

BÀI TẬP THỰC HÀNH

Bài 6:

Phân tích và xử lý dữ liệu với Pandas (phần 01)

Thực hành 1



Yêu cầu 1.1: Học viên đọc dữ liệu dạng CSV lưu trong file csv_Data_Loan.csv với các tham số mặc định

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	loan_amnt	term	int_rate	emp_length	home_ownership	annual_inc	purpose	addr_state	dti	delinq_2yrs	revol_util	total_acc	bad_loan	longest_cre	verification_status
2	5000	36 months	10.65	10	RENT	24000	credit_card	AZ	27.65	0	83.7	9	0	26	verified
3	2500	60 months	15.27	0	RENT	30000	car	GA	1	0	9.4	4	1	12	verified
4	2400	36 months	15.96	10	RENT	12252	small_business	IL	8.72	0	98.5	10	0	10	not verified
5	10000	36 months	13.49	10	RENT	49200	other	CA	20	0	21	37	0	15	verified
6	5000	36 months	7.9	3	RENT	36000	wedding	AZ	11.2	0	28.3	12	0	7	verified
7	3000	36 months	18.64	9	RENT	48000	car	CA	5.35	0	87.5	4	0	4	verified
8	5600	60 months	21.28	4	OWN	40000	small_business	CA	5.55	0	32.6	13	1	7	verified
9	5375	60 months	12.69	0	RENT	15000	other	TX	18.08	0	36.5	3	1	7	verified
10	6500	60 months	14.65	5	OWN	72000	debt_consolidation	AZ	16.12	0	20.6	23	0	13	not verified
11	12000	36 months	12.69	10	OWN	75000	debt_consolidation	CA	10.78	0	67.1	34	0	22	verified
12	9000	36 months	13.49	0	RENT	30000	debt_consolidation	VA	10.08	0	91.7	9	1	7	verified
13	3000	36 months	9.91	3	RENT	15000	credit_card	IL	12.56	0	43.1	11	0	8	verified
14	10000	36 months	10.65	3	RENT	100000	other	CA	7.06	0	55.5	29	1	20	verified
15	1000	36 months	16.29	0	RENT	28000	debt_consolidation	MO	20.31	0	81.5	23	0	4	not verified
16	10000	36 months	15.27	4	RENT	42000	home_improvement	CA	18.6	0	70.2	28	0	13	not verified
17	3600	36 months	6.03	10	MORTGAGE	110000	major_purchase	CT	10.52	0	16	42	0	18	not verified
18	6000	36 months	11.71	1	MORTGAGE	84000	medical	UT	18.44	2	37.73	14	0	8	verified
19	9200	36 months	6.03	6	RENT	77385.19	debt_consolidation	CA	9.86	0	23.1	28	0	10	not verified
20	21000	36 months	12.42	10	RENT	105000	debt_consolidation	FL	13.22	0	90.3	38	1	28	verified
21	10000	36 months	11.71	10	OWN	50000	credit_card	TX	11.18	0	82.4	21	0	26	verified
22	10000	36 months	11.71	5	RENT	50000	debt_consolidation	CA	16.01	0	91.8	17	0	8	not verified
23	6000	36 months	11.71	1	RENT	76000	major_purchase	CA	2.4	0	29.7	7	1	10	not verified
24	15000	36 months	9.91	2	MORTGAGE	92000	credit_card	IL	29.44	0	93.9	31	0	9	verified



Yêu cầu 1.2: Đọc dữ liệu từ file Data_Loan.CSV vào 2 biến DataFrame tương ứng.

- **df_number**: Chỉ chứa các cột dữ liệu số
- **df_object**: Chỉ chứa các cột dữ liệu Object

```
1 #Hiển thị 5 dòng dữ liệu đầu tiên của biến df_number
2 df_number.head()
```

	loan_amnt	int_rate	emp_length	annual_inc	dti	delinq_2yrs	revol_util	total_acc	bad_loan	longest_credit_length
0	5000	10.65	10.0	24000.0	27.65	0.0	83.7	9.0	0	26.0
1	2500	15.27	0.0	30000.0	1.00	0.0	9.4			
2	2400	15.96	10.0	12252.0	8.72	0.0	98.5			
3	10000	13.49	10.0	49200.0	20.00	0.0	21.0			
4	5000	7.90	3.0	36000.0	11.20	0.0	28.3			

```
1 #Hiển thị 5 dòng dữ liệu đầu tiên của biến df_object
2 df_object.head()
```

	term	home_ownership	purpose	addr_state	verification_status
0	36 months	RENT	credit_card	AZ	verified
1	60 months	RENT	car	GA	verified
2	36 months	RENT	small_business	IL	not verified
3	36 months	RENT	other	CA	verified
4	36 months	RENT	wedding	AZ	verified



