

Pontifícia Universidade Católica de Minas Gerais

Laboratório de Experimentação de Software

Caracterizando a atividade de code review

Giovanni Bogliolo Sirihal Duarte
Luiz Gustavo Mendes Santos
Pedro Ramos Vidigal

Belo Horizonte
2024

1. Introdução

O code review, componente essencial no desenvolvimento ágil, assegura a qualidade do código por meio da inspeção e colaboração entre desenvolvedores. No GitHub, o code review se dá pela análise de Pull Requests antes de sua integração à branch principal.

Este laboratório se propõe a investigar a atividade de code review em repositórios populares do GitHub, buscando identificar variáveis que influenciam a aprovação de um PR sob a perspectiva do desenvolvedor que submete o código.

Com base na análise nos dados coletados, serão respondidas questões como: PRs menores são aprovados com mais facilidade? O tempo de análise influencia na decisão final? Uma descrição detalhada aumenta as chances de aprovação? A interação entre desenvolvedores e revisores impacta na qualidade do código? As respostas a essas perguntas contribuirão para uma melhor compreensão do processo de code review e fornecerão informações valiosas para otimizar essa prática crucial no desenvolvimento de software.

1.1. Hipóteses

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

R: PRs menores tendem a ser aprovados com mais facilidade, pois são mais fáceis de revisar e entender, já PRs maiores podem gerar mais dúvidas e debates, levando a um maior tempo de análise e, possivelmente, rejeição.

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

R: PRs analisados rapidamente podem indicar clareza e qualidade no código, levando à aprovação, já PRs com longos tempos de análise podem sugerir a necessidade de revisões e correções, aumentando a chance de rejeição.

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

R: Descrições claras e detalhadas facilitam a compreensão do código e aumentam a chance de aprovação, descrições vagas ou incompletas podem gerar dúvidas e atrasar o processo de revisão.

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

R: PRs com muitas interações podem indicar um debate construtivo e colaboração, aumentando a chance de aprovação após as devidas correções, PRs com poucas interações podem sugerir falta de atenção ou dificuldade de entendimento do código.

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

R: PRs maiores podem demandar mais revisões para garantir a qualidade do código. PRs menores podem ter menos revisões, especialmente se forem simples e de fácil compreensão.

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

R: PRs com maior tempo de análise podem ter mais revisões, refletindo a necessidade de múltiplas iterações para aprovação. PRs com menor tempo de análise podem ter menos revisões, indicando clareza e qualidade do código.

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

R: Descrições claras e detalhadas podem reduzir a necessidade de múltiplas revisões. Descrições vagas ou incompletas podem gerar dúvidas e demandar mais revisões para esclarecimento.

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

R: PRs com muitas interações podem ter mais revisões, refletindo o debate e a colaboração entre desenvolvedores e revisores, já PRs com poucas interações podem ter menos revisões, especialmente se o código for claro e de fácil entendimento.

