

**Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра обчислювальної техніки**

**Лабораторна робота №3
з дисципліни
«Аналіз даних з використанням мови Python»**

Виконав:
студент групи ПІ-04
Пащенко Дмитро Олексійович

Перевірила:
Тимофєєва Ю. С.

Київ 2022

Код програми та результат виконання

```
In [133]: import numpy as np
import pandas as pd
```

```
In [134]: data = pd.read_csv('Birthweight.csv')
data
```

```
Out[134]:
```

	ID	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35
0	1360	56	4.55	34	44	0	20	0	162	57	23	10	35	179	0	0
1	1016	53	4.32	36	40	0	19	0	171	62	19	12	0	183	0	0
2	462	58	4.10	39	41	0	35	0	172	58	31	16	25	185	0	1
3	1187	53	4.07	38	44	0	20	0	174	68	26	14	25	189	0	0
4	553	54	3.94	37	42	0	24	0	175	66	30	12	0	184	0	0
5	1636	51	3.93	38	38	0	29	0	165	61	31	16	0	180	0	0
6	820	52	3.77	34	40	0	24	0	157	50	31	16	0	173	0	0
7	1191	53	3.65	33	42	0	21	0	165	61	21	10	25	185	0	0
8	1081	54	3.63	38	38	0	18	0	172	50	20	12	7	172	0	0
9	822	50	3.42	35	38	0	20	0	157	48	22	14	0	179	0	0
10	1683	53	3.35	33	41	0	27	0	164	62	37	14	0	170	0	0

Створити програму, яка за даними файлу відповідно до варіантів лабораторної No2, виконує наступні завдання:

1. Виділити один зі стовпців (на вибір) з файлу як об'єкт Series, виділити з нього підмасив. Задати назви індексів цього об'єкту. Виділити підмасиви за допомогою прямої та непрямої індексацій.

```
In [135]: birthweight = pd.Series(data['Birthweight'], index = [15, 16, 17, 18, 0, 1, 2, 3])
birthweight
```

```
Out[135]: 15    3.11
16    3.03
17    2.92
18    2.90
0     4.55
1     4.32
2     4.10
3     4.07
Name: Birthweight, dtype: float64
```

```
In [136]: birthweight.loc[15:17]
```

```
Out[136]: 15    3.11
16    3.03
17    2.92
Name: Birthweight, dtype: float64
```

```
In [137]: birthweight.iloc[0:3]
```

```
Out[137]: 15    3.11
16    3.03
17    2.92
Name: Birthweight, dtype: float64
```

2. До об'єкту DataFrame, в який записано вміст файлу, додати новий стовпець, що є результатом операцій над іншими стовпцями. Також продемонструвати додавання та видалення рядків, видалення стовпців.

```
In [138]: data
```

```
Out[138]:
```

	ID	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35
0	1360	56	4.55	34	44	0	20	0	162	57	23	10	35	179	0	0
1	1016	53	4.32	36	40	0	19	0	171	62	19	12	0	183	0	0
2	462	58	4.10	39	41	0	35	0	172	58	31	16	25	185	0	1
3	1187	53	4.07	38	44	0	20	0	174	68	26	14	25	189	0	0
4	553	54	3.94	37	42	0	24	0	175	66	30	12	0	184	0	0
5	1636	51	3.93	38	38	0	29	0	165	61	31	16	0	180	0	0
6	820	52	3.77	34	40	0	24	0	157	50	31	16	0	173	0	0
7	1191	53	3.65	33	42	0	21	0	165	61	21	10	25	185	0	0
8	1081	54	3.63	38	38	0	18	0	172	50	20	12	7	172	0	0
9	822	50	3.42	35	38	0	20	0	157	48	22	14	0	179	0	0
10	1683	53	3.35	33	41	0	27	0	164	62	37	14	0	170	0	0

```
In [139]: data['ageDiff'] = data['fage'] - (data['mage'])
data
```

```
Out[139]:
```

