



Análise Exploratória de Dados de Vinhos





Dedicatória

Dedico este trabalho, primeiramente, ao professor, cuja dedicação e excelência na metodologia de ensino foram fundamentais para minha trajetória. Seus ensinamentos não apenas facilitaram a compreensão dos conteúdos, mas também serviram de inspiração, tornando a jornada de aprendizado mais leve, eficiente e enriquecedora.

Aos meus pais, que estiveram ao meu lado, oferecendo amor, suporte e compreensão, não só durante este curso, mas em toda minha vida. Sem o apoio e incentivo de vocês, certamente este caminho teria sido muito mais difícil.

Aos meus amigos, que foram pilares importantes nessa caminhada, em especial ao Diogo. Mesmo enfrentando desafios pessoais, esteve presente, oferecendo apoio, motivação e companhia, mostrando que a verdadeira amizade se fortalece nos momentos mais difíceis.

A todos vocês, minha mais sincera gratidão.





Citação

**“Sem dados você é apenas mais uma pessoa com
uma opinião”**

Deming, William Edwards





Sumário

1. Introdução	4
2. Metodologia.....	5
3. Análise Exploratória dos Dados	6
3.1 Panorama Geral dos dados	6
3.2 Qualidade e Avaliação dos Vinhos	9
3.3 Preço e Economia	12
3.4 Custo-benefício e Eficiência.....	14
3.5 Variedades e Características dos Vinhos.....	15
3.6 Influência de Degustadores e Descrições	16
3.7 Tendências Temporais e Sazonais	17
4. Considerações Finais.....	18
4.1 Limitações do Estudo.....	19
5. Referências	19





1. Introdução

Este relatório apresenta uma análise exploratória dos dados do dataframe de avaliações de vinhos. A análise visa identificar padrões de qualidade, preço, preferências regionais e aspectos econômicos relacionados aos vinhos, além de compreender a influência dos degustadores e tendências ao longo do tempo.

Objetivos:

- Analisar a distribuição dos vinhos por país, variedade e popularidade.
- Avaliar a relação entre pontuação, preço e características regionais.
- Explorar indicadores de custo-benefício.
- Identificar tendências sazonais e temporais.
- Compreender a influência dos degustadores nas avaliações.





2. Metodologia

Foram utilizadas bibliotecas de análise de dados e visualização como Pandas, Seaborn, Matplotlib e numpy. O dataframe foi tratado para corrigir inconsistências como valores nulos e ausência de identificadores.

Em primeiro momento, a coluna 1, foi renomeada para ID, para melhor identificação de cada uma das avaliações.

Posteriormente, país, província e região tiveram seus dados em branco substituídos por “Desconhecido”, para assim, podermos considerar os valores numéricos de preços, notas, e identificar se estes dados sem nomenclaturas possuem um real impacto nas informações.

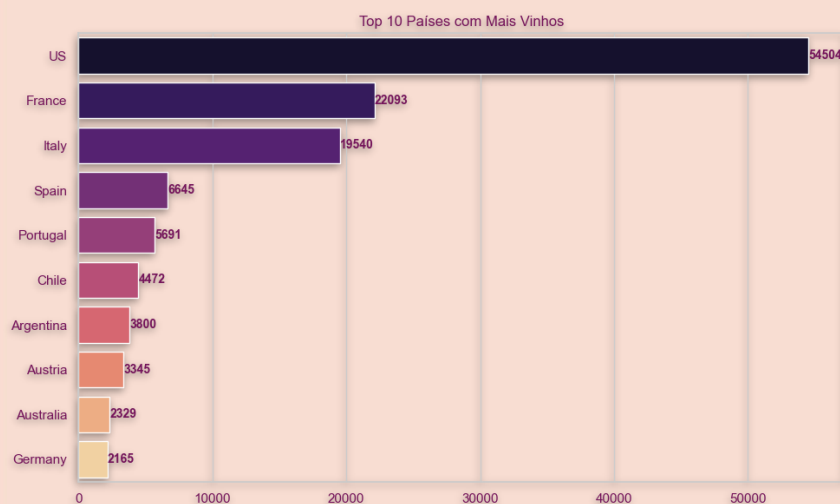
Por último, foi adicionada a nomenclatura “Anônimo” nos dados em branco das colunas “TasterName” e “TasterTwitterHandle”.



3. Análise Exploratória dos Dados

3.1 Panorama Geral dos dados

- O país com maior quantidade de vinhos no dataset é os **Estados Unidos**, seguido por **Itália** e **França**. Isso evidencia uma forte predominância do mercado norte-americano no banco de dados.



Análise:

Observa-se uma concentração significativa nos países com tradição vinícola. Outros países aparecem em menor escala, indicando menor representatividade no dataset.

Gráfico 1: Barplot da quantidade de vinhos por país

Como a quantidade de vinhos varia entre países?

Análise:

A visualização em treemap evidencia proporcionalmente a dominância dos EUA e países europeus tradicionais.

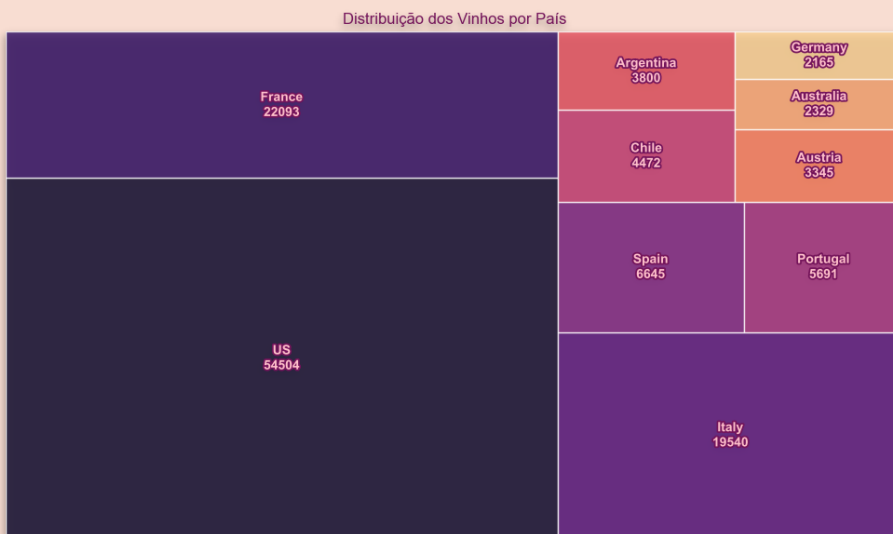


Gráfico 2: Treemap dos vinhos por país



Variedade de vinho mais comum:

A variedade 'Pinot Noir' é a mais representada, seguida de 'Chardonnay' e 'Cabernet Sauvignon', refletindo uma preferência global ou maior foco editorial nesses tipos.

