



VNU-HCMUS

# PROJECT 1 - PHÂN LOẠI CHẤT LƯỢNG RƯỢU VANG

---

**ĐƯỢC TRÌNH BÀY BỞI**  
Nhóm L

**ĐƯỢC TRÌNH BÀY CHO**  
Giảng viên: TS.Tô Đức Khanh

# MỤC LỤC



I      TỔNG QUAN ĐỒ ÁN

---

II      EDA

---

III      TIỀN XỬ LÝ DỮ LIỆU

---

IV      KIỂM ĐỊNH

---

V      XÂY DỰNG MÔ HÌNH

---

VI      KẾT QUẢ ĐẠT ĐƯỢC

---



# I. TỔNG QUAN ĐỒ ÁN

---

1  
Giới thiệu đồ án

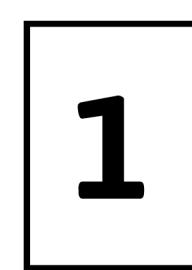
2  
Bài toán đặt ra

3  
Phương án xử lý

# 1. GIỚI THIỆU ĐỒ ÁN

---

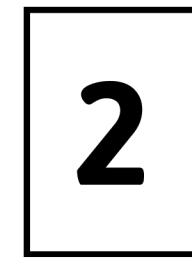
Đặc tính hóa học đóng vai trò quan trọng trong việc đánh giá chất lượng rượu vang



## Quyết định hương vị

1

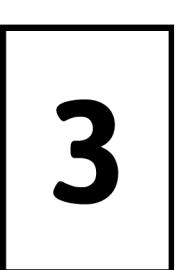
Các hợp chất hóa học trong rượu vang kết hợp tạo nên hương vị, màu sắc và cấu trúc độc đáo.



## Ảnh hưởng quá trình sản xuất

2

Quá trình lên men và lão hóa rượu vang là các phản ứng hóa học phức tạp, quyết định sự phát triển hương vị theo thời gian.



## Cung cấp cơ sở khoa học cho đánh giá

Phân tích hóa học giúp các chuyên gia đo lường, so sánh, và phân loại rượu một cách khách quan, tạo nên tiêu chuẩn đánh giá chất lượng.

**PHÂN LOẠI CHẤT LƯỢNG RƯỢU VANG  
DỰA TRÊN CÁC ĐẶC TÍNH HÓA HỌC**

## 2. BÀI TOÁN ĐẶT RA

---



### Mục tiêu 1

Xác định các feature quan trọng



### Mục tiêu 3

Xây dựng mô hình phân loại dự đoán chất lượng



### Mục tiêu 2

Xây dựng mô hình hồi quy dự đoán chất lượng



### Mục tiêu 4

Đề ra chiến lược cải tiến sản phẩm và kinh doanh dựa trên insight từ phân tích.

### **3. PHƯƠNG ÁN XỬ LÝ**

---

EDA

REGRESSION MODELS

PREPROCESSING

CLASSIFICATION MODELS

A/B TESTING

BUSINESS STRATEGY



## II. EDA

---

1

Thống kê mô tả

2

Trực quan hóa dữ liệu

# 1. THỐNG KÊ MÔ TẢ

BIEN	n	mean	median	sd	min	max	iqr
alcohol	6497	10.492	10.300	1.193	8.000	14.90	1.8000
chlorides	6497	0.056	0.047	0.035	0.009	0.61	0.0270
citric_acid	6497	0.319	0.310	0.145	0.000	1.66	0.1400
density	6497	0.995	0.995	0.003	0.987	1.04	0.0047
fixed_acidity	6497	7.215	7.000	1.296	3.800	15.90	1.3000
free_sulfur_dioxide	6497	30.525	29.000	17.749	1.000	289.00	24.0000
p_h	6497	3.219	3.210	0.161	2.720	4.01	0.2100
quality	6497	5.818	6.000	0.873	3.000	9.00	1.0000
residual_sugar	6497	5.443	3.000	4.758	0.600	65.80	6.3000
sulphates	6497	0.531	0.510	0.149	0.220	2.00	0.1700
total_sulfur_dioxide	6497	115.745	118.000	56.522	6.000	440.00	79.0000
volatile_acidity	6497	0.340	0.290	0.165	0.080	1.58	0.1700

## Dữ liệu gồm

- 1599 mẫu rượu vang đỏ & 4898 mẫu rượu vang trắng
- Từ hai file winequality-red.csv và winequality-white.csv
- Với 11 biến

## **2. TRỰC QUAN HÓA DỮ LIỆU**

---



- a. KIỂM TRA SỰ CÂN BẰNG CỦA DỮ LIỆU**

---
- b. BOOSTRAP**

---
- c. KHẢO SÁT PHÂN PHỐI CỦA CÁC BIẾN**

---
- d. TƯƠNG QUAN CỦA CÁC BIẾN VỚI QUALITY**

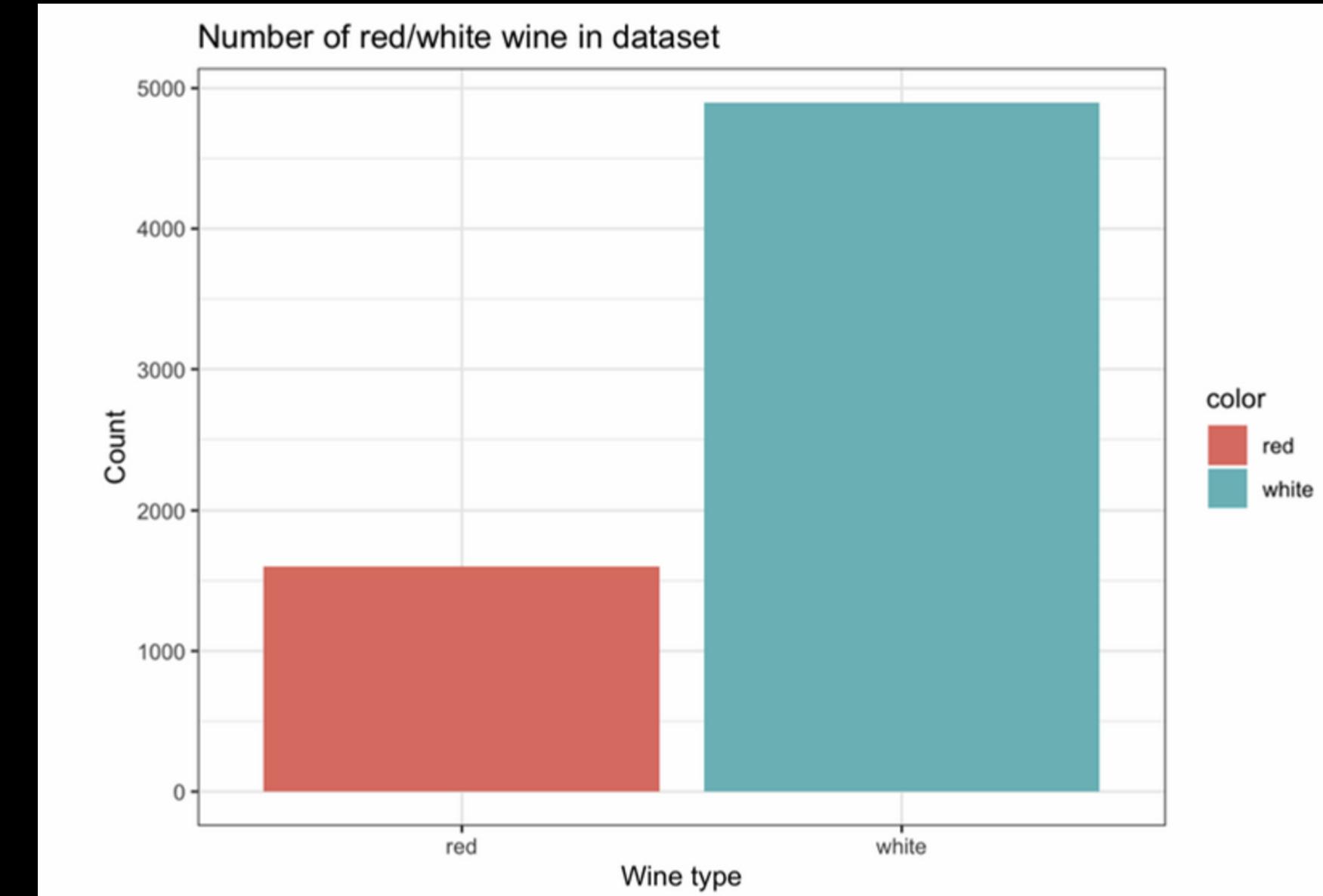
---
- e. RED WINE VÀ WHITE WINE**

---

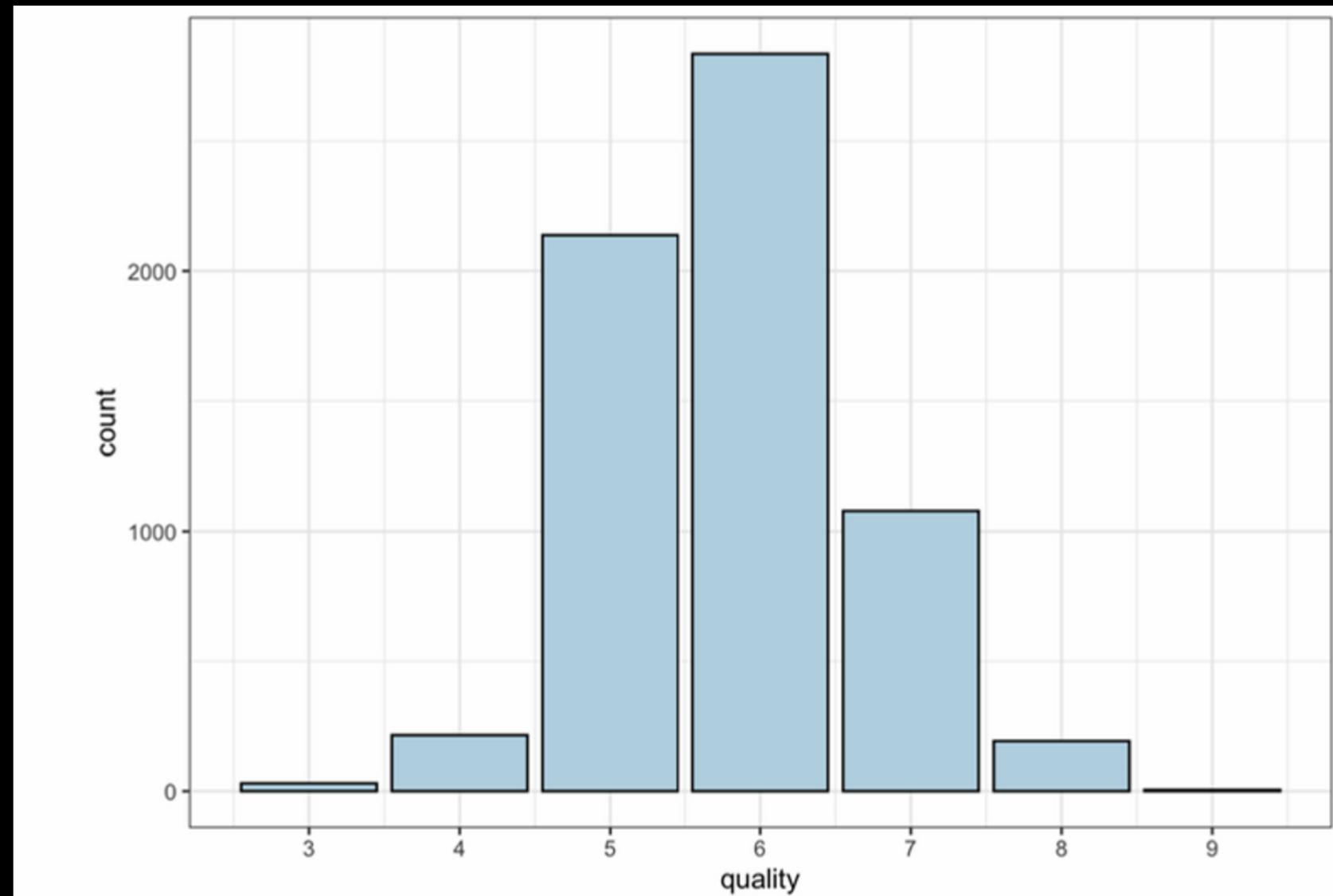
# KIỂM TRA SỰ CÂN BẰNG DỮ LIỆU

Số lượng mẫu lớp white nhiều gấp ba lần lớp red

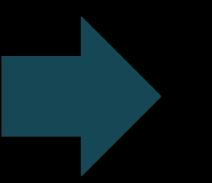
Bộ dữ liệu này không cân bằng  
(imbalanced data)



# KIỂM TRA SỰ PHÂN BỐ CỦA CHẤT LƯỢNG RƯỢU

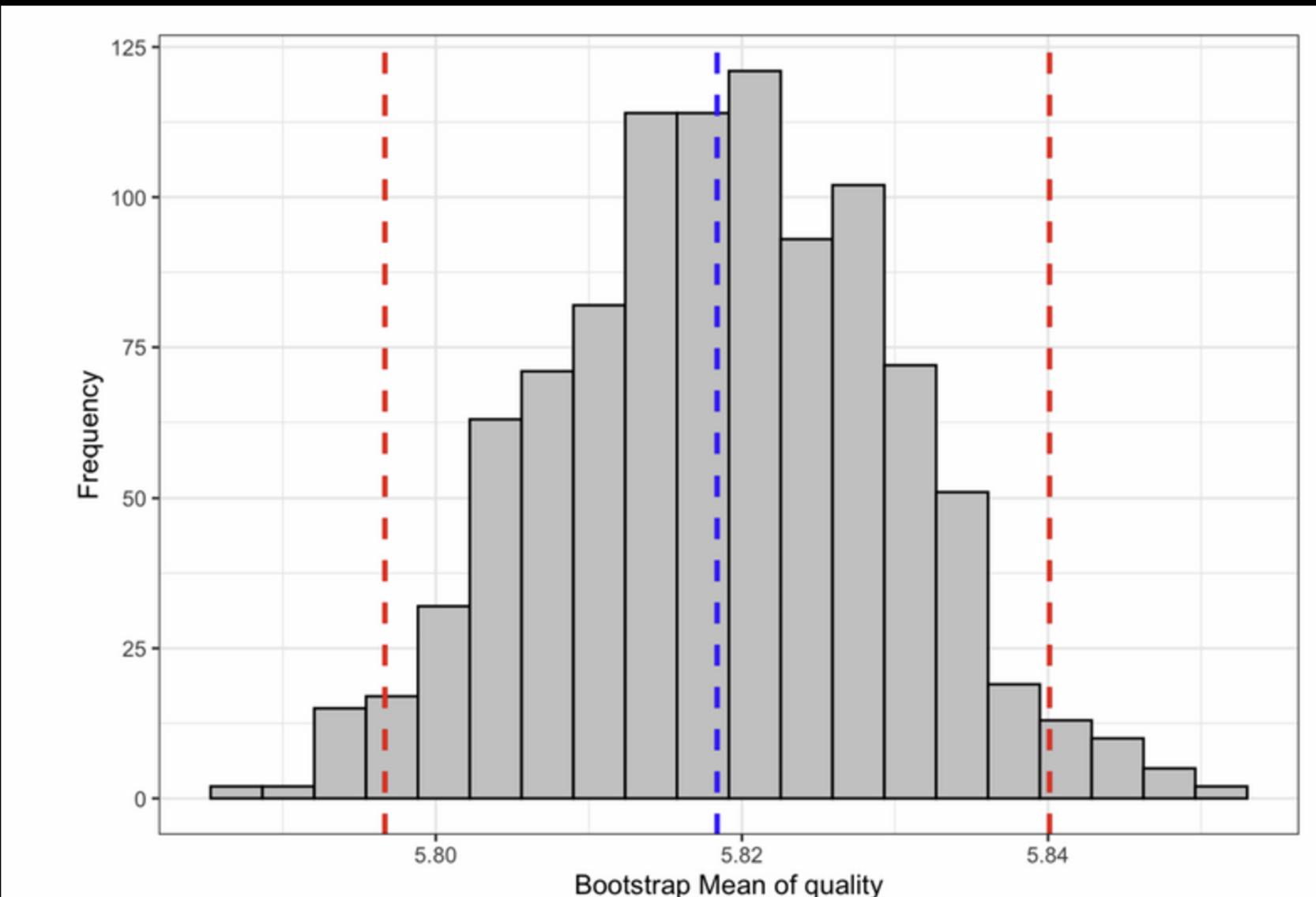


Dữ liệu quality chủ yếu nằm trong khoảng [5,7],  
rất ít dưới 3 và một số ít trên 8.



