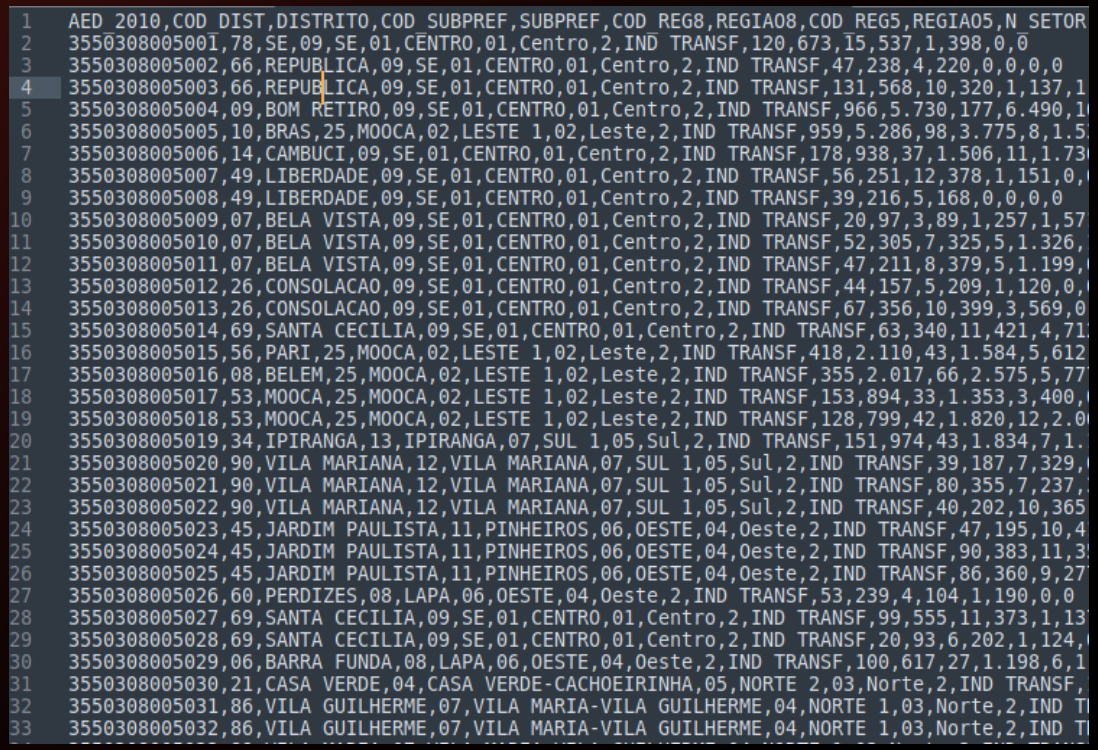
MINERAÇÃO DE DADOS

Análise descritiva de dados

ANÁLISE DESCRITIVA DE DADOS

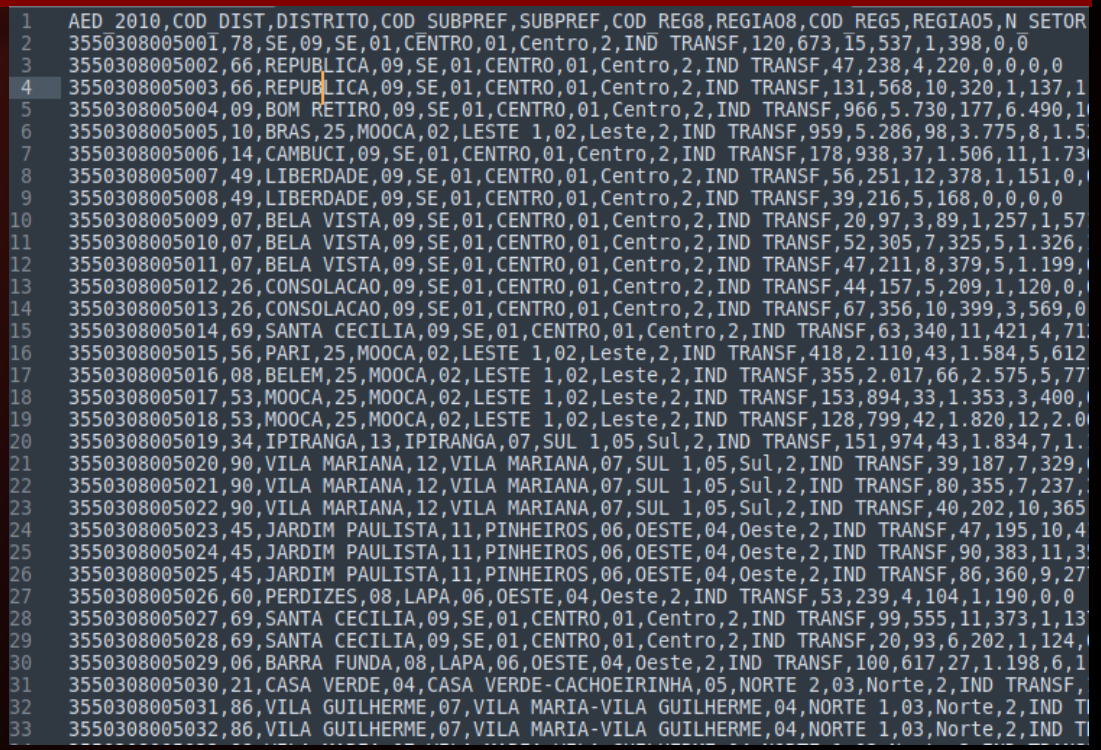
- **Bases de dados:**
  - **Centenas de atributos**
  - **Milhares de objetos**
- **Domínio desconhecido**



```
AED_2010,COD_DIST,DISTRITO,COD_SUBPREF,SUBPREF,COD_REG8,REGIAO8,COD_REG5,REGIAO5,N_SETOR
3550308005001,78,SE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,120,673,15,537,1,398,0,0
3550308005002,66,REPUBLICA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,47,238,4,220,0,0,0,0
3550308005003,66,REPUBLICA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,131,568,10,320,1,137,1
3550308005004,09,BOM RETIRO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,966,5.730,177,6.490,1
3550308005005,10,BRAS,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,959,5.286,98,3.775,8,1.5
3550308005006,14,CAMBUCI,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,178,938,37,1.506,11,1.73
3550308005007,49,LIBERDADE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,56,251,12,378,1,151,0,
3550308005008,49,LIBERDADE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,39,216,5,168,0,0,0,0
3550308005009,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,20,97,3,89,1,257,1,57
3550308005010,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,52,305,7,325,5,1.326,
3550308005011,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,47,211,8,379,5,1.199,
3550308005012,26,CONSOLACAO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,44,157,5,209,1,120,0,
3550308005013,26,CONSOLACAO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,67,356,10,399,3,569,0
3550308005014,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,63,340,11,421,4,71
3550308005015,56,PARI,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,418,2.110,43,1.584,5,612
3550308005016,08,BELEM,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,355,2.017,66,2.575,5,77
3550308005017,53,MOOCA,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,153,894,33,1.353,3,400,
3550308005018,53,MOOCA,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,128,799,42,1.820,12,2.0
3550308005019,34,IPIRANGA,13,IPIRANGA,07,SUL 1,05,Sul,2,IND TRANSF,151,974,43,1.834,7,1.
3550308005020,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,39,187,7,329,
3550308005021,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,80,355,7,237,
3550308005022,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,40,202,10,365
3550308005023,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,47,195,10,4
3550308005024,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,90,383,11,3
3550308005025,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,86,360,9,27
3550308005026,60,PERDIZES,08,LAPA,06,OESTE,04,Oeste,2,IND TRANSF,53,239,4,104,1,190,0,0
3550308005027,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,99,555,11,373,1,13
3550308005028,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,20,93,6,202,1,124,
3550308005029,06,BARRA FUNDA,08,LAPA,06,OESTE,04,Oeste,2,IND TRANSF,100,617,27,1.198,6,1
3550308005030,21,CASA VERDE,04,CASA VERDE-CACHOEIRINHA,05,NORTE 2,03,Norte,2,IND TRANSF,
3550308005031,86,VILA GUILHERME,07,VILA MARIA-VILA GUILHERME,04,NORTE 1,03,Norte,2,IND T
3550308005032,86,VILA GUILHERME,07,VILA MARIA-VILA GUILHERME,04,NORTE 1,03,Norte,2,IND T
```

ANÁLISE DESCRITIVA DE DADOS

- **Bases de dados:**
  - **Centenas de atributos**
  - **Milhares de objetos**
- **Domínio desconhecido**

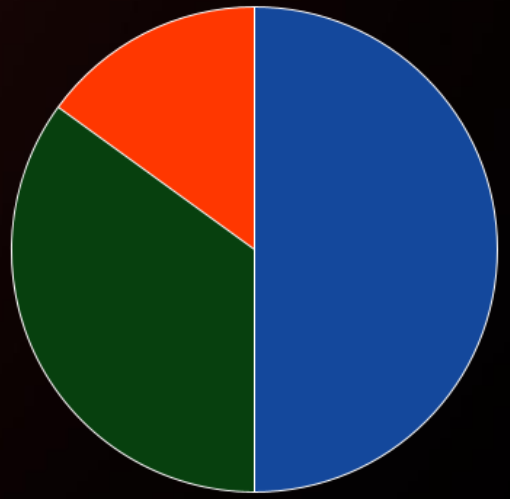


```
AED_2010,COD_DIST,DISTRITO,COD_SUBPREF,SUBPREF,COD_REG8,REGIAO8,COD_REG5,REGIAO5,N_SETOR
3550308005001,78,SE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,120,673,15,537,1,398,0,0
3550308005002,66,REPUBLICA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,47,238,4,220,0,0,0,0
3550308005003,66,REPUBLICA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,131,568,10,320,1,137,1
3550308005004,09,BOM RETIRO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,966,5.730,177,6.490,1
3550308005005,10,BRAS,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,959,5.286,98,3.775,8,1.5
3550308005006,14,CAMBUCI,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,178,938,37,1.506,11,1.73
3550308005007,49,LIBERDADE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,56,251,12,378,1,151,0,
3550308005008,49,LIBERDADE,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,39,216,5,168,0,0,0,0
3550308005009,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,20,97,3,89,1,257,1,57
3550308005010,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,52,305,7,325,5,1.326,
3550308005011,07,BELA VISTA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,47,211,8,379,5,1.199,
3550308005012,26,CONSOLACAO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,44,157,5,209,1,120,0,
3550308005013,26,CONSOLACAO,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,67,356,10,399,3,569,0
3550308005014,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,63,340,11,421,4,71
3550308005015,56,PARI,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,418,2.110,43,1.584,5,612
3550308005016,08,BELEM,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,355,2.017,66,2.575,5,77
3550308005017,53,MOOCA,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,153,894,33,1.353,3,400,
3550308005018,53,MOOCA,25,MOOCA,02,LESTE 1,02,Leste,2,IND TRANSF,128,799,42,1.820,12,2.0
3550308005019,34,IPIRANGA,13,IPIRANGA,07,SUL 1,05,Sul,2,IND TRANSF,151,974,43,1.834,7,1.
3550308005020,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,39,187,7,329,
3550308005021,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,80,355,7,237,
3550308005022,90,VILA MARIANA,12,VILA MARIANA,07,SUL 1,05,Sul,2,IND TRANSF,40,202,10,365
3550308005023,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,47,195,10,4
3550308005024,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,90,383,11,3
3550308005025,45,JARDIM PAULISTA,11,PINHEIROS,06,OESTE,04,Oeste,2,IND TRANSF,86,360,9,27
3550308005026,60,PERDIZES,08,LAPA,06,OESTE,04,Oeste,2,IND TRANSF,53,239,4,104,1,190,0,0
3550308005027,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,99,555,11,373,1,13
3550308005028,69,SANTA CECILIA,09,SE,01,CENTRO,01,Centro,2,IND TRANSF,20,93,6,202,1,124,
3550308005029,06,BARRA FUNDA,08,LAPA,06,OESTE,04,Oeste,2,IND TRANSF,100,617,27,1.198,6,1
3550308005030,21,CASA VERDE,04,CASA VERDE-CACHOEIRINHA,05,NORTE 2,03,Norte,2,IND TRANSF,
3550308005031,86,VILA GUILHERME,07,VILA MARIA-VILA GUILHERME,04,NORTE 1,03,Norte,2,IND T
3550308005032,86,VILA GUILHERME,07,VILA MARIA-VILA GUILHERME,04,NORTE 1,03,Norte,2,IND T
```

Como entender as características dessas bases?

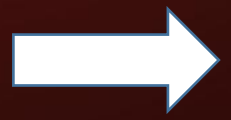
ANÁLISE DESCRITIVA DE DADOS (ADD)

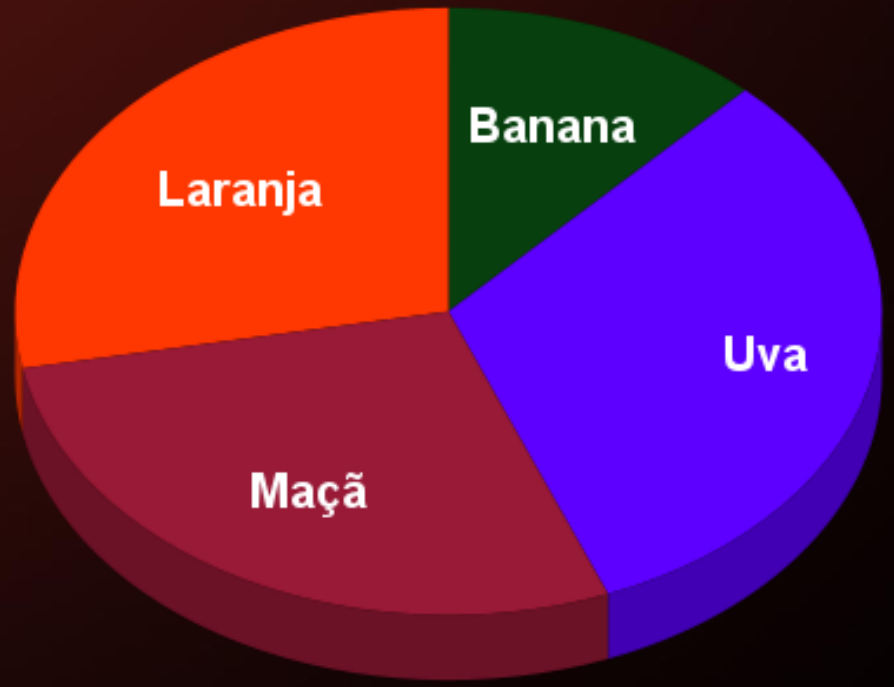
- Descrever, simplificar ou sumarizar as principais características de uma base de dados
- Análise quantitativa de dados
- Univariadas:
  - Tendência central, variação
- Bivariadas:
  - Relações entre atributos



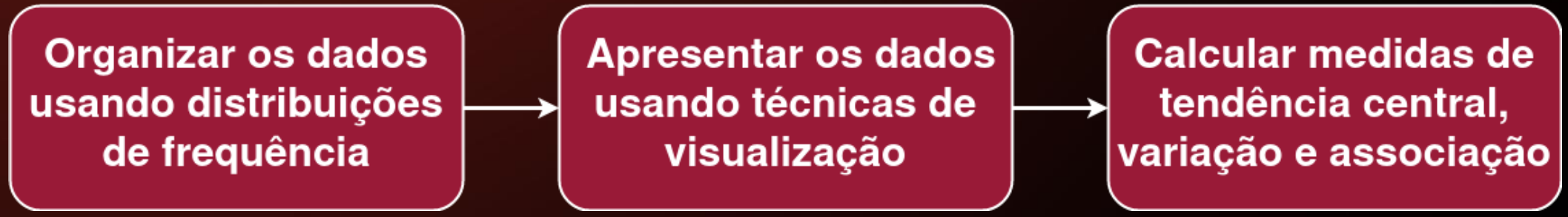
ANÁLISE DESCRITIVA DE DADOS

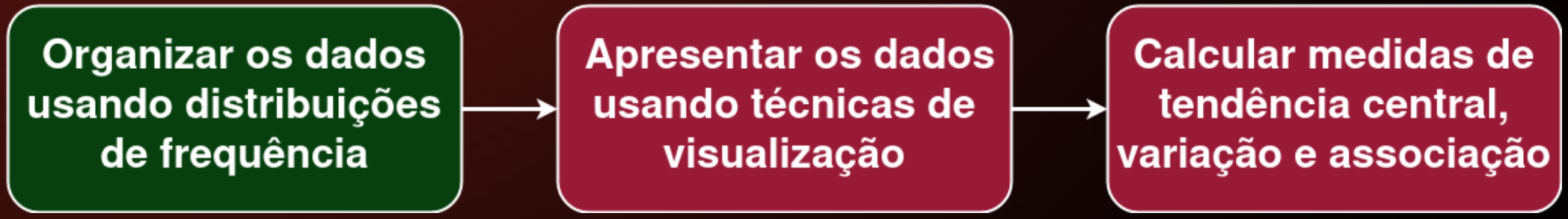
Frutas
Banana
Uva
Maçã
Laranja
...
Laranja
Uva
Banana
Maçã
Laranja
Uva





ANÁLISE DESCRITIVA DE DADOS - PROCESSO




Organizar os dados usando distribuições de frequência
Apresentar os dados usando técnicas de visualização
Calcular medidas de tendência central, variação e associação

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

- Listar os possíveis valores dos atributos
- Resumo dos dados agrupados por classes
- Base para construção de gráficos

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

- Classes: grupos / intervalos dos valores de atributos
- Limite inferior / superior de classes
- Fronteira de classes
- Ponto médio
- Amplitude
- Frequência absoluta / relativa

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Base: Mamografia

Atributo: idade

80 objetos

Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Escolher o número de classes

Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Amplitude = (maior valor - menor valor) / classes
= (81-23)/5 = 11,6 = teto(11,6)
= 12

Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	~~34,5~~	5	6,25%	6,25%
2	35	40,5	46	~~46,5~~	15	18,75%	25%
3	47	52,5	58	~~58,5~~	20	25%	50%
4	59	64,5	70	~~70,5~~	28	35%	85%
5	71	76,5	82		12	15%	100%

Separação entre as classes

Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Quantidade de objetos por classe

Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Percentual de objetos da classe

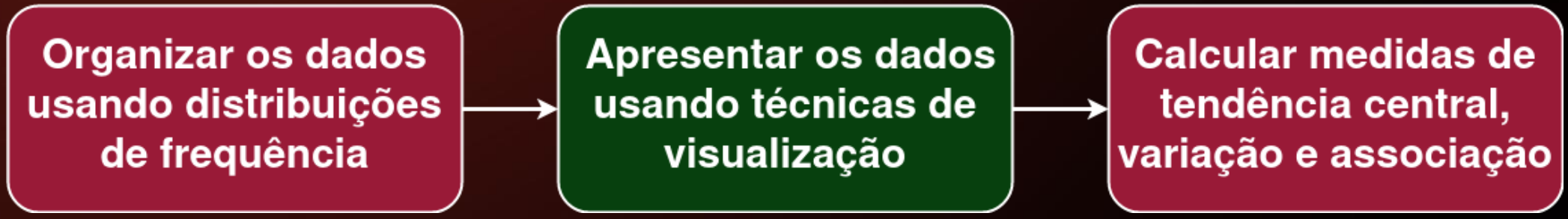
Fonte: Castro e Ferrari (2016)

ADD - DISTRIBUIÇÕES DE FREQUÊNCIA

Classe	Limite inferior	Ponto médio	Limite superior	Fronteira	Frequência absoluta	Frequência relativa	Frequência acumulada
1	23	28,5	34	34,5	5	6,25%	6,25%
2	35	40,5	46	46,5	15	18,75%	25%
3	47	52,5	58	58,5	20	25%	50%
4	59	64,5	70	70,5	28	35%	85%
5	71	76,5	82		12	15%	100%

Soma dos percentuais das classes

Fonte: Castro e Ferrari (2016)

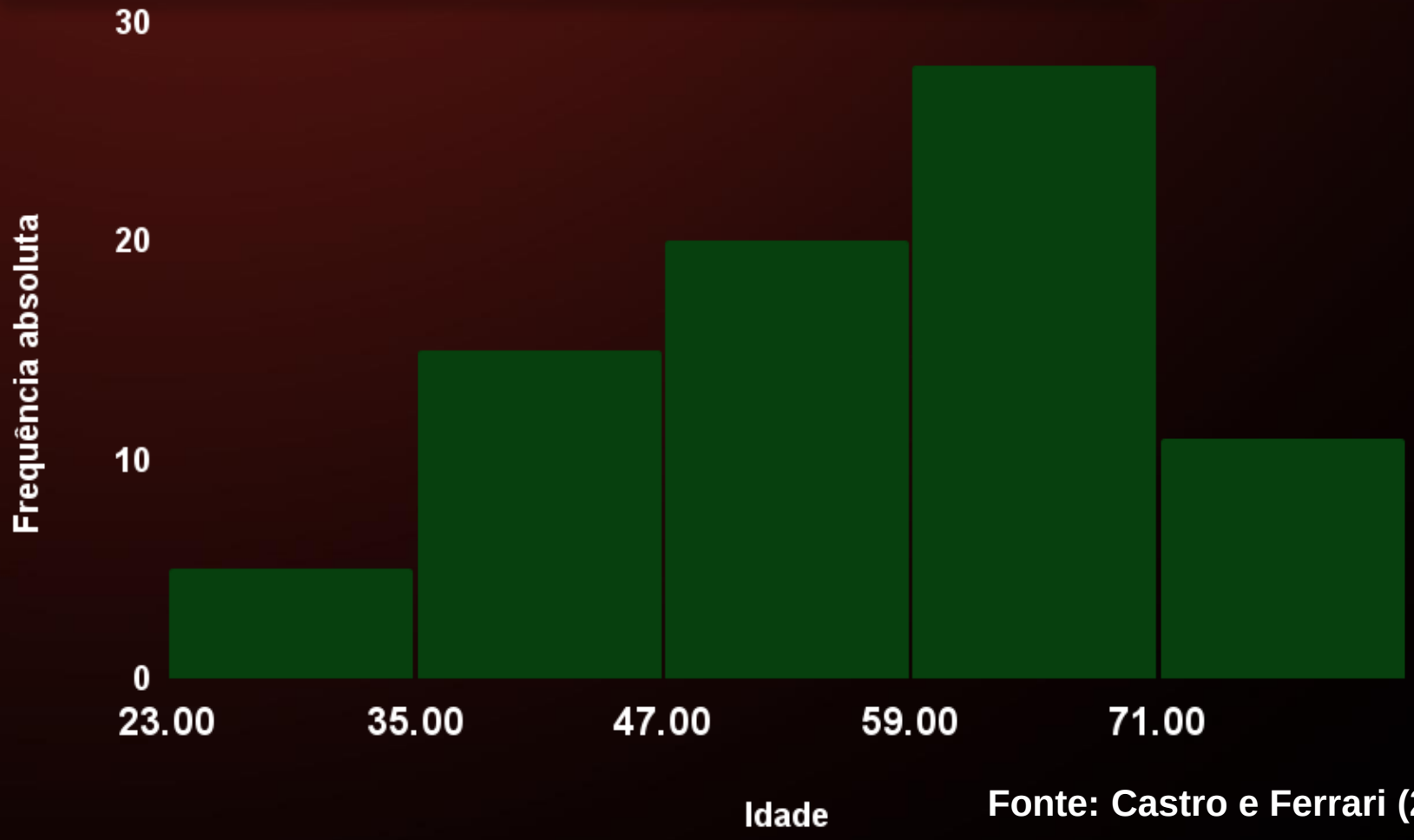

Organizar os dados usando distribuições de frequência
Apresentar os dados usando técnicas de visualização
Calcular medidas de tendência central, variação e associação

