Machine Learning com Python

Prof. Luciano Galdino

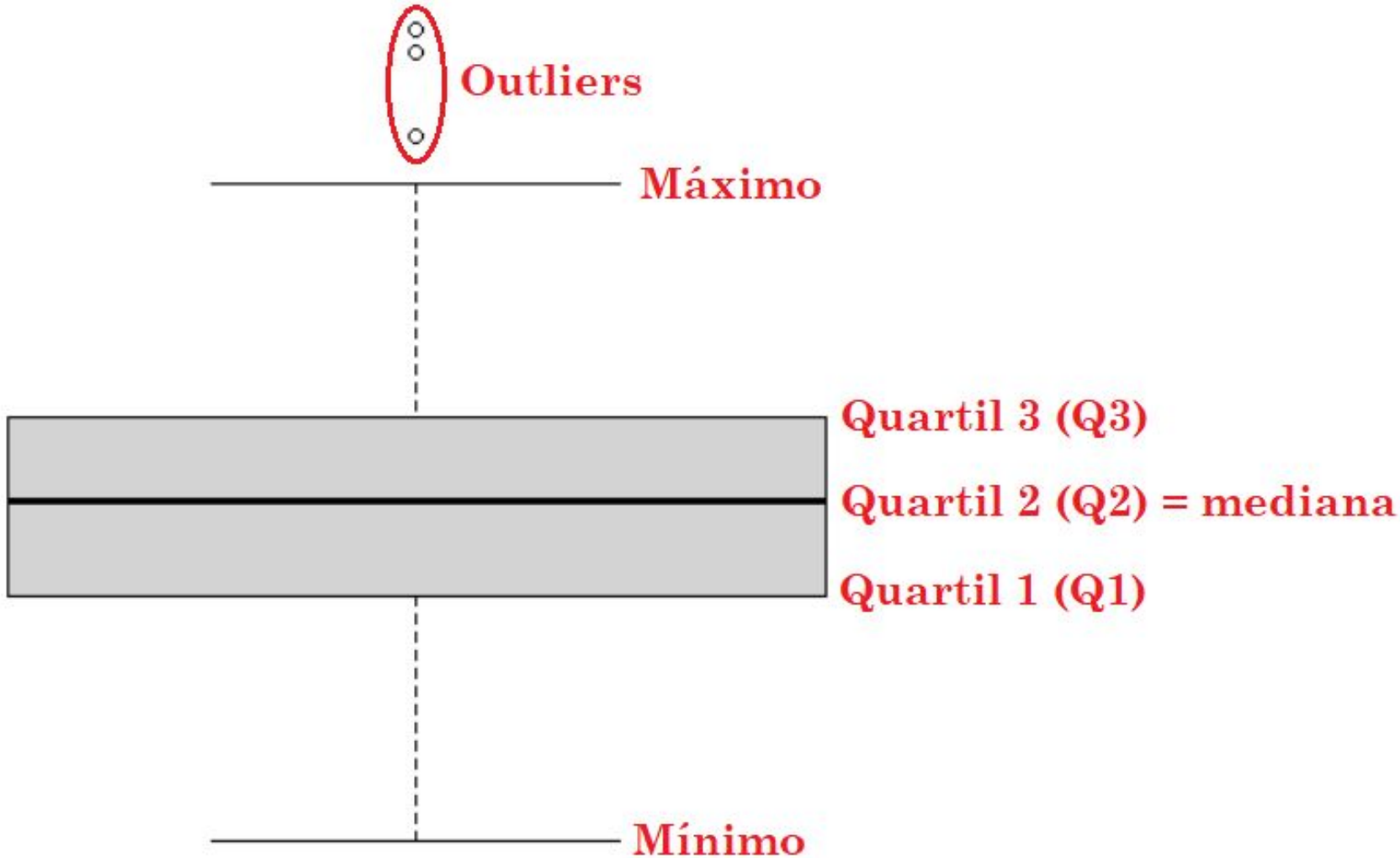
OUTLIERS

Outliers são dados discrepantes, isto é, são dados muito diferentes dos demais dados pertencentes à variável em análise.

A relevância deles deve ser analisada para definir se continuarão no dataset ou se devem ser tratados (corrigidos, excluídos ou substituídos), pois se não forem relevantes, podem interferir significativamente nos resultados das análises.

Eles podem ser identificados por observações diretas no dataset (quando a quantidade for pequena), por gráficos e por funções específicas.

O gráfico mais utilizado para identificar outliers é o BoxPlot.



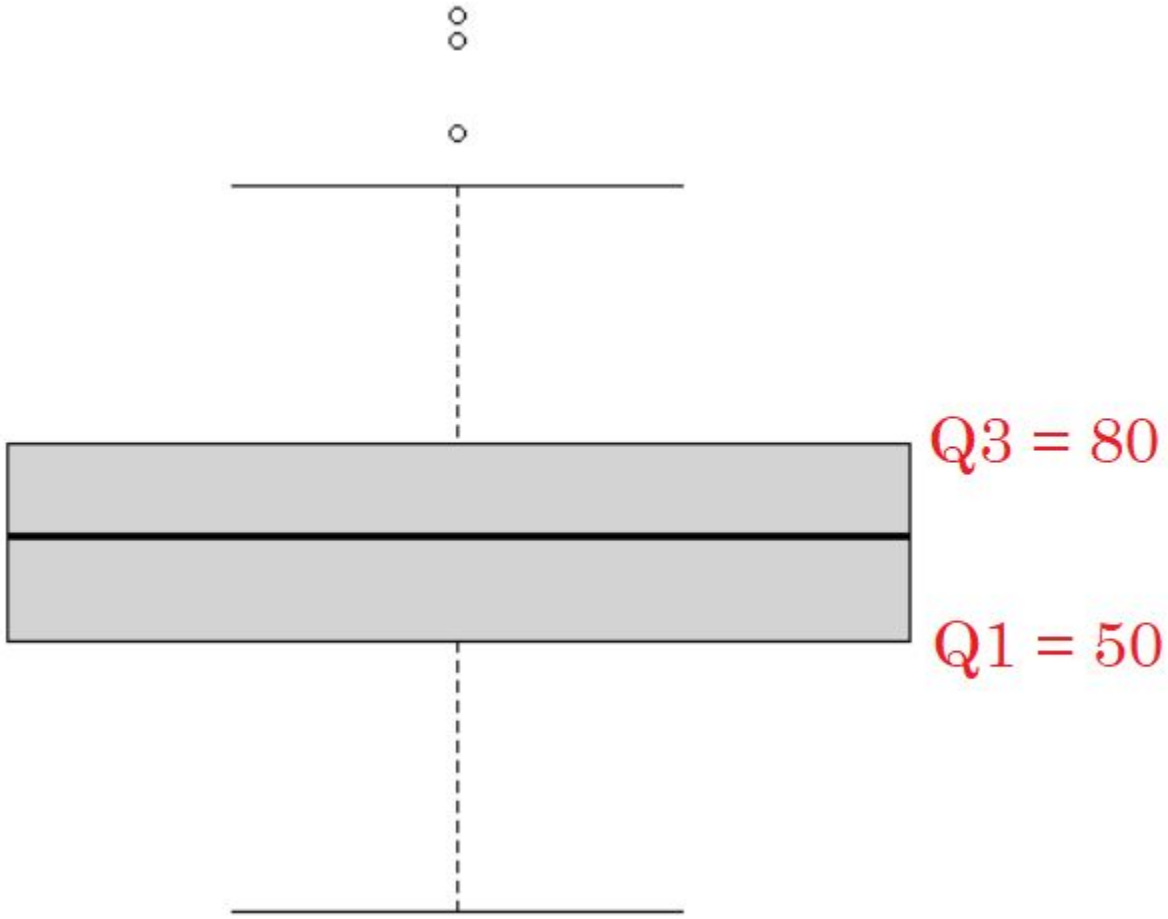
Cálculo da faixa de valores que limitam os outliers

A referência de cálculo é o intervalo interquartil (IQR) e obedece a seguinte equação:

Limite superior = Q3 + 1,5 x IQR

Limite inferior = Q1 - 1,5 x IQR

Exemplo



IQR = 80 - 50 = 30

Limite superior = 80+1,5x30 = 125

Limite inferior = 50 - 1,5x30 = 5

Conclusão:

Outliers > 125

Outliers < 5