Análise Exploratória da Qualidade dos Vinhos Tinto com Base em Propriedades Físico-Químicas e Sensoriais

1st Marina Vasques Rodrigues
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
marinavasq18@alu.ufc.br

2nd Fábio Gabriel Esteves Ivo Gomes
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
fabiogabriel@alu.ufc.br

3rd Caio Vinícius Pessoa Freires
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
caiopessoa145@gmail.com

4th Fábio Agostinho da Silva Nascimento Filho
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
fabinhosnf@gmail.com

***Abstract*—Este trabalho apresenta uma análise exploratória do conjunto de dados "Wine Quality", focada em vinhos tintos. Foram avaliadas propriedades físico-químicas e sensoriais, considerando seus principais valores estatísticos. A análise inclui quatro abordagens: univariada incondicional, univariada condicional por classe, bivariada incondicional e multivariada incondicional. O objetivo é compreender e identificar relações entre os preditores e a qualidade dos vinhos tintos, fornecendo insights para possíveis modelos preditivos.**

***Index Terms*—análise exploratória de dados, vinhos tintos, estatística descritiva, boxplot, histogramas**

I. Introdução

O estudo da qualidade dos vinhos é relevante para a indústria e para consumidores. Este trabalho realiza uma análise exploratória do dataset "Wine Quality", avaliando variáveis físico-químicas e sensoriais.

II. Métodos

A. Descrição do Dataset

O dataset "Wine Quality" [1] contém 6.497 amostras, sendo 1.599 vinhos **tintos** e 4.898 vinhos **brancos**. As variáveis de entrada são fatores físico-químicos (como pH e densidade), enquanto a saída representa a avaliação sensorial, obtida pela média de pelo menos três especialistas, em uma escala de 0 a 10. Para esta análise, serão considerados apenas os dados referentes aos vinhos **tintos**, uma vez que pesquisas indicam que eles são os mais consumidos e preferidos pelos brasileiros. [2]

As 11 variáveis de entrada são:

1) **Acidez fixa (g/L):** ácidos naturais predominantes, influenciam frescor e aroma [3].
2) **Acidez volátil (g/L):** ácidos que evaporam facilmente, impactam sabor e aroma [4].
3) **Ácido cítrico (g/L):** presente em menor quantidade, equilibra acidez [3],[6].
4) **Açúcar residual (g/L):** açúcar restante após fermentação, influencia doçura [4].
5) **Cloretos:** teor de cloretos, maior próximo ao mar [5].
6) **Dióxido de enxofre livre (mg/L):** influência na preservação e estabilidade do vinho.
7) **Dióxido de enxofre total (mg/L):** soma do livre e ligado, afeta conservação.
8) **Densidade:** concentração de ácidos, açúcares e outros compostos.
9) **pH:** nível de acidez total, influencia sabor e estabilidade.
10) **Sulfatos (g/L):** contribuem para sabor e antioxidante natural.
11) **Álcool (% vol):** impacto na percepção de corpo e sabor.

B. Análise Monovariada Incondicional

A análise monovariada incondicional avalia cada preditor X_d individualmente, usando todas as N observações. Os passos são:

1) Plotagem de histogramas (incondicional)
2) Cálculo da média μ_d:

$$\mu_d = \frac{1}{N}\sum_{i=1}^{N} X_{i,d}$$

3) Cálculo do desvio padrão populacional σ_d:

$$\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{i,d} - \mu_d)^2}$$

4) Cálculo da assimetria (skewness) γ_d:

$$\gamma_d = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_{i,d} - \mu_d)^3}{\sigma_d^3}$$

C. Análise Bivariada Incondicional

Nesta etapa, foi realizada uma **análise bivariada incondicional** com o objetivo de identificar a relação entre os preditores físico-químicos do vinho tinto. Essa análise permite investigar o grau de associação entre duas variáveis numéricas, fornecendo indícios de possíveis colinearidades e interdependências que podem influenciar a variável resposta (*quality*).

1) Correlação linear de Pearson: A medida utilizada para quantificar o grau de associação linear entre duas variáveis X e Y foi o **coeficiente de correlação de Pearson** (r_{xy}), definido como:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde:

- x_i e y_i são os valores observados das variáveis X e Y;
- $\bar{x}$ e $\bar{y}$ são as respectivas médias amostrais;
- n representa o número de observações.

O valor de r_{xy} varia entre -1 e 1, indicando:

- $r_{xy} \approx 1$: forte correlação linear positiva (as variáveis aumentam juntas);
- $r_{xy} \approx -1$: forte correlação linear negativa (uma aumenta enquanto a outra diminui);
- $r_{xy} \approx 0$: ausência de correlação linear significativa.