ADD - VISUALIZAÇÃO DE DADOS

- Representações visuais ajudam a extrair o conhecimento mais rapidamente
- Compartilhar o conhecimento com diferentes pessoas

ADD - VISUALIZAÇÃO DE DADOS

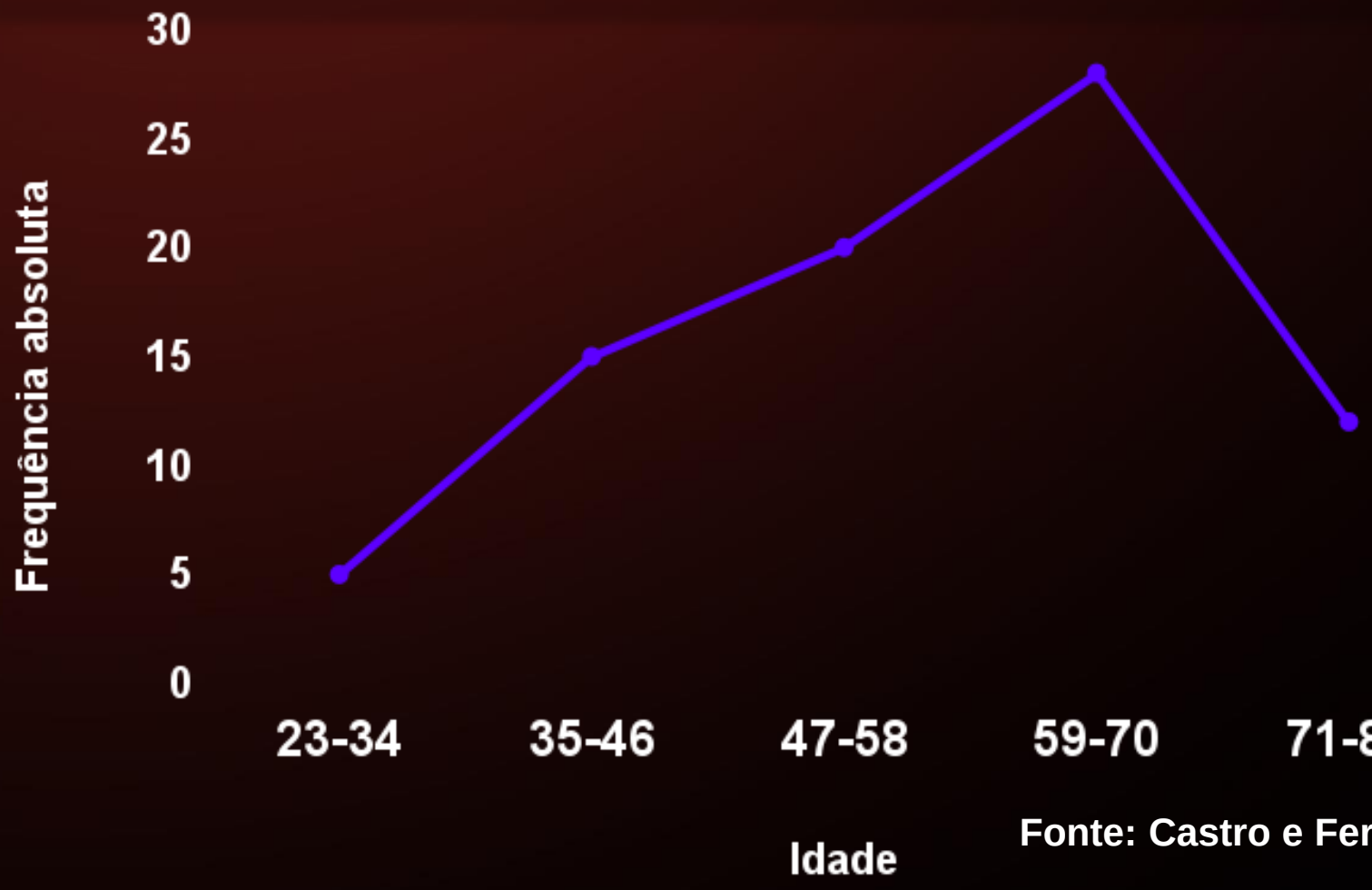
Histograma: mamografia (idade)



Fonte: Castro e Ferrari (2016)

ADD - VISUALIZAÇÃO DE DADOS

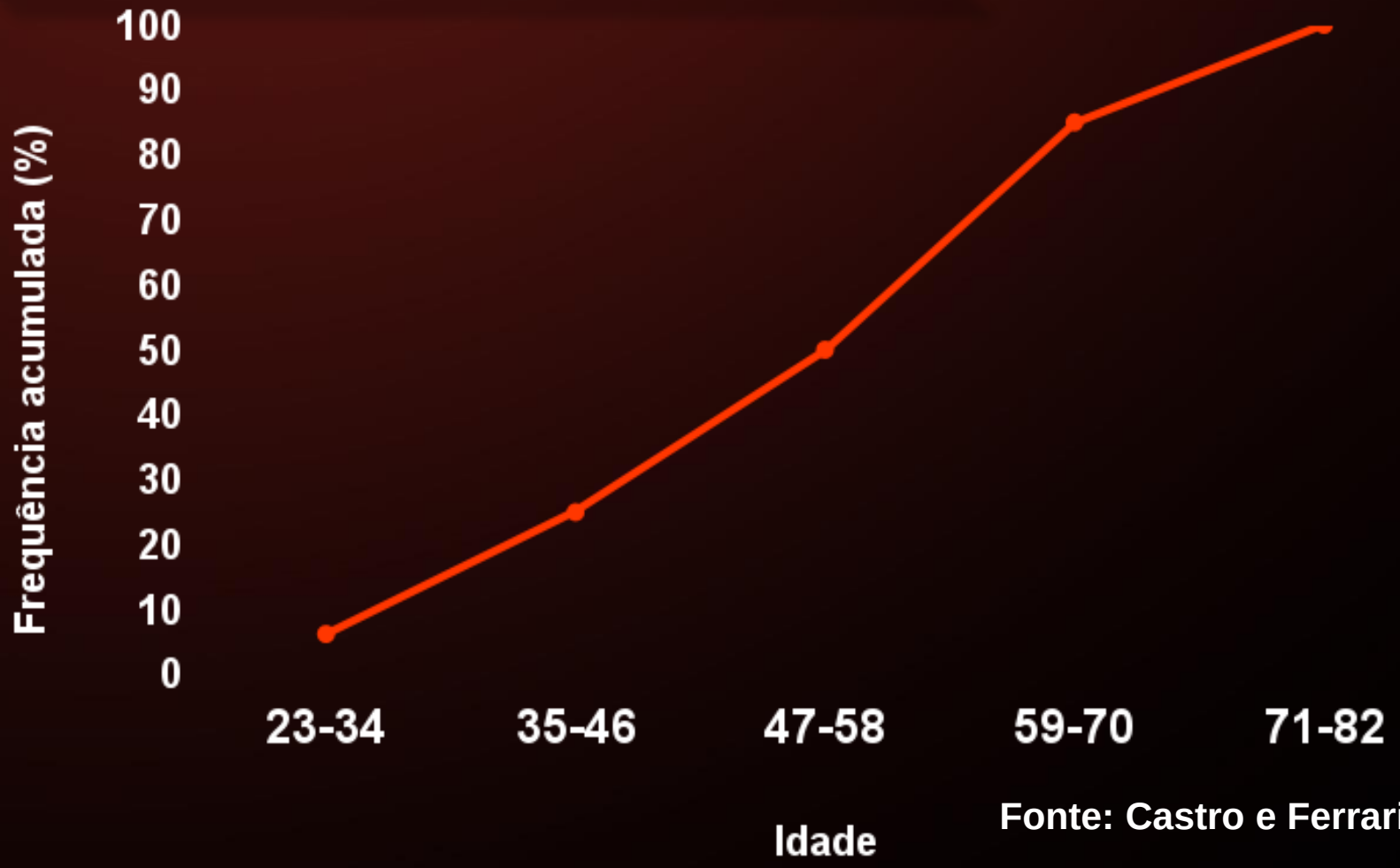
Polígono de frequências: mamografia (idade)



Fonte: Castro e Ferrari (2016)

ADD - VISUALIZAÇÃO DE DADOS

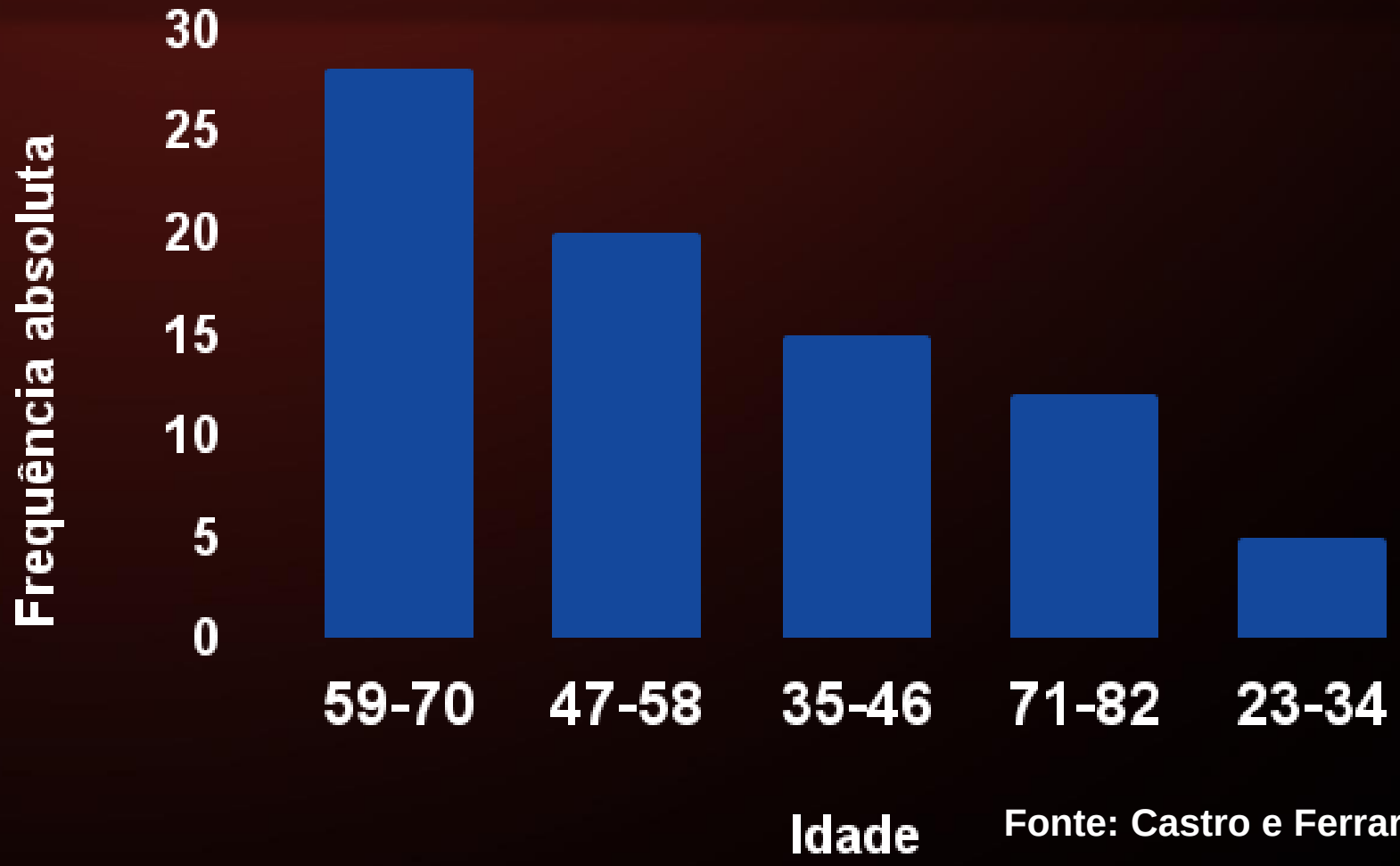
Ogiva: mamografia (idade)



Fonte: Castro e Ferrari (2016)

ADD - VISUALIZAÇÃO DE DADOS

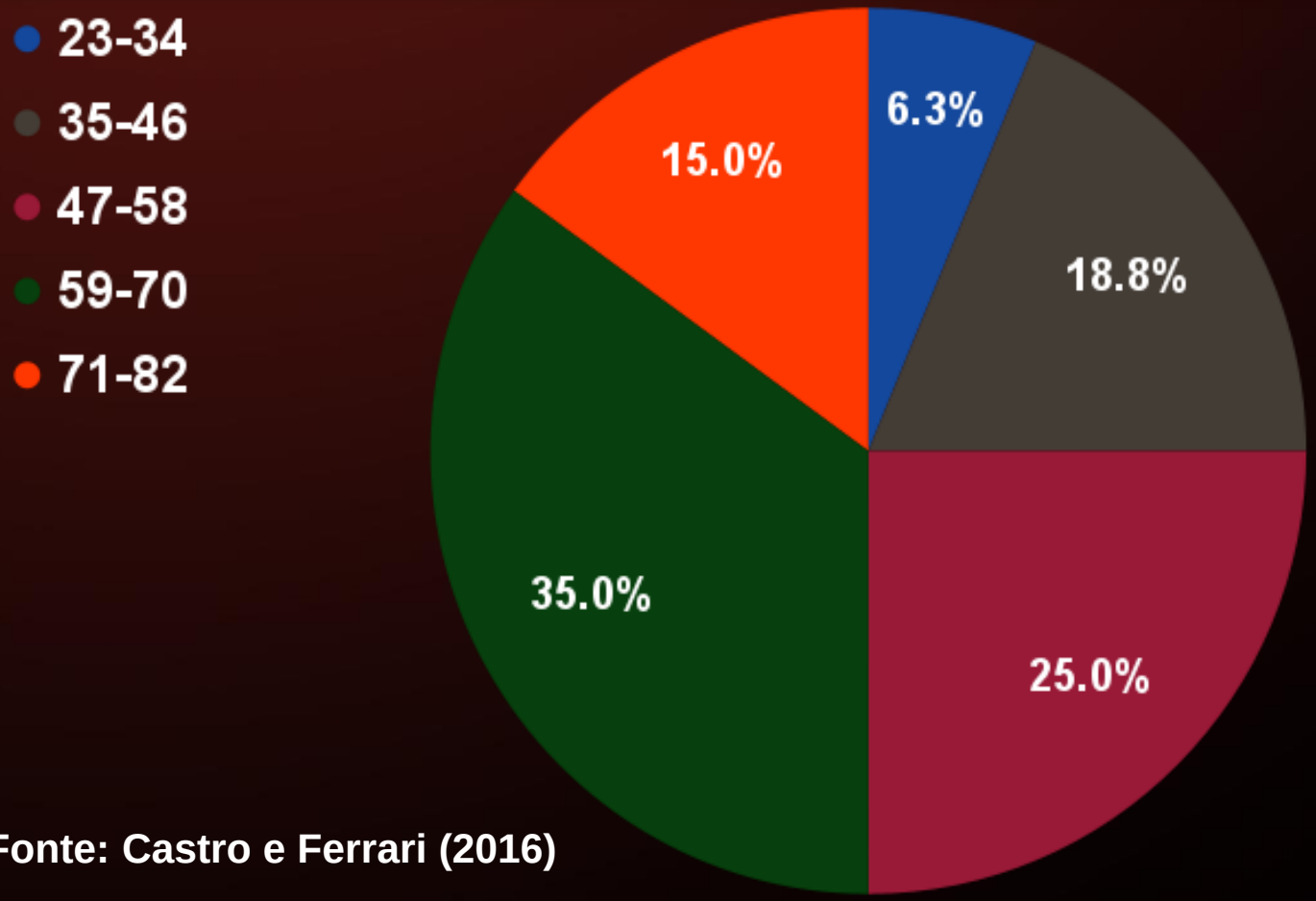
Gráfico de Pareto: mamografia (idade)



Fonte: Castro e Ferrari (2016)

ADD - VISUALIZAÇÃO DE DADOS

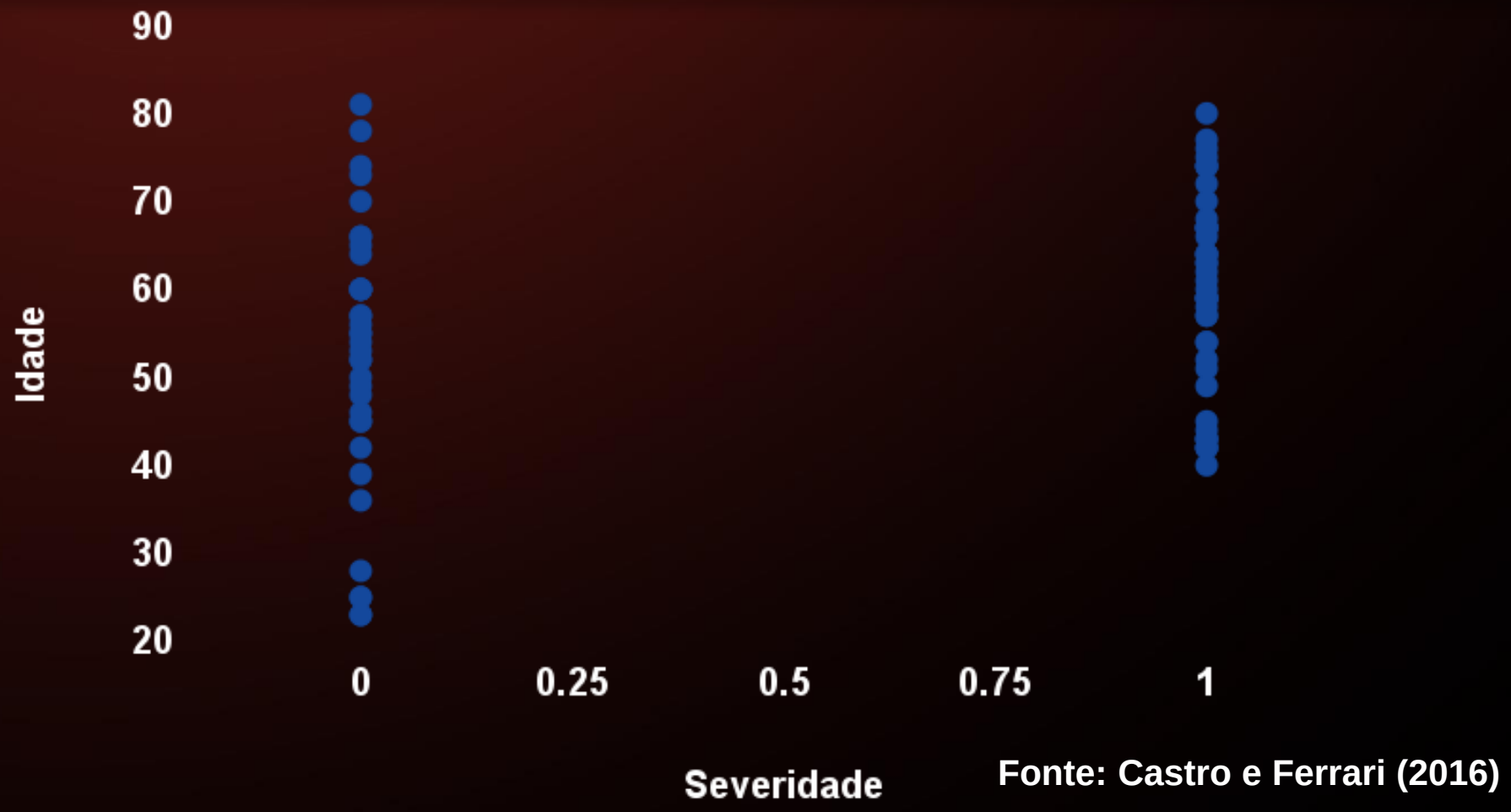
Gráfico de setores: mamografia (idade)



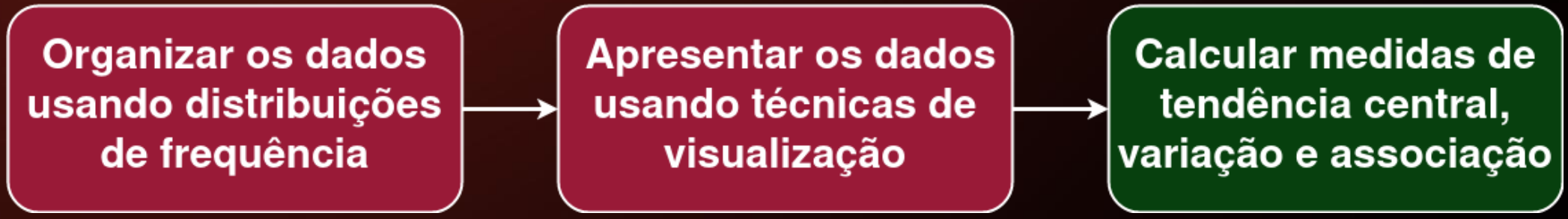
Fonte: Castro e Ferrari (2016)

ADD - VISUALIZAÇÃO DE DADOS

Gráfico de dispersão: mamografia (idade x severidade)

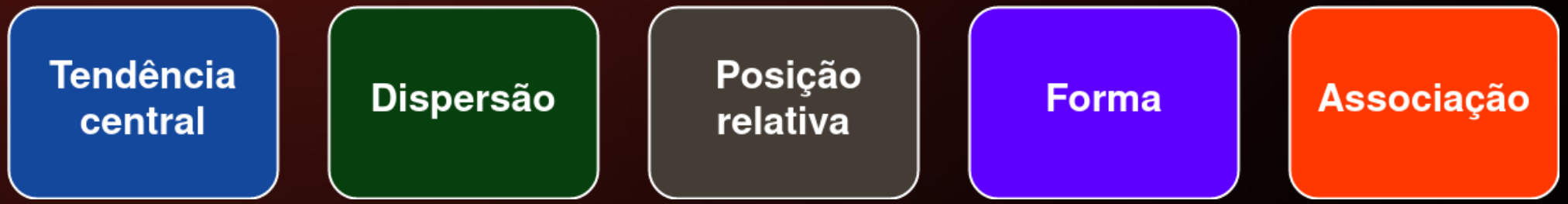


Fonte: Castro e Ferrari (2016)


Organizar os dados usando distribuições de frequência
Apresentar os dados usando técnicas de visualização
Calcular medidas de tendência central, variação e associação

ADD – MEDIDAS RESUMO

- **Mostram como os dados estão distribuídos**



ADD – MEDIDAS RESUMO

Medidas de tendência central

- **Média: sensível a valores extremos**
- **Mediana: valor do meio**
- **Moda: valor mais frequente**
- **Ponto médio: (menor + maior)/2**

ADD – MEDIDAS RESUMO

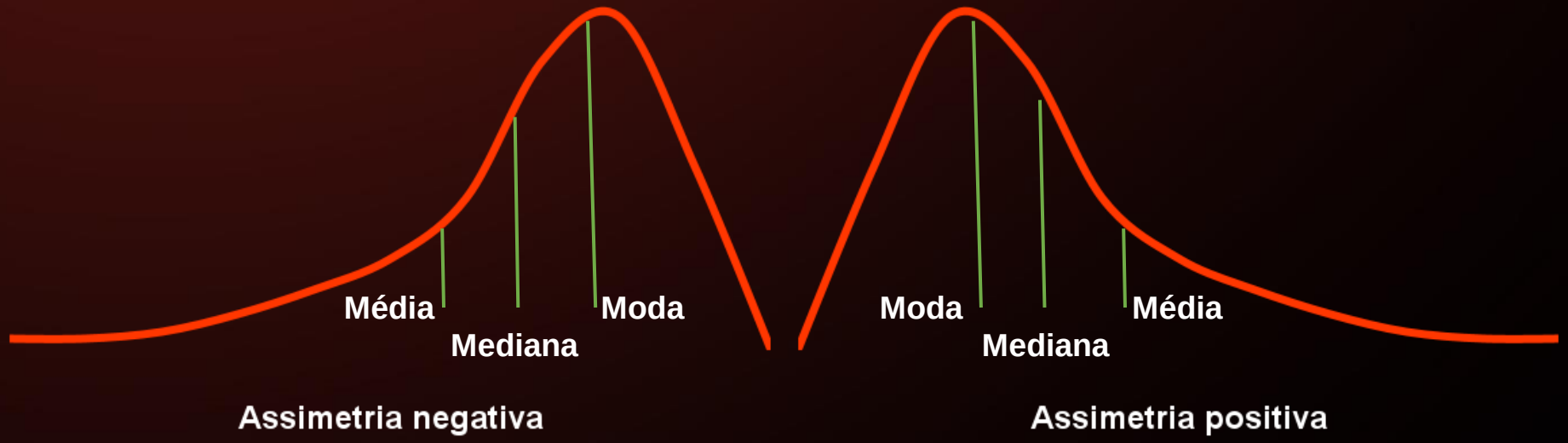
Medidas de dispersão

- **Amplitude: maior - menor**
- **Variância: dispersão dos valores de um conjunto**
- **Desvio padrão: variação em relação à média**

ADD – MEDIDAS RESUMO

Medidas de forma

- Assimetria: função de distribuição de probabilidade

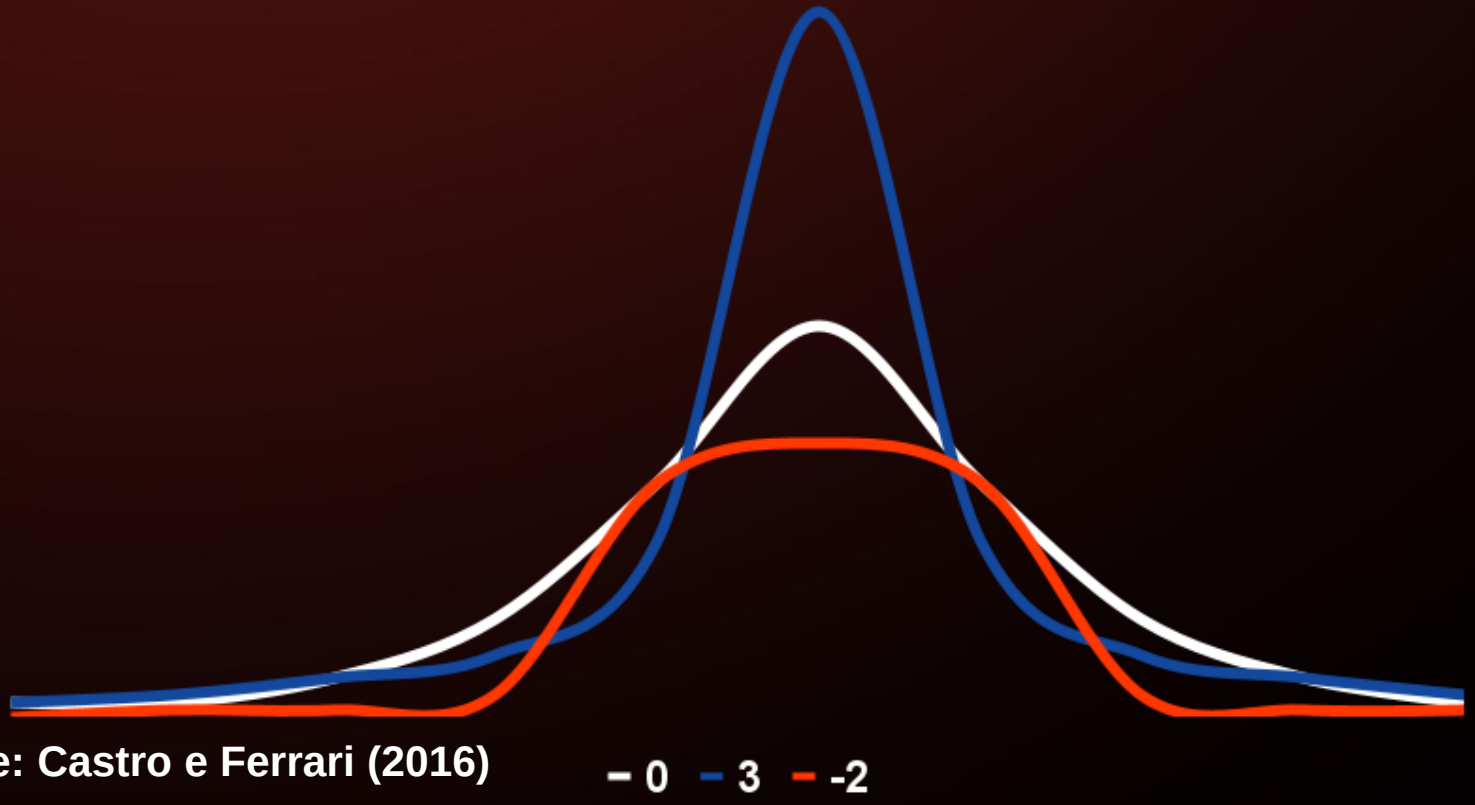


Fonte: Castro e Ferrari (2016)

ADD – MEDIDAS RESUMO

Medidas de forma

- Curtose: pico ou achatamento da função de distribuição

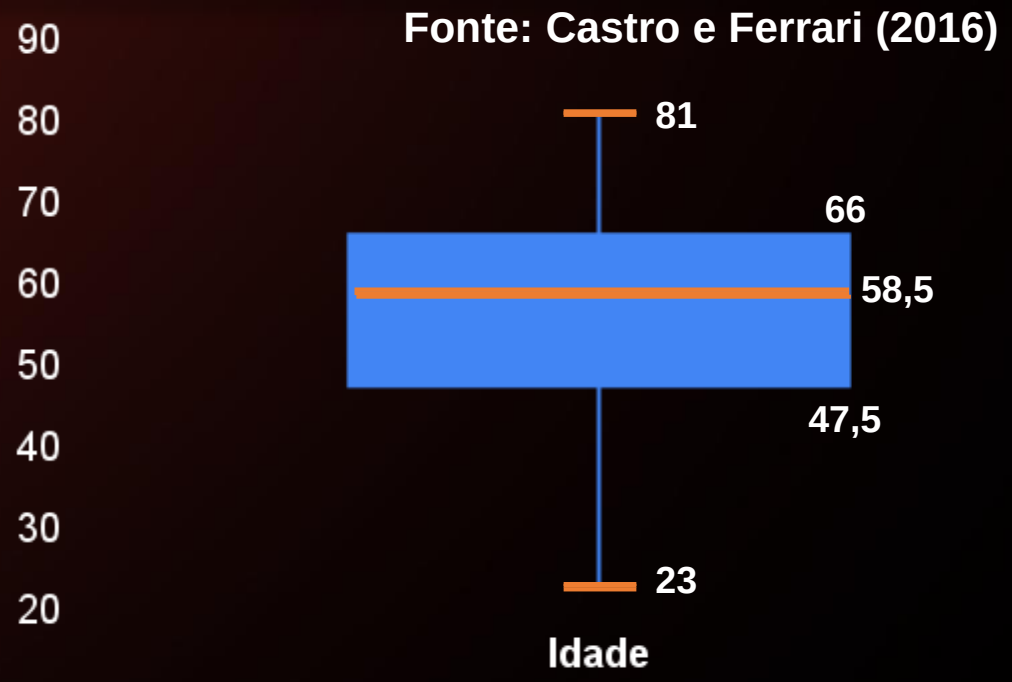


Fonte: Castro e Ferrari (2016)

ADD – MEDIDAS RESUMO

Medidas de posição relativa

- Quantis (q)
  - q = 4: quartis
  - Q_1: 25% menores valores
- Range interquartil: (Q_3 – Q_1)
- Diagrama de caixa



Fonte: Castro e Ferrari (2016)

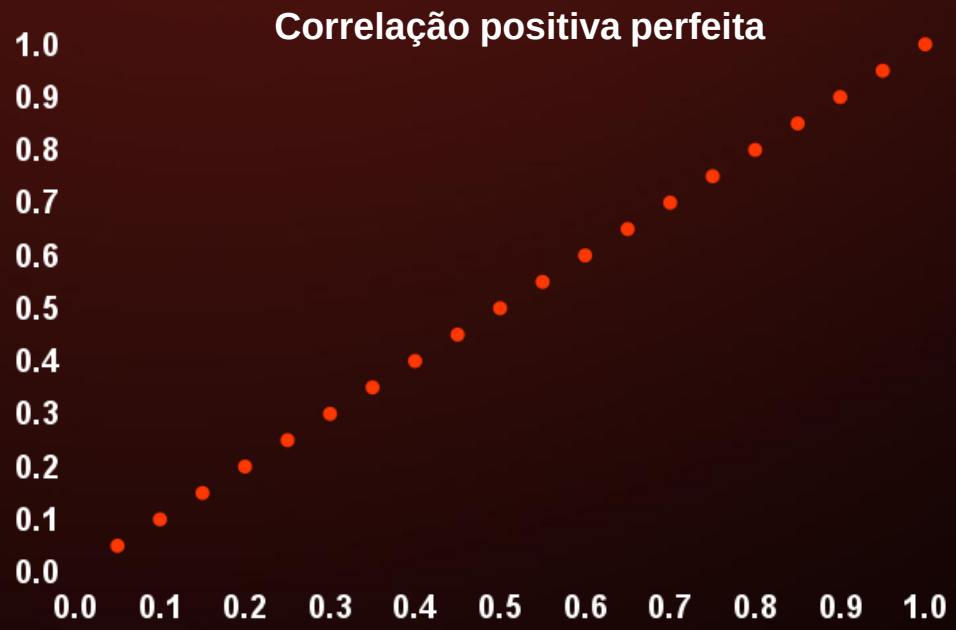
ADD – MEDIDAS RESUMO

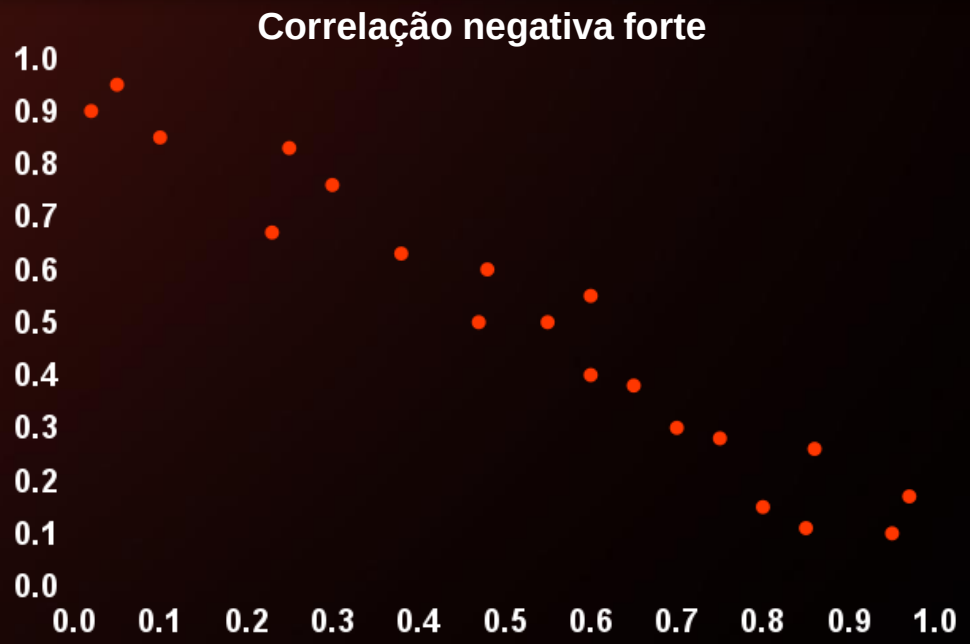
Medidas de associação

- **Verificar se há dependência entre dois atributos**
- **Coeficiente de correlação de Pearson (paramétrico)**
- **Correlação de Spearman (não paramétrico)**
- **1.0: correlação positiva perfeita**
- **0.0: sem correlação**
- **-1.0: correlação negativa perfeita**

ADD – MEDIDAS RESUMO

Medidas de associação

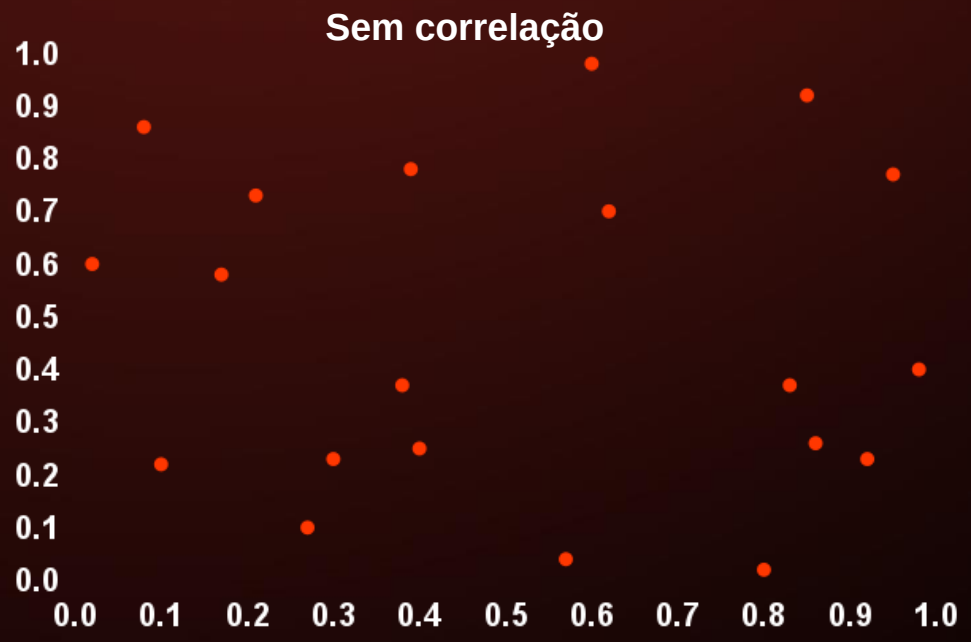




Fonte: Castro e Ferrari (2016)

ADD – MEDIDAS RESUMO

Medidas de associação



Fonte: Castro e Ferrari (2016)

REFERÊNCIAS

Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações: Cap. 3: Análise descritiva de dados.
Leandro Nunes de Castro e Daniel Gomes Ferrari. Editora Saraiva, 2016.