Curso: Ciências da Computação	Disciplina: M. A. Desempenho	Código/Turma: 36
Professor/a: Lyndainês		Data: 02/05/2025
Aluno/a: Cairo de Queirós Vidal		Matrícula:2217557

Análise Exploratória dos Dados

Neste relatório, iremos analisar um conjunto de dados referente a diferentes configurações de máquinas, abrangendo variáveis como quantidade de CPU, memória RAM, latência, tipo de armazenamento e características categóricas como sistema operacional e tipo de processador. O objetivo é gerar posteriormente um modelo de regressão para prever o tempo de resposta com base nessas variáveis.

Inicialmente, é essencial observar as características estatísticas fundamentais do dataset antes da análise. Nas tabelas abaixo são apresentadas as estatísticas descritivas para variáveis discretas e categóricas após o tratamento dos valores faltantes.

Antes da análise detalhada, é importante destacar como tratamos os valores nulos:

- **Para variáveis discretas (numéricas)**, optou-se por preencher valores faltantes utilizando a **mediana**, devido à sua robustez em relação à presença de outliers, reduzindo a influência de valores extremos sobre a distribuição dos dados.
- **Para variáveis categóricas**, escolhemos preencher os valores nulos com a **moda**, representando a categoria mais frequente observada para cada variável.

Essas escolhas garantem que os dados analisados estejam livres de lacunas, mantendo características estatísticas coerentes e minimizando a influência negativa de valores extremos ou desvios na distribuição.

Variáveis Discretas:

	cpu_cores	ram_gb	latencia_ms	armazenamento_tb	tempo_resposta
Média	8.88	23.46	142.39	0.97	102.64
Mediana	9.0	16.0	135.33	1.0	82.62
Desvio Padrão	3.88	21.09	81.8	0.63	60.04
Mínimo	2.0	4.0	3.41	0.26	24.14
Máximo	15.0	64.0	299.29	2.0	398.73
Q1	6.0	8.0	72.35	0.51	66.87
Q3	12.0	32.0	210.99	1.0	114.54
IQR	6.0	24.0	138.64	0.49	47.66
Qtd. Valores Nulos	0.0	0.0	0.0	0.0	0.0

Observando os dados numéricos, notamos que o número médio de núcleos de CPU é de aproximadamente 8,88, com uma mediana de 9 núcleos, indicando uma distribuição relativamente simétrica para essa variável. A RAM, por outro lado, apresenta grande variação com média de 23,46 GB e mediana de 16 GB, sugerindo assimetria, provavelmente devido à presença de valores mais altos que influenciam a média.

A variável "tempo de resposta", nosso alvo de predição, possui uma média de aproximadamente 102,64 ms e uma mediana significativamente menor, de 82,62 ms, indicando a presença de valores extremos elevados. O armazenamento médio é próximo de 1 TB (0,97 TB) com pouca variação (desvio padrão de 0,63 TB).

A latência apresenta uma média de aproximadamente 142,39 ms com considerável desvio padrão (81,8 ms), sugerindo heterogeneidade nos valores observados, possivelmente relacionada às diferentes configurações das máquinas.

Nenhum valor nulo foi observado após o tratamento com medianas nas variáveis numéricas, garantindo maior robustez para a etapa posterior de modelagem.

Variáveis Categóricas:

	sistema_operacional	tipo_hd	tipo_processador
Moda	MacOS	HDD	Apple Silicon
Frequência da Moda	68	113	68
Frequência Relativa da Moda (%)	36.96	61.41	36.96
Qtd. Valores Únicos	3	2	3
Qtd. Valores Nulos	0	0	0

Analisando as variáveis categóricas, o sistema operacional predominante é o MacOS, presente em cerca de 36,96% dos casos, seguido por outros dois sistemas. O tipo de armazenamento HDD é amplamente majoritário (61,41%), indicando que máquinas com HDD predominam claramente no dataset, fato relevante na análise do tempo de resposta, já que HDs tendem a ter desempenho inferior em comparação a SSDs.

Já em relação ao tipo de processador, destaca-se o Apple Silicon, também com frequência relativa de 36,96%, sugerindo uma distribuição equilibrada em relação às outras categorias presentes.

Assim como nas variáveis discretas, não existem valores nulos após o tratamento com a moda nas variáveis categóricas, garantindo integridade dos dados para análise subsequente.

Essas observações serão fundamentais para interpretar os resultados da regressão e avaliar o impacto dessas variáveis sobre o tempo de resposta.

Parte II – Modelo e Diagnóstico

Nesta etapa, realizaremos o ajuste de um modelo de regressão linear múltipla com as seguintes características:

- **Variável dependente:** tempo_resposta
- **Variáveis explicativas:** Todas as demais variáveis (numéricas e categóricas).

Inicialmente, precisamos garantir que todas as variáveis incluídas no modelo sejam numéricas. Para isso, convertemos as variáveis categóricas utilizando o método das **variáveis dummy**, gerando colunas binárias que representam a presença (valor 1) ou ausência (valor 0) de cada categoria específica.

No dataset original, as variáveis categóricas são:

- **sistema_operacional**
- **tipo_hd**
- **tipo_processador**

Essas variáveis foram convertidas utilizando a função `pd.get_dummies()` do Pandas com o parâmetro `drop_first=True`. Com isso, para cada variável categórica, a **primeira categoria em ordem alfabética** é excluída automaticamente, sendo adotada como **categoria de referência** no modelo, essa exclusão evita problemas de multicolinearidade.

As categorias base adotadas para cada variável categórica foram:

- **sistema_operacional:** Linux
- **tipo_hd:** HDD
- **tipo_processador:** AMD

Após a criação das variáveis dummy (utilizando como categoria base a mais frequente em cada variável categórica), obtivemos os seguintes resultados do modelo ajustado:

Métrica	Valor
R^2	0.728
R^2 Ajustado	0.715
Estatística F	58.51
Prob (F-stat)	1.62e-45
Log-Likelihood	-894.33
AIC	1806.66
BIC	1835.59
Nº de Observações	184
Df Model	8
Df Residual	175
Tipo de Covariância	nonrobust

	Coeficiente	Erro Padrão	t	P-valor	IC 95% Inf.	IC 95% Sup.
const	238.2073	10.3377	23.04	0.0	217.8047	258.6098
cpu_cores	-12.0533	0.6195	-19.46	0.0	-13.2759	-10.8307
ram_gb	-1.2289	0.1132	-10.85	0.0	-1.4524	-1.0054
latencia_ms	-0.0075	0.0292	-0.26	0.7973	-0.0652	0.0502
armazenamento_tb	-2.3841	3.8455	-0.62	0.5361	-9.9736	5.2053
sistema_operacional_MacOS	-0.2365	3.2179	-0.07	0.9415	-6.5874	6.1143
sistema_operacional_Windows	-1.2734	6.0518	-0.21	0.8336	-13.2174	10.6706
tipo_hd_SSD	9.1705	4.9375	1.86	0.0649	-0.5742	18.9151
tipo_processador_Apple Silicon	-0.2365	3.2179	-0.07	0.9415	-6.5874	6.1143
tipo_processador_Intel	2.081	6.027	0.35	0.7303	-9.8139	13.9759

Avaliação dos Coeficientes e Qualidade do Modelo

Testes Estatísticos Realizados:

Para avaliar a significância estatística dos coeficientes e do modelo como um todo, utilizamos dois testes principais:

- **Teste T (t-Student):**
  Utilizado para verificar se cada coeficiente individual da regressão é estatisticamente diferente de zero.
  - **Hipótese nula (H_0):** O coeficiente é igual a zero (a variável não tem efeito significativo no tempo de resposta).
  - **Hipótese alternativa (H_1):** O coeficiente é diferente de zero (a variável tem efeito significativo).
  - Se o **valor-p** associado ao teste T for **menor que 0,05**, rejeitamos H_0, indicando que a variável tem influência estatisticamente significativa sobre o tempo de resposta.
- **Teste F:**
  Utilizado para avaliar a significância global do modelo. Verifica se ao menos uma variável explicativa tem relação com a variável dependente.
  - **Hipótese nula (H_0):** Nenhuma variável explicativa contribui significativamente.
  - **Hipótese alternativa (H_1):** Pelo menos uma variável contribui significativamente.
  - Um **valor-p muito pequeno (menor que 0,05)** indica que o modelo como um todo é estatisticamente significativo.

Coeficientes destacados em vermelho (variáveis não significativas): Esses coeficientes possuem valores-p elevados (maiores que 0,05), indicando que suas associações com o tempo de resposta não são estatisticamente significativas no contexto do modelo atual.

Valores de R^2 e R^2 Ajustado:

- **R^2:** 0,728 (72,8%)
- **R^2 Ajustado:** 0,715 (71,5%)

Estes valores indicam que aproximadamente 72,8% da variabilidade no tempo de resposta pode ser explicada pelo conjunto de variáveis selecionadas para o modelo. O valor ajustado leva em consideração o número de variáveis incluídas, indicando um bom ajuste do modelo aos dados observados.

Teste global do modelo (Estatística F):

- **Estatística F:** 58,51
- **Probabilidade associada (p-valor):** 1,62e-45

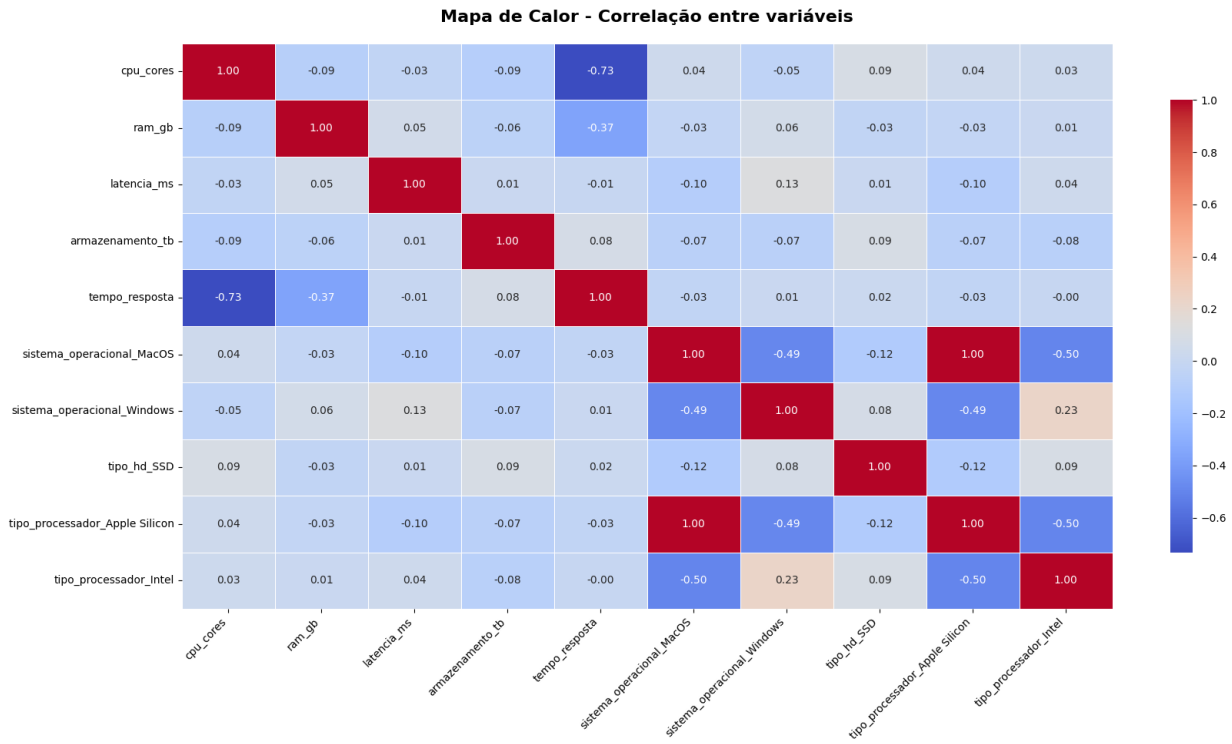
O valor altamente significativo (p-valor extremamente próximo de zero) indica que o modelo ajustado explica adequadamente a variabilidade observada na variável dependente, sendo globalmente significativo.

Interpretação dos Coeficientes das Variáveis Categóricas:

A interpretação desses coeficientes é feita sempre em comparação à categoria base:

- **sistema_operacional:** A categoria base é o Linux.
  - Windows tem um efeito médio negativo (-1,2734) em comparação ao Linux, porém não significativo, indicando que não há evidência estatística de diferença no tempo médio de resposta entre esses sistemas operacionais.
- **tipo_hd:** A categoria base é o HDD.
  - SSD apresentou um coeficiente positivo (9,1705) e marginalmente significativo (p = 0,0649), sugerindo que máquinas com SSD tendem a apresentar tempos de resposta ligeiramente maiores do que aquelas com HDD. Essa observação deve ser interpretada com cautela devido à marginalidade estatística.
- **tipo_processador:** A categoria base é AMD.
  - Processadores Intel possuem um coeficiente positivo (2,081), embora não significativo, indicando que, estatisticamente, não há diferença relevante no tempo de resposta entre máquinas com processadores Intel e AMD.

Diagnóstico inicial de Multicolinearidade (Mapa de Calor):



Ao analisar detalhadamente o mapa de calor das correlações atualizado, identificamos uma situação crítica em relação à multicolinearidade:

- As variáveis **tipo_processador_Apple Silicon** e **sistema_operacional_MacOS** apresentam uma correlação perfeita (valor de 1.00). Isso significa que essas duas variáveis fornecem exatamente a mesma informação para o modelo, o que gera um sério problema conhecido como **multicolinearidade perfeita**.

Nesse caso, é essencial tomar uma ação corretiva antes da interpretação dos resultados finais do modelo. A recomendação mais comum seria remover uma dessas variáveis, pois ambas transmitem a mesma informação ao modelo.

Além disso, também observamos uma correlação moderada negativa (-0,50) entre **sistema_operacional_MacOS** e **tipo_processador_Intel**, que reforça a necessidade de atenção especial a essas variáveis categóricas.

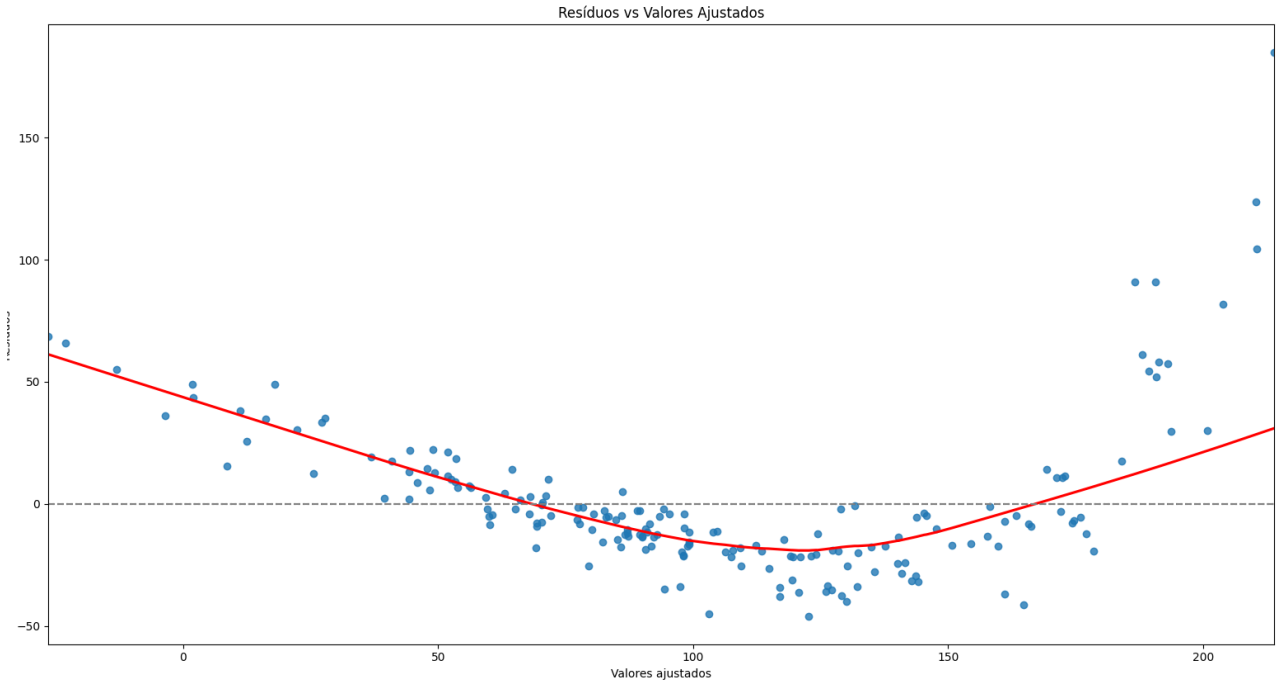
Portanto, recomenda-se fortemente que uma das variáveis (**tipo_processador_Apple Silicon** ou **sistema_operacional_MacOS**) seja excluída do modelo para resolver essa questão, garantindo a confiabilidade e robustez dos resultados do modelo de regressão.

Diagnóstico de Heterocedasticidade

Para avaliar a presença de **heterocedasticidade** no modelo de regressão linear múltipla, utilizamos duas abordagens complementares: **análise gráfica dos resíduos** e o **Teste de Breusch-Pagan**.

Análise Gráfica – Resíduos vs Valores Ajustados

Abaixo está o gráfico dos resíduos em função dos valores ajustados:



Este gráfico permite uma análise visual sobre a variância dos resíduos. Caso os resíduos estejam dispersos aleatoriamente ao redor da linha zero, com variância constante, isso sugere **homocedasticidade** (variância constante dos erros).

No entanto, no gráfico apresentado, é possível observar um **padrão de funil** (menor dispersão para valores médios e maior dispersão nas extremidades). Esse tipo de padrão é um indicativo de **heterocedasticidade**, ou seja, a variância dos erros não é constante ao longo dos valores ajustados.

Teste Estatístico – Breusch-Pagan

Para confirmar a evidência visual, foi aplicado o **Teste de Breusch-Pagan**, um teste formal que verifica a dependência da variância dos resíduos em relação às variáveis explicativas.

Métrica	Valor
LM Statistic	19.904848484236904
LM p-value	0.01850878480355333
F-Statistic	2.6534517112216793
F p-value	0.00900930285686676

Interpretação:
Os p-valores tanto do teste LM quanto do teste F estão **abaixo de 0,05**, o que nos leva a **rejeitar a hipótese nula de homocedasticidade**. Em outras palavras, há **evidências estatísticas significativas** de que a variância dos resíduos não é constante, confirmando a presença de heterocedasticidade no modelo.

Tanto a inspeção visual quanto o teste formal indicam que há heterocedasticidade nos resíduos do modelo. Isso **viola um dos pressupostos básicos da regressão linear clássica** e pode afetar a confiabilidade das inferências sobre os coeficientes, especialmente no que diz respeito aos erros padrão e valores-p.

Análise Crítica

Para avaliar o desempenho do modelo de regressão linear e sua capacidade preditiva, comparamos dois modelos distintos:

- **Modelo 1 – Modelo Completo:** inclui todas as variáveis numéricas e dummies geradas a partir das variáveis categóricas.
- **Modelo 2 – Modelo Reduzido:** considera apenas as variáveis `ram_gb` e `cpu_cores`, selecionadas com base em significância estatística pelo **teste t**.

Critério para exclusão de variáveis

A escolha das variáveis no modelo reduzido foi orientada pelo **teste t aplicado ao Modelo 1**. Variáveis com **valores-p superiores a 0,05** foram consideradas **estatisticamente insignificantes** e, portanto, excluídas. Além disso, também foram removidas variáveis com **multicolinearidade severa** (ex.: correlação perfeita entre `sistema_operacional_MacOS` e `tipo_processador_Apple Silicon`).

Dados do modelo Reduzido

Métrica	Valor
R^2	0.722
R^2 Ajustado	0.718
Estatística F	234.48
Prob (F-stat)	5.67e-51
Log-Likelihood	-896.46
AIC	1798.91
BIC	1808.56
Nº de Observações	184
Df Model	2
Df Residual	181
Tipo de Covariância	nonrobust

	Coeficiente	Erro Padrão	t	P-valor	IC 95% Inf.	IC 95% Sup.
const	237.1288	6.6398	35.71	0.0	224.0274	250.2302
ram_gb	-1.2314	0.1121	-10.98	0.0	-1.4526	-1.0102
cpu_cores	-11.8988	0.6089	-19.54	0.0	-13.1002	-10.6973

Comparação entre os Modelos

Comparativo entre os dois modelos

Métrica	Modelo Completo	Modelo Reduzido
R^2	0.728	0.722
R^2 Ajustado	0.715	0.718
Estatística F	58.51	234.48
AIC	1806.66	1798.91
BIC	1835.59	1808.56
Nº Variáveis (excluindo constante)	8	2

Interpretação:

- O **R^2 ajustado** do modelo reduzido (0.718) é **praticamente equivalente** ao do modelo completo (0.715), apesar de usar **muito menos variáveis**, o que é uma evidência de que o modelo reduzido é mais eficiente e generalizável.
- A **estatística F** do modelo reduzido é superior (234.48 vs 58.51), indicando maior robustez estatística com menor complexidade.
- Os valores de **AIC e BIC** também são menores no modelo reduzido, reforçando sua superioridade sob critérios de penalização por complexidade.

Recomendação Final

Recomenda-se o uso do Modelo Reduzido, pois:

- É mais simples e interpretável;
- Possui desempenho estatístico praticamente idêntico (ou superior) ao modelo completo;
- Elimina variáveis irrelevantes, reduzindo o risco de overfitting;
- Torna a manutenção e explicação do modelo mais objetiva para uso prático.

Ações Práticas Sugeridas

Com base nos resultados dos modelos, é possível sugerir medidas para **otimizar o tempo de resposta do sistema**:

1. **Aumentar a quantidade de núcleos da CPU:**
   O coeficiente negativo associado a `cpu_cores` indica que quanto maior o número de núcleos, menor tende a ser o tempo de resposta.
2. **Expandir a memória RAM:**
   Ainda que com impacto mais sutil do que a CPU, o coeficiente também negativo de `ram_gb` indica que mais memória contribui para tempos de resposta menores.