**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**”

**Виконав(ла)**

(шифр, прізвище, ім'я, по батькові)

*ІП-11 Тарасьонок Дмитро Євгенович*

**Перевірила**

(прізвище, ім'я, по батькові)

*Ліхоузова Т. А.*

Київ 2023

**ЗМІСТ**

[1 Мета лабораторної роботи 3](#_Toc135778108)

[2 Завдання 4](#_Toc135778109)

[2.1 Основне завдання 4](#_Toc135778110)

[2.2 Додаткове завдання 4](#_Toc135778111)

[3 Виконання основного завдання 5](#_Toc135778112)

[3.1 Побудувати та проаналізувати часовий ряд для статистики захворювань на Covid в двох сусідніх країнах по вашому вибору (дані взяти в інтернеті). 5](#_Toc135778113)

[3.2 Побудувати та проаналізувати часовий ряд для курсу гривня/долар або гривня/євро за останні 3 роки (дані взяти в інтернеті). 16](#_Toc135778114)

[4 Виконання додаткового завдання 29](#_Toc135778115)

[5 Висновок 41](#_Toc135778116)

# Мета лабораторної роботи

Мета роботи – ознайомитись з методами моделювання часових послідовностей.

# Завдання

## Основне завдання

* 1. Побудувати та проаналізувати часовий ряд для статистики захворювань на Covid в двох сусідніх країнах по вашому вибору (дані взяти в інтернеті).
  2. Побудувати та проаналізувати часовий ряд для курсу гривня/долар або гривня/євро за останні 3 роки (дані взяти в інтернеті).

## Додаткове завдання

Потрібно з'ясувати, чи є сезонна компонента в кількості опадів в Сіетлі. (https://www.kaggle.com/rtatman/did-it-rain-in-seattle-19482017/data,   або seattleWeather\_1948-2017.csv [Скачати потрібні дані](https://drive.google.com/drive/folders/1kYY4pxkeLXoszCX8icpFUrhFCnwgAyhz?usp=sharing)).

* 1. Градуси перевести в Цельсії.
  2. Чи є кореляція між температурою та опадами?
  3. Скласти прогноз опадів на 2018 рік, оцінити точність прогнозу

# Виконання основного завдання

## Побудувати та проаналізувати часовий ряд для статистики захворювань на Covid в двох сусідніх країнах по вашому вибору (дані взяти в інтернеті).

Для початку імпортуємо всі необхідні пакети. Для прогнозування будемо використовувати моделі ARIMA та Prophet.

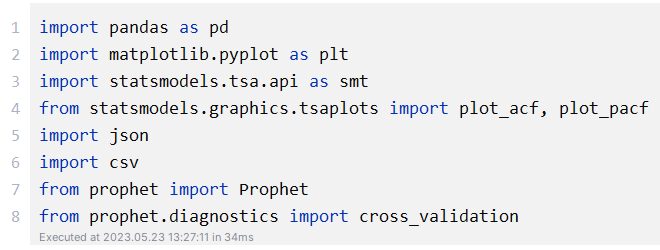


Рисунок 3.1 – Імпортування пакетів

Завантажимо дані про захворюваність на коронавірус у світі, переглянемо їх.

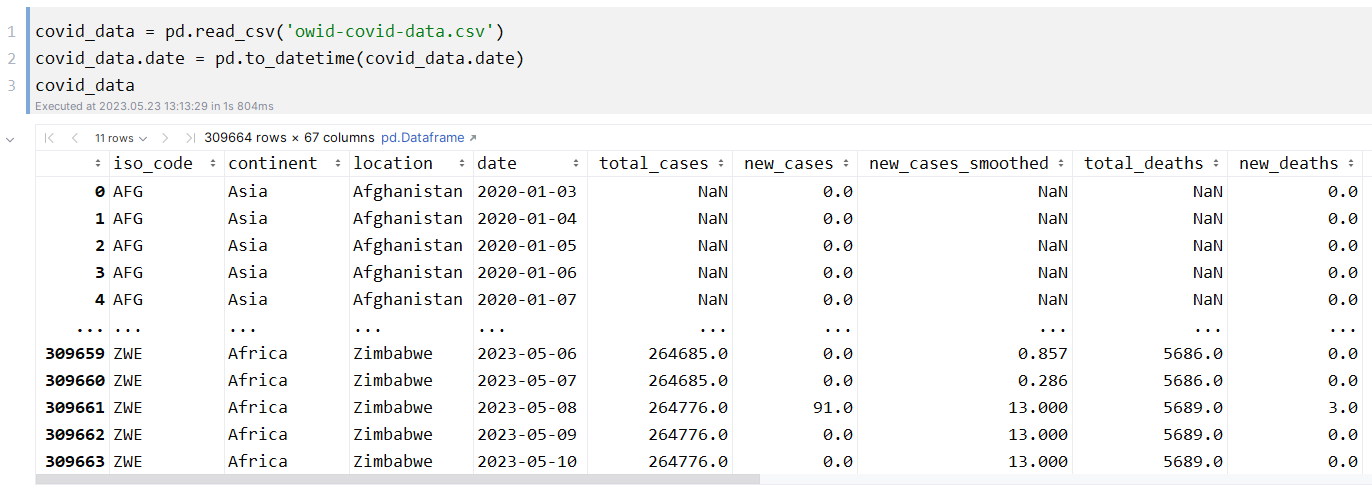


Рисунок 3.2 – Завантаження історії захворюваності на COVID-19

Відфільтруємо ці дані за двома країнами: Франція та Німеччина, залишимо тільки стовпці з датами та кількістю нових випадків, а також заповнимо пропущені значення нулями.

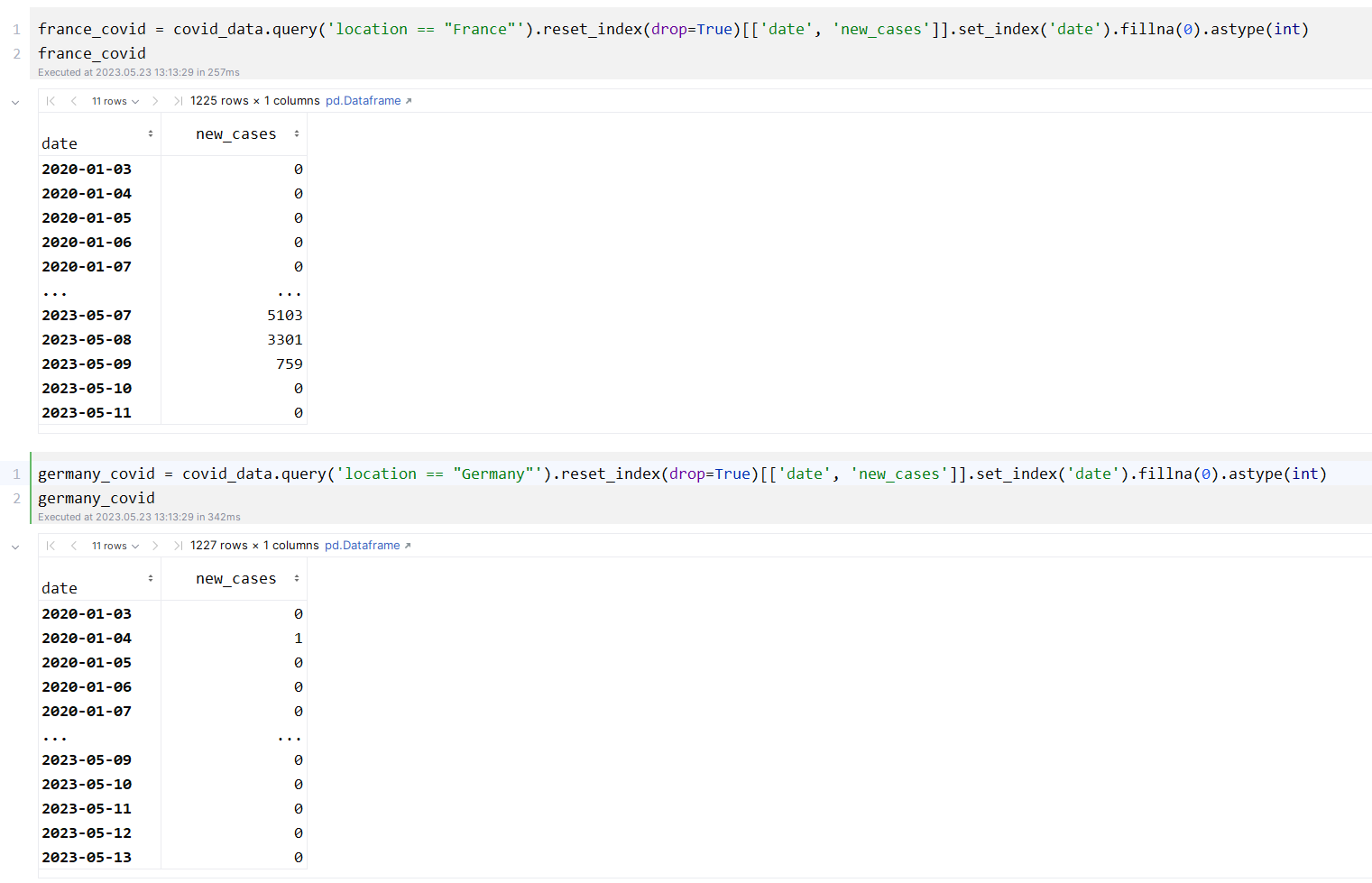


Рисунок 3.3 – Виокремлення нових випадків захворювань для Франції та Німеччини

Можемо зобразити ці часові послідовності на графіку.

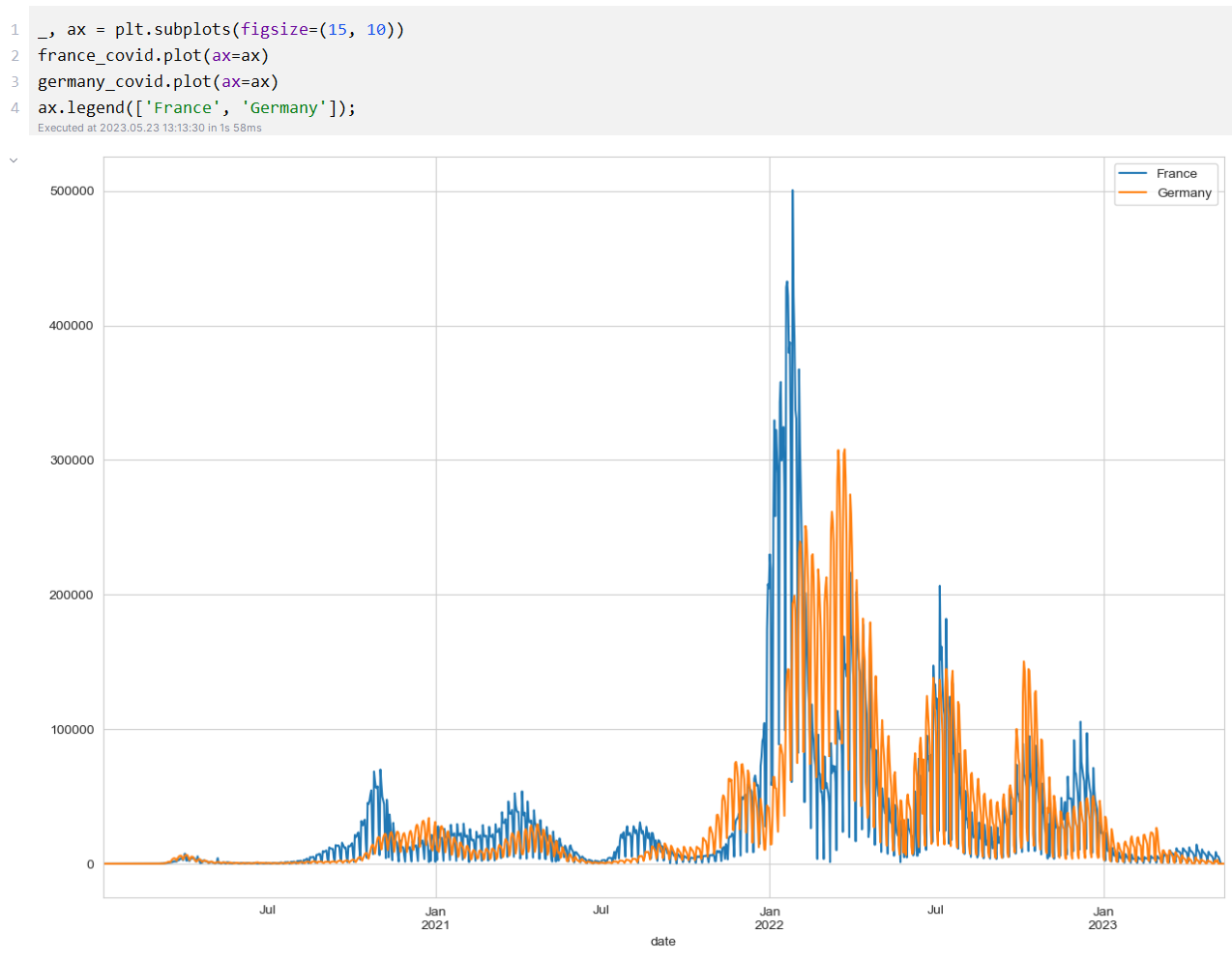


Рисунок 3.4 – Графік захворюваності на COVID-19 у Франції та Німеччині

Тепер відобразимо гістограму кількості нових випадків для Німеччини.

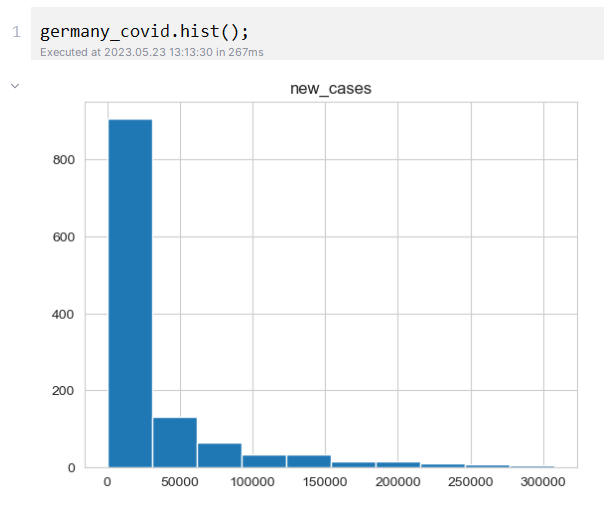


Рисунок 3.5 – Розподіл кількості нових випадків у Німеччині

Для більшої наочності графіків можемо скористатися ковзаючим середнім. В якості розміру вікна встановимо 5, 10, 20 та 40. Спочатку наведемо їх для Франції.

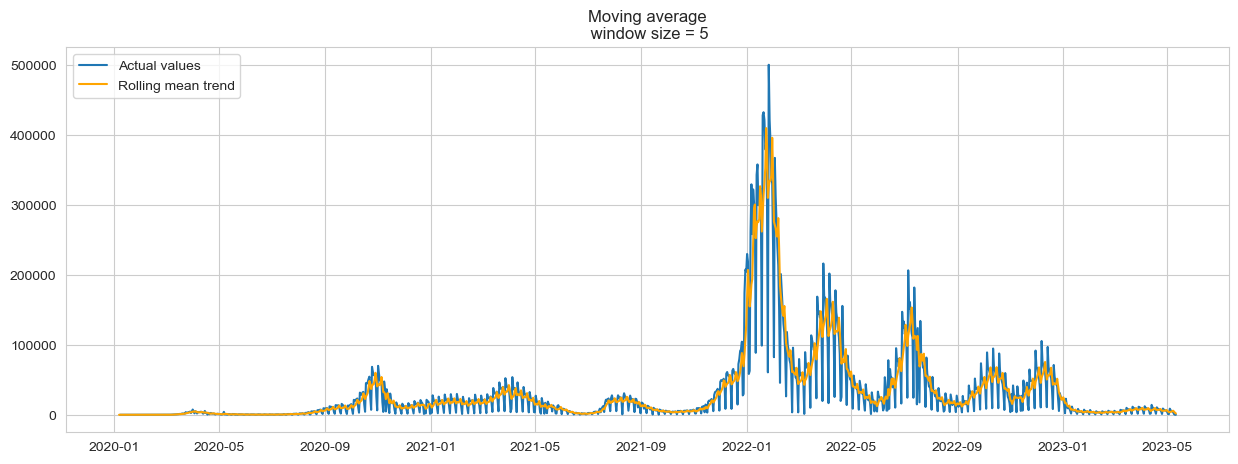


Рисунок 3.6 – Графік захворюваності на COVID-19 у Франції з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 5



Рисунок 3.7 – Графік захворюваності на COVID-19 у Франції з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 10

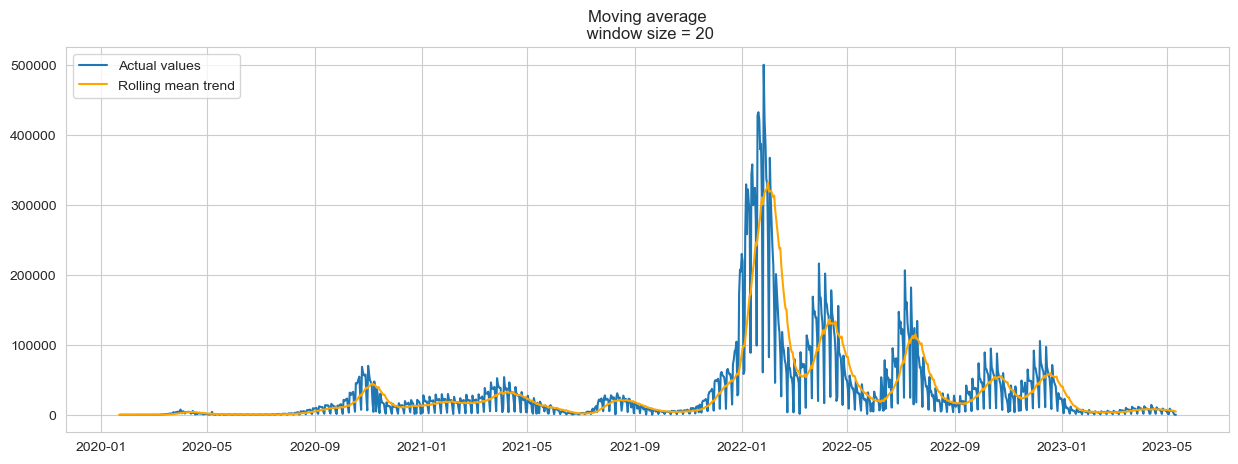


Рисунок 3.8 – Графік захворюваності на COVID-19 у Франції з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 20

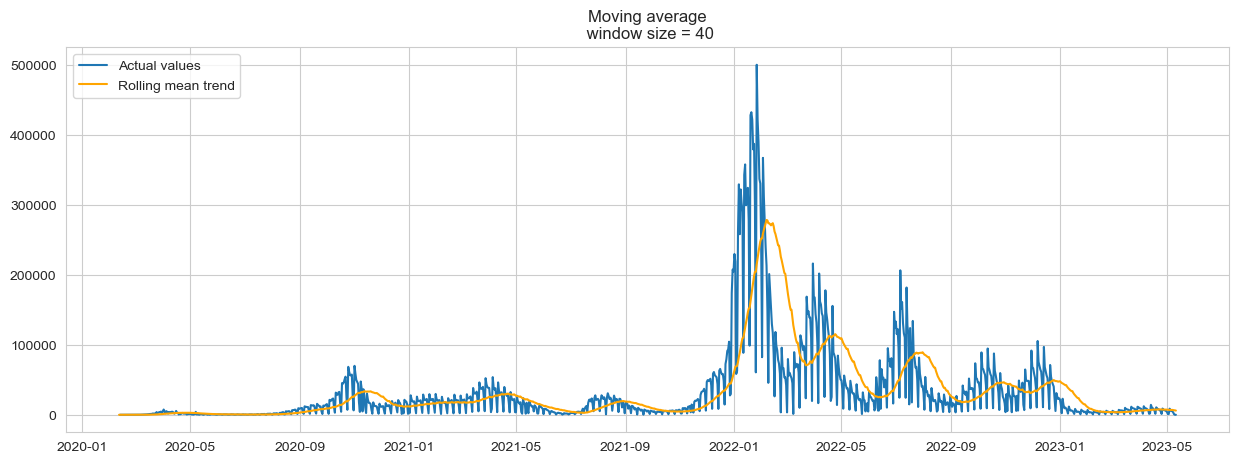


Рисунок 3.9 – Графік захворюваності на COVID-19 у Франції з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 40

Можемо побачити, що досить добре часовий ряд відображає графік із розміром вікна рівним 10. Тепер наведемо такі ж графіки для Німеччини.

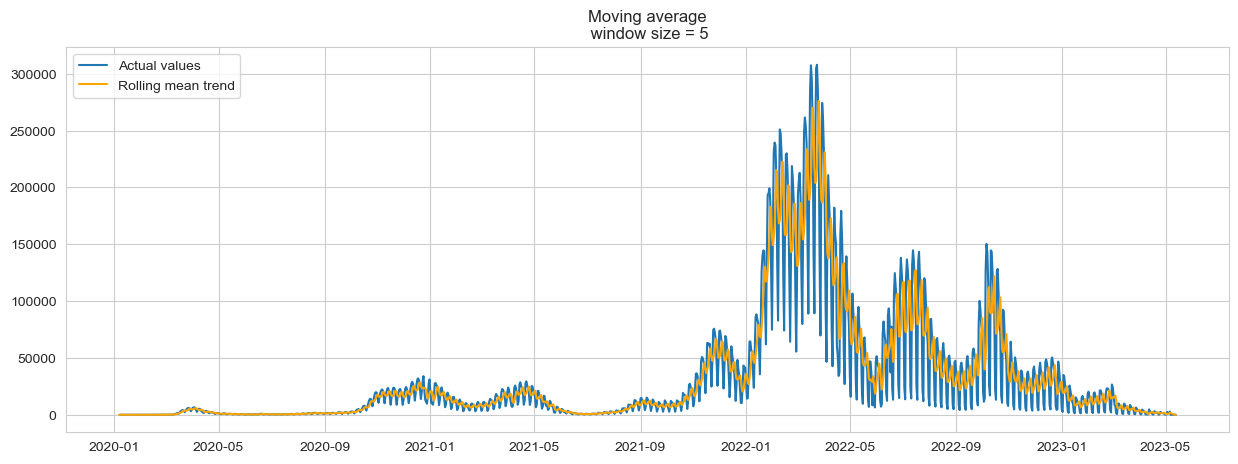


Рисунок 3.10 – Графік захворюваності на COVID-19 у Німеччині з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 5

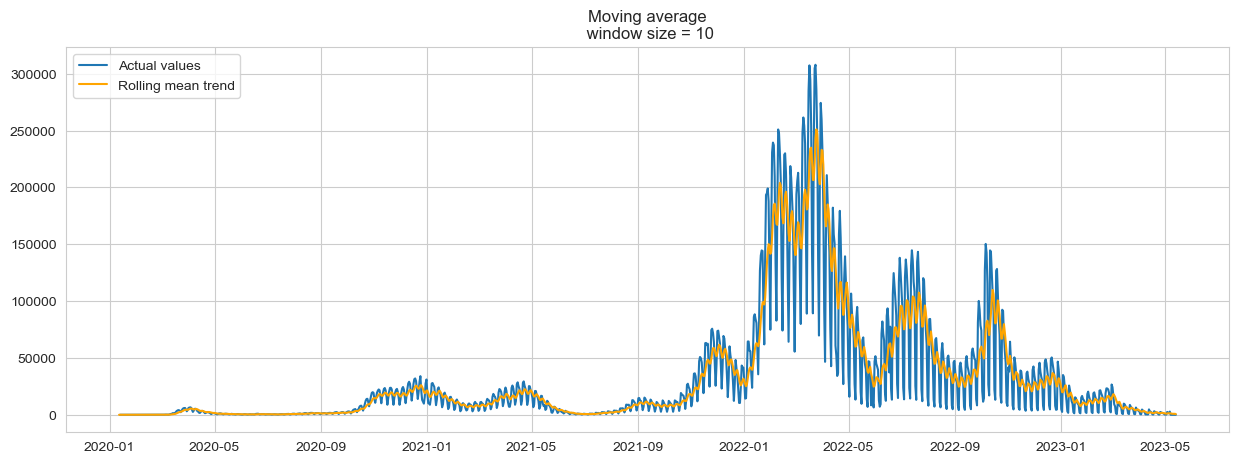


Рисунок 3.11 – Графік захворюваності на COVID-19 у Німеччині з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 10

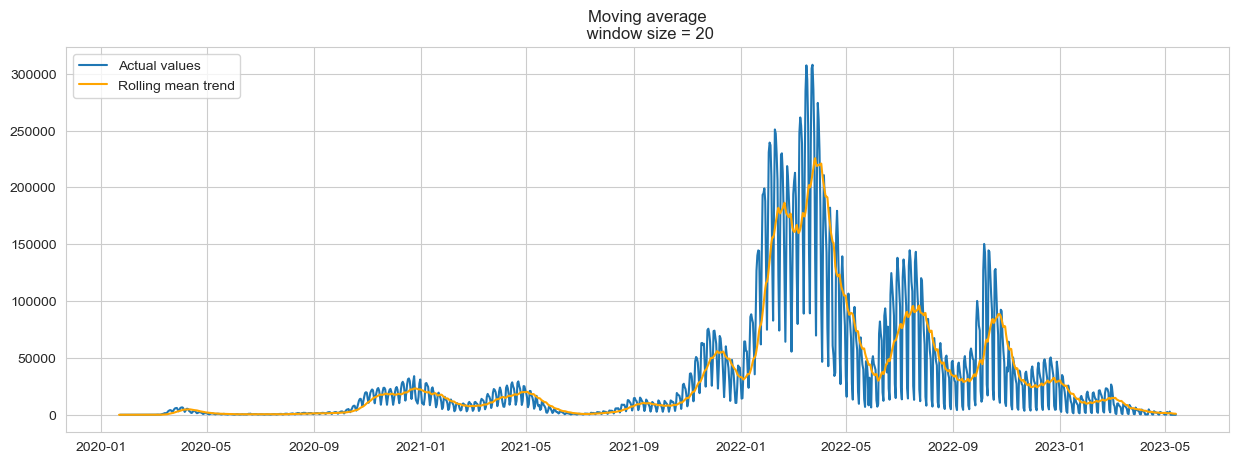


Рисунок 3.12 – Графік захворюваності на COVID-19 у Німеччині з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 20

Рисунок 3.13 – Графік захворюваності на COVID-19 у Німеччині з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 40

Наступним кроком можемо розбити часові послідовності на компоненти.

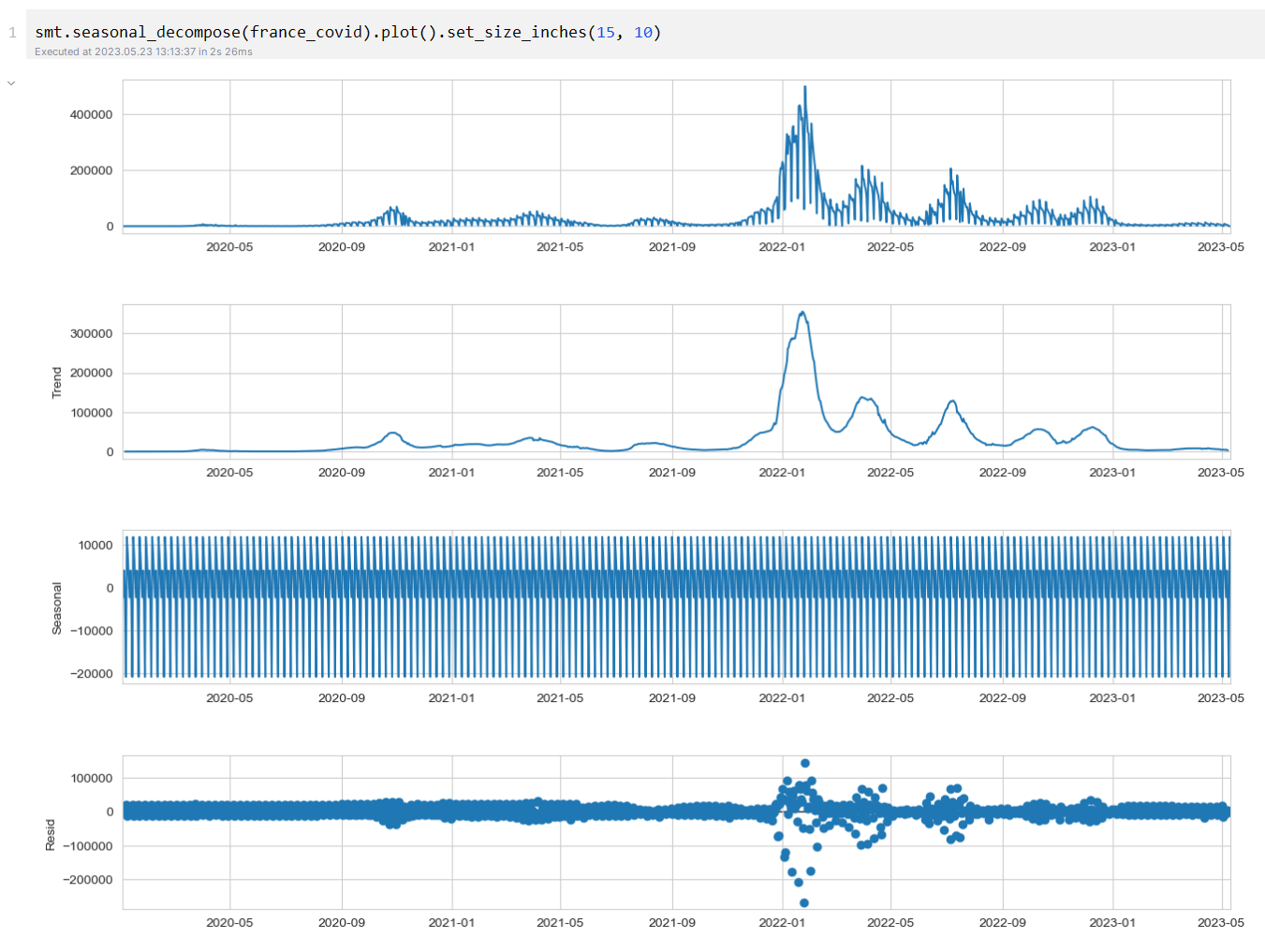


Рисунок 3.14 – Декомпозиція часової послідовності для Франції на тренд, сезонність та залишки

Бачимо, що чіткий тренд та сезонність відсутні.

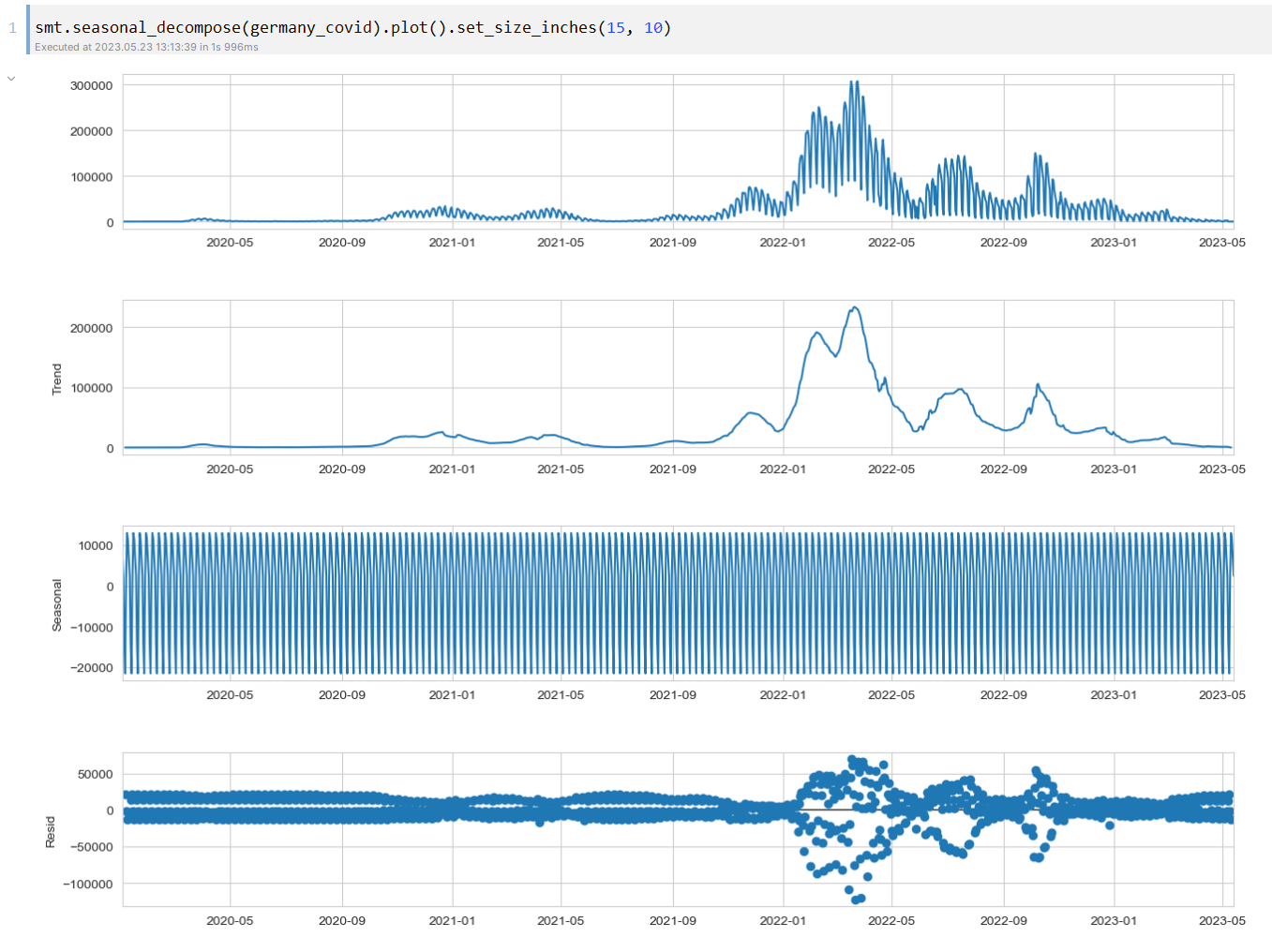


Рисунок 3.15 – Декомпозиція часової послідовності для Німеччини на тренд, сезонність та залишки

Така ж картина спостерігається й для Німеччини. Побудуємо графіки автокореляції для цих послідовностей.



Рисунок 3.16 – Графіки автокореляції та часткової автокореляції для Франції (зліва) та Німеччини (справа)

Наступним кроком можемо перевірити ряд на стаціонарність за допомогою тесту Дікі-Фуллера.

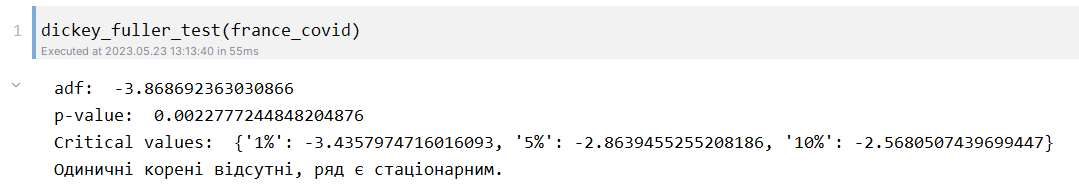


Рисунок 3.17 – Тест Дікі-Фуллера для історії захворювань на COVID-19 у Франції

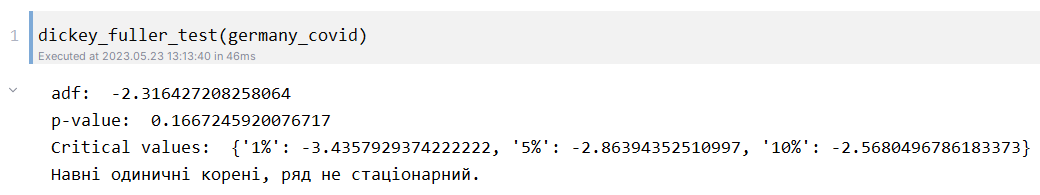


Рисунок 3.18 – Тест Дікі-Фуллера для історії захворювань на COVID-19 у Німеччині

Обидва ряди не є стаціонарними. Можемо зробити часову послідовність стаціонарною за допомогою методу диференціювання.

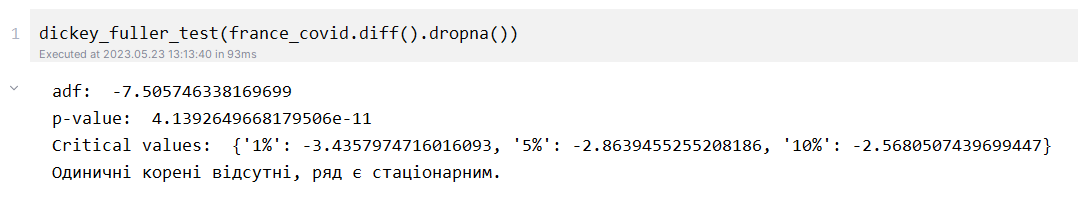


Рисунок 3.19 – Тест Дікі-Фуллера для історії захворювань на COVID-19 у Франції після взяття перших різниць

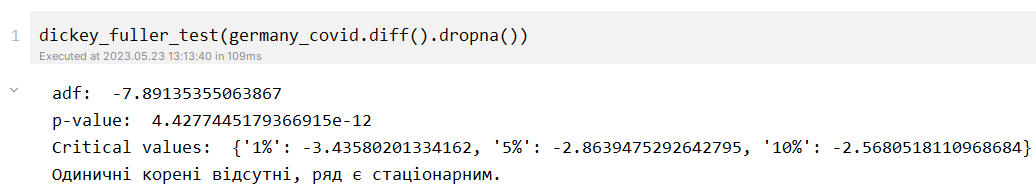


Рисунок 3.20 – Тест Дікі-Фуллера для історії захворювань на COVID-19 у Німеччині після взяття перших різниць

Після диференціювання послідовність стала стаціонарною.

## Побудувати та проаналізувати часовий ряд для курсу гривня/долар або гривня/євро за останні 3 роки (дані взяти в інтернеті).

Тепер виконаємо завдання для прогнозування курсів валют. Було завантажено дані з сайту  [https://minfin.com.ua/ua/currency/auction/archive/usd/all/#fromDate=24-02-2021&toDate=14-05-2023](%20https://minfin.com.ua/ua/currency/auction/archive/usd/all/#fromDate=24-02-2021&toDate=14-05-2023). Обрано було саме історію цін на чорному ринку, адже на ньому ціни сильно відрізняються від офіційного курсу. Так, наприклад, з початку війни курс НБУ зафіксований, а курс чорного ринку постійно змінюється, підлаштовуючись під поточну ситуацію. Завантажимо ці дані та виведемо графік курсу валют.

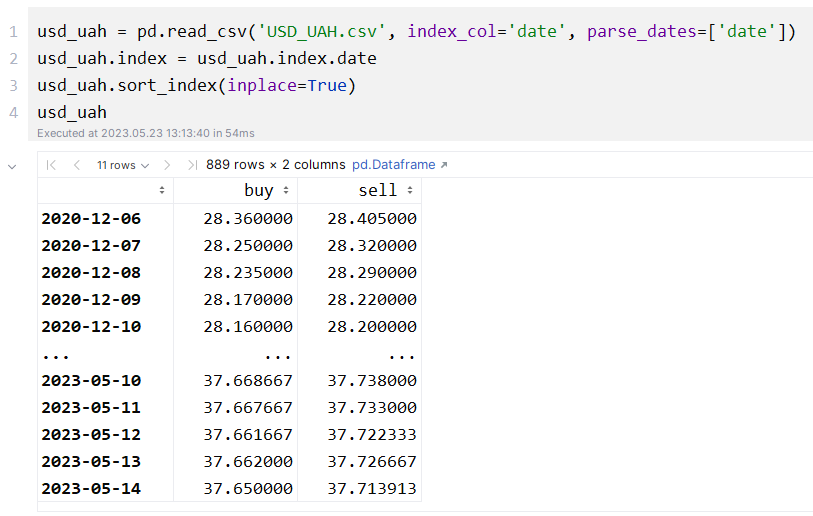


Рисунок 3.21 – Завантаження історичних даних про курс гривні до долара

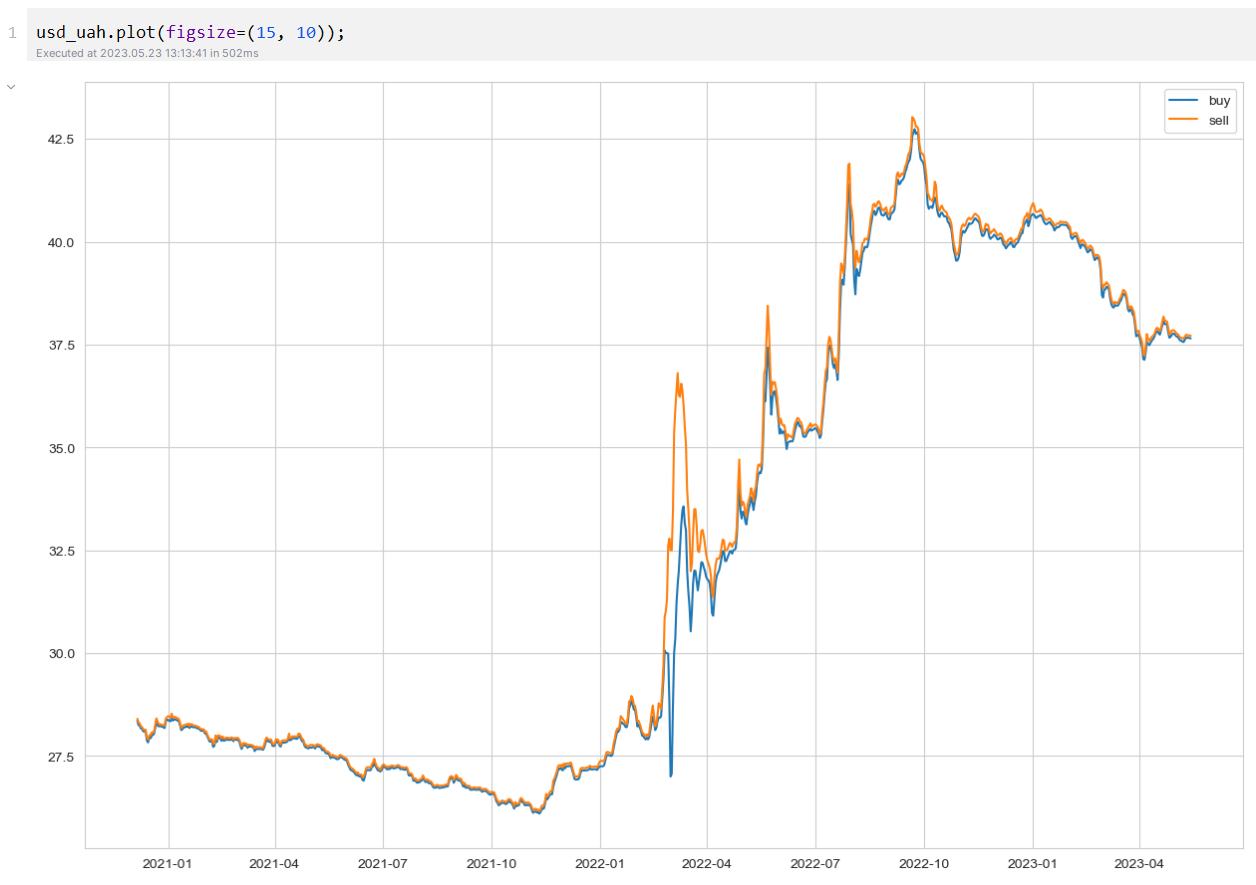


Рисунок 3.22 – Графік історичних даних про курс гривні до долара

Бачимо, що міститься інформація про ціну як на купівлю, так і на продаж. Використовуватимемо ціну на продаж для прогнозування, оскільки вона сильніше змінюється, але можна прогнозувати й ціну покупки.

