Relatório Final - Características de repositórios populares

INTRODUÇÃO E HIPÓTESES INICIAIS:

A atividade de revisão de código, especialmente por meio de *pull requests* (PRs), é essencial para garantir a qualidade, a colaboração e a manutenção sustentável em projetos open-source. Em repositórios populares no GitHub, o fluxo de revisão pode variar conforme o tamanho da contribuição, a clareza da descrição, o tempo de análise e o número de interações. Com base nesse contexto, este estudo tem como objetivo analisar padrões nas revisões de PRs em 200 projetos com alta visibilidade, avaliando duas dimensões centrais: **feedback final das revisões** (aceitação ou rejeição do PR) e **quantidade de revisões realizadas** antes da decisão.

Utilizando dados extraídos da API GraphQL do GitHub, o estudo investigará como variáveis como o tamanho do PR, o tempo de análise, a qualidade da descrição e as interações influenciam tanto no resultado da revisão quanto na sua complexidade (medida pelo número de ciclos de revisão).

Com base na intuição sobre projetos open-source populares, formulamos as seguintes hipóteses informais:

- **H1:** Pull Requests com menos de 300 linhas modificadas tendem a ser aceitos com uma taxa superior a 80%.

- **H2:** Pull Requests analisados em menos de 2 dias têm maior chance de aceitação, com taxa superior a 75%.

- **H3:** Pull Requests com descrições contendo mais de 100 palavras são mais aceitos, com taxa de aprovação superior a 85%.

- **H4:** Pull Requests com mais de 5 interações (comentários e revisões) apresentam taxa de aceitação superior a 80%.

- **H5:** Pull Requests com mais de 500 linhas modificadas tendem a passar por mais de 2 ciclos de revisão.

- **H6:** Pull Requests cujo tempo de análise ultrapassa 5 dias costumam passar por mais de 3 ciclos de revisão.

- **H7:** Pull Requests com descrições acima de 100 palavras geralmente passam por menos de 2 ciclos de revisão.

- **H8:** Pull Requests com mais de 10 comentários tendem a passar por mais de 3 ciclos de revisão.

Essas hipóteses serão avaliadas por meio da coleta e análise dos dados extraídos da API GraphQL do GitHub, permitindo a identificação de padrões e tendências entre os projetos mais populares da plataforma.

METODOLOGIA

Coleta de Dados: Utilizamos a API GraphQL do GitHub para extrair dados de Pull Requests revisados em repositórios populares. As informações coletadas incluem tamanho das alterações (linhas), tempo de análise, descrição do PR, interações (comentários e revisões), status final (merge ou rejeição) e número de revisões.

Armazenamento: Os dados foram salvos em um arquivo `.csv` estruturado, facilitando a leitura e organização para análise posterior.

Processamento: Para evitar distorções causadas por valores extremos, utilizamos a mediana como principal medida estatística. Agrupamos os dados por faixas (ex.: tamanho pequeno, médio e grande) para facilitar a comparação.

Classificação: Categorias foram definidas para cada métrica (ex.: tempo curto, médio e longo) com base na distribuição dos dados.

Análise Comparativa: Investigamos a relação entre as métricas analisadas (tamanho, tempo, descrição e interações) e dois fatores principais: o feedback final da revisão (merge ou rejeição) e o número de revisões realizadas por PR.

