PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Bacharelado em Engenharia de Software

CARACTERÍSTICAS DE REPOSITÓRIOS POPULARES

João Victor Tadeu Chaves Frois
Lucas Gabriel Padrão Rezende

Belo Horizonte, 2022

1. INTRODUÇÃO

O GitHub é uma plataforma que possibilita a hospedagem de códigos fontes e arquivos durante o desenvolvimento de um *software*. Desta forma, ele permite que programadores contribuam em projetos privados ou públicos de qualquer lugar do mundo possibilitando ao usuário ter acesso a diferentes versões do código de acordo com as etapas do desenvolvimento.

Além de ser uma plataforma de hospedagem, o git também é uma espécie de rede social para desenvolvedores. Reunindo mais de 65 milhões de usuários, 3 milhões de organizações e 200 milhões de repositórios, o GitHub permite que quaisquer usuários cadastrados na plataforma possam disponibilizar seus trabalhos à comunidade. Sendo assim o GitHub possibilita a seus usuários saberem quais são os repositórios *open-source* mais populares.

Outrossim, a plataforma disponibiliza através de uma API própria a extração de dados e métricas dos repositórios públicos. Sendo assim é possível estudar as principais características dos repositórios mais populares. Portanto, durante a execução deste trabalho serão analisados os 1000 repositórios *open-source* mais populares do GitHub em aspectos de desenvolvimento, frequência de contribuições externas, frequência de lançamentos de releases, dentre outras características que respondam às questões levantadas.

1.1. QUESTÕES E MÉTRICAS

No contexto apresentado algumas questões são levantadas em relação aos repositórios mais populares do GitHub, questões essas que visam entender características que tornam estes repositórios populares. Sendo assim as questões levantadas e as métricas que serviram de base para respondê-las são:

- **RQ 01.** Sistemas populares são maduros/antigos?
  Métrica: idade do repositório (calculado a partir da data de sua criação).
- **RQ 02.** Sistemas populares recebem muita contribuição externa?
  Métrica: total de *pull requests* aceitas.
- **RQ 03.** Sistemas populares lançam *releases* com frequência?
  Métrica: *total de releases*.
- **RQ 04.** Sistemas populares são atualizados com frequência?

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

- **RQ 05.** Sistemas populares são escritos nas linguagens mais populares?
  Métrica: linguagem primária de cada um desses repositórios.
- **RQ 06.** Sistemas populares possuem um alto percentual de *issues* fechadas?
  Métrica: razão entre número de *issues* fechadas pelo total de *issues*.

1.2. HIPÓTESES INFORMAIS

Para cada uma das questões levantadas hipóteses informais foram criadas antes da coleta de dados, buscando se estabelecer um ponto de comparação para os resultados obtidos posteriormente. Desta forma, será possível comparar as expectativas com os resultados finais.

As hipóteses levantadas para cada questão, juntamente com a ideia por trás de suas construções são:

- **RQ 01.** Os repositórios mais populares são sim mais antigos. Uma vez que estão a um período de tempo maior disponíveis a comunidade estes repositórios tendem a acumular mais visitas e avaliações do que repositórios que estão a pouco tempo disponíveis.
- **RQ 02.** Os repositórios mais populares recebem sim mais contribuições externas. Um alto número de contribuições significa um maior engajamento da comunidade, aumentando assim as chances de esse repositório ser popular uma vez que tem um alto número de indivíduos envolvidos em seu desenvolvimento.
- **RQ 03.** Os repositórios mais populares lançam sim *releases* com frequência. Já que há *releases* lançadas com frequência a conclusão estabelecida foi de que se está tendo um uso frequente da ferramenta desenvolvida dentro do repositório, portanto há uma comunidade que a utiliza e gera novas demandas, mostrando assim um bom engajamento com o que é produzido.
- **RQ 04.** Os repositórios mais populares são sim atualizados com maior frequência. Uma vez que se há atualizações frequentes, pode-se concluir que se está tendo um uso frequente da ferramenta desenvolvida dentro do repositório, portanto há uma comunidade que a utiliza e gera novas demandas, mostrando assim um bom engajamento com o que é produzido, trazendo assim mais popularidade ao repositório.

- **RQ 05.** Os repositórios mais populares são sim escritos nas linguagens mais populares. Já que exemplificam o uso de algo que já é popular, estes repositórios trazem consigo uma comunidade que já é engajada e populosa, que pode consumir seus recursos como exemplos, trazendo maior engajamento e consequentemente popularidade ao repositório.
- **RQ 06.** Os repositórios mais populares possuem sim um alto percentual de *issues* fechadas. Uma vez que possuem maior engajamento da comunidade e de seus desenvolvedores, os repositórios mais populares tendem a possuir mais colaborações gerando um número maior de *issues* fechadas em relação ao seu total.

2. METODOLOGIA

Para se responder os questionamentos levantados se fez necessário coletar as métricas indicadas dos 1000 repositórios mais populares do GitHub. Essa coleta de dados foi feita de forma automatizada através de um *script* construído em *TypeScript*.

O *script* desenvolvido consome uma API própria do GitHub, que possibilita a extração de dados de repositórios abertos ao público, mediante a passagem de um *token* de autenticação que certifica que quem está consumindo a API seja um usuário já cadastrado na plataforma. Além de ser construído em *TypeScript*, o *script* criado foi escrito de forma que consumisse a API do GitHub através de um *query* GraphQL sem a utilização de bibliotecas de terceiros.

A construção desse script foi feita sendo dividida em duas partes:

1. Extração dos dados necessários de 100 repositórios e a automatização da requisição destes dados. Nessa etapa os 100 repositórios consultados já deveriam ser os 100 mais populares.
2. Escalonamento do *script* já criado para 1000 repositórios e exportação dos dados coletados para um arquivo *.csv*. Nessa etapa o arquivo exportado deveria conter todos os dados necessários identificados e separados com base em seu repositório de origem.

Após a elaboração e construção do processo de coleta de dados, uma análise dos mesmos foi feita utilizando o Google Sheets, ferramenta principal utilizada para leitura, manipulação e análise dos dados.

Exportando os dados coletados para uma planilha foi possível se manipular os dados de: data de criação do repositório a fim de se calcular a idade (em anos) de cada um dos repositórios; data da última atualização a fim de se calcular o tempo (em horas) desde o último update do repositório e o número de *issues* a fim de se calcular a razão entre o número de *issues* fechadas e seu total para cada repositório.

Após as manipulações necessárias foram geradas tabelas e cálculos para se relacionar as mil métricas coletadas e gerar visualizações que possibilitam uma análise rápida e simplificada, mas suficientes para se responder cada uma das perguntas. As visualizações criadas foram tabelas e gráficos (do tipo de barras e de Candle) para cada uma das métricas, modelo adotado devido às diferentes escalas numéricas apresentadas por cada um dos dados coletados.

Os códigos produzidos, bem como os resultados extraídos e suas análises estão disponíveis no seguinte repositório: https://github.com/padraorezende/Lab.

3. RESULTADOS

Os resultados obtidos para cada uma das perguntas, juntamente das métricas coletadas e suas análises serão evidenciadas nas seções a seguir. Todas as imagens relacionadas a tabelas e gráficos foram geradas a partir da planilha disponibilizada no repositório supracitado.

3.1. RQ 01 - IDADE DOS REPOSITÓRIOS

Para análise desta métrica há um importante fator externo a ser considerado, o ano de criação do GitHub corresponde ao ano de 2008, desta forma nenhum o teto máximo para a idade dos repositórios avaliados é de 14 anos.

Sendo assim, os repositórios foram agrupados com base em suas idades, sendo maior idade possível 14 anos e a menor idade 0 anos (correspondente aos repositórios que ainda não completaram 1 ano de existência). O agrupamento se deu conforme a tabela a seguir:

IDADE (ANOS)	QUANTIDADE DE REPOSITÓRIOS
0	4
1	21
2	39
3	82
4	94
5	110
6	134
7	123
8	145
9	74
10	69
11	60
12	34
13	9
14	2

FIGURA 01 - Tabela montada dinamicamente pelo Google Sheets que correlaciona os valores de idade e a ocorrência de cada um dos valores.

Com base no agrupamento mostrado foi construído um gráfico de barras para melhor visualização dos dados levantados.

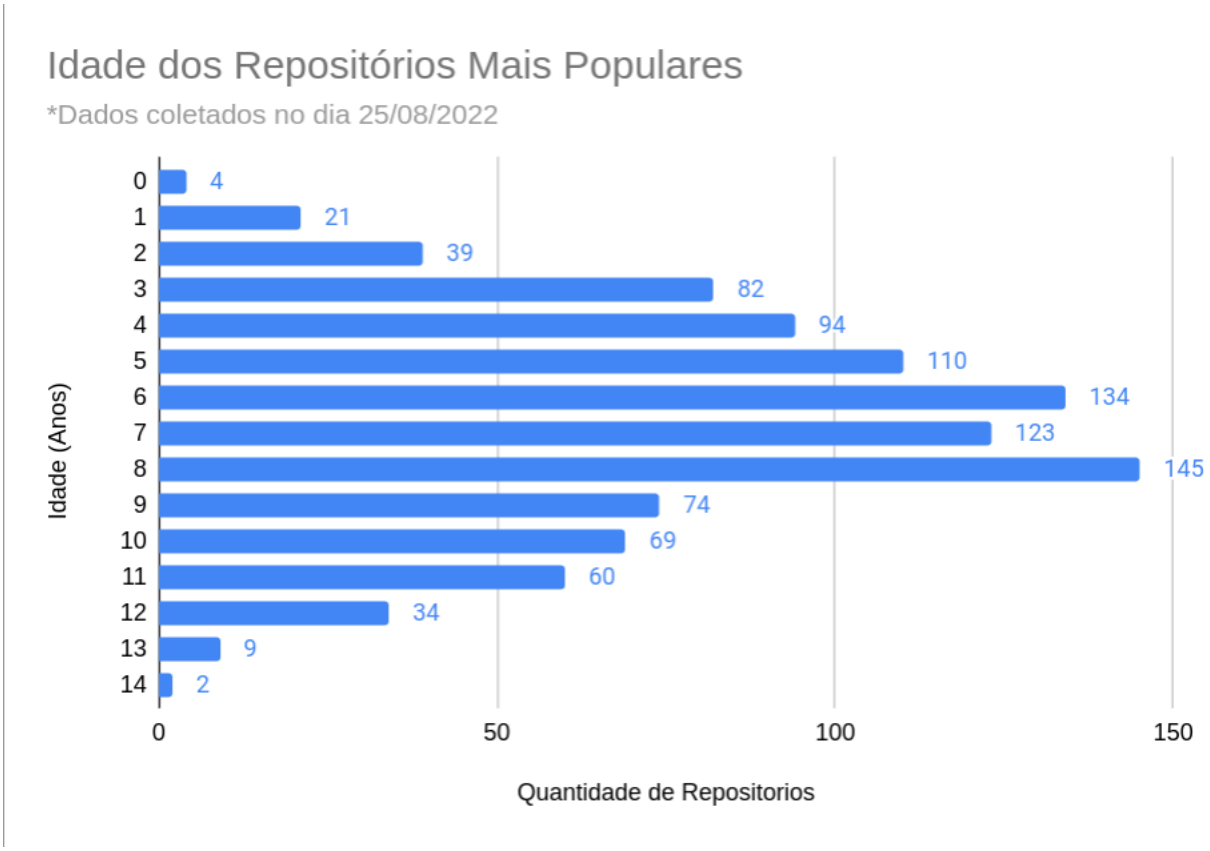

FIGURA 02 - Gráfico barras das Idades dos Repositórios montado pelo Google Sheets.

Além disso, também foram tabelados os valores da menor idade (função MIN), da idade referente ao primeiro quartil (função QUARTILE com paramento 1) , da idade referente ao terceiro quartil (função QUARTILE com paramento 3) e da maior idade (função MAX) conforme a tabela abaixo.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Idade (Anos)	0	5	8	14

FIGURA 03 - Tabela para montagem do gráfico de Candle.

Esses dados foram tabulados com intuito de se montar um gráfico de Candle montando de forma objetiva os extremos da métrica e em qual intervalo está localizada a maior incidência de repositórios.

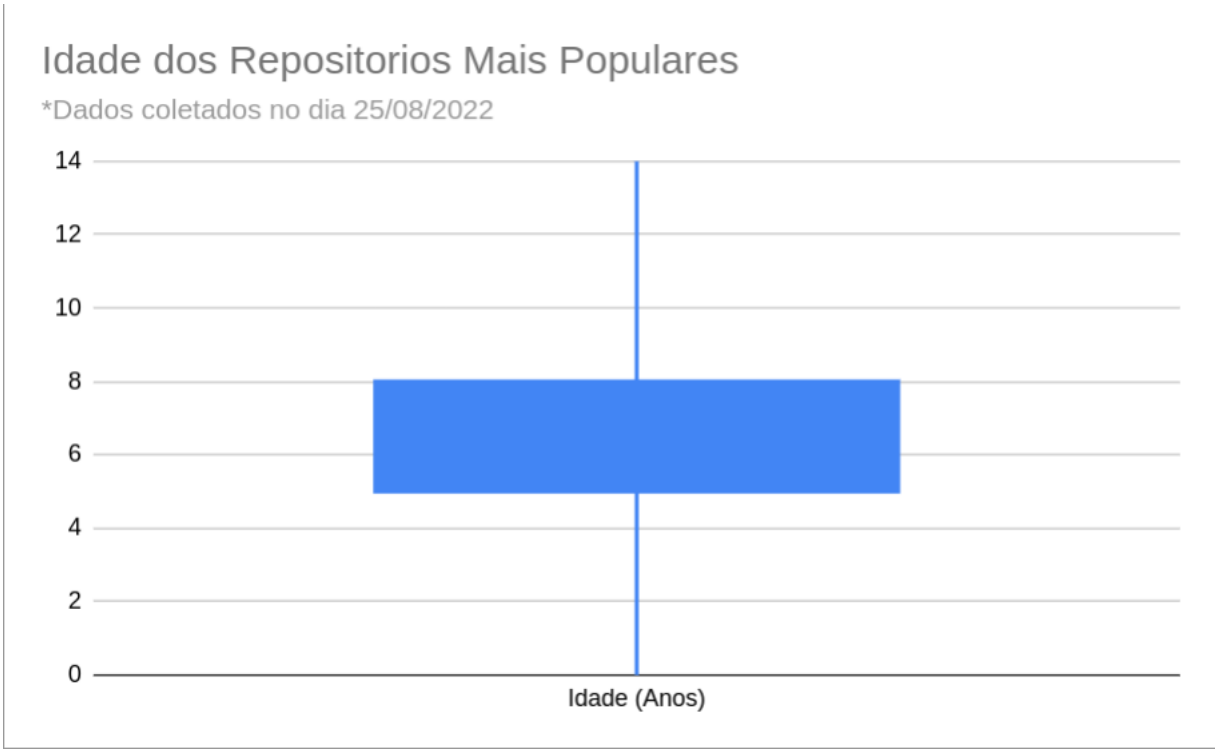

FIGURA 04 - Gráfico de Candle da Idade dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação a idade, sendo 512 repositórios que possuem idades de 5 à 8 anos.

Dados complementares gerados são a média, a moda e a mediana dos valores de idade dos repositórios, sendo eles respectivamente:

Media	Moda	Mediana
6,665	8	7

FIGURA 05 - Média, Moda e Mediana das idades dos repositórios.

3.2. RQ 02 - TOTAL DE PULL REQUESTS DOS REPOSITÓRIOS

Para análise do total de *Pull Requests* dos repositórios foram tabelados os valores da menor quantidade *Pull Requests* de um repositório (função MIN), primeiro quartil (função QUARTILE com paramento 1) da quantidade *Pull Requests* dos repositórios, terceiro quartil (função QUARTILE com paramento 3) da quantidade *Pull Requests* dos repositórios e da maior quantidade *Pull Requests* de um repositório (função MAX) conforme a tabela abaixo.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Total de Pull Requests	0	122	1485	101509

FIGURA 06 - Tabela para montagem do gráfico de Candle.

Esses dados foram tabulados com intuito de se montar um gráfico de Candle montando de forma objetiva os extremos da métrica e em qual intervalo está localizada a maior incidência de repositórios.

FIGURA 07 - Gráfico de Candle da Pull Requests dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação a quantidade de *Pull Requests*, sendo 503 repositórios que possuem a quantidade de *Pull Requests* entre 122 e 1485.

Dados complementares gerados são a média, a moda e a mediana dos valores da quantidade de *Pull Requests* dos repositórios, sendo eles respectivamente:

Media	Moda	Mediana
2211,684	0	435

FIGURA 08 - Média, Moda e Mediana das quantidades de Pull Requests dos repositórios.

3.3. RQ 03 - TOTAL DE RELEASES DOS REPOSITÓRIOS

Para análise do total de *Releases* dos repositórios foram tabelados os valores da menor quantidade *Releases* de um repositório (função MIN), primeiro quartil (função QUARTILE com paramento 1) da quantidade *Releases* dos repositórios, terceiro quartil (função QUARTILE com paramento 3) da quantidade *Releases* dos repositórios e da maior quantidade *Releases* de um repositório (função MAX) conforme a tabela abaixo.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Total de Releases	0	0	80	2352

FIGURA 09 - Tabela para montagem do gráfico de Candle.

Esses dados foram tabulados com intuito de se montar um gráfico de Candle montando de forma objetiva os extremos da métrica e em qual intervalo está localizada a maior incidência de repositórios.

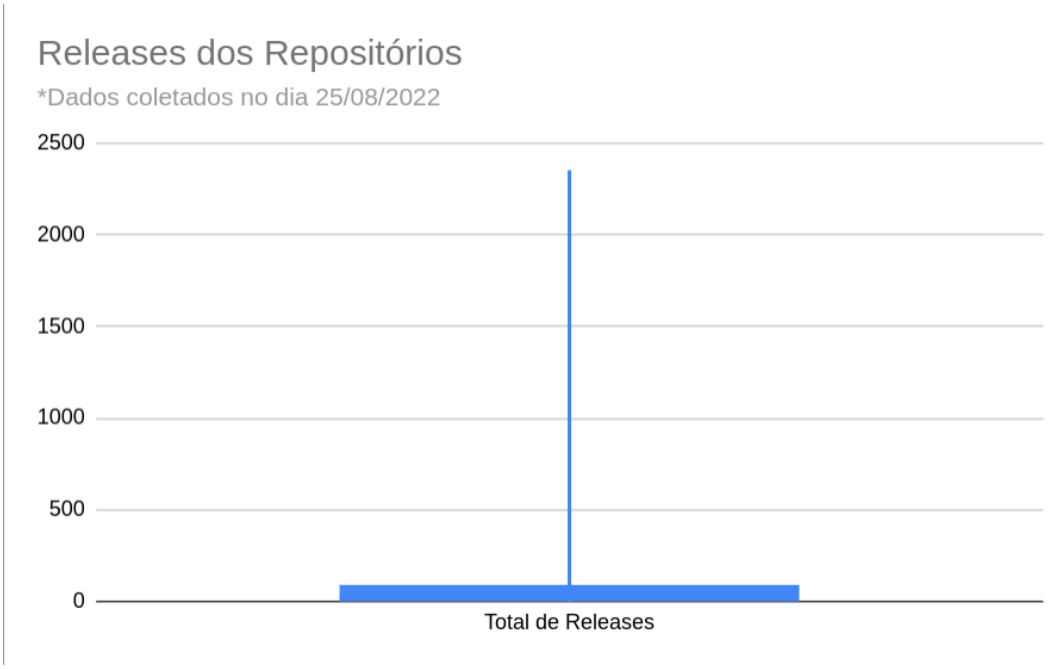

FIGURA 10 - Gráfico de Candle da Releases dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação a quantidade de *Releases*, sendo 752 repositórios que possuem a quantidade de *Releases* entre 0 e 80.

