Relatório Final - Laboratório 03 - Caracterizando a Atividade de Code Review no GitHub

Integrantes do grupo:

- Bernardo Cruz Rohlfs
- Leonardo Augusto Pereira do Carmo
- Matheus Belo Santos Mello
- Tarcísio Ney Martins Filho

Professora: Aline Norberta de Brito

Introdução

Resumo:
Este artigo aborda a prática de code review, fundamental nos processos de desenvolvimento ágeis, especialmente no contexto de sistemas open source hospedados no GitHub. A prática de code review envolve a revisão do código produzido por desenvolvedores antes de sua integração na base principal, visando garantir a qualidade e evitar a inclusão de defeitos. Neste estudo, o foco está na análise das atividades de code review em repositórios populares do GitHub, com o objetivo de identificar as variáveis que influenciam o merge de um Pull Request (PR) sob a perspectiva dos desenvolvedores que submetem código aos repositórios selecionados.

Introdução:
A prática de code review é essencial para garantir a qualidade do código integrado em projetos de desenvolvimento de software. No contexto de sistemas open source hospedados no GitHub, as atividades de code review ocorrem principalmente através da avaliação de Pull Requests (PRs). Um PR é uma solicitação de integração de código em uma branch principal do repositório, que é revisada e

discutida por colaboradores do projeto antes de ser aprovada ou rejeitada para merge.

Objetivo:
O objetivo deste trabalho é analisar a atividade de code review em repositórios populares do GitHub, com o intuito de identificar as variáveis que influenciam o merge de um PR, sob a perspectiva dos desenvolvedores que submetem código aos repositórios selecionados. Algumas das variáveis a serem consideradas incluem o tamanho dos PRs, o tempo de análise, a qualidade da descrição do PR, as interações durante o processo de revisão, entre outras.

Perguntas a serem respondidas:

RQ 01. Qual a relação entre o **tamanho** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs menores tendem a receber feedback mais rápido e com menos problemas identificados, levando a um processo de revisão mais rápido e a um merge mais suave.

RQ 02. Qual a relação entre o **tempo de análise** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs que são analisados mais rapidamente tendem a ter um feedback mais positivo, já que os problemas são identificados e resolvidos de maneira mais eficiente.

RQ 03. Qual a relação entre a **descrição** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs com descrições claras e detalhadas têm maior probabilidade de receber um feedback positivo, pois facilitam a compreensão do que foi implementado e quais são as expectativas de revisão.

RQ 04. Qual a relação entre as **interações** nos PRs e o feedback final das revisões?

Hipótese Informal: Um maior número de interações (comentários, revisões, sugestões) nos PRs pode indicar uma maior colaboração e engajamento da equipe, resultando em um feedback mais completo e construtivo.

RQ 05. Qual a relação entre o **tamanho** dos PRs e o número de revisões realizadas?

Hipótese Informal: PRs maiores tendem a exigir mais revisões devido à complexidade e ao número de mudanças, enquanto PRs menores podem ser revisados mais rapidamente e com menos iterações.

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

Hipótese Informal: PRs que são analisados mais rapidamente geralmente passam por menos revisões, pois os problemas são identificados e corrigidos mais cedo no processo.

RQ 07. Qual a relação entre a **descrição** dos PRs e o número de revisões realizadas?

Hipótese Informal: PRs com descrições completas e claras podem reduzir a necessidade de revisões, pois os revisores têm uma compreensão mais abrangente do que está sendo proposto.

RQ 08. Qual a relação entre as **interações** nos PRs e o número de revisões realizadas?

Hipótese Informal: Um maior número de interações nos PRs pode indicar um processo de revisão mais detalhado e abrangente, levando a um número maior de revisões antes do merge.

Metodologia

O dataset utilizado nesta pesquisa será composto por Pull Requests (PRs) submetidos a repositórios populares do GitHub. A seleção dos PRs seguirá critérios específicos:

Repositórios Populares: Serão considerados os 200 repositórios mais populares do GitHub, com base em métricas como estrelas, forks e atividade.

Quantidade Mínima de PRs: Serão selecionados apenas os repositórios que possuam pelo menos 100 PRs, incluindo tanto PRs MERGED quanto CLOSED.

PRs com Processo de Code Review: A inclusão ocorrerá somente para PRs com status MERGED ou CLOSED e que possuam ao menos uma revisão (conforme o campo review_total_count).

Exclusão de PRs Revisados Automaticamente: PRs revisados automaticamente por bots ou ferramentas de CI/CD serão excluídos do dataset. Serão considerados apenas os PRs cuja revisão tenha levado pelo menos uma hora, ou seja, a diferença entre a data de criação e a data de merge (ou close) seja maior que uma hora.

