МИНЕСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Кафедра «Измерительно-вычислительные комплексы»

«Технология обработки информации»

Отчет по лабораторной работе №3
Исследование библиотек CSV, pandas

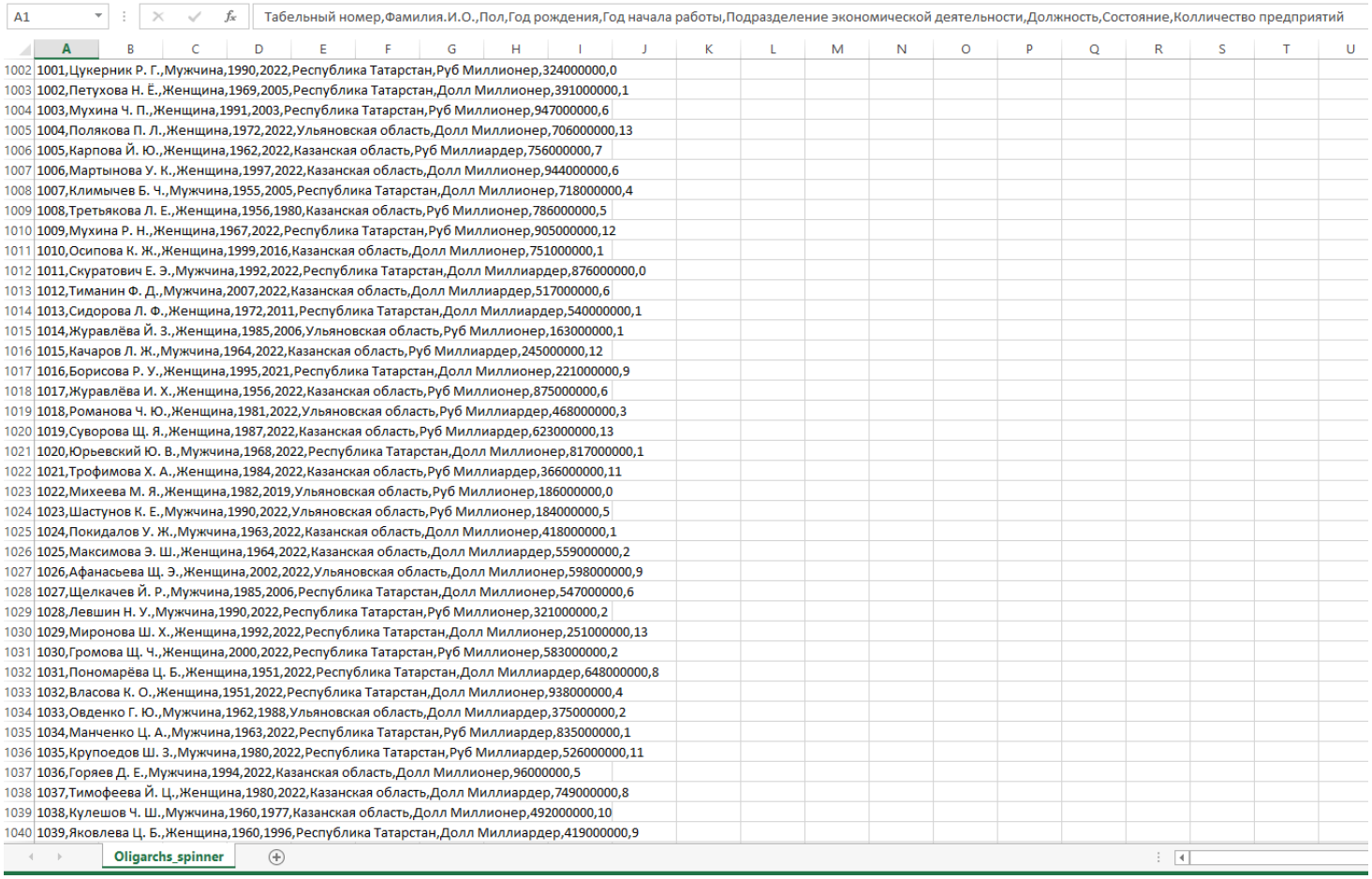
Выполнил:
Студент группы ИСТбд-41
Калашников М. А.
Проверил:
Шишкин В.В.

Ульяновск
2022

ИСТд-4

Л.р. №3. Исследование библиотек CSV, pandas

1. Создать симулированный набор данных и записать его на диск в виде csv файла со следующими параметрами:
- количество строк не менее 1000 (задается случайным образом);
- структура набора:
    - табельный номер;
    - Фамилия И.О.;
    - пол;
    - год рождения;
    - год начала работы в компании;
    - подразделение;
    - должность;
    - оклад;
    - количество выполненных проектов



A1 | Табельный номер,Фамилия.И.О.,Пол,Год рождения,Год начала работы,Подразделение экономической деятельности,Должность,Состояние,Колличество предприятий

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1002	1001,Цукерник Р. Г.,Мужчина,1990,2022,Республика Татарстан,Руб Миллионер,324000000,0																				
1003	1002,Петухова Н. Ё.,Женщина,1969,2005,Республика Татарстан,Долл Миллионер,391000000,1																				
1004	1003,Мухина Ч. П.,Женщина,1991,2003,Республика Татарстан,Руб Миллионер,947000000,6																				
1005	1004,Полякова П. Л.,Женщина,1972,2022,Ульяновская область,Долл Миллионер,706000000,13																				
1006	1005,Карпова Й. Ю.,Женщина,1962,2022,Казанская область,Руб Миллиардер,756000000,7																				
1007	1006,Мартынова У. К.,Женщина,1997,2022,Казанская область,Долл Миллионер,944000000,6																				
1008	1007,Климычев Б. Ч.,Мужчина,1955,2005,Республика Татарстан,Долл Миллионер,718000000,4																				
1009	1008,Третьякова Л. Е.,Женщина,1956,1980,Казанская область,Руб Миллионер,786000000,5																				
1010	1009,Мухина Р. Н.,Женщина,1967,2022,Республика Татарстан,Руб Миллионер,905000000,12																				
1011	1010,Осипова К. Ж.,Женщина,1999,2016,Казанская область,Долл Миллионер,751000000,1																				
1012	1011,Скуратович Е. Э.,Мужчина,1992,2022,Республика Татарстан,Долл Миллиардер,876000000,0																				
1013	1012,Тиманин Ф. Д.,Мужчина,2007,2022,Казанская область,Долл Миллиардер,517000000,6																				
1014	1013,Сидорова Л. Ф.,Женщина,1972,2011,Республика Татарстан,Долл Миллиардер,540000000,1																				
1015	1014,Журавлёва Й. З.,Женщина,1985,2006,Ульяновская область,Руб Миллионер,163000000,1																				
1016	1015,Качаров Л. Ж.,Мужчина,1964,2022,Казанская область,Руб Миллиардер,245000000,12																				
1017	1016,Борисова Р. У.,Женщина,1995,2021,Республика Татарстан,Долл Миллионер,221000000,9																				
1018	1017,Журавлёва И. Х.,Женщина,1956,2022,Казанская область,Руб Миллионер,875000000,6																				
1019	1018,Романова Ч. Ю.,Женщина,1981,2022,Ульяновская область,Руб Миллиардер,468000000,3																				
1020	1019,Суворова Щ. Я.,Женщина,1987,2022,Казанская область,Руб Миллиардер,623000000,13																				
1021	1020,Юрьевский Ю. В.,Мужчина,1968,2022,Республика Татарстан,Долл Миллионер,817000000,1																				
1022	1021,Трофимова Х. А.,Женщина,1984,2022,Казанская область,Руб Миллиардер,366000000,11																				
1023	1022,Михеева М. Я.,Женщина,1982,2019,Ульяновская область,Руб Миллионер,186000000,0																				
1024	1023,Шастунов К. Е.,Мужчина,1990,2022,Ульяновская область,Руб Миллионер,184000000,5																				
1025	1024,Покидалов У. Ж.,Мужчина,1963,2022,Казанская область,Долл Миллионер,418000000,1																				
1026	1025,Максимова Э. Ш.,Женщина,1964,2022,Казанская область,Долл Миллиардер,559000000,2																				
1027	1026,Афанасьева Щ. Э.,Женщина,2002,2022,Ульяновская область,Долл Миллионер,598000000,9																				
1028	1027,Щелкачев Й. Р.,Мужчина,1985,2006,Республика Татарстан,Долл Миллионер,547000000,6																				
1029	1028,Левшин Н. У.,Мужчина,1990,2022,Республика Татарстан,Руб Миллионер,321000000,2																				
1030	1029,Миронова Ш. Х.,Женщина,1992,2022,Республика Татарстан,Долл Миллионер,251000000,13																				
1031	1030,Громова Щ. Ч.,Женщина,2000,2022,Республика Татарстан,Руб Миллионер,583000000,2																				
1032	1031,Пономарёва Ц. Б.,Женщина,1951,2022,Республика Татарстан,Долл Миллиардер,648000000,8																				
1033	1032,Власова К. О.,Женщина,1951,2022,Республика Татарстан,Долл Миллионер,938000000,4																				
1034	1033,Овденко Г. Ю.,Мужчина,1962,1988,Ульяновская область,Долл Миллиардер,375000000,2																				
1035	1034,Манченко Ц. А.,Мужчина,1963,2022,Республика Татарстан,Руб Миллиардер,835000000,1																				
1036	1035,Крупоедов Ш. З.,Мужчина,1980,2022,Республика Татарстан,Руб Миллиардер,526000000,11																				
1037	1036,Горяев Д. Е.,Мужчина,1994,2022,Казанская область,Долл Миллионер,96000000,5																				
1038	1037,Тимофеева Й. Ц.,Женщина,1980,2022,Казанская область,Долл Миллиардер,749000000,8																				
1039	1038,Кулешов Ч. Ш.,Мужчина,1960,1977,Казанская область,Долл Миллионер,492000000,10																				
1040	1039,Яковлева Ц. Б.,Женщина,1960,1996,Республика Татарстан,Долл Миллиардер,419000000,9																				

