PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

IEC PUC MINAS

Disciplina:

MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA

Professor: Rafael Barroso

PUC Minas

PUC Minas

AVALIAÇÃO

Exercício: 15 pontos

Prática: 12 pontos

Prática: 12 pontos

Prática: 12 pontos

Prática: 12 pontos

Prática: 12 pontos

Final: 25 pontos

Aula 1
- Visão crítica de dados
- Dados estruturados e não-estruturados

Aula 2
- Principais trativas para organização e entendimento de dados
- Análise exploratória de dados
- Visualização analítica

Aula 3
- Tratamento de nulos
- *Encoding*
- Reescalagem de dados
- Discretização

Aula 4
- Limpeza de dados
- Tratamento de outliers
- Agrupamento de dados
- Amostragem

Aula 5
- Seleção de variáveis

Aula 6
- Redução de dimensionalidade

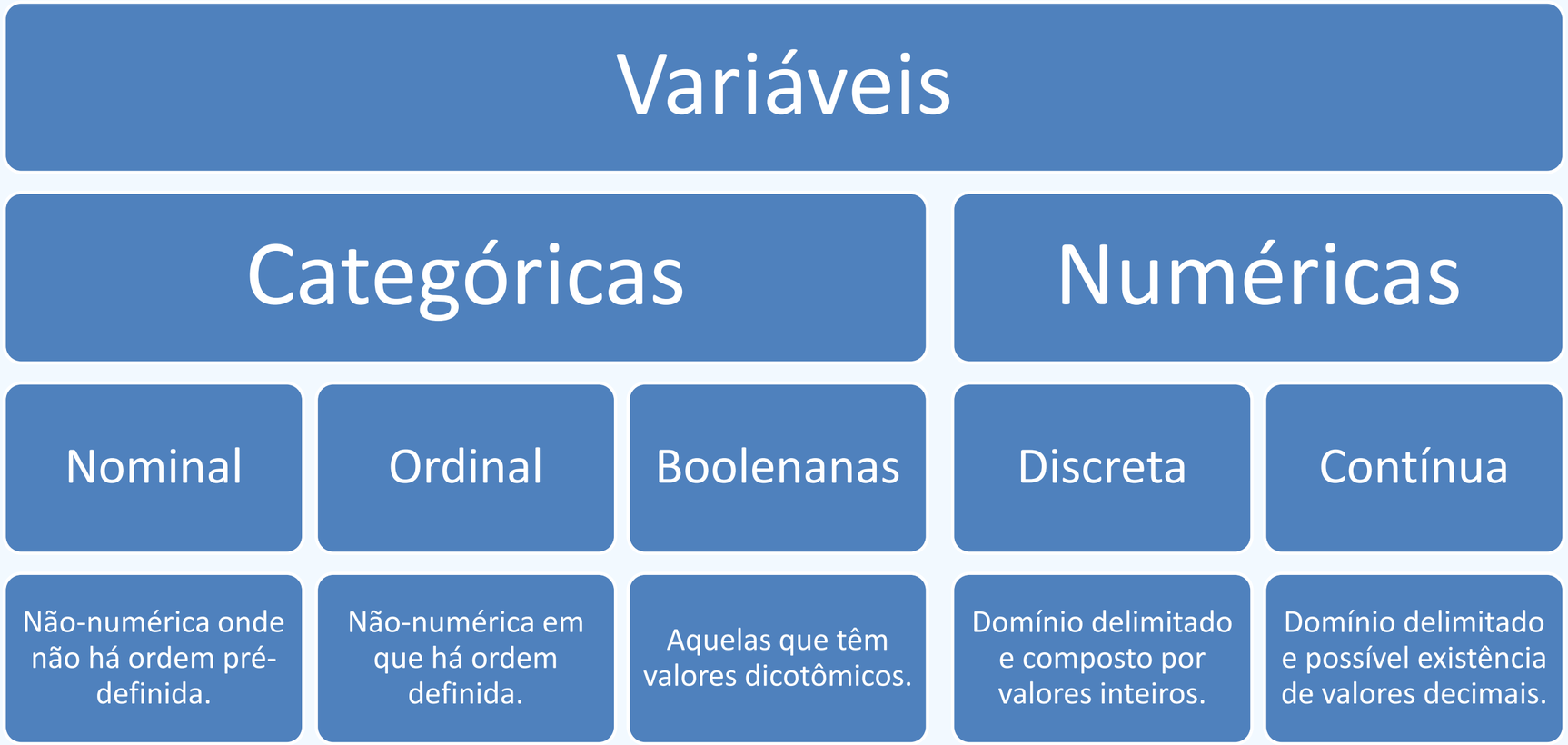
PUC Minas

PARTE 1

TIPOS DE VARIÁVEIS

PUC Minas

TIPOS DE VARIÁVEIS



PUC Minas

TIPOS DE VARIÁVEIS

Categórica Nominal	Categórica Ordinal	Numérica Discreta	Numérica Contínua
• Cor dos olhos, • Cor do cabelo, • Estado de saúde, • Cidade natal, • País, • Estilo musical, etc.	• Escolaridade, • Mês (de um dado evento).	• Número de filhos/pessoas, • Número de prédios, • Número de assentos, etc.	• Tempo, • Peso, • Altura, • Dinheiro, • Distância, etc.

PUC Minas

PARTE 2

ESTATÍSTICA DESCRITIVA

PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE TENDÊNCIA

Média

- Valor que resume as observações a um valor central.
- Resume a amostra a um valor que, se repetido *n* vezes (onde *n* é o tamanho da amostra), obtém-se o mesmo valor para a soma dos elementos da amostra.
- É fortemente influenciado por valores extremos (mínimos e máximos).
- Pode não ter um significado coerente para o domínio em questão.

Mediana

- Valor que divide as observações da amostra em dois grupos de mesmo tamanho.
- É equivalente ao percentil 50%.
- Demanda que a amostra esteja ordenada para sua identificação.
- Em amostras com número ímpar de elementos, é o próprio elemento central.
- Em amostras com número par de elementos, é a média dos dois elementos centrais.

Moda

- Valor mais recorrente na amostra.
- Pode não existir.

PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE TENDÊNCIA

12	15	23	32	4	93	14	12	13

Média

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

$$\bar{x} = \frac{12+15+23+32+4+93+14+12+13}{9}$$

$$\bar{x} = \frac{218}{9}$$

$$\bar{x} = 24{,}2$$

Mediana

4, 12, 12, 13, 14, 15, 23, 32, 93

M = 14

Moda

Número	Frequência
4	1
12	2
13	1
14	1
15	1
23	1
32	1
93	1

PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE TENDÊNCIA

76	34	64	74	42	23	59	79	26	86

Média

Mediana

Moda

PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE TENDÊNCIA

76	34	64	74	42	23	59	79	26	86

Média

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

$$\bar{x} = \frac{76+34+64+74+42+23+59+79+26+86}{10}$$

$$\bar{x} = \frac{563}{10}$$

$$\bar{x} = 56{,}3$$

Mediana

23, 26, 34, 42, 59, 64, 74, 76, 79, 86

M = 61,5

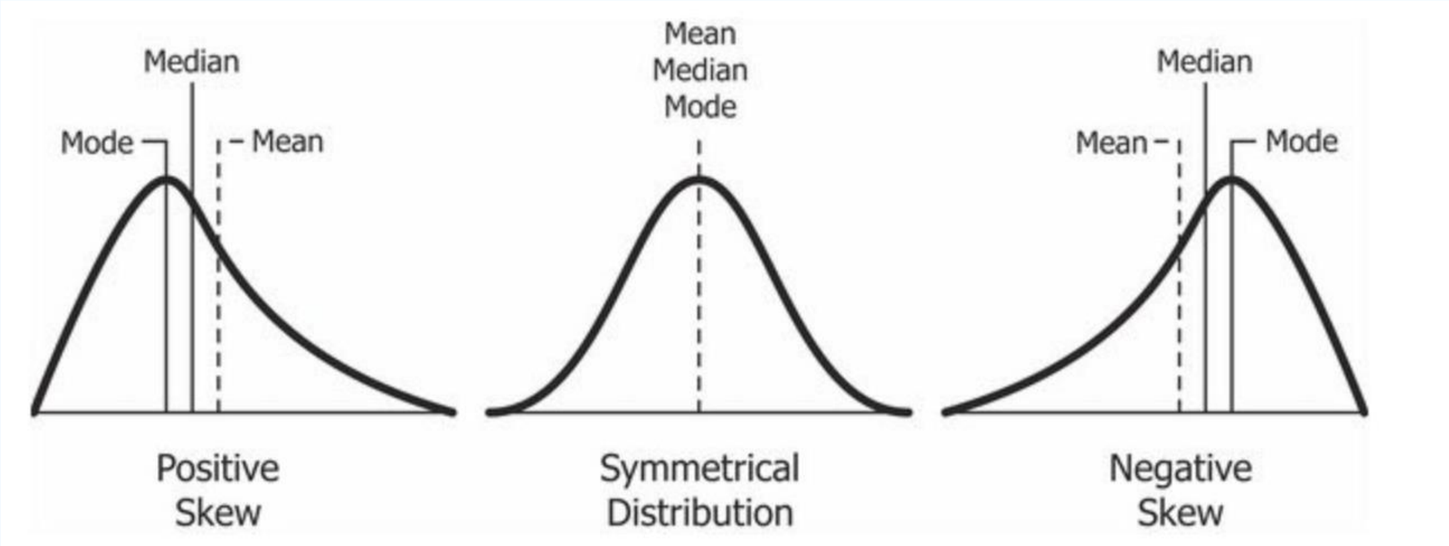
Moda

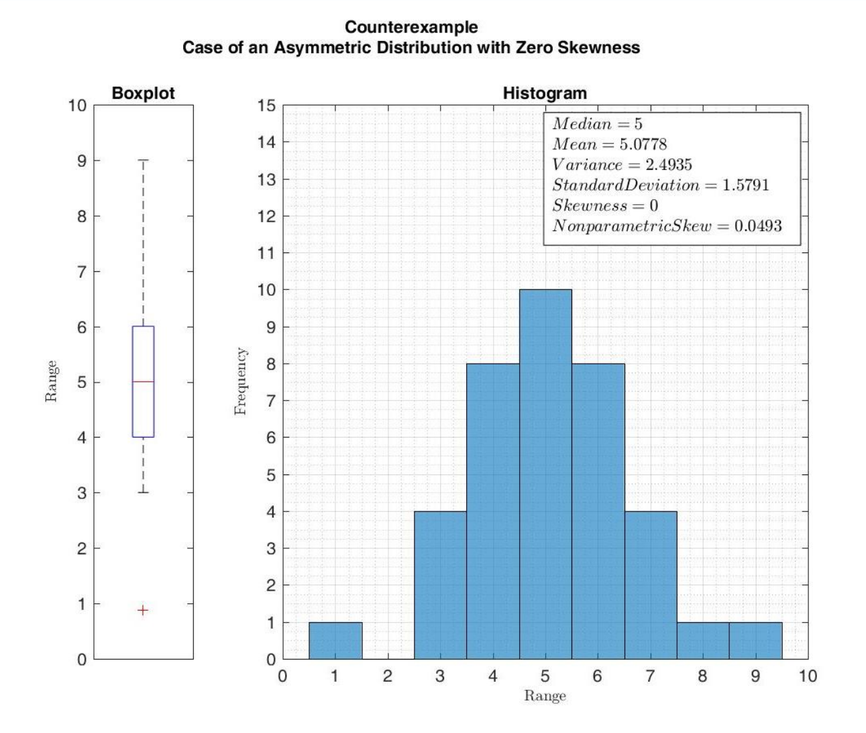
Número	Frequência
76	1
34	1
64	1
74	1
42	1
23	1
59	1
79	1
26	1
86	1

PUC Minas

ESTATÍSTICA DESCRITIVA

ASSIMETRIA





PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE VARIABILIDADE

Amplitude

- É a diferença entre os valores máximo e mínimo da amostra de dados.

Desvio Padrão

- Indica o grau de variação no conjunto de dados.
- Valores baixos indicam que os elementos se concentram próximo à média.
- Por conseguinte, valores altos indicam que os dados estão distribuídos/menos concentrados.
- É sempre positivo ou nulo (sempre nulo para constantes).

PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE TENDÊNCIA

76	34	64	74	42	23	59	79	26	86

Amplitude

$$A = x_{máx} - x_{min}$$

$$A = 86 - 23$$

$$A = 63$$

Desvio Padrão

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \overline{x})^2}{n}}$$

$$\sigma = 22{,}1$$

Variância

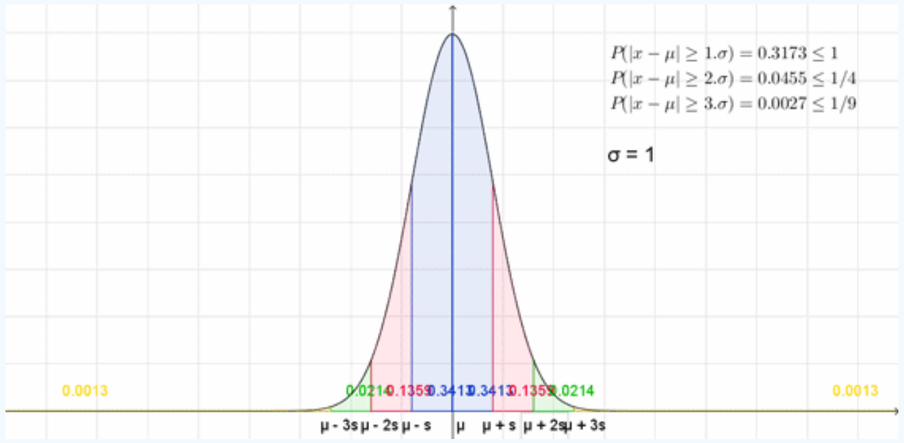
$$\sigma^2 = \frac{\sum_{i=0}^{n}(x_i - \overline{x})^2}{n}$$

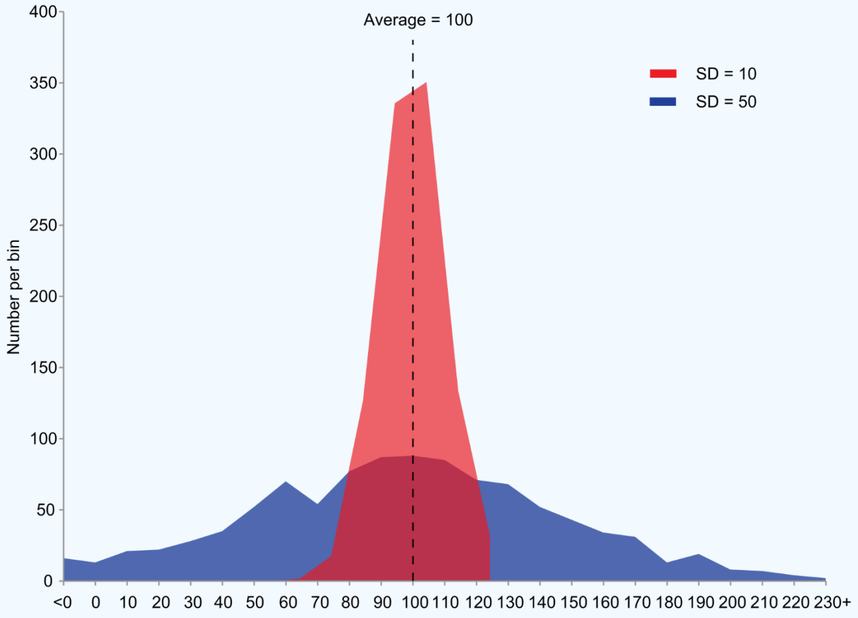
$$\sigma = 489{,}4$$

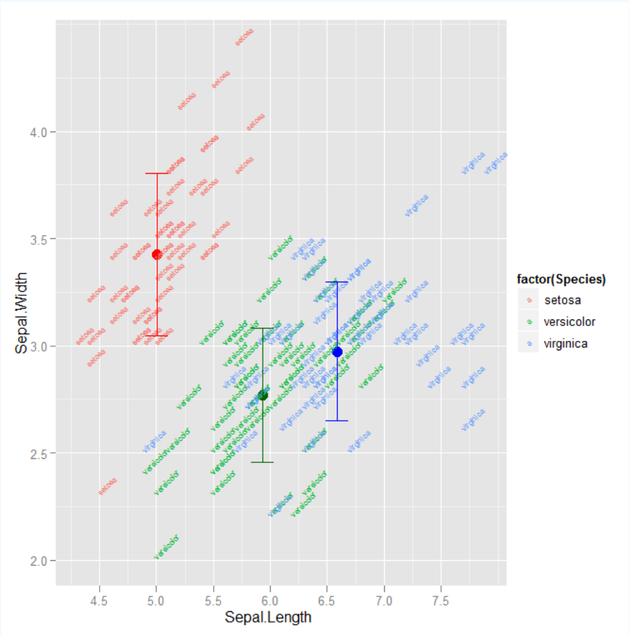
PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE VARIABILIDADE



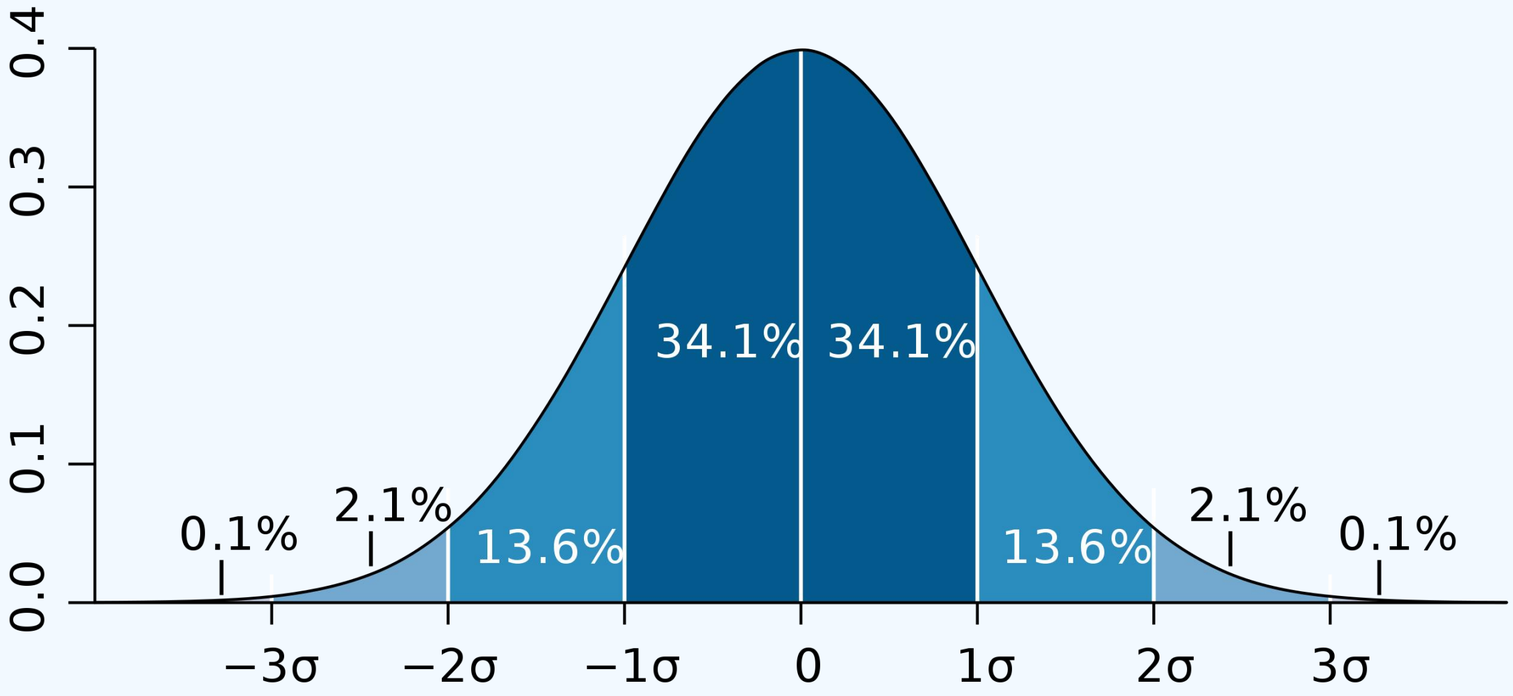




PUC Minas

ESTATÍSTICA DESCRITIVA

MEDIDAS DE VARIABILIDADE



ESTATÍSTICA DESCRITIVA

MEDIDAS DE POSIÇÃO

- Percentis, quartis, etc.
- Responsáveis por dividir a amostra ordenada de acordo com o percentual da população que desejamos;
- Ex: percentil 10% ➔ equivale ao valor (no domínio da amostra) que separa os primeiros 10% de observações dos outros 90%;
- Percentis notáveis:
  - 1º quartil ➔ percentil 25%,
  - 2º quartil ➔ percentil 50% ➔ mediana,
  - 3º quartil ➔ percentil 75%.

ESTATÍSTICA DESCRITIVA

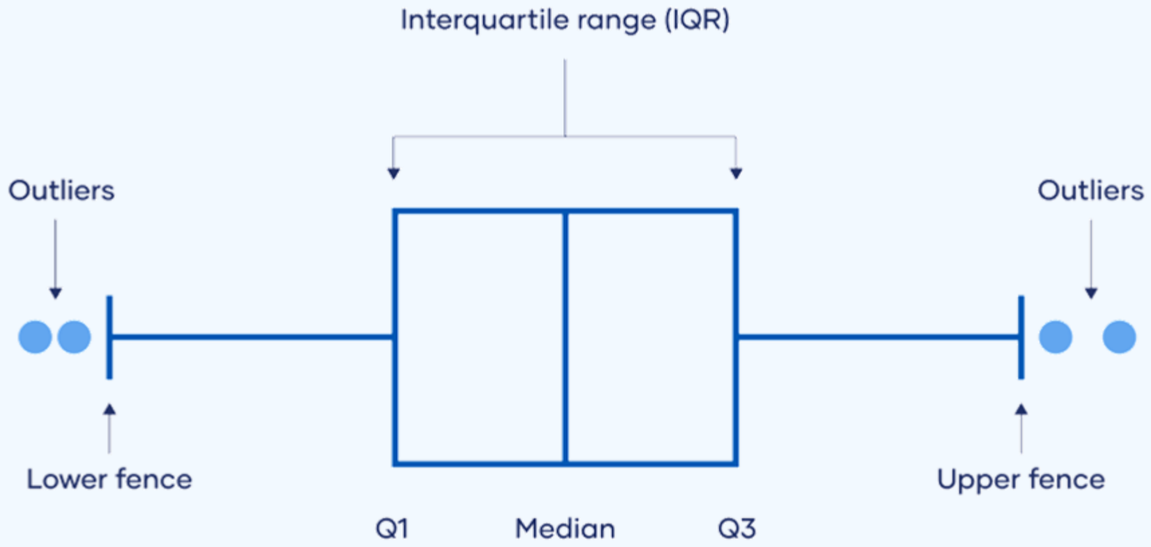
OUTLIERS

- São valores possíveis, mas incomuns;
- Em alguns casos, podem ser facilmente identificados no conjunto de dados;
- Em outros, pode-se aplicar algum teste estatístico para identifica-los;
- A principal ferramenta se baseia na distância interquartis.

PUC Minas

ESTATÍSTICA DESCRITIVA

OUTLIERS



Identificação de Outliers – Distância Interquartis:

$$Limite\ Inferior = Q_1 - 1{,}5 \times IQR$$

$$Limite\ Superior = Q_3 + 1{,}5 \times IQR$$

$$IQR = (Q_3 - Q_1)$$

Fonte: https://www.scribbr.com/statistics/outliers/#:~:text=Statistical%20outlier%20detection%20involves%20applying,can%20be%20considered%20an%20outlier.

PUC Minas

PARTE 3

ANÁLISE EXPLORATÓRIA

PUC Minas

ANÁLISE EXPLORATÓRIA

O QUE É:

- Processo inicial de análise crítica;
- Busca identificar padrões, anomalias, testar hipóteses e validar premissas;
- Tem como objetivo entender os dados antes de começar a usá-los;
- Ajuda a definir como melhor utilizar/trabalhar os dados;
- Além de validar informações conhecidas (ou supostas), pode revelar novos *insights*.

PUC Minas

ANÁLISE EXPLORATÓRIA

ABORDAGENS: