

**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.

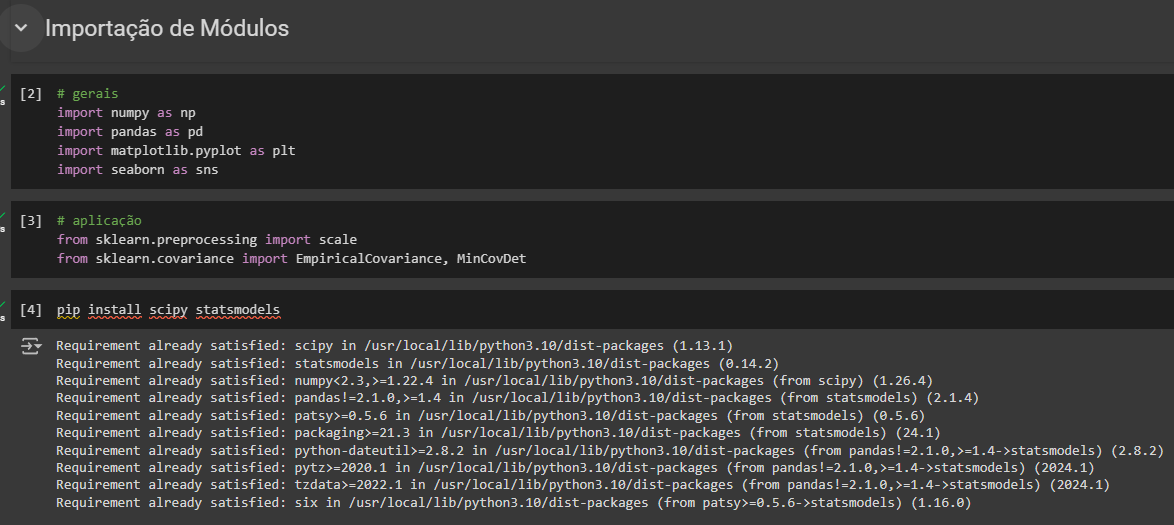


Figura 1 - Códigos referentes à importação e aplicação de módulos.

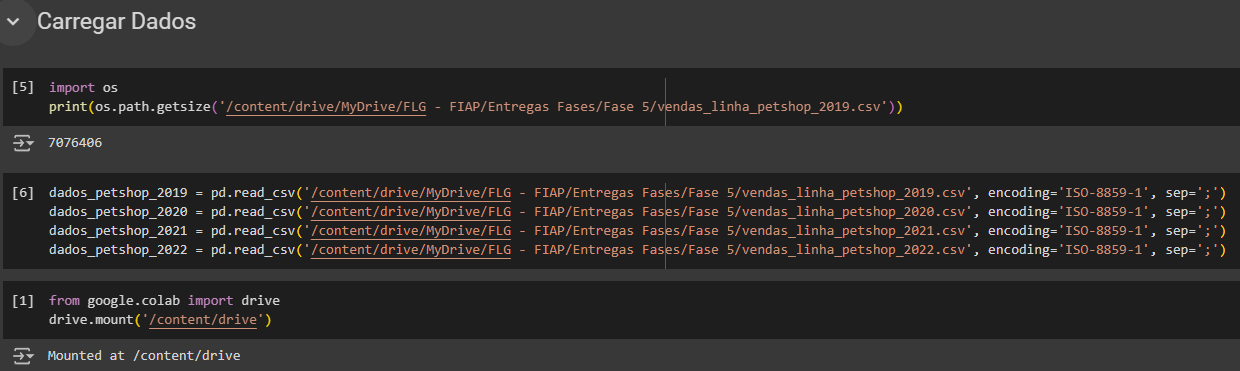


Figura 2 - Códigos para realizar o carregamento dos dados.

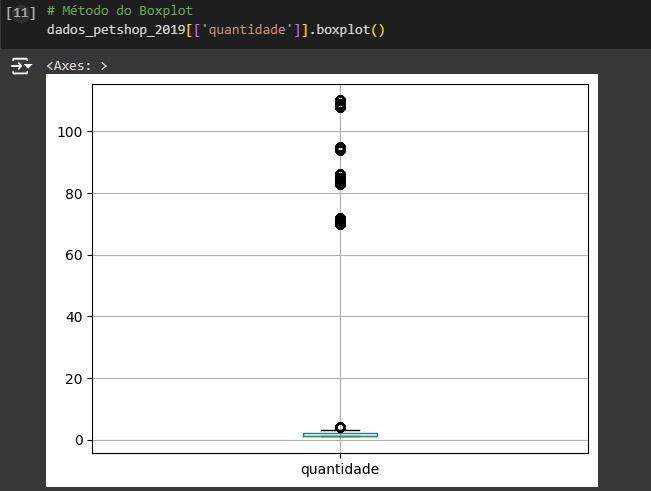


Figura 3 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2019.

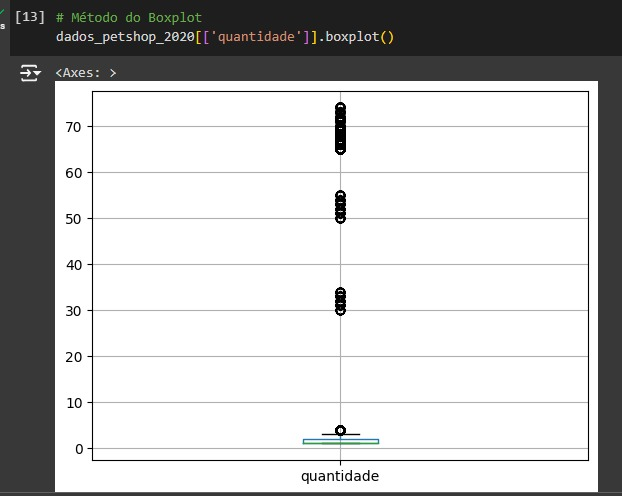


Figura 4 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2020.

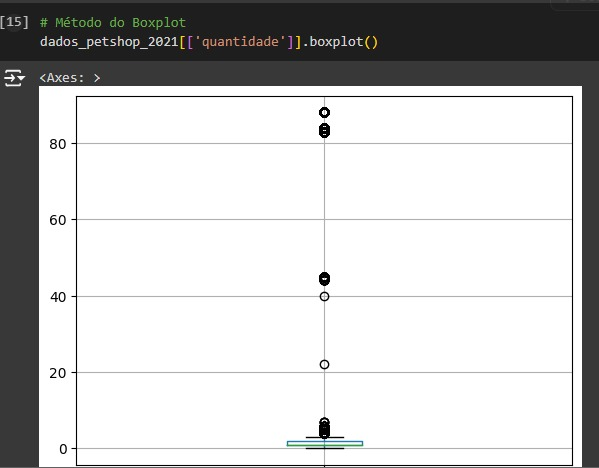


Figura 5 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2021.

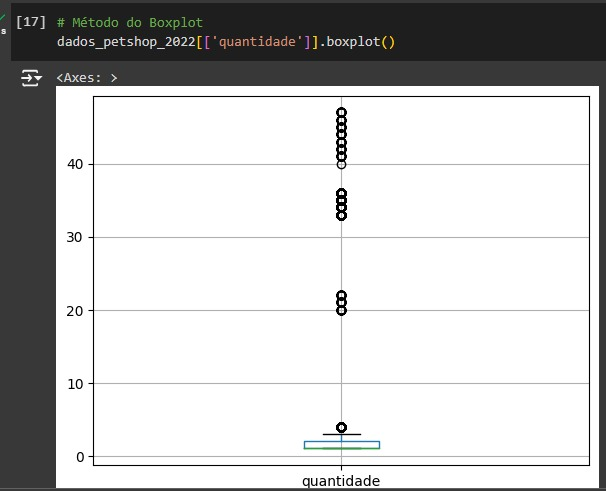


Figura 6 - Candidatos a outliers encontrados no boxplot referente aos dados de vendas de 2022.

**Estimativa de variabilidade que ignore o efeito desses outliers:**

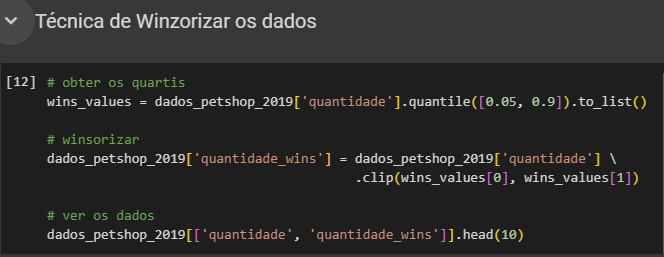


Figura 7 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados\_petshop\_2019.

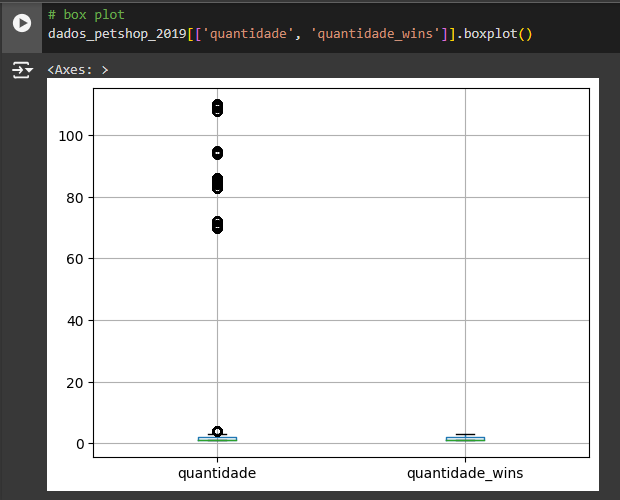


Figura 8 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2019.

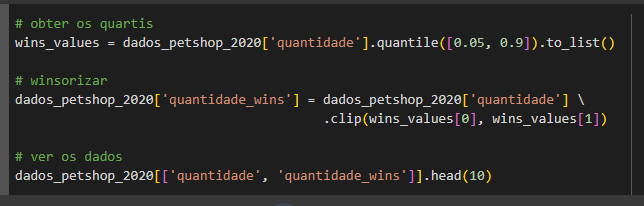


Figura 9 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados\_petshop\_2020.

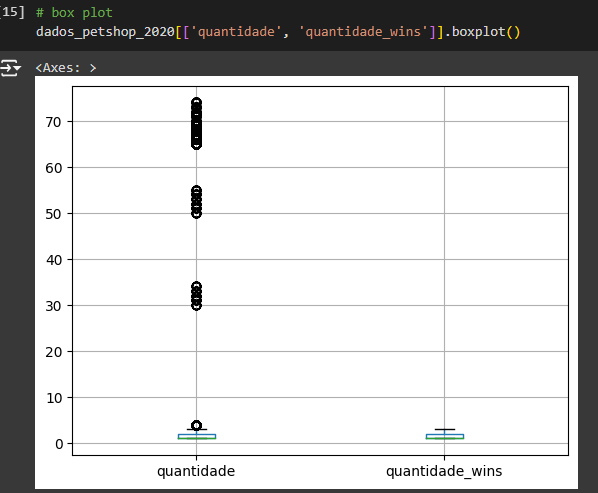


Figura 10 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2020.

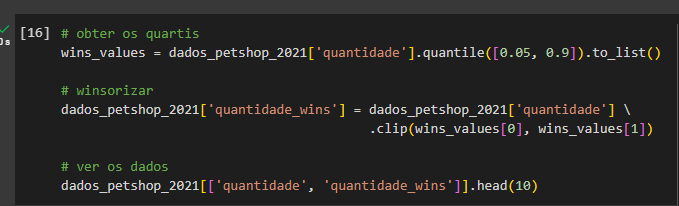


Figura 11 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados\_petshop\_2021.

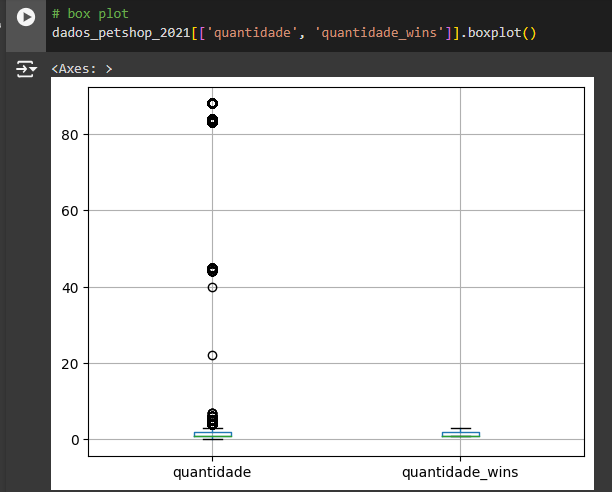


Figura 12 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2021.

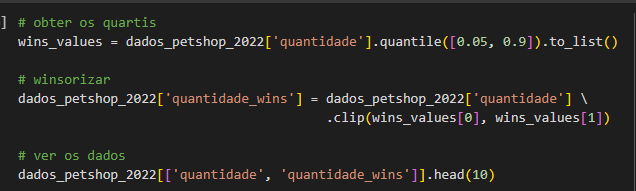


Figura 13 - Códigos com a aplicação da técnica de Winsorização na coluna ‘quantidade’, com a intenção de limitar valores extremos no arquivo denominado dados\_petshop\_2022.

