

**Laboratório de Experimentação de Software**  
**Laboratório II- Um Estudo das Características de Qualidade de Sistemas Java**  
**Grupo: Davi Santos; Rafael Duarte.**

**Instituto de Informática e Ciências Exatas– Pontifícia Universidade de Minas Gerais**  
**(PUC MINAS)**  
**Belo Horizonte – MG – Brasil**

## **1. Introdução**

O objetivo deste laboratório é analisar aspectos da qualidade de repositórios desenvolvidos na linguagem Java, correlacionando-os com características do seu processo de desenvolvimento, sob a perspectiva de métricas de produto calculadas através da ferramenta CK.

## **2. Hipóteses Informais**

Para o direcionamento do estudo foram levantadas as seguintes hipóteses:

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

Um grande número de contribuidores pode indicar que o software é bem documentado e fácil de entender, o que pode levar a um código mais limpo e de alta qualidade. A frequência das contribuições também pode ser um sinal de que o código é mantido ativamente e melhorado constantemente.

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

Um repositório mais maduro pode ter uma documentação mais completa e uma comunidade de suporte mais experiente e ativa, o que pode ajudar a garantir que o software seja utilizado corretamente e a resolver problemas comuns.

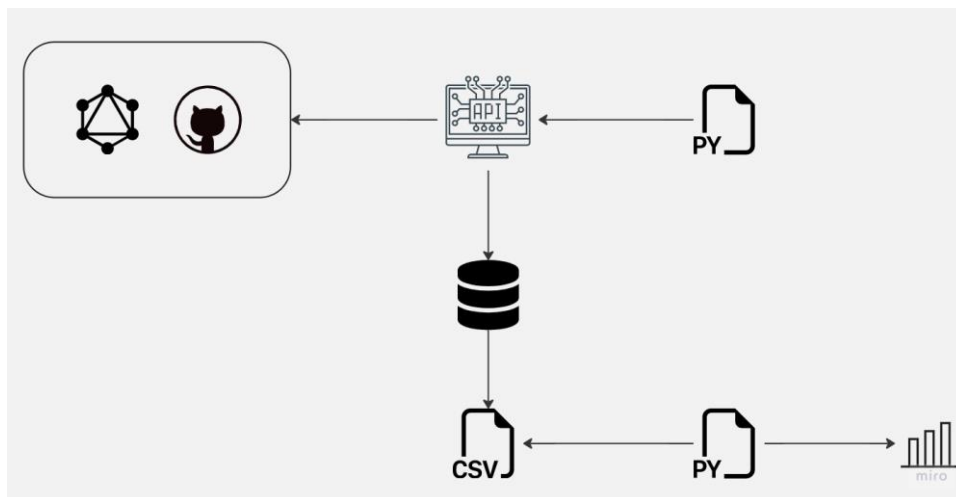
### **RQ 03. Qual a relação entre a atividade dos repositórios e as suas características de qualidade?**

Os projetos populares geralmente têm uma comunidade maior de contribuidores, revisores e usuários, que trabalham juntos para melhorar a qualidade do código. Tendem a ser bem mantidos, com atualizações frequentes e correções de bugs.

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

As características de qualidade de software são mais influenciadas por fatores como a qualidade do código, a cobertura de testes, a documentação, a colaboração e a manutenção do projeto, que não dependem diretamente do tamanho do repositório.

### 3. Metodologia



**Figura 1- Fluxograma da Metodologia**

Este estudo selecionou os 1000 repositórios Java mais populares do *GitHub* com base na medida de popularidade pelo número de estrelas. Para a coleta de dados, foi utilizada a *API GraphQL* do *GitHub* e um *script* Python, que permitiu a obtenção de informações como data de criação, número de releases, nome do repositório e URL. Esses dados foram salvos em um arquivo CSV para posterior análise.

A fim de analisar o código de cada um dos 1000 repositórios, foi realizada uma clonagem usando a URL fornecida e a ferramenta CK (<https://github.com/mauricioaniche/ck>) foi executada para a obtenção de métricas a nível de classe. Os dados obtidos pelo CK foram sumarizados para fins de análise e a sumarização seguiu os seguintes critérios: LCOM pela mediana, CBO pela mediana, LOC pelo somatório e DIT pelo maior valor em cada repositório. O processo de análise de código foi automatizado por meio de um script Python, e a execução foi dividida em quatro partes de 250 e foi realizada no *GitHub Actions*.

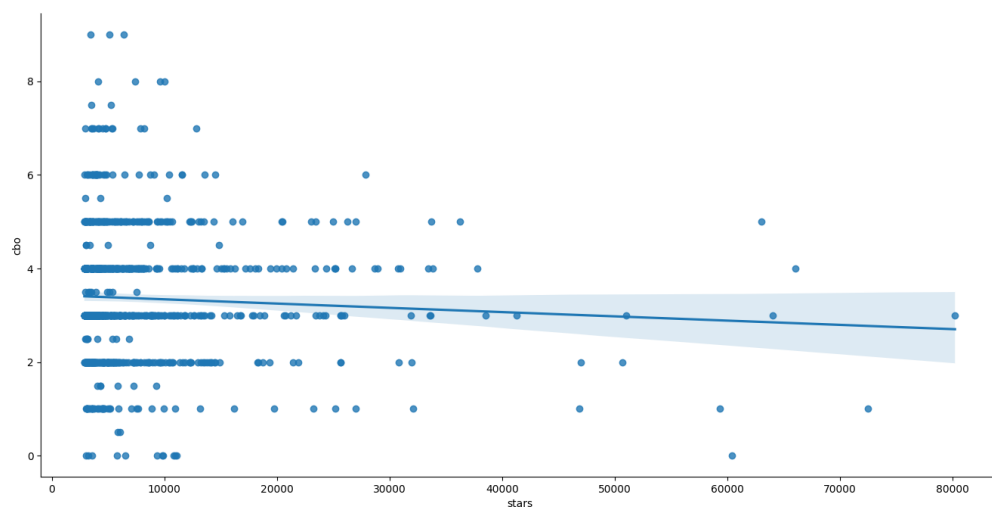
Os dados coletados foram submetidos a uma análise de correlação simples entre os dados de LOC com CBO, DIT e LCOM, estrelas com CBO, DIT e LCOM, idade com CBO, DIT e LCOM e *releases* com CBO, DIT e LCOM. Para a análise dos resultados obtidos, os dados que continham valores vazios de LOC, CBO, DIT e LCOM e os repositórios em que a ferramenta CK não pôde ser executada foram removidos. No entanto, o motivo do CK não conseguir executar em alguns repositórios não foi descoberto.

Os resultados finais foram utilizados para gerar *scatterplots*, matriz de *scatterplots*, matriz de correlação com testes de correlação de *Spearman* e *heatmap* para cada requisito, utilizando um script Python que utilizou as bibliotecas *Pandas* e *Matplotlib*.

## 4. Resultados

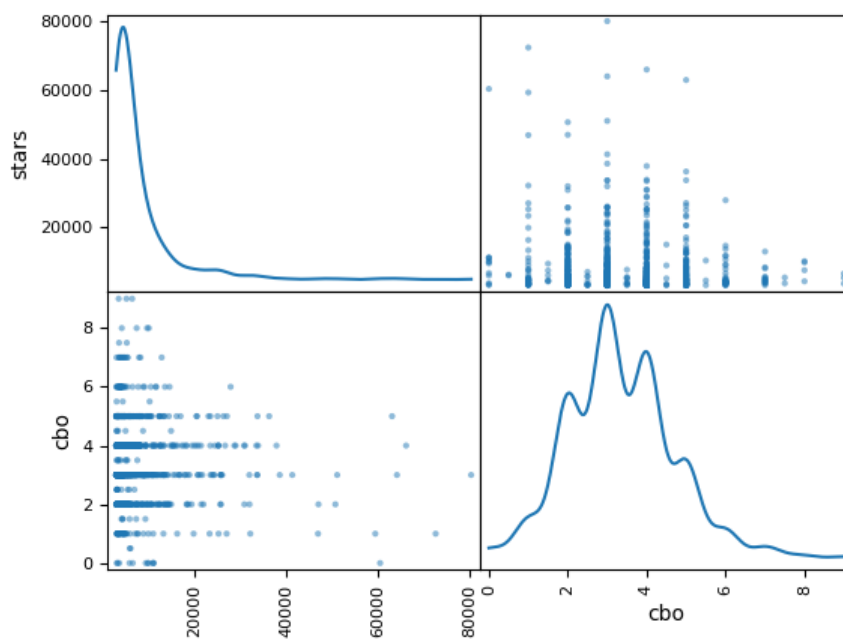
Este tópico descreve os resultados dos gráficos obtidos através da metodologia apresentada.

