



Pontifícia Universidade Católica de Minas Gerais

Campus Coração Eucarístico
Engenharia de Software

Trabalho apresentado à disciplina de
Laboratório de Experimentação de Software
Prof. João Paulo Carneiro Aramuni

Contagem – MG

Sumário

1	Introdução	2
2	Metodologia	3
2.1	Criação do Dataset	3
2.2	Definição e Coleta de Métricas	3
2.3	Análise Estatística	4
3	Resultados e Análise	4
3.1	A. Análise do Feedback Final das Revisões (Status do PR)	4
3.2	B. Análise do Número de Revisões	5
4	Análise da Relação com o Tamanho do PR	5
4.1	RQ01: Relação entre Tamanho e Status do PR	6
4.2	RQ05: Relação entre Tamanho e Número de Revisões	6
5	Análise da Relação com o Tempo de Análise do PR	7
5.1	RQ02: Relação entre Tempo de Análise e Status do PR	7
5.2	RQ06: Relação entre Tempo de Análise e Número de Revisões	8
6	Análise da Relação com a Descrição do PR	9
6.1	RQ03: Relação entre Descrição e Status do PR	9
6.2	RQ07: Relação entre Descrição e Número de Revisões	10
7	Análise da Relação com as Interações no PR	11
7.1	RQ04: Relação entre Interações e Status do PR	11
7.2	RQ08: Relação entre Interações e Número de Revisões	11
8	Conclusão	12

1 Introdução

A revisão de código (*code review*) é uma prática fundamental nos processos modernos de desenvolvimento de software, especialmente no ecossistema de código aberto (*open-source*). Realizada tipicamente por meio de *Pull Requests* (PRs) em plataformas como o GitHub, essa atividade visa garantir a qualidade, a manutenibilidade e a corretude do código antes de sua integração à base principal do projeto. Um processo de revisão eficaz pode prevenir a introdução de defeitos e alinhar as contribuições aos padrões arquiteturais e de estilo definidos pela equipe.

Neste contexto, este estudo busca caracterizar a atividade de *code review* em repositórios populares do GitHub. O objetivo principal é analisar empiricamente quais variáveis do processo de submissão e revisão influenciam o resultado final de um PR (se ele é aceito ou rejeitado) e o esforço de revisão (medido pelo número de revisões). Para isso, foi realizada uma análise em larga escala sobre PRs extraídos de repositórios de alta popularidade.

Para guiar esta investigação, o trabalho se propõe a responder às seguintes questões de pesquisa (RQs), divididas em duas dimensões principais:

A. Feedback Final das Revisões (Status do PR)

- **RQ 01:** Qual a relação entre o tamanho dos PRs e o feedback final das revisões?
- **RQ 02:** Qual a relação entre o tempo de análise dos PRs e o feedback final das revisões?
- **RQ 03:** Qual a relação entre a descrição dos PRs e o feedback final das revisões?
- **RQ 04:** Qual a relação entre as interações nos PRs e o feedback final das revisões?

B. Número de Revisões

- **RQ 05:** Qual a relação entre o tamanho dos PRs e o número de revisões realizadas?
- **RQ 06:** Qual a relação entre o tempo de análise dos PRs e o número de revisões realizadas?
- **RQ 07:** Qual a relação entre a descrição dos PRs e o número de revisões realizadas?
- **RQ 08:** Qual a relação entre as interações nos PRs e o número de revisões realizadas?

Com base na literatura e na experiência prática, foram formuladas as seguintes hipóteses informais, que serão validadas ou refutadas pela análise de dados:

- **Hipótese para RQs 01 e 05 (Tamanho):** Acredito que PRs maiores (mais arquivos e linhas) serão mais rejeitados e exigirão mais revisões, pois sua complexidade torna a análise mais difícil e a chance de conter erros é maior.
- **Hipótese para RQs 02 e 06 (Tempo):** Hipotetizo que PRs que demoram muito para serem analisados tendem a ser rejeitados (indicando abandono ou problemas complexos) e, paradoxalmente, podem ter menos revisões se a demora indicar falta de engajamento dos revisores.

- **Hipótese para RQs 03 e 07 (Descrição):** Minha suposição é que PRs com descrições detalhadas têm maior chance de serem aceitos e podem necessitar de menos revisões, pois o autor já forneceu o contexto necessário para a análise.
- **Hipótese para RQs 04 e 08 (Interações):** Por fim, acredito que um alto número de comentários e participantes pode ser um indicativo de controvérsia ou problemas, levando a uma maior chance de rejeição e a um número maior de revisões.

2 Metodologia

Para responder às questões de pesquisa, foi adotada uma metodologia quantitativa dividida em três etapas: (1) criação do dataset por meio da coleta de dados de PRs via API do GitHub; (2) definição e extração das métricas de análise; e (3) tratamento e análise estatística dos dados. Todo o processo foi automatizado com scripts em Python.

