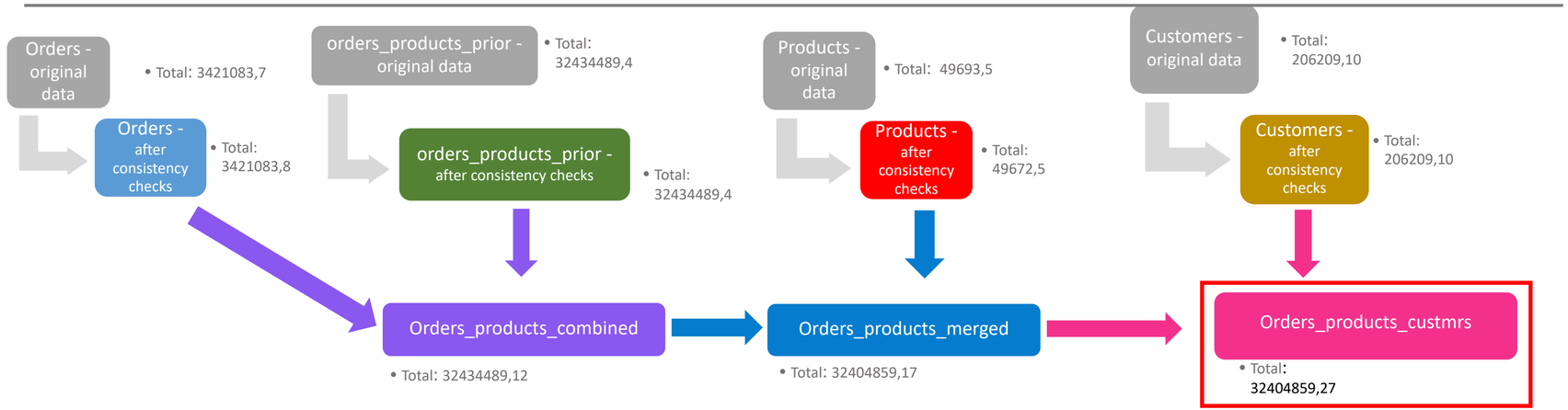
Instacart Grocery Basket Analysis

Project Name: Instacart Grocery Basket Analysis
Date: May 2023
Analyst Name: Jurgita Aciene
Citation: “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle in May 2023

Contents:

instacart

Population flow



Exclusion flag

Condition: max_order <5
Obervations to be removed: 1440295
Final total count of active_users: 30964964

instacart

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates
orders	206209 in ‘days_since_prior_order’ col	Left as is. It is 1st order of each customer	none
products	16 in ‘product_name ‘ col	deleted rows with missing	5 duplicates were removed
orders_products_prior	none	N/A	none
customers	11259 in ‘first_name’ col	Left as is. >5%, and all other info available	none

instacart

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
eval_set from orders			non useful data
_merge			After one merger, prior to another, so a new _merge col could be created.
3: Unnamed: 0_x,Unnamed: 0.1,Unnamed: 0_y			Columns were created during merging process - not necessary for analysis
	orders_dow, from orders		To: orders_day_of_week, for consistency
	First Name, from customers		To: first_name, for consistency
	Surnam, from customers		To: last_name, for consistency
	Gender, from customers		To: gender, for consistency
	STATE, from customers		To: state, for consistency
	Age, from customers		To: age, for consistency
	n_dependants, from customers		To: num_of_dependants, for consistency
		orders_dow	To: str from int
		first_name	To: str from mixed

instacart

Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
ords_prods_merge	price_range_loc	price	If <=5-Low-range product; if >5, <=15 - Mid-range product; if >15 - High-range product
ords_prods_merge	busiest_day	orders_day_of_week	If 0 - Busiest day; if 4 - Least busy; otherwise - Regular busy.
ords_prods_merge	busiest_days	orders_day_of_week	If 0 or 1 - Busiest days; if 3 or 4 - Slowest days; otherwise - Regularly busy.
ords_prods_merge	busiest_period_of_day	order_hour_of_day	If 10,11,14,15,13,12,16,9 - Most orders; if 3,4,2,5,1,0,6,23 - Fewest orders; otherwise - Average orders.
ords_prods_merge	max_order	user-id, order_number	Largest order_number for each user
ords_prods_merge	loyalty_flag	max_order	If >40 - Loyal customer; if >10, <=40 - Regular customer; if <=10 - New customer.
ords_prods_merge	average_price	user_id, prices	Mean price paid by each user
ords_prods_merge	spending_flag	average_price	If <10 - Low spender; if >=10 - High spender.
ords_prods_merge	median_order_frequency	user_id, days_since_prior_order	days_since_prior_order median for each user
ords_prods_merge	order_frequency_flag	median_order_frequency	If >20 - Non-frequent customer; if >10, <=20 - Regular customer; if <=10 - Frequent customer.
active_users	region	state	Based on provided website
active_users	activity_flag	max_order	If max_order < 5 - Low activity; If max_order >=5 - Normal activity
active_users	income_flag	income	If <50k - low earner; if >=50k, <150k - middle-class customer; if >=150k - upper-class
active_users	profile	age, num_of_dependants, fam_status	If age = 18-60, num_of_dependants >=1, fam_status = married - married with dependents; If age >60, num_of_dependants >=1 - senior with dependents; If age =18-60, num_of_dependants = 0 - single adult; If age >60, num_of_dependants = 0 - senior no dependents; If age = 18-60, num_of_dependants >=1, fam_status = living with parents and siblings - unmarried with dependents.

price_range_loc	
Mid-range product	21860860
Low-range product	10126321
High-range product	417678

busiest_day	
Regular busy	22416875
Busiest day	6204182
Least busy	3783802

busiest_days	
Regularly busy	12916111
Busiest days	11864412
Slowest days	7624336

busiest_period_of_day	
Most orders	21118071
Average orders	9997651
Fewest orders	1289137

loyalty_flag	
Regular customer	15876776
Loyal customer	10284093
New customer	6243990

income_flag	
Middle class	23706735
Upper class	3895275
Low earners	3362554

order_frequency_flag	
Frequent customer	21559853
Regular customer	7208564
Non-frequent custor	3636437
NaN	5

regions	
3.0	10791885
4.0	8292913
2.0	7597325
1.0	5722736

activity_flag	
Normal activity	30964564
Low activity	1440295

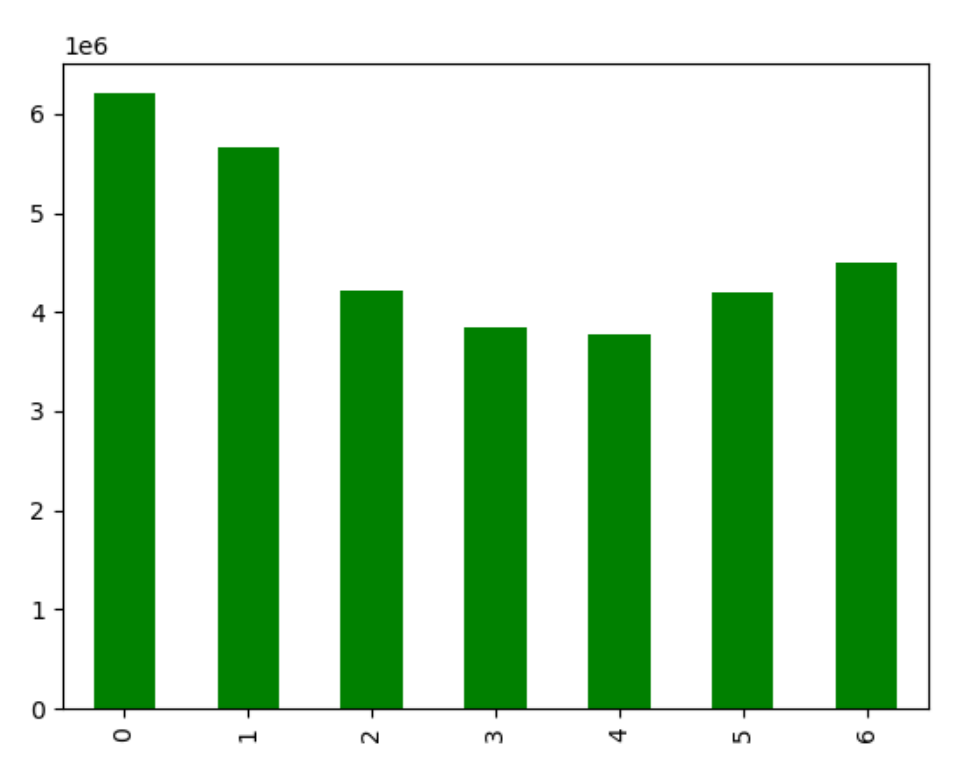
spending_flag	
Low spender	31770614
High spender	634245

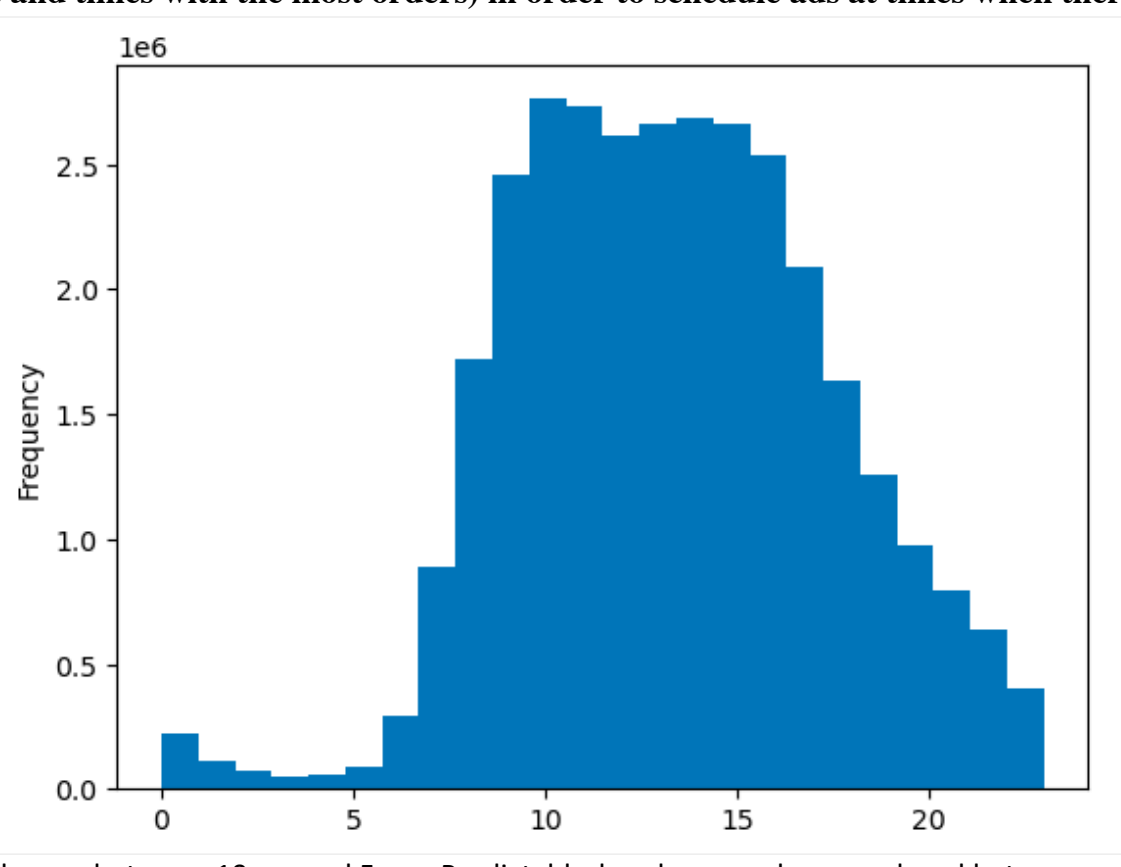
profile	
Married with dependents	14164205
Senior with dependents	7579506
Single adult	5206580
Senior no dependents	2533101
Unmarried with dependents	1481172

instacart

Visualisations

1. 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.





