# Отчет по ДЗ

Разработка программы на языке Python

Дисциплина: Парадигмы и конструкции языков программирования

Выполнил: Уйданов Лев

Группа: ИБМ3-34Б

Преподаватель: Гапанюк Ю.Е.

Москва

2025

# Домашнее задание

**Цель домашнего задания:** формирование и анализ датасета.

## Задание:

1. Осуществите разбор данных сайта с использованием библиотеки BeautifulSoup.
2. Сформируйте датасет табличного типа с использованием Python и сохраните датасет в формате CSV.
3. Проведите разведочный анализ данных для сформированного датасета.

```
import requests
from bs4 import BeautifulSoup
import pandas as pd

URL = "https://www.scrapethissite.com/pages/simple/"
HEADERS = {
    "User-Agent": "Mozilla/5.0"
}


def scrape_countries():
    response = requests.get(URL, headers=HEADERS)
    if response.status_code != 200:
        print("Страница не найдена")
        return []

    soup = BeautifulSoup(response.text, "html.parser")

    countries = []

    blocks = soup.find_all("div", class_="country")
    if not blocks:
        print("Данные не найдены")
        return []

    for block in blocks:
        def safe_text(parent, tag, class_name):
            el = parent.find(tag, class_=class_name)
            return el.get_text(strip=True) if el else ""

        name = safe_text(block, "h3", "country-name")
        capital = safe_text(block, "span", "country-capital")
        population = safe_text(block, "span", "country-population")
        area = safe_text(block, "span", "country-area")
        region = safe_text(block, "span", "country-region")

        if not name:
            continue

        countries.append({
            "country": name,
            "capital": capital,
            "population": int(population) if population.isdigit() else None,
            "area": float(area) if area else None,
```

```
            "region": region
        })

    return countries


def main():
    data = scrape_countries()
    print(f"\nСобрано стран: {len(data)}")

    if not data:
        print("Нет данных для анализа")
        return

    df = pd.DataFrame(data)

    df.to_csv("countries_dataset.csv", index=False, encoding="utf-8")
    print("CSV сохранён: countries_dataset.csv")

    print("\n=== INFO ===")
    print(df.info())

    print("\n=== DESCRIBE ===")
    print(df.describe())

    print("\n=== ТОП-5 СТРАН ПО НАСЕЛЕНИЮ ===")
    print(df.sort_values("population", ascending=False).head(5)[
        ["country", "population"]
    ])

    print("\n=== КОЛИЧЕСТВО СТРАН ПО РЕГИОНАМ ===")
    print(df["region"].value_counts())


if __name__ == "__main__":
    main()
```

Вывод:
Собрано стран: 250

CSV сохранён: countries_dataset.csv

=== INFO ===

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 250 entries, 0 to 249

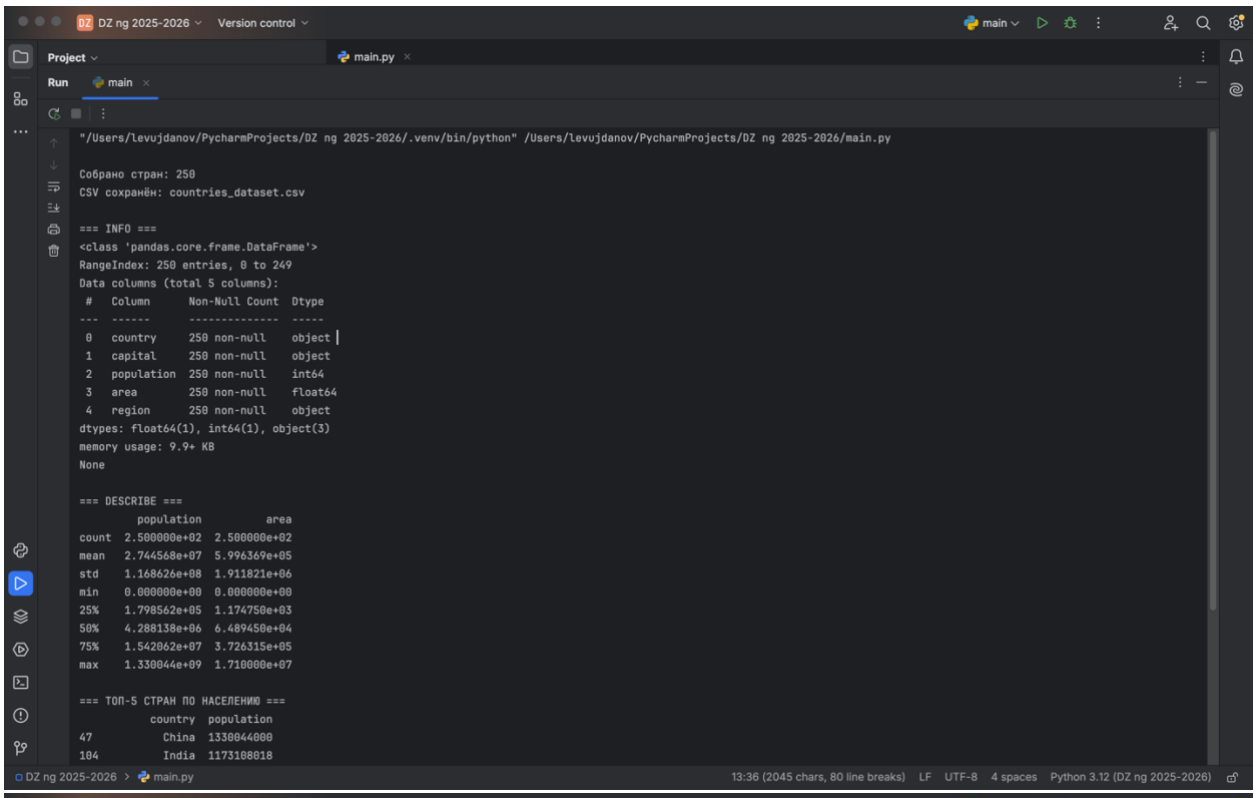
Data columns (total 5 columns):

