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
«Часові ряди і прості лінійна регресія»

Виконав *ІП-15, Дацьо Іван*
(шифр, прізвище, ім'я, по батькові)

Перевірив *Нестерук А.*
прізвище, ім'я, по батькові

Київ 2023

Завдання:

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 масштабування 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.

Виконання

1. Імпортуємо усі необхідні пакети:

```
In [1]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        from scipy import stats
```

Завантажимо дані із csv файлу та переглянемо їх:

```
In [2]: dataset = pd.read_csv("1895-2022.csv")
        dataset.head()
```

Out[2]:

	Date	Value	Anomaly
0	189501	17.6	-2.5
1	189601	17.2	-2.9
2	189701	19.3	-0.8
3	189801	20.8	0.7
4	189901	19.2	-0.9

Перейменуємо назви стовпчиків для простішого використання:

```
In [3]: dataset.columns = ['Date', 'Temperature', 'Anomaly']
        dataset
```

Out[3]:

	Date	Temperature	Anomaly
0	189501	17.6	-2.5
1	189601	17.2	-2.9
2	189701	19.3	-0.8
3	189801	20.8	0.7
4	189901	19.2	-0.9
...	...	...	...
123	201801	19.1	-1.0
124	201901	19.4	-0.7
125	202001	27.7	7.6
126	202101	24.0	3.9
127	202201	15.8	-4.3

128 rows × 3 columns

Оскільки усі місяць для усіх записів є однаковим, то для кращого читання даних видалимо місяць в колонці Date. Але спочатку перевіримо тип даних в цій колонці:

```
In [4]: dataset.Date.dtype
Out[4]: dtype('int64')
```

Оскільки дані в цій колонці є цілими числами то ми можемо розділити їх на 100 та прибрати останні два числа:

In [5]:
```
dataset.Date = dataset.Date.floordiv(100)
dataset
```

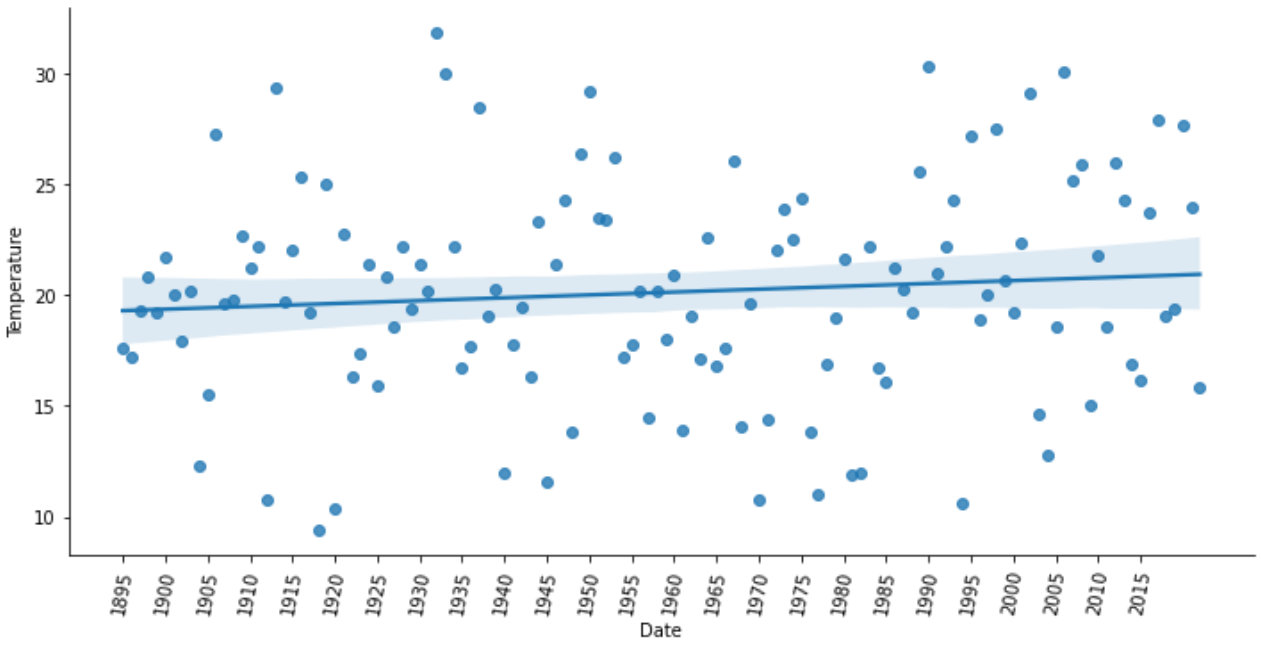
Out[5]:

	Date	Temperature	Anomaly
0	1895	17.6	-2.5
1	1896	17.2	-2.9
2	1897	19.3	-0.8
3	1898	20.8	0.7
4	1899	19.2	-0.9
...	...	...	...
123	2018	19.1	-1.0
124	2019	19.4	-0.7
125	2020	27.7	7.6
126	2021	24.0	3.9
127	2022	15.8	-4.3

128 rows × 3 columns

2. Бібліотеку Seaborn використати для графічного представлення даних DataFrame у вигляді регресійної прямої, за період 1895-2018 років.

In [6]:
```
sns.lmplot(x="Date", y="Temperature", data=dataset, aspect=2)
plt.xticks(range(1895, 2018, 5), rotation=80)
plt.show()
```



3. Спрогнозуємо дані температури на 2019-2022 рік.

Спрогнозуємо дані температури на 2019, 2020, 2021 та 2022 рік.

In [7]:
```
linear_regression = stats.linregress(x = dataset.Date, y = dataset.Temperature)
predictions = [(linear_regression.slope * x + linear_regression.intercept) for x in range(2019, 2023)]
for i, year in zip(predictions, range(2019,2023)):
    print(f'Прогнозована температура в {year} році: {i}')
```

```
Прогнозована температура в 2019 році: 20.90263172953671
Прогнозована температура в 2020 році: 20.91556262589269
Прогнозована температура в 2021 році: 20.92849352224867
Прогнозована температура в 2022 році: 20.941424418604647
```

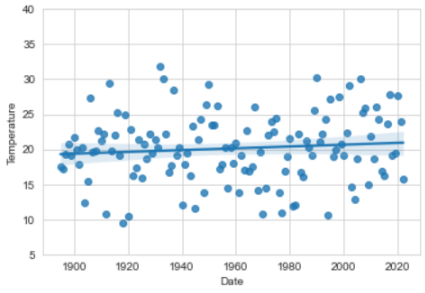
4. Оцінимо за формулою, яка середня температура могла бути отримана в 1886 році :

```
In [8]: prediction = linear_regression.slope * 1886 + linear_regression.intercept
        print(f'Можливе значення температури в 1886 році: {prediction}')

        Можливе значення температури в 1886 році: 19.182822514191535
```

5 та 6. Скористаємось функцією regplot бібліотеки Seaborn для виведення всіх точок даних та змінимо масштабування для осі Y від 5 до 40 для кращого відображення:

```
In [9]: plt.clf()
        sns.set_style('whitegrid')
        axes = sns.regplot(x='Date', y='Temperature', data=dataset)
        axes.set_ylim(5, 40)
        plt.show()
```



7. Порівняємо отримані результати із реальними даними:

Рік	NOAA «Climate at a Glance»	Отримані результати	Різниця
2019	19.4	20.90	+1.5

2020	27.7	20.91	-6.79
2021	24	20.92	-3.08
2022	15.8	20.94	+5.14

Можна помітити що прогнозовані дані відрізняються суттєво. Це зумовлено тим що даний підхід не враховує додаткові чинники, які могли вплинути на температуру. Даний спосіб є корисним, коли потрібно знайти приблизне значення величини.

Висновок

Виконавши дану лабораторну я ознайомився з бібліотекою seaborn, яка була використана для побудови лінійної регресії. Було спрогнозовано середні температури у січні, використовуючи для цього дані із попередніх років. Був зроблений висновок, що даний метод передбачає лише приблизні значення які в реальності значно відрізняються.