instacart [Project Name]

Project Name: Instacart Grocery Basket Analysis

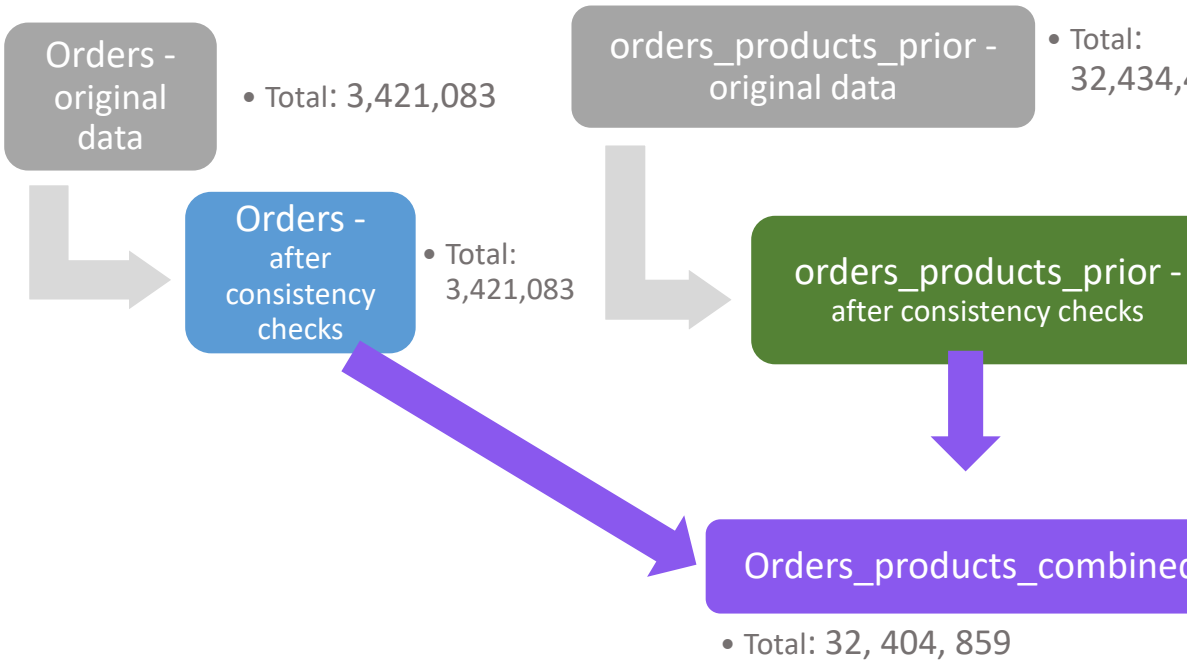
Date: January 17, 2024

Analyst Name: Sarah Boller

Contents:

instacart

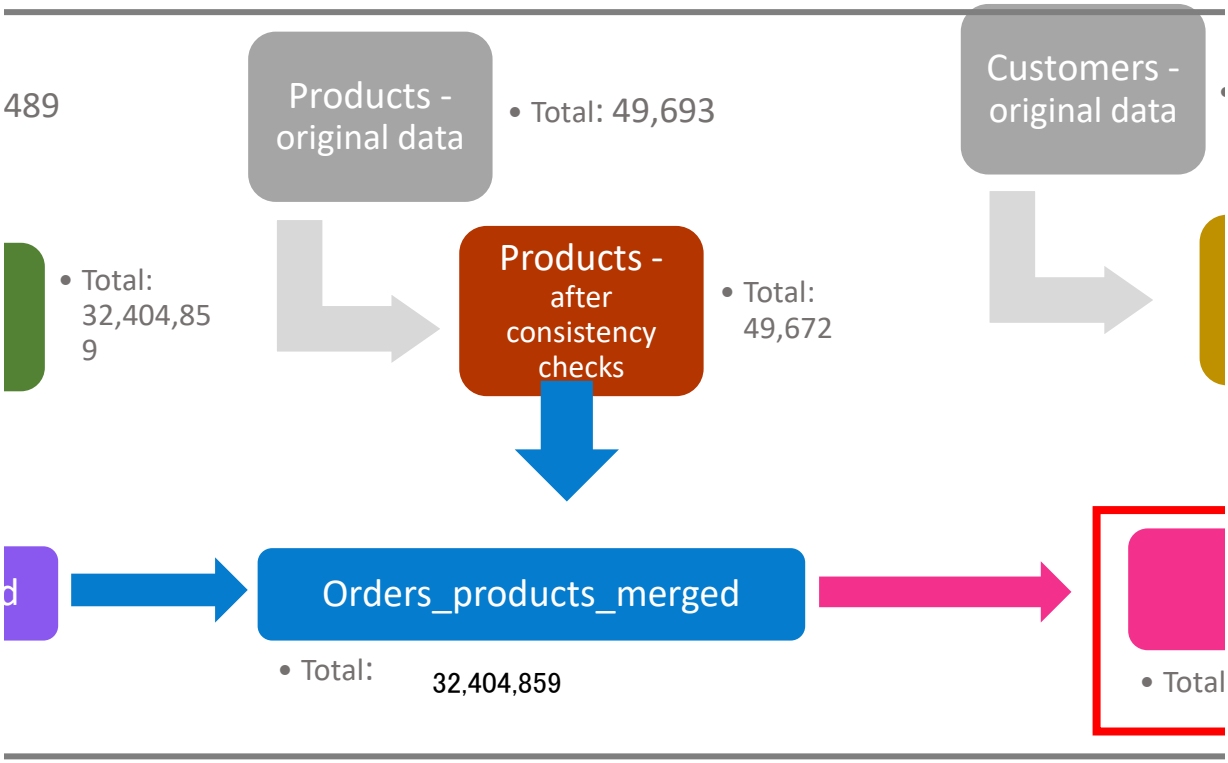
Population flow



1.) The grey boxes in the first row of the population flow represent the origi
imported the data set into Jupyter.

2.) The second row of boxes (coloured) represents the data sets **after** you m
conducting these operations. This offers a visual oveview of how the data *flo*

3.) The third row, where also the arrows are coloured, represents the merge
you end up with the final dataset (in the red box). Keep in mind the final dat



nal data sets as they were when you downloaded them. In the Total fields you ne

anipulated them, e.g., removed missing values and duplicates. In the Total fields
ows throughout the data consistency checks.

s you performed between the datasets. In the Total fields you need to add the c
aset should be without exclusions (based on the exclusion flag).

Title page

• Total: 206, 209

Customers - after consistency checks

• Total: 206,209

Orders_products_all

l: 32,404,859

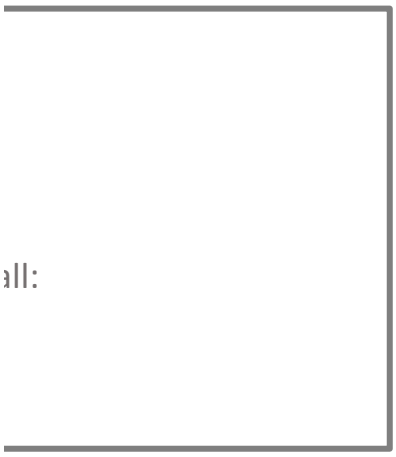
Exclusion flag

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

eed to add the count of the rows when you

you need to add the count of the rows after

ount of the rows in the merged datasets, so that


all:

instacart

Consistency checks

Dataset	Missing values	Missing values treatment
orders	day_since_prior_order: 206,20	This indicates a first order. I created a flag based on this information.
products	product_name: 16 missing valu	None
orders_products_prior	none	
customers	PII only.	Removed with the rest of PII data

Duplicates
none
none
none
none

instacart

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed
eval_set		
	order_dow -> order_day_of_week	order_day_of_week
		user_id
	n_dependants -> n_dependents	
	Gender -> gender	
	STATE -> state	
	Age -> age	
First Name		
Surnam		
	order_hour_of_day -> time_of_order_hr	
add_to_cart_order		
aisle_id		

Comment/Reason
unnecessary to data
int -> string
miss spelled
Consistancy
Consistancy
Consistancy
PII
PII
clearity
irrelevant
irrelevant

instacart

Column derivations and aggregations

Dataset	New column	Column/s it was derived from
final with profiles	returning customer	days_since_prior_order
final with profiles	loyalty_flag	max_order
	mean_expenses	prices, user_id
final with profiles	Spender_flag	mean_expenses
	median_freq	days_since_prior_order
final with profiles	freq_flag	days_since_prior_order
final with profiles	income_flag	income
final with profiles	parent_flag	*n_dependents*
final with profiles	Age_Group	*age*
final with profiles	customer_profiles	age_group, income_flag, parent_fl

Conditions
if nan returning customer is False. If there is a number returning customer is True
if max_orders is more than 40, then set Loyal customer else if max_orders is more than 10 and less than equal to 40, then set Regular customer else max_orders is less than equal to 10, set New customer
mean prices per user_id
if mean_expensess is less than 10, then set Low spender else if expensess is more than equal to 10, then set High spender
median of days_since_prior_order per user_id
if median_days_prior_purchase is more than 20, then set Non-frequent customer else if median_days_prior_purchase is more than 10 and less than equal to 20, then set Regular customer else median_days_prior_purchase is less than equal to 10, set Frequent customer
If income <= $67,584, then low income If income > $67,584 & < $127,912, then regular income if income >= $127,912, then high income
If n_dependents >= 1 then parent Else, not a parent
If age <= 33, then Age_Group 18-33 If age > 33 & <= 49, then 34-49 If age > 49 & <= 65, then 50 to 65 Else, Over 65
Age Group_income flag_ Parent_flag

In this tab
them. NB:

Title page

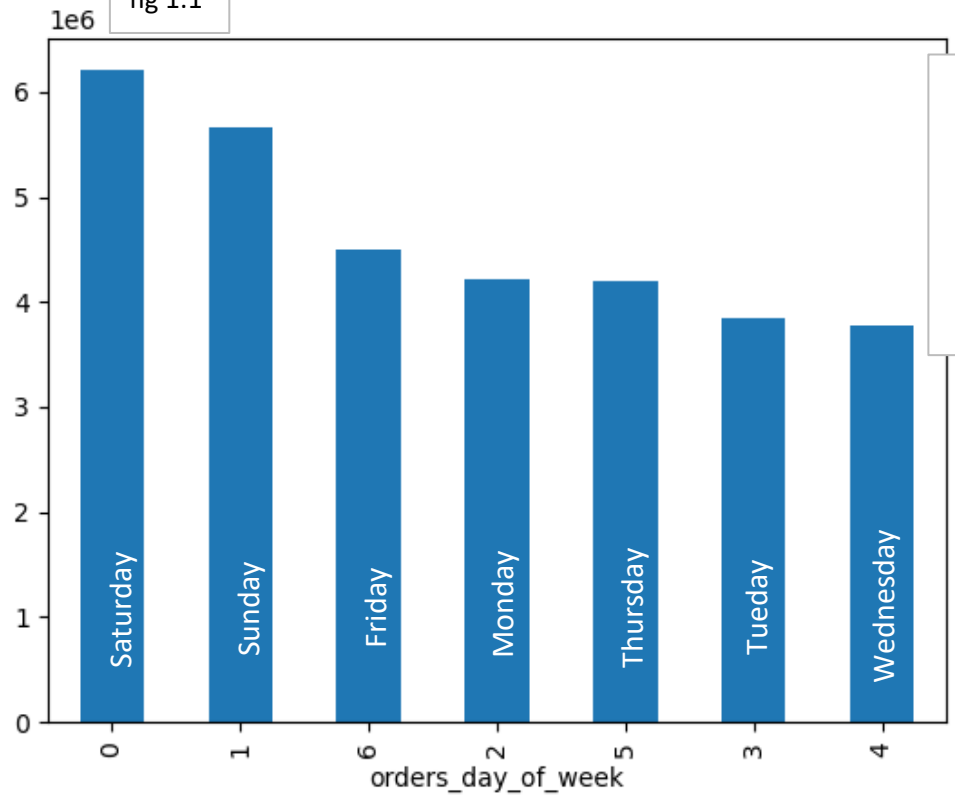
› you should add tables below that the frequences of flags/label variables that you produced after deriving
: don't do this for continuous variables, only for flags.

instacart

Visualisations

The sales team needs to know what the busiest days of the week and hours orders) in order to schedule ads at times when there are fewer orders.

