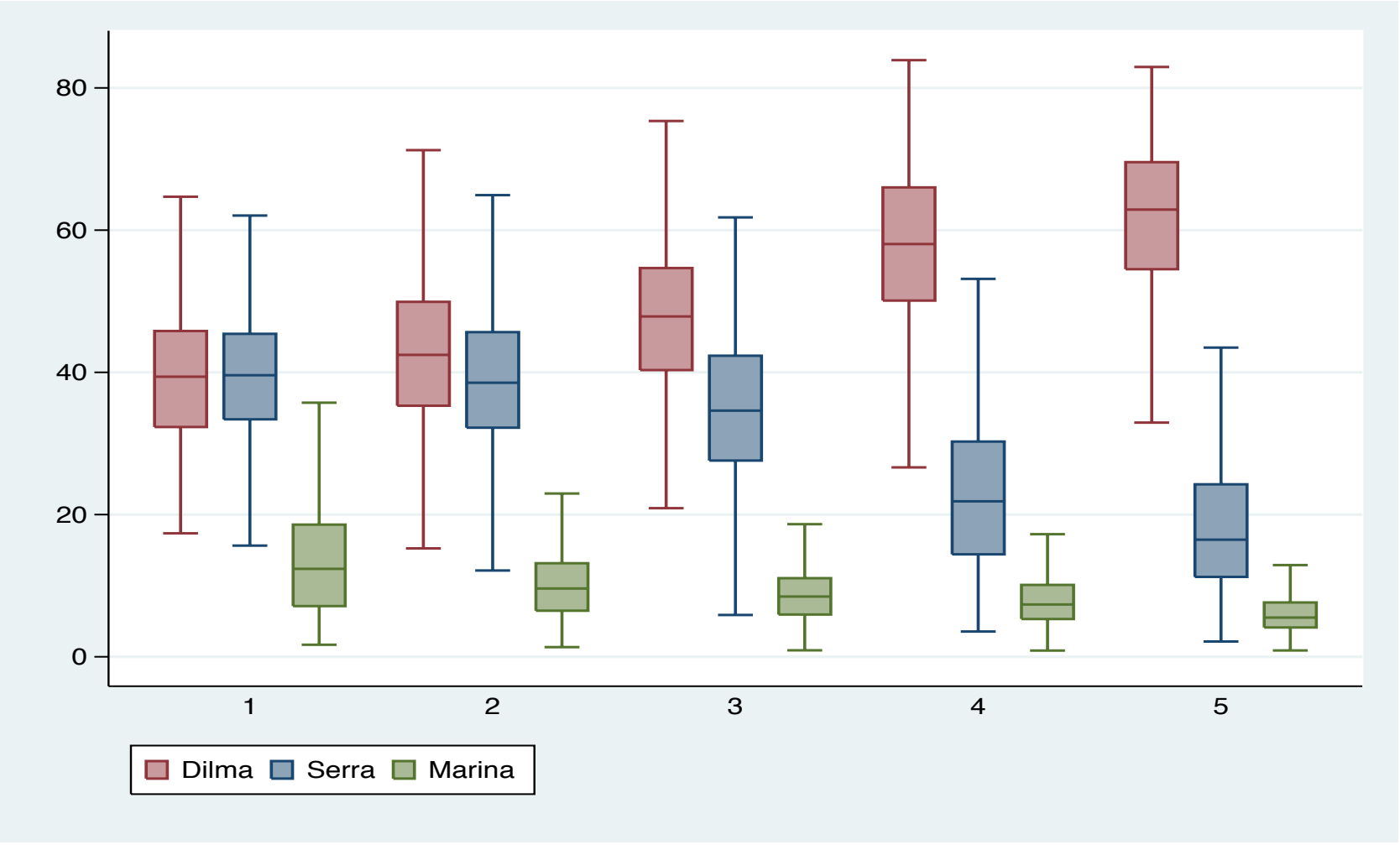
Aula 6: Explorando variáveis quantitativas com gráficos

Boxplot

Você já viu um boxplot?



Boxplot

- Um boxplot é uma apresentação gráfica do sumário de cinco números.
- Boxplots são úteis para comparar grupos.
- Boxplots são particularmente eficientes para assinalar os outliers.