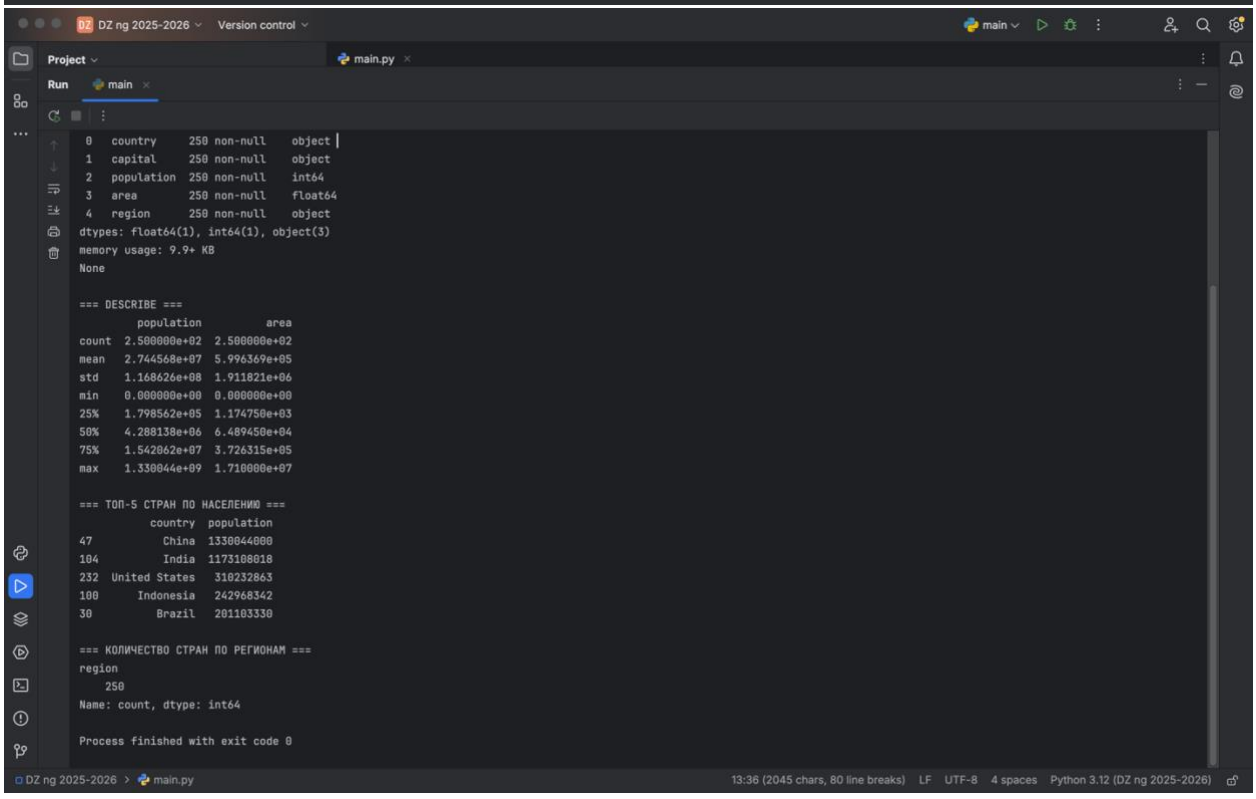
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | country | 250 non-null | object |

```
 1   capital     250 non-null    object
 2   population  250 non-null    int64
 3   area        250 non-null    float64
 4   region      250 non-null    object
dtypes: float64(1), int64(1), object(3)
memory usage: 9.9+ KB
None


=== DESCRIBE ===
         population          area
count  2.500000e+02  2.500000e+02
mean   2.744568e+07  5.996369e+05
std    1.168626e+08  1.911821e+06
min    0.000000e+00  0.000000e+00
25%    1.798562e+05  1.174750e+03
50%    4.288138e+06  6.489450e+04
75%    1.542062e+07  3.726315e+05
max    1.330044e+09  1.710000e+07


=== ТОП-5 СТРАН ПО НАСЕЛЕНИЮ ===
           country  population
47           China  1330044000
104          India  1173108018
232  United States   310232863
```

```
100      Indonesia   242968342
30          Brazil   201103330

=== КОЛИЧЕСТВО СТРАН ПО РЕГИОНАМ ===
region
   250
Name: count, dtype: int64

Process finished with exit code 0
```

Вывод скрином:




В рамках данной работы был выполнен парсинг веб-сайта scrapethissite.com с использованием библиотек requests и BeautifulSoup. В результате сбора данных был сформирован датасет, содержащий информацию о 250 странах мира, включая названия стран, столицы, численность населения и площадь территории. Полученные данные были сохранены в формате CSV и подготовлены для последующего анализа. Для первичного анализа была применена описательная статистика, позволившая оценить распределение населения и площади стран. В целом сформированный датасет является корректным и пригодным для дальнейшего статистического анализа.