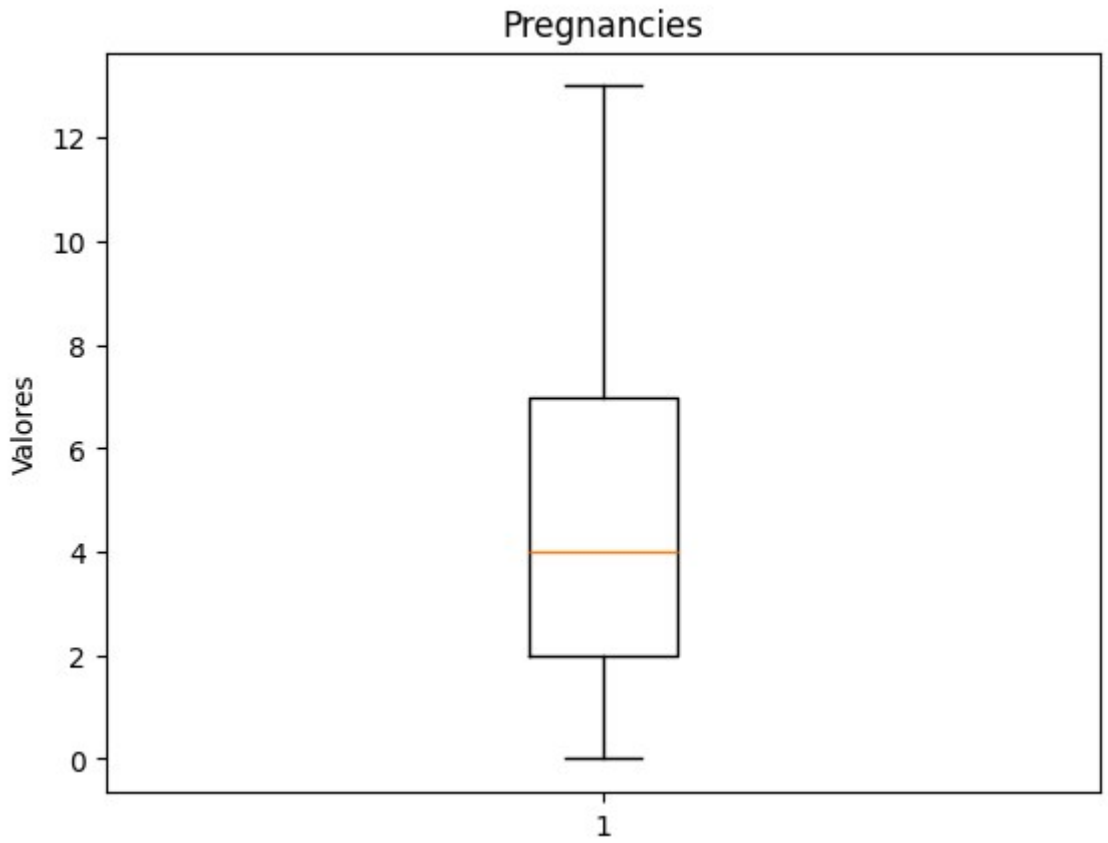
Avaliação da Escala dos Dados – “diabetes.csv”

Utilizando a plotagem do gráfico Boxplot da biblioteca MatPlotLib, foi possível analisar quais colunas do dataset possuíam outliers;

Evidências

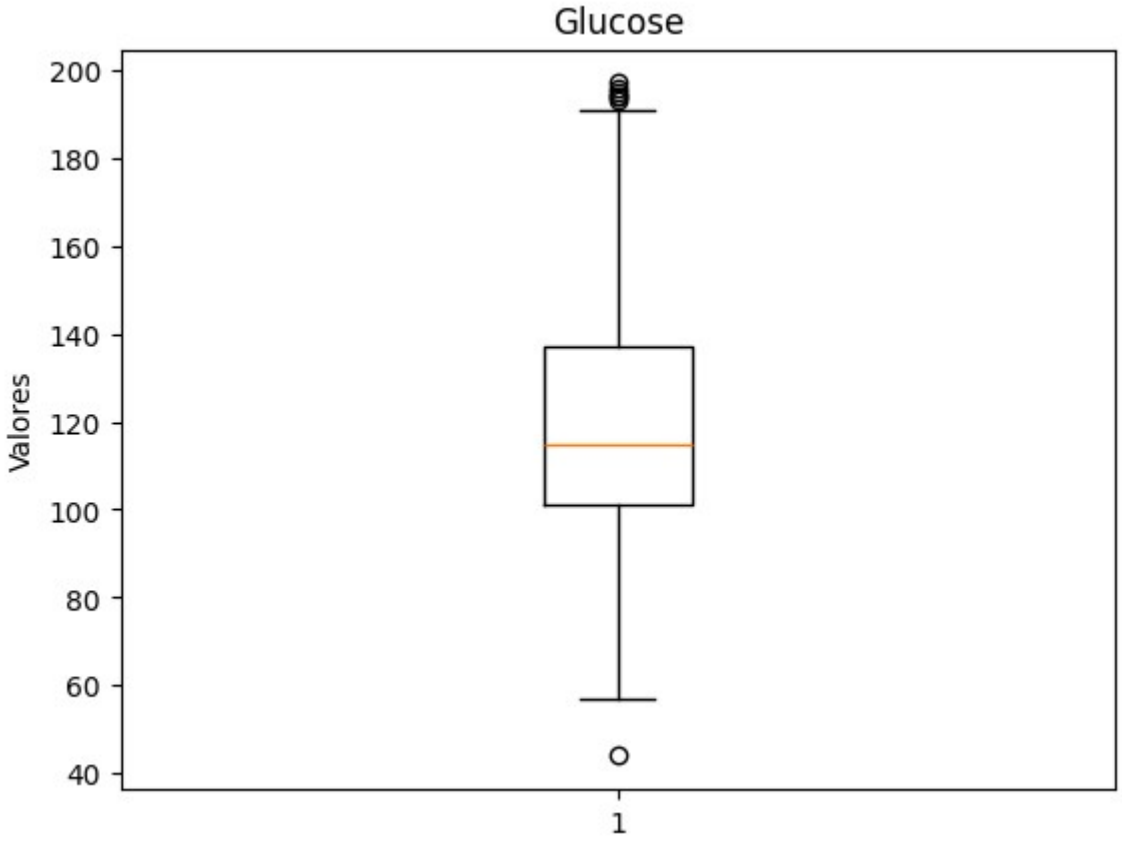
Pregnacies – NÃO há valores atípicos



count	375.000000
mean	4.333333
std	3.331639
min	0.000000
25%	2.000000
50%	4.000000
75%	7.000000
max	13.000000

Glucose

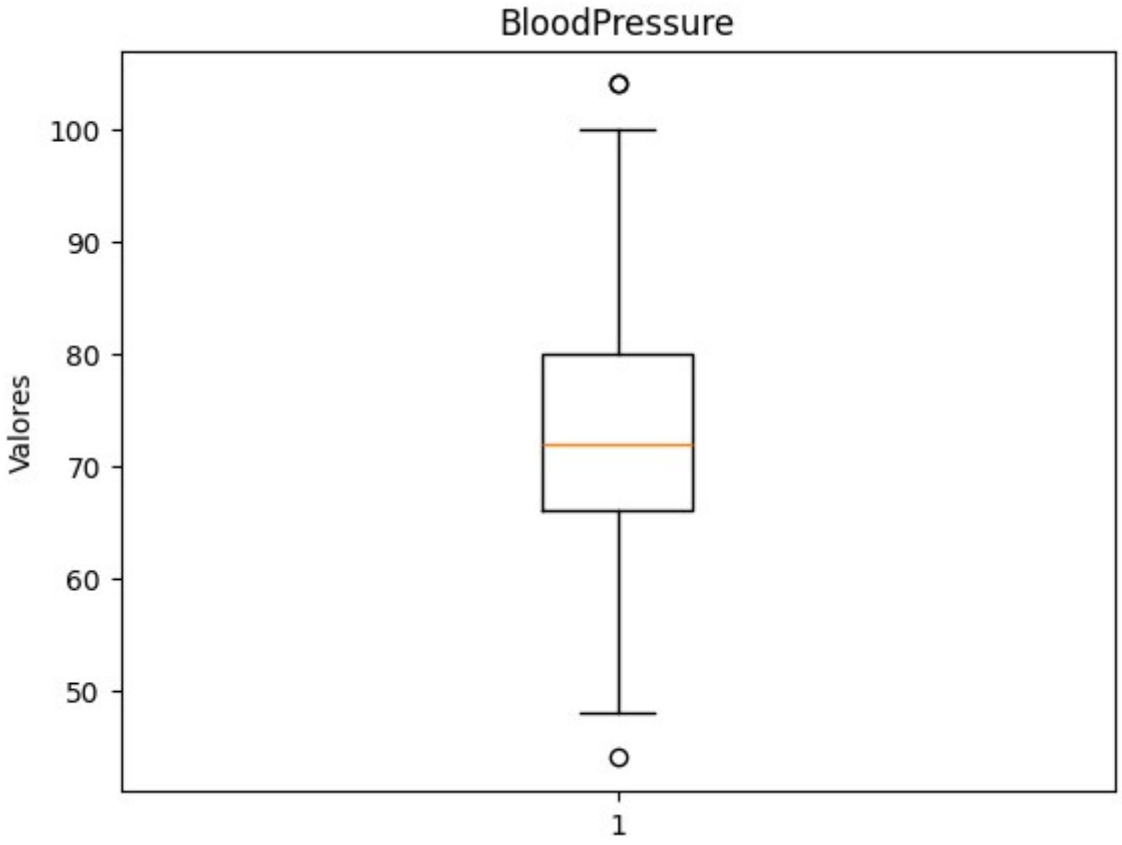
Conclusão sobre a Escala: Os dados possuem uma **dispersão considerável** (amplitude total de 153.0) e apresentam **assimetria positiva (ou à direita)**, indicando que a cauda da distribuição é mais alongada em direção aos valores mais altos (onde também foi detectado o outlier superior).



count	375.000000
mean	120.125333
std	28.786375
min	44.000000
25%	101.000000
50%	115.000000
75%	137.000000
max	197.000000

BloodPressure

Conclusão sobre Escala e Dispersão: A maior parte da amostra (50%) tem valores de pressão arterial entre 66.0 e 80.0. A escala dos dados varia de 44.0 a 104.0, e a distribuição é **aproximadamente simétrica**, mas com uma **leve inclinação/assimetria positiva (à direita)**, além da presença de outliers em ambas as extremidades que aumentam a dispersão total.



count	375.000000
mean	73.066667
std	10.268759
min	44.000000
25%	66.000000
50%	72.000000
75%	80.000000
max	104.000000

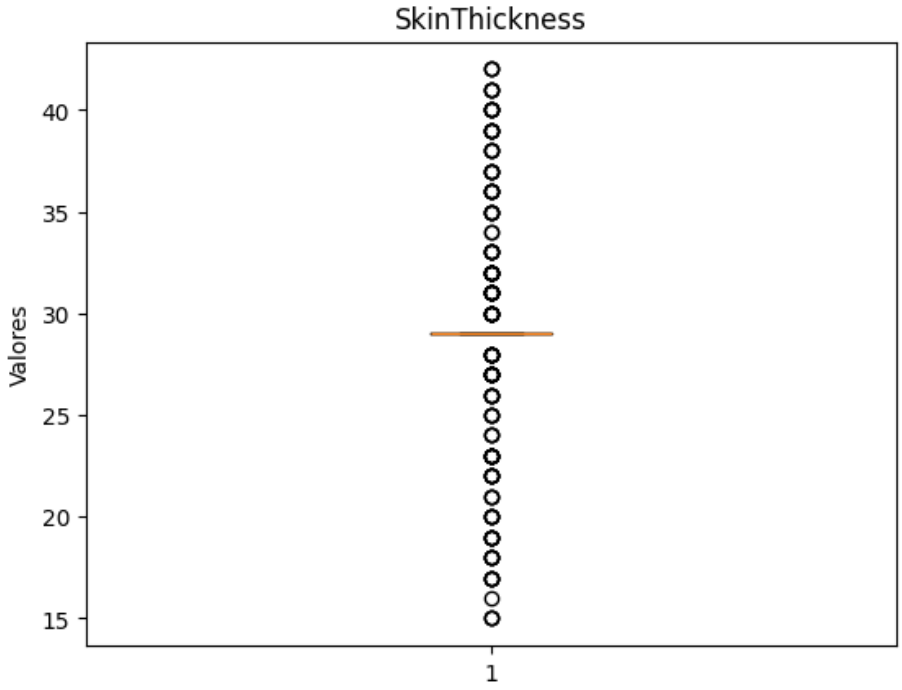
SkinThickness

Conclusão sobre Outliers:

A distribuição possui um **grande número de outliers** ou valores extremos. A concentração de 50% dos dados em 29.0 força os valores mais baixos (como 15.0) e mais altos (como 42.0) a serem classificados como extremos em relação à massa central de dados.

Observação Importante: A distribuição da coluna `SkinThickness` é, portanto, altamente atípica para uma variável contínua, sugerindo que:

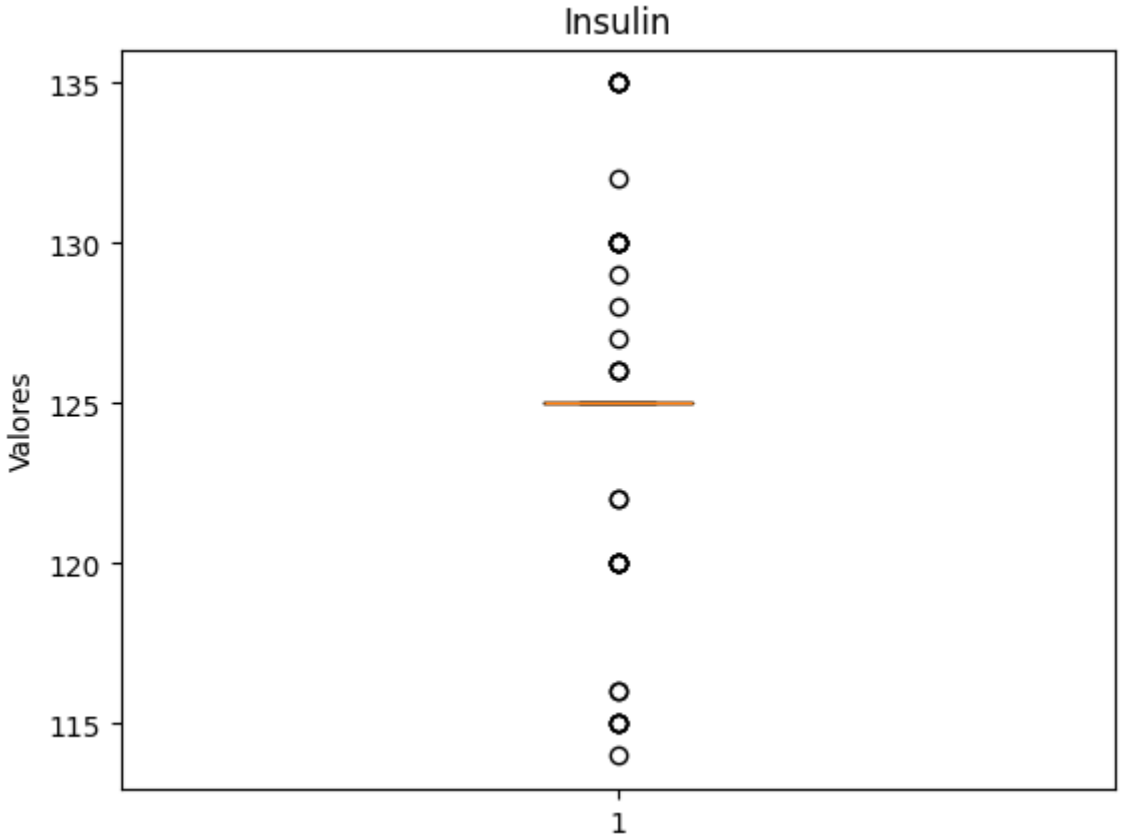
- Pode ser uma variável **discreta** com um forte pico em 29.0.
- Pode haver valores **zero** ou **ausentes** (representados incorretamente por um único valor constante) que precisariam de tratamento específico, mas o `min` de 15.0 sugere que os zeros não são a causa, a menos que os zeros tenham sido previamente filtrados.
- O valor 29.0 pode ser um **valor de substituição** (imputação) que foi usado para preencher muitos dados ausentes (missing values).



count 375.000000
mean 28.976000
std 4.618257
min 15.000000
25% 29.000000
50% 29.000000
75% 29.000000
max 42.000000
Name: SkinThickness, dtype: float64

Insulin

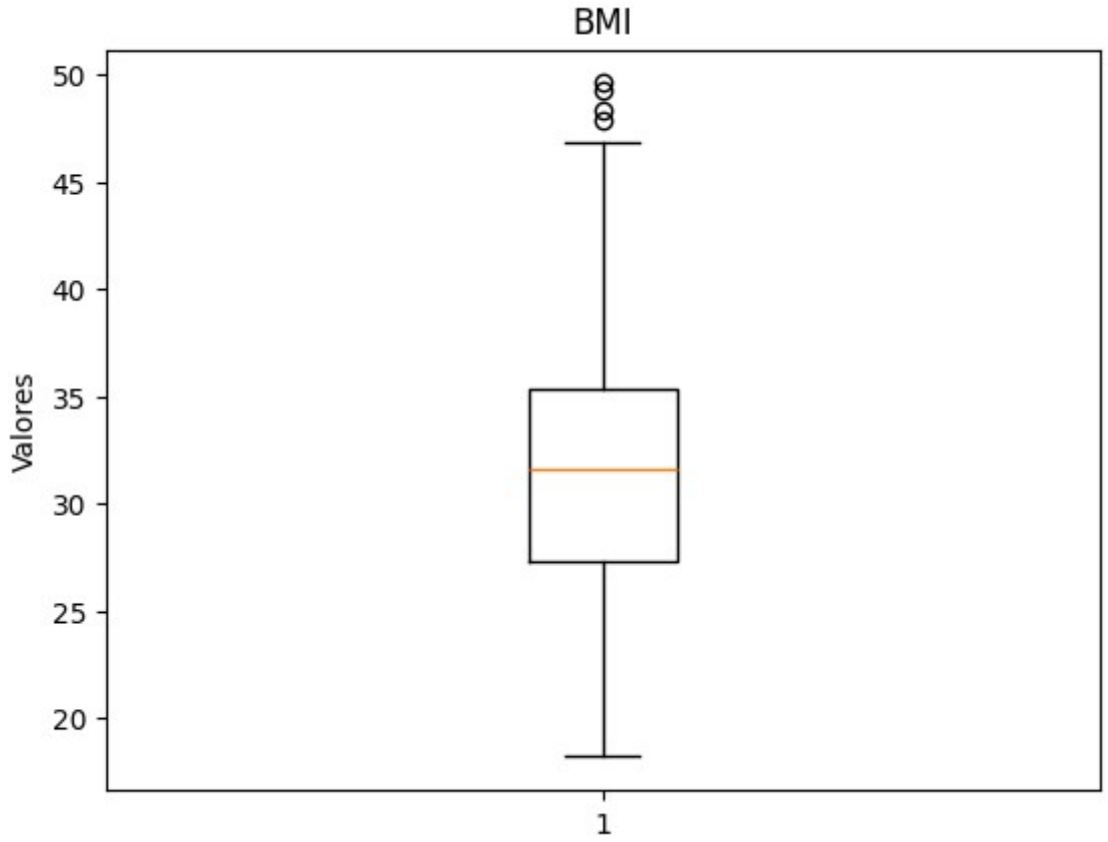
Conclusão sobre Escala: A escala dos dados é estreita, com a maior parte dos valores esmagadoramente concentrada em um único ponto (125.0). A pouca variabilidade que existe é puxada para as extremidades pelos valores de outlier.



count	375.000000
mean	124.984000
std	2.213174
min	114.000000
25%	125.000000
50%	125.000000
75%	125.000000
max	135.000000

BMI

Conclusão para Escala: A escala dos dados varia de 18.20 a 49.60, com uma amplitude total de 31.40. A maior parte dos dados (os 50% centrais) está concentrada em um intervalo de 8.05 unidades (27.30 a 35.35), indicando uma **dispersão moderada** com boa concentração em torno da mediana.

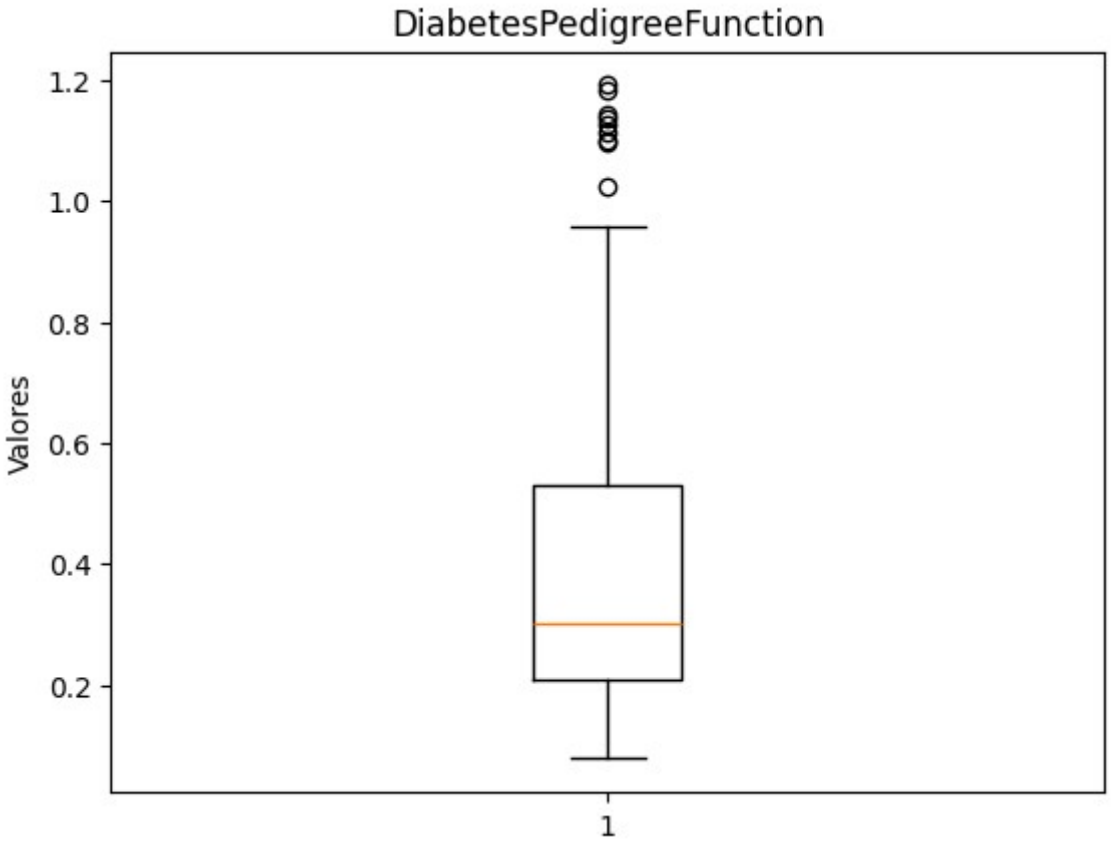


count	375.00000
mean	31.68240
std	6.22331
min	18.20000
25%	27.30000
50%	31.60000
75%	35.35000
max	49.60000

DiabetesPedigreeFunction

Conclusão sobre Outliers:

- O **valor máximo (Max = 1.191000)** é **maior** que o Limite Superior (1.014250).
    - **Conclusão:** Existem **outliers superiores** (valores extremos altos) no conjunto de dados.
- O **valor mínimo (Min = 0.078000)** é **maior** que o Limite Inferior (−0.275750).
    - **Conclusão:** Não existem **outliers inferiores** (valores extremos baixos) no conjunto de dados.

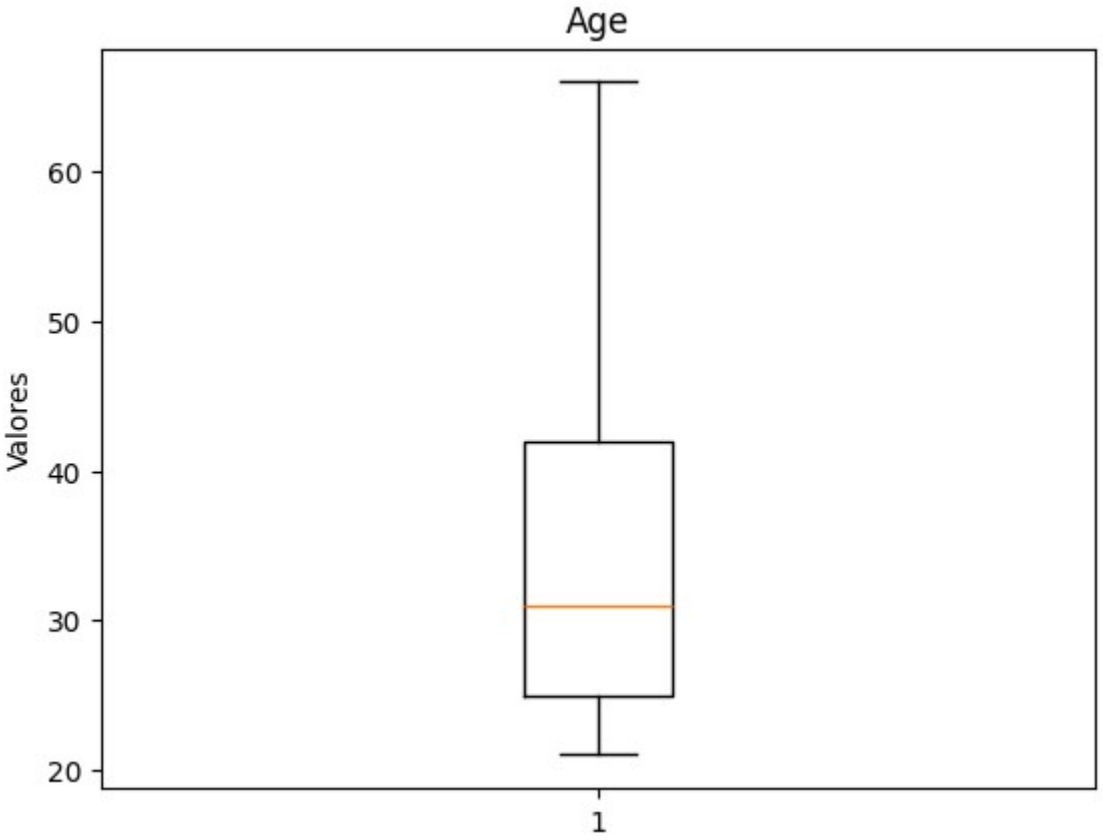


count	375.000000
mean	0.387872
std	0.240621
min	0.078000
25%	0.208000
50%	0.300000
75%	0.530500
max	1.191000

Age

Conclusão Geral

1. **Outliers: Não há *outliers*** (valores atípicos) detectados pela regra de 1.5×IQR. Isso indica que a distribuição dos dados não possui valores extremos que se desviem drasticamente da maioria.
2. **Escala/Dispersão:** Os dados apresentam uma **amplitude total de 45.00**. A **dispersão dos 50% centrais (IQR) é de 17.00**, o que sugere que a maior parte das observações está relativamente concentrada, mas há uma cauda ou dispersão notável nas extremidades, que, no entanto, ainda não são consideradas *outliers*.



count	375.000000
mean	34.530667
std	11.622332
min	21.000000
25%	25.000000
50%	31.000000
75%	42.000000
max	66.000000