Instacart Basket

ANALYSING SALES DATA FOR SUCCESSFUL US-BASED ONLINE GROCERY COMPANY
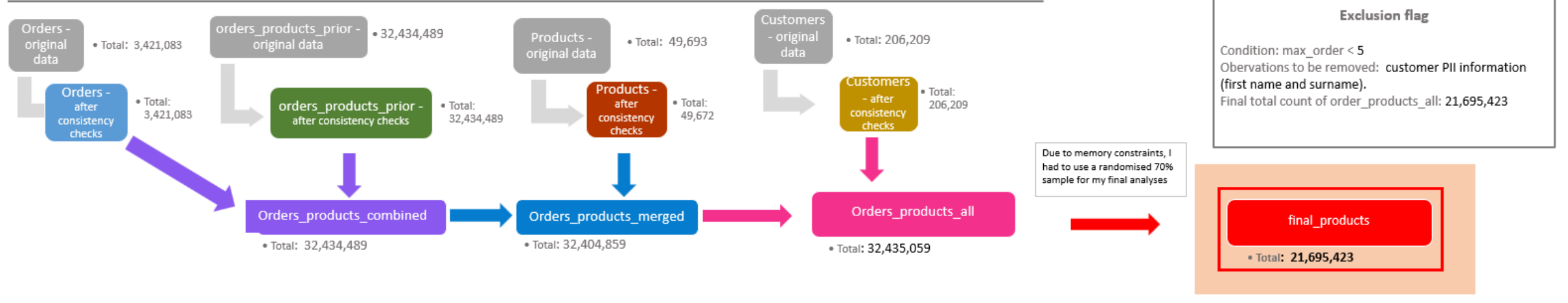
Project Overview

Objective:	To conduct exploratory analysis for Instacart Grocery, a US-based online grocery store. To derive insights into customer sales patterns, segment customers based on behavioural and demographic profiles to analyse spends, and then make recommendations for targeted marketing campaigns.
Data:	Opensource purchasing data was supplied by Instacart. Customer data was fabricated by CareerFoundry for the purposes of this exercise.
Tools & Skills:	Analysis was conducted in Python using Jupyter notebooks. Many libraries were used including pandas, numpy, matplotlib, scipy, and seaborn. Analysis involved: Data cleaning including data wrangling and consistency checks. Data manipulation including merging datasets, deriving new variables, and subsetting. Data analysis including aggregating data, creating visualisations, using crosstabs, and descriptive statistics.
Constraints:	Data supplied by Instacart is only from 2017 and some data had to be fabricated for the purposes of the project. A randomised 70% subset had to be used due to memory issues.

Key Insights: Data cleaning & Wrangling

Four datasets were cleaned, wrangled, and merged to create the final set we used for our analysis. The data contained over 21 million rows.

Population flow

Exclusion flag

Condition: max_order < 5
Obervations to be removed: customer PII information (first name and surname).
Final total count of order_products_all: 21,695,423

Orders - original data • Total: 3,421,083
Orders - after consistency checks • Total: 3,421,083

orders_products_prior - original data • 32,434,489
orders_products_prior - after consistency checks • Total: 32,434,489

Products - original data • Total: 49,693
Products - after consistency checks • Total: 49,672

Customers - original data • Total: 206,209
Customers - after consistency checks • Total: 206,209

Orders_products_combined • Total: 32,434,489
Orders_products_merged • Total: 32,404,859
Orders_products_all • Total: 32,435,059

Due to memory constraints, I had to use a randomised 70% sample for my final analyses

final_products • Total: 21,695,423

Consistency checks

| Dataset	Missing values	Missing values treatment	Anomalies	Anomalies treatment	Mixed-data types	Duplicates
orders	206,209 missing values found in the *days_since_prior_order* column	Values equated to less than 6% of dataset so they were left as presumed to be new customers.*	None	None	None	None
products	16 missing values were found in the *product_name* column	Created a new dataframe minus missing values	Maximum price was listed as 99,999 for dairy/eggs item.	Impacted results were set to NaNs and disregarded	None	5 duplicates were removed
orders_products_prior	0	N/A	0	N/A	0	0
customers	11,259 in *first_name* column	*First_name* is not needed for our analyses, so it was removed	0	N/A	0	0

*the client only wanted data from customers with more than 5 purchases, however, we could create a flag for new customers using this data point if we wanted to identify and target that group

