Qualidade De Vinhos: Exploração de Dados e Análise Estatística

1º João Lucas Oliveira Mota
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
jlucasoliveira2002@alu.ufc.br

2º Gabriel dos Reis Rodrigues
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
gabrieldosreis@alu.ufc.br

***Abstract*— Existem diferentes tipos de vinhos tintos e brancos. Considerando fatores sensoriais é possível classificá-los entre "bons" e "ruins", segundo especialistas. Nesse artigo, utilizamos de um dataset com medições físico-químicas de diferentes amostras de vinhos brancos e tintos, e sua classificação segundo esses especialistas, visando realizar o tratamento dos dados e a apresentação de relações entre as propriedades e suas classificações. Ademais, foi realizada uma análise exploratória dos dados, de forma univariada, bivariada e multivariada, elaborando gráficos e tabelas para melhor ilustrar os dados utilizados.**

Keywords—Análise estatística, Qualidade de Vinhos

I. Introdução

Devido à grande variedade de subtipos de vinhos, é possível tentar enxergar um padrão de qualidade, seguindo diferentes características físico-químicas e sensoriais deles.

No dataset Wine Quality [1], foram realizadas medições objetivas de características físico-químicas de exemplares de vinhos brancos e tintos. Para cada exemplar, também foi retirado um valor sensorial (uma média de pelo menos 3 avaliações feitas por especialistas em vinhos) que atribui um número entre 0 e 10, com zero sendo um vinho muito ruim e 10 um vinho excelente.

Neste artigo, realizamos uma análise estatística desses dados, buscando encontrar uma relação entre os valores obtidos nos testes objetivos e nas observações sensoriais.

II. Métodos

Nesta seção, vamos abordar os dados que estamos trabalhando, explicar as variáveis que vamos considerar nas análises e nos gráficos, e as técnicas que utilizamos para averiguar os dados.

A. Conhecimento necessário

Focamos, neste trabalho, na análise dos dados, com ênfase em seu desvio padrão, assimetria (*skewness*) e média.

O desvio padrão indica o quão os valores do conjunto estão dispersos. Ele mostra a distância dos valores em relação à média do conjunto. Para o cálculo do desvio padrão, utilizaremos a Equação (1):

$$\sigma_d = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n-1}} \qquad (1)$$

Nesta fórmula, temos n o número total de componentes, μ_x é a média dos componentes e x_i é o componente atual.

Também utilizaremos o valor de assimetria, que define o quão centralizado o preditor está, se a moda e a média estão na mesma posição. Para o cálculo da assimetria, utilizaremos a Equação (2):

$$Skewness = \frac{\sum(x_i - \mu_x)^3}{(n-1)(\sigma_d)^3} \qquad (2)$$

Sendo na fórmula 2, x_i o componente atual, μ_x a média dos componentes e σ_d o desvio padrão. Por fim, temos a fórmula da média, definida como o valor que demonstra a concentração dos dados de uma distribuição, como o ponto de equilíbrio das frequências em um histograma. Vide Equação (3):

$$\mu_x = \frac{\sum_{i=1}^{n} (x_i)}{n} \qquad (3)$$

Para as análises feitas neste artigo, utilizamos histogramas de frequência, como os gráficos 1 e 2, que indicam a distribuição dos valores do eixo X conforme os dados apresentam.

Também foram utilizados gráficos de dispersão, que possuem uma distribuição de pontos conforme a intersecção de valores de duas variáveis, e uma chamada reta de dispersão, que ilustra a relação entre essas variáveis e o seu comportamento. Elaboramos os gráficos com os dois eixos com preditores e pontos com uma escala de cores que varia do amarelo (para vinhos de qualidade baixa) e azul escuro (para vinhos de qualidade elevada).

Por fim, realizamos o procedimento de Principal Component Analysis (PCA).

O principal objetivo do PCA é conseguir representar o máximo de informações sobre um conjunto de dados em uma dimensão menor que a original com a menor perda possível de informações. Para tal é necessário termos uma matriz de covariância, para isso seja uma matriz X formada por "p" características (preditores) de "n" indivíduos, ou seja, uma matriz de ordem n x p. A partir de X obteremos a matriz de covariância simétrica e de tamanho p x p. Após isso é necessário normalizar os dados para que as características diferentes possam ser analisadas de maneira uniforme.

Por fim, é necessário calcular os autovetores, dessa forma teremos a expressão det(M - λ*I) = 0, sendo M nossa matriz de covariância normalizada, λ os autovalores e I a matriz identidade. Para cada λ existe um conjunto de autovetores (a_j) que são obtidos após solucionar a equação citada acima. O componente principal é obtido Equação 4:

$$Y_j = \sum_{i=1}^{p} a_i X_i \qquad (4)$$

B. *Apresentando os dados*

O conjunto de dados que utilizamos trata de um conjunto com 1599 vinhos tintos e 4898 vinhos brancos (N), e estabelece 11 variáveis preditoras (D), tais como pH, açúcar residual e etc. Também define 6 classes, para a qualidade do vinho (L), aferidas de 3 a 9, com as *class-distribution* dos dois tipos de vinho seguindo os gráficos abaixo:

Gráfico 1: Histograma de frequência da qualidade de vinhos brancos

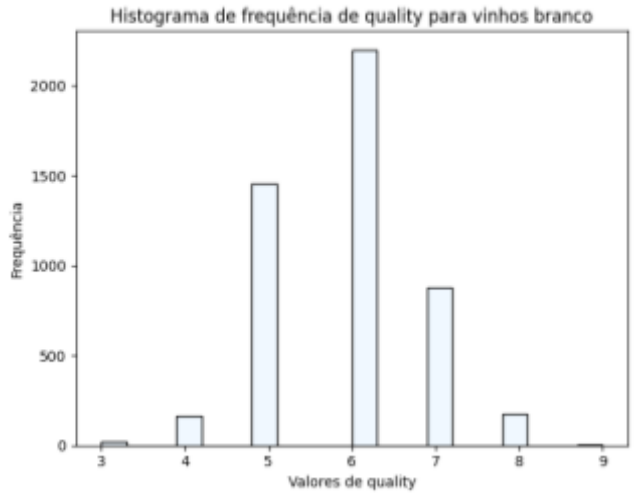
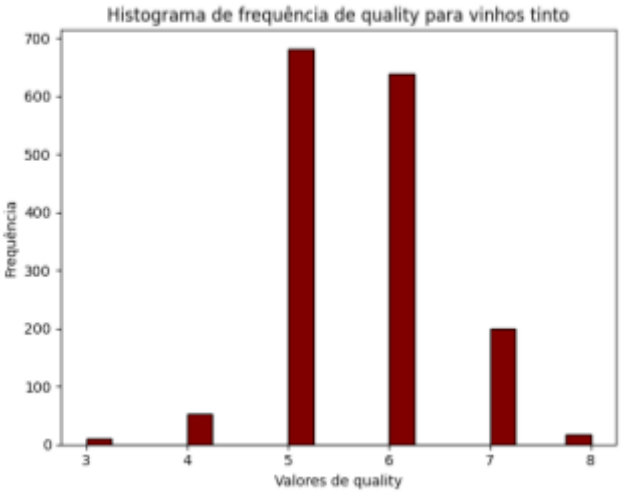


Gráfico 2: Histograma de frequência da qualidade de vinhos tintos



Tais gráficos mostram as frequências de qualidade elevadas em 6 e 5, para os conjuntos de vinhos brancos e tintos, respectivamente.

Os conjuntos de dados utilizados é completo, sem nenhum valor de atributo ausente, logo não foi necessário realizar qualquer manipulação dos dados a priori.

.

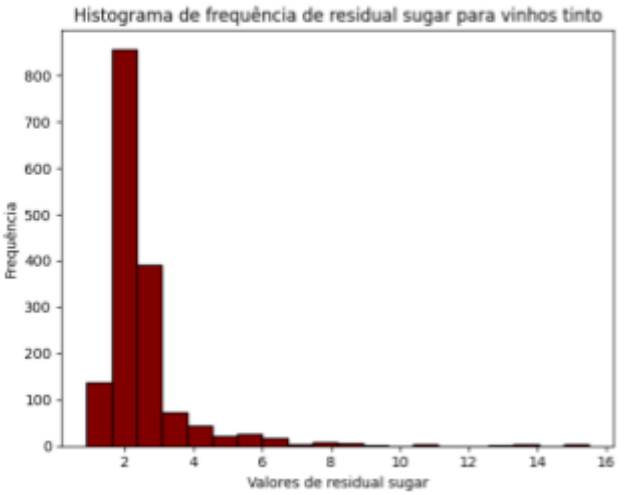
III. Resultados

A. *Análise Incondicional Univariada*

No código, apresentamos as análises mono-variadas para os dois tipos de vinho, além de suas respectivas tabelas com valores de média, desvio padrão e assimetria.

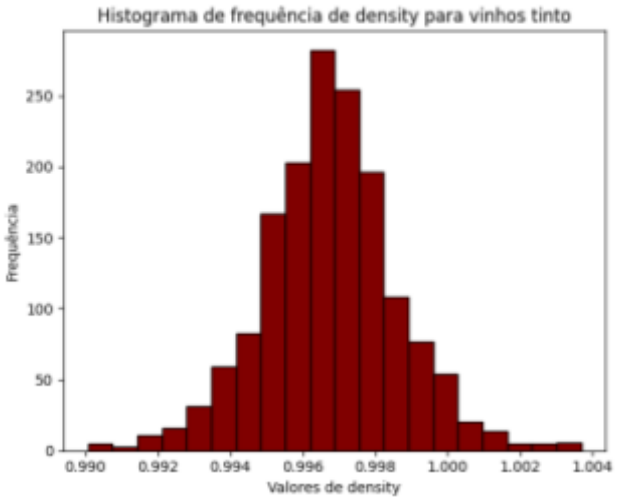
Dentre os resultados observados, é possível destacar alguns casos interessantes:

Gráfico 3: Histograma de frequência de açúcar residual para vinhos tintos



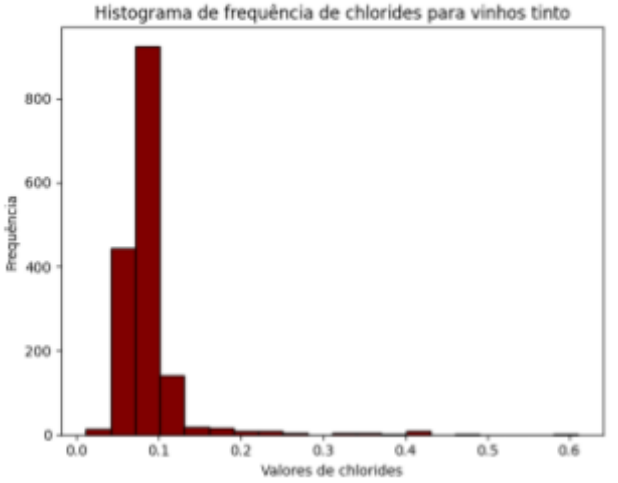
No gráfico 3, nota-se uma concentração de frequência de açúcar residual muito elevada em torno de 2 (média calculada μ_x = 2,53), com alguns outliers de valores mais elevados, mas em pouquíssima frequência quando comparado ao resto dos dados.

Gráfico 4: Histograma de frequência de densidade para vinhos tintos



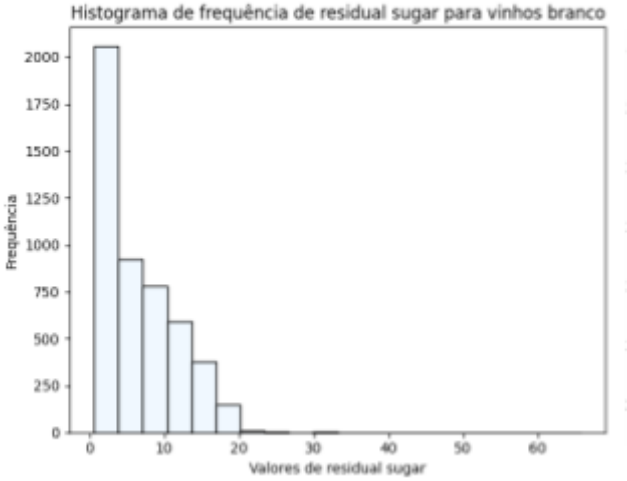
No gráfico 4, observa-se uma distribuição muito simétrica dos dados considerados (o valor de skewness é um dos mais baixos da tabela, com 0,07), e o menor desvio padrão (σ_d = 0,0018 segundo os cálculos).

Gráfico 5: Histograma de frequência de Cloretos para vinhos tintos



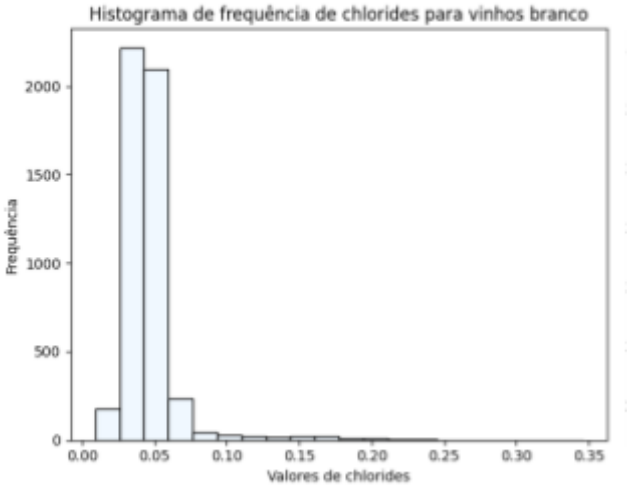
O gráfico 5, com os cloretos em vinhos tintos, apresenta um baixíssimo desvio padrão (σ_d = 0,04 segundo os cálculos), o que, em conjunto com uma frequência elevada em valores abaixo de 0,1, indica uma quantidade pequena de outliers, muito embora estes ainda estejam presentes em valores de cloretos mais elevados.

Gráfico 6: Histograma de frequência de açúcar residual para vinhos brancos



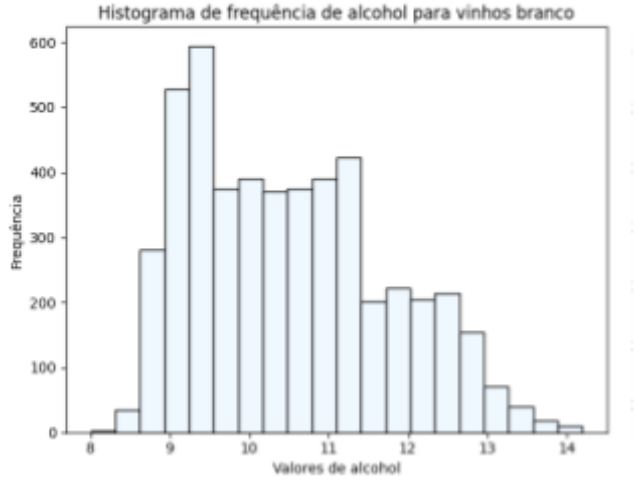
O gráfico 6 representa a frequência de açúcar residual em vinhos brancos. É interessante ressaltar a diferença entre os valores presentes nos vinhos brancos e tintos, visto que a média que o gráfico 6 indica (μ_x = 6,39) é mais elevada em comparação com a que o gráfico 3 (com μ_x = 2,53), que representa a frequência do mesmo preditor para vinhos tintos, indicando uma presença maior de açúcar residual nas amostras de vinho branco que nas amostras de vinho tinto.

Gráfico 7: Histograma de frequência de cloretos para vinhos brancos



O gráfico 7 mostra a frequência de cloretos para vinhos brancos. Nota-se uma semelhança nos padrões dos gráficos 7 e 5, ambos para o mesmo preditor e com baixíssimos desvios padrões (σ_d = 0,02 para o gráfico 7), porém o valor da média acaba caindo pela metade para os vinhos brancos (μ_x = 0,04, enquanto o μ_x = 0,08 para o gráfico 5), indicando uma presença muito menor de cloretos em vinhos brancos que em vinhos tintos.

Gráfico 8: Histograma de frequência de álcool para vinhos brancos



Por fim, temos o gráfico 8, que tem os valores de álcool para vinhos brancos. Note que este gráfico é o que possui a menor variação entre as distribuições de frequência, com valores distribuídos de forma bastante uniforme. Isso indica uma grande variação entre os níveis de álcool das amostras de vinho branco.

B. Análise Class-Conditional Univariada

Nessa seção, realizamos uma análise class-condicional do conjunto de dados.

Primeiro consideramos uma comparação de preditores para vinhos tintos. Observe os gráficos abaixo:

Gráfico 9: Histograma de frequência sulfatos para vinhos tintos (classe 8)

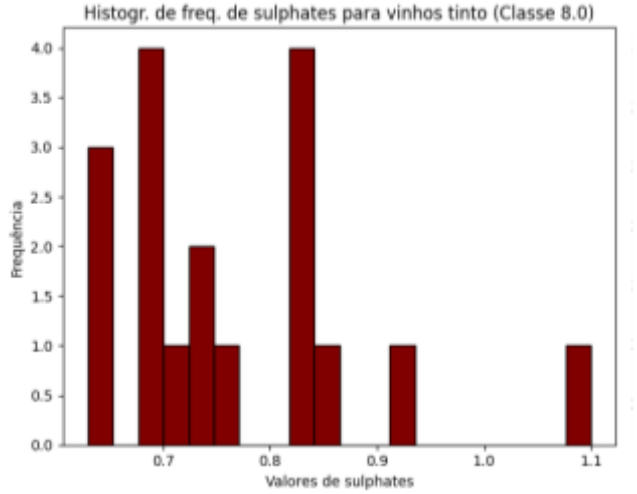
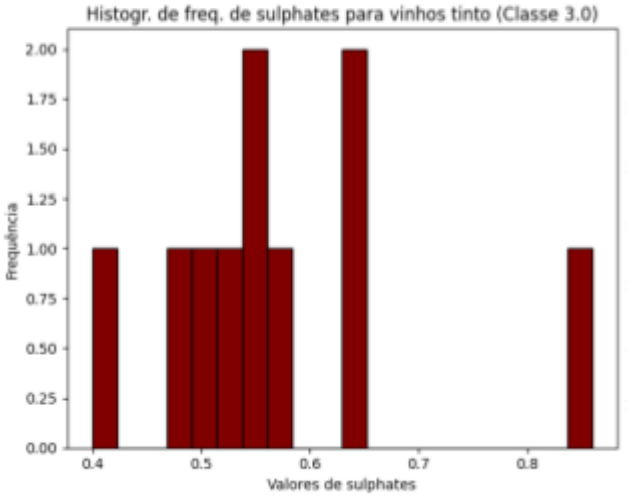


Gráfico 10: Histograma de frequência sulfatos para vinhos tintos (classe 3)



O gráfico 9 ilustra dados dos vinhos tintos de classe 8 com base no preditor da taxa de sulfatos. Segundo os dados de média (μ_x = 0,76) e desvio padrão (σ_d = 0,11), e considerando o gráfico do mesmo preditor para vinhos tintos de classe 3 (gráfico 10), com sua média (μ_x = 0,57) indicando um valor menor, temos que o valor de sulfatos mais elevado pode estar diretamente relacionado a uma maior qualidade do vinho tinto.

Em seguida, realizamos um procedimento parecido, dessa vez focando nos preditores de vinhos brancos. Novamente, observe os gráficos abaixo:

Gráfico 11: Histograma de frequência de álcool para vinhos brancos (classe 9)

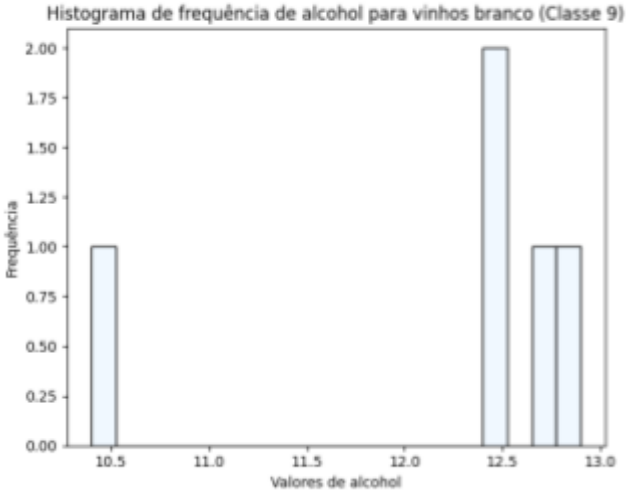
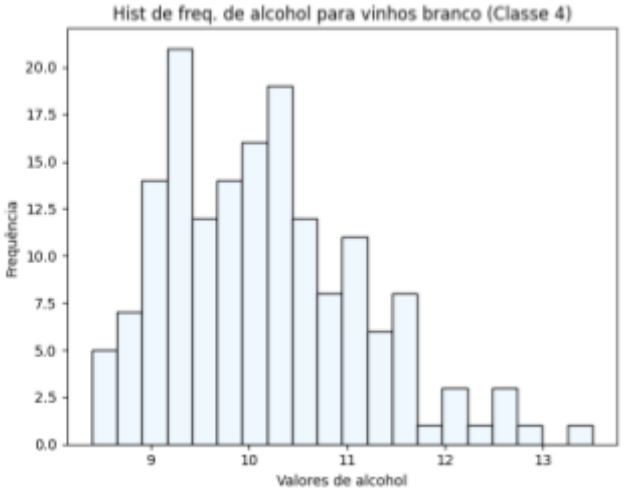


Gráfico 12: Histograma de frequência álcool para vinhos brancos (classe 4)



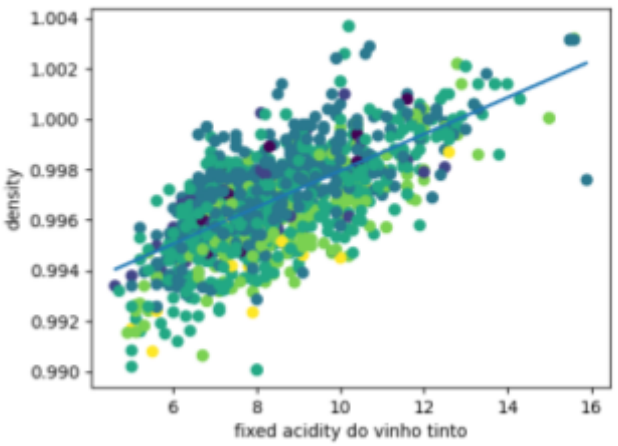
Os gráficos escolhidos (Gráficos 11 e 12) representam os histogramas de frequência para álcool em vinhos brancos, respectivamente para vinhos de classe 9 e 4.

Considerando os valores de média do gráfico 11 (μ_x = 12.18), e do gráfico 12 (μ_x = 10.15), e seus desvios padrões respectivamente (σ_d = 1,013 e σ_d =1,00), é possível relacionar de forma diretamente proporcional o aumento de nível de álcool com o aumento da qualidade do vinho branco.

C. Análise Bivariada

Na análise bivariada realizamos a plotagem dos gráficos abaixo, considerando a relação entre a variação simultânea de 2 preditores simultaneamente, em função da variação da qualidade dos vinhos testados. A qualidade está representada em formato de gradiente de cores, com pontos mais próximos do amarelo para vinhos de baixa qualidade e mais próximos do azul escuro para vinhos de alta qualidade.

Gráfico 13: Gráfico de dispersão de acidez fixa por densidade para vinhos tintos

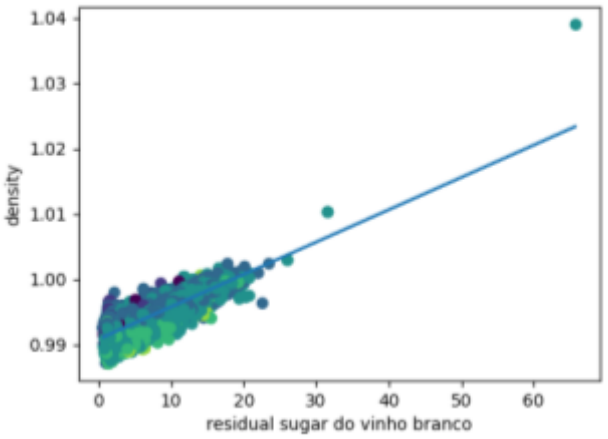


Observando o Gráfico 13, em conjunto com o valor presente na Matriz de Correlação do vinho tinto presente no código (r = 0,67) percebemos uma relação entre os dois preditores.

Também é possível observar um certo padrão na distribuição dos pontos no gráfico de dispersão, com pontos mais claros abaixo da linha e pontos mais escuros acima dela.

Considerando esse padrão, é possível definir uma correlação entre a proporção de valores dos preditores de densidade e acidez fixa e a qualidade do vinho tinto a ser observado.

Gráfico 14: Gráfico de dispersão de açúcar residual por densidade para vinhos brancos



Por fim, consideramos o gráfico 14, com os preditores de açúcar residual e densidade, dessa vez para vinhos tintos. Ao observar este gráfico, em conjunto com novamente o seu valor correspondente na Matriz de Correlações (r = 0.84) percebemos uma grande correlação entre os dois preditores.

Além disso, tendo em vista o padrão de dispersão dos pontos no gráfico, é possível inferir uma correlação entre a proporção de valores dos preditores de densidade e açúcar residual e a qualidade do vinho branco a ser observado.

D. Análise Multivariada

No conjunto de dados utilizado, temos uma gama diversa de preditores, portanto para uma melhor análise faz-se necessário realizar uma análise multivariada.

Para tal, usaremos do da técnica de PCA, explicada anteriormente, com a restrição de dimensão n = 2 chegando a dois autovalores. Aplicando o screen plot para os dois tipos de vinhos temos:

Gráfico 15: Screen plot para vinhos tintos

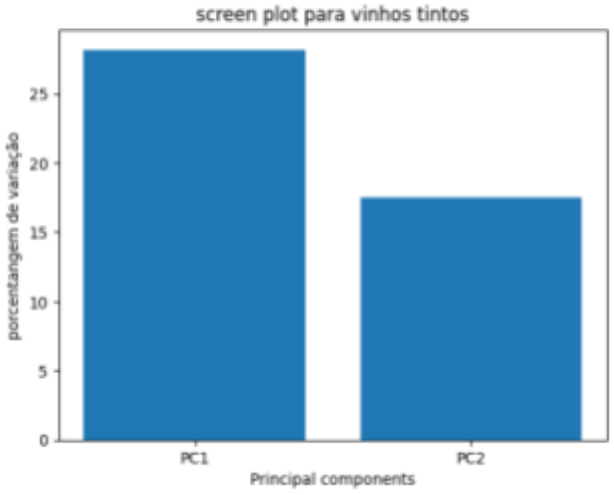
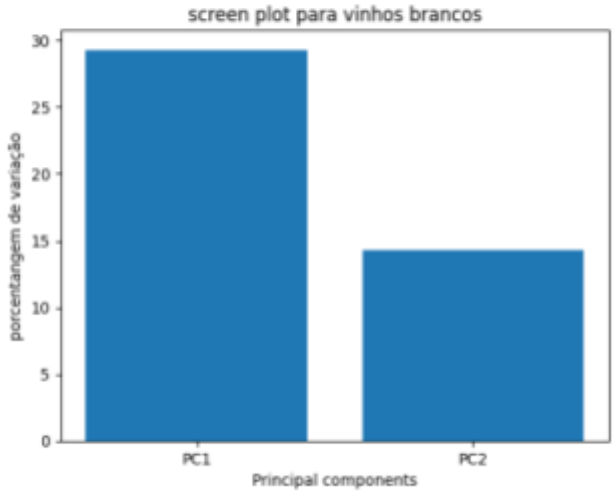


Gráfico 16: Screen plot para vinhos brancos



Como é possível ver, os dois componentes, tanto para o dataset do vinhos brancos quanto dos vinhos tintos, não são suficientes para uma boa representação dos dados.

IV. Referências

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. **Modeling wine preferences by data mining from physicochemical properties**. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. Disponível em: https://archive.ics.uci.edu/dataset/186/wine+quality

[2] JAMES, Gareth *et al*. **An Introduction to Statistical Learning with Applications in Python**. S.I: Springer, 2023. Disponível em: https://hastie.su.domains/ISLP/ISLP_website.pdf. Acesso em: 10 set. 2023.

[3] VARELLA, Carlos Alberto Alves. **Análise de Componentes Principais**. Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf. Acesso em: 10 set. 2023