TABLE OF CONTENTS

INTRODUCTION

Business Problem Statement

Swire Coca-Cola (SCCU) aims to optimize delivery logistics by shifting low-volume customers to cost-efficient Alternate Routes to Market (ARTM/ white truck delivery) using third-party services. However, this may inadvertently move high-growth potential customers to less personalized service, risking revenue loss and weaker relationships. The purpose of this project is to establish a reliable, data-informed systematic method to identify and predict high-potential customers, balancing efficiency with sustainable growth.

Analytics Approach

The project team will employ various analytical methods using historical sales data and customer characteristics.

Key tasks include:

- Identifying differentiating factors between customers above and below defined volume thresholds
- Predicting high-potential accounts among ARTM customers
- Evaluating customer segmentation for insights into growth strategies

Exploratory Data Analysis (EDA)

The purpose of this exploratory data analysis (EDA) is to:

- Understand what data is available for the project.
- Understand the scope of missing data and propose solutions
- Identify patterns within the available data and characteristics of each variable
- Understand relationships between variables

This EDA aims to answer the following questions:

- What data is available for this project?
- What data is missing? How should missing data be addressed?
- Are there potential variables that should be calculated?
- Do the values make sense? Are there mistaken values that should be cleaned or imputed?
- What are the differentiating factors between customers currently ordering above and below 400 units per year?

GROUP MEMBER CONTRIBUTIONS

Each member of the group created their own EDA and information was combined for final draft to answer next steps.

DESCRIPTION OF AVAILABLE DATA

Swire Coca-Cola (SCCU) provided three main data sets for the project:

1. Historical Sales Data

Transactional Data

- This dataset records detailed information, including order quantities and delivery metrics.
- 11 variables
- 1,045,540 rows

2. Customer Profiles

Customer Profile Data

- This dataset provides detailed information about customers, including onboarding and purchasing behavior.
- 11 variables
- 30,478 rows

Customer Address and Zip Mapping

- This dataset also includes a separate table which maps ZIP codes to full address information.
- 2 variables
- 1,801 rows

3. Delivery Costs

Delivery Cost Data

- This dataset provides unit of cost measurement for gallons and cases for various volume ranges and cold drink channels.
- This data can be used to calculate Total Delivery Cost.
- 5 variables
- 160 rows

Comments:

- It is not within this project's scope to include additional, externally-source data.
- The final model will likely not include all predictors from all available data sets.
- This EDA will focus on Transactional Data and Customer Profile Data.

DISCUSSION OF MISSING DATA

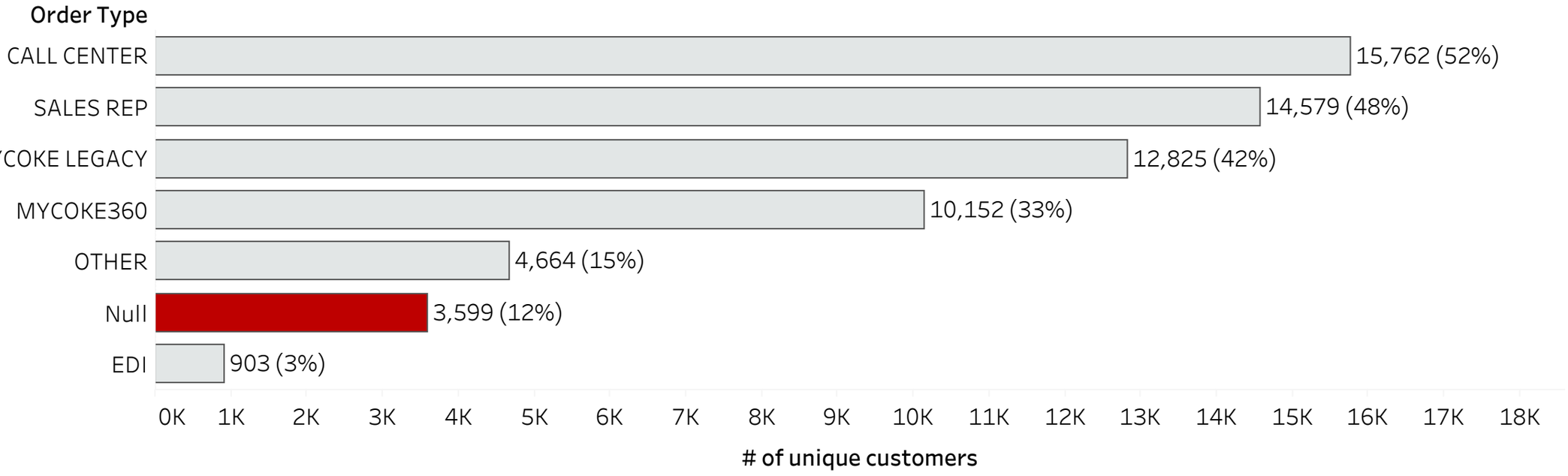
This portion of the EDA describes the scope of missing data and proposed solutions.

Questions of interest:

- What is the scope of missing data across data sets?
- What are possible solutions?
- Which solutions should be applied to which columns?

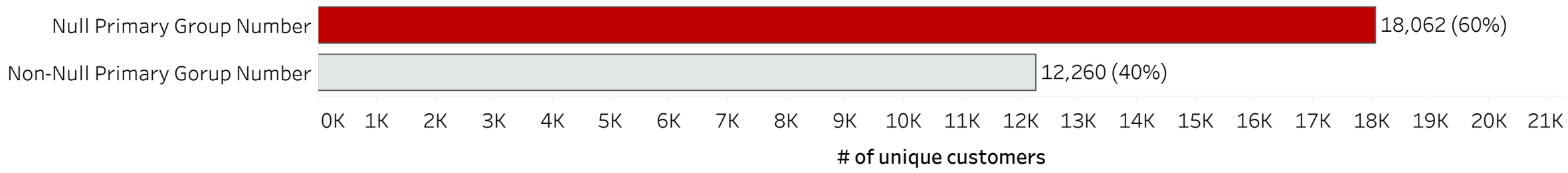
SCOPE OF MISSING DATA

Order Type - Null Values



Comments:
- About 12% of customers do not have an associated Order Type.
- This data was likely not gathered, or was lost along the way.
- This field should not be imputed, and can be left as 'null', or updated to 'unknown'.

Primary Group Number - Null Values



Comments:
- About 60% of customers do not have a Primary Group Number.
- A null Primary Group Number means that the customer does not belong to a larger conglomerate.
- A new calculated field could be created, pulling the primary group number (if available and the customer number otherwise).

POSSIBLE SOLUTIONS

There are two fields with missing (or null) values: Order Type and Primary Group Number. Both fields are categorical, and in some cases categorical variables can be imputed using the mode. This method does not make sense for either variable for this project. Alternatively, we suggest the following:

Order Type
- Replace null values with 'unknown', as to not lose 12% of the data points.

Primary Group Number
- Create a new variable called 'Customer Group Number (Rollup)'.
- This new variable will represent the primary group number, if available, and the customer number otherwise.
- This will allow the aggregation transactions up to primary group membership.

