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 №4 з дисципліни
«Аналіз даних в інформаційних системах»

«Вивідна статистика»

Варіант 7

Виконав студент ІП-12 Васильєв Єгор Костянтинович
(шифр, прізвище, ім'я, по батькові)

Перевірив Ліхоузова Тетяна Анатоліївна
(прізвище, ім'я, по батькові)

Київ 2023

Лабораторна робота № 4

Тема: Вивідна статистика

Для виконання лабораторної роботи було обрано мову програмування Python, та бібліотеки Numpy, Pandas і Matplotlib. Спочатку дані було завантажено та імпортовано до датафрейму.

	Country Name	Region	GDP per capita	Populatiion	CO2 emission	Area
0	Afghanistan	South Asia	561,7787463	34656032.00000	9809,225	652860
1	Albania	Europe & Central Asia	4124,98239	2876101.00000	5716,853	28750
2	Algeria	Middle East & North Africa	3916,881571	40606052.00000	145400,217	2381740
3	American Samoa	East Asia & Pacific	11834,74523	55599.00000	nan	200
4	Andorra	Europe & Central Asia	36988,62203	77281.00000	462,042	470
5	Angola	Sub-Saharan Africa	3308,700233	28813463.00000	34763,16	1246700
6	Antigua and Barbuda	Latin America & Caribbean	14462,17628	100963.00000	531,715	440
7	Argentina	Latin America & Caribbean	12440,32098	43847430.00000	204024,546	2780400
8	Armenia	Europe & Central Asia	3614,688357	2924816.00000	5529,836	29740
9	Aruba	Latin America & Caribbean	nan	104822.00000	872,746	180
10	Australia	East Asia & Pacific	49755,31548	24127159.00000	361261,839	7741220
11	Austria	Europe & Central Asia	44757,6349	8747358.00000	58712,337	83879
12	Azerbaijan	Europe & Central Asia	3878,709257	9762274.00000	37487,741	86600
13	Bahamas, The	Latin America & Caribbean	28785,47767	391232.00000	2416,553	13880
14	Bahrain	Middle East & North Africa	22579,09342	1425171.00000	31338,182	771
15	Bangladesh	South Asia	1358,779029	162951560.00000	73189,653	147630
16	Barbados	Latin America & Caribbean	15891,62655	284996.00000	1272,449	430
17	Belarus	Europe & Central Asia	4989,427763	9507120.00000	63497,772	207600
18	Belgium	Europe & Central Asia	41271,48215	11348159.00000	93350,819	30530
19	Belize	Latin America & Caribbean	4744,736397	366954.00000	495,045	22970
20	Benin	Sub-Saharan Africa	789,4404107	10872298.00000	6318,241	114760

Рисунок 3.1 – імпортовані дані у data frame

Було досліджено структуру даних та знайдено наступні помилки: граматична помилка у назві стовпця Population, числові дані, які частково зберігаються у вигляді рядкових, наявність від’ємних значень та пропущені значення, що не дозволяють коректно побудувати діаграму розмаху. Після виправлення цих помилок датафрейм отримав наступний вигляд:

	Country Name	Region	GDP per capita	Population	CO2 emission	Area
0	Afghanistan	South Asia	561.77875	34656032.00000	9809.22500	652860.00000
1	Albania	Europe & Central Asia	4124.98239	2876101.00000	5716.85300	28750.00000
2	Algeria	Middle East & North Africa	3916.88157	40606052.00000	145400.21700	2381740.00000
3	American Samoa	East Asia & Pacific	11834.74523	55599.00000	165114.11634	200.00000
4	Andorra	Europe & Central Asia	36988.62203	77281.00000	462.04200	470.00000
5	Angola	Sub-Saharan Africa	3308.70023	28813463.00000	34763.16000	1246700.00000
6	Antigua and Barbuda	Latin America & Caribbean	14462.17628	100963.00000	531.71500	440.00000
7	Argentina	Latin America & Caribbean	12440.32098	43847430.00000	204024.54600	2780400.00000
8	Armenia	Europe & Central Asia	3614.68836	2924816.00000	5529.83600	29740.00000
9	Aruba	Latin America & Caribbean	13445.59342	104822.00000	872.74600	180.00000
10	Australia	East Asia & Pacific	49755.31548	24127159.00000	361261.83900	7741220.00000
11	Austria	Europe & Central Asia	44757.63490	8747358.00000	58712.33700	83879.00000
12	Azerbaijan	Europe & Central Asia	3878.70926	9762274.00000	37487.74100	86600.00000
13	Bahamas, The	Latin America & Caribbean	28785.47767	391232.00000	2416.55300	13880.00000
14	Bahrain	Middle East & North Africa	22579.09342	1425171.00000	31338.18200	771.00000
15	Bangladesh	South Asia	1358.77903	162951560.00000	73189.65300	147630.00000
16	Barbados	Latin America & Caribbean	15891.62655	284996.00000	1272.44900	430.00000
17	Belarus	Europe & Central Asia	4989.42776	9507120.00000	63497.77200	207600.00000
18	Belgium	Europe & Central Asia	41271.48215	11348159.00000	93350.81900	30530.00000
19	Belize	Latin America & Caribbean	4744.73640	366954.00000	495.04500	22970.00000
20	Benin	Sub-Saharan Africa	789.44041	10872298.00000	6318.24100	114760.00000

Рисунок 3.2 – виправлені дані у data frame

За допомогою вбудованого методу бібліотеки scipy було досліджено чи є параметри у датафреймі, які розподілені за нормальним законом, використовуючи критерій Пірсона.

```
Statistics = 370.214, p = 0.000
Data does not follow a normal distribution
Statistics = 110.278, p = 0.000
Data does not follow a normal distribution
Statistics = 406.218, p = 0.000
Data does not follow a normal distribution
Statistics = 284.697, p = 0.000
Data does not follow a normal distribution
```

Рисунок 3.3 – перевірка параметрів на нормальність розподілу

За допомогою реалізованої функції було обчислено відсоткову різницю між середнім арифметичним та медіаною для кожного параметру датафрейма.

```
Difference between median and mean: 81.66%, mean = 34322559, median = 6293253
Difference between median and mean: 46.60%, mean = 13445, median = 7179
Difference between median and mean: 93.00%, mean = 165114, median = 11562
Difference between median and mean: 84.97%, mean = 618844, median = 93030
```

Рисунок 3.4 – різниця між середнім та медіаною для кожного параметра

Критерієм Шапіро-Уілка було перевірено нормальність розподілу викидів CO2 для кожного регіону

```
South Asia region does not have a normal distribution of CO2 emissions
Europe & Central Asia region does not have a normal distribution of CO2 emissions
Middle East & North Africa region does not have a normal distribution of CO2 emissions
East Asia & Pacific region does not have a normal distribution of CO2 emissions
Sub-Saharan Africa region does not have a normal distribution of CO2 emissions
Latin America & Caribbean region does not have a normal distribution of CO2 emissions
North America region has a normal distribution of CO2 emissions
```

Рисунок 3.5 – перевірка параметру кількості викидів CO2 на нормальність розподілу

За допомогою методу pie бібліотеки matplotlib було побудовано кругову діаграму розподілу населення по регіонам.

