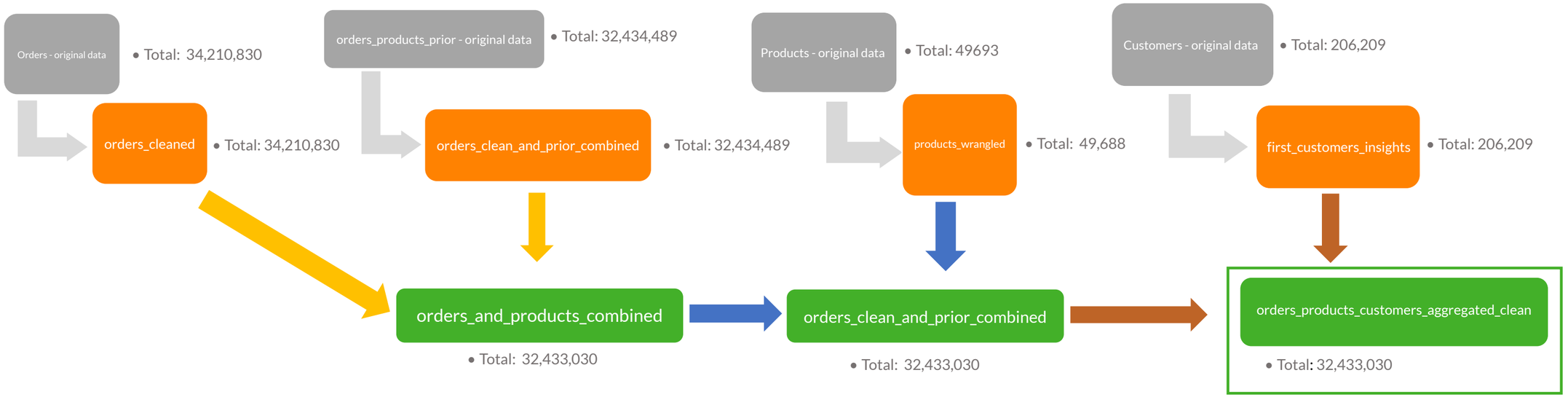
Project Name: Data exploration, and customers trends analysis
Date: 24.08.2025
Analyst Name: Álvaro Madrigal
Data source: *"The Instacart Online Grocery Shopping Dataset 2017"*, Accessed from *www.instacart.com/datasets/grocery-shopping-2017* via *Kaggle* on [24.08.2025].

Contents:

instacart

Population flow



1.) The first row of boxes (grey) of the population flow represent the original data sets from *"The Instacart Online Grocery Shopping Dataset 2017"*.

2.) The second row of boxes (orange) represents the data sets after data manipulation (e.g., missing values and duplicates removing).

3.) The third row of boxes (green) represents the merges between the datasets.

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates
orders	206,209 in column "since_prior_order column"	Kept them. Identified as first order, and create a new column to mark them	No duplicates
products	16 in column "product_name"	Imput "Unknown Product" in order to keep the 16 for analysis	5 duplicates
orders_products_prior	No missing values	-	No duplicates
customers	8 missing values in column "first_name"	Kept	No duplicates

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
eval_set	-	-	Result from merging data frames, not relevant for analysis
Unnamed: 0.1	-	-	Result from merging data frames, not relevant for analysis
Unnamed: 0	-	-	Result from merging data frames, not relevant for analysis
-	order_dow	-	Renamed to "order_day_of_week"
-	First Name	-	Renamed to "first_name"
-	Surname	-	Renamed to "last_name"
-	STATE		Renamed to "state"
-	GENDER		Renamed to "gender"
-	AGE		Renamed to "age"
-	-	date_joined	Changed data type to date
-	-	user_id	Changed to int32
-	-	age	Changed to int8
-	-	n_dependants	Changed to int8
-	-	income	Changed to int32
-	-	gender	Changed to Category
-	-	fam_status	Changed to Category
-	-	state	Changed to Category

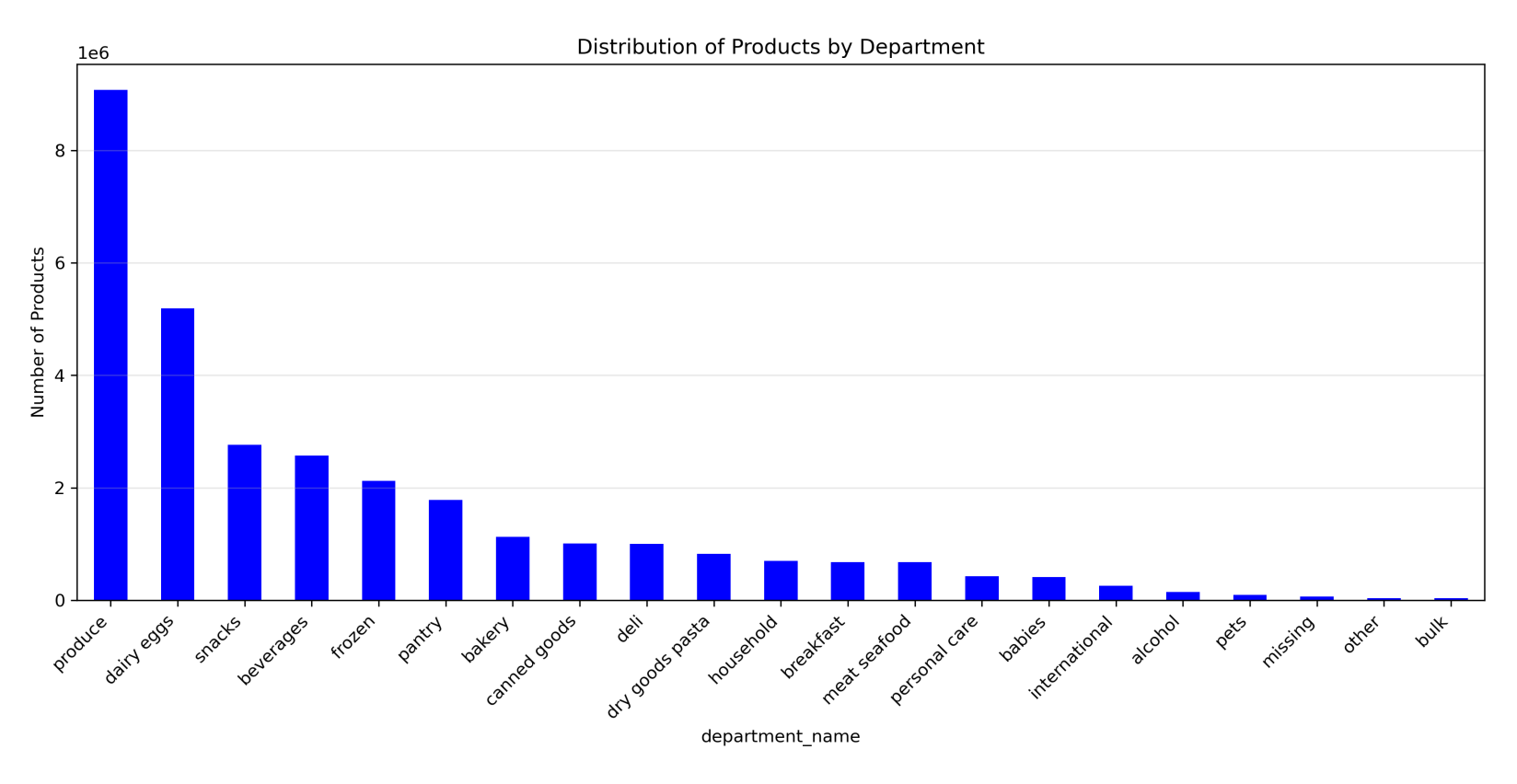
Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
orders_products_customers	first_order	days_since_prior_order	orders['first_order'] = orders['days_since_prior_order'].isna().astype(int)
orders_products_customers	max_order	order_id	orders_and_products_combined['max_order'] = orders_and_products_combine
orders_products_customers	region	state	region = {'Maine' : 'Northeast','New Hampshire' : 'Northeast','Vermont' : 'Nort
orders_products_customers	price_range	price	> 15, 'price_range_loc'] = 'High-range product' <= 15)] <= 5, 'price_range_loc'] =
orders_products_customers	loyalty_flag	max_order	orders_and_products_combined.loc[(ords_prods_merge['max_order'] <= 40) &
orders_products_customers	avg_spend_by_product_flag	avg_product_spend	<10 = 'Low spender', ≥10 = 'High spender'
orders_products_customers	customer_recurrence_flag	order_id	<6 = 'Inactive', ≥6 = 'Active Customer

