НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформатики та програмної інженерії

Лабораторна робота № 1

з дисципліни «Прикладні задачі машинного навчання»

Тема: «Часові ряди і прості лінійна регресія»

Прийняв:

Виконав:

студент групи ІП-13

Недельчев Є.О.

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. Завантажимо дані та продивимося їх структуру:

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

Out[2]:

	Date	Value	Anomaly
0	189507	61.50	-1.98
1	189607	63.69	0.21
2	189707	62.23	-1.25
3	189807	63.39	-0.09
4	189907	61.16	-2.32

Перейменуємо назви стовпчиків на більш інтуїтивно зрозумілі:

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

Out[3]:

	Date	Temperature	Anomaly
0	189507	61.50	-1.98
1	189607	63.69	0.21
2	189707	62.23	-1.25
3	189807	63.39	-0.09
4	189907	61.16	-2.32
...	...	...	...
119	201407	63.51	0.03
120	201507	64.77	1.29
121	201607	66.15	2.67
122	201707	66.24	2.76
123	201807	65.21	1.73

124 rows × 3 columns

Оскільки будемо обробляти тільки січневі дані, мітки осі x будуть краще читатися без позначення 01 (для січня); видалимо місяць з Date:

```
Ввод [4]: dataset.Date = dataset.Date.floordiv(100)
          dataset
```

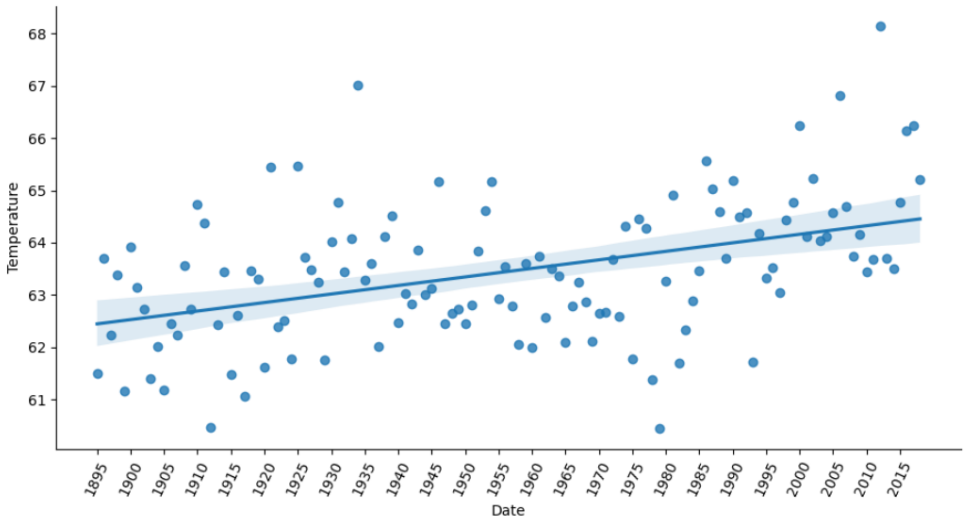
Out[4]:

	Date	Temperature	Anomaly
0	1895	61.50	-1.98
1	1896	63.69	0.21
2	1897	62.23	-1.25
3	1898	63.39	-0.09
4	1899	61.16	-2.32
...	...	...	...
119	2014	63.51	0.03
120	2015	64.77	1.29
121	2016	66.15	2.67
122	2017	66.24	2.76
123	2018	65.21	1.73

124 rows × 3 columns

2. Використаємо бібліотеку seaborn для графічного представлення даних DataFrame у вигляді регресійної прямої, що представляє графік зміни обраних показників за період 1895-2018 років.

Ввод [5]:
```
sns.lmplot(x="Date", y="Temperature", data=dataset, aspect=1.9)
plt.xticks(range(1895, 2020, 5), rotation=65)
plt.show()
```



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

Ввод [6]:
```
linear_regression = stats.linregress(x = dataset.Date, y = dataset.Temperature)
predicts = [(linear_regression.slope * x + linear_regression.intercept) for x in range(2019, 2023)]
for x, i in zip(predicts, range(2019,2023)):
    print(f'Очікувана температура в {i} році: {x}')
```

```
Очікувана температура в 2019 році: 64.4694256490952
Очікувана температура в 2020 році: 64.48575710464202
Очікувана температура в 2021 році: 64.50208856018882
Очікувана температура в 2022 році: 64.51842001573564
```

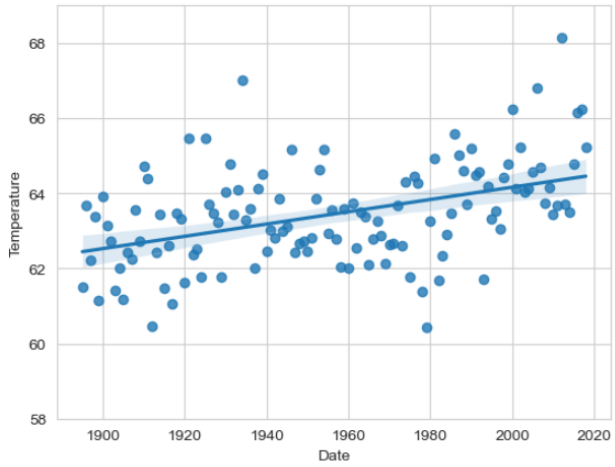
4. Оцінимо, які могли б бути дані у випадковий рік на проміжку з 1800 по 1890.

Ввод [7]:
```
random_year = random.randint(1800, 1890)
predict = linear_regression.slope * random_year + linear_regression.intercept
print(f'Можлива температура в {random_year} році: {predict}')
```

```
Можлива температура в 1822 році: 61.25212890637293
```

5 & 6. Скористаємось функцією regplot бібліотеки Seaborn для виведення всіх точок даних. Також виконаємо масштабування осі y від 58 до 69 для кращої візуалізації:

```
Ввод [8]: plt.clf()
sns.set_style('whitegrid')
axes = sns.regplot(x='Date', y='Temperature', data=dataset)
axes.set_ylim(58, 69)
plt.show()
```



7. Порівняємо отримані в результаті лабораторної роботи прогнози із реальними даними.

Рік	NOAA «Climate at a Glance»	Дані, отримані в ході роботи	Різниця
2019	62.78	64.46	+1.68
2020	65.34	64.49	-0.85
2021	64.82	64.50	-0.32
2022	65.19	64.52	-0.67

Досить легко помітити, що похибка є досить суттєвою, оскільки таких підхід прогнозування даних ігнорує безліч чинників. Такий спосіб прогнозування дозволяє лише приблизно оцінити необхідні дані.

Висновок

Виконуючи цю лабораторну роботу я ознайомився з бібліотекою seaborn та використав на практиці графік лінійної регресії. Також були спрогнозовані дані на підставі старих даних та було зроблено висновок, що прогнозування подій із використанням лінійної регресії є досить неточним і дозволяє лише приблизно оцінити необхідні нам дані.