Міністерство освіти і науки України

Західноукраїнський національний університет

Факультет комп’ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Звіт до лабораторної роботи №8

З дисципліни “Інтелектуальний аналіз даних”

Студента групи КН-31

Ковальковського В. В.

Тернопіль 2023

**Попередня обробка часового ряду**

**Завдання:** обрати набір даних, який містить часовий ряд з сайту <https://www.ukrstat.gov.ua/>, та провести попередню обробку. Набір даних, має бути за період в 10 років з кроком 1 місяць та мати не менше трьох параметрів.

**Хід роботи**:

Для цієї лабораторної роботи я обрав набір даних про «Індекси споживчих цін на товари та послуги за 2013 – 2023 рр.». Це показники, що відображають зміни в рівні цін, які споживачі платять за корзину товарів та послуг протягом певного періоду часу. Вони використовуються для вимірювання інфляції, тобто зростання загального рівня цін, що в свою чергу впливає на купівельну спроможність грошей.

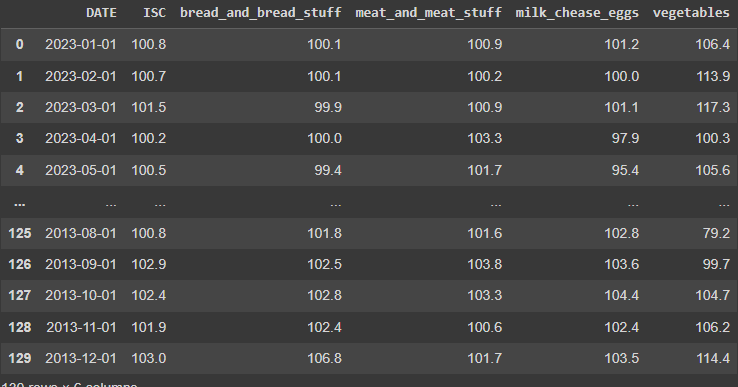


Рис.1 – Набір даних

На рис.1 показано датасет, який використовується у цій лабораторній роботі.

