Relatório de Análise de Pull Requests em Repositórios Populares do GitHub

Alunos: Pedro Reis, Gabriel Fernandes, Guilherme Augusto

Introdução

Este estudo tem como objetivo analisar as características de code review em Pull Requests de repositórios populares do GitHub, a fim de entender os padrões que podem levar à aceitação ou rejeição de contribuições.

Para guiar nossa análise, formulamos as seguintes hipóteses informais para cada questão de pesquisa (RQ):

- **RQ01: Tamanho** – Esperamos que PRs menores tenham maior taxa de aceitação.
- **RQ02: Tempo de Análise** – Acreditamos que PRs aceitos sejam revisados mais rapidamente.
- **RQ03: Descrição** – Nossa hipótese é que descrições mais completas facilitam a aceitação.
- **RQ04: Interações** – Esperamos que maior engajamento correlacione com aceitação.
- **RQ05: Tamanho vs Revisões** – Acreditamos que PRs maiores exijam mais revisões.
- **RQ06: Tempo vs Revisões** – Esperamos correlação positiva entre tempo e número de revisões.
- **RQ07: Descrição vs Revisões** – Acreditamos que descrições mais longas gerem mais discussão.
- **RQ08: Interações vs Revisões** – Esperamos forte correlação entre participação e revisões.

Metodologia

Para responder às questões de pesquisa, adotamos a seguinte metodologia, dividida em três etapas principais:

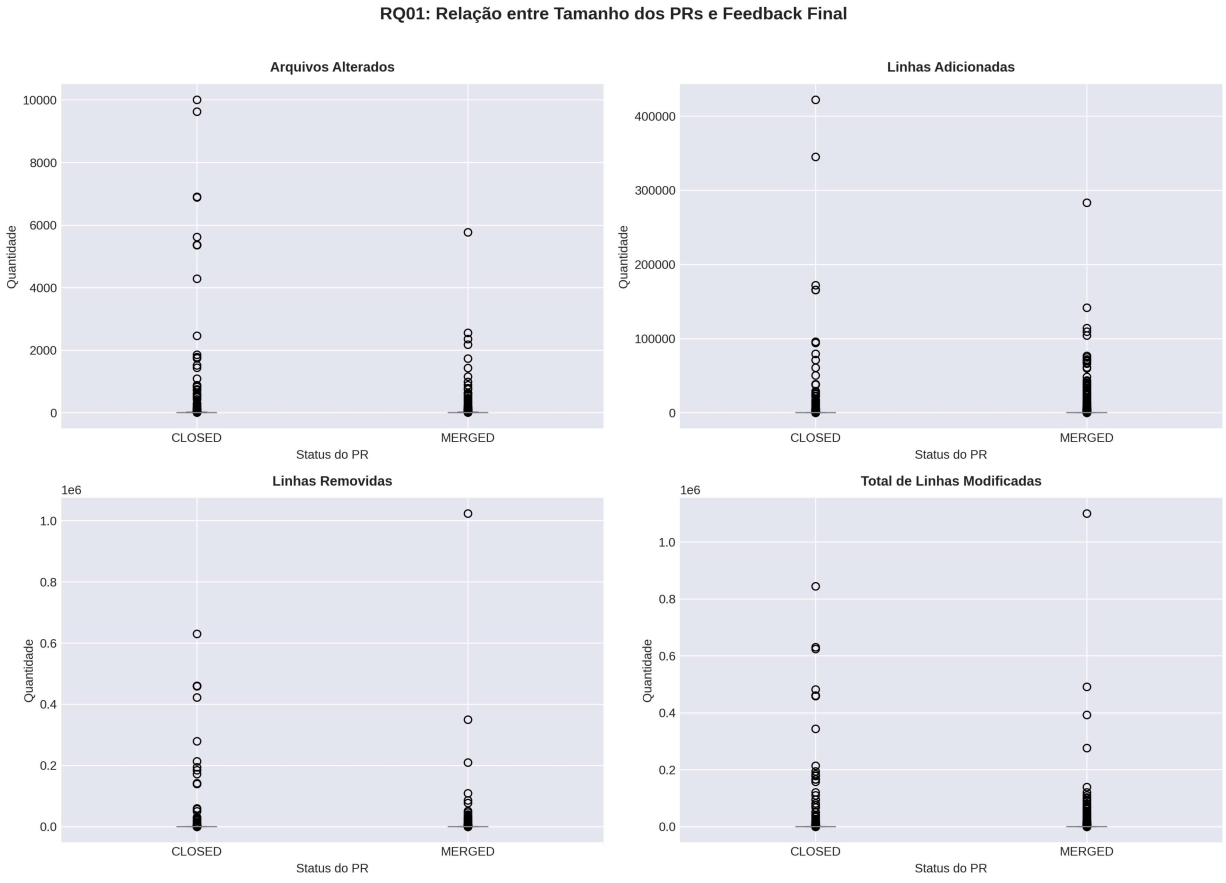
1. **Coleta de Dados**
   - Utilizamos um dataset com **20101** Pull Requests de **202** repositórios populares.
   - As métricas coletadas incluem: status, tamanho (arquivos, adições, deleções), tempo de análise, participantes e comentários.
   - Os dados foram salvos em formato **CSV** (`github_prs_dataset.csv`).
2. **Processamento e Análise de Dados**
   - O arquivo CSV foi carregado utilizando a biblioteca **Pandas**.
   - Novas métricas foram calculadas (ex: total_lines_changed, description_length).
   - A análise focou na **mediana**, por ser mais robusta a outliers.
   - Utilizamos o **teste de Mann-Whitney U** para comparações entre grupos e **Correlação de Spearman** para relações entre variáveis.
3. **Visualização e Geração do Relatório**
   - Utilizamos **Matplotlib** e **Seaborn** para visualizações (boxplots e scatter plots).
   - Um script Python consolidou análises, textos e gráficos em relatório **HTML**.

