Instacart Grocery Basket Analysis

David Carpenter
January 2024

Agenda

1. Overview and background information
2. Visualizations
3. Recommendations
4. Code examples

Overview and Background Information

Context

Instacart is a grocery delivery service operating in the United States.

Instacart's marketing team is interested in learning more about their customer segments and purchasing behaviors with the goal of applying targeted marketing strategies to the various segments.

The analysis that follows will inform what this strategy could look like to ensure Instacart targets the right customer profiles with the appropriate products.

Objective

1. Perform an initial data and exploratory analysis.

2. Derive insights from customer data.

3. Suggest strategies for better segmentation based on discoveries made from analyzing customers’ sales patterns.

Questions to Answer

1. What are the busiest days of the week and hours of the day (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.

2. What times of the day do people spend the most money, as this might inform the type of products they advertise at these times.

3. 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.

4. What are the different types of customers and how their ordering behaviors differ? For example:

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

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

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

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

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

   f.) 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.

Process

I created a Jupyter Notebook and used Python and Pandas to perform an exploratory data analysis on a fictional data set from Instacart with approximately 31 million rows.

I cleaned the data and ensured data types were consistent in each column, removed rows or imputed values when data was missing, renamed columns for consistency and removed personally identifiable information. I then derived columns using calculations that enabled me to more easily create visualizations.

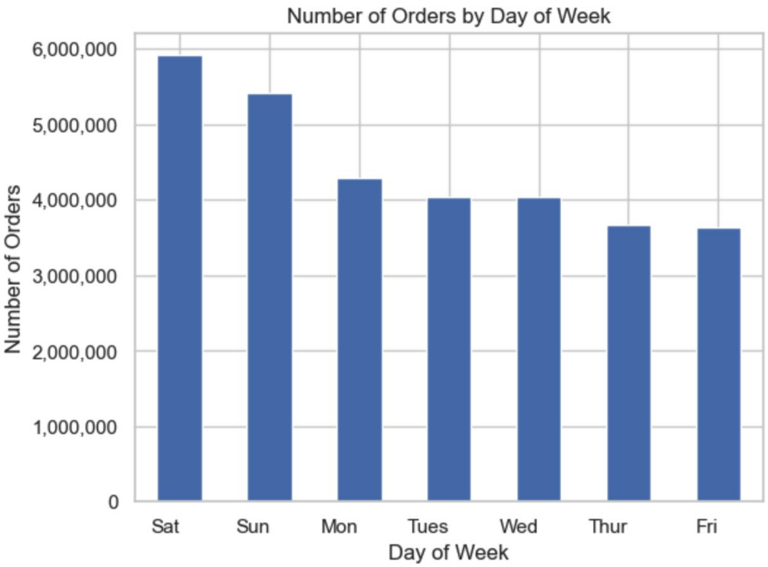
Next, I created visualizations using the pandas, numpy, os, matplotlib, seaborn and scipy libraries.

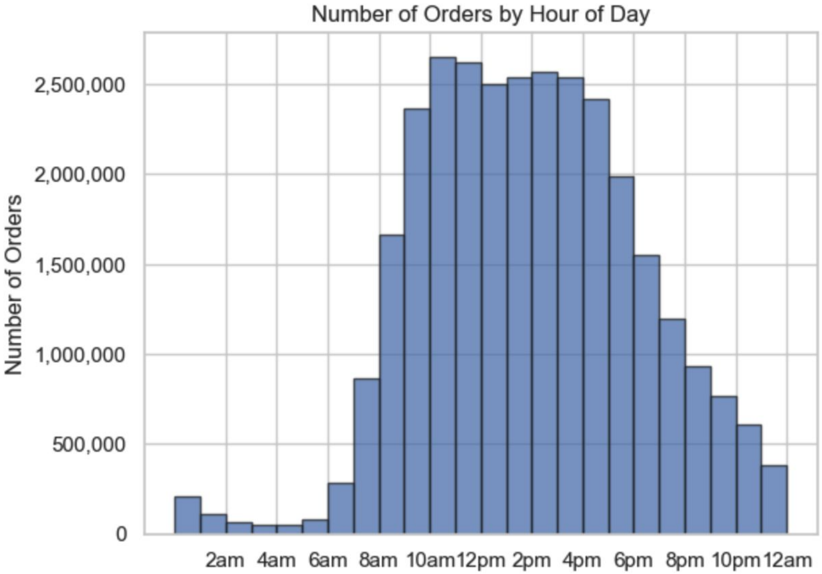
Finally, I created recommendations based on the visualizations I generated.