O sumário de cinco números

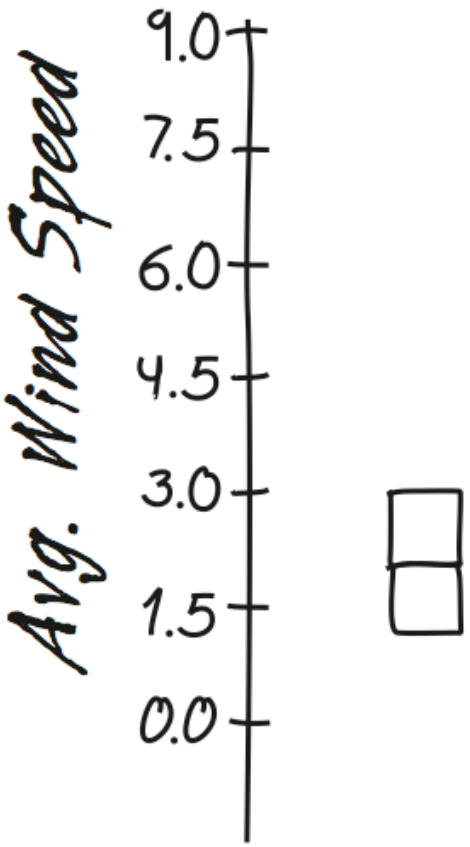
- O sumário de cinco números da distribuição reporta a mediana, os quartis e os extremos (máximo e mínimo).

Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

  - Exemplo: o sumário de cinco números para velocidade diária do vento é:

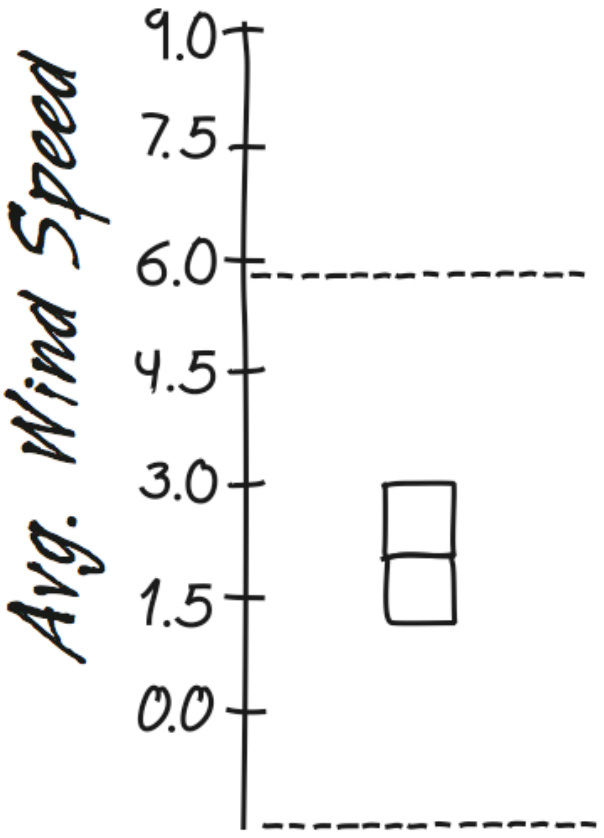
Construindo um boxplot

- Desenhe uma linha vertical que contenha todos os valores da distribuição.
- Desenhe pequenas linhas horizontais nos quartis inferior, superior e na mediana.
- Conecte-os com uma linha vertical para formar um caixa (box).



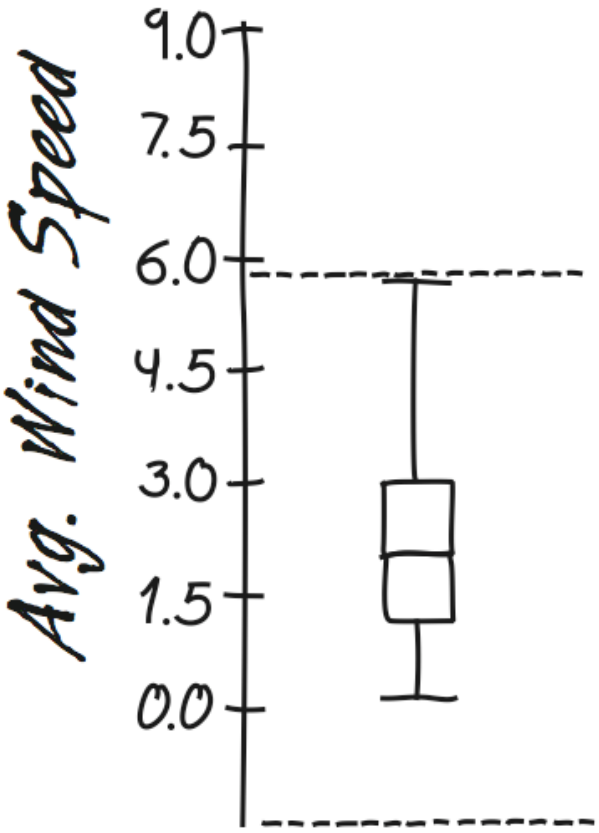
Construindo um boxplot (cont.)

- Eleve "cercas" que cobrirão a área principal dos dados.
  - A cerca superior está a 1.5 IQRs acima do quartil superior.
  - A cerca inferior está a 1.5 IQRs abaixo do quartil inferior.
  - Observe: as cercas servem apenas para construir o boxplot e não devem aparecer na apresentação final.



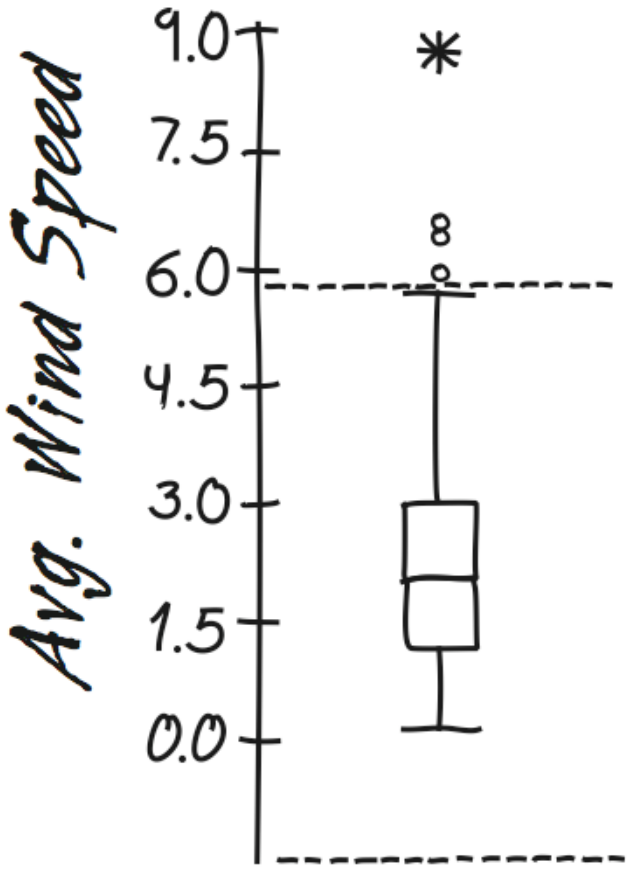
Construindo um boxplot (cont.)

- Use as cercas para desenhar as linhas verticais.
  - Desenhe linhas do fim da parte inferior da caixa até os valores mais extremos no interior das "cercas".
  - Se um dados cai fora da cerca, ele não será conectado pela linha vertical.



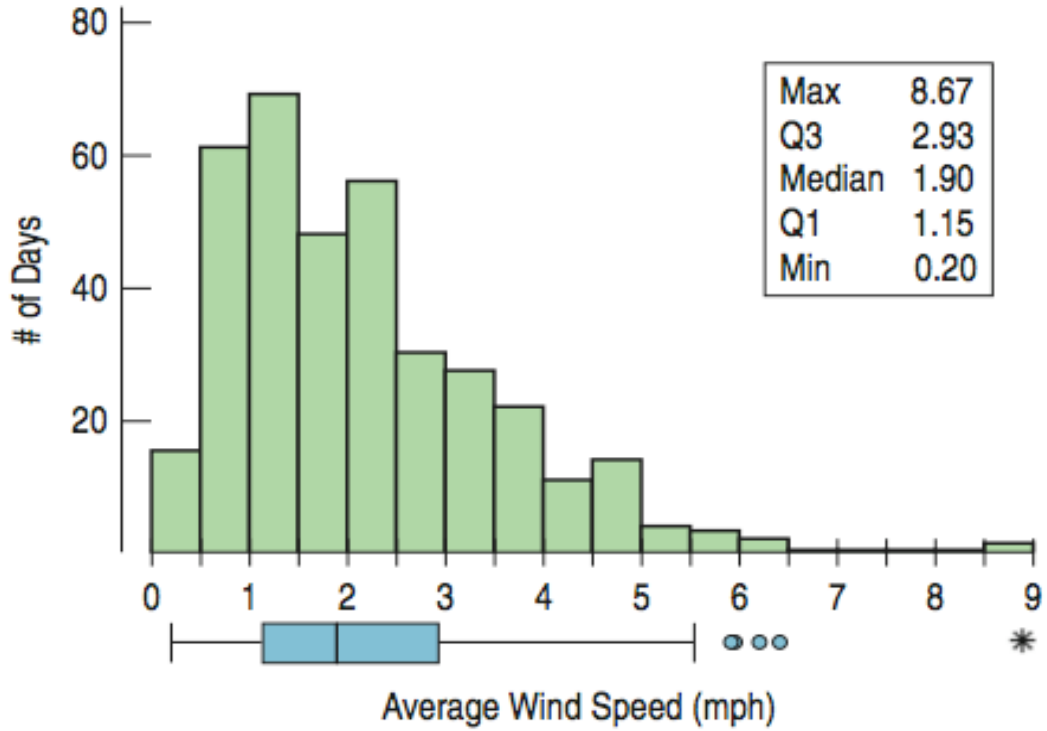
Construindo um boxplot (cont.)

- Assinale os outliers (qualquer valor que saia das cercas) com símbolos especiais.
  - Frequentemente utilizamos símbolos para outliers extremos que distam mais de 3 IQRs dos quartis.



