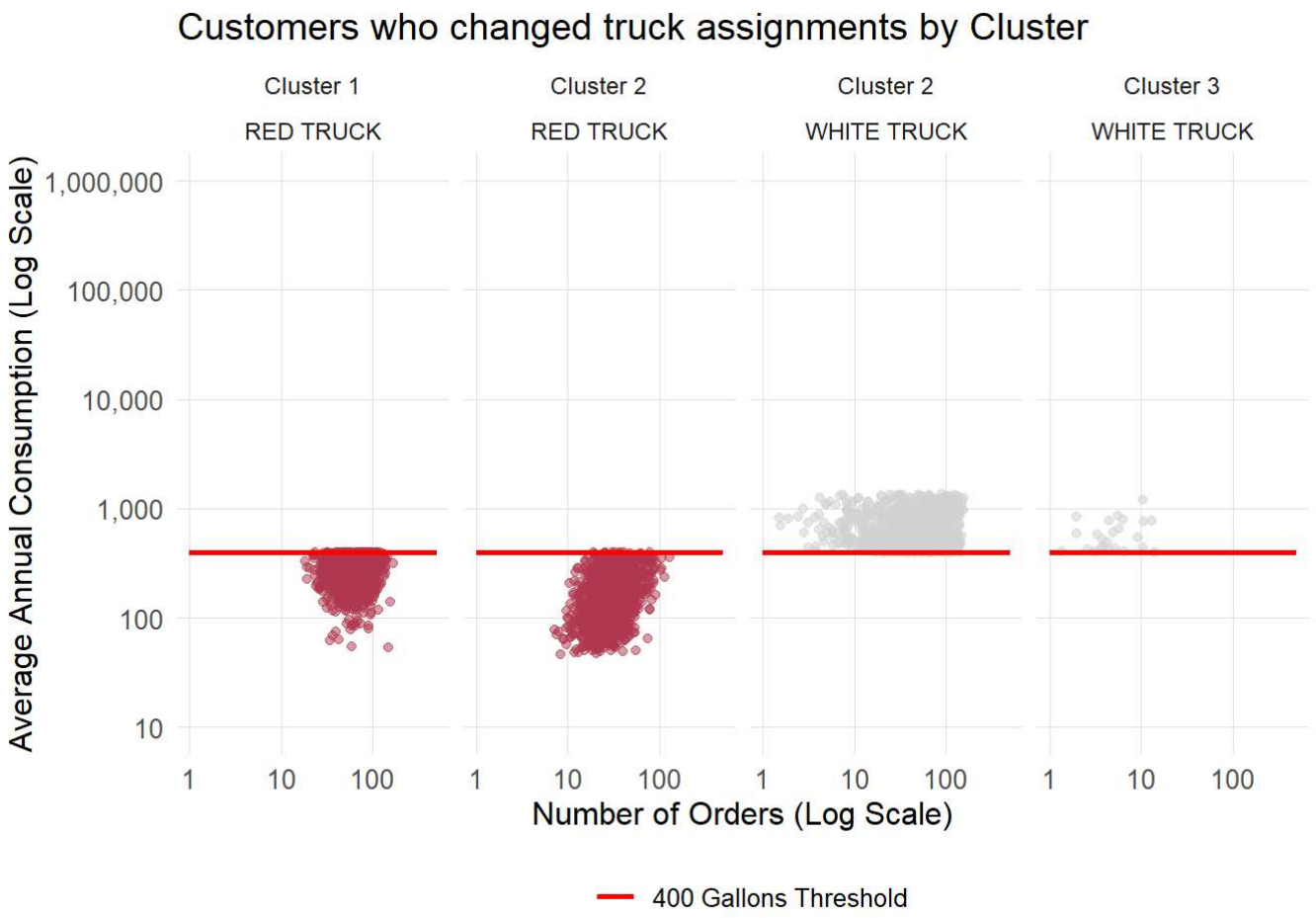
Among customers who are chain members (CHAIN_MEMBER = 1) and who, based on the 400-gallon threshold, should be served by red trucks, 17% would now be served by white trucks instead. This shift raises the question of whether there could be a negative impact due to the inconsistent service model within the same customer group.

