Características de Qualidade em repositórios de linguagem Java:

Grupo 01:
Nataniel Geraldo Mendes Peixoto
Nelson de Campos Nolasco
Rubia Coelho de Matos

Introdução

No processo de desenvolvimento de sistemas *open-source*, em que diversos desenvolvedores contribuem em partes diferentes do código, um dos riscos a serem gerenciados diz respeito à evolução dos seus atributos de qualidade interna. Isto é, ao se adotar uma abordagem colaborativa, corre-se o risco de tornar vulnerável aspectos como modularidade, manutenibilidade, ou legibilidade do software produzido.

Neste trabalho, analisamos 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. As seguintes hipóteses de partida foram consideradas:

1. Repositórios populares tem melhores atributos de qualidade, tendo em vista que atraem mais colaboradores, revisões e boas práticas de desenvolvimento.
2. Repositórios com maior maturidade têm melhores atributos de qualidade, tendo em vista que passaram por mais revisões, refatorações e melhorias ao longo do tempo
3. Repositórios com mais atividade (releases) têm atributos de qualidade medianos, tendo em vista que passam por atualizações frequentes, o que pode impactar a coesão e o acoplamento do código.
4. Repositórios com mais linhas de código (LOC) têm atributos de qualidade mediano, considerando que o aumento do tamanho do código pode levar a uma maior complexidade, impactando o acoplamento, a herança e a coesão.

Nosso objetivo é analisar um conjunto de 1000 repositórios e validar essas hipóteses com base em métricas extraídas da *API GraphQL* do *GitHub*.

Metodologia

1. **Coleta de Dados**: Utilizamos a *API GraphQL* do *GitHub* para coletar dados detalhados sobre 1000 repositórios populares na linguagem Java, incluindo informações sobre estrelas, releases e data de criação.

2. **Análise de Métricas**:
    - Analisamos a popularidade dos repositórios com base no número de estrelas.
    - Analisamos a maturidade dos repositórios com base em sua data de criação.
    - Analisamos o tamanho dos repositórios com base no número de linhas de código (LOC) e comentários.
    - Analisamos o grau de interação com repositórios por meio do número de releases.
3. **Visualizações**:
    - Estatística descritiva das variáveis em estudo.
    - Histogramas e boxplot das variáveis em estudo.
    - Matriz de correlação entre as variáveis em estudo.
    - Gráficos de dispersão entre métricas de processo (popularidade, tamanho, atividade e maturidade) e métricas de qualidade (CBO, DIT e LCOM).
    - Teste de correlação de *Spearman* e *Pearson*.

Resultados

➔ Estatística Descritiva:

A tabela abaixo evidencia a estatística descritiva das métricas utilizadas no estudo:

	CBO	DIT	LCOM	LOC	Comments	Maturity	Release	Stars
count	979	979	979	979	979	979	979	979
mean	5	1	115	180.318	60.372	9	14	9.281
std	2	1	1.797	897.840	437.577	3	13	11.052
min	0	0	0	0	0	0	0	3.301
25%	4	1	8	4.086	786	7	0	4.267
50%	5	1	22	21.252	5.120	9	10	5.646
75%	6	2	48	93.512	26.190	11	30	9.638
max	22	4	55.991	23.206.140	12.489.032	16	30	148.824

A seguir, apresentamos algumas análises selecionadas a partir dos resultados da estatística descritiva, dentre as várias possíveis interpretações.

- A análise foi realizada com 979 repositórios válidos, conforme indicado na linha *'count'* da estatística descritiva. Esse número representa a quantidade de repositórios que atenderam aos critérios de filtragem e limpeza dos dados.
- A média do CBO (*Coupling Between Objects*) foi de 5, com valores variando de 0 a 22. Isso sugere que, em média, as classes dos repositórios analisados possuem um nível de acoplamento baixo, o que pode facilitar a manutenção e os testes do código. No entanto, repositórios com CBO próximos de 22 indicam um alto acoplamento, o que pode dificultar a manutenção e aumentar a complexidade do código.
- A média do DIT (*Depth of Inheritance Tree*) foi de 4, com valores variando de 0 a 4. Isso indica que, em média, os repositórios analisados possuem uma estrutura de herança moderada, o que pode contribuir para uma maior complexidade no código. Embora um DIT de 4 não seja excessivamente alto, ele sugere que as classes de alguns repositórios podem estar mais acopladas, o que pode dificultar

a manutenção e a compreensão do código em comparação com uma estrutura de menos estruturas de herança.

- A média do LCOM (*Lack of Cohesion of Methods*) foi de 115, com valores variando de 0 a 55.991. A média elevada sugere que, em geral, os repositórios têm baixa coesão, com métodos pouco relacionados. No entanto, 75% dos repositórios apresentaram valores abaixo de 48, indicando que a maioria das classes possui melhor coesão. O valor máximo significativamente elevado pode indicar a presença de outliers na amostra.
- A média de linhas de código (LOC) nos repositórios analisados é de aproximadamente 180.318, com um mínimo de 0 e um máximo de 23.206.140 (referente ao repositório *aws-sdk-java*). Isso indica que, em média, os repositórios Java possuem um volume expressivo de código. No entanto, a grande variação entre o mínimo e o máximo sugere que alguns projetos são significativamente maiores do que outros, o que pode afetar a média.
- A média de maturidade (tempo de existência do repositório) é de 9 anos, com uma mínima de 0 anos e máxima de 16 anos. Isso indica que, em média, os repositórios populares em Java têm uma existência considerável. No entanto, o valor mínimo de 0 anos sugere que a popularidade também pode ser observada em repositórios mais novos, não necessariamente os mais antigos.
- A média de releases é de 14, com uma mínima de 0 e máxima de 30. Embora isso sugira que a maioria dos repositórios tenha uma quantidade razoável de atualizações, o valor mínimo de 0 indica que alguns repositórios não receberam atualizações. Assim, a atividade dos repositórios pode variar consideravelmente;
- A média de estrelas (indicador de popularidade) dos repositórios analisados foi de aproximadamente 9.281, com um mínimo de 3.301 e um máximo de 148.824. Esses valores reforçam o próprio critério de coleta dos repositórios, que considerou a quantidade de estrelas como parâmetro de seleção. A ampla variação sugere que, embora todos sejam relativamente populares, alguns se destacam significativamente mais do que outros.

➔ Matriz de correlação:

A tabela abaixo evidencia a matriz de correlações das métricas utilizadas no estudo:

	CBO	DIT	LCOM	LOC	Comments	Maturity	Release	Stars
CBO	1	0.626071	0.064981	0.0110517	0.0167707	0.0159292	0.277802	-0.138384
DIT	0.626071	1	0.0610657	-0.0795308	-0.0490112	0.135348	0.117152	-0.147474
LCOM	0.064981	0.0610657	1	0.0291935	0.0159351	0.0279079	-0.0284641	0.018344
LOC	0.0110517	-0.0795308	0.0291935	1	0.966024	0.0877223	0.0320939	0.0718547
Comments	0.0167707	-0.0490112	0.0159351	0.966024	1	0.0776211	0.00502366	0.0293311
Maturity	0.0159292	0.135348	0.0279079	0.0877223	0.0776211	1	-0.00966642	-0.0223417
Release	0.277802	0.117152	-0.0284641	0.0320939	0.00502366	-0.00966642	1	0.0857957
Stars	-0.138384	-0.147474	0.018344	0.0718547	0.0293311	-0.0223417	0.0857957	1

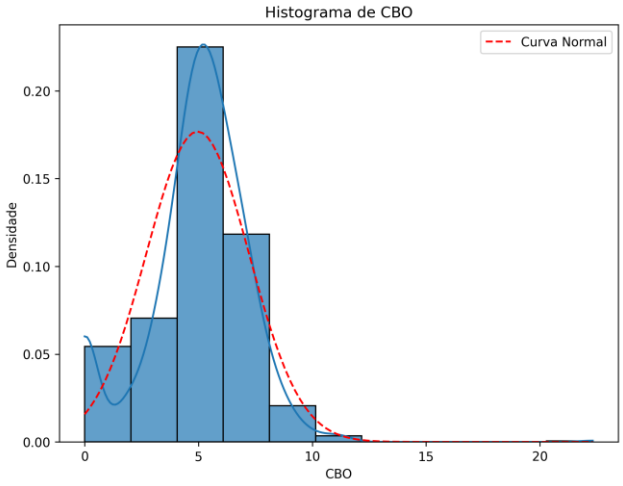
A seguir, apresentamos algumas análises selecionadas a partir dos resultados da matriz de correlações, dentre as várias possíveis interpretações.

- A correlação entre DIT e COB foi de 0,626071, indicando uma correlação moderada e positiva. Os resultados sugerem que repositórios com mais níveis de herança (maior DIT) geralmente têm também mais acoplamento entre as classes (maior CBO). Uma classe que herda de várias outras pode ter uma maior probabilidade de interagir com outras classes, aumentando o acoplamento.
- A correlação entre LOC e COMMENTS foi de 0,9966024, indicando uma correlação forte positiva. Os resultados sugerem que, à medida que o número de linhas de código (LOC) aumenta, o número de comentários também tende a aumentar. Ou seja, repositórios com mais linhas de código tendem a ter mais comentários associados.
- A correlação entre RELEASE e CBO foi de 0,277802, indicando uma correlação positiva fraca. Os resultados sugerem que repositórios com mais releases podem ter se tornado mais complexos à medida que novas funcionalidades foram adicionadas, o que tende a aumentar o CBO.

➔ Histogramas:

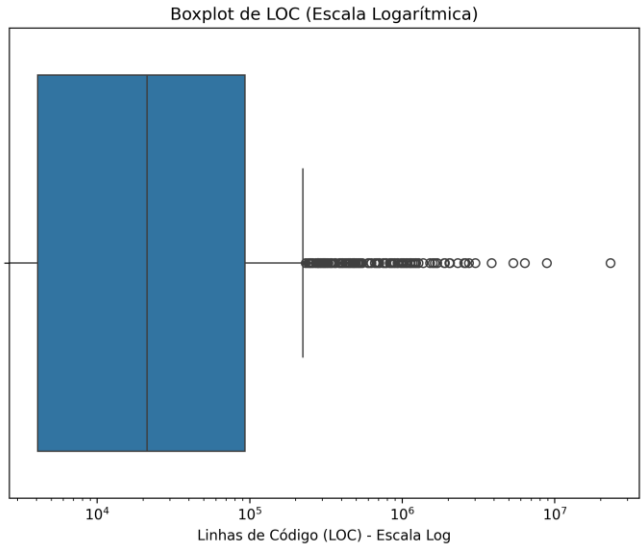
Foi elaborado o histograma para cada uma das 8 métricas em estudo, sendo: *CBO, DIT, LCOM, LOC, Comments, Maturity, Release e Stars*. A seguir, são apresentados os resultados de alguns histogramas. Os demais, embora não analisados de forma individualizada, podem ser consultados na pasta “Relatórios”.

➢ Histograma CBO:



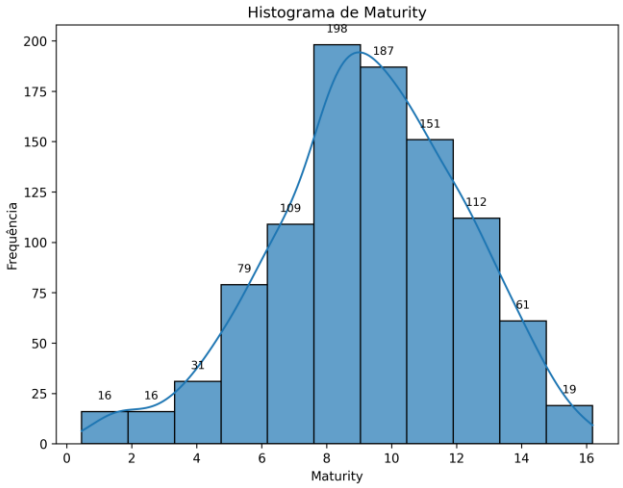
Os dados evidenciam que aproximadamente 50% dos repositórios têm média de CBO igual ou inferior a 5, tal como observado na estatística descritiva. Isso sugere que a maioria dos repositórios analisados apresenta um nível relativamente baixo de acoplamento entre suas classes, o que pode indicar uma maior modularização e facilitar a manutenção do código

➢ BoxPlot de LOC (Linhas de código):



Enquanto a maioria dos repositórios tem um número moderado de LOC, alguns repositórios têm quantidades extremas (altas ou baixas) de linhas de código, contribuindo para a grande variação observada, com expressivos *outliers*.

