

AULA 3: PRÉ-PROCESSAMENTO DOS DADOS

INTRODUÇÃO A CIÊNCIA DE DADOS NA ENGENHARIA DE PETRÓLEO

Calendário

DATA	ATIVIDADE
26/08	Introdução
02/09	Tipos de dados/ Pré-processamento
09/09	Aula Prática 1
16/09	Aula Prática 2
23/09	Introdução ML
30/09	ML Regressão
07/10	Aula Prática 3
14/10	ML Classificação
21/10	ML Agrupamento
28/10	Feriado
04/11	Aula Prática 4
11/11	Entrega dos Trabalhos

Tópicos

3

□ Pré-processamento dos dados:

▣ Análise Exploratória dos Dados:

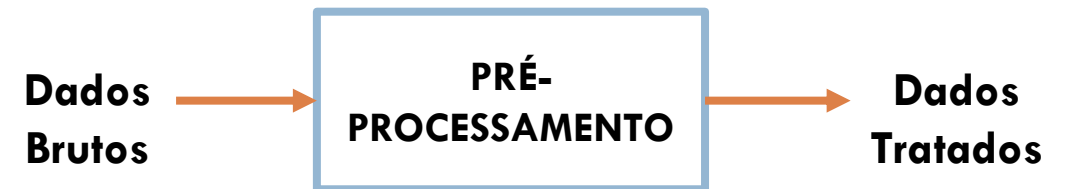
- Estatística Descritiva;
- Histograma;
- Matriz de Correlação.

▣ Limpeza dos dados:

- *Outliers*;
- Valores ausentes.

▣ Redução da Dimensionalidade:

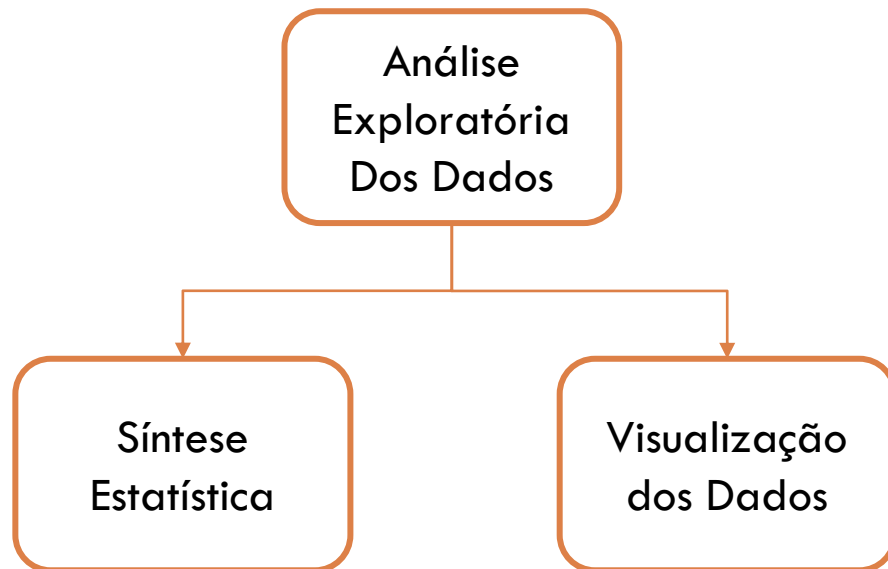
- PCA.



Análise Exploratória dos Dados

4

- Melhor entendimento da natureza e padrões dos dados.



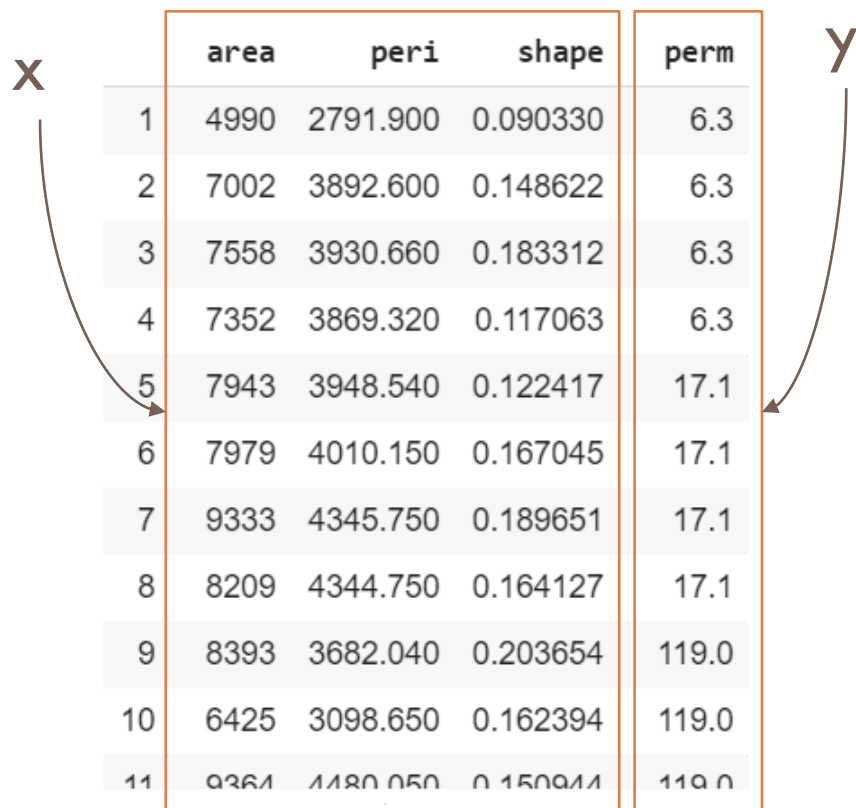
Respostas para:

- Dados estão adequados para aplicar modelos de previsão?
- Quais potenciais variáveis utilizar?
- Como as variáveis se correlacionam entre si?
- Existe redundância de variáveis?

Análise Exploratória dos Dados

5

Dataset: Permeabilidade das rochas de um reservatório de petróleo



	area	peri	shape	perm
1	4990	2791.900	0.090330	6.3
2	7002	3892.600	0.148622	6.3
3	7558	3930.660	0.183312	6.3
4	7352	3869.320	0.117063	6.3
5	7943	3948.540	0.122417	17.1
6	7979	4010.150	0.167045	17.1
7	9333	4345.750	0.189651	17.1
8	8209	4344.750	0.164127	17.1
9	8393	3682.040	0.203654	119.0
10	6425	3098.650	0.162394	119.0
11	9361	4180.050	0.150911	119.0

- Problema: Qual permeabilidade em função das demais variáveis?
- Dataset: 48 registros

Variável	Descrição
Área	Área do espaço dos poros em pixels de 256 por 256
Peri	perímetro em pixels
Shape	perímetro/raiz(área)
Perm	Permeabilidade em mili-Darcy

Análise Exploratória dos Dados - Univariada

6

□ Estatística básica das variáveis:

▣ Medidas de tendência central:

- Média, Mediana e Moda

▣ Dispersão dos Dados

- Range (máximo – mínimo), Quartis, Variância e Desvio-padrão.

Resumo estatístico do *Dataset* Permeabilidade das Rochas

	area	peri	shape	perm
count	48.000000	48.000000	48.000000	48.000000
mean	7187.729167	2682.211938	0.218110	415.450000
std	2683.848862	1431.661164	0.083496	437.818226
min	1016.000000	308.642000	0.090330	6.300000
25%	5305.250000	1414.907500	0.162262	76.450000
50%	7487.000000	2536.195000	0.198862	130.500000
75%	8869.500000	3989.522500	0.262670	777.500000
max	12212.000000	4864.220000	0.464125	1300.000000

Desvio-padrão

Range
(max – min)

mediana
quartis

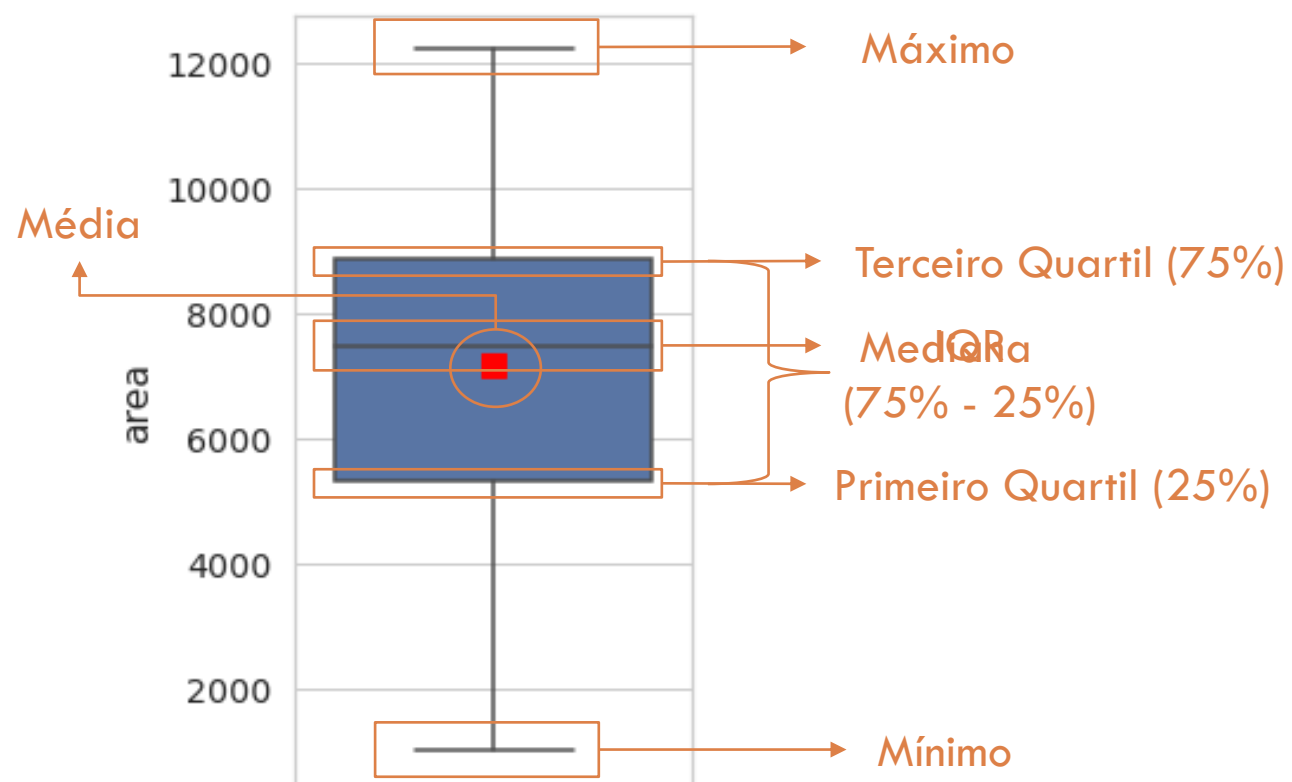
○ Por que analisar as variáveis individualmente?