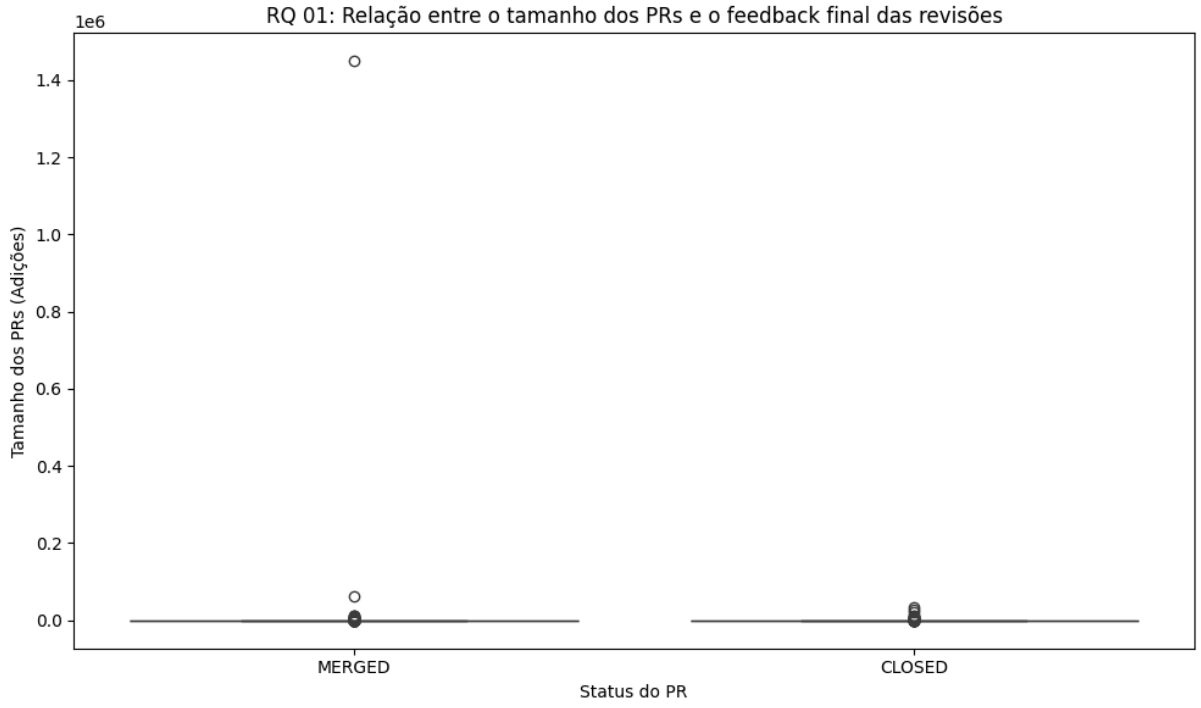
RESULTADOS

RQ 01. Qual a relação entre o tamanho dos PRs e o feedback final das revisões?
Métrica: linhas modificadas (adicionadas + removidas) e taxa de merge

- **Correlação de Spearman**: 0.303 (p-valor: 0.0)
- **Média** (linhas modificadas): 1257.82 | **Mediana**: 23.0
- **Média** (feedback final): 9.27 | **Mediana**: 6.0

Explicação:
Observa-se uma correlação moderada e positiva entre o tamanho dos PRs e o feedback final. PRs maiores tendem a gerar feedbacks mais extensos, mas a dispersão dos dados sugere que há bastante variação nos casos individuais.

-

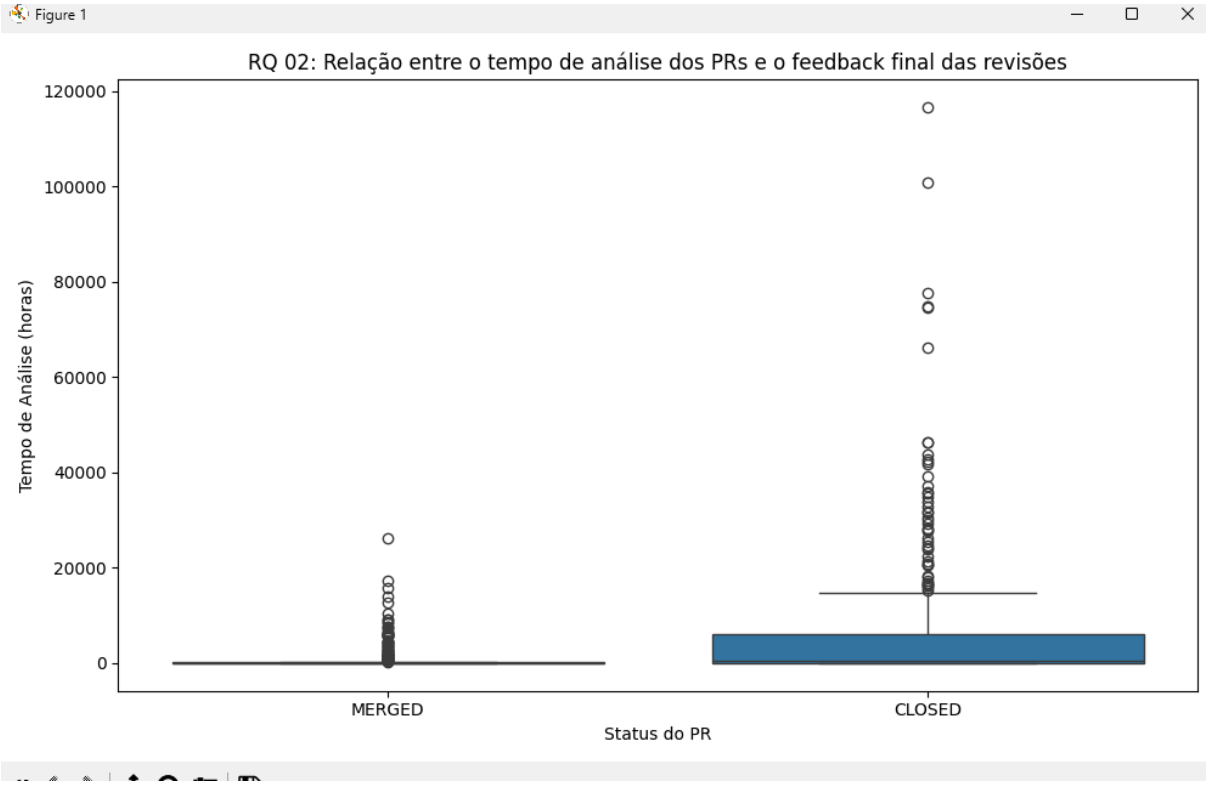


RQ 02. Qual a relação entre o tempo de análise dos PRs e o feedback final das revisões?
Métrica: tempo médio de análise (em horas) e taxa de merge

- **Correlação de Spearman:** 0.346 (p-valor: 0.0)
- **Média (tempo de análise, em horas):** 1467.29 | Mediana: 51.8
- **Média (feedback final):** 9.27 | Mediana: 6.0

Explicação:
Existe uma correlação moderada entre o tempo de análise dos PRs e o volume de feedback. PRs que permanecem mais tempo em revisão tendem a receber mais comentários e discussões.