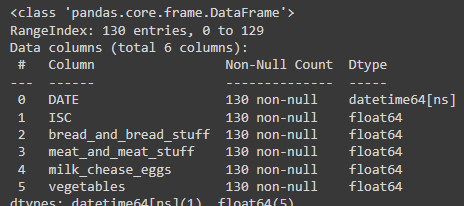


Рис.2 –Інформація про набір даних

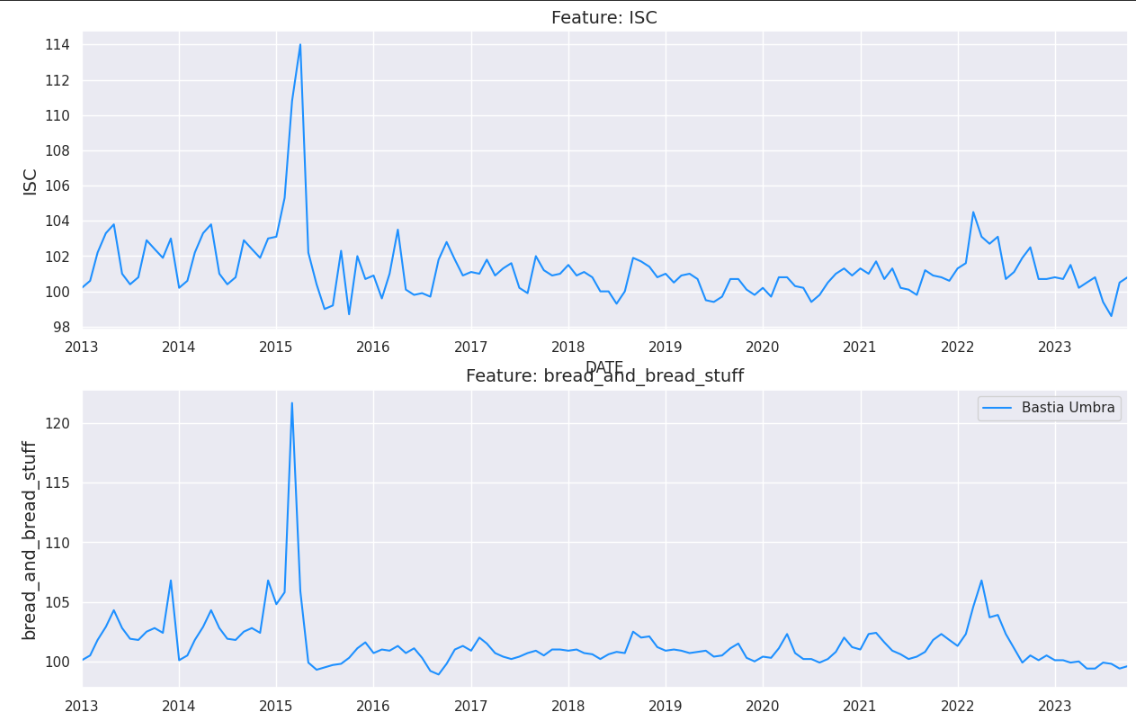


Рис.3 – Графік залежності середнього ІСЦ та ІСЦ хліба і хлібних продуктів

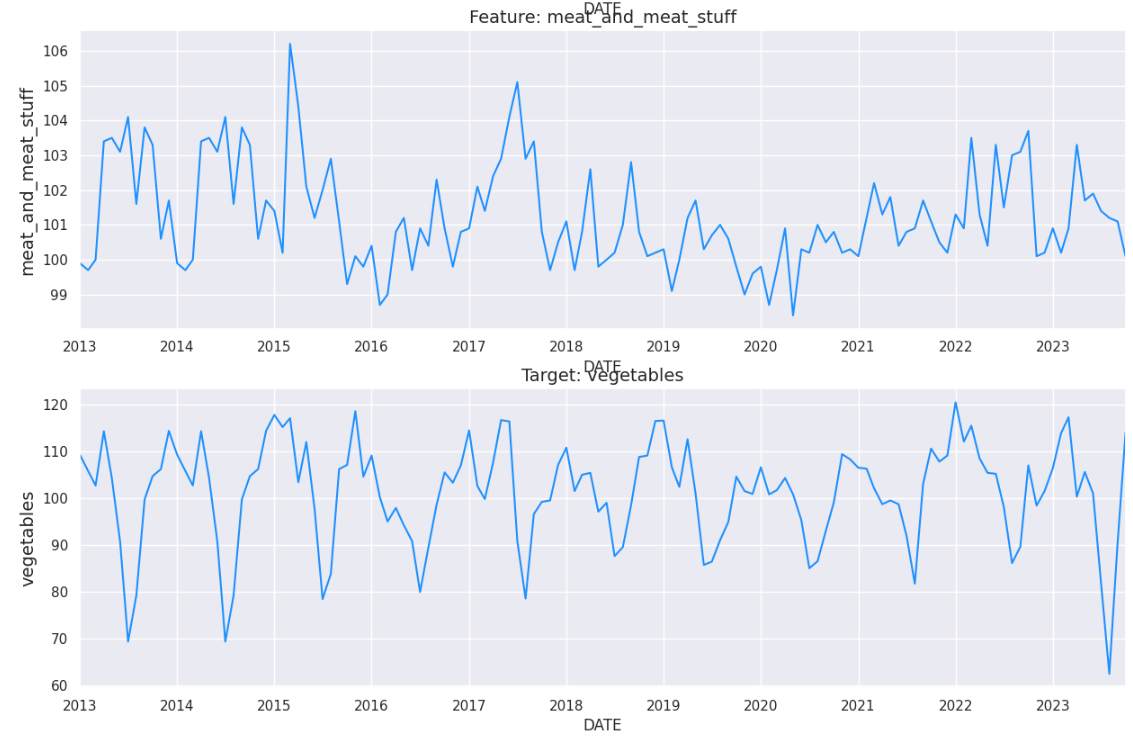


Рис.4 – Графік залежності ІСЦ м’яса та м’ясних продуктів і ІСЦ овочів від часу

На рис.3 та рис.4 зображено графіки залежності наших змінних від часу. З них можна зробити такі висновки:

**ISC (Загальний Індекс Споживчих Цін)**: Спостерігається загальна стабільність із невеликими коливаннями в індексі. Був різкий пік у 2015 році, після чого індекс повернувся до попередніх рівнів.

**Хліб та хлібні вироби:** Індекс для цієї групи також показує стабільність із значним піком у 2015 році, схожим на загальний ІСЦ.

**Молоко, сир, яйця:** Графік для цієї групи товарів має більш виражену волатильність з регулярними коливаннями, але без очевидного довгострокового тренду.

**М'ясо та м'ясні продукти:** Цей індекс показує тренд на зростання з 2013 року з деякими коливаннями. Пік у 2015 році був особливо високим порівняно з іншими роками.

**Овочі (цільовий показник):** Індекс овочів характеризується значною волатильністю з різкими коливаннями від року до року. Немає чіткого тренду зростання або зниження, але є декілька різких спадів, особливо помітний у 2023 році.

