DATA ANALYTICS IMMERSION

ANNEMARIE SAUERBIER 9.OCTOBER 2025

Instacard Project

ANNEMARIE SAUERBIER 7.OCTOBER 2025

Content

ORDERS ORIGINAL DATA

3421083, 7

ORDERS PRODUCTS PRIOR ORIGINAL DATA

49693, 5

PRODUCTS ORIGINAL DATA

49672, 5

CUSTOMER ORIGINAL DATA

206209, 10

ORDERS ORIGINAL DATA

(32434489, 11)

•ORDERS_PRODUCTS_PRIOR - after consistency checks

ORDERS ORIGINAL DATA After consistency checks

CUSTOMERS after consistency checks

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 the rows 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).

ORDERS PRODUCTS COMBINED

ORDERS PRODUCTS MERGED

(32404859, 15)

ORDERS PRODUCTS ALL

[5 rows x 27 columns]
(30629741, 27)

WRANGLING

COLUMNS DROPED	COLUMNS RENAMED	COLUMNS' TYPE CHANGED	COMMENT/REASON
Eval_Set new: object			Not important
	Order_DOW→oders_day_of_week new: int64		Clear naming convention
		oder_id new: int64	ID as strings for merge

COLUMN DERIVATIONS AND AGGREGATIONS

DATASET	NEW COLUMN	COLUMN/S IT WAS DERIVED FROM	CONDITIONS
df	mean_product_price	prices (grouped by user_id)	groupby('user_id')['prices'].transform('mean')
df	spending_flag	mean_product_price	'Low_spender' if < 10 else 'High_spender'
df	max_order	order_number (grouped by user_id)	groupby('user_id')['order_number'].transform('max')
df	Aggregated summary	order_number (grouped by department_id)	Aggregated mean, min, max using .agg({'order_number': ['mean', 'min', 'max']})

#The merged dataset contains 30.6 million entries across 27 columns, combining detailed order, product, and customer demographic information for in-depth analysis.

OBJECTIVE

Understand customer profiles, spending, and regional patterns to guide marketing strategies

CUSTOMER LOYALITY

Regular customers make up the largest group by far.
Loyal customers come second.
New customers are the smallest group. The business has a strong base of returning
(regular) customers, which is a positive sign.

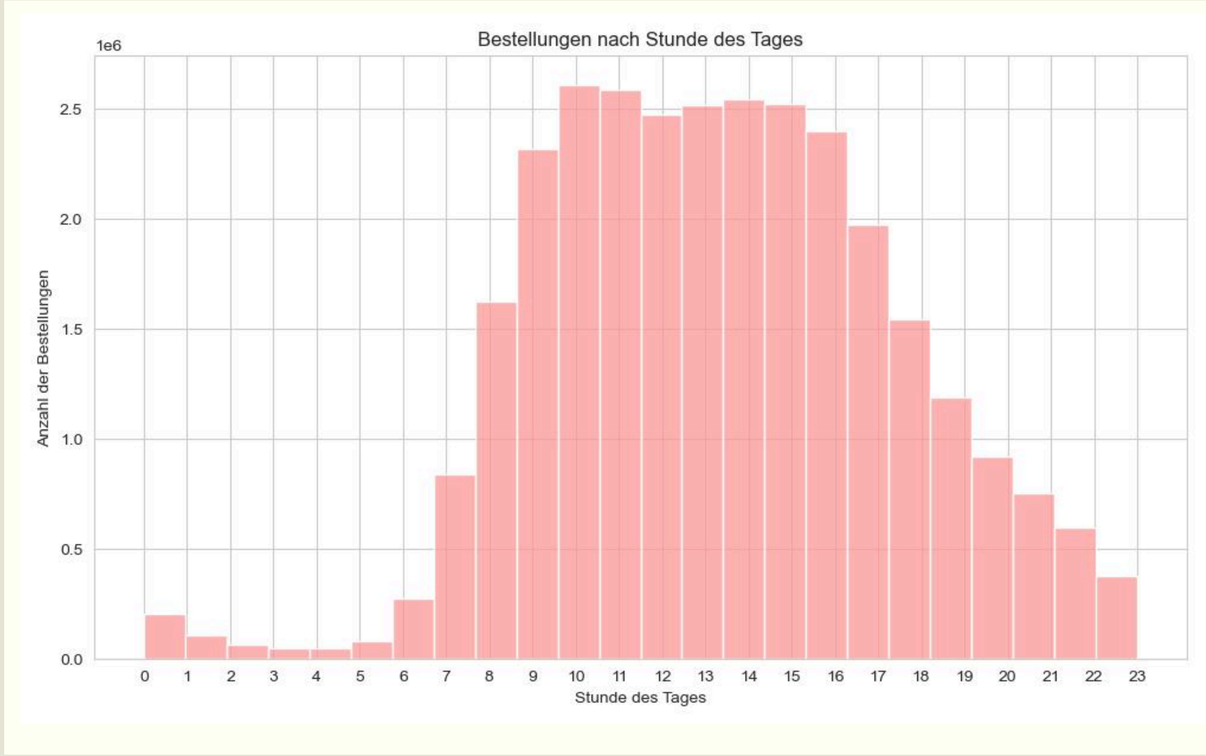
There may be room to convert more regulars into loyal customers through loyalty programs, incentives, or personalized offers.

Marketing efforts might also be focused on acquiring more new customers to grow the base.

This bar chart displays the number of orders placed at each hour of the day (from 0 to 23). The y-axis is labeled "Anzahl der Bestellungen" (Number of Orders).

The x-axis shows the hour of the day (in 24-hour format). Knowing peak hours helps in staffing, inventory planning, and system load handling.

In this case, almost all customers (about 99.7%) are high spenders with average #spending above $10, and only a very small fraction (around 0.27%) are low spenders. #It means my marketing and sales efforts should primarily focus on retaining and upselling to the high spender segment, since they make up almost the entire customer base and likely generate most of your revenue.

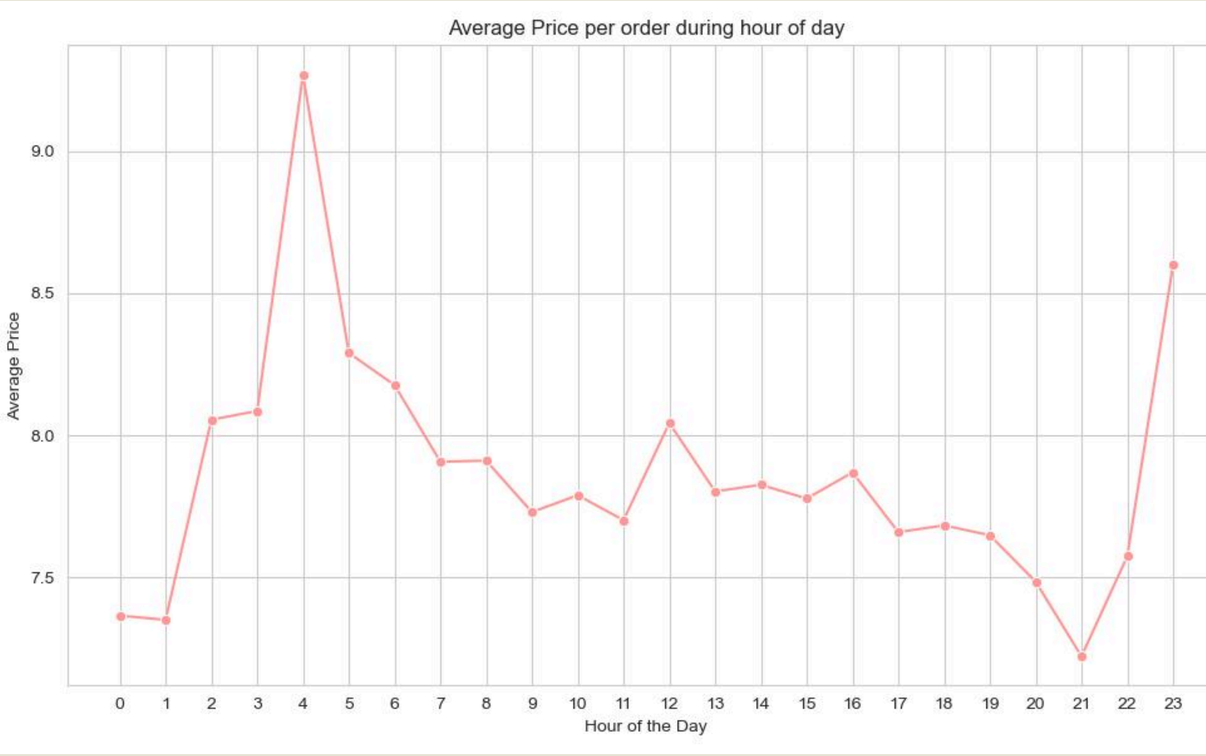


AVERAGE ORDER PRICE BY HOUR

Highest prices at 3–4 AM and 23:00. Lower prices during peak hours (9:00–17:00), suggesting smaller routine orders.