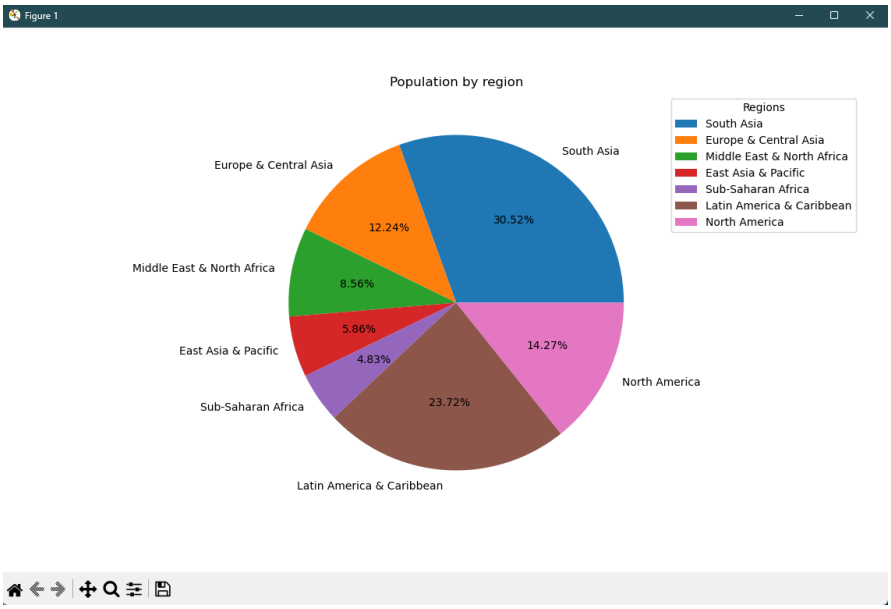

Рисунок 3.6 – діаграма розподілу населення по регіонам

Використовуючи методи для роботи з зображеннями бібліотеки matplotlib було розміщено бульбашки на мапі України, що відповідають кількості населення в обраних п’яти містах.

Рисунок 3.7 – візуалізація кількості населення по містам

З використанням методу distance бібліотеки scipy було обчислено найбільшу відстань між двома містами у пікселях та кілометрах, взявши протяжність України зі Заходу на Схід рівною 1316км.

```
The longest distance is between Lviv and Luhansk.
Distance in pixels: 703.53
Distance in kilometers: 1114.13
```

Рисунок 3.8 – найбільша відстань між містами

За допомогою бібліотек geopandas та geoviews було візуалізовано значення ВВП і прибутку на одну особу по регіонам за даними 2016 року, отриманими шляхом об’єднання shape-файлу з відповідними даними.

