Міністерство освіти і науки України

Національний технічний університет України «Київський політехнічний

інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної техніки

Катедра інформатики та програмної інженерії

Звіт

з лабораторної роботи № 2 з дисципліни

«Прикладні задачі машинного навчання»

«Часові ряди й проста лінійна регресія»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірив Нестерук Андрій Олександрович

Київ 2023

**Лабораторна робота 2**

**Часові ряди й проста лінійна регресія**

**Постановка завдання.**

1. Завантажити метеорологічні дані 1895-2022 років з CSV-файлу в DataFrame. Після цього дані відформатувати для використання.
2. Бібліотеку Seaborn використати для графічного представлення даних DataFrame у вигляді регресійної прямої, що представляє графік зміни обраних показників за період 1895-2018 років.
3. Спрогнозувати дані на 2019, 2020, 2021 та 2022 рік.
4. Оцінити за формулою, які могли б бути показники до 1895 року.
5. Скористатися функцією regplot бібліотеки Seaborn для виведення всіх точок даних; дати представляються на осі x, а показники на осі y.
6. Виконати масштабування осі y.
7. Порівняти отриманий прогноз для 2019, 2020, 2021 та за 2022 роки з даними на NOAA «Climate at a Glance»: https://www.ncdc.noaa.gov/cag/ і зробити висновок.

**Хід роботи.**

1. Завантажую метеорологічні дані 1895-2018 років за січень у Нью-Йорку з CSV-файлу в DataFrame. Після цього дані форматую. Для зручності перейменовую назви стовпців так, щоб вони називалися «Date», «Temperature» й «Anomaly». Далі забираю позначення місяця в даті, адже маємо і так лише січень. Тож методом floordiv виконую цілочисельне ділення з кожним елементом, тож у даті отримуємо лише рік. Також знаходжу основні статистичні показники методом describe.

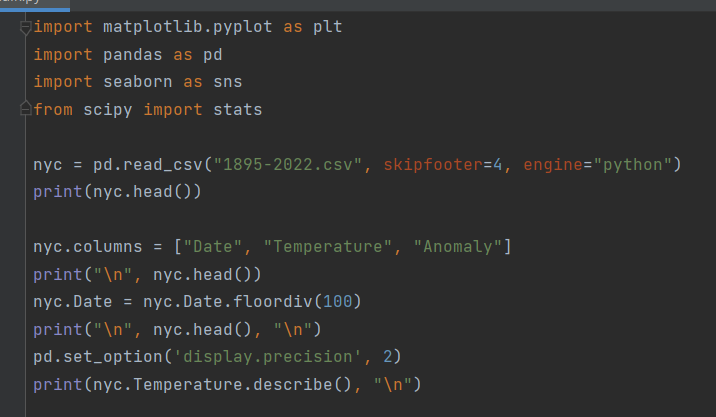


Рисунок 1 – Код до пункту 1.

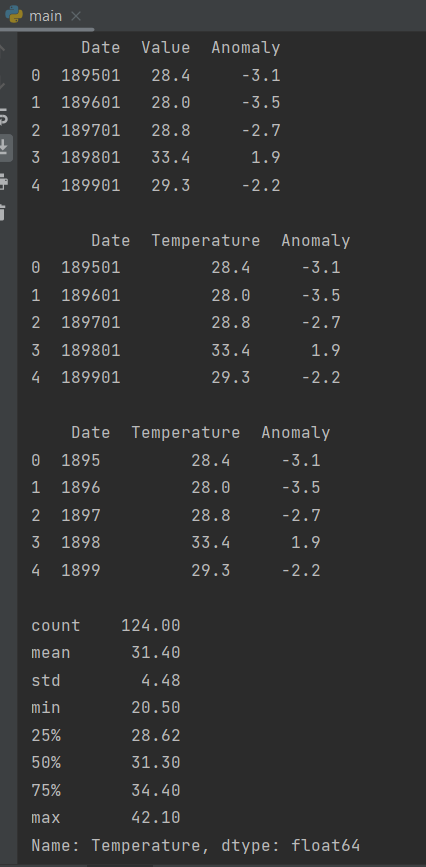


Рисунок 2 – Результат виконання пункту 1.

1. Бібліотеку Seaborn використовую для графічного представлення даних DataFrame у вигляді регресійної прямої, що представляє графік зміни обраних показників за період 1895-2018 років. Для відображення як зміни аномалій, так і температури, я обрав метод pairplot і вид графіку з регресивними прямими.

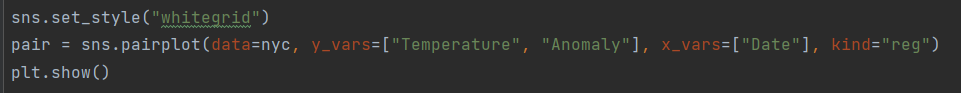


Рисунок 3 – Код до пункту 2.

