Análise de Correlação: Popularidade, Maturidade e Qualidade de Código em Repositórios Java

Relacionamos popularidade, maturidade, atividade e tamanho de repositórios Java com métricas de qualidade de código (CBO, DIT, LCOM).

H1: Popularidade ³ Qualidade

Projectos mais populares (mais stars) tenderiam a
melhor qualidade (menor CBO, menor LCOM, DIT
moderado).

H2: Maturidade ³ Qualidade

Projectos mais antigos tenderiam a acumular herança/camadas (DIT ±) e, com o tempo, ficar menos coesos (LCOM ±).

H3: Actividade ³ Qualidade

Mais actividade poderia reflectir mais mudanças e
pontos de acoplamento (CBO ±) e mais complexidade.

H4: Tamanho ³ Qualidade

Projectos maiores tenderiam a ter mais acoplame
(CBO ±), mais complexidade e menos coesão (LCO
±).

Metodologia

Base de dados

Utilizei os 1000 repositórios Java contendo, por repositório, métricas de processo (ex.: stars, forks, watchers, open_issues, age_years, size_kb, total_classes) e de qualidade (ex.: cbo_mean, dit_mean, lcom_mean, complexity_mean, cohesion_mean).

Roteiro analítico

Python

Linguagem de programação principal

Pandas

Manipulação e análise de dados

APIs

GraphQL e GitLab para colecta de dados

Sumarização Global das Métricas de Qualidade

As métricas foram calculadas sobre cbo_mean, dit_mean, lcom_mean, complexity_mean, cohesion_mean para análise estatística descritiva por repositório.

5.19

CBO Média

Mediana: 5,16 | Desvio: 1,78

1.46

DIT Média

Mediana: 1,39 | Desvio: 0,36

59.26

LCOM Média

Mediana: 22,55 | Desvio: 159,95

Complexity Mean

Média: 15,50

Mediana: 14,08

Desvio: 8,92

Cohesion Mean

Média: 0,645

Mediana: 0,635

Desvio: 0,122

Essas estatísticas revelam padrões interessantes: LCOM apresenta alta variabilidade (desvio de 159,95), sugerindo grande diversidade na coesão entre projectos, enquanto DIT mostra valores consistentemente baixos, indicando árvores de herança relativamente rasas na maioria dos repositórios.

RQ1: Popularidade vs Qualidade

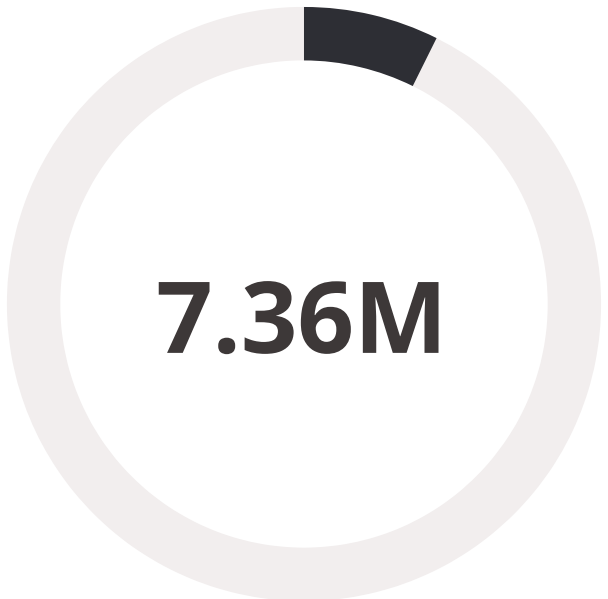
Achado Principal

Não há correlação significativa entre stars e as métricas de qualidade (Spearman próximo de zero para CBO, LCOM, Complexity e Cohesion; DIT com 20,03).

Exemplos de Correlação (Spearman, nj835)

Métrica	Correlação	Significância
stars × CBO_mean	j 0,00	ns
stars × DIT_mean	j 20,03	ns
stars × LCOM_mean	j 0,03	ns

> **Leitura:** Popularidade (atracção de utilizadores) não implica, em melhor (ou pior) qualidade interna.



Estrelas Agregadas

Volume total observado no Dashboard

Curiosidade: o Dashbord mostra volume agregado de popularidade (~7,36M estrelas), mas essa grandeza não se traduz directamente em efeito sobre CBO/DIT/LCOM na análise repositório-a-repositório.

RQ2 & RQ3: Maturidade e Atividade

RQ2: Maturidade

Idade em anos vs Qualidade

RQ3: Atividade

Open issues vs Qualidade

Achados RQ2 - Maturidade

- **age_years × DIT_mean:** +0,28 (p j)
- **age_years × LCOM_mean:** +0,18 (p j)
- **age_years × Cohesion_mean:** 20,12 ()

Repositórios mais antigos tendem a ter árvores de herança mais profundas e menor coesão com o tempo.

Achados RQ3 - Actividade

- **open_issues × CBO_mean:** +0,22 (j)
- **open_issues × LCOM_mean:** +0,26 (p j)
- **open_issues × Complexity_mean:** (pj)

Mais issues abertas correlacionam com mais acoplamento/complexidade 4 sistemas onde problemas acumulam.

Com a evolução, os projetos tendem a acumular camadas/linhagens de herança e "espalhar" responsabilidades, prejudicando levemente a coesão. A atividade elevada (medida por issues abertas) sugere bases de código "conturbadas" com maior acoplamento.

