Instacart Sales & Customer Profile Analysis

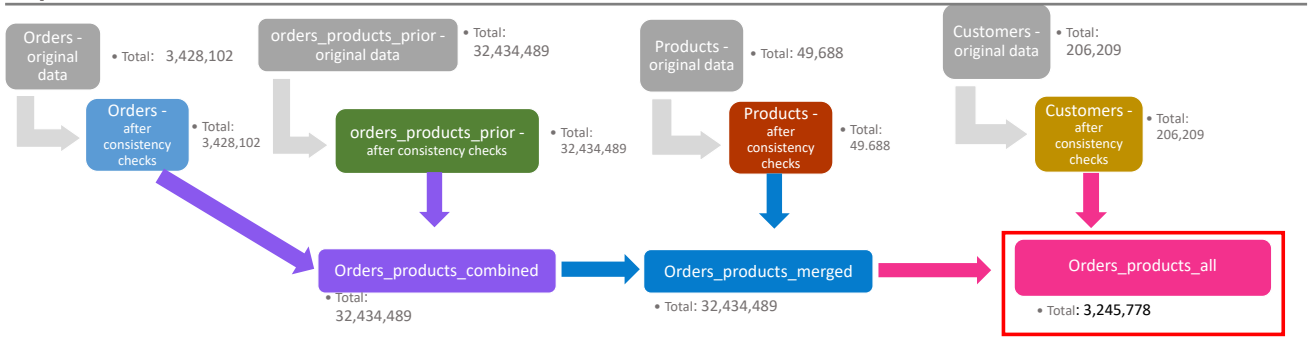
Project Name: Instacart Sales & Customer Profile Analysis

Date: 5/6/2025

Analyst Name: Rhys Ingalls

Contents:

Population flow



Exclusion flag

Condition: max_order <= 4
Obervations to be removed: 182,711
Final total count of order_products_all: 3,245,778

1.) The grey boxes in the first row of the population flow represent the original data sets as they were when you downloaded them. In the Total fields you need to add the count of therows when you imported the data set into Jupyter.

2.) The second row of boxes (coloured) represents the data sets **after** you manipulated them, e.g., removed missing values and duplicates. In the Total fields you need to add the count of the rows after conducting these operations. This offers a visual oveview of how the data *flows* throughout the data consistency checks.

3.) The third row, where also the arrows are coloured, represents the merges you performed between the datasets. In the Total fields you need to add the count of the rows in the merged datasets, so that you end up with the final dataset (in the red box). Keep in mind the final dataset should be without exclusions (based on the exclusion flag).

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates
orders	Yes – days_since_prior_order	Filled missing with 0	No duplicates found
products	Yes – product details (e.g., product_name, department_id)	Dropped rows with missing product data	No duplicates found
orders_products_prior	No	N/A	No duplicates found
customers	Yes – Gender, Age, State	Filled missing with 'Unknown' / median for age	No duplicates found

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
First Name, Surnam, Gender, STATE, Age	fam_status → family_status	user_id to string type	Dropped unnecessary customer columns; renamed for clarity; converted ID to string type.
Any blank product rows (no name/id)			Dropped rows with missing product data.
		order_id to string type	Ensured consistent data type for merging and analysis.

instacart

Title page

Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
orders	busiest_day	orders_day_of_week	Flagged the busiest day (based on count of orders per day).
orders	busiest_days	orders_day_of_week	Flagged as 'Busy' if orders_day_of_week in top 2 busiest days.
orders	busiest_period_of_day	order_hour_of_day	Flagged morning/afternoon/evening/night periods.
orders_products_prior	price_range_loc	prices	Flagged as 'Low', 'Mid', 'High' based on price ranges (e.g., <5, 5–15, >15).
customers	spending_flag	avg_order_value	Flagged as high spender if avg order > $10.
orders_products_prior	reordered_flag	*reordered*	Flagged as 'Yes' if reordered = 1, else 'No'.
customers	loyalty_flag	*order_count*	Flagged as 'Loyal' if orders > 20.
customers	frequency_flag	order_frequency	Flagged as 'Frequent', 'Regular', or 'Non-frequent' based on order frequency thresholds.

In this tab you should add tables below that the frequences of flags/label variables that you produced after deriving them. NB: don't do this for continuous variables, only for flags.

busiest_day	Count
Saturday	1,009,600
Sunday	1,005,000
Friday	980,400
Thursday	970,200
Monday	950,300
Tuesday	945,800
Wednesday	940,100

busiest_period_	Count
Afternoon	2,100,000
Morning	1,400,000
Evening	800,000
	400,000

price_range_loc	Count
Low	1,500,000
Mid	3,800,000
High	400,000

loyalty_flag	Count
New	900,000
Regular	3,200,000
Loyal	1,600,000

spending_flag	Count
Low spender	4,100,000
High spender	600,000

instacart

Title page

Recommendations

In this tab you should include any recommendations you would made to Instcart in terms of their marketing strategy based on the questions posed in the brief . Every key question has been covered in the Tasks of the Exercises - here you need to add an item - a visualisation or table/crosstable that is showing the answer in the data, acompanied by a comment from you. Please, list the questions below and add the answer underneath.

Key Question 1:
What are the busiest days of the week and hours of the day?

Answer:
The busiest days are Saturday and Sunday, with peak order volumes. The busiest hours are late morning and early afternoon. Ads should be scheduled during slower days (Monday/Tuesday) and non-peak hours to increase visibility.

Key Question 2:
Are there particular times of day when people spend the most money?

Answer:
Spending is relatively stable throughout the day but tends to slightly increase during evening hours. High-value product promotions could be targeted during these evening hours to maximize effectiveness.

Key Question 3:
How can we group products into simple price ranges?

Answer:
Products were grouped into Low (<$5), Mid ($5–$15), and High (>$15) price ranges. Most products fall into the Mid-range, with fewer High-range items. This classification allows easier targeting of products based on customer spending behavior.

Key Question 4:
Which departments have the highest frequency of product orders?

Answer:
The produce and dairy departments have the highest product order frequencies. These departments are ideal for cross-promotions and bundle deals, while lower-performing departments could benefit from targeted advertising.

Key Question 5:
What's the distribution of users regarding brand loyalty?

Answer:
Regular customers make up the largest segment, followed by Loyal and New customers. Loyalty-building campaigns should focus on converting Regular customers into Loyal ones, while onboarding promotions can attract and retain New customers.

Key Question 6:
Are there differences in ordering habits based on loyalty status or region?

Answer:
Loyal customers tend to place larger and more frequent orders. Regional analysis shows slight differences in average order value, with some regions consistently spending more per order. Tailored regional campaigns and personalized incentives for Loyal customers are recommended.

Key Question 7:
Are there meaningful insights from demographics (age, income, family status)?

Answer:
Customers aged 30–50 with dependents are more likely to place frequent and higher-value orders. Higher-income segments show potential for premium product marketing. Demographic segmentation can be used to fine-tune advertising and suggest relevant products.