### 4.1 Popularidade



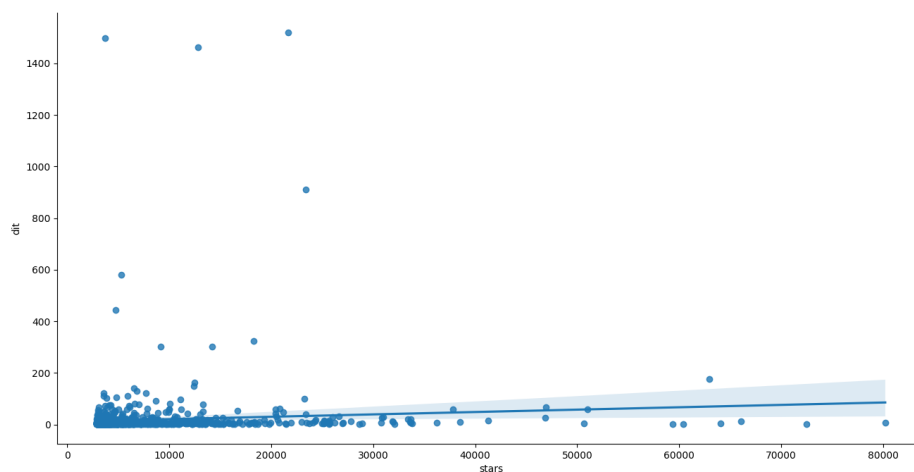
**Figura 2 – Popularidade dos repositórios correlacionando com a métrica CBO.**

Na figura 2, pode-se observar que existe uma fraca correlação entre a popularidade e o acoplamento entre os objetos (CBO), identificou-se o coeficiente de correlação igual a -0,053872.



**Figura 3 – Matriz com os dados de Popularidade dos repositórios em relação a métrica CBO.**

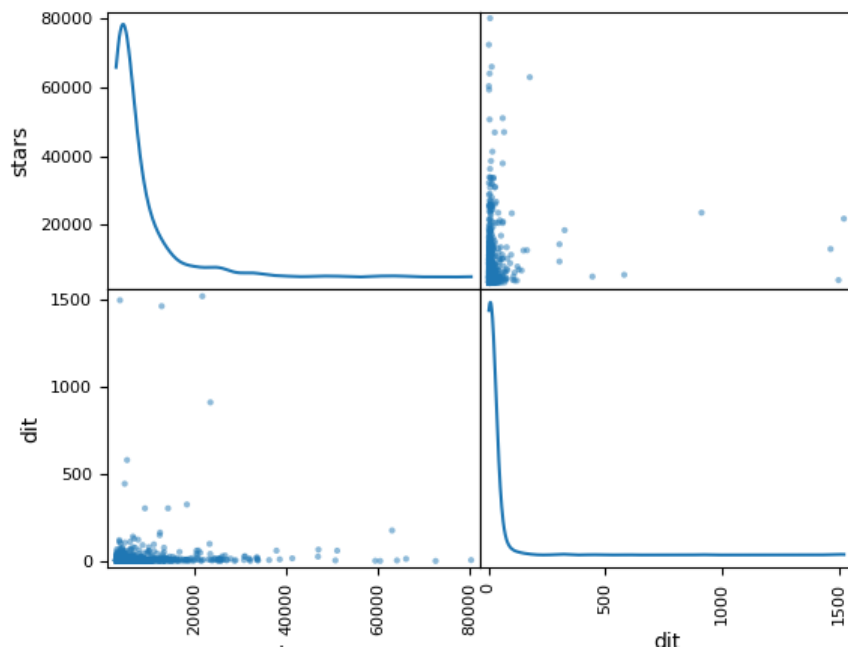
Através da reta traçada no gráfico, nota-se uma leve inclinação negativa, mesmo não tendo uma forte relação entre as variáveis, dentre os dados analisados, existem mais repositórios indicando um menor acoplamento. Como demonstrado na figura 3, em que existe maior frequência nos repositórios que tem o CBO com o valor igual a 3.



**Figura 4 – Popularidade dos repositórios correlacionando com a métrica DIT.**

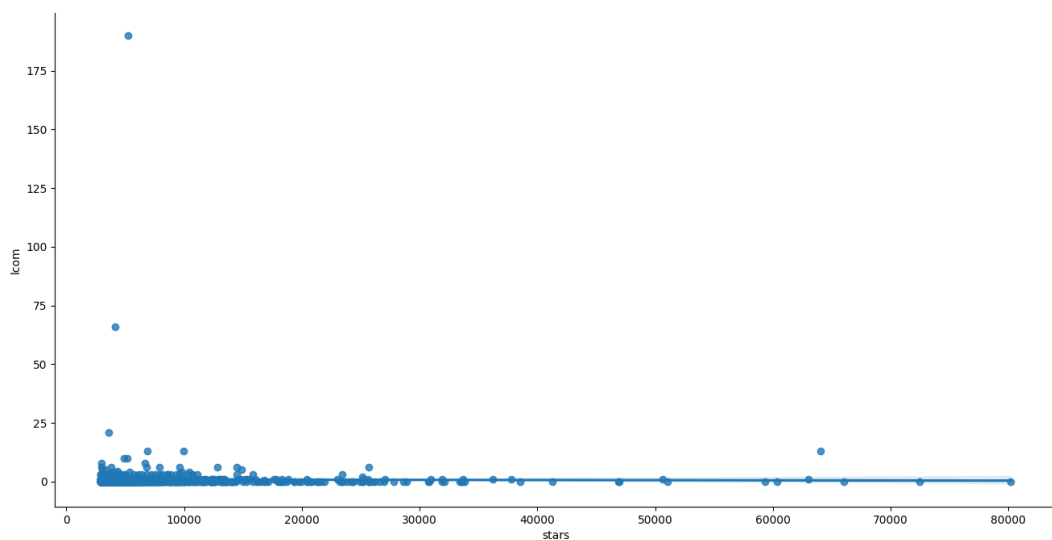
Na figura 4, existe uma fraca relação entre a popularidade dos repositórios e a métrica de DIT. O coeficiente de correlação é igual a 0,081. A partir dos dados analisados, é possível

concluir que a popularidade não influencia no número de níveis que uma classe herda métodos e atributos.