2) Procedimentos de análise: Para o conjunto de dados do vinho tinto (*red wine*), foram consideradas as 11 variáveis físico-químicas disponíveis, excluindo-se a variável resposta *quality*. Assim, foram avaliadas todas as possíveis combinações de pares de preditores, totalizando:

$$\frac{11 \times 10}{2} = 55 \text{ pares distintos.}$$

O cálculo das correlações foi realizado por meio da função `corr()` da biblioteca `pandas`, e os resultados foram arredondados para duas casas decimais.

Além disso, foram gerados gráficos de dispersão (*scatter plots*) para cada par de variáveis, com o objetivo de visualizar o padrão de relacionamento entre elas. Essa visualização auxilia na identificação de possíveis relações lineares, outliers e agrupamentos.

Os gráficos foram produzidos utilizando as bibliotecas `matplotlib` e `seaborn`, com parâmetros ajustados para facilitar a leitura, como transparência (`alpha = 0.5`) e tamanho reduzido dos pontos (`s = 20`).

3) Critério de interpretação: Para fins interpretativos, adotaram-se as faixas de intensidade de correlação linear indicadas por Dancey e Reidy (2006):

Valor de $\|r\|$	**Interpretação**
$0,00 \leq \|r\| < 0,10$	Correlação desprezível
$0,10 \leq \|r\| < 0,30$	Correlação fraca
$0,30 \leq \|r\| < 0,50$	Correlação moderada
$0,50 \leq \|r\| < 0,70$	Correlação forte
$\|r\| \geq 0,70$	Correlação muito forte

Esses critérios foram utilizados na seção de Resultados para destacar as relações mais relevantes entre os preditores físico-químicos.

III. Resultados

A. Descrição do Dataset

A Tabela 1 apresenta um resumo estatístico das variáveis do dataset de vinhos tintos. Observa-se que as médias de atributos como ácido fixo, açúcar residual, dióxido de enxofre livre e dióxido de enxofre total possuem variações relevantes. Esses valores indicam diferenças importantes nas características físico-químicas dos vinhos, que podem influenciar diretamente a avaliação de sua qualidade. Por exemplo, a média da qualidade dos vinhos tintos é 5,87, com valor máximo observado de 8,0, mostrando que há uma faixa relativamente limitada de variação na avaliação em comparação a outros tipos de vinho.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	2.208702	3.311113	0.658149	10.397805	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	9.664060	0.154386	0.169507	1.159925	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	1.000000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	100.369000	4.010000	2.000000	14.900000	8.000000

Fig. 1: Resumo estatístico das variáveis do Vinho Tinto

B. Análise Monovariada Incondicional

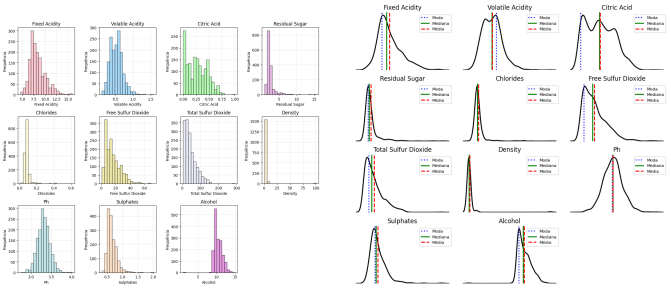
A análise monovariada incondicional permite estudar estatisticamente cada preditor individualmente. Para os vinhos tintos, observamos os histogramas de frequência, gráficos de assimetria e os valores de média, desvio padrão e assimetria apresentados na Tabela 1.

A partir desses resultados, podemos destacar:

1) **Média**: Os preditores com maiores valores médios nos vinhos tintos são o *Ácido Fixo* e o *Álcool*, indicando que, em geral, esses atributos apresentam níveis mais altos nos vinhos tintos.
2) **Desvio Padrão**: O *Dióxido de Enxofre Livre* e o *Dióxido de Enxofre Total* apresentam desvio padrão elevado, o que indica grande variabilidade entre os vinhos tintos para esses preditores.
3) **Assimetria**: O preditor *Sulfatos* apresenta a maior assimetria, sugerindo a presença de alguns vinhos com valores de cloreto significativamente maiores que a maioria. Além disso, a *Densidade* apresenta um valor de assimetria elevado, indicando que há vinhos tintos com densidade muito acima da média, o que contribui para a dispersão dos dados.

Preditor	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
fixed acidity	8.3196	1.7411	0.9828
volatile acidity	0.5278	0.1791	0.6716
citric acid	0.2710	0.1948	0.3183
residual sugar	2.5388	1.4099	4.5407
chlorides	0.0875	0.0471	5.6803
free sulfur dioxide	15.8749	10.4602	1.2506
total sulfur dioxide	46.4678	32.8953	1.5155
density	2.2087	9.6641	9.8039
pH	3.3111	0.1544	0.1937
sulphates	0.6581	0.1695	2.4287
alcohol	10.3978	1.1599	-0.5768

Fig. 2: Média, desvio padrão e assimetria das variáveis



(a) Histograma de frequência dos preditores do Vinho Tinto

(b) Assimetria das variáveis do Vinho Tinto

Fig. 3: Análise gráfica do Vinho Tinto

C. Análise Monovariada das Características Físico-Químicas dos Vinhos Tintos

A análise dos vinhos tintos considerou tanto os histogramas de frequência quanto os gráficos de assimetria das variáveis físico-químicas, permitindo uma avaliação mais detalhada das distribuições e do comportamento dos dados.

Acidez Fixa: Os histogramas indicam que a acidez fixa está concentrada entre 6 e 9, com leve assimetria à direita, mostrada pelos gráficos de assimetria. Isso evidencia que a maioria dos vinhos possui acidez moderada, com poucos casos de valores mais elevados.

Acidez Volátil: A distribuição concentra-se entre 0,3 e 0,7, com assimetria à direita observada tanto nos histogramas quanto nos gráficos de assimetria. Valores muito altos são raros, indicando que altos níveis de acidez volátil não são comuns em vinhos tinto.

Ácido Cítrico: A distribuição apresenta muitos valores próximos a zero, e a assimetria reforça que alguns vinhos possuem ácido cítrico significativamente mais alto, embora não sejam a maioria.

Açúcar Residual (Residual Sugar): Histogramas mostram forte concentração em valores muito baixos (0–2), caracterizando vinhos predominantemente secos. Os gráficos de assimetria confirmam a cauda à direita, representando poucos vinhos com açúcar residual mais elevado.

Cloretos (Chlorides): Distribuição extremamente concentrada próxima a zero, com baixa assimetria, indica baixo teor de cloretos na maioria dos vinhos.

Dióxido de Enxofre Livre (Free Sulfur Dioxide): Os histogramas e gráficos de assimetria indicam distribuição à direita, com a maioria dos vinhos até 20 mg/L e poucos valores elevados.

Dióxido de Enxofre Total (Total Sulfur Dioxide): Apresenta padrão semelhante ao enxofre livre, mas com maior variabilidade; gráficos de assimetria confirmam a tendência de valores extremos mais raros.

Densidade (Density): Distribuição altamente concentrada entre 0,990 e 1,005, com assimetria praticamente nula, evidenciando pequenas diferenças entre os vinhos.

pH: Distribuição aproximadamente normal, centrada em 3,3–3,5, com gráficos de assimetria mostrando leve tendência à direita, indicando estabilidade do perfil ácido.

Sulfatos (Sulphates): Distribuição assimétrica à direita, concentrada entre 0,4 e 0,8. Alguns vinhos apresentam valores maiores, como evidenciado pela cauda à direita nos gráficos de assimetria.

Álcool (Alcohol): Distribuição levemente assimétrica à direita, concentrando-se entre 9 e 12%, indicando teor alcoólico moderado na maioria das amostras.

Conclusão: A análise combinada dos histogramas e dos gráficos de assimetria revela que os vinhos tintos do dataset são majoritariamente secos, com baixo teor de acidez volátil, cloretos e dióxido de enxofre, e teor alcoólico moderado. A presença de assimetria à direita na maioria das variáveis evidencia que existem alguns vinhos com valores extremos, embora a maior parte das amostras se mantenha dentro de um perfil físico-químico predominante. Essa abordagem integrada proporciona uma compreensão mais robusta do perfil dos vinhos tintos, auxiliando em análises comparativas e decisões de produção.

D. Análise Bivariada Incondicional

1) Gráficos de Dispersão: A Figura 4 apresenta os gráficos de dispersão (*scatter plots*) gerados para todos os pares de variáveis físico-químicas do vinho tinto. Essa visualização permite identificar padrões lineares ou não lineares, agrupamentos e possíveis *outliers* nas variáveis.