Velocidade do vento: construindo boxplot (cont.)

- Compare o histograma e o boxplot da velocidade diária dos ventos:



Correlação

Propriedades da correlação

- O sinal de um coeficiente de correlação revela a direção da associação.
- Correlação é sempre entre -1 e +1.
- Correlação *pode* ser exatamente igual a -1 ou +1, mas esses valores são incomuns para dados reais, pois eles significam que todos os pontos de dados caem *exatamente* sobre uma linha reta.
- Uma correlação próxima de zero corresponde a uma fraca associação linear.

Propriedades da correlação

- Correlação mensura a força de uma relação linear entre duas variáveis.
- As variáveis podem ter uma forte associação, mas uma reduzida correlação se a associação não é linear.
- A correlação é sensitivel aos outliers. Um único valor outilier pode fazer uma reduzida correlação tornar- se forte, ou vice-versa.

Propriedades da correlação

- Outliers podem distorcer a correlação acentuadamente.
- Um outlier pode fazer uma pequena correlação parecer grande ou esconder uma forte correlação.
- Ele pode até mesmo transformar um coeficiente de uma associação positiva em negativa (e vice- versa).
- Quando encontrar um oulier é uma boa ideia apresentar as correlações com e sem aquele ponto.

Correlação de Pearson = r

- Correlação mensura a força de uma relação linear entre duas variáveis.

$$r = \frac{\sum \left[\frac{(x - \bar{x})}{s_x} \frac{(y - \bar{y})}{s_y} \right]}{n - 1}$$

Calculando o coeficiente de correlação (r)

Exemplo: PIB per Capita e expectativa de vida em países europeus selcionados

Country	Per Capita GDP (*x*)	Life Expectancy (*y*)
Austria	21.4	77.48
Bélgica	23.2	77.53
Finlandia	20.0	77.32
França	22.7	78.63
Alemanha	20.8	77.17
Irlanda	18.6	76.39
Italia	21.5	78.51
Holanda	22.0	78.15
Suíça	23.8	78.99
Reino Unido	21.2	77.37

Calculando o coeficiente de correlação (r)

Exemplo: PIB per Capita e expectativa de vida em países europeus selcionados

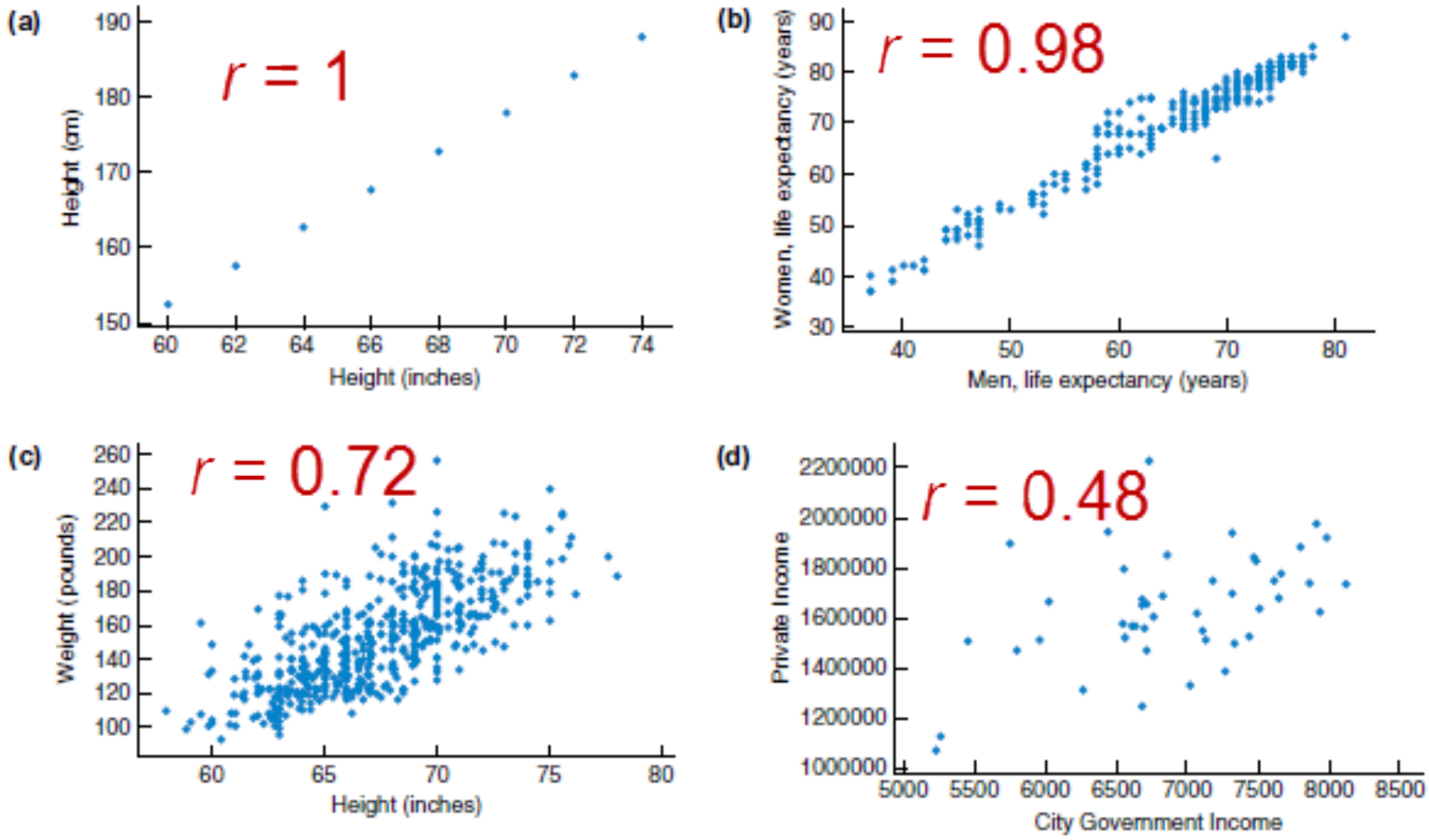
x	y	(a) xi-x/s	(b) yi –y/s	(a) X (b)
21.4	77.48	-0.078	-0.345	0.027
23.2	77.53	1.097	-0.282	-0.309
20.0	77.32	-0.992	-0.546	0.542
22.7	78.63	0.770	1.102	0.849
20.8	77.17	-0.470	-0.735	0.345
18.6	76.39	-1.906	-1.716	3.271
21.5	78.51	-0.013	0.951	-0.012
22.0	78.15	0.313	0.498	0.156
23.8	78.99	1.489	1.555	2.315
21.2	77.37	-0.209	-0.483	0.101
= 21.52	= 77.754			soma = 7.285
s_x =1.532	s_y =0.795			

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

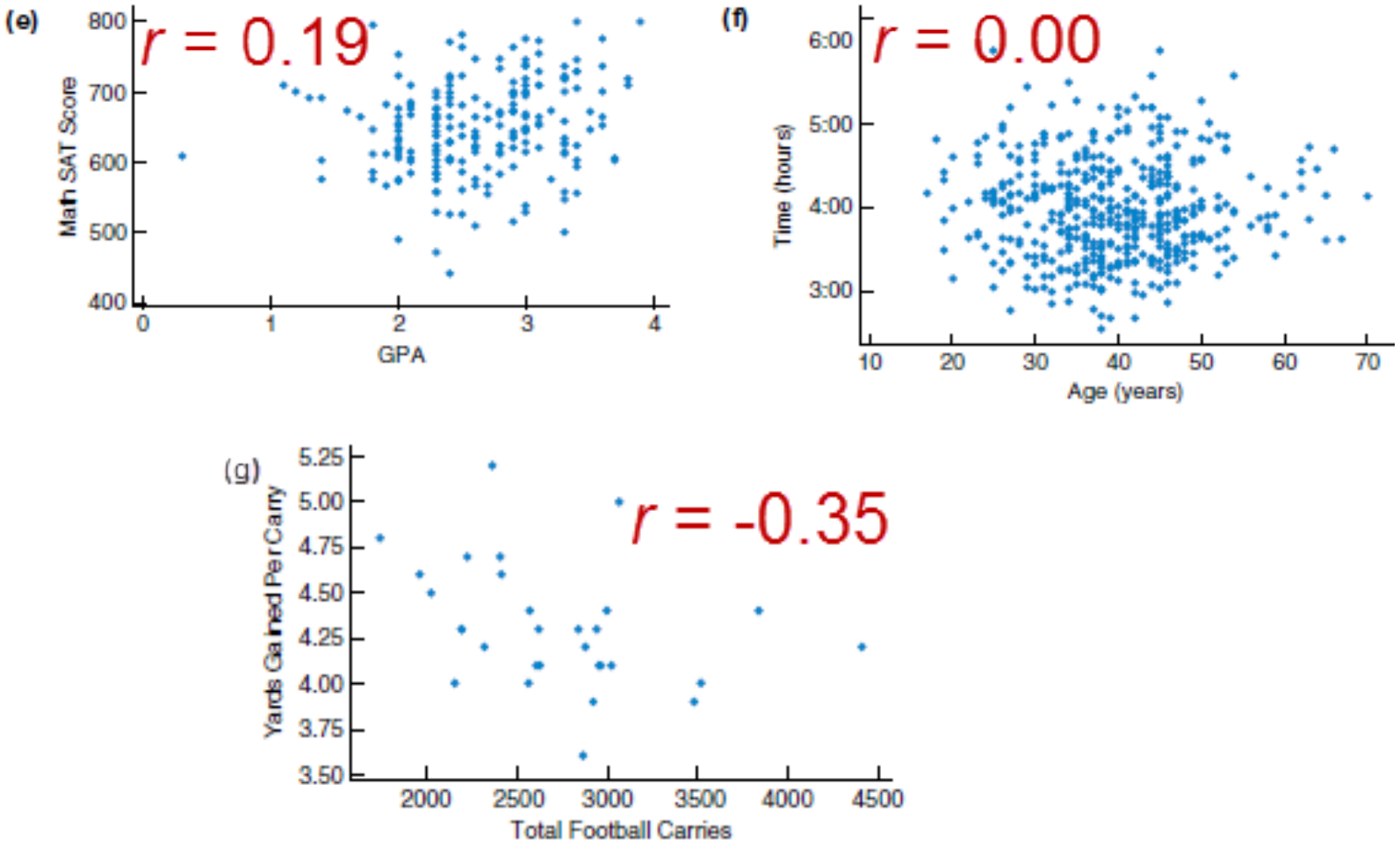
$$= \left(\frac{1}{10-1}\right)(7.285)$$

$$= 0.809$$

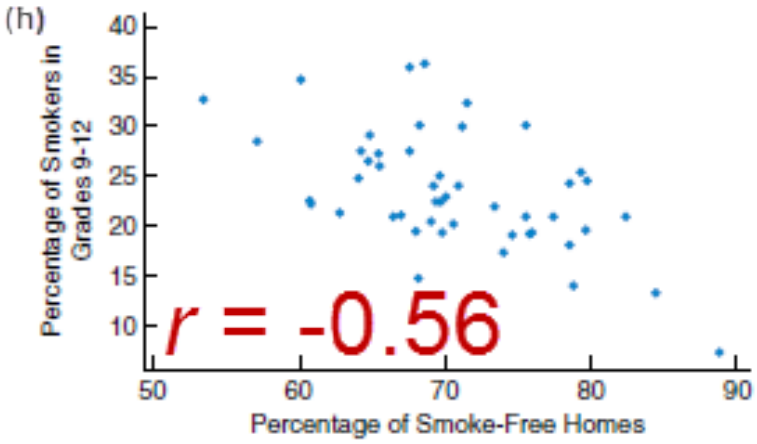
Correlação positiva

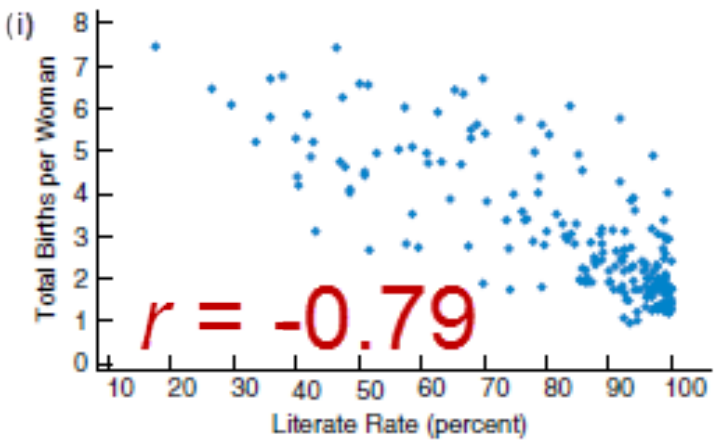


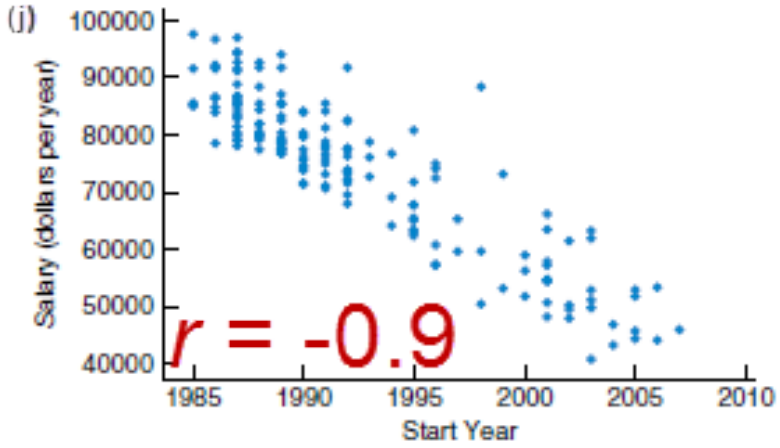
Correlação fraca ou ausente

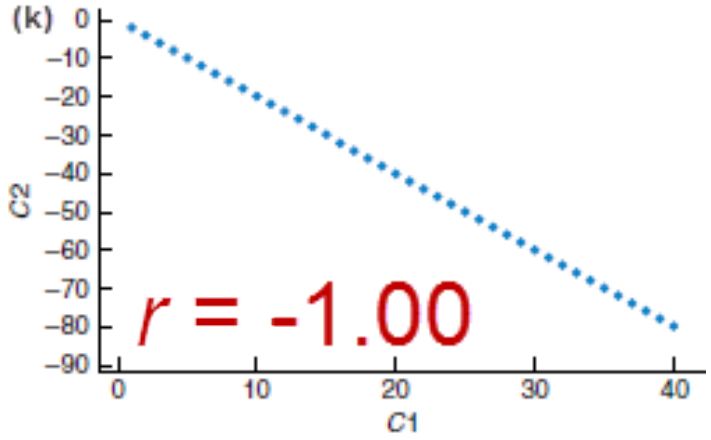


Correlação negativa

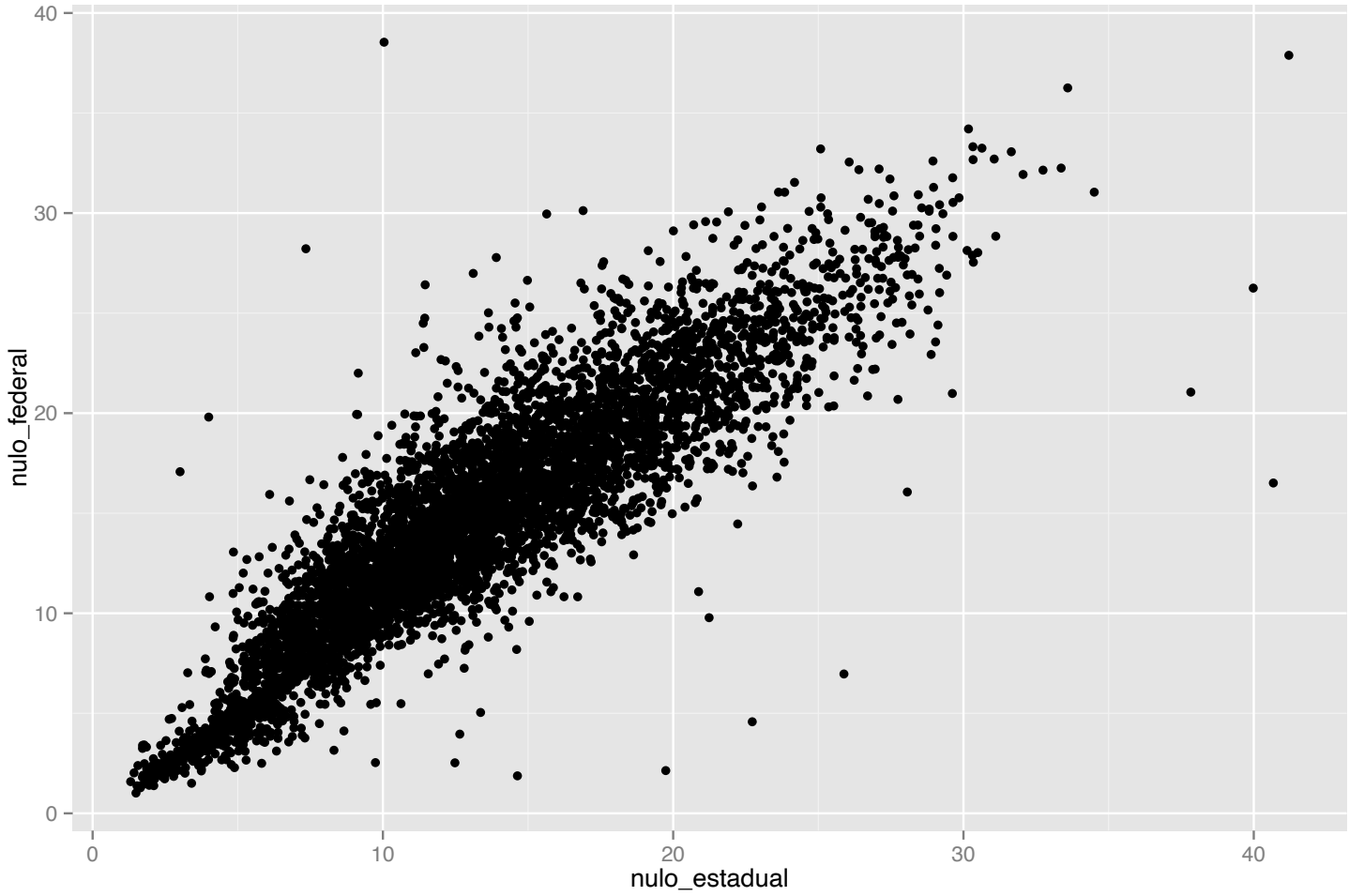


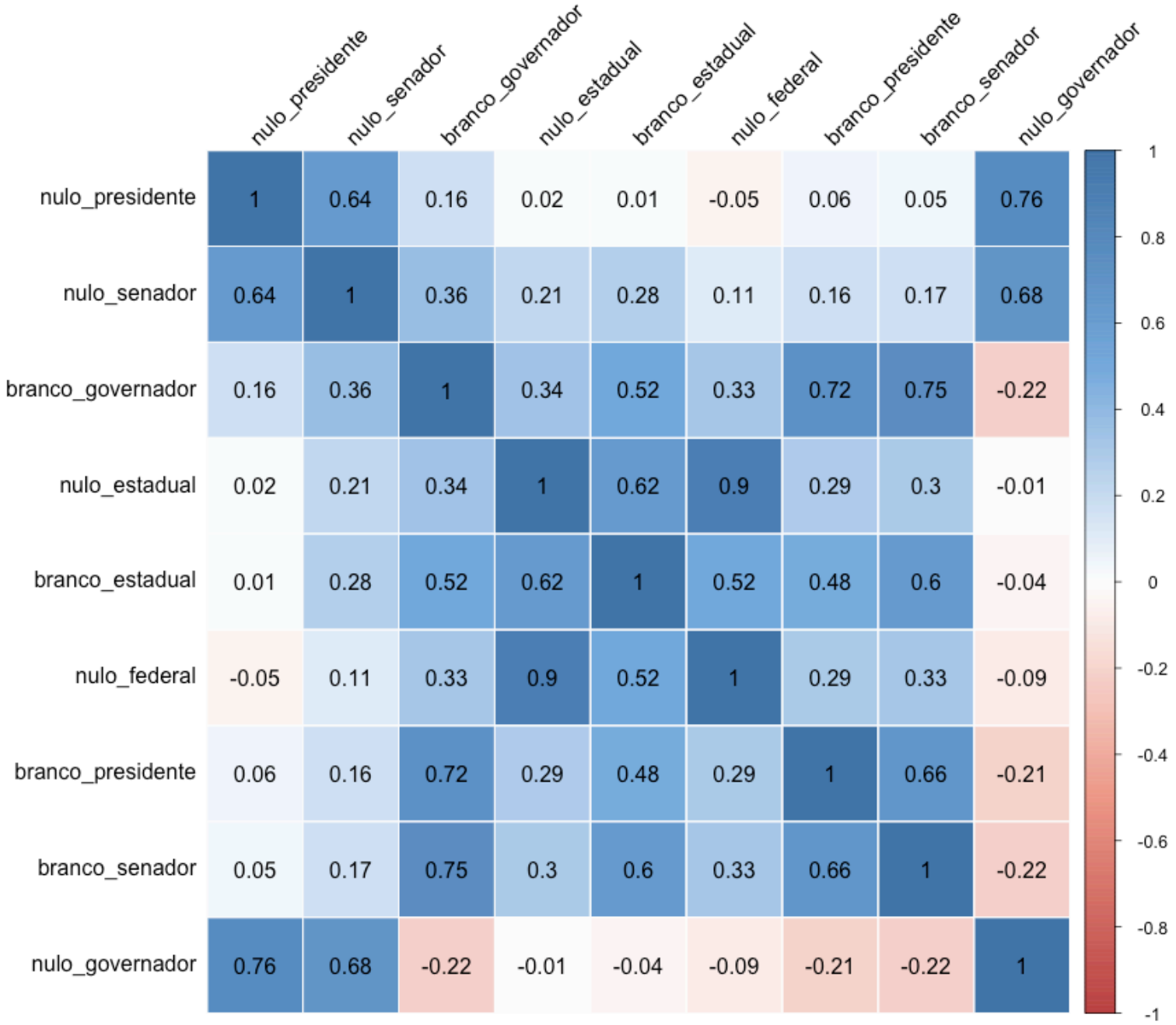






r = 0.90




nulo_presidente
nulo_senador
branco_governador
nulo_estadual
branco_estadual
nulo_federal
branco_presidente
branco_senador
nulo_governador
nulo_presidente 1 0.64 0.16 0.02 0.01 -0.05 0.06 0.05 0.76
nulo_senador 0.64 1 0.36 0.21 0.28 0.11 0.16 0.17 0.68
branco_governador 0.16 0.36 1 0.34 0.52 0.33 0.72 0.75 -0.22
nulo_estadual 0.02 0.21 0.34 1 0.62 0.9 0.29 0.3 -0.01
branco_estadual 0.01 0.28 0.52 0.62 1 0.52 0.48 0.6 -0.04
nulo_federal -0.05 0.11 0.33 0.9 0.52 1 0.29 0.33 -0.09
branco_presidente 0.06 0.16 0.72 0.29 0.48 0.29 1 0.66 -0.21
branco_senador 0.05 0.17 0.75 0.3 0.6 0.33 0.66 1 -0.22
nulo_governador 0.76 0.68 -0.22 -0.01 -0.04 -0.09 -0.21 -0.22 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1