FEATURE ENGINEERING

Total Units Ordered

Description: used to quantify total units ordered, which are then used to determine when a threshold is reached
Calculation: sum of two existing variables

[Ordered Gallons] + [Ordered Cases]

Total Units Loaded

Description: used to quantify total units loaded
Calculation: sum of two existing variables

[Loaded Gallons] + [Loaded Cases]

Total Units Delivered

Description: used to quantify total units delivered
Calculation: sum of two existing variables

[Delivered Gallons] + [Delivered Cases]

Order Grouping

Description: used to identify customers who are ordering less or more than 400 units per year
Calculation: logical statement utilizing level of detail (LOD) calculation functionality

```
IF {FIXED [Year], [Customer Number]:    SUM([Total Units Ordered])} < 400
THEN 'Customers Ordering Less than 400'
ELSE 'Customers Ordering More than 400'
END
```

Customer Group Number (Rollup)

Description: used to aggregate transactions up to the customer group number, if desired
Calculation: logical statement using two existing variables

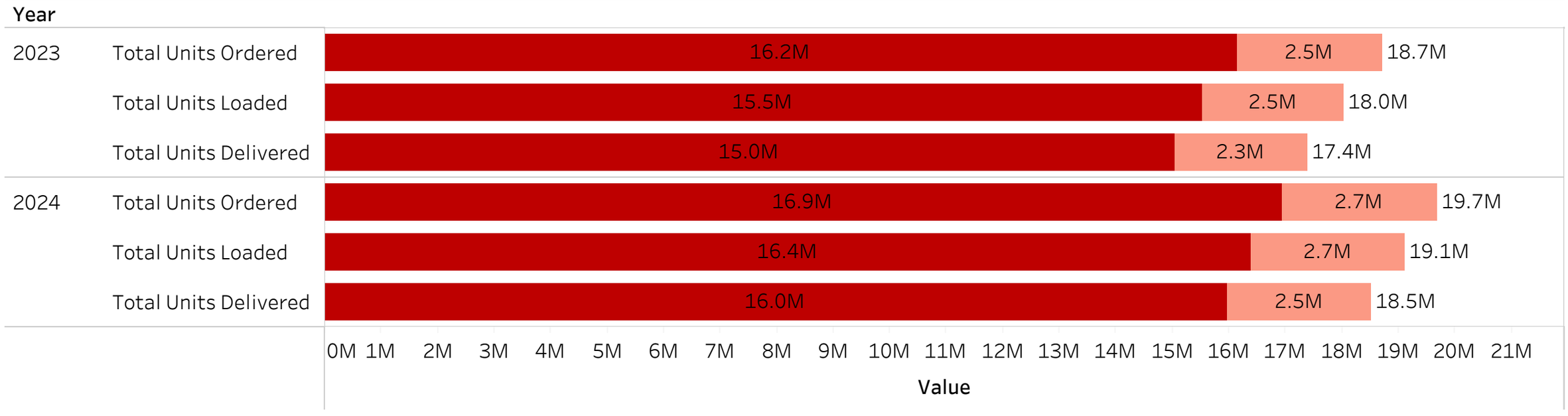
```
IF NOT ISNULL([Primary Group Number])
THEN [Primary Group Number]
ELSE [Customer Number]
END
```

EXPLORATORY VISUALIZATIONS & SUMMARY TABLES

GENERAL DATA EXPLORATION

Order Grouping
- Customers Ordering Less than 400 units
- Customers Ordering More than 400 units

Ordered, Loaded, Delivered Totals by Year



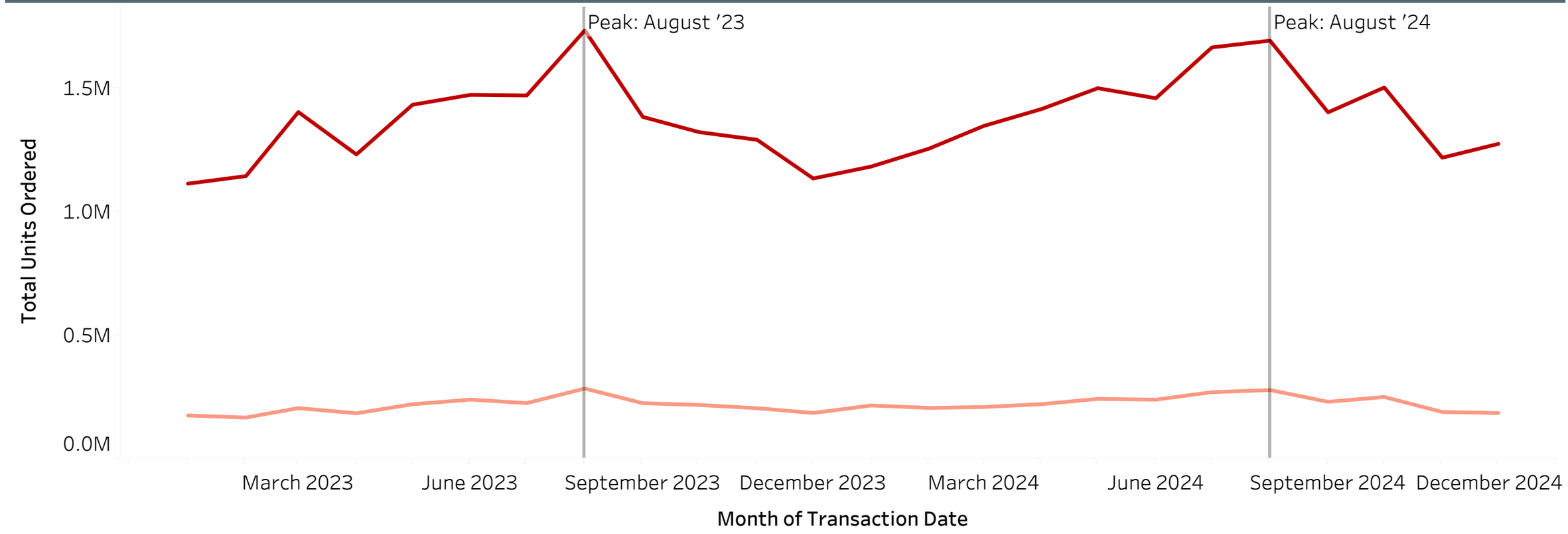
Comments:
- Most units (Cases + Gallons) are associated with Customers who order more than 400 units per year.
- At an annual level, the number of units ordered is consistently more than those loaded or delivered.

Summary of Total Ordered Units

	Year 2023	2024
Min. Total Units Ordered	0	0
Max. Total Units Ordered	8,177	8,480
Avg. Total Units Ordered	36	38
Std. dev. of Total Units Ordered	125	135

Comments:
- Since the Average Total Units Ordered is closer to the Min than the Max, there are likely many customers who order few total units.

Ordered Units by Transaction Date

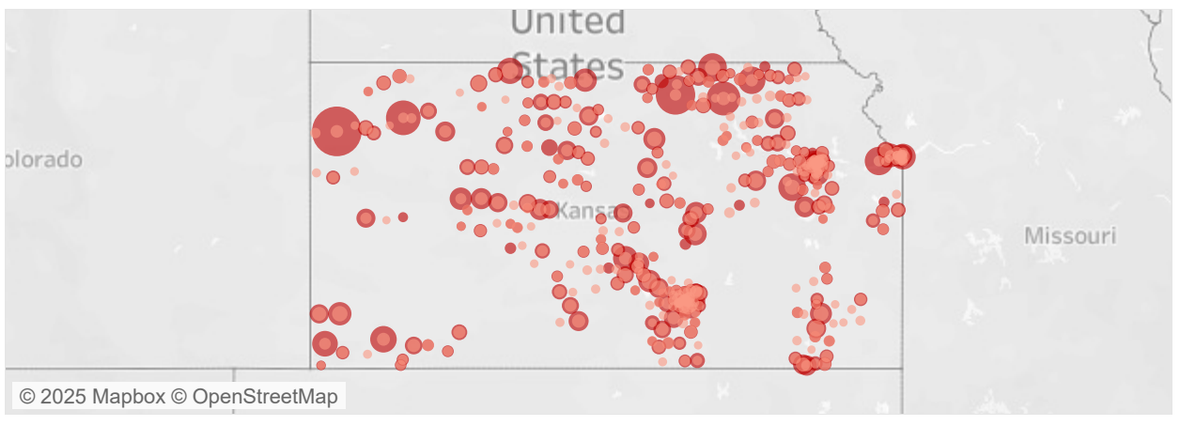


Comments:
- Both order groupings experience peaks in August for both 2023 and 2024.
- This peak is more prominent for customers ordering more than 400 units annually.

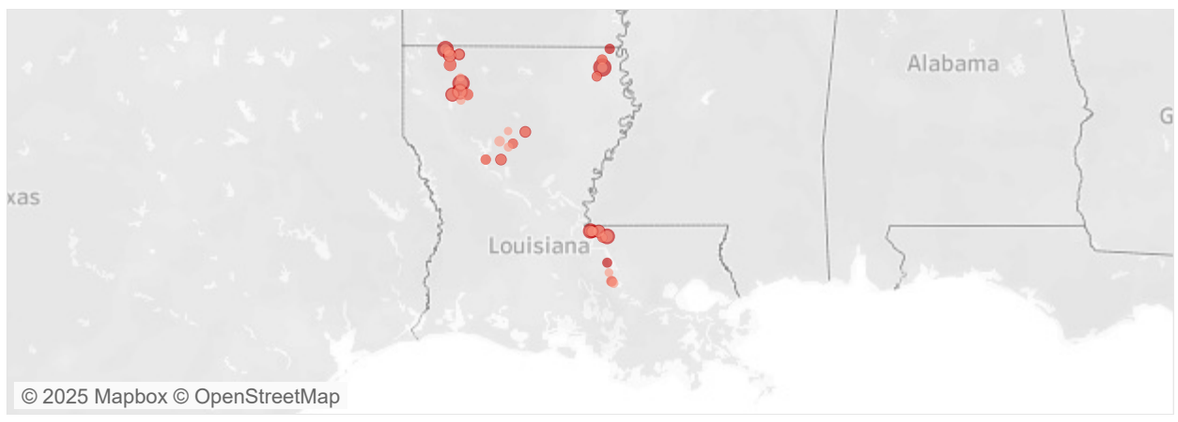
Ordered Units by Zip Code

size represents the total units ordered

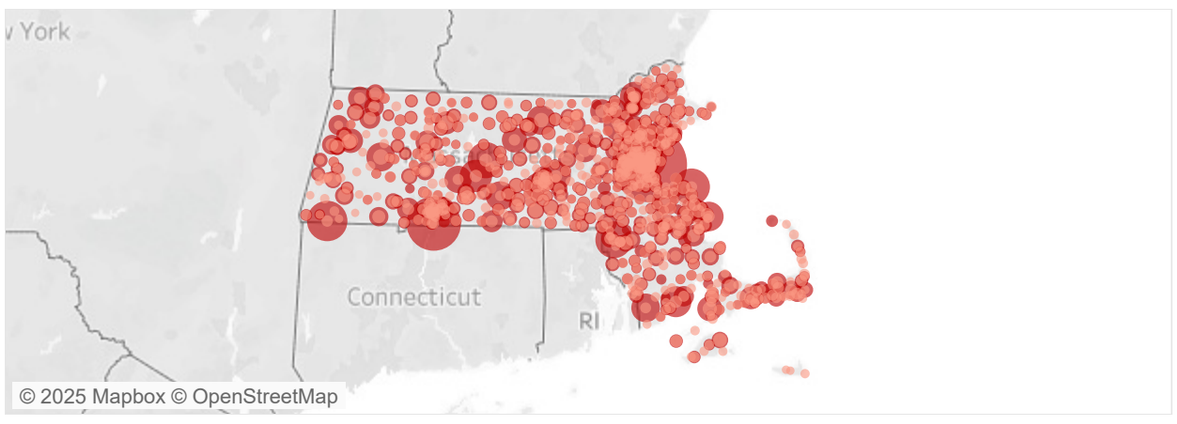
Kansas



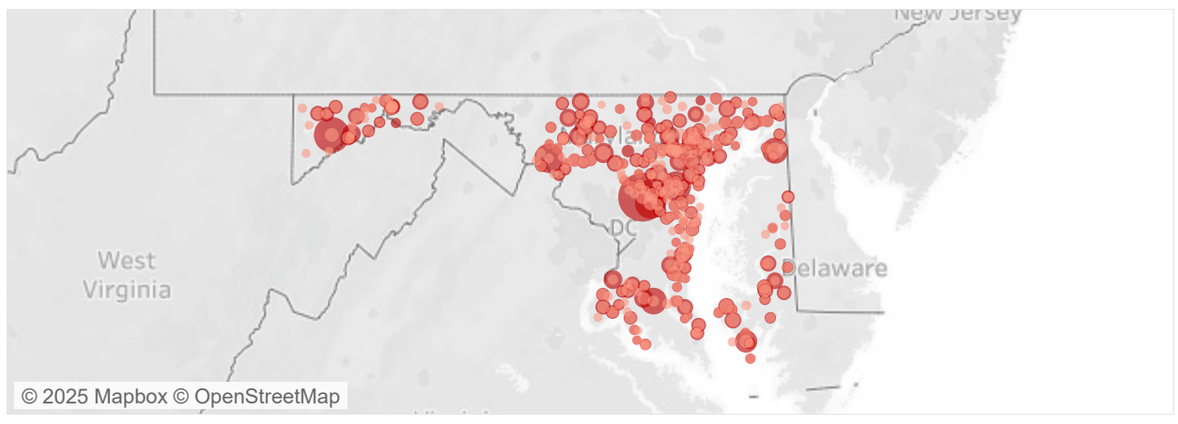
Louisiana



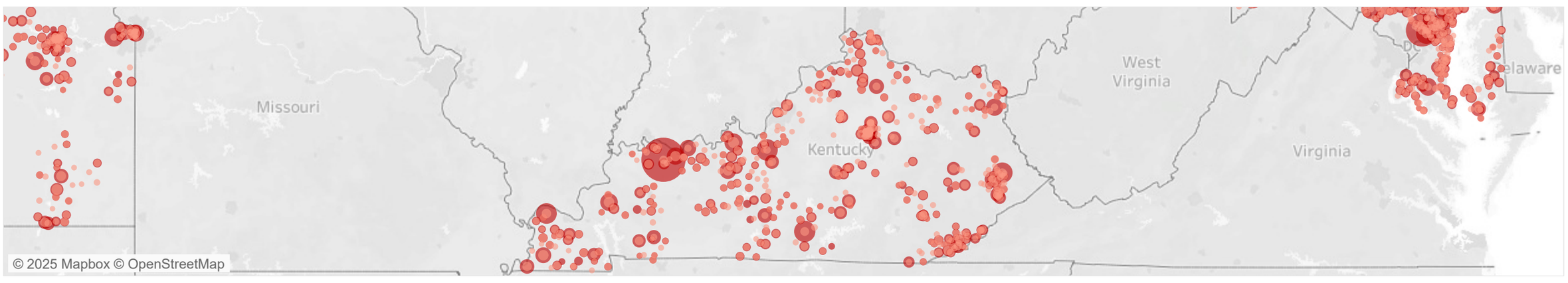
Massachusetts



Maryland



Kentucky

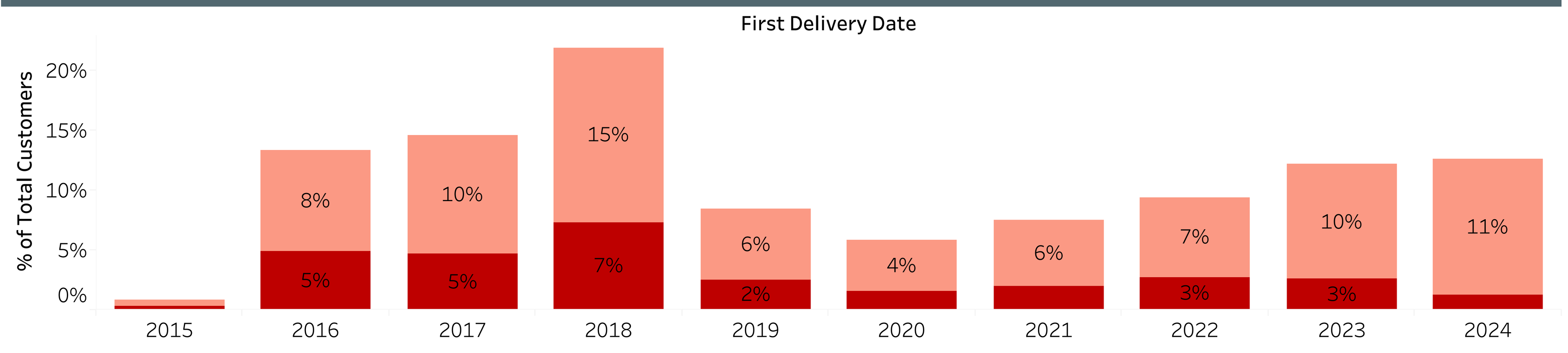


Comments:
- There seem to be some customers ordering less than 400 units annually that are not geographically close to other customers.
- Depending on the assignment of zip codes (randomization) this information may not be helpful in modeling.

EXPLORATION OF CUSTOMERS

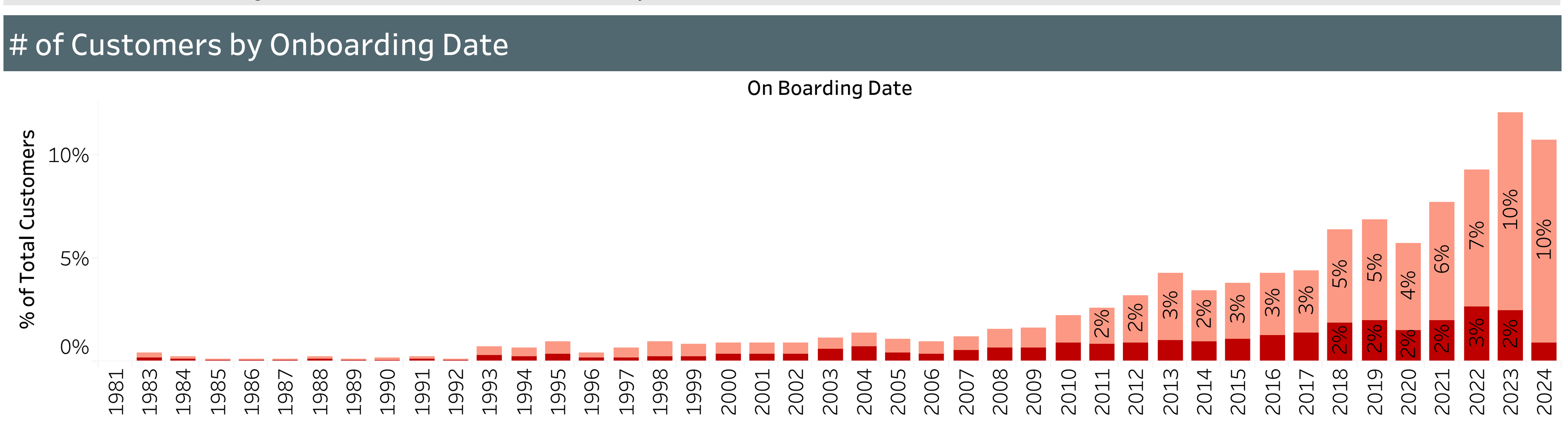
Order Grouping
- Customers Ordering Less than 400 units
- Customers Ordering More than 400 units

of Customers by First Delivery Date



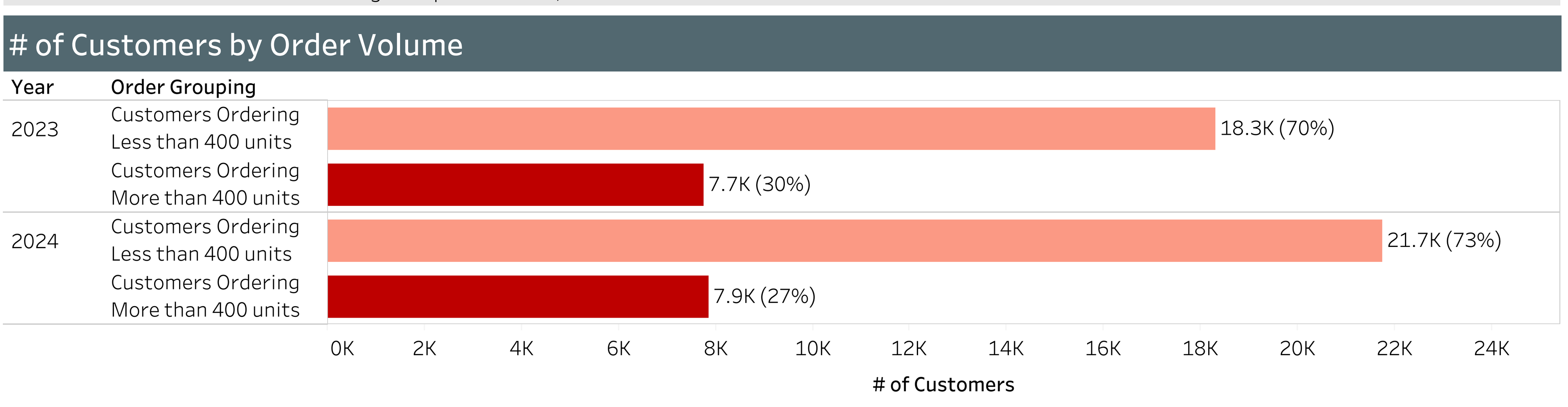
Comments:
- Half of customers have a first delivery date between 2015 and 2018, and half between 2019 and 2024.
- Most customers ordering more than 400 units have a first delivery between 2015 and 2018.

of Customers by Onboarding Date



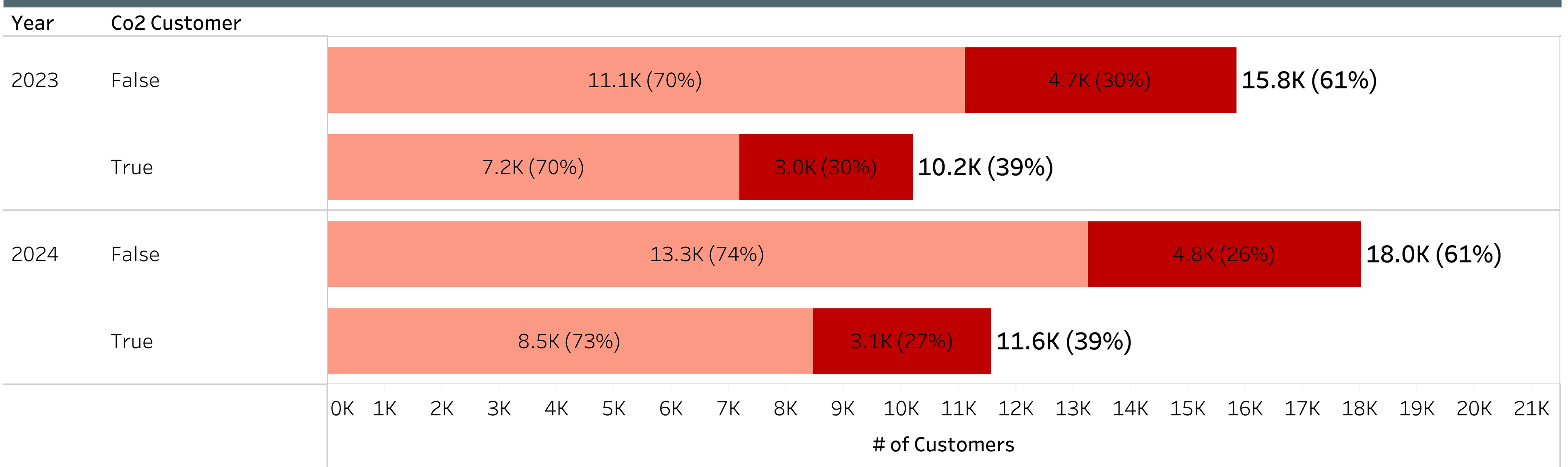
Comments:
- Half of customers have an onboarding date prior to 2018, and half between 2018 and 2024.

of Customers by Order Volume



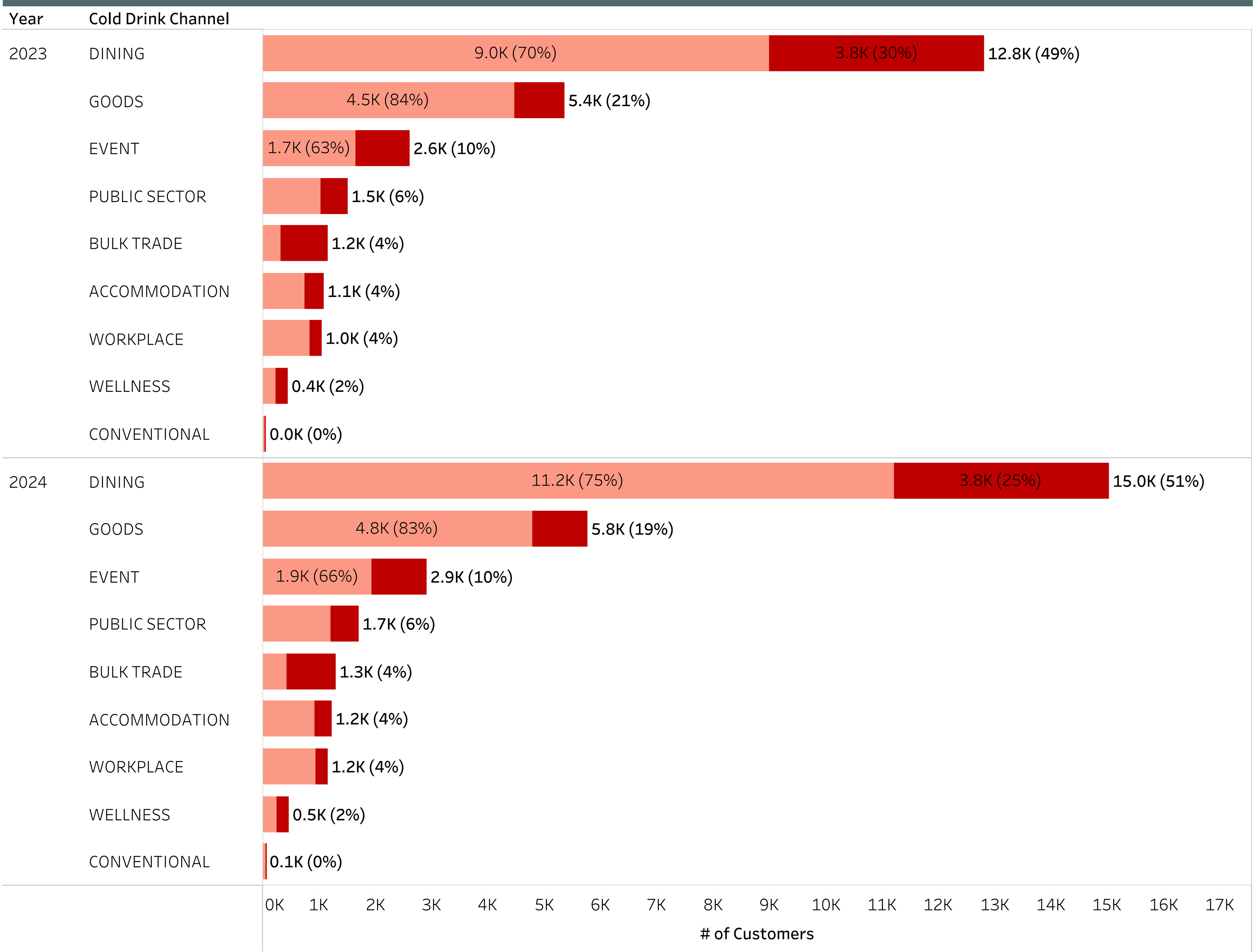
Comments:
- About 70% of customers order less than 400 units per year, with the remaining 30% of customers ordering more than 400 units per year.
- The percentage of customers ordering more than 400 units per year decreased from 2023 to 2024.

