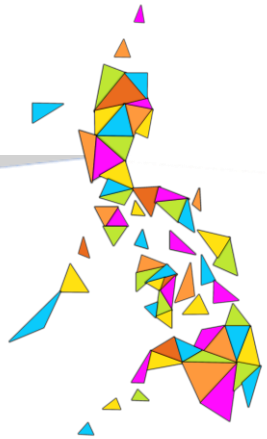
"Harnessing the Power of Data and Statistics for a Future-ready Filipino and Filipina Youth"
2ND PHILIPPINE DATA FESTIVAL
03-04 October 2023 | Century Park Hotel, Malate, Manila

Web Scraping of Commodities for Consumer Price Index in the National Capital Region, Philippines

Session 6: Big Data for Research, Policy and Decision-making

Desiree R. Robles
Senior Statistical Specialist
Philippine Statistics Authority



Outline of the Presentation

- Introduction
- Methodology
- Results
- Issues and Challenges
- Ways Forward

Introduction

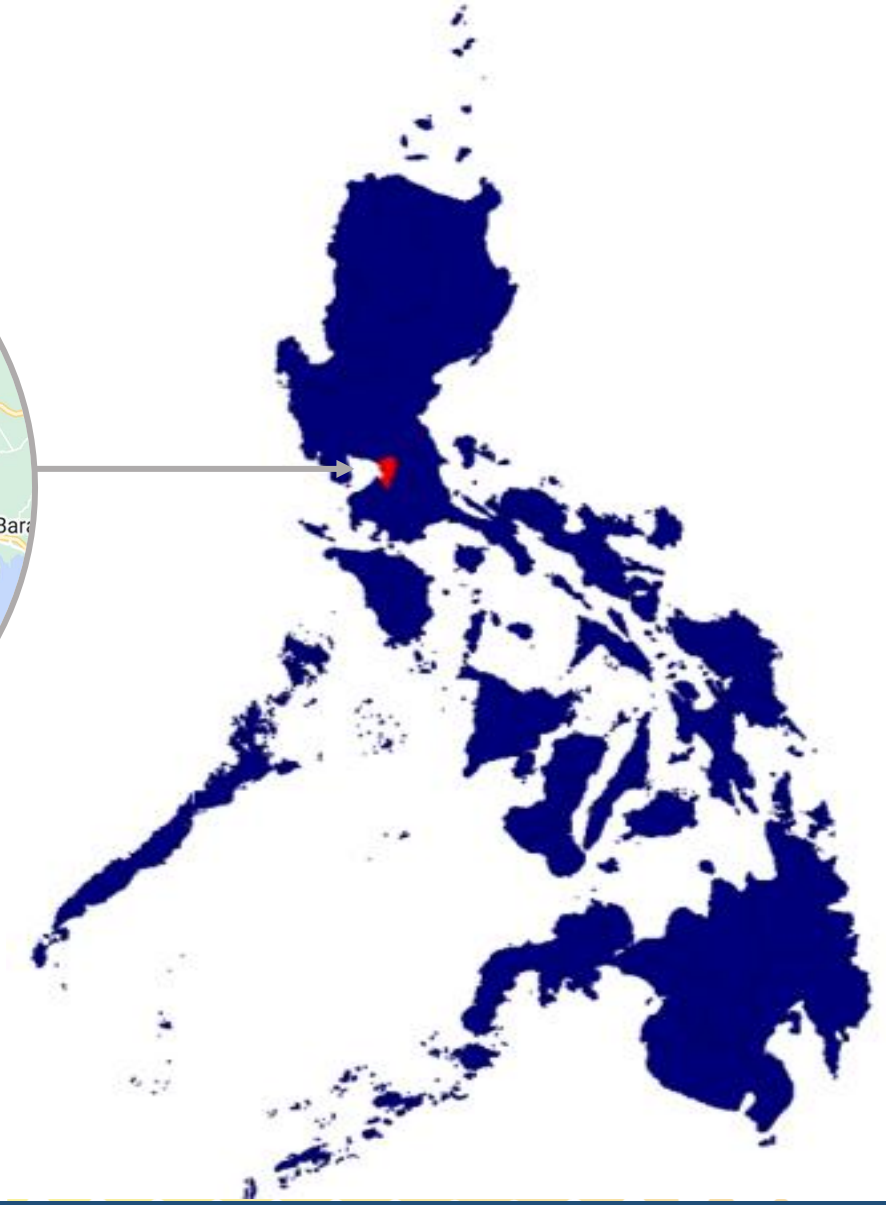
Objectives:

1. To know whether prices collected from websites via web scraping can be used as substitute for the data collected via traditional survey in computing the 2012-based CPI for National Capital Region, Philippines.
2. To be used as benchmark for the use of Big Data for official statistics

Methodology

Geographic Domain:

National Capital Region



Methodology

Frequency of Collection:

Daily (except Saturdays and Sundays)

Methodology

Sample Outlets/Websites:

ABENSON www.abenson.com.ph
Lazada
Shopee
WILCON BUILDING BIG IDEAS
ACE The helpful place.
National BOOK STORE
watsons look good. feel great.
Pharmacy
ZALORA
Anson's ansons.ph
pushkart.ph Online Grocery Delivery
WESTERN APPLIANCES
zagana .com

Methodology

Total No. of URLs Web Scraped: 1,354

Table 1. List of Stores and Number of URLs by Division Code.

Name of Online Stores	No. of URLs	Commodity Division Code									
		01	02	03	04	05	06	07	08	09	11
Total	**1,354**	**402**	**15**	**94**	**38**	**231**	**74**	**5**	**8**	**233**	**254**
Abensons	16					13				3	
Ace Hardware	15				4	11					
Ansons	12					11				1	
Lazada	552	155	11	39	14	87	16	2	3	107	118
National Bookstore	3	1	1				1				
PushKart	23									23	
Shopee	74	65	1			1					7
Watsons	539	151	2	43	14	84	16	3	5	96	125
Western Appliance	45						41				4
Wilcon	17					16				1	
Zalora	16				6	8				2	
Zagana	30	30									

Legend:

01 – Food and Non-Alcoholic Beverages
02 – Alcoholic Beverages and Tobacco
03 – Clothing and Footwear
04 – Housing, Water, Electricity, Gas and Other Fuels
05 – Furnishing, Household Equipment and Routine Household Maintenance
06 – Health
07 – Transport
08 – Communication
09 – Recreation and Culture
11 – Restaurant and Miscellaneous Goods and Services

Methodology

Total No. of Commodities Web Scraped: 517

Table 2. Distribution of Commodities Web Scraped per Division.

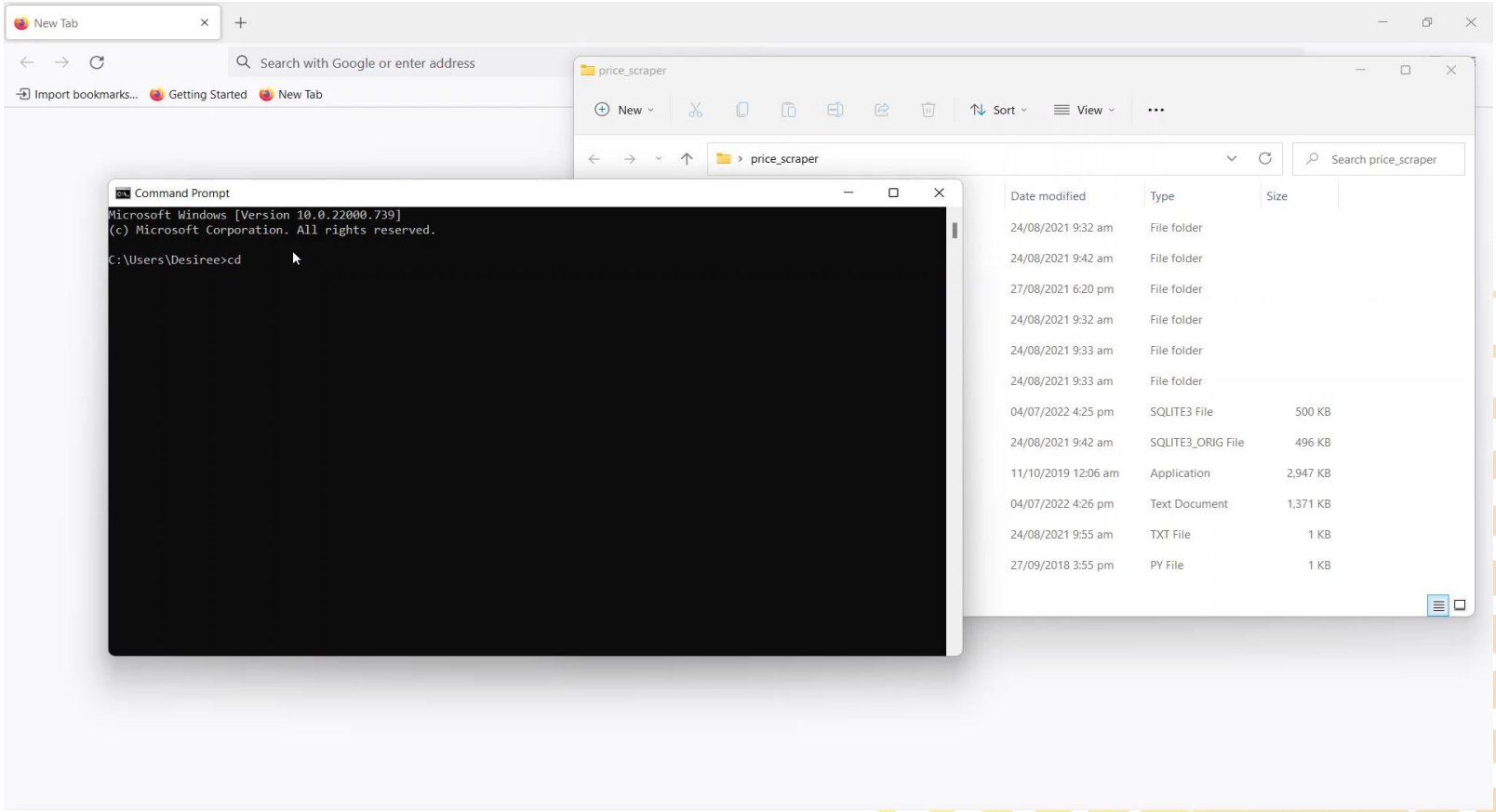
Division	No. of Commodities Web Scraped
01 - Food and Non-Alcoholic Beverages	183
02 - Alcoholic Beverages and Tobacco	10
03 - Clothing and Footwear	41
04 - Housing, Water, Electricity, Gas and Other Fuels	14
05 - Furnishing, Household Equipment, and Routine Household Maintenance	85
06 - Health	41
07 - Transport	2
08 - Communication	2
09 - Recreation and Culture	64
11 – Restaurants and Miscellaneous Goods and Services	75
Total	**517**