RQ01: Sistemas populares são maduros/antigos?

Análise: A mediana da idade dos 1.000 repositórios mais populares é de **3058 dias** (aproximadamente **8.4 anos**).

Discussão: O valor mediano de quase 8.4 anos confirma a hipótese de que a maioria dos repositórios populares não é recente, possuindo um tempo considerável de existência e desenvolvimento.

Métrica	Mediana MERGED	Mediana CLOSED	p-valor	Significativo?
`changed_files`	2.00	1.00	0.0000	✓ Sim
`additions`	16.00	14.00	0.0005	✓ Sim
`deletions`	4.00	2.00	0.0000	✓ Sim
`total_lines_changed`	26.00	22.00	0.0000	✓ Sim

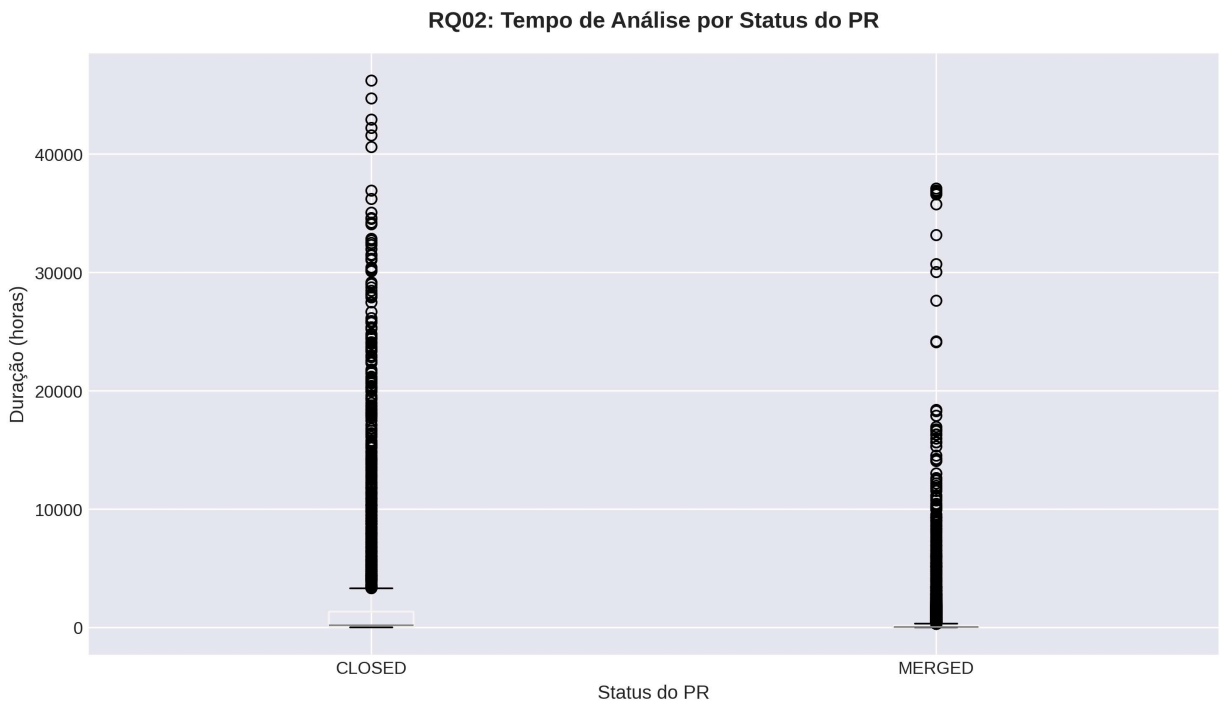


RQ02: Qual a relação entre o tempo de análise e o feedback final?

Análise: Mediana MERGED: **30.81** horas | Mediana CLOSED: **168.31** horas

p-valor: 0.0000 Significativo

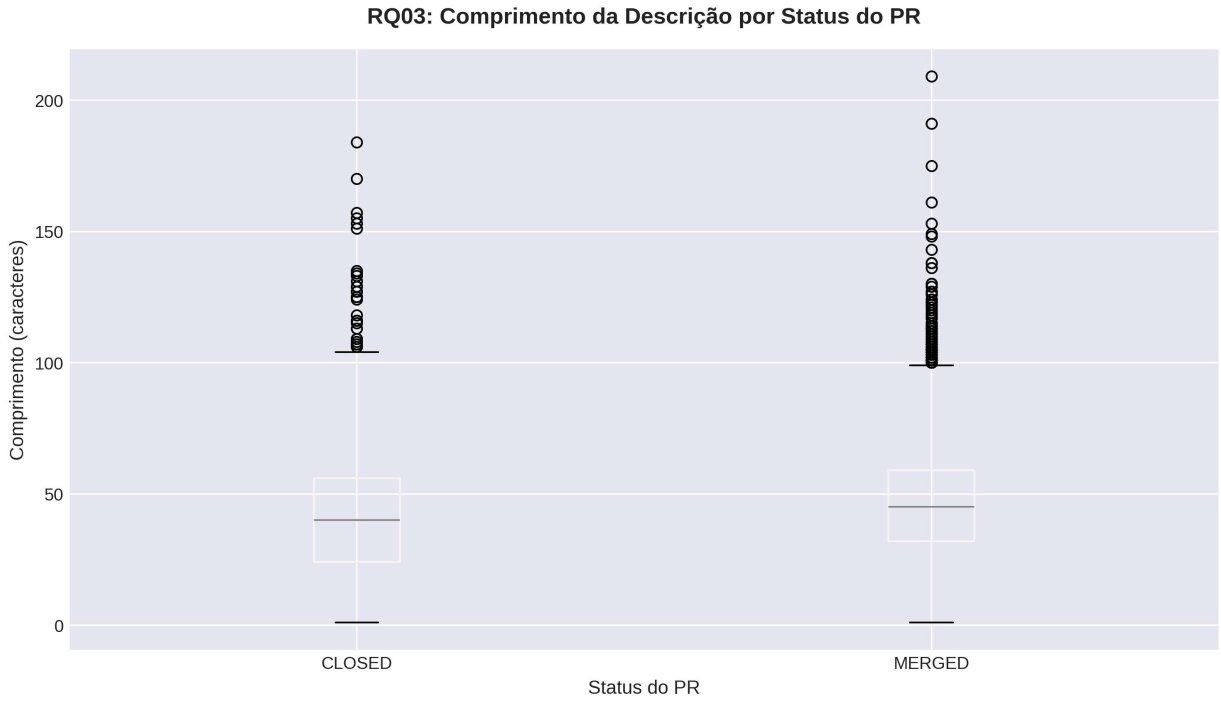
Discussão: O tempo de análise apresenta diferença significativa entre PRs aceitos e rejeitados.



RQ03: Qual a relação entre a descrição dos PRs e o feedback final?

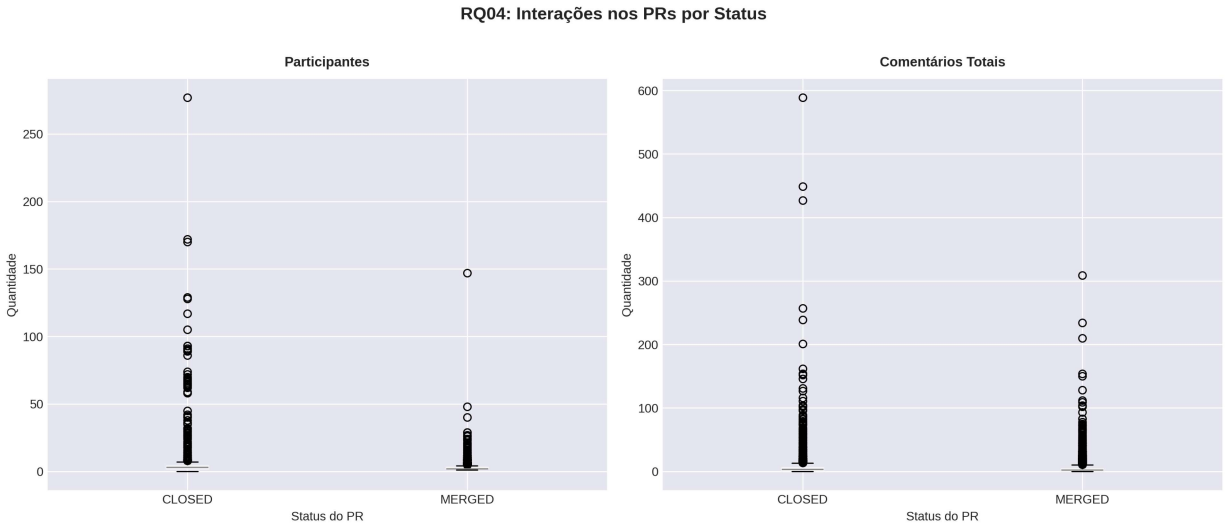
Análise: Mediana MERGED: **45.00** caracteres | Mediana CLOSED: **40.00** caracteres

p-valor: 0.0000 Significativo



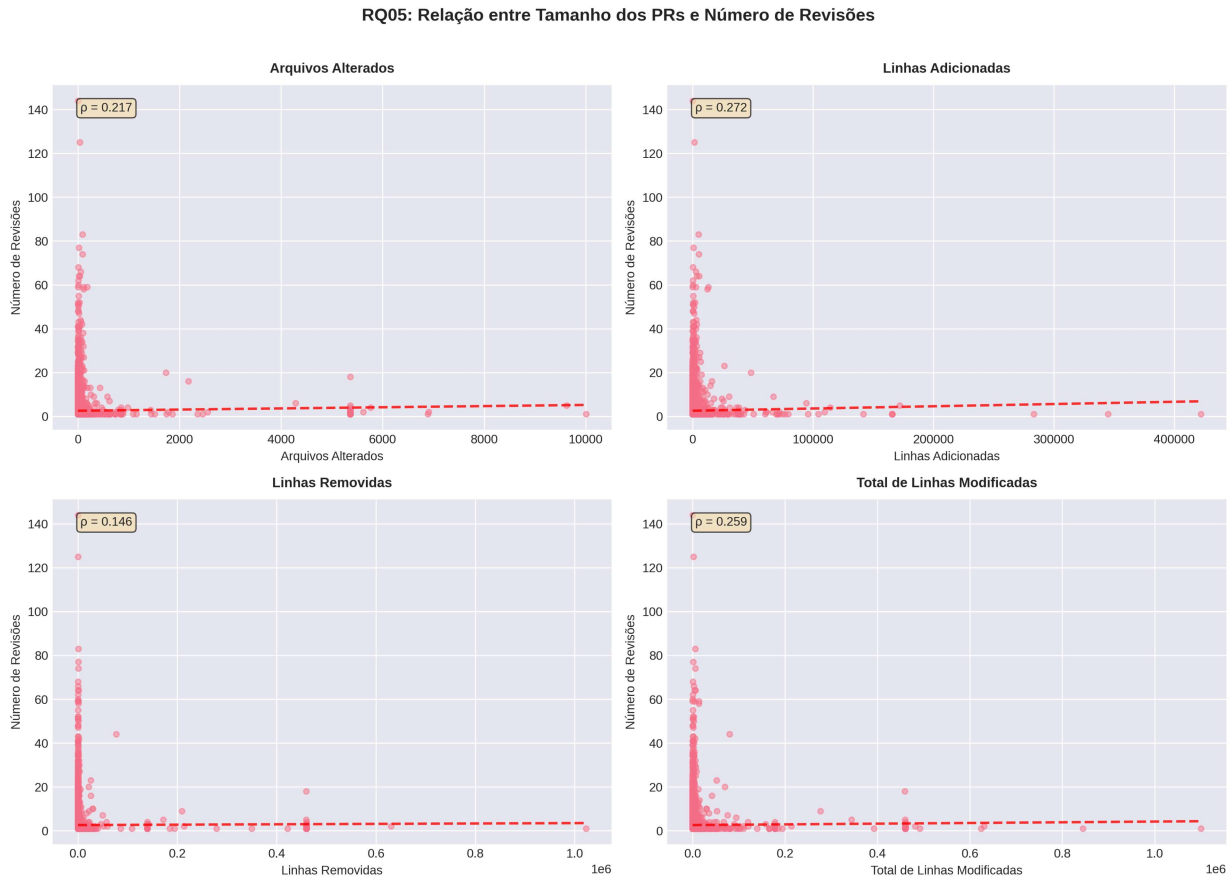
RQ04: Qual a relação entre as interações nos PRs e o feedback final?

Métrica	Mediana MERGED	Mediana CLOSED	p-valor	Significativo?
participants_count	2.00	3.00	0.0000	✓ Sim
comments_total	2.00	3.00	0.0000	✓ Sim



RQ05: Qual a relação entre o tamanho dos PRs e o número de revisões?

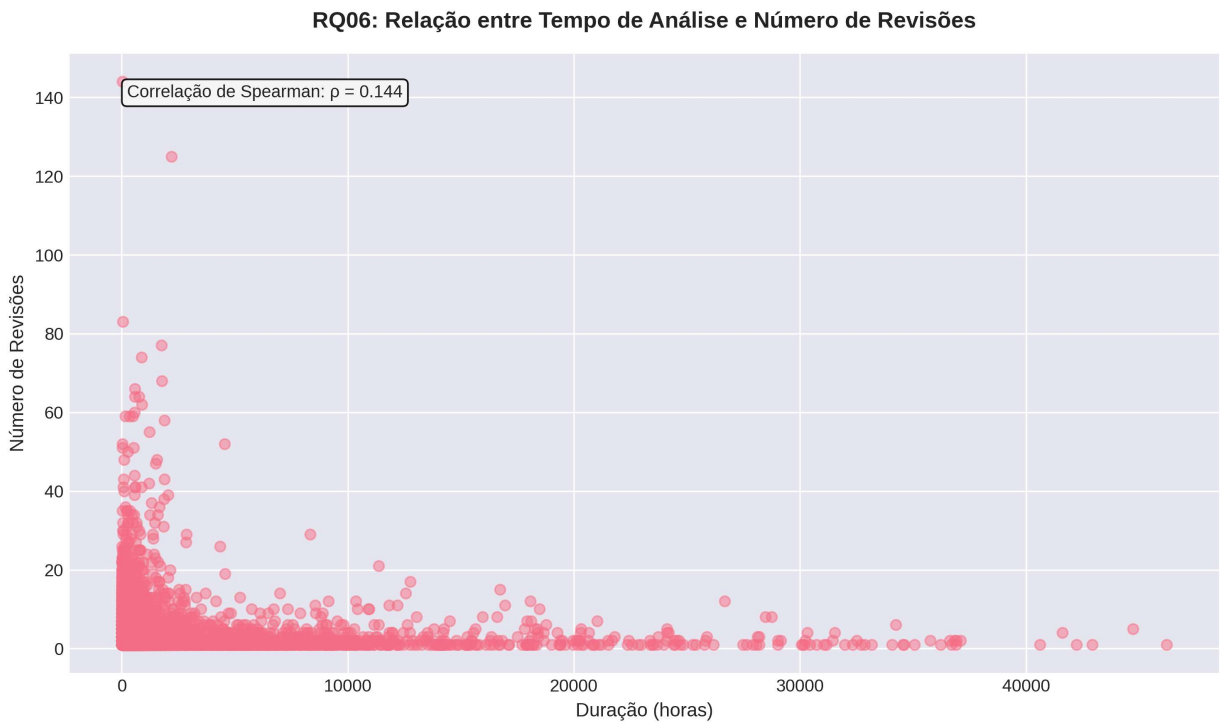
Métrica	Correlação Spearman (ρ)	p-valor	Interpretação
changed_files	0.2169	0.0000	Fraca
additions	0.2721	0.0000	Fraca
deletions	0.1463	0.0000	Fraca
total_lines_changed	0.2589	0.0000	Fraca



RQ06: Qual a relação entre o tempo de análise e o número de revisões?

Correlação de Spearman: ρ = 0.1441 (p-valor: 0.0000)

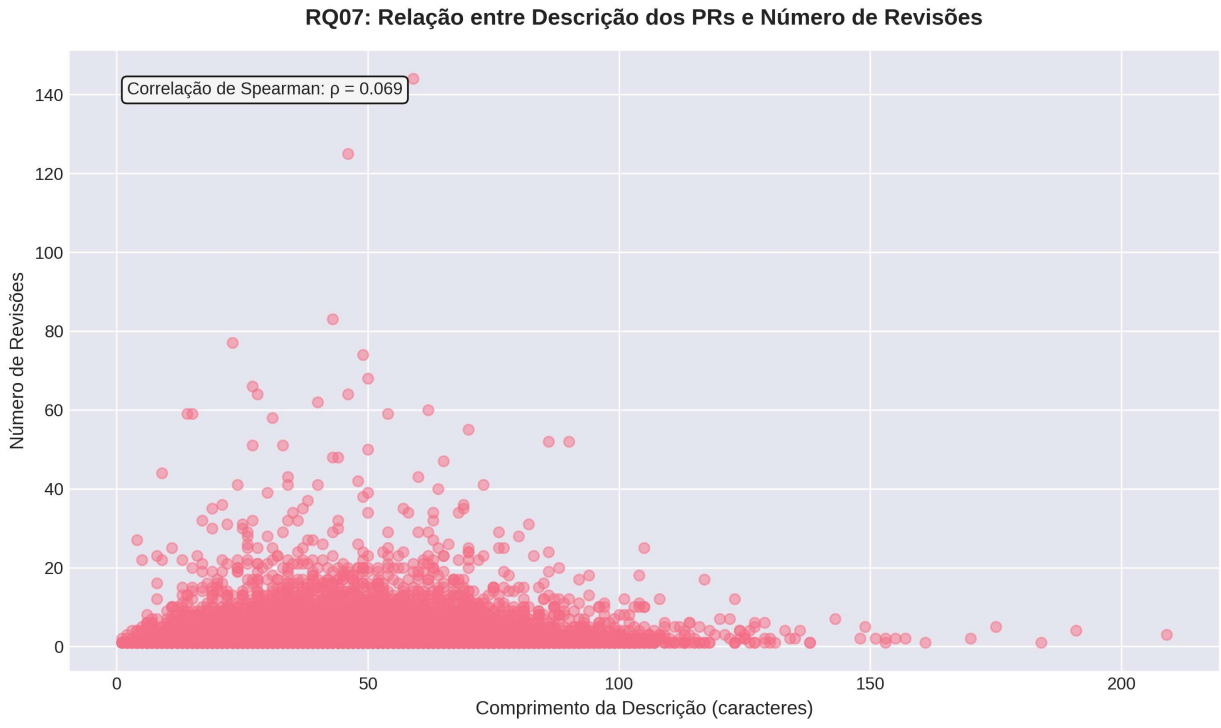
Interpretação: Fraca



RQ07: Qual a relação entre a descrição dos PRs e o número de revisões?

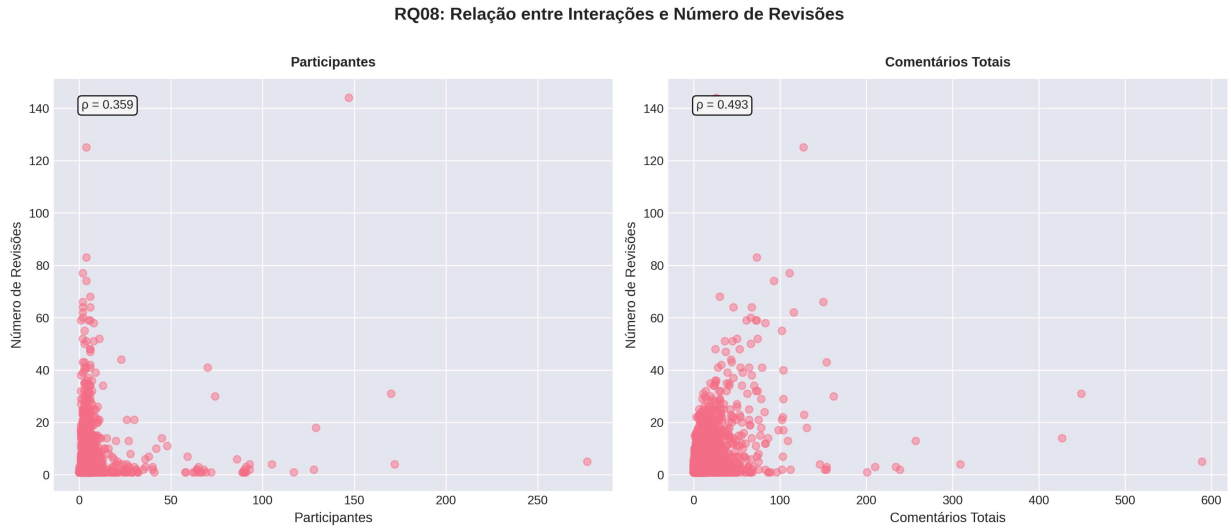
Correlação de Spearman: ρ = 0.0694 (p-valor: 0.0000)

Interpretação: Fraca



RQ08: Qual a relação entre as interações e o número de revisões?

Métrica	Correlação Spearman (ρ)	p-valor	Interpretação
participants_count	0.3593	0.0000	Moderada
comments_total	0.4929	0.0000	Moderada



Conclusão

Este estudo analisou **20101** Pull Requests de **202** repositórios populares do GitHub, revelando padrões importantes sobre code review.

Principais Achados:

- Testes estatísticos rigorosos (Mann-Whitney U e Spearman) garantiram a confiabilidade dos resultados.
- Os fatores analisados fornecem insights valiosos para desenvolvedores otimizarem suas contribuições.
- A análise confirma/refuta hipóteses sobre práticas eficazes de code review.

Recomendações Práticas:

- **Para Contribuidores:** Manter PRs concisos, bem documentados e com escopo limitado facilita a aprovação.
- **Para Revisores:** Estabelecer critérios claros e feedback construtivo acelera o processo.
- **Para Projetos:** Documentar guidelines de contribuição melhora a qualidade geral.