

Caracterizando a atividade de code review no GitHub

Gabriel Faria, João Victor Salim, Lucas
Garcia, Maisa Pires e Miguel Vieira

HIPÓTESES

H01

PRs menores
(menos arquivos e
linhas modificadas)
têm maior
probabilidade de
serem merged.

H02

PRs com
descrições mais
detalhadas têm
maior probabilidade
de serem merged.

H03

PRs que demoram
mais tempo para
serem analisados
têm menor
probabilidade de
serem merged.

HIPÓTESES

H04

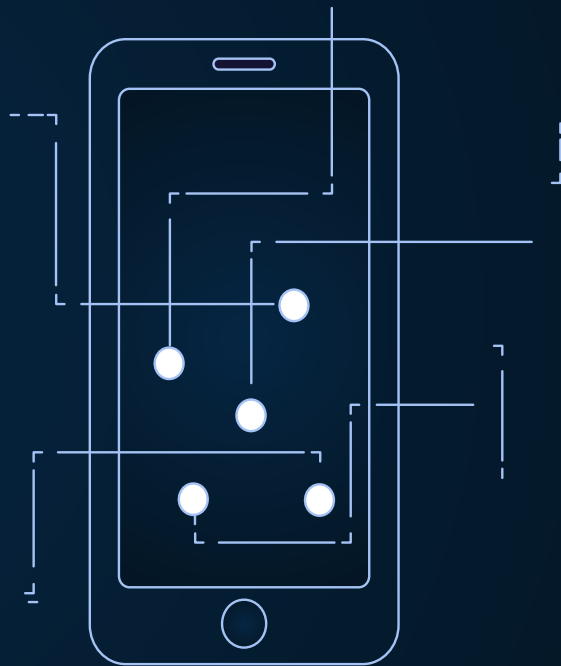
PRs com mais interações (comentários e participantes) têm maior probabilidade de serem merged.

H05

PRs maiores requerem mais revisões.

TECNOLOGIAS

- **Linguagens:**
 - a. Python 3.10+ (análise e geração de gráficos)
- **Bibliotecas**
 - a. requests, pandas, numpy, matplotlib, seaborn, scipy
- **Análise estática**
 - a. Teste de correlação de Spearman
- **API:**
 - a. GitHub GraphQL API (consulta de repositórios Java por popularidade)



METODOLOGIA

- **Coleta de dados**

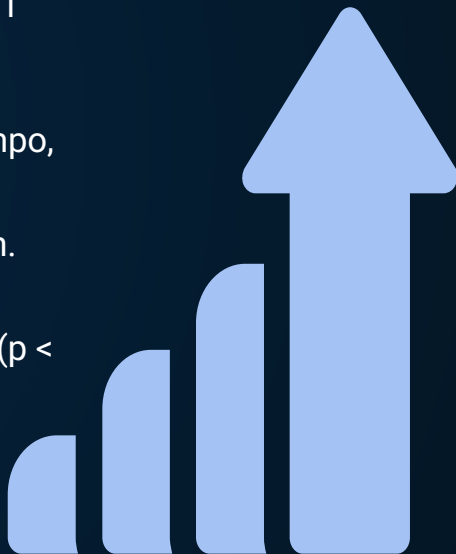
- Foram coletados dados dos 200 repositórios mais populares do GitHub.
- Resultado: 4.115 PRs brutos; 3.616 PRs válidos (status merged/closed e ≥ 1 interação).

- **Métricas definidas**

- Consolidação em pull_requests.csv com métricas derivadas (tamanho, tempo, interações).
- Resultados e figuras em results/ e resumo em results/analysis_results.json.