of Customers by CO2 Customer Status



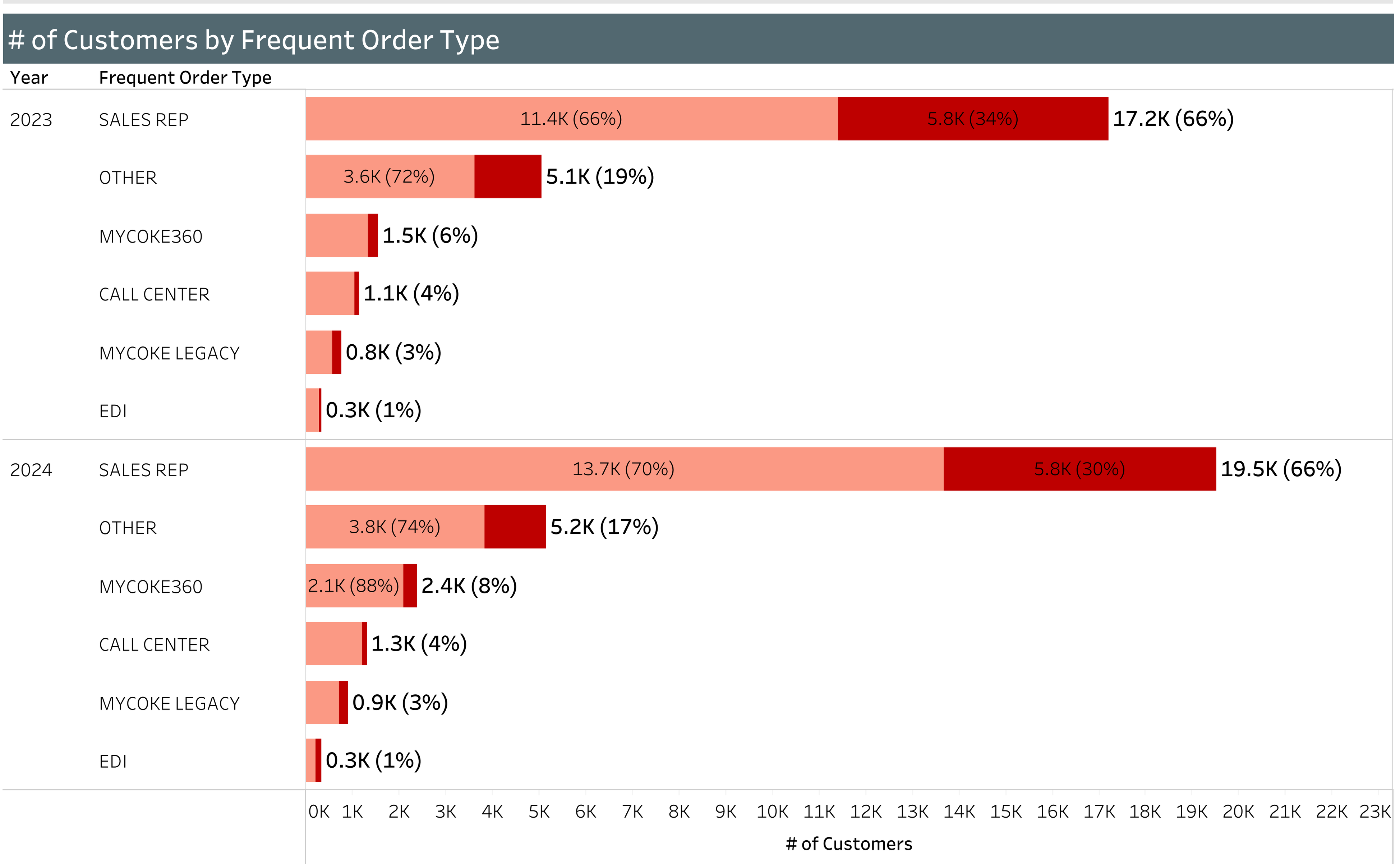
Comments:
- About 60% of customers are not classified as CO2.
- Customers ordering above and below 400 units are equally split across CO2 status.

of Customers by Cold Drink Channel



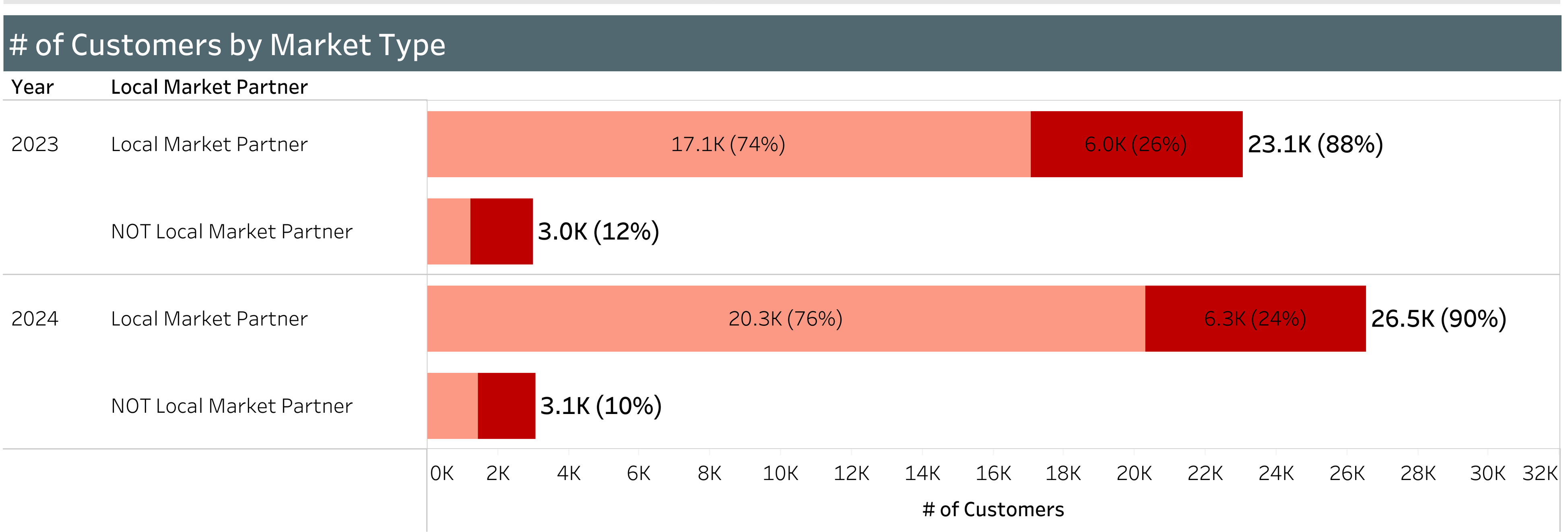
Comments:
- About 50% of customers belong to the 'Dining' Cold Drink Channel.
- In 2024, customers ordering less than 400 units per year disproportionately belong to the 'Dining' Cold Drink Channel.
- Customers in the 'Bulk Trade' Cold Drink Channel are disproportionately customers ordering more than 400 units per year.