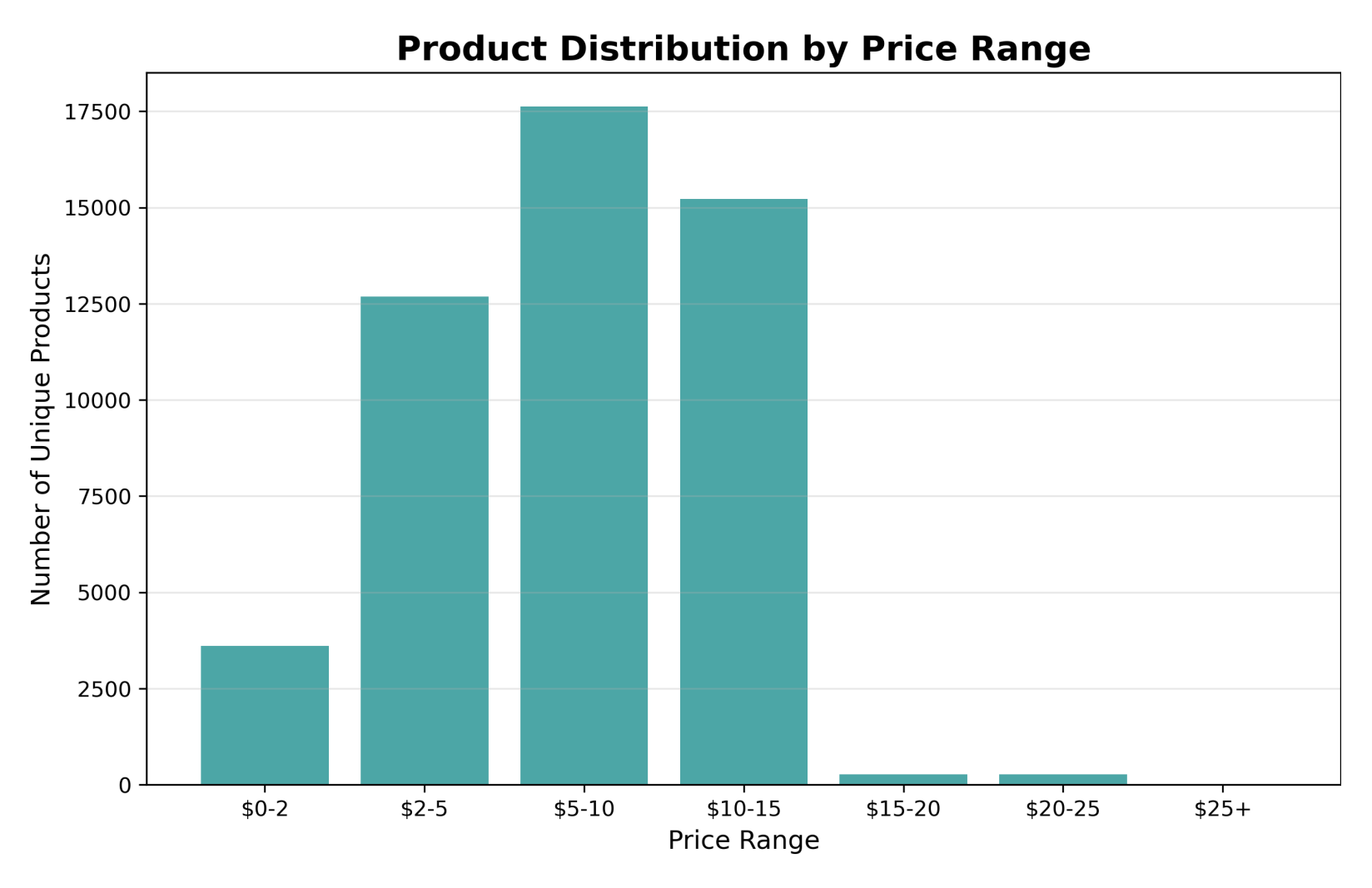
Visualisations

Before diving into the insights and recommendations from our data analysis, let's begin with a brief overview of the business through three key pillars: Products and Inventory, Customer Base and Distribution, and Primary Revenue Sources:

Products and Inventory:

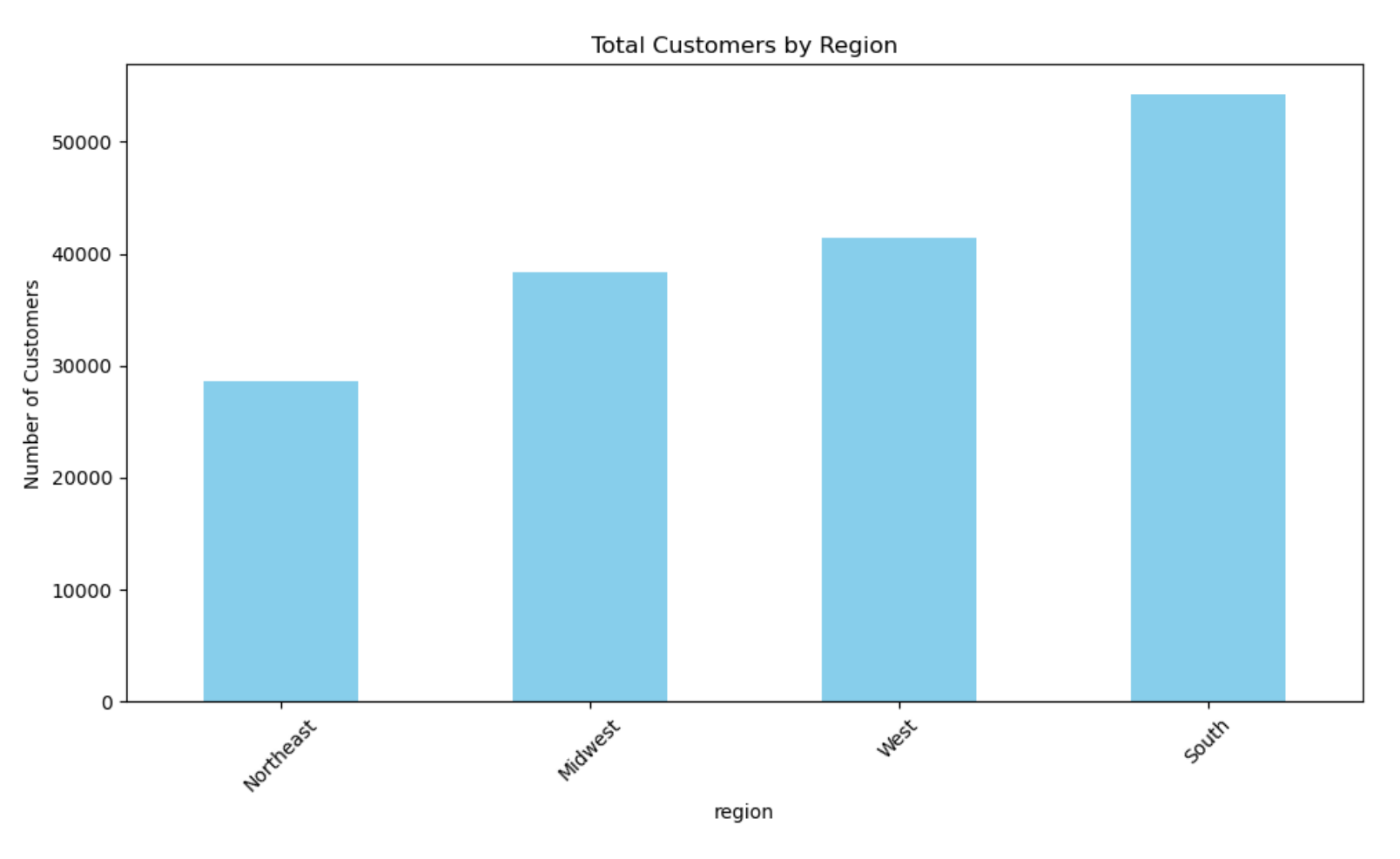


Considering the different categories (Departments), Produce, Daily Eggs, Snacks, and Beverages have the greatest variety of products.

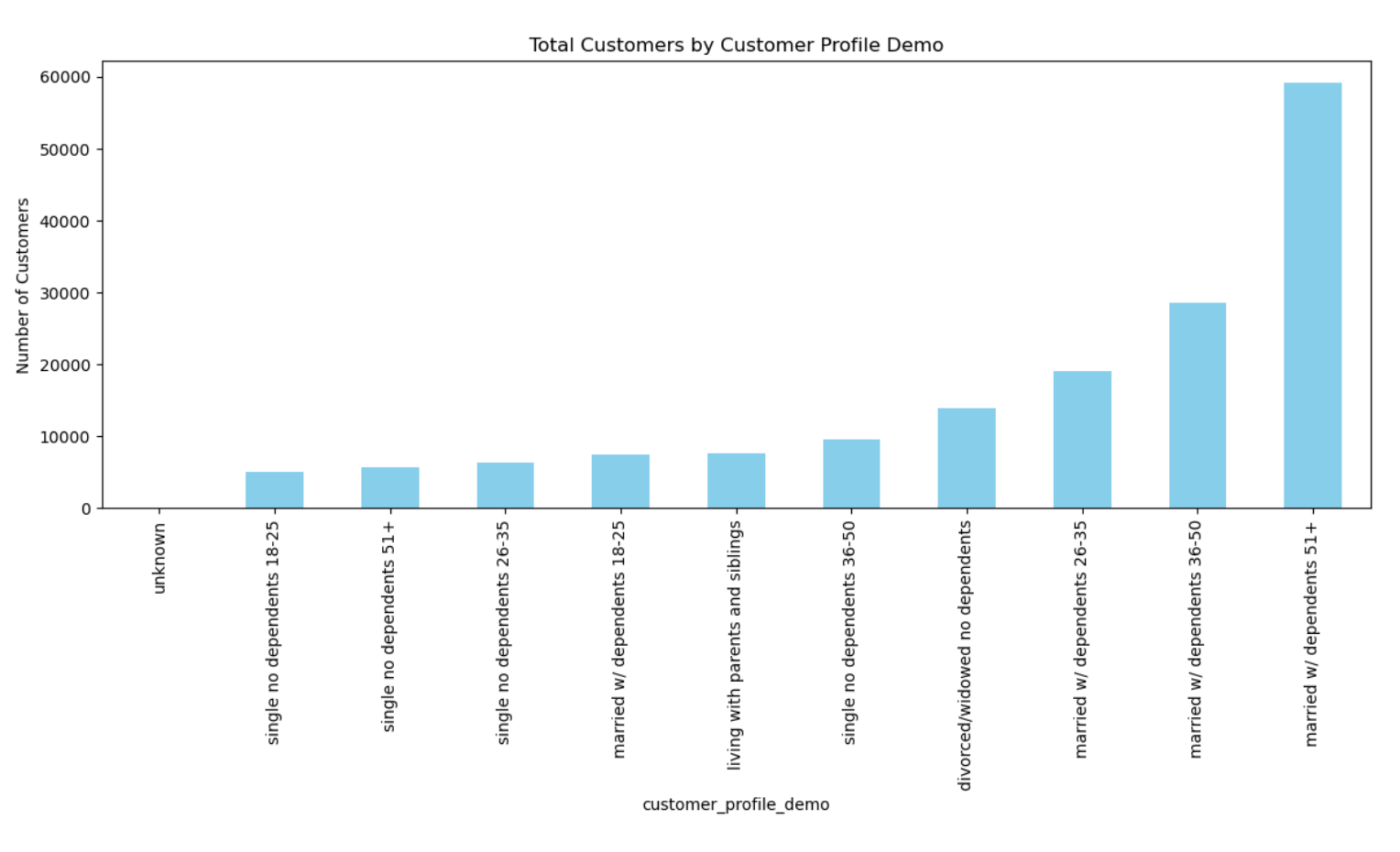


The majority of products are priced between $2 and $15, with a lower offer for products from $15 and more expensive.