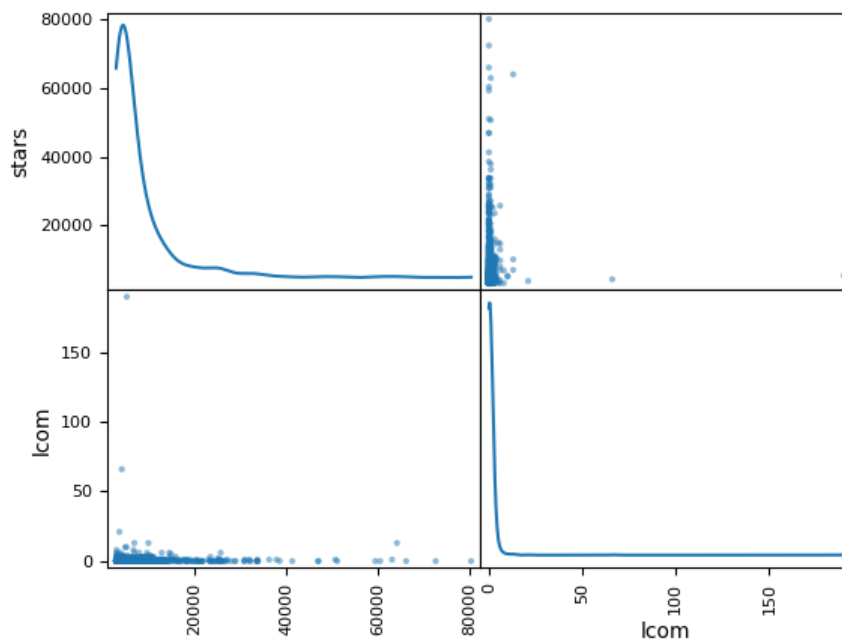
**Figura 5 – Matriz com os dados de Popularidade dos repositórios em relação a métrica DIT.**

Conforme demonstrado na figura 5, poucos repositórios possuem um número elevado na métrica de profundidade de herança.



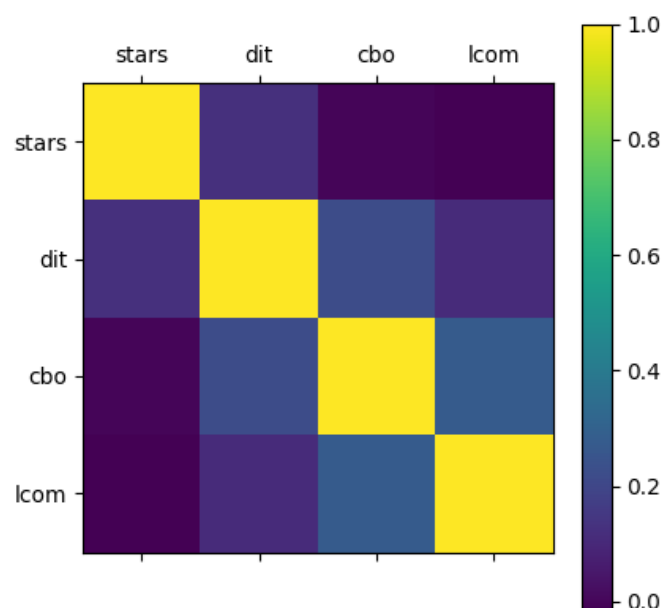
**Figura 6 – Popularidade dos repositórios correlacionando com a métrica LCOM.**

Na figura 6, existe uma fraca relação entre a popularidade dos repositórios e a métrica de LCOM. O coeficiente de correlação é igual a 0,007743. A partir dos dados analisados, é possível concluir que a popularidade não exerce influência sobre a falta de coesão.



**Figura 7 – Matriz com os dados de Popularidade dos repositórios em relação a métrica LCOM**

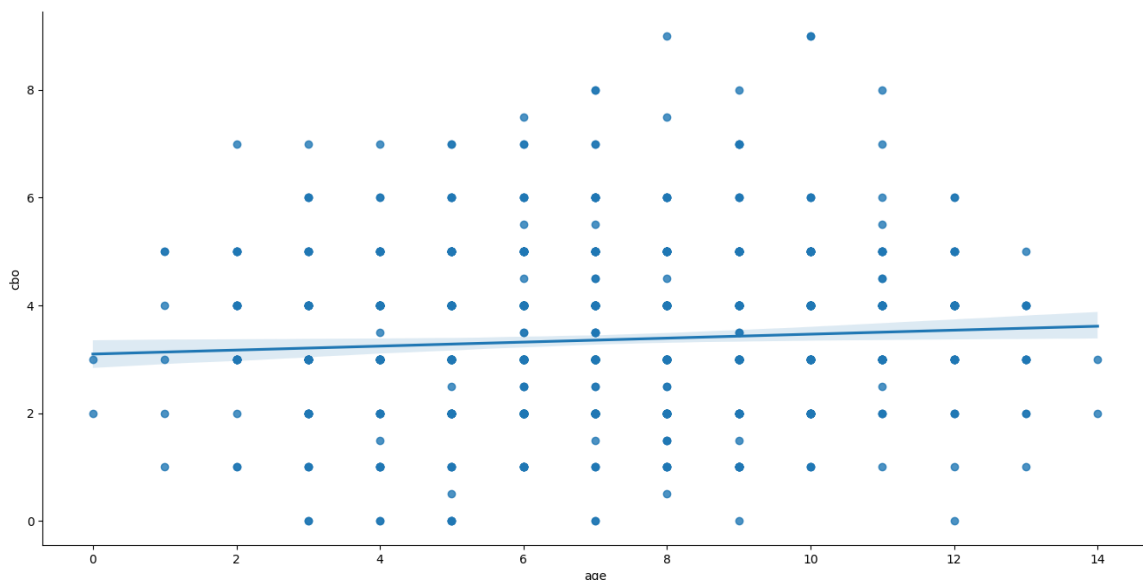
Conforme mostra a afigura 7, a maior parte dos dados coletados possui LCOM próximo de zero.



**Figura 8 – Heatmap relacionando as métricas de qualidade com a popularidade dos repositórios**

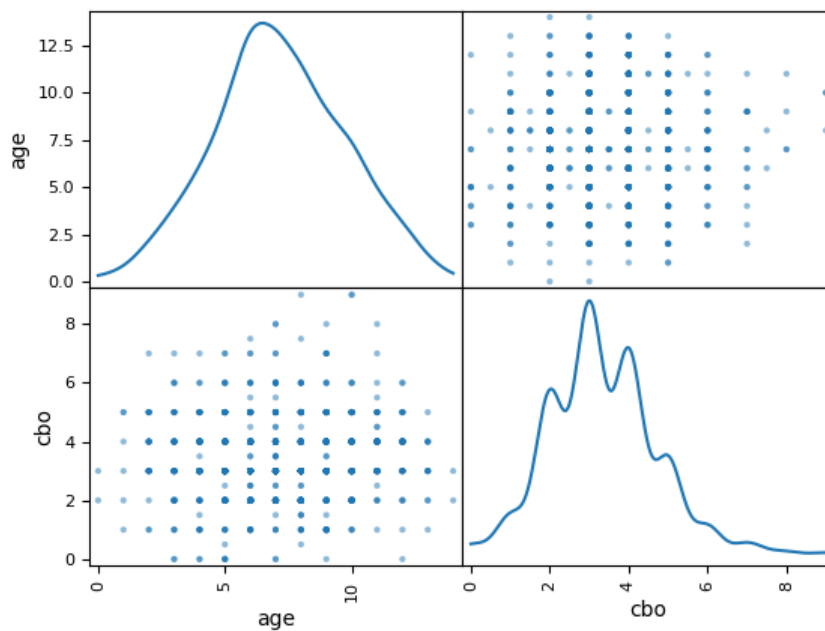
Através da figura 8, observa-se que os dados coletados referente a popularidade e as métricas de qualidade, representam uma fraca relação, sendo o coeficiente de correlação sempre tendendo a um valor próximo 0.

## 4.2 Maturidade

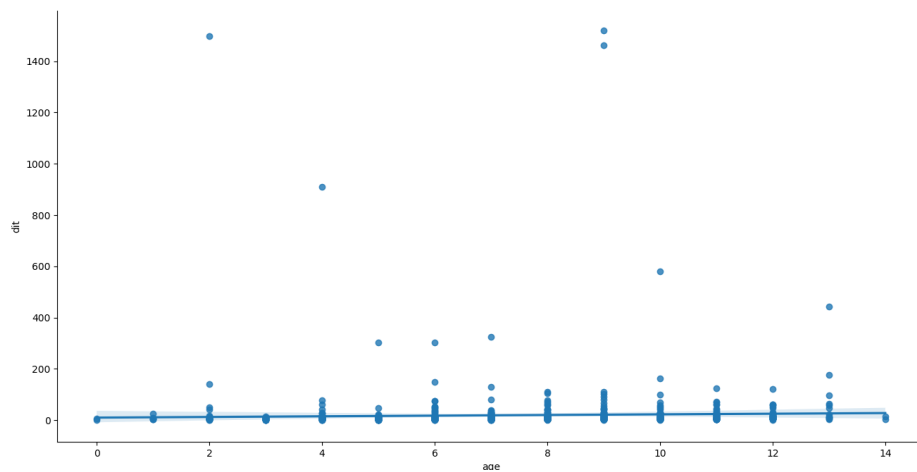


**Figura 9 – Idade dos repositórios correlacionando com a métrica CBO**

Na figura 9, pode-se destacar que existe uma fraca correlação entre a idade dos repositórios e o acoplamento entre os objetos (CBO), identificou-se o coeficiente de correlação igual a 0,070689.