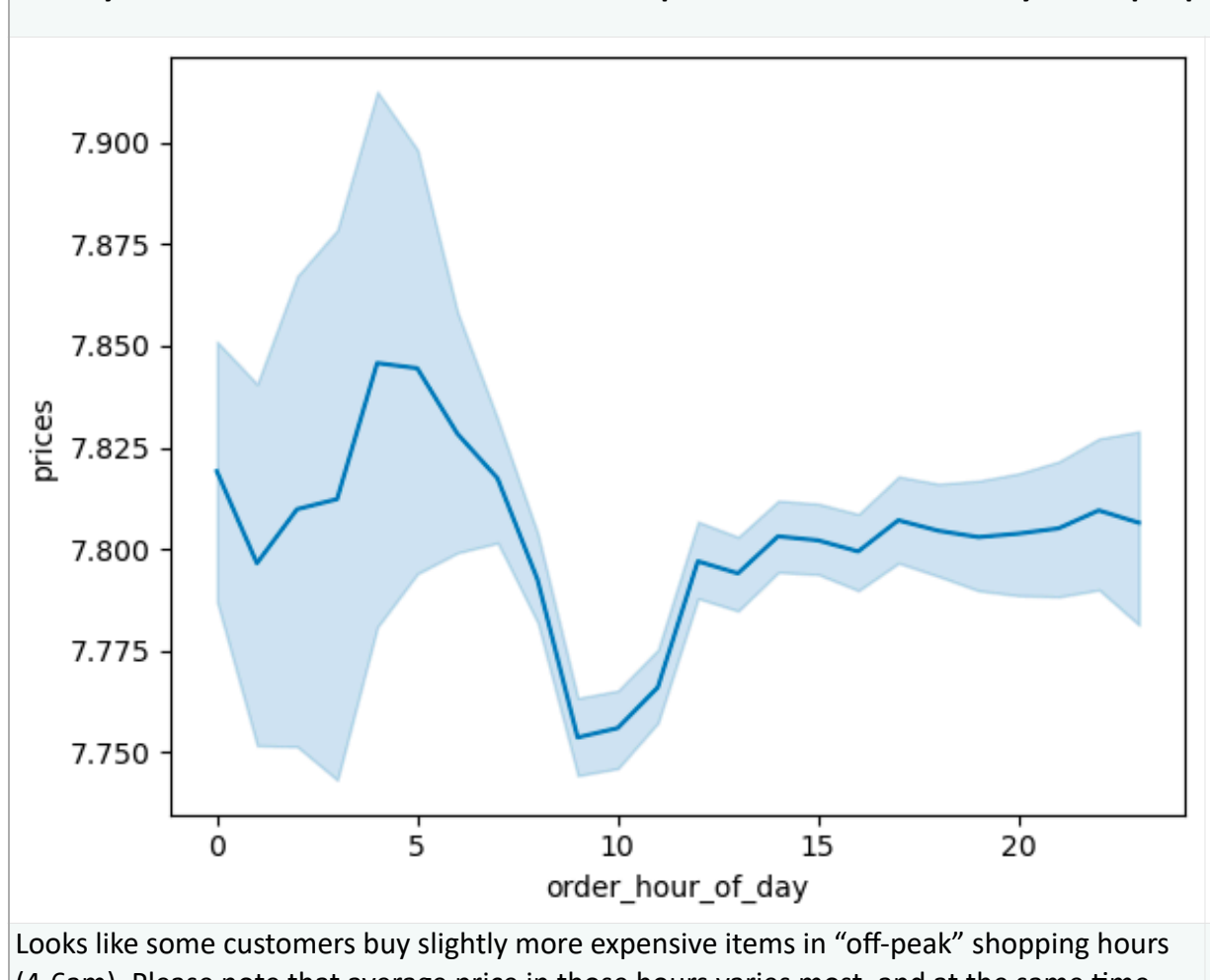
Saturdays (0), Sundays (1), and Fridays (6) are the busiest days of the week. Good days for ads are Tuesdays (3) and Wednesdays (4).

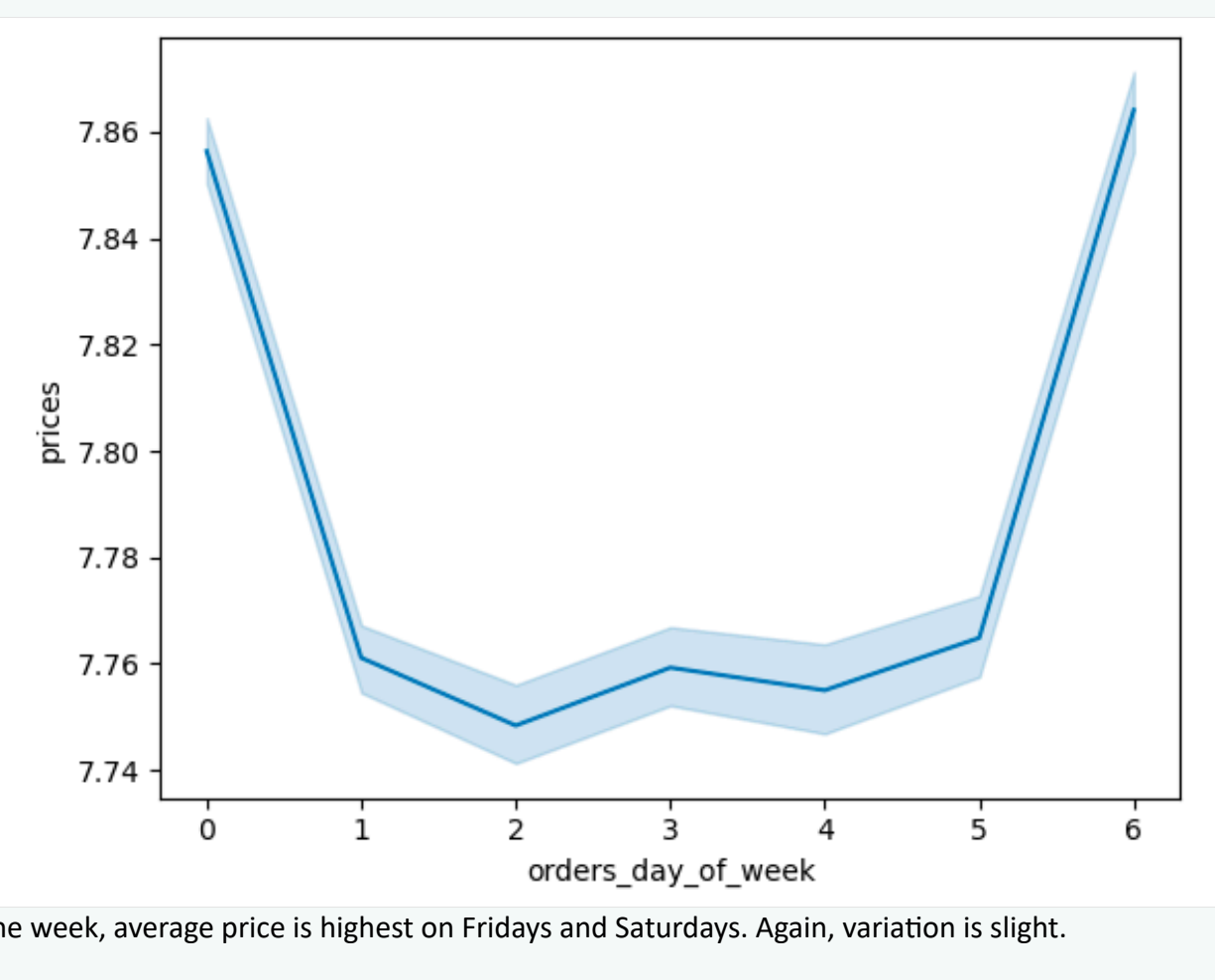
The busiest hours of the day are between 10 am and 5 pm. Predictably, barely any orders are placed between midnight and 7am.

Busiest hours

10	288418
11	284728
15	283639
14	283042
13	277999
12	272841
16	272553
9	257812
17	228795
18	182912
8	178201
19	140569
20	104292
7	91868
21	78109
22	61468
23	40043
6	30529
0	22758
1	12398
5	9569
2	7539
4	5527
3	5474

2. 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.