Yêu cầu 1.3: Đọc dữ liệu nhiệt độ của 6 thành phố [Hà Nội, Vinh, Đà Nẵng, Nha Trang, TP Hồ Chí Minh, Cà Mau] từ file txt_Data_Temp.txt vào biến DataFrame tương ứng

```
1 df_Temp.head()
```

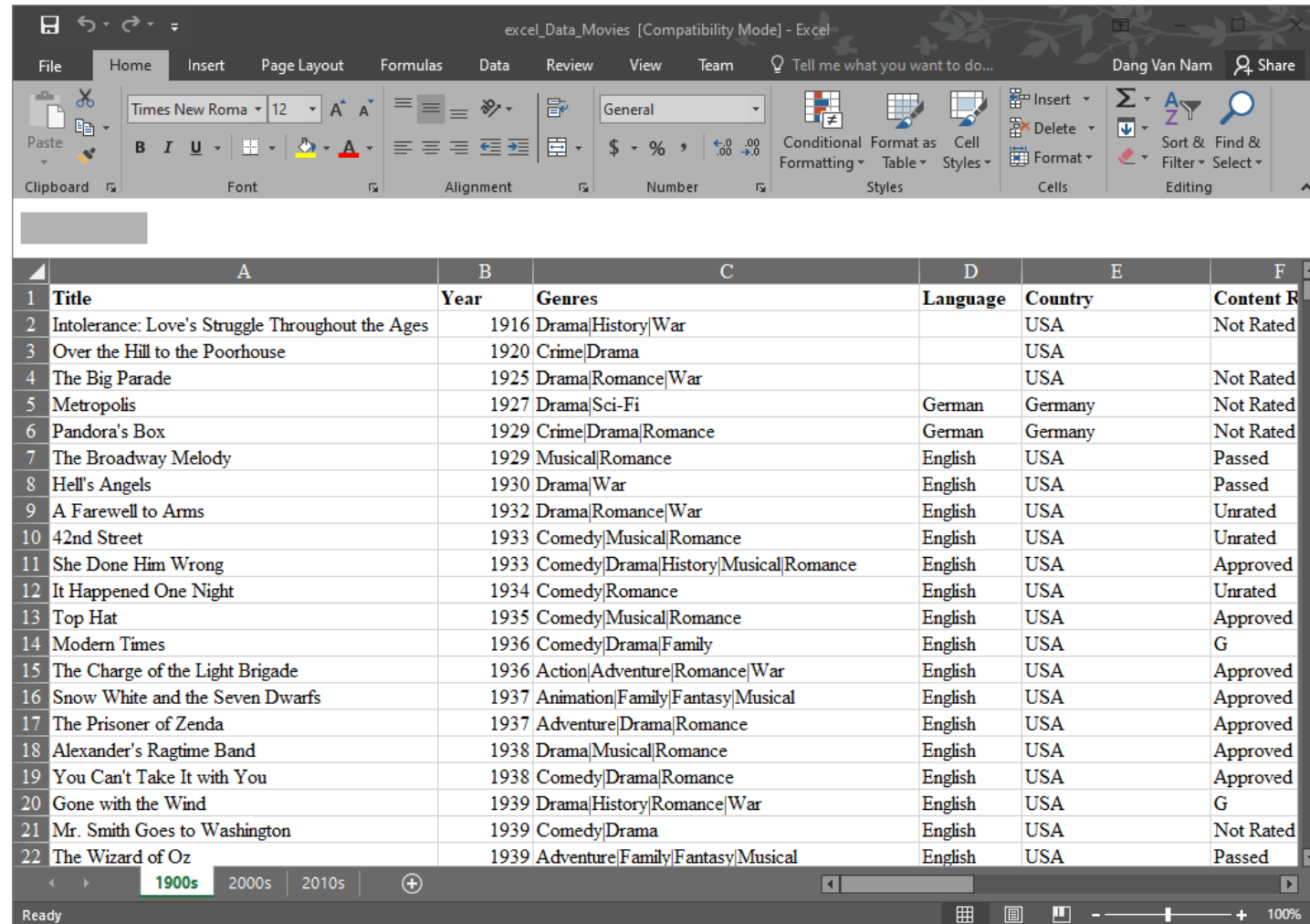
	HaNoi	Vinh	DaNang	NhaTrang	HCM	CaMau
0	25.65	24.79	24.01	25.06	25.48	24.97
1	25.31	24.21	24.02	24.93	25.16	24.83
2	25.05	23.73	23.89	24.79	24.80	24.55
3	24.79	23.36	23.83	24.84	24.74	24.48
4	24.59	23.05	23.69	24.82	24.80	24.38

Thực hành 2

Thực hành 2



Yêu cầu: Học viên đọc dữ liệu dạng excel lưu trong file excel_Data_Movies.xls theo từng sheet



	A	B	C	D	E	F
1	Title	Year	Genres	Language	Country	Content R
2	Intolerance: Love's Struggle Throughout the Ages	1916	Drama History War		USA	Not Rated
3	Over the Hill to the Poorhouse	1920	Crime Drama		USA	
4	The Big Parade	1925	Drama Romance War		USA	Not Rated
5	Metropolis	1927	Drama Sci-Fi	German	Germany	Not Rated
6	Pandora's Box	1929	Crime Drama Romance	German	Germany	Not Rated
7	The Broadway Melody	1929	Musical Romance	English	USA	Passed
8	Hell's Angels	1930	Drama War	English	USA	Passed
9	A Farewell to Arms	1932	Drama Romance War	English	USA	Unrated
10	42nd Street	1933	Comedy Musical Romance	English	USA	Unrated
11	She Done Him Wrong	1933	Comedy Drama History Musical Romance	English	USA	Approved
12	It Happened One Night	1934	Comedy Romance	English	USA	Unrated
13	Top Hat	1935	Comedy Musical Romance	English	USA	Approved
14	Modern Times	1936	Comedy Drama Family	English	USA	G
15	The Charge of the Light Brigade	1936	Action Adventure Romance War	English	USA	Approved
16	Snow White and the Seven Dwarfs	1937	Animation Family Fantasy Musical	English	USA	Approved
17	The Prisoner of Zenda	1937	Adventure Drama Romance	English	USA	Approved
18	Alexander's Ragtime Band	1938	Drama Musical Romance	English	USA	Approved
19	You Can't Take It with You	1938	Comedy Drama Romance	English	USA	Approved
20	Gone with the Wind	1939	Drama History Romance War	English	USA	G
21	Mr. Smith Goes to Washington	1939	Comedy Drama	English	USA	Not Rated
22	The Wizard of Oz	1939	Adventure Family Fantasy Musical	English	USA	Passed

Thực hành 3

Mô tả file dữ liệu: Data_Patient.csv

- File dữ liệu chứa thông tin của 300 bệnh nhân bị chứng đau ngực
- Mỗi dòng ứng với thông tin của một bệnh nhân, bao gồm 9 thuộc tính

	A	B	C	D	E	F	G	H	I
1	id	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8
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_06	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_09	63	Male	Asymptomatic	130	254	147	7	1
11	Patient_10	53	Male	Asymptomatic	140	203	155	7	1
12	Patient_11	57	Male	Asymptomatic	140	192	148	6	0
13	Patient_12	56	Female	Atypical angina	140	294	153	3	0
14	Patient_13	56	Male	Non-anginal pain	130	256	142	6	1
15	Patient_14	44	Male	Atypical angina	120	263	173	7	0
16	Patient_15	52	Male	Non-anginal pain	172	199	162	7	0
17	Patient_16	57	Male	Non-anginal pain	150	168	174	3	0
18	Patient_17	48	Male	Atypical angina	110	229	168	7	1
19	Patient_18	54	Male	Asymptomatic	140	239	160	3	0
20	Patient_19	48	Female	Non-anginal pain	130	275	139	3	0
21	Patient_20	49	Male	Atypical angina	130	266	171	3	0

Chi tiết thông tin của một bệnh nhân như sau:

- **id:** Mã của bệnh nhân (số)
- **Feature_1:** Tuổi của bệnh nhân (số)
- **Feature_2:** Giới tính của bệnh nhân (chuỗi: Male – Female)
- **Feature_3:** Cho biết loại triệu chứng đau ngực mà bệnh nhân này mắc phải, với 4 giá trị: (Typical angina, Atypical angina, Non-anginal pain, Asymptomatic)
- **Feature_4:** Huyết áp của bệnh nhân – đơn vị: mmhg (số)
- **Feature_5:** Chỉ số cholesterol của bệnh nhân – đơn vị: mg/dl (số)
- **Feature_6:** Thông số nhịp tim của bệnh nhân – đơn vị: lần/phút (số)
- **Feature_7:** Chỉ số Thalassemia của bệnh nhân chỉ gồm 3 giá trị (3: Bình thường | 4: Khiếm khuyết cố định | 7: Kiểm khuyết có thể đảo ngược)
- **Feature_8:** Cho biết bệnh nhân có bị bệnh tim hay không? (0: Không bị bệnh tim mạch | 1: Bị bệnh tim mạch)

Yêu cầu 3.1:

- Đọc dữ liệu từ file Data_Patient.csv vào biến kiểu dataframe: df_patient với cột id là cột chỉ số (index_col)
- Hiển thị thông tin tổng quan của tập dữ liệu
- Hiển thị thông tin của 10 bệnh nhân đầu tiên và 5 bệnh nhân cuối cùng của tập dữ liệu.
- Đặt lại tên các cột dữ liệu trong Dataframe như sau:
 - Feature_1 → Age
 - Feature_2 → Gender
 - Feature_3 → Type
 - Feature_4 → Blood_pressure
 - Feature_5 → Cholesterol
 - Feature_6 → Heartbeat
 - Feature_7 → Thalassemia
 - Feature_8 → Result



Yêu cầu 3.2:

- Sử dụng phương thức `.describe()` cho biết:
 - Thuộc tính Age:
 - Tuổi của bệnh nhân trẻ nhất
 - Tuổi của bệnh nhân già nhất
 - Thuộc tính Cholesterol:
 - Cholesterol trung bình của các bệnh nhân
 - Độ lệch chuẩn của giá trị này trong toàn bộ tập dữ liệu
 - Bao nhiêu bệnh nhân giới tính nam (Male)
 - Có bao nhiêu giá trị khác nhau của thuộc tính Type. Giá trị xuất hiện nhiều nhất là giá trị nào, bao nhiêu lần.

	Gender	Type
count	300	295
unique	2	4
top	Male	Asymptomatic
freq	205	139

Yêu cầu 3.3:

- Cho biết những cột nào trong dữ liệu có chứa missing data và số lượng missing là bao nhiêu?

Yêu cầu 3.4:

- Hiển thị thông tin của các bệnh nhân:
 - Bệnh nhân có index: **Patient_100; Patient_150; Patient_200**
 - Bệnh nhân ở vị trí 255 đến 260, với 3 thuộc tính: Age, Gender và Result**

	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
id								
Patient_100	45	Male	Asymptomatic	115	260	185	3.0	0
Patient_150	52	Male	Typical angina	152	298	178	7.0	0
Patient_200	50	Female	Asymptomatic	110	254	159	3.0	0

	Age	Gender	Result
id			
Patient_255	42	Female	0
Patient_256	67	Female	0
Patient_257	76	Female	0
Patient_258	70	Male	0
Patient_259	57	Male	1
Patient_260	44	Female	0

Yêu cầu 3.5:


- Thay đổi giá trị cho thuộc tính Gender: **Male → 0, Female → 1**
- Thay đổi giá trị cho thuộc tính Result: **0 → No, 1 → Yes**
- Cập nhật giá trị thuộc tính Thalassemia của bệnh nhân có index: **Patient_05 bằng giá trị 4.0**

	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
id								
Patient_01	63	0	Typical angina	145	233	150	6.0	No
Patient_02	67	0	Asymptomatic	160	286	108	3.0	Yes
Patient_03	67	0	Asymptomatic	120	229	129	7.0	Yes
Patient_04	37	0	Non-anginal pain	130	250	187	3.0	No
Patient_05	41	1	Atypical angina	130	204	172	4.0	No

Thực hành 4

Yêu cầu 4.1:

- Đọc dữ liệu từ file Data_Patient.csv vào biến kiểu dataframe: df_patient với cột id là cột chỉ số (index_col)
- Đặt lại tên các cột dữ liệu trong Dataframe như sau:




	A	B	C	D	E	F	G	H	I
1	id	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8
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_06	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_09	63	Male	Asymptomatic	130	254	147	7	1
11	Patient_10	53	Male	Asymptomatic	140	203	155	7	1
12	Patient_11	57	Male	Asymptomatic	140	192	148	6	0
13	Patient_12	56	Female	Atypical angina	140	294	153	3	0
14	Patient_13	56	Male	Non-anginal pain	130	256	142	6	1
15	Patient_14	44	Male	Atypical angina	120	263	173	7	0
16	Patient_15	52	Male	Non-anginal pain	172	199	162	7	0
17	Patient_16	57	Male	Non-anginal pain	150	168	174	3	0
18	Patient_17	48	Male	Atypical angina	110	229	168	7	1
19	Patient_18	54	Male	Asymptomatic	140	239	160	3	0
20	Patient_19	48	Female	Non-anginal pain	130	275	139	3	0
21	Patient_20	49	Male	Atypical angina	130	266	171	3	0

Ready Data_Patient

- Feature_1 → Age
- Feature_2 → Gender
- Feature_3 → Type
- Feature_4 → Blood_pressure
- Feature_5 → Cholesterol
- Feature_6 → Heartbeat
- Feature_7 → Thalassemia
- Feature_8 → Result

Yêu cầu 4.2:

- 
- Lọc dữ liệu trong df_patient thành các DataFrame:
 - **df_male**: chứa danh sách bệnh nhân Nam
 - **df_female**: chứa danh sách bệnh nhân nữ
 - **df_no**: danh sách những người không bị bệnh đau tim
 - **df_yes**: danh sách những người bị bệnh đau tim

Yêu cầu 4.3:

- Lọc trong df_patient đưa ra danh sách bệnh nhân thỏa mãn yêu cầu sau:
 1. Những người bị mắc bệnh **đau tim** và trên **70 tuổi**
 2. Người có giới tính **Female**, có huyết áp trên **170 mmhg** nhưng **không bị bệnh đau tim**.
 3. Những người có triệu chứng đau ngực là **Typical angina**, giới tính **Male** và **bị bệnh đau tim**.



Yêu cầu 4.4: Xác định:

1. Chỉ số huyết áp (**Blood_pressure**) thấp nhất, cao nhất, trung bình, trung vị và độ lệch chuẩn của tập dữ liệu
2. Chỉ số nhịp tim (**Heartbeat**) thấp nhất, cao nhất, trung bình, trung vị và độ lệch chuẩn của tập dữ liệu

1. Chỉ số huyết áp:

Min: 94

Max: 200

Mean: 131.68666666666667

Median: 130.0

Std: 17.682497692285477

2. Chỉ số nhịp tim:

Min: 71

Max: 202

Mean: 149.56333333333333

Median: 152.5

Std: 22.818595118151098

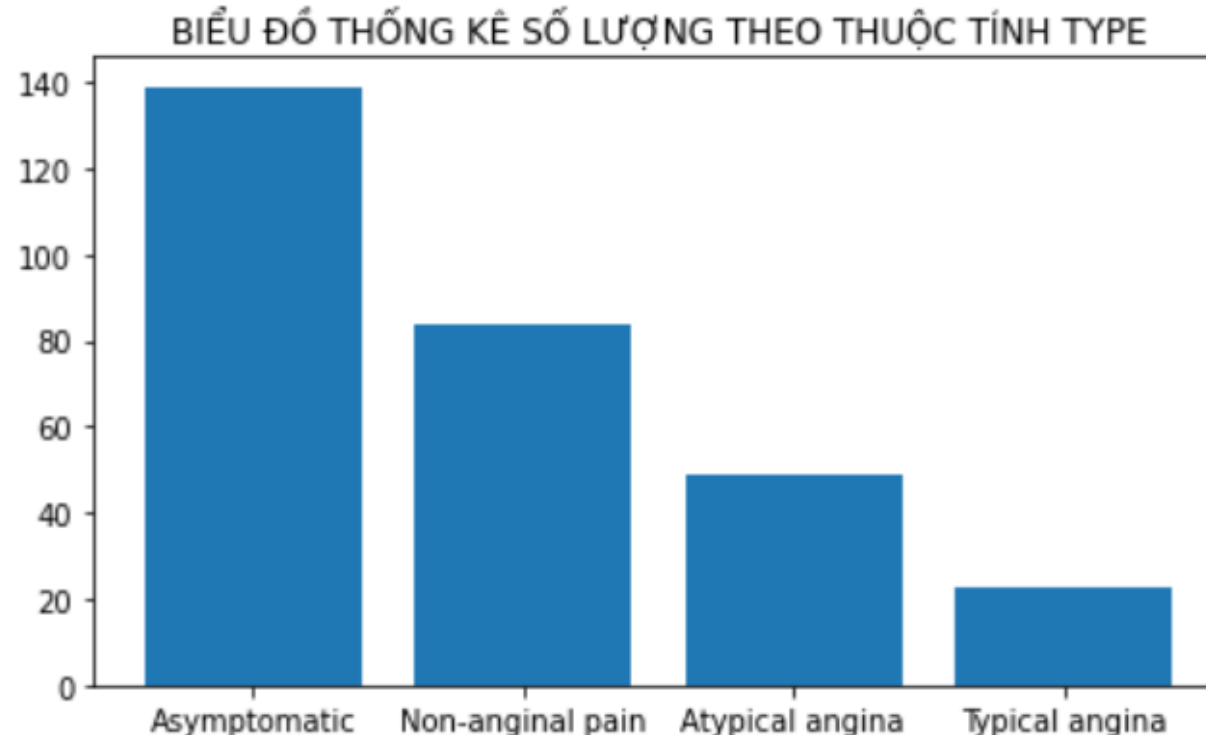




Yêu cầu 4.5: Xác định:

1. Số giá trị khác nhau của thuộc tính **Type**
2. Vẽ đồ thị dạng cột thể hiện kết quả thống kê số lượng theo từng giá trị khác nhau của thuộc tính Type

Asymptomatic	139
Non-anginal pain	84
Atypical angina	49
Typical angina	23



Thực hành 5

Thực hành 5



Yêu cầu: Dựa vào dữ liệu chuỗi thời gian quan trắc thông số nhiệt độ của Hà Giang và Cà Mau từ năm 2012 đến 2019 (Data_Temperature.csv), sử dụng các kỹ thuật để tìm ra những thông tin từ 2 bộ dữ liệu đó.

TimeVN	HaGiang	CaMau
2012-08-01 7:00	26.2	25.4
2012-08-01 10:00	31.4	30.8
2012-08-01 13:00	35.1	30
2012-08-01 16:00	35.8	27.7
2012-08-01 19:00	30.4	28.1
2012-08-01 22:00	24	27.6
2012-08-02 1:00	23.8	24.4
2012-08-02 4:00	23.9	24.5
2012-08-02 7:00	24	25.4
2012-08-02 10:00	28.8	23.3
2012-08-02 13:00	34.7	24
2012-08-02 16:00	26.7	27.5
2012-08-02 19:00	25.4	27.2
2012-08-02 22:00	24.8	26.6
2012-08-03 1:00	24.5	26.5
2012-08-03 4:00	24.4	26.4
2012-08-03 7:00	24.6	26.4
2012-08-03 10:00	30.1	26.5
2012-08-03 13:00	34.4	30.6
2012-08-03 16:00		28.9
2012-08-03 19:00		27.6

Data_Temperature



Thank you!