Preditiva.ai

Estatística Descritiva

Análise Bidimensional

Aula 4

17/12

Plano de aula

Preditiva.ai

	Carga horária	Dias
Módulo 1 - Estatística Descritiva	08h	08 a 15/12/21
Módulo 2 - Probabilidades	10h	12/01 a 26/01/22
Módulo 3 - Inferência Estatística	20h	31/01 a 14/03/22
Módulo 4 - BI e Metodologias de Projetos de Dados	04h	16/03 a 21/03/22
Módulo 5 - Bancos de Dados Relacionais - SQL	12h	23/03 a 04/04/22
Módulo 6 - Visualização de Dados (Data Viz) e Storytelling	10h	06/04 a 13/04/22
Módulo 7 - Apresentação de Projeto final (Capstone)	06h	18/04 a 20/04/22
Carga horária (h)	**70h**	

Estatística Descritiva
Análise Bidimensional

Preditiva.ai

Já sabemos como resumir e analisar cada variável de um conjunto de dados, mas:

"E se tivermos que analisar o comportamento de **2 variáveis simultaneamente**?

Por exemplo:

- Qual a taxa de turnover por nível de formação?
- Qual o percentual de compra do novo livro do Flávio Augusto por região do Brasil?

Estatística Descritiva
Análise Bidimensional

Preditiva.ai

A **Análise Bidimensional** é o nome dado a um conjunto de técnicas utilizadas para:

Analisar o **comportamento conjunto** de **duas variáveis**

Considerando os **diferentes tipos de variáveis**, podemos ter 3 situações:

1. Duas variáveis **quantitativas**
2. Uma variável **qualitativa** e outra variável **quantitativa**
3. Duas variáveis **qualitativas**

Análise Bidimensional

2 Variáveis Quantitativas

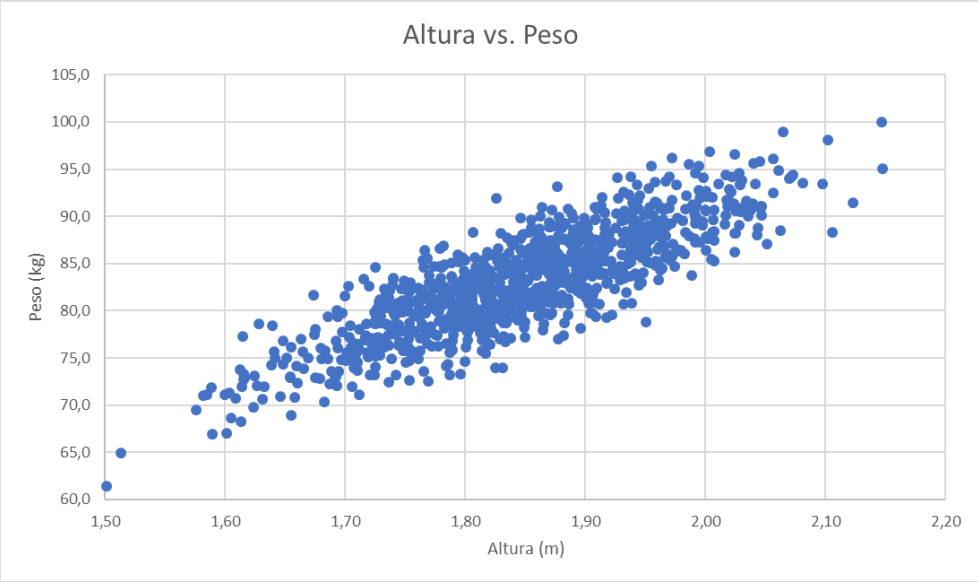
Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

Preditiva.ai

A análise de 2 variáveis quantitativas inicia-se com o **Gráfico de Dispersão**. Nele uma das variáveis fica no eixo X e a outra no eixo Y.

Exemplo: Em uma pesquisa com os suecos, foram obtidos os dados de peso e altura de cada habitante. O gráfico de dispersão com essas variáveis é apresentado abaixo.



Podemos perceber que existe uma **relação entre altura e peso**, ou seja, quanto **maior a altura**, **maior o peso**.

A essa relação damos o nome de **correlação**, logo **Altura e Peso estão correlacionados.**

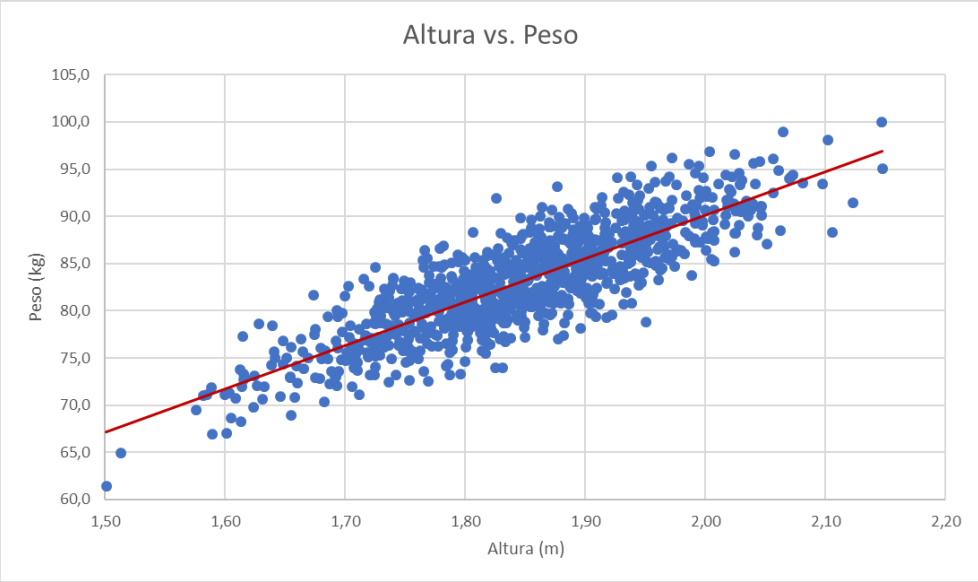
Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

Preditiva.ai

O **Gráfico de Dispersão** fornece a **distribuição conjunta** da Altura e Peso, e junto com ela a visualização de uma possível **correlação entre essas duas variáveis**.