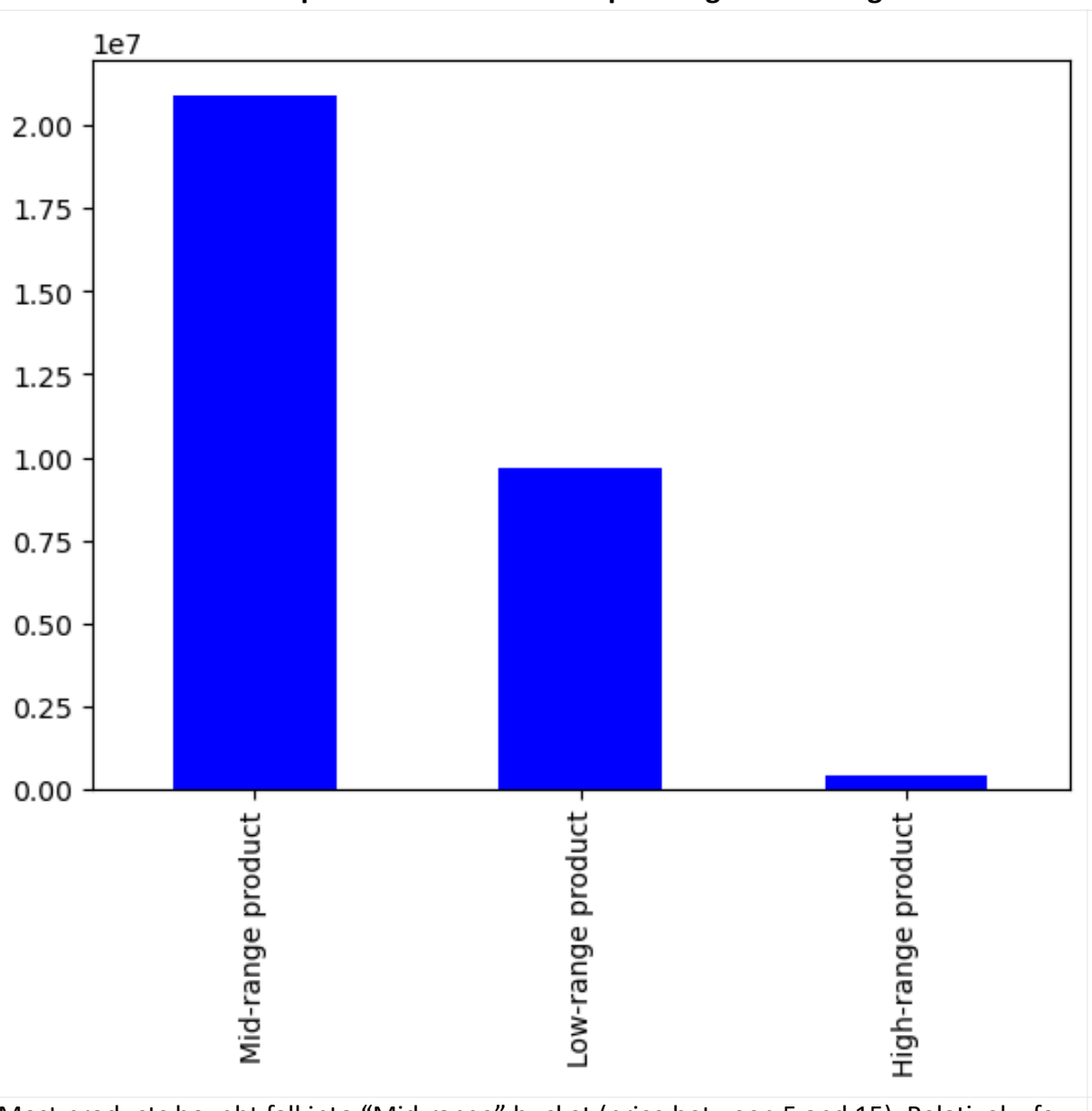


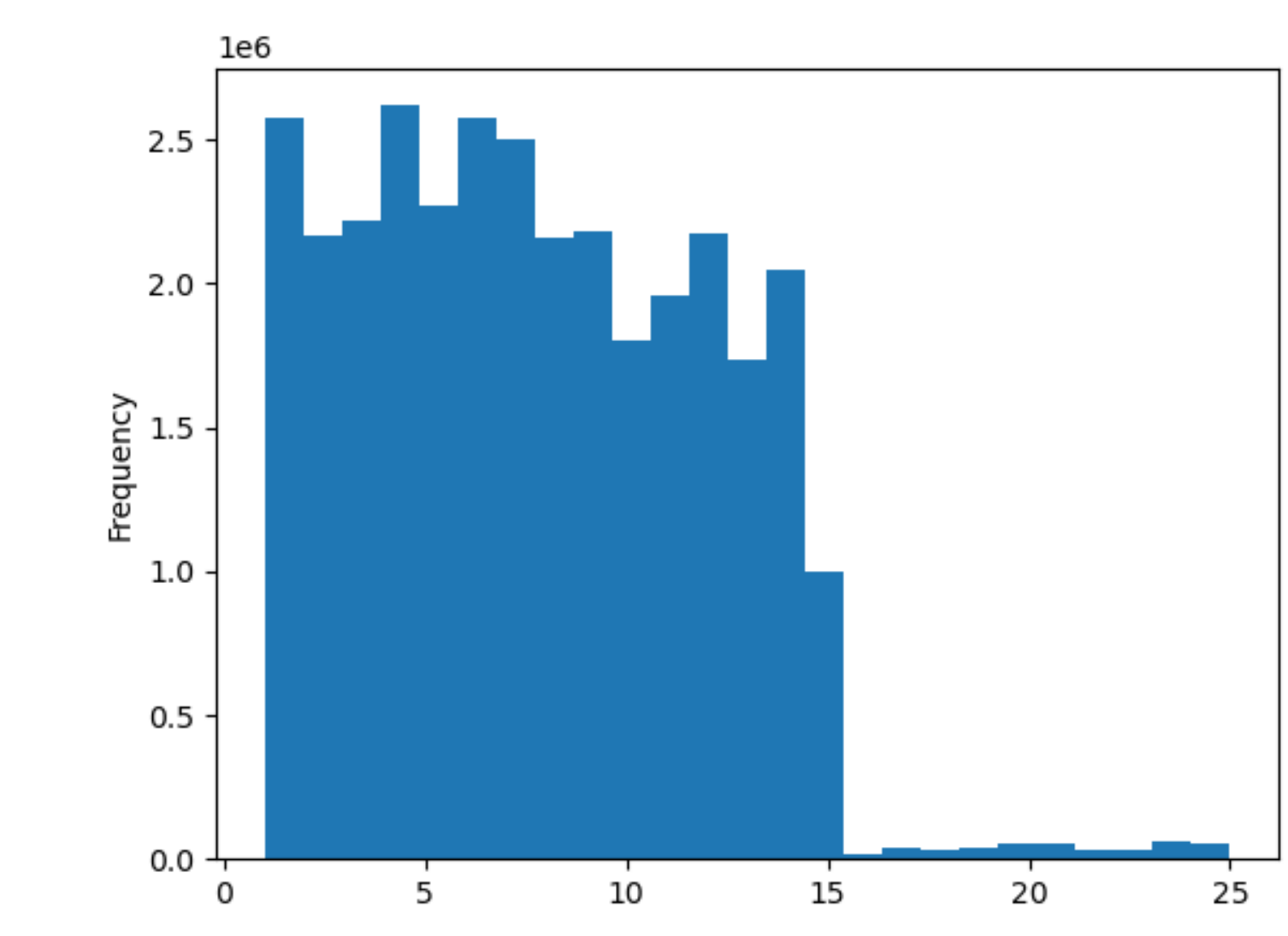


Looks like some customers buy slightly more expensive items in "off-peak" shopping hours (4-6am). Please note that average price in those hours varies most, and at the same time, the variation is not that significant (7.75 to 7.9 dollars). Cheapest items are bought between 9-11am.

In terms of days of the week, average price is highest on Fridays and Saturdays. Again, variation is slight.

3. 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.

