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 комп'ютерного практикуму № 4 з дисципліни

«Аналіз даних в інформаційних системах»

на тему: «Вивідна статистика»

Виконав студент ІП-13, Дойчев Костянтин Миколайович
(шифр, прізвище, ім'я, по батькові)

Перевірив Олійник Юрій Олександрович
(прізвище, ім'я, по батькові)

Київ 2023

Комп’ютерний практикум 4

Тема – Вивідна статистика.

Мета – ознайомитись з

- методами визначення точкових оцінок параметрів розподілу; дослідити, що впливає на якість точкових оцінок;
- методикою визначення інтервальних оцінок параметрів розподілу; дослідити, що впливає на якість інтервальних оцінок;
- методами перевірки статистичних гіпотез про вигляд закону розподілу; дослідити, що впливає на ширину критичної області.

Завдання

Основне:

1. Скачати дані із файлу Data2.csv
2. Подивитись, проаналізувати структуру
3. Вказати, чи є параметри, що розподілені за нормальним законом
4. Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів
5. Вказати, в якому регіоні розподіл викидів CO2 найбільш близький до нормального
6. Побудувати кругову діаграму населення по регіонам

Основне завдання
DataFrame та його структура

За допомогою Python бібліотеки Pandas прочитаємо та завантажимо дані з даного Data2.csv файлу в dataframe. Проінспектуємо структуру наших даних.

```
In 32
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
Executed at 2023.06.25 21:31:14 in 18ms

In 33
# read data
df = pd.read_csv('data/Data2.csv', sep=';', decimal=',', encoding='cp1252')
Executed at 2023.06.25 21:31:14 in 13ms

In 34
# Print rows
head_rows = 5
tail_rows = 6

h = df.head(5)
t = df.tail(6)

print('Head')
print(h)
print('Tail')
print(t)
Executed at 2023.06.25 21:31:14 in 28ms
```

```
Head
      Country Name                     Region  GDP per capita  Populatiion  \
0      Afghanistan                 South Asia      561.778746   34656032.0
1          Albania       Europe & Central Asia     4124.982390    2876101.0
2          Algeria  Middle East & North Africa     3916.881571   40606052.0
3   American Samoa         East Asia & Pacific    11834.745230      55599.0
4          Andorra       Europe & Central Asia    36988.622030      77281.0

    CO2 emission       Area
0       9809.225   652860.0
1       5716.853    28750.0
2     145400.217  2381740.0
3            NaN      200.0
4        462.042      470.0
Tail
              Country Name                     Region  GDP per capita  \
211               Vietnam        East Asia & Pacific     2170.648054
212  Virgin Islands (U.S.)   Latin America & Caribbean            NaN
213     West Bank and Gaza  Middle East & North Africa     2943.404534
214           Yemen, Rep.  Middle East & North Africa      990.334774
215                Zambia         Sub-Saharan Africa     1269.573537
216              Zimbabwe         Sub-Saharan Africa     1029.076649

     Populatiion  CO2 emission      Area
211   92701100.0    166910.839  330967.0
212     102951.0           NaN     350.0
213    4551566.0           NaN    6020.0
214   27584213.0     22698.730  527970.0
215   16591390.0      4503.076  752610.0
216   16150362.0     12020.426  390760.0
```