2.1 Criação do Dataset

O objeto de estudo deste trabalho foram os *Pull Requests* (PRs) submetidos a repositórios de código aberto. A seleção da amostra seguiu um conjunto de critérios rigorosos para garantir a relevância e a qualidade dos dados:

- **Popularidade do Repositório:** Foram selecionados PRs pertencentes aos 200 repositórios mais populares do GitHub (medido pelo número de estrelas).
- **Volume de Atividade:** Apenas repositórios com um histórico de pelo menos 100 PRs (somando aceitos e fechados) foram considerados.
- **Status do PR:** Foram coletados apenas PRs com status final definido, ou seja, MERGED (aceito) ou CLOSED (rejeitado).
- **Processo de Revisão:** Para garantir que o PR passou por uma revisão humana, foram filtrados apenas aqueles com pelo menos uma revisão (*review*) registrada e cuja diferença entre a data de criação e a de finalização foi superior a uma hora, mitigando a inclusão de revisões automáticas por *bots*.

2.2 Definição e Coleta de Métricas

Para cada PR do dataset, foram extraídas métricas para representar as dimensões de análise propostas (Tamanho, Tempo, Descrição e Interações):

- **Tamanho:** Número de arquivos modificados (`size_files`); e a soma de linhas de código adicionadas e removidas (`size_additions + size_deletions`).
- **Tempo de Análise:** O intervalo, em horas, entre a criação do PR e seu fechamento ou *merge* (`analysis_time_hours`).
- **Descrição:** O número de caracteres no corpo da descrição do PR (`description_chars`).
- **Interações:** O número de participantes distintos na discussão (`interaction_participants`) e o número total de comentários (`interaction_comments`).

As duas variáveis de resultado analisadas foram o **status final do PR** e o **número total de revisões** (`reviews_count`).

2.3 Análise Estatística

Os dados coletados foram consolidados em um único arquivo CSV. Para cada questão de pesquisa, a análise foi realizada sobre o conjunto de dados completo. Conforme solicitado na especificação do laboratório, a sumarização dos dados foi feita utilizando a **mediana** como medida de tendência central, por sua robustez a valores extremos.

Para analisar as correlações, foi utilizado o teste de **correlação de Spearman**. A escolha por este teste se justifica porque ele não assume uma distribuição normal dos dados e é mais adequado para relações monotônicas (não necessariamente lineares), características comuns em dados de repositórios de software.

3 Resultados e Análise

Nesta seção, são apresentados os resultados obtidos. A análise está organizada de acordo com as duas dimensões de pesquisa: o feedback final dos PRs e o número de revisões.

3.1 A. Análise do Feedback Final das Revisões (Status do PR)

Nesta subseção, investigamos as RQs 01 a 04, comparando as métricas entre os PRs com status **MERGED** e **CLOSED**. A Tabela 1 apresenta os valores medianos para cada grupo.

Tabela 1: Valores medianos das métricas, agrupados por status do PR.

Métrica	CLOSED (Mediana)	MERGED (Mediana)
Tamanho (Arquivos)	1.00	2.00
Tamanho (Linhas Add+Del)	26.00	30.00
Tempo de Análise (Horas)	295.16	35.31
Descrição (Caracteres)	462.50	305.00
Interações (Participantes)	4.00	3.00
Interações (Comentários)	3.00	2.00

Análise para RQ01 a RQ04: A análise da Tabela 1 revela tendências claras. PRs com status **CLOSED** apresentam uma mediana de tempo de análise (295.16 horas) drasticamente superior à dos PRs aceitos (35.31 horas), sugerindo que a demora excessiva é um forte indicador de que o PR será eventualmente rejeitado ou abandonado (RQ02). Além disso, PRs rejeitados tendem a ter mais participantes (4 vs. 3) e mais comentários (3 vs. 2), o que suporta a ideia de que um maior volume de interações pode sinalizar controvérsias ou problemas que dificultam a aprovação do PR (RQ04). Curiosamente, PRs aceitos são, na mediana, ligeiramente maiores em número de arquivos e linhas de código, o que contradiz a hipótese inicial de que tamanho levaria à rejeição (RQ01).

3.2 B. Análise do Número de Revisões

Para as RQs 05 a 08, investigamos a correlação entre as métricas e o número de revisões. A Figura 1 apresenta o mapa de calor com os coeficientes de correlação de Spearman.

