

Big Data Analytics

Tema da aula
Análise Exploratória de Dados

Big Data & Analytics



Professor:
Caio Felipe Andrade

Coordenadores:
Profª Drª Alessandra de Ávila Montini
Profª Dr. Adolpho Walter Pimazoni Canton





BUSINESS SCHOOL

Graduação, pós-graduação, MBA, Pós-MBA, Mestrado Profissional, Curso In Company e EAD



CONSULTING

Consultoria personalizada que oferece soluções baseadas em seu problema de negócio



RESEARCH

Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais **escolas de negócio do mundo**, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 **projetos de consultorias** em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única *Business School* brasileira a figurar no *ranking* LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação, utilizando técnicas de **Big Data, Analytics** e **Inteligência Artificial**.



Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil
Os diretores foram professores de grandes especialistas do mercado
+10 anos de atuação
+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R\$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Corpo Diretivo

COORDENADORES DO LABDATA | ATUAÇÃO ACADÊMICA E PROFISSIONAL

5



Profª Dra.
Alessandra Montini

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em estatística aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, Coordenadora de Grupos de Pesquisa no CNPQ, Parecerista da FAPESP e Colunista de grandes Portais de Tecnologia.

 [linkedin.com/in/alessandramontini/](https://www.linkedin.com/in/alessandramontini/)



Prof. Dr. **Adolpho
Walter Canton**

Diretor do LABDATA-FIA. Consultor em Projetos de *Analytics*, *Big Data* e Inteligência Artificial. Professor FEA – USP. PhD em Estatística Aplicada pela *University of North Carolina at Chapel Hill*, Estados Unidos.



Conteúdo da Aula

- 1. Introdução
- 2. Banco de dados
 - i. Dados de série histórica e seção transversal
 - ii. *Missing values*
 - iii. Tipos de variáveis
 - iv. Pacotes computacionais
 - v. População e Amostra
- 3. Sintetizando dados qualitativos
 - i. Distribuição de frequências
 - ii. Gráficos: barra e setores
- 4. Sintetizando dados quantitativos
 - i. Medidas de posição
 - ii. *Box plot* e *outlier*
 - iii. Histograma
 - iv. Formas da distribuição
 - v. Medidas de dispersão
- 5. Análise Bidimensional
 - i. Qualitativo x Qualitativo
 - ii. Qualitativo x Quantitativo
 - iii. Quantitativo x Quantitativo
- 6. Códigos em R
- 7. Exercícios de Fixação



1. Introdução



Era da “explosão” de dados

1. INTRODUÇÃO | TRANSFORMAR INFORMAÇÃO EM CONHECIMENTO

8

Ferramentas automáticas de coleta de dados e tecnologia madura de banco de dados têm provido uma quantidade gigantesca de dados armazenados.

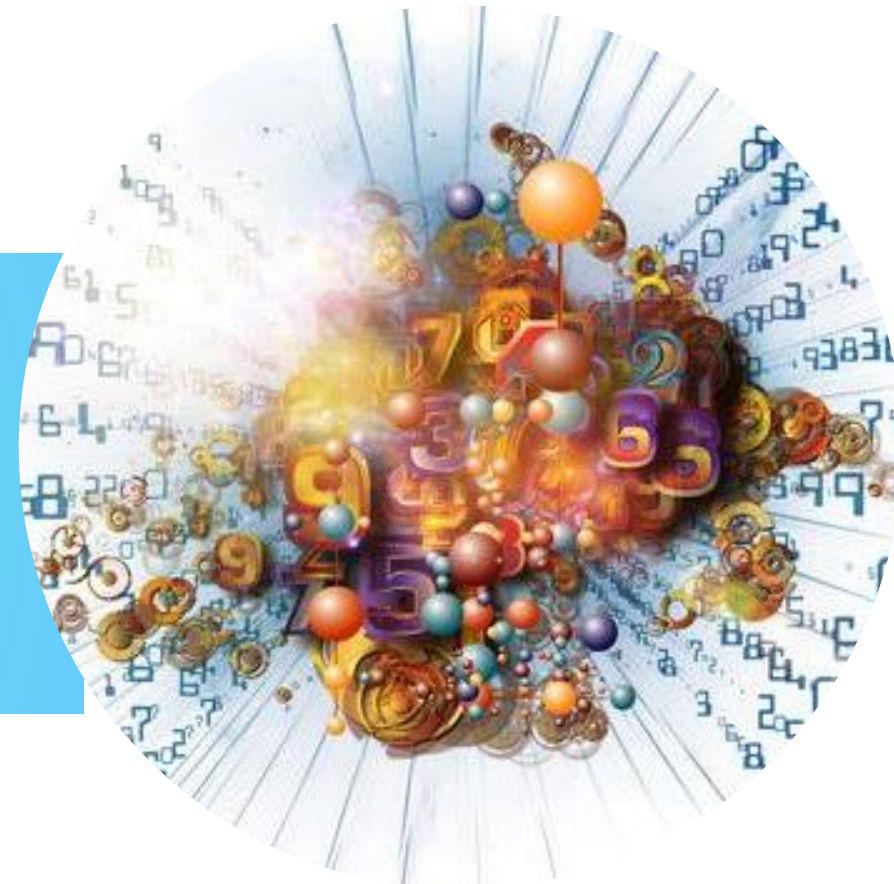
Telecomunicações: ligações recebidas/realizadas e uso de dados

Bancos: realização de empréstimos, investimentos, saques, acesso ao *internet banking*

Redes Sociais: *posts* curtidos e publicações realizadas

Internet: acesso a portais de notícias (Uol, Terra, Globo, R7 etc.)

E-commerce: produtos navegados e transações efetuadas



Como extrair conhecimento de interesse dos bancos de dados?

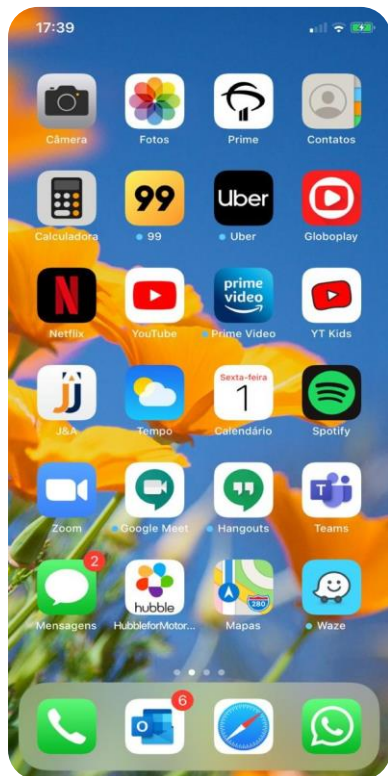


Nosso cotidiano

1. INTRODUÇÃO | TRANSFORMAR INFORMAÇÃO EM CONHECIMENTO

9

Nosso celular



Ao acordar



Deslocamento



Dia produtivo



Hora de dormir

Os dados fazem parte do nosso cotidiano. Nós, consumidores, utilizamos dados e também fornecemos dados do nosso comportamento para que as empresas possam melhorar seus serviços.

Uso consciente dos dados por parte das empresas com a LGPD (Lei Geral de Proteção de Dados), que entrará em vigor a partir de agosto/2021.



Profissional *Data Driven*

1. INTRODUÇÃO | SE DESTACAR NA SUA EMPRESA TOMANDO DECISÕES POR MEIO DE DADOS

10



PROFISSIONAL TRADICIONAL



PROFISSIONAL DATA DRIVEN



Metodologia de análise de dados

1. INTRODUÇÃO | TRANSFORMAR INFORMAÇÃO EM CONHECIMENTO

11



Banco de Dados

Preparação das informações

Extração, manipulação da base, definição de base para análise: histórico, público alvo, cálculo de variáveis e homologação.



Análise Exploratória

Entendimento da base analítica

Medidas descritivas para entendimento da base na visão de negócios e avaliação de consistência das informações.



Algoritmos Estatísticos e *Machine Learning*

Regras otimizadas

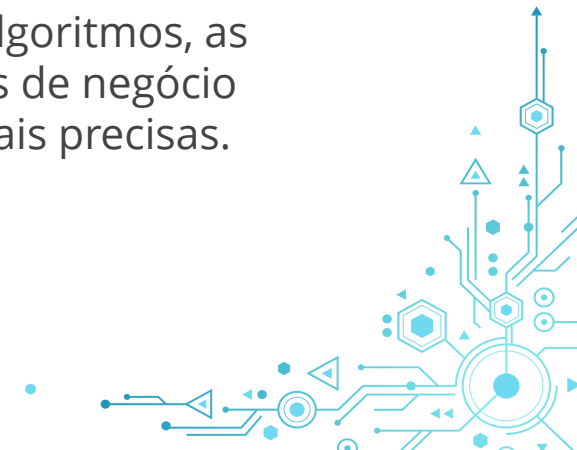
Encontram padrões de comportamentos históricos que podem ser aplicados para bases futuras.



Tomada de decisão

Resultados otimizados

Com base na análise histórica dos dados e uso de algoritmos, as decisões de negócio ficam mais precisas.



Análise Exploratória dos Dados

1. INTRODUÇÃO | TRANSFORMAR INFORMAÇÃO EM CONHECIMENTO

12

Tem o objetivo de realizar uma análise preliminar do banco de dados por meio de gráficos, tabelas, medidas de posição e de dispersão. Esta análise tem a finalidade de extrair conhecimento dos dados.



2. Banco de Dados



Banco de dados

2. BANCO DE DADOS | TIPOS DE BASE, VARIÁVEIS E PACOTES COMPUTACIONAIS

14

- i. Dados de Série Histórica e Seção Transversal
- ii. *Missing values*
- iii. Tipos de variáveis
- iv. Pacotes Computacionais
- v. População e Amostra