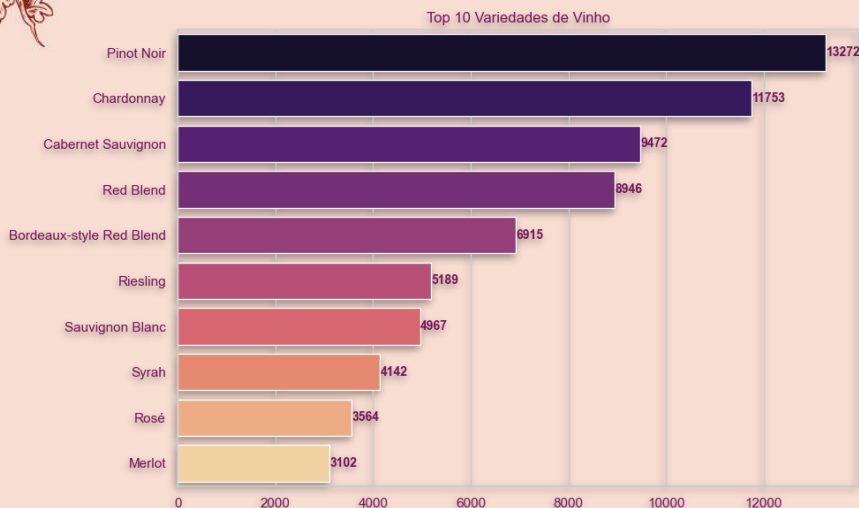


Gráfico 3: Barplot das 10 variedades mais comuns

Vinhos mais populares (número de avaliações):

Identificamos os vinhos com maior número de registros de avaliações, evidenciando aqueles que possuem maior impacto ou divulgação na base de dados.

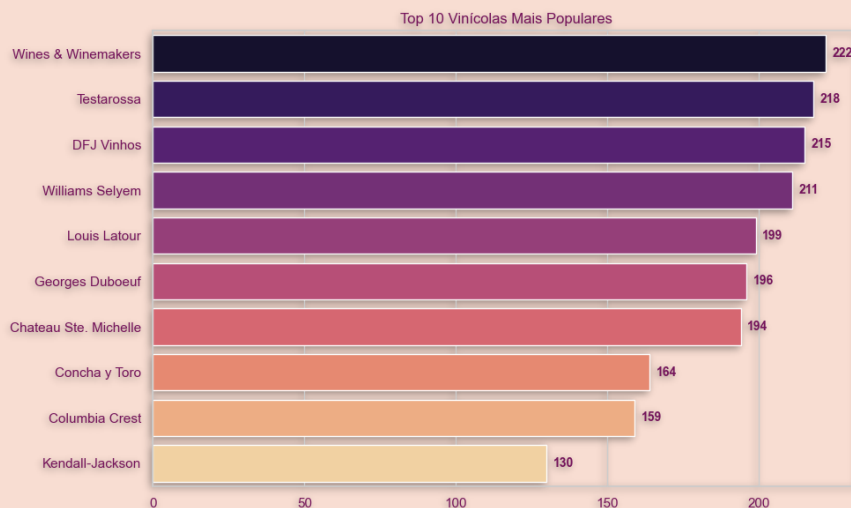


Gráfico 4: Barplot dos vinhos mais populares





Gráfico 5: WordCloud das palavras mais usadas nas descrições

Palavras como “*Wine*”, “*finish*”, “*palate*” e “*nose*” muito recorrentes, porém, muitas vezes estas podem ser usadas como ponto inicial para a dissertação do avaliador sobre o vinho provado. Assim como “*palate*” outras palavras freqüentemente usadas, que podem indicar características sensoriais que estão destacadas nas avaliações são “*aroma*”, “*flavor*”, “*blend*” e “*sweet*”.



3.2 Qualidade e Avaliação dos Vinhos

Média de pontos por país:

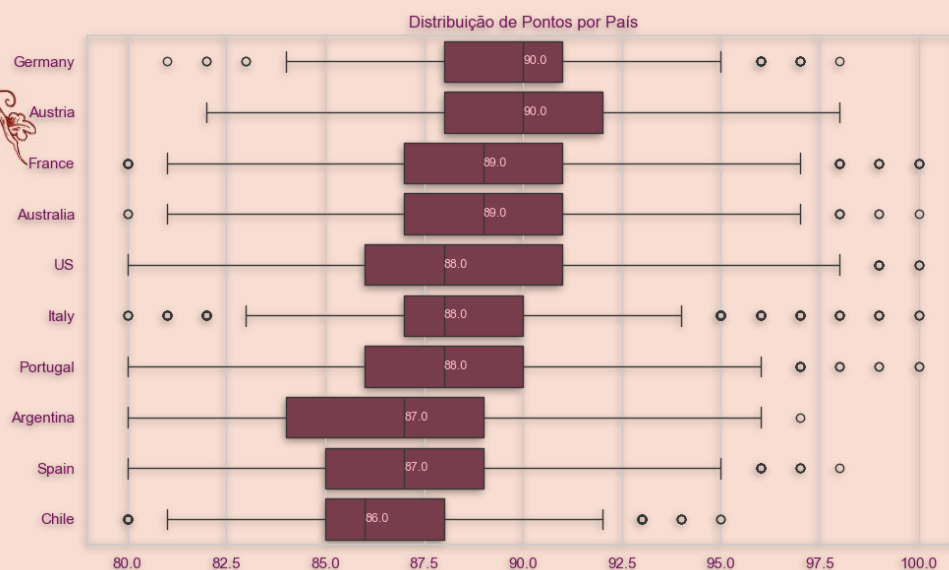


Gráfico 6: Boxplot da distribuição de pontos por país

Os 3 países com as maiores pontuações de medianas são **França, Alemanha e Áustria**, sugerindo uma percepção de qualidade superior dos vinhos europeus, sendo que destes 3, apenas a Áustria não é rica de outliers nos dados.