Dados complementares gerados são a média, a moda e a mediana dos valores da quantidade de *Releases* dos repositórios, sendo eles respectivamente:

Media	Moda	Mediana
68,356	0	19

FIGURA 11 - Média, Moda e Mediana das quantidades de Releases dos repositórios.

3.4. RQ 04 - TEMPO DESDE A ÚLTIMA ATUALIZAÇÃO DOS REPOSITÓRIOS

Para análise do tempo desde a última atualização dos repositórios foram tabelados os valores da menor quantidade de horas desde a última atualização de um repositório (função MIN), primeiro quartil (função QUARTILE com paramento 1) da quantidade de horas desde a última atualização dos repositórios, terceiro quartil (função QUARTILE com paramento 3) da quantidade de horas desde a última atualização dos repositórios e da maior quantidade de horas desde a última atualização de um repositório (função MAX) conforme a tabela abaixo.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Tempo Desde a Ultima Atualização (Horas)	120,213	131,797	134,347	213,750

FIGURA 12 - Tabela para montagem do gráfico de Candle.

Esses dados foram tabulados com intuito de se montar um gráfico de Candle montando de forma objetiva os extremos da métrica e em qual intervalo está localizada a maior incidência de repositórios.

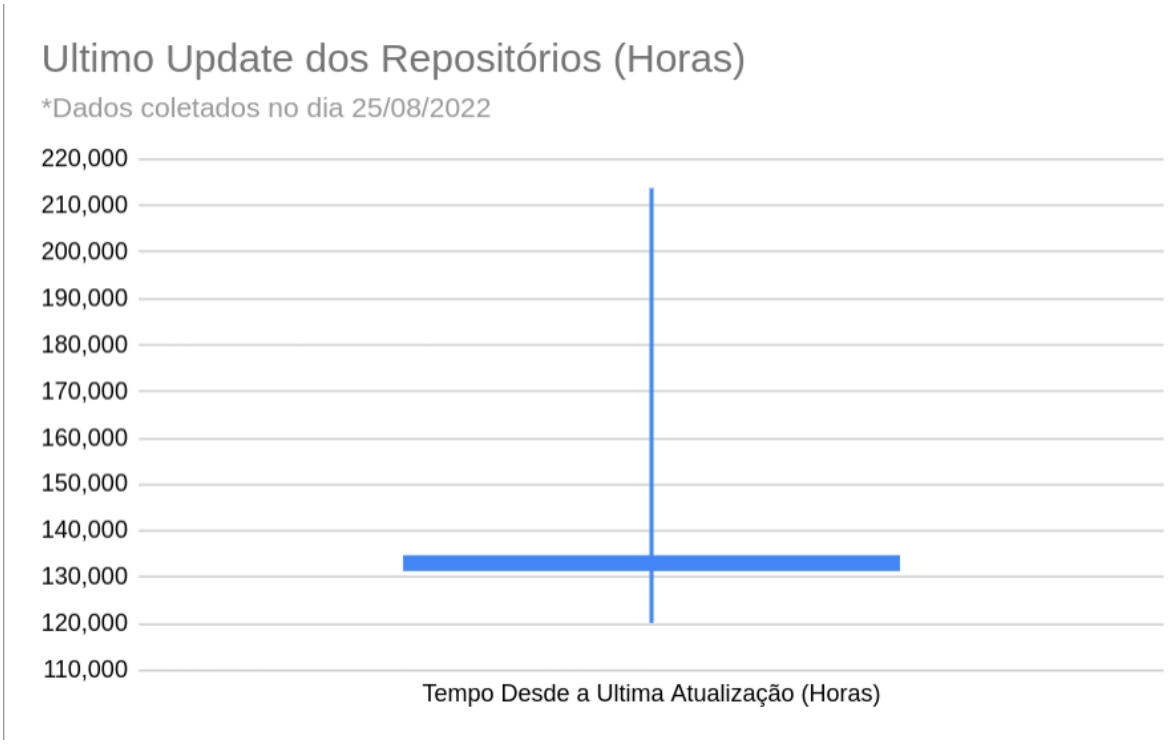

FIGURA 13 - Gráfico de Candle da quantidade de horas desde a última atualização dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação a quantidade de horas desde a última atualização, sendo 500 repositórios que possuem a quantidade de horas desde a última atualização entre 131,797 e 134,347.

Dados complementares gerados são a média, a moda e a mediana dos valores da quantidade de horas desde a última atualização dos repositórios, sendo eles respectivamente:

Media	Moda	Mediana
133,445	134,451	133,602

FIGURA 14 - Média, Moda e Mediana das quantidades de horas desde a última atualização dos repositórios.

3.5. RQ 05 - LINGUAGEM PRIMÁRIA DOS REPOSITÓRIOS

Para análise desta métrica há um importante fator externo a ser considerado, o GitHub disponibilizou uma listagem contendo as 10 linguagens de programação mais utilizadas ao longo dos anos.

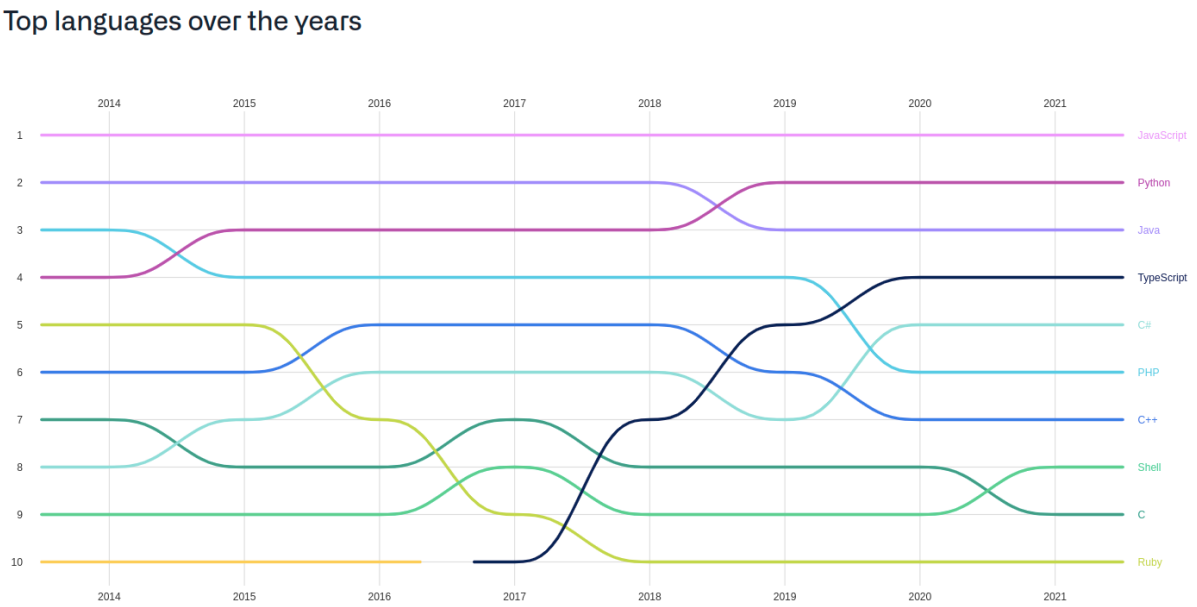

FIGURA 15 - Linguagens de programação mais utilizadas de 2014 a 2021.

Sendo assim, as linguagens mais utilizadas atualmente são respectivamente: JavaScript, Python, Java, TypeScript, C#, PHP, C++, Shell, C e Ruby.

Com base nisso, as linguagens primárias de cada um dos repositórios foram agrupadas com intuito de se gerar um gráfico de colunas que mostre a quantidade de repositórios por linguagem.

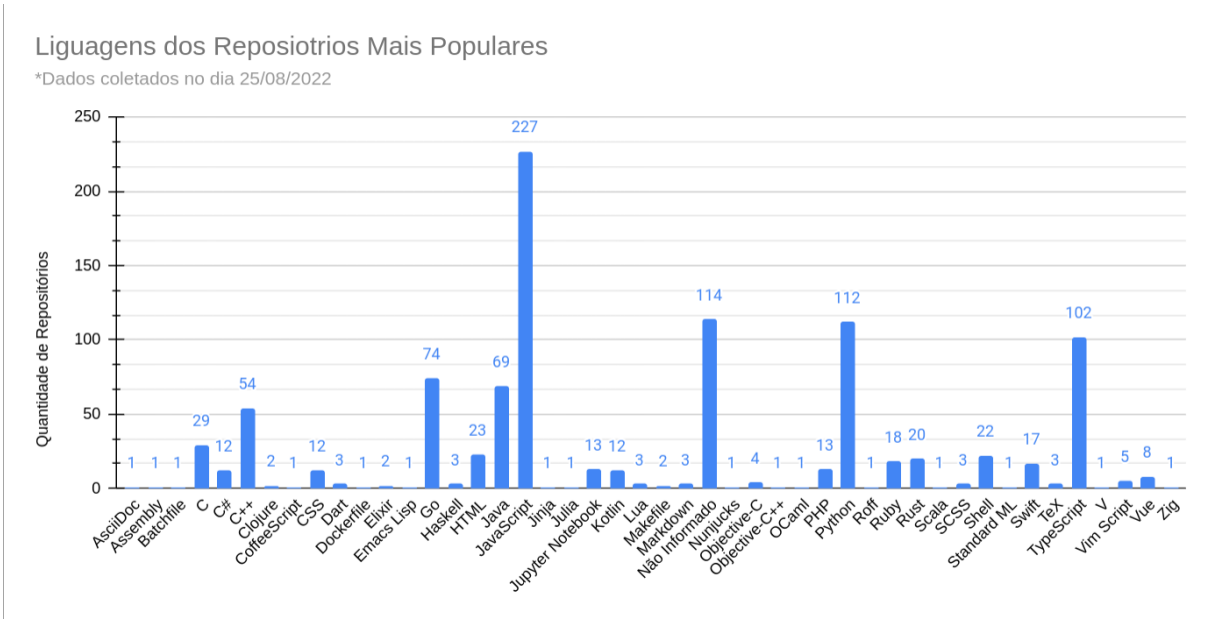

FIGURA 16 - Gráfico de colunas de Linguagem dos Repositórios.

Com base nisso, as linguagens mais populares e a sua incidência nos dados levantados tabuladas.

LINGUAGEM	QUANTIDADE DE REPOSITÓRIOS
JavaScript	227
Python	112
Java	69
TypeScript	102
C#	12
PHP	13
C++	54
Shell	22
C	29
Ruby	18

*Destaque para a linguagem GO que possui 74 repositórios, número superior a 7 das linguagens mais populares.

Desta forma as linguagens mais populares representam 658 repositórios, um percentual de 65,8% dos repositórios analisados.

3.6. RQ 06 - RAZÃO ENTRE O NÚMERO DE ISSUES FECHADAS E TOTAIS DOS REPOSITÓRIOS

Para análise do razão entre o número de *issues* fechadas e totais dos repositórios foram tabelados os valores da menor quantidade (função MIN), primeiro quartil (função QUARTILE com paramento 1), terceiro quartil (função QUARTILE com paramento 3) e da maior quantidade (função MAX) do Total de Issues Fechadas e do Total de Issues conforme a tabela abaixo.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Total de Issues Fechadas	0	160	3116	133178
Total de Issues	0	243	3517	140350

FIGURA 17 - Tabela para montagem do gráfico de Candle.

Esses dados foram tabulados com intuito de se montar um gráfico de Candle montando de forma objetiva os extremos da métrica e em qual intervalo está localizada a maior incidência de repositórios.

FIGURA 18 - Gráfico de Candle da quantidade de Issues Fechadas e Totais dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação a quantidade total de *Issues* fechadas, sendo 500 repositórios que possuem a quantidade total de *Issues* fechadas entre 160 e 3116. Além disso podemos ver também a relação referente ao total de *Issues*, sendo 500 repositórios que possuem a quantidade total de *Issues* entre 243 e 3517.

Dados complementares gerados são a média, a moda e a mediana dos valores da quantidade de *Issues* fechadas e totais, sendo eles respectivamente:

Metrica	Media	Moda	Mediana
Total de Issues Fechadas	2987,765	0	966
Total de Issues	3422,874	0	1175

FIGURA 19 - Média, Moda e Mediana das de Issues fechadas e totais dos repositórios.

Com esses dados foi possível se calcular a razão entre o total de *Issues* fechadas pelo Total de *Issues* de cada um dos repositórios, sendo o valor máximo de 1,000 para os repositórios que tem o total de *Issues* igual ao total de *Issues* fechadas (ou seja todas a *Issues* criadas estão fechadas) e o valor mínimo de 0,000 para os repositórios que não possuem nenhuma *Issue* fechada ou não possuem

nenhuma *Issues* (nenhuma aberta, ou seja nenhuma *Issue* existente). Feito esse cálculo para cada repositório, os valores foram tabulados com intuito de se montar um gráfico de Candle.

Metrica	Menor Valor	Primeiro Quartil	Terceiro Quartil	Maior Valor
Razão (Issues Fechadas por Total de Issues)	0,000	0,677	0,951	1,000

FIGURA 20 - Tabela para montagem do gráfico de Candle.

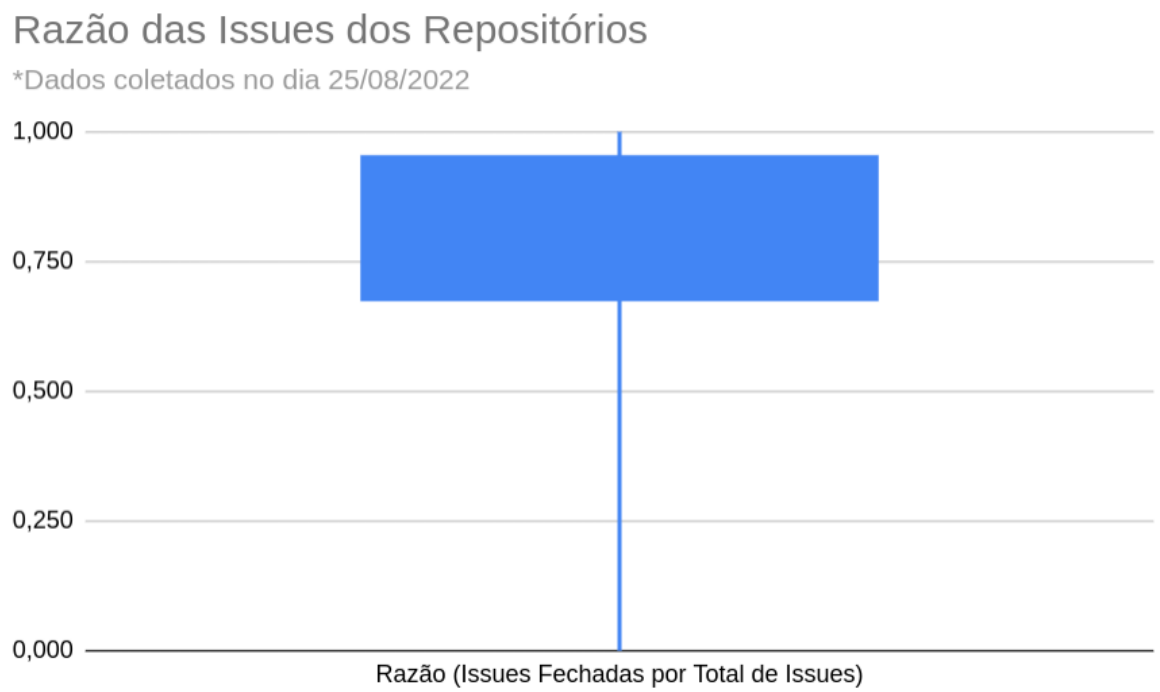

FIGURA 21 - Gráfico de Candle da razão de Issues Fechadas e Totais dos Repositórios.

Através do gráfico de Candle acima é possível visualizar onde se encontram a maior parte dos repositórios em relação à razão de Issues Fechadas e Totais, sendo 499 repositórios que possuem a razão entre 0,677 e 0,951.

4. CONCLUSÃO

As hipóteses de informais feitas nas etapas iniciais deste trabalho foram comparadas com os resultados obtidos a fim de serem validadas ou refutadas. Para essa análise foram observados os dados coletados e as visualizações criadas, em especial os gráficos de Candle que mostram o maior intervalo de agrupamento dos repositórios em cada uma das métricas colhidas. Desta forma as hipóteses e a conclusão sobre cada uma após análises foram agrupadas de forma resumida na tabelas a seguir.

HIPÓTESE INFORMAL	STATUS
RQ 01	Refutada
RQ 02	Refutada
RQ 03	Refutada
RQ 04	Validada
RQ 05	Validada
RQ 06	Validada

As três primeiras hipóteses, referentes aos três primeiros questionamentos levantados foram consideradas refutadas, uma vez que a maioria dos repositórios não se encaixou nas afirmações feitas, sendo:

- 39,3% dos repositórios mais populares possuem idade superior a 8 anos;
- 74,9% dos repositórios mais populares possuem contribuições externas inferiores a 1485;
- 80,4% dos repositórios mais populares possuem o número de releases lançados inferior a 80.

Portanto a conclusão acerca dos três primeiros questionamentos levantados são:

- **RQ 01 - Sistemas populares são maduros/antigos?** Os repositórios mais populares não são em sua maioria antigos.
- **RQ 02 - Sistemas populares recebem muita contribuição externa?** Os repositórios mais populares em sua maioria não recebem muitas contribuições externas.
- **RQ 03 - Sistemas populares lançam releases com frequência?** Os repositórios mais populares em sua maioria não lançam releases com frequência.

As três últimas hipóteses, referentes aos três últimos questionamentos levantados foram consideradas válidas, uma vez que a maioria dos repositórios se encaixou nas afirmações feitas, sendo:

- 96,9% dos repositórios mais populares foram atualizados em menos de 7 dias (168 horas), tempo esse inferior ao tempo de uma sprint convencional de desenvolvimento (14 dias ou 336 horas);
- 65,8% dos repositórios mais populares estão escritos nas 10 linguagens mais populares segundo o próprio GitHub;
- 81,6% dos repositórios mais populares possuem percentual de issues fechadas superior a 0,600.

Portanto a conclusão acerca dos três últimos questionamentos levantados são:

- **RQ 04 - Sistemas populares são atualizados com frequência?** Os repositórios mais populares são sim atualizados com maior frequência.
- **RQ 05 - Sistemas populares são escritos nas linguagens mais populares?** Os repositórios mais populares são sim escritos nas linguagens mais populares.
- **RQ 06 - Sistemas populares possuem um alto percentual de issues fechadas?** Os repositórios mais populares possuem sim um alto percentual de issues fechadas.