APPLICATION OF WEB SCRAPED DATA FOR THE CPI

CPI WEBINAR WITH UN REGIONAL HUB FOR AFRICA

WEBINAR - 27 MAY 2025

PETER KAMAU (KNBS) - PRESENTER

SERAH MWIKALI NDUNDA, PENINAH WAITHERA KAMAU, SAGIRE LUCAS, ROBI L. PALUMBO (BANK OF ITALY) AND F. POLIDORO (WORLD BANK)

WORLD BANK GROUP
Poverty & Equity

KNBS
KENYA NATIONAL BUREAU OF STATISTICS
Keeping you informed

OUTLINE

- Introduction
- Web scraping programs for CPI compilation
  - ✓ Common objectives
- Overview and CPI Coverage
- Negotiations and routines to capture price data from the web
- Data treatment and indices elaboration
- Results
- Conclusions

INTRODUCTION

- In July 2023 a training programme on modernization and innovation of CPI was held with the participation, amongst others, of experts from Kenya National Bureau of Statistics (KNBS) and Uganda Bureau of Statistics, UBOS
- This programmme was supported by the World bank and managed by Luigi Palumbo from Bank of Italy and Federico Polidoro from the World Bank and lasted the entire 2024 and the first quarter of 2025

COMMON OBJECTIVES

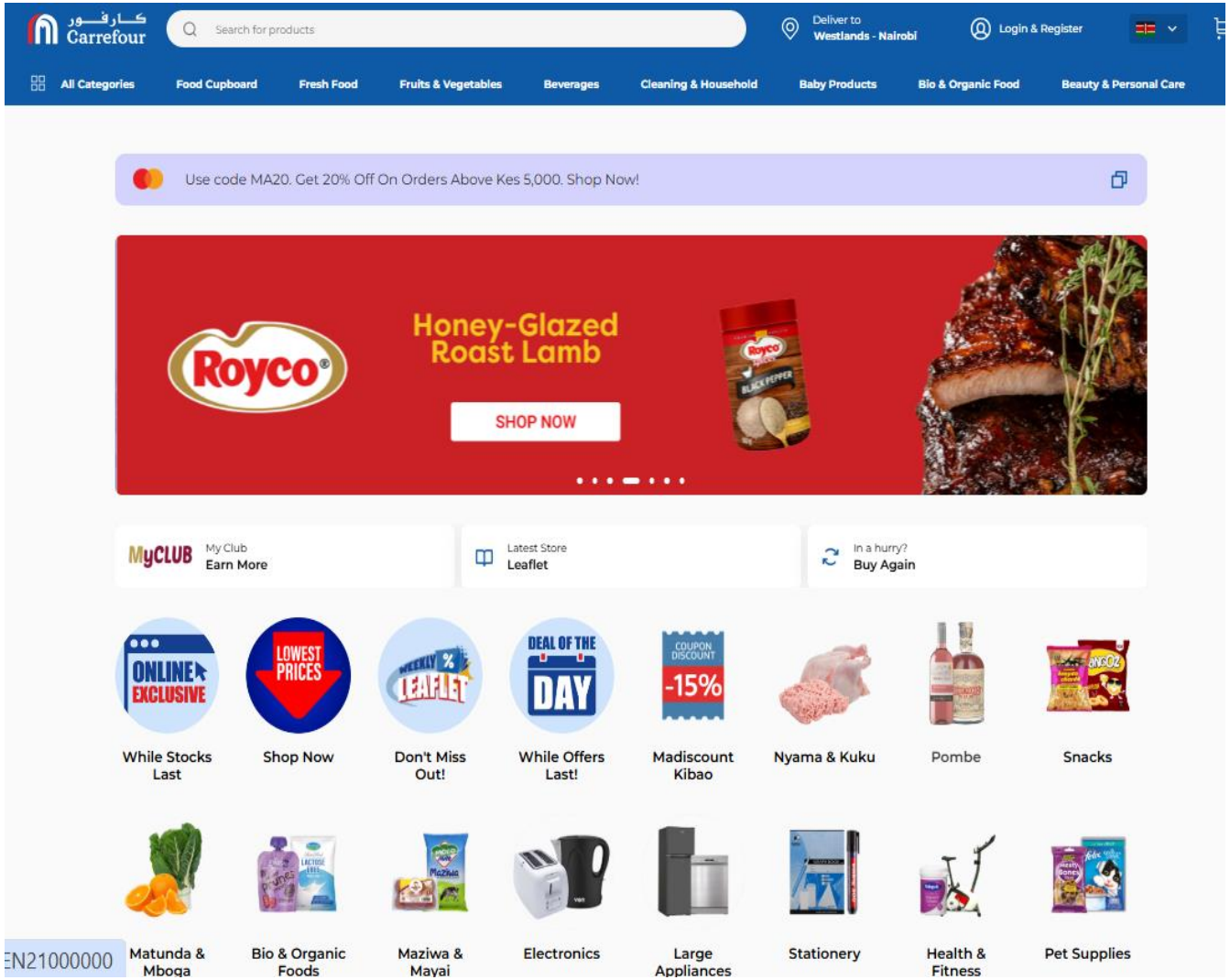
- Familiarizing with the general characteristics of the WS techniques 
- Having practical guidance to implement the use of WS techniques 
- Accessing data on the web through automatic procedures 
- Use web scraping data for CPI production (Implementing web scraping on experimental basis for at least one product in the CPI basket) 
- Familiarizing with the use of Machine Learning and Large Language Models for data processing 
- Consolidate and automate CPI calculation methodology, considering the new data sources 
- Familiarizing with data visualization and presentation 
- Research paper on the project 

Overview of Kenya and CPI Coverage

- 47 counties
- CPI basket items are collected in all the 47 counties

DATA SOURCES

- Web Scraping techniques explored for grocery products (food and products for home maintenance and personal care)
- Targeted approach but addressed to several products
- Carrefour (https://www.carrefour.ke/) supermarkets web sites were inquired



NEGOTIATIONS AND ROUTINES TO CAPTURE PRICE DATA FROM THE WEB

- Informal contacts and a letter sent asking to avoid any block to the KNBS robots
- Use of online robotstxt package to seek for permission to download data from the websites

NEGOTIATIONS AND ROUTINES TO CAPTURE PRICE DATA FROM THE WEB

Kenya – routines to capture price data from the web

- Routines launched more frequently than monthly (from 2 to 5 times a month) to capture grocery price data from Carrefour web site. Continuously from September 2024 (May and June also available)
- Prices in Kenya shillings, and main characteristics of the products captured (brand, variety, package)

IT tools infrastructure

Programming language	R
Code editor	R Studio currently locally hosed but looking into hosting it into a cloud server
Code collaboration environment	Shared folder and GitHub.
Scraping execution infrastructure	Scripts are executed locally but in the future from a server. All machines run windows. Packages and their dependencies will be installed at a go. Such packages include selenium and Rvest The machines are domiciled in the data science lab in the KNBS offices
Scheduling and orchestration infrastructure	Data mining done on a daily basis locally
Monitoring infrastructure	Use of dashboards and error pop ups messages
Data storage infrastructure	Locally in KNBS computers but looking to move it into a cloud storage
Data processing pipeline	Data compiled using an existing CPI compilation platform (objective using R to compile and do quality checks
Database	SQL database hosted on KNBS servers

CODE SNAPSHOT

```r
library(tidyverse)
library(RSelenium)
library(rvest)
library(netstat)
library(data.table)
library(robotstxt)
library(xlsx)
library(stringr)

robotstxt::paths_allowed("https://www.carrefour.ke")

# Start Selenium
free_port <- free_port()
rD <- rsDriver(browser = "firefox", verbose = FALSE, port = free_port, chromever = NULL)
remDr <- rD$client
remDr$open()

# List of categories and their names
categories <- list(
  "https://www.carrefour.ke/mafken/en/c/FKEN1660000" = "Fruits and Vegetables",
  "https://www.carrefour.ke/mafken/en/c/FKEN1701200" = "Rice Pasta and Pulses",
  "https://www.carrefour.ke/mafken/en/c/FKEN1701300" = "Sugar and Home Baking",
  "https://www.carrefour.ke/mafken/en/c/FKEN1710000" = "Biscuits Crackers and Cakes",
  "https://www.carrefour.ke/mafken/en/c/FKEN1714000" = "Canned Food",
  "https://www.carrefour.ke/mafken/en/c/FKEN1720000" = "Breakfast Cereals and Bars",
  "https://www.carrefour.ke/mafken/en/c/FKEN1730000" = "Chips Dips and Snacks",
  "https://www.carrefour.ke/mafken/en/c/FKEN1740000" = "Chocolate and Confectionary",
  "https://www.carrefour.ke/mafken/en/c/FKEN1770000" = "Jam Honey and Spread",
```