Com base no dataset criado, a pesquisa visa responder às seguintes questões de pesquisa, divididas em duas dimensões:

Feedback Final das Revisões (Status do PR):

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

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

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

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

Número de Revisões:

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

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

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

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

Para cada dimensão e questão de pesquisa, serão utilizadas as seguintes métricas para realizar as análises e correlações:

Tamanho: Número de arquivos modificados e total de linhas adicionadas e removidas nos PRs.

Tempo de Análise: Intervalo de tempo entre a criação do PR e a última atividade registrada (merge ou close).

Descrição: Tamanho do corpo de descrição dos PRs em caracteres, utilizando a versão markdown para essa contagem.

Interações: Número de participantes envolvidos nos PRs e número de comentários realizados durante o processo de revisão.

Essas métricas permitirão explorar e compreender as relações entre as variáveis estudadas e os resultados obtidos nos processos de code review em repositórios populares do GitHub.

Resultados

Na Figura 1 está representada a correlação entre as métricas utilizadas para a análise desse projeto. Sendo essas *Body, Addtions, Deletions, ReviewComments, participants* e *TimeToMergeOrClose.*

Se o valor for 0 as métricas não tem correlação entre elas, se for positivo indica que elas apresentam uma associação positiva.

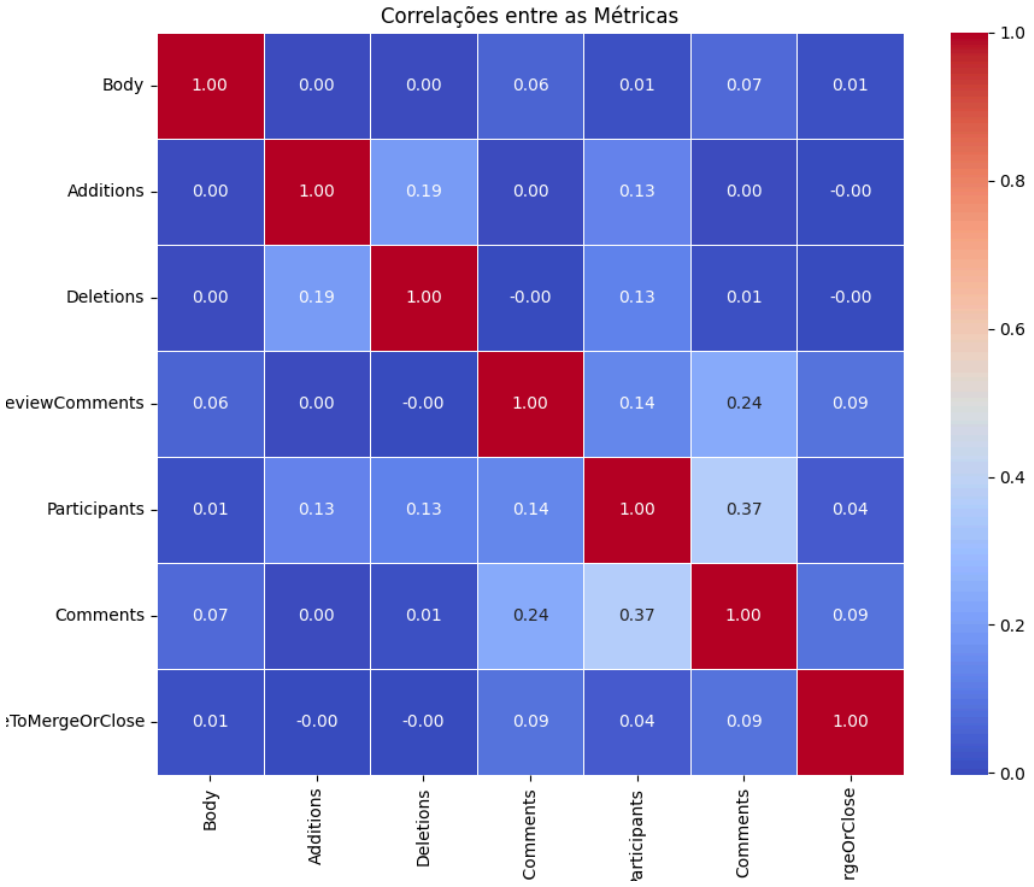

Figura 1

A seguir temos os resultados e conclusões a respeito das perguntas a serem respondidas nesse projeto.

RQ 01. Qual a relação entre o **tamanho** dos PRs e o feedback final das revisões?

As revisões são cruciais para garantir a qualidade do código e a integridade do projeto, detectando possíveis bugs, melhorando a legibilidade e garantindo a conformidade com os padrões estabelecidos. Quanto maior o tamanho do PR, mais complexa é a mudança, tornando a revisão ainda mais crucial.

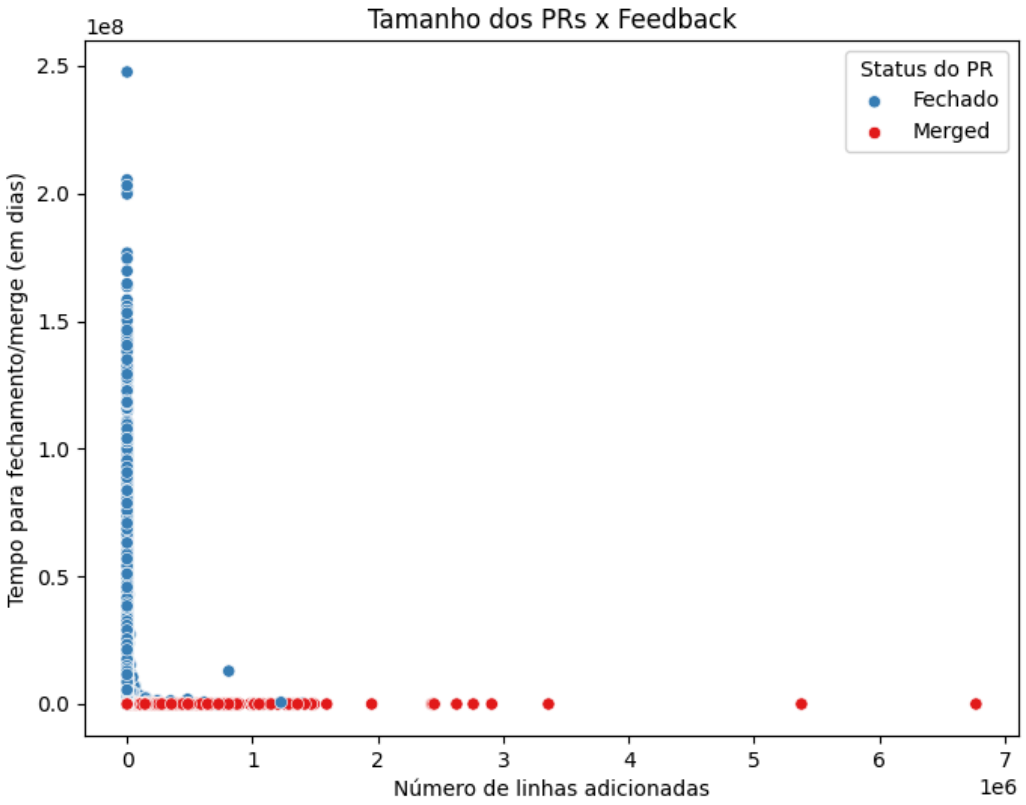

Figura 2

Vê-se na figura 2 que nossa hipótese informal está em de acordo com os resultados encontrados, pois, observando o gráfico, percebe-se que quanto menor o tamanho de linhas tendem a ser mais fáceis de revisar e têm uma probabilidade maior de receber feedback positivo, enquanto PRs maiores podem ser mais desafiadores de revisar e podem resultar em mais feedbacks ou feedbacks mais críticos

RQ 02. Qual a relação entre o **tempo de análise** dos PRs e o feedback final das revisões?

O tempo de análise representa a duração dedicada pelos revisores à avaliação minuciosa de um PR. Durante esse período, os revisores examinam o código. Essa fase é crucial, pois influencia diretamente a qualidade e a eficácia do feedback fornecido ao desenvolvedor, impactando diretamente na integridade e na qualidade do projeto de software.

