Relatório - Lab 02

Eduardo Bandeira de Melo Guimarães

1. Qual a relação entre a popularidade dos repositórios e as suas características de qualidade?
   - **Hipóteses:**
     - **Existe uma relação direta entre a popularidade dos repositórios e as suas métricas de qualidade**. Acredito que a quantidade de contribuintes em projetos é diretamente proporcional à sua popularidade. Assim, mais pessoas contribuindo para um projeto faz com que exista um filtro mais rigoroso para boas práticas de codificação. Mais pessoas contribuindo significa um aumento na probabilidade de existirem contribuidores mais experientes e com um conhecimento mais maduro de boas práticas.
   - **Metodologia:**
     - Para responder a esta questão, foi escrito um script em Python para fazer as seguintes tarefas: foram buscados os 1000 repositórios Java mais populares do GitHub usando a API GraphQL. Para cada um desses repositórios, foram extraídas as **quantidades de estrelas** de cada um deles para se avaliar o **grau de popularidade**, bem como a url do repositório. Então, com a url de cada repositório, clonou-se cada um deles e executou-se sobre eles a ferramenta de análise estática de código CK. Ao se executar a ferramenta, filtrou-se as métricas de qualidade: i) CBO's das classes; ii) DIT's das classes e iii) LCOM's dos métodos das classes. Foi feito um tratamento dos dados obtidos para eliminar repositórios que apresentaram algum erro durante a execução do script ou que apresentaram dados corrompidos. Dos 1000 repositórios analisados, 40 tiveram de ser eliminados. Então, finalmente, foi feito o teste estatístico de Pearson, usando a biblioteca Pandas, para se verificar a relação de correlação entre a **popularidade** dos repositórios e as características de qualidade analisadas.
   - **Resultados obtidos:**
     - Os resultados obtidos mostram que **praticamente não existe correlação entre a popularidade dos repositórios e as métricas de qualidade avaliadas**. Todos os resultados obtidos para o índice de correlação de Pearson estão na ordem de grandeza 10^{-2}. Para se ter uma ideia, para que uma correlação fraca comece a existir, é necessário um índice de correlação de Pearson de, pelo menos 0,20. O maior índice encontrado para os índices de correlação analisados foi de 0,08. Além disso, o coeficiente de Spearman encontrado (0,05) também foi muito baixo para encontrar qualquer relação de correlação (correlação fraca a partir de 0,1).
   - **Discussão sobre hipóteses vs resultados:**
     - Os resultados encontrados **refutam veementemente** a hipótese inicialmente levantada de que existe correlação entre a popularidade do repositório e seus índices de qualidade.

	Media de CBO's	Mediana de CBO's	Media de DIT's	Mediana de DIT's	Media de LCOM's	Mediana de LCOM's
Estrelas	-0.08197050241469327	-0.04218695727880081	-0.08400286947552586	-0.04266723411505075	0.024540298275037224	-0.020394316026645556
Idade em anos	0.01356522243601895	0.0028318186046523363	0.17004936269366178	0.07706968141696548	0.028162131247677382	0.012605600934052828
Releases	0.21017398625057038	0.13621887151534068	0.05580149468274014	-0.02827906537663338	-0.012945966610185098	-0.013526726736822819
Linhas totais	0.16933456613899478	0.1349458270429522	0.04012316552860754	-0.019747026296526658	0.052719035455903625	0.006378042118288096
Media de CBO's	1.0	0.835534498770266	0.2641345543117829	0.1336688568788439	0.06920768857245485	0.11271656185634955
Mediana de CBO's	0.835534498770266	1.0	0.1664481715236743	0.1790856224102235	0.07056418831463959	0.18280699716641585
Media de DIT's	0.2641345543117829	0.1664481715236743	1.0	0.6539537375207465	0.07360680374143984	0.0289330466866332
Mediana de DIT's	0.1336688568788439	0.1790856224102235	0.6539537375207465	1.0	0.09527941888910621	0.10089969055357151
Media de LCOM's	0.06920768857245485	0.07056418831463959	0.07360680374143984	0.09527941888910621	1.0	0.02354794203430376
Mediana de LCOM's	0.11271656185634955	0.18280699716641585	0.0289330466866332	0.10089969055357151	0.02354794203430376	1.0