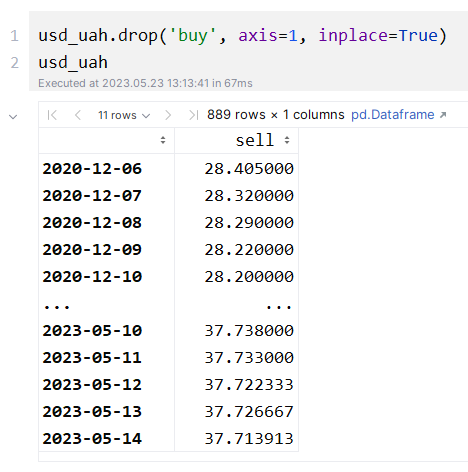


Рисунок 3.23 – Відкидання даних про ціну покупки долара

Виведемо графік із використанням ковзаючого середнього з розміром вікна рівним 10.



Рисунок 3.24 – Графік історичних даних про курс гривні до долара з використанням згладжування за допомогою ковзаючого середнього з розміром вікна 10

Тепер можемо декомпозувати послідовність на компоненти.

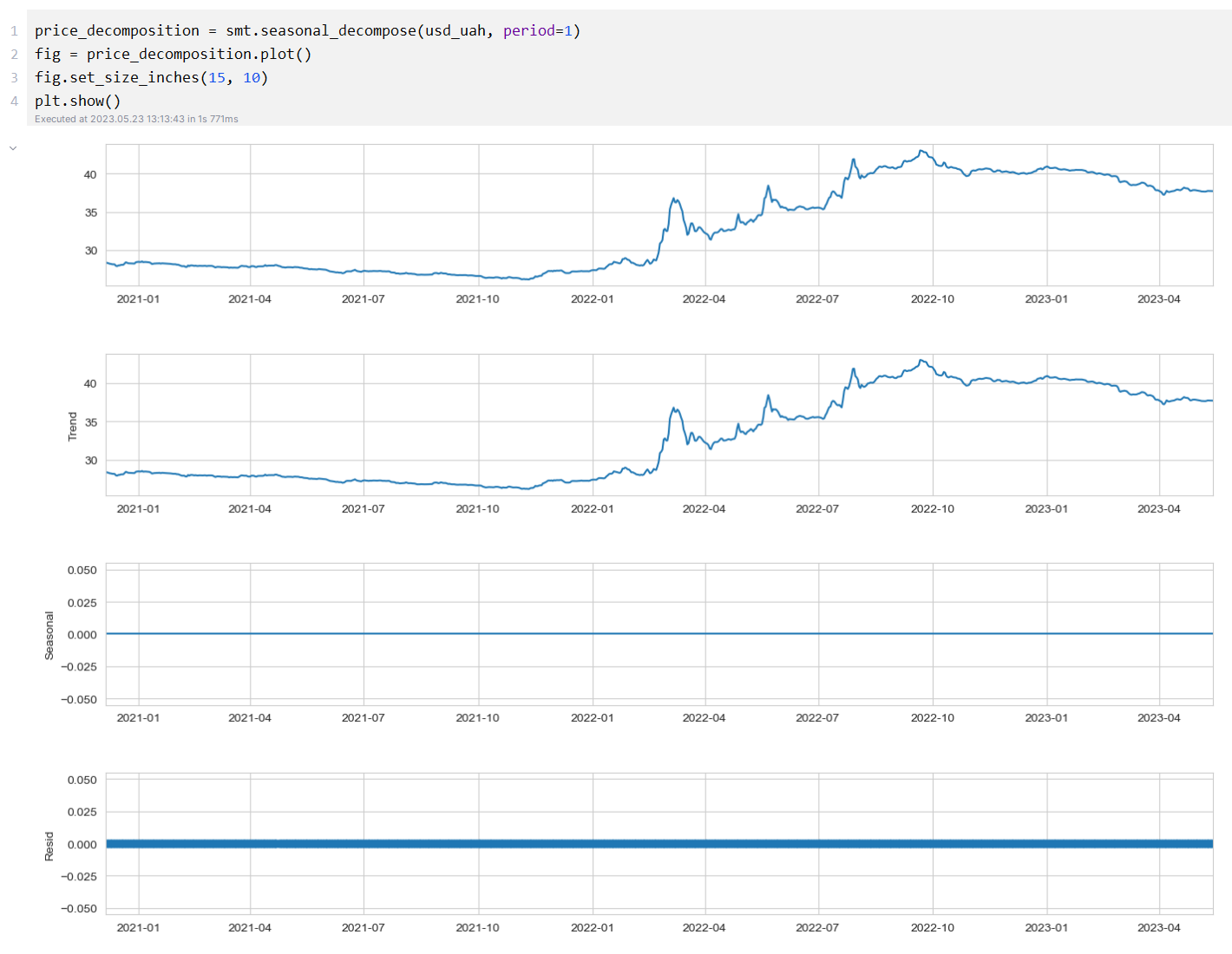


Рисунок 3.25 – Декомпозиція часової послідовності на тренд, сезонність та залишки

Бачимо, що якогось тренду, як і сезонності, немає. Тепер зобразимо графіки автокореляції та часткової автокореляції.

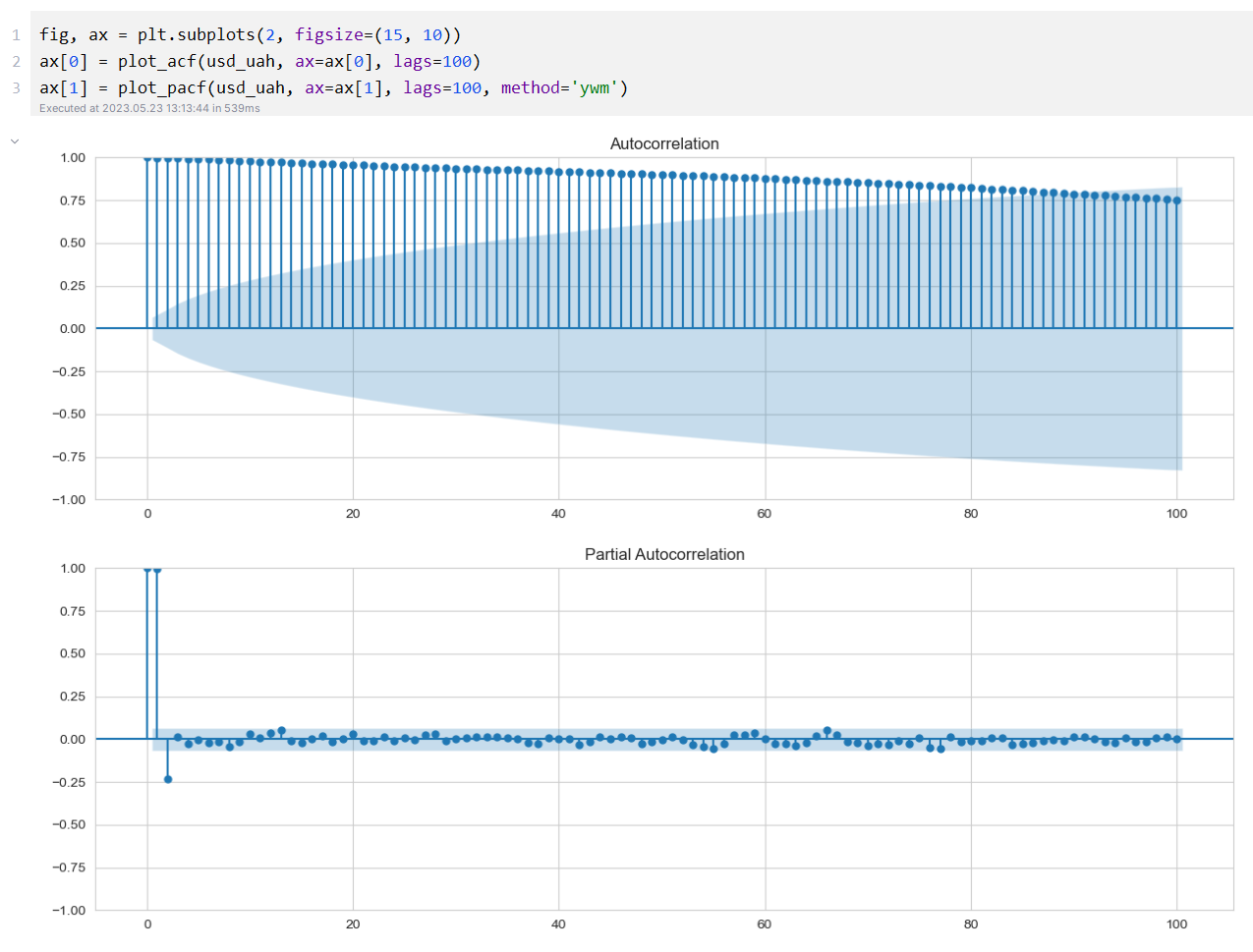


Рисунок 3.26 – Графік автокореляції та часткової автокореляції

Бачимо, що на графіку автокореляційної функції є дуже багато значущих лагів, а на графіку часткової автокореляції є тільки 1 значущий лаг, а це значить, що для того, щоб послідовність було приведено до стаціонарного вигляду, нам треба взяти перші різниці. Візуально ми можемо побачити, що ряд не є стаціонарним, але все одно виконаємо тест Дікі-Фуллера.

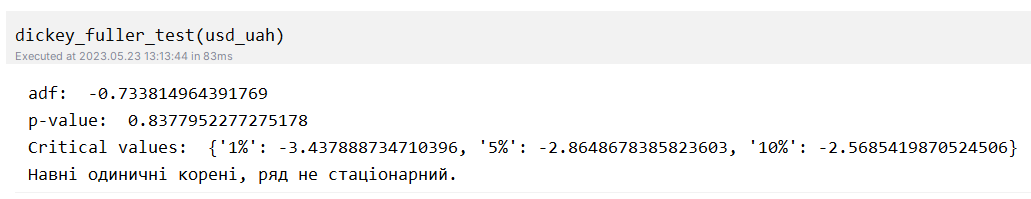


Рисунок 3.27 – Тест Дікі-Фуллера

Гіпотезу прийнято, ряд не є стаціонарним, як і було припущено. Візьмемо перші різниці й повторимо тест Дікі-Фуллера.

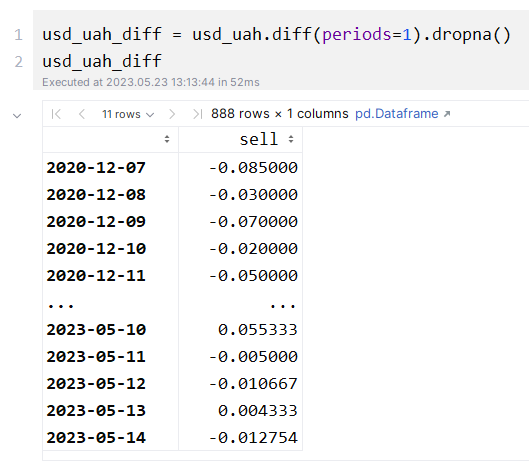


Рисунок 3.28 – Взяття перших різниць

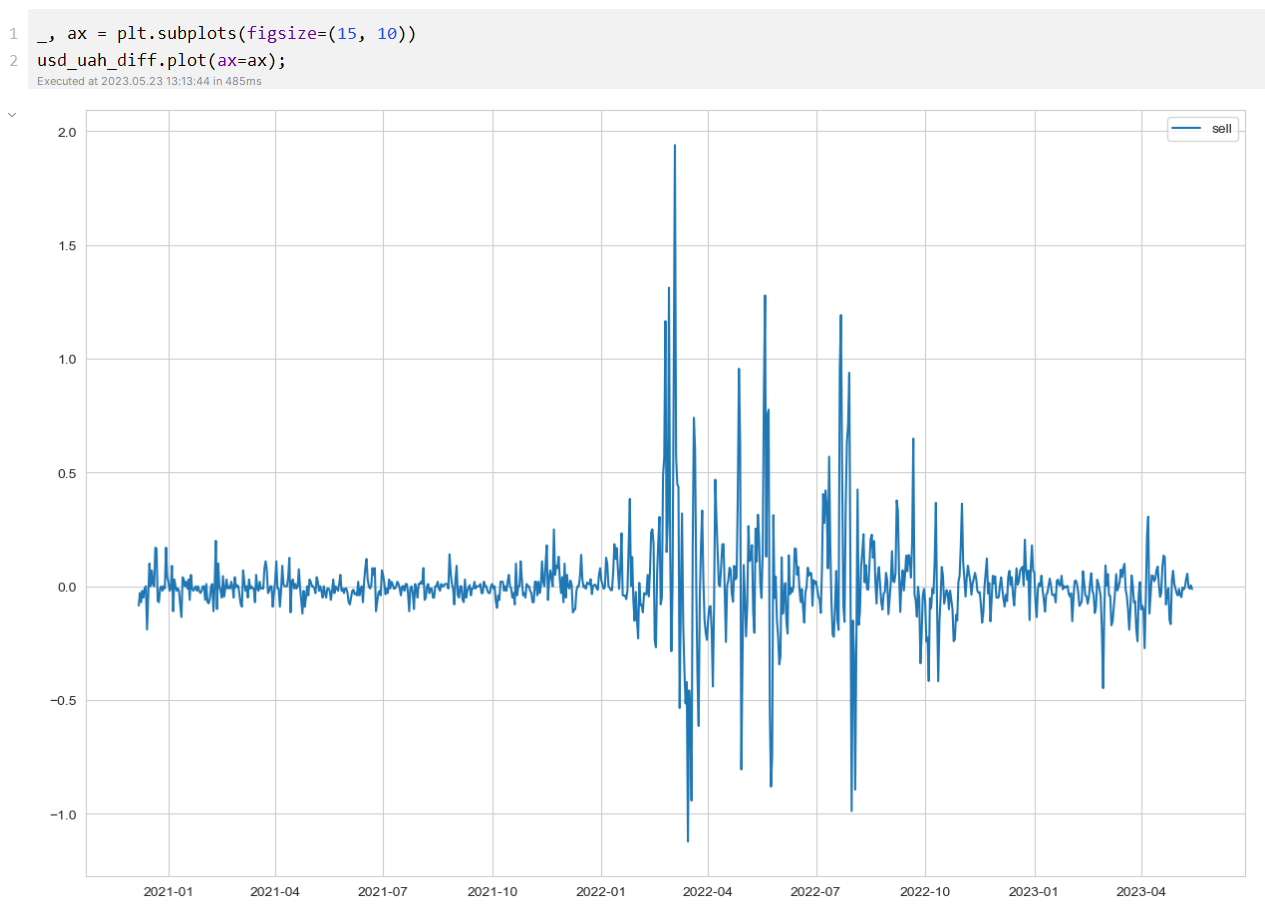


Рисунок 3.29 – Графік історичних даних про курс гривні до долара після взяття перших різниць

