Potencializando o desempenho com NoSQL

2 ANÁLISE DOS DADOS DE VENDAS

2.1 Quantidade

Analise os dados na perspectiva da coluna quantidade. Existem outliers nos dados disponibilizados? É possível identificar algo em relação às vendas associadas a estes outliers? Justifique sua resposta. Calcule uma estimativa de variabilidade que ignore o efeito desses outliers.

Sim, é possível encontrar outliers nos dados disponibilizados na perspectiva da coluna quantidade em todos os anos analisados (2019, 2020, 2021 e 2022). Esses outliers representam vendas com quantidades significativamente maiores que a maioria dos dados, o que sugere a ocorrência de eventos atípicos, como grandes compras ou possíveis erros de registro. Esses valores extremos podem influenciar de forma negativa a análise estatística, elevando a média e aumentando a variabilidade.

Importação de Módulos

```
[2]  # gerais
     import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns

[3]  # aplicação
     from sklearn.preprocessing import scale
     from sklearn.covariance import EmpiricalCovariance, MinCovDet

[4]  pip install scipy statsmodels

     Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (1.13.1)
     Requirement already satisfied: statsmodels in /usr/local/lib/python3.10/dist-packages (0.14.2)
     Requirement already satisfied: numpy<2.3,>=1.22.4 in /usr/local/lib/python3.10/dist-packages (from scipy) (1.26.4)
     Requirement already satisfied: pandas!=2.1.0,>=1.4 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (2.1.4)
     Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (0.5.6)
     Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (24.1)
     Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2.8.2)
     Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
     Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
     Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from patsy>=0.5.6->statsmodels) (1.16.0)
```

Figura 1 - Códigos referentes à importação e aplicação de módulos.

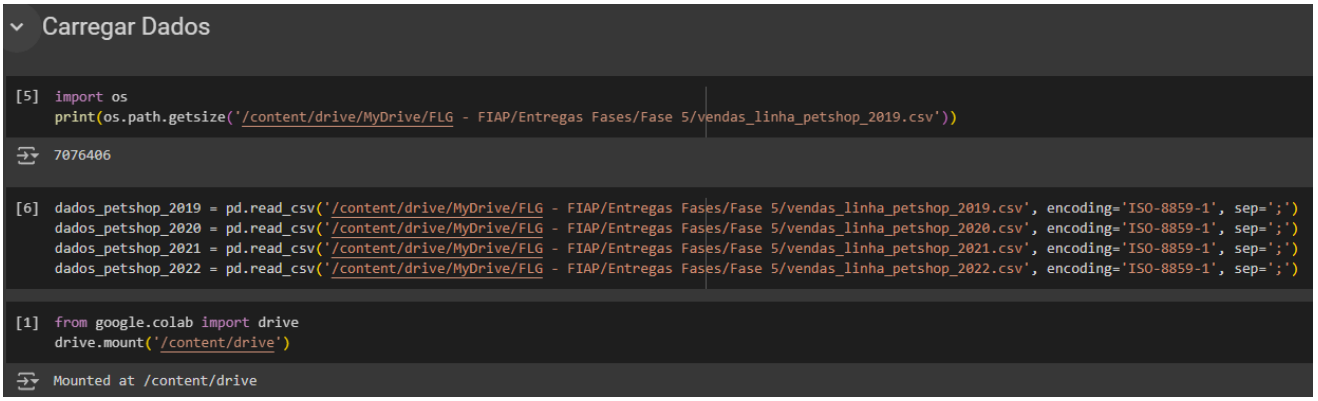

Figura 2 - Códigos para realizar o carregamento dos dados.