IDEAS

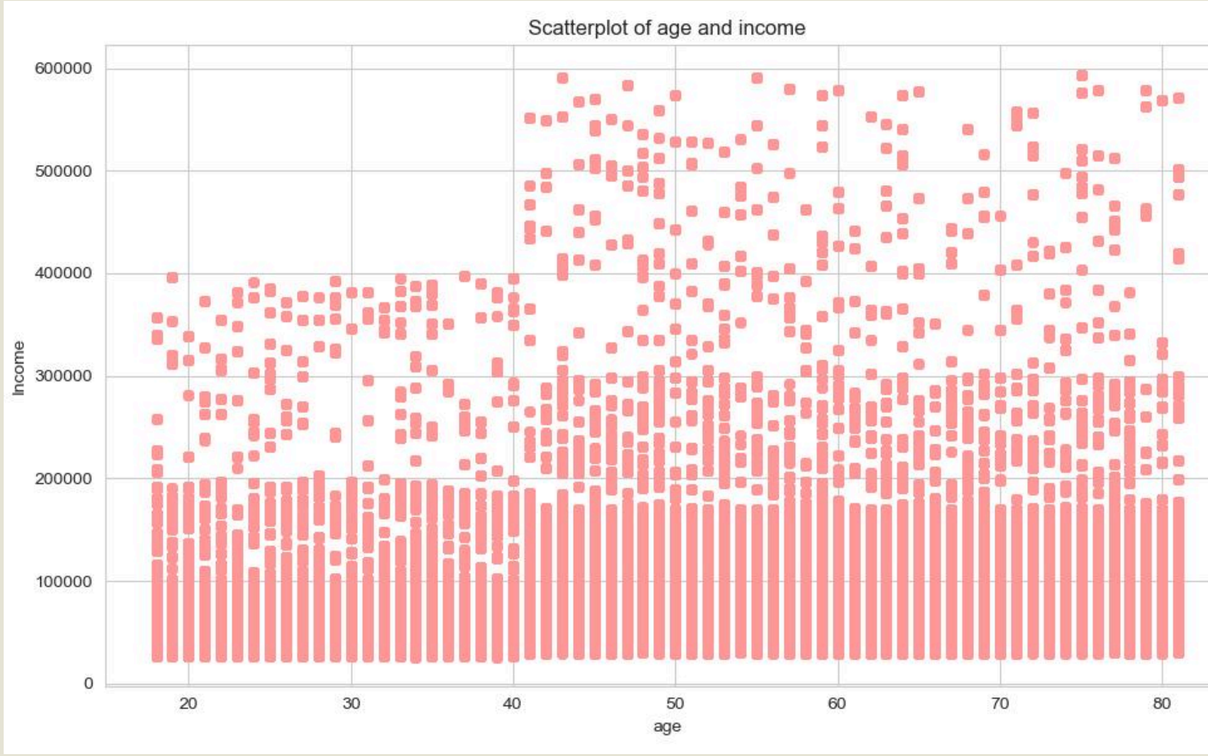
Run targeted promos during peak hours to boost average basket size and analyze late-night orders to understand why they're higher in value. Offer value deals during off-peak hours to increase order volume.