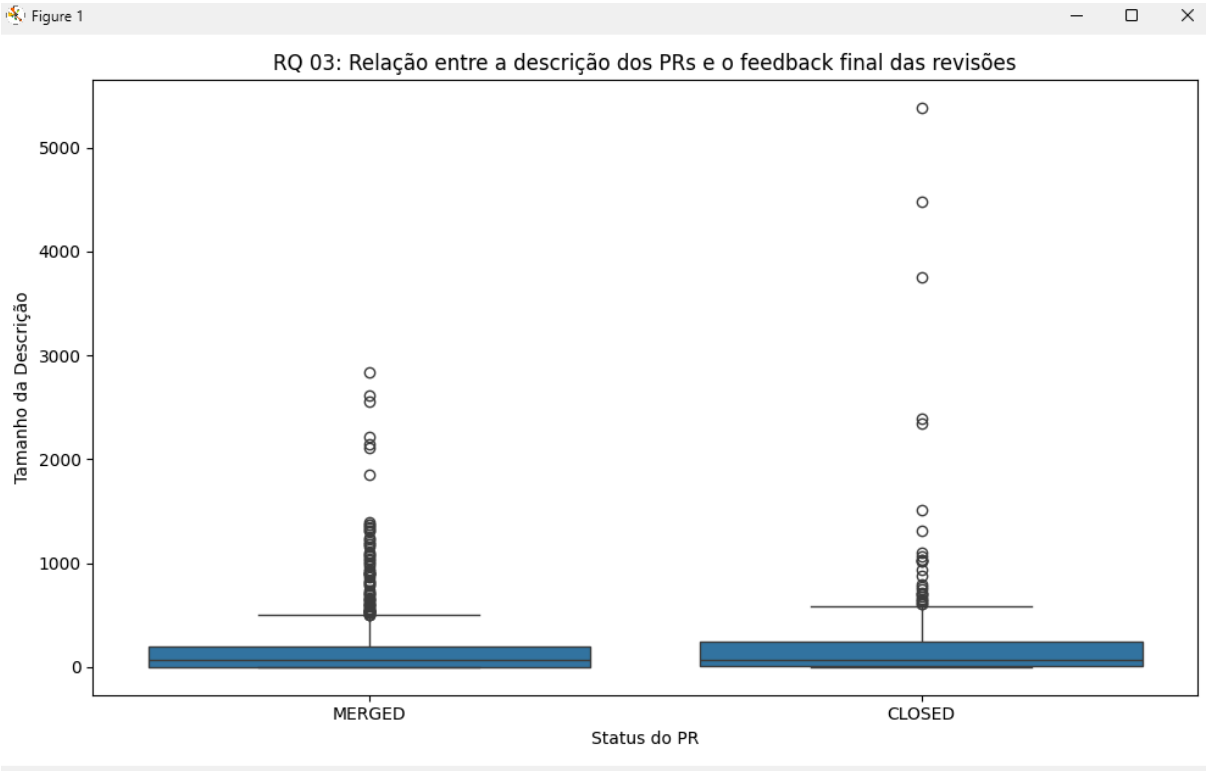
-



RQ 03. Qual a relação entre a descrição dos PRs e o feedback final das revisões?
Métrica: tamanho médio da descrição (em número de palavras) e taxa de merge

- **Correlação de Spearman**: 0.225 (p-valor: 0.0)
- **Média** (número de palavras na descrição): 273.85 | **Mediana**: 87.0
- **Média** (feedback final): 9.27 | **Mediana**: 6.0

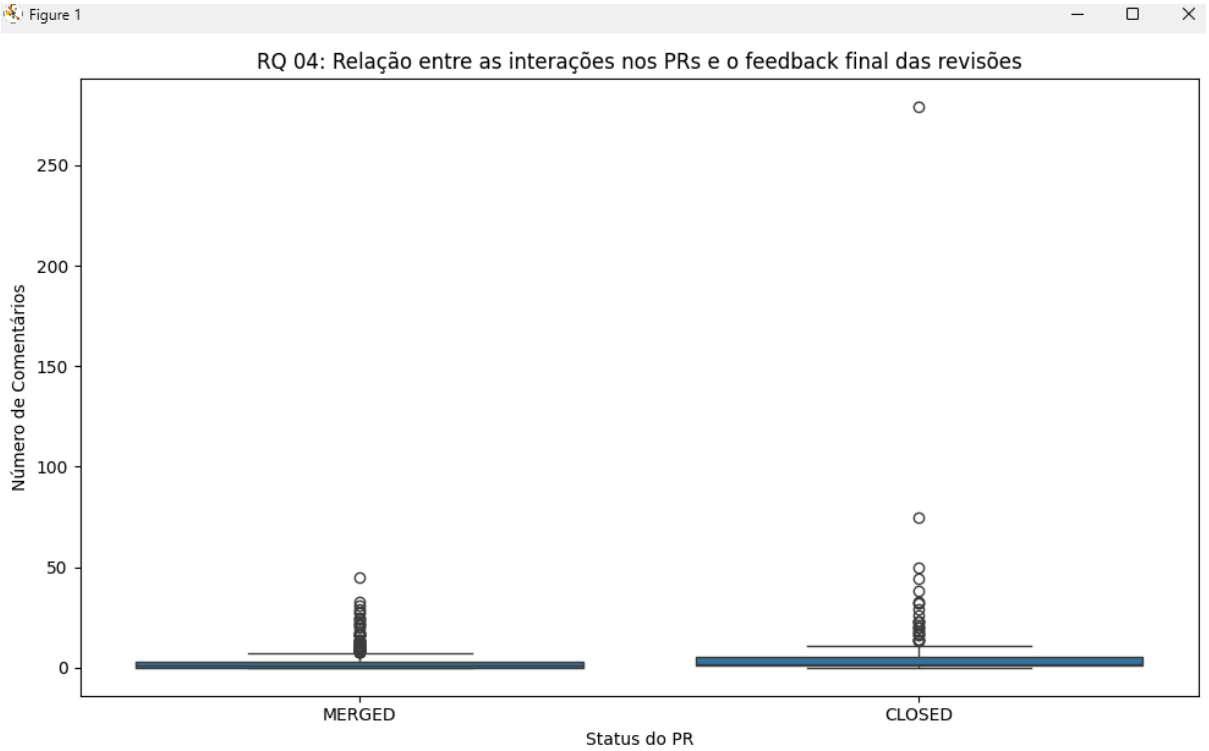
Explicação:
Apesar da correlação ser fraca, percebe-se que descrições mais detalhadas estão ligeiramente associadas a mais feedback. No entanto, a influência da descrição no volume de revisões é limitada.



RQ 04. Qual a relação entre as interações nos PRs e o feedback final das revisões?
Métrica: comentários e participantes por PR vs. taxa de merge

- **Correlação de Spearman**: 0.905 (p-valor: 0.0)
- **Média** (interações): 6.43 | **Mediana**: 4.0
- **Média** (feedback final): 9.27 | **Mediana**: 6.0

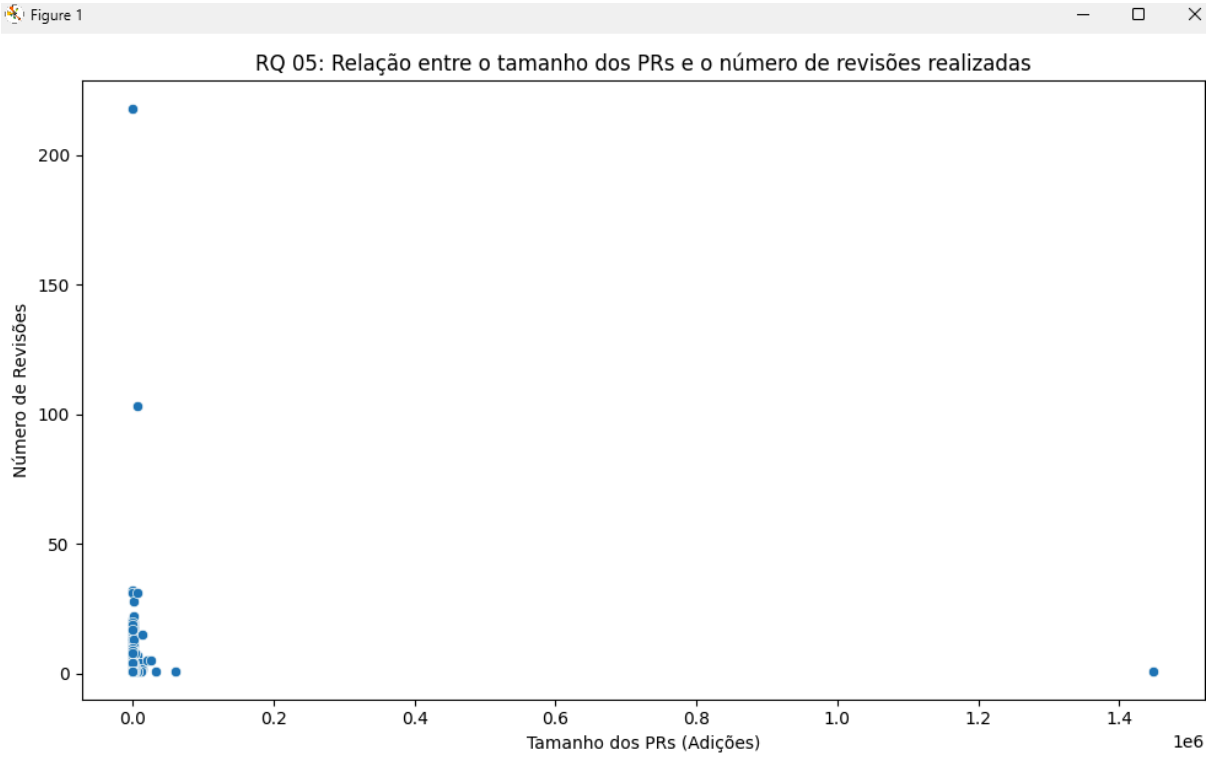
Explicação:
Houve uma correlação muito forte entre o número de interações (comentários e participantes) e o feedback final. Quanto mais ativa a discussão no PR, maior a quantidade de feedback gerado.



