BÀI TẬP THỰC HÀNH

Bài 7:

LẬP TRÌNH PYTHON CƠ BẢN

(Phân tích và xử lý dữ liệu với Pandas - 02)

Thực hành 1

Thực hành 1

Yêu cầu 1.1:

- Đọc dữ liệu từ file Data_Patient.csv vào biến kiểu dataframe: df_patient với cột đầu tiên (id) là cột chỉ số (index_col). Hiển thị 10 dòng dữ liệu đầu tiên.

Yêu cầu 1.2:

- Xóa cột dữ liệu có tên 'Thalassemia' và áp dụng thay đổi lên chính df_patient.

```
<class 'pandas.core.frame.DataFrame'>
Index: 300 entries, Patient_01 to Patient_300
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             300 non-null    int64
 1   Gender          300 non-null    object
 2   Type            295 non-null    object
 3   Blood_pressure  300 non-null    int64
 4   Cholesterol     300 non-null    int64
 5   Heartbeat       300 non-null    int64
 6   Result          300 non-null    int64
```

Thực hành 1

Yêu cầu 1.3:

- A) Tạo df_patient1 bằng cách loại bỏ đi 100 dòng dữ liệu đầu tiên từ df_patient.

id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Result
Patient_101	34	Male	Typical angina	118	182	174	0
Patient_102	57	Female	Asymptomatic	128	303	159	0
Patient_103	71	Female	Non-anginal pain	110	265	130	0
Patient_104	49	Male	Non-anginal pain	120	188	139	1
Patient_105	54	Male	Atypical angina	108	309	156	0

- B) Tạo df_patient2 bằng cách loại bỏ đi các dòng dữ liệu có thuộc tính Type = 'Non-anginal pain' và nhịp tim > 187 từ df_patient.

id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Result
Patient_54	44	Male	Atypical angina	130	219	188	0
Patient_112	52	Male	Typical angina	118	186	190	0
Patient_132	29	Male	Atypical angina	130	204	202	0
Patient_186	42	Male	Non-anginal pain	120	240	194	0
Patient_188	54	Male	Atypical angina	192	283	195	1
Patient_225	34	Female	Atypical angina	118	210	192	0

Thực hành 1

Yêu cầu 1.4:

- A) Sắp xếp lại dữ liệu cho df_patient theo chiều giảm dần của index, áp dụng thay đổi trực tiếp lên DataFrame này.
- B) Tạo df_patient3 bằng cách sắp xếp dữ liệu theo thuộc tính Gender tăng dần, Nếu trùng giá trị Gender thì sắp xếp theo thuộc tính Age giảm dần.

```
1 df_patient3.iloc[90:100]
```

id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Result
Patient_277	39	Female	Non-anginal pain	138	220	152	0
Patient_222	39	Female	Non-anginal pain	94	199	179	0
Patient_210	37	Female	Non-anginal pain	120	215	170	0
Patient_117	35	Female	Asymptomatic	138	183	182	0
Patient_225	34	Female	Atypical angina	118	210	192	0
Patient_161	77	Male	Asymptomatic	125	304	162	1
Patient_258	70	Male	Atypical angina	156	245	143	0
Patient_170	70	Male	Non-anginal pain	160	269	112	1
Patient_155	70	Male	Asymptomatic	130	322	109	1
Patient_136	70	Male	Asymptomatic	145	174	125	1

Thực hành 1

Yêu cầu 1.5:

- A) Nhóm bệnh nhân theo thuộc tính Gender và tìm tuổi lớn nhất, nhỏ nhất, trung bình của bệnh nhân theo giới tính.

```
1) Thống kê tuổi cao nhất theo giới tính:
Gender
Female    76
Male      77
Name: Age, dtype: int64

2) Thống kê tuổi thấp nhất theo giới tính:
Gender
Female    34
Male      29
Name: Age, dtype: int64

3) Thống kê tuổi trung bình theo giới tính:
Gender
Female    55.736842
Male      53.912195
Name: Age, dtype: float64
```