In parallel, 16% of customers who should be served by white trucks under the same threshold would now be served by red trucks. This inversion in fleet assignment suggests a possible misalignment with the intended operational segmentation, and should be further evaluated to ensure customer experience and operational efficiency are not compromised.

10.5 Impact on Customer Segments (clusters)

Below is the visualization of customers by cluster who would change their fleet assignment based on their consumption and number of orders.

▶ Code



Out of the 425 customers who would change their fleet assignment:

These customers represent 9.3% of the total volume.

- Cluster Breakdown:

1,273 customers from Cluster 1 will now be served by red trucks.

1,188 customers from Cluster 2 switched from white trucks to red trucks, reflecting high potential, recency, and order frequency.

- Additionally:

1,748 customers from Cluster 2 switched from red trucks to white trucks.

26 customers from Cluster 3 switched from red trucks to white trucks.

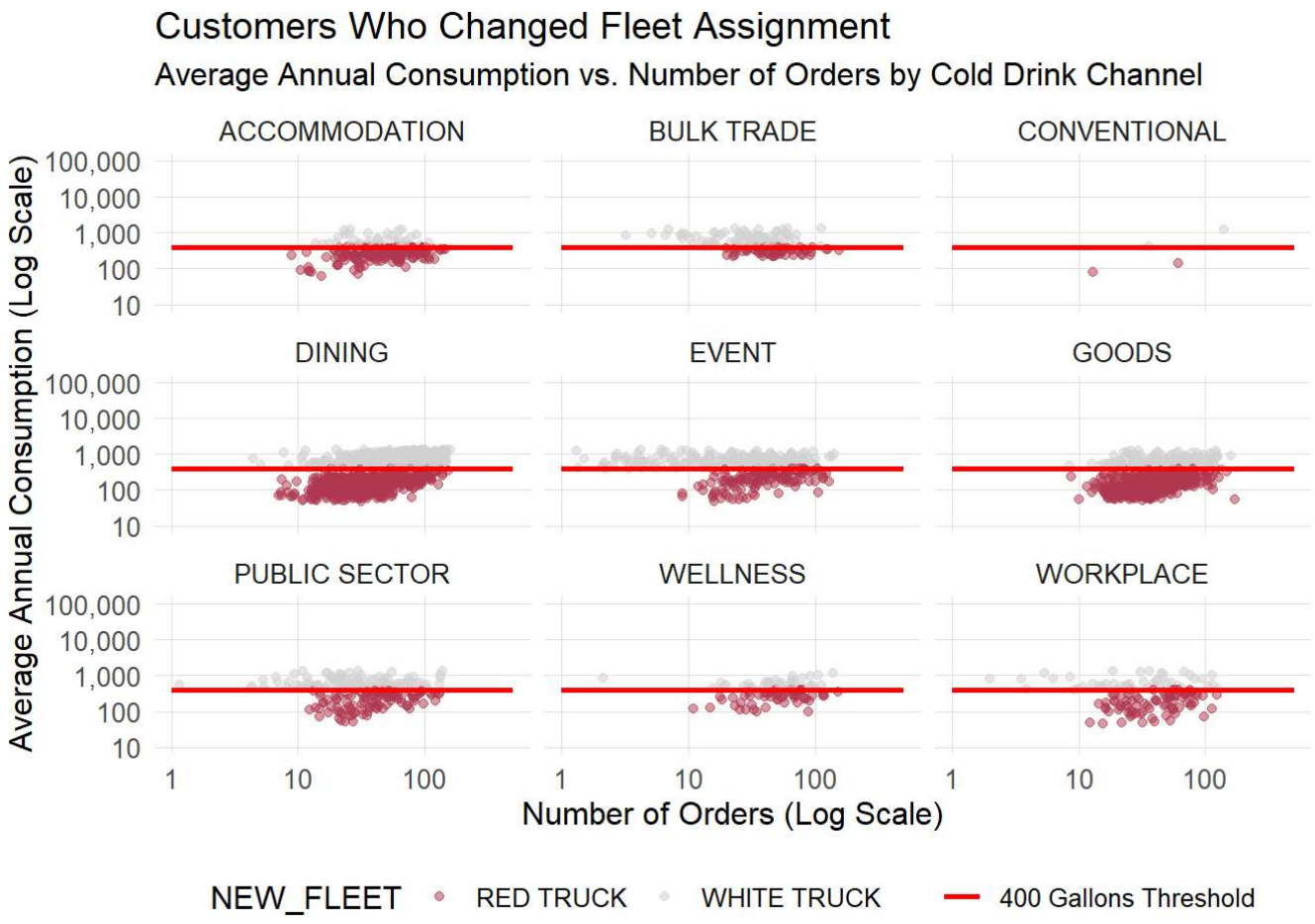
Customer Segmentation and Cold Drink Channel