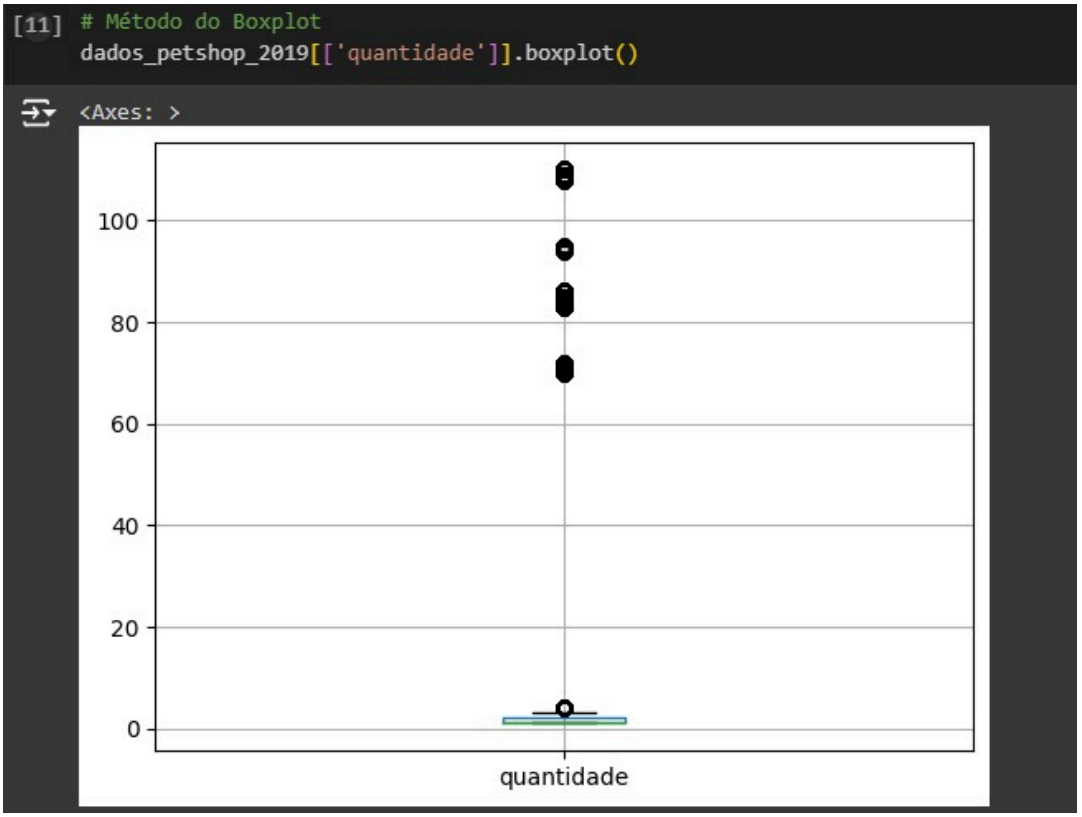

Figura 3 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2019.

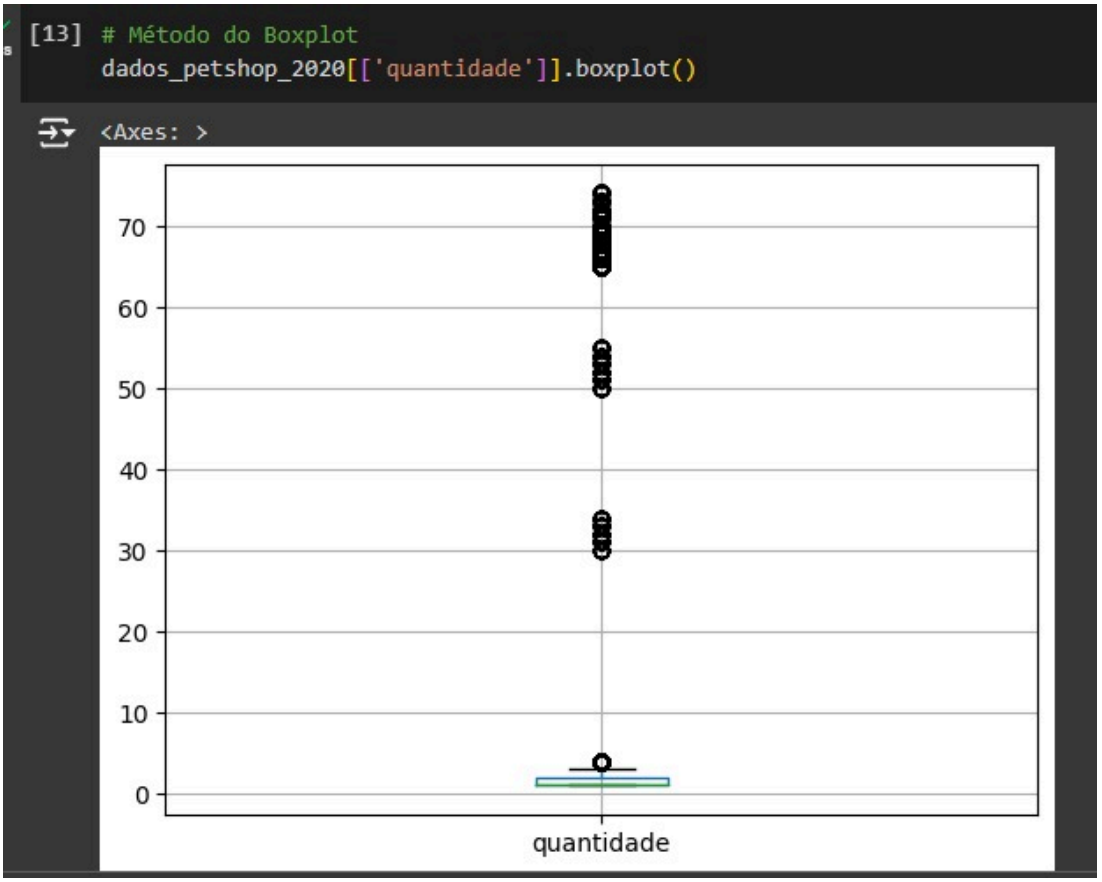

Figura 4 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2020.

Figura 5 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2021.

Figura 6 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2022.

Estimativa de variabilidade que ignore o efeito desses outliers:

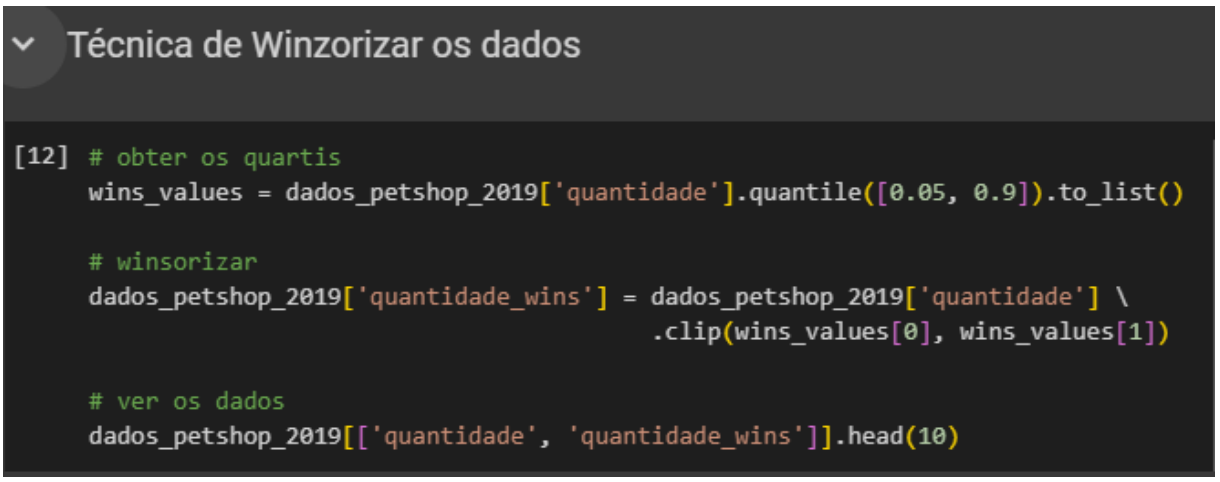

Figura 7 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados_petshop_2019.