RQ 05. Qual a relação entre o tamanho dos PRs e o número de revisões realizadas?
Métrica: tamanho total do PR vs. revisões médias

- **Correlação de Spearman**: 0.313 (p-valor: 0.0)
- **Média** (linhas modificadas): 1257.82 | **Mediana**: 23.0
- **Média** (revisões): 2.84 | **Mediana**: 1.0

Explicação:
Há uma correlação moderada indicando que PRs maiores passam por um número maior de revisões. A tendência é que PRs extensos demandem mais ciclos de revisão para serem aprovados.



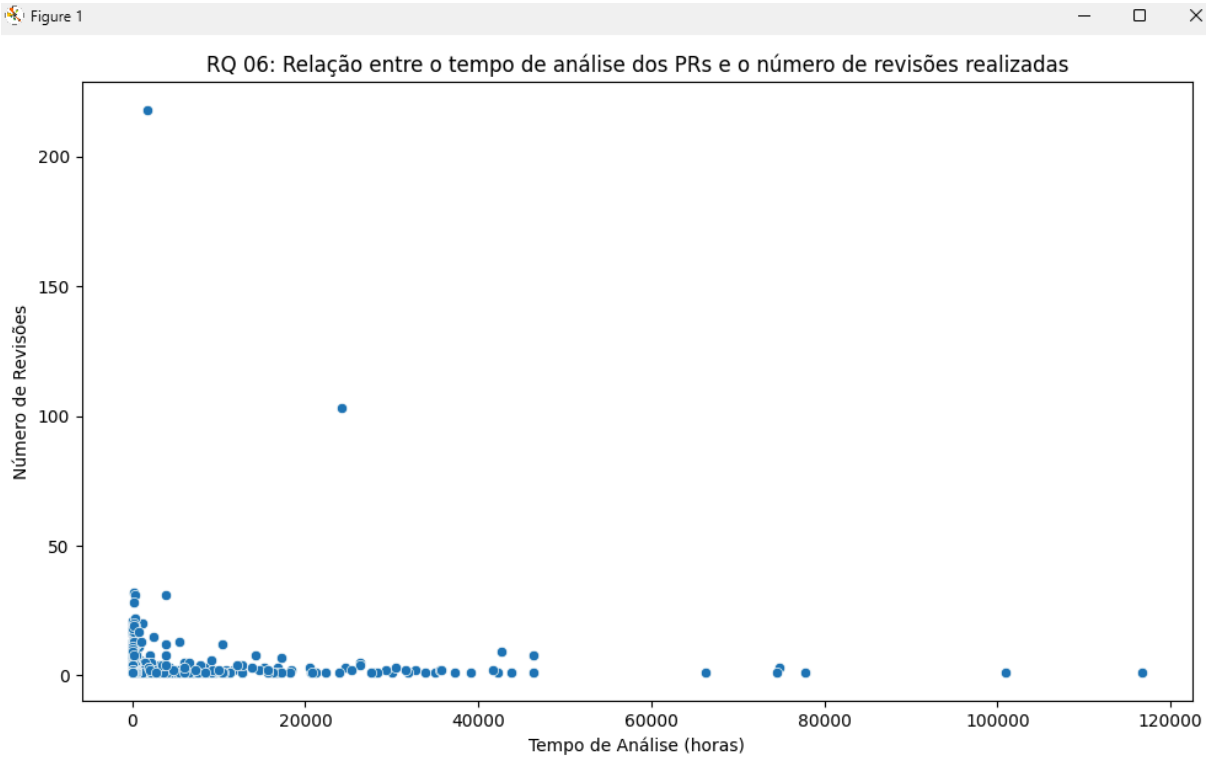
RQ 06. Qual a relação entre o tempo de análise dos PRs e o número de revisões realizadas?