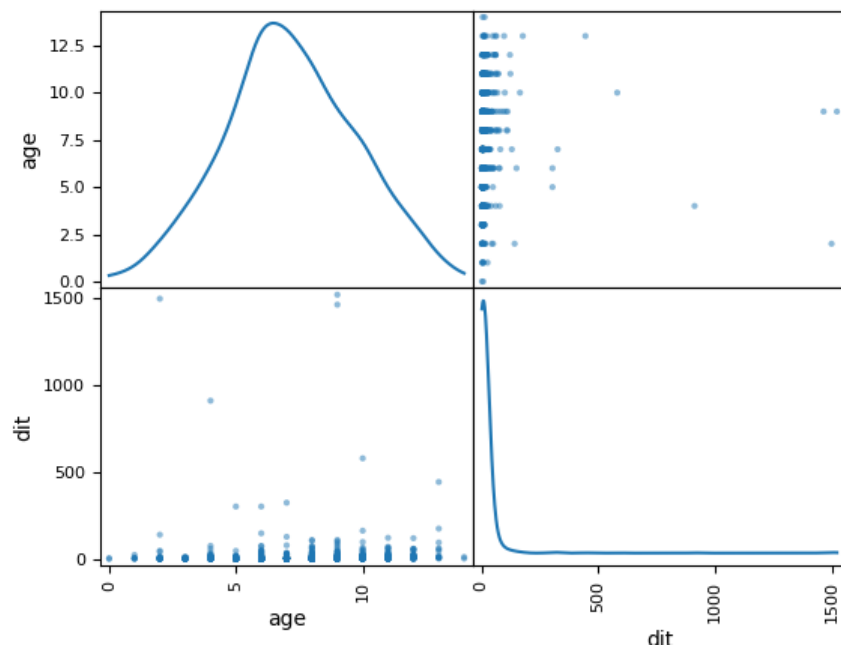
**Figura 10 – Matriz com os dados de idade do repositório em relação as métricas de CBO**



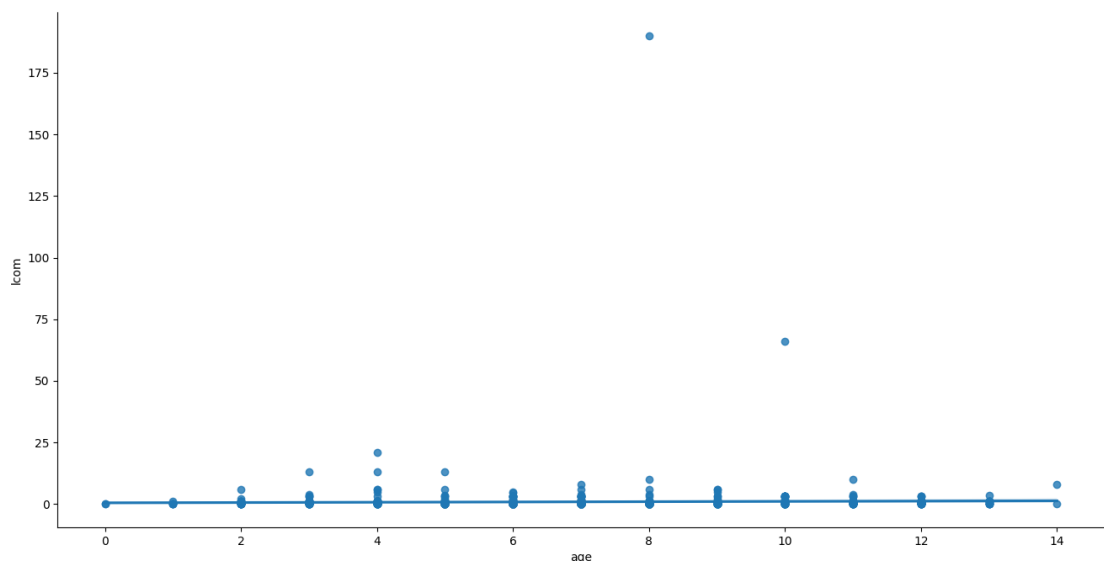
**Figura 11– Idade dos repositórios correlacionando com a métrica DIT**

Na figura 11, pode-se destacar que existe uma fraca correlação entre a idade dos repositórios e a métrica DIT, identificou-se o coeficiente de correlação igual a 0,276884. A partir dos dados analisados, é possível concluir que a popularidade não influencia no número de níveis que uma classe herda métodos e atributos.



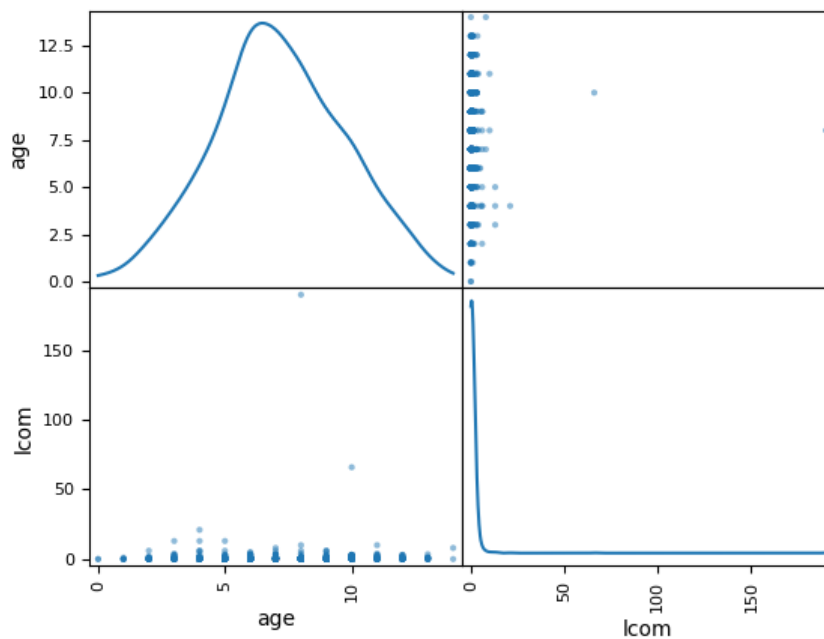


**Figura 12 – Matriz com os dados de idade do repositório em relação as métricas de DIT**

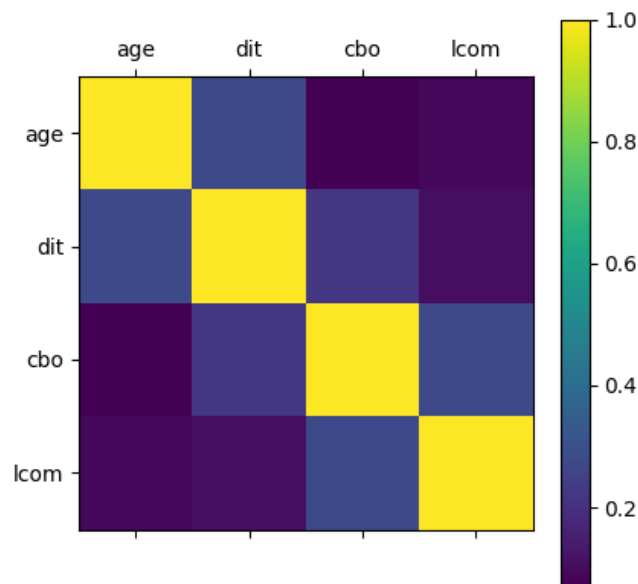


**Figura 13 – Idade dos repositórios correlacionando com a métrica LCOM**

Na figura 13, existe uma fraca relação entre a idade dos repositórios e a métrica de LCOM. O coeficiente de correlação é igual a 0,092109. A partir dos dados analisados, é possível concluir que a idade não exerce influência sobre a falta de coesão.



