|  |  |  |
| --- | --- | --- |
| Защищено:  Нардид А.Н.    "\_\_"\_\_\_\_\_\_\_\_\_\_\_\_\_2024 г. |  | Демонстрация:  Нардид А.Н.  "\_\_"\_\_\_\_\_\_\_\_\_\_\_\_\_2024 г. |

**Отчет по ДЗ по курсу**

**Парадигмы и конструкции языков программирования**

**ГУИМЦ**

#### Тема работы: " Формирование и анализ датасета. "

4

(количество листов)

|  |  |
| --- | --- |
| ИСПОЛНИТЕЛЬ: |  |
| студент группы ИУ5Ц-53Б | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | (подпись) |
| Пронин В.К. | "\_\_"\_\_\_\_\_\_\_\_\_\_\_\_\_2024 г. |

Москва, МГТУ - 2024 г.

# Тема и задание для выполнения лабораторной работы.

Тема работы: «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.

# Листинг программы

# data\_scraper.py

import requests

from bs4 import BeautifulSoup

import csv

def scrape\_data(url, output\_file):

    print(f"Connecting to {url}...")

    try:

        response = requests.get(url)

        response.raise\_for\_status()

    except requests.exceptions.RequestException as e:

        print(f"Error fetching URL: {e}")

        return

    soup = BeautifulSoup(response.content, 'html.parser')

    print("Parsing data...")

    # Пример парсинга таблицы (измените селектор для вашего сайта)

    data = []

    table = soup.find('table')  # Найдите таблицу

    if not table:

        print("Error: No table found on the page. Check the website structure.")

        return

    table\_rows = table.find\_all('tr')

    for row in table\_rows:

        columns = [col.text.strip() for col in row.find\_all(['th', 'td'])]

        if columns:

            data.append(columns)

    if not data:

        print("No data found. Ensure the selectors match the website structure.")

        return

    print(f"Saving data to {output\_file}...")

    try:

        with open(output\_file, 'w', newline='', encoding='utf-8') as csvfile:

            writer = csv.writer(csvfile)

            writer.writerows(data)

        print(f"Data successfully saved to {output\_file}")

    except Exception as e:

        print(f"Error saving file: {e}")

if \_\_name\_\_ == "\_\_main\_\_":

    url = input("Enter the URL to scrape: ")

    output\_file = "data.csv"  # Имя файла для сохранения

    scrape\_data(url, output\_file)

# data\_analysis.py

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

def load\_dataset(file\_path):

    try:

        print(f"Loading dataset from {file\_path}...")

        df = pd.read\_csv(file\_path)

        print("Dataset loaded successfully.")

        return df

    except FileNotFoundError:

        print(f"Error: File {file\_path} not found. Ensure the scraping script ran successfully.")

        return None

    except Exception as e:

        print(f"Error loading dataset: {e}")

        return None

def explore\_data(df):

    print("\n--- Dataset Overview ---")

    print(df.info())

    print("\n--- First 5 Rows ---")

    print(df.head())

    print("\n--- Dataset Statistics ---")

    print(df.describe(include='all'))

    print("\n--- Missing Values ---")

    print(df.isnull().sum())

def visualize\_data(df):

    numeric\_columns = df.select\_dtypes(include=['float64', 'int64']).columns

    for col in numeric\_columns:

        sns.histplot(df[col], kde=True)

        plt.title(f'Distribution of {col}')

        plt.show()

    if not numeric\_columns.empty:

        plt.figure(figsize=(10, 8))

        sns.heatmap(df.corr(), annot=True, cmap="coolwarm")

        plt.title("Correlation Matrix")

        plt.show()

if \_\_name\_\_ == "\_\_main\_\_":

    file\_path = "data.csv"

    df = load\_dataset(file\_path)

    if df is not None:

        explore\_data(df)

        visualize\_data(df)

# Результаты работы программы

C:\Users\Вячеслав\python\П\_и\_К\_ЯП\ДЗ>python data\_scraper.py

Enter the URL to scrape: https://docs.google.com/spreadsheets/d/167qAN1H8TW\_Og1-KK-qeTkW7mPcUAj\_FzKWMfVS\_QZ4/edit?gid=1854267334#gid=1854267334

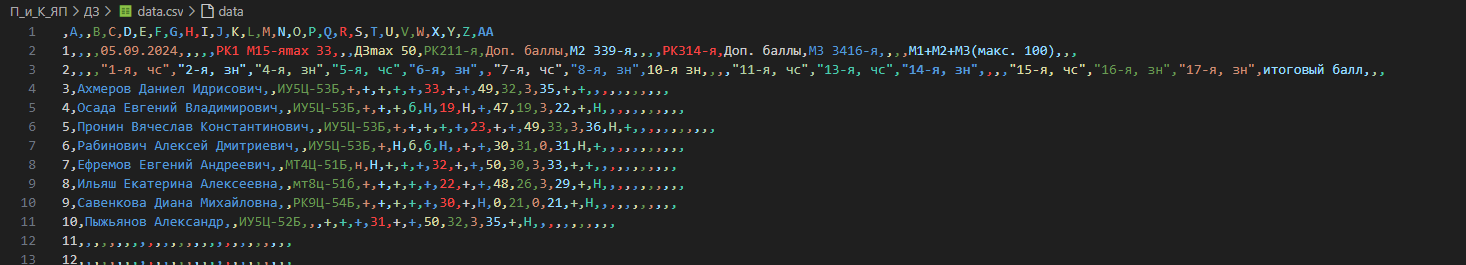
Connecting to https://docs.google.com/spreadsheets/d/167qAN1H8TW\_Og1-KK-qeTkW7mPcUAj\_FzKWMfVS\_QZ4/edit?gid=1854267334#gid=1854267334...

Parsing data...

Saving data to data.csv...

Data successfully saved to data.csv

**data.csv**

****