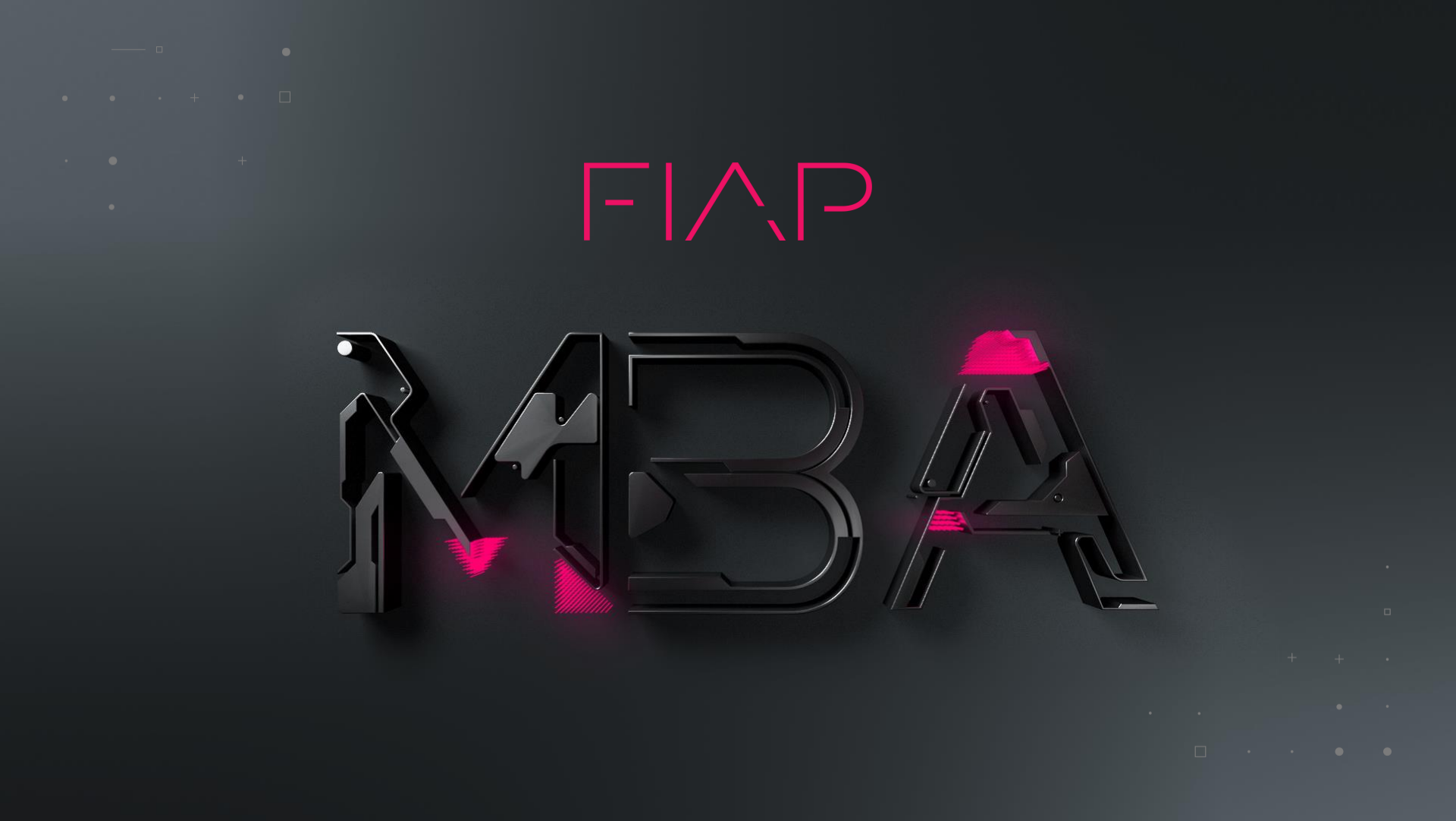

FIAP
MBA

ADELAIDE ALVES DE OLIVEIRA

PROFESSORA

profadelaide.alves@fiap.com.br

Formação Acadêmica

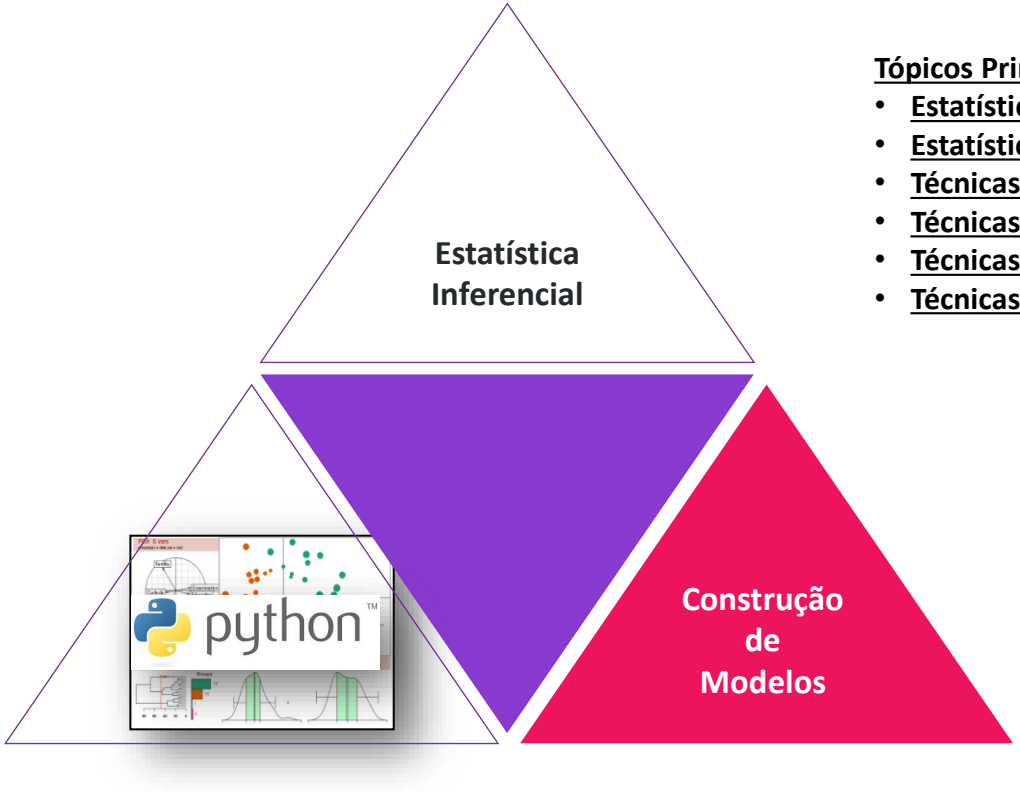
- Bacharel em Estatística – UNICAMP
- Mestre em Ciências – FSP/USP

Atividades Profissionais

- Diretora Técnica Estatística da empresa **SD&W - www.sdw.com.br**
- Professora na **FIAP** de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning nos cursos MBA: Big Data, Data Science, Business Intelligence & Analytics, Digital Data Marketing, AI & ML e Engenharia de Dados e nos Shift: People Analytics e Python Journey .

NOSSA JORNADA!!!

APPLIED STATISTICS

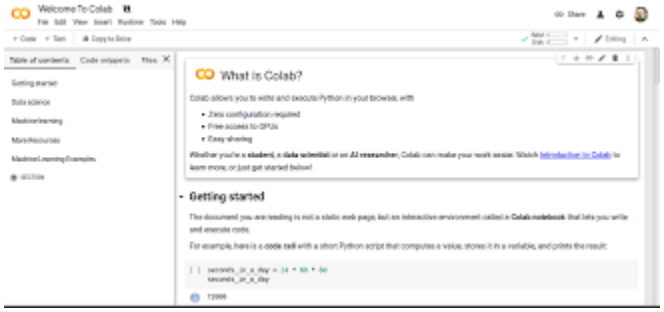


Tópicos Principais

- Estatística Descritiva e Amostragem
- Estatística Inferência Estatística
- Técnicas de Associação e Correlação
- Técnicas de Estimação
- Técnicas de Classificação
- Técnicas de Agrupamento

APPLIED STATISTICS

Formas de acesso



APPLIED STATISTICS

Avaliação	PESO
Exercícios individuais	30%
Projeto Integrado em grupo de até 4 alunos	Fase 1 20% Fase 2 50%

PROJETO INTEGRADO

Vamos começar !!!!!!!!!!!!

INTRODUÇÃO

INSIGHTS

#BUSINESS 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.

o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o insight e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o

o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o insight e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o
ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído
e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruído e o ruíd

INTRODUÇÃO

“ O Índice de Churn da nossa empresa está em 12%”

“A taxa de conversão de leads é de 4%”

DADOS OU INFORMAÇÕES?

INTRODUÇÃO

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

Conhecimento

Informação Estruturada e Contextualizada

Conjunto formado por experiências, valores e criatividade aplicado aos dados e informações

Informação

Dado com significado

Dados analisados com ajuda de ferramentas específicas, de modo a que sejam úteis para a tomada de decisão

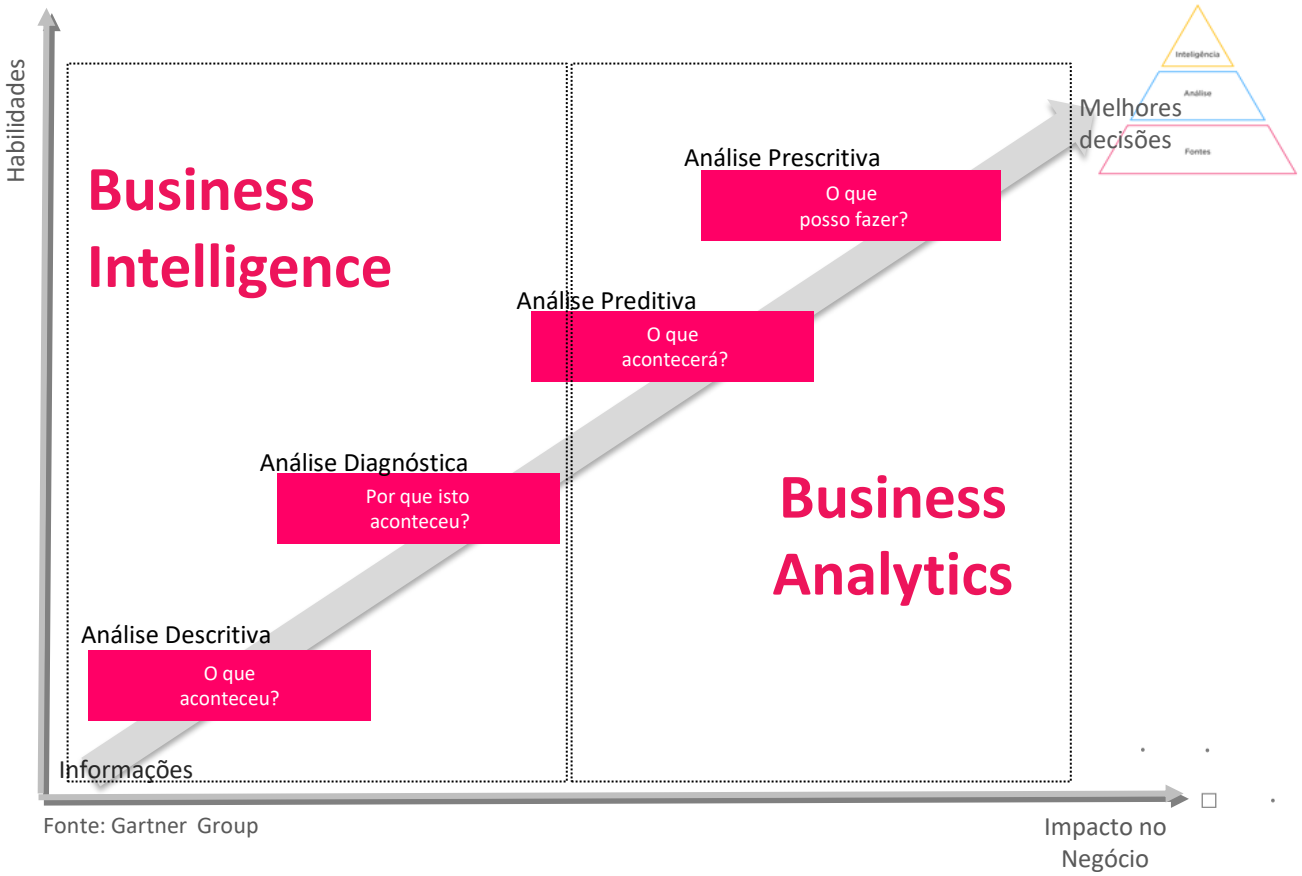
Dado

Valor sem significado

Fatos, números e estatísticas

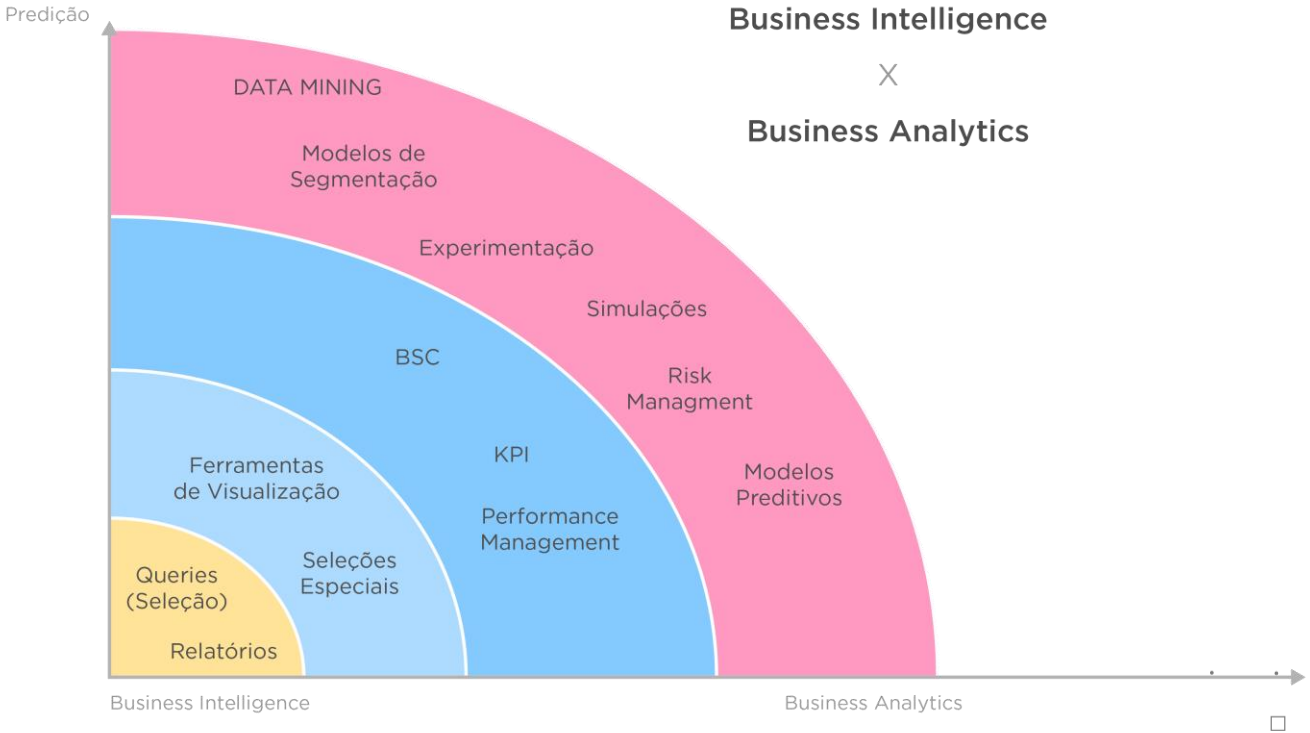
Ajudar o gestor a diminuir os riscos e aumentar as chances de sucesso!