Regiões com maiores classificações:

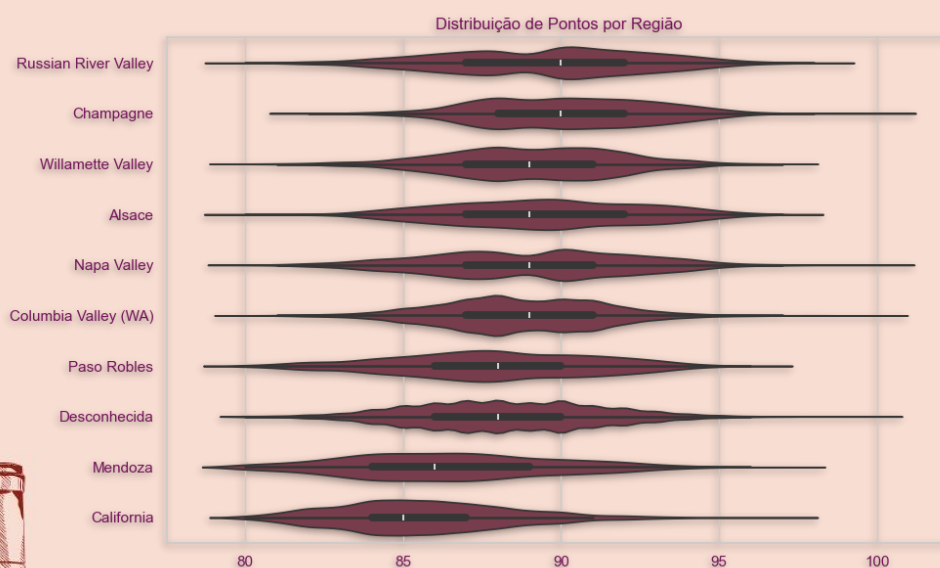


Gráfico 7: Violinplot da pontuação por região

É possível notar que uma grande distribuição de pontos foi registrada em região “desconhecida”, ou seja, apesar do avaliador ter preenchido as demais informações, deixou uma “lacuna” em aberto, sobre a região do vinho degustado. Também é possível notar destaque nas medianas das regiões como Champagne (França) e Russian River Valley (Eua).

Distribuição dos pontos:

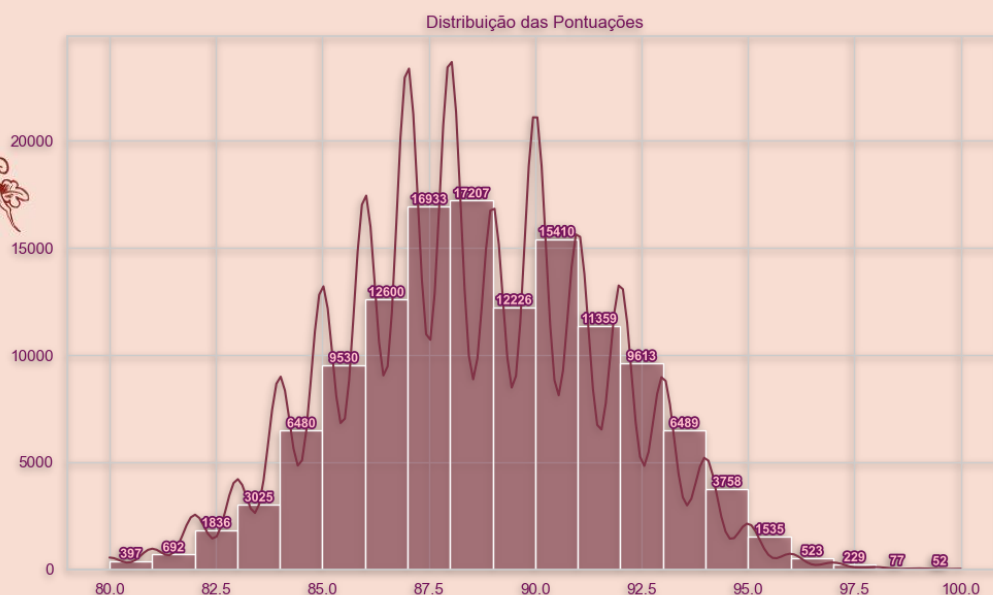


Gráfico 8: Histograma da distribuição de pontos

A maior parte dos vinhos possui pontuação entre **87 e 88 pontos**, possuindo um pico também por volta dos **90 pontos**, indicando uma tendência de avaliações concentradas em faixas intermediárias-altas.

Tendência ao longo dos anos:

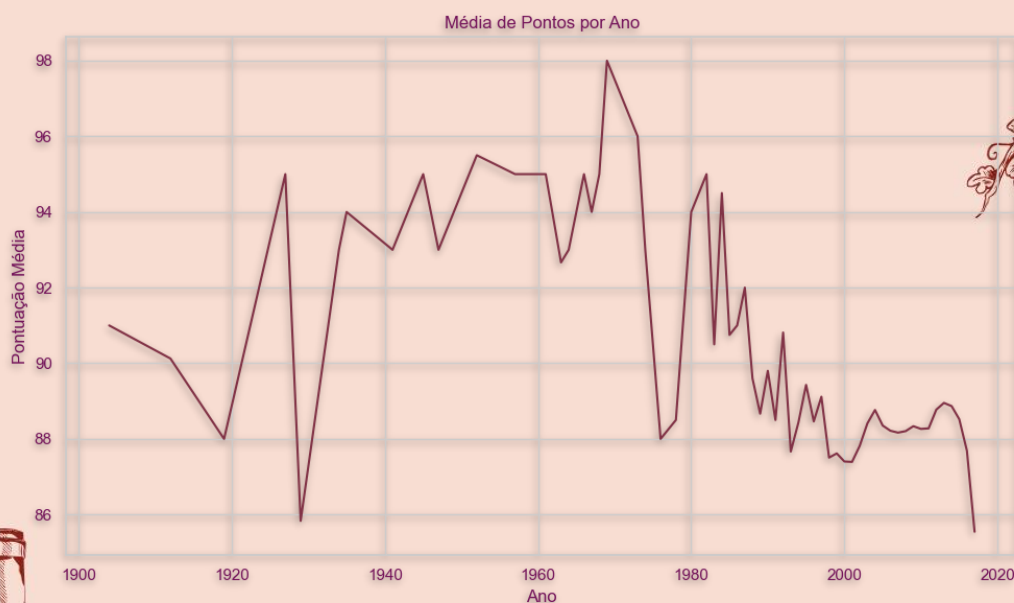


Gráfico 9: Lineplot da média de pontos por ano

Há uma leve tendência de aumento nas pontuações, possivelmente relacionada a melhorias nas técnicas de produção ou critérios mais generosos nas avaliações recentes, porém, ao longo dos anos, também há picos de quedas.

Influência dos degustadores:

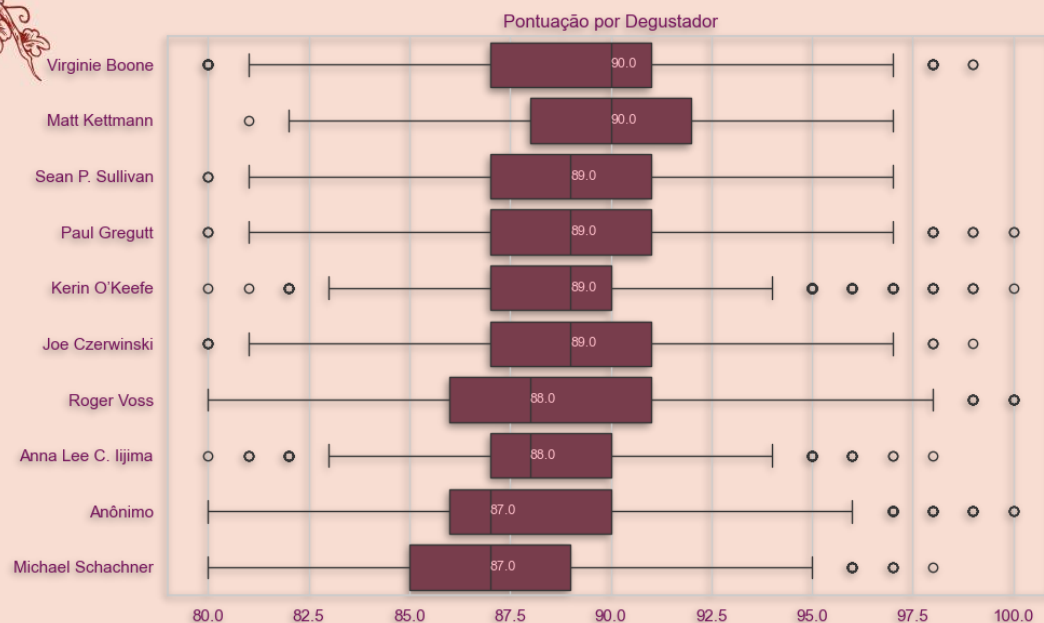


Gráfico 10: Boxplot da pontuação por taster_name

A análise mostra que alguns degustadores possuem médias consistentemente mais altas ou mais baixas, sugerindo subjetividade nas avaliações, porém quase todos, possuem outliers que podem mostrar picos em suas avaliações, sejam elas ligadas ao tipo de vinho provado ou não.



3.3 Preço e Economia

Variação de preços por país:

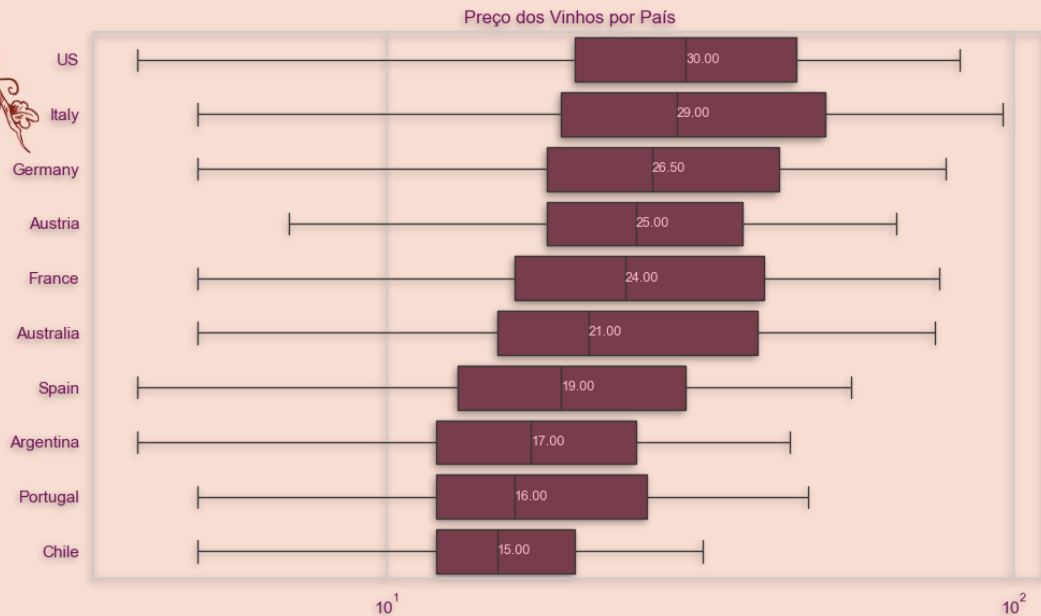


Gráfico 11:Boxplot do preço dos vinhos por país

Destaca-se que países como **França** e **EUA** apresentam tanto os vinhos mais caros considerando as medianas. Considerando os outliers não visuais no gráfico, existe uma ampla variedade de faixas de preços, onde **França** segue na liderança, porém, dessa vez seguida pela **Itália**.

Distribuição dos preços:

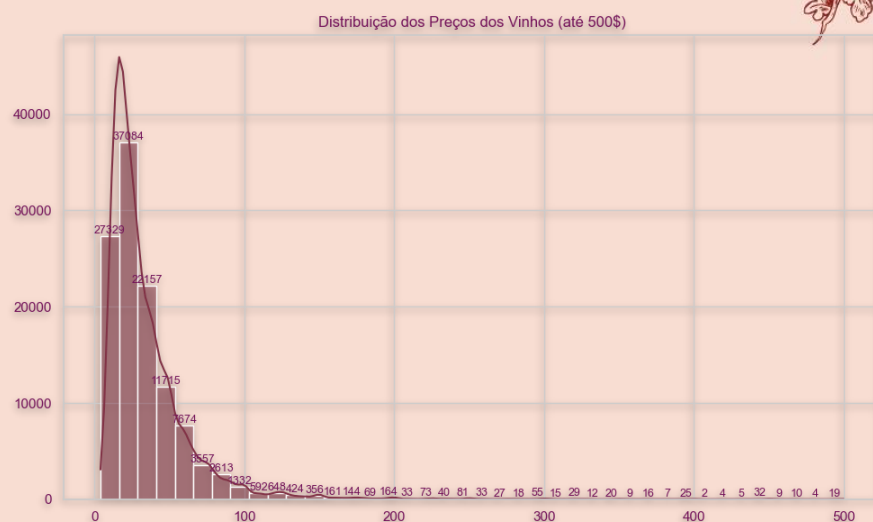


Gráfico 12:Histograma dos preços dos vinhos

Existe uma grande quantidade de vinhos com preços entre 10 e 80 **dólares**, sendo a faixa campeã, a próxima dos 20 dólares, mas alguns outliers não visuais no gráfico ultrapassam os **500 dólares** e chegam até **3000 dólares**.



