Міністерство освіти і науки України
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної

техніки Кафедра інформатики та програмної

інженерії

Звіт

з лабораторної роботи №7 з дисципліни
«Аналіз даних в інформаційних системах»

«Аналіз часових послідовностей»

Варіант 7

Виконав студент ІП-12 Васильєв Єгор Костянтинович
(шифр, прізвище, ім'я, по батькові)

Перевірив Ліхоузова Тетяна Анатоліївна
(прізвище, ім'я, по батькові)

Київ 2023

Лабораторна робота № 7

Тема: Аналіз часових послідовностей

Для виконання лабораторної роботи було обрано мову програмування Python. Спочатку дані про кількість нових захворювань covid-19 по країнам було завантажено та імпортовано до датафрейму, а після дослідження, була здійснена їх обробка для подальшої роботи: видалено усі стовпці окрім дати, країни та кількості нових захворювань; датафрейм було переформатовано наступним чином: дата була використана в якості індексу, а у двох стовпях «Франція» та «Італія» містилась відповідна кількість нових випадків. Також, пропущені значення були замінені сусідніми валідними значеннями. Після здійснення цих дій датафрейм отримав наступний вигляд:

Рисунок 3.1 – дані у data frame

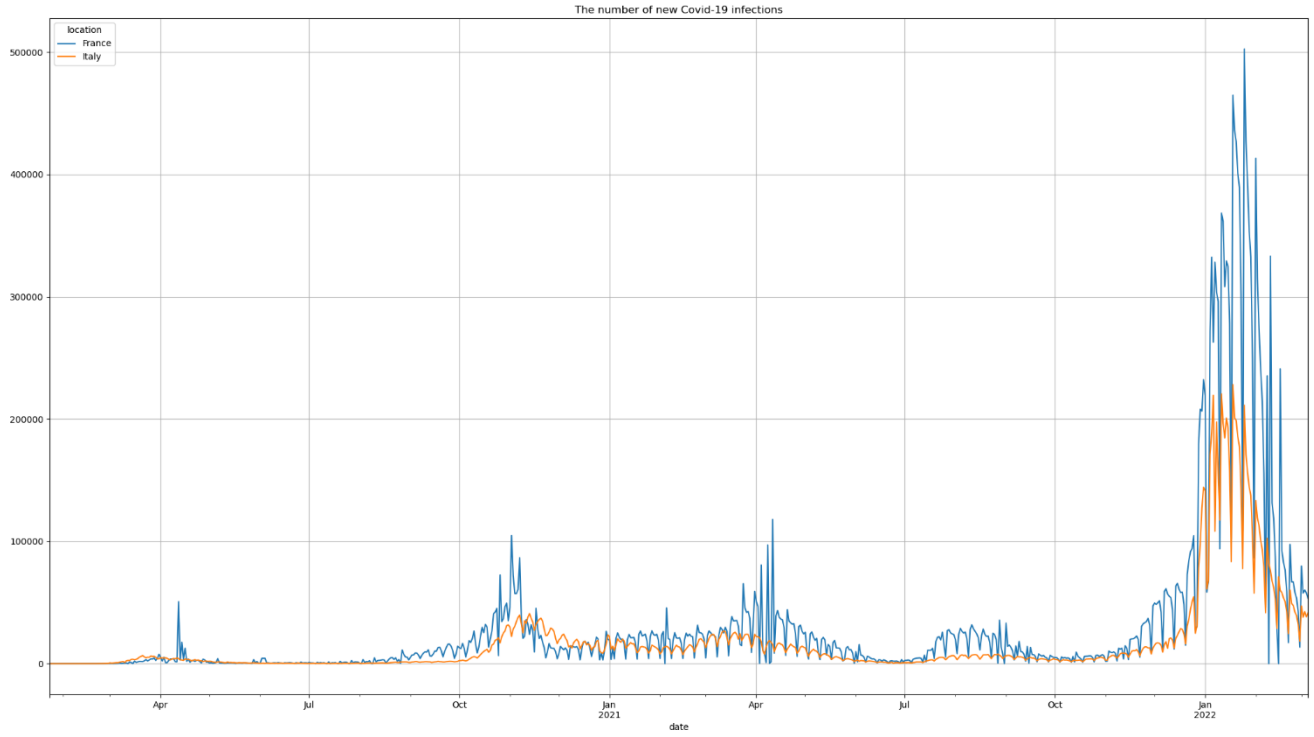

Рисунок 3.2 – порівняння кількості захворювань в Італії та Франції

Для зручності подальшої роботи дані були розділені окремо на захворювання в Італії та Франції:

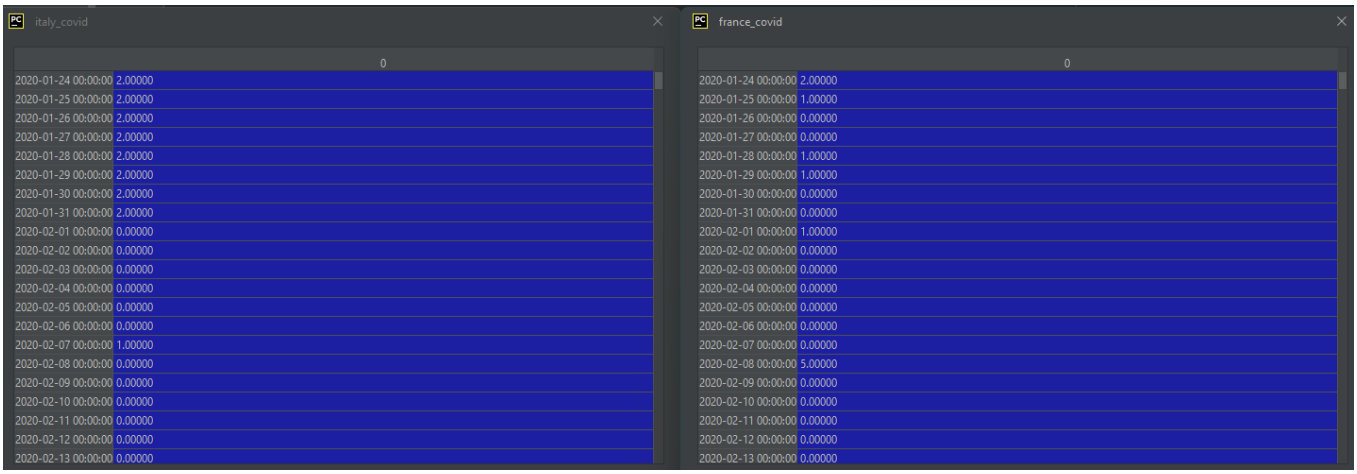

Рисунок 3.3 – data frame з даними по Італії та окремо по Франції

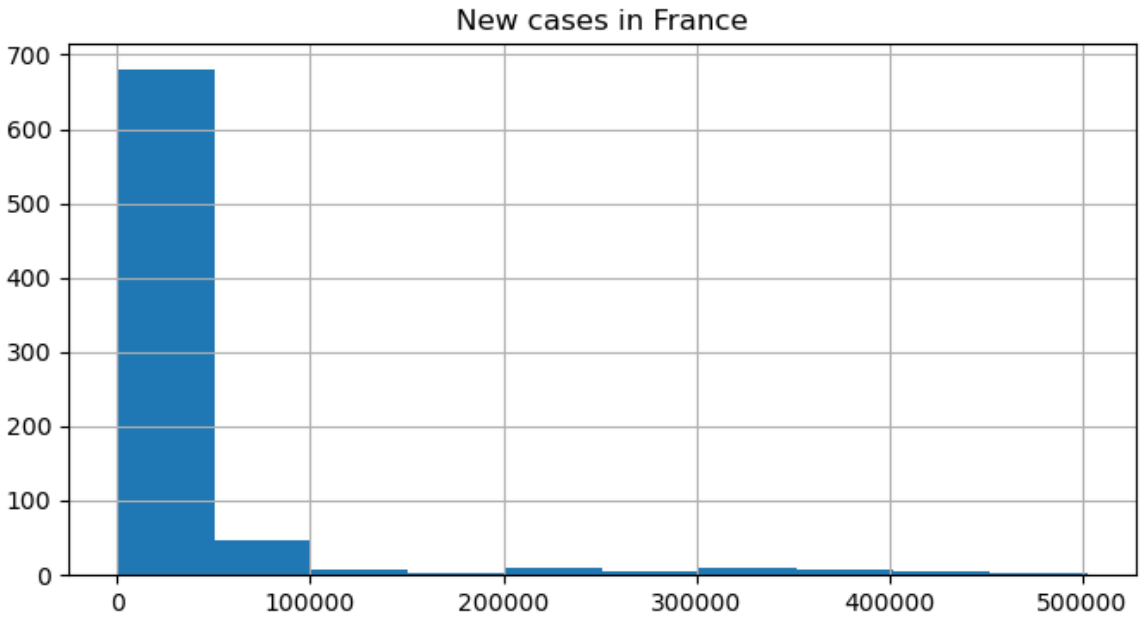

Рисунок 3.4 – кількісна гістограма по захворюванням Франції

Для кращої візуалізації властивостей часового ряду було застосовано згладжування за допомогою ковзаючого середнього:

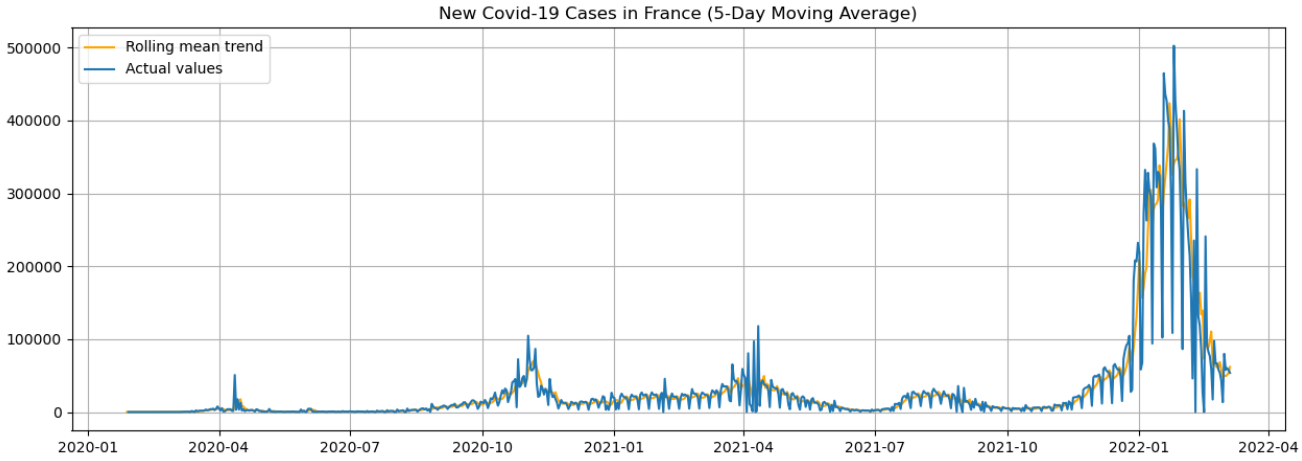

Рисунок 3.5 – згладжування ковзаючим середнім з window=5

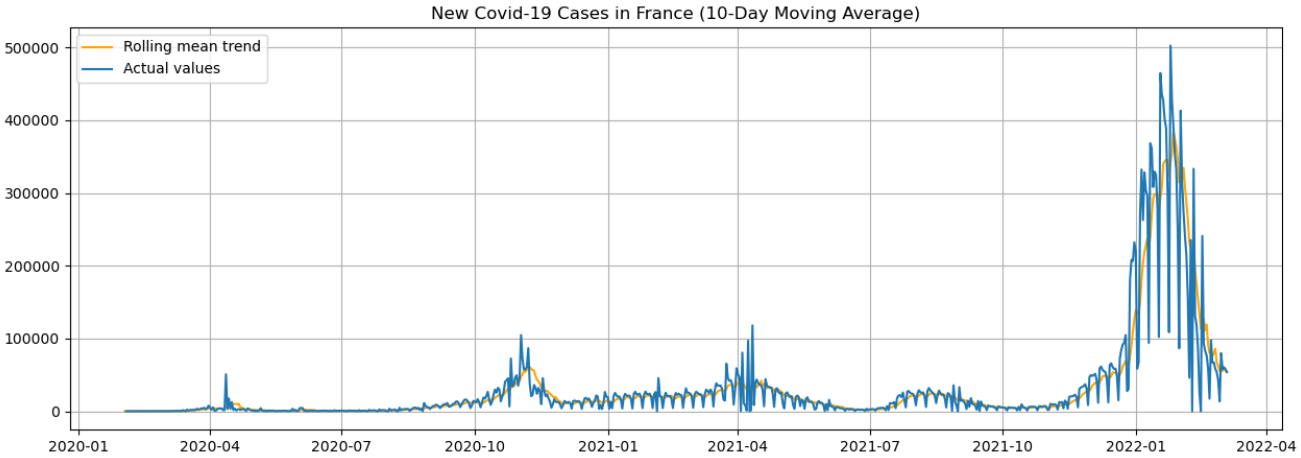

Рисунок 3.6 – згладжування ковзаючим середнім з window=10

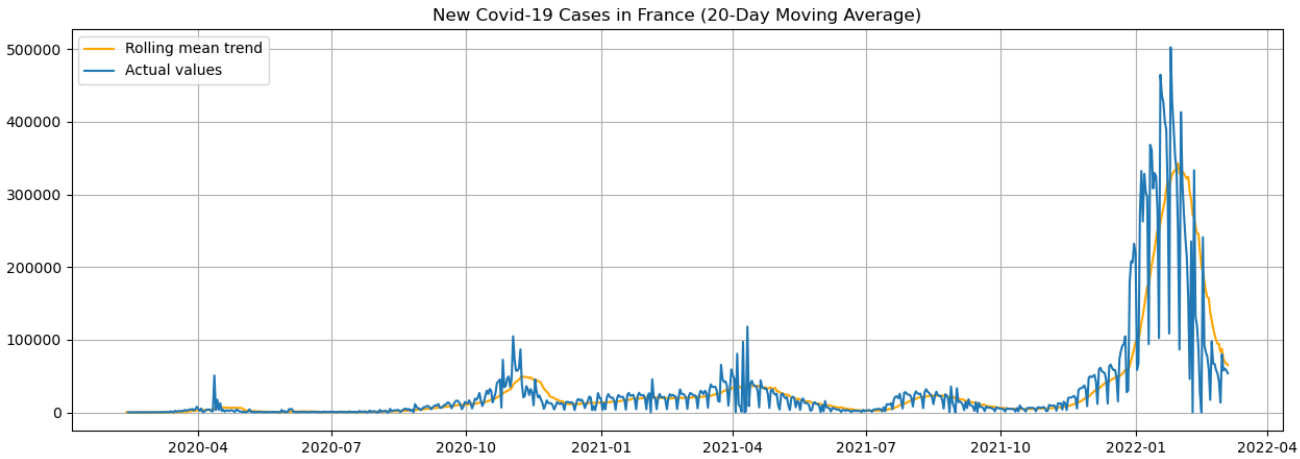

Рисунок 3.7 – згладжування ковзаючим середнім з window=20

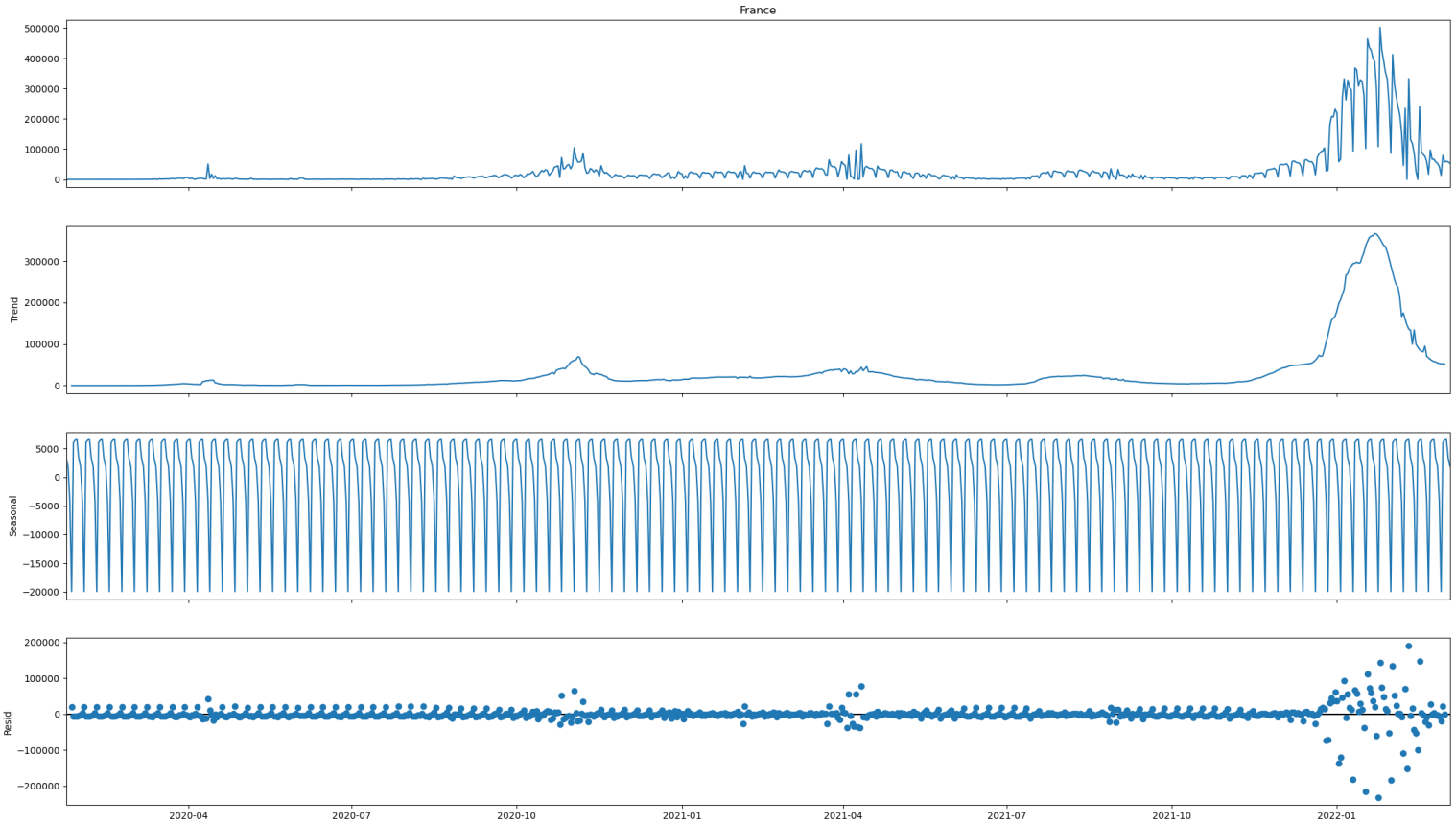

Рисунок 3.8 – візуалізація декомпозиції ряду на тренд, сезонність та залишки

Також була перевірена нульова гіпотеза про наявність одиничного кореня у часової послідовності, що означає нестаціонарність ряду

```
ADF Test Statistic: -3.0831219686639066
ADF p-value: 0.027827715204634444 < 0.05
Critical Value (1%): -3.439099096730074
Critical Value (5%): -2.8654013553540745
Critical Value (10%): -2.568826193777778
The time series is stationary.
```

Рисунок 3.9 – нульова гіпотеза відхилена

Аналогічні дії були виконані для Італії:

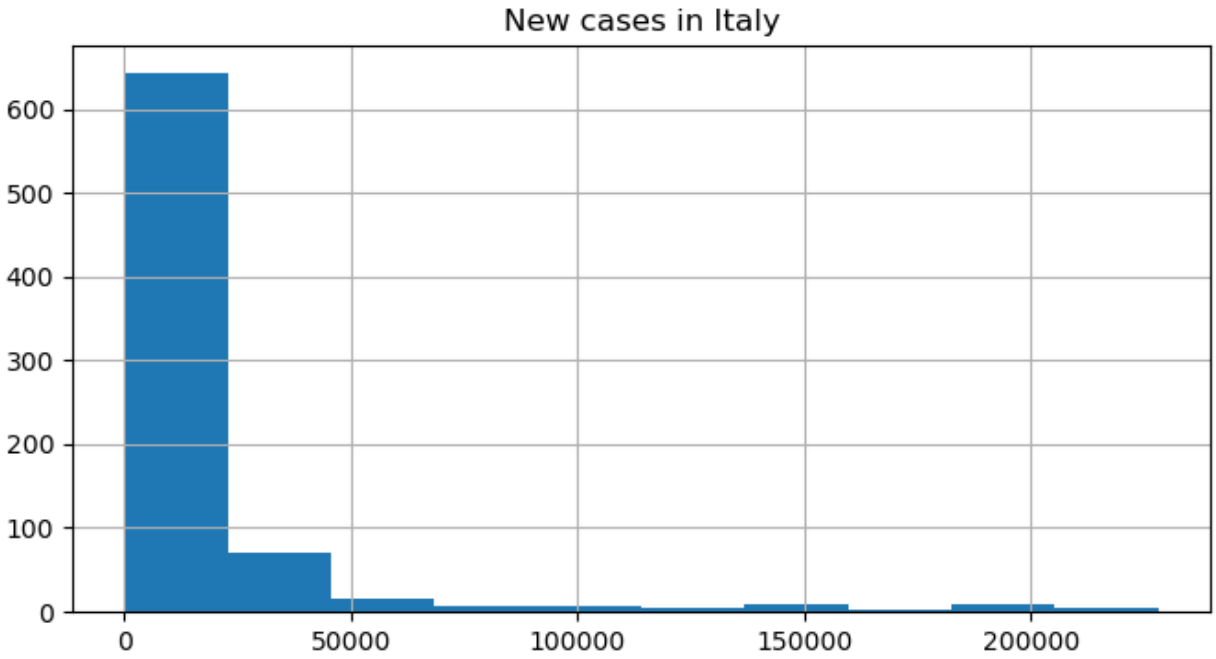

Рисунок 3.10 – кількісна гістограма по захворюванням Італії

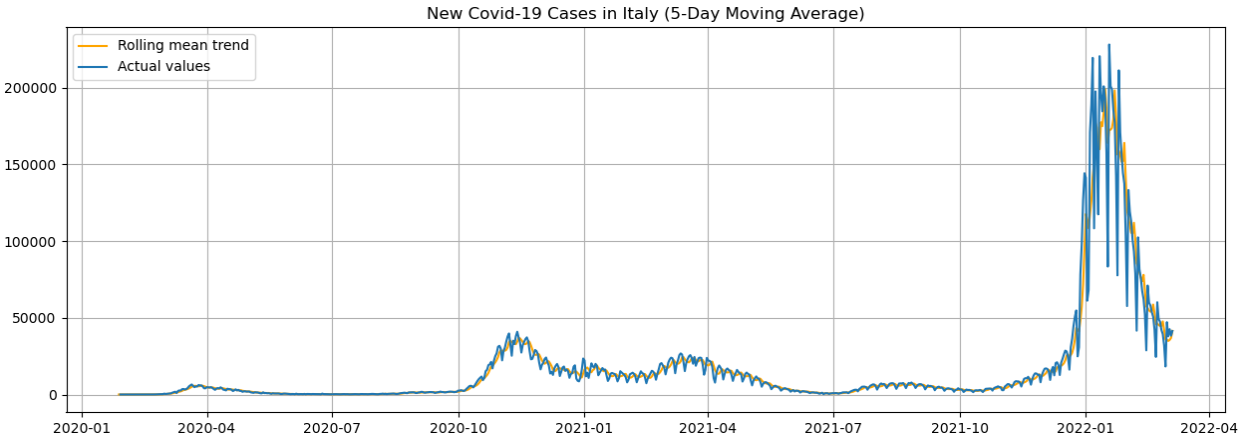

Рисунок 3.11 – згладжування ковзаючим середнім з window=5

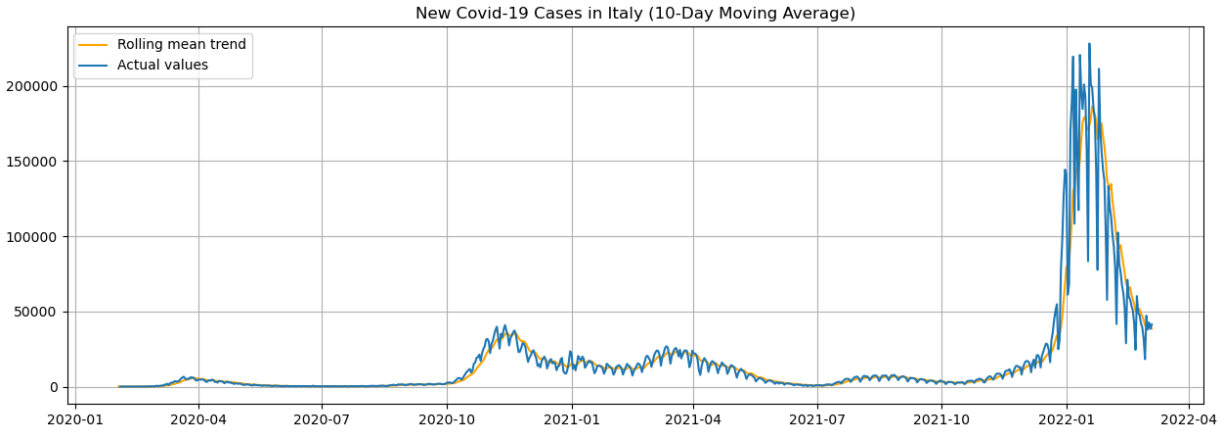

Рисунок 3.12 – згладжування ковзаючим середнім з window=10

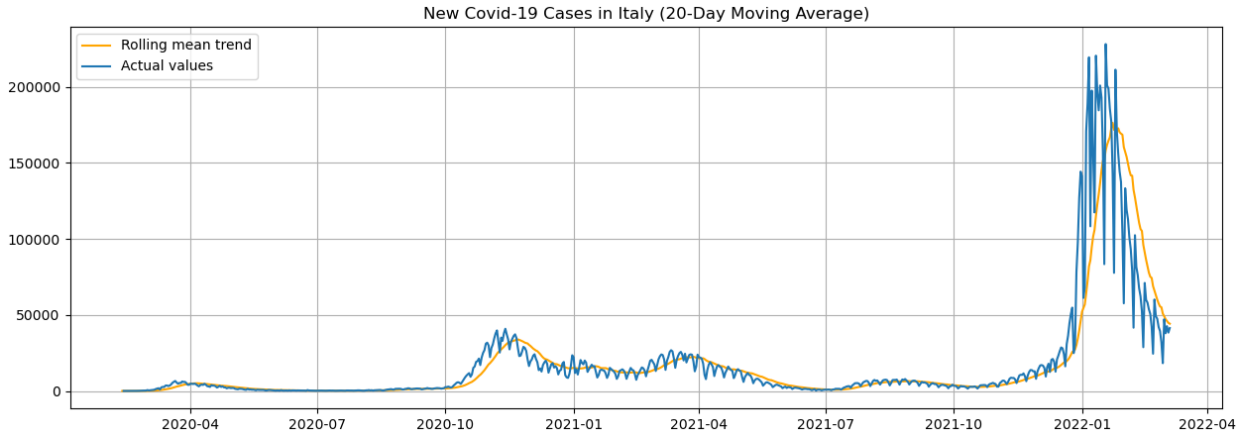

Рисунок 3.13 – згладжування ковзаючим середнім з window=20

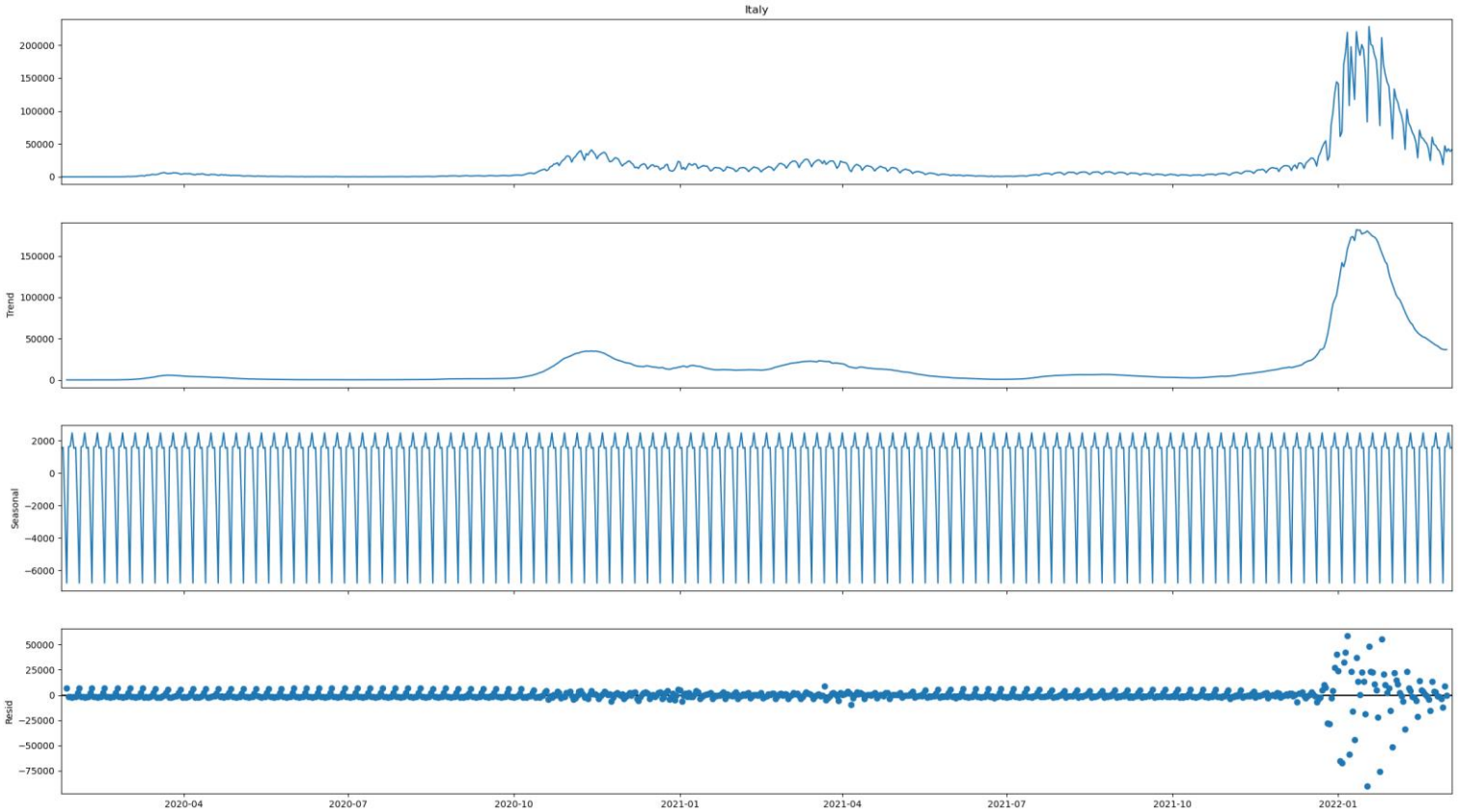

Рисунок 3.14 – візуалізація декомпозиції ряду на тренд, сезонність та залишки

```
ADF Test Statistic: -3.6283231701945566
ADF p-value: 0.005242072665219499 < 0.05
Critical Value (1%): -3.439029421541435
Critical Value (5%): -2.8653706489231876
Critical Value (10%): -2.568809835460933
The time series is stationary.
```

Рисунок 3.15 – нульова гіпотеза відхилена

Наступним кроком були завантажені та оброблені дані про курс гривні до євро з 01.01.2020 по 09.05.2023 які отримали наступний вигляд:

	0
2020-01-01 00:00:00	26.55650
2020-01-02 00:00:00	26.46650
2020-01-03 00:00:00	26.45160
2020-01-06 00:00:00	26.53690
2020-01-07 00:00:00	26.43020
2020-01-08 00:00:00	26.61020
2020-01-09 00:00:00	26.76530
2020-01-10 00:00:00	26.61520
2020-01-13 00:00:00	26.77490
2020-01-14 00:00:00	26.37100
2020-01-15 00:00:00	26.85240
2020-01-16 00:00:00	26.86880
2020-01-17 00:00:00	26.91520
2020-01-20 00:00:00	27.00280
2020-01-21 00:00:00	26.88490
2020-01-22 00:00:00	27.07040
2020-01-23 00:00:00	26.79560
2020-01-24 00:00:00	26.90710
2020-01-27 00:00:00	27.11530
2020-01-28 00:00:00	27.26350
2020-01-29 00:00:00	27.37390

