МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО”
Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем

Лабораторна робота №3

з дисципліни: «Технології оброблення великих даних»

на тему: «Кореляційний аналіз у Python»

Виконав

студент III курсу каф. ПЗКС ФПМ

групи КП-82

Мельничук Олексій Геннадійович

Перевірила

доц. каф. ПЗКС ФПМ

Олещенко Л.М.

Київ 2021

1. Індивідуальне завдання

Мета: продемонструвати свої навички кореляційного аналізу даних, використовуючи заданий набір даних та вказані інструменти

У цій лабораторній роботі ви дізнаєтесь, як використовувати Python для обчислення кореляції. У частині 1 ви налаштуєте набір даних. У частині 2 ви дізнаєтеся, як визначити, чи змінні в даному наборі даних є корельованими. У частині 3 ви будете використовувати Python для обчислення кореляції між двома наборами змінних. Нарешті, у частині 4 ви здійсните візуалізацію результатів дослідження.

2. Хід роботи

brainsize.txt

```
Gender,FSIQ,VIQ,PIQ,Weight,Height,MRI_Count
Female,133,132,124,118,64.5,816932
Male,140,150,124,NaN,72.5,1001121
Male,139,123,150,143,73.3,1038437
Male,133,129,128,172,68.8,965353
Female,137,132,134,147,65.0,951545
Female,99,90,110,146,69.0,928799
Female,138,136,131,138,64.5,991305
Female,92,90,98,175,66.0,854258
Male,89,93,84,134,66.3,904858
Male,133,114,147,172,68.8,955466
Female,132,129,124,118,64.5,833868
Male,141,150,128,151,70.0,1079549
Male,135,129,124,155,69.0,924059
Female,140,120,147,155,70.5,856472
Female,96,100,90,146,66.0,878897
Female,83,71,96,135,68.0,865363
Female,132,132,120,127,68.5,852244
Male,100,96,102,178,73.5,945088
Female,101,112,84,136,66.3,808020
Male,80,77,86,180,70.0,889083
Male,83,83,86,NaN,NaN,892420
Male,97,107,84,186,76.5,905940
Female,135,129,134,122,62.0,790619
Male,139,145,128,132,68.0,955003
Female,91,86,102,114,63.0,831772
Male,141,145,131,171,72.0,935494
Female,85,90,84,140,68.0,798612
Male,103,96,110,187,77.0,1062462
Female,77,83,72,106,63.0,793549
Female,130,126,124,159,66.5,866662
Female,133,126,132,127,62.5,857782
Male,144,145,137,191,67.0,949589
Male,103,96,110,192,75.5,997925
Male,90,96,86,181,69.0,879987
Female,83,90,81,143,66.5,834344
Female,133,129,128,153,66.5,948066
Male,140,150,124,144,70.5,949395
Female,88,86,94,139,64.5,893983
Male,81,90,74,148,74.0,930016
Male,89,91,89,179,75.5,935863
```

Частина 1: Набір даних

Крок 2: Перевірка дата фрейму.

```
#PART1

#step1
brainFile = './brainsize.txt'
brainFrame = pd.read_csv(brainFile)

#step2
print(brainFrame.head())
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  1: wsl
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$ python3 main.py
   Gender  FSIQ  VIQ  PIQ  Weight  Height  MRI_Count
0  Female   133  132  124   118.0    64.5     816932
1    Male   140  150  124     NaN    72.5    1001121
2    Male   139  123  150   143.0    73.3    1038437
3    Male   133  129  128   172.0    68.8     965353
4  Female   137  132  134   147.0    65.0     951545
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$
```