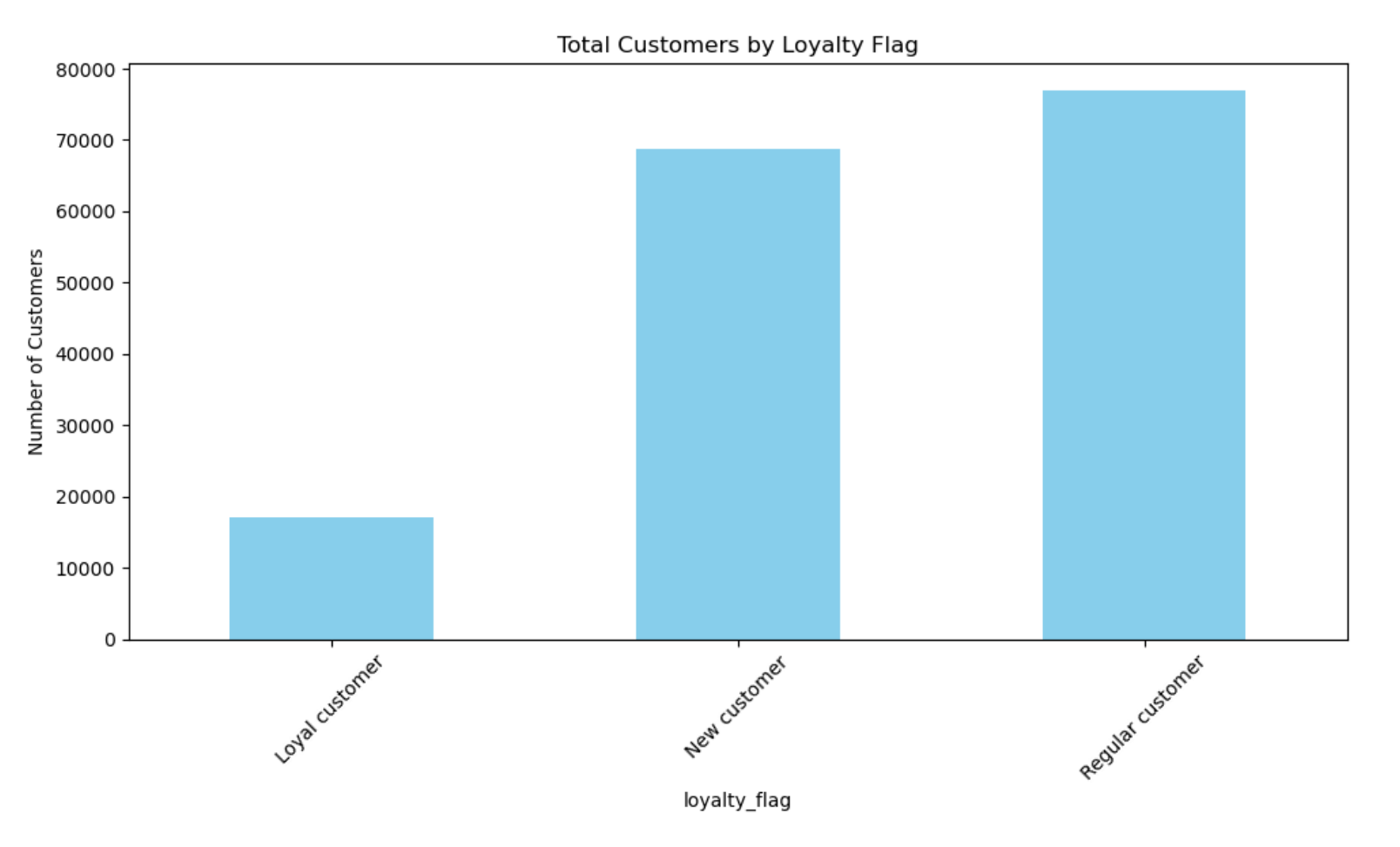
Customer Base and Distribution:



The South region has the highest number of customers, followed by the West, Midwest, and Northeast, in that order.

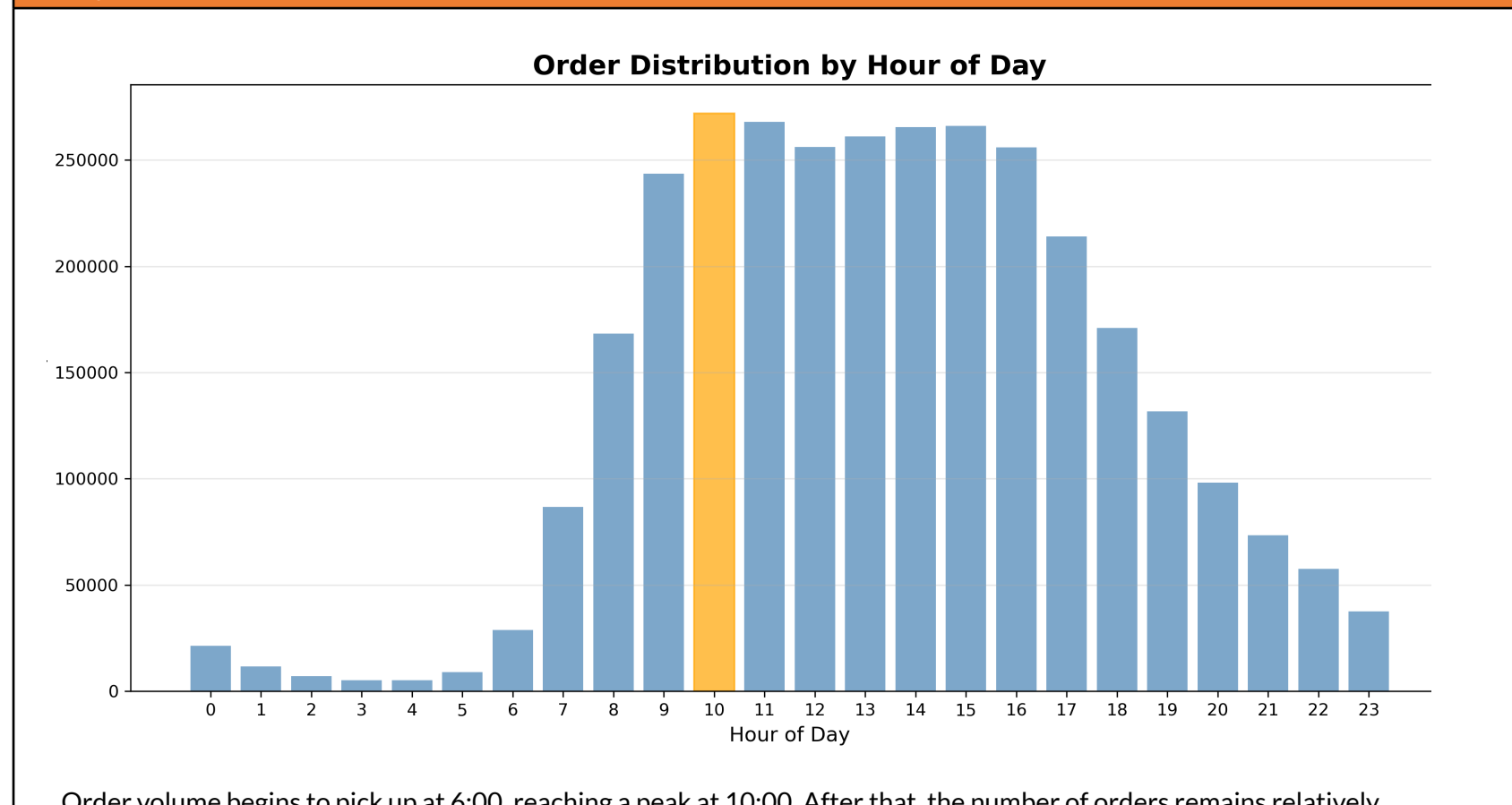


By combining age, family status, and number of dependents, she have created several demographic categories. Overall, customers who are married with dependents constitute the majority.

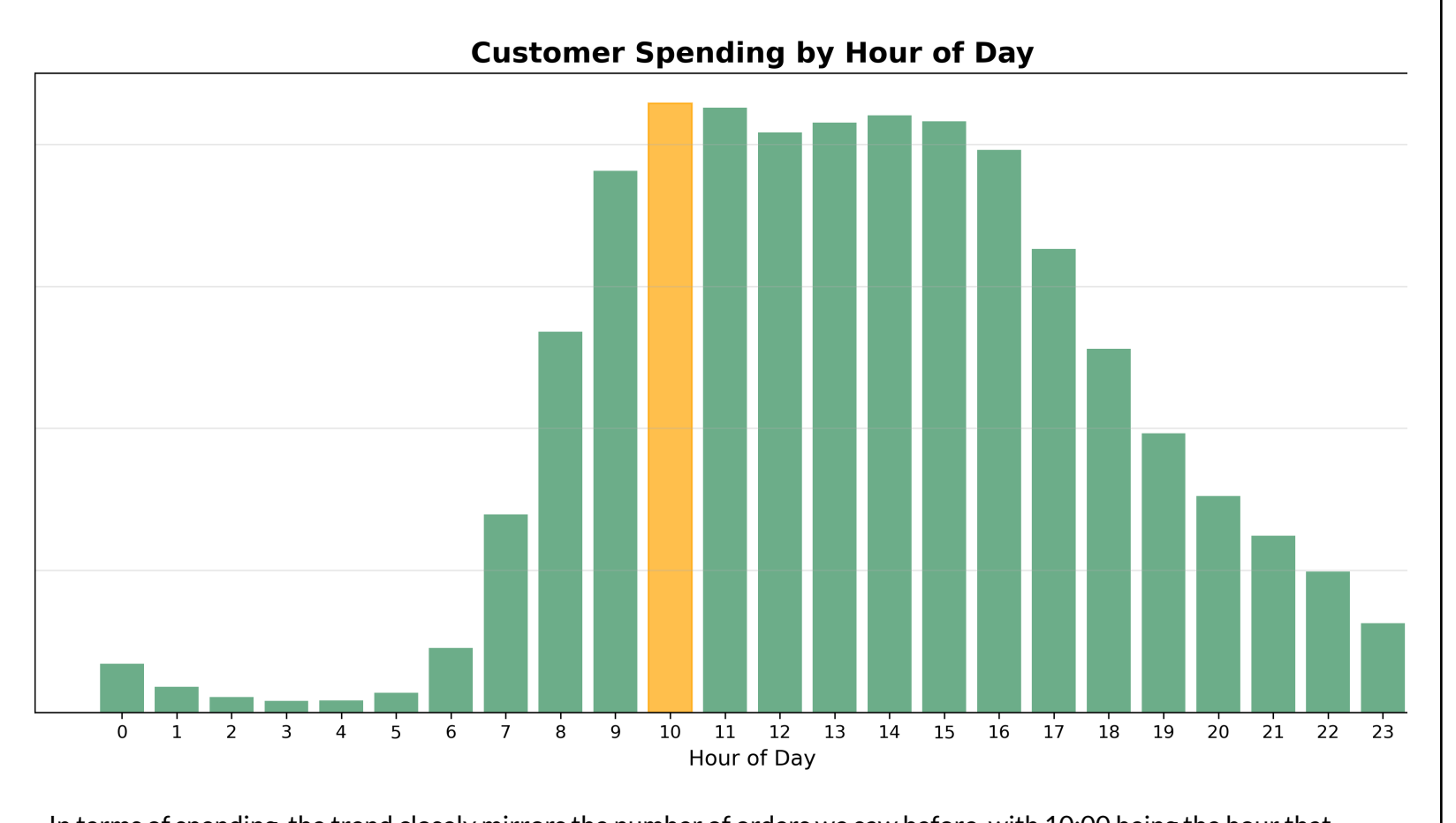


Division of total unique customers by order frequency, following the filters; Loyal customers have more than 40 orders, Regular customers have between 10 and 40 orders, and New customers have 10 or fewer orders.
The majority of customers fall into the Regular customer category, followed by New customers.

