Relatório Técnico: Justificativa para o Uso do Método IQR na Análise de Outliers

1. Introdução

A análise de dados financeiros, especialmente em grandes projetos, envolve frequentemente a identificação de valores discrepantes (outliers) que podem distorcer a interpretação de métricas importantes, como custos e atualizações monetárias. Neste laudo, justificaremos por que o **IQR (Intervalo Interquartil)** é o melhor método estatístico para detectar e analisar outliers nos dados fornecidos, em comparação com outras abordagens como **média**, **desvio padrão** e **regressão linear**.

Também serão apresentadas situações hipotéticas para ilustrar o impacto de cada método na análise, destacando a robustez e eficiência do IQR em contextos com dados financeiros.

2. Explicação dos Métodos Estatísticos

Nesta seção, descrevemos os principais métodos usados para análise estatística e como cada uma lida com outliers e variação de dados:

2.1. Média

A média aritmética é uma medida de tendência central que calcula a soma de todos os valores dividida pelo número total de observações. A fórmula básica é:

$$\text{Média} = \frac{\sum \text{Valores}}{n}$$

A média é amplamente usada, mas apresenta uma limitação significativa: ela é muito sensível a outliers. Se houver um valor extremamente alto ou baixo no conjunto de dados, ele pode distorcer a média e fazer com que a medida central não represente adequadamente a maioria dos dados.

2.2. Desvio Padrão

O desvio padrão mede a dispersão dos dados em torno da média. É útil para entender o quanto os valores se afastam da média, mas, assim como a média, o desvio padrão é suscetível a outliers. Em dados com variações extremas, ele pode inflacionar a percepção da variação. A fórmula do desvio padrão é:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Onde x_i são os valores, μ é a média, e n é o número total de observações.

2.3. Regressão Linear

A regressão linear é uma técnica usada para modelar a relação entre uma variável dependente (alvo) e uma ou mais variáveis independentes (preditoras). A equação da reta de regressão é:

$$y = \beta_0 + \beta_1 x$$

Onde y é o valor previsto, β é o intercepto, β é o coeficiente angular, e x é o valor da variável independente. A regressão linear é útil para prever tendências, mas não é adequada para identificar outliers diretamente, já que a presença de valores extremos pode distorcer os coeficientes, resultando em previsões imprecisas.

2.4. IQR (Intervalo Interquartil)

O IQR (Intervalo Interquartil) é uma medida robusta da dispersão central que considera a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), abrangendo os 50% centrais dos dados. A fórmula do IQR é:

$$IQR = Q3 - Q1$$

Os outliers são identificados ao calcular os limites inferior e superior:

$$LimiteInferior = Q1 - 1.5 \times IQR$$

$$LimiteSuperior = Q3 + 1.5 \times IQR$$

Qualquer valor fora desses limites é considerado um Outliers. O IQR é uma excelente escolha para detectar outliers porque não é influenciado por valores extremos e, portanto, reflete melhor a dispersão central dos dados.

2.5. Outras Técnicas de Detecção de Outliers

Além dos métodos discutidos anteriormente, existem outras técnicas amplamente utilizadas para a identificação de outliers, cada uma com características específicas que podem ser aplicadas em contextos diversos. Abaixo, descrevemos brevemente algumas dessas técnicas:

2.5.1. Métodos Baseados em Distância:

Distância Euclidiana e Mahalanobis: Essas técnicas calculam a distância de cada ponto em relação à média dos dados ou em relação a uma distribuição normal multivariada (no caso da Mahalanobis). São úteis para detectar outliers em conjuntos de dados com variáveis correlacionadas, mas podem ser sensíveis à escala e distribuição dos dados.

2.5.2. Modelos de Clusterização:
DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Este método identifica outliers como pontos que não pertencem a clusters densamente agrupados. É particularmente eficaz em dados com clusters naturais e onde os outliers estão em áreas de baixa densidade.

2.5.3. Análise de Componentes Principais (PCA):
A PCA transforma o conjunto de dados em componentes principais, concentrando a variabilidade em poucas dimensões. Outliers podem ser identificados como pontos que se afastam excessivamente ao longo dos principais componentes, sendo útil para dados multidimensionais.

2.5.4. Métodos Baseados em Máquinas de Aprendizado (Machine Learning):
Isolation Forest: Este método constrói várias árvores de decisão aleatórias e considera outliers os pontos isolados rapidamente. É eficiente para grandes conjuntos de dados e lida bem com distribuições desconhecidas.
Autoencoders (Redes Neurais): Modelos autoencoders tentam reconstruir os dados com base em uma arquitetura de rede neural. Outliers são identificados como pontos com erro de reconstrução alto. Esse método é útil em contextos com padrões complexos e dados não lineares.

3. Comparação Detalhada dos Métodos Utilizando Situações Hipotéticas

Situação 1: Dados sem Outliers

Considere um conjunto de dados sem outliers:

Valor Líquido Referência	Valor Atualizado
1.000.000	1.200.000
1.500.000	1.700.000
2.000.000	2.300.000
3.000.000	2.900.000
4.000.000	4.100.000

Média:
A média para o Valor Atualizado seria:

$$\text{Média} = \frac{1.200.000 + 1.700.000 + 2.300.000 + 2.900.000 + 4.100.000}{5} = 2.440.000$$

Neste caso, a média reflete bem o "centro" dos dados, já que não há outliers.

Desvio Padrão:
O desvio padrão, neste caso, seria relativamente baixo, mostrando que a variação dos dados é limitada.

Regressão Linear:
A regressão linear ajustaria bem os dados, com uma relação linear clara entre o Valor Líquido Referência e o Valor Atualizado.

IQR:
Calculamos os quartis:

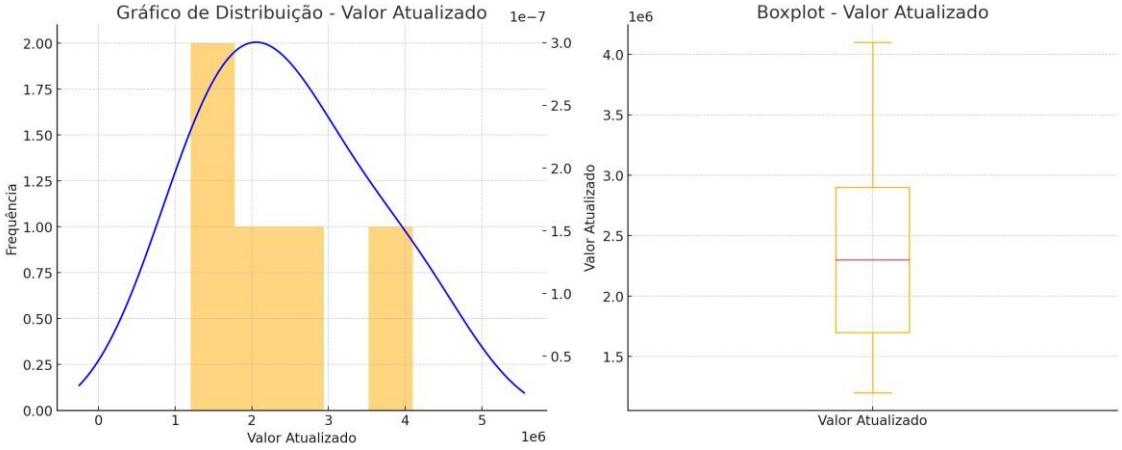
Q1: 1.700.000
Q3: 2.900.000
IQR: 1.200.000

Os limites seriam:

- **Limite Inferior:** Q1 - 1.5 x IQR = 1.700.000 - 1.5 x 1.200.000 = -100.000
- **Limite Superior:** Q3 + 1.5 x IQR = 2.900.000 + 1.5 x 1.200.000 = 4.700.000

Não há valores fora desse intervalo, logo, nenhum Outliers é detectado.

Gráfico de Distribuição e Boxplot



O gráfico de distribuição abaixo mostra a frequência dos valores atualizados, com uma curva de densidade sobreposta, ilustrando a simetria do conjunto de dados.

O boxplot apresenta a distribuição dos valores atualizados com a mediana bem centrada e sem outliers, visualizando claramente a dispersão dos dados.

Conclusão:
Consideramos um conjunto de dados sem outliers. Neste caso, todos os métodos (média, desvio padrão, regressão linear e IQR) apresentam resultados consistentes e adequados, pois os dados são simétricos e não há valores extremos.

Situação 2: Dados com Outliers

Agora, vamos considerar um conjunto de dados com um Outlier:

Valor Líquido Referência	Valor Atualizado
1.000.000	1.200.000
1.500.000	1.700.000
2.000.000	2.300.000
3.000.000	2.900.000
100.000.000	120.000.000

Neste caso, o valor **100.000.000** no **Valor Líquido Referência** e **120.000.000** no **Valor Atualizado** são outliers.

Média:
A média para o **Valor Atualizado** seria:

$$\text{Média} = \frac{1.200.000 + 1.700.000 + 2.300.000 + 2.900.000 + 120.000.000}{5} = 25.220.000$$

Problema: A média é fortemente distorcida pelo outlier, e não representa adequadamente os dados restantes. A maioria dos valores está entre 1.200.000 e 2.900.000, mas a média sugere um valor muito maior devido ao impacto do outlier.

Desvio Padrão:
O desvio padrão seria elevado, indicando uma dispersão maior do que realmente existe na maioria dos dados, já que o outlier influencia diretamente esse cálculo.

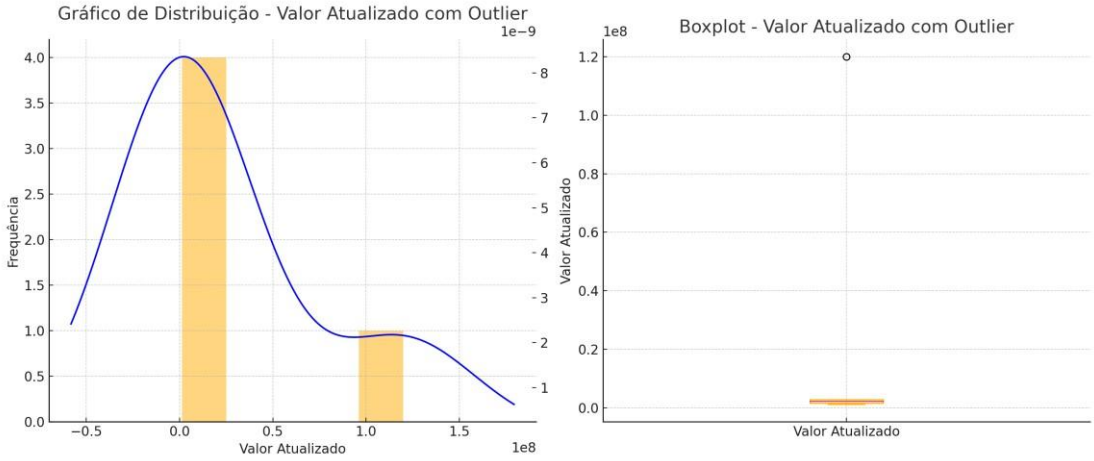
Regressão Linear:
A regressão linear seria distorto pelo outlier, inclinando a reta de regressão de forma que os valores normais fiquem subestimados, enquanto o outlier seria sobrestimado. O modelo de regressão linear não lida bem com outliers, resultando em previsões imprecisas para a maioria dos dados.

IQR:
Para o IQR, calculamos os quartis com os dados ordenados:

Q1: 1.700.000
Q3: 2.900.000
IQR: 1.200.000
Os limites seriam:

- **Limite Inferior:** Q1 - 1.5 x IQR = 1.700.000 **-** 1.5 x 1.200.000 = -100.000
- **Limite Superior:** Q3 + 1.5 x IQR = 2.900.000 + 1.5 x 1.200.000 = 4.700.000

Gráfico de Distribuição e Boxplot para Valores Atualizados com Outlier:



O gráfico de distribuição à esquerda mostra a frequência dos valores atualizados, com uma curva de densidade sobreposta. Observa-se que o outlier de alto valor distorce a distribuição, deslocando a curva e criando uma cauda longa para a direita.
O boxplot à direita ilustra a presença de um outlier significativo, destacado como um ponto fora da faixa principal dos dados. Esse valor extremo indica uma dispersão muito além dos demais valores, evidenciando a robustez do método IQR em identificar outliers.

Conclusão
Com isso, o valor 120.000.000 é claramente identificado como um outlier pelo IQR, uma vez que está muito acima do limite superior de 4.700.000. Portanto, o IQR oferece uma maneira eficaz de detectar esse valor extremo, enquanto os outros métodos, como média e desvio padrão, são impactados negativamente.