Частина 2: Діаграми розсіювання та корельовані змінні

Крок 1: Метод describe().

```
#PART 2

#step 1
print(brainFrame.describe())
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  1: wsl
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$ python3 main.py
              FSIQ         VIQ         PIQ      Weight     Height     MRI_Count
count    40.000000   40.000000   40.00000   38.000000  39.000000  4.000000e+01
mean    113.450000  112.350000  111.02500  151.052632  68.525641  9.087550e+05
std      24.082071   23.616107   22.47105   23.478509   3.994649  7.228205e+04
min      77.000000   71.000000   72.00000  106.000000  62.000000  7.906190e+05
25%      89.750000   90.000000   88.25000  135.250000  66.000000  8.559185e+05
50%     116.500000  113.000000  115.00000  146.500000  68.000000  9.053990e+05
75%     135.500000  129.750000  128.00000  172.000000  70.500000  9.500780e+05
max     144.000000  150.000000  150.00000  192.000000  77.000000  1.079549e+06
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$
```

Крок 2: Графіки діаграм розсіювання

b. Відокремте дані

```
# #step2 b
menDf = brainFrame[(brainFrame.Gender == 'Male')]
womenDf = brainFrame[(brainFrame.Gender == 'Female')]
print(menDf)
print(womenDf)
```

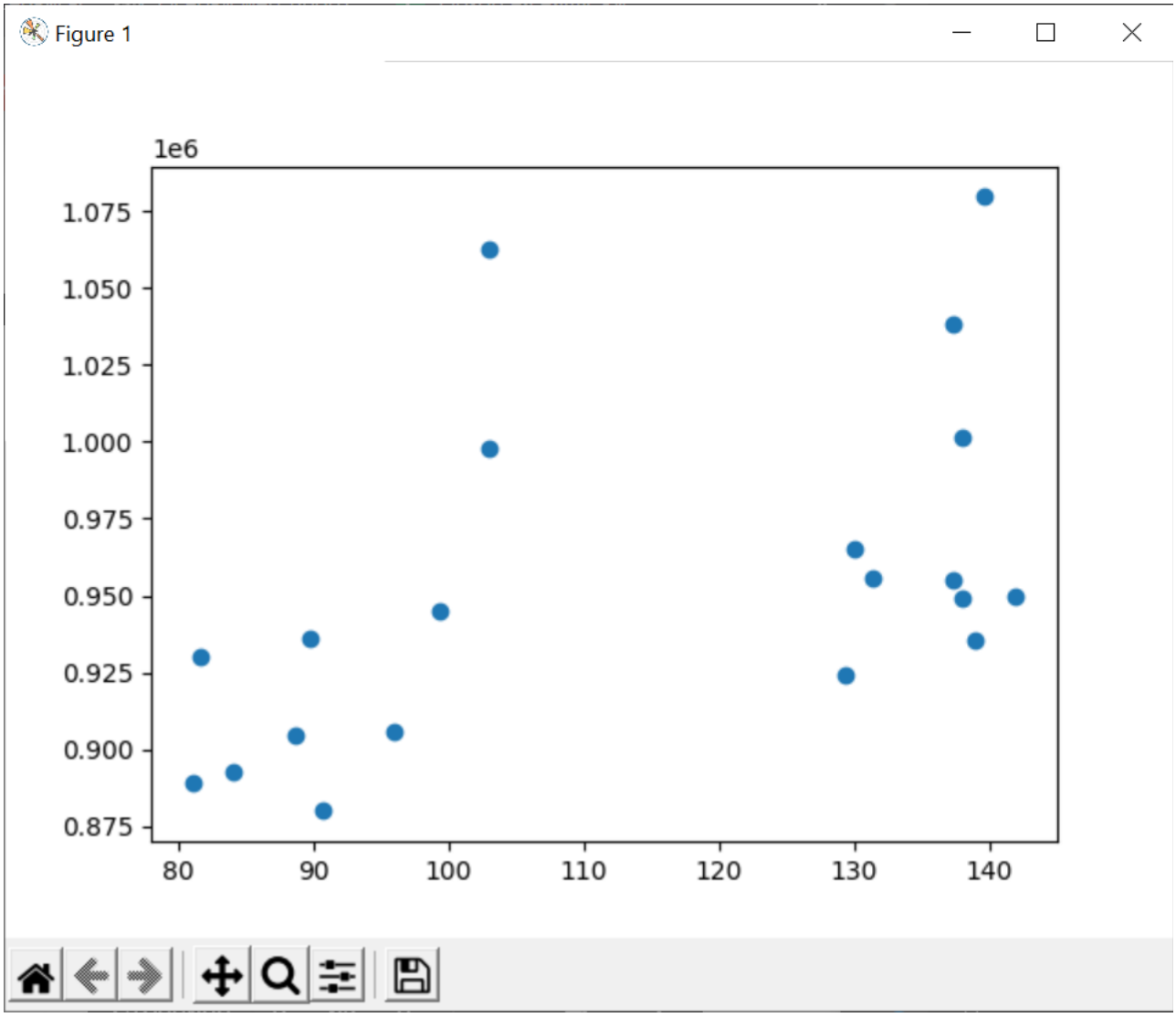
```
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$ python3 main.py
   Gender  FSIQ  VIQ  PIQ  Weight  Height  MRI_Count
1    Male   140  150  124     NaN    72.5    1001121
2    Male   139  123  150   143.0    73.3    1038437
3    Male   133  129  128   172.0    68.8     965353
8    Male    89   93   84   134.0    66.3     904858
9    Male   133  114  147   172.0    68.8     955466
11   Male   141  150  128   151.0    70.0    1079549
12   Male   135  129  124   155.0    69.0     924059
17   Male   100   96  102   178.0    73.5     945088
19   Male    80   77   86   180.0    70.0     889083
20   Male    83   83   86     NaN     NaN     892420
21   Male    97  107   84   186.0    76.5     905940
23   Male   139  145  128   132.0    68.0     955003
25   Male   141  145  131   171.0    72.0     935494
27   Male   103   96  110   187.0    77.0    1062462
31   Male   144  145  137   191.0    67.0     949589
32   Male   103   96  110   192.0    75.5     997925
33   Male    90   96   86   181.0    69.0     879987
36   Male   140  150  124   144.0    70.5     949395
38   Male    81   90   74   148.0    74.0     930016
39   Male    89   91   89   179.0    75.5     935863
```

```
    Gender  FSIQ  VIQ  PIQ  Weight  Height  MRI_Count
0   Female   133  132  124   118.0    64.5     816932
4   Female   137  132  134   147.0    65.0     951545
5   Female    99   90  110   146.0    69.0     928799
6   Female   138  136  131   138.0    64.5     991305
7   Female    92   90   98   175.0    66.0     854258
10  Female   132  129  124   118.0    64.5     833868
13  Female   140  120  147   155.0    70.5     856472
14  Female    96  100   90   146.0    66.0     878897
15  Female    83   71   96   135.0    68.0     865363
16  Female   132  132  120   127.0    68.5     852244
18  Female   101  112   84   136.0    66.3     808020
22  Female   135  129  134   122.0    62.0     790619
24  Female    91   86  102   114.0    63.0     831772
26  Female    85   90   84   140.0    68.0     798612
28  Female    77   83   72   106.0    63.0     793549
29  Female   130  126  124   159.0    66.5     866662
30  Female   133  126  132   127.0    62.5     857782
34  Female    83   90   81   143.0    66.5     834344
35  Female   133  129  128   153.0    66.5     948066
37  Female    88   86   94   139.0    64.5     893983
auvy@DESKTOP-8C29PGJ:/mnt/c/Users/datru/Desktop/study2021/big_data/lab3$
```

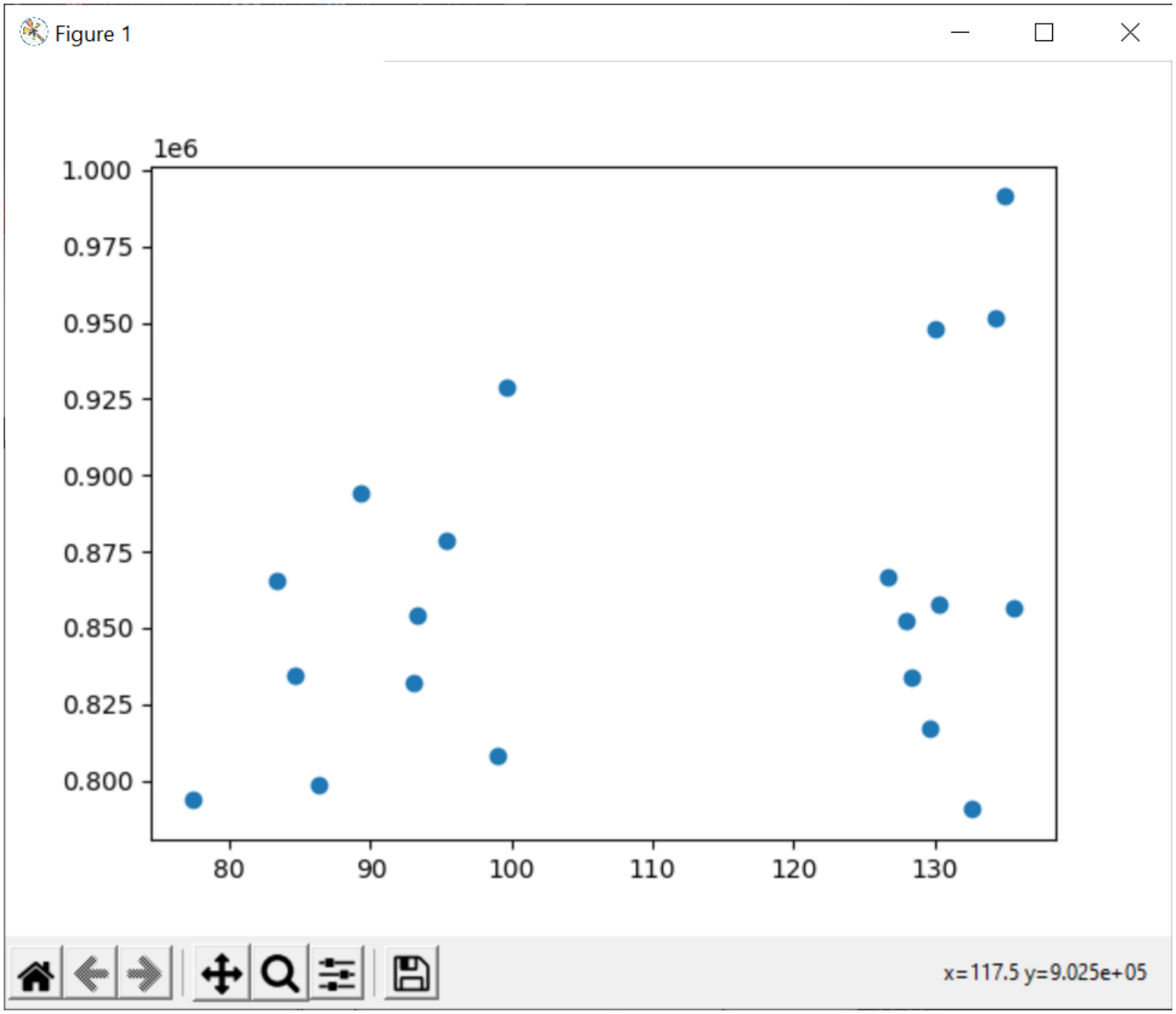
c. Побудуйте графіки.

```
#step2 c
menMeanSmarts = menDf[["PIQ", "FSIQ", "VIQ"]].mean(axis=1)
plt.scatter(menMeanSmarts, menDf["MRI_Count"])

# womenMeanSmarts = womenDf[["PIQ", "FSIQ", "VIQ"]].mean(axis=1)
# plt.scatter(womenMeanSmarts, womenDf["MRI_Count"])

plt.show()
%matplotlib inline
```

menMeanSmarts



womenMeanSmarts



Частина 3: Обчислення кореляції в Python

Крок 1: Обчисліть співвідношення між brainFrame.

```python
# #PART 3
# #step1

print(brainFrame.corr(method='pearson'))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab3>python3 main.py
              FSIQ       VIQ       PIQ    Weight    Height  MRI_Count
FSIQ      1.000000  0.946639  0.934125 -0.051483 -0.086002   0.357641
VIQ       0.946639  1.000000  0.778135 -0.076088 -0.071068   0.337478
PIQ       0.934125  0.778135  1.000000  0.002512 -0.076723   0.386817
Weight   -0.051483 -0.076088  0.002512  1.000000  0.699614   0.513378
Height   -0.086002 -0.071068 -0.076723  0.699614  1.000000   0.601712
MRI_Count  0.357641  0.337478  0.386817  0.513378  0.601712   1.000000

C:\Users\datru\Desktop\study2021\big_data\lab3>
```

Значення 1 по діагоналі тому що кожне значення залежить від себе.

Таблиця «зеркальна» тому що має однакові стовпці та рядки, тому значення повторюються.

```python
# #PART 3
# #step1

# print(brainFrame.corr(method='pearson'))

print('women')
print(womenDf.corr(method='pearson'))
print('men')
print(menDf.corr(method='pearson'))
```

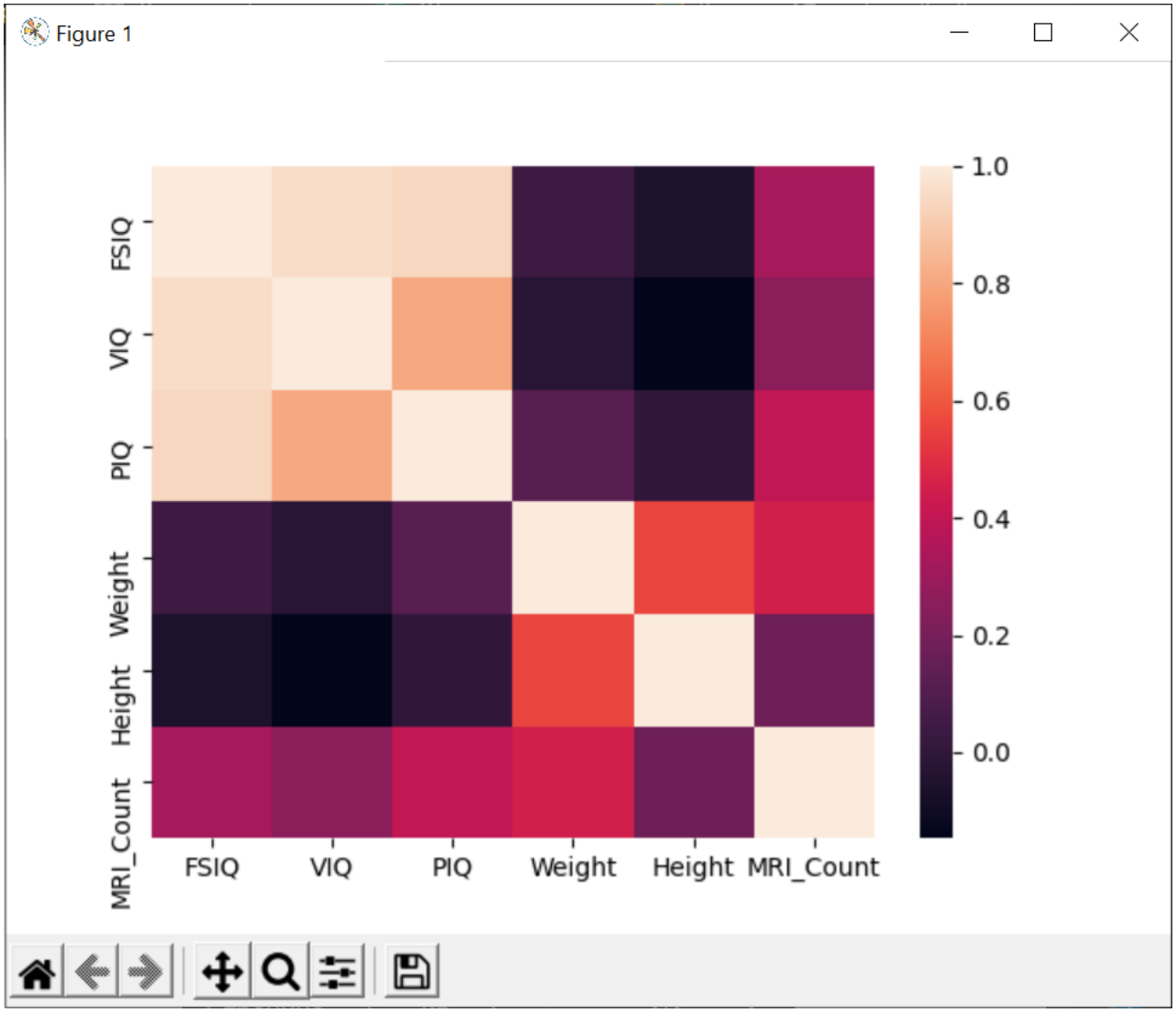
```
C:\Users\datru\Desktop\study2021\big_data\lab3>python3 main.py
women
              FSIQ       VIQ       PIQ    Weight    Height  MRI_Count
FSIQ      1.000000  0.955717  0.939382  0.038192 -0.059011   0.325697
VIQ       0.955717  1.000000  0.802652 -0.021889 -0.146453   0.254933
PIQ       0.939382  0.802652  1.000000  0.113901 -0.001242   0.396157
Weight    0.038192 -0.021889  0.113901  1.000000  0.552357   0.446271
Height   -0.059011 -0.146453 -0.001242  0.552357  1.000000   0.174541
MRI_Count  0.325697  0.254933  0.396157  0.446271  0.174541   1.000000
men
              FSIQ       VIQ       PIQ    Weight    Height  MRI_Count
FSIQ      1.000000  0.944400  0.930694 -0.278140 -0.356110   0.498369
VIQ       0.944400  1.000000  0.766021 -0.350453 -0.355588   0.413105
PIQ       0.930694  0.766021  1.000000 -0.156863 -0.287676   0.568237
Weight   -0.278140 -0.350453 -0.156863  1.000000  0.406542  -0.076875
Height   -0.356110 -0.355588 -0.287676  0.406542  1.000000   0.301543
MRI_Count  0.498369  0.413105  0.568237 -0.076875  0.301543   1.000000

C:\Users\datru\Desktop\study2021\big_data\lab3>
```