Nesse caso, a **correlação entre Altura e Peso aparenta ser linear**, ou seja, é possível **definir uma equação linear** em que dada a Altura encontramos o Peso, e vice-versa.



Entender a **correlação entre variáveis** é algo bastante **poderoso**! Mas é preciso ter cautela.

Vamos entender isso logo mais...

Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

Preditiva.ai

Uma forma de **medirmos a força da correlação** entre duas variáveis quantitativas, como Altura e Peso, é calculando o **Coeficiente de Correlação de Pearson**:

$$corr(X,Y) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{dp(X)}\right) \cdot \left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

O **Coeficiente de Correlação de Pearson** varia entre -1 e +1 e indica:

- **Correlação positiva forte**: coeficiente próximo a 1
- **Correlação inexistente**: coeficiente próximo a zero
- **Correlação negativa forte**: coeficiente próximo a -1

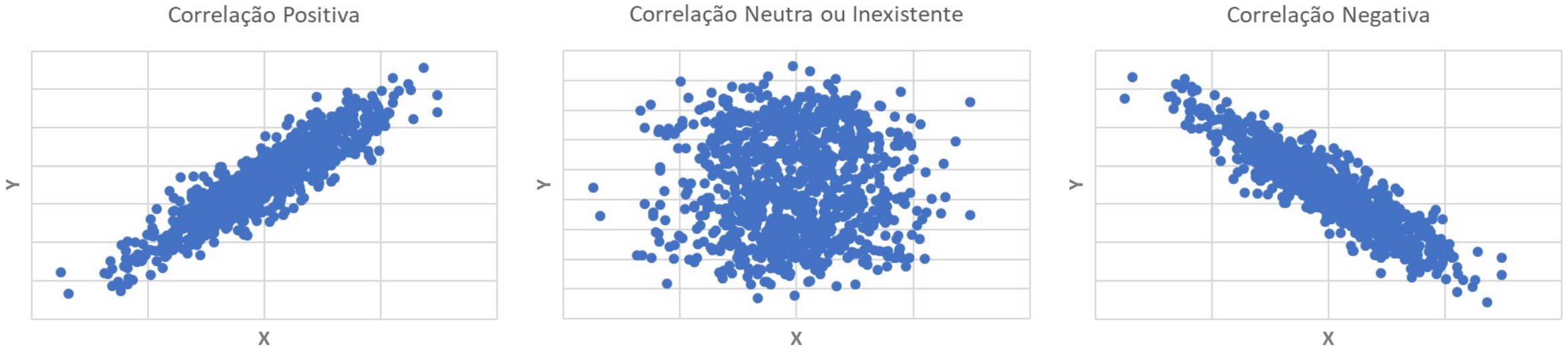
Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

Preditiva.ai

O **Coeficiente de Correlação de Pearson** varia entre -1 e +1 e indica:

- **Correlação positiva**: quando a variável X **aumenta**, a variável Y também **aumenta**
- **Correlação inexistente**: quando a variável X **aumenta**, a variável Y **não se altera**
- **Correlação negativa**: quando a variável X **aumenta**, a variável Y **diminui**



Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

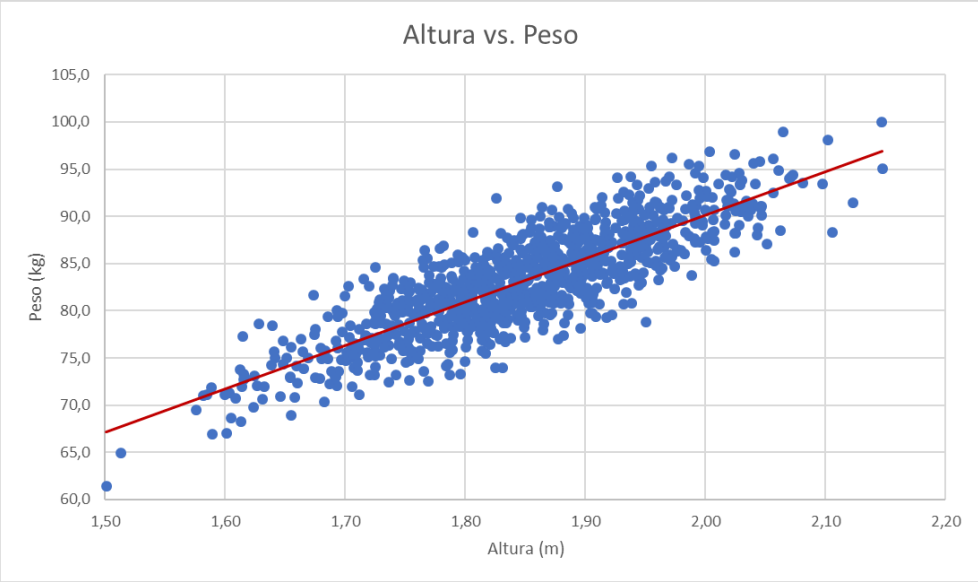
Preditiva.ai

No Excel: CORREL(X ; Y)

X: observações da variável 1

Y: observações da variável 2

Intervalo	Força
- 1,0 < r < - 0,7	Fortemente Negativa
- 0,6 < r < 0,6	Fraca
0,7 < r < 1,0	Fortemente Positiva



Neste exemplo, vamos considerar **X = Altura e Y = Peso**.