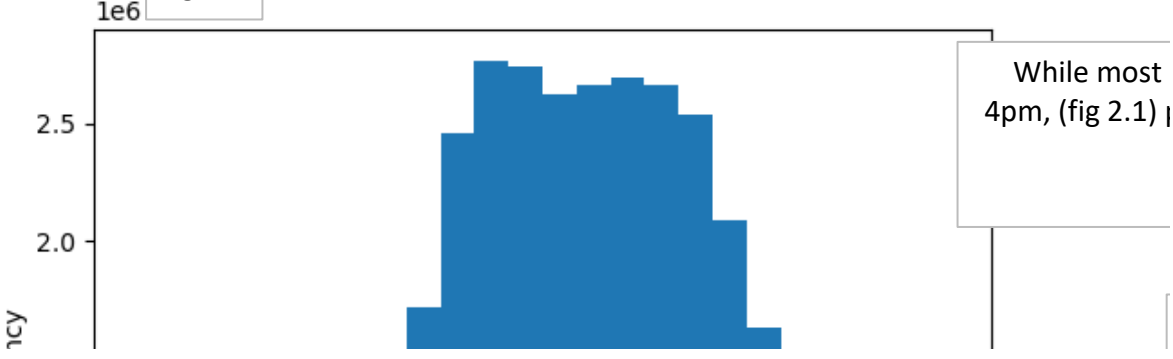
fig 1.1



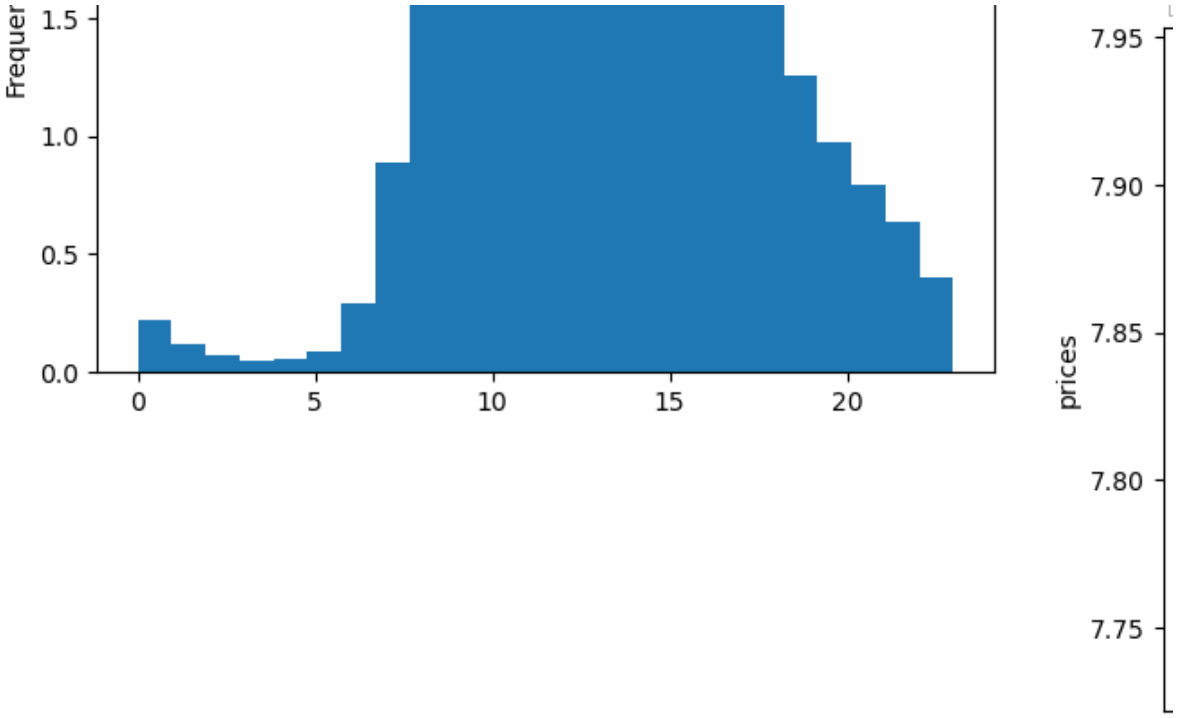
The busiest day
on the weeke
Peaking on Satu
and fanning out
there. With
Wednesday as
least busy day.

They also want to know whether there are particular times of the day when inform the type of products they advertise at these times.

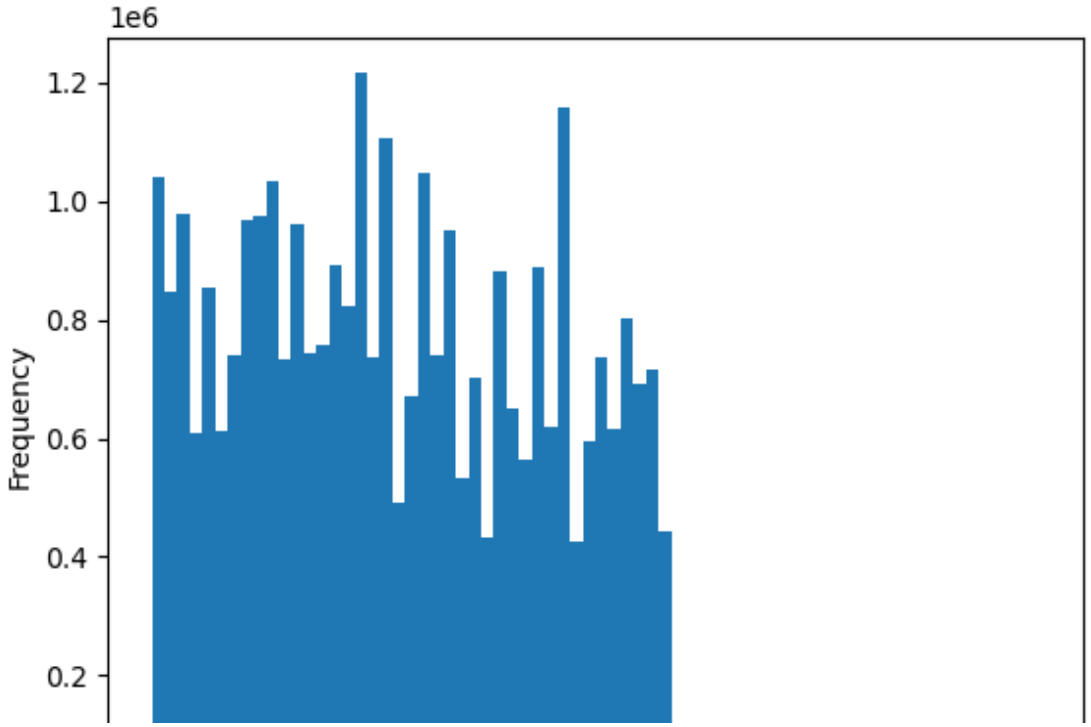
fig 2.1



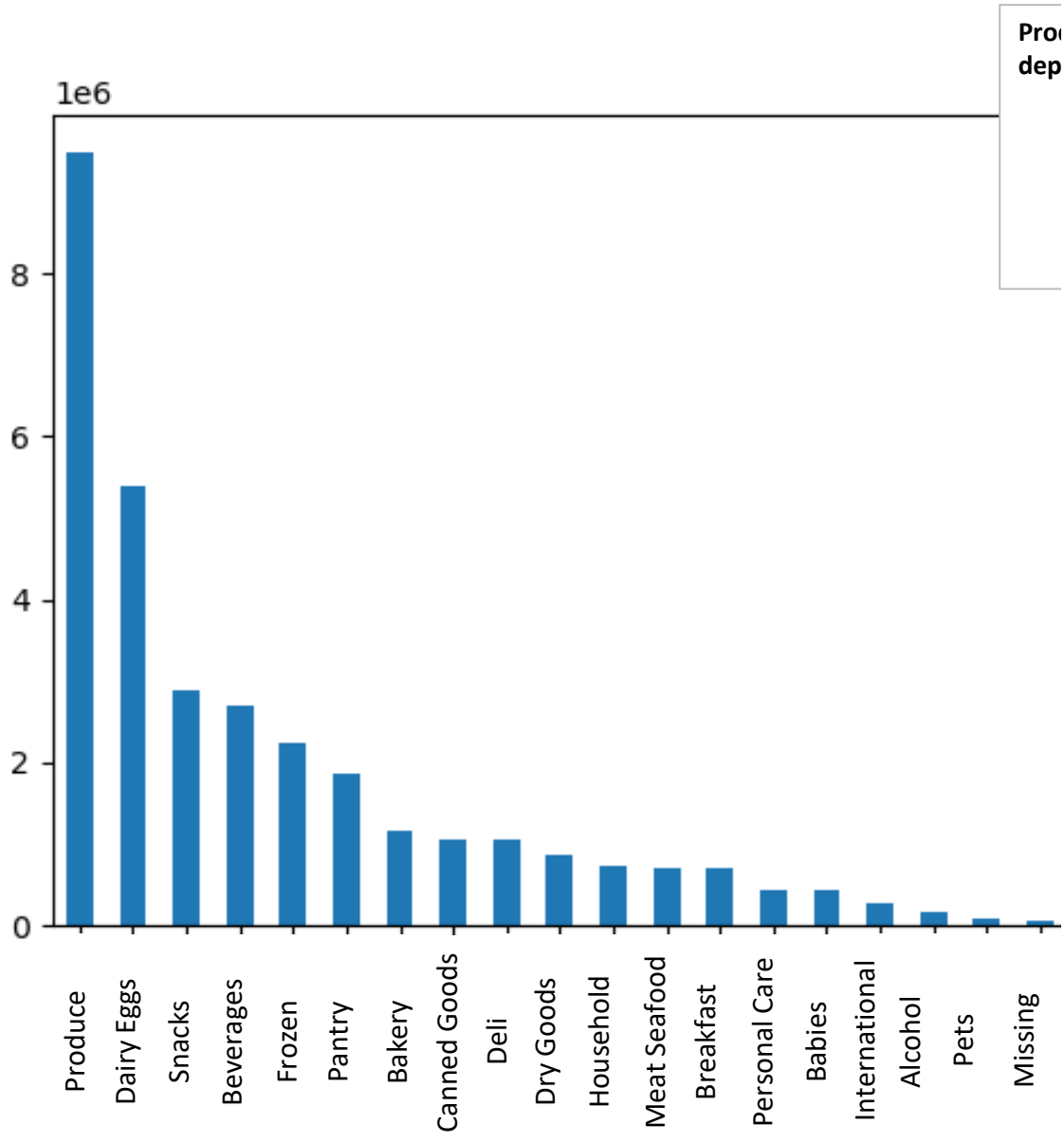
While most
4pm, (fig 2.1)



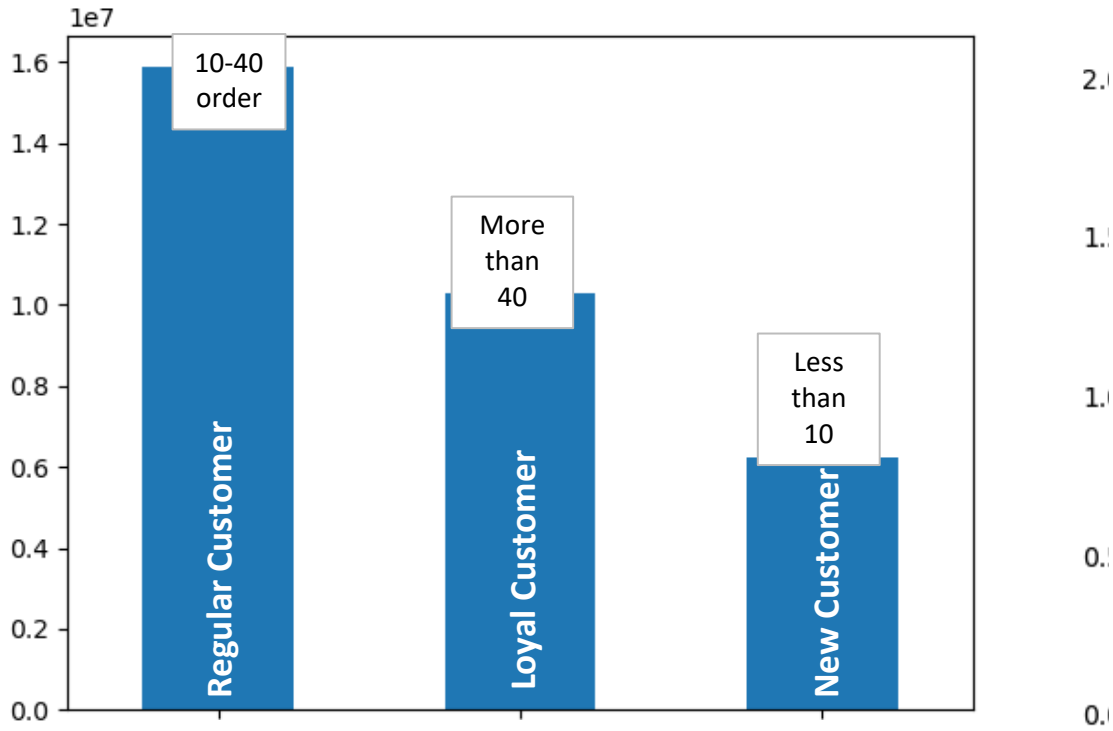
Instacart has a lot of products with different price tags. Marketing and sales w groupings to help direct their efforts



