аНАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО”
ФАКУЛЬТЕТ ІНФОРМАТИКИ ТА ОБЧИСЛЮВАЛЬНОЇ ТЕХНІКИ

Кафедра інформатики та програмної інженерії

Звіт до лабораторної роботи №3

з курсу

«Машине навчання»

студента 2 курсу
групи ІТ-02
Макарова Іллі Сергійовича

Викладач:
Оніщенко В.

Київ – 2022

Тема: Часові ряди і проста лінійна регресія

Завдання:

1. В цій лабораторній роботі Вам треба завантажити дані середніх січневих температур в Нью-Йорку в 1895-2018 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 середньої температури за період 1895-2018 років.

3. Спрогнозуйте середню температуру за Фаренгейтом за січень 2019 року, січень 2020 року та січень 2021 року та січень 2022 року.

4. Оцініть за формулою, якою могла бути середня температура до 1895 року.

5. Скористайтесь функцією regplot бібліотеки Seaborn для виведення всіх точок даних; дати представляються на осі x, а температури на осі y.

6. Виконайте маштабування вісі y від 20-градусного діапазону до 60- градусного діапазону

7. Порівняйте отриманий прогноз для січня 2019, січня 2020 та січня 2021 та за січень 2022 з даними на NOAA «Climate at a Glance»: https://www.ncdc.noaa.gov/cag/ і зробити висновок

Виконання:

Ну на початку як завжди імпортуємо бібліотеки та відкриваємо файл, аби подивитись, що там взагалі за дані

```
In [34]:
import pandas as pd
import seaborn as sns
from scipy import stats

pd.set_option('precision', 2)
```

```
In [2]:
df = pd.read_csv('data/us-new-york-avg-temp-1895-2018.csv')
```

```
In [3]:
df.head()
```

	Date	Value
0	189501	29.4
1	189601	29.0
2	189701	29.8
3	189801	34.4
4	189901	30.3

Там була невеличка проблема, в даних на сайті чомусь нема колонки Anomaly, дані наче ті, що треба, можливо її просто видалили, з самого data source. Але це не проблема, ми просто додаємо дану колонку самостійно:

```
In [4]:
df['Anomaly'] = df['Value'] - df['Value'].mean()
```

```
In [5]:
df = df.rename(columns={'Value': 'Temperature'})
```

```
In [6]:
df.head()
```

	Date	Temperature	Anomaly
0	189501	29.4	-2.73
1	189601	29.0	-3.13
2	189701	29.8	-2.33
3	189801	34.4	2.27
4	189901	30.3	-1.83

Так, тепер трошки змінемо формат дати, аби лишились лише роки:

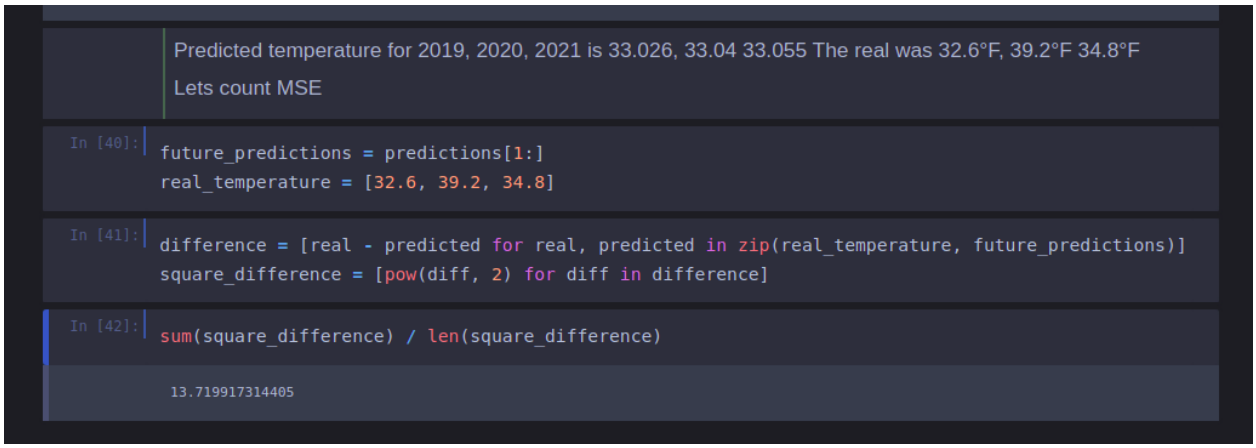
```
In [8]:
df['Date'] = df['Date'] // 100
```

```
In [9]:
df.head()
```

	Date	Temperature	Anomaly
0	1895	29.4	-2.73
1	1896	29.0	-3.13
2	1897	29.8	-2.33
3	1898	34.4	2.27
4	1899	30.3	-1.83

Настав час регресії, вирахуємо коефіцієнти регресійної моделі, та спробуємо передбачити температуру у 1890, 2019, 2020, 2021 роках

Тепер, давайте подивимось, на скільки наші predictions точні. І порівняємо з реальними даними наші розрахунки.

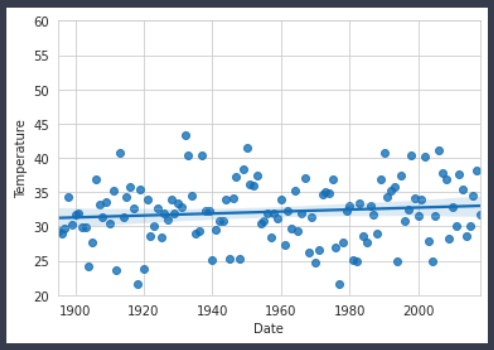


Ну і нарешті, давайте візуалізуємо наші дані, за допомогою seaborn. Тут відразу буде видно, як самі дані, так і регресійна пряма.

Також поставимо ліміт по осі y, аби графік виглядав красивіше

```
In [13]: sns.set_style('whitegrid')

In [14]: sns.regplot(x=df['Date'], y=df['Temperature']).set_ylim(20, 60)
         None
```



Ось і все, не знаю чи надо в роботі писати якийсь висновок. Як видно з порівнянь реальних даних, та наших передбачень, помилки є, та вони досить суттєві. Однак як на мене, лінійна регресія це все ж більше, про те, щоб побачити ту чи іншу тенденцію, та міру цієї тенденції, чекати від регресії точних передбачень марно.