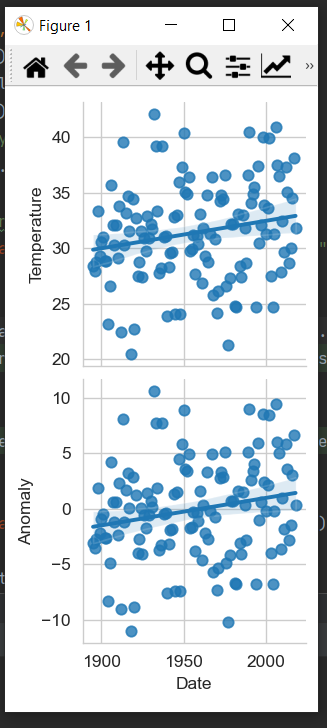


Рисунок 4 – Результат виконання пункту 2.

1. Прогнозую дані на 2019, 2020, 2021 та 2022 рік. Для цього використовую функцію scipy.stats.linregress(). У результаті отримуємо кут нахилу і точку перетину регресійної прямої, що можна об'єднати з рівнянням простої лінійної регресії для прямої лінії, *y = mx + b* при прогнозуванні середньої січневої температури.

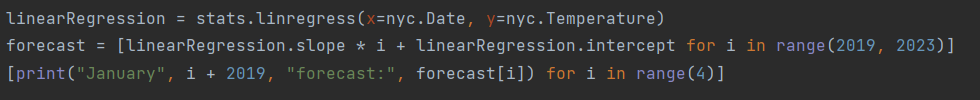


Рисунок 5 – Код до пункту 3.

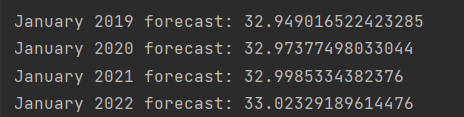


Рисунок 6 – Результат виконання пункту 3.

1. Оцінюю показники до 1895 року. Я вирішив оцінити останні недосліджені п’ять років, тобто 1890-1894 роки. Виконую дії, аналогічні до пункту 3.

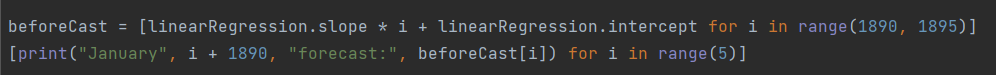


Рисунок 7 – Код до пункту 4.

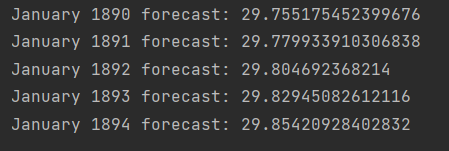


Рисунок 8 – Результат виконання пункту 4.

1. Користуюся функцією regplot бібліотеки Seaborn для виведення всіх точок даних; дати представляються на осі x, а показники температури - на осі y.

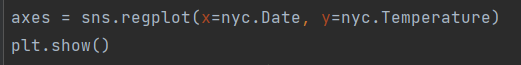


Рисунок 9 – Код до пункту 5.

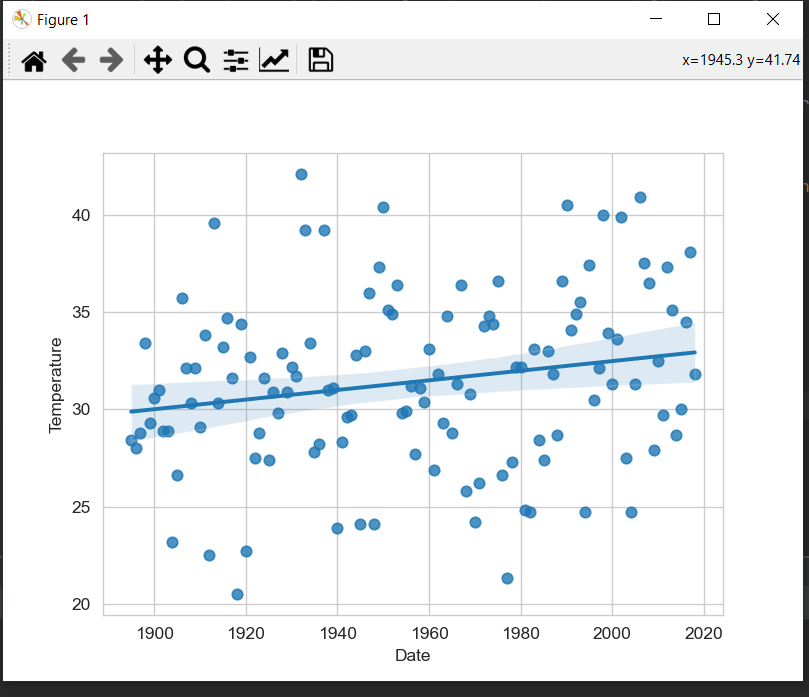


Рисунок 10 – Результат виконання пункту 5.

1. Виконую масштабування осі y за допомогою методу set\_ylim(10,70) для 10<=y<=70.

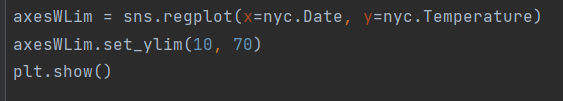
****

Рисунок 11 – Код до пункту 6.