Revenue:

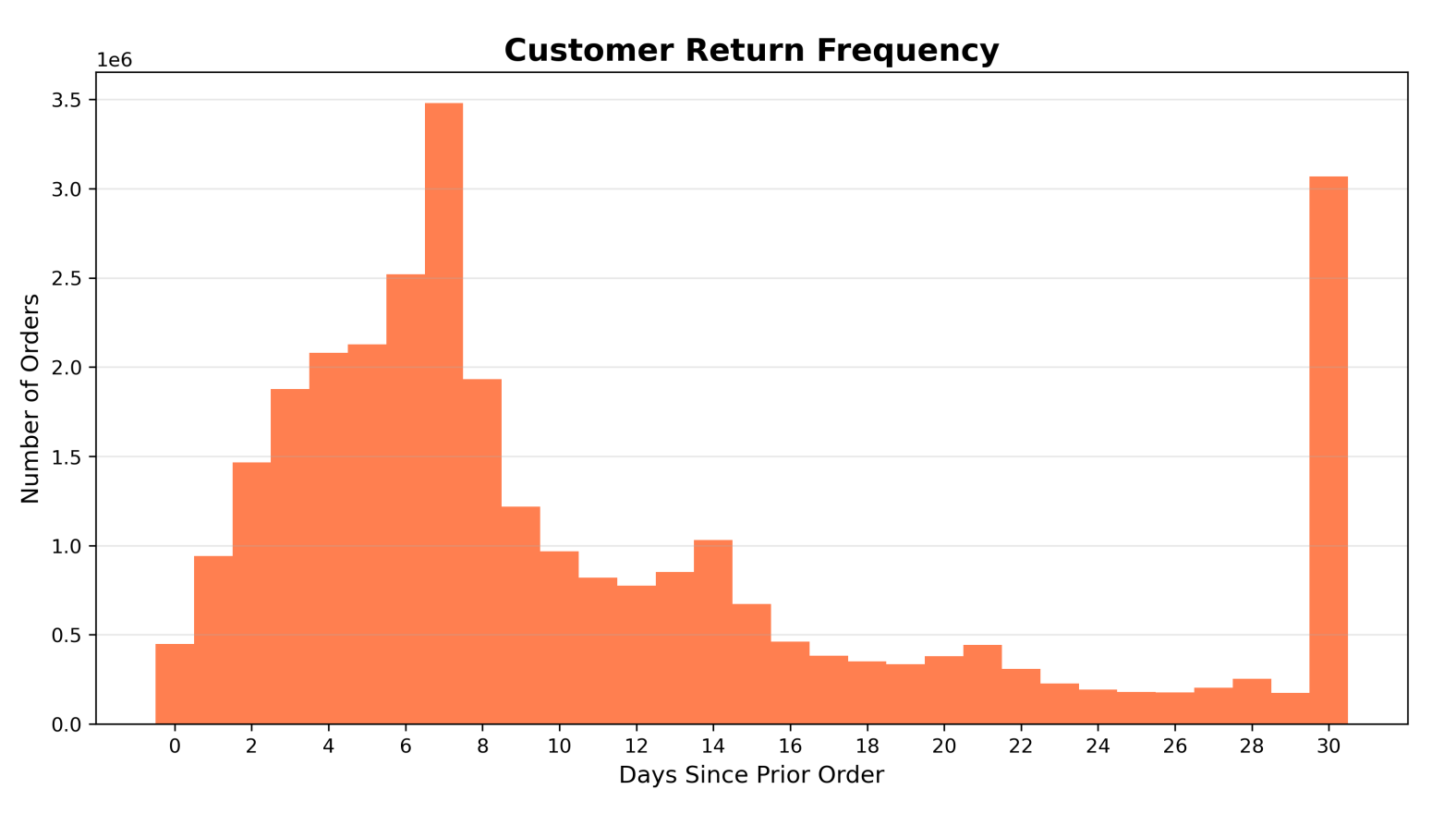


Order volume begins to pick up at 6:00, reaching a peak at 10:00. After that, the number of orders remains relatively stable until 16:00, when it starts to decline steadily until 5:00 next day again.



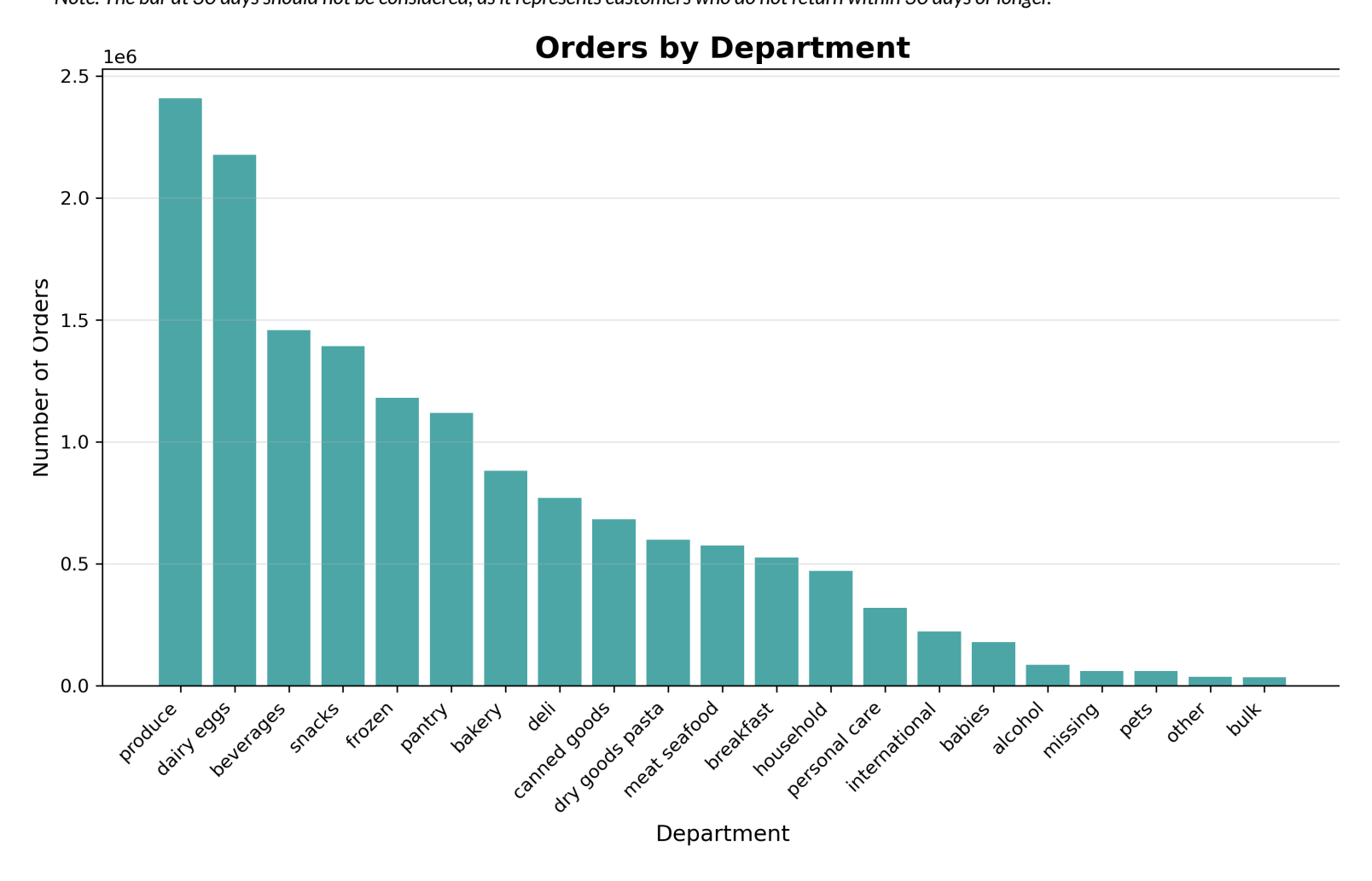
In terms of spending, the trend closely mirrors the number of orders we saw before, with 10:00 being the hour that generates the highest revenue of the day.

There is little difference in the average spending per product (total spent / total number of products ordered) depending on the hour of the day.

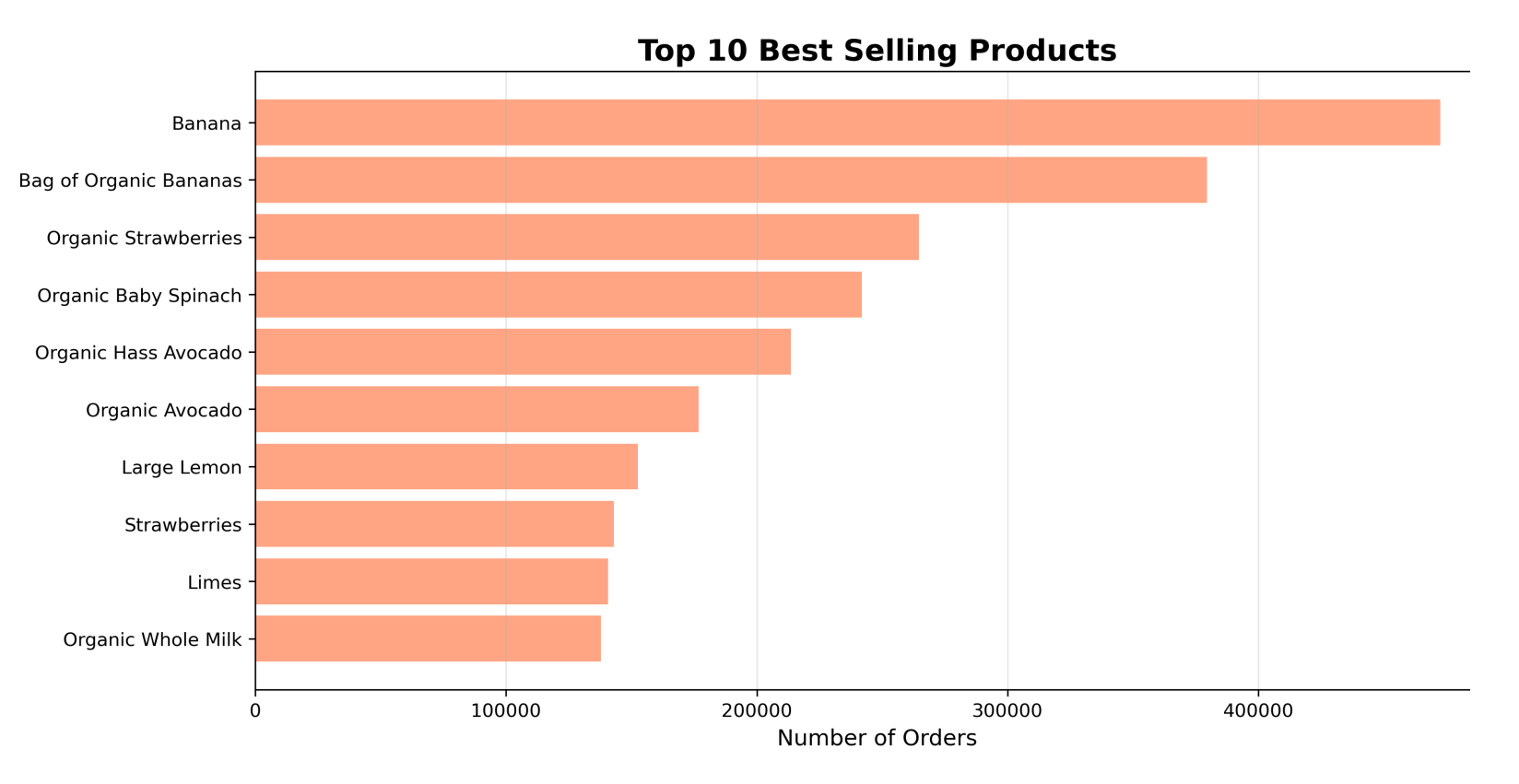


Analyzing the frequency of returning customers, we can see that most customers return on a weekly basis, every 7 days.

**Note: The bar at 30 days should not be considered, as it represents customers who do not return within 30 days or longer.*



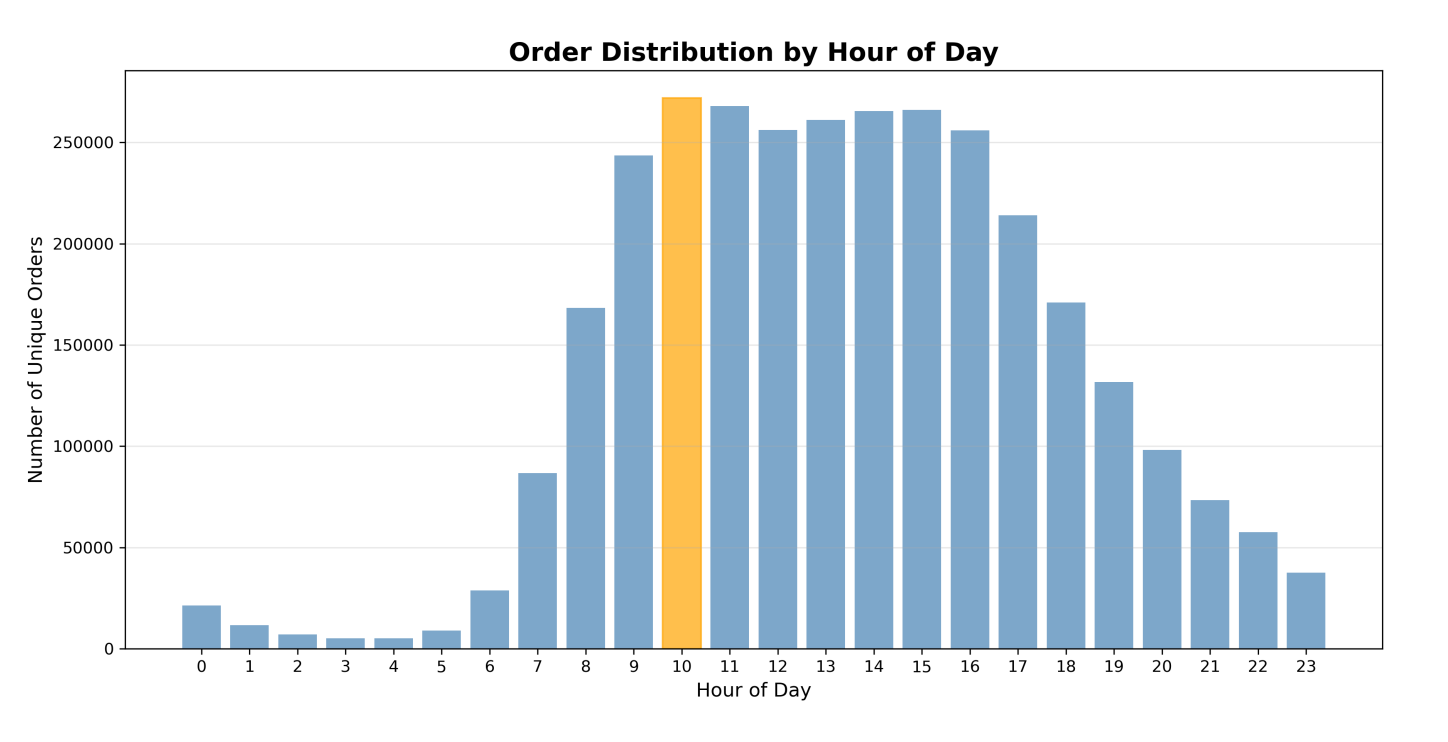
Looking at total orders by department, products from the Produce category are the most sold.



Looking at the products, fruits appear to be the top-selling items overall.

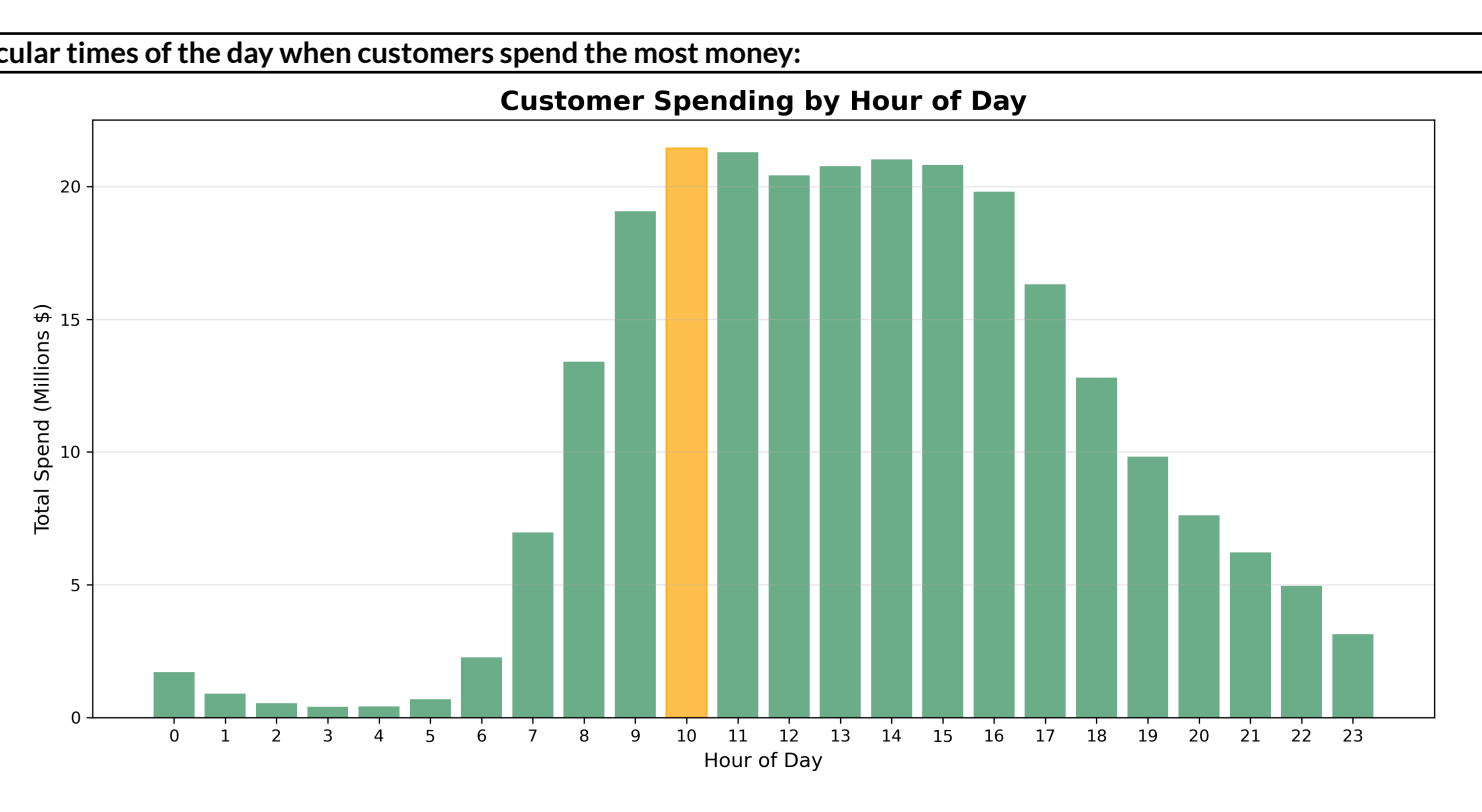
BUSINESS QUESTIONS

Busiest hour is for placing orders:



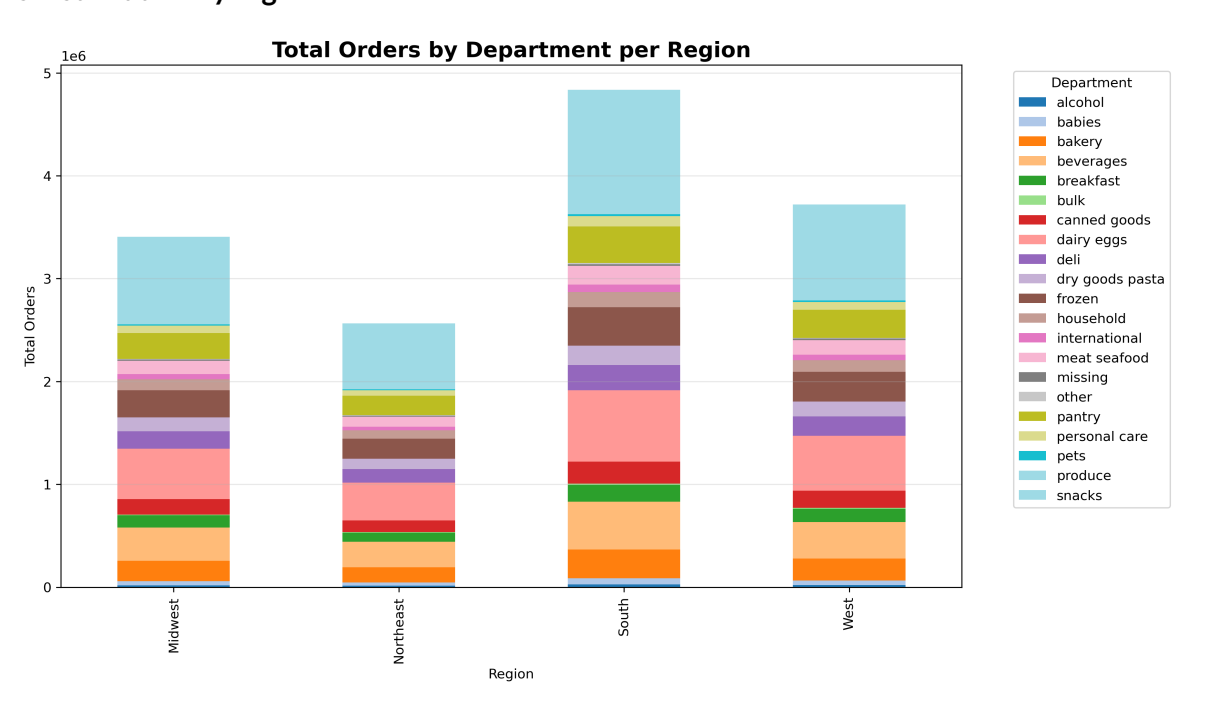
10:00 is the busiest hour.

