Ciência de Dados Aplicada ao Domínio dos Vinhos: Um Estudo com o *Dataset* X-Wines

Raíssa Gaiardo Girardi[1], Rogério Xavier de Azambuja[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul - *Câmpus* Farroupilha - Farroupilha, Rio Grande do Sul - Brasil

raissa.girardi@aluno.farroupilha.ifrs.edu.br;
rogerio.xavier@farroupilha.ifrs.edu.br;

Abstract. *This study performs an exploratory and statistical analysis on the X-Wines dataset, which contains thousands of wine labels and is widely used in data science. The aim was to identify consumption patterns based on user evaluations of wines, considering the physicochemical attributes found on the labels. Python code was developed using libraries such as Pandas, NumPy, and SciPy for data processing, statistical analysis, and validation, and executed on Google Colab. The results identified a weak correlation between attributes such as alcohol content and the average satisfaction score assigned to the wines. It is concluded that the public's perception of quality is defined by the sensory combination of wine attributes, reflecting its natural complexity.*

Keywords: *Data Science, Exploratory Analytics, Wines, X-Wines Dataset.*

Resumo*. O presente trabalho realiza uma análise exploratória e estatística no dataset X-Wines com milhares de rótulos de vinhos que é amplamente utilizado em ciência de dados. Buscou-se identificar padrões de consumo a partir da avaliação de vinhos por usuários considerando os atributos físico-químicos contidos nos rótulos. Foi desenvolvido um código na linguagem Python utilizando bibliotecas como Pandas, NumPy e SciPy para processamento, análise estatística e validação dos dados, executado no Google Colab. Os resultados identificaram uma fraca correlação entre atributos como o teor alcoólico e a nota média de satisfação atribuída aos vinhos. Conclui-se que a percepção de qualidade pelo público é definida pela combinação sensorial dos atributos do vinho, refletindo sua complexidade natural.*

Palavras-chave: *Ciência de Dados, Análise Exploratória, Vinhos, X-Wines Dataset.*

1.Introdução

O setor de vinhos representa um segmento relevante na economia global, associado tanto a aspectos culturais quanto comerciais, e que, nas últimas décadas, têm incorporado intensamente ferramentas de ciência de dados e inteligência artificial. Como por exemplo, sistemas de recomendação têm sido amplamente empregados em plataformas de e-commerce e serviços digitais, permitindo compreender padrões de consumo e oferecer experiências personalizadas aos usuários.

Nesse contexto, a utilização de bancos de dados específicos, como os produzidos pelo grupo de pesquisas X-Wines (de Azambuja, Morais e Filipe, 2023) - sistema de

recomendação de vinhos que utiliza informações dos rótulos dos vinhos para encontrar vinhos compatíveis com as preferências dos usuários, permitindo a busca por vinhos que harmonizam com pratos específicos, com um país de origem selecionado, entre outros - possibilita o desenvolvimento de análises exploratórias que não apenas contribuem para a evolução de técnicas de recomendação, mas também permitem investigar o comportamento de consumidores e as características intrínsecas dos vinhos avaliados.

Considerando a disponibilidade desses dados e a relevância das potenciais informações presentes no *dataset* X-Wines, objetivou-se realizar uma análise exploratória e estatística buscando identificar nos dados, correlações relevantes entre as avaliações proferidas pelos usuários aos vinhos e as características físico-químicas contidas nos rótulos dos mesmos. Para atingir esse objetivo, foram definidos os seguintes objetivos específicos: Explorar e pré-processar os dados disponíveis no *dataset* X-Wines e, investigar a relação entre variáveis físico-químicas dos vinhos e a percepção dos consumidores. A análise exploratória proposta pretende contribuir tanto para a literatura acadêmica quanto para aplicações práticas, ao fornecer subsídios para entender quais fatores impactam as preferências do consumidor.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a justificativa, evidenciando a relevância do tema e os motivos que fundamentam a realização deste estudo. A Seção 3 traz a revisão bibliográfica, reunindo os principais conceitos, teorias e contribuições pertinentes à pesquisa. Na Seção 4 são descritos os procedimentos metodológicos adotados, detalhando as etapas, métodos e instrumentos utilizados. A Seção 5 apresenta os resultados obtidos, enquanto a Seção 6 discute esses achados à partir do referencial teórico. Por fim, a Seção 7 reúne as considerações e perspectivas, destacando implicações, limitações e possíveis direções para trabalhos futuros.

2. Justificativa

A ciência de dados tem se consolidado como um campo multidisciplinar dedicado à coleta, mineração, tratamento, análise e interpretação de grandes volumes de dados, com o propósito de extrair conhecimento capaz de apoiar processos de tomada de decisão (Rautenberg & Carmo, 2019). A integração entre estatística, aprendizado de máquina e ferramentas computacionais tem permitido o desenvolvimento de métodos robustos para identificação de padrões, previsão de comportamento e compreensão de fenômenos complexos em áreas como saúde (Iwendi et al., 2020) e turismo (Li et al., 2018).

Uma das aplicações mais consolidadas da ciência de dados é o uso de modelos analíticos para compreender o comportamento de consumidores e avaliar atributos de produtos. Tais sistemas se baseiam em técnicas de análise que têm grande valor para empresas, especialmente na interpretação de padrões de consumo e preferências agregadas. Métodos como filtros colaborativos (Herlocker et al., 2004; Koren; Bell; Volinsky, 2009) utilizam dados de avaliações e características dos itens, permitindo identificar tendências coletivas, mapear segmentos de consumidores e compreender atributos de maior impacto na percepção de qualidade. Estratégias híbridas reforçam esse potencial ao combinar múltiplas fontes de informação (Shao; Li; Bian, 2021; Zhang; Lu; Jin, 2021).

No setor vitivinícola, esse tipo de conhecimento é particularmente valioso. A produção de vinhos envolve decisões relacionadas à escolha das uvas, definição de cortes, tempo de maturação, níveis de álcool, acidez e corpo, todos elementos que influenciam tanto a qualidade sensorial quanto a aceitação pelo mercado. Obras como *Wine Folly* (Puckette; Hammack, 2015) e *The Wine Bible* (MacNeil, 2015) destacam como as características químicas e organolépticas determinam estilos e percepções da bebida.

Assim, estudos recentes têm aplicado técnicas de ciência de dados à análise desses atributos, explorando correlações entre parâmetros físico-químicos e a avaliação dos consumidores, além de investigar modelos de classificação e predição de qualidade (Gkikas et al., 2023; He; Jiang; Gu, 2023).

Nesse cenário, destaca-se o *dataset* X-Wines proposto por de Azambuja, Morais e Filipe (2023), que reúne informações detalhadas contidas nos rótulos dos vinhos, avaliações de usuários e registros de navegação por uma plataforma Web desenvolvida com finalidade científica. Embora originalmente estruturado para testes de sistemas de recomendação e *machine learning*, o *dataset* constitui uma base valiosa para análises estratégicas para orientar a tomada de decisão empresarial, permitindo investigar como consumidores respondem a diferentes estilos de vinhos e suas características físico-químicas. Ao integrar múltiplas dimensões, atributos dos produtos, comportamento digital e avaliações, o *dataset* X-Wines possibilita análises exploratórias que ajudam a identificar preferências de mercado, avaliar a aceitação de determinados perfis sensoriais e detectar oportunidades para ajustes no portfólio de produção.

Assim, este trabalho se propõe a estudar como as características físico-químicas dos vinhos influenciam as preferências dos consumidores, buscando fornecer informações relevantes para empresas do setor vitivinícola, contribuindo para estratégias de produção, organização do portfólio e desenvolvimento de produtos mais alinhados às preferências identificadas nos dados.

3. Revisão Bibliográfica

3.1 Preferências de Consumo de Vinhos na Literatura

A compreensão das preferências dos consumidores de vinho tem sido objeto de diversos estudos, dada a relevância desses padrões para decisões produtivas, estratégias de mercado e desenvolvimento de novos rótulos. A literatura aponta que atributos físico-químicos, como corpo, acidez, teor alcoólico e composição varietal, influenciam diretamente a percepção sensorial e, consequentemente, as avaliações dos consumidores.