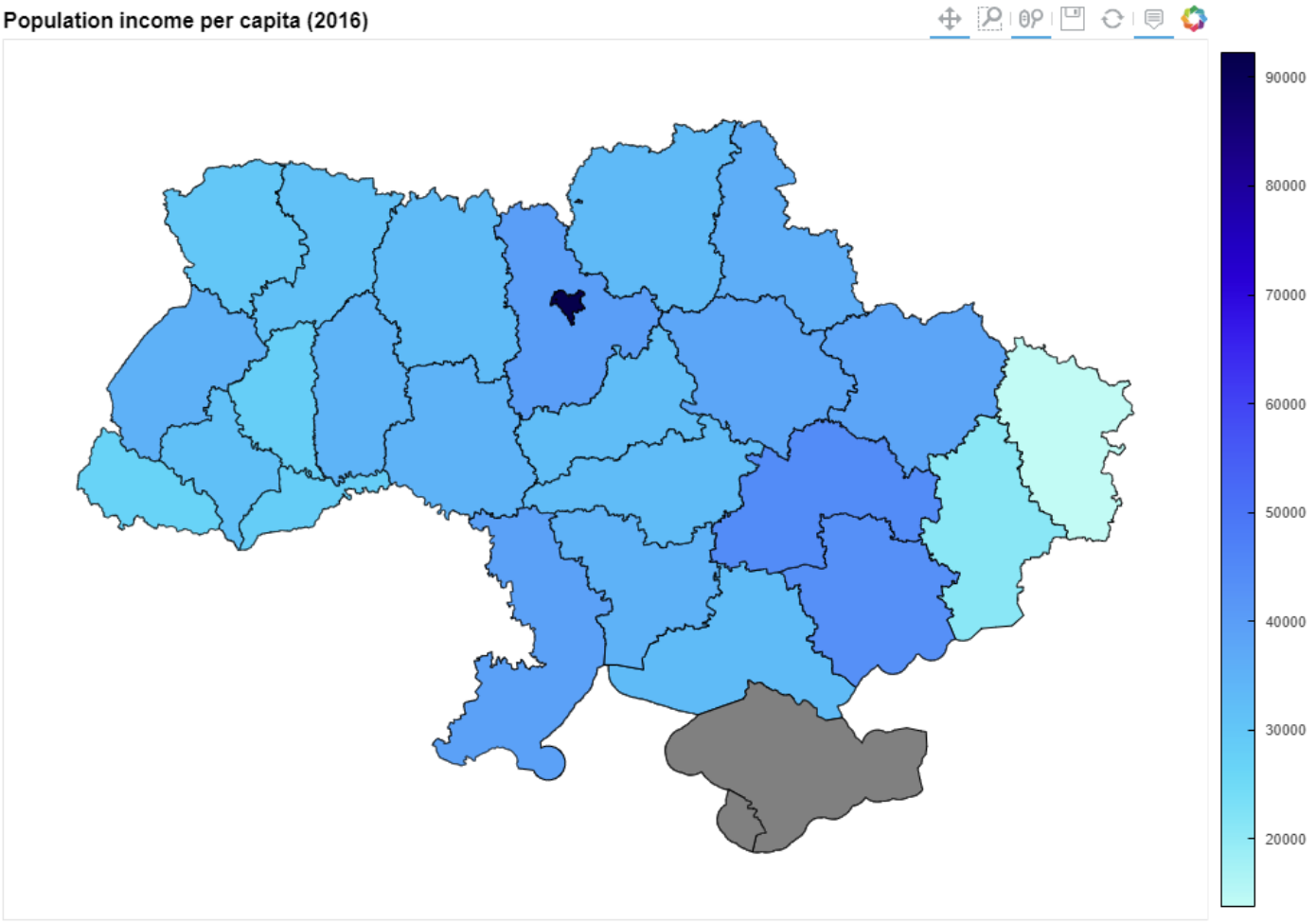

Рисунок 3.9 – рівень прибутку на душу населення по регіонам

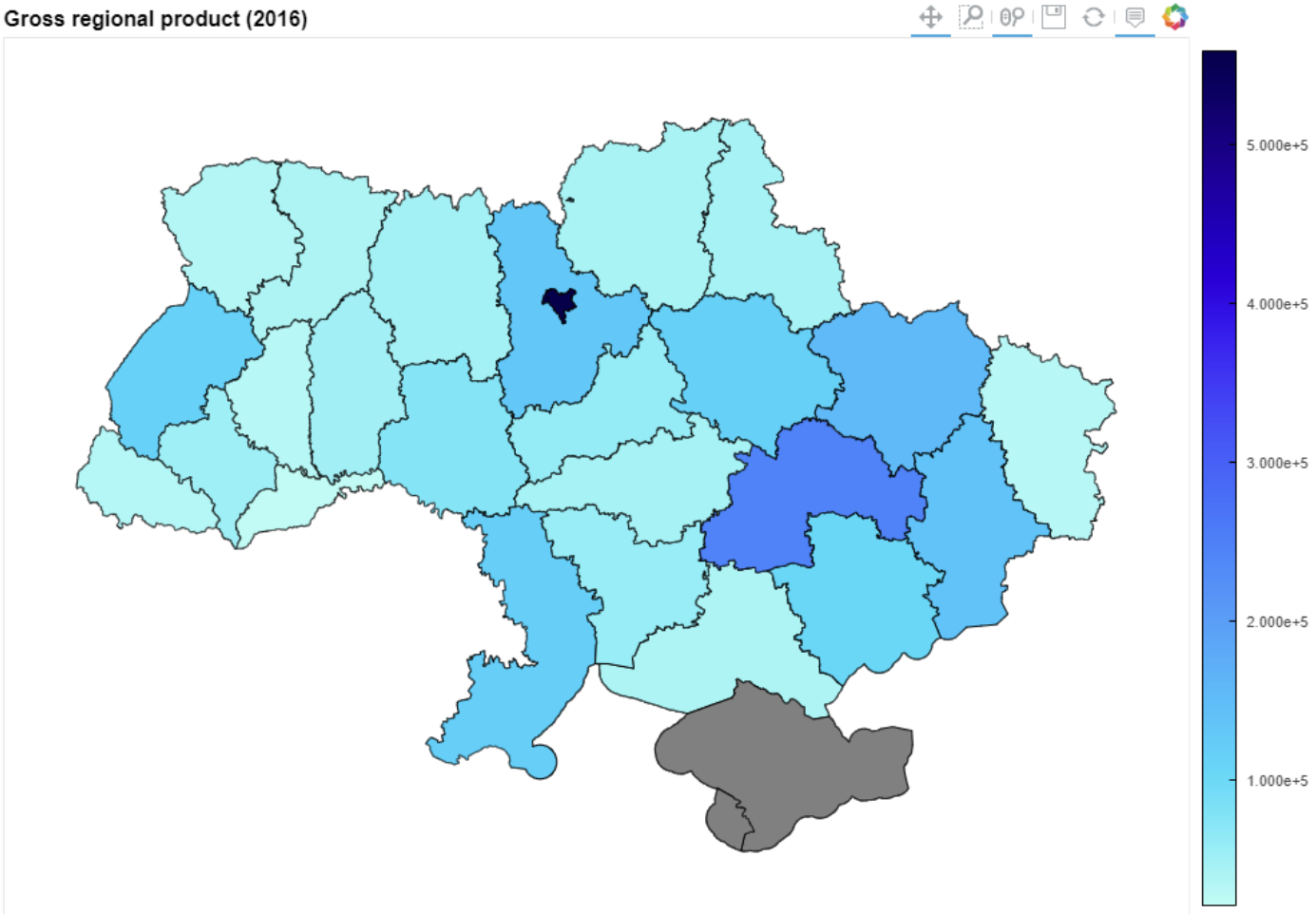

Рисунок 3.10 – рівень ВВП по регіонам

Для кожного регіону також було обраховано та візуалізовано коефіцієнт кореляції цих двох параметрів методом Пірсона.

	Name	UKRname	GDP-DPP Correlation
0	Autonomous Republic of Crimea	Автономна Республіка Крим	1.00000
1	Vinnytsia Oblast	Вінницька	0.98879
2	Volyn Oblast	Волинська	0.97171
3	Dnipropetrovsk Oblast	Дніпропетровська	0.99347
4	Donetsk Oblast	Донецька	0.76162
5	Zhytomyr Oblast	Житомирська	0.99003
6	Zakarpattia Oblast	Закарпатська	0.96040
7	Zaporizhia Oblast	Запорізька	0.98611
8	Ivano-Frankivsk Oblast	Івано-Франківська	0.95876
9	Kyiv Oblast	Київська	0.99541
10	Kirovohrad Oblast	Кіровоградська	0.98422
11	Luhansk Oblast	Луганська	0.88811
12	Lviv Oblast	Львівська	0.99144
13	Mykolaiv Oblast	Миколаївська	0.98253
14	Odessa Oblast	Одеська	0.98818
15	Poltava Oblast	Полтавська	0.99084
16	Rivne Oblast	Рівненська	0.95466
17	Sumy Oblast	Сумська	0.97795
18	Ternopil Oblast	Тернопільська	0.95332
19	Kharkiv Oblast	Харківська	0.99224
20	Kherson Oblast	Херсонська	0.98420

Рисунок 3.11 – dataframe з коефіцієнтом кореляції ВВП і прибутку на одиницю населення для кожного регіону

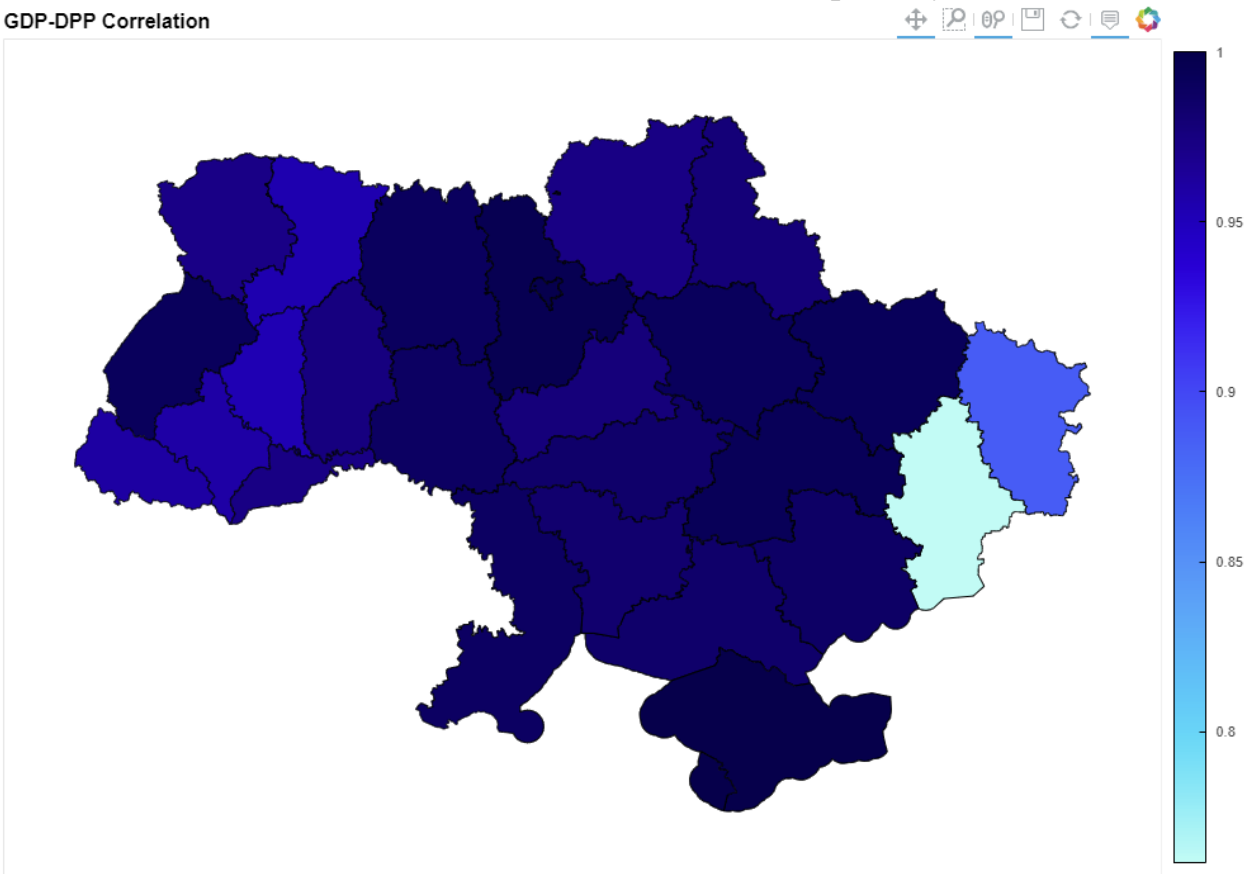

Рисунок 3.12 – візуалізація вище зазначеного dataframe

Висновок: я дослідив різні критерії нормальності на прикладі економіко-географічних даних різних країн світу та переконався у тому, що економічні показники майже ніколи не розподілені нормально, оскільки залежать від багатьох різноманітних факторів, однак ті з них, які схожі за суттю, наприклад ВВП та прибуток населення, можуть корелювати між собою.