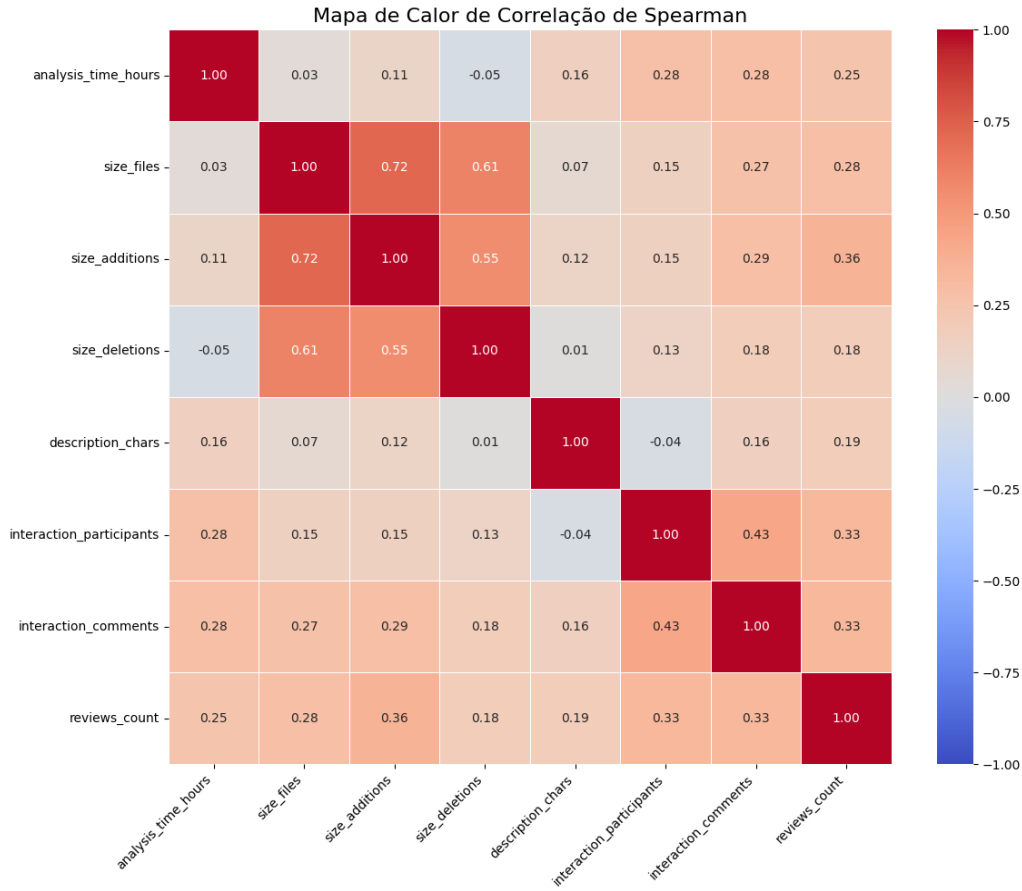


Figura 1: Mapa de calor da correlação de Spearman entre as métricas.

Análise para RQ05 a RQ08: A correlação mais forte no dataset é entre o número de comentários (`interaction_comments`) e o número de revisões (`reviews_count`), com um coeficiente de $r = 0.69$ (RQ08). Isso indica uma forte relação positiva: PRs que geram mais discussão também passam por mais ciclos de revisão. O número de participantes (`interaction_participants`) também tem uma correlação positiva forte com o número de revisões ($r = 0.58$). O tamanho do PR, medido por arquivos e linhas, também apresenta uma correlação positiva moderada com o número de revisões (RQ05), sugerindo que mudanças maiores exigem um esforço de revisão mais intenso. As demais correlações são fracas, indicando pouca relação entre tempo, descrição e o número de revisões.

4 Análise da Relação com o Tamanho do PR

Nesta seção, são investigadas as questões de pesquisa que relacionam o tamanho de um Pull Request (medido em número de arquivos e total de linhas modificadas) com seu resultado final (RQ01) o esforço de revisão (RQ05).

4.1 RQ01: Relação entre Tamanho e Status do PR

A primeira questão de pesquisa busca entender se o tamanho de um PR influencia sua aceitação. A hipótese inicial era de que PRs maiores seriam mais complexos e, portanto, teriam maior chance de serem rejeitados. Para investigar essa relação, foram gerados gráficos de box plot comparando a distribuição do tamanho para PRs com status **MERGED** e **CLOSED**, focando nos 95

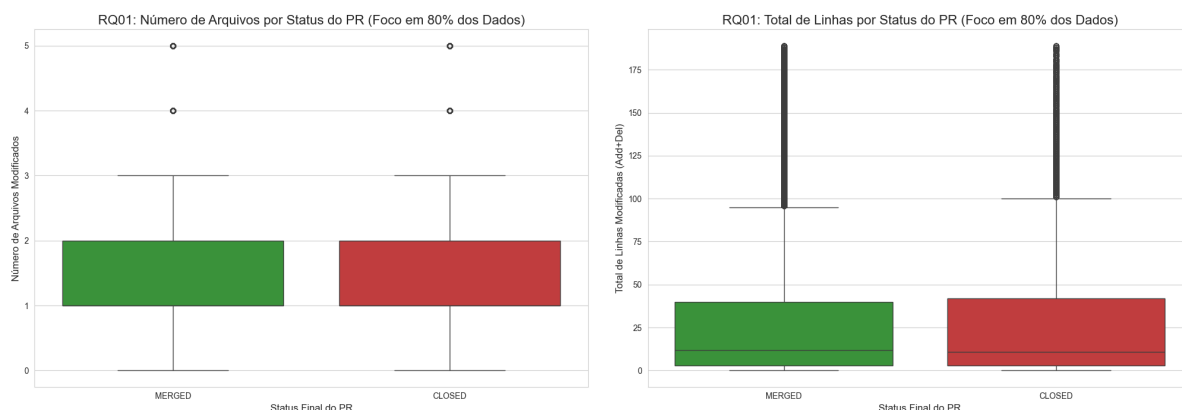


Figura 2: Distribuição do número de arquivos (esquerda) e total de linhas (direita) por status do PR.

A análise da Figura 2 revela um resultado que contradiz a hipótese inicial. Tanto para o número de arquivos quanto para o total de linhas modificadas, a mediana (a linha central na caixa) dos PRs com status **MERGED** é ligeiramente superior à dos PRs **CLOSED**. Isso indica que, para a maioria dos casos analisados, os PRs aceitos não são menores que os rejeitados. Não há, portanto, evidências nos dados de que um tamanho maior leva a uma maior probabilidade de rejeição.

Conclusão para RQ01: A hipótese de que PRs maiores são mais rejeitados é **refutada**. O tamanho do PR, por si só, não se mostrou um fator determinante para a sua rejeição.

4.2 RQ05: Relação entre Tamanho e Número de Revisões

A quinta questão de pesquisa avalia se contribuições maiores exigem um esforço de revisão mais intenso, medido pelo número de revisões. A hipótese era de que haveria uma correlação positiva entre essas variáveis. Para visualizar essa relação, foram utilizados gráficos de densidade (hexbin plots), que mostram a concentração de PRs em diferentes combinações de tamanho e número de revisões.

A análise da Figura 3 mostra uma concentração muito alta de PRs no canto inferior esquerdo de ambos os gráficos. Isso confirma que a vasta maioria das contribuições envolve poucos arquivos/linhas e recebe poucas revisões. No entanto, é possível observar uma tendência sutil: à medida que o tamanho aumenta (eixo X), os hexágonos com maior densidade tendem a se deslocar levemente para cima no eixo Y. Isso indica que, embora menos frequentes, PRs maiores estão consistentemente associados a um número maior de revisões.

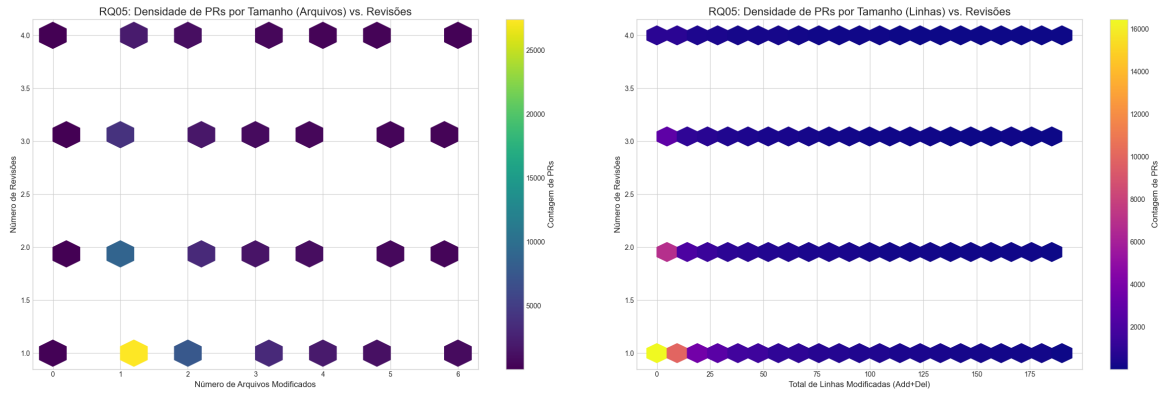


Figura 3: Densidade de PRs por tamanho (arquivos à esquerda, linhas à direita) e número de revisões.

Conclusão para RQ05: A hipótese de que PRs maiores exigem mais revisões é **validada**. Apesar da maioria dos PRs serem pequenos, há uma clara tendência de que o esforço de revisão aumenta proporcionalmente ao tamanho da contribuição, como mostrado no mapa de calor da Figura 1.

5 Análise da Relação com o Tempo de Análise do PR

Nesta seção, o foco da análise é o tempo, em horas, que um Pull Request permanece aberto. Investigamos a relação do tempo de análise com o resultado final do PR (RQ02) e com o esforço de revisão (RQ06).

5.1 RQ02: Relação entre Tempo de Análise e Status do PR

A segunda questão de pesquisa parte da hipótese de que PRs que demoram muito para serem analisados tendem a ser rejeitados, seja por abandono ou pela complexidade dos problemas discutidos. Para visualizar a distribuição dos dados, foi gerado um gráfico de violino, focando nos 80

A análise da Figura 4 oferece suporte claro à hipótese. O "violino" correspondente aos PRs com status **CLOSED** é visivelmente mais alongado para cima, e sua mediana (ponto branco) está em um patamar superior ao dos PRs **MERGED**. A parte mais larga da distribuição para PRs aceitos concentra-se em tempos de análise muito baixos, enquanto para os PRs rejeitados a distribuição é mais espalhada por durações maiores. Isso demonstra que, embora muitos PRs sejam resolvidos rapidamente, a probabilidade de rejeição aumenta consideravelmente com a demora no processo.

Conclusão para RQ02: A hipótese de que PRs com maior tempo de análise tendem a ser rejeitados é **validada**. A demora no processo de revisão é um forte indicador de que a contribuição pode não ser integrada.

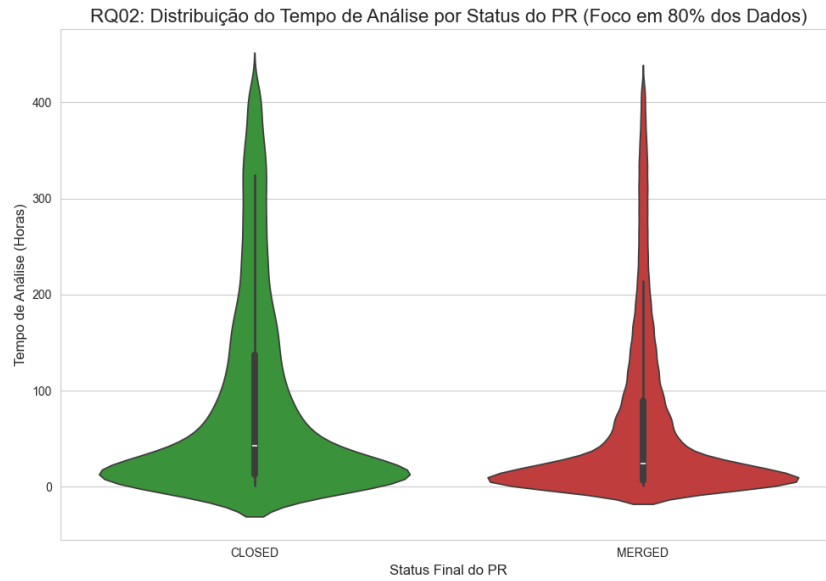


Figura 4: Distribuição do tempo de análise por status do PR (Foco em 80 por cento dos dados).

5.2 RQ06: Relação entre Tempo de Análise e Número de Revisões

A sexta questão de pesquisa investiga se o tempo que um PR permanece aberto se correlaciona com a quantidade de revisões que ele recebe. Para analisar a concentração dos dados, foi utilizado um gráfico de densidade 2D (KDE plot), também focado nos 80 por cento dos dados com menores valores.

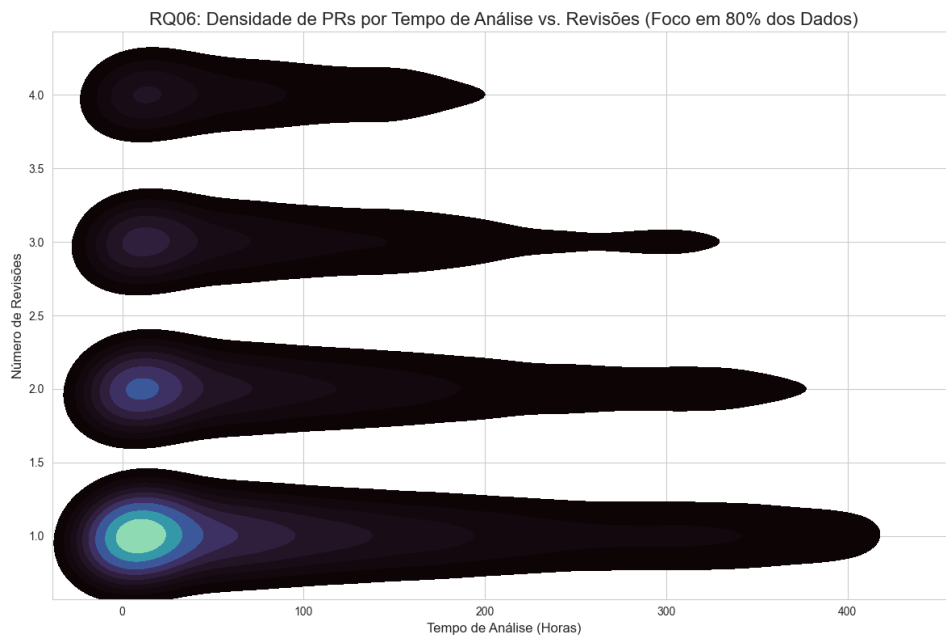


Figura 5: Densidade de PRs por tempo de análise e número de revisões (Foco em 80 por cento dos dados).

O gráfico de densidade na Figura 5 mostra que a esmagadora maioria dos Pull Requests (a área mais escura) se concentra em um tempo de análise muito baixo e um número de revisões também muito baixo. Não há uma tendência aparente de que, à medida que o tempo aumenta (movimento para a direita no eixo X), o número de revisões também aumente (movimento para cima no eixo Y). A densidade permanece concentrada na base do gráfico, independentemente do tempo de análise. Isso sugere que a demora em um PR não está necessariamente ligada a um processo com muitas rodadas de revisão, mas sim a outros fatores, como a falta de atenção dos revisores ou a complexidade do problema.

Conclusão para RQ06: A hipótese de que o tempo de análise se correlaciona com o número de revisões é **refutada**. Não há evidências de uma relação direta entre as duas variáveis; a maioria dos PRs, sejam eles rápidos ou demorados, recebe poucas revisões.

6 Análise da Relação com a Descrição do PR

Nesta seção, a análise foca no impacto do tamanho da descrição de um Pull Request. Investigamos se uma descrição mais detalhada se relaciona com o resultado final do PR (RQ03) e com o esforço de revisão necessário (RQ07).

6.1 RQ03: Relação entre Descrição e Status do PR

A terceira questão de pesquisa partiu da hipótese de que PRs com descrições mais detalhadas seriam mais fáceis de entender e, conseqüentemente, teriam maior chance de serem aceitos. Para comparar as distribuições, foi gerado um gráfico de densidade sobreposto (ridge plot), focando nos 90 por cento dos PRs com descrições menores.

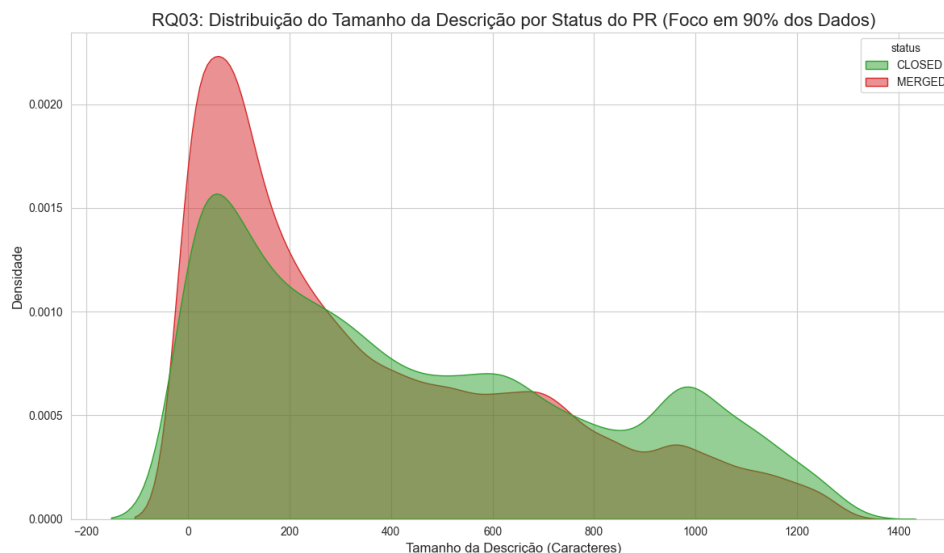


Figura 6: Distribuição do tamanho da descrição por status do PR (Foco em 90 por cento dos dados).

A análise da Figura 6 não suporta a hipótese inicial. A distribuição para PRs aceitos (**MERGED**, em verde) mostra um pico de densidade muito acentuado em tamanhos de

descrição próximos de zero, indicando que uma grande quantidade de PRs aceitos possui descrições mínimas. Em contrapartida, a distribuição para PRs rejeitados (CLOSED, em vermelho) é mais achatada e deslocada para a direita, sugerindo que PRs rejeitados tendem a ter descrições ligeiramente mais longas. Isso pode ocorrer porque PRs problemáticos ou controversos exigem mais explicações e discussões, que não necessariamente levam a um resultado positivo.

Conclusão para RQ03: A hipótese de que descrições mais detalhadas levam a uma maior chance de aceitação é **refutada**. Os dados sugerem que não há uma relação positiva, e que PRs rejeitados tendem a ter descrições um pouco mais longas que os aceitos.

6.2 RQ07: Relação entre Descrição e Número de Revisões

A sétima questão de pesquisa investiga se uma descrição mais completa poderia reduzir a necessidade de múltiplas revisões, esclarecendo previamente as dúvidas dos revisores. Para testar essa ideia, os PRs foram agrupados em quatro categorias de tamanho de descrição, e a média de revisões para cada grupo foi calculada, como mostra o gráfico de barras abaixo.

