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 №2 з дисципліни
«Прикладні задачі машинного навчання»

«Часові ряди і проста лінійна регресія»

Виконав: *ІП-13 Ал Хадам Мурат Резгович*
(шифр, прізвище, ім'я, по батькові)

Перевірив: *Нестерук Андрій Олександрович*
(прізвище, ім'я, по батькові)

Київ 2023

Лабораторна робота №2

Тема: Часові ряди і проста лінійна регресія

Мета: Завантажити метеорологічні дані, відформатувати, побудувати графік за допомогою Seaborn, спрогнозувати майбутні значення, оцінити минулі значення, порівняти з NOAA Climate at a Glance.

Постановка задачі

1. В даній лабораторній роботі Вам треба завантажити метеорологічні дані в 1895-2022 роках з CSV-файлу в DataFrame. Після цього дані треба буде 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. Спрогнозуйте дані на 2019, 2020, 2021 та 2022 рік.
4. Оцініть за формулою, якою могли б бути показники до 1895 року. Наприклад, оцінка середньої температури за січень 1890 року може бути отримана наступним чином:
5. Скористайтесь функцією regplot бібліотеки Seaborn для виведення всіх точок даних; дати представляються на осі x, а показники на осі y. Функція regplot будує діаграму розкиду даних, на якій точки представляють показники за заданий рік, а пряма лінія - регресійну пряму.
6. Виконайте маштабування осі *y* від (приклад від 10 до 70 градусів):
7. Порівняйте 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. В даній лабораторній роботі Вам треба завантажити метеорологічні дані в 1895-2022 роках з CSV-файлу в DataFrame. Після цього дані треба буде відформатувати для використання.

Завантажимо середні січневі температури в Нью-Йорку з 1895 по 2023 рік через часові ряди NOAA «Climate at a Glance». Імпортуємо потрібні бібілотеки та записуємо CSV файл у DataFrame:

```
In 35
import pandas as pd
pd.set_option('display.precision', 2)

df = pd.read_csv('https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USH00305801/tavg/1/1/1895-2023.csv?base_prd=true&begbaseyear=1901&endbaseyear=2000')
df
```

Out 35 — 1-10 — 133 rows × 4 columns

	New York	New York	Average Temperature	January
0	Units: Degrees Fahrenheit	NaN	NaN	NaN
1	Base Period: 1901-2000	NaN	NaN	NaN
2	Missing: -99	NaN	NaN	NaN
3	Date	Value	Anomaly	NaN
4	189501	29.6	-2.0	NaN
5	189601	28.8	-2.8	NaN
6	189701	29.6	-2.0	NaN
7	189801	34.2	2.6	NaN
8	189901	30.1	-1.5	NaN
9	190001	31.5	-0.1	NaN

Перевіримо чи має наш датафрейм пропущені значення:

```
In 36
df.isnull().any()
```

Out 36 — 4 rows — Length: 4, dtype: bool

	data
New York	False
New York	True
Average Temperature	True
January	True

Видалимо колонку “January” та 4 перших непотрібних рядки. Пронумеруємо рядки з нуля:

```
In 37
df = df.drop(columns=' January', axis=1)
df = df.drop(df.index[:4])
df.index = range(len(df))
df
```

Out 37

1-10 | 129 rows × 3 columns

	Year	Temperature	Anomaly
0	1895	29.6	-2.0
1	1896	28.8	-2.8
2	1897	29.6	-2.0
3	1898	34.2	2.6
4	1899	30.1	-1.5
5	1900	31.5	-0.1
6	1901	31.8	0.2
7	1902	29.7	-1.9
8	1903	29.7	-1.9
9	1904	24.0	-7.6

Перейменуємо колонки:

```
In 38
df.columns = ['Year', 'Temperature', 'Anomaly']
df
```

Out 38

1-10 | 129 rows × 3 columns

	Year	Temperature	Anomaly
0	1895	29.6	-2.0
1	1896	28.8	-2.8
2	1897	29.6	-2.0
3	1898	34.2	2.6
4	1899	30.1	-1.5
5	1900	31.5	-0.1
6	1901	31.8	0.2
7	1902	29.7	-1.9
8	1903	29.7	-1.9
9	1904	24.0	-7.6

Перевіримо типи даних значень наших колонок та змінимо типи данних колонок на числові:

In 39
```
df.dtypes
```

Out 39

3 rows Length: 3, dtype: object

	data
Year	object
Temperature	object
Anomaly	object

In 40
```
df['Year'] = df['Year'].str[:4].astype(int)
df['Temperature'] = df['Temperature'].astype(float)
df['Anomaly'] = df['Anomaly'].astype(float)
df
```

Out 40

1-10 129 rows × 3 columns

	Year	Temperature	Anomaly
0	1895	29.6	-2.0
1	1896	28.8	-2.8
2	1897	29.6	-2.0
3	1898	34.2	2.6
4	1899	30.1	-1.5
5	1900	31.5	-0.1
6	1901	31.8	0.2
7	1902	29.7	-1.9
8	1903	29.7	-1.9
9	1904	24.0	-7.6

In 41
```
df.dtypes
```

Out 41

3 rows Length: 3, dtype: object

	data
Year	int32
Temperature	float64
Anomaly	float64

Знайдемо описову статистику для даних температур:

```
In 42   1  df['Temperature'].describe()

Out 42      8 rows    Length: 8, dtype: float64
```

	Temperature
count	129.00
mean	31.92
std	4.54
min	21.20
25%	29.00
50%	31.70
75%	34.80
max	43.50

3. Спрогнозуйте дані на 2019, 2020, 2021 та 2022 рік.

З бібліотеки scipy імпортуємо модуль stats. Використаємо функцію linregress, яка обчислює нахил і точку перетину регресійної прямої для заданого набору точок даних:

Напишемо функцію, яка прийматиме рік та повертатиме спрогнозовану температуру. Спрогнозуємо температуру на січень 2019 – 2022 років:

```
Predict temp

In 44
def predict_temp(year):
    return linear_regression.slope * year + linear_regression.intercept


for year in range(2019, 2023):
    temp = predict_temp(year)
    print(f"Predicted temperature for {year} year is {temp}")

Predicted temperature for 2019 year is 33.21643932617769
Predicted temperature for 2020 year is 33.237979464818125
Predicted temperature for 2021 year is 33.25951960345855
Predicted temperature for 2022 year is 33.28105974209898
```

4. Оцініть за формулою, якою могли б бути показники до 1895 року.

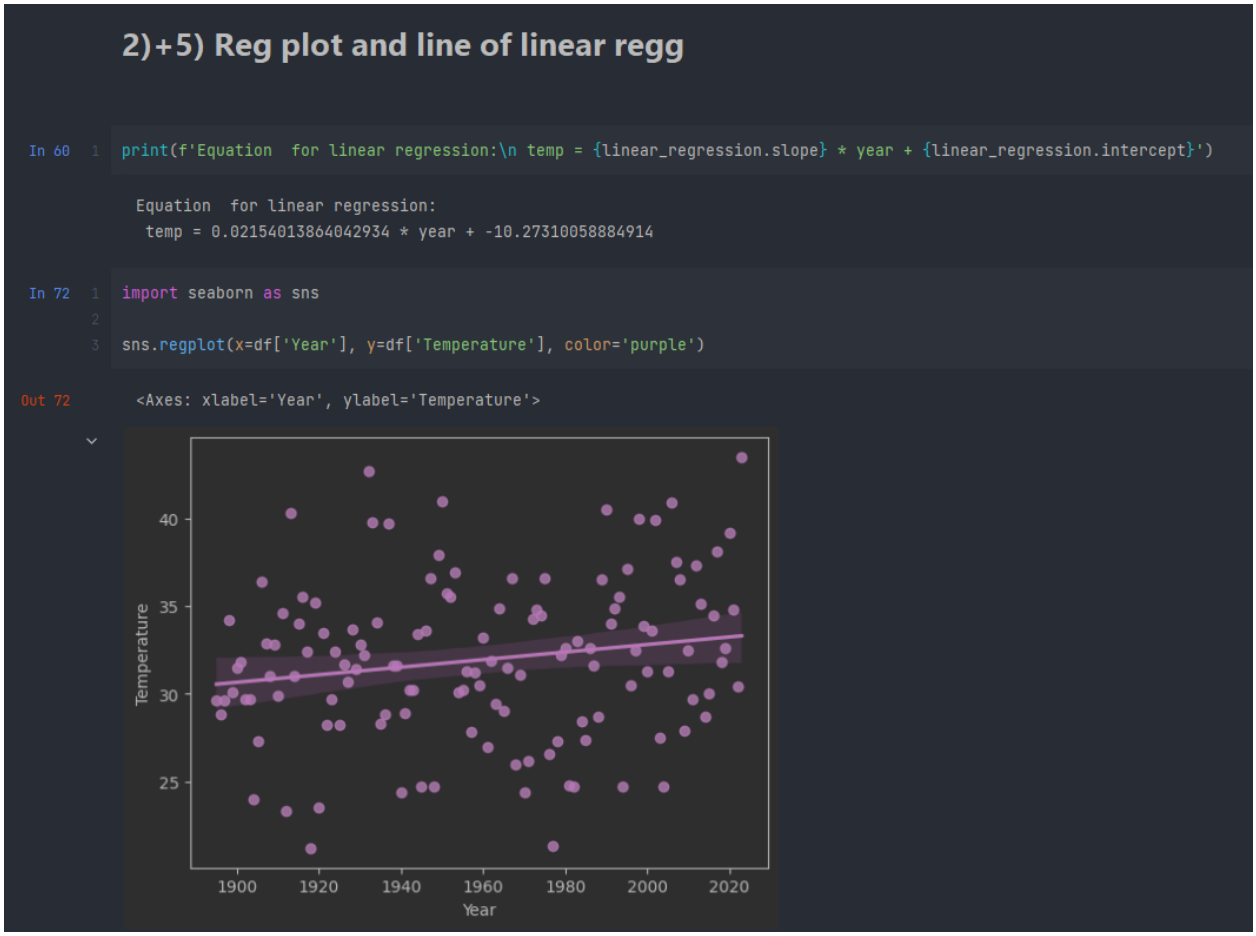
Спробуємо оцінити, якими показники температури могли б бути до 1895 року. Скористаємося функцією predict_temp. Знайдемо температури, наприклад, для 1885 – 1894 років:

```
4

In 45
for year in range(1885, 1895):
    temp = predict_temp(year)
    print(f"Predicted temperature for {year} year is {temp}")

Predicted temperature for 1885 year is 30.33006074836016
Predicted temperature for 1886 year is 30.351600887000593
Predicted temperature for 1887 year is 30.37314102564102
Predicted temperature for 1888 year is 30.39468116428145
Predicted temperature for 1889 year is 30.416221302921883
Predicted temperature for 1890 year is 30.43776144156231
Predicted temperature for 1891 year is 30.45930158020274
Predicted temperature for 1892 year is 30.480841718843166
Predicted temperature for 1893 year is 30.5023818574836
Predicted temperature for 1894 year is 30.523921996124024
```

2. **Бібліотеку Seaborn використати для графічного представлення даних DataFrame у вигляді регресійної прямої, що представляє графік зміни обраних показників за період 1895-2018 років.**
5. **Скористайтесь функцією regplot бібліотеки Seaborn для виведення всіх точок даних; дати представляються на осі x, а показники на осі y.**

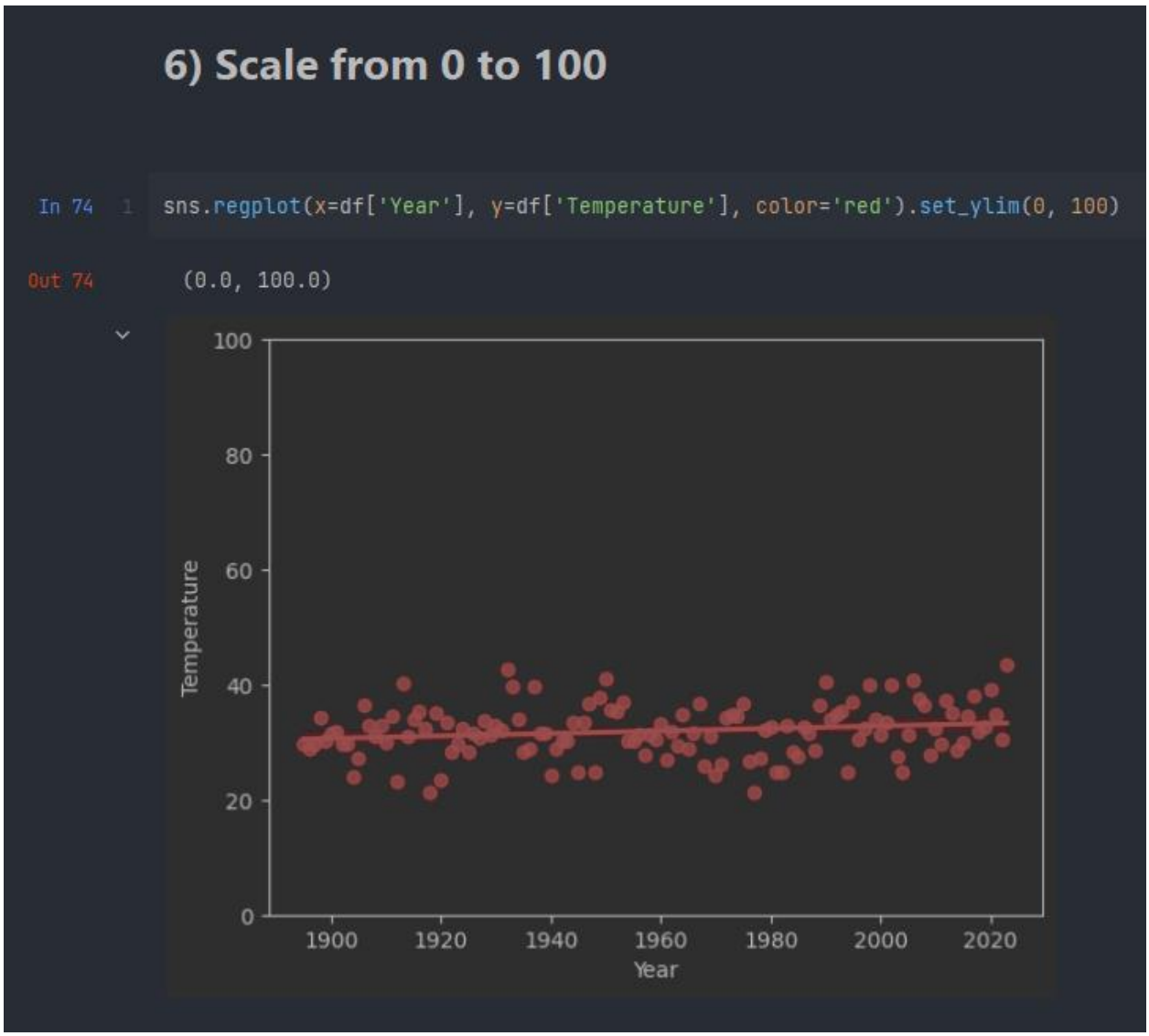
Виконаємо два пункта разом.

За допомогою функції regplot побудуємо діаграму розкиду даних, на якій точки представляють показники за заданий рік, а пряма лінія - регресійну пряму:



6. Виконайте маштабування осі *y*

Промасштабуємо вісь Temperature від 0 до 100:



7. Порівняйте 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.

Порівняємо наш прогноз з даними на сайті NOAA «Climate at a Glance»:

7) Conclusion for 2019-2022

In 87

```
for year in range(2019, 2023):
    temp = predict_temp(year)
    actual_temp = df.loc[df['Year'] == year]['Temperature'].values[0]
    diff = abs(round(temp - actual_temp, 2))

    print(f"Predicted temperature for {year} year is {temp}."
          f"\nThe actual temp is {actual_temp}"
          f"\nDifference by absolute: {diff}\n")
```

```
Predicted temperature for 2019 year is 33.21643932617769.
The actual temp is 32.6
Difference by absolute: 0.62

Predicted temperature for 2020 year is 33.237979464818125.
The actual temp is 39.2
Difference by absolute: 5.96

Predicted temperature for 2021 year is 33.25951960345855.
The actual temp is 34.8
Difference by absolute: 1.54

Predicted temperature for 2022 year is 33.28105974209898.
The actual temp is 30.4
Difference by absolute: 2.88
```

Побудована лінійна регресійна модель показала результати з середнім приблизним відхиленням у 1-2 градуси, але модель не врахувала аномально теплий січень 2020 року і отримане відхилення майже 6 градусів.

Висновок

У результаті проведеної лабораторної роботи, було ознайомлено з метеорологічними даними та застосовано бібліотеку Seaborn для їх візуалізації. Завантаживши дані з NOAA «Climate at a Glance» та використавши Pandas для їх обробки, було отримано перспективу змін температурних показників протягом тривалого періоду, починаючи з 1895 року.

Застосувавши метод лінійної регресії з використанням Seaborn, побудувано модель лінійної регресії та встановлено чітку тенденцію зростання температурних показників. Використання цієї моделі також дозволило зпрогнозувати значення показників на майбутні роки, зокрема на 2019-2022 роки.