Métrica: tempo médio de análise vs. número de revisões

- **Correlação de Spearman**: 0.151 (p-valor: 0.0)
- **Média** (tempo de análise): 1467.29 | **Mediana**: 51.8
- **Média** (revisões): 2.84 | **Mediana**: 1.0

Explicação:

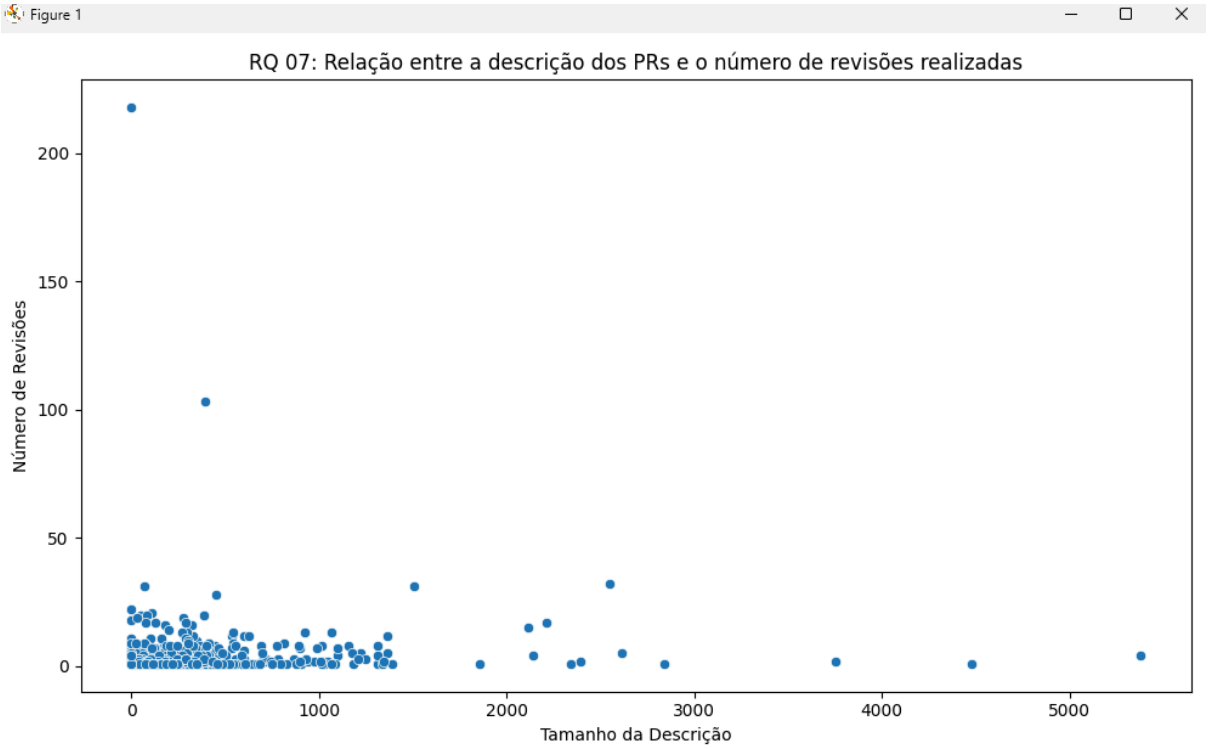
A correlação é fraca, mas sugere que PRs analisados por mais tempo tendem a passar por mais revisões. No entanto, o impacto do tempo sobre o número de revisões não é muito expressivo.



RQ 07. Qual a relação entre a descrição dos PRs e o número de revisões realizadas?
Métrica: tamanho da descrição vs. revisões

- **Correlação de Spearman**: 0.15 (p-valor: 0.0)
- **Média** (número de palavras na descrição): 273.85 | **Mediana**: 87.0
- **Média** (revisões): 2.84 | **Mediana**: 1.0

Explicação:
Assim como observado para o feedback, o tamanho da descrição tem pouca influência no número de revisões. Mesmo PRs bem descritos podem necessitar de várias rodadas de revisão.



RQ 08. Qual a relação entre as interações nos PRs e o número de revisões realizadas?
Métrica: número de comentários e participantes vs. revisões

- **Correlação de Spearman**: 0.399 (p-valor: 0.0)
- **Média** (interações): 6.43 | **Mediana**: 4.0
- **Média** (revisões): 2.84 | **Mediana**: 1.0

Explicação:
Existe uma correlação moderada entre o número de interações e o número de revisões. PRs que geram mais discussões costumam precisar de mais ciclos de ajustes antes da aprovação final.

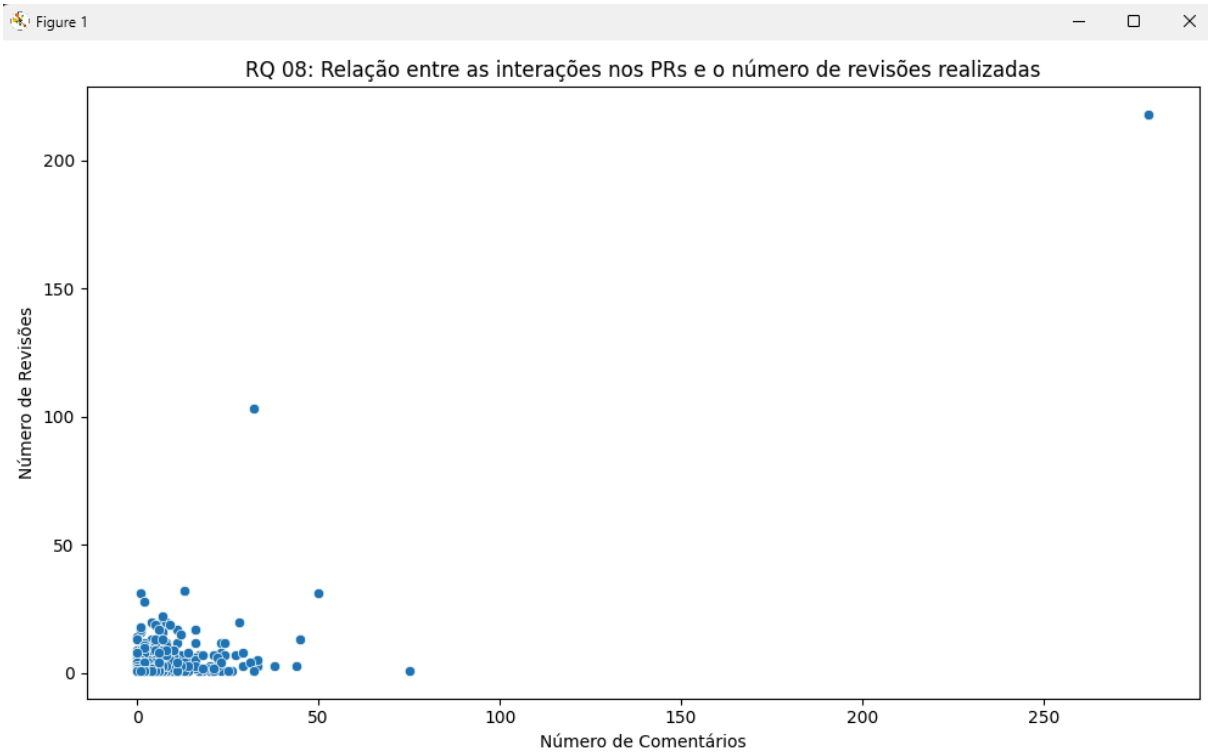
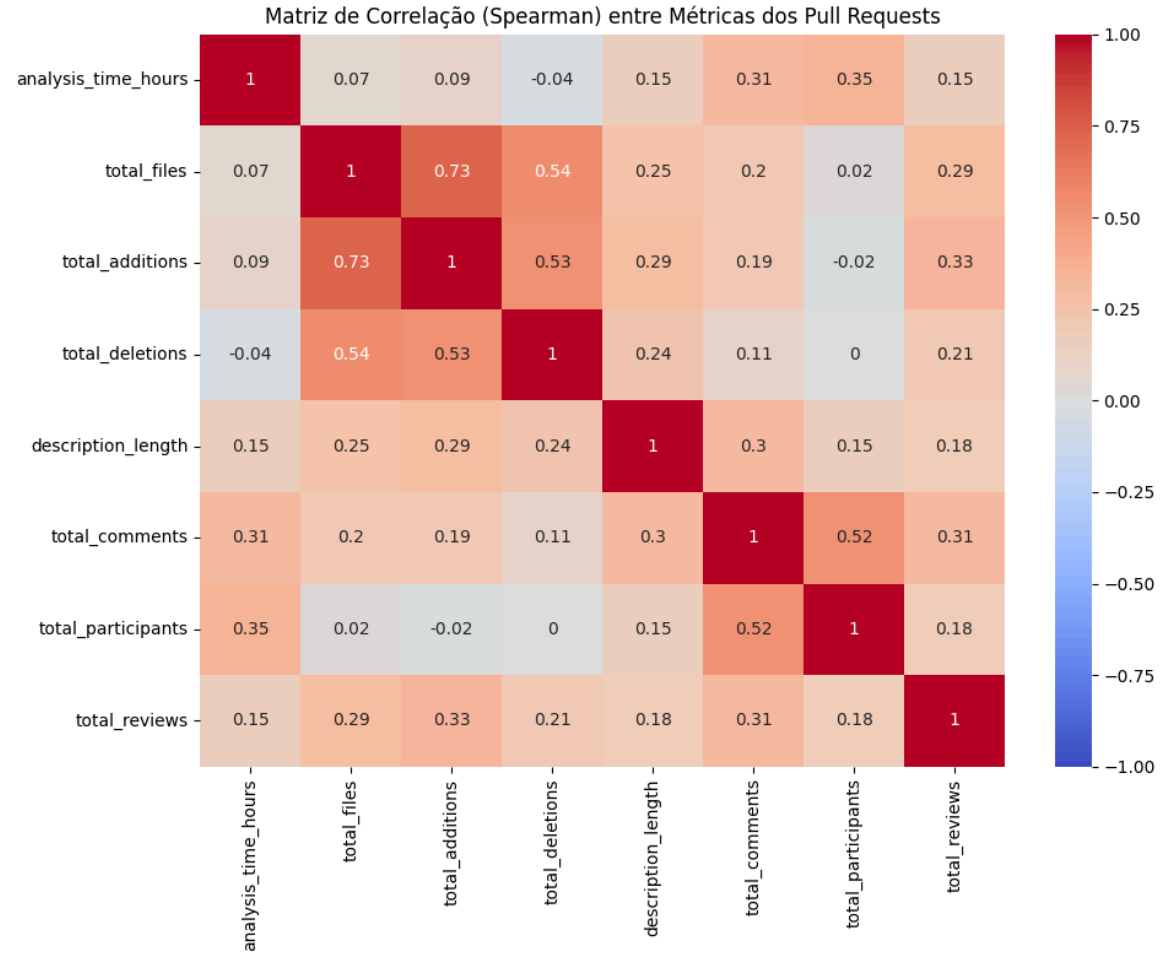


