



SHIFT

# PEOPLE **ANALYTICS**

## A TRANSFORMAÇÃO DO RH





## ADELAIDE DE OLIVEIRA PROFESSORA

- Mestre em Ciências (FSP/USP), graduada em Estatística (Unicamp).
- Diretora Técnica Estatística da empresa SD&W [www.sdw.com.br](http://www.sdw.com.br)
- Professora de DataMining e Análise Preditiva na FIAP dos cursos MBA: Big Data (Data Science), MBA Business Intelligence & Analytics, MBA Digital Data Marketing, IA & ML e Shift em People Analytics.

 profadelaide.alves@fiap.com.br



## **REGINA CLAUDIA CANTELE**

### PROFESSORA

- Doutora e mestra em Engenharia Elétrica (Poli/USP), graduada em Ciência da Computação e em Ciências Exatas (Universidade Caxias do Sul).
- Coordenadora dos cursos MBA em Engenharia de Dados e MBA em Digital Data Marketing da Fiap.
- Evangelista Big Data e Analytics.



[regina.cantele@fiap.com.br](mailto:regina.cantele@fiap.com.br)

DATA ANALYTICS

# ANÁLISE EXPLORATÓRIA DOS DADOS

INTRODUÇÃO (ANALYTICS / INSIGHTS)

# INTRODUÇÃO

#PEOPLE **INSIGHTS**

O INSIGHT representa a capacidade de compreender claramente a natureza interna das coisas, que surge quando se reconhece relações ou faz novas associações de algo que ainda não é óbvio, mas ao mesmo tempo reconhecível e real, e que fornece as bases para a construção de estratégias de negócios que sustentam uma real vantagem competitiva.

# INTRODUÇÃO

“ O Índice de Rotatividade(Turnover)  
da nossa empresa está em 17% ”

“O tempo médio de permanência da  
nossa empresa é de 2,5 anos”

DADOS OU INFORMAÇÕES?



# INTRODUÇÃO



## Conhecimento

Informação Estruturada  
e Contextualizada

## Informação

Dado com significado

## Dado

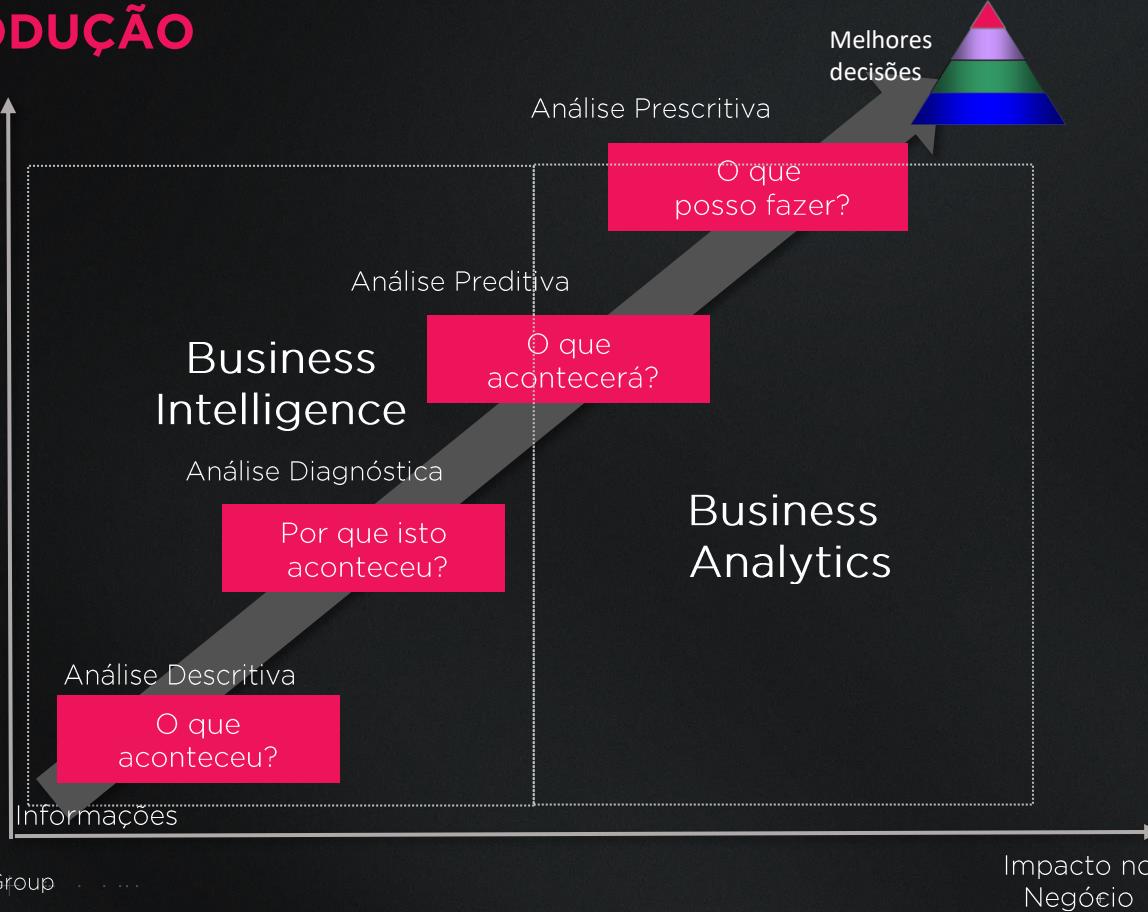
Valor sem significado



Ajudar o gestor a diminuir os riscos e aumentar as chances de sucesso!

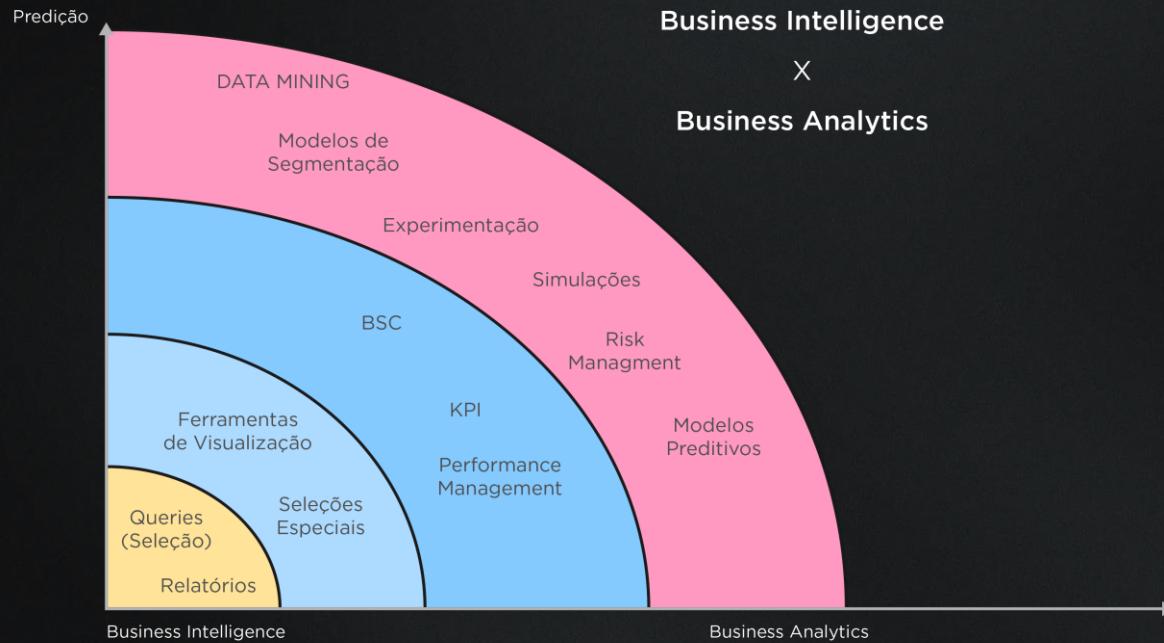
# INTRODUÇÃO

- Habilidades



Fonte: Gartner Group

# INTRODUÇÃO



# INTRODUÇÃO

... enfim, seus dados não servem para nada até que você saiba como tirar informações deles.

## DESCRITIVO

O que aconteceu?

Quantos colaborador(es) temos?  
Quantos são mulheres ou homens?  
Onde residem e qual a distância da empresa?  
Qual o tempo de casa de cada funcionário?

## DIAGNÓSTICO

Por que isto aconteceu?

Qual a relação entre o desligamento voluntário x sexo x tempo de deslocamento médio x distância de casa?

## PREDITIVO

O que acontecerá?

Quanto(a)s colaborador(es) precisaremos contratar nos próximos três anos considerando o perfil da população e comportamento do turnover?  
Qual a probabilidade de turnover em um determinado grupo?

## PRESCRITIVO

O que posso fazer?

Lista de ações para recrutar colaboradores por cargo nos canais A, B e C.  
Quem queremos reter? E, demitir???

# INTRODUÇÃO

## COMPONENTES

### Recursos Humanos/Financeiro

- Decisão estratégica
- Implantação da ação

### Tecnologia

- Infra estrutura
- Ferramentas
- Gerenciamento das informações

### Dados

- Informações das pessoas

### Estatística

- Análises que suportam a tomada de decisões

# INTRODUÇÃO

## Estatística

É a ciência que trata dados numéricos provenientes de mensuração em grupos de indivíduos.

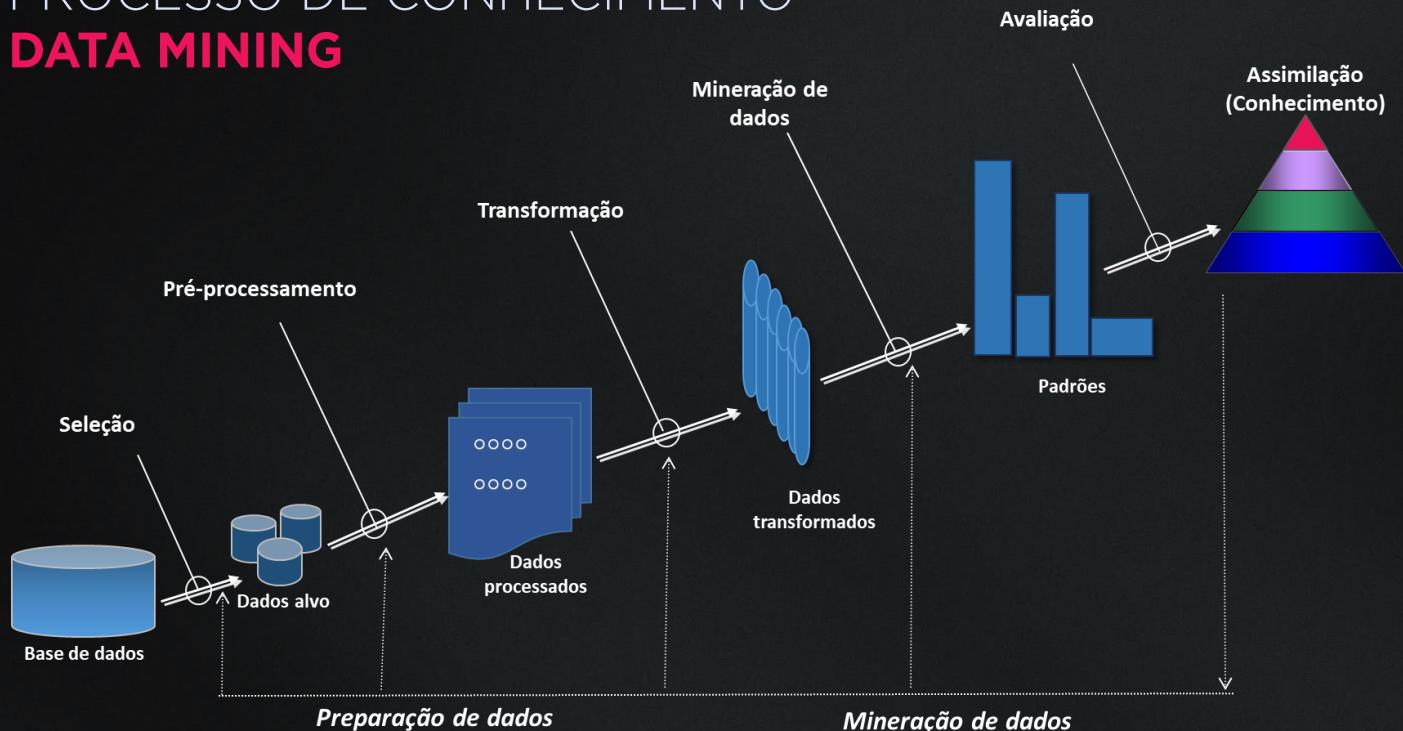
Trata da organização, descrição, apresentação, análise e interpretação de dados resultantes da observação de fenômenos coletivos. Produz métodos para inferência estatística.

A Estatística é uma ciência baseada na Teoria da Probabilidade, cujo objetivo principal é nos auxiliar a tomar decisões ou tirar conclusões em situações de incerteza, a partir dos dados.

**“Estatística é a Ciência que permite obter conclusões a partir de dados.”**

*(Paul Velleman)*

# PROCESSO DE CONHECIMENTO DATA MINING



Uma visão geral das etapas que compõe o processo KDD. – Knowledge Discovery in Databaset  
Fonte: Fayyad (1996)

# PASSOS PARA UMA ANÁLISE **DE UM PROBLEMA (MINERAÇÃO)**

- Identificar a questão/problema.
- Definir a base de análise.
- Como acessar. Qual o histórico. Precisa de pesquisa?
- Obter os dados.
- Consistência das informações.
- Análise exploratória dos dados.
- Modelos estatísticos / modelos preditivos / Outros Algoritmos: ML& IA.
- Interpretação dos resultados.
- Avaliação dos resultados.
- Compilação / resumo dos resultados.
- Algoritmo dos modelos.

## ANÁLISE EXPLORATÓRIA DE DADOS

O objetivo da análise exploratória de dados é examinar a estrutura subjacente dos dados e aprender sobre os relacionamentos sistemáticos entre muitas variáveis.

- Organizar e Descrever os Dados: **Estatística Descritiva**
- Analisar e Interpretar os Dados: **Estatística Inferencial**

# ANÁLISE EXPLORATÓRIA DE DADOS

## Tipos de Variáveis

### Qualitativas (Categóricas)

#### Nominal

Valores não apresentam ordenação, que é um atributo ou uma qualidade

- Cor;
- Sexo;
- Estado Civil;
- Vale transporte(Sim ou Não).

#### Ordinal

Valores apresentam uma ordem, uma escala pré-determinada

- Escala de questionário (adjetivos);
- Preferência;
- Faixa etária;
- Classe Social;
- Grau de Escolaridade.

### Quantitativas

#### Discreta (Contagem)

Valores fazem parte de um conjunto finito ou infinito numerável

- Escala de questionário numérica);
- Número de filhos;
- Quantidade de empregados;
- Quantidade de ligações.

#### Contínua (Medição)

Valores são os números reais

- Salário Mensal;
- Idade;
- Tempo de estudo;
- Taxas.
- Indicadores de RH.

# EXEMPLO DE TIPOS DE VARIÁVEIS

	Id	Sexo	Idade	Cor	Internet	Telefone móvel	Anos estudo	Rendimento
Categóricas	35000015	2	15	2	1	3	7	
	35000015	2	75	2	3	3	12	
	35000031	2	60	2	3	3	12	
	35000058	2	68	2	3	3	1	
Quantitativas	35000058	2	48	8	3	1	5	1.000
	35000058	2	42	2	3	3	6	
	35000066	2	36	2	1	3	9	1.000
	35000066	2	44	2	1	1	9	1.200
N = 1380	35000066	2	20	2	1	3	13	300
	35000066	4	26	2	1	3	12	
	35000074	4	14	2	1	3	8	
	35000074	4	71	2	3	3	5	
Exemplo: Internet: Acesso últimos 3 meses	35000090	4	20	2	1	1	12	
	35000090	2	19	8	1	3	12	620
	35000090	4	42	2	3	1	12	300
	35000090	4	17	2	1	1	11	
	35000090	4	25	2	1	1	12	433
	35000090	2	49	6	1	3	16	400
	35000104	2	38	2	3	1	6	600

Exemplo: Internet:  
Acesso últimos 3 meses

1 → 1  
3 → 0

# TRANSFORMANDO VARIÁVEIS **QUANTITATIVAS EM QUALITATIVAS**

Variável Quantitativa → **Critério** → Variável Qualitativa

	<b>Critério</b>	<b>Grau Instrução</b>
Anos de Estudo	→ 0	→ Analfabeto
	→ [1 - 9]	→ Fundamental
	→ [10 - 12]	→ Médio
	→ $> = 13$	→ Superior

## TRANSFORMANDO VARIÁVEIS QUANTITATIVAS EM QUALITATIVAS

**Exemplo:** Quantas classes serão necessárias para representar a idade?

Idade (anos) VARIÁVEL QUANTITATIVA CONTÍNUA

0 |-----| 90

Decidimos, antes, que desejávamos dividir a amplitude total em cinco segmentos de reta com amplitudes iguais.

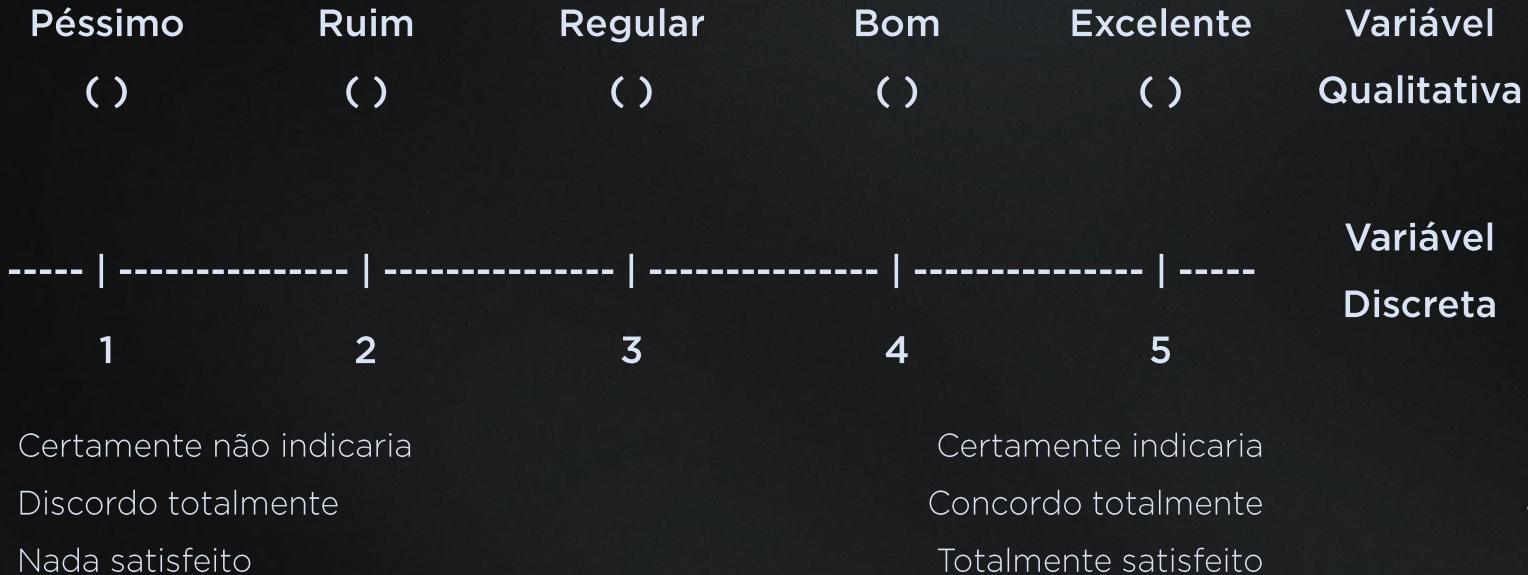
A amplitude de cada intervalo abaixo representa, portanto,

- amplitude =  $(90 - 0)/5 = 18$  anos.

0 |-----| 18 |-----| 36 |-----| 54 |-----| 72 |-----| 90

Faixa Etária	F
[0 - 18]	437
[19 - 36]	384
[37 - 54]	360
[55 - 72]	158
[73 - 90]	41
Total	1.380

## TIPOS DE VARIÁVEIS: **ESCALA DE QUESTIONÁRIO**

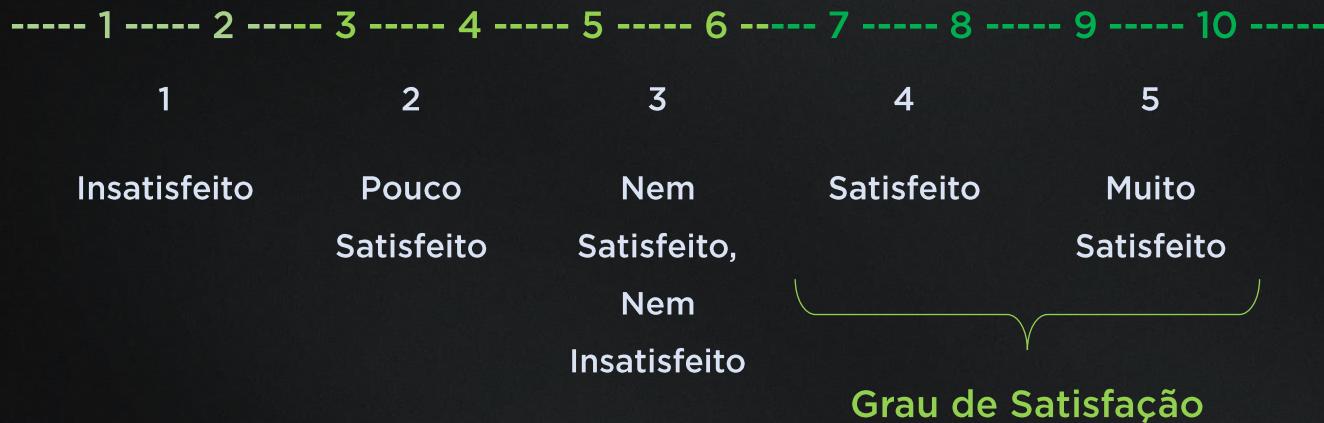


## TIPOS DE VARIÁVEIS: **ESCALA DE QUESTIONÁRIO**

- **Classificação:** “satisfeito” , “não satisfeito”
- **Grau de satisfação:** escala de 1 a 5 associada a adjetivos
- **Grau de satisfação:** escala de 0 a 10
- **Grau de satisfação:** escala construída com vários itens de um questionário

## TIPOS DE VARIÁVEIS: **ESCALA DE QUESTIONÁRIO**

Escala de Satisfação



## TRANSFORMANDO MEDIÇÕES **POR MEIO DE ESCALAS EM INDICADORES**

- Porcentagens de funcionários muito insatisfeitos, insatisfeitos, neutros, satisfeitos e muito satisfeitos em relação à média.
- Porcentagem de funcionários que se dizem dispostos a recomendar a empresa.
- Porcentagem de funcionários que identificam corretamente as intenções da empresa em termos de posicionamento e identificação.
- Percepção média a respeito da qualidade dos produtos da empresa em comparação aos dos principais concorrentes.

## TRANSFORMANDO MEDIÇÕES POR MEIO DE ESCALAS EM INDICADORES

Qual é a probabilidade de você recomendar nossa empresa a um amigo ou colega para aqui trabalhar?

A métrica obtida por essa pergunta é o Employer Net Promoter Score, **E-NPS**.

O NPS é baseado na crença fundamental de que os clientes/colaboradores da empresa podem ser divididos em três categorias:

- Os Promotores são os leais, entusiasmados, que recomendam a empresa para seus amigos, numa escala de 0 a 10, os promotores pontuam 9 ou 10.
- Os Neutros são os satisfeitos, mas pouco entusiasmados e que podem ser seduzidos pelo concorrente, que pontuam 7 ou 8.
- Já os Detratores são os infelizes, que se encontram presos a uma relação ruim. As respostas pontuam de 0 a 6.

# TRANSFORMANDO MEDIÇÕES POR MEIO DE ESCALAS EM INDICADORES

**Qual é a probabilidade de você recomendar esta empresa a um amigo ou colega?**

A melhor maneira de medir a eficiência desse motor de crescimento é obter o percentual de clientes/colaboradores promotores (P) e subtrair desse percentual os detratores (D). Dessa forma, obtemos o NPS: **NPS = %P - %D**

**Por exemplo:** %P=70% %D=20% NPS = 70% - 20% = 50%

- As empresas com motores de crescimento mais eficientes operam a uma taxa de eficiência de NPS de 50% a 80%;
- As empresas médias (grande maioria) ficam estagnadas em um índice NPS de apenas 5% a 10% (os promotores mal superam os detratores);
- Muitas empresas possuem NPS's negativos (a cada dia elas estão criando mais detratores do que promotores).



# APRESENTAÇÃO **DOS DADOS**

## Distribuição de frequênciа

O número de vezes em que ocorreram valores em cada classe ou valores chama-se frequência absoluta. O conjunto das ocorrências, com correspondentes frequências absolutas (FA) e relativas (FR), define a distribuição de frequências da variável.

### Exemplo:

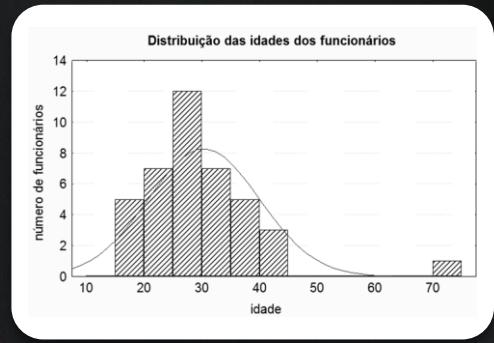
Faixa Etária	FA	FR (%)
[0 - 18]	437	$437/1380*100$
[19 - 36]	384	$384/1380*100$
[37 - 54]	360	$360/1380*100$
[55 - 72]	158	$158/1380*100$
[73 - 90]	41	$41/1380*100$
Total	1380	$1380/1380*100$



Faixa Etária	FA	FR (%)
[0 - 18]	437	31.7
[19 - 36]	384	27.8
[37 - 54]	360	26.1
[55 - 72]	158	11.4
[73 - 90]	41	3.0
Total	1380	100.0

# COMO EXPLORAR BASSES / VARIÁVEIS?

Ordem	idade
1	25
2	24
3	22
4	28
5	18
6	26
7	22
8	24
9	24
10	25
11	26
12	24
13	26
14	47
15	65
16	29
17	29
18	35
19	54
20	44
21	24
22	32
23	33
24	28
25	24
26	24
27	26
28	32

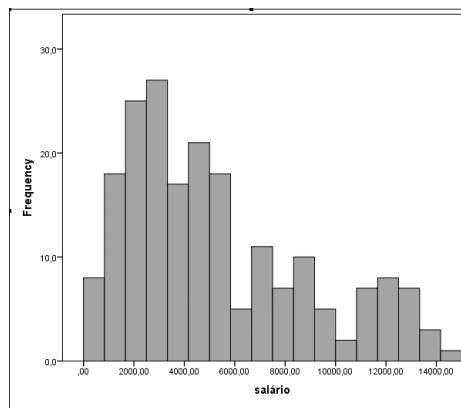
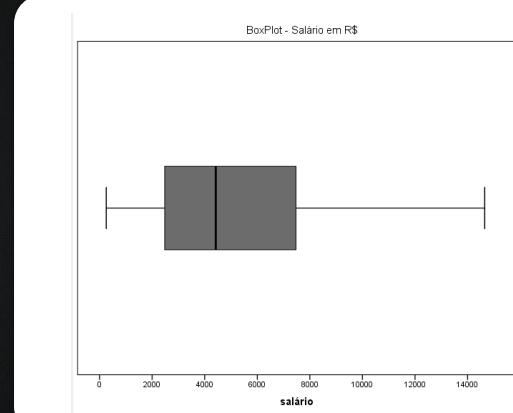


# APRESENTAÇÃO DOS DADOS

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	13.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.856,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Salário	200	254,19	14.649,52	5.259,30	3.615,36
Valid N (listwise)	200				



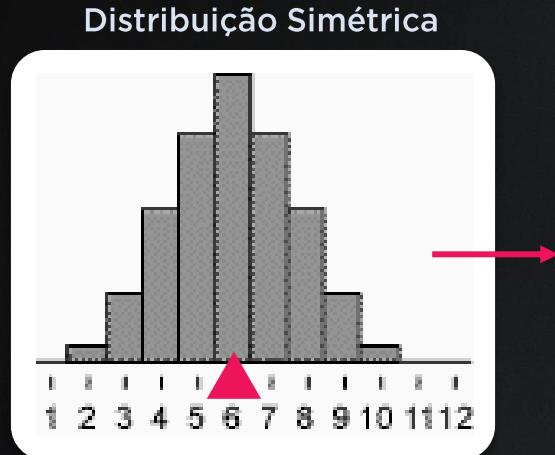
# APRESENTAÇÃO **DOS DADOS**

## **Medidas de Posição e Dispersão**

Medidas que resumam as características peculiares do fato em estudo (distribuições):

- Seu valor central;
- Seu grau de dispersão em torno do valor central (variabilidade);
- Seu grau de assimetria (forma de distribuição).

# APRESENTAÇÃO DOS DADOS



Distribuição do tempo de casa (anos)

## Medidas de tendência central:

Indicam o centro da distribuição de frequências ou a região de maior concentração de frequência na distribuição.



- Média
- Mediana
- Moda

## Medidas de dispersão:

Indicam o grau de homogeneidade dos valores, até que ponto eles se encontram concentrados ou dispersos da média.



- Variância
- Desvio padrão

# APRESENTAÇÃO **DOS DADOS**

## Medidas de Posição - Médias

- Média Aritmética Simples:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- Média Aritmética Ponderada:

$$\bar{x} = \frac{\sum_{i=1}^n x_i \cdot f_i}{n}$$

- Média Geométrica (evolução):

A média geométrica é muito usada no cálculo da taxa média de retorno de um investimento ou no cálculo da taxa equivalente de uma aplicação financeira.

$$Mg = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

- Média Quadrática:

$$\bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

# APRESENTAÇÃO **DOS DADOS**

## Medidas de Posição - Mediana

A mediana é uma quantidade que, como média, também procura caracterizar o centro da distribuição da frequência.

Ela é calculada com base na ordem dos valores que formam o conjunto dos dados.

- Número ímpar de observações - a mediana é definida como sendo igual ao valor de ordem  $(n+1)/2$  desse conjunto. **Ex.:** 2, 4, 11, 50, 18, 17, 26

$$n=7 \rightarrow (7+1)/2$$

Ordenado 2, 4, 11, **17**, 18, 26, 50

- Número par de observações - Para número par, a mediana será a média aritmética dos dois termos centrais do conjunto de dados ordenados. **Ex.:** 1, 3, 7, 10, 18, 20, 26, 35

$$n=8$$

Ordenado 1, 3, 7, **10, 18**, 20, 26, 35

**Mediana: 14**

# APRESENTAÇÃO **DOS DADOS**

## Exemplo:

Durante uma verificação de satisfação de funcionários, foram obtidas as seguintes avaliações:

6,03    5,59    6,40    6,00    5,99    6,02

Qual o valor encontrado da satisfação média e mediana?

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \rightarrow \bar{x} = \frac{6,03 + 5,59 + 6,40 + 6,00 + 5,99 + 6,02}{6} \rightarrow \bar{x} = 6,00$$

**Mediana:**    5,59    5,99    6,00    6,02    6,03    6,40

$$mediana = \frac{6,00 + 6,02}{2} = 6,01$$

# APRESENTAÇÃO **DOS DADOS**

## Exemplo Anterior:

Durante uma verificação de satisfação, foram obtidas as seguintes notas:

6,03 5,59 6,40 6,00 5,99 6,02  
**(6,04)**

Qual a nota média e mediana encontrada?  $\bar{x} = 6,00$  *mediana = 6,01*

Suponha que o terceiro valor tenha sido incorretamente medido e que, na verdade, seja de 6,04.

Determine novamente a nota média e mediana.

- Média aritmética:  $\bar{x} = \frac{6,03 + 5,59 + 6,04 + 6,00 + 5,99 + 6,02}{6} = 5,95$

- Mediana: 5,59 5,99 **6,00** 6,02 6,03 6,04

$$\text{mediana} = \frac{6,00 + 6,02}{2} = 6,01$$

# APRESENTAÇÃO DOS DADOS

## Medidas de Posição - Moda

A moda é o valor que ocorre com a maior frequência dentro de um conjunto de dados.

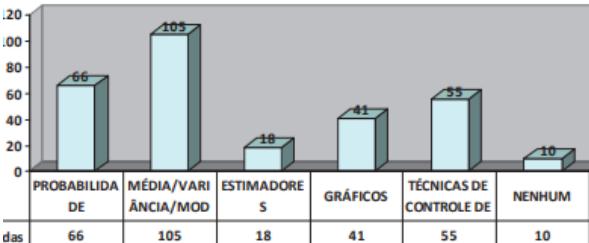
**Ex 1:** 2,2,5,7,**9,9,9**,10,10,11,12,18      = **9**

**Ex 2:** 3,5,8,10,12,15,16      = Amodal

**Ex 3:** 2,3,**4,4,4**,55,555,**7,7,7**,9      = 4 e 7 - Bimodal

A classe modal é representada, numa distribuição de frequências, como a classe com maior frequência.

## Ferramentas Estatísticas mais utilizadas em empresas em uma cidade de MG



Fonte: A IMPORTÂNCIA E O USO DA ESTATÍSTICA NA ÁREA EMPRESARIAL: uma pesquisa de campo com empresas do município de Elói Mendes - MG.

# APRESENTAÇÃO **DOS DADOS**

## Comparação entre Média, Mediana e Moda

	Vantagens	Limitações	Tipo de Variáveis
<b>Média</b>	Reflete todos os valores da amostra	É influenciada por valores extremos	Continua e Discreta
<b>Mediana</b>	Menos sensível a valores extremos que a média	Mais difícil de ser determinada para grandes quantidades de dados	Continua e Discreta
<b>Moda</b>	Representa um valor típico	Não tem função em certos conjuntos de dados	Continua, discreta, nominal e ordinal

# APRESENTAÇÃO **DOS DADOS**

- Medidas baseadas na ordenação dos dados:



# APRESENTAÇÃO **DOS DADOS**

- Medidas baseadas na ordenação dos dados:

**Quartis:**



**Decis:** dividem um conjunto de dados em dez partes iguais.



**Percentis (P1):** divide a série em cem partes, de modo que  $p\%$  ficam abaixo dele (P1).

# APRESENTAÇÃO **DOS DADOS**

## Medidas baseadas na ordenação dos dados:

Algumas aplicações:

- Construir classes ou faixas de variáveis;
- Curva de concentração.

# APRESENTAÇÃO DOS DADOS

## Exemplo:

Idade dos  
funcionários  
da Empresa A

Idade	
26,00	37,00
32,00	30,00
36,00	34,00
20,00	41,00
40,00	26,00
28,00	32,00
41,00	35,00
43,00	46,00
34,00	29,00
23,00	40,00
33,00	34,00
27,00	31,00
37,00	36,00
44,00	43,00
30,00	33,00
38,00	48,00
31,00	42,00
39,00	25,00

Média:

34,5 anos

Idade

Minimum	20,00
Percentile 25	30,00
Median	34,00
Percentile 75	40,00
Maximum	48,00
Minimum	20,00
Percentile 10	26,00
Percentile 20	29,00
Percentile 30	31,00
Percentile 40	33,00
Percentile 50	34,00
Percentile 60	36,00
Percentile 70	39,00
Percentile 80	41,00
Percentile 90	43,00
Maximum	48,00

# APRESENTAÇÃO **DOS DADOS**

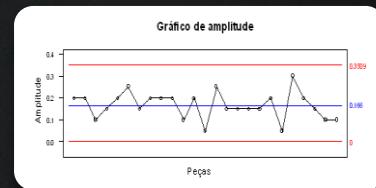
- **Amplitude:** É definida como a diferença entre o maior e o menor valor de um conjunto de dados.

Fortemente relacionado com a dispersão dos dados.

## Exemplo:

Idade	Minimum	20,00	Aplicações:
	Percentile 25	30,00	
	Median	34,00	
	Percentile 75	40,00	
	Maximum	48,00	

A = 28 anos



# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

As medidas de dispersão nos indicam o grau de homogeneidade dos valores, até que ponto eles se encontram concentrados ou dispersos.

**Exemplo:**

**A:** 35, 35, 35, 35, 35, 35

**B:** 32, 33, 34, 36, 37, 38

**C:** 30, 31, 32, 38, 39, 40

$$\bar{X}_A = \bar{X}_B = \bar{X}_C = 35$$

## Medidas de Dispersão

- Amplitude
- Amplitude Inter-Quartílica
- Variância
- Desvio-Padrão
- Coeficiente de Variação

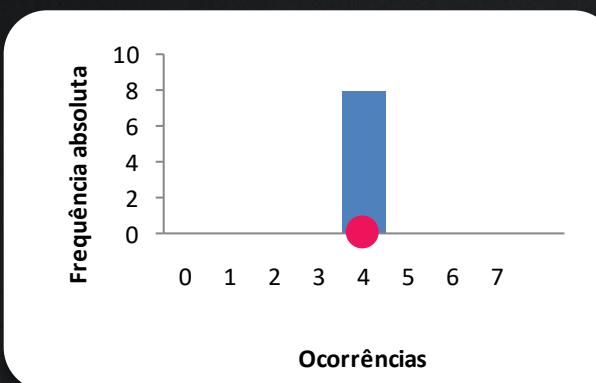
# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

Exemplos:

Qual o desvio padrão?

A: 4, 4, 4, 4, 4, 4, 4 ,4



$$\sigma = 0$$

Média

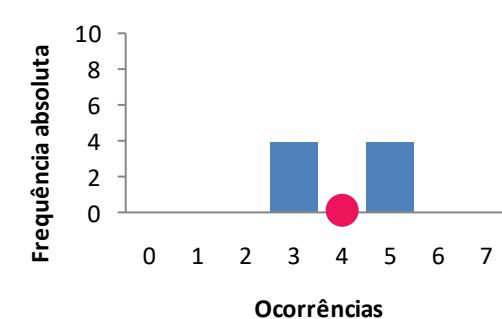
# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

Exemplos:

Qual o desvio padrão?

B: 3, 3, 3, 3, 5, 5, 5, 5



$$\sigma = 1$$

Média

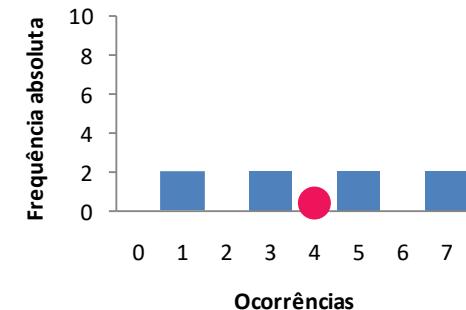
# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

Exemplos:

Qual o desvio padrão?

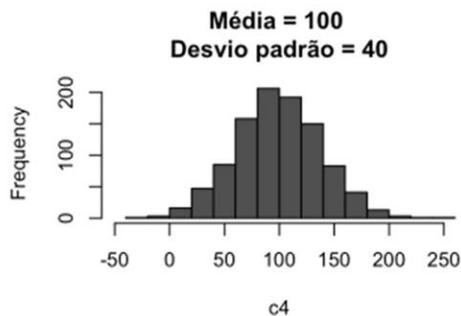
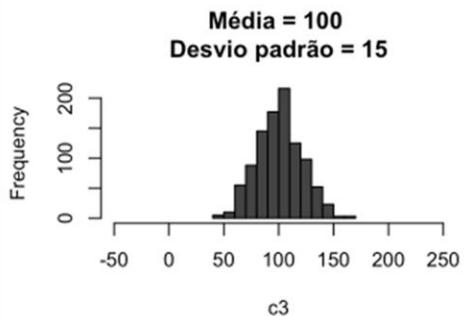
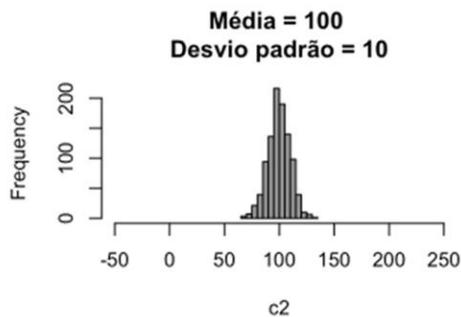
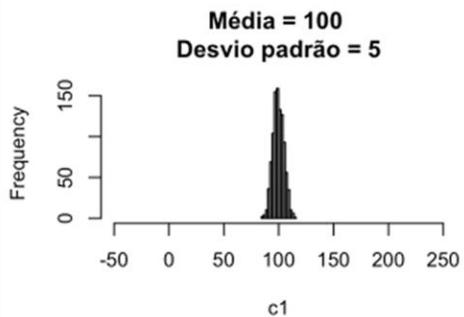
C: 1, 1, 3, 3, 5, 5, 7, 7



$$\sigma = 2.24$$

Média

# APRESENTAÇÃO DOS DADOS



# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

### Variância

O quanto os pontos estão distantes da média (ponto central). Mede, para cada ponto, a distância entre ele e a média e, ao final, obtém o valor médio dessas distâncias.

- Variância da população

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

- Variância amostral

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

### Desvio Padrão

É a raiz quadrada da variância.

Qual a vantagem do Desvio Padrão em relação a Variância?

- O desvio padrão se expressa **na mesma unidade da variável**, sendo, por isso, de maior interesse que a variância nas aplicações práticas. Quanto está distante do ponto central.

# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão

X	Média	X -Média(X)	( X -Média(X)) <sup>2</sup>
31	34,1	-3,1	9,8
32	34,1	-2,1	4,5
33	34,1	-1,1	1,3
33	34,1	-1,1	1,3
35	34,1	0,9	0,8
35	34,1	0,9	0,8
37	34,1	2,9	8,3
37	34,1	2,9	8,3
Soma	-	0,0	34,9

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

Variância:

$$\sigma^2 = \frac{34,9}{8} = 4,36$$

Desvio Padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{4,36} = 2.09$$

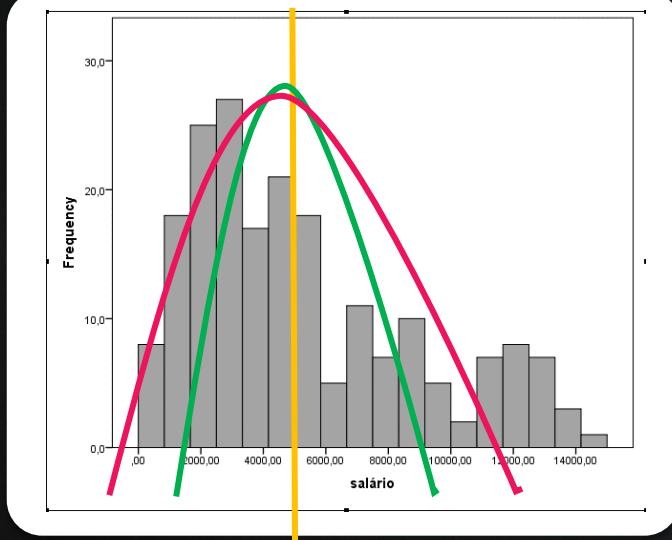
# APRESENTAÇÃO DOS DADOS

## EXEMPLO

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	15.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.858,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.755,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Salário	200	254,19	14.649,52	5.259,30	3.615,36
Valid N (listwise)	200				



Média +/- 1 desvio

$$137/200=68,5\%$$

Média +/- 2 desvio

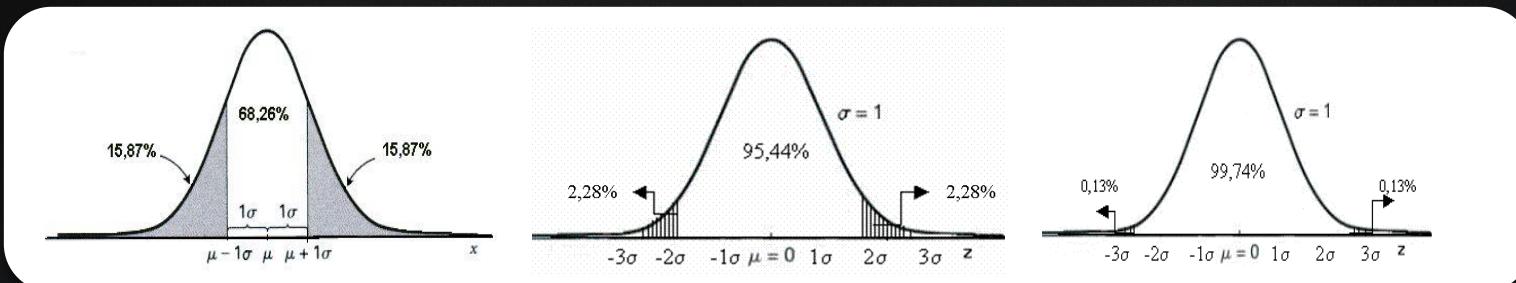
$$189/200=94,5\%$$

# APRESENTAÇÃO **DOS DADOS**

## Medidas de Dispersão: Interpretação do Desvio-Padrão

Regra Empírica

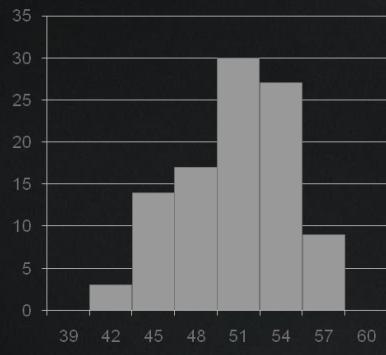
- Média (+/-) desvio 68 % dos casos
- Média (+/-) 2x desvio 95 % dos casos
- Média (+/-) 3x desvio 100 % dos casos



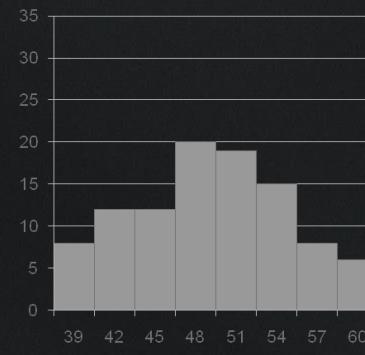
# APRESENTAÇÃO **DOS DADOS**

## Medidas de posição e dispersão

**Exemplo:** Ambos os conjuntos de dados representados a seguir têm média igual a 50. Um deles tem desvio-padrão de 3,8, e o outro, de 5,8. Qual é qual?



Valor (a)



Valor (b)

Desvio-padrão:

# APRESENTAÇÃO **GRÁFICA DOS DADOS**

## ■ Definições

Apesar da apresentação dos dados por meio de tabelas ser mais precisa, a representação gráfica tem a vantagem de transmitir os dados de uma maneira mais rápida, oferecendo um resultado imediato sobre o comportamento da variável que estamos descrevendo.

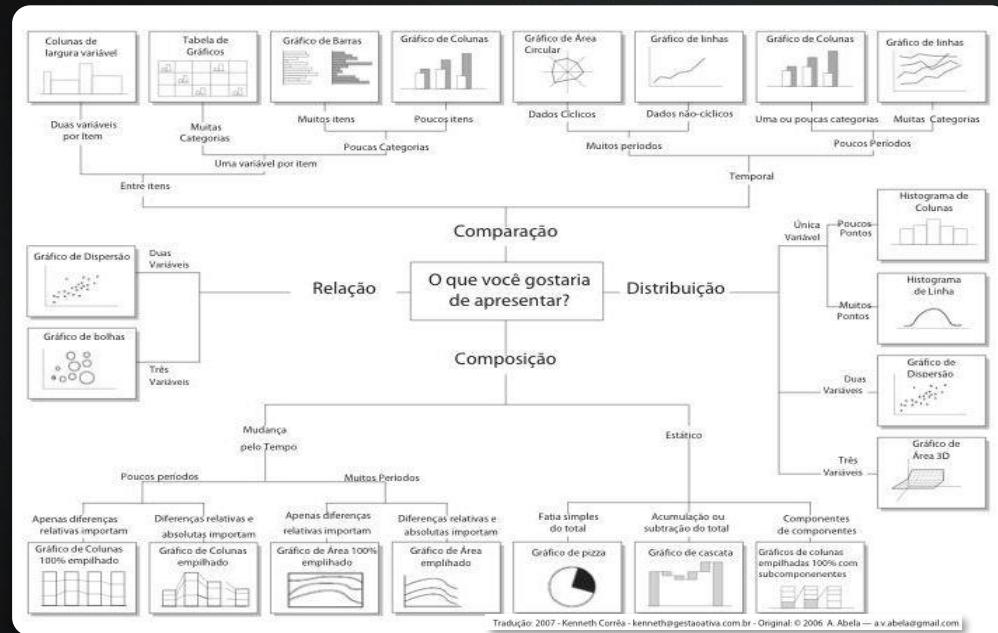
Elaborar gráficos é uma arte que somente pode ser adquirida por meio de prática, com os cuidados necessários para evitar posições tendenciosas, permitindo a visão clara dos pontos essenciais a serem notados.

Portanto, as regras básicas de elaboração de um gráfico são:

- simplicidade;
- clareza;
- veracidade.

# SUGESTÕES DE GRÁFICOS

## UMA IDEIA INICIAL



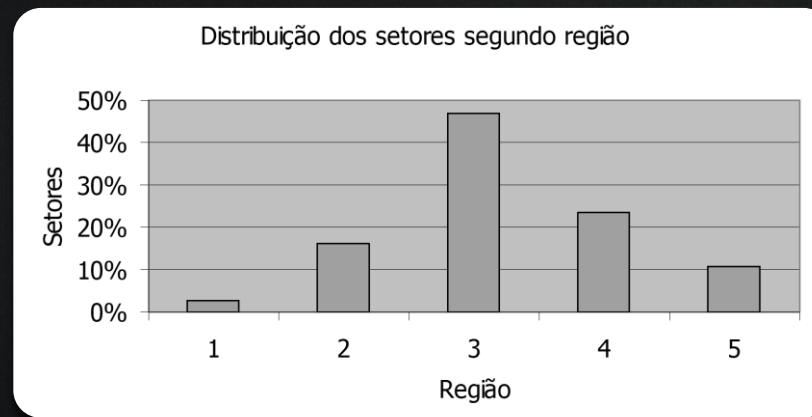
Tradução: 2007 - Kenneth Corrêa - kenneth@gestaoativa.com.br - Original: © 2006 A. Abela - a.abela@gmail.com

# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

- **Variáveis qualitativas ou discretas**

**Colunas:** um gráfico de colunas ilustra a comparações entre itens. As categorias são organizadas na horizontal e os valores são distribuídos na vertical.

**Exemplo:**

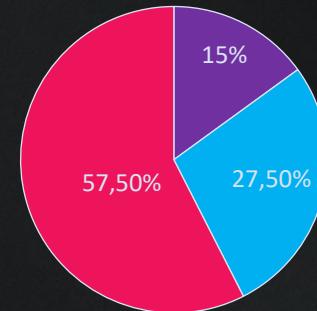
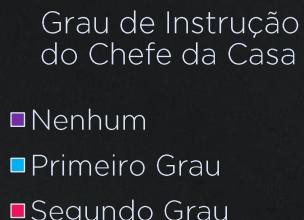
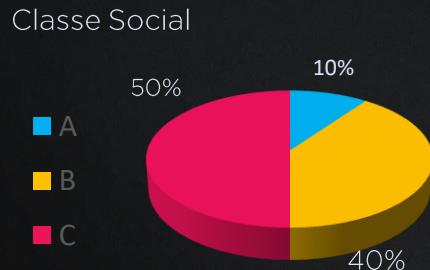


# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

- **Variáveis qualitativas ou discretas**

**Setores ou pizza:** um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens. A frequência relativa (%) transformada em graus mediante o cálculo proporcional.

## Exemplo:

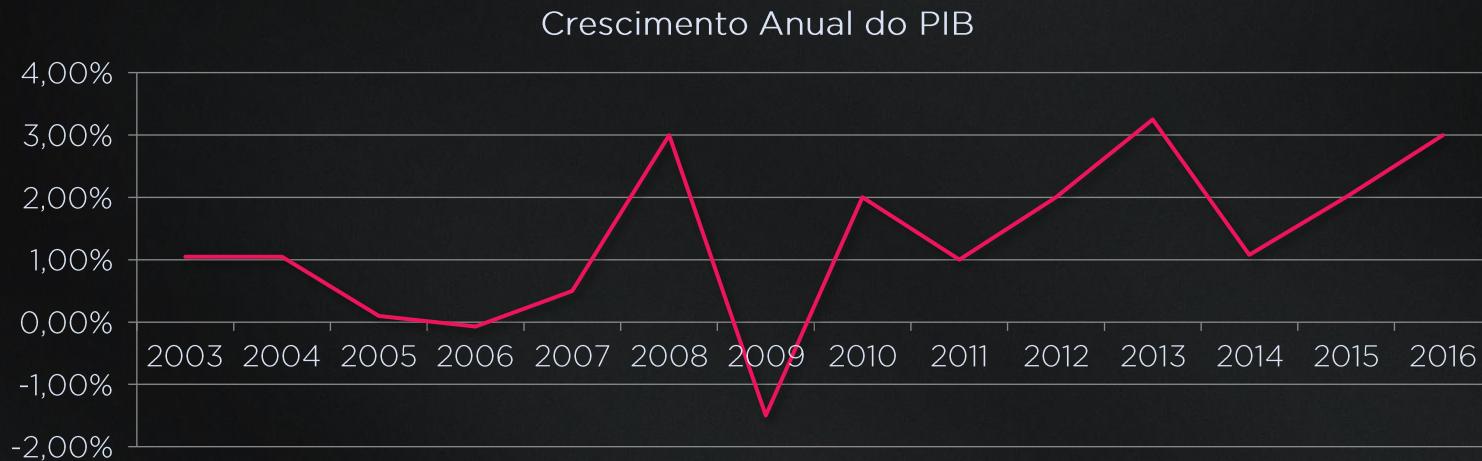


## GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

- **Variáveis qualitativas ou discretas**

**Linha:** um gráfico de linha mostra tendências nos dados em intervalos iguais.

**Exemplo:**



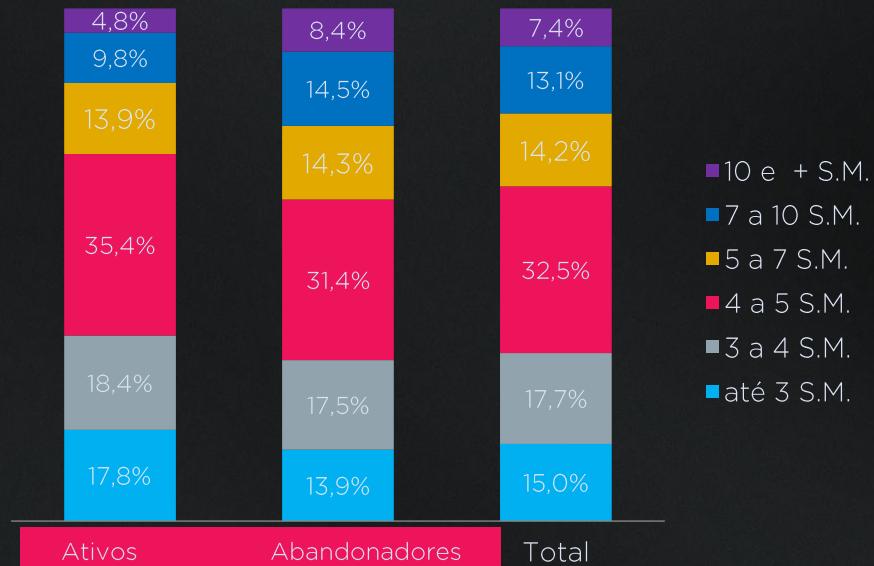
# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

## Variáveis qualitativas ou discretas

**Colunas sobrepostas:** nesta representação as barras estarão sobrepostas, com uso de duas ou mais variáveis. Sendo a soma 100%.

### Exemplo:

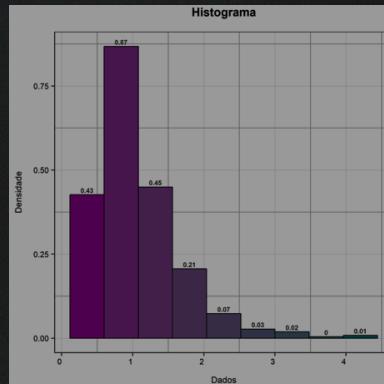
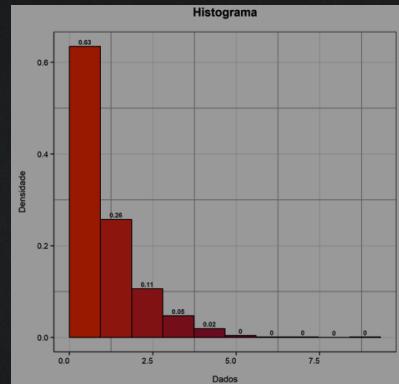
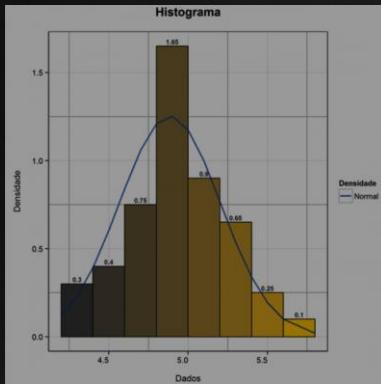
Distribuição de Salários (2017 em SM por situação)



# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

## ■ Variáveis contínuas

**Histograma:** o histograma é formado por retângulos, cujas áreas representam a frequências dos intervalos de suas classes. Essa apresentação é indicada para variáveis contínuas, portanto, não há espaço entre as barras.



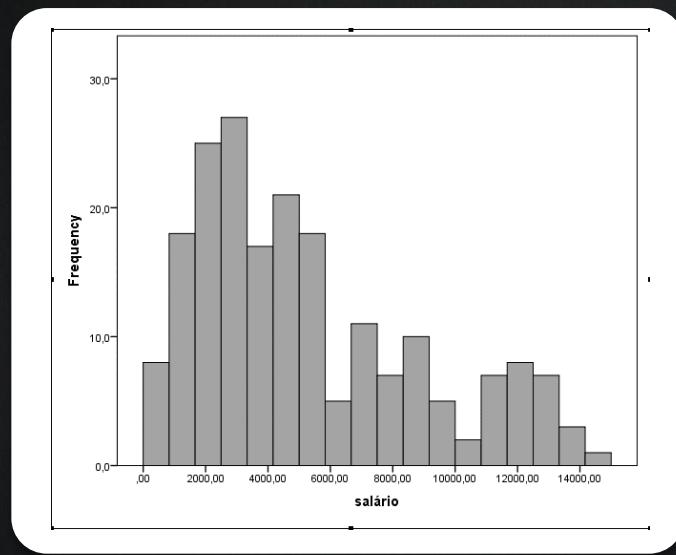
# GRÁFICOS EXISTENTES E SUA ADEQUAÇÃO

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	15.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.856,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.755,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Salário	200	254,19	14.649,52	5.259,30	3.615,36
Valid N (listwise)	200				

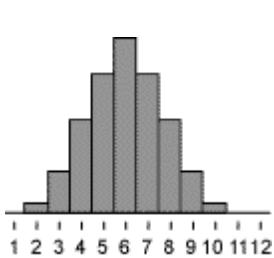
## EXEMPLO



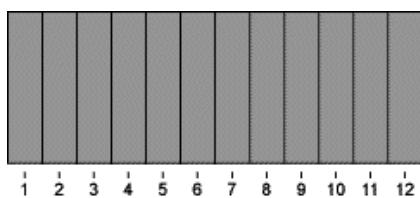
# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

## Histograma

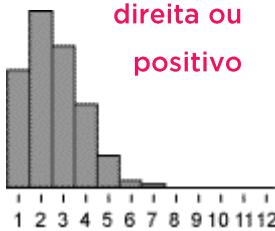
Simétrico



Uniforme



Assimétrico à  
direita ou  
positivo



Assimétrico  
à esquerda  
ou negativo



Moda < Mediana  
< Média

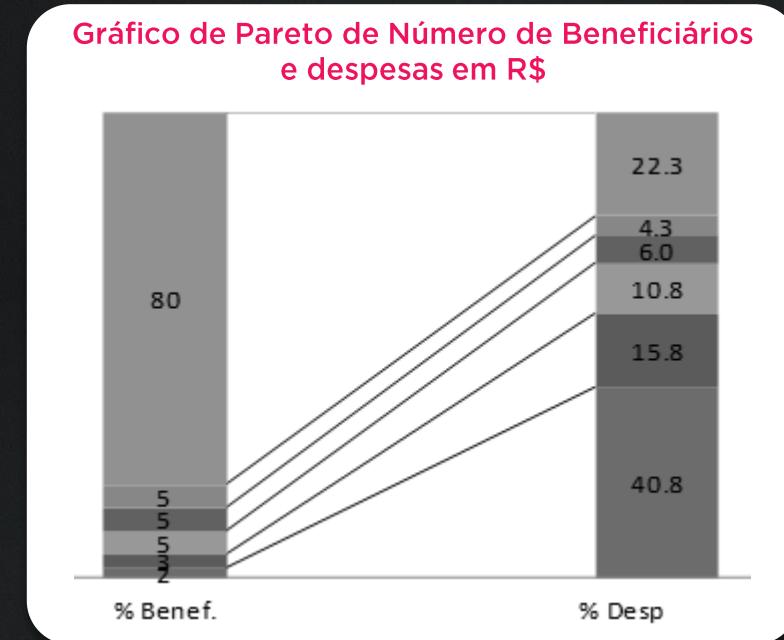
Média < Mediana  
< Moda

# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

- Medidas resumo

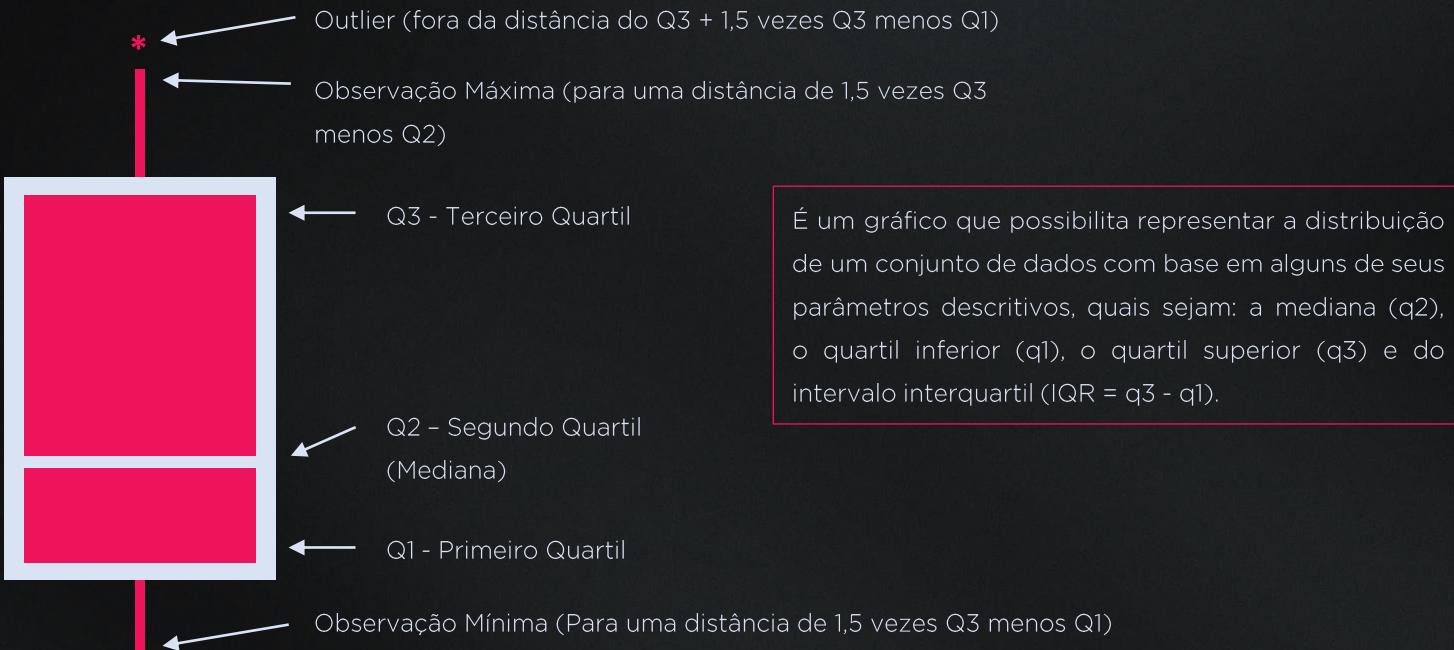
- Exemplo:

Número de Funcionários  
e Despesas em R\$



# GRÁFICOS EXISTENTES **E SUA ADEQUAÇÃO**

## Box Plot

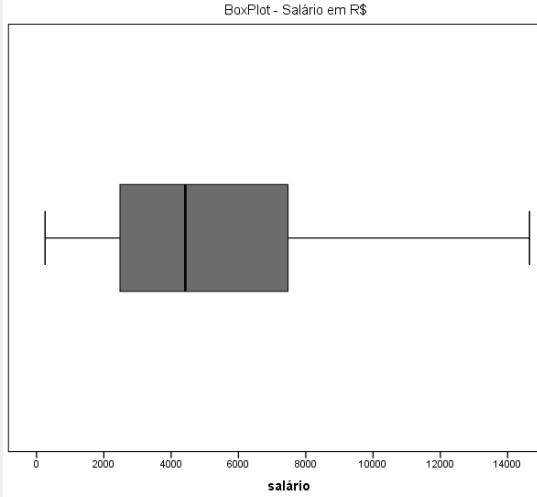


# APRESENTAÇÃO DOS DADOS

id	salário
1	4.763,75
2	7.591,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	15.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.856,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.755,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Salário	200	254,19	14.649,52	5.259,30	3.615,36
Valid N (listwise)	200				



## EXEMPLO

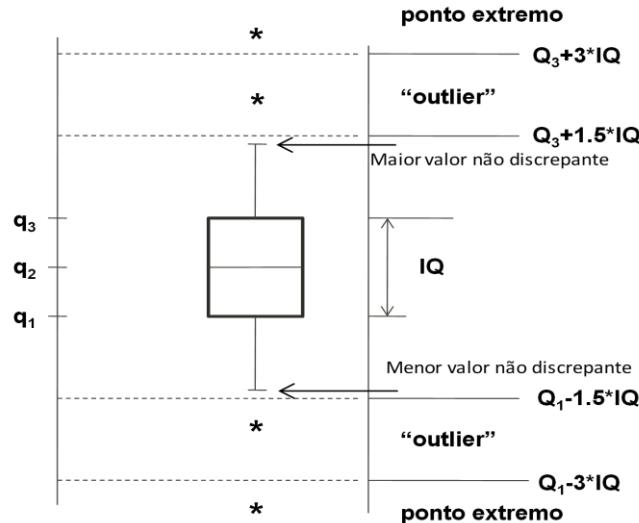
# DETECÇÃO DE OUTLIER

## Outliers

Observações que apresentam um grande afastamento das restantes ou são inconsistentes com elas.

- Dado incorreto
- População diferente
- Dado correto - Evento raro

Representação Gráfica na Análise dos Dados



DATA ANALYTICS

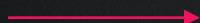
# MÉTRICAS, INDICADORES E MONITORAMENTO

INTRODUÇÃO (ANALYTICS / INSIGHTS)

# KPIs (KEY PERFORMANCE INDICATORS) OU INDICADORES DE DESEMPENHO

**Os indicadores são usados, basicamente, para:**

- Evidenciar a situação atual;
- Definir objetivos futuros;
- Avaliar a efetividade de ações e projetos;
- Para identificar gargalos, tendências e ameaças.



Os indicadores de RH são Instrumentos que servem para monitorar e avaliar a empresa, por meio de seus colaboradores, processos, programas e metas.

**Finalidades:**

- Descrever as atuais circunstâncias que envolvem a empresa.
- Oferecer condições para uma análise detalhada dessas circunstâncias, com a intenção de identificar problemas, fraquezas e desvios que precisam ser corrigidos.



**...são determinantes  
para a tomada de  
decisão.**

# PRINCIPAIS **INDICADORES**

## Índice de Rotatividade(Turnover)

Indica a **quantidade** de colaboradores **desligados** sobre um número total de funcionários ativos na empresa, em um **determinado período**.

## Custos de rotatividade

Abrangem todas as despesas referentes ao pagamento das rescisões contratuais, incluindo multas e tributos. Somam-se a esse valor os gastos para a reposição do profissional desligado, por meio de novos processos seletivos. Além disso, é preciso incluir o investimento feito em treinamentos e qualificação.

## Índice de retenção de talentos

É medido com base no **número de talentos perdidos** para o mercado ou concorrentes.

# PRINCIPAIS INDICADORES

## Tempo médio de empresa

É calculado o tempo médio de permanência, com base no headcount (número de colaboradores) total. Sendo que o detalhamento do perfil dos profissionais, incluindo idade, escolaridade, estado civil e progressão na empresa é importante.

Obs.: relacionado à taxa de rotatividade, à atratividade e à capacidade de retenção.

## Absenteísmo

Medir as taxas de ausência dos colaboradores é fundamental, bem como compreender quais são os motivos dessas faltas ou atrasos. Ex.: casos médicos, problemas pessoais etc.

# PRINCIPAIS INDICADORES

## Produtividade

As métricas de produtividade são construídas para cada tipo de processo e por atividade.

**Produtividade na área de vendas:** dois aspectos mais importantes que envolvem gestão de pessoas: o aumento da receita e a redução de custos. Como conversão, ticket médio, potencial de compra dos clientes .

## PRINCIPAIS INDICADORES

A produtividade da equipe de RH pode ser medida por meio de três elementos básicos:

tempo, qualidade e custos. Por exemplo:

- Atendimento aos prazos.
- A satisfação dos clientes internos.
- A incidência de falhas.
- A redução de despesas.
- A racionalização de recursos.
- A otimização de processos.
- Os períodos desperdiçados com paradas, manutenções corretivas, indisponibilidades e distrações também contribuem para esse indicador.

Do ponto de vista da produtividade da empresa, podem ser considerados como ponto de partida os indicadores de receita por colaborador e de lucro líquido por colaborador.

# PRINCIPAIS INDICADORES

## Avaliação de Aprendizagem

Mensurar a evolução das equipes. **Avaliar a melhora do rendimento de cada profissional** a fim confirmar a eficiência dos programas de treinamento.

## ROI em Treinamentos (Retorno sobre o Investimento em Treinamentos)

Compara os valores gastos em capacitação e melhorias obtidas nos processos e rotinas de trabalho.

## Relação entre horas extras e horas trabalhadas

Relação entre a quantidade de horas extras — pagas ou administradas em banco de horas — e as horas totais trabalhadas. Esse KPI também está ligado ao conceito de produtividade

## Custo per capita de benefícios corporativos

Calcula os custos da empresa com os benefícios corporativos concedidos, de forma per capita. Trata-se de uma comparação entre os gastos absolutos e o número de colaboradores presentes na folha de pagamento, em um mesmo período.

# PRINCIPAIS **INDICADORES**

## Índice de reclamações trabalhistas

O indicador aponta o número e a natureza das reclamações recebidas durante o ano. Algumas situações são bastante típicas e geram reclamações, como as condições de trabalho — envolvendo periculosidade, insalubridade, segurança, ergonomia e saúde —, além dos atrasos nos pagamentos e nos recolhimentos do INSS e FGTS, as horas extras e seus reflexos, os adicionais, a equiparação salarial e o assédio moral.

## Clima Organizacional

Resultado obtido [por meio de uma pesquisa específica](#), conduzida junto aos colaboradores. Utilizando notas, avalia o grau de satisfação em vários aspectos respondendo a questões sobre o relacionamento com os gestores, as oportunidades de crescimento profissional, o acesso a recursos necessários à realização das tarefas cotidianas, políticas de remuneração e benefícios, o incentivo ao aprendizado e às ações de reconhecimento e valorização do indivíduo, entre outras.

# APOIO

Satisfaction_level	Last_evaluation	Number_project	Average_montly_hours	time_spend_company	Work_accident	left	Promotion_last_5years	depto	salary
.380	.530	2	157	3	0	1	0	sales	low
.800	.860	5	262	6	0	1	0	sales	medium
.110	.880	7	272	4	0	1	0	sales	medium
.720	.870	5	223	5	0	1	0	sales	low
.370	.520	2	159	3	0	1	0	sales	low
.410	.500	2	153	3	0	1	0	sales	low
.110	.940	6	286	4	0	1	0	IT	medium
.810	.700	6	161	4	0	1	0	IT	medium
.430	.540	2	153	3	0	1	0	product_mng	medium
.830	.950	4	251	5	0	1	0	marketing	medium
.450	.570	2	148	3	0	1	0	marketing	high
.430	.510	2	141	3	0	1	0	sales	low
.580	.750	4	186	2	0	0	0	product_mng	low
.760	.500	3	258	3	0	0	0	IT	low
.500	.780	3	228	2	0	0	0	RandD	low

# APOIO

<b>id_colab</b>	<b>Potencial</b>	<b>Desempenho</b>	<b>satisfaction_level</b>	<b>last_evaluation</b>
1	7,30	7,66	0,93	0,76
2	7,24	7,61	0,91	0,79
3	7,30	7,66	0,57	0,67
4	7,30	7,66	0,91	0,67
5	7,34	7,61	0,69	0,53
6	7,67	8,11	0,68	0,81
7	7,18	7,29	0,17	0,73
8	7,35	7,66	0,11	0,83
9	5,37	5,90	0,85	0,63
10	7,12	7,41	0,95	0,78
11	7,03	6,39	0,63	0,85
12	5,37	6,24	0,95	0,61
13	6,95	6,05	0,62	0,67
14	6,95	7,41	0,53	0,75
15	8,26	7,80	0,37	0,45
16	5,03	3,47	0,45	0,54
17	8,26	7,75	0,11	0,90
18	8,70	8,53	0,49	0,52
19	7,55	6,07	0,66	0,54
20	7,39	7,61	0,68	0,61
21	7,87	7,87	0,89	0,93

# ANÁLISE **MULTIVARIADA**

## O que são dados multivariados?

- Amostra de indivíduos selecionados aleatoriamente: pessoas residentes em uma cidade, municípios do Brasil, domicílios, funcionários de uma empresa etc.
- Em cada indivíduo são observadas diversas dimensões (variáveis): sexo, idade, número de faltas por ano, tempo de empresa, reações a diferentes tipos de ações, avaliações da empresa em diferentes aspectos etc.
- Em função de essas variáveis serem medidas no mesmo indivíduo, existirão, provavelmente, relações de interdependência e correlações entre elas.

**Como analisar essas informações?**

## • Software Estatístico

- SAS
- SPSS
- Minitab
- STATISTICA
- STATA
- R
- Mplus
- Python
- KNIME
- WEKA



# ANÁLISE EXPLORATÓRIA DE DADOS: **VARIÁVEIS DEPENDENTES E INDEPENDENTES.**

## **Variável dependente**

Mede o fenômeno que se estuda e que se quer explicar. São aquelas cujos efeitos são esperados de acordo com as causas. Elas se situam, habitualmente, no fim do processo causal e são sempre definidas na hipótese ou na questão de pesquisa.

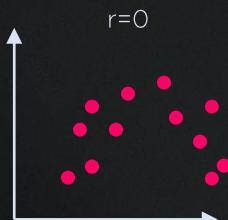
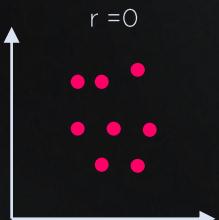
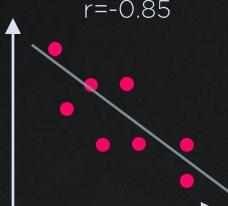
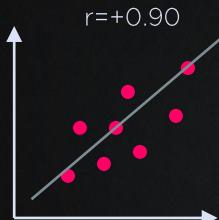
## **Variável independente**

São aquelas variáveis candidatas a explicar a(s) variável(eis) dependente(s), cujos efeitos queremos medir. Aqui, devemos ter cuidado, pois mesmo encontrando relação entre as variáveis, isso não significa, necessariamente, relação causal.

# ANÁLISE EXPLORATÓRIA DE DADOS: CORRELAÇÃO ENTRE VARIÁVEIS

Coeficiente de correlação ( $r$ ) representa a relação linear entre duas variáveis.

Valores de  $r$  e suas implicações.



Correlação Linear Simples ( $r$  de Pearson)

$$\frac{\sum_{i=1} (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sqrt{\sum_{i=1} (X_i - \bar{X})^2 * \sum_{i=1} (Y_i - \bar{Y})^2}}$$

- Para avaliar a correlação entre variáveis, é importante conhecer a magnitude ou força tanto quanto a significância da correlação.

APOIO

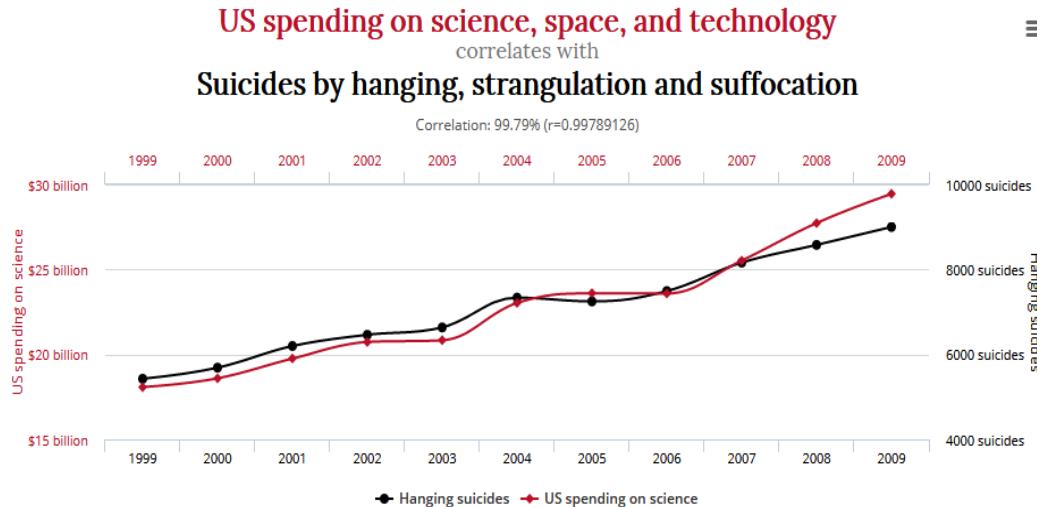
## ANÁLISE EXPLORATÓRIA DE DADOS: **CORRELAÇÃO ENTRE VARIÁVEIS**

### Associações Espúrias

- Associação entre dois fatores e quando queremos saber se um **causa** o outro.
- Big data: muitos resultados estatisticamente significativos que não fazem sentido causal.
- Variável de confusão: quando há muitas variáveis na análise.

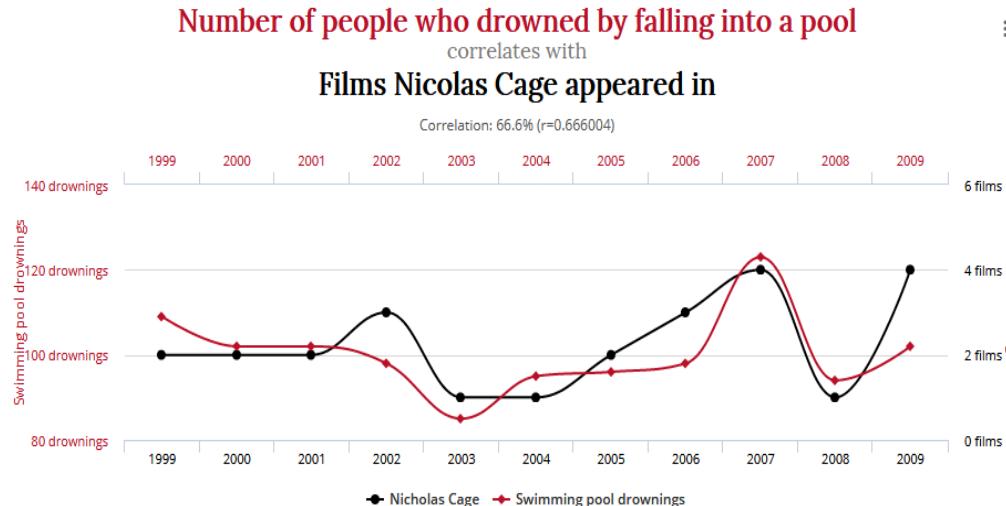
# ASSOCIAÇÕES ESPÚRIAS EXEMPLO

APOIO



# ASSOCIAÇÕES ESPÚRIAS EXEMPLO

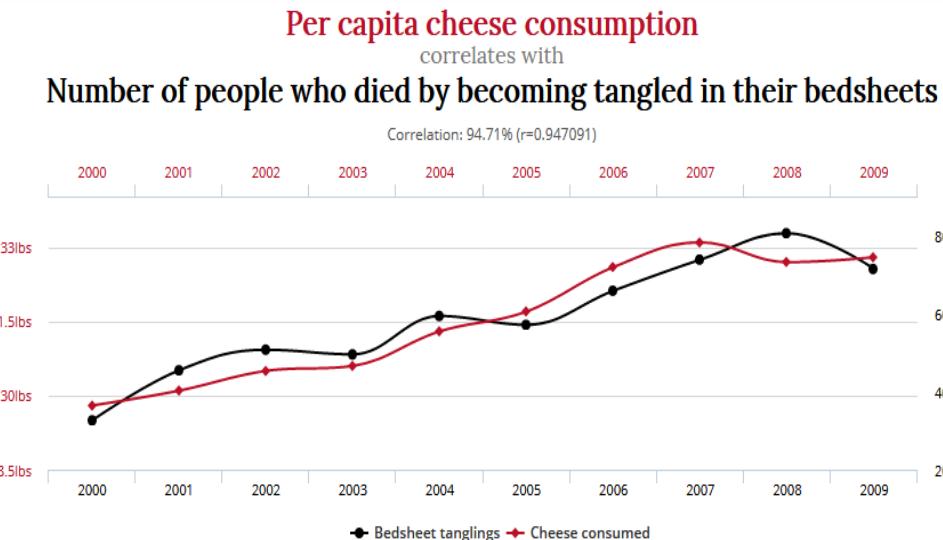
APOIO



# ASSOCIAÇÕES ESPÚRIAS

## EXEMPLO

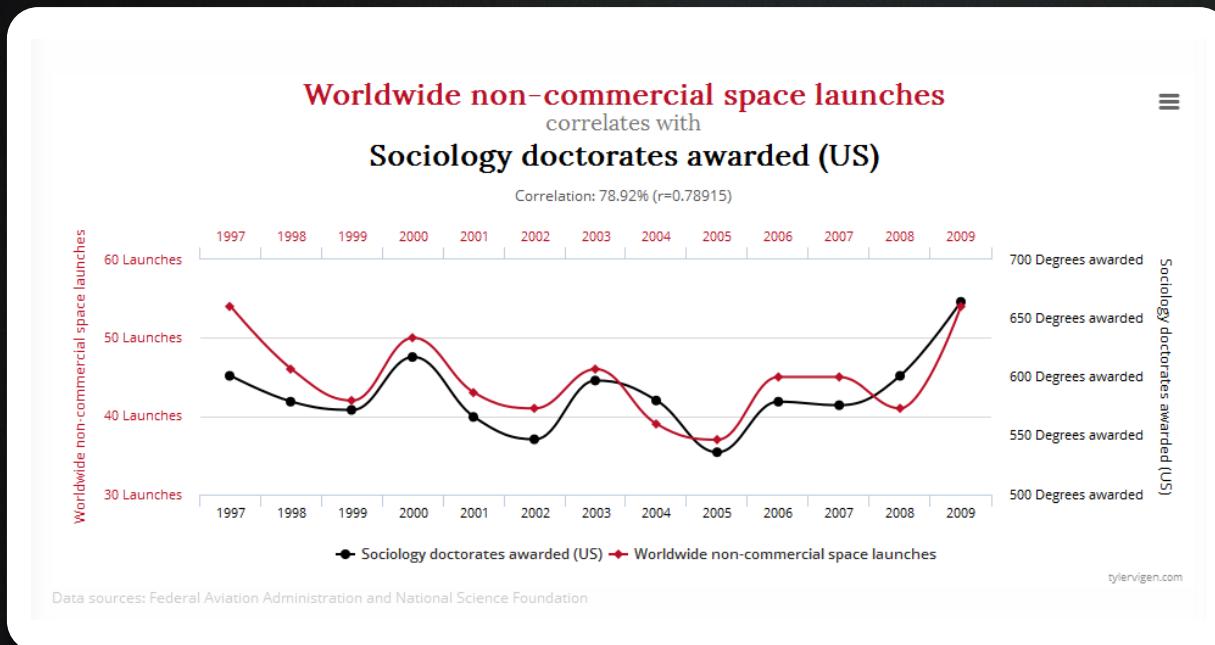
APOIO



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

# ASSOCIAÇÕES ESPÚRIAS EXEMPLO

APOIO



DATA ANALYTICS

**MODELOS ESTATÍSTICOS**

# INTRODUÇÃO

... enfim, seus dados não servem para nada até que você saiba como tirar informações deles

## DESCRITIVO

O que aconteceu?

Quantos colaborador(es) temos?  
Quantos são mulheres ou homens?  
Onde residem e qual a distância da empresa?  
Qual o tempo de casa de cada funcionário?

## DIAGNÓSTICO

Por que isto aconteceu?

Qual a relação entre o desligamento voluntário x sexo x tempo de deslocamento médio x distância de casa?

## PREDITIVO

O que acontecerá?

Quanto(a)s colaborador(es) precisaremos contratar nos próximos três anos considerando o perfil da população e comportamento do turnover?  
Qual a probabilidade de turnover em um determinado grupo?

## PRESCRITIVO

O que posso fazer?

Lista de ações para recrutar colaboradores por cargo nos canais A, B e C.  
Quem queremos reter? E, demitir???

# ANÁLISE MULTIVARIADA

## Análise Exploratória dos Dados

### Análise de Discriminação de Estrutura

- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis quando **uma das variáveis pode ser identificada como dependente (variável target)**, e as restantes como variáveis independentes.

### Análise Estrutural

- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. **Não há distinção entre variáveis dependentes e independentes.**

### Análise Supervisionada

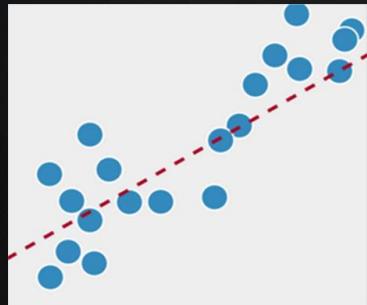
### Análise Não Supervisionada

# ANÁLISE SUPERVISIONADA

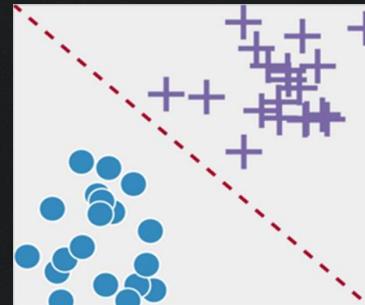


Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

- **Regressão:** Compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores numéricos reais. Esta tarefa é similar à tarefa de Classificação, com a diferença de que o **atributo alvo assume valores numéricos**.

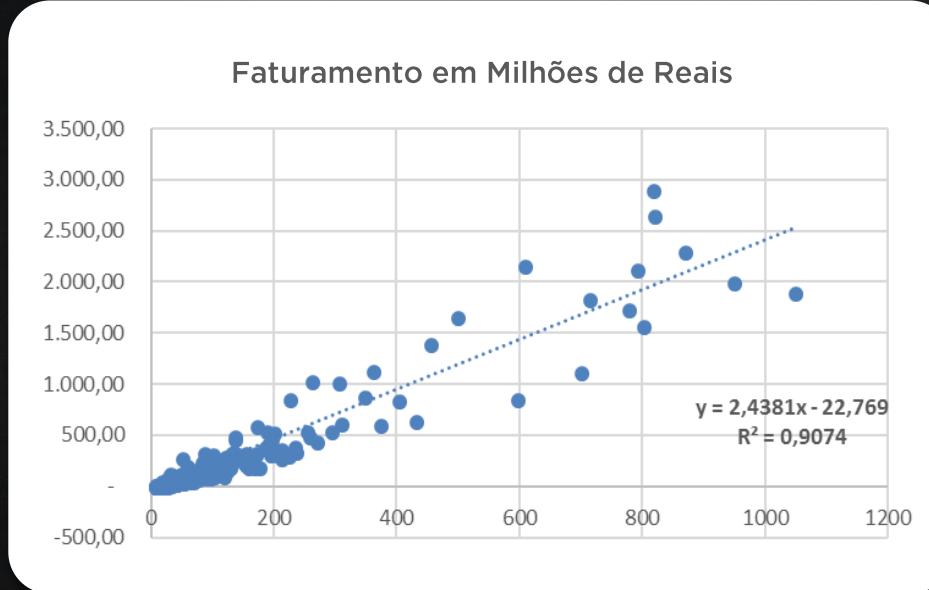


- **Classificação:** A tarefa de Classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram.



# ANÁLISE SUPERVISIONADA

- Exemplo: Faturamento anual (em milhões de Reais) por número de funcionários.



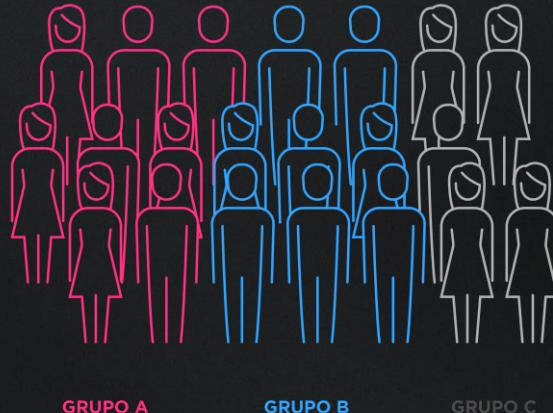
# ANÁLISE SUPERVISIONADA

- Como os funcionários com maiores avaliações se diferem, em seu perfil demográfico, dos demais?
- Como se diferem os funcionários que abandonarem a empresa dos que estão ativos?
- Quais benefícios fazem com que os meus funcionários prefiram a empresa?



Como separar grupos **previamente definidos**? Como definir critérios, funções das variáveis que discriminem os grupos?

## VARIÁVEL CATEGÓRICA



# ANÁLISE **SUPERVISIONADA**

- Existe uma estrutura, ou seja, um fato.
- **Entender esse fato.**

**Entendimento  
do fato**



**Previsão**



**Variável resposta contínua**

- Séries Temporais
- Regressão

**Classificação  
ou  
Discriminação**



**Variável resposta categórica**

- Logística
- Árvore de Decisão

# ANÁLISE **SUPERVISIONADA**

## **PREVISÃO**

**Exemplo:**

Estimar o **salário** de um cargo/ocupação em função:

- Localização, gênero, idade, cor e grau de escolaridade.

**Solução: Usar a Regressão linear**

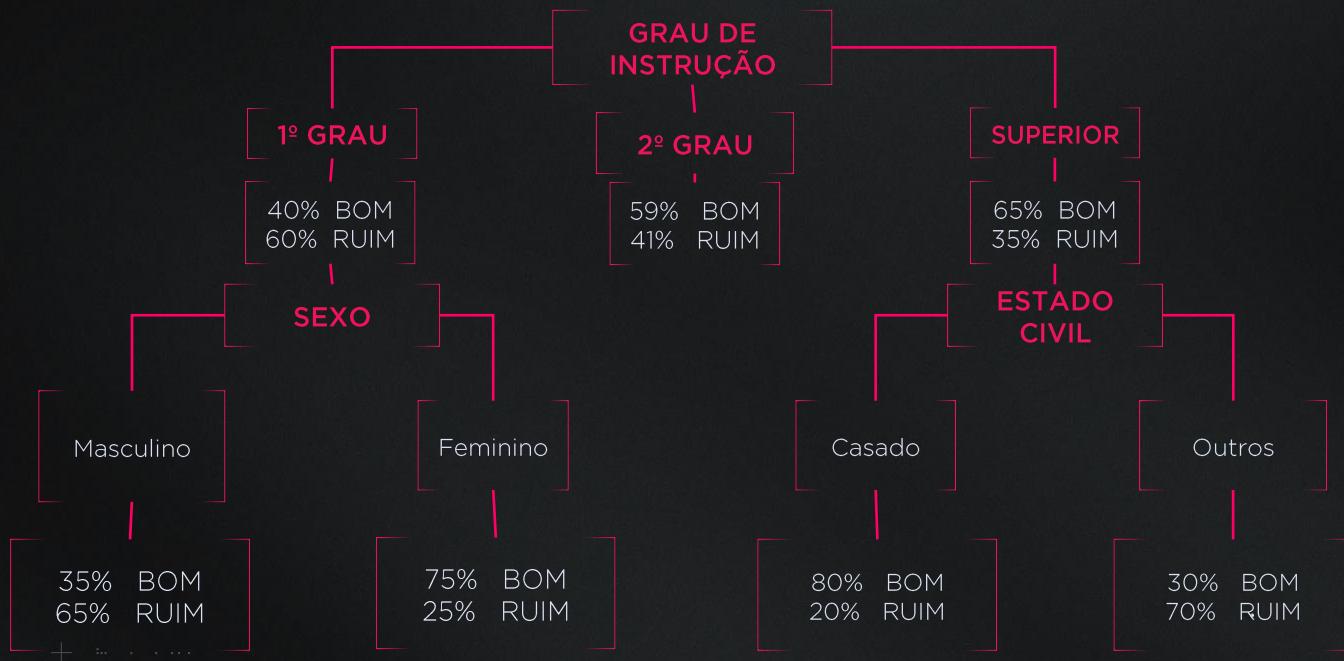
$$Y_i = \alpha + \beta * X_i$$

# ANÁLISE SUPERVISIONADA

## CLASSIFICAÇÃO

Não abandonou 50%  
Abandonou 50%

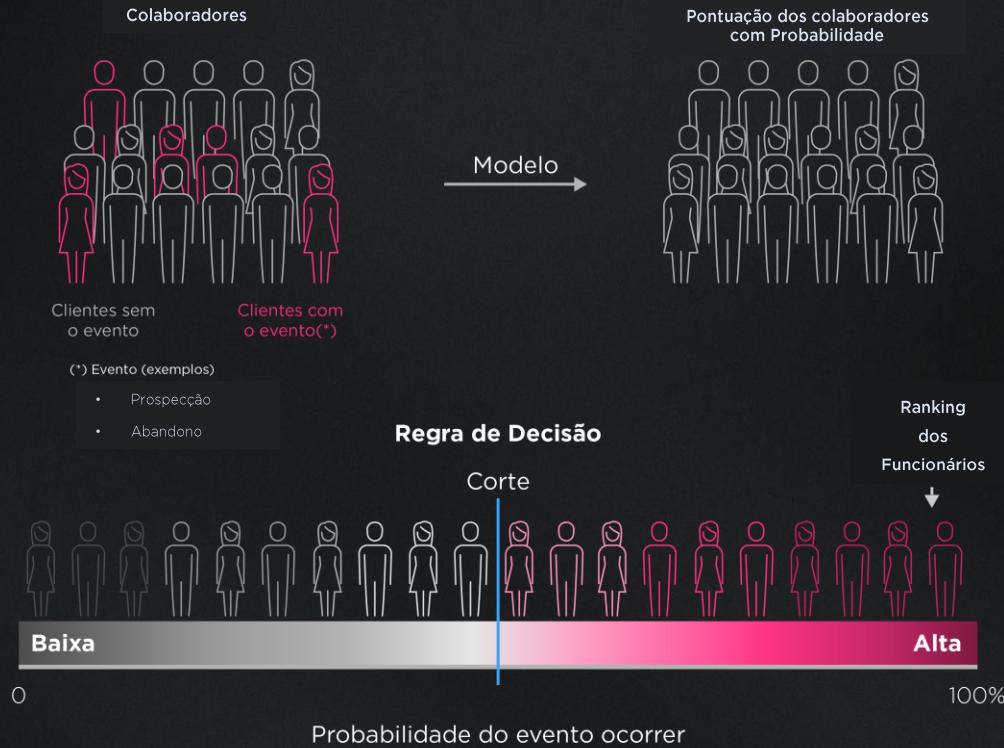
**EXEMPLO**  
MODELO DE ABANDONO  
ÁRVORES DE DECISÃO



# ANÁLISE SUPERVISIONADA

## CLASSIFICAÇÃO

### REGRESSÃO LOGÍSTICA



# ANÁLISE SUPERVISIONADA

## CLASSIFICAÇÃO

Pesos definidos na modelagem

### EXEMPLO

MODELO DE ABANDONO  
MODELO LOGÍSTICO

-0,24	Grupo 7	Departamento	Grupo 1	0,29
-1,84	Grupo 1	Grupo de CEP	Grupo 6	1,34
-0,63	46 ou mais	Tempo de cassa	Menos de 12	0,73
-0,63	0	Quantidade de Faltas	6 ou mais	
-0,59	0	Média de dias de Atraso	Mais de 24	0,88
	Mais de R\$6.600	Valor do Salário	Menos de 1.000	
-0,11	Acima de 59 anos	Faixa Etária	18 a 23 anos	0,18
	2 ou mais	Dependentes	0	
0,23		Constante		0,23
4%	Propensão			98%

Exemplo: risco de perder um funcionário

# ANÁLISE SUPERVISIONADA

## CLASSIFICAÇÃO

As redes neurais usam dados de entrada.

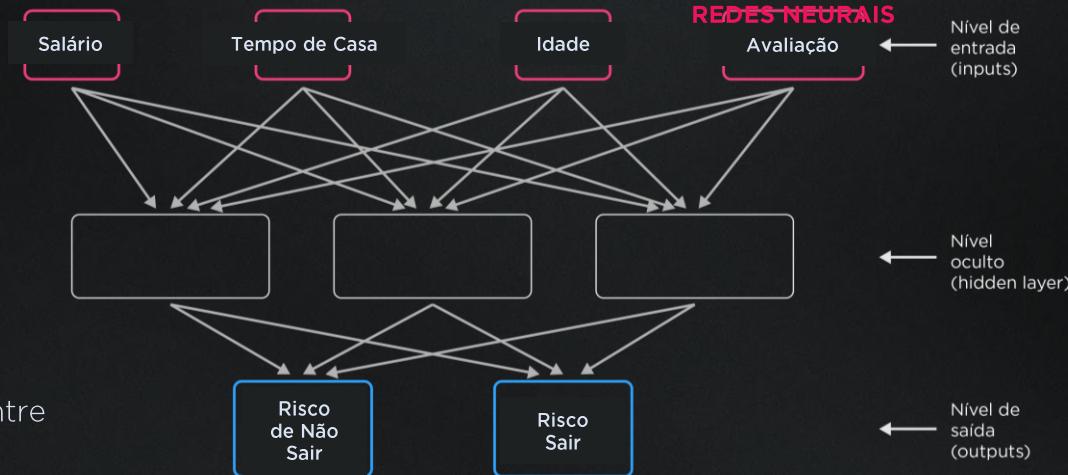
Atribui pesos nas conexões entre os atributos (neurônios).

E obtém um resultado (risco de perder um funcionário) - nível de saída.

### EXEMPLO

MODELO DE ABANDONO

**REDES NEURAIS**



**Exemplo:** risco de perder um funcionário

# ANÁLISE SUPERVISIONADA

- Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.
  - Previsão de Séries Temporais:** Uma série temporal é um conjunto de observações de um fenômeno (variável numérica) ordenadas no tempo. A previsão de uma série temporal tem como objetivo inferir valores que a variável da série deverá assumir no futuro considerando como base valores passados dessa série.



# ANÁLISE SUPERVISIONADA

## PREVISÃO

Séries Temporais → Conhecimento do Histórico

### Exemplo:

- Venda de protetor solar para o próximo verão: vendas( $t$ ) =  $v(t-1) + v(t-2) + \text{constante}$ .
- Quantidade de faltas de funcionários por dia.
- Vendas por ano/mês.

# ANÁLISE **SUPERVISIONADA** **PREVISÃO**

**Exemplo:** Qual a estimativa de vendas para 2017?

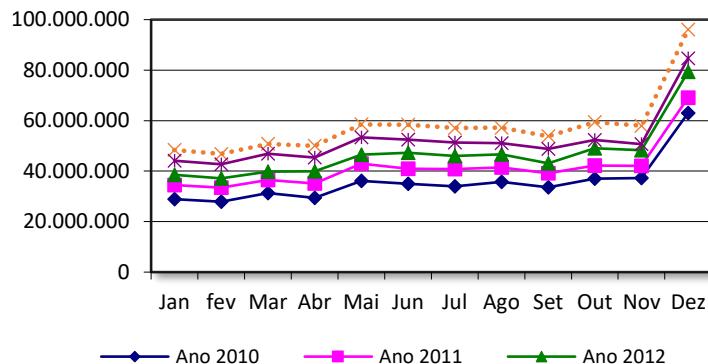


# ANÁLISE SUPERVISIONADA

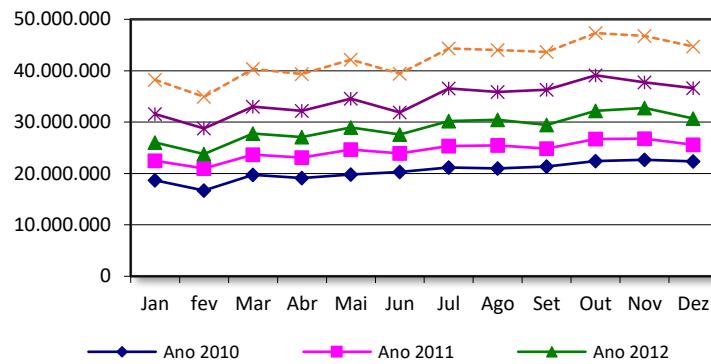
## PREVISÃO

Quantidade de transações mensais com cartões de crédito

Transações Crédito - Comércio Varejista



Transações Crédito - Turismo & Entretenimento



# ANÁLISE MULTIVARIADA

## Análise Exploratória dos Dados

### Análise de Discriminação de Estrutura

- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis quando uma das variáveis pode ser identificada como dependente (variável target), e as restantes como variáveis independentes.

### Análise Estrutural

- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. Não há distinção entre variáveis dependentes e independentes.

### Análise Supervisionada

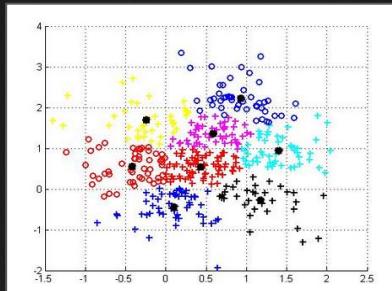
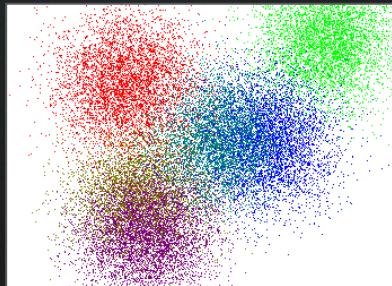
### Análise Não Supervisionada

# ANÁLISE NÃO SUPERVISIONADA



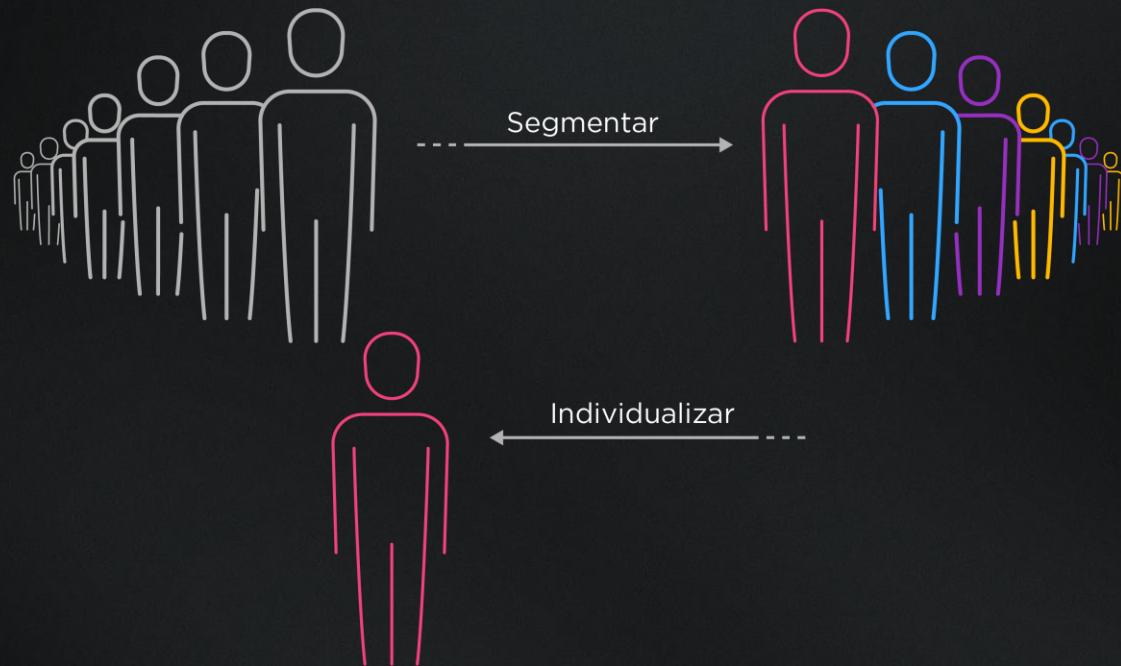
Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

- **Agrupamento (Clusterização):** Consiste em segmentar os registros do conjunto de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem propriedades comuns que os distingam de elementos nos demais clusters. O objetivo nesta tarefa é maximizar a similaridade intracluster e minimizar a similaridade intercluster.



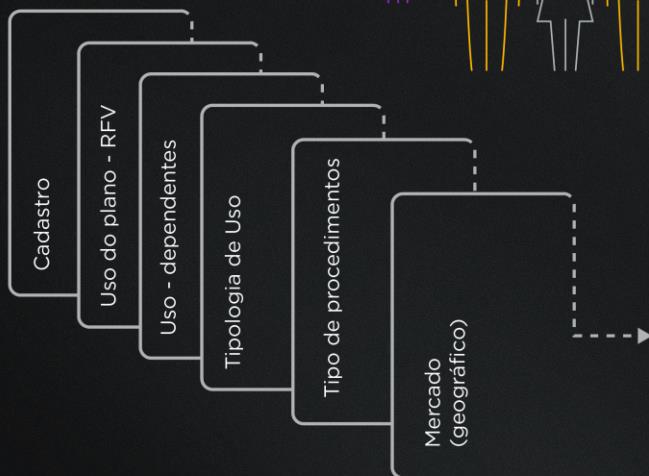
# ANÁLISE NÃO SUPERVISIONADA CLUSTERIZAÇÃO

Instrumentalização  
da Estratégia  
do Relacionamento

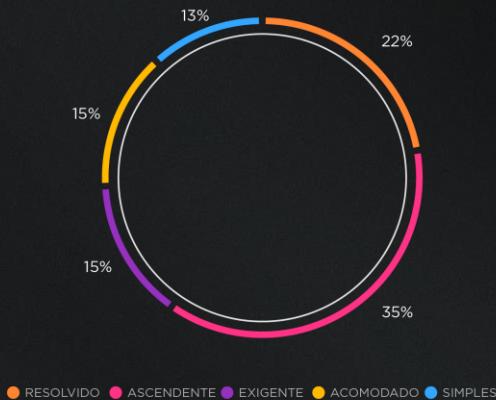


# ANÁLISE NÃO SUPERVISIONADA CLUSTERIZAÇÃO

Segmentação dos  
beneficiários com relação  
ao uso do plano de saúde



## EXEMPLO



# ANÁLISE NÃO SUPERVISIONADA



Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

- **Sumarização:** Consiste em identificar e indicar similaridades entre registros do conjunto de dados.

Construção do indicador  
de “Valor”

- Produtividade
- Nível de satisfação
- Quantidade de Projetos
- Última Avaliação
- Tempo de empresa
- Potencial
- Assiduidade
- outros.....

## ANÁLISE NÃO SUPERVISIONADA SUMARIZAÇÃO

- Analisa a estrutura de interrelações existentes entre certo número de variáveis contínuas ou discretas.

Redução da dimensionalidade dos dados multivariados, transformando-os em variáveis correlacionadas em variáveis não correlacionadas transformadas linearmente. Poucos fatores não correlacionados são extraídos, explicam a máxima quantidade de variância comum e são responsáveis pela correlação observada entre os dados multivariados. As relações entre as variáveis e os fatores são, então, estudadas (investigadas).

- Entendimento das variáveis latentes.
- Criação de Indicadores.

Exemplo:  
ANÁLISE **FATORIAL**

# ANÁLISE NÃO SUPERVISIONADA SUMARIZAÇÃO

## EXEMPLO

Acesso ao conhecimento: educação.

- Taxa de alfabetização da população acima de 15 anos.
- Proporção de pessoas com acesso aos níveis de ensino primário.

IDH

Direito a uma vida longa e saudável: longevidade.

- Expectativa de vida ao nascer.

Direito a um padrão de vida digno:

- Renda PIB per capita.

# ANÁLISE NÃO SUPERVISIONADA

## SUMARIZAÇÃO

### EXEMPLO

IDH



MUNICÍPIO	UF	Esp_Vida	Tx_alfab	Tx_freq_esc	rendacapita	IDH_M	Class_UF	Class_BR
São Caetano do Sul	SP	78,18	97,01	98,57	834,00	<b>0,919</b>	1	1
Águas de São Pedro	SP	77,44	97,06	85,75	954,65	<b>0,908</b>	2	2
Santos	SP	72,27	96,44	92,62	729,62	<b>0,871</b>	3	6
Vinhedo	SP	74,87	94,08	79,73	627,47	<b>0,857</b>	4	15
Jundiaí	SP	73,94	94,99	88,46	549,96	<b>0,857</b>	5	17
Ribeirão Preto	SP	74,40	95,56	84,21	539,84	<b>0,855</b>	6	22
Santana de Parnaíba	SP	71,35	92,06	87,55	762,05	<b>0,853</b>	7	25
Campinas	SP	72,22	95,01	87,54	614,86	<b>0,852</b>	8	26
Saltinho	SP	77,35	95,78	80,34	406,27	<b>0,851</b>	9	28
Ilha Solteira	SP	75,80	94,77	90,74	390,05	<b>0,850</b>	10	33
São José dos Campos	SP	73,89	95,42	89,20	470,01	<b>0,849</b>	11	36
Araçatuba	SP	74,52	93,69	85,34	503,17	<b>0,849</b>	12	41
Paulínia	SP	73,30	93,93	89,37	503,34	<b>0,847</b>	13	44
Presidente Prudente	SP	73,58	93,81	89,58	482,62	<b>0,846</b>	14	47
São João da Boa Vista	SP	76,92	93,56	79,51	408,33	<b>0,843</b>	15	56
Valinhos	SP	71,91	94,42	84,54	569,31	<b>0,842</b>	16	63
São Carlos	SP	73,08	94,36	89,61	456,25	<b>0,841</b>	17	65
São Paulo	SP	70,66	95,11	85,48	610,04	<b>0,841</b>	18	68
Americana	SP	72,46	95,62	87,15	473,23	<b>0,840</b>	19	71
Pirassununga	SP	75,16	93,95	84,33	402,30	<b>0,839</b>	20	77
Taubaté	SP	72,73	95,18	85,12	460,86	<b>0,837</b>	21	87
Piracicaba	SP	72,95	94,95	84,05	455,87	<b>0,836</b>	22	93
Santo André	SP	70,61	95,55	88,59	512,87	<b>0,836</b>	23	94
Caçapava	SP	74,88	93,88	86,86	363,53	<b>0,835</b>	24	96
Cordeirópolis	SP	76,82	93,28	77,86	367,03	<b>0,835</b>	25	97
Tremembé	SP	74,47	94,43	84,83	383,76	<b>0,834</b>	26	99
São José do Rio Preto	SP	71,31	94,61	85,55	512,01	<b>0,834</b>	27	102
São Bernardo do Campo	SP	69,93	95,02	91,93	505,45	<b>+ 0,834</b>	28	106

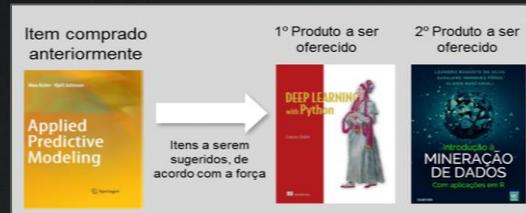
# ANÁLISE NÃO SUPERVISIONADA



Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

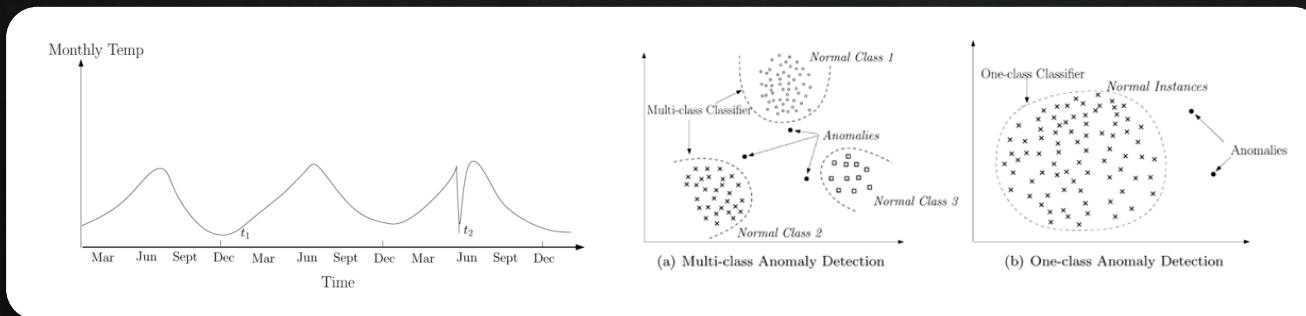
- **Descoberta de Associações:** Nesta tarefa, cada registro do conjunto de dados é normalmente chamado de transação. Cada transação é composta por um conjunto de itens. A tarefa de descoberta de associações compreende a busca por itens que frequentemente ocorrem de forma simultânea em uma quantidade mínima de transações do conjunto de dados.

- **Descoberta de Sequências:** É uma extensão da tarefa de Descoberta de Associações cujo propósito é identificar itens frequentes considerando um determinado período de tempo. Consideremos o exemplo das compras no supermercado. Se o banco de dados possui a identificação do cliente responsável por cada compra, a descoberta de associações pode ser ampliada de forma a considerar a ordem em que os produtos são comprados ao longo do tempo.



# ANÁLISE SUPERVISIONADA ou NÃO SUPERVISIONADA

- Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.
- **Detecção de Desvios:** Tal tarefa consiste em identificar registros do conjunto de dados cujas características destoem dos que se consideram a norma no contexto em análise. Tais registros são denominados valores atípicos (outliers).



DATA ANALYTICS

**MODELOS ESTATÍSTICOS**

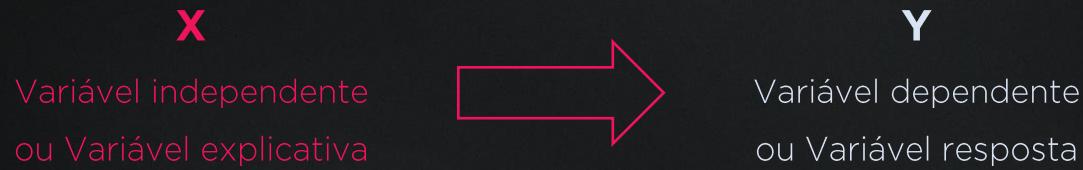
# DATA ANALYTICS **MODELOS ESTATÍSTICOS**

TÉCNICAS DE PREDIÇÃO E CLASSIFICAÇÃO  
**(REGRESSÃO LINEAR, LOGÍSTICA, ÁRVORE DE DECISÃO)**

# ANÁLISE DE REGRESSÃO

## MODELO LINEAR

A análise de regressão é, geralmente, feita sob um referencial teórico que justifique a adoção de alguma relação matemática de causalidade.



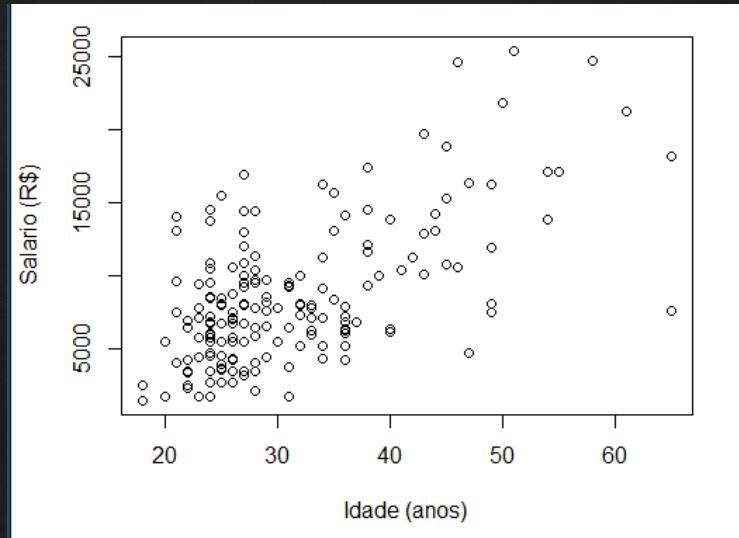
- Predizer valores de uma variável dependente (**Y**) em função de uma variável independente (**X**)
- Conhecer o quanto variações de **X** podem afetar **Y**

# ANÁLISE DE REGRESSÃO

## MODELO LINEAR

### Exemplo:

Salário (em Reais) em função  
da idade (em anos)



# ANÁLISE DE REGRESSÃO

## MODELO LINEAR

Técnica Estatística que relaciona funcionalmente, uma variável dependente das suas possíveis variáveis explicativas.

- Modelo Linear a Duas Variáveis
- Modelo Linear Múltiplo

### **Exemplos:**

- Estimar o faturamento de lojas a partir de suas características físicas;
- Estimar quantidade de faltas no ano em função de características do funcionário;
- Estimar quais são as variáveis que afetam a estimativa de satisfação.

# ANÁLISE DE REGRESSÃO

## MODELO LINEAR

### Análise de regressão - Exemplos

#### Variável Independente X

Características físicas: quantidade de funcionários, ckouts, vagas em estacionamento, área m<sup>2</sup> etc.

Características das campanhas: público, canal, investimento etc.

Grau de Escolaridade, idade, sexo etc.

#### Variável Dependente Y

Faturamento de Loja

Performance de campanhas

Salário

# ANÁLISE DE REGRESSÃO **MODELO LINEAR**

- O Modelo que relaciona Y com várias variáveis independentes

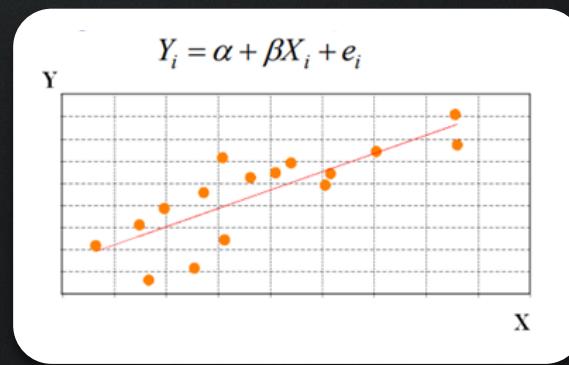
- **Modelo Linear Simples:**  $Y = B_0 + B_1 X + e$

X = variáveis independentes

Y = variável dependente

$B_0$  = constante

$B_1$  = coeficientes de regressão



# ANÁLISE DE REGRESSÃO **MODELO LINEAR**

- O Modelo que relaciona Y com várias variáveis independentes

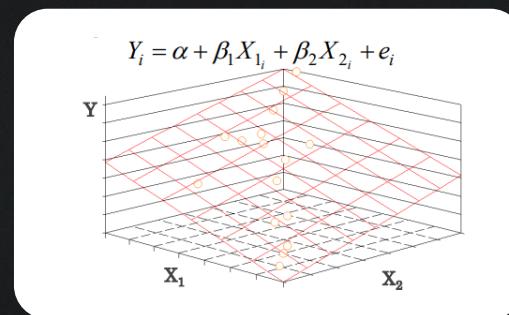
- **Modelo Linear Múltiplo:**  $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_nX_n + e$

$X_1, X_2, X_3, \dots, X_n$  = variáveis independentes

Y = variável dependente

$B_0$  = constante

$B_1, B_2, B_3, \dots, B_n$  = coeficientes de regressão associados às n variáveis



# ANÁLISE DE REGRESSÃO

## MODELO LINEAR

**Técnica Estatística quantitativa aplicada nas condições:**

- Informações do passado disponíveis;
- As relações entre as variáveis devem ser lineares;
- Informações quantificáveis em forma numérica;
- Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

**Modelo Causal permite:**

- Expressar as relações de Causa-Efeito entre variáveis;
- Entender melhor os mecanismos geradores do fato em estudo;
- Simular situações de forma a se avaliar o seu impacto na previsão;
- Analisar situações independentes do tempo.

# ANÁLISE DE REGRESSÃO

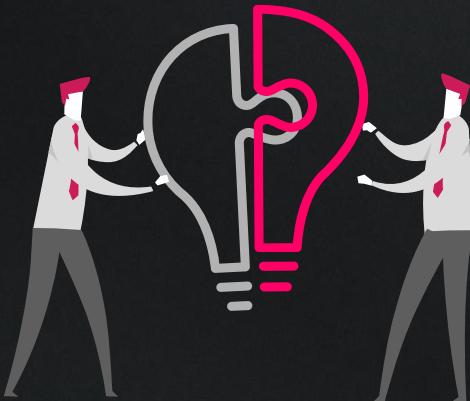
## MODELO LINEAR

O critério de Mínimos Quadrados é um bom ajuste:

- Escolhe a reta que minimiza a soma dos quadrados dos desvios;
- As distribuições amostrais são conhecidas;
- Sob certas condições, as distribuições amostrais dos estimadores de mínimos quadrados de  $B_0$  e  $B_1$  têm menores desvios padrão do que qualquer tipo de estimadores.

# ANÁLISE DE REGRESSÃO

## MODELO LINEAR



Base  
Salário 

# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO

Quando usar?

ÁRVORES DE DECISÃO



- Resultado: Regra/ Critério
- Variáveis preditoras: categóricas ou intervalares

REGRESSÃO LOGÍSTICA



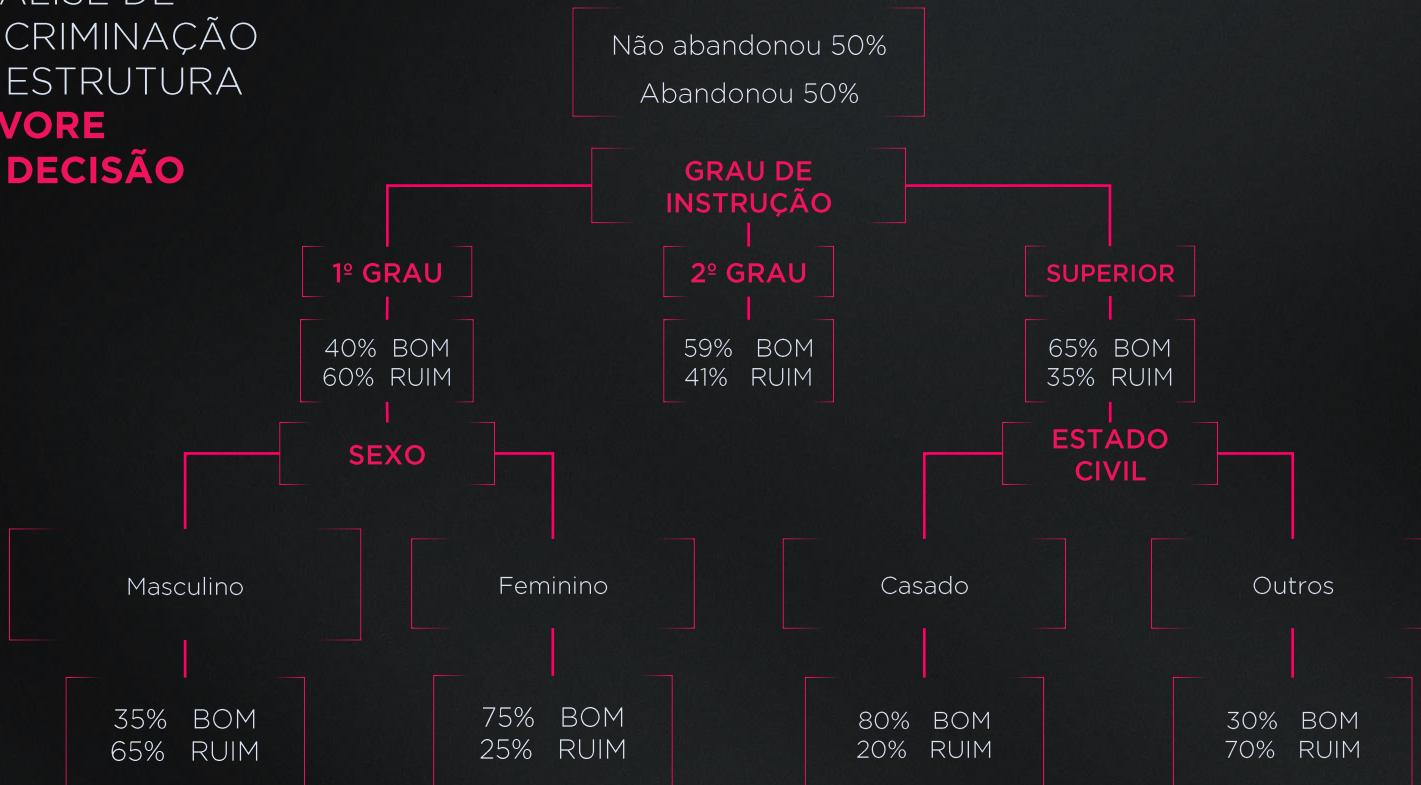
- Resultado: Modelo/ Função
- Variáveis preditoras: categóricas ou intervalares

## MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: **ÁRVORES DE DECISÃO**

- Metodologia estatística de fácil interpretação e utilização.
- São estruturas de dados compostas de um nó raiz e vários nós filhos que, por sua vez, têm seus filhos também e se interligam por ramos, cada um representando uma regra. Os nós que não possuem filhos são chamados de nós folhas e os que têm são chamados de nós pais, ou de decisão.
- Têm como objetivo encontrar regras que discriminem dois grupos previamente conhecidos.
- Exemplo: Encontrar uma regra que trace perfil de pessoas mais propensas a abandonar o emprego.

# ANÁLISE DE DISCRIMINAÇÃO DE ESTRUTURA

## ÁRVORE DE DECISÃO



# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: **ÁRVORES DE DECISÃO**

## Algoritmos utilizados:

- **CHAID:** CHi-square Automatic Interaction Detector
- **CART:** Classification And Regression Trees

## Tipos de Variáveis

- Variáveis Categóricas (nominais ou ordinais)
- Variável Frequência e Ponderada (Weight)

# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: **ÁRVORES DE DECISÃO**

## Exemplo

### **Segmento: Seguro Residencial**

A área de Seguros Residenciais deseja avaliar a estrutura de atendentes que tem como função a implantação das novas apólices . A área tem alocado um razoável número de funcionários e quer implementar um processo automático de aprovação das propostas, garantindo que somente as apólices de menor risco fossem liberadas automaticamente.

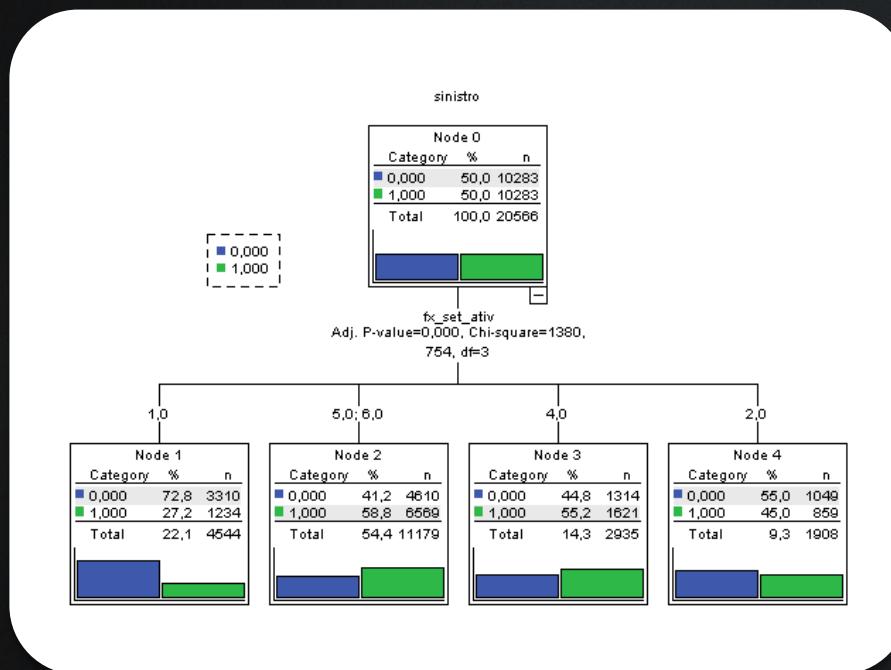
# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: ÁRVORES DE DECISÃO

apolice	parcelas	qtde_cob	tpconstr	tipmora	clasmora	corretor	corrent	uf	set_ativ	Impseg(R\$)	sinistro
925578	6	9	6	casa	moradia	2	N	MS	90	100000	1
395699	1	9	6	apto	moradia	1	S	ES	26	30000	0
863771	11	9	6	casa	moradia	1	S	SP	24	200000	0
892165	11	9	6	casa	moradia	1	S	MG	27	30000	0
923092	1	9	6	casa	veraneio	2	N	SP	90	70000	0
1003098	4	9	6	casa	veraneio	1	S	SP	7	150000	1
955644	11	9	6	casa	moradia	1	S	MG	11	30000	1
987421	1	9	6	casa	moradia	2	N	SP	90	65000	1
744959	4	9	6	casa	veraneio	1	S	RS	18	70000	1
920814	11	9	6	casa	moradia	2	S	SP	90	100000	0
395550	2	9	6	casa	moradia	1	S	ES	26	20000	0
972615	6	9	6	casa	veraneio	2	N	SP	90	87500	1
958900	11	9	6	casa	moradia	1	S	MG	23	85000	1
911272	4	9	6	casa	veraneio	2	N	SP	90	150000	0
895508	11	9	6	casa	moradia	1	S	MG	33	50000	0
374234	1	9	6	apto	moradia	1	N	DF	6	30000	0
883254	11	9	6	casa	moradia	1	S	SP	24	100000	0
727885	3	9	6	casa	moradia	2	S	RS	90	180000	1
327315	11	9	6	casa	moradia	1	S	BA	21	20000	0
910241	11	9	6	apto	moradia	1	S	SP	49	50000	0
956554	10	9	6	casa	moradia	1	S	MG	27	70000	1
1000162	3	9	6	casa	moradia	2	S	MS	90	80000	1
920421	1	9	6	casa	veraneio	1	S	SP	1	40000	1

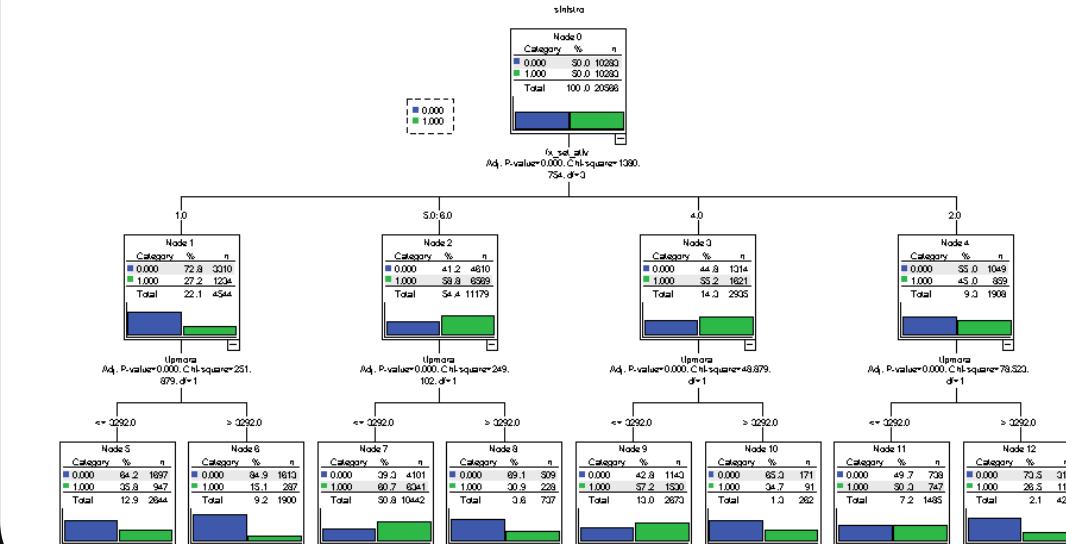
# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: ÁRVORES DE DECISÃO



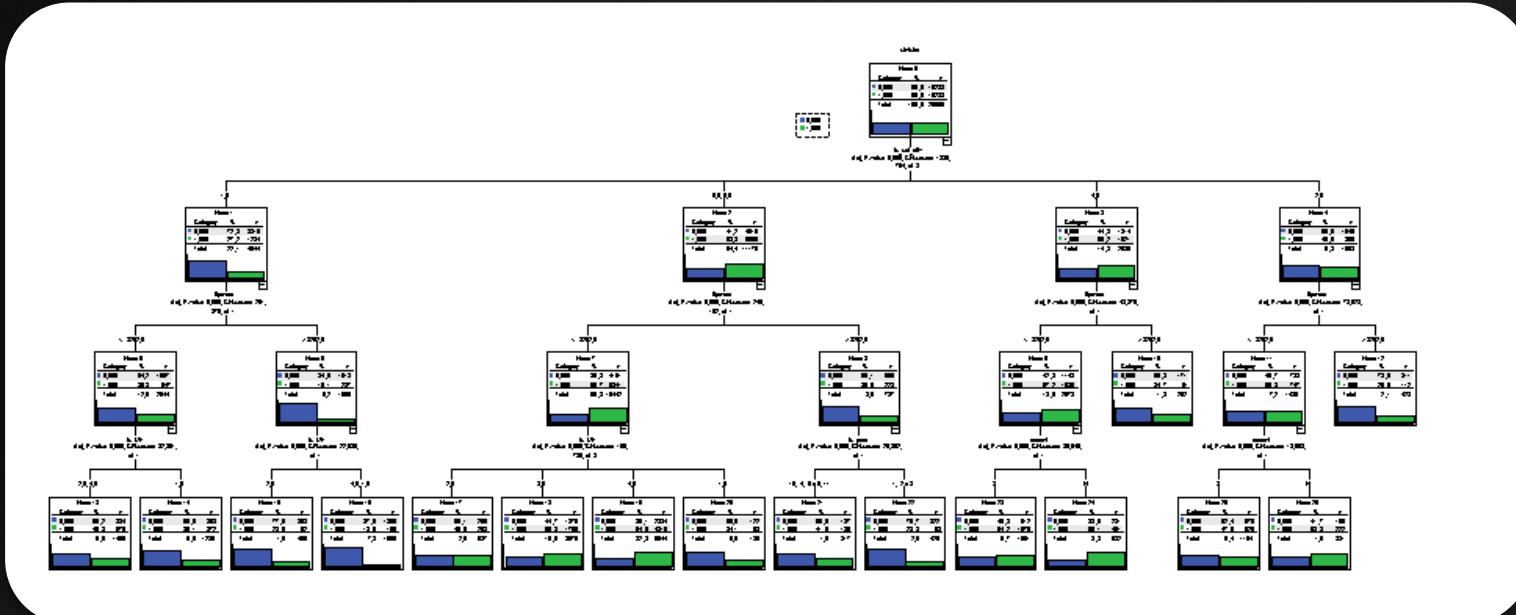
# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: ÁRVORES DE DECISÃO

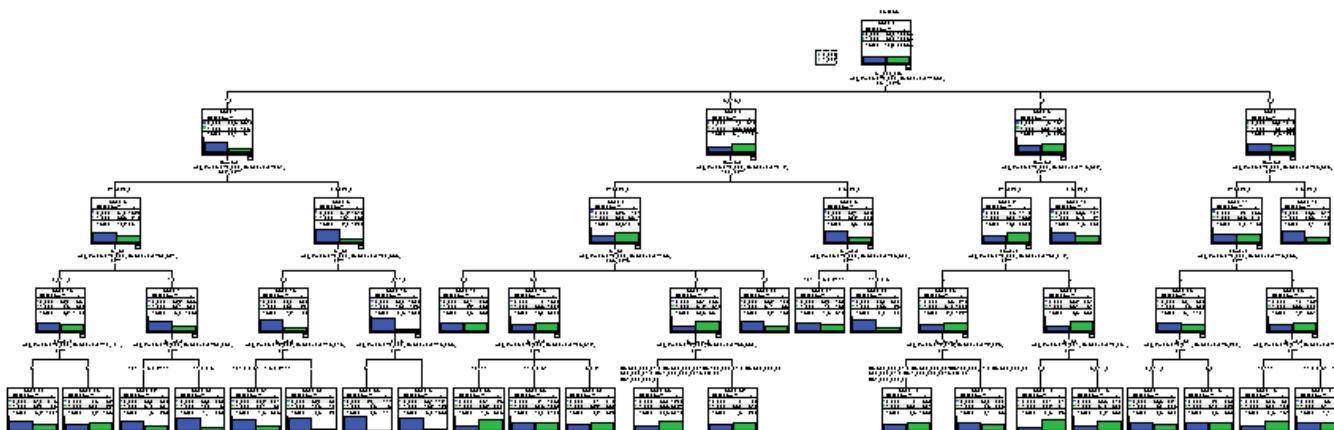


# MODELOS ESTATÍSTICOS

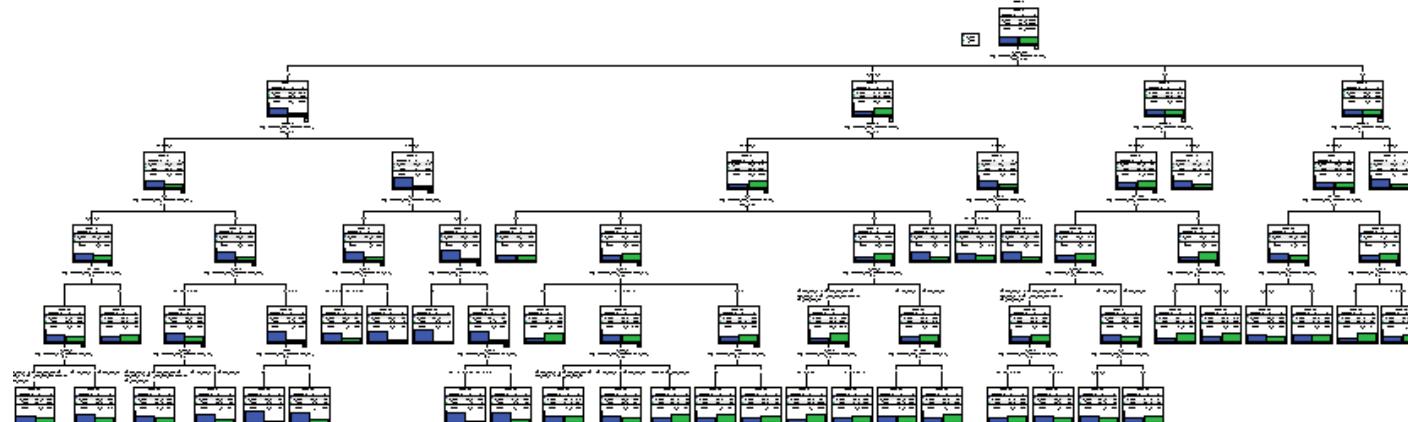
## DISCRIMINAÇÃO: ÁRVORES DE DECISÃO



# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: ÁRVORES DE DECISÃO



# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: ÁRVORES DE DECISÃO



# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: ÁRVORES DE DECISÃO

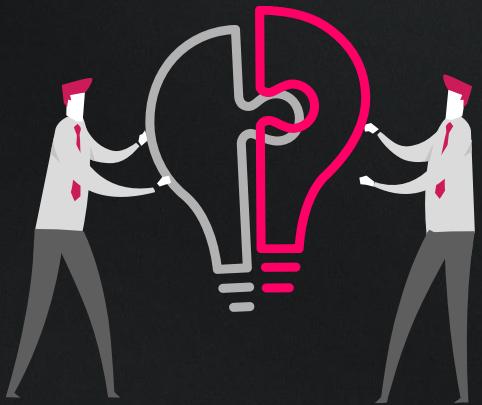
## Avaliação de Modelo

### Classification

Observed	Predicted		Percent Correct
	0	1	
0	<b>8.274</b>	2.009	80,5%
1	2.942	<b>7.341</b>	71,4%
Overall Percentage	73,8%	78,5%	76,1%

Dependent Variable: Evento Estudado

# Modelos Preditivos



Base  
Kaggle 

## MODELOS ESTATÍSTICOS **DESAFIO DO KAGGLE**

Por que nossos melhores e mais experientes funcionários deixaram prematuramente? Divirta-se com este banco de dados e tente prever quais empregados valiosos irão depois.

Os campos no conjunto de dados incluem:

- Nível de satisfação;
- Última avaliação;
- Número de projetos;
- Horas mensais médias;
- Tempo gasto na empresa;
- Se eles tiveram um acidente de trabalho;
- Se eles tiveram uma promoção nos últimos cinco anos;
- Departamentos;
- Salário;
- Se o funcionário deixou.

# MODELOS ESTATÍSTICOS

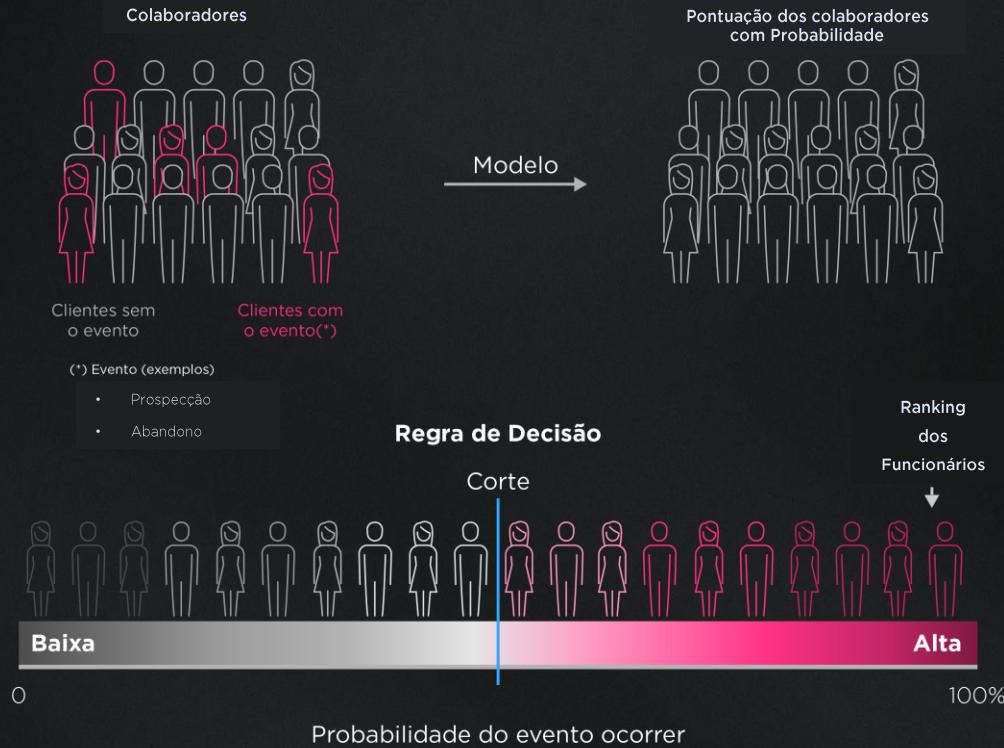
## DESAFIO DO KAGGLE

Satisfaction_level	Last_evaluation	Number_project	Average_montly_hours	time_spend_company	Work_accident	left	Promotion_last_5years	sales	salary
.380	.530	2	157	3	0	1	0	sales	low
.800	.860	5	262	6	0	1	0	sales	medium
.110	.880	7	272	4	0	1	0	sales	medium
.720	.870	5	223	5	0	1	0	sales	low
.370	.520	2	159	3	0	1	0	sales	low
.410	.500	2	153	3	0	1	0	sales	low
.110	.940	6	286	4	0	1	0	IT	medium
.810	.700	6	161	4	0	1	0	IT	medium
.430	.540	2	153	3	0	1	0	product_mng	medium
.830	.950	4	251	5	0	1	0	marketing	medium
.450	.570	2	148	3	0	1	0	marketing	high
.430	.510	2	141	3	0	1	0	sales	low
.580	.750	4	186	2	0	0	0	product_mng	low
.760	.500	3	258	3	0	0	0	IT	low
.500	.780	3	228	2	0	0	0	RandD	low

## MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: **REGRESSÃO LOGÍSTICA**

Encontrar uma **função logística**, formada por meio de ponderações das variáveis (atributos), cuja resposta permita estabelecer a **probabilidade de ocorrência** de determinado evento e a **importância das variáveis** (peso) para essa ocorrência.

# ANÁLISE DE DISCRIMINAÇÃO DE ESTRUTURA **REGRESSÃO LÓGICA**



# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Probabilidade

Sendo Y: a resposta à preferência por um evento (sim ou não),

a probabilidade de:

- Preferência (ou sucesso) será p
- Não-preferência (de fracasso) será (1-p)

“Chance de Ocorrência de um Evento”

Chance = (probabilidade de sucesso) / (probabilidade de fracasso)

Exemplo, se a probabilidade de sucesso é 0,65:

a chance é igual a:  $p / (1 - p) = p / q = 0,65 / 0,35 = 1,86$

# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

## Modelo de Regressão Logística

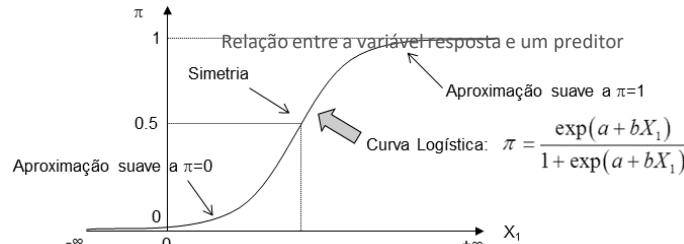
$$G = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

G: logit da resposta de preferência (sim) a :

Intersecção B1, B2, ..., Bn : coeficientes logísticos

- A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{\exp(x)}{1 + \exp(x)}$$



# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo  
Medidas de Avaliação

Previsão do modelo

		Total	
		Y = 1	Y = 0
Obs:	Y = 1	n1	n2
	Y = 0	n3	n4

$$\text{Sensibilidade} = n1 / (n1 + n2)$$

$$\text{Especificidade} = n4 / (n3 + n4)$$

Acurácia: É a proporção de previsões corretas.

É dada por::

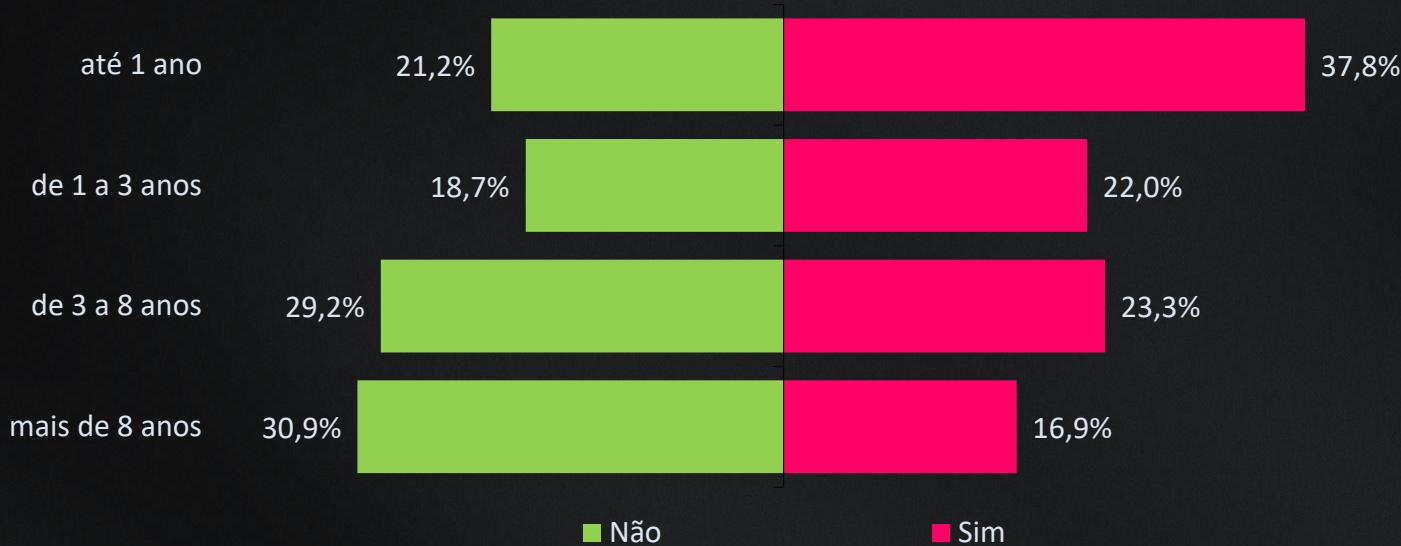
$$\rightarrow (n1+n4) / (n1+n2+n3+n4)$$

## MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: **REGRESSÃO LOGÍSTICA**

A área de RH deseja avaliar a propensão ao risco de seus funcionários pedirem demissão e quer implementar políticas para redução do seu turnover.

## MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

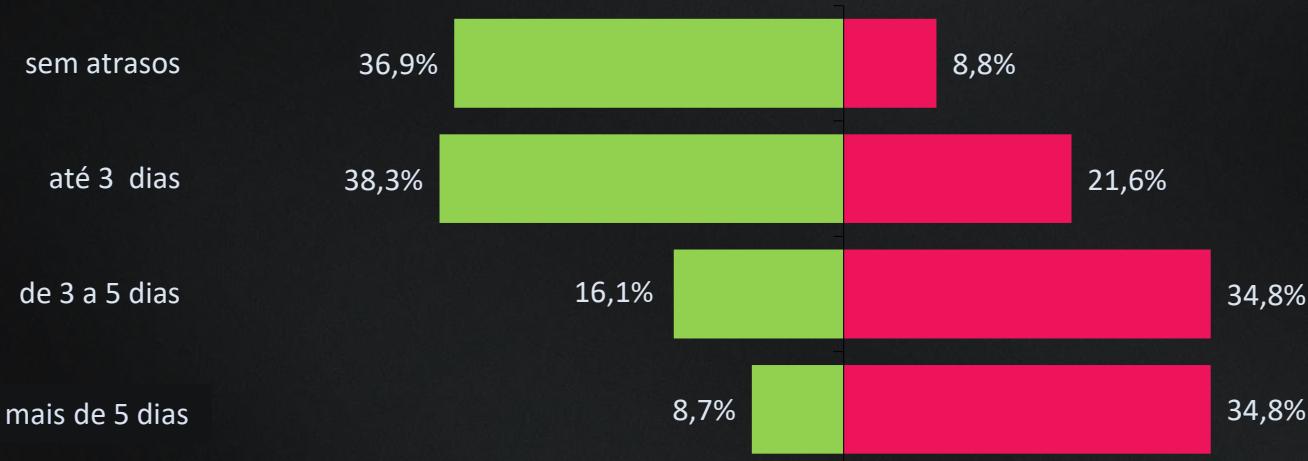
Tempo de contratação em anos



# MODELOS ESTATÍSTICOS

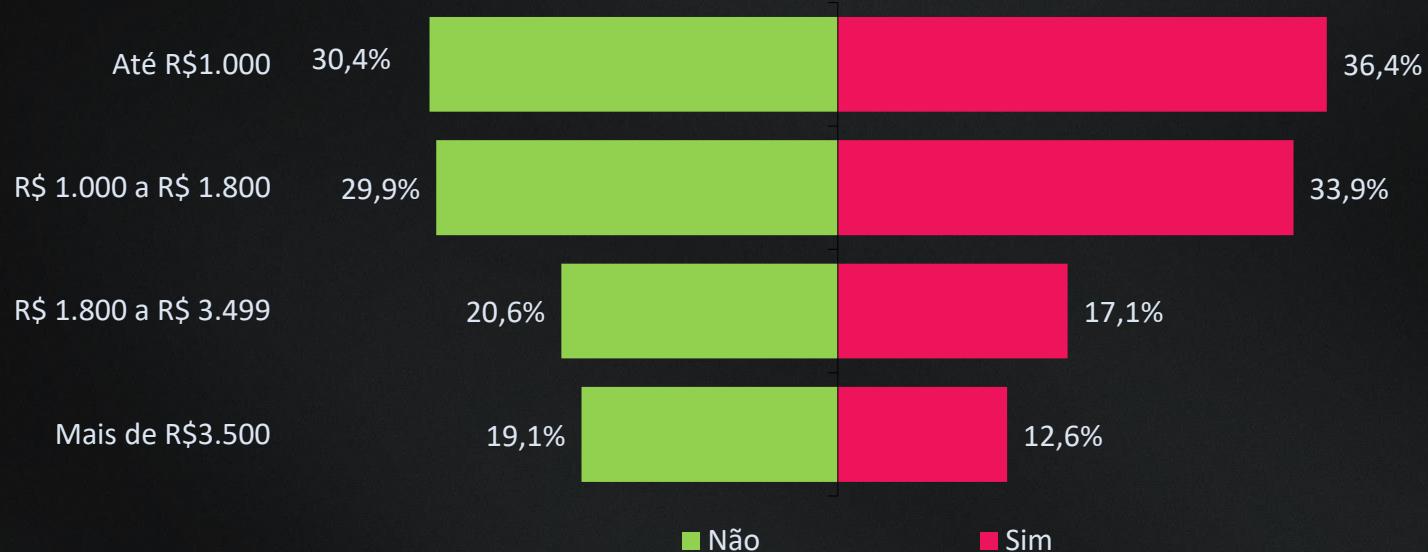
## DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Quantidade de dias com atrasos nos últimos seis meses



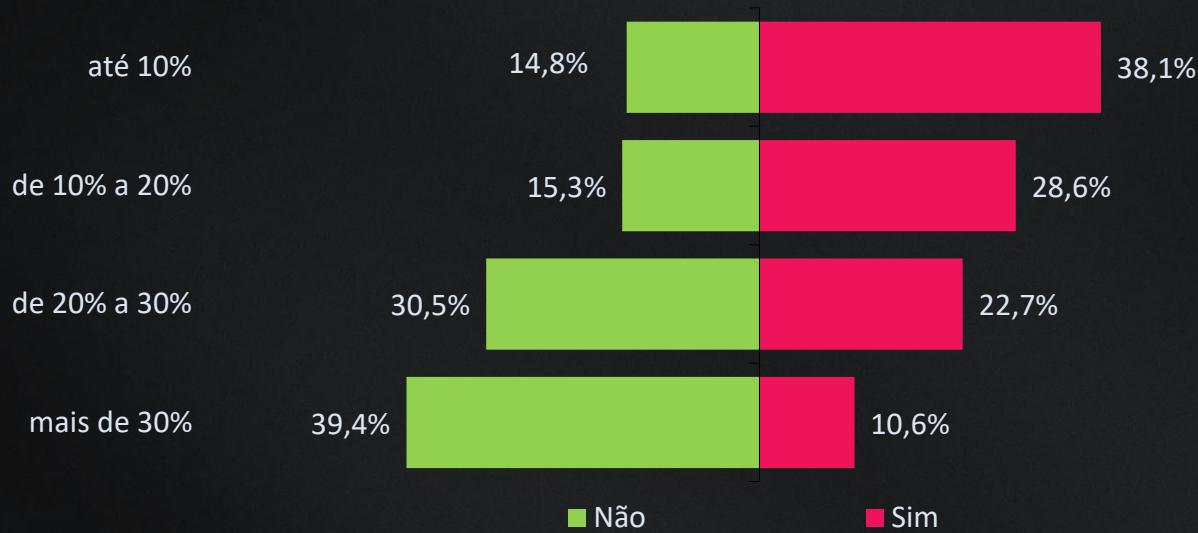
# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Salário R\$



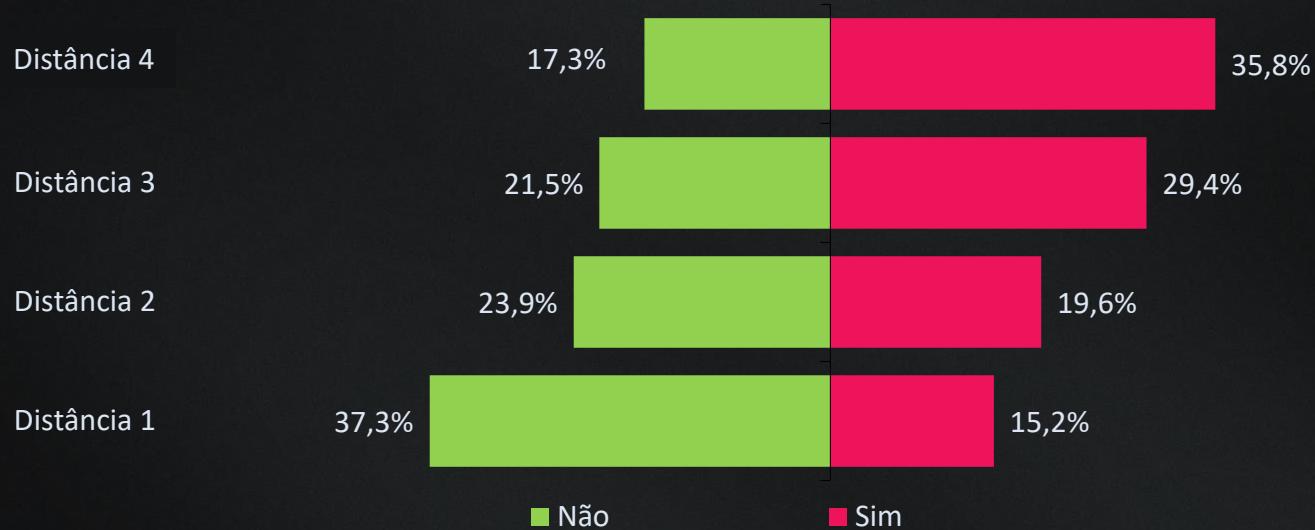
# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Percentual do valor de adiantamento sobre salário



## MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

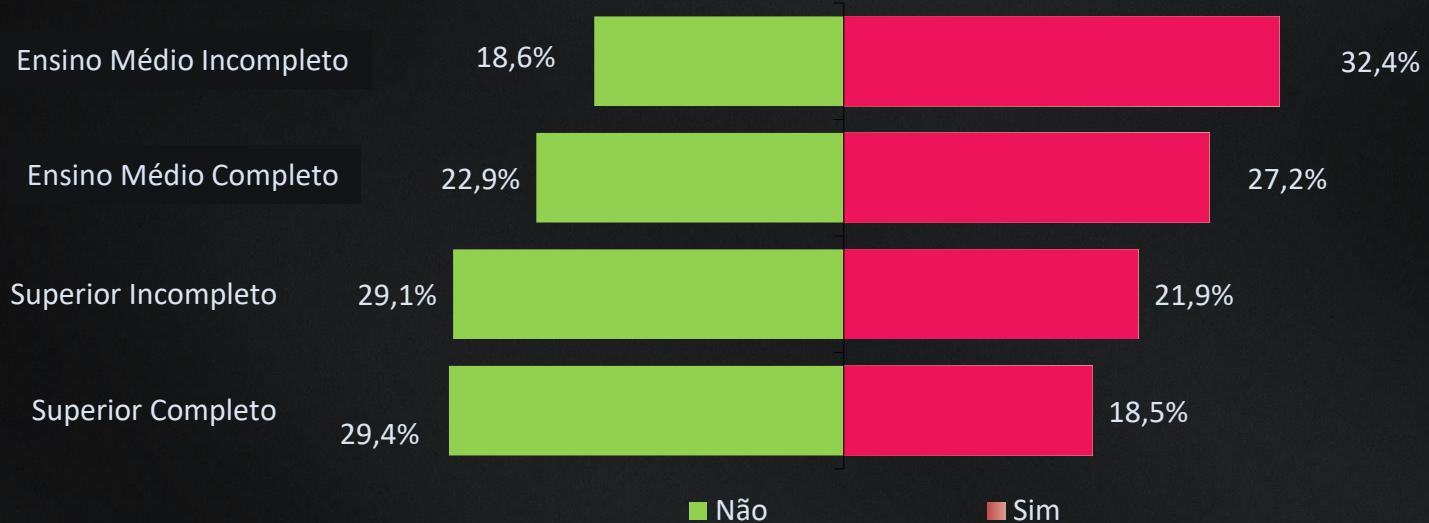
### Distância de moradia



# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

### Grau de escolaridade



# MODELOS ESTATÍSTICOS DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

**Tabela  
de Coeficientes  
do Modelo**

Variável	Categoría	Coeficientes
Média de dias com atrasos os últimos 6 meses	Sem atrasos	-1,276
	Até 3 dias	-0,611
	de 3 a 5 dias	0,580
	mais de 5 dias	1,308
Tempo de contratação em anos	até 1 ano	0,580
	de 1 a 3 anos	0,401
	de 3 a 8 anos	-0,264
	mais de 8 anos	-0,718
Salário	Até R\$1.000	0,262
	R\$ 1.000 a R\$ 1.800	0,103
	R\$ 1.800 a R\$ 3.499	-0,105
	Mais de R\$3.500	-0,261
Percentual do valor de adiantamento sobre salário	até 10%	0,581
	de 10% a 20%	0,401
	de 20% a 30%	-0,264
	mais de 30%	-0,718
Distância de moradia	Distância 4	1,067
	Distância 3	0,371
	Distância 2	-0,368
	Distância 1	-1,069
Grau de Escolaridade	Ensino Médio Incompleto	0,455
	Ensino Médio Completo	0,080
	Superior Incompleto	-0,122
	Superior Completo	-0,413
Constante		0,099

# MODELOS ESTATÍSTICOS

## DISCRIMINAÇÃO: REGRESSÃO LOGÍSTICA

Pesos definidos na modelagem (valores extremos)

-1,276	Sem atrasos	Qtde de dias com atrasos nos últimos seis meses	Mais de 5 dias	1,308
-0,718	Mais de 8 anos	Tempo de contratação em anos	Até 1 ano	0,580
-0,261	Mais de R\$3.500	Salário	Até R\$1.000	0,262
-0,718	Mais de 30%	% do valor de adiantamento	Até 10%	0,580
-1,069	Distância 1	Distância de residência	Distância 4	1,067
-0,413	Superior Completo	Grau de escolaridade	2.Grau Incompleto	0,455
0,099		Constante		0,099
4%	Propensão			98%

# DATA ANALYTICS

## MODELOS ESTATÍSTICOS

TÉCNICAS NÃO SUPERVISIONADAS - ANÁLISE DE AGRUPAMENTOS  
**(ANÁLISE DE CLUSTERS)**

# ANÁLISE ESTRUTURAL: CLUSTER ANALYSIS

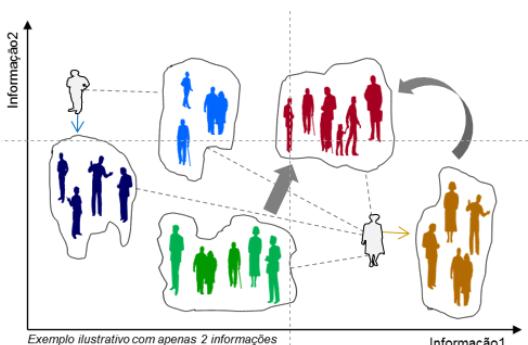
## Segmentação

Universo de Pessoas



Informações Comportamentais, por exemplo:  
Desenvolvimento do modelo estatístico de Segmentação de Pessoas

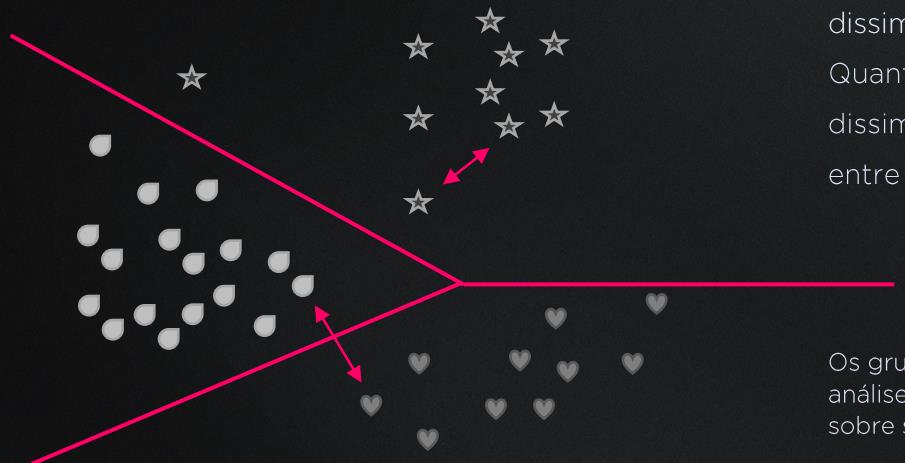
A Segmentação identifica os principais padrões de comportamento e permite que as pessoas sejam organizados em grupos.



- O Perfil dos Segmentos é dado pelas características médias das pessoas que o compõem.
- As ações e comunicações ficam mais interessantes quando são direcionadas pelos diferentes perfis.
- Com o tempo, as pessoas podem mudar de comportamento ou intensidade, logo, podem mudar de Segmento.
- É possível entender como ocorre a migração do cliente entre os segmentos, gerando ações de incentivo ou retenção.
- Os novos funcionários podem ser classificados nos Segmentos.

# ANÁLISE ESTRUTURAL: CLUSTER ANALYSIS

Distância Intra-Cluster Distância entre elementos de um mesmo grupo.



Distância Inter-Cluster Distância entre elementos de grupos distintos

A maioria dos algoritmos de análise de agrupamento tem como base medidas de dissimilaridade:

Quanto **MAIOR** for a medida de dissimilaridade **menor** será a semelhança entre os indivíduos.

Os grupos são “naturais”, isto é, surgem a partir da análise dos dados. Não existe suposição prévia sobre sua estrutura ou o número de grupos.

A decisão sobre o número de grupos depende de bom senso, embora existam critérios que dão suporte à tomada de decisão.

# ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

## Tipos de Segmentação

- Comportamento quanto aos indicadores
- Geodemográficos
- Valores, Hábitos e Atitudes das Pessoas

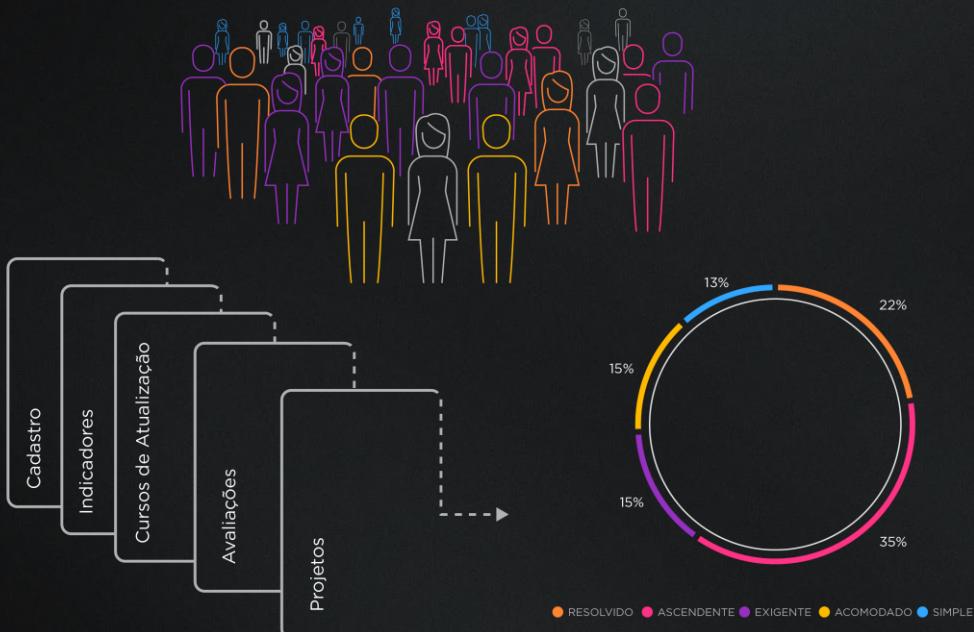
**Comportamental**

**Descritiva**

**Atitudinal**

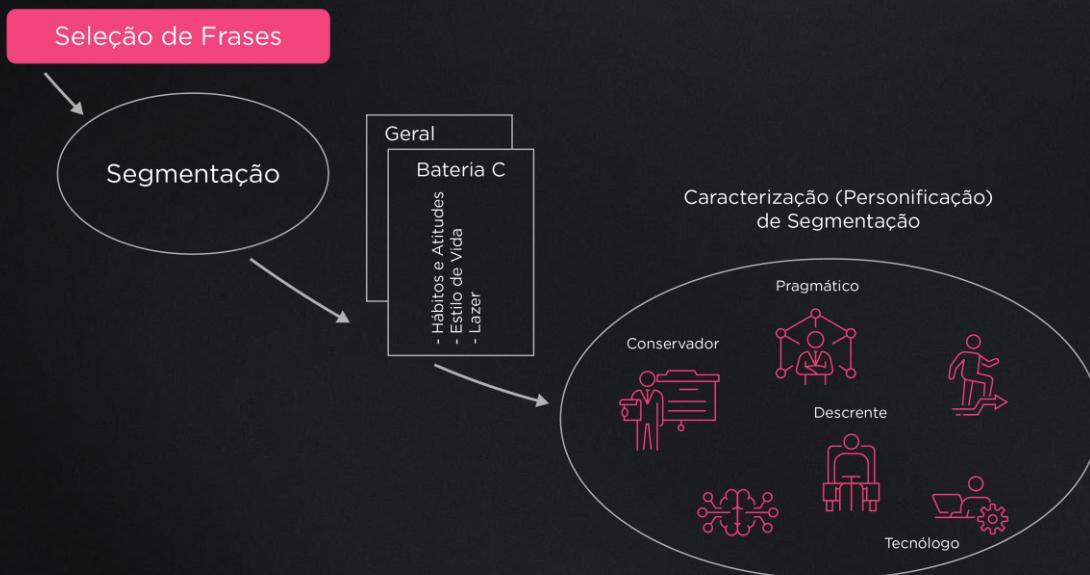
# ANÁLISE ESTRUTURAL: CLUSTER ANALYSIS

## Segmentação de Funcionários



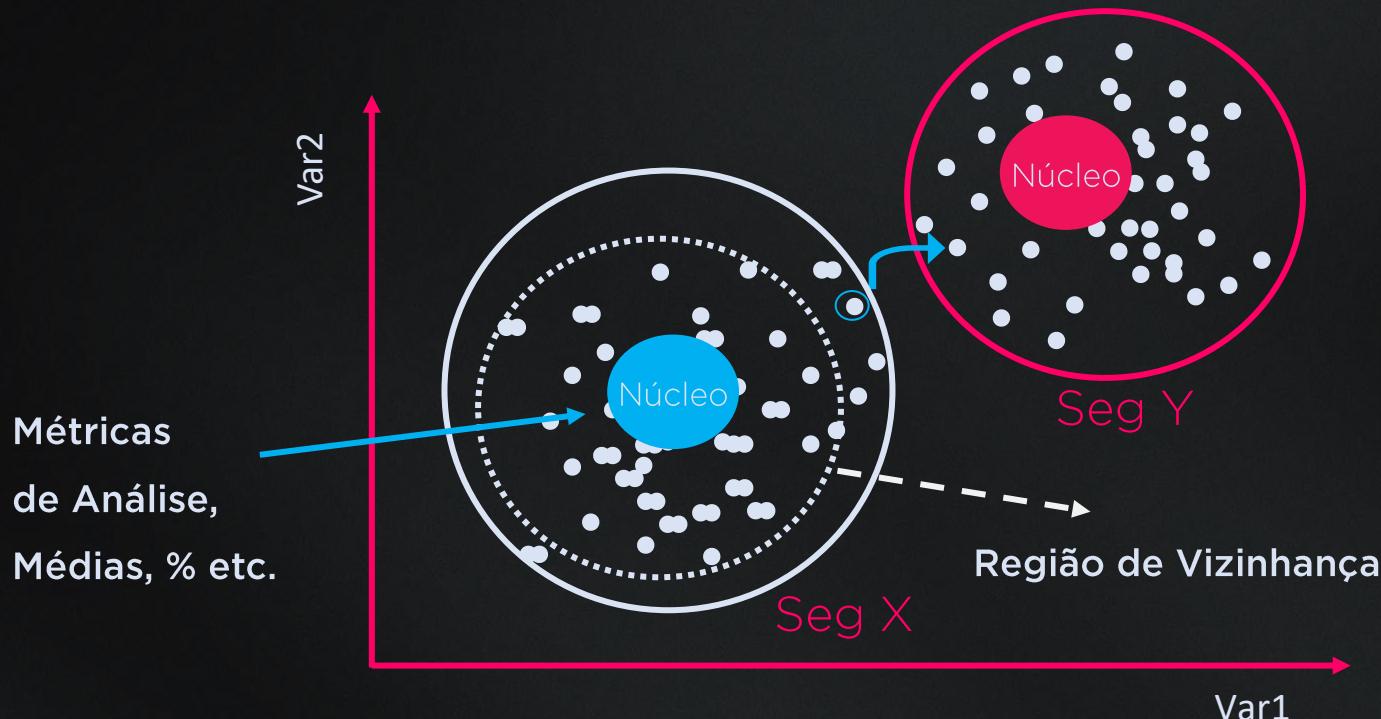
# ANÁLISE ESTRUTURAL: CLUSTER ANALYSIS

## Segmentação Atitudinal



# ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

## Modelos de Segmentação



# ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

## Elementos da Análise

### Entidades

Funcionários,  
Departamentos  
Etc.

### Atributos

Variáveis Discriminantes



# ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

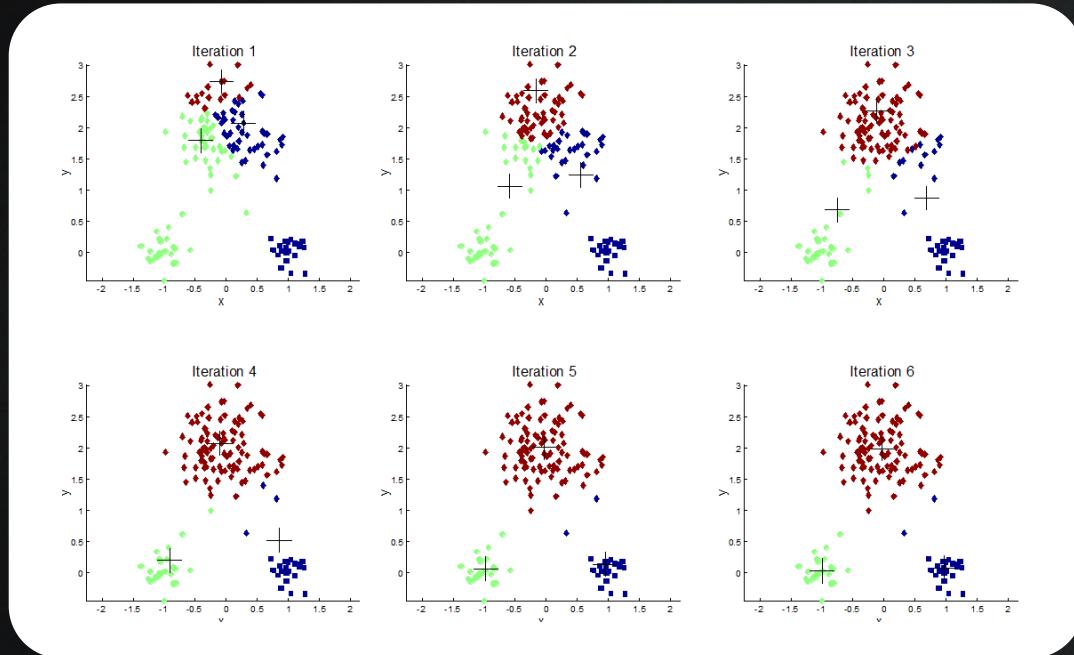
## Análise de conglomerados

### Cluster Analysis

**K-Means** - Uso intenso para grande volume de dados

- Parte de k sementes ou k clusters iniciais sobre os quais calcula as médias.
- Associa um item à semente/ média mais próxima (usando, por exemplo, a Distância Euclidiana). Recalcula a média desse novo cluster e repete iterativamente esta etapa até que não haja mais realocação de elementos.

# CLUSTER ANALYSIS - K-MEANS



# ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

## Medidas de distância

Por exemplo, a distância Euclidiana é calculada por:

$$d_{ij} = \sqrt{\left[ \sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]}$$

Em que  $x_{ik}$  é o valor da variável  $X_k$  para o indivíduo (registro)  $i$  e  $x_{jk}$  é o valor da mesma variável para o indivíduo  $j$ .

Usualmente, as variáveis são padronizadas antes de se calcular as distâncias, assim, as  $p$  variáveis serão igualmente importantes. Geralmente, a padronização feita é para que todas as variáveis (quantitativas) tenham média zero e variância 1.

## ANÁLISE ESTRUTURAL: **CLUSTER ANALYSIS**

### Padronização das variáveis :

Os métodos baseados em distância são afetados pela diferença de escala entre os valores das variáveis/atributos, sendo necessário normalizar os atributos.

**Padronização** - Transforma os valores em números de desvios padrões a partir da média.  
É dada por:

$$z = \frac{x - \bar{x}}{s}$$

Onde :  $\bar{x}$  = Média da variável  
 $s$  = desvio padrão

# ANÁLISE ESTRUTURAL: CLUSTER ANALYSIS

Base Original

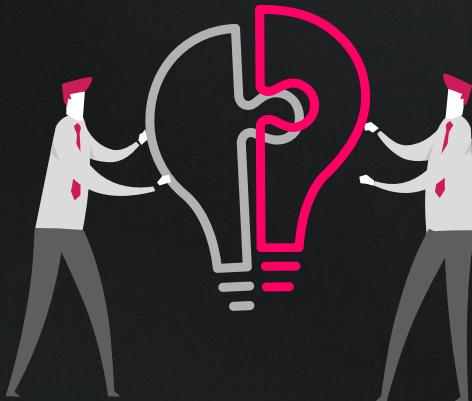
id	Salário	Idade
1	16.284	47
2	3.500	22
3	13.751	24
4	4.751	24
5	6.751	25
6	8.750	26

Base com Padronização da Variáveis

id	Salário	idade
1	1,64	1,71
2	-1,07	-0,97
3	1,10	-0,76
4	-0,80	-0,76
5	-0,38	-0,65
6	0,04	-0,54



## CLUSTER ANALYSIS - K-MEANS



Base de  
Indicadores



# OBRIGADA

 /regina.cantele

 /adelaide.alves

FIAP

Copyright © 2020 | Professoras Regina Cantele e Adelaide Alves

Todos os direitos reservados. A reprodução ou divulgação total ou parcial deste documento é expressamente proibida sem o consentimento formal, por escrito, do professor/autor.