4. Comparação Prática dos Métodos
Após essas situações hipotéticas, podemos fazer um resumo prático sobre os métodos:

4.1 Média
A média é útil quando os dados são simétricos e não possuem outliers. No entanto, quando os dados apresentam valores extremos, como na Situação 2, a média perde sua capacidade de representar o centro dos dados com precisão. Ela é influenciada pelos outliers, fazendo com que os resultados fiquem distorcidos.

4.2 Desvio Padrão

O desvio padrão, assim como a média, é sensível a outliers. Em dados com valores extremos, o desvio padrão se torna inflacionado, fazendo com que pareça haver uma maior variação nos dados do que realmente existe. Isso pode resultar em uma percepção errada da dispersão dos dados, especialmente em análises financeiras.

4.3 Regressão Linear

A regressão linear, embora útil para modelar relações entre variáveis, não é adequada para detectar outliers. Valores extremos, como visto na Situação 2, distorcem a linha de regressão, afetando negativamente a capacidade do modelo de fazer previsões precisas para a maioria dos dados. Quando há outliers presentes, a regressão pode subestimar ou superestimar as previsões, dependendo da magnitude do outlier.

4.4 IQR (Intervalo Interquartil)

O IQR se destaca como o método mais robusto e eficaz para a detecção de outliers, especialmente em conjuntos de dados assimétricos ou que contêm valores extremos. Ao focar nos quartis, o IQR ignora os valores fora do intervalo interquartil (os 50% centrais dos dados), tornando-o resistente a distorções causadas por outliers. Isso é particularmente útil em dados financeiros, onde outliers podem surgir devido a eventos excepcionais.

4.5 Métodos Baseados em Distância

- **Aplicabilidade**: Métodos como a Distância Euclidiana e a Distância de Mahalanobis são mais aplicáveis quando o conjunto de dados tem uma distribuição aproximadamente normal e as variáveis são correlacionadas. A Mahalanobis, em particular, é útil em dados multivariados e sensível à forma da distribuição.
- **Vantagens**: São métodos diretos e relativamente fáceis de calcular para dados com poucas variáveis, permitindo uma análise rápida.
- **Desvantagens**: São sensíveis a escalas e podem não ser eficazes para dados assimétricos ou com distribuições complexas. Outliers em variáveis individuais podem ser subestimados se os dados forem multivariados.
- **Comparação com IQR**: O IQR é mais robusto e menos sensível à escala dos dados, tornando-o preferível para dados financeiros assimétricos.

4.6 Modelos de Clusterização – DBSCAN

- **Aplicabilidade**: O DBSCAN é útil para conjuntos de dados que apresentam agrupamentos naturais e onde os outliers estão dispersos em áreas de baixa densidade. Ele não exige que os dados sigam uma distribuição normal.

- **Vantagens**: Identifica outliers com base na densidade, o que permite detectar valores anômalos em cenários de clusterização, especialmente em dados geoespaciais ou temporais.
- **Desvantagens**: A escolha dos parâmetros (raio de proximidade e número mínimo de pontos) é crucial e pode impactar a detecção. Pode não identificar outliers em conjuntos de dados homogêneos e sem clusters definidos.
- **Comparação com IQR**: O IQR é mais simples e independente de parâmetros específicos, o que o torna uma escolha mais prática e intuitiva em cenários financeiros onde a clusterização não é comum.

4.7 Análise de Componentes Principais (PCA)

- **Aplicabilidade**: A PCA é aplicável em conjuntos de dados multidimensionais onde se busca reduzir a complexidade e detectar outliers baseando-se em componentes principais. É especialmente eficaz em dados com alta correlação entre variáveis.
- **Vantagens**: Permite a redução de dimensionalidade, destacando variações nos dados que indicam possíveis outliers. Ajuda a visualizar e interpretar dados complexos.
- **Desvantagens**: Requer etapas adicionais de transformação, e a interpretação dos resultados pode ser menos intuitiva. Não é ideal para dados com poucas dimensões ou para dados financeiros onde as variáveis têm pouca correlação.
- **Comparação com IQR**: O IQR é mais direto e menos técnico, sendo fácil de implementar e interpretar para dados financeiros sem necessidade de redução de dimensionalidade.