Methodology

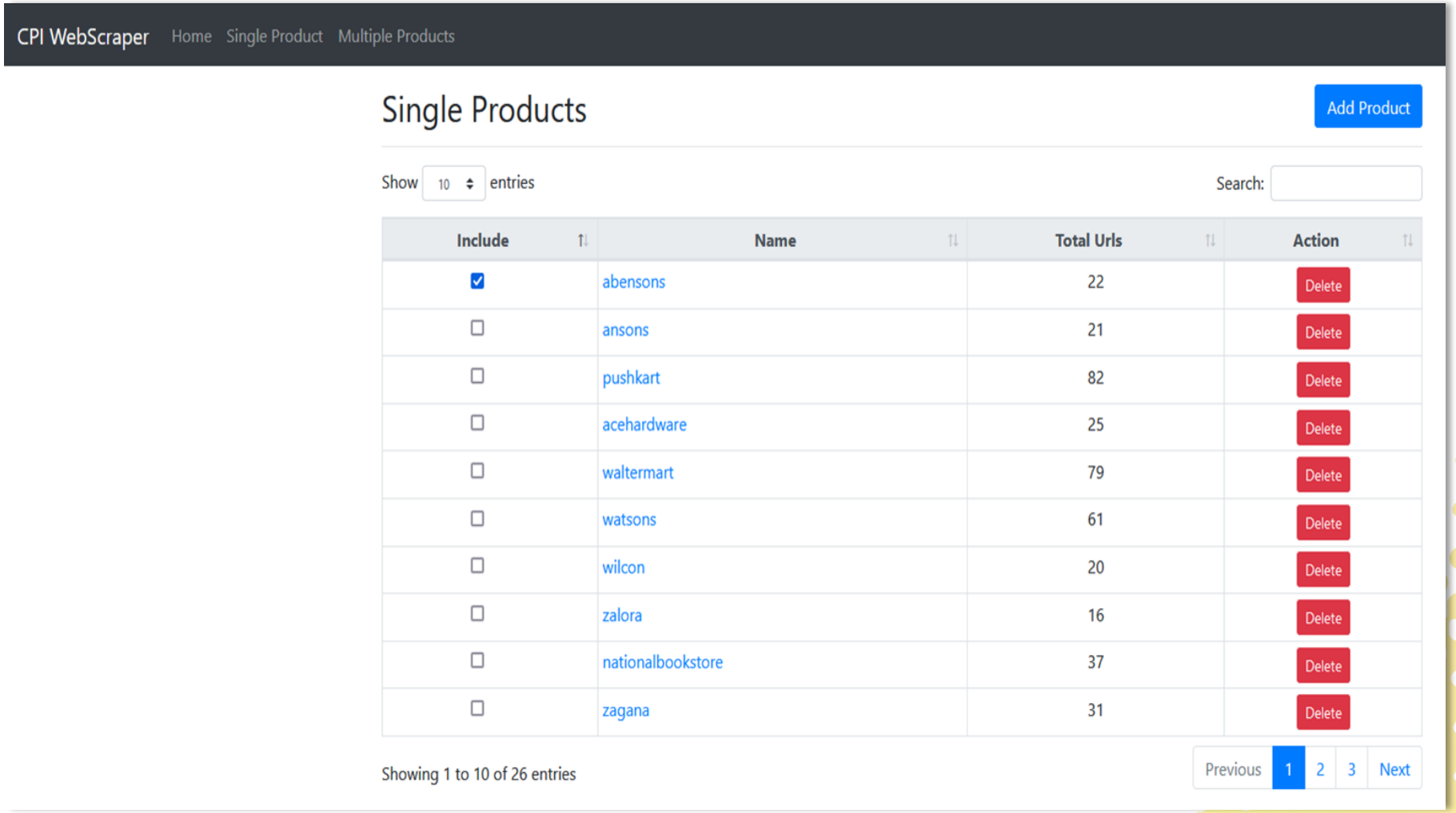
python™

BeautifulSoup

Firefox



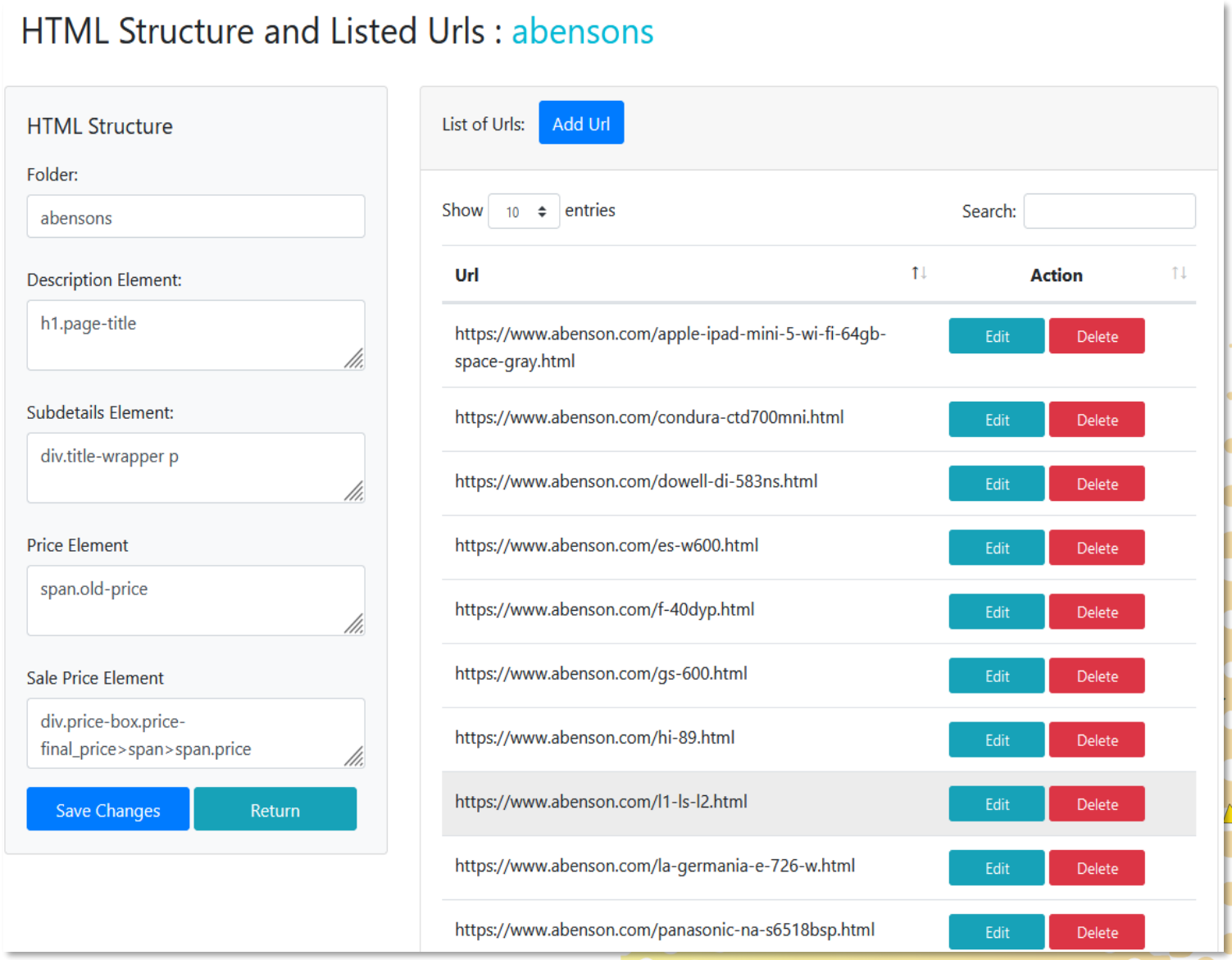
Methodology



Web Scraping Application: Folders

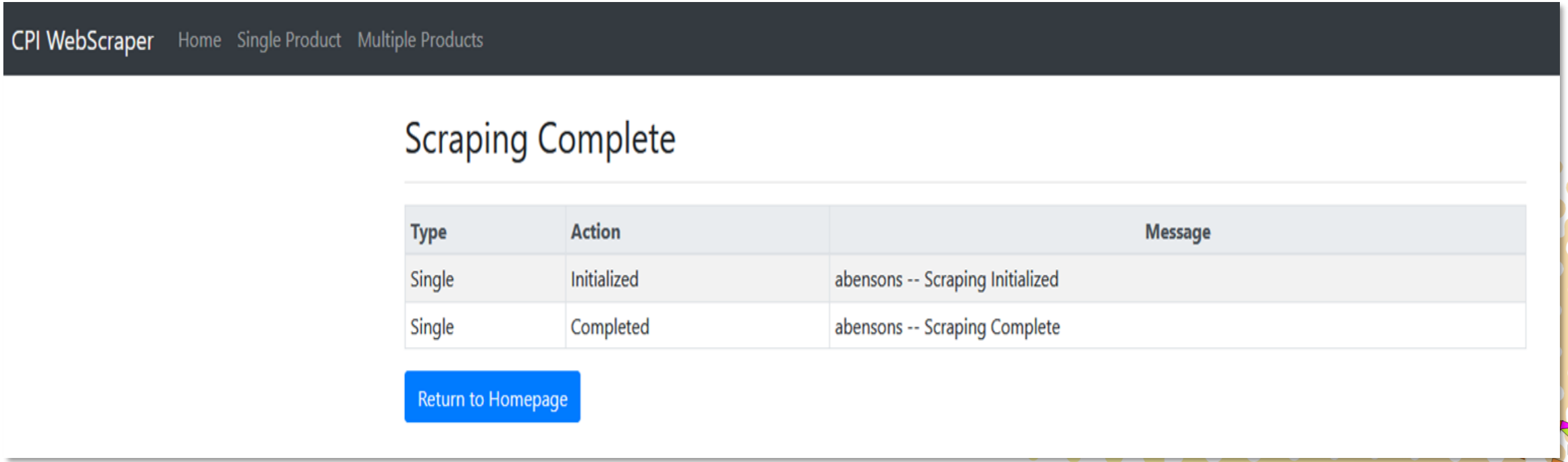
Methodology

Web Scraping Application: HTML Structure



Methodology

Web Scraping Application:
Completion Prompt



Methodology

Web Scraping Application: Sample Output

	A	B	C	D	E	F	G
1	Url	Description	Sub Details	Price	Sale Price		
2	https://wv	APPLE IPAD MINI 5 WI-FI	Item is discontinued.		23,990		
3	https://wv	CONDURA CTD700MNI	Item is discontinued.	19,997			
4	https://wv	DOWELL DI 583NS	SKU 161693	798			
5	https://wv	SHARP ES-W600	SKU 112944	3,997			
6	https://wv	PANASONIC F-40DYP	SKU 56136		1,748		
7	https://wv	HANABISHI GS 600	SKU 3776		648		
8	https://wv	HANABISHI HI-89	SKU 96012		698		
9	https://wv	PANASONIC NA-S6518BSP	SKU 161243	4,799			
10	https://wv	ASAHI RB-6004	SKU 118847		2,098		
11	https://wv	STANDARD SDS 12W	SKU 135746		1,298		
12	https://wv	STANDARD SGS 235S 2B	SKU 136929		1,998		
13	https://wv	SHARP SJ DTH55BS SL	Item is discontinued.	11,697			
14	https://wv	SONY KDL 32R307F	Item is discontinued.	14,499			
15	https://wv	CANON POWERSHOT SX620HS	SKU 144585	15,198			
16	https://wv	TEFAL RK104E	SKU 163548		3,895		
17	https://wv	TEFAL RK7405	SKU 161277		8,995		
18	https://wv	TEFAL RK8145	SKU 161278		10,995		
19	https://wv	LA GERMANIA E-726 W	SKU 170556	6,798			
20	https://wv	TEKNO TKX- 180	SKU 164815	648			
21	https://wv	TEKNO TKX-780	SKU 164814	1,278			
22	https://wv	KELVINATOR WKELH010EA	SKU 147117	18,498			
23							
24							
25							
26							
27							

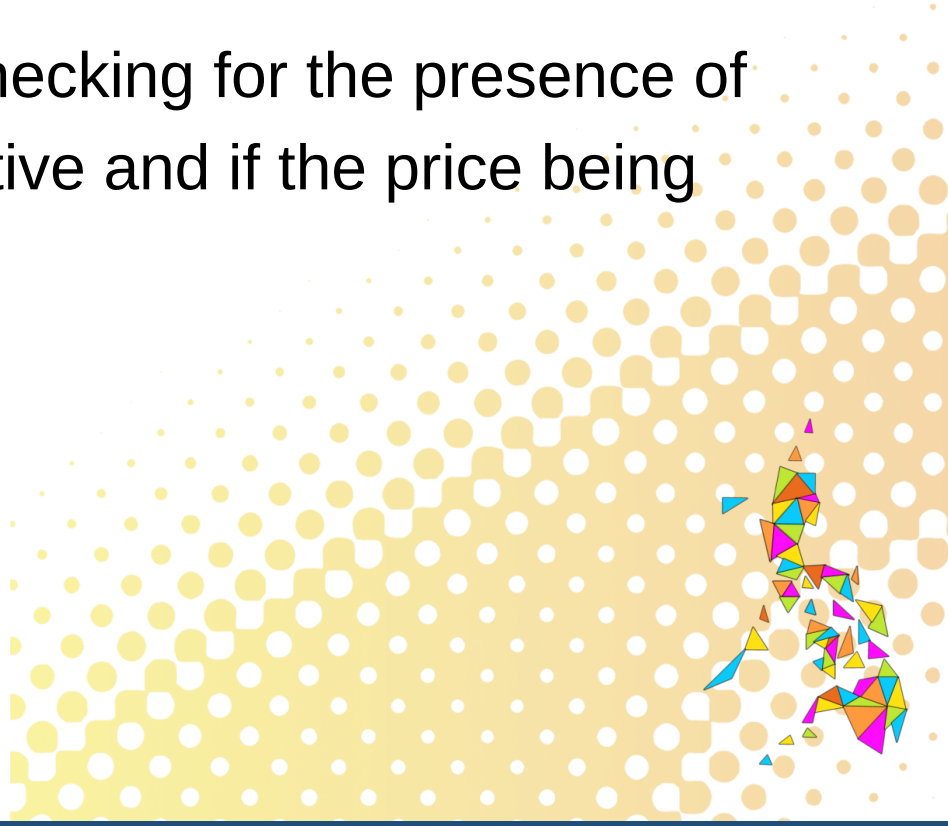
single-abensons-742022

Ready Accessibility: Unavailable

Methodology

Data Processing:

Validations are done daily: consistency checking, checking for the presence of web scraped prices, checking if the links are still active and if the price being collected is correct.



Methodology

CPI Computation:

- Computation of Average Prices, Indices, M-o-M Growth Rate, Y-o-Y Growth Rate follow the official CPI compilation.

