Data exploration of grocery purchase data

Introduction

This study aims to analyse the impact of households' adoption to online grocery shopping channels on the healthiness of their total (online & offline) grocery purchases. This report serves to provide information on the dataset (i.e. data structure and key variables) and will present several summary statistics to enhance understanding of the data.

The dataset used for this project takes form of panel data time series that captures detailed purchasing behavior at the household and retailer level for the year 2019 in the Netherlands. Each row represents a specific product purchased by a unique household on a given date, providing insights into individual consumption patterns over time. Key variables include the date of purchase, household and retailer identifier, and product details, such as barcode, quantity purchased and price. The panel structure allows for tracking changes in purchasing habits, retailer preferences, and product choices across time.

Note that the original dataset contains 15 variable, but not every one of those is relevant for this study. In de data preparation phase I will trim and recode the dataset in such a way that best allows for analysis of the study. The table below presents a brief description of the variables relevant for this study.

Variable	Description
Panelist	Unique identifier for each household
Date of purchase	Date on which the purchase was made
Barcode	Barcode of the product purchased
Retailer	Numeric indicator of the supermarket at which the purchase was made
Brand	The brand of the product
Total unit sales	Total units of the product sold
Total value sales	Total value of the sales for the product (in Euro cents)
Total volume sales	Total volume of the product sold
Purchase method	Method of purchase (i.e. offline or online)
category	Category indicator to group products (e.g. vegetables)
Measurement unit	The unit at which the volume of a product is measured
Volume per unit	The volume per unit
segment	A generalized category indicator, based on the 'category' variable

Table 1: Description of variables in dataset

Dataset structure

This research is focused on distribution of purchases by the online and offline channel. The dataset contains purchase data of 150 unique households in the Netherlands in 2019. These households made over 180.000 individual product purchases across 26 different retailers. Table 2 below summarizes some key metrics from the dataset. Out of the 150 households in the panel, 26 have purchased online at least once in 2019. These 26 households combine for just over 6000 online purchases, meaning that approximately 3% of the purchases is made online. The 'online households' are the households of interest for this study, and will serve as the treated group in the difference-in-differences analysis I will perform later on.

Metric	Count
Number of Households	150
Number of Retailers	26
Number of Purchases	182706

Table 2: Summary structure dataset

Sales across retailers

Before diving into the specifics of the purchased products, it is important to in broad terms where and how the household make their purchases. Figure 1 below shows an overview of the total expenditure per retailer. It becomes evident that there are 3 rather dominant retailers: Retailer 4, Retailer 15 and Retailer 17. These retailers combine for nearly 60% of the total expenditure.

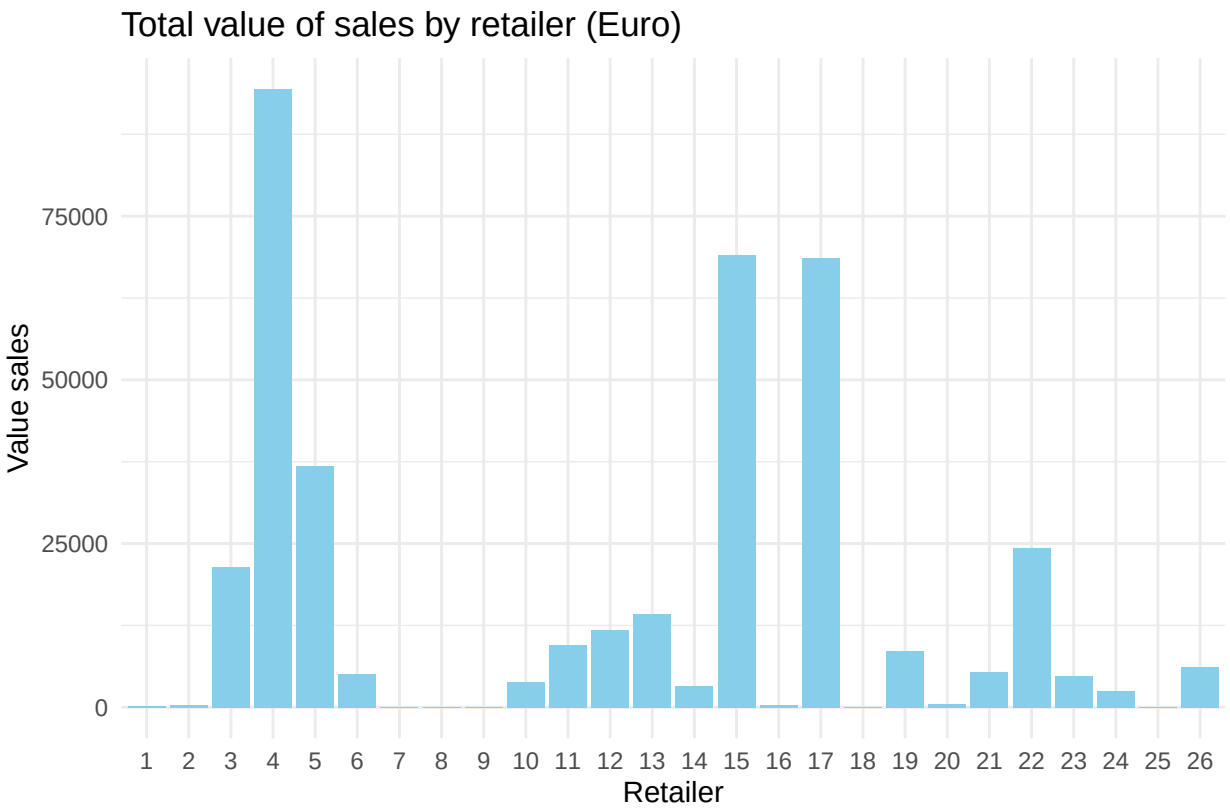

Figure 1: The distribution of total value sales across different retailers

Basket summary statistics

In this project I will aggregate the data to the basket level and eventually look at weekly purchases. The table below presents some general basket summary statistics. As can be seen, households purchase on average 10 products during a shopping trip and spend on average 22 euros per trip.

Mean basket size	Mean expenditure (Euro)	Mean volume (gram)
10.32	22.02	7305.05

Table 3: Summary Statistics of Average Basket Size, Expenditure, and Volume