Correlação preço x pontuação:

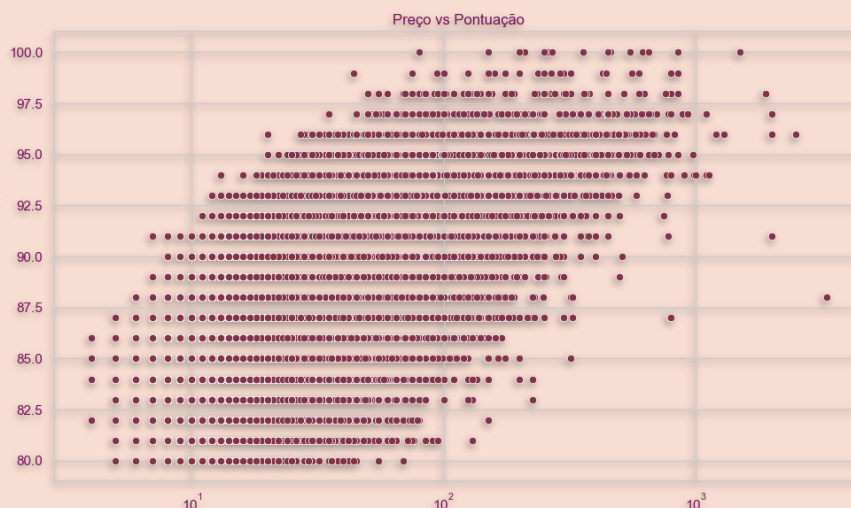


Gráfico 13: Scatterplot preço vs. Pontos

Há uma **correlação positiva moderada**, indicando que vinhos mais caros tendem a ter pontuações mais altas, mas não de forma determinística.

Análise temporal de preços:

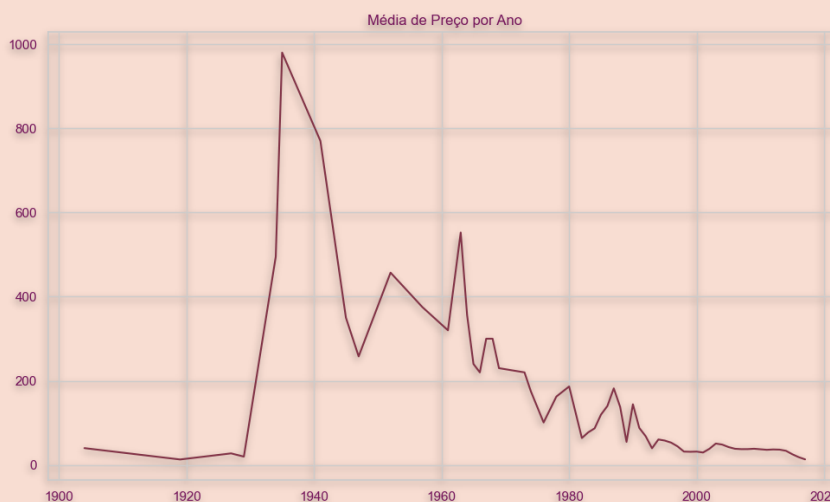


Gráfico 14: Lineplot da média de preço ao longo dos anos

Percebe-se um leve aumento nos preços médios, acompanhando tendências de mercado, assim também como uma diminuição periódica.

10 vinhos mais caros:

Nome do Vinho	Preço (USD)	Pontuação
Château les Ormes Sorbet 2013 Médoc	3300.0	88
Domaine du Comte Liger-Belair 2010 La Romanée	2500.0	96
Château Pétrus 2014 Pomerol	2500.0	96
Blair 2013 Roger Rose Vineyard Chardonnay (Arroyo Seco)	2013.0	91
Domaine du Comte Liger-Belair 2005 La Romanée	2000.0	96
Château Pétrus 2011 Pomerol	2000.0	97
Château Margaux 2009 Margaux	1900.0	98
Château Lafite Rothschild 2010 Pauillac	1500.0	100
Château Cheval Blanc 2010 Saint-Émilion	1500.0	100
Château Mouton Rothschild 2009 Pauillac	1300.0	96





3.4 Custo-benefício e Eficiência

Vinhos com melhor custo-benefício:

Os vinhos destacados são aqueles com alta pontuação e preço acessível. Sendo possível notar uma grande concentração nos valores entre 10 dólares e 100 dólares.

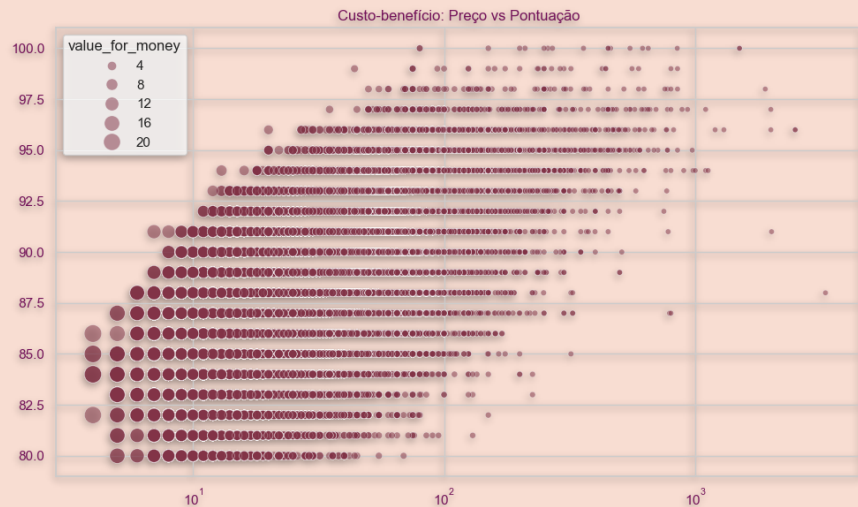


Gráfico 15: Scatterplot preço vs. pontuação, com tamanho representando popularidade

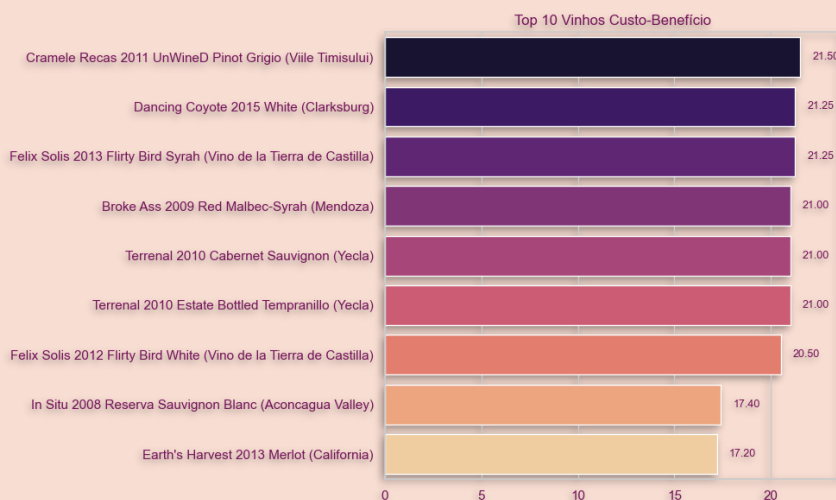
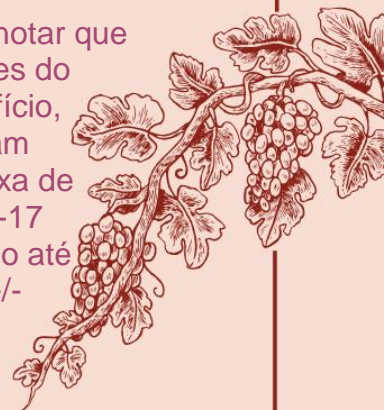


Gráfico 16: Barplot dos top 10 vinhos custo-benefício

É possível notar que os campeões do custo-benefício, se encontram dentro a faixa de preço de +/-17 dólares, indo até a faixa de +/-23 dólares.



Dessa forma, entende-se que, para aqueles que buscam vinhos de melhor qualidade com bom custo-benefício, há maior probabilidade de encontrá-los na faixa de preço entre US\$17 e US\$23.

Essa conclusão baseia-se na quantidade significativa de vinhos com altas pontuações nesse intervalo, quando comparada à de precificações mais elevadas, indicando que vinhos bem avaliados não estão necessariamente entre os mais caros.



3.5 Variedades e Características dos Vinhos

Pontuação por variedade:

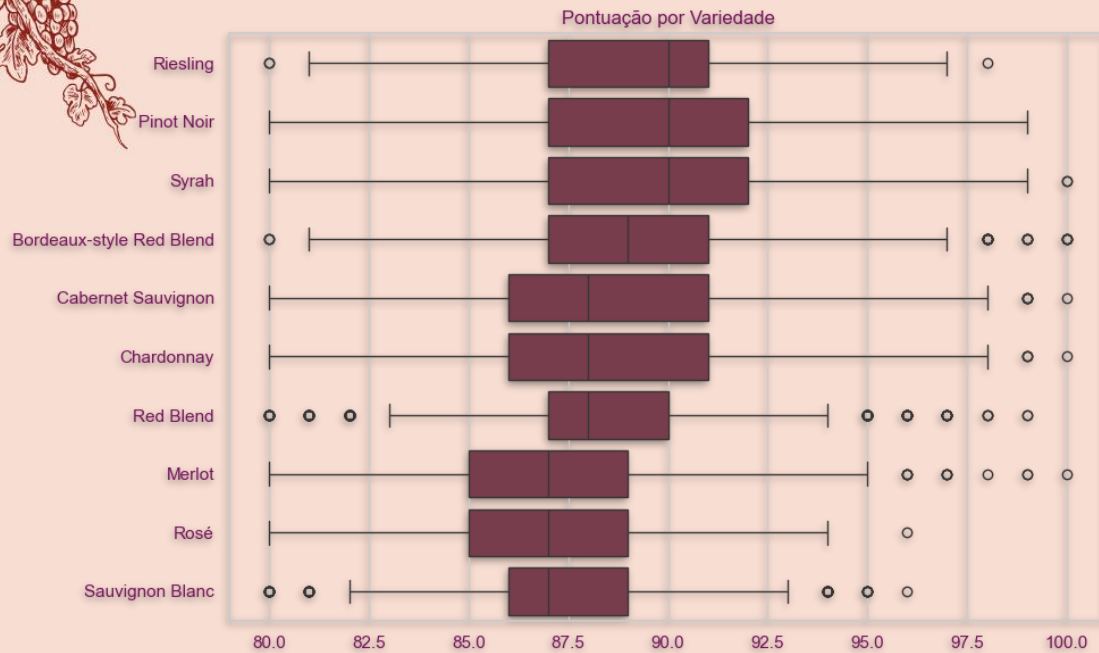
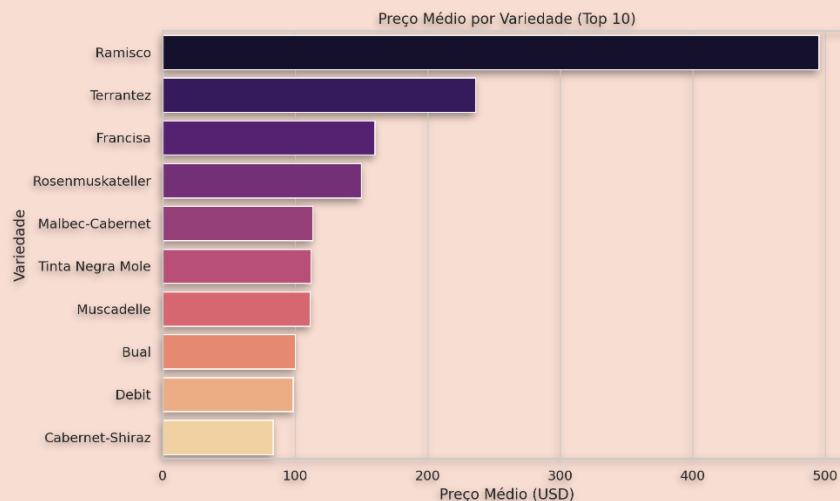


Gráfico 17: Boxplot das pontuações por variedade de uva

Variedades como 'Riesling', 'Pinot Noir' e 'Syrah' se destacam por medianas mais altas, porém todas com uma menor distância entre a mediana e o terceiro quartil, o que mostra que a distribuição dos dados está **assimétrica**, mais especificamente com uma **cauda mais longa para a esquerda**.



Aqui é possível notar que, o preço médio é diretamente relacionado ao tipo de vinho, sendo os "Ramisco" e "Terrantez" comumente mais caro do que um "Debit", por exemplo.

[illegible]

3.6 Influência de Degustadores e Descrições

Análise de textos:

WordCloud - Vinhos com Pontuação ≥ 95

Gráfico 18: WordCloud de descrições dos vinhos com mais de 95 pontos

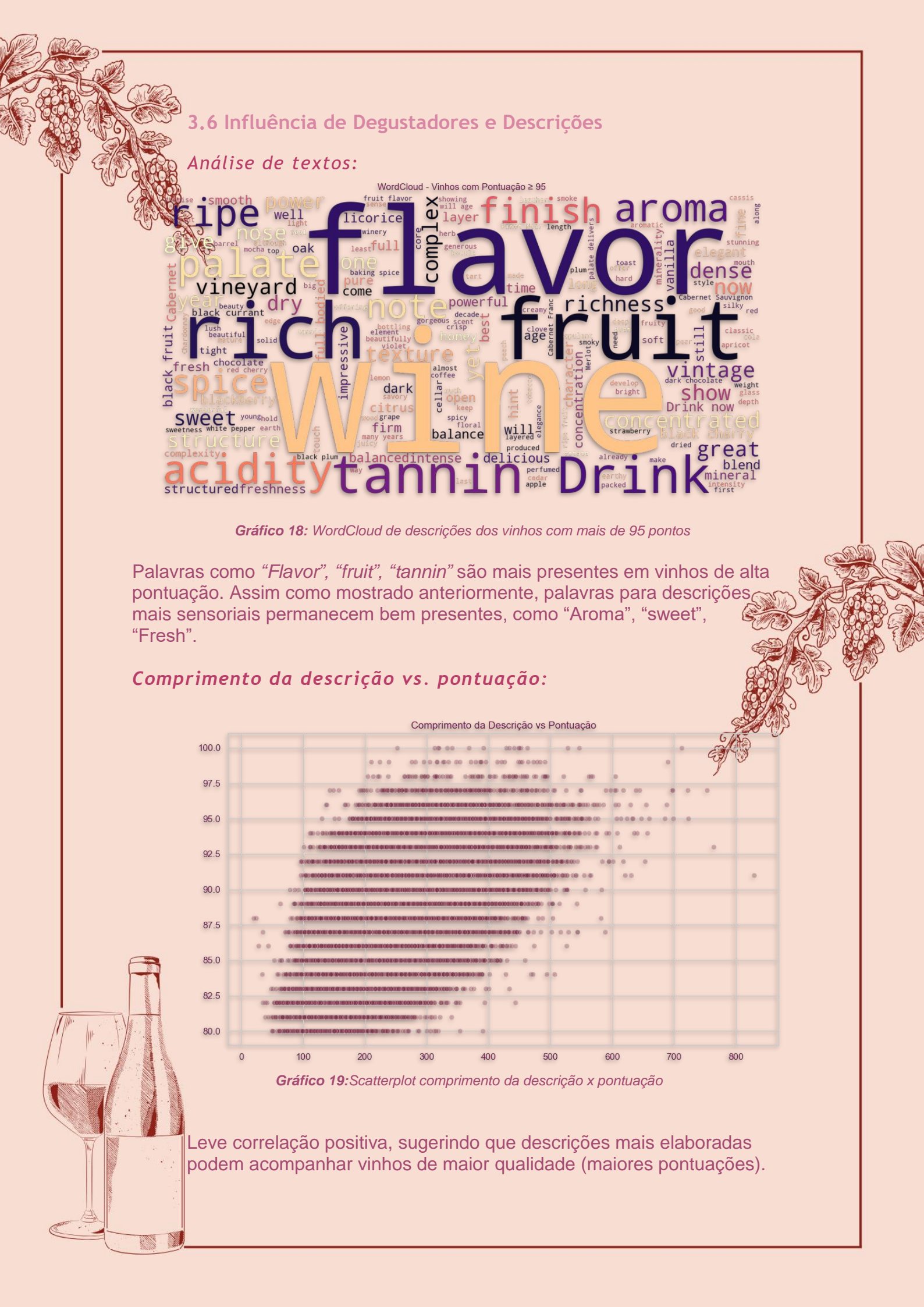
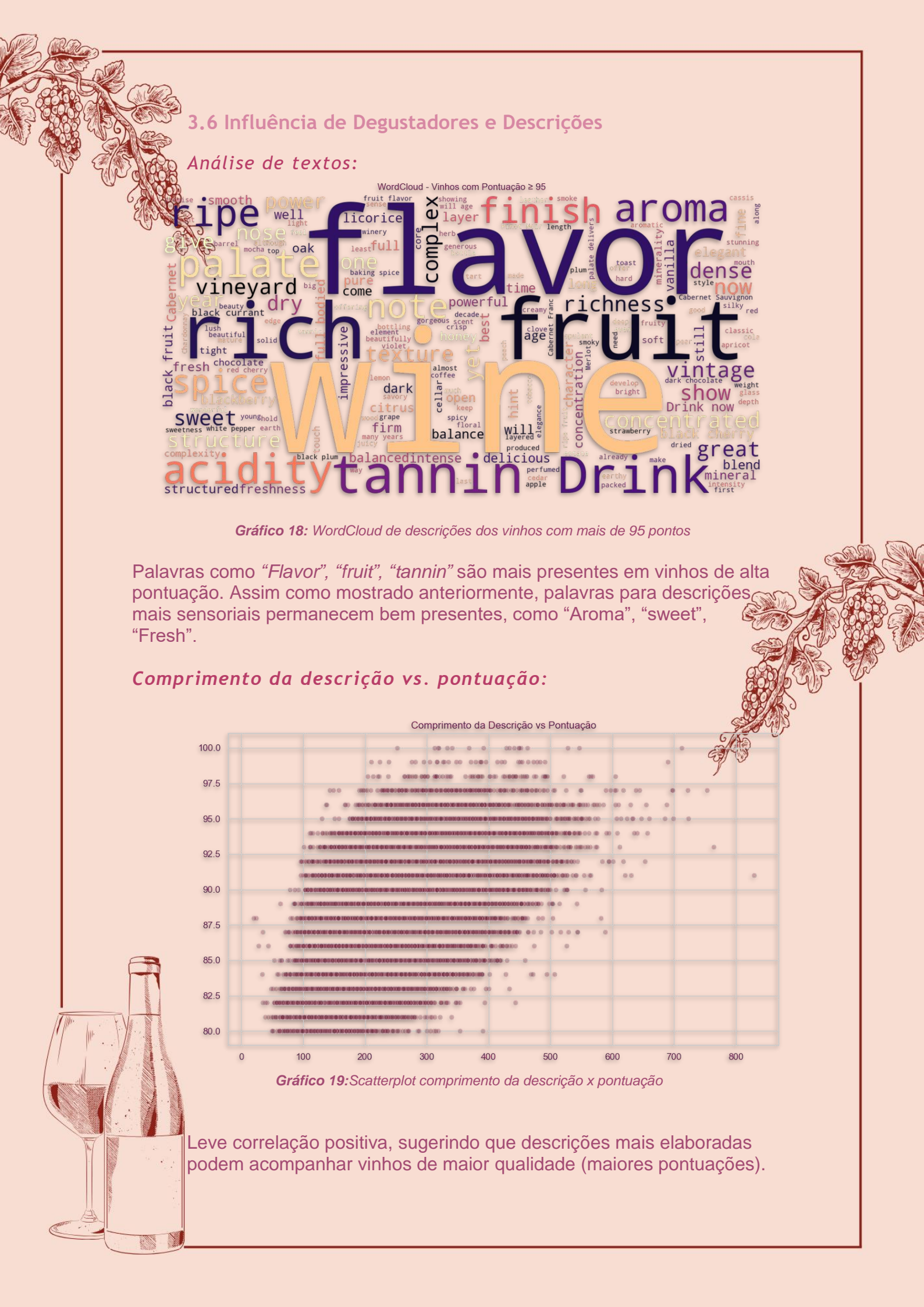
Palavras como “Flavor”, “fruit”, “tannin” são mais presentes em vinhos de alta pontuação. Assim como mostrado anteriormente, palavras para descrições mais sensoriais permanecem bem presentes, como “Aroma”, “sweet”, “Fresh”.

Comprimento da descrição vs. pontuação:

Comprimento da Descrição vs Pontuação

Gráfico 19: Scatterplot comprimento da descrição x pontuação

Leve correlação positiva, sugerindo que descrições mais elaboradas podem acompanhar vinhos de maior qualidade (maiores pontuações).

[illegible][illegible][illegible][illegible]

3.6 Influência de Degustadores e Descrições

Análise de textos:

WordCloud - Vinhos com Pontuação ≥ 95

3.7 Tendências Temporais e Sazonais

Análise temporal de pontos e preços:

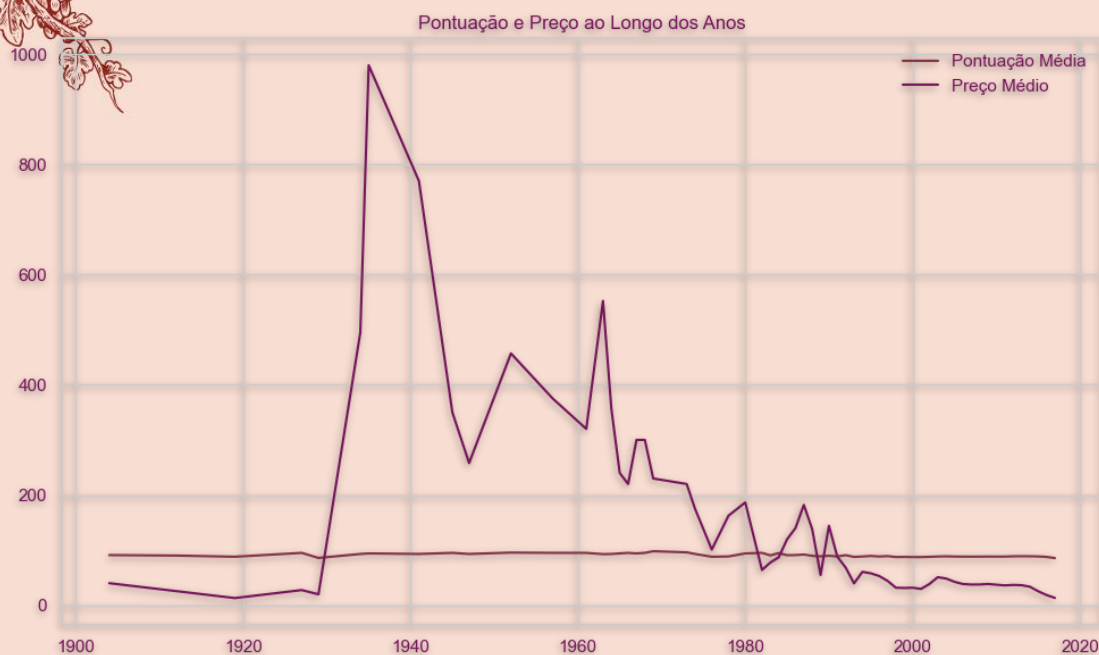


Gráfico 20: Heatmap de preços e pontuações ao longo dos anos

Ao analisar o gráfico é possível notar que periodicamente ocorrem variações nos preços, com períodos de aumento tanto em preço quanto na média de pontuação, porém, esse aumento e diminuição, não está diretamente correlacionado, existindo picos de preços, em momentos de pontuações médias “comuns”.





4. Considerações Finais

A análise confirma a dominância dos Estados Unidos em quantidade, porém países europeus como França, Itália e Portugal destacam-se em qualidade.

Existe uma relação positiva entre preço e pontuação, mas não é absoluta. Variedades como **'PinotNoir'**, **'Chardonnay'** e **'Cabernet Sauvignon'** são as mais populares.

Degustadores influenciam significativamente as avaliações, assim como a escolha das palavras nas descrições.

Existe uma tendência de leve aumento tanto nos preços quanto nas pontuações ao longo dos anos.

Tabela resumida dos dados:

Pergunta	Resumo da Resposta
País com mais vinhos	Estados Unidos (US)
Média de pontos por país (Top 5)	França, Alemanha, Áustria, Itália, Portugal
Correlação preço x pontuação	0.42 (positiva moderada)
Top 5 vinhos com maior pontuação	5 vinhos com 100 pontos (ex: Quinta do Noval 2011)
Top 5 vinhos com menor preço	Vinhos de US\$4 (ex: Terrenal 2010, Bandit NV)
Variedade mais comum	Pinot Noir
Vinho com descrição mais longa	Saggi 2007 Red (91 pts)
Top 10 vinhos mais caros	Ex: Château Pétrus (\$2500), Lafite Rothschild
Média de pontos por variedade (Top 5)	Riesling, Pinot Noir, Syrah, Chardonnay, Cabernet Sauvignon
Países com maior média de pontos (Top 5)	França, Alemanha, Áustria, Itália, Portugal
Média de preço por variedade (Top 5)	Ramisco, Terrantez, Francisa, Tinta Miúda, Merlot
Vinhos mais populares (títulos repetidos)	Ex: Gloria Ferrer NV (11x), Korbel NV (9x)






4.1 Limitações do Estudo

Apesar da riqueza do dataset utilizado, algumas limitações importantes foram identificadas e devem ser consideradas na interpretação dos resultados:

- **Ausência de dados temporais detalhados:** A base de dados não contém informações como safra do vinho, data de avaliação ou data de lançamento, o que inviabiliza análises de sazonalidade, envelhecimento e evolução histórica precisa.
- **Dados incompletos:** Algumas colunas possuem valores "desconhecido", o que pode distorcer médias ou análises por região e país.
- **Falta de informações químicas e sensoriais estruturadas:** Não há dados quantitativos sobre acidez, teor alcoólico, taninos ou outras propriedades físico-químicas dos vinhos, o que limita análises técnicas mais profundas.
- **Classificação por tipo de vinho (tinto, branco, rosé):** não está estruturada no dataset, sendo possível apenas inferir a partir da variedade da uva, o que exige mapeamento externo.
- **Informações sobre consumo e demografia do público consumidor:** não estão presentes, impedindo análises de comportamento ou preferências por perfil de consumidor.
- **Descrição textual subjetiva:** Embora as descrições dos vinhos sejam úteis para análises de linguagem, elas carregam um alto grau de subjetividade e variabilidade, dificultando padronizações.

Essas limitações reforçam a importância de tratar os achados como **análises exploratórias iniciais**, que podem ser aprofundadas em estudos futuros com bases mais completas.



5. Referências

- WineEnthusiast Magazine Dataset
- Documentações oficiais das bibliotecas Pandas, Seaborn, Matplotlib e WordCloud
- Storytelling com Dados: um Guia Sobre Visualização de Dados Para Profissionais de Negócios - Cole Nussbaumer Knaflic