Рисунок 3.16 – початкові дані

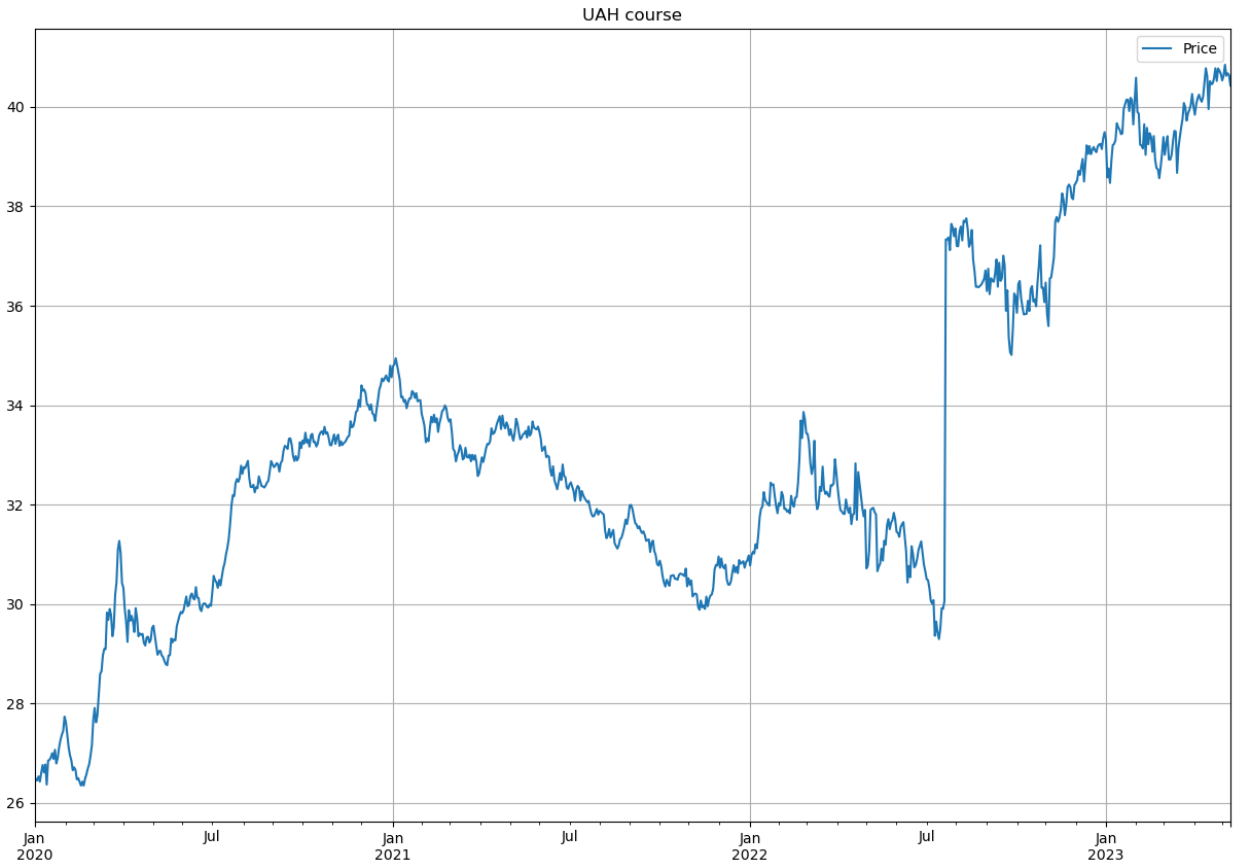

Рисунок 3.17 – візуалізація завантажених даних

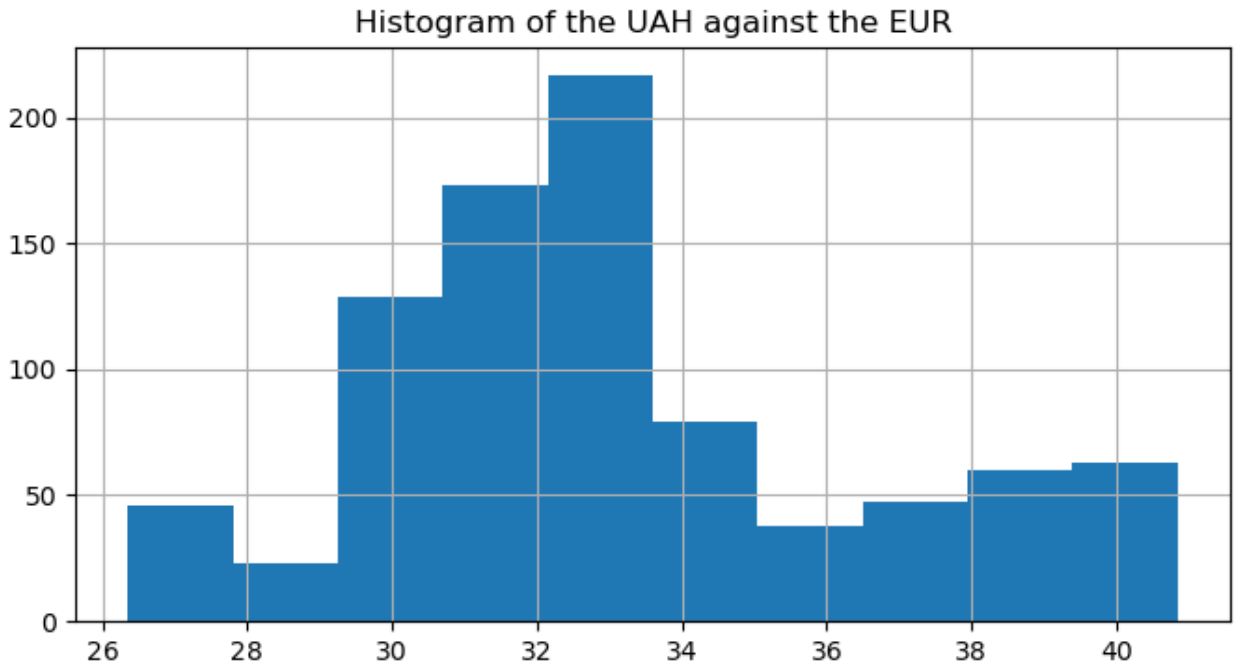

Рисунок 3.18 – кількісна гістограма курсів гривні до євро

Як і у попередніх даних, було візуалізовано згладжування за допомогою ковзаючого середнього:

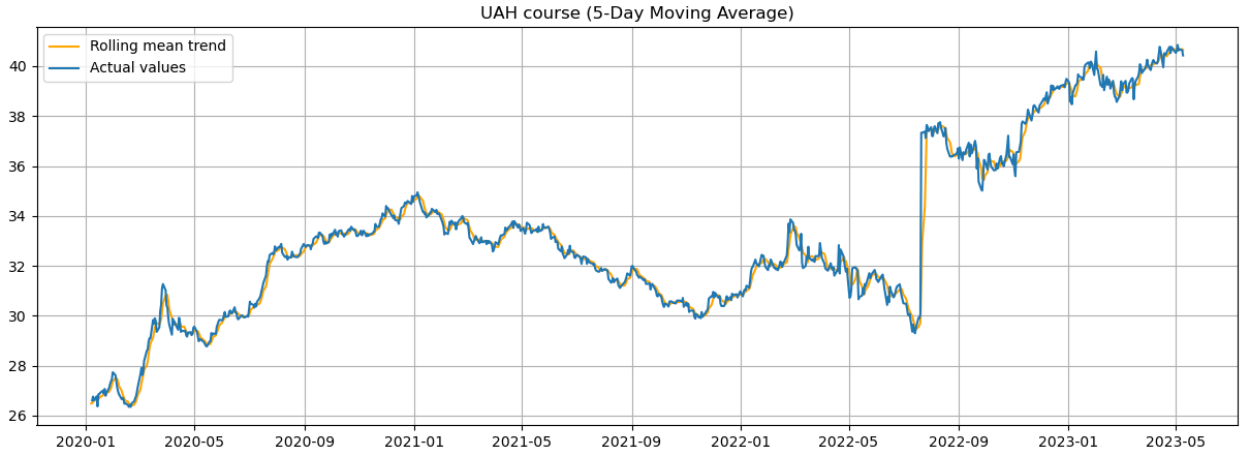

Рисунок 3.19 – згладжування ковзаючим середнім з window=5

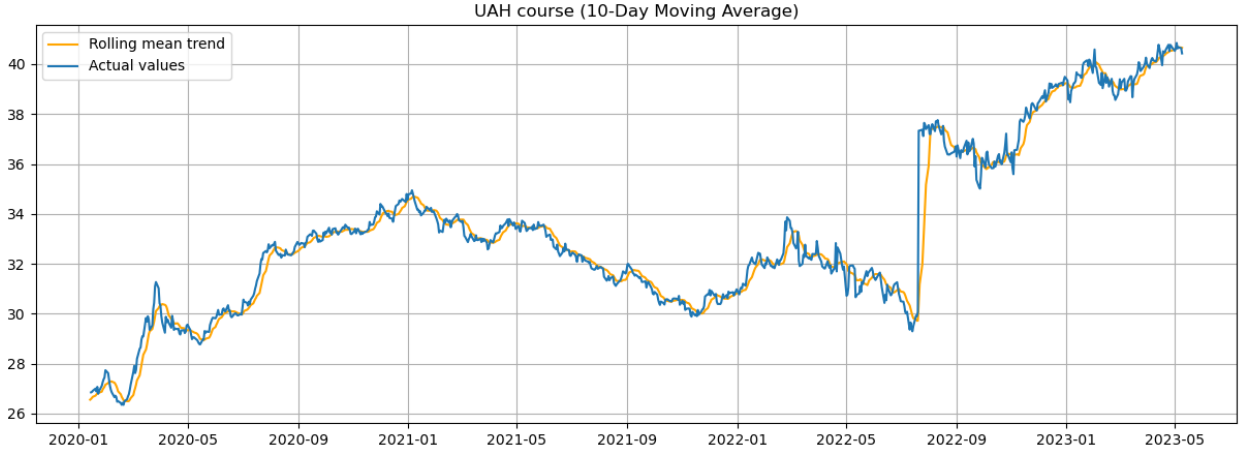

Рисунок 3.20 – згладжування ковзаючим середнім з window=10

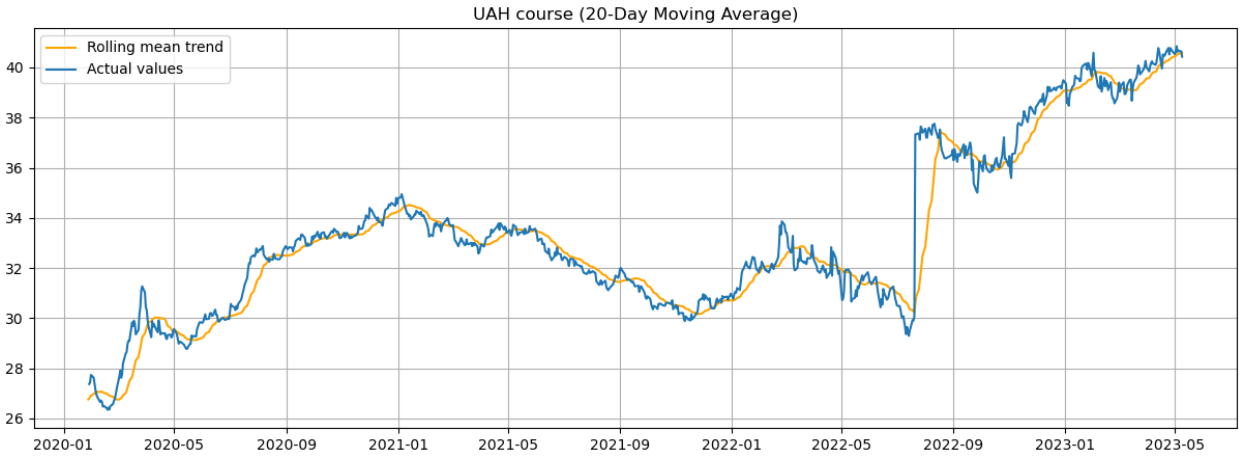

Рисунок 3.21 – згладжування ковзаючим середнім з window=20

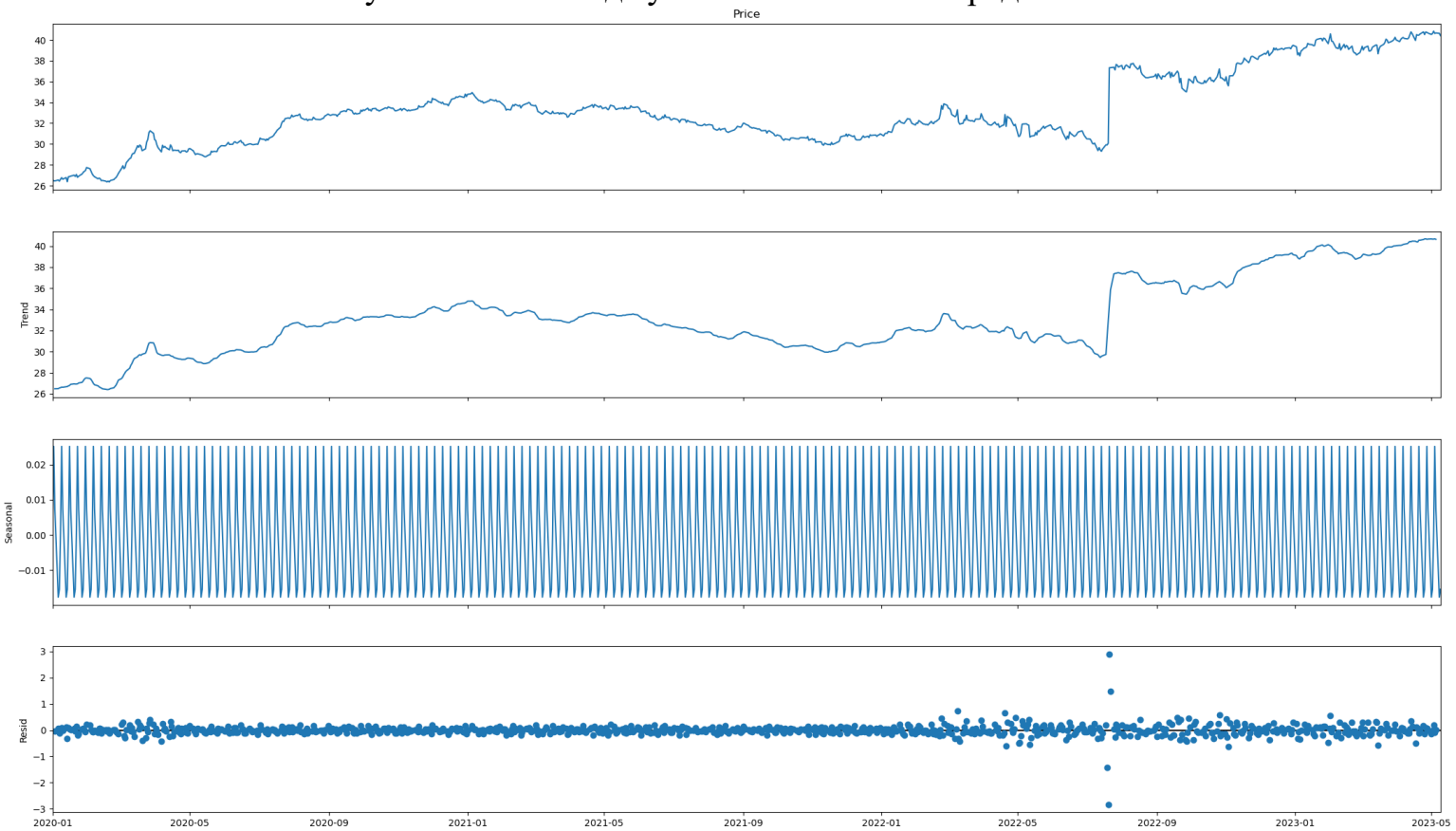

Рисунок 3.22 – візуалізація декомпозиції ряду на тренд, сезонність та залишки

Так само була перевірена нульова гіпотеза про нестаціонарність ряду:

```
ADF Test Statistic: -1.2890358976503298
ADF p-value: 0.634173356797348 > 0.05
Critical Value (1%): -3.4378627516320006
Critical Value (5%): -2.8648563831383322
Critical Value (10%): -2.568535885040459
The time series is non-stationary.
```

Рисунок 3.23 – неможливо відхилити нульову гіпотезу

Для додаткового завдання були завантажені дані про температуру та кількість опадів у Сіетлі. Були здійснені необхідні перетворення, заповнено нулями пропущену кількість опадів, переведено температуру у Цельсії та отримано наступний датафрейм:

	PRCP	TMAX	TMIN	RAIN
1948-01-01 00:00:00	0.47000	10.55556	5.55556	True
1948-01-02 00:00:00	0.59000	7.22222	2.22222	True
1948-01-03 00:00:00	0.42000	7.22222	1.66667	True
1948-01-04 00:00:00	0.31000	7.22222	1.11111	True
1948-01-05 00:00:00	0.17000	7.22222	0.00000	True
1948-01-06 00:00:00	0.44000	8.88889	3.88889	True
1948-01-07 00:00:00	0.41000	10.00000	4.44444	True
1948-01-08 00:00:00	0.04000	8.88889	1.66667	True
1948-01-09 00:00:00	0.12000	10.00000	-0.55556	True
1948-01-10 00:00:00	0.74000	6.11111	1.11111	True
1948-01-11 00:00:00	0.01000	5.55556	0.00000	True
1948-01-12 00:00:00	0.00000	5.00000	-3.33333	False
1948-01-13 00:00:00	0.00000	7.22222	-1.66667	False
1948-01-14 00:00:00	0.00000	3.33333	-3.33333	False
1948-01-15 00:00:00	0.00000	1.11111	-0.55556	False
1948-01-16 00:00:00	0.00000	1.11111	-2.22222	False
1948-01-17 00:00:00	0.00000	1.66667	-1.66667	False
1948-01-18 00:00:00	0.00000	0.55556	-2.22222	False
1948-01-19 00:00:00	0.00000	1.11111	-2.77778	False
1948-01-20 00:00:00	0.00000	2.22222	-1.66667	False
1948-01-21 00:00:00	0.00000	8.88889	0.00000	False

temps Format: %s

Рисунок 3.24 – початкові дані

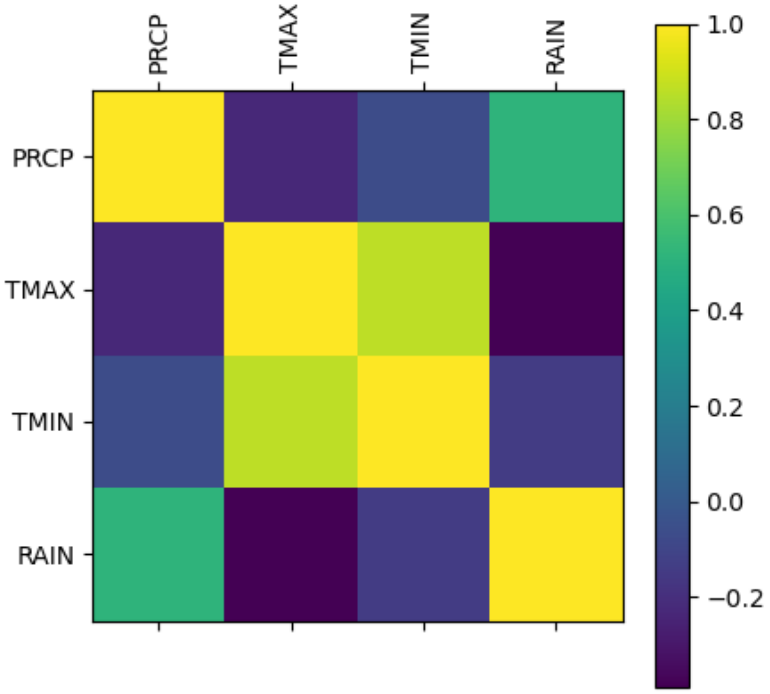

Рисунок 3.25 – візуалізація матриці кореляцій

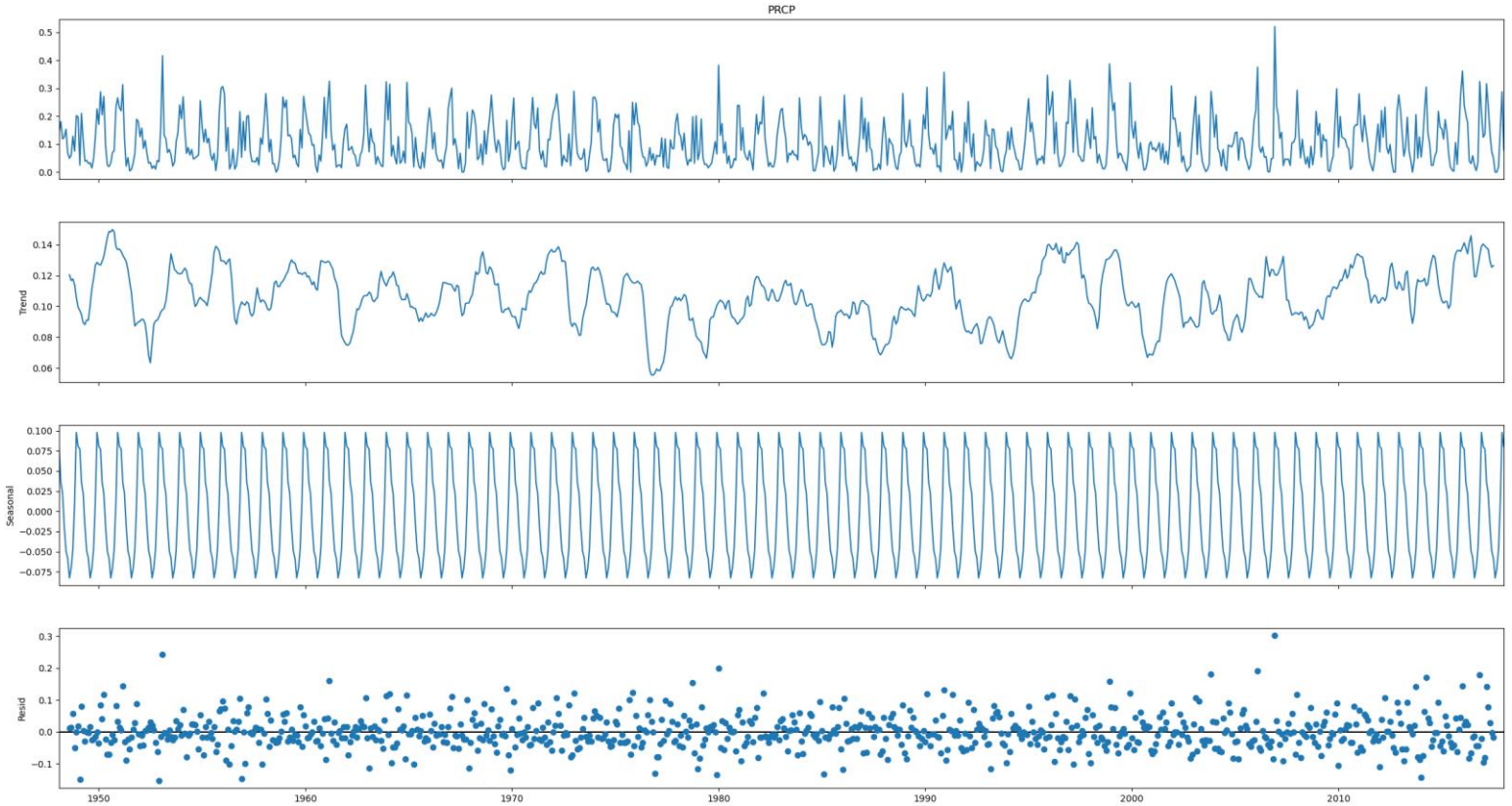

Рисунок 3.26 – візуалізація декомпозиції ряду на тренд, сезонність та залишки

Занадто багато даних для візуалізації, тому було взято лише дані за останні 5 років.

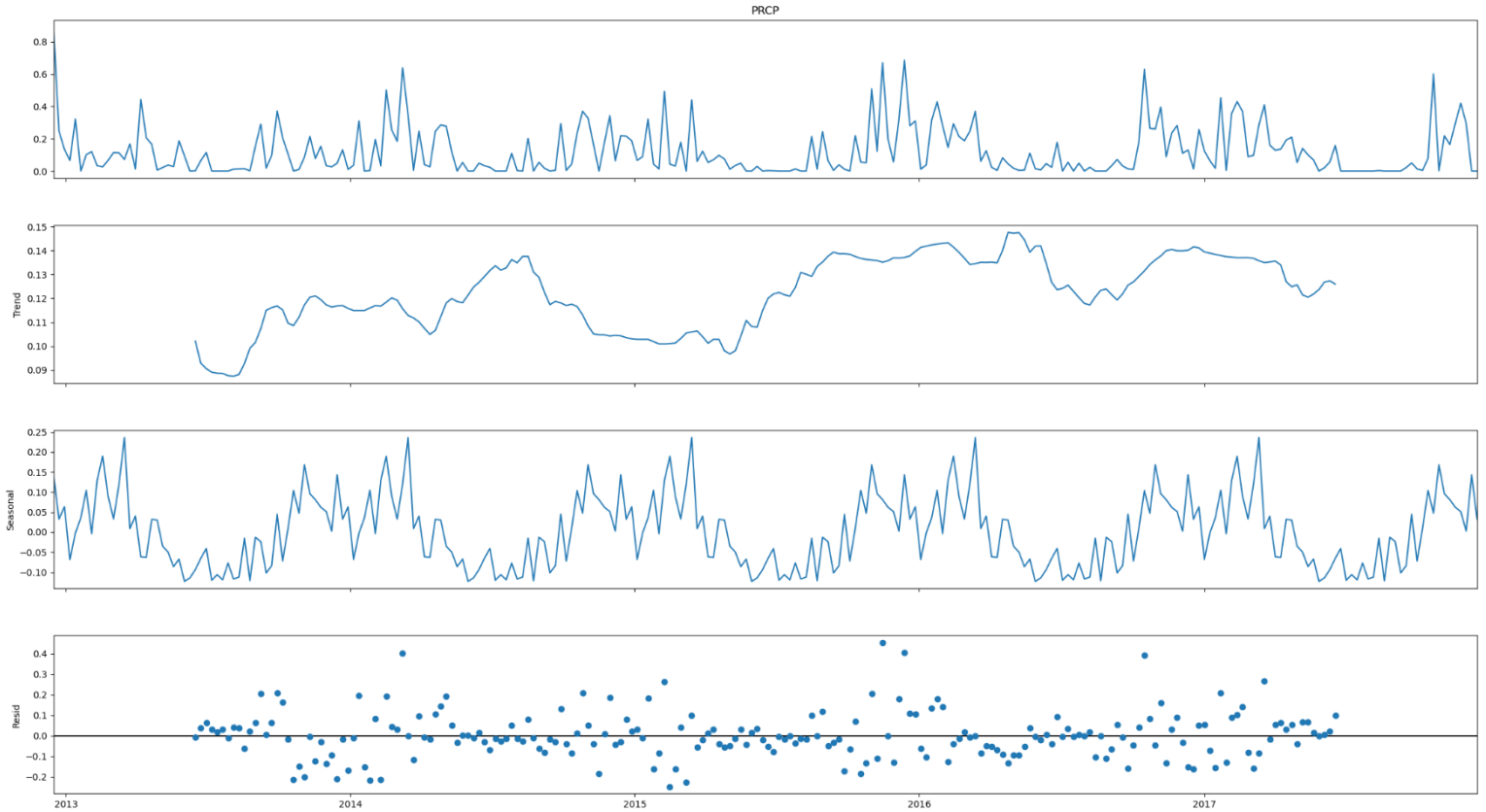

Рисунок 3.27 – візуалізація декомпозиції ряду що містить лише дані останніх років на тренд, сезонність та залишки