Utilizando **CORREL(Altura;Peso)** = **0,85**

Logo, **existe uma forte correlação positiva** entre Altura e Peso.

Demonstração

Análise Bidimensional

Correlação vs. Causalidade

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

É fundamental dominarmos a **diferença entre esses 2 conceitos** para não cairmos em algumas **armadilhas** de Analytics.

Vejamos a definição destes 2 termos:

- **Correlação**: relação de **dependência** ou **associação** entre duas variáveis.
- **Causalidade**: relação entre um evento A e um evento B, sendo que o evento B é **consequência** do evento A.

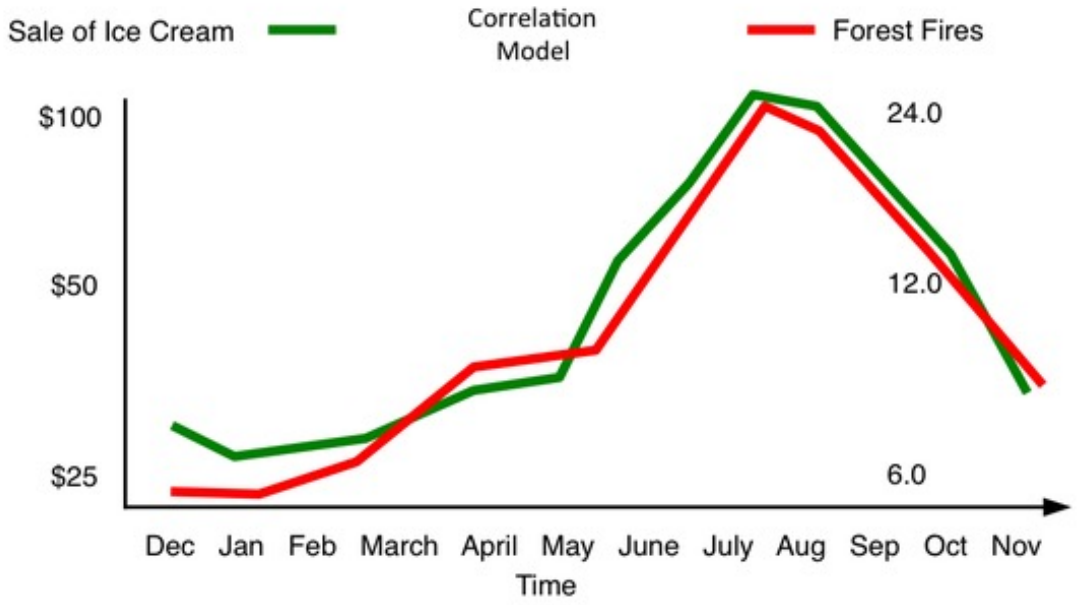
Ou seja, **Correlação** está relacionada com a **dependência ou associação** e a **Causalidade** relacionada a **consequência**.

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

Você acha que a **venda de sorvetes** pode **causar incêndios nas florestas**?

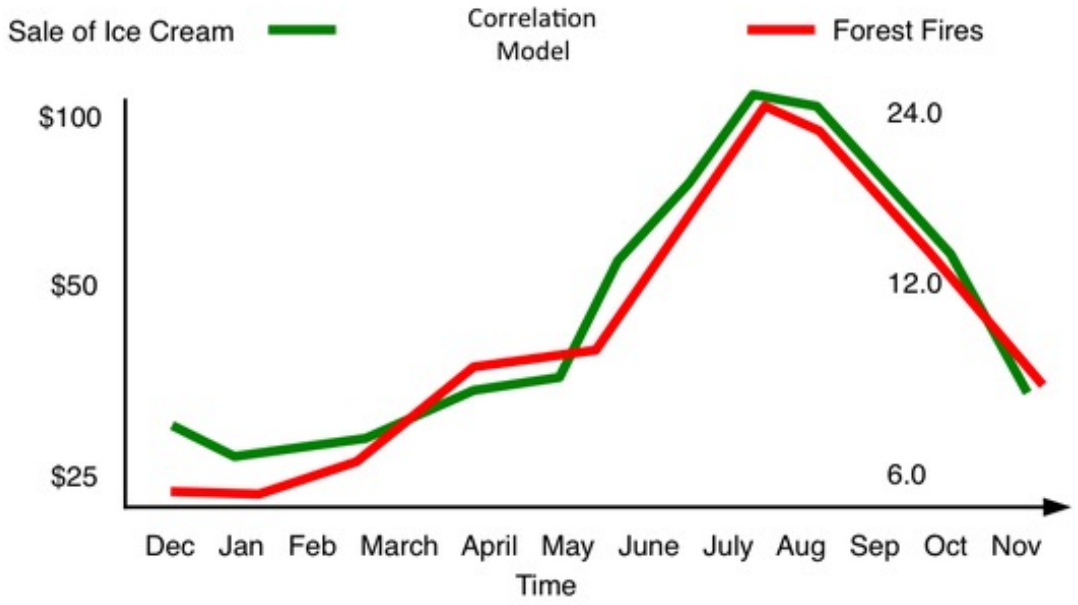
Neste caso há uma **3ª variável não avaliada** e que faz mais sentido ser a **causadora** do aumento no **consumo de sorvete** e dos **incêndios nas florestas**: **o clima quente!**

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

A **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas** é conhecida como **Correlação Espúria**.

As **Correlações Espúrias** podem ser uma armadilha para **falsas conclusões**.

