

**UỶ BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

—🙐🕮🙐—



**MÔN HỌC: PHÂN TÍCH DỮ LIỆU**

**BÀI TẬP NHÓM**

**ĐỀ TÀI: PHÂN TÍCH KHÁM PHÁ VỀ CHẤT LƯỢNG RƯỢU VANG ĐỎ**

**THÀNH VIÊN**

LƯƠNG THANH TUẤN – 3122410447

**GIẢNG VIÊN HƯỚNG DẪN: ĐỖ NGỌC TÀI**

**TP. HỒ CHÍ MINH, 10/2025**

**MỤC LỤC**

[**DANH MỤC HÌNH ẢNH 1**](#_Toc211025139)

[**DANH MỤC BẢNG BIỂU 1**](#_Toc211025140)

[**LỜI CẢM ƠN 2**](#_Toc211025141)

[**I. GIỚI THIỆU CHUNG 3**](#_Toc211025142)

[**1. Lý do chọn đề tài 3**](#_Toc211025143)

[**2. Mục tiêu xây dựng đề tài 3**](#_Toc211025144)

[**3. Bảng phân công công việc 4**](#_Toc211025145)

[**3.1. Bảng tiến độ làm việc 4**](#_Toc211025146)

[**3.2. Bảng phân công chi tiết 4**](#_Toc211025147)

[**II. PHÂN TÍCH DỮ LIỆU 5**](#_Toc211025148)

[**1. Lấy dữ liệu 5**](#_Toc211025149)

[**2. Khám phá dữ liệu 5**](#_Toc211025150)

[**3. Kiểm tra tính toàn vẹn của dữ liệu 10**](#_Toc211025151)

[**4. Các tính chất thống kê dữ liệu số 11**](#_Toc211025152)

[**5. Mối liên hệ giữa các tính chất 12**](#_Toc211025153)

[**6. Boxplot 13**](#_Toc211025154)

[**7. Histogram 14**](#_Toc211025155)

[**8. Scatter Plot 16**](#_Toc211025156)

[**TÀI LIỆU THAM KHẢO 21**](#_Toc211025157)

# DANH MỤC HÌNH ẢNH

[**Hình 1. Thông tin dữ liệu 5**](#_Toc211025127)

[**Hình 2. Tính toàn vẹn của dữ liệu 10**](#_Toc211025128)

[**Hình 3. Tính chất thống kê dữ liệu số 11**](#_Toc211025129)

[**Hình 4. Mối liên hệ giữa các biến 12**](#_Toc211025130)

[**Hình 5. Độ phân tán của các cột 13**](#_Toc211025131)

[**Hình 6.1. Phân phối của fixed\_acidity, volatile\_acidity, citric\_acid 14**](#_Toc211025132)

[**Hình 6.2. Phân phối của residual\_sugar, chlorides, free\_sulfur\_dioxide 14**](#_Toc211025133)

[**Hình 6.3. Phân phối của total\_SO2, density, pH 15**](#_Toc211025134)

[**Hình 6.4. Phân phối của sulphates, alcohol, quality 15**](#_Toc211025135)

[**Hình 7.1. Phân phối của fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar 16**](#_Toc211025136)

[**Hình 7.2. Phân phối của chlorides, free\_S02, total\_S02, density 17**](#_Toc211025137)

[**Hình 7.3. Phân phối của pH, sulphates, alcohol 19**](#_Toc211025138)

# DANH MỤC BẢNG BIỂU

[**Bảng 1. Bảng tiến độ làm việc 4**](#_Toc211025121)

[**Bảng 2. Bảng phân công chi tiết 4**](#_Toc211025122)

[**Bảng 3. Bảng phân chia mức ảnh hưởng 20**](#_Toc211025123)

# 

# LỜI CẢM ƠN

Nhóm em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đến thầy Đỗ Ngọc Tài, giảng viên khoa Công nghệ Thông tin – Trường Đại học Sài Gòn, vì đã tận tình giảng dạy và hướng dẫn nhóm em trong suốt quá trình học tập cũng như hoàn thành bài báo cáo cuối kỳ này. Những kiến thức chuyên môn cùng với những kinh nghiệm thực tế mà thầy đã truyền đạt cho chúng em trong suốt khóa học không chỉ là nền tảng vững chắc giúp nhóm em hoàn thành tốt bài báo cáo, mà còn là nguồn động lực để chúng em tiếp tục rèn luyện, phát triển trên con đường học tập và nghề nghiệp sau này.