- Histograma de Maturidade (*Maturity*):



O histograma evidencia a distribuição dos repositórios em termos de tempo de existência. Em síntese, é possível observar que 50% dos repositórios têm aproximadamente 9 anos ou menos. Ou seja, os repositórios mais populares em Java tendem a ser mais maduros.

Questões de pesquisa:

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

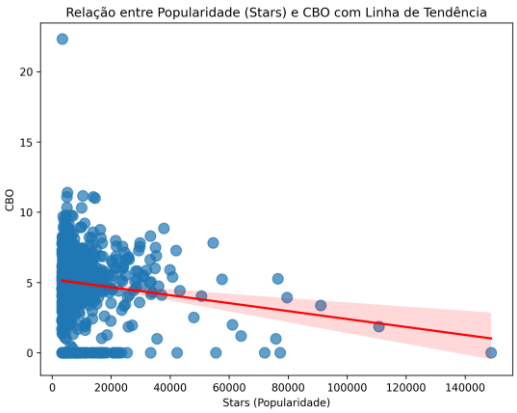
Hipótese**:** Repositórios populares tem melhores atributos de qualidade, tendo em vista que atraem mais colaboradores, revisões e boas práticas de desenvolvimento.

Resultados:

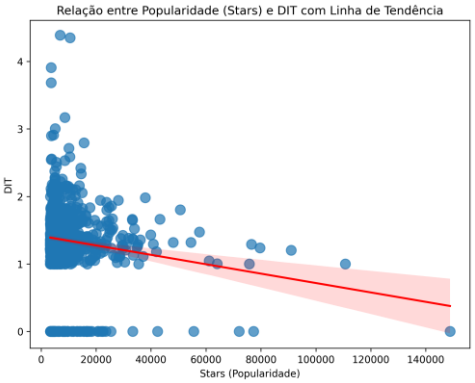
Os resultados a seguir evidenciam, por meio de gráficos de dispersão, a relação entre popularidade e atributos de qualidade. A significância estatística foi verificada pelos testes de Correlação de *Pearson* e *Spearman*:

	Metric	Pearson_Coefficient	Pearson_p-value	Spearman_Coefficient	Spearman_p-value
0	CBO	-0.138384	1.39091e-05	-0.0252824	0.429424
1	DIT	-0.147474	3.5904e-06	-0.0497261	0.119981
2	LCOM	0.018344	0.566455	0.00395571	0.901622

➔ Stars (popularidade) x CBO: Os resultados indicam que repositórios com maior acoplamento entre classes (CBO) tendem a ter menos estrelas (Stars). Esse comportamento é evidenciado pela correlação negativa significativa obtida pelo teste de *Pearson* (com *p-value* = 1.39e-05, menor que 0,05), sugerindo que repositórios com maior CBO têm, em média, menos popularidade. A correlação de *Spearman* não foi significativa (com *p-value* = 0.429424), indicando que a relação observada não segue uma tendência única e constante ao longo do tempo. A linha de tendência no gráfico também reflete essa relação negativa.

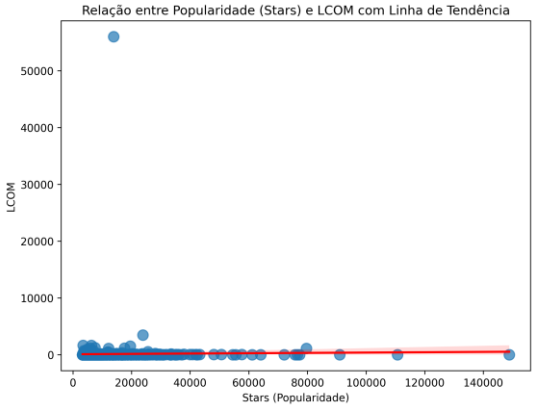


➔ Stars (popularidade) x DIT: Os resultados indicam que repositórios com maior profundidade na árvore de herança (DIT) tendem a ter menos estrelas (Stars). Esse comportamento é evidenciado pela correlação negativa significativa obtida pelo teste de *Pearson* (com *p-value* = 3.59e-06, menor que 0,05), sugerindo que repositórios com hierarquias mais profundas podem ser menos populares. No entanto, a correlação de *Spearman* não foi significativa (com *p-value* = 0.119981), indicando que a relação observada não segue uma tendência única e constante ao longo do tempo. Esse padrão também pode ser observado na linha de tendência no gráfico, que apresenta um leve declínio.