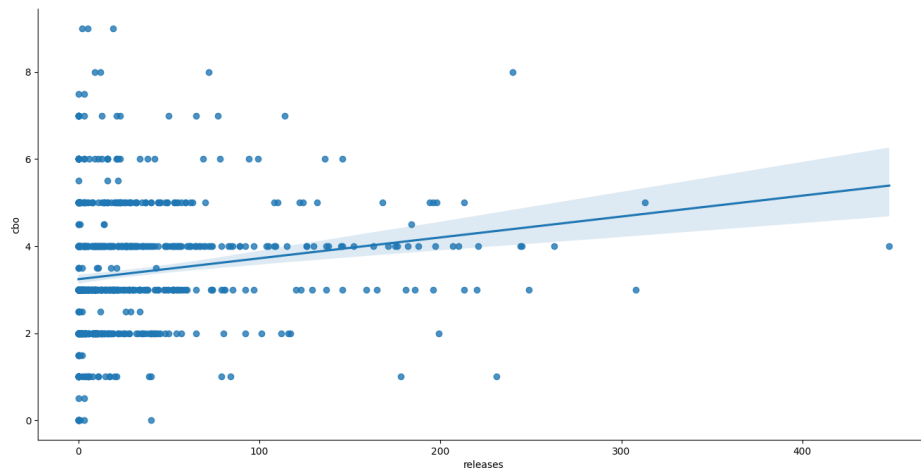
**Figura 14 – Matriz com os dados de idade do repositório em relação as métricas de LCOM**



**Figura 15 – Heatmap relacionando as métricas de qualidade com a idade dos repositórios**

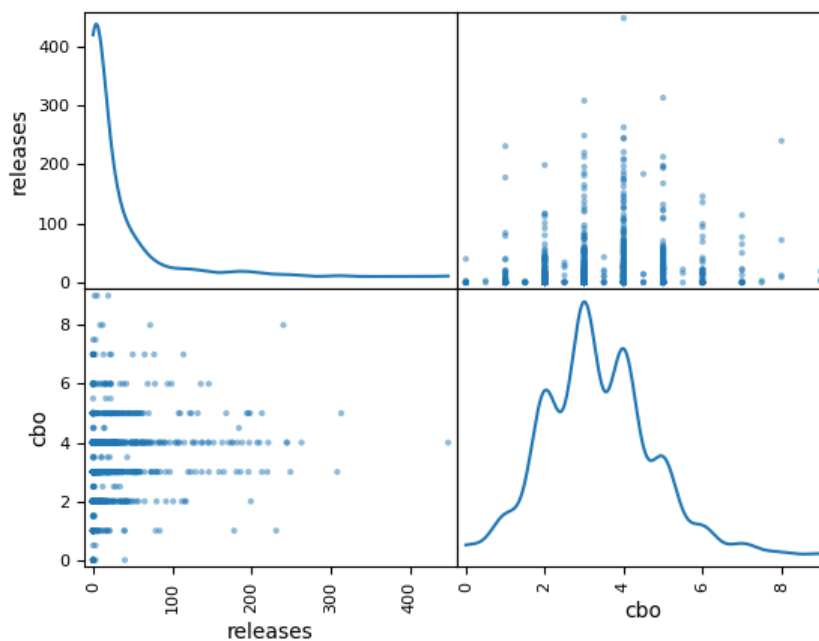
Através da figura 15, observa-se que os dados coletados referente a idade e as métricas de qualidade, representam uma fraca relação, sendo o coeficiente de correlação sempre tendendo a um valor próximo 0.

### 4.3 Atividade (*releases*)

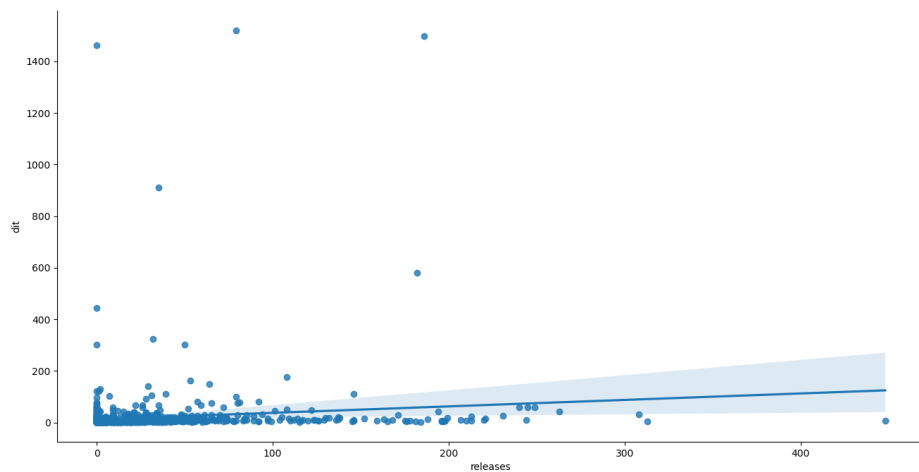


**Figura 16 – Número de *releases* dos repositórios correlacionando com a métrica CBO**

Na figura 16, pode-se observar que existe uma fraca correlação entre número de *releases* e o acoplamento entre os objetos (CBO), identificou-se o coeficiente de correlação igual a 0,230895.

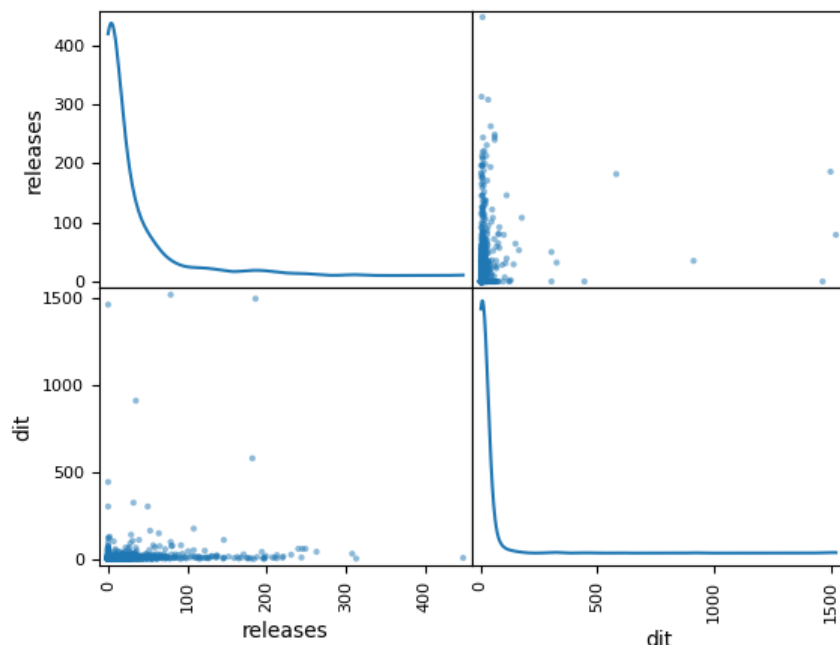


**Figura 17– Matriz com os dados de *releases* do repositório em relação as métricas de CBO**

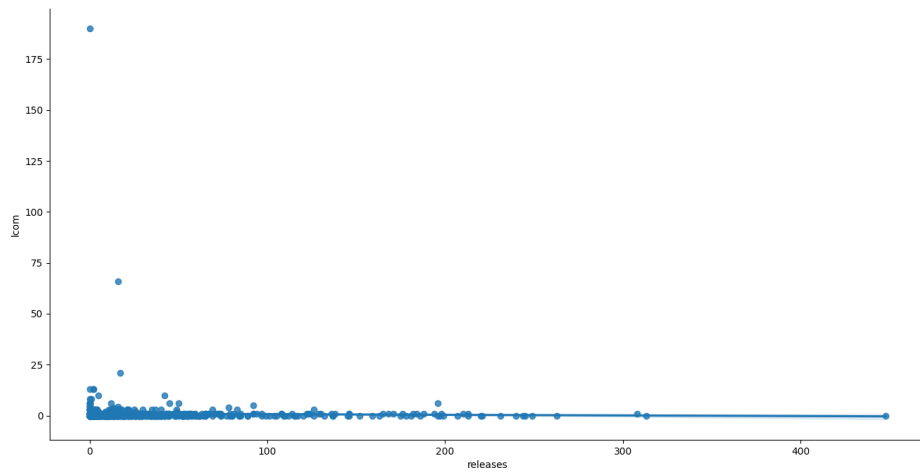


**Figura 18– Número de *releases* dos repositórios correlacionando com a métrica DIT**

Na figura 18, existe uma fraca relação entre o número de *releases* e a métrica de DIT. O coeficiente de correlação é igual a 0,363293. A partir dos dados analisados, é possível concluir que o número de *releases* não influencia no número de níveis que uma classe herda métodos e atributos.

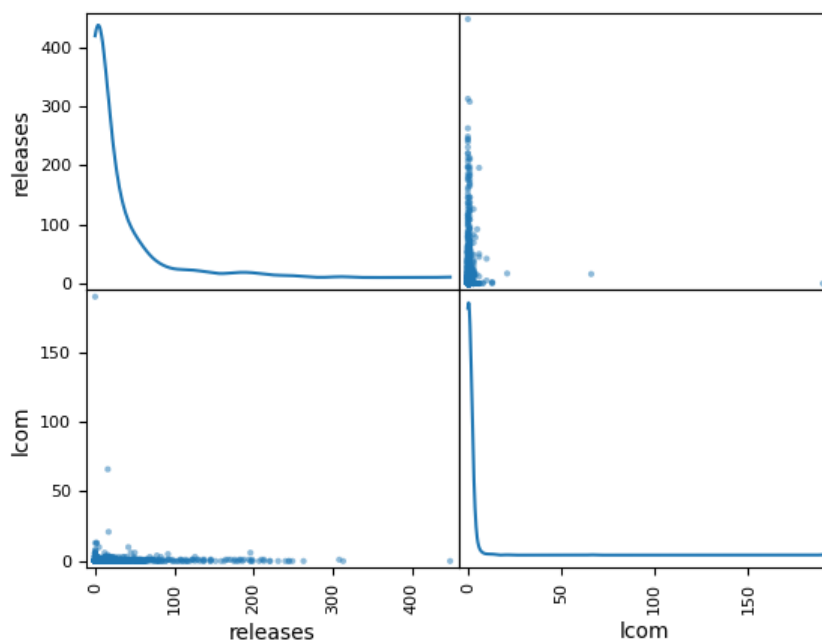


**Figura 19– Matriz com os dados de *releases* do repositório em relação as métricas de DIT**

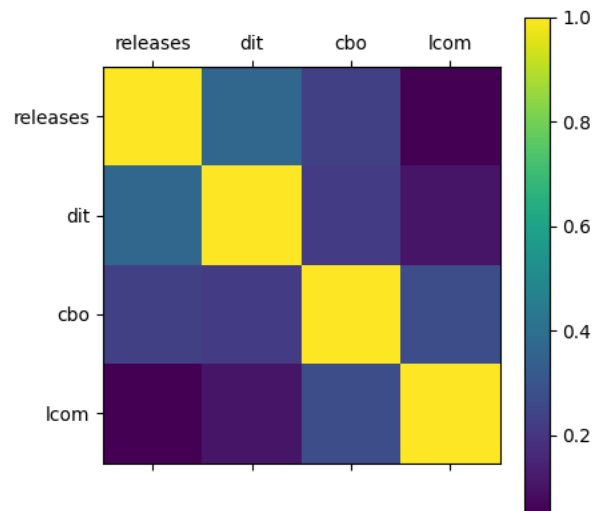


**Figura 20– Número de *releases* dos repositórios correlacionando com a métrica LCOM**

Na figura 20, existe uma fraca relação entre o número de *releases* dos repositórios e a métrica de LCOM. O coeficiente de correlação é igual a 0,051887. A partir dos dados analisados, é possível observar que a popularidade não exerce influência sobre a falta de coesão.



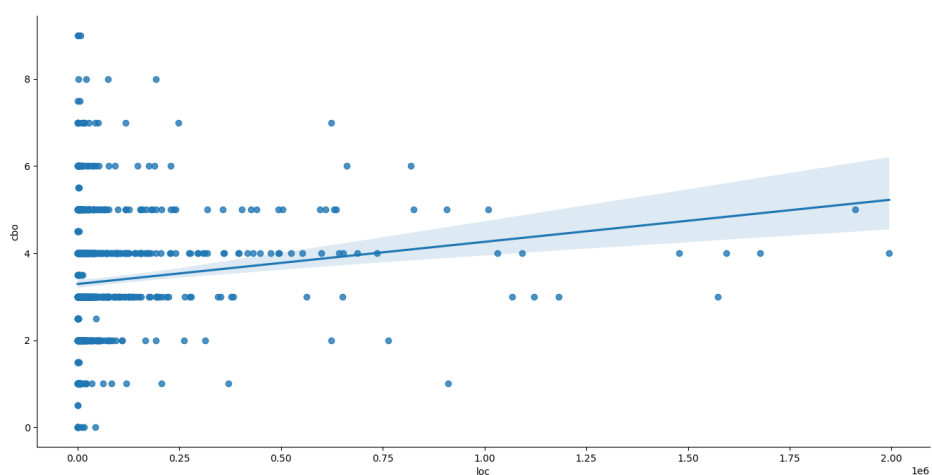
**Figura 21– Matriz com os dados de *releases* do repositório em relação as métricas de LCOM**



**Figura 22 – Heatmap relacionando as métricas de qualidade com o número de *releases* dos repositórios**

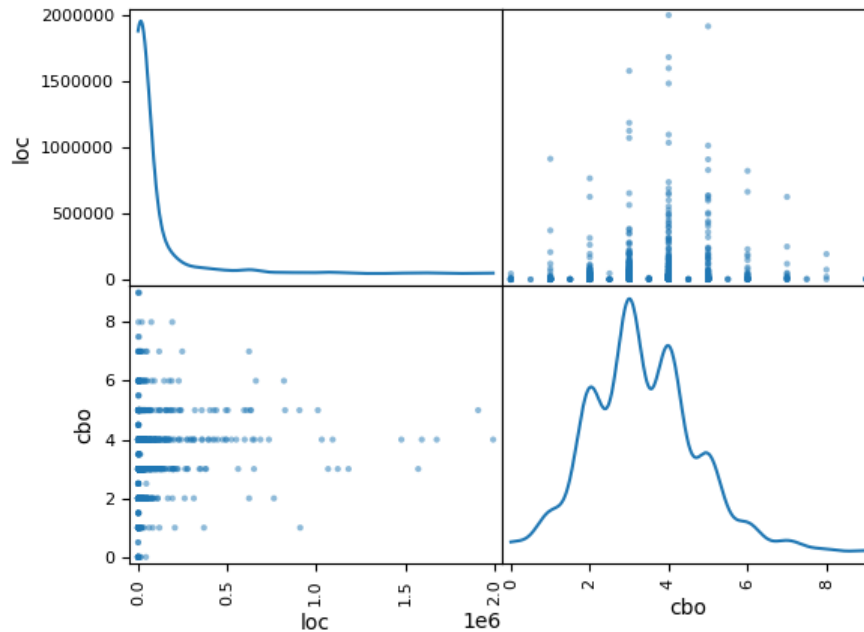
Através da figura 22, observa-se que os dados coletados referente ao número de *releases* e as métricas de qualidade, representam uma fraca relação, sendo o coeficiente de correlação sempre tendendo a um valor próximo 0.

#### 4.4 Tamanho

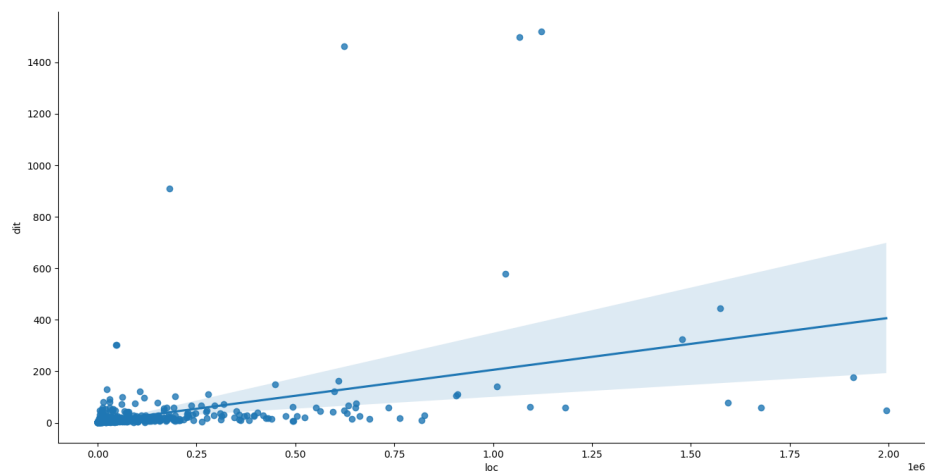


**Figura 23 – Tamanho dos repositórios correlacionando com a métrica CBO**

Na figura 23, pode-se destacar que existe uma fraca correlação entre o tamanho dos repositórios e o acoplamento entre os objetos (CBO), identificou-se o coeficiente de correlação igual a 0,241092.

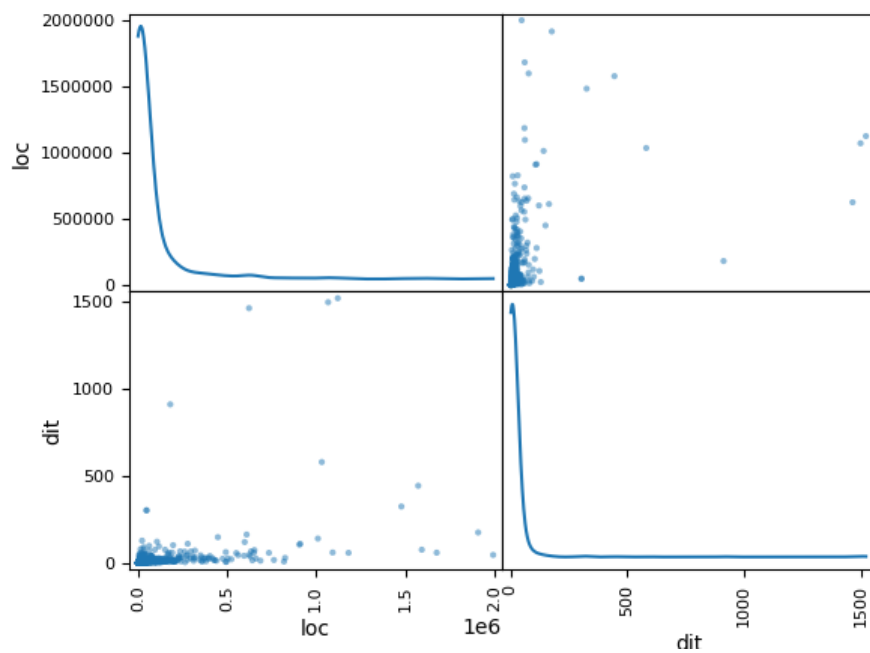


**Figura 24 – Matriz com os dados de tamanho do repositório em relação as métricas de CBO**

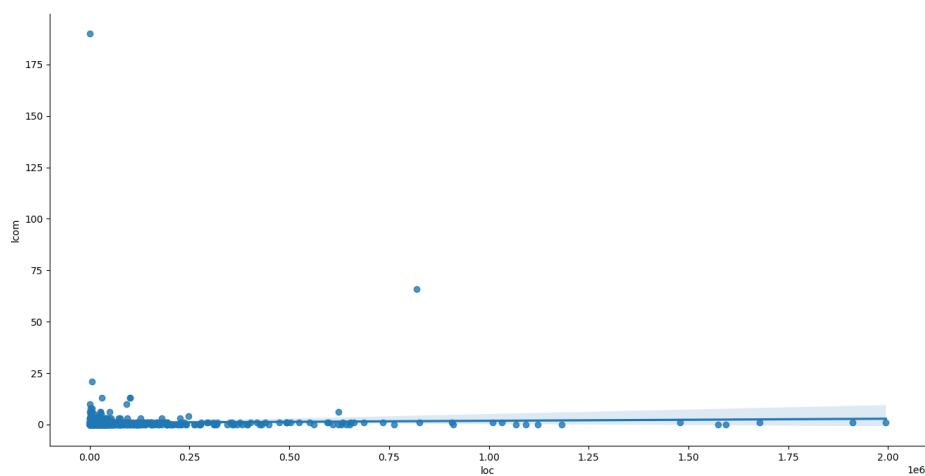


**Figura 25 – Tamanho dos repositórios correlacionando com a métrica DIT**

Pode-se destacar na figura 25, que existe uma fraca correlação entre o tamanho dos repositórios e a métrica DIT, identificou-se o coeficiente de correlação igual a 0,799229. A partir dos dados analisados, é possível concluir que a popularidade não influencia no número de níveis que uma classe herda métodos e atributos.



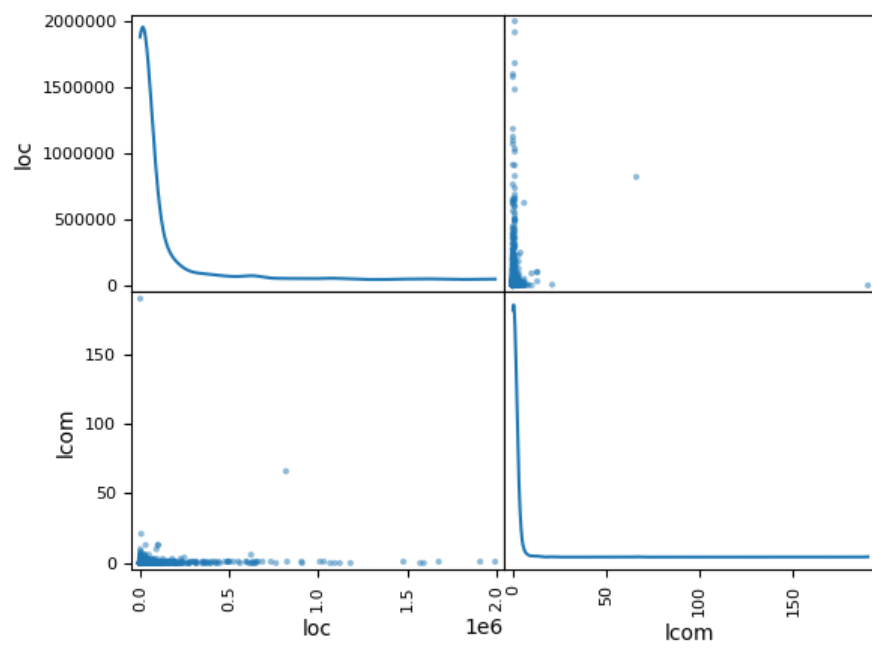
**Figura 26 – Matriz com os dados de tamanho do repositório em relação as métricas de DIT**



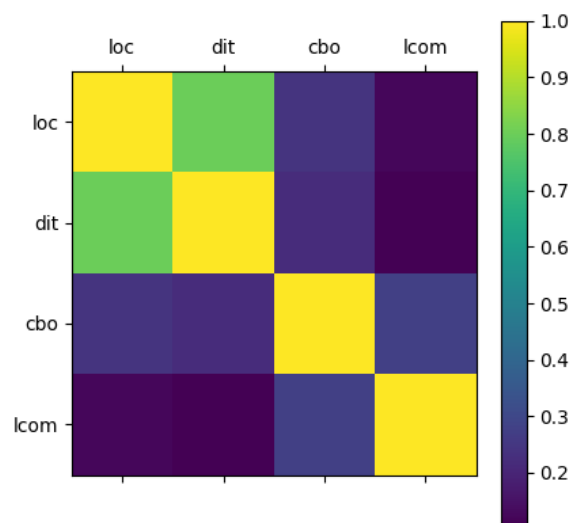
**Figura 27 – Tamanho dos repositórios correlacionando com a métrica LCOM**

Existe uma fraca relação entre o tamanho dos repositórios e a métrica de LCOM. O coeficiente de correlação é igual a 0,122350. A partir dos dados analisados, é possível concluir que a popularidade não exerce influência sobre a falta de coesão.





**Figura 28 – Matriz com os dados de tamanho do repositório em relação as métricas de LCOM**



**Figura 29 – Heatmap relacionando as métricas de qualidade com o tamanho dos repositórios**

Nota-se que os dados coletados referente ao tamanho dos repositórios e as métricas de qualidade, representam uma fraca relação, sendo o coeficiente de correlação sempre tendendo a um valor próximo 0.

## 5. Discussão

### 1. Qual a relação entre a popularidade dos repositórios e as suas características de qualidade?

O objetivo desta questão foi investigar a relação entre a popularidade de um repositório e suas características de qualidade. Para tanto, foram utilizados os parâmetros de qualidade DIT, CBO e LCOM, e como parâmetro de popularidade, o número de estrelas no *GitHub*. A hipótese inicial era de que um repositório popular teria características de qualidade elevadas.

Os resultados obtidos pela matriz de correlação não confirmaram totalmente a hipótese inicial. As correlações entre o número de estrelas e as métricas de qualidade foram relativamente baixas, e em especial, o parâmetro LCOM apresentou correlação negativa. Esses resultados indicam que não se pode afirmar que a popularidade influencia na qualidade ou vice-versa.

Apesar de não ter confirmado a hipótese inicial, é possível que um grande número de contribuidores esteja relacionado com outros aspectos da qualidade do software que não foram medidos neste estudo. Por exemplo, pode haver uma correlação positiva entre a popularidade e a confiabilidade do software, ou com sua capacidade de resposta às solicitações de suporte.

Os resultados também podem ser interpretados como indicando que outros fatores, além da popularidade, influenciam na qualidade do software. Além disso, o fato de que os parâmetros de qualidade estão mais correlacionados entre si sugere que há uma interação complexa entre diferentes aspectos da qualidade do software, que não pode ser facilmente explicada por uma única variável.

Portanto, não é possível estabelecer uma relação entre a popularidade de um repositório com suas características de qualidade.

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

Esta pergunta buscou descobrir qual o impacto que a idade de um repositório tem em suas características de qualidade. Para isso, foi utilizado o parâmetro de idade do repositório, em anos, calculado a partir da sua data de criação até o dia da coleta de dados, e foram utilizados os parâmetros de qualidade DIT, CBO e LCOM.

Com a análise dos dados de correlação e a distribuição das métricas de qualidade em relação à idade, é possível constatar que há uma correlação moderada entre a idade e o DIT. Isso indica que, conforme a idade de um repositório aumenta, o seu DIT também aumenta. Porém, como se trata de uma correlação moderada de aproximadamente 0,28, não é possível estabelecer que somente a idade afeta o aumento do DIT, podendo estar relacionado a outros fatores. Já as correlações entre idade e CBO e LCOM são muito próximas a zero, o que indica que não se pode apontar uma relação com a idade nessas métricas.

A hipótese inicial não se confirmou, uma vez que a tendência do DIT é aumentar, fazendo assim um pequeno declínio na qualidade. As demais características de qualidade não apresentaram relação com a idade, não sendo possível afirmar que a idade de um repositório afeta positiva ou negativamente a qualidade do código desse repositório.

Consequentemente, conclui-se que mais pesquisas são necessárias para entender completamente a relação entre a idade dos repositórios e suas características de qualidade, além de considerar outras métricas de qualidade e fatores que possam influenciar essas relações.

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

O propósito desta questão era investigar a relação entre a quantidade de *releases* publicadas de um repositório e suas características de qualidade. A hipótese inicial propôs que um repositório que publica várias versões tende a ter uma melhora em sua qualidade.

Considerando a análise da correlação de *Spearman* entre o número de *releases* de um repositório do *GitHub* e suas métricas de qualidade DIT, CBO e LCOM, é possível inferir que há uma relação moderada entre a atividade do repositório medida pelo número de *releases* e a métrica de qualidade DIT. Isso indica que quanto maior a atividade do repositório, maior é o valor de DIT, podendo indicar um aumento na complexidade do código, fazendo um contraponto com a hipótese inicial.

No entanto, não foi identificada uma relação significativa entre a atividade do repositório e as outras métricas de qualidade analisadas, como CBO e LCOM. Dessa forma, é preciso levar em consideração que outros fatores também podem influenciar a qualidade do código, além da atividade do repositório.

Assim, conclui-se que a atividade do repositório, medida pelo número de *releases*, pode ter uma relação moderada com a qualidade do código medida pelo DIT, mas não é um indicador significativo para as outras métricas de qualidade analisadas.

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

A questão 4 busca relacionar o tamanho de um repositório com sua qualidade. Para alcançar esse objetivo, foi utilizado o tamanho total do repositório em linhas de código e investigado sua relação com DIT, CBO e LCOM. A hipótese inicial era de que o tamanho não seria o fator mais relevante na qualidade de um repositório.

Com base na matriz de correlação de *Spearman* apresentada, é possível inferir que há uma forte relação positiva entre o tamanho do repositório, medido em linhas de código, e a métrica de qualidade DIT. Isso indica que quanto maior o tamanho do código, maior é a sua complexidade. Além disso, há uma correlação moderada entre o tamanho do repositório e a métrica de qualidade CBO. Porém, não foram identificadas relações significativas entre o tamanho do repositório e a métrica de qualidade LCOM.

A métrica DIT também apresenta forte correlação positiva com a métrica de qualidade LOC, indicando que quanto maior a complexidade do código, maior é o seu tamanho. As correlações entre DIT e CBO, bem como entre DIT e LCOM, são moderadas e fracas, respectivamente.

Em relação à métrica CBO, esta apresenta correlação moderada com o tamanho do código e com a complexidade medida por DIT, indicando uma relação entre a quantidade de classes e objetos e a complexidade do código. Já a correlação entre CBO e LCOM é fraca, indicando pouca relação entre a coesão entre os métodos e as demais características do código. Em resumo, a métrica CBO apresenta correlação moderada com algumas métricas de qualidade analisadas.

Em contraponto com a hipótese inicial, observou-se uma tendência do aumento da complexidade do código conforme o tamanho do código também aumenta. Como o CBO e LCOM também têm um aumento moderado conforme o aumento do tamanho, pode-se então afirmar que há uma tendência leve em abaixar a qualidade de um código conforme aumenta seu tamanho, levando em consideração que há outros fatores que não foram observados nessa pesquisa que têm maior influência na qualidade.

Conclui-se que o tamanho do repositório, medido em linhas de código, e a métrica de qualidade DIT apresentam correlações positivas significativas com as demais métricas de qualidade analisadas, enquanto a métrica CBO apresenta correlação moderada com algumas métricas e a métrica LCOM apresenta pouca correlação com as demais características do código.

