Instacart Grocery Basket Analysis

Date: 1st August 2021
Author: Senja P

Context and Goals

Background

Instacart in an online grocery store that operates through an app. In this study project, the marketing and sales teams wanted to get a better understanding of the customers and their purchasing behaviours in order to plan targeted marketing strategies. My role as a Data Analyst was to answer for the following key questions;

Key Business Questions

Who are the customers?

- What are the customer demographics? Are there differences in ordering habits based on a customer's loyalty status and regions?

What do they buy?

- Are certain types of products more popular, are there differences in customers' behaviour?

When do they shop?

- What are the busiest days of the week and hours of the day, and are there any particular times when customers spend most money?

Data and Analysis Process

Data

Size: 35 millions rows

Data Sets: products, orders and customers

Year: 2017

Source: Instacart Online Grocery Shopping Dataset 2017 – open source [Accessed from https://www.instacart.com/datasets/grocery-shopping-2017, on 4th July 2021]

The customer dataset and prices column in product dataset were fabricated for the training purposes by Career Foundry.

Approach

Descriptive analysis, the purpose was to use past data for marketing purposes.

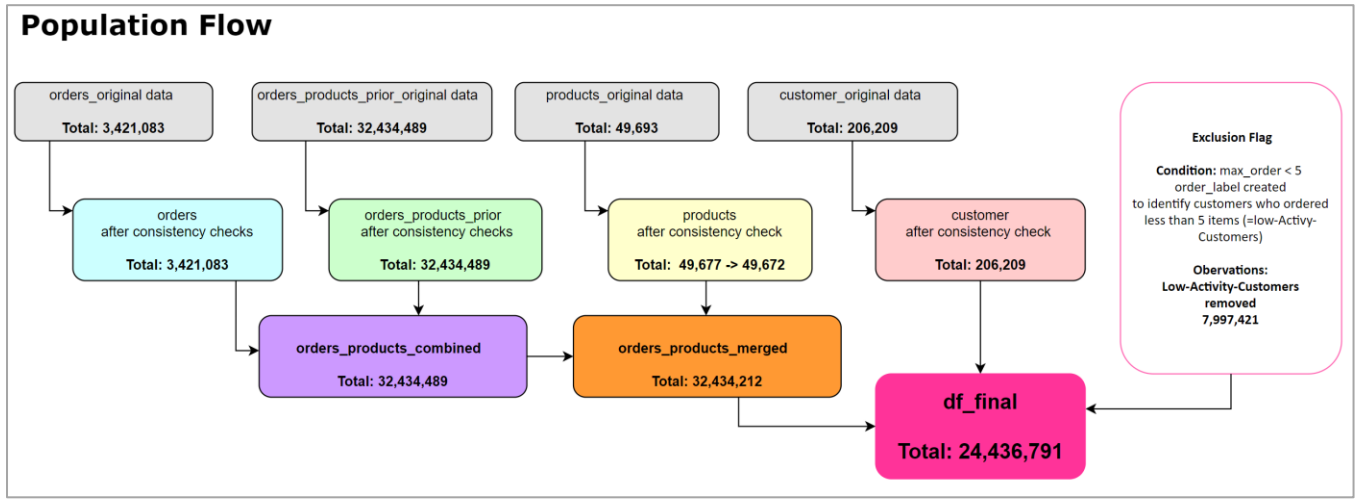

Figure 1: Population Flow – the number of items in original and cleaned and merged datasets

Analysis Steps

- Installing libraries and importing dataframes
- Completing consistency checks; missing values and duplicates
- Data wrangling, e.g. dropping and renaming columns
- Merging dataframes
- Grouping, aggregating and creating new variables to answer the business questions
- Creating visuals and exporting dataframes
- Creating a population flow

Analysis Tools and Techniques

Technical Tools Utilised

- Python/Jupyter notebook & Anaconda

pandas, numpy, os, matplotlib, seaborn & scipy

- Excel for visualisations and a data dictionary

Key Combined Variables

- Customer behaviours were analysed against their loyalty status and household types.
- The age and dependant status variables were used to form the household types.
- The loyalty categories based on the number of orders; new customers placed 12 or less orders, regular customers ordered more than 12 but less than 36 times and the loyal customers made 36 or more orders
- The US states were divided into four geographical regions; Northeast, Midwest, South and West

Challenges and learnings

- Working with a large data set, over 32M rows, I understood how large dataframes can lead to memory issues and I learned to use a subset dataset instead.
- I also learned how to ignore unnecessary columns when importing data and create flags and if-statements with the loc() function when sorting the products into low, medium and high price range categories.
- As this was the first time when I used Python I found it easier to use Excel for some visualisations, for example I was able to convert the Python summary data easily into percentages in Excel and use that information for graphs.
- If I started the project again I would consider choosing the different customer age categories. The more and smaller age groups might provide a better and more detailed information for marketing teams.

Findings - Who are the customers?

- The households with children use more Instacart Grocery online services than those without children in each household category. Approximately 50% of all orders were place by customers over the age of 40. The 30-40 year olds with children made 13% and young parents under 30s made 14% of all customers.
- The most orders were made in South region, where the most customers were living. The general regional shopping habits were in line with the percentages of customer populations in each region and there were not significant differences between customers' income and their loyalty status.

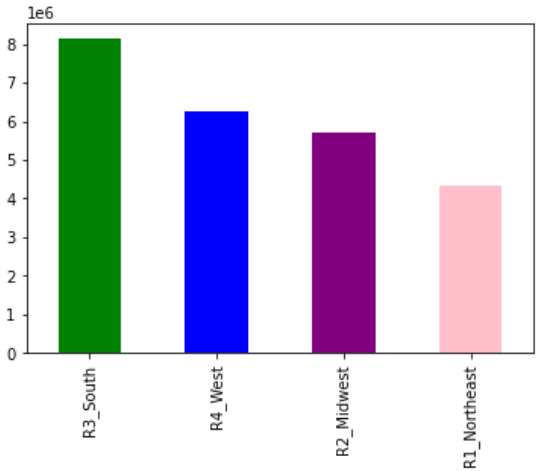

Figure 2: The number of customers (in millions) in each US region, excl. low activity customers who ordered 5 or less times (Python)

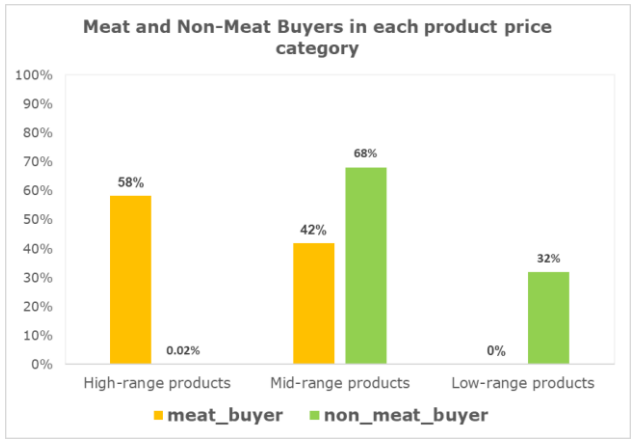

Figure 3: The shopping percentages of meat and non-meat buyers by product price categories

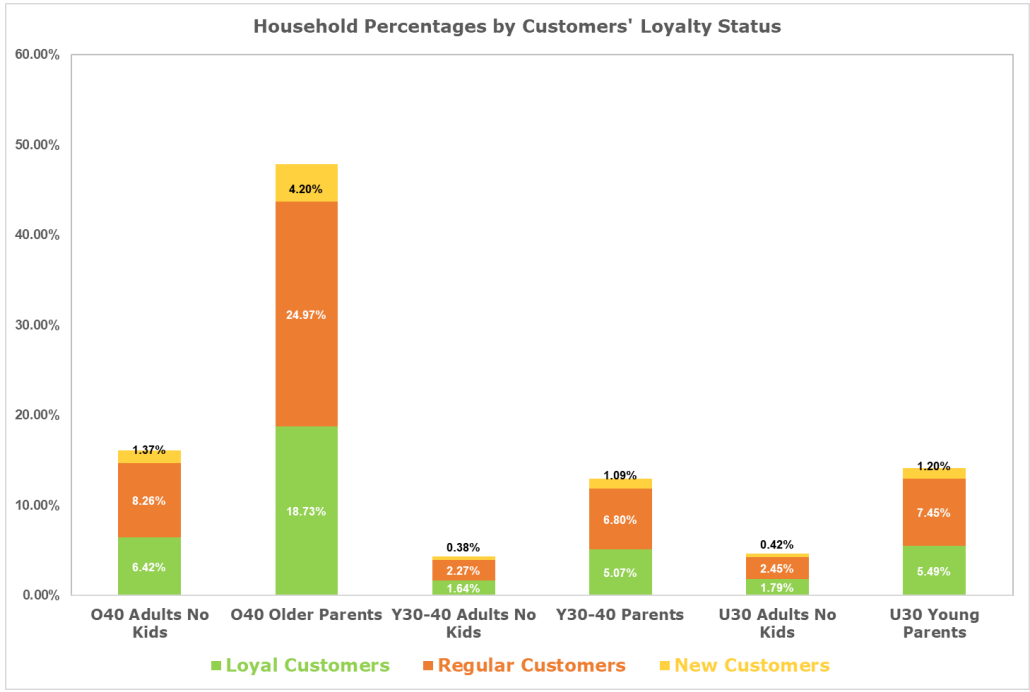

Figure 4: Percentage of Customers' household types by loyalty status

- The meat buyers bought 58% of high price products but they didn't spend their money on low price products or there were no products available in this category. The non-meat buyers spent only 0.02% on high price products. The most of products, 68% were on mid-range and 32% were on low-range products.

Findings - What do they buy?

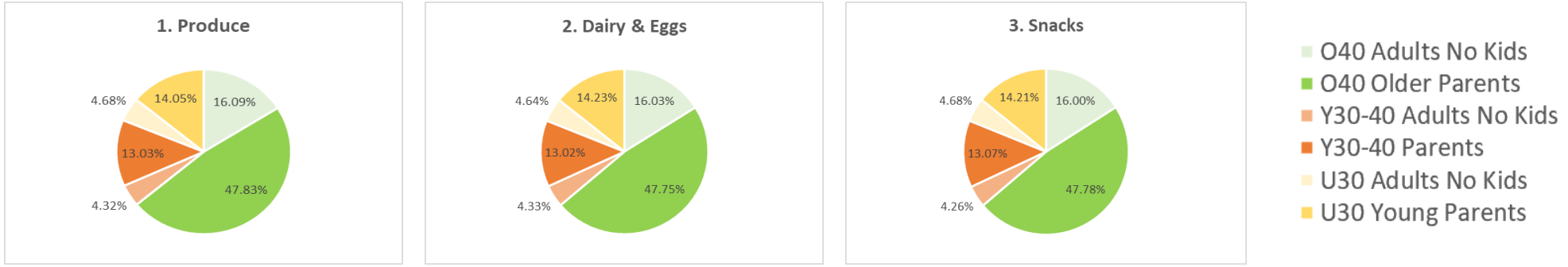

Figure 5: The top 3 departments and shopping percentages by household types (Produce includes farm-produced crops, also fruits and vegetables)

- The most popular departments were Produce, Dairy & Eggs and Snacks. Almost 50% of products in all three popular departments were bought by over 40s who have children.
- Marketing and sales team wanted to use simpler price range groupings to help direct their efforts. The products were categorised based low (price $5 or less), mid ($5-$15) and high (more than $15) price range products.
- Only Dairy and Egg products were found in high price range product category. As expected the most of the snack were in low-price category. Approximately 75% of all produce department products were classified in mid-range products.

