Aula de
Big Data

Name = **Leandro de Souza**
Instagram = **@prof.leandrodesouza**

AULA06

Preparando os

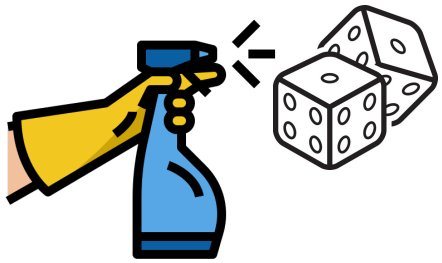
DADOS

Preparando os

DADOS

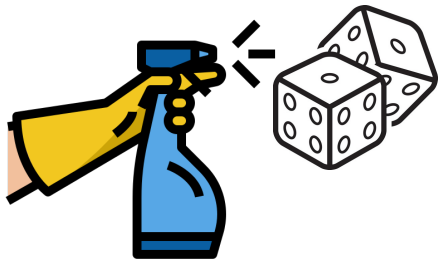
A **fase de preparação**, tratamento ou pré-processamento dos dados é essencial na análise de dados, sendo a tarefa que demanda maior tempo e trabalho.

Limpeza dos



O processo de limpeza requer uma inspeção minuciosa dos dados, bem como a realização de operações de **correção** e **remoção**, conforme a necessidade.

id	nome	idade	sexo	cidade
500	"pedro"	32	"M"	"São Paulo"
501	"maria"	41	"F"	"Curitiba"
502	"jonas"	25	"1"	"05360-152"
503	"lucia"	38	"2"	"Londrina"
504	"lucas"	29	"masc"	"Aracaju"
505	"lucas"	29	"masc"	"Aracaju"

Limpeza dos 

- Existem dados duplicados?
- Existem dados com informações incompletas?
- Existem dados com erros de digitação?
- Existem dados iguais representados de diferentes formas?
- Existem dados que violam as regras de negócio?

