Lab 02 - Sprint 03

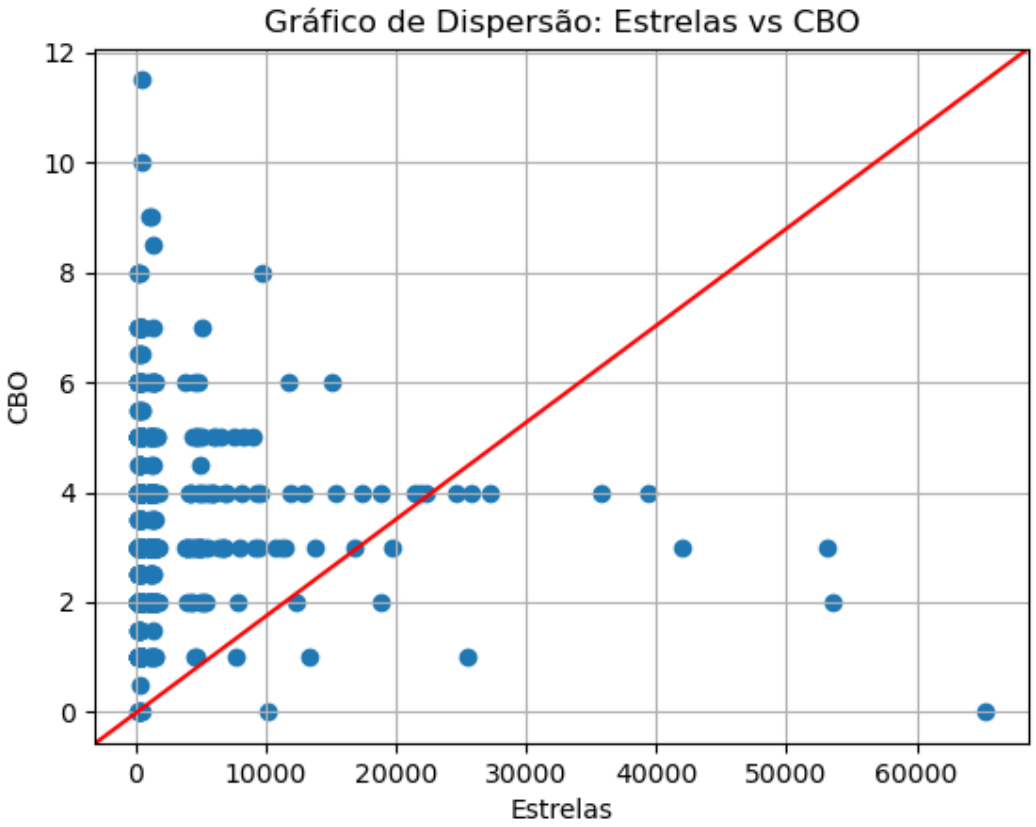
Gabriel Victor Couto, João Victor Guerra,
Luiz Gustavo

Metodologia

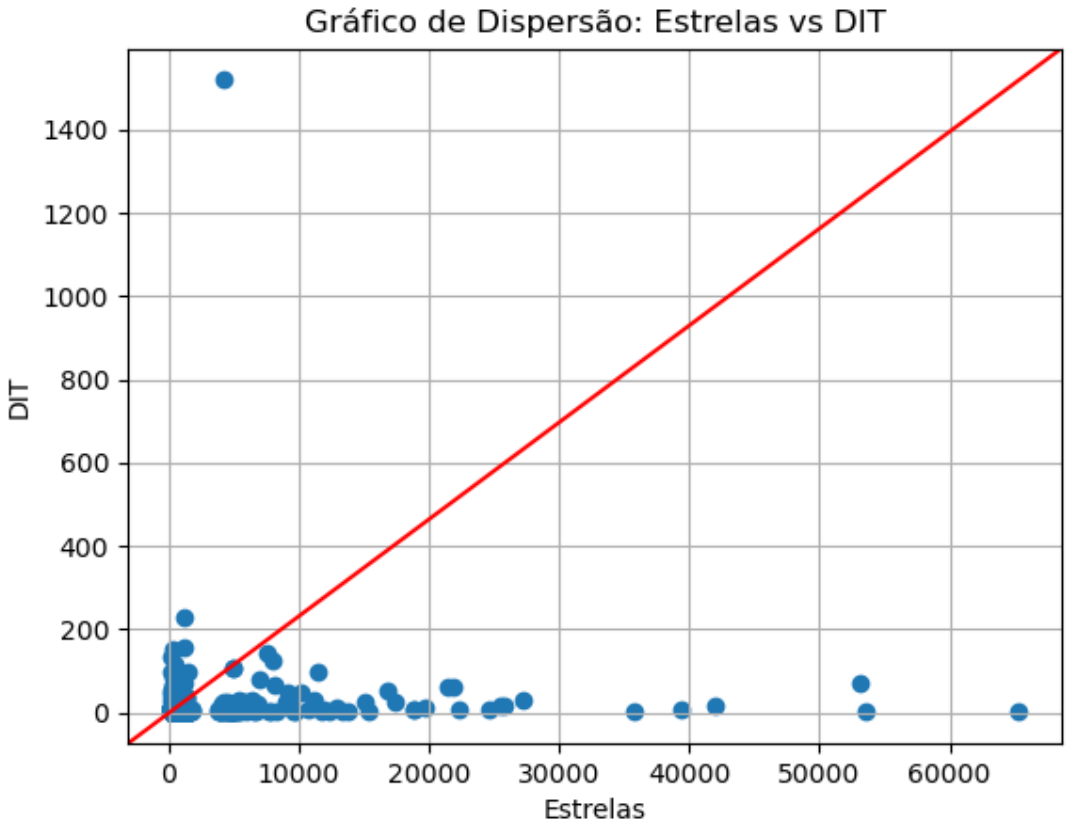
Com o objetivo de extrair métricas de qualidade de software dos 1000 repositórios mais populares da linguagem Java no GitHub e compará-las com características do repositório, o grupo adotou a seguinte estratégia;
Através da API GraphQL do próprio GitHub, obteve-se uma lista dos dados dos repositórios, salvando em um arquivo csv os respectivos dados: nome, idade do repositório, número de estrelas e número de releases.
Após isso, o grupo desenvolveu um script que possibilitou a leitura desse arquivo gerado e que cada repositório da lista fosse clonado em uma pasta específica, tendo suas métricas lidas através da ferramenta CK, gerando posteriormente outro arquivo csv contendo os novos dados obtidos. Após isso, de forma automatizada, os dois csvs foram concatenados, gerando um único csv com todos os dados necessários.
Para fins de performance e armazenamento, depois que um repositório era analisado pelo CK, o mesmo era excluído da pasta. Assim, não havia mais que um repositório clonado no projeto em tempo de execução.
Com os dados obtidos, foi necessário descartar aqueles que não contribuíam para a pesquisa, isto é, aqueles cujos valores eram nulos nas colunas de LOC, LCOM, DIT e CBO.
Por fim, através da biblioteca matplotlib, foram gerados gráficos que correlacionavam os aspectos de popularidade, tamanho, atividade e maturidade com os aspectos de qualidade CBO, DIT e LCOM. O relatório apresenta os resultados obtidos, tal como a análise de cada gráfico e discussões a respeito das hipóteses dos integrantes do grupo.

