

# **Trabalho prático II**

## **Matemática Computacional II**

Autores:

Bruno Dylan Pinto Ferreira 8200586  
Gonçalo André Fontes Oliveira 8200595  
Jorge Miguel Fernandes Correia 8200592  
Nuno de Figueiredo Brito e Castro 8200591

Grupo: 03

Licenciatura em Engenharia Informática  
2021/2022



**ESCOLA  
SUPERIOR  
DE TECNOLOGIA  
E GESTÃO**

## CONTENTS

<b>INTRODUÇÃO</b>	3
<b>TRATAMENTO DE OUTLIERS</b>	3
<b>BASE DE DADOS</b>	4
<b>Resultado e discussão</b>	5
Métrica 1	5
Métrica 2	10
Métrica 3	25
<b>Conclusão e trabalho futuro</b>	32
<b>AGRADECIMENTOS</b>	32
<b>Referências</b>	32

**Resumo:** Este relatório foi criado no âmbito do trabalho prático 2 da disciplina de MC2 e tem como objetivo a aplicação as técnicas de análise de dados abordadas nas aulas de Matemática Computacional II relativas aos conteúdos C3, C4, C5 e C6. O grupo teria de efetuar hipóteses de investigação, tais como:

- Correlação entre as diversas métricas;
- O impacto das métricas no desempenho do grupo e do projeto;
- Definição de clusters;
- Determinação dos fatores que aumentam a probabilidade de falha.

Os dados analisados neste relatório são provenientes da análise de dados (reporting) de vários Projetos de Engenharia de Software II (ESII). Cada grupo teve de efetuar recolhas semanais e deveriam apresentar os seguintes dados:

- M1) número de linhas de código (LOC);
- M2) Número de componentes (classes);
- M3) LOC, por componente;
- M4) número de testes, por componente;
- M5) número de testes;
- M6) número de testes que falharam;
- M7) code coverage (percentage of missed instructions);
- M8) cyclomatic complexity;
- M9) percentagem de documentação (conforme e não conforme), por componente;
- M10) duplicação de código.

## INTRODUÇÃO

Num projeto de programação, os programadores são levados a fazer várias decisões. Algumas dessas decisões podem comprometer o projeto. Por isso vamos fazer uma análise aos dados obtidos e ver o que afeta mais o projeto.

Neste trabalho iremos apresentar vários dados e o processo que usamos:

- DATABASE AND METHODOLOGY – resumo dos dados e metodologia usada
- RESULT AND DISCUSSION – valores obtidos ao realizamos os nossos testes e discussão deles mesmos
- CONCLUSIONS AND FUTURE WORK – conclusão sobre os dados apresentados neste documento

## TRATAMENTO DE OUTLIERS

Alguns dados importados possuem valores vazios (N.A) e por isso existe a necessidade de tratar estes dados de modo a que a tabela resultante seja uma tabela que permita efetuar as análises previstas. Na secção de cada métrica será explicado o respetivo tratamento de outliers efetuado.

## BASE DE DADOS

Na base de dados encontram-se 17 conjuntos de dados, que são:

- Recolha
- Grupo Anónimo
- M1) Número total de linhas de código (LOC)
- M2) Número de classes
- M3) Número de LOC, por classe
- M4) Número de testes, por classe
- M5) Número de testes
- M6) Número de testes que falharam
- M6.b) Número de testes que falharam, por classe
- M7) Code coverage (percentage of missed instructions)
- M7.b) Code coverage, por classe
- M8) Cyclomatic complexity (média de todos os métodos)
- M8.b) cyclomatic complexity, por classe (média de cada método)
- M9) Percentagem de documentação (conforme e não conforme), por classe
- M9.b) Número de documentação em falta, por classe
- M10) Duplicação de código, em %
- M10.b) Número de LOC comentadas, por classe

Software que serão usados :

- R
- R-Studio
- Excel
- Discord
- Microsoft Teams
- Latex
- Overleaf

## RESULTADO E DISCUSSÃO

### Métrica 1

A primeira métrica em estudo é a possibilidade de a quantidade de comentários ((M10.b) Número de LOC comentadas, por classe) influenciar no número de testes falhados ((M6.b) Número de testes que falharam, por classe). O estudo desta métrica permitirá perceber se os comentários têm um impacto significativo no sucesso dos testes pelo que, pode mostrar uma vez mais que as linhas de código comentadas não servem apenas de documentação.

**Tratamento de outliers:** Decidimos que seriam removidas as linhas em que as colunas "Número de testes que falharam, por classe" e "Número de LOC comentadas, por classe" estivesse vazias.

Também foram removidas as linhas em que a coluna "Número de LOC, por classe" estivesse vazia ou possuísse o valor 0. Os dados vazios que sobraram foram transformados em 0 de modo a poder prosseguir com a análise.

Podemos desde já observar que ambas as variáveis são quantitativas discretas mas devido a terem muitos valores possíveis, serão consideradas como contínuas. Trata-se de variáveis independentes uma vez que, os dados referentes às linhas de código comentadas por classe e ao número de testes falhados por classe, não se referem às mesmas classes visto que os dados são provenientes de grupos diferentes que de certo possuem desenvolvimentos distintos.

Após a identificação e classificação das variáveis em estudo, daremos início à análise descritiva de todas as variáveis:

Começamos por retirar as linhas em que o M5) Número de testes eram NA ou 0 e as que o M6.b) Número de testes que falharam, por classe e o M10.b) Número de LOC comentadas, por classe eram simultaneamente NA.

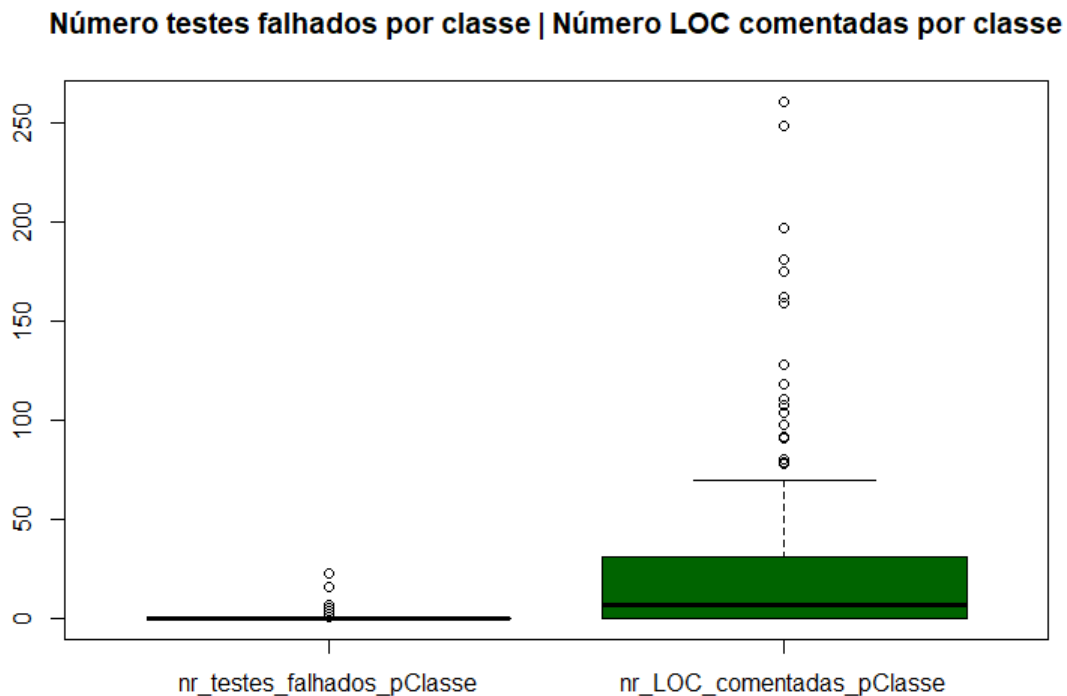
**Medidas de localização:** As medidas de localização podem-se dividir:

- **Medidas de tendência central:** média, mediana, moda
- **Medidas de tendência não central:** quartis, decis, percentis, quantis

Table I: Summary Métrica 1

	nr_testes_falhados_pClasse	nr_LOC_comentadas_pClasse
Minimo	0.0000	0.00
1º Quartil	0.0000	0.00
Mediana	0.0000	7.00
Media	0.5566	30.43
3º Quartil	0.0000	31
Maximo	23.0000	261.00

Com os dados obtidos através do script Summary, executado no RStudio podemos concluir que dos 203 dados recolhidos:



- 50% deles possuem um número de testes falhados igual a 0 e um número de linhas de código comentadas inferior ou igual a 7.
- A média de testes falhados por classe é de 0.5566 e a média de número de linhas de código comentadas é de 30.43.

```
> quantile(nr_testes_falhados_pClasse, type=2)
0% 25% 50% 75% 100%
0 0 0 0 23
> quantile(nr_LOC_comentadas_pClasse, type=2)
0% 25% 50% 75% 100%
0 0 7 31 261
```

- 75% dos dados possuem um número de testes falhados igual a 0 e um número de linhas de código comentadas inferior ou igual a 31.

**Medidas de dispersão:** As medidas de dispersão determinam a dispersão dos dados relativamente às medidas de localização, permitindo avaliar a representatividade da média num conjunto de dados. Caso seja notória uma grande variação nos valores observados, uma medida de tendência central, por si só, pouca informação fornece. A dispersão pode ser observada através:

- **Amplitude da amostra:** A amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo pelo que, quanto maior o resultado obtido maior dispersão deve existir, no entanto, a desvantagem é não ser sensível a valores intermédios. Na imagem seguinte, podemos observar que existe uma maior dispersão de dados no número de linhas de código comentadas por classe.

```
> diff(range(nr_testes_falhados_pClasse))
[1] 23
> diff(range(nr_LOC_comentadas_pClasse))
[1] 261
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais e, à semelhança da amplitude amostral quanto maior o valor da amplitude interquartil, maior é dispersão nos 50% das observações centrais. Como podemos analisar na imagem abaixo, existe uma maior dispersão nos dados centrais do número de linhas de código comentadas. Uma das desvantagens desta medida de dispersão é o facto de não se refletir no conjunto de todos os dados.

```
> IQR(nr_testes_falhados_pClasse)
[1] 0
> IQR(nr_LOC_comentadas_pClasse)
[1] 31
```

- **Desvio absoluto médio:** Mais uma vez, podemos concluir que existe uma maior variabilidade no número de linhas de código comentadas por classe.

```
> mad(nr_testes_falhados_pClasse)
[1] 0
> mad(nr_LOC_comentadas_pClasse)
[1] 10.3782
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados.

```
> var(nr_testes_falhados_pClasse)
[1] 4.832171
> var(nr_LOC_comentadas_pClasse)
[1] 2769.405
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como o resultado obtido, que pode ser observado na imagem seguinte, é maior que 0 significa que existe variabilidade nos dados e que estamos perante variáveis aleatórias.

```
> sd(nr_testes_falhados_pClasse)
[1] 2.19822
> sd(nr_LOC_comentadas_pClasse)
[1] 52.62513
```

- **Coefficiente de variação:** O coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes.

```
> cvTestes
[1] 394.9015
> cvComentarios
[1] 172.9464
```

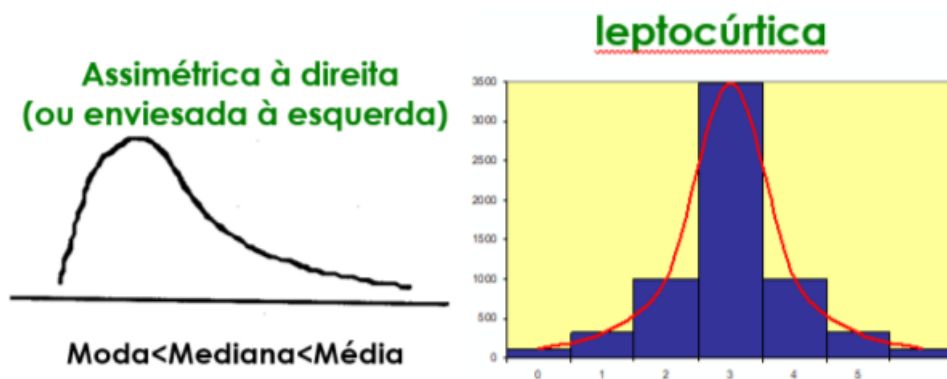
*Podemos concluir que apesar de terem sido obtidos valores diferentes, a análise final é a mesma. No entanto, é sempre vantajoso usar mais que uma medida de dispersão já que existem sempre fatores que influenciam os resultados dependendo da medida.*

Depois de analisadas as medidas estatísticas de localização e dispersão, podemos analisar a simetria e curtose.

```
> skewness(nr_testes_falhados_pClasse) > kurtosis(nr_testes_falhados_pClasse)
[1] 7.344033 [1] 63.54489
> skewness(nr_LOC_comentadas_pClasse) > kurtosis(nr_LOC_comentadas_pClasse)
[1] 2.479285 [1] 6.250527
```

Como se pode observar,  $g_1$  é maior que 0 para ambas as variáveis pelo que se pode concluir que são assimétricas à direita.

Quanto à curtose pode-se observar também que  $g_2$  são em ambas maiores que 0 o que significa que são leptocúrticas. Através da análise feita anteriormente, podemos concluir que não se trata de uma distribuição normal, uma vez que  $g_1$  e  $g_2$  não são aproximadamente 0.



Ainda referente à análise descritiva podemos ainda construir a tabela de frequências com ajuda do RStudio. Na imagem seguinte, que apresenta a tabela de frequências dos testes falhados por classe, podemos observar que os dados se concentram mais no intervalo  $[0, 3[$ .

Table II: Tabela Frequências Testes

Intervalo	ni	NI	fi	FI
$[0, 3[$	195	195	0.960591133	0.9605911
$[3, 6[$	4	199	0.019704433	0.9802956
$[6, 9[$	2	201	0.009852217	0.9901478
$[9, 12[$	0	201	0.000000000	0.9901478
$[12, 15[$	0	201	0.000000000	0.9901478
$[15, 18[$	1	202	0.004926108	0.9950739
$[18, 21[$	0	202	0.000000000	0.9950739
$[21, 24[$	1	203	0.004926108	1.0000000



Seguidamente, pode ser observada a tabela de frequências relativa ao número de linhas comentadas por classe, onde podemos observar que na amostra, a maior parte dos dados referentes a linhas de código comentadas por classe encontram-se entre 0 e 27.

Table III: Tabela Frequências Comentários

Intervalo	ni	NI	fi	FI
[0, 29[	143	143	0.704433498	0.7044335
[29, 58[	27	170	0.133004926	0.8374384
[58, 87[	8	178	0.039408867	0.8768473
[87, 116[	10	188	0.049261084	0.9261084
[116, 145[	4	192	0.019704433	0.9458128
[145, 174[	2	194	0.009852217	0.9556650
[174, 203[	5	199	0.024630542	0.9802956
[203, 232[	0	199	0.000000000	0.9802956
[232, 261[	2	201	0.009852217	0.9901478
[261, 290[	2	203	0.009852217	1.0000000

Neste momento, depois de feita a identificação e classificação, análise descritiva e identificação de outliers, podemos fazer o teste de correlação para verificar se as duas variáveis se relacionam.

```
> cor.test(nr_LOC_comentadas_pClasse, nr_testes_falhados_pClasse, method= "spearman", exact=FALSE)

spearman's rank correlation rho

data: nr_LOC_comentadas_pClasse and nr_testes_falhados_pClasse
s = 1994785, p-value = 1.406e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.4307695
```

Pelo resultado obtido, podemos concluir que as duas variáveis não se associam, uma vez que apenas se trata de uma associação caso seja aproximadamente 1 (associação direta perfeita) ou aproximadamente -1 (associação inversa perfeita).

Conclui-se com o estudo desta métrica que o número de linhas de código comentadas por classe, não influencia o número de testes falhados por classe.

## Métrica 2

A segunda métrica visa estudar se, a cada recolha que passa existe um aumento significativo do número de linhas de código comentadas, por classe e número de testes falhados por classe.

**Tratamento de outliers:** Decidimos que seriam removidas as linhas em que a coluna "Número de LOC comentadas, por classe" estivesse vazia. Para cada recolha foram selecionadas as linhas correspondentes e caso existissem células vazias, estas seriam transformadas em 0 de modo a poder prosseguir com a análise.

Tal como foi referido na métrica anterior, a variável de linhas de código comentadas por classe pode ser classificada como quantitativa discreta, no entanto, devido ao facto de serem apresentados muitos valores podemos afirmar de que trata-se de uma variável quantitativa contínua. Quanto à variável recolha, classifica-se como variável qualitativa nominal.

Primeiramente começamos por realizar uma pequena análise descritiva da variável em estudo.

Table IV: Summary Métrica 2 Recolha 1

	nr_LOL_comentadas_pClasse	nr_LOC_falhados_pClasse
Mínimo	0.00	0
1º Quartil	0.00	0
Mediana	4.00	0
Media	12.57	0
3º Quartil	21.00	0
Maximo	42.00	0

**Medidas de localização:** Podemos observar através dos resultados obtidos com a recolha 1 que :

- A média de linhas de código comentadas por classe é 12.57, já para o número de testes falhados por classe é 0.
- 50% dos dados da amostra são inferiores ou iguais a 4 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- 75% dos dados da amostra são inferiores ou iguais a 21 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- 100% dos testes falhados por classe são iguais a 0 nesta recolha.

**Medidas de dispersão:** As medidas de dispersão determinam a dispersão dos dados relativamente às medidas de localização, permitindo avaliar a representatividade da média num conjunto de dados. Caso seja notória uma grande variação nos valores observados, uma medida de tendência central, por si só, pouca informação fornece. A dispersão pode ser observada através:

- **Amplitude da amostra:** Tal como referido anteriormente, a amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo e quanto maior o resultado obtido maior dispersão deve existir, no entanto, não é sensível a valores intermédios o que se revela uma desvantagem.

```
> diff(range(nr_testes_falhados_pClasse))  
[1] 0  
> diff(range(nr_LOC_comentadas_pClasse))  
[1] 42
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais e, à semelhança da amplitude amostral quanto maior o valor da amplitude interquartil, maior é dispersão nos 50% das observações centrais. Como podemos analisar na imagem abaixo, existe uma maior dispersão nos dados centrais do número de linhas de código comentadas. Uma das desvantagens desta medida de dispersão é o facto de não se refletir no conjunto de todos os dados.

```
> IQR(nr_testes_falhados_pClasse)
[1] 0
> IQR(nr_LOC_comentadas_pClasse)
[1] 21
```

- **Desvio absoluto médio:** Mais uma vez, podemos concluir que existe uma maior variabilidade no número de linhas de código comentadas por classe.

```
> mad(nr_testes_falhados_pClasse)
[1] 0
> mad(nr_LOC_comentadas_pClasse)
[1] 5.9304
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados

```
> var(nr_testes_falhados_pClasse)
[1] 0
> var(nr_LOC_comentadas_pClasse)
[1] 286.2857
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como se pode ver no resultado obtido, que pode ser observado na imagem seguinte, é igual a 0 para o número de testes falhados por classe o que significa que não existe variabilidade nos dados e que não estamos perante variáveis aleatórias. O mesmo não se pode verificar para o número de linhas de código comentadas.

```
> sd(nr_testes_falhados_pClasse)
[1] 0
> sd(nr_LOC_comentadas_pClasse)
[1] 16.91998
```

- **Coeficiente de variação:** Como referido anteriormente, o coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes.

```
> cvTestes
[1] NaN
> cvComentarios
[1] 134.5907
```

*Apesar de terem sido obtidos valores diferentes, a análise final é a mesma. No entanto, é sempre vantajoso usar mais que uma medida de dispersão já que existem sempre fatores que influenciam os resultados dependendo da medida.*

*Por fim podemos concluir que na recolha 1 não existe variabilidade no número de testes falhados por classe, ao contrário do que acontece para o número de linhas de código comentadas por classe.*

Table V: Summary Métrica 2 Recolha 2

	nr_LOL_comentadas_pClasse	nr_LOC_falhados_pClasse
Minimo	0.000	0.0000
1º Quartil	0.000	0.0000
Mediana	2.000	0.0000
Media	7.108	0.6216
3º Quartil	9.000	0.0000
Maximo	42.000	16.0000

Podemos observar através dos resultados obtidos com a recolha 2 que :

- A média de linhas de código comentadas por classe é 7.11, já para o número de testes falhados por classe é 0.62 .
- 50% dos dados da amostra são inferiores ou iguais a 2 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- 75% dos dados da amostra são inferiores ou iguais a 9 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- **Amplitude da amostra:** A amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo. No entanto, a desvantagem é não ser sensível a valores intermédios.

```
> diff(range(nr_testes_falhados_pClasse))
[1] 16
> diff(range(nr_LOC_comentadas_pClasse))
[1] 42
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais.

```
> IQR(nr_testes_falhados_pClasse)
[1] 0
> IQR(nr_LOC_comentadas_pClasse)
[1] 9
```

- **Desvio absoluto médio:** Medida de dispersão absoluta que permite medir a variabilidade dos dados.

```
> mad(nr_testes_falhados_pClasse)
[1] 0
> mad(nr_LOC_comentadas_pClasse)
[1] 2.9652
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados

```
> var(nr_testes_falhados_pClasse)
[1] 7.51952
> var(nr_LOC_comentadas_pClasse)
[1] 116.8769
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como o resultado obtido, que pode ser observado na imagem seguinte, é maior que 0 significa que existe variabilidade nos dados e que estamos perante variáveis aleatórias.

```
> sd(nr_testes_falhados_pClasse)
[1] 2.742174
> sd(nr_LOC_comentadas_pClasse)
[1] 10.81096
```

- **Coefficiente de variação:** O coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes. É muito usado quando todas as observações têm o mesmo sinal.

```
> cvTestes
[1] 441.1324
> cvComentarios
[1] 152.0934
```

*Por fim podemos concluir que na recolha 2, em comparação à recolha anterior, apresenta maior dispersão de dados no número de testes falhados por classe e menor dispersão de dados em relação ao número de linhas de código comentadas por classe*

Table VI: Summary Métrica 2 Recolha 3

	nr_LOL_comentadas_pClasse	nr_LOC_falhados_pClasse
Mínimo	0.00	0.00000
1º Quartil	0.00	0.00000
Mediana	18.00	0.00000
Média	37.56	0.04951
3º Quartil	53.00	0.00000
Máximo	261.00	2.00000

Para a recolha 3 podemos concluir que:

- A média de linhas de código comentadas por classe é 37.56, já para o número de testes falhados por classe é 0.50.
- 50% dos dados da amostra são inferiores ou iguais a 18 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- 75% dos dados da amostra são inferiores ou iguais a 53 linhas de código comentadas por classe e igual a 0 testes falhados por classe.
- **Amplitude da amostra:** A amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo. No entanto, a desvantagem é não ser sensível a valores intermédios.

```
> diff(range(nr_testes_falhados_pClasse))
[1] 2
> diff(range(nr_LOC_comentadas_pClasse))
[1] 261
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais.

```
> IQR(nr_testes_falhados_pClasse)
[1] 0
> IQR(nr_LOC_comentadas_pClasse)
[1] 53
```

- **Desvio absoluto médio:** Medida de dispersão absoluta que permite medir a variabilidade dos dados.

```
> mad(nr_testes_falhados_pClasse)
[1] 0
> mad(nr_LOC_comentadas_pClasse)
[1] 26.6868
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados

```
> var(nr_testes_falhados_pClasse)
[1] 0.06752475
> var(nr_LOC_comentadas_pClasse)
[1] 2802.728
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como o resultado obtido para o número de testes falhados, que pode ser observado na imagem seguinte, é aproximadamente 0 o que significa que não existe variabilidade nos dados e que não estamos perante variáveis aleatórias. Ao contrário do resultado obtido para o número de linhas de código comentadas por classe que é superior a 0.

```
> sd(nr_testes_falhados_pClasse)
[1] 0.2598553
> sd(nr_LOC_comentadas_pClasse)
[1] 52.9408
```

- **Coefficiente de variação:** O coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes. É muito usado quando todas as observações têm o mesmo sinal.

```
> cvTestes
[1] 524.9076
> cvComentarios
[1] 140.9336
```

*Por fim podemos concluir que na recolha 3, em comparação à recolha anterior, apresenta menor dispersão de dados no número de testes falhados por classe e maior dispersão de dados em relação ao número de linhas de código comentadas por classe*

Table VII: Summary Métrica 2 Recolha 4

	nr_LOL_comentadas_pClasse	nr_LOC_falhados_pClasse
Minimo	0.00	0.00000
1º Quartil	6.50	0.00000
Mediana	26.00	0.00000
Media	39.32	0.1069
3º Quartil	48.00	0.00000
Maximo	261.00	2.00000

Para a recolha 4 podemos concluir que:

- A média de linhas de código comentadas por classe é 39.32, já para o número de testes falhados por classe é 0.11.
- 50% dos dados da amostra são inferiores ou iguais a 26 linhas de código comentadas por classe e igual a 0 testes falhados por classe.

- 75% dos dados da amostra são inferiores ou iguais a 48 linhas de código comentadas por classe e igual a 0 testes falhados por classe.

- **Amplitude da amostra:** A amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo. No entanto, a desvantagem é não ser sensível a valores intermédios.

```
> diff(range(nr_testes_falhados_pClasse))
[1] 2
> diff(range(nr_LOC_comentadas_pClasse))
[1] 261
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais.

```
> IQR(nr_testes_falhados_pClasse)
[1] 0
> IQR(nr_LOC_comentadas_pClasse)
[1] 41.5
```

- **Desvio absoluto médio:** Medida de dispersão absoluta que permite medir a variabilidade dos dados.

```
> mad(nr_testes_falhados_pClasse)
[1] 0
> mad(nr_LOC_comentadas_pClasse)
[1] 31.1346
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados

```
> var(nr_testes_falhados_pClasse)
[1] 0.1731063
> var(nr_LOC_comentadas_pClasse)
[1] 2410.296
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como o resultado obtido para o número de testes falhados por classe, que pode ser observado na imagem seguinte, é aproximadamente 0 pode-se concluir que não existe variabilidade nos dados e que não estamos perante variáveis aleatórias. O mesmo não se pode concluir para o número de linhas de código comentada por classe.

```
> sd(nr_testes_falhados_pClasse)
[1] 0.4160604
> sd(nr_LOC_comentadas_pClasse)
[1] 49.09477
```

- **Coefficiente de variação:** O coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes. É muito usado quando todas as observações têm o mesmo sinal.

```
> cvTestes
[1] 389.3137
> cvComentarios
[1] 124.8576
```

*Por fim podemos concluir que na recolha 4, em comparação à recolha anterior, apresenta menor dispersão de dados quer relativamente a número de testes falhados por classe, quer relativamente ao número de linhas de código comentadas por classe*

Após o estudo das medidas de localização e dispersão iremos proceder ao teste de hipóteses, onde vamos estudar se a média de linhas de código comentadas por classe e o número de testes falhados aumenta de recolha para recolha. Assumimos a normalidade uma vez que temos mais de 30 observações:

### Teste de hipótese para comparação de médias entre Recolha 1 e Recolha 2:

Teste de hipóteses para o número de linhas de código comentadas

Hipótese nula (H0): média da recolha 2 > média da recolha 1

Hipótese alternativa (H1): média da recolha 2 < média da recolha 1

Sendo que a média de linhas de código comentadas por classe da recolha 1 é de 12.57, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

#### *Teste de hipóteses para 90%:*

IC 90% =  $]-\infty; 9.43]$  Como se pode observar na imagem seguinte, o p-value < 10%, ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+         alternative="less",
+         mu = 12.57,
+         conf.level = 0.90)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = -3.0731, df = 36, p-value = 0.002012
alternative hypothesis: true mean is less than 12.57
90 percent confidence interval:
 -Inf 9.428412
sample estimates:
mean of x
 7.108108
```

#### *Teste de hipóteses para 95%:*

IC 95% =  $]-\infty; 10.11]$  Como se pode observar na imagem seguinte, o p-value < 5%, ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+         alternative="less",
+         mu = 12.57,
+         conf.level = 0.95)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = -3.0731, df = 36, p-value = 0.002012
alternative hypothesis: true mean is less than 12.57
95 percent confidence interval:
 -Inf 10.10874
sample estimates:
mean of x
 7.108108
```



**Teste de hipóteses para 99%:**

IC 99% =  $]-\infty; 11.43]$  Como se pode observar na imagem seguinte, o p-value < 1%, ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 12.57,
+        conf.level = 0.99)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = -3.0731, df = 36, p-value = 0.002012
alternative hypothesis: true mean is less than 12.57
99 percent confidence interval:
 -Inf 11.43496
sample estimates:
mean of x
 7.108108
```

**Teste de hipóteses para o número de testes falhados por classe:**

Hipótese nula (H0): média da recolha 2 > média da recolha 1

Hipótese alternativa (H1): média da recolha 2 < média da recolha 1

Sendo que a média de testes falhados por classe da recolha 1 é de 0, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

**Teste de hipóteses para 90%:**

IC 90% =  $]-\infty; 1.21]$  Como se pode observar na imagem seguinte, o p-value > 10%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0,
+        conf.level = 0.90)

One sample t-test

data:  nr_testes_falhados_pClasse
t = 1.3789, df = 36, p-value = 0.9118
alternative hypothesis: true mean is less than 0
90 percent confidence interval:
 -Inf 1.210161
sample estimates:
mean of x
 0.6216216
```

### **Teste de hipóteses para 95%:**

IC 95% = ] $-\infty$ ; 1.38] Como se pode observar na imagem seguinte, o p-value > 5%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0,
+        conf.level = 0.95)

One sample t-test

data:  nr_testes_falhados_pClasse
t = 1.3789, df = 36, p-value = 0.9118
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
 -Inf 1.382724
sample estimates:
mean of x
0.6216216
```

### **Teste de hipóteses para 99%:**

IC 99% = ] $-\infty$ ; 1.72] Como se pode observar na imagem seguinte, o p-value > 1%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0,
+        conf.level = 0.99)

One sample t-test

data:  nr_testes_falhados_pClasse
t = 1.3789, df = 36, p-value = 0.9118
alternative hypothesis: true mean is less than 0
99 percent confidence interval:
 -Inf 1.719118
sample estimates:
mean of x
0.6216216
```

**Conclusão:** Da recolha 1 para a recolha 2 não se notou um aumento significativo relativamente à média do número de linhas de código comentadas por classe e número de testes falhados por classe. O contrário acontece para a média de testes falhados por classe, onde se nota um aumento.

### Teste de hipótese para comparação de médias entre Recolha 2 e Recolha 3:

Teste de hipóteses para o número de linhas de código comentadas

Hipótese nula (H0): média da recolha 3 > média da recolha 2

Hipótese alternativa (H1): média da recolha 3 < média da recolha 2

Sendo que a média de linhas de código comentadas por classe da recolha 2 é de 7.108, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

#### Teste de hipóteses para 90%:

IC 90% = ]-∞; 44.36] Como se pode observar na imagem seguinte, o p-value > 10%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 7.108,
+        conf.level = 0.90)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = 5.7816, df = 100, p-value = 1
alternative hypothesis: true mean is less than 7.108
90 percent confidence interval:
 -Inf 44.36022
sample estimates:
mean of x
 37.56436
```

#### Teste de hipóteses para 95%:

IC 95% = ]-∞; 46.31] Como se pode observar na imagem seguinte, o p-value > 5%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 7.108,
+        conf.level = 0.95)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = 5.7816, df = 100, p-value = 1
alternative hypothesis: true mean is less than 7.108
95 percent confidence interval:
 -Inf 46.31015
sample estimates:
mean of x
 37.56436
```

**Teste de hipóteses para 99%:**

IC 99% =  $]-\infty; 50.02]$  Como se pode observar na imagem seguinte, o p-value  $> 1\%$ , ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 7.108,
+        conf.level = 0.99)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = 5.7816, df = 100, p-value = 1
alternative hypothesis: true mean is less than 7.108
99 percent confidence interval:
 -Inf 50.0186
sample estimates:
mean of x
 37.56436
```

**Teste de hipóteses para o número de testes falhados por classe:**

Hipótese nula (H0): média da recolha 3  $>$  média da recolha 2

Hipótese alternativa (H1): média da recolha 3  $<$  média da recolha 2

Sendo que a média de testes falhados por classe da recolha 2 é de 0.6216, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

**Teste de hipóteses para 90%:**

IC 90% =  $]-\infty; 0.83]$  Como se pode observar na imagem seguinte, o p-value  $< 10\%$ , ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0.6216,
+        conf.level = 0.90)

One sample t-test

data:  nr_testes_falhados_pClasse
t = -22.126, df = 100, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0.6216
90 percent confidence interval:
 -Inf 0.08286185
sample estimates:
mean of x
 0.04950495
```

**Teste de hipóteses para 95%:**

IC 95% =  $]-\infty; 0.09]$  Como se pode observar na imagem seguinte, o p-value < 5%, ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0.6216,
+        conf.level = 0.95)

One Sample t-test

data:  nr_testes_falhados_pClasse
t = -22.126, df = 100, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0.6216
95 percent confidence interval:
 -Inf 0.09243291
sample estimates:
mean of x
0.04950495
```

**Teste de hipóteses para 99%:**

IC 99% =  $]-\infty; 0.11]$  Como se pode observar na imagem seguinte, o p-value < 1%, ou seja, rejeita-se a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0.6216,
+        conf.level = 0.99)

One Sample t-test

data:  nr_testes_falhados_pClasse
t = -22.126, df = 100, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0.6216
99 percent confidence interval:
 -Inf 0.1106355
sample estimates:
mean of x
0.04950495
```

**Conclusão:** Da recolha 2 para a recolha 3 podemos observar que existiu um aumento significativo em relação ao número de linhas de código comentadas por classe, no entanto, não se nota o mesmo para o número de testes falhados por classe.

### Teste de hipótese para comparação de médias entre Recolha 3 e Recolha 4:

Teste de hipóteses para o número de linhas de código comentadas

Hipótese nula (H0): média da recolha 4 > média da recolha 3

Hipótese alternativa (H1): média da recolha 4 < média da recolha 3

Sendo que a média de linhas de código comentadas por classe da recolha 3 é de 37.56, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

#### Teste de hipóteses para 90%:

IC 90% = ]-∞; 44.85] Como se pode observar na imagem seguinte, o p-value > 10%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 37.56,
+        conf.level = 0.90)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = 0.41045, df = 130, p-value = 0.6589
alternative hypothesis: true mean is less than 37.56
90 percent confidence interval:
 -Inf 44.84582
sample estimates:
mean of x
 39.32061
```

#### Teste de hipóteses para 95%:

IC 95% = ]-∞; 46.43] Como se pode observar na imagem seguinte, o p-value > 5%, ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+        alternative="less",
+        mu = 37.56,
+        conf.level = 0.95)

One sample t-test

data:  nr_LOC_comentadas_pClasse
t = 0.41045, df = 130, p-value = 0.6589
alternative hypothesis: true mean is less than 37.56
95 percent confidence interval:
 -Inf 46.42674
sample estimates:
mean of x
 39.32061
```

**Teste de hipóteses para 99%:**

IC 99% =  $]-\infty; 49.42]$  Como se pode observar na imagem seguinte, o p-value  $> 1\%$ , ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_LOC_comentadas_pClasse,
+         alternative="less",
+         mu = 37.56,
+         conf.level = 0.99)

One Sample t-test

data:  nr_LOC_comentadas_pClasse
t = 0.41045, df = 130, p-value = 0.6589
alternative hypothesis: true mean is less than 37.56
99 percent confidence interval:
 -Inf 49.42382
sample estimates:
mean of x
39.32061
```

Teste de hipóteses para o número de testes falhados por classe:

Hipótese nula (H0): média da recolha 4  $>$  média da recolha 3

Hipótese alternativa (H1): média da recolha 4  $<$  média da recolha 3

Sendo que a média de testes falhados por classe da recolha 1 é de 0.04951, fazendo os testes de hipóteses podemos concluir se realmente houve um aumento significativo.

**Teste de hipóteses para 90%:**

IC 90% =  $]-\infty; 0.15]$  Como se pode observar na imagem seguinte, o p-value  $> 10\%$ , ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+         alternative="less",
+         mu = 0.04951,
+         conf.level = 0.90)

One Sample t-test

data:  nr_testes_falhados_pClasse
t = 1.5779, df = 130, p-value = 0.9415
alternative hypothesis: true mean is less than 0.04951
90 percent confidence interval:
 -Inf 0.1536943
sample estimates:
mean of x
0.1068702
```

**Teste de hipóteses para 95%:**

IC 95% =  $]-\infty; 0.17]$  Como se pode observar na imagem seguinte, o p-value  $> 5\%$ , ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0.04951,
+        conf.level = 0.95)

One sample t-test

data:  nr_testes_falhados_pClasse
t = 1.5779, df = 130, p-value = 0.9415
alternative hypothesis: true mean is less than 0.04951
95 percent confidence interval:
 -Inf 0.1670921
sample estimates:
mean of x
0.1068702
```

**Teste de hipóteses para 99%:**

IC 99% =  $]-\infty; 0.19]$  Como se pode observar na imagem seguinte, o p-value  $> 1\%$ , ou seja, não se rejeita a hipótese nula.

```
> t.test(nr_testes_falhados_pClasse,
+        alternative="less",
+        mu = 0.04951,
+        conf.level = 0.99)

One sample t-test

data:  nr_testes_falhados_pClasse
t = 1.5779, df = 130, p-value = 0.9415
alternative hypothesis: true mean is less than 0.04951
99 percent confidence interval:
 -Inf 0.1924913
sample estimates:
mean of x
0.1068702
```

**Conclusão:** Podemos concluir que, da recolha 3 para a recolha 4 notou-se um aumento significativo da média de número de linhas de código comentadas por classe e do número de testes falhados por classe.

Após estes 3 testes podemos concluir que houve um aumento significativo de ambas variáveis na maior parte dos testes realizados. Podemos encarar positivamente o aumento significativo da média de linhas de código comentadas por classe, ao contrário do aumento significativo do aumento da média de testes falhados por classe já que era suposto que este número fosse diminuindo significativamente.



### Métrica 3

A terceira métrica em estudo é a possibilidade de a quantidade de número total de linhas de código ((M1) Número total de linhas de código (LOC)) influencia a cyclomatic complexity ((M8) Cyclomatic complexity (média de todos os métodos)). O estudo desta métrica permitirá perceber se o número total de linhas de código tem um impacto negativo na cyclomatic complexity

**Tratamento de outliers:** Decidimos que seriam removidas as linhas em que a coluna "Número total de linhas de código (LOC)" ou a coluna "Cyclomatic complexity (média de todos os métodos)" estivesse vazia.

Podemos desde já observar que ambas as variáveis são quantitativas discretas mas devido a terem muitos valores possíveis, serão consideradas como contínuas. Tratam-se de variáveis independentes uma vez que, os dados referentes ao total de linhas de código e ao cyclomatic complexity, não se referem às mesmas classes visto que os dados são provenientes de grupos diferentes que de certo possuem desenvolvimentos distintos.

Após a identificação e classificação das variáveis em estudo, daremos início à análise descritiva de todas as variáveis:

Para esta amostra foram retiradas as linhas em que o (M1) Número total de linhas de código (LOC)) e a (M8) Cyclomatic complexity (média de todos os métodos) eram simultaneamente NA.

Tendo em conta as restrições referidas anteriormente, foram removidas 175 entradas ficando assim com uma amostra de 309 entradas

**Medidas de localização:** As medidas de localização podem-se dividir:

- **Medidas de tendência central:** média, mediana, moda
- **Medidas de tendência não central:** quartis, decis, percentis, quantis

Com os dados obtidos através do script Summary, executado no RStudio podemos concluir que dos 309 dados recolhidos:

Table VIII: Summary Métrica 3

	nr_total_linhas_codigo	cxy
Mínimo	221.00	0.00
1º Quartil	596.00	0.00
Mediana	1321.00	0.00
Media	1490.00	0.4693
3º Quartil	1728.00	0.00
Maximo	3147.00	23.00

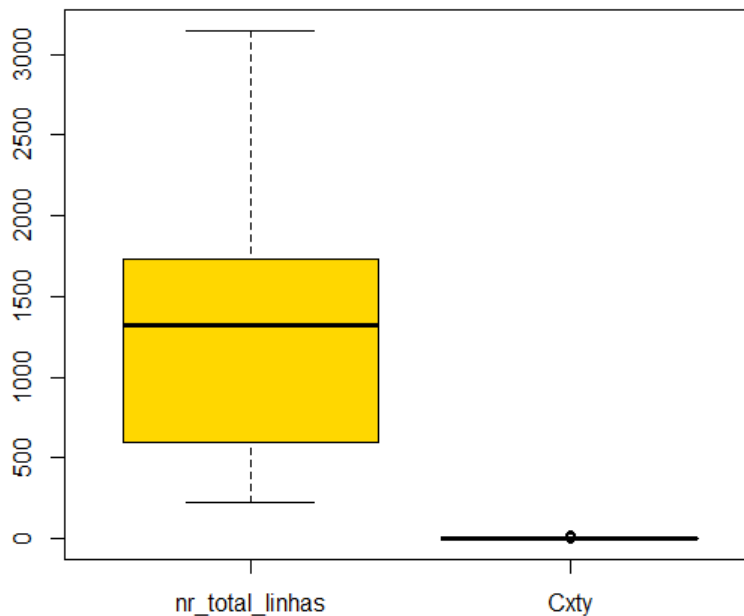
A média de testes falhados por classe é de 0.5566 e a média de número de linhas de código comentadas é de 30.43.

Realizando a análise dos quantis podemos concluir que:

```
> quantile (Cxty, type=2)
 0%   25%   50%   75%  100%
  0     0     0     0    23
> quantile (nr_total_linhas, type=2)
 0%   25%   50%   75%  100%
221   596 1321 1728 3147
```

- 50% deles possuem cyclomatic complexity igual a 0 e um número total de linhas de código inferior ou igual a 1321.
- 75% dos dados possuem cyclomatic complexity igual a 0 e um número total de linhas de código inferior ou igual a 1728.

### Número total de linhas de código | Cyclomatic Complexity



**Medidas de dispersão:** As medidas de dispersão determinam a dispersão dos dados relativamente às medidas de localização, permitindo avaliar a representatividade da média num conjunto de dados. Caso seja notória uma grande variação nos valores observados, uma medida de tendência central, por si só, pouca informação fornece. A dispersão pode ser observada através:

- **Amplitude da amostra:** A amplitude amostral consiste na diferença entre o valor máximo e o valor mínimo. No entanto, a desvantagem é não ser sensível a valores intermédios.

```
> diff(range(Cxty))
[1] 23
> diff(range(nr_total_linhas))
[1] 2926
```

- **Amplitude interquartil:** A amplitude interquartil consiste na diferença entre quartis que compreendem 50% das observações centrais.

```
> IQR(Cxty)
[1] 0
> IQR(nr_total_linhas)
[1] 1132
```

- **Desvio absoluto médio:** Medida de dispersão absoluta que permite medir a variabilidade dos dados.

```
> mad(Cxty)
[1] 0
> mad(nr_total_linhas)
[1] 1074.885
```

- **Variância:** Variância é a média dos quadrados dos desvios relativamente à média dos dados

```
> var(Cxty)
[1] 3.581032
> var(nr_total_linhas)
[1] 1043924
```

- **Desvio padrão:** O desvio padrão mede a variabilidade ou dispersão dos dados. Como o resultado obtido, que pode ser observado na imagem seguinte, é maior que 0 significa que existe variabilidade nos dados e que estamos perante variáveis aleatórias.

```
> sd(Cxty)
[1] 1.892362
> sd(nr_total_linhas)
[1] 1021.726
```

- **Coefficiente de variação:** O coeficiente de variação deve ser usado para comparar distribuições que estejam definidas em unidades de medida diferentes, ou distribuições que tenham médias diferentes. É muito usado quando todas as observações têm o mesmo sinal.

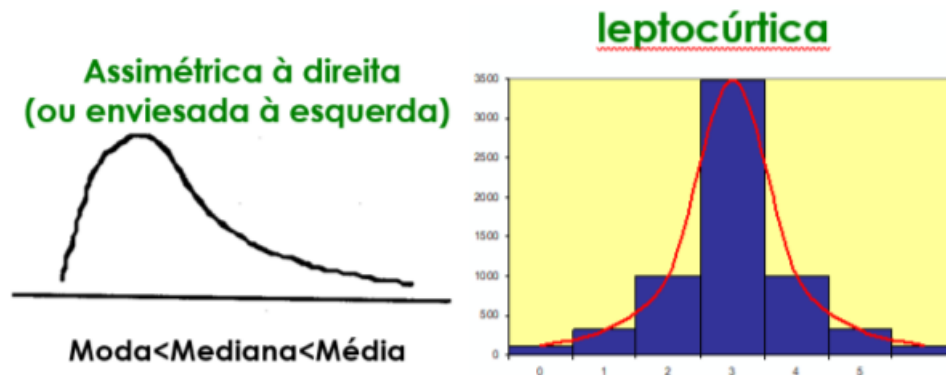
```
> cvCxty
[1] 403.2688
> cvNrTotalLinhas
[1] 68.584
```

Depois de analisadas as medidas estatísticas de localização e dispersão, podemos analisar a simetria e curtose.

```
> skewness(Cxty)
[1] 7.91432
> skewness(nr_total_linhas)
[1] 0.6183255
> kurtosis(Cxty)
[1] 78.67063
> kurtosis(nr_total_linhas)
[1] -1.006527
```

Como se pode observar, g1 é maior que 0 para ambas as variáveis pelo que se pode concluir que são assimétricas à direita.

Quanto à curtose pode-se observar também que g2 são em ambas maiores que 0 o que significa que são leptocúrticas. Através da análise feita anteriormente, podemos concluir que não se trata de uma distribuição normal, uma vez que g1 e g2 não são aproximadamente 0. No entanto como existem mais de 30 observações pode assumir-se que trata-se de uma distribuição normal.



Ainda referente à análise descritiva podemos ainda construir a tabela de frequências com ajuda do RStudio. Na imagem seguinte, que apresenta a tabela de frequências do número total de linhas de código.

Table IX: Tabela Frequências Número total linha de código

Intervalo	ni	NI	fi	FI
[221, 547[	74.00	74.00	0.24	0.24
[547, 873[	38.00	112.00	0.12	0.36
[873, 1199[	32.00	144.00	0.10	0.47
[1199, 1525[	43.00	187.00	0.14	0.61
[1525, 1851[	50.00	237.00	0.16	0.77
[1851, 2177[	0.00	237.00	0.00	0.77
[2177, 2503[	0.00	237.00	0.00	0.77
[2503, 2829[	0.00	237.00	0.00	0.77
[2829, 3155[	72.00	309.00	0.23	1.00

Seguidamente, pode ser observada a tabela de frequências relativa à cyclomatic complexity, onde podemos observar que na amostra, a maior parte dos dados referentes à cyclomatic complexity encontram-se entre 0 e 3.

Table X: Tabela Frequências Cxty

Intervalo	ni	NI	fi	FI
[0, 3[	296.00	296.00	0.96	0.96
[3, 6[	8.00	304.00	0.03	0.98
[6, 9[	3.00	307.00	0.01	0.99
[9, 12[	0.00	307.00	0.00	0.99
[12, 15[	0.00	307.00	0.00	0.99
[15, 18[	1.00	308.00	0.00	1.00
[18, 21[	0.00	308.00	0.00	1.00
[21, 24[	1.00	309.00	0.00	1.00

Neste momento, depois de feita a identificação e classificação, análise descritiva e identificação de outliers, podemos fazer o teste de correlação para verificar se as duas variáveis se relacionam.

Assumindo que não é uma distribuição normal, utiliza-se o método de spearman

```
> cor.test(nr_total_linhas, Cxty, method= "spearman", exact=FALSE)

Spearman's rank correlation rho

data:  nr_total_linhas and Cxty
S = 5870101, p-value = 0.0006145
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
-0.1937845
```

Assumindo que é uma distribuição normal, utiliza-se o método de pearson

```
> cor.test(nr_total_linhas, Cxty, method= "pearson", exact=FALSE)

Pearson's product-moment correlation

data:  nr_total_linhas and Cxty
t = -1.6831, df = 307, p-value = 0.09337
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.20501028  0.01612887
sample estimates:
      cor 
-0.09562037
```

Pelo resultado obtido, podemos concluir que as duas variáveis não se associam, uma vez que apenas se trata de uma associação caso seja aproximadamente 1 (associação direta perfeita) ou aproximadamente -1 (associação inversa perfeita).

### Estudar a relação linear entre as duas variáveis

Calculando o modelo linear e executando o comando do summary obtemos o output seguinte:

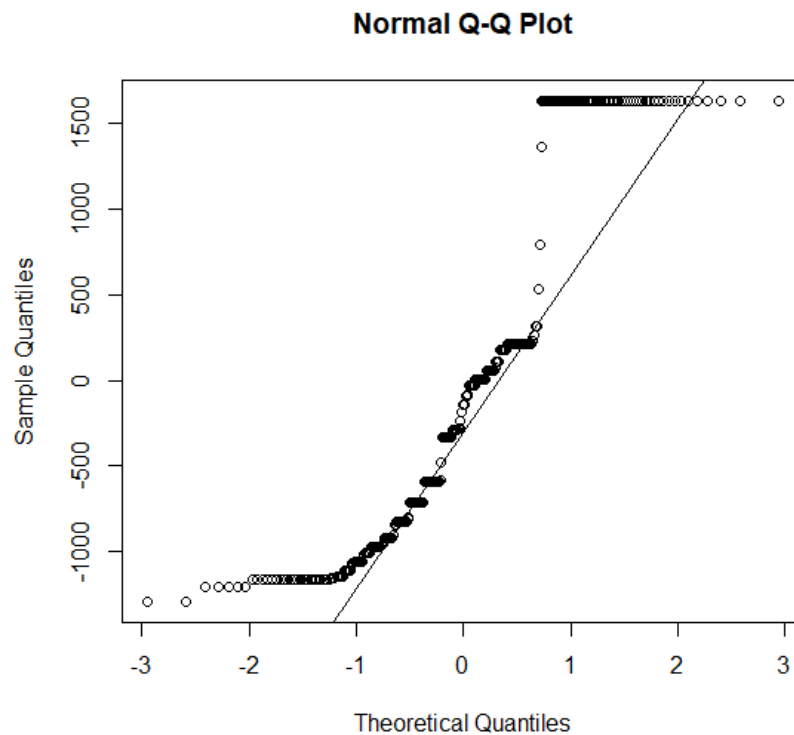
```
Call:
lm(formula = nr_total_linhas ~ Cxty)

Residuals:
    Min       1Q   Median       3Q      Max
-1293.0  -918.0  -141.3   317.3  1633.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1513.97     59.71   25.354  <2e-16 ***
Cxty         -51.63     30.67   -1.683   0.0934 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1019 on 307 degrees of freedom
Multiple R-squared:  0.009143, Adjusted R-squared:  0.005916
F-statistic: 2.833 on 1 and 307 DF, p-value: 0.09337
```

Realizando o gráfico dos resíduos do modelo linear, pode observar-se que este não segue uma distribuição normal, pois não coincide com a representação da linha reta do gráfico



Para avaliar a independência das variáveis aleatórias residuais, além de observar a ausência de padrões na configuração do gráfico dos resíduos com o índice das observações, pode-se realizar o teste de Durbin Watson que verifica se a auto correlação dos resíduos é nula

```
> durbinWatsonTest(modeloLinear)
lag Autocorrelation D-W Statistic p-value
1      0.9412651      0.1137661      0
Alternative hypothesis: rho ≠ 0
```

Como se pode observar a auto correlação dos resíduos é nula

Conclui-se com o estudo destas métricas que o número total de linhas de código, não influencia a cyclomatic complexity.

## CONCLUSÃO E TRABALHO FUTURO

Neste trabalho era proposto colocar em prática 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.

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. No entanto, devido ao facto dos dados serem bastante inconstantes, os estudos apresentados ao longo deste relatório podem não ser os mais precisos possíveis.

Referente a trabalho futuro, poderão ser adicionados mais valores às métricas existentes e, eventualmente a métricas que possam ser futuramente adicionadas.

## AGRADECIMENTOS

Gostaríamos de agradecer à professora Eliana Costa e Silva por todo o suporte ao longo do desenvolvimento do trabalho que nos auxiliou bastante.

## REFERÊNCIAS

1. Documentação overleaf. URL: <https://pt.overleaf.com/learn>.
2. Eliana Costa e Silva. Material disponibilizado no moodle.
3. Coefficient of variation in r. URL: <https://www.statology.org/coefficient-of-variation-in-r/>.
4. Métricas programação. URL: <http://www.mccabe.com/pdf/McCabeIQMetrics.pdf>.