- Two types of CPI were computed and compared with the Official CPI:
  - Online – All prices used are collected from websites (web scraped)
  - Hybrid – combination of offline (traditional survey) and online (web scraped) prices.

Results

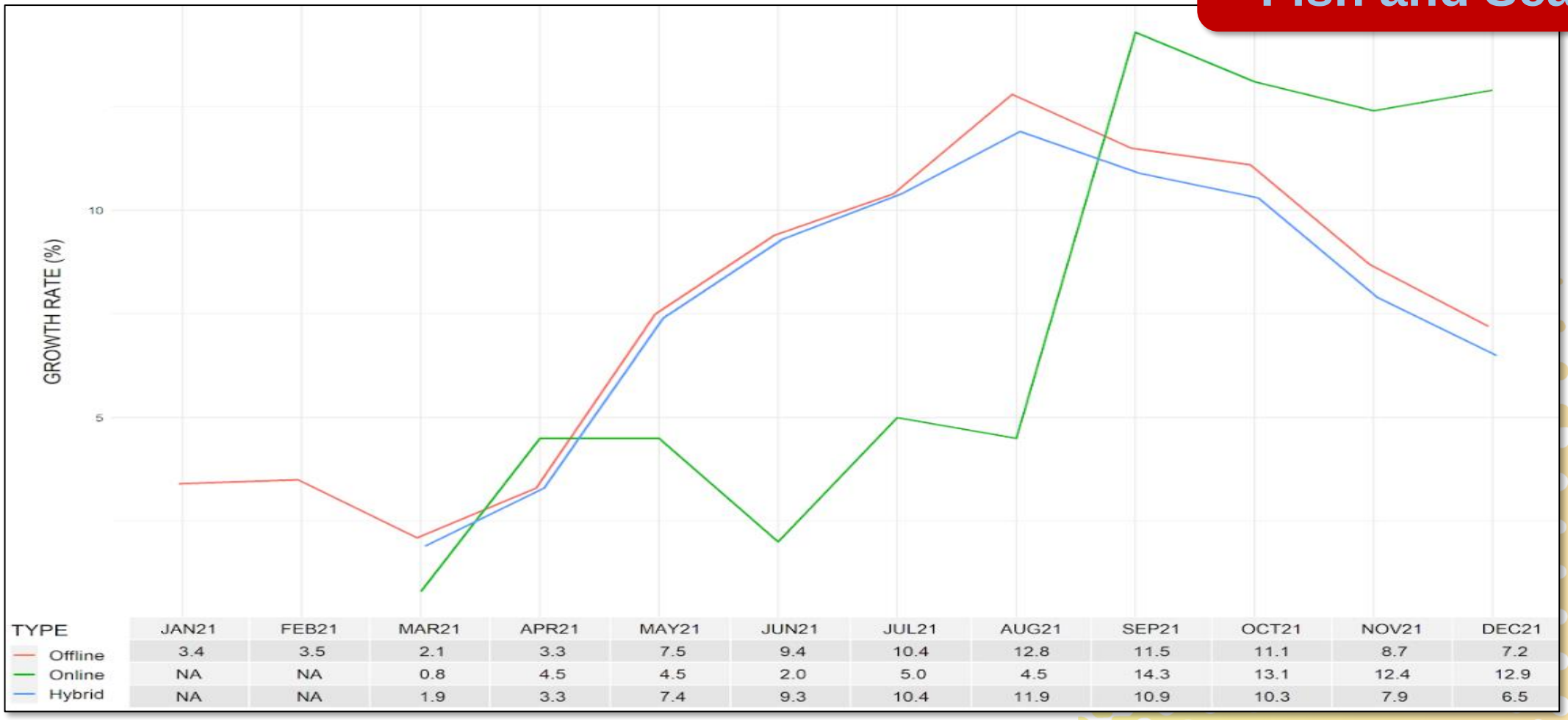
Month-on-Month: Fish and Seafood

TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	-0.4	-1.3	-1.9	-4.8	-2.7	-0.9	0.3	0.2	2.4	4.8	3.9	4.1	-0.3	-2.6	-0.7	-1.0	-1.0	0.0	2.5	-0.9	2.0	2.5	2.5
Online	NA	NA	0.2	0.0	1.5	-2.9	0.5	-8.6	0.5	0.6	-0.4	0.0	9.5	0.7	4.0	0.0	-1.0	0.0	0.0	0.0	-0.5	0.0	0.0
Hybrid	NA	NA	-1.8	-4.8	-2.7	-1.0	0.3	-0.1	2.4	4.7	3.8	4.1	0.0	-2.5	-0.5	-1.0	-1.1	0.0	1.7	-1.0	1.9	2.4	2.5

Figure 1. Month-on-Month Growth Rate of CPI for Fish and Seafood in NCR.

Results

Year-on-Year: Fish and Seafood

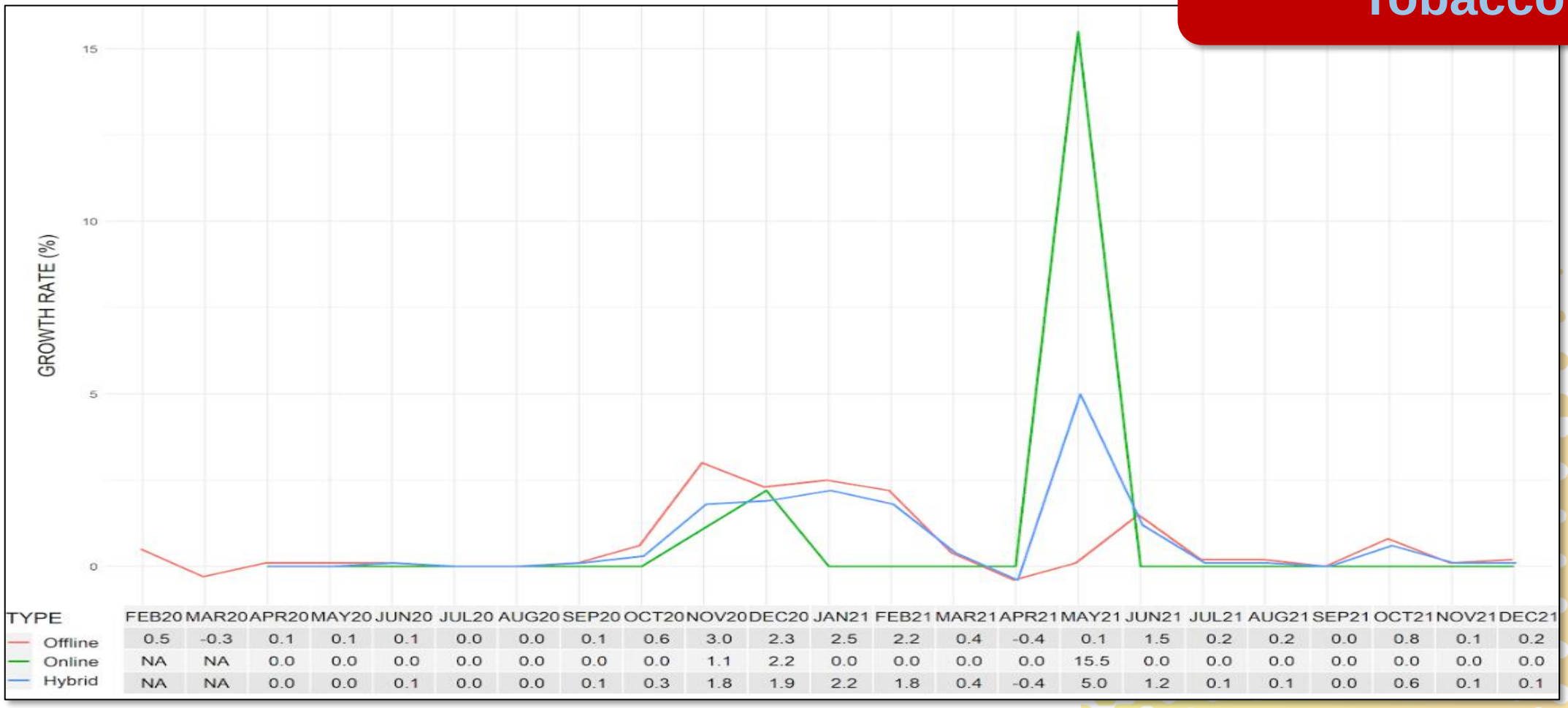


TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	3.4	3.5	2.1	3.3	7.5	9.4	10.4	12.8	11.5	11.1	8.7	7.2
Online	NA	NA	0.8	4.5	4.5	2.0	5.0	4.5	14.3	13.1	12.4	12.9
Hybrid	NA	NA	1.9	3.3	7.4	9.3	10.4	11.9	10.9	10.3	7.9	6.5

Figure 2. Year-on-Year Growth Rate of CPI for Fish and Seafood in NCR.

Results

Month-on-Month: Tobacco

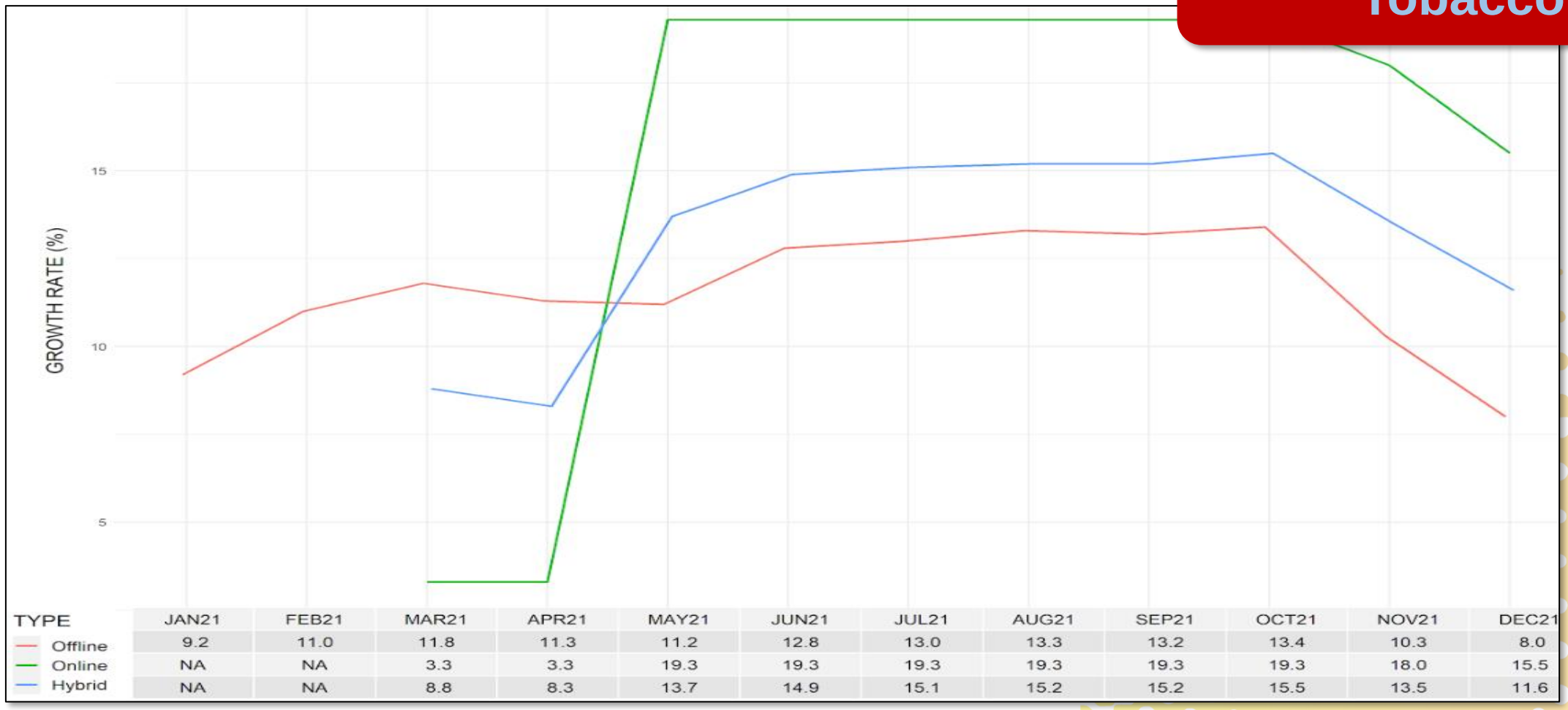


TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.5	-0.3	0.1	0.1	0.1	0.0	0.0	0.1	0.6	3.0	2.3	2.5	2.2	0.4	-0.4	0.1	1.5	0.2	0.2	0.0	0.8	0.1	0.2
Online	NA	NA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	2.2	0.0	0.0	0.0	0.0	15.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hybrid	NA	NA	0.0	0.0	0.1	0.0	0.0	0.1	0.3	1.8	1.9	2.2	1.8	0.4	-0.4	5.0	1.2	0.1	0.1	0.0	0.6	0.1	0.1

Figure 3. Month-on-Month Growth Rate of CPI for Tobacco in NCR.