Resultados Obtidos

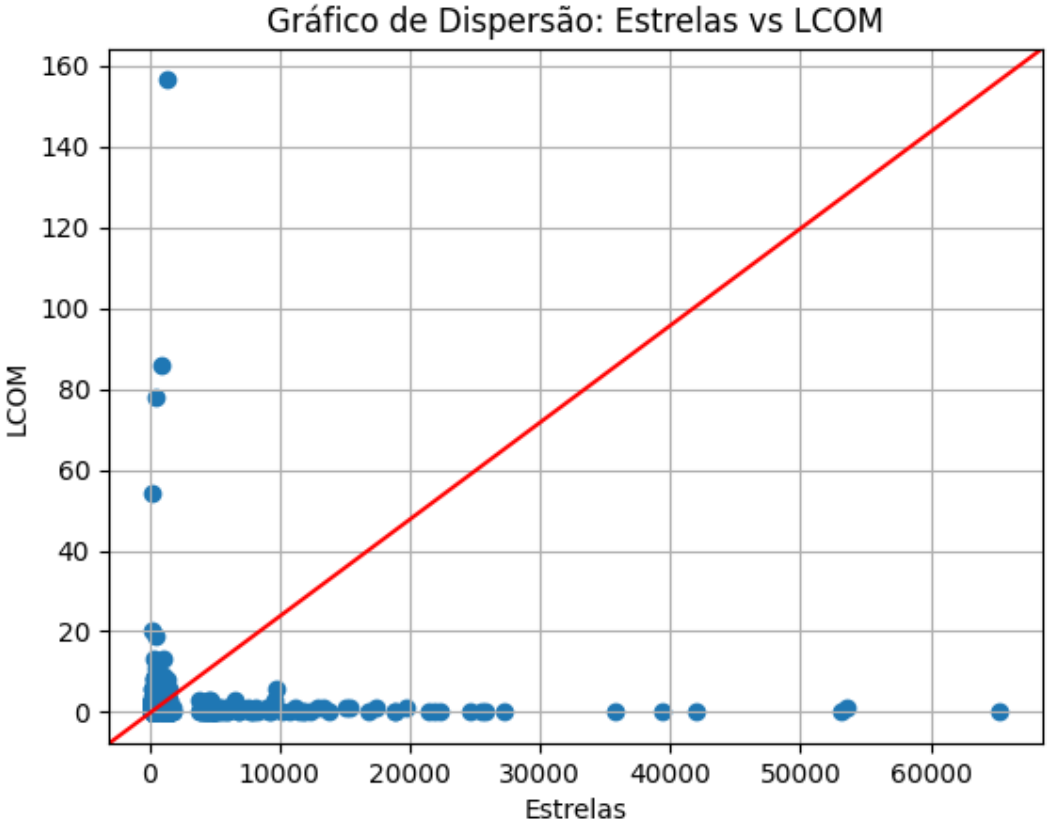
Popularidade:



Se considerarmos que o número de estrelas dita a popularidade de um repositório, conseguimos observar que para os repositórios de até 30000 estrelas, os valores de CBO são distribuídos de modo semelhante, apresentando muitas concentrações em quase todos os valores do gráfico, se aglutinando mais por volta do 4, um valor médio.
Os repositórios ainda mais populares (+30000 estrelas) não ultrapassam o valor 4 de CBO, e, como um todo, o CBO decai – indicando que se trata de projetos com baixo acoplamento