INTRODUÇÃO



Fonte: Gartner Group

INTRODUÇÃO

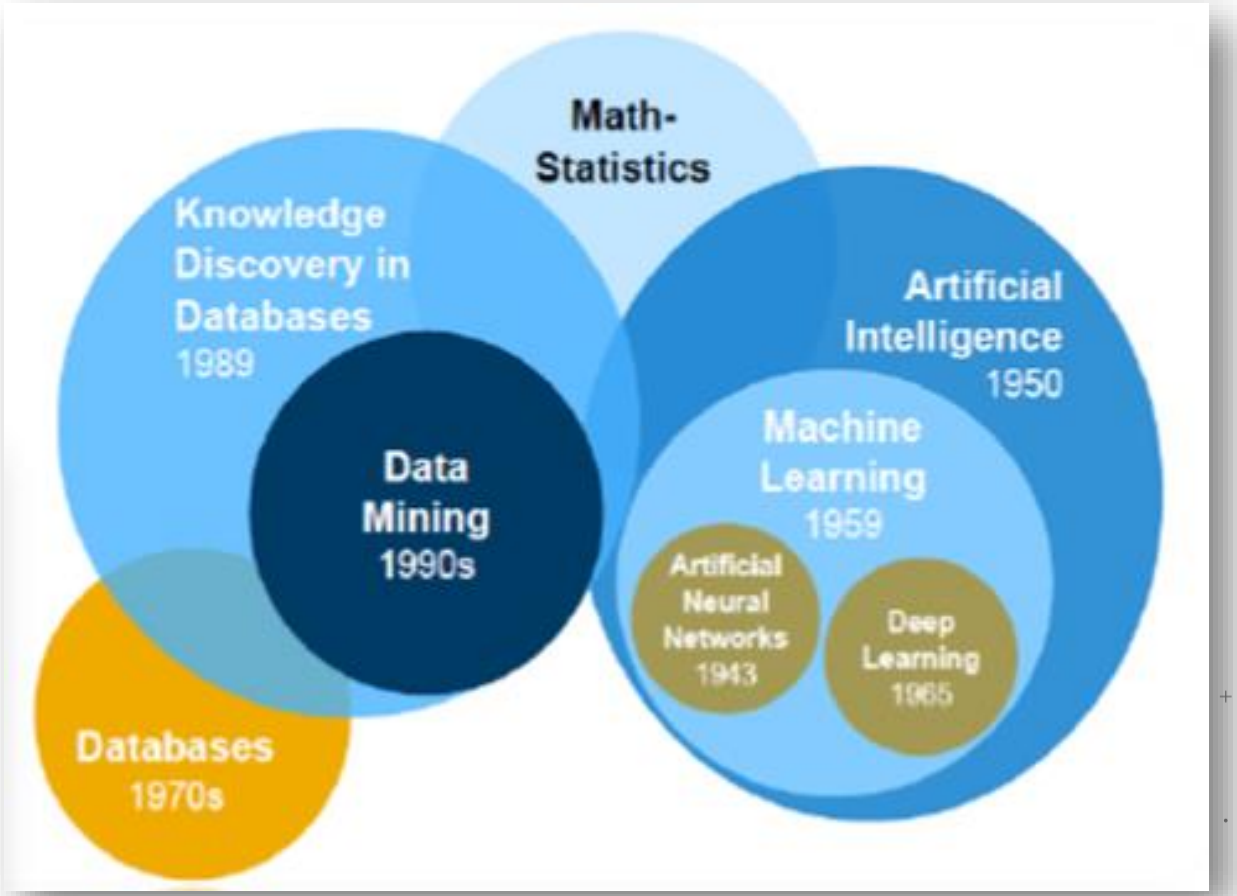

Predição
Business Intelligence
X
Business Analytics
DATA MINING
Modelos de Segmentação
Experimentação
Simulações
BSC
Risk Managment
KPI
Ferramentas de Visualização
Performance Management
Modelos Preditivos
Seleções Especiais
Queries (Seleção)
Relatórios
Business Intelligence
Business Analytics

INTRODUÇÃO

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

O que aconteceu?	DESCRITIVO	Quantos cancelamentos? Quantos clientes novos, quantos antigos? Qual região? Que tipo de cliente?
Por que isto aconteceu?	DIAGNÓSTICO	Qual a relação entre cancelamento voluntário por tempo de conta e tipo de cliente?
O que acontecerá?	PREDITIVO	Qual a probabilidade de um cliente cancelar o serviço para em uma certa região nos próximos 3 meses?
O que posso fazer?	PRESCRITIVO	Lista de ações para reter os clientes, segundo seu valor?

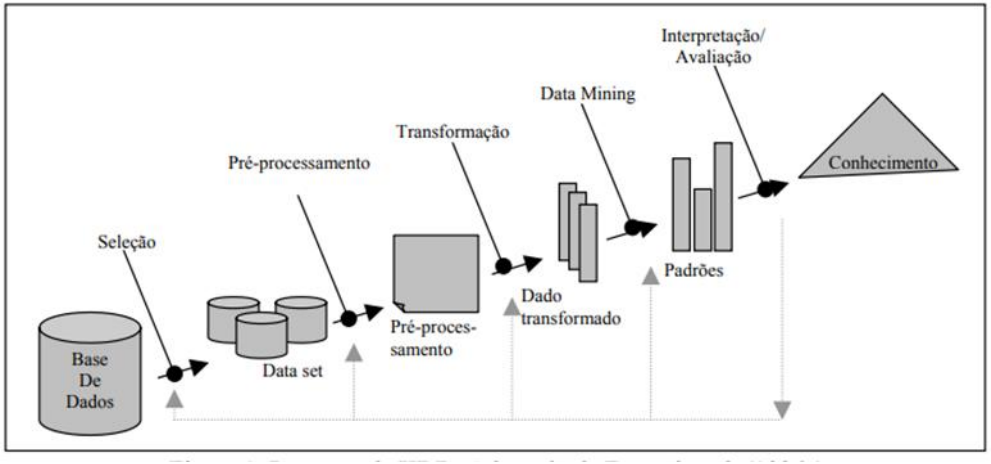
INTRODUÇÃO



- SPSS: 1968
- SAS: 1976
- R: ~1990
- Python: ~1990

PROCESSO KDD

KNOWLEDGE DISCOVERY IN DATABASES

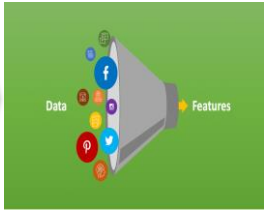


Fonte: Processo de KDD. Adaptado de Fayyad et al. (1996a).

CICLO ANALÍTICO

Entender o problema de Negócio

Coletar DADOS



Explorar/ Visualizar

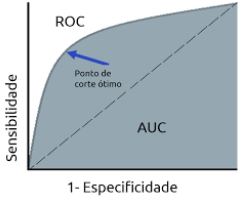
Feature Engineering

Preparar os dados

FeatureExtraction/ Selection

Machine Learning

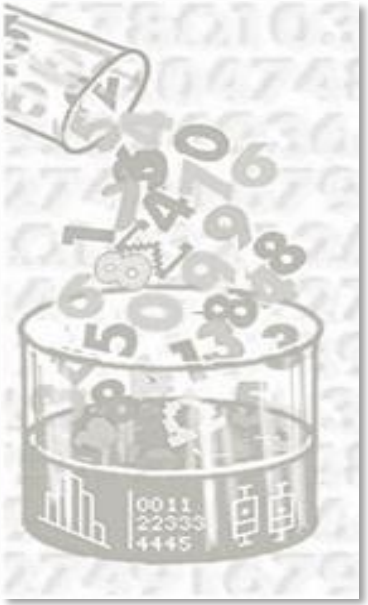
Validação / Monitoramento



Deploy / Implementar


DATA ANALYTICS

ESTATÍSTICA



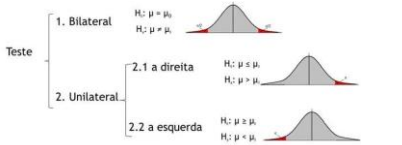
“Estatística é a Ciência que permite obter conclusões a partir de dados (Paul Velleman*)

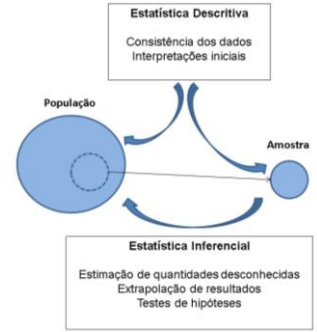
- É a ciência que trata dados provenientes de mensuração em grupos de indivíduos.
- Trata da organização, descrição, apresentação, análise e interpretação de dados resultantes da observação de fenômenos coletivos. Produz métodos para inferência estatística.

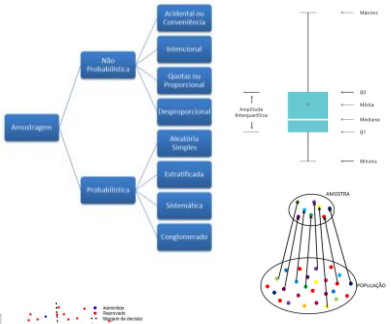
Propriedades:

Estuda as variações:

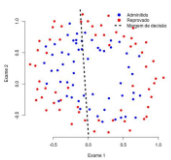
- Entre indivíduos;
- Em um mesmo indivíduo.

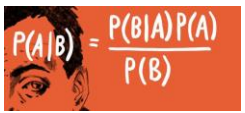






Classificação da Variável	Valor Numérico
Muito Bom	1,10
Bom	1,05
Razoável	1,00
Fraco	0,95
Mau	0,90





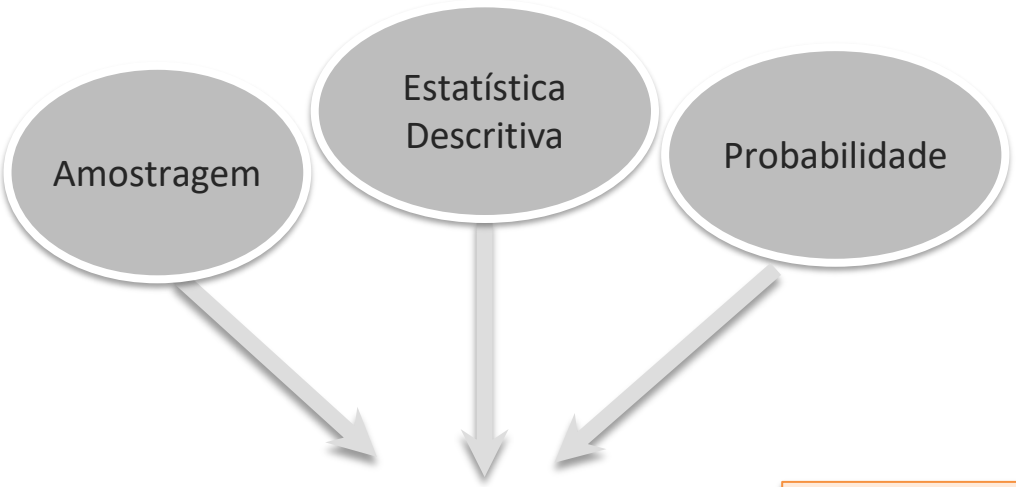
*Paul F. Velleman (born April 16, 1949) is a Professor of statistics at Cornell University

INTRODUÇÃO

EXEMPLO

O Market Share de um certo produto: 65%

INTRODUÇÃO



Estatística Inferencial

- Estimação de parâmetros
- Confiabilidade da informação
- Nível de significância (erro / precisão)

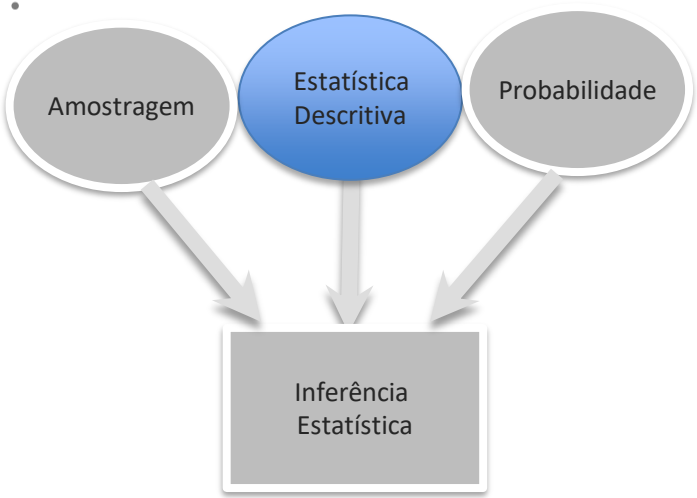
Exemplo:
Estimativa de share – 65%
Std Error - 4%
95% de confiança
IC(95%) = 57.2% ---- 72.8%
Margem de erro ± 7.8%

➔**Analisar e Interpretar os Dados**

Tem por objetivo chegar a conclusões sobre a população com base nos resultados obtidos em "amostras" extraídas dessa população.

A Estatística Inferencial nos diz em que ponto poderemos estar errando e com que probabilidade.

INTRODUÇÃO



Estatística Descritiva

➔ Organizar e Descrever os Dados

É utilizada para resumir as informações de dados obtidas a partir de uma pesquisa, examinar a estrutura subjacente dos dados e aprender sobre os relacionamentos sistemáticos entre muitas variáveis.

Inclui um conjunto de ferramentas gráficas e descritivas, para explorar os dados, como pré-requisito outras análises como: Estimação, Testes de Hipóteses e construção de Modelos.

Também chamada de

Análise Exploratória dos Dados (AED)

CONCEITOS

População →

Elementos (N=8) →

Variáveis*
(atividade física, sexo, idade, filhos, entre outros) →

QUAIS AS OCORRÊNCIAS POSSÍVEIS PARA ATIVIDADE FÍSICA ?

COMO VOCÊ REPRESENTARIA ESSAS OCORRÊNCIAS ?

*Variáveis/Atributos/Features/Características/Colunas

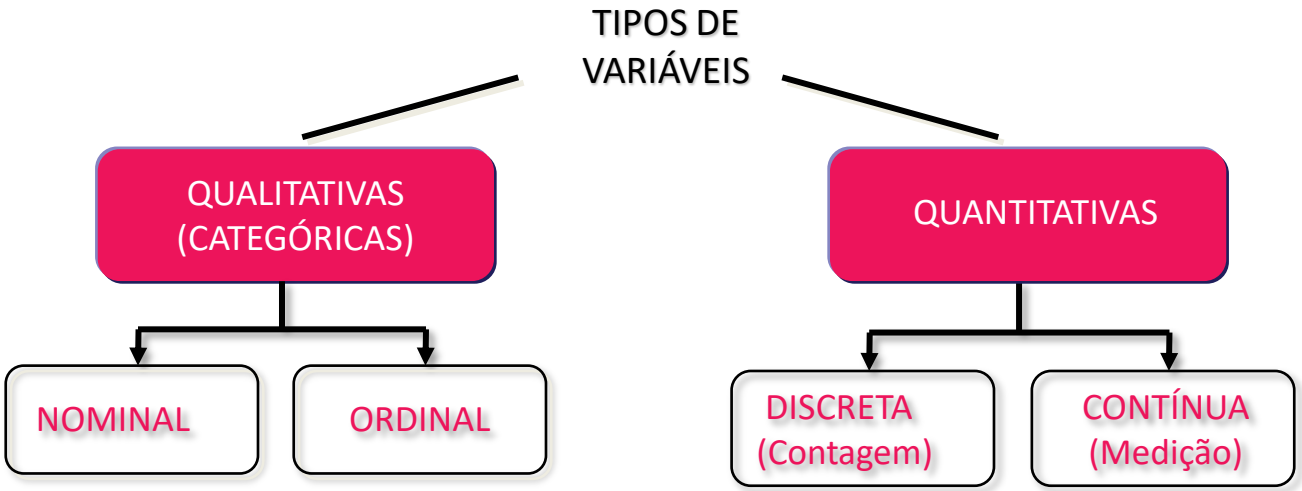
APRESENTAÇÃO DOS DADOS

ARQUIVO

Estrutura matricial: linhas e colunas

ORDEM	SEXO	ATIVIDADE FÍSICA	ESTADO CIVIL	GRUPO
1	F	SIM	CASADA	1
2	M	SIM	SOLTERIO	2
3	F	NÃO	SOLTEIRA	2
4	M	NÃO	CASADO	2
5	F	NÃO	CASADA	2
6	F	NÃO	CASADA	2
7	F	NÃO	SOLTEIRA	3
8	M	NÃO	SOLTEIRO	3

ESCALA DE MENSURAÇÃO



Valores não apresentam ordenação é um atributo ou uma qualidade

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

Valores apresentam uma ordem, uma escala pré-determinada

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

Valores fazem parte de um conjunto finito ou infinito numerável

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

Valores são os números reais

- Salário Mensal,
- Idade,
- Tempo de estudo,
- Taxas.

ESCALA DE MENSURAÇÃO

Exemplo de Tipos de Variáveis

Parte do arquivo da PNAD[1] do município de SP

id	sexo	idade	cor	internet	telefone_movel	anosestudo	Rendimento
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	
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
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	
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

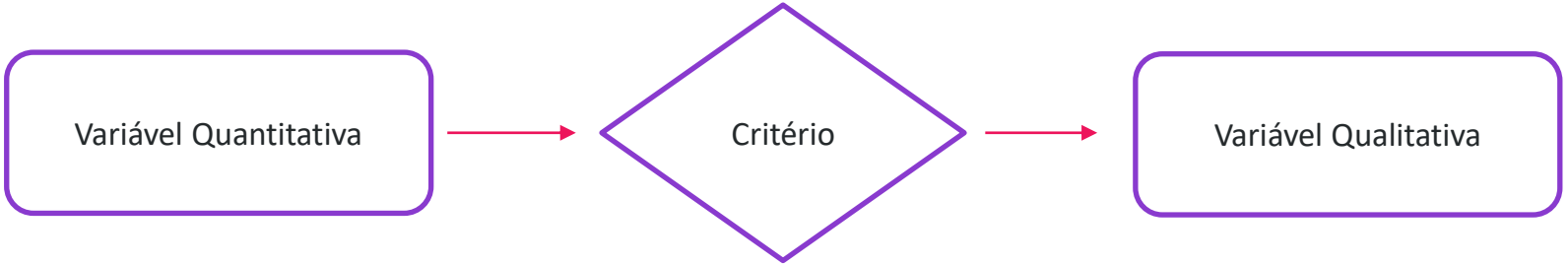
Categóricas | Quantitativas | n = 1380

Exemplo: Internet:
Acesso últimos 3 meses
1 ➔ 1
3 ➔ 0

[1]PNAD: Pesquisa Nacional por Amostra de Domicílios IBGE

ESCALA DE MENSURAÇÃO

Transformando variáveis quantitativas em qualitativas



Anos De Estudos

Critério	Grau De Instrução
0	Analfabeto
[1 - 8]	Fundamental
[9 – 11]	Médio
> = 12	Superior

ESCALA DE MENSURAÇÃO

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 5 segmentos de reta com amplitudes iguais. A amplitude de cada intervalo abaixo representa, portanto,

- amplitude = (90 – 0)/5 = 18 anos.

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

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

APRESENTAÇÃO DOS DADOS

Distribuição de Frequência

O número de vezes 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

Faixa Etária	FA	FR (%)
[0 - 18]	437	437/1380*100
[19 - 36]	384	384/1380*100
[37 - 54]	360	360/1380*100
[55 - 72]	180	180/1380*100
[73 - 90]	41	41/1380*100
Total	1380	1380/1380*100

Faixa Etária	FA	FR (%)
[0 - 18]	437	30,7
[19 - 36]	384	27,8
[37 - 54]	360	26,1
[55 - 72]	180	11,4
[73 - 90]	41	3,0
Total	1380	100,0

APRESENTAÇÃO DOS DADOS

Distribuição de Frequência

- Frequência Absoluta – Valor total das observações
- Frequência Relativa – Valor porcentual das observações
- Frequência Acumulada – Somatória das frequências de todos os intervalos

APRESENTAÇÃO DOS DADOS

Tipos de Variáveis: Escala de questionário:

péssimo ( ) ruim ( ) regular ( ) bom ( ) excelente ( )

Variável Qualitativa Ordinal

Variável Discreta

1 2 3 4 5

certamente não compraria

discordo totalmente

nada satisfeito

certamente compraria

concordo totalmente

totalmente satisfeito

APRESENTAÇÃO DOS DADOS

Tipos de Variáveis: Escala de questionário:

Exemplo: Como medir satisfação do cliente?

Escala de Satisfação

------- 1 ------- 2 ------- 3 ------- 4 ------- 5 ------- 6 ------- 7 ------- 8 ------- 9 ------- 10 -------

1	2	3	4	5
Insatisfeito	Pouco Satisfeito	Nem Satisfeito, Nem Insatisfeito	Satisfeito	Muito Satisfeito

Grau de Satisfação

- 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

APRESENTAÇÃO DOS DADOS

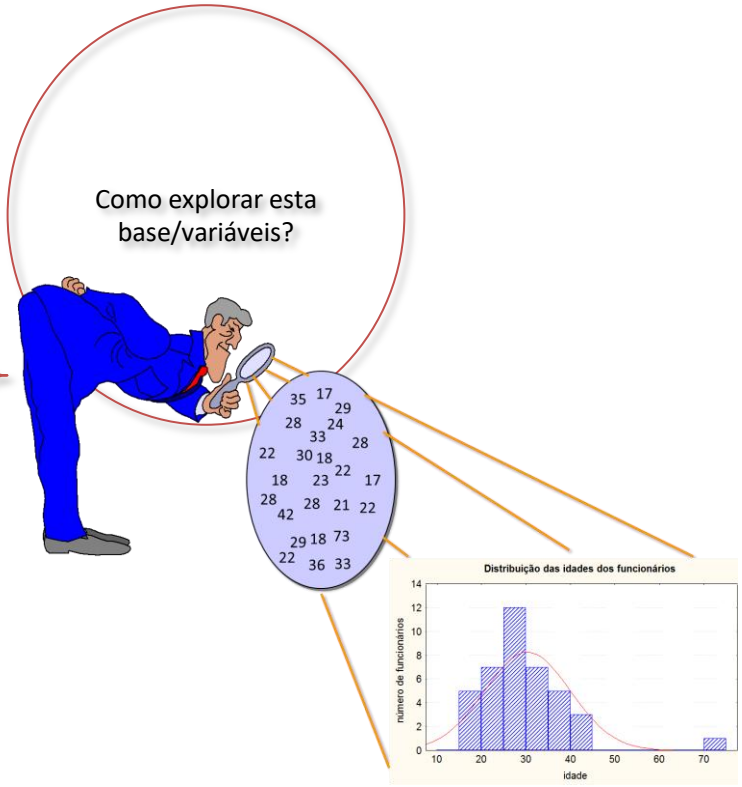
Tipos de Variáveis: Escala de questionário:

Transformando medições por meio de escalas em Indicadores

- Porcentagens de clientes muito insatisfeitos, insatisfeitos, neutros, satisfeitos e muito satisfeitos em relação à média total de clientes.
- Porcentagem de clientes que se dizem dispostos a voltar a comprar na empresa.
- Porcentagem de clientes que se dizem dispostos a recomendar a empresa.
- Porcentagem de clientes que declaram preferir os produtos da empresa.
- Porcentagem de clientes 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 com os dos principais concorrentes.
- NPS = %P - %D P:Promotores (9 e 10) D:Detratores(0 a 6)

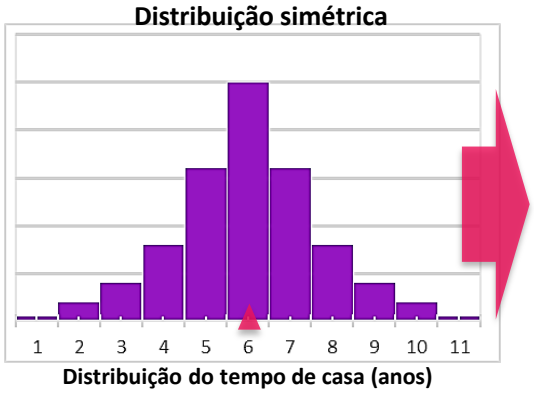
APRESENTAÇÃO DOS DADOS

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	4.858,72
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	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
39	9.072,00
40	3.273,02



APRESENTAÇÃO DOS DADOS

Medidas de Posição e Dispersão

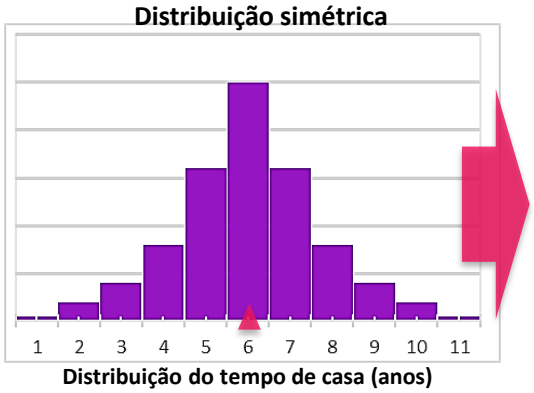


Medidas que resumam as características peculiares do fato em estudo (distribuições):

- seu valor central
- seu grau de dispersão em torno do valor central (variabilidade)
- seu grau de assimetria (forma de distribuição)

APRESENTAÇÃO DOS DADOS

Medidas de Posição e Dispersão



Medidas de tendência central:

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

- Mediana
- Moda
- Média

Medidas de dispersão:

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

- Variância
- Desvio padrão

MEDIDAS DE POSIÇÃO - MÉDIA

- 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): $Mg = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
- Média Quadrática: $\bar{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$
- Média Harmônica: $M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$

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.

MEDIDAS DE POSIÇÃO - MÉDIA

- Média Harmônica: $M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$

  - Exemplo:
    - Física : Velocidade Média / Hodrologia: Bombas / Circuitos Elétrico
    - Finanças: Preço/Ganho
    - Esportes: Beisebol
    - Genética
    - Modelos / MachineLearning – Score para aferir algoritmos

  - É a menor valor entre as 3 médias: Média Harmônica. Média Aritmética e Média Geométrica

    Uma vez que a média harmônica de uma lista de números tende para o mínimo dos elementos da lista, ela tende (em comparação com a média aritmética) a mitigar o impacto de grandes valores atípicos e agravar o impacto das pequenas

MEDIDAS DE POSIÇÃO - MÉDIA

- ´Média Harmônica: $M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$

Dois amigos em viagem revesam-se para chegar a um determinado destino. Um deles dirigiu até exatamente a metade do percurso, e depois o outro assumiu o volante terminando o percurso. O primeiro deles manteve uma velocidade v_1 = 80 km/h. Já o segundo, que estava com mais pressa, manteve uma velocidade de v_2 = 120 km/h.

Aplicando na fórmula com n = 2:

Assim, a média de velocidade nesse percurso foi de 96 km/h.

$$M_h = \frac{2}{\frac{1}{120} + \frac{1}{80}}$$

$$M_h = \frac{2}{\frac{2+3}{240}}$$

$$M_h = \frac{2}{\frac{5}{240}}$$

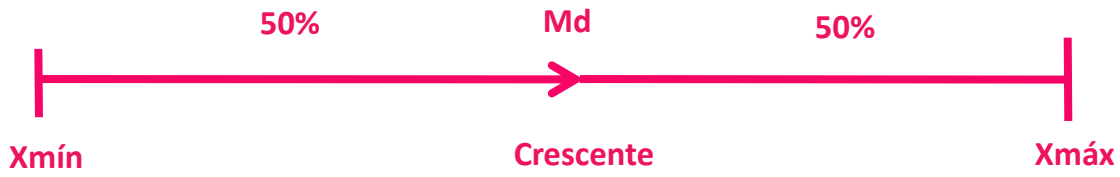
$$M_h = 2 \cdot \frac{240}{5}$$

$$M_h = \frac{480}{5} = 96 km/h$$

Fonte:https://www.preparaenem.com/matematica/media-harmonica.htm

MEDIDAS DE POSIÇÃO - MEDIANA

- A mediana é uma medida 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.
- Os modelos são diferentes para situações onde existem um número ímpar ou par de observações.



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 ➔ (7+1)/2

  ordenado ➔ 2, 4, 11, **17**,18, 26,50

- número par de observações - Para nº 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

MEDIDAS DE POSIÇÃO

- 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 da satisfação média e mediana encontrada?

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

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

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

MEDIDAS DE POSIÇÃO

- 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?

$$\bar{x} = 6{,}00 \qquad mediana = 6{,}01$$

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

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

MEDIDAS DE POSIÇÃO - MODA

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

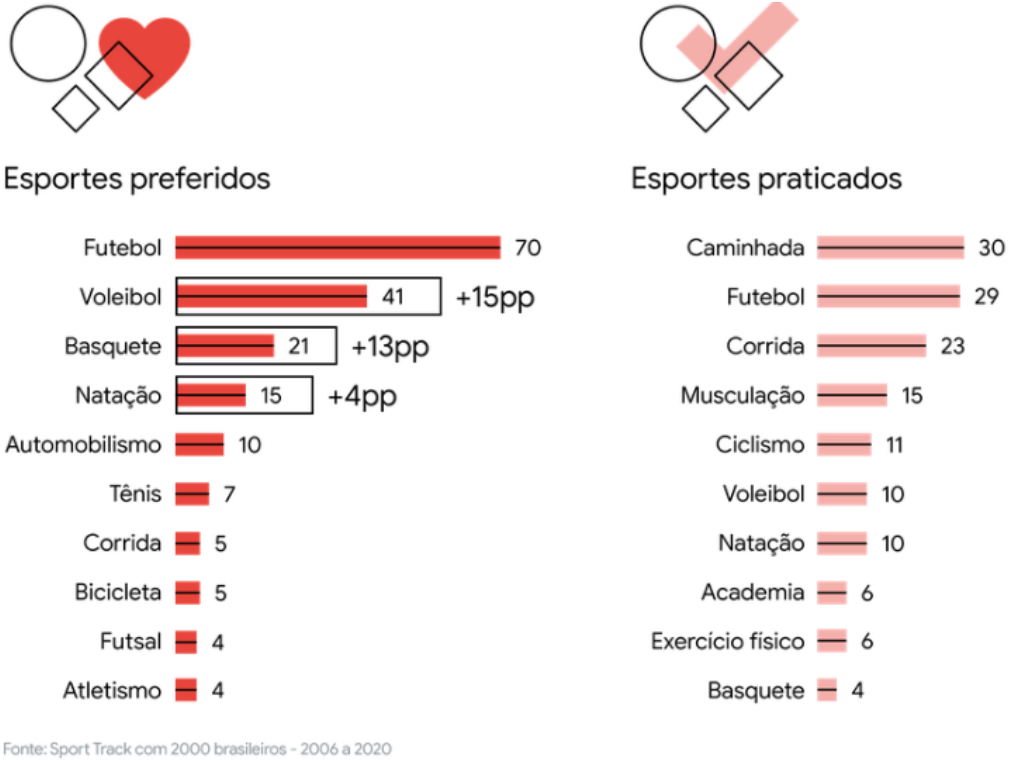
- Exemplo 1: 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 Mo = 9
- Exemplo 2: 3, 5, 8, 10, 12, 15, 16 Mo = Amodal
- Exemplo 3: 2, 3, 4, 4, 4, 55, 555, 7, 7, 7, 9 Mo = 4 e 7 Bimodal

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

MEDIDAS DE POSIÇÃO - MODA

Exemplo

Esportes e atividades físicas que vem ganhando o coração e a rotina dos brasileiros



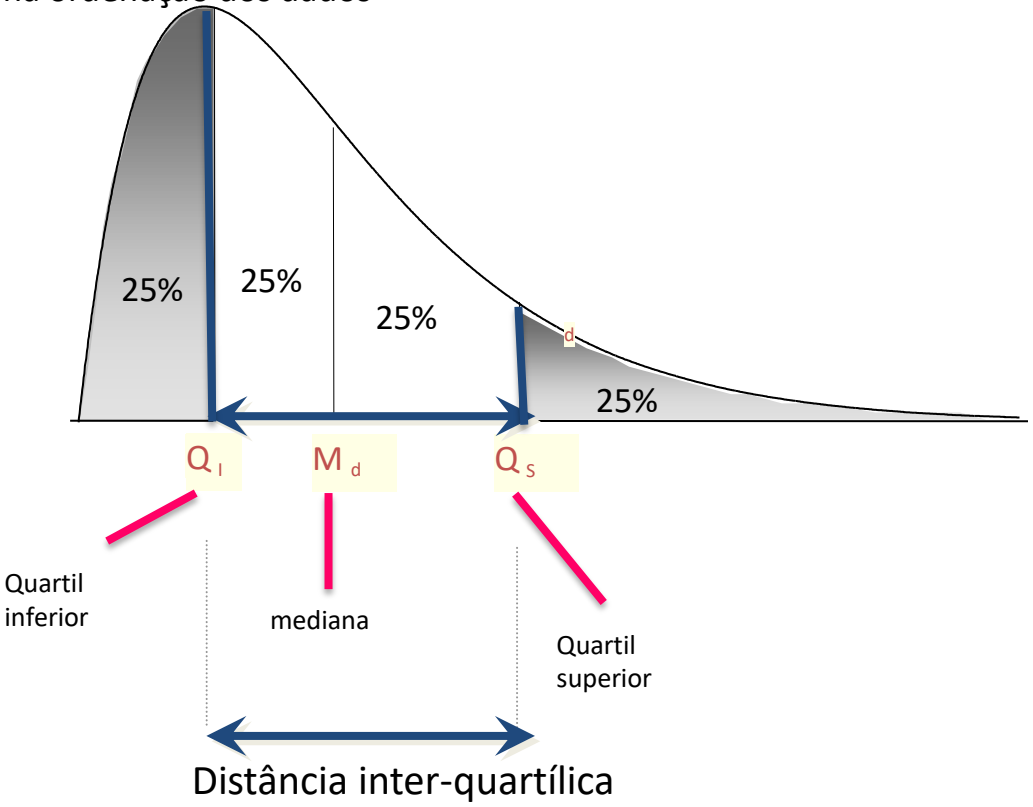
thinkwithgoogle.com/intl/pt-br/tendencias-de-consumo/tendencias-de-comportamento/os-brasileiros-e-o-esporte-como-as-pessoas-estao-se-exercitando-e-consumindo-conteudo-esportivo-no-pais/

MEDIDAS DE POSIÇÃO - RESUMO

	Vantagens	Limitações	Tipos de Variáveis
Média	Reflete todos os valores da amostra	É influenciada por valores extremos	Contínua e Discreta
Mediana	Menos sensível a valores extremos do que a média	Mais difícil de ser determinada para grande quantidade de dados	Contínua e Discreta
Moda	Representa um valor típico	Não tem função em certos conjuntos de dados	Contínua, Discreta, Nominal e Ordinal

MEDIDAS DE POSIÇÃO

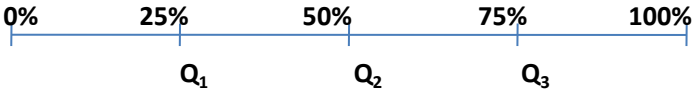
Medidas baseadas na ordenação dos dados



MEDIDAS DE POSIÇÃO

Medidas baseadas na ordenação dos dados

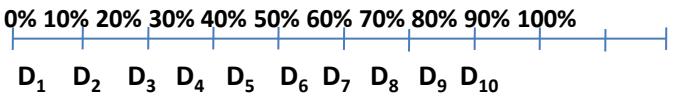
Quartis



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).

Decis



➔ Algumas aplicações:

- Construir classes ou faixas de variáveis
- Curva de concentração

MEDIDAS DE POSIÇÃO

Exemplo: Idade Pesquisa 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

idade	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

OUTRAS MEDIDAS 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.

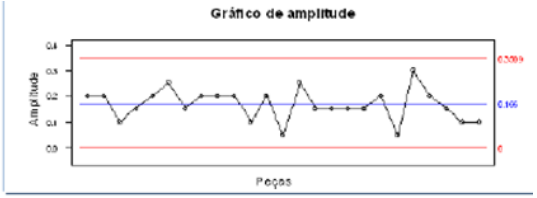
Exemplo

idade	Minimum	20,00
	Percentile 25	30,00
	Median	34,00
	Percentile 75	40,00
	Maximum	48,00

A = 28 anos

Aplicações:

- amplitude de temperatura em um dia
- controle de qualidad



A amplitude pode levar a erros de avaliação, pois não representa o conjunto dos dados.

Muitas vezes reflete muito mal a dispersão dos mesmos.

OUTRAS MEDIDAS AMPLITUDE INTER-QUARTÍLICA

É a diferença entre o terceiro e o primeiro quartil (Q3-Q1).

Usada em análise exploratória de dados – gráficos específicos.

idade	Minimum	20,00
	Percentile 25	30,00
	Median	34,00
	Percentile 75	40,00
	Maximum	48,00

IQ = 10 anos

- medida de dispersão em torno da mediana

MEDIDAS DISPERSÃO

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

Exemplo: **A: 5, 5, 5, 5, 5, 5**
B: 2, 3, 4, 6, 7, 8
C: 0, 1, 2, 8, 9, 10

$$\overline{X}_A = \overline{X}_B = \overline{X}_C = 5$$

Medidas de Dispersão

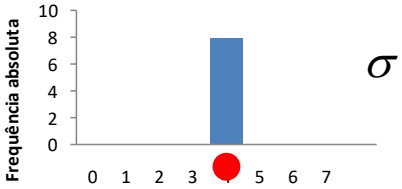
- Amplitude
- Amplitude Inter-Quartílica
- Variância
- Desvio-Padrão
- Coeficiente de Variação

MEDIDAS DISPERSÃO

Qual o desvio padrão?

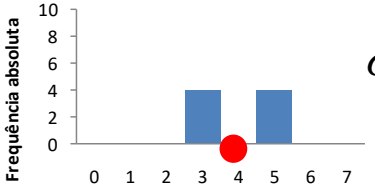
Exemplos:

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



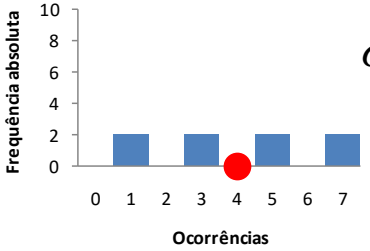
$\sigma = 0$

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



$\sigma = 1$

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



$\sigma = 2.24$

Média

MEDIDAS 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 destas distâncias.

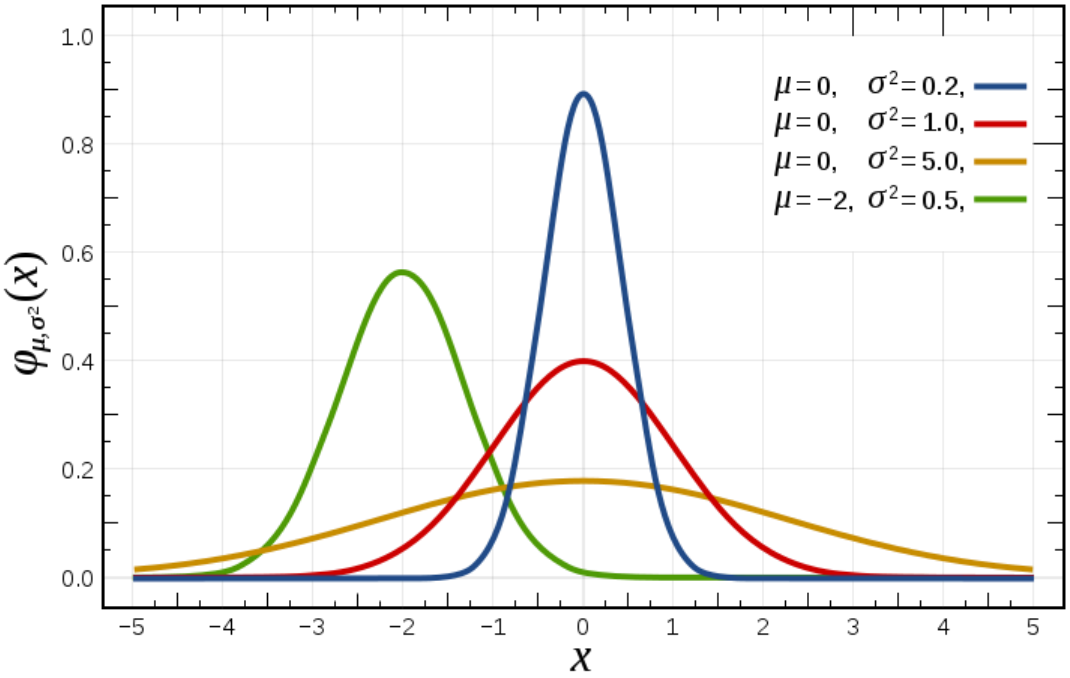
- variância da população

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

- variância amostral

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

MEDIDAS DISPERSÃO



MEDIDAS DISPERSÃO DESVIO PADRÃO

Desvio Padrão

É a raiz quadrada da variância.

$$s = \sqrt{S^2}$$

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.

MEDIDAS DISPERSÃO

Exemplo:

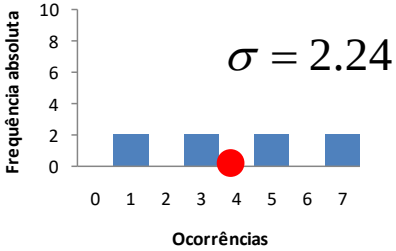
X	Média	(X-Média)	$(X\text{-Média})^2$
1	4	-3	9
1	4	-3	9
3	4	-1	1
3	4	-1	1
5	4	1	1
5	4	1	1
7	4	3	9
7	4	3	9
Soma	-	0	40

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

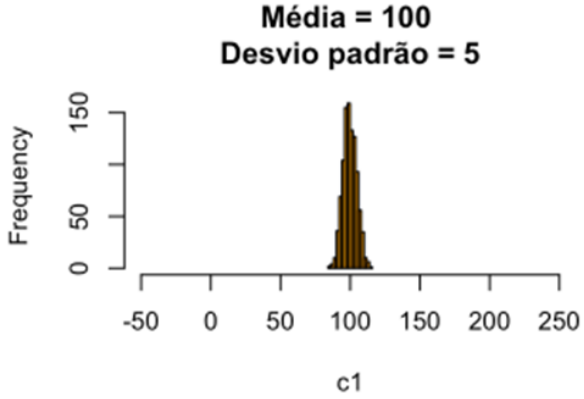
$$\sigma^2 = \frac{40}{8} = 5$$

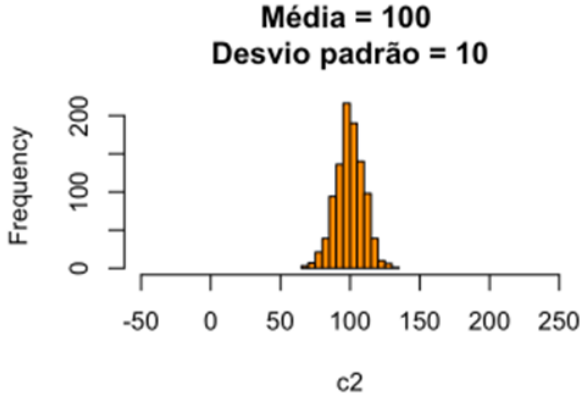
Desvio-Padrão:

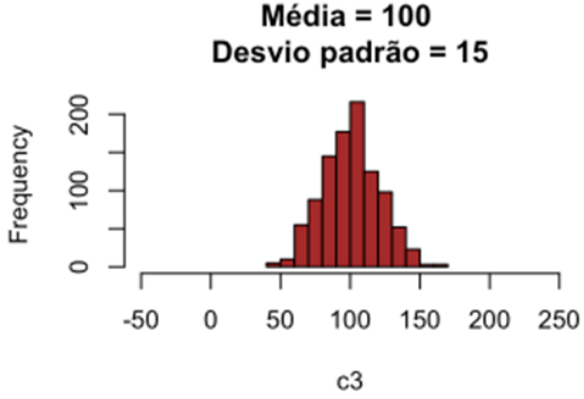
$$\sigma = \sqrt{\sigma^2} = \sqrt{5} = 2.24$$

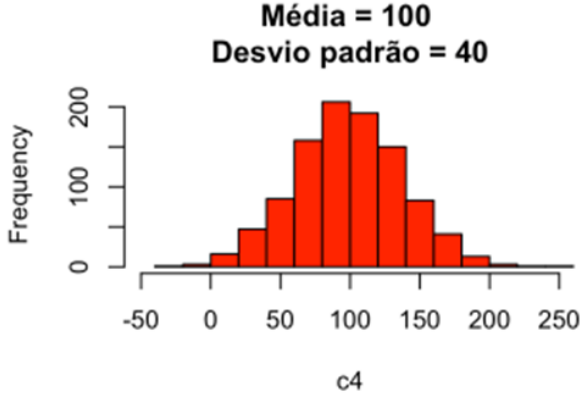


MEDIDAS DISPERSÃO DESVIO PADRÃO



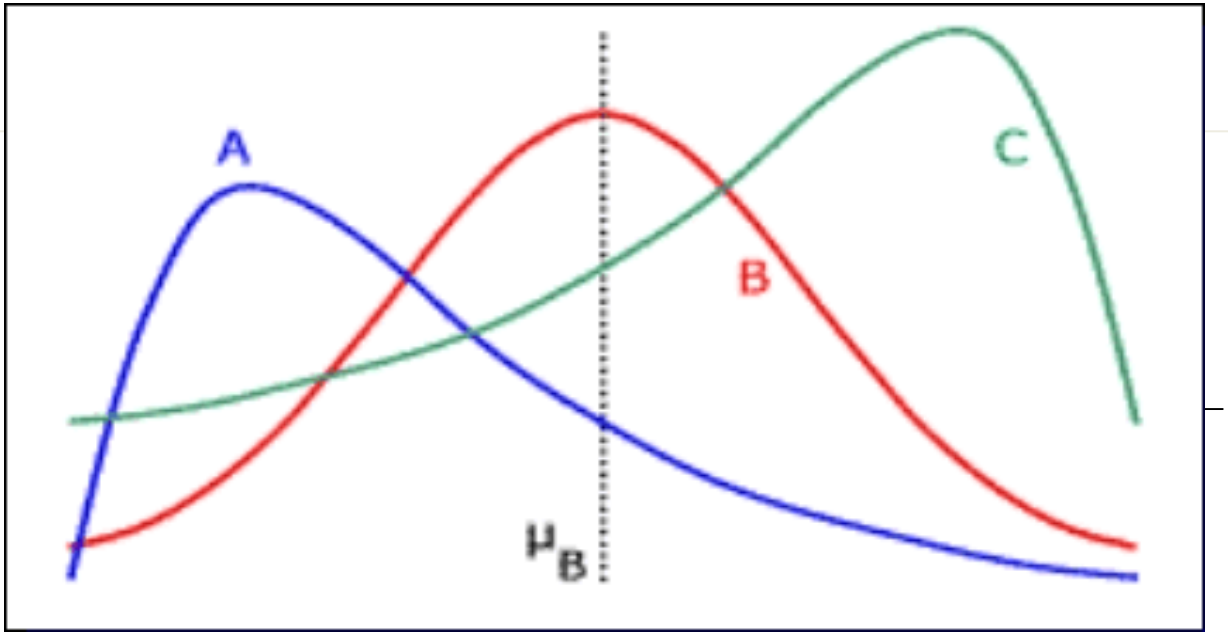






ASSIMETRIA

Uma distribuição ou uma curva é simétrica quando existe uma exata repartição de valores em torno do ponto central, ou seja, a média, a mediana e a moda coincidem. Os valores se agrupam mais acima ou mais abaixo do ponto central, e este “desvio” (ou viés) da simetria denomina-se assimetria.



GRAU DE DISPERSÃO COEFICIENTE DE VARIAÇÃO

Usado quando queremos **comparar duas variáveis quantitativas** quanto ao seu grau de dispersão.

$$CV = \frac{\sigma}{\overline{x}} * 100$$

É o quociente entre o desvio-padrão e a média.

Vantagem: caracterizar a dispersão dos dados em termos relativos a seu valor médio.

De uma forma geral, se o CV:

For menor ou igual a 15% → baixa dispersão: dados homogêneos
For entre 15 e 30% → média dispersão
For maior que 30% → alta dispersão: dados heterogêneos

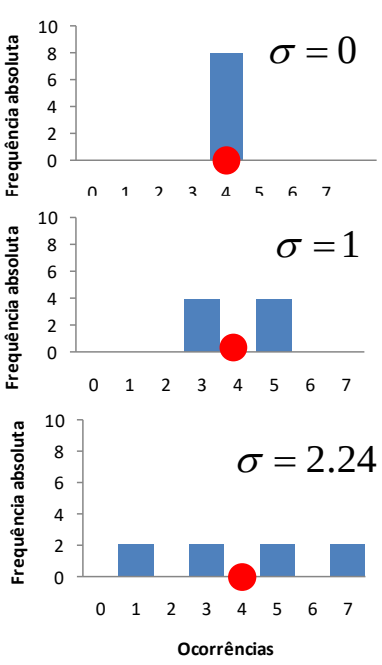
COEFICIENTE DE VARIAÇÃO

Exemplo: Qual o coeficiente de variação?

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

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

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



$CV = 0$

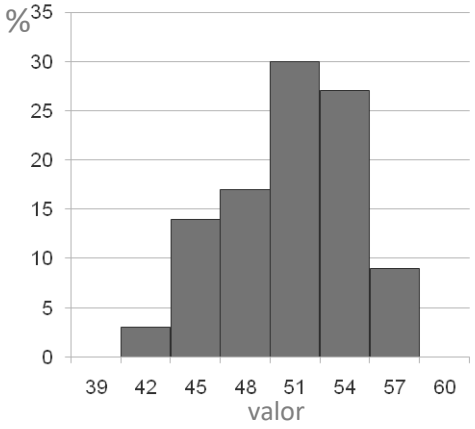
$CV = 25\%$

$CV = 56\%$

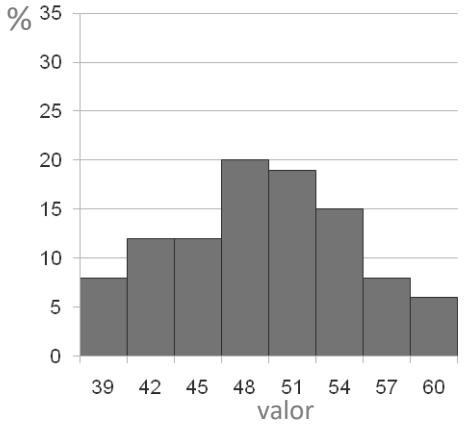
Média

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 outro, de 5,8. Qual é qual?



(a)



(b)

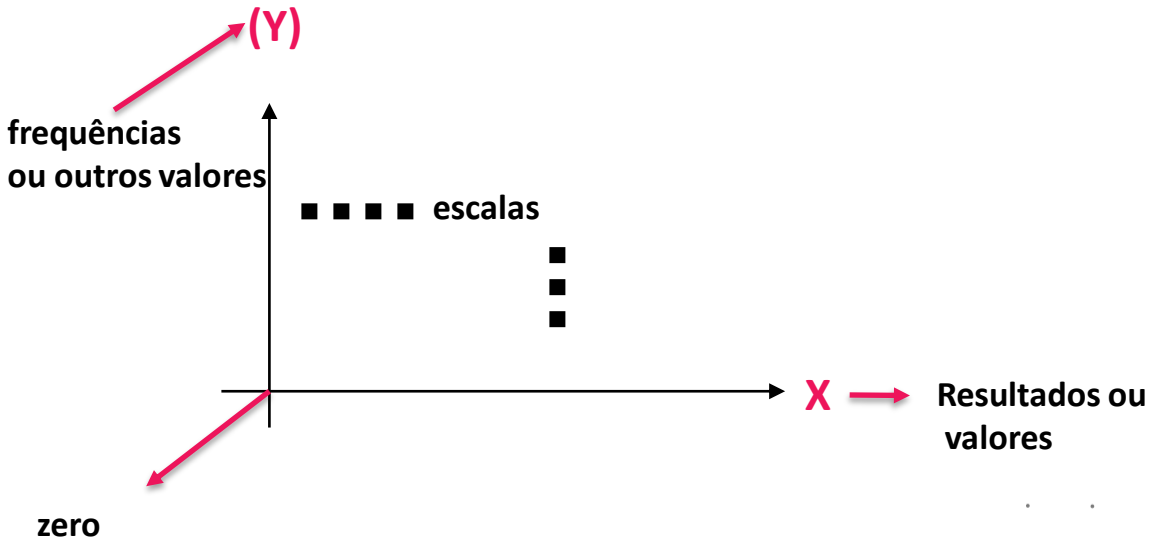
Desvio-padrão:

Coeficiente
de Variação:

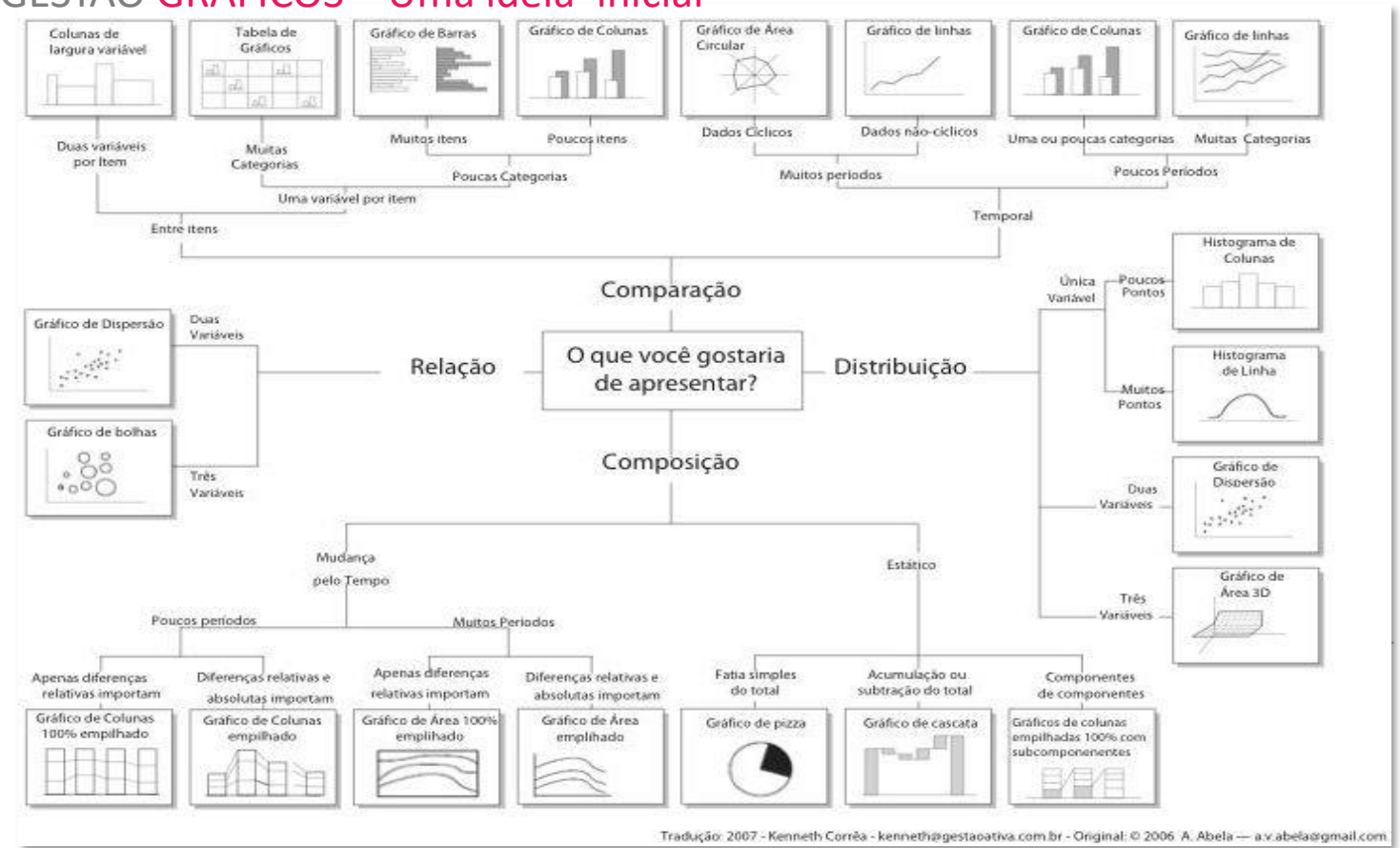
APRESENTAÇÃO GRÁFICA DOS DADOS

REPRESENTAM TABELAS (distribuições, coeficientes, séries)

➢ **EIXOS CARTESIANOS**



SUGESTÃO GRÁFICOS – Uma ideia inicial

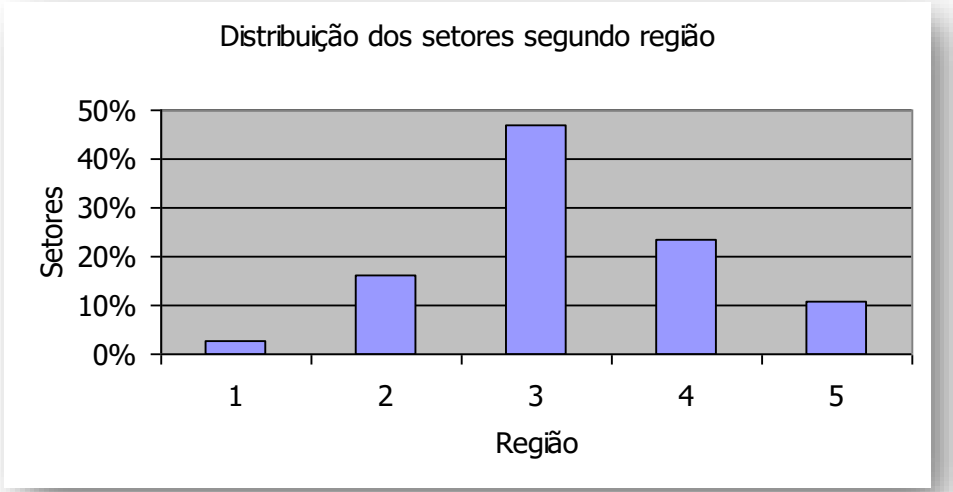


GRÁFICOS: Variáveis qualitativas ou discretas

Colunas

Um gráfico de colunas ilustra comparações entre itens. As categorias são organizadas na horizontal e os valores são distribuídos na vertical.

Exemplo:



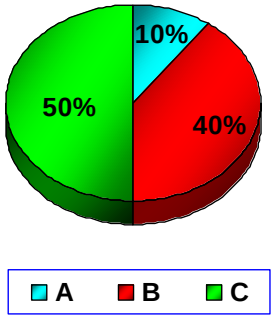
GRÁFICOS: Variáveis qualitativas ou discretas

- **Setores ou pizza**

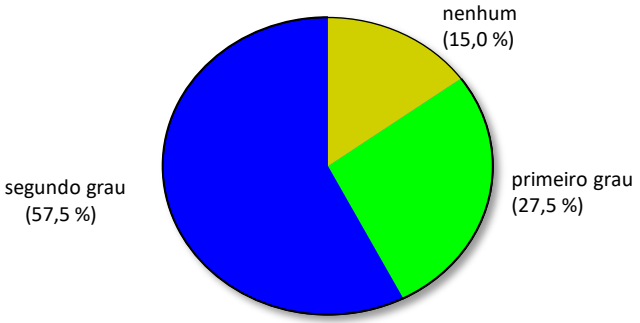
  Um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens. A frequência relativa (%) transformada em graus mediante o cálculo proporcional.

Exemplos:

Classe Social



Grau de Instrução do Chefe da Casa

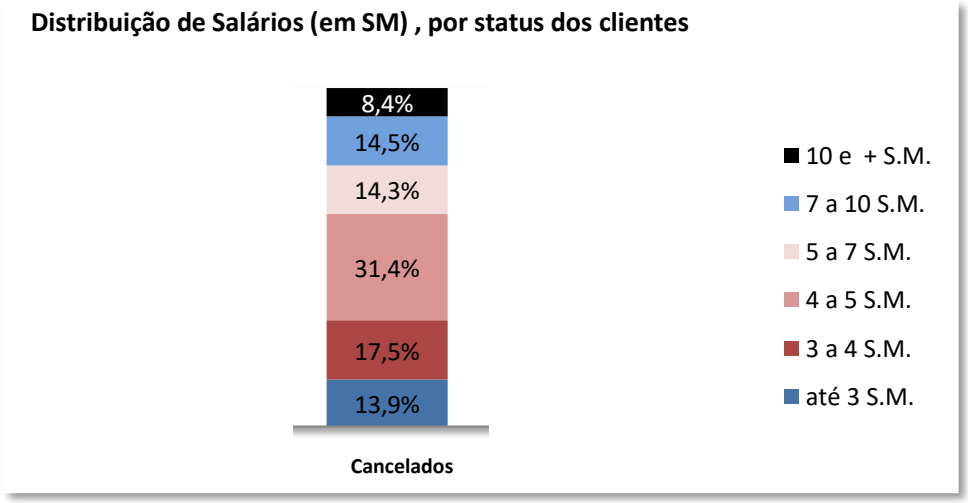


GRÁFICOS: Variáveis qualitativas ou discretas

- **Colunas sobrepostas**

  Nesta representação as barras estarão sobrepostas, com uso de duas ou mais variáveis. Sendo a soma 100%.

Exemplo:

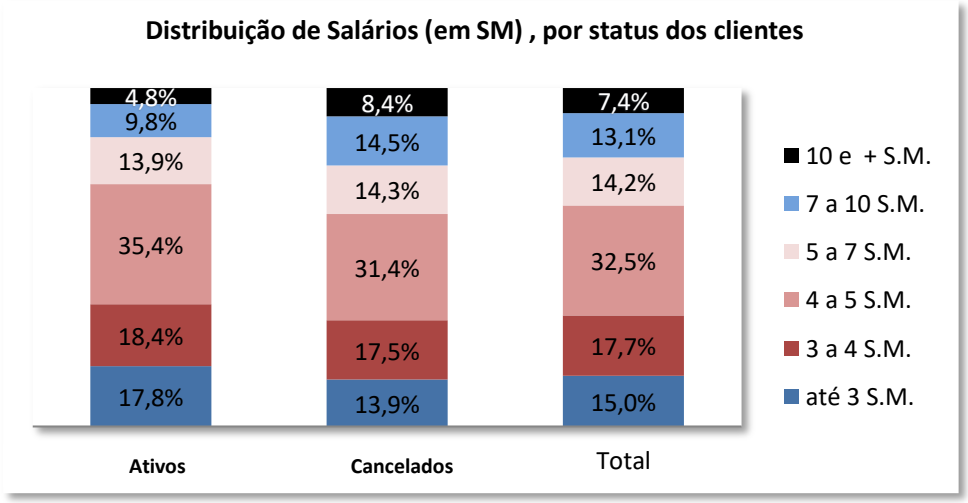


GRÁFICOS: Variáveis qualitativas ou discretas

- **Colunas sobrepostas**

  Nesta representação as barras estarão sobrepostas, com uso de duas ou mais variáveis. Sendo a soma 100%.

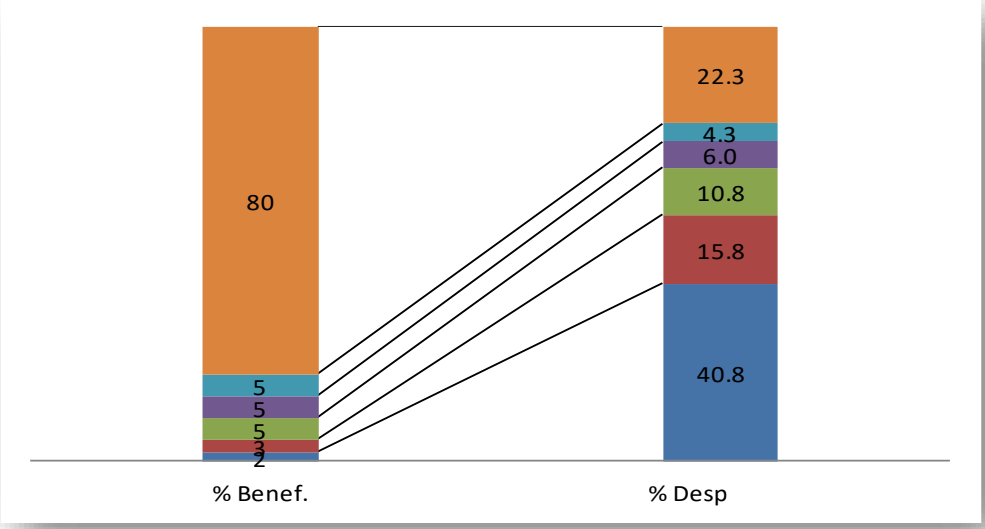
Exemplo:



OUTROS GRÁFICOS

Exemplo: Número de Funcionários e Despesas em R$

Gráfico de Pareto de despesas em R$ e Número de Beneficiários

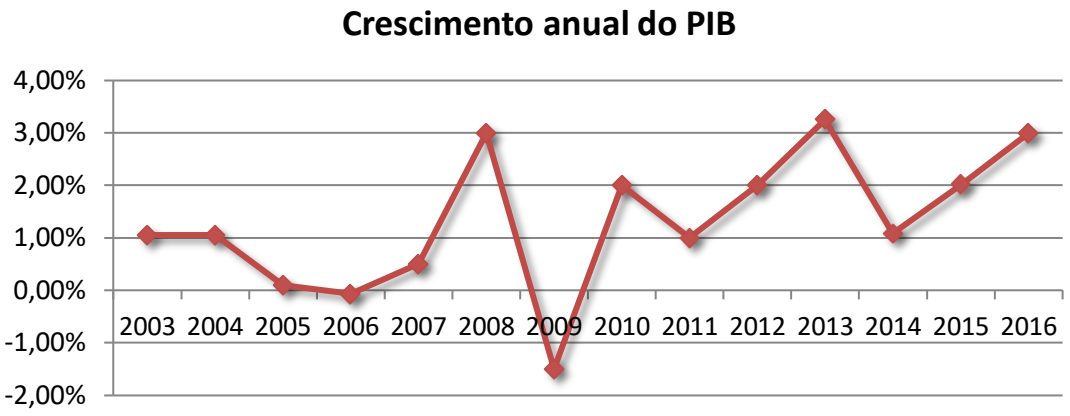


GRÁFICOS: Variáveis quantitativas

Linha

Um gráfico de linha mostra tendências nos dados em intervalos iguais.

Exemplo:

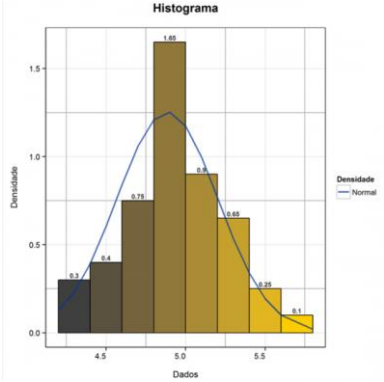


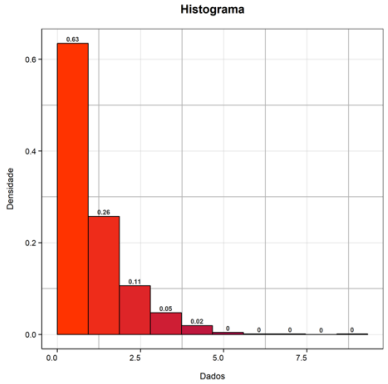
Fonte: IBGE, 2017.

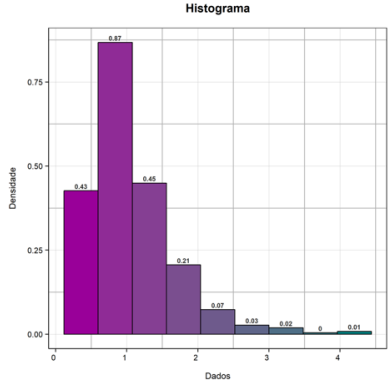
GRÁFICOS: Variáveis quantitativas

Histograma

O histograma é formado por retângulos cujas áreas representam frequências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.







http://www.portalaction.com.br/estatistica-basica/17-histograma

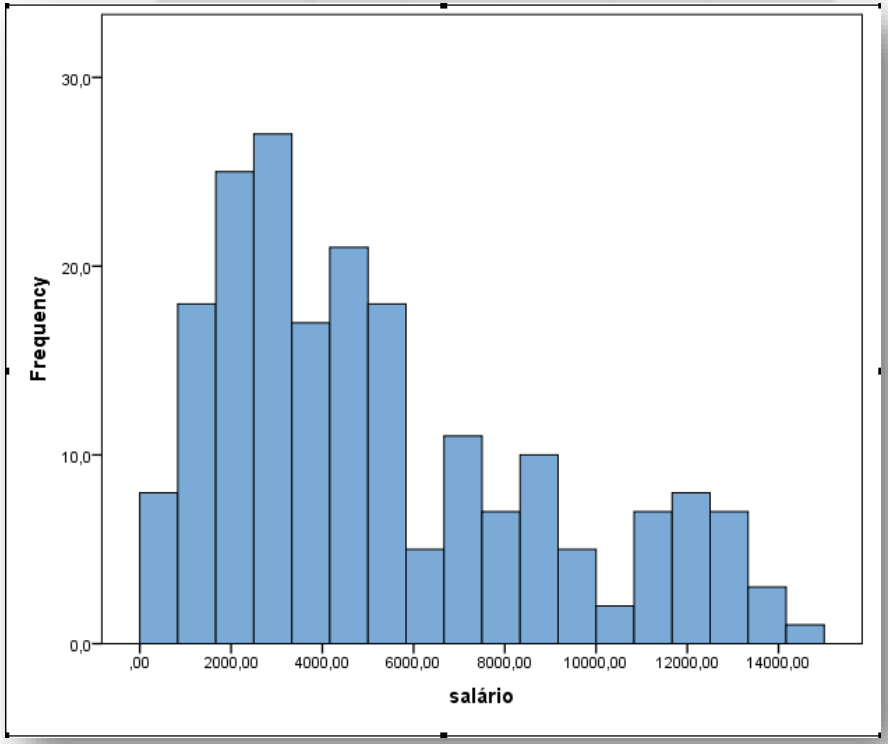
Exemplo

GRÁFICOS: Variáveis quantitativas

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	4.858,72
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	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
40	3.273,02

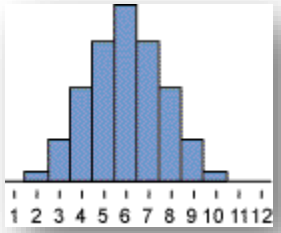
Descriptive Statistics

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

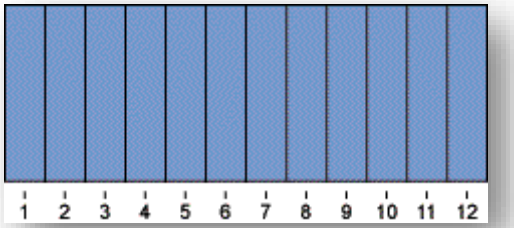


GRÁFICOS: HISTOGRAMA

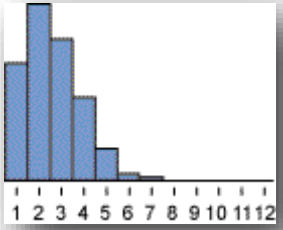
Simétrico



Uniforme

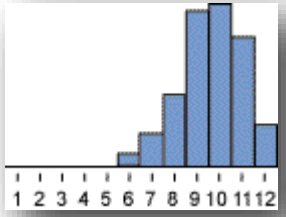


Assimétrico à direita ou positiva



Moda <Mediana <Média

Assimétrico à esquerda ou negativa



Média< Mediana <Moda

Fonte: Larson/Farber cap.2

Exemplo

GRÁFICOS: HISTOGRAMA

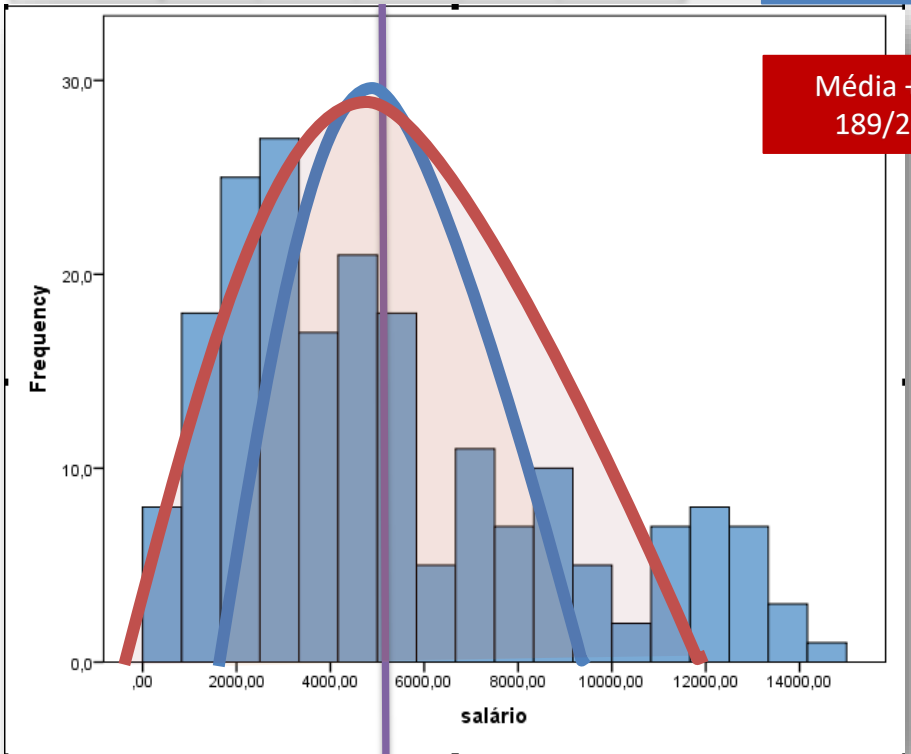
id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	13.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.858,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.753,71
29	3.426,82
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%



GRÁFICOS: HISTOGRAMA

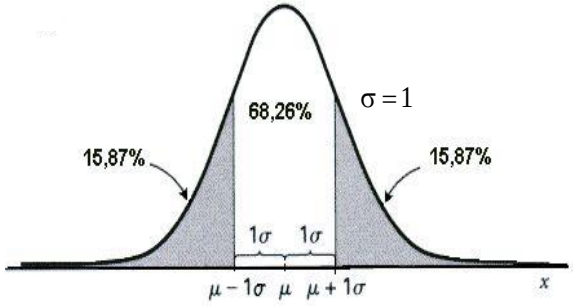
Medidas de Dispersão : Interpretação do Desvio- Padrão

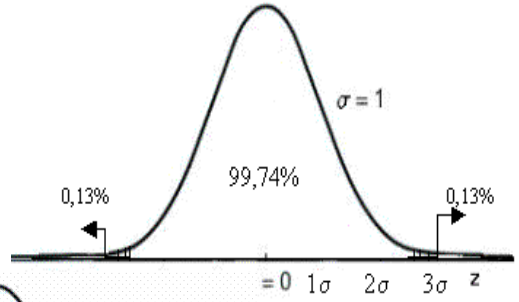
Regra Empírica

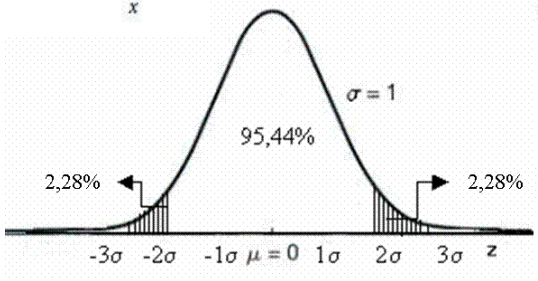
média (+/-) desvio 68 % dos casos

média (+/-) 2x desvio 95 % dos casos

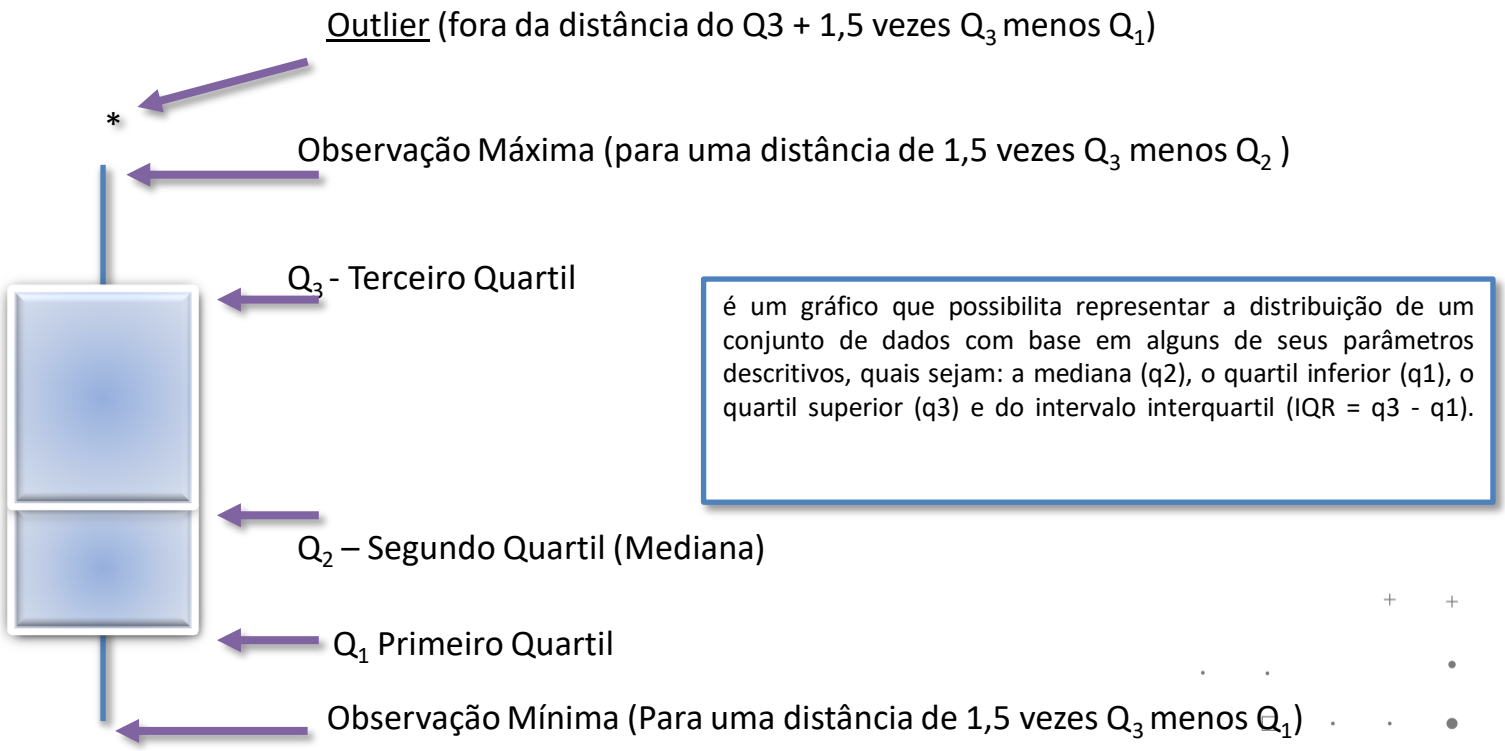
média (+/-) 3x desvio 100 % dos casos







GRÁFICOS: BOX-PLOT



é um gráfico que possibilita representar a distribuição de um conjunto de dados com base em alguns de seus parâmetros descritivos, quais sejam: a mediana (q2), o quartil inferior (q1), o quartil superior (q3) e do intervalo interquartil (IQR = q3 - q1).

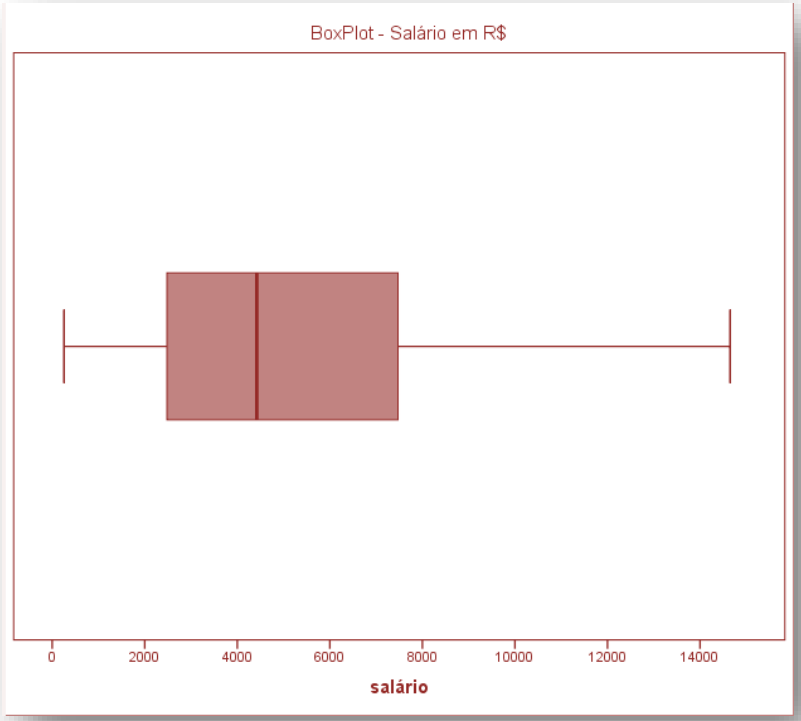
GRÁFICOS: BOX-PLOT

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	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	4.858,72
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	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
34	5.889,54
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				

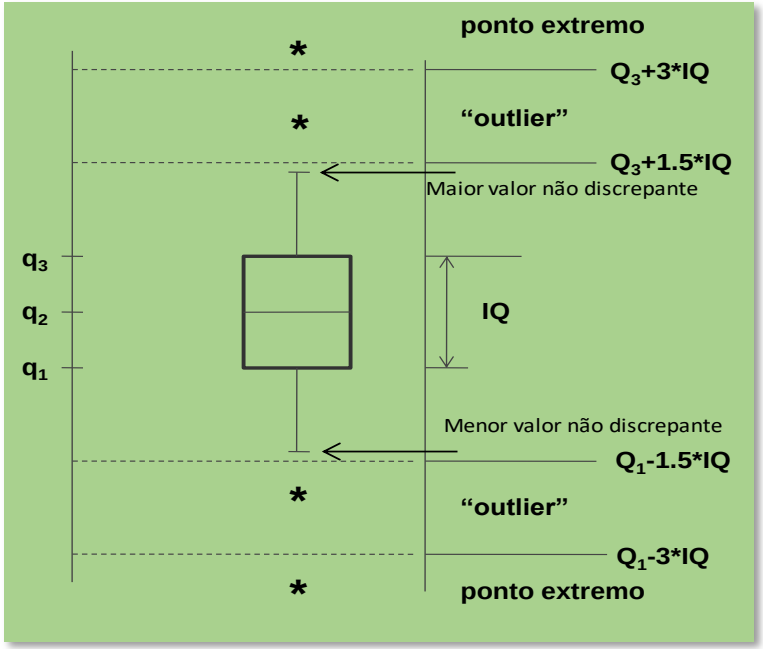


BOX-PLOT: DETECÇÃO DE OUTLIERS

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

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

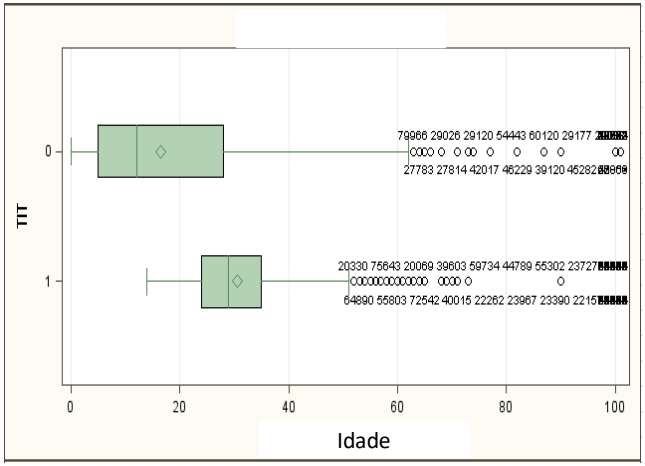
Representação Gráfica na Análise dos Dados

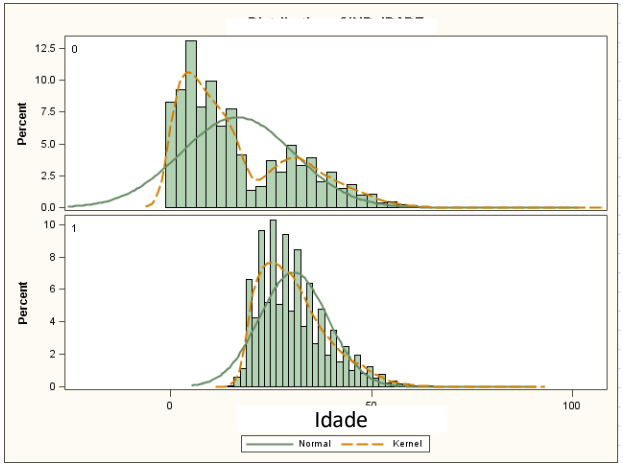


BOX-PLOT: DETECÇÃO DE OUTLIERS

Exemplo

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





BOX-PLOT: DETECÇÃO DE OUTLIERS

Exemplo

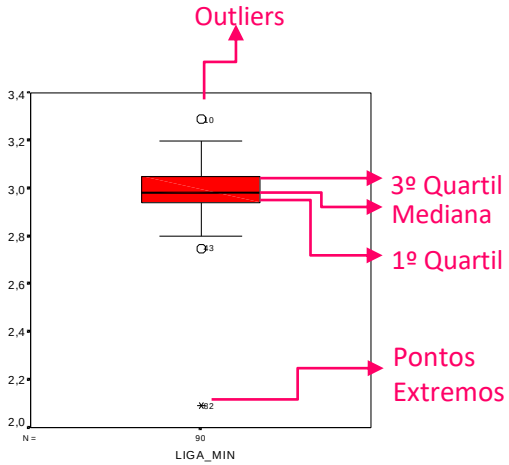
Outliers

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

Gráfico Box-Plot

Exemplo: Preço médio do produto A

N	90
Range	1,2
Mean	2,99
Median	2,98
Percentil 25	2,94
Percentil 75	3,05
Interquartile Range	0,11
Variance	0,02
Skewness	-2,92
Kurtosis	19,64



Preço médio do produto A

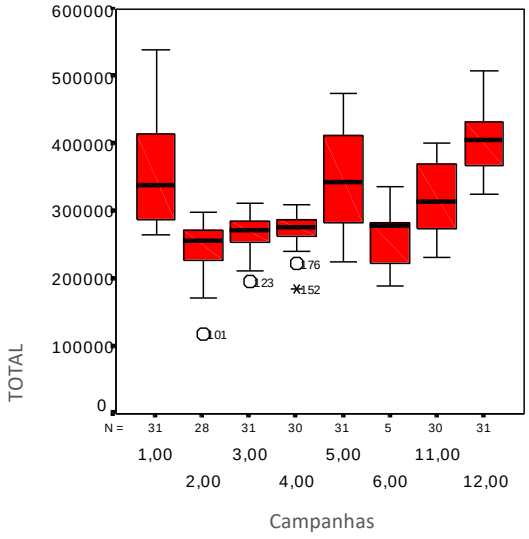
BOX-PLOT: DETECÇÃO DE OUTLIERS

Exemplo

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

Gráfico Box-Plot

Exemplo: “Total de unidades vendidas – Campanha 1 a 12 do ano YY



TRATAMENTO DOS DADOS

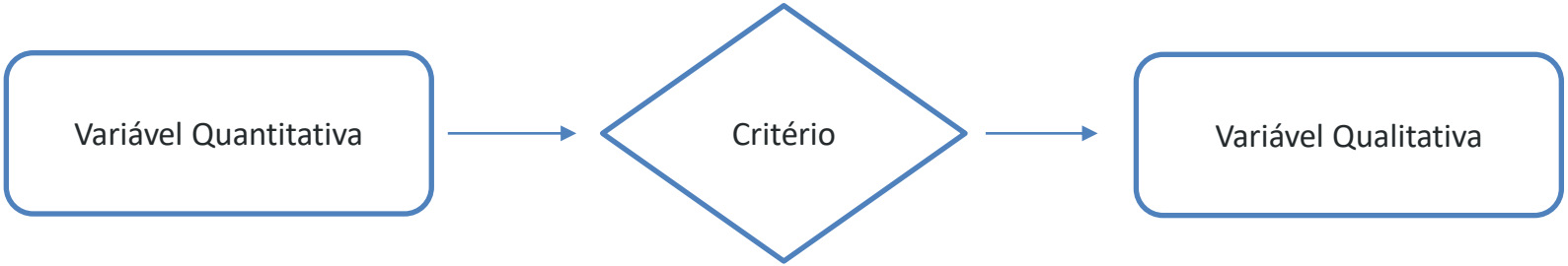
Transformar a variável quantitativa em qualitativa

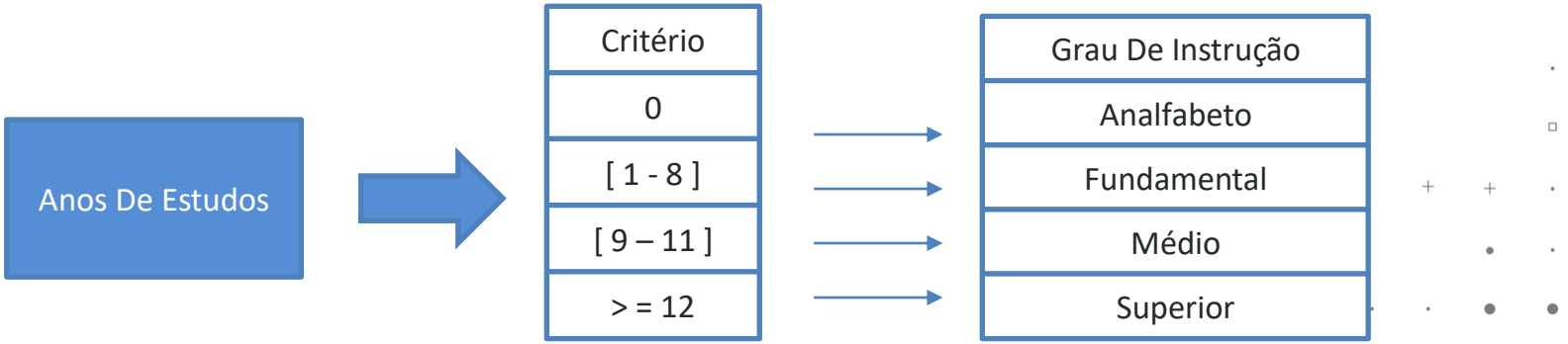
Detecção de outliers

Funções de normalizações dos dados

TRATAMENTO DOS DADOS

Transformando variáveis quantitativas em qualitativas





TRANSFORMANDO VARÍAVEIS QUANTITATIVAS EM QUALITATIVAS

Exemplo:

Quantas classes serão necessárias para representar a despesa anual?

Medidas resumo da despesa anual

Mean	Std Dev	Minimum	Maximum	Mode	Range	Sum	N
265,22	537,55	0	4491,19	0	4491,19	16118247,5	60773

Fórmula de Sturges[1] ➡ K = 1 + 3,322 * log n

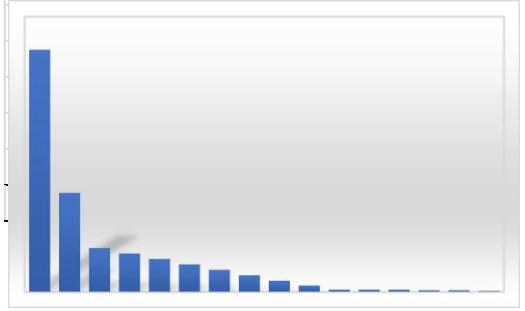
K = número de classes

➔ Substituindo:

➔ K = 1 + 3,3 * log (60.773) = 16,78 ~ 17

$$Intervalo = \frac{(Máximo - Mínimo)}{K} = \frac{4491,19}{17} = 264,18 \cong 265$$

Despesa	N	%	%ac
[0 - 265)	26740	44,0	44,0
[265 - 530)	10939	18,0	62,0
[530 - 795)	4862	8,0	70,0
[795 - 1060)	4254	7,0	77,0
[1060 - 1325)	3646	6,0	83,0
[1325 - 1590)	3039	5,0	88,0
[1590 - 1855)	2431	4,0	92,0
[1855 - 2120)	1823	3,0	95,0
[2120 - 2385)	1215	2,0	97,0
[2385 - 2650)	608	1,0	98,0
[2650 - 2915)	243	0,4	98,4
[2915 - 3180)	243	0,4	98,8



[1] A fórmula de Sturges relaciona os tamanhos dos intervalos de classes a partir da extensão dos dados

TRANSFORMANDO VARIÁVEIS

Construir novas variáveis - a partir de um conjunto de outras variáveis.

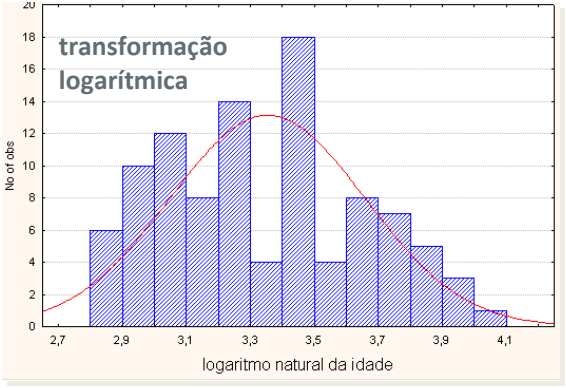
Exemplo de transformação: criar proporção, transformações logarítmicas, etc

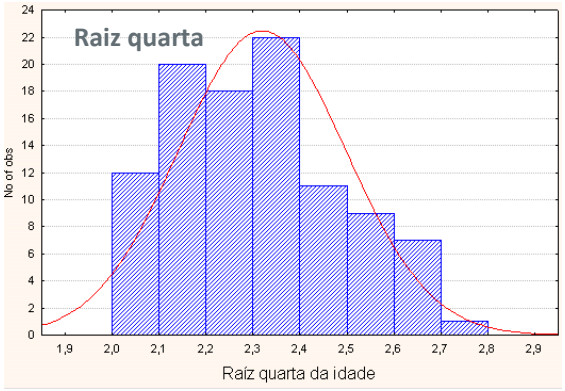
Renda anual	Renda (em MM)	LogRenda	RaizQuadradaRenda
R$ 12.000	0,01	4,08	109,54
R$ 24.000	0,02	4,38	154,92
R$ 60.000	0,06	4,78	244,95
R$ 120.000	0,12	5,08	346,41
R$ 240.000	0,24	5,38	489,9
R$ 360.000	0,36	5,56	600
R$ 600.000	0,6	5,78	774,6
R$ 900.000	0,9	5,95	948,68

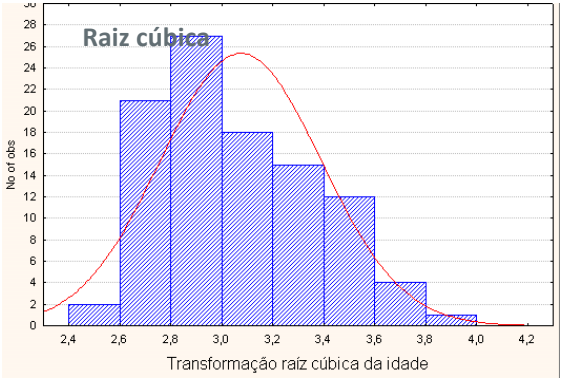
Uma variável no estudo que tem distribuição assimétrica, ou possuir um viés, ou seja, uma das extremidades elevadas e uma cauda longa, pode influenciar nos resultados de algumas análises, como em medidas de correlações ou análises de regressão. A aplicação de uma transformação pode reduzir esse viés.

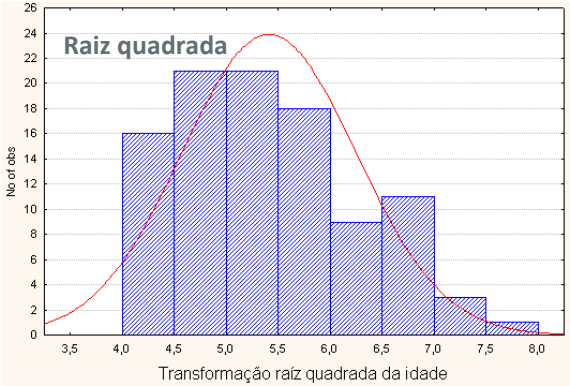
TRANSFORMANDO VARIÁVEIS

CONSTRUÇÃO DE NOVAS VARIÁVEIS









TRANSFORMANDO VARIÁVEIS

Normatização Min-Max

Normalização Min-Max - transformação, onde os dados de um atributo são normalizados gerando valores entre 0,0 a 1,0.

$$valor\ transformado = \frac{\text{valor original} - \text{valor mínimo}}{\text{valor máximo} - \text{valor mínimo}}$$

Por exemplo: suponha que os valores mínimo e máximo da variável rendimento são R$ 360,00 e R$ 15.800,00, respectivamente. Transformar a variável rendimento na faixa [0,0; 1,0].Um valor de rendimento igual a R$ 5.300,00, transforma-se em:

$$w = \frac{5.300 - 360}{15.800 - 360} = 0{,}32$$

TRANSFORMANDO VARIÁVEIS

Normatização Min-Max

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	4.858,72
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	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
40	3.273,02

$$sal\_norm1 = \frac{salário - mínimo}{máximo - mínimo}$$

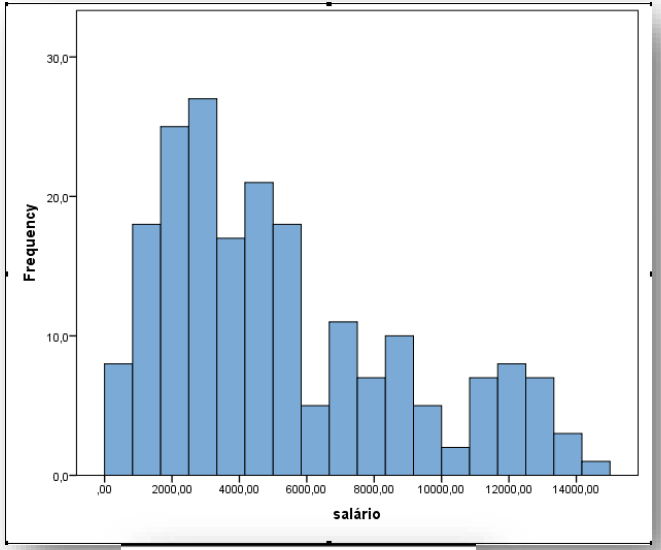
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
salário	200	254,19	14649,52	5259,30	3615,36
sal_norm1	200	0,00	1,00	0,35	0,25
Valid N (listwise)	200				

id	salário	sal_norm1
1	4.763,75	0,31
2	7.391,72	0,5
3	729,33	0,03
4	2.376,28	0,15
5	1.887,72	0,11
6	1.207,36	0,07
7	4.745,39	0,31
8	3.635,80	0,23
9	8.119,15	0,55
10	2.356,41	0,15
11	13.502,54	0,92
12	2.655,92	0,17
13	3.920,45	0,25
14	853,32	0,04
15	12.819,59	0,87
16	10.088,13	0,68
17	4.414,62	0,29
18	7.293,00	0,49
19	11.445,93	0,78
20	8.339,63	0,56
21	4.858,72	0,32
22	1.616,16	0,09
23	1.339,24	0,08
24	7.108,82	0,48
25	2.054,73	0,13
26	1.441,01	0,08
27	8.981,38	0,61
28	8.753,71	0,59
29	3.426,82	0,22
30	3.873,20	0,25
31	1.165,56	0,06
32	5.431,64	0,36
33	12.541,13	0,85
40	3.273,02	0,21

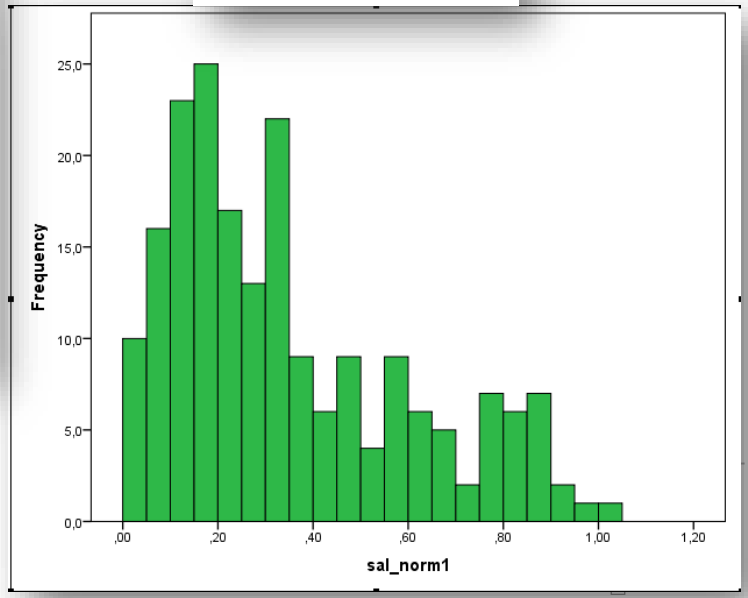
TRANSFORMANDO VARIÁVEIS

Normatização Min-Max



Mean = 5259,3047
Std. Dev. = 3615,36167
N = 200

Mean = ,3477
Std. Dev. = ,25115
N = 200



TRANSFORMANDO VARIÁVEIS

PADRONIZAÇÃO

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

$$z = \frac{X - \bar{X}}{S}$$

Na fórmula, x é o valor original da variável; $\bar{X}$ é a média da variável x; e s é o seu respectivo desvio-padrão.

Por exemplo: suponha que os valores média e desvio da variável rendimento são R$ 5.259,30 e R$ 3.615,36, respectivamente. Transformar a variável rendimento em uma nova variável padronizada.

Um valor de rendimento igual a R$ 5.300,00 transforma-se em:

$$z = \frac{5.300 - 5.259,30}{3.615,36} = 0,011$$

TRANSFORMANDO VARIÁVEIS

PADRONIZAÇÃO

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	4.858,72
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	8.753,71
29	3.426,82
30	3.873,20
31	1.165,56
32	5.431,64
33	12.541,13
40	3.273,02

$$zsalário = \frac{salário - média}{desvio}$$

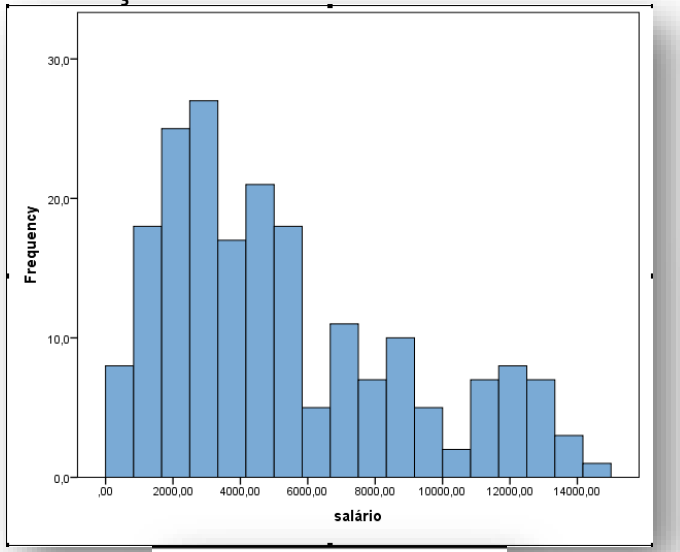
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
salário	200	254,19	14.649,52	5.259,30	3.615,36
Zscore(salário)	200	-1,3844	2,5973	0,0000	1,0000
Valid N (listwise)	200				

id	salário	Zsalário
1	4.763,75	-0,1371
2	7.391,72	0,5898
3	729,33	-1,2530
4	2.376,28	-0,7974
5	1.887,72	-0,9326
6	1.207,36	-1,1208
7	4.745,39	-0,1422
8	3.635,80	-0,4491
9	8.119,15	0,7910
10	2.356,41	-0,8029
11	13.502,54	2,2801
12	2.655,92	-0,7201
13	3.920,45	-0,3703
14	853,32	-1,2187
15	12.819,59	2,0912
16	10.088,13	1,3356
17	4.414,62	-0,2336
18	7.293,00	0,5625
19	11.445,93	1,7112
20	8.339,63	0,8520
21	4.858,72	-0,1108
22	1.616,16	-1,0077
23	1.339,24	-1,0843
24	7.108,82	0,5116
25	2.054,73	-0,8864
26	1.441,01	-1,0561
27	8.981,38	1,0295
28	8.753,71	0,9665
29	3.426,82	-0,5069
30	3.873,20	-0,3834
31	1.165,56	-1,1323
32	5.431,64	0,0477
33	12.541,13	2,0141
40	3.273,02	-0,5494

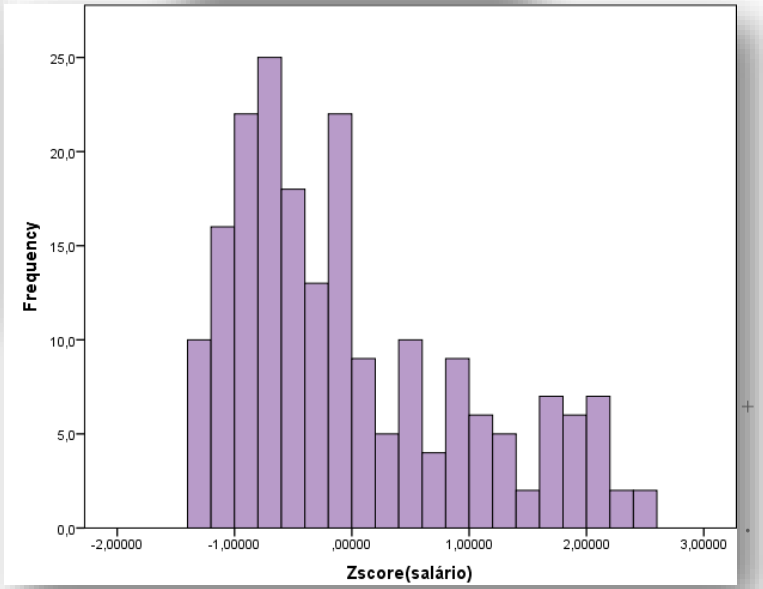
TRANSFORMANDO VARIÁVEIS

PADRONIZAÇÃO



Mean = 5259,3047
Std. Dev. = 3615,36167
N = 200

Mean = ,00000
Std. Dev. = 1,00000
N = 200



TRANSFORMANDO VARIÁVEIS

PADRONIZAÇÃO

Observação:

- A escala das variáveis pode afetar muito a qualidade das predições.
- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.
- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1 ou normalizar gerando valores dentre 0,0 a 1,0.

Exemplos de aplicações da normalização dos dados:

Convolutional Neural Networks (CNNs) e Algoritmos de Machine Learning (Regressão Penalizada, SVM, Cluster e outros)

TRANSFORMANDO VARIÁVEIS QUALITATIVAS EM NUMÉRICAS

Variáveis Categorizadas : Os algoritmos trabalham melhor com poucas categorias. Para reduzir o número de categorias pode-se usar atributos dos códigos, ou variáveis binárias para cada categoria.

“One-hot encoding”: Alguns algoritmos têm dificuldade em entender variáveis que têm mais de uma categoria. Acham que é uma variável contínua (0,1,2,3...) ➔ porém não tem significado contínuo. A solução é transformar todas as categorias em uma variável diferente de valores 0 e 1 (one-hot encoding)Variável com n categorias ➔ criar n variáveis .

Pode trazer problemas em alguns modelos, como na regressão linear (solução criar dummy) (n-1 variáveis).

TRANSFORMANDO VARIÁVEIS QUALITATIVAS EM NUMÉRICAS

Exemplo: Variável categórica SEXO {Masculino, Feminino}

- Variável Transformada: Masculino
- Masculino=1 se SEXO= “Masculino”;Masculino=0, caso contrário
- A variável Masculino é uma variável numérica, também chamada de variável DUMMY muito utilizada na construção de modelos.

Análise Exploratória de Dados

Base
Colaboradores

BIBLIOGRAFIA

- KUHN, M. / JOHNSON K. Applied Predictive Modeling, 1st ed. 2013, Corr. 2nd printing 2018 Edition
- LESKOVEC, RAJAMARAM, ULLMAN. Mining of Massive Datasets, 2014. http://mmds.org.
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. Análise multivariada de dados, 2009
- TORGO, L. Data Mining with R: Learning with Case Studies, 2.a ed. Chapman and Hall/CRC , 2007
- BUSSAB, W.O.; MORETTIN, P. A., Estatística básica, 5a. ed., São Paulo: Saraiva, 2006.
- MINGOTI, S.A.; Análise de dados através de métodos de estatística multivariada, UFMG, 2005
- CARVALHO, L.A.V., Datamining – A mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Editora Ciência Moderna, 2005.
- BERRY,M.J.A., LINOFF,G. Data Mining Techniques For Marketing, Sales and Customer Support. 3a. ed. New York: John Wiley & Sons, Inc., 2011.
- DUNHAM, M.H. Data Mining - Introductory and Advanced Topics. Prentice Hall, 2002.
- DINIZ,C.A.R. , NETO F.L. Data Mining: Uma Introdução. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.

OBRIGADA!

 /AdelaideAlves

 profadelaide.alves@fiap.com.br


FIAP