Outro aspecto explorado na literatura é o tipo de vinho. Estudos de Lesschaeve (2007) e Pickering et al. (2010) indicam que varietais tendem a ser mais facilmente compreendidos e aceitos por consumidores menos experientes, enquanto *blends* costumam atrair perfis mais familiarizados com nuances sensoriais.

A acidez também exerce papel significativo, influenciando a percepção de frescor e equilíbrio; consumidores iniciantes normalmente preferem acidez moderada, ao passo que entusiastas demonstram maior tolerância a níveis mais elevados (Reich; Maguire, 2018).

Além dos atributos físico-químicos, fatores culturais e contextuais também influenciam as preferências. Trabalhos como os de Silveira, Monticelli e Barbosa (2024) o perfil do consumidor varia entre regiões, refletindo tradições gastronômicas, contexto cultural e vínculos afetivos com produtos locais.

No conjunto, esses estudos indicam que preferências por corpo, acidez, teor alcoólico e estilo do vinho não apenas moldam o comportamento do consumidor, mas constituem insumos valiosos para a indústria vitivinícola. Assim, compreender como avaliações se relacionam com características físico-químicas pode apoiar decisões de produção, ajustes de portfólio e desenvolvimento de vinhos mais alinhados às demandas do mercado.

3.2 Principais Técnicas na Área da Ciência e Análise de Dados

A ciência de dados, além de integrar diferentes domínios, apoia-se em técnicas consolidadas de análise. Os métodos estatísticos fornecem a base inicial para compreender os dados, tanto pela estatística descritiva, que resume variáveis por meio de medidas de tendência central e dispersão (Montgomery; Runger, 2018), quanto pela estatística inferencial, que possibilita a generalização de resultados e testes de hipóteses.

No campo do aprendizado de máquina, técnicas supervisionadas, como regressão linear, árvores de decisão e redes neurais, são aplicadas para classificação e previsão, enquanto métodos não supervisionados, como o clustering, permitem identificar padrões ocultos sem rótulos pré-definidos (Hastie; Tibshirani; Friedman, 2009; Han; Kamber; Pei, 2011). Qualquer um desses métodos deve ser combinado à visualização de dados, um recurso indispensável para interpretação, que auxilia na identificação de tendências e relações complexas (Few, 2009).

3.3 Ferramentas Disponíveis no Mercado para Ciência e Análise de Dados

A aplicação das técnicas mencionadas é viabilizada por um ecossistema robusto de ferramentas. Entre as linguagens mais utilizadas, destaca-se Python que, com bibliotecas como pandas, numpy, scikit-learn e matplotlib, tornou-se referência para análise e visualização de dados (Python, 2024). O R também possui forte presença, especialmente em análises estatísticas avançadas e exploração de dados (Wickham; Grolemund, 2017).

Ambientes em nuvem como Google Colab democratizam o acesso, permitindo a execução de códigos Python em notebooks interativos gratuitos, com suporte a processamento paralelo (Bisong, 2019).

Em contextos corporativos, ferramentas de *Business Intelligence* como Tableau e Power BI ampliam a capacidade de visualização interativa, permitindo análises dinâmicas e painéis de monitoramento (Chaudhuri; Dayal; Narasayya, 2011).

Além disso, sistemas de gerenciamento de bancos de dados SQL (Structured Query Language) e NoSQL (Not Only SQL) sustentam o armazenamento e recuperação de informações em larga escala.

Complementarmente, plataformas de versionamento como GitHub (GITHUB, 2025) e GitLab (GITLAB, 2025) possibilitam colaboração distribuída e rastreabilidade de experimentos, o que se mostra fundamental para pesquisas aplicadas.

Esse conjunto de recursos torna possível a implementação de técnicas estatísticas e de aprendizado de máquina no estudo do consumo de vinhos, conferindo rigor e reprodutibilidade às análises desenvolvidas.

4. Procedimentos Metodológicos

4.1 Análise e Classificação dos Requisitos

A primeira etapa do desenvolvimento do sistema de informação consistiu na análise, identificação e classificação dos requisitos, que definem o comportamento esperado da solução proposta. Os requisitos foram levantados considerando as tarefas necessárias para realizar a análise estatística dos vinhos, incluindo leitura e processamento do *dataset*, padronização de variáveis, execução de cálculos estatísticos, geração de visualizações e armazenamento dos resultados. Também foram consideradas questões relacionadas à execução em *notebooks*, organização dos arquivos gerados e uso de ferramentas de versionamento.

A Tabela 1 apresenta os requisitos funcionais e a Tabela 2 os não funcionais identificados para o sistema proposto.

Tabela 1. Requisitos Funcionais

Requisitos Funcionais	
Código	**Descrição**
RF01	O sistema deve ler os arquivos no formato Comma Separated Values (CSV).
RF02	O sistema deve padronizar nomes de colunas e tratar valores ausentes.
RF03	O sistema deve calcular a média de notas e o número de avaliações por vinho.
RF04	O sistema deve identificar se o vinho é Varietal ou Blend com base nas uvas utilizadas em sua produção.
RF06	O sistema deve gerar e salvar gráficos (boxplot e dispersão) em formato PNG.
RF07	O sistema deve registrar logs de execução com informações e resultados.
RF08	O sistema deve salvar imagens e logs organizados em pastas no Google Drive.

Fonte: Elaboração própria, 2025

Tabela 2. Requisitos Não Funcionais

Requisitos Não Funcionais	
Código	**Descrição**
RNF01	O sistema deve processar datasets sem travar.
RNF02	O código deve ser executado facilmente no Google Colab, em células organizadas.
RNF03	Deve funcionar tanto no Colab quanto em ambiente local (Jupyter).
RNF04	O código deve ser modular, com funções auxiliares centralizadas em utils.py.
RNF05	Deve exibir mensagens de erro claras e salvar logs completos.
RNF06	As análises devem produzir os mesmos resultados a partir dos mesmos dados.
RNF07	Os gráficos devem ter boa legibilidade e serem salvos automaticamente.
RNF08	Os resultados devem ser salvos em estrutura de pastas padronizada (resultados//*logs*/imagens).

Fonte: Elaboração própria, 2025

4.2 Diagrama de Sequência

Com o objetivo de representar de forma clara o fluxo de interação entre o analista e o sistema, foi desenvolvido um diagrama de sequência, conforme ilustrado na Figura 1.

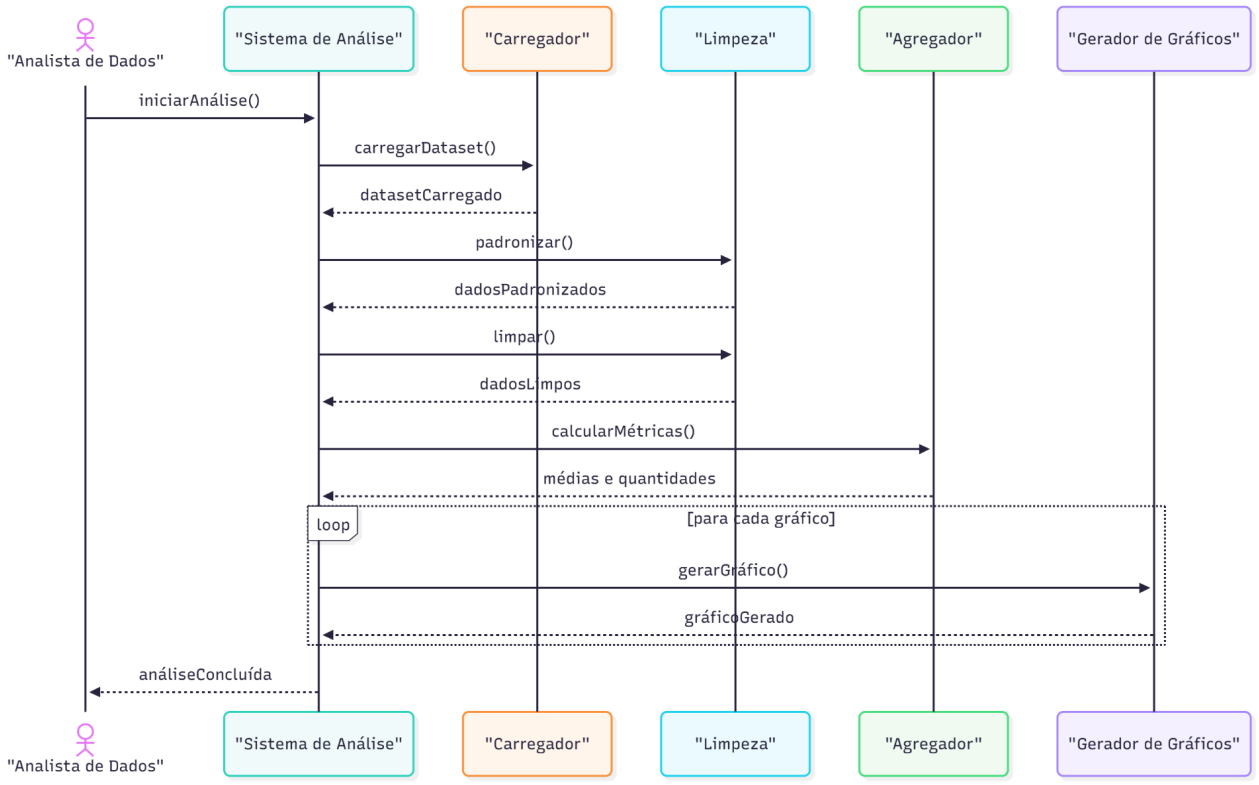

Figura 1. Diagrama de Sequência.

Fonte: Elaboração própria, 2025

A Figura 1 apresenta o diagrama de sequência do processo de análise de dados. O fluxo inicia quando o analista de dados solicita ao sistema de análise a execução da análise, que aciona o módulo carregador para importar o dataset. Em seguida, os dados são enviados ao módulo de Limpeza, onde passam por padronização e remoção de inconsistências. Após tratados, são encaminhados ao agregador, responsável pelo cálculo de métricas, médias e quantidades. Com essas informações, o gerador de gráficos cria as visualizações em um loop para cada gráfico solicitado. Por fim, o sistema de análise retorna ao analista a confirmação de conclusão do processo.

4.3 Metodologia

A análise de dados foi conduzida com o intuito de responder às perguntas de pesquisa sobre a influência das características físico-químicas dos vinhos (como teor alcoólico, acidez e corpo) e a comparação entre os estilos varietal e *blend* nas notas atribuídas pelos consumidores. A abordagem seguiu as etapas descritas abaixo:

4.3.1 Descrição do Conjunto de Dados

O banco de dados é composto por múltiplas tabelas e possui diferentes versões organizadas de acordo com o volume de informações e o propósito de uso. As informações contemplam atributos físico-químicos, características dos rótulos e avaliações de usuários. O *dataset* X-Wines é disponibilizado em três versões hierárquicas, definidas conforme o tamanho e o objetivo analítico (Tabela 3).

Tabela 3 – Caracterização do *dataset* X-Wines

Versão	Vinhos	Tipos de vinho	Países Vinícolas	Avaliações	Usuários	Avaliação de vinhos por múltiplos usuários
Test	100	6	17	1.000	636	Não
Slim	1.007	6	31	150.000	10.561	Não
Full	100.646	6	62	21.013.536	1.056.079	Sim

Fonte: de Azambuja, Morais e Filipe (2023).

Para a escolha das versões do *dataset* utilizadas nesta pesquisa, adotou-se exclusivamente a versão *Slim* do dataset, uma vez que ela reúne um conjunto de informações abrangente, organizado e adequado ao propósito analítico do estudo. A planilha ratings dessa versão contém 150 mil avaliações, volume que assegura representatividade estatística e permite identificar padrões consistentes no comportamento dos consumidores. A planilha wines, por sua vez, reúne cerca de mil registros e contempla todos os atributos relevantes (grapes, abv, acidity, body, entre outros), possibilitando uma caracterização completa das propriedades dos vinhos analisados. A utilização da versão *Slim* proporcionou um equilíbrio adequado entre profundidade analítica, clareza metodológica e eficiência no processamento, garantindo a integridade dos resultados obtidos.

Foram utilizadas duas tabelas principais:

Tabela *wines* versão *Slim*: Contém informações detalhadas sobre cada vinho. Possui 17 colunas, listadas a seguir:

- *WineID* – Identificador único do vinho

- *WineName* – Nome do vinho
- *Type* – Tipo do vinho (tinto, branco, rosé, espumante, sobremesa)
- *Elaborate* – Descrição do processo de elaboração
- *Grapes* – Variedade(s) de uva utilizadas
- *Harmonize* – Sugestões de harmonização com alimentos
- *ABV* – Teor alcoólico
- *Body* – Corpo do vinho (de muito leve até muito encorpado)
- *Acidity* – Acidez (baixa, média ou alta)
- *Code* – Código interno para o país (ISO-3166)
- *Country* – País de origem (ISO-3166)
- *RegionID* – Identificador da região
- *RegionName* – Nome da região produtora
- *WineryID* – Identificador da vinícola
- *WineryName* – Nome da vinícola
- *Website* – Página oficial do produtor
- *Vintages* – Safras disponíveis

Tabela *ratings* versão *Slim*: Contém historicamente as avaliações dos usuários, com a seguinte estrutura:

- *RatingID* – Identificador da avaliação
- *UserID* – Identificador do usuário (anonimizado)
- *WineID* – Identificador do vinho avaliado
- *Vintage* – Safra avaliada
- *Rating* – Nota dada ao vinho (escala de 1 a 5, incluindo variações de 0,5)
- *Date* – Data da avaliação

4.3.2 Pré-processamento e Limpeza de Dados

A primeira etapa envolveu a análise e o pré-processamento do *dataset* X-Wine. O objetivo foi preparar os dados para garantir sua qualidade e integridade para posterior análise. As principais ações realizadas foram:

- Carregamento dos dados: O conjunto de dados foi carregado a partir de arquivos no formato CSV contendo informações sobre os vinhos e as avaliações atribuídas pelos consumidores.
- Padronização: As colunas dos *datasets* foram padronizadas para evitar inconsistências nos nomes, como espaços e caracteres especiais. Algumas variáveis categóricas, como corpo e acidez, foram transformadas em valores numéricos para permitir a análise quantitativa, enquanto a variável *rating* foi tratada como numérica no formato *float* (Figuras 2 e 3). Essas padronizações foram feitas para garantir uma análise uniforme dos dados.

```
# Mapeamento de BODY (5 níveis)
wines = traduzir_coluna(
    wines,
    'body',
    {
        'very light-bodied': 1,
        'light-bodied': 2,
        'medium-bodied': 3,
        'full-bodied': 4,
        'very full-bodied': 5
    },
    'body_num'
)

# Mapeamento de ACIDITY (3 níveis)
wines = traduzir_coluna(wines, 'acidity', {'low': 1, 'medium': 2, 'high': 3})
```

Figura 2. Mapeamento numérico das variáveis.

Fonte: Elaboração própria, 2025

```
ratings = ratings.dropna(subset=['rating'])
ratings['rating'] = ratings['rating'].astype(float)
wines['abv'] = pd.to_numeric(wines['abv'], errors='coerce')
```

Figura 3. Tratamento das variáveis rating e teor alcoólico (ABV).

Fonte: Elaboração própria, 2025

- Limpeza e Filtragem de dados: Antes da aplicação das análises estatísticas, foi necessário realizar uma etapa de filtragem dos dados para otimizar e garantir a consistência e confiabilidade nos resultados.