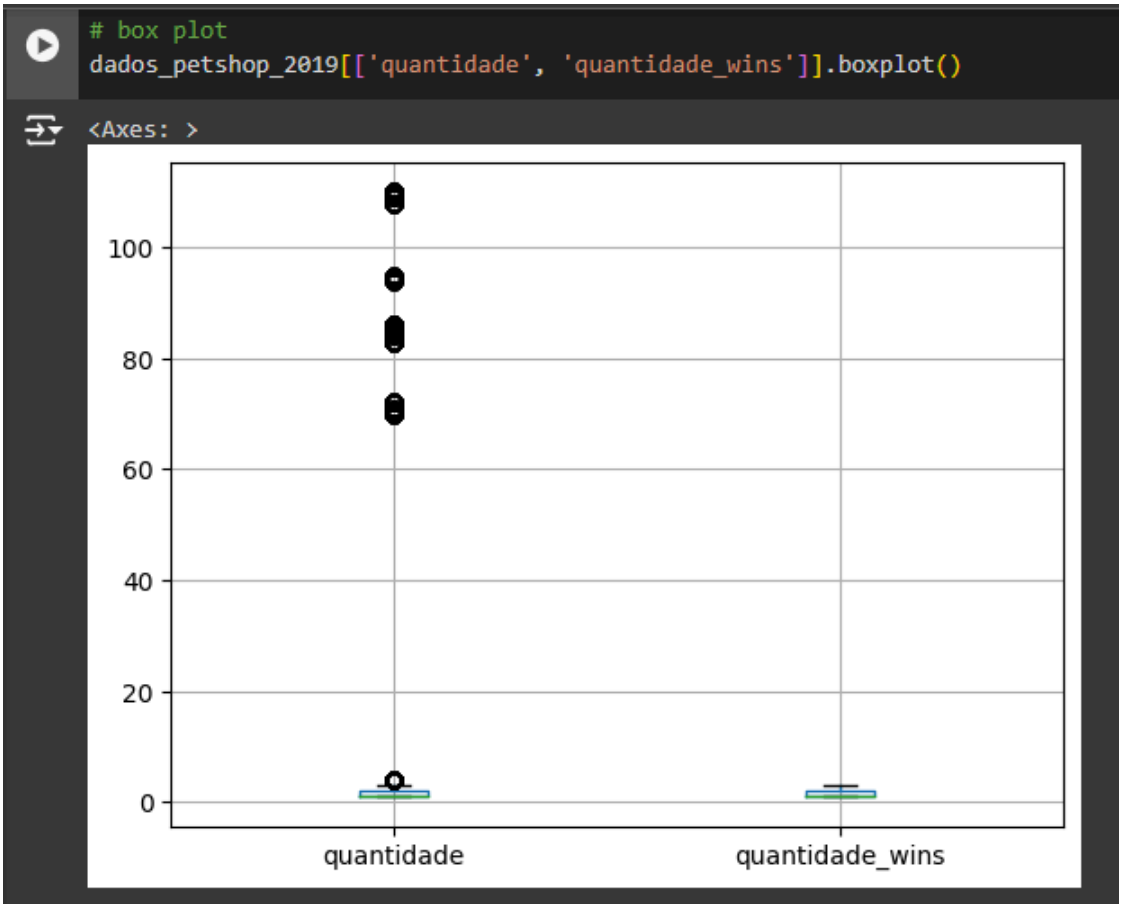

Figura 8 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2019.

```
# obter os quartis
wins_values = dados_petshop_2020['quantidade'].quantile([0.05, 0.9]).to_list()

# winsorizar
dados_petshop_2020['quantidade_wins'] = dados_petshop_2020['quantidade'] \
                                        .clip(wins_values[0], wins_values[1])

# ver os dados
dados_petshop_2020[['quantidade', 'quantidade_wins']].head(10)
```

Figura 9 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados_petshop_2020.

```
15] # box plot
    dados_petshop_2020[['quantidade', 'quantidade_wins']].boxplot()
```

```
<Axes: >
```

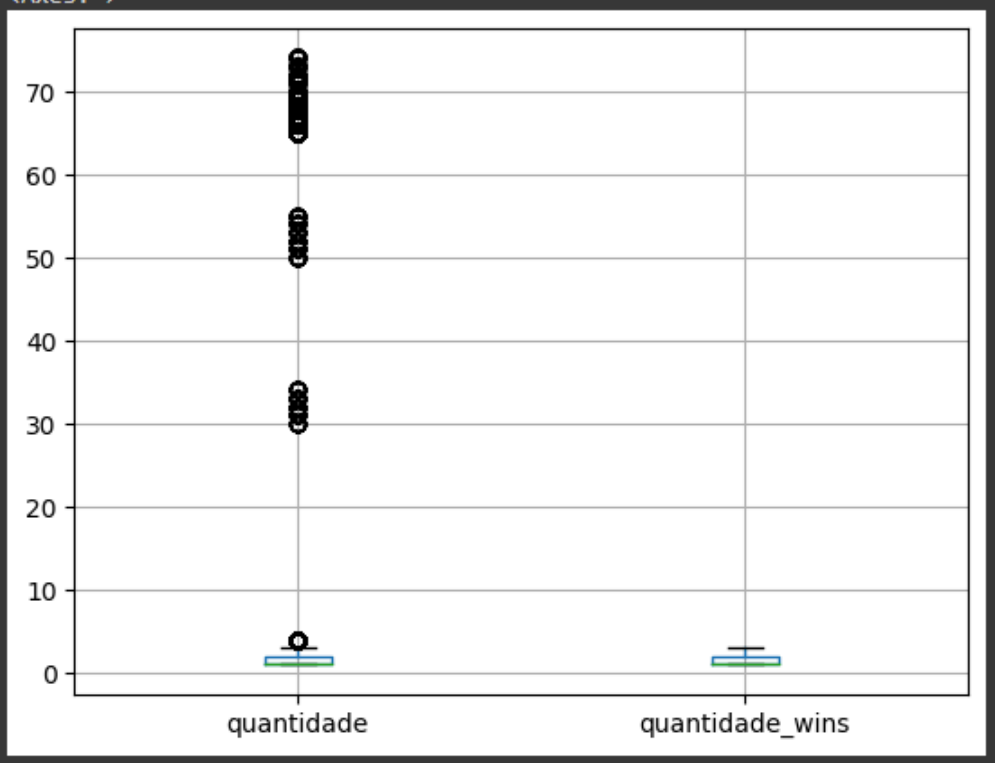

Figura 10 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2020.

```
[16] # obter os quartis
     wins_values = dados_petshop_2021['quantidade'].quantile([0.05, 0.9]).to_list()

     # winsorizar
     dados_petshop_2021['quantidade_wins'] = dados_petshop_2021['quantidade'] \
                                             .clip(wins_values[0], wins_values[1])

     # ver os dados
     dados_petshop_2021[['quantidade', 'quantidade_wins']].head(10)
```

Figura 11 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados_petshop_2021.

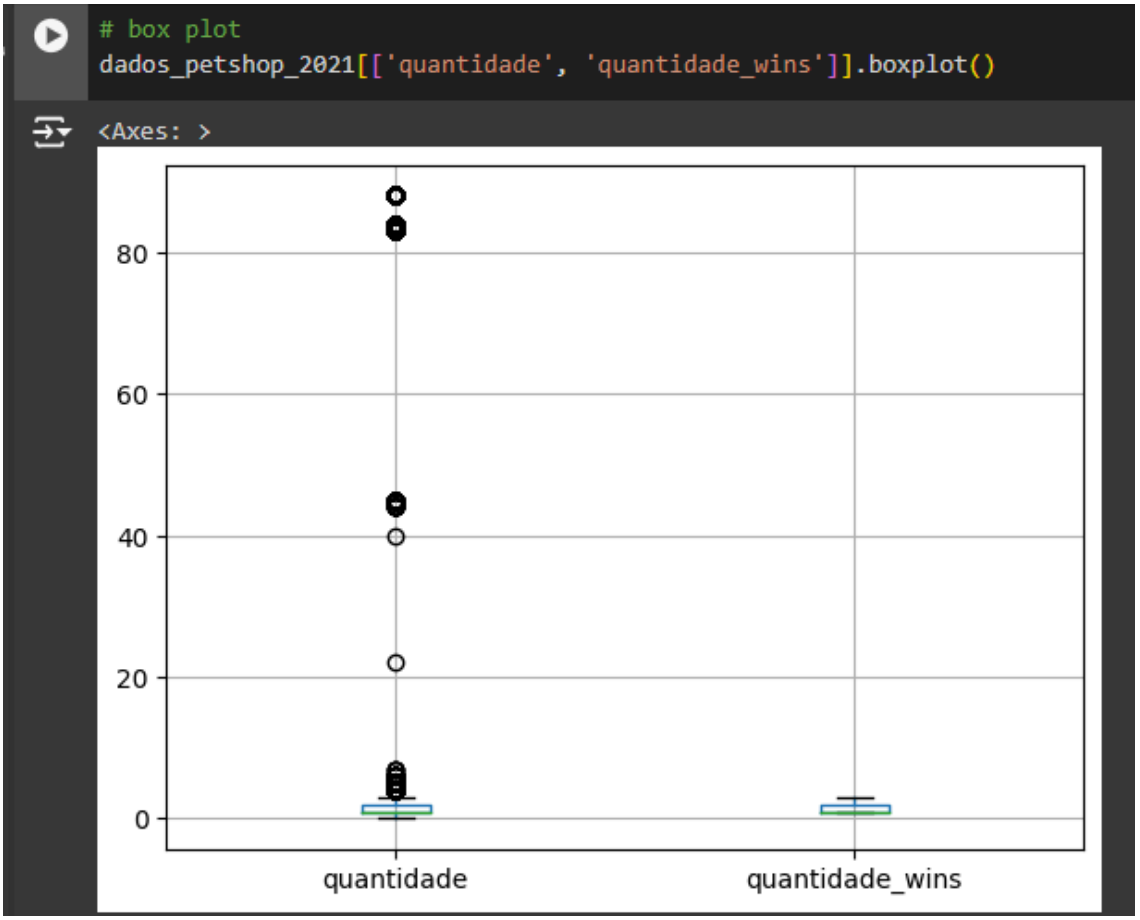

Figura 12 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2021.

```
# obter os quartis
wins_values = dados_petshop_2022['quantidade'].quantile([0.05, 0.9]).to_list()

# winsorizar
dados_petshop_2022['quantidade_wins'] = dados_petshop_2022['quantidade'] \
                                        .clip(wins_values[0], wins_values[1])

# ver os dados
dados_petshop_2022[['quantidade', 'quantidade_wins']].head(10)
```

Figura 13 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados_petshop_2022.