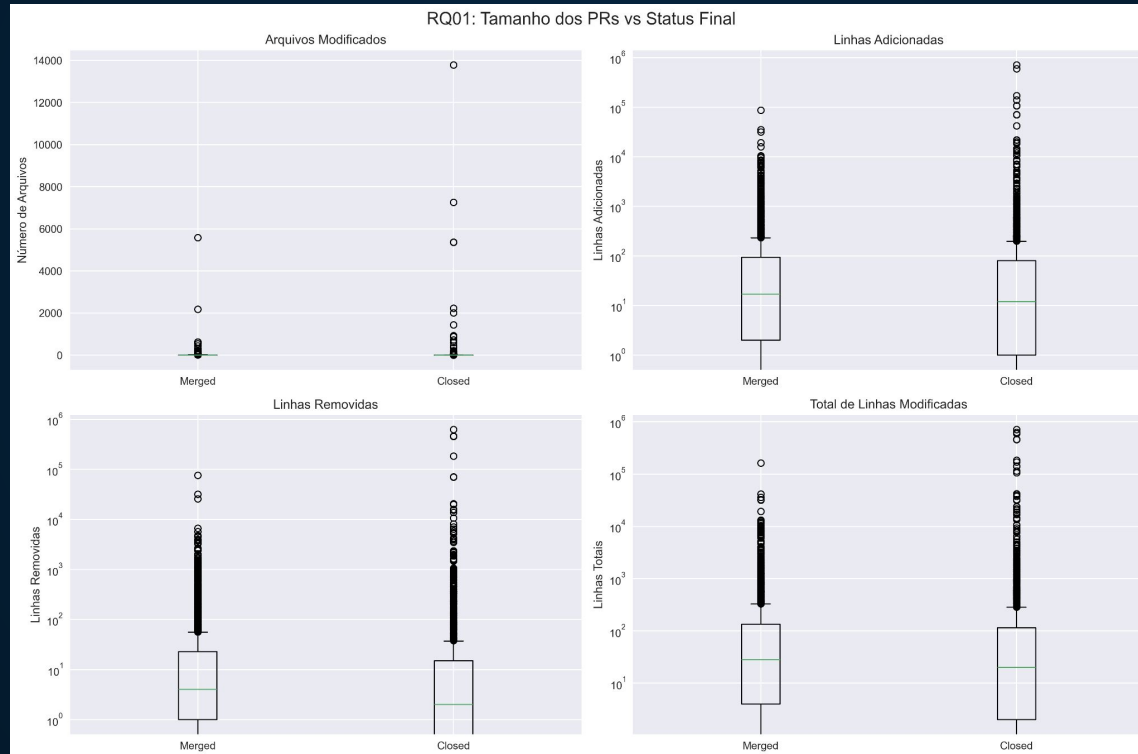
- **Análise**

- Foi realizada estatística descritiva (medianas) e correlações de Spearman ($p < 0,05$). Foram gerados gráficos por RQ e uma matriz de correlação das principais variáveis.