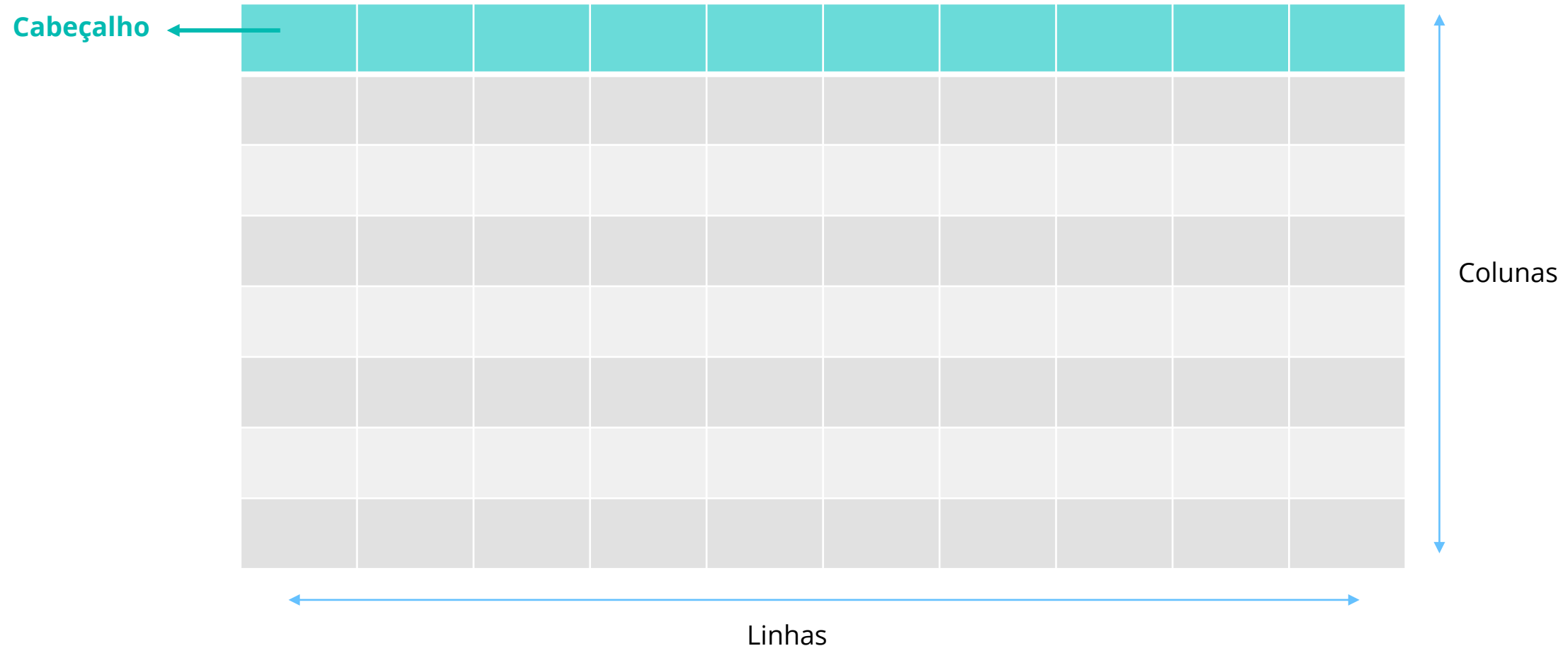
Estrutura do banco de dados

2. BANCO DE DADOS | CARACTERÍSTICAS

15

Geralmente, espera-se que o banco de dados seja formado como o apresentado:

- Na primeira 'linha-coluna' inicia-se a base de dados e na primeira linha adiciona-se o nome da variável.



Estrutura de dados de série histórica

2.i. DADOS DE SÉRIE HISTÓRICA | CARACTERÍSTICAS

16

EXEMPLO: preço do dólar (diário) – 24/03/2006 a 24/04/2020 (últimos 14 anos).

- Nesta base, as linhas representam períodos (diários) e a coluna 'Compra' representa os valores de compra e venda, em R\$.

Data	Compra
01/04/2020	5,2399
02/04/2020	5,2645
03/04/2020	5,2991
06/04/2020	5,2465
07/04/2020	5,2211
08/04/2020	5,2117
09/04/2020	5,0773
13/04/2020	5,1818
14/04/2020	5,1852
15/04/2020	5,2573
16/04/2020	5,2371
17/04/2020	5,2567
20/04/2020	5,2831
22/04/2020	5,3841
23/04/2020	5,4461
...	...



Dados de série histórica

São dados coletados ao longo de diversos períodos de tempo. Neste banco de dados, o valor da compra do dólar foi analisado nos últimos 14 anos, na periodicidade diária.



Estrutura de dados de seção transversal

2.i. DADOS DE SEÇÃO TRANSVERSAL | CARACTERÍSTICAS

17

Case: *People Analytics*

Um gestor deseja analisar as variáveis relacionadas a equipe de 36 colaboradores.



Arquivo: People_Analytics.xlsx



Estrutura de dados de seção transversal

2.i. DADOS DE SEÇÃO TRANSVERSAL | CARACTERÍSTICAS

18

EXEMPLO: Base de dados para um estudo de *People Analytics*.

- Nesta base, as linhas representam os colaboradores e, nas colunas, exibe-se suas respectivas características.

N	estado_civil	grau_instrucao	n_filhos	salario	idade_anos	reg_procedencia
1	solteiro	ensino fundamental		4	26	interior
2	casado	ensino fundamental	1	4,56	32	capital
3	casado	ensino fundamental	2	5,25	36	capital
4	solteiro	ensino médio		5,73	20	outra
5	solteiro	ensino fundamental		6,26	40	outra
6	casado	ensino fundamental	0	6,66	28	interior
7	solteiro	ensino fundamental		6,86	41	interior
8	solteiro	ensino fundamental		7,39	43	capital
9	casado	ensino médio	1	7,59	34	capital
10	solteiro	ensino médio		7,44	23	outra
11	casado	ensino médio	2	8,12	33	interior
12	solteiro	ensino fundamental		8,46	27	capital
13	solteiro	ensino médio		8,74	37	outra
14	casado	ensino fundamental	3	8,95	44	outra
15	casado	ensino médio	0	9,13	30	interior
16	solteiro	ensino médio		9,35	38	outra
17	casado	ensino médio	1	9,77	31	capital
18	casado	ensino fundamental	2	9,8	39	outra
19	solteiro	superior		10,53	25	interior
...	solteiro	ensino médio		10,76	37	interior

Dados de seção transversal (*cross-sectional*)

São dados coletados no mesmo intervalo de tempo. Neste banco de dados, a empresa coletou os dados, por exemplo, no mês de referência específico.

Arquivo: People_Analytics.xlsx



Missing values ou dados faltantes

2.ii. MISSING VALUES

EXEMPLO: Base de dados para um estudo de *People Analytics*.

- Nesta base, as linhas representam os colaboradores e, nas colunas, exibe-se suas respectivas características.

N	estado_civil	grau_instrucao	n_filhos	salario	idade_anos	reg_procedencia
1	solteiro	ensino fundamental		4	26	interior
2	casado	ensino fundamental	1	4,56	32	capital
3	casado	ensino fundamental	2	5,25	36	capital
4	solteiro	ensino médio		5,73	20	outra
5	solteiro	ensino fundamental		6,26	40	outra
6	casado	ensino fundamental	0	6,66	28	interior
7	solteiro	ensino fundamental		6,86	41	interior
8	solteiro	ensino fundamental		7,39	43	capital
9	casado	ensino médio	1	7,59	34	capital
10	solteiro	ensino médio		7,44	23	outra
11	casado	ensino médio	2	8,12	33	interior
12	solteiro	ensino fundamental		8,46	27	capital
13	solteiro	ensino médio		8,74	37	outra
14	casado	ensino fundamental	3	8,95	44	outra
15	casado	ensino médio	0	9,13	30	interior
16	solteiro	ensino médio		9,35	38	outra
17	casado	ensino médio	1	9,77	31	capital
18	casado	ensino fundamental	2	9,8	39	outra
19	solteiro	superior		10,53	25	interior
...	solteiro	ensino médio		10,76	37	interior

O *missing value*, ou simplesmente *missing*, trata-se de uma informação faltante em um campo da base de dados.

Na prática, o *missing* significa “desconheço a informação”.

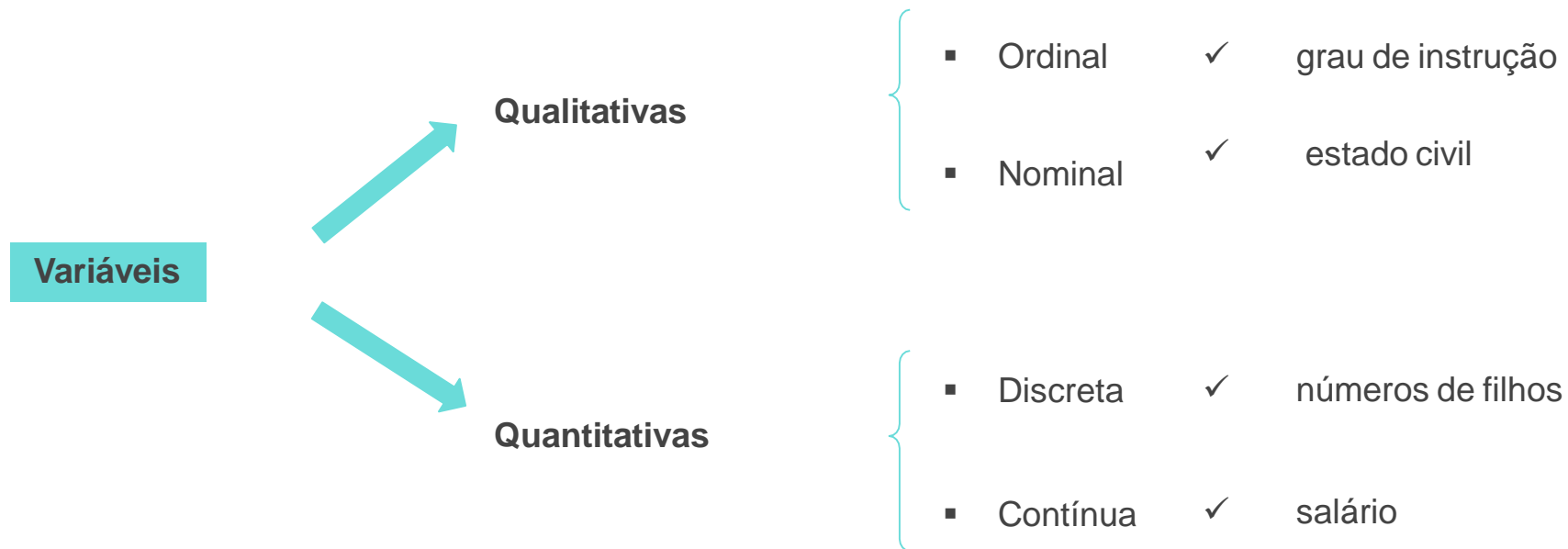
Arquivo: People_Analytics.xlsx



Tipos de variáveis

2.iii. TIPOS DE VARIÁVEIS | QUALITATIVAS E QUANTITATIVAS

20



Pacotes de análise de dados

2.iv. PACOTES COMPUTACIONAIS | SOFTWARES MAIS UTILIZADOS

21

- Pacotes computacionais para análise de dados são programas computacionais de aplicação analítica, mineração de dados, estatística e algoritmos computacionais.
- Utilizados desde uma simples análise exploratória até algoritmos mais sofisticados. Incluem uma série de funcionalidades para viabilizar de forma rápida o apoio à tomada de decisão que transformam os dados em informações de negócio para as empresas.

Exemplos





<https://www.r-project.org/>

- Software amplamente utilizado para análise de dados.
- **R** é um *software open source* (aberto e gratuito), com contribuição contínua da comunidade acadêmica e corporativa disponibilizando novas funções de análise de dados para comunidade R.
- **R** é uma linguagem e seu ambiente computacional fornece uma grande variedade de técnicas estatísticas e gráficos.

- O **Rstudio** é um ambiente de desenvolvimento integrado para o R.
- Ele inclui um console, editor de sintaxe que suporta execução direta de código, além de ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho.
- O RStudio está disponível na versão *open source* e comercial.



