Laboratório 03

Gabriel Victor Couto, João Victor Guerra,
Luiz Gustavo

Metodologia

Code Review é uma prática do desenvolvimento de software que consiste na avaliação prévia das alterações feitas em um código antes de fundi-lo à versão principal. Para mais, a prática se baseia em terceiros praticando a análise com o intuito de encontrar erros com maior facilidade. No github esta prática é fortemente abordada pelos mais diversos repositórios através das Pull Requests (PR), ferramenta que permite um integrante do time mostrar as alterações que realizou e disponibilizar essas para análise de colegas.

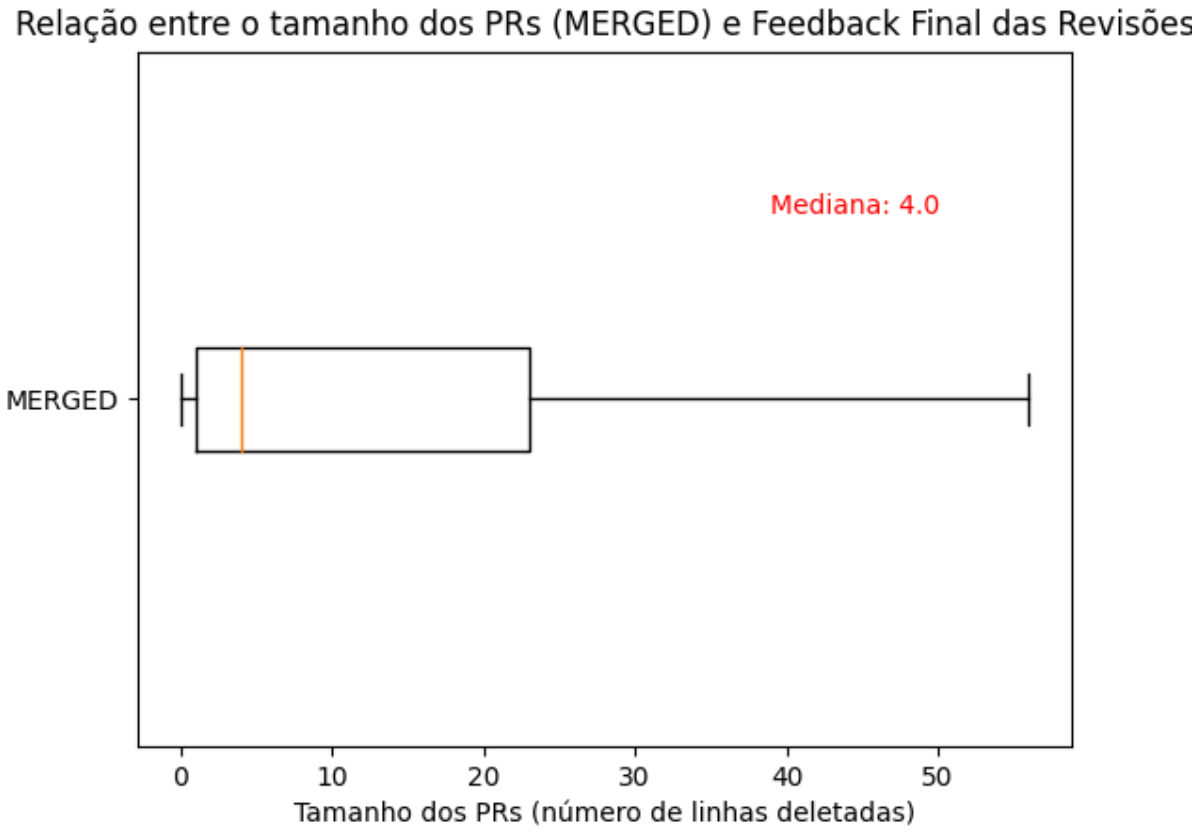
Neste trabalho buscamos analisar a prática do code review por meio da análise de pull requests dentro do github. São utilizados para esse estudo dados como tempo para fechamento ou merge da PR - sendo válidas apenas aquelas cujo tempo é superior a 1 hora, com propósito de ignorar as avaliações por ferramentas automáticas - e que possuam pelo menos uma revisão dos 200 repositórios mais populares com pelo menos 100 PRs.

Devido a problemas relacionados à ultrapassagem do limite de requisições imposto pela API do GitHub e demais problemas identificados no código, não foi possível realizar a mineração para todos os dados necessários. Sendo assim, escolhemos apurar apenas os dados dos 100 primeiros repositórios que se enquadram nas características supracitadas.

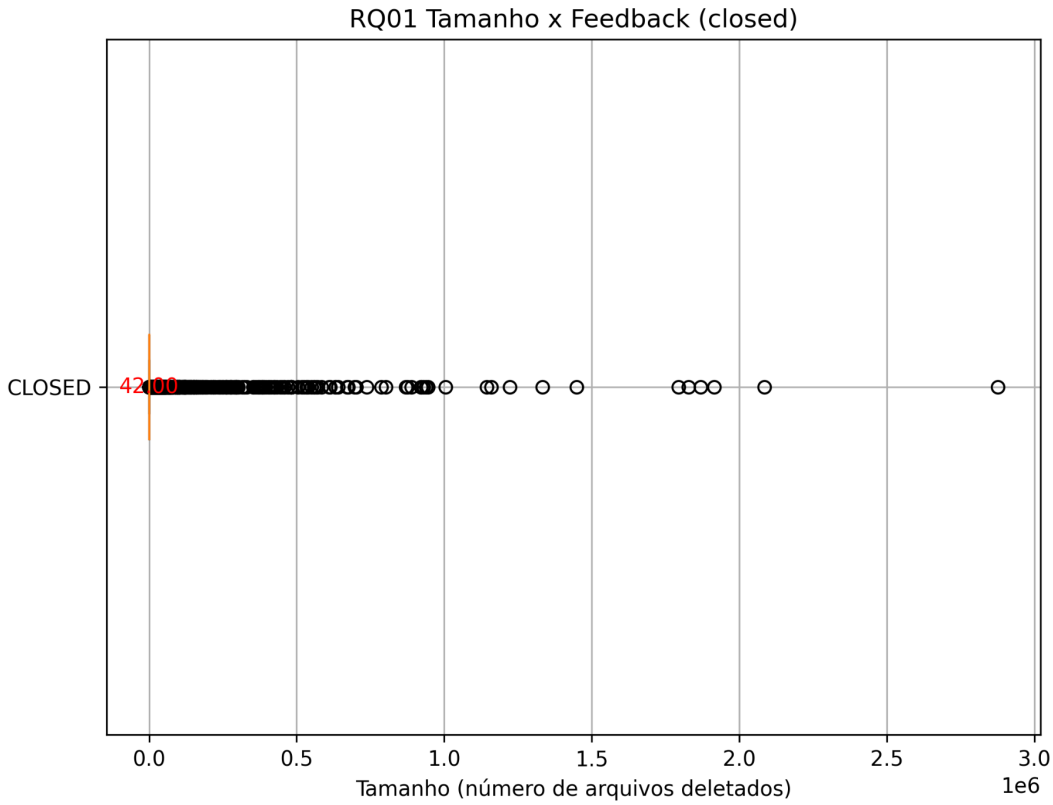
A partir disso, com base em cada RQ, foi criado um script para plotar gráficos de dispersão e boxplots. Para as perguntas do conjunto A foram plotados dois boxplots por pergunta, um para PRs merged e outro para PRs closed. Para as perguntas do conjunto B, foram usados gráficos de dispersão

Resultados Obtidos

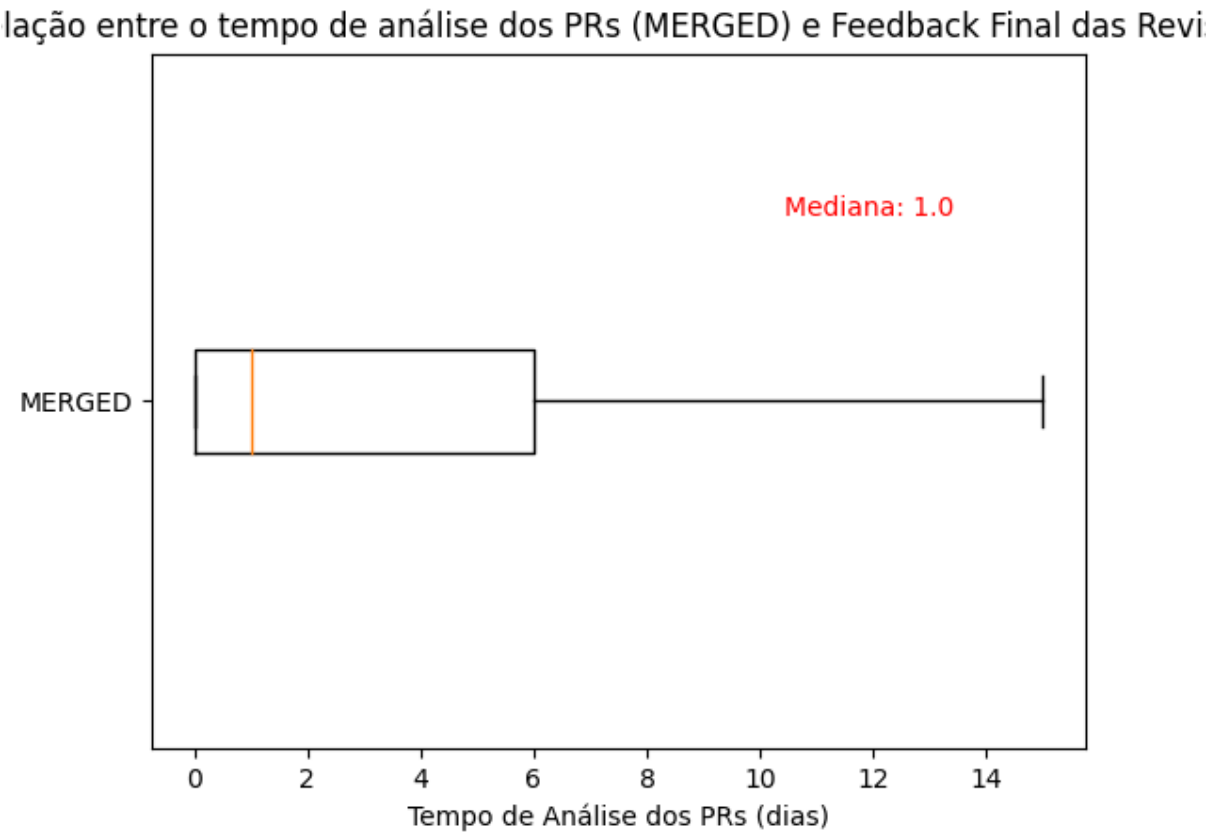
Para responder às oito questões levantadas, foram formulados os seguintes gráficos:



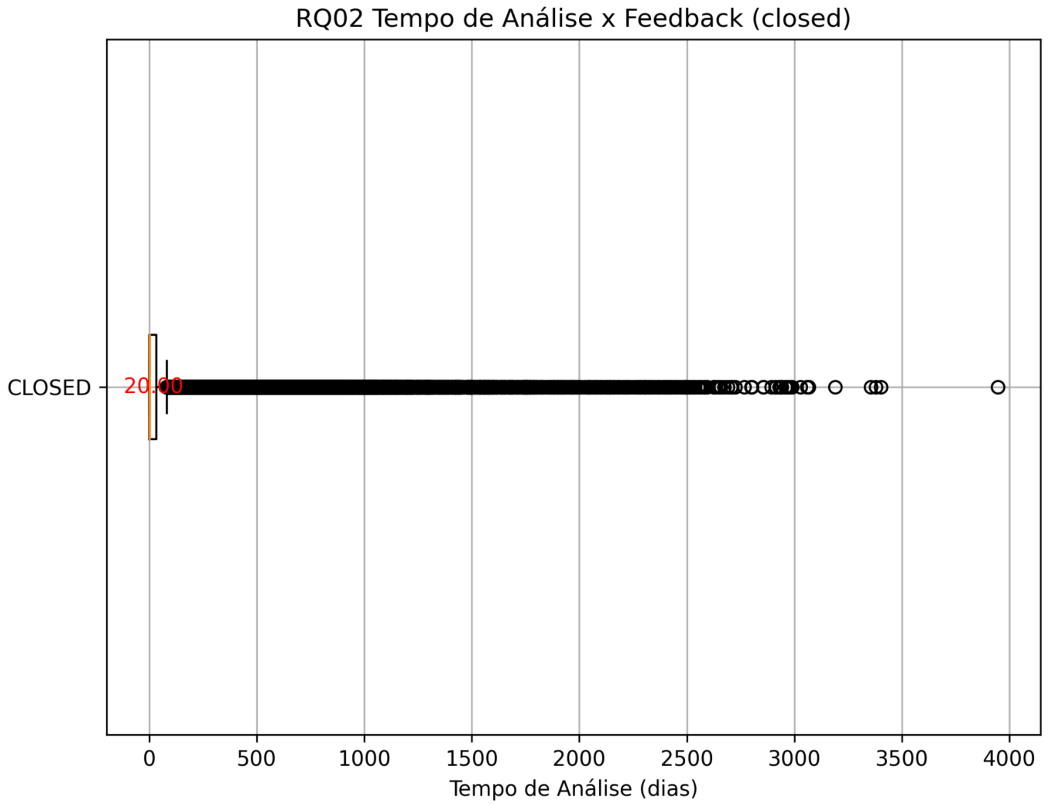
Neste gráfico boxplot podemos ver a distribuição do tamanho dos PRs que foram margeados, apresentando uma mediana de 3.0, mas podendo ir de 0 a quase 60 linhas excluídas.



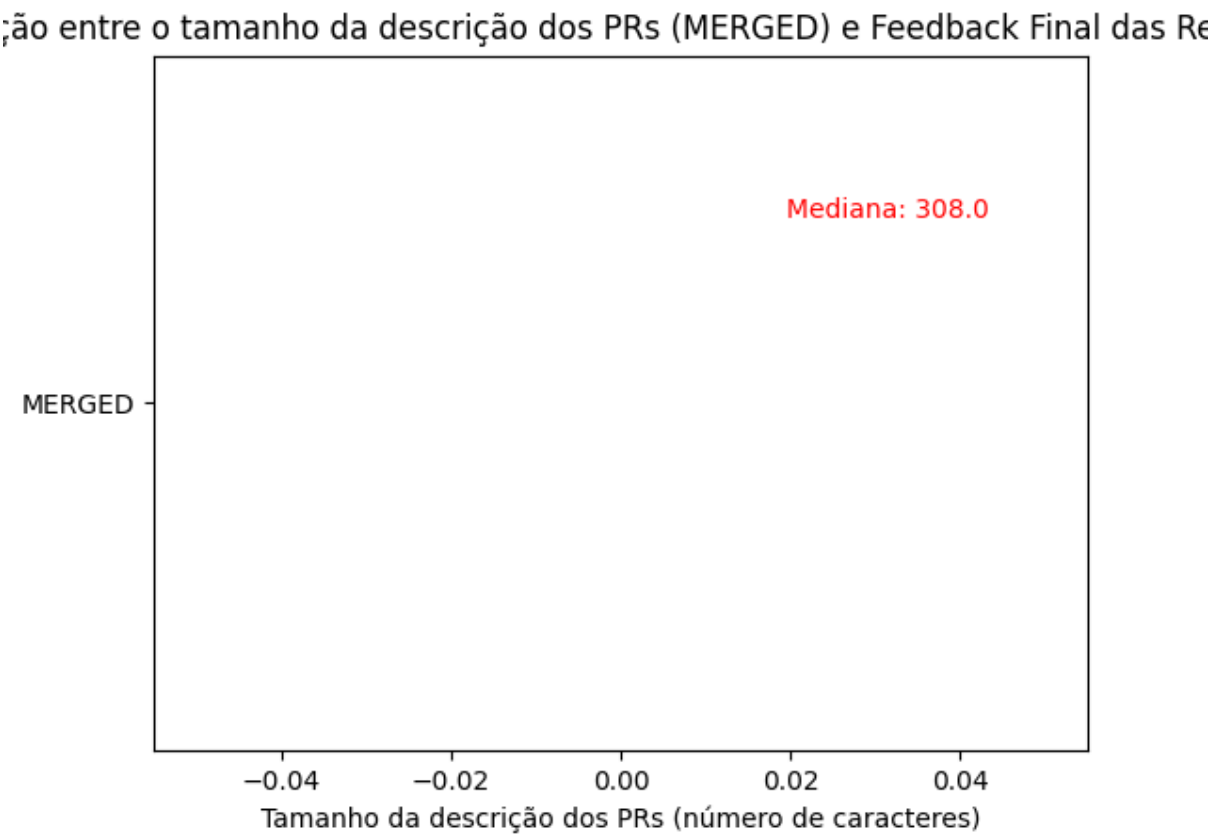
Já para as PRs closed os arquivos tendem a ter muito mais linhas excluídas, podendo alcançar quase 3 milhões. Apesar disso, a maior concentração dos dados está entre 0 e 500 mil linhas.



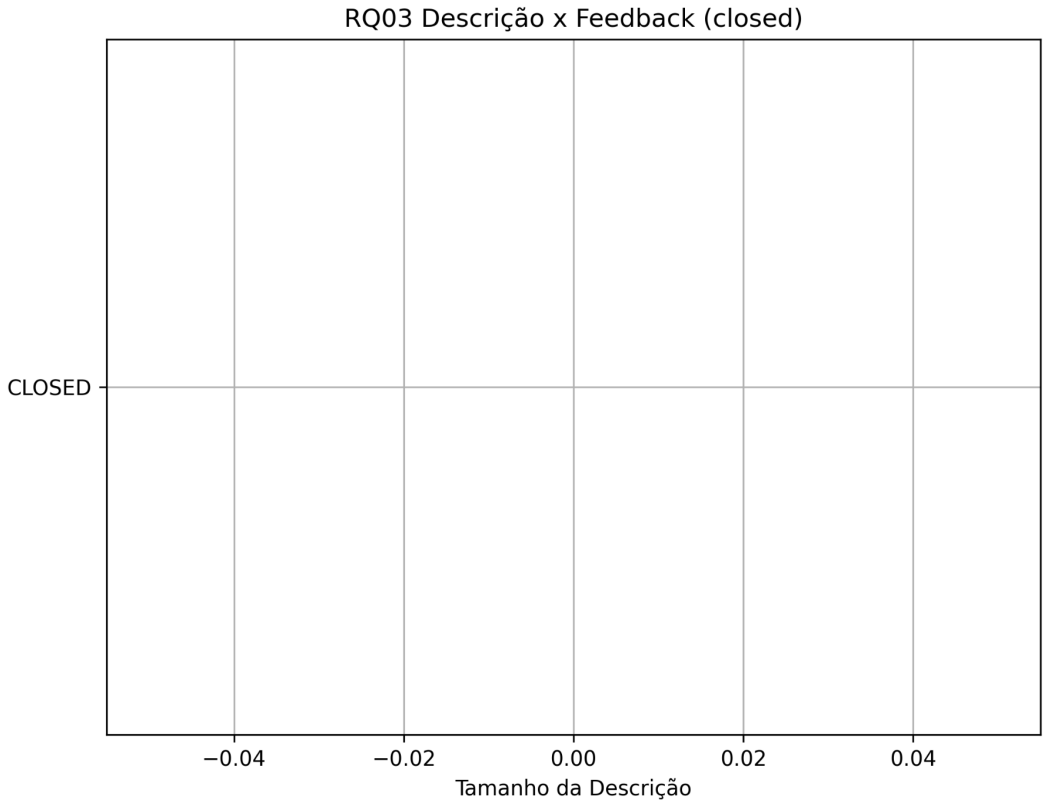
Para as PRs merged, o tempo de análise em dias teve como mediana 1.0, e os dados se concentraram de 0 a 6 dias.



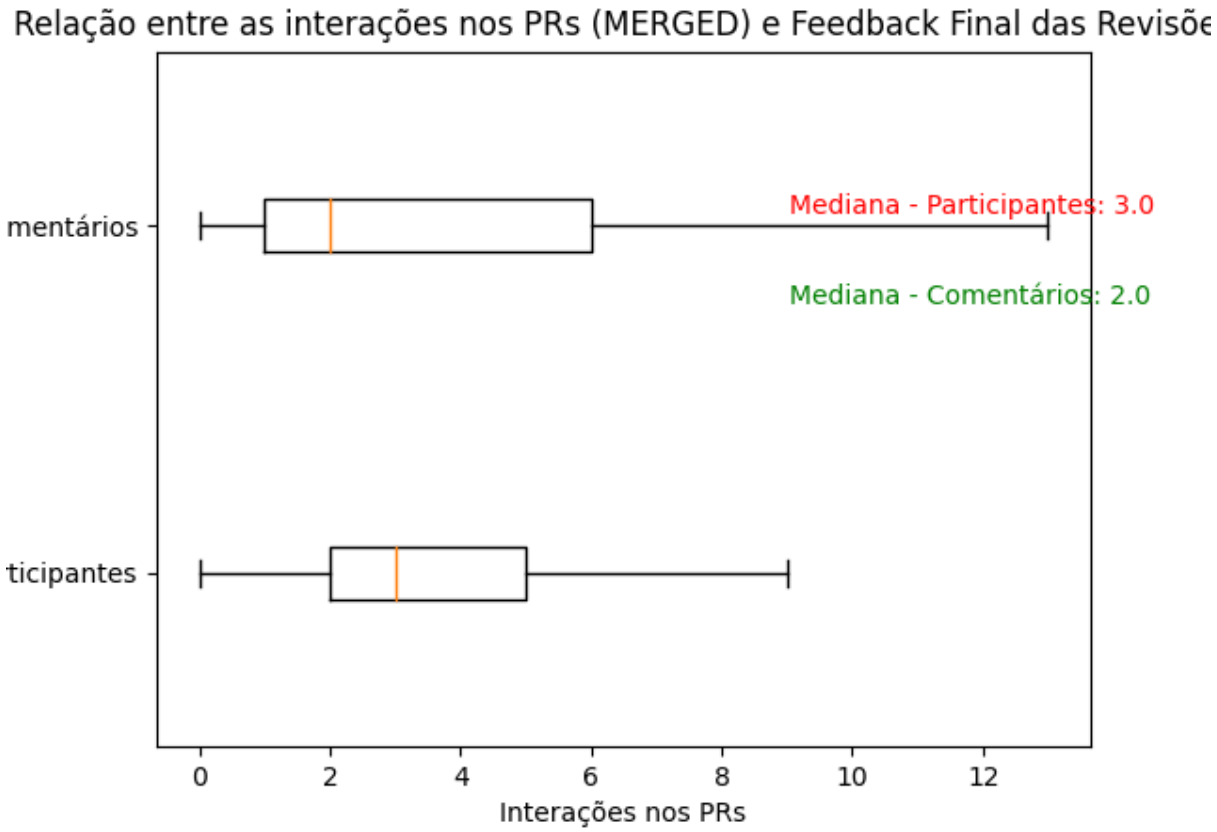
Enquanto isso, as PRs closed possuem valores muito distintos, enquanto a mediana é de 20 dias, os valores podem se aproximar de 4000 dias, apresentando múltiplos casos que demoraram mais de um ano para serem fechados.



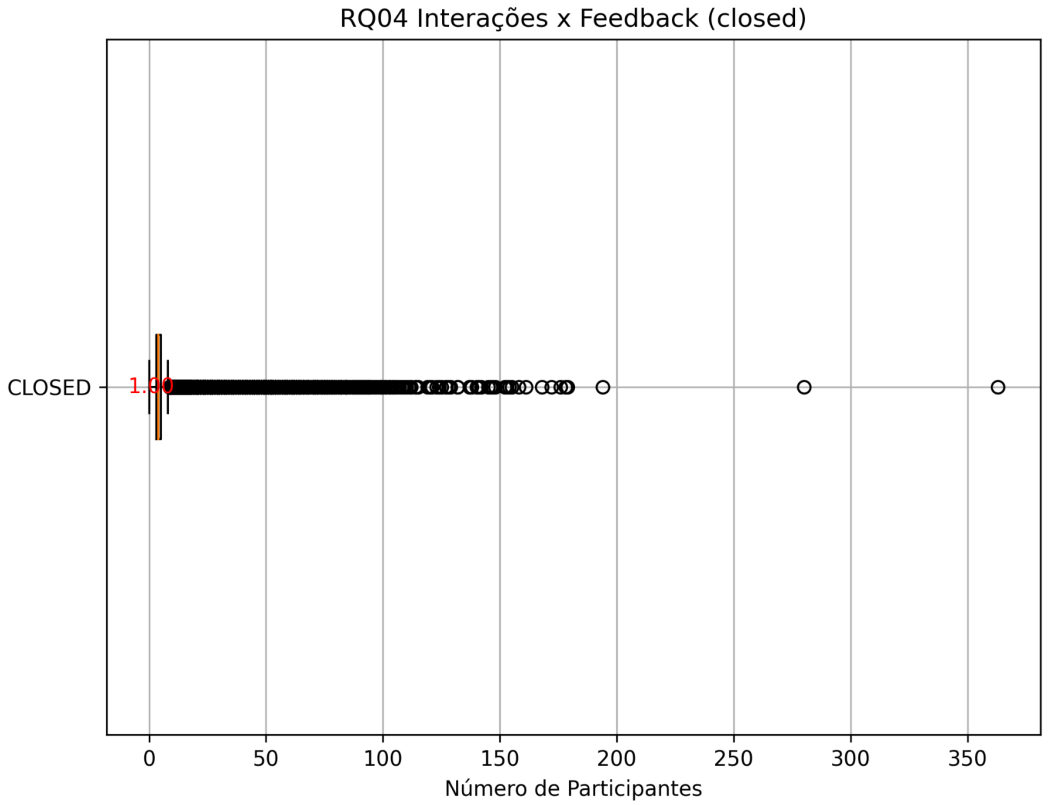
Quanto à relação de PRs merged e o tamanho das descrições não foi possível plotar um gráfico, contudo a mediana de 308.0 foi obtida.



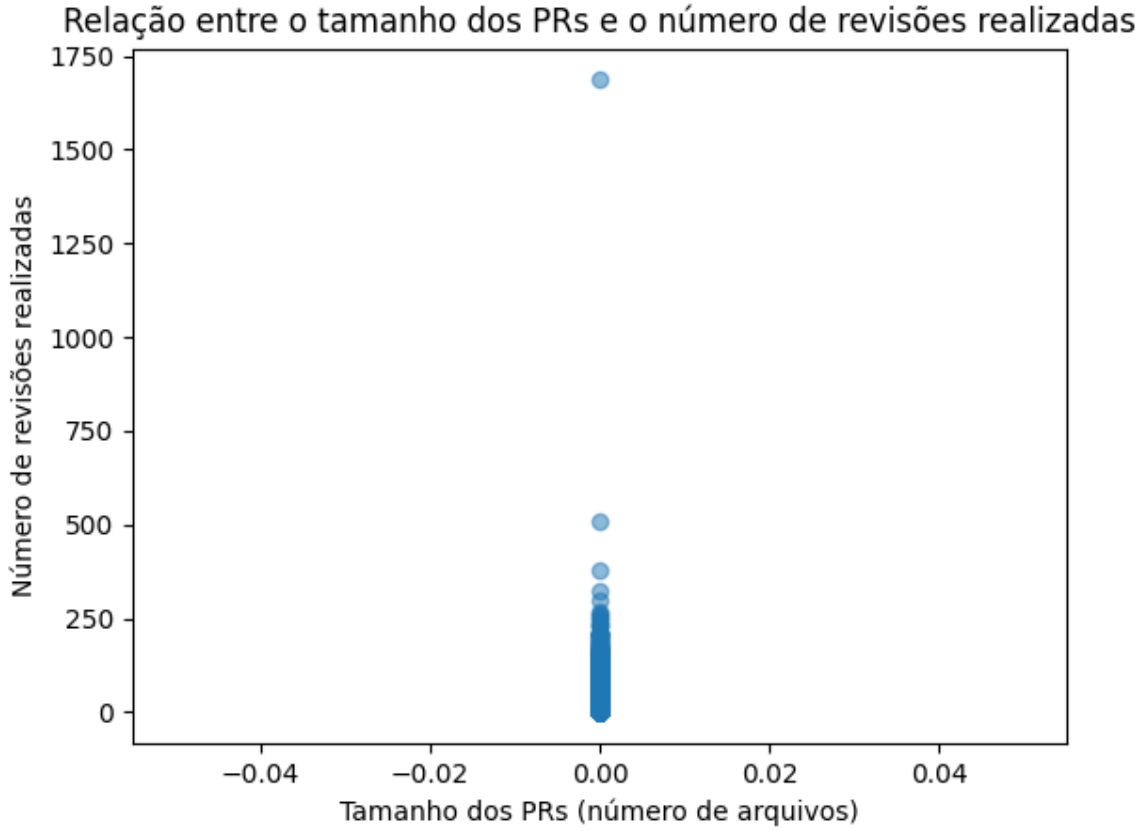
Não foi possível plotar o gráfico para as PRs closed, além disso, neste caso também não houve o cálculo da mediana.



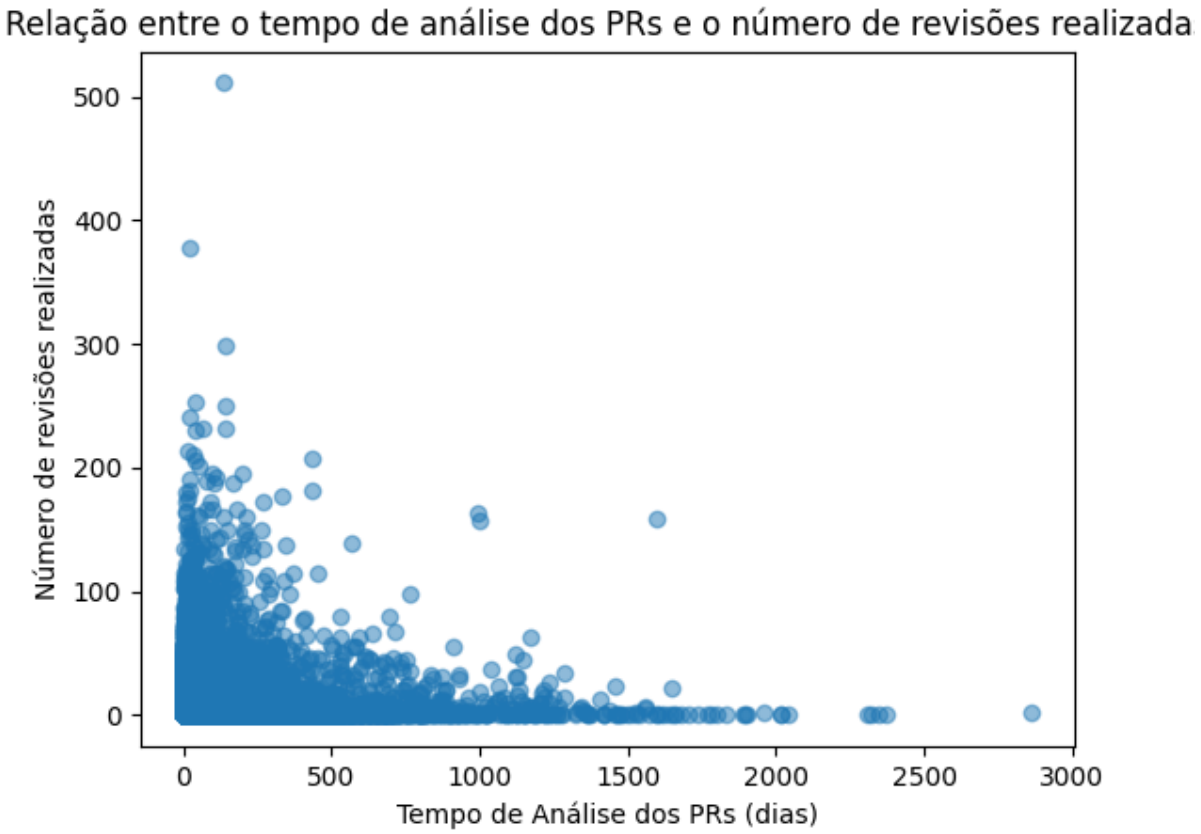
Já para a interação com PRs merged foram criados dois boxplots, um para quantidades de comentários, com mediana igual a 2 e valores se agrupando, principalmente, entre 1 e 6, e sobre a quantidade de participantes, com mediana igual a 3 e valores agrupando entre 2 e 5.



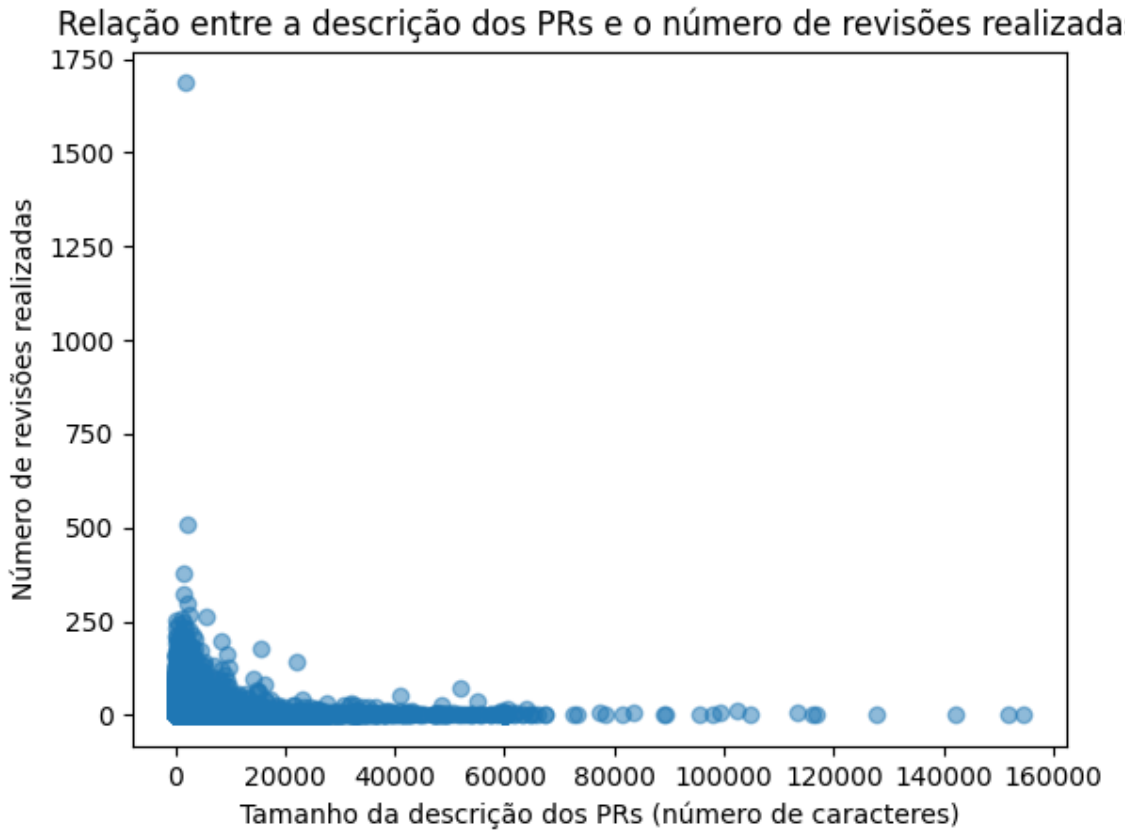
Nas PRs closed e o número de participantes, temos valores podendo passar de 350, contudo a maior parte se concentra entre 0 e 150, com mediana igual a 1.



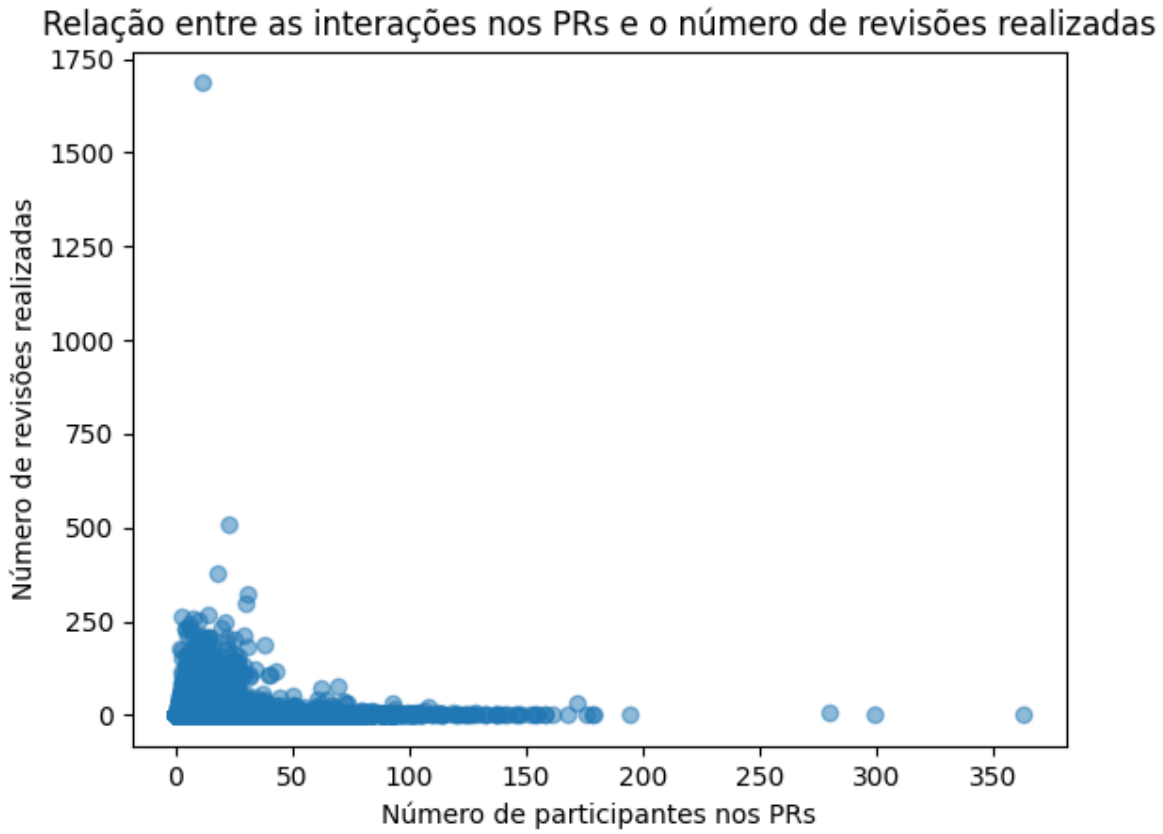
Apesar do número de revisões ser projetado adequadamente no gráfico, o tamanho das PRs foi representado sempre como 0, independente de seu valor real.



A partir da análise do gráfico podemos ver que os dados se concentram principalmente entre 0 e 100 revisões realizadas, mas podendo chegar a 500 revisões, e 0 e 500 dias para análise, podendo alcançar 3000 dias.



Quanto à relação entre o tamanho da descrição e o número de revisões vemos valores que podem chegar próximos a 160 mil caracteres na descrição, e os dados se concentram principalmente de 0 a 60 mil. Já o número de revisões se concentra abaixo de 250.



O número de participantes em uma PR se concentra abaixo de 150, podendo alcançar valores maiores de 350 em casos especiais.

Hipóteses

A. Feedback Final das Revisões (Status do PR):

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

R: É de se esperar que não haja uma correlação direta entre o tamanho de uma PR e o feedback final.

Conclusão: A partir dos gráficos a maior parte das PRs merged é pequena, ou seja, possui poucas linhas deletadas, dito isto, pode-se concluir que uma PR menor ajude a concluir o merge, já que a chance de haver conflitos ou erros sutis- como erros de padrão de escrita de código- diminui.

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

R: Espera-se que o tempo de análise não interfira no feedback final das revisões, pois são características muito distintas que, a princípio, não traçam nenhuma correlação entre si.

Conclusão: A partir da análise dos gráficos foi visto que existe sim uma relação entre feedback final e tempo de análise, pois PRs merged possuem em grande maioria valores menores de dias para análise quando comparados as PRs closed.

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

R: Pode-se esperar que quanto maior for o número de caracteres no corpo de uma PR (descrição), mais chances há de que seja aprovada, partindo do pressuposto de que o autor da PR se preocupou em documentar o esforço realizado, e esse esforço se refletiu na resolução de uma issue.

Conclusão: Houveram dificuldades na produção dos gráficos para esta questão, sendo assim foi impossível concluir algo de valor.

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

R: Espera-se que quanto mais interações uma PR possui, menores são as chances de que seja aprovada, pois vários usuários podem ter pontuados erros encontrados e inconsistências.

Conclusão: A hipótese está correta, visto que os dados mostram que PRs closed possuem muito mais participantes do que as merged.

B. Número de Revisões:

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

R: É esperado que quanto maior o tamanho de uma PR, maior também seja o número de revisões, no intuito de que todas as alterações sejam revisadas correta e eficientemente.

Conclusão: Não foi possível tirar conclusões relevantes devido ao erro na produção do gráfico.

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

R: Espera-se que PRs com muitas revisões também tenham um tempo de análise maior, visto que a mesma deve se adequar aos comentários pontuados pelos revisores.

Conclusão: Não podemos afirmar a relação entre os dados a partir do gráfico visto a quantidade grande de dados.

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

R: Espera-se que a descrição dos PRs e o número de revisões realizadas não tenham uma correlação direta, pois são métricas distintas e separadas.

Conclusão: Os dados na realidade estão relacionados, de modo que as análises com mais revisores se concentram em descrições pequenas (menos de 20 mil caracteres). Sendo assim, quanto maior a descrição menor a probabilidade de feedback negativo a PR.

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

R: Espera-se que quanto mais interações numa PR, maior o número de revisões realizadas, para que os erros pontuados sejam corrigidos e adequados.

Conclusão: A partir da análise do gráfico não podemos definir uma existência de correlação entre esses dois dados.