Big data

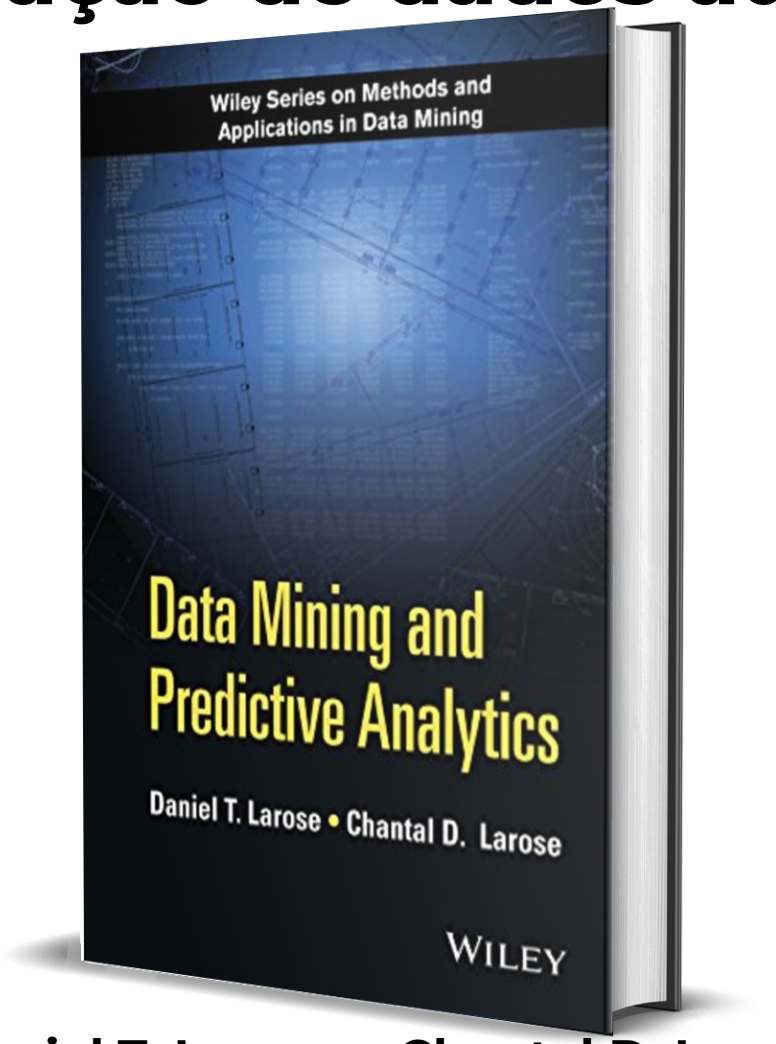
Big data

Manipulação de dados ausentes

ID	DATA	VALOR	FRETE	PAGAMENTO
101	2022-03-02	500,00	30,00	boleto
102	2022-03-03	420,00	-	cartão
103	2022-03-03	108,00	15,50	boleto
104	2022-03-04	100,00	5,85	-
...	...	...	...	...

O que fazemos com esses registros em nossa análise?

Manipulação de dados ausentes



Daniel T. Larose e Chantal D. Larose

Manipulação de dados ausentes

Para não descartar os registros com dados ausentes em nossa análise, **Daniel T. Larose e Chantal D. Larose**, autores do livro ***Data Mining and Predictive Analytics***, indicam as seguintes abordagens:

- Substituir o dado ausente com **alguma constante**, especificada pelo analista;
- Substituir o dado ausente pela **média ou moda** do campo;
- Substituir o dado ausente com **um valor gerado aleatoriamente a partir de uma distribuição observada;**
- Substituir o dado ausente a partir de **valores baseados em outras características do registro.**

Identificacão de anomalias

id	data	valor	frete	pagamento
106	2016-03-05	120,00	10,00	boleto
107	2016-03-05	350,00	14,00	cartão
108	2016-03-06	400,00	22,50	boleto
110				
111	2016-03-06	135,00	20,00	cartao
112	2016-03-06	280,00	15,00	cartao
113	2016-03-06	350,00	18,00	cartao
114	2016-03-06	310,00	50,00	cartao
115	2016-03-06	120,00	10,00	cartao
116	2016-03-06	5000,00	65,00	cartao

Por que identificar anomalias é uma tarefa importante na preparação de dados?

Big data

Identificação de anomalias

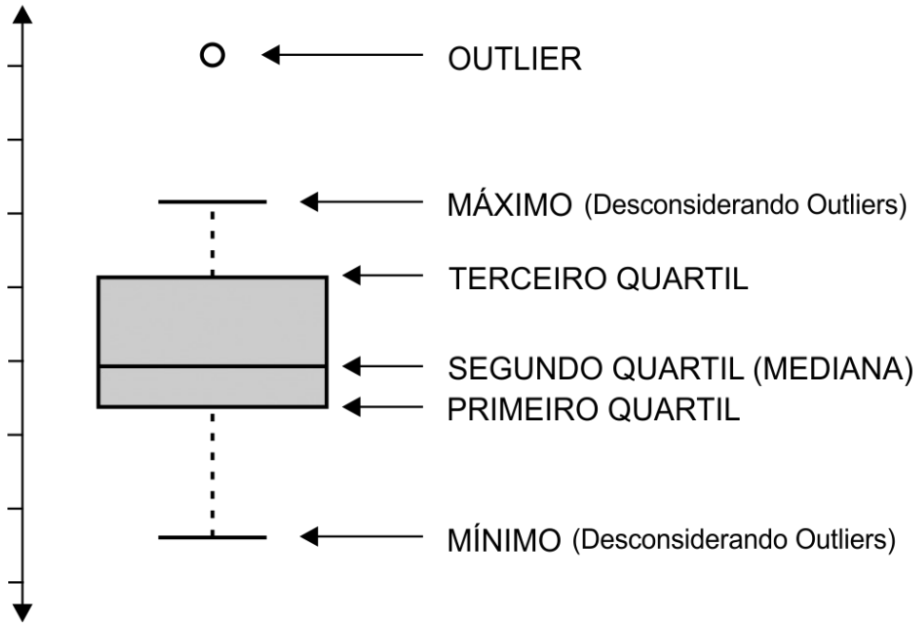
Média dos 10 primeiros registros

R$ 262,50

Média de todos os registros

R$ 693,20

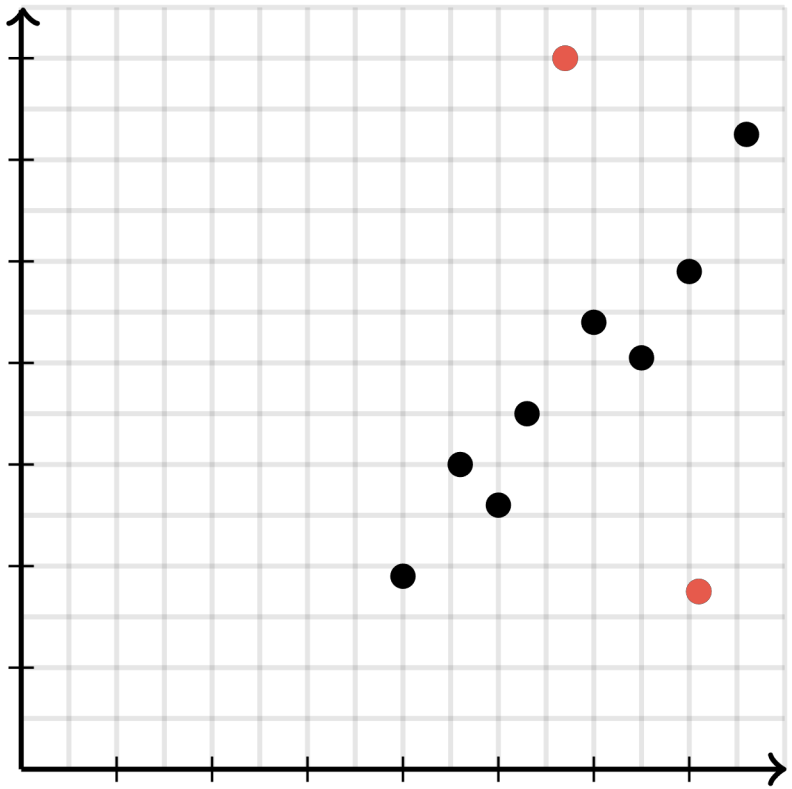
Identificação de anomalias



boxplot

Big data

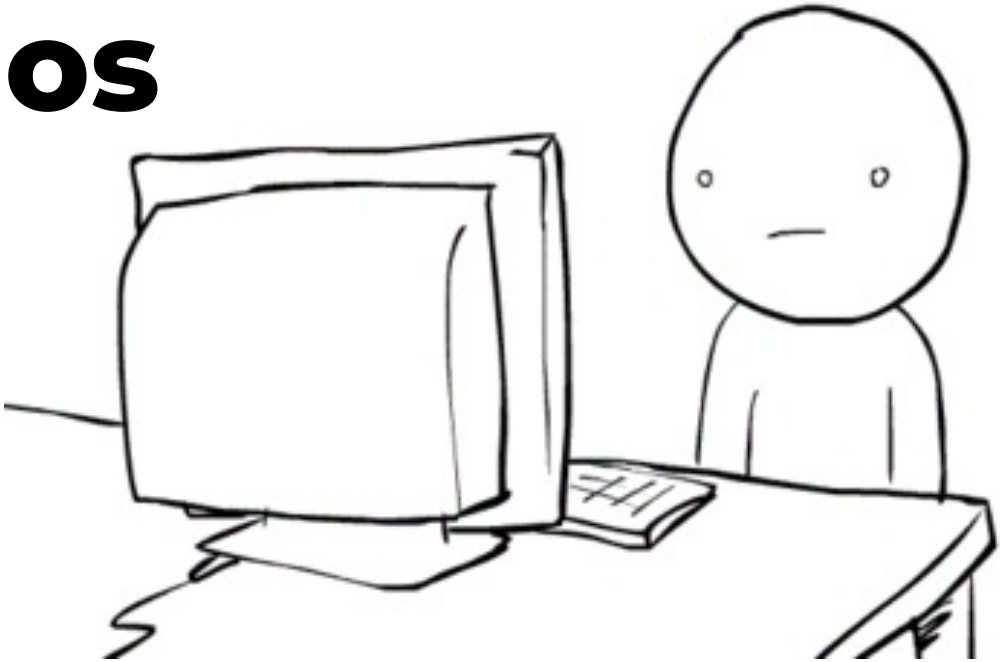
Identificação de anomalias



scatterplot

BIG DATA

Transformação
dos dados



BIG DATA

Variância e Desvio Padrão

Como calcular e para que serve?

BIG DATA

Variância e Desvio Padrão

Como calcular e para que serve?

A variância (V) é útil para determinar o afastamento da média que os dados de um conjunto analisado apresentam. Para isso, determina-se o valor médio das diferenças quadradas da média.

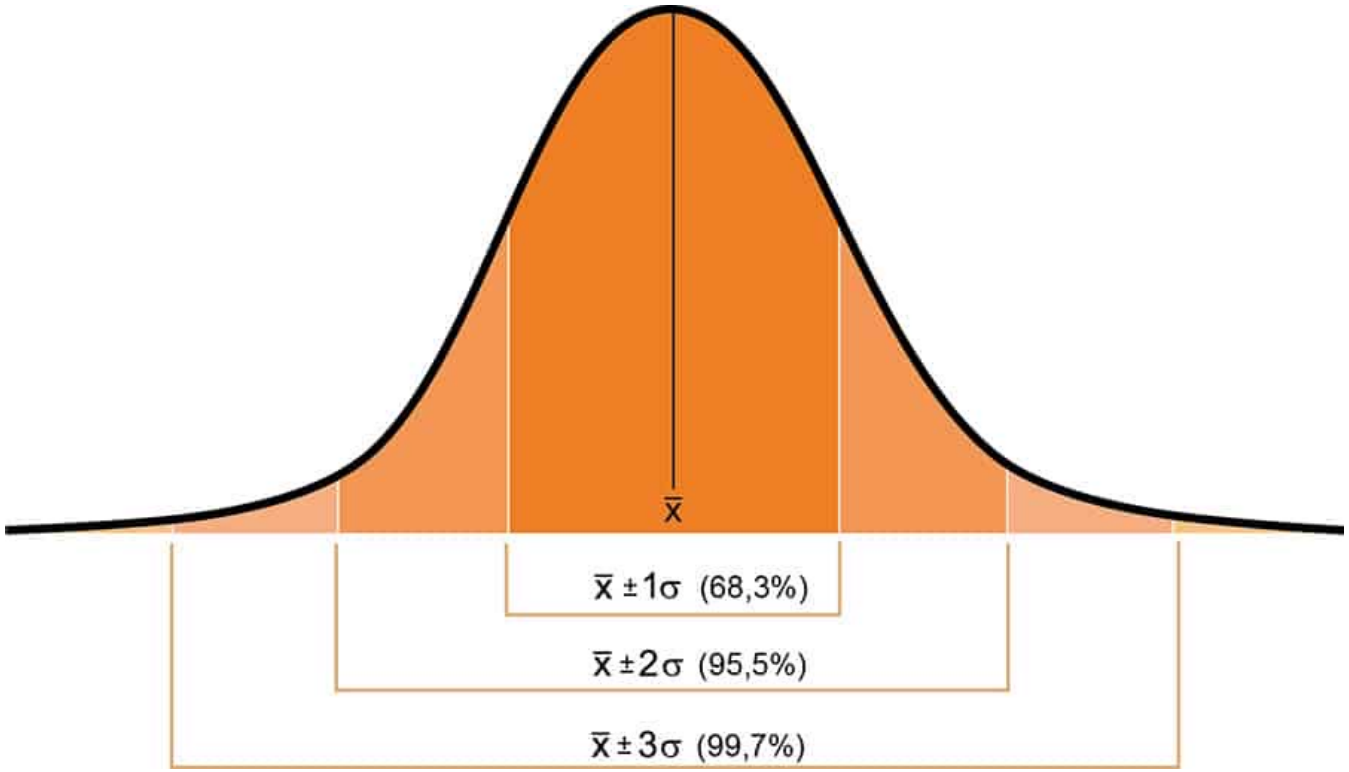
O desvio padrão (DP) é calculado a partir da variância, pois é a raiz quadrada desse parâmetro.

BIG DATA

Variância e Desvio Padrão

Como calcular e para que serve?

BIG DATA



BIG DATA

Transformação dos dados

Normalização dos dados

O objetivo da normalização é mudar os valores das colunas numéricas no conjunto de dados para usar uma escala comum, sem distorcer as diferenças nos intervalos de valores nem perder informações. A normalização também é necessária para alguns algoritmos para modelar os dados corretamente.

BIG DATA

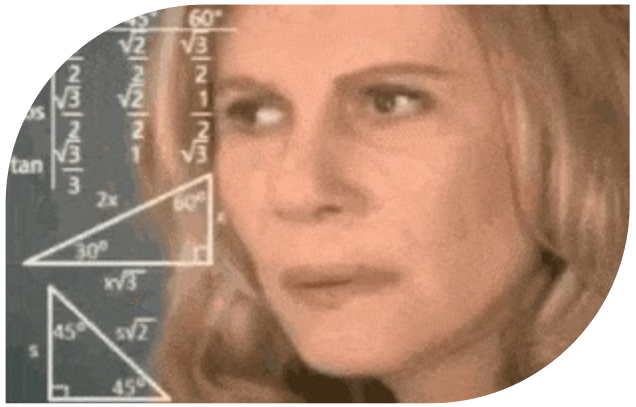
Transformação
dos dados

Normalização dos dados

id	preço
001	20,00
002	180,00
003	30,00
004	65,00
005	52,00
006	23,00
007	97,00
008	82,00
009	261,00
010	347.00

BIG DATA

Transformação dos dados



Normalização dos dados.

$$z = \frac{\mathrm{x} - min(x)}{[\max(x) - min(x)]}$$

Transformação linear, também conhecida como normalização *min-max*

Se o valor de X estiver entre o valor mínimo e o valor máximo, então x' estará entre 0 e 1

BIG DATA

Transformação
dos dados

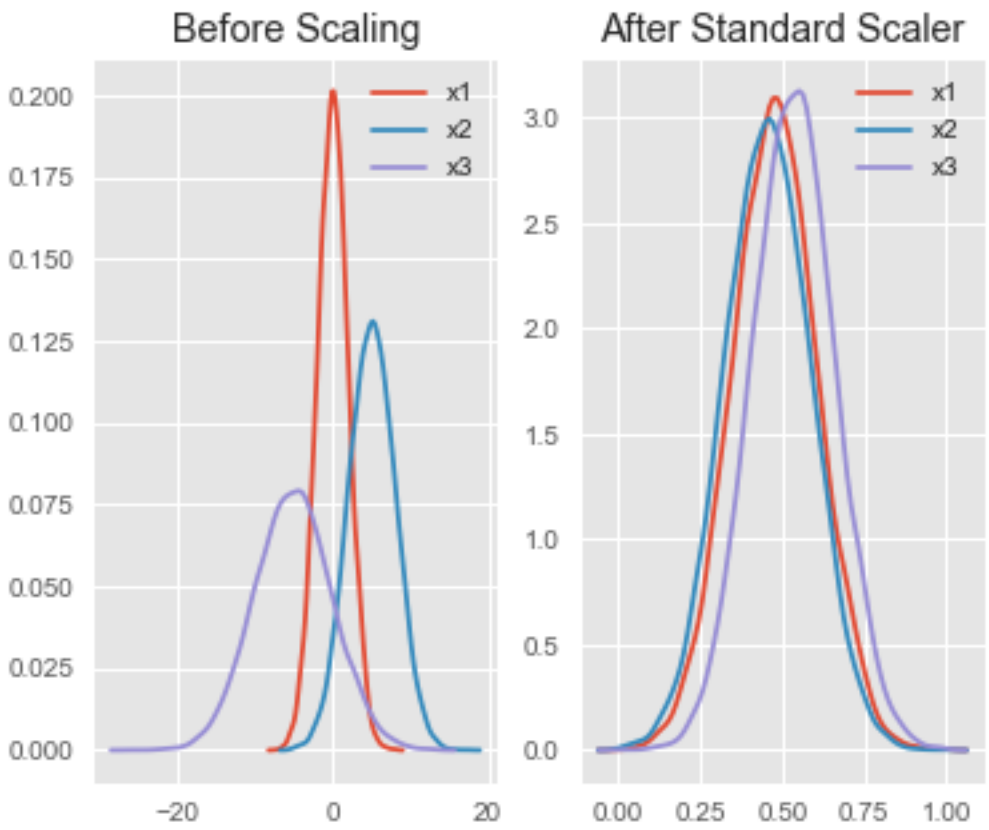
Normalização dos dados

id	preço	preço normalizado
001	20,00	0
002	180,00	0,49
003	30,00	0,03
004	65,00	0,14
005	52,00	0,1
006	23,00	0,01
007	97,00	0,25
008	82,00	0,19
009	261,00	0,74
010	347.00	1

BIG DATA

Transformação
dos dados

Normalização dos dados



Transformação dos dados

Normalização dos dados.

Outros ajustes podem ser necessários:

- Transformação de dados numéricos para categóricos;
- Transformação de dados categóricos para numéricos;
- Agregação de dados, por meio da combinação de dados de diferentes conjuntos em uma única fonte, de forma coerente;
- Criação de novos atributos

BIG DATA

Transformação
dos dados

Normalização dos dados.

EXEMPLO:

ALTURA	PESO	PORTE
1, 65m	75 kg	Pequeno
1,85m	90 kg	Grande
1,88m	89 kg	?

REGRA SUGERIDA:
Somar Altura com o Peso

ALTURA	PESO	PORTE	A + P
1, 65m	75 kg	Pequeno	
1,85m	90 kg	Grande	
1,88m	89 kg	?	

Esta mais próximo de quem?

O problema esta na GRANDEZA DAS UNIDADES

NORMALIZANDO OS DADOS

Dividir cada valor pela sua média

ALTURA	PESO	Pe. Norm	PORTE	Po. Norm	A + P
1, 65m	75 kg		Pequeno		
1,85m	90 kg		Grande		
1,88m	89 kg		?		

BIG DATA

Transformação dos dados

Normalização dos dados.

Diferentes Técnicas

Min-Max

$$n_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Z-Score

$$n_i = \frac{x_i - mean(x)}{std(x)}$$

Tanh

$$n_i = \frac{1}{2}\left[\tanh\left(001\frac{x_i - mean(x)}{std(x)}\right) + 1\right]$$

Soma

$$n_i = \frac{x_i}{\sum x}$$

BIG DATA

Transformação dos dados

Padronização dos dados.

Os valores são centralizados em torno da média com um desvio padrão da unidade. Isso significa que a média do atributo torna-se zero e a distribuição restante tem um desvio padrão por unidade.

BIG DATA

Transformação
dos dados

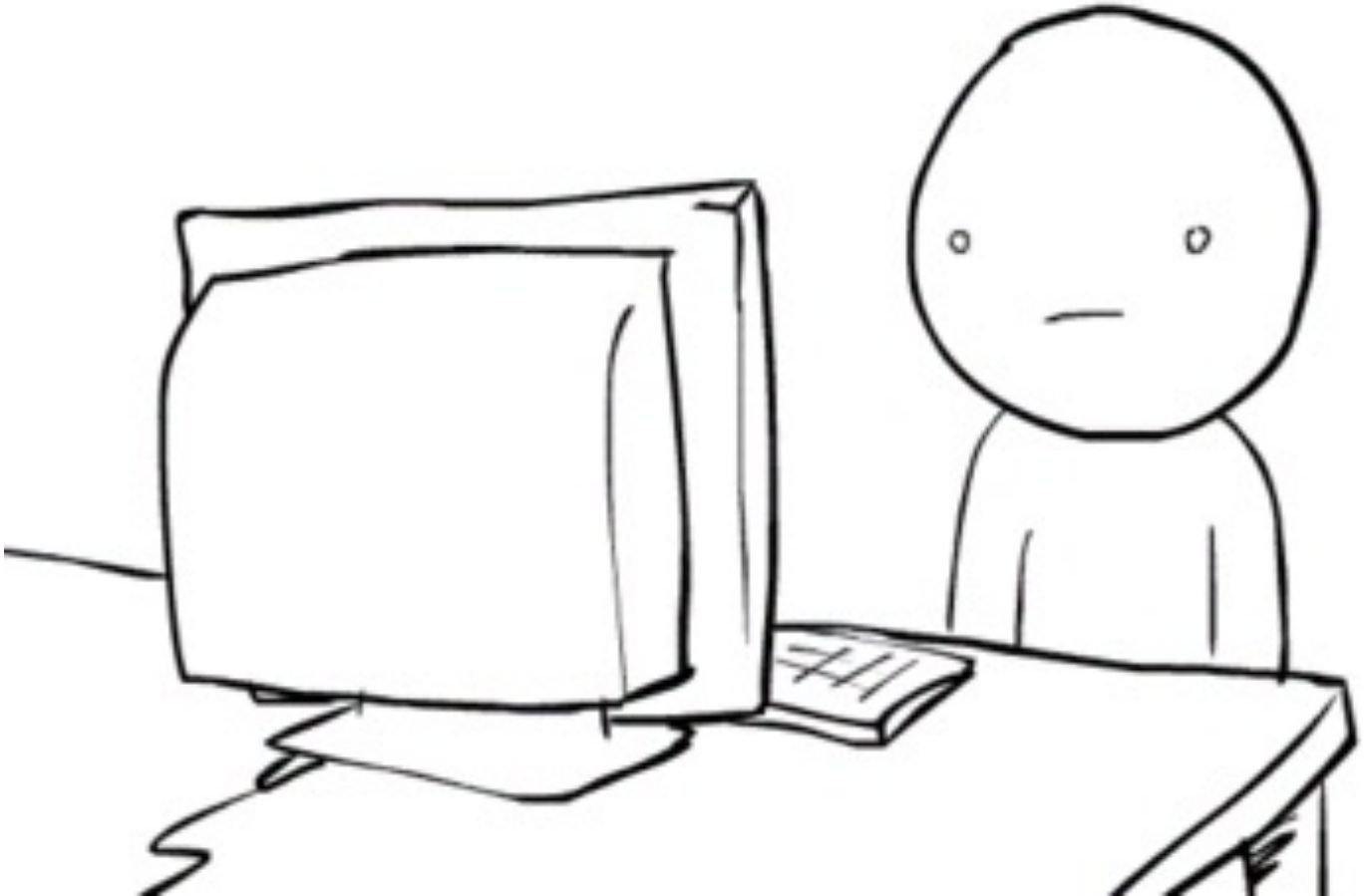
Padronização dos dados.

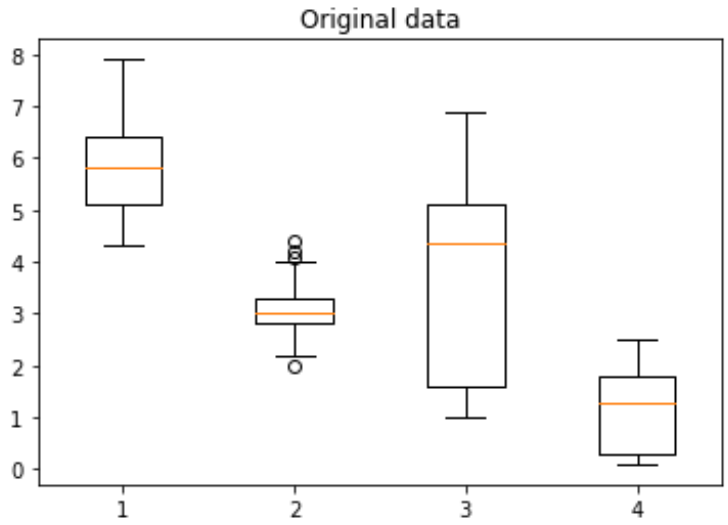
Padronização dos dados normalmente é feita usando a fórmula **z-score**:

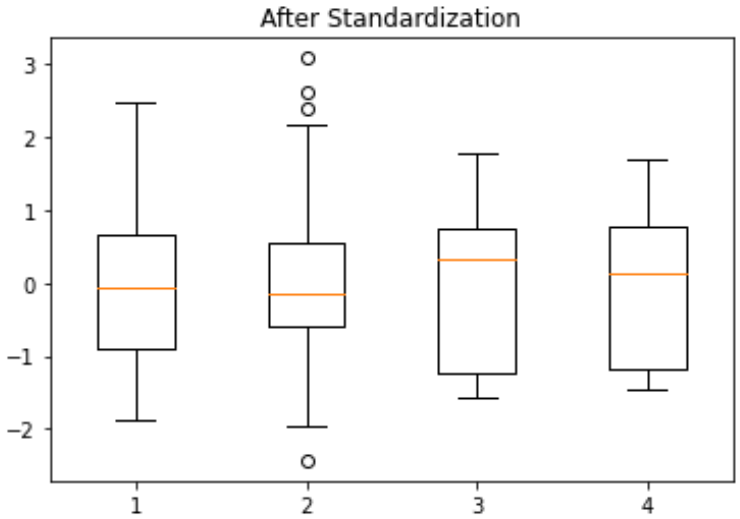
$$z = \frac{x - \mu}{\sigma}$$

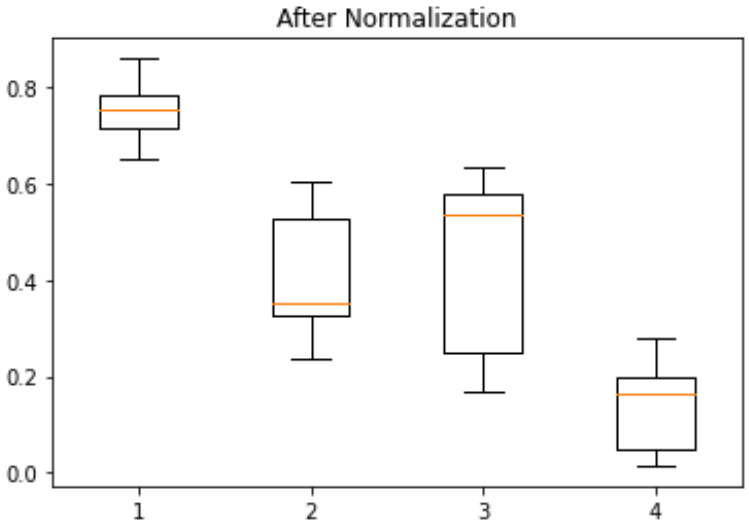
Neste caso, obtemos a **Média sendo 0 e Desvio Padrão como 1**

Professor, dê um exemplo!




Original data
8
7
6
5
4
3
2
1
0
1
2
3
4


After Standardization
3
2
1
0
−1
−2
1
2
3
4


After Normalization
0.8
0.6
0.4
0.2
0.0
1
2
3
4

BIG DATA

Transformação
dos dados

Normalizar ou padronizar as variáveis?