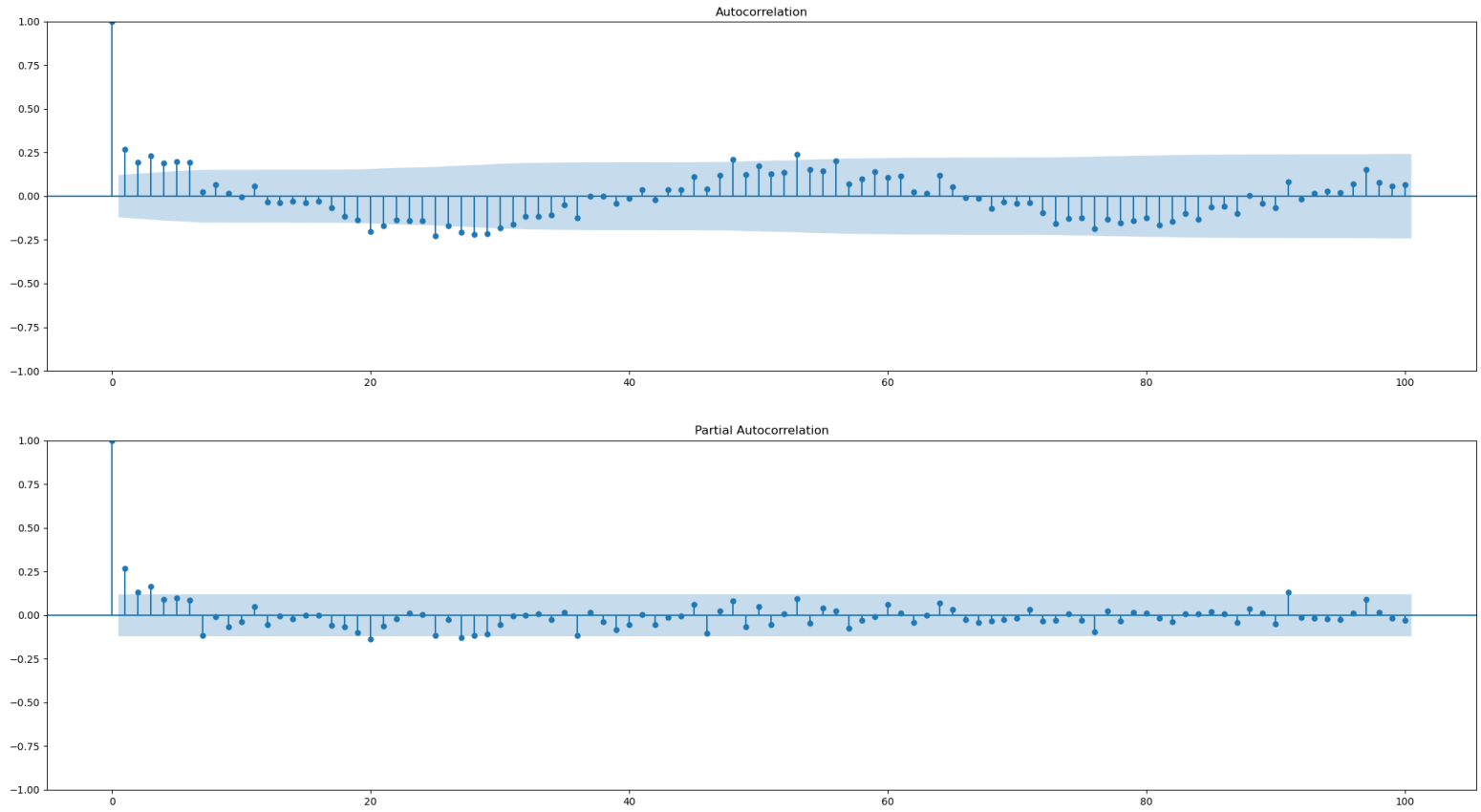

Рисунок 3.28 – графіки автокореляції та часткової автокореляції для опадів за останні 5 років

Бачимо, що перший лаг є найбільш значущим, тому для ARIMA моделі значення 1 для параметру p буде достатнім. Також, для її побудови було розділено дані на тренувальні (1948/01/01-2017/01/01) та тестові (2017/01/02-2017/12/14). Окрім цього було перевірено ряд на стаціонарність:

```
ADF Test Statistic: -16.928289719325942
ADF p-value: 9.82108979209427e-30 < 0.05
Critical Value (1%): -3.4306063382613226
Critical Value (5%): -2.8616532938508263
Critical Value (10%): -2.5668303029867414
The time series is stationary.
```

Рисунок 3.29 – нульова гіпотеза відхилена

Отже, параметр d дорівнюватиме 0 в нашому випадку. Після тренування моделі було порівняно прогнозовану та фактичну кількість опадів для 2017 року, а також зроблено передбачення на 2018 рік.

```
Mean Squared Error: 0.0778
```

Рисунок 3.30 – отримана середньоквадратична похибка

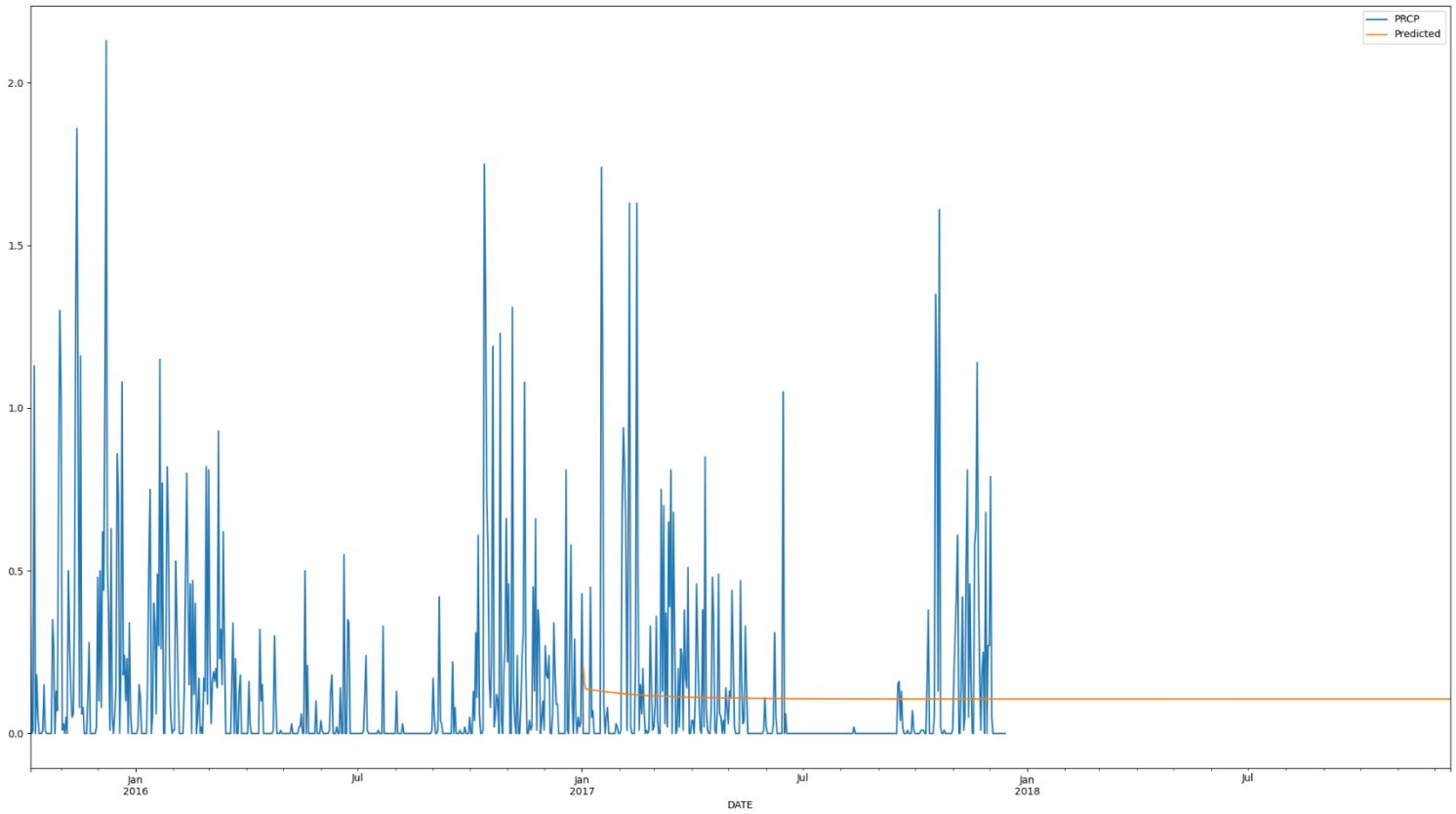

Рисунок 3.31 – візуалізація передбачення

Висновок: досліджуючи дані про нові захворювання covid-19 у двох сусідніх країнах, було встановлено, що часові послідовності мають однаковий тренд, однак дещо різняться за масштабом (кількістю випадків); зміна курсу євро до гривні не має чіткої сезонності, однак має зростаючий тренд за останній рік, пов’язаний з повномасштабним вторгненням; після аналізу даних про погоду в Сіетлі було встановлено, що: кореляція між температурою та кількістю опадів практично відсутня, однак ці опади проявляють чітко виражену сезонність; побудована модель ARIMA для їх прогнозування виявилась доволі точною, про що свідчить близька до нуля середньо квадратична похибка.