Oligarchs_spinner

2. Прочитать сгенерированный набор данных в виде списков и получить с помощью программирования и методов библиотеки numpy для разных по типу признаков столбцов (не менее 3) основные статистические характеристики (например для порядкового типа: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, мода).

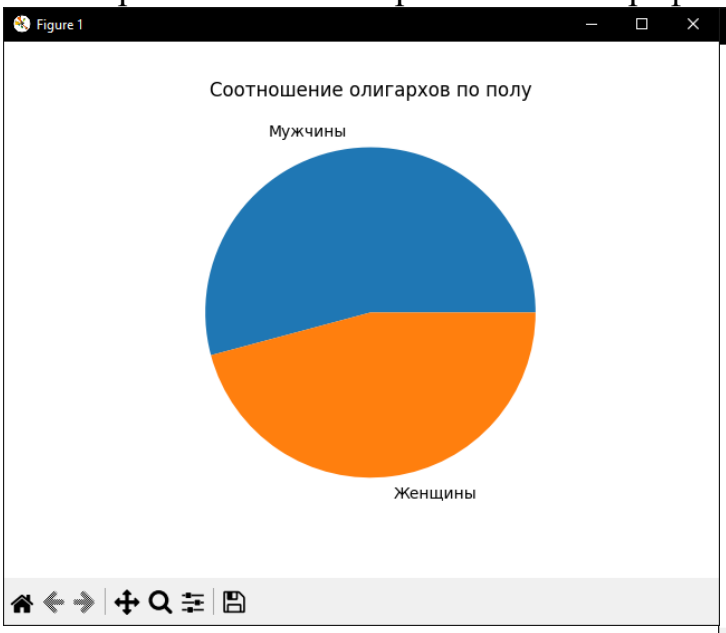
```
D:\labs3\venv\Scripts\python.exe D:\labs3\main.py
Задание по NUMPY
 ---------------
 Статистика по миллионерам и миллиардерам
 Количество олигархов всего: 1039
Подъём экономики был в 2022 году.
Так-как колличество новых олигархов больше в этом году
---------------
Доля олигархов мужского пола: 0.496
Доля олигархов женского пола: 0.503
---------------
Денежное состояние олигархов
 Максимальное состояние олигархов: 999000000
Минимальная состояние олигархов: 1000000
Среднее денежное состояние олигархов: 510487969.201
Дисперсия денежного состояния олигархов: 8.482377824794515e+16
Ст. откл. состояния олигархов: 291245220.129
Медиана денежного состояния олигархов: 517000000.0
Мода денежного состояния олигархов: 79000000
---------------
Группировка олигархов по регионам
Количество олигархов в Ульяновске: 333
Количество олигархов в Республике Татарстан: 338
Количество олигархов в Казанской области: 368
Больше всего сотрудников в: Казанской области
```

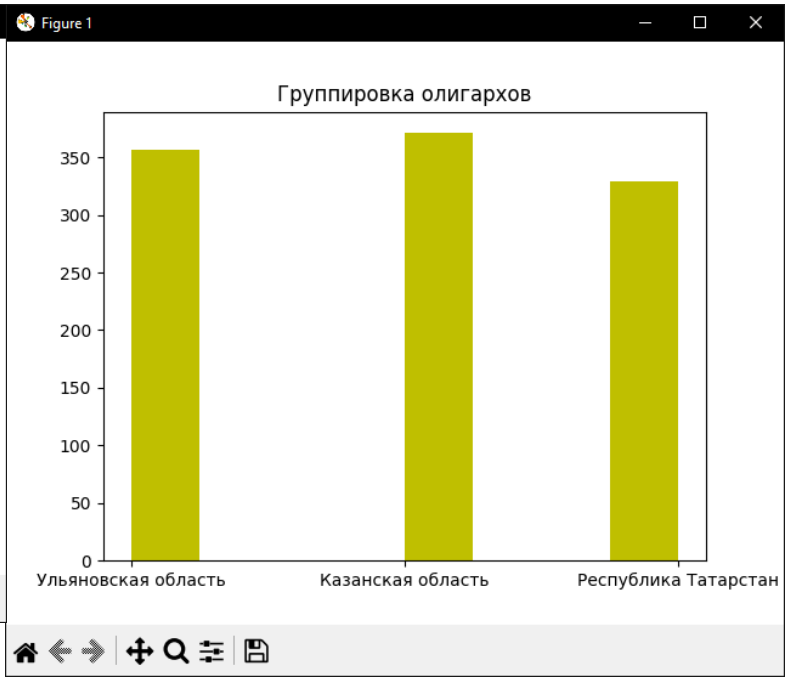
3. Прочитать сгенерированный набор данных в виде датафрейма и получить с помощью методов библиотеки pandas для тех же столбцов те же статистические характеристики. Продемонстрировать применение не менее 3 методов библиотеки pandas.

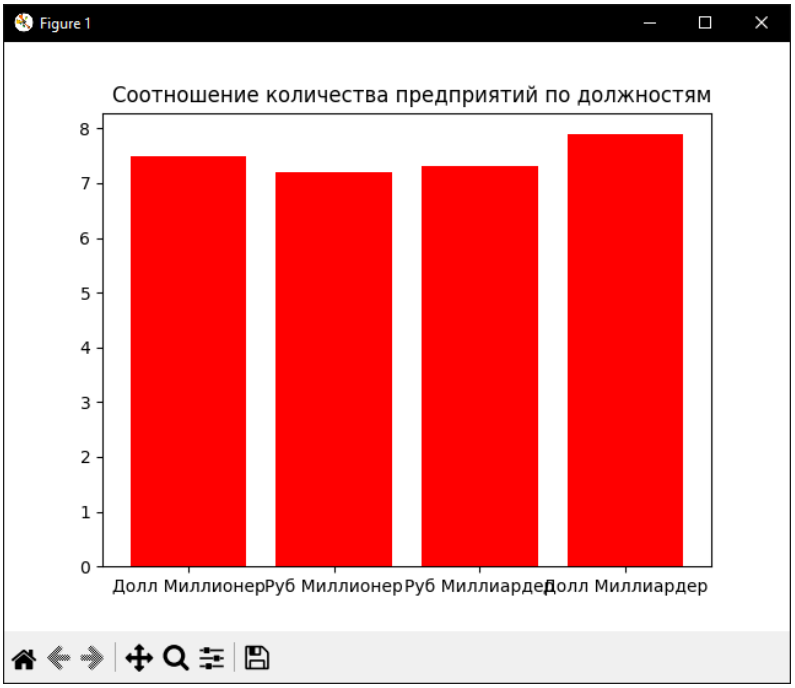
```
Задание PANDAS
---------------
Статистика олигархов
Общее колличество олигархов: 1039
Доля олигархов мужского пола: 0.497
Доля олигархов женского пола: 0.503
Больше всего олигархов появилось в 2022 году
---------------
Денежное состояние
Максимальное денежное состояние олигархов: 999000000
Минимальное денежное состояние олигархов: 1000000
Дисперсия денежного состояния: 8.490549672409923e+16
Ст. откл. денежного состояния олгархов: 291385477.888
Медиана кол-ва денежного состояния олигархов: 517000000.0
Мода денежного состояния 79000000
---------------
Группировка олигархов
Количество олигархов в Ульяновской области: 333
Количество олигархов в Республике Татарстан: 338
Количество олигархов в Казанской области: 368
Больше всего олигархов в: Казанской области

Process finished with exit code 0
```

4. Построить не менее 3 разнотипных графиков.







5. Оценить возможности библиотек csv, numpy, pandas в форме отчета по лабораторной работе.

CSV.

Формат CSV (Comma Separated Values) является наиболее распространенным и удобным форматом для работы с данными в виде таблиц. Формат является достаточно универсальным и позволяет хранить разносортные данные. Модуль CSV позволяет работать с CSV файлами не задумываясь о их структуре и т.д.

NumPy.

Библиотека NumPy позволяет использовать различные математические вычисления. Библиотека является достаточно быстрой т.к. базируется на коде написанном на языках C/C++/Fortran. Из положительных сторон можно отметить: использование высокоуровневых функций при работе с библиотекой; наличие разнообразных математических средств ; высокая скорость работы.

Pandas.

Библиотека Pandas позволяет осуществлять анализ данных. Библиотека является достаточно быстрой т.к. базируется на библиотеке NumPy. Из положительных сторон можно отметить: использование высокоуровневых функций при работе с библиотекой; наличие разнообразных средств для анализа данных; высокая скорость работы.

Листинг программы

```
import numpy as np
import pandas as pd
import csv
import random
from datetime import date
import matplotlib.pyplot as plt
```

```
    def table():
        with open('Maleling.txt', 'r', encoding='utf-8') as f:
            Male = [i.rstrip() for i in f]

        with open('Femaling.txt', 'r', encoding='utf-8') as f:
            Female = [i.rstrip() for i in f]

        OligarchList = {
            "Ульяновская область": ["Руб Миллионер", "Руб Миллиардер",
                                    "Долл Миллионер", "Долл Миллиардер"],
            "Республика Татарстан": ["Руб Миллионер", "Руб Миллиардер",
                                     "Долл Миллионер", "Долл Миллиардер"],
            "Казанская область": ["Руб Миллионер", "Руб Миллиардер",
                                  "Долл Миллионер", "Долл Миллиардер"],
        }
        # Заполнение CSV
        with open("Oligarchs_spinner.csv", 'w', newline='') as csvfile:
            writer = csv.writer(csvfile)
            writer.writerow(
                (
                    "Табельный номер", "Фамилия.И.О.", "Пол", "Год рождения", "Год начала
работы",
                    "Подразделение экономической деятельности", "Должность", "Состояние",
"Колличество предприятий"
                )
            )
        # Генерация Имени и Отчества
        def First_and_middle_name():
            return (random.choice('АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЭЮЯ') + ". " +
random.choice(
                'АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЭЮЯ') + ".")
        # Генерация таблицы олигархов
        for Employees in range(random.randint(1000, 1111)):
            Oligarchs = random.choice(list(OligarchList))
            sex = random.choice(["Мужчина", "Женщина"])
            Birthday = random.randint(1950, 2010)
            StartingYear = min(date.today().year, Birthday + random.randint(12, 99))
            MoneyCondition = random.randrange(1000000, 1000000000, 1000000)
            EnterprisesAmount = random.randint(0, 15)

            if (sex == "Мужчина"):
                Name = Male[random.randint(0, len(Male) - 1)]
            else:
                Name = Female[random.randint(0, 111)]
            # Заполнение CSV
            with open("Oligarchs_spinner.csv", 'a', newline='') as csvfile:
                writer = csv.writer(csvfile)
```

```
                writer.writerow(
                    [
                        Employees + 1,
                        (Name + " " + First_and_middle_name()), sex, Birthday, StartingYear,
                        Oligarchs, random.choice(OligarchList[Oligarchs]),
                        MoneyCondition, EnterprisesAmount
                    ]
                )
        # Работа с помощью NUMPY
        def numpy():
            with open("Oligarchs_spinner.csv") as csvfile:
                metadata = [list(row) for row in csv.reader(csvfile)]
            data = np.array(metadata)
            sexs = data[:, 2]
            Oligarchs = data[:, 5]
            MoneyCondition = data[1:, 7].astype("int32")
            StartingYear = data[1:, 4].astype("int32")
            count_MoneyCondition = np.bincount(MoneyCondition)
            count_StartingYear = np.bincount(StartingYear)
            # Вывод данных в консоль(Numpy)
            print("Задание по NUMPY \n ---------- \n Статистика по миллионерам и миллиардерам \n "
                  "Количество олигархов всего:", np.count_nonzero(MoneyCondition), "\nПодъём экономики был в",
                  np.argmax(count_StartingYear), "году. \nТак-как колличество новых олигархов больше в этом году"
                  "\n----------" "\nДоля олигархов мужского пола:", round(np.sum(sexs == "Мужчина") / np.size(sexs), 3),
                  "\nДоля олигархов женского пола:", round(np.sum(sexs == "Женщина") / np.size(sexs), 3), "\n----------"
                  "\nДенежное состояние олигархов \n Максимальное состояние олигархов:", np.max(MoneyCondition),
                  "\nМинимальная состояние олигархов:", np.min(MoneyCondition), "\nСреднее денежное состояние олигархов:",
                  round(np.average(MoneyCondition), 3), "\nДисперсия денежного состояния олигархов:",
                  round(np.var(MoneyCondition), 3), "\nСт. откл. состояния олигархов:", round(np.std(MoneyCondition), 3),
                  "\nМедиана денежного состояния олигархов:",np.median(MoneyCondition), "\nМода денежного состояния олигархов:",
                  np.argmax(count_MoneyCondition), "\n----------\nГруппировка олигархов по регионам"
                                                   "\nКоличество олигархов в Ульяновске:",
                  np.count_nonzero(Oligarchs == "Ульяновская область"), "\nКоличество олигархов в Республике Татарстан:",
                  np.count_nonzero(Oligarchs == "Республика Татарстан"), "\nКоличество олигархов в Казанской области:",
```

```
        np.count_nonzero(Oligarchs == "Казанская область"), "\nБольше всего сотрудников в:",
        "Ульяновской области" if ((np.count_nonzero(Oligarchs == "Ульяновская область") \
                                   > np.count_nonzero(Oligarchs == "Республика Татарстан")) and (
                                           np.count_nonzero(Oligarchs == "Ульяновская область") \
                                           > np.count_nonzero(Oligarchs == "Казанская область")))
        else ("Республике Татарстан" if (np.count_nonzero(Oligarchs == "Республика Татарстан") \
                                         > np.count_nonzero(Oligarchs == "Казанская область"))
              else "Казанской области"), "\n----------------")


    # Работа с помощью PANDAS
    def pandas():
        data = pd.read_csv("Oligarchs_spinner.csv", encoding='cp1251')
        # Вывод данных в консоль(Pandas)
        print("Задание PANDAS\n----------------\nСтатистика олигархов\nОбщее колличество олигархов:", data["Табельный номер"].count(),
        "\nДоля олигархов мужского пола:", round(data["Пол"].value_counts()["Мужчина"] / data["Пол"].shape[0], 3),
        "\nДоля олигархов женского пола:", round(data["Пол"].value_counts()["Женщина"] / data["Пол"].shape[0], 3),
        "\nБольше всего олигархов появилось в", data["Год начала работы"].mode()[0], "году\n----------------"
        "\nДенежное состояние\nМаксимальное денежное состояние олигархов:", data["Состояние"].max(),
        "\nМинимальное денежное состояние олигархов:", data["Состояние"].min(), "\nДисперсия денежного состояния:",
        round(data["Состояние"].var(), 3), "\nСт. откл. денежного состояния олгархов:", round(data["Состояние"].std(), 3),
        "\nМедиана кол-ва денежного состояния олигархов:", data["Состояние"].median(), "\nМода денежного состояния",
        data["Состояние"].mode()[0], "\n----------------\nГруппировка олигархов"
                                     "\nКоличество олигархов в Ульяновской области:",
        (data["Подразделение экономической деятельности"].value_counts()["Ульяновская область"]),
        "\nКоличество олигархов в Республике Татарстан:",
        (data["Подразделение экономической деятельности"].value_counts()["Республика Татарстан"]),
        "\nКоличество олигархов в Казанской области:",
        (data["Подразделение экономической деятельности"].value_counts()["Казанская область"]),
        "\nБольше всего олигархов в:",
```

```
            "Ульяновской области" if (
                    (data["Подразделение экономической деятельности"].value_counts()["Ульяновская область"] \
                    > data["Подразделение экономической деятельности"].value_counts()["Республика Татарстан"]) and (
                        data["Подразделение экономической деятельности"].value_counts()[
                            "Ульяновская область"] \
                        > data["Подразделение экономической деятельности"].value_counts()[
                            "Республика Татарстан"]))
                else ("Республике Татарстан" if (data["Подразделение экономической деятельности"].value_counts()
                                                 ["Республика Татарстан"] \
                      > data["Подразделение экономической деятельности"].value_counts()[
                          "Казанская область"]) else "Казанской области"))
        # Работа с графиками(3шт)
        def graphics():
            data = pd.read_csv("Oligarchs_spinner.csv", encoding='cp1251')
            proportion = {}
            positions = data['Должность'].unique()

            for item in positions:
                pos = data[data['Должность'] == item]
                proportion[item] = round(pos['Колличество предприятий'].sum() / pos['Колличество предприятий'].count(), 2)
            # График распределения соотношения по полу(мужской или женский)
            sex = [data["Пол"].value_counts()["Мужчина"], data["Пол"].value_counts()["Женщина"]]
            plt.pie(sex, labels=["Мужчины", "Женщины"])
            plt.title('Соотношение олигархов по полу', loc='center')
            plt.show()
            # График группировки по области нахождения(Ульяновская область, Казанская область, Республика Татарстан)
            plt.hist(data['Подразделение экономической деятельности'], bins=8, color='y')
            plt.title('Группировка олигархов', loc='center')
            plt.show()
            # График по колличеству олигархов в обределённой должности(миллионер или миллиардер, рублейвой или доллоровый)
            plt.bar(proportion.keys(), proportion.values(), color='r')
            plt.title('Соотношение количества предприятий по должностям', loc='right')
            plt.show()

        table()
        numpy()
        pandas()
        graphics()
```