of Customers by Frequent Order Type



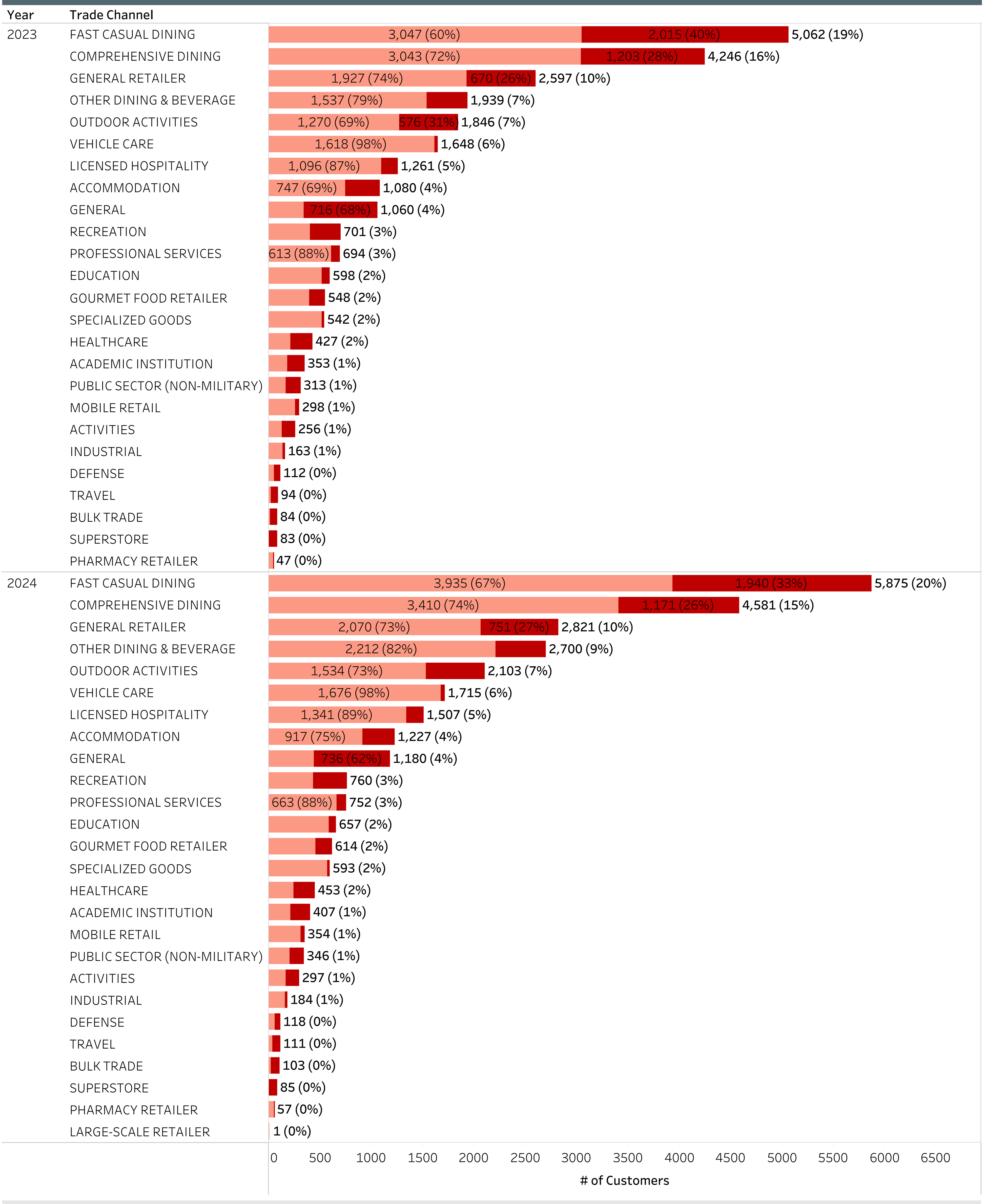
Comments:
- About 66% of customers belong to the 'Sales Rep' Frequent Order Type.
- Most 'Call Center' Frequent Order Type customers are customers ordering less than 400 units per year.

of Customers by Market Type



Comments:
- Most customers (88%) are Local Market Partners.
- Customers who are Not-Local Market Partners are disproportionately those who order more than 400 units per year.

of Customers by Trade Channel



Comments:
- The Trade Channel with the most customers is 'Fast Casual Dining'.

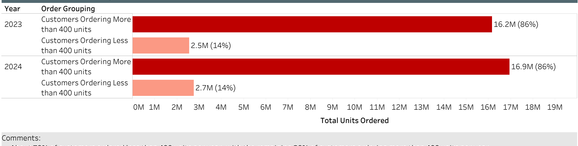
SWIRE COCA-COLA | USA

EXPLORATION OF ORDERED UNITS

units = gallons + cases

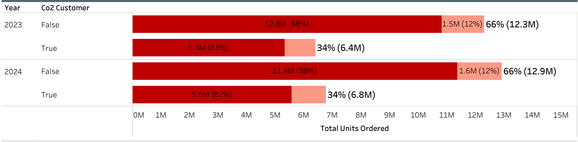
Order Grouping
- Customers Ordering More than 400 units
- Customers Ordering Less than 400 units

of Ordered Units by Order Volume



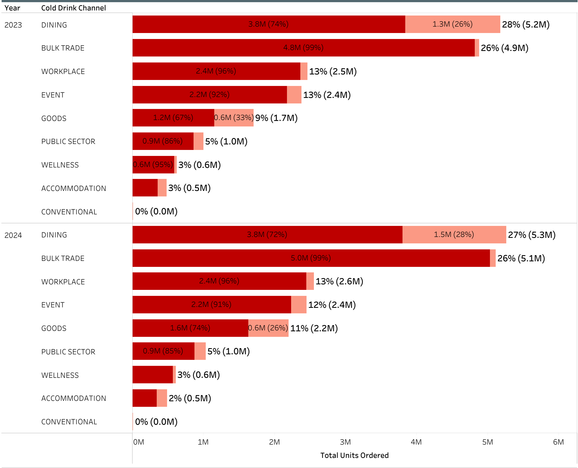
Comments:
- About 70% of customers ordered less than 400 units per year, with the remaining 30% of customers ordering more than 400 units per year.
- About 86% of units are ordered by customers who ordered more than 400 units per year, with the remaining 14% of units being ordered by customers who ordered less than 400 units per year.
- This inverted relationship is a point of interest, indicating that 30% of customers are ordering 86% of units, while 70% of customers are ordering 14% of units.

of Ordered Units by CO2 Customer Status



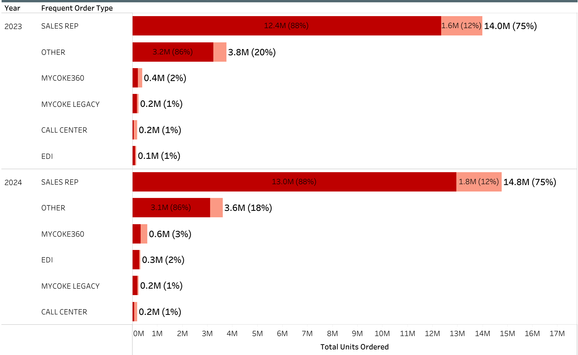
Comments:
- About 66% of ordered units are not classified as CO2.

of Ordered Units by Cold Drink Channel



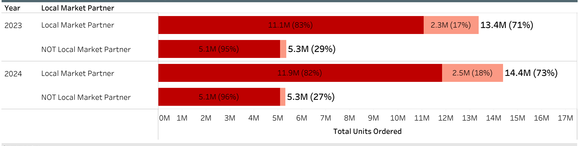
Comments:
- Units ordered via the 'Dining' and 'Goods' Cold Drink Channels are disproportionately ordered by customers ordering less than 400 units annually.
- Units ordered via the 'Bulk Trade', 'Workplace', 'Event', and 'Wellness' Cold Drink Channels are disproportionately ordered by customers ordering more than 400 units annually.

of Ordered Units by Frequent Order Type



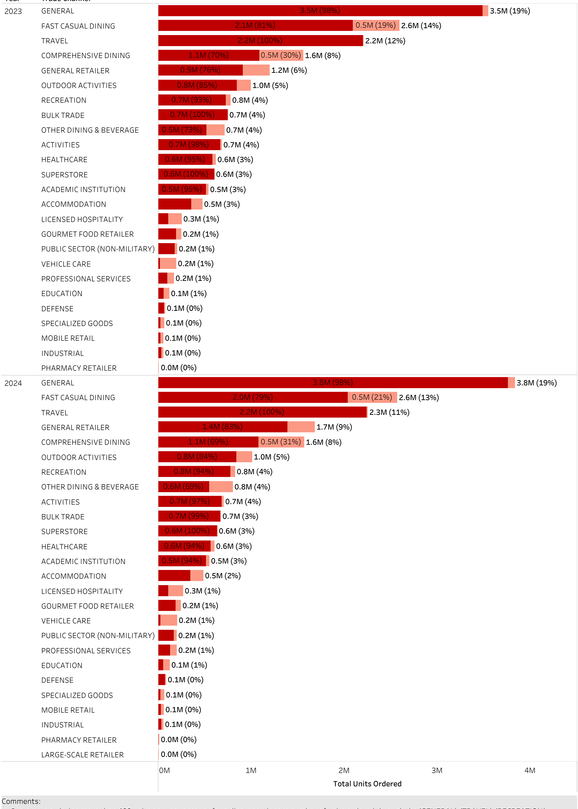
Comments:
- Most (75%) of ordered units have a frequent order type of 'Sales Rep'.
- Relatively few ordered units (~5%) have a frequent order type of 'MYCOKE360', 'MYCOKE LEGACY', or 'EDI'.

of Ordered Units by Market Type



Comments:
- Most ordered units are 'Local Market Partners'
- Ordered units by non-local market partners are disproportionately ordered by customers ordering more than 400 units per year.

of Ordered Units by Trade Channel



Comments:
- Customers ordering more than 400 units per year account for a disproportionate number of units ordered through the 'GENERAL', 'TRAVEL', 'RECREATION', 'BULK TRADE', 'ACTIVITIES', 'HEALTHCARE', 'SUPERSTORE', and 'ACADEMIC INSTITUTION' Trade Channels.
- Customers ordering less than 400 units per year account for a disproportionate number of units ordered through the 'COMPREHENSIVE DINING', 'OTHER DINING & BEVERAGE', 'GENERAL RETAILER', 'ACCOMMODATION', 'LICENSED HOSPITALITY', and 'VEHICLE CARE' Trade Channels.

RESULTS

Description of Available Data

Swire Coca-Cola (SCCU) provided three main data sets for the project:

- Transactional Data
- Customer Profiles
- Delivery Costs

Comments:

- It is not within this project's scope to include additional, externally-sourced data.
- The final model will likely not include all predictors from all available data sets.
- This EDA focused on Transactional and Customer Profile Data.

Discussion of Missing Data

There are two fields with missing (or null) values:

Order Type

- Replace null values with 'unknown', as to not lose 12% of the data points.

Primary Group Number

- Create a new variable called 'Customer Group Number (Rollup)'.
- This new variable will represent the primary group number, if available, and the customer number otherwise.
- This will allow the aggregation transactions up to primary group membership.

Feature Engineering

Five new variables were created using existing variables:

- Total Units Ordered - used to quantify total units ordered
- Total Units Loaded - used to quantify total units loaded
- Total Units Delivered - used to quantify total units delivered
- Order Grouping - used to identify customers who are ordering less or more than 400 units per year
- Customer Group Number (Rollup) - used to aggregate transactions up to the customer group number, if desired

Exploratory Visualizations

Interesting observations include:

- About 70% of customers ordered less than 400 units per year, with the remaining 30% of customers ordering more than 400 units per year. While about 86% of units are ordered by customers who order more than 400 units per year, with the remaining 14% of units being ordered by customers who ordered less than 400 units per year.
    - This inverted relationship is a point of interest, indicating that 30% of customers order 86% of units, while 70% of customers are ordering 14% of units.
- The number of units ordered peaked in August '23 and '24, this peak was more prominent for customers ordering more than 400 units annually.
- The percentage of customers ordering more than 400 units per year decreased from '23 to '24.

NEXT STEPS

Potential questions to address in the modeling portion of this project:

- What is the current average growth rate for Local Market Partners?
- What is the current average growth rate for All Customers?
- Are there specific customers with exceptionally low or high growth rates? What types of differentiating factors are present?
- What is the average Total Delivery Cost per Local Market Partner?
- What is the Total Delivery Cost per customer (all)?

NEXT STEPS continued...

A supervised machine learning model will be used to predict total annual delivery cost per customer. By leveraging historical sales and delivery data, it will help identify ARTM customers with potential for volume growth, ensuring cost-efficient truck assignments and forecasting high-cost customers before they exceed thresholds. The model is trained using one year of data as the training set and the next year as the test set, with key features such as order volume, delivery success rate, and trade channel. Performance will be evaluated based on key metrics such as RMSE, R^2 , and MAPE to evaluate the model's accuracy.

Once the calculations and predictions are complete, unsupervised learning (clustering) can be used to segment deliveries into White Truck vs. Red Truck categories based on key logistics variables.

The goal of this dashboard is to help logistics teams and operations managers classify deliveries into White Truck vs. Red Truck based on key order attributes. By integrating clustering techniques, this dashboard will automate truck assignments, reduce delivery costs, and improve operational efficiency. Ultimately, the desire is to determine how to integrate the insights into a routine strategy to support long-term growth while maintaining logistical efficiency.