Trong quá trình thực hiện bài báo cáo, thầy đã không chỉ giúp nhóm em định hướng đúng đắn, giải quyết những khó khăn vướng mắc, mà còn truyền đạt những kỹ năng quý báu trong cách tiếp cận vấn đề và phương pháp làm việc hiệu quả. Nhờ có sự chỉ bảo tận tình và sự quan tâm sát sao của thầy, nhóm em đã có cơ hội học hỏi thêm nhiều kiến thức mới, từ lý thuyết cho đến thực hành. Những gì nhóm em học được không chỉ nằm trong khuôn khổ của bài báo cáo này, mà còn là hành trang cần thiết để ứng dụng vào các dự án thực tế và trong suốt quá trình làm việc sau này.

Dù đã cố gắng hết sức, nhưng với sự hạn chế về kiến thức chuyên môn và kinh nghiệm thực tiễn, nhóm em khó tránh khỏi những thiếu sót trong bài báo cáo. Nhóm em xin kính mong thầy thông cảm và bỏ qua cho những sai sót đó. Nhóm em rất mong nhận được những góp ý, nhận xét từ thầy, không chỉ để hoàn thiện tốt hơn bài báo cáo này mà còn giúp nhóm em rút ra những bài học quý báu, bổ sung và củng cố thêm những kỹ năng, kiến thức cần thiết để chuẩn bị tốt hơn cho những dự án, bài tập lớn và đặc biệt là bài báo cáo tốt nghiệp sắp tới.

Một lần nữa, nhóm em xin gửi lời cảm ơn sâu sắc đến thầy Đỗ Ngọc Tài. Kính chúc thầy luôn dồi dào sức khỏe, hạnh phúc và thành công trong sự nghiệp giảng dạy, tiếp tục dẫn dắt nhiều thế hệ sinh viên trên con đường chinh phục tri thức và vươn tới những thành công mới. Chúng em luôn trân trọng và biết ơn sự tận tâm của thầy.

*Nhóm em chân thành cảm ơn thầy!*

# GIỚI THIỆU CHUNG

## Lý do chọn đề tài

Mặc dù đề tài này do giảng viên phân công, việc thực hiện phân tích khám phá dữ liệu (EDA) cho bộ dữ liệu “Red Vinho Verde” (Bồ Đào Nha) thuộc bộ Wine Quality gốc UCI, được mirror trên Kaggle. Dữ liệu chỉ gồm thử nghiệm hoá lý và điểm cảm quan “quality”.

Bộ dữ liệu Wine Quality (Vinho Verde) được công bố trên UCI Machine Learning Repository và mirror trên Kaggle, gồm 11 chỉ tiêu hoá–lý (axit, đường dư, SO₂, pH, tỉ trọng, cồn, ion) cùng điểm cảm quan “quality”. Mục tiêu là dùng các phép đo phòng thí nghiệm để giải thích và dự đoán chất lượng rượu, từ đó hiểu cân bằng acid–cồn–đường–SO₂–ion ảnh hưởng đến mùi vị và ổn định vi sinh. Khi phân tích, ta rèn được trọn quy trình EDA, mô hình hoá (logistic, tree ensemble, v.v.), giải thích (feature importance/SHAP), đồng thời học cách đánh giá đúng (F1 macro, balanced accuracy) và tránh rò rỉ dữ liệu. Kết quả mang lại khuyến nghị thực hành cho nhà làm rượu (kiểm soát VA, đặt Free/Total SO₂ theo pH, cân bằng RS–alcohol) và một case study nhỏ gọn, dễ tái lập cho người học Data/ML.

## Mục tiêu xây dựng đề tài

* Mục tiêu chung
* Phân tích khám phá về chất lượng rượu vang đỏ.
* Mục tiêu cụ thể
* Nhìn tổng quan về dữ liệu và ý nghĩa của các cột
* Đặt ra các câu hỏi để khai phá dữ liệu và trả lời dựa vào biểu đồ
* Xây dựng câu chuyện về dữ liệu

## Bảng phân công công việc

### Bảng tiến độ làm việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MÃ SV** | **HỌ TÊN** | **Email** | **TỈ LỆ PHẦN TRĂM** |
| 1 | 3122410447 | Lương Thanh Tuấn | aquattda@gmail.com | 100% |

Bảng 1. Bảng tiến độ làm việc

### Bảng phân công chi tiết

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ TÊN** | **CÔNG VIỆC** | **TRẠNG THÁI** |
| 1 | Lương Thanh Tuấn | Xây dựng quy trình khai phá dữ liệu | Hoàn thành |
| Viết báo cáo | Hoàn thành |
| Tìm hiểu và đặt câu hỏi | Hoàn thành |
| Xác định vấn đề | Hoàn thành |

Bảng 2. Bảng phân công chi tiết

# PHÂN TÍCH DỮ LIỆU

## Lấy dữ liệu

* Bộ dữ liệu được thầy Đỗ Ngọc Tài chia sẻ đường link kaggle

## Khám phá dữ liệu

* **Thông tin dữ liệu:**

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1. Thông tin dữ liệu

* Bộ dữ liệu bao gồm 1599 dòng, 12 cột. Tất cả các cột đều không Null và giá trị số.
* Dữ liệu có 11 tính chất để phân lớp: Fixed acidity (g/L), Volatile acidity (g/L), Citric acid (g/L), Residual sugar (g/L), Chlorides (g/L), Free sulfur dioxide (mg/L), Total sulfur dioxide (mg/L)(mg/L), Density (g/cm³), pH, Sulphates (g/L), Alcohol (%)
* Dữ liệu để phân lớp ở cột **qualit**.
* **Mô tả chi tiết dữ liệu:**

1. **Fixed acidity (axit cố định)**

**Ý nghĩa:**

* Tổng các axit không bay hơi trong rượu (tartaric, malic, lactic,…)
* Phạm vi ~ 4-8 g/L (cao: vị chua gắt, thấp: rượu nhẹ nhưng nhạt)

**Phương pháp:** Tách phần bay hơi (chưng cất), sau đó lấy cần cặn đem chuẩn độ kiềm (NaOH ~ 0.10 N) tới pH ~ 8.2

1. **Volatile acidity (axit bay hơi)**

**Ý nghĩa:**

* Tổng các axit bay hơi trong rượu, thành phần chính là acetic acid. Ngoài ra có thể có một ít formic, propionic,…
* Phạm vi ~0.1-1.5 g/L (cao: có mùi giấm giảm chất lượng, thấp: ổn định)

**Phương pháp:** Thực hiện phương pháp chưng cất giống Fixed ecidity.

1. **Citric acid**

**Ý nghĩa:**

* Là một axit hữu cơ yếu, không màu (phổ biến ở họ cam quýt).
* Giúp tạo độ chua tươi (“freshness”) cho rượu, giảm độ pH và cân bằng cấu trúc cồn/đường.
* Phạm vi ~0–1.0 g/L. (cao vừa: tươi; quá cao: bị gắt)

**Phương pháp:**

Sử dụng Enzymatic assay (UV–Vis 340 nm):

1. Cho enzyme phản ứng với citric acid trong mẫu.
2. Trộn mẫu với dung dịch NADH + enzyme trong cuvet.
3. Đo ánh sáng ở 340 nm trước khi cho enzyme tác động (A₁).
4. Thêm enzyme (khởi động phản ứng), chờ vài phút, đo lại (A₂).
5. Tính ΔA = A₁ − A₂ → tra/bấm công thức hoặc đường chuẩn để ra g/L citric acid (nhân hệ số pha loãng nếu có).

**ΔA càng lớn thì citric acid càng nhiều**

1. **Residual sugar (glucose + fructose)**

**Ý nghĩa:**

* Là lượng đường dư còn lại chưa lên men trong rượu vang thành phẩm.
* Tạo cảm giác ngọt cho rượu.
* Phạm vi ~0.5-8 g/L

**Phương pháp:**

Sử dụng Enzymatic assay (UV–Vis 340 nm):

1. Cho enzyme phản ứng với citric acid trong mẫu.
2. Trộn mẫu với dung dịch NADH + enzyme trong cuvet.
3. Đo ánh sáng ở 340 nm với glucose (A₁).
4. Đo ánh sáng ở 340 nm với fructose (A2).
5. Tính RS = A₁ + A₂ (nhân hệ số pha loãng nếu có).
6. **Chlorides (clorua)**

**Ý nghĩa:**

* Lượng ion clorua trong rượu
* Phạm vi ~0.005–0.30 g/L (Clorua cao do quá trình sản xuất rượu vang đỏ là do các ion chiết xuất từ vỏ nho trong quá trình lên men.)

**Phương pháp:**

Chuẩn độ bạc potentiometric (AgNO₃ + điện cực Ag/AgCl)

Nhỏ vài giọt AgNO₃ vào mẫu rượu, lợi dụng Cl có trong rượu để kết hợp với Ag tại ra kết tủa Agcl (trắng).

Khi Cl hết thì thêm một chút AgNO₃, do không còn Cl để tạo kết tủa nên khi đo điện thế sẽ nhảy mạnh (điểm tương đương)

Khi bắt được điểm tương đương sẽ ra được thể tích AgNO₃

Sử dụng công thức thể tính nồng độ Cl⁻:

Trong đó:

CAgNo3​​: nồng độ AgNO₃ chuẩn (mol/L)

Veq​: thể tích AgNO₃ (L)

Mcl=35.45 (mg/mmol)

Vmẫu: thể tích mẫu đem chuẩn độ (L)

1. **Free sulfur dioxide (SO₂ tự do)**

**Ý nghĩa:**

* Là lượng SO2 chưa liên kết với các phân tử có trong rượu, nên có khả năng phản ứng để kháng khuẩn, chống oxy hóa.
* Phạm vi ~15-35 mg/L, ngoài ra có thể đo phạm vi bằng chỉ số pH:

**Phương pháp:**

Aeration–Oxidation (AO)

Dùng Axit hóa để tách Free SO2 bay ra khỏi dung dịch và đẩy SO2 sang bình chứa H₂O₂ để thành H₂SO₄ và chuẩn độ để ra được Free SO2 ban đầu.

1. **Total sulfur dioxide (SO₂ tổng)**

**Ý nghĩa:**

* Là tổng của SO2 tự do và SO2 đã liên kết, phản ánh mức độ bảo quản tổng thể, siết chặt an toàn và chất lượng.
* Phạm vi ~50-400 ml/L (Cao: Độ bảo quản tốt nhưng vi phạm quy định)

**Phương pháp:**

Aeration–Oxidation (AO)

Kiềm hoá/ủ để biến phần “bound” thành có thể giải phóng, rồi trong cùng phép đo dùng axit hoá + thổi khí để kéo tất cả SO₂ (free + bound đã giải phóng) sang bẫy và chuẩn độ.

1. **Density (tỉ trọng)**

**Ý nghĩa:**

* Là khối lượng riêng của rượu đo tại nhiệt độ chuẩn 20 °C, phản ánh sự cân bằng cồn và đường dư. Khi tỉ trọng cao thì lượng đường dư tăng, khi tỉ trọng giảm thì cồn giảm.
* Phạm vi ~0.987–1.040 g/cm³. (Density tăng thì residual tăng/ alcohol giảm và ngược lại).

**Phương pháp:**

Densimeter ống U dao động

Khử CO2 (lắc/siêu âm) để tránh sai số đo. Cho vào máy bơm mẫu cell ống U, trả về kết quả.

1. **pH**

**Ý nghĩa:**

* Cho biết mức độ axit/bazơ của rượu (thấp sẽ có vị chua. Ảnh hưởng mạnh tới ổn định vi sinh, màu, vị, và hiệu lực SO₂ (pH thấp → cùng một lượng free SO₂ cho nhiều SO₂ phân tử hơn → bảo quản tốt hơn).
* Phạm vi ~3.3 – 3.7. pH < 3.3 ổn định vi sinh nhưng dễ “gắt” nếu acid cao, pH > 3.7 nguy cơ vi sinh cao nên cần lượng Free SO2 nhiều hơn và cảm giác mềm/”phẳng”

**Phương pháp:**

Sử dụng máy HI84502 để pH-meter với điện cực thủy tinh, đo ở 20 °C.

Hiệu chuẩn 2–3 điểm bằng dung dịch đệm pH 7.00 → 4.00 (→ 3.00) trước khi đo.

Khử CO₂ nhẹ và khuấy đều để số ổn định; nhúng điện cực, chờ ổn, đọc giá trị.

1. **Sulphates (SO₄²⁻)**

**Ý nghĩa:**

* Là hàm lượng ion sunfat trong rượu, phụ gia liên quan hình thành SO₂.
* Phạm vi ~0.3-0.8 g/L (Cao: gây đắng)

**Phương pháp:**

Turbidimetric/Gravimetric (BaSO₄):

Thêm BaCl₂ tạo kết tủa BaSO₄, đo độ đục (máy quan với bước sóng phù hợp) và so với đường chuẩn để ra nồng độ SO₄²⁻ ban đầu.

1. **Alcohol (độ cồn)**

**Ý nghĩa:**

* Là lượng ethanol theo thể tích trong rượu, ảnh hưởng đến hương vị.
* Phạm vi ~12.0 – 14.5 % (tùy thuộc vào loại nho và cao sẽ gây nóng làm mất cân bằng).

**Phương pháp:**

Densimeter ống U dao động

Khử CO2 (lắc/siêu âm) để tránh sai số đo. Cho vào máy bơm mẫu cell ống U, trả về kết quả.

1. **quality (điểm cảm quan)**

**Ý nghĩa:**

* Là điểm đánh giá cảm quan tổng thể của rượu (mùi–vị–cấu trúc–cân bằng–hậu vị).
* Phạm vi: ~3-8

**Phương pháp:**

Do các chuyên gia và giám khảo có kinh nghiệm lâu năm.

## Kiểm tra tính toàn vẹn của dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.

Hình 2. Tính toàn vẹn của dữ liệu

* Dữ liệu không bị trùng và không có giá trị rỗng (NaN, Null).
* Tuy nhiên dữ liệu có 244 dòng bị trùng.

## Các tính chất thống kê dữ liệu số

A screenshot of a graph

AI-generated content may be incorrect.

Hình 3. Tính chất thống kê dữ liệu số

**Nhiều biến có khả năng đuôi phải:**

* 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).

**Trường hợp khác:**

- density (0.9), pH (3.3) → cân bằng

- quality (6 > 5.63) → có khả năng lệch trái

## Mối liên hệ giữa các tính chất

A screenshot of a graph

AI-generated content may be incorrect.

Hình 4. Mối liên hệ giữa các biến

|  |  |
| --- | --- |
| **Cột tương quan dương** | **Cột tương quan âm** |
| * fixed\_acidity - citric\_acid (0.67) * fixed\_acidity - density (0.67) * citric\_acid - density (0.36) * citric\_acid - sulphates (0.31) * residual\_sugar - density (0.36) * chlorides - sulphates (0.37) * totalSO2 - freeSO2 (0.67) * alocohol - quality (0.48) | * fixed\_acidity - pH (-0.68) * ctric\_acid - volatile\_acidity   (-0.55)   * ctric\_acid - pH (-0.54) * density - alcohol (-0.5) * density - pH (-0.34) * volatile\_acidity - quality (-0.39) |

## Boxplot

A screenshot of a graph

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

Hình 5. Độ phân tán của các cột

* Độ lệch đa số sẽ phân phối lệch phải và theo hướng giá trị lớn. Tuy nhiên, có vài cột density và pH sẽ phân phối thêm bên trái.
* Theo biểu đồ, hầu hết các cột chiếm đa số giá trị outlier ngoại trừ (citric\_acid, alcohol) chiếm phần nhỏ, cần xem xét các cột có giá trị outlier cao này.

## Histogram

A graph of a number of velocipedia

AI-generated content may be incorrect.

Hình 6.1. Phân phối của fixed\_acidity, volatile\_acidity, citric\_acid

* Từ biểu đồ của fixed\_acidity, volatile\_acidity và citric\_acid ta thấy rằng độ phân bố bị lệch phải.
* Tuy nhiên với biểu đồ fixed\_acidity lại cân đối khi có 1 đỉnh, 2 biểu đồ còn lại thì xuất hiện nhiều đỉnh.
* Đối với fixed\_acidity các giá trị nằm trong khoảng [7-8]
* Đối với volatile\_acidity các giá trị nằm trong khoảng [0.4 - 0.7]
* Đối với citric\_acid các giá trị nằm trong khoảng [0.0 - 0.1]

A graph of a graph with text

AI-generated content may be incorrect.

Hình 6.2. Phân phối của residual\_sugar, chlorides, free\_sulfur\_dioxide

* Từ biểu đồ của residual\_sugar, chlorides và free\_sulfur\_dioxide ta thấy rằng độ phân bố bị lệch phải
* Tuy nhiên cả 3 biểu đồ này đều khá cân đối
* Đối với residual\_sugar các giá trị nằm trong khoảng [1 - 2.5]
* Đối với chlorides các giá trị nằm trong khoảng [0.085 - 0.15]
* Đối với free\_sulfur\_dioxide các giá trị nằm trong khoảng [0 - 15]

A graph of a number of data

AI-generated content may be incorrect.

Hình 6.3. Phân phối của total\_SO2, density, pH

* Từ biểu đồ của total\_sulfur\_dioxide, density và pH ta thấy rằng độ phân bố bị lệch phải
* Tuy nhiên cả 3 biểu đồ này đều khá cân đối
* Đối với total\_sulfur\_dioxide các giá trị nằm trong khoảng [0 - 30]
* Đối với density các giá trị nằm trong khoảng [0.995 - 0.998]
* Đối với pH các giá trị nằm trong khoảng [3.2 - 3.4]

A graph of a graph

AI-generated content may be incorrect.

Hình 6.4. Phân phối của sulphates, alcohol, quality

* Từ biểu đồ của sulphates, alcohol ta thấy rằng độ phân bố bị lệch phải
* Tuy nhiên sulphates phân phối đều, đối với alcohol thì khá cân bằng. Ngoại trừ quality thì nhiều đỉnh khiến phân bố lộn xộn
* Đối với sulphates các giá trị nằm trong khoảng [0.50 - 0.75]
* Đối với alcohol các giá trị nằm trong khoảng [9 - 10]
* Đối với quality các giá trị nằm trong khoảng 5 và 6

## Scatter Plot

A screenshot of a test

AI-generated content may be incorrect.

Hình 7.1. Phân phối của fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar

**Dựa trên phân tích sơ bộ, ta phân tích từng yếu tố:**

**Fixed Acidity vs quality**

* Đối với Fixed Acidity biểu đồ cho thấy sự phân tán rộng của các điểm dữ liệu. Các loại rượu có chất lượng từ thấp đến cao (cụ thể là từ 4 đến 7) xuất hiện trên toàn bộ dải giá trị của fixed\_acidity, tuy nhiên với lượng Fixed Acidity [12-14 g/L] vẫn có tập trung chất lượng rượu khá (5,6).
* Axit cố định (Fixed Acidity) không phải là một yếu tố có ảnh hưởng rõ rệt hoặc một chỉ báo đáng tin cậy để dự đoán chất lượng rượu vang.

**Volatile Acidity vs quality**

* Đối với Volatile Acidity biểu đồ thể hiện một mối tương quan nghịch (âm) rõ ràng. Các loại rượu có chất lượng cao (điểm 6, 7, 8) tập trung chủ yếu ở vùng có nồng độ volatile\_acidity thấp (dưới 0.7), trong khi rượu chất lượng thấp hơn lại có xu hướng phân bố ở vùng nồng độ cao hơn.
* Axit dễ bay hơi (Volatile Acidity) là một chỉ số quan trọng, ảnh hưởng tiêu cực đến chất lượng rượu. Về mặt hóa học, nồng độ cao của axit axetic (thành phần chính của volatile\_acidity) sẽ tạo ra mùi giống giấm, một đặc tính không mong muốn, từ đó làm giảm chất lượng cảm quan của sản phẩm.

**Axit Citric vs quality**

* Đối với Axit Citric mặc dù dữ liệu có độ phân tán, một xu hướng tương quan thuận (dương) có thể được nhận thấy. Cụ thể, khi nồng độ citric\_acid vượt ngưỡng 0.5, mật độ các loại rượu có chất lượng cao tăng lên đáng kể.
* Axit citric (Citric Acid) là một yếu tố có ảnh hưởng tích cực đến chất lượng rượu. Về mặt cảm quan, axit citric góp phần tạo nên sự tươi mát (freshness) và hương vị trái cây cho rượu, giúp nâng cao trải nghiệm của người dùng.

**Residual sugar vs quality**

* Tương tự như fixed\_acidity, biểu đồ của residual\_sugar cho thấy các điểm dữ liệu phân tán rộng mà không tuân theo một quy luật hay xu hướng cụ thể nào.
* Đường tồn dư (Residual Sugar) không cho thấy mối liên hệ rõ ràng với chất lượng rượu vang. Điều này cho thấy rằng vị ngọt không phải là yếu tố quyết định đến việc một loại rượu được đánh giá là ngon hay không trong tập dữ liệu này.

A group of graphs showing different colors

AI-generated content may be incorrect.

Hình 7.2. Phân phối của chlorides, free\_S02, total\_S02, density

**Dựa trên phân tích sơ bộ, ta phân tích từng yếu tố:**

**Clorides vs quality**

* Biểu đồ Clorides **không** cho thấy một mối tương quan. Các loại rượu có chất lượng **từ thấp đến cao tuy tập trung vào nồng độ muối thấp (dưới 0.1 g/L) nhưng không cho thấy được là yếu tố gây ảnh hưởng đến chất lượng rượu.**
* Nồng độ muối là một yếu tố **không ảnh hưởng** đến chất lượng rượu**.**

**Free SO₂ vs quality**

* Tương tự như fixed\_acidity và residual\_sugar, biểu đồ của free\_SO₂ cho thấy các điểm dữ liệu phân tán rộng mà không hình thành một xu hướng rõ ràng. Ở mọi mức chất lượng, đều có sự hiện diện của rượu với hàm lượng SO₂ tự do từ thấp đến cao.
* Dựa trên biểu đồ, lưu huỳnh đioxit tự do không phải là một yếu tố dự báo đáng tin cậy cho chất lượng rượu vang.

**Total SO₂ vs quality**

* Mặc dù biểu đồ cũng cho thấy sự phân tán rộng, một khuynh hướng nhẹ có thể được nhận thấy. **Tuy vậy nó vẫn không đánh giá được ảnh hưởng đến chất lượng rượu.**
* Tổng nồng độ SO₂ không phải là yếu tố **ảnh hưởng đến chất lượng rượu. Tuy vào đó cần thắt chặt lượng** SO₂ **để đảm bảo an toàn sức khỏe người sử dụng**

**Density vs quality**

* Biểu đồ thể hiện một mối tương quan nghịch rất rõ ràng và mạnh mẽ. Rượu chất lượng cao tập trung dày đặc ở vùng có tỷ trọng thấp (dưới 0.993 g/cm³), trong khi rượu chất lượng thấp hơn có xu hướng ở vùng tỷ trọng cao hơn.
* Tỷ trọng là một yếu tố dự báo chất lượng quan trọng. Về mặt vật lý, tỷ trọng của rượu là một hàm số của hai thành phần chính: nồng độ cồn và đường tồn dư (đường làm tăng tỷ trọng). Như đã phân tích trước đó, rượu chất lượng cao thường có nồng độ cồn cao và lượng đường tồn dư không quá lớn. Sự kết hợp này dẫn đến tỷ trọng thấp hơn, biến density trở thành một chỉ báo tổng hợp hiệu quả cho chất lượng rượu.

A screenshot of a test

AI-generated content may be incorrect.

Hình 7.3. Phân phối của pH, sulphates, alcohol

**Dựa trên phân tích sơ bộ, ta phân tích từng yếu tố:**

**Density vs quality**

* Mặc dù trên toàn dải giá trị, độ pH không cho thấy một xu hướng tuyến tính rõ ràng, việc phân tích sâu hơn vào các khoảng cụ thể lại phát hiện ra một insight quan trọng. Trong khoảng pH từ 2.**9** đến dưới 3.**2**, có một sự tập trung đáng chú ý của các loại rượu chất lượng cao và không có sự hiện diện của rượu chất lượng thấp.
* Độ pH trong khoảng hẹp [2.**9**, 3.**2]** có thể là một yếu tố ảnh hưởng tích cực đến chất lượng. Điều này cho thấy một độ axit cao và được kiểm soát chặt chẽ có thể là điều kiện cần để đạt được chất lượng tốt. Yếu tố này cần được nghiên cứu sâu hơn để xác nhận đây là một "khoảng giá trị tối ưu".

**Sulphates vs quality**

* Biểu đồ cho thấy một vùng các điểm dữ liệu rõ rệt. Trong khoảng nồng độ từ 0.75 đến 1.2**0** g/L, có sự tập trung mạnh mẽ của các loại rượu có chất lượng tốt (điểm 6, 7). Ngoài khoảng này, dữ liệu trở nên phân tán hơn.
* Hàm lượng sunfat là một yếu tố có ảnh hưởng mạnh, đặc biệt trong khoảng [0.75, 1.2**0**]. Về mặt hóa học, sunfat (SO₄²⁻) là một nguồn cung cấp SO₂, một chất bảo quản và chống oxy hóa quan trọng. Việc duy trì nồng độ sunfat trong khoảng tối ưu này có thể giúp bảo vệ hương vị và cấu trúc của rượu một cách hiệu quả nhất.

**Alcohol vs quality**

* Tương tự các yếu tố trên, việc xác định khoảng giá trị quan trọng mang lại nhiều thông tin hơn là nhìn vào xu hướng chung. Dữ liệu cho thấy trong khoảng nồng độ cồn từ 1**1**% đến 14%, có sự hội tụ của các loại rượu chất lượng từ trung bình đến cao.
* Nồng độ cồn là một yếu tố then chốt, với khoảng tối ưu nằm trong khoảng [1**1**%, 14%]. Nồng độ cồn cao ảnh hưởng đến cảm nhận về hương vị và sự cân bằng tổng thể. Việc đạt được nồng độ cồn trong khoảng này dường như là một yếu tố quan trọng để sản xuất ra rượu vang chất lượng.
* **Từ tất cả các nhận định trên, ta phân chia các yếu tố thành các loại:**

|  |  |  |
| --- | --- | --- |
| **Ảnh hưởng mạnh** | **Ảnh hưởng nhẹ** | **Không ảnh hưởng** |
| Volatile axidity [<0.7]  Citric\_acid [>0.5]  Sulphates [0.75-1.2**0**]  Alcohol [1**1**-14] | Fixed axidity [12-14]  Total\_SO2 [70-120]  Density [< 0.993]  pH [2.8-3.0] | Chlories  Resudal Sugar  Free SO2 |

Bảng 3. Bảng phân chia mức ảnh hưởng

# TÀI LIỆU THAM KHẢO

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2017.

[2] W. Shields, *SQL QuickStart Guide: The Simplified Beginner’s Guide to Managing, Analyzing, and Manipulating Data With SQL*, 1st ed. ClydeBank Media LLC, 2019.

[3] M. Shron, *Thinking with Data: How to Turn Information into Insights*, 1st ed. O'Reilly Media, 2014.

[4] C. N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*, 1st ed. Wiley, 2015.

[5] A. Maheshwari, *Data Analytics Made Accessible*, 1st ed. DMM Direct, 2014.

[6] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, 2016.

[7] H. Chen, *Practical Data Science with Python*, Apress, 2019.

[8] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O'Reilly Media, 2019.

[9] N. C. Zakas, *Understanding ECMAScript 6: The Definitive Guide for JavaScript Developers*, No Starch Press, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[11] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, *Data Mining for Scientific and Engineering Applications*, Springer, 2001.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[13] J. Heaton, *Deep Learning and Neural Networks: Python, Theano, and Keras*, Heaton Research, 2018.

Dựa trên phân tích sơ bộ, **nồng độ cồn (alcohol)** có vẻ là yếu tố có tương quan rõ ràng nhất với chất lượng rượu vang. Các yếu tố khác như **đường tồn dư (residual sugar)**, **độ pH**, và **tỷ trọng (density)** không cho thấy mối liên hệ rõ ràng."

**2. Phân tích chi tiết từng biểu đồ (Các slide/phần tiếp theo)**

Đi vào từng biểu đồ một, giải thích ngắn gọn về những gì chúng ta thấy.

**Biểu đồ 1: Đường tồn dư (Residual Sugar) và Chất lượng (Quality)**

* **Mô tả:** Biểu đồ này thể hiện mối quan hệ giữa lượng đường còn lại và điểm chất lượng rượu.
* **Nhận định:** Không có xu hướng rõ ràng. Các điểm chất lượng từ thấp đến cao (3 đến 8) xuất hiện ở mọi mức đường tồn dư.
* **Kết luận:** Lượng đường tồn dư **không phải là một chỉ số tốt** để dự đoán chất lượng rượu vang. Chúng ta không thể nói rằng rượu ngọt hơn hay ít ngọt hơn sẽ có chất lượng cao hơn.

**Biểu đồ 2: Nồng độ cồn (Alcohol) và Chất lượng (Quality)**

* **Mô tả:** Biểu đồ thể hiện mối quan hệ giữa nồng độ cồn và chất lượng rượu.
* **Nhận định:** Có một **xu hướng tương quan thuận nhẹ**. Các loại rượu có chất lượng cao hơn (điểm 7, 8 - màu sẫm) có xu hướng tập trung ở khu vực nồng độ cồn cao hơn (từ 11% trở lên). Ngược lại, rượu chất lượng thấp hơn (điểm 3, 4, 5) lại phổ biến ở mức cồn thấp hơn.
* **Kết luận:** Nồng độ cồn **là một yếu tố tiềm năng** ảnh hưởng đến chất lượng. Rượu có nồng độ cồn cao hơn có khả năng được đánh giá chất lượng tốt hơn.

**Biểu đồ 3: Độ pH và Chất lượng (Quality)**

* **Mô tả:** Biểu đồ này cho thấy mối liên hệ giữa độ pH và chất lượng rượu.
* **Nhận định:** Các điểm dữ liệu phân bổ dày đặc ở trung tâm và không tạo ra bất kỳ xu hướng nào. Hầu hết các loại rượu, bất kể chất lượng, đều có độ pH trong khoảng 3.1 đến 3.5.
* **Kết luận:** Độ pH **không cho thấy mối liên hệ** với chất lượng rượu vang trong tập dữ liệu này.

**Biểu đồ 4: Tỷ trọng (Density) và Chất lượng (Quality)**

* **Mô tả:** Biểu đồ thể hiện mối quan hệ giữa tỷ trọng của rượu và điểm chất lượng.
* **Nhận định:** Có một **xu hướng tương quan nghịch rất nhẹ**. Các loại rượu chất lượng cao hơn (màu sẫm) dường như có tỷ trọng thấp hơn một chút.
* **Kết luận:** Mối quan hệ này không đủ mạnh để đưa ra kết luận chắc chắn, nhưng nó gợi ý rằng tỷ trọng thấp hơn có thể là một đặc điểm của rượu vang chất lượng cao.

**3. Đề xuất & Bước tiếp theo (Slide/Phần cuối cùng)**

Phần này quan trọng nhất vì nó cho thấy bạn không chỉ đọc biểu đồ mà còn suy nghĩ về hướng đi cho công việc.

* **Phát hiện chính:** Trong 4 yếu tố được phân tích, **nồng độ cồn** là yếu tố có ảnh hưởng tích cực và rõ ràng nhất đến chất lượng rượu vang.
* **Đề xuất:**
  1. **Tập trung vào Nồng độ cồn:** Đề nghị bộ phận sản xuất hoặc thu mua nên ưu tiên các dòng sản phẩm có nồng độ cồn cao hơn (ví dụ: trên 11%) nếu mục tiêu là nâng cao điểm chất lượng trung bình.
  2. **Phân tích sâu hơn:** Các yếu tố này có thể ảnh hưởng lẫn nhau. Đề xuất thực hiện các phân tích đa biến (ví dụ: mô hình hồi quy) để xem xét sự kết hợp của nhiều yếu tố cùng lúc sẽ ảnh hưởng đến chất lượng như thế nào.
  3. **Không ưu tiên các yếu tố khác:** Dựa trên dữ liệu này, việc điều chỉnh độ pH hay lượng đường tồn dư có thể sẽ không mang lại hiệu quả rõ rệt trong việc cải thiện chất lượng.