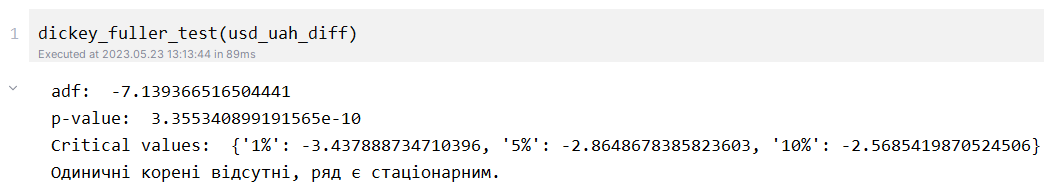


Рисунок 3.30 – Тест Дікі-Фуллера після взяття перших різниць

Тепер ряд і справді є стаціонарним. побудуємо для нього графіки автокореляції та часткової автокореляції.

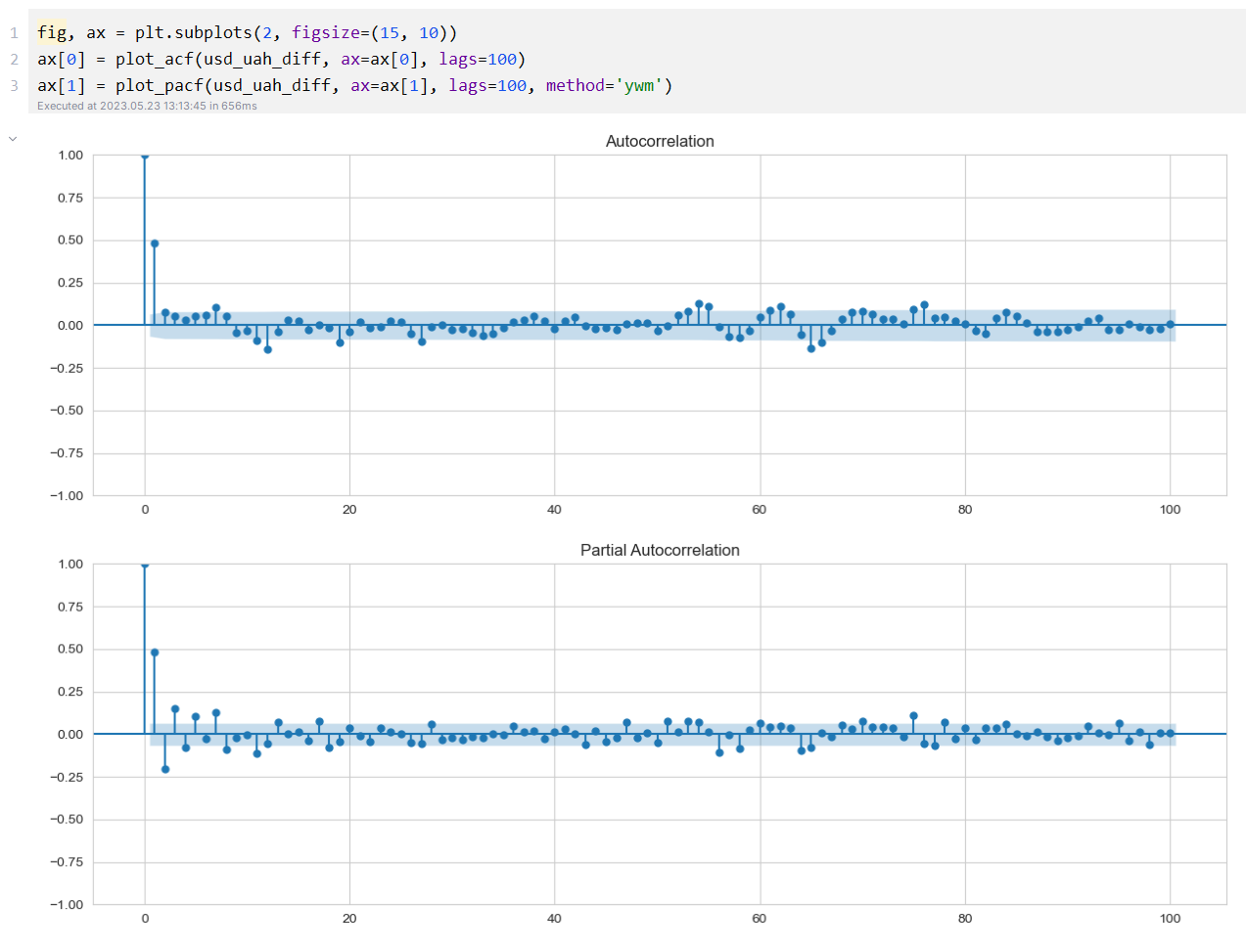


Рисунок 3.31 – Графік автокореляції та часткової автокореляції після взяття перших різниць

Бачимо на графіку автокореляції 1 значущий лаг, а значить параметр q для моделі ARIMA буде 1. На графіку часткової автокореляції бачимо 1 значущий лаг, а значить параметр p буде рівним 1. Побудуємо таку модель.

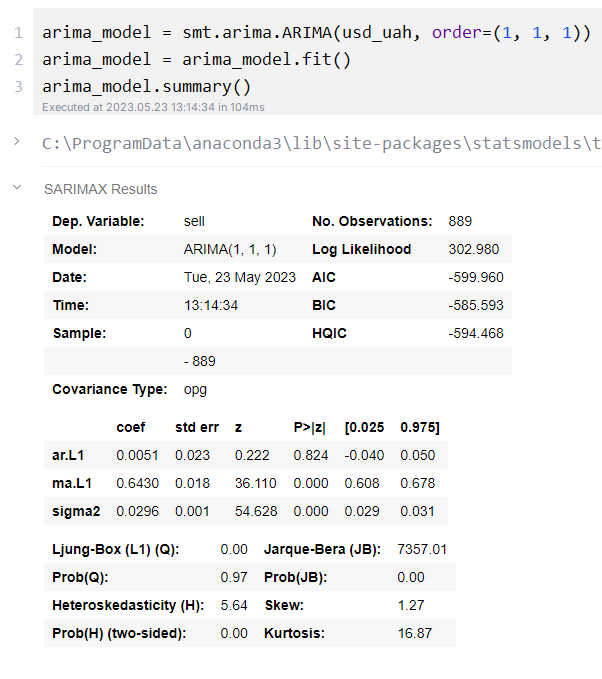


Рисунок 3.32 – Створення моделі ARIMA

Бачимо, що метрики вказують на доволі хороший результат. Тепер побудуємо передбачення й зобразимо їх на графіку.

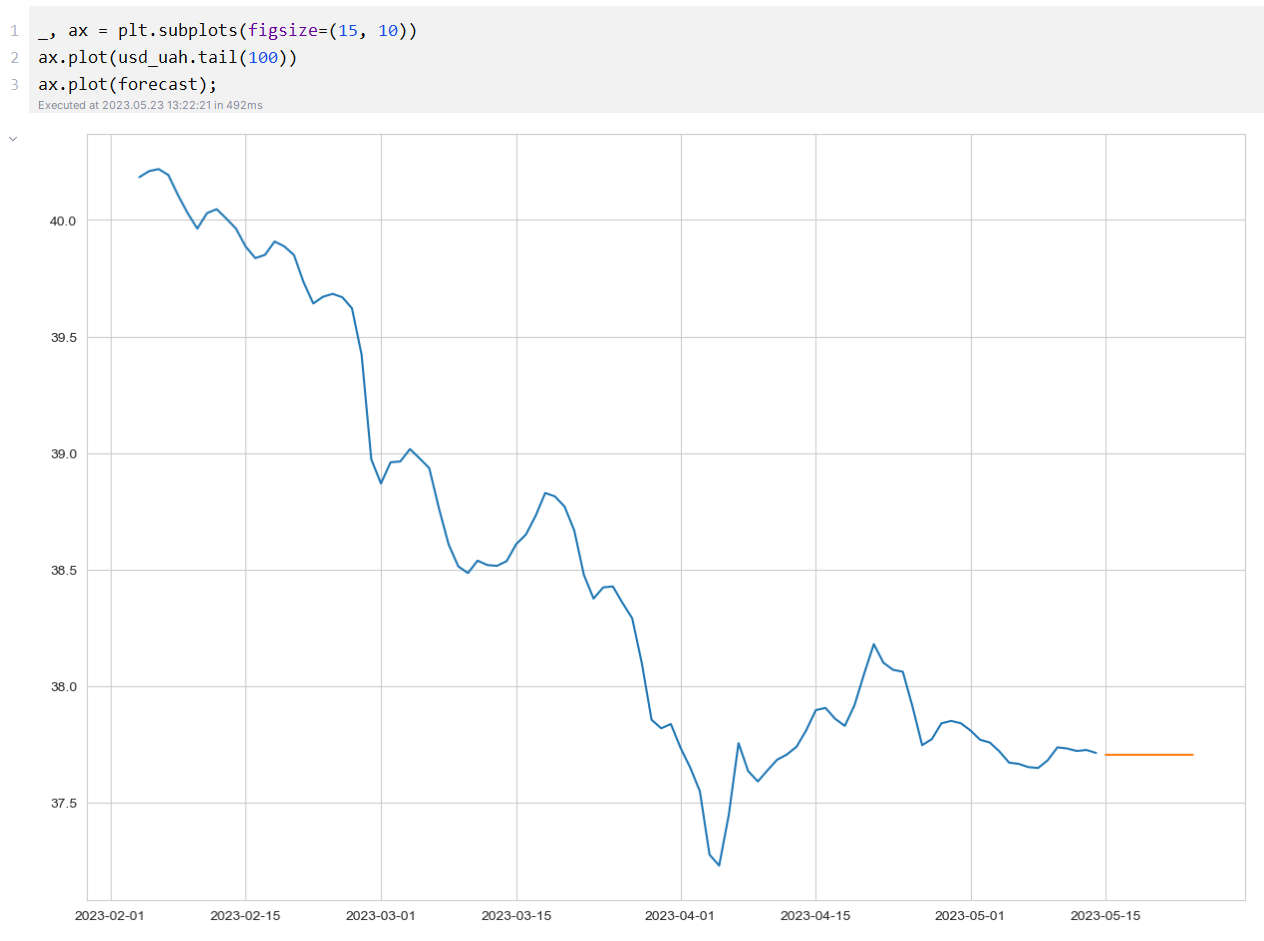


Рисунок 3.33 – Прогнозування курсу валют на наступні 100 днів моделлю ARIMA

На жаль, модель ARIMA не придатна для якісного довгострокового прогнозування, а тому спробуємо модель Prophet для вирішення цієї ж задачі. Для цієї моделі дуже важливо налаштувати гіперпараметр changepoint\_prior\_scale, який впливає на те, як модель буде реагувати на зміни в часовій послідовності. Для підбору цього параметру використаємо крос-валідацію. Будемо підбирати з такого набору: 0.001, 0.01, 0.1, 0.5, 5, 10.

