



# **Probabilidade e inferência estatística com R - Módulo 3**

Prof. Suellen Teixeira Zavadzki de Pauli

# Objetivos





- Compreender os conceitos de correlação e regressão;
- Avaliar a correlação entre variáveis por meio de gráficos e teste;
- Estimar e visualizar um modelo de regressão;
- Interpretar coeficientes de regressão e estatísticas no contexto de problemas reais;
- Compreender os conceitos da Análise de Variância.

# Correlação e Regressão Linear Simples

- Em determinadas situações, estamos interessados em descrever a relação entre duas variáveis ou até prever o valor de uma a partir da outra.
- **Exemplos:**
  - Qual o peso de determinado indivíduo se sabemos que a altura dele é  $X$ ?
  - Qual o consumo de combustível, em litros, dado que o carro percorreu uma distância de  $X$  km?
  - Qual a relação entre a renda semanal de uma família e as despesas de consumo?

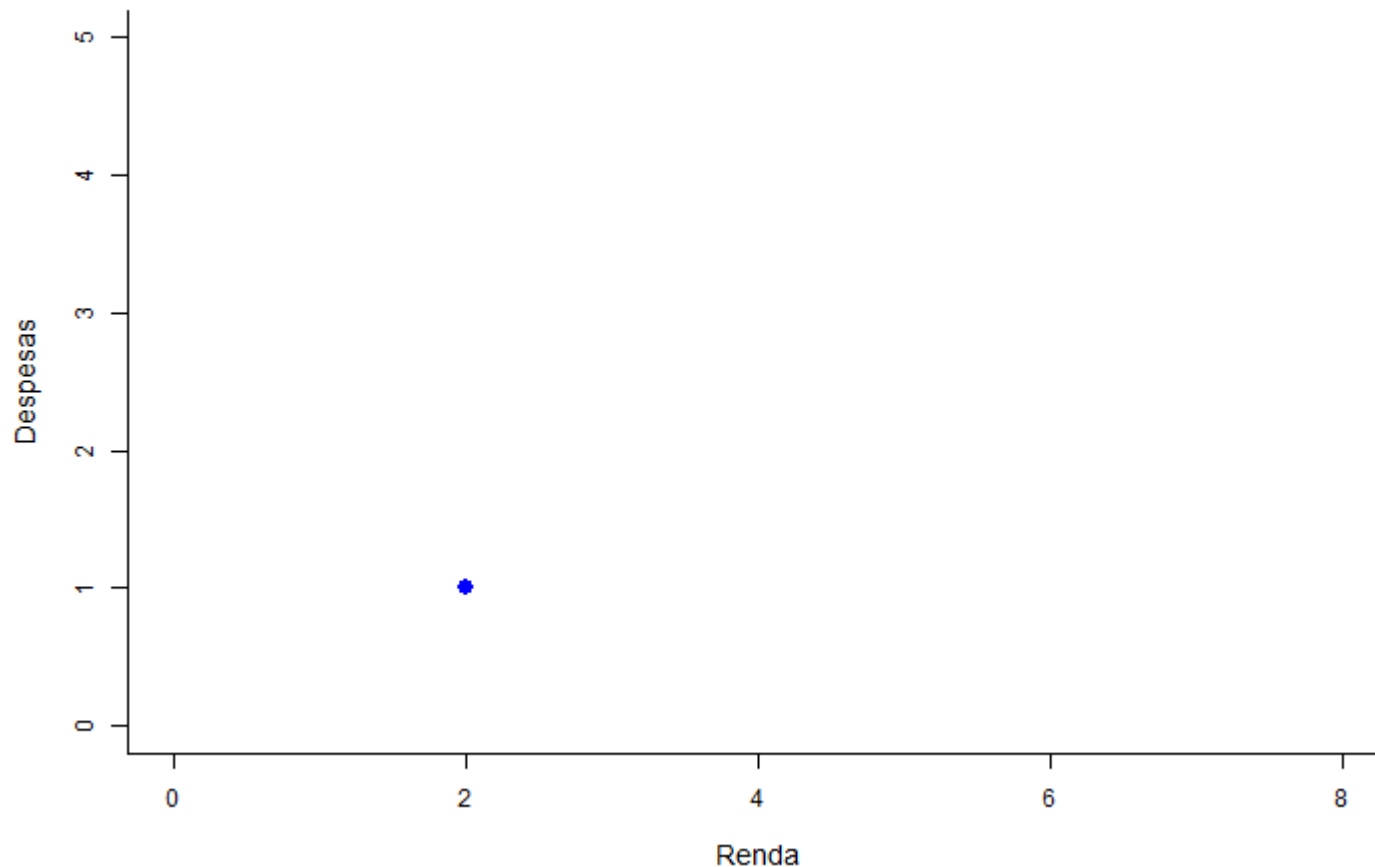
# Introdução













	RENDA	DESPESAS
FAMÍLIA 1		

# Introdução

Renda x Despesas

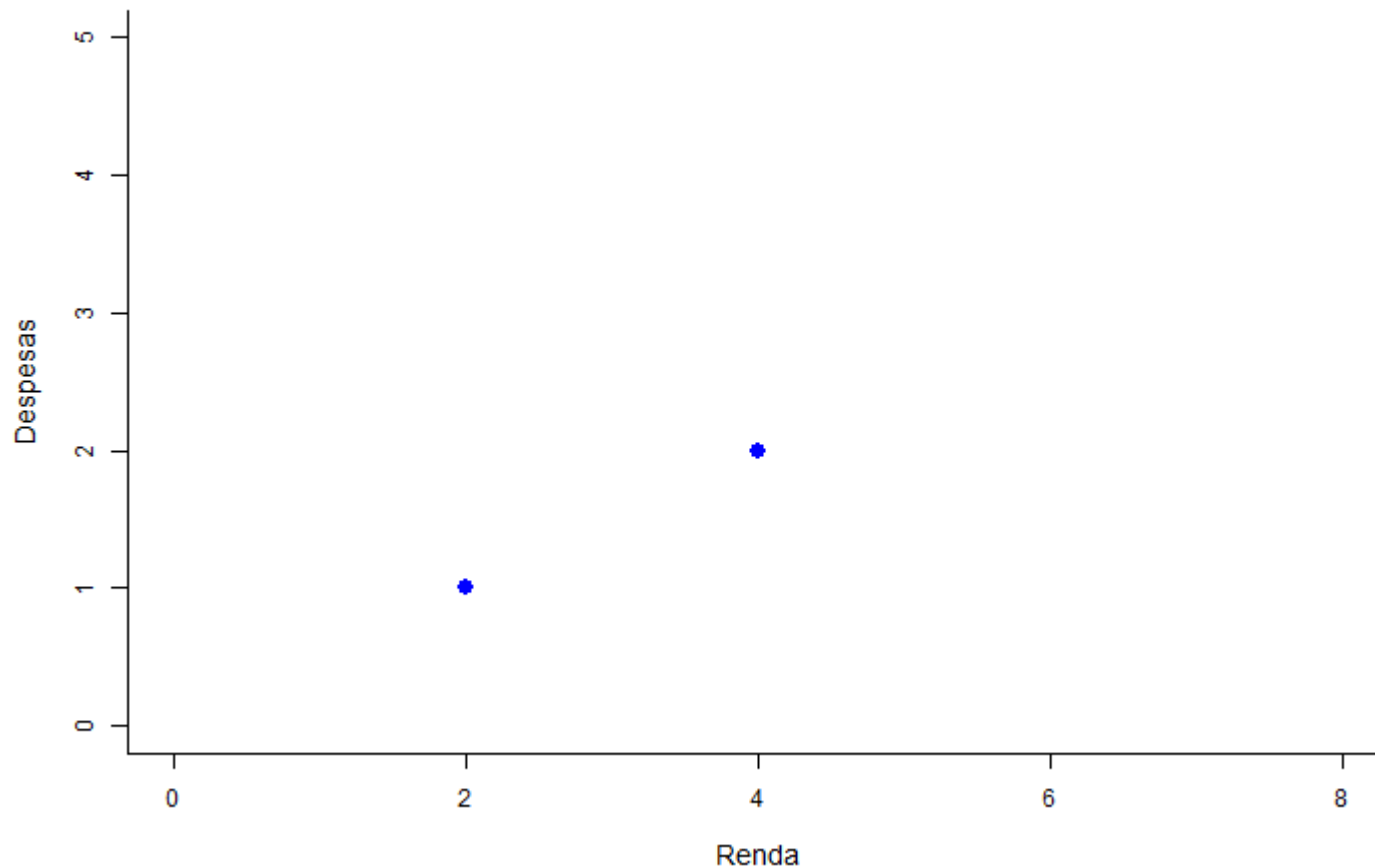


# Introdução

	RENDA	DESPESAS
FAMÍLIA 1	 	
FAMÍLIA 2	    	 







# Introdução

Renda x Despesas



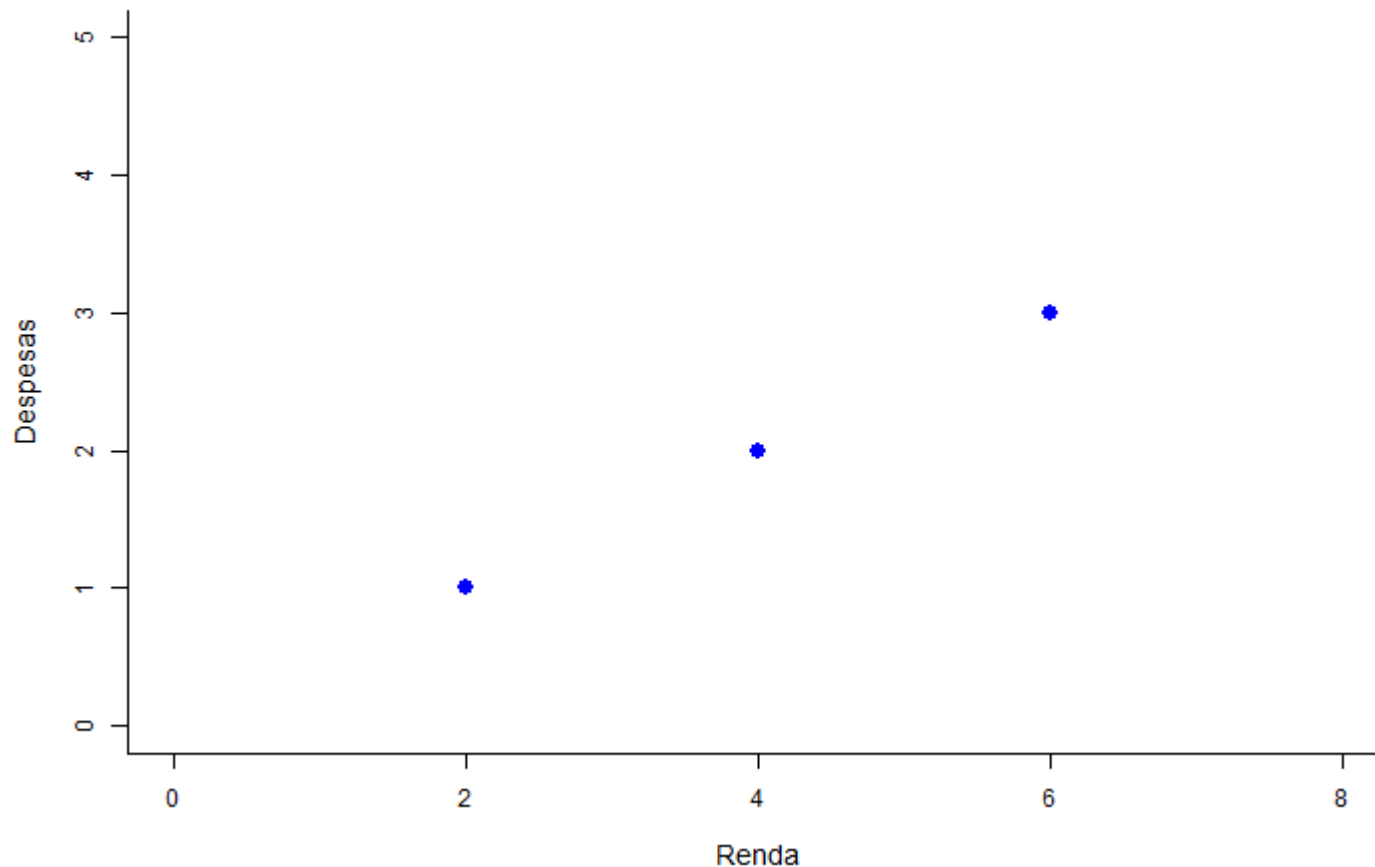


# Introdução

	RENDA	DESPESAS
FAMÍLIA 1		
FAMÍLIA 2		
FAMÍLIA 3		

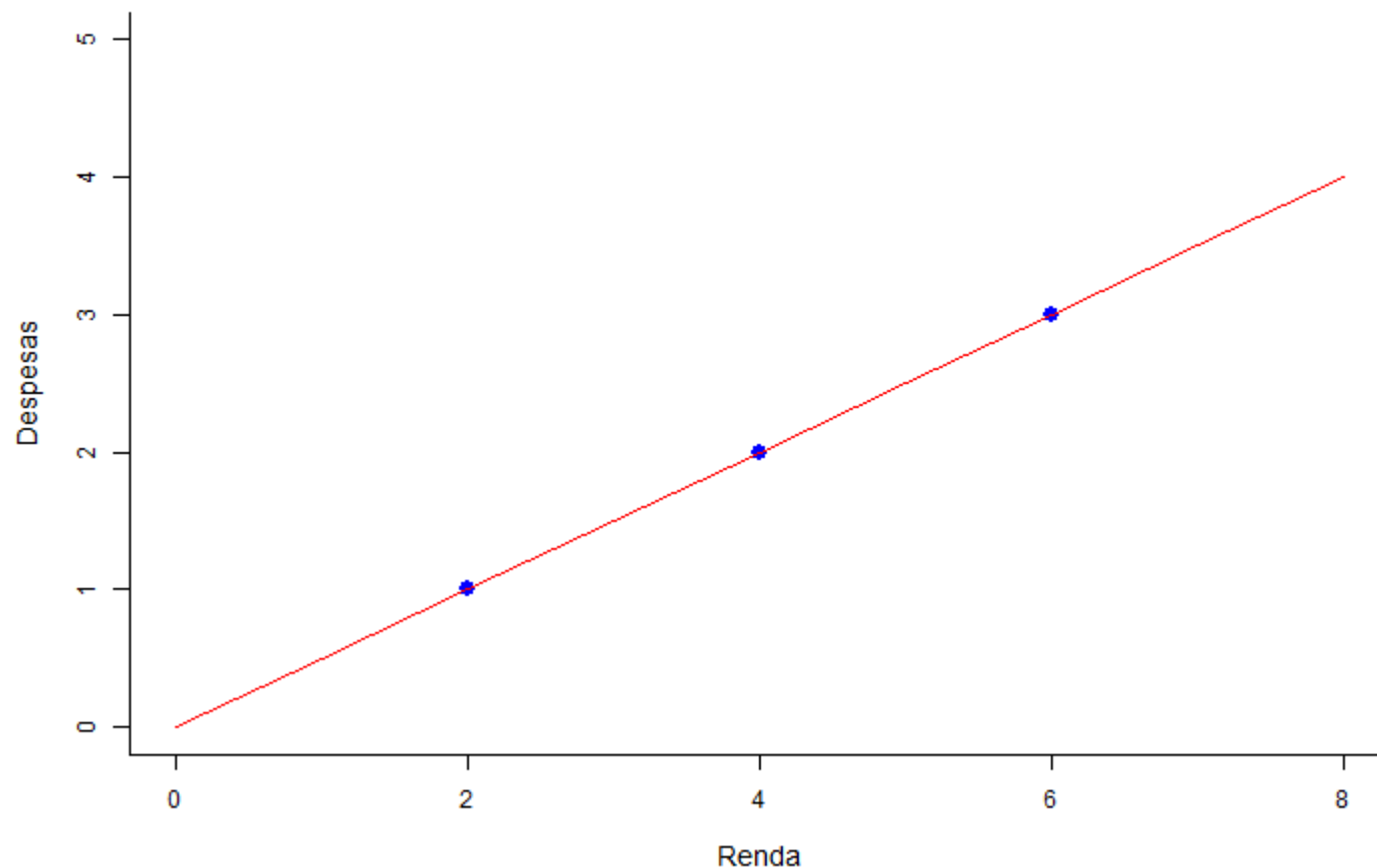
# Introdução

Renda x Despesas











# Introdução

**Renda x Despesas**

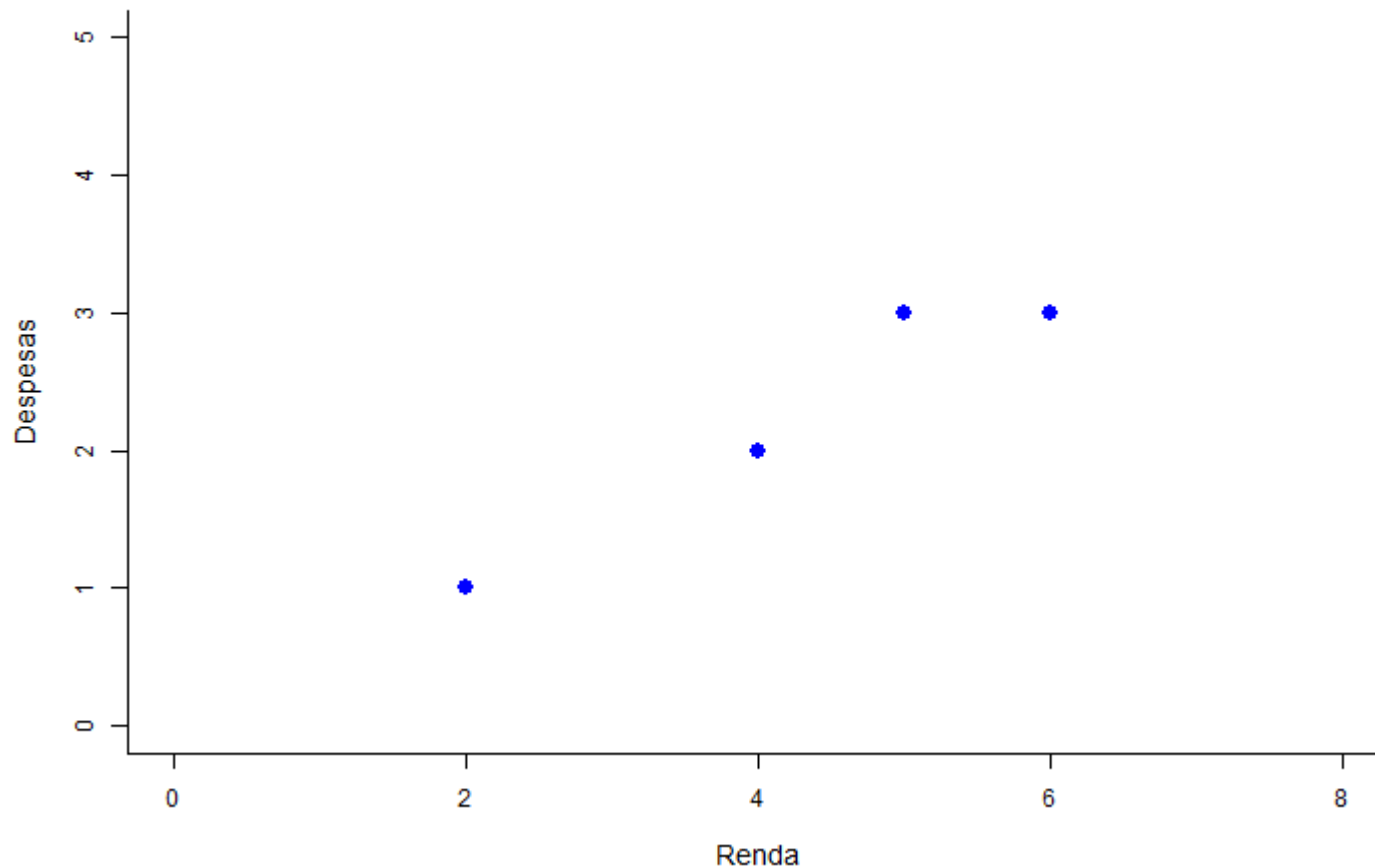


# Introdução

	RENDA	DESPESAS
FAMÍLIA 1		
FAMÍLIA 2		
FAMÍLIA 3		
FAMÍLIA 4		

