

oooo



TRƯỜNG ĐẠI HỌC MỞ TP. HỒ CHÍ MINH
HO CHI MINH CITY OPEN UNIVERSITY

BÁO CÁO BÀI TẬP LỚN

MÔN HỌC: PHÂN TÍCH DỮ LIỆU

oooo

CHẨN ĐOÁN BỆNH TIỂU ĐƯỜNG DỰA VÀO BỘ
DỮ LIỆU KẾT QUẢ XÉT NGHIỆM MÁU

Giảng viên hướng dẫn: Hồ Hướng Thiên

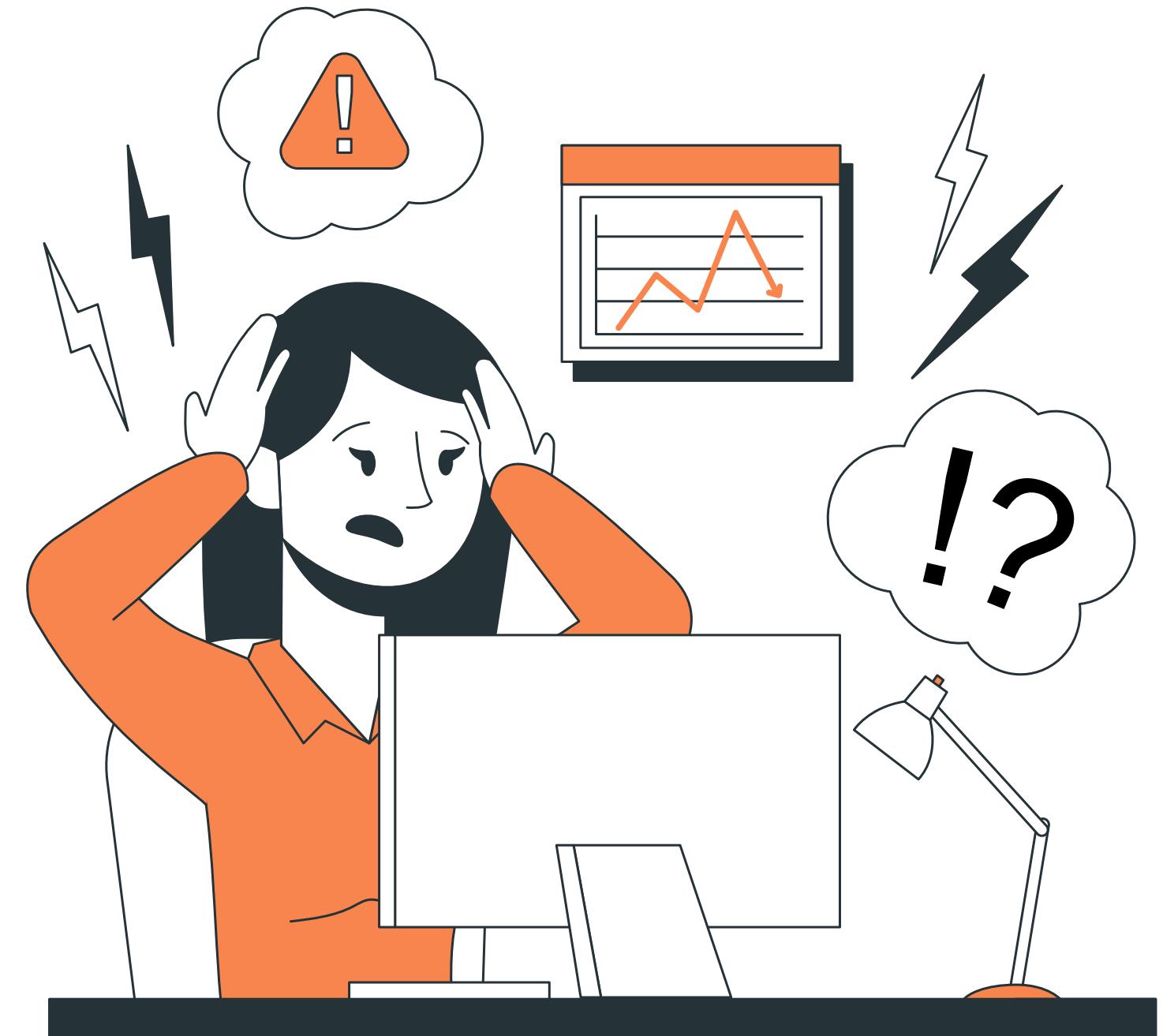
Sinh viên thực hiện: Phạm Công Thuận - 2151013097

Tạ Thị Thiên Thanh - 2151013088



NỘI DUNG

1. Mở đầu
2. Tổng quan về phân tích dữ liệu
3. Mô tả dữ liệu
4. Tiền xử lý dữ liệu
5. Tìm luật kết hợp bằng Apriori
6. Thuật toán phân lớp Naive Bayes
7. Thuật toán phân lớp SVM
8. Kết luận



o o o o

○ ○ ○ ○

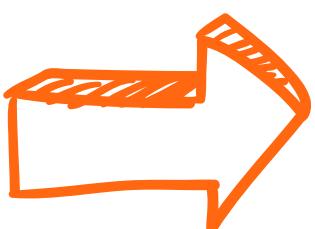
++ + + +

+ +
+ +
+ +
+ +
+ +

Mô tả dữ liệu

Bộ dữ liệu lâm sàng của hơn 5000 bệnh nhân

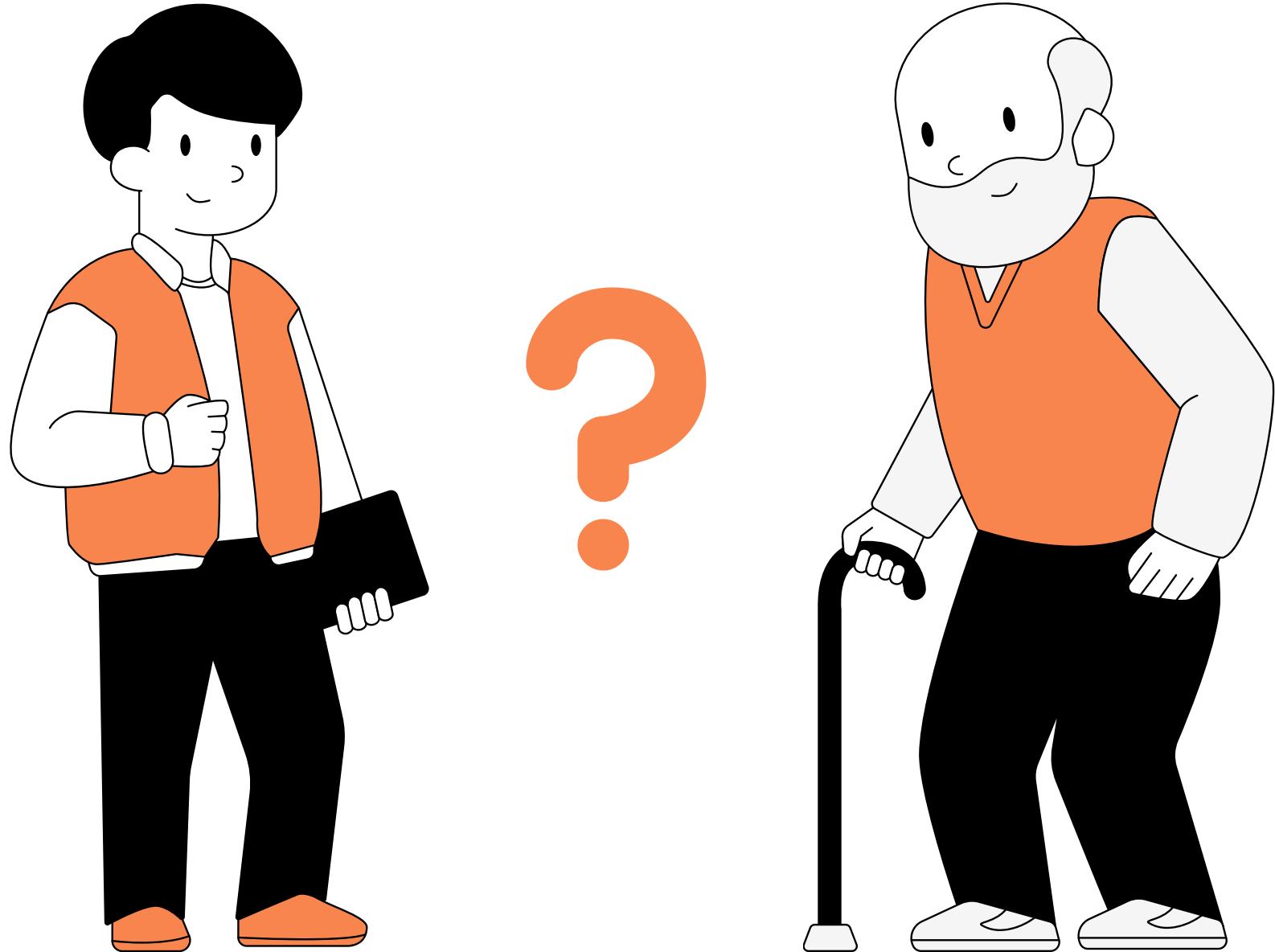
Phục vụ nghiên cứu chức năng thận, sức khỏe tim mạch, chẩn đoán tiểu đường



Đề tài tập trung nghiên cứu khả năng chẩn đoán bệnh tiểu đường

*Dữ liệu lâm sàng là dữ liệu thu được tại một cơ sở khám chữa bệnh chằng hạn như trạm y tế, phòng khám, bệnh viện

Thuộc tính ‘Age’



Số tuổi của bệnh nhân

Thuộc tính ‘Age’

mean	48.637853	Số tuổi trung bình
std	15.253447	Độ lệch chuẩn
min	-94.000000	Số tuổi nhỏ nhất
25%	36.000000	Tứ phân vị Q1
50%	49.000000	Giá trị trung vị (mean)
75%	59.000000	Tứ phân vị Q3
max	93.000000	Số tuổi cao nhất

Số lượng giá trị
không null



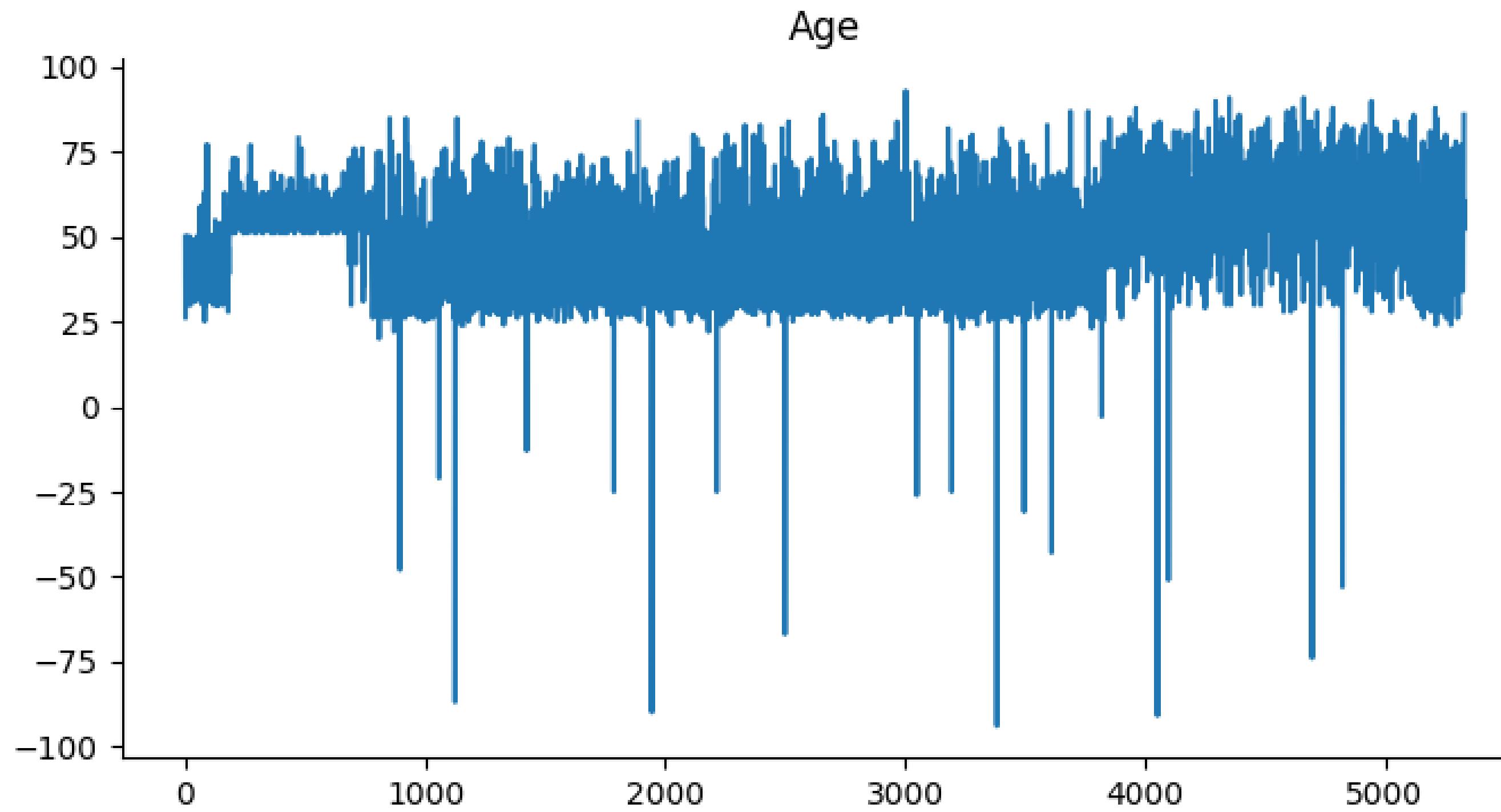
5310 non-null



float64

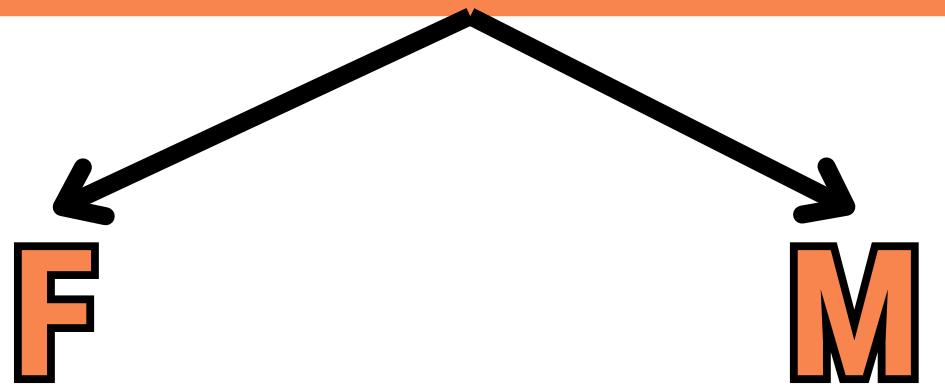
Kiểu dữ liệu
(datatype)

Thuộc tính ‘Age’

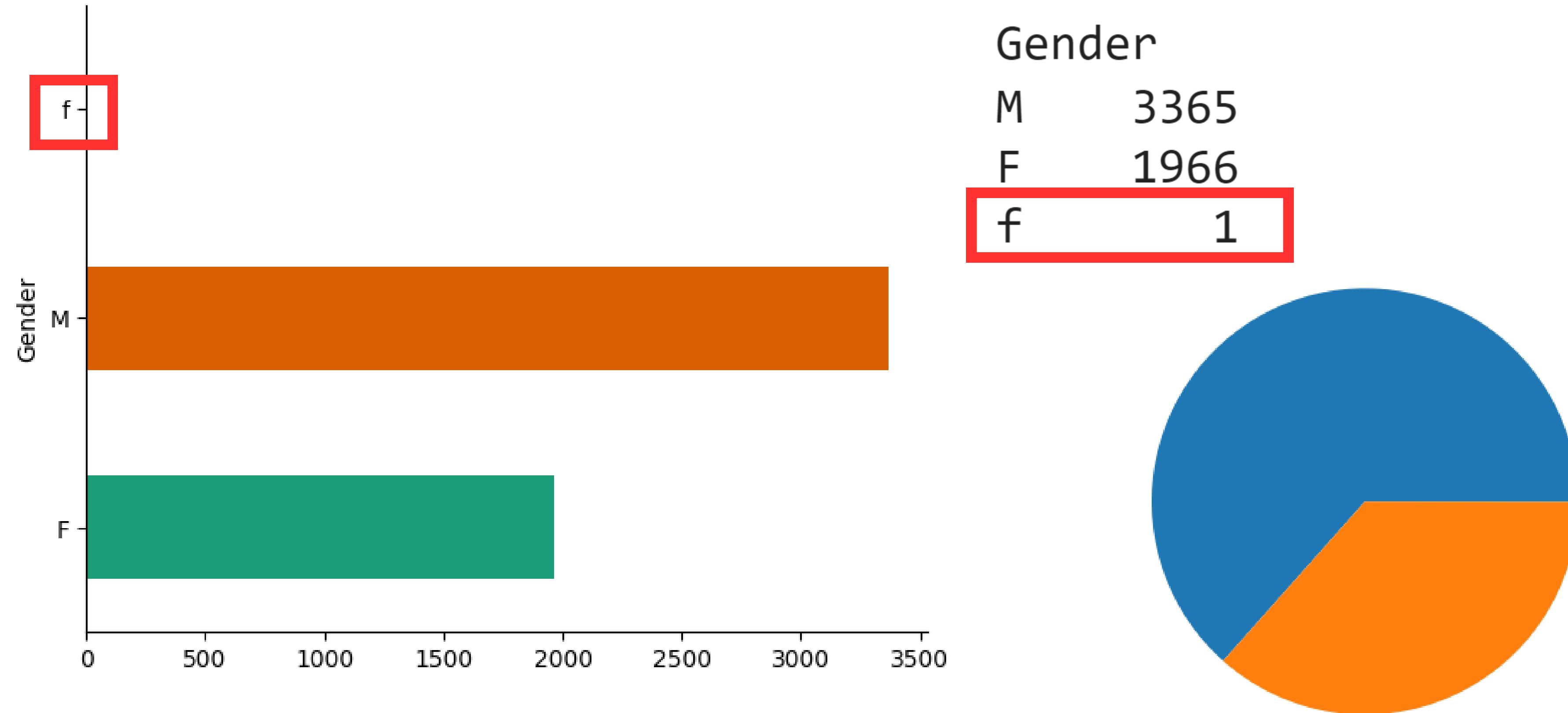


Thuộc tính ‘Gender’

Giới tính của bệnh nhân
bao gồm 2 giá trị

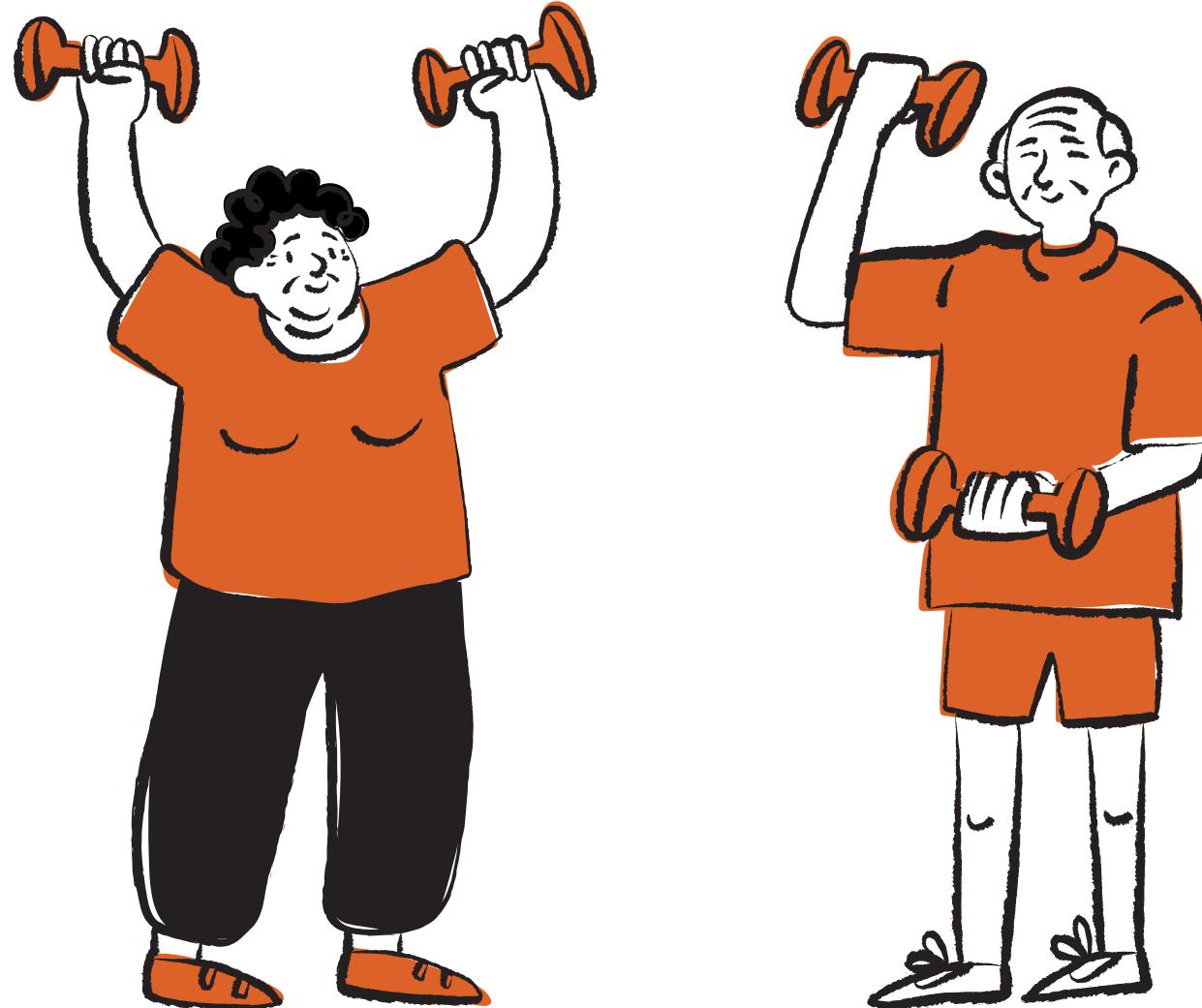
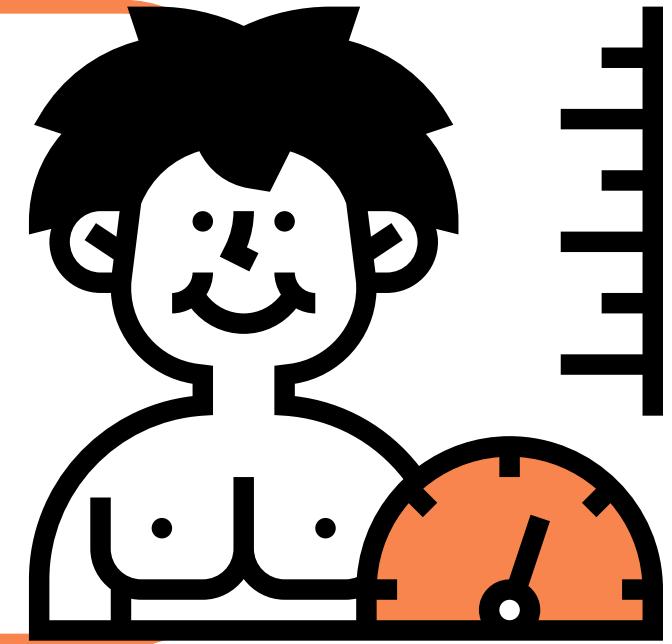


Thuộc tính ‘Gender’



Thuộc tính ‘BMI’

Body Mass Index (BMI): Chỉ số BMI của bệnh nhân.
BMI là chỉ số khối cơ thể dùng để xác định cân nặng
của một người đang thiếu cân, thừa cân hay cân đối



$$\frac{\text{weight}}{\text{height}^2}$$

Thuộc tính ‘BMI’

mean	24.130345	Giá trị BMI trung bình
std	7.718847	Độ lệch chuẩn
min	-99.000000	Giá trị BMI nhỏ nhất
25%	22.000000	Tứ phân vị Q1
50%	24.000000	Giá trị trung vị (mean)
75%	27.000000	Tứ phân vị Q3
max	47.000000	Giá trị BMI cao nhất

Số lượng giá trị
không null

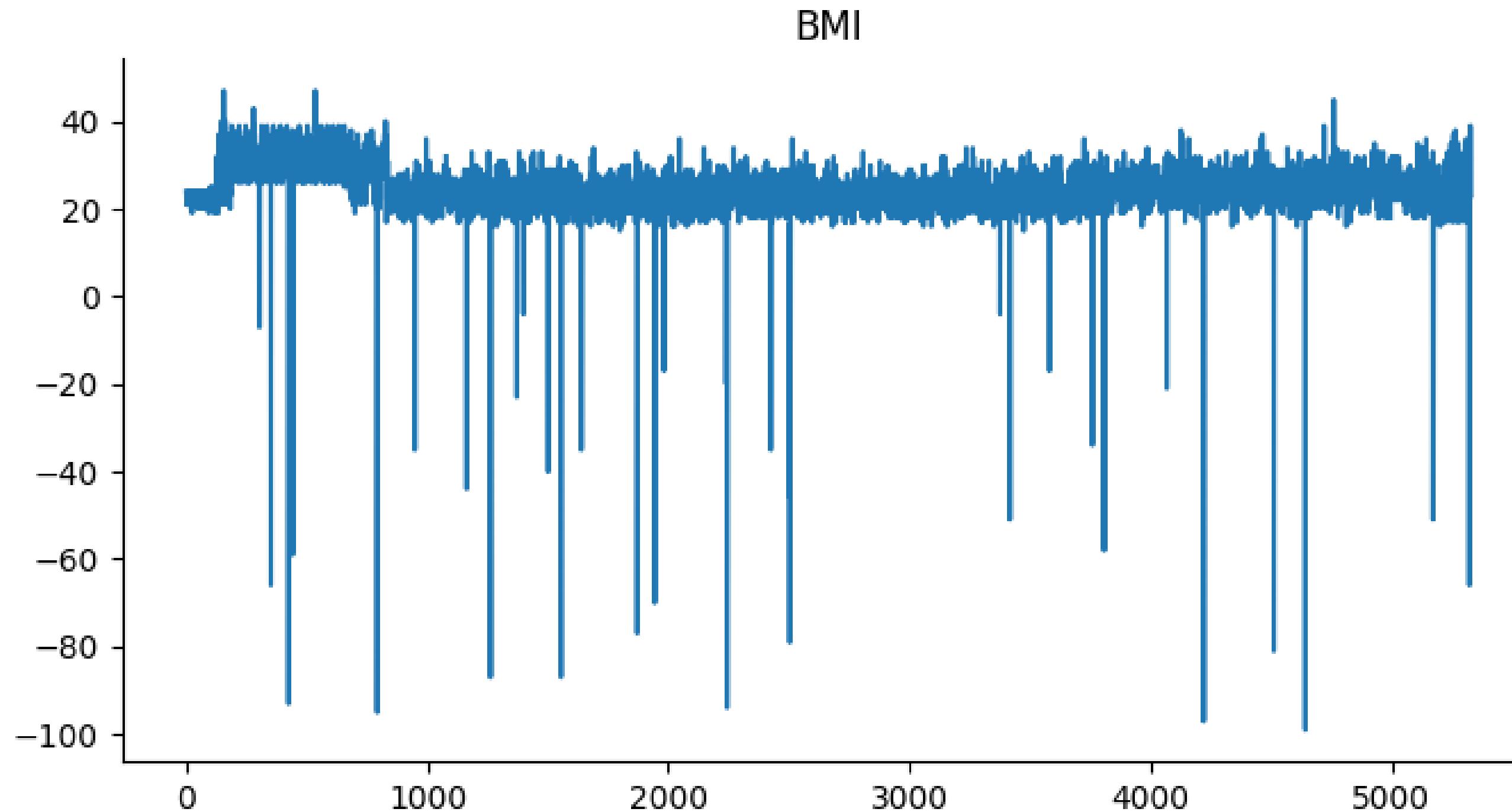


5309 non-null

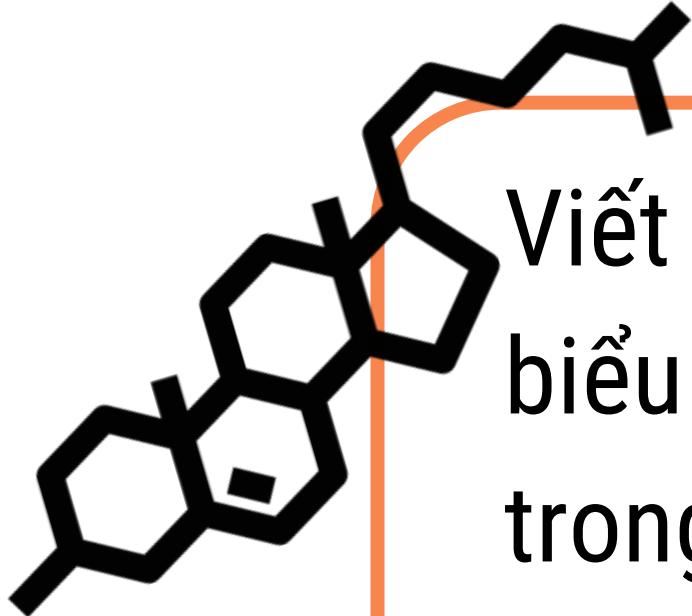
float64 →

Kiểu dữ liệu
(datatype)

Thuộc tính ‘BMI’



Thuộc tính ‘Chol’



Viết tắt của Cholesterol, biểu thị tỷ lệ cholesterol có trong máu.

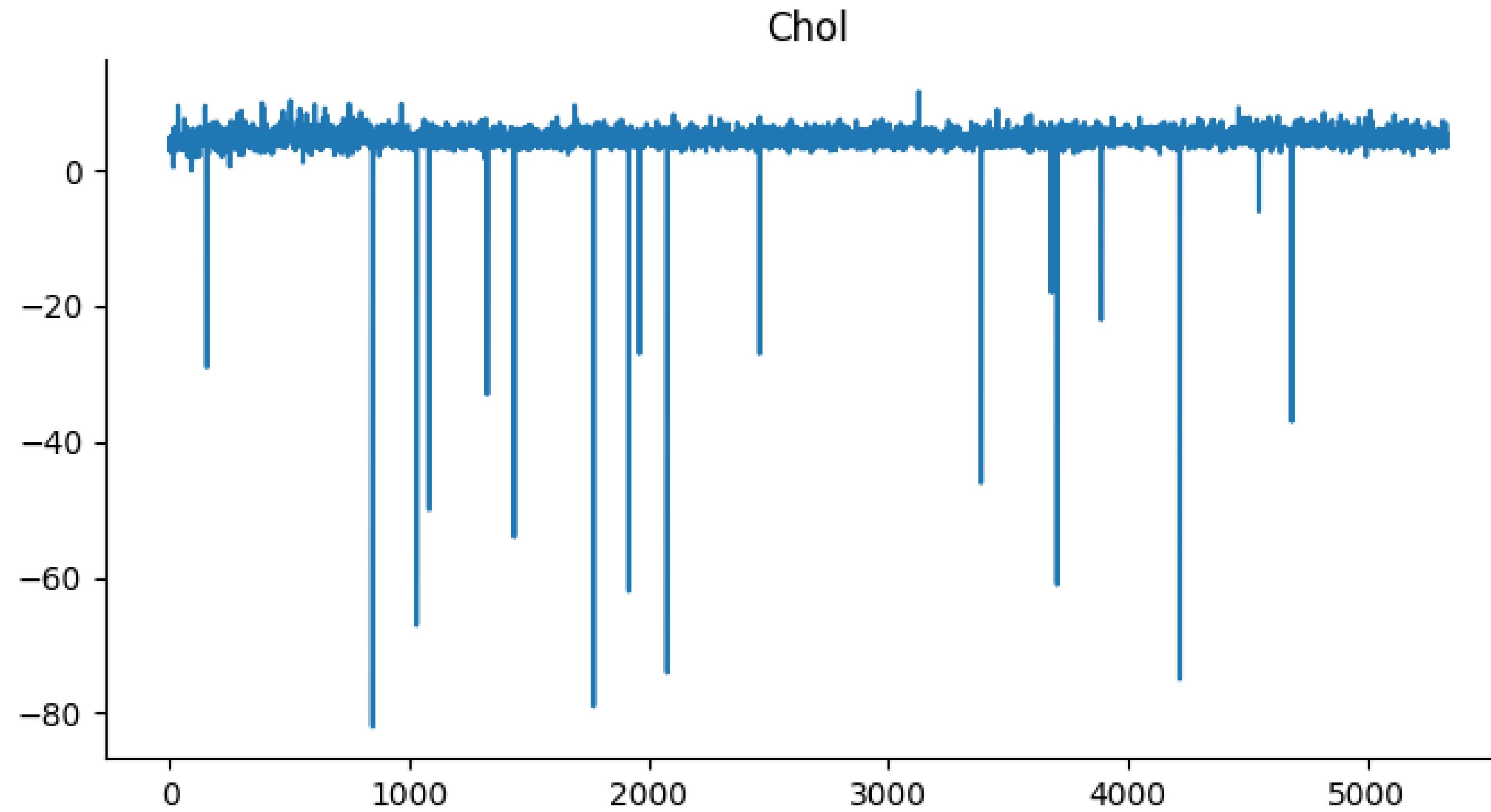
Cholesterol là một loại chất béo được sản sinh ra từ việc tiêu thụ thức ăn hoặc cơ thể tự sản xuất



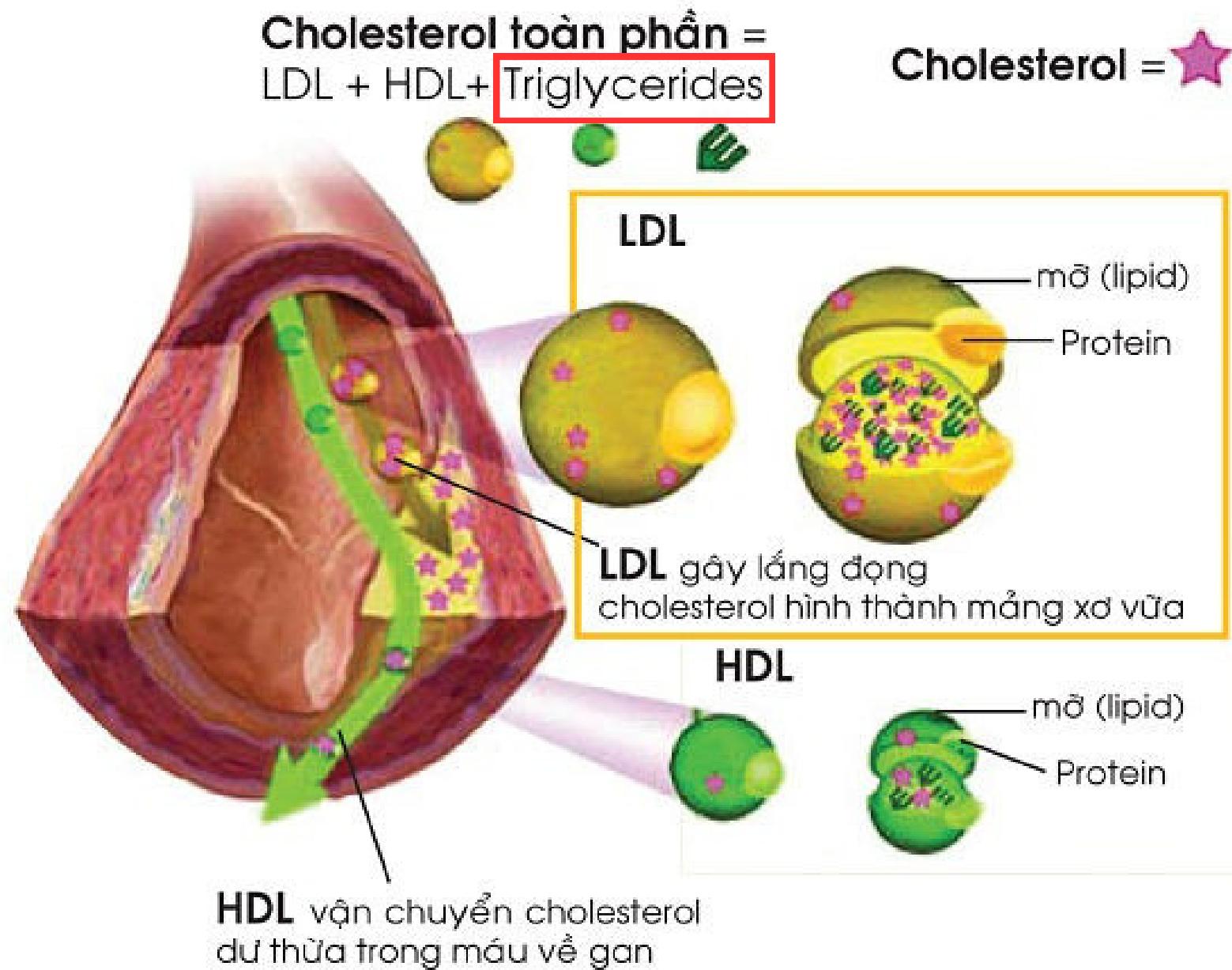
Thuộc tính ‘Chol’

mean	4.689629	Tỷ lệ Cholesterol trung bình
std	3.448396	Độ lệch chuẩn
min	-82.000000	Tỷ lệ Cholesterol thấp nhất
25%	4.180000	Tứ phân vị Q1
50%	4.800000	Giá trị trung vị (mean)
75%	5.460000	Tứ phân vị Q3
max	11.650000	Tỷ lệ Cholesterol cao nhất
Số lượng giá trị không null	← 5306 non-null →	Kiểu dữ liệu (datatype)

Thuộc tính ‘Chol’



Thuộc tính 'TG'



Là tỷ lệ triglycerides có trong máu.

Triglycerides là một dạng chất béo trung tính chứa 3 axit béo và có nguồn gốc từ mỡ động vật, thực vật mà bệnh nhân tiêu thụ.

CÁC THÀNH PHẦN MỠ MÁU

Thuộc tính ‘TG’

mean	1.503966	Tỷ lệ Triglycerides trung bình
std	3.983595	Độ lệch chuẩn
min	-94.000000	Tỷ lệ Triglycerides thấp nhất
25%	0.900000	Tứ phân vị Q1
50%	1.370000	Giá trị trung vị (mean)
75%	2.100000	Tứ phân vị Q3
max	32.640000	Tỷ lệ Triglycerides cao nhất

Số lượng giá trị
không null



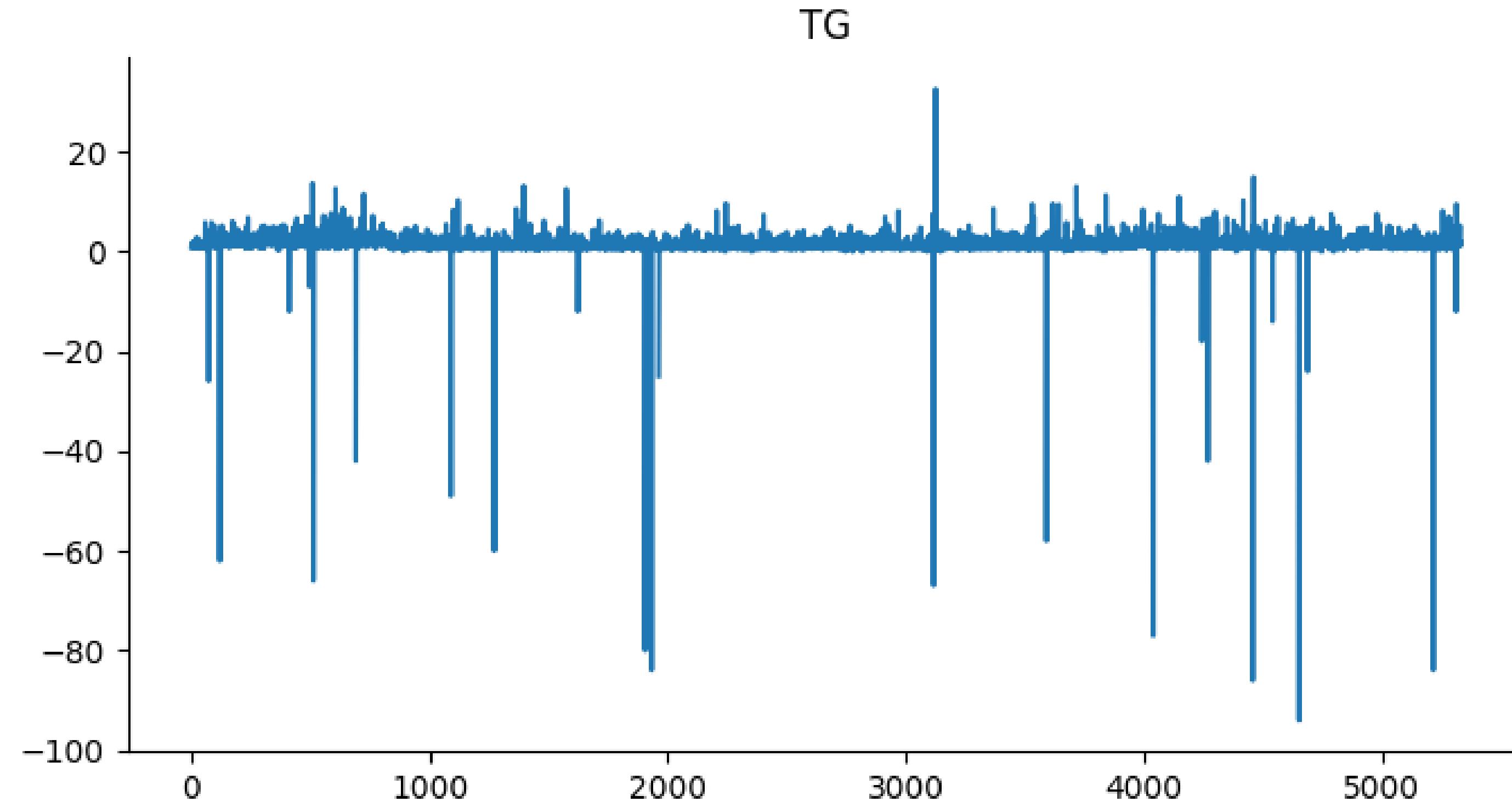
5300 non-null

float64

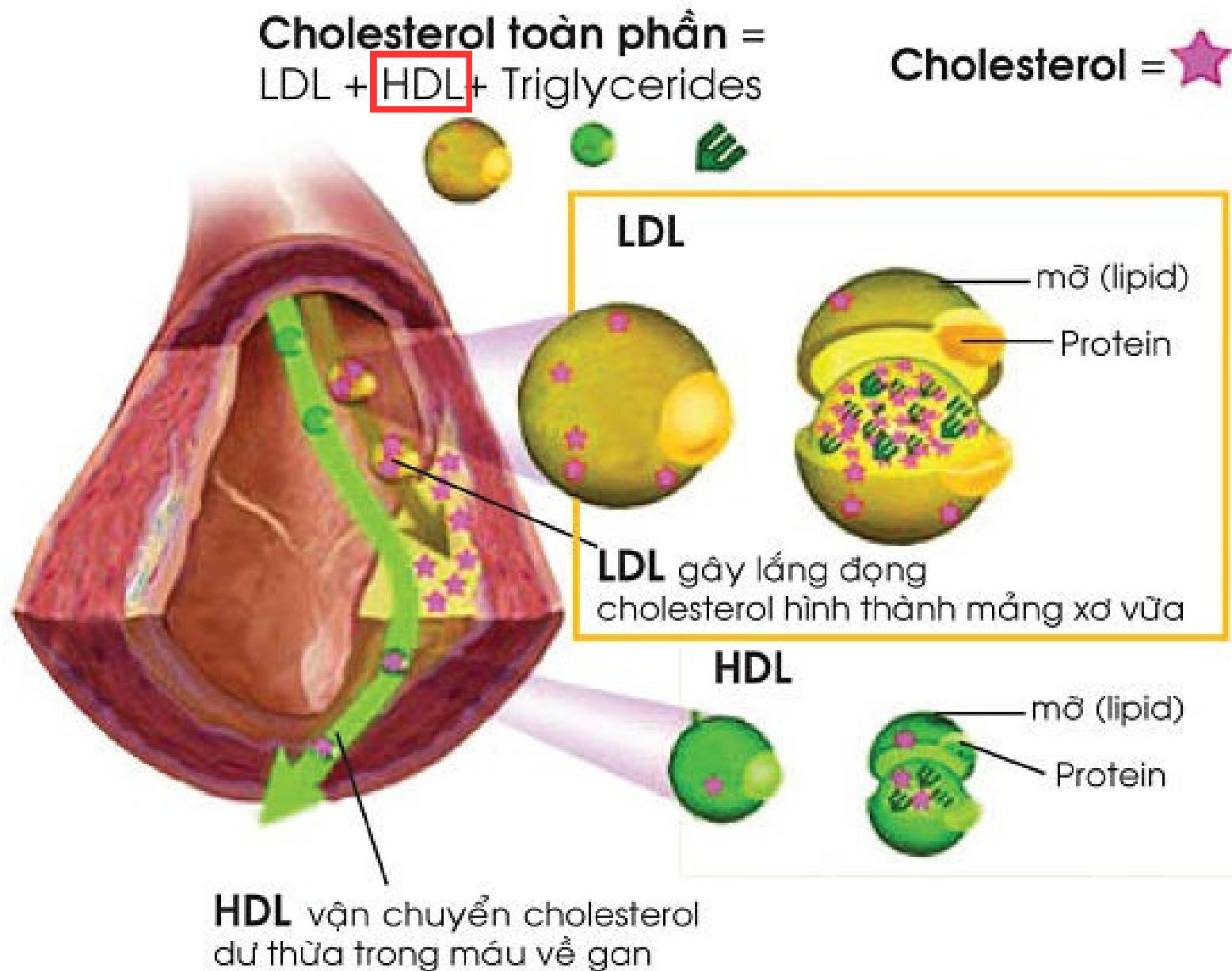


Kiểu dữ liệu
(datatype)

Thuộc tính 'TG'



Thuộc tính 'HDL'



CÁC THÀNH PHẦN MỠ MÁU

HDL (High-Density Lipoprotein) là chỉ số Lipoprotein tỷ trọng cao.

Lipoprotein tỷ trọng cao được xem như một loại cholesterol có lợi giúp vận chuyển cholesterol dư thừa tích trữ dưới mạch máu về gan để xử lý và đào thải ra ngoài.

Thuộc tính ‘HDL’

mean	1.346343	Chỉ số HDL trung bình
std	4.349107	Độ lệch chuẩn
min	-95.000000	Chỉ số HDL thấp nhất
25%	1.090000	Tứ phân vị Q1
50%	1.300000	Giá trị trung vị (mean)
75%	1.590000	Tứ phân vị Q3
max	9.900000	Chỉ số HDL cao nhất

Số lượng giá trị
không null

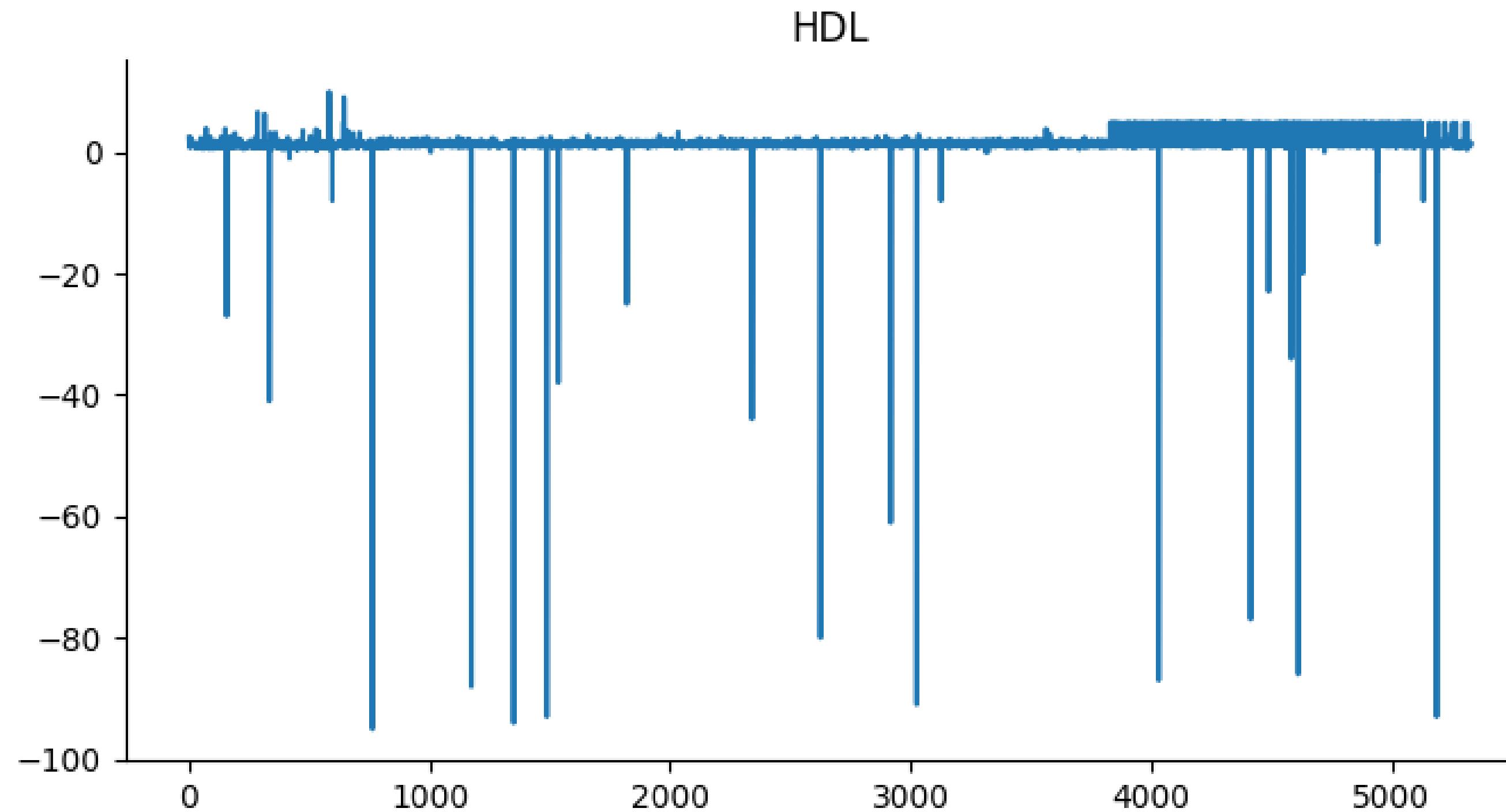


5310 non-null

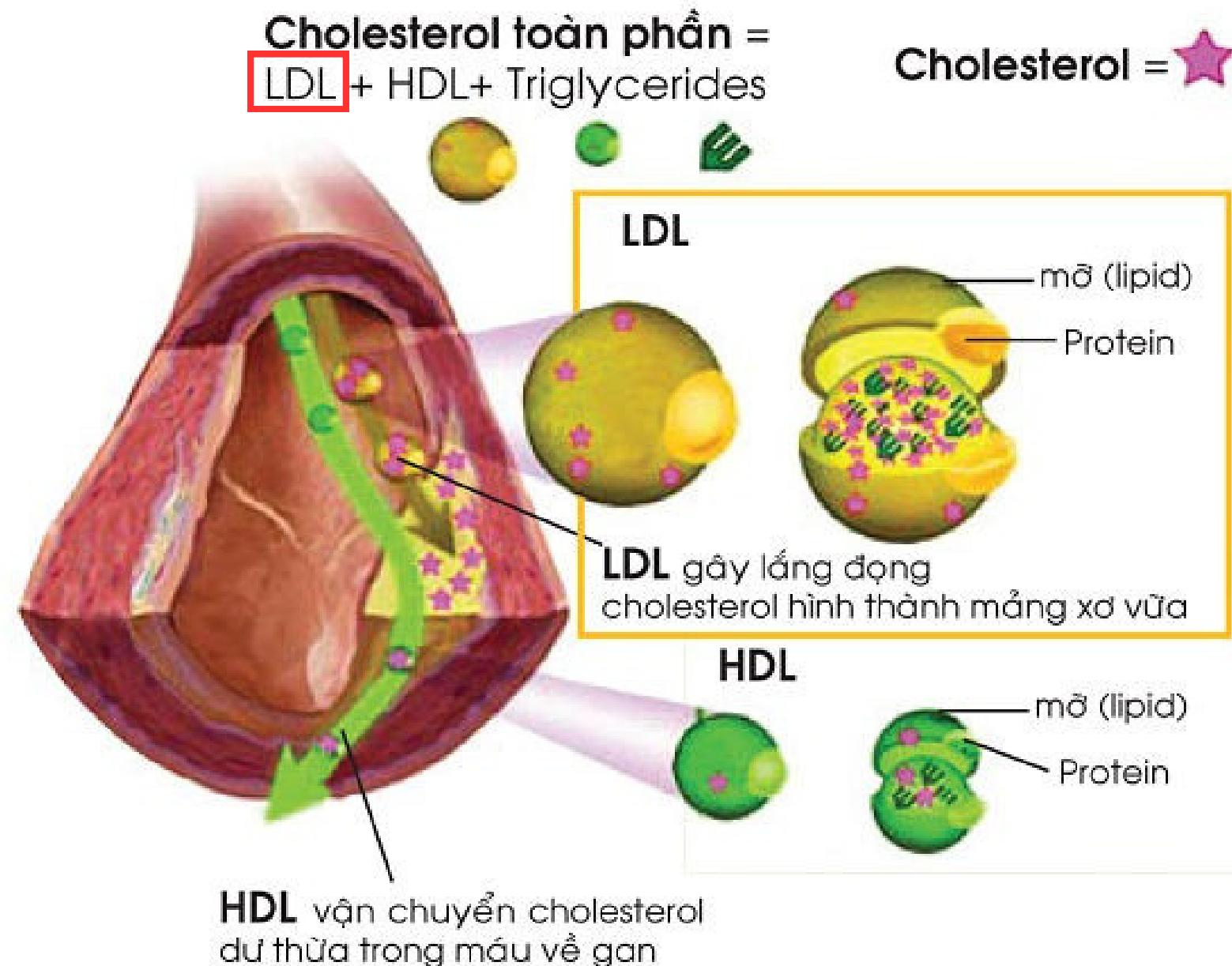
float64 →

Kiểu dữ liệu
(datatype)

Thuộc tính ‘HDL’



Thuộc tính ‘LDL’



CÁC THÀNH PHẦN MỠ MÁU

LDL (Low-Density Lipoprotein) là chỉ số Lipoprotein tỷ trọng thấp.

Lipoprotein tỷ trọng thấp các cholesterol có hại làm tăng nguy cơ xơ vữa động mạch.

Thuộc tính ‘LDL’

mean	2.705066	Chỉ số LDL trung bình
std	3.910821	Độ lệch chuẩn
min	-98.000000	Chỉ số LDL thấp nhất
25%	2.270000	Tứ phân vị Q1
50%	2.780000	Giá trị trung vị (mean)
75%	3.390000	Tứ phân vị Q3
max	9.900000	Chỉ số LDL cao nhất

Số lượng giá trị
không null

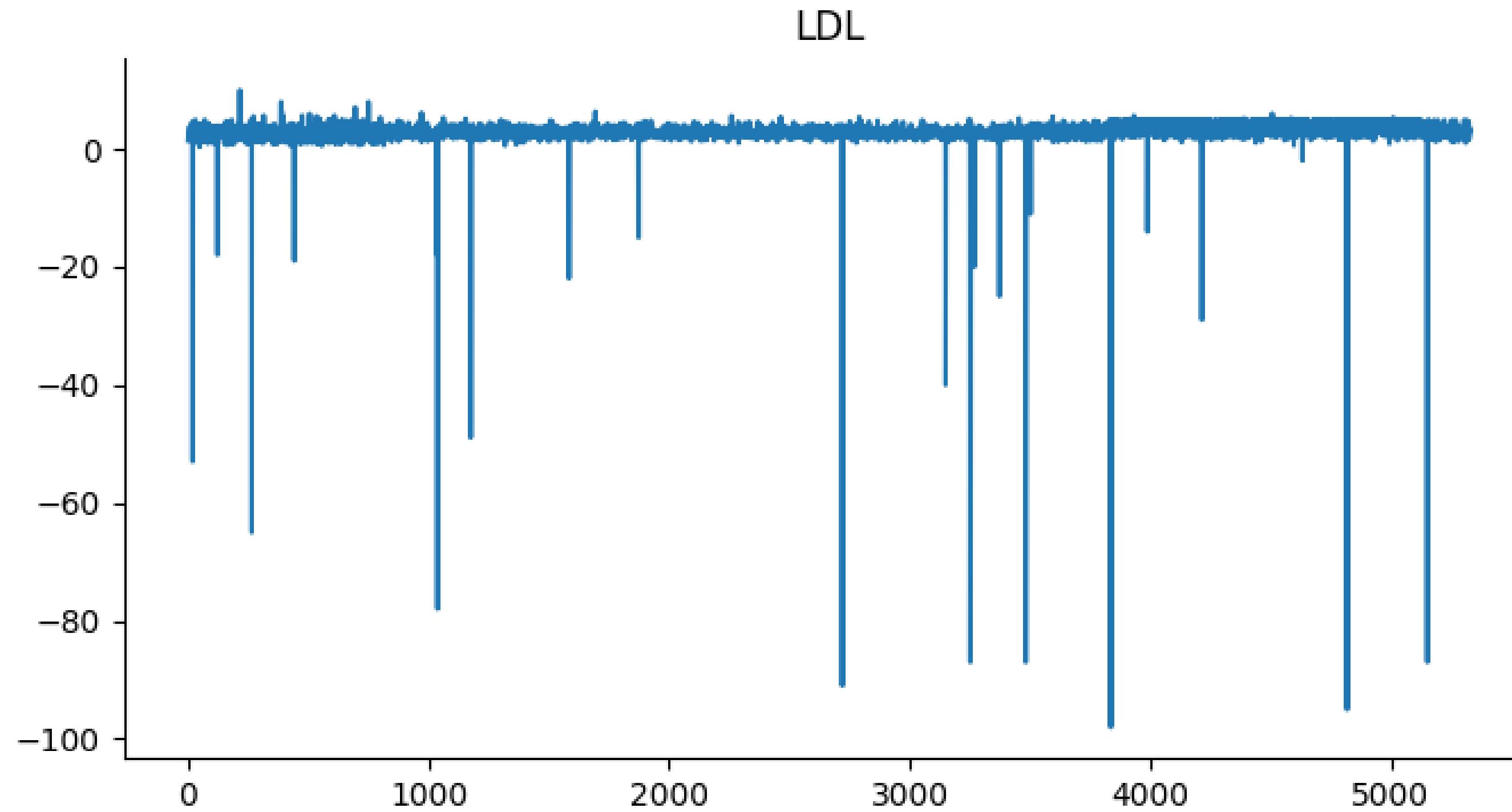


5314 non-null

float64 →

Kiểu dữ liệu
(datatype)

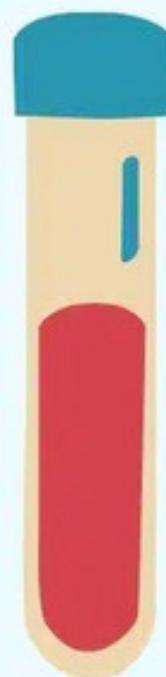
Thuộc tính ‘LDL’



Thuộc tính 'Cr'



CHỈ SỐ CREATININ MÁU BÌNH THƯỜNG LÀ BAO NHIÊU?



Bình thường:

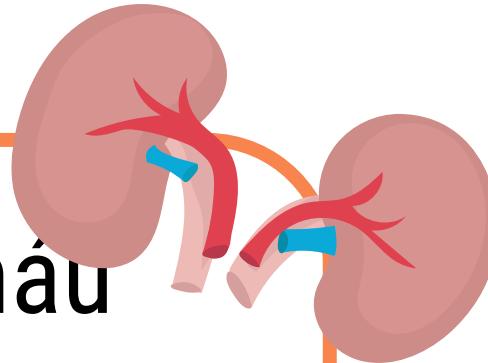
Nam là: 0,7 - 1,3 mg/dL (62 - 115 µmol/L)
Nữ là: 0,5 - 1,0 mg/dL (44 - 88 µmol/L)

Bình thường

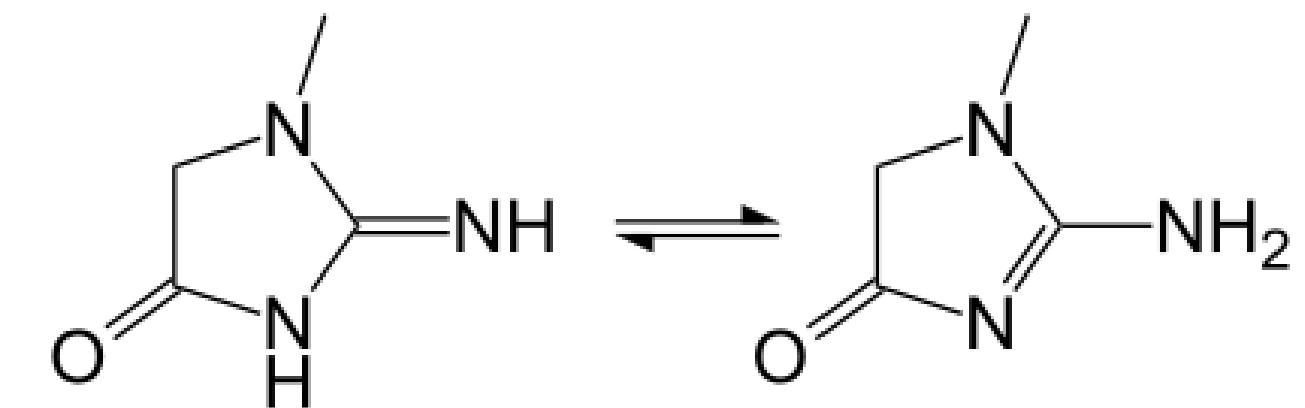
Thấp

Cao

Là chỉ số creatinin trong máu



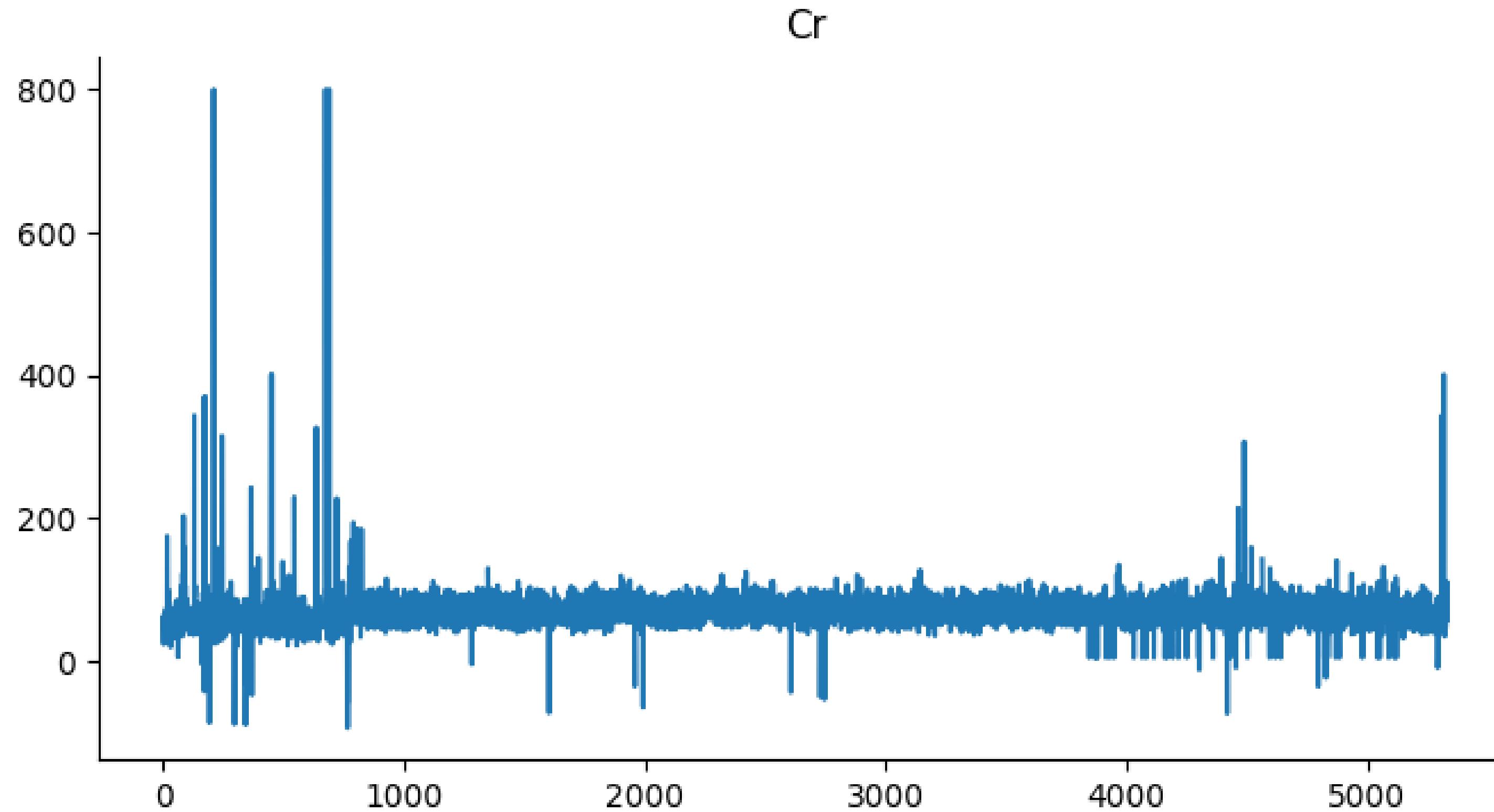
Creatinin là một chất cặn bã được đào thải thông qua thận, từ đó phản ánh chức năng thận



Thuộc tính ‘Cr’

mean	70.626207	Chỉ số creatinin trung bình
std	29.763911	Độ lệch chuẩn
min	-93.000000	Chỉ số creatinin thấp nhất
25%	57.750000	Tứ phân vị Q1
50%	70.000000	Giá trị trung vị (mean)
75%	81.400000	Tứ phân vị Q3
max	800.000000	Chỉ số creatinin cao nhất
Số lượng giá trị không null	← 5311 non-null →	Kiểu dữ liệu (datatype)

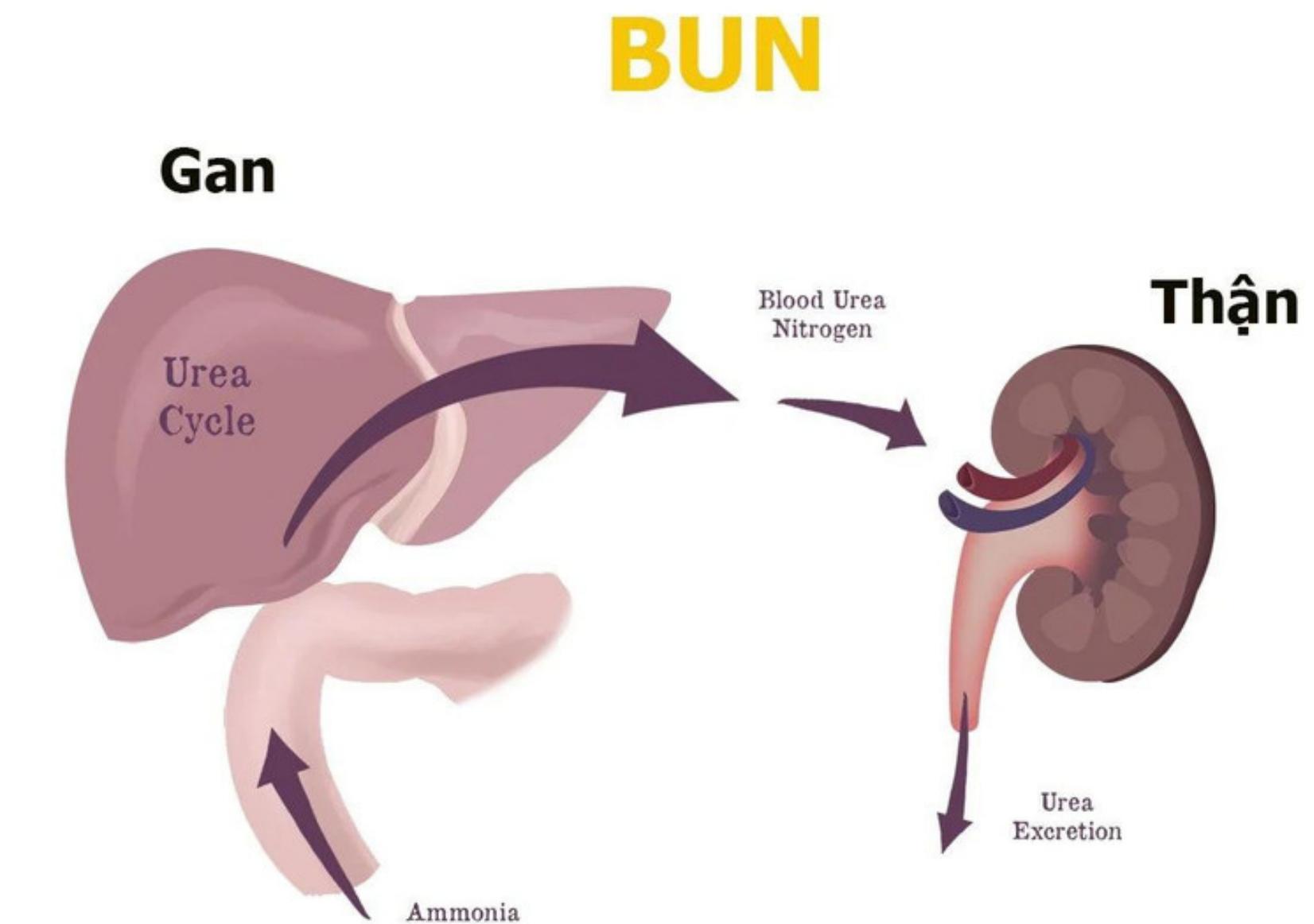
Thuộc tính 'Cr'



Thuộc tính 'BUN'

BUN (Blood Urea Nitrogen) là chỉ số BUN dùng để đánh giá chức năng gan và thận

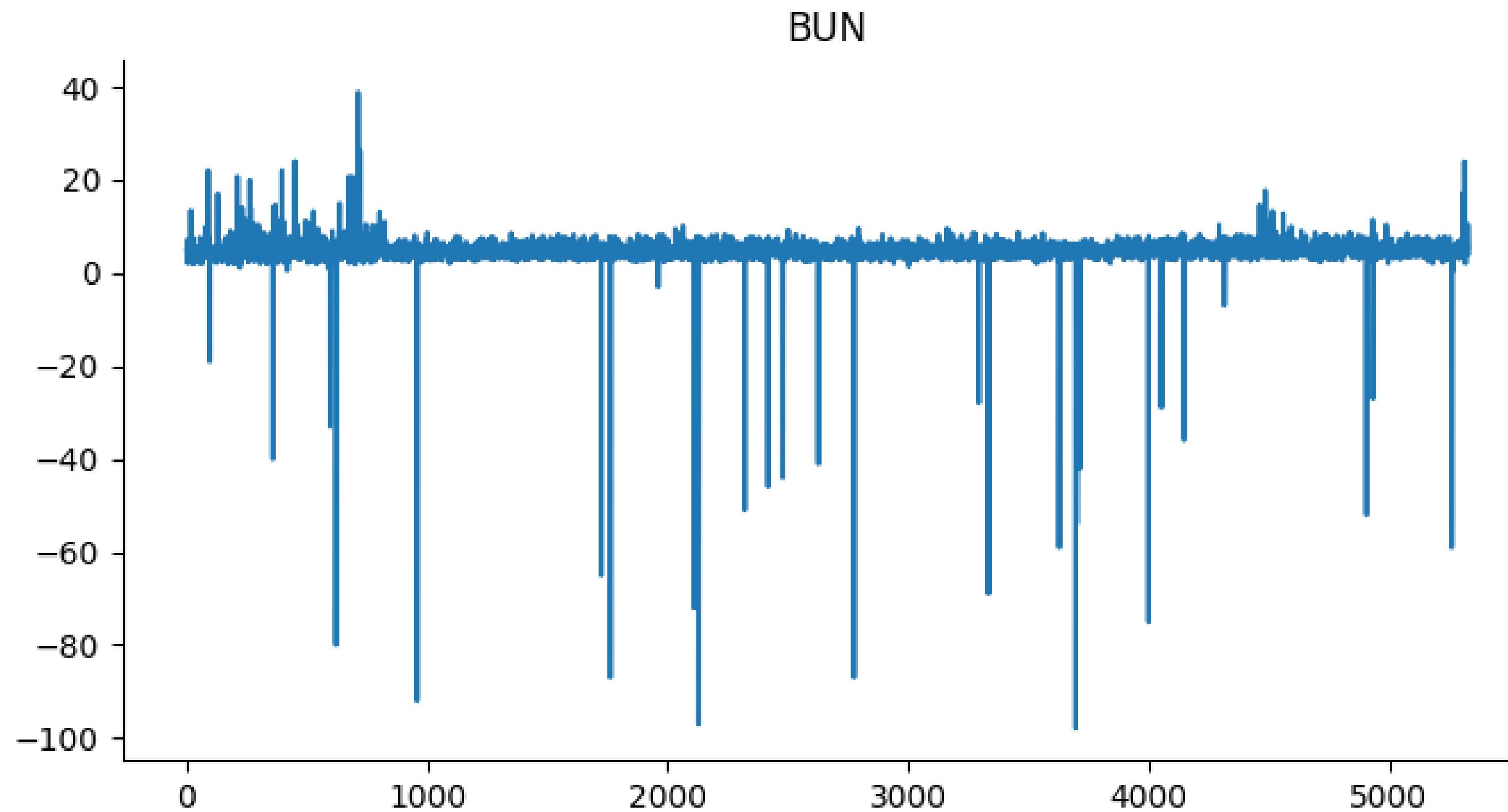
Xét nghiệm BUN sẽ cho biết nồng độ urea nitrogen trong máu đang ở mức bình thường hay bất thường



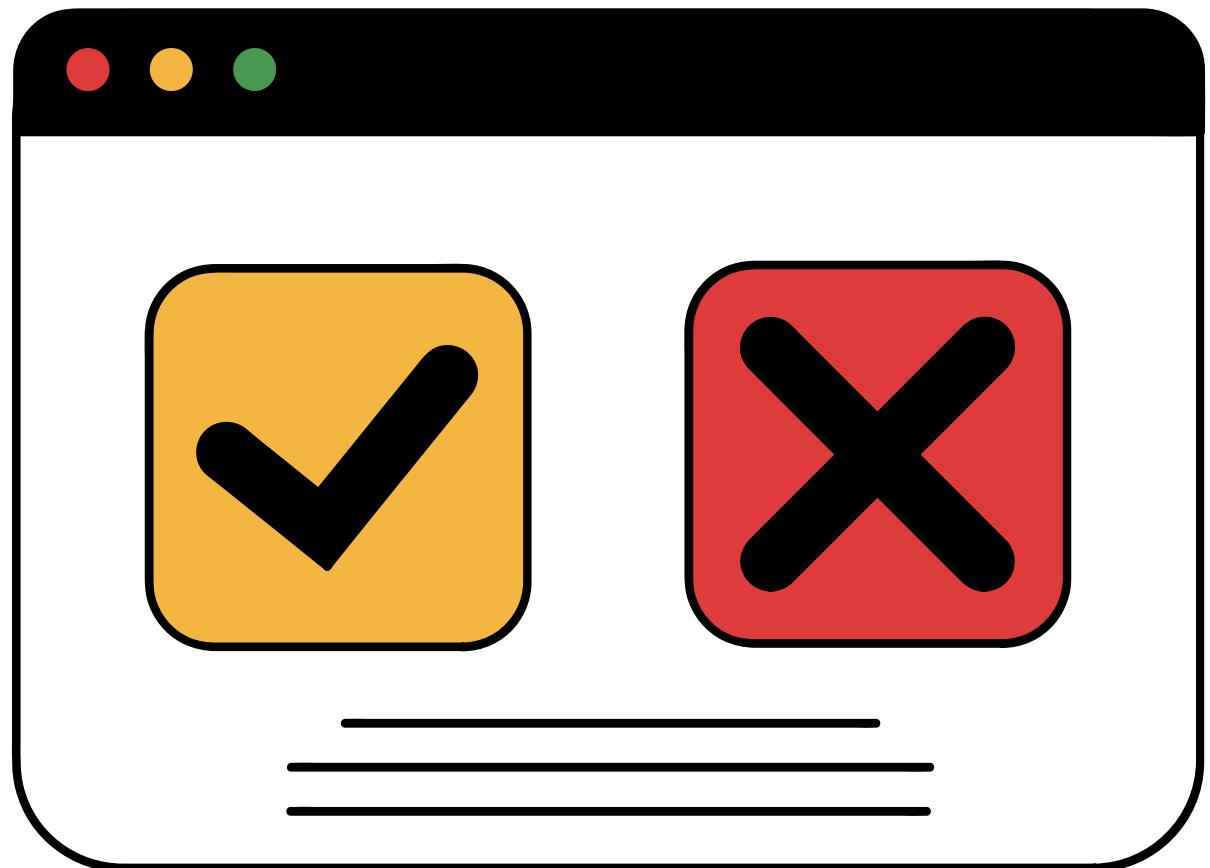
Thuộc tính ‘BUN’

mean	4.596416	Chỉ số BUN trung bình
std	4.919204	Độ lệch chuẩn
min	-98.000000	Chỉ số BUN thấp nhất
25%	3.900000	Tứ phân vị Q1
50%	4.710000	Giá trị trung vị (mean)
75%	5.600000	Tứ phân vị Q3
max	38.900000	Chỉ số BUN cao nhất
Số lượng giá trị không null	← 5311 non-null →	Kiểu dữ liệu (datatype)

Thuộc tính 'BUN'



Thuộc tính ‘Diagnosis’



Là lớp của từng bệnh nhân bao gồm 2 nhãn 0 và 1. Bệnh nhân được chẩn đoán mắc bệnh tiểu đường được gán nhãn là 1 và ngược lại có nhãn là 0.

Thuộc tính ‘Diagnosis’

mean	0.203087	Giá trị trung bình
std	3.366983	Độ lệch chuẩn
min	-83.000000	Gía trị thấp nhất
25%	0.000000	Tứ phân vị Q1
50%	0.000000	Giá trị trung vị (mean)
75%	1.000000	Tứ phân vị Q3
max	1.000000	Gia trị cao nhất

Số lượng giá trị
không null



5313 non-null

float64



Kiểu dữ liệu
(datatype)

oooo

#	Column	Non-Null Count	Dtype
0	Age	5310 non-null	float64
1	Gender	5332 non-null	object
2	BMI	5309 non-null	float64
3	Chol	5306 non-null	float64
4	TG	5300 non-null	float64
5	HDL	5310 non-null	float64
6	LDL	5314 non-null	float64
7	Cr	5311 non-null	float64
8	BUN	5311 non-null	float64
9	Diagnosis	5313 non-null	float64

Four small black circles are arranged horizontally in a row.

	Age	BMI	Chol	TG	HDL	LDL	Cr	BUN	Diagnosis
count	5310.000	5309.000	5306.000	5300.000	5310.000	5314.000	5311.000	5311.000	5313.000
mean	48.638	24.130	4.690	1.504	1.346	2.705	70.626	4.596	0.203
std	15.253	7.719	3.448	3.984	4.349	3.911	29.764	4.919	3.367
min	-94.000	-99.000	-82.000	-94.000	-95.000	-98.000	-93.000	-98.000	-83.000
25%	36.000	22.000	4.180	0.900	1.090	2.270	57.750	3.900	0.000
50%	49.000	24.000	4.800	1.370	1.300	2.780	70.000	4.710	0.000
75%	59.000	27.000	5.460	2.100	1.590	3.390	81.400	5.600	1.000
max	93.000	47.000	11.650	32.640	9.900	9.900	800.000	38.900	1.000

o o o o



```
1 print(df.isna().sum())
```



```
Age           22  
Gender        0  
BMI           23  
Chol          26  
TG            32  
HDL           22  
LDL           18  
Cr            21  
BUN           21  
Diagnosis     19  
dtype: int64
```



```
1 duplicated_rows = df[df.duplicated()]  
2 print(len(duplicated_rows))
```



```
200
```

.....



TIỀN XỬ LÝ DỮ LIỆU

CÁC TRƯỜNG HỢP CẦN XỬ LÝ

Chuẩn hóa thuộc tính ‘Gender’ về dạng binary F và M

Loại bỏ các dòng dữ liệu có giá trị không hợp lệ
(vd: giá trị âm)

Loại bỏ các dòng dữ liệu có giá trị rỗng (NaN)

Loại bỏ các dòng dữ liệu bị trùng lặp

CHUẨN

HÓA

THUỘC

TÍNH

'GENDER'

VỀ DẠNG

BINARY

```
df['Gender'] = df['Gender'].replace('f', 'F')
```

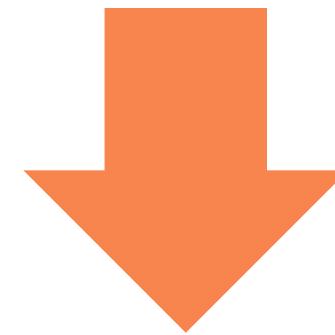
Thuộc tính 'Gender' sau khi chuẩn hóa:

Gender	Gender
M	3365
F	1966
f	1



**LOẠI BỎ
CÁC
DÒNG
DỮ LIỆU
BỊ
TRÙNG
LẶP**

```
1 df = df.drop_duplicates()  
2 df = df.reset_index()  
3 df = df.drop(df.columns[0], axis = 1)
```



```
1 duplicated_rows = df[df.duplicated()]  
2 print(len(duplicated_rows))
```

0 ✓

**LOẠI BỎ
CÁC
DÒNG
DỮ LIỆU
BỊ
TRÙNG
LẶP**

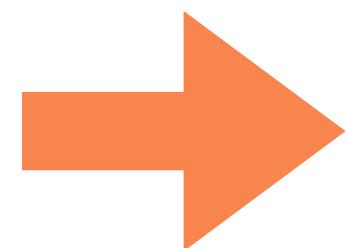
#	Column	Non-Null Count	Dtype
0	Age	5110 non-null	object
1	Gender	5132 non-null	object
2	BMI	5109 non-null	object
3	Chol	5107 non-null	object
4	TG	5101 non-null	object
5	HDL	5110 non-null	object
6	LDL	5114 non-null	object
7	Cr	5111 non-null	object
8	BUN	5112 non-null	object
9	Diagnosis	5114 non-null	object

**LOẠI BỎ
CÁC
DÒNG
DỮ LIỆU
CÓ
GIÁ TRỊ
RỖNG**

```
1 df = df.dropna()  
2 df = df.reset_index()  
3 df = df.drop(df.columns[0], axis = 1)
```

```
1 print(df.isna().sum())
```

Số lượng giá trị NaN
ở mỗi thuộc tính sau khi xử lý:



Age	0
Gender	0
BMI	0
Chol	0
TG	0
HDL	0
LDL	0
Cr	0
BUN	0
Diagnosis	0



**LOẠI BỎ
CÁC
DÒNG
DỮ LIỆU
có
GIÁ TRỊ
RỖNG**

#	Column	Non-Null Count	Dtype
0	Age	4933 non-null	object
1	Gender	4933 non-null	object
2	BMI	4933 non-null	object
3	Chol	4933 non-null	object
4	TG	4933 non-null	object
5	HDL	4933 non-null	object
6	LDL	4933 non-null	object
7	Cr	4933 non-null	object
8	BUN	4933 non-null	object
9	Diagnosis	4933 non-null	object

LOẠI BỎ

CÁC
DÒNG
DỮ LIỆU
CÓ
GIÁ TRỊ
KHÔNG
HỢP LỆ

```
1 df = df.drop(df[(df['Age'].astype(float) < 0) |  
2                   (df['BMI'].astype(float) < 0) |  
3                   (df['Chol'].astype(float) < 0) |  
4                   (df['TG'].astype(float) < 0) |  
5                   (df['HDL'].astype(float) < 0) |  
6                   (df['LDL'].astype(float) < 0) |  
7                   (df['Cr'].astype(float) < 0) |  
8                   (df['BUN'].astype(float) < 0) |  
9                   (df['Diagnosis'].astype(float) < 0)].index)  
10 df = df.reset_index()  
11 df = df.drop(df.columns[0], axis = 1)
```

LOẠI BỎ

CÁC

DÒNG

DỮ LIỆU

có

GIÁ TRỊ

KHÔNG

HỢP LỆ

#	Column	Non-Null Count	Dtype
0	Age	4743 non-null	float64
1	Gender	4743 non-null	object
2	BMI	4743 non-null	float64
3	Chol	4743 non-null	float64
4	TG	4743 non-null	float64
5	HDL	4743 non-null	float64
6	LDL	4743 non-null	float64
7	Cr	4743 non-null	float64
8	BUN	4743 non-null	float64
9	Diagnosis	4743 non-null	float64

LOẠI BỎ

CÁC

DÒNG

DỮ LIỆU

CÓ

GIÁ TRỊ

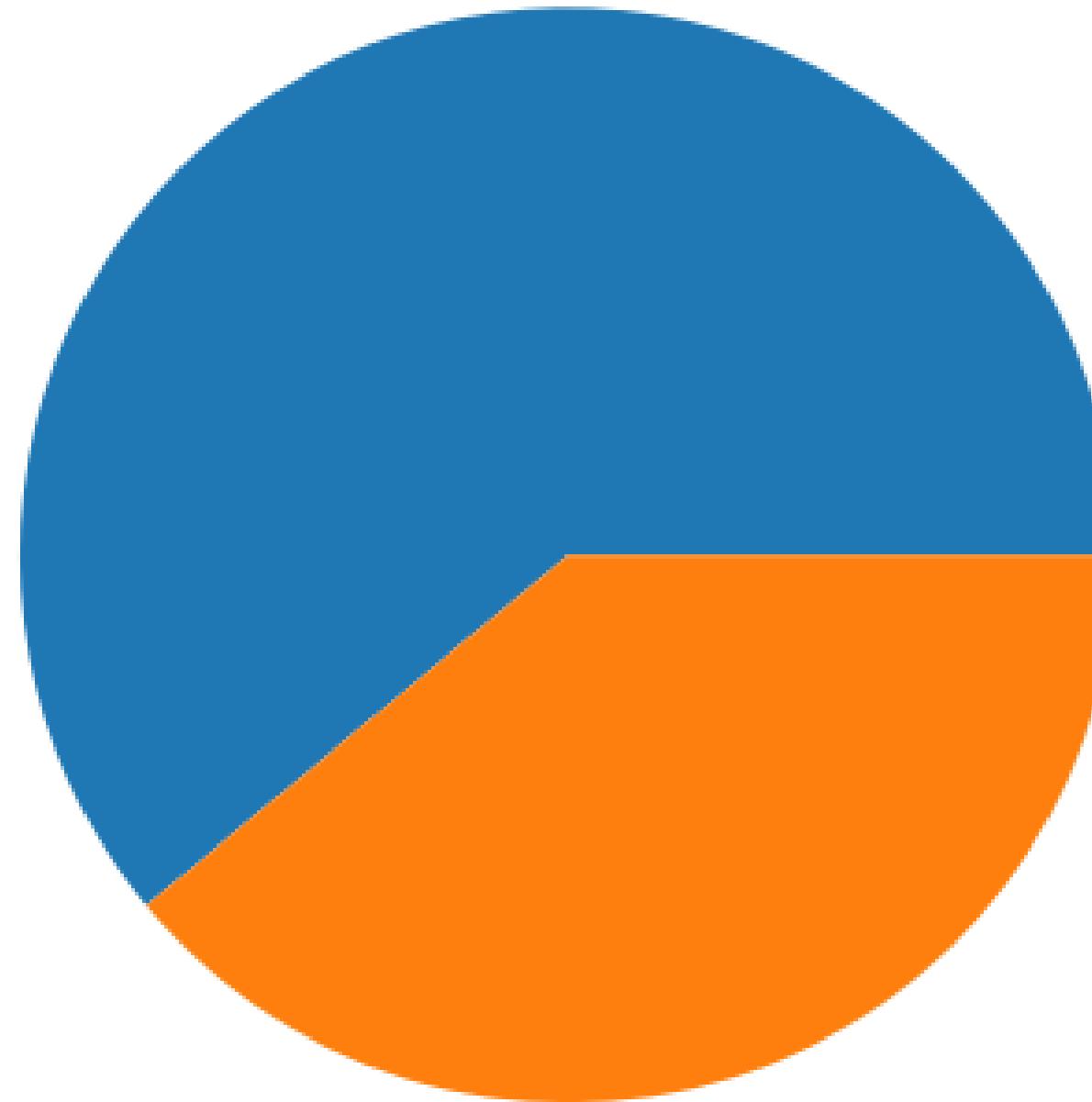
KHÔNG

HỢP LỆ

	Age	BMI	Chol	TG	HDL	LDL	Cr	BUN	Diagnosis
count	4743.000	4743.000	4743.000	4743.000	4743.000	4743.000	4743.000	4743.000	4743.000
mean	48.980	24.613	4.875	1.723	1.595	2.917	71.201	4.892	0.390
std	14.035	4.284	1.006	1.337	1.040	0.950	29.124	1.697	0.488
min	20.000	15.000	0.000	0.000	0.000	0.300	4.861	0.500	0.000
25%	36.000	22.000	4.200	0.910	1.100	2.290	58.000	3.910	0.000
50%	49.000	24.000	4.800	1.380	1.300	2.800	70.200	4.730	0.000
75%	59.000	27.000	5.470	2.100	1.590	3.400	81.700	5.580	1.000
max	93.000	47.000	11.650	32.640	9.900	9.900	800.000	38.900	1.000

KIỂM TRA SỰ CÂN BẰNG CỦA DỮ LIỆU

Diagnosis	
0.0	2895
1.0	1848



=>

oooo

+++++

SMOTE

dsdfghjkl

XỬ LÝ

```
df['Gender'] = df['Gender'].replace('F', 0)
df['Gender'] = df['Gender'].replace('M', 1)
```

VẤN ĐỀ

DỮ LIỆU

KHÔNG

```
from imblearn.over_sampling import SMOTE

X = df[['Age', 'Gender', 'BMI', 'Chol', 'TG', 'HDL', 'LDL', 'Cr', 'BUN']]
smote = SMOTE(k_neighbors=5)
df_resampled, label = smote.fit_resample(X, df['Diagnosis'])
```

CÂN

BẰNG

XỬ LÝ

VĂN ĐỀ

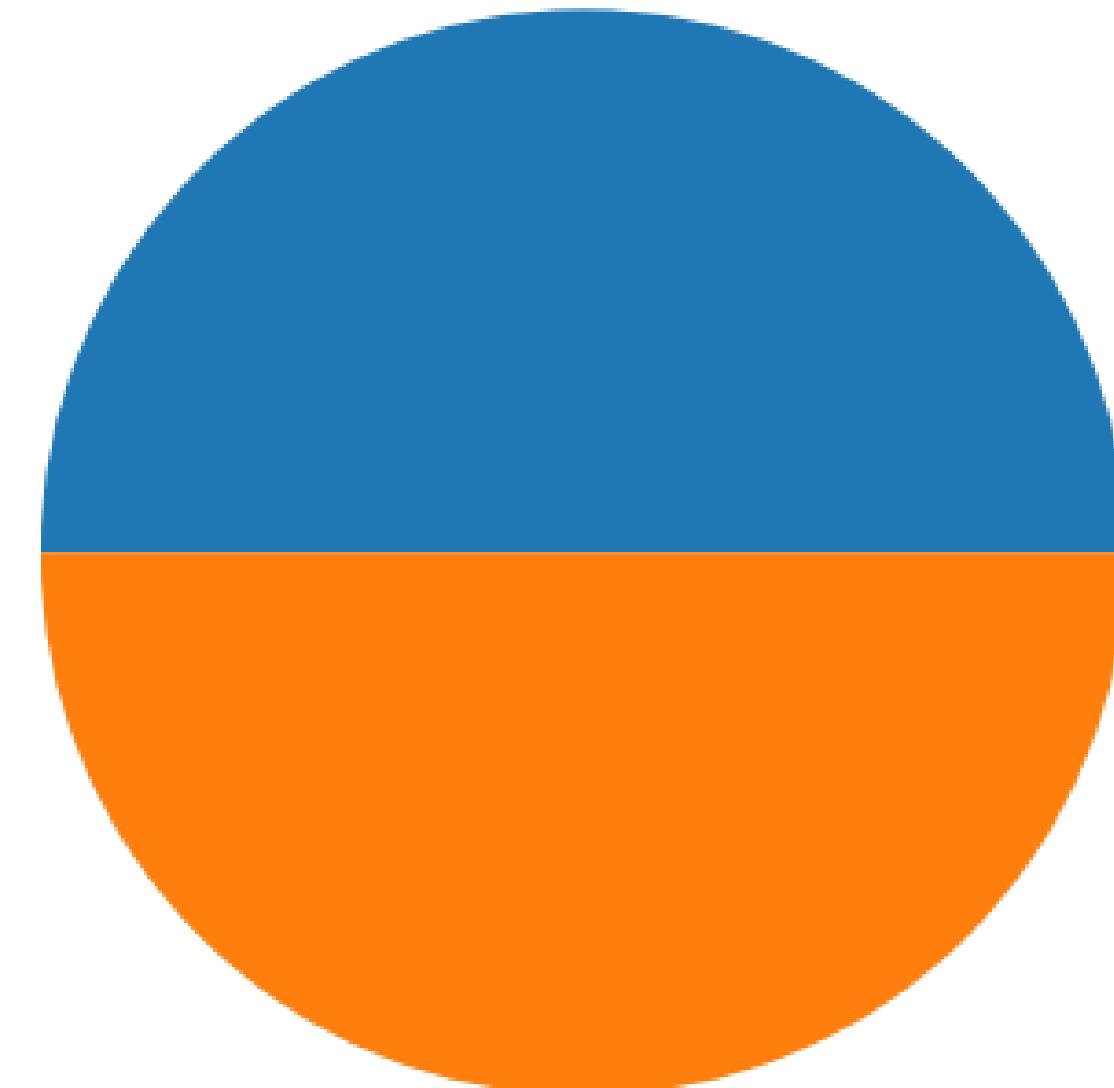
DỮ LIỆU

KHÔNG

CÂN

BẰNG

Diagnosis	
0.0	2895
1.0	2895



ooo

DỮ LIỆU

SAU KHI TIỀN XỬ LÝ (LÀM SẠCH)



#	Column	Non-Null Count	Dtype
0	Age	5790 non-null	int64
1	Gender	5790 non-null	int64
2	BMI	5790 non-null	float64
3	Chol	5790 non-null	float64
4	TG	5790 non-null	float64
5	HDL	5790 non-null	float64
6	LDL	5790 non-null	float64
7	Cr	5790 non-null	float64
8	BUN	5790 non-null	float64