SPENDING VS AGE

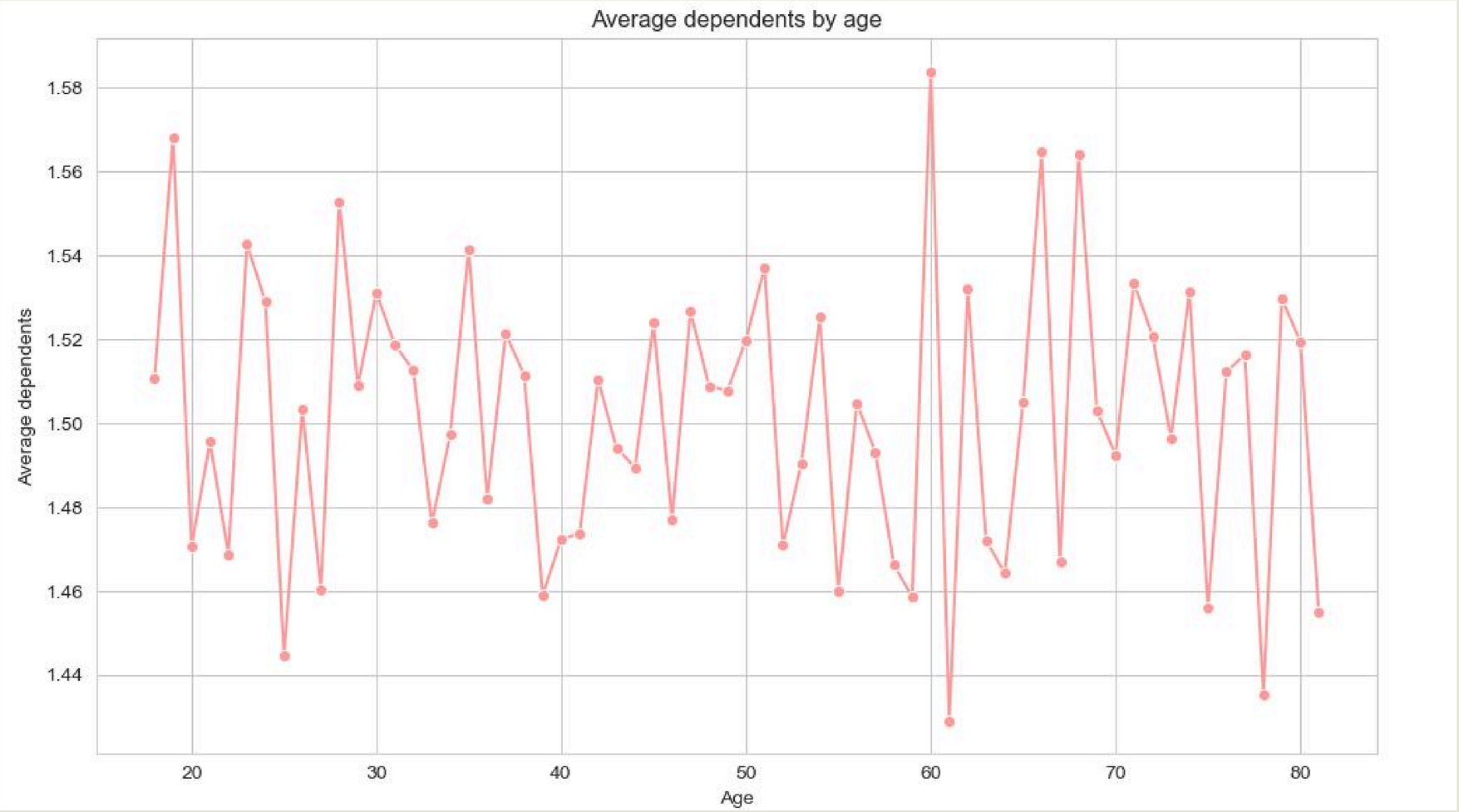
In this case, almost all customers (about 99.7%) are high spenders with average. spending above $10, and only a very small fraction (around 0.27%) are low spenders.

It means marketing and sales efforts should primarily focus on retaining and upselling to the high spender segment, #since they make up almost the entire customer base and likely generate most of your revenue.



DEPENDENCIES

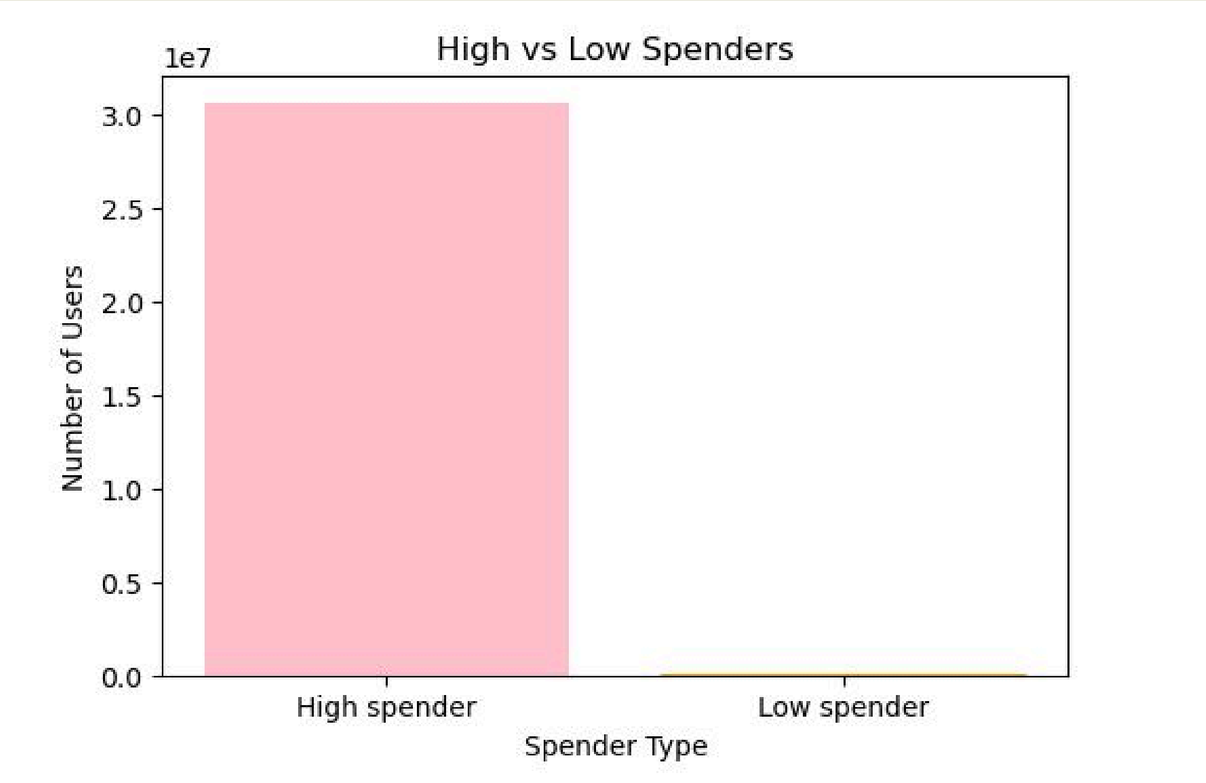
Most age groups cluster around 1.4–1.5 dependents.
Slight increases at early (20s) and later (60) ages may reflect young families and multi-generational households

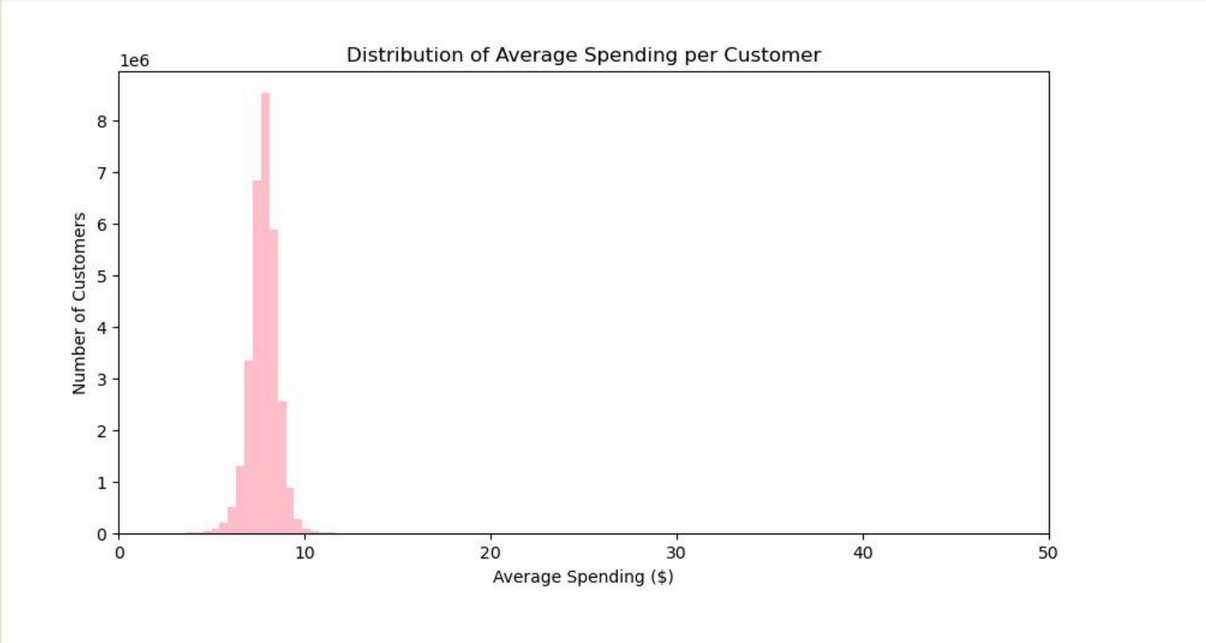


AVERAGE SPENDING

In this case, almost all customers (about 99.7%) are high spenders with average. spending above $10, and only a very small fraction (around 0.27%) are low spenders.

It means marketing and sales efforts should primarily focus on retaining and upselling to the high spender segment,
since they make up almost the entire customer base and likely generate most of your revenue.





Most customers spend consistently around $7.79 per order, with very few outliers spending significantly more or less.

THANK YOU!