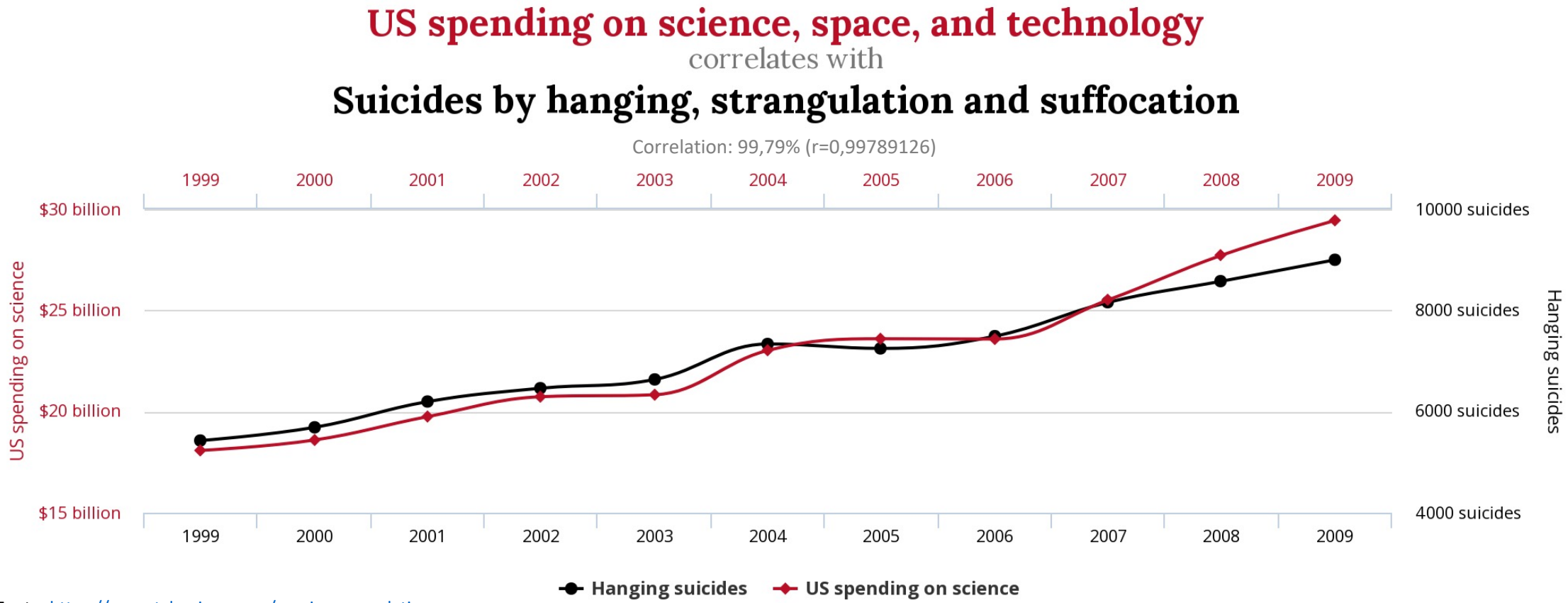
Vejamos alguns outros exemplos.

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Exemplo 1: Gasto em Pesquisa no EUA vs. Suicídios



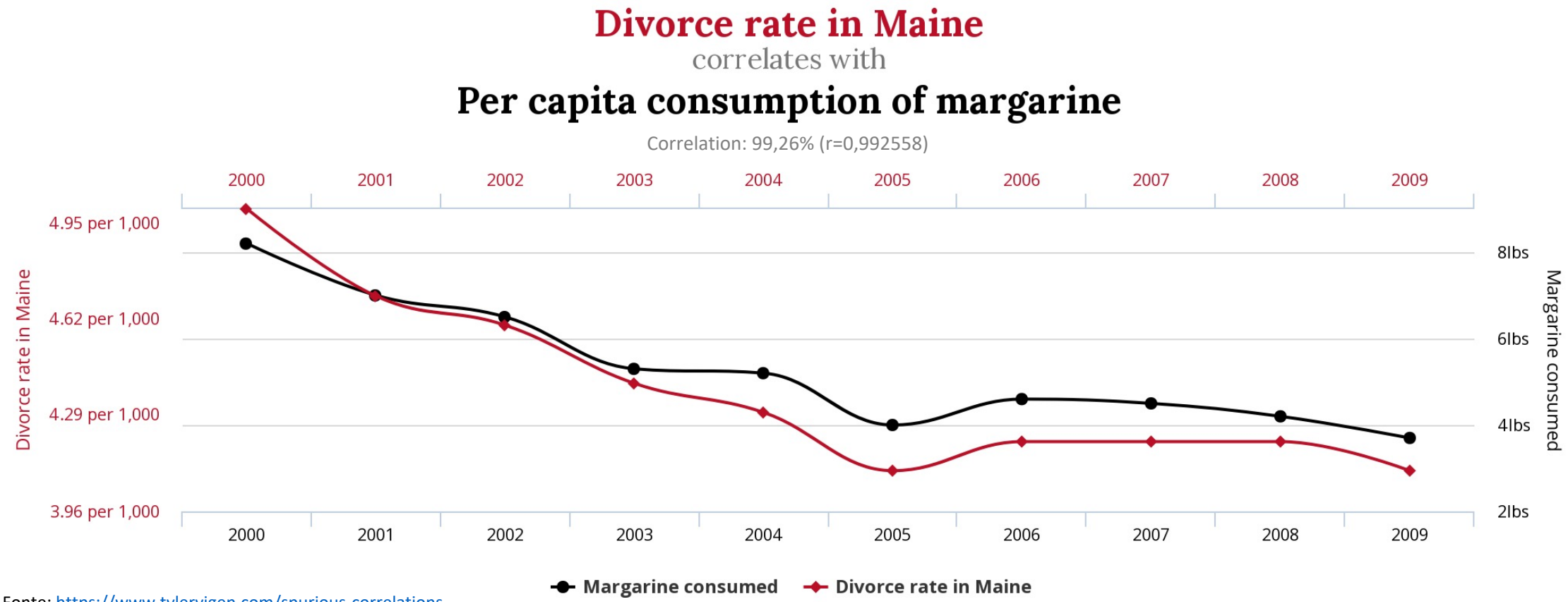
Fonte: https://www.tylervigen.com/spurious-correlations

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Exemplo 2: Divórcios em Maine vs. Consumo de margarina



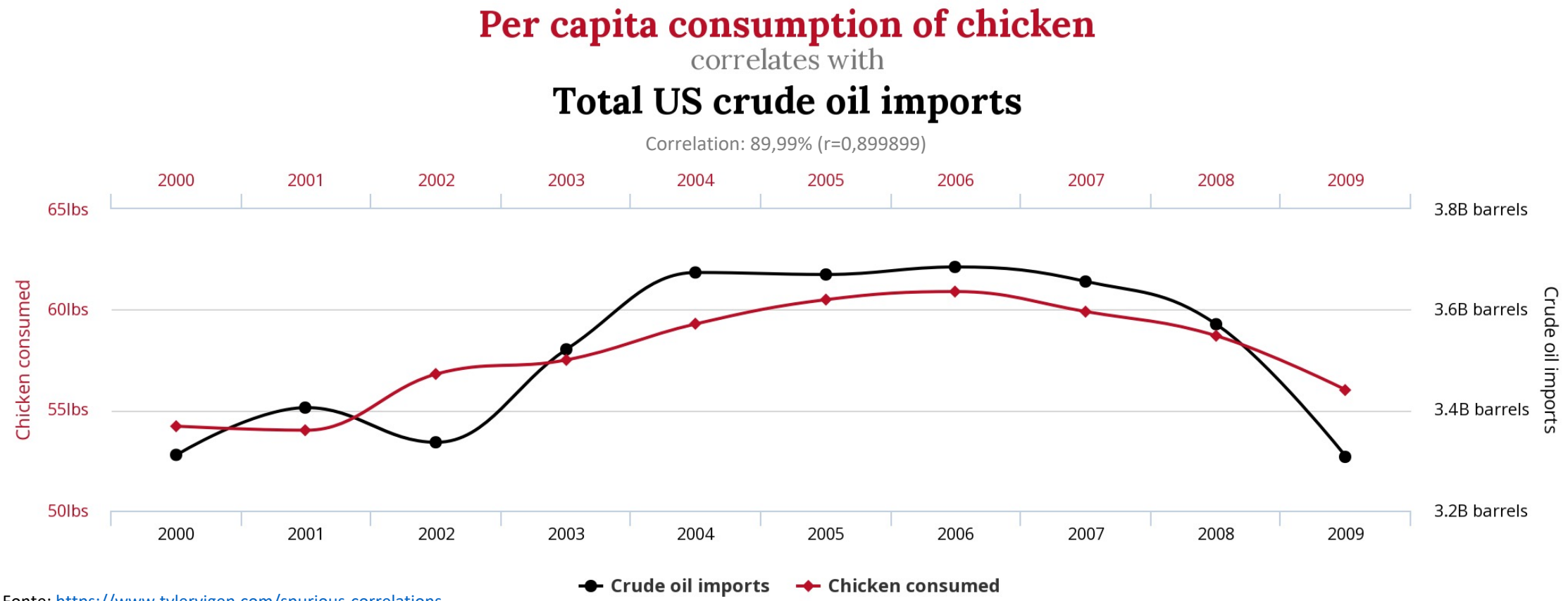
Fonte: https://www.tylervigen.com/spurious-correlations

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Exemplo 3: Consumo de frango vs. Importação de petróleo



Fonte: https://www.tylervigen.com/spurious-correlations

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Conclusões:

- Se 2 variáveis estão **correlacionadas**, **pode ou não** haver **causalidade**
- Se houver **correlação** e não houver **causalidade** entre essas 2 variáveis, possivelmente há uma **3ª variável que não foi observada**
- **Mantenha-se cético**: busque **fortes evidências** para assumir a **causalidade**
- Antes de assumir a **causalidade** responda as seguintes perguntas:
  - **Por que** a variável A **causa** a variável B?
  - **Como** a variável A **causa** a variável B?

Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

Preditiva.ai

Portanto, a existência de **correlação entre duas variáveis** indica que elas estão de alguma forma **associadas**, mas nem sempre isso quer dizer que **uma variável "causa" a outra**.

Em nosso exemplo:

- As pessoas são mais **pesadas** porque são mais **altas**?
- As pessoas são mais **altas** porque são mais **pesadas**?
- As pessoas são mais **altas** e mais **pesadas** devido a outro fator não observado? Genético, por exemplo.

Essa é a grande diferença entre **correlação** e **causalidade**!

Ou seja, **nem toda correlação é causalidade** mas **toda causalidade gera uma correlação**.

Análise Bidimensional

1 Variável Qualitativa e

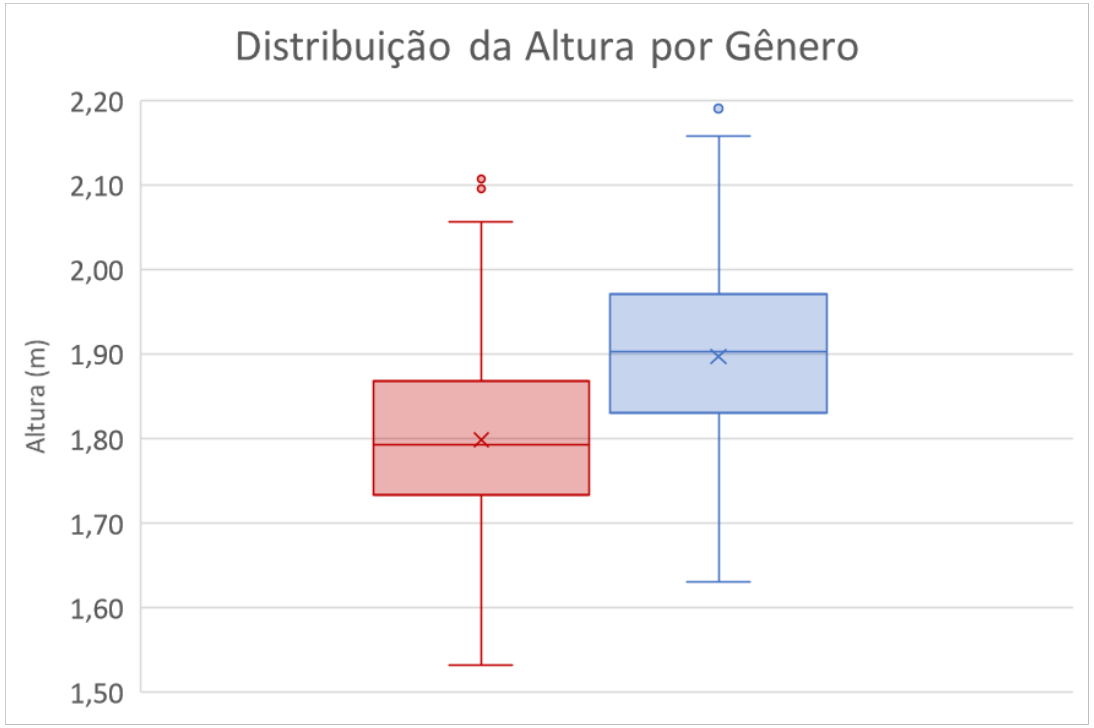
1 Variável Quantitativa

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa

Preditiva.ai

Voltemos a nossa pesquisa com os suecos, mas agora vamos analisar a variável Altura por sexo. No gráfico, o ***boxplot* vermelho** corresponde a distribuição da altura das mulheres e o ***boxplot* azul** corresponde a distribuição da altura dos homens.



Insights

1. As **medidas de posição** da altura dos homens são ligeiramente superiores as das mulheres.
2. As duas distribuições de altura possuem **dispersão semelhante**.
3. **Poucas observações discrepantes** (*outliers*).

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa

Preditiva.ai

Uma outra forma de analisarmos os dados é criar uma tabela com as **medidas resumo** separadas por sexo.

sexo	N	Média de altura	Variância de altura	Desvio Padrão de altura	Mínimo de altura	Máximo de altura	1º Quartil de altura	Mediana de altura	3º Quartil de altura
Feminino	600	1,60	0,010	0,10	1,33	1,91	1,53	1,59	1,67
Masculino	400	1,85	0,010	0,10	1,45	2,14	1,78	1,85	1,92
Total	**1.000**	**1,70**	**0,025**	**0,16**	**1,33**	**2,14**	**1,57**	**1,68**	**1,82**

Utilizando essa tabela podemos:

1. Identificar que a quantidade de mulheres **é maior** do que a quantidade de homens.
2. Confirmamos as informações extraídas anteriormente utilizando os ***boxplots***.

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa

Preditiva.ai

Para medir a associação entre variáveis qualitativas e quantitativas utilizamos o **Coeficiente de Determinação**, também conhecido como **R^2**. Neste exemplo, calcularíamos:

$$R^2 = 1 - \frac{\overline{var(Altura)}}{var(Altura)}$$

Sendo, $$\overline{var(Altura)} = \frac{\sum_{i=1}^{k} n_i \cdot var_i(Altura)}{\sum_{i=1}^{k} n_i}$$

(Média das variâncias para cada sexo)

Intuitivamente, o **R^2** mede quanto da **variância total** é explicada pela **introdução da variável qualitativa** e é uma medida que **varia entre 0 e 1**.

Dessa forma:

- **R^2** igual a zero: indica a **inexistência** de associação entre as variáveis
- **R^2** igual a 1: indica **forte associação** entre as variáveis

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa

Preditiva.ai

Calculando o R^2 para este exemplo:

$$\overline{var(Altura)} = \frac{600 \cdot 0{,}010 + 400 \cdot 0{,}010}{600 + 400} = 0{,}010 \qquad R^2 = 1 - \frac{0{,}010}{0{,}025} = 0{,}6 = \mathbf{60\%}$$

sexo	N	Média de altura	Variância de altura	Desvio Padrão de altura	Mínimo de altura	Máximo de altura	1º Quartil de altura	Mediana de altura	3º Quartil de altura
Feminino	600	1,60	0,010	0,10	1,33	1,91	1,53	1,59	1,67
Masculino	400	1,85	0,010	0,10	1,45	2,14	1,78	1,85	1,92
Total	**1.000**	**1,70**	**0,025**	**0,16**	**1,33**	**2,14**	**1,57**	**1,68**	**1,82**

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa

Preditiva.ai

Calculando o R^2 para este exemplo:

$$\overline{var(Altura)} = \frac{600 \cdot 0{,}010 + 400 \cdot 0{,}010}{600 + 400} = 0{,}010 \qquad R^2 = 1 - \frac{0{,}010}{0{,}025} = 0{,}6 = \mathbf{60}\%$$

Nesse exemplo, vemos que a associação entre Altura e sexo existe, e o R^2 igual a 60% indica que essa associação **é forte**.

Ou seja, **o sexo explica 60% da diferença** de altura entre homens e mulheres.

Análise Bidimensional

2 Variáveis Qualitativas,

1 Binária

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Outro tipo de medida de associação **muito útil é o Information Value (IV).** Essa medida é responsável por fornecer o "**poder de separação**" que uma variável qualitativa de duas ou mais categorias possui sobre outra variável de duas categorias (variável binária).

Exemplos de variáveis binárias:

1. Bons clientes x Maus clientes
2. Comprou x Não comprou
3. Doente x Não doente
4. Verdadeiro x Falso
5. Entre outros...

Em vários problemas de Analytics, estamos interessados em descobrir quais fatores, isto é, **quais variáveis são responsáveis por separar as categorias das variáveis binárias**.

Ex: Qual variável separa um Bom cliente de um Mau cliente para um empréstimo? Profissão? Comprometimento de Renda?

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Comprometimento de Renda** é útil para separar os clientes que **Pagaram e Atrasaram** em um financiamento de veículos:

Comprometimento	Classificação				Frequência	% Freq
	Qte Pagou	Qte Atrasou	% Pagou	% Atrasou		
Maior que 40%	4	127	0,8%	6,2%	131	5%
Entre 30 e 40%	92	710	17,4%	34,4%	802	31%
Entre 20 e 30%	102	432	19,2%	20,9%	534	21%
Menor que 20%	332	795	62,6%	38,5%	1127	43%
Total Geral	**530**	**2064**	**100%**	**100%**	**2594**	**100%**

Frequência: Quantidade de clientes em cada um dos níveis de comprometimento. Por exemplo, na categoria Maior que 40% temos **131** clientes.

Qte Pagou: Quantidade de clientes que fizeram o pagamento em cada um dos níveis de comprometimento. Por exemplo, na categoria Maior que 40%, dos **131** clientes, temos **4** que pagaram.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Comprometimento de Renda** é útil para separar os clientes que **Pagaram e Atrasaram** em um financiamento:

Comprometimento	Classificação				Frequência	% Freq
	Qte Pagou	Qte Atrasou	% Pagou	% Atrasou		
Maior que 40%	4	127	0,8%	6,2%	131	5%
Entre 30 e 40%	92	710	17,4%	34,4%	802	31%
Entre 20 e 30%	102	432	19,2%	20,9%	534	21%
Menor que 20%	332	795	62,6%	38,5%	1127	43%
Total Geral	**530**	**2064**	**100%**	**100%**	**2594**	**100%**

% Pagou: Percentual de clientes pagantes em cada nível de comprometimento em relação ao total de clientes pagantes. Por exemplo, na categoria Maior que 40% temos **4** clientes pagantes de um total de **530** clientes pagantes, logo **0,8%** (4/530) dos clientes pagantes estão na categoria Maior que 40% .

% Atrasou: Mesmo conceito do **% Pagou**, mas aplicado aos clientes que atrasaram o pagamento.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Comprometimento de Renda** é útil para separar os clientes que **Pagaram e Atrasaram** em um financiamento:

Comprometimento	Classificação				Frequência	% Freq	Taxa Pagou
	Qte Pagou	Qte Atrasou	% Pagou	% Atrasou			
Maior que 40%	4	127	0,8%	6,2%	131	5%	3,1%
Entre 30 e 40%	92	710	17,4%	34,4%	802	31%	11,5%
Entre 20 e 30%	102	432	19,2%	20,9%	534	21%	19,1%
Menor que 20%	332	795	62,6%	38,5%	1127	43%	29,5%
Total Geral	**530**	**2064**	**100%**	**100%**	**2594**	**100%**	**20,4%**

Taxa Pagou: Percentual de clientes pagantes em relação ao total de clientes em cada nível de comprometimento. Por exemplo, no Maior que 40% temos **4** clientes pagantes de um total de **131** clientes neste nível de comprometimento, logo **3,1%** (4/131) dos clientes dessa categoria realizaram o pagamento.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Comprometimento de Renda** é útil para separar os clientes que **Pagaram e Atrasaram** em um financiamento:

Comprometimento	Classificação				Frequência	% Freq	Taxa Pagou	Odds	LN(Odds)
	Qte Pagou	Qte Atrasou	% Pagou	% Atrasou					
Maior que 40%	4	127	0,8%	6,2%	131	5%	3,1%	0,12	-2,10
Entre 30 e 40%	92	710	17,4%	34,4%	802	31%	11,5%	0,50	-0,68
Entre 20 e 30%	102	432	19,2%	20,9%	534	21%	19,1%	0,92	-0,08
Menor que 20%	332	795	62,6%	38,5%	1127	43%	29,5%	1,63	0,49
Total Geral	**530**	**2064**	**100%**	**100%**	**2594**	**100%**	**20,4%**		

Odds: Razão entre %Pagou e %Atrasou. Por exemplo, na categoria Maior que 40% temos **0,8%** no %Pagou e **6,2%** no %Atrasou, logo a Odds de **0,12** (0,8%/6,2%) é a chance de encontrarmos um cliente que pagou nessa categoria. Ou seja, a proporção na categoria Maior que 40% é de aproximadamente 1 cliente que pagou para 6 clientes que atrasaram.

LN(Odds): Logarítimo Natural da Odds. Por exemplo, na categoria Maior que 40% temos LN(**0,12**) = **-2,10**.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Comprometimento de Renda** é útil para separar os clientes que **Pagaram e Atrasaram** em um financiamento:

Comprometimento	Classificação				Frequência	% Freq	Taxa Pagou	Odds	LN(Odds)	IV
	Qte Pagou	Qte Atrasou	% Pagou	% Atrasou						
Maior que 40%	4	127	0,8%	6,2%	131	5%	3,1%	0,12	-2,10	0,11
Entre 30 e 40%	92	710	17,4%	34,4%	802	31%	11,5%	0,50	-0,68	0,12
Entre 20 e 30%	102	432	19,2%	20,9%	534	21%	19,1%	0,92	-0,08	0,00
Menor que 20%	332	795	62,6%	38,5%	1127	43%	29,5%	1,63	0,49	0,12
Total Geral	**530**	**2064**	**100%**	**100%**	**2594**	**100%**	**20,4%**			**0,35**

IV: Produto da diferença entre %Pagou e %Atrasou pelo LN(Odds). Por exemplo, na categoria Maior que 40% temos (**0,8%** - **6,2%**) * **-2,10** = **0,11**.

Para fins de medida de associação, estamos interessados na soma dos **IV's** de cada categoria da variável.

Neste exemplo, o **IV Total** = 0,11 + 0,12 + 0,00 + 0,12 = **0,35**

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Preditiva.ai

Após calcularmos o IV, como avaliamos se a variável possui um **alto poder de separação**?

Abaixo apresentamos uma referência bastante **utilizada na prática**:

IV Total	Poder de separação
< 0,02	Muito fraco
0,02 a 0,1	Fraco
0,1 a 0,3	Médio
0,3 a 0,5	Forte
> 0,5	Muito bom pra ser verdade...Verifique!

Em nosso exemplo, a variável **Comprometimento** obteve um IV = 0,35, ou seja, ela possui um **forte** poder de separação entre os bons e maus clientes.

Logo, ao perguntar o comprometimento de renda de um cliente, é possível ter uma estimativa se ele pagará em dia ou atrasará *.

* Cuidado com a questão da causalidade. Não é possível inferir com **certeza** que o comprometimento é a real CAUSA da inadimplência. É preciso olhar a relação com outras variáveis, como bens, tipo de profissão, por exemplo.

Demonstração

Arquivo: “Demonstração - Análise Bidimensional.xlsx”

Revisão

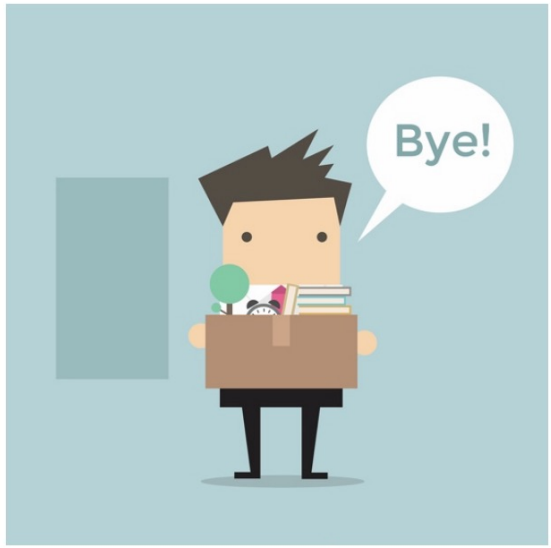
Nesta última seção aprendemos:

- Como realizar **Análises Bidimensionais** com **variáveis qualitativas** e **quantitativas**.
- Visualizar a **associação entre duas variáveis quantitativas** utilizando o **Gráfico de Dispersão**.
- Medir a **força da associação** entre essas variáveis utilizando o **Coeficiente de Correlação de Pearson**, o **Coeficiente de Determinação** ou **R^2** e o **Information Value (IV)**.
- **Nem toda correlação entre variáveis representa causalidade!**

Hands On
Turn Over RH

Preditiva.ai

Turnover de funcionários em uma empresa de Tecnologia



Para quarta dia 12/01/22:

Parte 1) Faça um análise unidimensional das variáveis da base.

a) Para variáveis qualitativas, crie tabelas de frequência absoluta, relativas e acumuladas.
b) Para variáveis quantitativas, crie histogramas e box plots.
c) Comente o mais relevante.

Parte 2) Faça a análise bidimensional das variáveis da base em relação à TurnOver. Dica: Utilize a análise de IV

a) Crie um ranking de IV das variáveis da base
b) Argumente a possível causalidade de cada variável acima de 0,1
c) Será que não existe correlação entre as variáveis da base? Ex: As pessoas saem mais por que são solteiras ou saem mais por que ganham pouco e, nesta empresa, quem ganha pouco é em geral solteiro(a)?

Qual a sua conclusão? Quais as variáveis que mais têm relação de causalidade com o TurnOver na sua opinião? Dica: Tente buscar evidências externas (estudos, matérias de RH) que corroboram com sua conclusão.

Framework de Data Analytics e Data Science

Etapas de um projeto

Preditiva.ai

Vamos encapsular todo o conhecimento que vimos anteriormente em uma metodologia de desenvolvimento. Uma abordagem comum é a chamada **CRISP-DM**. Veja:

Entendimento do negócio	Entendimento dos dados	Preparação dos dados	Modelagem/ Análise	Validação	Deploy (Implantação)
❑ Objetivo	❑ Coleta	❑ Seleção de variáveis	❑ Escolha da técnica	❑ Verificação dos critérios de sucesso	❑ Plano de implantação
❑ Premissas	❑ Descrição dos dados	❑ Limpeza dos dados	❑ Desenvolvi-mento do modelo ou estudo	❑ Revisão do processo/Validação do modelo	❑ Monitoramento do modelo/KPI's
❑ Risco envolvidos	❑ Análise Exploratória	❑ Formatação dos dados	❑ Finalização	❑ Determinação dos próximos passos	❑ Criação da documentação
❑ Custos	❑ Sanity Check				
❑ Critérios de sucesso					
❑ Planejamento					

Próximos passos

Agora que já sabemos como realizar uma boa **Análise Exploratória** de um conjunto de dados, pode surgir a dúvida: “E se eu quiser saber qual a **Probabilidade** de observar um determinado valor em uma variável do meu conjunto de dados?”

Esse será exatamente o assunto da próxima aula! Veremos como o estudo da **Probabilidade** pode nos ajudar a responder a essa e outras perguntas sobre **eventos de caráter aleatório**.

Preditiva.ai