Частина 4: Візуалізація

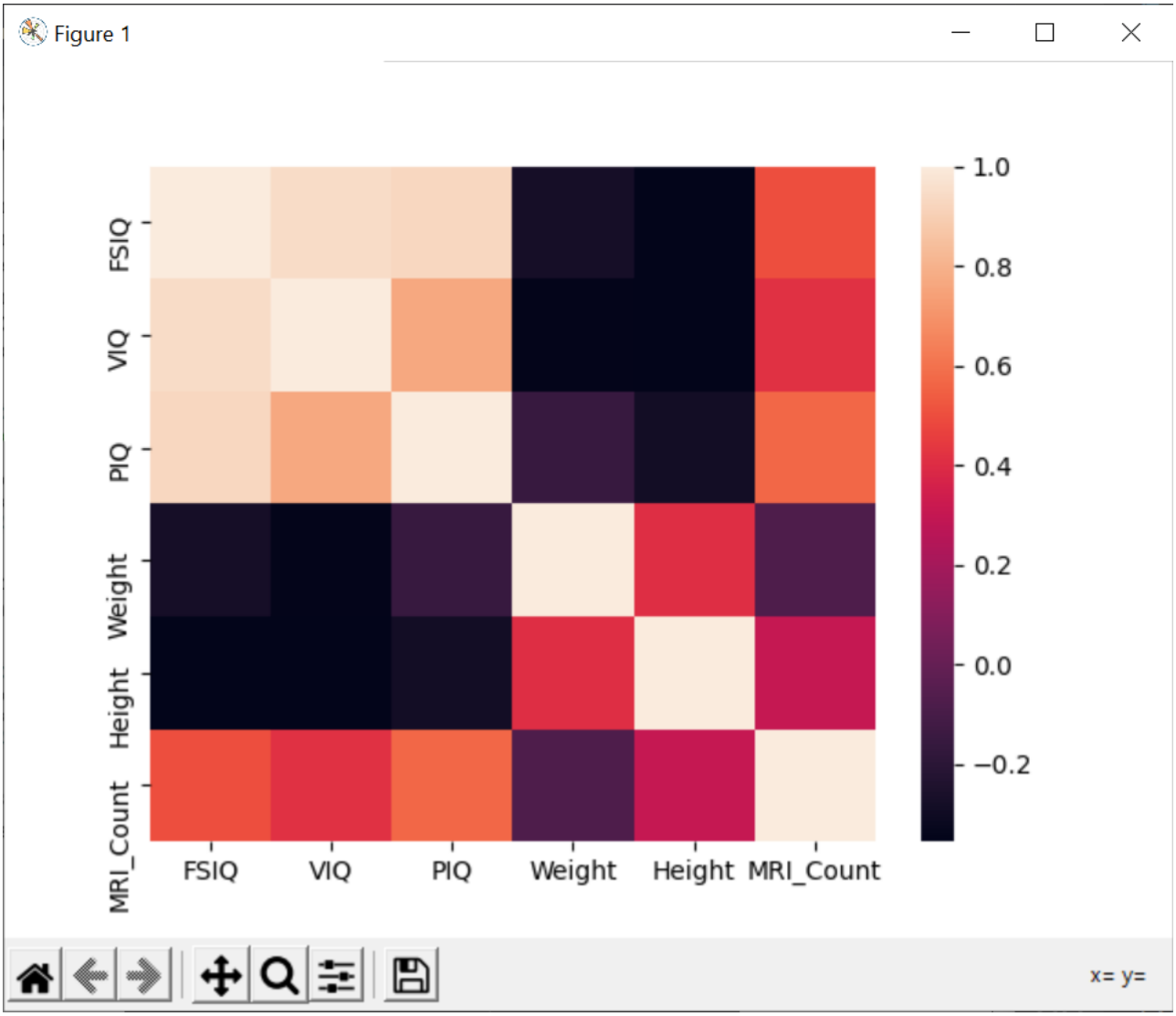
Крок 2: Побудуйте графік кореляційної теплової карти.

```
# #PART 4
# #step 2
wcorr = womenDf.corr()
sns.heatmap(wcorr)
plt.show()
```

Women correlation map



Men correlation map



Кореляція що близька до нуля означає що змінні не є залежними одна від одної. Розділ на стать можливо була зроблена для полегшення роботи з великим об'ємом даних або для додаткових уточнюючих досліджень. Зазвичай окрім статі також розділяють по віку та етнічній принадлежності.

Сильнішу кореляцію мають PIQ (performance iq) та FSIQ (full scale IQ), можливо трохи VIQ (verbal iq). Можливо це очікується, однак кореляція не достатньо сильна щоб твердо це стверджувати.

Висновки

В ході виконанні лабораторної роботи були використані навички обробки набору даних з мозгових сканів 40 добровольців з допомогою бібліотек для мови Python, а саме Numpy, Pandas, Matplotlib та Seaborn. Ці бібліотеки були використані для зручної побудови графів та теплових карт кореляції даних з набору.

PIQ, VIQ та FSIQ корелюють найсильніше з MRI count.