3

Initial analysis

Most popular day for purchases
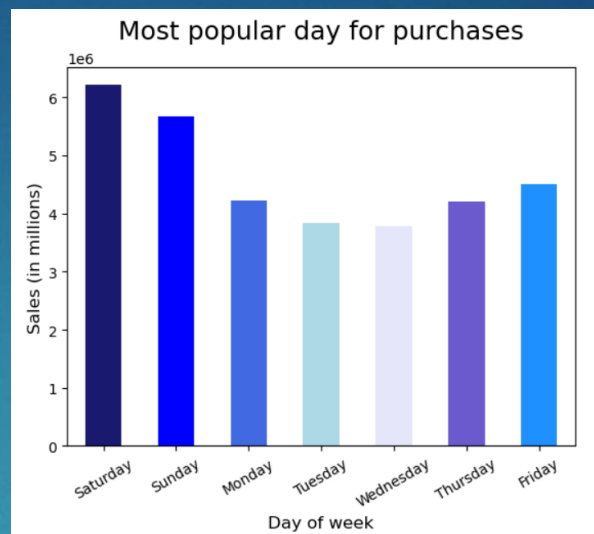
The most popular days of the week for purchasing are Saturday and Sunday. The most popular hours are within a standard hours, 9am until 5pm.

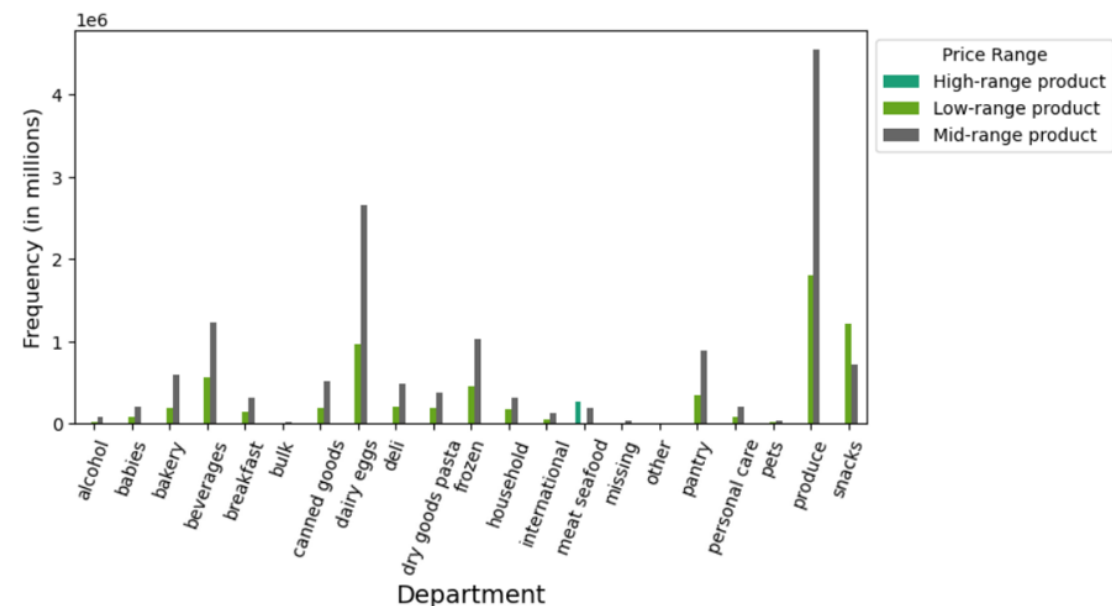The most popular departments are produce and dairy/eggs. Most products fall within mid or low-range price groups.

Distribution of product prices
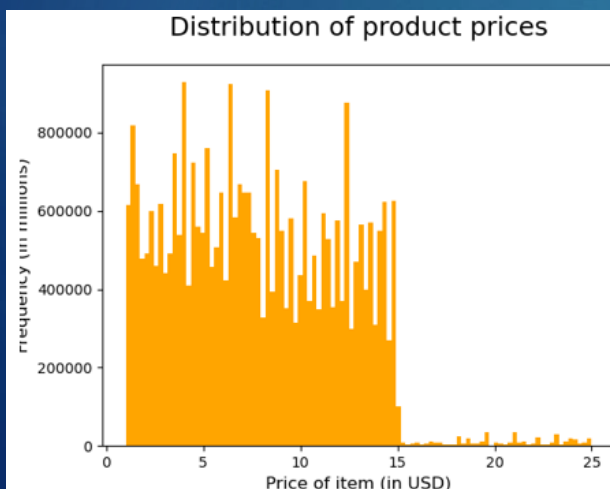
Price grouping conditions

<= 5 : 'Low-range product'

(> 5) & (<= 15) : 'Mid-range product'

> 15: 'High range'

else: 'Not enough data'