	ID	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35	ageDiff
0	1360	56	4.55	34	44	0	20	0	162	57	23	10	35	179	0	0	3
1	1016	53	4.32	36	40	0	19	0	171	62	19	12	0	183	0	0	0
2	462	58	4.10	39	41	0	35	0	172	58	31	16	25	185	0	1	-4
3	1187	53	4.07	38	44	0	20	0	174	68	26	14	25	189	0	0	6
4	553	54	3.94	37	42	0	24	0	175	66	30	12	0	184	0	0	6
5	1636	51	3.93	38	38	0	29	0	165	61	31	16	0	180	0	0	2
6	820	52	3.77	34	40	0	24	0	157	50	31	16	0	173	0	0	7
7	1191	53	3.65	33	42	0	21	0	165	61	21	10	25	185	0	0	0
8	1081	54	3.63	38	38	0	18	0	172	50	20	12	7	172	0	0	2
9	822	50	3.42	35	38	0	20	0	157	48	22	14	0	179	0	0	2
10	1683	53	3.35	33	41	0	27	0	164	62	37	14	0	170	0	0	10

```
In [140]: data.drop('Gestation', inplace = True, axis = 1)
data
```

```
Out[140]:
```

	ID	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35	ageDiff
0	1360	56	4.55	34	0	20	0	162	57	23	10	35	179	0	0	3
1	1016	53	4.32	36	0	19	0	171	62	19	12	0	183	0	0	0
2	462	58	4.10	39	0	35	0	172	58	31	16	25	185	0	1	-4
3	1187	53	4.07	38	0	20	0	174	68	26	14	25	189	0	0	6
4	553	54	3.94	37	0	24	0	175	66	30	12	0	184	0	0	6
5	1636	51	3.93	38	0	29	0	165	61	31	16	0	180	0	0	2
6	820	52	3.77	34	0	24	0	157	50	31	16	0	173	0	0	7
7	1191	53	3.65	33	0	21	0	165	61	21	10	25	185	0	0	0
8	1081	54	3.63	38	0	18	0	172	50	20	12	7	172	0	0	2
9	822	50	3.42	35	0	20	0	157	48	22	14	0	179	0	0	2
10	1683	53	3.35	33	0	27	0	164	62	37	14	0	170	0	0	10

```
In [141]: data.drop(9, inplace = True, axis = 0)
data
```

```
Out[141]:
```

	ID	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35	ageDiff
0	1360	56	4.55	34	0	20	0	162	57	23	10	35	179	0	0	3
1	1016	53	4.32	36	0	19	0	171	62	19	12	0	183	0	0	0
2	462	58	4.10	39	0	35	0	172	58	31	16	25	185	0	1	-4
3	1187	53	4.07	38	0	20	0	174	68	26	14	25	189	0	0	6
4	553	54	3.94	37	0	24	0	175	66	30	12	0	184	0	0	6
5	1636	51	3.93	38	0	29	0	165	61	31	16	0	180	0	0	2
6	820	52	3.77	34	0	24	0	157	50	31	16	0	173	0	0	7
7	1191	53	3.65	33	0	21	0	165	61	21	10	25	185	0	0	0
8	1081	54	3.63	38	0	18	0	172	50	20	12	7	172	0	0	2
10	1683	53	3.35	33	0	27	0	164	62	37	14	0	170	0	0	10
11	1088	51	3.27	36	0	24	0	168	53	29	16	0	181	0	0	5

```
In [144]: data.loc[len(data.index)] = [1111, 54, 4.00, 32, 0, 22, 0, 170, 53, 33, 10, 3, 180, 0, 0, 1]
data
```

30	1023	52	3.00	35	1	30	12	165	64	38	14	50	180	0	0	8
31	57	51	3.32	38	1	23	17	157	48	32	12	25	169	0	0	9
32	1522	50	2.74	33	1	21	17	156	53	24	12	7	179	0	0	3
33	223	50	3.87	33	1	28	25	163	54	30	16	0	183	0	0	2
34	272	52	3.86	36	1	30	25	170	78	40	16	50	178	0	0	10
35	27	53	3.55	37	1	37	25	161	66	46	16	0	175	0	1	9
36	365	52	3.53	37	1	26	25	170	62	30	10	25	181	0	0	4
37	619	52	3.41	33	1	23	25	181	69	23	16	2	181	0	0	0
38	1369	49	3.18	34	1	31	25	162	57	32	16	50	194	0	0	1
39	1262	53	3.19	34	1	27	35	163	51	31	16	25	185	0	0	4
40	516	47	2.66	33	1	20	35	170	57	23	12	50	186	1	0	3
41	1111	54	4.00	32	0	22	0	170	53	33	10	3	180	0	0	1

```
In [159]: data_two = data
data_three = pd.concat([data, data_two], ignore_index = True)
data_three
```

	ID	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35	ageDiff
0	1360	56	4.55	34	0	20	0	162	57	23	10	35	179	0	0	3
1	1016	53	4.32	36	0	19	0	171	62	19	12	0	183	0	0	0
2	462	58	4.10	39	0	35	0	172	58	31	16	25	185	0	1	-4
3	1187	53	4.07	38	0	20	0	174	68	26	14	25	189	0	0	6
4	553	54	3.94	37	0	24	0	175	66	30	12	0	184	0	0	6
...
77	619	52	3.41	33	1	23	25	181	69	23	16	2	181	0	0	0
78	1369	49	3.18	34	1	31	25	162	57	32	16	50	194	0	0	1
79	1262	53	3.19	34	1	27	35	163	51	31	16	25	185	0	0	4
80	516	47	2.66	33	1	20	35	170	57	23	12	50	186	1	0	3
81	1111	54	4.00	32	0	22	0	170	53	33	10	3	180	0	0	1

3. Встановити один зі стовпців індексом. Визначити основні статистичні характеристики та типи даних всіх стовпців. Змінити тип даних для одного з стовпців. Згрупувати дані за одним зі стовпців, застосувати кілька агрегуючих функцій, виділити підмасив за певними ознаками.

```
In [160]: data_three = data_three.set_index(['ID'])
data_three
```

Out[160]:

	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	mage35	ageDiff
ID															
1360	56	4.55	34	0	20	0	162	57	23	10	35	179	0	0	3
1016	53	4.32	36	0	19	0	171	62	19	12	0	183	0	0	0
462	58	4.10	39	0	35	0	172	58	31	16	25	185	0	1	-4
1187	53	4.07	38	0	20	0	174	68	26	14	25	189	0	0	6
553	54	3.94	37	0	24	0	175	66	30	12	0	184	0	0	6
...
619	52	3.41	33	1	23	25	181	69	23	16	2	181	0	0	0
1369	49	3.18	34	1	31	25	162	57	32	16	50	194	0	0	1
1262	53	3.19	34	1	27	35	163	51	31	16	25	185	0	0	4
516	47	2.66	33	1	20	35	170	57	23	12	50	186	1	0	3

```
In [161]: data_three.describe()
```

Out[161]:

	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedys	fnocig	fheight	lowbwt	
count	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000
mean	51.390244	3.340732	34.585366	0.512195	25.317073	8.439024	164.682927	57.536585	29.121951	13.512195	17.682927	180.707317	0.146341	
std	2.963798	0.609883	2.413702	0.502927	5.363061	10.809344	6.469068	7.101026	6.839230	2.212445	17.129051	6.914804	0.355623	
min	43.000000	1.920000	30.000000	0.000000	18.000000	0.000000	149.000000	45.000000	19.000000	10.000000	0.000000	169.000000	0.000000	
25%	50.000000	3.000000	33.000000	0.000000	21.000000	0.000000	161.000000	53.000000	23.000000	12.000000	0.000000	176.000000	0.000000	
50%	52.000000	3.320000	34.000000	1.000000	24.000000	2.000000	165.000000	57.000000	30.000000	14.000000	25.000000	181.000000	0.000000	
75%	53.000000	3.770000	36.000000	1.000000	29.000000	12.000000	170.000000	62.000000	33.000000	16.000000	25.000000	185.000000	0.000000	
max	58.000000	4.570000	39.000000	1.000000	41.000000	35.000000	181.000000	78.000000	46.000000	16.000000	50.000000	200.000000	1.000000	

```
In [162]: data_three.dtypes
```

Out[162]:

```
Length          int64
Birthweight     float64
Headcirc        int64
smoker          int64
mage            int64
mnocig          int64
mheight         int64
mppwt           int64
fage            int64
fedys           int64
fnocig          int64
fheight         int64
lowbwt          int64
mage35          int64
ageDiff         int64
dtype: object
```

```
In [165]: data_three.astype({'smoker': 'bool'})
```

```
Out[165]:
```

	Length	Birthweight	Headcirc	smoker	mage	mnocig	mheight	mppwt	fage	fedysr	fnocig	fheight	lowbwt	mage35	ageDiff
ID															
1380	56	4.55	34	False	20	0	162	57	23	10	35	179	0	0	3
1016	53	4.32	36	False	19	0	171	62	19	12	0	183	0	0	0
462	58	4.10	39	False	35	0	172	58	31	16	25	185	0	1	-4
1187	53	4.07	38	False	20	0	174	68	26	14	25	189	0	0	6
553	54	3.94	37	False	24	0	175	66	30	12	0	184	0	0	6
...
619	52	3.41	33	True	23	25	181	69	23	16	2	181	0	0	0
1369	49	3.18	34	True	31	25	162	57	32	16	50	194	0	0	1
1262	53	3.19	34	True	27	35	163	51	31	16	25	185	0	0	4
516	47	2.66	33	True	20	35	170	57	23	12	50	186	1	0	3

```
In [166]: data_three.astype({'smoker': 'bool'}).dtypes
```

```
Out[166]: Length          int64
Birthweight      float64
Headcirc         int64
smoker           bool
mage             int64
mnocig           int64
mheight          int64
mppwt            int64
fage             int64
fedysr           int64
fnocig           int64
fheight          int64
lowbwt           int64
mage35           int64
ageDiff          int64
dtype: object
```

```
In [189]: data_three.groupby('smoker').agg([sum, np.min, np.max])
```

```
Out[189]:
```

smoker	Length			Birthweight			Headcirc			mage			fheight			lowbwt			mage35			ageDiff		
	sum	amin	amax	sum	amin	amax	sum	amin	amax	sum	amin	amax	sum	amin	amax	sum	amin	amax	sum	amin	amax			
0	2080	43	58	141.54	2.65	4.55	1396	32	39	976	...	193	2	0	1	2	0	1	126	-4	11			
1	2134	46	58	132.40	1.92	4.57	1440	30	39	1100	...	200	10	0	1	4	0	1	166	-4	10			

2 rows x 42 columns

4. Створити декілька власних об'єктів DataFrame за такою ж тематикою, що й файл. Наприклад, якщо тема файлу – жаб, можна створити об'єкти, що містять розміри жаб, вагу, стать, кількість особин в популяції і т.д. Використати описані в теоретичних відомостях параметри методів merge та concat для різних видів злиття та об'єднання даних цих об'єктів.

```
In [200]: # Один-до-одного
a = pd.DataFrame({'mage': [20, 19, 32, 25], 'Birthweight': [2.4, 4.2, 3.3, 2.7]})
b = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
c = pd.merge(a, b)
```

```
Out[200]:
```

	mage	Birthweight	smoker
0	20	2.4	1
1	19	4.2	0
2	32	3.3	1
3	25	2.7	1

```
In [197]: # Один-до-одного, але назва стовпчика інша
a = pd.DataFrame({'mage': [20, 19, 32, 25], 'Birthweight': [2.4, 4.2, 3.3, 2.7]})
b = pd.DataFrame({'mother_age': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
c = pd.merge(a, b, left_on="mage", right_on="mother_age").drop('mother_age', axis = 1)
```

```
Out[197]:
```

	mage	Birthweight	smoker
0	20	2.4	1
1	19	4.2	0
2	32	3.3	1
3	25	2.7	1

```
In [201]: # Багато-до-одного
a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1], 'Birthweight': [2.4, 4.2, 3.3, 2.7]})
b = pd.DataFrame({'mage': [20, 19, 32, 25], 'fage': [24, 18, 36, 30]})
c = pd.merge(a, b)
```

```
Out[201]:
```

	mage	smoker	Birthweight	fage
0	20	1	2.4	24
1	19	0	4.2	18
2	32	1	3.3	36
3	25	1	2.7	30

```
In [202]: # Багато-до-багатох
a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1], 'Birthweight': [2.4, 4.2, 3.3, 2.7]})
b = pd.DataFrame({'mage': [20, 19, 32, 25], 'fage': [24, 18, 36, 30], 'Length': [48, 50, 53, 49]})
c = pd.merge(a, b)
```

```
Out[202]:
```

	mage	smoker	Birthweight	fage	Length
0	20	1	2.4	24	48
1	19	0	4.2	18	50
2	32	1	3.3	36	53
3	25	1	2.7	30	49

```
In [203]: a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
b = pd.DataFrame({'mage': [21, 19, 33, 26], 'fage': [24, 18, 36, 30]})
c = pd.merge(a, b)
```

```
Out[203]:
```

	mage	smoker	fage
0	19	0	18

```
In [204]: a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
b = pd.DataFrame({'mage': [21, 19, 33, 26], 'fage': [24, 18, 36, 30]})
c = pd.merge(a, b, how = 'outer')
```

```
Out[204]:
```

	mage	smoker	fage
0	20	1.0	NaN
1	19	0.0	18.0
2	32	1.0	NaN
3	25	1.0	NaN
4	21	NaN	24.0
5	33	NaN	36.0
6	26	NaN	30.0

```
In [205]: c = pd.merge(a, b, how = 'left')
```

```
Out[205]:
```

	mage	smoker	fage
0	20	1	NaN
1	19	0	18.0
2	32	1	NaN
3	25	1	NaN

```
In [206]: c = pd.merge(a, b, how = 'right')
```

```
Out[206]:
```

	mage	smoker	fage
0	21	NaN	24
1	19	0.0	18
2	33	NaN	36
3	26	NaN	30

```
In [213]: a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
b = pd.DataFrame({'mage': [21, 19, 33, 26], 'smoker': [1, 1, 1, 0]})
c = pd.concat([a, b])
c
```

Out[213]:

	mage	smoker
0	20	1
1	19	0
2	32	1
3	25	1
0	21	1
1	19	1
2	33	1
3	26	0

```
In [214]: a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
b = pd.DataFrame({'mage': [21, 19, 33, 26], 'smoker': [1, 1, 1, 0]})
c = pd.concat([a, b], axis = 1)
c
```

Out[214]:

	mage	smoker	mage	smoker
0	20	1	21	1
1	19	0	19	1
2	32	1	33	1
3	25	1	26	0

```
In [215]: a = pd.DataFrame({'mage': [20, 19, 32, 25], 'smoker': [1, 0, 1, 1]})
b = pd.DataFrame({'mage': [21, 19, 33, 26], 'smoker': [1, 1, 1, 0]})
c = pd.concat([a, b], ignore_index = True)
c
```

Out[215]:

	mage	smoker
0	20	1
1	19	0
2	32	1
3	25	1
4	21	1
5	19	1
6	33	1
7	26	0