[Instacart Basket Analysis]

Project Name: Instacart Basket Analysis

Date: 18 June 2022

Analyst Name: Elsa Ekevall

Contents:

Objective:

Instacart - an online grocery store that operates through an app - already has very good sales, but they want to uncover more information about their sales patterns. This project is tasked with performing an initial data and exploratory analysis of some of their data in order to derive insights and suggest strategies for better segmentation based on the provided criteria.

Context:

The Instacart stakeholders are most interested in the variety of customers in their database along with their purchasing behaviors. They assume they can't target everyone using the same methods, and they're considering a targeted marketing strategy. They want to target different customers with applicable marketing campaigns to see whether they have an effect on the sale of their products. This analysis will inform what this strategy might look like to ensure Instacart targets the right customer profiles with the appropriate products. The stakeholders would like to be able to answer the following key questions:

Key Questions:

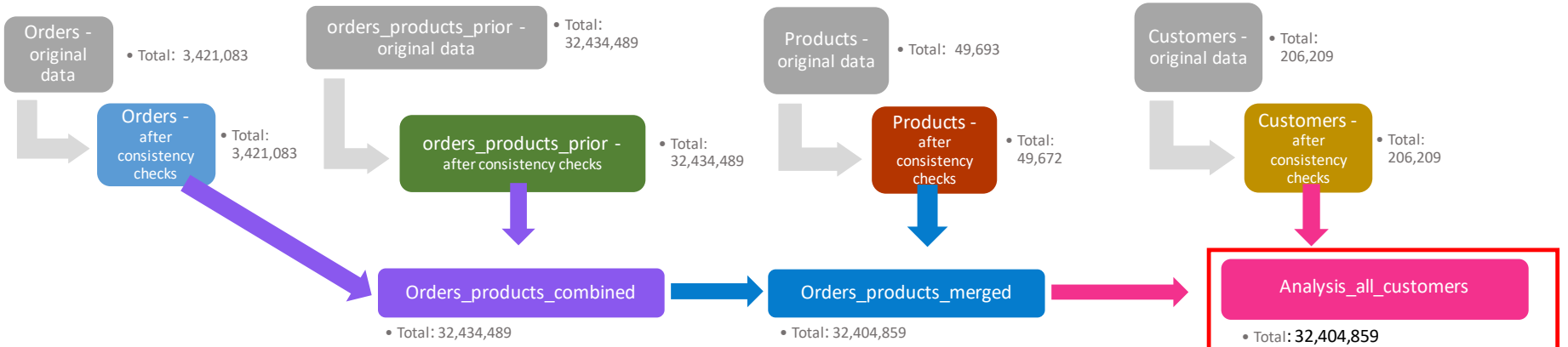
- The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.
- They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.
- 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.
- 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.
- 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)?*
  - *Are there differences in ordering habits based on a customer's loyalty status?*
  - *Are there differences in ordering habits based on a customer's region?*
  - *Is there a connection between age and family status in terms of ordering habits?*
  - *What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?*

Data Sources:

The Instacart Online Grocery Shopping Dataset 2017	
"The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 18 Jun 2022.	
Files Downloaded:	*Date:*
departments.csv	04-Feb-20
orders_products_prior.csv	01-May-17
orders.csv	02-May-17
products.csv	10 Feb 202
The Customers dataset created by Career Foundry	
This fictional dataset was accessed from https://s3.amazonaws.com/coach-courses-us/public/courses/data-immersion/A4/A4_Data_Assets/customers.zip on 18 June 2022.	
Files Downloaded:	*Date:*
customers.csv	08-Apr-20

instacart

Population flow



Exclusion flag

Condition: max_order < 5
Obervations to be removed: 1440295
Final total count of Analysis_active_customers: 30964564

instacart

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates	Outliers
orders	206209 in column days_since_prior_order	none - added new column first_order with flags for missing value	none	
products	16 in product_name column	dropped from new dataset df_prods_clean	5 (df_dups) dropped from new dataset df_prods_clean_no_dups	
orders_products_prior	none	n/a	none	
customers	none	n/a	none	
orders_products_merged_grouped.pkl	2076096 in column days_since_prior_order	none		values in the prices column above $100 changed to NaN *
	5 in customer_median_prior_order_days	none		

* The price data was not available at the time of the analysis.

instacart

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
Original Data file: orders.csv			
eval_set			not required for this analysis
	dow - order_dow to orders_day_of_week		"dow" not self-explanatory
		order_id	changed from integer to string
		user_id	changed from integer to string
		days_since_prior_order	changed from float to integer
Original Data file: products.csv			
		product_id	changed from integer to string
		aisle_id	changed from integer to string
		department_id	changed from integer to string
Original Data file: orders_products_prior.csv			
		order_id	changed from integer to string
		product_id	changed from integer to string
		reordered	changed from integer to boolean
Original Data file: customers.csv			
First Name			contains sensitive data not required for this analysis
Surnam			contains sensitive data not required for this analysis
	Gender - gender		other columns not capitalised
	STATE - US_state		other columns not capitalised and add country for clatification
	Age - age		other columns not capitalised
	n_dependents - no_dependents		added no. for clarification
	fam_status - family_status		added family for clarification
		user_id	changed from integer to string
		gender	changed from string to category
		US_state	changed from string to category
		date_joined	changed from string to datetime
		no_dependents	changed from integer to category
		family_status	changed from string to category
Merged Data file: orders_products_merged_grouped.pkl			
_merge			no longer required
		price_range_loc	changed from string to category
		busiest_day	changed from string to category
		Busiest_days	changed from string to category
		busiest_period_of_the_day	changed from string to category
		loyalty_flag	changed from string to category
		spending_flag	changed from string to category
		frequency_flag	changed from string to category

instacart

Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
orders.csv	first_order	days_since_prior_order	NaN values in the days_since_prior_order column labelled True in this column and all other values labelled false
orders_products_merged.pkl	price_range_loc	prices	# Find the high range products in the full dataframe df_ords_prods_merge.loc[df_ords_prods_merge['prices'] > 15, 'price_range_loc'] = 'High-range product' # Find the mid range products in the full dataframe df_ords_prods_merge.loc[(df_ords_prods_merge['prices'] <= 15) & (df_ords_prods_merge['prices'] > 5), 'price_range_loc'] = 'Mid-range product' # Find the low range products in the full dataframe df_ords_prods_merge.loc[df_ords_prods_merge['prices'] <= 5, 'price_range_loc'] = 'Low-range product'

orders_products_merged.pkl	busiest_day	orders_day_of_week	# For-loop to find the busiest day of the week
			# create empty list for results
			result = []
			# loop through the orders_days_of_week columna and if 0 append "Busiest day", if 4 append "Least day", otherwise append "Regularly busy",
			for value in df_ords_prods_merge["orders_day_of_week"]:
			  if value == 0:
			  result.append("Busiest day")
			  elif value == 4:
			  result.append("Least busy")
			  else:
			  result.append("Regularly busy")

orders_products_merged.pkl	Busiest_days	orders_day_of_week	# For-loop to find the Busiest days of the week
			# create new empty list for results
			result1 = []
			# loop through the orders_days_of_week columna and if 0 append "Busiest day", if 4 append "Least day", otherwise append "Regularly busy",
			for value in df_ords_prods_merge["orders_day_of_week"]:
			  if value == 0 or value == 1:
			  result1.append("Busiest days")
			  elif value == 4 or value == 3:
			  result1.append("Least busy days")
			  else:
			  result1.append("Regular days")

orders_products_merged.pkl	busiest_period_of_the_day	order_hour_of_day	# For-loop to find the Busiest periods of the day
			# create new empty list for results
			result2 = []
			# loop through the orders_days_of_week columna and if 0 append "Busiest day", if 4 append "Least day", otherwise append "Regularly busy",
			for value in df_ords_prods_merge['order_hour_of_day']:
			  if 9<= value < 18:
			  result2.append('Most orders')
			  elif 1<= value < 6:
			  result2.append('Fewest orders')
			  else:
			  result2.append('Average orders')

orders_products_merged_derived.pkl	max_order	order_number grouped by user_id	# Create a new column that shows the max-order per customer df_ords_prods_merge_grp['max_order'] = df_ords_prods_merge_grp.groupby(['user_id'])['order_number'].transform(np.max)
orders_products_merged_derived.pkl	loyalty_flag	max_order	# Create a loyalty flag for the three different customer groups using the iloc function # Loyal customers with maximum orders over 40 df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['max_order'] > 40, 'loyalty_flag'] = 'Loyal customer' # Regular customers with maximum orders over 10 and less than or equal to 40 df_ords_prods_merge_grp.loc[(df_ords_prods_merge_grp['max_order'] <= 40) & (df_ords_prods_merge_grp['max_order'] > 10), 'loyalty_flag'] = 'Regular customer' # New customers with maximum orders equla to or less than 10 df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['max_order'] <= 10, 'loyalty_flag'] = 'New customer'
orders_products_merged_derived.pkl	customer_average_prices	prices grouped by user_id	# Create a new column that shows the average price per customer (user_id) rounded to two decimal places df_ords_prods_merge_grp['customer_average_prices'] = df_ords_prods_merge_grp.groupby(['user_id'])['prices'].transform(np.mean).round(2)
orders_products_merged_derived.pkl	spending_flag	customer_average_prices	# Create a spending flag for the two different customer_average_prices groups using the iloc function # Low spenders with an average prices of less than 10 df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['customer_average_prices'] < 10, 'spending_flag'] = 'Low spender' # High spenders with an average prices of 10 or above df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['customer_average_prices'] >= 10, 'spending_flag'] = 'High spender'
orders_products_merged_derived.pkl	customer_median_prior_order_days	days_since_prior_order grouped by user_id	# Create a new column that shows the median days_since_prior_order per user_id (customer) df_ords_prods_merge_grp['customer_median_prior_order_days'] = df_ords_prods_merge_grp.groupby(['user_id'])['days_since_prior_order'].transform(np.mean).round(0)

Frequencies of flags/label variables

The frequencies of flags/label variables after deriving them.

```
Mid-range product    21860860
Low-range product    10126321
High-range product    417678
Name: price_range_loc, dtype: int64
```

```
Regularly busy   22416875
Busiest day      6204182
Least busy       3783802
Name: busiest_day, dtype: int64
```

```
Regular days     12916111
Busiest days     11864412
Least busy days   7624336
Name: Busiest_days, dtype: int64
```

```
Most orders     23205725
Average orders   8821575
Fewest orders     377559
Name: busiest_period_of_the_day, dtype: int64
```

```
Regular customer   15876776
Loyal customer     10284093
New customer        6243990
Name: loyalty_flag, dtype: int64
```

```
Low spender    31769965
High spender     634894
Name: spending_flag, dtype: int64
```

orders_products_merged_derived.pkl	frequency_flag	customer_median_prior_order_days	# Create a frequency flag for the three different customer_median_prior_order_days groups using the iloc function # Non frequent customers with the median "days_since_prior_order" greater than 20 df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['customer_median_prior_order_days'] > 20, 'frequency_flag'] = 'Non frequent customer' # Regular customers with the median "days_since_prior_order" higher than 10 and lower than or equal to 20 df_ords_prods_merge_grp.loc[(df_ords_prods_merge_grp['customer_median_prior_order_days'] <= 20) & (df_ords_prods_merge_grp['customer_median_prior_order_days'] > 10), 'frequency_flag'] = 'Regular Customer' # Frequent customers with the median "days_since_prior_order" equal to or lower than 10 df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['customer_median_prior_order_days'] <= 10, 'frequency_flag'] = 'Frequent customer' # Create a frequency flag 'First time customer' for the NaN customer_median_prior_order_days group using the iloc function # First time customers with the median "days_since_prior_order" = Nan df_ords_prods_merge_grp.loc[df_ords_prods_merge_grp['customer_median_prior_order_days'].isnull(), 'frequency_flag'] = 'First time customer'	Frequent customer 17495801 Regular Customer 11812857 Non frequent customer 3096196 NaN 5 Name: frequency_flag, dtype: int64

analysis_all_customers.csv
analysis_active_customers.csv
	region	US-state	# Set up region lists based on https://simple.wikipedia.org/wiki/List_of_regions_of_the_United_States
			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']
			west = ['Idaho', 'Montana', 'Wyoming', 'Nevada', 'Utah', 'Colorado', 'Arizona', 'New Mexico', 'Alaska', 'Washington', 'Oregon', 'California', 'Hawaii']
			south = ['Delaware', 'Maryland', 'District of Columbia', 'Virginia', 'West Virginia', 'North Carolina', 'South Carolina', 'Georgia', 'Florida', 'Kentucky', 'Tennessee', 'Mississippi', 'Alabama', 'Oklahoma', 'Texas', 'Arkansas', 'Louisiana']
			
			# create empty lists
			region = []
			
			for state in df_analysis['US_state']: # for each state in the state column in the dataframe
			  if state in northeast:
			  region.append('Northeast') # if it is in the northwest list add the variable 'Northeast' to the region column
			  elif state in midwest:
			  region.append('Midwest') # or if it is in the midwest list add the variable 'Midwest' to the region column
			  elif state in west:
			  region.append('West') # or if it is in the west list add the variable 'West' to the region column
			  else:
			  region.append('South') # else add the variable 'South' to the region column
				Active customer 30964564
				Low-activity customer 1440295
				Name: exclusion_flag, dtype: int64

analysis_all_customers.csv analysis_active_customers.csv	exclusion_flag	max_order	# Using loc statement to create an exclusion flag df_analysis.loc[df_analysis['max_order'] < 5, 'exclusion_flag'] = 'Low-activity customer' df_analysis.loc[df_analysis['max_order'] >= 5, 'exclusion_flag'] = 'Active customer'	South 10791885 West 8292913 Midwest 7597325 Northeast 5722736 Name: region, dtype: int64
analysis_all_customers.csv analysis_active_customers.csv	generation_flag	age	# Create a profiling variable based on age where the age groups are defined by PEW Research centre (https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/) # 'Generation_Z' "age" 18-25 (born between 1997 and 2004) df_analysis_active.loc[(df_analysis_active['age'] >= 18) & (df_analysis_active['age'] <= 25), 'generation_flag'] = 'Generation_Z' # 'Millennial' "age" 26-41 (born between 1981 and 1996) df_analysis_active.loc[(df_analysis_active['age'] >= 26) & (df_analysis_active['age'] <= 41), 'generation_flag'] = 'Millenial' # 'Generation_X' "age" 42-57 (born between 1965 and 1980) df_analysis_active.loc[(df_analysis_active['age'] >= 42) & (df_analysis_active['age'] <= 57), 'generation_flag'] = 'Generation_X' # 'Baby_Boomer' "age" 58-76 (born between 1945 and 1964) df_analysis_active.loc[(df_analysis_active['age'] >= 58) & (df_analysis_active['age'] <= 77), 'generation_flag'] = 'Baby_Boomer' # 'Silent' "age" 77-94 (born between 1944 and 1928) df_analysis_active.loc[(df_analysis_active['age'] >= 78) & (df_analysis_active['age'] <= 94), 'generation_flag'] = 'Silent'	Baby_Boomer 9577008 Generation_X 7829801 Millenial 7735184 Generation_Z 3864518 Silent 1958053 Name: generation_flag, dtype: int64
analysis_all_customers.csv analysis_active_customers.csv	income_flag	income	# Create a profiling variable based on income (Low earner below 25 percentile, Middle earner 25 to 75 percentile, and Top earner above 75 percentile based on the dataframe income statistics in cell above) # 'Low earner' 'income' less than 6.729200e+04 df_analysis_active.loc[df_analysis_active['income'] < 6.729200e+04, 'income_flag'] = 'Low earner' # 'Middle earner''income' between 6.729200e+04 and 1.281020e+05 df_analysis_active.loc[(df_analysis_active['income'] >= 6.729200e+04) & (df_analysis_active['income'] <= 1.281020e+05), 'income_flag'] = 'Middle earner' # 'High earner' 'income' greater than 6.729200e+04 df_analysis_active.loc[df_analysis_active['income'] > 1.281020e+05, 'income_flag'] = 'High earner'	Middle earner 15482468 High earner 7741091 Low earner 7741005 Name: income_flag, dtype: int64

analysis_all_customers.csv
analysis_active_customers.csv
	diet_flag	department_id	# Function to create a variable 'diet_flag' based on goods in the "department_id" column (Vegan - no diary eggs (16) or meat seafood (12), Vegetarian - no meat seafood (12) and None)
			
			# create empty list
			diet_flag = []
			
			for department in df_analysis_active['department_id']:
			  if (department != '16' and department != '12'):
			  diet.append('Vegan')
			  elif department != '12':
			  diet.append('Vegetarian')
			  else:
			  diet.append('None')
				Vegan 25112601
				Vegetarian 5177182
				None 674781
				Name: diet_flag, dtype: int64

analysis_all_customers.csv analysis_active_customers.csv	Parent with baby profile	no_dependents and department_id	# Creating parent with baby profile df_analysis_active.loc[(df_analysis_active['no_dependents'] >= 1) & (df_analysis_active['department_id'] == '18'), 'profile'] = 'Parent with baby'	
analysis_all_customers.csv analysis_active_customers.csv	Pet parent	department_id and no_dependents	# Creating pet parent profile df_analysis_active.loc[(df_analysis_active['department_id'] == '8') & (df_analysis_active['no_dependents'] < 1), 'profile'] = 'Pet parent'	Parent older children 22917819 High earnings no children/pets 5791130 NaN 1924949 Parent with baby 307064 Pet parent 23602 Name: profile, dtype: int64
analysis_all_customers.csv analysis_active_customers.csv	Parent older children	department_id and no_dependents	# Creating parent older children profile df_analysis_active.loc[(df_analysis_active['department_id'] != '18') & (df_analysis_active['no_dependents'] >= 1), 'profile'] = 'Parent older children'	
analysis_all_customers.csv analysis_active_customers.csv	High earner no children/pets	income, department_id and no_dependents	# Creating high earner no children/pets profile df_analysis_active.loc[(df_analysis_active['income'] >= 6.729200e+04) & (df_analysis_active['department_id'] != '8') & (df_analysis_active['no_dependents'] < 1), 'profile'] = 'High earnings no children/pets'	

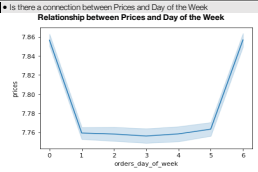
instacart

Facts and Figures

+30.96 million orders | **206,209 customers in the USA** | **49,628 products** | **21 departments**

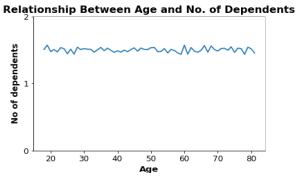
Visualisations

• Is there a connection between Prices and Day of the Week



The lineplot shows that **customer spenditure is the highest (around $7.86) on Friday (day 6) and Saturday (day 0). While during the remainder of the week the spenditure is around $7.76**

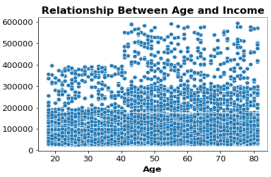
• Determine whether there's a connection between age and family situation



For all customer ages the **average number of dependents (children) ranges from 1 to 2.**

The number of dependents is evenly spread with **a quarter of the customers having either 0, 1, 2, or 3 dependents**.

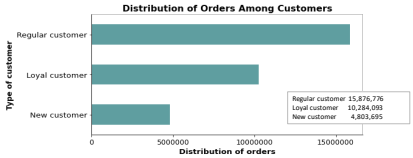
• Explore whether there's a connection between age and spending power (income)



The scatterplot shows that **most customers 40 and under earn up to 200k dollars** with an upper earnings limit of 400k dollars. While **most customers over 40 earn up to 300k dollars** with an upper earnings limit of around 600k dollars.

Analysis from this point onwards carried out only on the active customers files: analysis_active_customer and with profiles analysis_active_customer_profiles.csv

• The marketing team is curious about the distribution of orders among customers in terms of loyalty



Regular customers place the highest number of orders among the three customer groups. 51% of the orders placed are by Regular customers , 33% by Loyal customers and 16% by New customers .

• The Instacart officers are interested in comparing customer behavior in different geographic areas. Create a regional segmentation of the data. Determine whether there's a difference in spending habits between the different U.S. regions.

Spending habits based on individual user_id

Region	Spending Flag			
	High spender		Low spender	
	%	Count	%	Count
Midwest	3	1,267	97	47,252
Northeast	2	884	98	35,504
South	3	1,823	97	66,814
West	3	1,396	97	51,169

Most (33%) Instacart customers live in the South region, followed by the West and Midwest regions, while the Northeast region has the least (18%). **Although the number of customers in the regions differ, the spending habits across the regions are similar** with the proportion of high spenders (around 3%) and low spenders (around 97%).

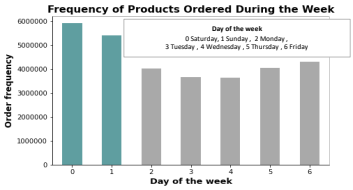
Spending habits across the whole dataset based on orders

Region	Spending Flag			
	High spender		Low spender	
	%	Count	%	Count
Midwest	2	156,154	98	7,441,171
Northeast	2	108,245	98	5,814,491
South	2	210,017	98	10,581,868
West	2	160,478	98	8,132,435

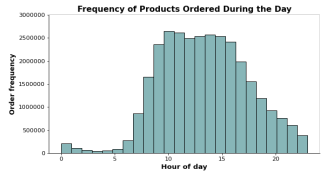
Instacart customers in the South region place the most orders, followed by the West and Midwest regions, while the Northeast region has the least. Although the number of customers in the regions differ, the spending habits based on the number of ordes placed across the regions are similar with the proportion of high spenders (around 2%) and low spenders (around 98%). Therefore **low spenders are placing the maority of the orders**

Key Questions and Answers

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.

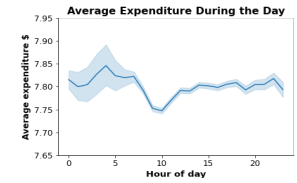


The weekend, **Saturday (day 0) and Sunday (day 1), are the busiest days of the week**. The least busy days are around the middle of the week Tuesday (day 2) and Wednesday (day 3).

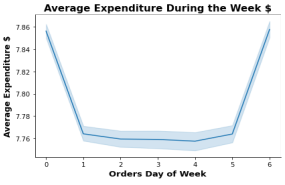


The histogram shows that **most of the orders are placed between 9 am and 4 pm** (around 2.5 million orders per hour). This peak tails off around 5 pm and there are fewer orders (below 5 million orders) between 11 pm and 6 am.

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.

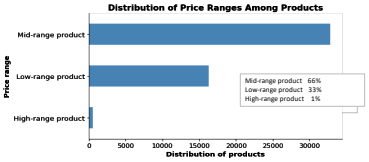


Across the week the average expenditure during the day is around 7.80 dollars. There is a slight decrease from **the high (7.85 dollars) around 4 am** to the low point (7.75 dollars) around 9 am. *[NB this chart was produced using a representative sample (70%) of the data.]*

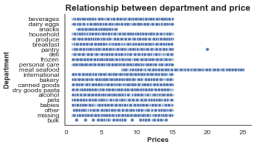


The lineplot shows that **customer spenditure is the highest (around $7.86) on Friday (day 6) and Saturday (day 0). While during the remainder of the week the spenditure is around $7.76**

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.

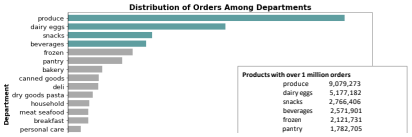


Around two thirds of the Instacart products are Mid-range products (32801) priced between $5 and $15, one third are Low-range products (16280) and only 1% High-range products (547) . *[Excluding the outlier products priced above $25.]*

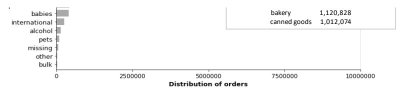


Within the deaprtments **only 'pantry' and 'meat seafood' have products above $15.** Prices of products in most departments range from $1.00 to $15.00 with the exception of the 'snacks' department $1.60 to $7.00 and bulk $1.4 to $14.

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.



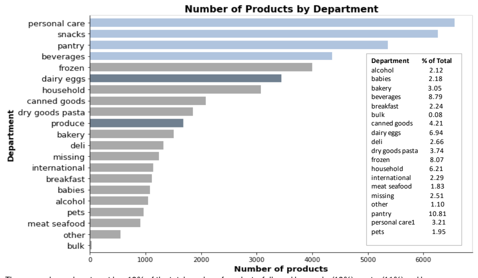
Product	% of Total Orders	Mean Price $
produce	29%	7.98
dairy eggs	17%	8.34
snacks	9%	4.28
beverages	8%	7.68
frozen	7%	7.73
pantry	6%	8.01
bakery	4%	7.86
canned goods	3%	7.55
deli	3%	7.78
dry goods pasta	3%	7.35
household	2%	7.38
meat seafood	2%	16.30



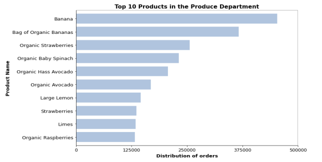
breakfast	2%	8.03
personal care	1%	8.00
babies	1%	7.63
international	1%	7.68
alcohol	0%	8.15
pets	0%	7.89
missing	0%	8.86
other	0%	6.96
bulk	0%	8.35

In descending order the **four most popular departments with over 2.5 million orders are 'produce', 'dairy eggs', 'snacks'** and **'beverages'**. Alcohol, pets, missing, other and bulk have the lowest product orders.

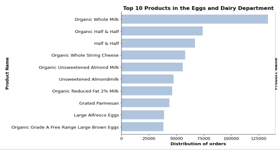
The **'snack' department has the lowest average for prices ($4.28)** and the **'meat seafood' department the highest average prices($16.30)**.

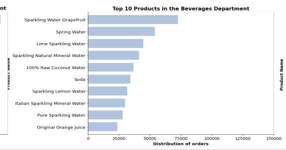


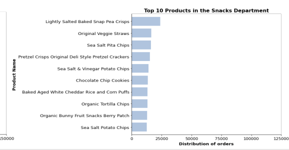
The personal care department has 13% of the total number of products, followed by snacks (13%), pantry (11%) and beverages (9%). The dairy and eggs department has 7% of the total number of products and produce has 3%. **Despite having the most products the personal care department only accounts for 1% of the total orders** compared to produce with 29% of the total orders. **Although there are more items in the personal care department they are not selling well.**



The three charts below and the chart above show the top 10 products in each of the top four departments. The **product items in the produce department are the highest selling products** with the tenth highestt (Organic Raspberries) having a similar number of orders to the top product (Organic Whole Milk) in the Eggs and Dairy department. Although overall Snacks has more orders than Beverages, the top selling product in the Snack department sells in roughly the same quamtities as the tenth highest selling product in the Beverages department. **The snacks department accounts for 9% of the total orders due to the number of products the department stocks.**

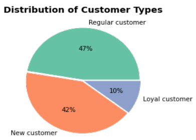






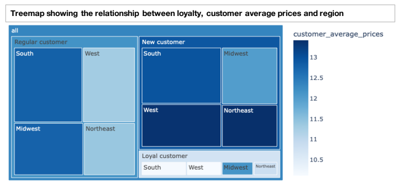
Key Question 5

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



Only 10% (17017) are Loyal customers with maximum orders over 40, while nearly half (76864) of the active customers are classified as Regular customers i.e. customers with maximum orders over 10 and less than or equal to 40. The remaining 42% (68750) are classified as new customers.

Treemap showing the relationship between loyalty, customer average prices and region

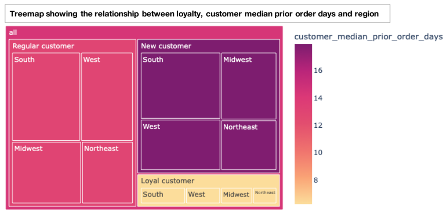


The averge expenditure for all customers is $12.33. **Loyal customers tend to spend less ($10.73) than average** , but there are regional difffernces with customers in the Midwest spending the same as all customers. New customers in the West and Northwest spend more than average.

The interactive file can be downloaded and then viewed using this link:

https://drive.google.com/file/d/1lUolEFmhMMCGY9HrRJbny5i-QdGf5ouk/view?usp=sharing

Treemap showing the relationship between loyalty, customer median prior order days and region

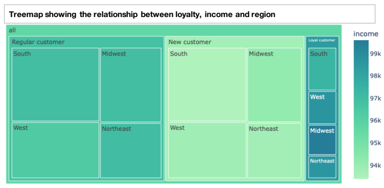


Customers in all four regions are similar with **loyal customers placing orders roughly once a week**, regular customers place orders approx. every fortnight, and new customers have around 18 days between orders.

The interactive file can be downloaded and then viewed using this link:

https://drive.google.com/file/d/1hYigpfwiger0xTIJE84ykxUO8Boz-qYx/view?usp=sharing

Treemap showing the relationship between loyalty, income and region

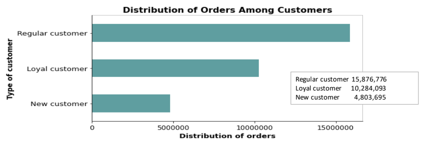


New customers tend to earn slighlty less ($93,737) than the average income for all users ($95,686) and regular customers slightly more ($96,812. **Loyal customers have the highest average income** ($98,458). Within this group there are regional differences with **Loyal customers in the south earning less** than the group average ($97,387) and **Loyal customers in the Midwest earning more** ($99,592).

The interactive file can be downloaded and then viewed using this link:

https://drive.google.com/file/d/1wbJpR5hv9fp-T4QTQnLqKRP0KD08qWmJ/view?usp=sharing

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

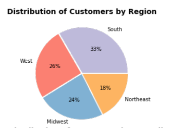


Regular customers place the highest number of orders among the three customer groups. 51% of the orders placed are by Regular customers , 33% by Loyal customers and 16% by New customers .

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

Spending habits based on individual user_id

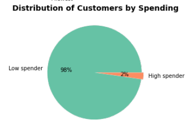
Region	Spending Flag			
	High spender		Low spender	
	%	Count	%	Count
Midwest	3	885	97	37,491
Northeast	2	614	98	27,967
South	3	1,264	97	52,929
West	3	948	97	40,533



Most (33%) Instacart customers live in the South region, followed by the West and Midwest regions, while the Northeast region has the least (18%). **Although the number of customers in the regions differ, the spending habits across the regions are similar** with the proportion of high spenders (around 3%) and low spenders (around 97%).

Spending habits across the whole dataset based on orders

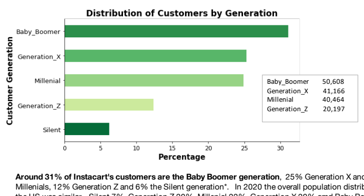
Region	Spending Flag			
	High spender		Low spender	
	%	Count	%	Count
Midwest	2	148,784	98	7,112,729
Northeast	2	103,140	98	5,361,545
South	2	199,618	98	10,111,521
West	2	152,412	98	7,774,815



Instacart customers in the South region place the most orders, followed by the West and Midwest regions, while the Northeast region has the least. Although the number of customers in the regions differ, the spending habits based on the number of ordes placed across the regions are similar with the proportion of high spenders (around 2%) and low spenders (around 98%). **Low spenders are placing more orders**

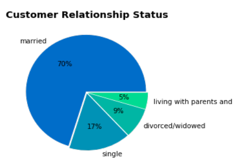
When examined in more detail by the different classsifications, such as loyalty, family status and generation, there are regional differences. However the overall percentages per region for each classification is more or less the same. This is shown by comparing the size of the region boxes relative to each other for each classifcation in the treemaps above and below.

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



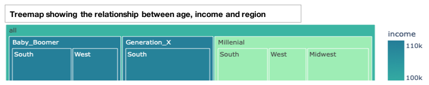
Around 31% of Instacart's customers are the Baby Boomer generation, 25% Generation X and Millenials, 12% Generation Z and 6% the Silent generation*. In 2020 the overall population distribution in the US was similar - Silent 7%, Generation Z 20%, Millenial 22%, Generation X 20% amd Baby Boomer 21% *[Ref: https://www.statista.com/statistics/296974/us-population-share-by-generation/ * Under 100% due to rounding]*

** The Silent generation were born between 1928 and 1944, the Baby Boomer generation were born between 1945 and 1964, Generation X were born between 1965 and 1980, Millenials were born between 1981 and 1996, and Generation Z were born between 1997 and 2004.*

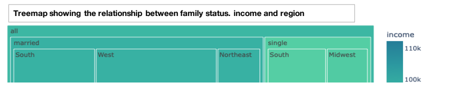


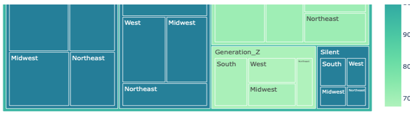
The majority (70%) of Instacart customers are married (114296), 17% are single (26,896), 9% are divorced/widowed (13,831 and 5% living with parents and siblings (7608).

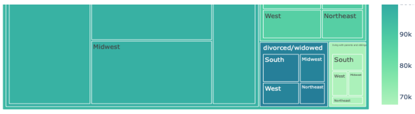
Treemap showing the relationship between age, income and region



Treemap showing the relationship between family status. income and region







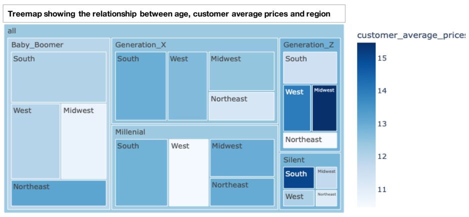
Generation X ($110, 927), Baby Boomer ($111,077) and Silent ($110,910) customer groups earn more than the average income for all Instacart customers ($95,686). **Millenial ($70, 843) and Generation Z ($68,142) customer groups earn less. Within these groups there are no regional differences.**

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1-rf45dTvJGH7_N-Z8UznqPS6qAeHjQ9D/view?usp=sharing

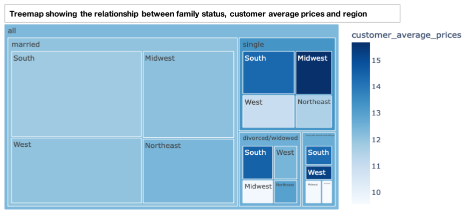
Divorced/widowed customers ($111, 202) earn more than the average income for all Instacart customers ($95,686). The average income of the Married group ($97, 537) is close to the average forfor all Instacart customers, the **Single customer group ($87,584) and Living with parents and siblings group ($68,310) earn less. Within these groups there are no regional differences.**

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1cCEqrFQmRI9QAI73vqELLeJlpSq9d1cl/view?usp=sharing

Treemap showing the relationship between age, customer average prices and region



Treemap showing the relationship between family status, customer average prices and region



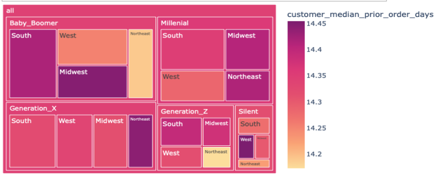
All groups tend towards the average customer spenditure of $12.33. **There are regional differences within all the groups, but they do not follow a pattern.** Generation Z customers in Midwest on average spend the most $15.46, yet have one of the lowest incomes, and Silent customers in the South spend $15.05. Generation Z customers in the Northeast ($10.45) and Millenials ($10.55) in the West spend the least.

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1IwvmU1KOjPqbHI1RWrxXn6Y8y6mH2AhG/view?usp=sharing

The average customer spenditure is $12.33 and **on avaerage all the groups spend around this amount with the exception of the single group who spend slighlty more ($13.31)**. Within the single, divorced/widowed and living with parents and siblings groups there are regional diffferences with customers in the South consistently spending more than average. In the single group customers in the Midwest spend the most ($15.67), however customers in this regiont spend well below average around $9.70 in the divorced/widowed and living with parents and siblings groups.

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/16TILQazS3dW-5oGcAGVNVHx7N8azmkLv/view?usp=sharing

Treemap showing the relationship between age, customer median prior order days and region



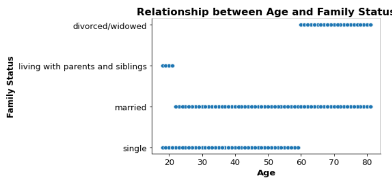
Treemap showing the relationship between family status, customer median prior order days and region

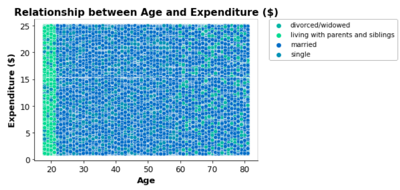
All groups tend towards the median number of prior order days - 2 weeks between orders. The regional differences within the groups are less than a day.

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/16fG_pU6sNFmWkzBrlbQqfrLTZsaqn8H4/view?usp=sharing

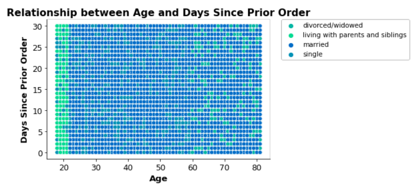
All groups tend towards the median number of prior order days - 2 weeks between orders. The regional differences within the groups are less than a day.

The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1sYvPFYOqK5VzsUHCLd4RBoavYVahOjlv/view?usp=sharing





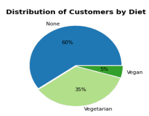
There are clear age ranges within the family status groups with the married group having the largest age range and the living with parents and siblings group the smallest age range. The scatterplots above and below show that **no relationship exists between age and the maximum order number per customer,** or **age and the number of days since prior orde**r, or **age and expenditure**. In all three cases the data shows the same parttern with the age and family status groups spanning the whole range of these three variables.



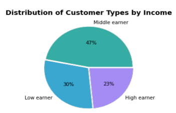
• What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?

Some of the other classification distributions, such as loyalty, family status, and generation have been analysed above.

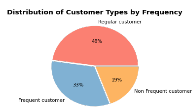
Diet



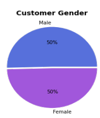
Income

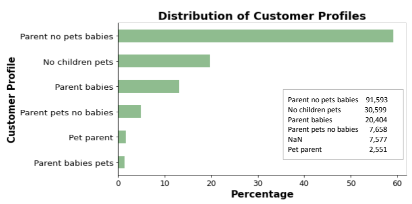


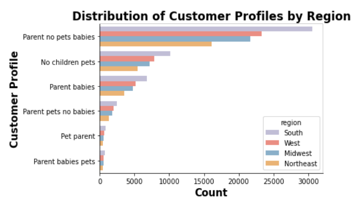
Frequency



Gender







The table shows that there are no regional differences in the percentage of customers with each profile.

The total number and percentage of customers classified by profile in the four regions

Region	Profile											
	No children pets		Parent babies		Parent babies pets		Parent no pets babies		Parent pets no babies		Pet parent	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Midwest	7,150	20	4,817	13	544	1	21,618	59	1,831	5	572	2
Northeast	5,441	20	3,552	13	419	1	16,062	59	1,293	5	460	2
South	10,187	20	6,828	13	729	1	30,594	59	2,505	5	854	2
West	7,821	20	5,207	13	557	1	23,319	59	2,029	5	665	2

• 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 maximum, mean, and minimum values on a customer-profile level for usage frequency (order number) and expenditure (prices). With the exception of diet_flag and order number for loyalty_flag (which is derived from order_number) and profile, most of the cusomer classifications have similar aggregations. The approximations for the majority of the classfications are shown below.

Maximum, minimum and mean values for order number

Order Number	max	mean	min
	99	18	1

Maximum, minimum and mean values for prices

Prices	max	mean	min
	25	8	1

Maximum, minimum and mean values for profile order number

Profile	Order Number		
	max	mean	min
No children pets	99	15.87	1
Parent babies	99	21.41	1
Parent babies pets	99	23.15	1
Parent no pets babies	99	15.83	1
Parent pets no babies	99	17.63	1
Pet parent	99	17.54	1

The profile groups 'parents with babies and pets' and 'parents with babies' place more orders. While the profile groups 'parents without pets and babies' place fewer orders.

Maximum, minimum and mean values for loyalty flag order number

Loyalty flag	Order Number		
	max	mean	min
Loyal customer	99	32.64	1
New customer	10	4.27	1
Regular customer	40	12.35	1

The loyalty flag was derived from orders and the conditiion can be seen in the maximum and mean order amounts.

Maximum, minimum and mean values for diet flag order number

Diet flag	Order Number		
	max	mean	min
None	99	18.58	1

Maximum, minimum and mean prices for diet flag order number

Diet flag	Prices		
	max	mean	min
None	25	7.9	1

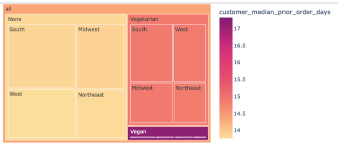
Vegan	99	11.43	1
Vegetarian	99	15.83	1

Vegan	20	6.96	1
Vegetarian	20	7.49	1

The diet flag group 'Vegan' place the lowest number of orders (11) and spend less ($7) compared to the other groups. The vegetarian group also place less orders (16) compared to other groups.

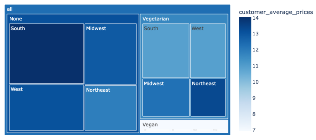
Diet and income are the two main classifications that immediately stand out.

Treemap showing the relationship between diet, customer median prior order days and region



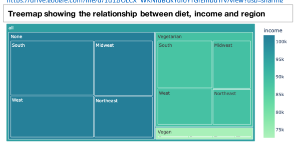
The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/15HgjhDFFKl6tdKX12BeqN-xY2iG_s_iH/view?usp=sharing

Treemap showing the relationship between diet, customer average prices and region



The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1d1ziOLCX_WKNiqBQkYuIoYTGiEmbdTrv/view?usp=sharing

Treemap showing the relationship between diet, income and region



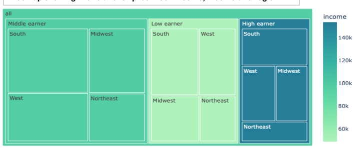
The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1BDZSe9RHXlse3nqdmXD115mWYZOs2L_6/view?usp=sharing

Diet

In all three charts there are clear differences between the diet classifcation groups. The **Vegan group on average wait more than 2 weeks (17 days) to place an order, they also spend less ($7) and earn less** ($73, 518) than the other groups in this classification. The vegetarian group earns ($88,682) less than the average income ($95,686).

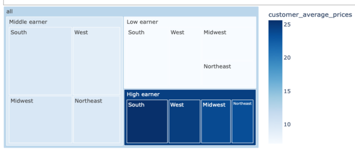
The diet flag classification was derived by grouping customers based on which departments they ordered from. The vegan group excludes the 'meat seafood and 'dairy eggs' department, and the Vegetarian group only excludes the 'meat seafood' department. The Vegan group excludes one of the most popular deaprtments 'dairy eggs' and the one with the highest prices 'meat seafood'.

Treemap showing the relationship between income, income and region



The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1opUtIlycD9UAQSOxDQ_MLZGASLpceo6f/view?usp=sharing

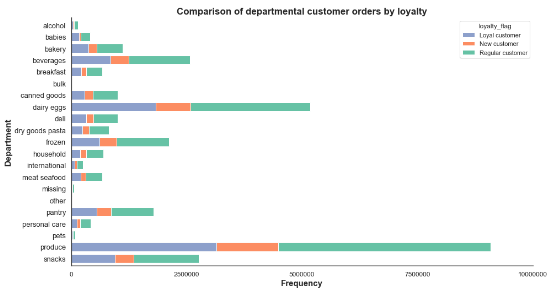
Treemap showing the relationship between income, customer average prices and region



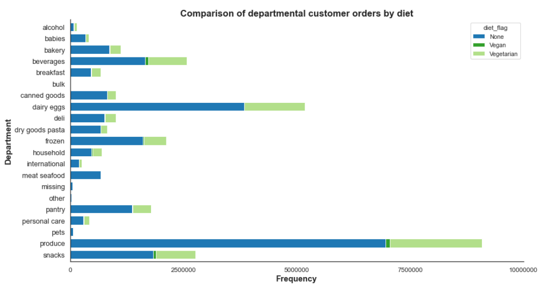
The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/16nzOS0GD_GxdYWGAjiZEkE0v7gI3QZsQ/view?usp=sharing

Income

As expected High earners earn ($153,204) more than average ($95,686) and low earners ean less ($48,748). **High earners spend nearly double ($24.59) what the average customer would spend** ($12.33), while middle earners spend around a quarter less ($9.55) and low earners around a third less ($7.05). There is no difference in the order period, with all three groups placing orders around every two weeks, and no regional differences.

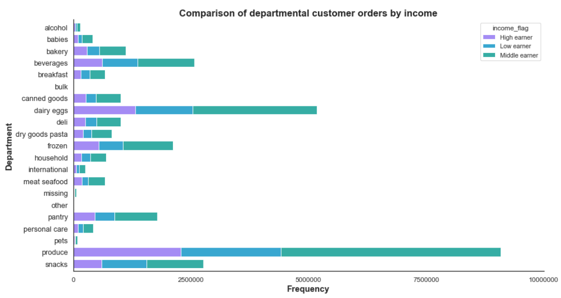


No noticeable differences in departmental ordering habits by loyalty. Customers spend similar percentages in each department.



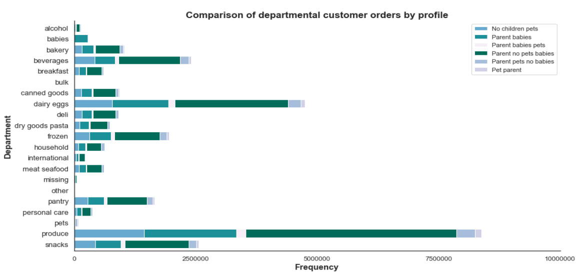
The percentage differences in ordering habits by the three diet groups are shown in the table below. The departments with the most noitceable differences are highlighted.

Department	None	Vegan	Vegetarian
alcohol	0.35%	3.43%	0.70%
babies	1.47%	0.54%	0.93%
bakery	3.76%	1.66%	3.29%
beverages	7.19%	19.78%	11.23%
breakfast	2.00%	2.13%	2.67%
bulk	0.10%	0.19%	0.14%
canned goods	3.53%	1.53%	2.53%
dairy eggs	16.65%	0.00%	17.61%
deli	3.29%	1.68%	3.16%
dry goods pasta	2.91%	0.79%	1.96%
frozen	6.98%	5.53%	6.50%
household	2.07%	6.25%	2.69%
international	0.86%	0.38%	0.74%
meat seafood	2.93%	0.00%	0.00%
missing	0.20%	0.18%	0.23%
other	0.10%	0.24%	0.13%
pantry	5.92%	4.27%	5.31%
personal care	1.26%	2.60%	1.67%
pets	0.30%	0.29%	0.29%
produce	30.17%	31.29%	26.68%
snacks	7.97%	17.24%	11.53%
	100.00%	*100.00%*	*100.00%*



Low earners order a higher percentage of snacks than the other groups in this classification.

Department	High earner	Low earner	Middle earner
snacks	8%	12%	8%



No noticeable differences in departmental ordering habits between customer profiles. Customers spend similar percentages in each department.

instacart

Title page

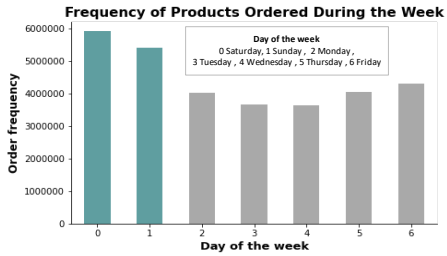
Recommendations

Reccommendations based on the analysis carried out only on the active customers files: analysis_active_customer and with profiles analysis_active_customer_profiles.csv

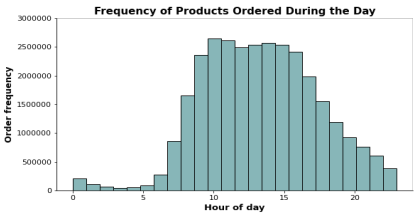
Further visualisations and analysis that informed the key questions can be found here

Key Questions and 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.



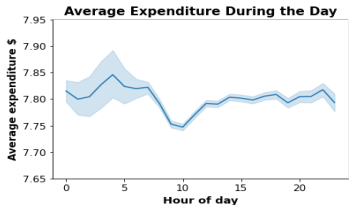
The weekend, **Saturday (day 0) and Sunday (day 1), are the busiest days of the week**. the least busy days are around the middle of the week Tuesday (day 2) and Wednesday (day 3).



The histogram shows that **most of the orders are placed between 9 am and 4 pm** (around 2.5 million orders per hour). This peak tails off around 5 pm and there are fewer orders (below 5 million orders) between 11 pm and 6 am.

Recommendation: Tuesday and Wednesday are the least busy days and there are fewer orders (below 1.5 million) in the period between 6 pm and 9 am, which could be considered a good time to schedule ads. The adverts will reach more customers between 6 pm and 12 am.

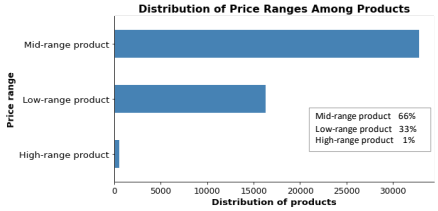
Key Question 2 • They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.



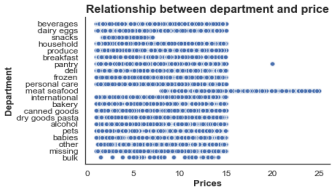
The average expenditure during the day is around 7.80 dollars. There is a slight decrease from **the high (7.85 dollars) around 4 am** to the low point (7.75 dollars) around 9 am. *[NB this chart was produced using a representative sample (70%) of the data.]*

Recommendation: Customers spend slightly more money $7.85 (as opposed to $7.80) around 4 am. However the thicker light blue band shows there is also more uncertainty around this estimate with the range varying from around $7.79 to $7.89. It might be worthwhile investigating the reason for the dip between 7 am and 12 pm, where there is very littke uncertainty, and to target adds to increase spending during this period.

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.



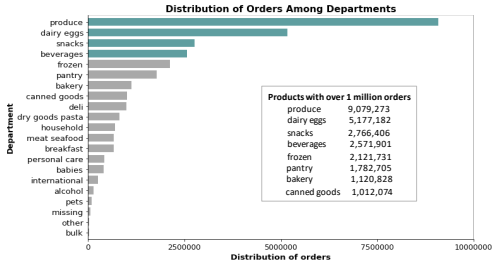
Around two thirds of the Instacart products are Mid-range products (32801) priced between $5 and $15, one third are Low-range products (16280) and only 1% High-range products (547) . *[Excluding the outlier products priced above $25.]*



Within the deaprtments **only 'pantry' and 'meat seafood' have products above $15**. Prices of products in most departments range from $1.00 to $15.00 with the exception of the 'snacks' department $1.60 to $7.00 and bulk $1.4 to $14.

Recommendation: Only 1% of the products are above $15 and they are mainly in the 'meat seafood' department, while around two thirds are Mid-range products. Where possible increase the maximum price in more departments and increase the number of products in the High-range product group.

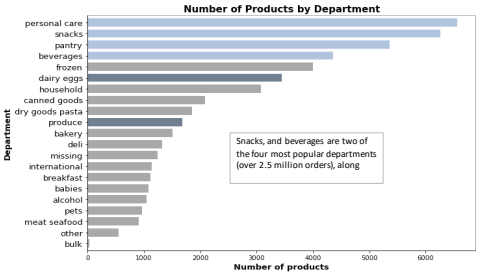
Key Question 4 • Are there certain types of products that are more popular than others? The marketing and sales teams want to know which departments have the highest frequency of product orders.



In descending order **the four most popular departments with over 2.5 million orders are produce, dairy eggs, snacks and beverages**. Alcohol, pets, missing, other and bulk have the lowest product orders.

Product	Percentage	Mean Price $	
produce	29%	7.98	
dairy eggs	17%	8.34	
snacks	9%	4.28	lowest
beverages	8%	7.68	
frozen	7%	7.73	
pantry	6%	8.01	
bakery	4%	7.86	
canned goods	3%	7.55	
deli	3%	7.78	
dry goods pasta	3%	7.35	
household	2%	7.38	
meat seafood	2%	16.30	highest
breakfast	2%	8.03	
personal care	1%	8.00	
babies	1%	7.63	
international	1%	7.68	
alcohol	0%	8.15	
pets	0%	7.89	
missing	0%	8.86	
other	0%	6.96	
bulk	0%	8.35	

Recommendation: The four departments with the highest number of orders are 'produce' (29%), 'dairy eggs' (17%), 'snacks' (9%) and 'beverages' (8%). The 'snacks' department has a mean price of $4.28 and the other three departments around $8.00. The 'meat seafood' department with a mean price of $16.00 only accounts for 2% of the total orders. As recommended above, where possible increase prices, especially in the top departments and also the frequency that customers purchase from the other departments.

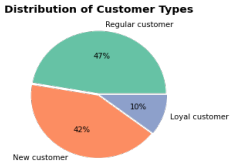


The personal care department has 13% of the total number of products, followed by snacks (13%), pantry (11%) and beverages (9%). The dairy and eggs department has 7% of the total number of products and produce has 3%. **Despite having the most products the personal care department only accounts for 1% of the total orders** compared to produce with 29% of the total orders. **The items in the personal care department are not as popular.**

Recommendation: Look at increasing orders or reducing the number of items stocked in departments such as personal care, where orders are low, but there is a large range of products.

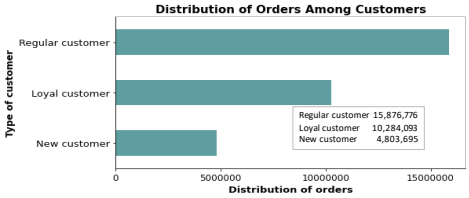
Key Question 5

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



Only 10% (17017) of the active customers are Loyal customers with maximum orders over 40, while nearly half (76864) are classified as Regular customers i.e. customers with maximum orders over 10 and less than or equal to 40. The remaining 42% (68750) are classified as New customers.

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



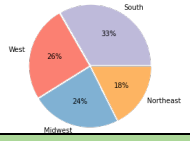
Regular customers place the highest number of orders among the three customer groups. 51% of the orders placed are by Regular customers, 33% by Loyal customers and 16% by New customers .

Recommendation: Although Loyal customers are only 10% of the total number of customers they account for one third of the orders placed. Therefore an analysis on the profiles of these customers and their ordering habits is recommended and then to target similar customers in the Regular and New customer groups with the aim of converting them to Loyal customers.

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

Spending habits based on individual user_id

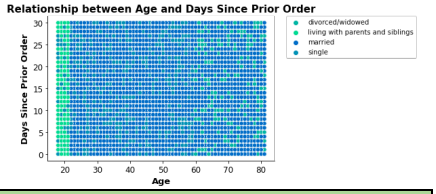
Region	Spending Flag			
	High spender		Low spender	
	%	Count	%	Count
Midwest	3	885	97	37,491
Northeas	2	614	98	27,967
South	3	1,264	97	52,929
West	3	948	97	40,533



Most (33%) Instacart customers live in the South region, followed by the West and Midwest regions, while the Northeast region has the least (18%). **Although the number of customers in the regions differ, the spending habits across the regions are similar** with the proportion of high spenders (around 3%) and low spenders (around 97%).

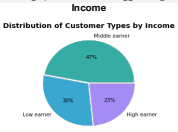
When examined in more detail by the different classsifications, such as loyalty, family status and generation, there are some regional differences. However similar to spending the overall percentages per region for each classification are more or less the same.

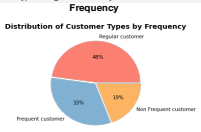
- Is there a connection between age and family status in terms of ordering habits?

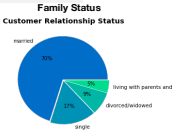


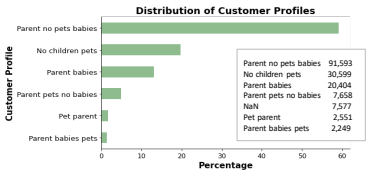
There are clear age ranges within the family status groups with the married group having the largest age range and the living with parents and siblings group the smallest age range. **No relationship exists between age and the maximum order number per customer** or **age and the number of days since prior order**.

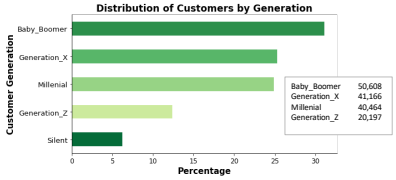
- What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?





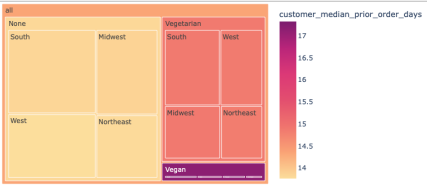






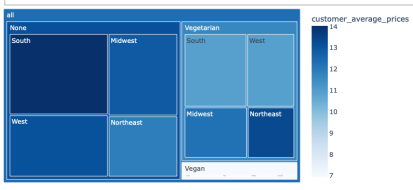
- What differences can you find in ordering habits of different customer profiles? Consider the price of orders, the frequency of orders, the products customers are ordering, and anything else you can think of.

Treemap showing the relationship between diet, customer median prior order days and region



The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/15HgjhDFFKI6tdKX12BeqN-xY2iG_s_iH/view?usp=sharing

Treemap showing the relationship between diet, customer average prices and region



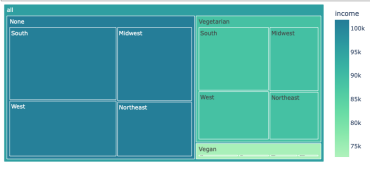
The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1d1ziOLCX_WKNjqBQkYuloYTGiEmbdTry/view?usp=sharing

Diet

In all three charts there are clear differences between the diet classifcation groups. The **Vegan group on average wait more than 2 weeks (17 days) to place an order, they also spend less ($7) and earn less** ($73, 518) than the other grops in this classification. The vegetarian group earns ($88,682) less than the average income $95,686.

The diet flag classification was derived by grouping customers based on which departments they ordered from. The vegan group excludes the 'meat seafood and 'dairy eggs' department, and the Vegetarian group only excludes the 'meat seafood' department. The Vegan group excludes one of the most popular deaprtments 'dairy eggs' and the one with the highest prices 'meat seafood'.