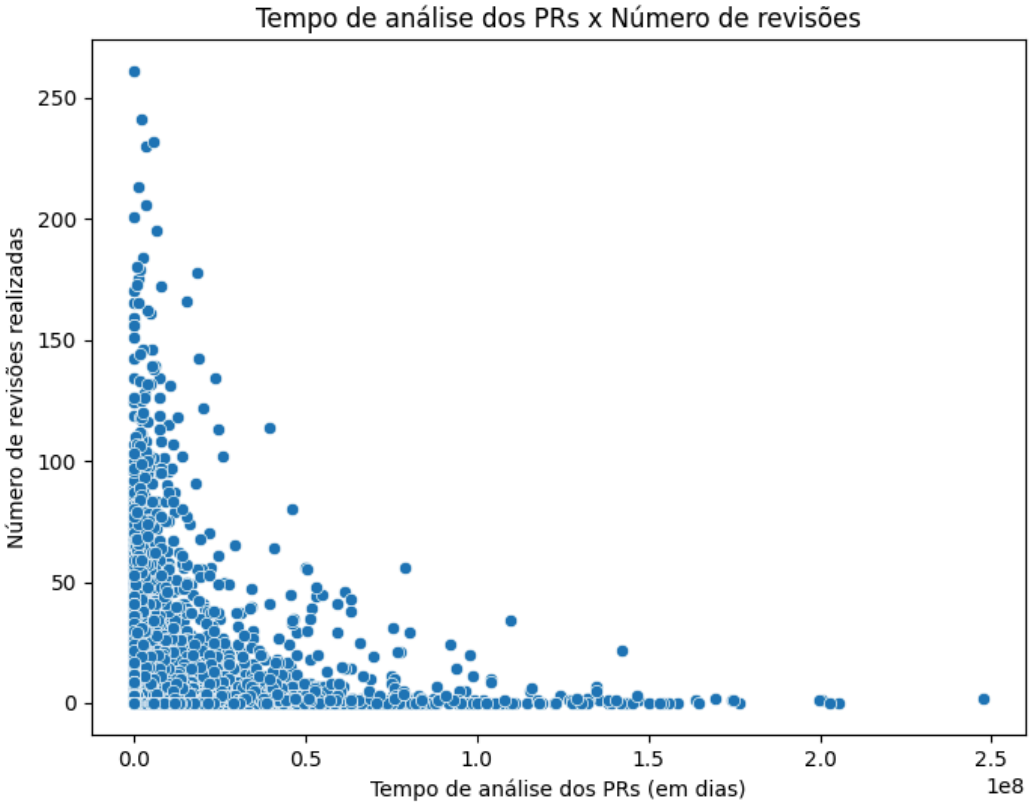

Figura 3

De acordo com a figura 3 e os resultados encontrados, percebe-se que quanto menor o tempo de análise maior foi o número de revisões realizadas. Nesse sentido, os dados observados corroboram a importância de um tempo de análise adequado, porém rápido o suficiente para manter um fluxo de trabalho eficiente e iterativo. Isso sugere que equipes que conseguem revisar PRs de maneira ágil e eficaz tendem a obter um feedback final mais construtivo e produtivo, contribuindo para a qualidade geral do código e para o progresso do projeto.

RQ 03. Qual a relação entre a **descrição** dos PRs e o feedback final das revisões?

A descrição de um Pull Request (PR) desempenha um papel fundamental na qualidade do feedback final das revisões. Uma descrição clara e informativa pode fornecer contexto essencial sobre as mudanças propostas, objetivos e possíveis impactos. Isso ajuda os revisores a entenderem melhor o propósito do PR e focarem em áreas específicas durante a revisão. Por outro lado, descrições vagas ou ausentes podem resultar em interpretações equivocadas e feedbacks menos úteis.