2. Qual a relação entre a maturidade dos repositórios e as suas características de qualidade?
   - **Hipóteses:**
     - **Existe uma relação direta entre a maturidade dos repositórios e as suas métricas de qualidade**. Acredito que a quantidade de contribuintes em projetos é diretamente proporcional à sua maturidade. Assim, mais pessoas contribuindo para um projeto faz com que exista um filtro mais rigoroso para boas práticas de codificação. Mais pessoas contribuindo significa um aumento na probabilidade de existirem contribuidores mais experientes e com um conhecimento mais maduro de boas práticas.
   - **Metodologia:**
     - Para responder a esta questão, foi escrito um script em Python para fazer as seguintes tarefas: foram buscados os 1000 repositórios Java mais populares do GitHub usando a API GraphQL. Para cada um desses repositórios, foram extraídas as **idades em anos** de cada um deles para se avaliar a **maturidade**, bem como a url do repositório. Então, com a url de cada repositório, clonou-se cada um deles e executou-se sobre eles a ferramenta de análise estática de código CK. Ao se executar a ferramenta, filtrou-se as métricas de qualidade: i) CBO's das classes; ii) DIT's das classes e iii) LCOM's dos métodos das classes. Foi feito um tratamento dos dados obtidos para eliminar repositórios que apresentaram algum erro durante a execução do script ou que apresentaram dados corrompidos. Dos 1000 repositórios analisados, 40 tiveram de ser eliminados. Então, finalmente, foi feito o teste estatístico de Pearson, usando a biblioteca Pandas, para se verificar a relação de correlação entre a **maturidade** dos repositórios e as características de qualidade analisadas.
   - **Resultados obtidos:**
     - Os resultados obtidos mostram que **praticamente não existe correlação entre a maturidade dos repositórios e as métricas de qualidade avaliadas**. Os resultados obtidos para o índice de correlação de Pearson são muito baixos, variando de 0,0027 a 0,1696. Para que uma correlação fraca comece a existir, é necessário um índice de correlação de Pearson de, pelo menos 0,20. O maior índice encontrado para os índices de correlação analisados foi de aproximadamente 0,17. Além disso, o coeficiente de Spearman encontrado (-0,07) também foi muito baixo para encontrar qualquer relação de correlação (correlação fraca a partir de 0,1).
   - **Discussão sobre hipóteses vs resultados:**
     - Os resultados encontrados **refutam veementemente** a hipótese inicialmente levantada de que existe correlação entre a maturidade do repositório e seus índices de qualidade.

	ia de CBO's	Mediana de CBO's	Media de DIT's	Mediana de DIT's	Media de LCOM's	Mediana de LCOM's
Estrelas	3197050241469327	-0.04218695727880081	-0.08400286947552586	-0.04266723411505075	0.024540298275037224	-0.020394316026645556
Idade em anos	356522243601895	0.0028318186046523363	0.17004936269366178	0.07706968141696548	0.028162131247677382	0.012605600934052828
Releases	017398625057038	0.13621887151534068	0.05580149468274014	-0.02827906537663338	-0.012945966610185098	-0.013526726736822819
Linhas totais	933456613899478	0.1349458270429522	0.04012316552860754	-0.019747026296526658	0.052719035455903625	0.006378042118288096
Media de CBO's		0.835534498770266	0.2641345543117829	0.1336688568788439	0.06920768857245485	0.11271656185634955
Mediana de CBO's	5534498770266	1.0	0.1664481715236743	0.1790856224102235	0.07056418831463959	0.18280699716641585
Media de DIT's	41345543117829	0.1664481715236743	1.0	0.6539537375207465	0.07360680374143984	0.0289330466866332
Mediana de DIT's	36688568788439	0.1790856224102235	0.6539537375207465	1.0	0.09527941888910621	0.10089969055357151
Media de LCOM's	920768857245485	0.07056418831463959	0.07360680374143984	0.09527941888910621	1.0	0.02354794203430376
Mediana de LCOM's	271656185634955	0.18280699716641585	0.0289330466866332	0.10089969055357151	0.02354794203430376	1.0