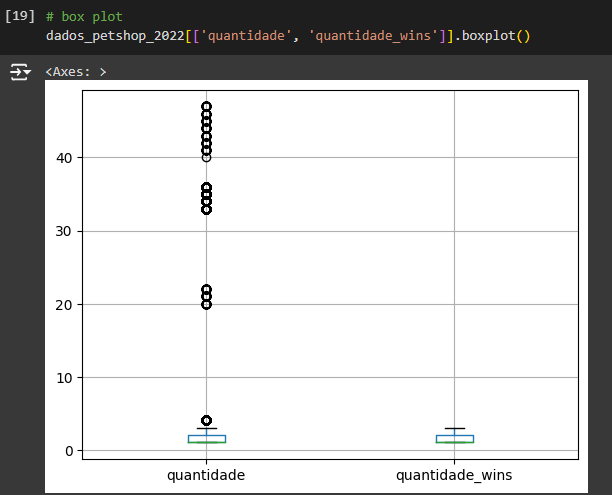


Figura 14 - Estimativa de variabilidade de 0 e 1, elimina o efeito dos candidatos a outliers da venda de 2022.

## 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.

## 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.

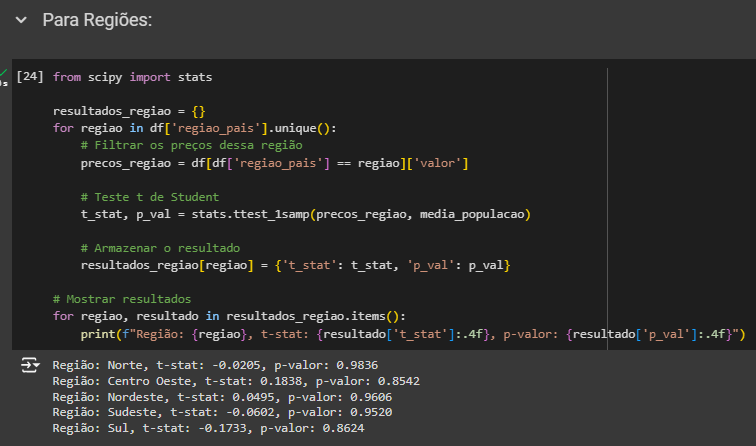


Figura 16 - O p-valor das regiões está bastante acima de 0,05 (5%), portanto, **não há** diferença estatisticamente significativa entre a média de preço de alguma região e a média da população.

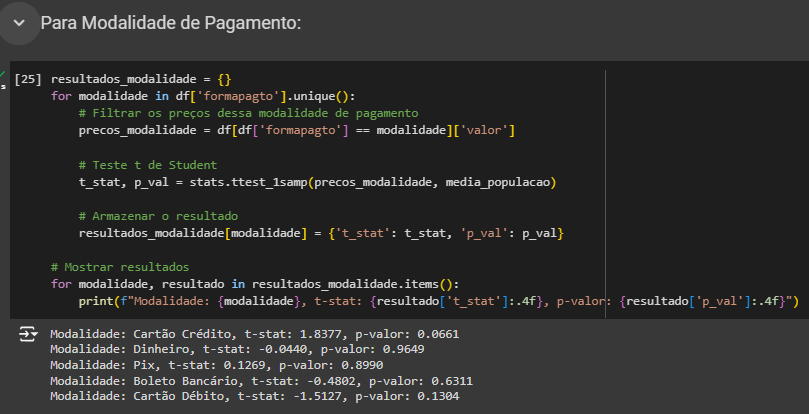


Figura 17 - O p-valor está bastante acima de 0,05 (5%) na maioria das modalidades. Na modalidade Cartão Crédito está apenas um pouco acima, mas ainda **não** **há** diferença estatisticamente significativa entre a média de preço de alguma modalidade de pagamento e a média da população.

## 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.

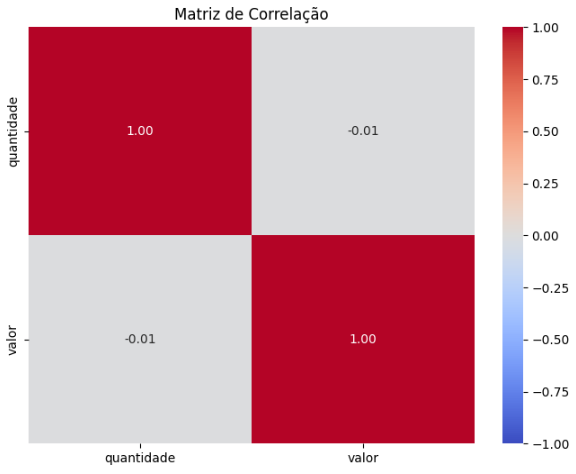
****

Figura 18 - Matriz de correlação.