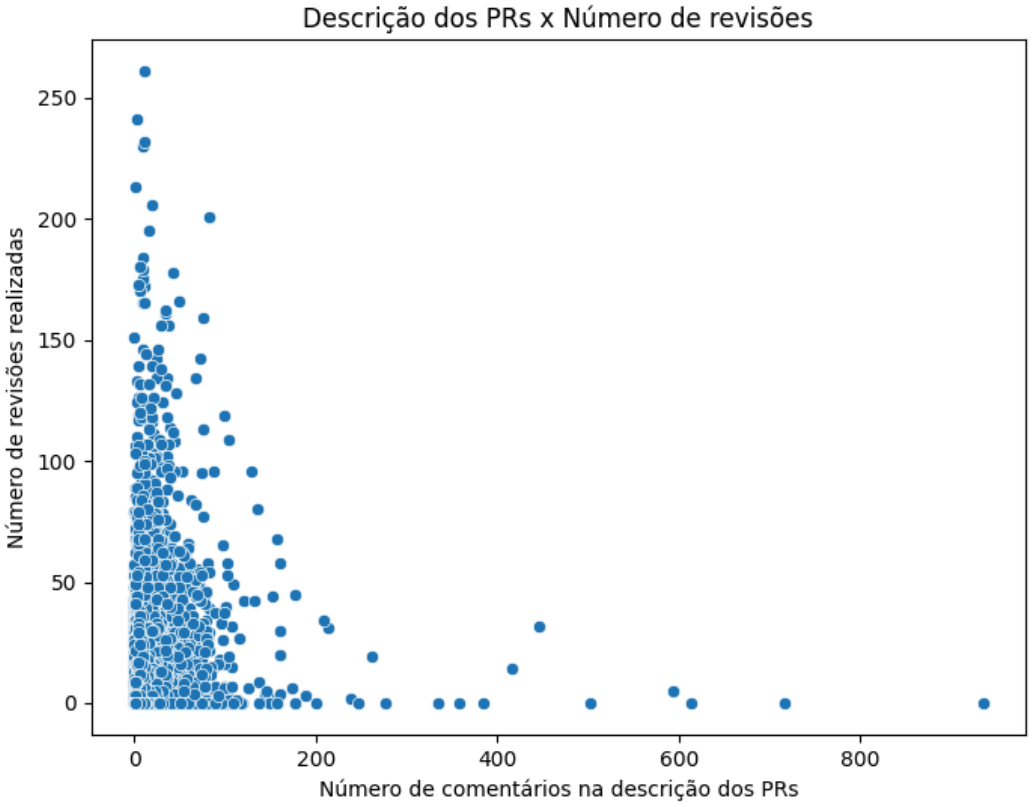

Figura 4

Segundo a figura 4 e os resultados obtidos, com o crescimento do número de descrições dos PRs um projeto tende a ter maior número de feedback, inclusive revisões positivas, entretanto, a partir de uma certa quantidade de comentários, as revisões começa a cair mostrando que descrições vagas podem resultar em interpretações equivocadas, menor quantidade de feedbacks e menos úteis.

RQ 04. Qual a relação entre as **interações** nos PRs e o feedback final das revisões?

No processo de revisão de código em Pull Requests (PRs), as interações entre os revisores e o autor desempenham um papel crucial na qualidade do feedback final. Quanto mais interações ocorrem, maior é a oportunidade para esclarecimento de dúvidas, discussão sobre as mudanças propostas e refinamento do código.Por outro lado, a falta de interações pode resultar em revisões superficiais ou em feedbacks menos abrangentes, comprometendo a qualidade final das revisões.

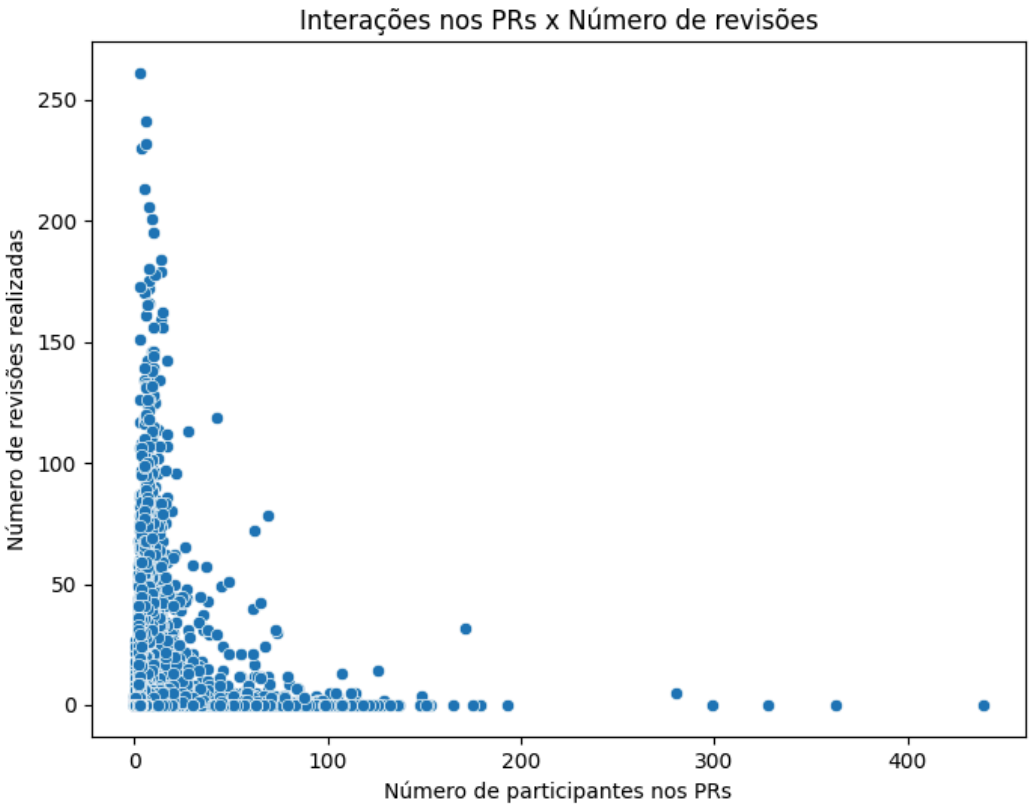

Figura 5

Não necessariamente quanto maior o número de participantes maior será o número de revisões, como dito na hipótese informal. Segundo a figura 5 e os resultados, chega um momento que quanto maior o número de participantes nos PRs menor é o número de revisões, demonstrando que um número de participantes menor, porém mais comprometido com o projeto, resulta em maiores números de revisões.

RQ 05. Qual a relação entre o **tamanho** dos PRs e o número de revisões realizadas?

No contexto do desenvolvimento de software, é observado que o tamanho dos Pull Requests (PRs) influencia diretamente o número de revisões realizadas.

Geralmente, PRs maiores requerem revisões mais detalhadas e consomem mais tempo dos revisores, o que pode resultar em um número menor de revisões.

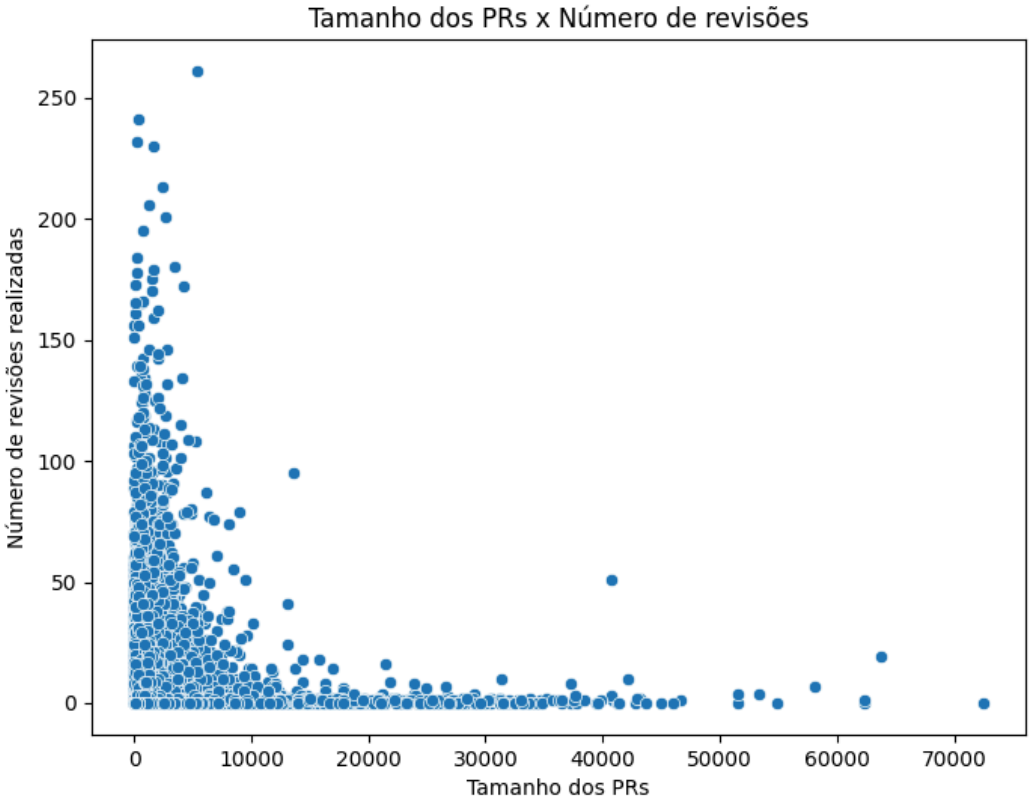

Figura 6

De acordo com a figura 6 e os resultados encontrados vê-se que, até certa quantidade, quanto maior o número de PRs, maior é o número de revisões realizadas, entretanto após esse ponto as revisões começam a diminuir por requerer mais tempo de revisão e não necessariamente quantidade de revisão.

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

O tempo dedicado à análise dos PRs impacta diretamente o número de revisões realizadas. Revisões mais rápidas podem resultar em uma quantidade menor de revisões, pois podem ser menos abrangentes e detalhadas. Por outro lado, revisões mais demoradas tendem a permitir uma análise mais aprofundada, potencialmente resultando em um maior número de revisões para garantir a qualidade do código.

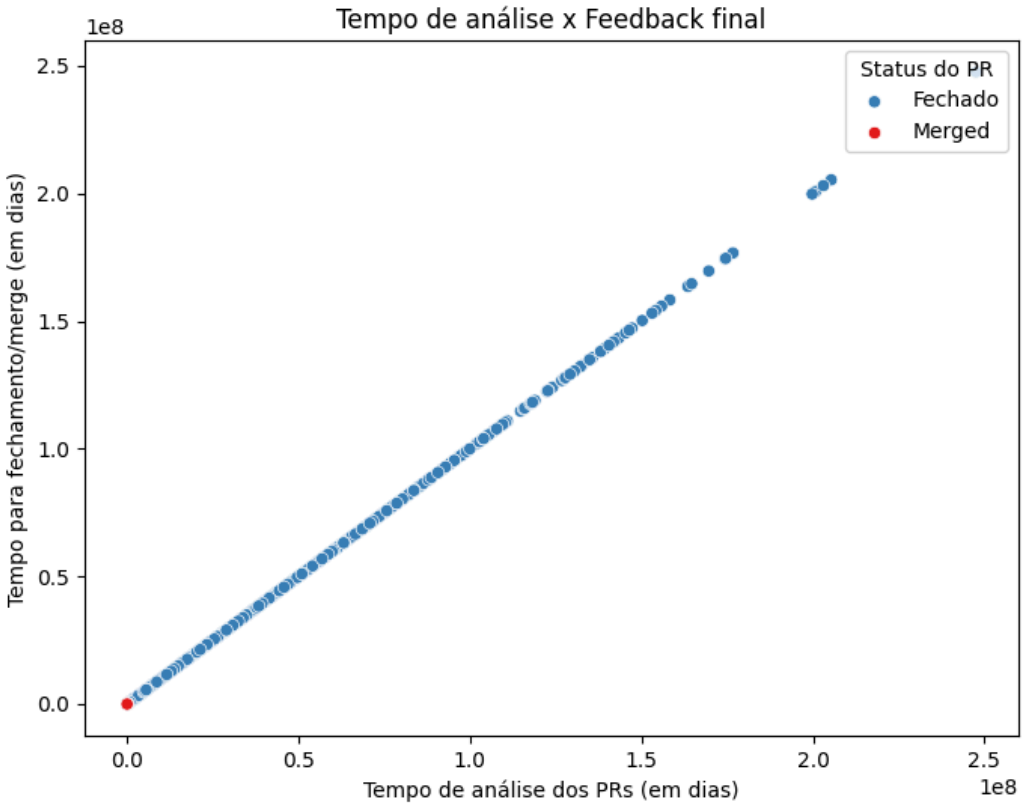

Figura 7

Segundo a figura 7 e os resultados quanto maior o tempo de análise dos PRs, maior o tempo destinado para fechamento, demonstrando que PRs com tempo de análise maior “obriga” a um número maior de revisões a fim de garantir a efetividade da subida.

RQ 07. Qual a relação entre a **descrição** dos PRs e o número de revisões realizadas?

A qualidade da descrição de um PR influencia diretamente o número de revisões realizadas. Uma descrição clara e informativa facilita o entendimento das mudanças propostas pelos revisores, o que pode levar a mais revisões para garantir a integridade e a qualidade do código. Por outro lado, descrições vagas fazem com que os revisores possam ter dificuldade em compreender o propósito das alterações e, consequentemente, dedicar menos tempo à revisão.

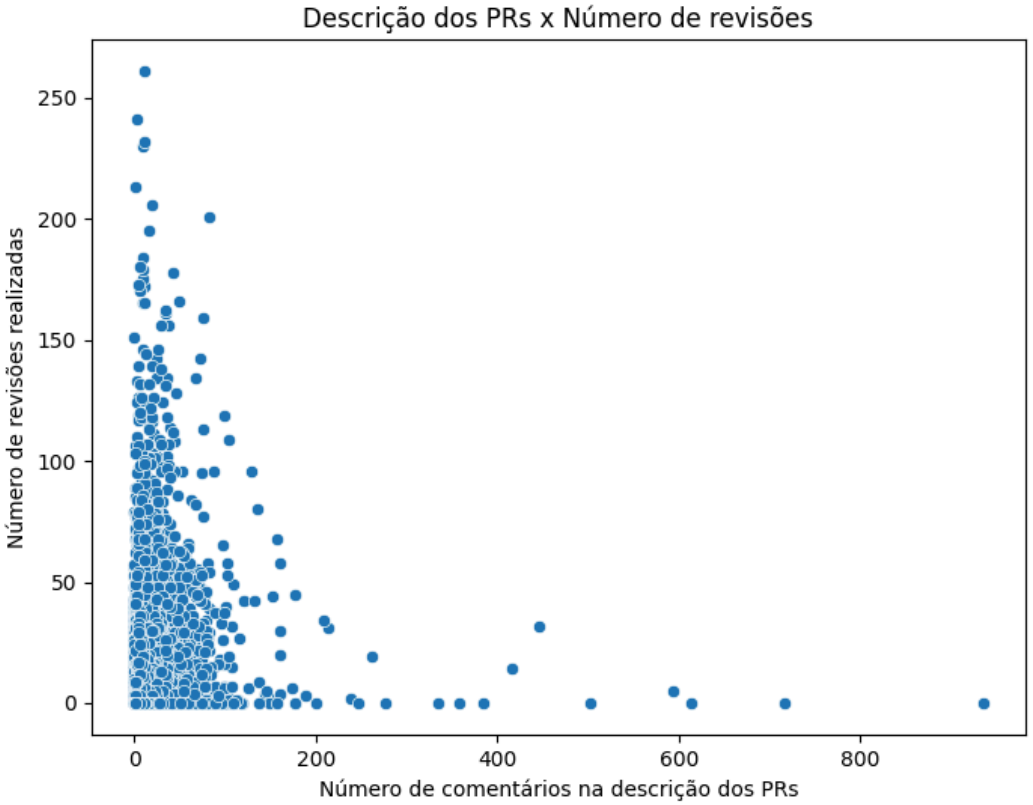

Figura 8

Conforme a figura 8 e os resultados encontrados, observa-se que o número de comentários é importante para maior quantidade de revisões, entretanto, comentários demais acabam diminuindo a eficiência da descrição, desmotivando o número de revisões

RQ 08. Qual a relação entre as **interações** nos PRs e o número de revisões realizadas?

