

CIÊNCIA DE DADOS EDUCACIONAIS

Tema: Análise Estatística

Renata Pitta Barros
Maio 2019



Renata Pitta Barros

CIÊNCIA DE DADOS EDUCACIONAIS

TEMA: ANÁLISE ESTATÍSTICA DE DADOS

- Introdução
- Divisão da Estatística
- Características dos dados
- Pré-processamento dos dados
- Distribuição dos dados





Motivação

O uso crescente de ambientes
de aprendizagem baseados
na web

Os dados educacionais tem origem em todos os lugares



Arquivos de dados

(XML, CSV, Excel, JSON)



Banco de dados

(MySQL, Oracle, MongoDB ...)



API



Sites



Textos e Relatórios



Mapas



Imagens e Vídeos



Mídias Sociais

NECESSIDADE

- Feedback para os alunos
- E métodos de avaliação adequados para esse tipo de ambiente

VANTAGENS

- Captura é automática sobre os registros das interações dos alunos
- A coleta contínua de dados, que permite visões longitudinais do comportamento do aprendizado do aluno
- Captura o que realmente os alunos fizeram, em vez de apenas instantes que são fornecidos em outros técnicas de coletas de dados

DESAFIO

- É analisar esse enorme volume de dados
- Nesses dados estão os registros do comportamento do aprendizado dos alunos
- Tem informações do que foi feito, quando foi feito e quanto tempo durou

TÉCNICAS INDUTIVAS



Focado nos métodos de análises, tais como: classificação, regressão, agrupamento e análise de associação.

O seu objetivo é fazer descobertas sobre o comportamento dos estudantes e o ambiente no qual a aprendizagem ocorre, fornecendo insumos para o professor ou aluno investigar eventuais padrões descobertos.

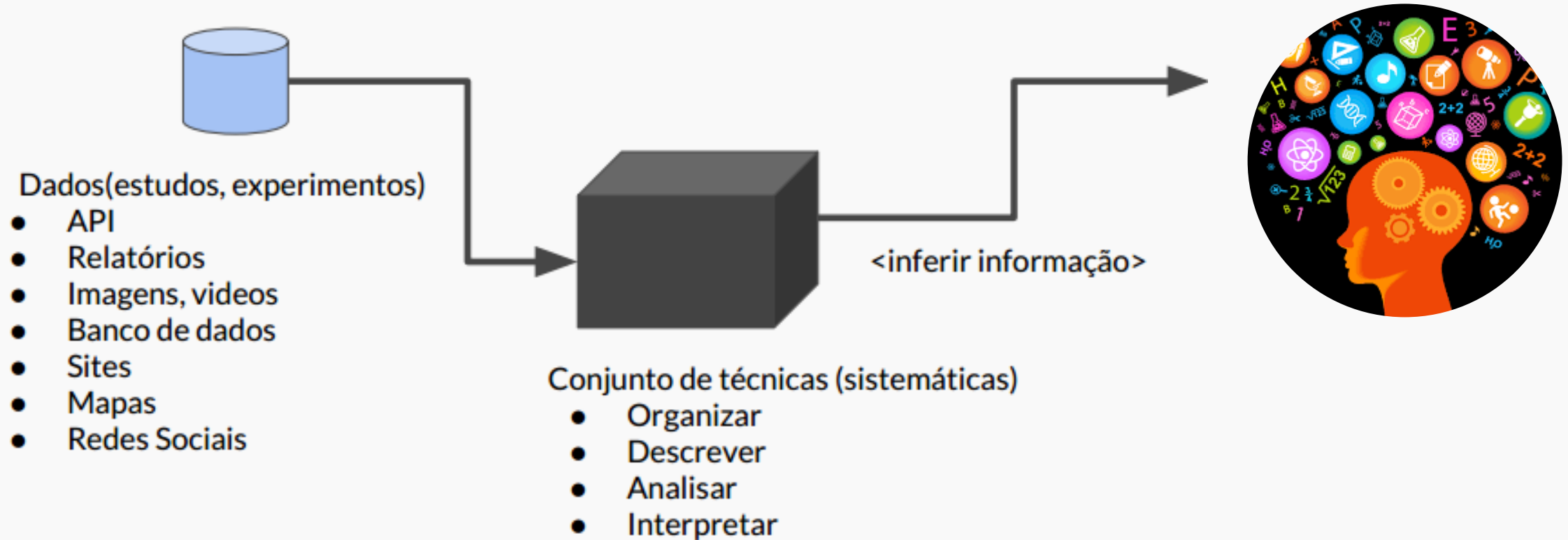
TÉCNICAS DEDUTIVAS

Focado em descrever os dados e fazer testes de inferências.

O seu objetivo é fazer descobertas como: frequência de acesso, o tempo de acesso aos materiais, mostrando comportamentos por períodos de tempo



ONDE É QUE ENTRA A ESTATÍSTICA?