➔ Stars (popularidade) x LCOM: Os resultados indicam que não há uma relação estatisticamente significativa entre o número de estrelas (Stars) e a falta de coesão entre

métodos (LCOM). Tanto o teste de *Pearson* (*p-value* = 0.566455) quanto o teste de *Spearman* (*p-value* = 0.901622) apresentaram valores acima do nível de significância de 0,05, indicando que a coesão dos métodos em uma classe não influencia diretamente a popularidade do repositório. Esse comportamento também pode ser observado no gráfico, onde a linha de tendência não apresenta uma inclinação clara.



Diante destes resultados, a hipótese inicial foi parcialmente validada. Os resultados indicam que repositórios mais populares podem ter um menor acoplamento (CBO) e estruturas de herança mais simples (DIT), mas não necessariamente apresentam maior coesão interna entre métodos (LCOM). Assim, embora a popularidade possa estar associada a alguns aspectos positivos de qualidade do código, ela não garante qualidade em todos os atributos analisados.

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

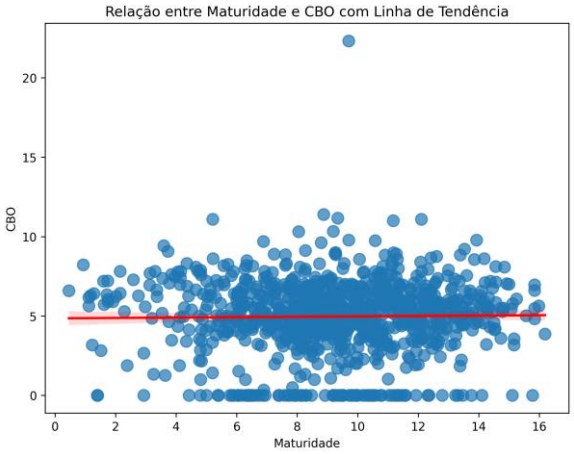
Hipótese**:** Repositórios com maior maturidade têm melhores atributos de qualidade, tendo em vista que passaram por mais revisões, refatorações e melhorias ao longo do tempo.

Resultados:

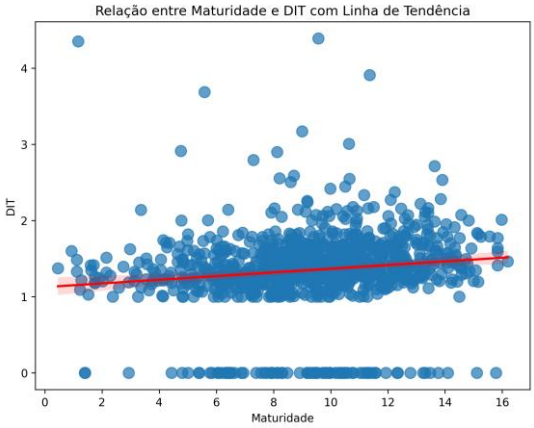
Os resultados a seguir evidenciam, por meio de gráficos de dispersão, a relação entre maturidade e atributos de qualidade. A significância estatística foi verificada pelos testes de Correlação de *Pearson* e *Spearman*:

	Metric	Pearson_Coefficient	Pearson_p-value	Spearman_Coefficient	Spearman_p-value
0	CBO	0.0159292	0.618622	0.00824382	0.796704
1	DIT	0.135348	2.14674e-05	0.244598	8.38454e-15
2	LCOM	0.0279079	0.383064	0.1823	9.19791e-09

➔ Maturidade x CBO: Os resultados indicam que não há uma relação estatisticamente significativa entre a maturidade do repositório e o acoplamento entre classes (CBO). Tanto o coeficiente de *Pearson* (0,0159, *p-value* = 0,6186) quanto o coeficiente de *Spearman* (0,0082, *p-value* = 0,7967) são próximos de zero e apresentam p-valores elevados, sugerindo que o tempo de existência do repositório não influencia diretamente o nível de acoplamento entre classes.

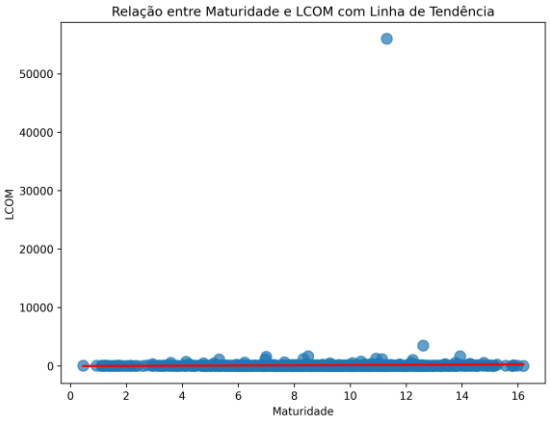


➔ Maturidade x DIT: Os resultados indicam uma relação positiva entre a maturidade do repositório e a profundidade da hierarquia de herança (DIT). O coeficiente de *Pearson* (0,1353, *p-value* < 0,001) e, principalmente, o coeficiente de *Spearman* (0,2446, *p-value* < 0,001) mostram uma correlação significativa, sugerindo que repositórios mais antigos tendem a apresentar uma hierarquia de classes mais profunda.



➔ Maturidade x LCOM: A relação entre maturidade e coesão das classes (LCOM) também apresenta significância estatística no teste de *Spearman* (0,1823, *p-value* < 0,001), enquanto no teste de *Pearson* não há um indicativo claro de correlação (0,0279, *p-value* = 0,3831). Isso sugere que repositórios mais antigos tendem a ter classes menos coesas (LCOM mais alto), possivelmente devido à adição de novas funcionalidades ao longo do

tempo, o que pode levar a um aumento da dispersão de responsabilidades dentro das classes.



Diante destes resultados, a hipótese inicial foi parcialmente validada. Repositórios mais maduros apresentam maior DIT, indicando maior uso de herança, enquanto a relação com CBO não foi significativa. Já o aumento de LCOM sugere menor coesão em classes mais antigas, possivelmente devido à evolução do software. Isso indica que maturidade nem sempre resulta em melhor qualidade em todas as métricas.

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

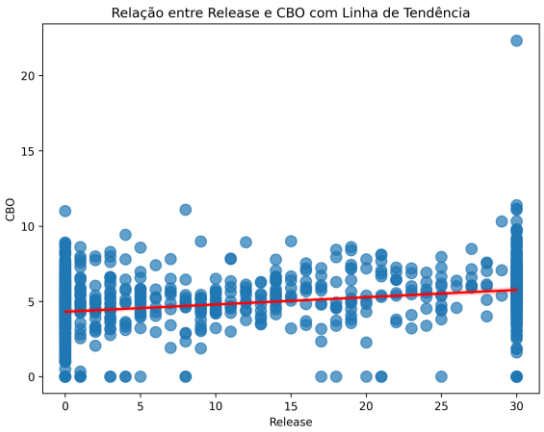
Hipótese**:** Repositórios com mais atividade (releases) têm atributos de qualidade medianos, tendo em vista que passam por atualizações frequentes, o que pode impactar a coesão e o acoplamento do código.

Resultados:

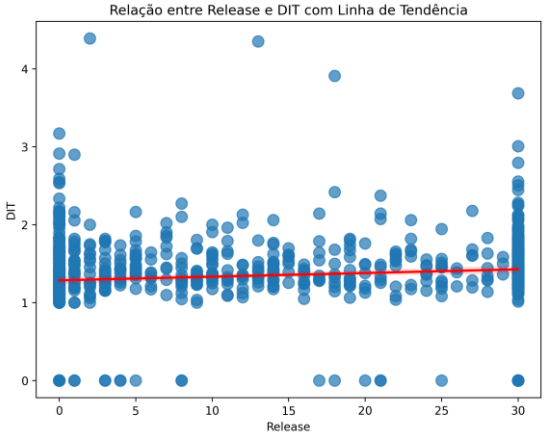
Os resultados a seguir evidenciam, por meio de gráficos de dispersão, a relação entre maturidade e atributos de qualidade. A significância estatística foi verificada pelos testes de Correlação de *Pearson* e *Spearman*:

	Metric	Pearson_Coefficient	Pearson_p-value	Spearman_Coefficient	Spearman_p-value
0	CBO	0.277802	8.29941e-19	0.321142	6.37099e-25
1	DIT	0.117152	0.000239224	0.214059	1.30836e-11
2	LCOM	-0.0284641	0.373652	0.27468	2.08909e-18

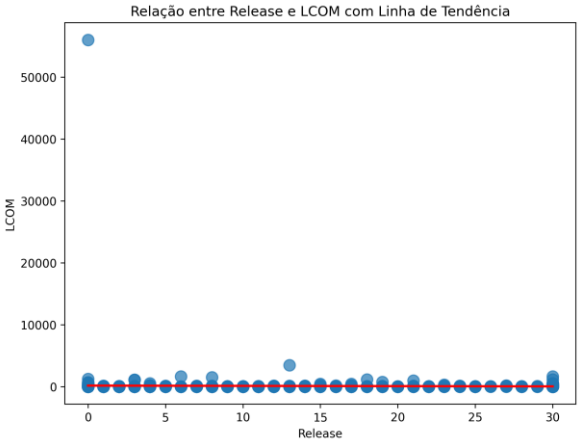
➔ Releases (atividades) x CBO: A correlação positiva e estatisticamente significativa (*Pearson* = 0.278, *Spearman* = 0.321) indica que repositórios com mais releases tendem a apresentar maior acoplamento entre classes, possivelmente devido à evolução contínua do código e à introdução de novas dependências ao longo do tempo.



➔ Releases (atividades) x DIT: A relação positiva (*Pearson* = 0.117, *Spearman* = 0.214) e estatisticamente significativa sugere que repositórios com mais releases podem ter estruturas de herança ligeiramente mais profundas, indicando uma organização hierárquica mais elaborada conforme o projeto cresce.



➔ Releases (atividades) x LCOM: Embora a correlação de *Pearson* seja não significativa e próxima de zero (-0.028), a correlação significativa de *Spearman* (0.275) sugere que repositórios com mais releases podem ter um leve aumento na falta de coesão entre métodos, o que pode refletir a complexidade crescente de projetos mais ativos.



Diante destes resultados, a hipótese inicial foi validada. Ou seja, repositórios com mais atividade tendem a ter atributos de qualidade medianos, indicando que a frequência de atualizações pode impactar a qualidade do código. A relação observada entre releases e CBO sugere um aumento no acoplamento do código, enquanto a relação com DIT indica um crescimento moderado na profundidade da árvore de herança. Embora a correlação com LCOM mostre uma leve tendência de aumento na falta de coesão, a relação não é muito forte.

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

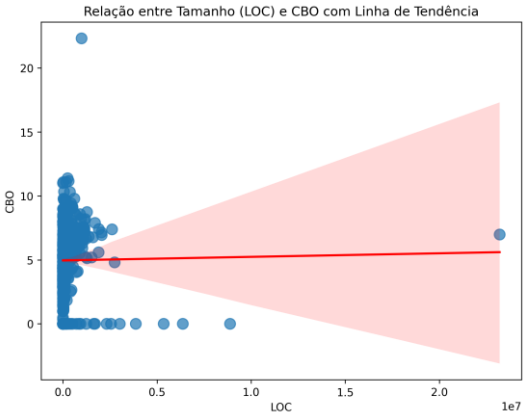
Hipótese**:** Repositórios com mais linhas de código (LOC) têm atributos de qualidade mediano, considerando que o aumento do tamanho do código pode levar a uma maior complexidade, impactando o acoplamento, a herança e a coesão.

Resultados:

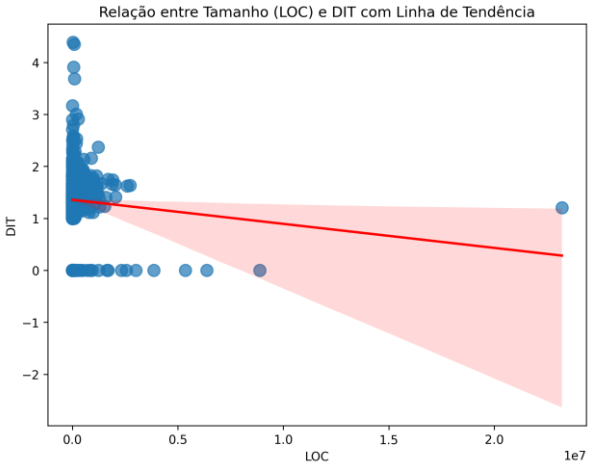
Os resultados a seguir evidenciam, por meio de gráficos de dispersão, a relação entre maturidade e atributos de qualidade. A significância estatística foi verificada pelos testes de Correlação de *Pearson* e *Spearman*:

	Metric	Pearson_Coefficient	Pearson_p-value	Spearman_Coefficient	Spearman_p-value
0	CBO	0.0110517	0.72982	0.439882	1.37695e-47
1	DIT	-0.0795308	0.012803	0.300109	7.97406e-22
2	LCOM	0.0291935	0.361525	0.461813	7.10159e-53

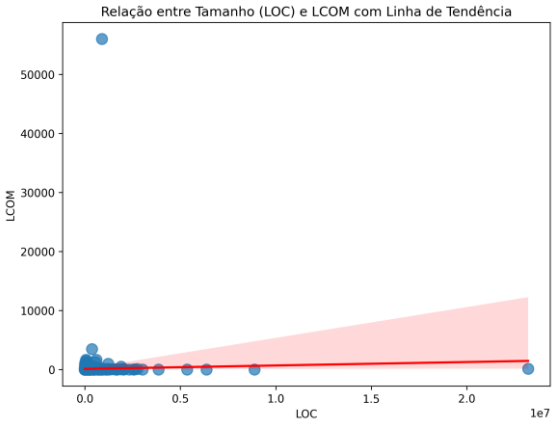
➔ LOC (tamanho dos repositórios) x CBO: A correlação entre as variáveis foi fraca, com o coeficiente de *Pearson* sendo 0.011 (*p-value* = 0.729), indicando falta de significância. No entanto, o teste de *Spearman* apresentou uma correlação moderada de 0.44 (*p-value* = 1.37e-47), sugerindo que, em termos de ordenação, repositórios maiores em LOC tendem a ter mais acoplamento.



➔ LOC (tamanho dos repositórios) x DIT: A correlação entre as variáveis foi fraca a moderada, com o coeficiente de *Pearson* de -0.079 (*p-value* = 0.0128), indicando uma correlação significativa, mas fraca. O teste de *Spearman* mostrou uma correlação moderada positiva de 0.30 (*p-value* = 7.97e-22), sugerindo que repositórios maiores em LOC tendem a ter maior profundidade na árvore de herança.



➔ LOC (tamanho dos repositórios) x LCOM: A correlação de *Pearson* entre as variáveis foi fraca (0.029, *p-value* = 0.361), sem significância. Porém, o teste de *Spearman* apresentou uma correlação moderada de 0.46 (*p-value* = 7.1e-53), indicando que repositórios maiores em LOC tendem a ter maior falta de coesão entre os métodos.



Diante dos resultados, a hipótese inicial foi validada. Repositórios com mais linhas de código (LOC) mostram correlações moderadas com as métricas de qualidade. O aumento de LOC está relacionado a maior acoplamento (CBO) e maior profundidade na árvore de herança (DIT), além de menor coesão (LCOM). Esses resultados sugerem que repositórios maiores tendem a ter maior complexidade, impactando a qualidade do código.

Conclusão

- ➔ Repositórios mais populares (mais estrelas) tendem a ter menor acoplamento (CBO) e menor profundidade de herança (DIT), indicando melhor modularização. Não há relação clara com a coesão (LCOM).
- ➔ Repositórios mais maduros (com mais tempo de existência) não mostram correlação com CBO ou LCOM, mas apresentam uma correlação positiva com DIT, sugerindo maior complexidade com o tempo.
- ➔ Repositórios com mais releases tendem a ter maior acoplamento (CBO) e profundidade de herança (DIT), e menor coesão (LCOM), sugerindo que a atividade de manutenção aumenta a complexidade e impacta a organização do código.
- ➔ Repositórios maiores (mais linhas de código) têm maior acoplamento (CBO) e profundidade de herança (DIT), refletindo maior complexidade. A coesão (LCOM) tende a ser mais baixa, indicando que repositórios grandes podem ser menos coesos e mais difíceis de manter.

A popularidade, maturidade, nível de atividade e tamanho do repositório estão todos associados de maneira variada com as métricas de qualidade. Enquanto repositórios populares e maduros tendem a ter melhor modularização, repositórios grandes e com mais releases geralmente exibem maior complexidade e menor coesão. Isso sugere que a qualidade de código pode ser impactada por múltiplos fatores, com *trade-offs* entre complexidade e qualidade em repositórios maiores e mais ativos.