МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ

ГОМЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ ИМЕНИ П.О.СУХОГО

Факультет автоматизированных и информационных систем

Кафедра «Информационные технологии»

направление специальности 1-40 05 01-12 Информационные системы

и технологии (в игровой индустрии)

ЛАБОРАТОРНАЯ РАБОТА №5

по дисциплине: « Скриптовые языки программирования»

на тему: «**Применение скриптовых языков для обработки данных.**»

Выполнил: студент гр. ИТИ-21

Важаня Я И

Принял: преподаватель

Гомель 2024

**Цель работы:** изучить средства Python для анализа данных; библиотеку Pandas; очистка данных; агрегирование и преобразование данных; исследование и классификацию данных.

**Задание:**

**Часть 1**

*Series*

1. Создайте объект Series, содержащий список группы. В качестве индексов используйте номер в журнале старосты. Индекс старосты задать «starosta».
2. Выведите: 1) список всех элементов, 2) каждого второго из списка, 3) набор индексов.

*DataFrame*

1. Создайте DataFrame cо столбцами Name, IsStarosta, Phone, Gender, YearBirth. Внесите 10 записей.
2. Выведите первые три записи DataFrame
3. Выведите столбик с именами.
4. Выведите 3 столбика вместе — имя и пол.
5. Выведите данные по старосте.
6. По умолчанию индексы нумеруются от 0. Сделайте в качестве индексов FIO — первые буквы фамилии, имени, отчества на латинском.
7. Добавьте столбик Age как разницу между текущим годом и годом рождения.
8. Выведите минимальный и максимальный возраст.
9. Выведите сводную таблицу: 1) с количеством девушек и юношей, 2) средним возрастом девушек и юношей.
10. Выведите распределение студентов по годам рождения.
11. Сохраните DataFrame в виде csv-файла.
12. Создайте новый DataFrame, заполните его данными из созданного ранее csv-файла.

**Часть 2**

*Задание 1*

Загрузите таблицу из файла polit.csv и сохраните датафрейм в переменную polit. Если в датафрейме есть строки с пропущенными значениями, удалите их. Сохраните изменения в polit.

*Задание 2*

Выберите в датафрейме строки, которые соответствуют странам со значениями индекса . Сохраните их в датафрейм not\_free.

*Задание 3*

Выберите в датафрейме строки, которые соответствуют странам Африки (afri) с процентом женщин в парламенте (fparl08) выше 30%. Сохраните их в датафрейм af\_w.

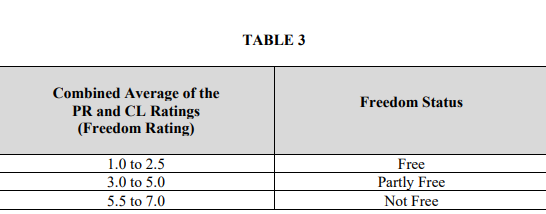
*Задание 4*

Выберите в датафрейме строки, которые соответствуют странам Африки или Латинской Америки (afri и lati) со значением индекса *Polity2* (polity09) больше или равным 8. Сохраните их в датафрейм la\_dem.

*Задание 5*

Добавьте в датафрейм polit столбец corr\_round, в котором будут храниться округленные до 2 знака после запятой значения индекса *Control of Corruption* (corr0509).

*Задание 6*

Добавьте в датафрейм polit столбец fh\_status, в котором будут храниться типы стран в зависимости от значения индекса *Freedom House* (значения *free*, *partly free*, *not free*). Соответствие значений fh09 с типами стран см.  (*Table 3* ).

*Задание 7*

Сгруппируйте строки в таблице в соответствии со значениями столбца fh\_status, полученного в предыдущей части и выведите минимальное, среднее и максимальное значение показателя gini (индекс Джини) по каждой группе.

*Задание 8*

Сгруппируйте строки в таблице в соответствии со значениями столбца fh\_status и запишите строки, относящиеся к разным группам, в отдельные csv-файлы.

**Ход работы**

**Часть 1**

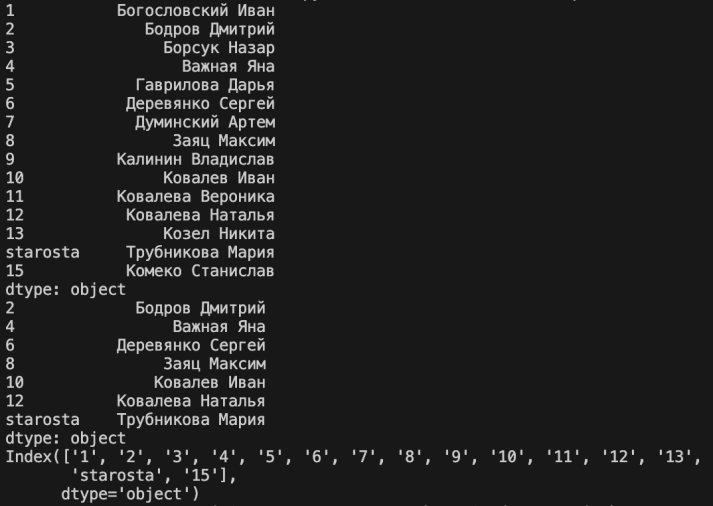
На рисунке 1 изображен объект Series, содержащий список группы(индекс старосты – «starosta»), каждый второй элемент из списка и набор индексов.

Рисунок 1 – Объект Series, содержащий список группы

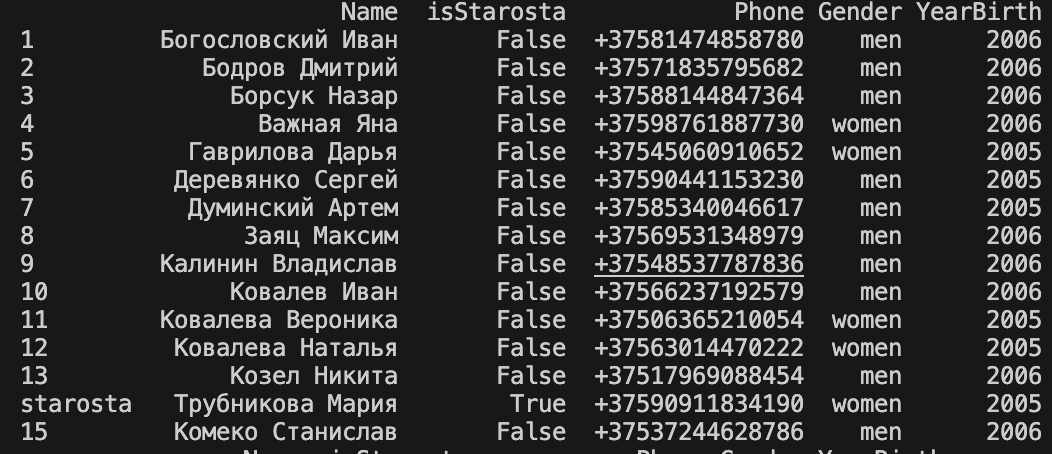
На рисунке 2 изображен *DataFrame* cо столбцами *Name*, *IsStarosta*, *Phone*, *Gender*, *YearBirth*.

Рисунок 2 – *DataFrame* c данными

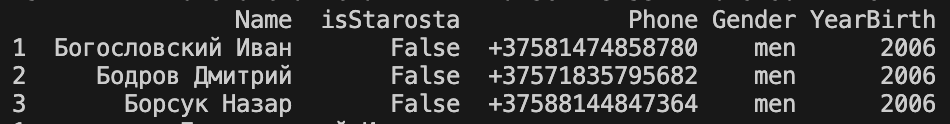
На рисунке 3 изображены первые три записи *DataFrame*.

Рисунок 3 – Первые три записи *DataFrame*

На рисунке 4 изображен столбик с именами.

Рисунок 4 – Столбик с именами

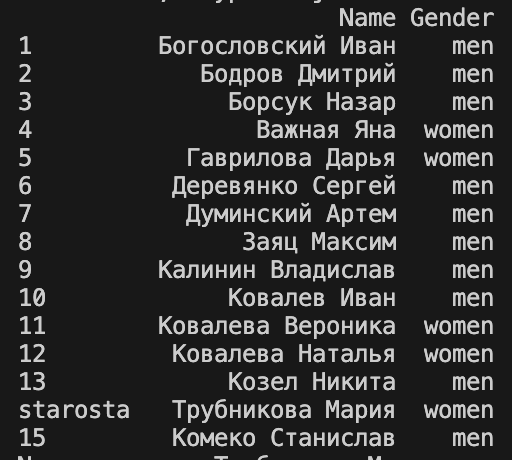
На рисунке 5 изображены 3 столбика вместе — имя и пол.

Рисунок 5 – 3 столбика вместе — имя и пол

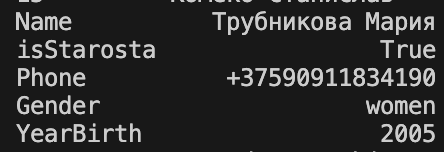
На рисунке 6 изображены данные по старосте.

Рисунок 6 – Результат выполнения 4-го пункта меню

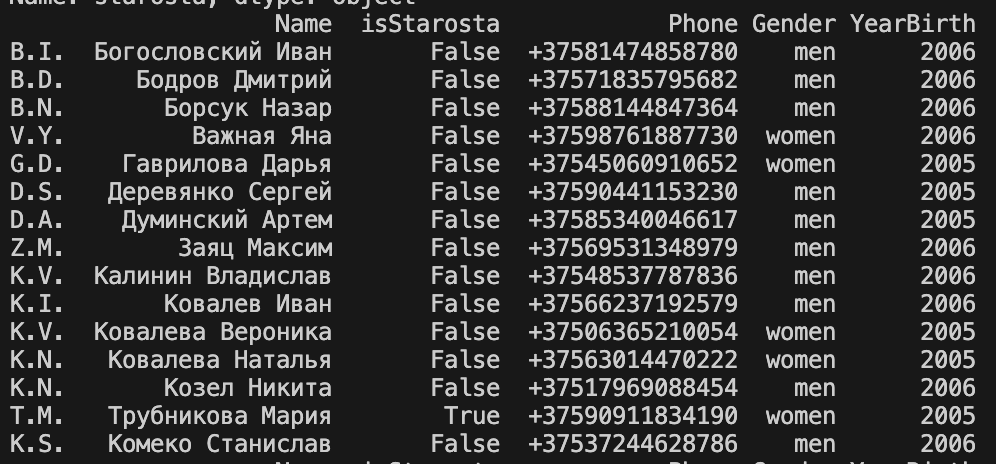
На рисунке 7 изображен *DataFrame,* в котором в качестве индексов FIO.

Рисунок 7 – *DataFrame,* в котором в качестве индексов FIO

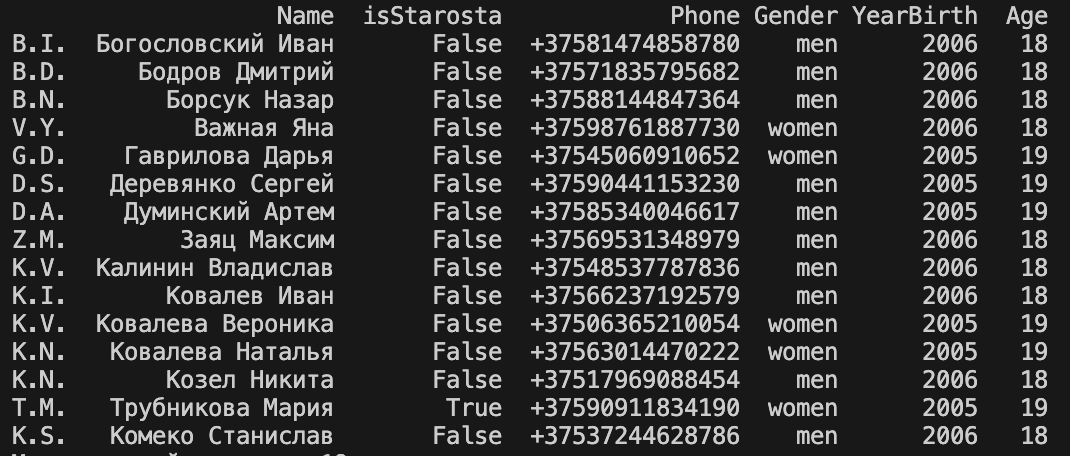
На рисунке 8 изображен *DataFrame* со столбиком Age как разница между текущим годом и годом рождения.

Рисунок 8 – *DataFrame* со столбиком Age

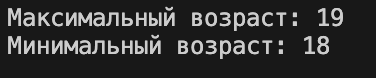
 На рисунке 9 изображен минимальный и максимальный возраст.

Рисунок 9 – Минимальный и максимальный возраст

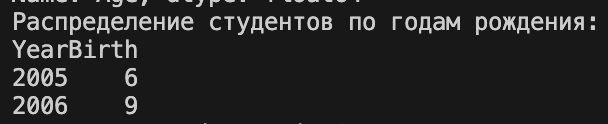
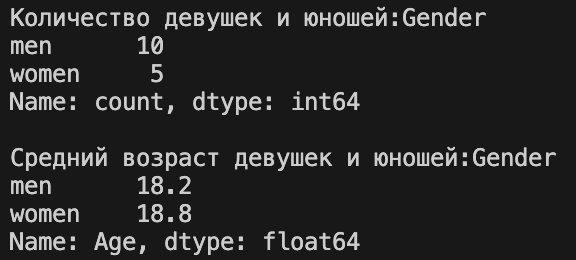
На рисунке 10 изображена сводная таблица: 1) с количеством девушек и юношей, 2) средним возрастом девушек и юношей.

Рисунок 10 – Сводная таблица

На рисунке 11 изображена распределение студентов по годам рождения.

Рисунок 11 – Распределение студентов по годам рождения

**Часть 2**

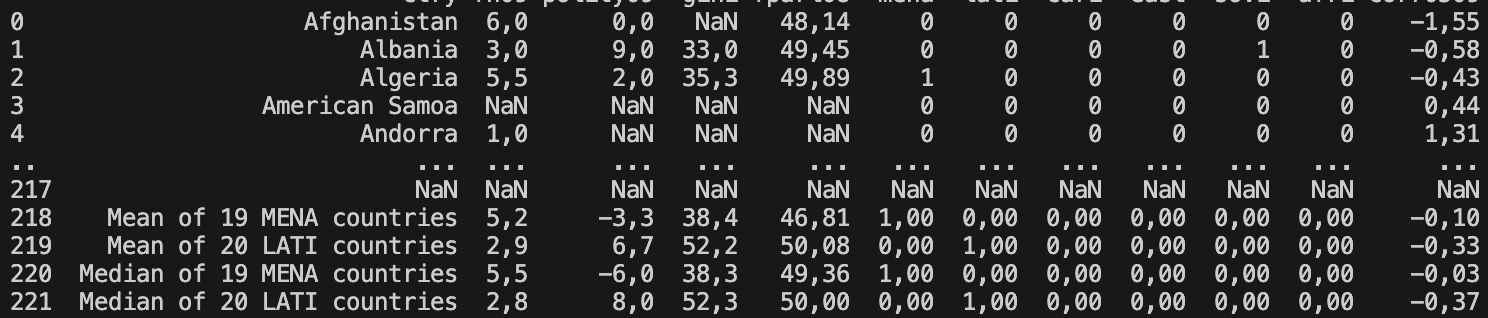
На рисунке 12 изображена таблица из файла polit.csv, которая сохранена как датафрейм в переменную polit.

Рисунок 12 – Таблица из файла polit.csv

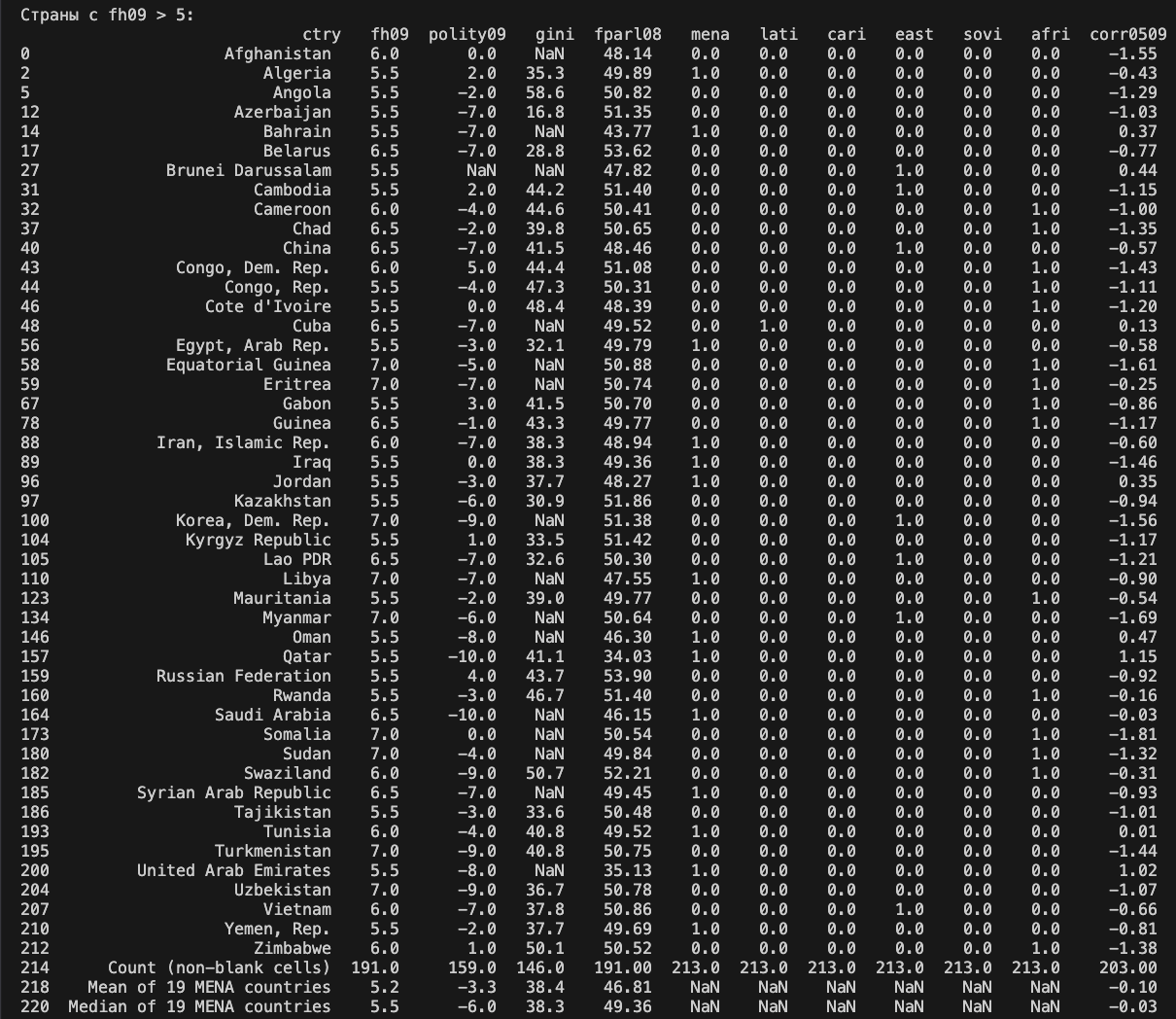
На рисунке 13 изображен датафрейм not\_free, строки в котором соответствуют странам со значениями индекса *Freedom House* (fh09) выше 5.

Рисунок 13 – Датафрейм not\_free

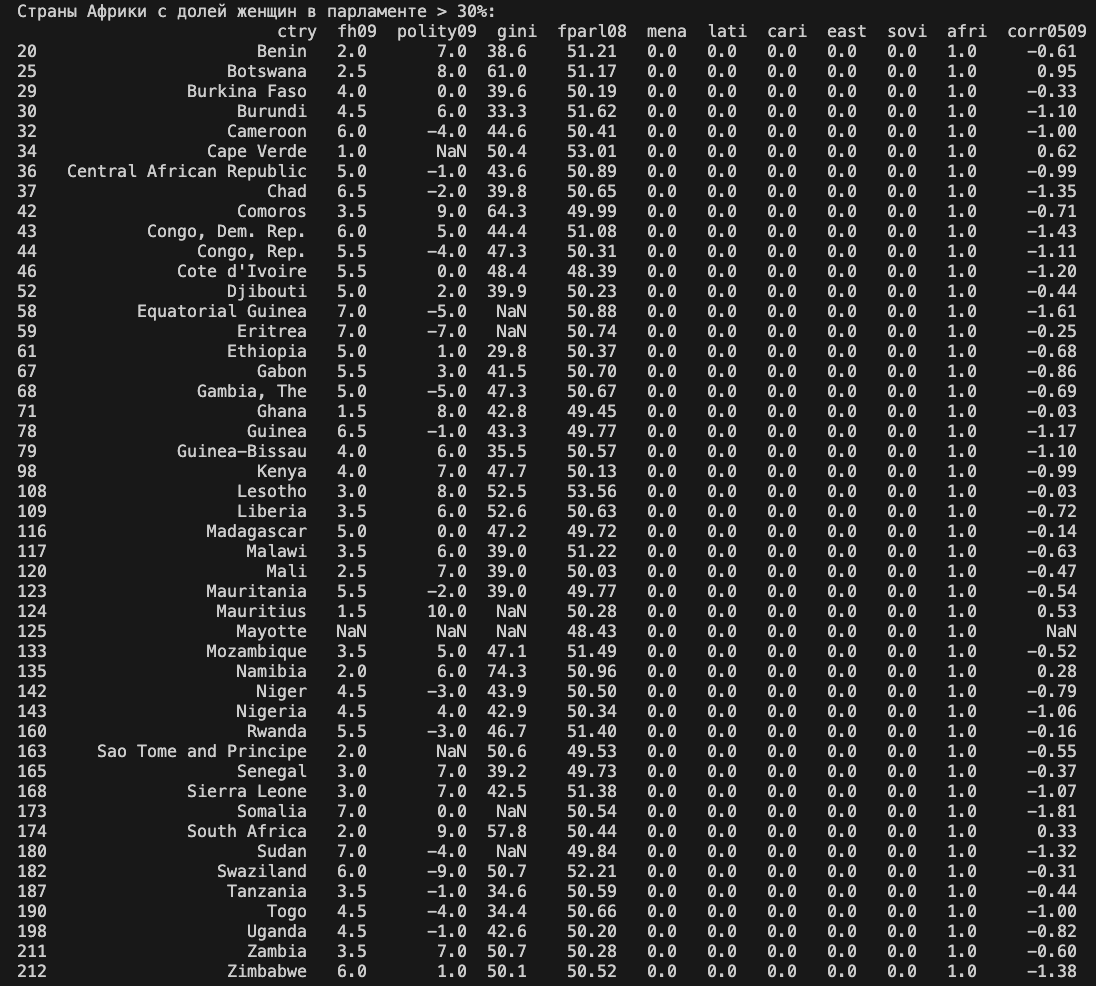
На рисунке 14 изображен датафрейм af\_w, строки в котором соответствуют странам Африки (afri) с процентом женщин в парламенте (fparl08) выше 30%.

Рисунок 14 – Датафрейм af\_w

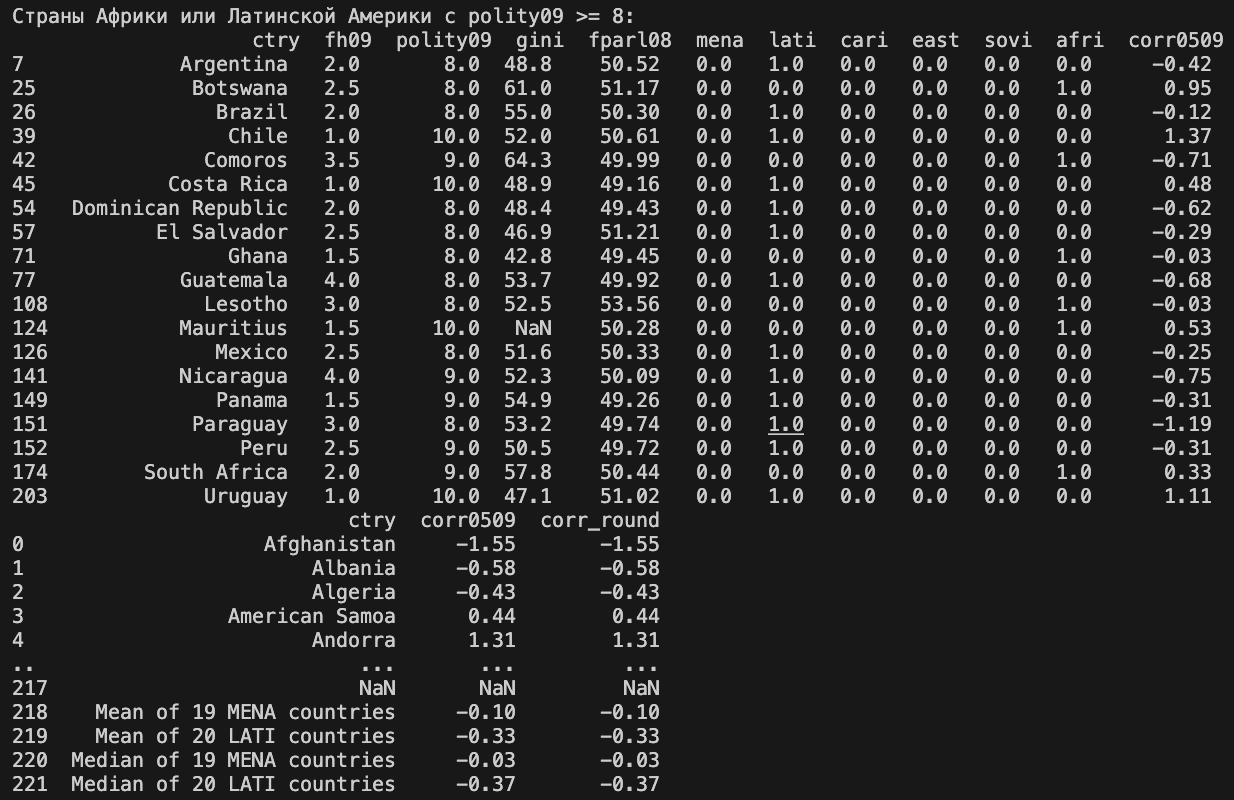
На рисунке 15 изображен датафрейм la\_dem, строки в котором соответствуют странам Африки или Латинской Америки (afri и lati) со значением индекса *Polity2* (polity09) больше или равным 8.

Рисунок 15 – Распределение студентов по годам рождения

На рисунке 16 изображена таблица со столбецом corr\_round, в котором будут храниться округленные до 2 знака после запятой значения индекса *Control of Corruption* (corr0509).

Рисунок 16 – Таблица со столбцом corr\_round

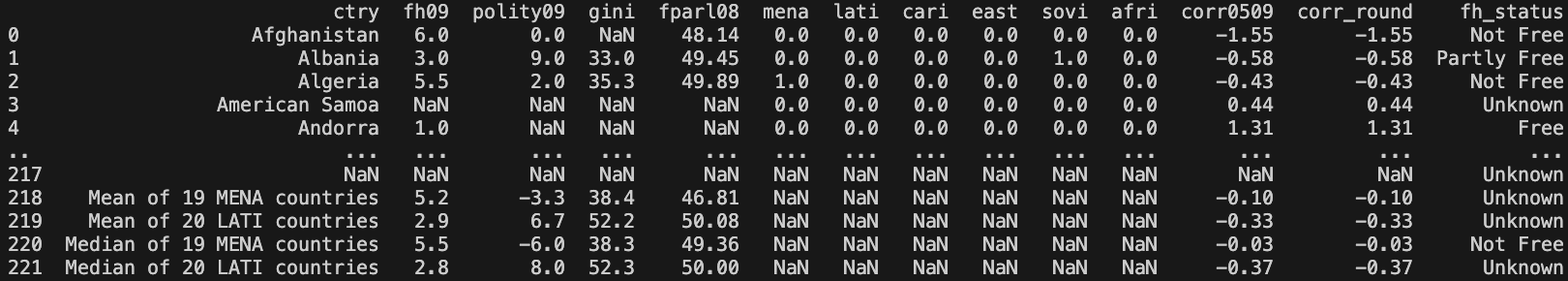
На рисунке 17 изображена таблица со столбцом fh\_status, в котором храняться типы стран в зависимости от значения индекса *Freedom House* (значения *free*, *partly free*, *not free*)

Рисунок 17 – Таблица со столбцом fh\_status

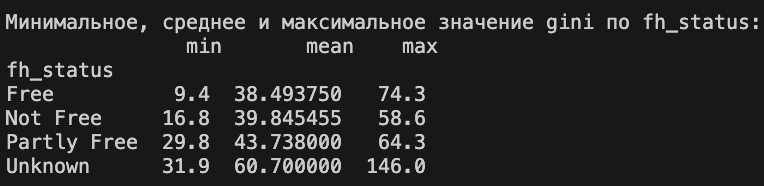
На рисунке 18 изображено минимальное, среднее и максимальное значение показателя gini (индекс Джини).

Рисунок 18 – Таблица со столбцом corr\_round

Вывод: в результате выполнения лабораторной работы изучили средства Python для анализа данных; библиотеку Pandas; очистку данных; агрегирование и преобразование данных; исследование и классификацию данных.

**ПРИЛОЖЕНИЕ А**

(обязательное)

Текст программы

**5lab.py:**

import pandas as pd

# 1 Список группы

index = ['1', '2', '3',

'4', '5', '6',

'7', '8', '9',

'10', '11', '12',

'13', 'starosta', '15']

members = ['Богословский Иван', 'Бодров Дмитрий', 'Борсук Назар',

'Важная Яна', 'Гаврилова Дарья', 'Деревянко Сергей',

'Думинский Артем', 'Заяц Максим', 'Калинин Владислав',

'Ковалев Иван', 'Ковалева Вероника', 'Ковалева Наталья',

'Козел Никита', 'Трубникова Мария', 'Комеко Станислав']

name = pd.Series(members, index=index)

# 2

print(name)

print(name[1::2])

print(name.index)

# 1 Создание DAtaFrane

is\_starosta = ([False for \_ in range(13)] + [True, False])

phone = (["+37581474858780", "+37571835795682", "+37588144847364",

"+37598761887730", "+37545060910652", "+37590441153230",

"+37585340046617", "+37569531348979", "+37548537787836",

"+37566237192579", "+37506365210054", "+37563014470222",

"+37517969088454", "+37590911834190", "+37537244628786"])

gender = (["men", "men", "men",

"women", "women", "men",

"men", "men", "men",

"men", "women", "women",

"men", "women", "men"])

year\_birth = (["2006", "2006", "2006",

"2006", "2005", "2005",

"2005", "2006", "2006",

"2006", "2005", "2005",

"2006", "2005", "2006"])

group = pd.DataFrame({'Name': members, 'isStarosta': is\_starosta, 'Phone': phone, 'Gender': gender, 'YearBirth': year\_birth},index=index)

print(group)

# 2 Первые три записи

print(group.head(3))

# 3 столбик с именами

print(group['Name'])

# 4 3 столбика вместе — имя и пол

print(group[['Name','Gender']])

# 5 Данные по старосте

print(group.loc['starosta'])

# 6 FIO в качестве индексов FIO

new\_index = ['B.I.', 'B.D.', 'B.N.',

'V.Y.', 'G.D.', 'D.S.',

'D.A.', 'Z.M.', 'K.V.',

'K.I.', 'K.V.', 'K.N.',

'K.N.', 'T.M.', 'K.S.']

new\_group = pd.DataFrame({'Name': members, 'isStarosta': is\_starosta, 'Phone': phone, 'Gender': gender, 'YearBirth': year\_birth},index=new\_index)

print(new\_group)

# 7 + столбик Age

age = [2024-int(i) for i in year\_birth]

new\_group['Age'] = age

print(new\_group)

# 8 Mинимальный и максимальный возраст

print("Максимальный возраст: " + str(new\_group['Age'].max()))

print("Минимальный возраст: " + str(new\_group['Age'].min()))

# 9 Cводная таблица с количеством девушек, юношей и средним возрастом девушек и юношей

print("\nКоличество девушек и юношей:" + str(new\_group['Gender'].value\_counts()))

print("\nСредний возраст девушек и юношей:" + str(new\_group.groupby('Gender')['Age'].mean()))

# 10 Pаспределение студентов по годам рождения

print("Распределение студентов по годам рождения:\n" + str(new\_group['YearBirth'].value\_counts().sort\_index()))

# 11 DataFrame в виде csv-файла

new\_group.to\_csv('iti-21.csv')

# 12 Новый DataFrame, заполненный данными из созданного ранее csv-файла.

new\_data\_frame = pd.read\_csv("iti-21.csv")

print(new\_data\_frame.head())

**5.2lab.py**:

import pandas as pd

# Задание 1(загрузка таблицы и удаление пропущенных значений)

polit = pd.read\_csv("polit.csv", sep=';')

polit.dropna()

print(polit)

numeric\_columns = ['fh09', 'polity09', 'gini', 'fparl08', 'corr0509']

for col in numeric\_columns:

polit[col] = polit[col].str.replace(',', '.').astype(float)

numeric\_columns = ['fh09', 'polity09', 'gini', 'fparl08', 'mena', 'lati', 'cari', 'east', 'sovi', 'afri', 'corr0509']

for col in numeric\_columns:

polit[col] = pd.to\_numeric(polit[col], errors='coerce')

#Задание 2(Freedom House (fh09) выше 5)

not\_free = polit[polit['fh09'] > 5]

print("\nСтраны с fh09 > 5:\n", not\_free)

#Задание 3(Выбрать в датафрейме строки, которые соответствуют странам Африки (afri) с процентом женщин в парламенте (fparl08) выше 30%. Сохранить их в датафрейм af\_w.

af\_w = polit[(polit['afri']==1.0) & (polit['fparl08']>30)]

print("\nСтраны Африки с долей женщин в парламенте > 30%:\n", af\_w)

#Задание 4(Выбрать в датафрейме строки, которые соответствуют странам Африки или Латинской Америки (afri и lati) со значением индекса Polity2 (polity09) больше или равным 8. Сохраните их в датафрейм la\_dem.

la\_dem = polit[((polit['afri']==1)|(polit['lati']==1)) & (polit['polity09']>=8)]

print("\nСтраны Африки или Латинской Америки с polity09 >= 8:\n", la\_dem)

#Задание 5(Добавить в датафрейм polit столбец corr\_round, в котором будут храниться округленные до 2 знака после запятой значения индекса Control of Corruption (corr0509).)

polit['corr\_round'] = polit['corr0509'].round(2)

print(polit[['ctry', 'corr0509', 'corr\_round']])

#Задание 6(Добавьте в датафрейм polit столбец fh\_status, в котором будут храниться типы стран в зависимости от значения индекса Freedom House (значения free, partly free, not free).)

def get\_country\_type(value):

if 1.0 <= value <= 2.5:

return 'Free'

elif 3.0 <= value <= 5.0:

return 'Partly Free'

elif 5.5 <= value <= 7.0:

return 'Not Free'

else:

return 'Unknown'

polit['fh\_status'] = polit['fh09'].apply(get\_country\_type)

print(polit)

#Задание 7(Сгруппировать строки в таблице в соответствии со значениями столбца fh\_status, полученного в предыдущей части и выведите минимальное, среднее и максимальное значение показателя gini (индекс Джини) по каждой группе.)

gini\_stats = polit.groupby('fh\_status')['gini'].agg(['min', 'mean', 'max'])

print("\nМинимальное, среднее и максимальное значение gini по fh\_status:\n", gini\_stats)

#Задание 8(Сгруппируйте строки в таблице в соответствии со значениями столбца fh\_status и запишите строки, относящиеся к разным группам, в отдельные csv-файлы.)

for status, group in polit.groupby('fh\_status'):

filename = f'polit\_{status.replace(" ", "\_").lower()}.csv'

group.to\_csv(filename, index=False)

print(f"\nДанные для статуса {status} сохранены в файл {filename}")