# Introdução

Renda x Despesas

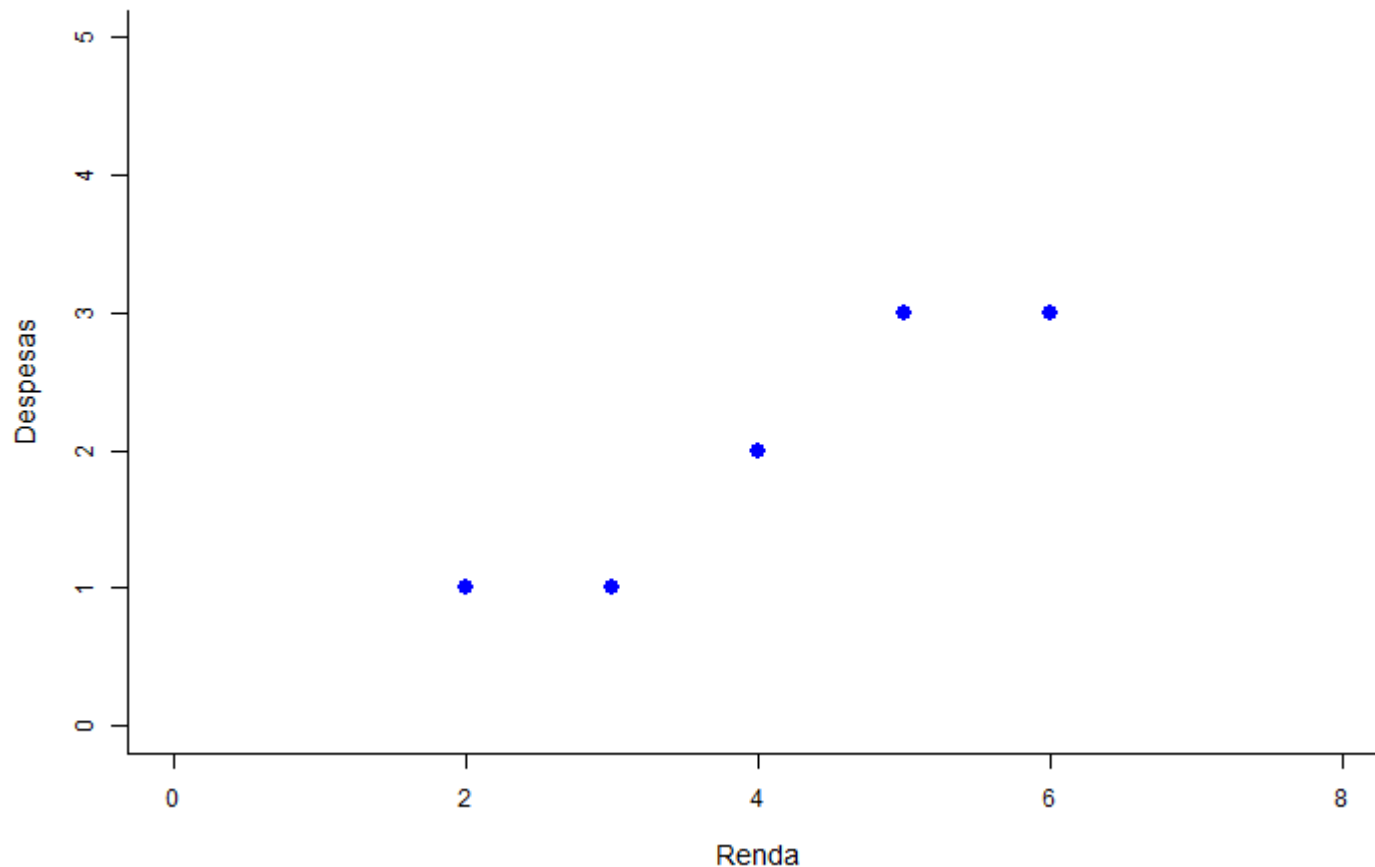


# Introdução

	RENDA	DESPESAS
FAMÍLIA 1	2	1
FAMÍLIA 2	4	2
FAMÍLIA 3	6	3
FAMÍLIA 4	5	3
FAMÍLIA 5	3	1

# Introdução

Renda x Despesas



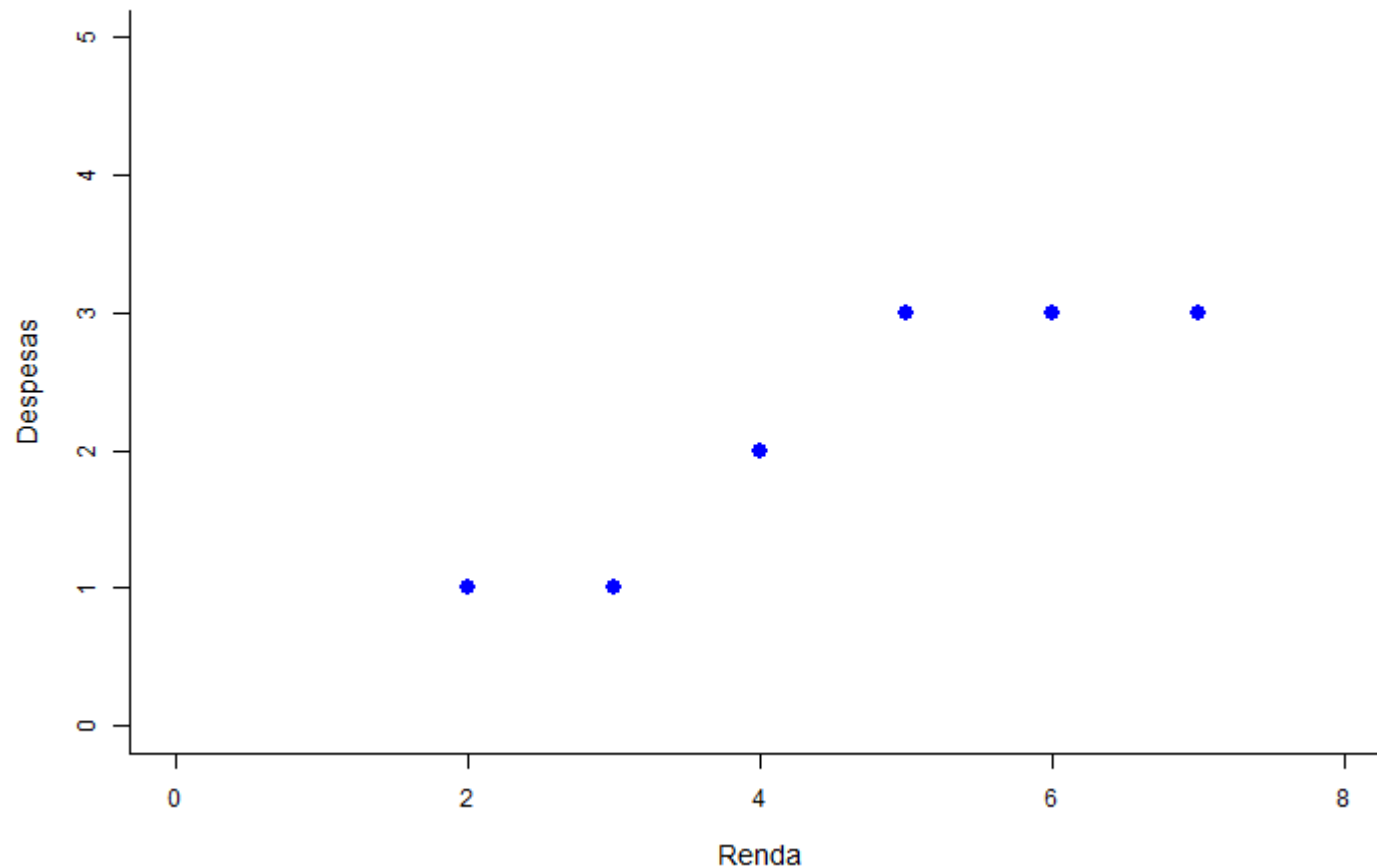
# Introdução

	RENDA	DESPESAS
FAMÍLIA 1	2	1
FAMÍLIA 2	4	2
FAMÍLIA 3	6	3
FAMÍLIA 4	5	3
FAMÍLIA 5	3	1
FAMÍLIA 6	6	3

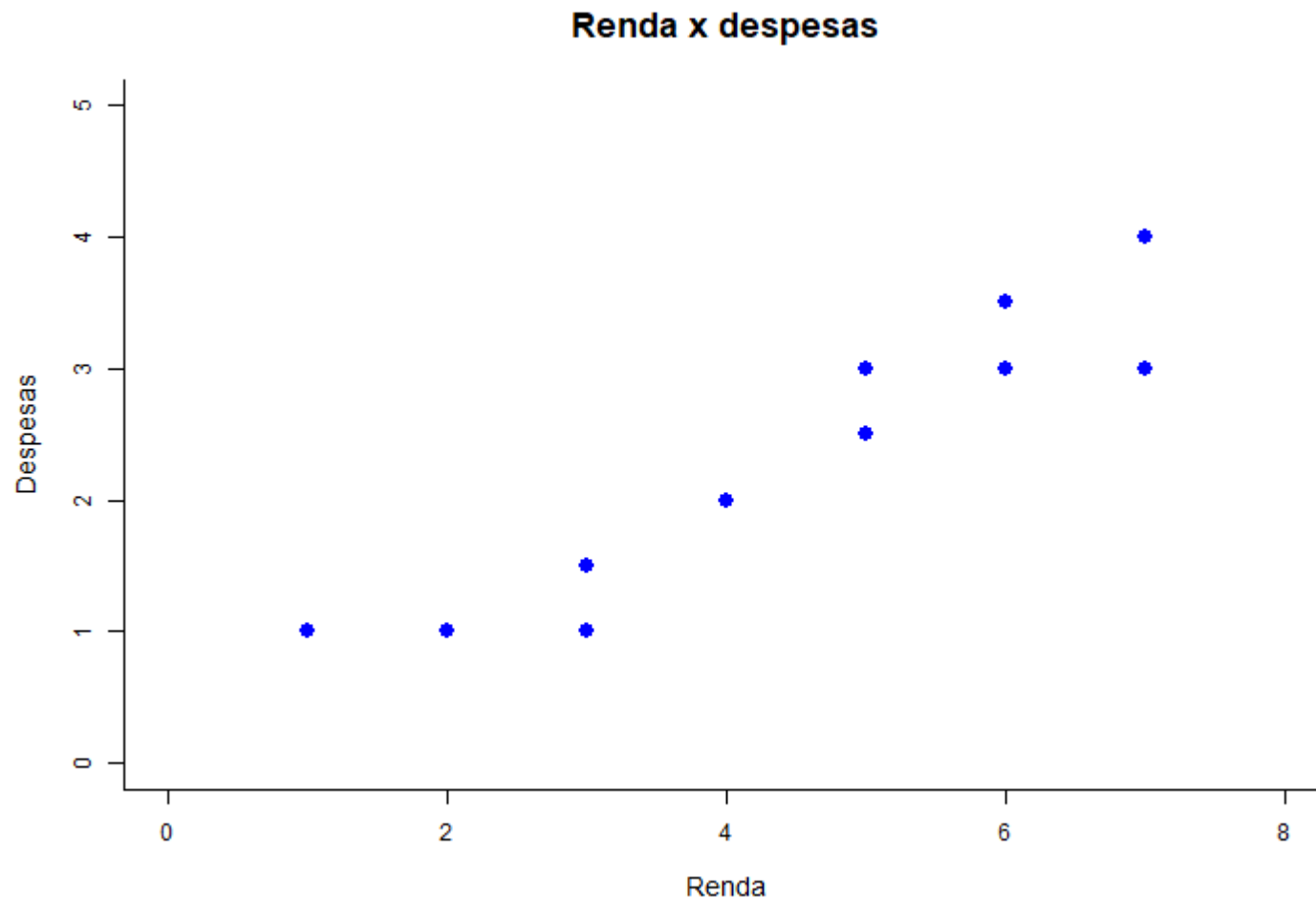


# Introdução

Renda x Despesas

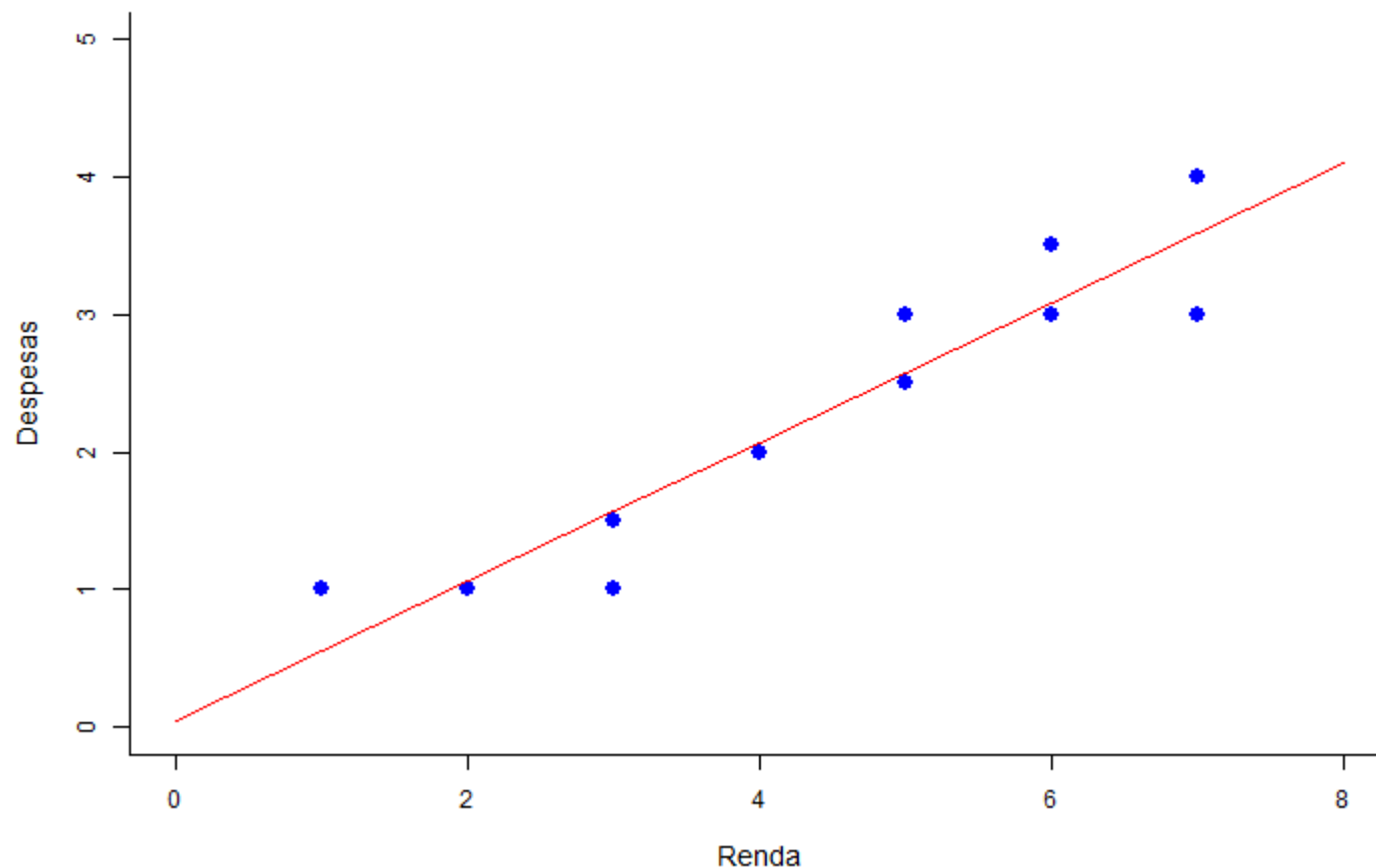


# Introdução



# Introdução

Renda x despesas



"All models are wrong but some are useful"

**George Box**

- Estudar a relação linear entre duas variáveis quantitativas:
  - Explicitando a forma dessa relação: **regressão**
    - É indispensável identificar qual variável é a variável dependente.
  - Quantificando a força ou o grau dessa relação: **correlação**
    - Não é necessário identificar qual variável é a variável dependente, pois queremos estudar o grau de relacionamento entre as variáveis  $X$  e  $Y$ , ou seja, uma medida de covariabilidade entre elas.
    - A correlação é considerada como uma medida de influência mútua entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.

# Diagrama de dispersão



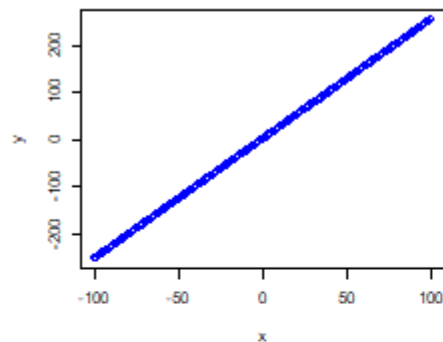
- Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

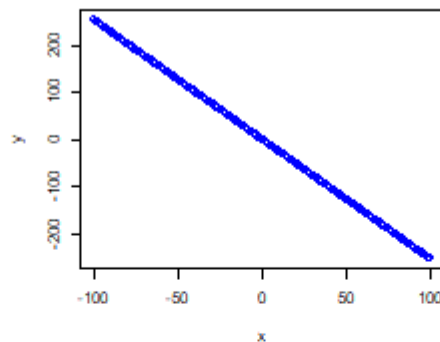
- Com base no conjunto de dados é possível construir um diagrama de dispersão, o qual deve exibir uma tendência linear para que se possa usar a regressão linear;
