TRƯỜNG ĐẠI HỌC SÀI GÒN KHOA CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH KHÁM PHÁ VỀ CHẤT LƯỢNG RƯỢU VANG ĐỎ

Sinh viên thực hiện: Lương Thanh Tuấn

Giảng viên hướng dẫn : TS.Đỗ Ngọc Tài

TP. HỒ CHÍ MINH, 10/2025

Nội dung chính

1. Giới thiệu tổng quát về dữ liệu

2. Phân tích khám phá

3. Kết luận

1. Giới thiệu tổng quát về dữ liệu

- Bộ dữ liệu **“Red Vinho Verde”** (Bồ Đào Nha) thuộc bộ Wine Quality gốc UCI, được công bố trên UCI Machine Learning Repository và mirror trên Kaggle và được hiến tặng (06/10/2009)
- Mục tiêu là mô hình hóa chất lượng rượu vang dựa trên các thử nghiệm hóa lý

UCI Machine Learning Repository

1. Giới thiệu tổng quát về dữ liệu

- Bộ dữ liệu gồm 12 cột, trong đó có 11 thuộc tính features và thuộc tính quality (target) và 1599 dòng. Ngoài ra tất cả các cột đều có kiểu dữ liệu số.

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed_acidity         1599 non-null   float64
 1   volatile_acidity      1599 non-null   float64
 2   citric_acid           1599 non-null   float64
 3   residual_sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free_sulfur_dioxide   1599 non-null   float64
 6   total_sulfur_dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
```

- Bộ dữ liệu không có giá trị Null, NaN. Tuy nhiên có tới 240 dòng bị trùng.

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

1. Giới thiệu tổng quát về dữ liệu

Nhiều biến có khả năng lệch phải (Mean > Median):

- fixed_acidity (7.9 < 8.32), volatile_acidity (0.52 < 0.528), citric_acid (0.26 < 0.27), residual_sugar (2.2 < 2.54), chlorides (0.08 < 0.09), free_SO2 (14 < 15.9), total_SO2 (38 < 46.47), sulphates (0.62 < 0.66), alcohol (10.2 < 10.42)

Nhiều biến có khả năng lệch trái (Mean < Median):

- quality (6 > 5.63)

	count	mean	std	min	25%	50%	75%	max
fixed_acidity	1599.0	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000
volatile_acidity	1599.0	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000
citric_acid	1599.0	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000
residual_sugar	1599.0	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100
free_sulfur_dioxide	1599.0	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000
total_sulfur_dioxide	1599.0	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000
density	1599.0	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369
pH	1599.0	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000
quality	1599.0	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000

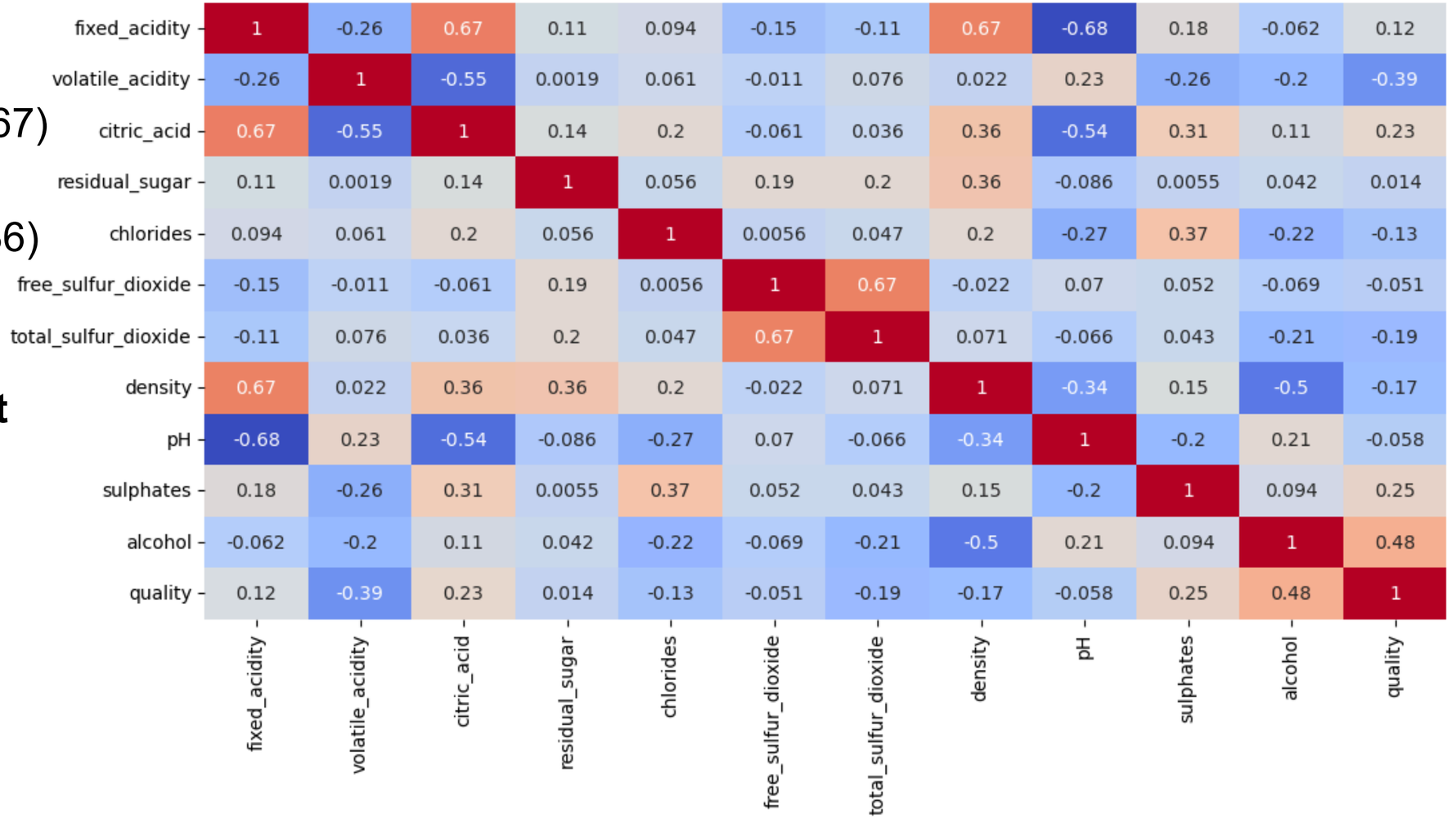
2. Phân tích khám phá

- Có nhiều cột tương quan dương rõ rệt như:

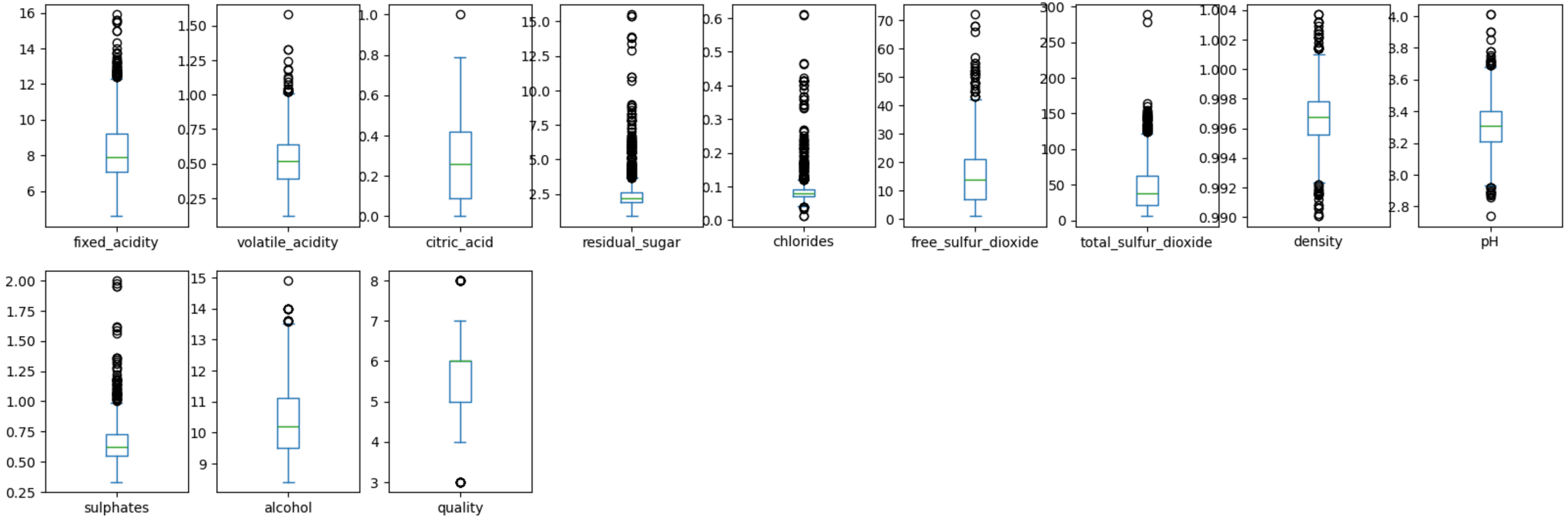
- fixed_acidity - citric_acid (0.67)
- fixed_acidity - density (0.67)
- residual_sugar - density (0.36)
- totalSO2 - freeSO2 (0.67)
- alocohol - quality (0.48)

- Bên cạnh đó cũng có các cột tương quan âm:

- fixed_acidity - pH (-0.68)
- ctric_acid - volatile_acidity (-0.55)
- ctric_acid - pH (-0.54)
- density - alcohol (-0.5)

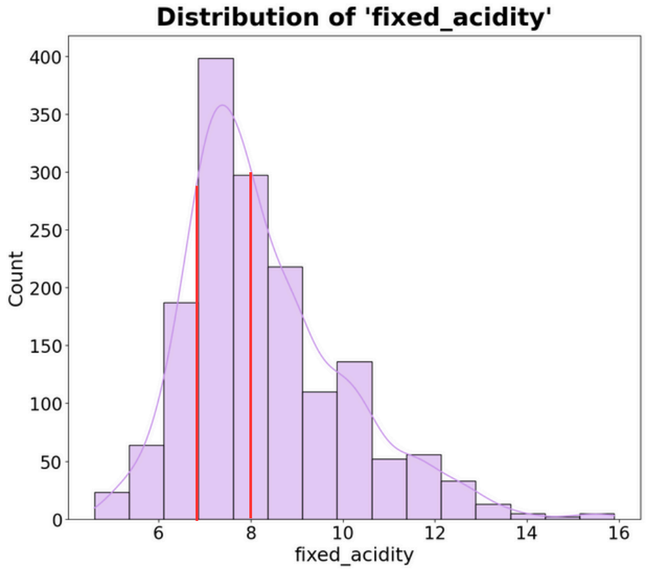


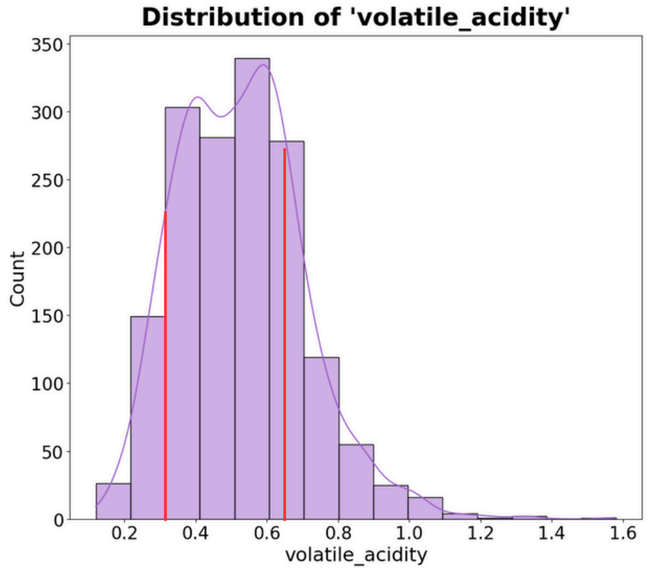
2. Phân tích khám phá

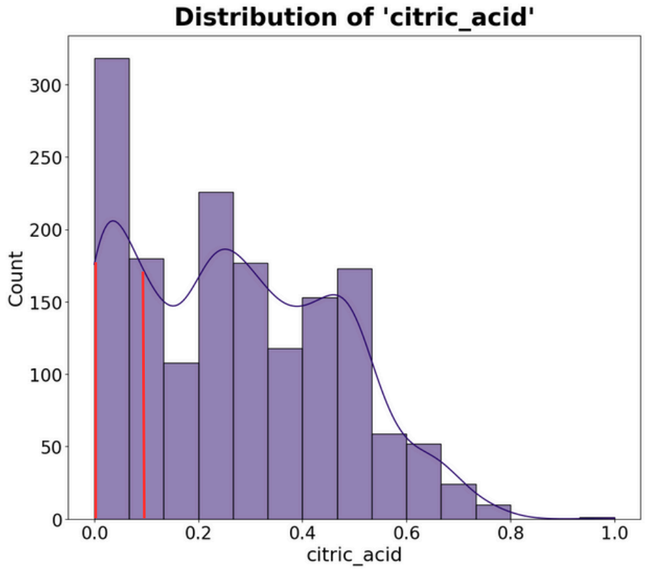


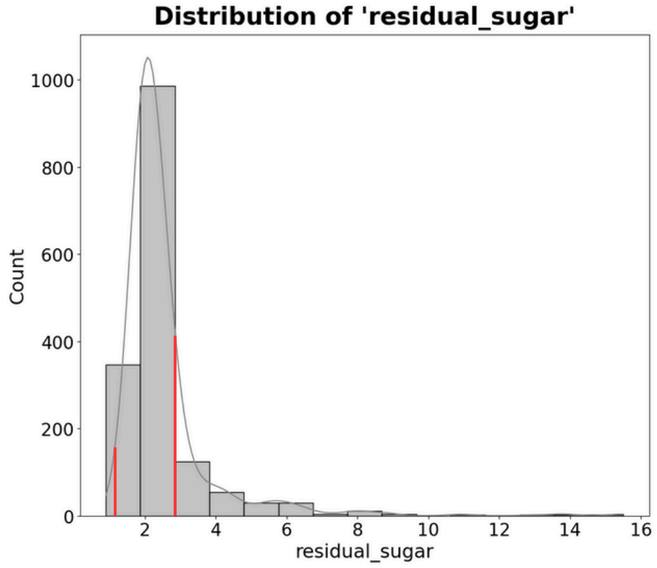
- **Ngoại trử citric_acid ra các yếu tố khác đều có nhiều giá trị outlier, khiến chúng bị lệch phải và trái**

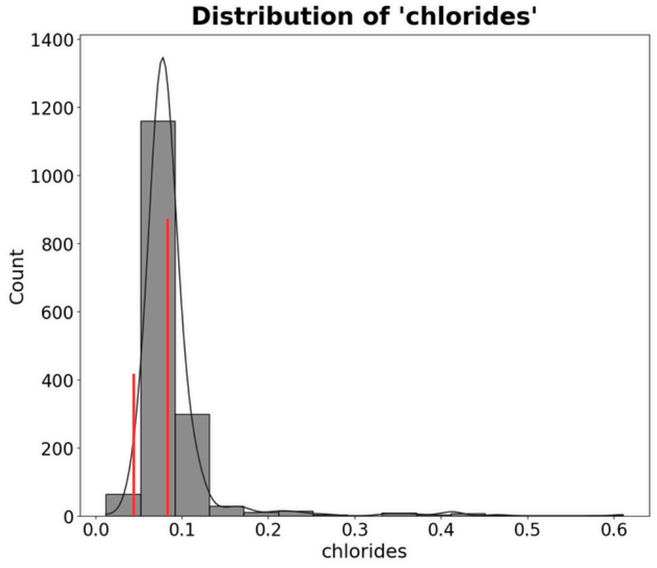
2. Phân tích khám phá

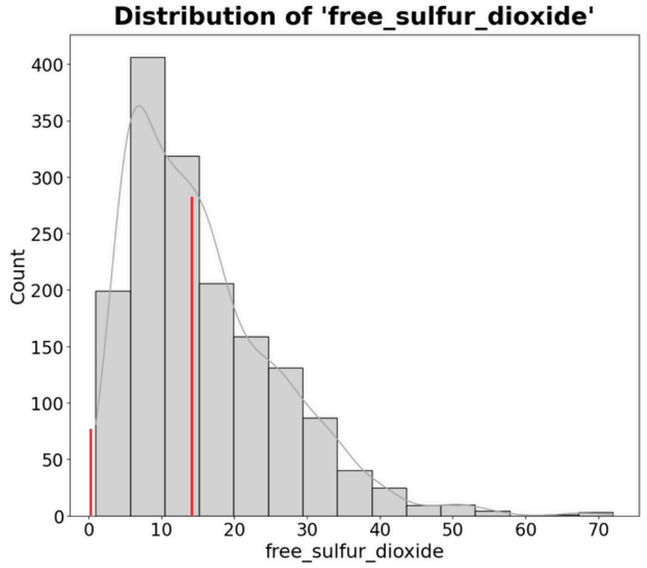




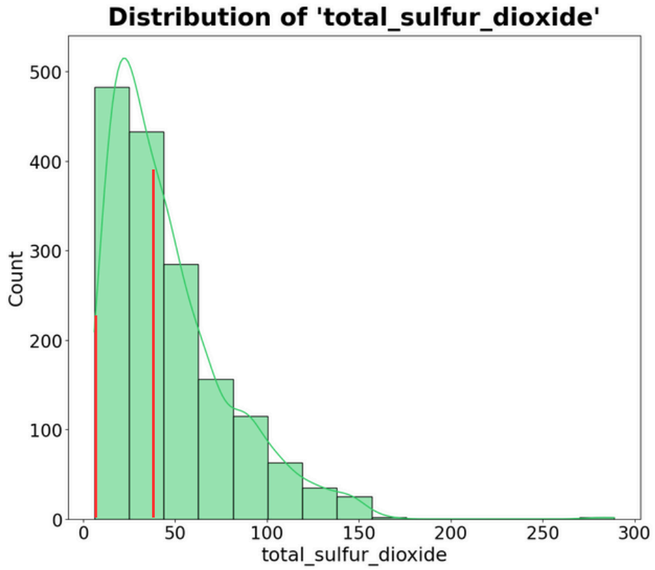


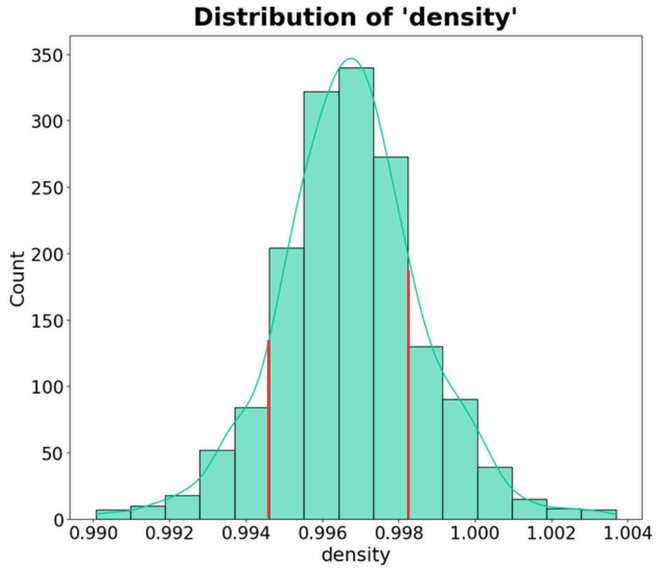


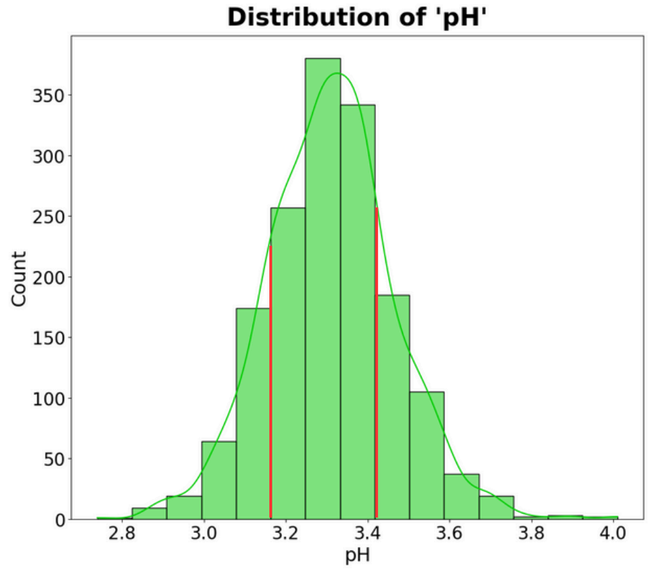


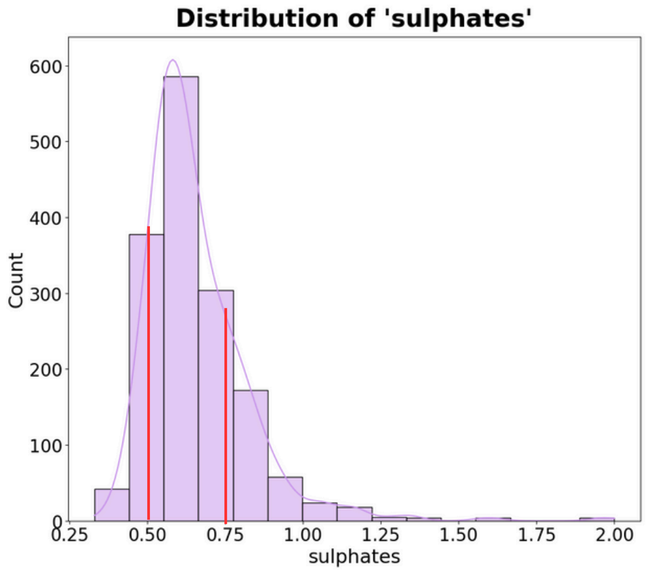


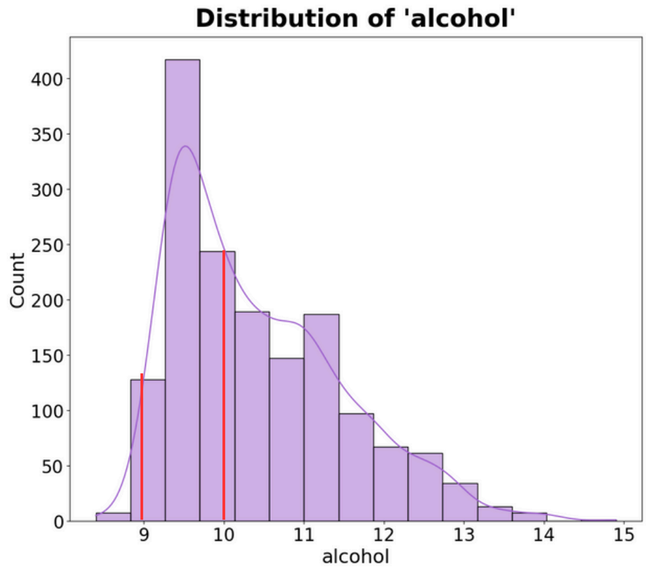
2. Phân tích khám phá

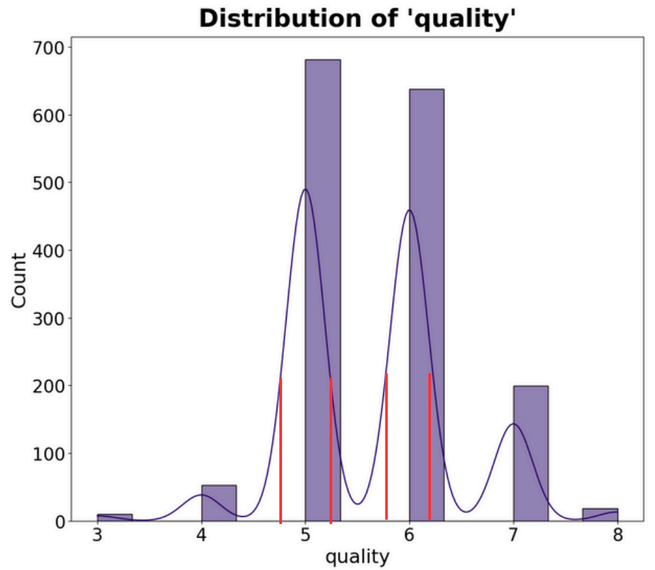




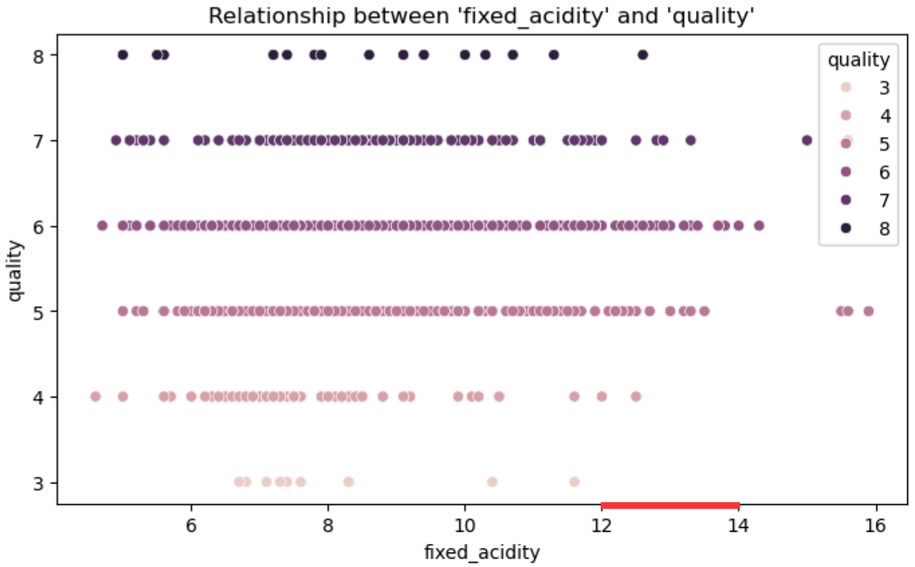


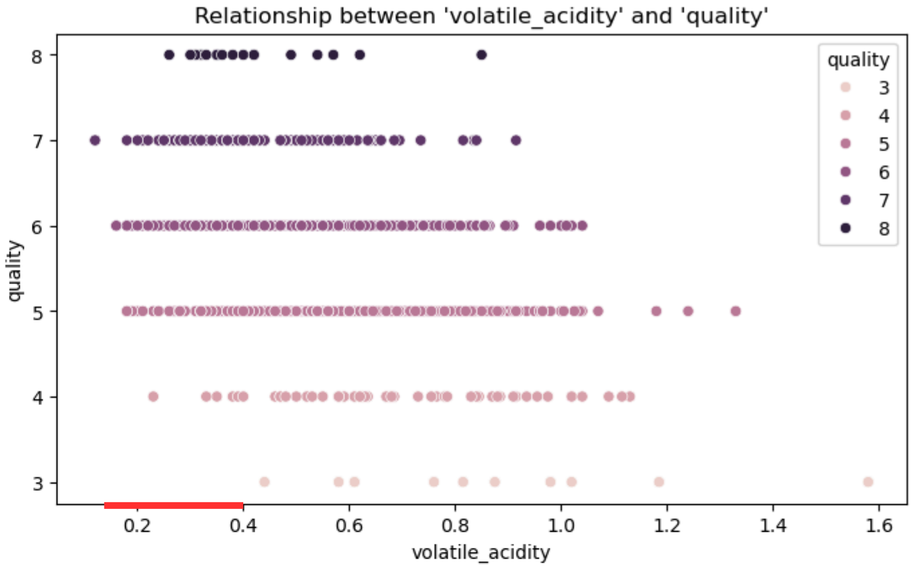


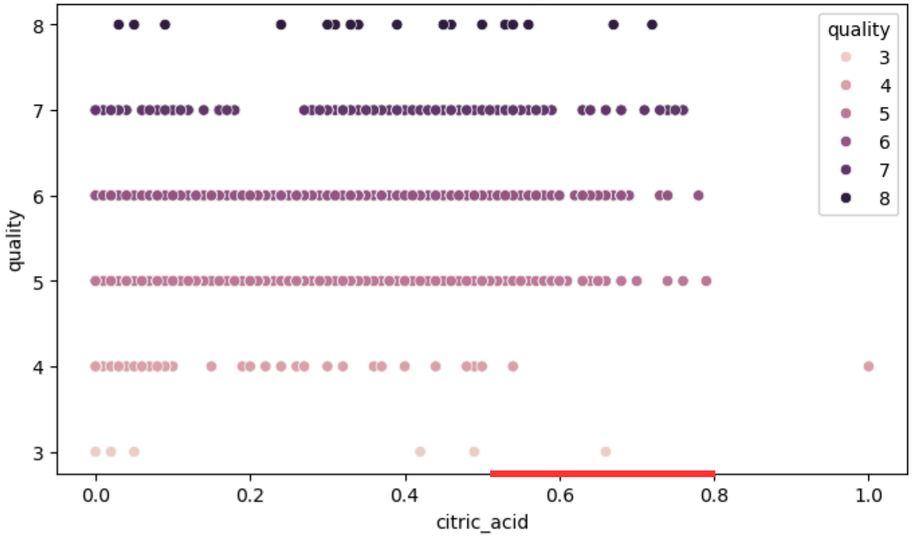


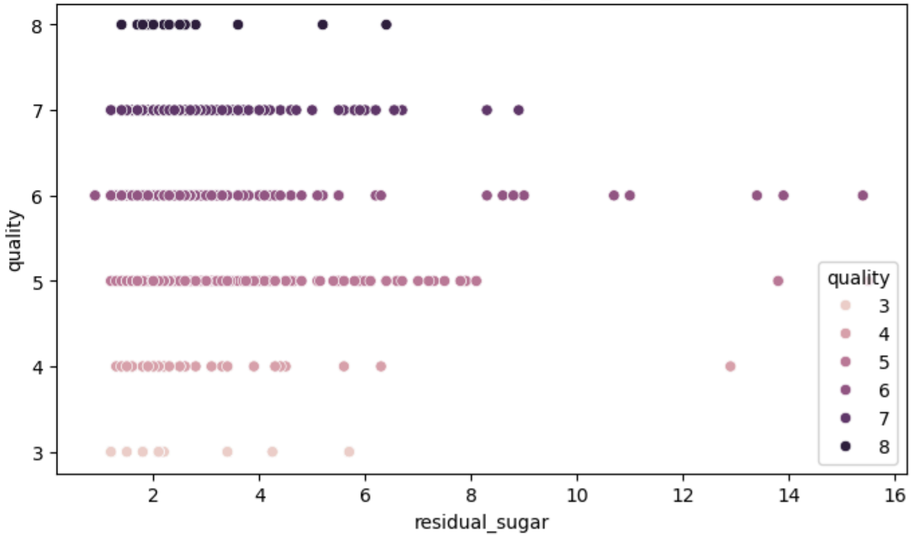


2. Phân tích khám phá

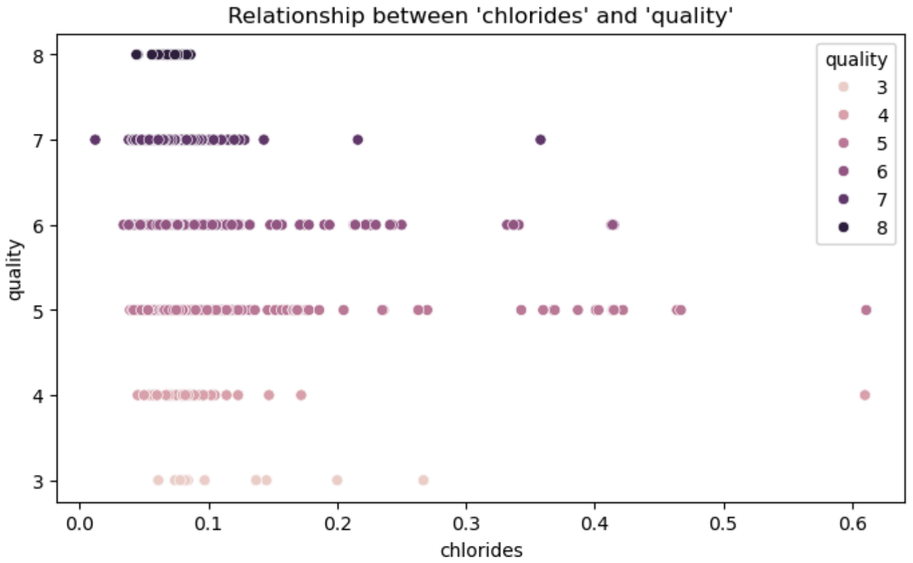


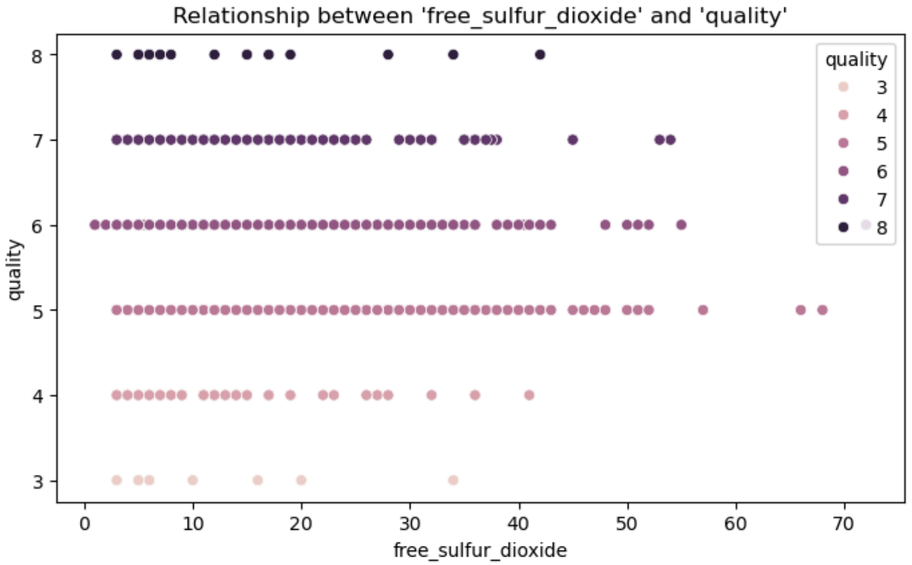


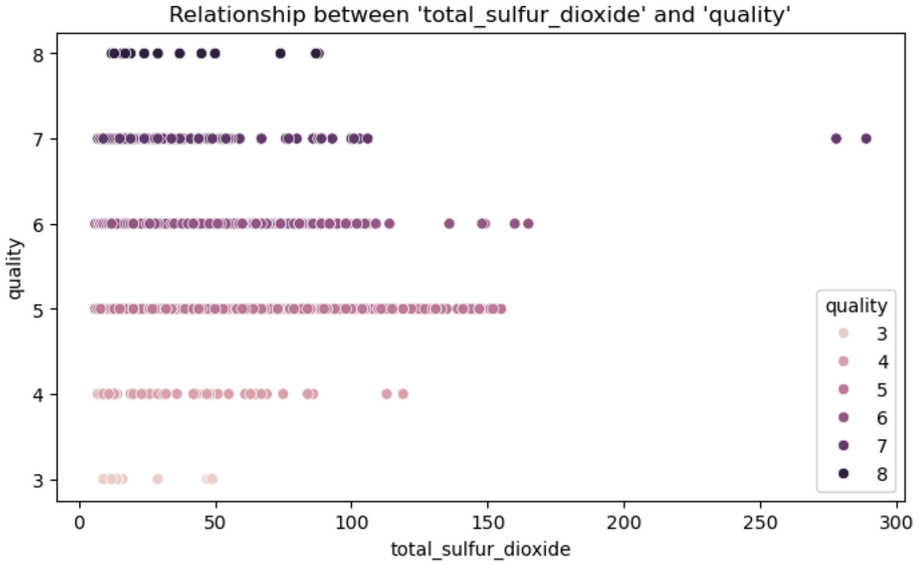


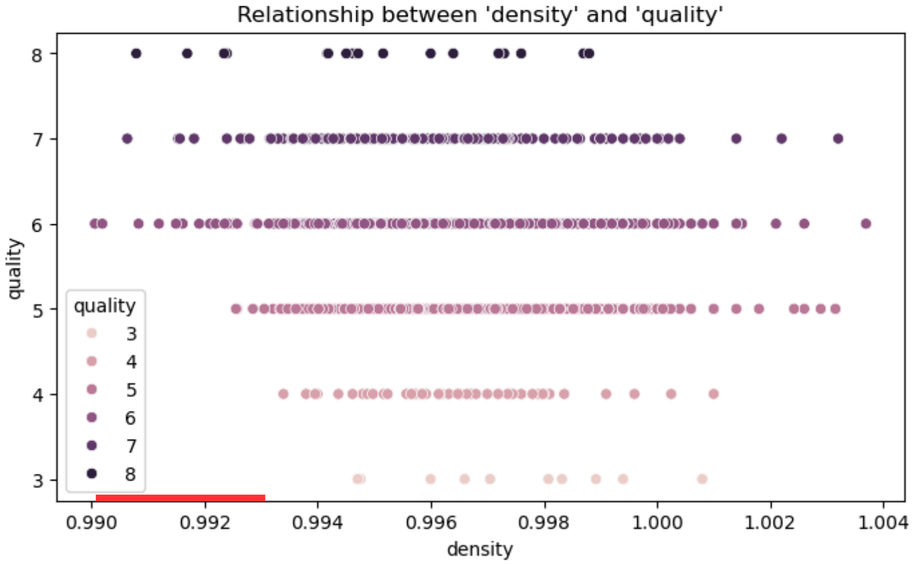


2. Phân tích khám phá

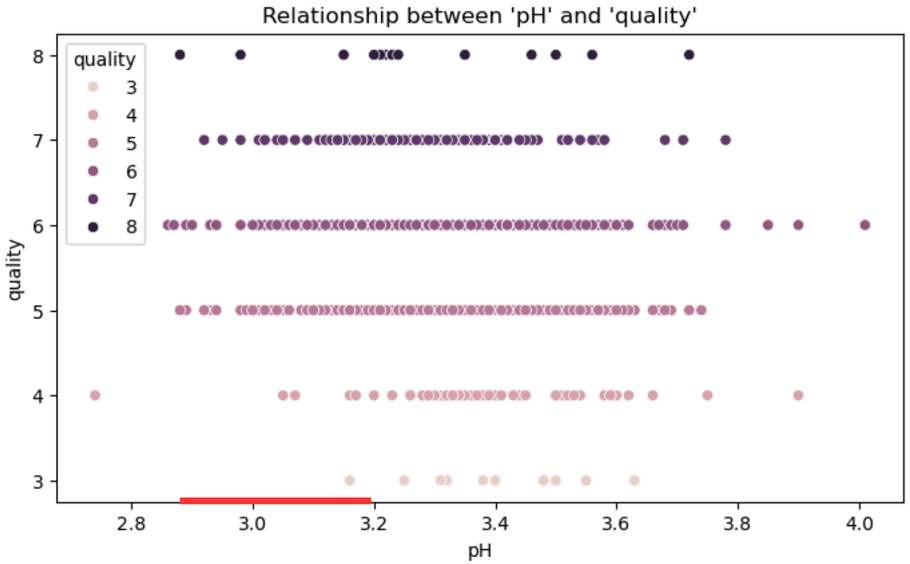


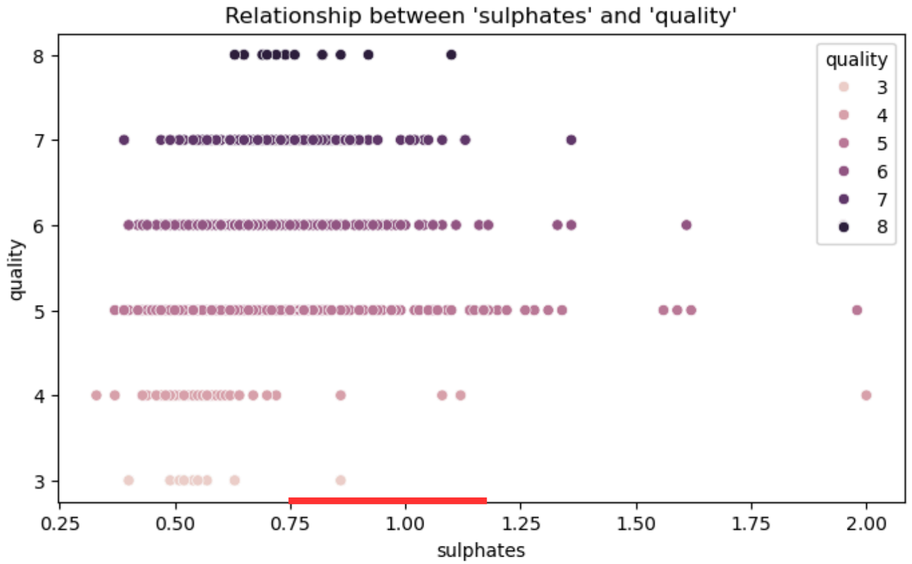


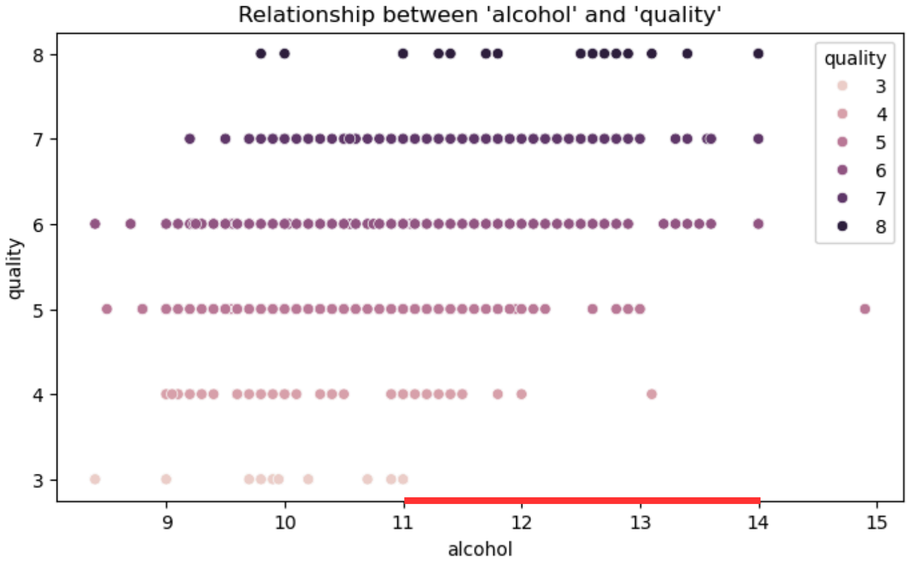




2. Phân tích khám phá







3. Kết luận

Dựa trên các phát hiện sau khi đã khám phá, thì chúng ta cần tập trung quan sát các yếu tố ảnh hưởng mạnh đến chất lượng của rượu (Volatile_acidity,

Ảnh hưởng mạnh	Ảnh hưởng nhẹ	Không ảnh hưởng
Volatile axidity [<0.7]	Fixed axidity [12-14]	Chlories
Citric_acid [>0.5]	Total_SO2 [70-120]	Resudal Sugar
Sulphates [0.75-1.20]	Density [< 0.993]	Free SO2
Alcohol [11-14]	pH [2.8-3.0]	

Ngoài ra cũng cần xem xét các yếu tố ít ảnh hưởng, vì chúng cũng có thể là mối nguy cơ ảnh hưởng đến chất lượng của rượu (giả sử như đảm bảo lượng SO2 phù hợp để đảm bảo sức khỏe người dùng, ...)

THANK YOU

FOR LISTENING