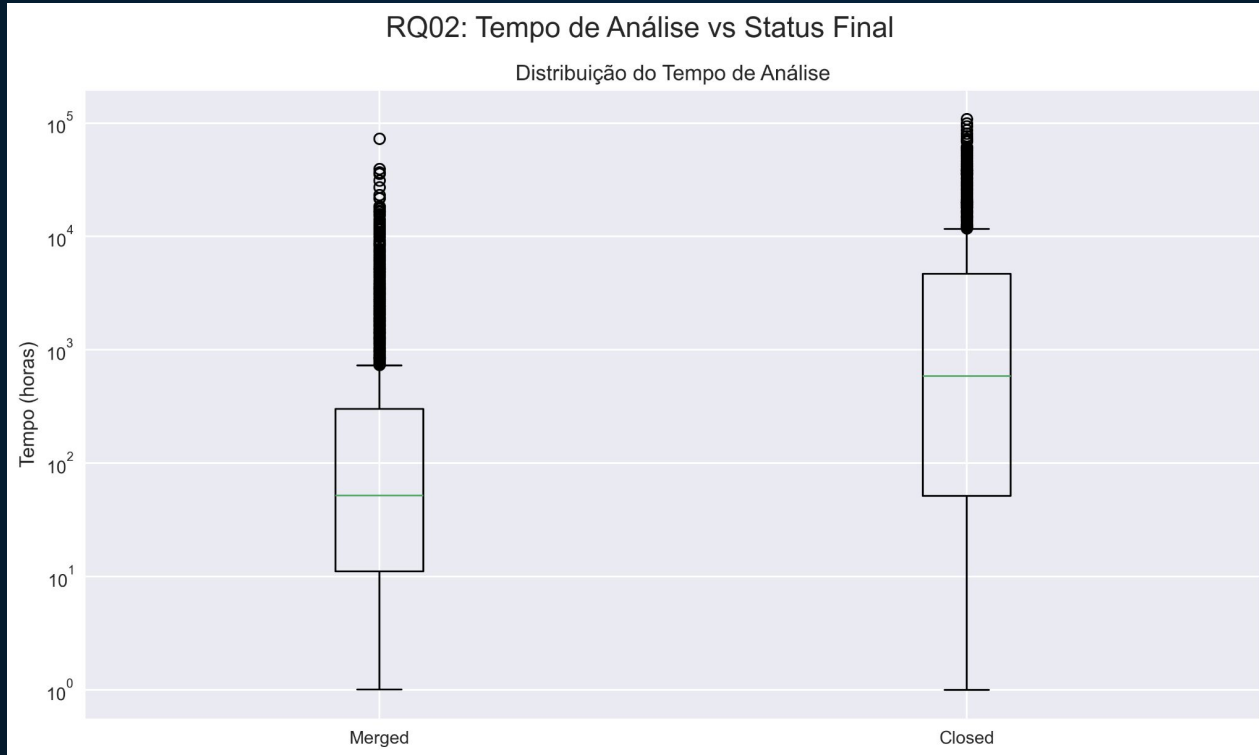
RESULTADO 1

Qual a relação entre o tamanho dos PRs e o feedback final das revisões?



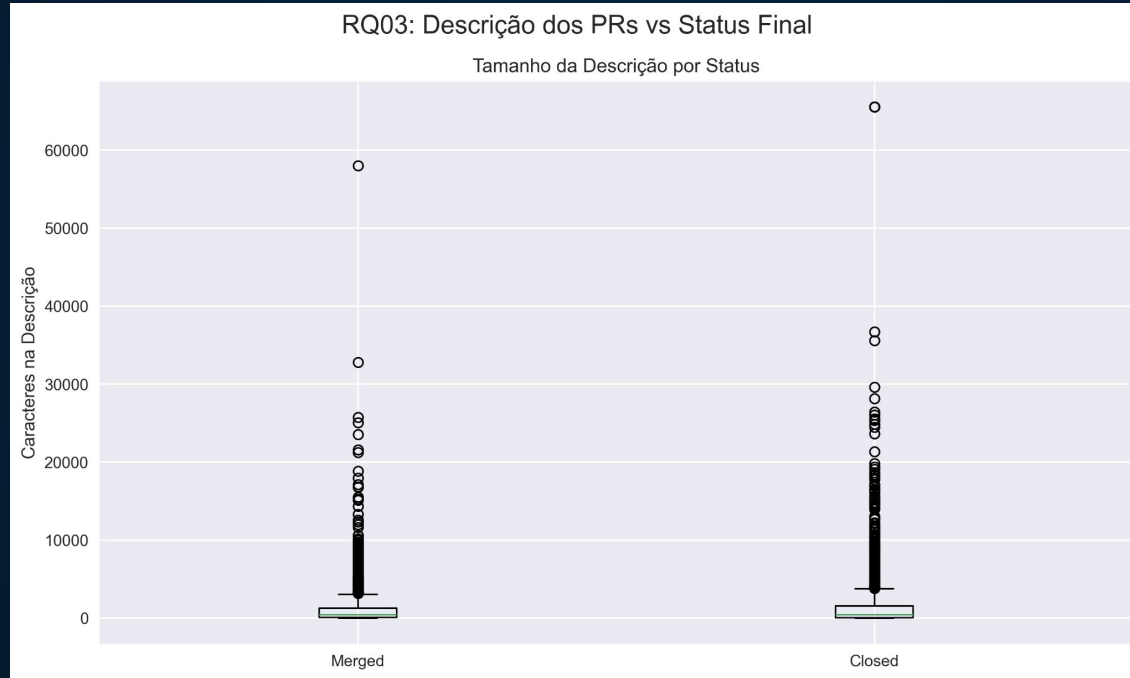
RESULTADO 2

Qual a relação entre o tempo de análise dos PRs e o feedback final das revisões?



