InstaCart Basket Analysis Case Study

—

Analysis by Ellen Johnston

About the Analyst

Hi, my name is Ellie. I am a data analyst with a unique blend of experience in the restaurant industry and a strong academic background in mathematics and data analytics. Through working in restaurants for the last 8 years, I've mastered the skills of problem-solving, multitasking, and ensuring customer satisfaction in high pressure environments.

In my freetime I perform as a hula hoop dancer at various events around the Twin Cities, showcasing my creative perspective. I am choosing data analytics as my next career path because I want to take my talents in math and creativity and use them in a tangible, real-world context.

Goal of Analysis

Instacart is an app-based grocery delivery service that wants to analyze the behavior of their customers to target advertisements more efficiently and generate more revenue. This analysis was conducted by looking at habits of customers such as when orders are most often placed, what items are most commonly ordered, and whether the demographic of a customer has an affect on ordering habits.

Project Deliverables

- Final Report
- GitHub Repository

Key Questions

- What is the busiest day of the week?
- What are the busiest hours of the day?
- Which day/time of day are customers spending the most money?
- How does a customer’s loyalty to the company affect ordering habits?
- How do ordering habits differ among various demographics?

Data Overview

There were 3 different data sets used to conduct this analysis.

Orders

Contains data on individual orders including customer name, order number of individual customer, time of day, day of week, and days since customers last prior order

Products

Contains data on each products in store including product ID number, name, which aisle it's located in, and which department it belongs to and price

Customers

Contains data on customers such as unique user ID, name, gender, state of residence, age, date joined, number of dependants, marital status, and income

Python Processes Used

- Data Wrangling
  - Removed unnecessary columns and renamed unintuitive columns
- Data Consistency Checks
  - Checked for and handled missing values, duplicates, and mixed type data in columns
- Combining Data
  - Merged datasets together using shared keys
- Deriving New Variables
  - Variables derived by using If-Statements with User Defined Functions, Loc Functions, and For Loops
- Grouping Data and Aggregating Variables
  - Grouped and aggregated variables using groupby(), agg(), and transform() functions
- Data Visualizations
  - Used visualization libraries such as matplotlib, seaborn, and scipy to create bar charts, histograms, scatterplots, and line charts

Order Frequency & Time

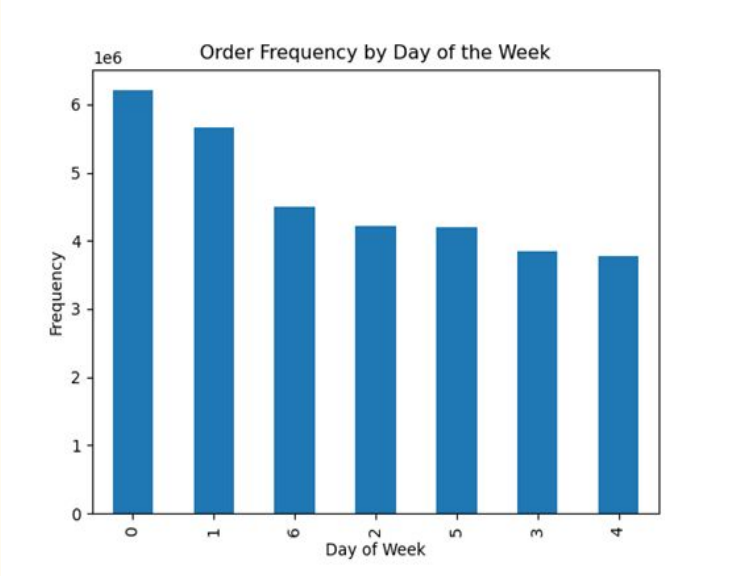
Two key questions can be answered by looking at order frequency and time; what are the busiest days for ordering on the app and what time of day is busiest.

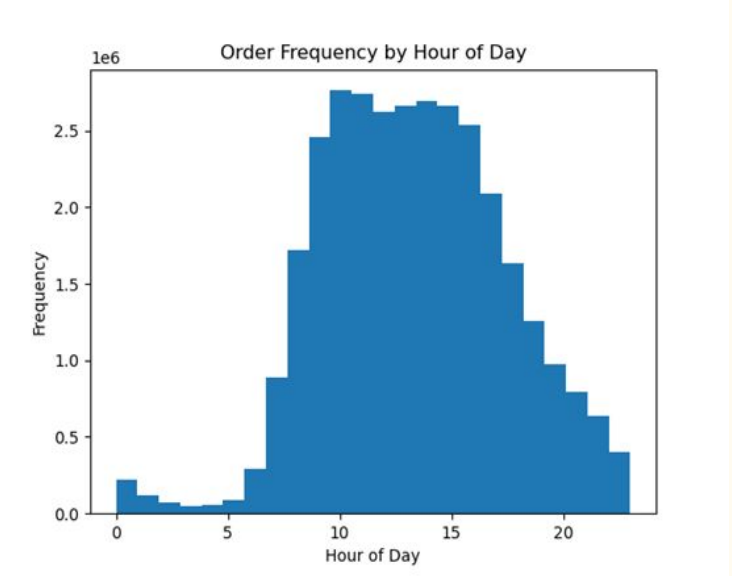
To conduct this part of the analysis, charts displaying the distribution of orders over the days and of the week and the hours of the day were created using the matplotlib library.

```
# Create a bar chart displaying the frequency of the orders_day_of_week column
dow_freq = df['orders_day_of_week'].value_counts().plot.bar()
plt.title('Order Frequency by Day of the Week')
plt.xlabel('Day of Week')
plt.ylabel('Frequency')
```

```
# Create histogram
hist = df['order_hour_of_day'].plot.hist(bins = 24)
plt.title('Order Frequency by Hour of Day')
plt.xlabel('Hour of Day')
plt.ylabel('Frequency')
```

Order Frequency & Time





Saturdays (day 0) and Sundays (day 1) are the days when orders are most often placed. Together they account for 36.6% of all orders.

The busiest hours of the day are from 10am-3pm. Though these hours make up only one quarter of the day, but account for 50% of all sales.

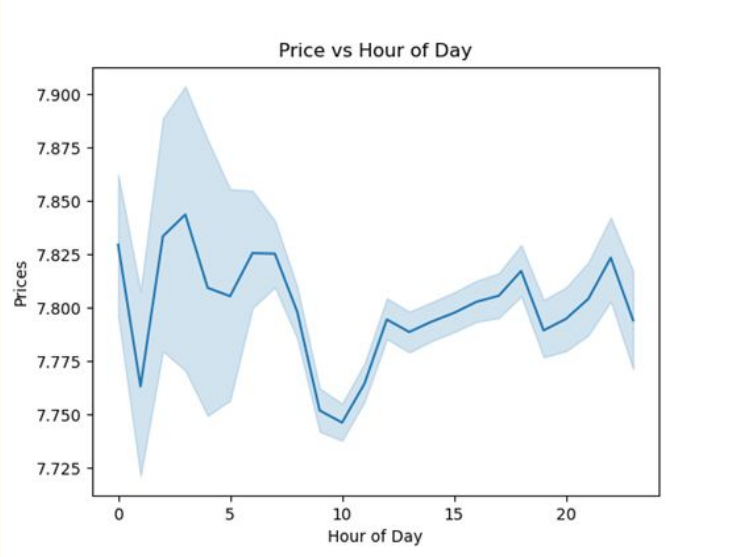
Expenditure & Time

Another main point of the analysis was to determine whether there are certain days of the week or hours of the day in the average price of an item ordered is greater.

Line charts were used to conduct this part of the analysis. The data frame being used was quite large so a random function was used to create a 70/30 split of the data. The line charts were created front from the smaller data frame.

```
# Assign random numbers between 0 and 1 to each row in dataframe
np.random.seed(4)
dev = np.random.rand(len(df)) <= 0.7
```

Expenditure & Time



The average price of an item ordered on Saturday or Friday is exponentially higher than other days of the week.

In the early hours of the morning when orders are infrequent the price spent on an item is unpredictable. Items order mid-morning are lowest price and items ordered later in the evening are more expensive.

Most Commonly Ordered Items

To determine which items are most commonly ordered a bar chart was created to determine the frequency of the departments that the items belong to. The top 10 items ordered were found by creating a new dataframe with value counts.

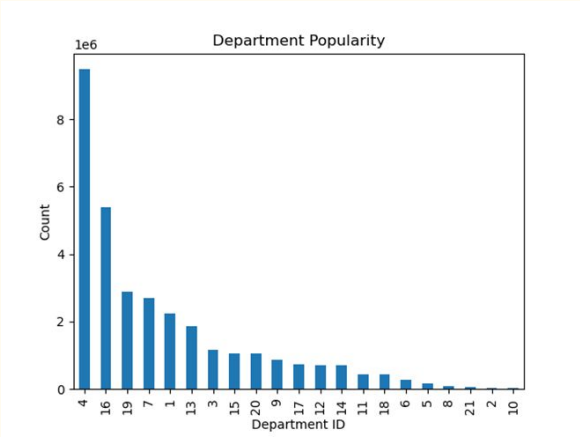
```
# Create bar chart from department_id column
bar = df['department_id'].value_counts().plot.bar()
plt.title('Department Popularity')
plt.xlabel('Department ID')
plt.ylabel('Count')
```

```
# Create dataframe containing value counts for products
prod_counts = df['product_name'].value_counts()
```

```
# Create dataframe to find top 10 most popular product
top_10_prods = prod_counts.head(10)
```

Most Commonly Ordered Items

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pasta
10	bulk
11	personal care
12	meat seafood
13	pantry
14	breakfast
15	canned goods
16	dairy eggs
17	household
18	babies
19	snacks
20	deli
21	missing



Value Counts for 10 Top Products

product_name	
Banana	472565
Bag of Organic Bananas	379450
Organic Strawberries	264683
Organic Baby Spinach	241921
Organic Hass Avocado	213584
Organic Avocado	176815
Large Lemon	152657
Strawberries	142951
Limes	140627
Organic Whole Milk	137905

The top departments are produce and dairy/eggs, which both consist of perishable items. Nine of the top 10 products belong to the produce department.

Demographic & Order Behavior

To analyze how different demographics behave the customers were split into various groups; the first divided them by age and income and the second by marital status and number of dependants. In both cases, the new variable were derived by using If-Statements and Loc Functions.

The age vs income group was defined as follows:

- young adult: <30
- adult: 29-65
- senior: >65

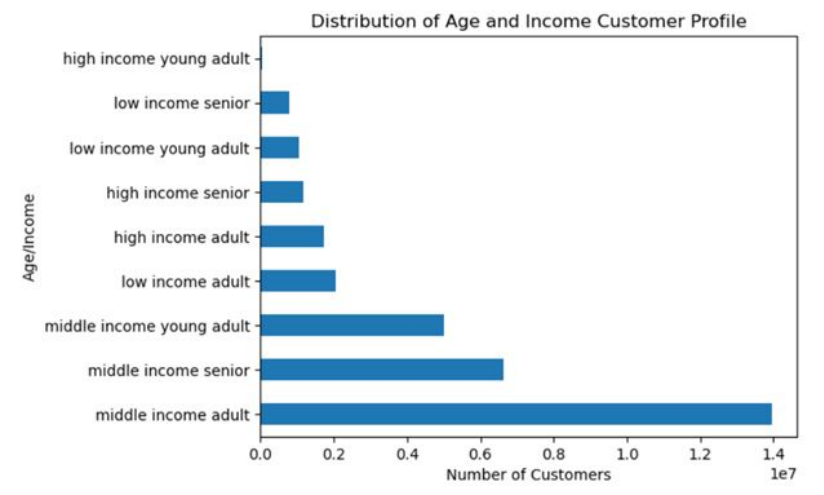
- low income: <$52,000
- middle income: $52,000-$156,000
- high income: >$156,000

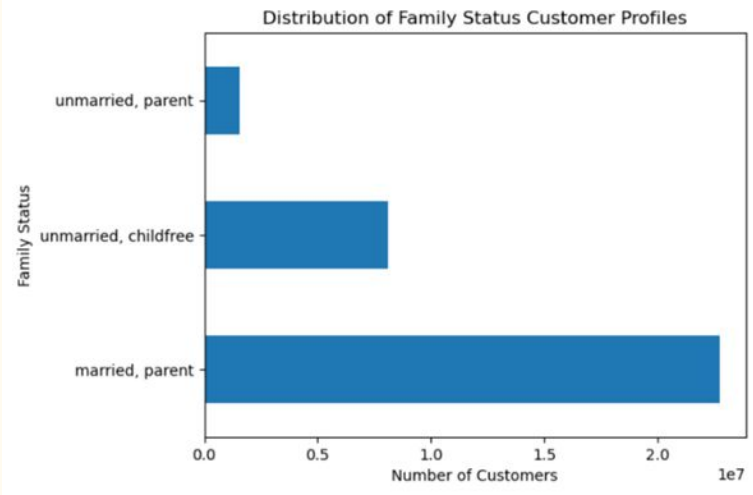
```
# Define low income young adult
ords_prods_merge.loc[(ords_prods_merge['income'] < 52000) & (ords_prods_merge['Age'] < 30), 'age_income_loc'] = 'low
```

```
# Define middle income young adult
ords_prods_merge.loc[(ords_prods_merge['income'] >= 52000) & (ords_prods_merge['income'] <= 156000) & (ords_prods_mer
```

```
# Define high income young adult
ords_prods_merge.loc[(ords_prods_merge['income'] >156000) &  (ords_prods_merge['Age'] < 30), 'age_income_loc'] = 'hig
```

Demographic & Order Behavior

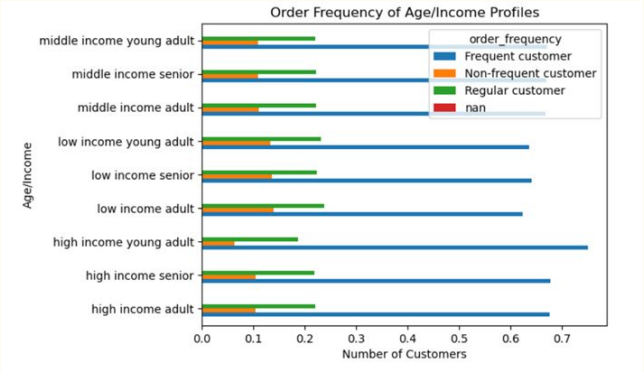


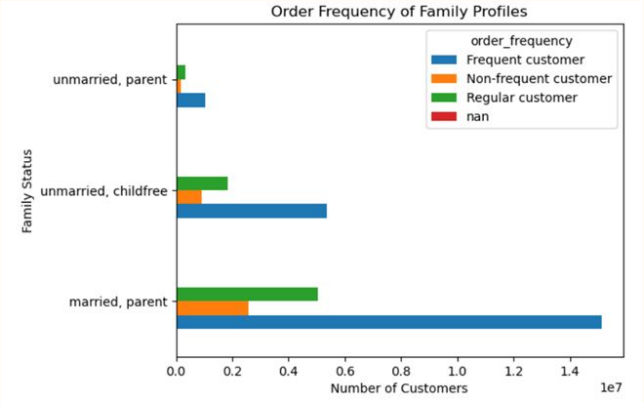


Middle income adults is the largest age vs. income group and accounts for 43% of all customers.

Married parents is the largest family status group and accounts for 70% of all customers.

Demographic & Order Behavior





Average Price Spent on an Item

age_income_loc	mean
high income adult	7.940265
high income senior	7.954955
high income young adult	7.803276
low income adult	6.720412
low income senior	6.500395
low income young adult	6.962268
middle income adult	7.930591
middle income senior	7.920826
middle income young adult	7.954097

family_status	mean
married, parent	7.790695
unmarried, childfree	7.789975
unmarried, parent	7.800712

When it comes to the different demographic groups, there is not significant difference on the order frequency or the average price spent per item.

Recommendations

Throughout the day, customers are more frequently placing orders between 10am and 3pm. However, they are likely to spend more per item between 5pm and 10pm. The frequency of ads should be increased for mid morning to afternoon, and then in the evening the focus should be on advertising more expensive items.

The weekends when we see customers ordering more frequently and also when they are spending the most on average per item. Overall weekend advertising should be increased and on the forefront of focus.

The most commonly ordered items come from the produce and dairy departments, both of which contain perishable items. Because these items need to be purchased so frequently, if some kind of sale was instituted for them it would incentivize customers to order instead of going to the store themselves and hopefully they would order other items as well.

There is no significant difference between how the different demographics order. However, the distribution of the demographics is significant. Ads should be targets towards items that would be popular among middle income adults or married parents.

Challenges

The biggest challenge I faced during this project was not being able to find any significant difference between the different demographic groups and order behavior. I felt that was one of the main questions to be answered and through all of my analysis I could not find anything besides the distribution of the demographics.

Overall I found learning python and the coding portion to be easy to learn, however when dealing with notebooks that were overly large I struggled with how to handle them. For the last few exercises in this unit I learned that it would be easier for me to submit a few smaller notebooks rather than one large one.