DIVISÃO DA ESTATÍSTICA

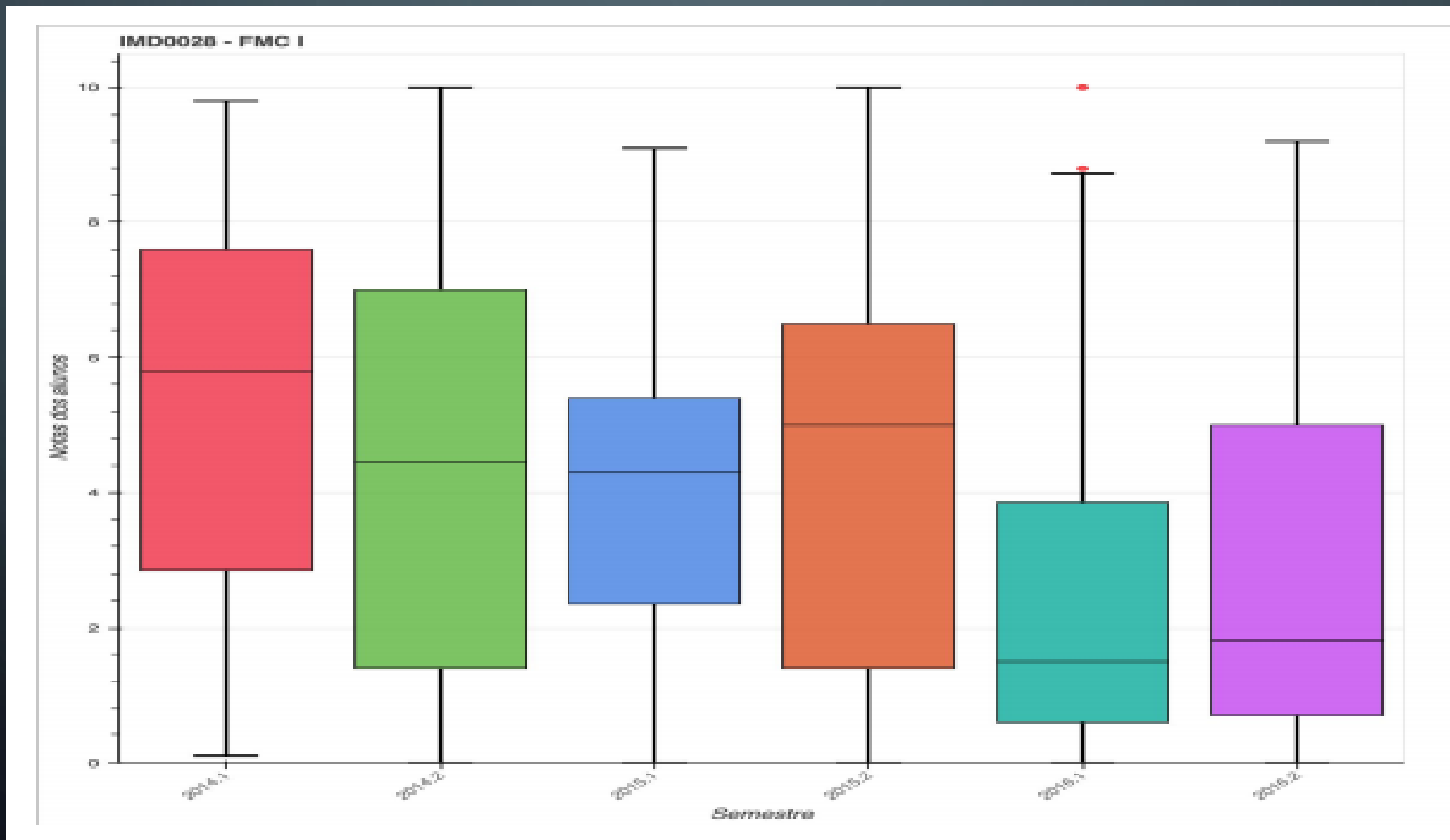
Estatística Descritiva

Inferência estatística



ESTATÍSTICA DESCRITIVA

Resumir e Descrever





NETFLIX

INFERÊNCIA ESTATÍSTICA

Extrapolar e obter inferências

Documentaries



CARACTERÍSTICA DOS DADOS




- ✓ Representados por uma matriz X^D_N atributos e objetos
- ✓ Atributos de Entrada
- ✓ Atributos de Saída
- ✓ Diferentes formas: tipo e escala

Sistema LOP - gerenciamento de exercícios de programação ECT -UFRN

Bem vindo!

Pronto para começar a praticar programação? 🤖

 Novidades

 Listas de Exercício

 Listas de Laboratório

 Provas anteriores

 Questões

Lista de Exercícios 2 - Estruturas Condicionais

Questões: 19

[Ver questões](#)

Lista de Exercícios 3 - Laço Condicional

Questões: 15

[Ver questões](#)

Lista de Exercícios 4 - Laço contado

Questões: 16

[Ver questões](#)

Lista de Exercícios 4.1 - Problemas tradicionais com repetição contada

Questões: 8

[Ver questões](#)

Ativar o Windows

Acesse as configurações do computador para ativar o Windows.

Atributos

Objetos

titulo	questoesFeitas	quantidade	NotaLista	ano	descricaoHorario	descricaoTurma	id_turma	matricula	nome
Lista de Exercícios 1 - Expressões e variáveis	11	15	73.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01A	598e15296d8650eb27d52e3d	20170039453	AMANDA
Lista de Exercícios 1 - Expressões e variáveis	5	15	33.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B	598e15296d8650eb27d52e5f	20170038394	EMANUEL FERNANDES P. DA ROCHA
Lista de Exercícios 1 - Expressões e variáveis	6	15	40.000000	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B	598e15296d8650eb27d52e5f	20170001812	ANDREY COSTA
Lista de Exercícios 1 - Expressões e variáveis	4	15	26.666667	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B	598e15296d8650eb27d52e5f	20170102735	FERNANDA
Lista de Exercícios 1 - Expressões e variáveis	2	15	13.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B	598e15296d8650eb27d52e5f	20170043590	JAYEDSON BRITO



Lop.csv

CARACTERÍSTICA DOS DADOS

- ✓ Tipo: Grau de quantização dos dados
- ✓ Escala: Significância relativa dos valores



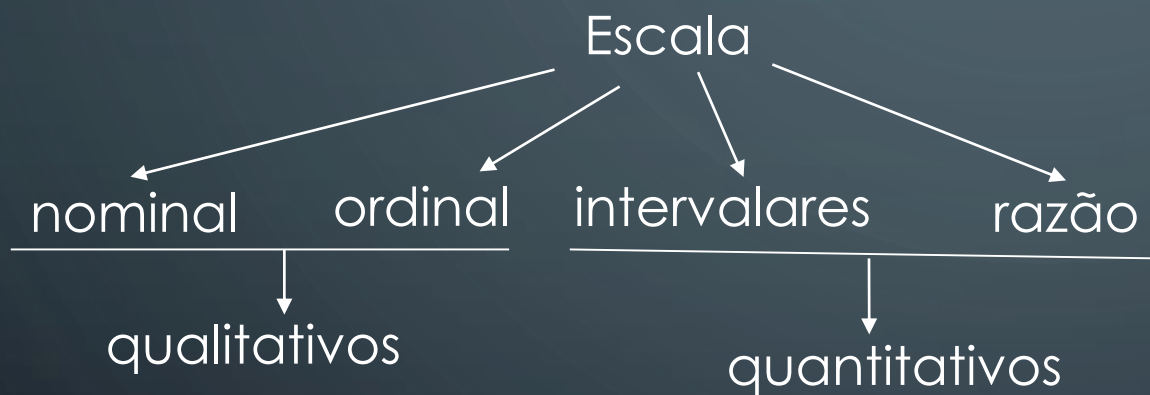
TIPOS : QUANTITATIVOS X QUALITATIVOS

	Quantitativas	Qualitativas
Descevre quantidade	SIM	NÃO
Descreve qualidade	NÃO	SIM
Usa números	SIM	SIM
Os numeros são quantidades reais	SIM	NÃO
Usa palavras	SIM	NÃO
As palavras expressão quantidade	SIM	NÃO

questoesFeitas	titulo
quantidade	descricaoHorario
NotaLista	descricaoTurma
ano	id_turma
	Matricula
	nome

CARACTERÍSTICA DOS DADOS

- ✓ Escala: define as operações que podem ser realizadas com os atributos



Nominal: Não existe uma relação de ordem

Operações: =, !=, classificadas ou agrupadas

Ordinal: Existe uma ordem das categorias

Operações: maior, menor, maior igual, menor igual, igual, diferente

Intervalar: Medem com ordem, com intervalos iguais em uma escala.

Operações: definir a ordem da diferença entre dois valores

Razão: Medem com ordem, com intervalos iguais em uma escala, e tem um verdadeiro zero

Operações: aritmética e booleanas

ESCALAS

	titulo	NotaLista
Podemos dizer se dois indivíduos são diferentes	SIM	SIM
Podemos dizer o tamanho da diferença	NÃO	SIM
Podemos dizer a direção da diferença	NÃO	SIM

Nominal
Ordinal
Intervalar
Razão

CARACTERÍSTICA DOS DADOS

	Nominal	Ordinal	Intervalar	Razão
Podemos dizer se dois indivíduos são diferentes	SIM	SIM	SIM	SIM
Podemos dizer a diferença da direção	NÃO	SIM	SIM	SIM
Podemos dizer o tamanho da diferença?	NÃO	NÃO	SIM	SIM
Podemos descrever a quantidade?	NÃO	SIM	SIM	SIM
Podemos descrever a qualidade?	SIM	NÃO	NÃO	NÃO

CARACTERÍSTICA DOS DADOS

titulo	questoesFeitas	quantidade	NotaLista	ano	descricaoHorario	descricaoTurma
Lista de Exercícios 1 - Expressões e variáveis	11	15	73.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01A
Lista de Exercícios 1 - Expressões e variáveis	5	15	33.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	6	15	40.000000	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	4	15	26.666667	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	2	15	13.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B

questoesfeitas

muito

medio

medio

pouco

pouco

Ordinal

Nominal

Razão

Intervalar ou ordinal?

Nominal

CARACTERÍSTICA DOS DADOS

Discreto ← **Contínua x Discreto** → Contínuo

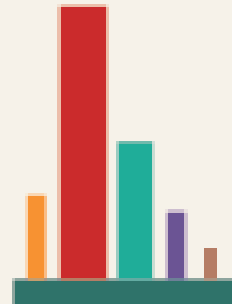
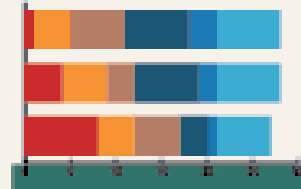
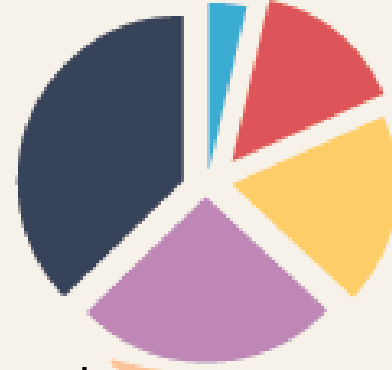
titulo	questoesFeitas	quantidade	NotaLista	ano	descricaoHorario	descricaoTurma
Lista de Exercícios 1 - Expressões e variáveis	11	15	73.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01A
Lista de Exercícios 1 - Expressões e variáveis	5	15	33.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	6	15	40.000000	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	4	15	26.666667	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B
Lista de Exercícios 1 - Expressões e variáveis	2	15	13.333333	2017.2	24M12 2T12 (24/07/2017 - 26/11/2017)	LÓGICA DE PROGRAMAÇÃO - Turma 01B

EXPLORAÇÃO DOS DADOS

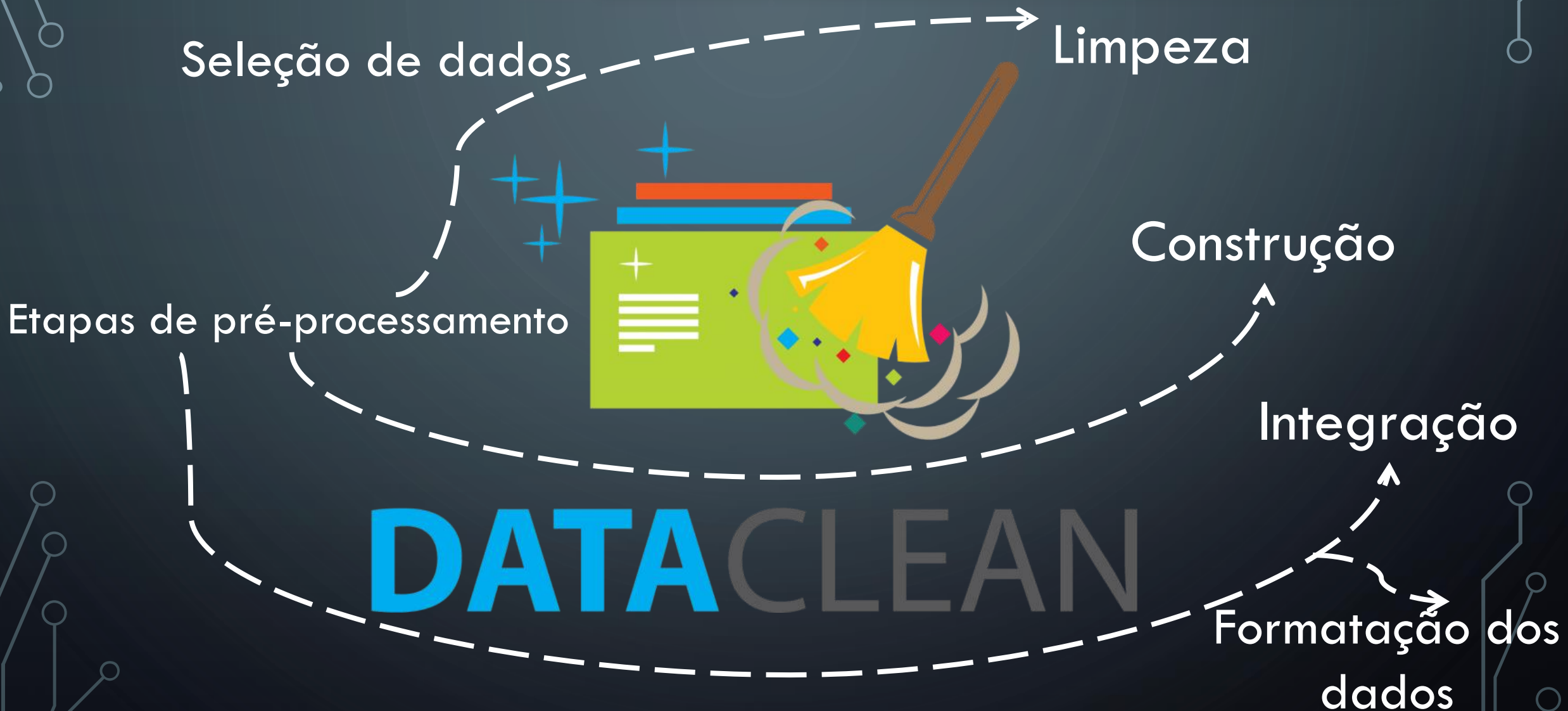


Extração das
informações

As técnicas variam de
acordo com a natureza
da variável



PREPARAÇÃO DOS DADOS



EXPLORAÇÃO DOS DADOS

- A estatística descritiva é utilizada nas variáveis quantitativas para resumir os dados, o que permite interpretações e comparações.
- Podem ser representados em forma de tabela ou gráfica
- Envolve a descrição da distribuição dos dados e o relacionamento entre as distribuições
- As distribuições são conjunto de valores de uma variável

DISTRIBUIÇÕES DAS VARIÁVEIS

- Frequência
- Localização ou tendência central (média)
- Dispersão ou espalhamento (desvio padrão)
- Distribuição ou formato

The background features a dark blue gradient with several large, faint, concentric circles centered in the upper half. In the corners, there are white line-art elements resembling electronic circuit traces, with small circles at various points along the lines.

FREQUÊNCIA

FREQUÊNCIA

Mede a quantidade de vezes que um atributo assume um dado valor em um determinado conjunto de dados.

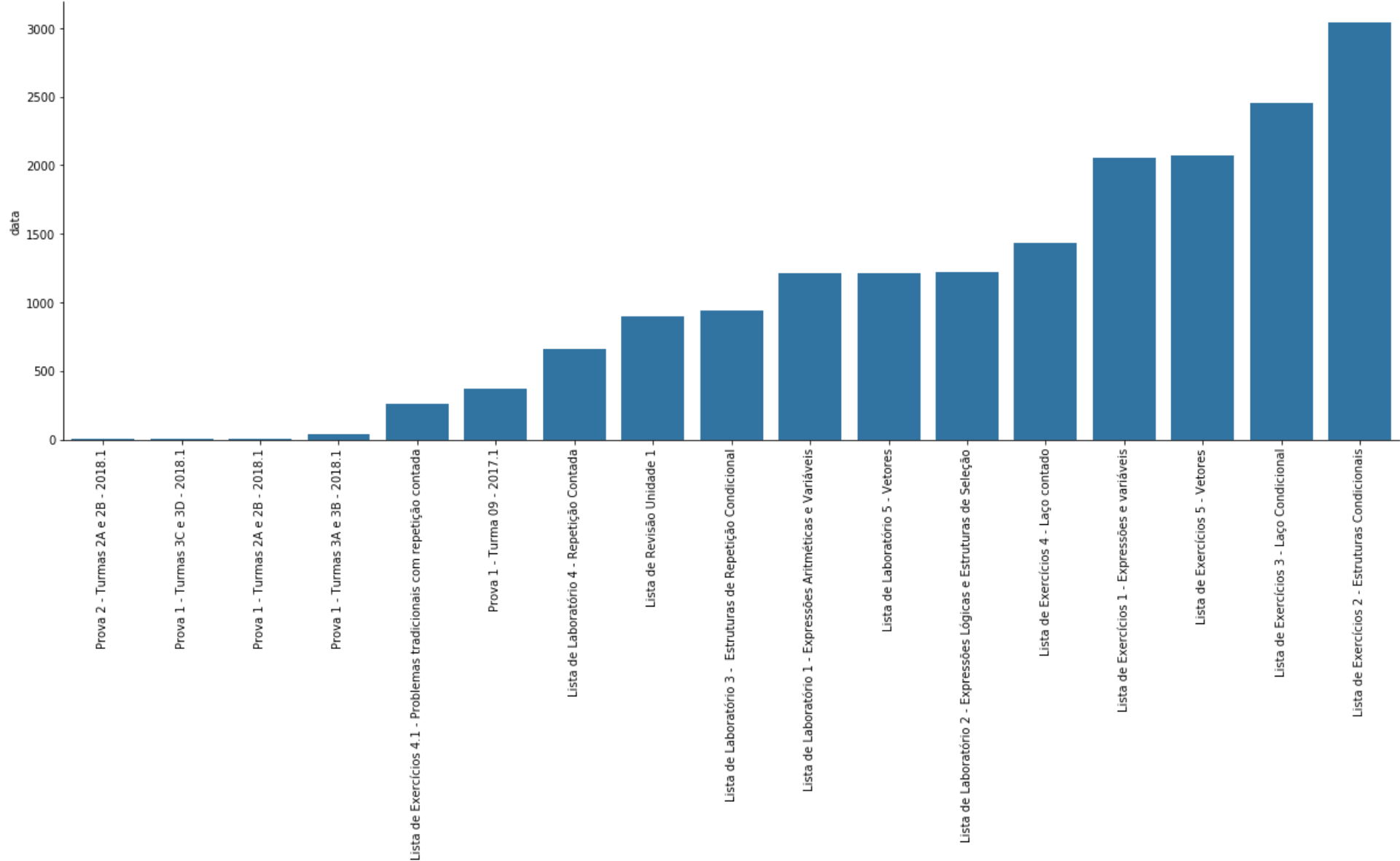
	titulo	id_lista
0	Lista de Exercícios 1 - Expressões e variáveis	2051
1	Lista de Exercícios 2 - Estruturas Condicionais	3042
2	Lista de Exercícios 3 - Laço Condicional	2451
3	Lista de Exercícios 4 - Laço contado	1430
4	Lista de Exercícios 4.1 - Problemas tradiciona...	260
5	Lista de Exercícios 5 - Vetores	2074
6	Lista de Laboratório 1 - Expressões Aritmética...	1209
7	Lista de Laboratório 2 - Expressões Lógicas e ...	1225
8	Lista de Laboratório 3 - Estruturas de Repeti...	944
9	Lista de Laboratório 4 - Repetição Contada	659
10	Lista de Laboratório 5 - Vetores	1213
11	Lista de Revisão Unidade 1	902
12	Prova 1 - Turma 09 - 2017.1	368
13	Prova 1 - Turmas 2A e 2B - 2018.1	7
14	Prova 1 - Turmas 3A e 3B - 2018.1	39
15	Prova 1 - Turmas 3C e 3D - 2018.1	5
16	Prova 2 - Turmas 2A e 2B - 2018.1	1

Nominal

100.000000	616
20.000000	418
40.000000	245
60.000000	212
80.000000	206
0.000000	169
33.333333	165
11.111111	160
6.666667	109
31.250000	96
13.333333	80
6.250000	78
26.666667	75
12.500000	69
22.222222	58
25.000000	56
66.666667	43
27.777778	42
44.444444	41
18.750000	36
37.500000	36
50.000000	35
43.750000	32
5.555556	30
53.333333	28
73.333333	23

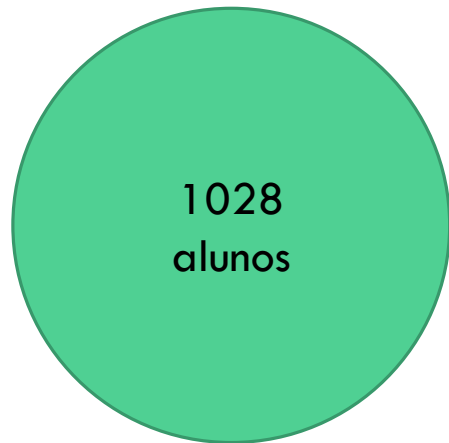
Razão

FREQUÊNCIA

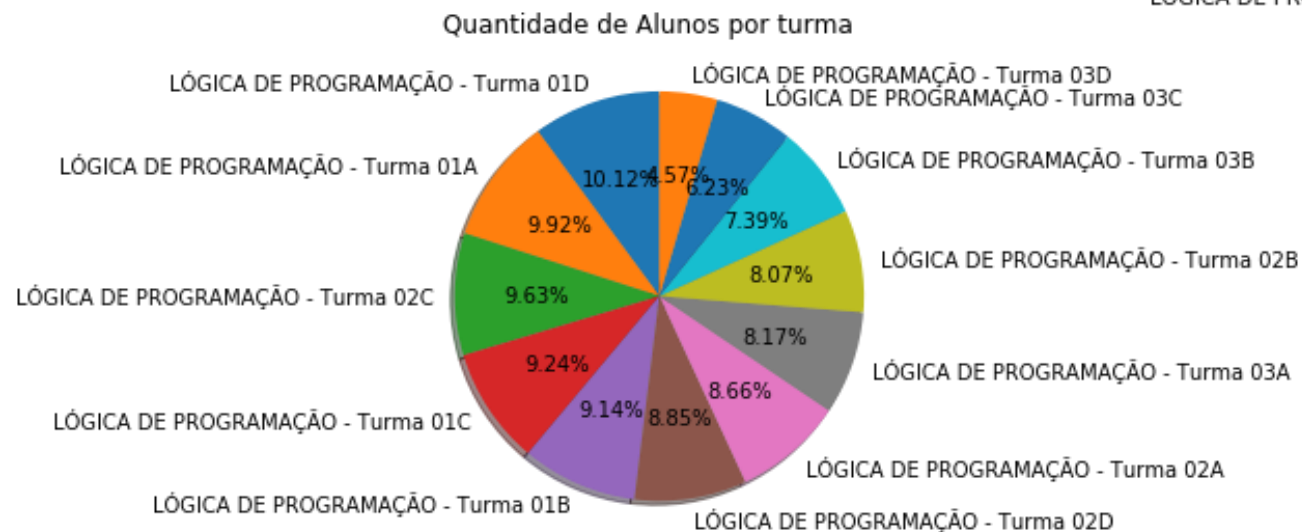
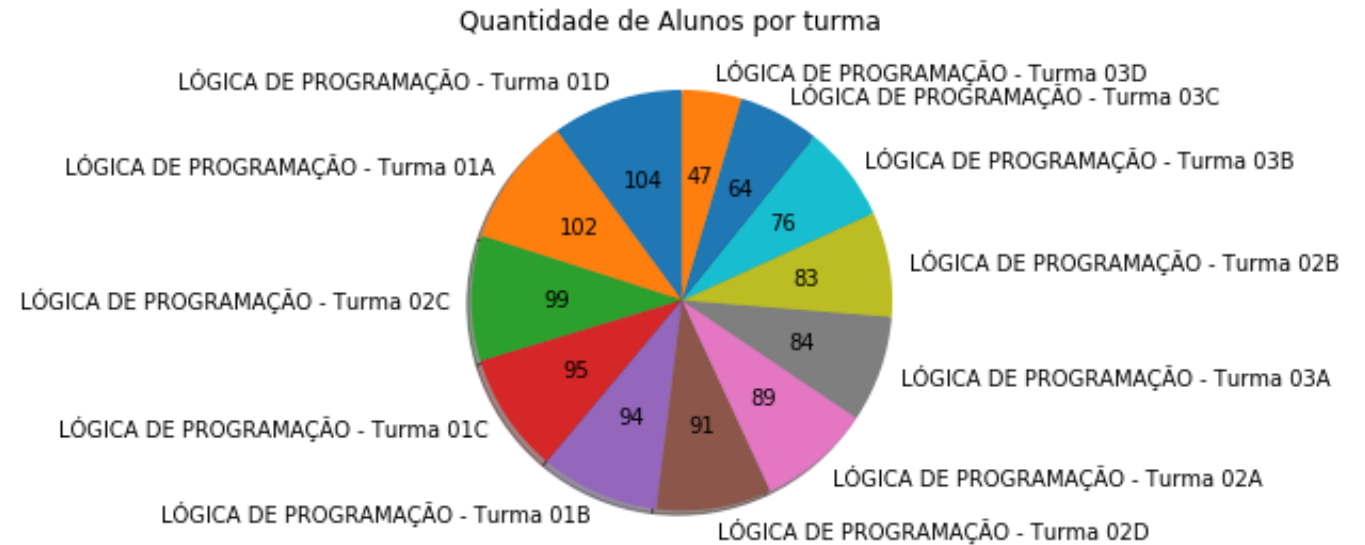


FREQUÊNCIA

Mede a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados.



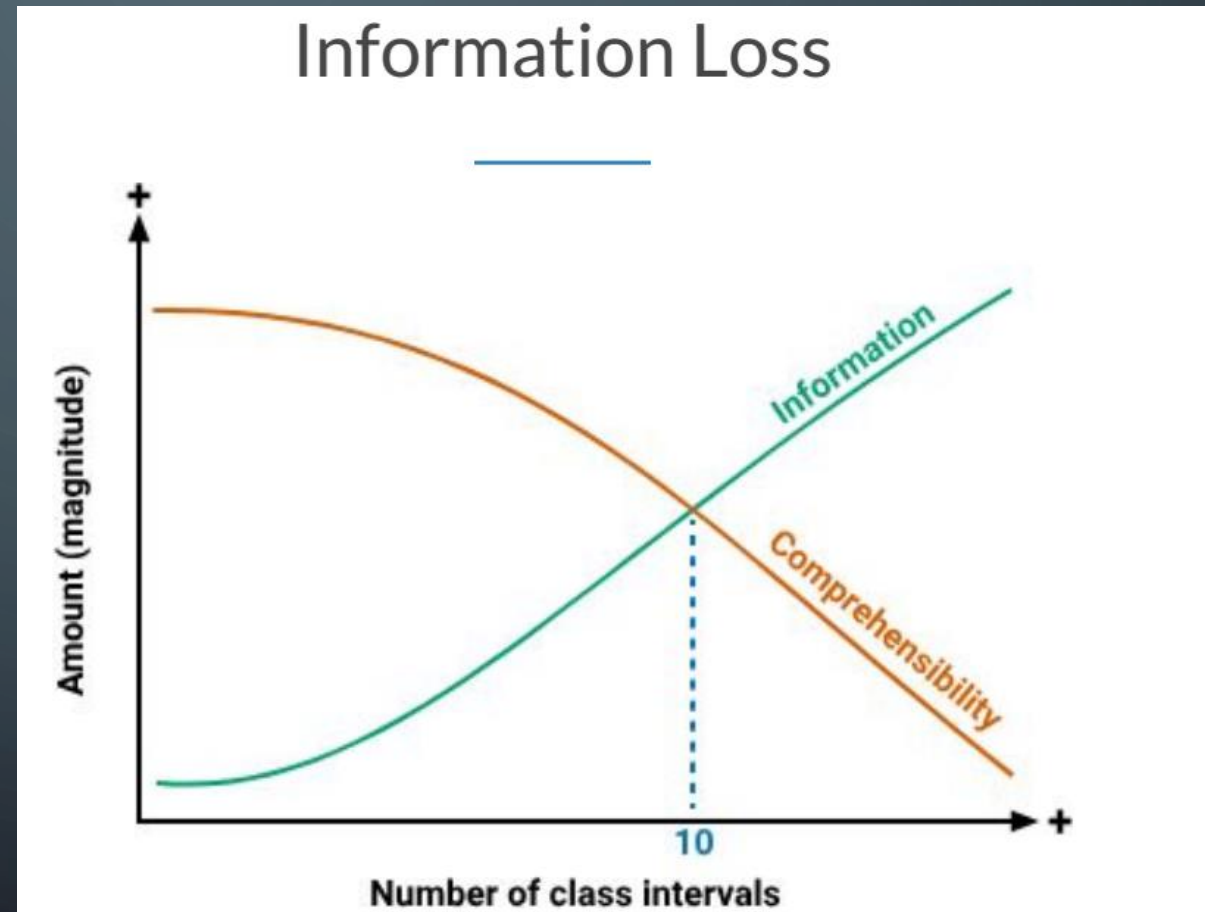
Contagem de alunos
por turma



Prorporção
Porcentagem

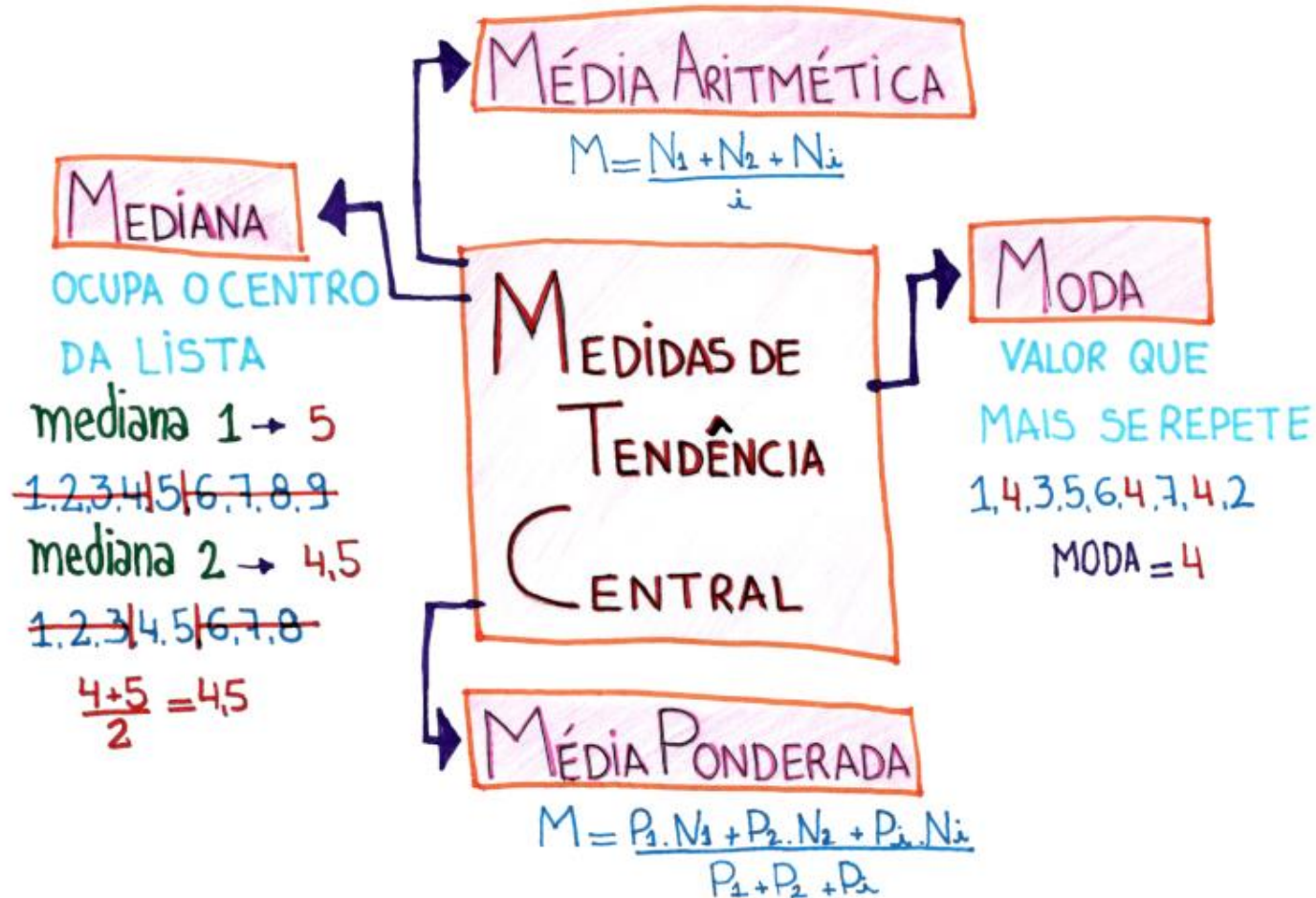
PERDA DE INFORMAÇÃO

NotaLista	...
100.000000	7.812500
99.277778	6.666667
99.166667	6.600000
99.000000	6.250000
98.500000	5.833333
98.437500	5.555556
97.875000	5.000000
97.388889	4.611111
97.133333	4.111111
96.400000	4.000000
96.000000	2.200000
95.000000	2.000000
94.444444	1.666667
93.750000	1.562500
93.611111	1.250000
	0.000000

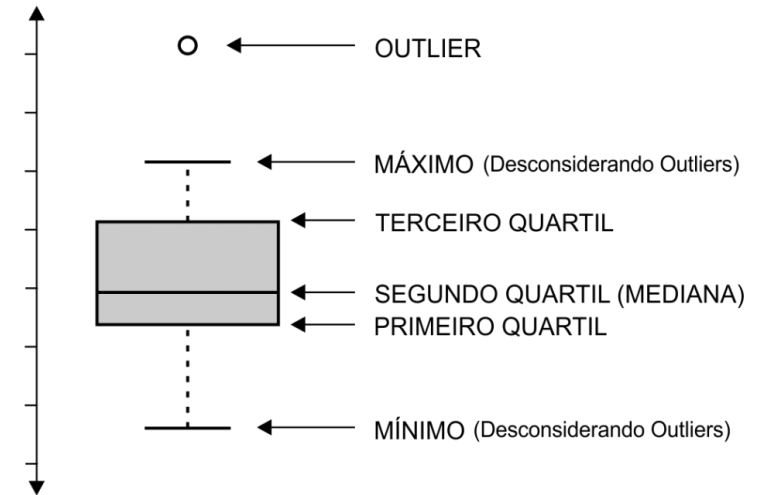


LOCALIZAÇÃO OU TENDÊNCIA CENTRAL (MÉDIA)

Localização ou tendência central (média)



Quartis



Ordenar os valores e dividir em quartos. Assim o 1º quartil tem 25% do valores abaixo dele...

