Pontifícia Universidade Católica de Minas Gerais

Laboratório de Experimentação de Software

Pesquisa: Características de Repositórios populares

Giovanni Bogliolo Sirihal Duarte
Luiz Gustavo Mendes Santos
Pedro Ramos Vidigal

Belo Horizonte
2023

1. Introdução

Este relatório propõe investigar as características que definem os repositórios populares open-source no GitHub. Para isso, foram realizadas coletas de dados, incluindo métricas como idade do repositório, número de pull requests aceitas, total de releases e tempo decorrido desde a última atualização. A partir dessa base de informações, foram formuladas hipóteses, padrões e tendências relevantes para essa classe de projetos.

A análise foi viabilizada por um processo de mineração de dados, implementado por meio de um script em Python que utiliza a linguagem de consulta GraphQL. Esse script coletou dados de 1000 repositórios, gerando respostas individuais que foram armazenadas em arquivos .csv, além de calcular a média dos valores obtidos. Com base nesses dados, foram construídos gráficos boxplot e histogramas para uma visualização mais clara das informações.

É importante ressaltar que todos os dados foram coletados na data de envio deste relatório e, portanto, podem não refletir o estado atual dos repositórios analisados.

1.1. Hipóteses

RQ 01. Sistemas populares são maduros/antigos?

R: Sim, sistemas mais antigos têm mais tempo para adquirir popularidade através do desenvolvimento contínuo, aprimoramento e reconhecimento pela comunidade.

RQ 02. Sistemas populares recebem muita contribuição externa?

R: Não, sistemas populares provavelmente recebem inúmeros pull requests. Porém, um padrão alto de código deve ser exigido, o que resulta em um número baixo de contribuições externas no total de pull requests aceitos.

RQ 03. Sistemas populares lançam releases com frequência?

R: Provavelmente um desenvolvimento ativo, correção de bugs e adição de novas funcionalidades, mantém o projeto atualizado e atraem novos usuários para o Repositório, fazendo com que ele mantenha sua popularidade alta.

RQ 04. Sistemas populares são atualizados com frequência?

R:Sim, sistemas populares são provavelmente atualizados com frequência. A atualização diária mantém o projeto relevante em um ambiente tecnológico em constante evolução.

RQ 05. Sistemas populares são escritos nas linguagens mais populares?

R: Sim, é provável que linguagens populares possuem comunidades maiores com mais recursos e bibliotecas disponíveis, criar repositórios com essas linguagens facilita o desenvolvimento e o aprendizado de novos usuários.

RQ 06. Sistemas populares possuem um alto percentual de issues fechadas?

R: Sim, provavelmente sistemas populares tendem a ter um alto percentual de issues fechadas. Isso indicaria uma equipe de desenvolvimento comprometida em resolver os Problemas, o que acabaria melhorando a qualidade do sistema e aumentando a sua popularidade.

RQ 07. Sistemas escritos em linguagens mais populares recebem mais contribuição externa, lançam mais releases e são atualizados com mais frequência?

R: Sim, a tendência é que linguagens populares atraiam inúmeros desenvolvedores, aumentando a probabilidade de contribuições externas, elas geralmente são linguagens fáceis de se aprender e de usar, elas possuem grande demanda no mercado o que pode explicar o grande número de releases e de atualizações.

RQ Extra. Sistemas populares são escritos nas linguagens mais populares(idiomas) ?

R: Sim, pelo número alto de falantes nativos e por serem línguas globalmente utilizadas, podemos deduzir que grande parte dos repositórios serão marcados com linguagens populares, mesmo por pessoas que não tem aquele idioma como primário.

2. Metodologia

A metodologia adotada para a análise dos resultados baseou-se no desenvolvimento de um script em Python, utilizando a biblioteca requests, que realizou uma consulta GraphQL para extrair dados relevantes dos repositórios mais populares do GitHub. Os dados obtidos foram exportados para arquivos CSV, no qual os repositórios foram ordenados de forma decrescente pelo número de estrelas, servindo como único critério de filtragem para a seleção dos 1.000 repositórios analisados. Em seguida, as médias das métricas de interesse (idade do repositório, total de pull requests aceitas, total de releases e tempo até a última atualização) foram calculadas individualmente para cada repositório. A visualização dos resultados foi realizada por meio da construção de gráficos boxplot e histogramas no Google Planilhas, permitindo a identificação da distribuição, dispersão e possíveis outliers nos dados.

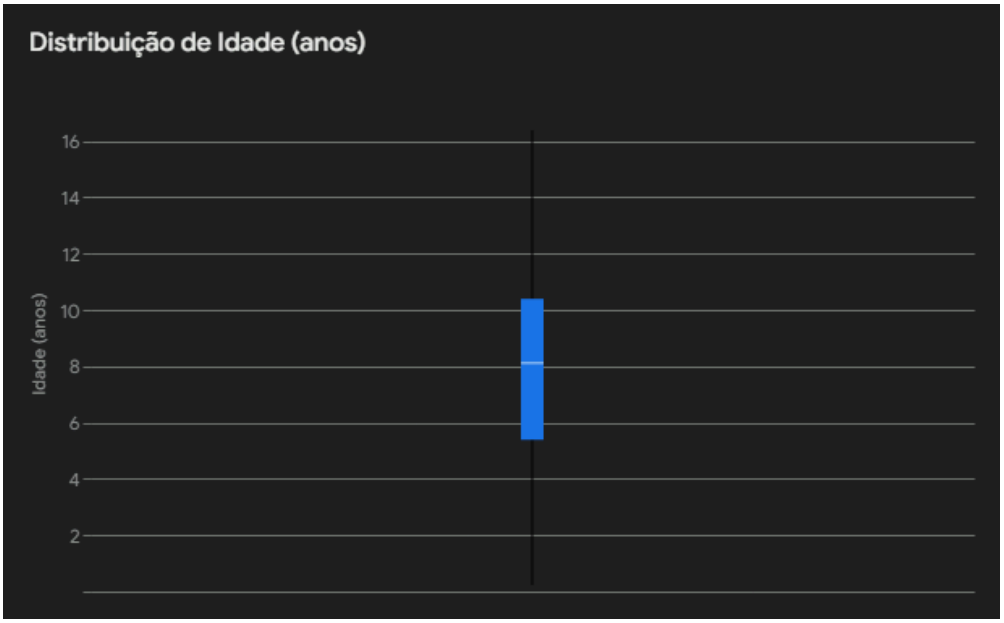
3. Resultados obtidos

RQ 01. Sistemas populares são maduros/antigos?

R: Mediana da idade dos repositórios: 8.11 anos

R: Média da idade dos repositórios: 7.87 anos

R: Moda da idade dos repositórios: 8.11 anos

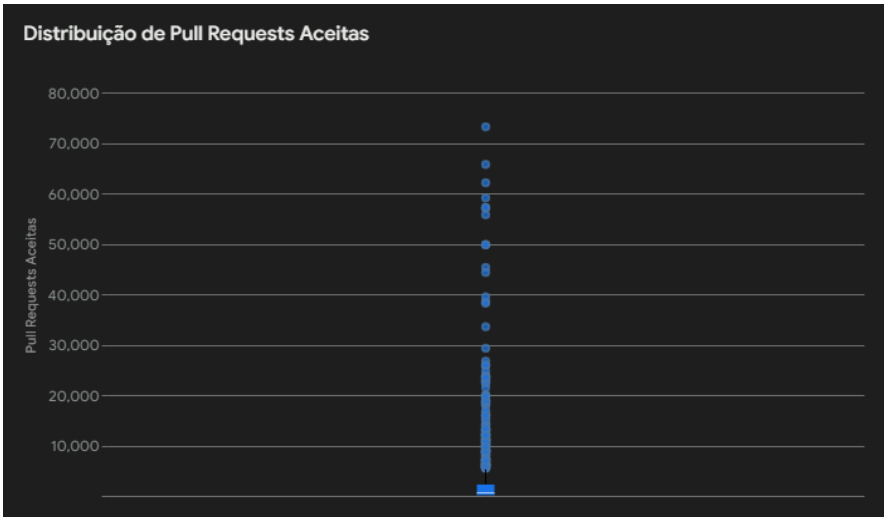


Q 02. Sistemas populares recebem muita contribuição externa?

R: Mediana de Pull Requests Aceitas: 580.00

R: Média da Pull Requests Aceitas: 2998.46

R: Moda da Pull Requests Aceitas: 1.00

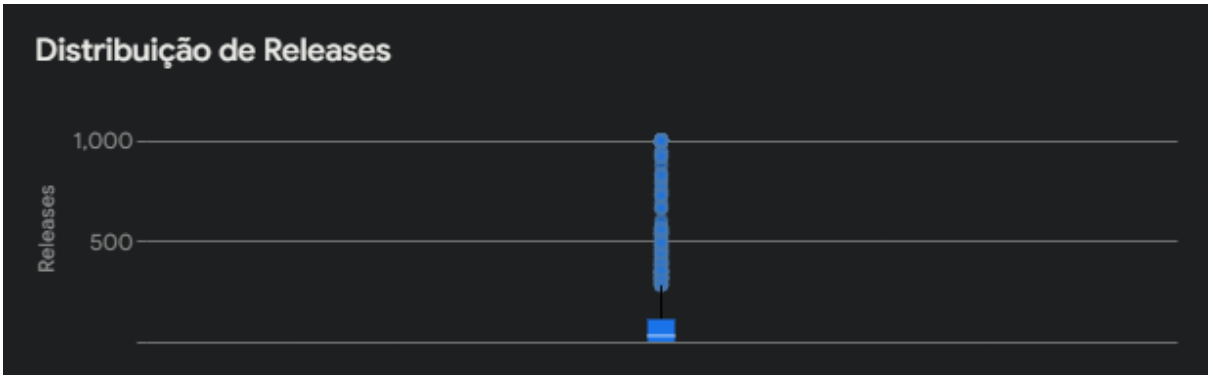


RQ 03. Sistemas populares lançam releases com frequência?

R: Mediana de Releases por Repositório: 30.50

R: Média de Releases por Repositório: 95.88

R: Moda de Releases por Repositório: 0

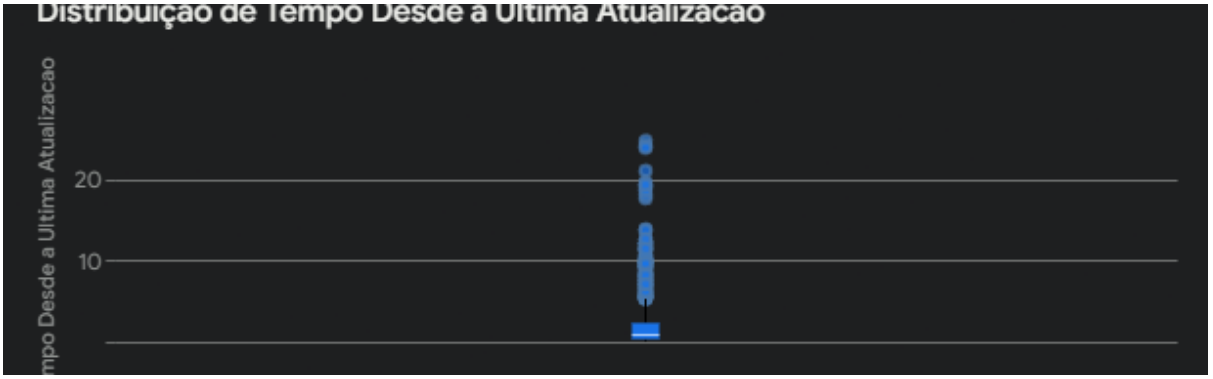


RQ 04. Sistemas populares são atualizados com frequência?

R: Mediana do Tempo desde a Última Atualização: 1.62 horas

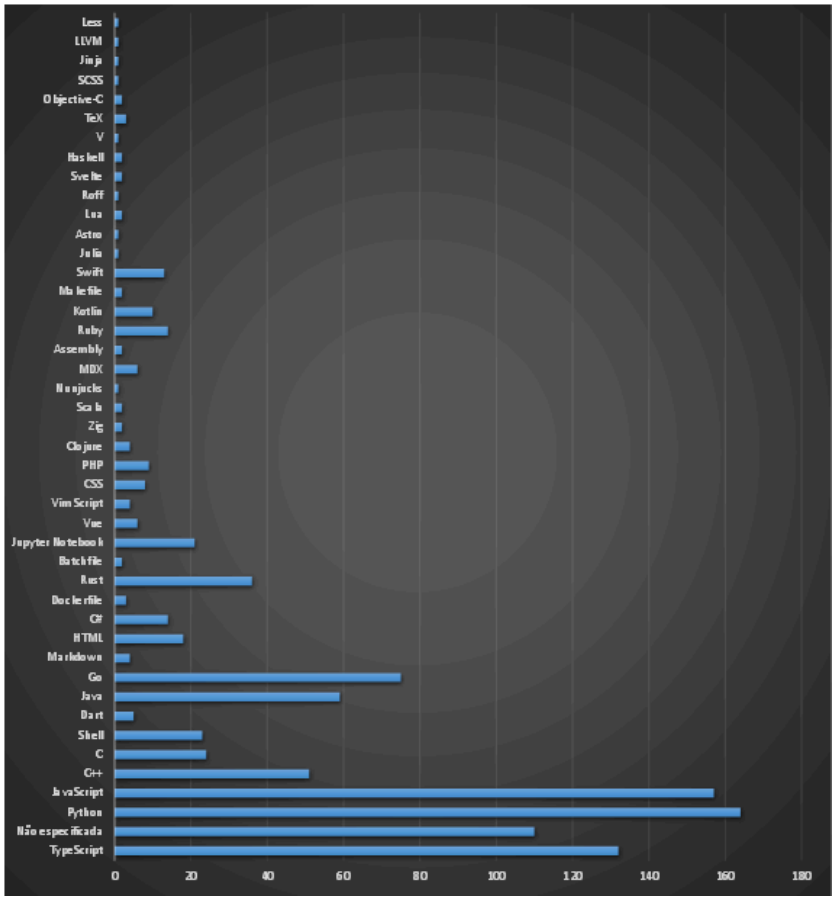
R: Média do Tempo desde a Última Atualização: 3.25 horas

R: Moda do Tempo desde a Última Atualização: 0.13 horas



RQ 05. Sistemas populares são escritos nas linguagens mais populares? (linguagem de programação)

R: Conforme a análise dos dados, 51% dos que contêm dados estão distribuídos nas linguagens python, javascript e typescript.

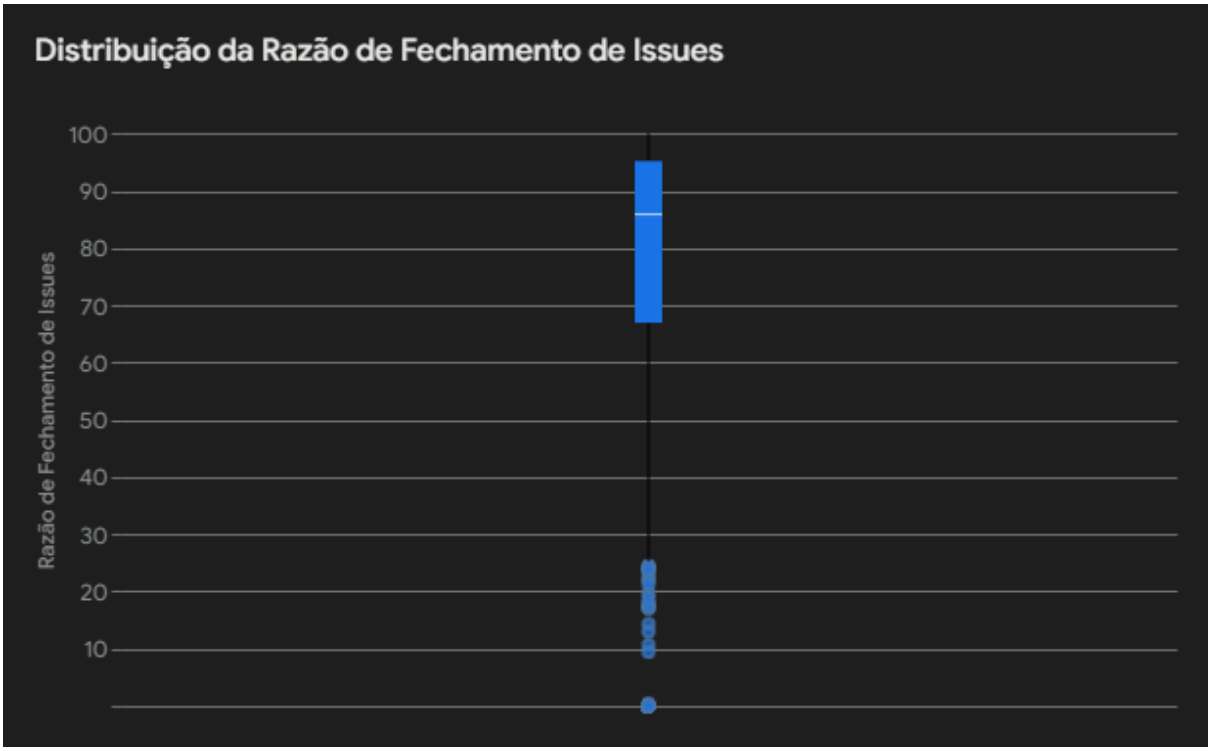


RQ 06. Sistemas populares possuem um alto percentual de issues fechadas?

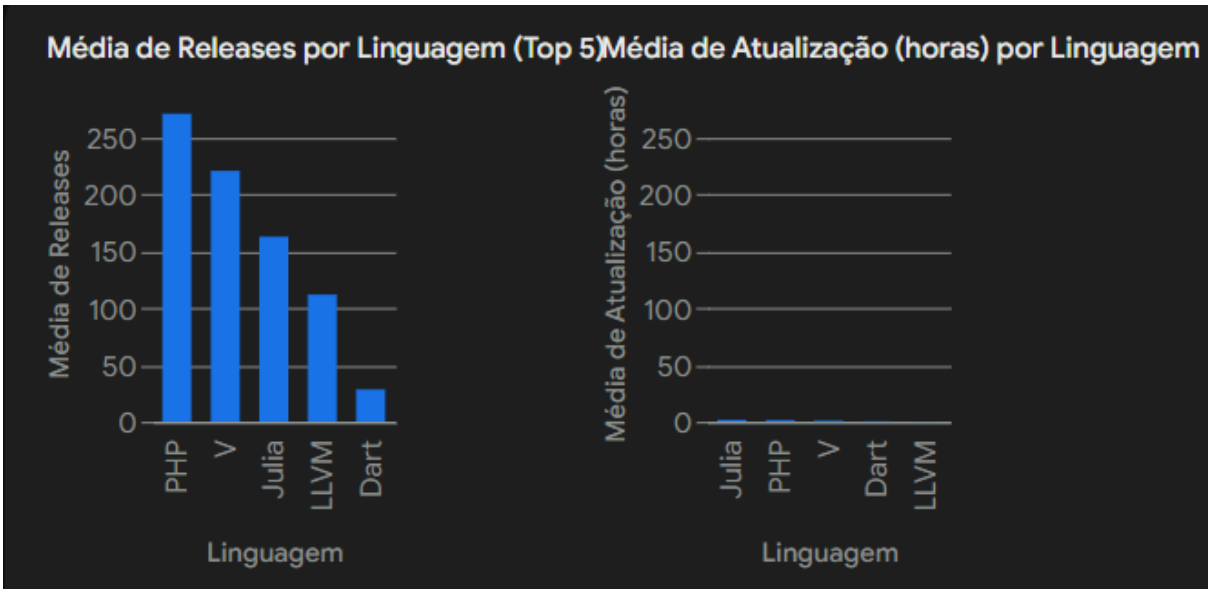
R: Mediana da Razão de Fechamento de Issues: 85.91%

R: Média da Razão de Fechamento de Issues: 76.51%

R: Moda da Razão de Fechamento de Issues: 0%

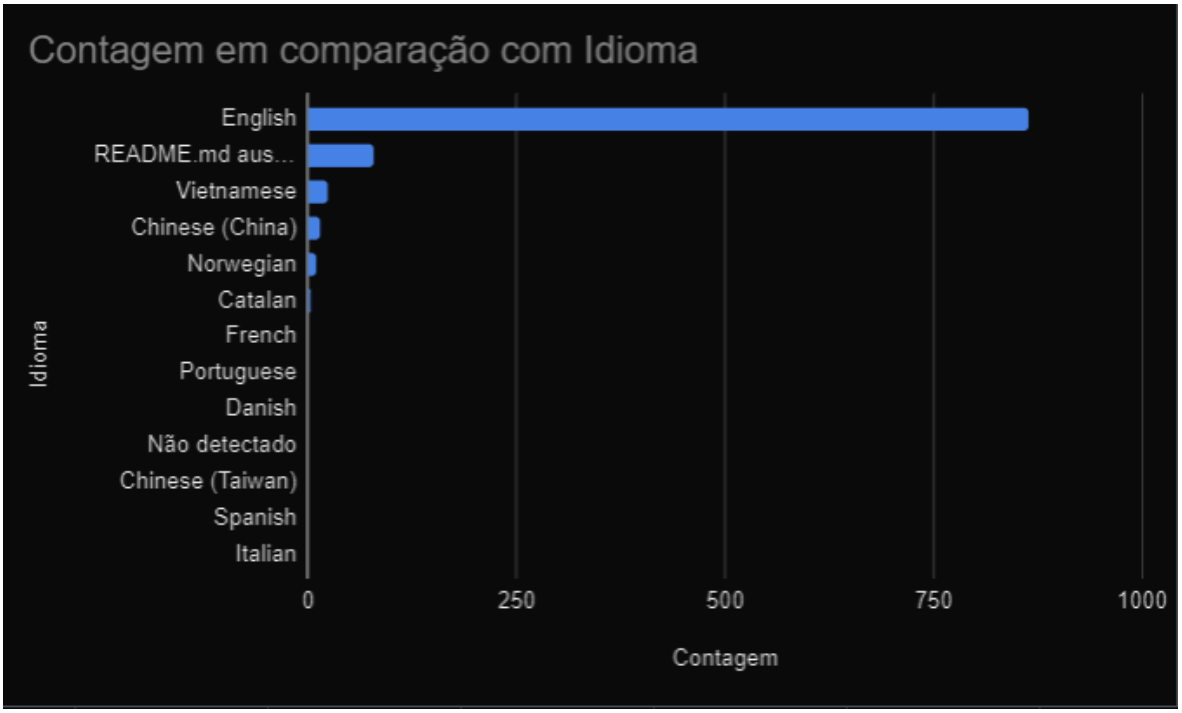


RQ 07. Sistemas escritos em linguagens mais populares recebem mais contribuição externa, lançam mais releases e são atualizados com mais frequência?



RQ Extra. Sistemas populares são escritos nas linguagens mais populares? (Idiomas)

R: Conforme a análise dos dados, 87% dos repositórios são feitos em inglês



4. Análise dos Resultados

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

RQ 01.

Análise: Os dados coletados confirmam que a hipótese inicial estava correta, mesmo os sistemas com menor número de estrelas entre os 1000 analisados possuem pelo menos 5 anos e meio de idade. O trabalho contínuo nesses repositórios ao longo dos anos contribui significativamente para o aumento de sua popularidade.

RQ 02.

Análise: Para confirmarmos se a hipótese está certa, temos primeiro que ressaltar que a média de pull requests aceitos pelos repositórios sugere um alto número de aceitações. No entanto, a maioria dos repositórios populares apresenta uma quantidade baixa de pull requests aceitos em relação ao total de pull requests realizados. Essa discrepância na média ocorre devido à existência de alguns

repositórios com mais de 20 mil pull requests aceitos, elevando assim a média geral além do esperado. Quando observamos valores como Moda e Mediana chegamos a um valor mais realista de pull requests aceitos e podemos então confirmar que a hipótese estaria correta e que, na verdade, quando comparado com o total de pull requests realizados, os aceitos são minoria.

RQ 03.

Análise: A média alta de releases diárias de 95 sugere que as comunidades mais populares conseguem se manter, pois tem um grande volume de releases, porém é válido ressaltar que este valor também pode ter sido alterado por um dúzia de repositórios que possuem uma média muito alta de releases e que quando levamos em consideração a moda e a mediana temos um valor mais realista que serve de base para confirmar melhor a hipótese inicial.

RQ 4.

Análise: Com base nos valores encontrados pela média, mediana, moda e na visualização do boxplot podemos confirmar que a hipótese está correta, as comunidades mais populares estão constantemente atualizando o repositório, tanto na moda quanto na mediana temos valores baixos que ficam em torno de 3 horas, isso mostra que a atualização diária consegue manter o projeto constantemente em evidência.

RQ 5.

Análise: Foram observados a partir do histograma que, mesmo tendo uma grande variedade de linguagens existe uma dominância nas linguagens populares como Python e JavaScript, sendo mais de 50% dos repositórios estando condensados em apenas 3 linguagens, isso nos leva a concluir que, sim as linguagens mais populares possuem comunidades maiores o que reflete nos extremos polos percebidos por essa análise.

RQ 6.

Análise: Podemos observar que os 1000 repositórios mais populares tem uma porcentagem alta quando se trata de fechamento de issues com uma mediana que chega a mais de 85%, isso demonstra o comprometimento em resolver os problemas e acaba também confirmando a hipótese, o que com certeza acaba sendo refletido diretamente na popularidade e qualidade do software criado.

RQ 7.

Analise: Após analisar a tabela podemos tirar algumas conclusões
LLVM e Julia, apesar de terem um número menor de releases em média, apresentam as maiores médias de Pull Requests, o que pode indicar um desenvolvimento mais focado em grandes atualizações em vez de releases frequentes.

V e PHP se destacam com as maiores médias de releases, o que sugere novas funcionalidades e melhorias sendo disponibilizadas com maior frequência.

Dart, apesar de ter a menor média de Pull Requests entre as top 5, ainda apresenta uma média de releases e atualizações consideráveis, o que indica uma comunidade ativa e um desenvolvimento constante.

Os dados e o gráfico apoiam a hipótese de que linguagens mais populares tendem a ter maior atividade de releases e atualizações. No entanto, a frequência de releases e atualizações também pode ser influenciada por outros fatores, como a natureza do projeto, a estratégia de desenvolvimento e a maturidade da linguagem.

RQ Extra.

Análise: Podemos observar a partir do histograma que temos uma discrepância considerável entre os repositórios que estão registrados como em inglês e os que estão em outras línguas, isso nos leva a conclusão de que, por ser uma língua popular, repositórios em inglês conseguem mais facilmente ter estrelas