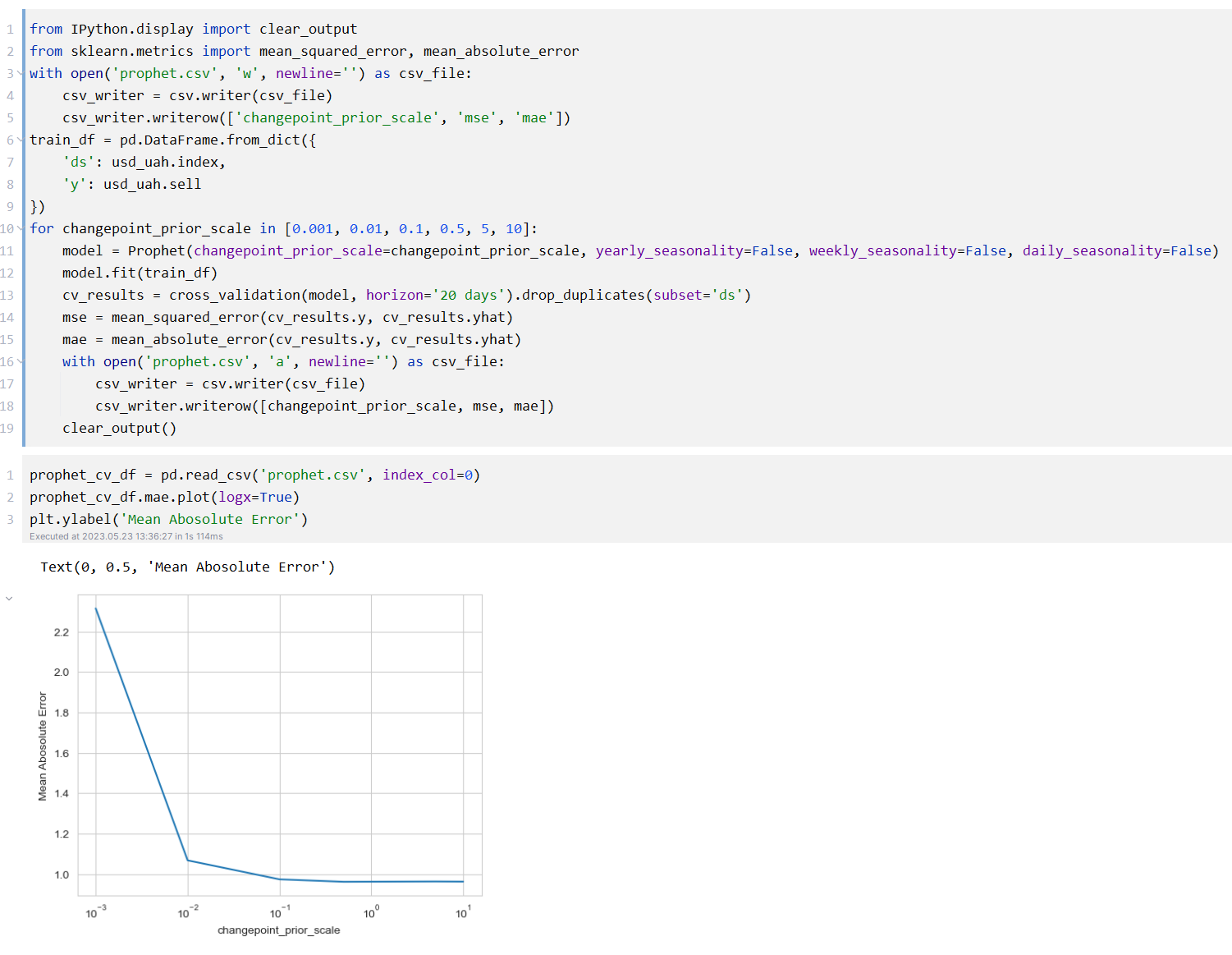


Рисунок 3.34 – Крос-валідація для підбору гіперпараметру changepoint\_prior\_scale моделі Prophet

Бачимо, що найкращим значенням параметру є будь-яке значення між 0.1 та 10. Будемо використовувати 5. Навчимо модель та виведемо наші прогнози на графіку.

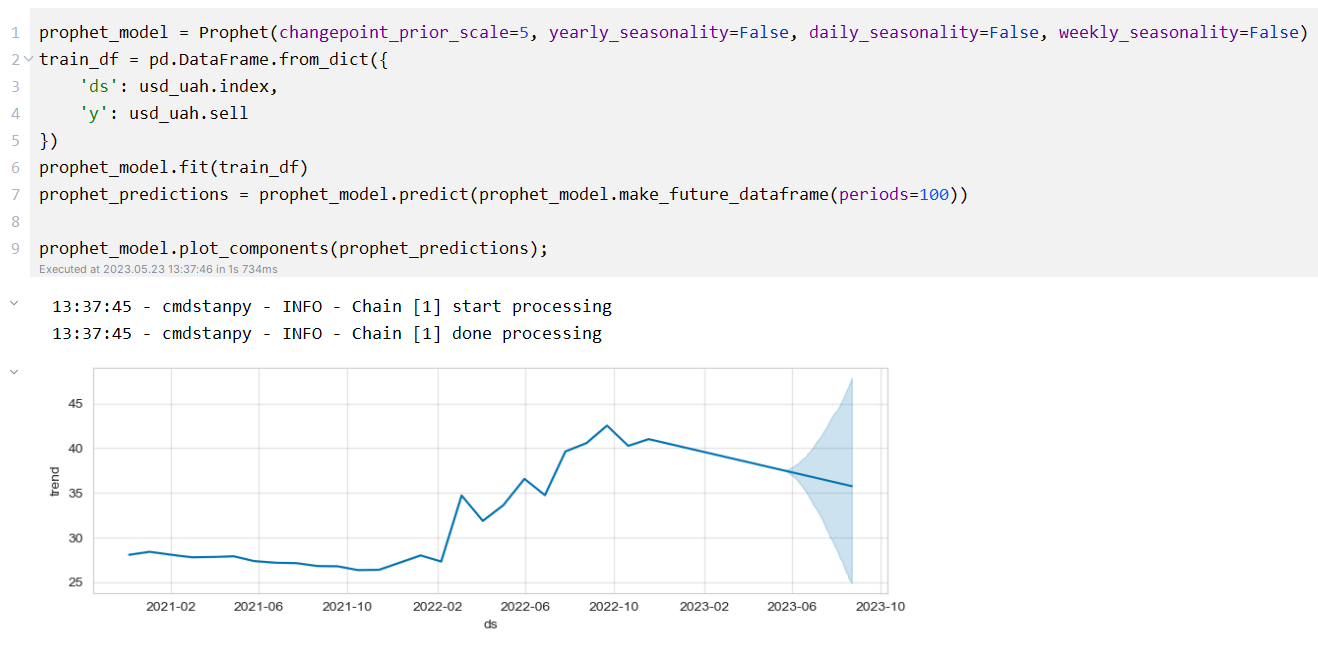


Рисунок 3.35 – Прогнозування курсу гривні до долара в наступні 100 днів



Рисунок 3.36 – Зображення прогнозу курсу гривні до долара в наступні 100 днів

Модель прогнозує подальший спад курсу долара. Ідеальний час, щоб вкластися в гривню.

# Виконання додаткового завдання

Для початку імпортуємо необхідні бібліотеки. Скористаємося двома моделями: ARIMA та Prophet.

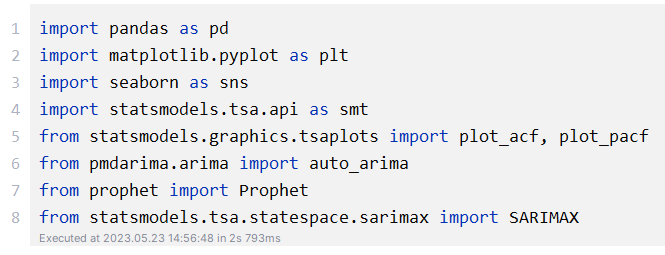


Рисунок 4.1 – Імпортування необхідних пакетів

Слідом завантажимо дані, перейменуємо стовпці для простоти роботи, встановимо дату як індекс, заповнимо пропущені значення опадів.

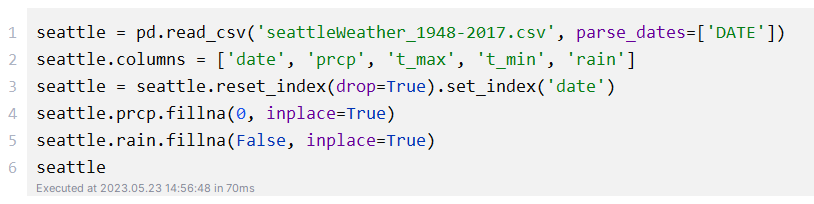


Рисунок 4.2 – Завантаження даних про погоду в Сіетлі

Після цього переведемо градуси Фаренгейта в градуси Цельсія за простою формулою: C = (F – 32) \* 5/9.

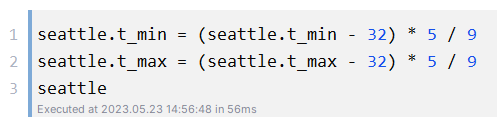


Рисунок 4.3 – Переведення температури зі шкали Фаренгейта в шкалу Цельсія

Виведемо матрицю кореляції.

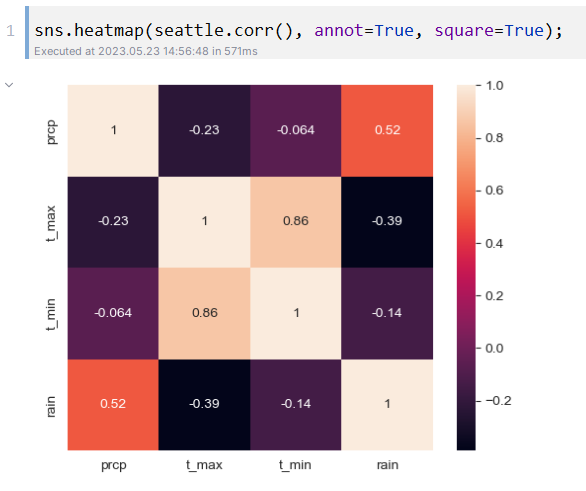


Рисунок 4.4 – Матриця кореляції

Бачимо, що кореляція слабка. Тепер виведемо всі дані на графіку.

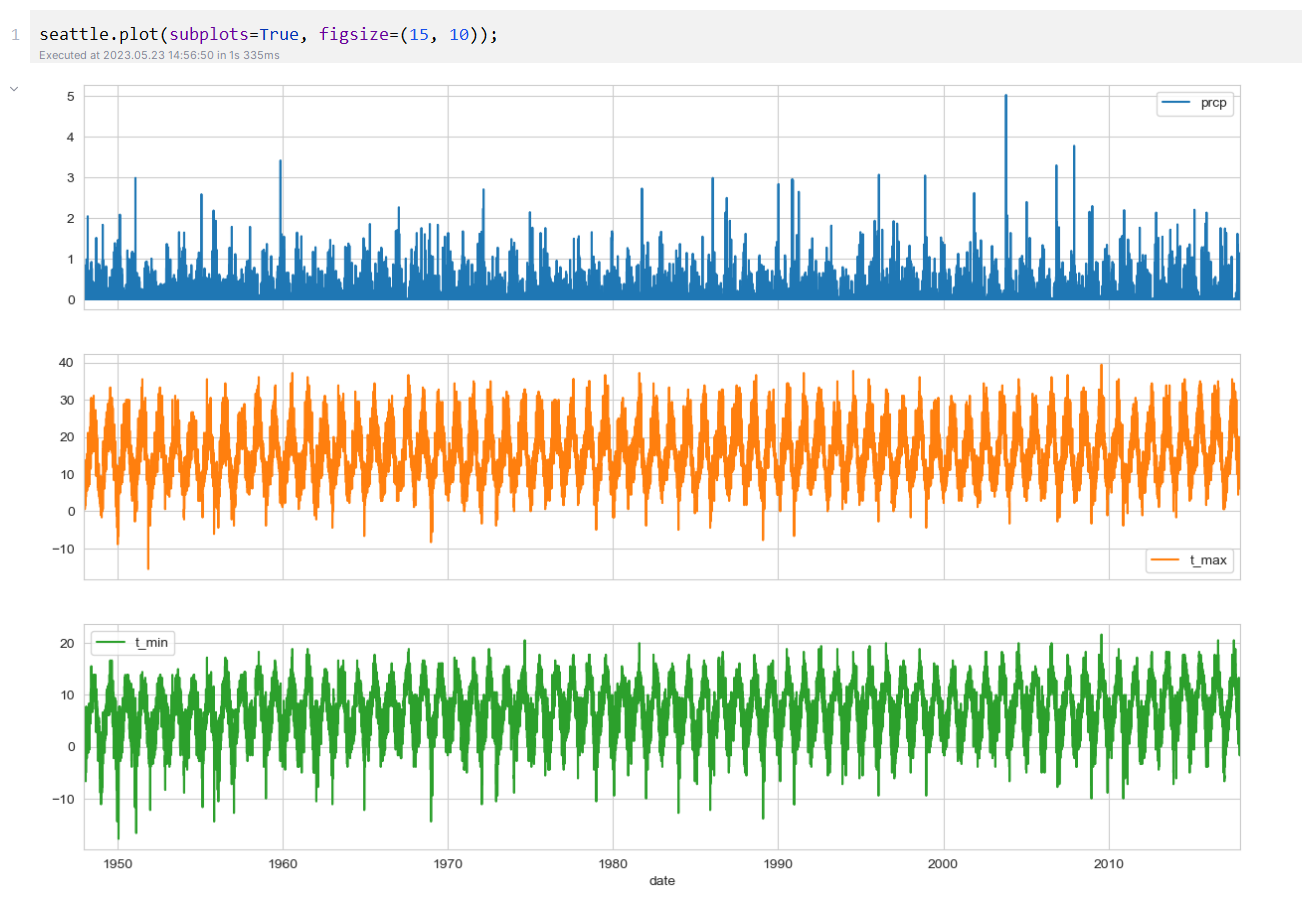


Рисунок 4.5 – Графік історичних даних про погоду

Оскільки часова послідовність занадто велика, виведемо графік тільки для останніх 2000 днів.

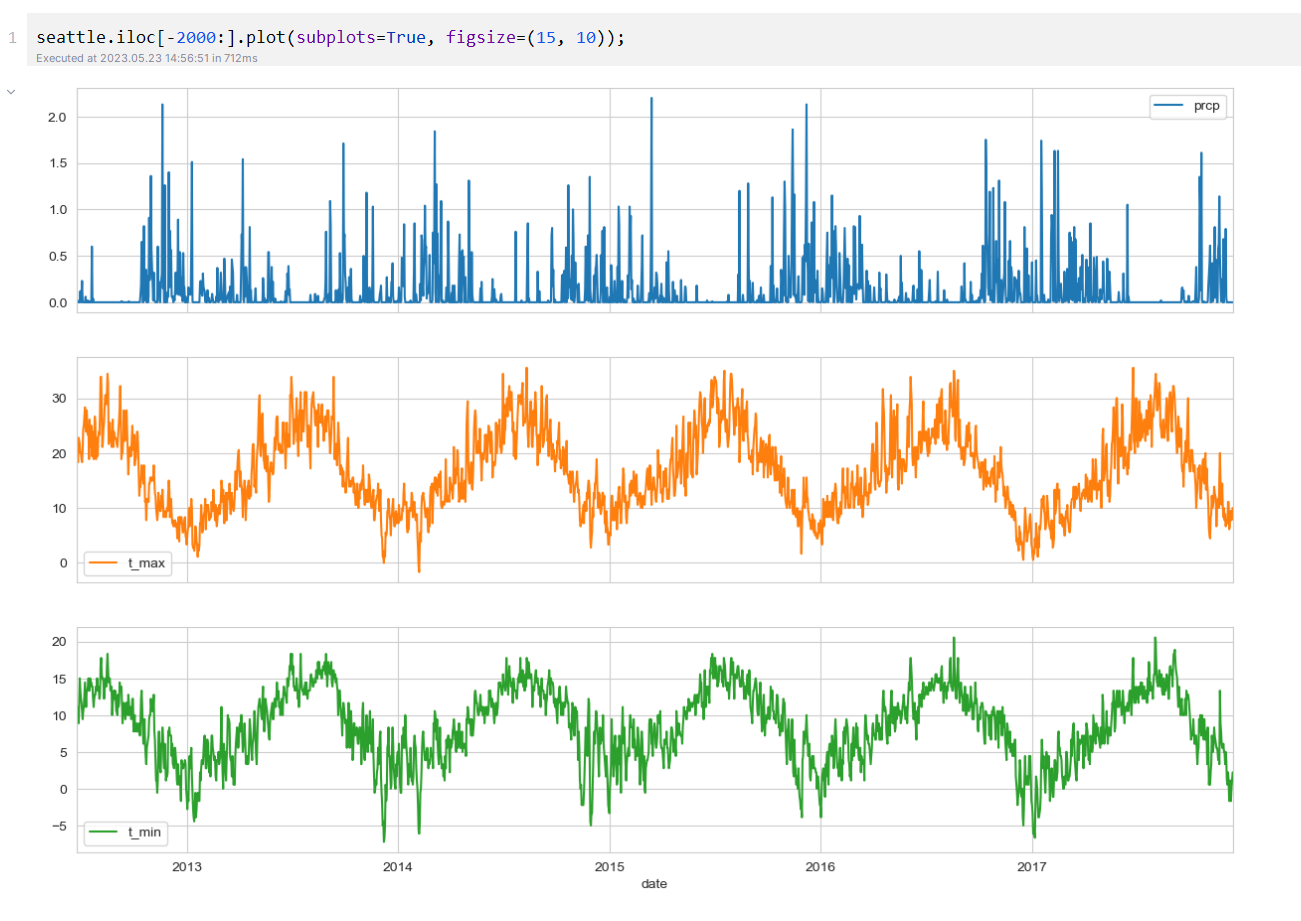


Рисунок 4.6 – Графік історичних даних про погоду за останні 2000 днів

Тут одразу прослідковується сезонна компонента. Масштабуємо дані до тижнів (обрахувавши середні показники для кожного дня тижня) та накреслимо такі ж графіки.

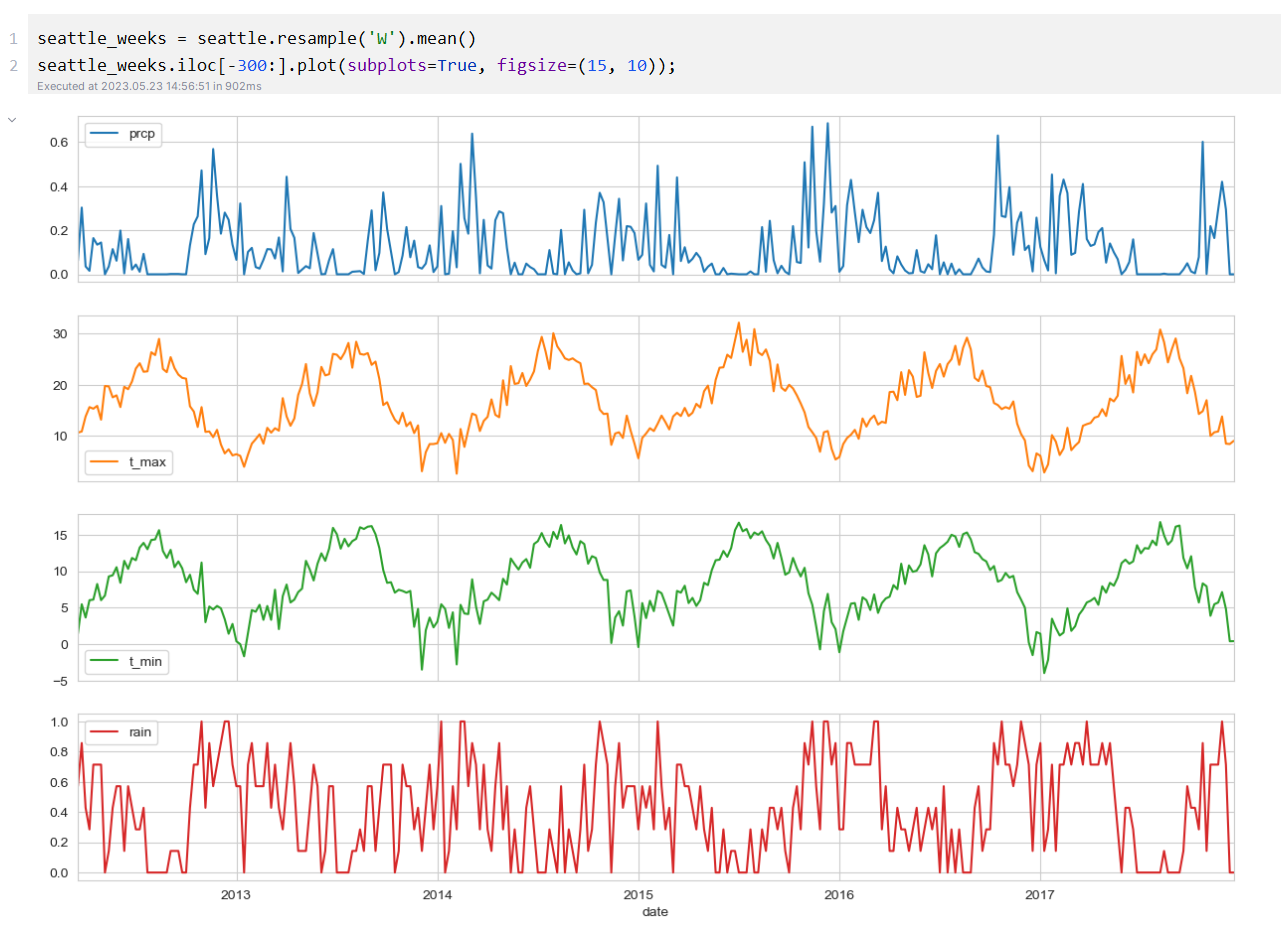


Рисунок 4.7 – Графік історичних даних за останні 300 днів

Далі декомпозуємо послідовність на тренд, сезонність та залишки.

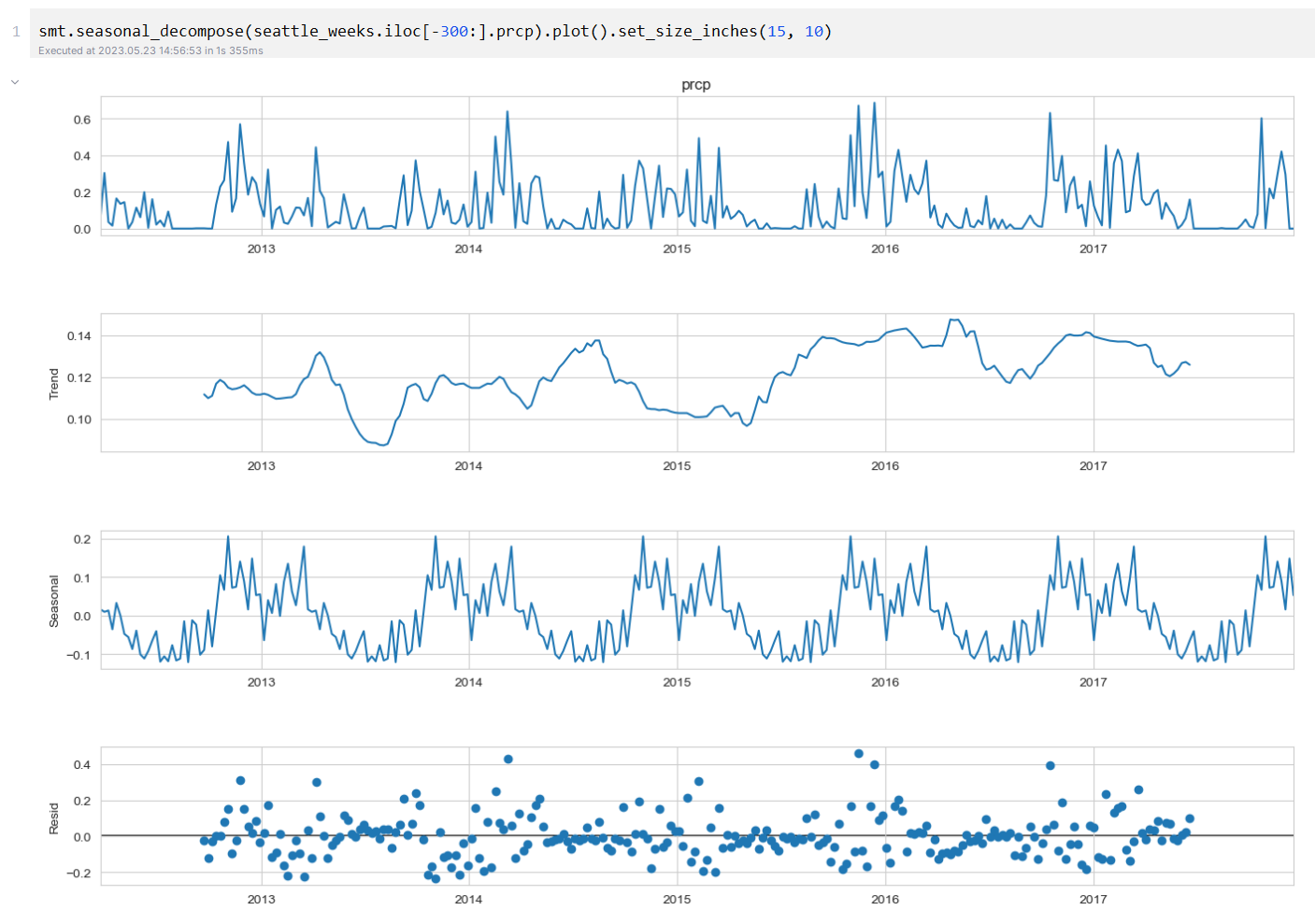


Рисунок 4.8 – Тренд, сезонність та залишки

Тут знову ж прослідковується сезонність. Проведемо для послідовності тест Дікі-Фуллера.

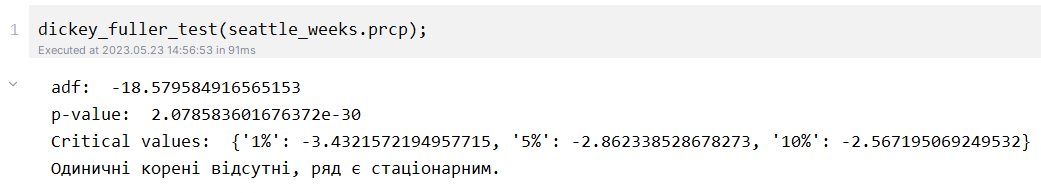


Рисунок 4.9 – Тест Дікі-Фуллера

Оскільки це історичні дані погоди, ряд, є стаціонарним. Побудуємо графіки автокореляції та часткової автокореляції.



Рисунок 4.10 – Графік автокореляції та часткової автокореляції

Далі поділимо дані про опади на дві вибірки: навчальну (до 2010 року) та тестову з (1 січня 2010 року до 31 грудня 2016 року), після чого спробуємо підібрати ідеальну модель за допомогою пакету pmdarima.

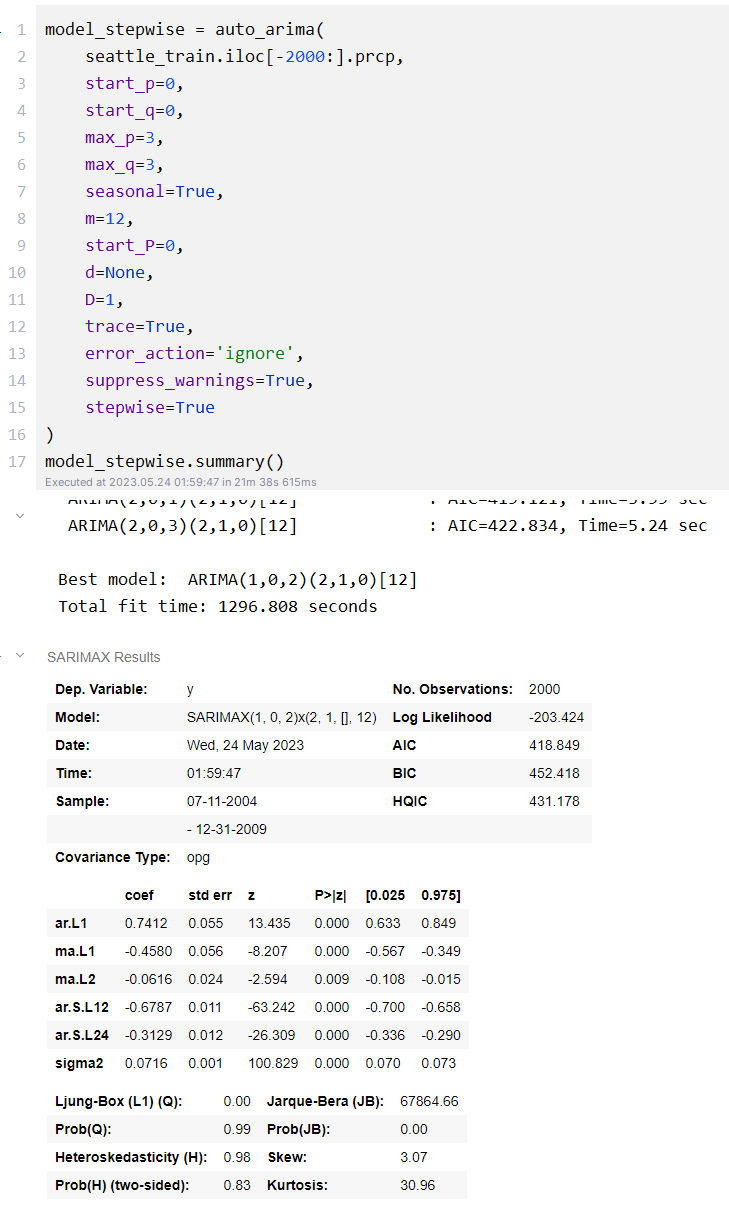


Рисунок 4.11 – Підбір моделі ARIMA

Функція auto\_arima повернула модель із такими параметрами: (p, d, q) (P, D, Q, s) = (1, 0, 2) (2, 1, 0, 12). Оскільки послідовність містить сезонну компоненту, функція й повернула модель SARIMAX. Спрогнозуємо значення на 4 місяці вперед.



Рисунок 4.12 – Прогнозування майбутніх значень опадів моделлю ARIMA

Бачимо, що ця модель погано прогнозує значення на велику дистанцію, але більш-менш добре – на малу.

Тепер скористаємося моделлю Prophet. В якості гіперпараметра changepoint\_prior\_scale знову використаємо значення 5, а сезонність залишимо шукати моделі самостійно. Зробимо передбачення на 4 роки вперед та оцінимо їхню точність.

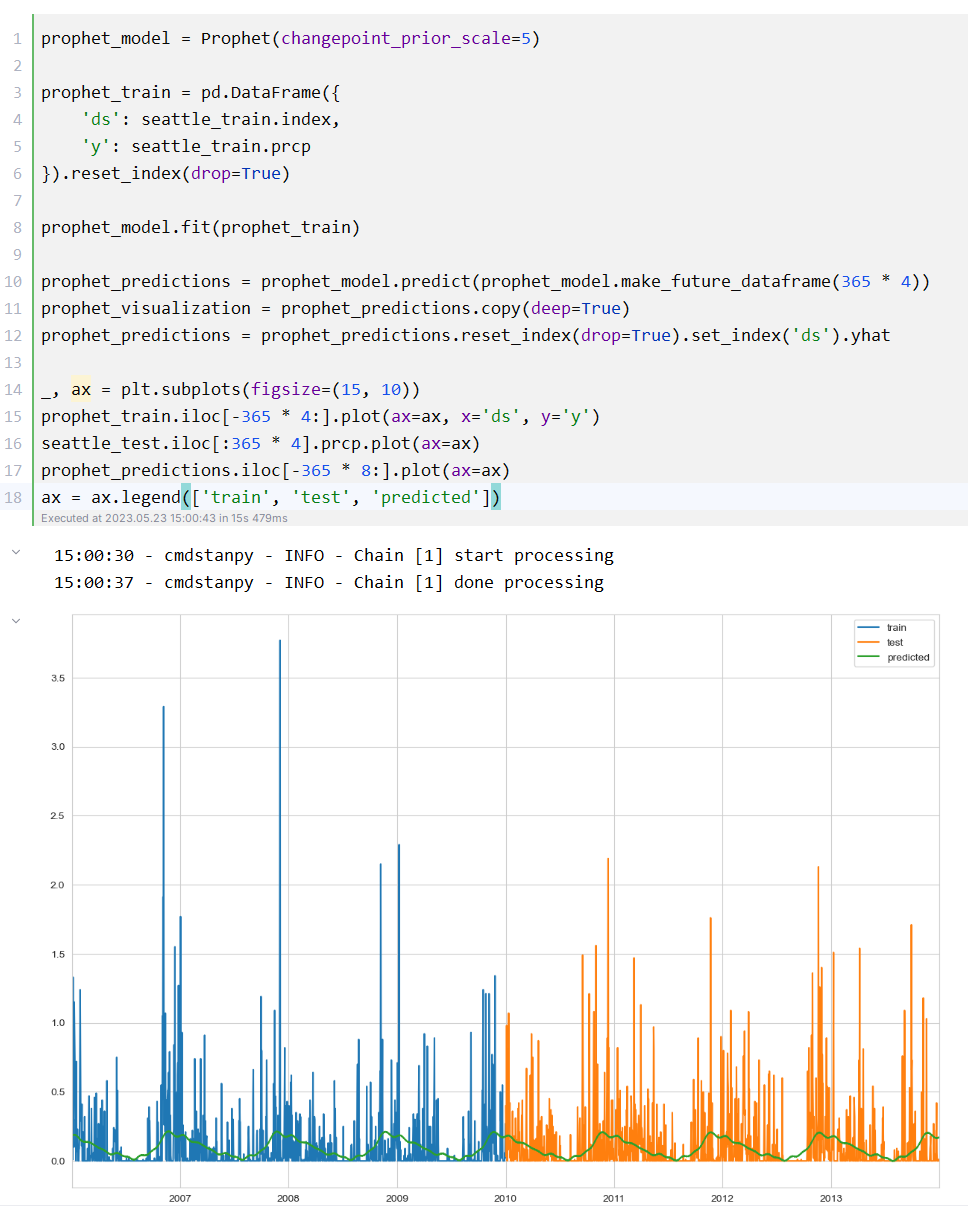


Рисунок 4.13 – Прогнозування майбутніх значень опадів моделлю Prophet

Як бачимо, прогноз дещо занижений, але сезонність відслідкована чітко. Виведемо компоненти прогнозу.

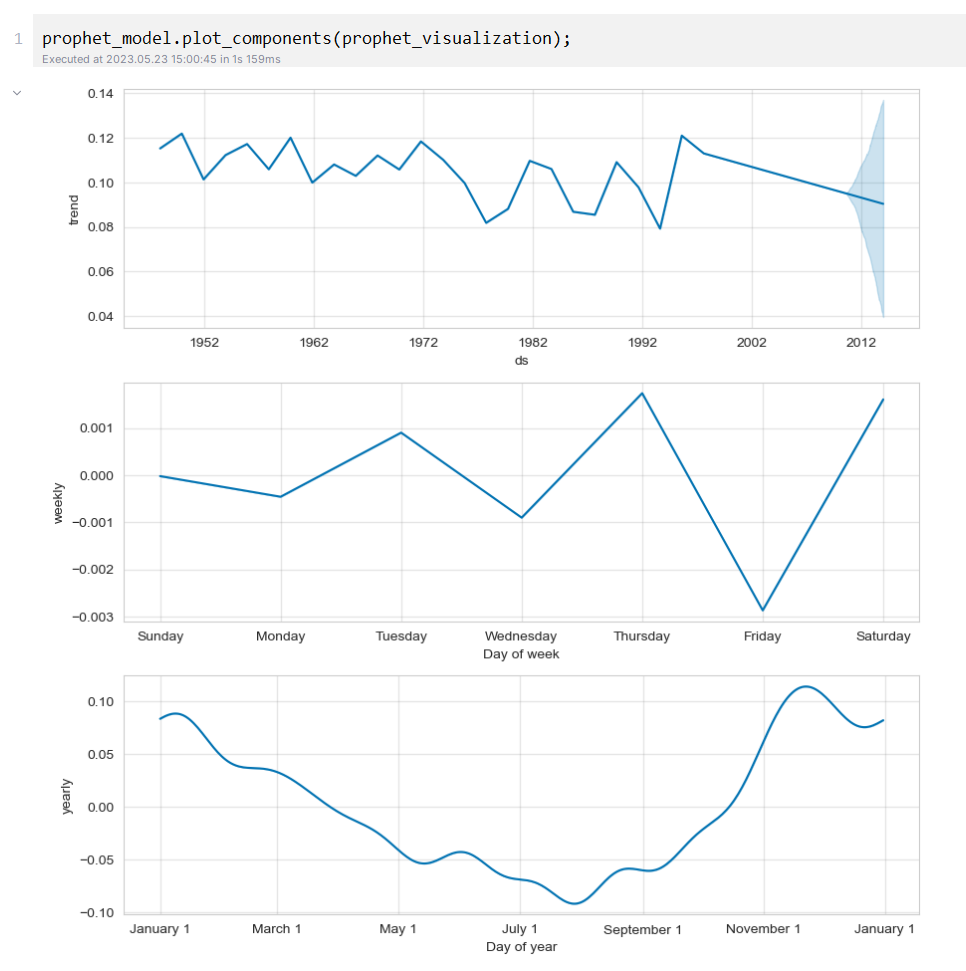


Рисунок 4.14 – Декомпозиція прогнозу моделі Prophet на тренд та сезонність

Останні 20 років спостерігається тренд на зменшення кількості опадів – це прослідковується абсолютно чітко на графіку тренду. Також наведено сезонність у межах тижня та року. У межах тижня вона вкрай незначна, фактично модель показує, що її не існує, а от у межах року вона прослідковується дуже чітко й можна якісно оцінити, коли в Сіетлі найбільше та найменше опадів.

При прогнозуванні кількості опадів у 2018 році картина залишиться незмінною, адже абсолютно очевидно, що моделі просто повторюють один і той самий «візерунок» через певні періоди часу. Відобразимо ці прогнози на графіку.

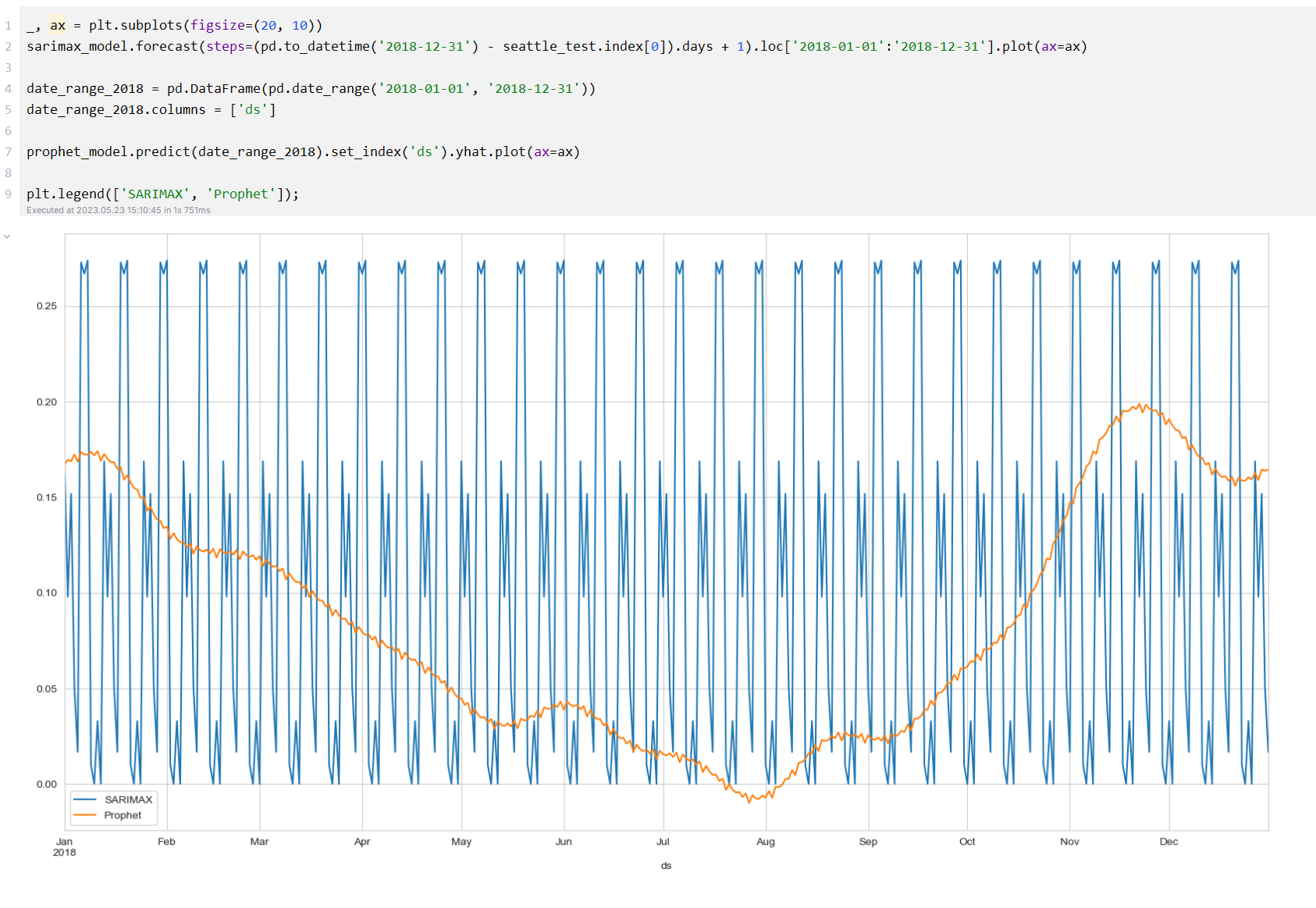


Рисунок 4.15 – Прогнозування кількості опадів на 2018 рік моделями ARIMA та Prophet

Бачимо, що ні про яку точність мова навіть не йде, але Prophet хоча б відобразив сезонність, у той час як SARIMAX просто прийшла до збіжності й почала прогнозувати одні й ті самі некоректні значення.

# Висновок

У ході даної лабораторної роботи було досліджено три часові ряди: захворюваності на COVID-19, курсу валют та опадів у Сіетлі. Було зображено графіки автокореляції та автокореляції, за якими визначалися параметри моделей класу ARIMA. Також було зображено графіки з використанням ковзаючого середнього й проведено тест Дікі-Фуллера для перевірки часової послідовності на стаціонарність. Набори даних про коронавірус та курс валют виявилися нестаціонарними, а от набір даних про опади в Сіетлі – стаціонарним. Було також проаналізовано ці часові послідовності на наявність сезонної компоненти. Вона виявилася знову тільки в останній часовій. послідовності.

Для часових послідовностей курсів валют та кількості опадів у Сіетлі було побудовано також моделі прогнозування: класу ARIMA та Prophet. Було виявлено їхні переваги та недоліки: на часовій послідовності курсів валют ARIMA не могла прогнозувати на великий період, а прогноз був доступний на дуже малий час уперед, у той час як Prophet робив приблизний прогноз майбутніх значень на досить велику дистанцію. Тим не менш, Prophet погано впорався з завданням прогнозування кількості опадів у Сіетлі, сильно занижуючи показники, а от SARIMAX добре прогнозувала значення, але знову на досить коротку дистанцію. Варто зауважити, що, не дивлячись на низькі значення, Prophet точно відображав сезонність: показував, що опади будуть у тому часовому проміжку, де вони й справді були, але показував, що опадів не буде в тому проміжку, у якому їх справді не було, тобто сезонність цією моделлю була відслідкована дуже добре.

Підсумовуючи, можна зазначити, що вибір моделі залежить від конкретної задачі й немає якогось універсального ключа до вирішення задач прогнозування часових послідовностей, а сам цей процес – дуже складний і має купу нюансів.