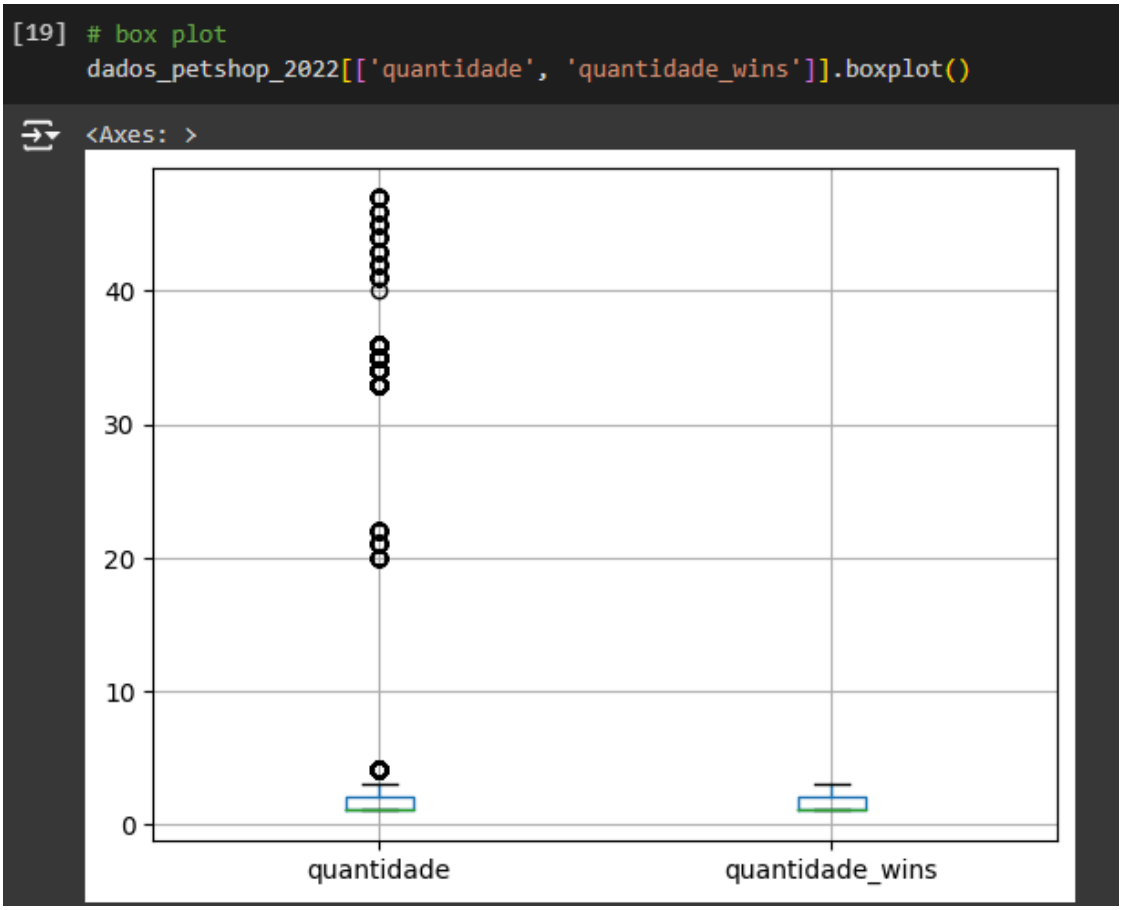

Figura 14 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2022.

2.2 Preço

Em relação à média de preço, há diferença estatisticamente significativa entre a média de preço de alguma região e a média da população? E em relação à média de preço de alguma modalidade de pagamento e à média da população? Justifique a hipótese.

Conforme será possível visualizar na análise das regiões (Figura 16), todos os p-valores são bem maiores que 0,05 (5%), indicando que **não há diferença** estatisticamente significativa entre a média de preço de qualquer região e a média da população. Concluindo, então, que as médias de preço das regiões não diferem da média geral.

Na análise das modalidades de pagamento (Figura 17), os p-valores para a maioria das modalidades também estão muito acima de 0,05, indicando que **não há diferença** estatisticamente significativa entre a média de preço de alguma modalidade de pagamento e a média da população.

Questão 2 - Teste de Hipóteses

```
[ ] # Carregar os dados em DataFrames separados
    df_vendas_petshop_2019 = pd.read_csv('/content/drive/MyDrive/FLG - FIAP/Entregas Fases/Fase 5/vendas_linha_petshop_2019.csv', encoding='ISO-8859-1', se
    df_vendas_petshop_2020 = pd.read_csv('/content/drive/MyDrive/FLG - FIAP/Entregas Fases/Fase 5/vendas_linha_petshop_2020.csv', encoding='ISO-8859-1', se
    df_vendas_petshop_2021 = pd.read_csv('/content/drive/MyDrive/FLG - FIAP/Entregas Fases/Fase 5/vendas_linha_petshop_2021.csv', encoding='ISO-8859-1', se
    df_vendas_petshop_2022 = pd.read_csv('/content/drive/MyDrive/FLG - FIAP/Entregas Fases/Fase 5/vendas_linha_petshop_2022.csv', encoding='ISO-8859-1', se

    # Concatenar todos os DataFrames em um único DataFrame
    df = pd.concat([df_vendas_petshop_2019, df_vendas_petshop_2020, df_vendas_petshop_2021, df_vendas_petshop_2022])

    # Converter a coluna 'valor' para float
    df['valor'] = df['valor'].str.replace(',', '.').astype(float)

    # Remover os missings
    df = df.dropna()

    # Média da população
    media_populacao = df['valor'].mean()

    # Média por região
    media_por_regiao = df.groupby('regiao_pais')['valor'].mean()

    # Média por modalidade de pagamento
    media_por_modalidade = df.groupby('formapagto')['valor'].mean()
```

Figura 15 - Código que carrega e combina dados de vendas dos anos de 2019, 2020, 2021 e 2022 em um único DataFrame e converte a coluna de valores string para float, além de remover todos os missings.

Para Regiões:

```
[24] from scipy import stats

     resultados_regiao = {}
     for regiao in df['regiao_pais'].unique():
         # Filtrar os preços dessa região
         precos_regiao = df[df['regiao_pais'] == regiao]['valor']

         # Teste t de Student
         t_stat, p_val = stats.ttest_1samp(precos_regiao, media_populacao)

         # Armazenar o resultado
         resultados_regiao[regiao] = {'t_stat': t_stat, 'p_val': p_val}

     # Mostrar resultados
     for regiao, resultado in resultados_regiao.items():
         print(f"Região: {regiao}, t-stat: {resultado['t_stat']:.4f}, p-valor: {resultado['p_val']:.4f}")

Região: Norte, t-stat: -0.0205, p-valor: 0.9836
Região: Centro Oeste, t-stat: 0.1838, p-valor: 0.8542
Região: Nordeste, t-stat: 0.0495, p-valor: 0.9606
Região: Sudeste, t-stat: -0.0602, p-valor: 0.9520
Região: Sul, t-stat: -0.1733, p-valor: 0.8624
```

Figura 16 - O p-valor das regiões está bastante acima de 0,05 (5%), portanto, **não há** diferença estatisticamente significativa entre a média de preço de alguma região e a média da população.

Para Modalidade de Pagamento:

```
[25] resultados_modalidade = {}
     for modalidade in df['formapagto'].unique():
         # Filtrar os preços dessa modalidade de pagamento
         precos_modalidade = df[df['formapagto'] == modalidade]['valor']

         # Teste t de Student
         t_stat, p_val = stats.ttest_1samp(precos_modalidade, media_populacao)

         # Armazenar o resultado
         resultados_modalidade[modalidade] = {'t_stat': t_stat, 'p_val': p_val}

     # Mostrar resultados
     for modalidade, resultado in resultados_modalidade.items():
         print(f"Modalidade: {modalidade}, t-stat: {resultado['t_stat']:.4f}, p-valor: {resultado['p_val']:.4f}")
```

```
Modalidade: Cartão Crédito, t-stat: 1.8377, p-valor: 0.0661
Modalidade: Dinheiro, t-stat: -0.0440, p-valor: 0.9649
Modalidade: Pix, t-stat: 0.1269, p-valor: 0.8990
Modalidade: Boleto Bancário, t-stat: -0.4802, p-valor: 0.6311
Modalidade: Cartão Débito, t-stat: -1.5127, p-valor: 0.1304
```

Figura 17 - O p-valor está bastante acima de 0,05 (5%) na maioria das modalidades. Na modalidade Cartão Crédito está apenas um pouco acima, mas ainda **não há** diferença estatisticamente significativa entre a média de preço de alguma modalidade de pagamento e a média da população.

2.3 Correlações

Calcule a matriz de correlação dos dados fornecidos. Quais as variáveis que apresentam forte correlação positiva ou negativa? Acrescente a matriz de correlação como uma imagem e anexe-a ao seu relatório.

É possível observar que as variáveis valor e quantidade apresentam **forte correlação positiva**, por estarem todas iguais a escala: 1. Isso indica que,

proporcionalmente, conforme o valor aumenta, a quantidade também tende a aumentar.

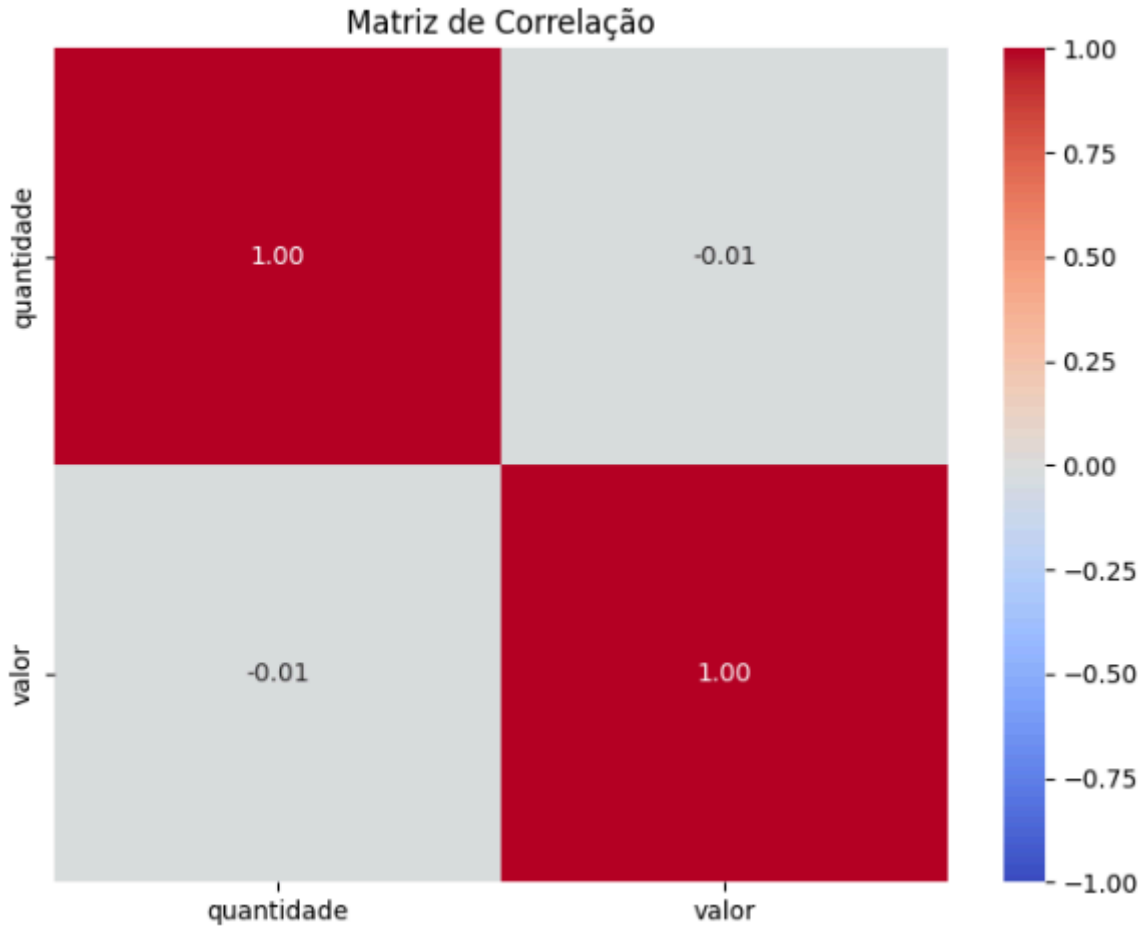

Figura 18 - Matriz de correlação.