Trabalho Prático de Matemática Computacional II

Hugo Guimarães,[a] João Santos,[b] Pedro Pinho,[c] and Sónia Oliveira[d]
CIICESI,
ESTG,
Polytechnic of Porto

(Dated: 24 February 2024)

Abstract. Este relatório responde a quatro questões investigações sobre a disciplina de Engenharia de Software II para a disciplina de matemática computacional II. A primeira pergunta é saber como a taxa de cobertura dos testes evoluíram ao longo do trabalho e se a maior parte dos grupos cumpriu o critério definido no Quality Gate a cerca da taxa de cobertura dos mesmos. A segunda pergunta é se houve alguma evolução do número de linhas de código totais, e se estas estão correlacionadas com a taxa de cobertura dos testes. A terceira questão é saber quais os fatores que aumentam ou diminuem o número de testes falhados. A última questão é saber os fatores que aumentam ou diminuem o número de testes falhados

INTRODUÇÃO

Este relatório foi criado no âmbito do trabalho prático da disciplina de Matemática computacional II, e tem como objetivo a aplicação as técnicas de análise de dados abordadas nas aulas de teóricas relativas aos conteúdos C3 C4, C5 e C6. Os dados analisados neste relatório são provenientes da análise de dados (reporting) de vários Projetos de Engenharia de Software II (ESII) em que cada grupo teve de efetuar recolhas semanais ao longo do desenvolvimento do projeto de ESII. Ao longo do desenvolvimento do trabalho de Matemática Computacional II, foi possível retirar conclusões interessantes, que irão ser apresentadas ao longo deste relatório, além disso, este projeto serve como forma de desenvolver a componente métrica e estatística do trabalho de Engenharia de Software II (ESII).

Neste trabalho irá ser apresentado vários dados e o processo que se usou para o estudo dos mesmos:

- BASE DE DADOS E METODOLOGIA – resumo dos dados e metodologia usada;
- RESULTADOS E DISCUSSÕES – valores obtidos da realização dos testes e discussão deles mesmos;
- CONCLUSÕES E TRABALHO FUTURO – conclusões sobre os dados apresentados neste documento.

[a]Corresponding author: 8220337@estg.ipp.pt
[b]Corresponding author: 8220256@estg.ipp.pt
[c]Corresponding author: 8220307@estg.ipp.pt
[d]Corresponding author: 8220114@estg.ipp.pt

TRATAMENTO DE OUTLIERS

Alguns dados importados possuem valores vazios (N.A) e para a resolução desses valores vazios optou-se por deixar como (N.A) quando importamos. Como foram importados os NA's, então, quando se fez os testes tentou-se sempre procurar uma opção que remova os remova.

IDENTIFICAÇÃO DAS VARIÁVEIS

Na base de dados encontram-se 28 conjuntos de dados, que são:

- Grupo Anónimo
- M1) Número de Sprint
- M2) Número de técnicos no projeto de ESII
- M3) Número de User Stories abertas
- M4) Número de User Stories fechadas - Entrega de SW funcional
- M5) Número de pedidos de alterações abertos
- M6) Número de pedidos de alterações rejeitados
- M7) Número de pedidos de alterações aprovados
- M8) Número de testes totais
- M9) Número de testes falhados, por sprint
- M10) Número de Merge Requests, por sprint
- M11) Número de Merge Requests falhados ou com conflitos não solucionados, por sprint
- M12) Número de Classes/componentes
- M13) Número médio de métodos por classe
- M14) Cyclomatic complexity - número de itens (métodos)<10 - O plugin PMD tem regras para obtenção deste valor (CyclomaticComplexity)
- M15) CognitiveComplexity - número de itens (métodos) <10 - O plugin PMD tem regras para obtenção deste valor (CognitiveComplexity)
- M16) Class Coupling - ExcessiveImports <30; CouplingBetweenObjects<20 - Class Coupling existe quando uma classe usa outra de alguma forma. O plugin PMD tem regras para obtenção deste valor (CouplingBetweenObjects; ExcessiveImports)
- M17) Class Cohesion - número de itens <10 - O plugin PMD tem regras para obtenção deste valor
- M18) Número de Linhas de código (LOC) totais
- M19) Número médio de Linhas de código (LOC) por classe
- M20) Número de linhas de Código duplicado totais
- M21) Tempo de ciclo, em dias - tempo médio entre o primeiro commit e a versão final do sprint
- M22) Número de correções necessárias por tempo de ciclo
- M23) Número de alterações por linha de código, ou seja, ao longo dos vários commits, quantas vezes o código já foi alterado, por sprint

- M24) Taxa de cobertura de código nos testes unitários [0, 1], por sprint
- M25) Taxa de cobertura de código na fase de integração contínua [0, 1], por sprint
- M26) Média das notas da UC de PP de todos os técnicos do projeto
- M27) Média das notas da UC de ESI de todos os técnicos do projeto

As variáveis quantitativas discretas são: M1, M2, M3, M4,M5, M6, M7 , M8, M9 e M10

As variáveis quantitativas continuas são: M11, M12, M13, M14, M15, M16, M17, M18, M19, M20, M21, M22, M23 M24, M25, M26 e M27

População: A população utilizada é baseadas nos alunos que entregaram recolhas vindas do projeto de ESII, e cujos resultados nos foram fornecidos para análise pelas docentes.

Amostras: As amostras utilizadas diferem de questão para questão, representando, de um modo geral, um subgrupo da população cuja informação é relevante ao contexto.

Software usados :

- R-Studio
- Excel
- Microsoft Teams
- Overleaf

Linguagens usadas :

- R
- Latex

RESULTADOS E DISCUSSÕES

Antes de se começar a investigação, foi preciso importar o ficheiro excel com a informação das recolhas para o RStudio, para isso, removeu-se a primeira linha do ficheiro original, para que o nome das variáveis fosse igual aos descritos na secção de identificação de variáveis, e resolveu-se alguns problemas de incoerência quanto ao carácter usado como divisor decimal. Por fim foi só exportar o ficheiro para csv, e dentro do RStudio usar o comando read-csv2 para o ler.

```
recolhas <- read.csv2("23_24_Recolhas_TP_MCII_ESII_CSV.csv", sep=";", stringsAsFactors=TRUE)
recolhas$X = NULL
```

FIGURE 1. Função usada para ler o ficheiro das recolhas

```
> recolhas
   Grupo M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12  M13 M14 M15 M16 M17  M18 M19 M20 M21 M22 M23  M24 M25   M26   M27
1      1  1  4 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 14.00 11.00
2      1  2  4  9  1  0  0  0  7  4   3   0  10  3.0   0   0   0   0 1021 204   0   7   0  17 1.00   1 14.00 11.00
3      1  3  4  4  1  0  0  0  8  2   7   0   1  2.0   7   0   0   1  325  43   0   7   0   8 1.00   1 14.00 11.00
4      3  1  4 NA NA NA NA NA NA NA  NA  NA  10  3.0  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 13.00 11.00
5      3  2  4 10 NA  5  0  5 12 NA   7   0  10  8.0  NA  NA  NA  NA  808  81   0   5   3   5   NA  NA 13.00 11.00
6      3  3  4 10 13  7  0  6  3 NA   8   1  12  9.0  NA  NA  NA  NA 1264 105   0   5   3   6   NA  NA 13.00 11.00
7      5  1  4  3  0  0  0  0  0  0   1   0   0  0.0  NA  NA  NA  NA    0   0   0   0   0   0   NA  NA 11.00 11.00
8      5  2  4  2  2  5  1  4 32  6   9   1  17  9.0  NA  NA  NA  NA  988  58   0   5   8   4 1.00  NA 11.00 11.00
9      5  3  4  0  1  1  0  1 80  0  11   2  21 13.0  28 104   0   0 3828 294   0   5   6   5 1.00  NA 11.00 11.00
10    11  1  4  0  0  0  0  0  0  0   0   0   0  0.0   0   0   0   0    0   0   0   0   0   0 0.00   0 12.00 12.00
11    11  2  4  2  1  4  0  0  7  0   8   1  11  9.0  10  15  16   7 2500 150   0   7   0   0 1.00   1 12.00 12.00
12    11  3  4  2  1  1  0  4 21  0   2   0  12  9.0  12  17   8   9 3200 266   0   7   0   0 1.00   1 12.00 12.00
13    15  1  4 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 13.00 13.00
14    15  2  4  4  4  2  0  2 11  0   2   0  16  3.0  34  36  NA  NA  361  20   0   1   3   0 1.00  NA 13.00 13.00
15    15  3  4  5  6  5  1  4 21  0   4   0  19  4.0  81  84   0   0  445  21   0   6   3   2 1.00  NA 13.00 13.00
16    17  1  4 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 11.00 11.00
17    17  2  4  3 NA NA NA NA  7 NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 11.00 11.00
18    17  3  4  1  8  0  0  4 10  0   1   0  12  1.0  12  12  12  12  717  60   0   7   2  39 1.00   1 11.00 11.00
19    19  1  2 NA NA NA NA NA NA NA  NA  NA   6   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 11.00 12.00
20    19  2  2  1 NA NA NA NA NA NA  NA  NA   6 10.0  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 11.00 12.00
21    19  3  2  1 NA NA NA NA NA NA  NA  NA   6 10.0  NA  NA  NA  NA   NA  NA  NA  NA  NA  49   NA  NA 11.00 12.00
22    23  1  4 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 10.00 12.00
23    23  2  4  1  1  0  0  0 38  0   6   2   8  7.0   0   0   0   0  169  21   0   5  NA 590 1.00  NA 10.00 12.00
24    23  3  4  1  0  4  0  4 53  0   8   1   7 18.0   0   0   0   0  622 207   0   3   3  NA   NA  NA 11.00 11.00
25    25  1  4  6  6 NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 13.00    NA
26    25  2  4 10 10 NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 13.00    NA
27    25  3  4  2  2 NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 13.00    NA
28    25  4  4  5  3 NA NA NA  0  0  NA  NA  NA   NA  NA  NA  NA  NA   15  NA  NA  NA  NA  NA   NA  NA 13.00    NA
29    25  5  4 10  1 NA NA NA NA NA  NA  NA   2   NA  NA  NA  NA  NA   35  10  NA  NA  NA  NA   NA  NA 13.00    NA
30    31  1  1 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA  6.00  5.00
31    31  2  1  7  0 NA NA NA NA NA   5   3   7  2.0  NA  NA  NA  NA   NA  NA  NA   7  NA  NA   NA  NA  5.75  4.75
32    31  3  4  7  0  0  3  9 69  0   3   0  11  8.2 149  NA  NA  NA  520  79   0   7   1   1 0.27  NA    NA    NA
33    33  2  4  0  0  0  0  0 34  0   0   0   6 11.0   0   0   0   0  585  97   0   0   0   0 0.75   0 11.00 10.00
34    33  1  4 NA NA NA NA NA NA NA  NA  NA  NA   NA  NA  NA  NA  NA   NA  NA  NA  NA  NA  NA   NA  NA 11.00 10.00
35    33  3  4  0  0  0  0  0 34  0   0   0   6 11.0   0   0   0   0  585  97   0   0   0   0 0.75   0 11.00 10.00
```