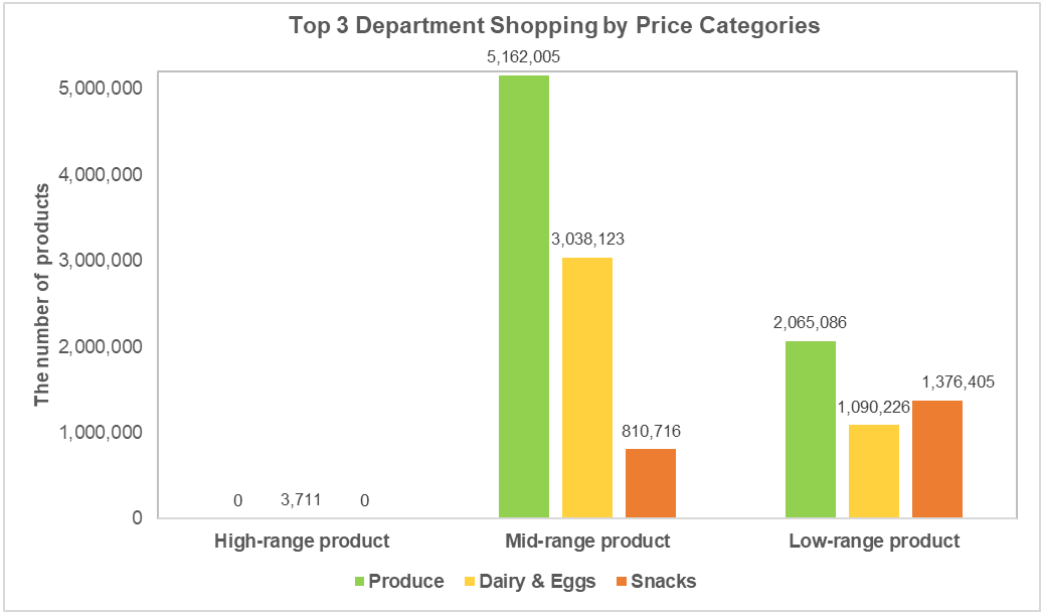

Figure 6: The top 3 department shopping by price categories

Findings - When do customers shop?

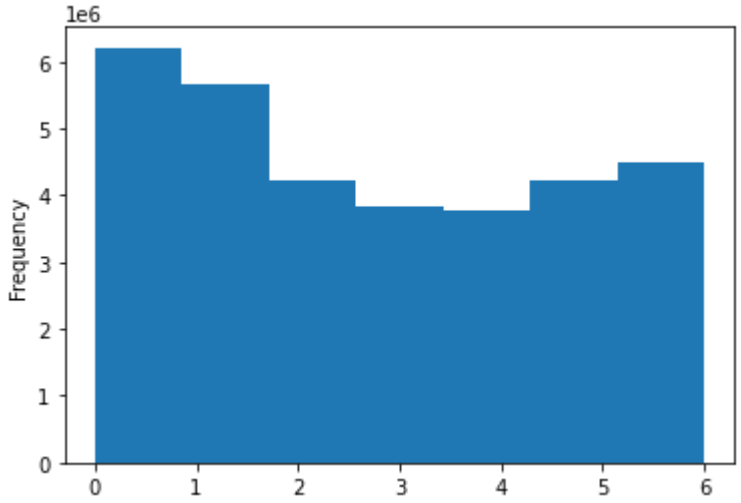

Figure 7: The shopping frequency by days (from Sat to Fri)

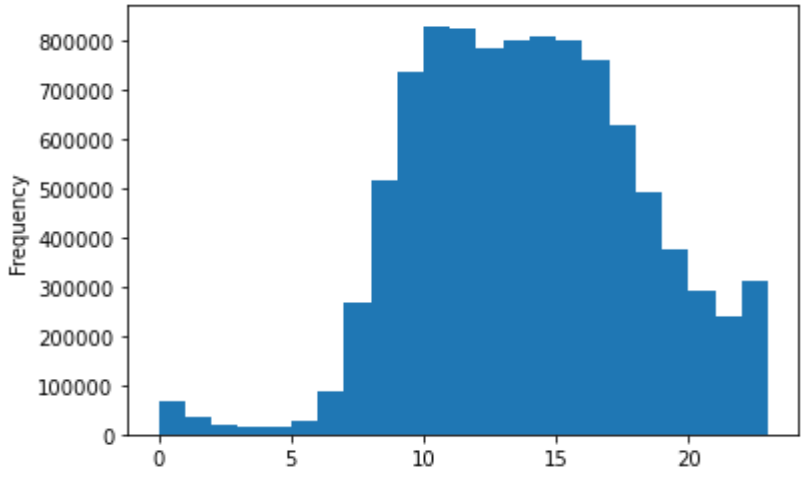

Figure 8: Orders by hours of the day (Python)

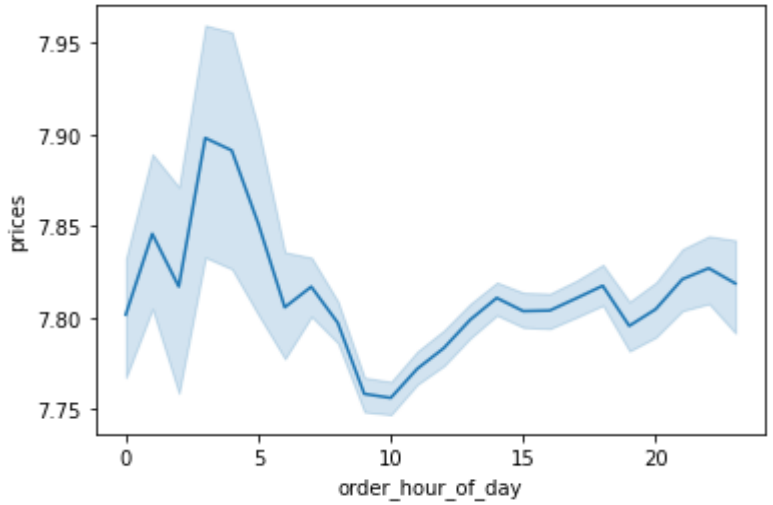

Figure 9: The relationship between product prices and hours when orders were placed (Python)

The weekend is the busiest time for placing online orders.
Saturday (=0) is the most popular day, followed by Sunday (1) and Friday (6). Tuesday and Wednesday are the less busiest days but there are not big differences on numbers between Tue (3) and Wed (4) or Monday (2) and Thu (5).

Histogram shows that the most orders are placed between 10am and 5pm, the peak being at 10-11am. The lowest number of products have been bought between 3am and 5am.

The line chart indicates that the most expensive items are bought in early hours; between 3am and 4am.

Recommendations

- Loyal and regular customers; The most of the marketing budget should be spent on those customers who have children as they place the most orders in all age groups.
- Finding New Online Customers; I would recommend to focus on the customers under 40s with no children as this group made up only 8% of current customers. However, further analysis is required to better understand their shopping habits.

This data didn't support the growing trend on vegetarianism as the egg and dairy products were the second most popular department. However, meat buyers made up only approximately 2% of total sales and the most popular department was produce. Therefore I would recommend to focus on marketing the vegetarian/produce classified products.

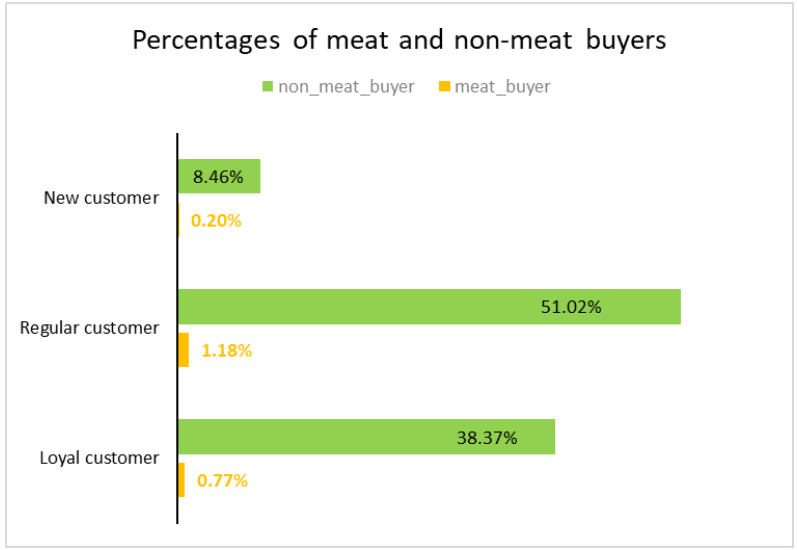

Figure 10: Total percentages for meat and non-meat buyers by loyalty groups

- The orders are declining from 5pm to 10pm but increasing from 10pm. Therefore I would recommend to increase marketing between 5pm to 9pm.
- Although the most expensive items are bought in early hours I do not recommend to promote any products or services in early hours as this might disturb customers when they are sleeping.

Appendix

Consistency Checks, Wrangling Steps and Column derivations available on the Excel spreadsheet

Instacart_Data
ning Steps_2021_0