PUC Minas

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Instituto de Ciências Exatas e Informática — Unidade Praça da Liberdade

Curso: Engenharia de Software.

Prof: José Laerte Pires Xavier Junior

Disciplina: Laboratório de Experimentação de Software

Alunos: Bruno Henrique Armanelli e Lucas Alves Gusmão

Características de Repositórios Populares do GitHub

1. Introdução

O GitHub é a maior coleção de projetos *open-source* na internet. A plataforma é utilizada tanto por desenvolvedores que desejam publicar seus projetos pessoais de pequeno porte, quanto por empresas e organizações que disponibilizam projetos e ferramentas para a comunidade. Por conta disso, ele torna-se uma fonte interessante para análises e pesquisas relacionadas à área da engenharia de software, sendo possível tirar conclusões com base nos dados armazenados na plataforma.

Com base em análises de dados por meio da API pública do GitHub, foi feito um trabalho na disciplina de Laboratório de Medição e Experimentação de Software com o objetivo de definir e verificar hipóteses sobre repositórios populares na plataforma. A popularidade será definida com base no total de "estrelas" que cada repositório recebeu de usuários na plataforma.

2. Metodologia

Este trabalho é uma pesquisa descritiva com base em uma abordagem quantitativa sob a base de dados pública do GitHub. Os dados foram coletados através de consultas na API em GraphQL disponibilizada pela plataforma. Na consulta foram limitados os primeiros 1000 últimos repositórios que atendem à regra de ordenação por estrelas. Um pequeno programa foi escrito na linguagem JavaScript, utilizando a biblioteca Apollo como cliente GraphQL, para executar a consulta e percorrer todas as páginas na API, depois os dados foram extraídos

para um arquivo no formato CSV e importados para a ferramenta DataStudio do Google, para que pudessem ser analisados.

As RQs, assim como suas métricas e hipóteses são as seguintes:

- **RQ 01. Sistemas populares são maduros/antigos?**

  Métrica: idade do repositório (calculado a partir da data de sua criação)

  Hipótese: um repositório será considerado maduro quando ele possuir mais de cinco anos de existência. Sistemas mais maduros possuem avaliações mais altas, pois tiveram mais tempo para conquistarem *stargazers* e gerar mais engajamento da comunidade. O resultado esperado é que a maioria dos repositórios sejam maduros.

- **RQ 02. Sistemas populares recebem muita contribuição externa?**

  Métrica: total de pull requests aceitas

  Hipótese: repositórios populares recebem um volume alto de contribuições externas, por possuírem alta visibilidade pela comunidade. Para medir a quantidade de contribuições externas, foram contabilizados os *pull requests* com *status "merged"*, isto é, que foram aceitos pelos mantenedores do repositório e agregaram valor à ele. O resultado esperado é que os repositórios tenham mais de 800 pull requests *"merged"*.

- **RQ 03. Sistemas populares lançam releases com frequência?**

  Métrica: total de releases

  Hipóteses: repositórios populares tendem a ter um número mais elevado de releases, por conta da maturidade e quantidade de contribuições externas também serem mais elevadas. O esperado é que repositórios populares tenham ao menos 400 releases.

- **RQ 04. Sistemas populares são atualizados com frequência?**

  Métrica: tempo até a última atualização (calculado a partir da data de última atualização)

  Hipótese: os repositórios mais populares, com base nas hipóteses anteriores, são atualizados com mais frequência. O resultado esperado é que possuam a última contribuição há pelo menos uma semana (7 dias).

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

  Métrica: linguagem primária de cada um desses repositórios

  Hipótese: um motivo que ajuda a determinar a popularidade de um repositório é as tecnologias utilizadas por ele. Tecnologias mais populares atraem mais contribuidores, que podem contribuir com o intuito de aprendizado ou para ajudar no desenvolvimento de uma ferramenta que gostam, mas que ao mesmo tempo possuem familiaridade com as tecnologias usadas em seu desenvolvimento. Por conta disso, é esperado que os repositórios sejam no mínimo 60% dos repositórios sejam desenvolvidos utilizando as linguagens mais populares, com base no ranking do GitHub Octoverse 2020[1].