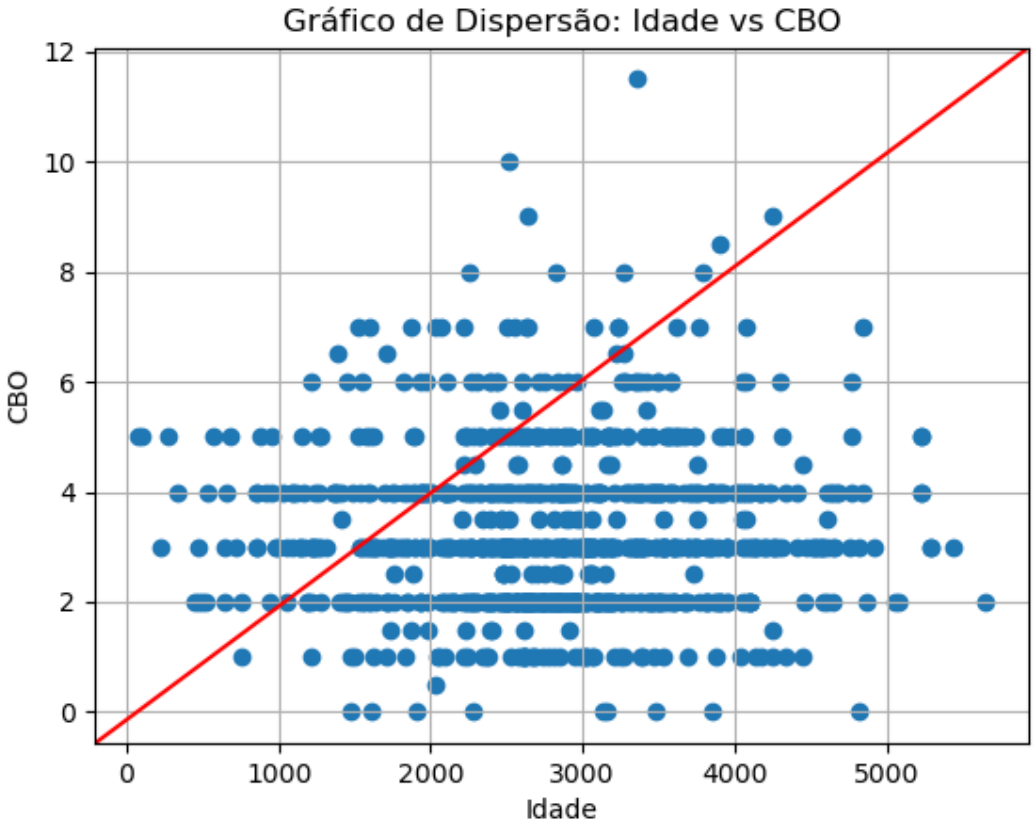


Correlacionando-se popularidade e DIT, percebe-se que a popularidade dos repositórios não tende a alterar o DIT, e que o mesmo, como um todo, se concentra entre os valores 0 e 180, com apenas duas anomalias – uma que ultrapassa a marca de 200 e outra que ultrapassa a de 1400.

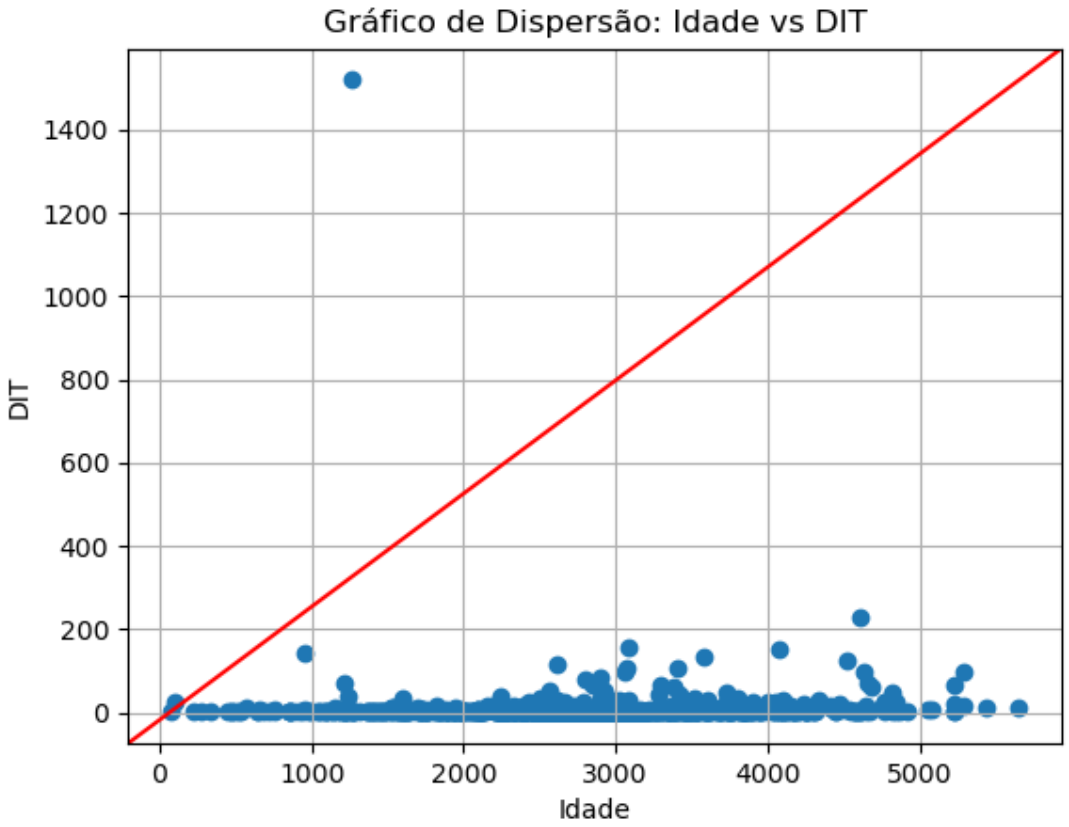


Nesse caso, correlacionando-se popularidade e LCOM, pode-se perceber que o valor é abaixo de 20 para maioria dos repositórios, com apenas 4 exceções, ou seja, o valor de LCOM é independente do valor de estrelas. Há apenas algumas exceções que ultrapassam o valor 20.

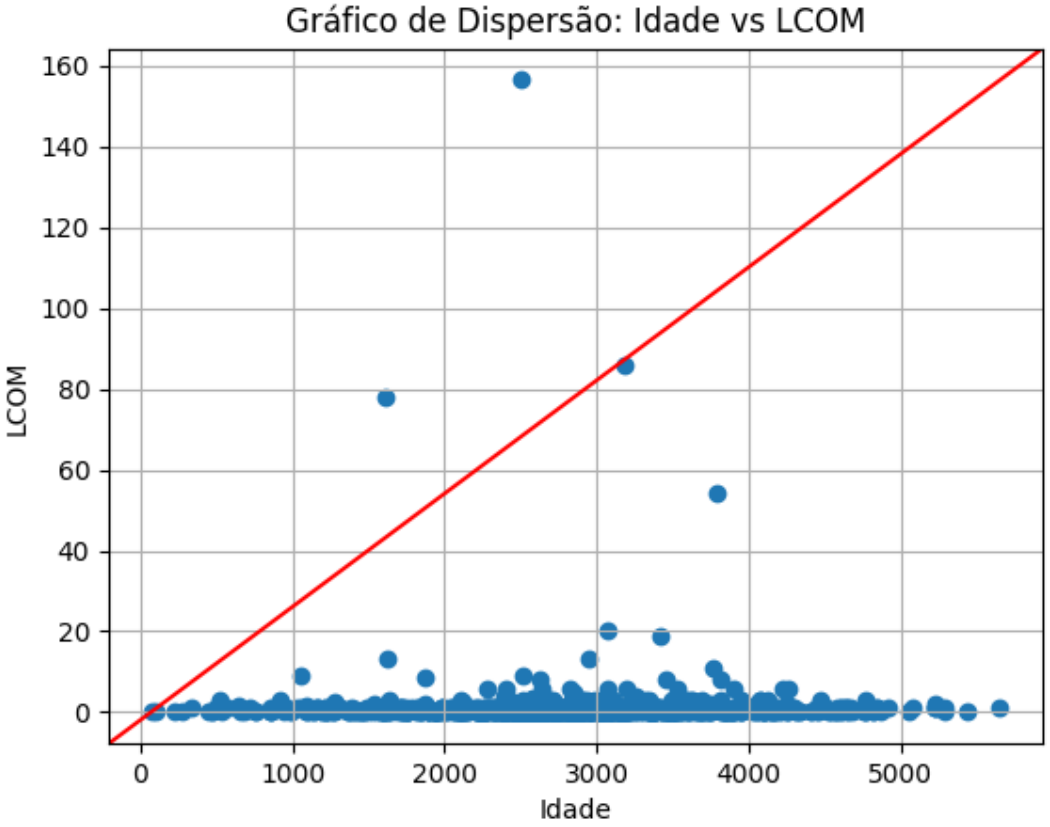
Idade:



Correlacionando-se a idade dos repositórios e o CBO, percebe-se um padrão de distribuição equilibrado, levando a crer que o CBO não é afetado pela idade de um repositório. O CBO médio se concentra entre 2 e 4, e é possível notar alguns repositórios que atingiram valores altos, como 9, 10 e 11. Os valores se aglutinam em posições semelhantes.

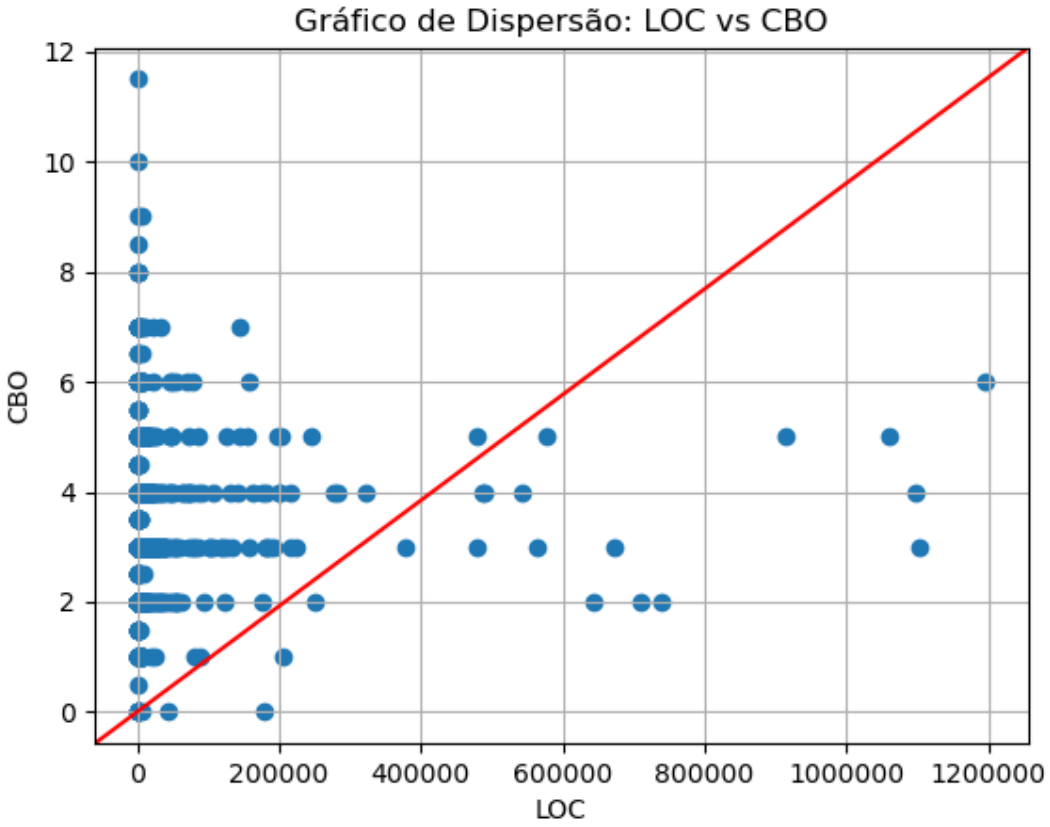


Correlacionando-se a idade dos repositórios e DIT, nota-se uma concentração nos valores de DIT entre 0 e ~120, o que nos leva a crer que a idade de um repositório não afeta diretamente no DIT. O DIT fugiu desse padrão em apenas dois casos, com um valor acima de 200 e outro acima de 1400.

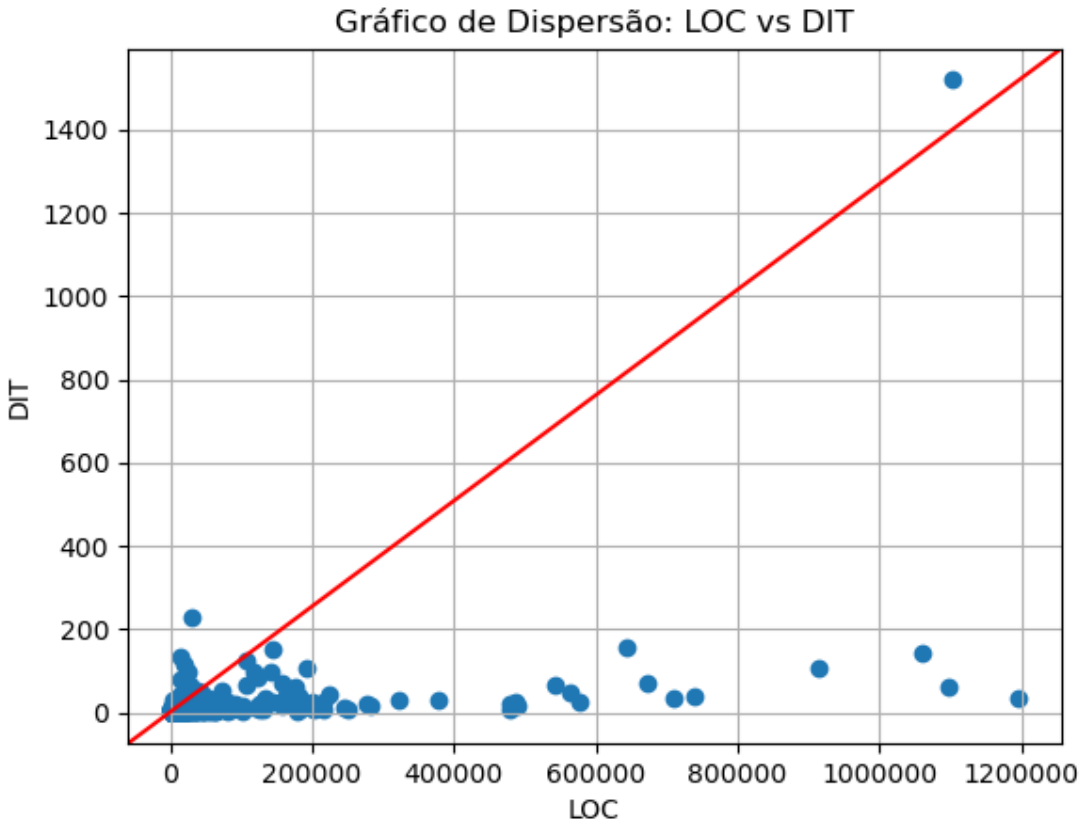


Correlacionando-se LCOM e a idade dos repositórios, percebe-se que, a tendência do LCOM é de se manter abaixo de vinte, independentemente da idade dos repositórios, logo, ambos não possuem correlação.

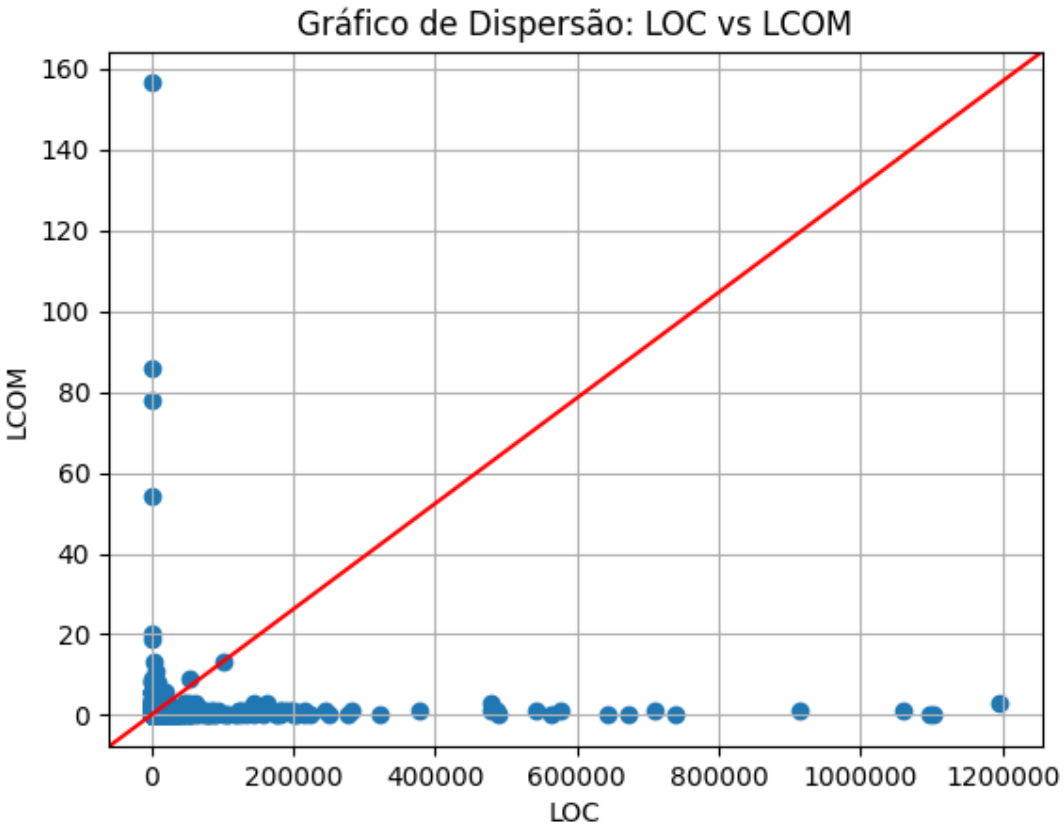
LOC:



Correlacionando-se CBO e LOC não se pode assumir que o tamanho do repositório influencie em seu valor de CBO, visto que a maior parte dos repositórios está entre 2 e 6 de CBO, desde 0 a 200 mil linhas - onde se encontram a maior parte do repositório - até o maior de todos, com 1,2 milhão de linhas.

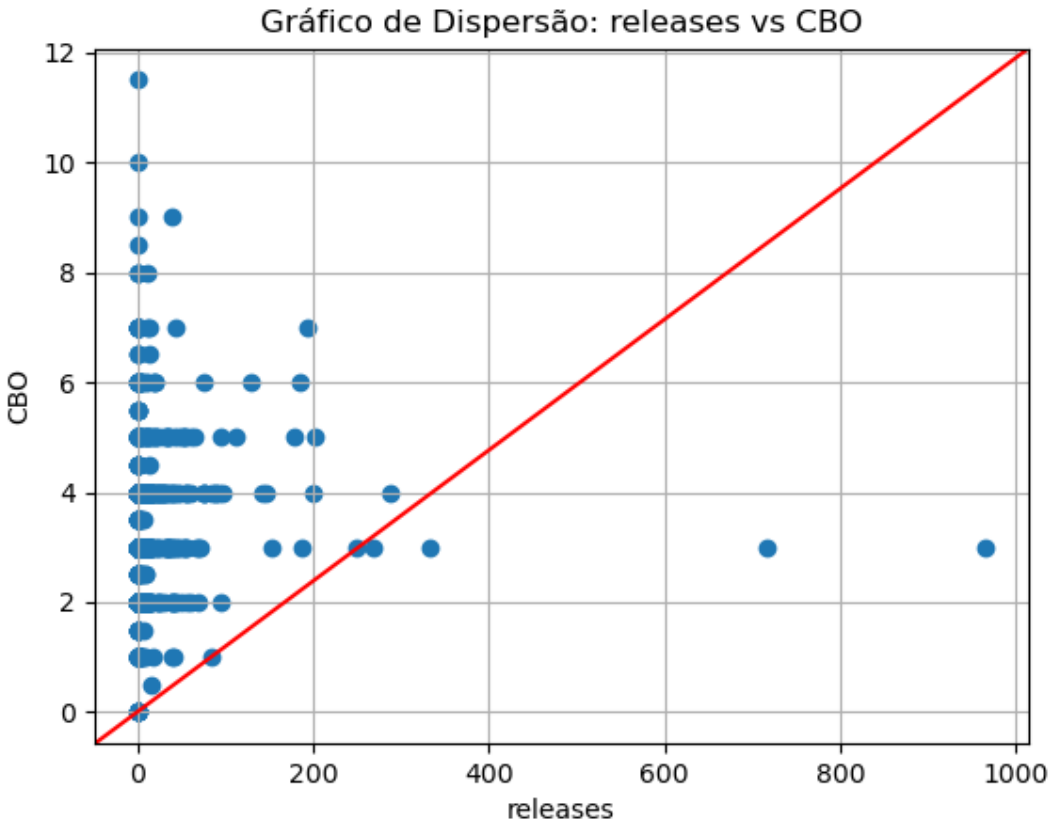


Quando analisados DIT e LOC, podemos assumir que não existe relação entre eles, visto que o valor de DIT se manteve quase que totalmente abaixo de 200, sendo assim, o tamanho de um repositório não influencia diretamente em DIT.

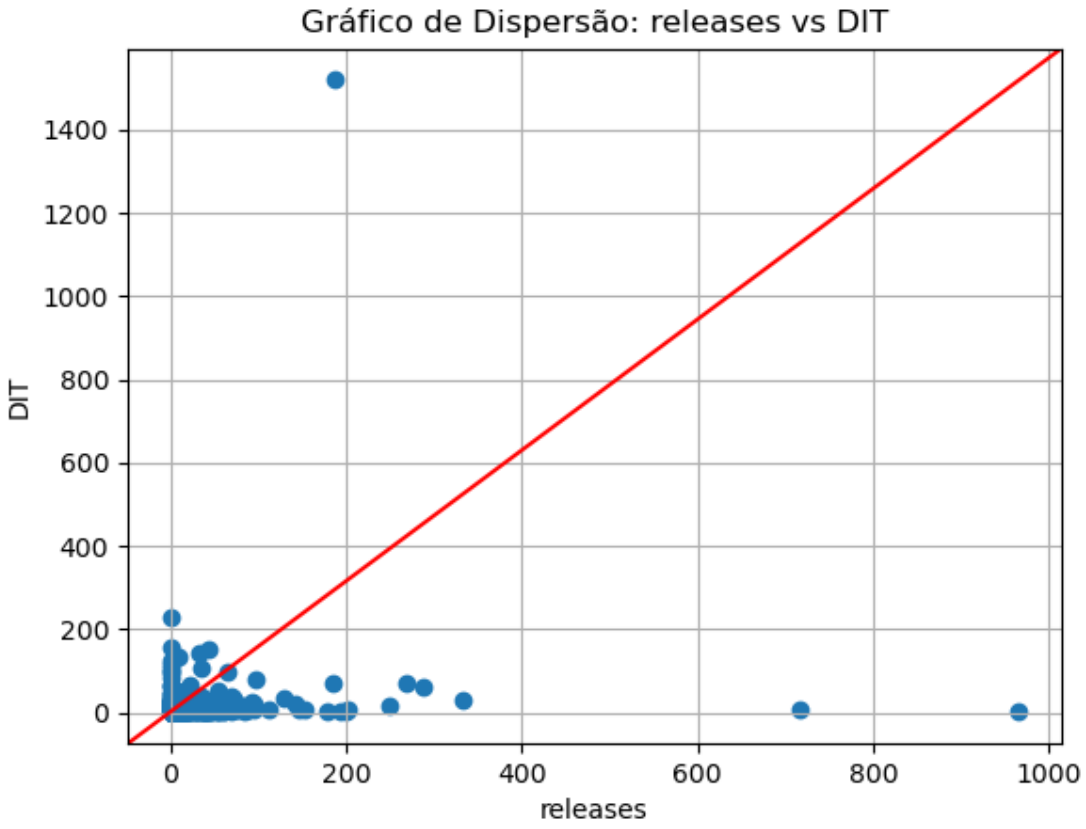


O valor de LOC, embora podendo alcançar altos valores de 1,2 milhão de linhas, não se expressa em conjunto com LCOM, ou seja, repositório maiores não possuem valor de LCOM maior ou menor, pois não possuem relação direta.

Releases:

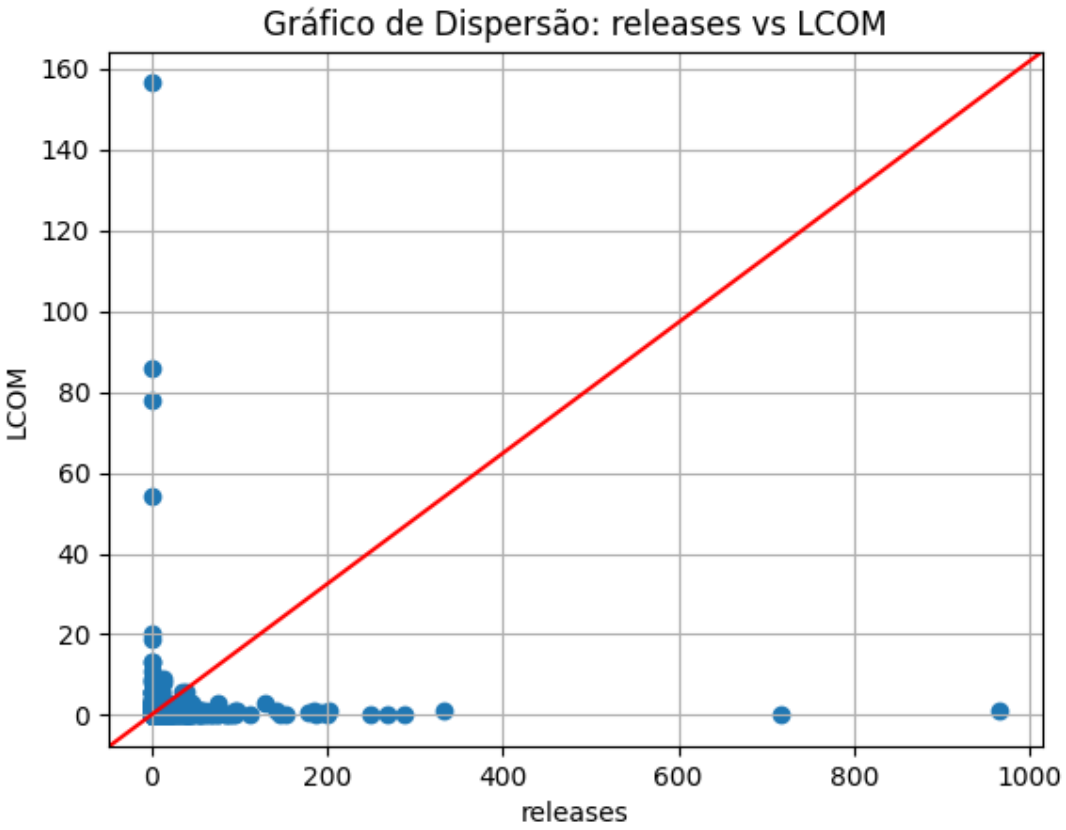


,
Correlacionando-se releases e CBO, pode-se apontar o fato de que repositórios com menos releases alcançaram níveis mais altos de CBO, mas não se pode traçar seguramente uma correlação entre ambos, visto que, por padrão, a maioria dos repositórios obtidos possuíam 0 ou poucas releases. O CBO médio concentrou-se entre 2 e 4.



,

Correlacionando-se DIT e releases, percebe-se uma aglutinação e valores semelhantes, onde o DIT se concentra entre 0 e ~180. Isso ocorre pois a maioria dos repositórios obtiveram valores semelhantes nas duas grandezas.



Apesar da concentração de LCOM estar entre 0 e 20 e a de releases estar entre 0 e 200, ou seja, ambas estão proporcionalmente próximas de 0-. não se pode afirmar que tenham relação.

Hipóteses

RQ 01. Qual a relação entre a ***popularidade*** dos repositórios e as suas características de qualidade?

R: Espera-se que quanto maior a popularidade de um repositório, mais qualidade ele tenha, tanto em aspectos técnicos quanto organizacionais. Por ser alvo de atenção de diferentes públicos, incluindo a comunidade desenvolvedora, é natural que o projeto aceite contribuições e que tenha mais preocupação com a qualidade.

Desse modo, baixo LCOM, CBO moderado (devido ao porte de projetos populares) e DIT baixado/controlado.

Comentário: Após a análise dos gráficos que relacionam popularidade com qualidade, não se pode aferir nenhuma relação direta entre os dados coletados. As

métricas de qualidade parecem se manter concentrados em valores médios, acumulando, assim, vários repositórios em um mesmo ponto do gráfico

RQ 02. Qual a relação entre a ***maturidade*** dos repositórios e as suas características de qualidade ?

R: É de se esperar que quanto mais maduro for um repositório, maior a experiência técnica dos desenvolvedores e a familiaridade com problemas. Dessa forma, repositórios mais maduros tendem a possuir maior qualidade, desde que sejam constantemente atualizados.

Desse modo, baixo LCOM, baixo CBO e DIT baixado/controlado.

Comentário: Na comparação das métricas de qualidade com a idade do repositório, podemos afirmar que as métricas não se relacionam. Enquanto o CBO apresentou uma dispersão ocupando todo gráfico, LCOM e DIT se mantiveram em valores reduzidos- de 0 a 20 e de 0 a 180, respectivamente- independente da idade do repositório.

RQ 03. Qual a relação entre a ***atividade*** dos repositórios e as suas características de qualidade?

R: Espera-se que desde que as atividades (contribuições internas e externas) passem por um processo de controle de qualidade e testes, quanto maior a atividade de um repositório, maior será sua qualidade, visto que os contribuintes se empenharão para resolver problemas e criar novas soluções.

Comentário: Para análise de atividade é importante destacar que a maior parte dos repositórios possui um baixo valor de releases, ou seja, menor de 200, sendo assim, os valores se acumularam todos próximos ao eixo das ordenadas. Tendo isso isso em vista, não foi possível concluir quaisquer relações entre as informações coletadas.

RQ 04. Qual a relação entre o ***tamanho*** dos repositórios e as suas características de qualidade?

R: Pode-se deduzir que uma parte considerável dos repositórios grandes são populares, e consequentemente recebem contribuições que ajudam-nos a manter um bom nível de qualidade. Entretanto, descartada essa hipótese, não se pode inferir a qualidade de um projeto a partir de seu tamanho.

Comentário: Assim como nas análises anteriores, os dados de qualidade não expressam correlação com o fator de referência, visto que as métricas se

mantiveram sempre com distribuição quase que constante e independente do outro fator.