4.8 Isolation Forest

- **Aplicabilidade**: Isolation Forest é ideal para conjuntos de dados grandes e heterogêneos. É amplamente utilizado em detecção de anomalias com grandes volumes de dados, especialmente em cenários de Machine Learning.
- **Vantagens**: Robusto contra outliers e eficaz em dados de alta dimensão. Trabalha bem com dados complexos e permite detectar anomalias em múltiplas variáveis.
- **Desvantagens**: Requer processamento computacional adicional e uma implementação baseada em machine learning. A interpretação dos resultados pode ser desafiadora para públicos não técnicos.
- **Comparação com IQR**: O IQR é mais simples e menos dependente de processamento intensivo. Para dados financeiros com poucas variáveis e estruturas simples, o IQR é mais conveniente.

4.9 Autoencoders (Redes Neurais)

- **Aplicabilidade**: Autoencoders são úteis em cenários complexos onde os dados possuem padrões não lineares e alta dimensionalidade. São populares em análises de big data e padrões complexos.
- **Vantagens**: Podem detectar anomalias sutis ao modelar padrões de dados complexos. São eficazes para dados com padrões não lineares e distribuídos de forma irregular.
- **Desvantagens**: Requerem conhecimento técnico avançado e poder computacional elevado. A configuração e interpretação podem ser complexas para análises financeiras tradicionais.
- **Comparação com IQR**: O IQR é mais direto e menos exigente computacionalmente. Em dados financeiros, onde a complexidade é limitada, o IQR é mais prático e intuitivo.

5. Vantagens do IQR em Dados Financeiros

5.1 Robustez contra Outliers

O IQR é resistente a outliers, o que o torna ideal para detectar valores que fogem à tendência geral dos dados, sem que esses valores extremos distorçam a análise. No contexto de dados financeiros, isso é crucial, pois grandes discrepâncias nos custos ou valores podem surgir devido a variações de mercado, ou eventos inesperados.

5.2 Adequação para Dados Assimétricos

Os dados financeiros raramente seguem uma distribuição normal. Em vez disso, muitas vezes são assimétricos, com grandes valores ou picos. O IQR é particularmente adequado para esse tipo de distribuição, pois se concentra nos quartis e não assume simetria nos dados. Isso é um diferencial importante em comparação com a média e o desvio padrão.

5.3 Simplicidade e Clareza

O método IQR é simples de aplicar e os resultados são facilmente interpretáveis. Em um gráfico boxplot, os outliers são claramente visualizados como pontos fora dos limites interquartis. Isso facilita a comunicação dos resultados com equipes técnicas e não-técnicas, tornando as decisões baseadas nesses dados mais claras e fundamentadas.

5.4 Melhor Análise de Dispersão

O IQR oferece uma análise mais robusta da dispersão dos dados, focando nos 50% centrais e excluindo valores extremos que possam comprometer a interpretação.

Isso é especialmente útil para definir faixas realistas de custos ou valores financeiros.

5.5. Praticidade e Economia de Recursos
O método IQR (Intervalo Interquartil) é amplamente valorizado pela sua facilidade de aplicação e pelo baixo custo de implementação. Como ele se baseia apenas nos quartis dos dados, a execução e interpretação do IQR são diretas, sem a necessidade de softwares avançados ou configurações complexas. Esses fatores tornam o método IQR uma opção prática e econômica, especialmente quando comparado com técnicas mais complexas, como métodos de Machine Learning ou modelagem multivariada.

Além disso, o cálculo do IQR e dos limites para detecção de outliers requer apenas operações básicas (subtração e multiplicação), o que permite a análise em tempo real, ideal para processos financeiros ou orçamentários onde a resposta rápida é essencial. O IQR também é independente de suposições sobre a distribuição dos dados, facilitando sua aplicação em uma variedade de cenários sem necessidade de ajustes adicionais.

Dessa forma, o método IQR se destaca não apenas por sua robustez e adequação para dados financeiros assimétricos, mas também por sua praticidade, garantindo economia de recursos e tempo sem comprometer a precisão dos resultados.

6. Situação Hipotética Aplicada aos Dados Fornecidos

Supondo que os dados fornecidos sigam uma estrutura semelhante às situações hipotéticas discutidas anteriormente, o uso do IQR para identificar outliers no **Valor Atualizado** dos dados se mostraria a abordagem mais eficaz. Isso pode ser ilustrado na seguinte comparação:

- **Valor Líquido Referência**: 1.000.000, 1.500.000, 2.000.000, 3.000.000, 100.000.000

- **Valor Atualizado**: 1.200.000, 1.700.000, 2.300.000, 2.900.000, 120.000.000

Ao aplicar o método IQR, detectamos que o valor **120.000.000** está muito acima do limite superior. Embora a média e o desvio padrão desses valores pudessem sugerir que o centro dos dados está perto de **25.220.000**, o IQR demonstra que o

verdadeiro centro dos dados está entre **1.700.000** e **2.900.000**, revelando o impacto negativo do outlier.

7. Ações Recomendadas após a Detecção de Outliers

Uma vez detectados os outliers utilizando o método IQR, recomendamos seguir as ações abaixo para garantir uma análise adequada dos dados financeiros:

1. **Ação Default: Análise e Investigação do Contexto dos Outliers** ○ Descrição: Como ação padrão, sugerimos iniciar com a análise do contexto dos dados e a investigação da justificativa dos outliers. Isso envolve verificar se os outliers refletem eventos específicos ou fatores legítimos, como variações de mercado ou circunstâncias excepcionais no projeto. Compreender o contexto é essencial para decidir se esses valores devem ser mantidos ou descartados da análise.
    - Exemplo: Caso um outlier esteja associado a uma mudança significativa nos preços ou custos, pode ser relevante mantê-lo como parte das flutuações naturais do cenário financeiro.
2. **Eliminação e Reexecução da Análise (Caso Necessário)** ○ Descrição: Se, após a análise contextual, os outliers forem considerados anomalias indesejáveis (por exemplo, erros de entrada de dados ou eventos não representativos), recomenda-se a eliminação desses valores e a reexecução da análise.
    - Recomendação de Comparação: Para uma análise mais robusta, sugerimos tabelar e visualizar as métricas com e sem a presença de outliers. Esta comparação pode incluir:
        - Média com e sem outliers
        - Regressão linear com e sem outliers
        - Desvio padrão com e sem outliers
        - Mediana e valores de quartis com e sem outliers ○ Vantagens: Esse procedimento oferece uma visão clara sobre como os outliers impactam as principais métricas e facilita a tomada de decisão sobre sua exclusão.
3. **Análise Comparativa com Outliers** ○ Em alguns casos, realizar uma análise comparativa incluindo e excluindo os outliers fornece insights sobre a extensão do impacto dos outliers nas métricas de interesse. Isso é particularmente útil em decisões onde o efeito do outlier precisa ser justificado para a tomada de decisão (por exemplo, se afetar significativamente o custo ou orçamento).

Essas ações, com destaque para a análise contextual como abordagem padrão, garantem que a detecção de outliers seja seguida de uma avaliação cuidadosa, permitindo uma análise financeira mais precisa e adaptada ao contexto dos dados.

8. Exemplo prático:

Tabela de exemplo:

Índice	Com Outlier	Sem Outlier
1	1.700.000	1.700.000
2	1.800.000	1.800.000
3	2.000.000	2.000.000
4	2.100.000	2.100.000
5	2.200.000	2.200.000
6	2.300.000	2.300.000
7	2.400.000	2.400.000
8	2.500.000	2.500.000
9	2.700.000	2.700.000
10	2.900.000	2.900.000
11	30.000.000	-

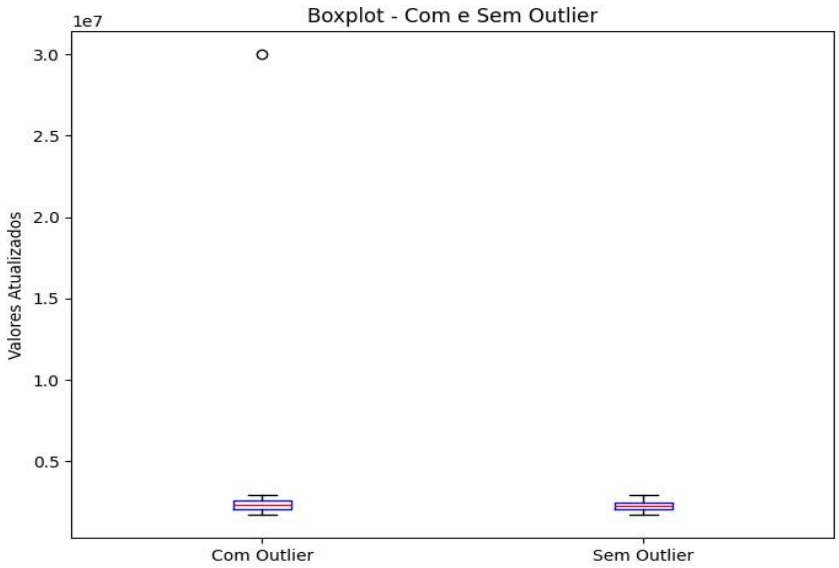
Cálculo das Médias • **Média com Outlier:** 5.591.000 • **Média sem Outlier:** 2.260.000

A presença do outlier de 30.000.000 aumenta significativamente a média dos dados para 5.591.000, enquanto a média sem esse valor extremo é de apenas 2.260.000. Isso mostra o impacto que outliers podem ter na média, distorcendo o valor central e influenciando a interpretação dos dados.

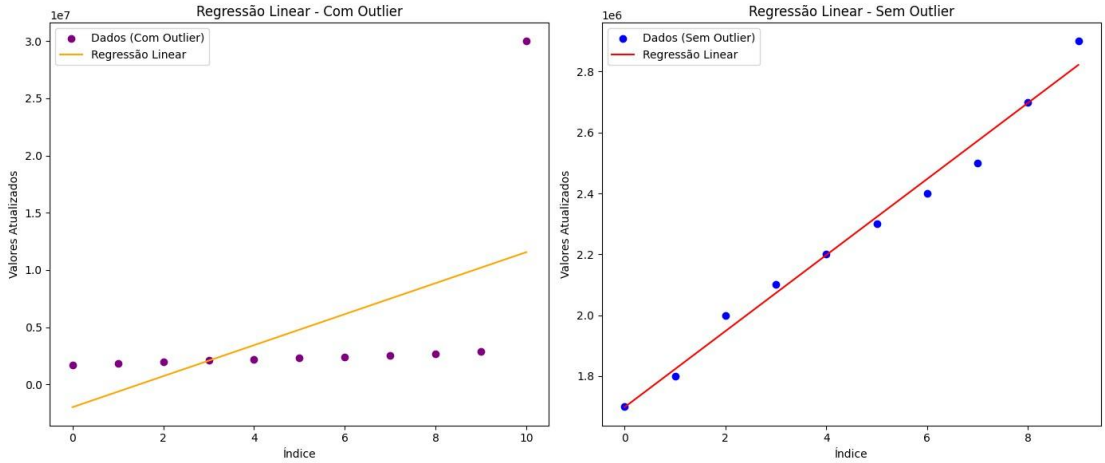
Boxplot - Visualização com e sem Outlier

O boxplot mostra a dispersão dos dados para as duas situações:

- **Com Outlier:** Apresenta uma caixa mais expandida e um valor extremo destacado, elevando a variabilidade.
- **Sem Outlier:** Exibe uma dispersão menor, concentrando os dados próximos ao valor central.



Análise de Regressão Linear



1. **Com Outlier**: A linha de regressão é inclinada para cima, influenciada pelo valor extremo, o que pode induzir uma tendência de crescimento que não representa a maioria dos dados.

2. **Sem Outlier**: A linha de regressão ajusta-se bem aos valores centrais, fornecendo uma tendência mais representativa dos dados sem a distorção do outlier.

O outlier influencia significativamente tanto a média quanto a inclinação da regressão. Para análises onde valores extremos podem distorcer a interpretação, é recomendável remover esses outliers, especialmente em métricas de tendência central e de dispersão. Caso os outliers sejam mantidos, é importante interpretálos com cautela, pois eles podem representar eventos anômalos ou erros de dados.

9. Conclusão

Com base nas comparações feitas, o **IQR (Intervalo Interquartil)** se destaca como a melhor abordagem para a análise de outliers nos dados fornecidos. Enquanto métodos como média, desvio padrão e regressão linear podem ser úteis em diferentes contextos, eles são fortemente influenciados por valores extremos e, portanto, não fornecem uma análise robusta quando há outliers presentes.

O IQR, por sua vez, é robusto contra outliers, sendo adequado para dados assimétricos e permitindo uma interpretação mais precisa da dispersão central dos dados. Para a análise dos dados financeiros fornecidos, onde podem ocorrer variações significativas em custos e valores, o IQR oferece a melhor combinação de precisão, clareza e resistência a distorções.

Recomenda-se, portanto, que o método IQR seja adotado como a abordagem estatística principal para a análise de outliers e dispersão dos dados, garantindo resultados confiáveis e robustos.