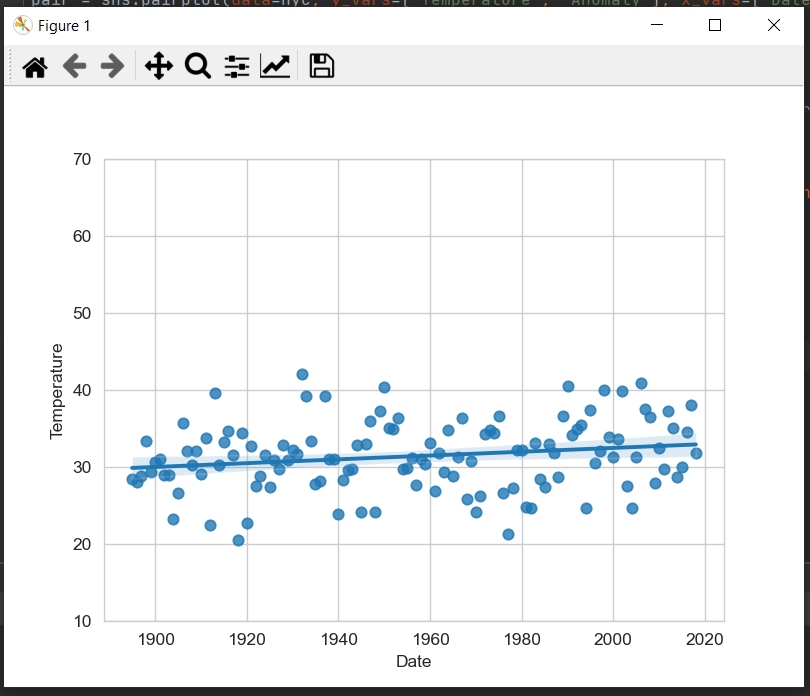


Рисунок 12 – Результат виконання пункту 6.

1. Порівнюю отриманий прогноз для 2019, 2020, 2021 та за 2022 роки з даними на NOAA «Climate at a Glance». Отже, на сайті отримуємо такі дані:

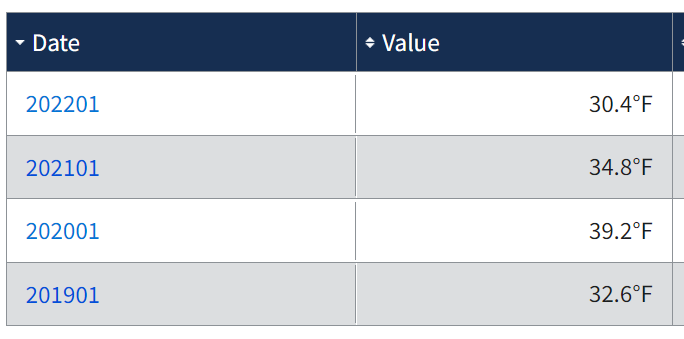


Рисунок 13 – Дані середньомісячної температури 2019-2022 років.

Тоді як спрогнозовані дані такі:

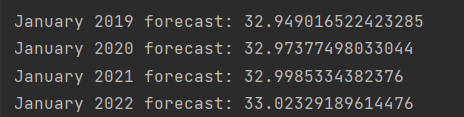


Рисунок 14 – Спрогнозовані дані за 2019-2022 роки.

Бачимо, що фактичні дані не зовсім збігаються з прогнозованими як за самими значеннями температури, так і за тенденцією, якщо рахувати лише за ці 4 роки. Це пов’язано з тим, що прогноз базується на регресійній прямій, відповідно значення на прямій мають тенденцію до зростання, а отже щорічний прогноз буде мати все більші і більші значення температури. Водночас реальні дані за 4 роки, звісно, коливаються так, як це фактично сталося фізично, тому хоч температура за кожен рік все ж не набагато відрізняється від прогнозу, але все ж різниться, а також тенденція у розрізі двох років відбувається то на зростання, то на спадання. Це нормально, адже глобальну тенденцію на зростання можливо помітити в розрізі десятиліть і століть, і неможливо на настільки малій вибірці.

**Висновок.**

Отже, у цій роботі я отримав навички роботи з часовими рядами й простими лінійними регресіями, їх обробки, виведення статистики, візуалізації та вчинення відповідних дій. Часові ряди разом з регресіями дійсно стають в нагоді під час прогнозування даних, які пов’язані з часовими проміжками, такими як щорічна температура в конкретному місяці й місці, що було досліджено в лабораторній роботі. У результаті лабораторної роботи було вивчено роботу з такими засобами, як scipy.stats.linregress, seaborn.pairplot(), seaborn.regplot(), створено різні графіки з регресійними прямими, проведено дослідження різних статистик, прогнозування даних на наступні роки на основі попередніх та порівняння їх з наявними. Використовуючи програмний засіб PyCharm, мову програмування Python з модулями Pandas, matplotlib, seaborn та scipy, отримуємо коректний результат.