Visualizations

Question 1

1. What are the busiest days of the week and hours of the day (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.



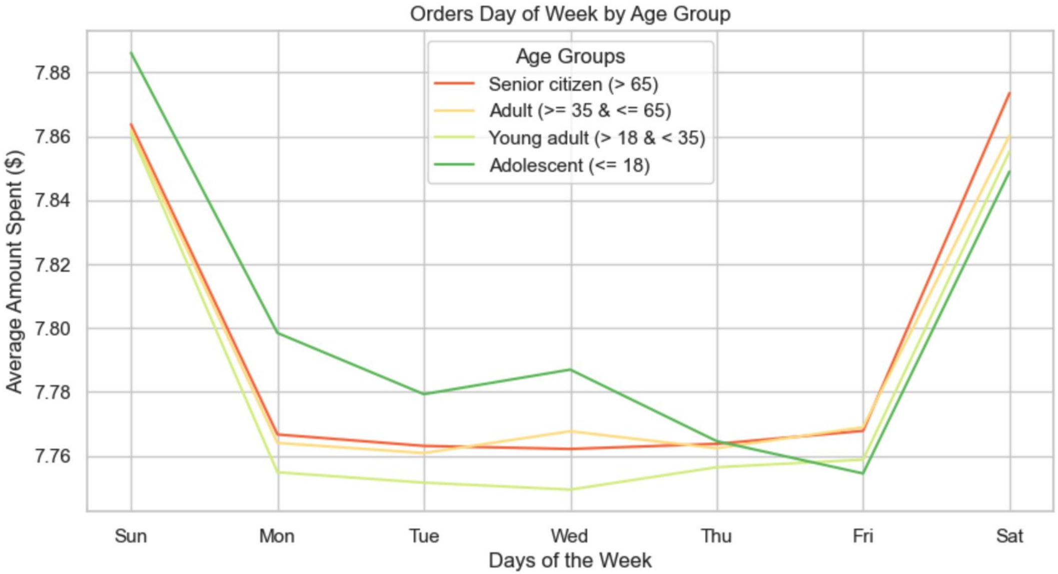


Saturday and Sunday are the busiest days. As the week progresses, fewer orders are placed.

The busiest hours of the day are generally between 10:00 a.m. and 4:00 p.m.

Question 2

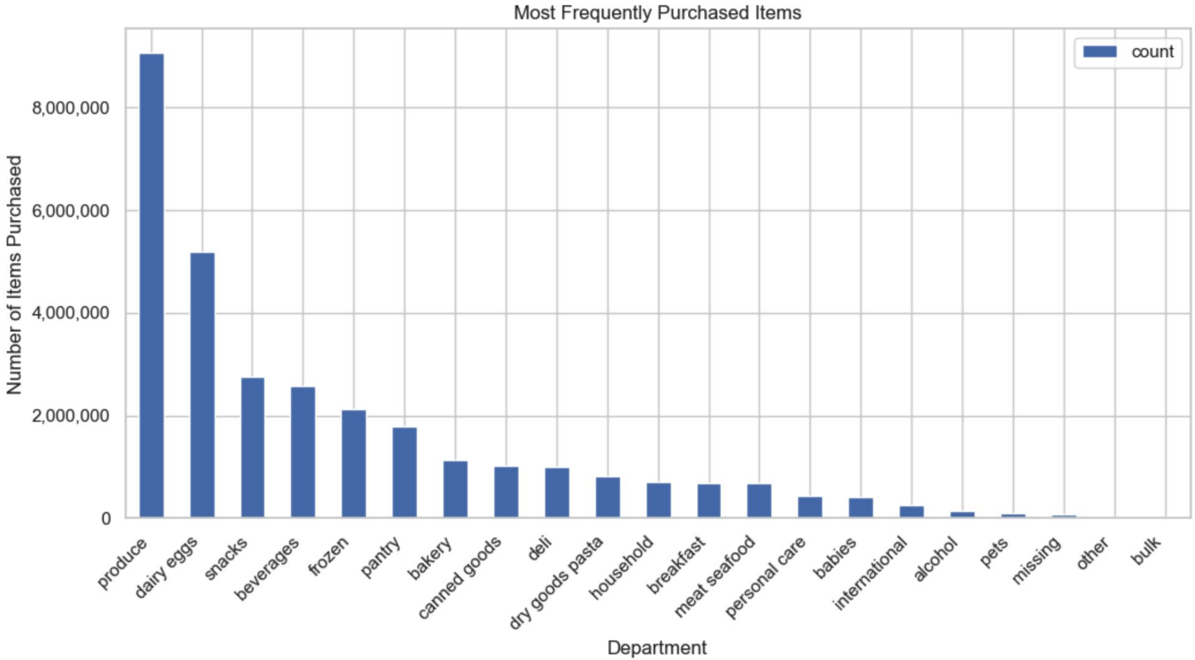
2. What times of the day do people spend the most money, as this might inform the type of products they advertise at these times.



All age groups generally spend more on Sundays and Saturdays than any other day of the week. Adolescents also spend a sizeable but comparatively small amount on Wednesdays. However, Adolescents make up a small number of Instacart's total customer base.

Question 3

3. 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.

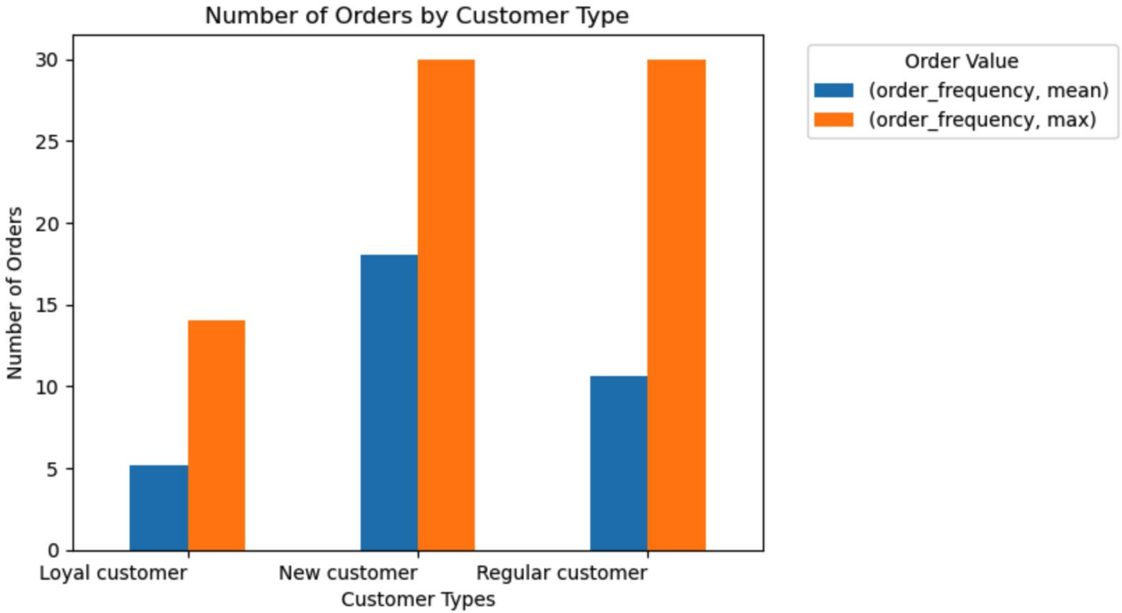


Produce, dairy and snacks are the most frequently purchased items.

Question 4a

4. What are the different types of customers and how their ordering behaviors differ? For example:

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



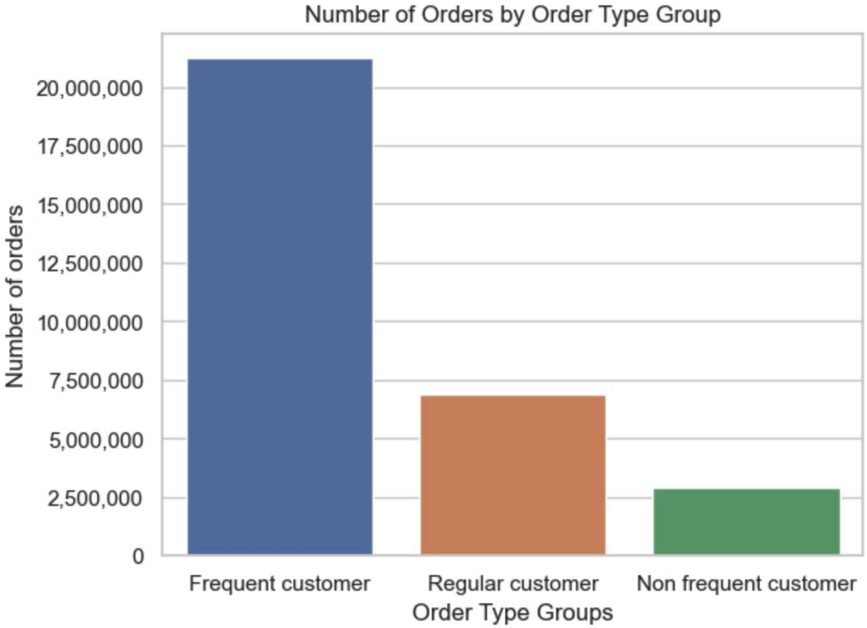
New customers generally purchase more frequently, on average than the other customer groups.

The customer groups are defined as:
- Loyal customer: > 40 orders
- Regular customer: > 10 and <= 40 orders
- New customer: <= 10 orders

Question 4b

4. What are the different types of customers and how their ordering behaviors differ? For example:

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



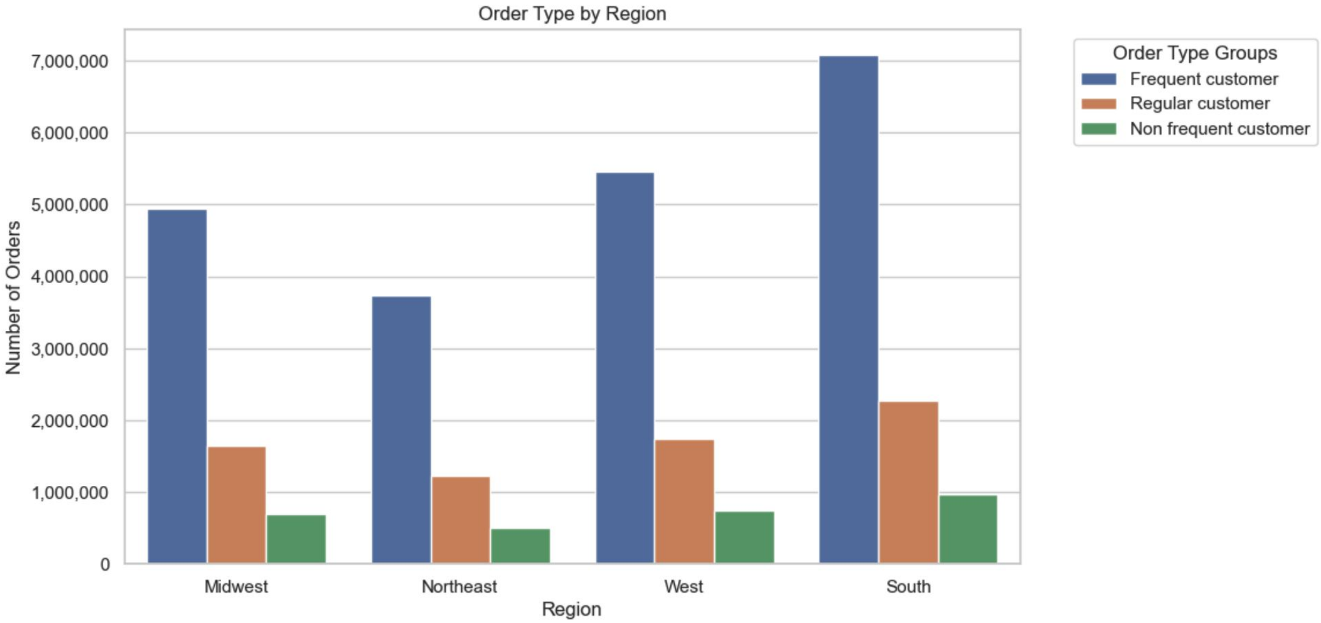
Customers labeled as "Frequent" place about three times as many orders as the "Regular" customer group and about nine times as many orders as "Non frequent" customers.

The customer groups are defined as:

- Frequent: Order frequency <= 10
- Regular: Order frequency > 10 and <= 20
- Non frequent: Order frequency > 20

Question 4c

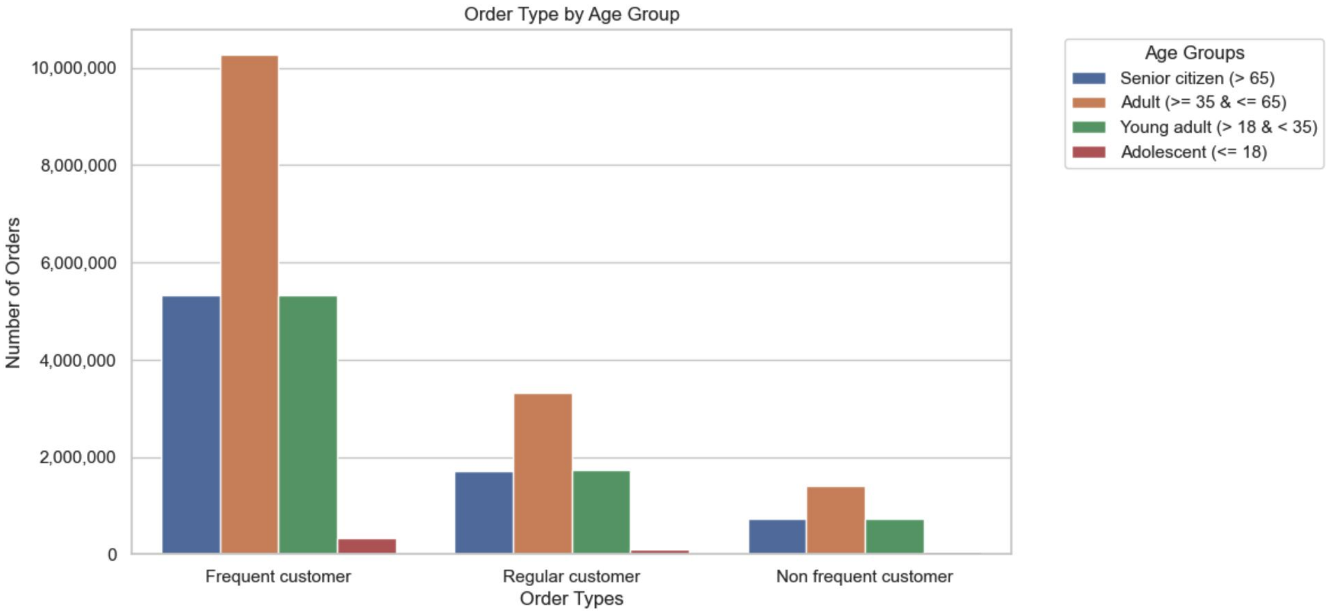
4. What are the different types of customers and how their ordering behaviors differ? For example:
   c.) Are there differences in ordering habits based on a customer's region?



There are more of each customer type in the South region of the United States.

Question 4d - Part 1 - Age

4. What are the different types of customers and how their ordering behaviors differ? For example:
    d.) Is there a connection between age and family status in terms of ordering habits?

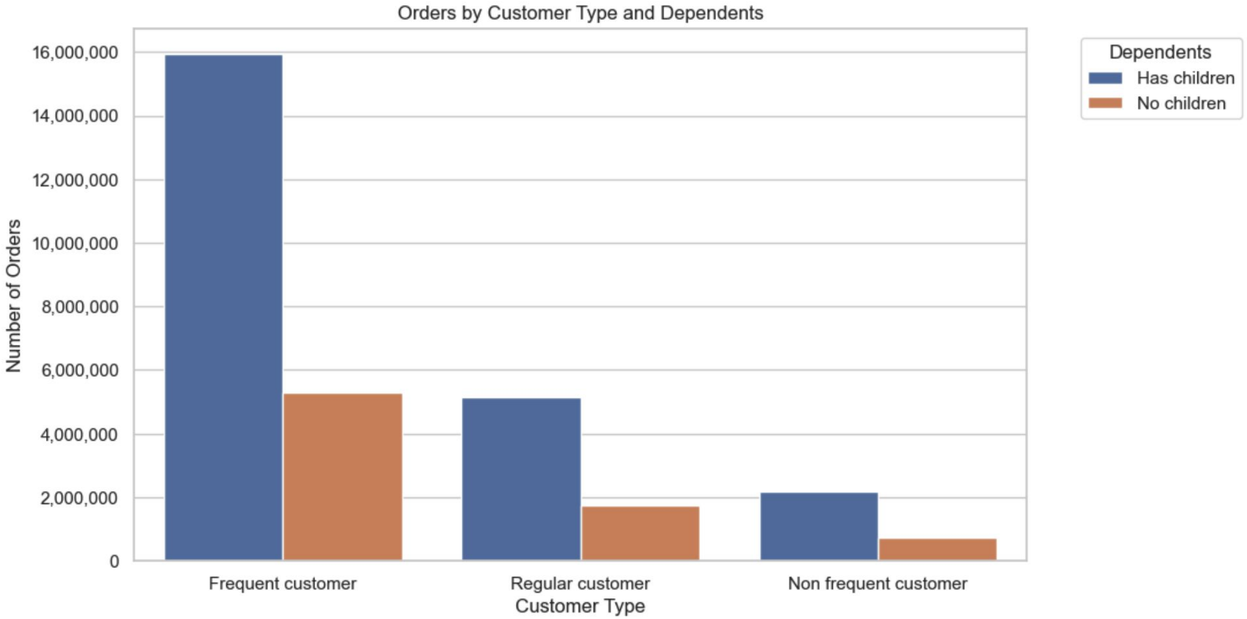


Adults account for the most number of orders for all customer types.

Question 4d - Part 2 - Family Status

4. What are the different types of customers and how their ordering behaviors differ? For example:

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

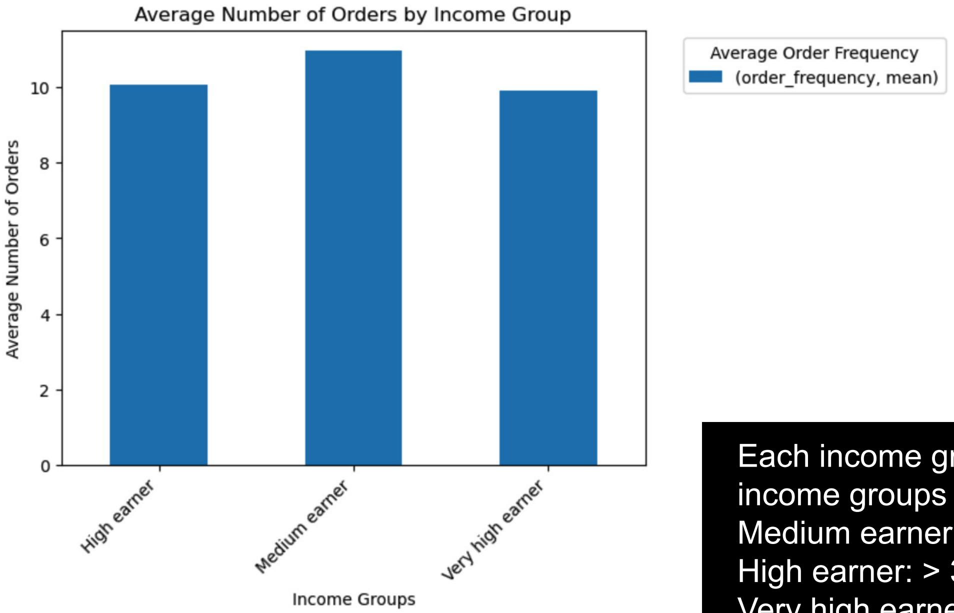


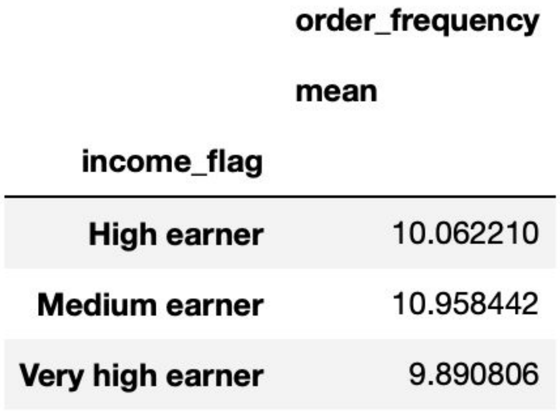
Customers with children account for the most number of orders for all customer types.

Question 4e Part 1 - Income

4. What are the different types of customers and how their ordering behaviors differ? For example:

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





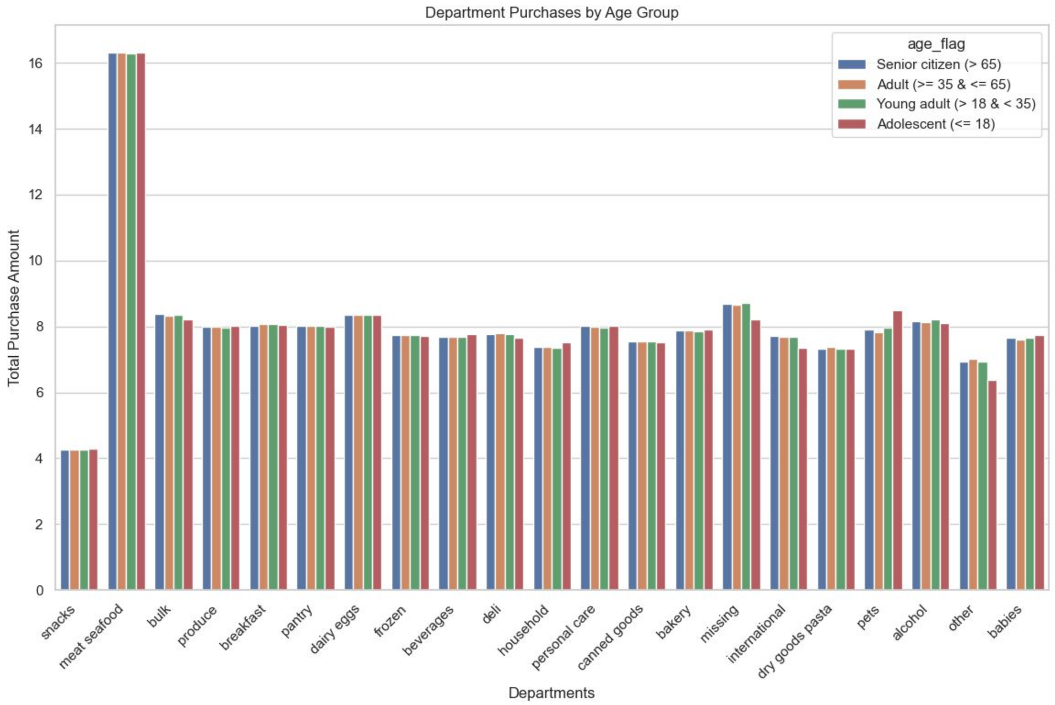
	order_frequency
	mean
income_flag	
High earner	10.062210
Medium earner	10.958442
Very high earner	9.890806

Each income group places about the same number of orders. The income groups are defined as:
Medium earner: <= 30,000
High earner: > 30,000 and <= 80,000
Very high earner: > 80,000

Question 4e - Part 2 - Types of Goods

4. What are the different types of customers and how their ordering behaviors differ? For example:

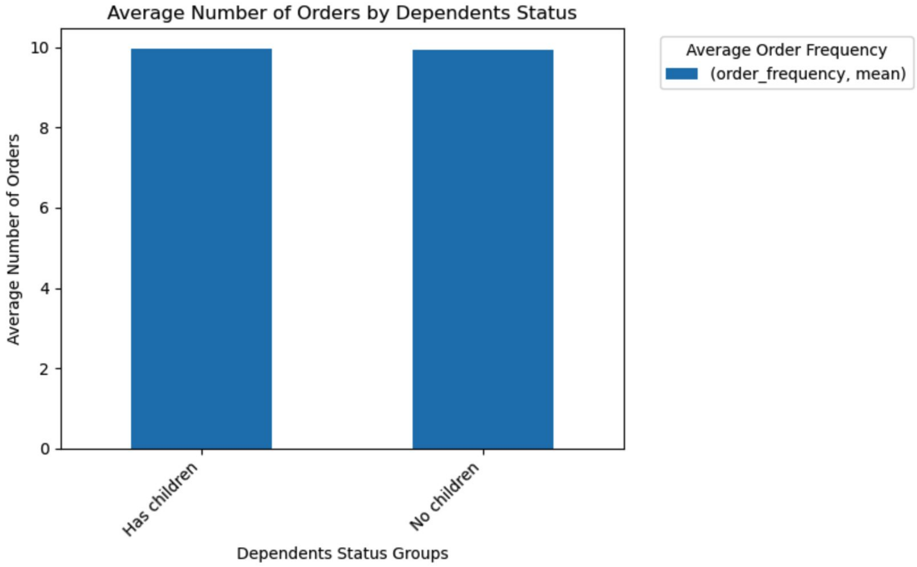
e.) What different classifications does the demographic information suggest? Income? Certain types of goods? Family status?

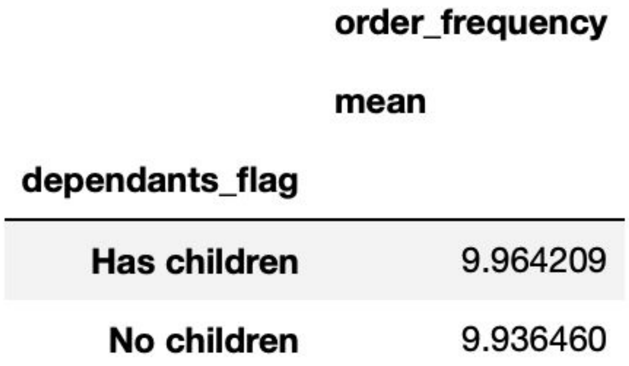


All age groups have similar buying behaviors for all types of foods.

Question 4e - Part 4 - Family Status

4. What are the different types of customers and how their ordering behaviors differ? For example:
    e.) What different classifications does the demographic information suggest? Income? Certain types of goods? Family status?





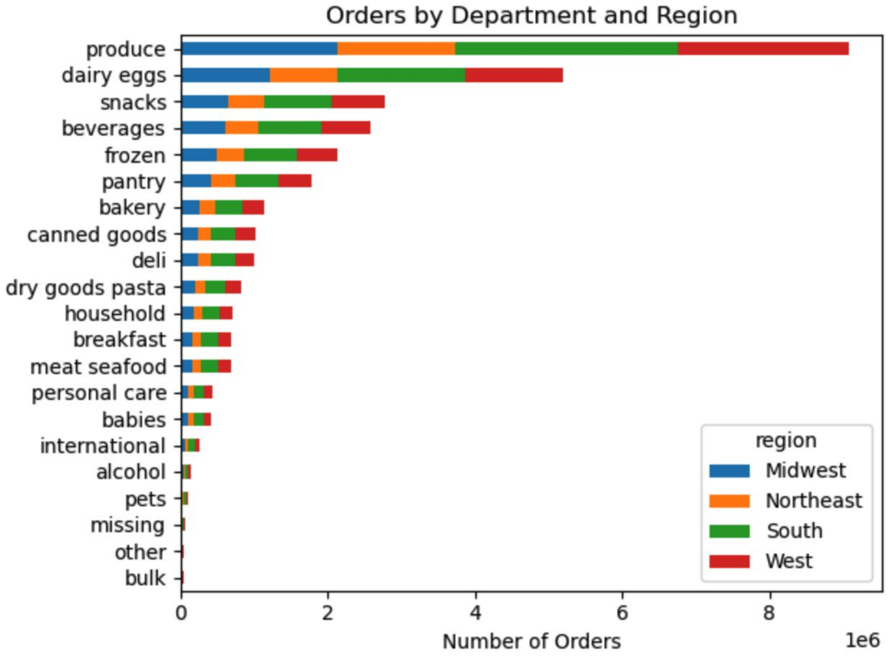
	order_frequency
	mean
dependants_flag	
Has children	9.964209
No children	9.936460

Order frequency does not differ significantly between customers who have and do not have children.

Question 4f - Departments & Regions

4. What are the different types of customers and how their ordering behaviors differ? For example:

f.) 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.



Buying behaviors from various departments does not differ significantly across geographic regions.

Recommendations

Recommendations

Adults in the "very high" earning group in the South region make up a large number of overall customers. This group primarily buy groceries on both weekend days. Categories such as produce, dairy/eggs, beverages and snacks make up the most of the types of foods that they purchase.

Customers with children comprise a large number of customers, too, and outnumber people without children by more than 2 to 1. However, both groups order at about the same frequency. The goods that these two groups purchase, however, likely differs and targeted strategies such as offering loyalty programs, coupons or referrals could be used to increase sales figures from these as well as all groups.

Adult customers comprise the largest demographic. This group also contains the most number of people labeled as those who make purchases frequently. The other age groups do not contain as many customers and do not have the same purchasing power as adults. More investigation should be performed to determine goods that these groups want in order to increase both the frequency and amount they spend on products.

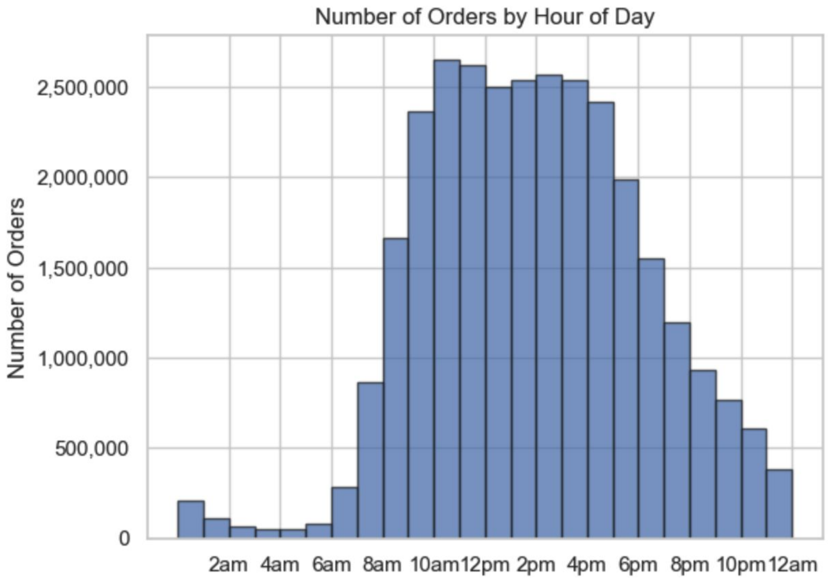
Ideally, additional data should be investigated to determine which products have the highest sales margin. Also, further investigation should be performed to determine why some products do not sell as well as others and steps would then be taken to either improve their sales or discontinue them if they do not generate profit for Instacart.

Code Examples

Code Examples - Explanation

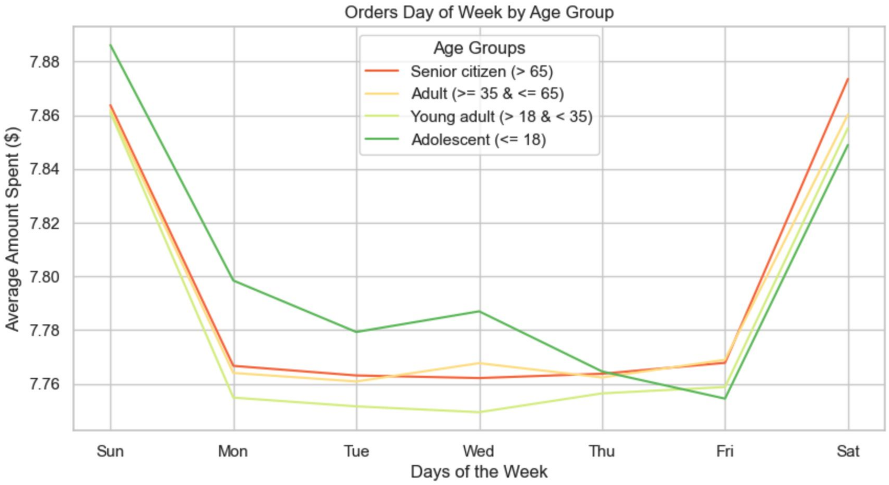
The following slides contain code I wrote to generate some of the corresponding visualizations I included in this presentation.

Code Examples - From Slide 8



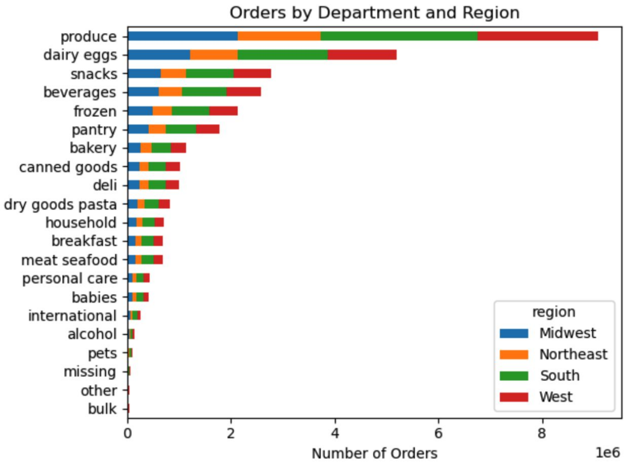
```
# Define custom bin edges to create separated bars
bins = np.arange(0, 25, 1)

histogram = df_2['order_hour_of_day'].plot.hist(bins = bins,
edgecolor = 'k', alpha = 0.7, color = 'b')

# Add a title to the histogram
histogram.set_title('Number of Orders by Hour of Day')

# Add a label to the y-axis
plt.ylabel('Number of Orders')

# Customize the x-axis ticks and labels
custom_xticks = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24]

# Define custom tick positions for the x-axis
custom_xlabels = ['2am', '4am', '6am', '8am', '10am', '12pm',
'2pm', '4pm', '6pm', '8pm', '10pm', '12am']

# Apply the custom tick labels
histogram.set_xticks(custom_xticks)
histogram.set_xticklabels(custom_xlabels)

# Format y-axis tick labels to display full numbers
def format_func(value, tick_number):
    # Format as integers with thousands separators
    return f'{int(value):,}'

# Apply the formatting to the y-axis
histogram.yaxis.set_major_formatter(tick.FuncFormatter(format_func))

# Show the plot
plt.show()
```

Code Examples - From Slide 10



```
# Line chart for days of the week (order_dow) when a purchase was
made and the age of
# the person who made the purchase (age_flag).
plt.figure(figsize = (10, 5))

line = sns.lineplot(data = df_2,
             x = 'order_dow',
             y = 'prices',
             hue = 'age_flag',
             palette = 'RdYlGn',
             errorbar = None
            )

# Customize the x-axis ticks and labels
# Define the custom tick positions (e.g., for days of the week)
custom_xticks = range(7)
# Define the custom tick labels
custom_xlabels = ["Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"]

# Apply the x-axis ticks and labels
line.set_xticks(custom_xticks)
line.set_xticklabels(custom_xlabels)

plt.title('Orders Day of Week by Age Group')
plt.xlabel('Days of the Week')
plt.ylabel('Average Amount Spent ($)')
plt.legend(title = 'Age Groups')

# Show the plot
plt.show()
```

Code Examples - From Slide 19



```
# Create the plot
department_region_counts = df.groupby(['department', 'region']).size().unstack()

# Calculate the total number of orders for each department
total_orders = department_region_counts.sum(axis = 1)

# Sort the data by total orders in descending order
sorted_data = department_region_counts.loc[total_orders.sort_values(ascending =
False).index]

# Plot the stacked horizontal bar chart
ax = sorted_data.plot.barh(stacked = True)

# Add title, x- and y-axis labels and a legend
plt.title('Orders by Department and Region')
plt.xlabel('Number of Orders')
plt.ylabel('')

# Reverse the order of the y-axis
plt.gca().invert_yaxis()

# Show the plot
plt.tight_layout()
```

Thank You