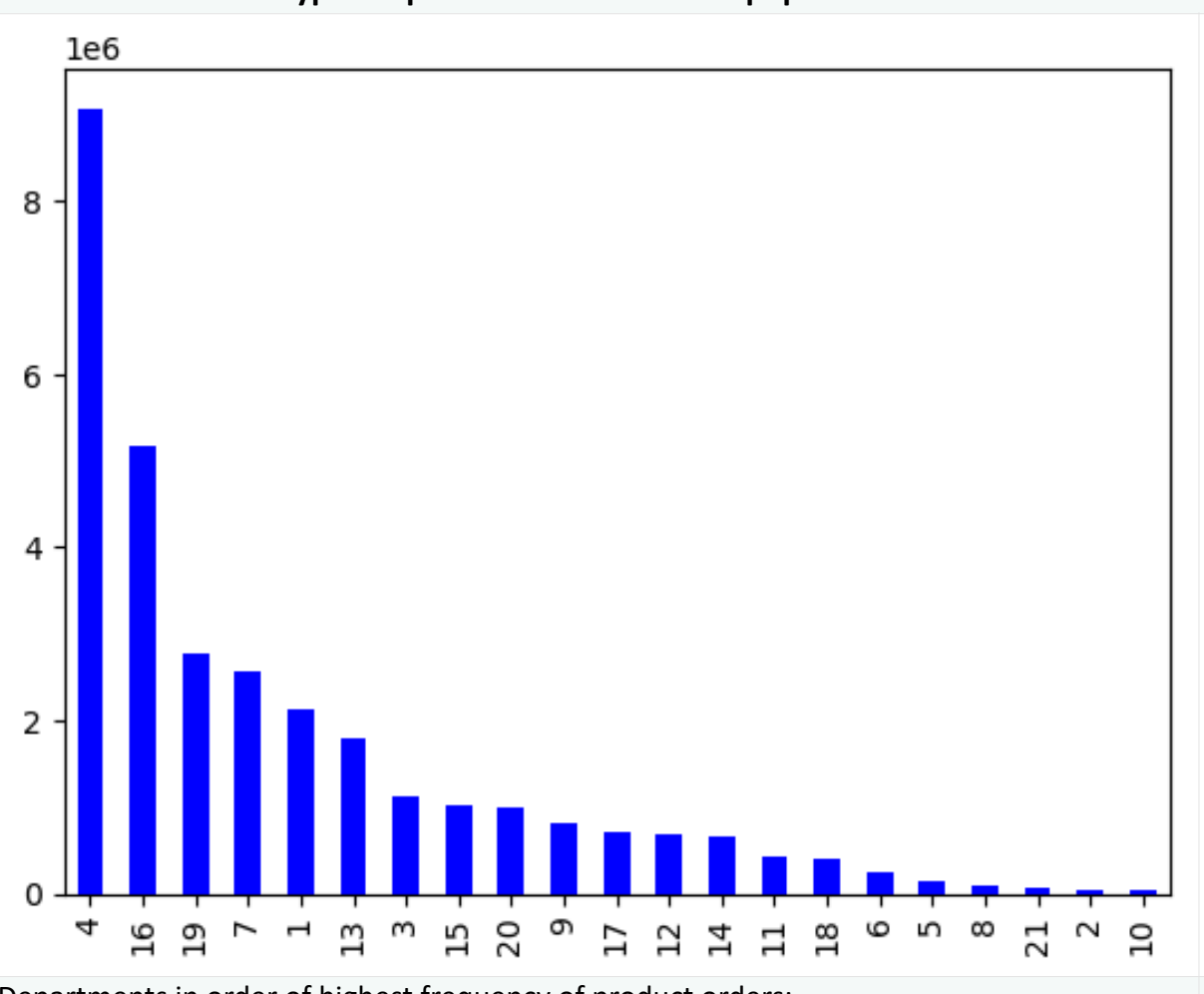


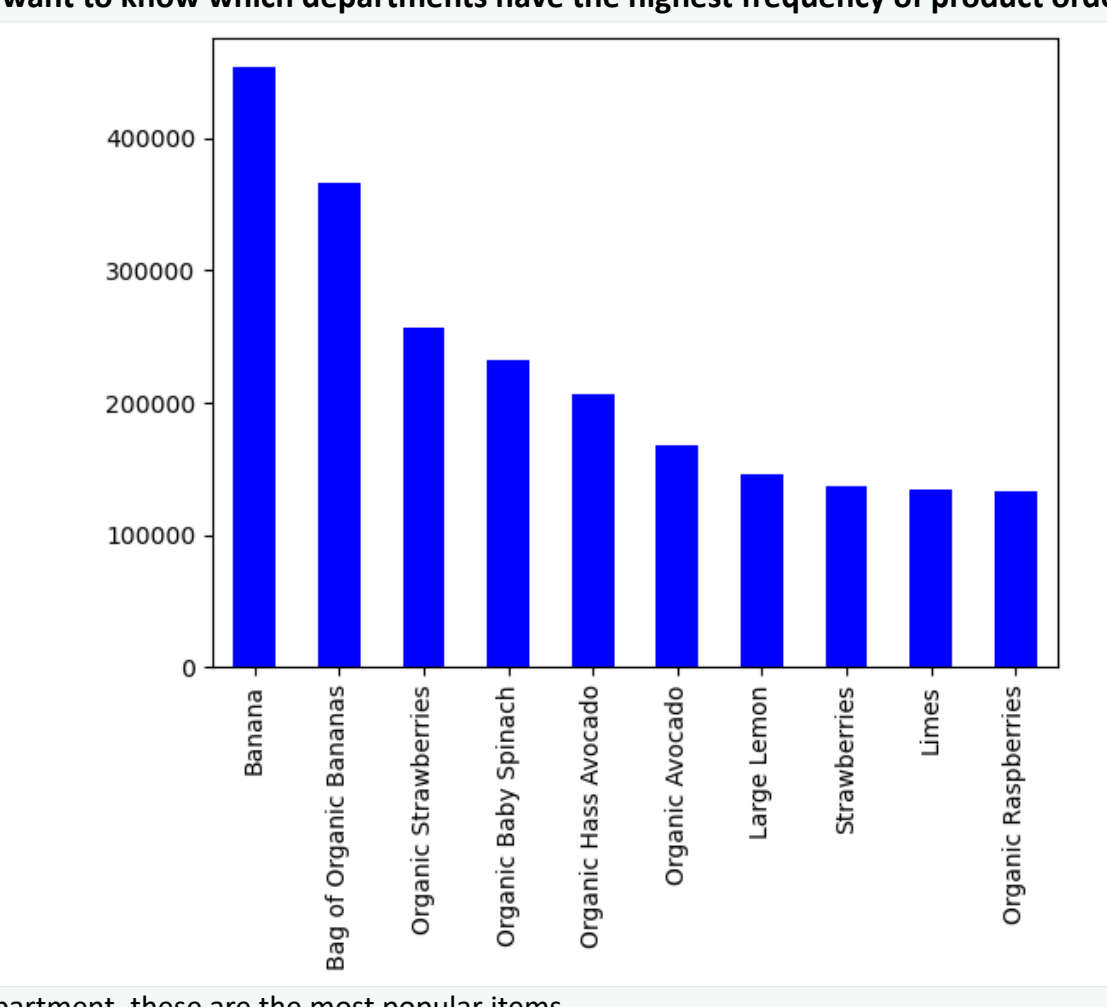


Most products bought fall into "Mid-range" bucket (price between 5 and 15). Relatively few are "High-range" (over 15 dollars). "Low-range" orders fall in between the two.

This histogram shows that actually most orders would fall into least expensive bucket, if we picked equal size (in terms of prices) buckets. The bar graph on the left is skewed towards mid-range, as it is the largest bucket.

4. 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.



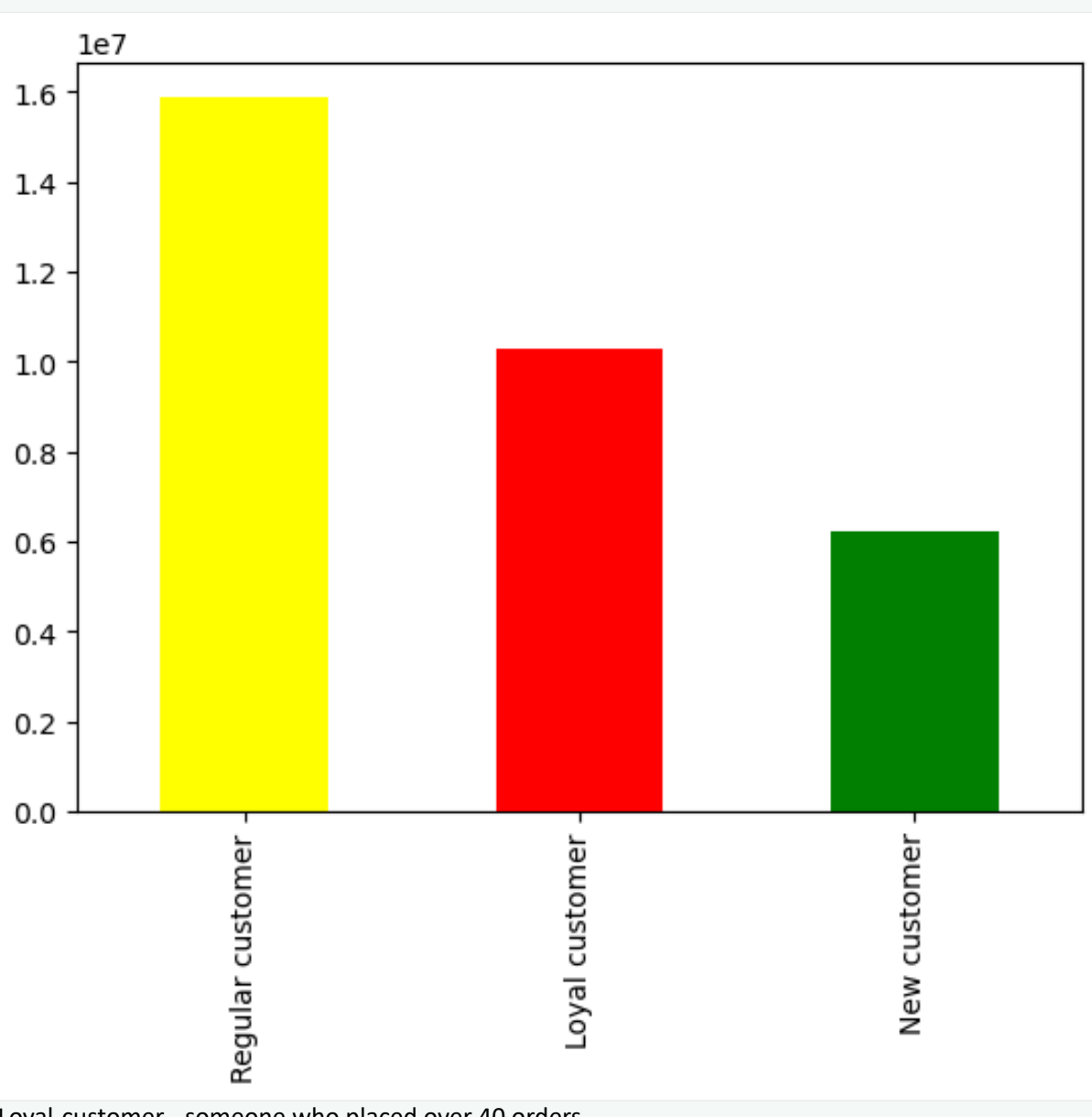


Departments in order of highest frequency of product orders:
4 = produce
16 = dairy eggs
19 = snacks
7 = beverages
1 = frozen
13 = pantry
3 = bakery
15 = canned goods
20 = deli
9 = dry good pasta
17 = household
12 = meat seafood
14 = breakfast
11 = personal care
18 = babies
6 = international
5 = alcohol
8 = pets
21 = missing
2 = other
10 = bulk

Within the produce department, these are the most popular items.

5. 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:

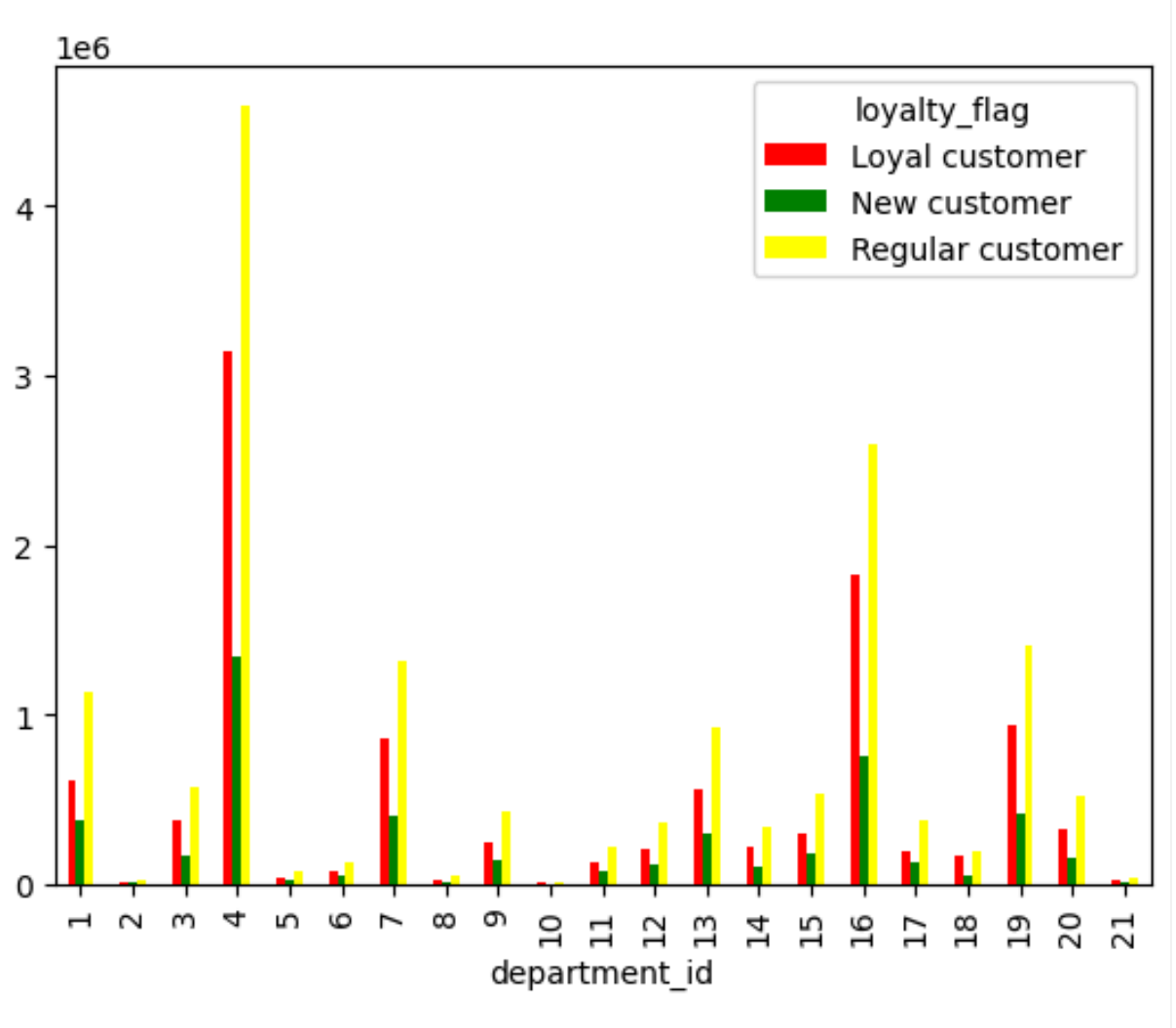
5a. What's the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?



Loyal-customer - someone who placed over 40 orders
Regular-customer - 10-40 orders under the belt
New-customer - 10 or less orders

While each individual Loyal customer placed most orders (>40 each), the largest number of overall orders was placed by regular customers (roughly 16M out of a total of 32.4M), followed by Loyal and only then - New.

