instacart

Instacart Grocery Basket Analysis

Project Name: **Instacart Grocery Basket Analysis**

Date: 2023-08-04

Analyst Name: David Ey

Contents:

Population Flow

Consistency checks

Wrangling steps

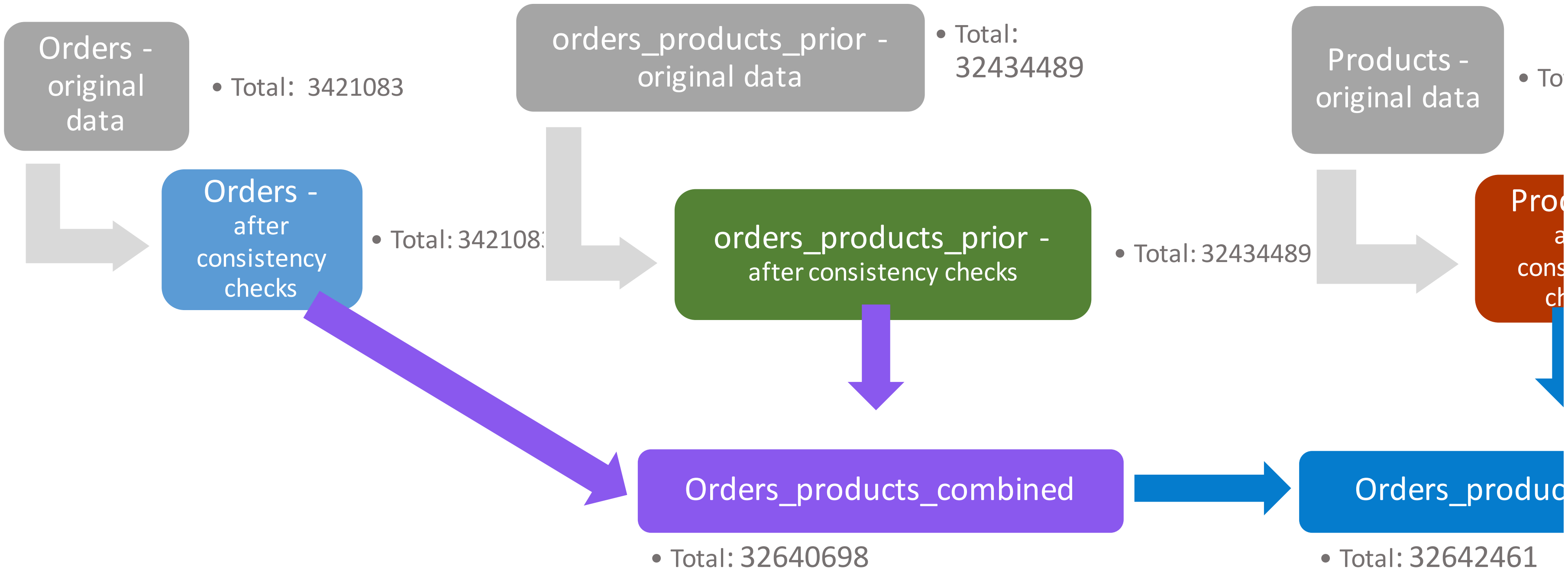
Column derivations

Visualizations

Recommendations

instacart

Population flow



Orders - original data

- Total: 3421083

Orders - after consistency checks

- Total: 342108

orders_products_prior - original data

- Total: 32434489

orders_products_prior - after consistency checks

- Total: 32434489

Products - original data

- To

Pro
a
cons
ch

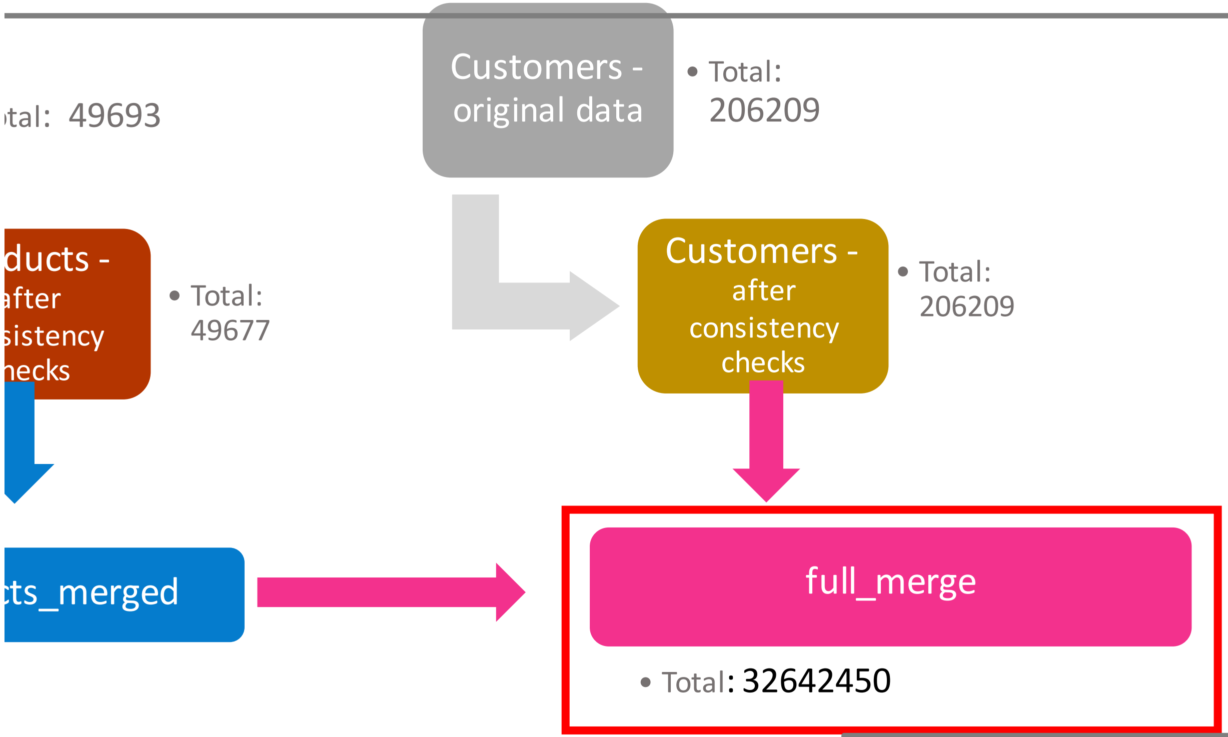
Orders_products_combined

- Total: 32640698

Orders_produc

- Total: 32642461

Title page



ıtal: 49693

Customers - original data

- Total: 206209

ducts - after sistency hecks

- Total: 49677

Customers - after consistency checks

- Total: 206209

cts_merged

full_merge

- Total: 32642450

Exclusion flag

Condition: max_order <
Obervations to be removed:
Final total count of full_merge1:

instacart

Title page

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates
orders	206209	as expected; no action taken	0
products	16	delete	5
orders_products_prior	0	n/a	0
customers	0	n/a	0

instacart

Title page

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
df_ords.drop(columns = ['eval_set'])			no useful information
	df_ords.rename(columns = {'order_dow' : 'orders_day_of_week'}, inplace = True)		clearer name
		df_ords_2['user_id'] = df_ords_2['user_id'].astype('str')	Serves no purpose as a number; is a discrete identifier
	df_ords_3.rename(columns = {'order_number' : 'user_order_number'}, inplace = True)		clearer name
df_ords = df_ords.drop(columns = ['Unnamed: 0'])			artifact cleared. Enacted whenever appeared, also in other dataframes
df_merged_large = df_merged_large.drop(columns = ['reordered'])			no useful information
df_orders_products_combined = df_orders_products_combined.drop(columns = ['_merge'])			was in place for merge flag, but afterwards unnecessary. Dropped all such columns after checks
		ords_prods_merge['order_id'] = ords_prods_merge['order_id'].astype('str')	Serves no purpose as a number; is a discrete identifier

		ords_prods_merge['user_id'] = ords_prods_merge['user_id'].astype('str')	Serves no purpose as a number; is a discrete identifier
		ords_prods_merge['product_id'] = ords_prods_merge['product_id'].astype('str')	Serves no purpose as a number; is a discrete identifier
		ords_prods_merge['aisle_id'] = ords_prods_merge['aisle_id'].astype('str')	Serves no purpose as a number; is a discrete identifier
		ords_prods_merge['department_id'] = ords_prods_merge['department_id'].astype('str')	Serves no purpose as a number; is a discrete identifier
		customer['user_id'] = customer['user_id'].astype('str')	Serves no purpose as a number; is a discrete identifier & needed to merge
full_merge = full_merge.drop(columns = ['first_name'])			Personally identifiable information
full_merge = full_merge.drop(columns = ['last_name'])			Personally identifiable information
full_merge2 = full_merge[full_merge['low_activity']!=True]			As management does not want to look at low-activity customers with fewer than 5 orders, this column was dropped from the remaining data following the separation of the data into subsets.

Column derivations and aggregations

Dataset	New column	lumn/s it was derived fro	Conditions
orders_products_merged	price_label	prices	> 15: High-range product <= 15 & > 5: Mid-range product <= 5: Low-range product
orders_products_merged	busiest_day	orders_day_of_week	0: Busiest day 4: Least busy otherwise: Regularly busy
orders_products_merged	busiest_days	orders_day_of_week	0 or 1: Busiest days 3 or 4: Slowest days otherwise: Average days
orders_products_merged	busiest_period_of_day	user_order_number	>= 9 and <= 16: Most orders 7 or 8 or between 17 and 22 = Average orders otherwise: Fewest orders
orders_products_merged	max_order	user_id user_order_number	maximum value of the user_order_number for each user id
orders_products_merged	loyalty_flag	max_order	> 40: Loyal customer <= 40 and > 10): Regular customer <= 10: New customer
orders_products_merged	user_prices_mean	user_id prices	average value of prices for each user id
orders_products_merged	spender_category	user_prices_mean	< 10: low spender otherwise: high spender

orders_products_merged	days_since_mean	user_id days_since_prior_order	average value of days since prior order for every user id
orders_products_merged	frequency_flag	days_since_mean	>20: Non-frequent customer <= 10: Frequent customer <= 20 and > 10: Regular customer

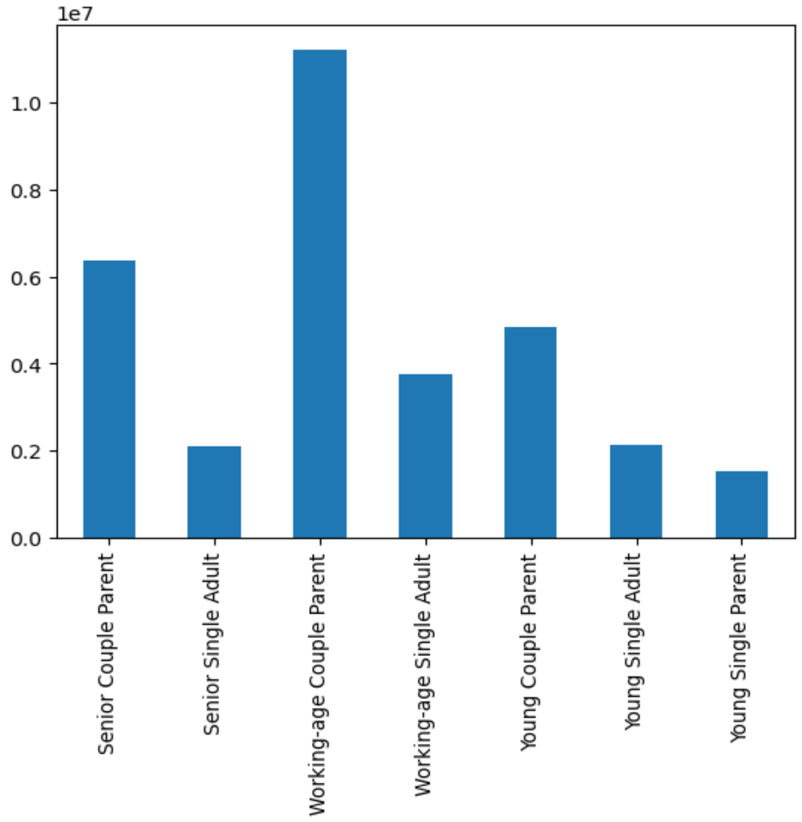
full_merge	region	state	Northeast = ["Maine", "New Hampshire", "Vermont", "Massachusetts", "Rhode Island", "Connecticut", "New York", "Pennsylvania", "New Jersey"]
			Midwest = ["Wisconsin", "Michigan", "Illinois", "Indiana", "Ohio", "North Dakota", "South Dakota", "Nebraska", "Kansas", "Minnesota", "Iowa", "Missouri"]
			South = ["Delaware", "Maryland", "District of Columbia", "Virginia", "West Virginia", "North Carolina", "South Carolina", "Georgia", "Florida", "Kentucky", "Tennessee", "Mississippi", "Alabama", "Oklahoma", "Texas", "Arkansas", "Louisiana"]
			West = ["Idaho", "Montana", "Wyoming", "Nevada", "Utah", "Colorado", "Arizona", "New Mexico", "Alaska", "Washington", "Oregon", "California", "Hawaii"]

full_merge	age_group	age	<= 34: 18-34 > 34 and <= 44: 35-44 > 44 and value <= 54: 45-54 > 54 and value <= 64: 55-64 otherwise 65+
full_merge	income_group	income	< 40000: Low (<$40k) >= 40000 and < 60000: Mid ($40-60k) >= 60000 and < 90000: Mid-High ($60-90k) >= 90000 and < 150000: High ($90-150k) otherwise: Highest ($150k+)
full_merge	consumer_type	department_id	5: Drinker 8: Pet-owner 18: Baby caregiver otherwise: none

full_merge	caregiver_flag	dependants_count	0: False otherwise: True
full_merge	marketing_profile	age_groupcaregiver_flagf	Young Single Adult: 18-34, has no dependants, single or divorced/widowed or living with parents and siblings Young Single Parent: 18-34, has dependants, single or divorced/widowed or living with parents and siblings Young Couple Parent: 18-34, has dependants, married Working-age Single Adult: 35-64, no dependants, single or divorced/widowed or living with parents and siblings Working-age Single Parent: 35-64, has dependants, single or divorced/widowed or living with parents and siblings Working-age Couple Parent: 35-64, has dependants, married Senior Single Adult: 65+: has no dependants, single or divorced/widowed or living with parents and siblings Senior Single Parent: 65+: has dependants, single or divorced/widowed or living with parents and siblings Senior Couple Parent: 65+: has dependants, married

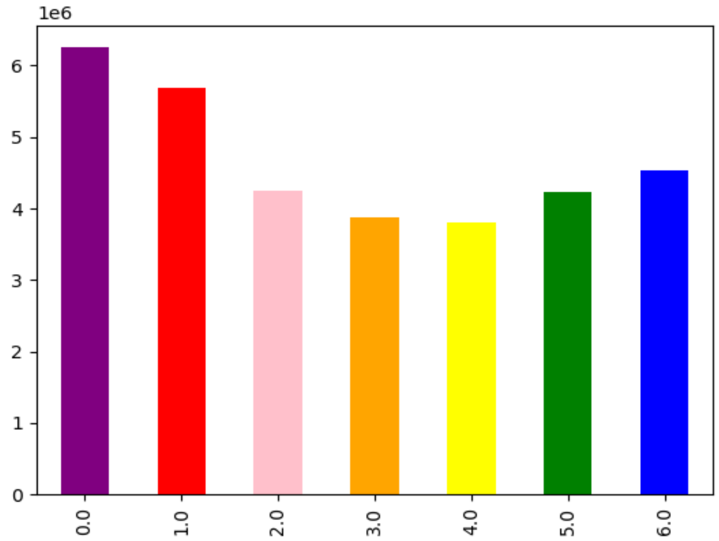
Visualisations

Marketing Profiles:

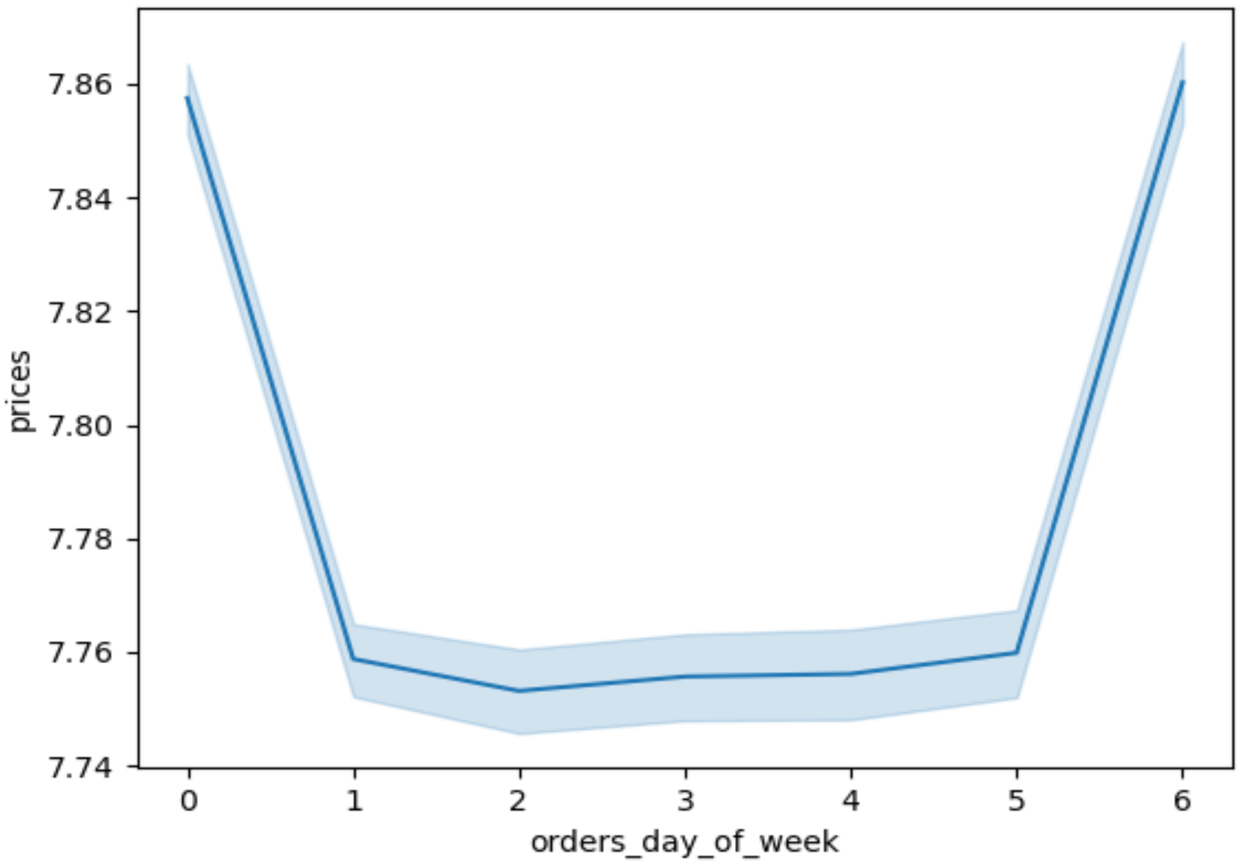


We created the following proifles based on our customer information:
Young Single Adult: 18-34, has no dependants, single or divorced/widowed or living with parents and siblings
Young Single Parent: 18-34, has dependants, single or divorced/widowed or living with parents and siblings
Young Couple Parent: 18-34, has dependants, married
Working-age Single Adult: 35-64, no dependants, single or divorced/widowed or living with parents and siblings
Working-age Single Parent: 35-64, has dependants, single or divorced/widowed or living with parents and siblings. **None of these were found**
Working-age Couple Parent: 35-64, has dependants, married
Senior Single Adult: 65+: has no dependants, single or divorced/widowed or living with parents and siblings
Senior Single Parent: 65+: has dependants, single or divorced/widowed or living with parents and siblings. **None of these were found**
Senior Couple Parent: 65+: has dependants, married

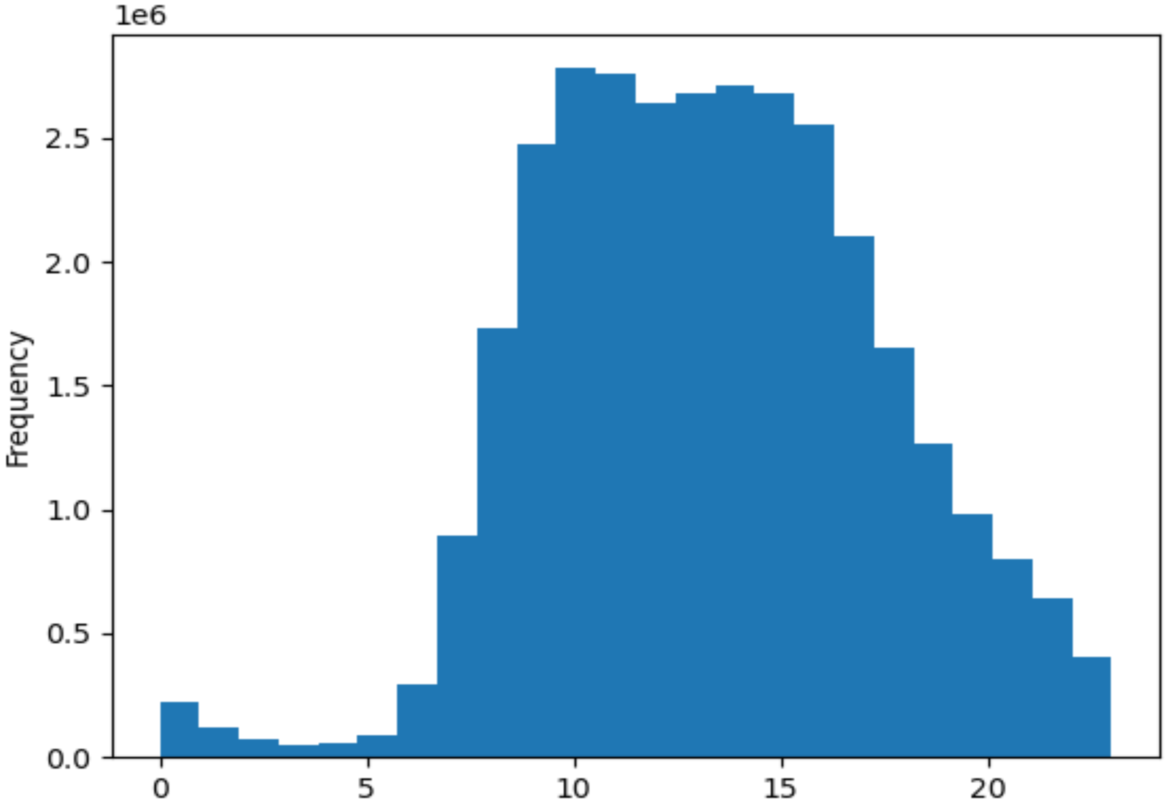
The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.



This chart displays the orders by day of the week (0 is Saturday, 1 is Sunday). Notably Saturday, Sunday, and Friday are the most frequent dates.

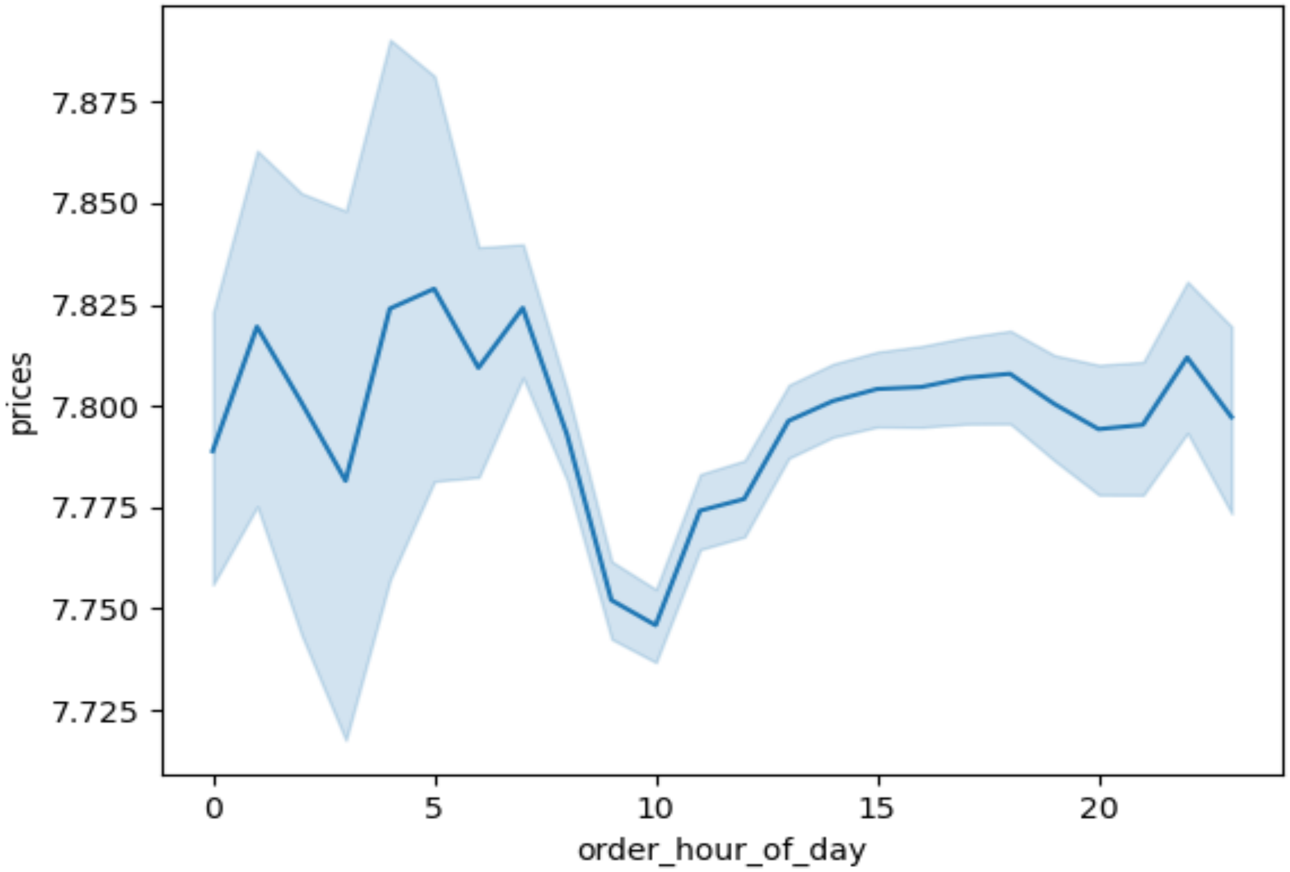


This chart displays the prices paid on the various days, reflecting the Saturday and Sunday have also the highest purchases prices. Note that the variation is tiny; all within 10 cents



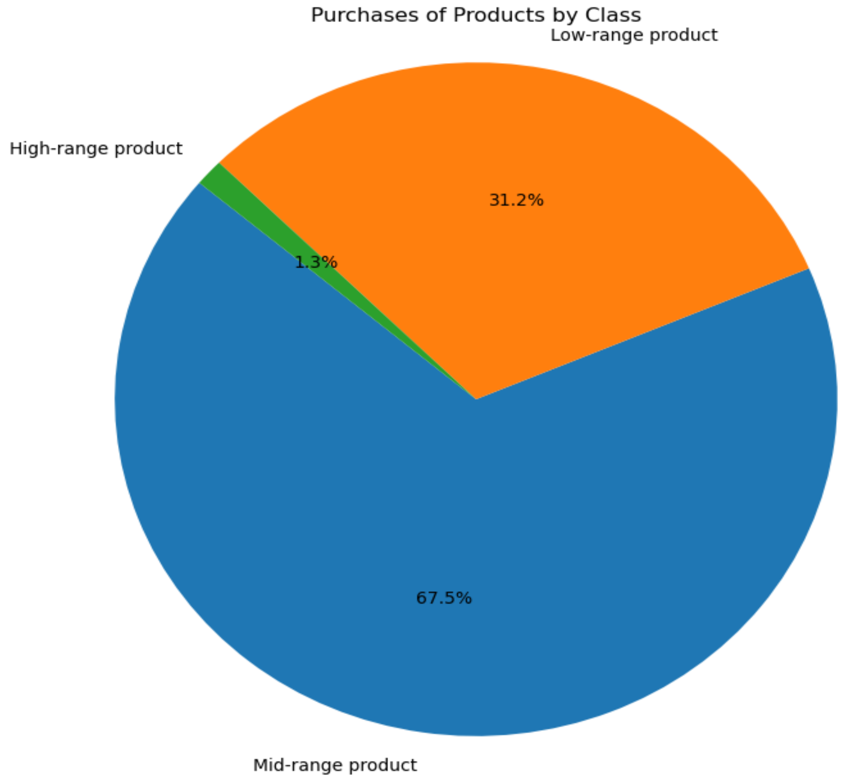
This histogram demonstrates when orders are made and in what frequency in millions.
This shows rather clearly that most orders are made in the middle hours of the day, from 9AM to 5PM

They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.

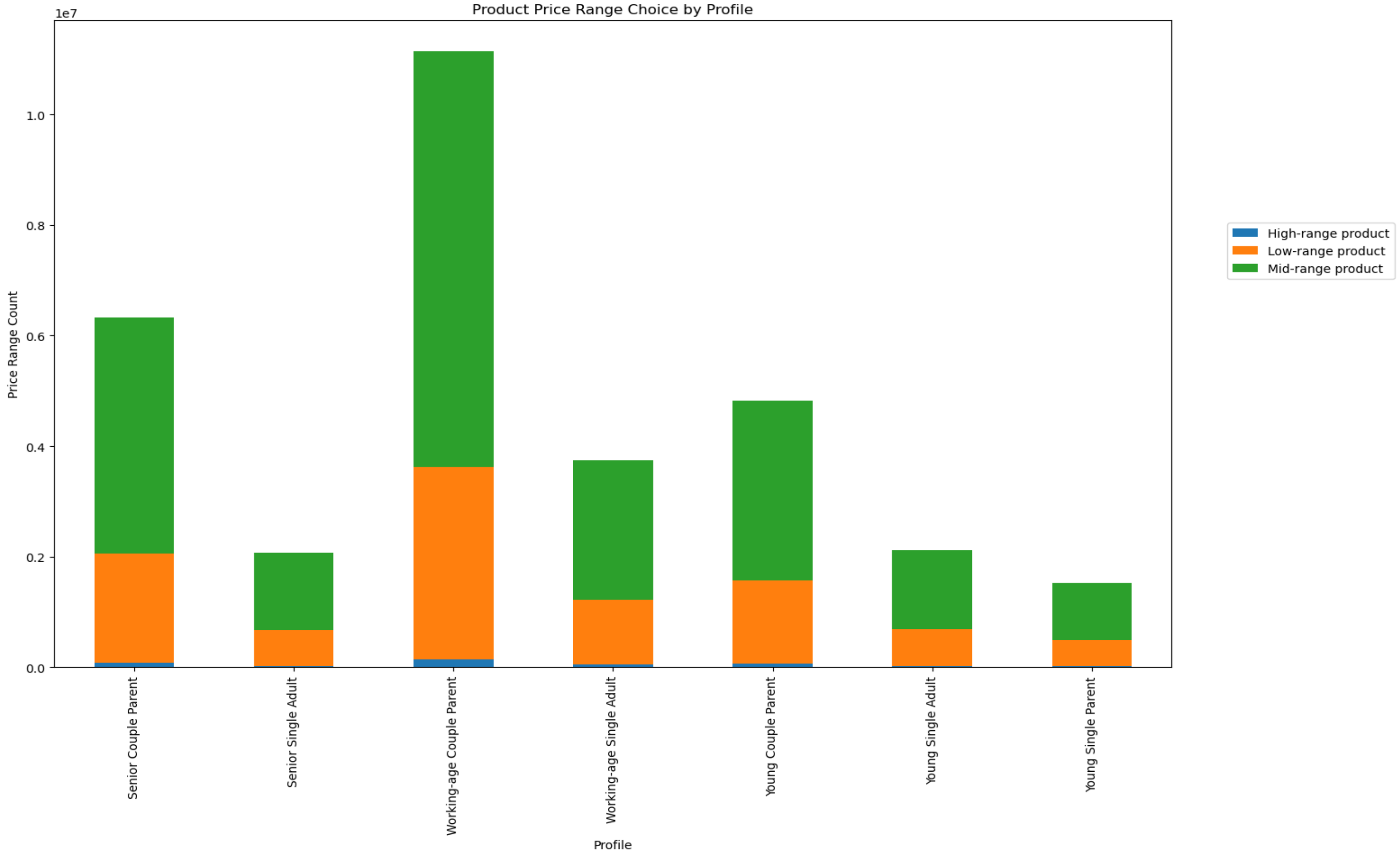


Overall, the average price per order does show some variance over the time of day on average, but notably the difference is within a range of less than 10 cents. There is not a significant variance.

Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.

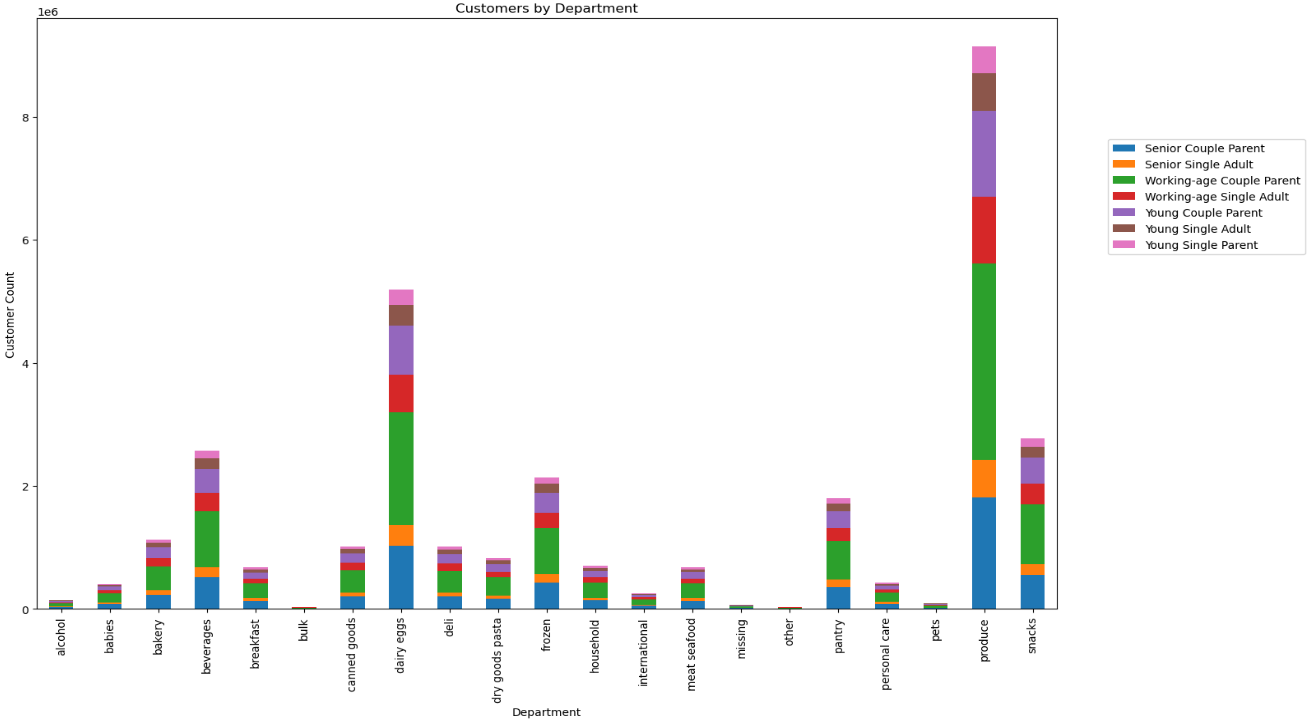


'>15: High-range product
<=15 & >5: Mid-range product
<=5: Low-range product
Mid-range products are by far the most common choice



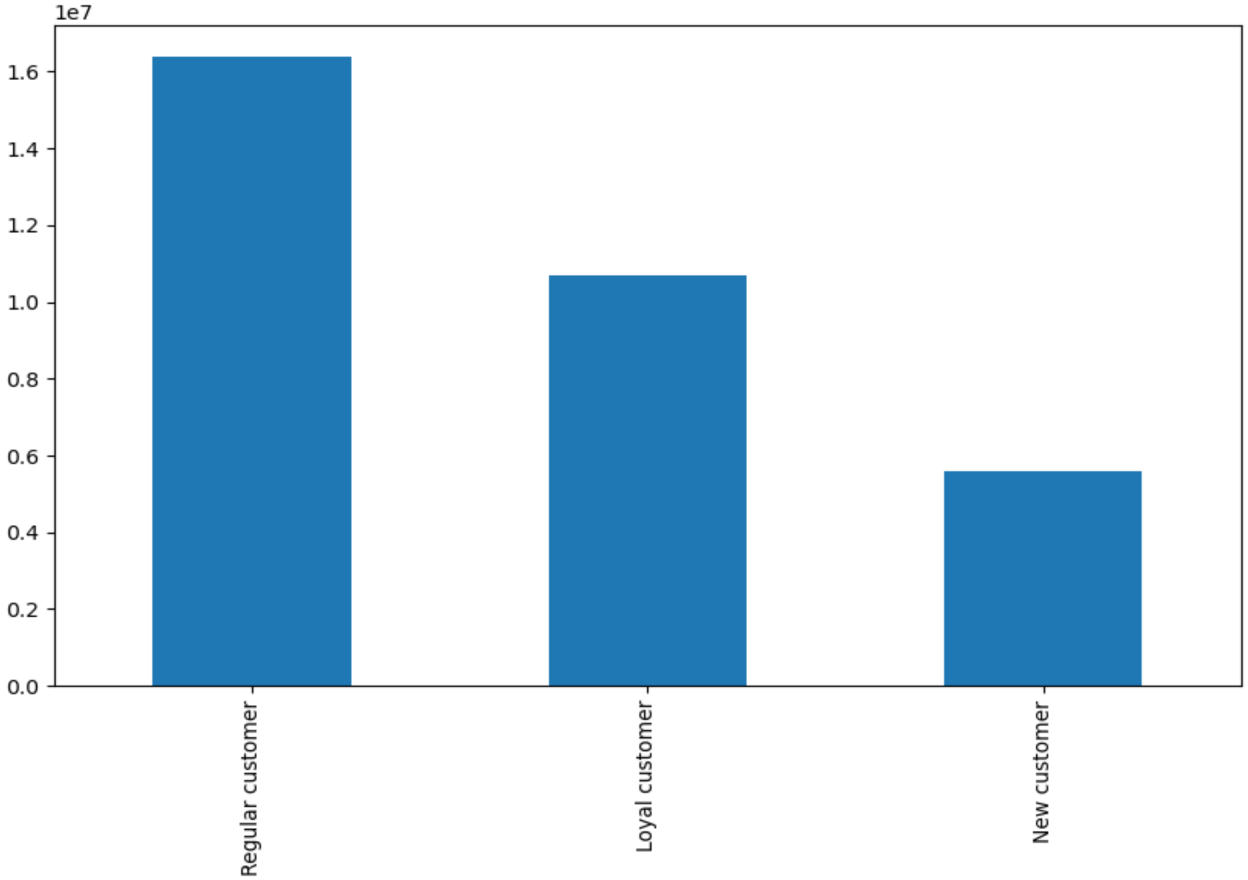
The different profiles buy products at proportional rates.

Are there certain types of products that are more popular than others? The marketing and sales teams want to know which departments have the highest frequency of product orders.



-Only working-age people buy alcohol.
-There are single working-age adults and single seniors who buy baby products, who are not parents.
-Pets, bulk, babies, and alcohol are tiny categories here.

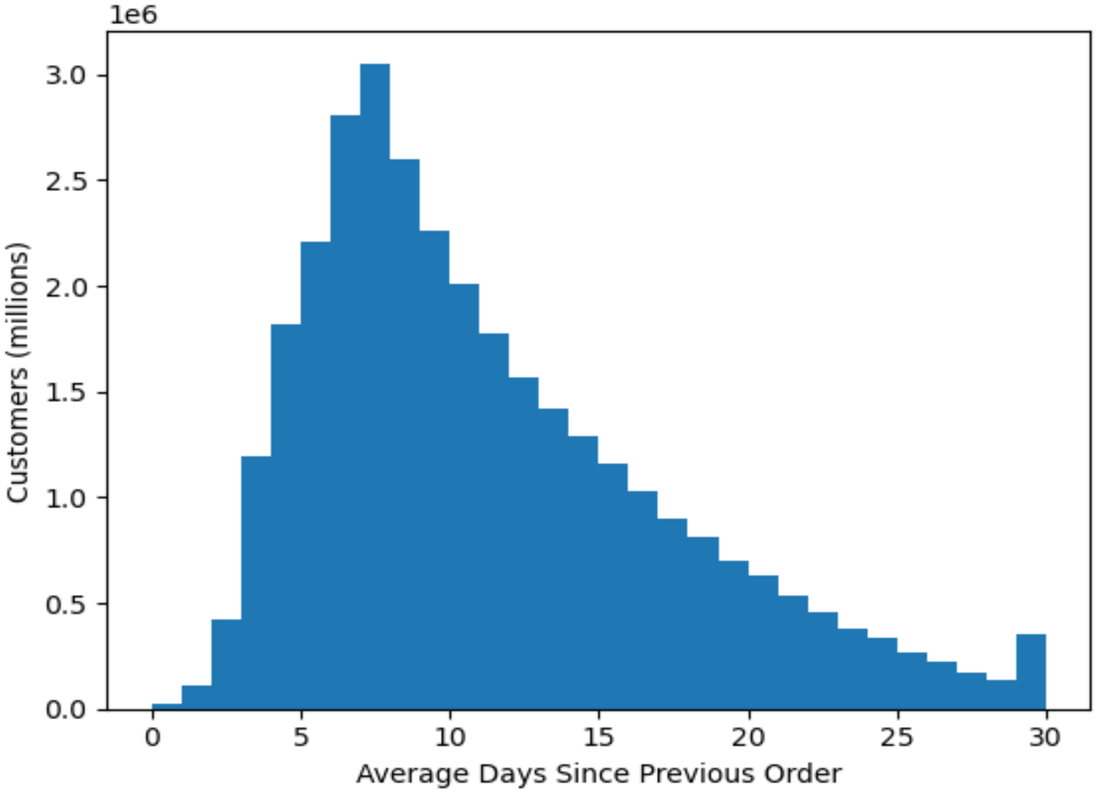
-Dairy eggs, and produce are by far the two largest categories while the meat/seafood is pretty tiny, which could suggest that these customers are more likely to be vegetarian. However the canned goods and frozen departments are also sizable, and would best be broken down into further categories in order to make the best judgment there.



Though the categories of babies, pets, and alcohol are tiny in terms of overall orders, baby caregivers are the largest of these three specialized columns. This also suggests that most dependants are not babies.

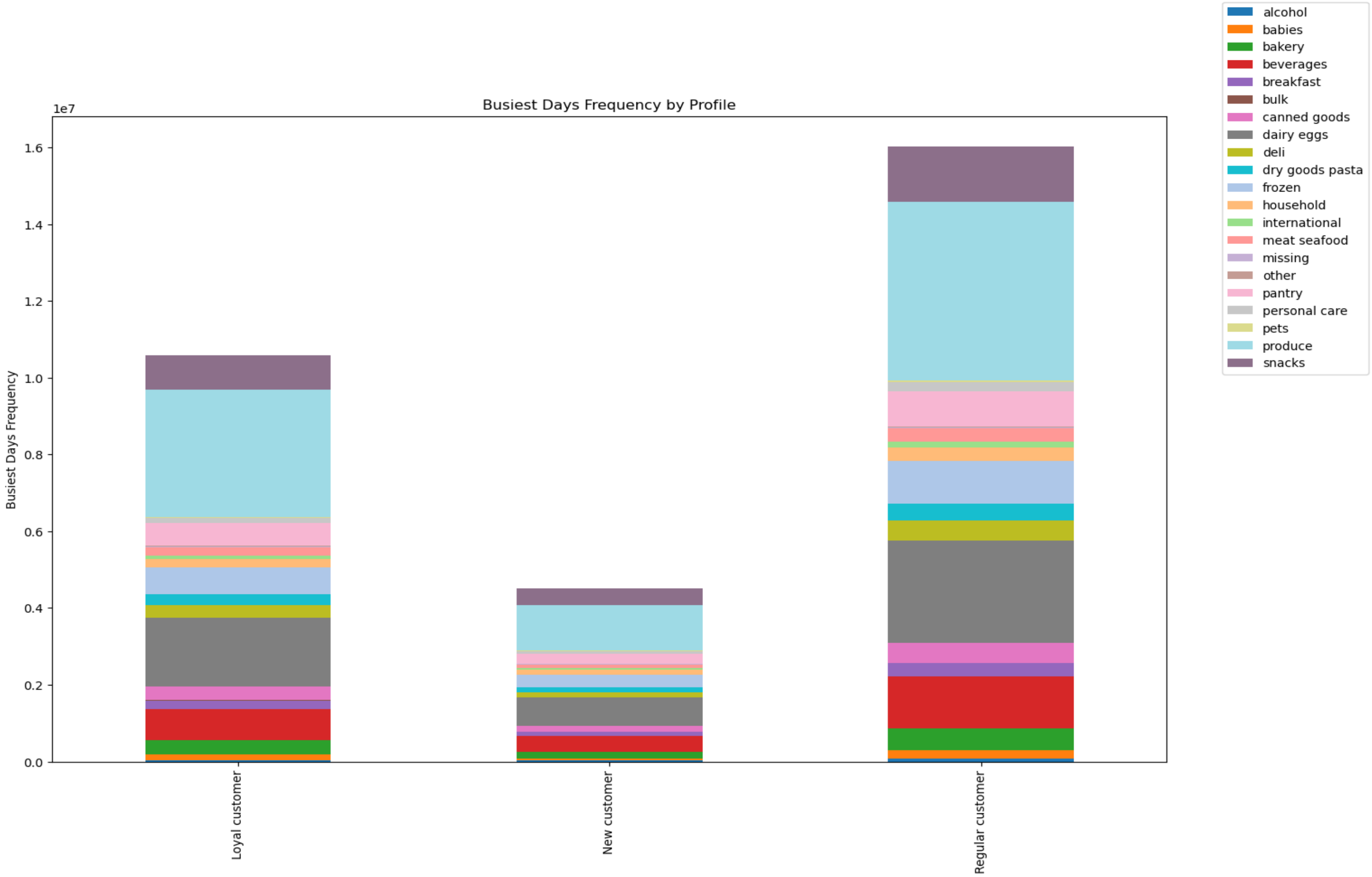
The marketing and sales teams are particularly interested in the different types of customers in their system and how their ordering behaviors differ. For example:

What's the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?

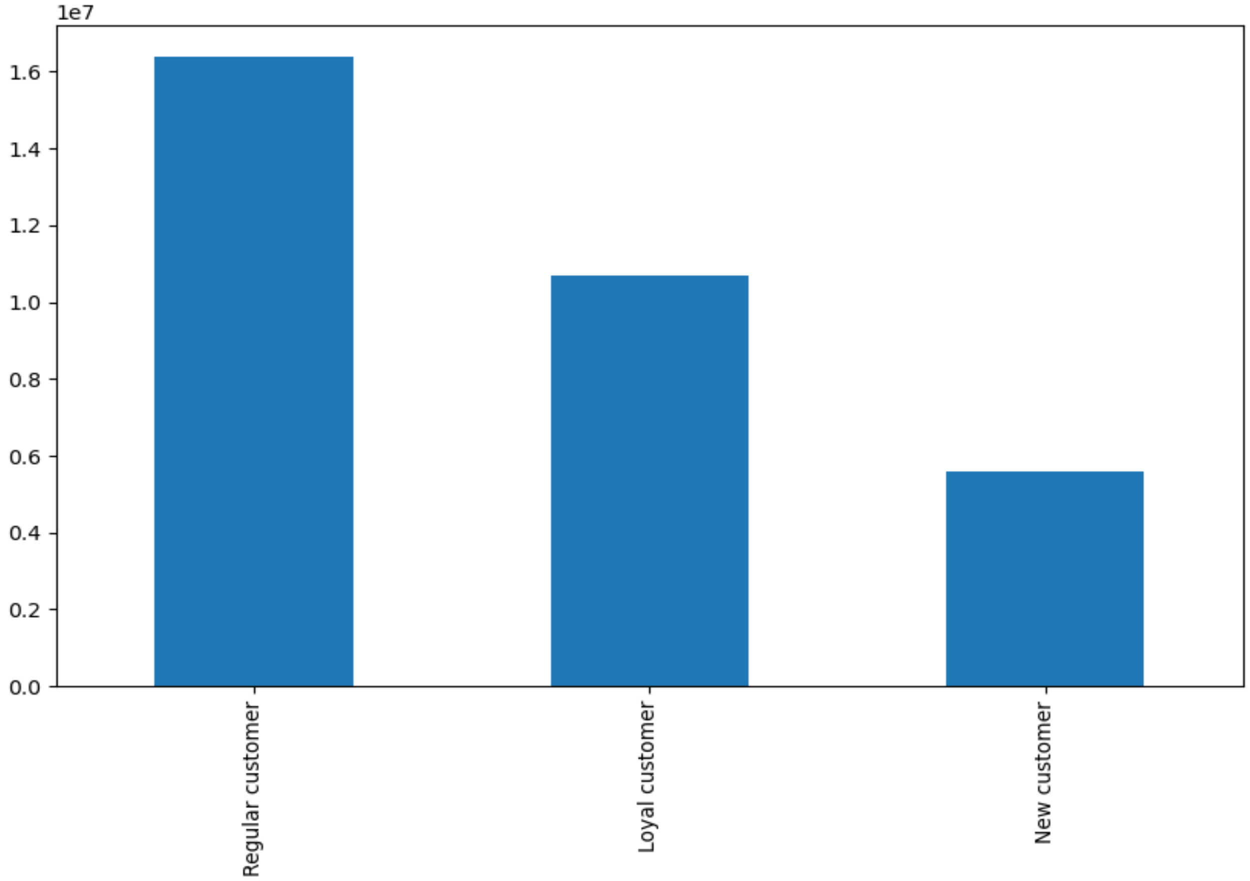


This histogram demonstrates the distribution of the average amount of days customers order from Instacart. Most shop approximately weekly, as shown by the skew.

Are there differences in ordering habits based on a customer’s loyalty status?

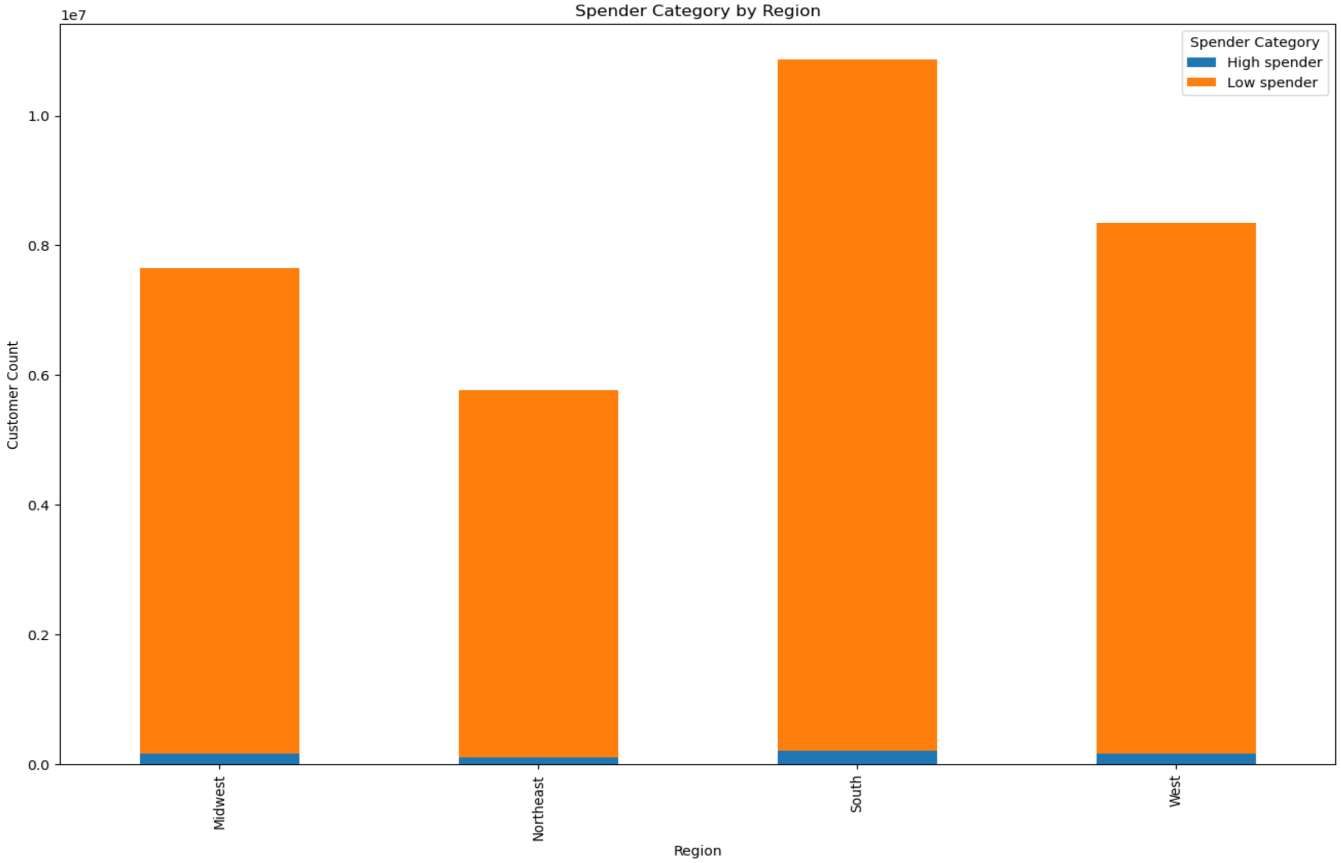


All of this is very proportional.

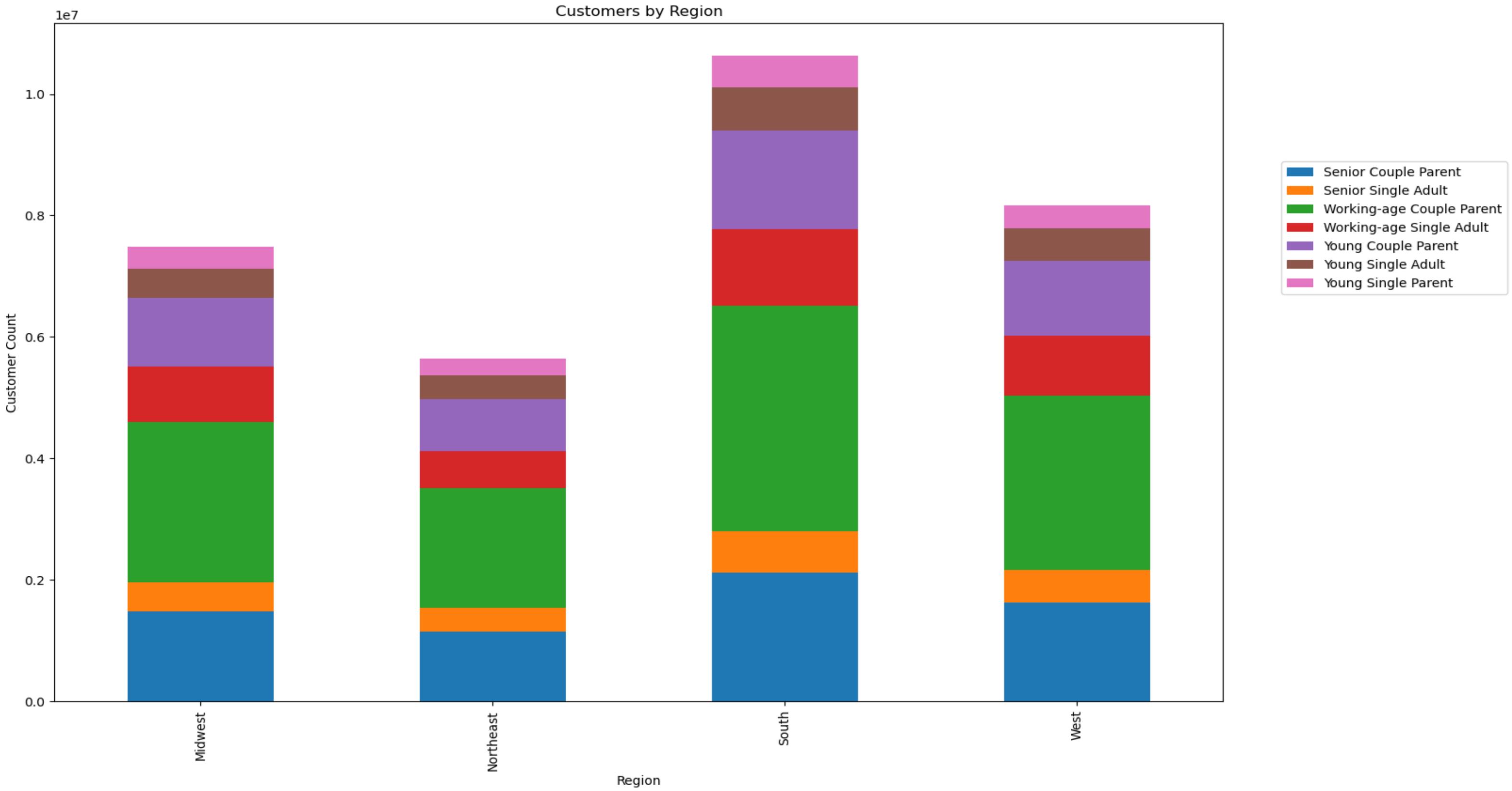


Most users are regular customers (more than 10 and less than 40 orders), while loyal customers (more than forty orders) and new customers (less than 10 orders) are smaller categories.

- **Are there differences in ordering habits based on a customer's region?**

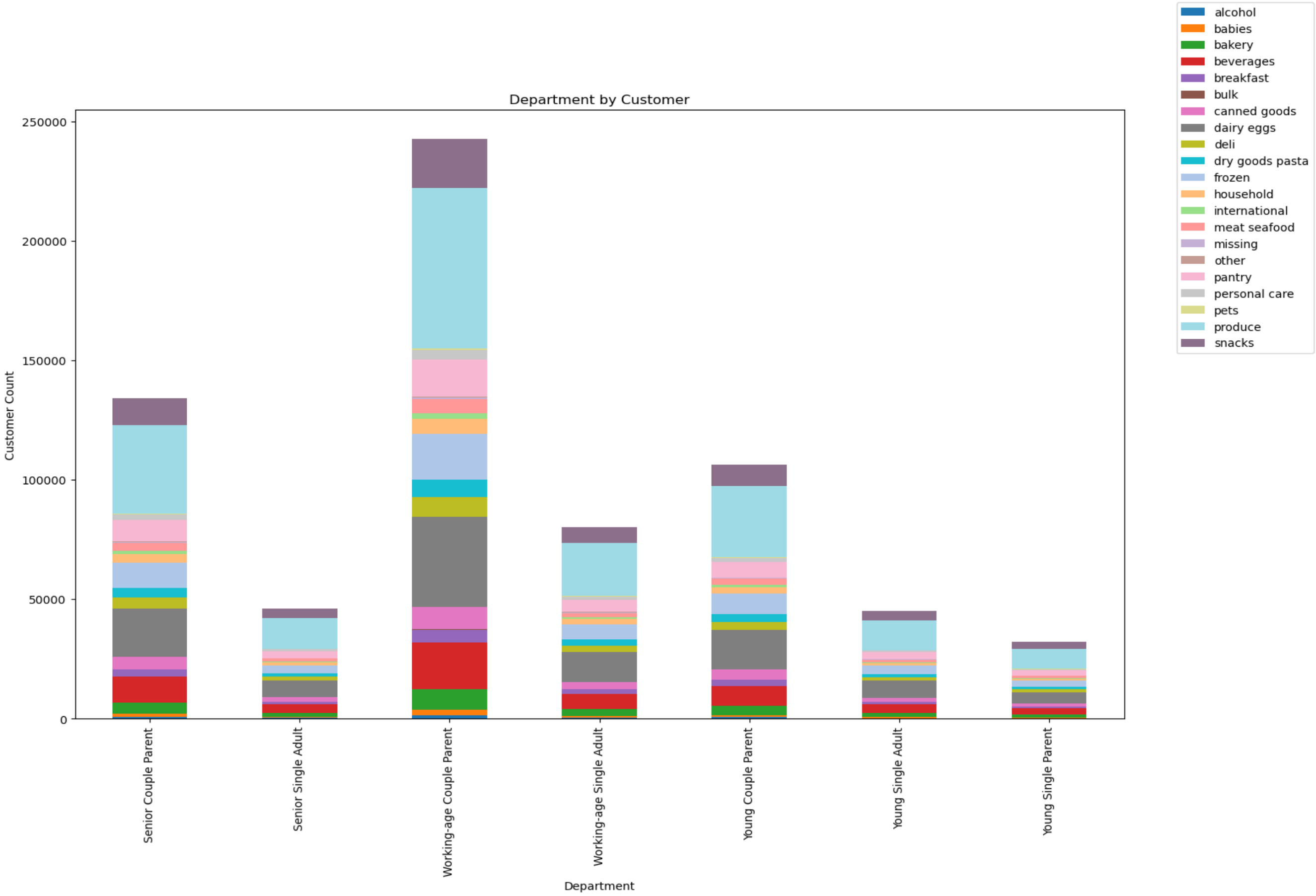


In general, most customers are low spenders, spending less than $10 on an order. This does not vary greatly by region.



