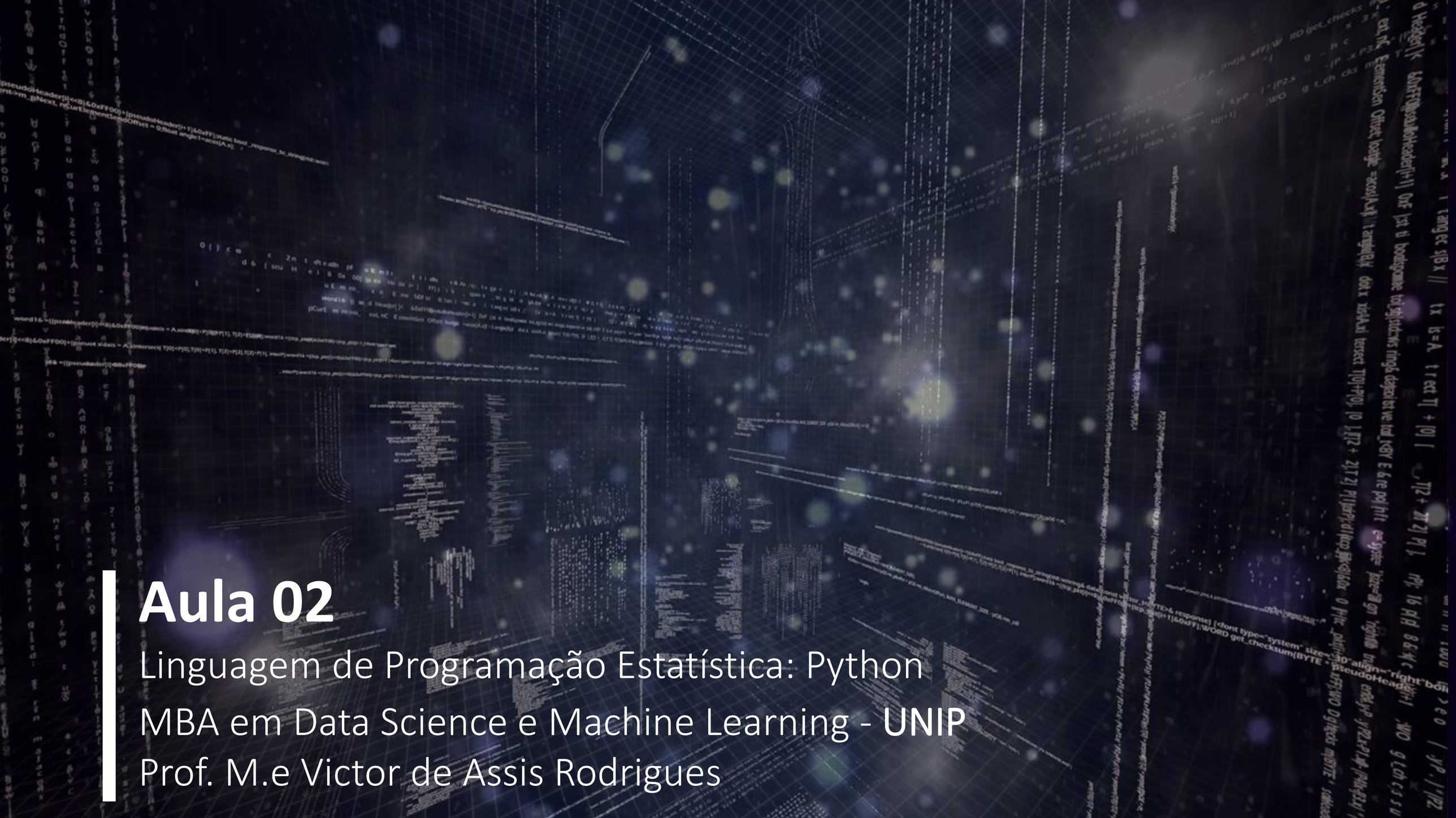


Aula 02

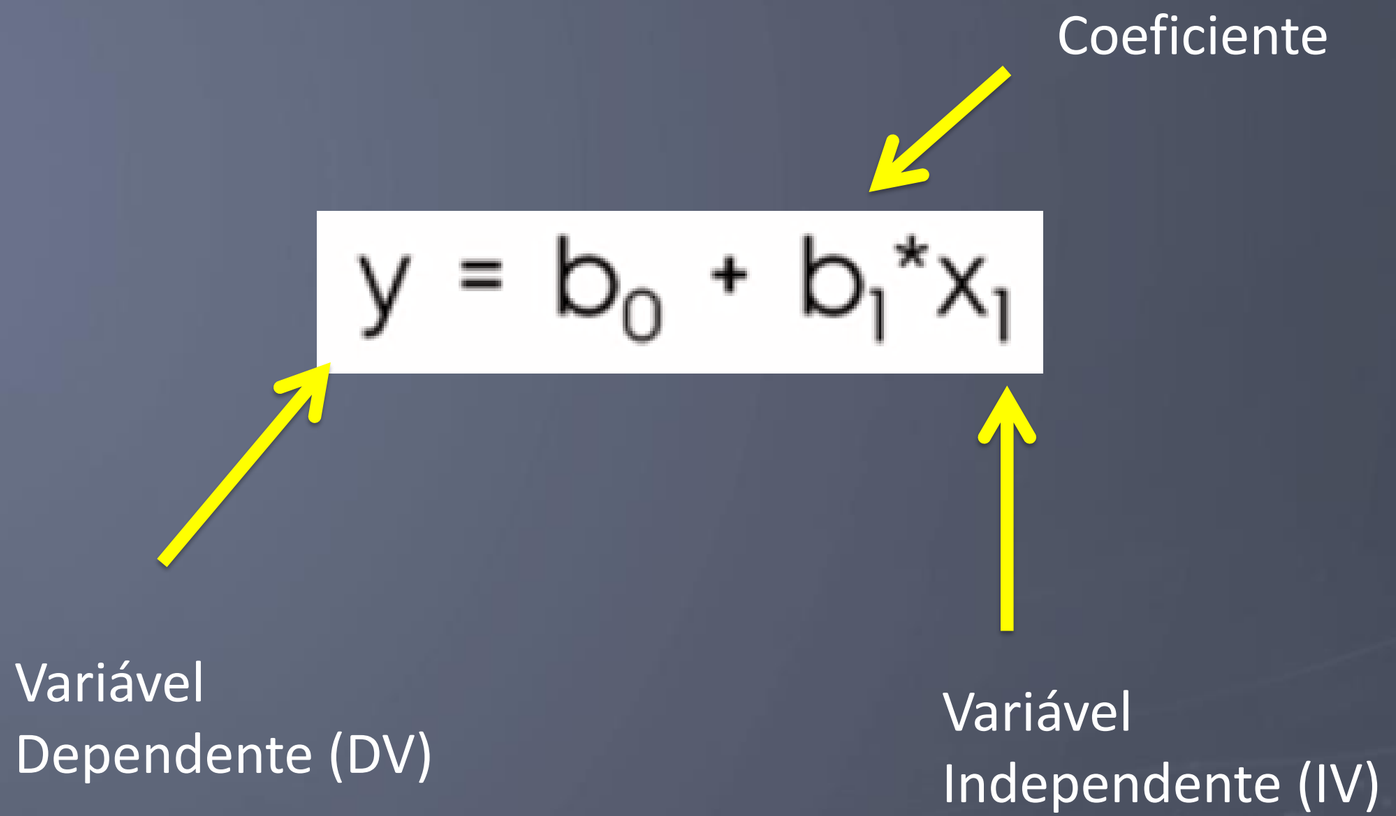
Linguagem de Programação Estatística: Python

MBA em Data Science e Machine Learning - **UNIP**

Prof. M.e Victor de Assis Rodrigues

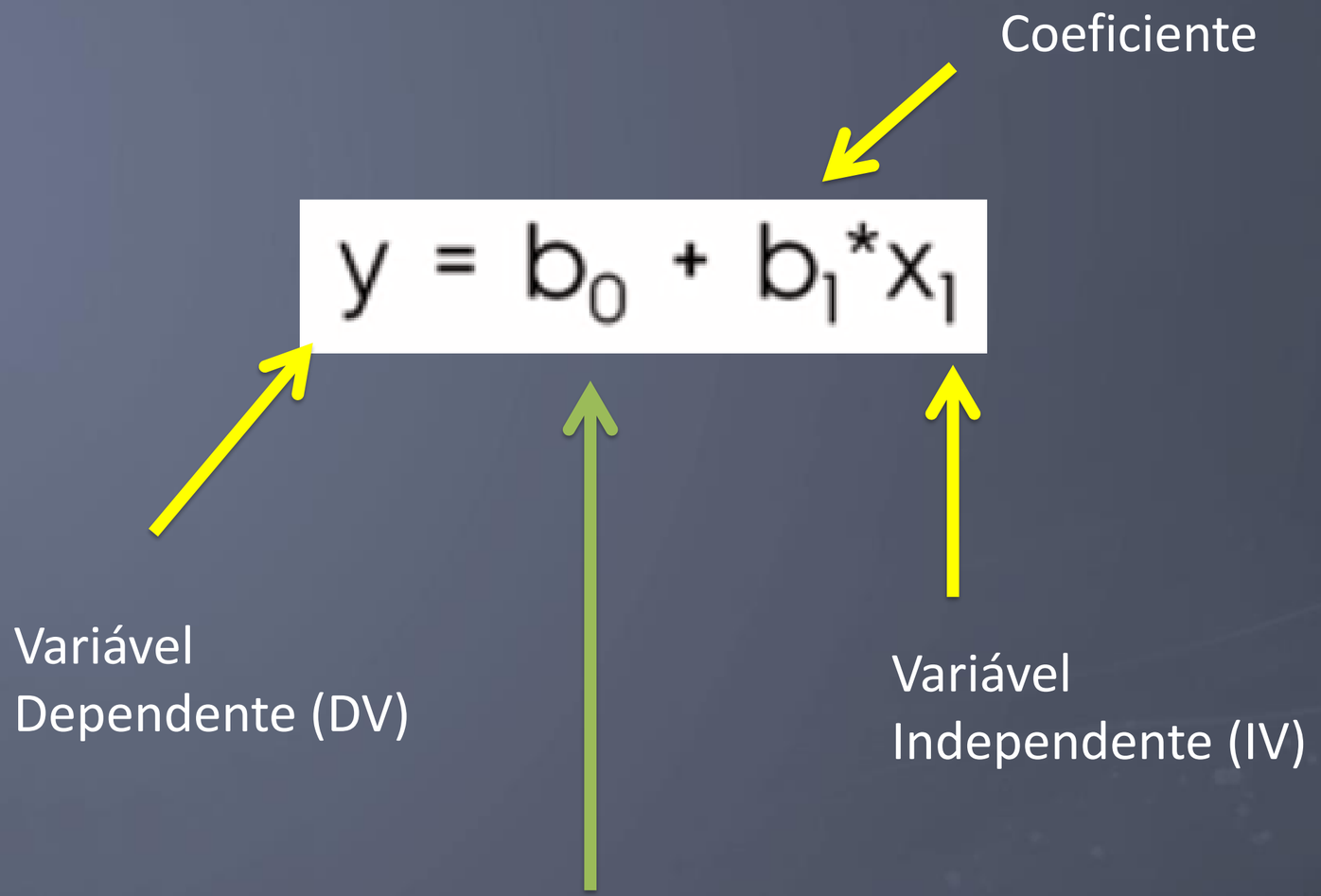
Regressão Linear Simples

Relembrando:



Regressão Linear Simples

Relembrando:



Regressão Linear Simples

Relembrando:

Coeficiente

$$y = b_0 + b_1 x_1$$

Variável Independente (IV)

Define o valor quando x = 0

Regressão Linear Múltipla

Regressão Linear Simples:

$$y = b_0 + b_1 * x_1$$

Regressão Linear Múltipla:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Regressão Linear Múltipla

Regressão Linear Múltipla:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

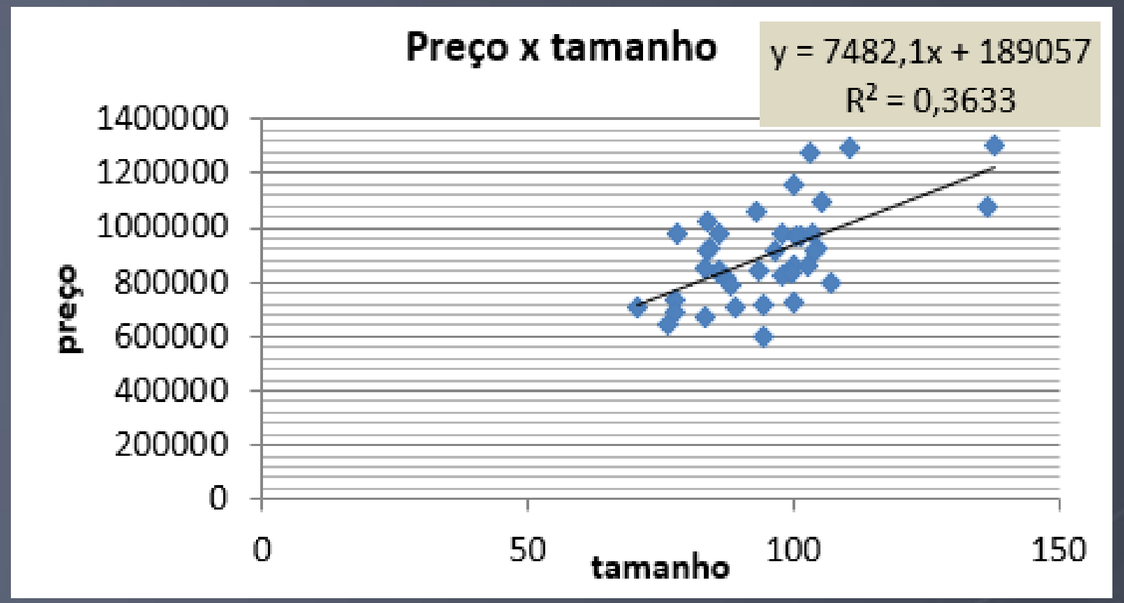
Apto	preço	tamanho	idade do prédio	andar	número de quartos	vagas garagem	piscina?	bairro
1	814364	87	9	9	2	2	0	A
2	837887	86	10	1	2	2	1	A
3	1094109	105	8	12	4	2	1	C
4	727129	100	11	7	3	1	0	A
5	784800	88	8	13	2	1	0	B
6	1158339	100	9	8	3	2	1	C
7	1080046	136	9	6	4	1	1	A
8	839743	86	10	8	2	2	0	B
9	920737	84	11	9	2	2	0	C
10	713176	94	6	6	3	1	1	A

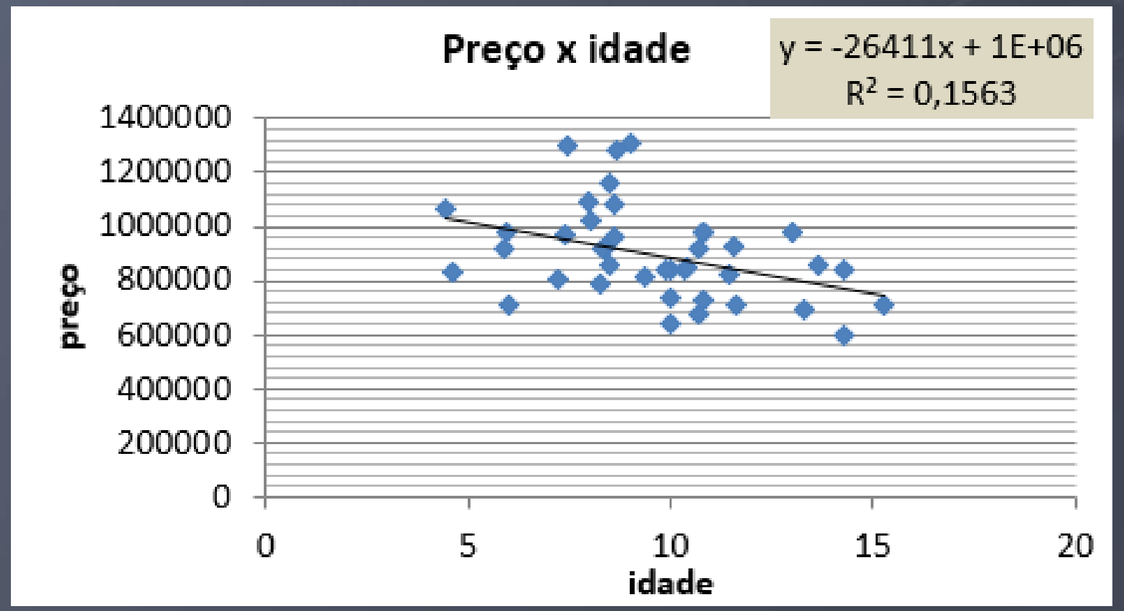
Fonte: https://www.statmeup.com.br/#exercicios?eid=372&p=2

Regressão Linear Múltipla

Regressão Linear Múltipla:

Apto	preço	tamanho	idade do prédio
1	814364	87	9
2	837887	86	10
3	1094109	105	8
4	727129	100	11
5	784800	88	8
6	1158339	100	9
7	1080046	136	9
8	839743	86	10
9	920737	84	11
10	713176	94	6

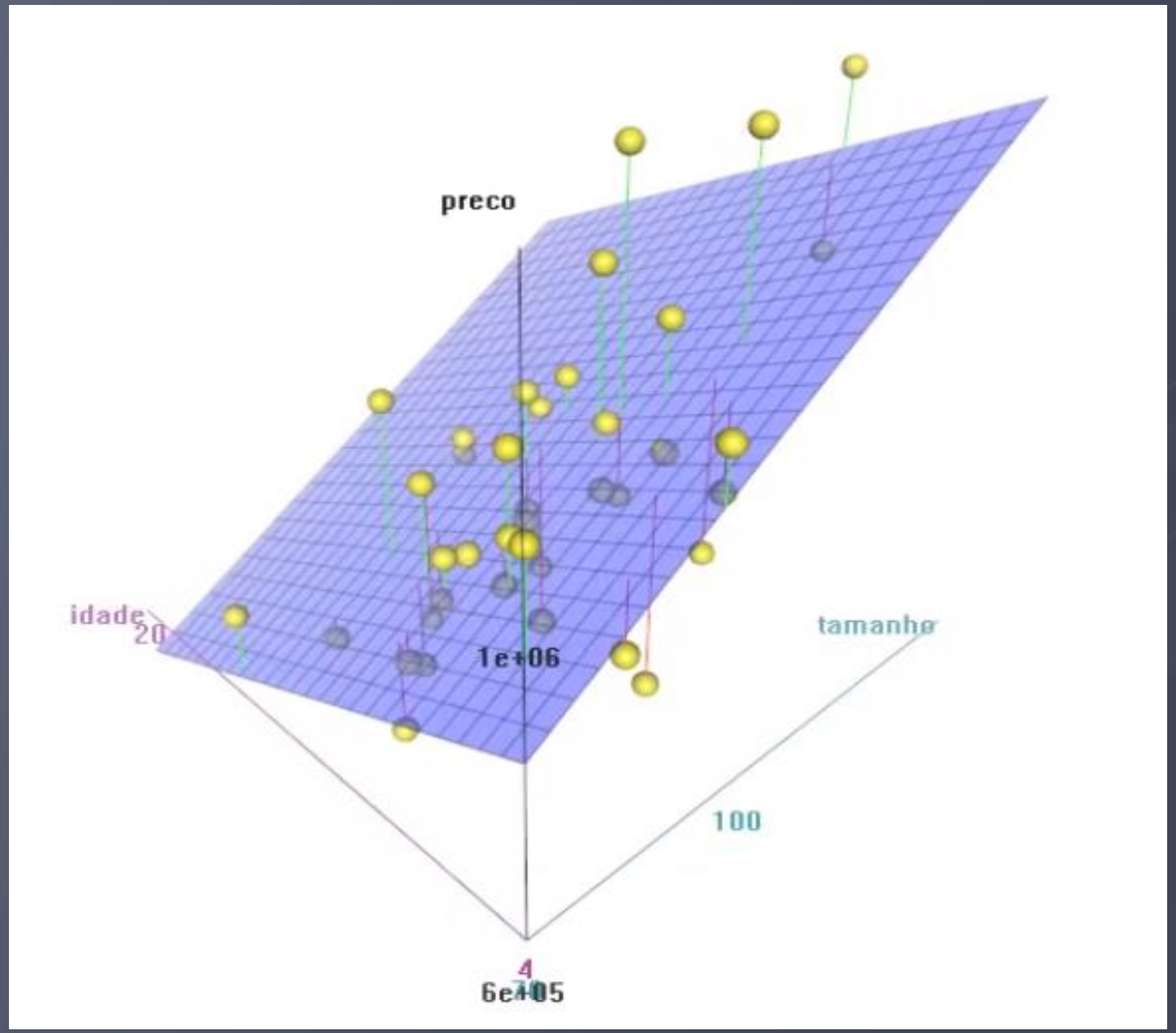




Regressão Linear Múltipla – Visualização 3D

Regressão Linear Múltipla:

Apto	preço	tamanho	idade do prédio
1	814364	87	9
2	837887	86	10
3	1094109	105	8
4	727129	100	11
5	784800	88	8
6	1158339	100	9
7	1080046	136	9
8	839743	86	10
9	920737	84	11
10	713176	94	6



Fonte: https://www.youtube.com/watch?v=TLlzToeIpGc

Regressão Linear Múltipla

Tratamento de dados categóricos

Apto	preço	tamanho	idade do prédio	andar	número de quartos	vagas garagem	piscina?	bairro
1	814364	87	9	9	2	2	0	A
2	837887	86	10	1	2	2	1	A
3	1094109	105	8	12	4	2	1	C
4	727129	100	11	7	3	1	0	A
5	784800	88	8	13	2	1	0	B
6	1158339	100	9	8	3	2	1	C
7	1080046	136	9	6	4	1	1	A
8	839743	86	10	8	2	2	0	B
9	920737	84	11	9	2	2	0	C
10	713176	94	6	6	3	1	1	A

Como tratar esses valores na Regressão Linear???

Regressão Linear Múltipla

Tratamento de dados categóricos – Coluna Bairro

Nesse caso Bairro possui apenas três valores distintos (A, B e C)

Poderíamos codificar cada valor diferente e um número correspondente.

bairro
A
A
C
A
B
C
A
B
C
A

bairro
0
0
2
0
1
2
0
1
2
0

Regressão Linear Múltipla

Tratamento de dados categóricos – Coluna Bairro

PROBLEMA: Se analisarmos a função com uma visão prática, ela está refletindo graus de importância entre os bairros: bairroC > bairroB > bairroA

bairro
0
0
2
0
1
2
0
1
2
0

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Regressão Linear Múltipla

“Dummy” Values

SOLUÇÃO: Separar cada valor em uma nova coluna.

Os valores criados por essa solução chamamos de valores “Dummy” (*falsos*)

bairro
0
0
2
0
1
2
0
1
2
0

Bairro A	Bairro B
1	0
1	0
0	0
1	0
0	1
0	0
1	0
0	1
0	0
1	0

Regressão Linear Múltipla

“Dummy” Values

Pergunta: Porque criamos apenas duas colunas para representar 3 valores?

bairro
0
0
2
0
1
2
0
1
2
0

Bairro A	Bairro B
1	0
1	0
0	0
1	0
0	1
0	0
1	0
0	1
0	0
1	0

Regressão Linear Múltipla

“Dummy” Values

Pergunta: Porque criamos apenas duas colunas para representar 3 valores?

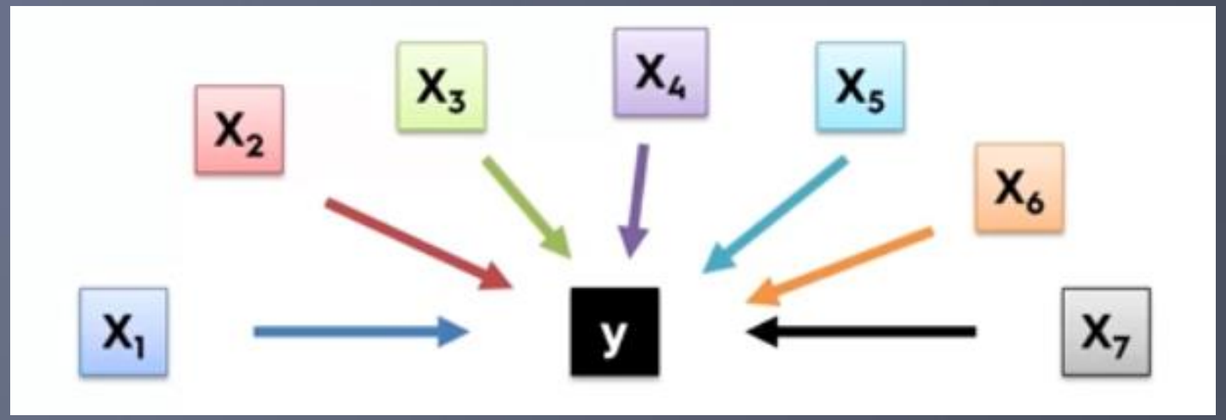
Resposta: Pois com 2 colunas já representamos todos os valores possíveis *(teoria dos conjuntos)*

bairro
0
0
2
0
1
2
0
1
2
0

Bairro A	Bairro B
1	0
1	0
0	0
1	0
0	1
0	0
1	0
0	1
0	0
1	0

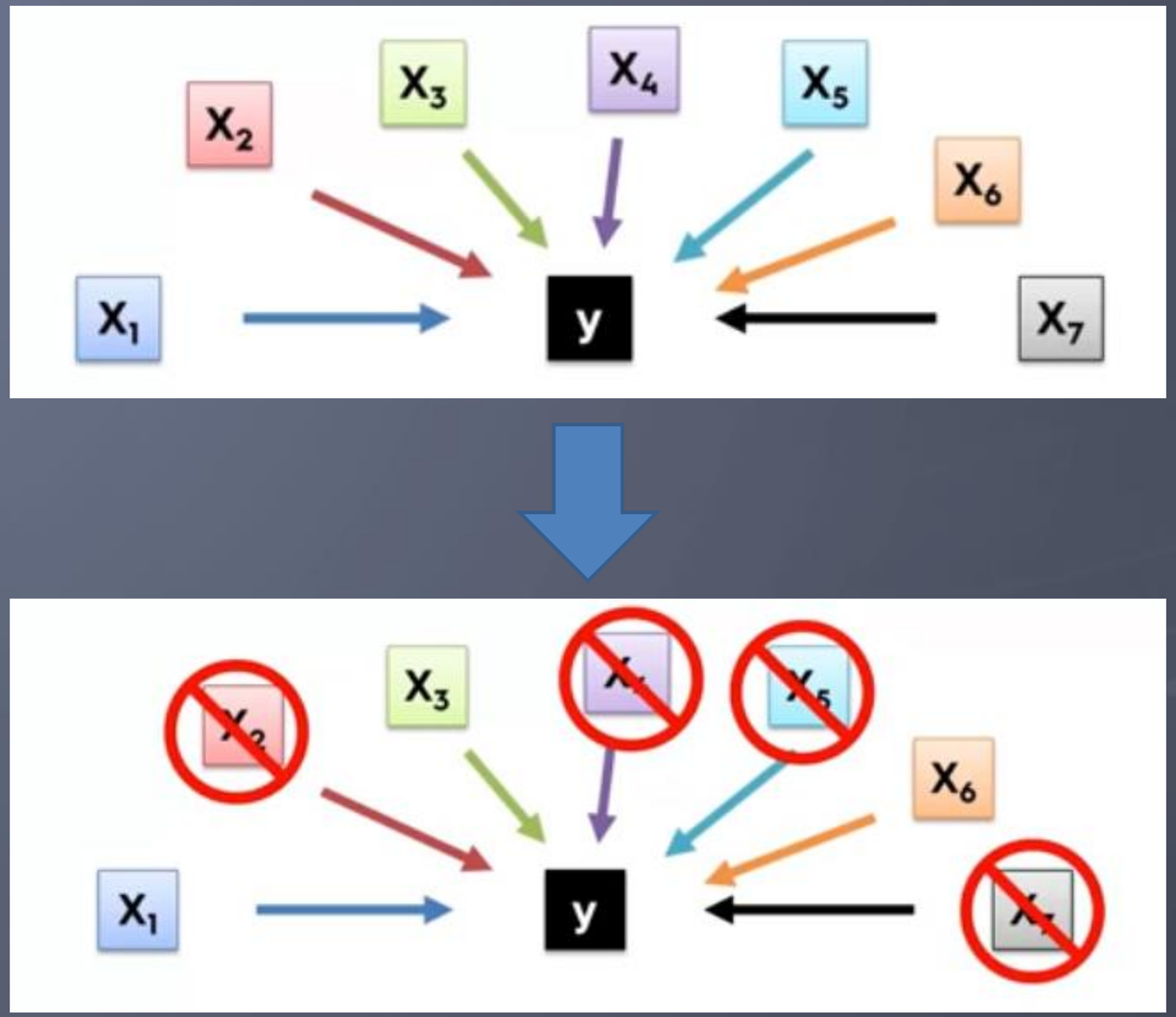
Regressão Linear Múltipla

Como construir modelos de Regressão Linear Múltipla?



Regressão Linear Múltipla

Como construir modelos de Regressão Linear Múltipla?



Regressão Linear Múltipla

Como construir modelos de Regressão Linear Múltipla?

5 Métodos:

1. All-in
2. *Backward Elimination*
3. *Forward Elimination*
4. *Bidirectional Elimination*
5. *Score Comparison*

Stepwise Elimination

Regressão Linear Múltipla

1. All-in

Usar todas as variáveis no modelo

Usamos esse método para:

Obtermos conhecimento prévio do problema;

Compararmos com algum outro modelo;

Passo inicial para construção utilizando Backward Elimination.

Regressão Linear Múltipla

2. Backward Elimination

PASSO 1. Escolhermos o nível de significância para uma coluna (*predictor*) permanecer

(ex. p-value <= 0.05)

PASSO 2. Criar os modelos com todos os *predictors*

L O O P

PASSO 3. Selecionar o *predictor* com o maior p-value, desde que seja > 0.05

PASSO 4. Eliminar *predictor* selecionado.

PASSO 5. Criar um novo modelo sem essa variável.

Regressão Linear Múltipla

3. Forward Elimination

PASSO 1. Escolhermos o nível de significância para uma coluna (*predictor*) entrar no modelo

(ex. p-value <= 0.05)

PASSO 2. Gerar todos os modelos de Regressão Linear Simples. Selecionar a que tenha o menor p-value.

LOOP

PASSO 3. Manter essa variável no modelo e gerar todos os possíveis modelos com essa(s) variável(is).

PASSO 4. Selecionar o modelo que tenha menor valor, desde que p-value <= 0.05

Regressão Linear Múltipla

4. Bidirecional Elimination (Eliminação Gradual)

PASSO 1. Escolhermos o nível de significância para uma coluna (*predictor*) entrar no modelo e ficar no modelo (ex. p-valueEnter = p-valueStay = 0.05)

L O O P

PASSO 2. Executar a seleção do método Forward (novas variáveis no modelo).

PASSO 3. Executar a eliminação do método Backward

PASSO 4. Nenhuma nova variável pode entrar e nenhum variável antiga pode sair

Regressão Linear Múltipla

4. Score comparison

PASSO 1. Selecionar um bom critério de construção (ex. Akaike)

PASSO 2. Construir todos os possíveis modelos de regressão. Combinações: 2^n-1

PASSO 3. Selecionar o modelo com melhor critério

OBS: Se tivermos 10 colunas, precisamos construir 1.023 modelos

Exemplo: Regressão Linear Múltipla

DEMO - Jupyter Notebook

Análise Exploratória de Dados

Não é possível “PULAR” essa etapa.

“A qualidade de seus outputs depende da qualidade de seus “inputs”.

Tudo começa da definição do problema a ser resolvido.

Possui algumas etapas:

- Identificação de variáveis
- Tratamento de Valores Missing
- Tratamento de Outliers
- Transformações de Variáveis
- Criação de Variáveis

Identificação de variáveis

Definir as variáveis preditoras (inputs) e variáveis target (outputs).

Definir os tipos de dados e/ou categorias de cada variável.

Tratamento de valores Missing

Valores que estejam faltando em determinada(s) variável(is).

Pode ter ocorrido algum erro de importação, falha de sistema, gerado por consulta ou pode ser que o valor realmente não exista.

Esse tipo de valor “tira o poder” de um modelo preditivo.

SOLUÇÕES:

- Remover linhas (ou pares de linhas);
- Preencher com um valor que não altera a ideia do projeto.

Tratamento de Outliers

“Outliers” são valores extremos.

Existem dois tipos: univariado e multivariado.

Normalmente são causados por entrada errada de dados, erros de experimentos ou intencional.

SOLUÇÕES:

- Remover linhas (ou pares de linhas);
- Tratá-los separadamente;
- Transformá-los.

Transformações de Variáveis

Refere-se a Engenharia de Dados

Uma única variável pode ter valores Missing, Outliers, valores incorretos, etc.

SOLUÇÕES:

Tratar Data Missing e Outliers;

Alterar distribuição das variáveis;

Mudar simetria;

Criação de novas variáveis unindo outras, ou até mesmo separando uma mesma variável.

Criação de variáveis

Criação de variáveis ajudam a gerar relacionamentos e definir melhor a análise.

Transformando variáveis existentes em novas variáveis ou coletando de outros locais.

Exemplo: Análise Exploratória

DEMO - Jupyter Notebook

Análise de Correlação

Correlação é a forma que as variáveis x e y se relacionam.

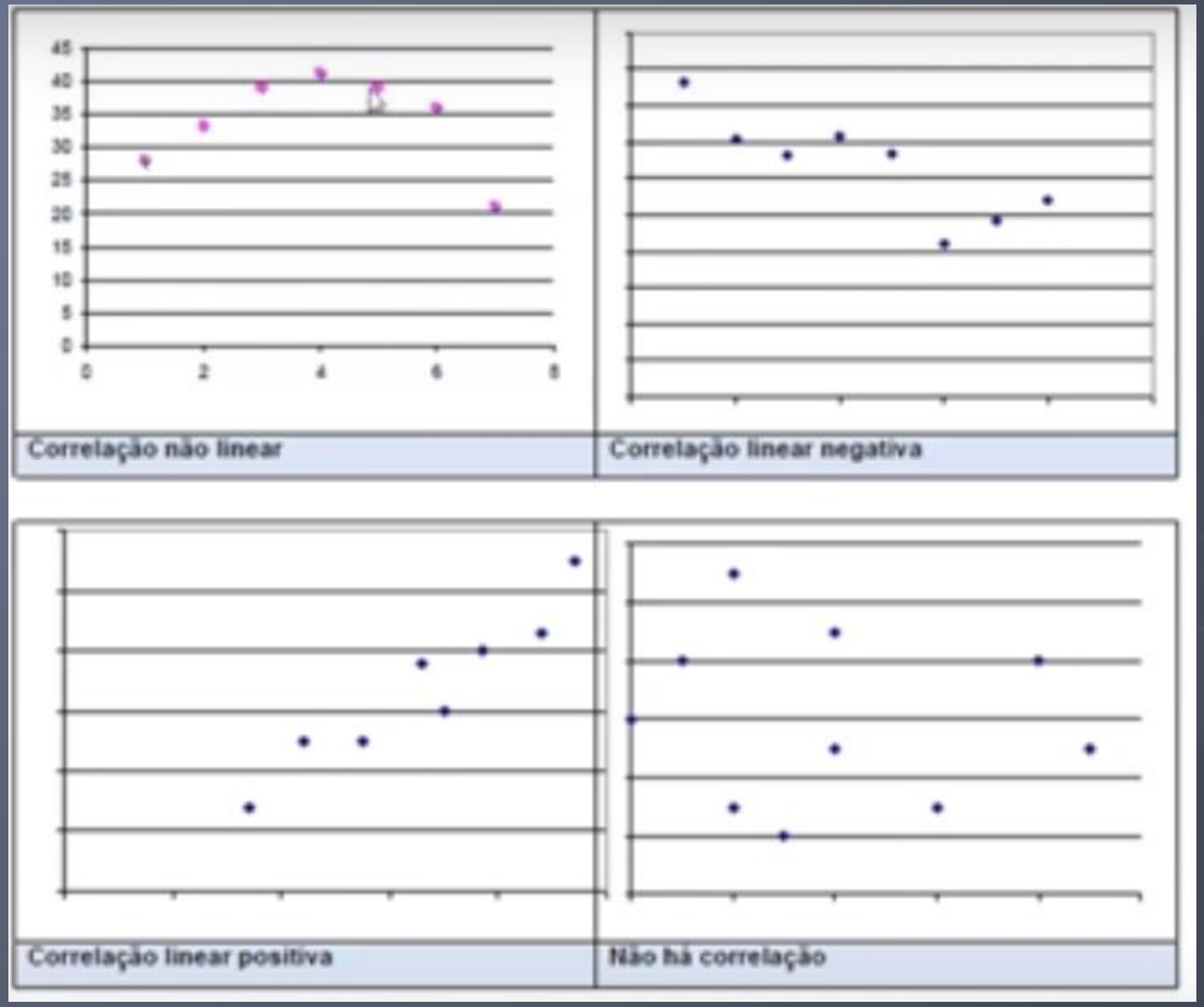
Podemos determinar se é adequado utilizar um modelo linear para determinado fenômeno.

Pode ser:

- Positiva: quando as variáveis são diretamente proporcionais;
- Negativa: quando as variáveis são inversamente proporcionais.

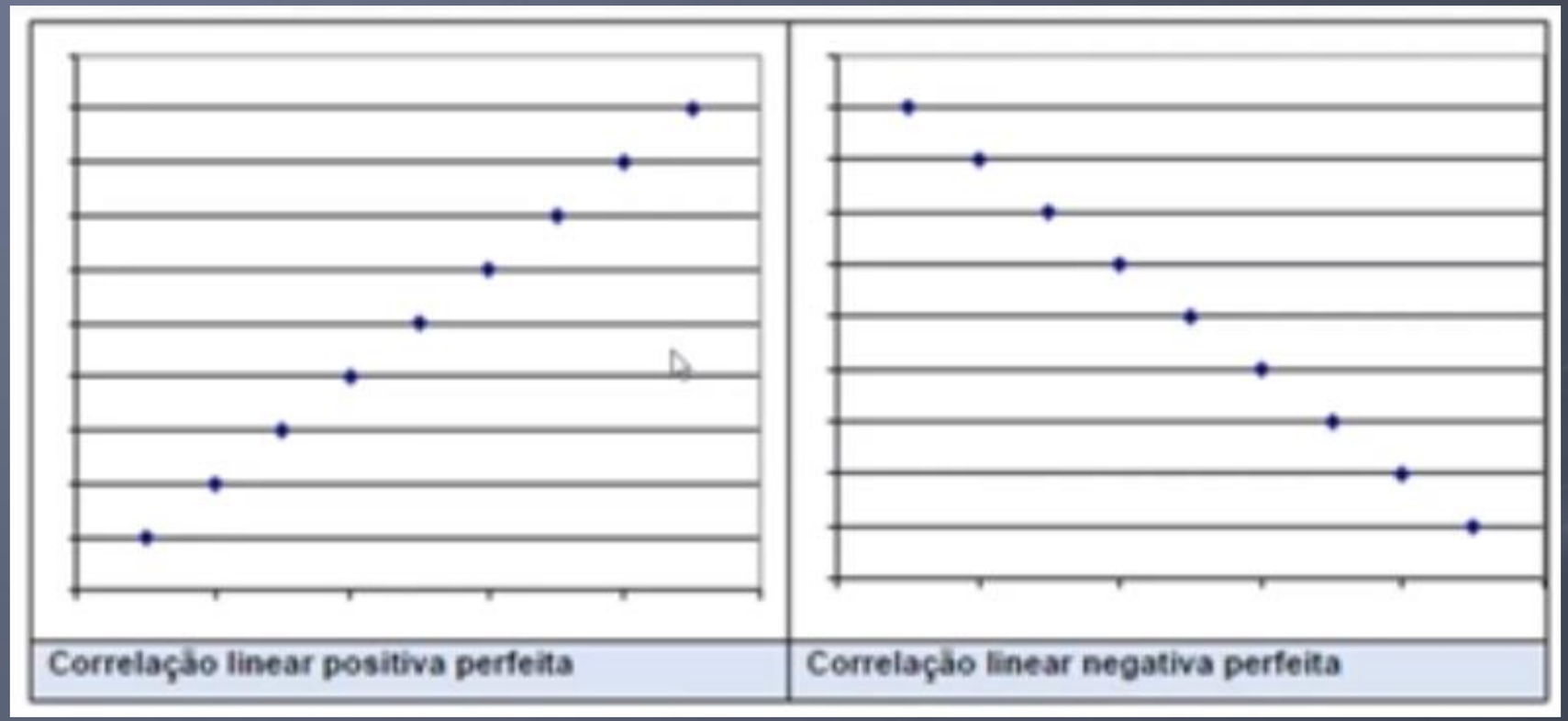
Análise de Correlação

Fonte: https://goo.gl/tLYmMd



Análise de Correlação

Fonte: https://goo.gl/tLYmMd



Análise de Correlação - Pearson

A técnica mais utilizada é a correlação linear de Pearson.

Chamado de *r*, o coeficiente linear de Pearson é calculado pela formula.

$$r = \frac{n \cdot \sum x \cdot y - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{\left[n \cdot \sum x^2 - \left(\sum x\right)^2\right] \cdot \left[n \cdot \sum y^{\beta} - \left(\sum y\right)^2\right]}}$$

Análise de Correlação - Pearson

Coeficiente	Classificação
$0,9 < \|r\| \leq 1,0$	Ótima
$0,8 < \|r\| \leq 0,9$	Boa
$0,7 < \|r\| \leq 0,8$	Razoável
$0,6 < \|r\| \leq 0,7$	Medíocre
$0,5 < \|r\| \leq 0,6$	Péssima
$\|r\| \leq 0,5$	Imprópria

Análise de Correlação - Pearson

Preço (*x*)	36	43	49	55	61	63	69	72	74	77
Demanda (*y*)	350	330	296	252	230	218	203	196	188	167

$$r = \frac{n \cdot \sum x \cdot y - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{\left[n \cdot \sum x^2 - \left(\sum x\right)^2\right] \cdot \left[n \cdot \sum y^2 - \left(\sum y\right)^2\right]}}$$

$$r = \frac{10 \times 137808 - 599 \times 2430}{\sqrt{\left[10 \times 37611 - 599^2\right] \times \left[10 \times 625802 - 2430^2\right]}}$$

$$r = -0{,}9912$$

Preço (*x*)	Demanda (*y*)	x^2	y^2	x . y
36	350	1296	122500	12600
43	330	1849	108900	14190
49	296	2401	87616	14504
55	252	3025	63504	13860
61	230	3721	52900	14030
63	218	3969	47524	13734
69	203	4761	41209	14007
72	196	5184	38416	14112
74	188	5476	35344	13912
77	167	5929	27889	12859
Σ = 599	Σ = 2430	Σ = 37611	Σ = 625802	Σ = 137808

Coeficiente	Classificação
$0{,}9 < \lvert r \rvert \leq 1{,}0$	Ótima

Exemplo: Análise de Correlação

DEMO - Jupyter Notebook

Referências

Livro: Python for Data Analysis

Livro: Python Data Science Handbook

https://www.datasciencecentral.com/profiles/blogs/top-20-python-libraries-for-data-science-in-2018

https://becode.com.br/porque-aprender-python/

http://mindbending.org/pt/a-historia-do-python

https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages

https://blog.liveedu.tv/top-3-most-popular-programming-languages-2018/