Among the customers who would change fleet assignments, the majority belong to the Dining segment (52%), where 962 would switch from red trucks to white trucks, and 1,229 would switch from white trucks to red trucks.

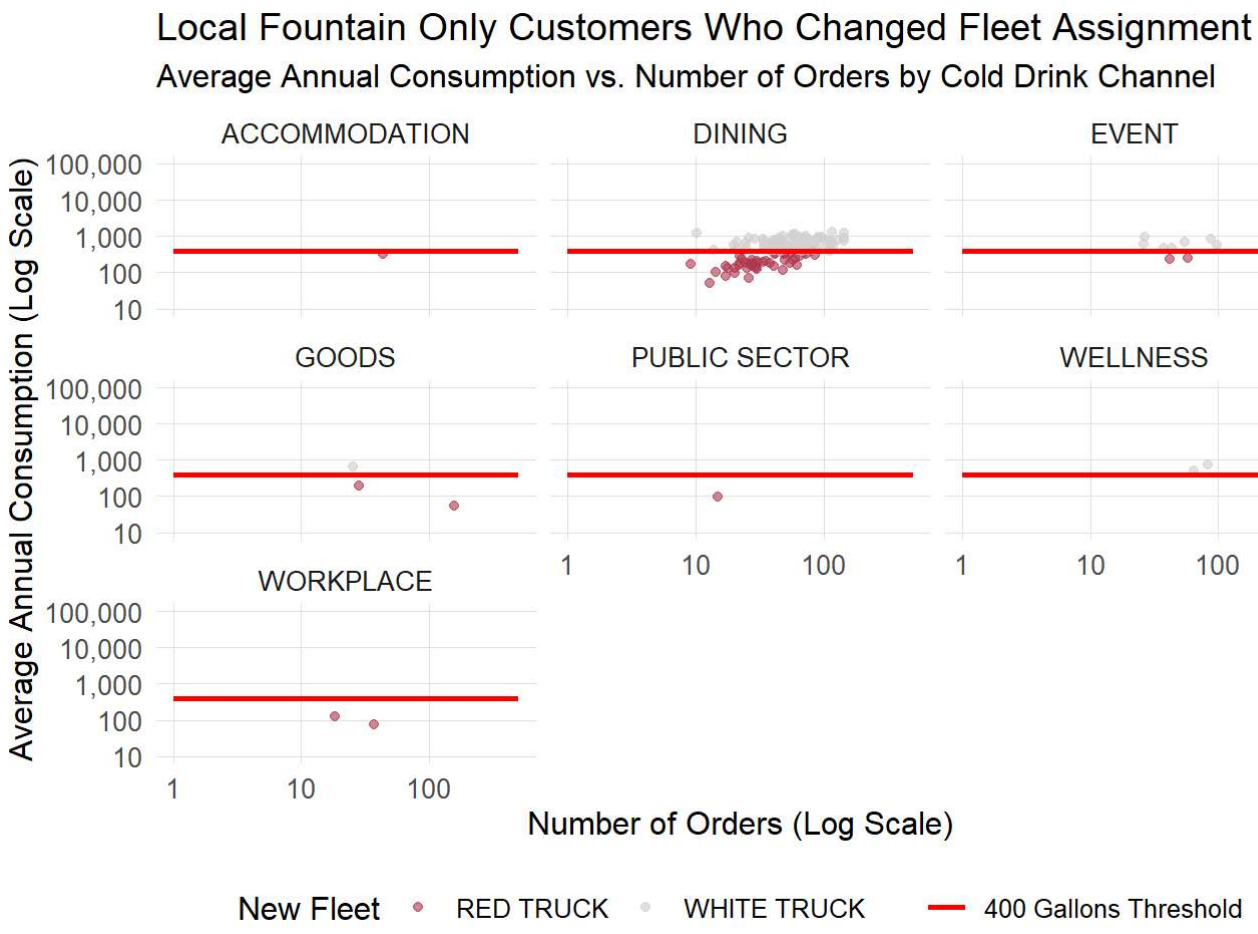
The second-largest segment with changes is GOODS (19%), where 251 customers would switch from red trucks to white trucks, and 549 would switch from white trucks to red trucks.

The EVENT segment (9%) would have 226 customers switching from red trucks to white trucks, while 159 would switch from white trucks to red trucks.

▶ Code



▶ Code



▶ Code

Among the local market partners with fountain drink only, nearly 90% of the fleet changes would occur in the Dining segment. In this group, 85 customers would switch from white trucks to red trucks, and 45 would switch from red trucks to white trucks.

11. Business Value and Final Conclusions

The proposed fleet reassignment strategy has the potential to save approximately $770,000 for the company over the past two years by increasing the number of customers served by red trucks, optimizing their usage frequency, and reducing their volume by 3%, which would allow for the eventual redeployment to strategic customers.

The proposal was quite conservative, redesigning the delivery method for only 14% of the customers and was able to assign the fleet not only based on volume but on several intrinsic customer characteristics. Therefore, the expectation is that after its implementation, there will be gains not only in cost reduction but also in sales increase, mainly for customers with greater growth potential. In addition, the proposal allowed the identification of three main customer groups, two of which showed good homogeneity.

When measuring the impacts of the new fleet assignment, the dining segment was the most impacted by these changes, particularly for the local market partners classified as fountain only. There was no significant impact on the activities of sales representatives, but there was a significant impact in reducing the volumes

delivered by red trucks (-20%) when orders are placed through call centers, which is actually a good outcome since orders through call centers no longer had a strong relationship with customers.

A differentiator for the process was the feature engineering, which brought robustness to the clustering. The supervised models, Decision Tree and Multinomial Logistic Regression, were very important in explaining the variables that influenced the clusters and, with their accuracy being raised (close to 90%), they have the potential to predict segments for new customers.

Limitations, Improvements, and Lessons

One of the main limitations of this project was the short two-year historical data, which made it difficult to predict the future impact of the recommendation. Analytical approaches were challenging due to the wide probability ranges, meaning that any outcome was possible.

Another challenge was the asynchrony between customer orders, which made it hard to track individual customer growth tied to specific times of the year. With a longer historical series, we could have made more accurate future predictions.

The census data could have been better utilized. The way it was applied in this project didn't deliver the expected results, but with adjustments and more historical data, it could provide valuable insights for future analysis.

It's clear that predicting future growth, even with extensive data, is a complex task. These predictions should only be emphasized if the process is robust, with strong statistical support and a consistent range of possible outcomes. Otherwise, it might be better to refrain from highlighting them.

Looking ahead, I strongly recommend conducting further tests to measure the impact of fleet allocation and the way customers place orders. This will be crucial in validating or refining the current approach. Additionally, analyzing revenue could provide deeper business insights, especially in understanding margins across different customer segments.

A key takeaway from this project is that data doesn't always provide all the answers we need for decision-making. In these cases, history shows that there will be both successes and setbacks, but decisions still need to be made. My role was to make responsible recommendations and take a clear stance, even when faced with uncertainties.