Localização ou tendência central (média)

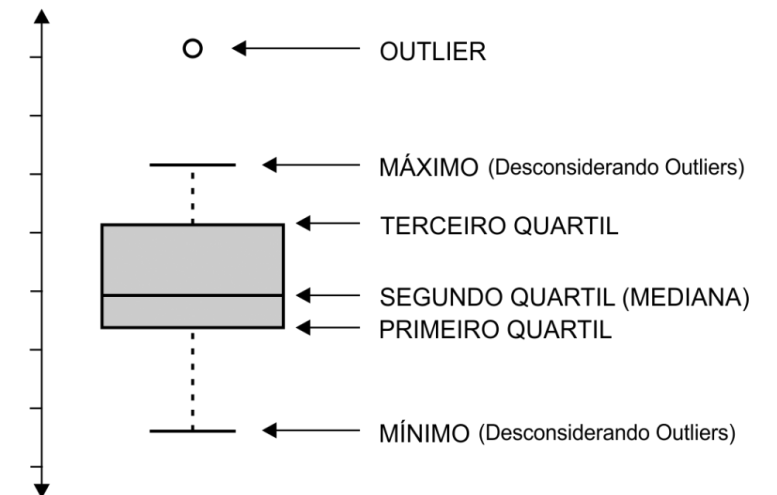
Notas de todas as listas dos alunos

```
↳ 100.000000 616
   20.000000 418
   40.000000 245
   60.000000 212
   80.000000 206
   0.000000 169
   33.333333 165
   11.111111 160
   6.666667 109
   31.250000 96
   13.333333 80
   6.250000 78
   26.666667 75
   12.500000 69
   22.222222 58
   25.000000 56
   66.666667 43
   27.777778 42
   44.444444 41
   18.750000 36
   37.500000 36
   50.000000 35
   43.750000 32
   5.555556 30
   53.333333 28
   73.333333 23
```

```
▶ tabelaFinal["NotaLista"].describe()
```

```
↳ count      3755.000000
   mean        46.860368
   std         33.417835
   min          0.000000
   25%         20.000000
   50%         40.000000
   75%         80.000000
   max        100.000000
   Name: NotaLista, dtype: float64
```

Quartis



Outlier: são valores muito distantes do mínimo ou do máximo

DISPERSÃO OU ESPALHAMENTO (DESVIO PADRÃO)

MEDIDAS DE ESPALHAMENTO

- Os valores estão amplamente espalhados ou relativamente concentrados em torno de um valor, por exemplo, a média.

Intervalo

Variância

Desvio Padrão

A variância é a medida mais utilizada.
Ela é a média da diferença de cada valor do atributo pela média dos valores desse atributo

```
tabelaFinal["NotaLista"].describe()
```

count	3755.000000
mean	46.860368
std	33.417835
min	0.000000
25%	20.000000
50%	40.000000
75%	80.000000
max	100.000000

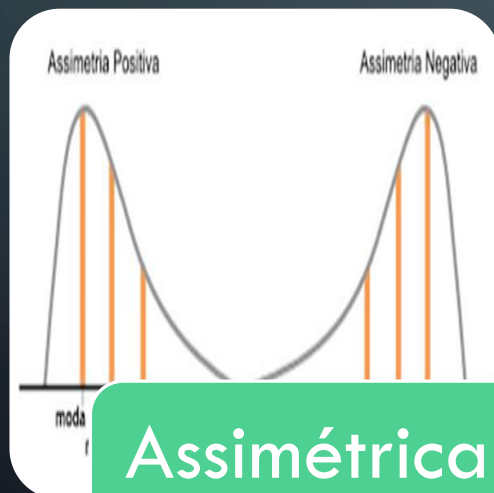
Name: NotaLista, dtype: float64

The background is a dark blue gradient. In the center, there are three concentric circles of increasing size, each with a lighter blue tint. The corners of the image are decorated with white, stylized circuit board traces and small circles, resembling electronic components or data paths.

DISTRIBUIÇÃO OU FORMATO

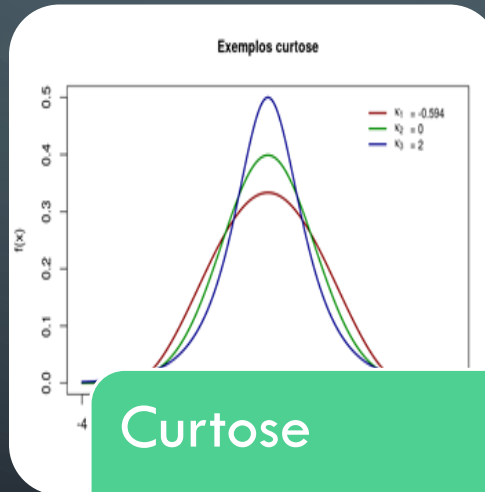
DISTRIBUIÇÃO OU FORMATO

- São medidas em torno da média. Chamados de Momentos
- A forma é importante para determinar o tipo de análise que pode ser executada nos dados



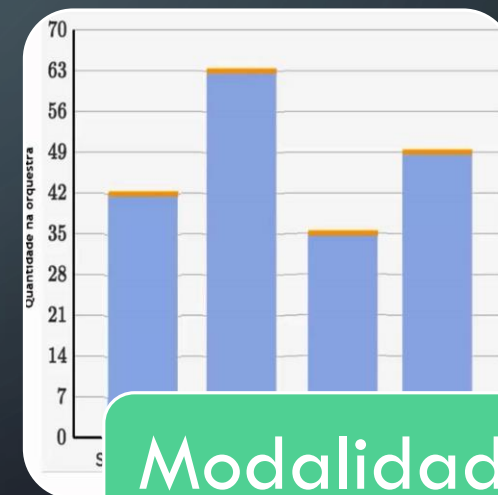
Assimétrica

- esquerda
- direita



Curtose

- Noraml
- Positiva
- negativa



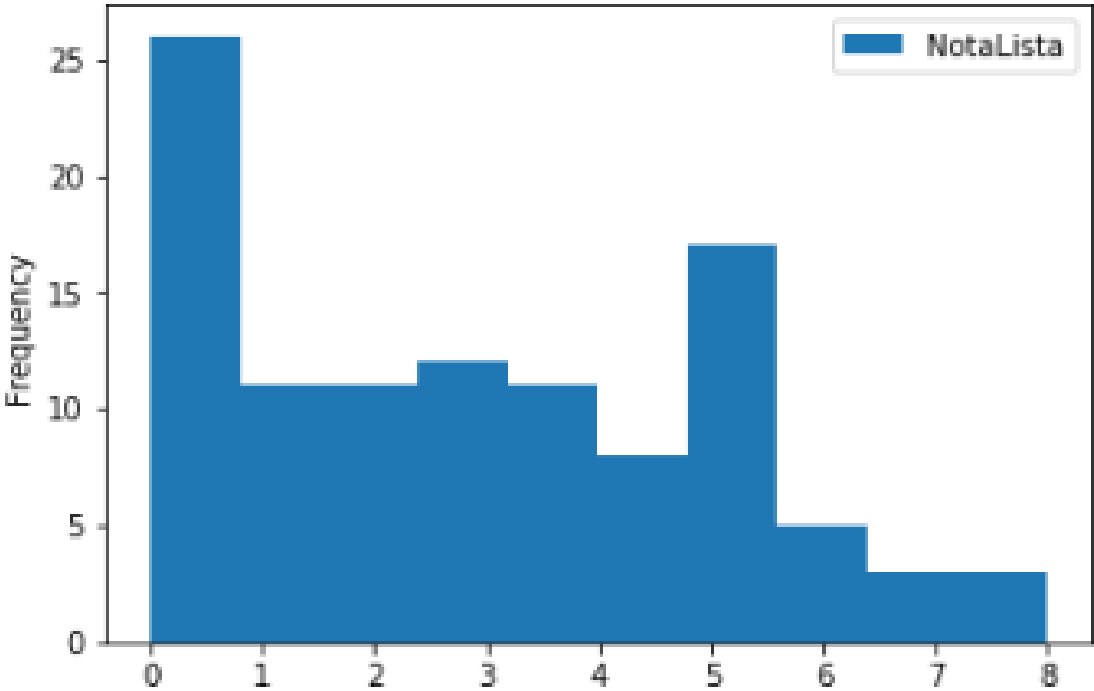
Modalidade

- Picos

DISTRIBUIÇÃO DAS NOTAS DOS ALUNOS

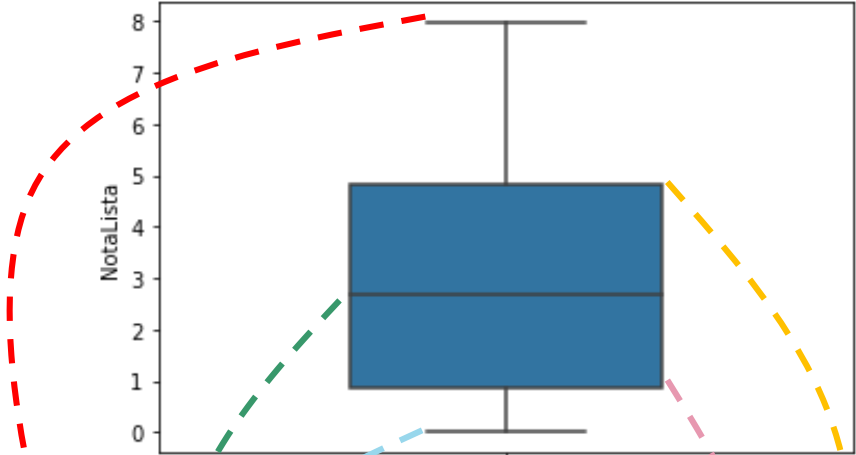
Tabela de frequência

7.977778	1
7.511111	1
7.317460	1
6.592593	1
6.499861	1
6.422222	1
6.370370	1
6.296296	1
6.148148	2
5.647664	1
5.481481	2
5.481481	1
5.333333	2
5.318086	1
5.317460	1
5.259259	1
5.259259	1
5.203704	1
5.153333	1
5.037037	1
5.000000	1
4.814815	3
4.814815	1
4.740741	1
4.592593	1
4.415278	1
4.370370	2
4.314815	1
4.288889	1
4.000000	1



Histograma

Gráfico de caixa



Medidas de espalhamento

NotaLista	
count	107.000000
mean	2.962544
std	2.101925
min	0.000000
25%	0.847656
50%	2.666667
75%	4.814815
max	7.977778

DADOS MULTIVARIADOS

- Mais de um atributo de entrada
- Estamos interessados na relação entre as variáveis

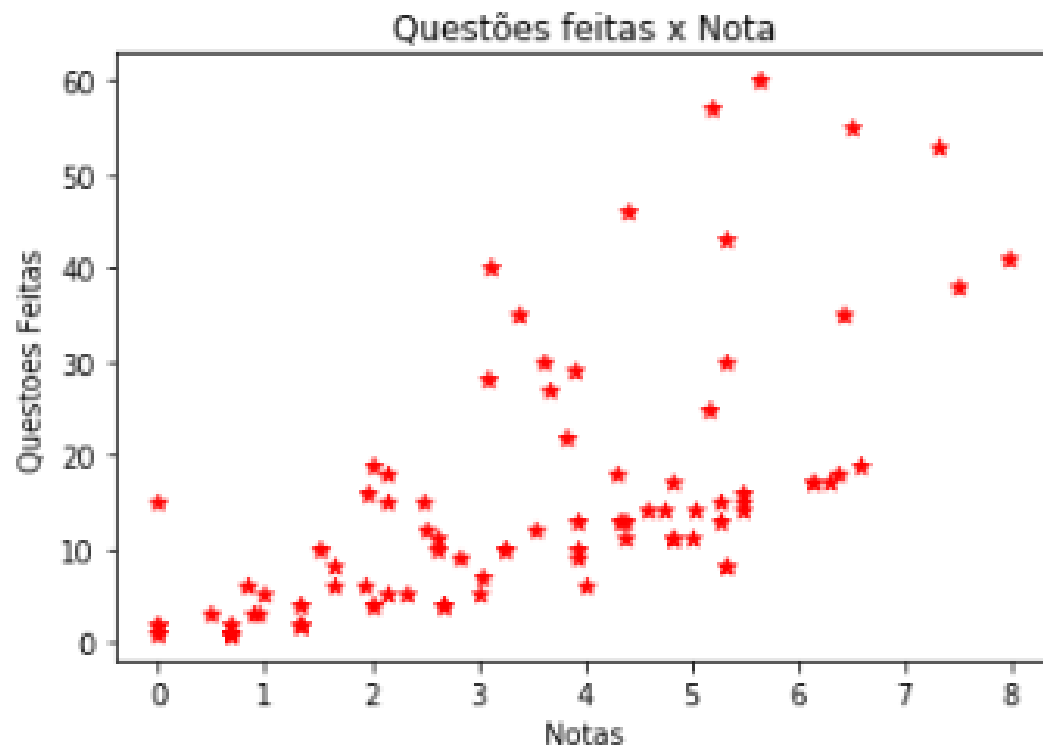
Ex: A quantidade de questões feitas está associada com o aumento da nota final?

- O grau de relacionamento é chamado de correlação

CORRELAÇÃO

- A correlação pode variar os seus valores entre -1 a $+1$
- O valor absoluto indica a força do relacionamento
- O sinal indica um relacionamento positivo ou negativo e o zero indica que não há correlação

GRÁFICO DE DISPERSÃO



	NotaLista	questoesFeitas
NotaLista	1.00000	0.69745
questoesFeitas	0.69745	1.00000

Escada de medição

Gráficos que podemos utilizar para mostrar as distribuições

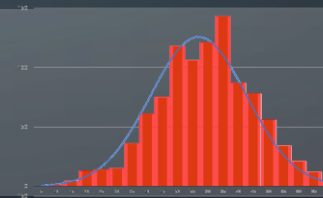
Nominal



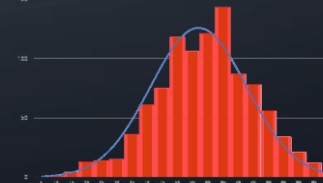
Ordinal



Intervalar



Razão





VISUALIZAÇÃO DE DADOS

É o uso de representações visuais, interativas e sustentadas por computador, de dados abstratos para amplificar a cognição”

10 RAZÕES

pelas quais deve investir na

VISUALIZAÇÃO DE DADOS