RESULTADO 3

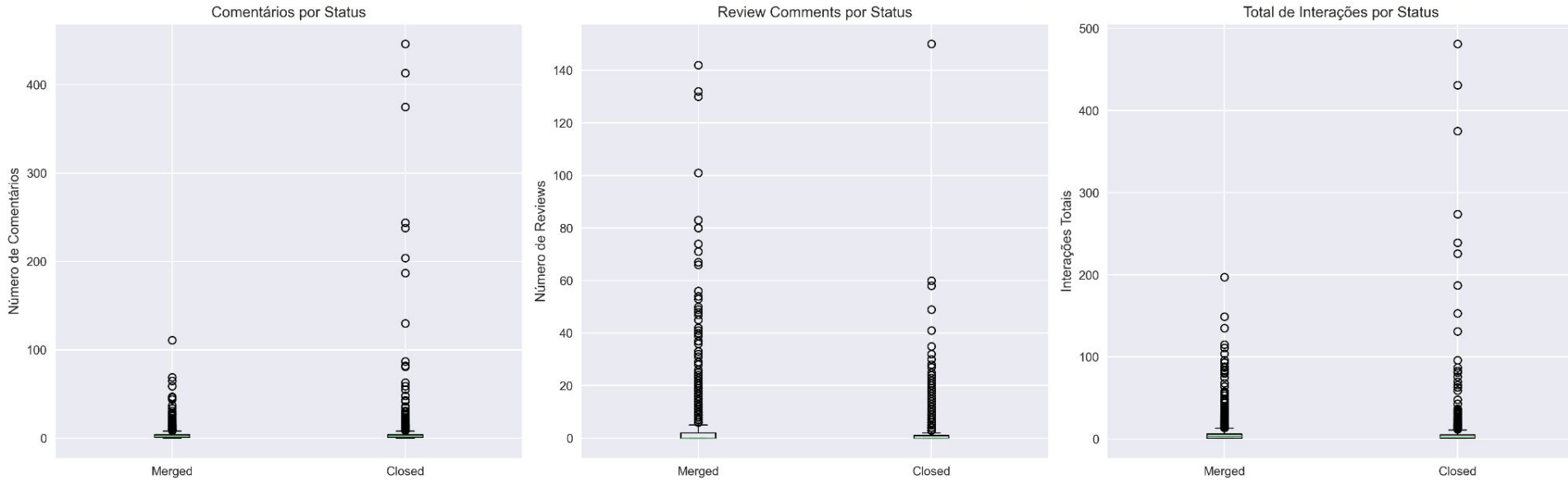
Qual a relação entre a descrição dos PRs e o feedback final das revisões?



RESULTADO 4

Qual a relação entre as interações nos PRs e o feedback final das revisões?

RQ04: Interações nos PRs vs Status Final



RESULTADO 5

Qual a relação entre o tamanho dos PRs e o número de revisões realizadas?



RESULTADO 6

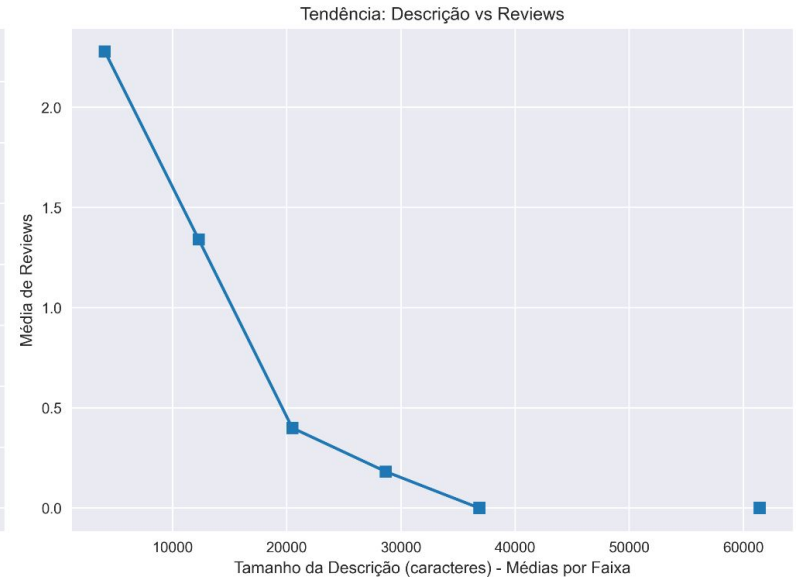
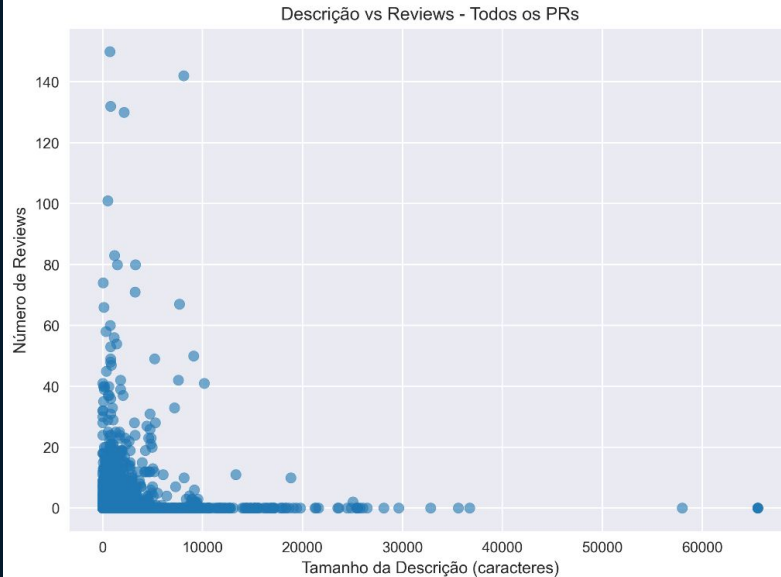
Qual a relação entre o tempo de análise dos PRs e o número de revisões realizadas?



RESULTADO 7

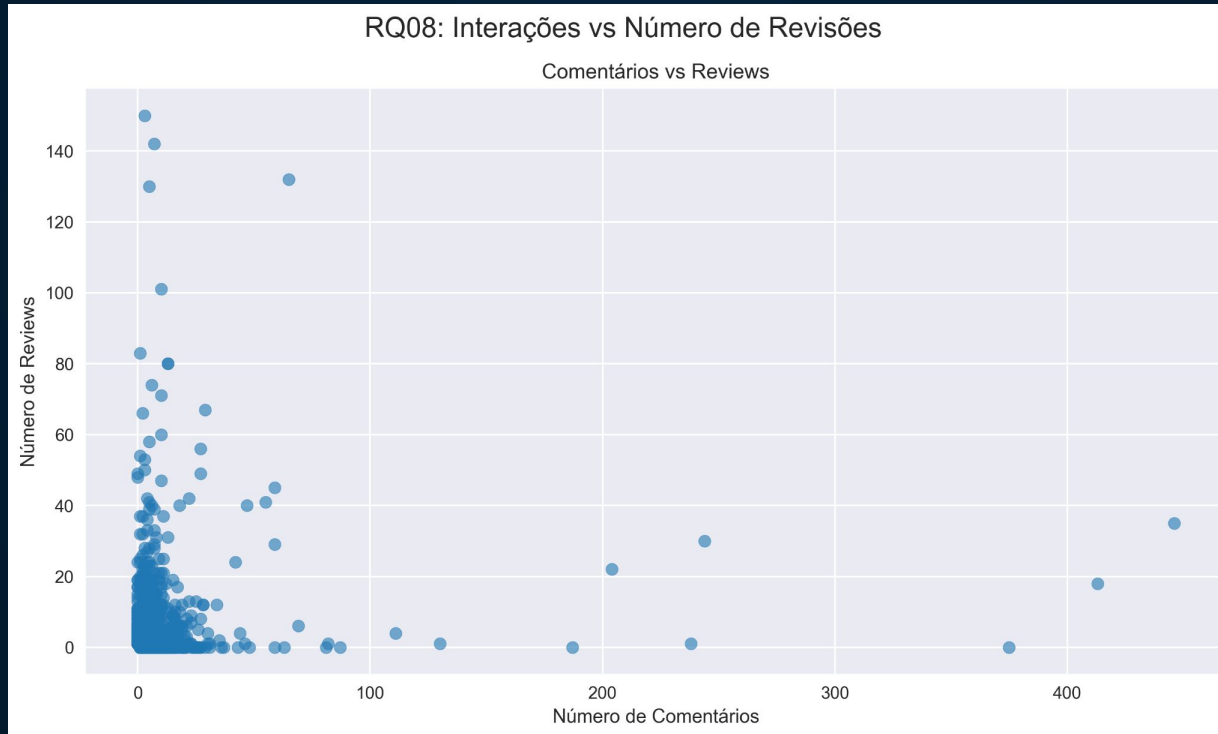
Qual a relação entre a descrição dos PRs e o número de revisões realizadas?

RQ07: Descrição dos PRs vs Número de Revisões

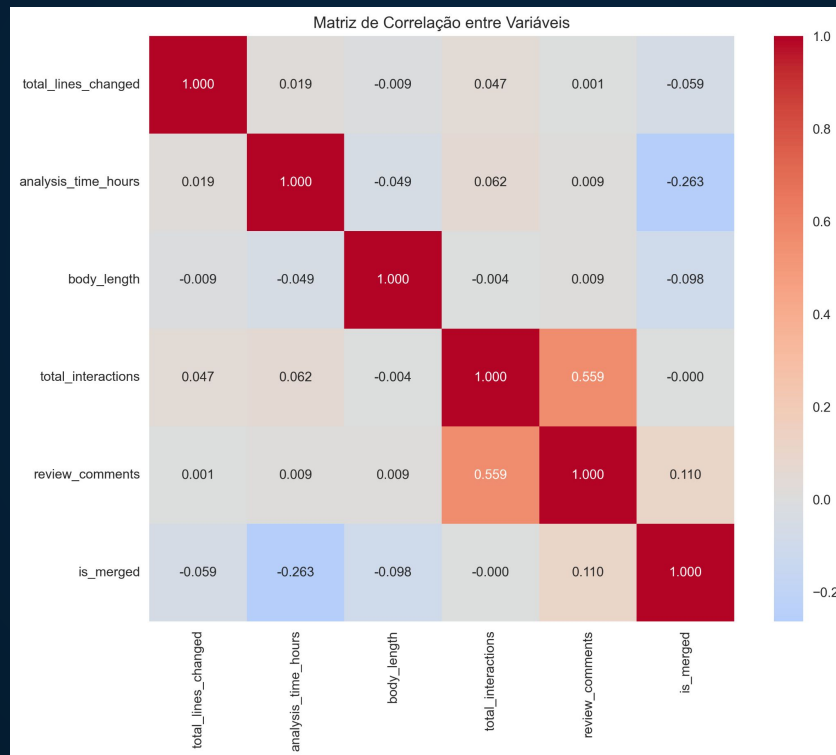


RESULTADO 8

Qual a relação entre as interações nos PRs e o número de revisões realizadas?



Correlação Geral



DISCUSSÃO

Tempo é o fator mais determinante: A correlação mais forte encontrada ($r = -0.355$) foi entre tempo de análise e probabilidade de merge. PRs que demoram mais para serem analisados têm menor chance de serem integrados.

Tamanho influencia revisões: PRs maiores (mais linhas adicionadas) requerem significativamente mais revisões ($r = 0.298$), confirmando a hipótese IH05.

Reviews vs Comments: Interessantemente, review comments aumentam a chance de merge ($r = 0.196$), enquanto comments gerais a diminuem ($r = -0.061$), sugerindo que feedback estruturado é mais efetivo.

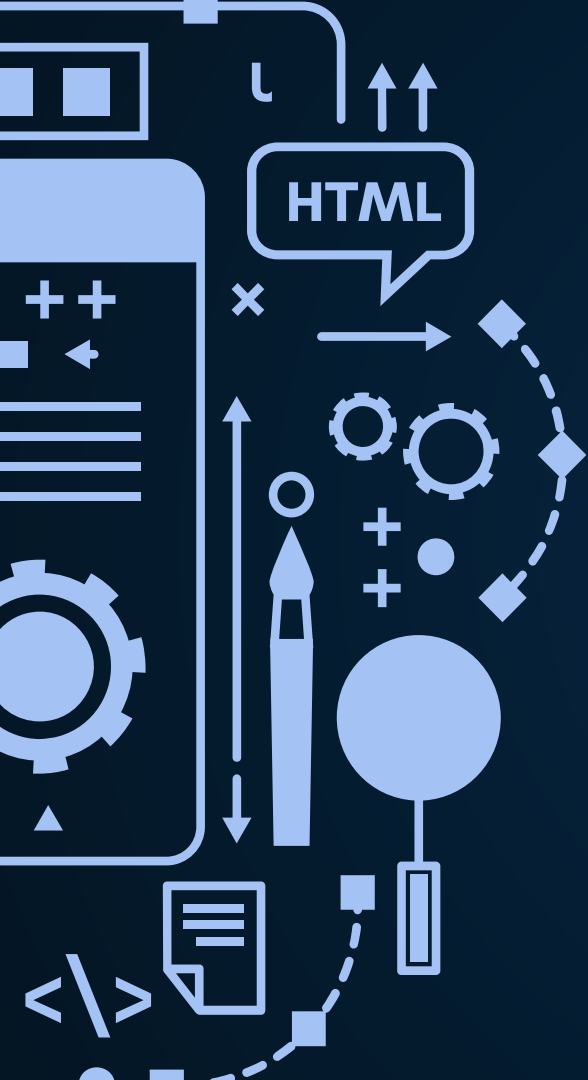
- IH01 NÃO CONFIRMADA - Correlações observadas indicam leve tendência oposta
- IH02 NÃO CONFIRMADA - Sem correlação significativa entre tamanho da descrição e merge.
- IH03 CONFIRMADA - Correlação negativa forte ($r = -0.355$).
- IH04 PARCIALMENTE CONFIRMADA - Depende do tipo de interação
- IH05 CONFIRMADA - Correlação positiva ($r \approx 0.27-0.30$ para tamanho).

CONCLUSÃO

O estudo de 3.616 PRs de repositórios populares do GitHub revelou que:

- O tempo de análise é o fator mais crítico para o sucesso de um PR ($r = -0.355$)
- PRs maiores demandam mais revisões, mas não necessariamente são rejeitados
- A qualidade das interações importa mais que a quantidade (reviews > comments)
- Descrições detalhadas aumentam ligeiramente as chances de merge

A taxa geral de merge de 55.5% indica um processo seletivo, mas equilibrado nos repositórios analisados.



Obrigado!

Alguma Pergunta?

Bibliografía

- GitHub REST API: <https://docs.github.com/en/rest>
- Pandas: <https://pandas.pydata.org/>
- Matplotlib: <https://matplotlib.org/>
- Seaborn: <https://seaborn.pydata.org/>
- SciPy: <https://scipy.org/>