RQ4 & Discussão dos Resultados

RQ4: Tamanho vs Qualidade

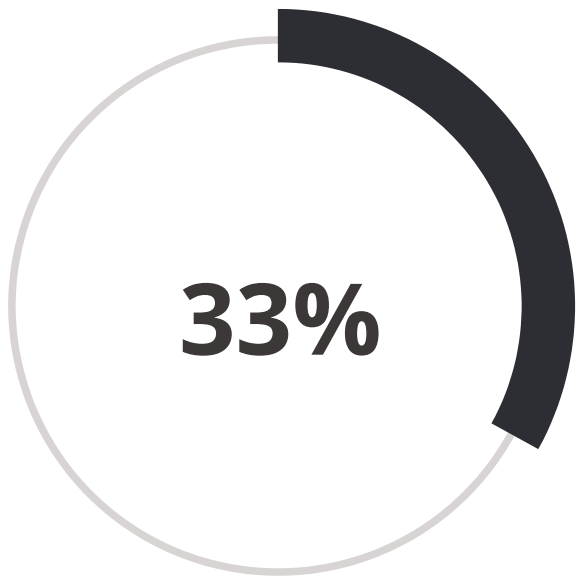
Correlações significativas encontradas

1

2

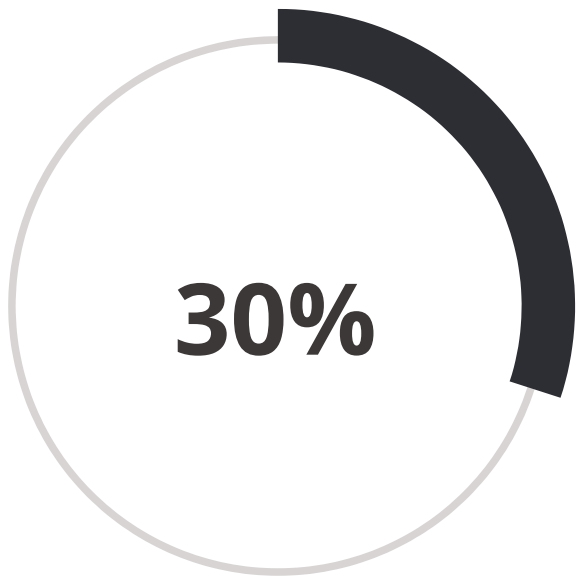
Validação das Hipóteses

Confronto entre expectativas e resultados



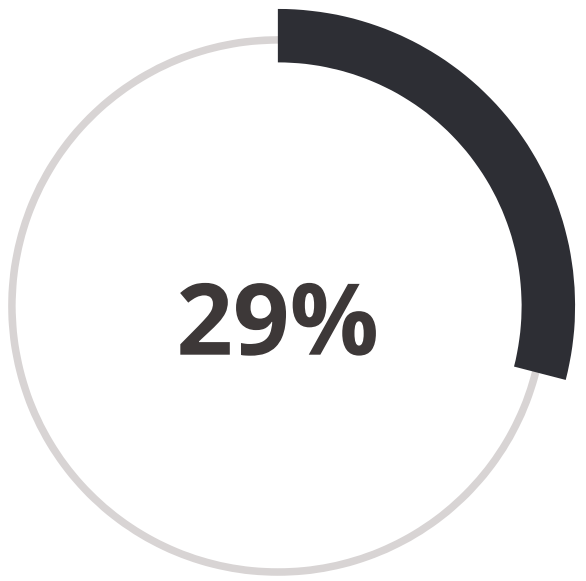
size_kb × CBO_mean

Correlação de Spearman (j)



size_kb × Complexity

Projetos maiores = mais complexos



total_classes × CBO

Mais classes = mais acoplamento

Projetos maiores (em tamanho do repo/quantidade de classes) correlacionam com mais acoplamento, mais complexidade e menor coesão 4 alinhado com H4.

☆

H1: Popularidade

Não confirmada. Stars praticamente neutro em relação a CBO/DIT/LCOM. Popularidade reflecte

🕐

H2: Maturidade

Parcialmente confirmada. Idade correlaciona positivamente com DIT e LCOM, sugerindo "peso

Conclusões e Análise Estatística

Gráficos de Correlação

Gerei gráficos de dispersão para RQ13RQ4 com análise de Spearman & Pearson consolidada em CSV.

Testes Estatísticos

Report rho de Spearman (mais robusto) e r de Pearson (linearidade). N 835 repositórios após descartar NaN.

835

Repositórios

Amostra final analisada

4

RQs Testadas

Questões de pesquisa validadas

Insights Principais

- Tamanho é o fator mais correlacionado com degradação da qualidade
- Maturidade traz complexidade estrutural acumulada
- Popularidade não garante qualidade interna
- Atividade alta pode indicar problemas arquiteturais

> No Dashboard, o panorama por licença/popularidade ajuda a ler o "contexto macro" (ex.: Apache 2.0, MIT e GPL v3 concentram grande parte da fama), mas a relação direta licença qualidade não foi testada aqui.

Os resultados sugerem que o crescimento de projetos de software traz desafios inevitáveis de modularidade e legibilidade, independentemente da sua popularidade. A gestão da qualidade arquitetural requer atenção contínua, especialmente em projetos de maior escala e longevidade.