Treemap showing the relationship between diet, income and region



The interactive file can be downloaded and then viewed using this link:
https://drive.google.com/file/d/1BDZ5e9RHXIse3nqdmXD115mWYZOs2L_6/view?usp=sharing

Some of the questions raised: Which groups are online when - target specific groups? Which items are not selling at all or in very low numbers - why? What do the loyal customers have in common? Are some departments, such as personal care, overstocked with similar items? What customer classifications would help marketing?

department	No children pets	Parent babies	Parent babies pets	Parent no pets babies	Parent pets no babies	Pet parent
alcohol	26030	10971	4056	76580	16193	4643
babies	0	280085	26979	0	0	0
bakery	165799	238671	34684	497625	68453	23262
beverages	427509	417482	69966	1268880	170232	54429
breakfast	102166	138617	19912	310026	35863	11984
bulk	5487	6803	699	16992	957	296
canned goods	150797	208689	30773	458299	63293	21018
dairy eggs	778821	1159385	135849	2329901	257671	85386
deli	157539	201224	25988	467243	55997	18803
dry goods pasta	114659	192937	25484	350039	49380	16584
frozen	317332	446236	66637	934751	144759	46150
household	93245	133264	32022	294474	68739	21603
international	39724	48986	7404	119357	15783	5179
meat seafood	100867	141449	18654	307649	41151	13366
missing	9859	13648	1924	29521	3740	1155
other	4897	7056	1324	14802	2515	787
pantry	276276	344024	50757	832824	111196	36656
personal care	58454	84892	19177	176704	37328	12659
pets	0	0	17416	0	52042	17280
produce	1444096	1895319	193135	4334468	380910	129225
snacks	439359	525112	76472	1322639	151745	50209

department	None	Vegan	Vegetarian
alcohol	80863	10624	53140
babies	338021	1682	70689
bakery	866170	5135	249523
beverages	1658656	61321	851924
breakfast	461316	6591	202943
bulk	22225	580	10646
canned goods	815089	4738	192247
dairy eggs	3840766	0	1336416
deli	758555	5211	240068
dry goods pasta	670580	2455	149101
frozen	1611110	17155	493466
household	476404	19389	204064
international	198740	1178	56073
meat seafood	674781	0	0
missing	46989	556	17223
other	23646	747	10018
pantry	1366262	13230	403213
personal care	289736	8073	126497
pets	69897	894	22269
produce	6957738	97004	2024531
snacks	1837788	53433	875185

department	High earner	Low earner	Middle earner
alcohol	40862	36825	66940
babies	105674	83835	220883
bakery	290423	264802	565603
beverages	618517	743159	1210225
breakfast	162877	187541	320432
bulk	7541	9400	16510
canned goods	265861	222464	523749
dairy eggs	1314008	1229206	2633968
deli	251016	248143	504675
dry goods pasta	212921	179984	429231
frozen	541072	511039	1069620
household	177257	183419	339181
international	65340	57726	132925
meat seafood	182880	127575	364326
missing	16334	15938	32496
other	8711	8573	17127
pantry	457174	417879	907652
personal care	106120	107752	210434
pets	24944	21751	46365
produce	2289572	2124737	4664964
snacks	601987	959257	1205162

department	Loyal customer	New customer	Regular customer
alcohol	38723	28629	77275
babies	167108	46708	196576
bakery	379873	172104	568851
beverages	855017	396649	1320235
breakfast	221577	106179	343094
bulk	13834	4149	15468
canned goods	294594	180196	537284
dairy eggs	1830707	750056	2596419
deli	324046	158335	521453
dry goods pasta	242580	144836	434720
frozen	609460	372271	1140000
household	196873	128117	374867
international	78394	43841	133756
meat seafood	202480	112464	359837
missing	18350	12713	33705
other	10715	5839	17857
pantry	560402	299576	922727
personal care	123365	75442	225499
pets	25469	15880	51711
produce	3147376	1336604	4595293
snacks	943150	413107	1410149

Row Labels	Sum of No children pets	Sum of Parent babies	Sum of Parent babies pets	Sum of Paren	Sum of Parent pe	Sum of Pet parent
alcohol	0.55%	0.17%	0.47%	0.54%	0.94%	0.81%
babies	0.00%	4.31%	3.14%	0.00%	0.00%	0.00%
bakery	3.52%	3.67%	4.04%	3.52%	3.96%	4.08%
beverages	9.07%	6.43%	8.14%	8.97%	9.85%	9.54%
breakfast	2.17%	2.13%	2.32%	2.19%	2.08%	2.10%
bulk	0.12%	0.10%	0.08%	0.12%	0.06%	0.05%
canned goods	3.20%	3.21%	3.58%	3.24%	3.66%	3.68%
dairy eggs	16.53%	17.85%	15.81%	16.47%	14.91%	14.96%
deli	3.34%	3.10%	3.02%	3.30%	3.24%	3.29%
dry goods pasta	2.43%	2.97%	2.97%	2.48%	2.86%	2.91%
frozen	6.73%	6.87%	7.75%	6.61%	8.38%	8.09%
household	1.98%	2.05%	3.73%	2.08%	3.98%	3.79%
international	0.84%	0.75%	0.86%	0.84%	0.91%	0.91%
meat seafood	2.14%	2.18%	2.17%	2.18%	2.38%	2.34%
missing	0.21%	0.21%	0.22%	0.21%	0.22%	0.20%
other	0.10%	0.11%	0.15%	0.10%	0.15%	0.14%
pantry	5.86%	5.30%	5.91%	5.89%	6.44%	6.42%
personal care	1.24%	1.31%	2.23%	1.25%	2.16%	2.22%
pets	0.00%	0.00%	2.03%	0.00%	3.01%	3.03%
produce	30.64%	29.18%	22.48%	30.65%	22.04%	22.64%
snacks	9.32%	8.09%	8.90%	9.35%	8.78%	8.80%
Grand Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Row Labels	Sum of None	Sum of Vegan	Sum of Vegetarian
alcohol	0.35%	3.43%	0.70%
babies	1.47%	0.54%	0.93%
bakery	3.76%	1.66%	3.29%
beverages	7.19%	19.78%	11.23%
breakfast	2.00%	2.13%	2.67%
bulk	0.10%	0.19%	0.14%
canned goods	3.53%	1.53%	2.53%
dairy eggs	16.65%	0.00%	17.61%
deli	3.29%	1.68%	3.16%
dry goods pasta	2.91%	0.79%	1.96%
frozen	6.98%	5.53%	6.50%
household	2.07%	6.25%	2.69%
international	0.86%	0.38%	0.74%
meat seafood	2.93%	0.00%	0.00%
missing	0.20%	0.18%	0.23%
other	0.10%	0.24%	0.13%
pantry	5.92%	4.27%	5.31%
personal care	1.26%	2.60%	1.67%
pets	0.30%	0.29%	0.29%
produce	30.17%	31.29%	26.68%
snacks	7.97%	17.24%	11.53%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Row Labels	Sum of High earner	Sum of Low earner	Sum of Middle earner
alcohol	0.53%	0.48%	0.43%
babies	1.37%	1.08%	1.43%
bakery	3.75%	3.42%	3.65%
beverages	7.99%	9.60%	7.82%
breakfast	2.10%	2.42%	2.07%
bulk	0.10%	0.12%	0.11%
canned goods	3.43%	2.87%	3.38%
dairy eggs	16.97%	15.88%	17.01%
deli	3.24%	3.21%	3.26%
dry goods pasta	2.75%	2.33%	2.77%
frozen	6.99%	6.60%	6.91%
household	2.29%	2.37%	2.19%
international	0.84%	0.75%	0.86%
meat seafood	2.36%	1.65%	2.35%
missing	0.21%	0.21%	0.21%
other	0.11%	0.11%	0.11%
pantry	5.91%	5.40%	5.86%
personal care	1.37%	1.39%	1.36%
pets	0.32%	0.28%	0.30%
produce	29.58%	27.45%	30.13%
snacks	7.78%	12.39%	7.78%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Row Labels	Sum of Loyal customer	Sum of New customer	Sum of Regular customer
alcohol	0.38%	0.60%	0.49%
babies	1.62%	0.97%	1.24%
bakery	3.69%	3.58%	3.58%
beverages	8.31%	8.26%	8.32%
breakfast	2.15%	2.21%	2.16%
bulk	0.13%	0.09%	0.10%
canned goods	2.86%	3.75%	3.38%
dairy eggs	17.80%	15.61%	16.35%
deli	3.15%	3.30%	3.28%
dry goods pasta	2.36%	3.02%	2.74%
frozen	5.93%	7.75%	7.18%
household	1.91%	2.67%	2.36%
international	0.76%	0.91%	0.84%
meat seafood	1.97%	2.34%	2.27%
missing	0.18%	0.26%	0.21%
other	0.10%	0.12%	0.11%
pantry	5.45%	6.24%	5.81%
personal care	1.20%	1.57%	1.42%
pets	0.25%	0.33%	0.33%
produce	30.60%	27.82%	28.94%
snacks	9.17%	8.60%	8.88%
Grand Total	**100.00%**	**100.00%**	**100.00%**