- B) Nhóm bệnh nhân theo thuộc tính Gender và Type và tìm tuổi lớn nhất, nhỏ nhất, trung bình của bệnh nhân theo giới tính và loại đau ngực.

```
1) Thống kê tuổi cao nhất theo giới tính và loại:
Gender  Type
Female  Asymptomatic        71
        Atypical angina     74
        Non-anginal pain    76
        Typical angina      69
Male    Asymptomatic        77
        Atypical angina     70
        Non-anginal pain    70
        Typical angina      69
Name: Age, dtype: int64
```

Thực hành 1

Yêu cầu 1.6:

- Sử dụng df_patient.reset_index(inplace=True) để bỏ cột index. Sau đó thực hiện xóa các bệnh nhân có giá trị trong cột id trùng nhau, giữ lại bệnh nhân có id trùng nhau đầu tiên, áp dụng cho chính dataframe hiện tại.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 292 entries, 0 to 299
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              292 non-null    object
 1   Age             292 non-null    int64
 2   Gender          292 non-null    object
 3   Type            287 non-null    object
 4   Blood_pressure  292 non-null    int64
 5   Cholesterol     292 non-null    int64
 6   Heartbeat       292 non-null    int64
 7   Result          292 non-null    int64
dtypes: int64(5), object(3)
```

Thực hành 2

Thực hành 2

Yêu cầu 2.1:

- Đọc dữ liệu từ file Data_Point.xlsx vào biến kiểu dataframe:
  - df_lop1 dữ liệu điểm sheet 0 (4080130_01)
  - df_lop2 dữ liệu điểm sheet 1 (4080130_02)
  - df_lop3 dữ liệu điểm sheet 2 (4080130_03)

	Code	A	B1	B2	C1	C2
2	1621050322	8	0	5	7.5	8
3	1621050512	6	3	7.5	8.5	9
4	1621050211	6.7	4	6.5	3	5
5	1621050827	8	6.5	8	10	9
6	1621050298	7	5	8	8.5	9
7	1621050351	4.3	5	5	6	6
8	1621050422	7	6.5	9	10	10
9	1621050281	5.3	3.5	6	8.5	8
10	1621050753	6	5	6.5	10	10
11	1621050283	6	5.5	7	8.5	8
12	1621050122	5.3	2	6	8.5	8
13	1621050203	6	8	8	10	10
14	1621050090	6	5	6.5	10	8
15	1621050802	7	0	8	0	5
16	1621050434	6	9	7	10	9.5
17	1621050240	6	9	7	10	9.5

4080130_01 | 4080130_02 | 4080130_03 | Code

Thực hành 2

Yêu cầu 2.2:

- Nối 3 DataFrame df_lop1, df_lop2, df_lop3 thành một DataFrame df_full chứa tất cả danh sách bảng điểm của 3 lớp

```
1  df_full.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144 entries, 0 to 143
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Code    144 non-null    int64
 1   A       144 non-null    float64
 2   B1      144 non-null    float64
 3   B2      144 non-null    float64
 4   C1      144 non-null    float64
 5   C2      144 non-null    float64
dtypes: float64(5), int64(1)
memory usage: 6.9 KB
```

	Code	A	B1	B2	C1	C2
0	1621050322	8.0	0.0	5.0	7.5	8.0
1	1621050512	6.0	3.0	7.5	8.5	9.0
2	1621050211	6.7	4.0	6.5	3.0	5.0
3	1621050827	8.0	6.5	8.0	10.0	9.0
4	1621050298	7.0	5.0	8.0	8.5	9.0
...	...	...	...	...	...	...
139	1721050290	7.0	8.0	8.0	10.0	9.0
140	1621050162	6.3	7.0	8.5	10.0	9.0
141	1721050199	6.3	0.0	7.5	10.0	6.0
142	1621050308	0.0	5.0	0.0	10.0	9.0
143	1621050034	8.0	8.0	7.5	8.5	8.0

144 rows × 6 columns

Thực hành 2

Yêu cầu 2.3:

- Trong df_full: Tạo một cột Diem_he10 được tính dựa vào các cột tương ứng của từng hàng dữ liệu, theo công thức sau:

  Diem_he10 = 0.6*A + 0.3*((B1+B2)/2) + 0.1 * ((C1+C2)/2)

- Làm tròn đến 1 số sau dấu phẩy

	Code	A	B1	B2	C1	C2	Diem_he10
0	1621050322	8.0	0.0	5.0	7.5	8.0	6.3
1	1621050512	6.0	3.0	7.5	8.5	9.0	6.0
2	1621050211	6.7	4.0	6.5	3.0	5.0	6.0
3	1621050827	8.0	6.5	8.0	10.0	9.0	7.9
4	1621050298	7.0	5.0	8.0	8.5	9.0	7.0
...	...	...	...	...	...	...	...
139	1721050290	7.0	8.0	8.0	10.0	9.0	7.6
140	1621050162	6.3	7.0	8.5	10.0	9.0	7.1
141	1721050199	6.3	0.0	7.5	10.0	6.0	5.7
142	1621050308	0.0	5.0	0.0	10.0	9.0	1.7
143	1621050034	8.0	8.0	7.5	8.5	8.0	8.0

144 rows × 7 columns

Thực hành 2

Yêu cầu 2.4:

- Từ cột Diem_he10 trong df_full tạo một cột Diem_chu, Diem_so theo quy đổi dưới đây:

Điểm theo thang 10	Điểm theo hệ 4	
	Điểm chữ	*Điểm số*
Từ 9,0 đến 10,0	A^+	4,0
Từ 8,5 đến cận 9,0	A	3,7
Từ 8,0 đến cận 8,4	B^+	3,5
Từ 7,0 đến cận 7,9	B	3,0
Từ 6,5 đến cận 7,0	C^+	2,5
Từ 5,5 đến cận 6,5	C	2,0
Từ 5,0 đến cận 5,5	D^+	1,5
Từ 4,0 đến cận 5,0	D	1,0
Từ 0,0 đến cận 4,0	F	0

	Code	A	B1	B2	C1	C2	Diem_he10	Diem_chu	Diem_so
0	1621050322	8.0	0.0	5.0	7.5	8.0	6.3	C	2.0
1	1621050512	6.0	3.0	7.5	8.5	9.0	6.0	C	2.0
2	1621050211	6.7	4.0	6.5	3.0	5.0	6.0	C	2.0
3	1621050827	8.0	6.5	8.0	10.0	9.0	7.9	B	3.0
4	1621050298	7.0	5.0	8.0	8.5	9.0	7.0	B	3.0
...	...	...	...	...	...	...	...	...	...
139	1721050290	7.0	8.0	8.0	10.0	9.0	7.6	B	3.0
140	1621050162	6.3	7.0	8.5	10.0	9.0	7.1	B	3.0
141	1721050199	6.3	0.0	7.5	10.0	6.0	5.7	C	2.0
142	1621050308	0.0	5.0	0.0	10.0	9.0	1.7	F	0.0
143	1621050034	8.0	8.0	7.5	8.5	8.0	8.0	B+	3.5

144 rows × 9 columns

Thực hành 2

Yêu cầu 2.5:

- Tạo một DataFrame df_diem_ok chỉ lấy dữ liệu các cột [‘Code’, ‘Diem_he10’, ‘Diem_chu’, ‘Diem_so’] từ df_full

```
1  df_diem_ok.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144 entries, 0 to 143
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Code       144 non-null    int64
 1   Diem_he10  144 non-null    float64
 2   Diem_chu   144 non-null    object
 3   Diem_so    144 non-null    float64
dtypes: float64(2), int64(1), object(1)
memory usage: 4.6+ KB
```

	Code	Diem_he10	Diem_chu	Diem_so
0	1621050322	6.3	C	2.0
1	1621050512	6.0	C	2.0
2	1621050211	6.0	C	2.0
3	1621050827	7.9	B	3.0
4	1621050298	7.0	B	3.0
...	...	...	...	...
139	1721050290	7.6	B	3.0
140	1621050162	7.1	B	3.0
141	1721050199	5.7	C	2.0
142	1621050308	1.7	F	0.0
143	1621050034	8.0	B+	3.5

144 rows × 4 columns

Thực hành 2

Yêu cầu 2.6:

- Đọc dữ liệu trong sheet: code của file excel Data_point vào DataFrame df_code.
- Trộn (merge) dữ liệu của df_code và df_diem_ok để ghép phách cho bảng điểm và lưu vào DataFrame df_finaly.
- Thực hiện lưu dữ liệu trong DataFrame df_finaly ra file excel: Diem_4080130.xlsx

```
1 df_finaly
```

	Code	Name	Birth	Class	Diem_he10	Diem_chu	Diem_so
0	1421050452	Nguyễn Duy Khánh	28/03/1995	DCCTPM59_1	0.0	F	0.0
1	1421050514	Vũ Trà My	01/01/1995	DCCTPM59_1	7.6	B	3.0
2	1521020083	Tạ Văn Được	20/08/1996	DCCTPM60_1	7.1	B	3.0
3	1521050138	Nguyễn Hữu Trang	04/10/1997	DCCTPM60_1	5.2	D+	1.5
4	1521050164	Phí Đình Thành	19/05/1997	DCCTPM60_1	0.0	F	0.0
...	...	...	...	...	...	...	...
139	1721050290	Nguyễn Hoài Thương	15/01/1999	DCCTPM62A	7.6	B	3.0
140	1721050401	Nguyễn Đức Nguyên	20/06/1999	DCCTPM62B	7.8	B	3.0
141	1721050524	Nguyễn Thị Anh	18/05/1999	DCCTPM62A	7.5	B	3.0
142	1721050707	Nguyễn Thị Lý	21/08/1994	DCCTPM62B	9.3	A+	4.0
143	1931050001	Lưu Quang Linh	17/05/1988	LCCTCT64HN	7.4	B	3.0

144 rows × 7 columns

Thực hành 3

Thực hành 3

Yêu cầu:

- Đọc dữ liệu từ file Data_Patient.csv vào biến kiểu dataframe
- Cho biết các cột chứa giá trị khuyết thiếu, xác định các vị trí thiếu dữ liệu và đề xuất phương án xử lý giá trị thiếu.
- Phát hiện và xử lý ngoại lai trong tập dữ liệu (nếu có)

	A	B	C	D	E	F	G	H	I
1	id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
2	Patient_01	63	Male	Typical angina	145	233	150	6	0
3	Patient_02	67	Male	Asymptomatic	160	286	108	3	1
4	Patient_03	67	Male	Asymptomatic	120	229	129	7	1
5	Patient_04	37	Male	Non-anginal pain	130	250	187	3	0
6	Patient_05	41	Female	Atypical angina	130	204	172		0
7	Patient_16	56	Male	Atypical angina	120	236	178	3	0
8	Patient_07	62	Female	Asymptomatic	140	268	160	3	1
9	Patient_08	57	Female	Asymptomatic	120	354	163	3	0
10	Patient_19	63	Male	Asymptomatic	130	254	147	7	1
11	Patient_10	53	Male	Asymptomatic	140	203	155	7	1
12	Patient_110	57	Male	Asymptomatic	140	192	148	6	0
13	Patient_120	56	Female	Atypical angina	140	294	153	3	0
14	Patient_130	56	Male	Non-anginal pain	130	256	142	6	1
15	Patient_140	44	Male	Atypical angina	120	263	173	7	0
16	Patient_150	52	Male	Non-anginal pain	172	199	162	7	0

Data_Patient

Ready

VINBIGDATA VINGROUP Academy Vietnam

Thank you!