Results

Year-on-Year: Tobacco

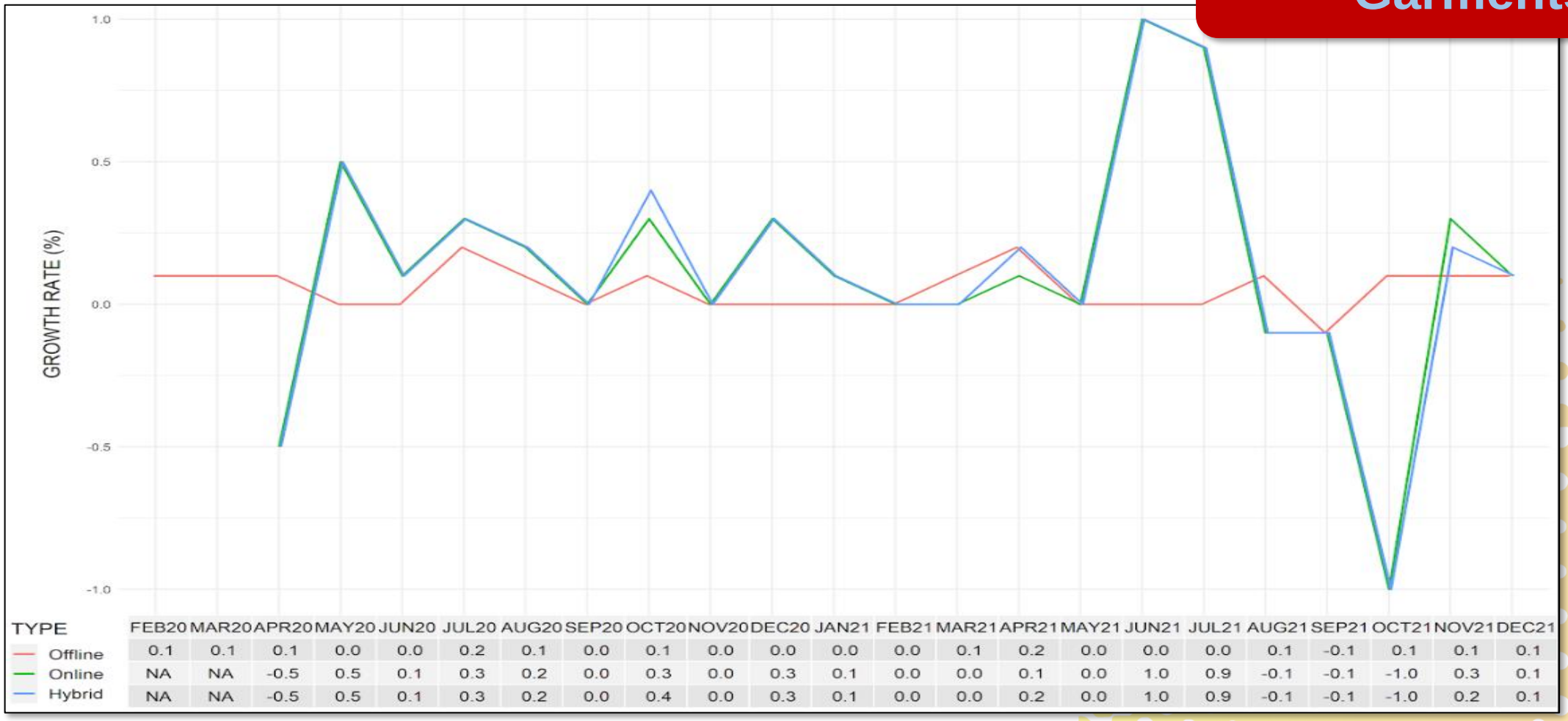


TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	9.2	11.0	11.8	11.3	11.2	12.8	13.0	13.3	13.2	13.4	10.3	8.0
Online	NA	NA	3.3	3.3	19.3	19.3	19.3	19.3	19.3	19.3	18.0	15.5
Hybrid	NA	NA	8.8	8.3	13.7	14.9	15.1	15.2	15.2	15.5	13.5	11.6

Figure 4. Year-on-Year Growth Rate of CPI for Tobacco in NCR.

Results

Month-on-Month: Garments

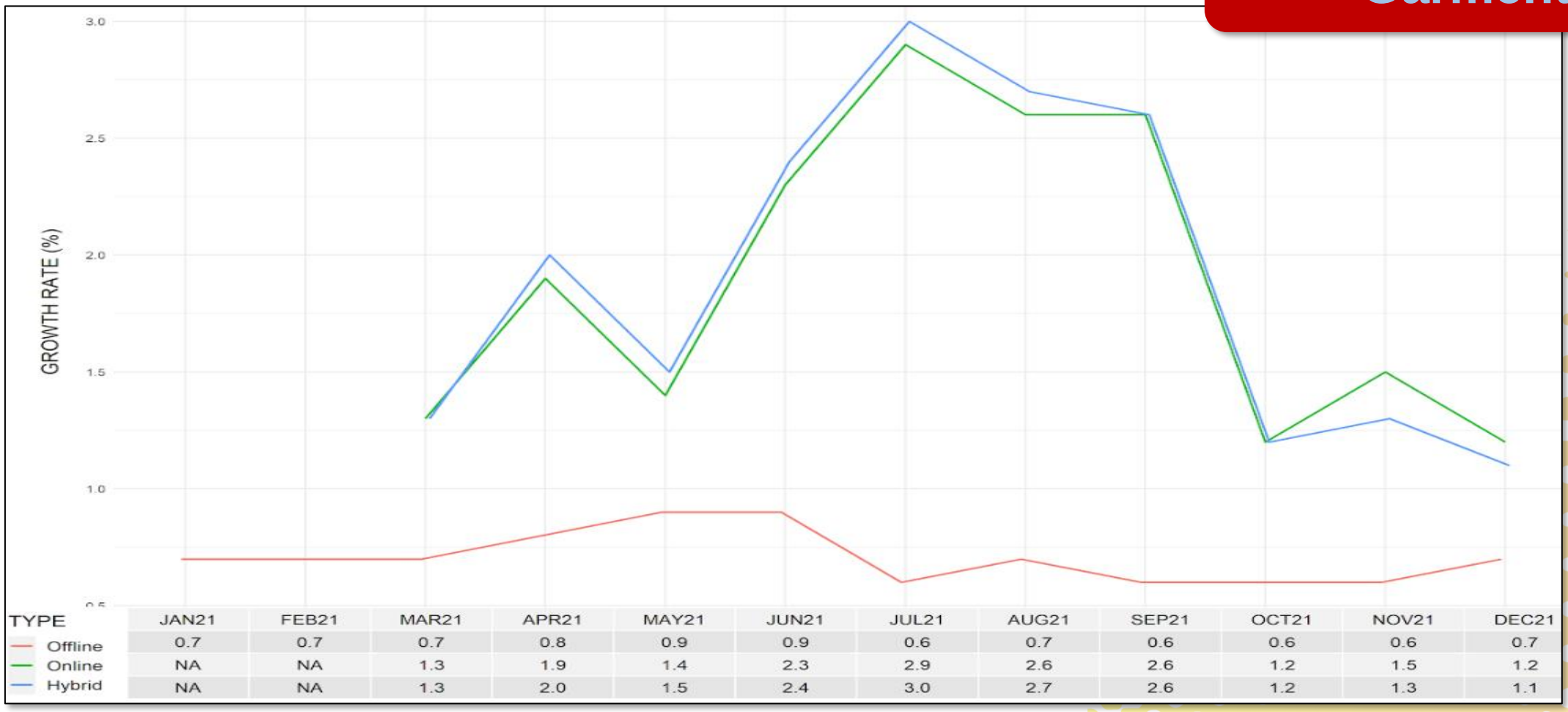


TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.1	0.1	0.1	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.1	-0.1	0.1	0.1	0.1
Online	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.3	0.0	0.3	0.1	0.0	0.0	0.1	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.3	0.1
Hybrid	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.4	0.0	0.3	0.1	0.0	0.0	0.2	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.2	0.1

Figure 5. Month-on-Month Growth Rate of CPI for Garments in NCR.

Results

Year-on-Year: Garments



TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.7	0.7	0.7	0.8	0.9	0.9	0.6	0.7	0.6	0.6	0.6	0.7
Online	NA	NA	1.3	1.9	1.4	2.3	2.9	2.6	2.6	1.2	1.5	1.2
Hybrid	NA	NA	1.3	2.0	1.5	2.4	3.0	2.7	2.6	1.2	1.3	1.1

Figure 6. Year-on-Year Growth Rate of CPI for Garments in NCR.

Results

Unit Root Test

Table 3. Summary of the Unit Root Test by Type of Data Collection for All Items and Division Levels

2009 PCOICOP	Description	Type of Data Collection	Test Statistic at I(0)	Test Statistic at I(1)	Conclusion
0	All items	Offline	-0.945	-3.150**	Stationary at I(1)
		Hybrid	0.241	-2.903*	Stationary at I(1)
01	Food and Non-alcoholic Beverages	Offline	-1.754	-2.765*	Stationary at I(1)
		Hybrid[a]	-1.146	-2.491	Not stationary
02	Alcoholic Beverages and Tobacco	Offline[a]	-1.133	-2.313	Not stationary
		Hybrid	0.297	-3.210**	Stationary at I(1)
03	Clothing and Footwear	Offline	-1.040	-5.057***	Stationary at I(1)
		Hybrid	1.621	-3.899***	Stationary at I(1)
04	Housing, Water, Electricity, Gas and Other Fuels	Offline	0.395	-2.810*	Stationary at I(1)
		Hybrid	0.379	-2.866*	Stationary at I(1)
05	Furnishings, Household Equipment and Routine Household Maintenance	Offline	0.503	-5.714***	Stationary at I(1)
		Hybrid	-0.628	-4.260***	Stationary at I(1)
06	Health	Offline	-0.797	-4.971***	Stationary at I(1)
		Hybrid	-1.175	-4.980***	Stationary at I(1)
07	Transport	Offline	-5.330***		Stationary at I(0)
		Hybrid	-3.303**		Stationary at I(0)
08	Communication	Offline	-0.010	-5.847***	Stationary at I(1)
		Hybrid	-1.281	-5.736***	Stationary at I(1)
09	Recreation and Culture	Offline	-2.650	-4.854***	Stationary at I(1)
		Hybrid	-1.200	-4.490***	Stationary at I(1)
11	Restaurants and Miscellaneous Goods and Services	Offline	0.114	-3.580**	Stationary at I(1)
		Hybrid[a]	-0.412	-2.200	Not stationary

*Note: ***, **, and * indicate 1%, 5%, and 10% significance level, respectively. [a] denotes logarithmic transformation also not stationary.*

Results

Cointegration Test

Table 4. Summary Results of Johansen Cointegration Test Between Offline and Hybrid CPI at the Major Commodity Group Level

2009 PCOICOP	Description	Dependent	Cointegrating Equations	Trace Statistic	Critical Value (α = 0.05)	ECT	Conclusion
0	All items	Offline vs Hybrid	1	15.32	12.53	-0.0208	Not significant cointegration
03	Clothing and Footwear	Offline vs Hybrid	1	19.62	12.53	0.0566***	Diverging series
04	Housing, Water, Electricity, Gas and Other Fuels	Offline vs Hybrid	0	10.27	15.41		No cointegration
05	Furnishings, Household Equipment and Routine Household Maintenance	Offline vs Hybrid	1	14.70	12.53	-0.0167***	Significant cointegration
06	Health	Offline vs Hybrid	0	4.22	15.41		No cointegration
07	Transport	Offline vs Hybrid	1	17.58	12.53	-2.8628***	Significant cointegration
08	Communication	Offline vs Hybrid	0	6.11	15.41		No cointegration
09	Recreation and Culture	Offline vs Hybrid	0	8.84	15.41		No cointegration

Issues and Challenges

1. Websites selected for scraping are not CPI sample outlets. Chosen based on availability of commodities listed in the market basket
2. Not all web scraped commodities have exactly similar specifications with those from the market-basket.
3. Not all of the subclass (5-digit level PCOICOP) and class (4-digit level PCOICOP) have complete commodities.
4. There is an issue with legality and ethics.

Ways Forward

1. Start the web scraping simultaneous with price collection for the new CPI series
2. Collect prices from the websites of the CPI sample outlets

Authors:

Divina Gracia L. Del Prado, Deputy National Statistician
Elena G. Varona (ret.), Chief, Price Statistics Division (PSD)
Desiree R. Robles, Senior Statistical Specialist
Glen G. Polo, Officer-in-Charge, PSD
Rosario S. Lodovice, Statistical Specialist II
Jo Louise L. Buhay, Statistical Specialist I

Thank you!

http://www.psa.gov.ph

http://openstat.psa.gov.ph

https://twitter.com/PSAgovph

https://www.facebook.com/PhilippineStatisticsAuthority