<https://rstudio.com/>



Conceitos: População e Amostra

2.v. POPULAÇÃO E AMOSTRA | IMPORTANTE CONCEITO ESTATÍSTICO

23

- **População:** todas as observações do universo de referência.
- O tamanho da **população** será denotado por **N**.
- **Amostra:** parte de uma população de referência.
- O tamanho da **amostra** será denotado por **n**.



Todos os brasileiros



Pesquisa eleitoral para presidência da República no Brasil



Todos os carros produzidos em um montadora



Controle de qualidade de veículos para avaliar defeitos na produção



Todos os clientes de um banco



Treino de algoritmos para identificar padrão de cancelamento de conta corrente



3. Sintetizando dados qualitativos



Sintetizando dados qualitativos

3. SINTETIZANDO DADOS QUALITATIVOS | FREQUÊNCIAS E GRÁFICOS

25

- i. Distribuição de Frequências
- ii. Gráficos: barra e setores



Distribuições de Frequências

3. SINTETIZANDO DADOS QUALITATIVOS | CASE PEOPLE ANALYTICS

26

A tabela apresenta a **frequência absoluta** e a **frequência relativa** relacionadas a uma amostra de 36 funcionários do case *People Analytics*.

Grau de instrução	Frequência absoluta	Frequência Relativa	Porcentagem
ensino fundamental	12	0,33	33,33%
ensino médio	18	0,50	50,00%
superior	6	0,17	16,67%
Total Geral	36	1	100%

Porcentagem =
Frequência Relativa x
100.

Analisando a tabela, descritivamente, verificamos que da amostra de 36 funcionários analisados, metade tem ensino médio, seguido por 33% dos funcionários com ensino fundamental, e apenas 17% com ensino superior.

@2020 LABDATA FIA. Copyright all rights reserved.

A soma das frequências relativas deve ser igual a um.



A frequência pode ser representada por meio de um gráfico de **barra** ou gráfico de **pizza (ou setor)**.

Gráfico de Barra para Frequência Absoluta
- Escolaridade -

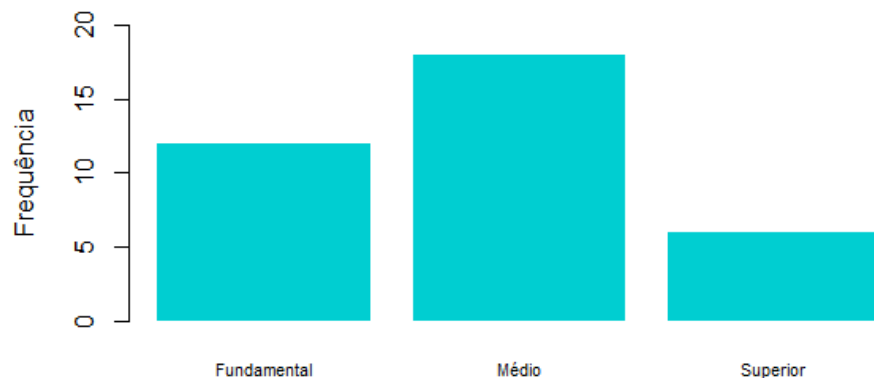
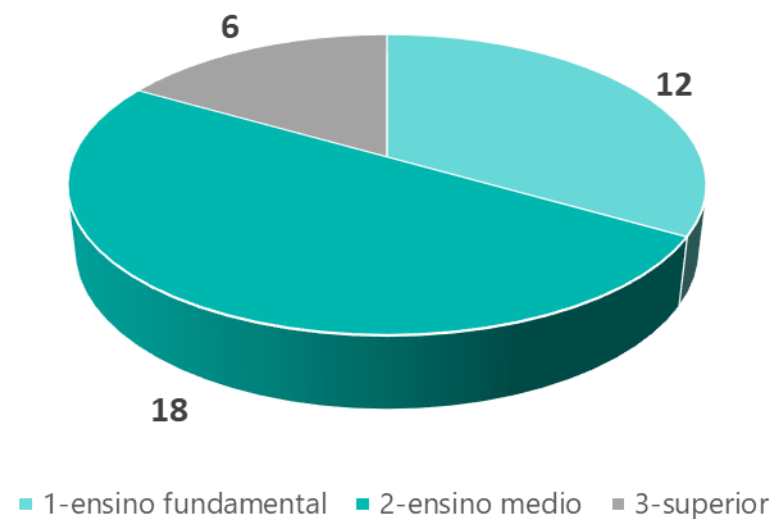


Gráfico de Pizza para Frequência Absoluta
- Escolaridade -



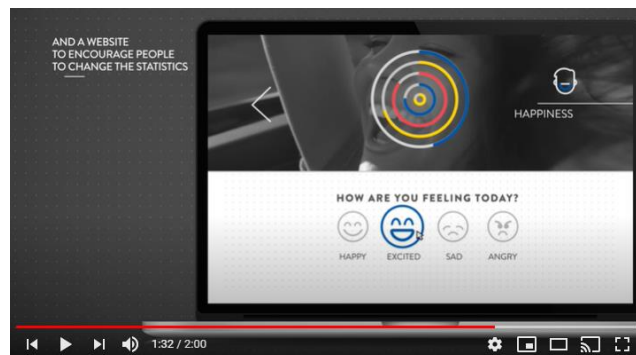
Discussão: Case Logo do Equador

3. SINTETIZANDO DADOS QUALITATIVOS | TRANSFORMANDO A VIDA DAS PESSOAS COM NÚMEROS

28

Qual foi o principal resultado alcançado com o logotipo?

<https://www.youtube.com/watch?v=1I39X--1NTs>
Video espanhol (legenda em inglês)



Fonte: YouTube - Prêmio Cannes Lions 2016



4. Sintetizando dados quantitativos



Sintetizando dados quantitativos

4. SINTETIZANDO DADOS QUANTITATIVOS | MEDIDAS DE POSIÇÃO, DISPERSÃO E GRÁFICOS

30

- i. Medidas de posição
- ii. *Box plot* e *outlier*
- iii. Histograma
- iv. Formas da distribuição
- v. Medidas de dispersão



Mínimo e Máximo









4.i. MEDIDAS DE POSIÇÃO | DADOS QUANTITATIVOS

31

São os valores extremos das observações, sendo o **mínimo** o menor valor de uma determinada variável e o **máximo** o maior valor de uma determinada variável.

Ver preço geladeira electrolux dfn41 371

Patrocinados ⓘ

							
Refrigerador Electrolux 371L 2 Portas Frost Free Branco...	Refrigerador Electrolux DFN41 Frost Free com Pain...	Refrigerador Electrolux DFN41 Frost Free com Pain...	Refrigerador Electrolux DFN41 Frost Free com Pain...	Geladeira/Refrige Frost Free 371 litros (DFN41) 127V	Geladeira Electrolux Frost Free Duplex 2 Portas Dfn41...	Geladeira/refrige Frost Free 371 Litros (dfn41)	Geladeira/refrige Frost Free 371 Litros (dfn41)
R\$ 1.804,05	R\$ 1.899,00	R\$ 1.899,00	R\$ 1.899,00	R\$ 1.999,00	R\$ 1.999,00	R\$ 1.935,12	R\$ 2.039,00
Magazine Luiza	Casas Bahia	Extra.com.br	Pontofrio.com	Loja Electrolux	Carrefour	Shoptime	Americanas.com

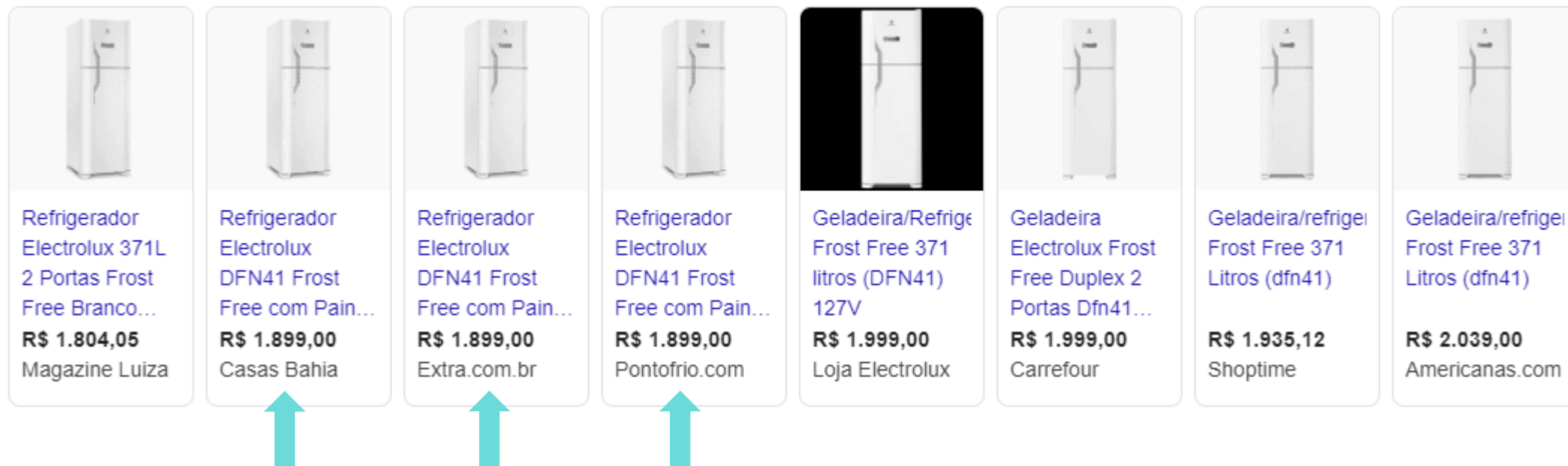
O menor valor da pesquisa (**mínimo**)
por preço de geladeira do modelo
DFN41 371 litros na internet foi
R\$1.804,05, no Maganize Luiza.

O maior valor da pesquisa (**máximo**)
por preço de geladeira do modelo
DFN41 371 litros na internet foi
R\$2.039,00, nas Americanas.

A **moda** é a observação mais frequente em um conjunto de observações.

Ver preço geladeira electrolux dfn41 371

Patrocinados ⓘ



O valor mais frequente (**moda**) da busca por preço de geladeira do modelo DFN41 371 litros na internet foi **R\$1.899,00**



A **média** é obtida a partir da soma das observações dividindo-se pelo total de observações.

Exemplo: Cálculo da **média (aritmética)** para o conjunto de dados de 5 elementos:

salario
5,25
4,56
4
5,73
6,26

A média é dada por: $\bar{X} = \frac{5,25 + 4,56 + 4 + 5,73 + 6,26}{5} = 5,16$



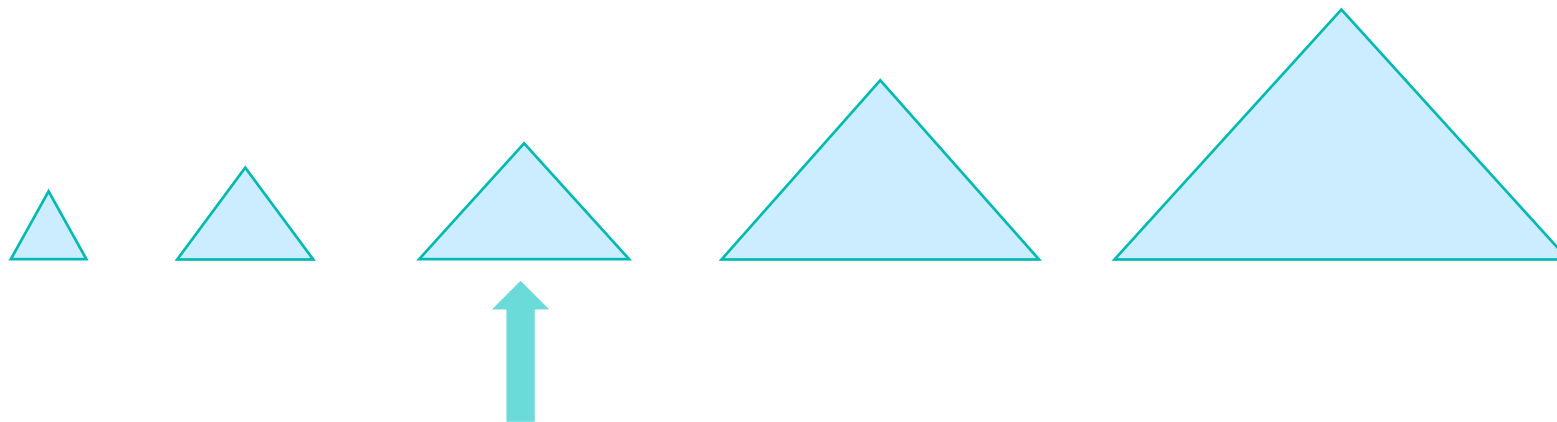
Mediana

4.i. MEDIDAS DE POSIÇÃO | DADOS QUANTITATIVOS

34

A **mediana** é a observação que ocupa a posição central de um conjunto de observações, ou seja, 50% das observações são maiores do que a mediana e 50 % das observações são menores do que a mediana.

Exemplo:



Este triângulo representa a **mediana** pois 50 % dos triângulos são maiores do que ele e 50 % dos triângulos são menores do que ele.



Mediana

4.i. MEDIDAS DE POSIÇÃO | DADOS QUANTITATIVOS

35

A **mediana** é a observação que ocupa a posição central de um conjunto de observações, ou seja, 50% das observações são maiores do que a mediana e 50 % das observações são menores do que a mediana.

Exemplo:

salario
5,25
4,56
4
5,73
6,26



salario
4
4,56
5,25
5,73
6,26

A mediana é **5,25** salários mínimos

Para se obter a mediana, inicialmente, deve-se **ordenar** a base de dados em ordem crescente.

- 50% dos funcionários ganham menos do que 5,25 salários mínimos;
- 50% dos funcionários ganham mais do que 5,25 salários mínimos.



Mediana

4.i. MEDIDAS DE POSIÇÃO | DADOS QUANTITATIVOS

36

Caso tenhamos um número par de observações, a mediana é calculada como a média das duas observações que ocupam a posição central.

Exemplo:

salario
5,25
4,56
4
5,73



salario
4
4,56
5,25
5,73

A mediana é $\frac{4,56+5,25}{2} = 4,905$ salários mínimos

Para se obter a mediana, inicialmente, deve-se **ordenar** a base de dados em ordem crescente.

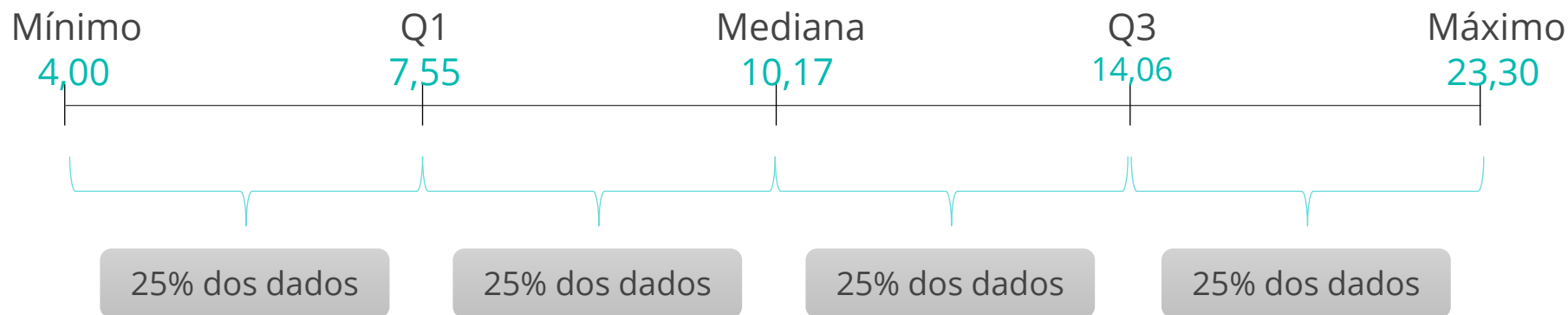
- 50% dos funcionários ganham menos do que 4,905 salários mínimos;
- 50% dos funcionários ganham mais do que 4,905 salários mínimos.



Exemplo: Salários no case de *People Analytics*

A mediana é a observação que ocupa a posição central, também conhecida como **Q2** (ou **segundo quartil**). Seguindo a mesma lógica, existem duas outras medidas:

- **Q1 (primeiro quartil)**: 25% das observações são menores do que o Q1.
- **Q3 (terceiro quartil)**: 75% das observações são menores do que o Q3.



Os dados indicam que, para o salário dos 36 funcionários analisados:

- 25% ganham menos que 7,55 salários mínimos;
- 75% ganham menos do que 14,06 salários mínimos.



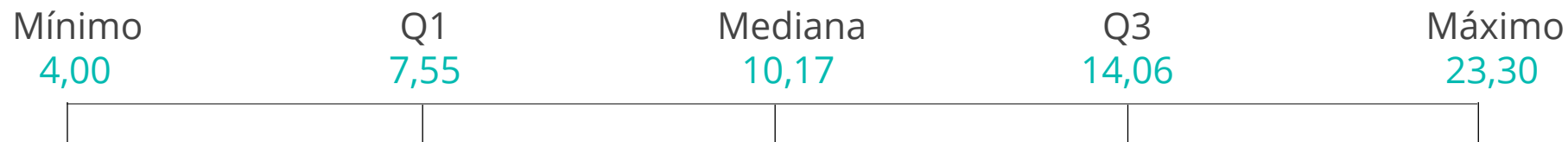
Intervalo interquartil e Amplitude

4.i. MEDIDAS DE POSIÇÃO | DADOS QUANTITATIVOS

38

O **IIQ (intervalo interquartil)** é igual a $IIQ = Q3 - Q1$, indicando que 50% dos dados centrais estão entre Q1 e Q3. Quanto menor o IIQ, menos dispersos estão os dados.

A **amplitude** é igual a **Máximo - Mínimo**, indicando a comprimento do intervalo de valores.



Exemplo: Salários no case de *People Analytics*

$$IIQ = 14,06 - 7,55 = \mathbf{6,51}$$

$$\mathbf{Amplitude} = 23,30 - 4 = \mathbf{19,30}$$

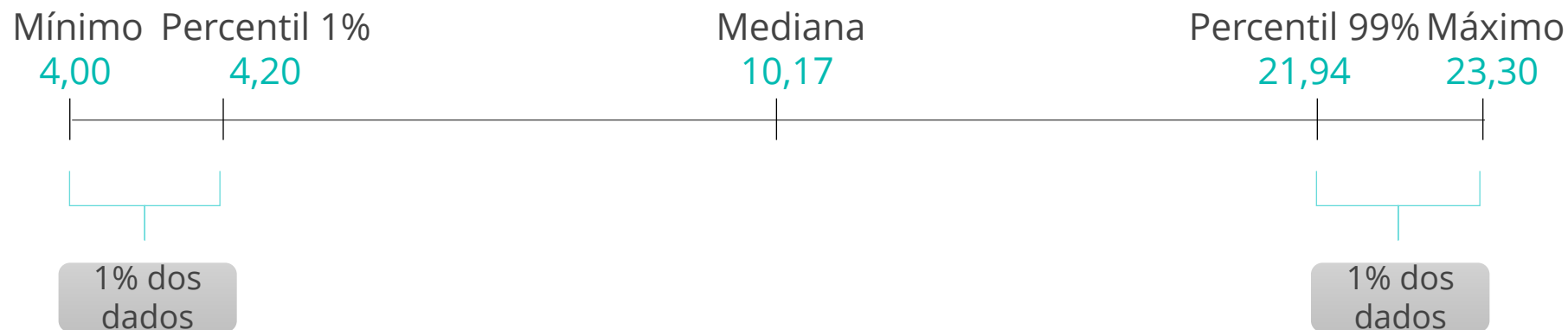
As **medidas de dispersão** são interessantes para serem utilizadas para **comparar duas ou mais populações (grupos)** de uma mesma variável.



Exemplo: Salários no case de *People Analytics*

Quando trabalhamos com grandes bancos de dados o máximo e o mínimo representam apenas 1 ponto, então comumente utilizamos o **percentil 1% (P1)** e o **percentil 99% (P99)** para complementar a análise.

A interpretação dos percentis segue a mesma lógica dos quartis.



Os dados indicam que, para o salário dos 36 funcionários analisados:

- 1% ganham menos que 4,20 salários mínimos;
- 99% ganham menos do que 21,94 salários mínimos.



Exemplo de funções para medidas de posição em Excel.

Medidas	Função do Excel
Média	=MÉDIA()
Mediana	=MED()
Moda	=MODO.ÚNICO() ou MODO()
1º quartil	=QUARTIL.INC(;1) ou QUARTIL(;1)
2º quartil	=QUARTIL.INC(;2) ou QUARTIL(;2)
3º quartil	=QUARTIL.INC(;3) ou QUARTIL(;3)
Percentil 1	=PERCENTIL.INC (;0,01) ou PERCENTIL(;0,01)
Percentil 99	=PERCENTIL.INC (;0,99) ou PERCENTIL(;0,99)
Mínimo	=MÍNIMO()
Máximo	=MÁXIMO()

As fórmulas em fonte menor correspondem a versões do Excel anteriores ao ano de 2007.