2. Metodologia

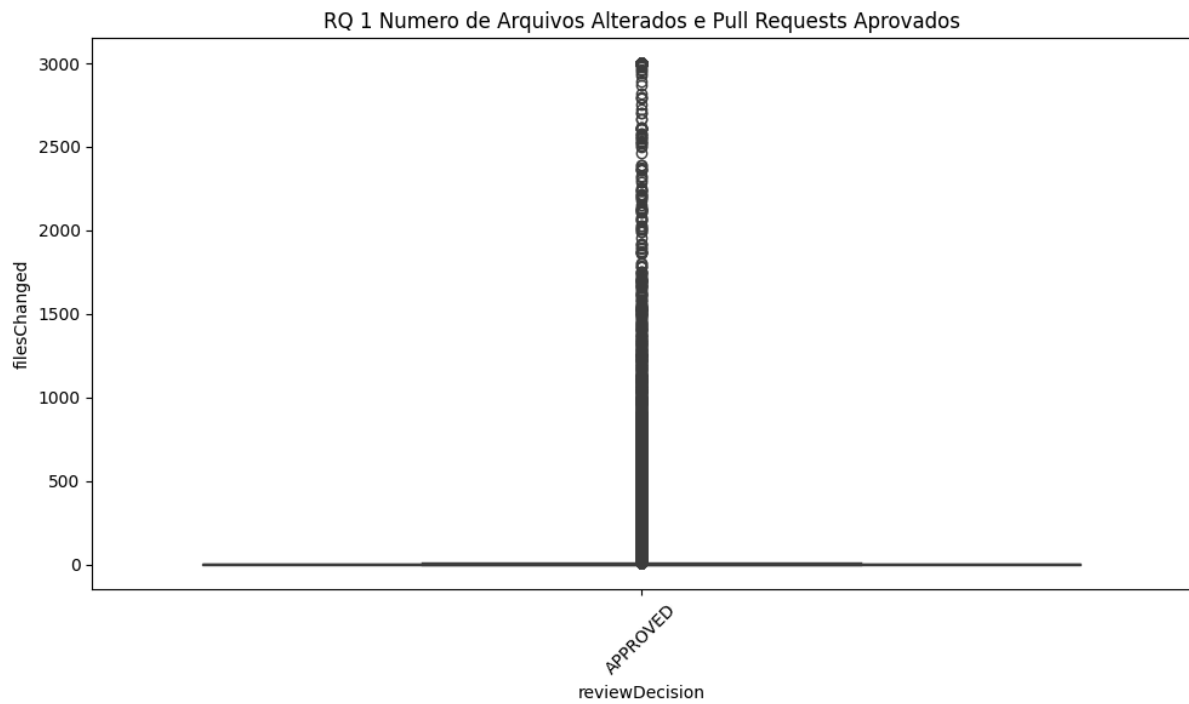
Será utilizada a API do GitHub para coletar dados sobre os Pull Requests (PRs) dos 200 repositórios open source mais populares da plataforma. Para garantir a robustez da análise, serão selecionados apenas repositórios com pelo menos 100 PRs incluindo aqueles com status MERGED e CLOSED. A linguagem de programação Python será usada para automatizar a coleta, processamento e análise dos dados e serão utilizadas também bibliotecas como requests e pandas para manipulação de dados, análise estatística e visualização.

A fim de analisar PRs que passaram por um processo de code review, serão aplicados os seguintes filtros: apenas PRs com status MERGED ou CLOSED, pois estes representam PRs que foram efetivamente revisados e avaliados, incluir apenas PRs com pelo menos uma revisão registrada, garantindo que houve análise humana e considerar apenas PRs com tempo de revisão mínimo de uma hora, excluindo PRs com revisões automáticas ou superficiais.

