

Analise de dados e Estatística

Fatec 2025

Estatística, a chave da nova revolução

1

Revolução da Agricultura → atualmente o mundo produz mais alimentos através de cultivos, do que através de caça.

2

Revolução Industrial → fábricas de produção em massa deram ao mundo uma imensa variedade de opções de produtos.

3

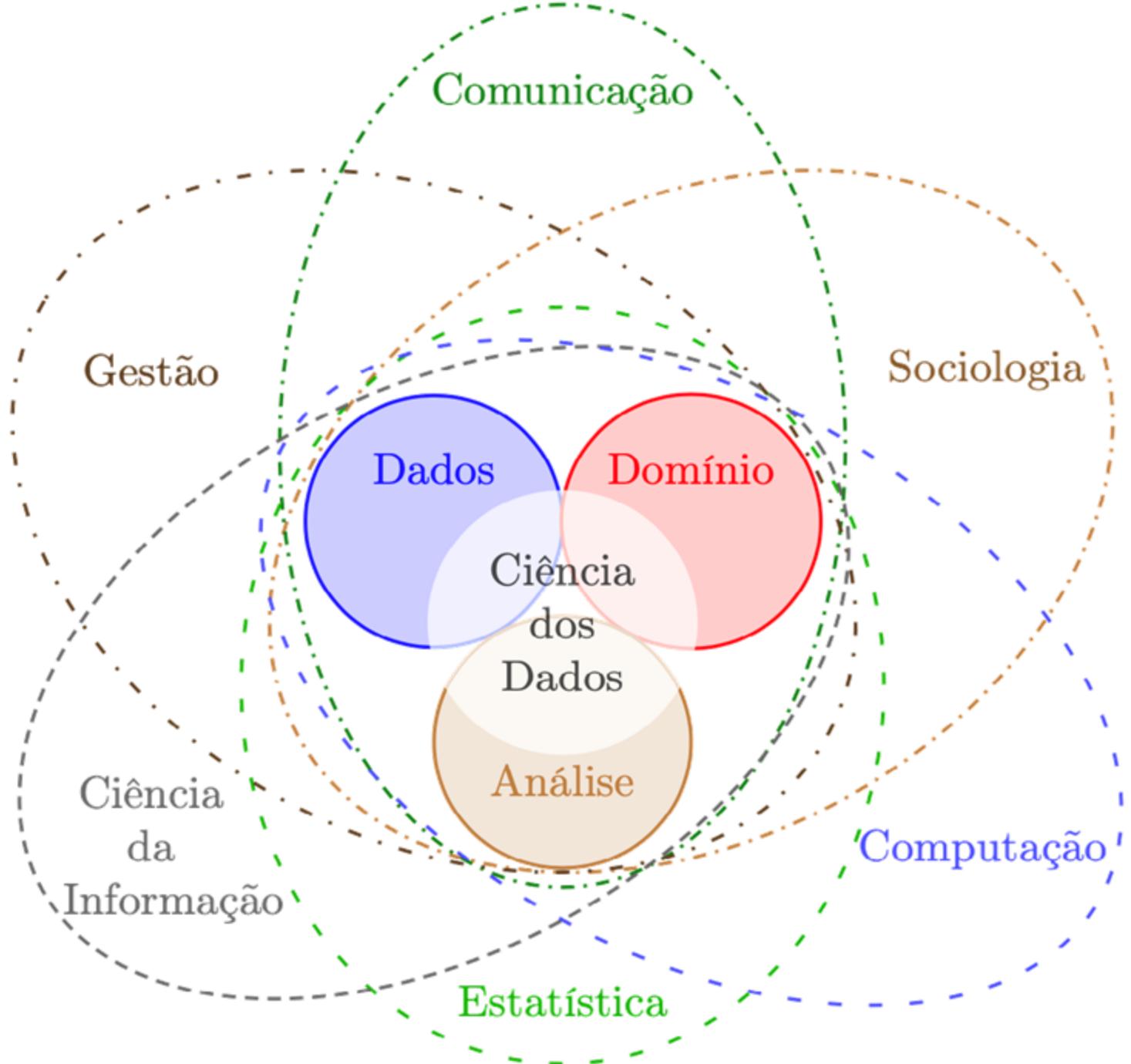
Revolução da Informação → a tecnologia nos deu uma grande variedade de produtos eletrônicos, tornou a indústria mais eficiente e aumentou consideravelmente a quantidade de informação a nossa disposição.

4

Revolução Digital → estamos no meio desta revolução e o volume de dados gerados pela humanidade, nos traz o desafio de conseguir extrair informação útil. A análise estatística é a chave desta revolução.

Habilidades

Ciência dos Dados



O que é Analise de dados

- Arte de analisar dados crus e transformá-los em informação
- O objetivo da análise é sempre otimizar os resultados de um modelo

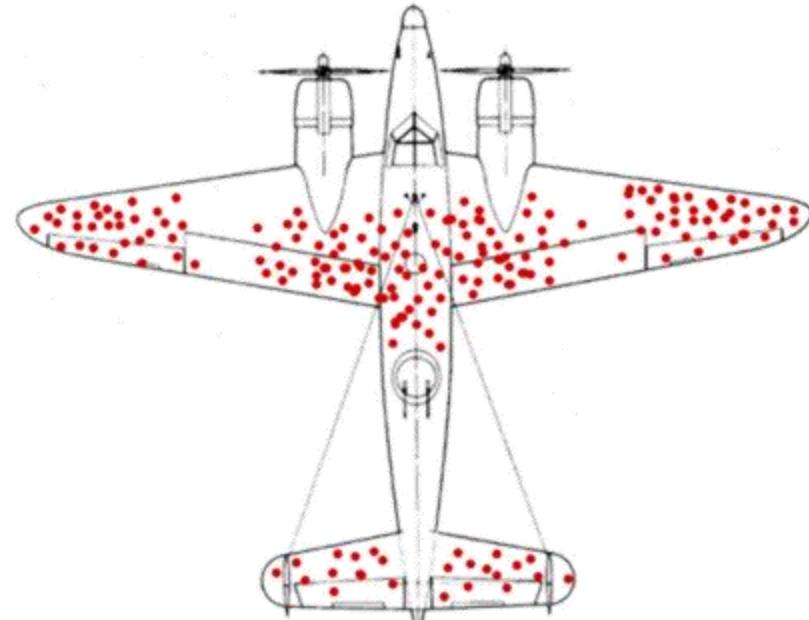


Viés de sobrevivência

Abraham Wald – Segunda Guerra Mundial



- Estatísticos da equipe de *Abraham Wald* analisavam todos os aviões que voltavam das batalhas
- Objetivo era identificar onde a estrutura dos aviões deveria ser reforçada da maneira mais eficiente
- Seria óbvio reforçar os locais onde há perfurações
- Mas esses aviões **conseguiram voltar**
- Ou seja, deve-se reforçar os locais que **não tinham** perfurações
- Esse caso ilustra o viés da sobrevivência (*survival bias*), um tipo de viés de seleção



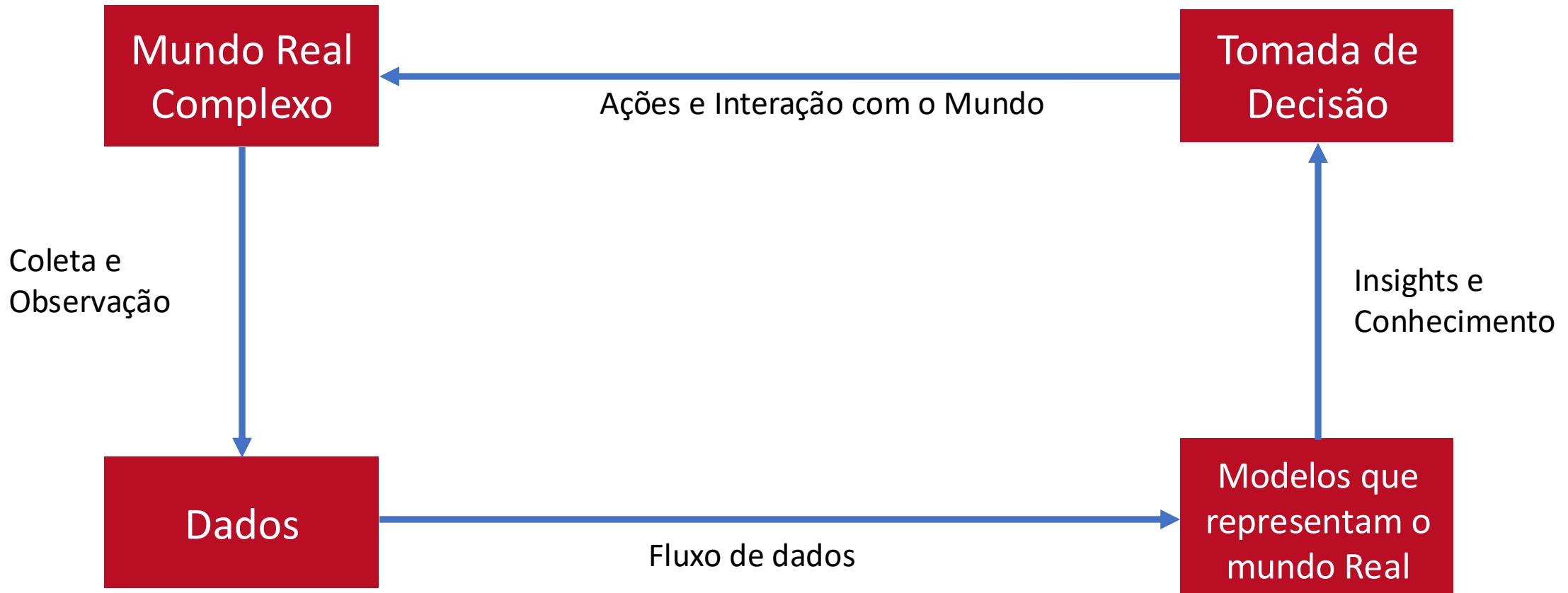
Se usarmos a única fonte de informação disponível como sendo suficiente para uma hipótese, vamos ignorar grande parte das causas dos problemas. As vezes, a resposta mais importante está na informação que **está faltando**

Estabelecendo Questionamentos

- Uma das tarefas mais comuns de um cientista de dados é descobrir padrões e obter insights a partir da **análise exploratória de dados**
- Durante este processo, os dados são avaliados de modo a se encontrar padrões, detectar anomalias e formular hipóteses
- Normalmente as técnicas de EDA são feitas por meio da **visualização analítica**
- A qualidade dos achados depende da habilidade do cientista em **formular as questões corretas** e realizar uma sequência de análises que possibilitem responder a elas!

Framework de Análise Exploratória

Modelo Mental



Quais são as perguntas certas?

- Formule perguntas que possibilite compreender os dados e seus processos de transformação, ao passo em que constrói um background sobre o problema
- Por onde começar?
 - **Data Scrutiny / Description**
 - Avalie a estrutura e qualidade dos dados
 - Quantos registros? Quantas variáveis?
 - Como os dados foram coletados?
 - Qual o formato dos dados? Tipos de dados?
 - Há erros nos dados? Outliers? Valores únicos?
 - Missing data?
 - Calcule as estatísticas descritivas para os dados
 - Verifique a distribuição e comportamento dos dados
 - **Interaja com os dados** (auxiliado por estatística descritiva e visualização)

a transformation can sometimes work wonders in deciphering the data

Chatfield (1985)

Quais são as perguntas certas?

- Qual o motivo da sua análise?
 - Explorar os dados ...
 - Otimizar algum resultado
 - Predizer algum desfecho
 - Etc.
- Além disso, você deve avaliar os dados e o problema que está formulando sobre os seguintes aspectos:
 - Riscos
 - Benefícios
 - Contingências
 - Regulamentações
 - Recursos
 - Requisitos

Estabelecendo Questionamentos

- Análise Geral
 - Qual pergunta eu quero responder com essa estatística?
 - O que esta estatística significa?
 - Com o que uma observação mediana se parece? E quanto a uma observação extrema?
 - Qual o horizonte temporal?
- Desafiando premissas:
 - Valores maiores ou menores são melhores?
 - Valores estão dentro da expectativa esperada?
 - Quais premissas estou usando?
 - Quais os vieses da coleta de dados?

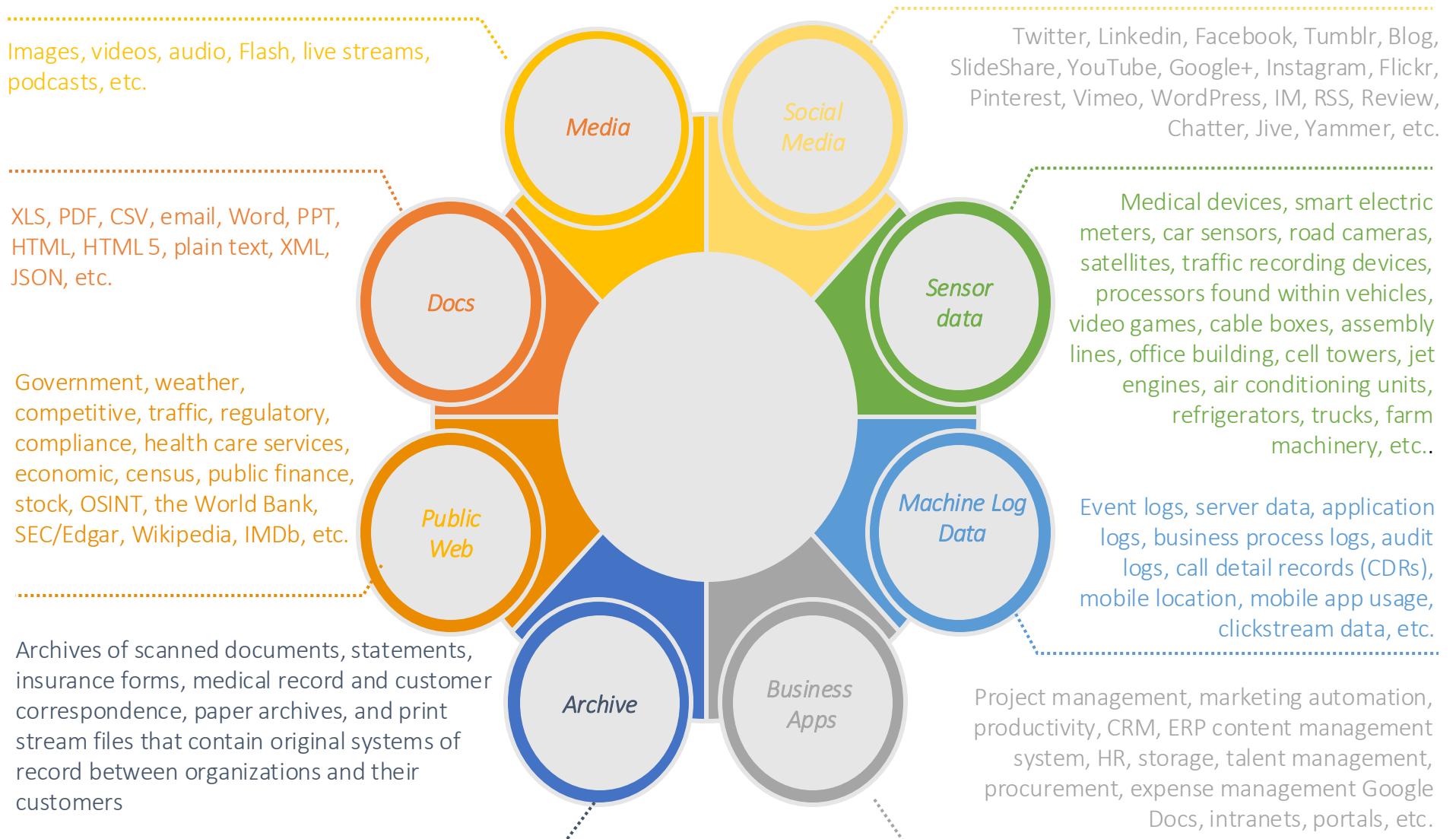
Estabelecendo Questionamentos: evidências e ponto de vista

- Evidências
 - O quão forte é essa evidência?
 - Significância estatística x Relevância de negócio
 - O que está faltando nos meus dados? Quais dimensões não estão sendo levadas em conta?
- Perspectiva e ponto de vista
 - O que eu gostaria que os dados me falassem?
 - Como seriam os dados se a conclusão fosse diferente?
 - Se não existissem dados, o que eu faria?

Os dados

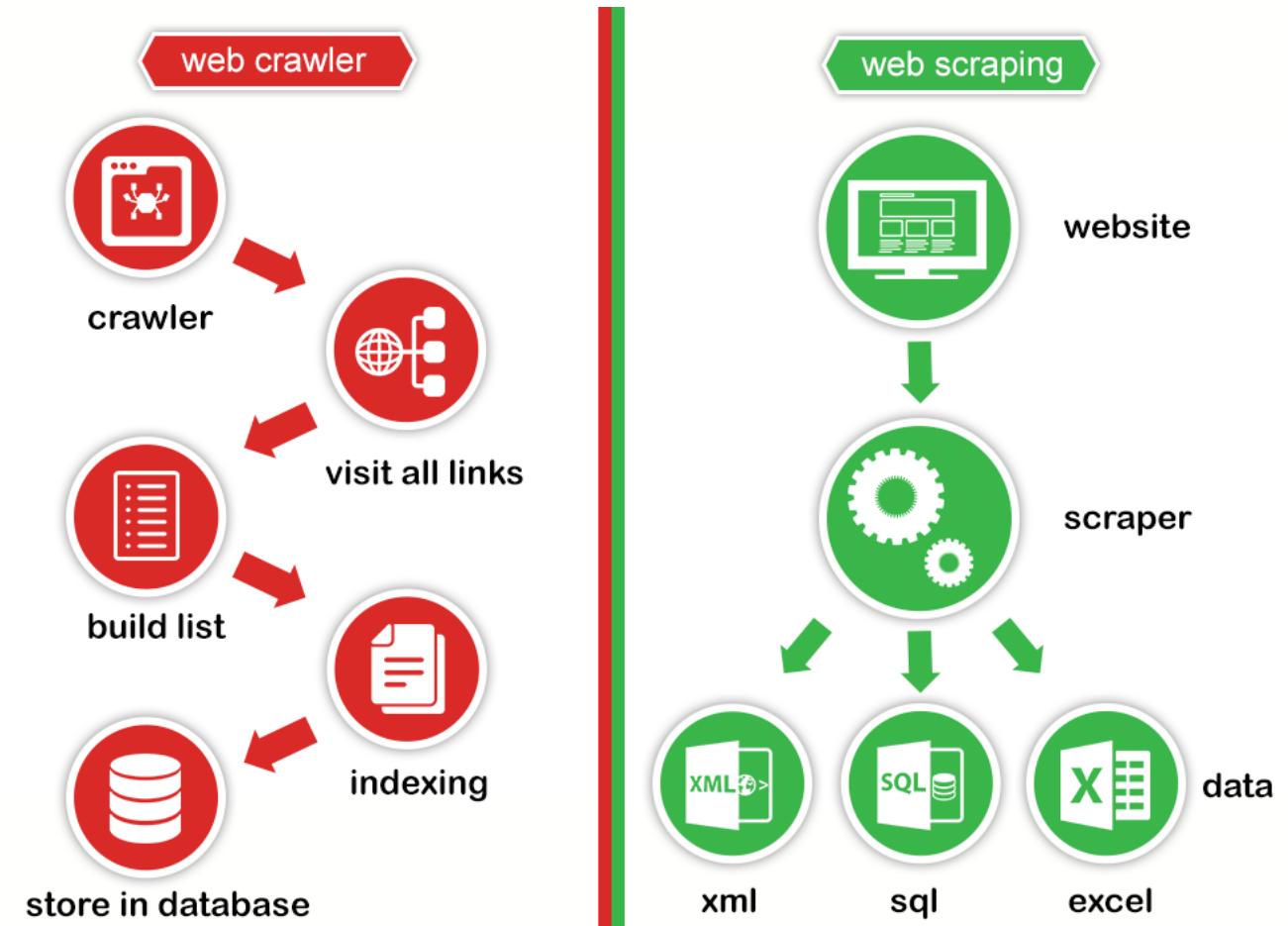
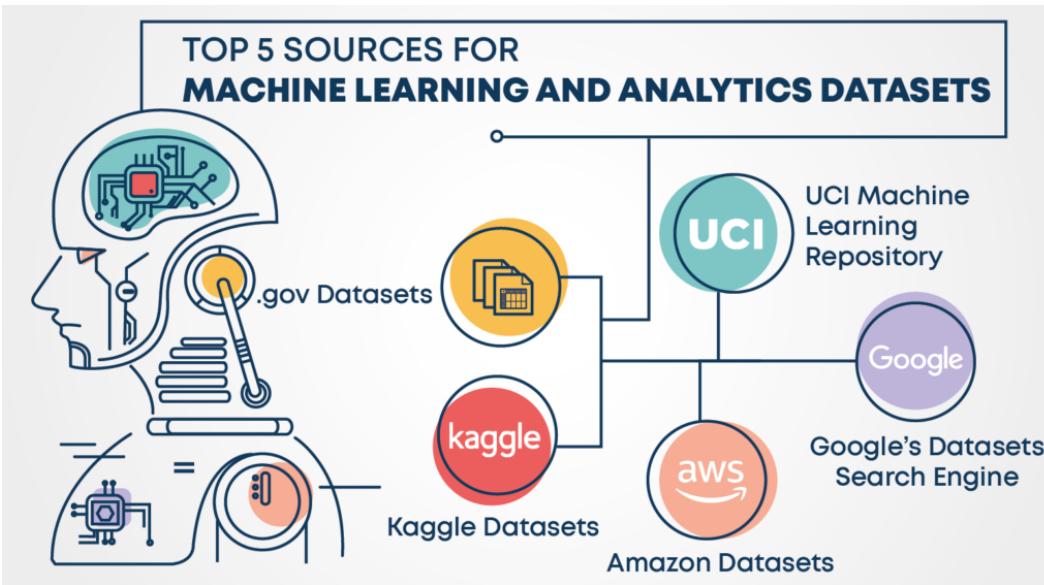
Fatec 2025

Fonte e formatos de dados



Dados

Repositórios, Crawling e Scraping



Dados

Formato

■ Tabular

x	y	tempo	red	green	blue
0	0	0	255	0	0
0	1	0	200	10	6
...					
0	0	0.1	255	50	100
0	1	0.1	255	200	190
...					



➤ **Tupla**,
ponto multidimensional,
Vetor,
Linha
Instância

Dados

Organização

```
calls_for_service.tsv
1 CASENO OFFENSE EVENTDT EVENTTM CVLEGEND CVDOW InDbDate Block_Location
BLKADDR City State
2 18000273 VEHICLE STOLEN 01/01/2018 12:00:00 AM 20:30 MOTOR VEHICLE THEFT
1 01/24/2018 03:30:18 AM "1100 PARKER ST
3 Berkeley, CA
4 (37.859364, -122.288914)" 1100 PARKER ST Berkeley CA
5 17092476 BURGLARY AUTO 12/12/2017 12:00:00 AM 13:30 BURGLARY - VEHICLE
2 01/24/2018 03:30:17 AM "2300 LE CONTE AVE
6 Berkeley
7 (37.874867, -122.263689)" 2300 LE CONTE AVE,Berkeley,CA
8 17092534
9 Berkeley
10 (37.857495, -122.275256)
11 17091517,THEFT MISD. (U
12 Berkeley, CA
13 (37.876791, -122.280472)
14 17048102,THEFT FROM AUT
```

TSV
Tab separated values

```
calls_for_service.csv
1 CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,Stat
2 18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018
3 01/24/2018 03:30:18 AM,"1100 PARKER ST
4 Berkeley, CA
5 (37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
6 17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018
7 01/24/2018 03:30:17 AM,"2300 LE CONTE AVE
8 Berkeley, CA
9 (37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
10 17092534,BURGLARY AUTO,
11 03:30:17 AM,"1700 STUAR
12 Berkeley, CA
13 (37.857495, -122.275256)
14 17091517,THEFT MISD. (U
15 03:30:11 AM,"1600 CALIF
16 Berkeley, CA
17 (37.876791, -122.280472)
18 17048102,THEFT FROM AUT
```

CSV
Comma separated
values

```
1 {
2   "field1": "value1",
3   "field2": ["list", "of", "values"],
4   "myfield3": {"is_recursive": true, "a null value": null}
5 }
```

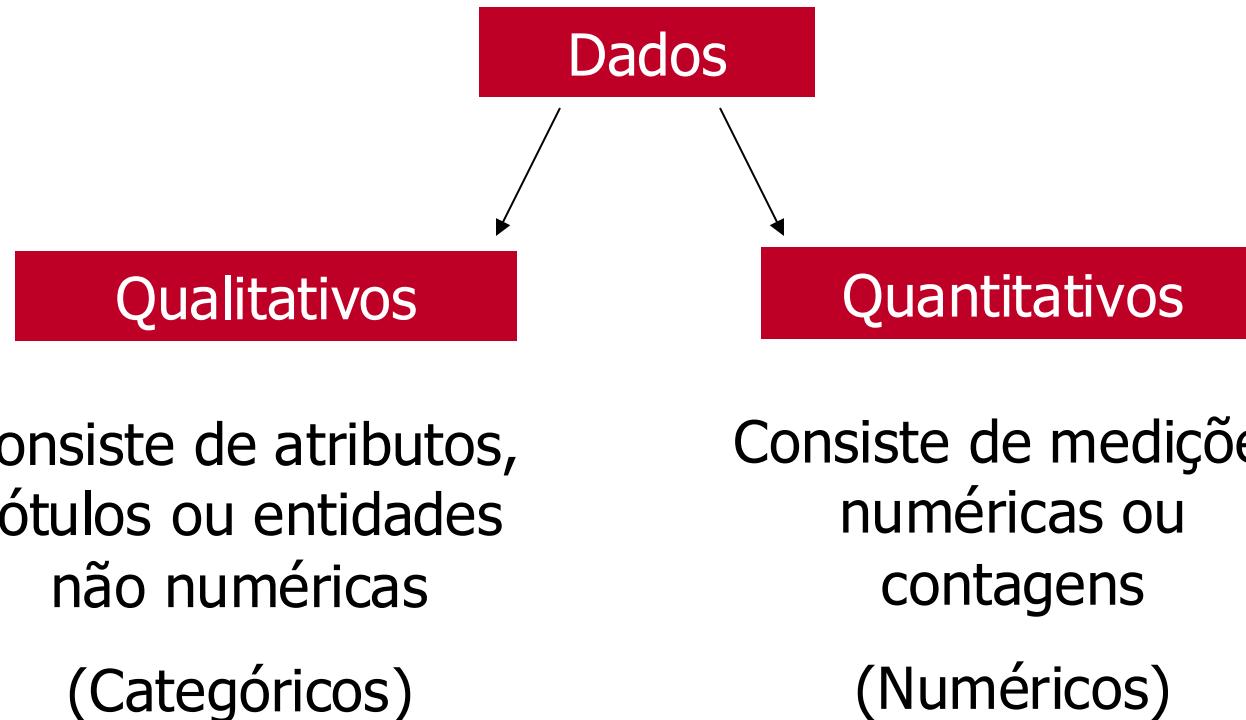
JSON

JavaScript
Object
Notation

Dados

Tipos de Dados

- Os conjuntos de dados são compostos por dois tipos de dados: dados **qualitativos** e dados **quantitativos**



Dados

Tipos de Dados

- Considere o seguinte exemplo
 - Notas de alunos em uma determinada disciplina

Aluno	Nota
Sally	3.22
Bob	3.98
Cindy	2.75
Mark	2.24
Kathy	3.84

Qualitativo **Quantitativo**



Dados Quantitativos

Tipos de Variáveis

Um dado classificado como "**idade**" é **quantitativo**
Ex.: 11, 15, 18, 25, 42 anos.

Entretanto, se esse dado for informado por "**faixa etária**" ele é **qualitativo** (ordinal).

Ex: 0 – 5 anos

6 – 12 anos

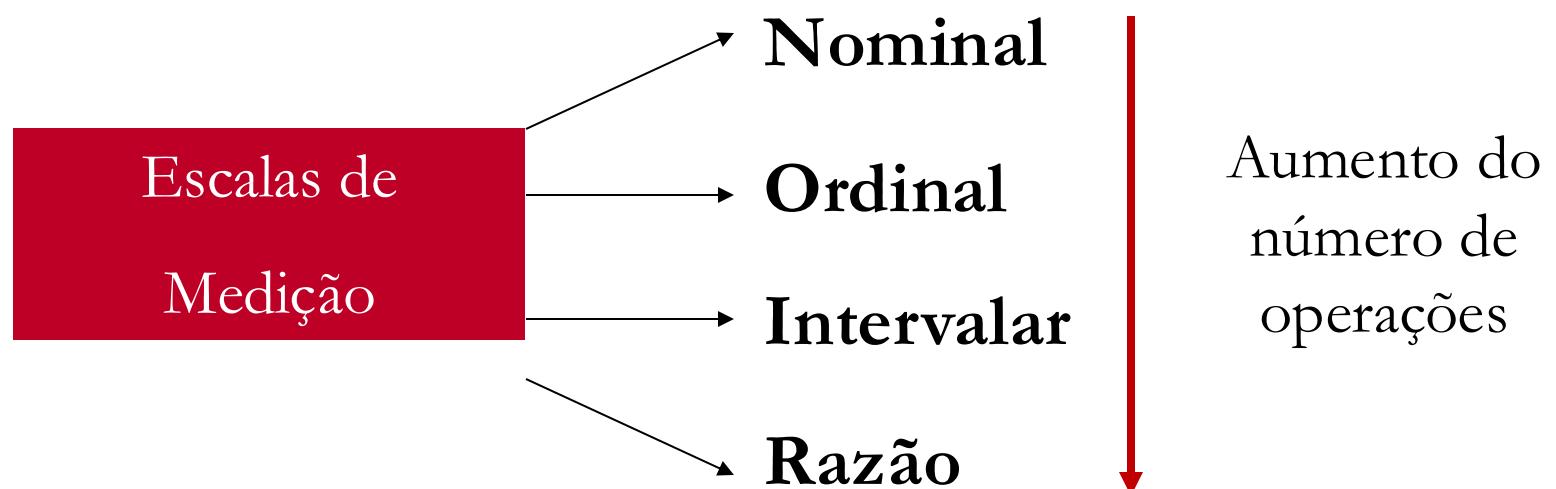
13 – 18 anos

19 – 28 anos

Dados

Escalas de Medição

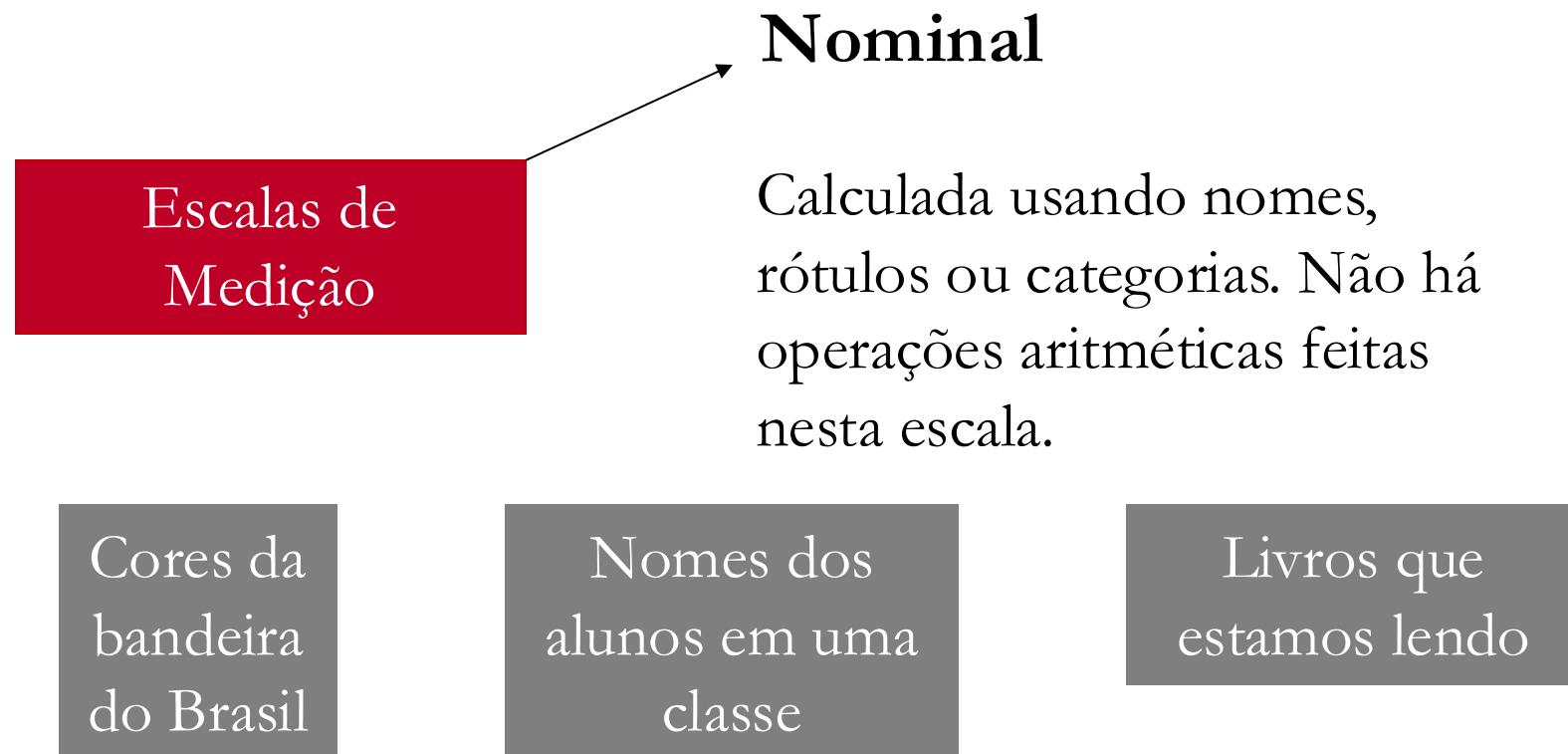
- Há quatro escalas de medição de um dado, quais sejam:



Dados

Escala Nominal

- Representam **categorias** que não mantém necessariamente relação entre elas
- Não é possível realizar operações aritméticas (soma, média, etc.)
- Normalmente realiza-se apenas a contagem das observações em cada categoria



Escala Ordinal

- **Categorias** podem ser representadas por nomes, símbolos ou números, porém há uma ordenação de uma categoria em relação à outra
- A distância entre uma categoria e a outra não pode ser medida numericamente
- Além da operação de contagem, permitem operações que envolvam ordenação (maior/menor)

Escalas de Medição

Ordinal

Podem ser ordenadas entre si, mas não é possível diferenciá-las numericamente.

Nível de
experiência: junior,
pleno e senior

Números das camisas
dos jogadores da
seleção

Top 10 músicas
mais tocadas no
momento

Dados

Escala Intervalar

- Escala **quantitativa**
- O valor nulo não corresponde à ausência da característica medida
- A escala possui um zero arbitrário
- Exemplo: $0\text{ }^{\circ}\text{C}$ não significa ausência de temperatura ($-273\text{ }^{\circ}\text{C}$)
- Operação de divisão é ilegítima em dados intervalares

Escalas de
Medição

Temperatura

Intervalar

Ordenadas, a diferença entre duas medidas pode ser calculada

Linha do tempo em
anos

40 ° não é o dobro de $20\text{ }^{\circ}\text{C}$.

Veja em Fahrenheit:

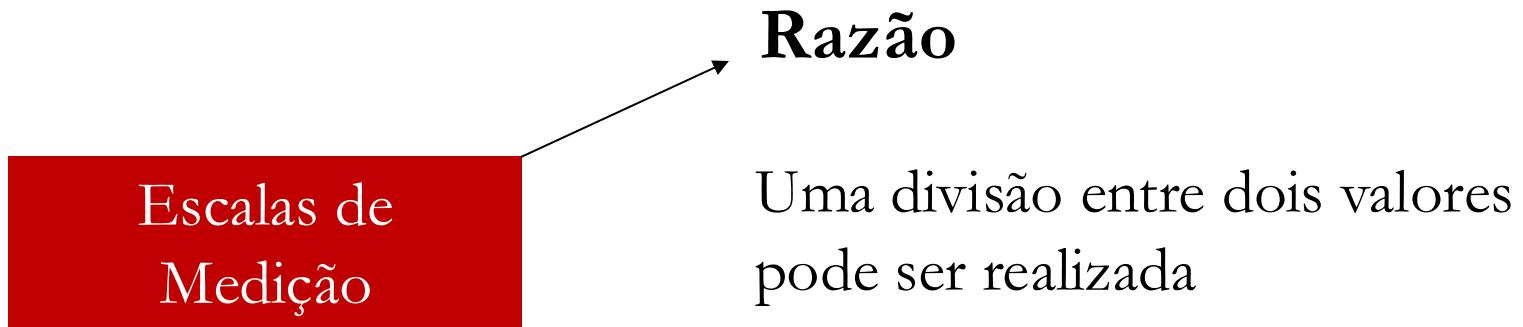
$20\text{ }^{\circ}\text{C} \rightarrow 68^{\circ}\text{F}$

$40\text{ }^{\circ}\text{C} \rightarrow 104\text{ }^{\circ}\text{F}$

Dados

Escala Razão

- Escala **quantitativa**
- O zero corresponde à ausência da característica medida
- É possível realizar todas as operações aritméticas em dados dessa escala



Idade

Peso

Altura

Dinheiro

Análise exploratória de dados

Análise exploratória

Nominal	Ordinal	Discreta	Contínua
<ul style="list-style-type: none">- Cor- Sexo- Estado Civil- Vale transporte (Sim ou Não)- Grau de Escolaridade	<ul style="list-style-type: none">- Escala de questionário- Preferência- Faixa etária- Classe Social	<ul style="list-style-type: none">- Escala de questionário (numérica)- Número de filhos- Quantidade de empregados- Quantidade de ligações	<ul style="list-style-type: none">- Salário Mensal- Idade- Anos de estudo- Taxas

Análise Exploratória

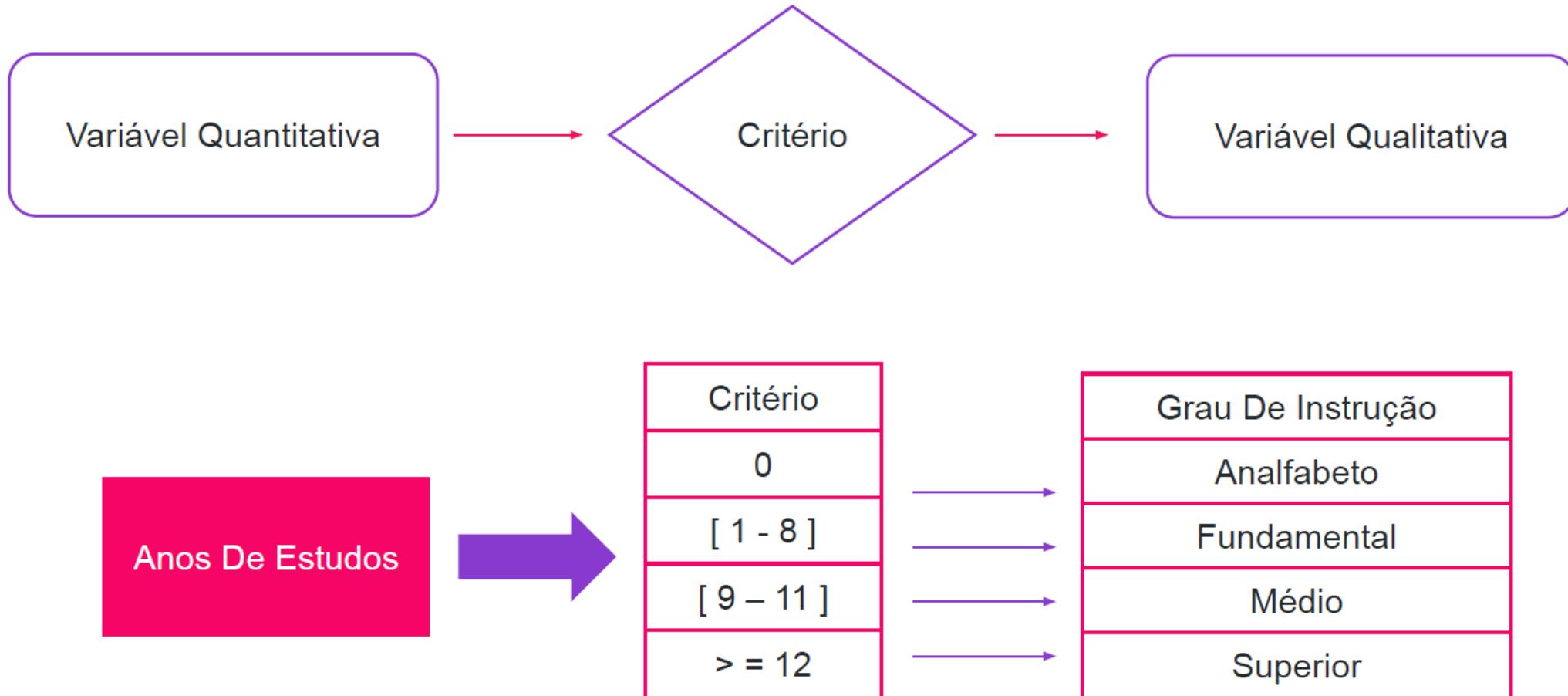
Qualitativas	id	sexo	idade	cor	internet	Telefone móvel	Anos de estudo	Rendimento
	35000015	2	15	2	1	3	7	800
	35000015	2	75	2	3	3	12	2.100
Quantitativa (categóricas)	35000031	2	60	2	3	3	12	1.800
	35000058	2	68	2	3	3	1	
	35000058	2	48	8	3	1	5	1.000
	35000058	2	42	2	3	3	6	900
	35000066	2	36	2	1	3	9	1.000
	35000066	2	44	2	1	1	9	1.200
	35000066	2	20	2	1	3	13	300
	35000066	4	26	2	1	3	12	1.900
	35000074	4	14	2	1	3	8	700
	35000074	4	71	2	3	3	5	450
	35000090	4	20	2	1	1	12	1.890
	35000090	2	19	8	1	3	12	620
	35000090	4	42	2	3	1	12	300
	35000090	4	17	2	1	1	11	1.100
	35000090	4	25	2	1	1	12	433
	35000090	2	49	6	1	3	16	400
	35000104	2	38	2	3	1	6	600

Tipos de variáveis Qualitativas

Os dados podem conter variáveis:

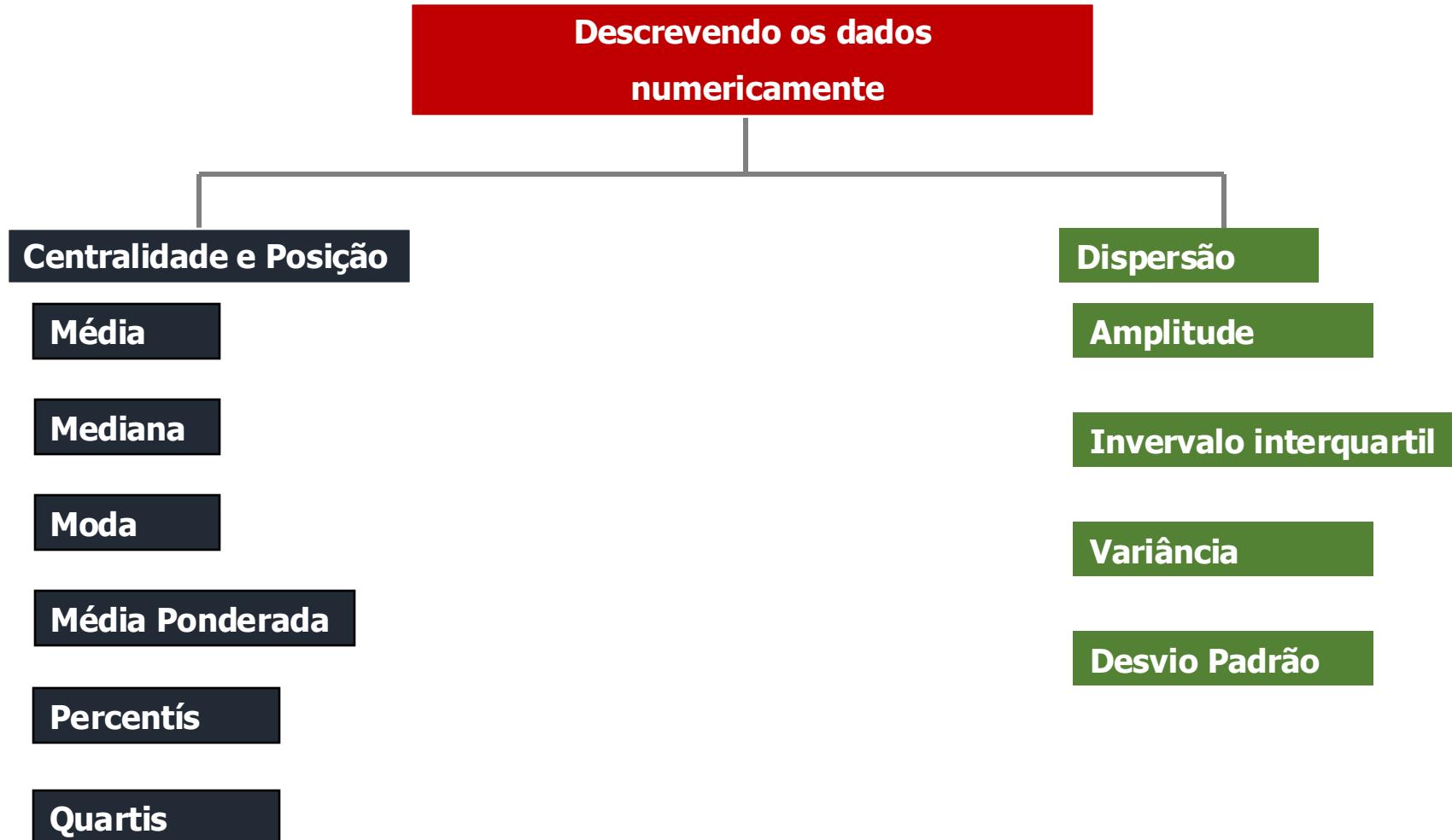
Qualitativas – utilizam termos **descritivos** para descrever algo de interesse. Ex: cor dos olhos, estado civil, religião, sexo, grau de escolaridade, classe social, tipo sanguíneo, cor da pele, etc...

Transformando



Dados Quantitativos

Descrevendo os dados numericamente



Variáveis e observações

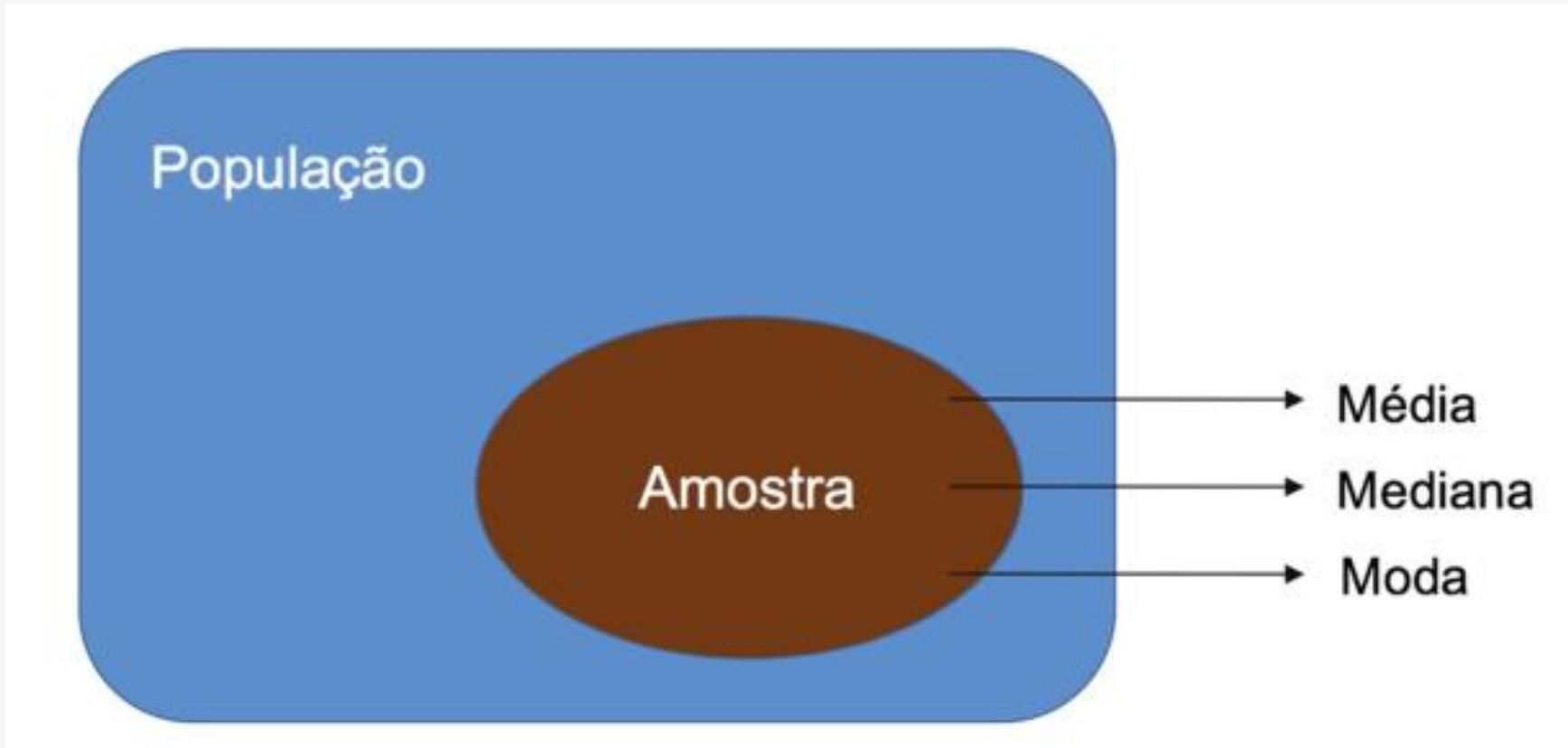
Observações

Variáveis

	Idade	Sexo	Peso	Cor dos olhos
Indivíduo 1	42	M	59	Verde
Indivíduo 2	34	M	54	Castanho
Indivíduo 3	56	F	89	Azul
Indivíduo 4	41	M	76	Castanho
Indivíduo 5	23	F	65	Castanho

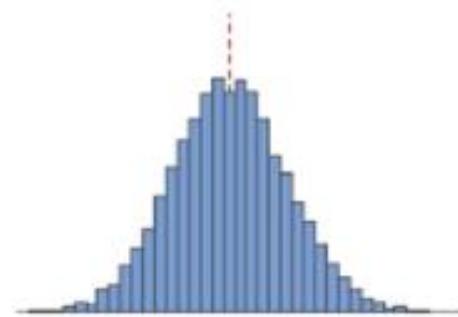
Medidas de posições com Power BI

Medidas de Posição



Medida de posição: Média

Média



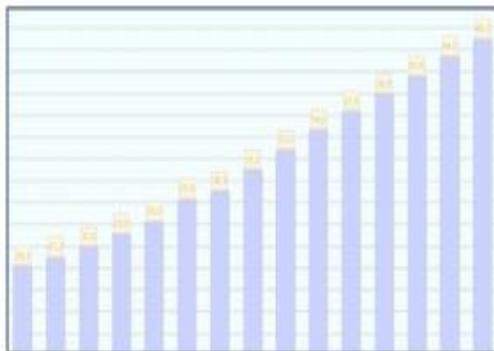
Sem dúvida, **médias** são as formas mais simples de identificar tendências em um conjunto de dados

Cuidado! Outliers podem impactar o valor

Medidas de posição: Mediana

Divide os elementos da amostra ao meio. 50% são menores que a mediana e 50% da amostra maiores.

Mediana



Se o número de elementos n na amostra for ímpar, a Mediana será: $(n + 1) / 2$

Se o número de elementos n na amostra for par, a Mediana será: $(n / 2) + 1$

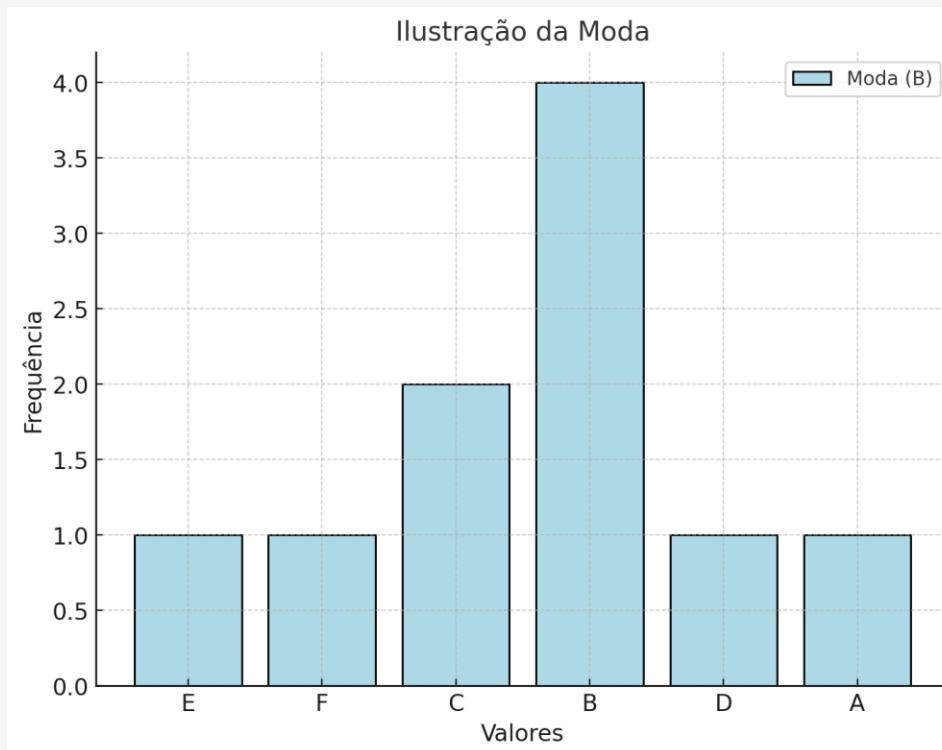
- A **mediana** pode ser mais representativa, pois não é afetada por outliers.

Exemplo:

Se um grupo de 10 pessoas tem salários de R\$ 2.000 a R\$ 5.000, mas um CEO ganha R\$ 100.000, a média será **muito maior** do que a maioria ganha, distorcendo a realidade.

Moda

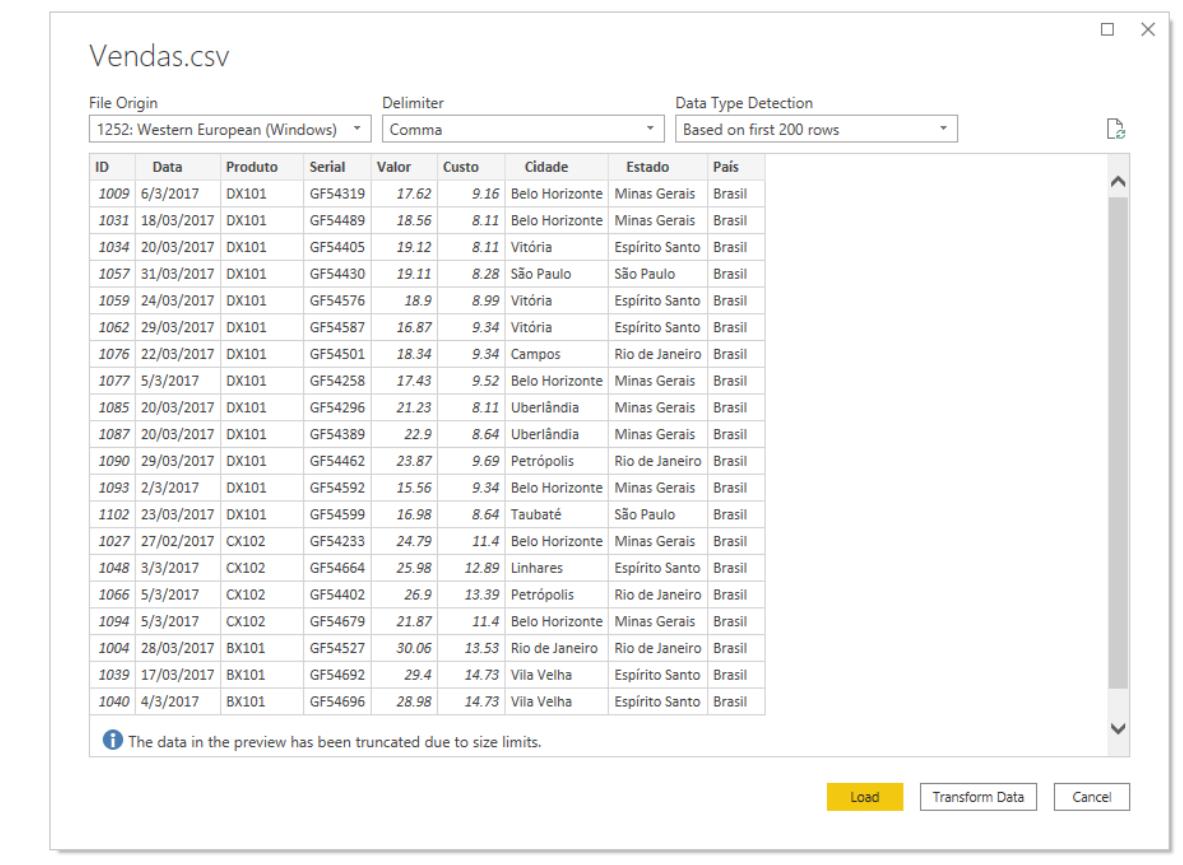
- Valor de maior Frequência na amostra



O gráfico de barras mostra a **frequência** dos valores em um conjunto de dados

- O valor **B** aparece **4 vezes**, sendo o mais frequente
- Esse valor é a **moda**, ou seja, o elemento que mais se repete na distribuição

Calculando posições com Power BI



Carregue o arquivo Vendas.CSV

Clique no data view e crie as medidas (modelagem)

which return only the minimum value when there are multiple modes in the set of values considered. The Excel function MODE.MULT would return all of the modes, but you cannot implement it as a measure in DAX.

```
1 Mode :=  
2 MINX (  
3     TOPN (  
4         1,  
5         ADDCOLUMNS (  
6             VALUES ( Data[Value] ),  
7             "Frequency", CALCULATE ( COUNT ( Data[Value] ) )  
8         ),  
9         [Frequency],  
10        0  
11    ),  
12    Data[Value]  
13 )
```

Substitua por
Vendas[valor]

O código pode ser dividido nas seguintes partes:

- Criar uma tabela com valores únicos e contar a frequência de cada valor
- Ordenar os valores pela frequência e selecionar o mais frequente
- Retornar o menor valor caso haja empate

Visualizando as medidas



Exercício

Use o arquivo com dados de pacientes e crie uma tabela de análise descritiva no Power BI com a média, mediana e moda da idade, altura e peso dos pacientes.

Além disso, responda:

- Qual a moda do tipo sanguíneo?
- Qual a moda do estado civil?

Medidas de posições relativas

Posição relativa

Os dados podem ser medidos em termos de posição relativa, que compara a posição de um valor, em relação a outro valor dentro do conjunto de dados.

Percentil e quartil são as medidas mais comuns de posição relativa

Percentil x Porcentagem

Percentil e Porcentagem **não** são a mesma coisa.



Porcentagem (%): Proporção calculada em relação a uma grandeza de cem unidades. A porcentagem pode ser encontrada multiplicando o valor numérico por 100.

Percentil: É o ponto da distribuição dos resultados ordenados da amostra (por ordem crescente dos dados) em 100 partes de igual amplitude.. Por exemplo, um resultado no percentil 90 significa que 90% dos resultados se situam nesse ponto ou abaixo dele.

Percentil

A maneira mais fácil de informar a posição relativa é por meio do uso do **percentil**

Percentil



Percentil x Porcentagem

Suponha que um aluno tenha conseguido nota **36** em um exame de admissão em uma universidade, cujo valor máximo era **45**

Supondo que além de informar a você que o aluno conseguiu nota 36, eu dissesse que ele ficou em

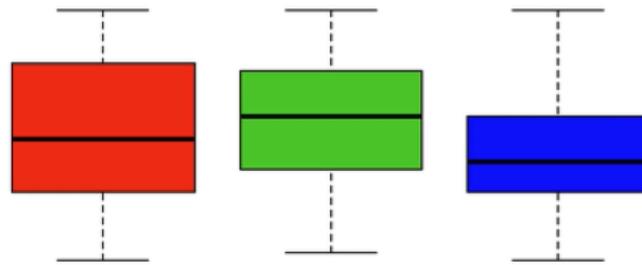
97º percentil

Isso significa que o aluno foi melhor que **97%** dos outros alunos que prestaram o mesmo exame

Perceba que se dividirmos **36/45**, o aluno teve um aproveitamento de **80%**

Esta informação **NÃO** é a mesma coisa que o **percentil**

Quartil



Quartil é simplesmente um específico percentil de interesse

Quartis são valores que dividem uma tabela de dados em quatro partes iguais:

O **primeiro quartil** é o valor que constitui **25% percentil**.

O **segundo quartil** é o valor que constitui **50% percentil**.

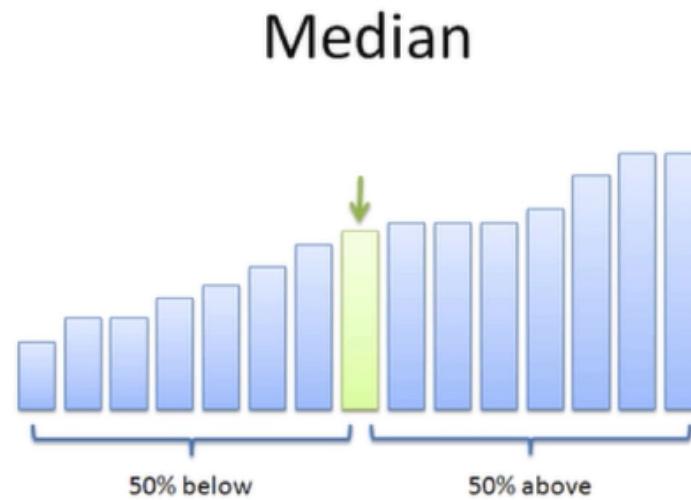
O **terceiro quartil** é o valor que constitui **75% percentil**.

O **quarto quartil** é o valor que constitui **100% percentil**.

Quartil e mediana

Perceba que o **segundo quartil** é a **mediana**, ou seja,

50º percentil



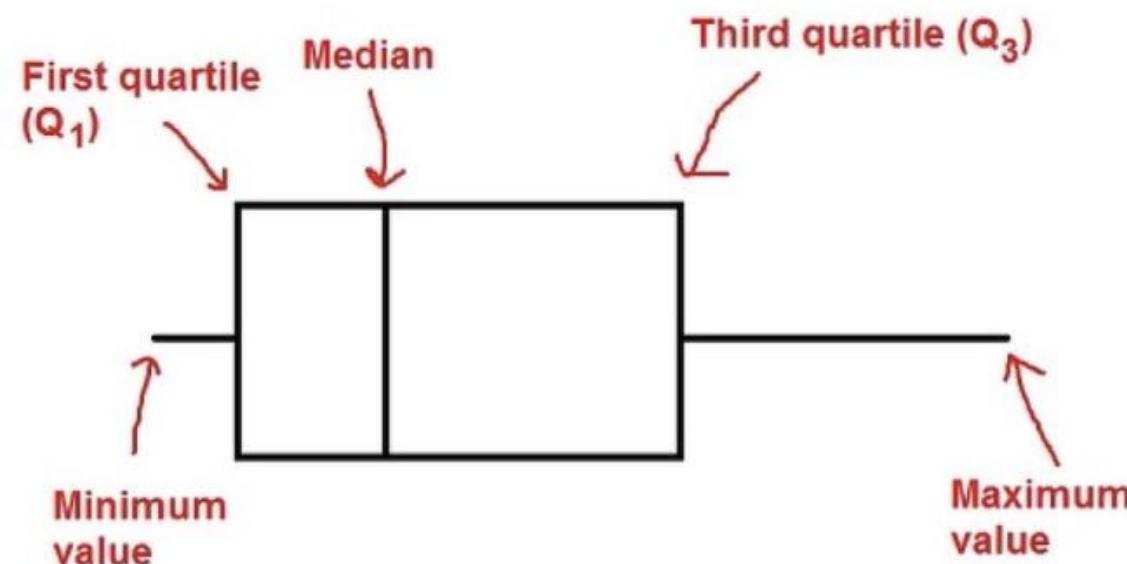
Quartil - intervalos

Temos ainda os intervalos interquartis:

- Intervalo interquartil $\rightarrow Q_3 - Q_1$
- Intervalo semi-interquartil $\rightarrow (Q_3 - Q_1)/2$
- Quartil médio $\rightarrow (Q_3 + Q_1)/2$

Quartil e BoxPlot

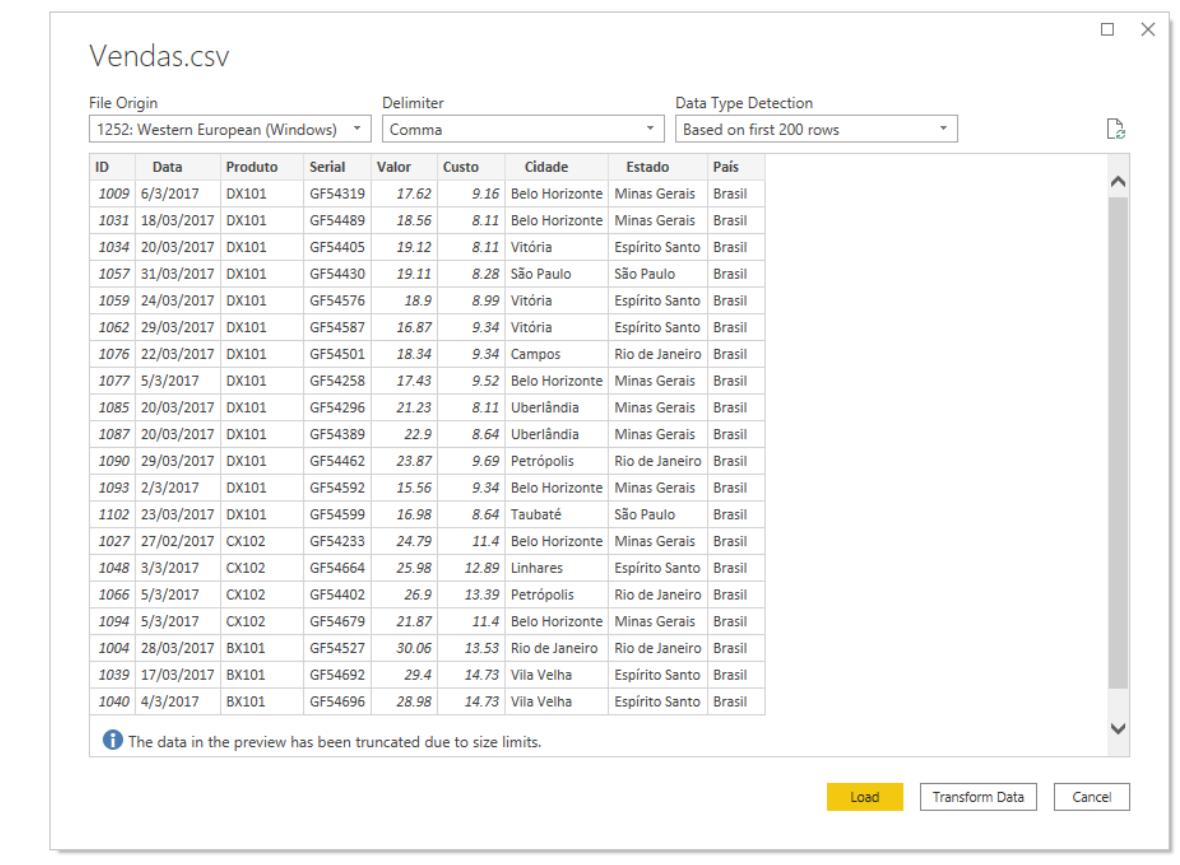
Os intervalos interquartis são fundamentais para saber interpretar um boxplot:



Percentis em Dax

- Max_Value = MAX (Vendas[Valor])
- Q3_Value = PERCENTILEX.INC (Vendas, Vendas[Valor], 0.75)
- Median_Value = MEDIAN (Vendas[Valor])
- Q1_Value = PERCENTILEX.INC (Vendas,Vendas[Valor], 0.25)
- Min_Value = MIN (Vendas[Valor])

Usando percentil com PBI



Carregue o arquivo Vendas.CSV

Clique no data view e crie as medidas (modelagem)

Calculando a media das 20% maiores vendas

Calcule a media das vendas entre as top 20, ou seja, 20% das maiores vendas, calcule a media.

ID	Data	Produto	Serial	Valor	Custo	Cidade	Estado	País
1009	6/3/2017	DX101	GF54319	\$17.62	\$9.16	Belo Horizonte	Minas Gerais	Brasil
1031	18/03/2017	DX101	GF54489	\$18.56	\$8.11	Belo Horizonte	Minas Gerais	Brasil
1034	20/03/2017	DX101	GF54405	\$19.12	\$8.11	Vitória	Espírito Santo	Brasil

Retorna tudo
EXCluindo o
percentil

Métodos gráficos

- ➡ • Tabela de Frequência
- ➡ • Tabela de Contingência
- ➡ • Gráficos de Linhas
- ➡ • Gráficos de Barras
- ➡ • Gráfico de Pareto
- ➡ • Histogramas
- ➡ • Gráficos de Caixa (boxplots)
- ➡ • Diagramas de dispersão
- ➡ • Gráfico Temporal
- ➡ • Ogiva (frequência cumulativa)
- ➡ • Ramo e folhas
- ➡ • Gráficos de Pontos
- ➡ • Gráfico de Quartis

Métodos Gráficos ou Tabulares



Tabela de frequência

- Medidas de posição:
 - Média, moda e mediana
- Medidas de dispersão (quanto dispersos em relação à medida central):
 - Desvio padrão e variância
- Estas medidas são boas em amostras e para selecionar uma amostra é necessário usar uma Tabela de Frequência

A Tabela de Frequência indica a frequência observada, ou seja, mostra a frequência com que cada observação aparece nos dados.

Tabela de frequência

Para descrevermos um conjunto de dados, definiremos o que são classes de frequência, isto é, intervalos da variável de interesse, e verificaremos o número de dados neste intervalo.

Isso nos dá a Distribuição de Frequência, que é a associação das frequências aos valores obtidos correspondentes.

Para criar uma tabela de frequência, precisamos definir:

- Número de classes
- Amplitude das classes
- Ponto inicial

Tabela de frequência

A frequência pode ser:

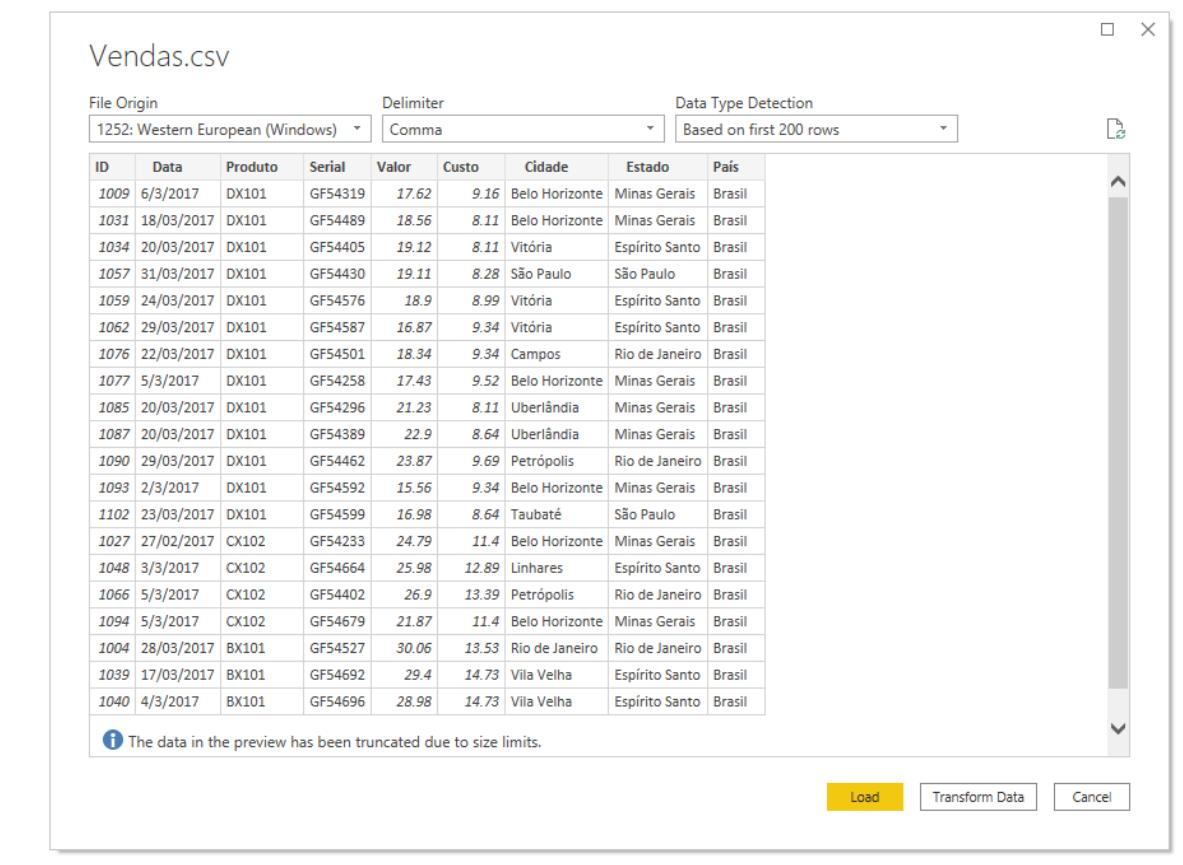
- Absoluta
- Relativa

<i>Exercício</i>	<i>frequência absoluta</i>	<i>frequência relativa</i>
nenhum	185	40,04%
mudando	213	46,10%
baixo/ moderado	49	10,61%
alto	15	3,25%

Frequência Acumulada

É o total acumulado (soma) de todas as classes anteriores até a classe atual.

Criando tabela de frequênciа com o Power BI



Carregue o arquivo Vendas.CSV

Clique no data view e crie as medidas (modelagem)

Tabela de frequência de vendas por estado

- Insira a visualização TABELA
- Insira os campos estado e valor
- Veja que as estatísticas (media, variância e mediana) podem ser obtidas facilmente através das propriedades dos campos

The screenshot shows a Power BI interface with a table visualization and its context menu open.

Table Visualization:

Estado	Valor
Espírito Santo	\$991.74
Minas Gerais	\$1,183.49
Rio de Janeiro	\$1,267.48
São Paulo	\$2,091.64
Total	\$5,534.35

Context Menu (Open at Total row):

- Remove field
- Rename
- Move
- Conditional formatting
- Remove conditional formatting
- Don't summarize
- Sum** (selected)
- Average
- Minimum
- Maximum
- Count (Distinct)
- Count
- Standard deviation
- Variance
- Median
- Show value as
- New quick measure

Tabela de frequência de vendas por estado

- Altere a medida para CONTAGEM
- Altere para visualizar em percentual

Estado	Count of Valor
Espírito Santo	21
Minas Gerais	24
Rio de Janeiro	25
São Paulo	36
Total	106

Frequência
absoluta

The screenshot shows the Power BI interface with a context menu open over a 'Count of Valor' measure. The menu includes options like 'Remove field', 'Rename', 'Move', 'Conditional formatting', 'Remove conditional formatting', 'Don't summarize', 'Sum', 'Average', 'Minimum', 'Maximum', 'Count (Distinct)', 'Count' (which is checked), 'Standard deviation', 'Variance', and 'Median'. At the bottom of the menu, there are two options: 'Show value as' (with 'Percent of grand total' selected) and 'New quick measure'.

Estado	%GT Count of Valor
Espírito Santo	19.81%
Minas Gerais	22.64%
Rio de Janeiro	23.58%
São Paulo	33.96%
Total	100.00%

Frequência
relativa

Configurando o Colab para linguagem R

- Ambiente de Execução

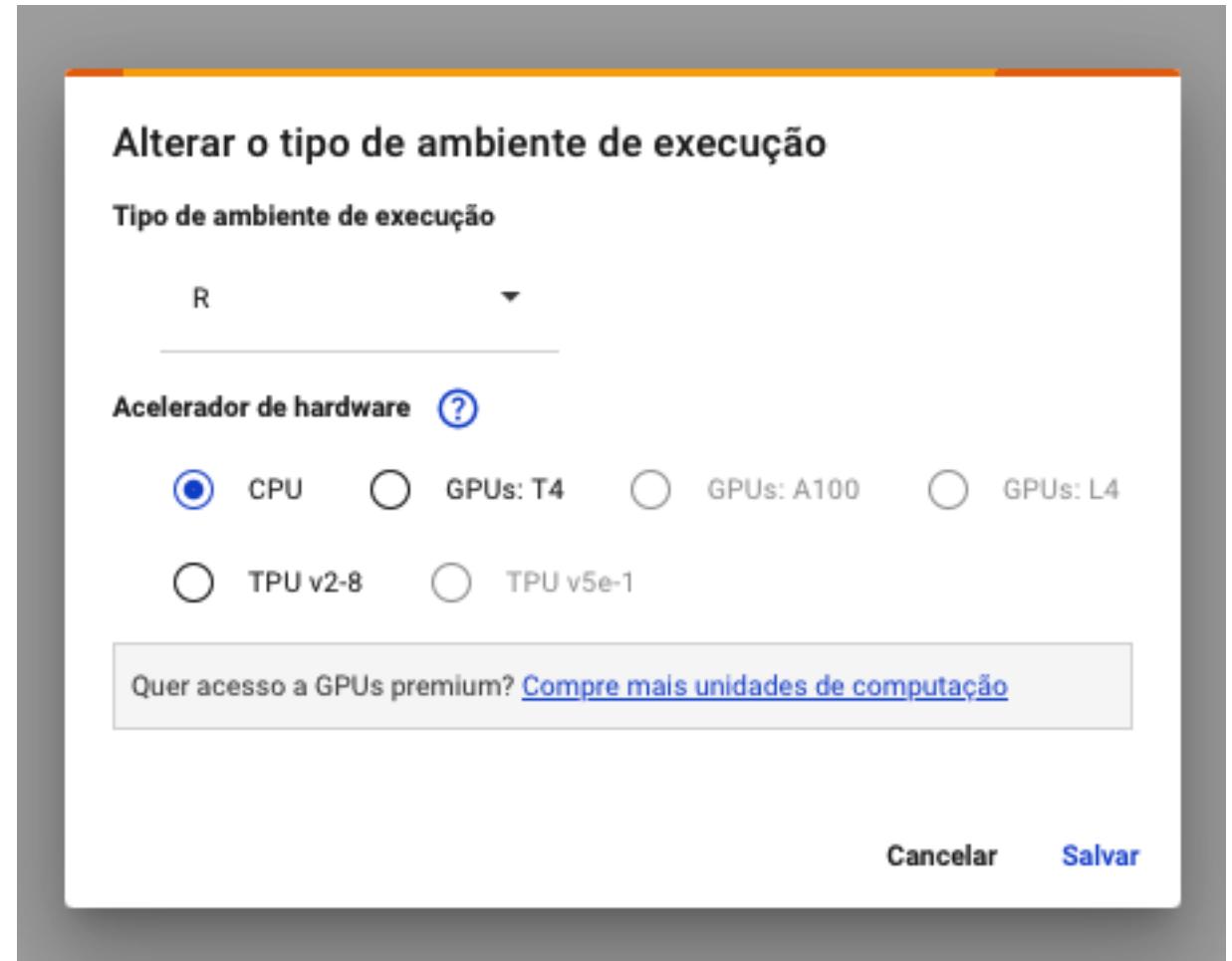


Tabela de frequência em R

```
1 work.dir <- "C:/dados"
2 setwd(work.dir)
3 # carrega o arquivo na memória - dec= separador de decimais sep = separador de colunas
4 # h = cabeçalho
5 # encoding é o mesmo usado no PowerBI
6 usuarios <- read.table("usuarios.csv",dec=". ", sep=",", h= T, fileEncoding = "windows-1252")
7
8 #nomes das colunas da tabela
9 names(usuarios)
10 str(usuarios)
11 summary(usuarios$salario)
12
13 #tabela de frequencia absoluta
14 freq <- table(usuarios$grau_instrucao)
15 freq
16
17 #tabela de frequencias relativas
18 freq_rel <- prop.table(freq)
19 freq_rel
20
21 #melhorando a apresentacao, em porcentagem
22 p_freq_rel <- prop.table(freq) * 100
23 p_freq_rel
24
25 # criando a tabela e incluindo a linha de total em cada coluna
26 freq <- c( freq, sum(freq))
27 freq_rel <- c( freq_rel, sum(freq_rel))
28 p_freq_rel <- c(p_freq_rel, sum(p_freq_rel))
29 names(freq)[4] <- "Total"
30
31 #agora, juntando todos os campos acima em uma unica tabela
32 tabela_final <- cbind(freq, freq_rel, p_freq_rel)
33 tabela_final
34
35 #deixando apenas mais bonitinha, com formatacao
36 tabela_final <- cbind(freq,
37                         freq_rel = round(freq_rel, digits=2),
38                         p_freq_rel = round(p_freq_rel, digits=2))
```

Medidas de dispersão

Medidas de dispersão

Uma maneira de descrever um conjunto de dados, é através de **medidas de dispersão**. Elas descrevem a amplitude dos dados, ou seja, quanto espalhados os dados estão dentro de um conjunto.

Medidas de dispersão: variância

A **variância** mede a amplitude (variabilidade) dos dados em relação à média.

Medidas de dispersão: desvio padrão

O **desvio padrão** é usado para medir a variabilidade entre os números em um conjunto de dados. Assim como o nome sugere, o desvio padrão é um padrão de desvio (distância) da média.

É a raíz quadrada da variância

Exemplo: Anderson x Patrícia

Anderson – cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	100
Disciplina 2	100
Disciplina 3	100
Disciplina 4	50
Disciplina 5	50
Disciplina 6	50

Média final = 75

Média para aprovação = 60

Exemplo: Anderson x Patrícia

Patrícia – também cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	75
Disciplina 2	74
Disciplina 3	76
Disciplina 4	77
Disciplina 5	75
Disciplina 6	74

Média final = 75

Qual a
diferença
destas 2
distribuições?

Veja a variabilidade da amostra

Resultado do Anderson:

Média Amostra =	75
Variância da Amostra =	750
Desvio Padrão da Amostra =	27.39

As notas relevam comportamentos diferentes de estudo.

Mais variabilidade nas notas.

O desvio padrão foi alto, entre 27 pontos para mais ou para menos.

Terá de cursar mais 3 disciplinas para receber o diploma final.

Resultado da Patrícia

Média Amostra =	75
Variância da Amostra =	1.37
Desvio Padrão da Amostra =	1.17

As notas relevam comportamentos diferentes de estudo.

Manteve a variabilidade baixa com notas mais uniformes.

O desvio padrão foi baixo, entre 1,17 pontos para mais ou para menos.

Receberá o diploma do curso ao final do semestre.

Exemplo Anderson x Patricia em R

```
> #Exemplo Anderson x patricia
> #criando dataset Anderson
> anderson <- c(100,100,100,50,50,50)
> #criando dataset patricia
> patricia <- c(75,74,76,77,75,74)
> #calculando a média
> mean(anderson)
[1] 75
> mean(patricia)
[1] 75.16667
> #calculando a variância
> var(anderson)
[1] 750
> var(patricia)
[1] 1.366667
> #calculando o desvio padrao
> sd(anderson)
[1] 27.38613
```

Identificando outliers

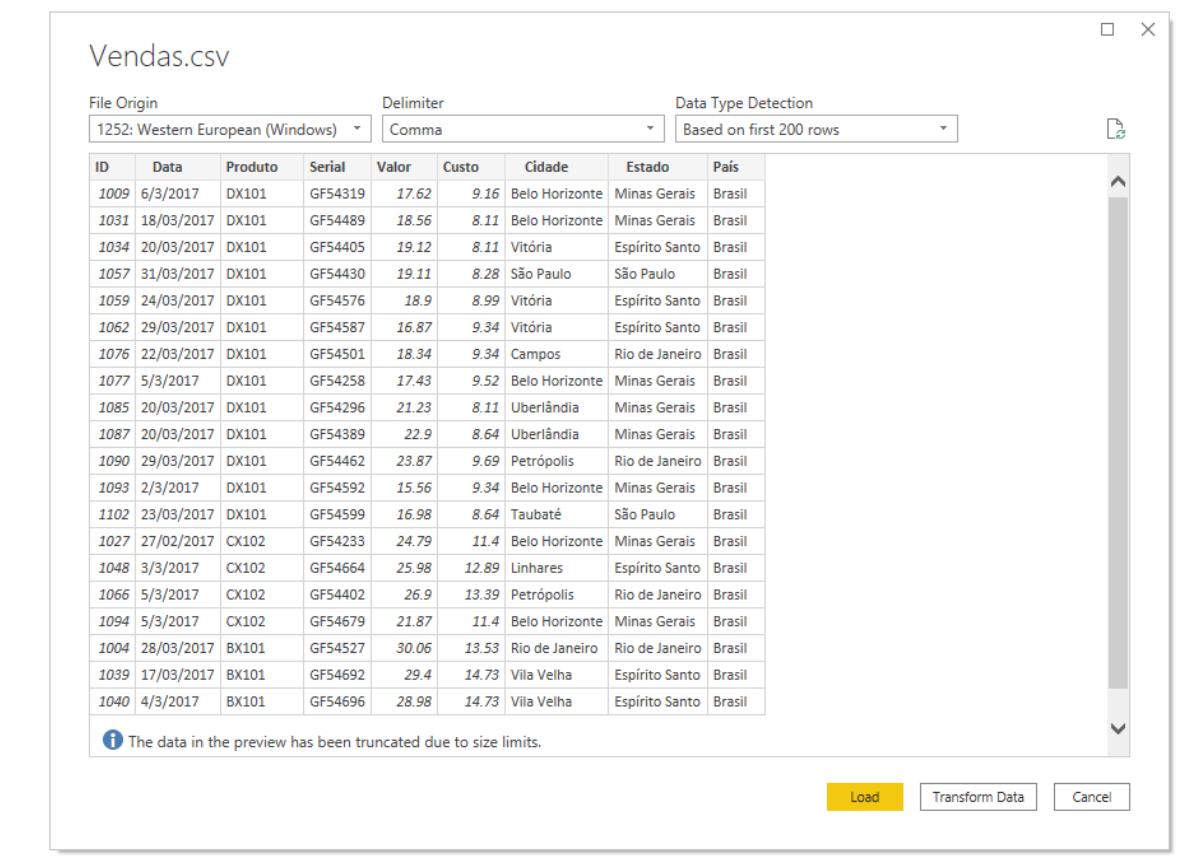
```
vetor <- c(501,504,493,499,497,503,525,495,506,502)
mean(vetor) #media
sort(vetor) #classifica para melhorar a localização do valor do meio
median(vetor) #mediana
sd(vetor) # desvio padrão, quanto mais próximo de zero, mais homogêneo
var(vetor) #variância não usa a mesma unidade dos dados
max(vetor) #valor máximo
min(vetor) # valor mínimo
summary(vetor) # localize os quartis
quantile(vetor) # exibe somente os quartis
plot(vetor) # exibe o gráfico de dispersão
lines(vetor, plot=TRUE) #gráfico de linhas
abline(mean(vetor),0, col="red") # linha do valor medio
abline(median(vetor),0, col="blue")
abline(max(vetor),0, col="purple")
abline(min(vetor),0, col="purple3")
quartis <- quantile(vetor)
abline(quartis[[2]],0, col="green1")
abline(quartis[[4]],0, col="green3")
amplitude <- quartis[[4]]-quartis[[2]]
limsup <- mean(vetor)+ 1.5 * amplitude #interpolação para encontrar limites
liminf <- mean(vetor)- 1.5 * amplitude
print(liminf)
print(limsup)
abline(limsup,0, col="red3")
abline(liminf,0, col="red3")

# Como o valor 525 é superior a 511,8, podemos afirmar com um alto grau de certeza de que esse ponto é um outlier.
```

Verifique se há outliers na base da vazão do Rio Nilo

```
str(Nile)
print(Nile)
vazao_nilo <- Nile
plot(vazao_nilo, xlab="Ano", ylab="Vazão do Rio Nilo",
main="Rio Nilo", las=1, yaxt="n")
axis(side=1, at=seq(1850, 1970, 20))
axis(side=2, at=seq(400, 1400, 100), las=1)
```

Calculando posições com Power BI



Carregue o arquivo Vendas.CSV

Clique no data view e crie as medidas (modelagem)

Desvio padrão

Standard Deviation

You can use standard DAX functions to calculate the standard deviation of a set of values.

- **STDEV.S**: returns the standard deviation of values in a column representing a sample population.
- **STDEV.P**: returns the standard deviation of values in a column representing the entire population.
- **STDEVX.S**: returns the standard deviation of an expression evaluated over a table representing a sample population.
- **STDEVX.P**: returns the standard deviation of an expression evaluated over a table representing the entire population.



The screenshot shows the Power BI Data View ribbon with three tabs: Structure, Formatting, and Properties. The Structure tab is active, displaying a table with columns: ID, Data, Produto, Serial, Valor, Custo, Cidade, Estado, and País. The table has three rows of data. The Properties tab is also visible on the right side of the ribbon.

ID	Data	Produto	Serial	Valor	Custo	Cidade	Estado	País
1009	6/3/2017	DX101	GF54319	\$17.62	\$9.16	Belo Horizonte	Minas Gerais	Brasil
1031	18/03/2017	DX101	GF54489	\$18.56	\$8.11	Belo Horizonte	Minas Gerais	Brasil
1024	20/03/2017	DX101	GF54408	\$10.10	\$0.11	Vitória	Espírito Santo	Brasil

Fonte: daxpatterns.com

Variância

Variance

You can use standard DAX functions to calculate the variance of a set of values.

- **VAR.S**: returns the variance of values in a column representing a sample population.
- **VAR.P**: returns the variance of values in a column representing the entire population.
- **VARX.S**: returns the variance of an expression evaluated over a table representing a sample population.
- **VARX.P**: returns the variance of an expression evaluated over a table representing the entire population.

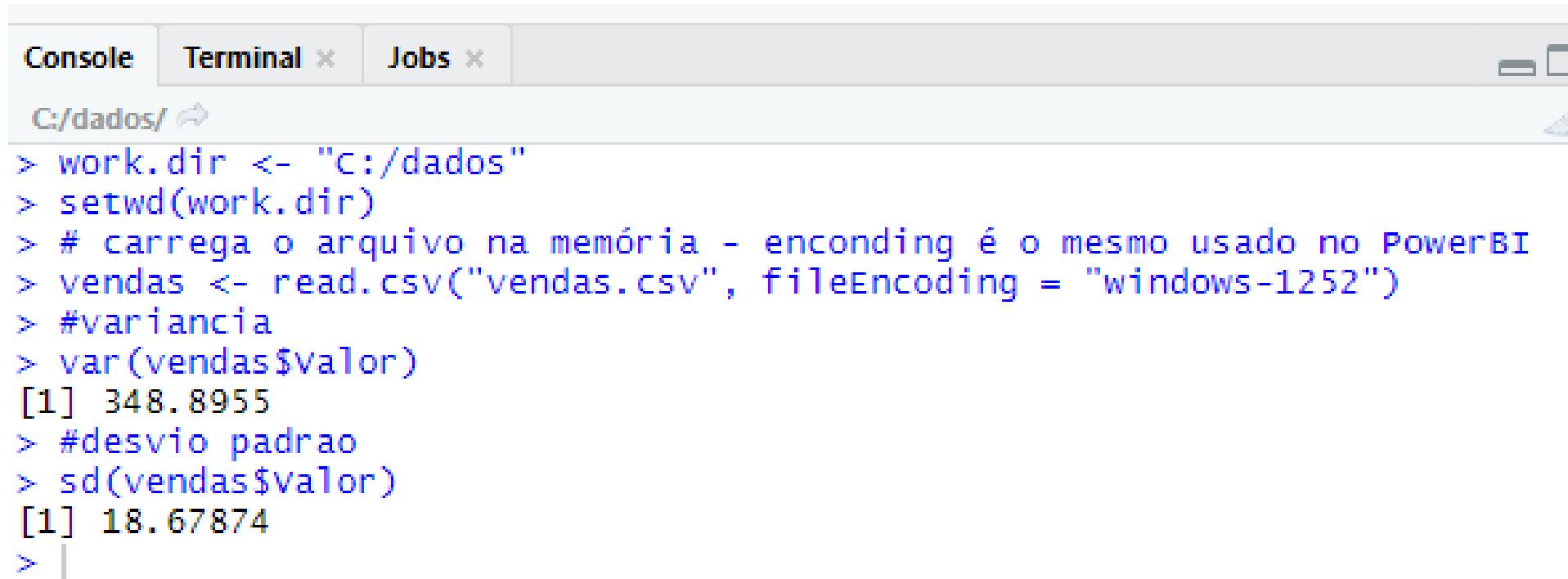
ID	Data	Produto	Serial	Valor	Custo	Cidade	Estado	País
1009	6/3/2017	DX101	GF54319	\$17.62	\$9.16	Belo Horizonte	Minas Gerais	Brasil
1031	18/03/2017	DX101	GF54489	\$18.56	\$8.11	Belo Horizonte	Minas Gerais	Brasil
1034	20/03/2017	DX101	GF54405	\$19.12	\$8.11	Vitória	Espírito Santo	Brasil
1057	21/03/2017	DX101	GF54420	\$10.11	\$0.10	Porto Alegre	Rio Grande do Sul	Brasil

Fonte: daxpatterns.com

Resultado no Power BI



Variância e desvio padrão em R



The screenshot shows the RStudio interface with the 'Console' tab selected. The console window displays the following R code and its execution results:

```
C:/dados/
> work.dir <- "C:/dados"
> setwd(work.dir)
> # carrega o arquivo na memória - encoding é o mesmo usado no PowerBI
> vendas <- read.csv("vendas.csv", fileEncoding = "windows-1252")
> #variancia
> var(vendas$valor)
[1] 348.8955
> #desvio padrao
> sd(vendas$valor)
[1] 18.67874
> |
```

Coeficiente de variação

O **coeficiente de variação** (CV), mede o desvio padrão em termos de percentual da média. Um CV alto, indica alta variabilidade dos dados, ou seja, menos consistência dos dados. Um CV menor, indica mais consistência dentro do conjunto de dados.

Quando comparamos a consistência entre 2 conjuntos de dados em relação a suas médias, é melhor feito quando utilizamos **coeficiente de variação**.

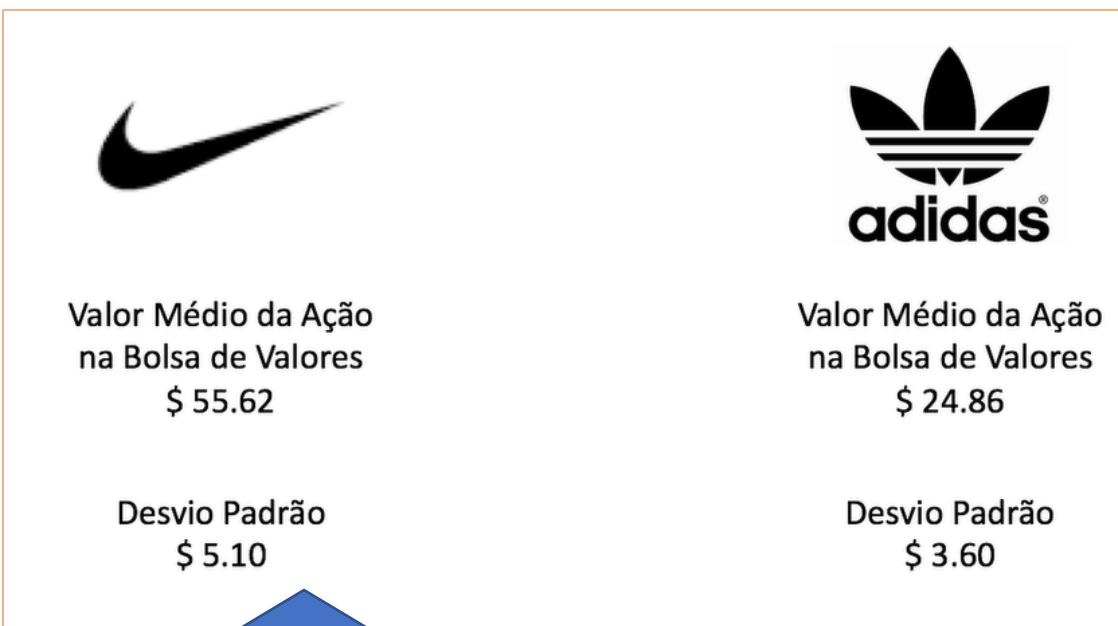
Como calculamos o Coeficiente de Variação = CV

$$CV = \frac{s}{x} \times 100$$

Onde: **S** = Desvio Padrão
X = Média

Exemplo: coeficiente de variação

- Quais ações você compraria?



Os seguintes dados foram coletados:



Nike → $CV = S / x (100) = \$5.10 / \$55.62 (100) = 9.2\%$



Adidas → $CV = S / x (100) = \$3.60 / \$24.86 (100) = 14.5\%$

Dados mais consistentes

Um investidor se sentiria mais seguro em adquirir ações da **Nike**, pois o preço das ações teria uma variação menor, podendo assim evitar perdas e permitindo ao investidor ter um investimento mais seguro.

Análise de correlação

- Permite estudar a relação entre 2 conjuntos de valores
- Quantifica-se o quanto um conjunto de valores está relacionado com outro
- Determina a intensidade e a direção dessa relação
- Com que intensidade os valores de uma variável aumentam ou diminuem enquanto os valores da outra variável aumenta ou diminuem?

Exemplo para correlação

- Um restaurante deseja comparar a quantidade de vendas do prato Feijoada (FJ) e da bebida caipirinha (CP), no decorrer dis nesses de novembro a julho, que ocorrem conforme a tabela:

Mês	11	12	1	2	3	4	5	6	7
FJ	24	27	29	30	32	58	64	64	65
CP	10	25	20	36	28	38	50	60	69

- A correlação encontrada é alta e positiva.
- Quando a feijoada tem mais procura,
- Vende-se mais caipirinha.

```
Console ~/ ↵
> fj<- c(24,27,29,30,32,58,64,64,65)
> cp<-c(10,25,20,36,28,38,50,60,69)
> cor(fj,cp)
[1] 0.8949842
> |
```

Dados Faltantes

- R indica um valor faltante como NA (not available ou indisponível)
`capacity <- c(14, 13, 14, 13, 16, NA, NA, 20, NA)`
 - Três veículos são vans e o termo capacidade de bagagem não se aplica a eles
`mean(capacity)`
- Para encontrar a média é preciso remover os NA
`mean(capacity, na.rm=TRUE)`

`is.na(capacity)`

Moda em R

- R não fornece uma função para encontrar a moda.
- Teste mode()
- Use mfv() do pacote modeest

```
install.packages("modeest")
library(modeest)
scores <- c(1, 2, 2, 2, 3, 4, 4, 4, 5, 6)
moda <- mfv(scores)
print(moda)
```

Z-Score

- Um número isolado não fornece muitas informações
- Para comparações, é necessário estar na mesma escala
- Podemos usar a média e o desvio-padrão para padronizar um conjunto, usar a média como ponto zero e seu desvio-padrão como medida

Se a média das notas for **77.5** e o desvio padrão for **15.14**, um aluno com nota **90** terá um **Z-Score** de aproximadamente **0.82**. Isso significa que essa nota está **0.82 desvios padrão acima da média**.

- **Z-Score positivo (+)** → O valor está acima da média.
- **Z-Score negativo (-)** → O valor está abaixo da média.
- **Z-Score próximo de 0** → O valor está muito próximo da média.

```
notas <- c(55, 60, 65, 70, 75, 80, 85, 90, 95, 100)
media_notas <- mean(notas)
print(media_notas)
desvio_padrao <- sd(notas)
print(desvio_padrao)
z_scores <- (notas - media_notas) / desvio_padrao
df <- data.frame(Notas = notas, Z_Score = z_scores)
print(df)
```

FiveNum

- A função `fivenum()` em R é utilizada para calcular o **resumo de cinco números** de um conjunto de dados numéricos
- fornece uma visão geral da distribuição dos dados através de cinco estatísticas-chave:
 1. **Valor mínimo**: O menor valor no conjunto de dados.
 2. **Primeiro quartil (Q1)**: Também conhecido como **lower-hinge**, é o valor abaixo do qual 25% dos dados se encontram
 3. **Mediana (Q2)**: O valor central que divide o conjunto de dados em duas metades iguais
 4. **Terceiro quartil (Q3)**: Conhecido como **upper-hinge**, é o valor abaixo do qual 75% dos dados se encontram
 5. **Valor máximo**: O maior valor no conjunto de dados

```
dados <- c(7, 15, 36, 39, 40, 41, 42, 43, 47, 49, 50, 51, 52, 73, 80, 81, 82, 83, 85, 87)
resumo <- fivenum(dados)
print(resumo)
boxplot(dados, main = "Boxplot dos Dados", ylab = "Valores")
```

Qual a diferença entre `fivenum` e `summary`?