Análise Exploratória de Dados

1. Carregamento dos Dados

Em primeiro lugar, foi feito o processo de carregamento dos dados através da biblioteca scikit-learn, com o objetivo de visualizar as variáveis e seus significados com seus valores.

Temos como variáveis de entrada e saída do dataset:

Entrada	Saída (apenas uma variável)
alcohol	target
malic_acid	
ash	
alcalinity_of_ash	
magnesium	
total_phenols	
flavanoids	
nonflavanoid_phenols	
proanthocyanins	
color_intensity	
hue	
od280/od315_of_diluted_wines	
proline	

2. Descrição Estatística dos Dados

A partir do comando .describe() em Python é possível visualizar métodos estatísticos do DataFrame, que é útil para nossa análise:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	0.957449	2.611685	746.893258	0.938202
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	0.228572	0.709990	314.907474	0.775035
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	0.480000	1.270000	278.000000	0.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	0.782500	1.937500	500.500000	0.000000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	0.965000	2.780000	673.500000	1.000000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	1.120000	3.170000	985.000000	2.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	1.710000	4.000000	1680.000000	2.000000

A partir dessa descrição percebemos que temos 178 amostras do dataset (count = 178).

Também é possível visualizar uma descrição estatística de seus dados através da propriedade DESCR. Ao utilizarmos o comando abaixo:

```
print(wine_data.DESCR)
```

obtemos as seguintes informações sobre nosso dataset:

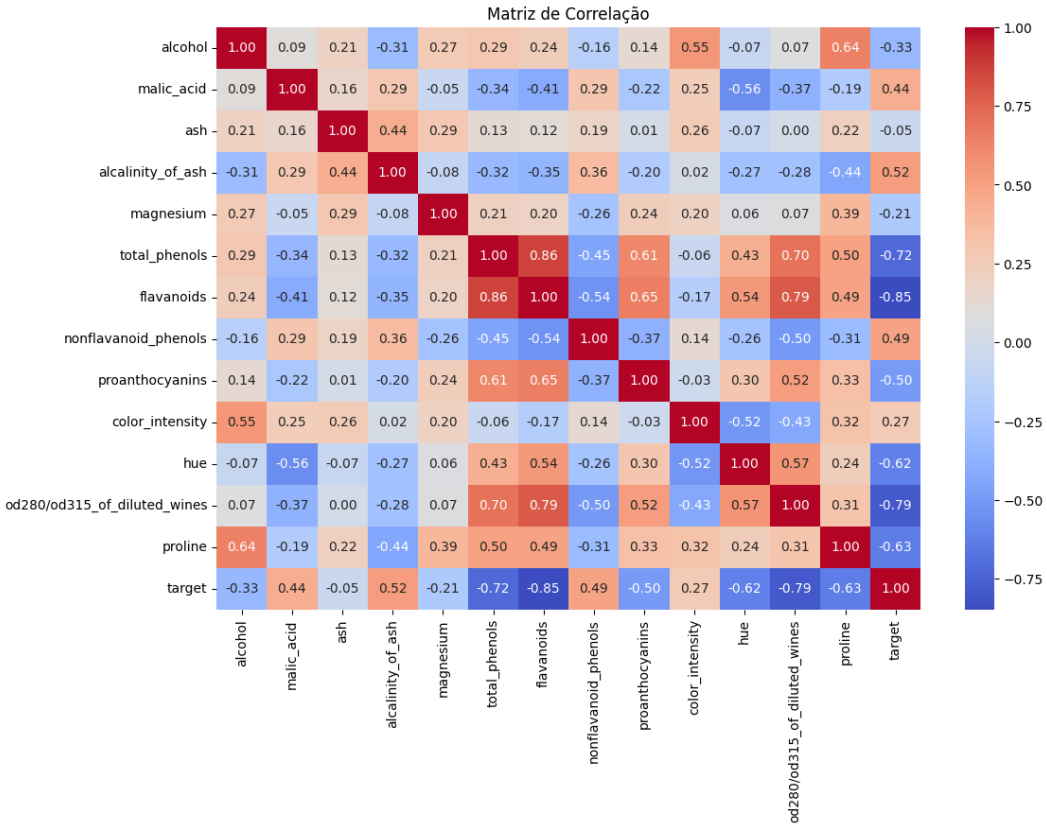
```
**Data Set Characteristics:**

:Number of Instances: 178
:Number of Attributes: 13 numeric, predictive attributes and the class
:Attribute Information:
    - Alcohol
    - Malic acid
    - Ash
    - Alcalinity of ash
    - Magnesium
    - Total phenols
    - Flavanoids
    - Nonflavanoid phenols
    - Proanthocyanins
    - Color intensity
    - Hue
    - OD280/OD315 of diluted wines
    - Proline
    - class:
        - class_0
        - class_1
        - class_2

:Summary Statistics:

============================== ==== ===== ======= =====
                                Min   Max   Mean     SD
============================== ==== ===== ======= =====
Alcohol:                      11.0  14.8    13.0   0.8
Malic Acid:                   0.74  5.80    2.34  1.12
Ash:                          1.36  3.23    2.36  0.27
Alcalinity of Ash:            10.6  30.0    19.5   3.3
Magnesium:                    70.0 162.0    99.7  14.3
Total Phenols:                0.98  3.88    2.29  0.63
Flavanoids:                   0.34  5.08    2.03  1.00
Nonflavanoid Phenols:         0.13  0.66    0.36  0.12
Proanthocyanins:              0.41  3.58    1.59  0.57
Colour Intensity:              1.3  13.0     5.1   2.3
Hue:                          0.48  1.71    0.96  0.23
OD280/OD315 of diluted wines: 1.27  4.00    2.61  0.71
Proline:                       278  1680     746   315
============================== ==== ===== ======= =====
```

O número de instâncias (178), bate com a propriedade “count” que obtivemos com o método describe(), que significa o número de amostras que temos do nosso conjunto de dados.

Além disso, é mostrado também que não temos **nenhum atributo faltante dos dados**, e vemos que a classe 1 (target = 1) é a classe predominante do nosso dataset:

```
:Missing Attribute Values: None
:Class Distribution: class_0 (59), class_1 (71), class_2 (48)
```

3. Matriz de Correlações



A partir da matriz de correlações foi identificado as seguintes variáveis com forte correlação:

- *flavanoids* e *total_phenols* (Correlação positiva): 0.86
- flavanoids / od280/od315_of_diluted_wines: 0.79
- *target* e *flavanoids* (Correlação negativa): -0.85 (variável de entrada e saída)

Altos valores de correlação entre variável de entrada e saída devem ser mantidas, pois significam variáveis relevantes.

Altos valores de correlação entre variáveis de entrada devem retiradas, pois significam informações redundantes.

Portanto, se a redução de features for necessária, podemos manter apenas **flavanoids** e não utilizar *total_phenols* e *od280/od315_of_diluted_wines.*

4. Matriz de Scatterplots

A partir da matriz acima, é possível concluir que as variáveis que mostram padrões semelhantes são *flavanoids* e *total_phenols.* É possível tirar essa conclusão a partir do comportamento linear e conciso que os dados das duas variáveis apresentam. Isso significa que elas apresentam redundância dentro do nosso conjunto de dados também.

É possível perceber também que a variável `proline` separa bem as classes também, especialmente a combinação de `proline` e `d280/od315_of_diluted_wines,` e `proline` e `hue`.

5. Boxplots

O boxplot nos ajuda a analisar a correlação entre os valores de uma variável com medidas estatísticas.

Pontos importantes para a interpretação de um Boxplot:

Posição

Para ver a posição dos dados, analisamos a mediana (2o quartil) do nosso boxplot.

Dispersão

Pode ser verificada pelos intervalos interquartis ou pela amplitude (valor máx. - valor mín.)

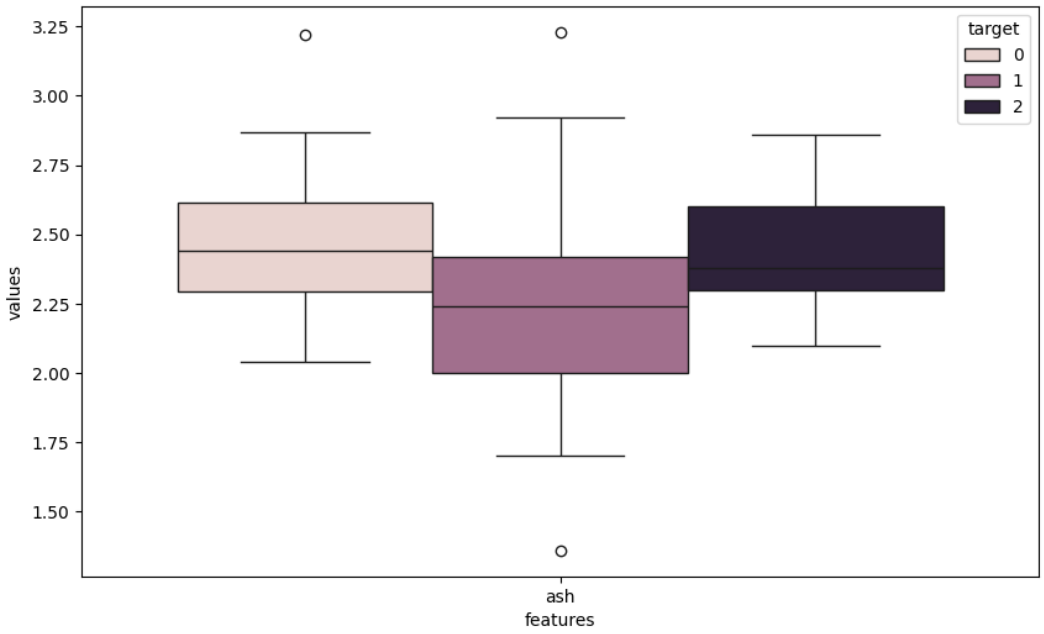
Simetria

Podemos analisar a simetria dos valores da variável através da mediana (2o quartil). Quando a linha da mediana está mais próxima do primeiro quartil, os dados são assimétricos positivamente. Caso esteja mais próximo do terceiro quartil, irão ser assimétricos negativamente.

Foi feita uma readaptação do código para plotar o **boxplot de cada variável individualmente,** ao invés de todas juntas, para facilitar a visualização e a análise de cada variável.

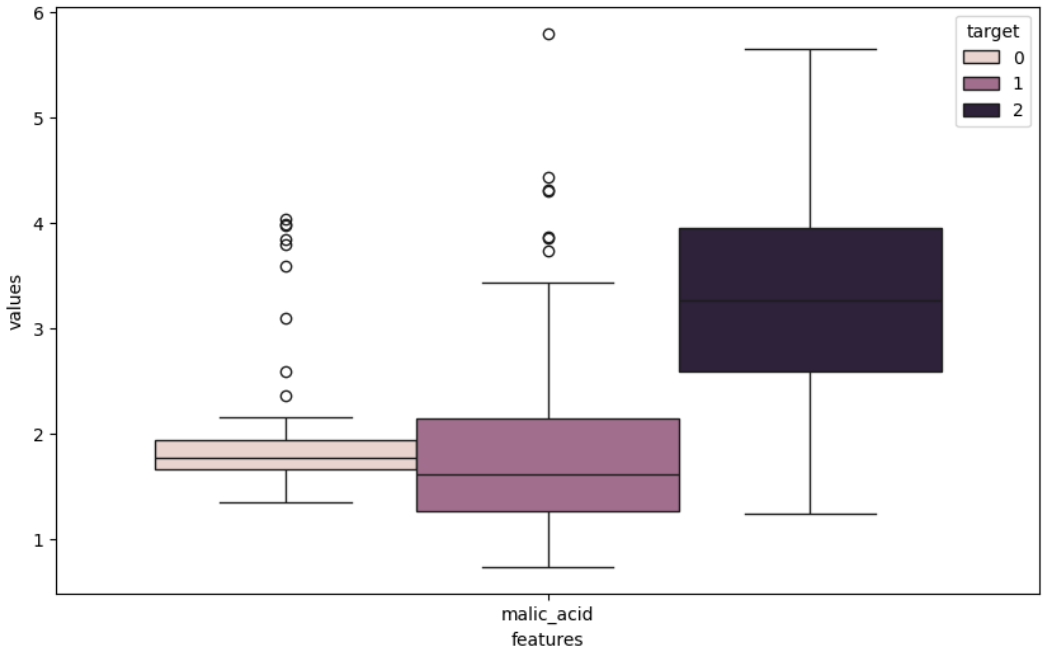
Variáveis irrelevantes

Abaixo se encontra o Boxplot da variável `ash`, que apresenta distribuições muito similares entre as classes, indicando que pode ser irrelevante para o problema de classificação. O `max_value` das 3 classes dessa variável são praticamente iguais.

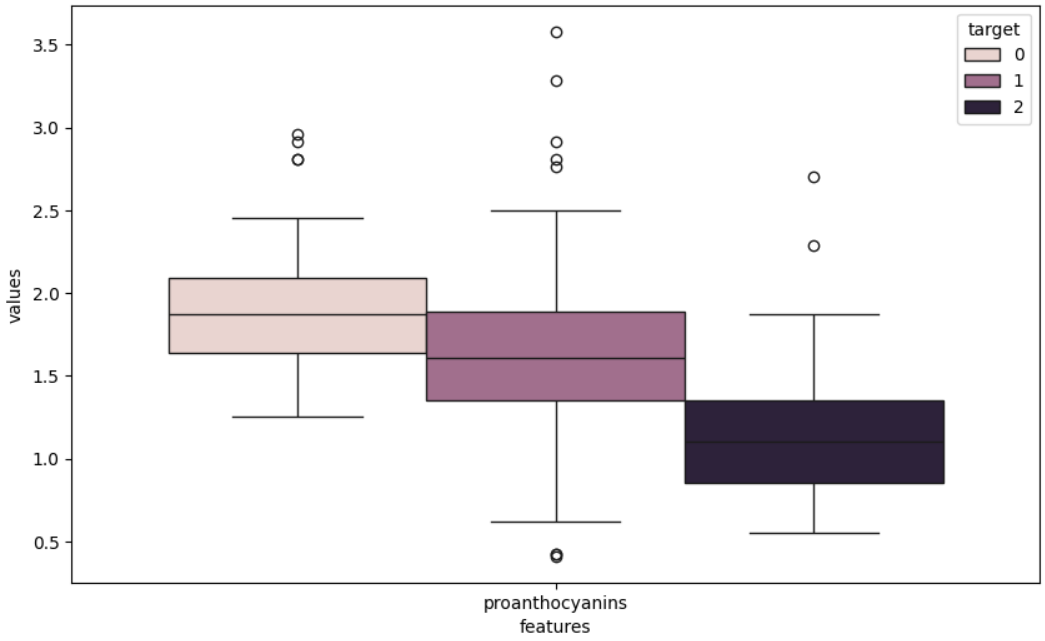


Outliers

Foi identificado também que a variável `malic_acid` apresenta bastantes outliers em duas classes de seu conjunto, como mostra a imagem abaixo:



Outra variável que apresenta outliers significativos é a `proanthocyanins`, que é perceptível em suas 3 classes abaixo:



6. Conclusão

1. Quais variáveis apresentam alta correlação entre si? Explique por que você acredita que são redundantes.

Nota-se que as variáveis `flavanoids` e `total_phenols`

`flavanoids` e `od280/od315_of_diluted_wines` possuem alta correlação entre si, a partir dos valores obtidos com a matriz de correlação entre essas variáveis.

Elas são redundantes por possuirem alto valor de correlação entre si, sendo diretamente proporcionais (ambos valores de correlação são positivos).

2. Há variáveis que, com base nos scatterplots e boxplots, parecem não ajudar a distinguir as classes? Quais você considera irrelevantes?

A partir dos scatterplots e boxplots, concluimos que a variável `ash` é irrelevante para nossa análise, pois não ajuda a disntinguir as classes nos **scatterplots,** além

de que as distribuições para suas 3 classes no **boxplot** são muito semelhantes entre si, indicando que pode ser uma variável irrelevante se formos treinar um modelo para **classificação.**

3. Quais variáveis você consideraria remover para otimizar o modelo de classificação, baseado nas observações feitas?

Bom, a partir das análises acima, acredito que três variáveis do dataset poderiam ser removidas, inicialmente, para treinar o modelo e verificar como ele iria se comportar. As variáveis são:

`total_phenols`: por possuir uma alta correlação linear com falavoids.

`od280/od315_of_diluted_wines`: por possuir uma alta correlação linear com falavoids.

`ash`: por possuir distribuições muito parecidas entre as diferentes classes no Boxplot.

Dessa forma, o modelo poderia ficar mais otimizado durante a classificação do problema.