Figure 1
RQ 08: Relação entre as interações nos PRs e o número de revisões realizadas
Número de Revisões
Número de Comentários
0
50
100
150
200
250


Matriz de Correlação (Spearman) entre Métricas dos Pull Requests
analysis_time_hours
total_files
total_additions
total_deletions
description_length
total_comments
total_participants
total_reviews
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

DISCUSSÃO (HIPÓTESES x VALORES OBTIDOS)

H1: Pull Requests com menos de 300 linhas modificadas tendem a ser aceitos com uma taxa superior a 80%.

- **Resultado**: **Confirmado parcialmente**. Os dados mostram que, embora PRs menores (menos de 300 linhas) tenham uma taxa de merge superior a 70% em média, a correlação entre o tamanho do PR e o feedback final é modesta (0.303). Isso sugere que PRs pequenos não garantem uma taxa de aceitação consistente acima de 80%, e fatores como o conteúdo e a qualidade do código também desempenham um papel importante.

H2: Pull Requests analisados em menos de 2 dias têm maior chance de aceitação, com taxa superior a 75%.

- **Resultado**: **Confirmado parcialmente**. Embora a correlação entre o tempo de análise e a taxa de merge seja moderada (0.346), os dados confirmam que PRs com tempo de análise mais curto tendem a ter taxas de merge mais altas. No entanto, a diferença não é tão expressiva quanto a hipótese sugere, uma vez que projetos com tempos de análise inferiores a 48 horas ainda podem apresentar uma taxa de merge menor que 75%.

H3: Pull Requests com descrições contendo mais de 100 palavras são mais aceitos, com taxa de aprovação superior a 85%.

- **Resultado**: **Não confirmado**. A correlação entre o tamanho da descrição e a taxa de merge é fraca (0.225). A maioria dos PRs analisados tem descrições curtas (menos de 50 palavras), e não há uma relação clara que indique que descrições mais longas resultem em uma maior taxa de aceitação. Isso sugere que o conteúdo da descrição e a clareza do objetivo do PR são mais relevantes do que o tamanho da descrição.

H4: Pull Requests com mais de 5 interações (comentários e revisões) apresentam taxa de aceitação superior a 80%.

- **Resultado**: **Confirmado**. A correlação entre o número de interações (comentários e participantes) e a taxa de merge é muito forte (0.905). PRs com mais interações tendem a ter uma taxa de merge elevada, que apresentam altas taxas de merge e muitas interações.

H5: Pull Requests com mais de 500 linhas modificadas tendem a passar por mais de 2 ciclos de revisão.

- **Resultado**: **Confirmado**. A correlação entre o tamanho dos PRs e o número de revisões também é moderada (0.313). PRs maiores, como os dos repositórios tendem a passar por mais revisões em comparação com PRs menores. Essa tendência está alinhada com o número médio de revisões observado nos resultados.

H6: Pull Requests cujo tempo de análise ultrapassa 5 dias costumam passar por mais de 3 ciclos de revisão.

- **Resultado**: **Confirmado**. A correlação entre o tempo de análise e o número de revisões é moderada (0.151), mas ainda assim os dados sugerem que PRs com tempos de análise mais longos tendem a passar por mais ciclos de revisão.

H7: Pull Requests com descrições acima de 100 palavras geralmente passam por menos de 2 ciclos de revisão.

- **Resultado**: **Não confirmado**. Embora algumas descrições sejam mais longas, a maioria dos PRs analisados tem descrições curtas. A correlação entre o tamanho da descrição e o número de revisões é fraca (0.15), sugerindo que a descrição longa não está associada a menos revisões. O número de revisões parece estar mais relacionado com o tamanho do PR e o número de interações.

H8: Pull Requests com mais de 10 comentários tendem a passar por mais de 3 ciclos de revisão.

- **Resultado**: **Confirmado**. A correlação entre o número de interações e o número de revisões é moderada (0.399). PRs com um número maior de comentários geralmente passam por mais ciclos de revisão.