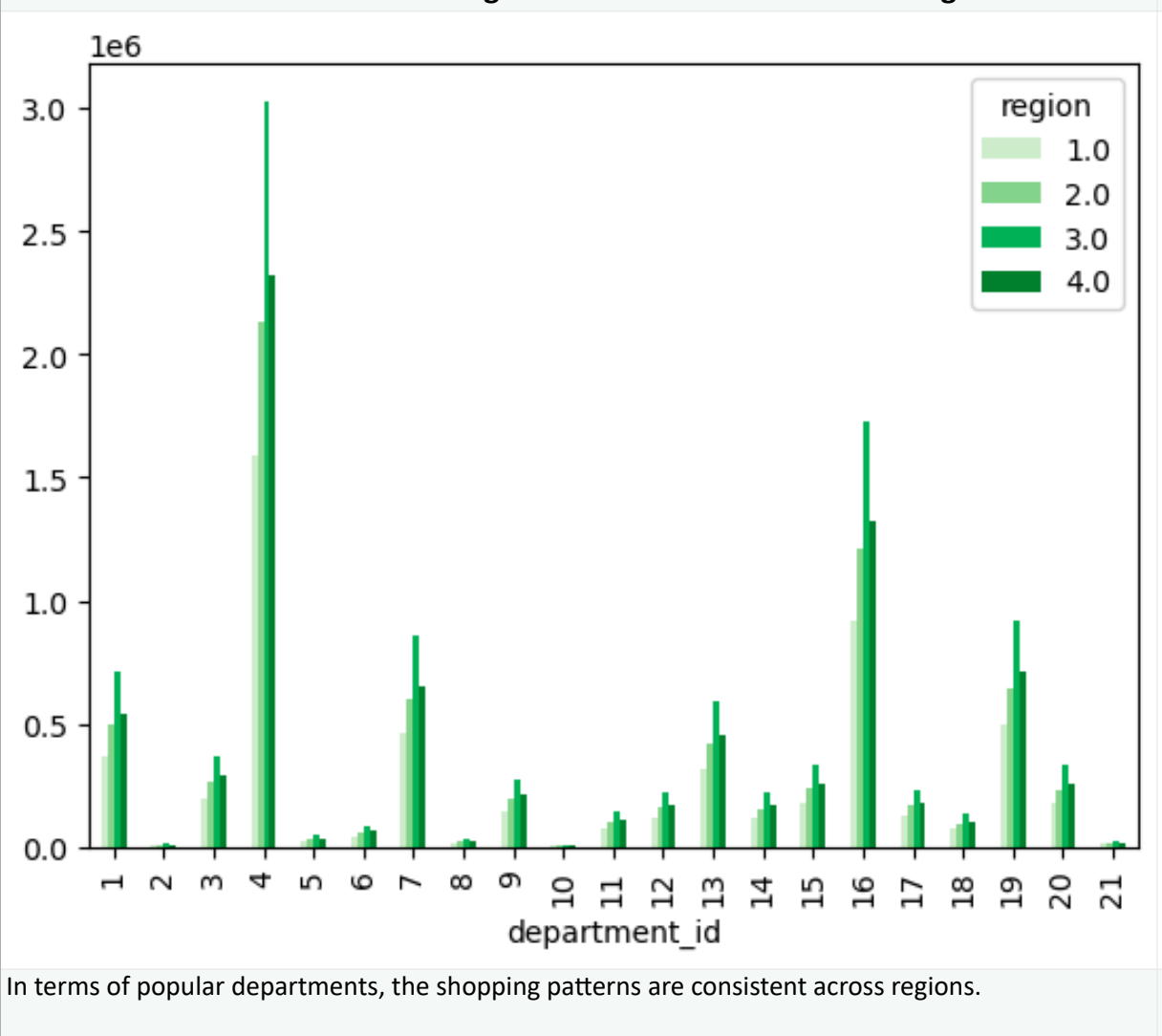
5b. Are there differences in ordering habits based on a customer's loyalty status?



In terms of popular departments, the shopping patterns are consistent across loyalty buckets.

In terms of day of week, all loyalty buckets spend most on Fridays and Saturdays and average price of orders on those days is very similar (7.86). Loyal customers spend on average slightly less on the other days of the week (7.72-7.75 compared to 7.76-7.78). New customers spend on average more than others on Sundays.
In terms of time of the day, all loyalty buckets pay on average least around 9-10am. For Loyal customers average price in those hours is the lowest of the three groups. Loyal and regular customers pay on average most during the night shoppings between 11pm and 7am. New customers pay most for a couple of hours right after midnights, but also mid-day between 1 and 6pm. The largest variation of price can be seen during the night shopping.

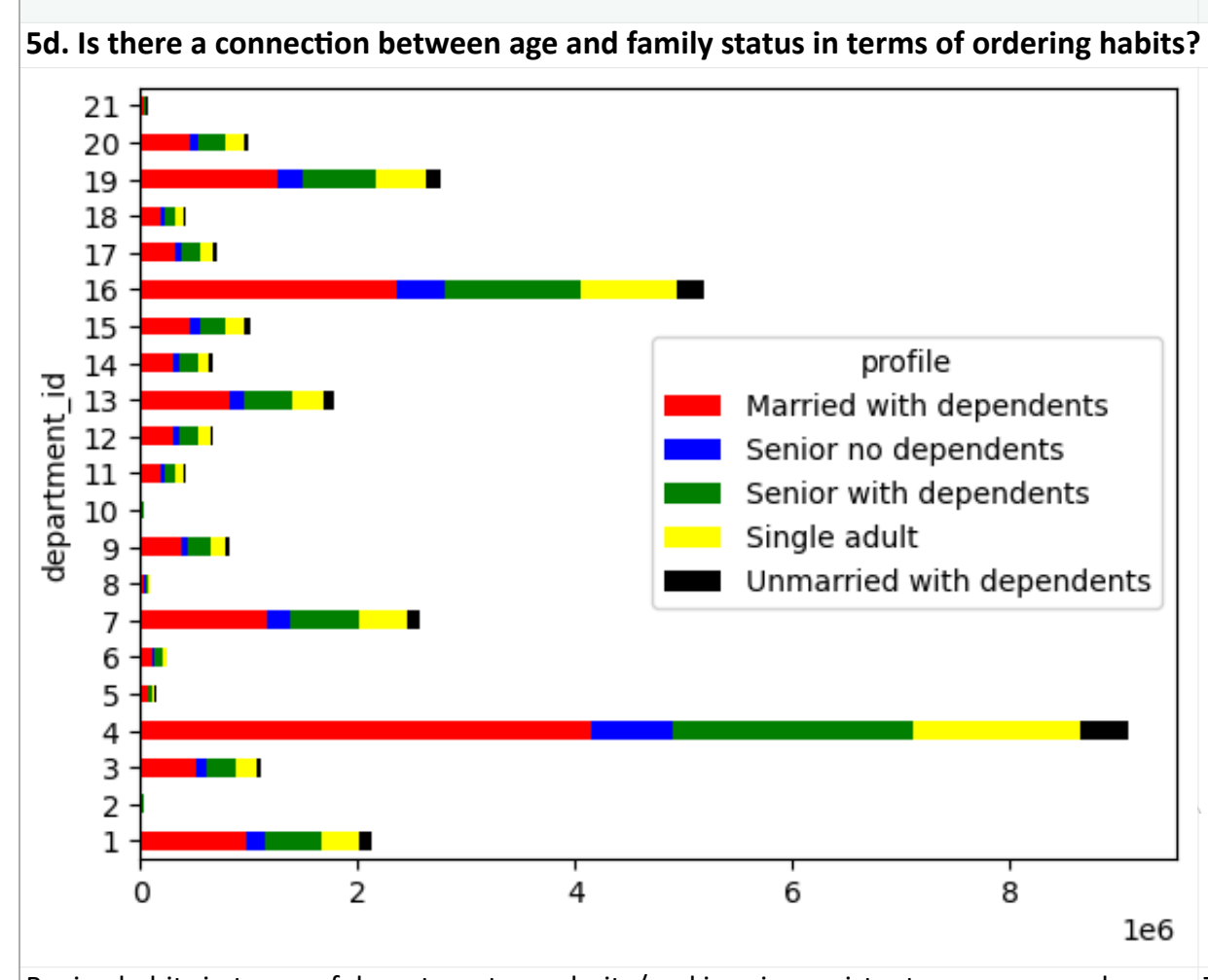
5c. Are there differences in ordering habits based on a customer's region?

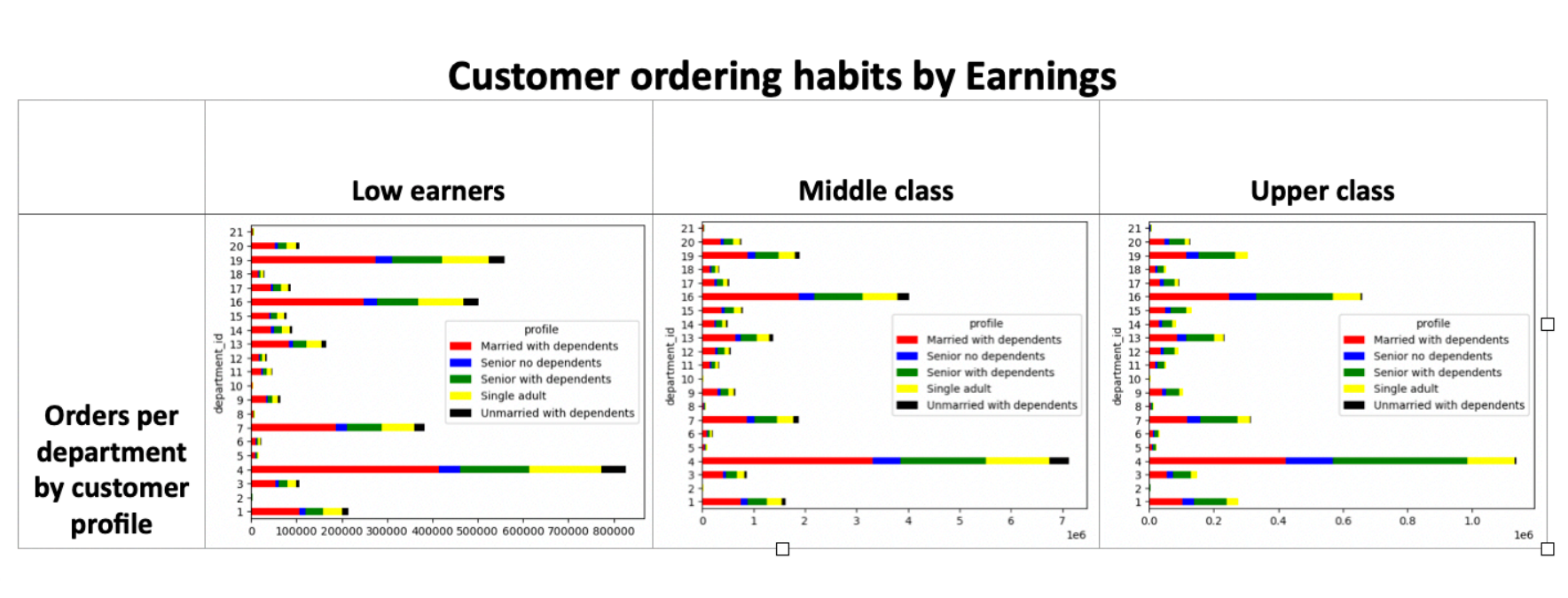


In terms of popular departments, the shopping patterns are consistent across regions.