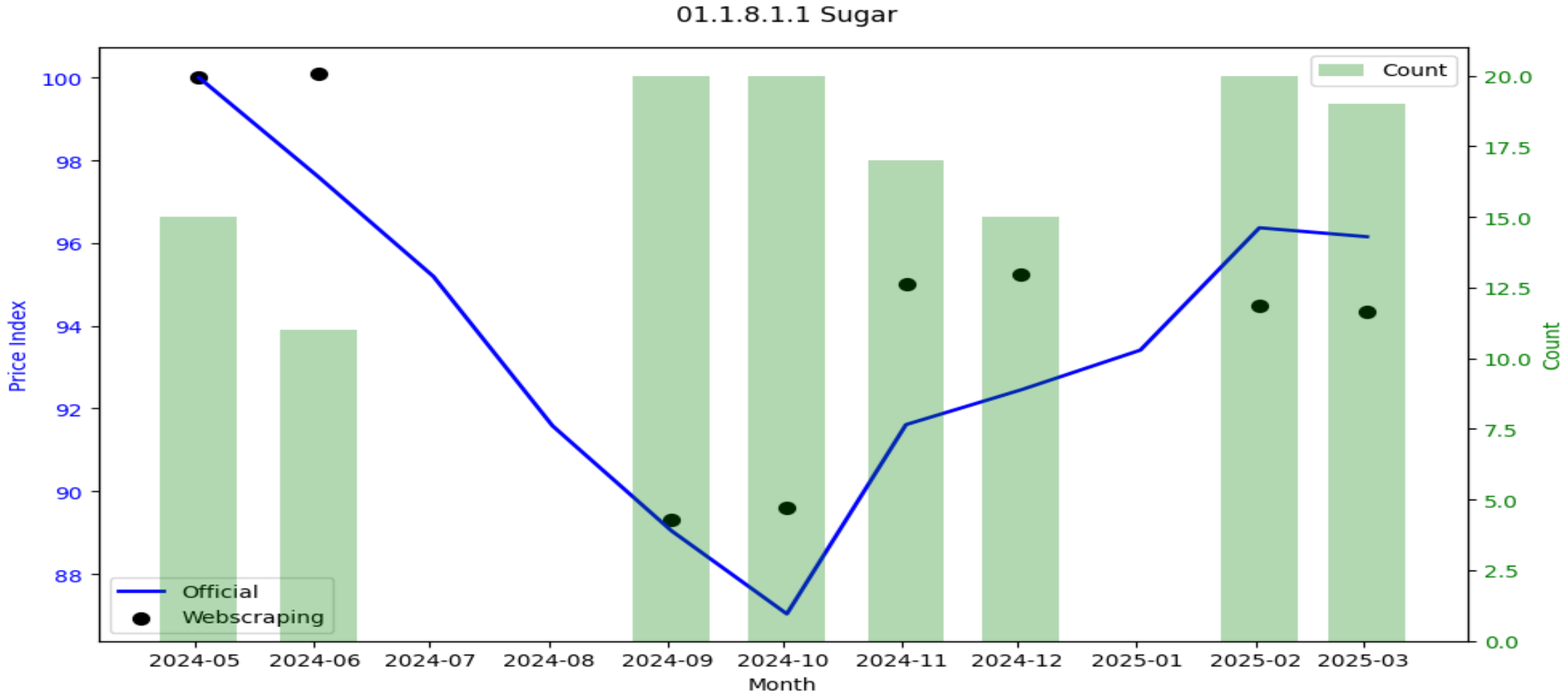
DATA TREATMENT AND INDICES ELABORATION

- 120k elementary prices downloaded for Kenya grocery products in 9 months (100k from Carrefour web site, the remaining from Naivas)

- Unique data sets progressively increasing

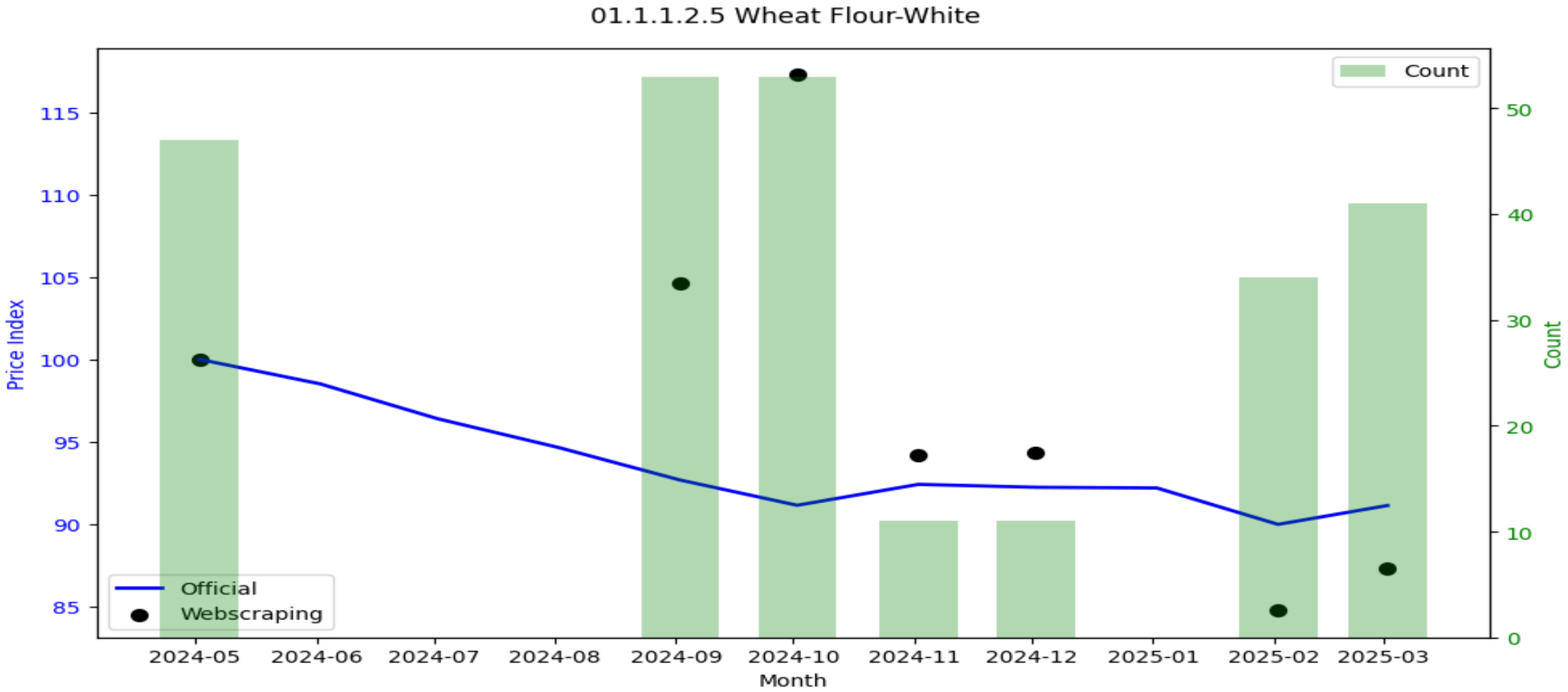
RESULTS - KENYA

Kenya sugar CPI (web scraped prices and official index). Indices (May 2024=100)



RESULTS - KENYA

Kenya Wheat Flour White CPI (web scraped prices and official index). Indices (May 2024=100)



RESULTS

- Good correlation (except for a couple of months for Wheat Flour – White) of the inflation profiles for Sugar and Wheat Flour – White between official indices and those based on the web scraped data (despite some discontinuity in the elementary items data availability in particular for Wheat Flour; reasons to investigate)

CONCLUSIONS AND WAY FORWARD

- The route started is correct and promising despite mixed results
- Reorganizing the data collection for Carrefour (Kenya) on weekly basis and ensuring a good coverage of the routines of the different grocery products
- Design the integration of web scraped data in the current production pipeline considering a general market analysis
- Starting from January 2026 the regular compilation of parallel indices
- Extending the coverage of the web scraping procedures

Thank you for your attention