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

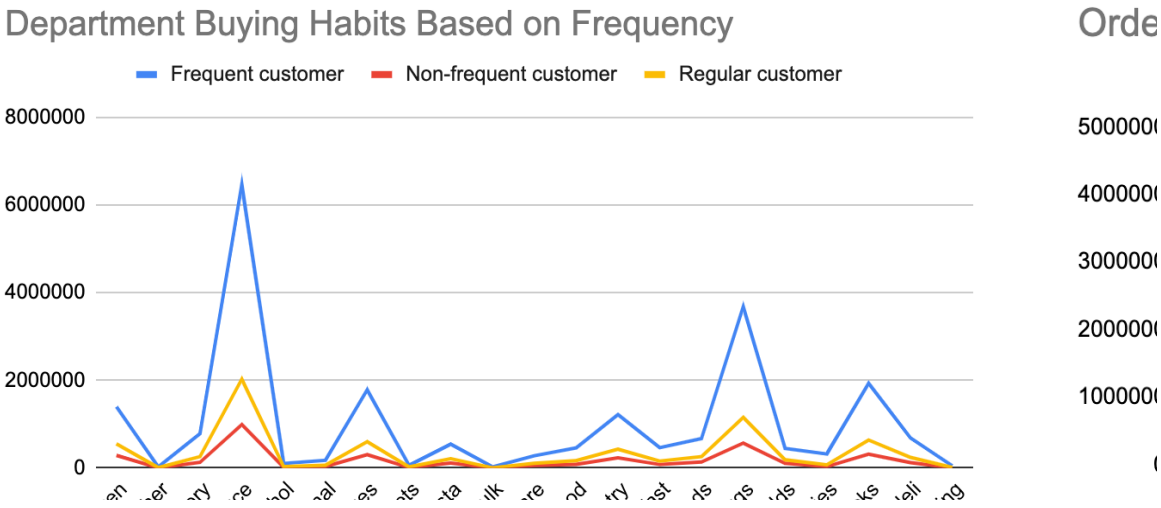


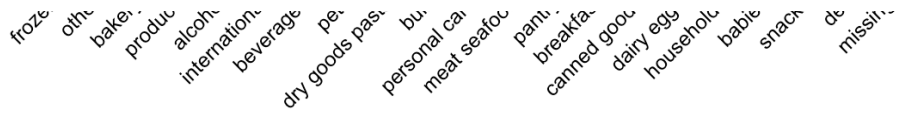
What's the distribution among users in regards to their brand loyalty (i.e., ho

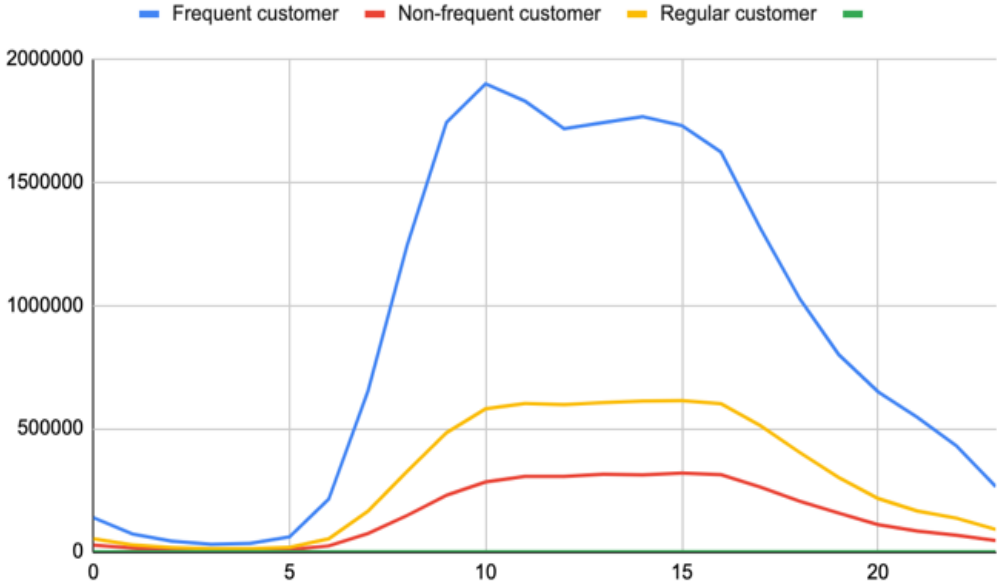


Are there differences in ordering habits based on customer's loyalty statu

There is no outstanding difference in spending habits based on loyalty or fr

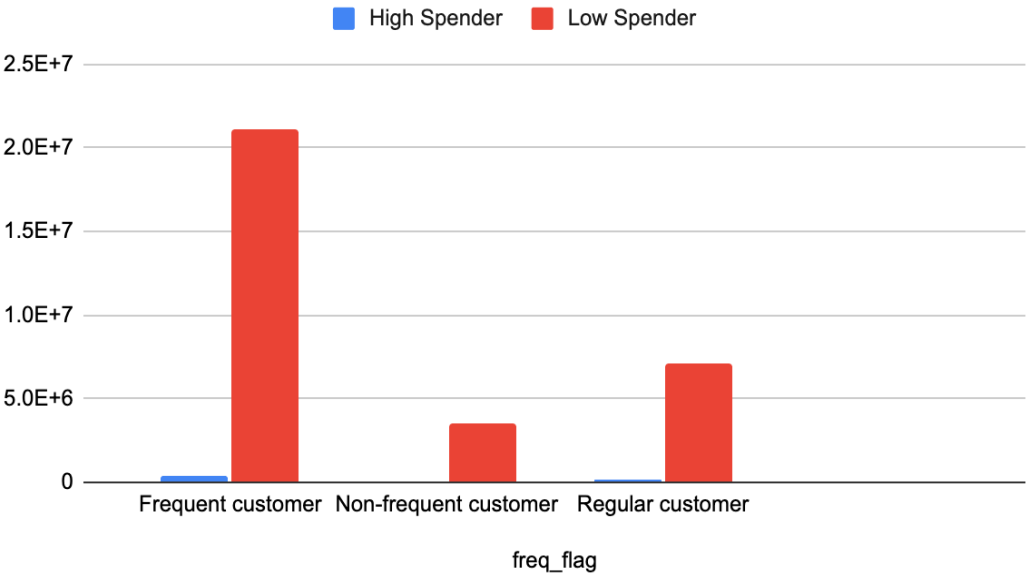






Loya
150000
100000
50000

High Spender and Low Spender

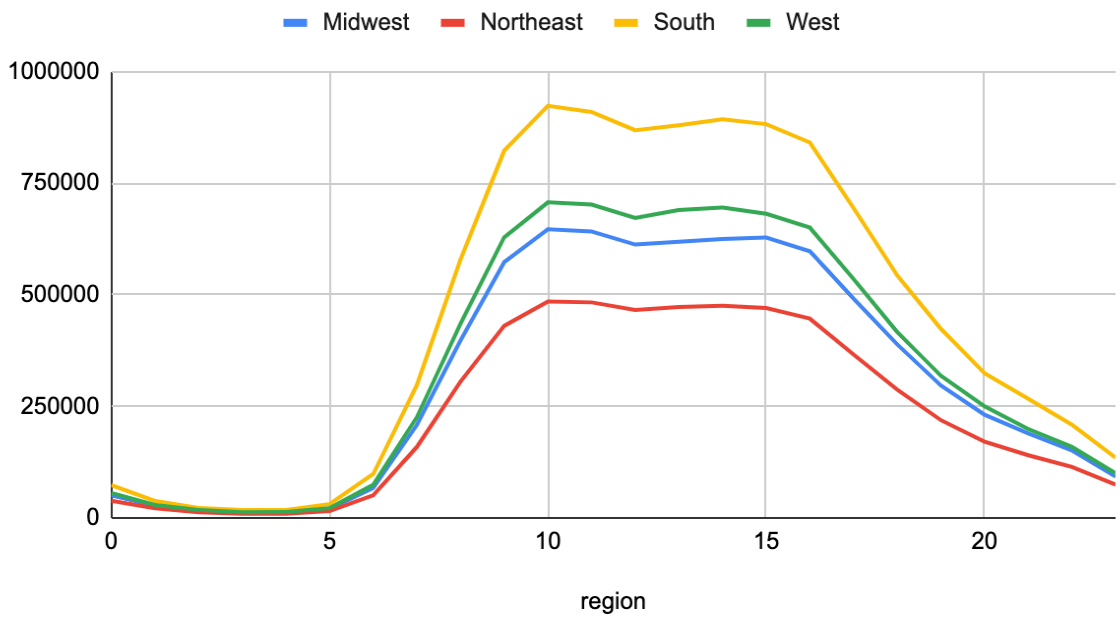


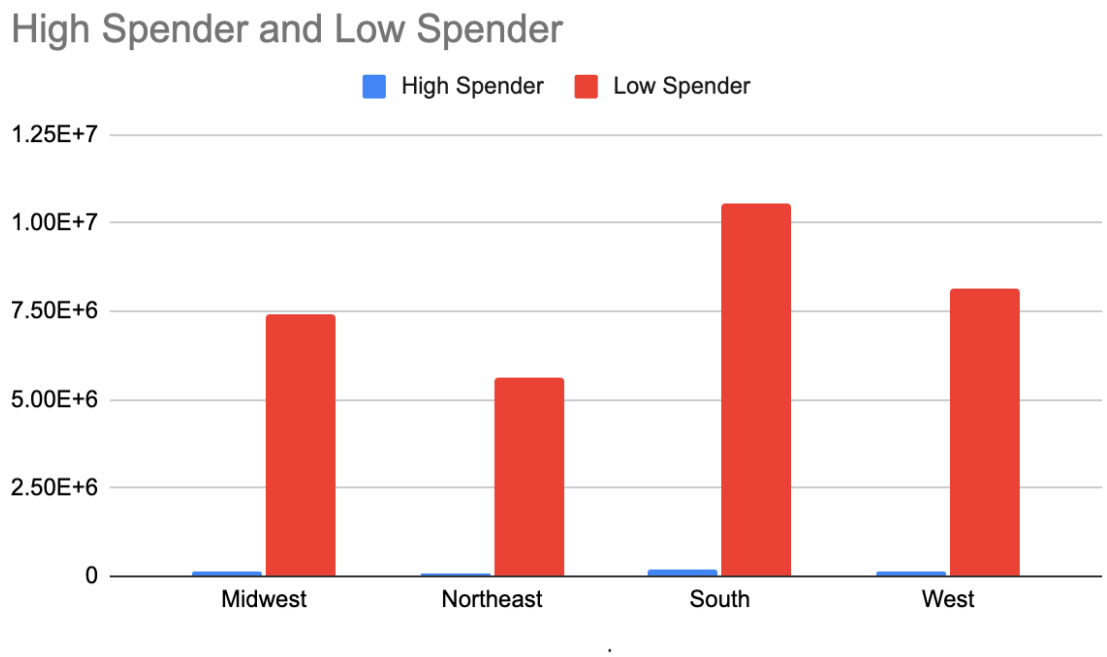
Hig
2.0E
1.5E
1.0E
5.0E

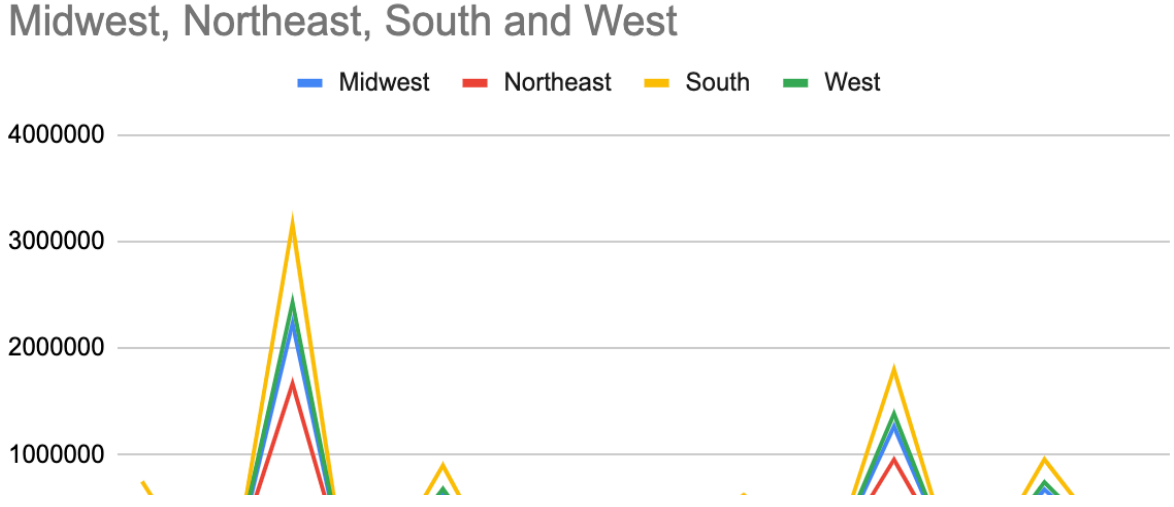
Are there differences in ordering habits based on a customer's region