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

  Métrica: razão entre número de issues fechadas pelo total de issues

  Hipótese: *issues* são uma das principais formas de contribuir com um repositório de código aberto. Qualquer usuário que encontrar um problema ou possuir uma sugestão sobre o projeto pode abrir uma issue em um repositório. Assim que uma issue é aberta, outro usuário pode tentar solucioná-la, e após isso ser feito, ela é fechada. Por conta do volume de contribuições, estima-se que repositórios mais populares possuam um percentual alto de issues fechadas, acima de 75%.

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

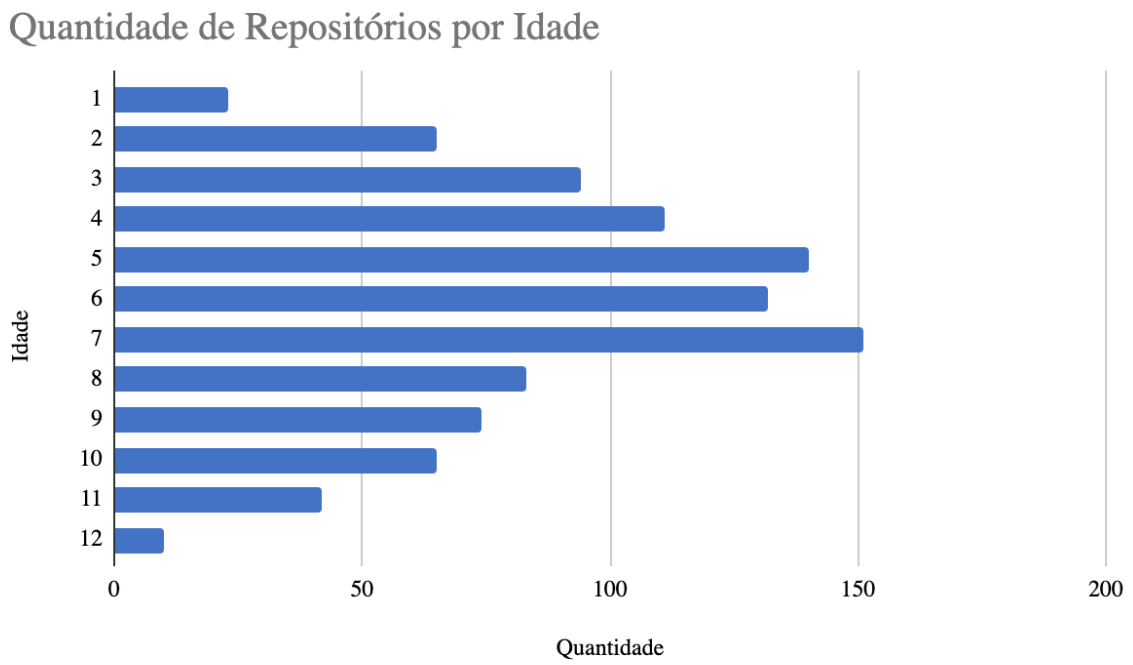
  Métrica: mediana das pull requests com status "merged", de dias percorridos desde a última atualização, da quantidade de releases feitos em linguagens populares e não populares, segundo a pesquisa do Octoverse[1].

  Hipótese: é esperado que os repositórios escritos em linguagens mais populares tenham uma mediana maior de pull-requests aceitos e de releases. Em relação aos dias desde a última atualização, é esperado que os repositórios de linguagens mais populares sejam atualizados mais frequentemente do que os outros. Isso se deve à hipótese de que repositórios escritos em linguagens populares possuem mais engajamento da comunidade, o que os garante mais contribuições e atividade em geral.

3. Resultados Obtidos

Com os dados obtidos na execução do script gerado de acordo com as métricas definidas anteriormente, foram elaboradas respostas com os resultados para cada *Research Question*. A execução e coleta dos dados ocorreu em 22 de agosto de 2021 às 21h12.

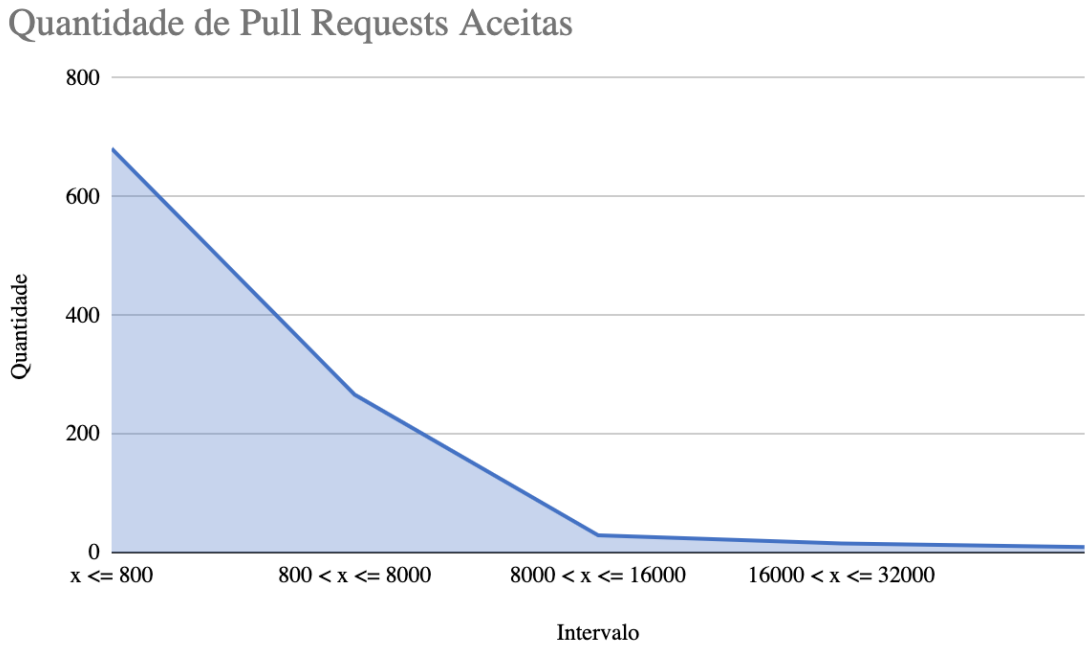
- **RQ 01:** Com base nos dados coletados, foi montado o seguinte gráfico que ilustra a quantidade de repositórios por idade:



A quantidade de repositórios com mais de 5 anos de idade é 697. A média de idade é 6,042 e a mediana é 6.

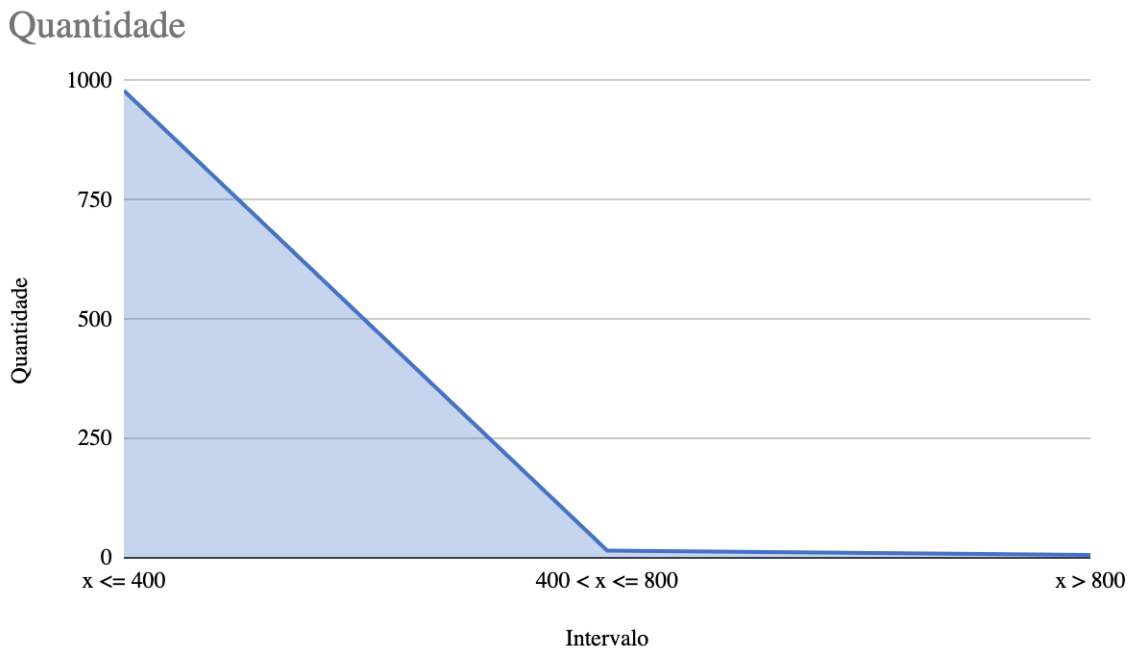
Quantidade > 5	697
Média	6.042
Mediana	6

- **RQ 02:** A maioria dos repositórios possuía menos de 800 pull requests aceitos, com somente 31,9% dos repositórios possuindo acima de 800 contribuições externas. A média de pull requests aceitos foi de 1890,89, enquanto a mediana foi 380,5.



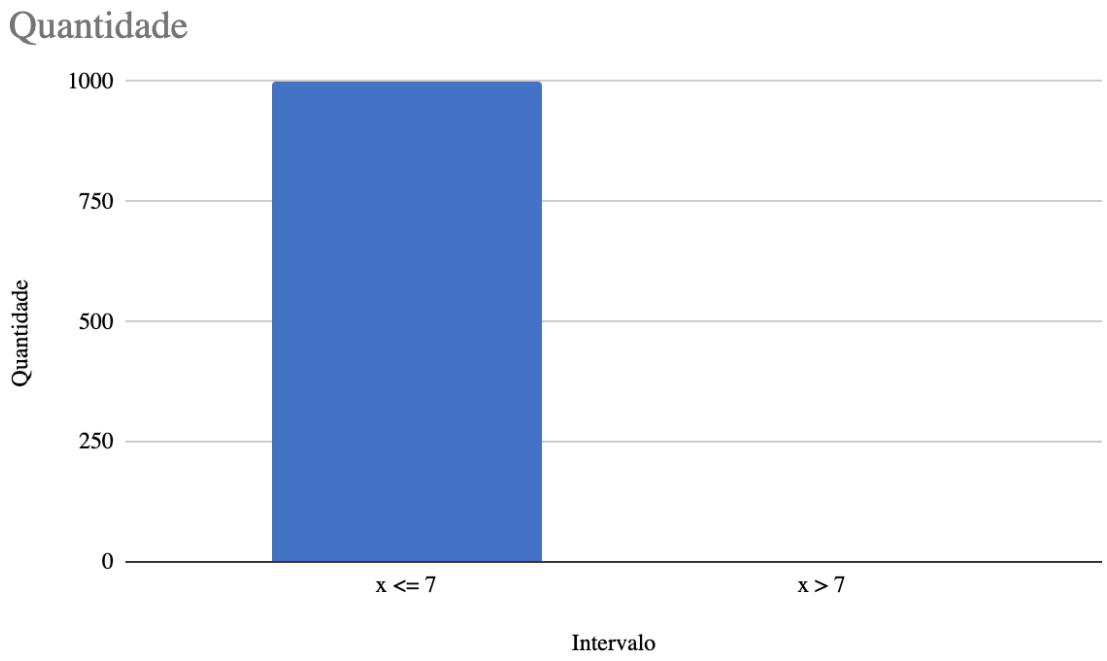
Quantidade > 800	319
Média	1890.898
Mediana	380.5

- **RQ 03:** Analisando os dados coletados, foi identificado que grande parte dos repositórios não utilizava a função de releases e, portanto, não possuíam um número elevado de releases. Apenas 21 dos repositórios analisados possuíam mais de 400 releases. A média de releases é 54,21 e a mediana 14.



Quantidade > 400	21
Média	54.21
Mediana	14

- **RQ 04:** Analisando os dados referentes às atualizações nos repositórios pesquisados, obtivemos tabela abaixo contendo o tempo dias corridos desde a última atualização no respectivo repositório. Foi possível concluir que 100% dos repositórios haviam sido atualizados há menos de 7 (sete) dias. A mediana é zero porque a maior parte dos repositórios foi atualizado menos de 24h do momento da coleta de dados.

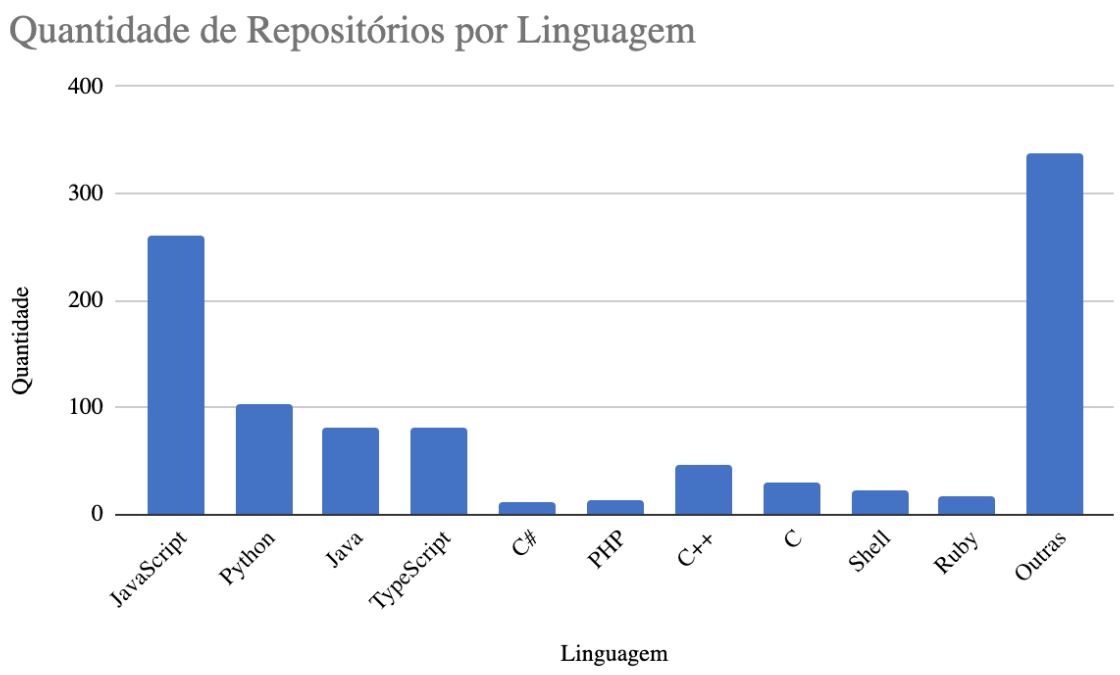


Quantidade < 7	1000
Média	0.05405405405
Mediana	0

- **RQ 05:** Com base nos dados referentes à linguagem mais utilizada em cada repositório, foi possível analisar se os repositórios mais populares também utilizavam as linguagens mais populares, com base na pesquisa do Octoverse 2020[1], que indicou as linguagens mais populares do ano de 2020 segundo dados do GitHub.

Segundo os resultados obtidos, 66,30% dos repositórios eram de projetos escritos nas linguagens mais populares. As 3 linguagens mais populares foram: JavaScript, Python e Java, respectivamente.

Linguagem	Quantidade
JavaScript	260
Python	102
Java	81
TypeScript	80
C#	11
PHP	13
C++	46
C	30
Shell	23
Ruby	17
Outras	337
Total	**1000**



- **RQ 06:** Analisando a porcentagem de issues fechadas em relação ao total de issues de cada repositório e após o cálculo da média destes valores obtivemos o valor de 77,28% de issues fechadas, enquanto o cálculo da mediana para os mesmos valores resultou em 85,71% de issues fechadas.

Média	77.28706
Mediana	85.715

- **RQ 07:** Separando os dados de repositórios escritos em linguagens populares de repositórios escritos em linguagens não-populares, foi possível perceber que os repositórios de linguagens mais populares possuíam uma mediana mais elevada de contribuições externas aceitas, assim como de releases. Ambos possuíram uma mediana de 0 dias desde a última atualização.

-	Linguagens Populares	Linguagens Não-Populares
Pull Requests Merged	485.5	213
Última Atualização	0	0
Quant. Releases	22	1

4. Conclusão

Considerando os resultados obtidos e as hipóteses levantadas, pode-se concluir, para cada pergunta, respectivamente:

- **RQ 01:** Foi definido na hipótese levantada anteriormente que o resultado esperado era que a maioria dos repositórios existissem há no mínimo 5 anos. Os resultados corroboraram a hipótese: apenas 293 dos 1000 repositórios buscados possuíam menos de cinco anos.
- **RQ 02:** Em comparação com a hipótese levantada, os valores coletados no resultado foram menores do que o esperado. Nos 1000 repositórios analisados, 319 apresentaram mais de 800 pull requests com *status "merged"*, o que representa 31,9% dos repositórios coletados.
- **RQ 03:** Repositórios populares possuem menos *releases* do que o valor esperado na hipótese. O valor da mediana de releases é 14 e a média 54,22. Foi observado que muitos repositórios não utilizam a ferramenta de *release* da plataforma, pois muitos possuíam 0 releases.
- **RQ 04:** A hipótese levantada em relação a esta RQ está correta. Todos os repositórios apresentaram a última atualização há menos de 7 dias. Foi observado que a maioria dos repositórios haviam sido atualizados no mesmo dia que a coleta de dados foi realizada.

- **RQ 05:** A hipótese levantada foi validada ao verificar durante os resultados do experimento que 66,3% dos repositórios estão escritos nas linguagens mais populares, segundo o GitHub Octoverse[1].
- **RQ 06:** Os repositórios possuíam uma taxa alta de issues fechadas, que corresponde ao valor estabelecido na hipótese. A mediana é 85,71%, acima do 75% esperado, mas a média foi de 77,28%, que é um valor bem próximo do esperado na hipótese. Com base nisso, é possível afirmar que os repositórios populares possuem uma alta taxa de contribuições da comunidade que resolvem os problemas detalhados nas *issues*.
- **RQ 07:** Com base na análise realizada sob os dados, foi possível corroborar a hipótese estabelecida de que repositórios escritos em linguagens populares recebem mais contribuições externas e possuem mais atividade/engajamento da comunidade do que os repositórios de linguagens não-populares. Os repositórios de linguagens mais populares tiveram uma mediana de 485,5 pull requests aceitas, contra 213 de linguagens não-populares.

5. Referência Bibliográfica:

[1] GitHub Octoverse 2020, https://octoverse.github.com/. Acesso em: 29 de agosto de 2021.