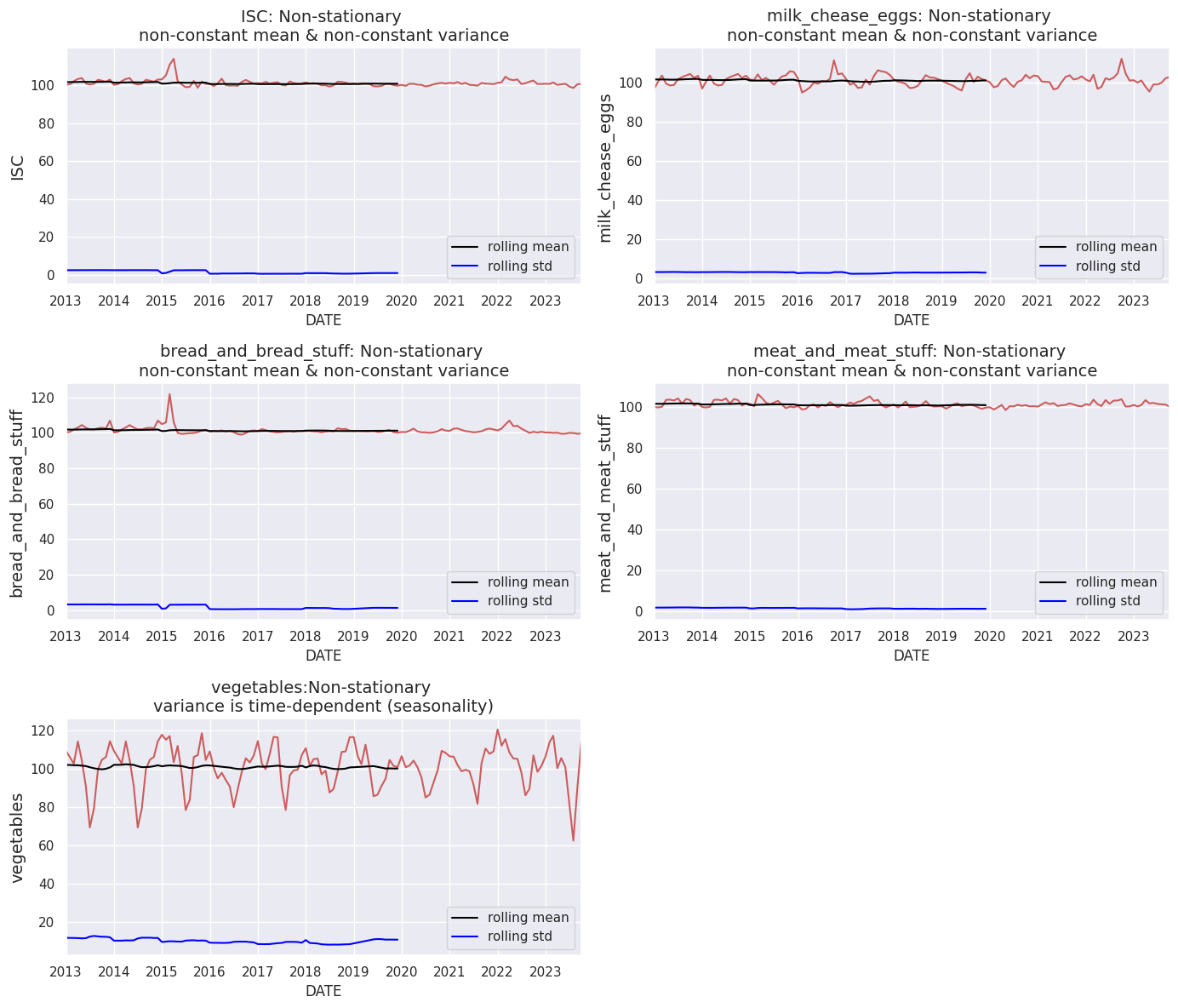


Рис.5 – Перевірка на стаціонарність змінних.

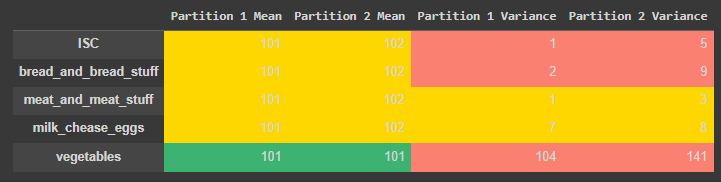


Рис.6 – Перевірка базової статистики

На рис. 6 представлена таблиця з базовими статистиками для різних категорій товарів, розділених на дві партії:

**Середнє (Mean):** Для обох партій середні значення є досить схожими для категорій ISC, bread\_and\_bread\_stuff, meat\_and\_meat\_stuff, і milk\_chease\_eggs, з невеликим збільшенням від 101 у першій партії до 102 у другій. Однак для овочів середнє значення залишається стабільним на рівні 101 для обох партій.

