

# **MBA Business Analytics e Big Data**

## **Análise Exploratória de Dados**

Prof. Dr. João Rafael Dias

2º semestre - 2022

Manipulação de *strings* com `stringr`  
Manipulação de datas com `lubridate`  
Manipulação de *dataframes* com `dplyr`  
Prática no RStudio

Técnicas de visualização  
Tipos de gráficos  
Análise univariada  
Análise bivariada  
Prática no RStudio

Apresentação do curso  
Introdução ao R/Rstudio  
Estrutura de dados  
Comando básicos  
Leitura e escrita de dados  
Prática no RStudio

Tipos de variáveis  
Medidas de Centralidade  
Medidas de Dispersão  
Medidas de Associação  
Prática no RStudio

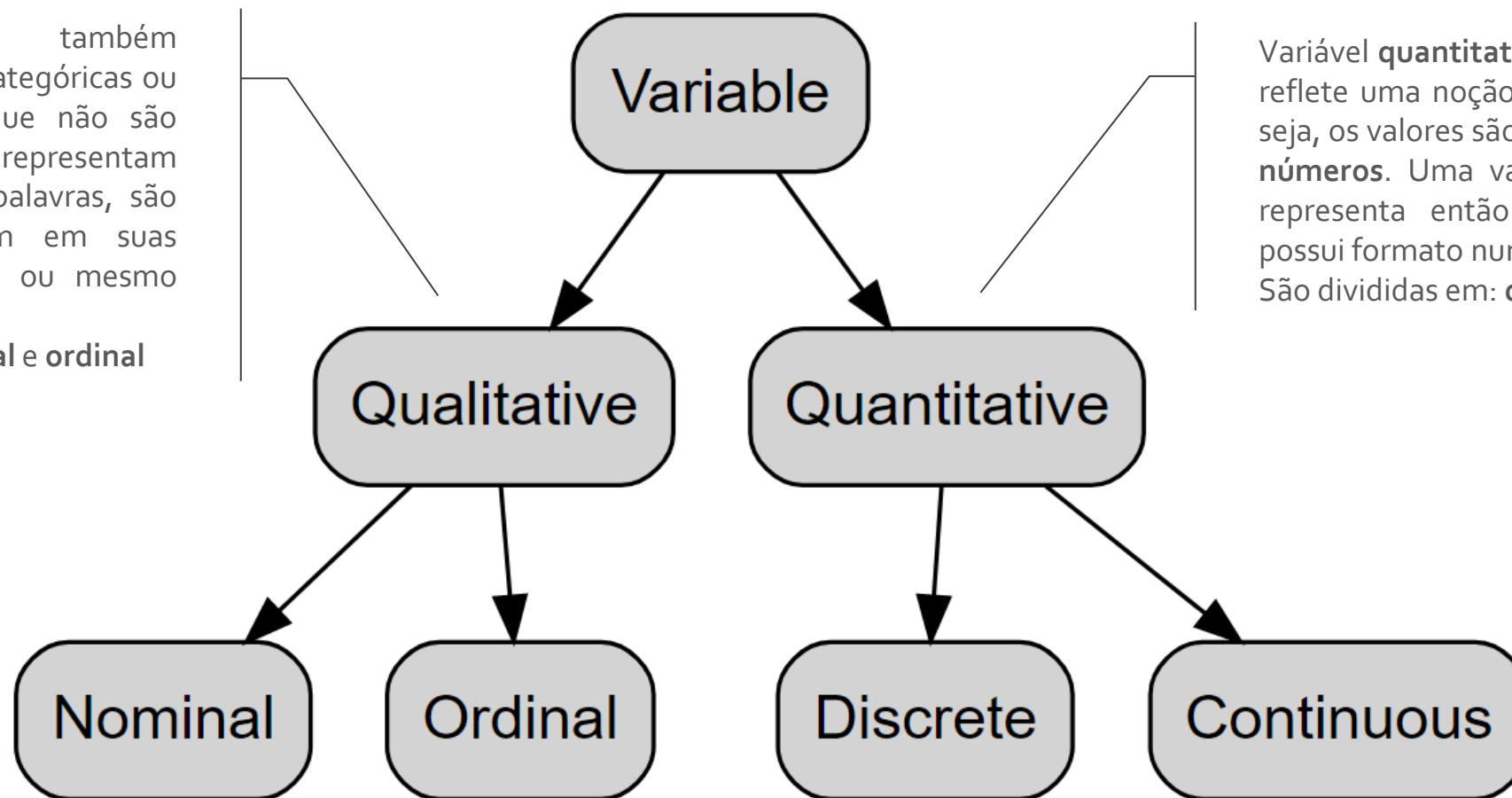
Trabalhos práticos  
Aplicação do conteúdo

# Tipos de variáveis

- Em estatística, variáveis são classificadas em 2 tipos básicos: quantitativas e qualitativas. Estas por suas vez possuem dois subgrupos distintos.

Variável **qualitativa**, também denominadas variáveis categóricas ou *factors*, são variáveis que não são numéricas e que representam **categorias**. Em outras palavras, são variáveis que assumem em suas modalidades, categorias ou mesmo níveis.

