Análise da Distribuição da Variável Target: Número de Requisições (reqs)

Introdução

Este relatório apresenta uma análise detalhada da distribuição da variável `reqs` (número de requisições) do dataset `data_center_limpo.parquet`. A variável `reqs` foi identificada como a variável target principal no arquivo `motivacao.ipynb`, sendo crucial para a gestão preditiva de infraestrutura digital em datacenters. A compreensão de sua distribuição é fundamental para o desenvolvimento de modelos preditivos eficazes, que visam otimizar o desempenho operacional e garantir a continuidade dos serviços em cenários de alta demanda ou eventos críticos.

O objetivo desta análise é fornecer uma visão clara sobre o comportamento do número de requisições, incluindo suas tendências centrais, dispersão e a presença de valores atípicos, através da visualização de um histograma e da interpretação de estatísticas descritivas.

Metodologia

Para esta análise, o dataset `data_center_limpo.parquet` foi carregado utilizando a biblioteca `pandas` em Python. A variável `reqs` foi selecionada como foco de estudo. Para visualizar a distribuição desta variável, um histograma foi gerado utilizando a biblioteca `matplotlib`, com 50 bins para uma representação detalhada da frequência das requisições. O histograma foi salvo como `reqs_histogram.png`.

Adicionalmente, foram calculadas estatísticas descritivas básicas para a variável `reqs`, incluindo contagem, média, desvio padrão, valores mínimo e máximo, e os quartis (25%, 50% - mediana, e 75%). Estas estatísticas foram salvas em um arquivo de texto (`reqs_descriptive_stats.txt`) para facilitar a análise textual.

Estatísticas Descritivas da Variável `reqs`

A seguir, são apresentadas as estatísticas descritivas para a variável `reqs`:

Plain Text

```
count       720.000000
mean      74045.897222
std       43237.087995
min       15127.000000
25%       37414.500000
50%       62077.000000
```

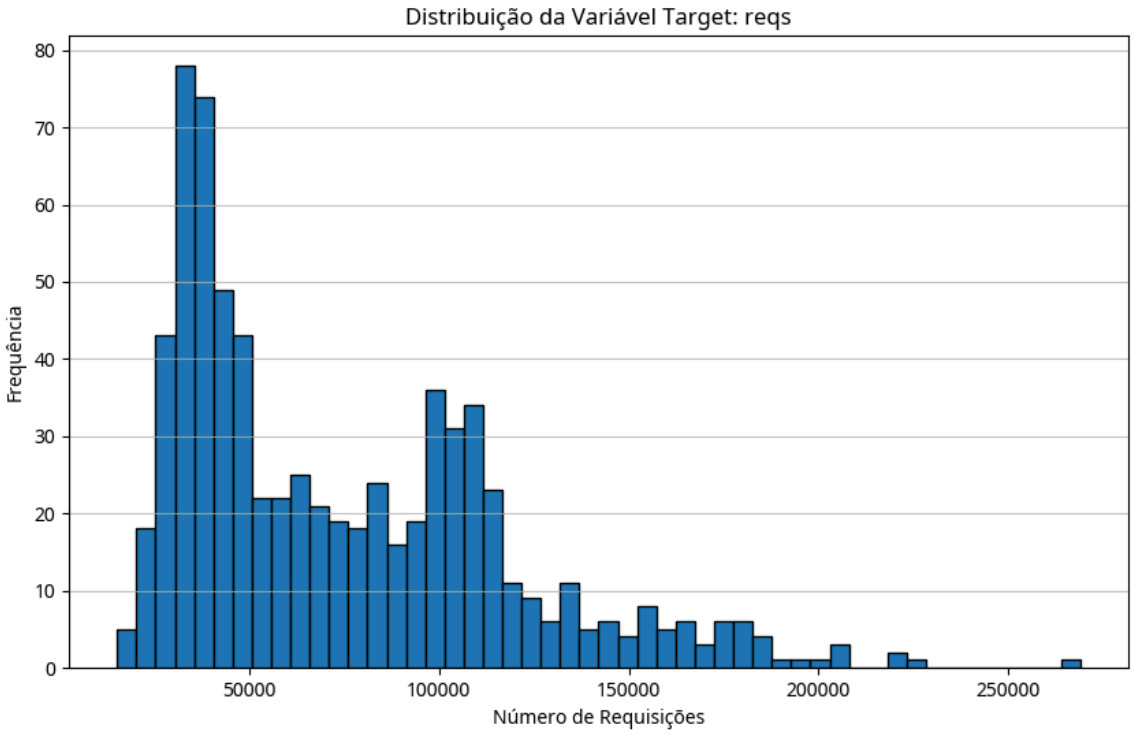
```
75%      102059.750000
max      268996.000000
```

- **Contagem (count)**: O dataset contém 720 observações para a variável `reqs`, indicando um número razoável de pontos de dados para análise.
- **Média (mean)**: O número médio de requisições é de aproximadamente 74.045,90. Este valor representa o centro da distribuição.
- **Desvio Padrão (std)**: Com um desvio padrão de aproximadamente 43.237,09, observa-se uma dispersão considerável dos dados em torno da média. Isso sugere que o número de requisições pode variar bastante.
- **Mínimo (min)**: O menor número de requisições registrado é de 15.127.
- **25º Percentil (25%)**: 25% das observações têm um número de requisições igual ou inferior a 37.414,50.
- **Mediana (50%)**: A mediana, ou 50º percentil, é de 62.077,00. Este valor é inferior à média, o que pode indicar uma assimetria positiva (cauda à direita) na distribuição, onde há mais valores menores e alguns valores maiores puxando a média para cima.
- **75º Percentil (75%)**: 75% das observações têm um número de requisições igual ou inferior a 102.059,75.
- **Máximo (max)**: O maior número de requisições registrado é de 268.996,00, um valor significativamente alto em comparação com a média e a mediana, reforçando a ideia de assimetria e a presença de picos de demanda.

Essas estatísticas sugerem que a distribuição de `reqs` não é simétrica e pode apresentar uma concentração de valores em faixas mais baixas, com ocorrências menos frequentes de valores muito altos, o que é comum em dados de demanda e tráfego. A visualização do histograma a seguir confirmará essa hipótese.

Histograma da Variável `reqs`

A imagem a seguir apresenta o histograma da variável `reqs`, ilustrando visualmente a distribuição do número de requisições.



Análise do Histograma

O histograma da variável `reqs` revela características importantes sobre a distribuição do número de requisições:

- **Assimetria Positiva (Cauda à Direita)**: A distribuição é claramente assimétrica à direita, o que significa que a maioria das observações de `reqs` se concentra em valores mais baixos, enquanto um número menor de observações atinge valores muito mais altos. Isso é consistente com a observação de que a média (74.045,90) é maior que a mediana (62.077,00).
- **Picos de Demanda**: A presença de uma cauda longa à direita indica a ocorrência de picos de demanda significativos. Estes picos representam momentos em que o datacenter experimenta um volume excepcionalmente alto de requisições, o que é um ponto crítico para a gestão preditiva, conforme destacado no `motivacao.ipynb`.
- **Concentração de Dados**: A maior frequência de requisições ocorre nas faixas de valores mais baixos, sugerindo que o datacenter opera na maior parte do tempo sob uma carga de trabalho moderada. No entanto, a existência de valores na faixa superior (próximo a 270.000 requisições) demonstra a capacidade do sistema de lidar com eventos de alta demanda, mas também a necessidade de monitoramento e previsão para evitar sobrecargas.
- **Implicações para Modelagem Preditiva**: A natureza assimétrica da distribuição de `reqs` é um fator importante a ser considerado na modelagem preditiva. Modelos que

assumem uma distribuição normal podem não ser os mais adequados. Técnicas que lidam bem com dados assimétricos ou a transformação da variável `reqs` (por exemplo, logarítmica) podem ser necessárias para melhorar a performance dos modelos. Além disso, a previsão dos picos de demanda será um desafio crucial, exigindo modelos capazes de capturar esses eventos raros, mas de alto impacto.

Conclusão

A análise da variável `reqs` através de estatísticas descritivas e do histograma fornece insights valiosos sobre o comportamento da carga de trabalho no datacenter. A distribuição assimétrica com picos de demanda reforça a importância da gestão preditiva para garantir a estabilidade e a eficiência operacional. O próximo passo no desenvolvimento do modelo preditivo deve considerar essas características da distribuição para selecionar as abordagens de modelagem mais apropriadas e eficazes.