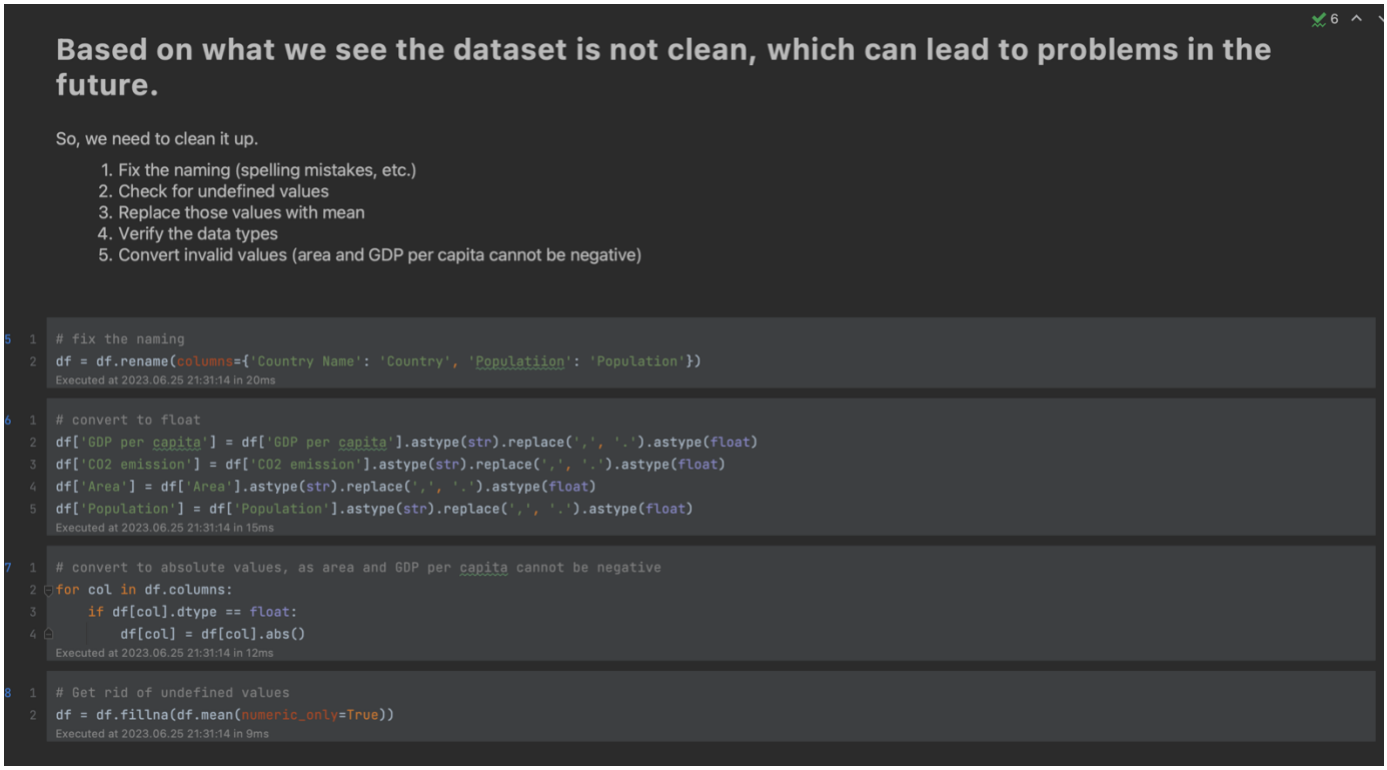
На даному рисунку можна помітити загальну інформацію про датафрейм: кількість рядків та колонок, назви всіх колонок, кількість записів в кожній з них, тип даних колонки та використання пам’яті.

Висновки:

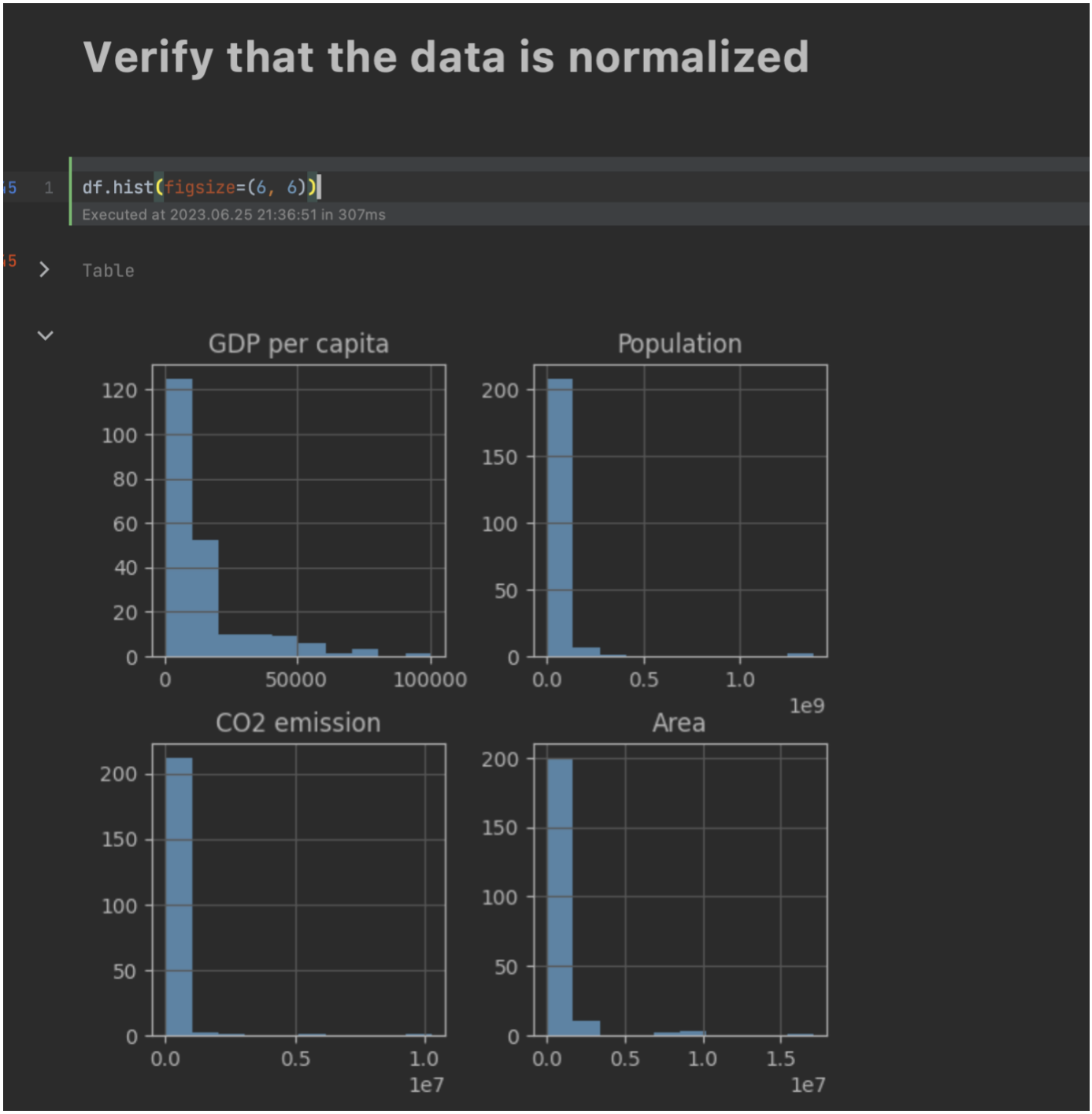
1. Помилки в правописі (spelling) колонки Population. Populatiion -> Population
2. Пусті дані (NaN)
3. Від’ємні значення де вони не мають бути. Площа не може бути від’ємною

Виправлення помилок

Змінимо назву колонки та перевіримо наявність пустих значень



Верифікація нормальності



Аналізуючи гістограми є велика підозра, що дані не нормалізовані. Тому перевіримо їх за допомогою тесту Шапіро-Уілкса (Shapiro-Wilk test)

```
In 40
# based on the histogram, we can see that the data is not normalized. However, it would be better to verify it.

alpha = 0.05

# Shapiro-Wilk Test
for col in df.columns:
    if df[col].dtype == float:
        stats_per_col, p_per_col = stats.shapiro(df[col])
        print('Column name:', col)
        print(f'Stats: {round(stats_per_col, 5)}, p: {round(p_per_col, 5)}')

        if p_per_col > alpha:
            print('normal distribution')
        else:
            print('not normal distribution')
        print()
Executed at 2023.06.25 21:31:14 in 2ms
```

```
Column name: GDP per capita
Stats: 0.73067, p: 0.0
not normal distribution

Column name: Population
Stats: 0.2171, p: 0.0
not normal distribution

Column name: CO2 emission
Stats: 0.17369, p: 0.0
not normal distribution

Column name: Area
Stats: 0.33839, p: 0.0
not normal distribution
```

Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів

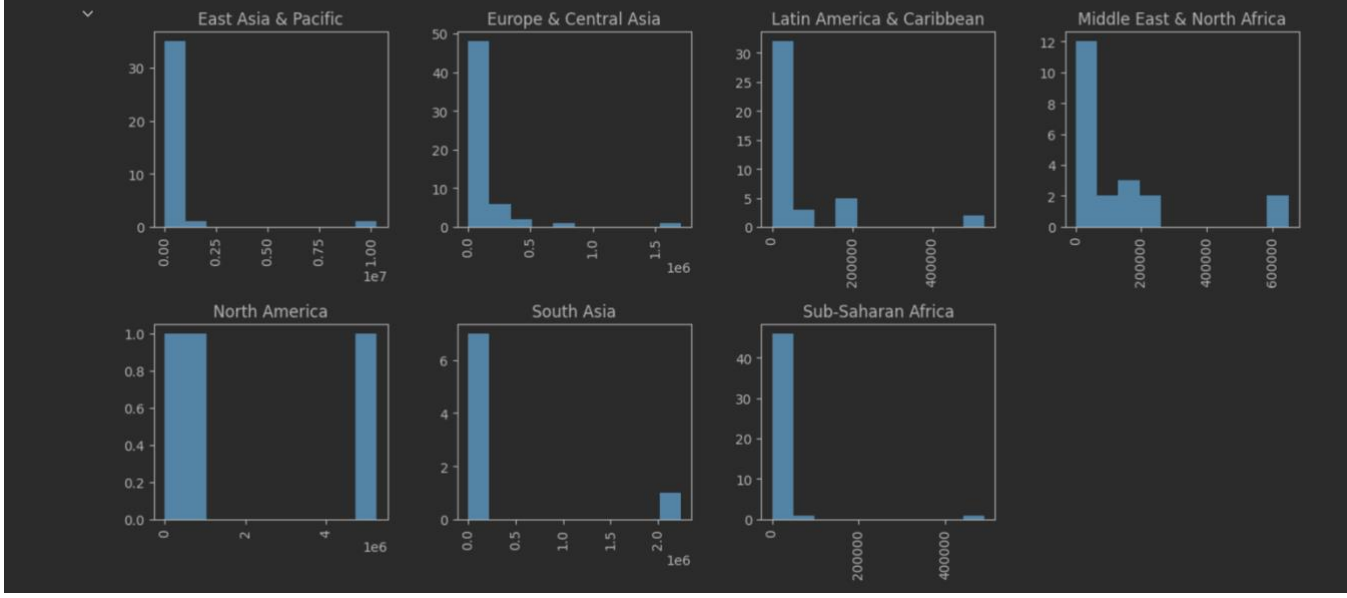
```
In 65
# let's check the hypothesis that the mean and median are the same

alpha = 0.05
stats, p = stats.ttest_1samp(df['Area'], df['Area'].median())

print('Area')
print(f'Stats: {round(stats, 5)}, p: {round(p, 5)}')
if p > alpha:
    print('equal')
else:
    print('different')
Executed at 2023.06.25 21:46:18 in 2ms
```

```
Area
Stats: 4.23766, p: 3e-05
different
```

Вказати, в якому регіоні розподіл викидів CO2 найбільш близький до нормального

```
In _ 1  # region with co2 emission that is the closest to a normal distribution
    2  df['CO2 emission'].hist(by=df['Region'], layout=(4, 4), figsize=(15, 15))
    3  plt.show()
```

Executed at 2023.06.25 21:50:02 in 483ms




```
In _ 1  for region in df['Region'].unique():
    2      is_similar_region = df['Region'] == region
    3      stats_per_col, p_per_col = stats.shapiro(df[is_similar_region]['CO2 emission'])
    4
    5      print('Region: ', region)
    6      if p_per_col > alpha:
    7          print('normally distributed')
    8      else:
    9          print('not normally distributed')
   10      print()
```

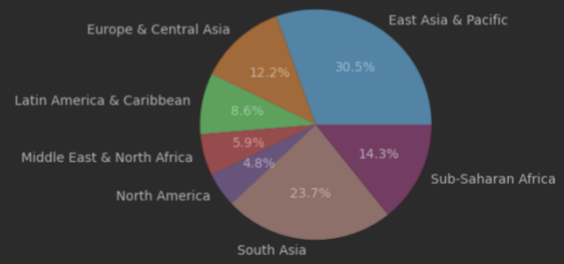
Executed at 2023.06.25 21:50:02 in 1ms

```
Region:  South Asia
not normally distributed

Region:  Europe & Central Asia
not normally distributed

Region:  Middle East & North Africa
not normally distributed

Region:  East Asia & Pacific
not normally distributed

Region:  Sub-Saharan Africa
not normally distributed

Region:  Latin America & Caribbean
not normally distributed

Region:  North America
normally distributed
```

Бачимо, що тільки регіон північної Америки нормально розподілений

Побудувати кругову діаграму населення по регіонам

```
In 85  1  # pie chart with population per region
       2
       3  df.groupby('Region')['Population'].sum().plot(kind='pie', figsize=(4, 4), autopct='%1.1f%%')
       4  plt.ylabel(None)
       5  plt.show()
          Executed at 2023.06.25 21:50:02 in 182ms
```



Висновок

У цьому комп’ютерному практикумі я використав очистку даних з роботи №3. Але тут підтвердилася підозра про не нормалізовані дані. У минулій роботі було багато викидів і тут ми бачимо що дата сет не нормалізовий. Також я побудував різні графіки для кращого аналізу даних