- Com isso podemos decidir empiricamente se um relacionamento linear entre  $X$  e  $Y$  pode ser assumido;
- É possível verificar se o grau de relacionamento linear entre as variáveis é forte ou fraco.

# Diagrama de dispersão

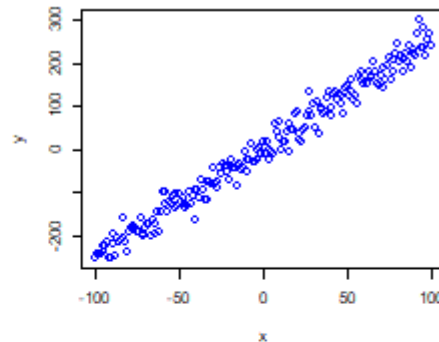
Correlação perfeita positiva



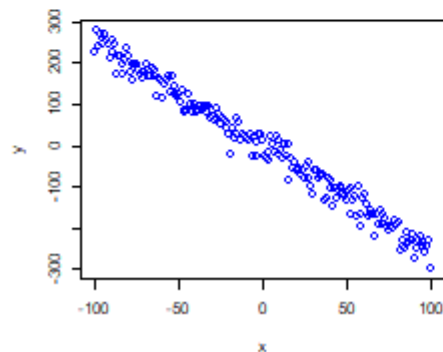
Correlação perfeita negativa



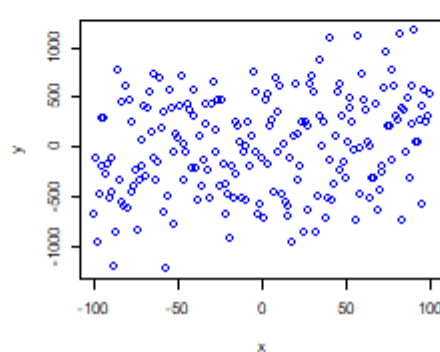
Alta correlação positiva



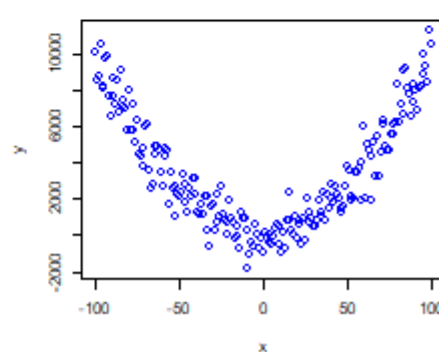
Alta correlação negativa



Baixa correlação



Correlação não linear



# Coeficiente de correlação linear

O grau de relação entre duas variáveis pode ser medido através do coeficiente de correlação linear ( $r$ ), dado por

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}}$$

onde  $-1 \leq r \leq 1$ ;

- $r = 1$ : relação linear perfeita positiva entre  $X$  e  $Y$ ;
- $r = 0$ : : inexistência de relação linear entre  $X$  e  $Y$ ;
- $r = -1$ : relação linear perfeita negativa entre  $X$  e  $Y$ ;
- $r > 0$ : relação linear positiva entre  $X$  e  $Y$ ;
- $r < 0$ : relação linear negativa entre  $X$  e  $Y$ ;



# Coeficiente de determinação



- Existem muitos tipos de associações possíveis, e o coeficiente de correlação avalia o quanto uma nuvem de pontos no gráfico de dispersão se aproxima de uma reta;
- O coeficiente de determinação (  $r^2$  ) é o quadrado do coeficiente de correlação, por consequência;

$$0 \leq r^2 \leq 1$$

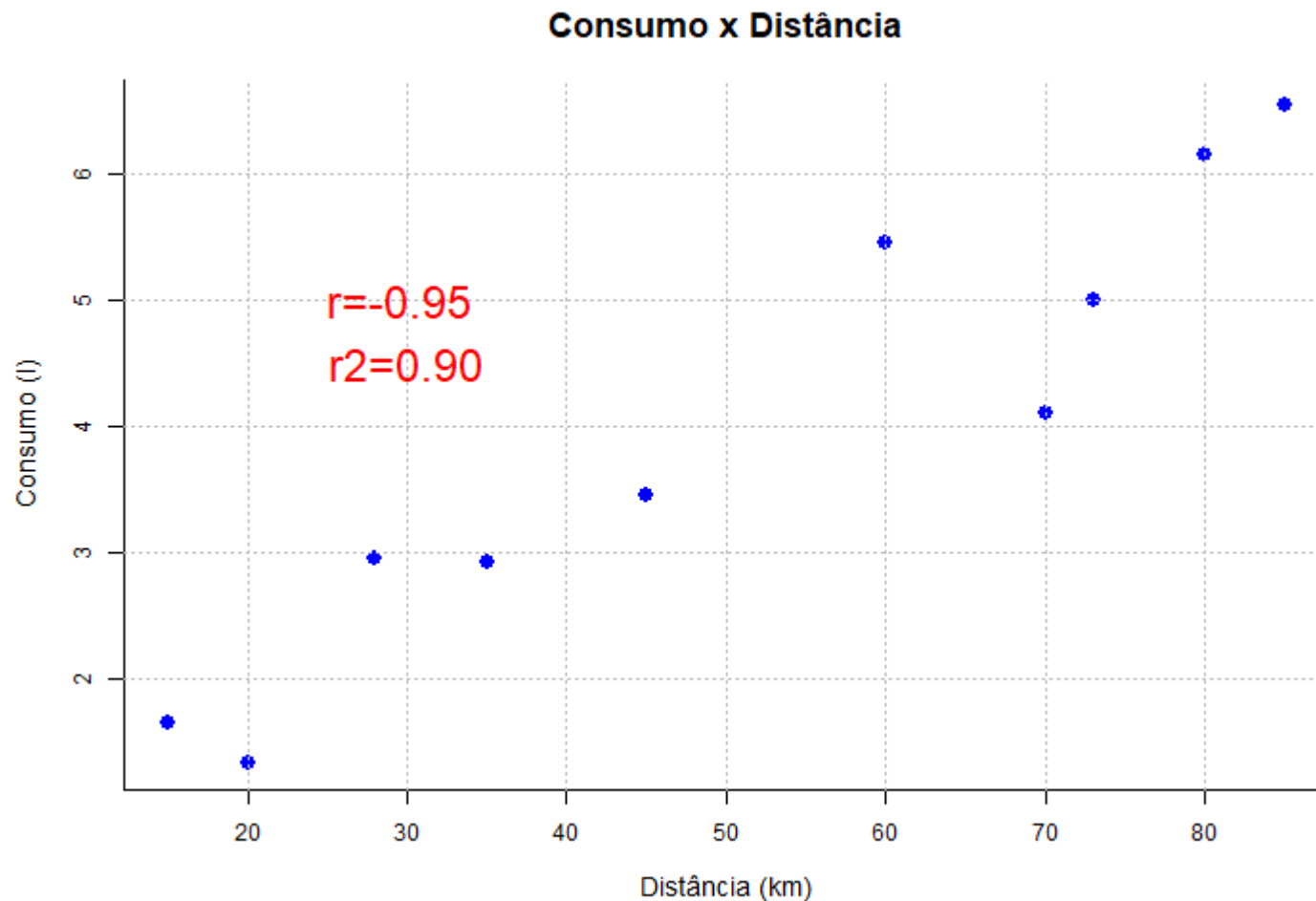
- O  $r^2$  nos dá a porcentagem de variação em  $Y$  que pode ser explicada pela variável independente  $X$ .
- Quanto mais próximo de 1, maior é a explicação da variável  $Y$  pela variável  $X$ .

# Exercício 1

- A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos.
  - Calcule o coeficiente de correlação linear e o coeficiente de determinação.
  - Faça um diagrama de dispersão.

<i>Distância</i>	20.00	60.00	15.00	45.00	35.00	80.00	70.00	73.00	28.00	85.00
<i>Consumo</i>	1.33	5.45	1.66	3.46	2.92	6.15	4.11	5.00	2.95	6.54

# Exercício 1



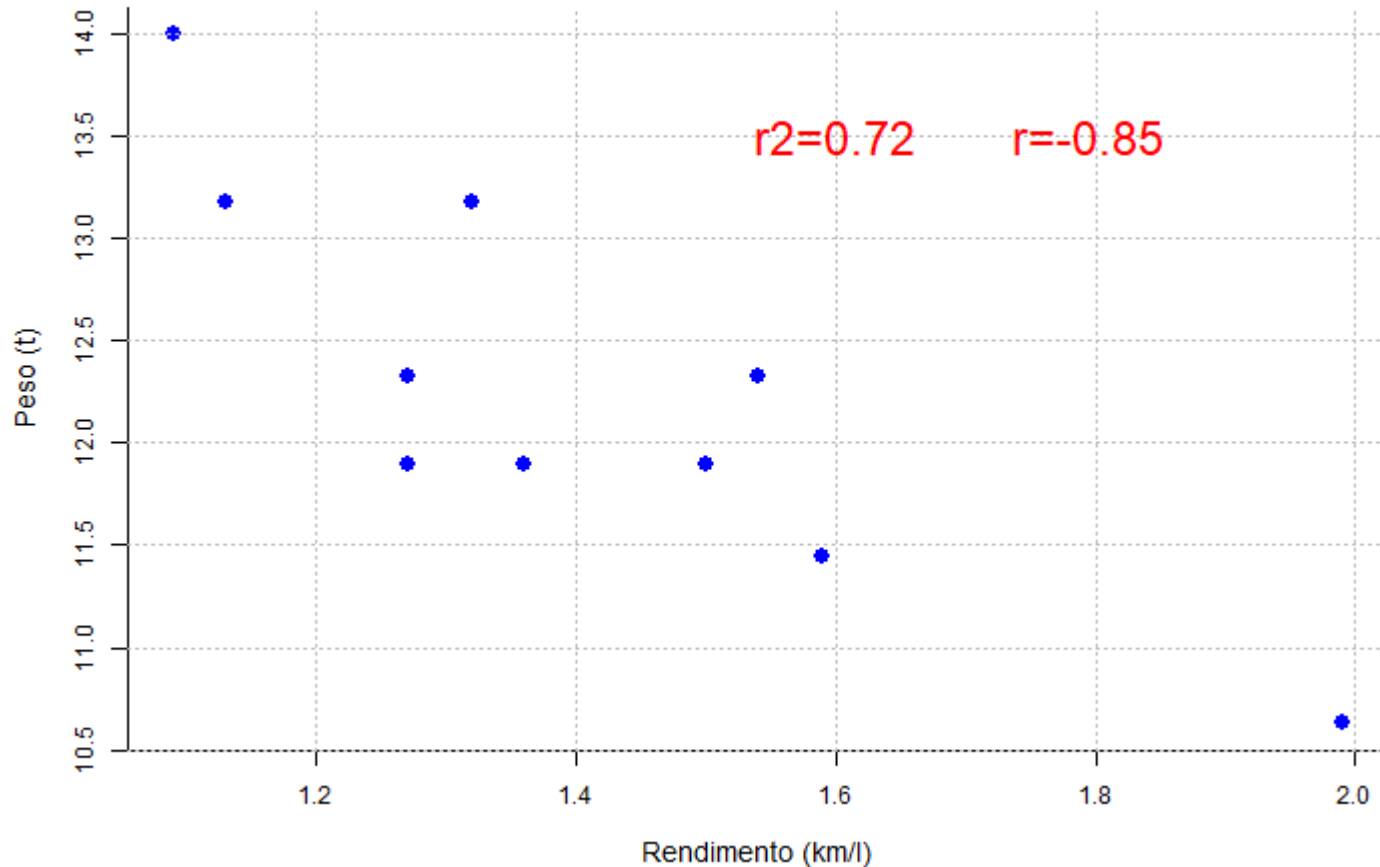
# Exercício 2

- A tabela a seguir relaciona os pesos de carros (t) e o rendimento de combustível (em km/l), para uma amostra de 10 carros.
  - Calcule o coeficiente de correlação linear e o coeficiente de determinação.
  - Faça um diagrama de dispersão.

<i>Peso</i>	1.32	1.59	1.27	1.99	1.13	1.54	1.36	1.5	1.27	1.09
<i>Rendimento</i>	13.18	11.45	12.33	10.63	13.18	12.33	11.90	11.9	11.90	14.00

# Exercício 2

Peso x Rendimento



# Teste para o coeficiente de correlação



- Usualmente definimos o coeficiente de correlação para uma amostra, pois desconhecemos esse valor para a população.
- Uma população que tenha duas variáveis não correlacionadas pode produzir uma amostra com coeficiente de correlação diferente de zero.
- Para testar se uma amostra foi colhida de uma população para a qual o coeficiente de correlação entre duas variáveis é nulo, precisamos obter a distribuição amostral da estatística  $r$ .

# Teste para o coeficiente de correlação

- Seja  $\rho$  o verdadeiro coeficiente de correlação populacional desconhecido. Seja  $\rho$  o verdadeiro coeficiente de correlação populacional desconhecido.
- Para testar se o coeficiente de correlação populacional é igual a zero, realizamos um teste de hipótese com

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

A estatística de teste utilizada é

$$t_{calc} = r \sqrt{\frac{n - 2}{1 - r^2}}$$

que tem distribuição  $t$  de Student com  $n - 2$  graus de liberdade.

# Teste para o coeficiente de correlação



## Procedimentos gerais

- Hipóteses  $H_0 : \rho = 0, H_1 : \rho \neq 0$
- Nível de significância  $\alpha$
- Verificar a região de rejeição com base no nível de significância  $t_{crit}$ , com  $n - 2$  graus de liberdade
- Cálculo da estatística do teste sob a hipótese nula

$$t_{calc} = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Rejeitar a hipótese nula se a estatística de teste calculada estiver dentro da região de rejeição ou  $|t_{calc}| > |t_{crit}|$



# Exercício 1 - continuação



Com base nas informações do exercício 1, realize o teste de hipótese para o coeficiente de correlação  $\rho$ , usando um nível de 5% de significância.

$$r = 0.95$$

# Exercício 1 - continuação



Com base nas informações do exercício 1, realize o teste de hipótese para o coeficiente de correlação  $\rho$ , usando um nível de 5% de significância.

$$r = 0.95$$

$$t_{calc} > t_{crit}$$

$$8.605 > 2.306$$

Rejeitamos a hipótese nula de que não há correlação entre as variáveis, com 95% de confiança.

# Exercício 2 - continuação



Com base nas informações do exercício 2, realize o teste de hipótese para o coeficiente de correlação  $\rho$ , usando um nível de 5% de significância.

$$r = 0.85$$

# Exercício 2 - continuação



Com base nas informações do exercício 2, realize o teste de hipótese para o coeficiente de correlação  $\rho$ , usando um nível de 5% de significância.

$$r = 0.85$$

$$t_{calc} = -4.563861$$

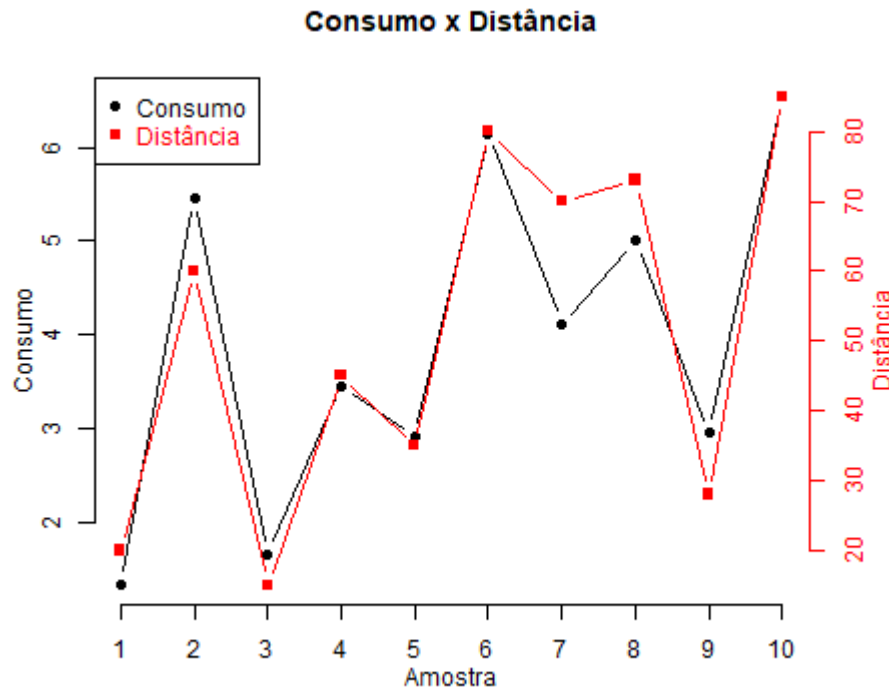
$$|t_{calc}| > |t_{crit}|$$

$$4.56 > 2.306$$

Rejeitamos a hipótese nula de que não há correlação entre as variáveis, com 95% de confiança.

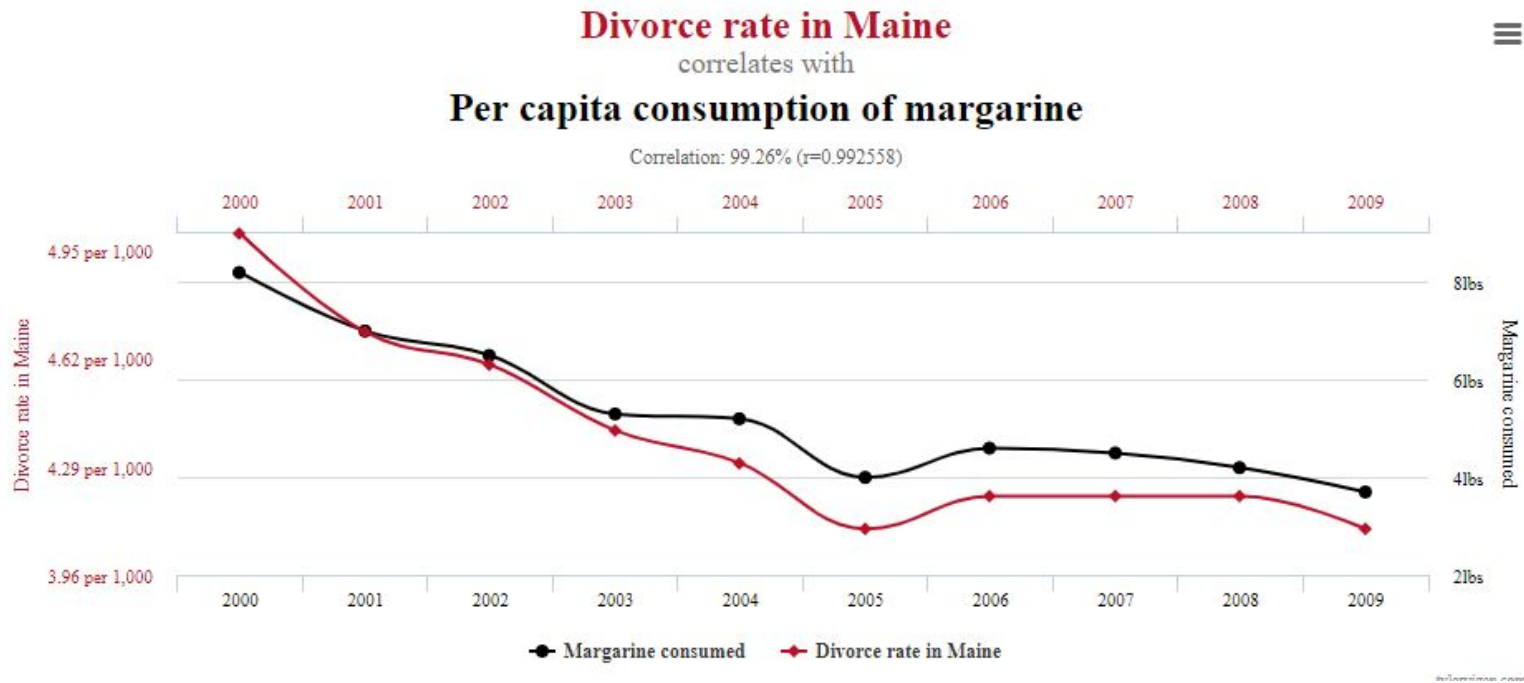
# Exercício 1 - continuação

- Construa um gráfico no qual seja possível visualizar os valores das duas variáveis no eixo y.



# Correlação x Causalidade ?

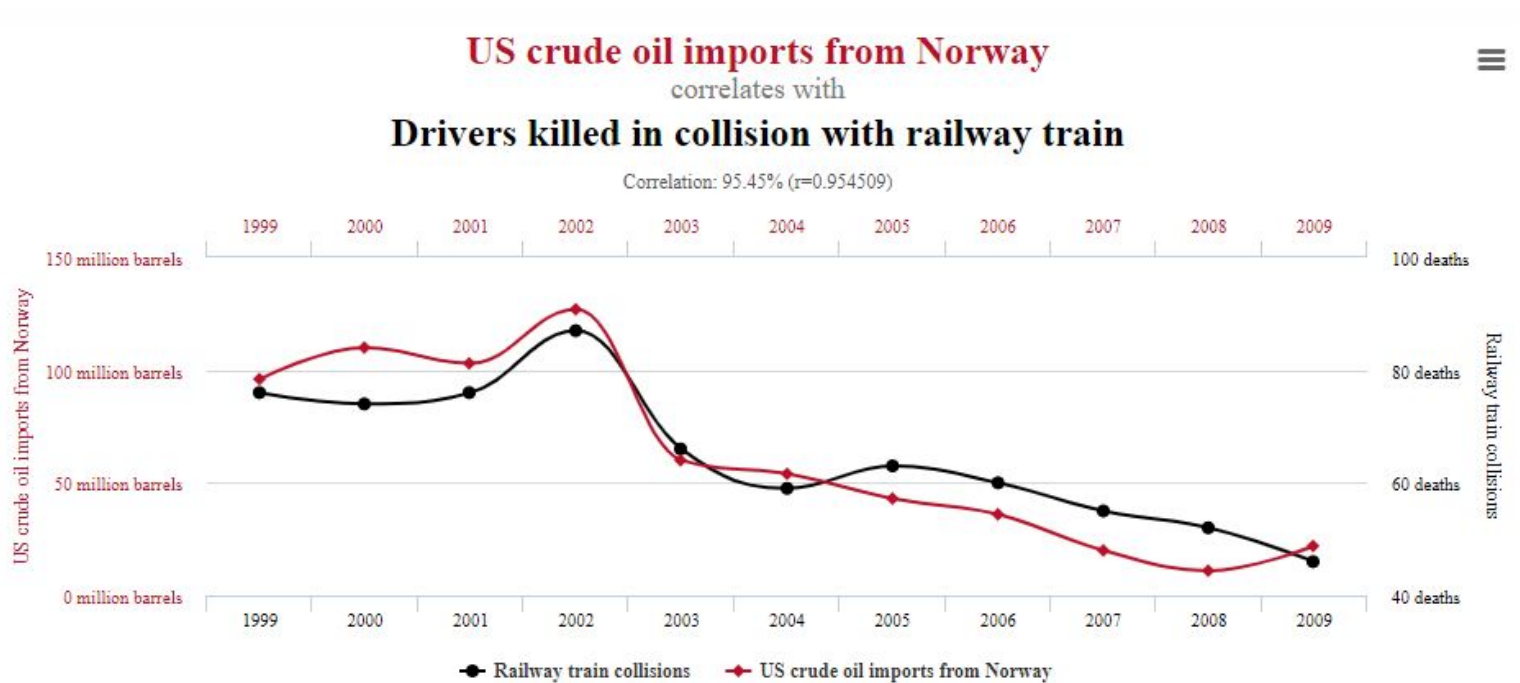
Taxa de divórcio no Maine se correlaciona com o consumo per capita de margarina



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

# Correlação x Causalidade ?

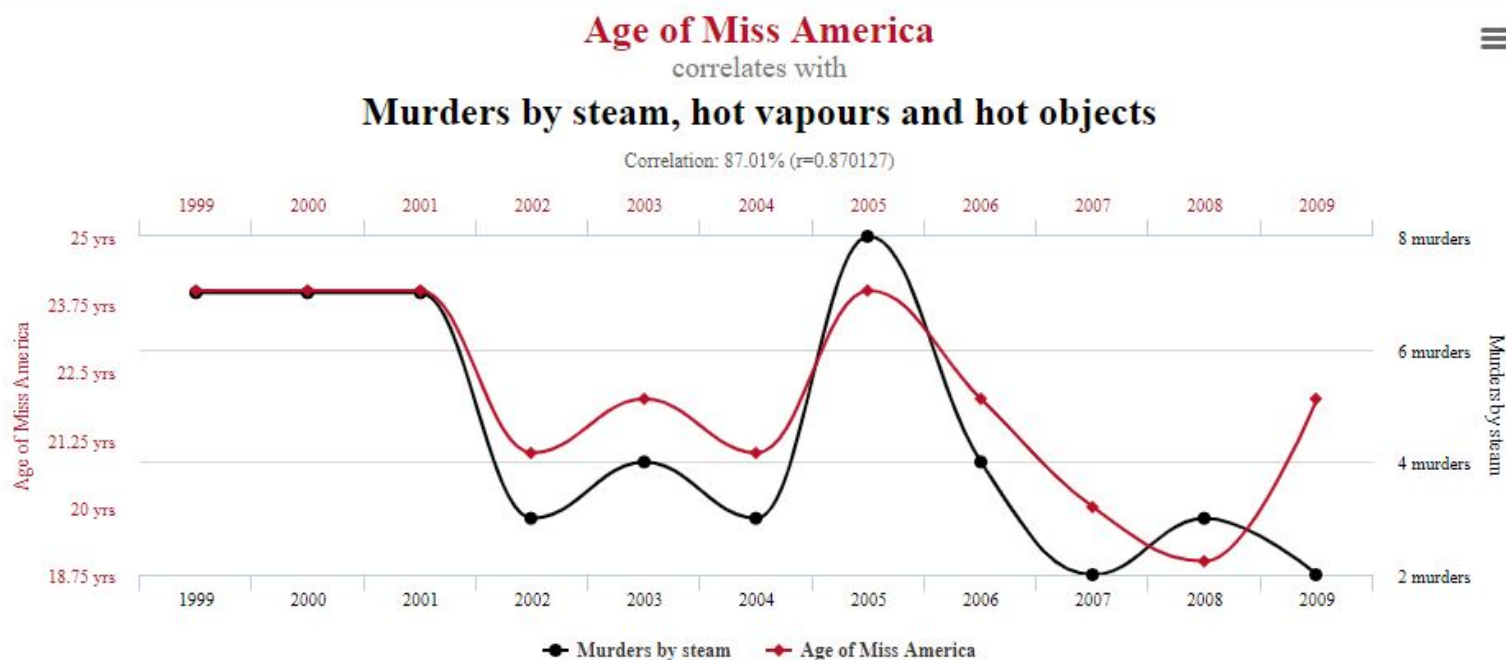
As importações de petróleo bruto dos EUA da Noruega se correlacionam com motoristas mortos em colisão com trem ferroviário



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

# Correlação x Causalidade ?

A idade da miss America correlaciona-se com assassinatos por vapor, vapores quentes e objetos quentes



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



# Matriz de correlação

- Base de dados mtcars

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
## Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
## Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
## Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
## Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1