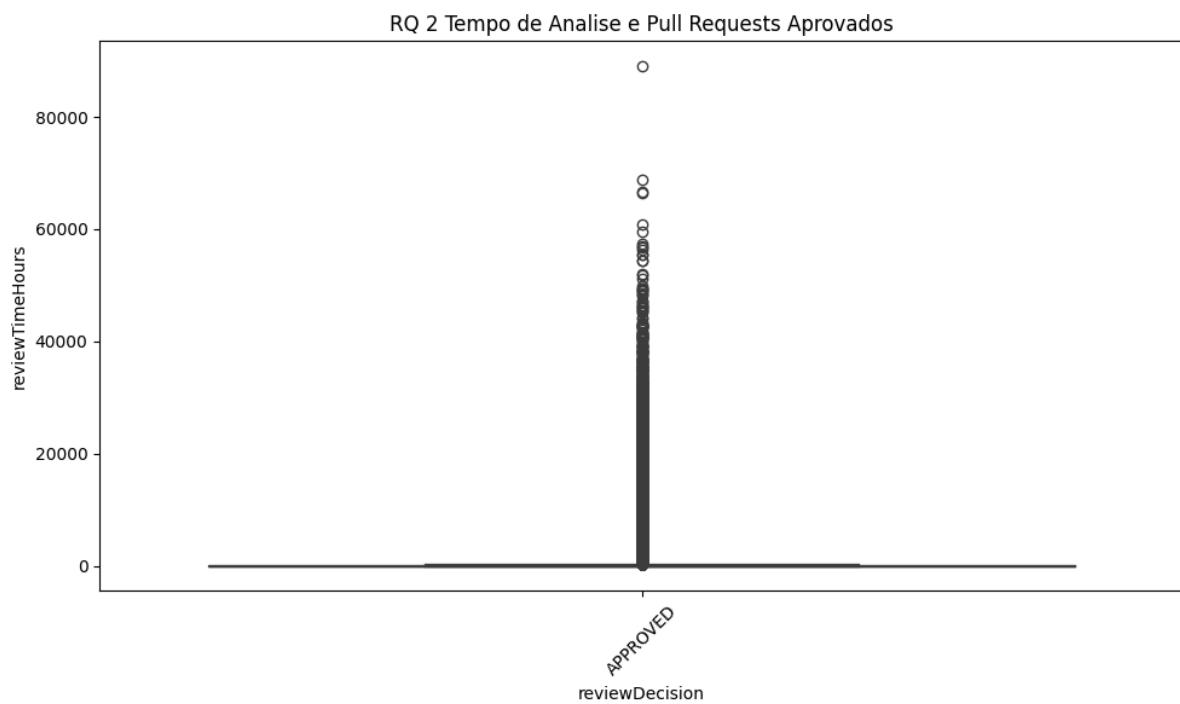
Para cada PR selecionado, serão extraídas as seguintes métricas: número de arquivos modificados, total de linhas adicionadas e removidas, intervalo entre a criação do PR e a última atividade, número de caracteres no corpo da descrição do PR, número de participantes e número de comentários no PR, quantidade de revisões realizadas no PR e status do PR (MERGED ou CLOSED).

Com base nas métricas extraídas a partir daí, serão realizadas análises quantitativas e qualitativas para investigar as relações entre as variáveis e responder às questões de pesquisa. E por fim os resultados obtidos serão comparados com as hipóteses iniciais, buscando identificar padrões e tendências na atividade de code review.

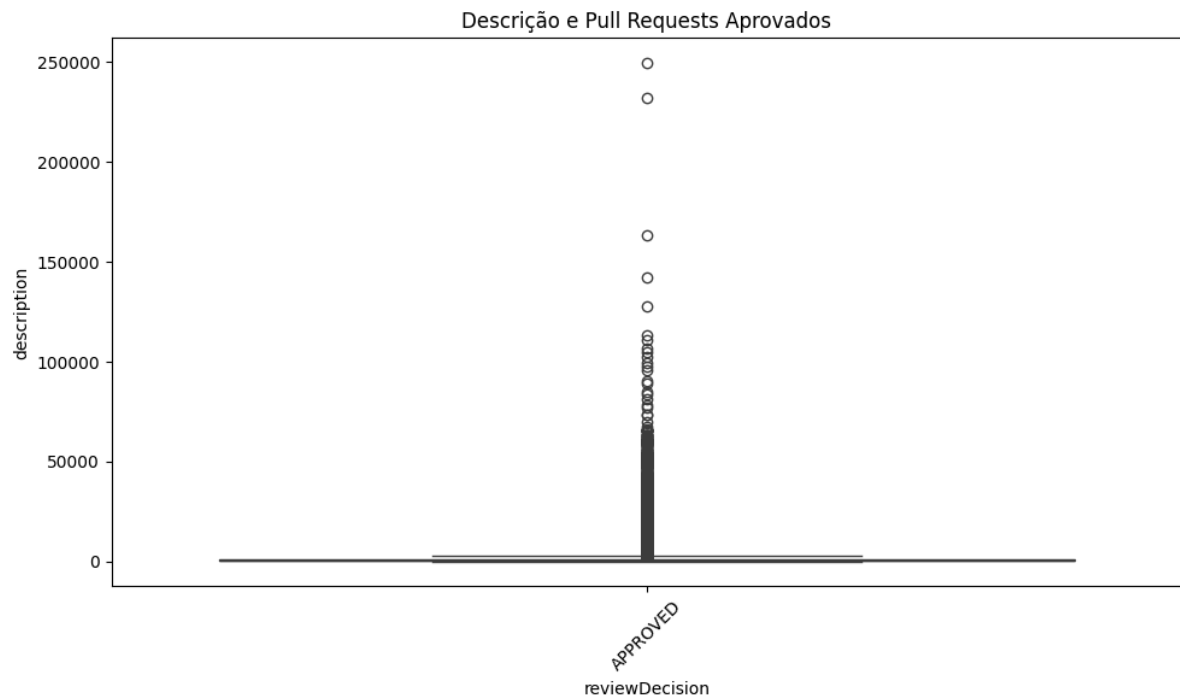
RQ 01. Decisão final: APPROVED, Mediana de tamanho de Arquivos: 2,0



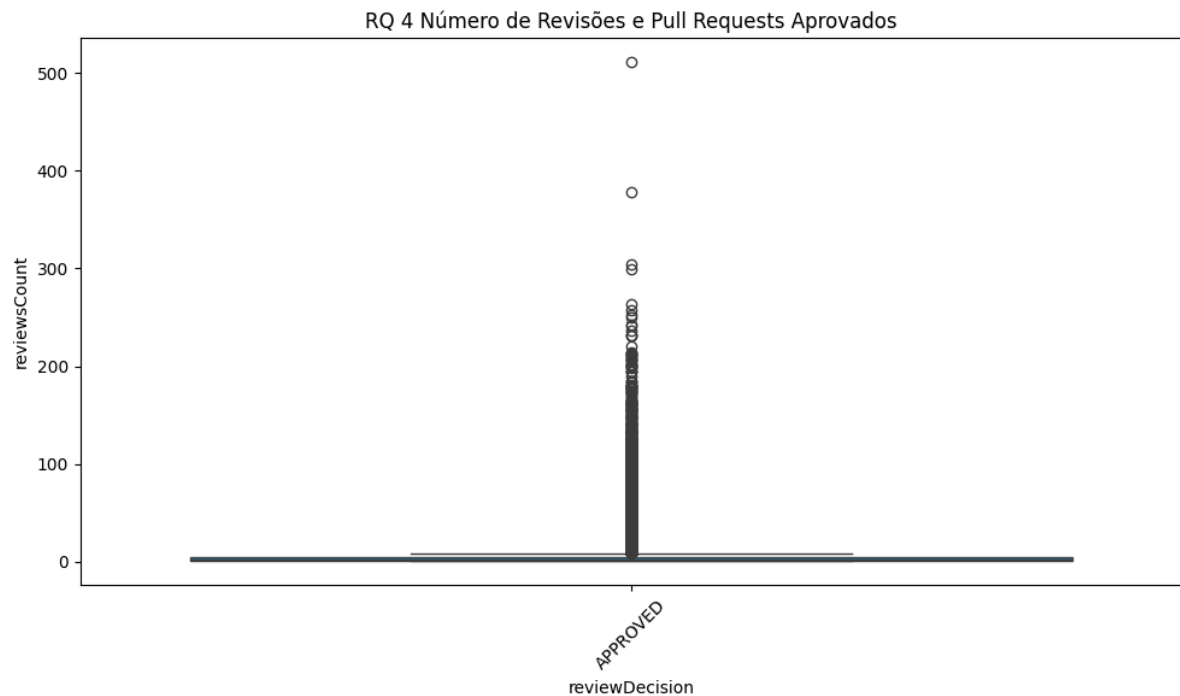
RQ 02. Decisão final: APPROVED, Mediana de tempo em horas: 35,292222