Dùng bootstrap để ước lượng phân phối mẫu và  
khoảng tin cậy

# BOOTSTRAP

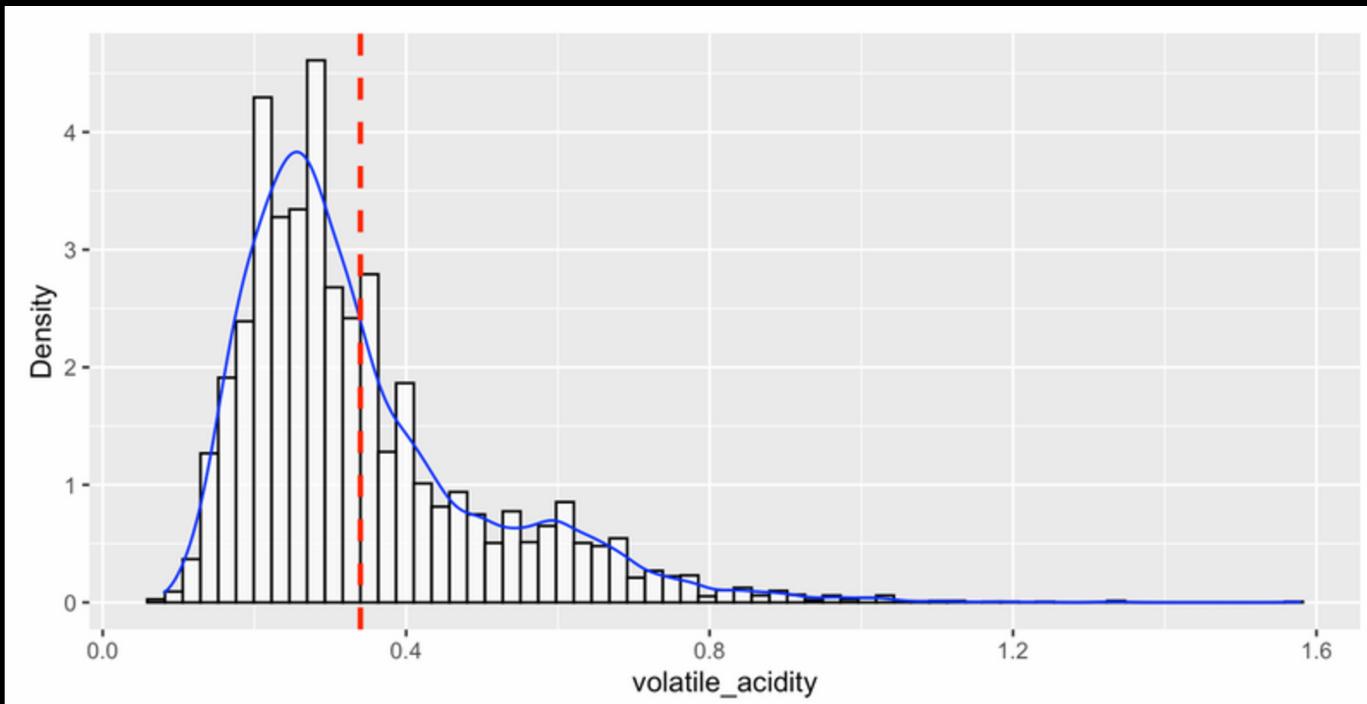
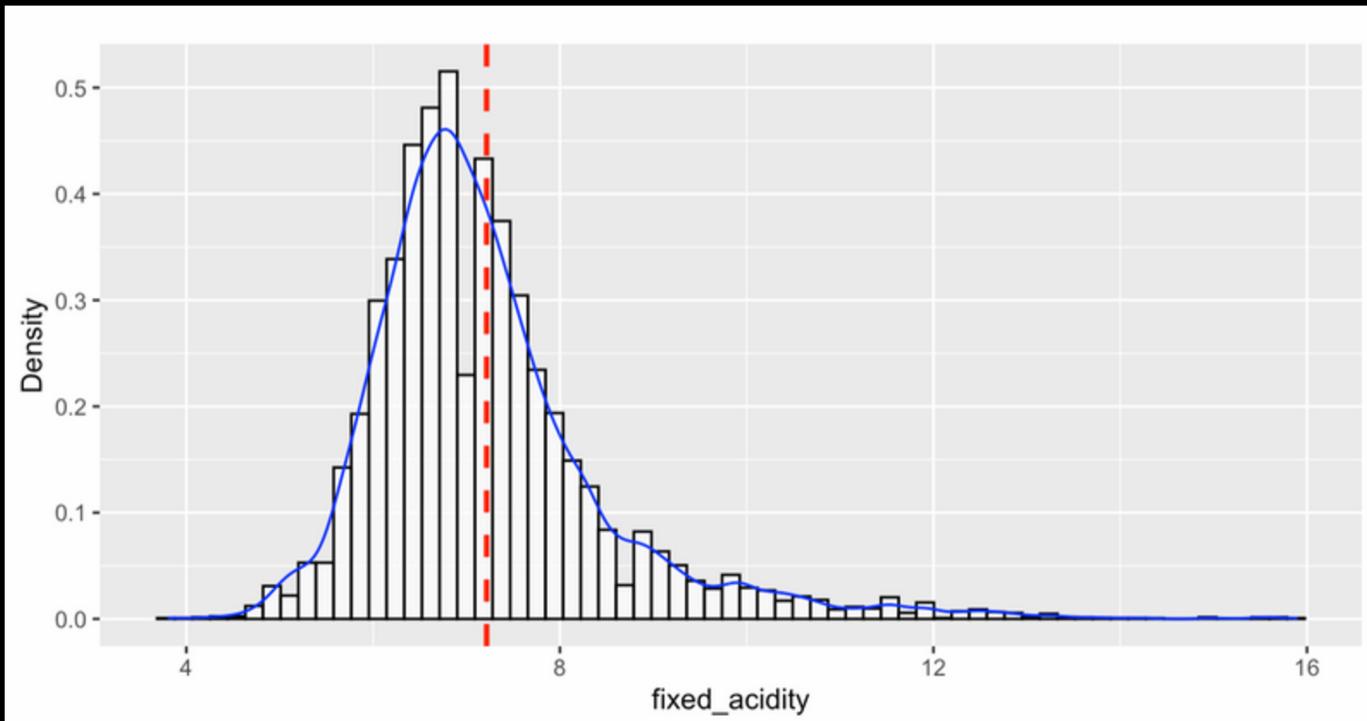


Phân phối trung bình mẫu của quality  
gần phân phối chuẩn

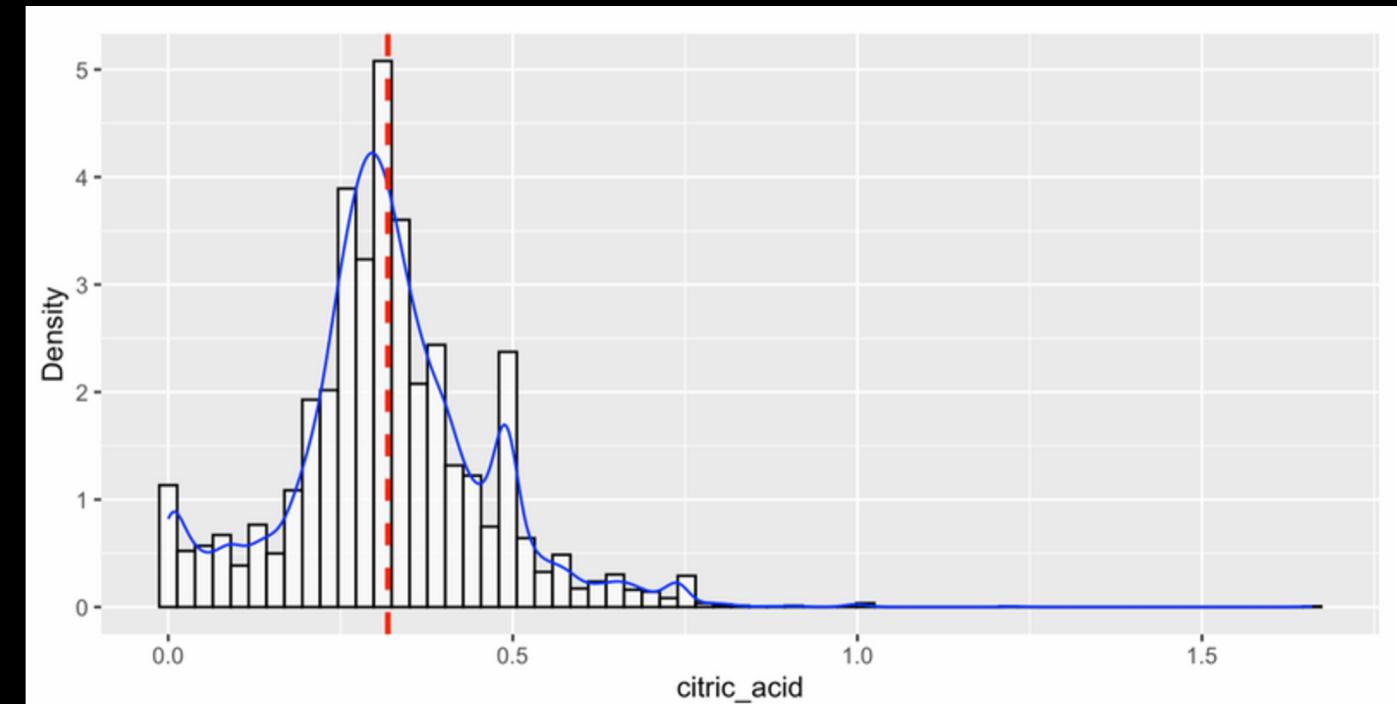
Trung bình: 5.82  
KTC 95%: [5.979; 5.841]

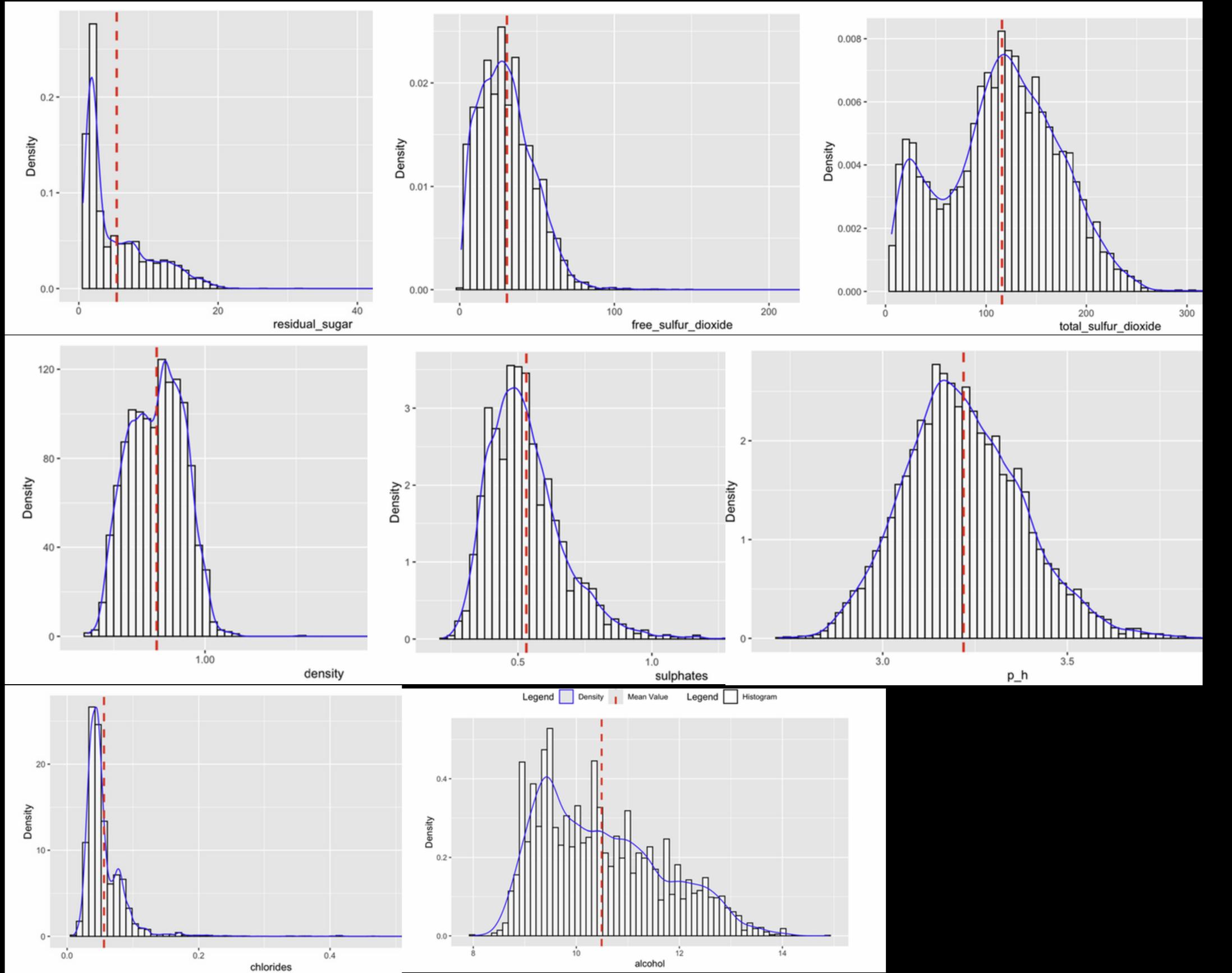
Tương tự, kiểm tra phân phối và khoảng tin cậy cho các biến khác

# HISTOGRAM & DENSITY



- fixed và volatile acidity phân phối lệch phải với đuôi dài.
- citric acid có phân phối hai đỉnh và một số outliers
- Nồng độ cao bất thường từ 0.0 đến 0.5 phổ biến hơn.

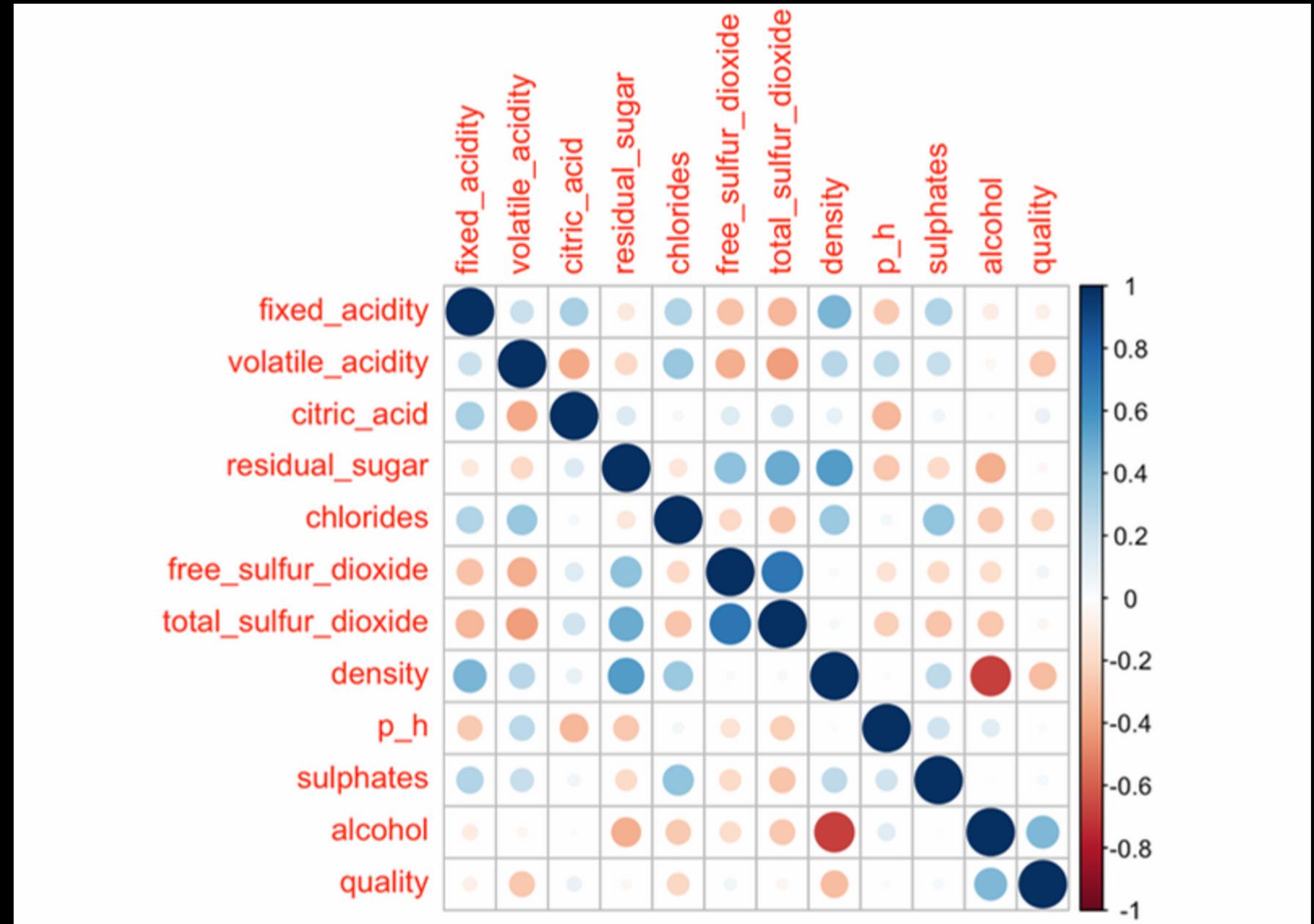




## Rút ra một số nhận xét

- Residual sugar bị lệch phải nặng
- Chlorides có 2 đỉnh có thể ứng với xu hướng của hai loại rượu
- Free và total SO<sub>2</sub> đều bị lệch phải. Trong đó: - free SO<sub>2</sub> có nhiều dạng gai
- pH có phân phối gần giống phân phối chuẩn, cho thấy hai loại rượu có nồng độ pH gần giống nhau.
- Alcohol bị lệch phải nhẹ, với một số nồng độ cao bất thường so với phần còn lại