Particular times of the day when customers spend the most money:



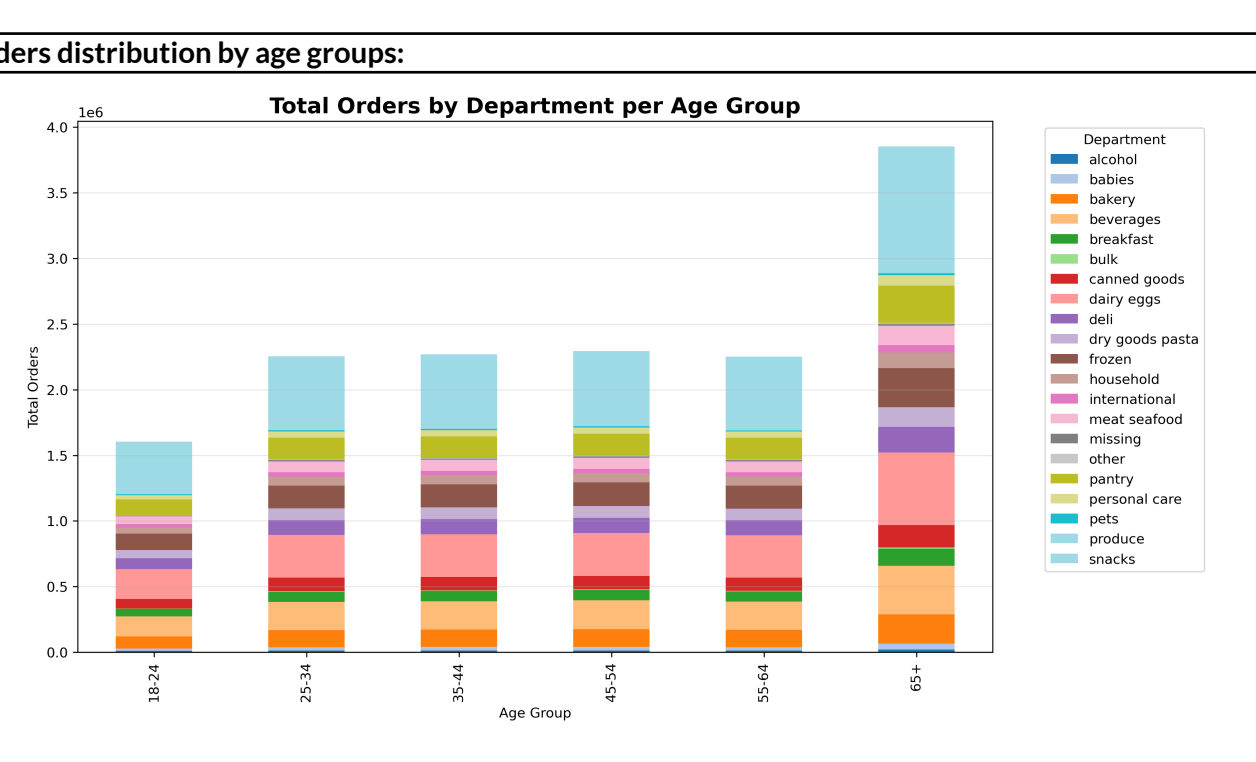
10:00 is the hour when customers spend the most.

Departments orders distribution by region:



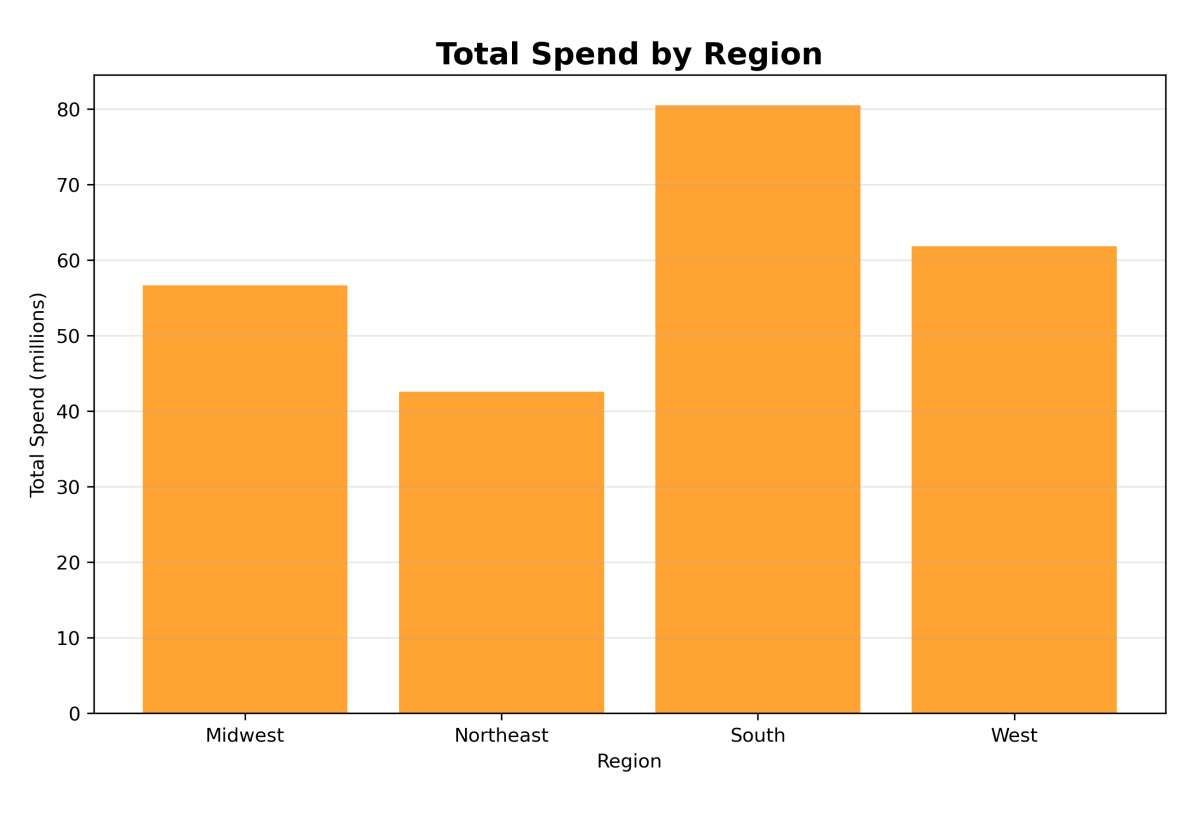
The Produce department leads in sales across all regions.

Departments orders distribution by age groups:



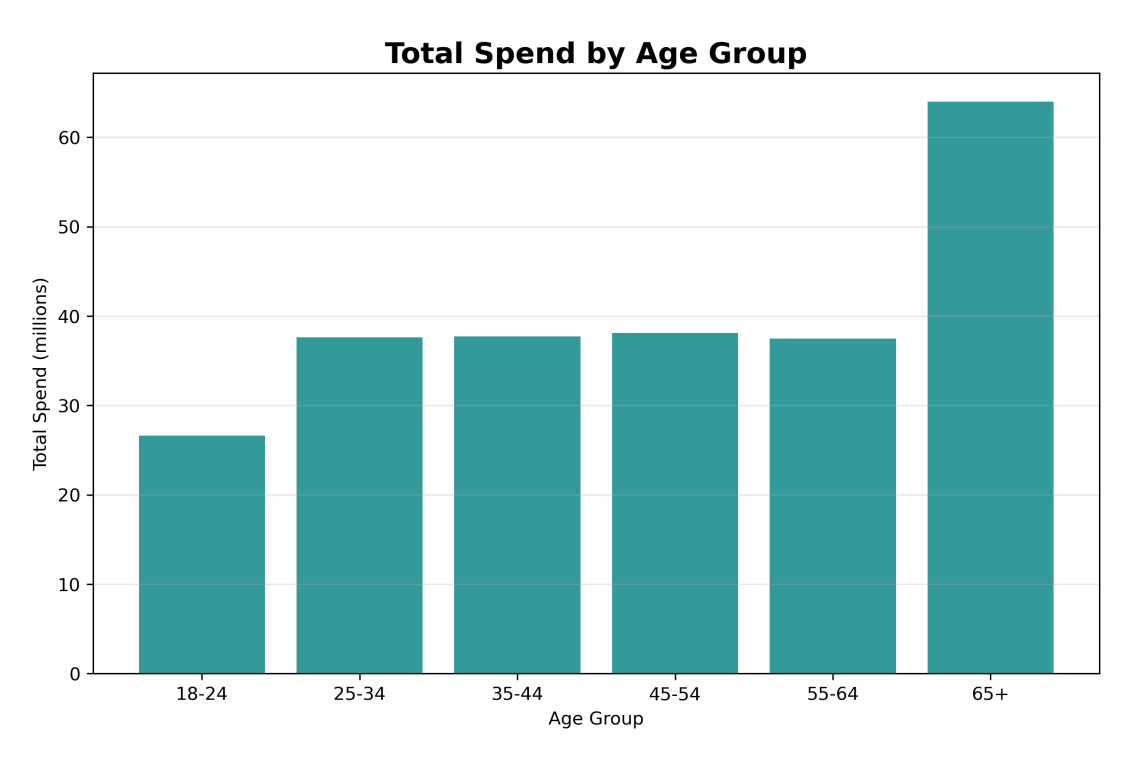
The Produce department is the leader across all age groups.

Total spend by region:



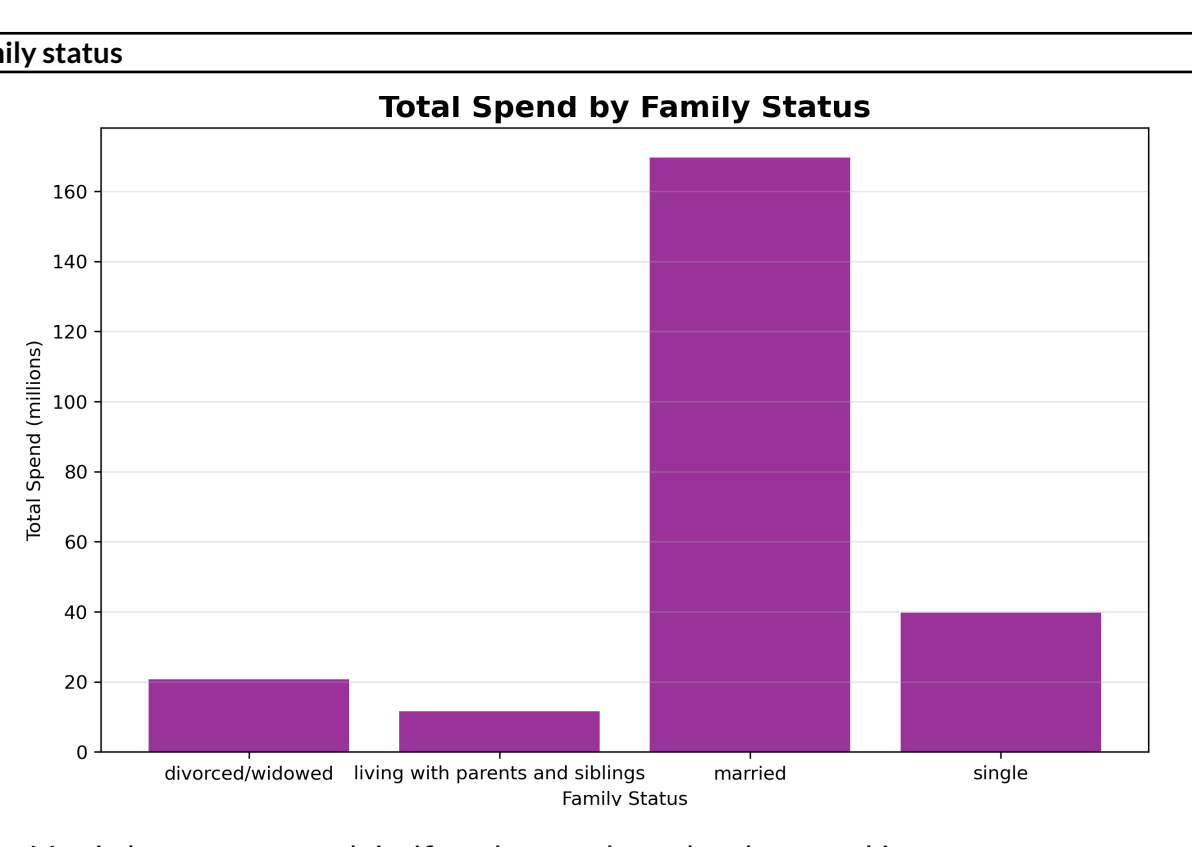
The South is the region generating the most revenue, followed by the West, Midwest, and Northeast, mirroring the same distribution observed in total users.

Total spend by age groups:



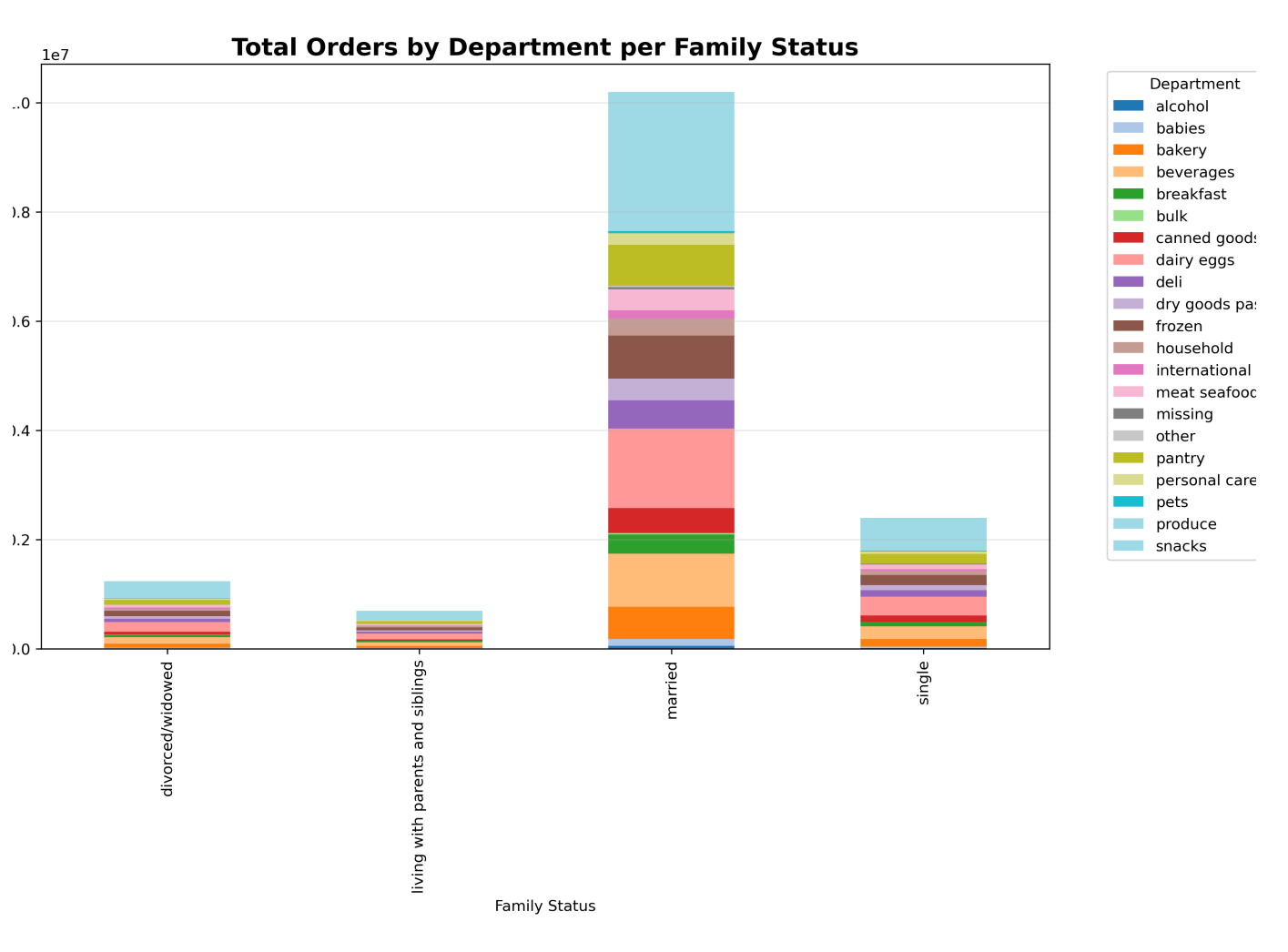
"The 65+ age group has the highest spending.

Total spend by family status



Married customers spend significantly more than other demographic groups.

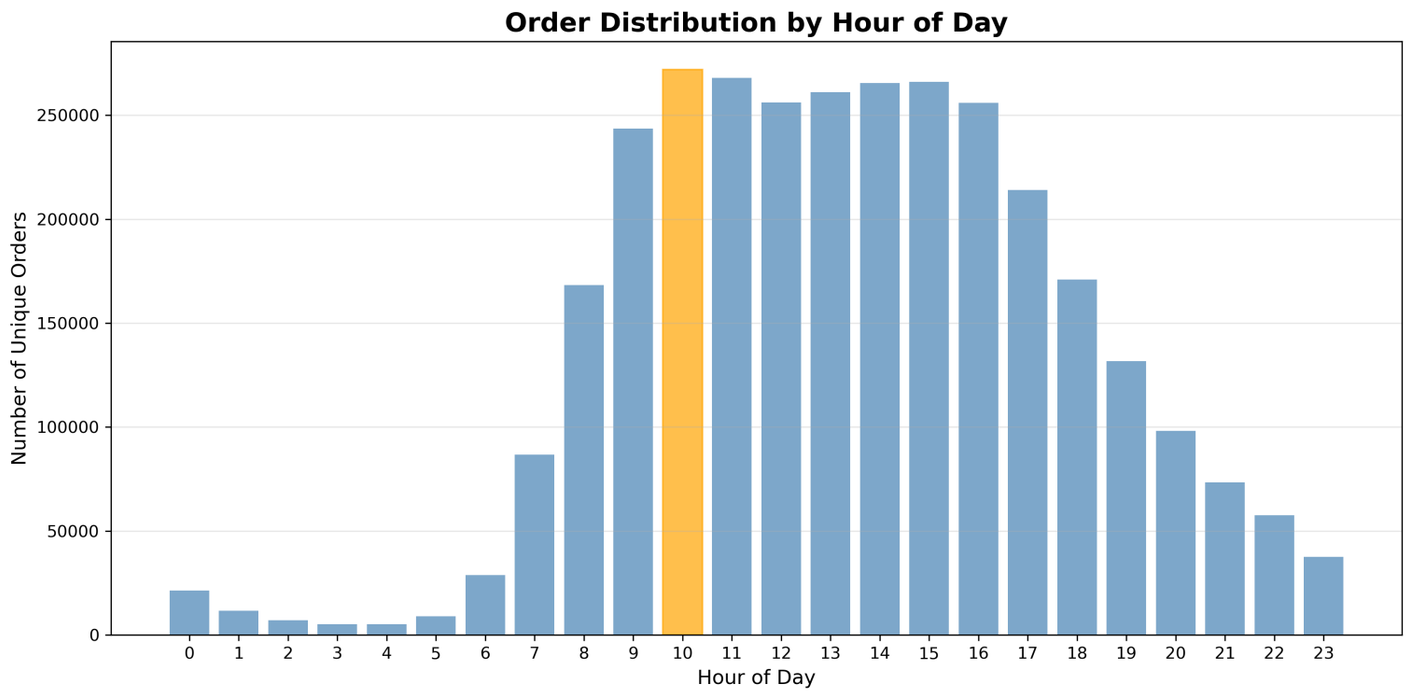
Departments orders distribution by family status:



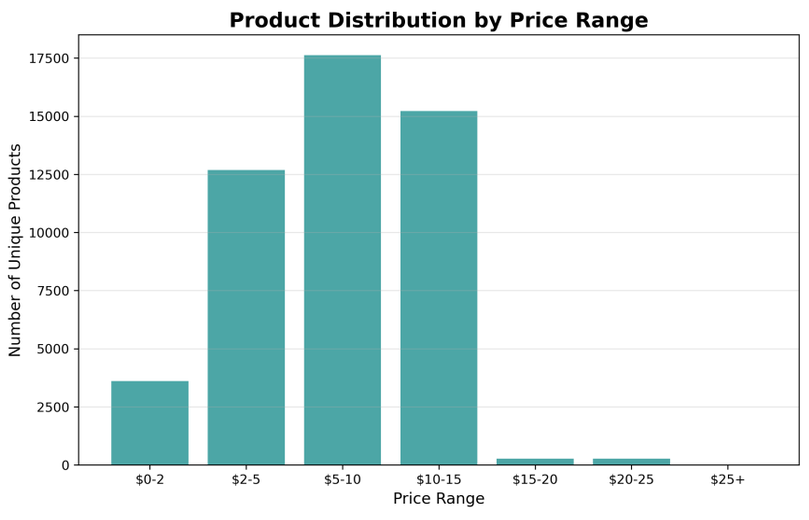
The Produce category leads in total orders, making it the most popular product group among customers.

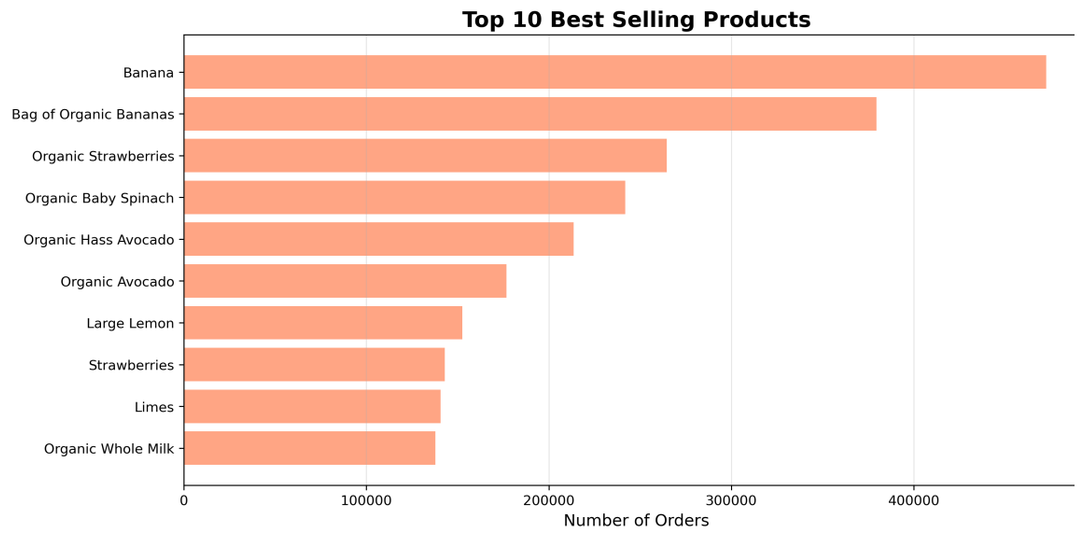
Recommendations

Taking into account the analyzed data, we developed some recommendations for different departments:



Given that peak hours occur between 9:00 and 16:00, it's essential that the Engineering team ensures our services operate flawlessly to prevent technical issues that could result in lost orders or customers.

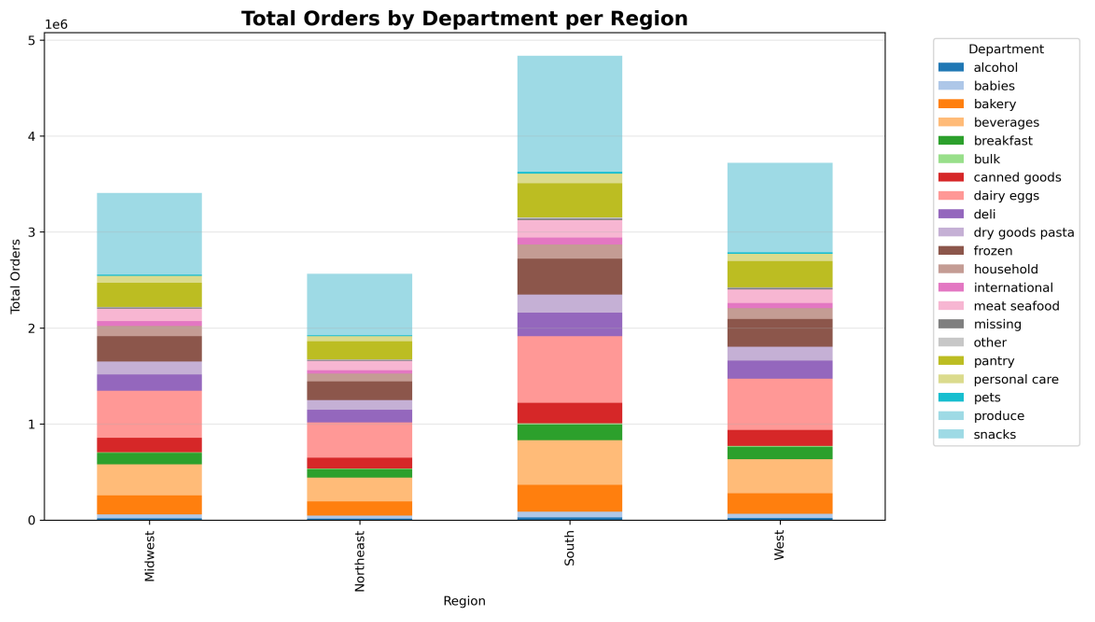




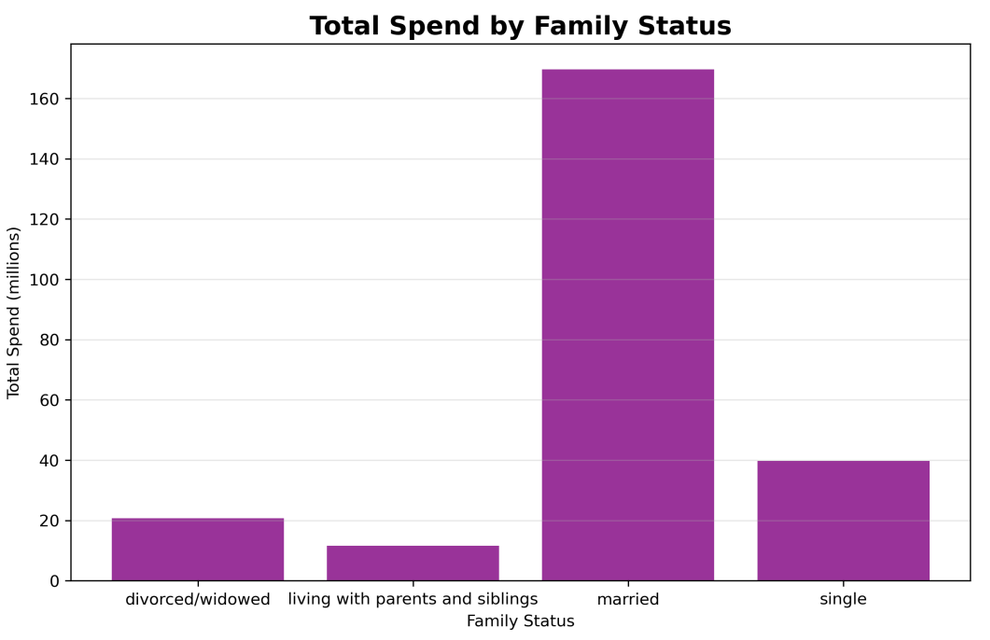
Using the insights about price and most demanded product, the Pricing team can work on a new product strategy:

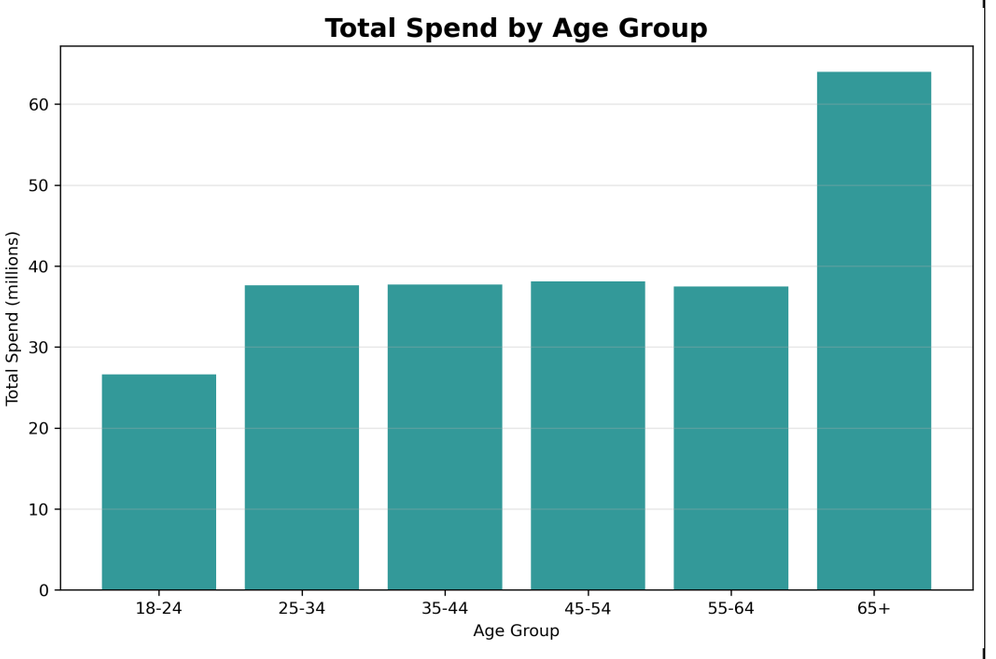
· Increasing variety in the top Produce products, or even in Daily Eggs, Snacks, and Beverages. Focusing on fruits within Produce could be a good start.

· Since most products fall in the $2–$15 range, consider small premium options or packages ($16–$25) to increase revenue. Especially with those cheaper products we saw that sell quite well.



Using the insights about regional behavior, we can run some activities for the different U. S regions. Concentrate marketing in the South region since it's the first revenue source, while optimizing strategies in West, Midwest, and Northeast to capture potential growth based on promotion and most consumed products.





Promotions based on demographic data, customer behavior, and loyalty:

· Target 65+ customers since they are the ones spending the most: We could activate options tailored to this age group, capitalizing on their higher spending tendencies and the peak hours of orders.

· Marriage focused with and without dependents promotions: Family-oriented deals or packages (cross-selling opportunities, we could investigate further complementary products or products that are usually bought together).

·Loyalty-based Incentives: Implement rewards for repeat customers, such as points, discounts, or exclusive offers. Particularly to engage those in the "Regular" group, as well as the New customers.

· Behavior-Driven campaigns: The Marketing team can use purchase frequency, product preferences trends, and time-of-day/hour-of-day trends to design personalized offers that drive engagement during peak hours or even slow periods.