In terms of day of week, all regions spend most on Fridays and Saturdays and average price of orders on those days is very similar (7.86). Midweek, the days when customers spend least on average vary very slightly. Two days with lowest average price are as follow:
* Northeast: Monday and Tuesday,
* Midwest: Monday and Wednesday,
* South: Sunday and Monday,
* West: Wednesday and Thursday.
In terms of time of the day, while the pattern varies slightly, the lowest average price paid is around 9-10am across all regions. Night shoppers pay most on average, peaking between 2am and 6am. The lowest price is around 7.74 across regions, the highest differs by a couple of cents only from 7.85 in the South to 7.89 in the West.

5d. Is there a connection between age and family status in terms of ordering habits?

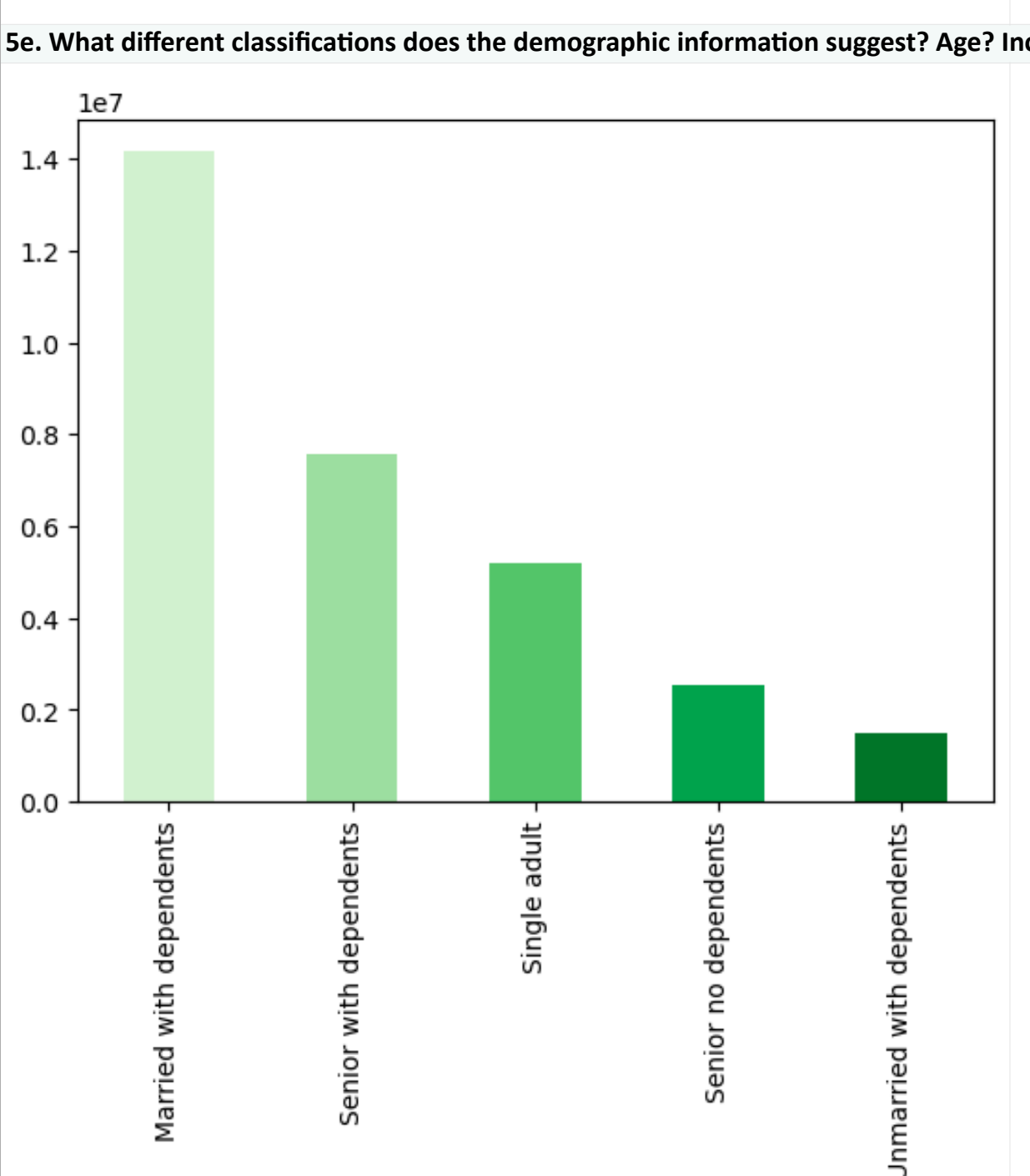


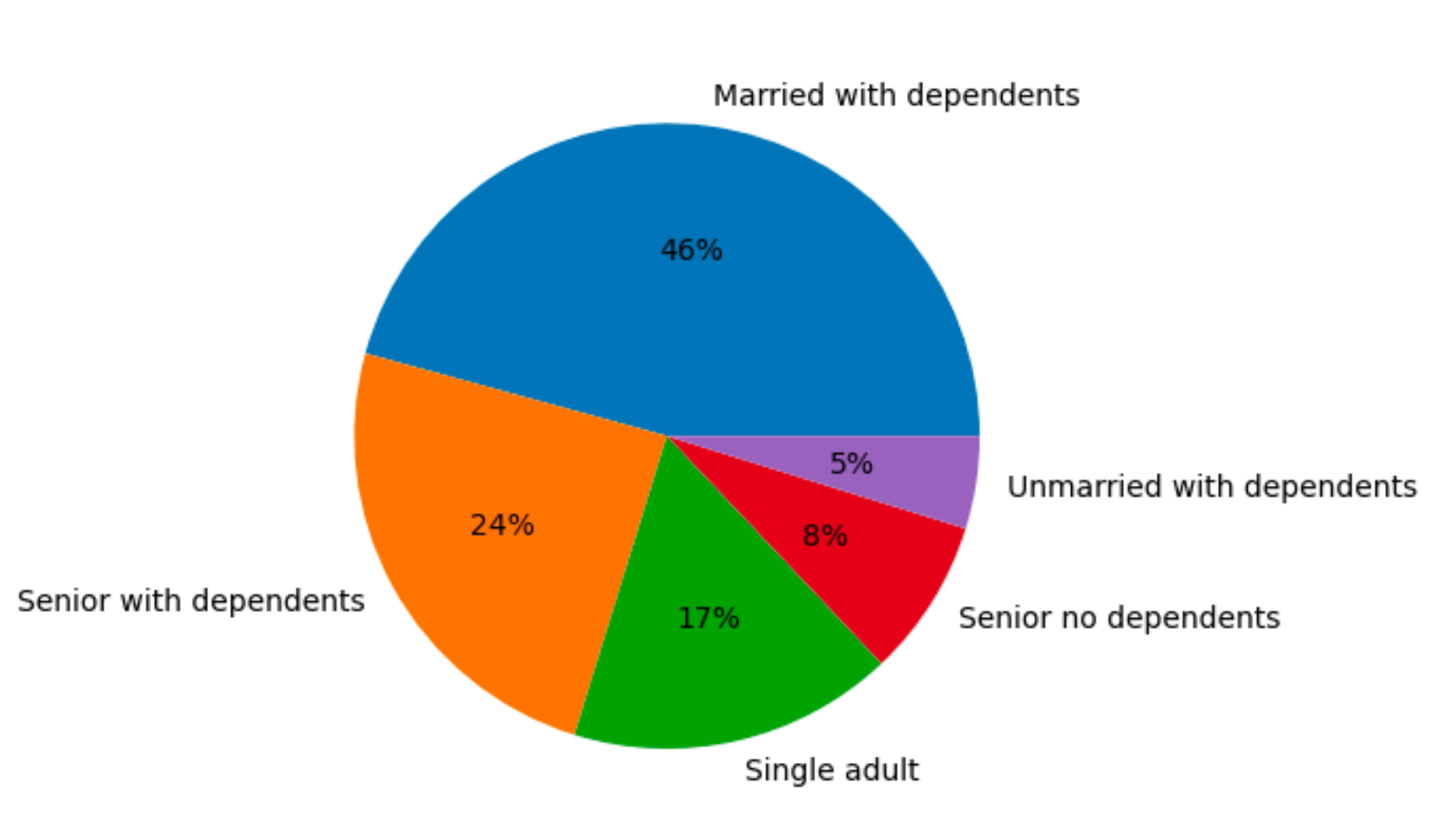


Buying habits in terms of department popularity/rankings is consistent across age and family status (including buying for babies (department # 18).

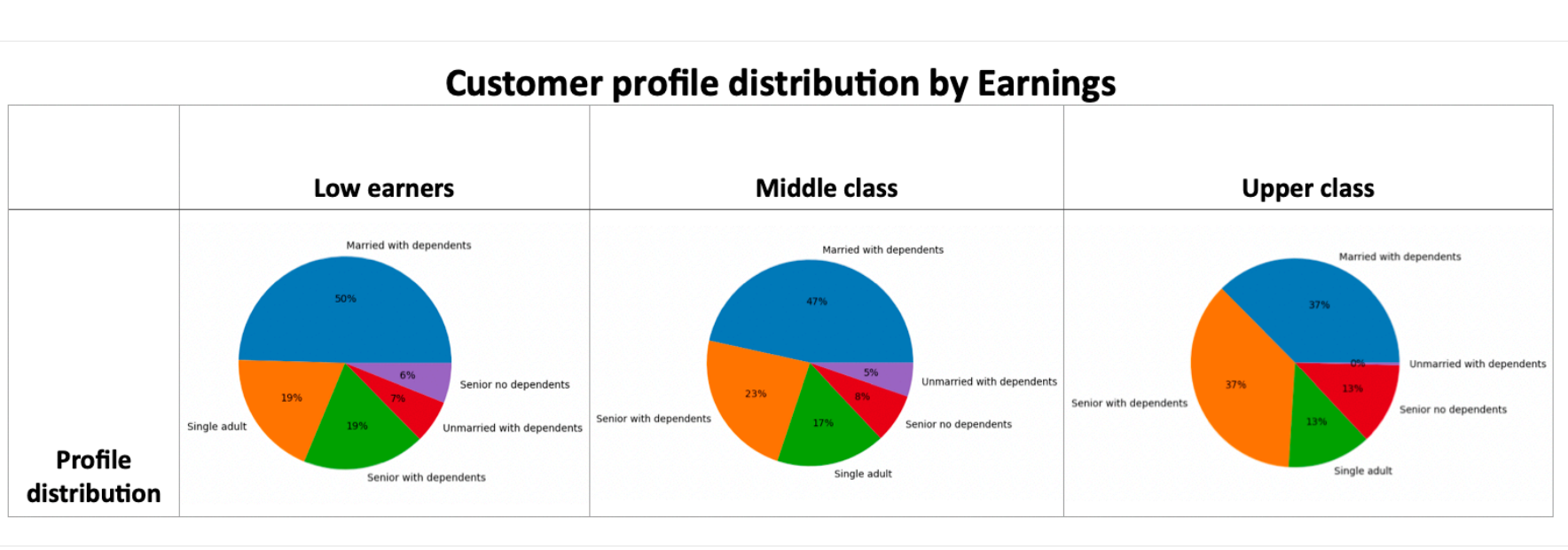
The only difference I noticed after looking at buying habits across departments several different ways, that low income customers (under 50k) buy more snacks (department # 19) than other groups (middle class 50-150k, and upper class >150k). For low income group snack outrank dairy & eggs. While middle class and upper class graphs closely resemble entire population one, lower earners order less produce and more snacks (19), dairy & eggs (16), and beverages (7).

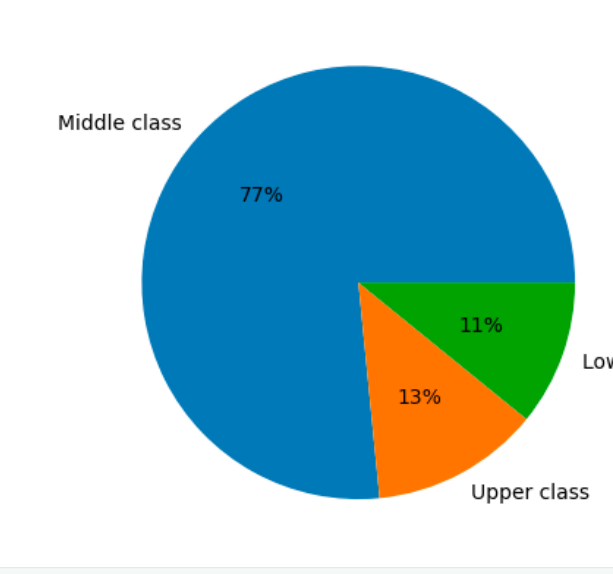
5e. What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?

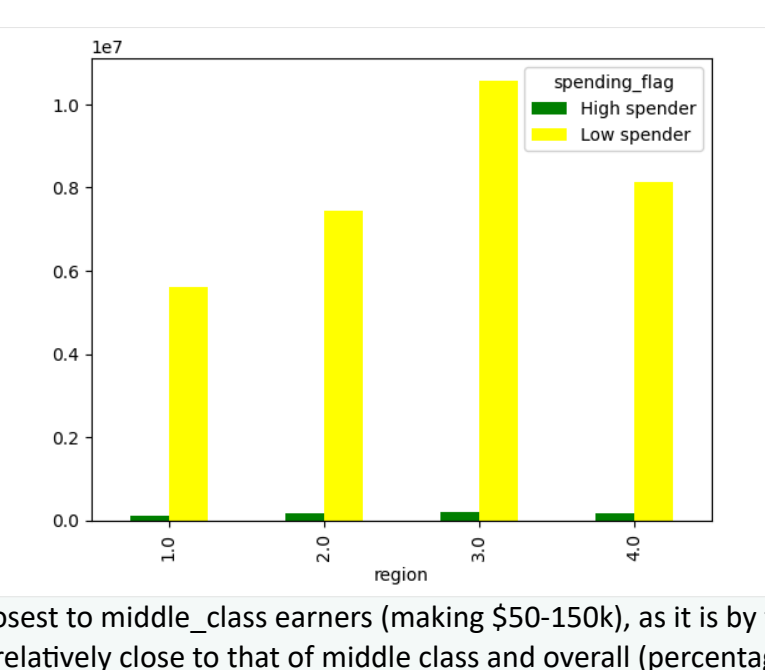




income_flag	Low earners	Middle class	Upper class
profile			
Married with dependents	1666765	11039673	1457767
Senior no dependents	203163	1837936	492002
Senior with dependents	628327	5525648	1425531
Single adult	643757	4057818	505005
Unmarried with dependents	220542	1245660	14970

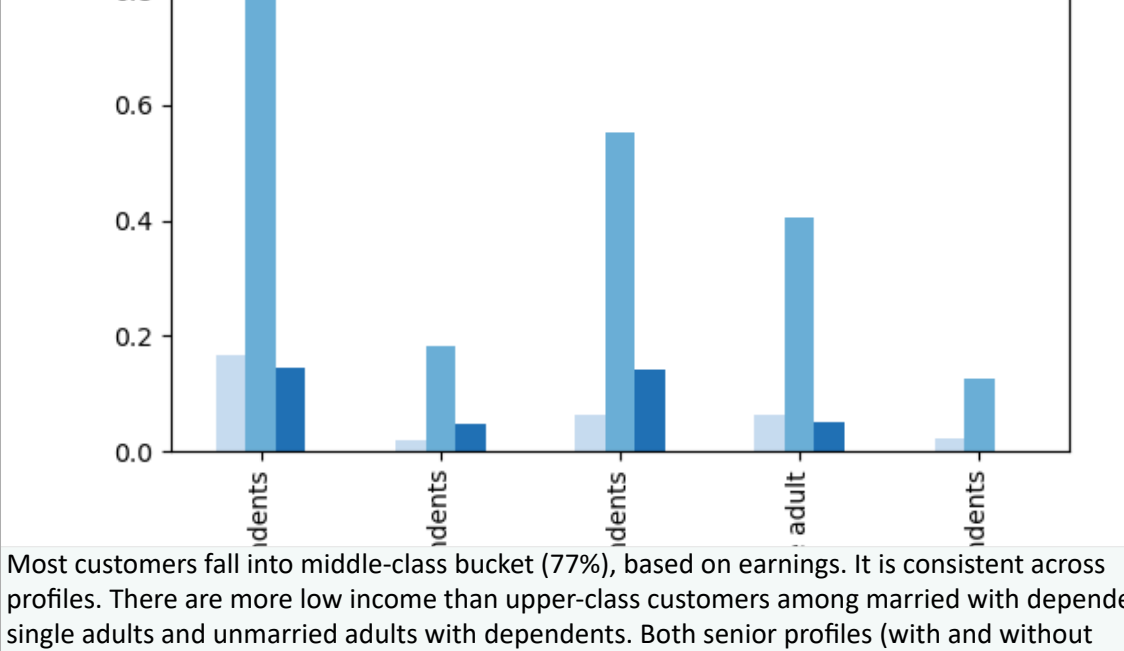


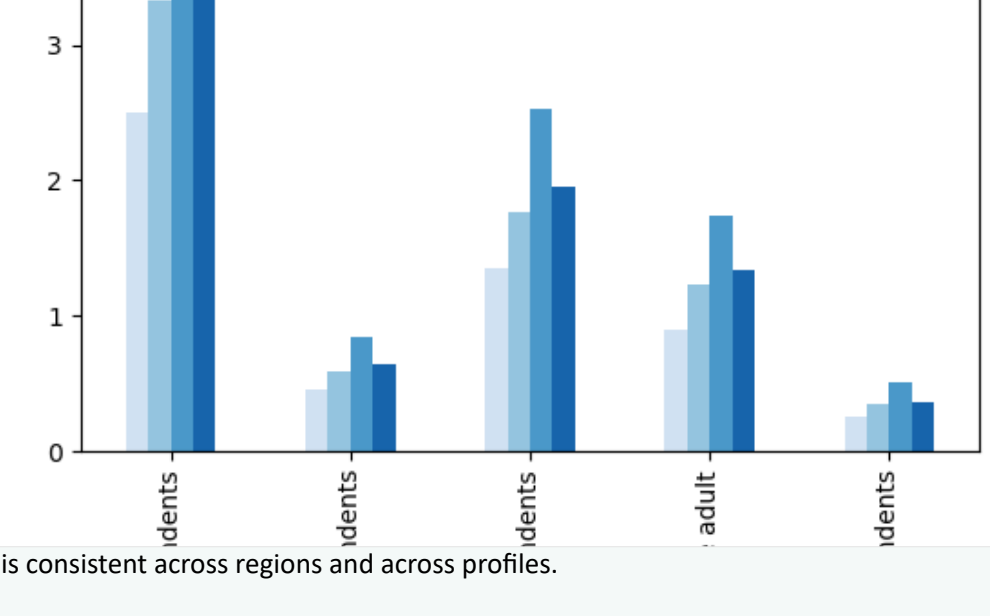




Based on age, number of dependents, and family status, I profiled customers into 5 groups:
* Married with dependents: 18-60, have dependents, married.
* Senior with dependents: 61+, have dependents.
* Single adult: 18-60, no dependents.
* Senior no dependents: 61+, no dependents.
* Unmarried with dependents: 18-60, have dependents, living with parents and siblings.
Separately, I grouped customers by earning power into Low earners, Middle class and Upper class (mentioned in previous question)

Overall distribution of profiles looks closest to middle_class earners (making $50-150k), as it is by far the lasgest group. Profile distribution among low earners is still relatively close to that of middle class and overall (percentage of seniors is lower, compared to other income levels and overall population, and percentage of unmarried with dependants is higher in this group). Upper class profile distribution differs most from the other two: seniors (both with and without dependents) make up a larger piece (half of all upper class customers are seniors, compared to only a quarter of low income pie), while all the other profile groups are smaller (almost no unmarried with dependents).
I also cheched profile distribution for each reagion separately, and it is almost the same across regions.
Most customers are low spenders (average order price under 10 dollars).





Most customers fall into middle-class bucket (77%), based on earnings. It is consistent across profiles. There are more low income than upper-class customers among married with dependents, single adults and unmarried adults with dependents. Both senior profiles (with and without dependents) have more upper-class customers than low earners.

Number of orders (customers) is consistent across regions and across profiles.

5f. 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.

	prices			
	mean	max	min	sum
profile				
Married with dependents	7.787682	25.0	1.0	110288074.4
Senior no dependents	7.799175	25.0	1.0	19752706.6
Senior with dependents	7.795131	25.0	1.0	59074137.4
Single adult	7.784867	25.0	1.0	40526786.1
Unmarried with dependents	7.800604	25.0	1.0	11552531.1

	prices			
	mean	max	min	sum
loyalty_flag				
Loyal customer	7.773575	25.0	1.0	79933825.2
New customer	7.801665	25.0	1.0	37469804.2
Regular customer	7.798262	25.0	1.0	123790606.2

	prices		
	mean	max	min
department_id			
1	7.733725	15.0	1.0
2	6.961995	15.0	1.1
3	7.857482	15.0	1.0
4	7.980348	15.0	1.0
5	8.150699	15.0	1.0
6	7.678797	15.0	1.0
7	7.682556	15.0	1.0
8	7.887489	15.0	1.0
9	7.349370	15.0	1.0
10	8.346919	14.1	1.4
11	7.998744	15.0	1.0
12	16.300034	25.0	8.0
13	8.014482	20.0	1.0
14	8.031647	14.9	1.0
15	7.546660	15.0	1.0
16	8.336372	15.0	1.0
17	7.379951	15.0	1.0
18	7.634399	15.0	1.0
19	4.275555	7.0	1.6
20	7.783347	15.0	1.0
21	8.660655	15.0	1.0

	days_since_prior_order		
	mean	max	min
profile			
Married with dependents	10.827289	30.0	0.0
Senior no dependents	10.741025	30.0	0.0
Senior with dependents	10.808885	30.0	0.0
Single adult	10.816736	30.0	0.0
Unmarried with dependents	10.750995	30.0	0.0

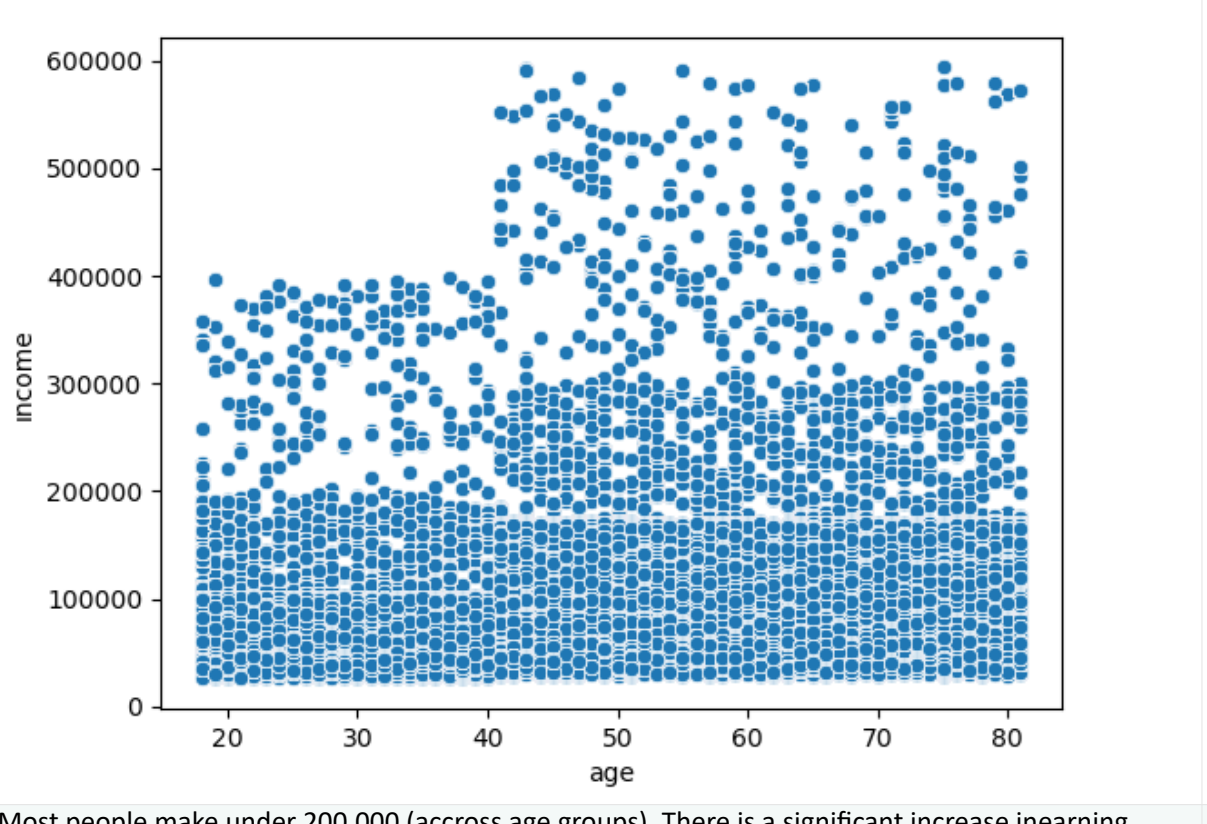
Single adults spend less per order on average thank other profile groups, while unmarried with dependents spend most. The difference is in cents (1.5 to be exact), so not significant.

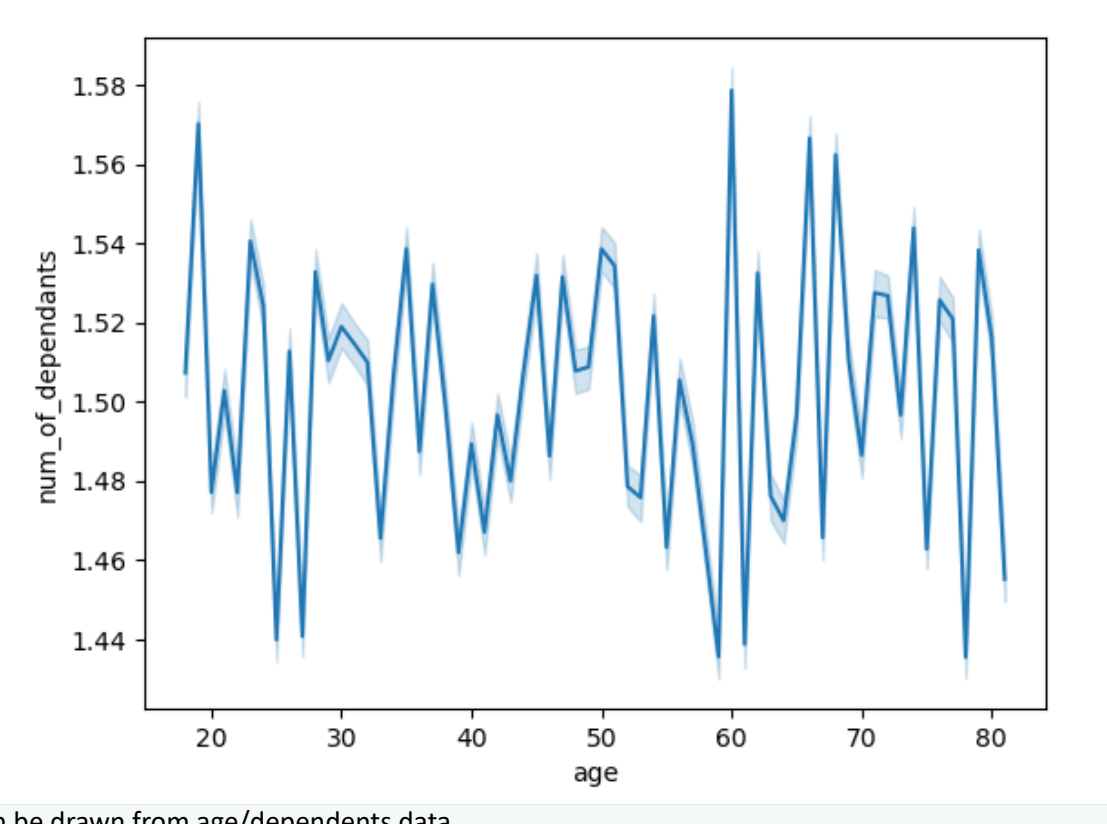
Loyal customers pay slightly less on average than both New and Regular customers: 7.77 versus 7.8.

Seniors with dependents shop most frequently, while least frequent customer profile group is married with dependents. Having said that, there is really no difference between the profile groups, as it is less than a day on average. Max is capped at to days, so less frequent shoppers all fall into that 30d since prior order bucket, skewing real results here.

I checked a number of other relationships, with no significant differences, one exception being orders by earnings power. As covered previously in 5d, low earners order more from snacks department. Average, as well as max order price for this department is the lowest of all 21 departments.

Graphs not used to answer key questions:





Most people make under 200,000 (accross age groups). There is a significant increase inearning power at around 40. No one under 40 makes over 400,000, and at the same time many more people make between 200,000 and 300,000 once they reach 40+.

No significant insights can be drawn from age/dependents data.

instacart

Recommendations

Key Question 1: 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.

The busiest days are Saturday and Sunday and the busiest time of the day is between 10am and 5pm. Customers place fewest number of orders on Tuesdays and Wednesdays, so these days might be best for ramping up ads. In terms of hours of the day with fewest orders, those are predictably between midnight and 7am. Since most customers sleep during that time, ads would not be helpful. Instead, it could be done sometime between 5pm and 10pm.

Key Question 2: 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.

Customers spend between 7.75 and 7.85 on average. Fridays and Saturdays seem to be on the higher end of this range, so that could be a good time for marketing more expensive items. Some customers buy slightly more expensive items in the “off-peak” shopping hours between 4-6am, however, this is also a time, when relatively few orders are placed (even though at higher prices). Orders on the lower end of the range fall between 9-10am and on Mondays in terms of days of the week.

Key Question 3: 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.

A simpler price grouping can be as follow:
* Low-range products: <5 dollars
* Mid-range products: 5 to 15 dollars
* High-range products: >15 dollars

Currently, most products ordered fall into mid-range bucket and very few high-range.

Key Question 4: 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.

Produce department has the highest frequency of product orders, followed by dairy & eggs, snacks, beverages and frozen. The most popular products are bananas, followed by organic bananas, organic strawberries, organic baby spinach and organic Hass avocados. Customers would appreciate discount reminders on their favorite items.

Key Question 5: What’s the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?

While each individual Loyal customer placed most orders (>40 each), the largest number of overall orders was placed by regular customers (roughly 16M out of a total of 32.4M), followed by Loyal and only then - New. Regular and new customers should be targeted by ads and discount offers to make them loyal customers as well.

Key Question 6: Are there differences in ordering habits based on a customer’s loyalty status?

Loyal customers shop often, but pay on average less than regular or new customers. New customers tend to pay most per order. The difference new customers pay is most pronounced on Sundays and Mondays. There seems to be no significant difference in department popularity rankings between different loyalty groups.

Key Question 7: Are there differences in ordering habits based on a customer’s region?

In terms of popular departments, the shopping patterns are consistent across regions.
In terms of day of week, all regions spend most on Fridays and Saturdays and average price of orders on those days is very similar (7.86). Midweek, the days when customers spend least on average vary very slightly. Two days with lowest average price are as follow:
* Northeast: Monday and Tuesday,
* Midwest: Monday and Wednesday,
* South: Sunday and Monday,
* West: Wednesday and Thursday.

In terms of time of the day, while the pattern varies slightly, the lowest average price paid is around 9-10am across all regions. Night shoppers pay most on average, peaking between 2am and 6am. The lowest price is around 7.74 across regions, the highest differs by a couple of cents only from 7.85 in the South to 7.89 in the West.

K**ey Question 8:** Is there a connection between age and family status in terms of ordering habits?

Buying habits in terms of department popularity/rankings is consistent across age and family status. Earnings power seems to have an effect on shopping habits by department: low earners order relatively less produce and more snacks (19), dairy & eggs (16), and beverages (7), than either middle-class or upper-class customers. Average, as well as max order price for snack department is the lowest of all 21 departments.

Key question 9: What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?

Age:
70% adults (18-60)
30% of customers are seniors (61+)

Profile:
46% married with dependents
24% seniors with dependents
17% single adults
8% seniors without dependents
5% unmarried with dependents

Dependents:
75% have dependents
25% no dependents

Earnings power:
77% middle-class ($50-150)
13% upper-class (over $150k)
11% low earners (under $50k)

Regionally:
33% South
26% West
23% Midwest
18% Northeast

Loyalty:
49% Regular
32% Loyal
19% New

Predominantly (98%) **low spenders** (average order price <10 dollars)

Key question 10: 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.

No significant difference across customer profiles. New customers spend slightly more, than loyal and low earners prefer snacks more than others, but other than that, spending habits are quite consistent across customers and regions.