# MA TRẬN TƯƠNG QUAN



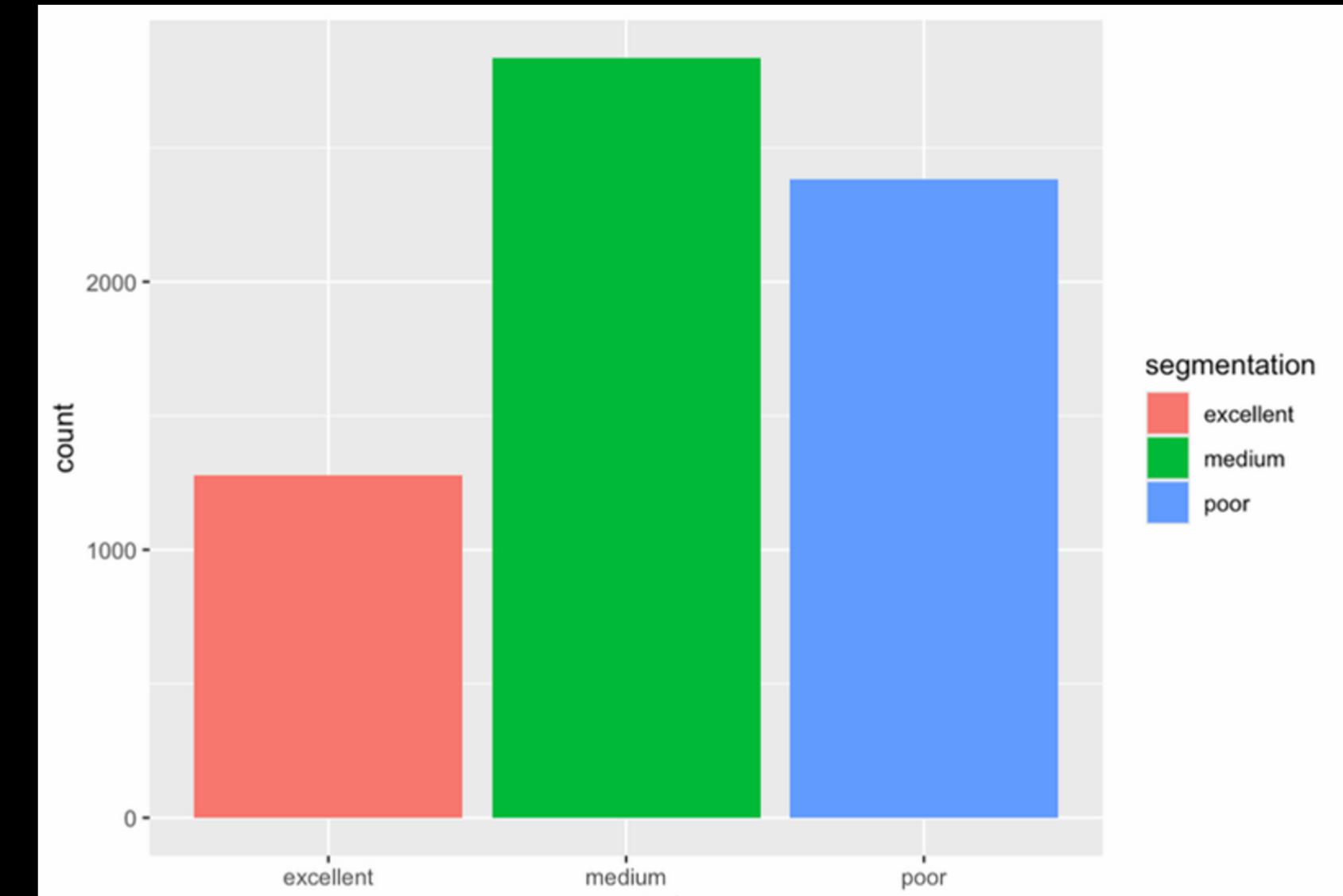
## Rút ra một số nhận xét

- Quality tương quan dương cao với alcohol
- Quality tương quan âm cao với density, volatile acidity và chlorides
- Tuy nhiên density và alcohol tương quan âm cao với nhau -> chỉ giữ lại một biến để dự đoán quality

# SO SÁNH VỚI SEGMENTATION

## Feature segmentation

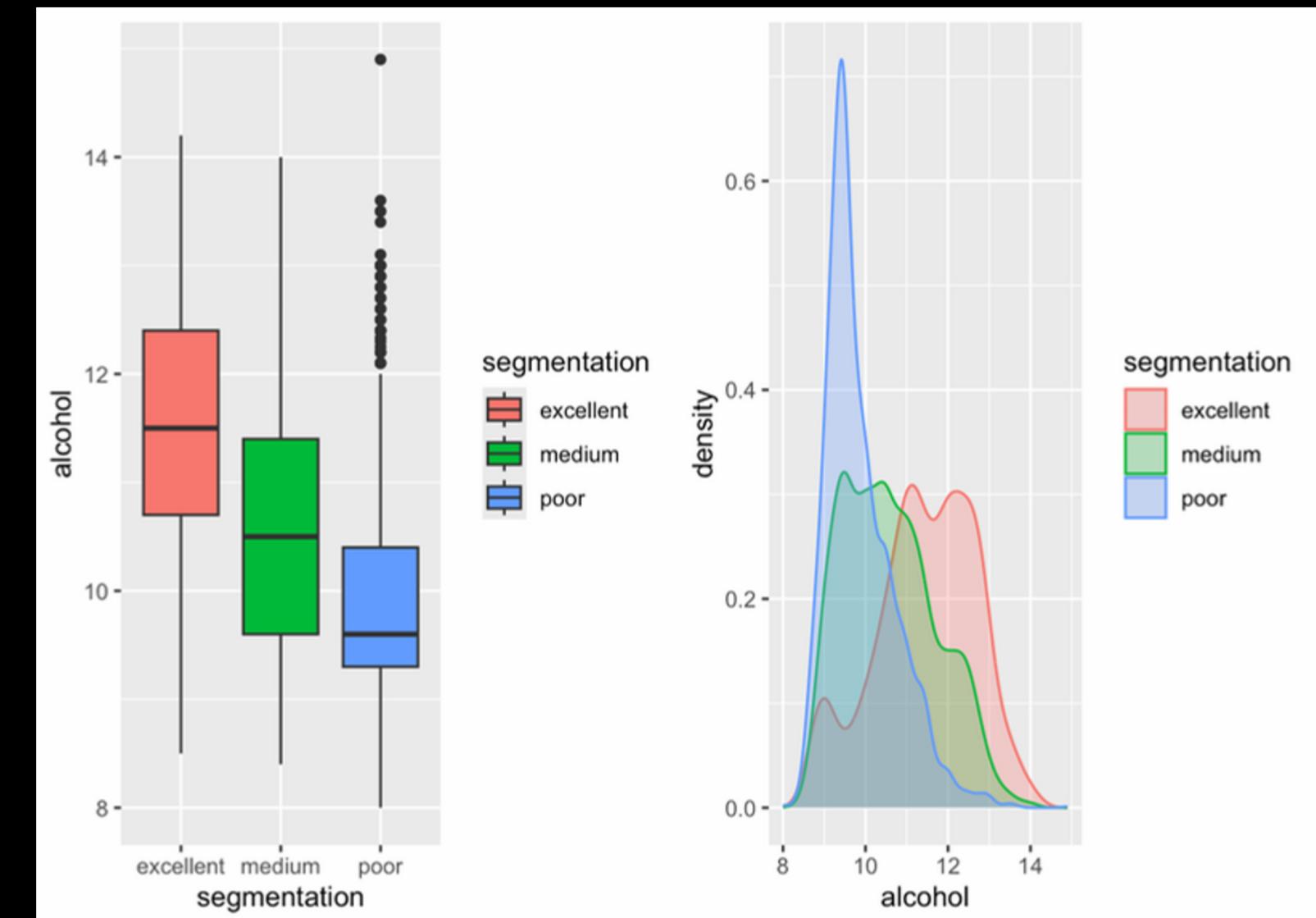
- Tạo feature mới bằng cách rời rạc hóa biến quality**
- **Poor** (phân khúc tầm thấp): có điểm quality từ 3-5
  - **Medium** (phân khúc tầm trung): có điểm quality bằng 6
  - **Excellent** (phân khúc cao cấp): có điểm quality từ 7 trở lên



# SO SÁNH VỚI SEGMENTATION

## Alcohol (nồng độ cồn)

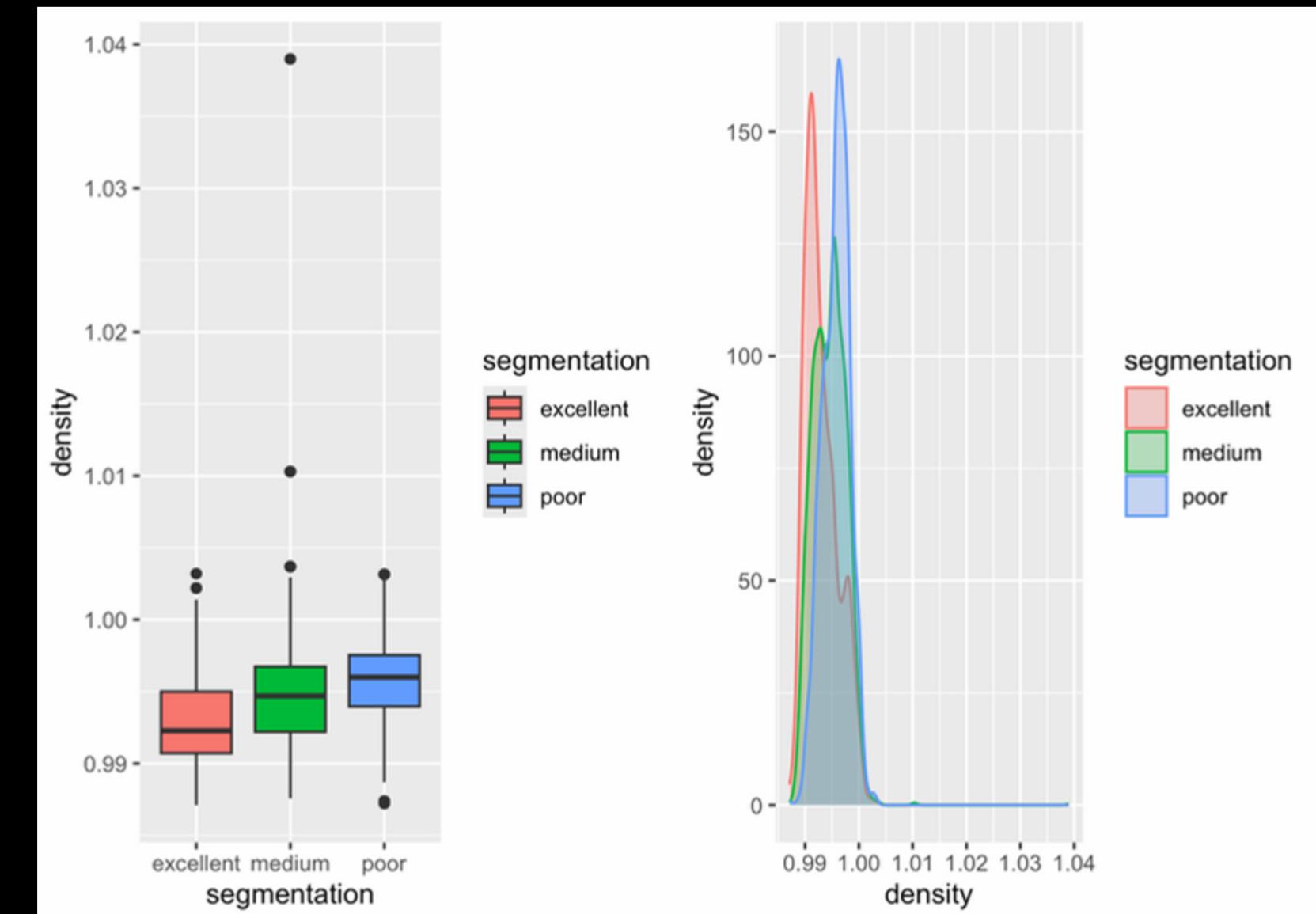
- Rượu vang phân khúc cao hơn xu hướng có alcohol cao hơn
- Alcohol là một yếu tố quan trọng ảnh hưởng đến chất lượng rượu vang



# SO SÁNH VỚI SEGMENTATION

## Density (mật độ trong rượu)

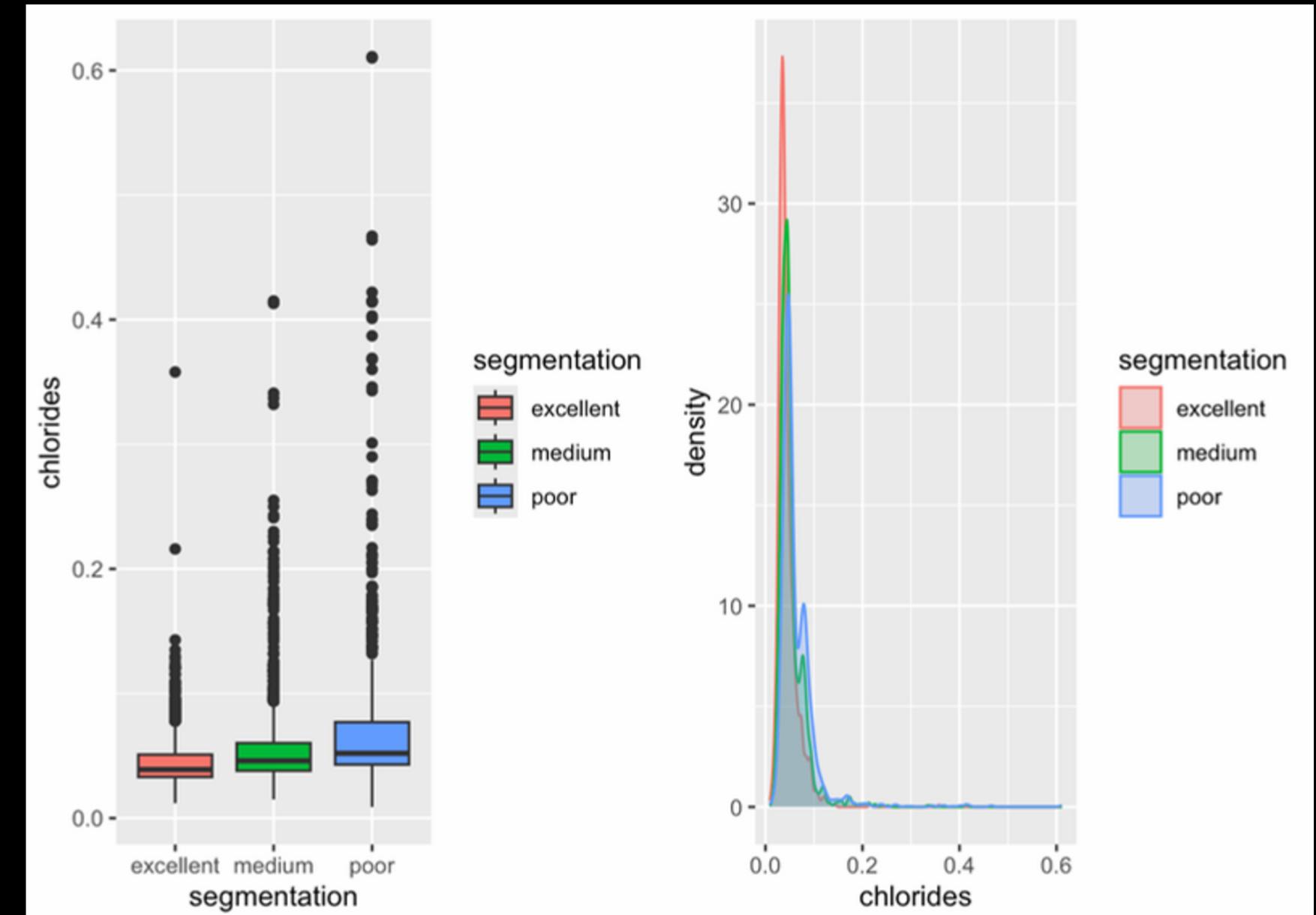
- Rượu vang chất lượng cao có density thấp hơn so với loại rượu có chất lượng thấp hơn
- Density có thể ám chỉ đến mức độ nguyên chất có rượu (tỷ lệ được pha với nước)



# SO SÁNH VỚI SEGMENTATION

## Chlorides (muối hay độ mặn)

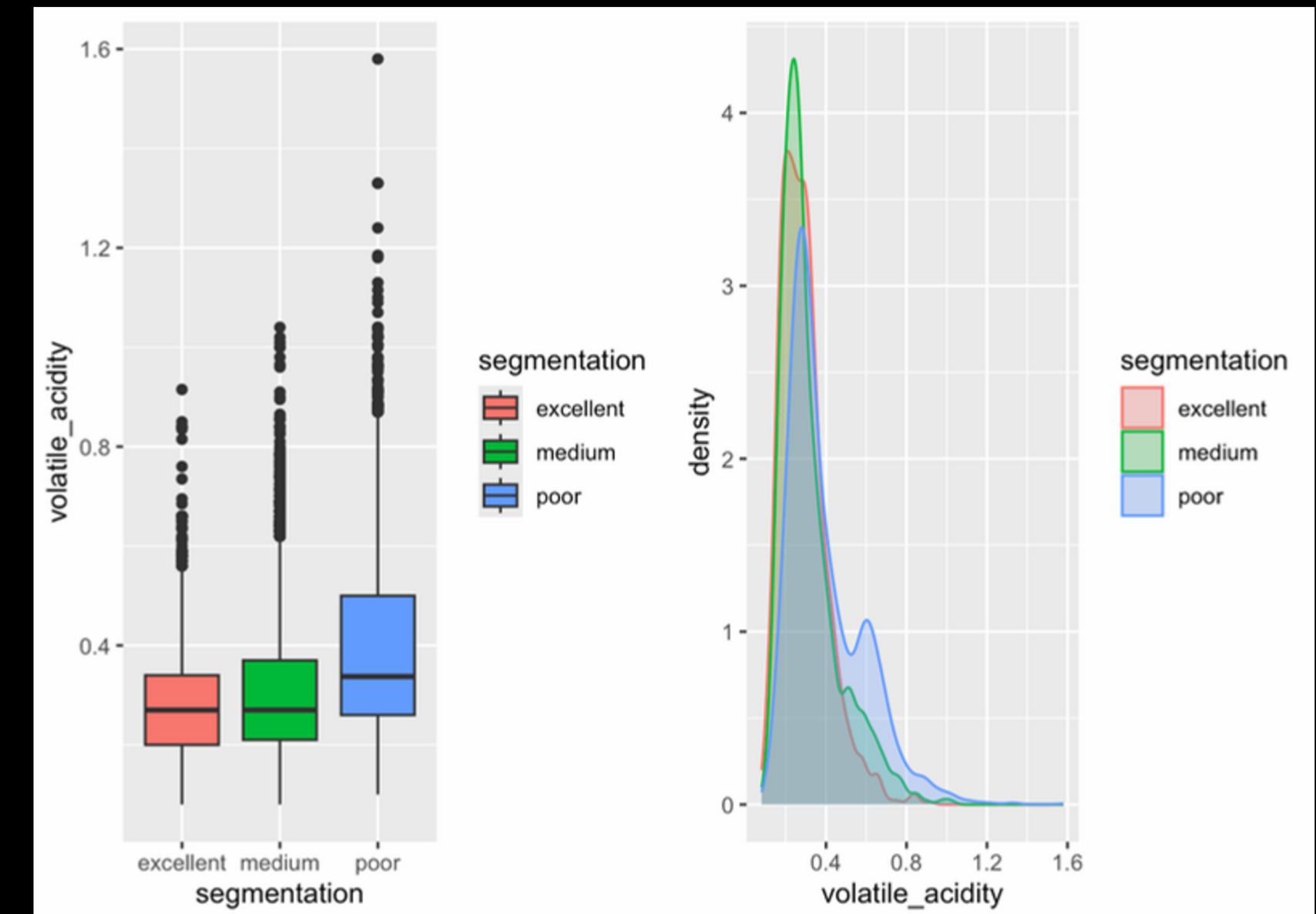
- Rượu chất lượng cao có hàm lượng chlorides thấp và đồng nhất hơn (khoảng IQR hẹp hơn)
- Có nhiều điểm có khả năng là outliers



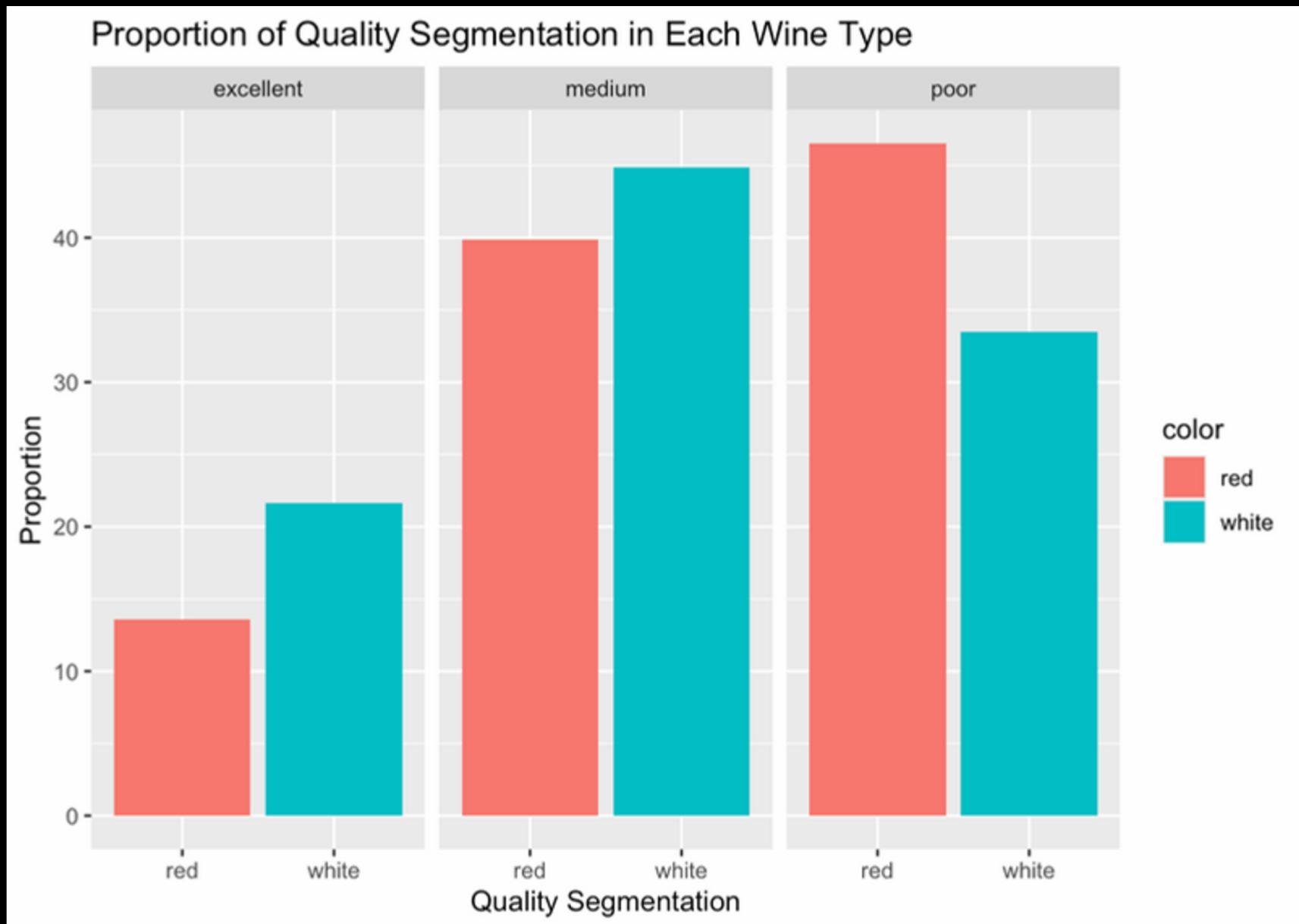
# SO SÁNH VỚI SEGMENTATION

## Volatile Acidity (axit dễ bay hơi)

- Rượu vang chất lượng cao có hàm lượng volatile acidity thấp hơn và khoảng IQR hẹp hơn
- Thể hiện hàm lượng axit này tỉ lệ nghịch với chất lượng rượu vang và rượu vang chất lượng cao có sự đồng nhất cao hơn trong khâu sản xuất khi có khoảng IQR hẹp hơn



# RED WINE & WHITE WINE



## Segmentation trong color

- Phân bố của hai loại rượu khá tương đồng khi rượu tầm trung và thấp chiếm đa số
- Ở rượu vang trắng có cho tỷ lệ rượu cao cấp nhiều hơn và rượu tầm thấp ít hơn
- Vì vậy màu sắc của rượu có thể là một tiêu chí đánh giá quan trọng



### III. TIỀN XỬ LÝ DỮ LIỆU

---

1

Xử lý missing & duplicated values

2

Xử lý imbalanced data

# MISSING & DUPLICATED VALUES

---

1

## Missing values

Không có missing values

2

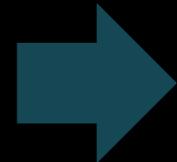
## Duplicated values

- Có 1177 hàng trùng lặp
- Loại bỏ các hàng trùng lặp

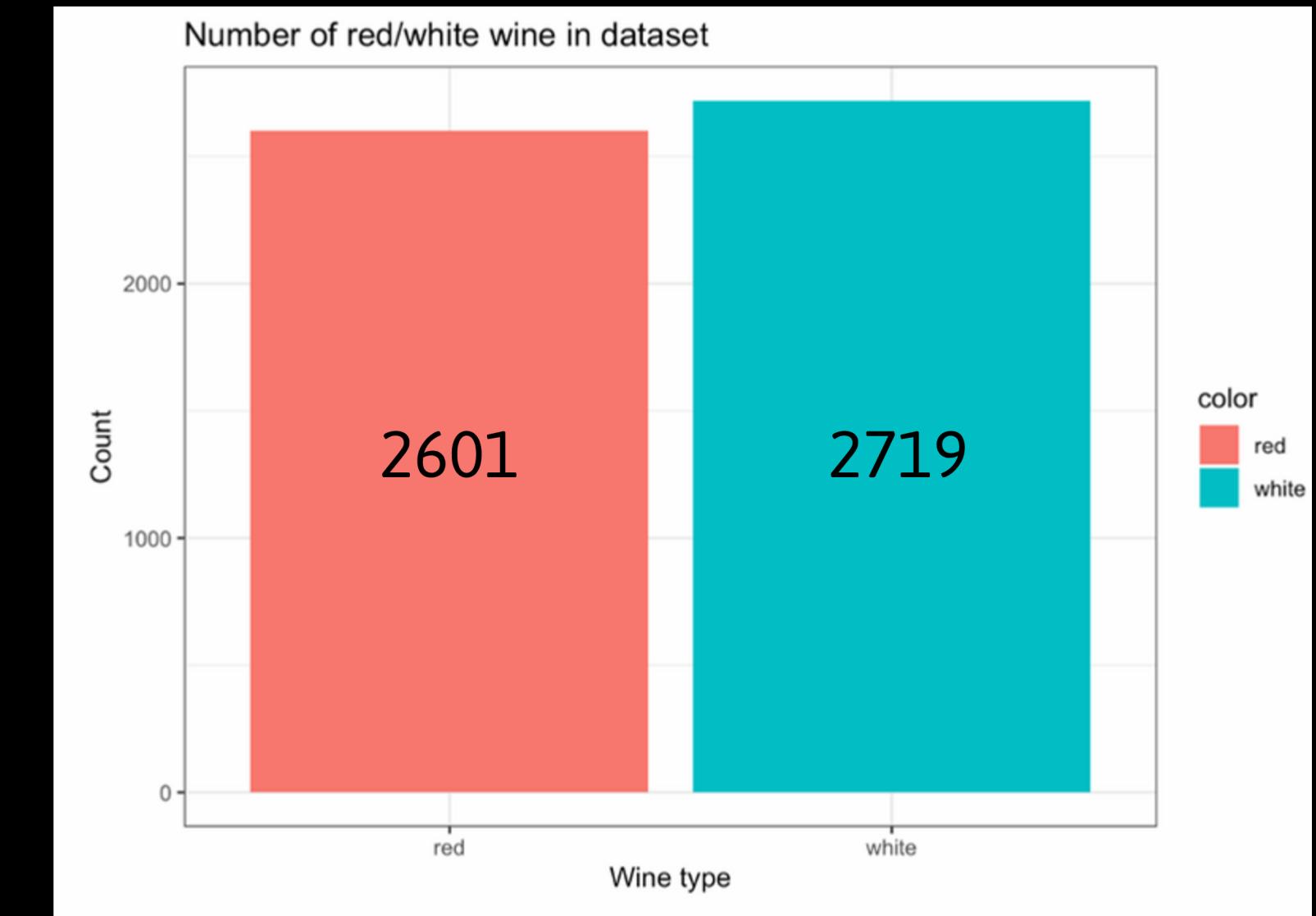
# XỬ LÝ DỮ LIỆU MẤT CÂN BẰNG

**Kết hợp phương pháp over sampling  
và under sampling**

Resampling đến khi dữ liệu cân bằng (2 lớp  
xấp xỉ bằng nhau)



Kết quả cho thấy số lượng mẫu của rượu đỏ và  
rượu trắng gần như cân bằng.



# FAKE || FACT

## IV. KIỂM ĐỊNH

---

1  
Kiểm định trung bình

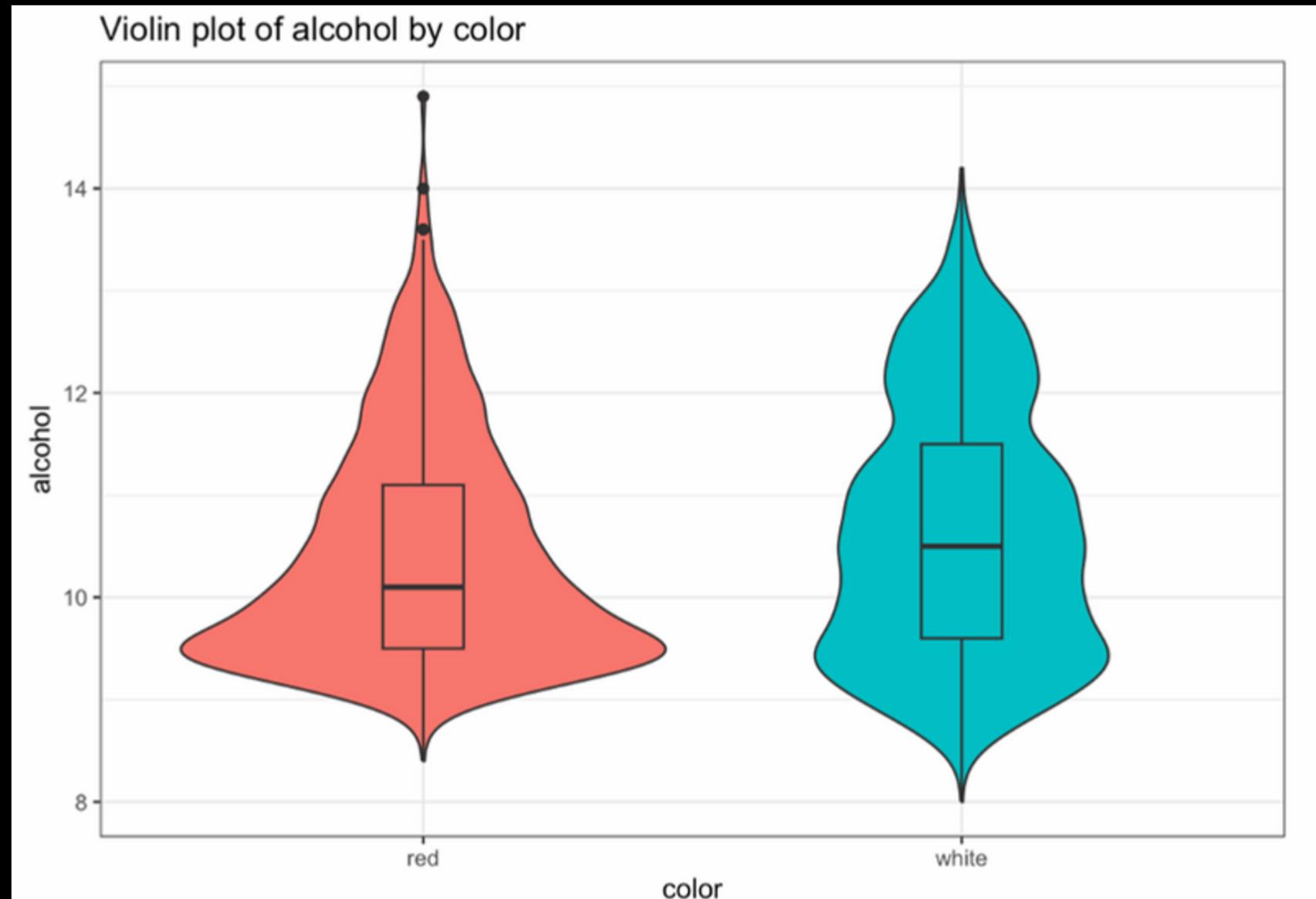
2  
Kiểm định ANOVA

# KIỂM ĐỊNH TRUNG BÌNH

**Kiểm định trung bình cho nồng độ cồn giữa  
rượu đỏ và rượu trắng**

$$H_0 : \mu_{alcohol\_red} = \mu_{alcohol\_white}$$
$$H_1 : \mu_{alcohol\_red} \neq \mu_{alcohol\_white}$$

- mean\_A (trung bình nồng độ cồn của rượu đỏ): 10.62329
- mean\_B (trung bình nồng độ cồn của rượu trắng): 10.4976
- p\_value: 0.922

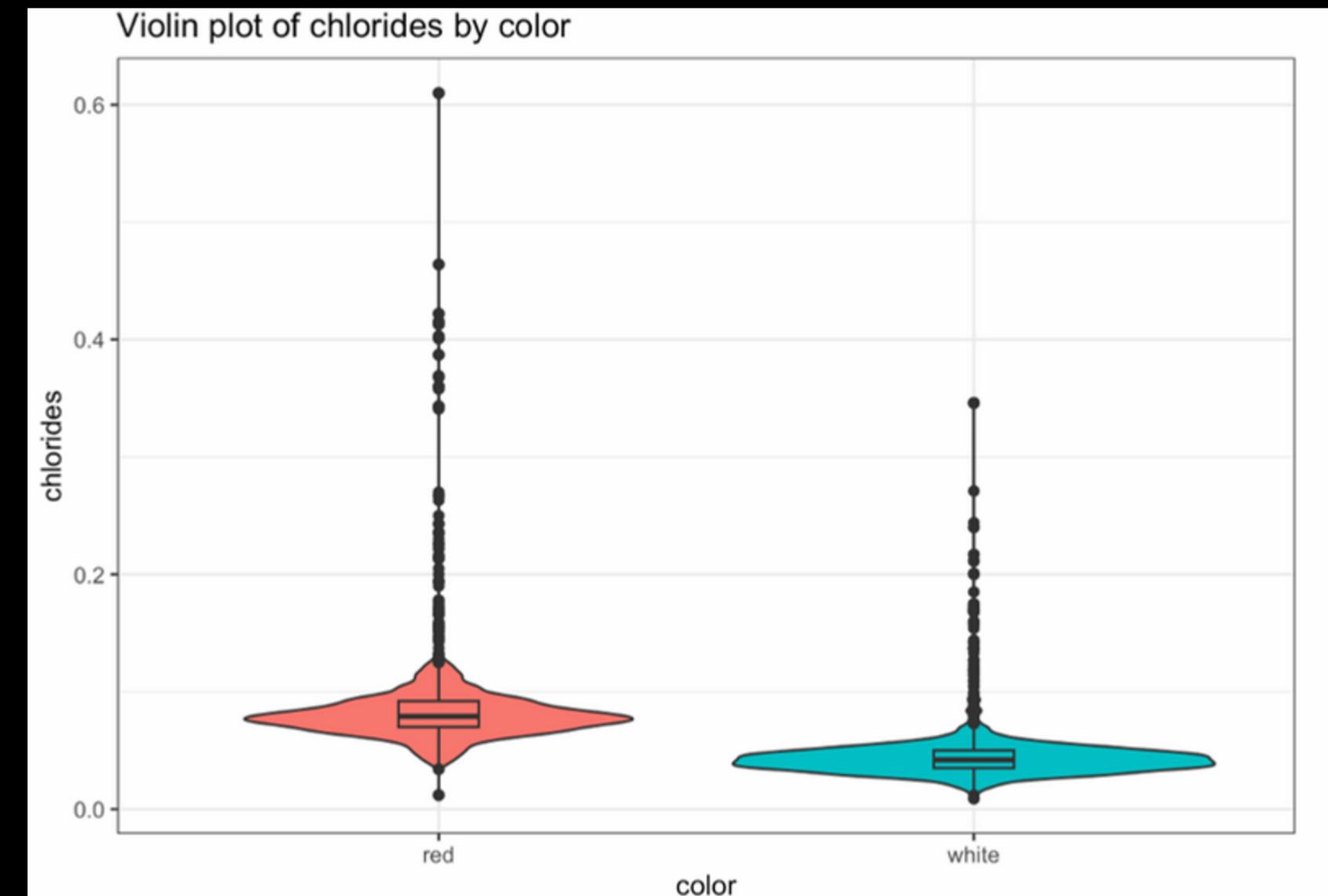


# KIỂM ĐỊNH TRUNG BÌNH

## Kiểm định trung bình cho hàm lượng clorua giữa rượu đỏ và rượu trắng

$$H_0 : \mu_{chlorides\_red} = \mu_{chlorides\_white}$$
$$H_1 : \mu_{chlorides\_red} > \mu_{chlorides\_white}$$

- mean\_A (trung bình hàm lượng clorua của rượu đỏ): 0.04521045
- mean\_B (trung bình hàm lượng clorua của rượu trắng): 0.074923
- p\_value: 0.836



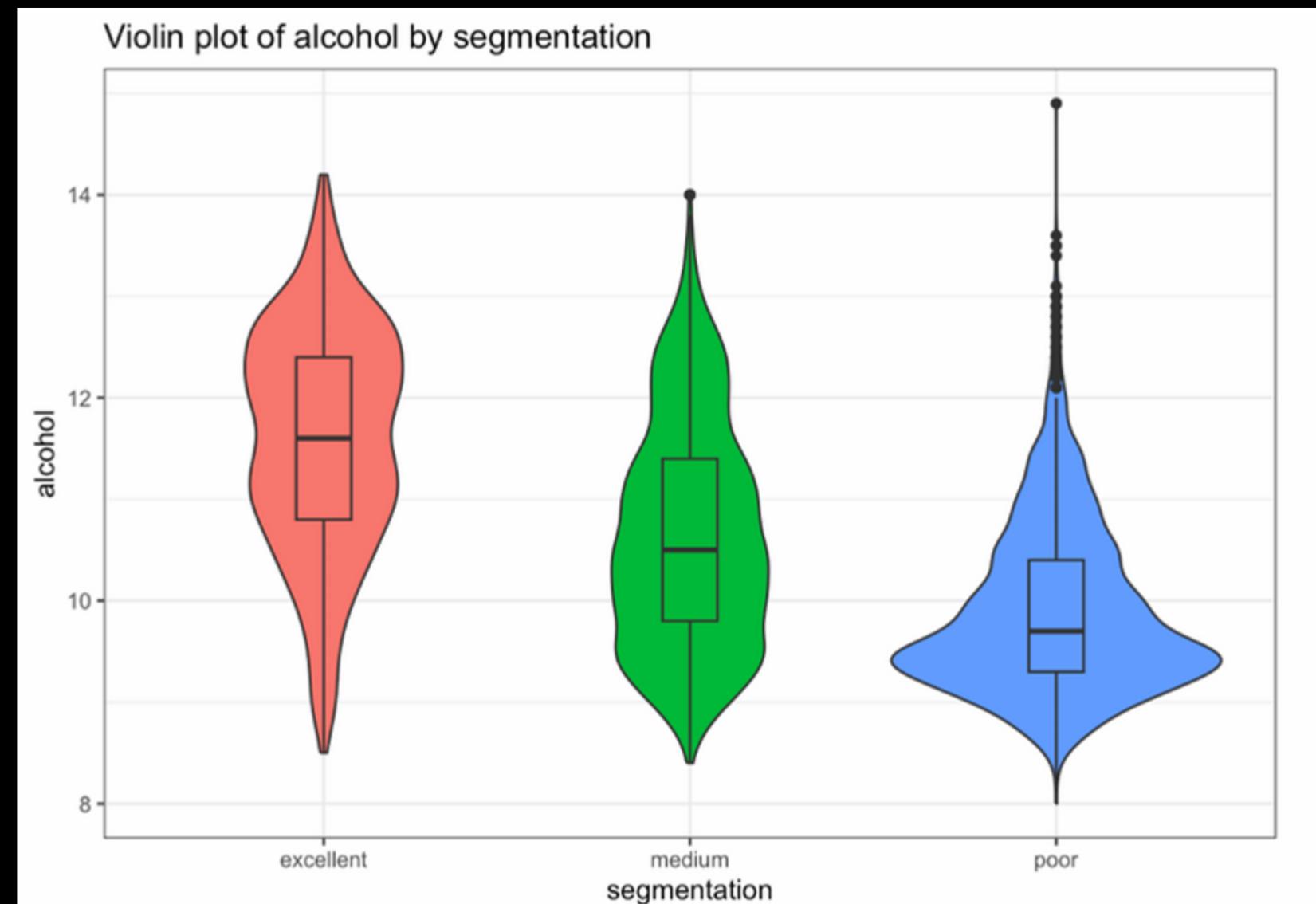
# KIỂM ĐỊNH ANOVA

## Kiểm định ANOVA cho nồng độ cồn giữa các phân khúc chất lượng rượu

$$H_0 : \mu_{alcohol\_poor} = \mu_{alcohol\_medium} = \mu_{alcohol\_excellent}$$
$$H_1 : \mu_{alcohol\_poor} \neq \mu_{alcohol\_medium} \neq \mu_{alcohol\_excellent}$$

### Kết quả

```
## Component 1 :  
##                 Df R Sum Sq R Mean Sq Iter Pr(Prob)  
## segmentation      2 1886.0    943.00 5000 < 2.2e-16 ***  
## Residuals       5317 5594.8      1.05  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



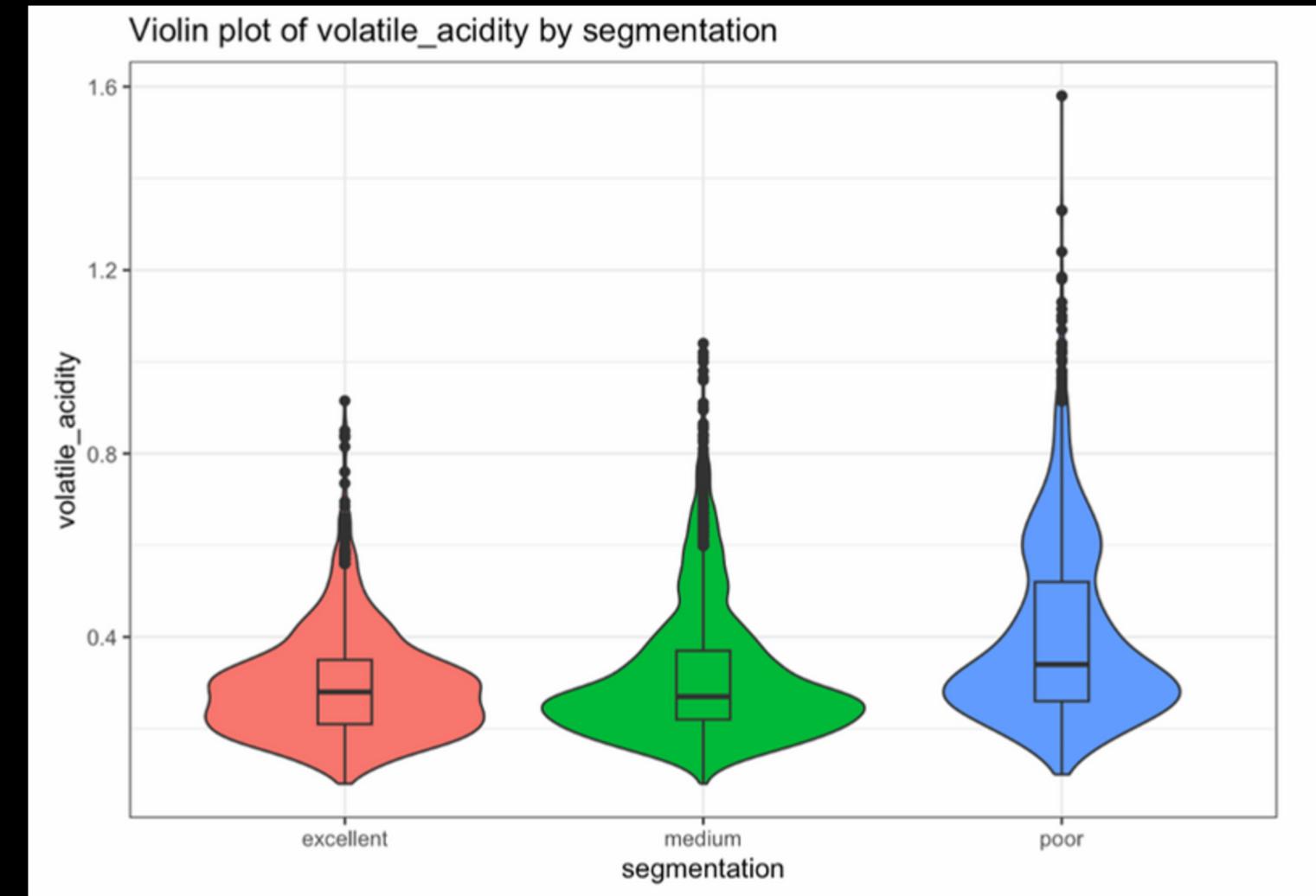
# KIỂM ĐỊNH ANOVA

## Kiểm định ANOVA cho volatile acidity giữa các phân khúc chất lượng rượu

$$H_0 : \mu_{volatile\_poor} = \mu_{volatile\_medium} = \mu_{volatile\_excellent}$$
$$H_1 : \mu_{volatile\_poor} \neq \mu_{volatile\_medium} \neq \mu_{volatile\_excellent}$$

### Kết quả

```
## Component 1 :  
##                 Df R Sum Sq R Mean Sq Iter Pr(Prob)  
## segmentation     2   11.30    5.6500 5000 < 2.2e-16 ***  
## Residuals      5317  139.27    0.0262  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





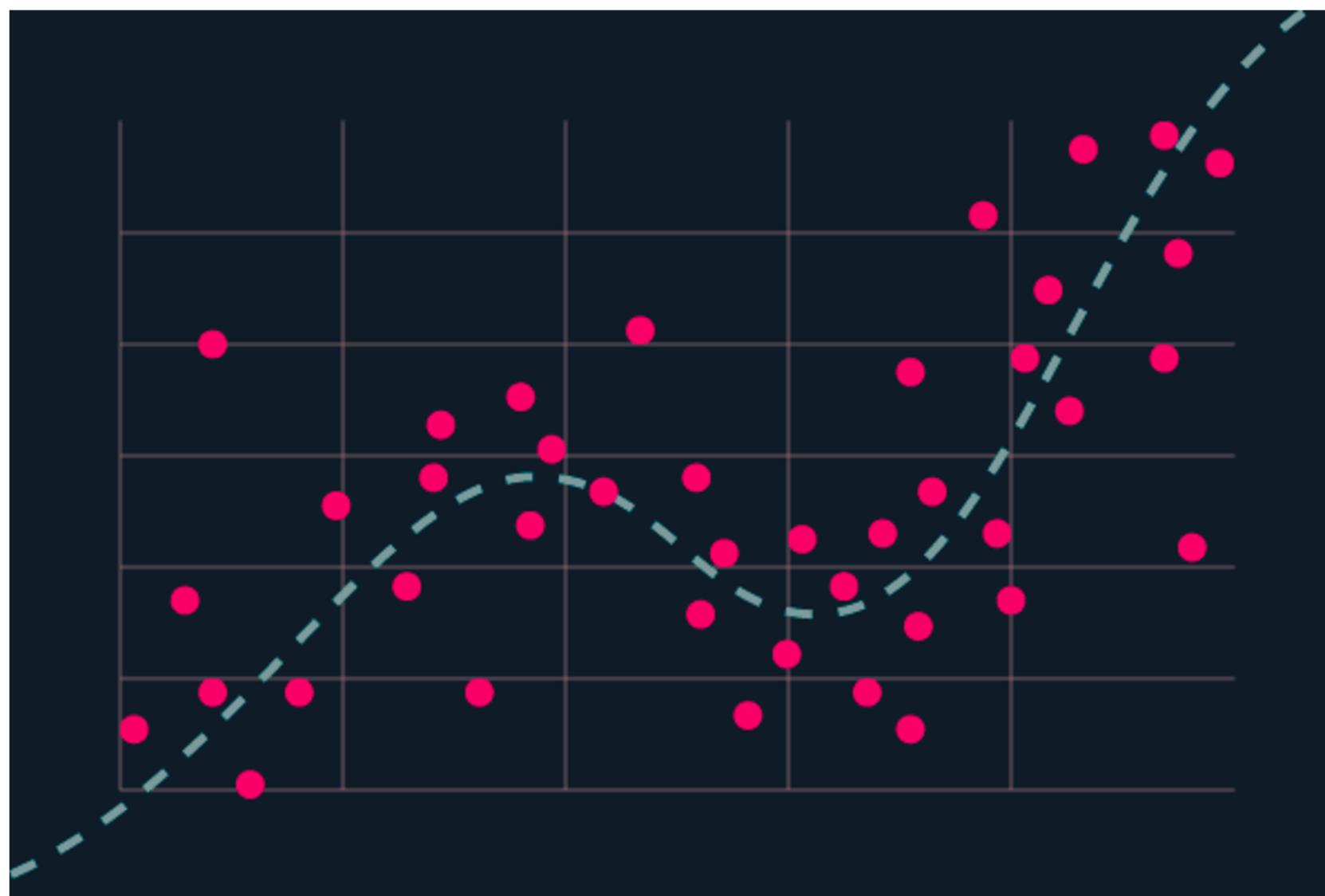
## V. XÂY DỰNG MÔ HÌNH

---

1  
Mô hình hồi quy

2  
Mô hình phân loại

# 1. MÔ HÌNH HỒI QUY



1. XÂY DỰNG MÔ HÌNH

2. LỰA CHỌN MÔ HÌNH

3. CHUẨN ĐOÁN MÔ HÌNH

4. MỞ RỘNG MÔ HÌNH

5. DỰ ĐOÁN KẾT QUẢ

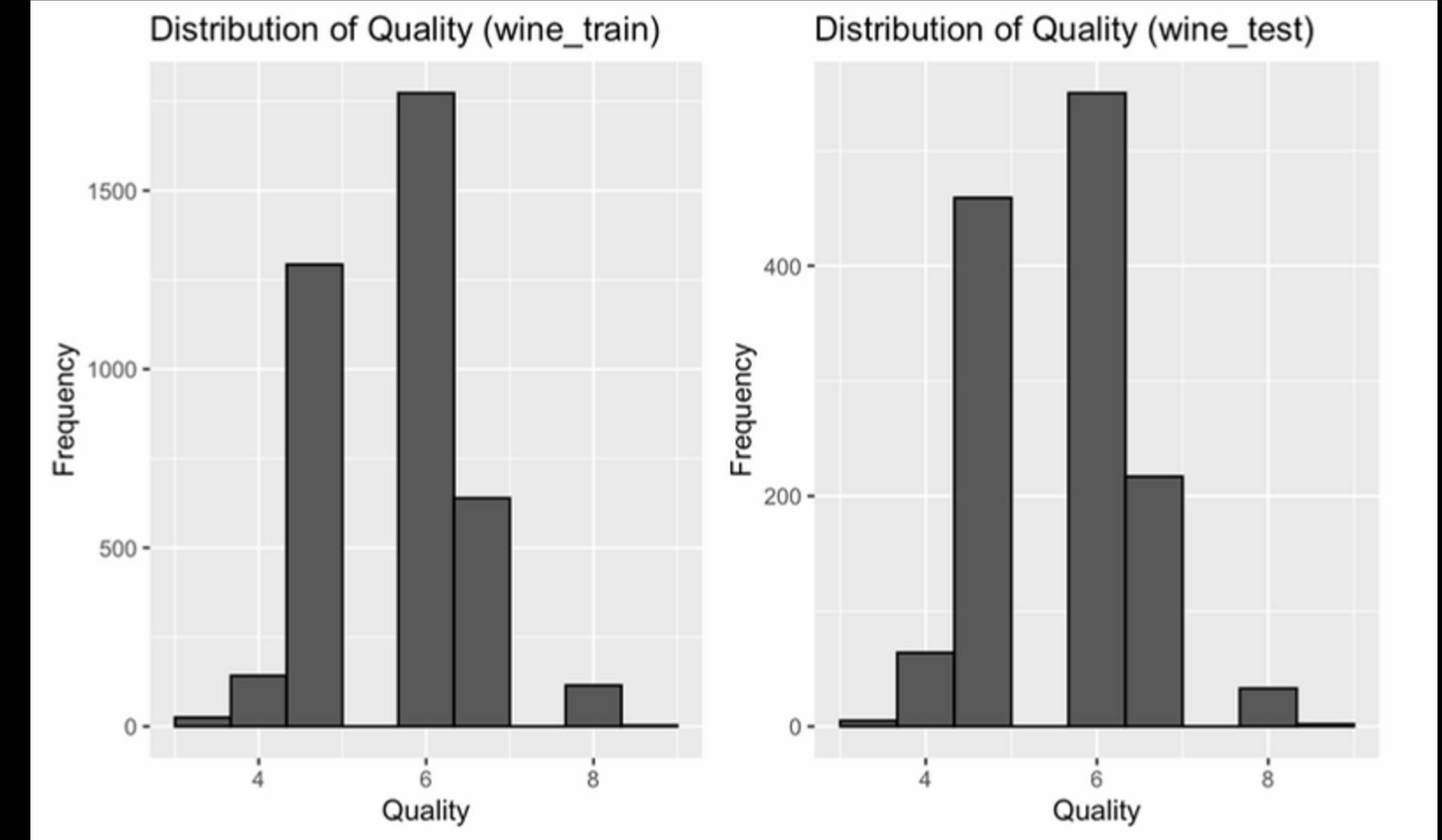
6. ĐÁNH GIÁ MÔ HÌNH

# CHUẨN BỊ DỮ LIỆU

## Train - Test set split

Test size = 0.25

## Data normalization



Trainingset, Testset có cùng phân phối điểm chất lượng rượu

# XÂY DỰNG MÔ HÌNH

---

Xây dựng mô hình hồi quy chất lượng rượu theo tất cả các biến

quality ~ fixed\_acidity + volatile\_acidity + citric\_acid + residual\_sugar + chlorides +  
free\_sulfur\_dioxide + total\_sulfur\_dioxide + density + p\_h + sulphates + alcohol + color

Tồn tại hiện tượng đa cộng tuyến

VIF

density: 25.3  
fixed\_acidity: 5.7

Bootstrap kiểm định hệ số hồi quy  
( pValue > 0.05)

pValue

$\beta$  citric\_acid: 0.376  
 $\beta$  fixed\_acidity: 0.814

# LỰA CHỌN MÔ HÌNH

1

**Loại bỏ lần lượt các biến đa cộng tuyến, các biến có pValue  
cho hệ số hồi quy lớn hơn 0.05**

quality ~ volatile\_acidity + residual\_sugar + **chlorides** + **free\_sulfur\_dioxide** +  
**total\_sulfur\_dioxide** + sulphates + alcohol + color

Linear Model 5



R2\_adj = 0.3003

2

**Both-direction Stepwise**

forward + backward

quality ~ alcohol + volatile\_acidity + sulphates + residual\_sugar + color

R2\_adj = 0.2915

3

**Ridge Regression**

quality ~ .

alpha = 0, cross-validation --> best\_lambda

R2\_adj = 0.2119

# CHUẨN ĐOÁN MÔ HÌNH

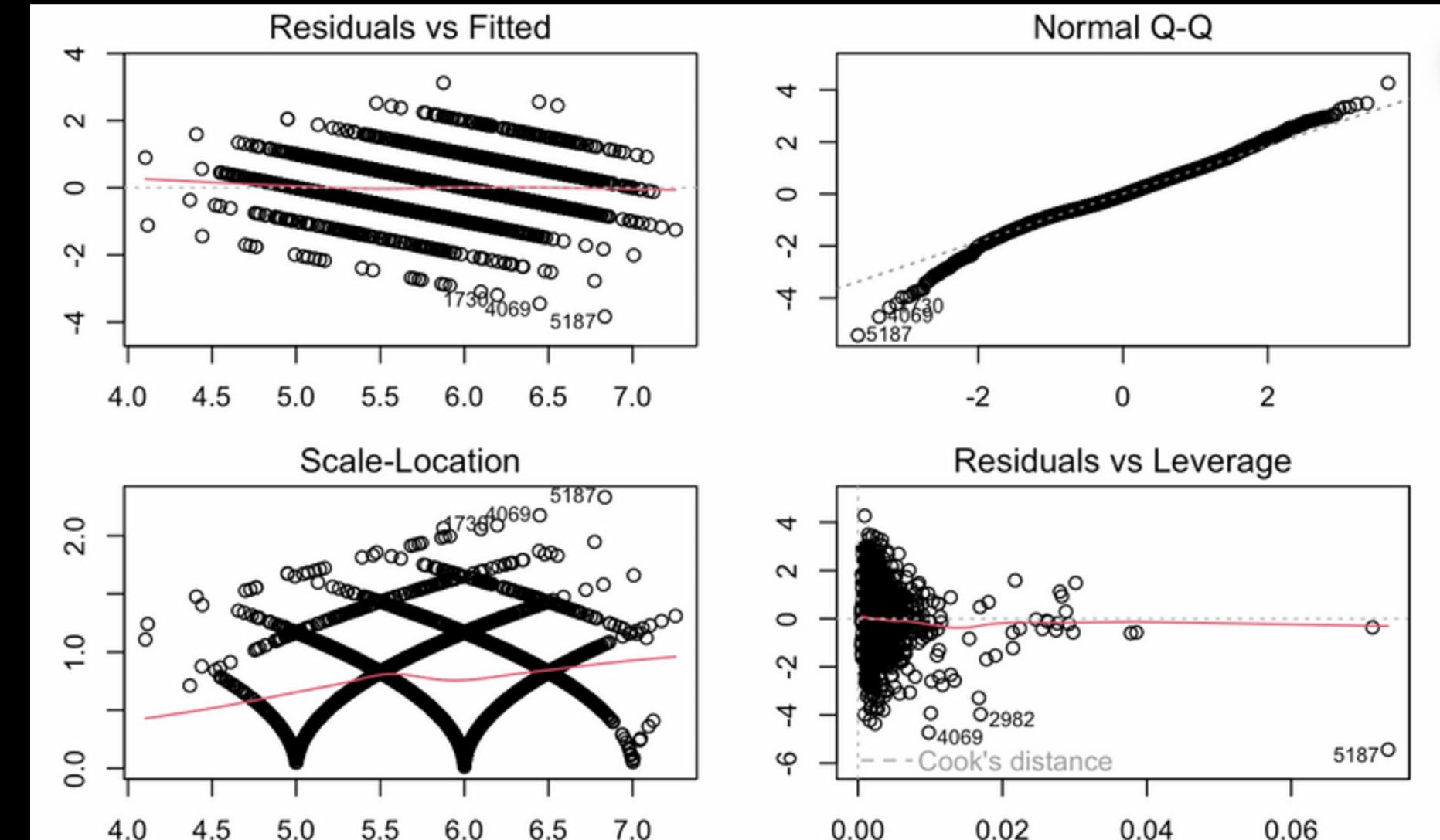
## F - test

- Có thực sự tồn tại mối liên hệ giữa biến  $X_j$  và  $Y$  ( $pValue \beta_i < 0.05$ )
- Tính tuyến tính của dữ liệu
- Phần dư có trung bình bằng 0
- Phần dư có phân phối chuẩn
- Phương sai của phần dư là không đổi

## ANOVA

## Kiểm định khoảng tin cậy cho trung bình

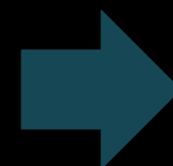
## Ước lượng khoảng tiên đoán



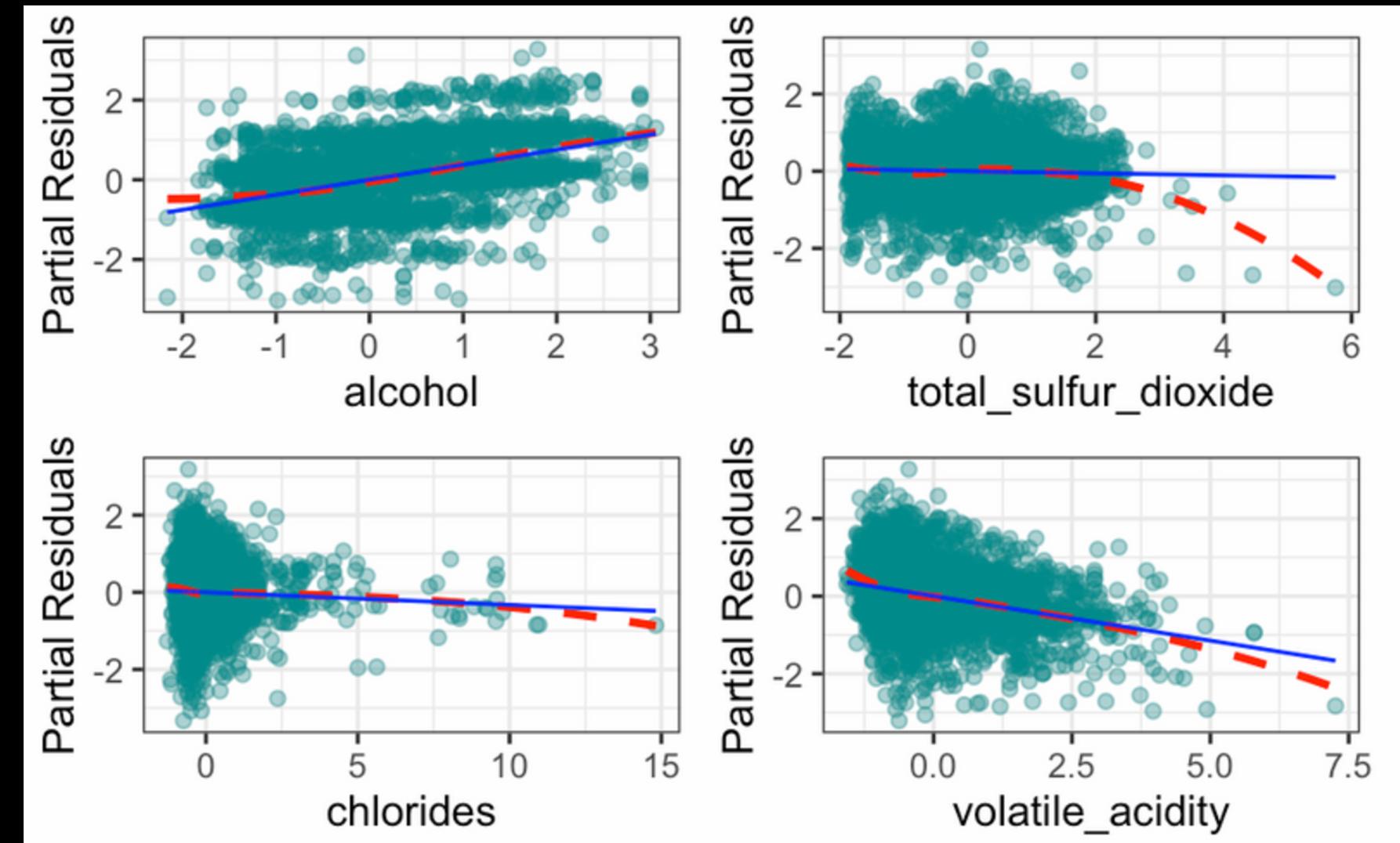
# CHUẨN ĐOÁN MÔ HÌNH

## Kiểm tra tính tuyến tính từng phần

Có một số feature mà mối quan hệ giữa nó và biến mục tiêu quality không hẳn là tuyến tính



Sử dụng mô hình hồi quy mở rộng



# MỞ RỘNG MÔ HÌNH



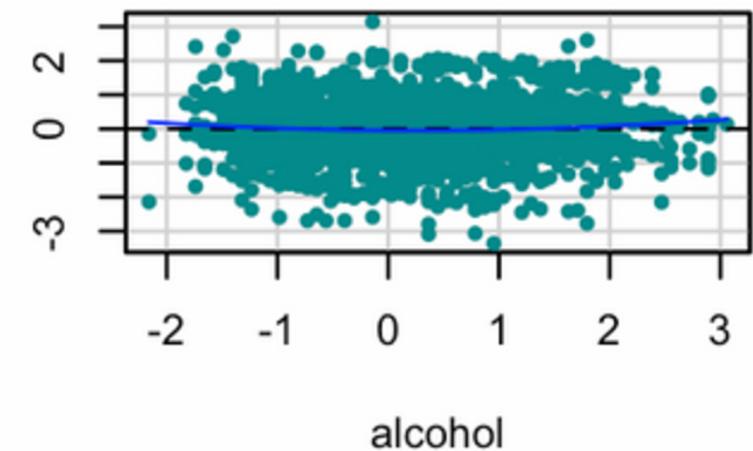
## Hồi quy đa thức

Áp dụng dạng đa thức bậc 2 cho các biến hồi quy từ linear model 5



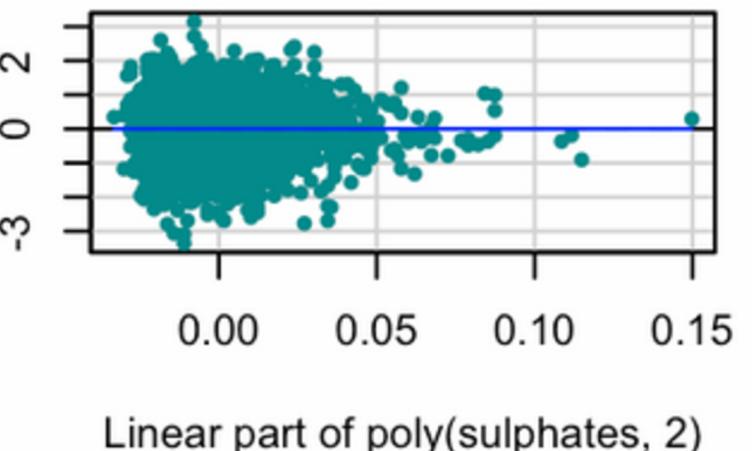
Tính tuyến tính từng phần

Pearson residuals



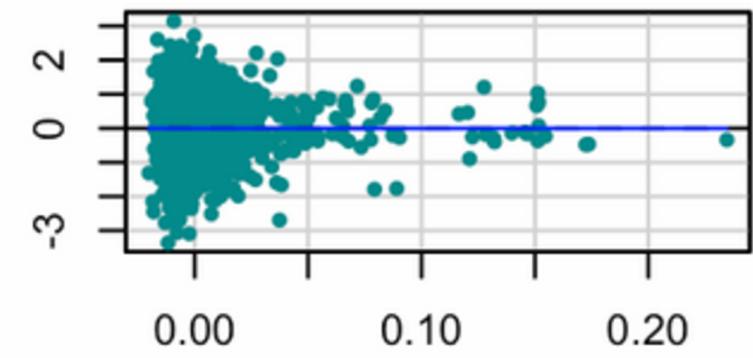
alcohol

Pearson residuals



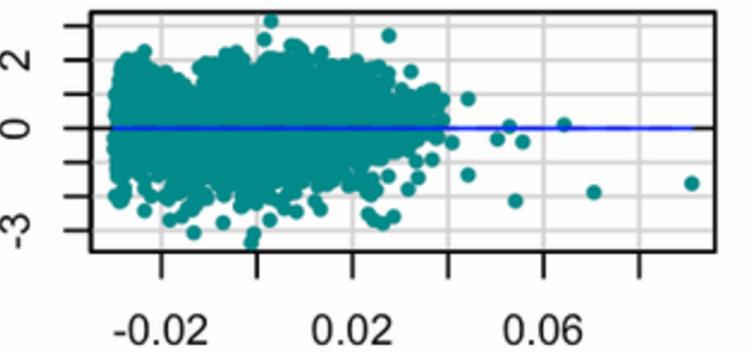
Linear part of poly(sulphates, 2)

Pearson residuals



Linear part of poly(chlorides, 2)

Pearson residuals



Linear part of poly(total\_sulfur\_dioxide, 2)