The distribution of marketing profiles is proportional across regions. As all other measures, including ordering habits, have been proportional, it is likely that ordering habits behave accordingly (see charts below for further details).

Is there a connection between age and family status in terms of ordering habits?

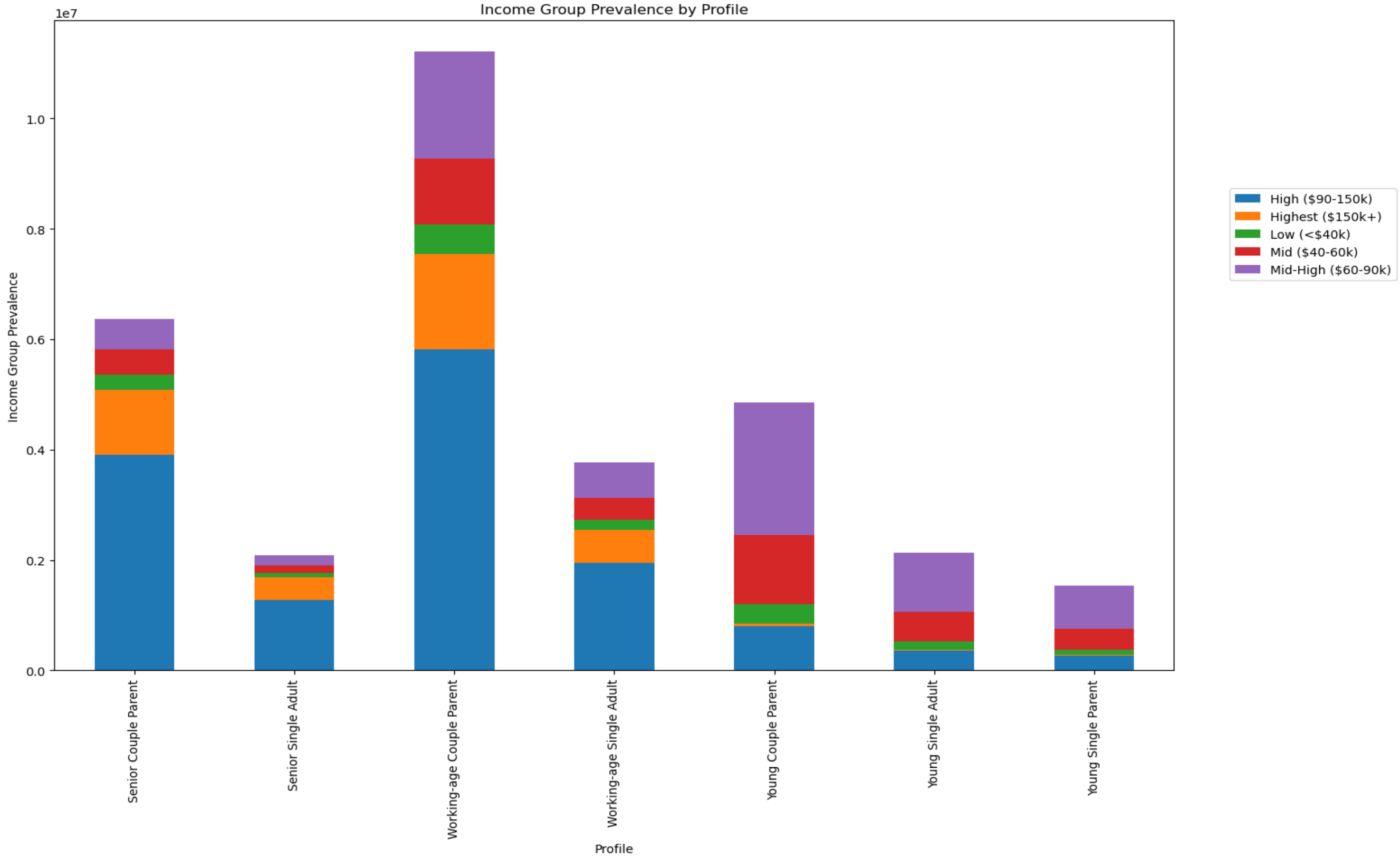


All in all, their purchase habits by department seem very proportional.

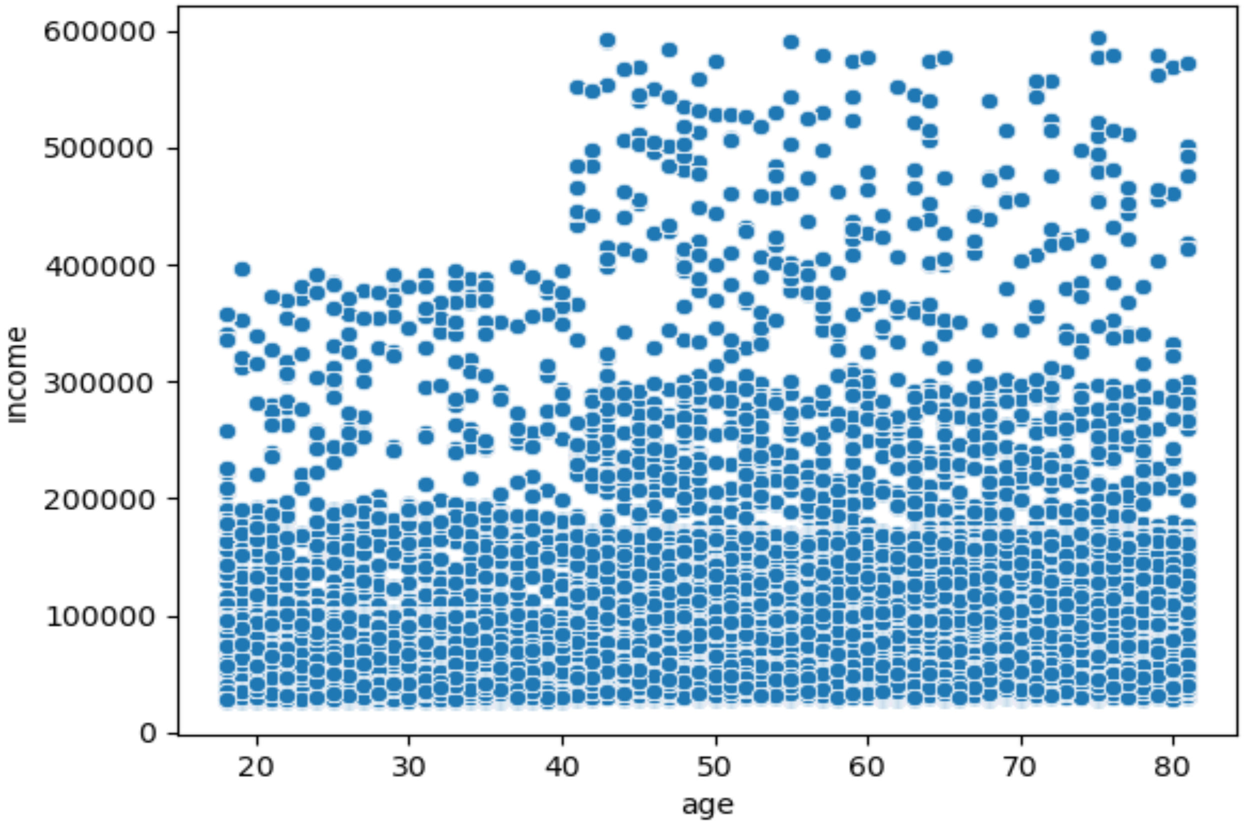
Most of those in this data are parents, and this could make sense given they purchase more items. They also mostly are of quite high income (as detailed below). There are also generally more young people (18-34) and seniors (65+) then the categories in between.

What different classifications does the demographic information suggest?

Age? Income? Certain types of goods? Family status?

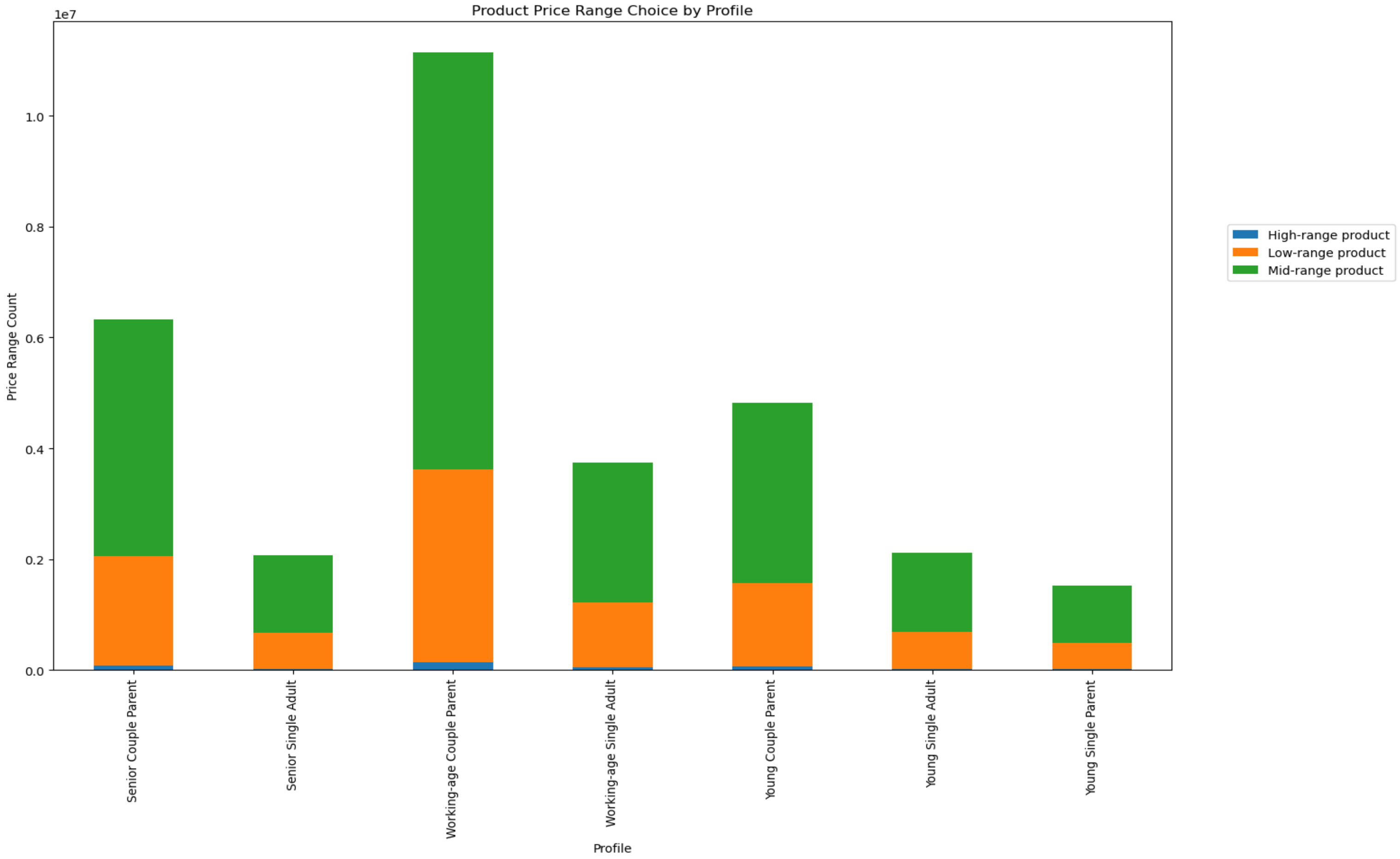


In general, as above, there isn't a lot of variation between different groups, no matter how much we divide them. The most notable aspect is that most customers are parents.

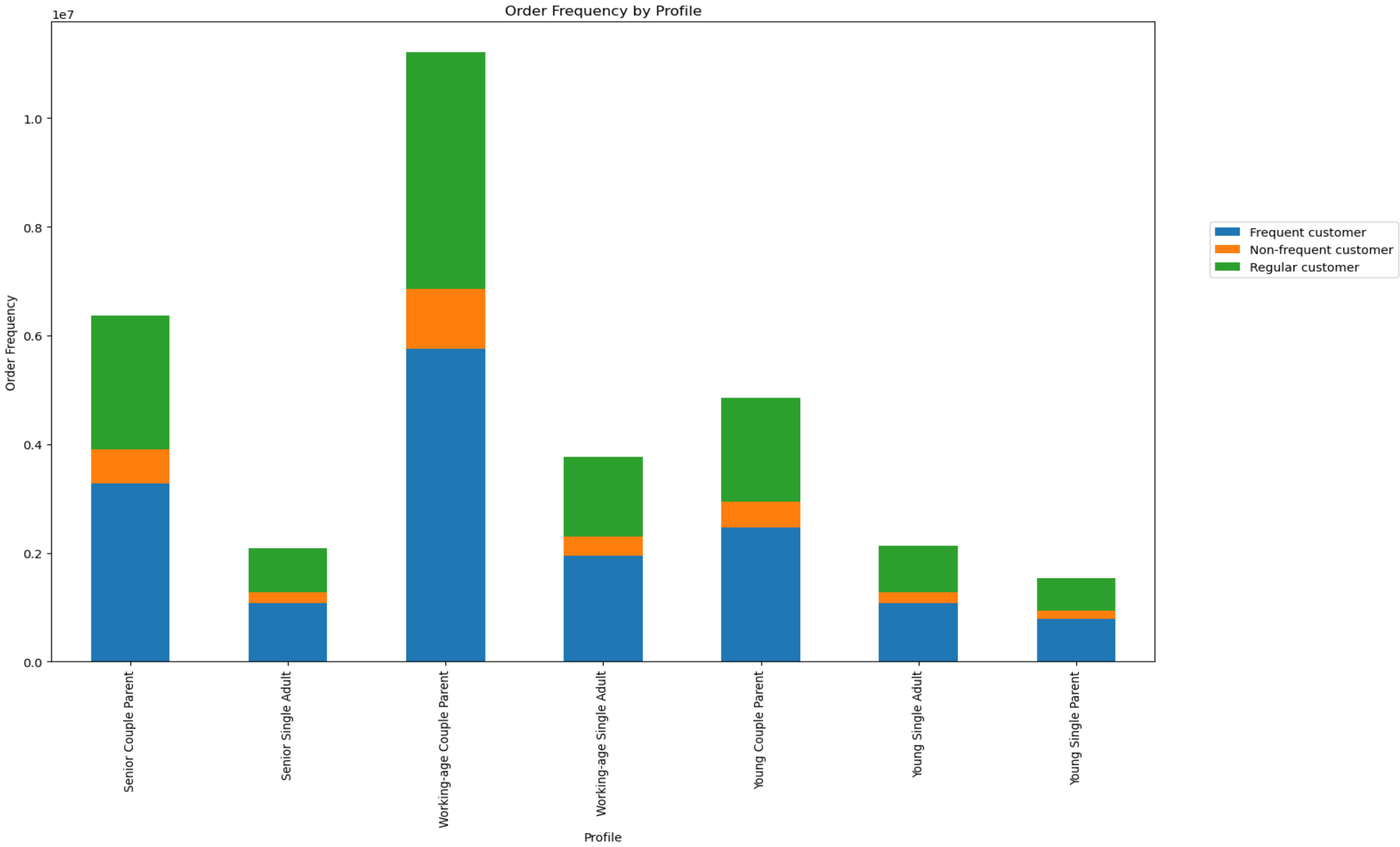


There is a marked increase in maximum and median income starting in the late 30's age, though it does not really get higher with age from there on. There is a consistent amount of people earning less than 20,000 across all ages. There are more people in their 20's and 30's earning between 30-40,000 than those earning 20-30,000 in the same age group.

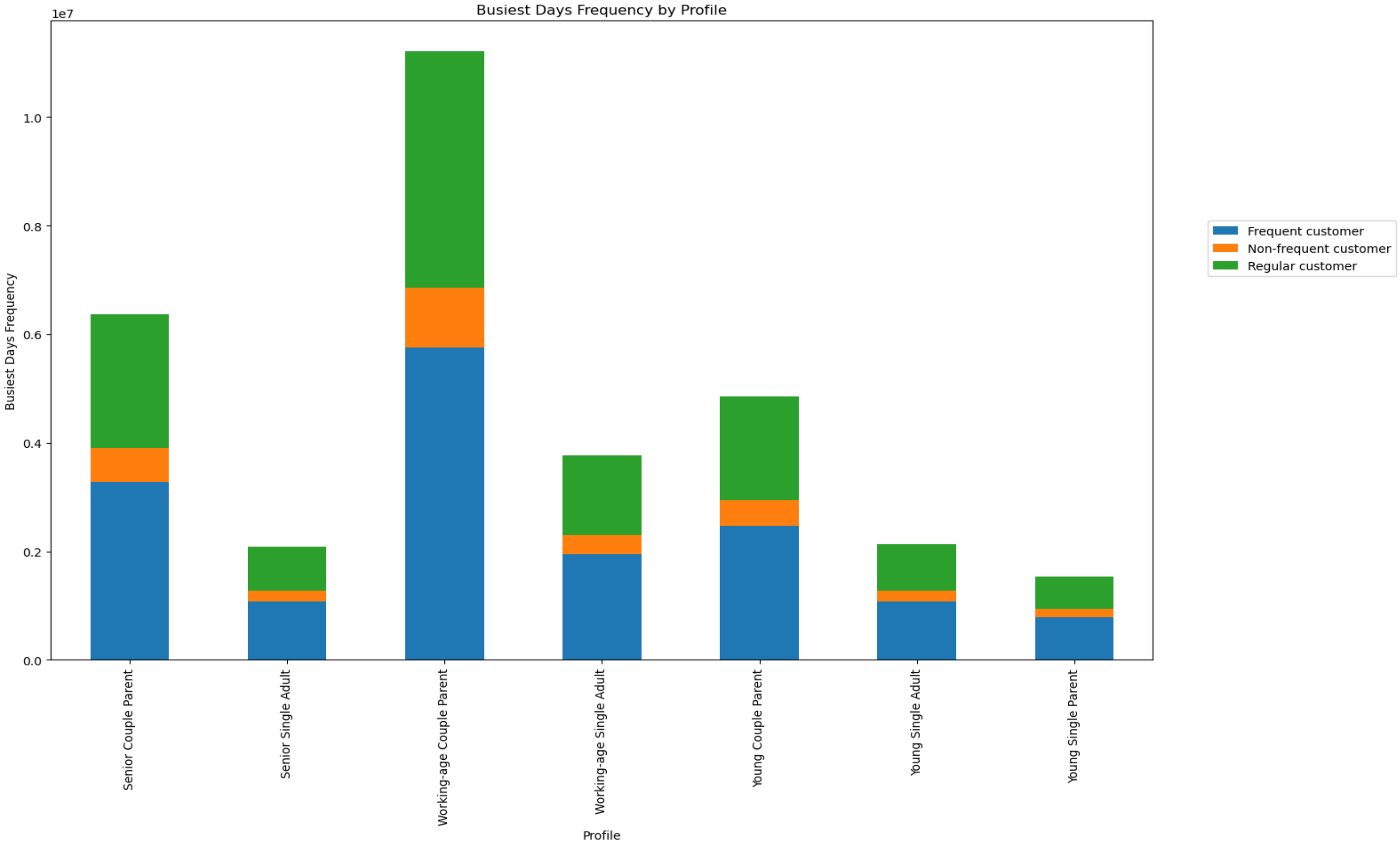
What differences can you find in ordering habits of different customer profiles? Consider the price of orders, the frequency of orders, the products customers are ordering, and anything else you can think of.



The different profiles buy products of the different price ranges at a proportional rate.

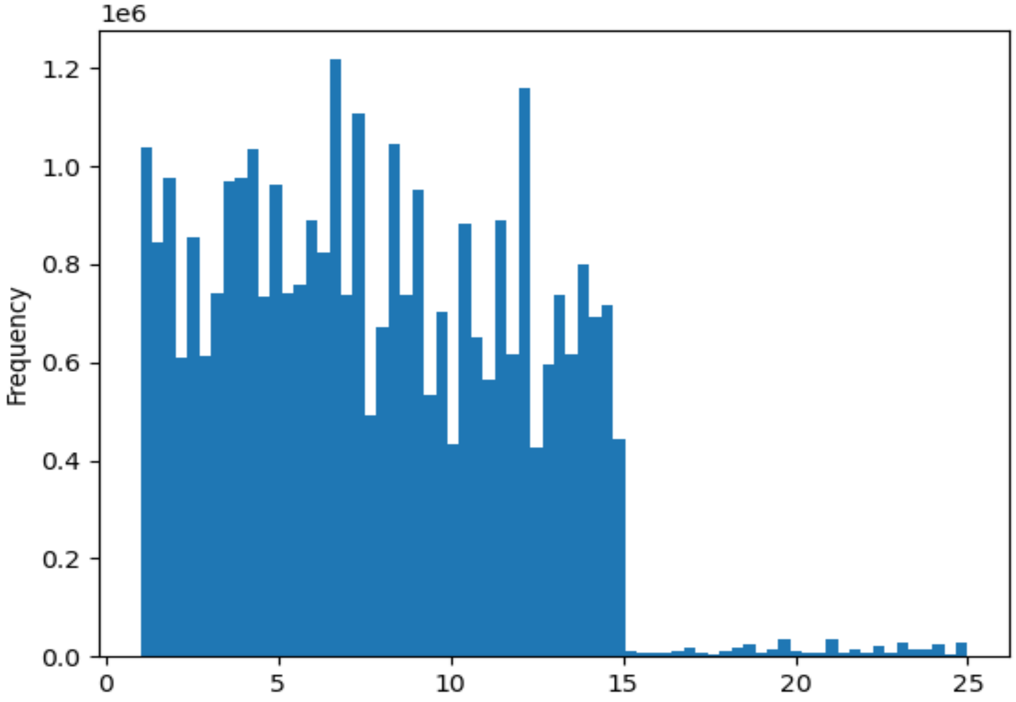


The same is true of order frequency.

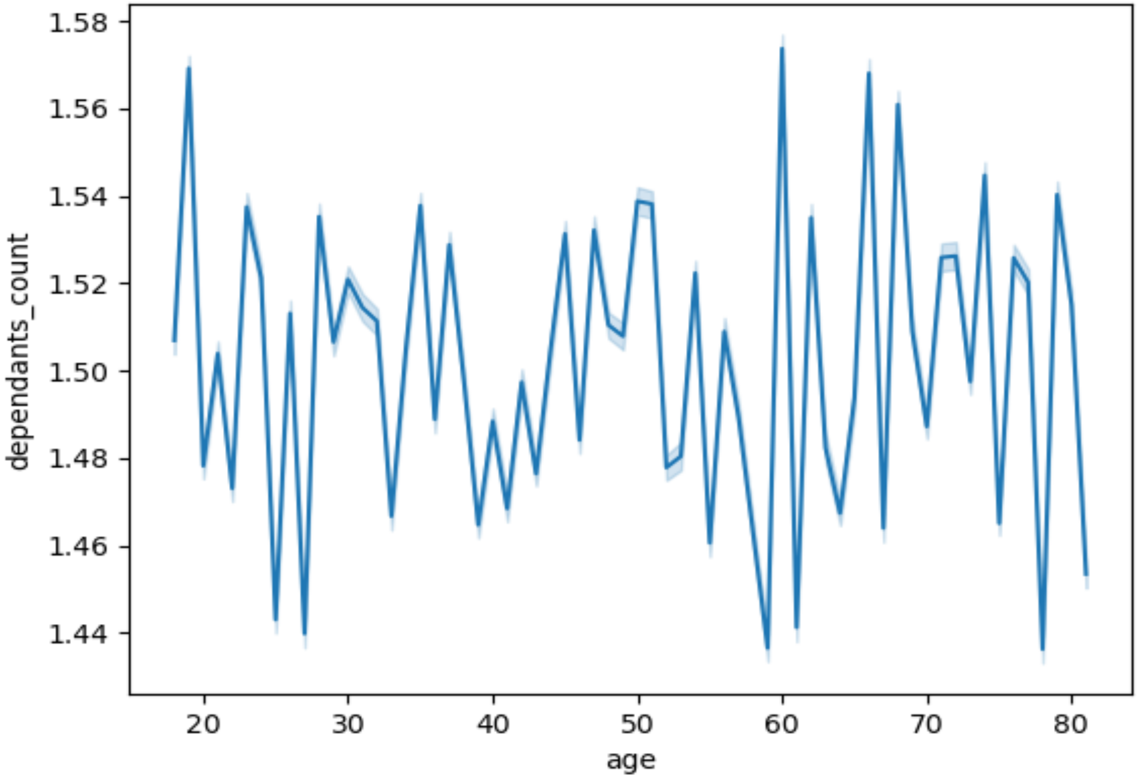


As well as their order behavior in terms of busiest vs least busiest days

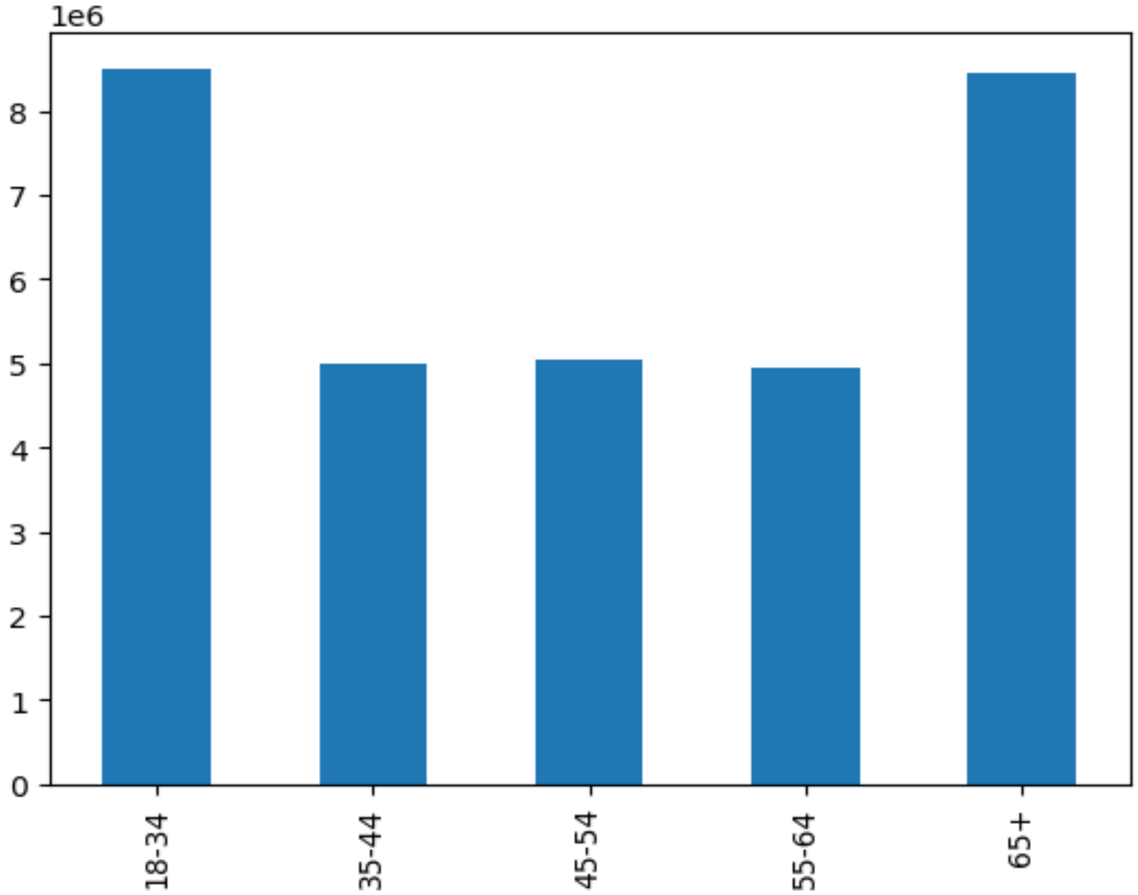
Other



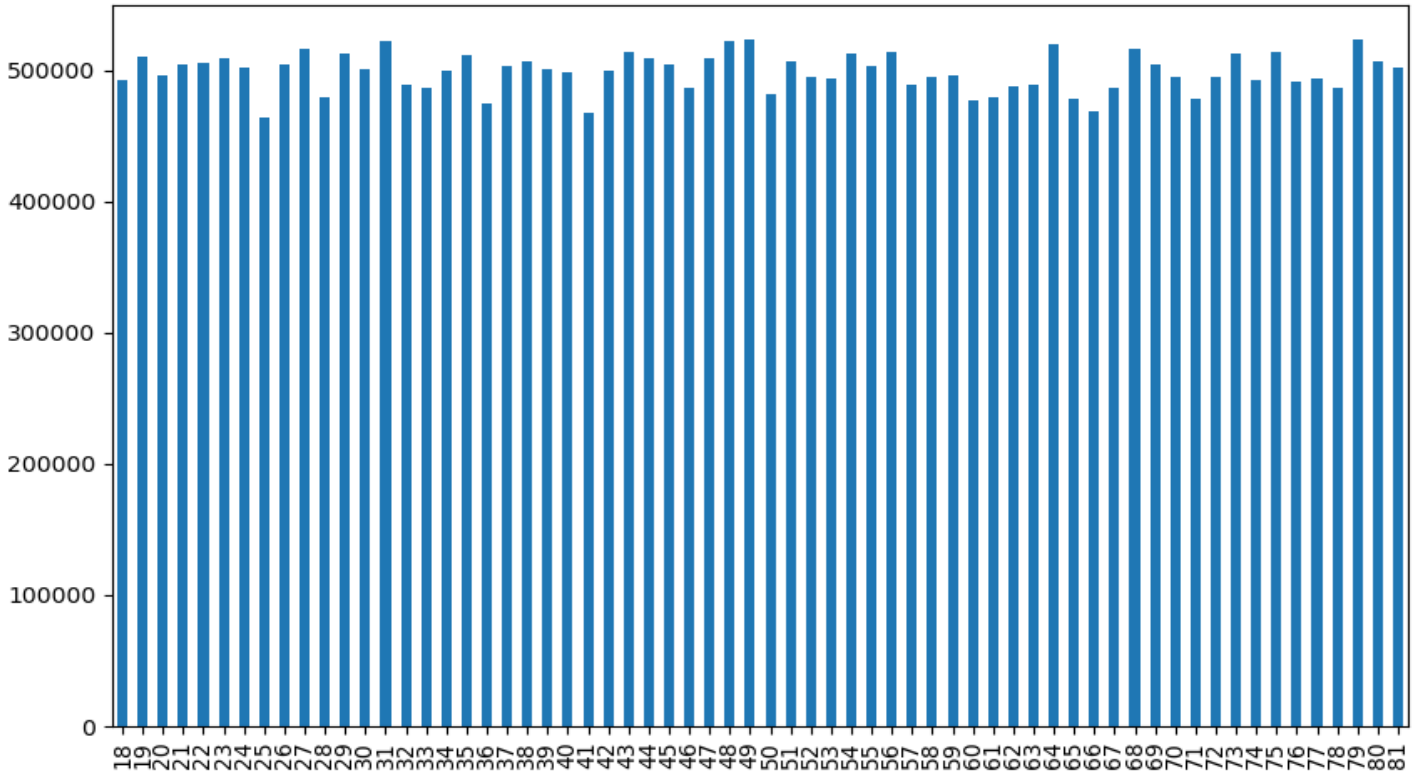
This demonstrates the price frequency of individual item purchases. Most are less than $15



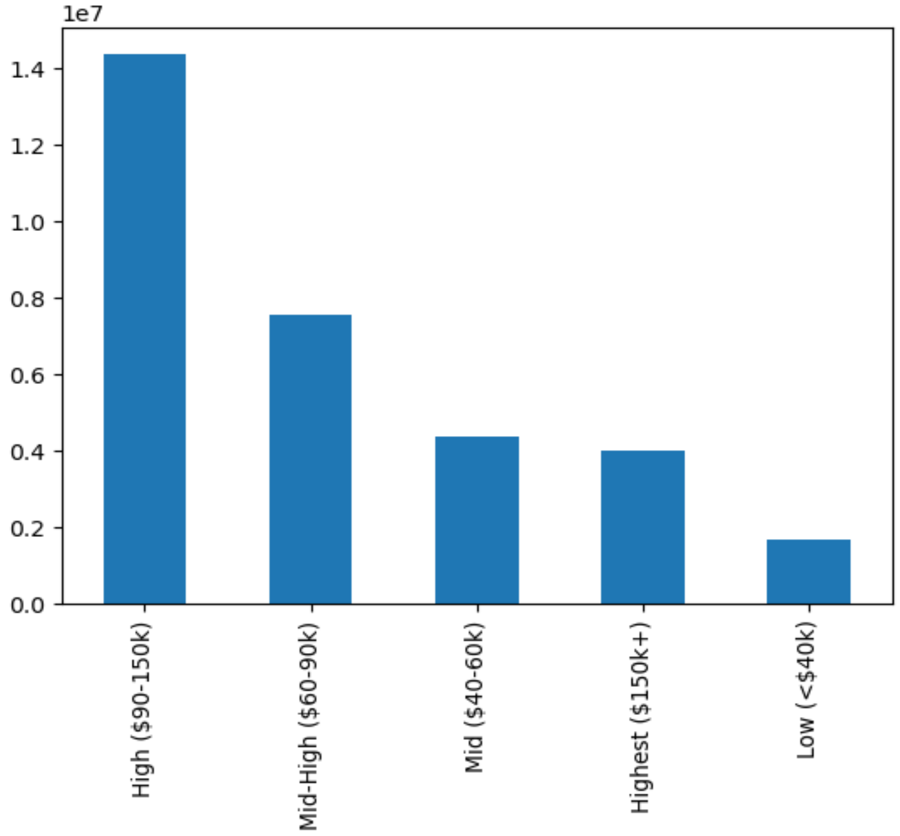
This line graph demonstrates age by number of dependants on average. Though it seems to It looks vary wildly, but as the range is between 1.44 and 1.56 dependants, it actually means that Instacart shoppers have about 1.5 children on average.



Count of customers by age group. The 18-34 and 65+ age groups are the largest, but also have a greater variety of different ages within them.



When broken down to individual ages however, the age distribution is pretty even



The largest groups are high, but not the highest, earners

instacart

Recommendations

The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.

Given that Saturday, Sunday and Friday are the busiest days for orders and the busiest times are from 9AM to 5PM, it's best to run ads Tuesday-Friday, ideally in the evenings.

They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.

The price difference is very small overall; it would be best to have a balanced strategy as to which products to advertise.

Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.

Most orders (68%) are for mid-priced items, from $5-15. About 31% are cheaper, and 1.3% are more expensive. It's likely best to target this mid-range category, but allow for lower-priced items as some simply are cheap enough to still make a reasonable profit.

Are there certain types of products that are more popular than others? The marketing and sales teams want to know which departments have the highest frequency of product orders.

Dairy eggs, and produce are by far the two largest categories while the meat/seafood is pretty tiny, which could suggest that these customers are more likely to be vegetarian. This would be a good market to target However the canned goods and frozen departments are also sizable, and would best be broken down into further categories in order to make the best judgment there. It is recommended to divide these departments to that we can ensure accuracy in future data.

Additionally, most customers have dependents, but do not purchase for babies. It would be worthwhile to research the competitiveness of our baby products' prices or otherwise see if competitors edge us out in selling baby products. However, this also suggests that we can effectively market items for children who are not babies and would likely have a receptive audience.

The marketing and sales teams are particularly interested in the different types of customers in their system and how their ordering behaviors differ. For example:

What’s the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?

Most customers shop Instacart on a weekly basis. Offering discounts for larger total purchases would likely encourage/maintain loyalty, and encourage customers to do all of their shopping with Instacart and taking advantage of their current habits.

Are there differences in ordering habits based on a customer’s loyalty status?

They generally do not. Consider introducing a loyalty program.

Are there differences in ordering habits based on a customer’s region?

There are not. Notably the south, west, and midwest are where most customers live, but the northeast has considerable population and wealth. It would be worthwhile to research why we are behind there.

Is there a connection between age and family status in terms of ordering habits?

All of the different groups spend proportionally. Most customers have dependents, so marketing to parents/children is advisable.

What different classifications does the demographic information suggest?
Age? Income? Certain types of goods? Family status?

None particularly as so much data reflects proportionality between groups. However, as above, most customers are parents.

What differences can you find in ordering habits of different customer profiles? Consider the price of orders, the frequency of orders, the products customers are ordering, and anything else you can think of.

All of these are proportional and do not reveal particular differences.