Mais interações entre revisores e autores geralmente levam a um maior número de revisões, pois promovem esclarecimentos, discussões e ajustes no código proposto. Essa troca de feedbacks contribui para uma revisão mais abrangente e interativa, melhorando a qualidade do código.

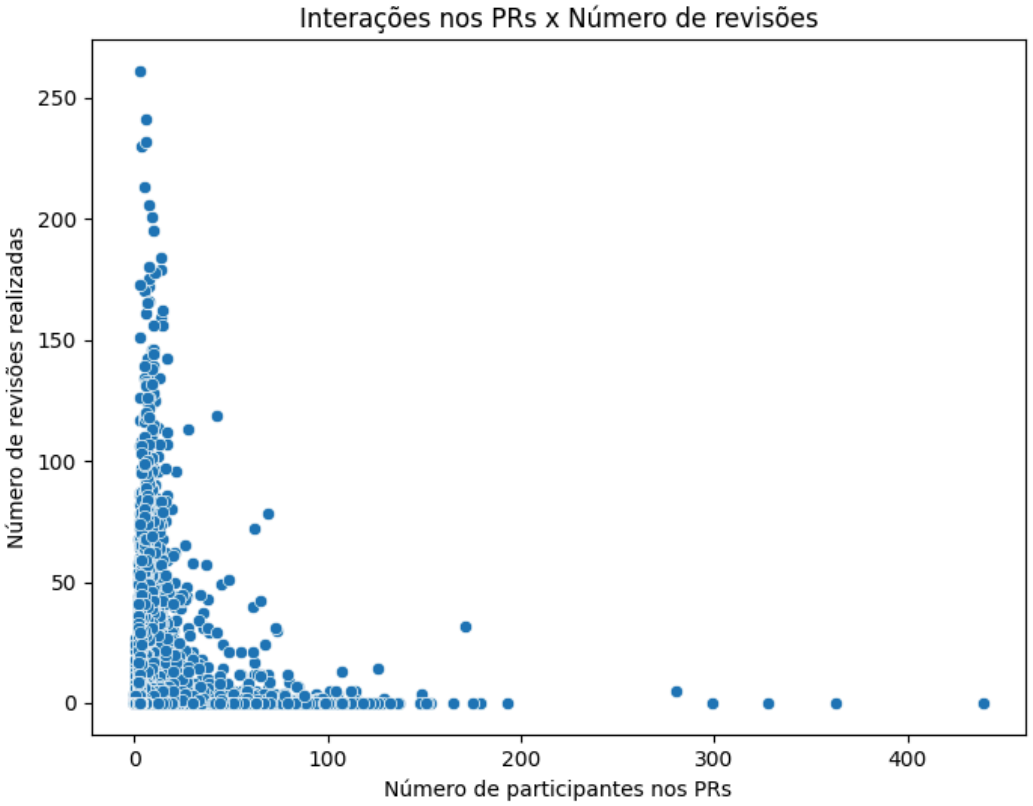

Figura 9

Segundo os resultados e a figura 9 pode-se observar que números altos de participantes nos PRs tendem a ter números maiores de revisões os quais melhoram a qualidade de código. Entretanto, a partir de uma quantidade começa-se ter um número menor de revisões que pode demonstrar menor engajamento quando se tem um número muito alto de participantes, como se fossem participantes “observadores”.

Conclusão

Ao investigar as relações entre diferentes aspectos dos Pull Requests (PRs) e o feedback final das revisões, bem como o número de revisões realizadas, várias conclusões podem ser destacadas.

Os PRs menores tendem a receber feedback final mais positivo, sugerindo que a concisão pode ser vantajosa no processo de revisão de código. Além disso, PRs com descrições detalhadas têm vantagem em receber um feedback final mais positivo, destacando a importância da comunicação clara entre os colaboradores.

Quanto ao número de interações nos PRs, observa-se uma relação com o feedback final das revisões. Isso sugere que uma comunicação mais intensiva pode influenciar positivamente na qualidade da revisão.

Ainda, foram identificadas relações entre o tempo de análise dos PRs e o feedback final das revisões, assim como entre a descrição dos PRs e o número de revisões realizadas. Além disso, a quantidade de PRs estão associados diretamente ao número de revisões realizadas, indicando uma possível maior eficiência no processo de revisão.

Por fim, essas análises ressaltam a importância de uma abordagem flexível na gestão das revisões de código, da concisão nos PRs, da comunicação clara e da interação eficaz entre os colaboradores para melhorar a qualidade das revisões de código e otimizar o processo de desenvolvimento de software.