Para realizar as análises, foram utilizadas apenas as colunas essenciais de cada tabela do *dataset*. Na planilha de ratings, utilizaram-se as informações de *wineid*, que identificam o vinho avaliado e a nota atribuída pelo usuário. Já na planilha *wines*, a seleção das colunas variou conforme o objetivo da análise. Para identificar se os vinhos eram varietais ou *blends*, utilizaram-se as colunas *wineid* e *grapes*, e para a análise das características físico-químicas, foram consideradas as colunas *wineid*, *abv*, *acidity* e *body*.

A primeira etapa consistiu na remoção de todos os registros com valores ausentes em variáveis essenciais, como as características físico-químicas (*ABV*, acidez, corpo) e as avaliações dos vinhos. Em seguida, os vinhos foram agrupados pelo identificador *wineid*, e para cada rótulo foi computado o número total de avaliações disponíveis.

Para aumentar a estabilidade estatística e evitar distorções causadas por vinhos com poucos dados, foi aplicado um critério mínimo de 10 avaliações por vinho, assim, apenas os vinhos que atenderam a esse limite foram mantidos no conjunto final analisado.

- Classificação dos vinhos: Para análise de comparação entre vinhos varietais (feitos com uma única variedade de uva) e vinhos *blend* (feitos com mais de uma variedade de uva), foi realizada a classificação dos vinhos com base na variável grapes da planilha *wines* versão *Slim*, que contém informações sobre as variedades de uvas utilizadas. Os vinhos foram classificados como varietais quando eram produzidos com apenas uma uva e como *blends* quando continham duas ou mais variedades de uvas na composição (Figura 4). Após a filtragem adotada na análise, identificaram-se 591 vinhos varietais e 260 vinhos classificados como *blends*, totalizando 851 rótulos analisados.

```
wines['n_grapes'] = wines['grapes'].apply(contar_uvas)
wines['is_varietal'] = (wines['n_grapes'] == 1).astype(int)
wines['wine_style'] = np.where(wines['is_varietal']==1, "Varietal", "Blend")

logging.info("Colunas derivadas 'n_grapes', 'is_varietal' e 'wine_style' criadas")
```

Figura 4. Identificação de varietal e *blend*.

Fonte: Elaboração própria, 2025

- Dados gerados: Foram calculados para cada vinho duas métricas fundamentais, a nota média e o número total de avaliações. Para isso, foi utilizada a biblioteca Pandas, aplicando o método *groupby("wineid")* sobre o conjunto de avaliações e, em seguida, as funções de agregação *mean*() e *count*() para obter, respectivamente, a média aritmética das notas e a quantidade de avaliações por rótulo (Figura 5). Esse procedimento permitiu consolidar em um único *dataset* informações essenciais sobre os vinhos e as avaliações dos consumidores, servindo de base tanto para as análises estatísticas (correlação e teste *t-student*) quanto para as comparações entre grupos de vinhos.

```
logging.info("Calculando nota média e número de avaliações por vinho...\n")

ratings_summary = (
    ratings.groupby("wineid")["rating"]
           .agg(num_ratings="count", avg_rating="mean")
           .reset_index()
)

wines = wines.merge(ratings_summary, on="wineid", how="left")
wines["num_ratings"] = wines["num_ratings"].fillna(0).astype(int)
wines["avg_rating"] = wines["avg_rating"].fillna(0).round(2)
```

Figura 5. Cálculo da Nota Média e Quantidade de Avaliações por Vinho.

Fonte: Elaboração própria, 2025

4.3.3 Método de análises de significância estatística e correlação das características físico-químicas, estilo e notas médias atribuídas

A próxima etapa focou na análise da correlação entre as características físico-químicas dos vinhos (ABV, acidez e corpo), estilo e as notas médias atribuídas pelos consumidores. As principais ações realizadas nesta etapa foram:

- Teste *t-student*: Para avaliar a representatividade das notas médias obtidas para os diferentes níveis de corpo (*body*) e acidez (*acidity*) foi utilizado teste *t-student*, que é um método estatístico usado para verificar se a diferença entre duas médias é significativa, considerando a variação dos dados e o tamanho de uma amostra. Ainda, foi utilizado teste *t-student* de duas amostras para avaliar se existe diferença estatisticamente significativa entre as notas médias atribuídas ao estilo de vinhos (varietal e *blend*) (Figura 6).

```
varietal = wines[wines['is_varietal']==1]['avg_rating']
blend    = wines[wines['is_varietal']==0]['avg_rating']

t_teste, p_val = ttest_ind(varietal, blend, equal_var=False)

print("\nTeste t-Student (Welch) — Vinhos (nota média por vinho)")
print("\nT = %.3f", t_teste)
print("\np-valor = %.5f", p_val)
```

Figura 6. Aplicação do Teste *t-student* para comparação entre vinhos varietais e *blends*.

Fonte: Elaboração própria, 2025

- Cálculo da correlação: Para avaliar a relação entre as características físico-químicas dos vinhos e a nota média atribuída pelos usuários, foi desenvolvido um código-fonte em Python utilizando as bibliotecas *pandas* para a manipulação dos dados e *numpy* para as operações numéricas auxiliares. O coeficiente de correlação de Pearson foi calculado por meio da função *pearsonr()*, disponível na biblioteca SciPy, que retorna simultaneamente o valor da correlação *(r)* e o p-value. O coeficiente r representa a força e a direção da relação linear entre duas variáveis, sendo amplamente utilizado em análises estatísticas (Figueiredo Filho; Paranhos; Santos, 2014). Já o p-value indica a probabilidade de que o resultado observado tenha ocorrido apenas ao acaso, assumindo que não exista uma relação real entre as variáveis. Esses cálculos fundamentaram as análises apresentadas (Figura 7).

```
r, p = pearsonr(sub[var], sub['avg_rating'])
resultados_corr.append({
    'Variável': var,
    'Correlação (r)': round(r, 4),
    'p-valor': p
})
```

Figura 7. Cálculo da Correlação de Pearson entre Variáveis e Nota Média dos Vinhos.

Fonte: Elaboração própria, 2025

4.3.4 Interpretação dos Resultados

A interpretação dos resultados estatísticos foi conduzida considerando diferentes métricas analíticas, permitindo avaliar tanto a intensidade das relações quanto sua significância estatística.

A classificação adotada para interpretar o coeficiente de correlação (r) foi: Correlação forte: $|r| \geq 0,7$; Correlação moderada: $0,4 \leq |r| < 0,7$; Correlação fraca: $0,2 \leq |r| < 0,4$; Correlação muito fraca ou inexistente: $|r| < 0,2$.

Além disso, o *p-value* associado ao coeficiente de correlação foi considerado significante se o valor de $p < 0,05$, correspondente a um intervalo de confiança de 95%, indicando que a associação observada é improvável de ter ocorrido ao acaso.

Para investigar diferenças dentro do grupo, como níveis das categorias de corpo e acidez, aplicou-se o teste *t-student*. Assim como na análise de correlação, os resultados foram interpretados com base no *p-value*.

4.3.5 Visualização dos Dados Gerados

A visualização dos resultados foi realizada com apoio das bibliotecas *matplotlib* e *seaborn*, permitindo representar graficamente os padrões identificados nas análises. Para a etapa de correlação entre características físico-químicas e notas médias, foram criados gráficos de dispersão com linha de regressão, facilitando a observação da relação entre ABV, acidez, corpo e as avaliações dos usuários (Figura 8). Já na análise comparativa entre vinhos varietais e *blends*, foi elaborado um *boxplot*, possibilitando comparar de forma clara a distribuição das notas entre os dois grupos. Essas visualizações atuam como suporte interpretativo às análises estatísticas, tornando os resultados mais intuitivos e acessíveis.

```python
plt.figure(figsize=(7,5))
sns.regplot(
    data=wines_complete,
    x="abv",
    y="avg_rating",
    scatter_kws={"alpha":0.4},
    line_kws={"linewidth":2}
)

plt.title("Correlação entre ABV e Nota Média", fontsize=14)
plt.xlabel("ABV", fontsize=12)
plt.ylabel("Nota Média", fontsize=12)

plt.tight_layout()
plt.savefig(os.path.join(IMGS_DIR, "correlacao_scatter_abv.png"), dpi=300)
plt.show()
```

Figura 8. Geração do gráfico de correlação entre ABV e nota média.

Fonte: Elaboração própria, 2025

5. Resultados

5.1 Características físico-químicas e nota média dos vinhos

A primeira etapa da análise investigou a relação entre as características físico-químicas dos vinhos (ABV, acidez e corpo) e a nota média atribuída pelos avaliadores.

5.1.1 Teor alcoólico (ABV) e nota média

Inicialmente, foi gerado um gráfico de dispersão considerando todos os vinhos da amostra (Figura 9).

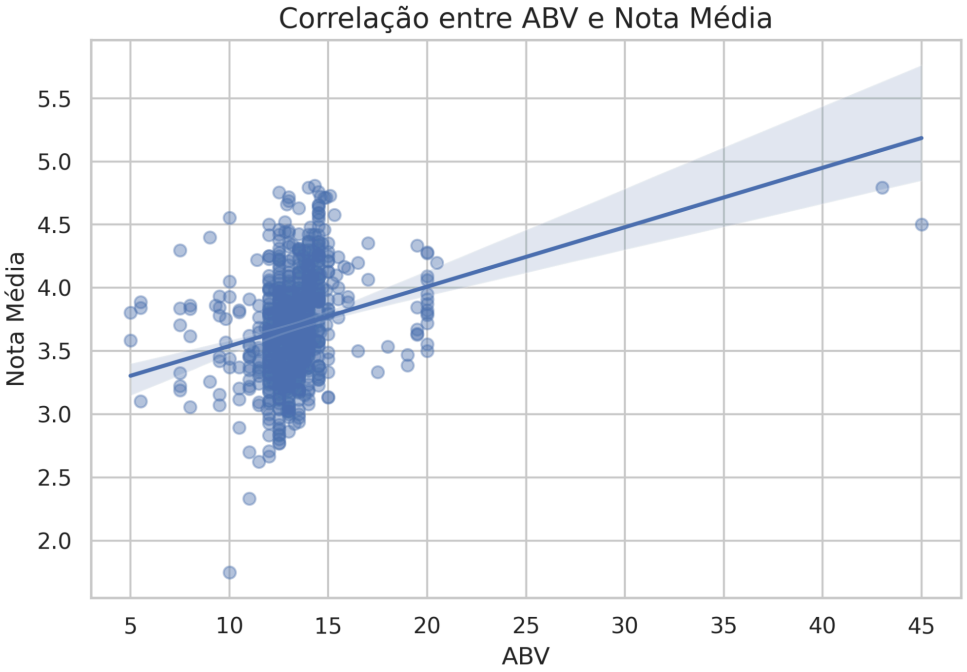

Figura 9. Gráfico de dispersão entre ABV e nota média — com *outliers* (ABV > 40%)

Fonte: Elaboração própria, 2025

A correlação de Pearson entre o ABV e a nota média dos vinhos gerou um coeficiente de r = 0,2690 e um *p-value* = $1{,}39e^{-15}$. Entretanto, observou-se que dois vinhos apresentavam um teor alcoólico muito acima da faixa da amostra, o que distorcia a distribuição dos dados. Por esse motivo, foram excluídos para gerar uma visualização mais representativa.

Após a exclusão dos dois *outliers*, gerou-se novamente o gráfico de dispersão (Figura 10).

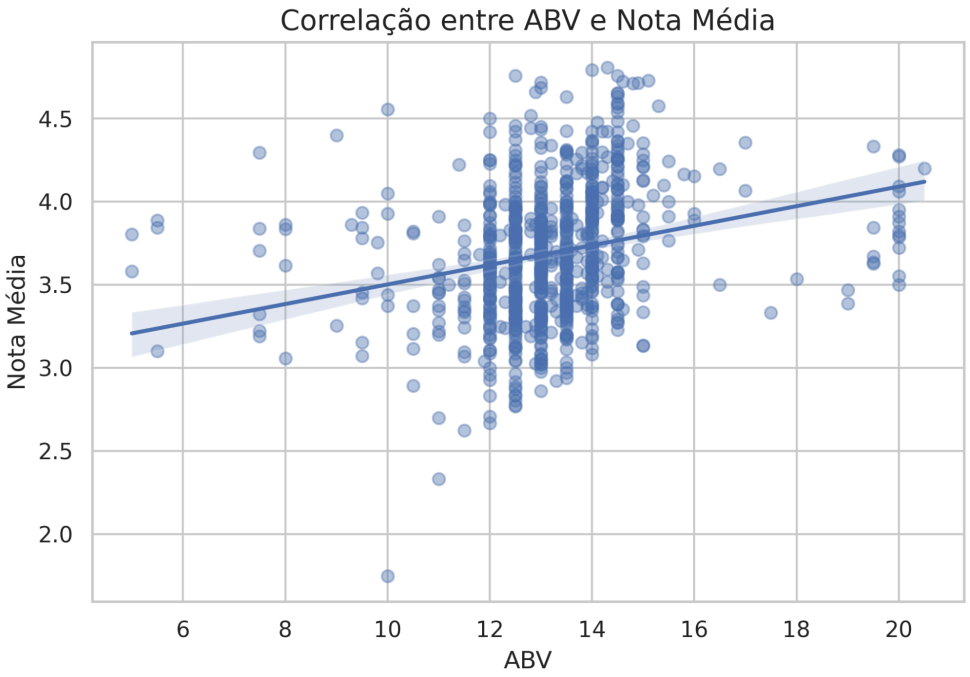

Figura 10. Gráfico de dispersão entre ABV e nota média — sem *outliers* (ABV < 40%)

Fonte: Elaboração própria, 2025

A correlação de Pearson entre o ABV e a nota média dos vinhos foi calculada, resultando em um coeficiente de r = 0.25 e um *p-value* = $2.45e^{-14}$.

5.1.2 Acidez/Corpo dos vinhos e nota média

Para as variáveis categóricas acidez (*acidity*) e corpo (*body_num*), foram calculadas as estatísticas descritivas das notas médias, incluindo número de vinhos (N), média, desvio padrão, teste *t-student* e p-value para cada categoria (Tabela 4, Figura 11 e Figura 12).

Tabela 4. Estatísticas descritivas por categoria e resultados do teste *t-student*

Variável	Categoria	N	Média	Desvio Padrão	Teste *t-student*	p-value
acidity	Baixa acidez 1	11	3.394	0,488	-2,0325	0,069509
	Média acidez 2	202	3.725	0,446	1,026	0,306127
	Alta acidez 3	638	3.688	0,383	-0,3319	0,740108
body_num	Encorpado 1	12	3.574	0,294	-1,4016	0,1886
	Leve encorpado 2	86	3.643	0,281	-1,6409	0,1045
	Médio corpo 3	294	3.622	0,4	-3,052	0,0025
	Muito encorpado 4	340	3.690	0,4	-0,1186	0,9056
	Muito pouco encorpado 5	119	3.924	0,409	6,1572	0,00000001

Fonte: Elaboração própria, 2025

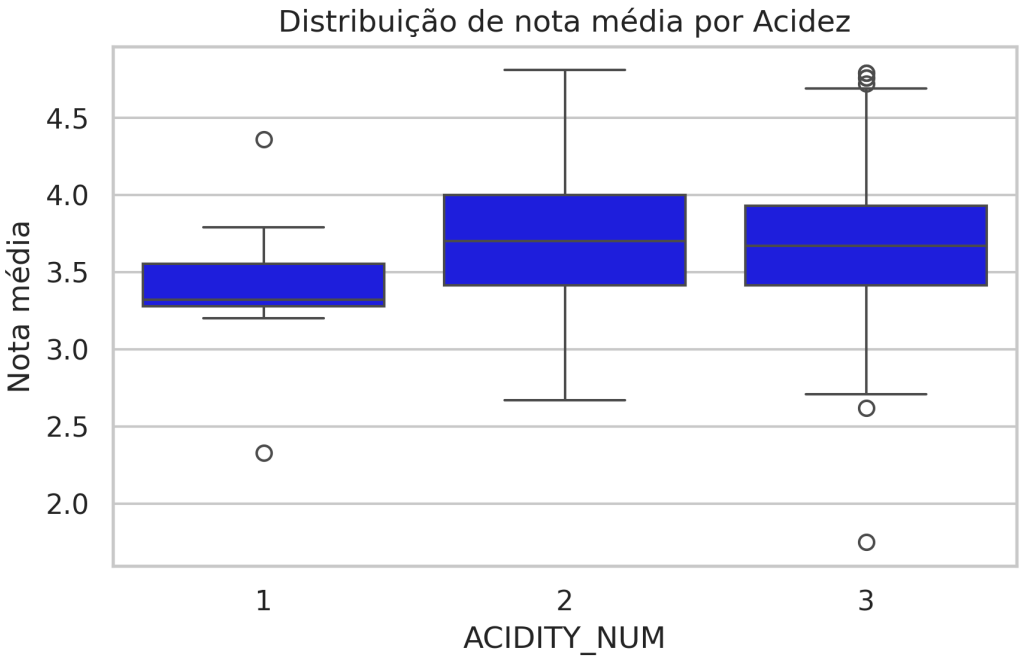

Figura 11. Gráfico comparativos das notas médias por categoria de acidez (acidity_num).

Fonte: Elaboração própria, 2025

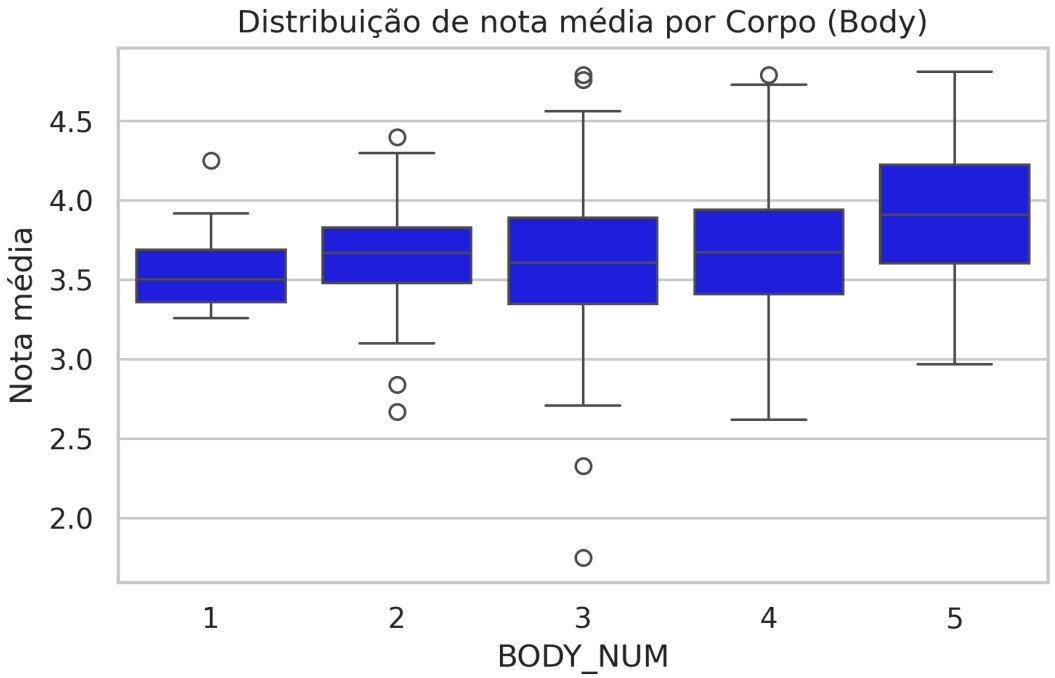

Figura 12. Gráfico comparativo das notas médias por categoria de corpo (*body_num)*.

Fonte: Elaboração própria, 2025

Ainda foram calculados r de Pearson e *p-value* para identificar se há correlação entre as categorias de cada característica e a nota média (os valores encontrados estão na Tabela 5).

Tabela 5. Correlação e p-value para acidez e corpo

Variável	r de Pearson	p-value
acidity	0,001	0,37
body	0,199	0,24

Fonte: Elaboração própria, 2025

5.2 Estatísticas descritivas por estilo

Para a variável categórica estilo (varietal e *blend*), foram calculadas as estatísticas descritivas das notas médias, incluindo o número de vinhos em cada grupo (N), a média das avaliações e o desvio padrão (Tabela 5, Figuras 13 e 14).

Tabela 5. Estatísticas descritivas por estilo (varietal e *blend*)

Estilo	N	Média	Desvio Padrão
Varietal	591	3.690	0,409
Blend	260	3.699	0,384

Fonte: Elaboração própria, 2025

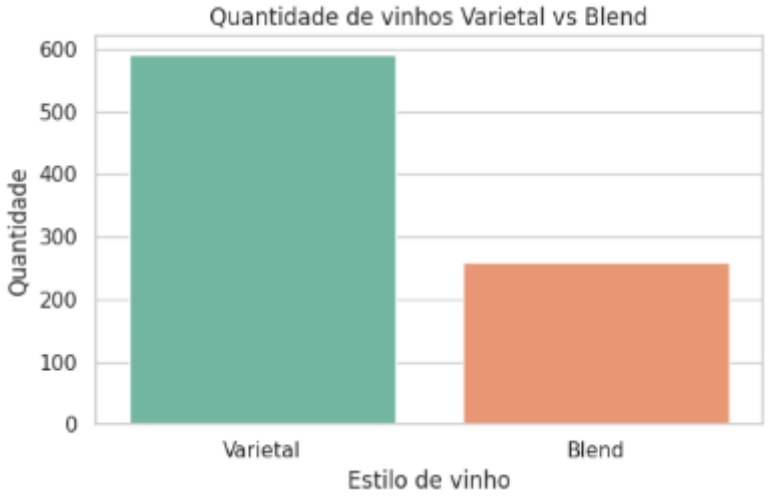

Figura 13. Gráfico de quantidade de vinhos varietal vs *blend*.

Fonte: Elaboração própria, 2025

As médias das notas atribuídas foram comparadas entre os dois grupos: Varietal apresentou nota média de 3,690, e *blend*, média de 3,699.

O teste *t-student* indicou uma estatística t = -0,2771 e um *p-value* = 0,7818.

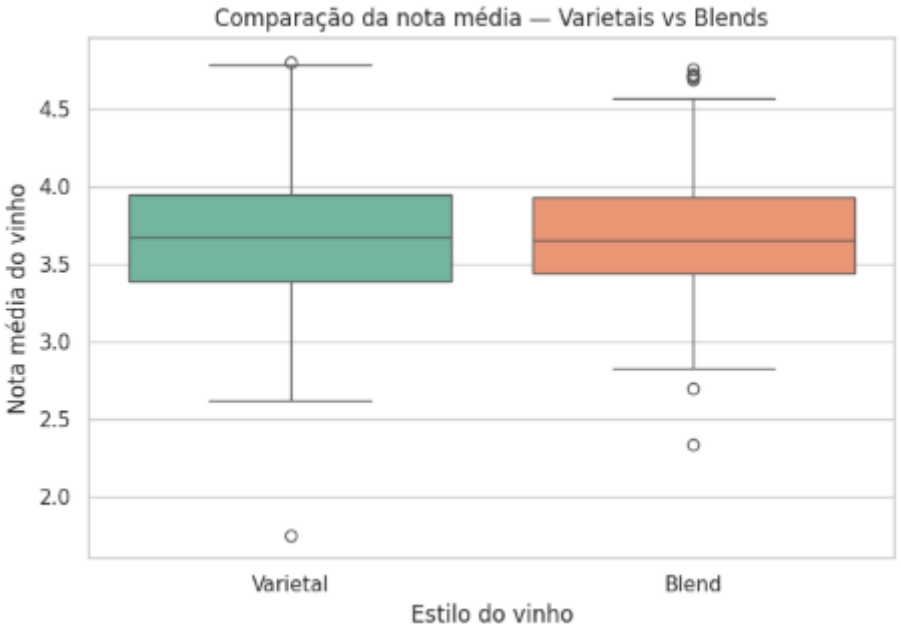

Figura 14. Gráfico comparativo das notas médias de varietal vs *blend*.

Fonte: Elaboração própria, 2025

6. Discussão

Resgatando os objetivos definidos na seção 1, buscou-se interpretar e discutir os resultados obtidos.

6.1 Vinhos com um maior teor alcoólico tendem a ter nota média maiores

Analisando os resultados obtidos ao correlacionar o teor alcoólico de cada vinho com a nota atribuída, percebemos que o coeficiente de Pearson (r=0,2690) e *p-value* (p<0,05) poderiam estar sendo influenciados por dois vinhos com teor alcoólico acima de 40% - *outliers* (Figura 9). Excluíram-se então estes vinhos das análises, com o objetivo de visualizar de forma mais homogênea os resultados e impedir que os índices avaliados fossem tendencionados por duas amostras muito destoantes. Após essa exclusão, gerou-se uma nova figura (Figura 10), com coeficiente de Pearson = 0,25 e *p-value* <0,05, indicando uma correlação fraca, mas significativa. Desse modo, a presença dos *outliers* não alterou a interpretação do resultado.

Ainda que a correlação tenha sido estatisticamente significativa e indique que quanto maior o teor alcoólico, maior a nota média do vinho, observamos que entre o teor alcoólico de 10% e 15% temos uma distribuição similar de notas com aumento

gradual da menor nota média atribuída a um vinho conforme aumenta-se o teor alcoólico. Desse modo, observa-se que podem existir vinhos com um teor alcoólico menor que seja mais bem avaliado quando comparado com um vinho com teor alcoólico maior, entretanto, o vinho com a pior nota entre os com maior teor alcoólico tende a ter uma nota melhor que o pior vinho com teor alcoólico menor.

6.2 Características físico-químicas não influenciam a avaliação dos usuários

Avaliamos ainda se as características de corpo e acidez influenciavam a nota média atribuída pelos usuários aos vinhos. Como a classificação dessas características são níveis, foi utilizado teste *t-student* e p-value de cada nível das duas características, para verificar se as notas médias encontradas refletiam dados homogêneos, e dessa forma, se saberia se as notas médias eram representativas do grupo ou se eram o resultado de vinhos muito mal avaliados e muito bem avaliados num mesmo grupo.

6.2.1 Acidez

O teste *t-student* para acidez mostrou que a diferença entre qualquer nota atribuída aos três níveis de acidez (baixo, médio, alto) em relação a média não é estatisticamente significativa, desse modo, os valores de cada grupo são homogêneos.

O valor de r de Pearson mostrou que não há correlação entre as variáveis nota média e acidez ($r<0,2$ e *p-value*$>0,05$). Desse modo, ainda que os vinhos com maior acidez tenham uma nota maior, não há diferença significativa em relação a vinhos com média e baixa acidez. Com isso, os resultados apontam que a acidez do vinho não impactou na avaliação dos usuários e, portanto, não é uma característica determinante para avaliações melhores/piores.

Cabe ressaltar que o N de vinhos analisados para cada grupo é muito diferente (Tabela 4) e que esse pode ser um fator determinante para os resultados obtidos.

6.2.2 Corpo

O teste *t-student* para corpo apontou que a diferença entre qualquer nota atribuída aos níveis 1-Encorpado; 2-Levemente Encorpado; 4-Muito Encorpado; em relação a média não é estatisticamente significativa, desse modo, os valores que compõem

cada grupo são homogêneos. Entretanto, os níveis 3-Médio Corpo e 5-Muito Pouco Corpo tiveram avaliações heterogêneas. Desse modo, sabe-se que nesses grupos as notas estão sendo puxadas por alguns vinhos específicos e não pela característica do corpo em si. Os valores de r de Pearson (0,199) e *p-value* > 0,05 mostram que não há correlação entre a nota e o corpo.

Com isso, os resultados apontam que o corpo do vinho não foi responsável por determinar avaliações melhores/piores para os usuários.

6.3 A variação de vinhos varietais e *blend* não impacta a nota média dada pelos consumidores

Conforme apontado pelos resultados, os vinhos classificados como varietal e *blend* possuem notas médias muito similares (varietal = 3,690; *blend* = 3,699). Para confirmar esses dados, foi realizado teste *t-student*, obtendo um valor de t = -0,2771 e *p-value* = 0,7818, indicando que não há diferença estatisticamente significativa entre as notas médias dos dois tipos de vinho.

Esses resultados nos mostram que os usuários do X-wine atribuíram notas similares a vinhos compostos por uma única uva ou a vinhos compostos por mais de um tipo de uva, avaliando igualmente ambos os estilos.

Apesar da similaridade, o número de avaliações para vinhos varietais é bem maior (N = 591) quando comparado aos vinhos *blend* (N = 260), não permitindo a extrapolação dos resultados para concluir que ambos os vinhos são consumidos de maneira proporcional. Também não se pode concluir que os vinhos varietais são mais consumidos que os vinhos *blend*, uma vez que o número de avaliações não corresponde necessariamente ao número de consumidores e à quantidade consumida.

Considerando os resultados acima e o contexto de produção de vinhos, indicam que há espaço no mercado para ambos os estilos.

6.4 As características físico-químicas isoladamente não explicam completamente as notas

Ainda que os resultados apontem correlações estatisticamente significativas entre o teor alcoólico e as notas médias dos vinhos, essa correlação foi classificada como fraca. Assim, uma única característica pode não explicar completamente as notas atribuídas pelos usuários, pois deve-se considerar que um vinho *blend* com acidez

baixa, teor alcoólico alto e encorpado vai compor um sabor diferente de outro vinho *blend* com acidez alta, teor alcoólico alto e pouco encorpado. Desse modo, a soma das características físico-químicas pode explicar melhor as notas atribuídas pelos usuários do que elas isoladamente.

Ainda, sabe-se que as características organolépticas do vinho são compostas por mais variáveis além das compreendidas em características físico-químicas, como a safra das uvas que o compõem, o tempo de maturação, a região em que foi produzido, dentre outras. Desse modo, existem muitas características que foram desconsideradas nesta análise de correlação simples.

7. Considerações e Perspectivas

Essa análise preliminar foi realizada utilizando a versão *Slim* do dataset, que reúne um volume de dados adequado para as análises estatísticas propostas nesta etapa do estudo. A escolha por essa versão decorre de sua maior praticidade para exploração inicial, permitindo uma avaliação clara das relações entre as variáveis e garantindo agilidade no processamento e na verificação dos resultados.

A planilha ratings da versão *Slim* contém 150 mil avaliações, quantidade suficiente para identificar padrões relevantes no comportamento dos consumidores. Já a planilha wines possui aproximadamente mil registros, abrangendo todos os atributos essenciais (*grapes, abv, acidity, body*, entre outros), o que possibilitou uma análise completa dos dados disponíveis.

Por outro lado, a planilha *wines* da versão *Slim* possui aproximadamente mil registros, quantidade totalmente suportável pelos ambientes de manipulação utilizados (Python), sem risco de truncamento. Assim, sua versão *Slim* foi empregada normalmente na pesquisa, garantindo acesso à totalidade dos atributos essenciais (*grapes*, *abv*, *acidity*, *body*, entre outros).

Desse modo, o conjunto de dados completo pode fornecer informações ainda mais robustas e identificar correlações ainda mais fortes do que as encontradas até aqui. Assim, pretende-se ampliar essas análises, considerando o *dataset* completo, o conjunto das características físico-químicas e outras variáveis que possam contribuir para o entendimento das preferências dos consumidores.

Entretanto, para considerar mais de uma característica físico-química com a nota da avaliação do usuário, fazem-se necessários testes estatísticos multidimensionais, aumentando a complexidade das relações avaliadas e dos cálculos realizados.

Além disso, ainda que relevante a análise e compreensão das relações das múltiplas características físico-químicas com a avaliação dada pelos consumidores, existem variáveis que não são possíveis de quantificar, como as questões culturais de cada pessoa, tradições familiares e memórias ligadas ao vinho que influenciam diretamente as preferências dos consumidores.

Ainda, os vinhos possuem uma complexidade natural e uma combinação de atributos sensoriais únicos, que fazem com que vinhos com as mesmas características físico-químicas tenham sabores completamente diferentes e forneçam uma experiência singular a cada garrafa até mesmo do mesmo vinho.

O conhecimento de cada avaliador sobre vinhos também contribui para sua percepção da qualidade dos vinhos, tornando vinhos de entrada com avaliações mais pautadas nos gostos pessoais que vinhos mais complexos, destinados a apreciadores com paladar mais apurado e que atribuem valor também ao processo de produção.

Observando os dados, é possível ver também que a quantidade de avaliações é relevante para essas análises porque uma única avaliação ruim tem impactos diferentes num vinho avaliado por 10 pessoas e noutro avaliado por 100 pessoas, podendo não representar a real opinião do público e tendenciar os dados, levando a uma interpretação equivocada.

Na prática, considerando a análise de dados para orientação da tomada de decisão, deve-se lembrar que os dados explorados neste trabalho são oriundos dos usuários do X-wine, que é composto por usuários de vários países do mundo e limitado a todos que conhecem a ferramenta.

E ainda que para o aprendizado de ciência de dados, este trabalho mostrou a possibilidade de obter informações importantes no *dataset* X-Wines, utilizando fatores comuns à produção de vinho e entendendo seu papel na percepção de qualidade pelos usuários, direcionando análises mais avançadas nos estudos dessas correlações e apontando o potencial da utilização de ciência de dados na área de vitivinícolas.

Os resultados deste trabalho fornecem informações importantes ao setor vitivinícola ao indicar que, embora algumas características físico-químicas apresentem

relação estatística com a avaliação dos consumidores, nenhuma delas, isoladamente, determina a percepção de qualidade. Assim, produtores podem usar esses resultados como apoio para posicionamento de produtos, desenvolvimento de novos rótulos e compreensão de que fatores como acidez, corpo ou tipo (varietal/blend) não são, por si só, determinantes das notas atribuídas pelos usuários.

Por fim, além dos resultados estatísticos obtidos, o próprio algoritmo desenvolvido representa um produto relevante deste trabalho. A estrutura criada, capaz de tratar dados, gerar métricas, realizar testes estatísticos e produzir visualizações de forma automática, pode ser aplicada a qualquer *dataset*, independentemente da área. Assim, a *pipeline* desenvolvida se configura como uma solução reutilizável e extensível, útil tanto para análises vitivinícolas quanto para outros setores que dependem de métodos quantitativos e que os utilizam de apoio à tomada de decisão.

8.Referências

de AZAMBUJA, A.; MORAIS, R.; FILIPE, J. *X-Wines: A dataset for wine recommendation and analysis*. **Big Data and Cognitive Computing** 7, no. 1: 20. 2023 https://doi.org/10.3390/bdcc7010020

BISONG, E. Building Machine Learning and Deep Learning Models on Google Cloud Platform. **Berkeley: Apress**, 2019. DOI: 10.1007/978-1-4842-4470-8.

CHAUDHURI, S.; DAYAL, U.; NARASAYYA, V. An overview of business intelligence technology. **Communications of the ACM**, v. 54, n. 8, p. 88–98, 2011. DOI: 10.1145/1978542.1978562.

RAUTENBERG, S.; CARMO, P. R. V. do. Big data e ciência de dados: complementaridade conceitual no processo de tomada de decisão. Brazilian **Journal of Information Science: Research Trends**, v. 13, n. 1, p. 56–70, 2019. DOI: 10.36311/1981-1640.2019.v13n1.06.p56.

LI, J. et al. Big data in tourism research: **A literature review. Tourism Management,** 2018. Disponível em:

https://linkinghub.elsevier.com/retrieve/pii/S0261517718300591. Acesso em: nov. 2025.

FEW, S. Now You See It: Simple Visualization Techniques for Quantitative Analysis. **Oakland: Analytics Press**, 2009.

FIGUEIREDO FILHO, R. F.; PARANHOS, D. B.; SANTOS, M. L. W. D. Desvendando os mistérios do coeficiente de correlação de Pearson: o retorno. ***Leviathan (São Paulo)***, 2014. DOI: 10.11606/issn.2237-4485.lev.2014.132346.

GKIKAS, D. C. et al. Machine learning methods for wine quality prediction. **Informatics**, v. 10, n. 3, 2023. DOI: 10.3390/informatics10030063.

GITHUB. GitHub — plataforma de hospedagem de código-fonte e controle de versão. Disponível em: https://github.com . Acesso em: nov. out 2025.

GITLAB. GitLab — plataforma DevOps de hospedagem de repositórios Git e integração contínua. Disponível em: https://gitlab.com. Acesso em: out. 2025.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. **Burlington: Morgan Kaufmann,** 2011. eBook ISBN: 9780123814807.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. 2. ed. **New York: Springer,** 2009.

HE, J.; JIANG, Y.; GU, G. Wine quality classification using machine learning algorithms. **International Journal of Computer Applications Technology and Research**, v. 11, n. 6, 2023. DOI: 10.7753/IJCATR1106.1010.

HERLOCKER, J. et al. Evaluating collaborative filtering recommender systems. **ACM Transactions on Information Systems**, v. 22, n. 1, p. 5–53, 2004. DOI: 10.1145/963770.963772.

IWENDI, C. et al. COVID-19 patient health prediction using machine learning. **Frontiers in Public Health**, v. 8, 2020. DOI: 10.3389/fpubh.2020.00357.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. **Computer**, v. 42, n. 8, p. 30–37, 2009. DOI: 10.1109/MC.2009.263.

LESSCHAEVE, I. Sensory evaluation of wine and commercial realities: review of current practices and perspectives. **Food Quality and Preference**, v. 18, n. 1, p. 1–13, 2007. DOI: 10.5344/ajev.2007.58.2.25.

MACNEIL, K. The Wine Bible. 2. ed. **New York: Workman Publishing**, 2015.

PYTHON. Website oficial. Disponível em: https://www.python.org/. Acesso em: out. 2025.

MONTGOMERY, D.; RUNGER, G. Applied Statistics and Probability for Engineers. 7. ed. **Hoboken: Wiley,** 2018.

SILVEIRA, A. B.; MONTICELLI, J. M.; BARBOSA, F. S. Wine consumer profile in wine regions of Brazil. **Consumer Behavior Review**, v. 8, n. 1, e-257580, 2024. DOI: 10.51359/2526-7884.2024.257580.

PUCKETTE, M.; HAMMACK, J. Wine Folly: The Essential Guide to Wine. **New York: Avery**, 2015.

REICH, N.; MAGUIRE, M. The role of acidity in wine preference and consumer segmentation. **Journal of Sensory Studies**, 2018. DOI: 10.32628/CSEIT217628.

SHAO, Y.; LI, X.; BIAN, J. Hybrid recommendation model combining collaborative and content-based approaches. **Expert Systems with Applications,** v. 183, 2021. DOI: 10.1016/j.eswa.2021.115164.

WICKHAM, H.; GROLEMUND, G. R for Data Science. **Sebastopol: O'Reilly**, 2017.

ZHANG, Z.; LU, C.; JIN, X. A survey on hybrid recommender systems. **Knowledge-Based Systems**, v. 215, 2021.