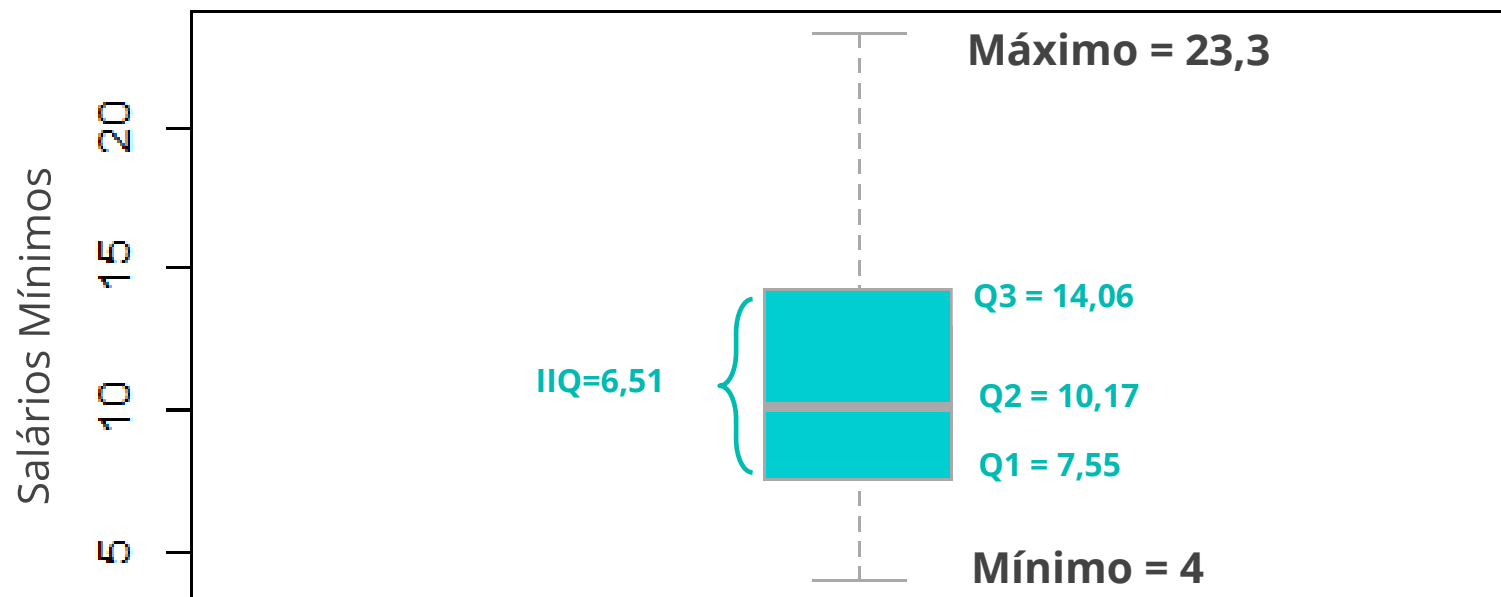
Box Plot

4.ii. BOX PLOT E OUTLIERS | DADOS QUANTITATIVOS

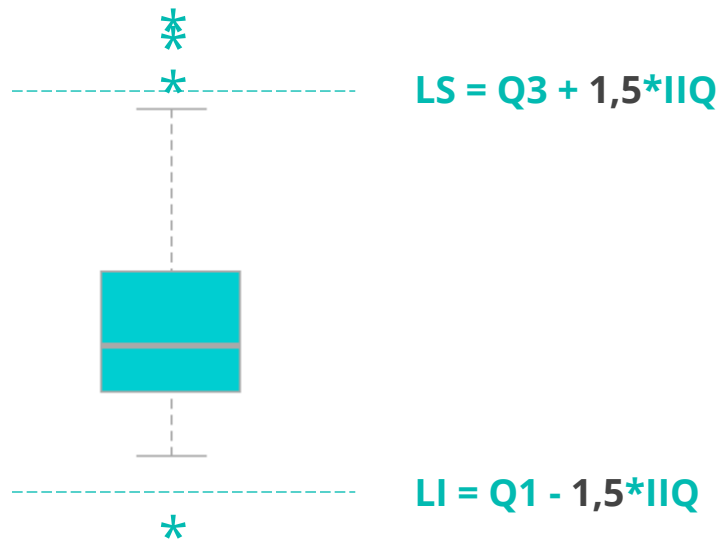
41

Box Plot é uma representação gráfica muito utilizada para apresentar a distribuição dos dados, sua composição apresenta medidas resumos muito importantes como quartis, IIQ, máximo e mínimo, além de ser útil para detecção de *outliers*.

Exemplo: Salários no case de *People Analytics*



Outliers ou dados discrepantes são as observações muito diferentes das demais, consideradas pontos fora da curva. Geralmente, são classificadas como *outliers* quando estão acima do Limite Superior (LS) ou abaixo do Limite Inferior (LI).



O valor **1,5** é comumente sugerido na literatura; porém, na prática ele pode ser parametrizado de acordo com a severidade desejada na detecção do *outlier*.



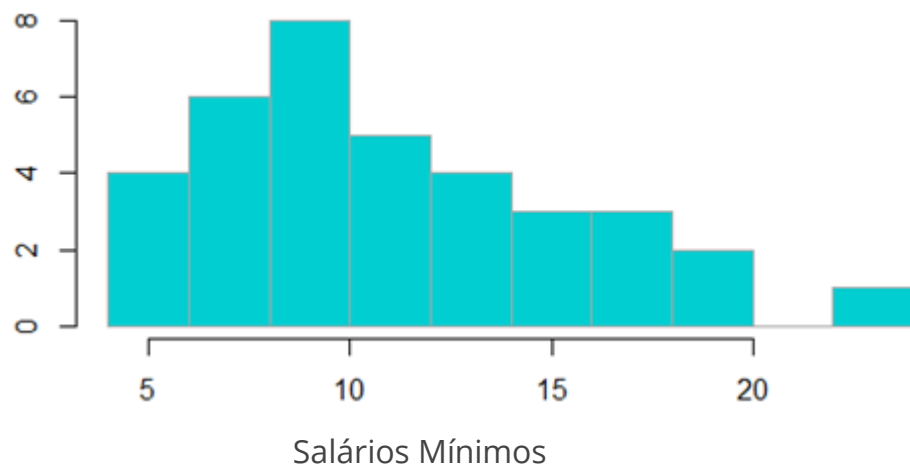
Histograma

4.iii. HISTOGRAMA | DADOS QUANTITATIVOS

43

Histograma é um gráfico conjunto de retângulos justapostos, que têm as bases sobre o eixo x (horizontal). A área é proporcional às frequências de classe ou agrupamentos de uma variável quantitativa.

Exemplo: Salários no case de *People Analytics*



É uma representação gráfica importante para avaliar o **formato** dos dados, **distribuição**, **amplitude** e **simetria**.



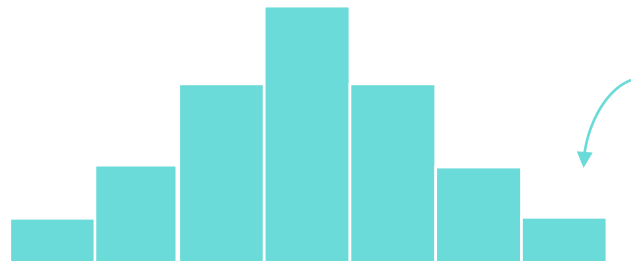
Simetria e Assimetria

4.iv. FORMATOS DA DISTRIBUIÇÃO | DADOS QUANTITATIVOS

44

Uma medida importante da **forma de uma distribuição** é chamada **coeficiente de assimetria**. Pode ser facilmente computada utilizando o R, com a função 'skewness()', ou no Excel, com '=DISTORCAO.P()'.

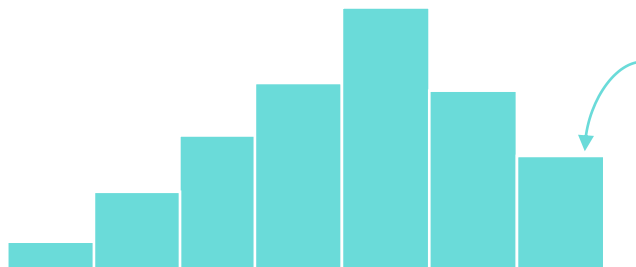
Distribuição Simétrica



Coeficiente de Assimetria = 0

- Média = Mediana

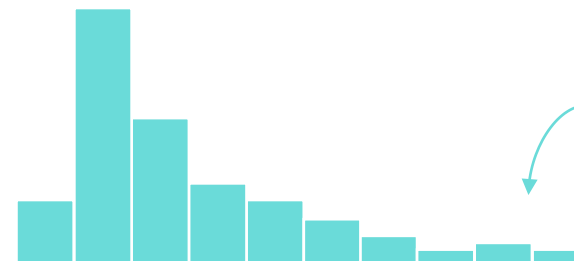
Distribuição Assimétrica a Esquerda



Coeficiente de Assimetria = -0,31

- Média < Mediana

Distribuição Assimétrica a Direita



Coeficiente de Assimetria = 1,25

- Média > Mediana



Variância Populacional

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

45

A **variância populacional** (σ^2) é a média aritmética dos desvios elevados ao quadrado. Mede a dispersão dos dados em torno da média, com referência ao N populacional.

Exemplo: Suponha que os dados abaixo são oriundos de uma população.

	Casados	Desvio	Solteiros	Desvio
	R\$ 5.200	R\$ 200	R\$ 4.000	-R\$ 1.000
	R\$ 5.000	R\$ -	R\$ 5.000	R\$ -
	R\$ 4.800	-R\$ 200	R\$ 6.000	R\$ 1.000
	R\$ 5.000	R\$ -	R\$ 5.000	R\$ -
Média	R\$ 5.000		R\$ 5.000	

A variância populacional do salário dos **casados** é dada por:

$$Variância = \frac{(200)^2 + (0)^2 + (-200)^2 + (0)^2}{4} = \frac{80.000}{4} = 20.000$$

A variância populacional do salário dos **solteiros** é dada por: 500.000



Desvio padrão Populacional

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

46

O **desvio padrão populacional** (σ) é a raiz quadrada da variância populacional, utilizado por voltar para mesma unidade de medida da média.

Exemplo:

A variância populacional do salário dos **casados** é dada por:

$$\text{Variância} = \frac{(200)^2 + (0)^2 + (-200)^2 + (0)^2}{4} = \frac{80.000}{4} = 20.000$$

O desvio padrão populacional dos **casados** é dado por:

$$\text{Desvio Padrão} = \sqrt{20.000} = 141,42$$

A variância populacional do salário dos **solteiros** é dada por: 500.000

O desvio padrão populacional dos **solteiros** é dado por:

$$\text{Desvio Padrão} = \sqrt{500.000} = 707,11$$



Variância Amostral

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

47

A **variância amostral** (s^2) é a média aritmética dos desvios elevados ao quadrado. Mede a dispersão dos dados em torno da média, com referência ao n amostral.

Exemplo: Suponha que os dados abaixo são oriundos de uma amostra.

	Casados		Desvio	Solteiros		Desvio
	R\$	5.200	R\$ 200	R\$ 4.000	-R\$ 1.000	
	R\$	5.000	R\$ -	R\$ 5.000	R\$ -	
	R\$	4.800	-R\$ 200	R\$ 6.000	R\$ 1.000	
	R\$	5.000	R\$ -	R\$ 5.000	R\$ -	
Média	R\$	5.000		R\$ 5.000		

A variância amostral do salário dos **casados** é dada por:

$$Variancia = \frac{(200)^2 + (0)^2 + (-200)^2 + (0)^2}{4 - 1} = \frac{80.000}{3} = 26.666,67$$

A variância amostral do salário dos **solteiros** é dada por: 666.666,67



Desvio Padrão Amostral

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

48

O **desvio padrão amostral** (s) é a raiz quadrada da variância amostral, utilizado por voltar para mesma unidade de medida da média.

Exemplo:

A variância amostral do salário dos **casados** é dada por:

$$\text{Variância} = \frac{(200)^2 + (0)^2 + (-200)^2 + (0)^2}{4-1} = \frac{80.000}{3} = 26.666,67$$

O desvio padrão amostral dos **casados** é dado por:

$$\text{Desvio Padrão} = \sqrt{26.666,67} = 163,30$$

A variância amostral do salário dos **solteiros** é dada por: 666.666,67

O desvio padrão amostral dos **solteiros** é dado por:

$$\text{Desvio Padrão} = \sqrt{666.666,67} = 816,50$$



Coeficiente de variação

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

49

O **Coeficiente de Variação (CV)** é uma medida de dispersão relativa, utilizada na comparação da variabilidade para dados que **diferem em seu valor médio**.

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}} * 100$$

Exemplo: Considerando o desvio padrão populacional:

	R\$ 5.200	R\$ 200	R\$ 4.000	-R\$ 1.000
	R\$ 5.000	R\$ -	R\$ 5.000	R\$ -
	R\$ 4.800	-R\$ 200	R\$ 6.000	R\$ 1.000
	R\$ 5.000	R\$ -	R\$ 5.000	R\$ -
Média	R\$ 5.000		R\$ 5.000	
CV	2,83		14,14	

O **CV** indica que o salário dos solteiros está mais disperso ao redor da média do que o salário dos casados.



Coeficiente de variação

4.v. MEDIDAS DE DISPERSÃO | DADOS QUANTITATIVOS

50

Quando as médias dos grupos a serem comparados diferem, a utilização do coeficiente de variação é muito importante para a comparação entre as variabilidades dos grupos.

O grupo com o **maior coeficiente de variação** pode ser considerado o grupo com **maior variabilidade**.

Exemplo: Considere os salários dos diferentes cargos abaixo:

Cargo	An. Júnior	An. Pleno	An. Sênior	Gerente	Diretor
Média	R\$ 2.000	R\$ 4.000	R\$ 10.000	R\$ 20.000	R\$ 50.000
Desvio Padrão	R\$ 200	R\$ 200	R\$ 200	R\$ 200	R\$ 200
Coeficiente de Variação	10,00	5,00	2,00	1,00	0,40

↓
Maior Variabilidade

↓
Menor Variabilidade



Exemplo de funções para medidas de dispersão em Excel.

Medidas	Função do Excel
Variância (populacional) Desvio Padrão (populacional)	=VAR.P() =DESVPAD.P()
Coeficiente de Variação (populacional)	=DESVPAD.P()/MÉDIA()*100
Variância (amostral) Desvio Padrão (amostral)	=VAR.A() =DESVPAD.A()
Coeficiente de Variação (amostral)	=DESVPAD.A()/MÉDIA()*100



5. Análise Bidimensional



Análise Bidimensional

5. ANÁLISE BIDIMENSIONAL | 3 TÓPICOS

53

- i. Qualitativo x Qualitativo
- ii. Qualitativo x Quantitativo
- iii. Quantitativo x Quantitativo



Investigação da Relação entre Variáveis

5. ANÁLISE BIDIMENSIONAL | 2 DIMENSÕES

54

Como podemos relacionar as informações da base de dados entre si?

- O grau de instrução sofre alguma influência de acordo com o estado civil?
- Pode-se afirmar que os funcionários que têm maior grau de instrução ganham mais?
- Pode-se afirmar que os funcionários mais velhos ganham mais?

N	estado_civil	grau_instrucao	n_filhos	salario	idade_anos	idade_meses	reg_procedencia
1	solteiro	ensino fundamental		4	26	3	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	5	capital
4	solteiro	ensino médio		5,73	20	10	outra
5	solteiro	ensino fundamental		6,26	40	7	outra
6	casado	ensino fundamental	0	6,66	28	0	interior
7	solteiro	ensino fundamental		6,86	41	0	interior
8	solteiro	ensino fundamental		7,39	43	4	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio		7,44	23	6	outra
11	casado	ensino médio	2	8,12	33	6	interior
12	solteiro	ensino fundamental		8,46	27	11	capital
13	solteiro	ensino médio		8,74	37	5	outra
14	casado	ensino fundamental	3	8,95	44	2	outra
15	casado	ensino médio	0	9,13	30	5	interior
16	solteiro	ensino médio		9,35	38	8	outra
17	casado	ensino médio	1	9,77	31	7	capital
18	casado	ensino fundamental	2	9,8	39	7	outra
19	solteiro	superior		10,53	25	8	interior
20	solteiro	ensino médio		10,76	37	4	interior



Distribuição de Frequências

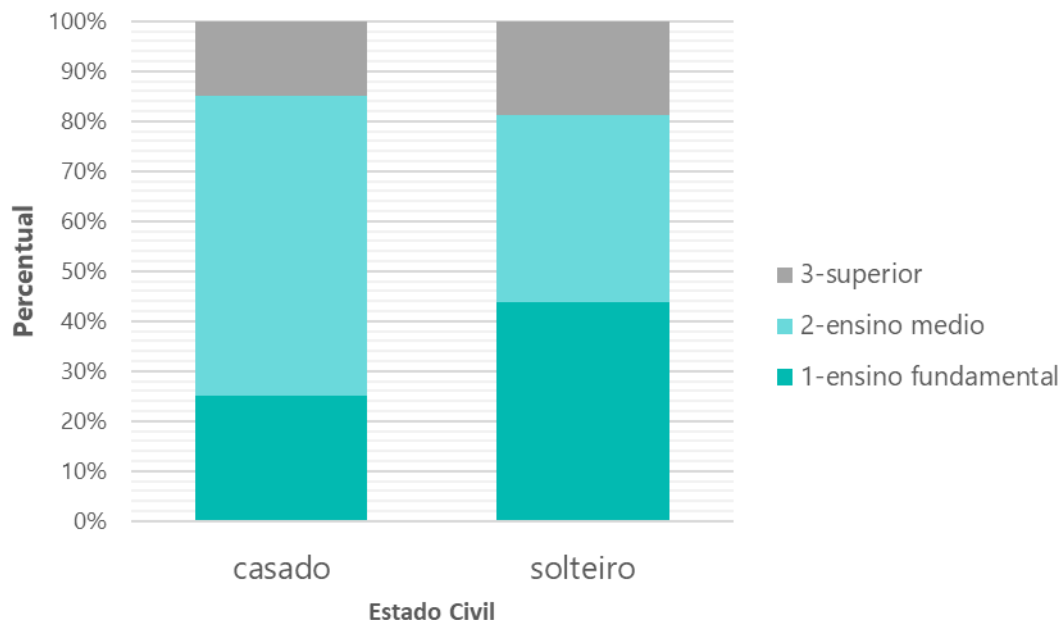
5. QUALITATIVO X QUALITATIVO | ANÁLISE BIDIMENSIONAL

55

Construir uma distribuição conjunta de 2 variáveis qualitativas é descrever a associação entre elas. Quando não há relação entre as informações, é esperado que **a distribuição das categorias de uma variável** dentro dos **grupos da outra variável** seja a mesma.

Exemplo: Case de People Analytics

- O grau de instrução sofre alguma influência de acordo com o estado civil?



Não havendo dependência entre as variáveis, esperaríamos as mesmas proporções de escolaridade para casados e solteiros.

No gráfico ao lado, descritivamente, verificamos que os solteiros possuem maior proporção de pessoas com ensino fundamental do que os casados.

Box Plot por Categoria

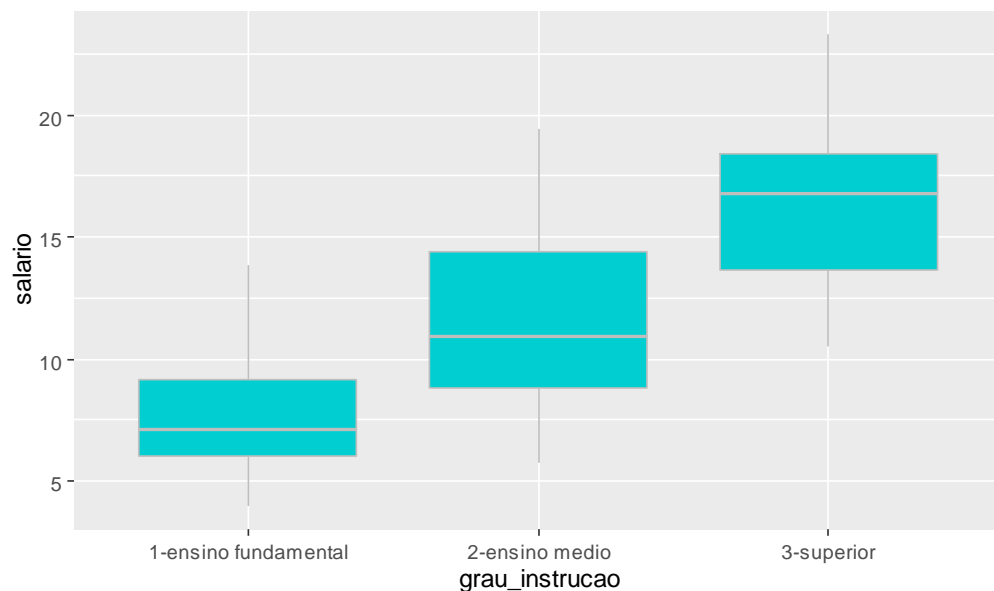
5. QUALITATIVO X QUANTITATIVO | ANÁLISE BIDIMENSIONAL

56

Pode-se analisar as medidas resumo da variável quantitativa dentro das categorias da variável qualitativa. Quando não há relação entre as informações, é esperado que as **medidas resumo da variável quantitativa** sejam semelhantes dentro das **categorias da variável qualitativa**. Pode-se utilizar medidas resumo, histogramas e *box plots* por grupo.

Exemplo: Case de People Analytics

- Pode-se afirmar que os funcionários que têm maior grau de instrução ganham mais?



Não havendo dependência entre as variáveis, esperaríamos que as medidas resumo de salários mínimos fossem similares, independentemente da categoria de grau de instrução.

No gráfico ao lado, descritivamente, verificamos que, à medida que os funcionários da empresa têm maior grau de instrução, maiores são seus salários.



Gráfico de Dispersão

5. QUANTITATIVO X QUANTITATIVO | ANÁLISE BIDIMENSIONAL

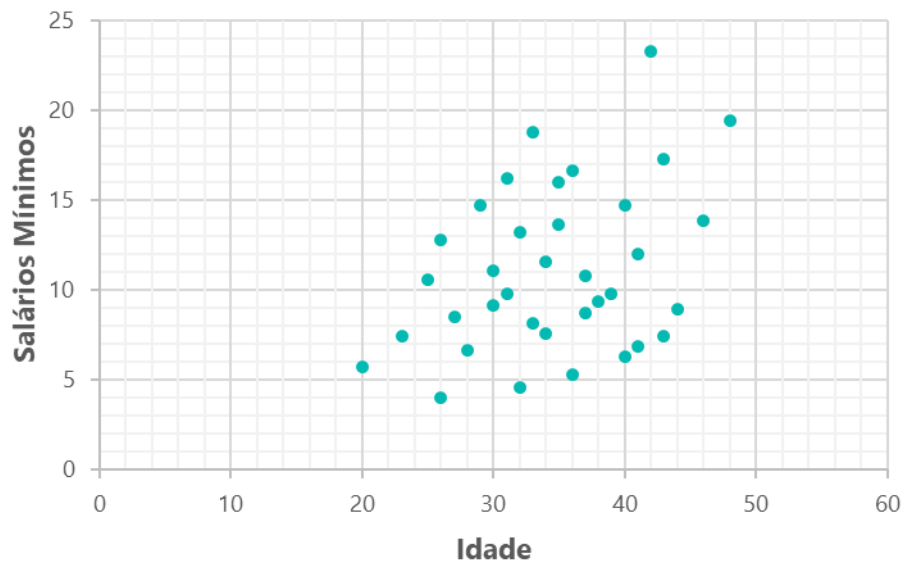
57

Para construir uma distribuição conjunta de 2 variáveis quantitativas, geralmente, utiliza-se o **gráfico de dispersão**.

Quando não há relação entre as informações, é esperado que os pontos apresentem comportamento aleatório. Quando uma variável aumenta à medida que a outra aumenta, dizemos que há associação **positiva**. Por outro lado, quando uma variável aumenta à medida que a outra decresce, existe uma associação **negativa** (ou inversa).

Exemplo: Case de People Analytics

- Pode-se afirmar que os funcionários mais velhos ganham mais?



Não havendo dependência entre as variáveis, esperaríamos que os pontos do gráfico não mostrassem tendência alguma.

No gráfico ao lado, descritivamente, verificamos que, à medida que aumenta a idade, aumenta também os salários dos funcionários.



Estudo de Caso: Limite de Crédito

ANÁLISE EXPLORATÓRIA DE DADOS | ESTUDO DE CASO

58

Esta base de dados traz informações dos clientes de um banco que solicitaram limite de crédito. Responda às seguintes questões de negócio:



- Qual a idade média dos clientes presentes no banco de dados?
- Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável idade? Interprete os valores.
- Existem clientes com idades discrepantes? Analise o boxplot.
- Existem clientes que possuem rendimento total discrepante em relação aos demais clientes? Analise o boxplot.
- A partir de qual valor o rendimento é considerado discrepante?
- A variável rendimento total pode ser considerada simétrica?
- Existem clientes que possuem salário discrepante em relação aos demais clientes? Analise o boxplot.
- A partir de qual valor o salário é considerado discrepante?
- A variável salário pode ser considerada simétrica?
- Existem clientes que possuem limite de cheque especial discrepante em relação aos demais clientes? Analise o boxplot.
- A partir de qual valor o limite de cheque especial é considerado discrepante?
- A variável limite de cheque especial pode ser considerada simétrica?

Arquivo “Exercícios.xlsx”, Aba “Problema 1” e “Base de dados 1”.
Você pode utilizar a planilha “Medidas Descritivas.xls” para obter algumas medidas e boxplot.



- Quantos clientes a base de dados possui? Quantos são mulheres? E de forma relativa, quantas são mulheres?
- Quais são os valores da média, mediana, mínimo, máximo e quartis do tempo de relacionamento?
- Com base na distribuição de frequências do tempo de relacionamento, qual a proporção de clientes que ainda não completaram 1 ano de relacionamento?
- Qual a proporção de clientes que possuem 10 anos de relacionamento?
- Qual o % de clientes que têm 1 produto? E 2 produtos? Utilize a variável Num_de_Produtos.
- Qual o total de clientes que já cancelaram os produtos? E que não cancelaram? Qual a frequência relativa de cada categoria? Considere 1 para o cliente que cancelou e 0 para o cliente que não cancelou.

@2020 LABDATA FIA. Copyright all rights reserved.



Estudo de Caso: Imóveis

ANÁLISE BIDIMENSIONAL | ESTUDO DE CASO

60

De acordo com a localização de um imóvel, sabe-se que o valor do mesmo pode variar substancialmente. Na base de dados disponibilizada são fornecidas informações sobre o valor do imóvel (R\$ mil) por m^2 , a distância para estação de metrô (km), a idade e a região.



- a) Faça a distribuição de frequências da variável idade.
- b) Faça a distribuição de frequências da variável região.
- c) Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável distância ao metrô? Interprete os valores.
- d) Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável valor do imóvel (R\$ mil) por m^2 ? Interprete os valores.

Arquivo “Exercícios.xlsx”, Aba “Problema 3” e “Base de dados 3”.



6. Códigos em R



CASE: People Analytics

6. CÓDIGOS EM R | CASE

62

```
# Tabela de frequências: grau de instrução  
table(dados$grau_instrucao) # frequências absolutas
```

1-ensino fundamental	2-ensino medio	3-superior
12	18	6

```
# prop.table: Frequências relativas  
prop.table(table(dados$grau_instrucao))
```

1-ensino fundamental	2-ensino medio	3-superior
0.3333333	0.5000000	0.1666667

Os comandos **table** e **prop.table** apresentam as frequências absolutas e relativas, respectivamente.

Note que separamos os nomes da base de dados e da variável que será analisada por meio do símbolo \$.



Arquivo: People_Analytics.xlsx
Aba "Código R"

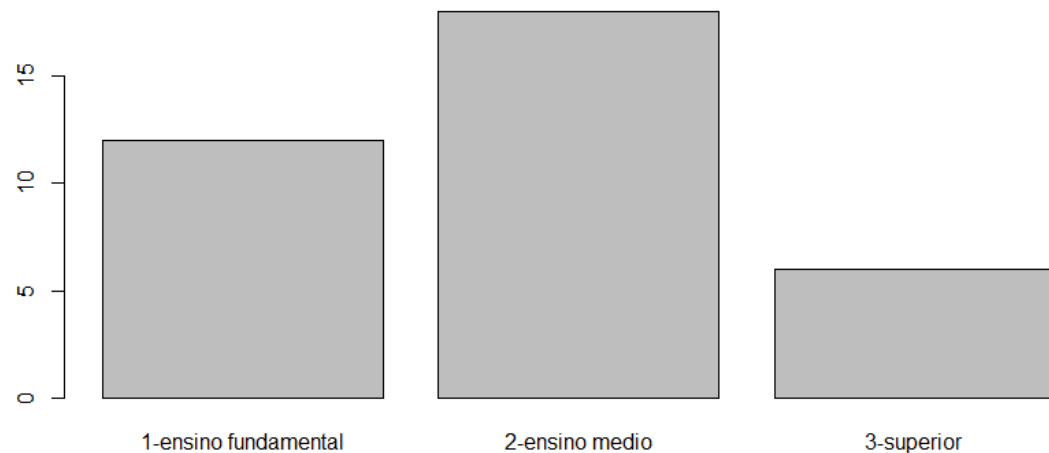


CASE: People Analytics

6. CÓDIGOS EM R | CASE

63

```
# Gráfico de Barras: grau de instrução  
barplot(table(dados$grau_instrucao))
```



O comando **barplot** cria o gráfico de barras.



Arquivo: People_Analytics.xlsx
Aba "Código R"

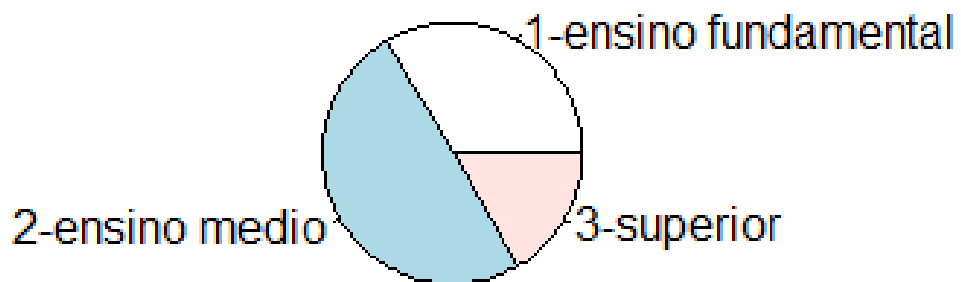


CASE: People Analytics

6. CÓDIGOS EM R | CASE

64

```
# Gráfico de pizza ou setores: grau de instrução  
pie(table(dados$grau_instrucao))
```



O comando **pie** cria o gráfico de pizza.



Arquivo: People_Analytics.xlsx
Aba "Código R"



CASE: People Analytics

6. CÓDIGOS EM R | CASE

65

```
# Medidas resumo: salário
```

```
# summary: Min, Q1, Mediana, Média, Q3, Máx  
summary(dados$salario)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.000	7.553	10.165	11.122	14.060	23.300

```
# probs: define os percentis a serem exibidos (1 e 99)  
quantile(dados$salario, probs = c(0.01,0.99))
```

1%	99%
4.196	21.935

O comando **summary** apresenta as principais medidas descritivas de uma variável, sendo:

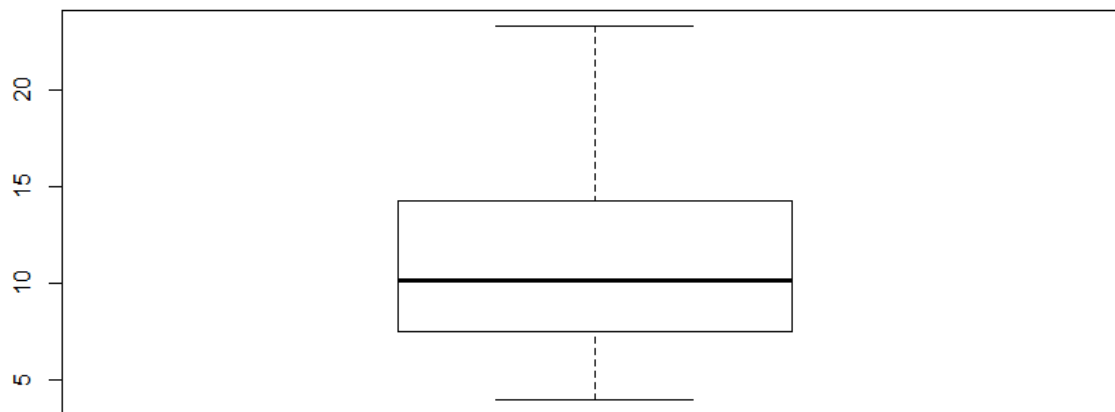
Min = Mínimo	Mean = Média
1st Qu. = Quartil 1	3rd Qu. = Quartil 3
Median = Mediana	Max = Máximo

O comando **quantile** apresenta os percentis desejados. Ao lado, selecionamos os percentis 1% e 99%.

Arquivo: People_Analytics.xlsx
Aba "Código R"



```
# Box Plot: salário  
boxplot(dados$salario)
```



O comando **boxplot** cria o gráfico de boxplot.



Arquivo: People_Analytics.xlsx
Aba "Código R"



CASE: People Analytics

6. CÓDIGOS EM R | CASE

67

```
# install.packages(moments)
library(moments)

# Coeficiente de assimetria: salário
skewness(dados$salario)

[1] 0.625682
```

O comando **skewness** calcula o coeficiente de assimetria. É necessária a instalação da biblioteca **moments**.



Arquivo: People_Analytics.xlsx
Aba "Código R"



CASE: People Analytics

6. CÓDIGOS EM R | CASE

68

```
# Análise de dados faltantes (missings)
```

```
# Resumo da variável n_filhos
```

```
summary(dados$n_filhos)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	1.00	2.00	1.65	2.00	5.00	16

```
# Na presença de missings, a média não será calculada
```

```
mean(dados$n_filhos)
```

```
[1] NA
```

```
mean(dados$n_filhos, na.rm = TRUE)
```

```
[1] 1.65
```

O comando **summary** realiza os cálculos das estatísticas descritivas, desconsiderando os *missings*.

Caso você queira realizar algum cálculo específico (fora do *summary*), é necessária a inclusão do comando **na.rm = TRUE** para exclusão dos *missings*.

R Studio®

Arquivo: People_Analytics.xlsx
Aba "Código R"



CASE: People Analytics

6. CÓDIGOS EM R | CASE

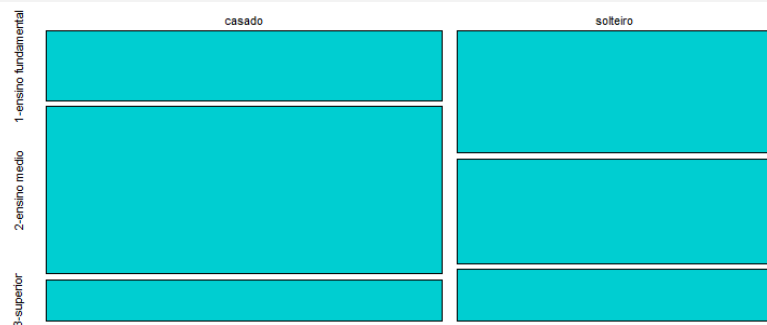
69

```
# Criar tabela de dados das variáveis
Tabela <- table(dados$estado_civil, dados$grau_instrucao)

# Visualizar a variável como tabela ou gráfico
Tabela

      1-ensino fundamental 2-ensino medio 3-superior
casado                   5          12         3
solteiro                 7           6         3

plot(Tabela, col = "darkturquoise")
```



Para a análise bivariada de variáveis qualitativas, podemos calcular a tabela de frequências também pelo comando **table**.

Para o gráfico de barras, pode-se realizar o **plot** da própria tabela.

Argumento **col**: Controla a cor da gráfico.

Arquivo: People_Analytics.xlsx
Aba "Código R"

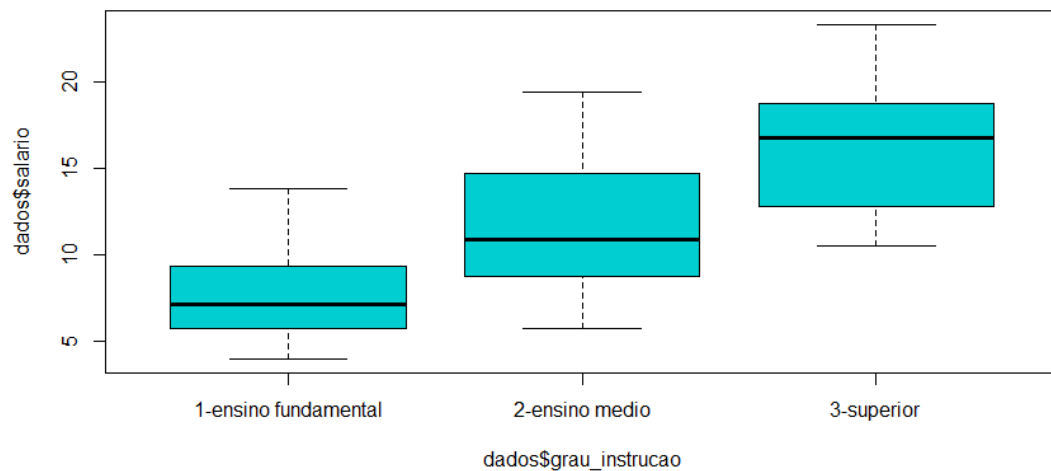


CASE: People Analytics

6. CÓDIGOS EM R | CASE

70

```
# Box plots do salário, por grau de instrução  
boxplot(dados$salario~dados$grau_instrucao,  
        col="darkturquoise")
```



Para a análise bivariada entre uma variável quantitativa e outra qualitativa, pode-se construir o boxplot por meio do comando **boxplot** com as variáveis separadas por ~.

R Studio®

Arquivo: People_Analytics.xlsx
Aba "Código R"

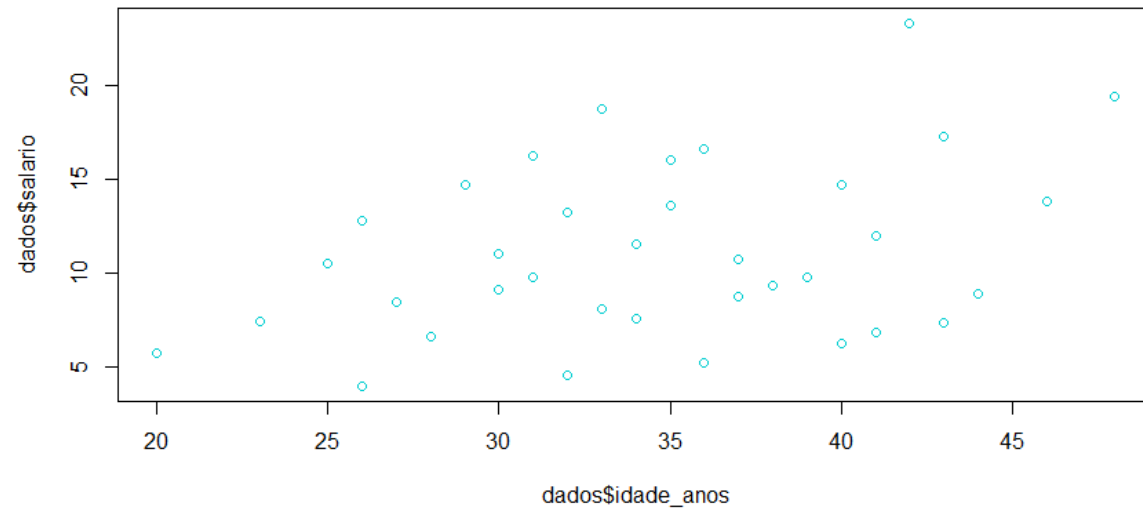


CASE: People Analytics

6. CÓDIGOS EM R | CASE

71

```
# Gráfico de dispersão entre idade e salário  
plot(dados$idade_anos,dados$salario,  
      col="darkturquoise")
```



Arquivo: People_Analytics.xlsx
Aba "Código R"

Para a análise bivariada entre duas variáveis quantitativas, deve-se construir o gráfico de dispersão pelo comando **plot**.



7. Exercícios de Fixação



Escolha uma das alternativas para cada questão

ANÁLISE EXPLORATÓRIA DE DADOS | EXERCÍCIOS DE FIXAÇÃO

73

1. Qual a medida que avalia os dados mais frequentes de uma variável?

- (a) Média
- (b) Moda
- (c) Mediana
- (d) Desvio Padrão

2. Qual medida de posição está envolvida na afirmação: “25% dos maiores salários de uma empresa, são de pessoas que ganham acima de R\$10.000”?

- (a) Mediana
- (b) Q1
- (c) Q3
- (d) Média



Escolha uma das alternativas para cada questão

ANÁLISE EXPLORATÓRIA DE DADOS | EXERCÍCIOS DE FIXAÇÃO

74

3. Qual alternativa contempla todas as medidas de dispersão?

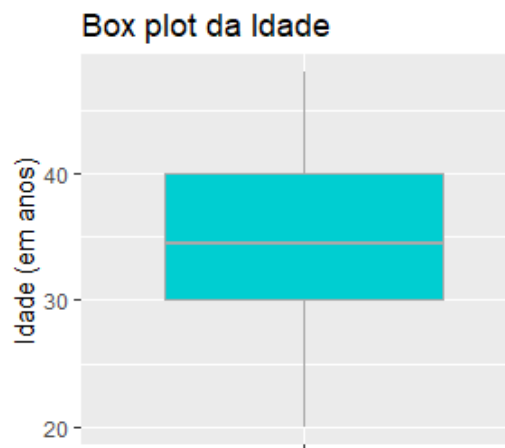
- (a) Média, amplitude, desvio padrão
- (b) Moda, intervalo inter-quartil, coeficiente de variação
- (c) Desvio padrão, intervalo inter-quartil e amplitude
- (d) Amplitude, coeficiente de variação e P99

4. O valor 35 do boxplot do lado é referente a qual medida de posição?

- (a) Q3
- (b) Mediana
- (c) Q1
- (d) Moda

5. O intervalo inter-quartil do boxplot ao lado é:

- (a) 20
- (b) 40
- (c) 10
- (d) 30



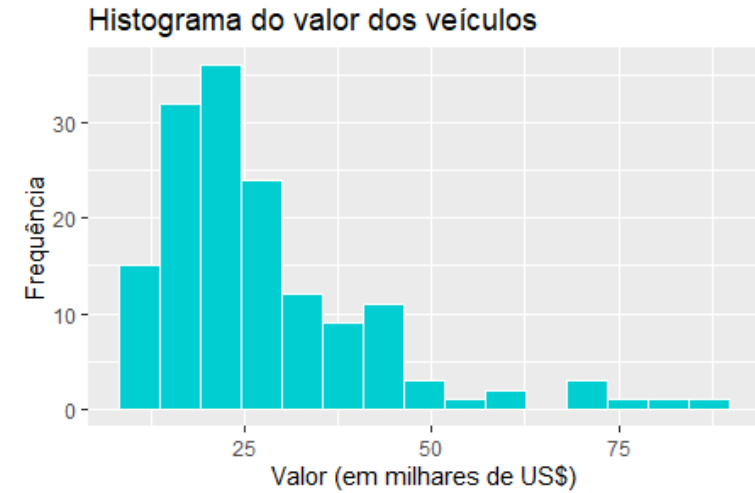
Escolha uma das alternativas para cada questão

ANÁLISE EXPLORATÓRIA DE DADOS | EXERCÍCIOS DE FIXAÇÃO

75

6. A distribuição da variável no histograma ao lado indica:

- (a) Assimetria à direita
- (b) Simetria
- (c) Assimetria à esquerda
- (d) N/A



Referências

LIVROS-TEXTO | ANÁLISE EXPLORATÓRIA DE DADOS ATÉ REGRESSÃO LINEAR MÚLTIPLA

76



1. Anderson, R. A., Sweeney, J. D. e Williams, T. A. *Estatística Aplicada à Administração e Economia*. Editora Cengage. 4ª edição, 2019.