# MỞ RỘNG MÔ HÌNH

Hồi quy đa thức bậc 2



Tính tuyến tính của dữ liệu



Phân dư có trung bình bằng 0

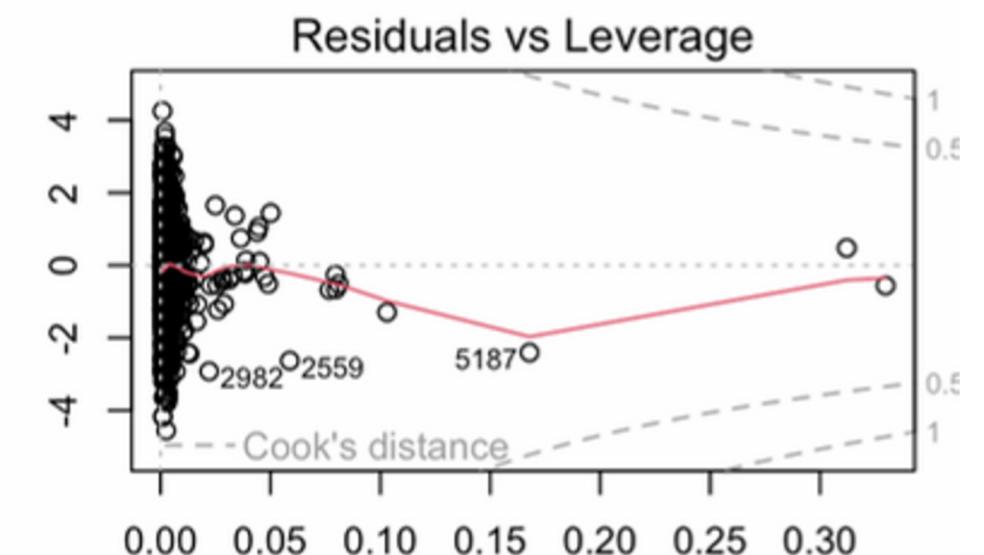
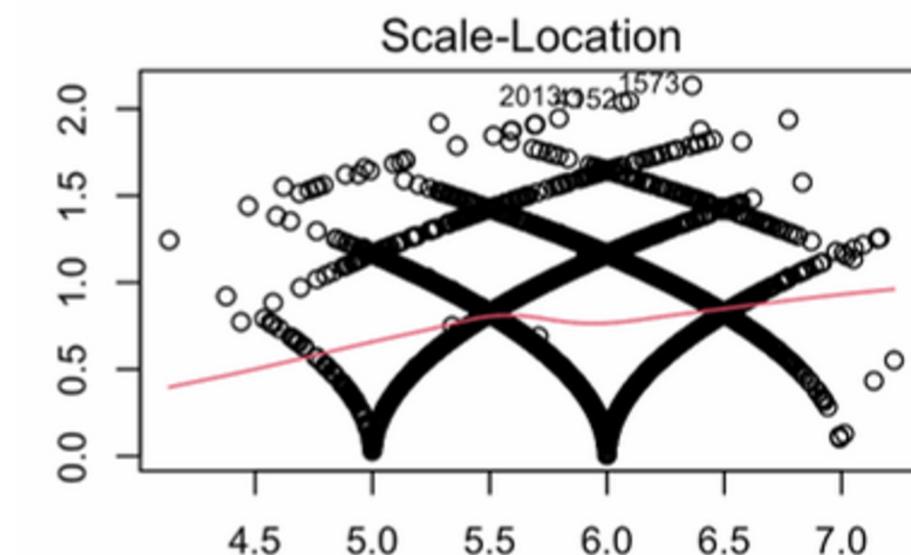
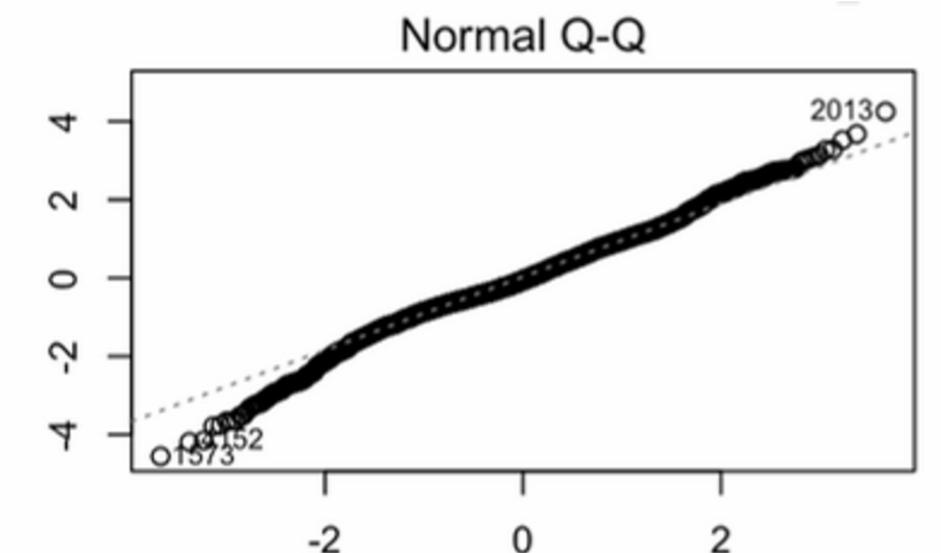
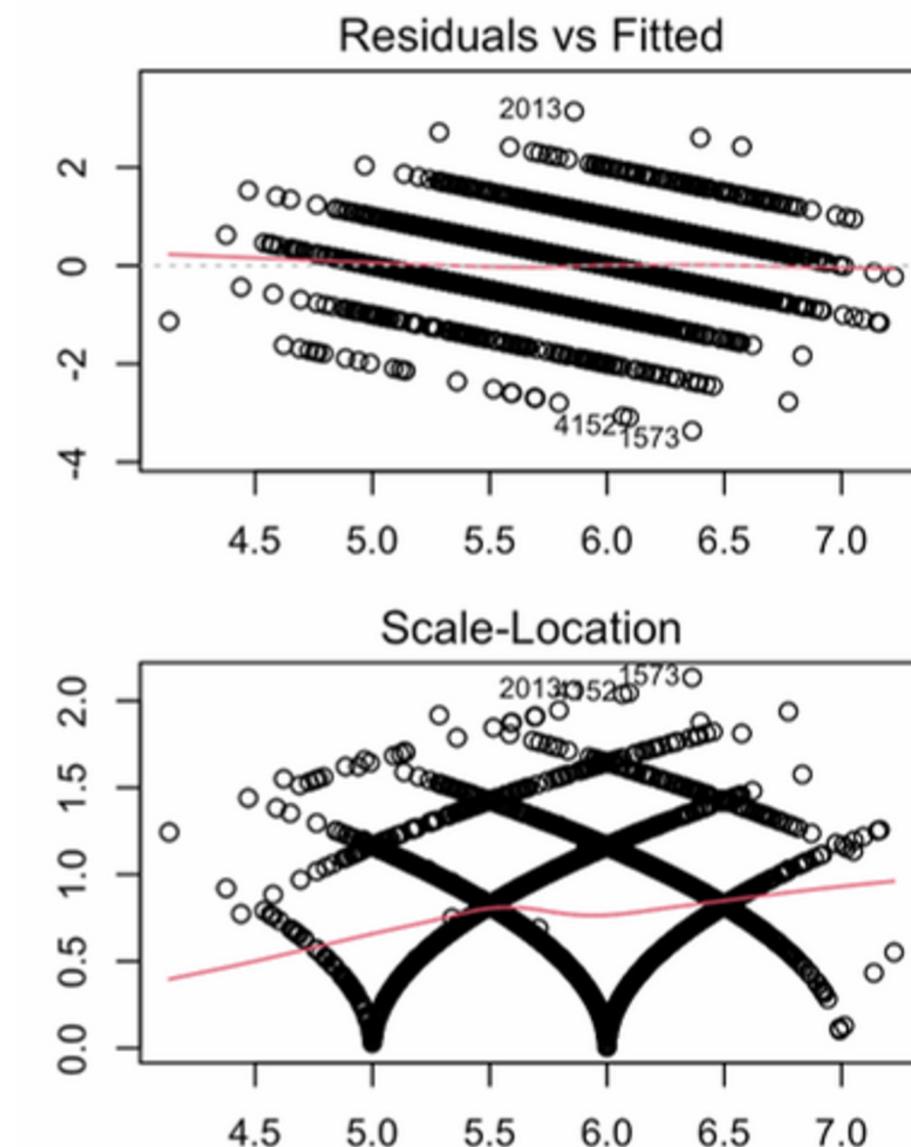


Phân dư có phân phối chuẩn



Phương sai của phân dư là không đổi

Mô hình là có ý nghĩa



fit quá tốt trên trainingset

# DỰ ĐOÁN KẾT QUẢ

R

Kết quả hồi quy

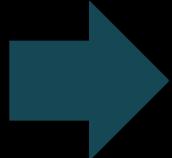
Số thực

N

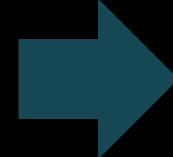
Điểm chất lượng rượu

Số nguyên

Index	5189	2218	179	153	2511
Kết quả hồi quy	6.012	5.702	5.629	5.344	5.995
Round()	6	6	6	5	6



Làm tròn kết quả hồi quy



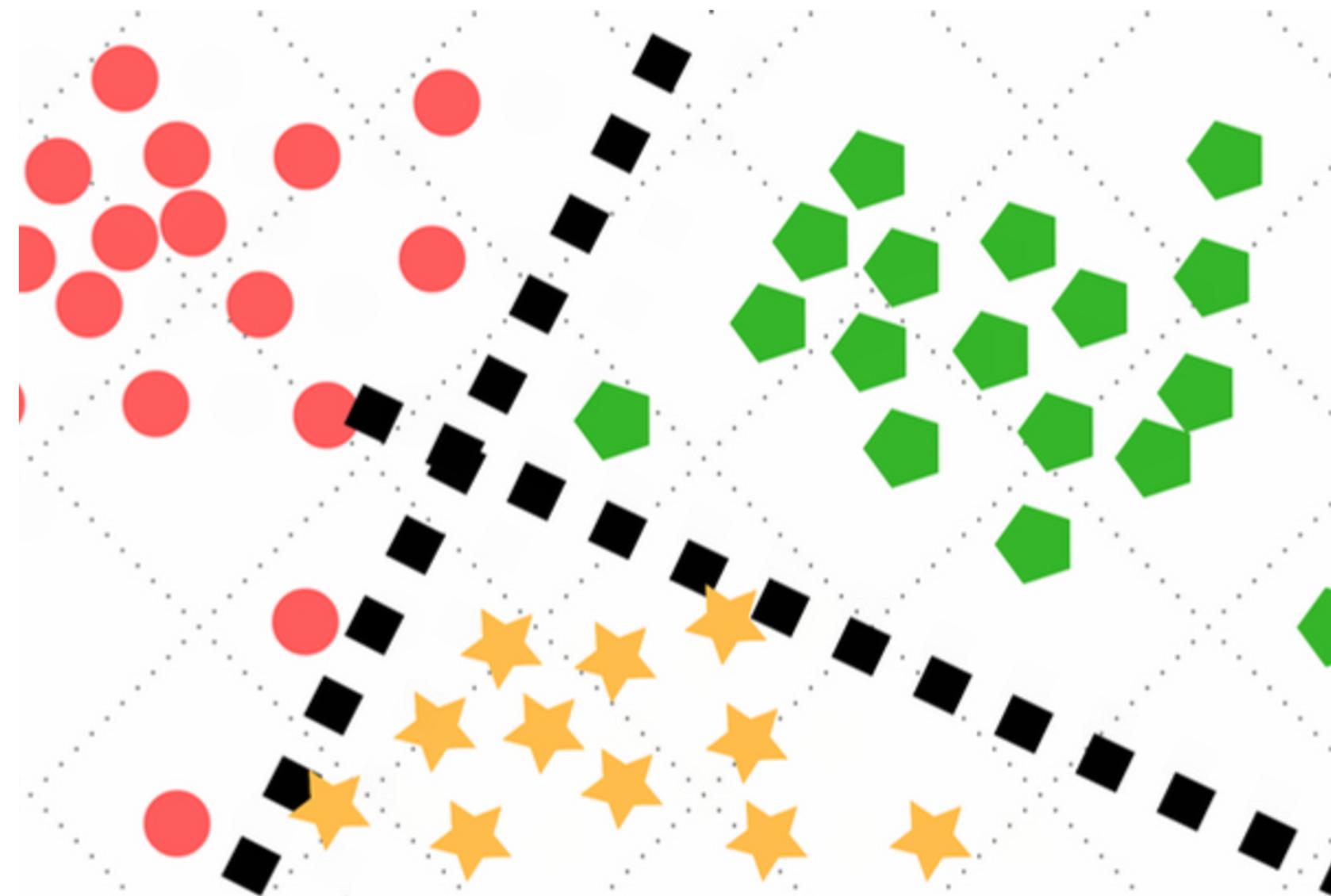
Dễ dàng phân loại chất lượng rượu & đánh giá mô hình

# XÁC ĐỊNH TOP 5 CÁC BIẾN QUAN TRỌNG

Dựa trên hệ số hồi quy của các biến trong linear model 5

	Ảnh hưởng tích cực	Ảnh hưởng tiêu cực	
Chống oxi hóa	free_sulfur_dioxide	total_sulfur_dioxide	Chất bảo quản
Cải thiện hương vị	residual_sugar	volatile_acidity	Giảm hương thơm, mùi vị
Cân bằng cấu trúc	alcohol		

## 2. MÔ HÌNH PHÂN LOẠI



1. CHUẨN BỊ DỮ LIỆU

2. XÂY DỰNG MÔ HÌNH

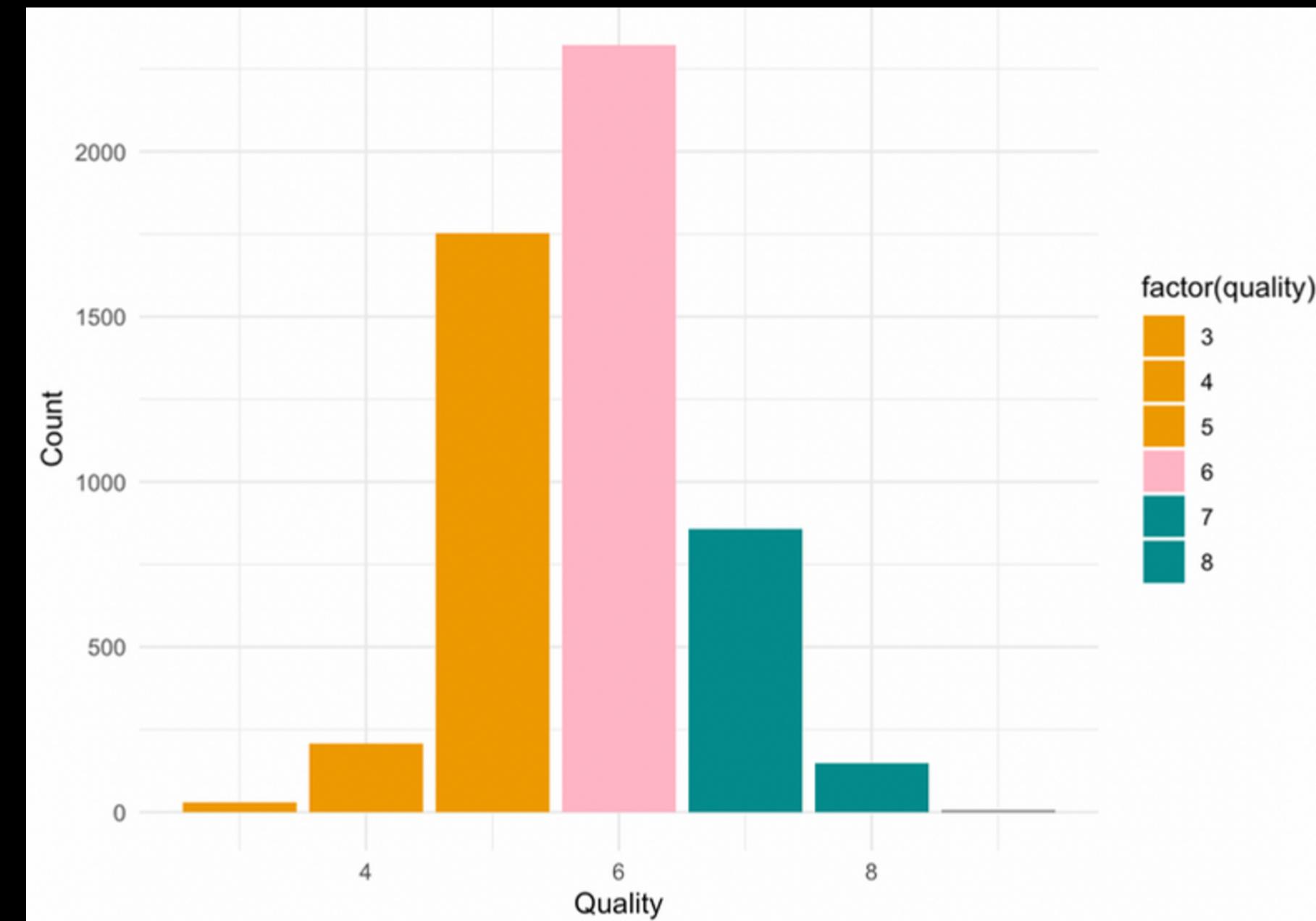
# CHUẨN BỊ DỮ LIỆU

Tạo một feature mới là ‘segmentation’ nhận vào ba giá trị:

- poor - điểm đánh giá từ 3 đến 5
- medium - điểm đánh giá là 6
- excellent - điểm đánh giá là 7 hoặc 8

Train-Test set split (`test_size = 0.25`)

Chuẩn hoá dữ liệu

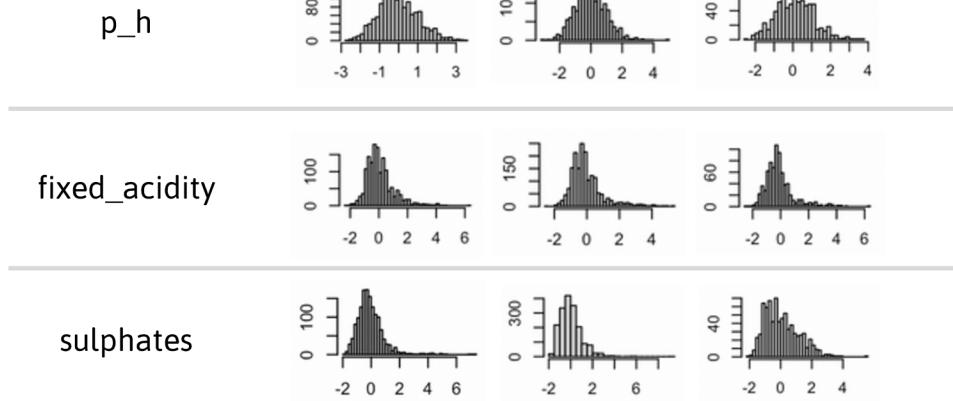


# XÂY DỰNG MÔ HÌNH

## Naive Bayes

segmentation ~ density +  
volatile\_acidity + p\_h +  
fixed\_acidity + sulphates

## Tính phân phối chuẩn



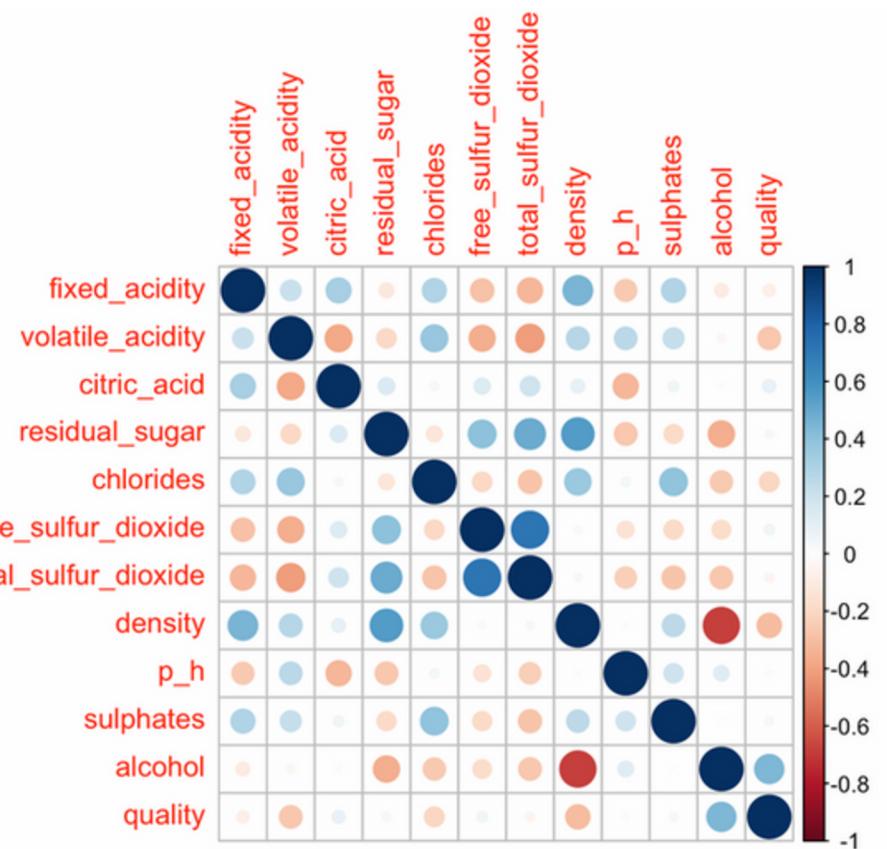
## LDA

## Xây dựng mô hình LDA theo tất cả các biến định lượng

segmentation ~ fixed\_acidity + volatile\_acidity + citric\_acid + residual\_sugar + chlorides +  
free\_sulfur\_dioxide + total\_sulfur\_dioxide + density + p\_h + sulphates + alcohol

## Tính độc lập

Hệ số tương quan của  
các cặp biến giải thích  
này đều nhỏ hơn 0.3



# XÂY DỰNG MÔ HÌNH

---

Multinomial  
Logistic

## Xây dựng mô hình phân loại theo tất cả các biến trong tập dữ liệu

segmentation ~ fixed\_acidity + volatile\_acidity + citric\_acid + residual\_sugar + chlorides + free\_sulfur\_dioxide + total\_sulfur\_dioxide + density + p\_h + sulphates + alcohol + color

## Thống kê suy luận

Kiểm định hệ số hồi quy cho các biến thành phần

## Xây dựng một mô hình tinh gọn hơn với số lượng biến ít hơn

Sử dụng 4 biến quan trọng nhất để làm các biến giải thích xây dựng mô hình  
segmentation ~ density + residual\_sugar + alcohol + color

# Xây dựng mô hình multinomial logistic theo tất cả các biến

---

```
## # weights: 42 (26 variable)
## initial value 4383.463032
## iter 10 value 3519.710380
## iter 20 value 3465.543810
## iter 30 value 3349.835695
## final value 3349.814554
## converged

## Call:
## nnet::multinom(formula = segmentation ~ ., data = datatrain_scaled,
##                 maxit = 500)
##
## Coefficients:
##             (Intercept) fixed_acidity volatile_acidity citric_acid residual_sugar
## medium      0.9171971   -0.6543810      0.2764179 -0.06031218     -0.7769028
## poor       -0.5473110   -0.7794531      0.9585154 -0.03831687     -1.2163011
## chlorides   chlorides free_sulfur_dioxide total_sulfur_dioxide density      p_h
## medium     0.2619832           -0.2075470      0.3214175 1.160668 -0.4873156
## poor       0.3012093           -0.5219812      0.6422195 1.540384 -0.6530659
## sulphates   sulphates alcohol      color
## medium    -0.3030049  -0.3151711  0.3185663
## poor      -0.5864910 -1.1156337  0.8730507
##
## Residual Deviance: 6699.629
## AIC: 6751.629
```

Mô hình hội tụ sau 30 vòng lặp, với giá trị log-likelihood cuối cùng là 3349.81

Multinomial Logistic

# THỐNG KÊ SUY LUẬN

---

```
##           (Intercept) fixed_acidity volatile_acidity citric_acid residual_sugar
## medium          0.1      2.0e-08        0.0013       0.37      4.2e-07
## poor            0.4      5.8e-09        0.0000       0.61      2.5e-12
##               chlorides free_sulfur_dioxide total_sulfur_dioxide density      p_h
## medium         0.023      3.4e-03        2.8e-03    7.1e-06   2.5e-10
## poor           0.011      8.1e-10        9.9e-08   1.4e-07   6.3e-13
##               sulphates alcohol color
## medium     9.2e-08 1.5e-02  0.324
## poor      0.0e+00 9.1e-14  0.019
```

p-Value > 0.05 cho thấy biến axit citric không có tác động có ý nghĩa thống kê.

Multinomial Logistic

# Mô hình multinomial logistic tinh gọn

---

```
##          Overall    Variables  
## density      2.701052   density  
## residual_sugar 1.993204 residual_sugar  
## fixed_acidity 1.433834 fixed_acidity  
## alcohol       1.430805   alcohol  
## volatile_acidity 1.234933 volatile_acidity  
## color         1.191617   color
```

Bốn biến quan trọng nhất được tìm thấy gồm có: density, residual\_sugar, fixed\_acidity và alcohol.

## Mô hình multinomial logistic với số lượng biến ít hơn

```
segmentation ~ density + residual_sugar  
+ alcohol + fixed_acidity
```

```
## # weights: 18 (10 variable)  
## initial value 4383.463032  
## iter 10 value 3694.689844  
## final value 3586.318867  
## converged
```

```
## Call:  
## nnet::multinom(formula = segmentation ~ density + residual_sugar +  
##   color + alcohol, data = datatrain_scaled, maxit = 1000)  
##  
## Coefficients:  
## (Intercept) density residual_sugar color alcohol  
## medium 1.728342 -0.07695876 0.03837985 -0.2903843 -0.8234473  
## poor 1.779757 -0.06121696 -0.18162521 -0.5803544 -1.8295888  
##  
## Std. Errors:  
## (Intercept) density residual_sugar color alcohol  
## medium 0.3952343 0.1292338 0.08898406 0.2214409 0.08431454  
## poor 0.4283464 0.1431318 0.09871566 0.2404099 0.09762952  
##  
## Residual Deviance: 7172.638  
## AIC: 7192.638
```



## VI. KẾT QUẢ ĐẠT ĐƯỢC

---

1

Nhận xét và kết luận về kết quả  
đạt được sau quá trình phân tích

2

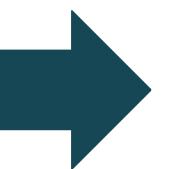
Các phương pháp cải tiến

# NHẬN XÉT

# QUÁ TRÌNH PHÂN TÍCH

Có sự khác biệt giữa rượu vang đỏ  
và rượu vang trắng

Các thành phần hóa học ảnh  
hưởng đến chất lượng rượu vang  
một cách độc lập với màu rượu



## Rượu vang đỏ

Nồng độ clorua, axit và  
sunfat cao

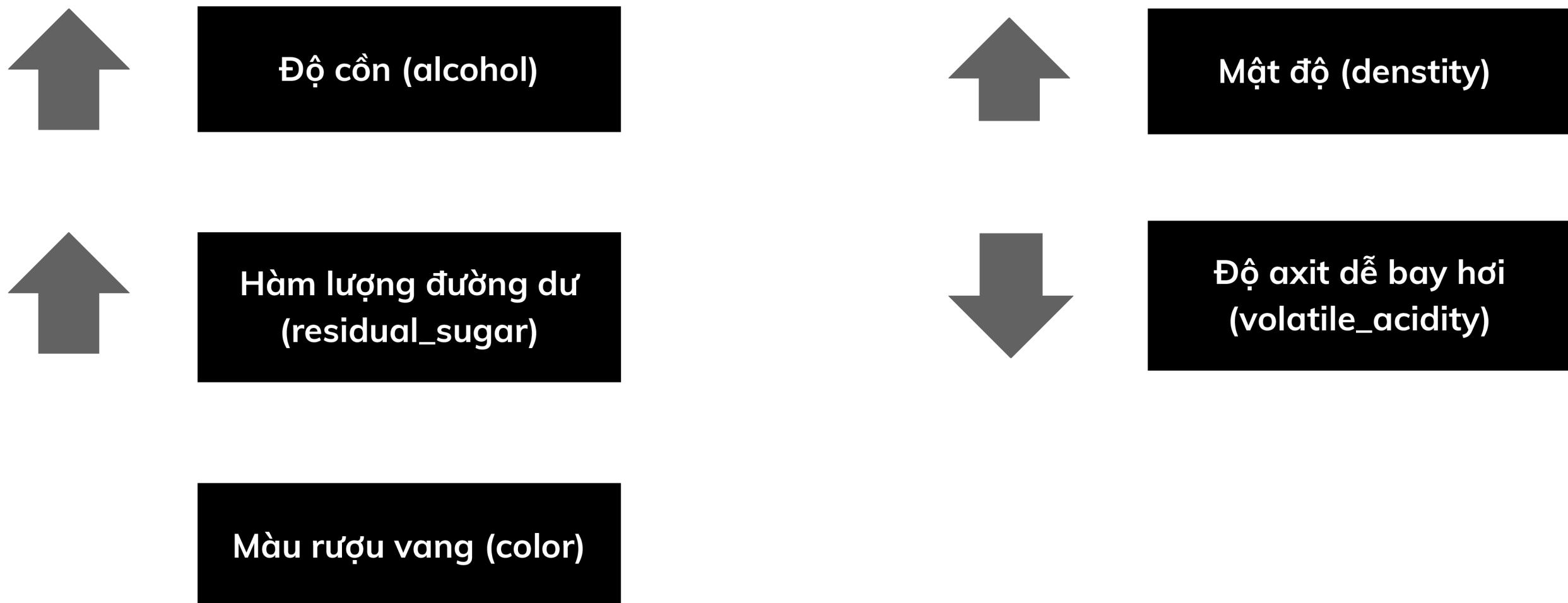
## Rượu vang trắng

Hàm lượng axit dễ bay hơi  
thấp.

Có thể sử dụng chung một model dự đoán  
chất lượng cho cả hai loại rượu vang

# NHẬN XÉT

# CÁC BIẾN QUAN TRỌNG



# NHẬN XÉT

# MÔ HÌNH HỒI QUY

	RMSE	MAE	R-squared	Accuracy
Linear Regression	0.8064	0.5271	0.2156	0.2135
Hồi quy đa thức	0.7625	0.4812	0.2463	0.2256
Ridge Regression	0.8032	0.5293	0.2120	0.2105

Hồi quy đa thức là mô hình đạt hiệu suất tốt nhất với điểm RMSE và MAE thấp nhất (0.7625 và 0.4812) chỉ ra khả năng dự đoán chất lượng rượu vang với sai số ít hơn so với các mô hình khác, và độ chính xác (0.2256) cao hơn so với các mô hình khác. Điều này đúng như dự đoán, vì hồi quy đa thức có khả năng mô hình hóa mối quan hệ phi tuyến, đây có thể là yếu tố giúp mô hình dự đoán chính xác hơn với thực tế.

# NHẬN XÉT

# MÔ HÌNH PHÂN LOẠI

Model	Accuracy
Gaussian Naive Bayes	0.5023
Kernel Naive Bayes	0.5564
LDA	<b>0.5805</b>
Multinomial Logistic	0.5759

LDA là mô hình phân loại đạt độ chính xác cao nhất (0.5805), chỉ ra khả năng phân loại chất lượng rượu vang chính xác hơn so với các mô hình khác. LDA là mô hình phân loại tuyến tính, cho phép giải thích dễ dàng hơn về mối quan hệ giữa các thành phần lý hóa sinh và chất lượng rượu vang. Ngoài ra, LDA có khả năng tính toán hiệu quả, phù hợp cho các ứng dụng thực tế cần xử lý lượng dữ liệu lớn.

# NHẬN XÉT

# CÁC VẤN ĐỀ CẦN QUAN TÂM

---



Không có thông số nào ảnh hưởng đáng kể đến chất lượng rượu vang (corr với quality < 0.5)



Chất lượng rượu vang phức tạp và chủ quan, phụ thuộc vào nhiều yếu tố.



Các mô hình dự đoán hiện tại có tiềm năng nhưng còn hạn chế

# ĐỀ XUẤT GIẢI PHÁP CẢI THIỆN MÔ HÌNH

---

Bổ sung thêm các  
yếu tố quan trọng

Loại nho, năm sản xuất, thời  
gian ủ, điều kiện bảo quản,  
nhiệt độ uống, độ ẩm,...

Phân tích kỹ hơn

Phân tích kỹ lưỡng hơn về mối  
liên hệ giữa các yếu tố được xác  
định và chất lượng rượu vang.

Thu thập thêm dữ liệu

Thu thập thêm dữ liệu về các  
loại rượu vang khác nhau và  
đánh giá của chuyên gia để cải  
thiện độ chính xác của mô hình.

# PHƯƠNG PHÁP CẢI TIẾN SẢN PHẨM



## Tinh chỉnh thành phần

Điều chỉnh nồng độ cồn, độ axit và pH có thể cải thiện chất lượng rượu vang



## Đảm bảo quy trình kiểm tra chất lượng

Kiểm tra tính nhất quán  
Kiểm soát chất lượng



## Tối ưu hoá quy trình sản xuất

Giám sát thời gian thực  
Phân tích dự đoán các vấn đề chất lượng tiềm ẩn



## Tối ưu hoá quá trình lên men

Giám sát và tối ưu hóa các thành phần chính trong quá trình lên men để đạt được hương vị và mùi thơm mong muốn.

## Giảm thiểu các thành phần không mong muốn

Giải quyết các yếu tố tạo ra thuộc tính không mong muốn, như độ axit dễ bay hơi, có thể nâng cao chất lượng rượu vang.

# CHIẾN LƯỢC KINH DOANH

---

## 1. Chiến lược tiếp thị

### Phân khúc khách hàng/ thị trường

phân loại rượu thành các phân khúc khác nhau, cho phép tiếp thị có mục tiêu đến tập khách hàng phù hợp.

### Lấy các thuộc tính chính làm USP (Unique Selling Point) của sản phẩm

Lấy các thuộc tính có sức ảnh hưởng được xác định trong mô hình để lên kế hoạch tiếp thị thu hút người tiêu dùng đang tìm kiếm những đặc tính cụ thể.

# CHIẾN LƯỢC KINH DOANH

---

## 2. Chiến lược thu hút khách hàng

### Đề xuất rượu vang

Cung cấp các đề xuất rượu được cá nhân hóa cho khách hàng dựa trên sở thích, nhu cầu và các giao dịch trước đó của họ.

### Tạo nội dung có giá trị giáo dục

Tổ chức các buổi workshop giúp người tiêu dùng hiểu hơn về rượu vang, thúc đẩy sự đánh giá và hiểu biết sâu sắc hơn về rượu.

---

# THANK YOU

NHÓM L - 21KDL1

---