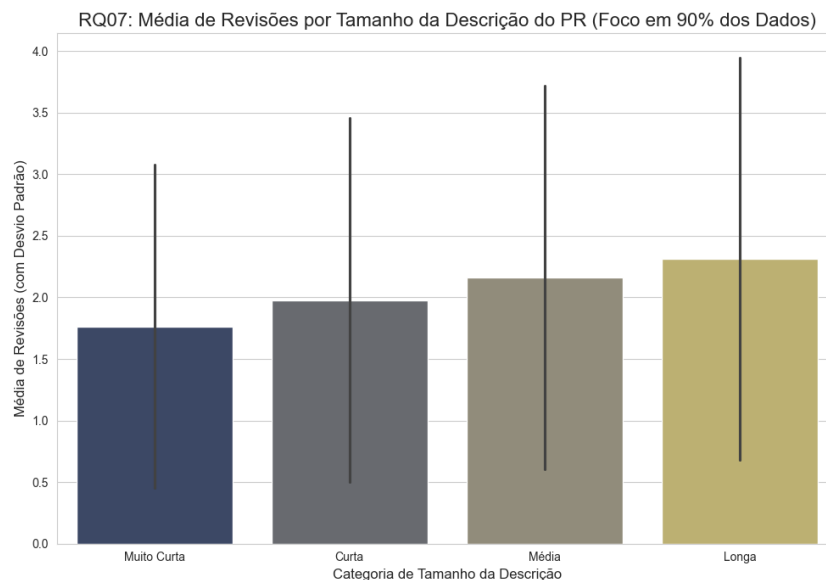


Figura 7: Média de revisões por categoria de tamanho da descrição (Foco em 90 por cento dos dados).

O gráfico de barras na Figura 7 mostra que a média de revisões é praticamente idêntica para todas as quatro categorias de tamanho de descrição, de "Muito Curta" a "Longa". As barras de erro (que indicam o desvio padrão) também são grandes e se sobrepõem consideravelmente, reforçando que não há uma diferença estatisticamente significativa no número de revisões entre os grupos. Isso indica que o esforço de revisão não é influenciado pelo tamanho da descrição inicial do PR.

Conclusão para RQ07: A hipótese de que descrições mais longas diminuem o número de revisões é **refutada**. Não foi encontrada nenhuma correlação entre o tamanho da descrição e a quantidade de esforço de revisão.

7 Análise da Relação com as Interações no PR

Nesta seção final, a análise se aprofunda no fator humano do processo de revisão: as interações, medidas pelo número de participantes e de comentários. Investigamos como essas interações se relacionam com a aceitação do PR (RQ04) e com o esforço total de revisão (RQ08).

7.1 RQ04: Relação entre Interações e Status do PR

A quarta questão de pesquisa explora a hipótese de que um alto nível de interação pode ser um sintoma de problemas, controvérsias ou complexidade, aumentando a chance de um PR ser rejeitado. Para testar isso, foram utilizados box plots para comparar a distribuição de participantes e comentários entre os PRs aceitos e rejeitados.

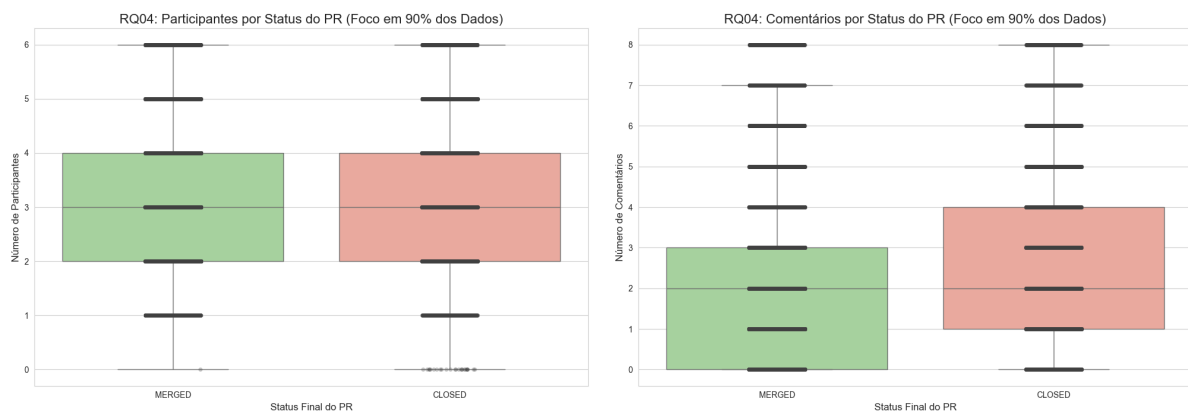


Figura 8: Distribuição do número de participantes (esquerda) e comentários (direita) por status do PR.

A análise da Figura 8 valida fortemente a hipótese. Em ambos os gráficos, a mediana (linha central) para os PRs com status **CLOSED** é visivelmente mais alta do que para os PRs **MERGED**. Isso indica que, em geral, PRs rejeitados envolvem mais participantes e geram mais comentários. A sobreposição dos pontos de dados individuais (stripplot) reforça que, embora a maioria dos PRs tenha pouca interação, os casos com maior volume de discussão estão predominantemente na categoria de rejeitados.

Conclusão para RQ04: A hipótese de que um maior volume de interações está associado à rejeição do PR é **validada**. Tanto o número de participantes quanto o de comentários se mostraram fortes indicadores de um processo de revisão contencioso que frequentemente resulta na não aceitação do código.

7.2 RQ08: Relação entre Interações e Número de Revisões

A oitava questão de pesquisa investiga a relação direta entre as interações e o esforço de revisão. A hipótese é que discussões mais ativas (mais comentários e participantes) demandam um processo de revisão mais rigoroso e, portanto, um maior número de revisões.

A análise dos heatmaps na Figura 9 refuta a hipótese inicial. As áreas de maior densidade (cores mais claras) estão fortemente concentradas na linha inferior de ambos os gráficos, indicando que a esmagadora maioria dos Pull Requests passa por apenas uma

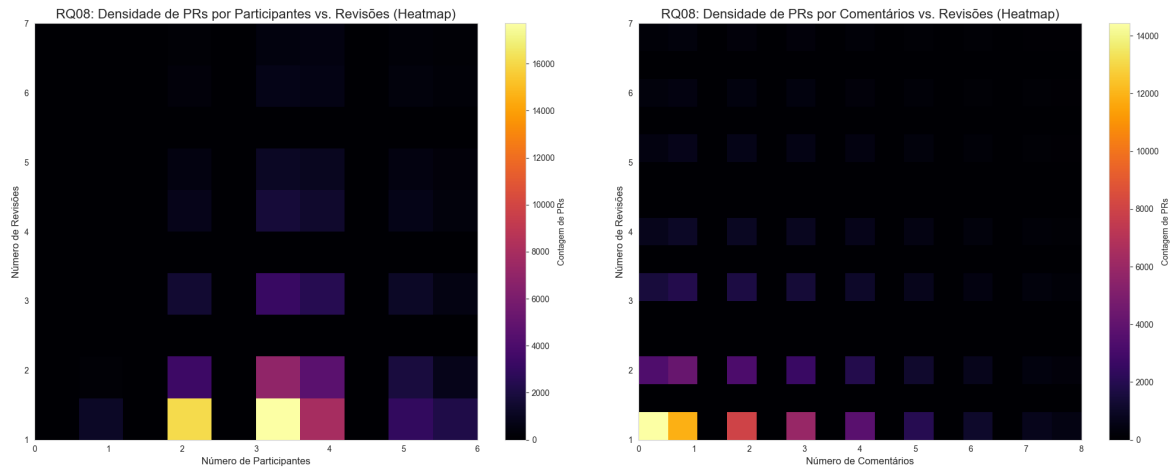


Figura 9: Densidade de PRs por interações (participantes à esquerda, comentários à direita) e número de revisões.

revisão, independentemente do número de participantes ou comentários. Não se observa uma tendência diagonal positiva; pelo contrário, o número de revisões permanece baixo mesmo com o aumento das interações no eixo X. Isso sugere que o esforço de revisão (medido pelo número de revisões) não escala com o volume de discussão.

Conclusão para RQ08: A hipótese de que mais interações levam a um maior número de revisões é **refutada**. Os dados demonstram que a grande maioria das contribuições é resolvida com um único ciclo de revisão, e um aumento no número de participantes ou comentários não está correlacionado com um aumento no número de revisões.

8 Conclusão

Este estudo buscou caracterizar a atividade de *code review* no GitHub, investigando como o tamanho, o tempo de análise, a descrição e as interações de um Pull Request se correlacionam com seu resultado final e com o esforço de revisão. A análise empírica dos dados permitiu validar algumas hipóteses intuitivas, mas também refutou outras, revelando uma dinâmica complexa no processo de revisão.

Os resultados demonstraram que os fatores sociais e de processo são indicadores de resultado mais fortes do que as características técnicas da contribuição. As principais conclusões foram:

- **Interações e Demora São Sinais de Alerta:** O volume de interações (comentários e participantes) e o tempo que um PR permanece aberto emergiram como os preditores mais fortes de rejeição. PRs que geram longos debates ou que ficam parados por muito tempo têm uma chance significativamente maior de serem fechados sem *merge*.
- **Esforço de Revisão Escala com Interação e Tamanho:** O número de revisões — um proxy para o esforço — está fortemente correlacionado com o volume de interações e, de forma moderada, com o tamanho do PR. Isso confirma que discussões ativas e contribuições grandes demandam um trabalho de revisão mais intenso.

- **Descrição e Tamanho Não Preveem Rejeição:** Contraintuitivamente, o tamanho do PR (em arquivos ou linhas de código) e o detalhamento de sua descrição não se mostraram fatores determinantes para a sua aceitação ou rejeição. Contribuições grandes não são necessariamente rejeitadas, e descrições longas não garantem a aceitação.

Em suma, este trabalho sugere que a complexidade da comunicação e a agilidade do processo de revisão são mais críticas para o sucesso de uma contribuição do que sua complexidade técnica isolada. Para trabalhos futuros, seria valioso aplicar análise de sentimento aos comentários para distinguir entre discussões construtivas e debates contenciosos, aprofundando o entendimento de por que PRs com muitas interações tendem a ser rejeitados.