Caracterizando a atividade de code review no GitHub

Introdução

A análise de code review em projetos open source no GitHub é fundamental para entender os fatores que influenciam na aceitação ou rejeição de contribuições. Este estudo focou em Pull Requests (PRs) dos 200 repositórios mais populares, avaliando como diferentes características dos PRs afetam seu resultado e o processo de revisão.

Metodologia

Coleta de Dados

1. Seleção de Repositórios:

○ Os 200 repositórios mais populares do GitHub (com mais de 10.000 estrelas).

○ Repositórios com pelo menos 100 PRs (MERGED + CLOSED).

2. Filtragem de PRs:

○ PRs com status MERGED ou CLOSED.

○ Pelo menos uma revisão (não automática).

○ Tempo mínimo de revisão de 1 hora (para excluir revisões

automatizadas).

Teste de Correlação

Para a análise estatística das relações entre as variáveis, optamos por utilizar a correlação de **Spearman**.
Essa escolha foi feita porque:

- **Spearman** é uma correlação não paramétrica, ou seja, **não assume que os dados seguem uma distribuição normal**.

- Muitas das métricas analisadas (como número de arquivos modificados, número de linhas alteradas, número de participantes) **não apresentam distribuições normais** e podem conter **outliers**.

- O teste de Spearman **mede relações monotônicas** (não necessariamente lineares), sendo mais adequado para capturar padrões em dados com comportamento não-linear ou escalas diferentes.

Assim, a correlação de Spearman oferece **uma análise mais robusta e segura** para o perfil dos dados coletados neste estudo.

Hipóteses e Resultados

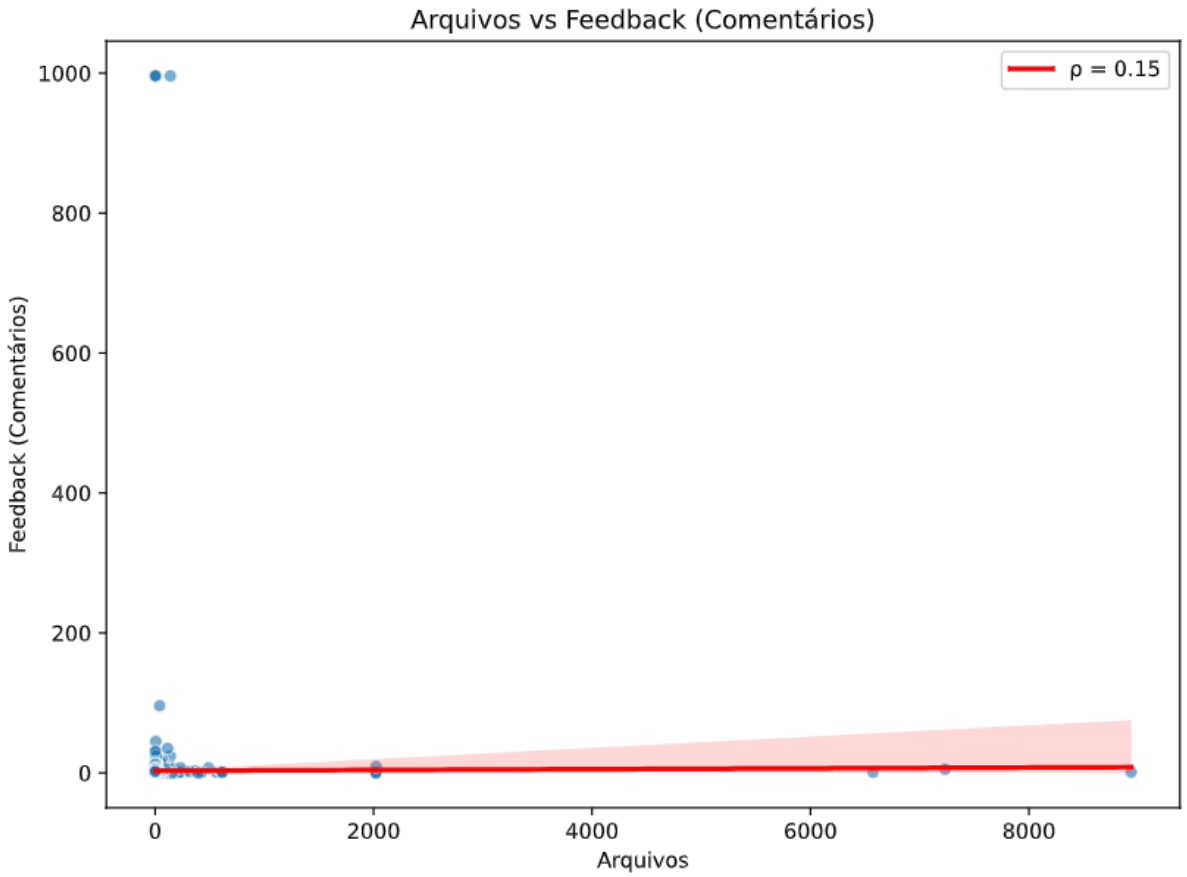
Dimensão A: Feedback Final das Revisões (Status do PR)

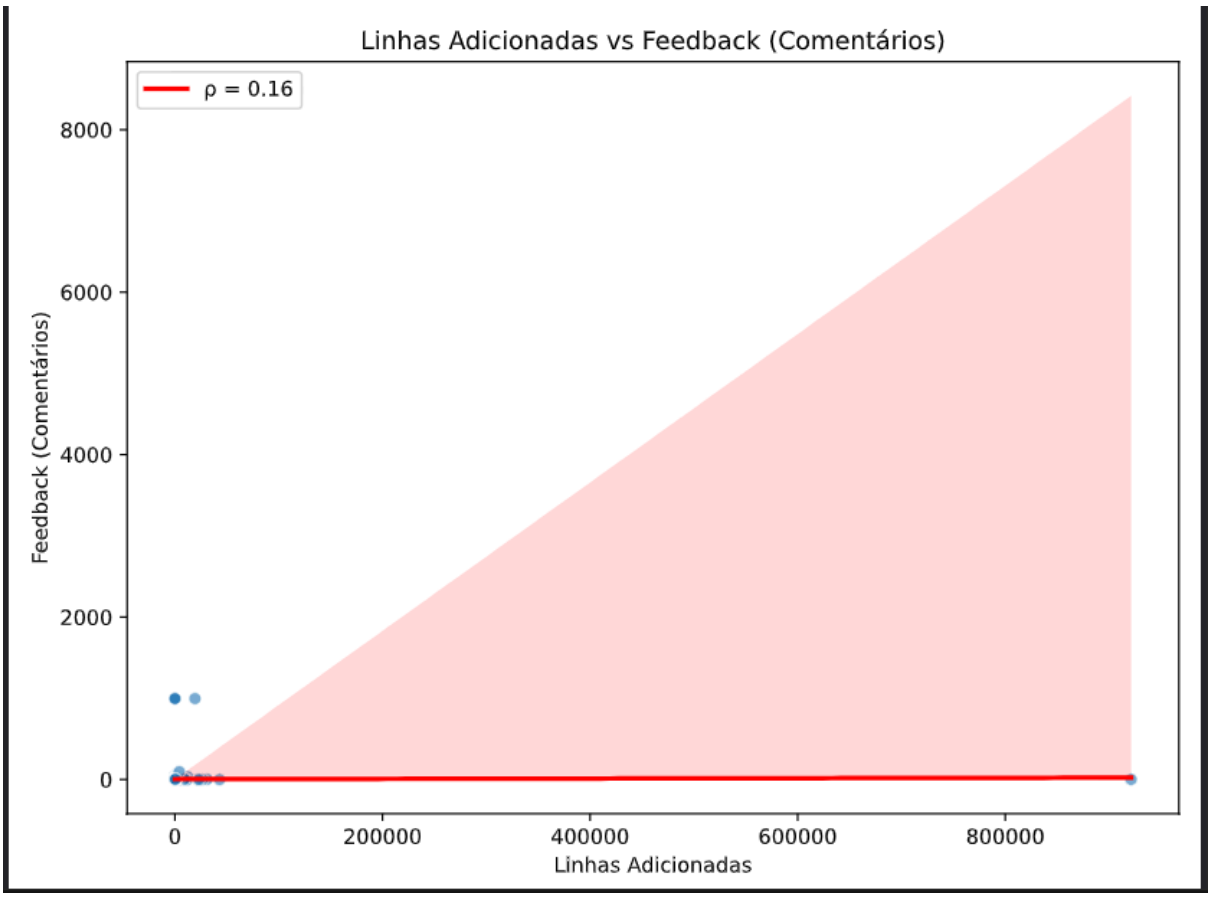
RQ01 - Tamanho dos PRs vs Feedback Final
Hipótese: PRs menores têm maior chance de serem aceitos.
Resultado:

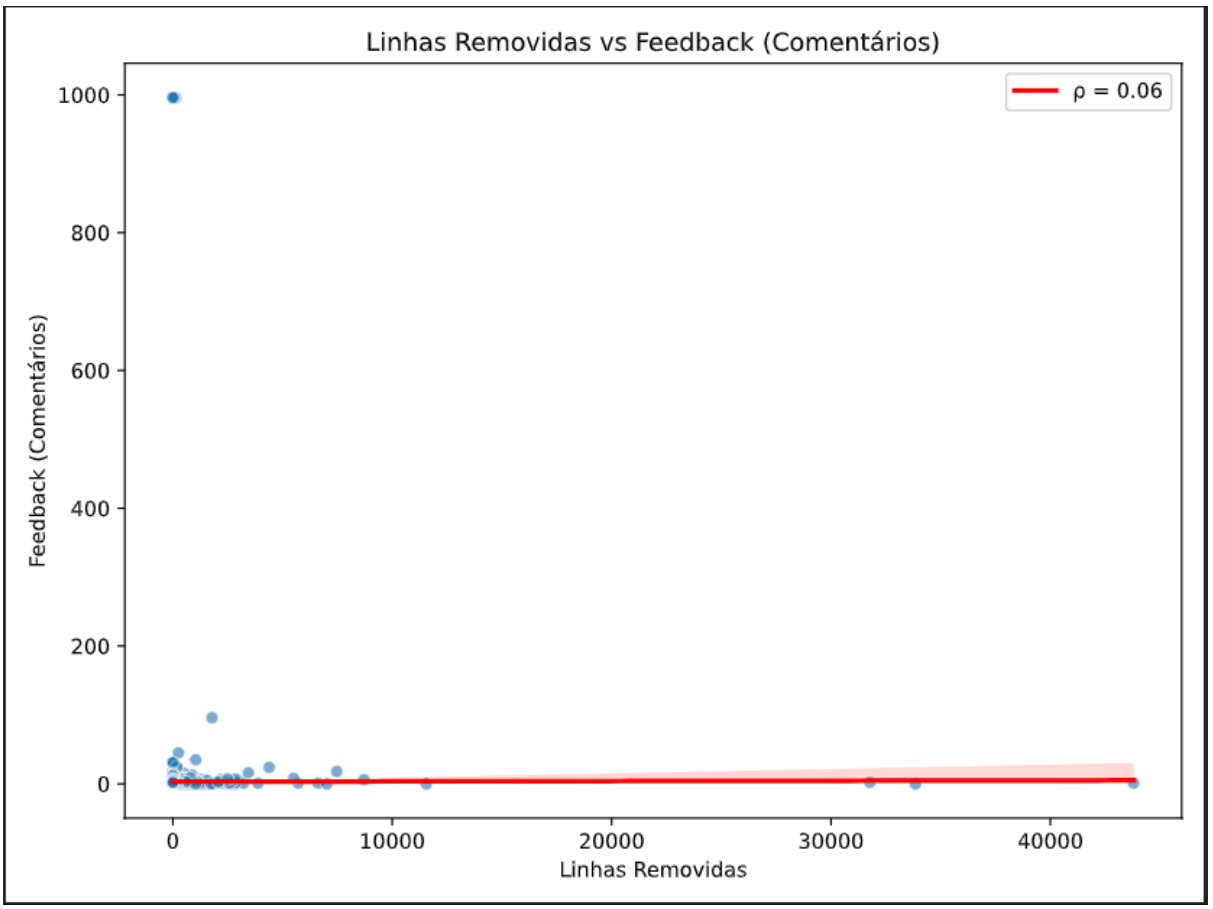
- Correlação entre **Arquivos modificados** e **Feedback (Comentários)**: **ρ = 0.15**

- Correlação entre **Linhas Adicionadas** e **Feedback**: **ρ = 0.16**

- Correlação entre **Linhas Removidas** e **Feedback**: **ρ = 0.06**


Arquivos vs Feedback (Comentários)
ρ = 0.15
Feedback (Comentários)
1000
800
600
400
200
0
0
2000
4000
6000
8000
Arquivos


Linhas Adicionadas vs Feedback (Comentários)
ρ = 0.16
Feedback (Comentários)
8000
6000
4000
2000
0
0
200000
400000
600000
800000
Linhas Adicionadas

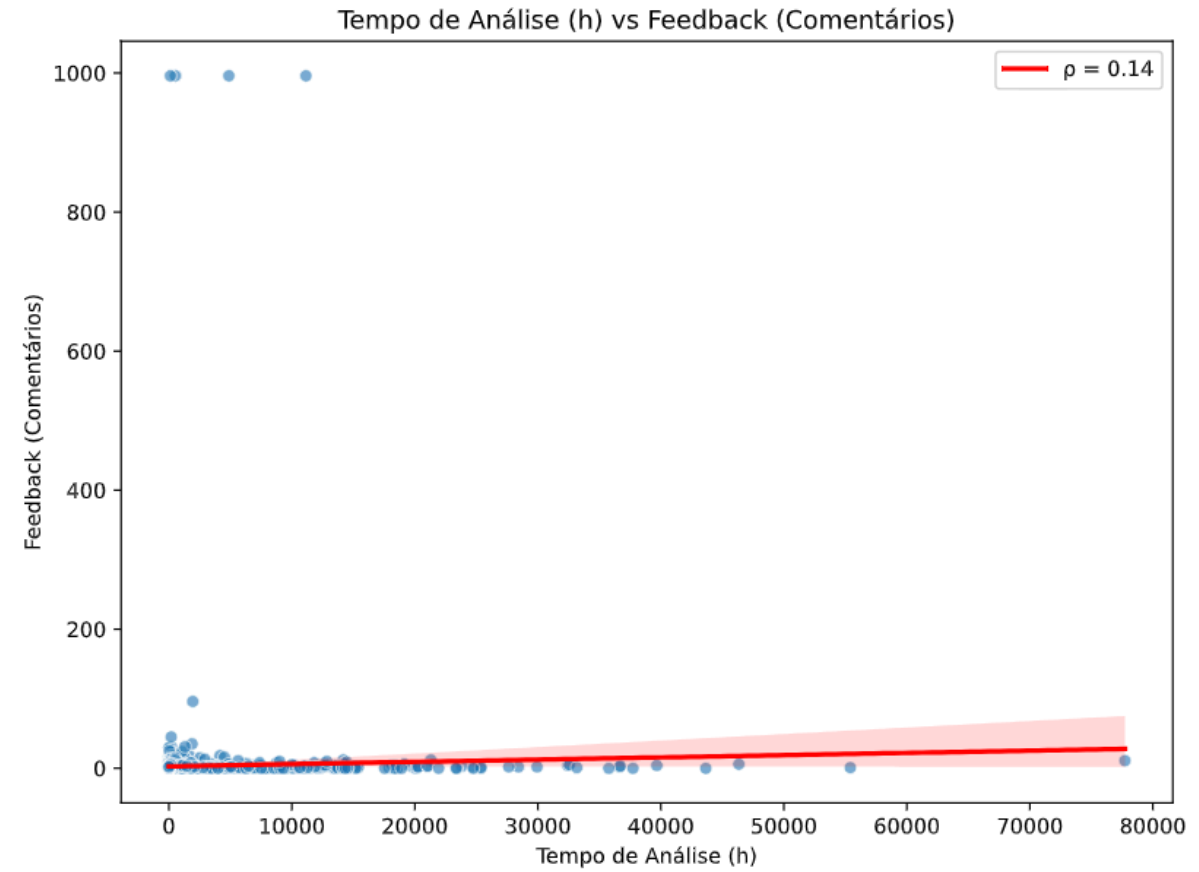

Linhas Removidas vs Feedback (Comentários)
ρ = 0.06
Feedback (Comentários)
1000
800
600
400
200
0
0
10000
20000
30000
40000
Linhas Removidas

As correlações são positivas mas fracas, sugerindo que PRs maiores tendem a ter ligeiramente mais feedback, mas o efeito é fraco. A hipótese não foi fortemente confirmada.

RQ02 - Tempo de Análise vs Feedback Final
Hipótese: PRs com maior tempo de análise têm maior chance de rejeição.
Resultado:

- Correlação entre **Tempo de Análise** e **Feedback**: **ρ = 0.14**

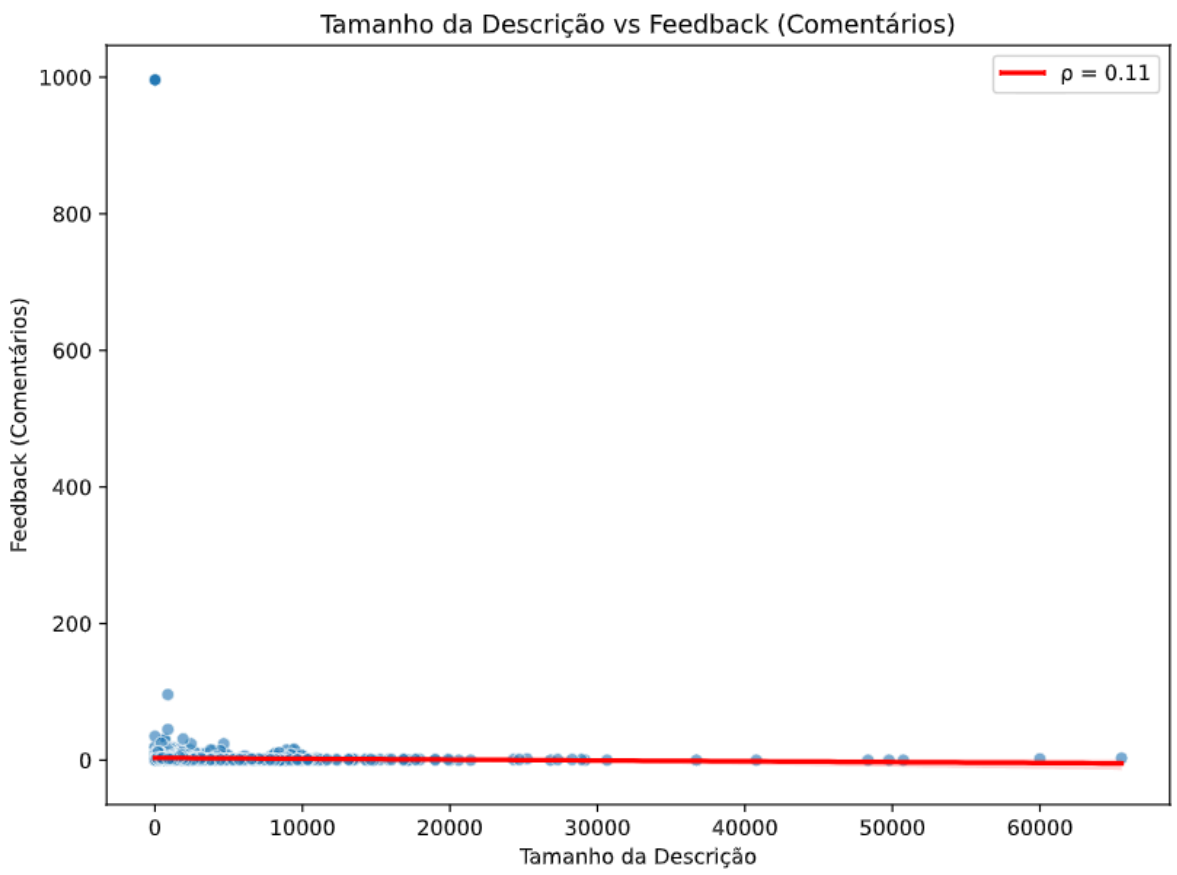
Existe uma correlação fraca e positiva. PRs com mais tempo de análise tendem a ter um pouco mais de feedback. A hipótese é parcialmente sustentada.



RQ03 - Descrição dos PRs vs Feedback Final
Hipótese: PRs com descrições mais detalhadas têm maior chance de serem aceitos.
Resultado:

- Correlação entre **Tamanho da Descrição** e **Feedback**: **ρ = 0.11**

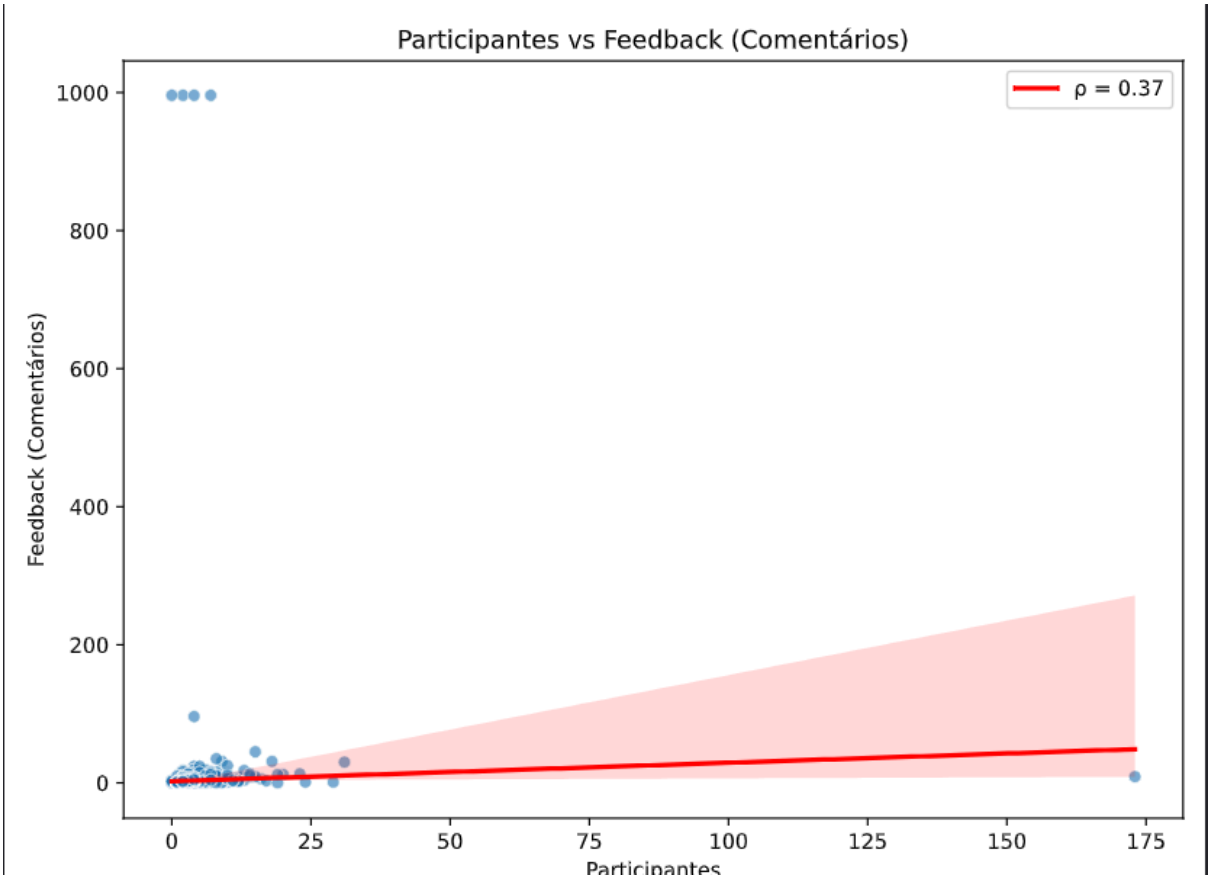
Correlação fraca positiva. PRs com descrições maiores geram um pouco mais de feedback. A hipótese é parcialmente sustentada, mas o efeito é fraco.



RQ04 - Interações nos PRs vs Feedback Final
Hipótese: PRs com mais interações têm maior chance de rejeição.
Resultado:

- Correlação entre **Participantes** e **Feedback**: **ρ = 0.37**

Correlação moderada positiva. PRs com mais participantes tendem a gerar mais feedback, sustentando a hipótese de que mais interações podem dificultar o merge.

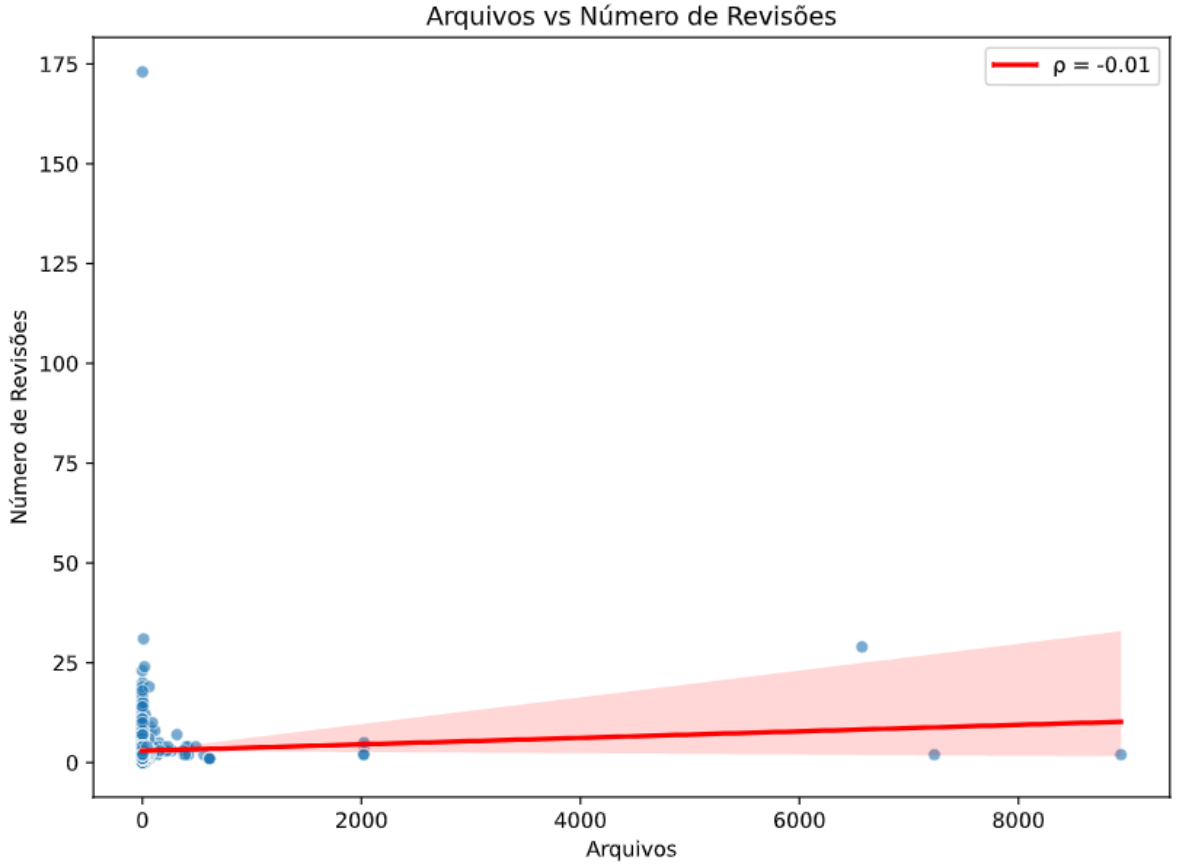


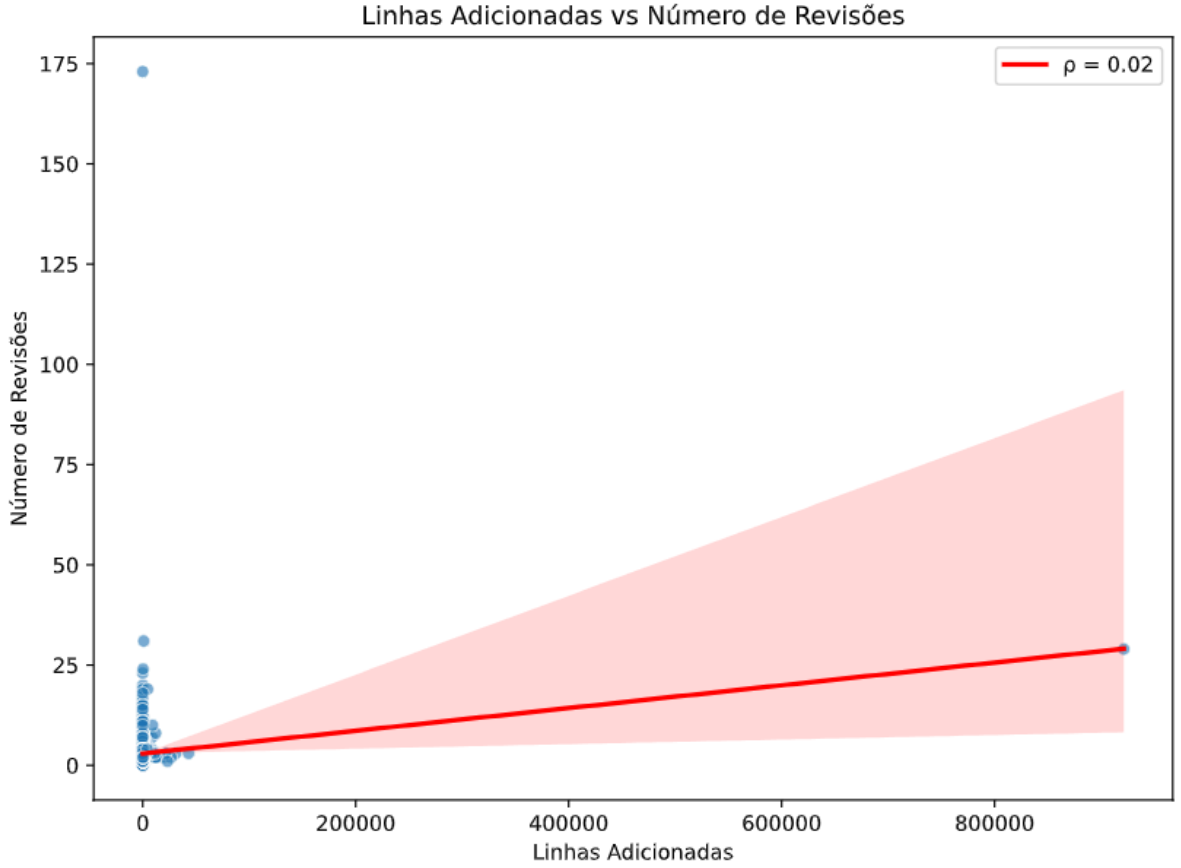
Dimensão B: Número de Revisões

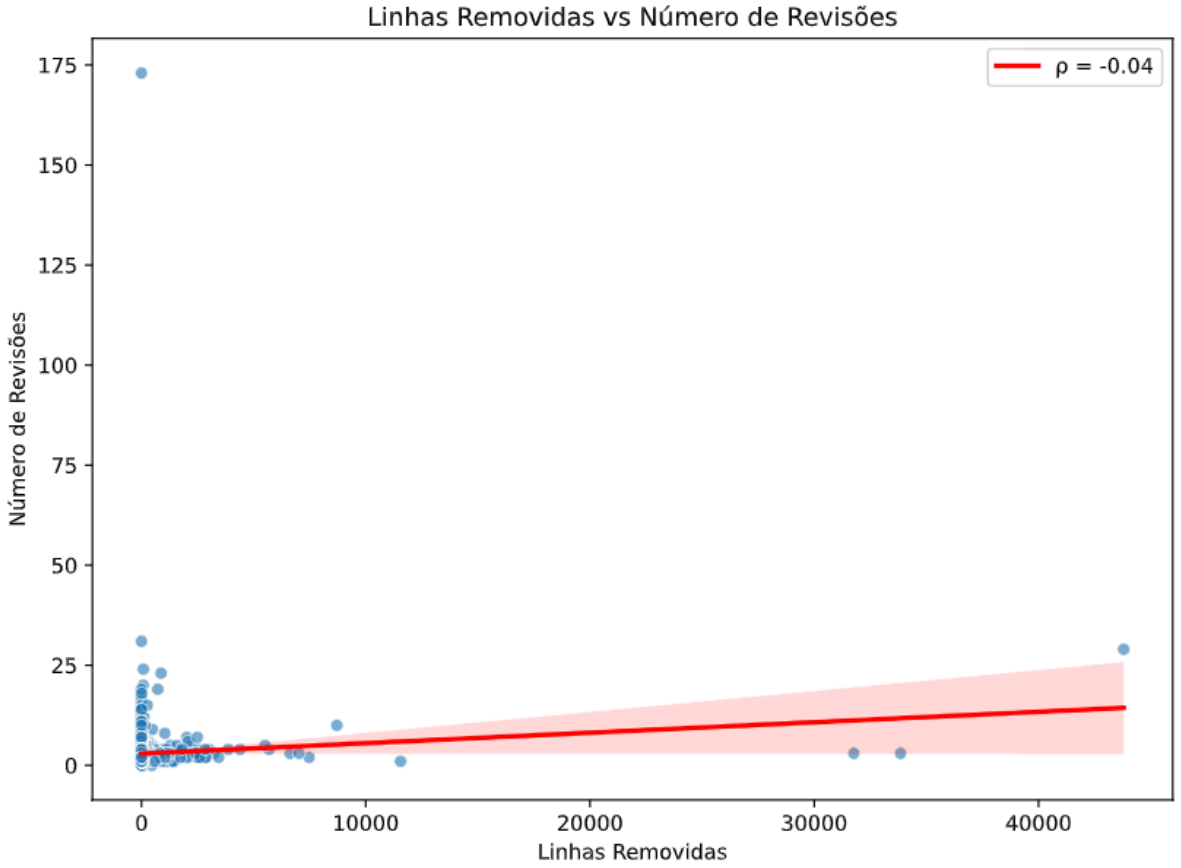
RQ05 - Tamanho dos PRs vs Número de Revisões
Hipótese: PRs maiores exigem mais revisões.
Resultado:

- Correlação entre **Arquivos modificados** e **Número de Revisões**: **ρ = -0.01**
- Correlação entre **Linhas Adicionadas** e **Número de Revisões**: **ρ = 0.02**
- Correlação entre **Linhas Removidas** e **Número de Revisões**: **ρ = -0.04**

Não há correlação relevante. A hipótese foi refutada: PRs maiores não exigem mais revisões.


Arquivos vs Número de Revisões
ρ = -0.01
Número de Revisões
Arquivos
0
25
50
75
100
125
150
175
0
2000
4000
6000
8000

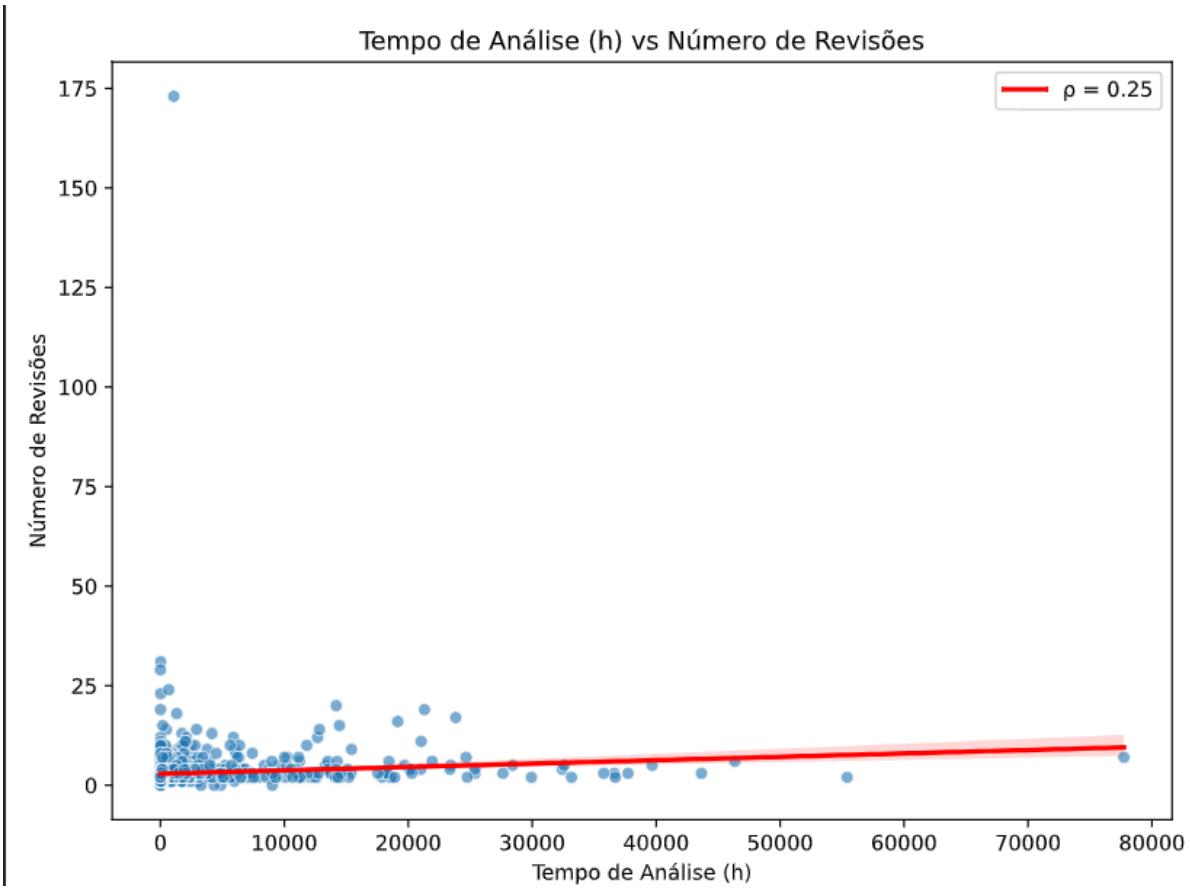

Linhas Adicionadas vs Número de Revisões
ρ = 0.02
Número de Revisões
Linhas Adicionadas
0
25
50
75
100
125
150
175
0
200000
400000
600000
800000


Linhas Removidas vs Número de Revisões
ρ = -0.04
Número de Revisões
Linhas Removidas
0
25
50
75
100
125
150
175
0
10000
20000
30000
40000

RQ06 - Tempo de Análise vs Número de Revisões
Hipótese: PRs com mais revisões tendem a ter maior tempo de análise.
Resultado:

- Correlação entre **Tempo de Análise** e **Número de Revisões**: **ρ = 0.25**

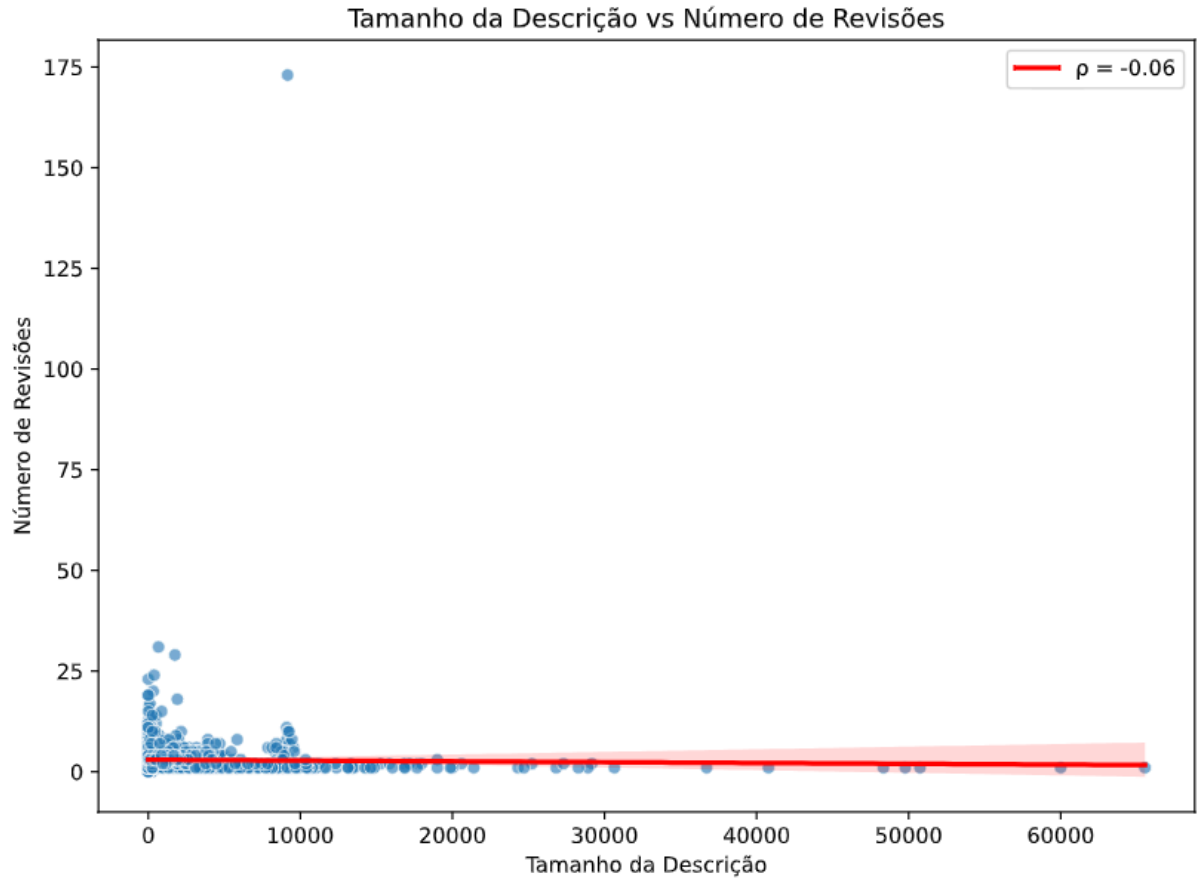
Correlação moderada positiva. A hipótese foi confirmada.



RQ07 - Descrição dos PRs vs Número de Revisões
Hipótese: PRs com descrições mais claras exigem menos revisões.
Resultado:

- Correlação entre **Tamanho da Descrição** e **Número de Revisões**: **ρ = -0.06**

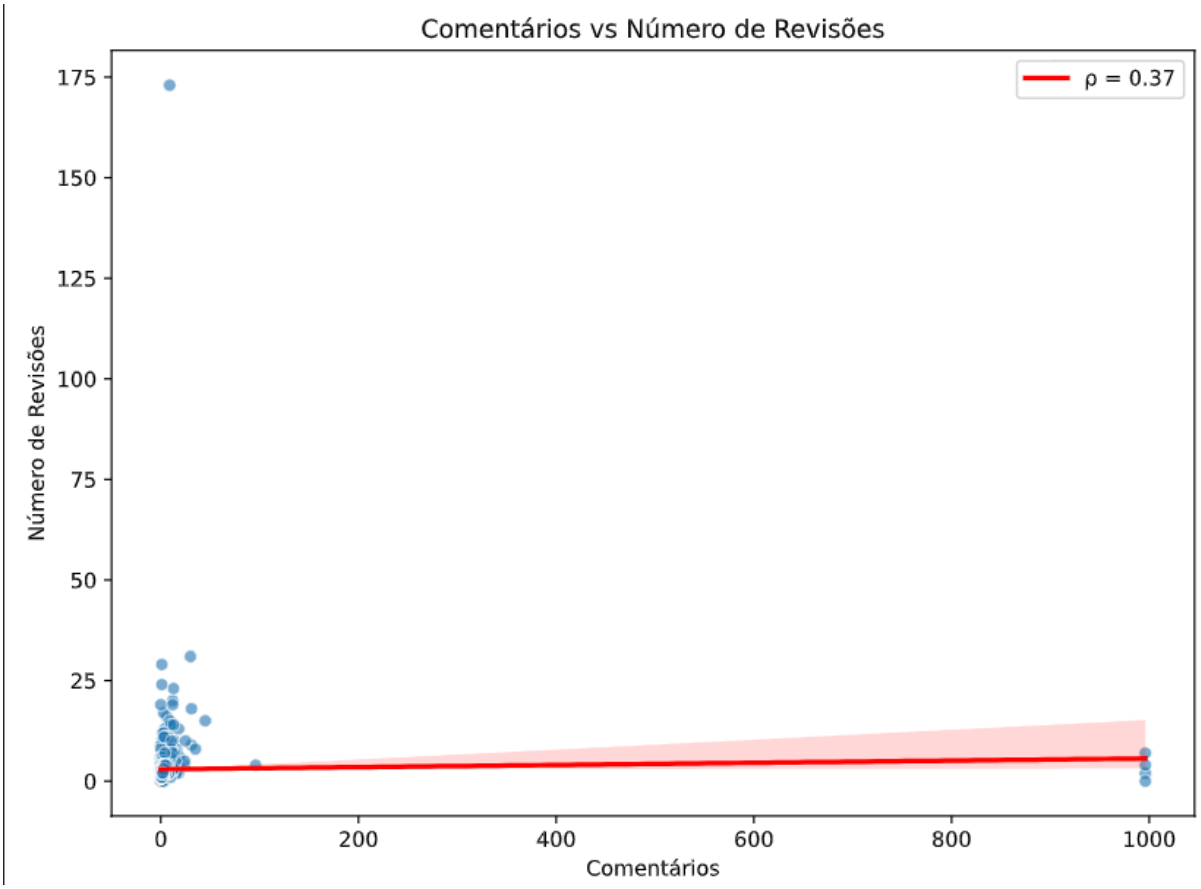
Correlação negativa fraca. A hipótese foi parcialmente confirmada.



RQ08 - Interações nos PRs vs Número de Revisões
Hipótese: PRs com mais interações têm mais revisões.
Resultado:

- Correlação entre **Comentários** e **Número de Revisões**: **ρ = 0.37**

Correlação moderada positiva. A hipótese foi confirmada.



Conclusão

- As características que mais se relacionam com aumento no feedback e número de revisões são o número de **participantes** e **comentários** nos PRs.
- Tamanho dos PRs (arquivos e linhas) não mostrou grande impacto.
- Tempo de análise e tamanho da descrição mostraram impactos leves.
- No geral, o **nível de interação** no PR parece ser o fator mais importante no processo de revisão.