FIGURE 2. Valor da variável data

Questão de Investigação 1

A primeira questão de investigação trata-se de **saber a evolução da taxa de cobertura dos testes e saber se os grupos cumpriram o critério definido no Quality Gate dos trabalhos de Engenharia de Software II para a taxa de cobertura admissível dos mesmos**. Para o estudo dessa métrica, foi analisado o ficheiro Excel e verificou-se que dever-se-ia usar a variável M24 (taxa de cobertura de código nos testes unitários [0, 1], por sprint). Após essa verificação percebeu-se que deveria de ser um teste de proporção, pois serão analisadas as taxas de cobertura da segunda e da quarta semana, e este teste permite-nos saber se existiu mudanças significativas entre as duas. Para a realização de todos os testes optou-se por um nível de significância de 5%.

Primeiramente, removeu-se todas as linhas com valores omissos (NA), depois, selecionou-se as colunas relevantes (M1 e M24) para a segunda e a quarta semana. Após a seleção das colunas resolveu-se fazer o summary para cada semana, o resultado desse comando pode ser visualizado na Tabela I e na Tabela II, para além do summary, também se criou um Boxplot para poder haver um apoio visual ao tirar conclusões, esse gráfico pode ser visto na Figura 3

TABLE I. Medidas de Localização dos dados da taxa de cobertura da segunda semana

Semana 2					
Min	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Max**
0	0,7425	1	0,749	1	1

TABLE II. Medidas de Localização dos dados da taxa de cobertura da quarta semana

Semana 4					
Min	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Max**
0,53	0,685	0,81	0,7895	0,8625	1

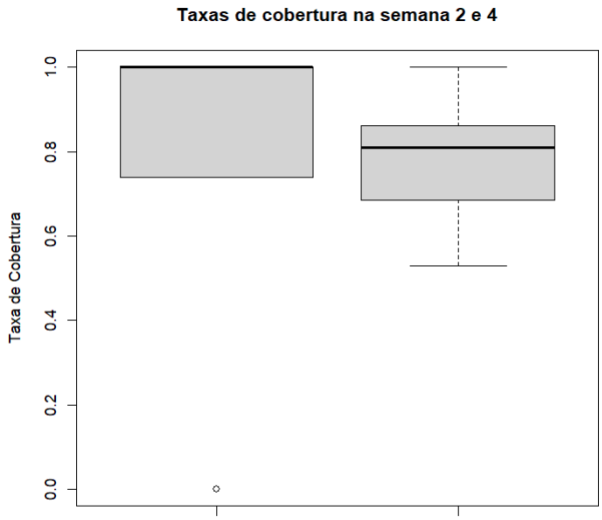

FIGURE 3. Boxplot da distribuição da taxa de cobertura da semana 2 (à esquerda) e da semana 4 (à direita)

Com essa informação, conseguiu-se observar que na segunda semana a mediana está com o valor 1, isso significa que metade das pessoas têm uma taxa de cobertura de 100%. Já na quarta semana consegue-se observar que a taxa de cobertura diminuiu, logo, a taxa de cobertura dos testes da segunda semana é significativamente maior do que na quarta semana, provavelmente, isso deve-se pois na quarta semana existe muito mais código para ser testado.

No entanto, estas são as medidas de localização das amostras e não da população, não sendo suficientes para tirar conclusões.

Antes da realização do t-test, primeiro precisa-se que um conjunto de pressupostos sejam cumpridos, para isso, realizou-se um teste de variâncias, pois é preciso especificar no teste se as variâncias são iguais ou não, já que estamos a fazer um teste com duas variáveis, e um teste de normalidade, pois para poder fazer um t-test, precisa-se que os dados sigam uma distribuição normal, e como temos menos que trinta amostras, não podemos assumi-la.

$$H0 : \sigma^2_{semana2} = \sigma^2_{semana4} \qquad vs : \qquad H1 : \sigma^2_{semana2} \neq \sigma^2_{semana4}$$

```
> # Teste de igualdade de variâncias (usando o teste de Fligner-Killeen, uma alternativa robusta)
> var_test_result <- var.test(semana2, semana4, alternative = "two.sided")
> # Exibir o resultado do teste de igualdade de variâncias
> var_test_result

        F test to compare two variances

data:  semana2 and semana4
F = 7.572, num df = 9, denom df = 10, p-value = 0.003959
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  2.003721 30.014340
sample estimates:
ratio of variances
          7.571988
```

FIGURE 4. Output do teste da variância das taxas de cobertura dos testes da segunda e da quarta semana

Na Figura 4 verifica-se que o p-value do teste de variâncias(0.004) é menor que que o nível de significância estipulado de 5%, assim, rejeitando a hipótese nula, isto quer dizer que existem evidências estatísticas que as variâncias das taxas de cobertura das duas semanas são diferentes. Com esta informação, sabemos que, se um t-test for usado, precisa-se de especificar que as variâncias não são iguais

H0: A taxa de cobertura da semana i segue uma distribuição normal
vs:
H1: A taxa de cobertura da semana i não segue uma distribuição normal

$i \in \{2,3,4\}$

```
> # Teste de normalidade
> shapiro_test_semana2 <- shapiro.test(semana2)
> shapiro_test_semana4 <- shapiro.test(semana4)
> # Exibir os resultados dos testes de normalidade
> shapiro_test_semana2

        Shapiro-Wilk normality test

data:  semana2
W = 0.65219, p-value = 0.0002332

> shapiro_test_semana4

        Shapiro-Wilk normality test

data:  semana4
W = 0.95752, p-value = 0.7405
```

FIGURE 5. Output do teste da normalidade dos dados das taxas de cobertura dos testes da segunda e da quarta semana

Na Figura 5 verifica-se que na segunda semana o p-value(0.002) é menor que o nível de significância estipulado de 5%, assim a hipótese nula é rejeitada, isto quer dizer que existem evidências estatísticas que os dados sobre as taxas de cobertura da segunda semana não seguem uma distribuição normal, portanto, já não se pode usar o t-test, pois os pressupostos do mesmo não são cumpridos, vai ter de usar um teste não paramétrico para o substituir, pois estes não precisam de cumprir tantos pressupostos. Para substituir o t-test, optou-se por usar um teste de Wilcoxon, pois é um teste não paramétrico bom para comparar duas amostras independentes, mas ao contrário do t-test que testava as médias, este testa as medianas.

```
> wilcox.test(semana2_taxaCobertura, semana4_taxaCobertura, alternative = "two.sided",  exact = FALSE)

        Wilcoxon rank sum test with continuity correction

data:  semana2_taxaCobertura and semana4_taxaCobertura
W = 68, p-value = 0.3652
alternative hypothesis: true location shift is not equal to 0
```

FIGURE 6. Output do teste de Wilcoxon sobre os dados das taxas de cobertura dos testes da segunda e da quarta semana

$$H0 : \eta_{semana2} = \eta_{semana4} \qquad vs : \qquad h1 : \eta_{semana2} \neq \eta_{semana4}$$

Na Figura 6, verifica-se que o p-value dos teste de Wilcoxon(0.37) é maior que o nível de significância estipulado de 5%, assim a hipótese nula não é rejeitada, concluindo-se que não existem evidencias estatísticas que as medianas das taxas de cobertura da segunda semana e da quarta semana sejam diferentes. Como o a hipótese nula não foi rejeitada, então não será necessário fazer um teste unilateral para saber qual a semana com maior e menor mediana.

Por fim, para a realização dos testes de proporção foi usado um cálculo para somar todos elementos cuja taxa de cobertura fosse superior a 80%, esse valor é definido no Quality Gate para os trabalhos da disciplina de Engenharia de Software II.

```
> sucess2
[1] 6
> sucess4
[1] 6
>
```

FIGURE 7. Número de casos onde a taxa de cobertura é maior que 80% nas duas semanas

```
> resultado_teste

        2-sample test for equality of proportions with continuity correction

data:  c(sucess2, sucess4) out of c(length(semana2_taxaCobertura), length(semana4_taxaCobertura))
X-squared = 1.9949e-31, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.4228231  0.5319140
sample estimates:
   prop 1    prop 2
0.6000000 0.5454545
```

FIGURE 8. Output do testes de proporção sobre os dados das taxas de cobertura dos testes da segunda e da quarta semana

$$\text{H0: } P_1 = P_2$$
$$\text{vs:}$$
$$\text{H1: } P_1 \neq P_2$$

p_1 - grupos com taxas de cobertura superiores a 80% na primeira semana
P_2 - grupos com taxas de cobertura superiores a 80% na segunda semana

Observou-se na Figura 7 que, tanto na segunda semana, como na quarta semana houveram 6 casos onde as taxas de cobertura foram maiores ou iguais a 80%. Com o teste de proporção observa-se que na segunda semana tem uma proporção de 0.60 e já na quarta semana tem uma proporção de 0.55, e que existe um intervalo de confiança de 95% para a diferença nas proporções que varia entre aproximadamente -0.42 a 0.53. Algo importante a destacar, é que embora a segunda e a quarta semana tenham o mesmo número de casos onde a taxa de cobertura é maior que 80%, as proporções de ambas são diferentes, isto deve-se ao facto que as ficaram com tamanhos diferentes depois que os valores omissos foram descartados. A segunda semana tem 10 valores, já a quarta semana tem 11 valores.

No output da Figura 7 pode-se ver que o p-value (1.00) é maior que o nível de significância estabelecido de 5%, não rejeitando a hipótese nula, mostrando que existem evidências estatísticas que a proporção da primeira semana é igual à proporção da segunda semana

Para finalizar, para verificar se os grupos atenderam à métrica definida no Quality Gate para os trabalhos de Engenharia de Software II de 80%, foi feito um teste de Wilcoxon onde se vai comprar se a mediana das taxas de cobertura dos testes da quarta semana foi inferior a 80%.

$$H0 : \eta_{semana2} = 80\% \qquad vs : \qquad H1 : \eta_{semana2} < 80\%$$

```
> wilcox.test(semana4_taxaCobertura, alteranitve="less", mu = 0.8, conf.level = 0.95, exact = FALSE)

        Wilcoxon signed rank test with continuity correction

data:  semana4_taxaCobertura
V = 27, p-value = 1
alternative hypothesis: true location is not equal to 0.8
```

FIGURE 9. Teste de Wilcoxon para saber se os grupos respeitaram o Quality Gate para a taxa de cobertura dos testes

A partir da Figura 9 consegue-se ver que o p-value (1.00) é maior que o nível de significância estabelecido de 5%, logo a hipótese nula não é rejeitada, havendo evidências estatísticas que a taxa de cobertura dos testes da quarta semana é igual a 80%

Conclusão: Com os testes realizados, provou-se que não existe evolução significativa na taxa de cobertura dos testes entre a segunda e a quarta semana, e que a maior parte dos alunos respeitou a métrica defendia no Quality Gate de Engenharia de Software II, onde se deve ter uma taxa de cobertura dos testes de no mínimo 80%

Questão de Investigação 2

Na primeira questão, após serem analisados os dados das taxas de cobertura dos testes da segunda e da quarta semana, observou-se que segundo esses dados a taxa de cobertura da segunda para a quarta semana diminuiu, e afirmou-se se que provavelmente isso aconteceu pois havia mais código para testar, portanto para esta questão de investigação vai-se **verificar se houve um aumento no número de linhas de código totais (M18), e se elas estão correlacionadas com a taxa de cobertura dos testes (M24)**

Para verificar se houve um aumento no número de linhas de código totais, começou-se por separar o número de linhas de código escritas na segunda, terceira e quarta semana e escreveu-se um sumário para cada uma e um boxplot para estudar os dados das três amostras. Não serão estudadas o número de linhas de código totais da primeira semana, pois, como ainda não era preciso escrever código para o trabalho de Engenharia de Software II, a maior parte dos grupos deixou esse campo em branco.

TABLE III. Medidas de localização do número de linhas de código totais da segunda semana

Linhas de Código - 2ª semana					
Min	1º Quartil	Mediana	Média	3º Quartil	Max
0	265	505	665,5	898	2500

TABLE IV. Medidas de localização do número de linhas de código totais da terceira semana

Linhas de Código - 3ª semana					
Min	1º Quartil	Mediana	Média	3º Quartil	Max
191	415	552,5	1040,9	853,8	3828

TABLE V. Medidas de localização do número de linhas de código totais da quarta semana

Linhas de Código - 4ª semana					
Min	1º Quartil	Mediana	Média	3º Quartil	Max
15	393	1063	1442,7	2042	4365

Como dá para ver pelos sumários das Tabelas III, IV, V, e sobretudo no boxplot da Figura 10, consegue-se ver que consoante as semanas passavam, o número de linhas de código aumentavam.

Ao analisar o sumário do número de linhas de código totais na segunda semana (Tabela III), percebe-se que o primeiro quartil (Q_1) é de 265.0, que pelo menos 25% dos grupos tiveram pelo menos 265 linhas de código totais o segundo quartil (Q_2, ou mediana) é de 505.0, que 50% dos grupos tiveram pelo menos 505 linhas de código totais, e o terceiro quartil (Q_3) atinge 898.0, que pelo menos 75% dos grupos tiveram pelo menos 898 linhas de código totais. Já na terceira semana (Tabela IV), consegue-se ver que houve um aumento significativo nos quartis, com $Q_1 = 415.0$, $Q_2 = 552.5$, e $Q_3 = 853.8$. No final, na quarta semana (Tabela V), os valores continuaram a crescer com os quartis $Q_1 = 393.5$, $Q_2 = 1063.0$, e $Q_3 = 2042.0$.

Estas observações são confirmadas visualmente com o boxplot da Figura 10, onde é possível perceber uma tendência no aumento no número de linhas de código ao longo das semanas, podendo ver isso através da linha mais escura no maio das caixas, que é a mediana, que vai aumentando consoante as semanas passavam. Este comportamento ascendente sugere que o desenvolvimento de código tornou-se mais extenso à medida que o projeto avançava, possibilitando uma correlação entre o tempo e o número de linhas de código. Outro ponto que também é possível ver a partir do boxplot, é que a caixa da terceira semana é muito maior do que a da segunda e terceira semana, isso significa que existe uma maior dispersão nos dados da quarta semana, que tem uma amplitude inter-quartil maior

No entanto, estes dados refletem apenas as medidas de localização das amostras, não representado os valores reais da população, portanto vai ter de se fazer mais testes para verificar se houve mesmo um aumento do número de linhas de código totais. Para fazer esse estudo, decidiu-se fazer um teste de ANOVA, pois, ele permite comparar as médias populacionais de mais de dois grupos, no entanto para poder usar esse teste é preciso saber se as amostras seguem uma distribuição normal, para isso vai-se fazer o teste de Shapiro-Wilk para as três amostras.

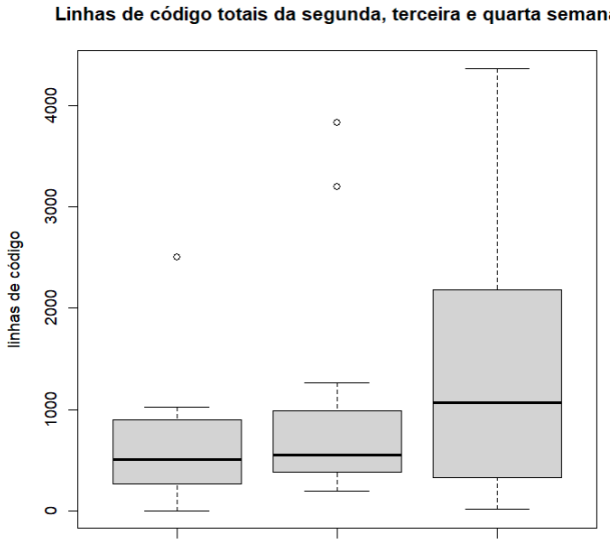

FIGURE 10. BoxPlot do número de linhas de código totais da 2ª, 3ª e 4ª semana

```
> shapiro.test(linhasCodigoSegundaSemana)

        Shapiro-Wilk normality test

data:  linhasCodigoSegundaSemana
W = 0.809, p-value = 0.01236
```

FIGURE 11. Teste de normalidade aos dados do número de linhas de código da segunda semana

```
> shapiro.test(linhasCodigoTerceiraSemana)

        Shapiro-Wilk normality test

data:  linhasCodigoTerceiraSemana
W = 0.66459, p-value = 0.0003924
```

FIGURE 12. Teste de normalidade aos dados das linhas de código totais da terceira semana

```
> shapiro.test(linhasCodigoQuartaSemana)

        Shapiro-Wilk normality test

data:  linhasCodigoQuartaSemana
W = 0.87123, p-value = 0.06775
```

FIGURE 13. Teste de normalidade aos dados do número de linhas de código totais da quarta semana

H0: O número de linhas de código totais da semana i seguem uma distribuição normal

vs:

H1: O número de linhas de código totais da semana i não seguem uma distribuição normal

$i \in \{2,3,4\}$

Nas Figuras 11 e 12, consegue-se ver que o p-value (0.01236 para a segunda semana e 0.0003924 para a terceira semana) é menor que o nível de significância estabelecido de 5%, rejeitando a hipótese nula, existindo evidências estatísticas que os dados sobre o número de linhas de código totais da segunda e terceira semana não seguem uma distribuição normal. No entanto, na Figura 13, o p-value (0.06775) é maior que o nível de significância estabelecido de 5%, não rejeitando a hipótese nula, existindo evidências estatísticas que os dados sobre o número de linhas de código da quarta semana seguem uma distribuição normal. Como apenas os dados da quarta semana seguem uma distribuição normal, já não dá para usar um teste de ANOVA, assim, vai ter de ser substituído por um teste não paramétrico, sendo o teste escolhido o de Kruskal-Wallis, visto que as amostras são independentes, nele é testado a igualdade das medianas de todos os grupos, ao contrário do teste de ANOVA que testava as médias. Caso as amostras fossem emparelhadas, ter-se-ia de usar um teste de Friedman.

$$H0 : \eta_{semana2} = \eta_{semana3} = \eta_{semana4} \qquad vs : \qquad h1 : \eta_{semana2} \neq \eta_{semana3} \neq \eta_{semana4}$$

```
> kruskal.test(list(linhasCodigoSegundaSemana, linhasCodigoTerceiraSemana, linhasCodigoQuartaSema
na), na.action = na.omit)

        Kruskal-Wallis rank sum test

data:  list(linhasCodigoSegundaSemana, linhasCodigoTerceiraSemana, linhasCodigoQuartaSemana)
Kruskal-Wallis chi-squared = 1.5319, df = 2, p-value = 0.4649
```

FIGURE 14. Teste de Kruskall-Wallis para comparar as medianas da segunda, terceira e quarta semana

A partir do output da Figura 14, consegue-se ver que o p-value (0.4649) é maior que o nível de significância estabelecido de 5%, logo a hipótese nula não é rejeitada, verificando-se que ao contrário do que foi visto na análise análise das amostras, existem evidências estatísticas que as medianas do número de linhas de código totais das três semanas são estatisticamente iguais.

Após se verificar que não houve aumento significativo no número de linhas de código totais, vai-se confirmar se estas correlacionam com a taxa de cobertura dos testes, e provar se a afirmação da primeira pergunta é válida.

Para poder estudar a validade dessa afirmação, vai ter de se fazer um teste de correlação para saber se os dados sobre o número de linhas de código totais estão correlacionados com os dados da taxa de cobertura dos testes, mas antes é preciso fazer um teste de normalidade às duas amostras para saber se se vai usar um teste de correlação de Pearson ou um teste de correlação de Spearman.

Nos dados a cerca do número de linhas de código totais e da taxa de cobertura dos testes, foram ignorados os dados da primeira semana, pois nessa altura ainda não era preciso escrever código, havendo assim muito pouca informação, e a escassa informação que tem pode não ser credível.

H0: O número de linhas de código totais seguem uma distribuição normal
vs:
H1:O número de linhas de código totais não seguem uma distribuição normal

```
> shapiro.test(numeroLinhasCodigo)

        Shapiro-Wilk normality test

data:  numeroLinhasCodigo
W = 0.77655, p-value = 4.548e-06
```

FIGURE 15. Teste de normalidade de Shapiro-Wilk para os dados sobre o numero de linhas de código totais

```
> shapiro.test(taxaCoberturaTestes)

        Shapiro-Wilk normality test

data:  taxaCoberturaTestes
W = 0.74413, p-value = 5.583e-06
```

FIGURE 16. Teste de normalidade de Shapiro-Wilk para os dados sobre a taxa de cobertura dos testes

H0: A taxa de cobertura dos testes seguem uma distribuição normal
vs:
H1:A taxa de cobertura dos testes não seguem uma distribuição normal

Como dá para ver nas Figuras 15 e 16, os p-values dos dois testes ($4.548 * 10^{-6}$ e $5.583 * 10^{-6}$) são menores que o nível de significância estabelecido de 5%, rejeitando a hipótese nula, comprovando que existem evidências estatísticas que os dados do número de linhas de código totais e da taxa de cobertura dos testes não seguem uma distribuição normal, assim será preciso usar um teste de correlação de Spearman, pois este é uma alternativa não paramétrica ao teste de Pearson, que precisa que os dados sigam uma distribuição normal.

```
> cor.test(numeroLinhasCodigo, taxaCoberturaTestes, method = "spearman", conf.level = 0.95, na.action = na.omit, exact = FALSE)

        Spearman's rank correlation rho

data:  numeroLinhasCodigo and taxaCoberturaTestes
S = 3753.3, p-value = 0.1872
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.243296
```

FIGURE 17. Teste de correlação de Spearman para os dados do número de linhas de código totais e para a taxa de cobertura dos testes

H0: Coeficiente de correlação é igual a 0
vs:
H1: Coeficiente de correlação é diferente a 0

A partir do output da Figura 17 é possível verificar que o p-value (0.1872) é menor que o nível de significância estabelecido de 5%, não rejeitando a hipótese nula, existindo evidências estatísticas que o coeficiente de correlação é igual a 0. O coeficiente de correlação de Spearman, denominado por ρ (rho), é um número que varia entre -1 e 1. Quanto mais próximo dos extremos (-1 ou 1), maior é a força da correlação. Já os valores próximos de 0 implicam em correlações mais fracas ou inexistentes. No output apresentado, é possível ver que o coeficiente de correlação (0.243296) está muito próximo de zero.

Outra forma de ver a correlação entre os dados é através de um gráfico de dispersão, assim é possível identificar padrões e tendências nos mesmos, servindo como guia visual entre a relação entre as duas variáveis. Na Figura 18, é apresentado um gráfico de dispersão que representa os pontos correspondentes ao número de linhas de código totais no eixo x, e à taxa de cobertura dos testes no eixo y, onde cada ponto no gráfico representa uma observação. A inclinação geral da dispersão dos pontos pode indicar a direção da relação entre as variáveis: se os pontos estão inclinados para cima, sugere uma correlação positiva, enquanto uma inclinação para baixo indica uma correlação negativa.

A adição da linha de regressão linear, representada vermelho, ajuda a destacar a tendência geral dos dados, mesmo que, segundo o resultado do teste de correlação de Spearman, não tenha sido encontrada uma correlação significativa entre as variáveis.

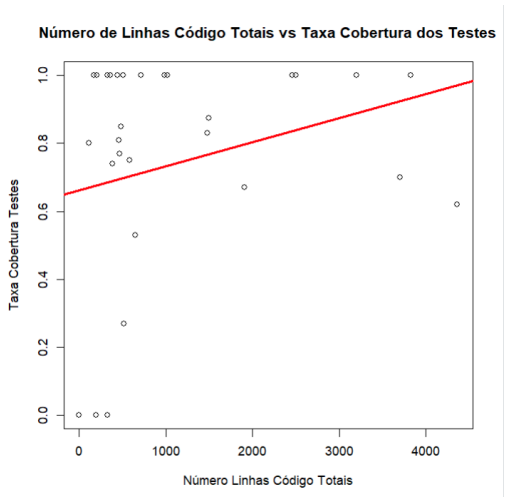

FIGURE 18. Gráfico de dispersão dos dados do número linhas de código totais com a taxa de cobertura dos testes

Conclusão: Com os testes realizados, provou-se que para além de não ter havido uma evolução significativa no número de linhas de código totais no decorrer do projeto de Engenharia de Software II, também se provou que a afirmação feita na primeira pergunta não tem validade, os dados sobre o número de linhas de código totais e a taxa de cobertura dos testes não estão correlacionados

Questão de Investigação 3

A terceira questão de investigação trata-se de **saber os fatores que aumentam ou diminuem o número de testes falhados (M9)**, para isso, será preciso montar um modelo de regressão linear para saber quais as variáveis que influenciam o valor da variável *M*9.

Para montar o modelo de regressão linear, primeiramente realizou-se uma matriz de correlações com todas as variáveis das recolhas, com essa matriz, é possível visualizar quais variáveis têm uma maior correlação com quais variáveis. Como existem muitas variáveis onde os dados não seguem uma distribuição normal, decidiu-se usar o método de *Spearman* para a montagem da matriz

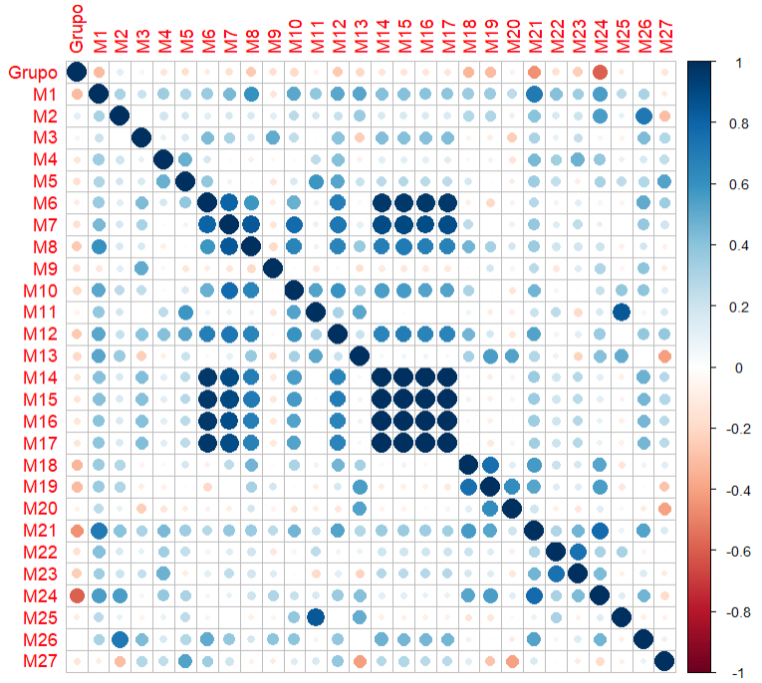

FIGURE 19. Matriz de correlações de todos os dados recolhidos

Como dá para ver na Figura 19, as correlações para a variável M9, o número de testes falhados, são bastante fracas. O nível de correlações consegue ser visualizado a partir da cor, onde, quanto mais clara é a cor indica valores próximos de zero ou correlações mais fracas, enquanto que cores mais escuras indicam valores mais extremos ou correlações mais fortes.

Mesmo com correlações bastante fracas, vai-se tentar montar um modelo de regressão linear com as variáveis *M*03, *M*04, *M*05 e *M*22, pois, dentro de todas as correlações vistas na coluna *M*09 na matriz de correlações, estas são as que têm uma maior correlação.

- *M*03 - Número de User Stories abertas
- *M*04 - Número de User Stories fechadas
- *M*05 - Número de pedidos de alterações abertos
- *M*22 - Número de correções necessárias por tempo de ciclo

Como a questão em estudo trata-se de fatores que aumentam e diminuem o número de testes, a melhor forma para o analisar seria através de um modelo de regressão linear, neste caso em questão, uma regressão linear múltipla, onde numa primeira etapa foi montado da seguinte forma: a variável *M*09 é a variável dependente, e as variáveis *M*03, *M*04, *M*05 e *M*22 são as variáveis independentes.

Modelo de regressão linear:

$$M09 = a0 + a1 * M03 + a2 * M04 + a3 * M05 + a4 * M22 + \varepsilon$$

```
> summary(regression_model)

Call:
lm(formula = data$M9 ~ data$M3 + data$M4 + data$M5 + data$M22,
    data = data, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19376 -0.25056  0.05371  0.16246  2.39322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.16246    0.30128  -0.539   0.5972
data$M3      0.19375    0.07282   2.661   0.0171 *
data$M4      0.02553    0.11280   0.226   0.8238
data$M5     -0.10765    0.11592  -0.929   0.3668
data$M22    -0.06096    0.09010  -0.677   0.5083
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8767 on 16 degrees of freedom
Multiple R-squared:  0.3275,    Adjusted R-squared:  0.1594
F-statistic: 1.948 on 4 and 16 DF,  p-value: 0.1514
```

FIGURE 20. Primeiro modelo de correlação

Na Figura 20 é possível fazer um teste global para saber se o modelo é válido, para isso, tem que se olhar para a secção F-statistic e verificar se o p-value de lá é menor que o nível de significância estabelecido, como o p-value (0.1514) é maior que o nível de significância estabelecido, isso significa que a hipótese nula não é rejeitada, havendo evidências estatísticas que as variáveis independentes não têm um efeito significativo sobre a variável dependente

Teste Global:

$$H0 : a_1 = a_2 = ... = a_i = 0 \qquad \text{vs.} \qquad H1 : \text{existe pelo menor um i, onde } a_i \neq 0$$

Para encontrar um modelo de regressão válido, ter-se-ia que descartar variáveis dependentes e executar de novo o modelo de regressão linear sucessivamente até encontrar o modelo mais adequado, para fazer isso utilizou-se uma técnica de seleção de variáveis denominada de *stepwise*.

O método *stepwise* tenta automatizar o processo de seleção das variáveis mais otimizadas para o modelo de regressão linear, considerando critérios como AIC (Akaike Information Criterion) ou BIC (Bayesian Information Criterion). Dessa forma, ao aplicar o *stepwise* o modelo será ajustado consecutivamente, adicionando ou removendo variáveis até encontrar um modelo adequado.

Como é possível ver na Figura 21, a função iterou até encontrar as melhores variáveis para o modelo de regressão linear, que coincidentemente foi a única variável que mostrou impactar a variável dependente no antigo modelo de regressão ($M3$), tendo um p-value menor que o nível de significância estabelecido (Teste Marginal).

Modelo de regressão linear final:

$$M09 = a0 + a1 * M03 + \varepsilon$$

```
> stepwise_model <- step(regression_model)
Start:  AIC=-1.24
data$M9 ~ data$M3 + data$M4 + data$M5 + data$M22

           Df Sum of Sq    RSS     AIC
- data$M4   1    0.0394 12.336 -3.1717
- data$M22  1    0.3519 12.649 -2.6464
- data$M5   1    0.6629 12.960 -2.1363
<none>                  12.297 -1.2388
- data$M3   1    5.4407 17.738  4.4543

Step:  AIC=-3.17
data$M9 ~ data$M3 + data$M5 + data$M22

           Df Sum of Sq    RSS     AIC
- data$M22  1    0.3150 12.651 -4.6422
- data$M5   1    0.6541 12.990 -4.0868
<none>                  12.336 -3.1717
- data$M3   1    5.4491 17.785  2.5108

Step:  AIC=-4.64
data$M9 ~ data$M3 + data$M5

          Df Sum of Sq    RSS     AIC
- data$M5  1    0.9216 13.573 -5.1656
<none>                 12.651 -4.6422
- data$M3  1    5.2891 17.940  0.6931

Step:  AIC=-5.17
data$M9 ~ data$M3

          Df Sum of Sq    RSS     AIC
<none>                 13.573 -5.1656
- data$M3  1    4.7129 18.286 -0.9064
```

FIGURE 21. Método Stepwise para selecionar as melhores variáveis para o modelo de regressão linear

```
> summary(regression_model)

Call:
lm(formula = data$M9 ~ data$M3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0121 -0.4800 -0.1252  0.2295  2.6332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.22950    0.27249  -0.842   0.4101
data$M3      0.17737    0.06905   2.569   0.0188 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8452 on 19 degrees of freedom
Multiple R-squared:  0.2577,    Adjusted R-squared:  0.2187
F-statistic: 6.597 on 1 and 19 DF,  p-value: 0.0188
```

FIGURE 22. Modelo de regressão linear final, com as variáveis indicadas pelo stepwise

Ao remover as restantes variáveis do modelo, ficando apenas a variável *M*3, o modelo passou de uma regressão linear múltipla para uma regressão linear simples. Agora, para provar a validade do modelo, irão ser feitas as seguintes validações:

- Testes globais e Marginais

- Análise de Resíduos
- Interpretação do R^2

Teste Global

O Teste Global é usado para avaliar conjuntamente a influência de todas as variáveis independentes no modelo da regressão. Para um modelo de regressão linear simples o teste de hipótese é formulado da seguinte maneira:

$$H_0 : a_1 = 0$$
$$\text{vs.}$$
$$H_1 : a_1 \neq 0$$

A hipótese nula H_0 diz que todos os coeficientes a_i são iguais a zero, indicando que as variáveis independentes não têm efeito significativo na variável dependente. Já a hipótese alternativa H_1 sugere que pelo menos uma variável independente tem um efeito significativo no modelo. Para obter o resultado deste teste precisa-se de ver o p-value que está presente na secção do "F-statistic" na Figura 22 caso o p-value lá for menor que o alpha pré-definido (5%), o que não é o caso, então rejeitar-se-ia a hipótese nula, logo não existe pelo menos um i tal que $a_i \neq 0$

Testes Marginais

Nos testes marginais, são apenas os testes que se realizaram na análise da regressão, onde se analisa o p-value de cada uma das variáveis independentes, verificando se todas elas têm um p-value ($Pr(>|t|)$) menor que o nível de significância (α), caso alguma seja maior, o teste marginal falha.

Para cada $i = 1, 2, 3, ..., n$, o teste individual é formulado da seguinte maneira:

$$H_0 : a_i = 0$$
$$\text{vs:}$$
$$H_1 : a_i \neq 0$$

No modelo de regressão linear final, como só tem uma variável, e como o p-value dessa variável (0.0014) é menor que o nível de significância estabelecido de 5%, então a hipótese nula é rejeitada, e mais mais uma das validações é realizada.

Análise de resíduos - Teste de normalidade

O teste de normalidade pode ser realizado a partir de um teste de Shapiro-Wilk, onde se vai testar a normalidade dos resíduos da regressão, pois um dos pressupostos de uma regressão linear é que os seus resíduos sigam uma distribuição normal. O output do teste realizado pode ser visto na Figura 23, onde se consegue ver que os resíduos do modelo de regressão linear não seguem uma distribuição normal, logo não passam no teste. O insucesso desse teste pode ser visto a partir do p-value(0.0009263), onde um valor menor que o nível de significância pré-estabelecido (5%), significa se rejeita a hipótese nula, existindo evidências estatísticas que os resíduos não seguem uma distribuição normal.

Teste de Shapiro-Wilk:

H0: os resíduos seguem uma distribuição normal
vs:
H1: os resíduos não seguem uma distribuição normal

```
> shapiro.test(regression_model$residuals)

        Shapiro-Wilk normality test

data:  regression_model$residuals
W = 0.80969, p-value = 0.0009263
```

FIGURE 23. Output do teste de Shapiro-Wilk para os resíduos da regressão linear

Outra maneira de verificar a normalidade dos resíduos é através de uma representação gráfica da distribuição dos resíduos pela construção de um histograma. No contexto da regressão linear, os resíduos ideais seguiriam uma distribuição normal, resultando em um histograma simétrico com uma forma de um sino, onde os dados se concentram mais no centro, que seria a média. Ao criar um histograma dos resíduos, é possível visualizar a sua distribuição e identificar se ela se assemelha a uma distribuição normal.

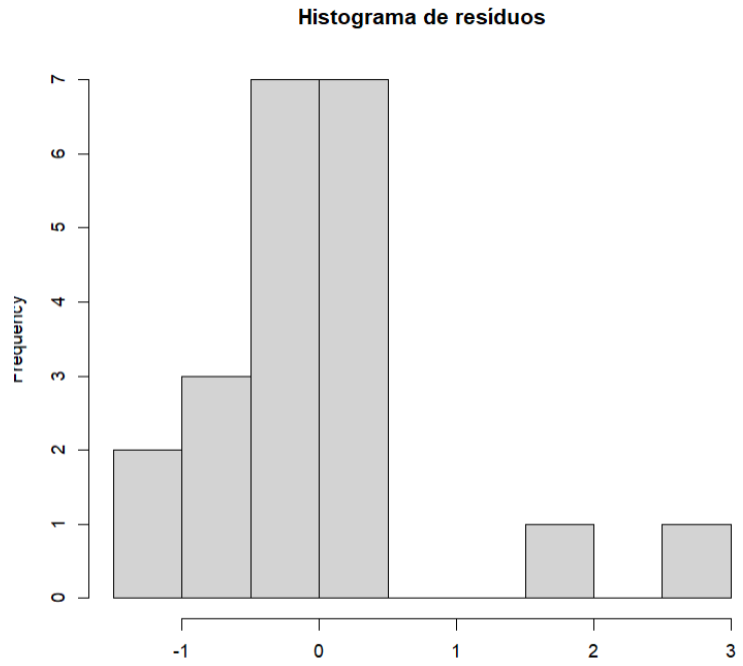

FIGURE 24. Histograma da regressão Linear

Como dá para ver na Figura 24, mais uma vez foi comprovado que os resíduos da regressão linear não seguem uma distribuição normal, pois o histograma exibe uma assimetria significativa e não apresenta uma forma aproximadamente gaussiana

Análise de resíduos - Testes de resíduos

Para os testes de resíduos, pode-se usar um teste de Durbin Wattson, onde se vai testar se os resíduos estão relacionados, se estiverem, então a regressão linear não é válida, pois numa regressão linear, para além dos resíduos terem de seguir uma distribuição normal, eles não devem estar correlacionados. No Output da Figura 25 dá para ver que o p-value (0.002) é menor que o nível de significância pré-estabelecido (5%), então a hipótese nula é rejeitada, existindo evidências estatísticas que os resíduos não estão correlacionados, fazendo com que o teste não falhe.

Teste de Durbin Wattson:

H0: a auto-correlação dos resíduos é igual a 0
vs:
H1: a auto-correlação dos resíduos é diferente a 0

ou

H0: resíduos estão correlacionados
vs:
H1: resíduos não estão correlacionados

```
> durbinWatsonTest(regression_model)
 lag Autocorrelation D-W Statistic p-value
   1       0.3430393     0.7517375   0.002
 Alternative hypothesis: rho != 0
```

FIGURE 25. Output do teste de Durbin Watson para a correlação dos resíduos

Interpretação do R^2

O R^2 é uma medida de quão bem o modelo se ajusta aos dados. A interpretação do R^2 é essencialmente uma medida da variabilidade explicada pelo modelo em relação à variabilidade total dos dados. Um R^2 de 1.0 indica que o modelo explica toda a variabilidade, enquanto um R^2 de 0 indica que o modelo não explica nada.

Para a análise do R^2 da regressão linear proposta anteriormente, precisamos de ir à Figura 22, na secção *"Multiple R-squared: "*, lá é possível ver que o R^2 é igual a 0.2577 ou 26%, que é um valor relativamente pequeno, o que significa que o modelo ajusta-se mal aos dados.

Conclusão: com o modelo de regressão linear final, pode-se afirmar que a variável $M3$ (número de User Stories abertas) correlaciona com a variável $M9$ (número de testes falhados, por sprint), e ao ter um coeficiente de 0.17737, significa que ao aumento de um valor na variável M4, leva a um aumento de 0.17737 na variável $M9$.

Questão de Investigação 4

Para a quarta questão de investigação, decidiu-se investigar os **fatores que fazer aumentar ou diminuir a métrica do PMD do Class Cohesion (M17)**, para isso, tal como na terceira questão de investigação, será preciso montar um modelo de regressão linear para saber quais variáveis influenciam o valor da variável $M17$.

Para poder montar o modelo de regressão linear, primeiramente observou-se a matriz de correlações realizada na terceira questão de investigação (Figura 19), e a partir daí foi possível observar que as variáveis $M6$, $M7$, $M14$, $M15$ e $M16$ têm uma forte correlação com a variável $M17$, logo são boas candidatas para o modelo de regressão linear.

- $M06$ - número de pedidos de alterações rejeitados
- $M07$ - número de pedidos de alterações aprovados
- $M14$ - Cyclomantic Complexity, número de items menores que 10
- $M15$ - Cognitive Complexity, número de items menores que 10
- $M16$ - Class Coupling

Modelo de regressão linear:

$$M17 = a0 + a1 * M06 + a2 * M07 + a3 * M14 + a4 * M15 + a5 * M16 + \varepsilon$$

```
> summary(regression_model)

Call:
lm(formula = data$M17 ~ data$M6 + data$M7 + data$M14 + data$M15 +
    data$M16, data = data, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1508 -0.9581  1.0688  1.0688  1.4852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.06879    0.49408  -2.163  0.04708 *
data$M6      0.85156    2.78515   0.306  0.76400
data$M7     -0.13315    0.21304  -0.625  0.54137
data$M14     0.64960    0.17082   3.803  0.00173 **
data$M15     0.03762    0.20582   0.183  0.85742
data$M16     0.32246    0.16248   1.985  0.06579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.017 on 15 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9989
F-statistic:  3648 on 5 and 15 DF,  p-value: < 2.2e-16
```

FIGURE 26. Regressão linear múltipla com as variáveis tiradas da matriz de correlações

Na Figura 26 é possível fazer um teste global para saber se o modelo é válido, para isso, tem que se olhar para a secção F-statistic e verificar se o p-value de lá é menor que o nível de significância estabelecido, como o p-value (0.1514) é maior que o nível de significância estabelecido, isso significa que a hipótese nula não é rejeitada, havendo evidências estatísticas que as variáveis independentes não têm um efeito significativo sobre a variável dependente

Teste Global:

$$H0 : a_1 = a_2 = ... = a_i = 0 \qquad \text{vs.} \qquad H1 : \text{existe pelo menor um i, onde } a_i \neq 0$$

Para encontrar um modelo de regressão válido, ter-se-ia que descartar variáveis dependentes e executar de novo o modelo de regressão linear sucessivamente até encontrar o modelo mais adequado, para fazer isso utilizou-se uma técnica de seleção de variáveis denominada de *stepwise*.

```
> stepwise_model <- step(regression_model)
Start:  AIC=34.39
data$M17 ~ data$M6 + data$M7 + data$M14 + data$M15 + data$M16

           Df Sum of Sq     RSS    AIC
- data$M15  1     0.136  61.134 32.440
- data$M6   1     0.380  61.379 32.523
- data$M7   1     1.588  62.587 32.933
<none>                   60.998 34.393
- data$M16  1    16.017  77.015 37.289
- data$M14  1    58.811 119.809 46.569

Step:  AIC=32.44
data$M17 ~ data$M6 + data$M7 + data$M14 + data$M16

           Df Sum of Sq     RSS    AIC
- data$M6   1     0.596  61.730 30.643
- data$M7   1     1.931  63.066 31.093
<none>                   61.134 32.440
- data$M16  1    27.056  88.191 38.135
- data$M14  1    80.225 141.360 48.043

Step:  AIC=30.64
data$M17 ~ data$M7 + data$M14 + data$M16

           Df Sum of Sq     RSS    AIC
- data$M7   1     4.263  65.993 30.045
<none>                   61.730 30.643
- data$M16  1    26.470  88.200 36.137
- data$M14  1   101.432 163.162 49.055

Step:  AIC=30.05
data$M17 ~ data$M14 + data$M16

           Df Sum of Sq     RSS    AIC
<none>                   65.993 30.045
- data$M16  1    40.183 106.176 38.032
- data$M14  1   105.741 171.733 48.130
```

FIGURE 27. Método stepwise para selecionar as melhores variáveis para o modelo de regressão linear

O método *stepwise* tenta automatizar o processo de seleção das variáveis mais otimizadas para o modelo de regressão linear, considerando critérios como AIC (Akaike Information Criterion) ou BIC (Bayesian Information Criterion). Dessa forma, ao aplicar o *stepwise* o modelo será ajustado consecutivamente, adicionando ou removendo variáveis até encontrar um modelo adequado.

Como é possível ver na Figura 21, a função iterou até encontrar as melhores variáveis para o modelo de regressão linear, ficando da seguinte forma:

Modelo de regressão linear:

$$M17 = a0 + a1 * M14 + a2 * M15 + \varepsilon$$

Como mostra a Figura 28, todas as variáveis independentes têm um p-value menor que o nível de significância estabelecido de 5%, existindo evidências estatísticas que todas as variáveis independentes têm impacto sobre a variável dependente.

No entanto, ainda falta comprovar se o modelo de regressão linear é valido com um conjunto de validações, sendo elas:

- Testes globais e Marginais
- Análise de Resíduos
- Interpretação do R^2
- Interpelação dos coeficientes do modelo de regressão

```
> summary(regression_model)

Call:
lm(formula = data$M17 ~ data$M14 + data$M16, data = data, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1798 -0.3663  1.1552  1.1552  1.1552

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1552     0.4539  -2.545  0.02030 *
data$M14      0.6220     0.1158   5.370 4.19e-05 ***
data$M16      0.3822     0.1154   3.311  0.00389 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.915 on 18 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.999
F-statistic: 1.011e+04 on 2 and 18 DF,  p-value: < 2.2e-16
```

FIGURE 28. Modelo de regressão linear final, com as variáveis indicadas pelo stepwise

- Existência de multicolinariedade entre as variáveis independentes

Teste Global

O Teste Global é usado para avaliar conjuntamente a influência de todas as variáveis independentes no modelo da regressão. Para cada $i = 1,2,3,...,n$, o teste de hipótese é formulado da seguinte maneira:

$$H_0 : a_1 = a_2 = a_3 = a_4 = 0$$
vs.
$$H_1 : \text{Existem pelo menos um } i \text{ tal que } a_i \neq 0$$

A hipótese nula H_0 diz que todos os coeficientes a_i são iguais a zero, indicando que as variáveis independentes não têm efeito significativo na variável dependente. Já a hipótese alternativa H_1 sugere que pelo menos uma variável independente tem um efeito significativo no modelo. Para obter o resultado deste teste precisa-se de ver o p-value que está presente na secção do "F-statistic" na Figura 28 caso o p-value lá for menor que o alpha pré-definido (5%), o que é o caso, então não se rejeita a hipótese nula, havendo evidências que existe pelo menos i tal que $a_i \neq 0$, que existe pelo menos uma variável independente com um efeito significativo sobre a variável dependente

Testes Marginais

Nos testes marginais, são apenas os testes que se realizaram na análise da regressão, onde se analisa o p-value de cada uma das variáveis independentes, e todas elas têm de ser menores que o nível de significância (α), caso alguma seja maior, o teste marginal falha.

Para cada $i = 1,2,3,...,n$, o teste individual é formulado da seguinte maneira:
$$H_0 : a_i = 0$$
vs:
$$H_1 : a_i \neq 0$$

No modelo de regressão linear final, como o p-value de todas as variáveis independentes são menores que o nível de significância estabelecido de 5%, então a hipótese nula é rejeitada, e mais mais uma das validações é realizada.

Análise de resíduos - Teste de normalidade

O teste de normalidade pode ser realizado a partir de um teste de Shapiro-Wilk, onde se vai testar a normalidade dos resíduos da regressão, pois um dos pressupostos de uma regressão linear é que os seus resíduos sigam uma distribuição normal. O output do teste realizado pode ser visto na Figura 29, onde se consegue ver que os resíduos do modelo de regressão linear não seguem uma distribuição normal, logo não passam no teste. O insucesso desse teste pode ser visto a partir do p-value($1.621 * 10^{-5}$), onde um valor menor que o nível de significância pré-estabelecido (5%), significa se rejeita a hipótese nula, existindo evidências estatísticas que os resíduos não seguem uma distribuição normal.

Teste de Shapiro-Wilk:

H0: os resíduos seguem uma distribuição normal
vs:
H1: os resíduos não seguem uma distribuição normal

```
> # teste de normalidade:
> shapiro.test(regression_model$residuals)

        Shapiro-Wilk normality test

data:  regression_model$residuals
W = 0.68122, p-value = 1.621e-05
```

FIGURE 29. Output do teste de Shapiro-Wilk para os resíduos da regressão linear

Outra maneira de verificar a normalidade dos resíduos é através de uma representação gráfica da distribuição dos resíduos pela construção de um histograma. No contexto da regressão linear, os resíduos ideais seguiriam uma distribuição normal, resultando em um histograma simétrico com uma forma de um sino, onde os dados se concentram mais no centro, que seria a média. Ao criar um histograma dos resíduos, é possível visualizar a sua distribuição e identificar se ela se assemelha a uma distribuição normal.

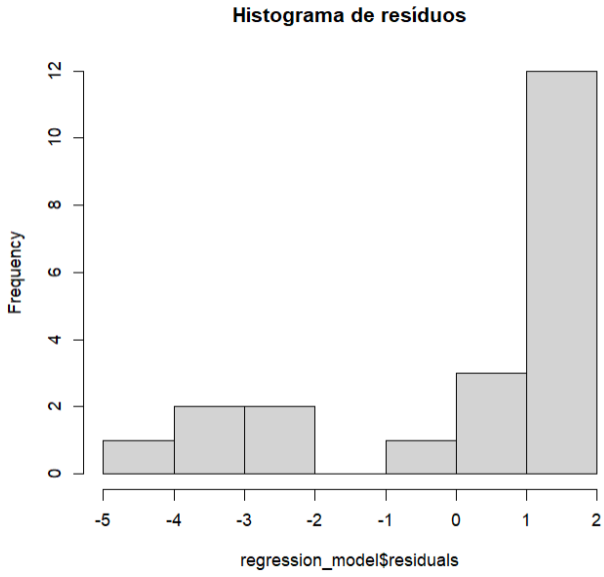

FIGURE 30. Histograma da regressão Linear

Análise de resíduos - Testes de resíduos

Para os testes de resíduos, pode-se usar um teste de Durbin Wattson, onde se vai testar se os resíduos estão relacionados, se estiverem, então a regressão linear não é válida, pois numa regressão linear, para além dos resíduos terem de seguir uma distribuição normal, eles não devem estar correlacionados. No Output da Figura 31 dá para ver que o p-value (0.786) é maior que o nível de significância pré-estabelecido (5%), então a hipótese nula não é rejeitada, existindo evidências estatísticas que os resíduos estão correlacionados, fazendo com que o teste falhe.

Teste de Durbin Wattson:

H0: a auto-correlação dos resíduos é igual a 0
vs:
H1: a auto-correlação dos resíduos é diferente a 0

ou

H0: resíduos estão correlacionados
vs:
H1: resíduos não estão correlacionados

```
> # testes de resíduos
> durbinWatsonTest(regression_model)
 lag Autocorrelation D-W Statistic p-value
   1      0.03154486      1.909897   0.786
 Alternative hypothesis: rho != 0
```

FIGURE 31. Output do teste de Durbin Watson para a correlação dos resíduos

Como dá para ver na Figura 30, mais uma vez foi comprovado que os resíduos da regressão linear não seguem uma distribuição normal, pois o histograma exibe uma assimetria significativa e não apresenta uma forma aproximadamente gaussiana

Interpretação do R^2

O R^2 é uma medida de quão bem o modelo se ajusta aos dados. A interpretação do R^2 é essencialmente uma medida da variabilidade explicada pelo modelo em relação à variabilidade total dos dados. Um R^2 de 1.0 indica que o modelo explica toda a variabilidade, enquanto um R^2 de 0 indica que o modelo não explica nada.

Para a análise do R^2 da regressão linear proposta anteriormente, precisamos de ir à Figura 28, na secção *"Multiple R-squared: "*, lá é possível ver que o R^2 é igual a 0.9991 ou 99%, que é um valor muito grande, o que significa que o modelo ajusta-se muito bem aos dados.

Existência de multicolinariedade entre as variáveis independentes

A multicolinariedade é a existência de uma forte correlação entre as variáveis independentes de uma regressão linear múltipla, a existência de multicolinariedade pode implicar numa má interpretação dos coeficientes e consequentemente na qualidade da regressão. Para medir a multicolinariedade da regressão linear calculou-se o fator de variação (*VIF*) que nos dá um valor que nos permite avaliar a multicolinariedade das variáveis independentes, onde valores altos referem-se a uma maior multicolinariedade. Como pode ser visto na Figura 32, as variáveis $M14$ e $M16$ têm um VIF bastante elevado (268.673), havendo uma forte correlação entre as duas variáveis, existindo evidências da presença de multicolinariedade na regressão linear.

```
> vif(regression_model) # usado nas regressões lineares múltiplas
data$M14 data$M16
 268.673  268.673
```

FIGURE 32. Fator de variação da regressão linear múltipla

Clusters

Para poder observar padrões na regressão linear, pode-se dividi-la em grupos ou clusters, assim, é possível fazer uma visão das relações entre as variáveis independentes e as variáveis dependentes. Para a regressão linear desta questão de investigação, vai-se usar técnicas de clustering dos resíduos, deste modo vai-se agrupar os resíduos semelhantes em clusters específicos, revelando informações importantes sobre a qualidade do modelo.

Para fazer os clusters dos resíduos da regressão linear usou-se um algoritmo denominado de *k-means*, que itera entre atribuir observações aos clusters mais próximos e recalcular os centros dos clusters, até que a atribuição de clusters se estabilize.

Antes de realizar o k-means, é necessário saber quantos clusters são necessários, isso pode ser visto ao interpretar visualmente os dados, ou pelo uso de técnicas como a que se vai usar a seguir, denominada de *método de Elbow*. A partir desse método foi gerado o gráfico da Figura 33, e lá é possível ver quantos clusters vão ser precisos a partir da linha vermelha, 3 clusters, que é quando começa a gerar uma curva no gráfico, a linha azul é um valor alternativo para o número de clusters.

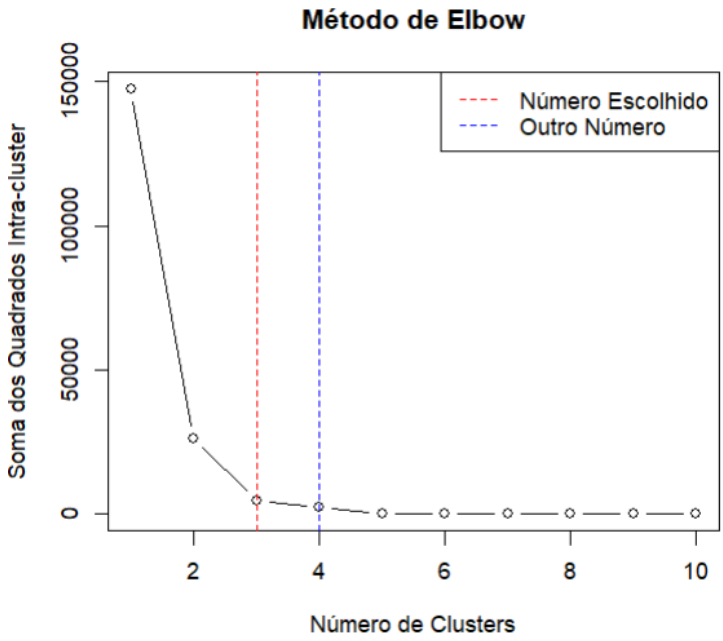

FIGURE 33. Gráfico do método de Elbow para a seleção do número de clusters

Após se saber que serão precisos no mínimo três clusters, bastou usar apenas a função *k-means* para dividir os resíduos da regressão em clusters, e depois montou-se um boxplot com cada cluster e um gráfico que mostra os dados divididos em cada cluster, a separação é vista a partir da cor de cada ponto.

Como dá para ver na Figura 35 os pontos entre os clusters encontram-se dispersos, sobretudo no terceiro cluster que só tem dois pontos, e estão muito dispersos, isso pode sugerir que a variabilidade dos resíduos não é fortemente influenciada pela associação a um determinado cluster, no segundo cluster isso também acontece, mas acaba por ser mais fácil de o identificar através da análise do boxplot da Figura 34, em que esse cluster apresenta uma caixa maior, logo uma maior variabilidade entre os resíduos do cluster. A partir do boxplot, é possível ver que as caixas dos clusters têm tamanhos bastante diferentes, isso significa que os resíduos não são homogéneos.

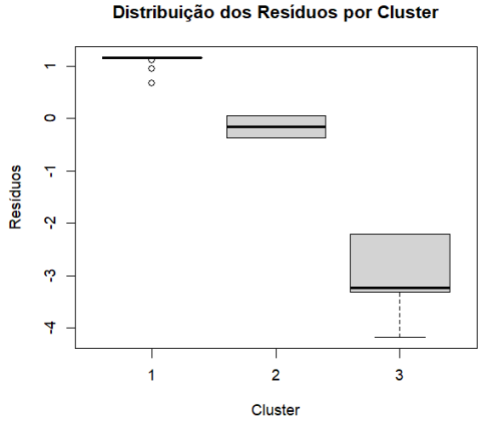

FIGURE 34. BoxPlot dos clusters gerados dos resíduos da regressão

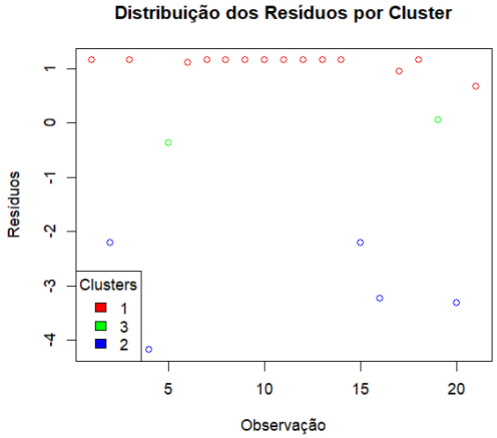

FIGURE 35. Gráfico com a distribuição dos resíduos da regressão linear por cluster

Por último, foi calculada a média para cada cluster, apresentando os valores presentes na Figura 36, onde é possível ver que a média dos três cluster é muito próxima de zero, isso significa que os resíduos não apresentam uma tendência significativa, isto é, que se ajustam mal ao modelo.

```
> aggregate(regression_model$residuals, by = list(data$cluster), FUN = mean)
  Group.1          x
1       1  1.1025578
2       2 -0.1540689
3       3 -3.0255343
```

FIGURE 36. Médias de cada cluster dos resíduos da regressão linear

Conclusão: Com o modelo de regressão linear final, pode-se afirmar que as variáveis $M14$ (Cyclomantic Complexity) e $M16$ (Class Coupling) correlacionam com a variável $M17$ (Class Cohesion). A variável $M14$ ao ter um coeficiente de 0.6220, e a variável $M16$ ao ter um coeficiente de 0.3822, significa que ao aumento de um valor nessas variáveis, leva a um aumento de 0.6220 ou 0.3822 na variável $M9$.

CONCLUSÕES E TRABALHO FUTURO

O objetivo da realização deste projeto foi colocar em pratica todos os tópicos lecionados na unidade curricular de Matemática Computacional II, sendo que os dados para estudo foram retirados do trabalho prático realizado no âmbito da Unidade Curricular de Engenharia de Software II, servindo também para retirar conclusões e complementar a componente estatística desta unidade curricular.

Concluímos que este trabalho nos ajudou a perceber a evolução notória de algumas métricas ao longo de cada recolha. Podemos ainda tirar conclusões sobre se algumas métricas têm alguma influência sob outras o que se revela bastante interessante e importante uma vez que, através dessas conclusões podem ser evitados erros/problemas. Isto foi especialmente importante no que se refere ao desenvolvimento do projeto de Engenharia de Software II, uma vez que nos permitiu acompanhar o crescimento do nosso trabalho, servindo de indicador do nosso próprio progresso.

No entanto, devido ao facto dos dados serem bastante inconstantes, os estudos apresentados ao longo deste relatório podem não ser os mais precisos possíveis. Outra limitação foi o facto de termos mais trabalhos para realizar o que nos fez ficar divididos e não nos focar tanto no neste trabalho. Apesar disto, tomámos decisões e foram adotadas medidas para tornar este relatório o mais completo e preciso possível, mesmo havendo incongruências nos dados, e limitações temporais.

Em relação ao desenvolvimento de trabalhos e relatórios futuros, idealmente seria possível tratar os dados de uma forma mais uniforme, permitindo ter em conta toda a informação e desta forma apresentar resultados mais úteis e precisos. Seria também interessante conseguir avaliar eventuais diferentes métricas.

ACKNOWLEDGMENTS

Gostaríamos de agradecer às professoras Eliana Costa e Silva e Maria João Polidoro pelo apoio fundamental durante o trabalho. Suas orientações foram essenciais para a análise estatística, contribuindo significativamente para o desenvolvimento deste projeto.

REFERENCES

1. Documentacão overleaf. `https://pt.overleaf.com/learn`.
2. Material disponibilizado no moodle. `https://moodle2.estg.ipp.pt/course/view.php?id=310`.