**Варіативність (Variance):** Варіативність значень у другій партії є вищою для всіх категорій порівняно з першою партією. Наприклад, для ISC варіативність зростає з 1 у першій партії до 5 у другій, що свідчить про більший розкид даних. Найвищий рівень варіативності спостерігається у категорії овочів у другій партії з значенням 141, що значно перевищує будь-яку іншу категорію.

Загалом, ці статистики можуть вказувати на розбіжності в змінності та розподілі цін або інших вимірюваних характеристик товарів між двома різними партіями чи періодами часу. Більша варіативність у другій партії може бути ознакою більшої непередбачуваності або розмаїтості у даних.

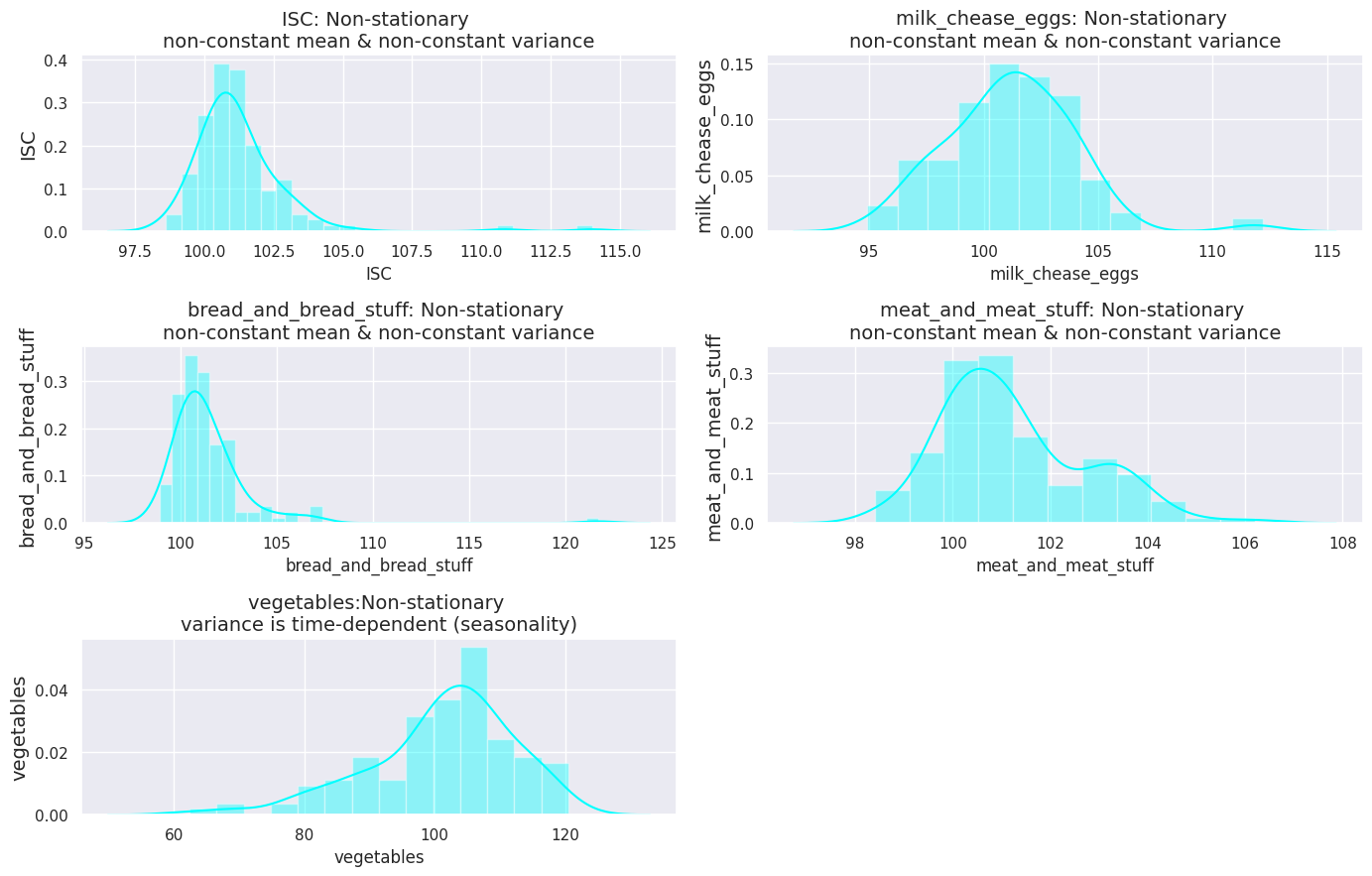


Рис.7 – Гістограми середнього та дисперсії даних.

Оцінимо гістограми на рис.7. Оскільки ми розглядаємо середнє та дисперсію, ми очікуємо, що дані відповідають гаусівському розподілу (дзвоноподібний розподіл) у випадку стаціонарності.

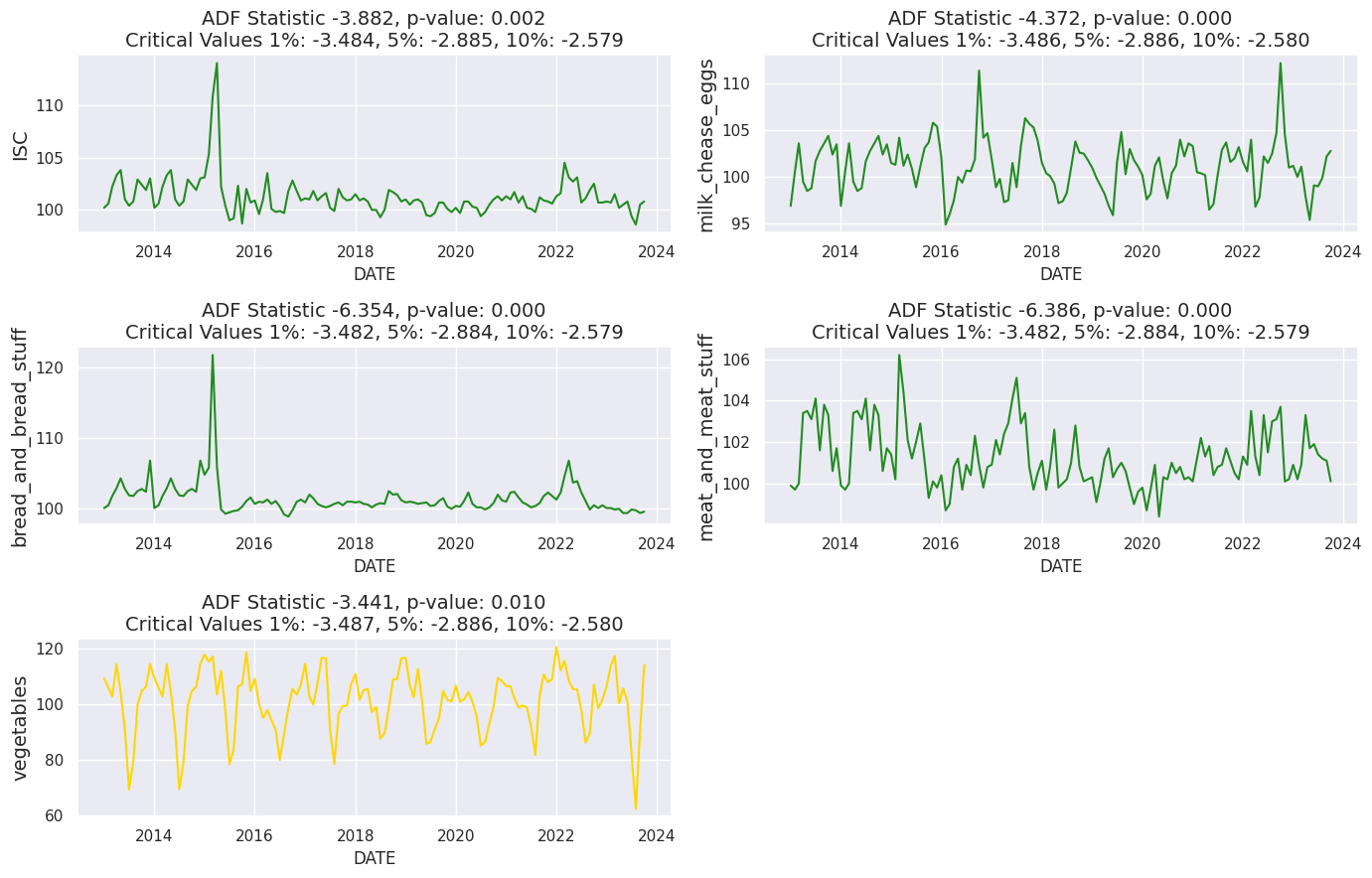


Рис.8 – Розширений тест Дікі-Фуллера

Результати Розширеного тесту Дікі-Фуллера (ADF), які представлені на рис. 8, показують наступне:

* **ISC (Індекс споживчих цін):** ADF статистика становить -3.882, а p-значення — 0.002. Оскільки ADF статистика менша за критичні значення на рівнях значущості 1%, 5% і 10%, і p-значення менше за 0.05, нульову гіпотезу про наявність одиничного кореня (нестаціонарність) відхиляють. Таким чином, дані ІСЦ можна вважати стаціонарними.
* **Молоко, сир, яйця:** ADF статистика становить -4.372, а p-значення — 0.000. Ці результати також вказують на стаціонарність цього часового ряду, оскільки ADF статистика менша за критичні значення, і p-значення дуже мале.
* **Хліб та хлібні вироби:** ADF статистика дорівнює -6.354, з p-значенням 0.000. Результати значно менші за критичні значення, тому нульову гіпотезу про нестаціонарність відхиляють.
* **М'ясо та м'ясні продукти:** ADF статистика дорівнює -6.386, з p-значенням 0.000. Ці результати також підтверджують стаціонарність часового ряду.
* **Овочі:** ADF статистика становить -3.441, а p-значення — 0.010. Оскільки ADF статистика нижча за критичне значення на рівні значущості 1% та 5%, але вища за критичне значення на рівні 10%, можна зробити висновок про стаціонарність часового ряду на рівні значущості 1% і 5%, але не на рівні 10%.

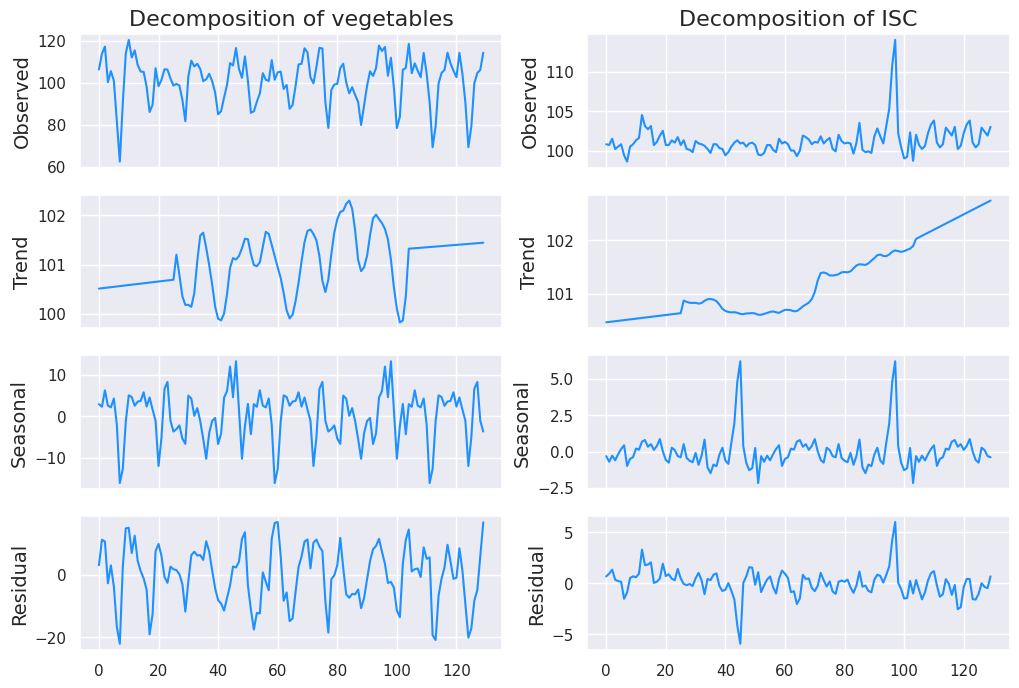


Рис.9 – Декомпозиція змінних Овочі та ІСЦ

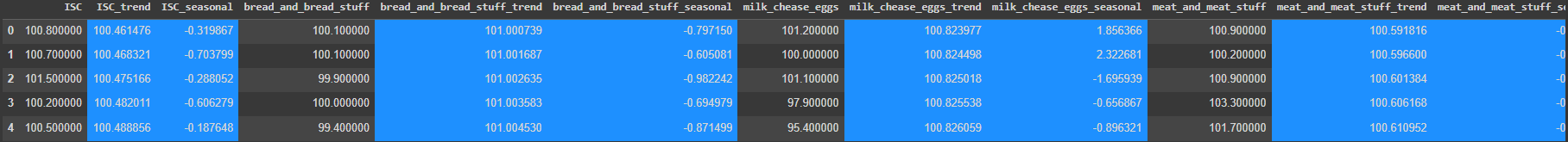


Рис.10 – Вигляд даних після декомпозиції

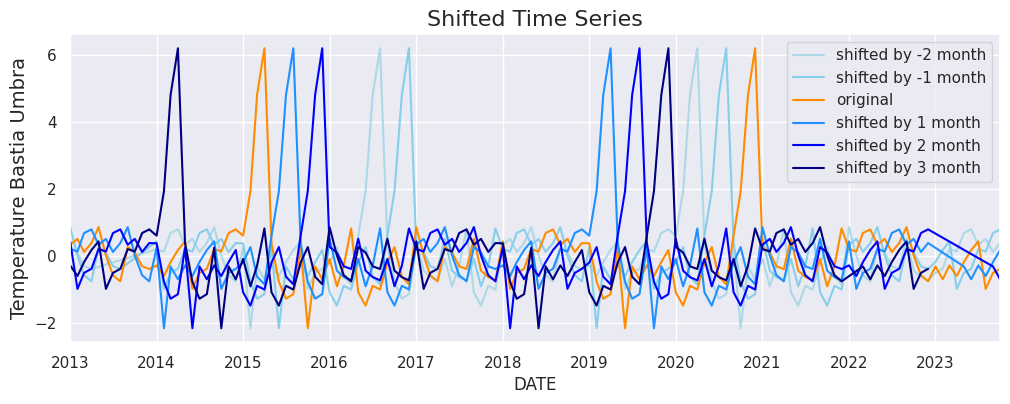


Рис.11 – Визначення лагу

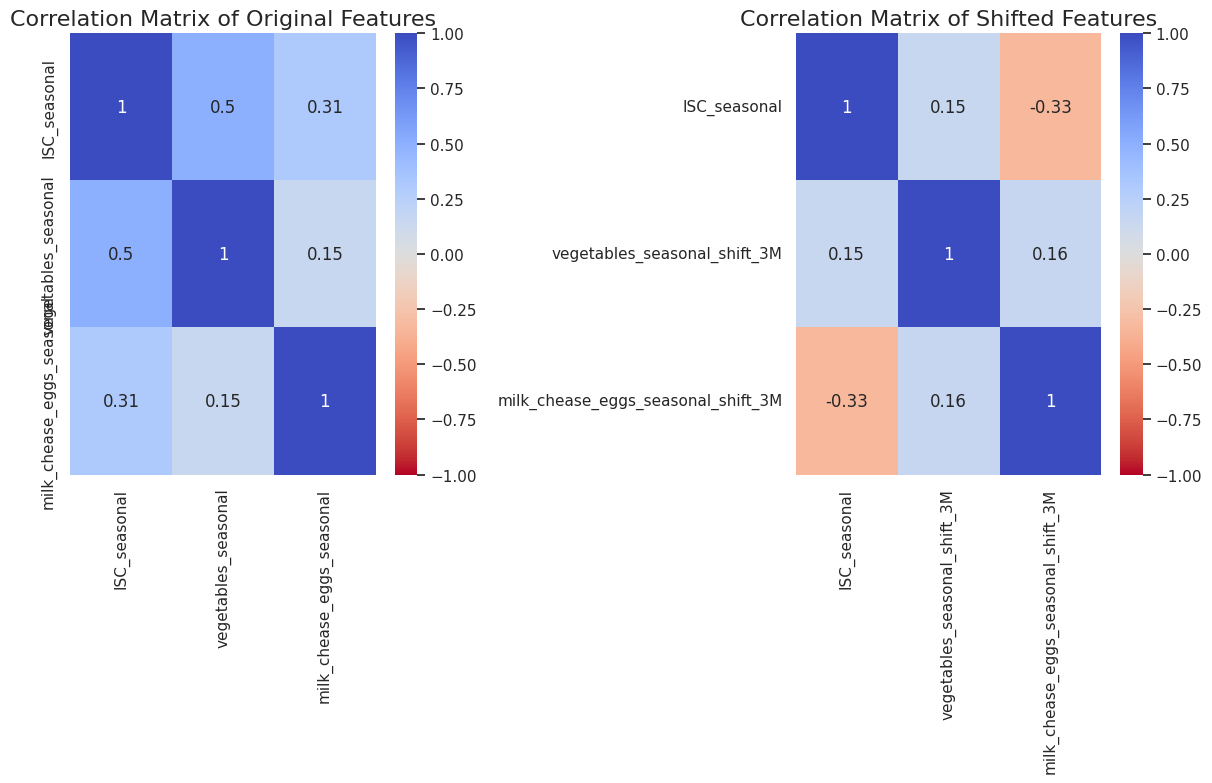


Рис.12 – Кореляція між змінними

На рис.12 ми можемо бачити кореляцію між нашими змінними. З неї можна зробити висновок, що дані мають низьку кореляцію.

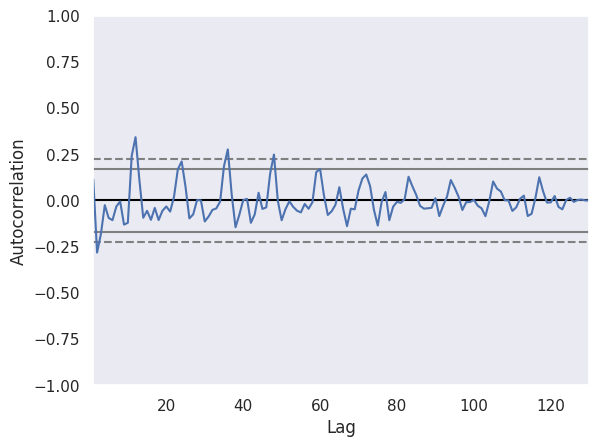


Рис.13 – Автокореляція змінної Овочі

На графіку автокореляції зображено, як значення змінної "овочі" корелюють самі з собою з різними часовими затримками (lags). Автокореляція вимірюється в діапазоні від -1 до 1, де 1 означає ідеальну позитивну кореляцію, -1 — ідеальну негативну кореляцію, а 0 вказує на відсутність кореляції.

Основні спостереження з графіка:

* Коливання в межах конфіденційних інтервалів: Більшість автокореляційних коефіцієнтів знаходяться в межах конфіденційних інтервалів (показані пунктирними лініями), що свідчить про те, що більшість автокореляційних зв'язків можуть бути результатом випадковості, а не мають статистично значущого зв'язку.
* Періодичні сплески: Існують деякі сплески, які виходять за межі конфіденційних інтервалів, особливо на малих затримках. Це може свідчити про наявність короткочасної автокореляції в даних.
* Зниження автокореляції з часом: Загальна тенденція показує, що автокореляція зменшується, коли затримка збільшується. Це зазвичай буває в часових рядах, де вплив попередніх значень швидко зменшується з плином часу.
* Відсутність сильної автокореляції: Жоден з автокореляційних коефіцієнтів не наближається до 1 або -1, що вказує на відсутність сильної довготривалої автокореляції в даних змінної "овочі".

Такий розподіл автокореляційних коефіцієнтів може бути ознакою неперіодичної сезонності або інших нестаціонарних елементів у поведінці часового ряду для змінної "овочі".

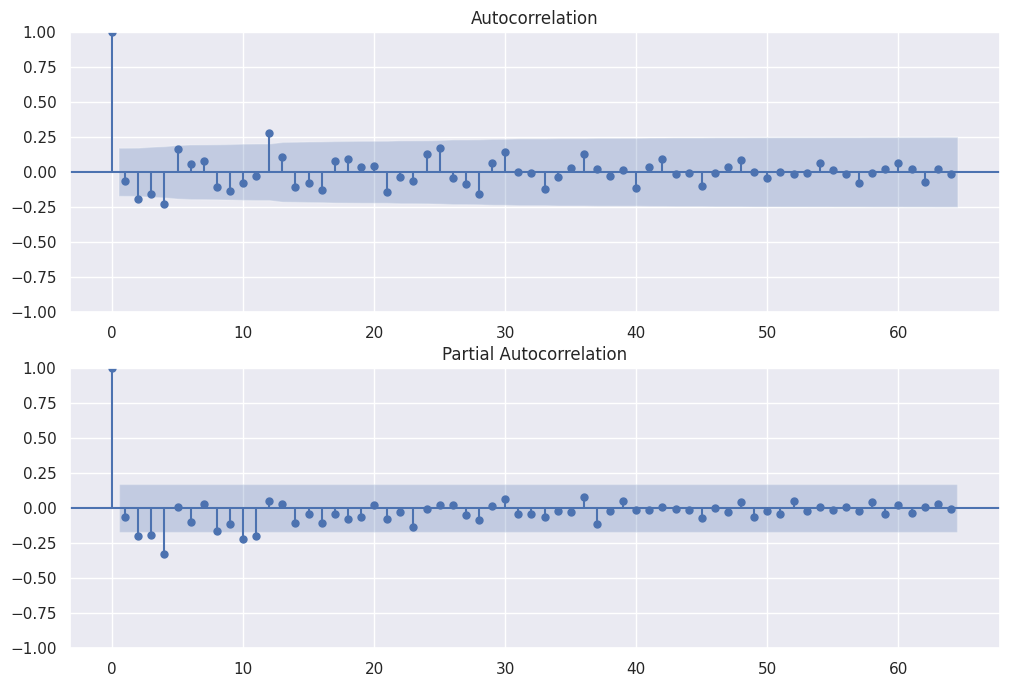


Рис.14 – Автокореляція та часткова автокореляція змінної Овочі

Ми бачимо деяку синусоїдальну форму як у функції ACF, так і в функції PACF. Це свідчить про те, що присутні як АР, так і МА процеси.

**Висновок:** під час лабораторної роботи я обрав набір даних, який містить часовий ряд з сайту <https://www.ukrstat.gov.ua/>, провів попередню обробку та аналіз даних.