- Cada variável possui uma natureza estatística diferente.

Boxplot

7

- Resumo de análise descritiva da variável área:

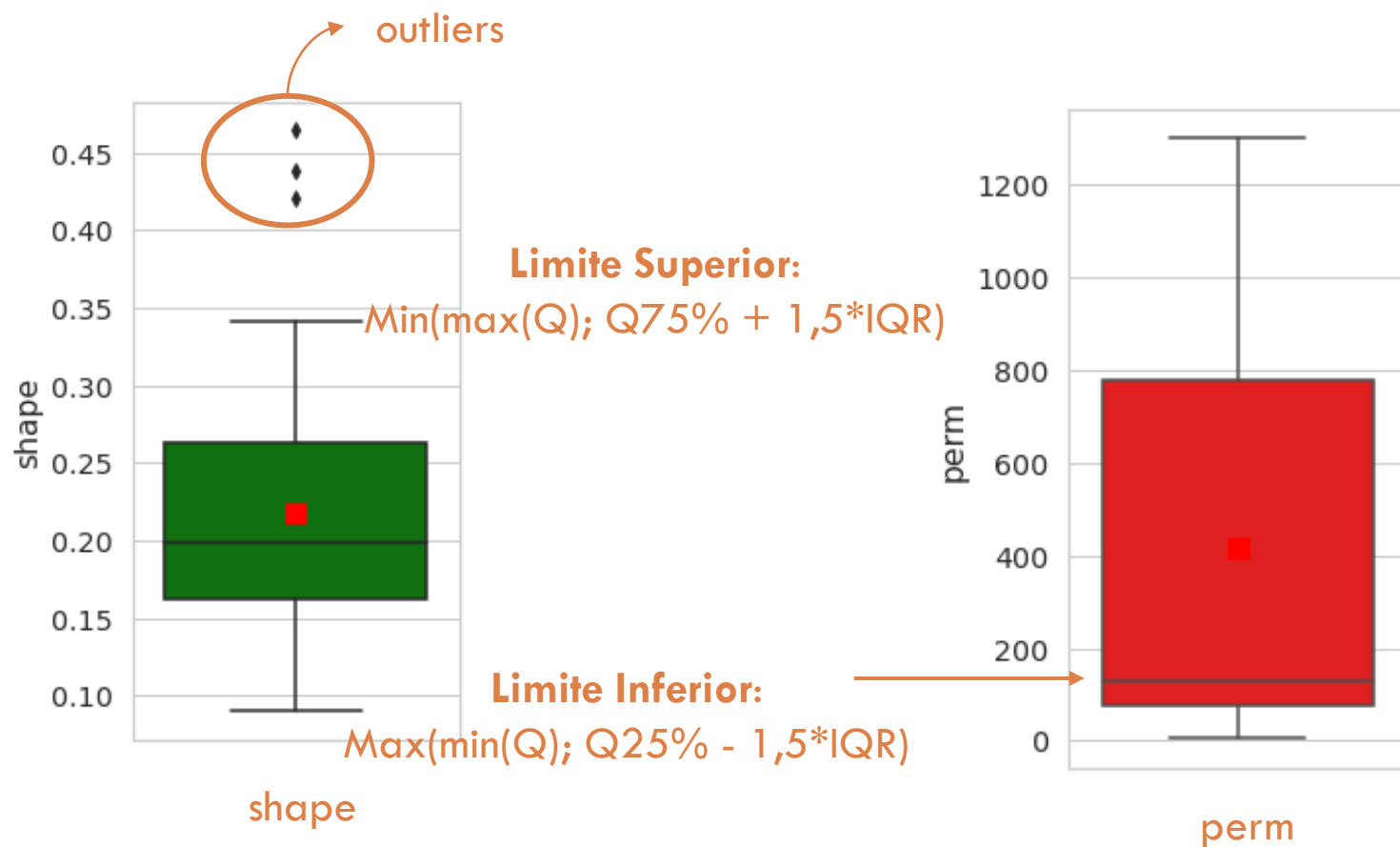


	area
count	48.000000
mean	7187.729167
std	2683.848862
min	1016.000000
25%	5305.250000
50%	7487.000000
75%	8869.500000
max	12212.000000

Boxplot

8

- Resumo de análise descritiva das variáveis shape e perm:

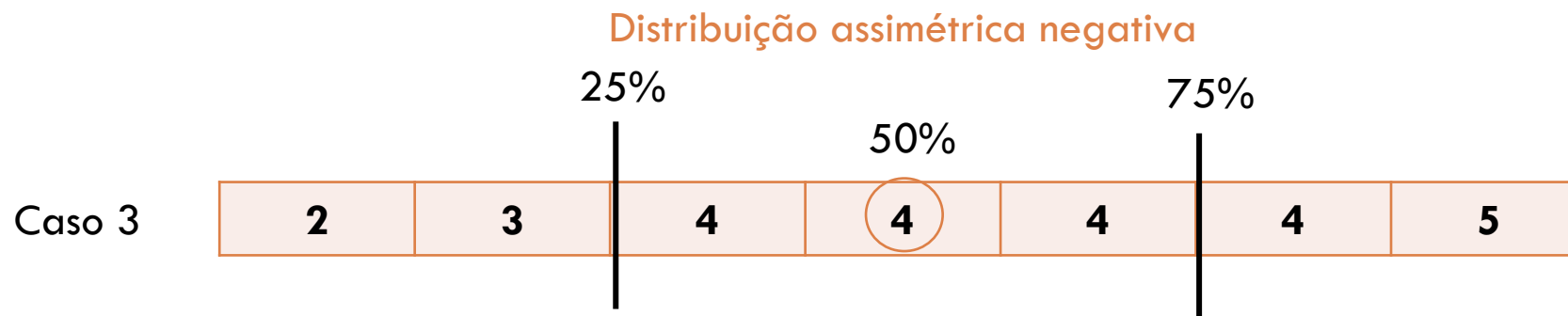
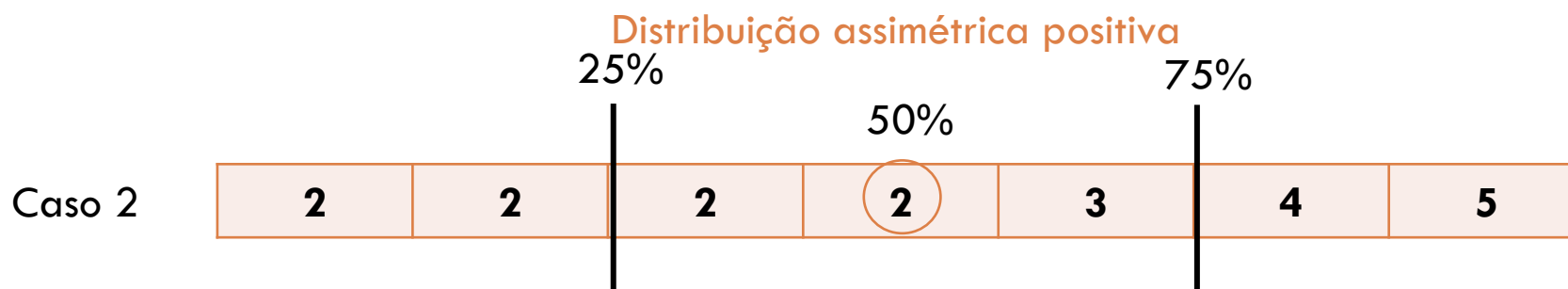
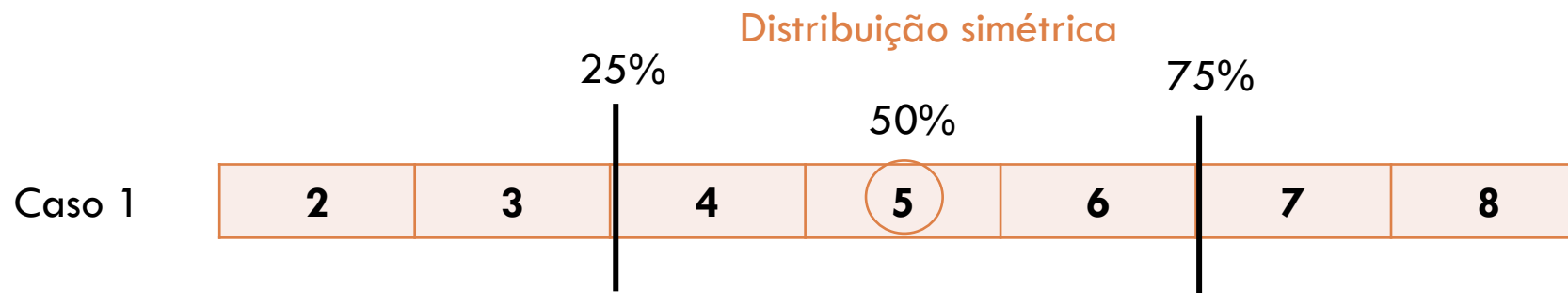


Simetria dos dados:

- Distribuição simétrica:
 - Mediana no centro do retângulo.
- Distribuição assimétrica positiva:
 - Mediana próxima ao primeiro quartil.
- Distribuição assimétrica negativa:
 - Mediana próxima ao terceiro quartil.

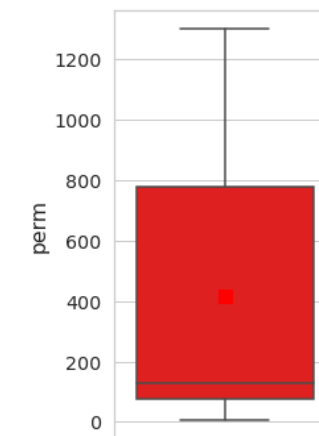
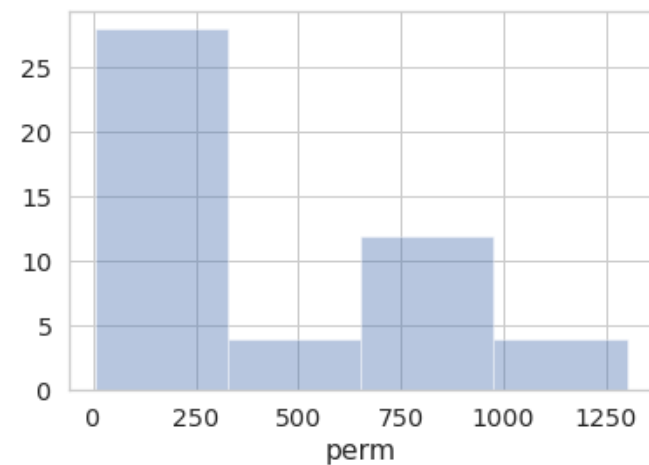
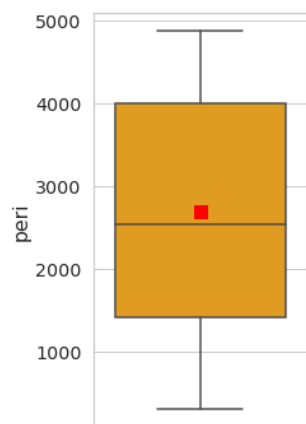
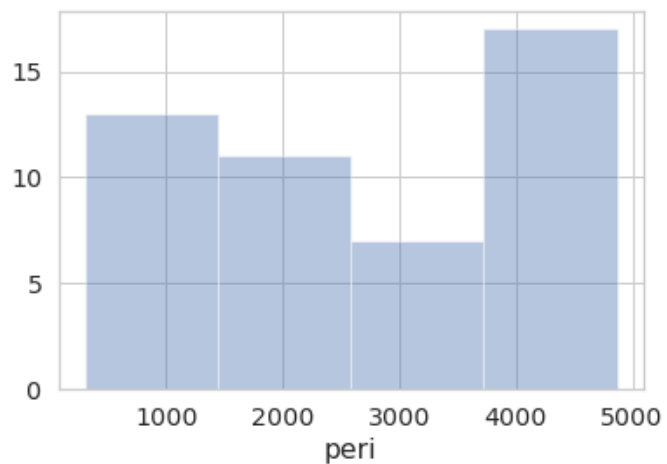
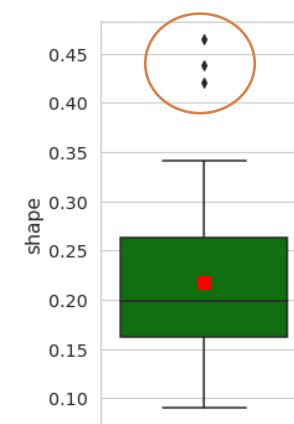
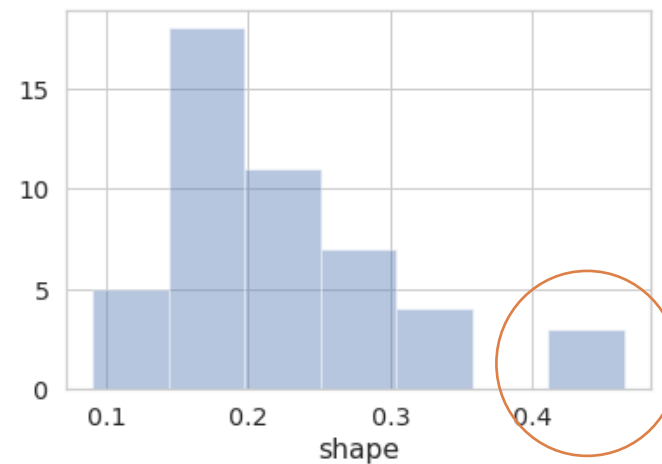
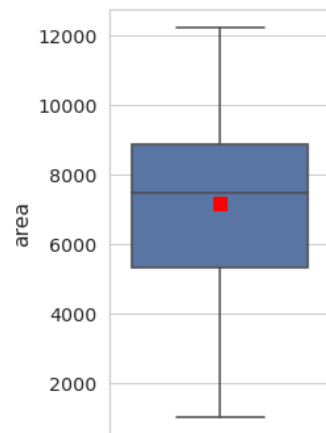
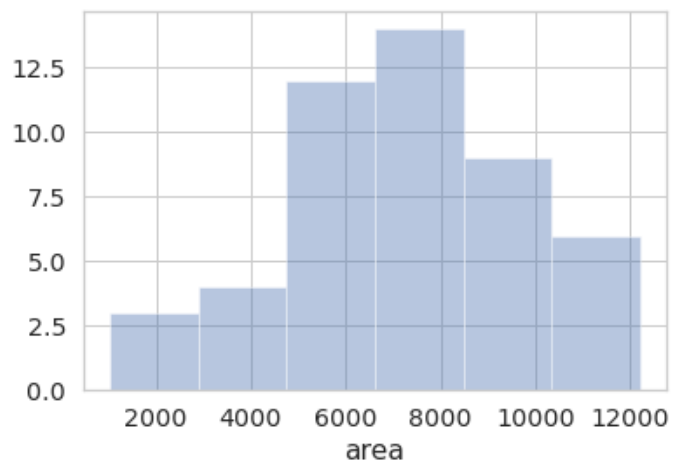
Simetria dos dados

9



Histogramas

10



Matriz de Correlação

11

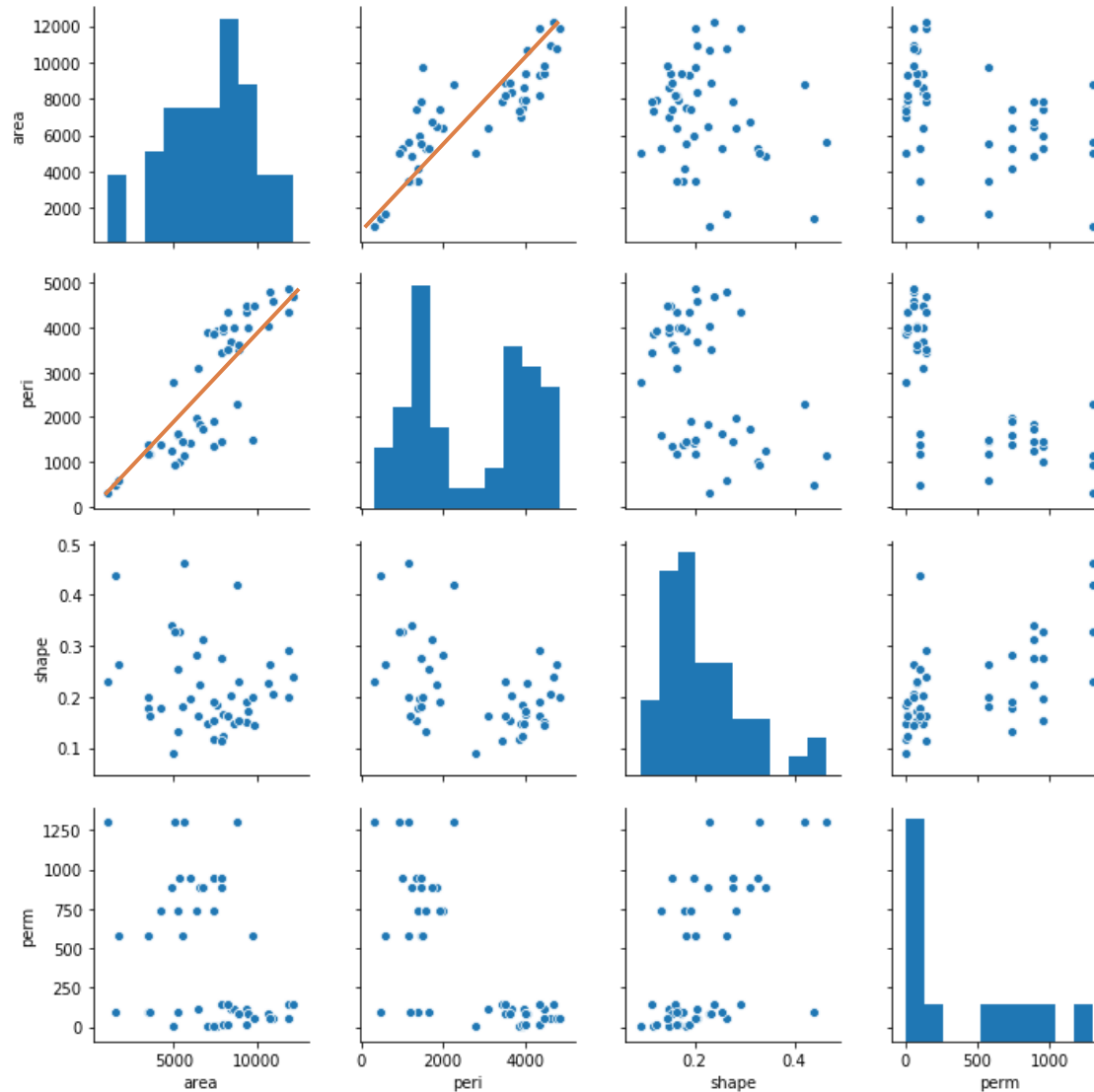
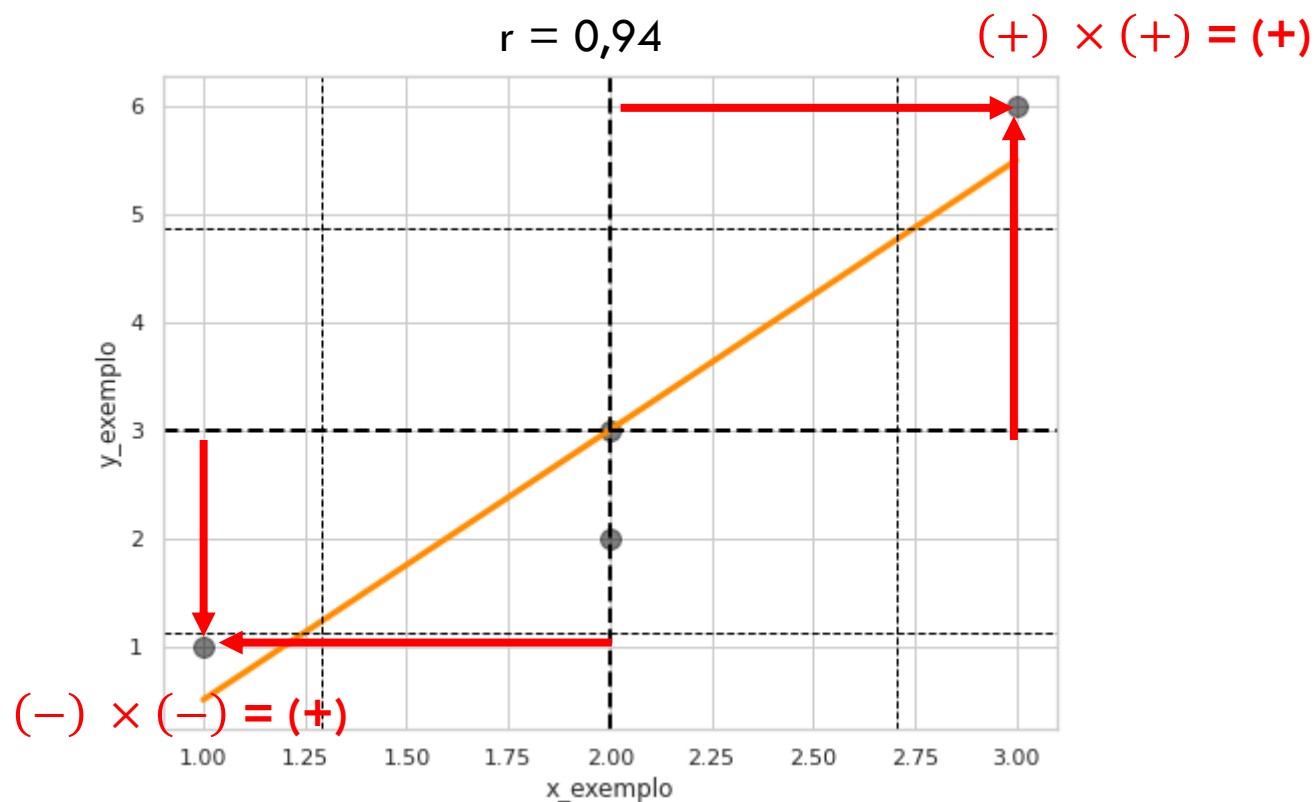


Gráfico de Dispersão

- Quais variáveis parecem estar correlacionadas?
 - Perímetro e área

Coeficiente de Correlação

12



Correlação de Pearson

- Quanto a relação entre duas variáveis pode ser descrita por uma reta.

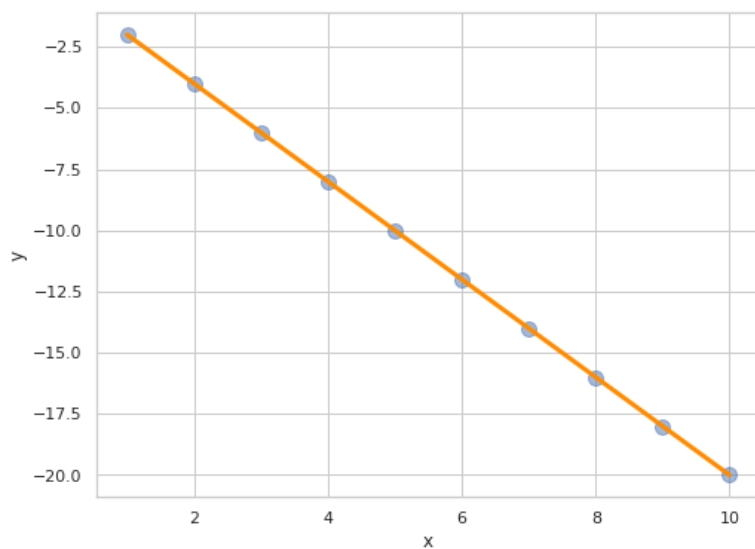
$$r = \frac{1}{n-1} \sum \left(\underbrace{\frac{x_i - \bar{x}}{s_x}}_{Z_{xi}} \right) \left(\underbrace{\frac{y_i - \bar{y}}{s_y}}_{Z_{yi}} \right)$$

- Z_i : quanto a medida se afasta da média em termos de desvios padrão.

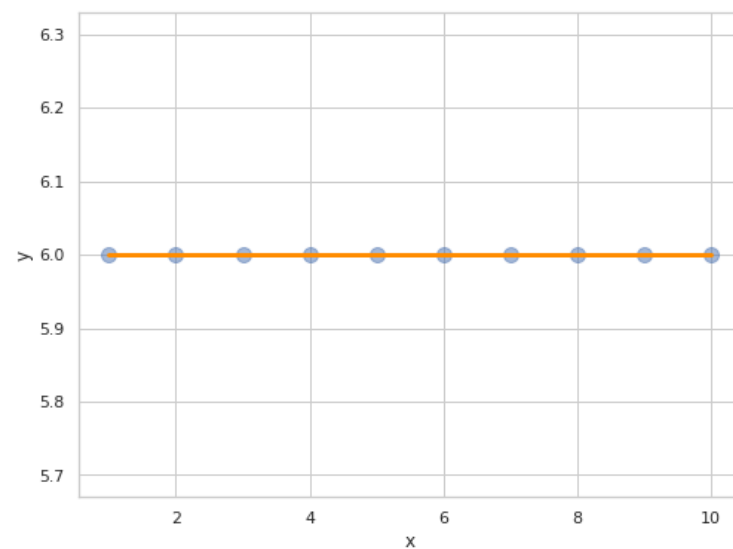
$$-1 \leq r \leq +1$$

Coeficiente de Correlação

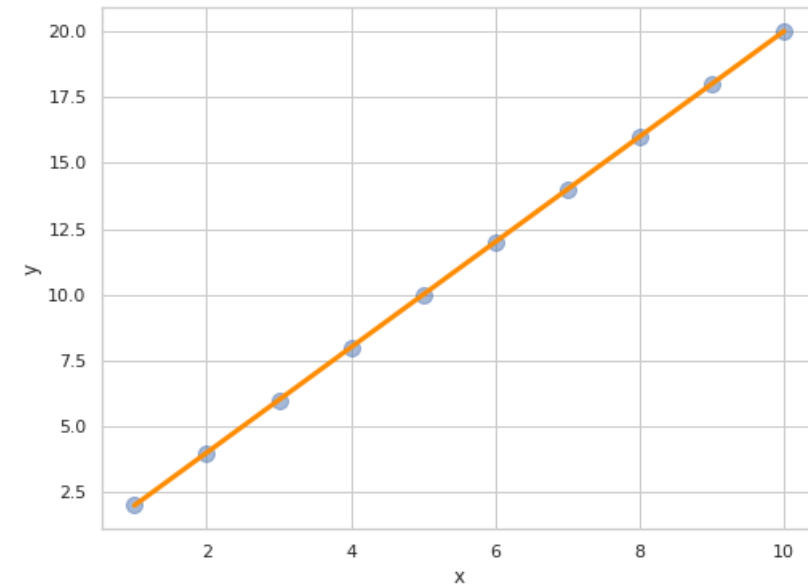
13



$r = -1$



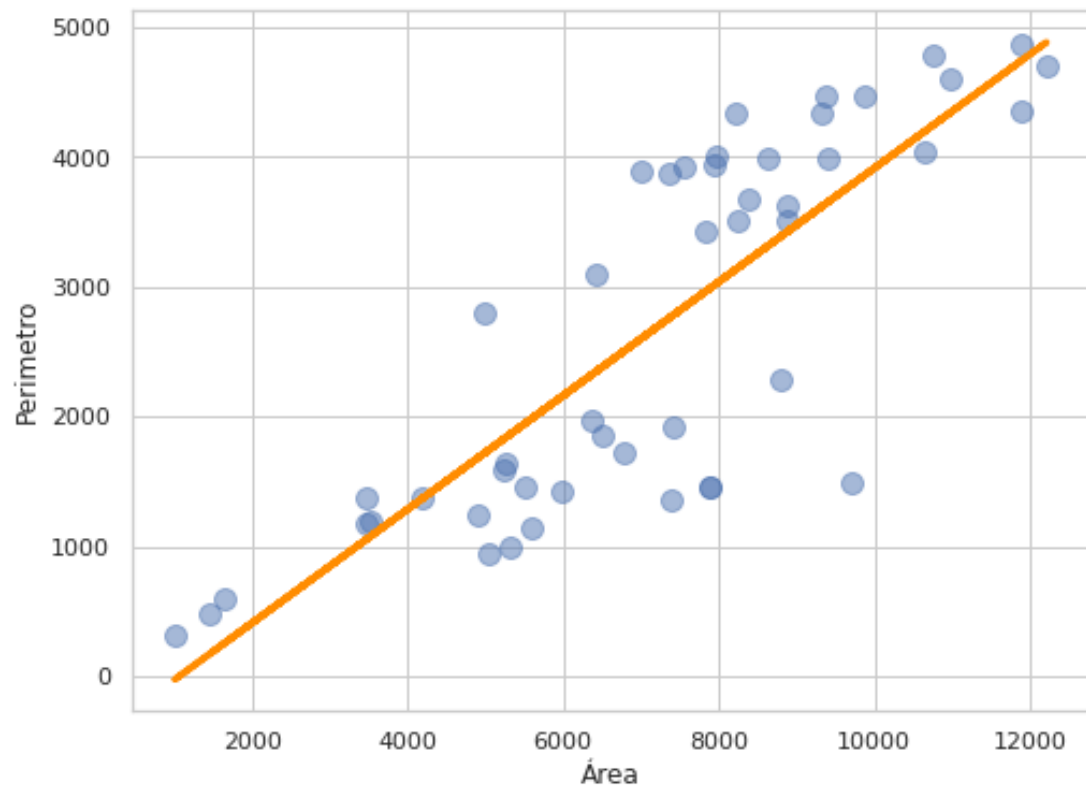
$r = 0$



$r = +1$

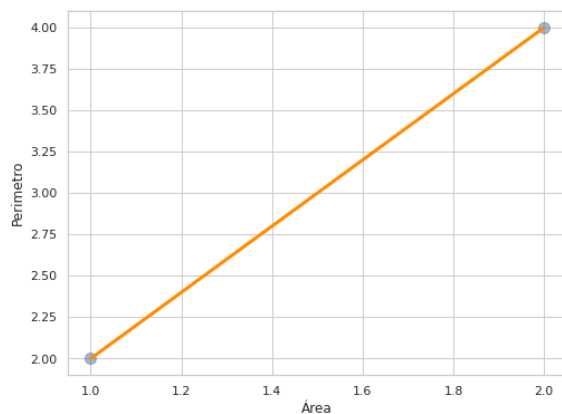
Correlação nas variáveis área e perímetro

14

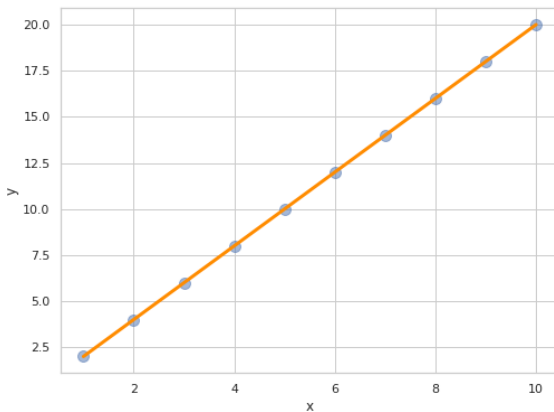


R	0,82
Valor-p	$7,50 \times 10^{-13}$

○ $\downarrow \text{valor} - p \rightarrow \uparrow \text{confiança}$



R	1
Valor-p	1

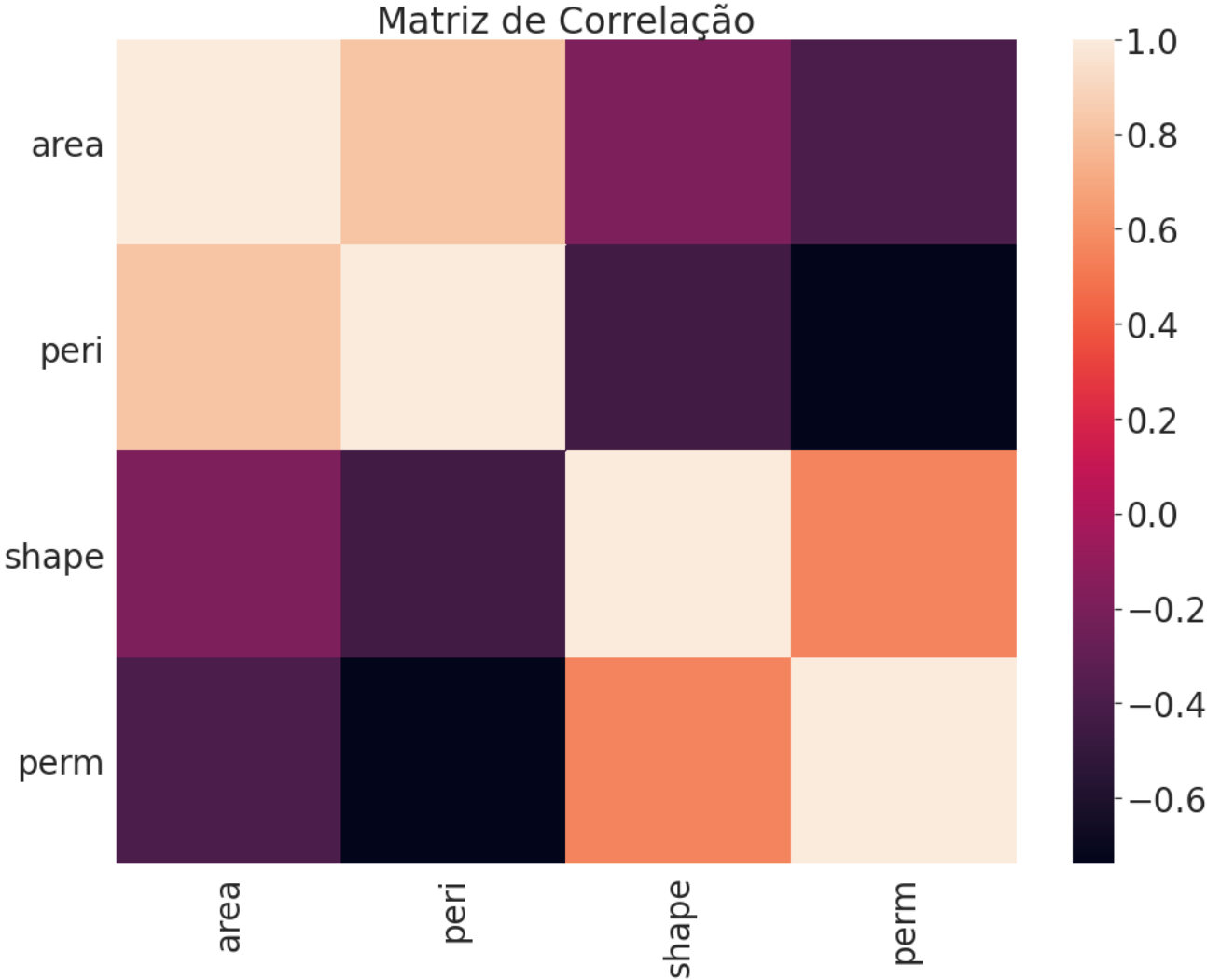


R	1
Valor-p	0

Matriz de Correlação

15

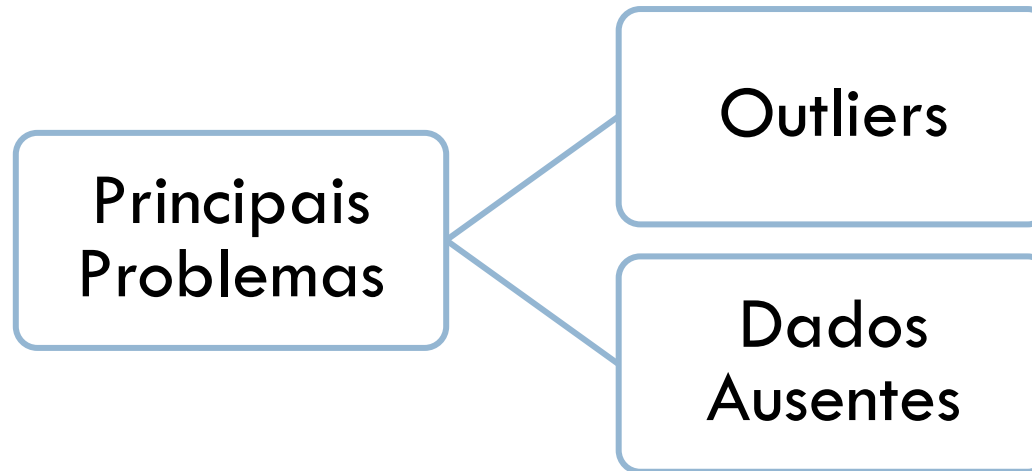
	AREA	PERI	SHAPE	PERM
AREA	1.00	0.82	-0.18	-0.40
PERI	0.82	1.00	-0.43	-0.74
SHAPE	-0.18	-0.43	1.00	0.56
PERM	-0.40	-0.74	0.56	1.00



Limpeza dos dados

16

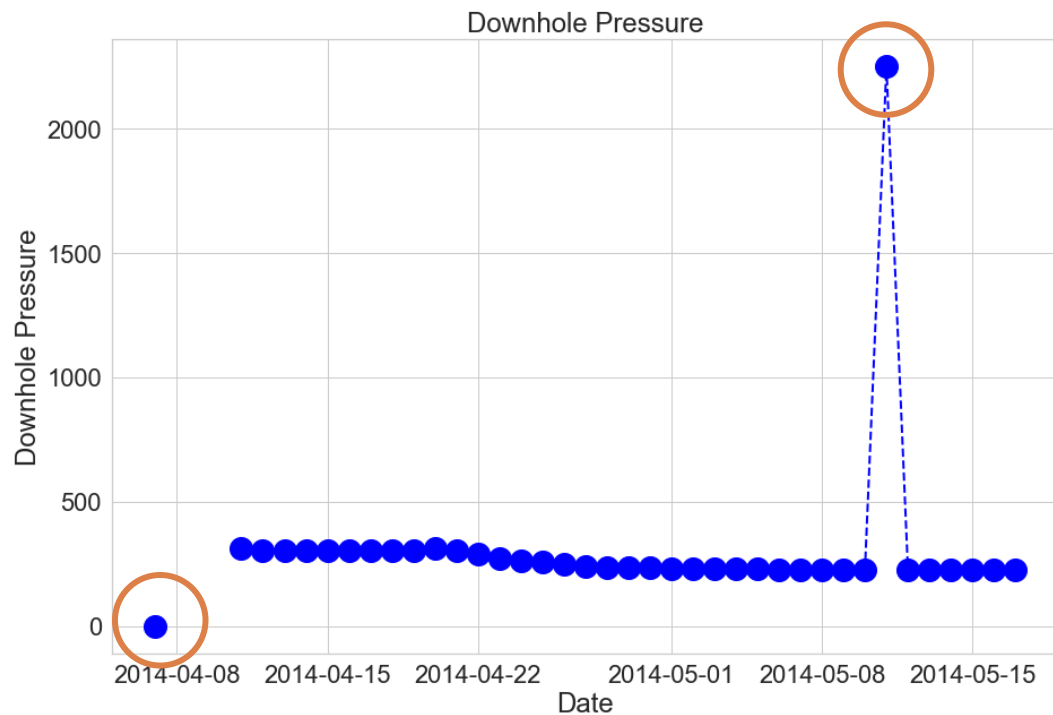
- “Garbage in, garbage out”.
- Etapa fundamental ANTES de qualquer análise preditiva nos dados.



Limpeza dos dados

17

- Considerar um *dataset* com P_PDG e T_PDG:



Dados anômalos
(*outliers*)

DATEPRD	AVG_DOWNHOLE_ PRESSURE	AVG_DOWNHOLE_ TEMPERATURE	AVG_DP_ TUBING
07-abr-14	0,00	0,00	0,00
08-abr-14			
09-abr-14			
10-abr-14			
11-abr-14	310,38	96,88	277,28
12-abr-14	303,50	96,92	281,45
13-abr-14	303,53	96,96	276,03
14-abr-14	303,78	96,97	282,79
15-abr-14	303,86	97,02	289,94
16-abr-14	303,79	97,07	299,67
17-abr-14	304,34	96,92	282,90
18-abr-14	304,85	96,72	273,70

Dados ausentes

Outliers

18

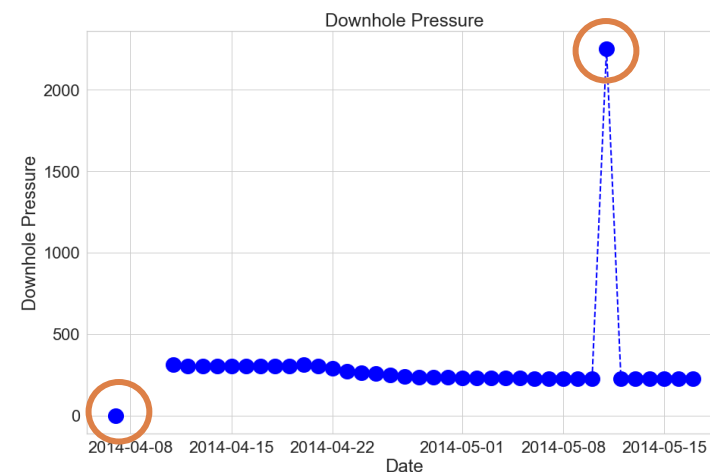
- Registros significativamente destoantes dos demais em um mesmo conjunto de dados.
- Dados não acurados, ruídos ou inconsistentes.

Qual origem?

- Equipamentos com falhas.
- Erros de digitação ou humanos
- Erros de transmissão.
- Formatos inconsistentes (por exemplo, data).
- Fenômenos naturais.

Como resolver?

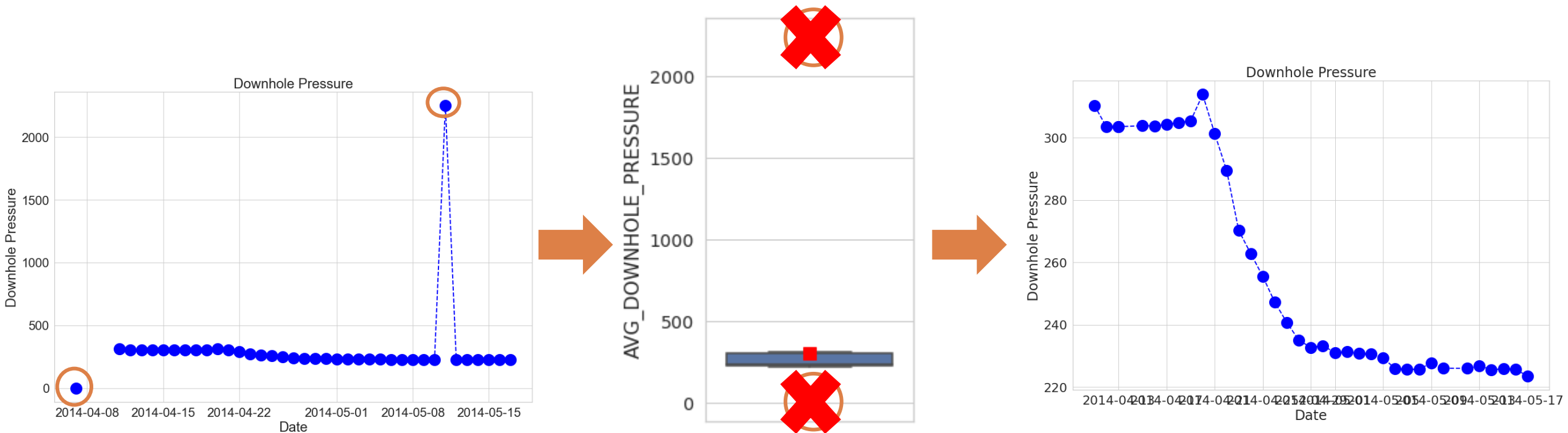
- Visualmente.
- Métodos estatísticos.
- Métodos baseados em distância.
- Métodos baseados em densidade.
- Técnicas de Agrupamento.



**Dados anômalos
(outliers)**

Remoção de Outliers

19



Outliers:

$$Q < Q_{25\%} - 1,5 \cdot IQR$$

$$Q > Q_{75\%} + 1,5 \cdot IQR$$

Dados Ausentes

20

Qual origem?

- Equipamentos param de funcionar.
- Medidas não estão sempre disponíveis.
- Remoção de Outliers.

Como resolver?

- Remover linha/coluna.
- Métodos de inserção

Por que precisamos lidar com dados ausentes?

DATEPRD	AVG_DOWNHOLE_ PRESSURE	AVG_DOWNHOLE_ TEMPERATURE	AVG_DP_ TUBING
07-abr-14	0,00	0,00	0,00
08-abr-14			
09-abr-14			
10-abr-14			
11-abr-14	310,38	96,88	277,28
12-abr-14	303,50	96,92	281,45
13-abr-14	303,53	96,96	276,03
14-abr-14	303,78	96,97	282,79
15-abr-14	303,86	97,02	289,94
16-abr-14	303,79	97,07	299,67
17-abr-14	304,34	96,92	282,90
18-abr-14	304,85	96,72	273,70

Dados ausentes

Dados Ausentes

21

1

Remover
linha/coluna

2

Preencher o valor
ausente
manualmente

Não é viável
para muitos
dados

3

Utilizar a média ou
mediana de todas
as amostras de
determinada classe

Normalmente
distribuídos: média
Não normalmente
distribuídos: mediana

4

Preencher com
medida de
tendência central
(média ou mediana)

5

Preencher com valor
de algum modelo

Regressão,
árvore de
decisão,
interpolação,
médias móveis

- Métodos 3 a 5 – insere bias nos dados
- Método 5 é a estratégia mais utilizada.

Dimensionalidade dos Dados

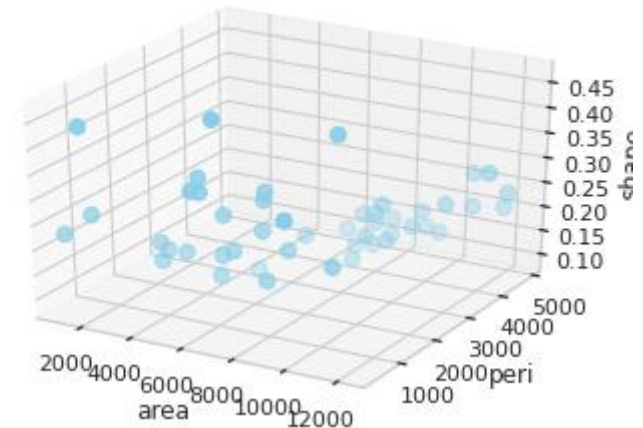
22

- Voltando para conjunto de dados iniciais referente a permeabilidade das rochas.

	area	peri	shape	perm
1	4990	2791.900	0.090330	6.3
2	7002	3892.600	0.148622	6.3
3	7558	3930.660	0.183312	6.3
4	7352	3869.320	0.117063	6.3
5	7943	3948.540	0.122417	17.1
6	7979	4010.150	0.167045	17.1
7	9333	4345.750	0.189651	17.1
8	8209	4344.750	0.164127	17.1
9	8393	3682.040	0.203654	119.0
10	6425	3098.650	0.162394	119.0
11	9364	4480.050	0.150944	119.0

Quais são as dimensões do *dataset*?

- Area
- Peri
- Shape
- perme



Como representar graficamente dados com mais de 3 dimensões?

Dataset: Permeabilidade das rochas

Dimensionalidade dos Dados

23

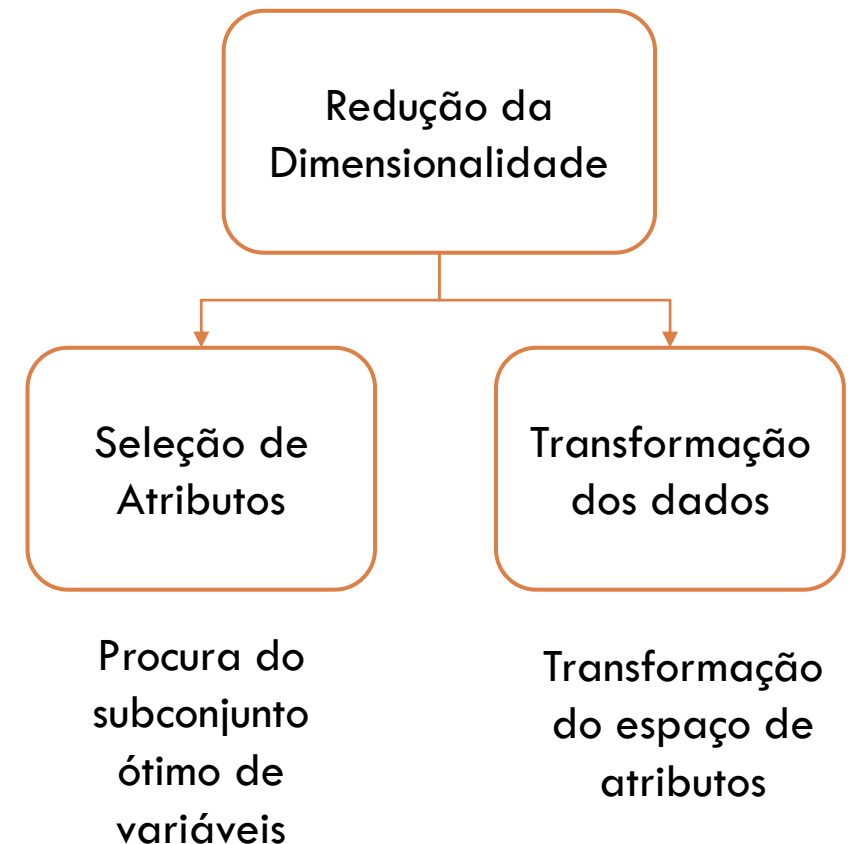
- Qual o problema da alta dimensionalidade dos dados?

Maior capacidade de armazenagem e computacional

Dificuldade de visualização

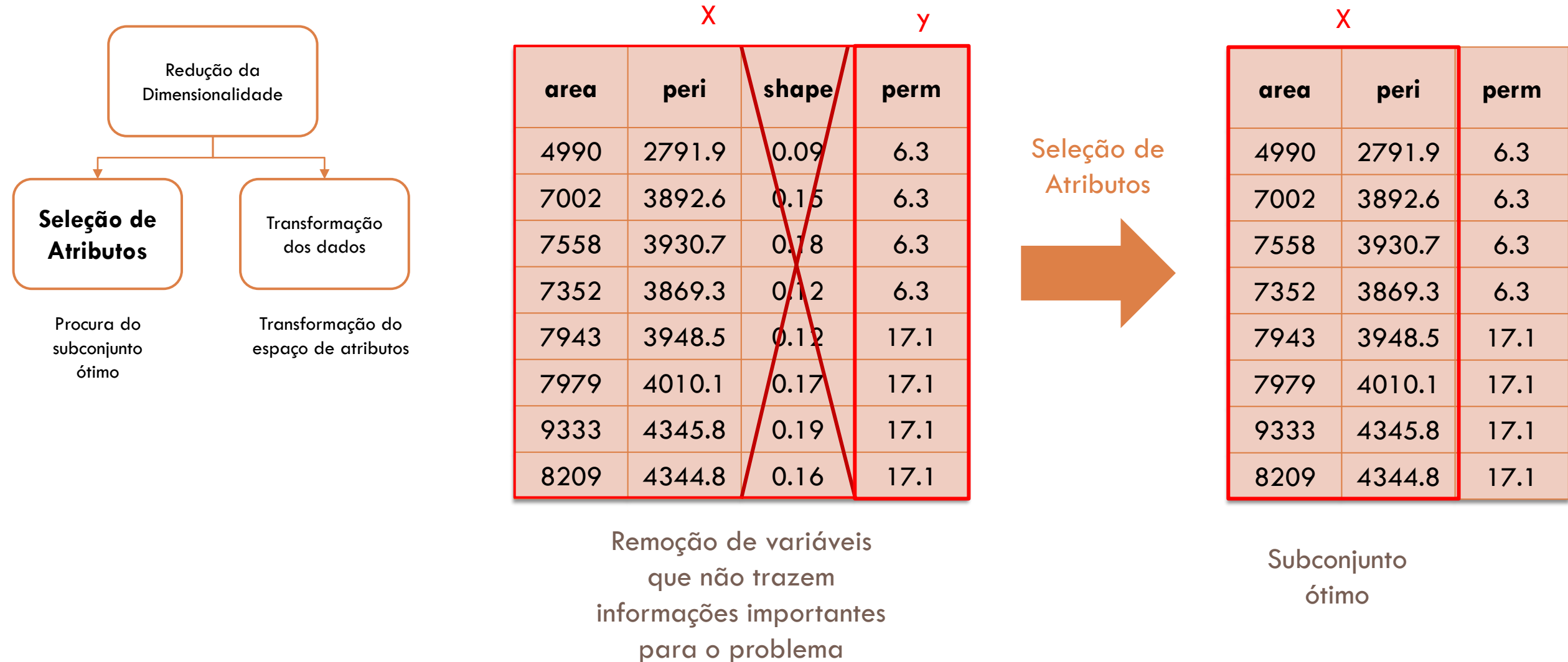
Maior dimensão → maior superajuste dos modelos.

Maior número de registros para treinar os modelos



Dimensionalidade dos Dados

24



Dimensionalidade dos Dados

25



X			Y
area	peri	shape	perm
4990	2791.9	0.09	6.3
7002	3892.6	0.15	6.3
7558	3930.7	0.18	6.3
7352	3869.3	0.12	6.3
7943	3948.5	0.12	17.1
7979	4010.1	0.17	17.1
9333	4345.8	0.19	17.1
8209	4344.8	0.16	17.1

Transformação dos dados



X		
PC1	PC2	perm
0.53	- 0.64	6.3
- 0.55	- 0.65	6.3
- 0.72	- 0.52	6.3
- 0.64	- 0.43	6.3
- 0.87	- 0.45	17.1
- 1.41	- 0.26	17.1
- 1.10	- 0.56	17.1
- 0.82	- 0.18	17.1

Transformação em combinações lineares das variáveis

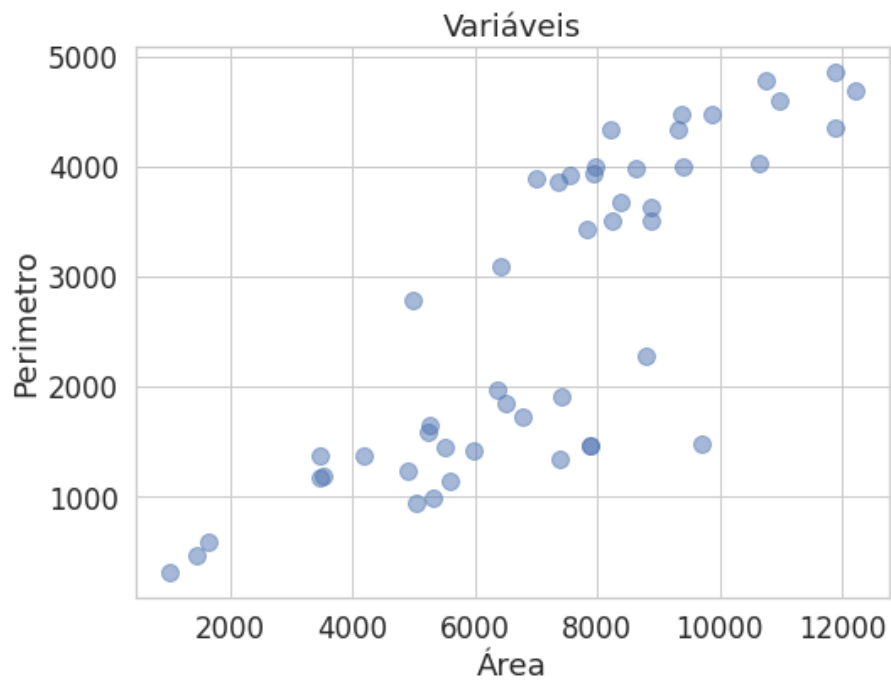
Transformação do espaço de atributos

- Método mais comum para transformação dos dados.
- Etapas:
 1. Normalização;
 2. Encontrar os componentes principais;
 3. Avaliar quanto cada componente explica a variância total dos dados.

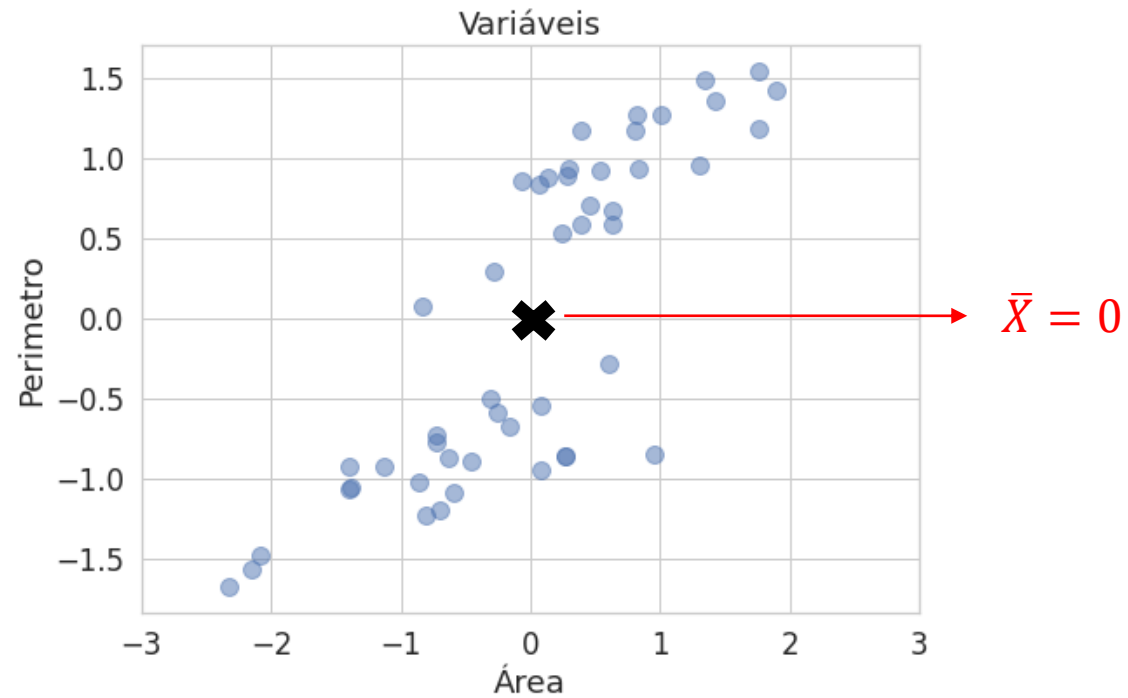
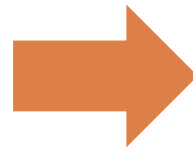
1 - Normalização

27

□ Passo 1: Normalizar os dados



Sem normalização



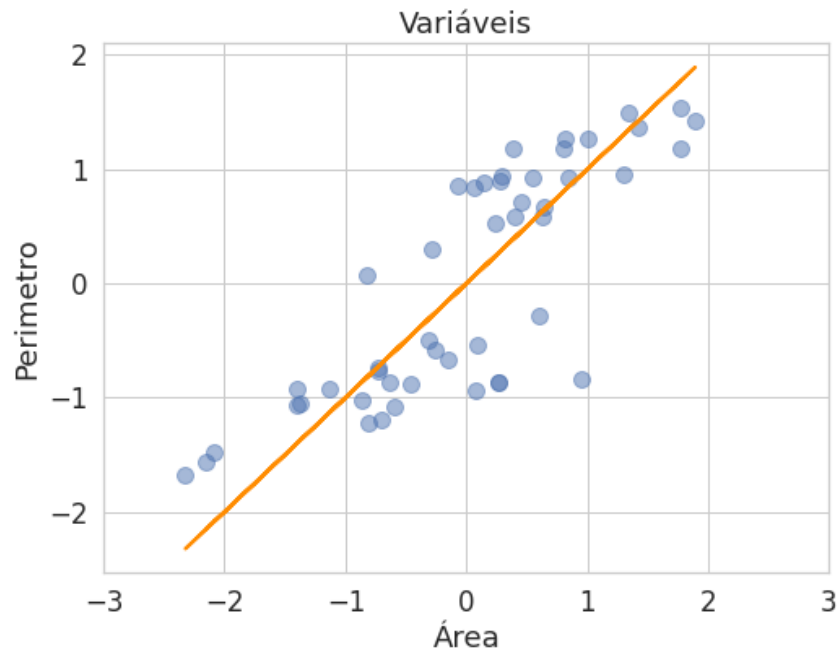
Com normalização

Distribuição dos dados igual, mas a média é 0

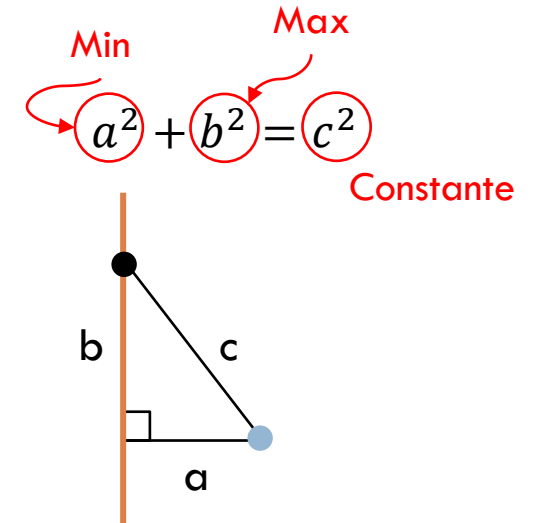
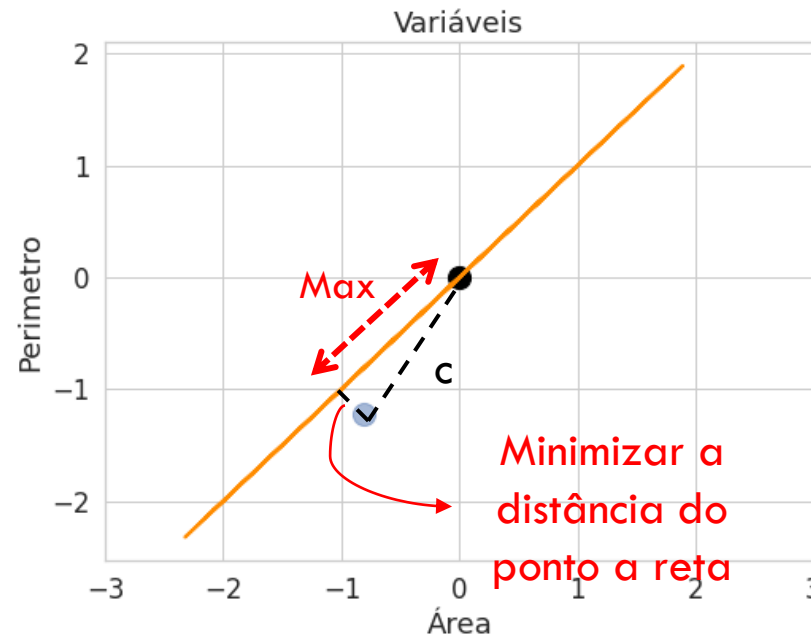
2 – Encontrar Componente Principal 1 (CP1)

28

- Passo 2: Encontrar reta que passa pela origem que maximize a variância.



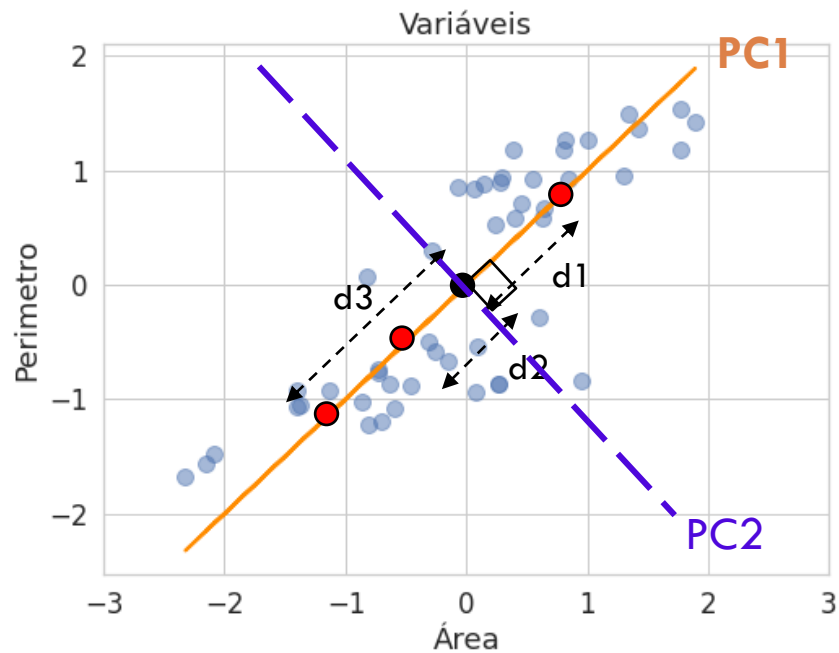
Reta que maximize a soma dos quadrados das distâncias
Max(SS(distancias))



3 – Encontrar Componentes Principais

29

Passo 3: Encontrar Componentes Principais



Soma do quadrado das distâncias:

$$SS(dist) = d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2$$

$$\max SS(dist)$$

Autovalor do
PC 1 (λ_1)

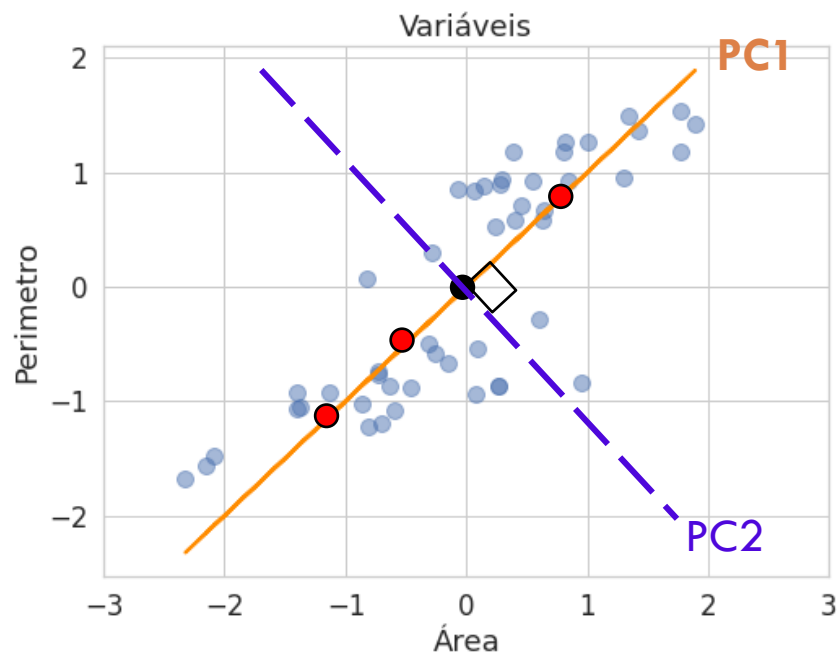
$$PC1 = a_{i1}X_1 + a_{i2}X_2 \longrightarrow \text{Componente Principal 1}$$

$$PC2 = b_{i1}X_1 + b_{i2}X_2 \longrightarrow \text{Componente Principal 2}$$

4 – Avaliar o quanto cada PC explica a variância dos dados

30

□ Passo 3: Encontrar Componentes Principais



Como saber quanto cada componente explica o total da variância dos dados?

Componente Principal 1 $\frac{\lambda_1}{n-1}$

Componente Principal 2 $\frac{\lambda_2}{n-1}$

Qual número máximo de componentes principais que podem ser feitos?

- O número máximo de dimensões do problema.

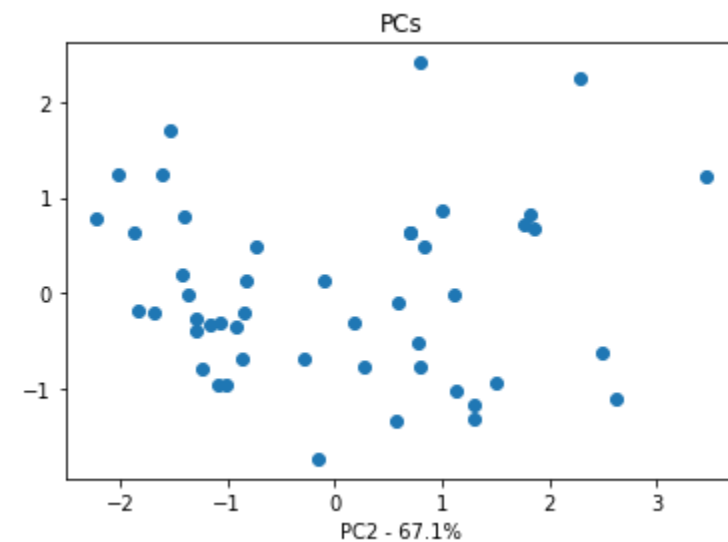
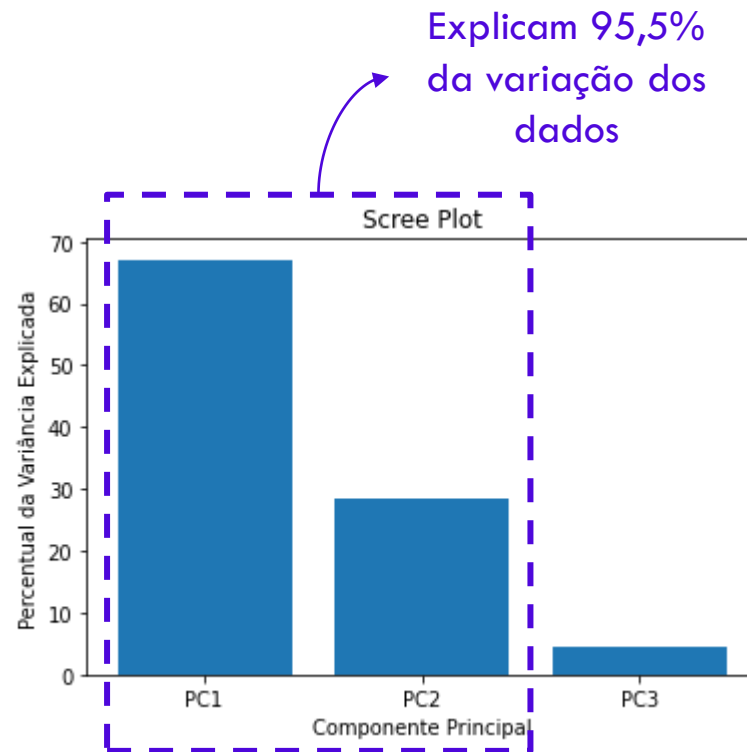
Aplicando PCA no nosso problema

31

Dados originais

area	peri	shape	perm
4990	2791.9	0.09	6.3
7002	3892.6	0.15	6.3
7558	3930.7	0.18	6.3
7352	3869.3	0.12	6.3
7943	3948.5	0.12	17.1
7979	4010.1	0.17	17.1
9333	4345.8	0.19	17.1
8209	4344.8	0.16	17.1

X



Referências Bibliográficas

32

- Jiawei Han, Micheline Kamber, and Jian Pei. 2011. **Data Mining: Concepts and Techniques** (3rd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Steven S. Skiena. 2017. **The Data Science Design Manual** (1st. ed.). Springer Publishing Company, Incorporated.
- Vídeos:
 - ▣ <https://www.youtube.com/watch?v=FgakZw6K1QQ>
 - ▣ <https://www.youtube.com/watch?v=UVHneBUBW0>