3. Qual a relação entre a atividade dos repositórios e as suas características de qualidade?
   - **Hipóteses:**
     - Levantando-se uma hipótese pura sobre a relação entre a atividade de projetos e suas métricas de qualidade, diria que essa relação não existe. Entretanto, os dados analisados estão “contaminados”. Ou seja, os repositórios que estão sendo analisados já estão no TOP-1000 mais populares do GitHub e provavelmente essa característica por si só já exerce influência sobre suas características de qualidade. Dessa forma, levanto-se em conta que os repositórios analisados já estão entre os mais populares do GitHub, diria que, dentre eles, **existe** sim **uma relação direta entre a atividade e as métricas de qualidade** pelo mesmo motivo já citado anteriormente: mais atividade provavelmente significa mais contribuintes e isso levaria a uma participação mais provável de desenvolvedores mais experientes, que manteriam níveis elevados de qualidade.
   - **Metodologia:**
     - Para responder a esta questão, foi escrito um script em Python para fazer as seguintes tarefas: foram buscados os 1000 repositórios Java mais populares do GitHub usando a API GraphQL. Para cada um desses repositórios, foram extraídas as **quantidades de releases** de cada um deles para se avaliar a **atividade**, bem como a url do repositório. Então, com a url de cada repositório, clonou-se cada um deles e executou-se sobre eles a ferramenta de análise estática de código CK. Ao se executar a ferramenta, filtrou-se as métricas de qualidade: i) CBO’s das classes; ii) DIT’s das classes e iii) LCOM’s dos métodos das classes. Foi feito um tratamento dos dados obtidos para eliminar repositórios que apresentaram algum erro durante a execução do script ou que apresentaram dados corrompidos. Dos 1000 repositórios analisados, 40 tiveram de ser eliminados. Depois, foi feita uma outra “limpeza” dos dados, removendo repositórios sem nenhuma release. Provavelmente, para estes repositórios, o lançamento de releases não é feito de forma sistemática e como etapa de lançamento de atualizações, o que faz com que a métrica “quantidade de releases” não sirva como ferramenta de medição de atividade para esses repositórios. Então, finalmente, foi feito o teste estatístico de Pearson, usando a biblioteca Pandas, para se verificar a relação de correlação entre a **atividade** dos repositórios e as características de qualidade analisadas.
   - **Resultados obtidos:**
     - Os resultados obtidos mostram que praticamente **não existe correlação entre a atividade dos repositórios e as métricas de qualidade avaliadas, com exceção da média de CBO’, que apresenta uma correlação baixa**. Os resultados obtidos para o índice de correlação de Pearson são muito baixos, variando de 0,013 a 0,136. Para que uma correlação fraca comece a existir, é necessário um índice de correlação de Pearson de, pelo menos 0,20. Apenas a métrica “média de CBO’s” apresenta um índice cujo valor pode representar certo grau de correlação baixa (0,210).
   - **Discussão sobre hipóteses vs resultados:**
     - Os resultados encontrados **refutam com ressalvas** a hipótese inicialmente levantada de que existe correlação entre a atividade do repositório e seus índices de qualidade.

	Media de CBO's	Mediana de CBO's	Media de DIT's	Mediana de DIT's	Media de LCOM's	Mediana de LCOM's
Estrelas	-0.08197050241469327	-0.04218695727880081	-0.08400286947552586	-0.04266723411505075	0.024540298275037224	-0.020394316026645556
Idade em anos	0.01356522243601895	0.0028318186046523363	0.17004936269366178	0.07706968141696548	0.028162131247677382	0.012605600934052828
Releases	0.21017398625057038	0.13621887151534068	0.05580149468274014	-0.02827906537663338	-0.012945966610185098	-0.013526726736822819
Linhas totais	0.16933456613899478	0.1349458270429522	0.04012316552860754	-0.019747026296526658	0.052719035455903625	0.006378042118288096
Media de CBO's	1.0	0.835534498770266	0.2641345543117829	0.1336688568788439	0.06920768857245485	0.11271656185634955
Mediana de CBO's	0.835534498770266	1.0	0.1664481715236743	0.1790856224102235	0.07056418831463959	0.18280699716641585
Media de DIT's	0.2641345543117829	0.1664481715236743	1.0	0.6539537375207465	0.07360680374143984	0.0289330466866332
Mediana de DIT's	0.1336688568788439	0.1790856224102235	0.6539537375207465	1.0	0.09527941888910621	0.10089969055357151
Media de LCOM's	0.06920768857245485	0.07056418831463959	0.07360680374143984	0.09527941888910621	1.0	0.02354794203430376
Mediana de LCOM's	0.11271656185634955	0.18280699716641585	0.0289330466866332	0.10089969055357151	0.02354794203430376	1.0

4. Qual a relação entre o tamanho dos repositórios e as suas características de qualidade?
   - **Hipóteses:**
     - Não existe nenhuma relação entre o tamanho de um projeto e suas características de qualidade. Essa ausência de relação já é bastante conhecida na Engenharia de Software pelos mais diversos motivos, sendo o principal deles o fato de que linhas de código não é uma boa métrica para tamanho ou complexidade de um software. Além disso, a complexidade de um sistema não determina de forma alguma sua qualidade. O que define a qualidade são diversos outros fatores, como: i) projeto bem feito e bem organizado; ii) experiência da equipe que trabalha no projeto; iii) condução do projeto; entre vários outros.
   - **Metodologia:**
     - Para responder a esta questão, foi escrito um script em Python para fazer as seguintes tarefas: foram buscados os 1000 repositórios Java mais populares do GitHub usando a API GraphQL. Para cada um desses repositórios, foram extraídas as url's do repositório. Então, com a url de cada repositório, clonou-se cada um deles e executou-se sobre eles a ferramenta de análise estática de código CK. Ao se executar a ferramenta, filtrou-se as métricas de qualidade: i) CBO's das classes; ii) DIT's das classes e iii) LCOM's dos métodos das classes, assim como as LOC's de cada classe. Foi feito um tratamento dos dados obtidos para eliminar repositórios que apresentaram algum erro durante a execução do script ou que apresentaram dados corrompidos. Dos 1000 repositórios analisados, 40 tiveram de ser eliminados. Somou-se, então, para cada repositório, as linhas da métrica "LOC" para se obter a **quantidade total de linhas de código** do projeto como métrica de **tamanho**. Então, finalmente, foi feito o teste estatístico de Pearson, usando a biblioteca Pandas, para se verificar a relação de correlação entre o **tamanho** dos repositórios e as características de qualidade analisadas.
   - **Resultados obtidos:** Os resultados obtidos mostram que praticamente não existe correlação entre o tamanho dos repositórios e as métricas de qualidade avaliadas. Os resultados obtidos para o índice de correlação de Pearson são muito baixos, variando de 0,006 a 0,169. Para que uma correlação fraca comece a existir, é necessário um índice de correlação de Pearson de, pelo menos 0,20.
   - **Discussão sobre hipóteses vs resultados:** Os resultados encontrados **refutam veementemente** a hipótese inicialmente levantada de que existe correlação entre a maturidade do repositório e seus índices de qualidade.

	Media de CBO's	Mediana de CBO's	Media de DIT's	Mediana de DIT's	Media de LCOM's	Mediana de LCOM's
Estrelas	-0.08197050241469327	-0.04218695727880081	-0.08400286947552586	-0.04266723411505075	0.024540298275037224	-0.020394316026645556
Idade em anos	0.01356522243601895	0.0028318186046523363	0.17004936269366178	0.07706968141696548	0.028162131247677382	0.012605600934052828
Releases	0.21017398625057038	0.13621887151534068	0.05580149468274014	-0.02827906537663338	-0.012945966610185098	-0.013526726736822819
Linhas totais	0.16933456613899478	0.1349458270429522	0.04012316552860754	-0.019747026296526658	0.052719035455903625	0.006378042118288096
Media de CBO's	1.0	0.835534498770266	0.2641345543117829	0.1336688568788439	0.06920768857245485	0.11271656185634955
Mediana de CBO's	0.835534498770266	1.0	0.1664481715236743	0.1790856224102235	0.07056418831463959	0.18280699716641585
Media de DIT's	0.2641345543117829	0.1664481715236743	1.0	0.6539537375207465	0.07360680374143984	0.0289330466866332
Mediana de DIT's	0.1336688568788439	0.1790856224102235	0.6539537375207465	1.0	0.09527941888910621	0.10089969055357151
Media de LCOM's	0.06920768857245485	0.07056418831463959	0.07360680374143984	0.09527941888910621	1.0	0.02354794203430376
Mediana de LCOM's	0.11271656185634955	0.18280699716641585	0.0289330466866332	0.10089969055357151	0.02354794203430376	1.0

Representação complementar das relações de correlação estudadas: