1. Định nghĩa vấn đề

Bộ dữ liệu về chất lượng rượu vang đỏ (Wine Quality - Red Wine) được thu thập từ quá trình phân tích các chỉ số hóa học của rượu vang và đánh giá chất lượng bởi chuyên gia.

Mục tiêu là phân tích thống kê mô tả để hiểu rõ đặc điểm của các biến đầu vào và biến mục tiêu, từ đó làm cơ sở cho các bước phân tích và mô hình hóa sau này.

Dữ liệu vào (Input features):

- **Fixed acidity**: Nồng độ axit cố định (g/dm³)
- **Volatile acidity**: Nồng độ axit bay hơi (g/dm³)
- **Citric acid**: Hàm lượng axit citric (g/dm³)
- **Residual sugar**: Lượng đường còn dư (g/dm³)
- **Chlorides**: Nồng độ chloride (g/dm³)
- **Free sulfur dioxide**: Lượng SO_2 tự do (mg/dm³)
- **Total sulfur dioxide**: Tổng lượng SO_2 (mg/dm³)
- **Density**: Khối lượng riêng (g/cm³)
- **pH**: Chỉ số pH
- **Sulphates**: Nồng độ sulphate (g/dm³)
- **Alcohol**: Nồng độ cồn (% thể tích)

Kết quả (Output):

- **Quality**: Điểm đánh giá chất lượng rượu (giá trị nguyên từ 0–10, càng cao thì chất lượng càng tốt)

2. Đọc và hiểu dữ liệu

2.1. Import thư iện cần thiết

Nhấp đúp (hoặc nhấn Enter) để chỉnh sửa

ỌC VÀ HIỂU DỮ LIỆU

2.1. Import thư viện cần thiết

```
import pandas as pd
import numbers as pd
```

```
from IPython import display
```

2.2. Tải dữ liệu

```
import pandas as pd
df = pd. read_csv("winequality-red.csv")
X = df.drop("quality", axis=1)
y = df["quality"]
```

3. Phân tích dữ liệu

3.1. Thống kê mô tả

(1) Thông tin chung về dữ liệu

- Số dòng và số cột trong dữ liệu.
- Kiểu dữ liệu của từng biến.
- Ý nghĩa và đơn vị đo lường của các biến:
  - **fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, sulphates**: g/dm³
  - **free sulfur dioxide, total sulfur dioxide**: mg/dm³
  - **density**: g/cm³
  - **pH**: chỉ số pH (không có đơn vị)
  - **alcohol**: % (phần trăm thể tích)
  - **quality**: điểm đánh giá chất lượng rượu (thang đo rời rạc)

(2) Kiểm tra chất lượng dữ liệu

- Kiểm tra dữ liệu trùng lặp.
- Kiểm tra giá trị Null/NaN.

(3) Thống kê tóm tắt các biến số

- Các thước đo thống kê cơ bản: giá trị nhỏ nhất, lớn nhất, trung bình, trung vị, độ lệch chuẩn, các phân vị.
- Bổ sung median vì describe() không mặc định tính.

(4) Phân bố biến phân loại (quality)

- Tính tần số xuất hiện của từng mức chất lượng rượu.
- Tính thêm tỷ lệ phần trăm để thấy mức độ phân bố.

(5) Nhận xét sơ bộ

- So sánh giá trị trung bình và độ biến thiên giữa các biến.
- Nhận diện các biến có khoảng dao động rộng hoặc khả năng có ngoại lệ.
- Đặc điểm phân bố của biến mục tiêu quality.

```
# (1) Thông tin chung
print("Kích thước dữ liệu:", df.shape)
print("\nKiểu dữ liệu:\n", df.dtypes)
print("\n5 dòng đầu:",df.head() )
print("\n5 dòng cuối:",df.tail())
print("\nThông tin tổng quan:")
df.info()

# (2) Kiểm tra chất lượng dữ liệu
print("\n--- Kiểm tra dữ liệu ---")
print("Số dòng trùng lặp:", df.duplicated().sum())
print("Có giá trị Null:", df.isnull().sum().any())
print("Có giá trị NaN:", df.isna().sum().any())

# (3) Thống kê tóm tắt các biến số
print("\n--- Bảng thống kê describe() ---", df.describe().T)

print("\n--- Median của các biến ---", df.median(numeric_only=True))

# (4) Phân bố biến phân loại (quality)
print("\n--- Tần số xuất hiện của quality ---", df['quality'].value_counts().sort_index(

print("\n--- Tỷ lệ phần trăm của quality ---", df['quality'].value_counts(normalize=True
```

```
Kích thước dữ liệu: (1599, 12)

Kiểu dữ liệu:
 fixed acidity          float64
volatile acidity       float64
citric acid            float64
residual sugar         float64
chlorides              float64
free sulfur dioxide    float64
total sulfur dioxide   float64
density                float64
pH                     float64
sulphates              float64
alcohol                float64
quality                  int64
dtype: object

5 dòng đầu:    fixed acidity  volatile acidity  citric acid  residual sugar  chlo
0            7.4              0.70         0.00            1.9      0.076
1            7.8              0.88         0.00            2.6      0.098
2            7.8              0.76         0.04            2.3      0.092
3           11.2              0.28         0.56            1.9      0.075
```

```
4            7.4             0.70         0.00            1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5

5 dòng cuối:       fixed acidity  volatile acidity  citric acid  residual sugar
1594            6.2             0.600         0.08             2.0      0.090
1595            5.9             0.550         0.10             2.2      0.062
1596            6.3             0.510         0.13             2.3      0.076
1597            5.9             0.645         0.12             2.0      0.075
1598            6.0             0.310         0.47             3.6      0.067

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
1594                 32.0                  44.0  0.99490  3.45       0.58
1595                 39.0                  51.0  0.99512  3.52       0.76
1596                 29.0                  40.0  0.99574  3.42       0.75
1597                 32.0                  44.0  0.99547  3.57       0.71
1598                 18.0                  42.0  0.99549  3.39       0.66

      alcohol  quality
1594     10.5        5
1595     11.2        6
1596     11.0        6
1597     10.2        5
```

- Dữ liệu có 11 tính chất để phân lớp: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide,total sulfur dioxide, density, pH, sulphates, alcohol.
- Đơn vị nồng độ: Hầu hết các thành phần hóa học (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, sulphates) đều được đo bằng g/dm3.
- SO2: Các chỉ số sulfur dioxide thường được đo bằng mg/dm3 vì chúng thường có nồng độ thấp hơn nhiều.
- Kích thước tập dữ liệu gồm 1599 hàng và 12 cột
- Dữ liệu để phân lớp ở cột quality

(2) Kiểm tra tính toàn vẹn của dử liệu

- Dữ liệu có bị trùng lặp không? Hiển thị dòng bị vi phạm.
- Dữ liệu có tồn tại giá trị Null không? Hiển thị dòng bị vi phạm.

- Dữ liệu có tồn tại giá trị NaN không? Hiển thị dòng bị vi phạm

```
has_null = df.isnull().sum().any()
has_nan  = df.isna().sum().any()
n_duplicated = df.duplicated().sum()
print(f'Tính toàn vẹn dữ liệu:')
print(f'+ Có giá trị Null: {has_null}')
if has_null:
    display.display(df[df.isnull().any(axis=1)])
print(f'+ Có giá trị Nan: {has_nan}')
if has_nan:
    display.display(df[df.isna().any(axis=1)])
print(f'+ Số dòng trùng: {n_duplicated}')
if n_duplicated>0:
    display.display(df[df.duplicated()])
```

```
Tính toàn vẹn dữ liệu:
+ Có giá trị Null: False
+ Có giá trị Nan: False
+ Số dòng trùng: 240
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	p
4	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.5
11	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.99780	3.3
27	7.9	0.430	0.21	1.60	0.106	10.0	37.0	0.99660	3.1
40	7.3	0.450	0.36	5.90	0.074	12.0	87.0	0.99780	3.3
65	7.2	0.725	0.05	4.65	0.086	4.0	11.0	0.99620	3.4
...	...	...	...	...	...	...	...	...	
1563	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.2
1564	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.2
1567	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.2
1581	6.2	0.560	0.09	1.70	0.053	24.0	32.0	0.99402	3.5
1596	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.4

** Nhận xét:**

- Dự liệu có 240 dòng bị trùng
- DỰ liệu không có giá trị bị rỗng

Thống kê mô tả tổng quát

```
import pandas as pd


# --- 1. Tính toán các chỉ số thống kê cơ bản và tùy chỉnh ---

# Tính Mean, Median, Variance, và Standard Deviation
basic_stats = df.agg(['mean', 'median', 'var', 'std']).T
basic_stats = basic_stats.rename(columns={'mean': 'Mean', 'median': 'Median', 'var': 'Va

# Tính Min, Max, Quartiles (Q1, Q3)
desc_stats = df.describe().T[['min', 'max', '25%', '75%']]
desc_stats = desc_stats.rename(columns={'min': 'Min', 'max': 'Max', '25%': 'Q1 (25th Pct

# Tính Range (Phạm vi)
desc_stats['Range'] = desc_stats['Max'] - desc_stats['Min']

# Tính IQR (Interquartile Range: Q3 - Q1)
desc_stats['IQR'] = desc_stats['Q3 (75th Pct)'] - desc_stats['Q1 (25th Pct)']

# Tính thêm Percentiles (ví dụ: 10th và 90th percentile)
percentiles = df.quantile([0.1, 0.9]).T
percentiles = percentiles.rename(columns={0.1: '10th Percentile', 0.9: '90th Percentile'

# Tính Mode (Yếu vị)
# Lấy giá trị mode đầu tiên nếu có nhiều hơn một mode
mode_series = df.mode().iloc[0]
mode_df = mode_series.to_frame(name='Mode').T.rename(index={0: 'Mode'})

# --- 2. Kết hợp tất cả các kết quả vào một DataFrame duy nhất ---

# Kết hợp các bảng
final_stats = pd.merge(basic_stats, desc_stats, left_index=True, right_index=True)
final_stats = pd.merge(final_stats, percentiles, left_index=True, right_index=True)
final_stats = pd.merge(final_stats, mode_df.T, left_index=True, right_index=True)

# Sắp xếp lại và làm tròn các cột cho dễ đọc
column_order = [
    'Mean', 'Median', 'Mode', 'Standard Deviation', 'Variance',
    'Min', 'Max', 'Range', 'Q1 (25th Pct)', 'Q3 (75th Pct)', 'IQR',
    '10th Percentile', '90th Percentile'
]
final_stats = final_stats[column_order]

# --- 3. Hiển thị kết quả ---
print("### Bảng Tổng Hợp Thống Kê Mô Tả Toàn Diện ###")
# Hiển thị kết quả, làm tròn 3 chữ số thập phân
print(final_stats.round(3))
```

```
### Bảng Tổng Hợp Thống Kê Mô Tả Toàn Diện ###
                      Mean  Median    Mode  Standard Deviation  Variance  \
fixed acidity        8.320   7.900   7.200               1.741     3.031
```

```
                      volatile acidity       0.528   0.520   0.600              0.179     0.032
                      citric acid            0.271   0.260   0.000              0.195     0.038
                      residual sugar         2.539   2.200   2.000              1.410     1.988
                      chlorides              0.087   0.079   0.080              0.047     0.002
                      free sulfur dioxide   15.875  14.000   6.000             10.460   109.415
                      total sulfur dioxide  46.468  38.000  28.000             32.895  1082.102
                      density                0.997   0.997   0.997              0.002     0.000
                      pH                     3.311   3.310   3.300              0.154     0.024
                      sulphates              0.658   0.620   0.600              0.170     0.029
                      alcohol               10.423  10.200   9.500              1.066     1.136
                      quality                5.636   6.000   5.000              0.808     0.652

                                            Min      Max    Range  Q1 (25th Pct)  Q3 (75th Pct)  \
                      fixed acidity         4.600   15.900   11.300          7.100          9.200
                      volatile acidity      0.120    1.580    1.460          0.390          0.640
                      citric acid           0.000    1.000    1.000          0.090          0.420
                      residual sugar        0.900   15.500   14.600          1.900          2.600
                      chlorides             0.012    0.611    0.599          0.070          0.090
                      free sulfur dioxide   1.000   72.000   71.000          7.000         21.000
                      total sulfur dioxide  6.000  289.000  283.000         22.000         62.000
                      density               0.990    1.004    0.014          0.996          0.998
                      pH                    2.740    4.010    1.270          3.210          3.400
                      sulphates             0.330    2.000    1.670          0.550          0.730
                      alcohol               8.400   14.900    6.500          9.500         11.100
                      quality               3.000    8.000    5.000          5.000          6.000

                                              IQR  10th Percentile  90th Percentile
                      fixed acidity         2.100            6.500           10.700
                      volatile acidity      0.250            0.310            0.745
                      citric acid           0.330            0.010            0.522
                      residual sugar        0.700            1.700            3.600
                      chlorides             0.020            0.060            0.109
                      free sulfur dioxide  14.000            5.000           31.000
                      total sulfur dioxide 40.000           14.000           93.200
                      density               0.002            0.995            0.999
                      pH                    0.190            3.120            3.510
                      sulphates             0.180            0.500            0.850
                      alcohol               1.600            9.300           12.000
                      quality               1.000            5.000            7.000
```

Trung tâm dữ liệu:

- Fixed acidity trung bình ≈ 8.32, alcohol ≈ 10.42.
- Quality trung vị = 6 ⟶ phần lớn rượu chất lượng trung bình.
- Citric acid mode = 0 ⟶ nhiều rượu không có citric acid.

Biến thiên:

- Total sulfur dioxide SD = 32.9 ⟶ phân tán lớn, nhiều outlier.
- Density SD = 0.002 ⟶ hầu hết rượu đồng nhất về mật độ.
- Residual sugar range = 14.6 ⟶ có rượu ngọt cực đoan.

Phân bố:

- Alcohol IQR = 1.6 ⟶ 50% rượu nằm trong 9.5–11.1 độ cồn.
- Free sulfur dioxide 90th percentile = 31 ⟶ rượu chứa SO_2 cao.
- pH ổn định, SD = 0.154.

Nhận xét:

- Độ axit lệch phải, nhiều rượu axit cao.
- Đường, SO_2 phân bố lệch phải, xuất hiện giá trị cực đoan.
- Density đồng đều, alcohol biến thiên vừa phả

Nhấp đúp (hoặc nhấn Enter) để chỉnh sửa