No. of Purchases by Departments and Product Price Range
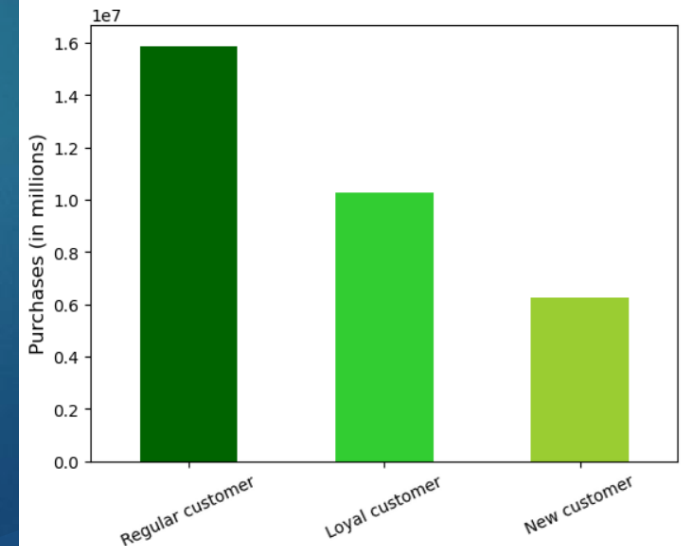
Customer profiling

Demographic grouping allowed us to gain insights such as:

 - Families with dependants - irrespective of income bracket - account for more purchases than those who have no dependants.
- Age appears to have no impact on purchasing habits.
- Time of purchase is pretty consistent across all age groups
- Departmental spend is broadly consistent across all income brackets

Purchases by family status and department
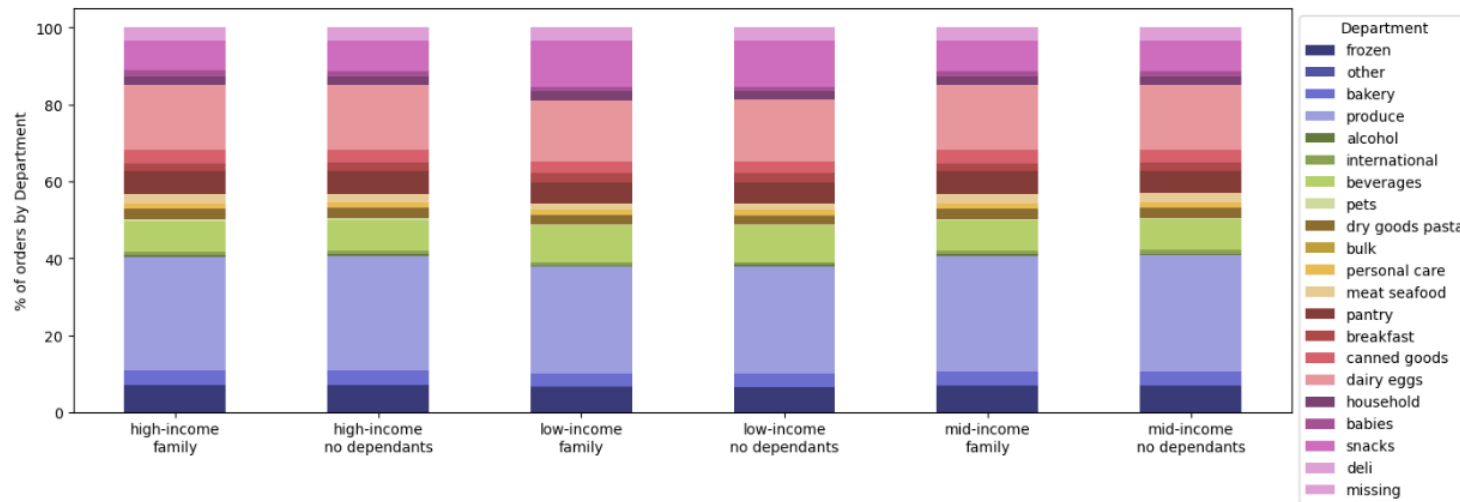
Purchases by loyalty status

Recommendations

▶ The analysis revealed limited differences in spending habits when looking at a variety of demographic or customer profiles. Some differences, however, show that low-income families purchase more snacks, or families with dependants spend more than those with no dependants. These insights could be incorporated into targeted marketing campaigns.

▶ Instacart should look to take advantage of peak hours over the weekend by focusing advertising in the run up to the weekend, and then in-app push notifications on the day.

▶ There is significant brand loyalty so Instacart should investigate a rewards scheme to incentivise regular and loyal customers and maximise profits.

▶ Instacart could investigate whether targeted campaigns to students and seniors could increase purchasing at non-peak hours.

▶ Access to time-series data would allow us to look at seasonal and annual trends and could reveal further insights.

Final Thoughts

- *If I could change something, I would have made my career change earlier. I decided to switch to technology after fifteen years in retail and healthcare, and I am thankful for the skills I learned over the years, this is my true passion. I could have had more time in this industry that I adore if I had made the change sooner, but I'm happy to be here now.*