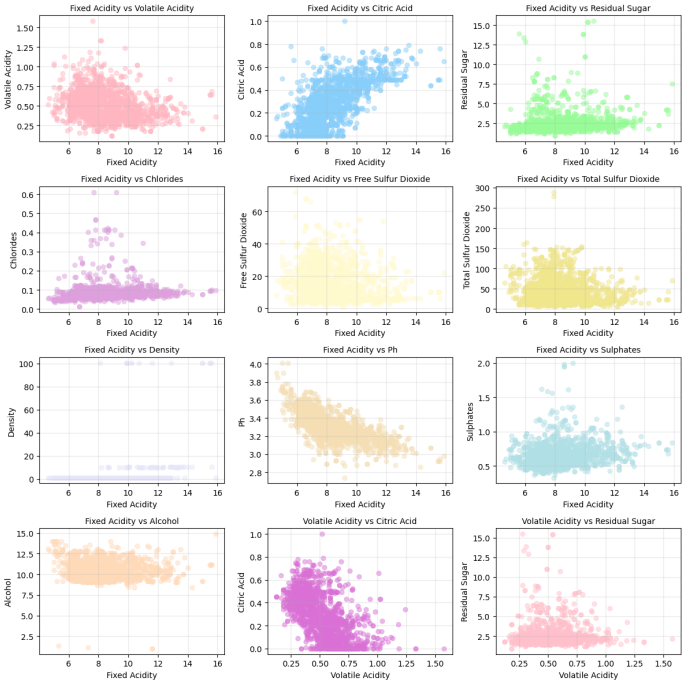


Fig. 4: Gráficos de dispersão entre os pares de variáveis físico-químicas do vinho tinto.

2) Matriz de Correlação: Para quantificar as relações lineares entre os preditores, foi calculada a matriz de correlação de Pearson (Tabela I). Observa-se que algumas variáveis possuem correlação forte, como *free sulfur dioxide* e *total sulfur*

dioxide ($r = 0.67$), enquanto outras apresentam correlação negativa, como *fixed acidity* e *pH* ($r = -0.68$).

TABLE I: Matriz de correlação de Pearson entre as variáveis físico-químicas do vinho tinto.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.26	-0.68	0.18	-0.06
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.00	0.23	-0.26	-0.19
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.09	-0.54	0.31	0.11
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.30	-0.09	0.01	0.05
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.07	-0.27	0.37	-0.20
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	0.09	0.07	0.05	-0.06
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.04	-0.07	0.04	-0.20
density	0.26	0.00	0.09	0.30	0.07	0.09	0.04	1.00	-0.12	0.05	-0.05
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.12	1.00	-0.20	0.19
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.05	-0.20	1.00	0.08
alcohol	-0.06	-0.19	0.11	0.05	-0.20	-0.06	-0.20	-0.05	0.19	0.08	1.00

Observa-se, de maneira geral, que:

- Variáveis relacionadas ao enxofre (*free sulfur dioxide* e *total sulfur dioxide*) apresentam forte correlação positiva.
- Algumas variáveis químicas possuem correlação negativa, como *fixed acidity* e *pH*, indicando que vinhos mais ácidos tendem a ter pH menor.
- A maioria das demais variáveis apresenta correlações fracas ou moderadas, sugerindo relações menos lineares ou mais complexas entre si.

IV. Referências

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[2] PORTAL INSIGHTS. Qual o vinho preferido dos brasileiros? Disponível em: ¡https://www.portalinsights.com.br/perguntas-frequentes/qual-o-vinho-preferido-dos-brasileiros¿. Acesso em: 19 out. 2025.

[3] Caveroyale, "Ácido Cítrico: Importância e Aplicações em Vinhos Premium," [Online]. Available: https://www.caveroyale.com.br/glossario/acido-citrico-importancia-aplicacoes-vinhos-premium/, acesso em: 28 set. 2025.

[4] Caveroyale, "Acidez Volátil: Entenda seu Impacto nos Vinhos Premium," [Online]. Available: https://www.caveroyale.com.br/glossario/acidez-volatil-vinhos-premium/, acesso em: 28 set. 2025.

[5] Agrovin, "Técnicas para corrigir a acidez do vinho," [Online]. Available: https://agrovin.com/pt-pt/tecnicas-para-corrigir-a-acidez-do-vinho/, acesso em: 28 set. 2025.

[6] Embrapa, "Metodologia de Análise de Vinho Tinto," [Online]. Available: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/887323/1/Metodologiaanalisevinhotintoed012010.pdf, acesso em: 28 set. 2025.

[7] Famiglia Valduga, "A importância da acidez no vinho," [Online]. Available: https://blog.famigliavalduga.com.br/qual-a-importancia-da-acidez-no-vinho/, acesso em: 28 set. 2025.