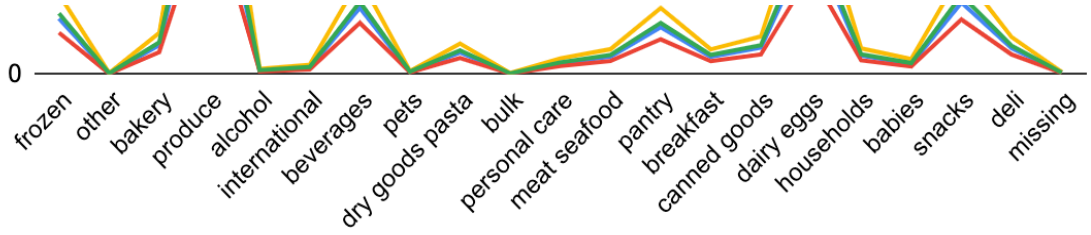
There is no outstanding difference in spending habits based on region. They f

Midwest, Northeast, South and West


Midwest
Northeast
South
West
1000000
750000
500000
250000
0
0
5
10
15
20
region

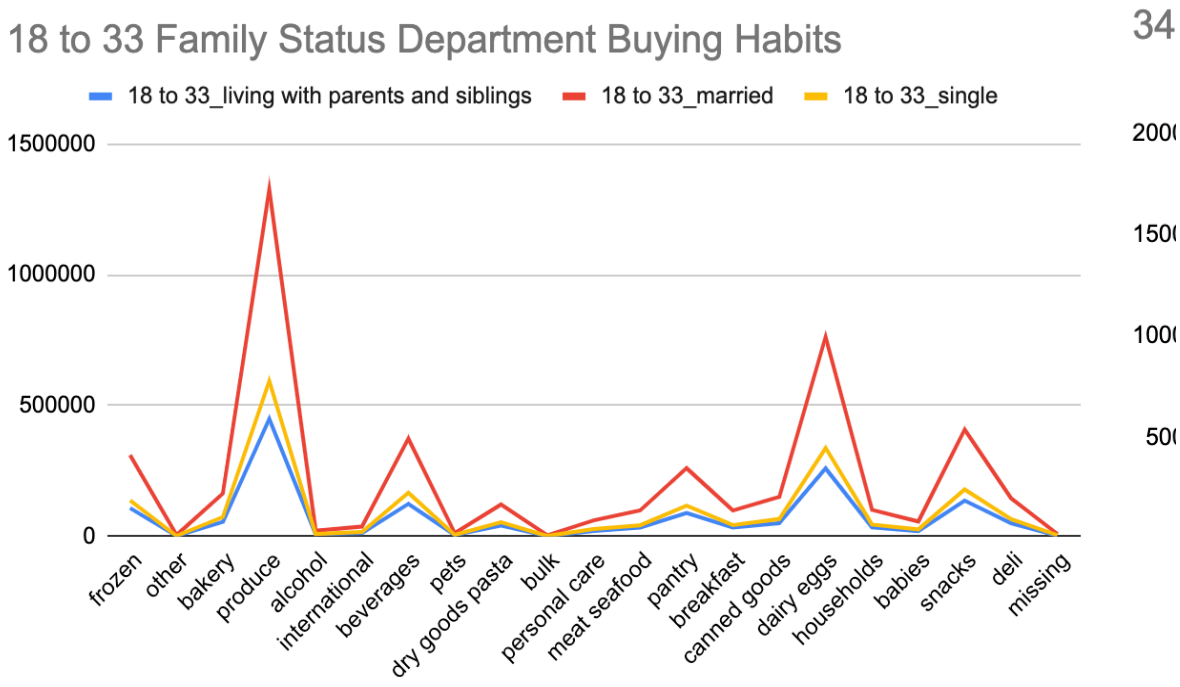

High Spender and Low Spender
High Spender
Low Spender
1.25E+7
1.00E+7
7.50E+6
5.00E+6
2.50E+6
0
Midwest
Northeast
South
West

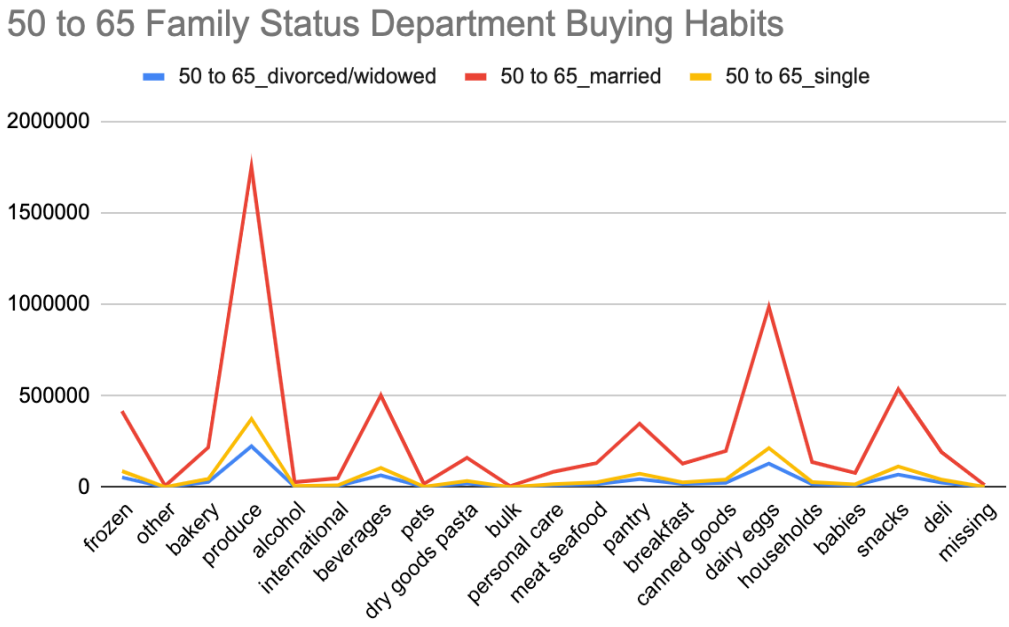

Midwest, Northeast, South and West
Midwest
Northeast
South
West
4000000
3000000
2000000
1000000

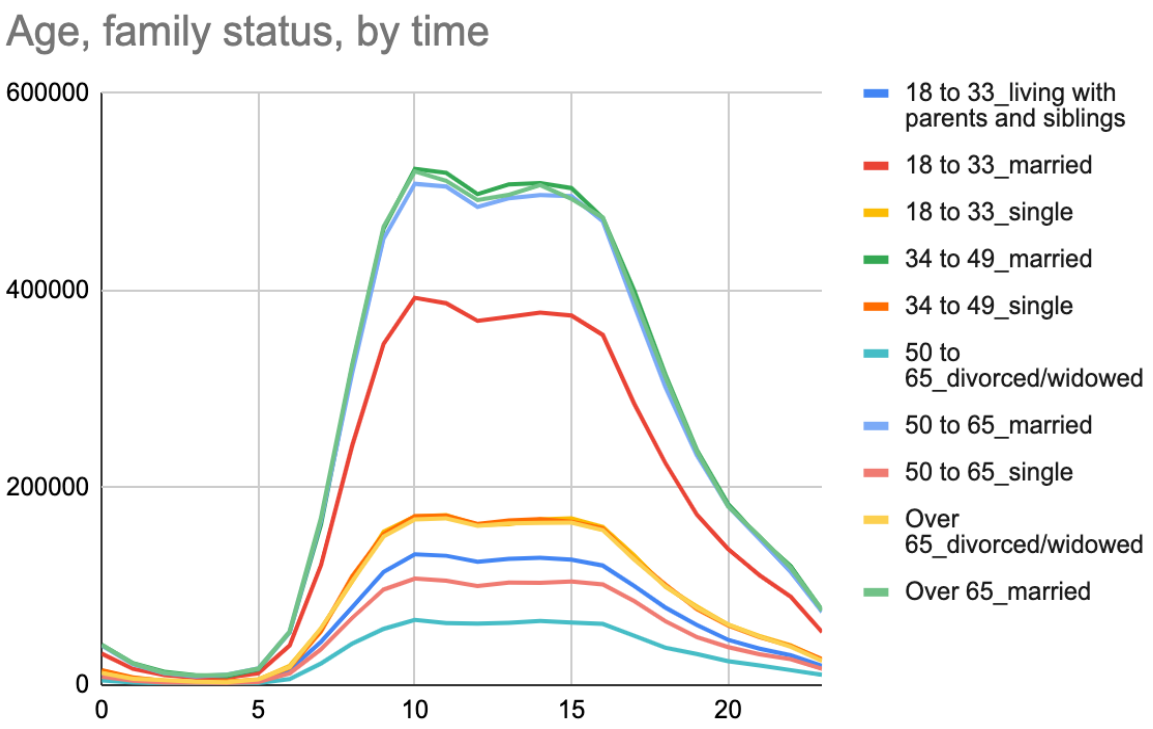


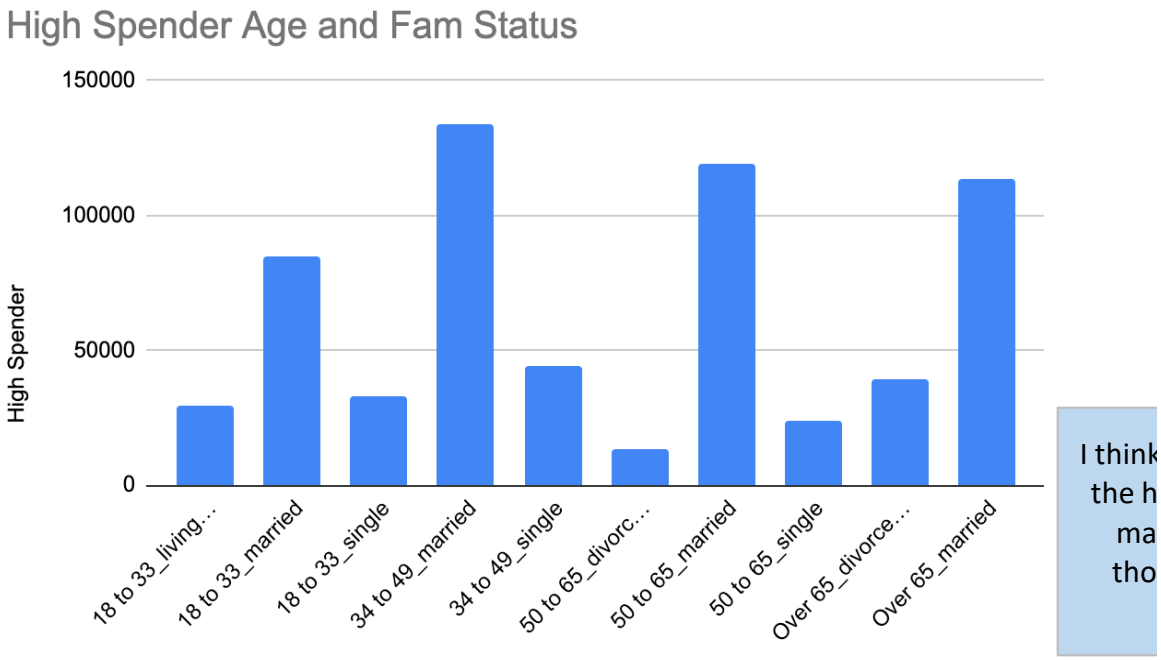
Is there a connection between age and family status in terms of ordering hab

There is no outstanding difference in spending habits based on age and family statu





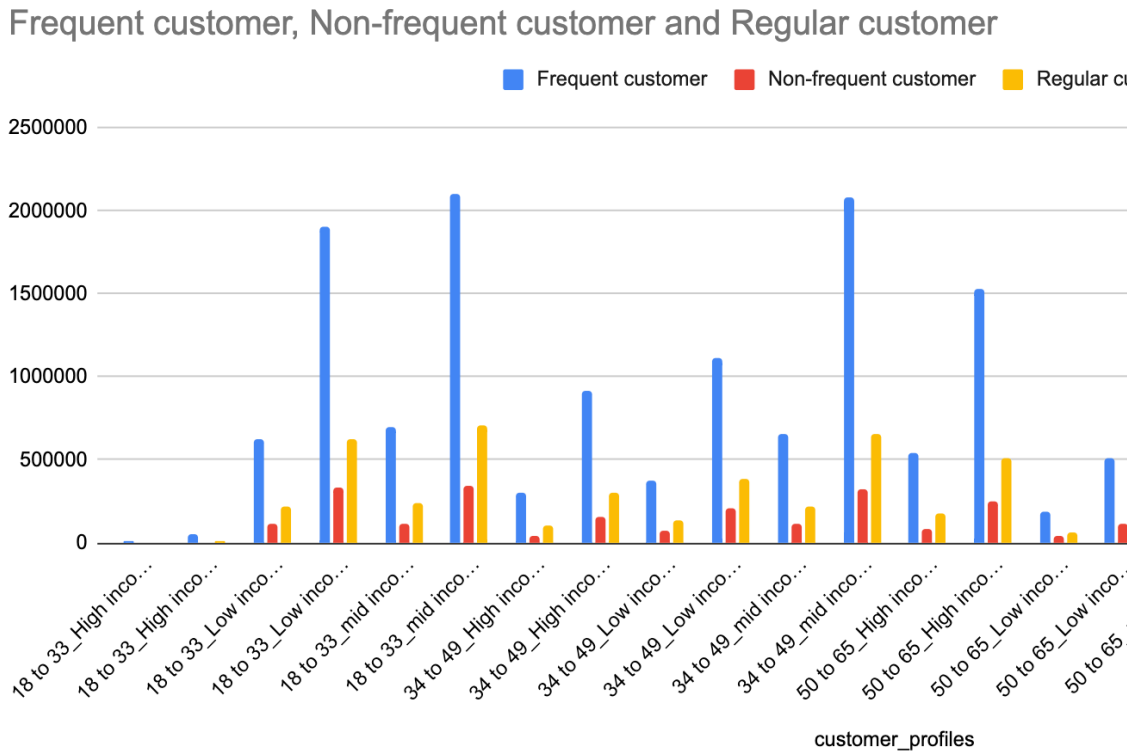


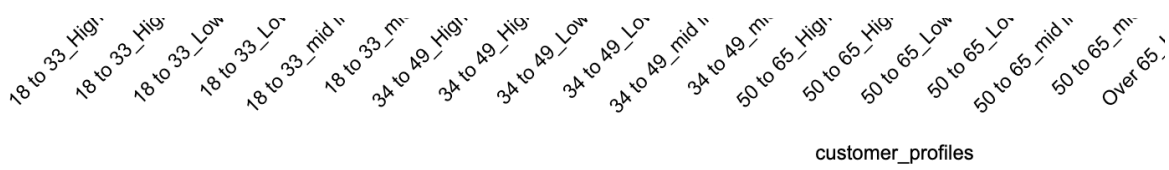


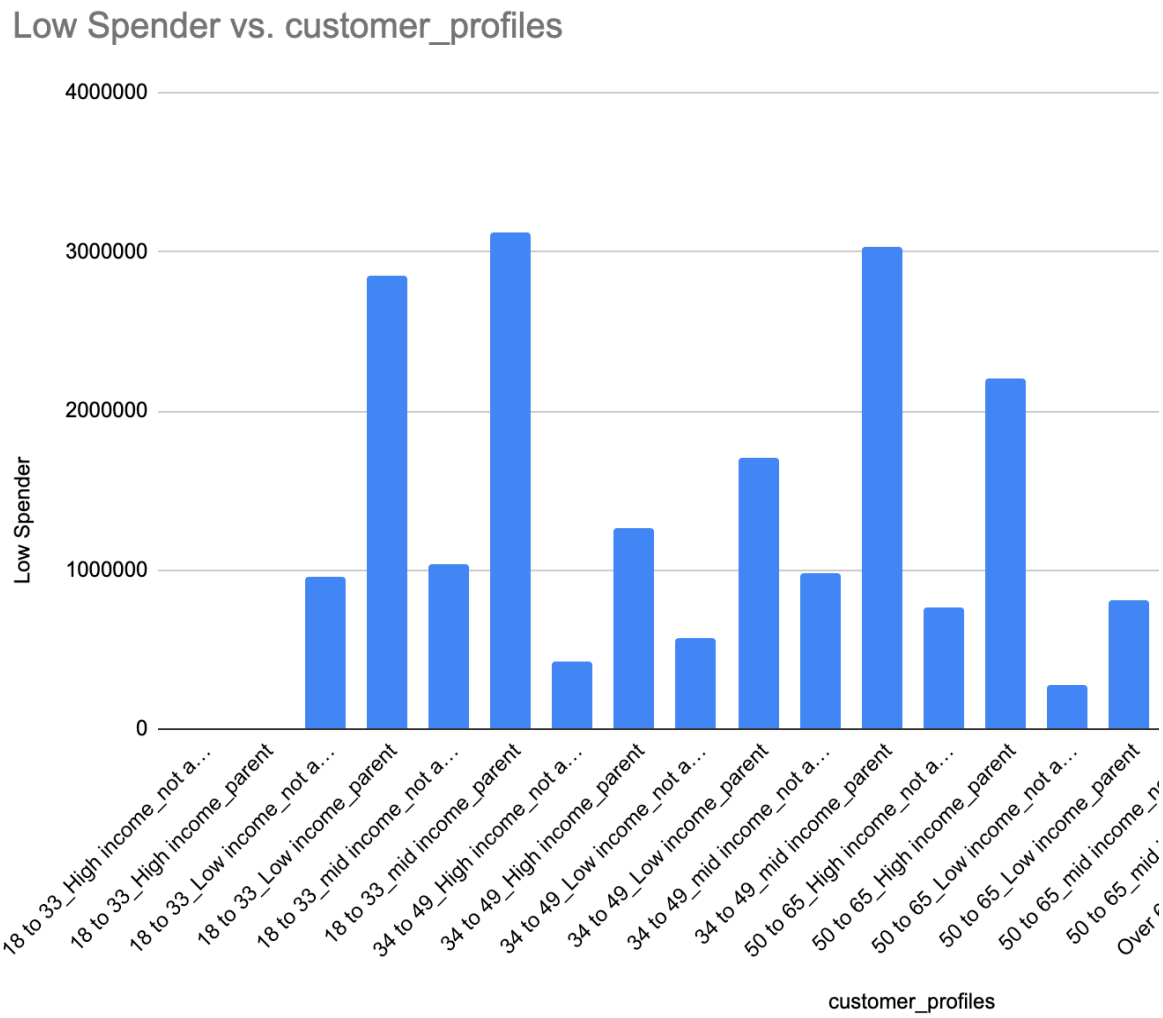
I think
the h
ma
tho

What different classifications does the demographic information suggest
goods? Family status?

goods? Family status?

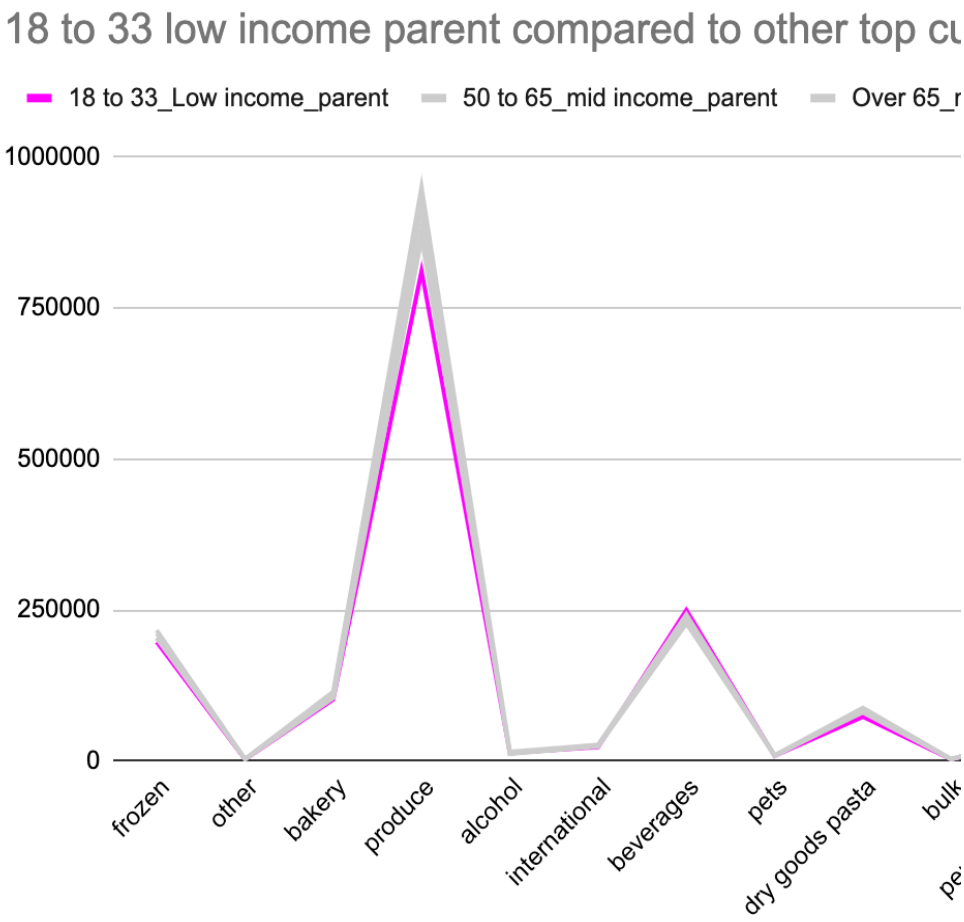






What differences can you find in ordering habtis of different custom
the frequency of orders, the products customers are ordering, and a

One profile stand out is low income parents in the 18-33 age group. They tend to purchase more snacks, beverages, and breakfast items, over other customer profiles


18 to 33 low income parent compared to other top cu
18 to 33_Low income_parent
50 to 65_mid income_parent
Over 65_r
1000000
750000
500000
250000
0
frozen
other
bakery
produce
alcohol
international
beverages
pets
dry goods pasta
bulk

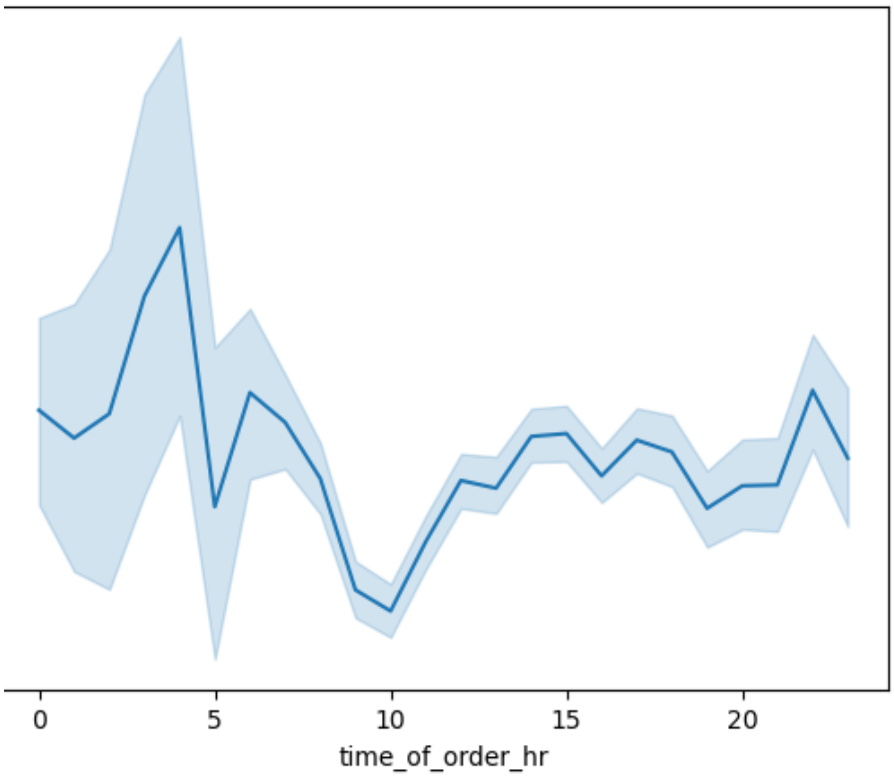
of the day are (i.e., the days and times with the most

's are
nd.
ırday
from
ı
the
(fig

people spend the most money, as this might

people are shopping between 9am and
people spend the most money between 2
and 4am (fig 2.2)

fig 2.2



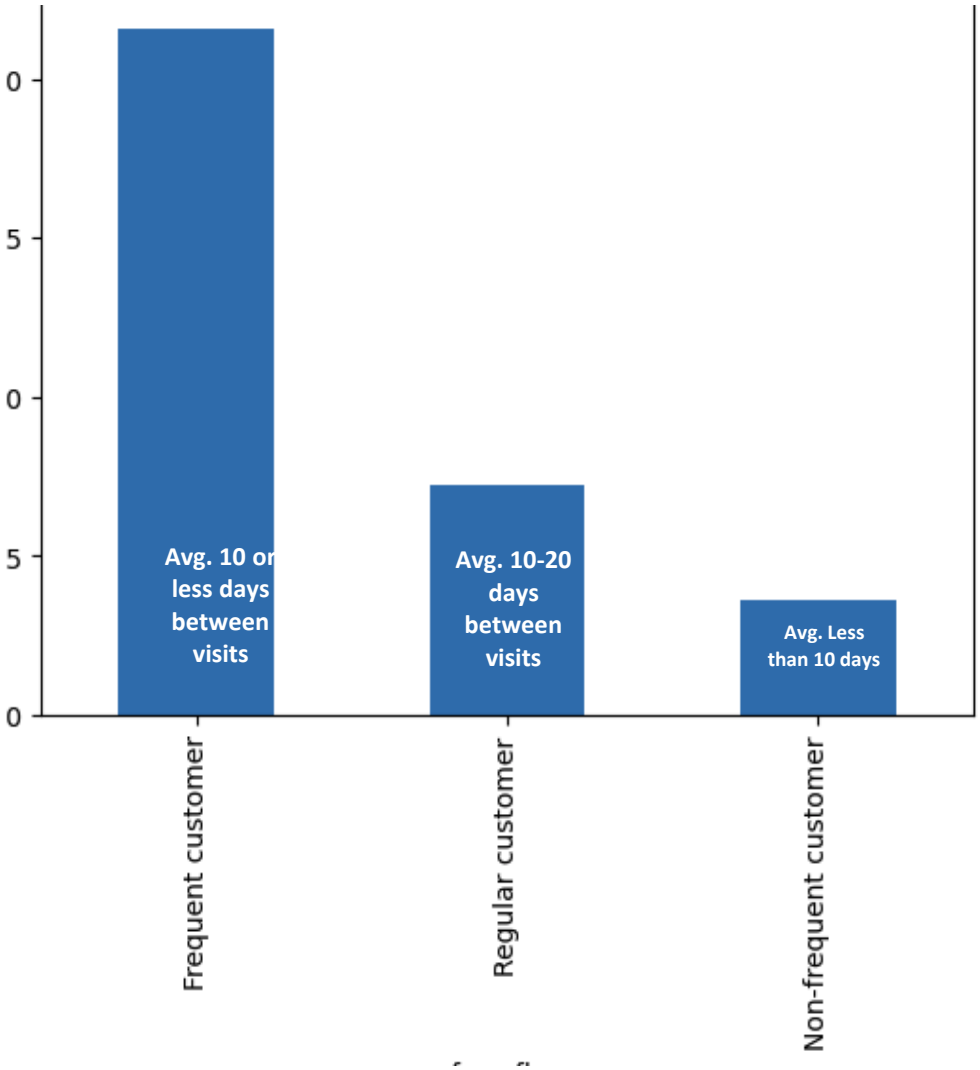
vant to use simpler price range

Price Range	Number of Products
mid-range ($5-$15)	22,086,764
low-range (less than $5)	9,900,417
high-range (more than $	714,678

arketing and sales teams want

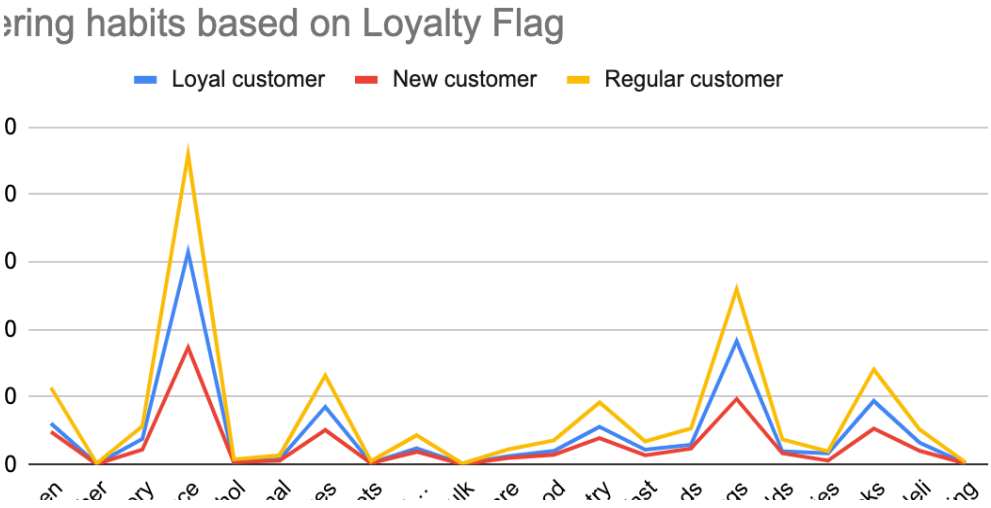
duce is by far the most popular
artment, overall.

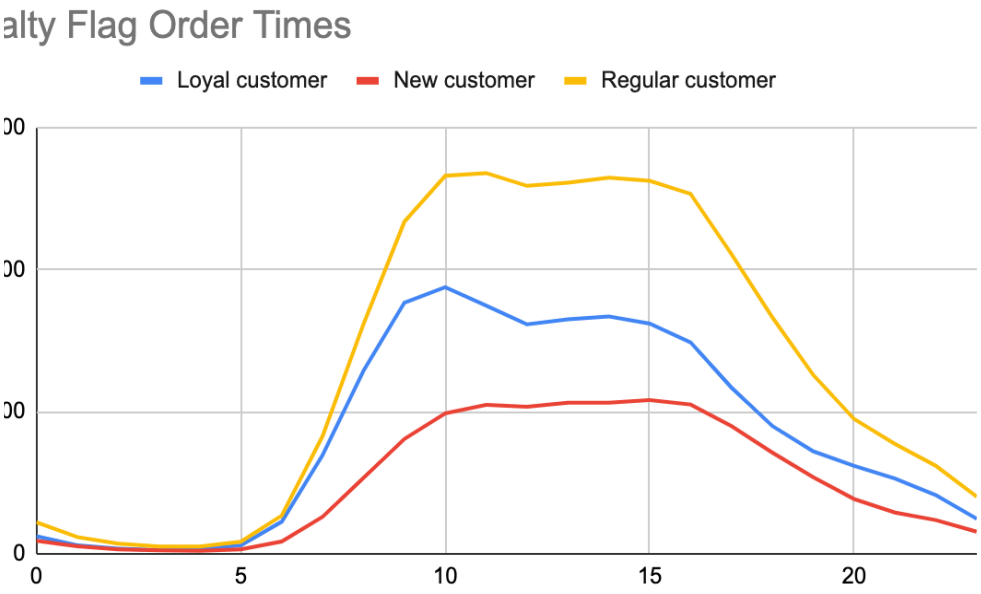
ow often do they return to

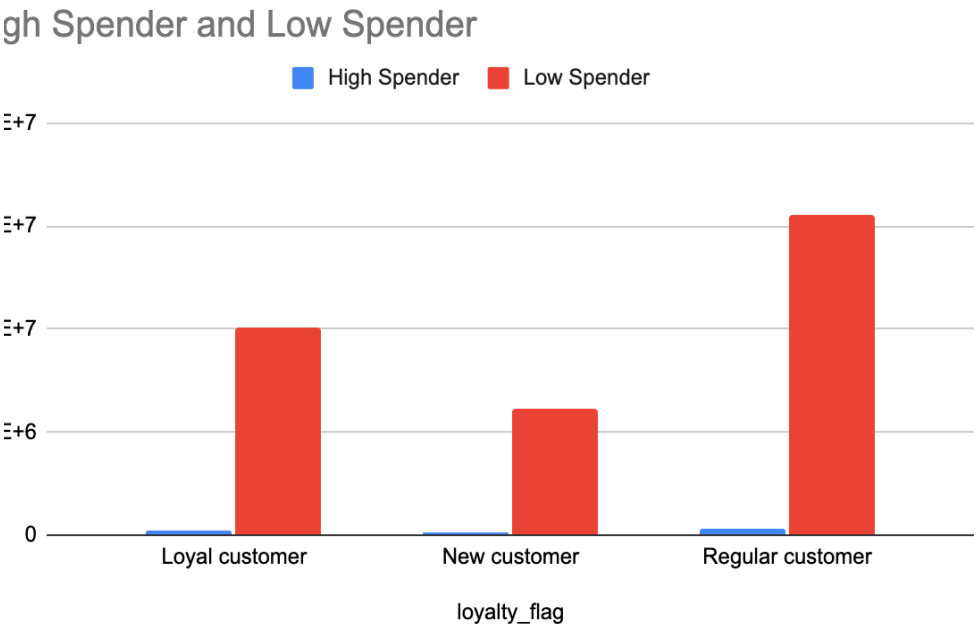


ıs?

equency. They follow the same patterns as a whole.



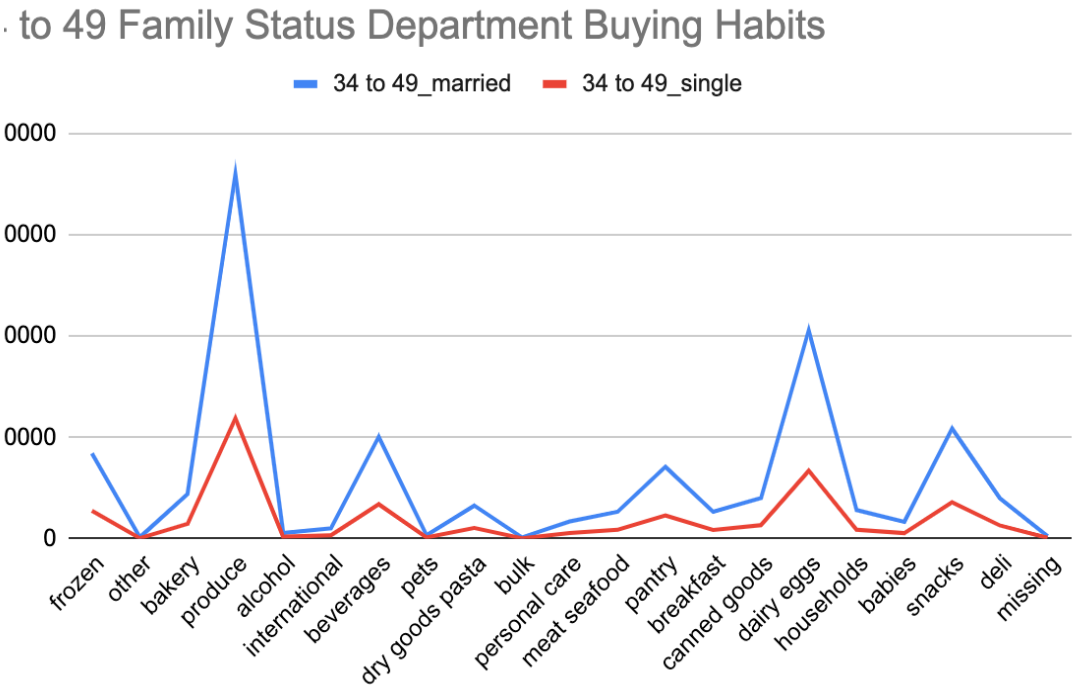


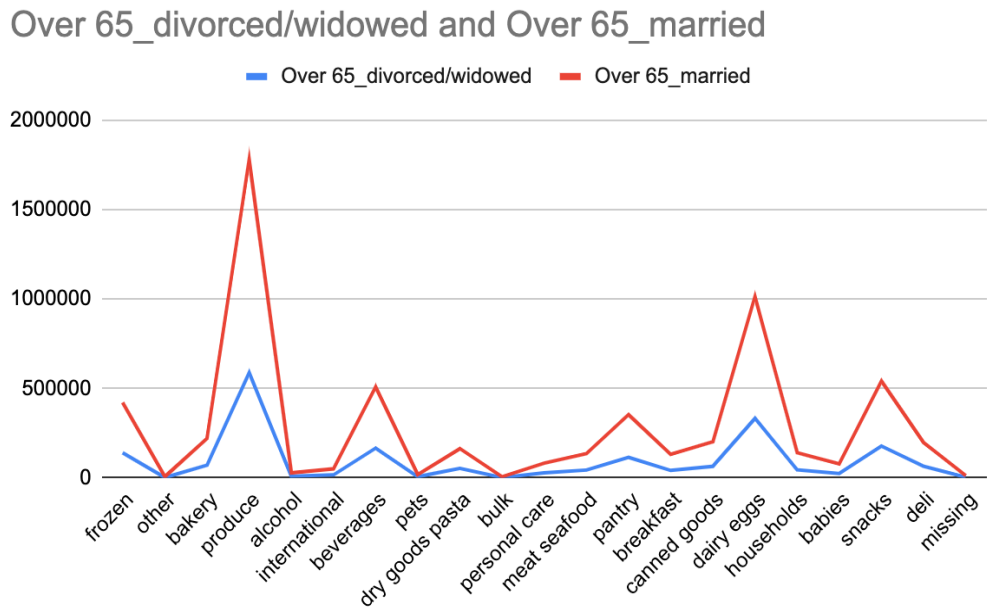


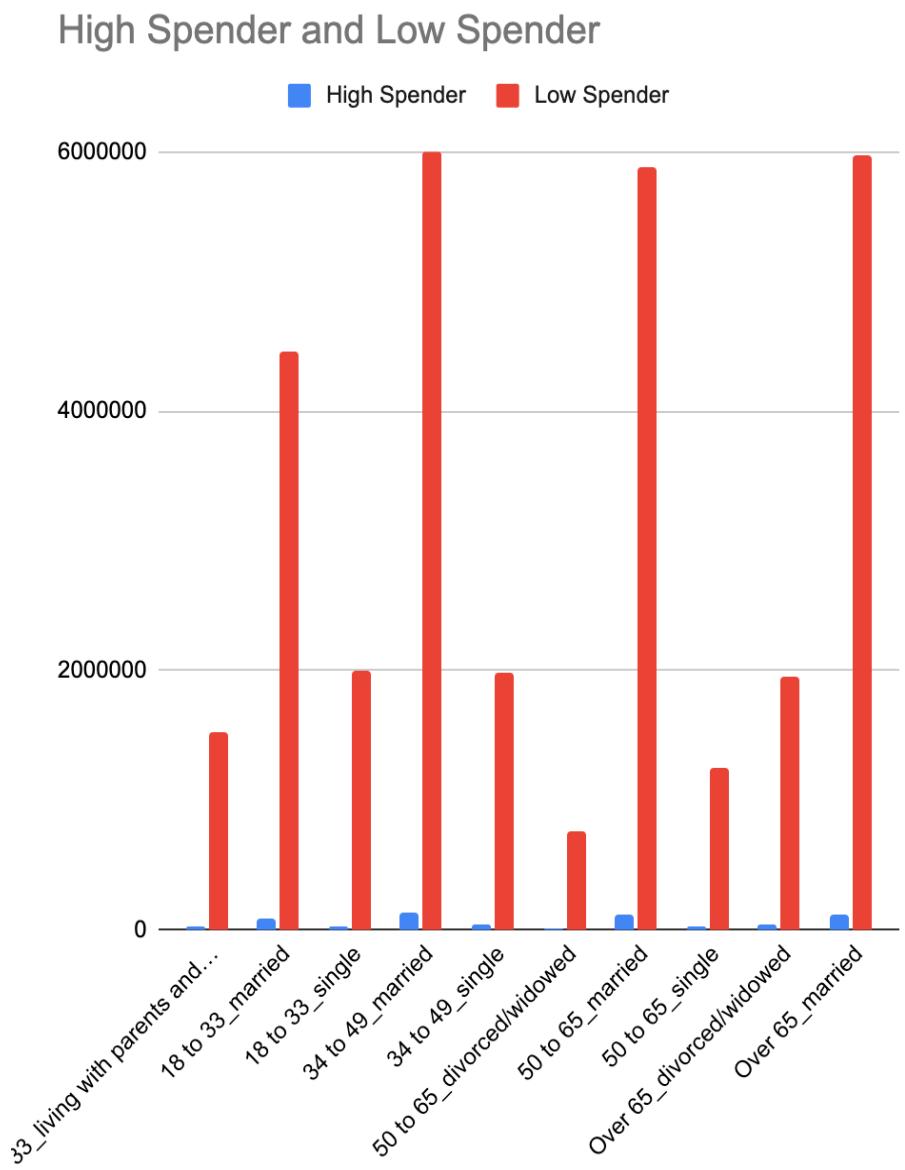
ollow the same patterns as the data as a whole.

ıits?

ıs. They follow the same patterns as the data as a whole.

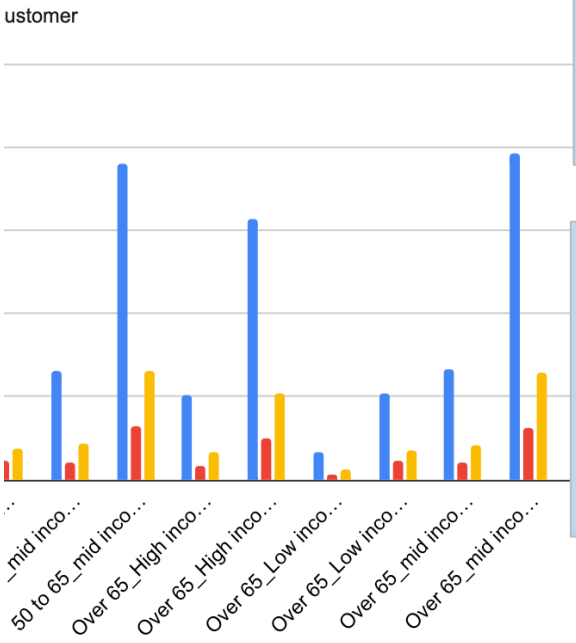






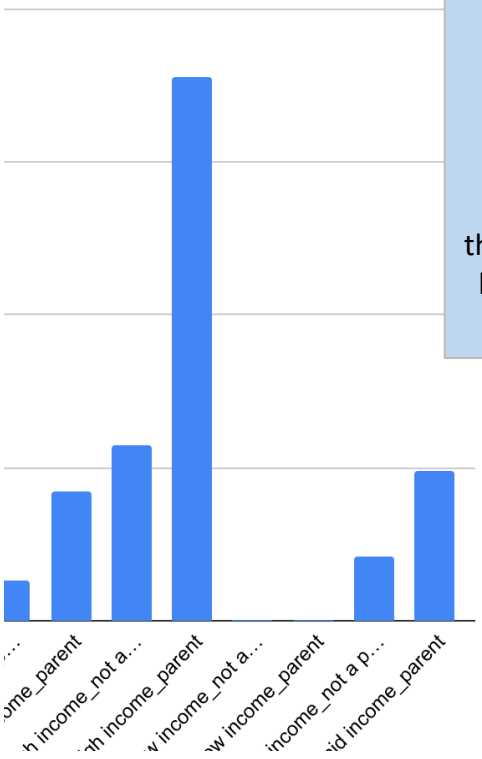
k it is important to note that
ighest spenders tend to be
rried, with the highest of
se being between 34 - 49
years of age.

? Age? Income? Certain types of



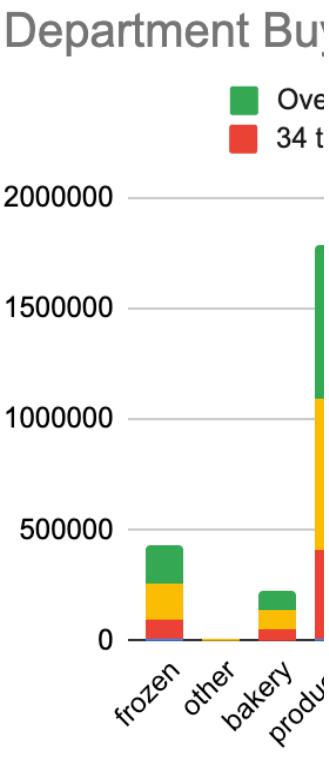
The most frequent customers tend to be parents between 18 and 65 in the mid income range.

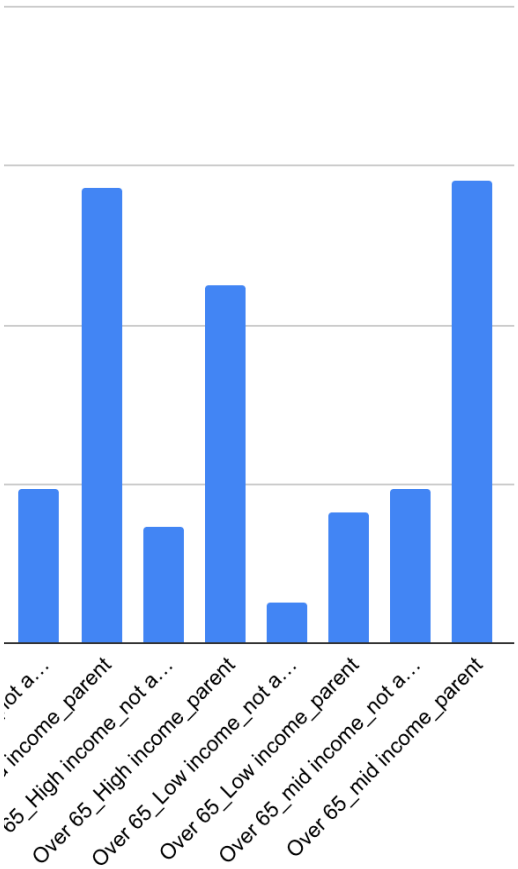
An important note is the higher the income the less frequent the customer, and 18-33 high income have some of the lowest frequencies, while thier mid to low income profiles have some of the



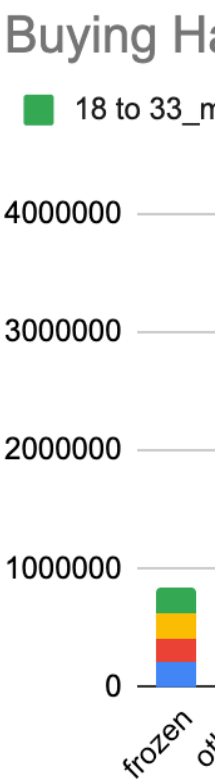
The highest spenders are parents in the high income range.

There are no departments that deviate from the general buying patterns of the data



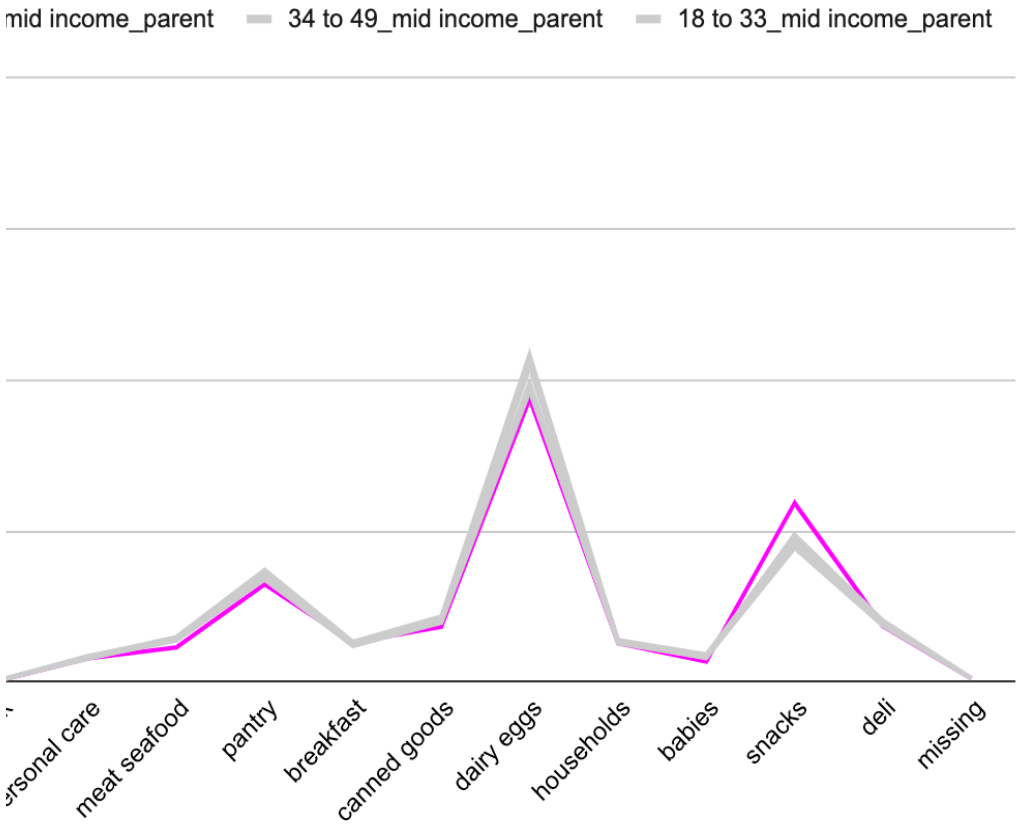


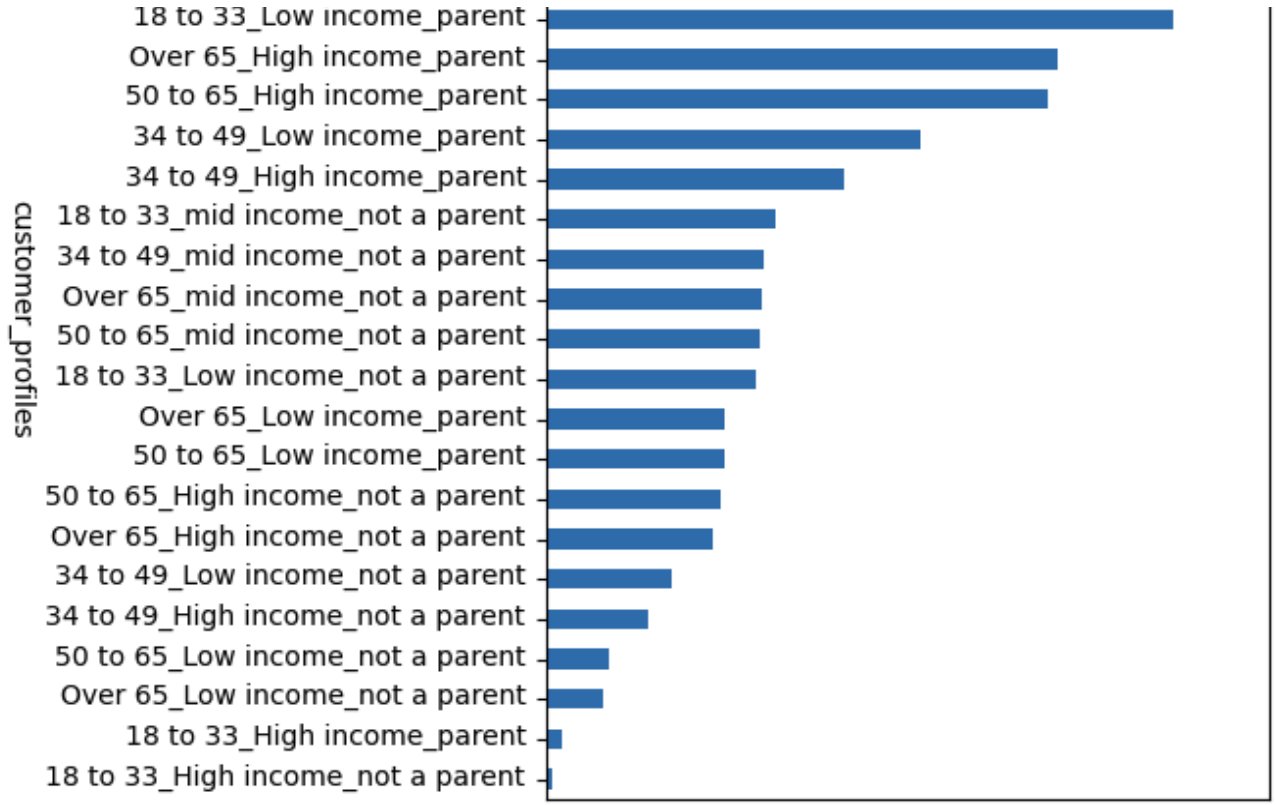
The lowest spenders are parents in the middle income range, but they buy the most amount of items.

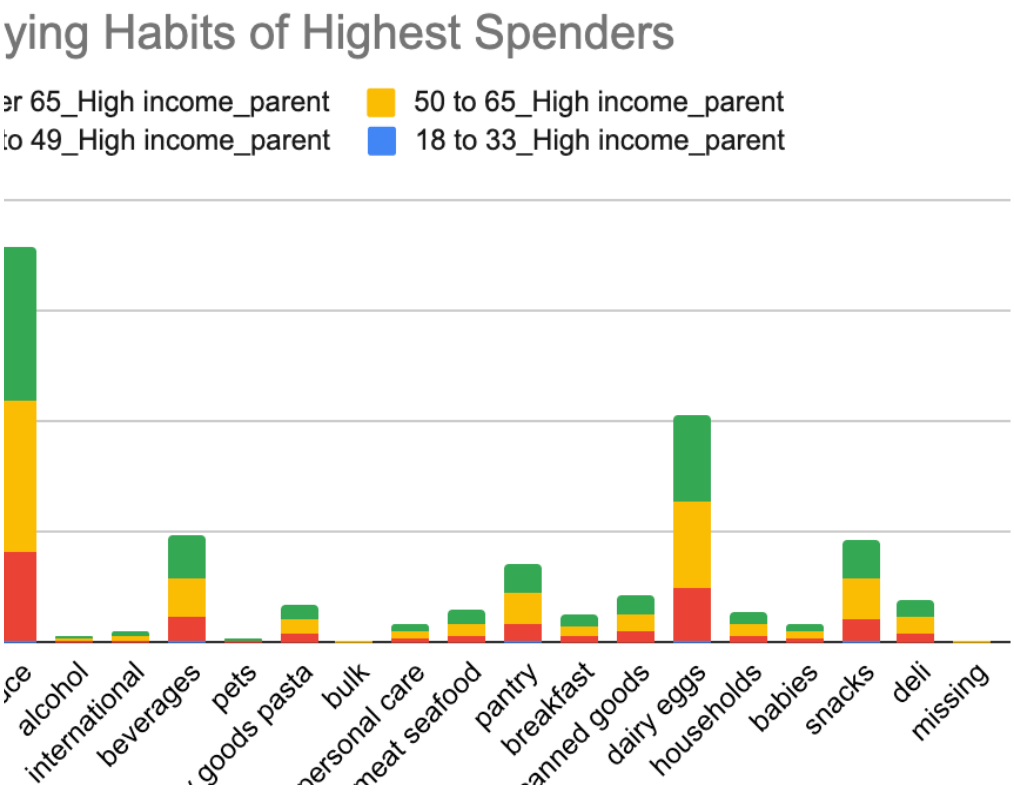


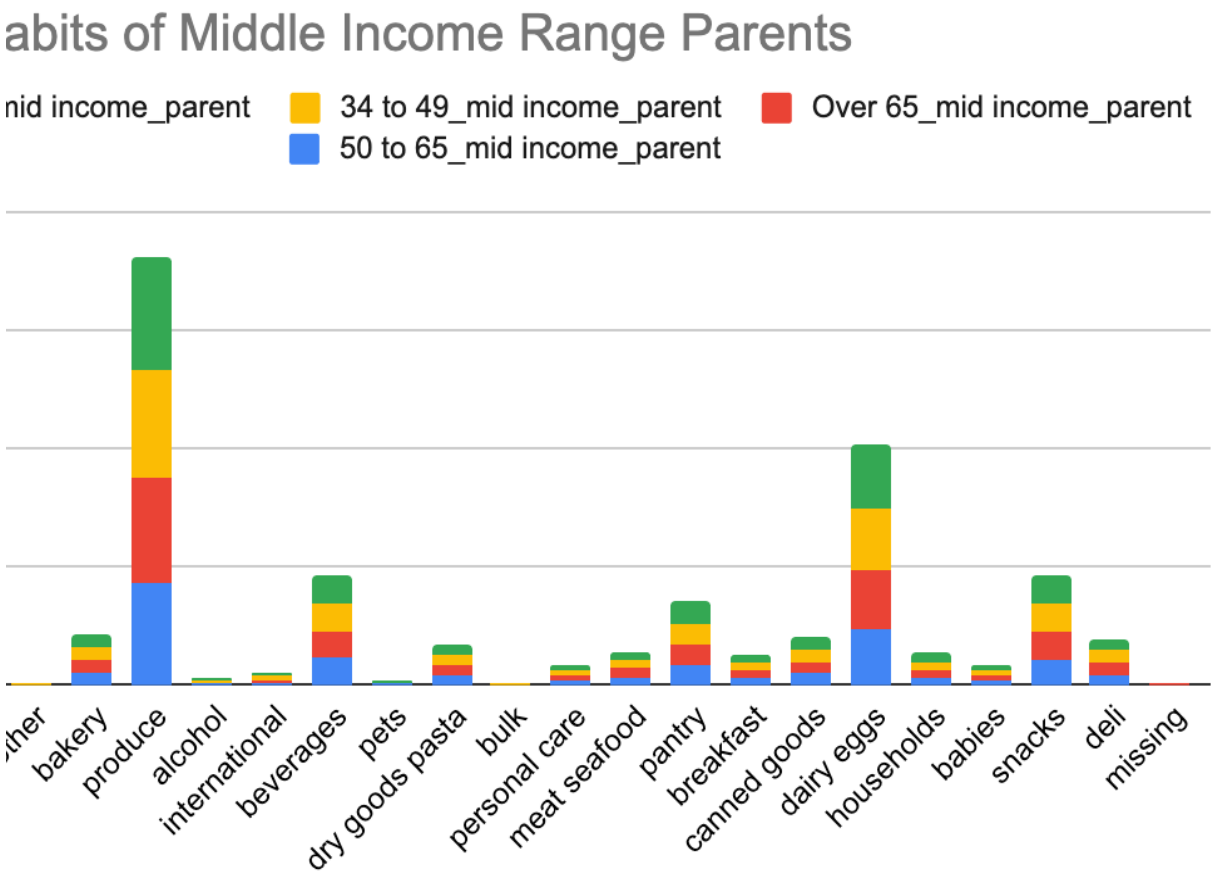
ner profiles? Consider the price of orders, anything else you can think of.

ustomer profiles




customer_profiles
18 to 33_Low income_parent
Over 65_High income_parent
50 to 65_High income_parent
34 to 49_Low income_parent
34 to 49_High income_parent
18 to 33_mid income_not a parent
34 to 49_mid income_not a parent
Over 65_mid income_not a parent
50 to 65_mid income_not a parent
18 to 33_Low income_not a parent
Over 65_Low income_parent
50 to 65_Low income_parent
50 to 65_High income_not a parent
Over 65_High income_not a parent
34 to 49_Low income_not a parent
34 to 49_High income_not a parent
50 to 65_Low income_not a parent
Over 65_Low income_not a parent
18 to 33_High income_parent
18 to 33_High income_not a parent


ying Habits of Highest Spenders
er 65_High income_parent
50 to 65_High income_parent
to 49_High income_parent
18 to 33_High income_parent
alcohol
international
beverages
pets
bulk
personal care
meat seafood
pantry
breakfast
canned goods
dairy eggs
households
babies
snacks
deli
missing


abits of Middle Income Range Parents
nid income_parent
34 to 49_mid income_parent
50 to 65_mid income_parent
Over 65_mid income_parent
bakery
produce
alcohol
international
beverages
pets
dry goods pasta
bulk
personal care
meat seafood
pantry
breakfast
canned goods
dairy eggs
households
babies
snacks
deli
missing