Lab 01

Carlos Eduardo Aquino
Pedro Rodrigues

1. Introdução

Esse laboratório consiste em responder seis perguntas relacionadas aos repositórios mais populares no github, sendo elas:

1. Sistemas populares são maduros/antigos?
2. Sistemas populares recebem muita contribuição externa?
3. Sistemas populares lançam releases com frequência?
4. Sistemas populares são atualizados com frequência?
5. Sistemas populares são escritos nas linguagens mais populares?
6. Sistemas populares possuem um alto percentual de issues fechadas?

Para responder essas perguntas foram definidas uma série de métricas, sendo elas:

1. idade do repositório
2. total de pull requests aceitas
3. total de releases
4. tempo até a última atualização
5. linguagem primária de cada um desses repositórios
6. razão entre número de issues fechadas pelo total de issues

2. Hipóteses informais

Foram definidas seis respostas que o grupo julgou serem prováveis respostas para a pergunta realizada. Elas são:

Hipótese 1: Sistemas antigos tendem a ser mais populares, já que um repositório provavelmente tende a se manter ativo caso ele seja popular. Além disso, a chance de pessoas conhecerem repositórios antigos é maior que repositórios novos, considerando que ela tem um espaço de tempo maior para descobrir o repositório.

Hipótese 2: Sistemas populares recebem mais contribuição externa, considerando que mais pessoas conhecem o sistema, portanto irão existir mais ideias de como melhorar o sistema.

Hipótese 3: Sistemas populares não lançam releases com mais frequência, já que a quantidade de pessoas que contribuem em um repositório não aceleram a velocidade na qual releases saem, por mais que provavelmente a quantidade de contribuições aumentará

Hipótese 4: Sistemas populares são atualizados com frequência pois mais pessoas conhecem o sistema, o que faz mais pessoas quererem contribuir no projeto.

Hipótese 5: Sistemas populares são escritos nas linguagens mais populares, pois as pessoas tendem a preferir sistemas escritos em linguagens que elas entendam quando se trata de um repositório do github, assim como elas tendem a conhecer as linguagens mais populares.

Hipótese 6: Sistemas populares não possuem um alto número de issues fechadas, considerando que existem mais pessoas tanto para abrir issues quanto para solucioná-las

3. Metodologia

A coleta de dados foi feita utilizando a API GraphQL do GitHub. Foram coletados os seguintes dados: nome do repositório, data de criação, total de pull requests aceitas, total de estrelas, total de issues fechadas, total de issues, data da última atualização, linguagem primária, e total de releases.

O total de pull requests aceitas, o total de releases e a linguagem primária de cada repositório serão utilizados sem qualquer tipo de processamento relacionado ao seu valor.

A data de criação será subtraída da data atual para determinar a idade do repositório em dias. Já a data da última atualização será subtraída da data atual para determinar qual foi o tempo desde a última atualização.

O total de issues fechadas será dividido do total de issues para determinar a métrica definida para a pergunta 6.

A partir disso, foi feito um boxplot e um gráfico de dispersão para as perguntas 1, 2, 3, 4 e 6, e foi feito um gráfico de barras para a pergunta 5, com o objetivo de analisar os dados coletados e tirar uma conclusão sobre eles.

Para analisar a pergunta 1, será avaliado se a maioria dos repositórios analisados são antigos. Repositórios com 2000 dias ou mais serão considerados antigos.

Para analisar a pergunta 2, será avaliado se a maioria dos repositórios possuem muitas pull requests aceitas. Repositórios com 10.000 pull requests ou mais serão considerados com muitas pull requests aceitas.

Para analisar a pergunta 3, será avaliado se a maioria dos repositórios possuem muitas releases. Repositórios com 100 ou mais releases serão considerados com muitas releases.

Para analisar a pergunta 4, será avaliado se a maioria dos repositórios foram atualizados recentemente. Repositórios que foram atualizados há 30 dias ou menos serão considerados como atualizados recentemente.

Para analisar a pergunta 5, será avaliado se a maioria dos repositórios possuem como linguagem primária uma linguagem popular. Linguagens populares foram determinadas a partir do link [1]. O gráfico de barras montado foi filtrado de modo que ele apenas mostre as linguagens mais populares.

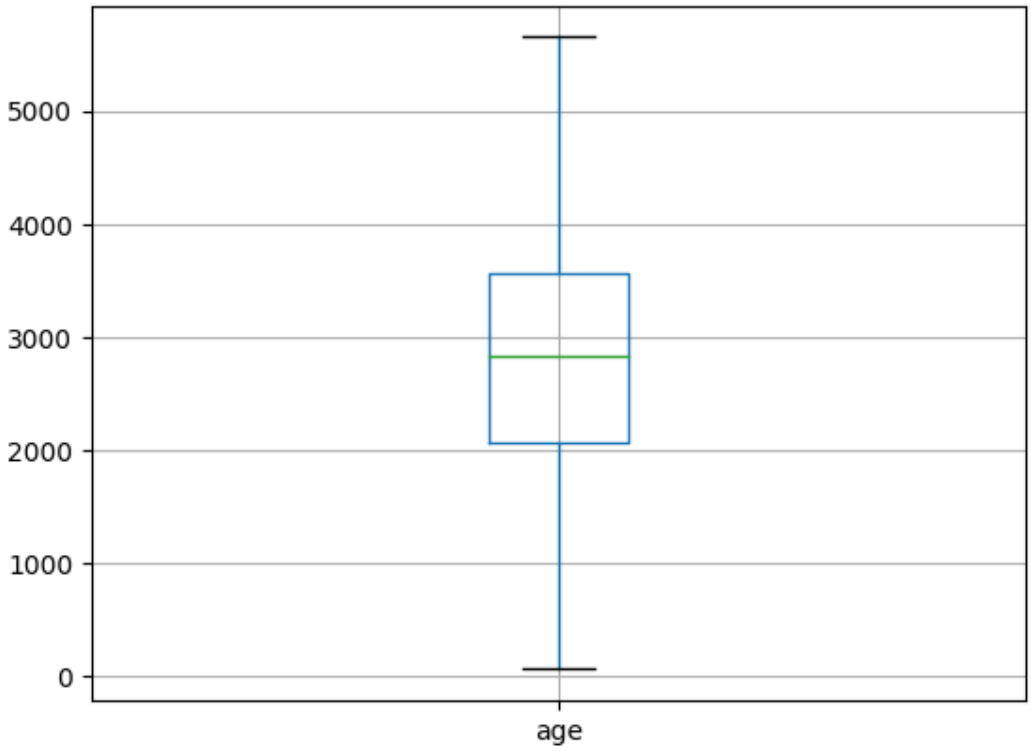
Para analisar a pergunta 6, será avaliado se a maioria dos repositórios possuem um percentual alto de issues fechadas. Repositórios que possuem um percentual de 90% de issues fechadas serão considerados com um percentual alto de issues fechadas.

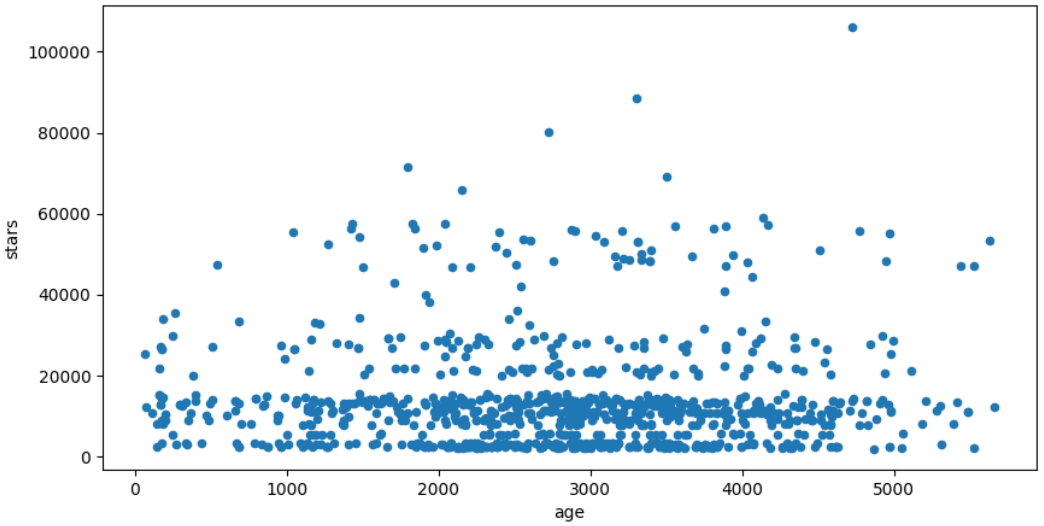
[1]https://octoverse.github.com/2022/top-programming-languages

4. Resultados obtidos

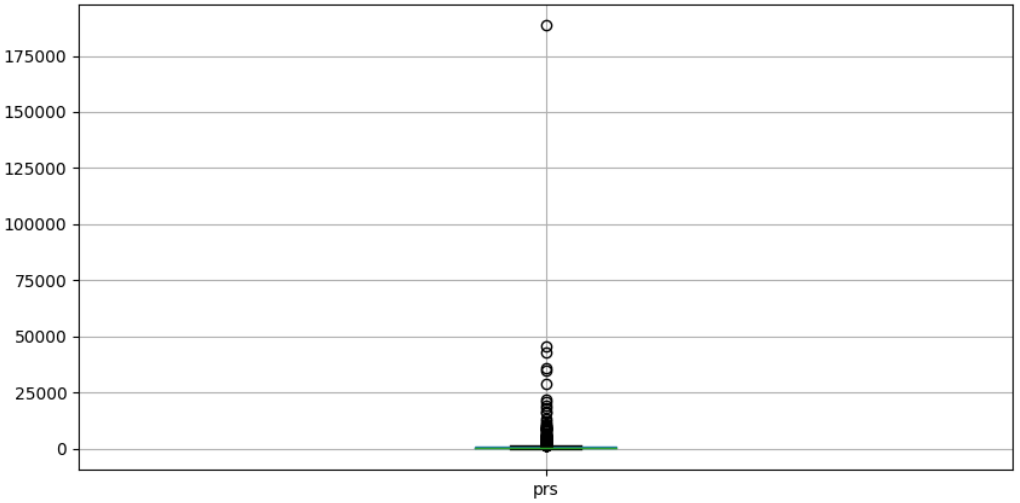
Os resultados são apresentados a seguir:

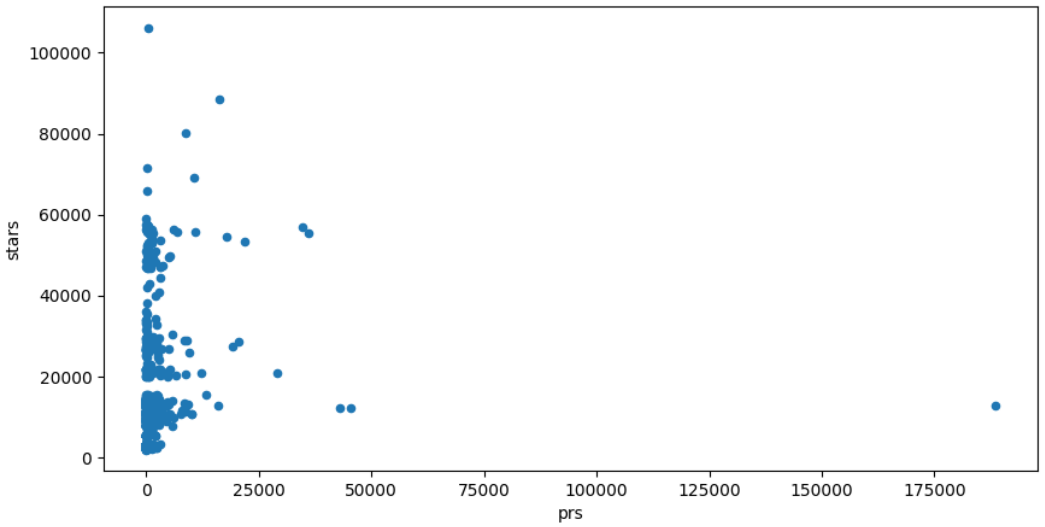
Pergunta 1:



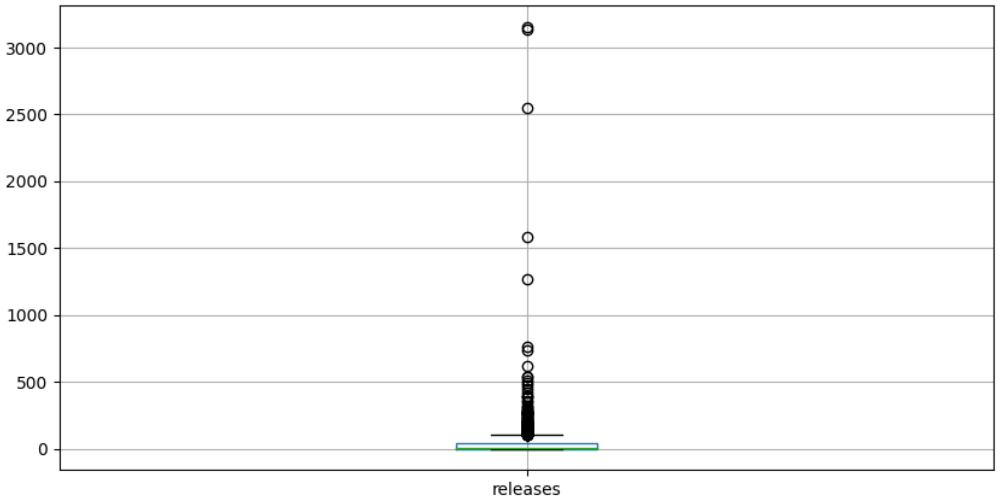

100000
80000
60000
40000
20000
0
stars
0
1000
2000
3000
4000
5000
age

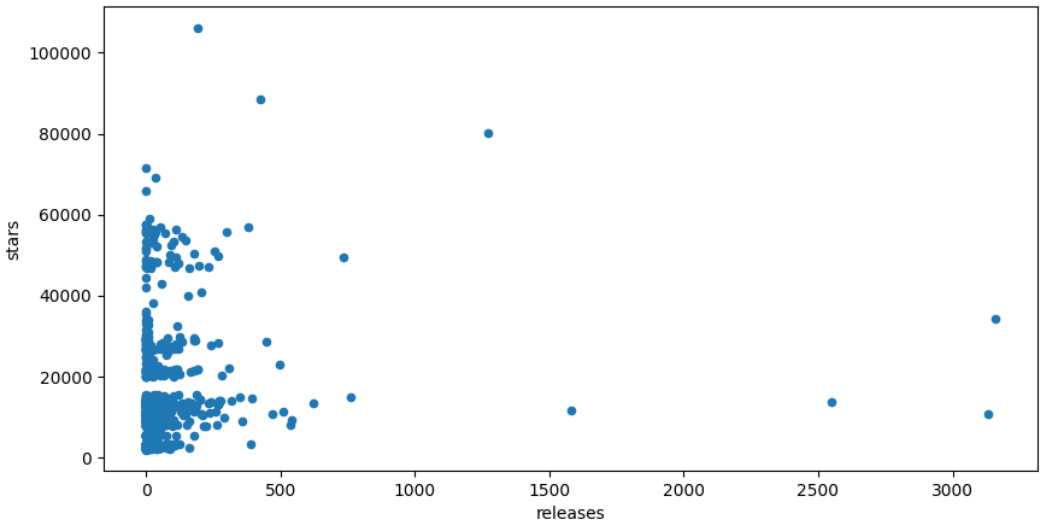
Pergunta 2:


175000
150000
125000
100000
75000
50000
25000
0
prs

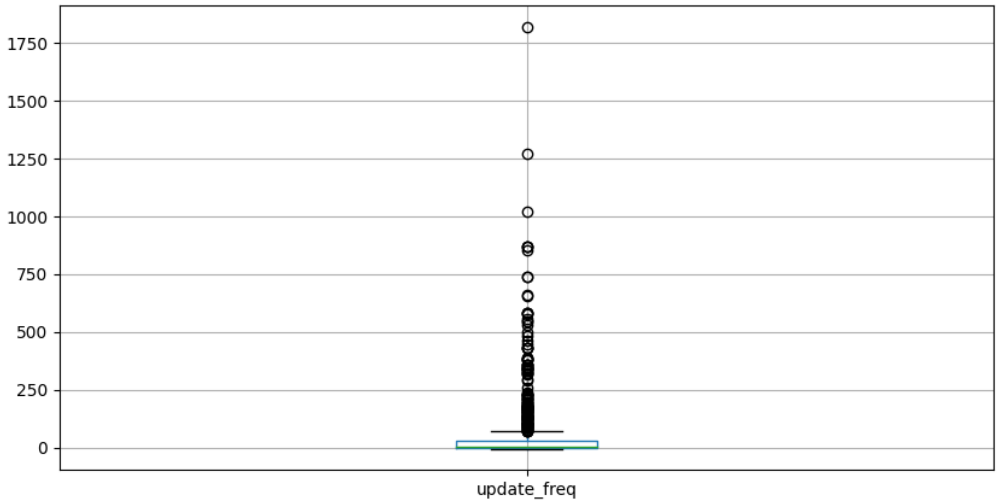

stars
prs
100000
80000
60000
40000
20000
0
0
25000
50000
75000
100000
125000
150000
175000

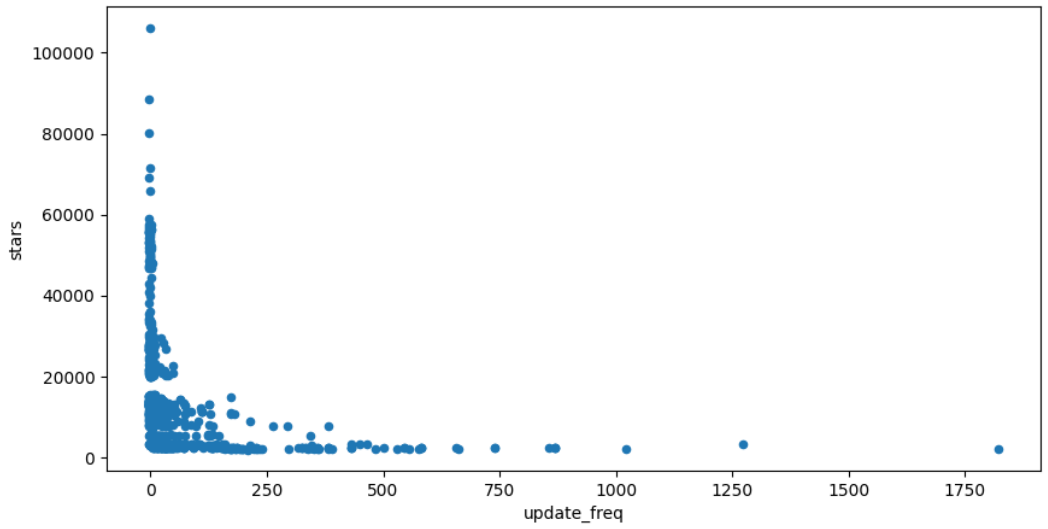
Pergunta 3:


releases
3000
2500
2000
1500
1000
500
0

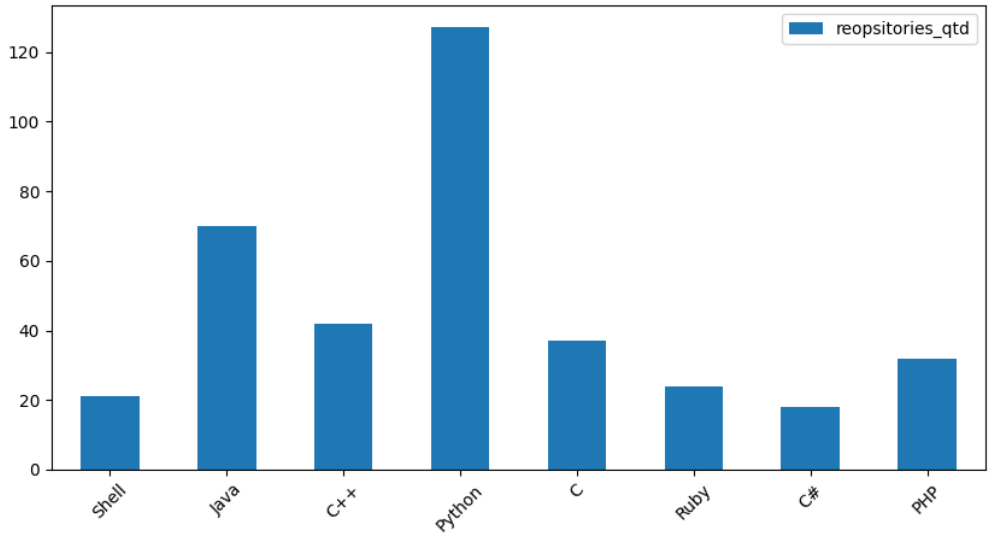

100000
80000
60000
40000
20000
0
stars
0
500
1000
1500
2000
2500
3000
releases

Pergunta 4:

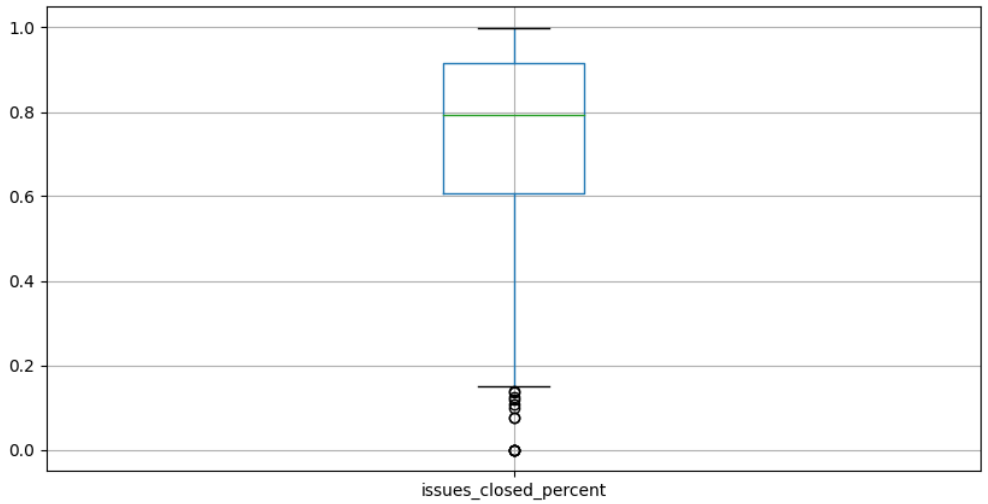

1750
1500
1250
1000
750
500
250
0
update_freq

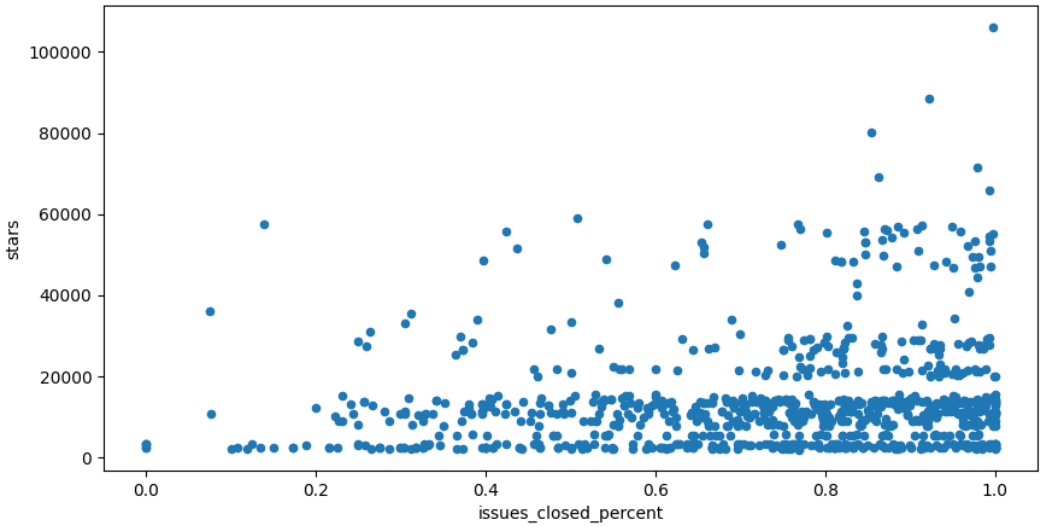


Pergunta 5:



Pergunta 6:





5. Apresentação dos resultados e discussão das hipóteses

Pergunta 1:

É possível perceber que a maioria dos repositórios de fato são antigos, assim como previsto, já que a maioria dos repositórios encontrados possuem 2000 dias ou mais de idade.

Pergunta 2:

Observa-se que a maioria dos repositórios não possuem muita contribuição externa, ao contrário do previsto, pois a maioria dos repositórios encontrados possuem

menos de 10.000 pull requests aceitas. Isso provavelmente se deve ao fato de que uma grande parte dos repositórios populares filtra de forma rigorosa as contribuições externas, reduzindo o número de contribuições.

Pergunta 3:

É possível observar que a maioria dos repositórios não lançam releases com frequência, assim como previsto, considerando que a maioria dos repositórios não possuem mais de 100 releases.

Pergunta 4:

É possível observar que repositórios populares são atualizados com muita frequência, assim como previsto, já que a maioria dos repositórios foram atualizados há menos de 30 dias.

Pergunta 5:

É possível observar que repositórios populares não tendem a utilizar as linguagens mais populares, contradizendo o que havia sido previsto, já que aproximadamente 380 repositórios dos 1.000 repositórios minerados utilizam linguagens populares.

Pergunta 6:

É possível observar que a maioria dos repositórios não possuem uma alta quantidade de issues fechadas, assim como previsto, pois o percentual de issues fechadas da maioria dos repositórios não ultrapassa 80%.