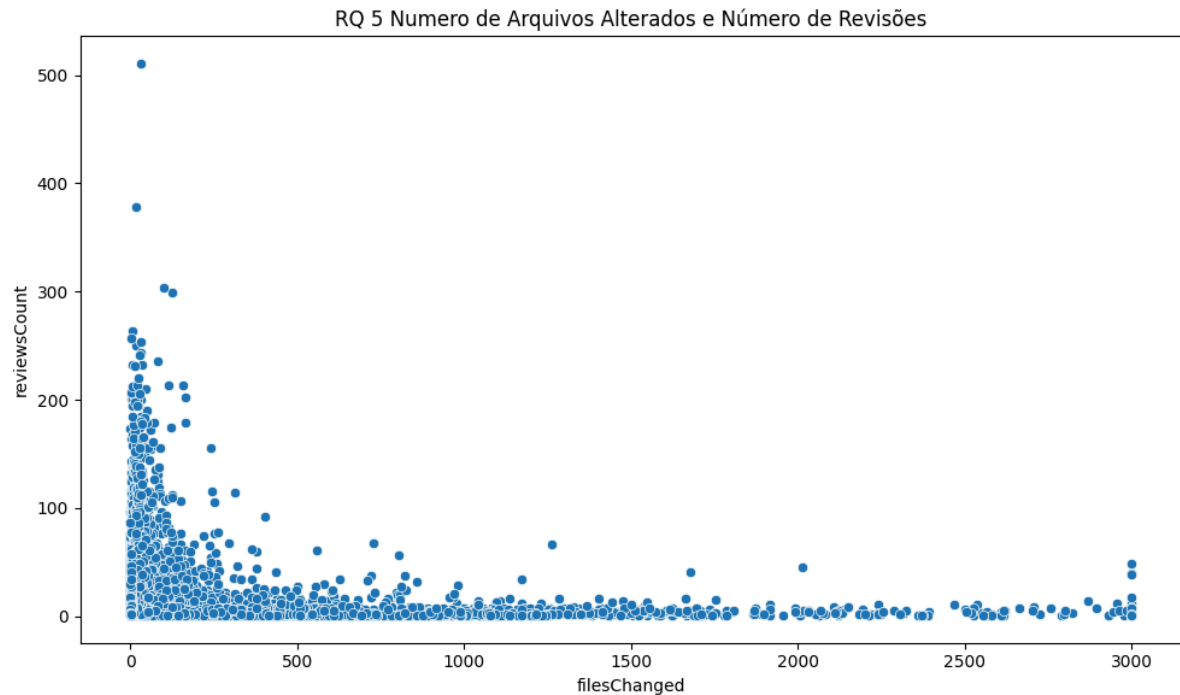
R: RQ 03. Decisão final: APPROVED, Mediana de tamanho da descrição: 468,0



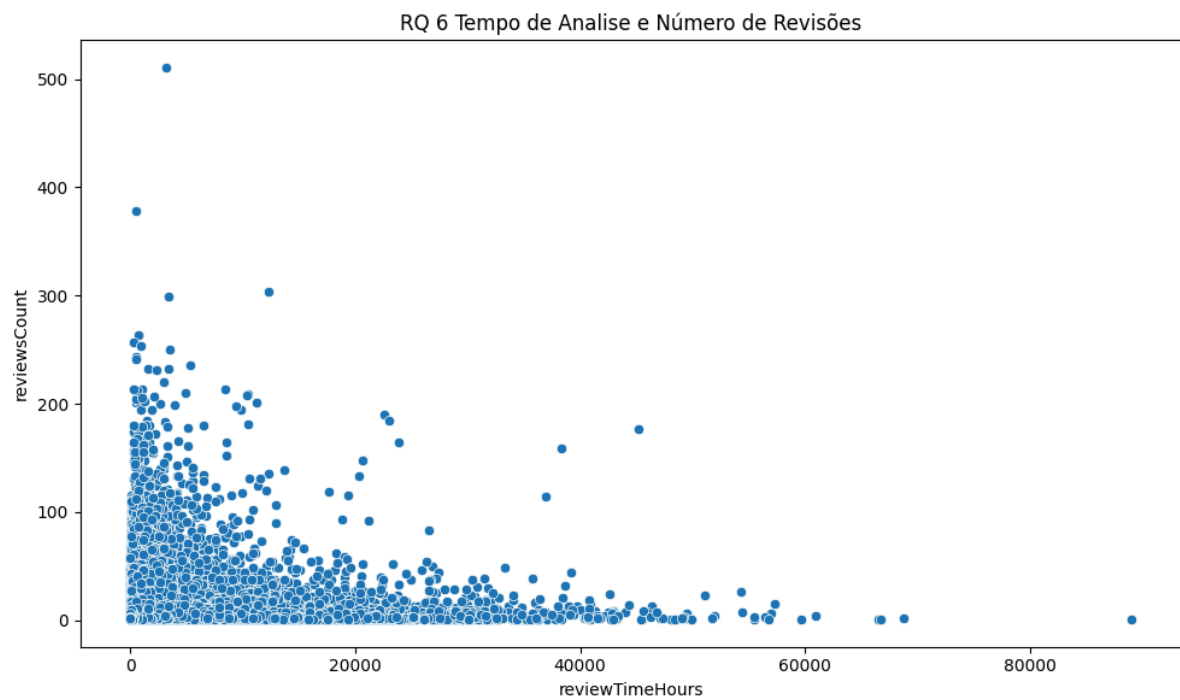
RQ 04. Decisão final: APPROVED, Mediana do número de reviews: 2,0



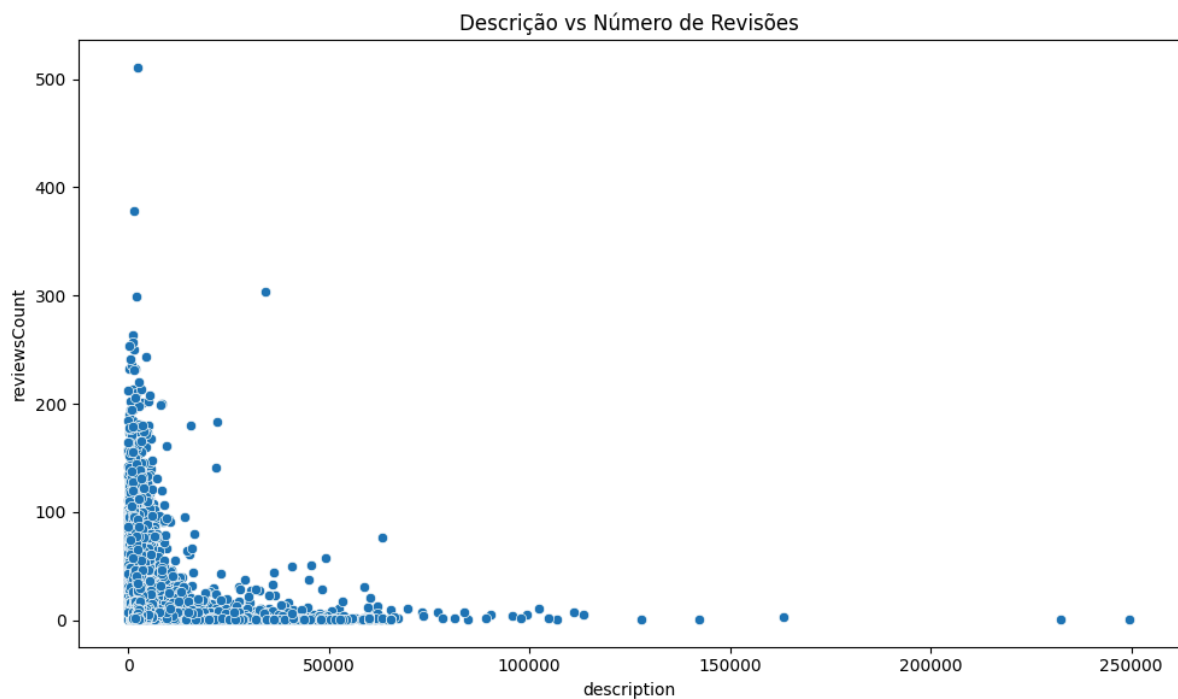
RQ 05. Decisão final: APPROVED, Mediana do número de arquivos para o número de revisões: 2,0



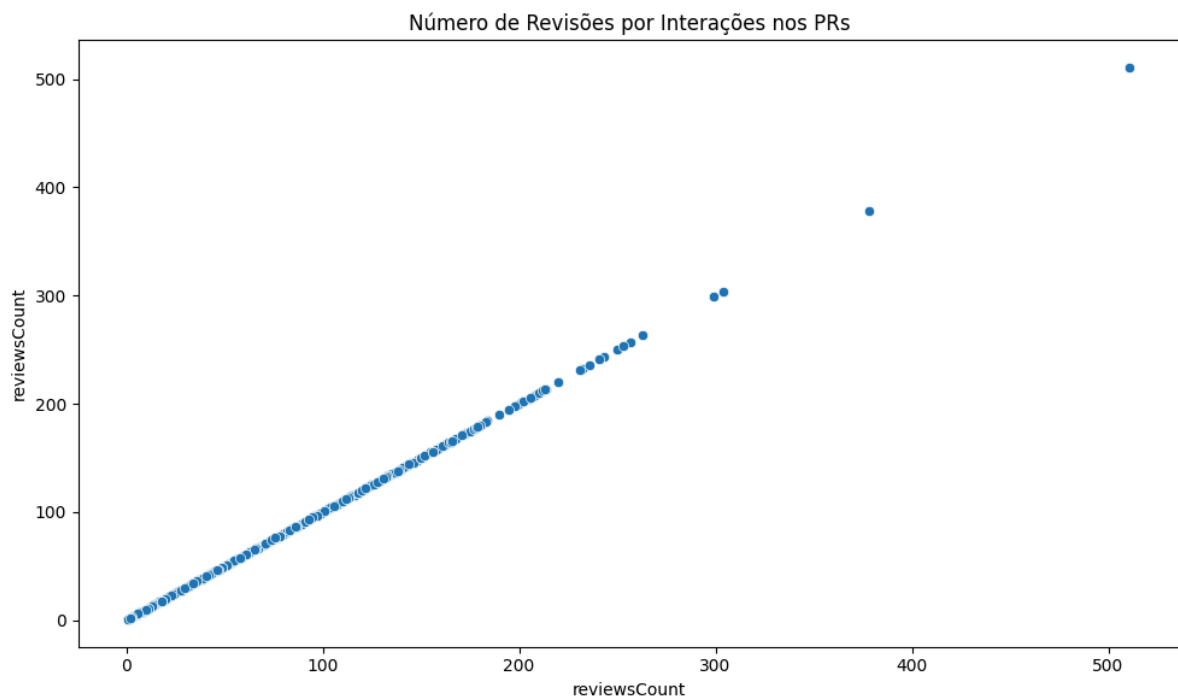
RQ 06. Decisão final: APPROVED, Mediana do tempo de revisão para a mediana do número de revisões: 25,913472



RQ 07. Decisão final: APPROVED, Mediana do tamanho da descrição para a mediana do número de revisões: 437,0



RQ 08. Decisão final: APPROVED, Mediana do número de revisões para a mediana do número de revisões: 2,0



4. Análise dos resultados

Nesta seção, serão apresentadas as análises dos resultados obtidos para cada pergunta de pesquisa.

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

R: A partir do gráfico de distribuição do tamanho dos PRs, nota-se que a maioria dos PRs aprovados tem um tamanho reduzido, conforme indicado pela mediana de 2 arquivos modificados. No entanto, o gráfico também revela uma dispersão significativa, com alguns PRs aprovados contendo um número bem maior de arquivos modificados, evidenciado pela presença de outliers. Essa variação sugere que, embora PRs menores tenham maior chance de aprovação, PRs maiores também podem ser aprovados, desde que apresentem qualidade ou justificativas adequadas no processo de revisão.

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

R: O gráfico de tempo de análise dos PRs mostra uma concentração de PRs revisados rapidamente, com a mediana em torno de 35 horas. Porém, a média é maior, o que sugere a existência de PRs que levaram consideravelmente mais tempo para serem analisados. Isso é evidenciado pelos outliers no gráfico, que representam PRs analisados em prazos muito mais longos, possivelmente devido à complexidade ou quantidade de arquivos modificados. Esses PRs que demoram mais para serem revisados tendem a exigir mais discussões e iterações, como indicado por um maior número de comentários ou revisões. A análise gráfica apoia a hipótese inicial, mas também destaca que PRs complexos não seguem essa regra de maneira uniforme.

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

R: No gráfico de tamanho das descrições, observa-se que a mediana de 468 caracteres está acompanhada por uma variação significativa nas descrições. PRs com descrições mais detalhadas tendem a ser aprovados com mais facilidade, mas o gráfico indica que alguns PRs aprovados tinham descrições menores, sugerindo que a complexidade do código também pode ser um fator importante. PRs simples, que não requerem descrições extensas, ainda são aprovados com frequência, como indicado pelos pontos concentrados na faixa de descrições curtas.

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

R: O gráfico de interações entre desenvolvedores e revisores sugere uma correlação entre o número de interações e o feedback positivo. PRs aprovados tendem a ter um nível moderado de interações (mediana de 2 interações), mas há exceções. Alguns PRs com poucas interações foram aprovados, possivelmente por serem mais simples ou bem descritos desde o início. Já PRs com muitas interações podem ter passado por um processo mais rigoroso de revisões e ajustes antes de serem aprovados, o que é refletido nos outliers do gráfico. Isso sugere que PRs complexos exigem mais discussão, mas interações em excesso também podem ser um indicativo de problemas ou falta de clareza no código original.

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

R: A maioria dos Pull Requests (PRs) com um menor número de arquivos alterados (até cerca de 500 arquivos) teve uma quantidade variável de revisões, com muitos PRs sendo revisados de forma rápida, e alguns exigindo mais revisões. Para PRs com um número muito grande de arquivos modificados (acima de 1000), há uma redução considerável na quantidade de revisões. Isso sugere que revisores podem preferir aprovar PRs maiores com menos iterações, talvez para evitar processos de revisão extensos, ou esses PRs possam estar sendo divididos em revisões subsequentes para gerenciar melhor a complexidade.

Existe uma tendência de correlação inversa clara, conforme o número de arquivos alterados aumenta, a quantidade de revisões tende a diminuir, o que pode parecer contraintuitivo. A hipótese inicial poderia sugerir que PRs maiores requerem mais revisões devido à complexidade, mas os dados mostram o contrário. Isso indica que a hipótese pode estar incorreta ou incompleta, uma vez que PRs com mais arquivos alterados podem ter sido aprovados após revisões mais rápidas ou por revisores experientes, diminuindo o número de revisões.

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

R: A análise mostra que a relação entre o tempo de análise e o número de revisões é mais complexa do que o previsto. Embora a hipótese original sugerisse que PRs que levam mais tempo para serem analisados passam por mais revisões, o gráfico indica que, na prática, PRs com maior tempo de análise frequentemente passam por menos revisões. Isso pode ocorrer porque PRs mais complexos são revisados com mais cautela ou porque a aprovação pode ser mais direta após um longo tempo de espera.

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

R: O gráfico de descrição versus número de revisões sugere uma relação inversa entre o tamanho da descrição e o número de revisões. PRs com descrições mais curtas tendem a apresentar uma maior variação no número de revisões, com alguns necessitando de mais de 100 revisões, indicando que descrições sucintas ou vagas podem levar a múltiplas iterações e ajustes. A mediana das descrições parece estar

abaixo de 5.000 caracteres, e, nesse intervalo, observa-se uma alta concentração de PRs com poucas revisões.

Entretanto, há outliers significativos: PRs com descrições muito curtas e número elevado de revisões, o que pode indicar falta de clareza nas descrições ou alta complexidade do código. Por outro lado, PRs com descrições mais longas (acima de 50.000 caracteres) geralmente apresentam um número reduzido de revisões, sugerindo que uma explicação detalhada favorece a compreensão e, consequentemente, reduz o número de iterações necessárias.

Isso sugere que PRs bem descritos tendem a passar por menos revisões, enquanto descrições vagas ou incompletas exigem mais iterações. Contudo, mesmo PRs com descrições longas não estão imunes a revisões, especialmente se o código em questão for particularmente complexo.

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

R: O gráfico reforça a hipótese de que o número de interações em PRs está diretamente relacionado ao número de revisões, sugerindo que PRs com mais interações também passam por mais iterações de revisão, possivelmente devido à complexidade ou à necessidade de ajustes contínuos.