São divididas em: **nominal** e **ordinal**



Variável **quantitativa** é a variável que reflete uma noção de magnitude, ou seja, os valores são representados por **números**. Uma variável quantitativa representa então uma medição e possui formato numérico. São divididas em: **discreta** e **contínua**

# Tipos de variáveis

## Conceitos



	Qualitativa	Nominal	Variável qualitativa onde não há ordenação possível ou implícita nos níveis. Pode possuir apenas dois níveis ou múltiplos níveis. <b>Exemplo:</b> sexo, cor dos olhos, sim/não (para vars indicadoras), UF, estado civil, profissão, marca do carro
		Ordinal	Variável qualitativa que possui ordenação implícita nos níveis. <b>Exemplo:</b> grau de escolaridade, segmento do cliente, rating de crédito, severidade/intensidade, estágio da doença, mês de observação
	Quantitativa	Discreta	Variável quantitativa onde os valores assumidos são contagem e possuem um finito número de possibilidade. <b>Exemplo:</b> número de filhos por família, população de um país, número de cigarros fumados, bactérias/litro
		Contínua	Variável quantitativa na qual os valores não são de contagem e possui um infinito número de possibilidades. São resultado de medições. <b>Exemplo:</b> idade, peso, altura, renda, faturamento, salário, nota de avaliação, preço, gastos com cartão

# Tipos de variáveis


## Conceitos



	Qualitativa	Nominal	Variável qualitativa onde não há ordenação possível ou implícita nos níveis. Pode possuir apenas dois níveis ou múltiplos níveis. <b>Exemplo:</b> sexo, cor dos olhos, sim/não (para vars indicadoras), UF, estado civil, profissão, marca do carro
		Ordinal	Variável qualitativa que possui ordenação implícita nos níveis. <b>Exemplo:</b> grau de escolaridade, segmento do cliente, rating de crédito, severidade/intensidade, estágio da doença, mês de observação
	Quantitativa	Discreta	Variável quantitativa onde os valores assumidos são contagem e possuem um finito número de possibilidade. <b>Exemplo:</b> número de filhos por família, população de um país, número de cigarros fumados, bactérias/litro
		Contínua	Variável quantitativa na qual os valores não são de contagem e possui um infinito número de possibilidades. São resultado de medições. <b>Exemplo:</b> idade, peso, altura, renda, faturamento, salário, nota de avaliação, preço, gastos com cartão

# Tipos de variáveis


## Conceitos

	Qualitativa	Nominal	Variável qualitativa onde não há ordenação possível ou implícita nos níveis. Pode possuir apenas dois níveis ou múltiplos níveis. <b>Exemplo:</b> sexo, cor dos olhos, sim/não (para vars indicadoras), UF, estado civil, profissão, marca do carro
		Ordinal	Variável qualitativa que possui ordenação implícita nos níveis. <b>Exemplo:</b> grau de escolaridade, segmento do cliente, rating de crédito, severidade/intensidade, estágio da doença, mês de observação
	Quantitativa	Discreta	Variável quantitativa onde os valores assumidos são contagem e possuem um finito número de possibilidade. <b>Exemplo:</b> número de filhos por família, número de hospitalizações, número de cigarros fumados, bactérias/litro
		Contínua	Variável quantitativa na qual os valores não são de contagem e possui um infinito número de possibilidades. São resultado de medições. <b>Exemplo:</b> idade, peso, altura, renda, faturamento, salário, nota de avaliação, preço, gastos com cartão

# Tipos de variáveis

## Conceitos

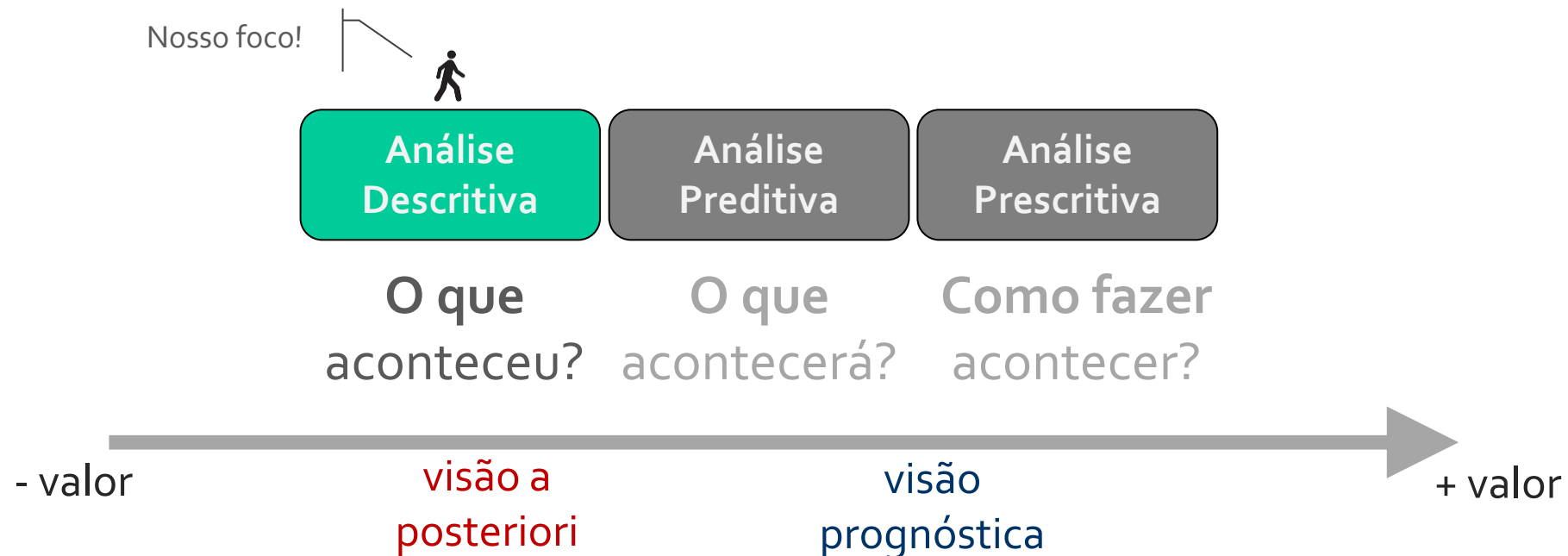


	Qualitativa	Nominal	Variável qualitativa onde não há ordenação possível ou implícita nos níveis. Pode possuir apenas dois níveis ou múltiplos níveis. <b>Exemplo:</b> sexo, cor dos olhos, sim/não (para vars indicadoras), UF, estado civil, profissão, marca do carro
		Ordinal	Variável qualitativa que possui ordenação implícita nos níveis. <b>Exemplo:</b> grau de escolaridade, segmento do cliente, rating de crédito, severidade/intensidade, estágio da doença, mês de observação
	Quantitativa	Discreta	Variável quantitativa onde os valores assumidos são contagem e possuem um finito número de possibilidade. <b>Exemplo:</b> número de filhos por família, população de um país, número de cigarros fumados, bactérias/litro
		Contínua	Variável quantitativa na qual os valores não são de contagem e possui um infinito número de possibilidades. São resultado de medições. <b>Exemplo:</b> idade, peso, altura, renda, faturamento, salário, nota de avaliação, preço, gastos com cartão



# Estatística descritiva

- A estatística descritiva preocupa-se em sumarizar e descrever qualquer conjunto de dados, sendo o primeiro passo para gerar informação relevante.
- É a ferramenta que tenta responder: “o que aconteceu?”, “como aconteceu?”, “em que região aconteceu?”
- Com ela podemos conhecer a distribuição das variáveis, entender o comportamento e o relacionamento entre elas.
- Existem diversas estatísticas que denotam **localidade** (ou centralidade) e **dispersão** dos dados.



- Para essa etapa será utilizada uma base de dados de preços de carros usados na Europa em 2015.
- 111.726 observações e 8 variáveis.
- Adaptado de: <https://www.kaggle.com/mirosval/personal-cars-classifieds>

	maker	car_age	mileage	engine_power	transmission	door_count	seat_count	price_eur
1	nissan	10	149465	121	auto	5	5	4811.25
2	ford	7	99713	74	man	5	5	6476.68
3	toyota	0	5	97	man	5	5	14985.20
4	toyota	0	5	51	man	5	4	8878.57
5	citroen	9	91960	65	man	5	5	3885.64
6	bmw	7	107763	225	auto	5	4	22205.77
7	hyundai	5	29934	57	man	5	5	5773.50
8	citroen	8	52542	92	man	5	5	5551.44
9	citroen	9	99039	65	man	5	5	4071.06
10	fiat	11	63588	76	man	5	5	2960.77
11	hyundai	2	23196	66	man	5	5	9807.55
12	ford	6	73103	74	man	5	5	7179.87
13	ford	6	67232	74	man	5	5	6809.77
14	hyundai	3	29413	99	man	5	5	14211.70
15	mercedes-benz	7	86000	85	man	5	5	6550.70

### Metadados

**maker:** fabricante do veículo  
**car\_age:** idade do veículo (anos)  
**mileage:** distância percorrida (milhas)  
**engine\_power:** potência do motor (kW)  
**transmission:** tipo de transmissão  
**door\_count:** número de portas  
**seat\_count:** número de assentos  
**price\_eur:** preço do veículos (Euros)

- Para essa etapa será utilizada uma base de dados de preços de carros usados na Europa em 2015.
- 111.726 observações e 8 variáveis.
- Adaptado de: <https://www.kaggle.com/mirosval/personal-cars-classifieds>

	maker	car_age	mileage	engine_power	transmission	door_count	seat_count	price_eur
1	nissan	10	149465	121	auto	5	5	4811.25
2	ford	7	99713	74	man	5	5	6476.68
3	toyota	0	5	97	man	5	5	14985.20
4	toyota	0	5	51	man	5	4	8878.57
5	citroen	9	91960	65	man	5	5	3885.64
6	bmw	7	107763	225	auto	5	4	22205.77
7	hyundai	5	29934	57	man	5	5	5773.50
8	citroen	8	52542	92	man	5	5	5551.44
9	citroen	9	99039	65	man	5	5	4071.06
10	fiat	11	63588	76	man	5	5	2960.77
11	hyundai	2	23196	66	man	5	5	9807.55
12	ford	6	73103	74	man	5	5	7179.87
13	ford	6	67232	74	man	5	5	6809.77
14	hyundai	3	29413	99	man	5	5	14211.70
15	mercedes-benz	7	86000	85	man	5	5	6550.70

### Metadados

**maker:** fabricante do veículo

**car\_age:** idade do veículo (anos)

**mileage:** distância percorrida (milhas)

**engine\_power:** potência do motor (kW)

**transmission:** tipo de transmissão

**door\_count:** número de portas

**seat\_count:** número de assentos

**price\_eur:** preço do veículos (Euros)

## Medidas de localização

- Permitem ver “onde” os dados estão localizados e entre quais valores, dando um entendimento de qual é a tendência central e a “posição” como um todo.
- As medidas mais comuns são: **mínimo**, **máximo**, **média**, mediana, quartis e percentis.

## Mínimo Máximo

Mínimo e máximo são simplesmente o menor e maior valor, respectivamente, que uma variável quantitativa assume.

```
> # MINIMO E MAXIMO
> # dplyr
> data %>% summarise(min = min(price_eur), max = max(price_eur))
      min      max
1 2500.11 119953.3
>
> # base R
> min(data$price_eur)
[1] 2500.11
> max(data$price_eur)
[1] 119953.3
>
```

## Exemplo



Aqui iremos calcular as métricas da variável **price\_eur**

## Medidas de localização

- Permitem ver “onde” os dados estão localizados e entre quais valores, dando um entendimento de qual é a tendência central e a “posição” como um todo.
- As medidas mais comuns são: **mínimo, máximo, média**, mediana, quartis e percentis.

### Média

Média é a estatística mais comum. Dá a ideia do valor central do dado, ou em outras palavras seu “centro de gravidade”. Demonstra a concentração dos dados de uma distribuição.

```
> # MEDIA
> # dplyr
> data %>% summarise(avg = mean(price_eur))
      avg
1 14234.2
>
> # base R
> mean(data$price_eur)
[1] 14234.2
```



É dada pela soma de todos os valores dividida pelo número de observações:

$$\bar{x} = \frac{\sum x_i}{n}$$

### Exemplo

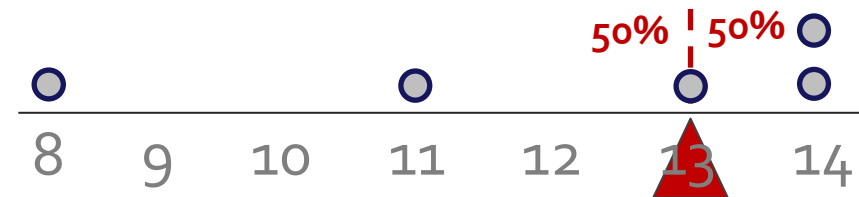
## Medidas de localização

- Permitem ver “onde” os dados estão localizados e entre quais valores, dando um entendimento de qual é a tendência central e a “posição” como um todo.
- As medidas mais comuns são: mínimo, máximo, média, **mediana**, **quartis** e **percentis**.

## Mediana

Mediana é uma medida de localização e também dá uma ideia de tendência central dos dados. Porém está implícito que mesmo número de observação acima e abaixo desse valor. Dessa forma, 50% dos dados estão acima e 50% estão abaixo.

```
> # MEDIANA
> # dplyr
> data %>% summarise(avg = median(price_eur))
      avg
1 11166.62
>
> # base R
> median(data$price_eur)
[1] 11166.62
>
```



$$\begin{array}{ll} n \text{ ímpar} & n \text{ par} \\ med = x_{\frac{n+1}{2}} & med = \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \end{array}$$

Para calcular basta ordenar os dados do menor para o maior e tomar o ponto do meio como a mediana

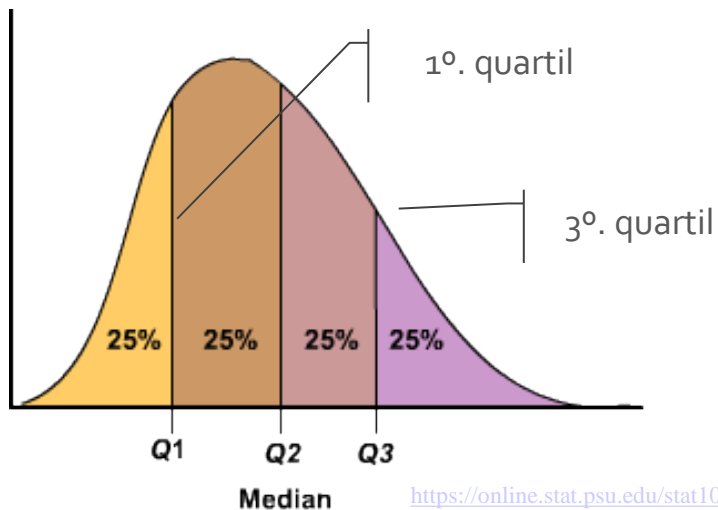
## Exemplo

## Medidas de localização

- Permitem ver “onde” os dados estão localizados e entre quais valores, dando um entendimento de qual é a tendência central e a “posição” como um todo.
- As medidas mais comuns são: mínimo, máximo, média, **mediana**, **quartis** e **percentis**.

## Quartis

Os quartis são similares à mediana no sentido de dividirem os dados em duas partes. Porém aqui, eles não dividem os dados em partes iguais. O 1º. quartil divide os dados de tal forma que 25% estão abaixo e 75% estão acima. Já o 3º. quartil representa o valor no qual 75% dos dados estão abaixo e 25% acima



```
> # 1o e 3o quartil
> # dplyr
> data %>% summarise(Q1 = quantile(price_eur,0.25),
+                    Q3 = quantile(price_eur,0.75))
      Q1      Q3
1 6500.74 17952.66
>
> # base R
> quantile(data$price_eur,probs = c(0.25,0.75))
      25%      75%
6500.74 17952.66
>
```

<https://online.stat.psu.edu/stat100/book/export/html/639>

## Exemplo



## Centralidade vs dispersão

- A média não revela alguns fatos interessantes sobre a variabilidade dos dados.
- Observe o valor da temperatura máxima em duas localidades diferentes do globo (qual variou mais?).

Dia (Mês de Abril)	Temp. máx. (°C) - Manaus	Temp. máx. (°C) - NYC
3	31	19
4	32	8
5	33	12
6	32	16
7	31	11
8	32	18
9	30	15
10	30	14
11	32	24
12	32	23
<b>Média</b>	<b>31,5</b>	<b>16,0</b>

Em Manaus, por exemplo, a temperatura máxima registrada foi **33°C** enquanto que a mínima **30°C**

Já em NY no mesmo período de tempo, foi observado o valor mínimo de **8°C** e máximo de **24°C**

Em Manaus a temperatura média do dia 3 – 12 foi de **31,5°C**

Em NY a temperatura média para os 9 dias foi de **16,0°C**

## Medidas de dispersão

- Medidas de dispersão permitem que se tenha uma sensibilidade da variabilidade dos dados (no sentido de se a distribuição é mais “alongada” ou “achatada”).
- As medidas mais comuns são: **amplitude**, **desvio padrão**, variância, e intervalo interquartil.

Amplitude é o tamanho do intervalo que contém todo o dado e provê uma ideia de dispersão. É útil para pequeno conjunto de dados.

O cálculo é feito pela diferença entre o maior e menor valor.

$$\textit{amplitude} = \textit{máx} - \textit{mín}$$

## Amplitude

```
> # range
> # dplyr
> data %>% summarise(amp = max(price_eur) - min(price_eur))
      amp
1 117453.1
>
> # base R
> range(data$price_eur)
[1] 2500.11 119953.26
> max(data$price_eur) - min(data$price_eur)
[1] 117453.1
>
```

É uma métrica de fácil determinação, porém ela baseia-se apenas em dois valores, que são os valores extremos do dado.

## Exemplo

## Medidas de dispersão

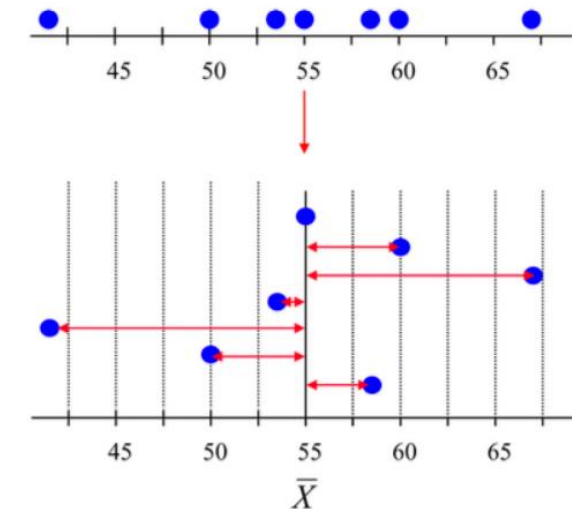
- Medidas de dispersão permitem que se tenha uma sensibilidade da variabilidade dos dados (no sentido de se a distribuição é mais “alongada” ou “achatada”).
- As medidas mais comuns são: **amplitude**, **desvio padrão**, variância, e intervalo interquartil.

## Desvio Padrão

O desvio padrão é a métrica mais comum para descrever a dispersão de um dado. Ele representa o desvio médio dos dados em relação à média. Maiores valores indicam dados mais espalhados em relação a média, enquanto valores menores denotam dados menos concentrados.

```
> # desvio padrao
> # dplyr
> data %>% summarise(sd = sd(price_eur))
      sd
1 11965.79
>
> # base R
> sd(data$price_eur)
[1] 11965.79
>
```

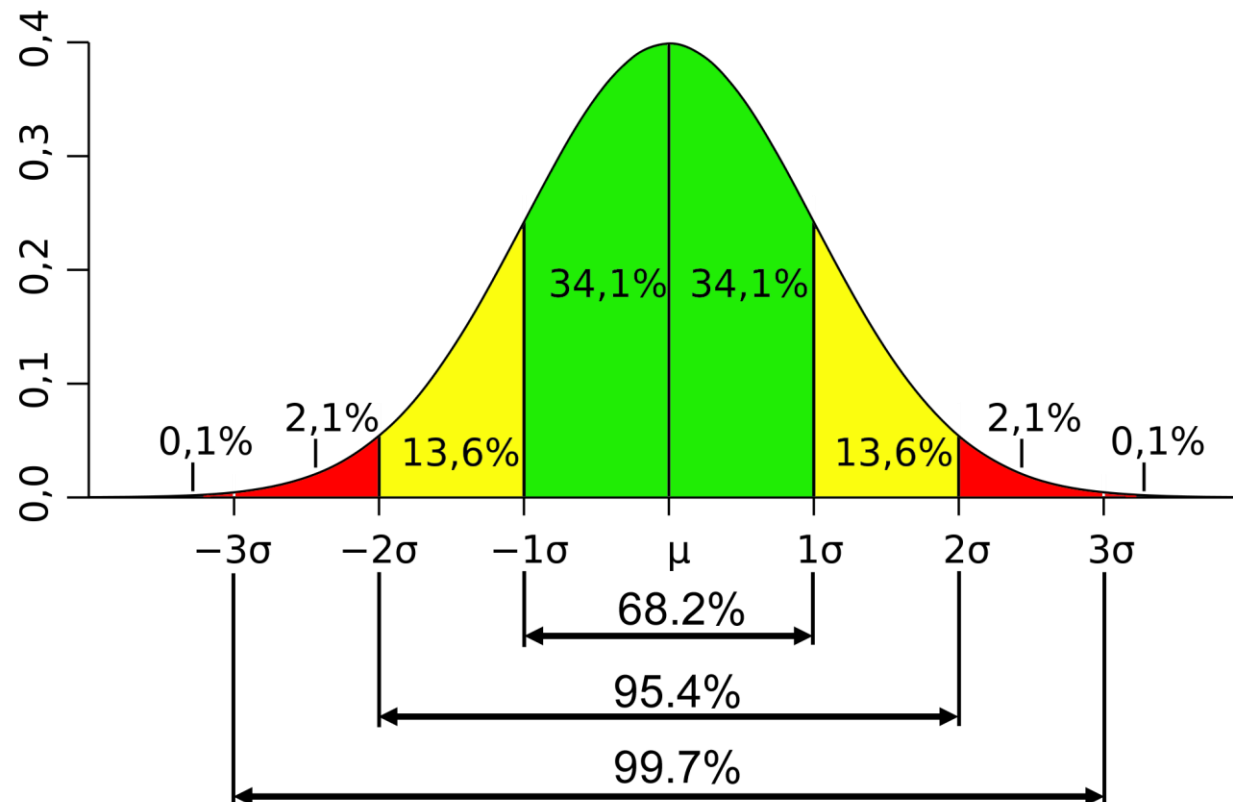
$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$



## Exemplo

## Propriedades do desvio padrão [extra]

- Apesar de ser difícil de se interpretar, o desvio padrão tem propriedades que são muito utilizadas na estatística.
- Se a distribuição da variável for simétrica, podemos usar a regra empírica, que permite saber qual é a probabilidade de um valor estar muito distante da média (em desvios padrão).



<https://br.pinterest.com/pin/378443174930440633/>

### Exemplo

Se a média das vendas é 15 e o desvio padrão é 3, sabemos que, com aproximadamente **95% de probabilidade** as vendas do mês que vem estarão entre **9** e **21** unidades.

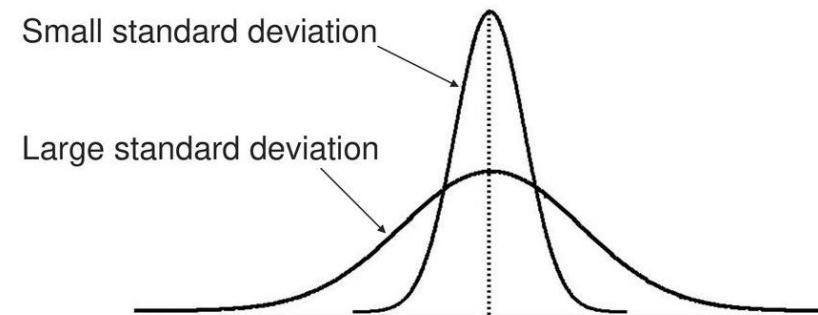
## Medidas de dispersão

- Medidas de dispersão permitem que se tenha uma sensibilidade da variabilidade dos dados (no sentido de se a distribuição é mais “alongada” ou “achatada”).
- As medidas mais comuns são: amplitude, desvio padrão, **variância**, e **intervalo interquartil**.

Variância é outra medida de dispersão dos dados muito utilizada na análise de volatilidade. Possui a mesma ideia que o desvio padrão.

Ela é representada pelo quadrado do desvio padrão (note que ela possui a unidade da variável porém ao quadrado).

```
> # variance
> # dplyr
> data %>% summarise(var = var(price_eur))
      var
1 143180134
>
> # base R
> var(data$price_eur)
[1] 143180134
>
```



$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

## Variância

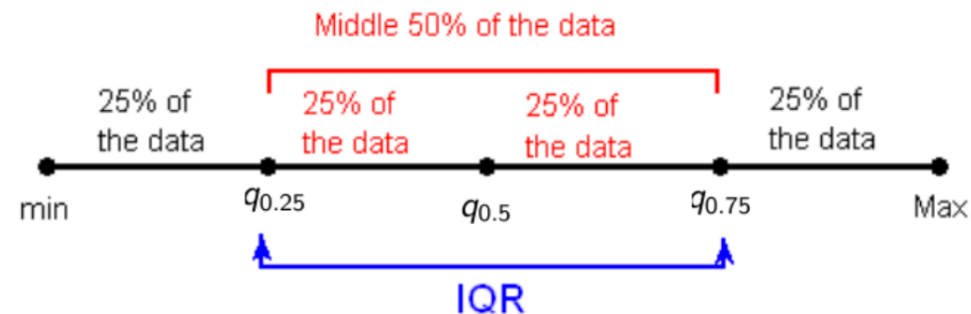
## Exemplo

## Medidas de dispersão

- Medidas de dispersão permitem que se tenha uma sensibilidade da variabilidade dos dados (no sentido de se a distribuição é mais “alongada” ou “achatada”).
- As medidas mais comuns são: amplitude, desvio padrão, **variância**, e **intervalo interquartil**.

Trata-se de mais uma medida de dispersão e é relacionada com o 1º. e 3º. quartil. Ele pode ser interpretado como uma amplitude (i.e. a diferença entre dois valores extremos), porém do meio da distribuição dos dados.

```
> # iqr
> # dplyr
> data %>% summarise(iqr = IQR(price_eur))
      iqr
1 11451.92
>
> # base R
> IQR(data$price_eur)
[1] 11451.92
>
```



$$IQR = Q_{0.75} - Q_{0.25}$$

## Intervalo Interquartil

## Exemplo

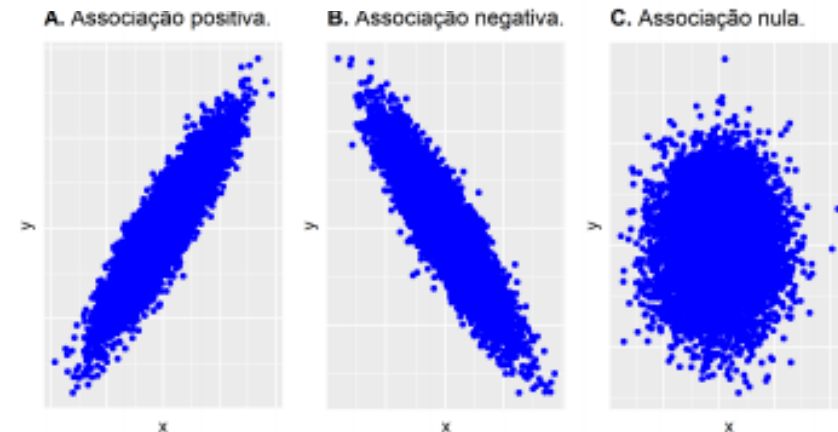
## Medidas de associação

- Medidas de associação mostram como cada variável está relacionada a outra variável.
- Ajuda a responder a pergunta: o conhecimento de uma variável X ajuda a entender uma variável Y?
- Existem diversas medidas, por isso iremos focar em: correlação de **Pearson** e V de Crámer.

O coeficiente de correlação de Pearson mede o grau de associação entre duas variáveis quantitativas. Além disso, ele é usado para mensurar a força de relação linear e a direção entre as variáveis. Assume valores entre -1 e +1, onde valores próximos de zero indicam que não há associação entre as variáveis.

```
> cor(data$mileage, data$price_eur, method = 'pearson')  
[1] -0.4159104
```

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



[https://metodologia.ceie-br.org/wp-content/uploads/2019/02/livro2\\_cap9.pdf](https://metodologia.ceie-br.org/wp-content/uploads/2019/02/livro2_cap9.pdf)

## Exemplo

## Correlação de Pearson

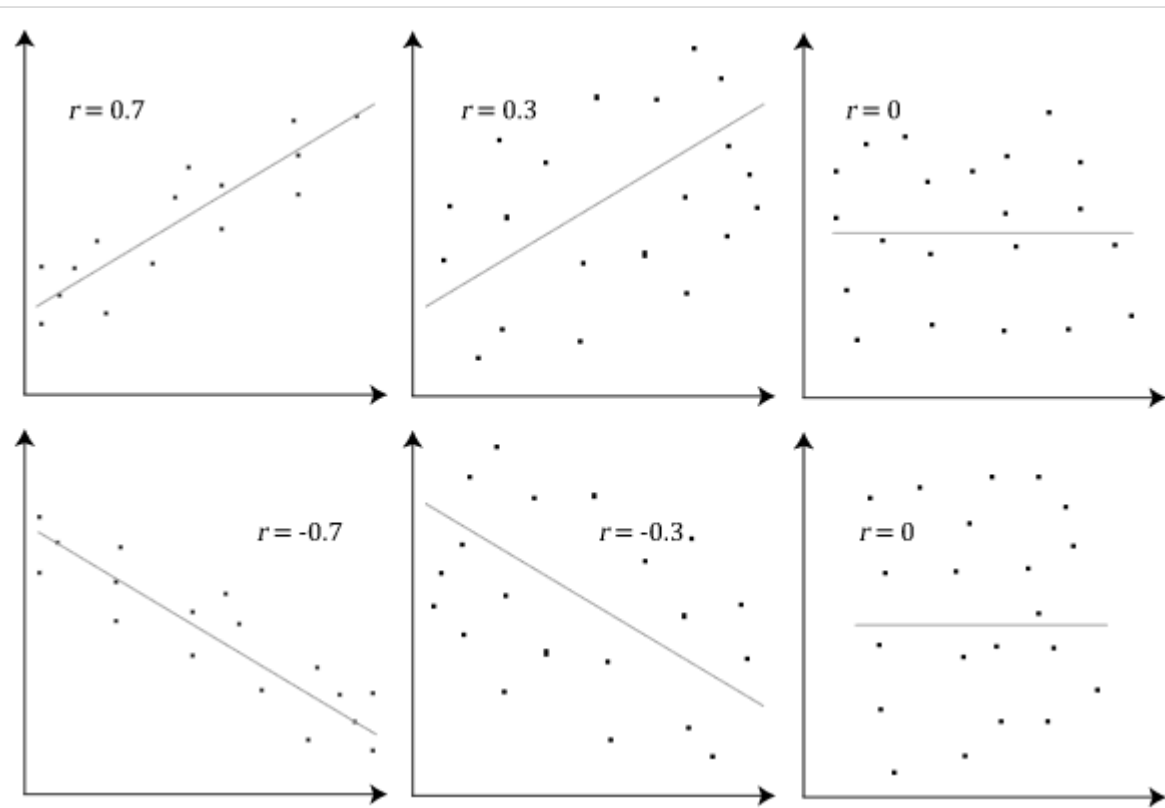
## Medidas de associação

- Medidas de associação mostram como cada variável está relacionada a outra variável.
- Ajuda a responder a pergunta: o conhecimento de uma variável X ajuda a entender uma variável Y?
- Existem diversas medidas, por isso iremos focar em: correlação de **Pearson** e V de Crámer.

## Correlação de Pearson

Se  $r > 0$  temos associação positiva  
Se  $r < 0$  temos associação negativa  
Se  $r \sim 0$  temos uma associação nula

Baixa associação:  $|0.1 - 0.3|$   
Média associação:  $|0.3 - 0.5|$   
Alta associação:  $|0.5 - 1.0|$



## Exemplo



## Medidas de associação

- Medidas de associação mostram como cada variável está relacionada a outra variável.
- Ajuda a responder a pergunta: o conhecimento de uma variável X ajuda a entender uma variável Y?
- Existem diversas medidas, por isso iremos focar em: correlação de **Pearson** e V de Crámer.

O coeficiente de correlação de Pearson mede o grau de associação entre duas variáveis quantitativas. Além disso, ele é usado para mensurar a força de relação linear e a direção entre as variáveis. Assume valores entre -1 e +1, onde valores próximos de zero indicam que não há associação entre as variáveis.

## Correlação de Pearson

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

<https://towardsdatascience.com/everything-you-need-to-know-about-interpreting-correlations-2c485841c0b8>

## Exemplo

## Medidas de associação

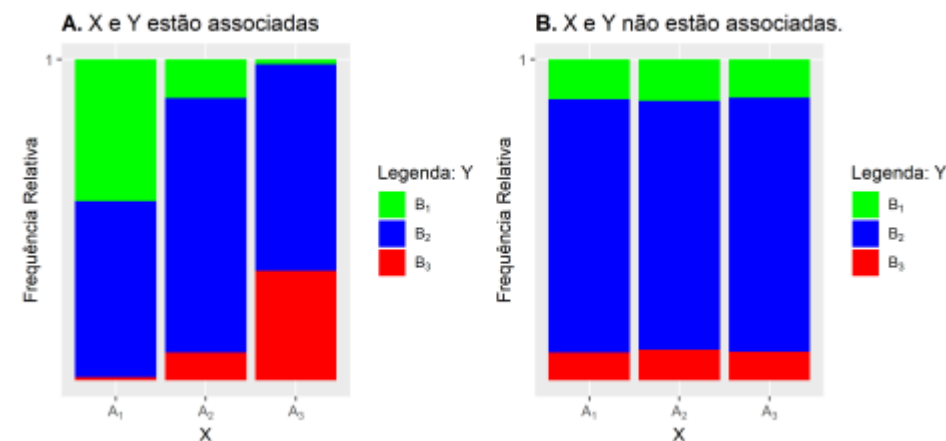
- Medidas de associação mostram como cada variável está relacionada a outra variável.
- Ajuda a responder a pergunta: o conhecimento de uma variável X ajuda a entender uma variável Y?
- Existem diversas medidas, por isso iremos focar em: correlação de **Pearson** e V de Crámer.

## V de Crámer

O coeficiente V de Crámer é uma medida de associação entre duas variáveis nominais e é relacionado com o teste T de Tschuprow para tabelas de contingência quadradas. Quanto mais próximo de 1 maior a associação entre as duas variáveis qualitativas, oposto para valores próximos de 0.

```
> library(rcompanion)
> cramerv(data$transmission, data$maker)
Cramer V
0.4245
```

$$\hat{V} = \sqrt{\frac{\chi^2 / n}{\min(k - 1, r - 1)}}$$



[https://metodologia.ceie-br.org/wp-content/uploads/2019/02/livro2\\_cap9.pdf](https://metodologia.ceie-br.org/wp-content/uploads/2019/02/livro2_cap9.pdf)

## Exemplo

## Medidas de associação

- Medidas de associação mostram como cada variável está relacionada a outra variável.
- Ajuda a responder a pergunta: o conhecimento de uma variável X ajuda a entender uma variável Y?
- Existem diversas medidas, por isso iremos focar em: correlação de **Pearson** e V de Crámer.

O coeficiente V de Crámer é uma medida de associação entre duas variáveis nominais e é relacionado com o teste T de Tschuprow para tabelas de contingência quadradas. Quanto mais próximo de 1 maior a associação entre as duas variáveis qualitativas, oposto para valores próximos de 0.

## V de Crámer

Phi and Cramer's V	Interpretation
> 0.25	Very strong
> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
> 0	No or very weak

## Exemplo

...foco de hoje

- **CASE 3: Série de vendas de supermercados da rede Walmart nos EUA (pt. 2)**

Explorando as variáveis da base de dados do case anterior, calculando as métricas descritivas de localidade, dispersão e associação

