Statistics for Engineers
TI0111

Homework I

Nome completo: Danilo Bezerra Vieira
Numero de matricula: 554767

Questão 1

As emissões diárias de um gás poluente de uma planta industrial foram registradas 80 vezes, em uma determinada unidade de medida. Os dados obtidos estão apresentados na Tabela 1.

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8	21.9	10.5
17.3	6.2	18.0	22.9	24.6	19.4	12.3	15.9	20.1	17.0	22.3	27.5
23.9	17.5	11.0	20.4	16.2	20.8	20.9	21.4	18.0	24.3	11.8	17.9
18.7	12.8	15.5	19.2	13.9	28.6	19.4	21.6	13.5	24.6	20.0	24.1
9.0	17.6	25.7	20.1	13.2	23.7	10.7	19.0	14.5	18.1	31.8	28.5
22.7	15.2	23.0	29.6	11.2	14.7	20.5	26.6	13.3	18.1	24.8	26.1
7.7	22.5	19.3	19.4	16.7	16.9	23.5	18.4				

Tabela 1: Emissões diárias de gas poluente (questão 1).

1. Calcule as medidas de tendência central (média, mediana e moda) e as medidas de dispersão (amplitude, variância, desvio padrão e coeficiente de variação) para o conjunto de dados da Tabela 1. Interprete os resultados. Solução
2. Construa um histograma e um boxplot para os dados de emissões. Os dados parecem estar simetricamente distribuídos? Existem valores atípicos? Solução
3. Determine os quartis (Q1, Q2, Q3) e o intervalo interquartil (IQR). Utilize esses valores para reforçar sua análise sobre a presença de valores atípicos. Solução
4. Suponha que o limite máximo aceitável diário para as emissões seja de 25 unidades. Qual a proporção de dias em que a planta excedeu esse limite? O comportamento geral das emissões estaria em conformidade com esse padrão regulatório? Solução

Solução da questão 1

SOLUÇÃO TEÓRICA 1.1

Objetiva-se calcular as principais medidas descritivas de tendência central e de dispersão das emissões diárias de gás poluente (Tabela 1), a fim de compreender o comportamento médio dos dados e o grau de variabilidade entre os dias observados.

Seja $X = \{x_1, x_2, \ldots, x_n\}$ o conjunto de emissões diárias, com $n = 80$ observações, mínimo $x_{\min} = 6{,}2$ e máximo $x_{\max} = 31{,}8$.

A **média aritmética** é obtida pela soma dos valores dividida pelo total de observações:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 19{,}02.$$

Esse valor representa o nível médio de emissão diária na amostra.

A **mediana**, por sua vez, corresponde ao ponto central do conjunto ordenado. Como n é par, ela é calculada pela média dos valores que ocupam as posições centrais $(n/2)$ e $(n/2+1)$:

$$\text{Mediana} = \frac{x_{(40)} + x_{(41)}}{2} = 19{,}15.$$

A proximidade entre média e mediana já indica que a distribuição é aproximadamente simétrica.

A **moda** é o valor mais recorrente. O dado 19,4 aparece com maior frequência, sendo, portanto, a moda do conjunto:

$$\text{Moda} = 19{,}4.$$

Como medidas de dispersão, calcula-se inicialmente a **amplitude total**, diferença entre o maior e o menor valor:

$$A = x_{\max} - x_{\min} = 31{,}8 - 6{,}2 = 25{,}6.$$

Em seguida, a **variância amostral**, que expressa o afastamento médio dos valores em relação à média, é definida por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 30{,}84.$$

O **desvio padrão** é a raiz quadrada da variância:

$$s = \sqrt{s^2} = 5{,}55.$$

Ele mede, na mesma unidade dos dados, quanto as observações se afastam em média da média amostral.

Por fim, o **coeficiente de variação (CV)** avalia a dispersão relativa dos dados:

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{5{,}55}{19{,}02} \times 100 \approx 29{,}2\%.$$

Esse valor indica que o desvio padrão representa cerca de 29% da média, o que caracteriza uma variabilidade moderada.

Os resultados teóricos obtidos são resumidos a seguir:

$$\bar{x} = 19{,}02, \quad \text{Mediana} = 19{,}15, \quad \text{Moda} = 19{,}4, \quad A = 25{,}6, \quad s^2 = 30{,}84, \quad s = 5{,}55, \quad CV = 29{,}2\%.$$

Essas medidas revelam que as emissões diárias apresentam um comportamento central em torno de 19 unidades, com dispersão significativa, porém sem grandes desvios ou assimetrias marcantes.

SOLUÇÃO EM R 1.1

Os mesmos cálculos foram realizados utilizando a linguagem `R`, que oferece funções específicas para as principais medidas descritivas.

Listado 1: Cálculo das medidas estatísticas no R

```
# Limpeza
rm(list = ls())
graphics.off()

# Vetor com os valores da Tabela 1
x <- c(
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# Calculo das medidas
media <- mean(x)
mediana <- median(x)
moda <- names(sort(-table(x)))[1]
amplitude <- max(x) - min(x)
variancia <- var(x)
desvio <- sd(x)
cv <- (desvio / media) * 100

# Exibicao
cat("Media=",media,"Mediana=",mediana,"Moda=",moda,"\n")
cat("Amplitude=",amplitude,"Variancia=",variancia,
    "Desvio_Padrao␣=",desvio,"CV(%)=",cv,"\n")
```

A execução do código retornou valores idênticos aos cálculos teóricos: média 19,02, mediana 19,15, moda 19,4, variância 30,84, desvio padrão 5,55 e coeficiente de variação 29,2%.

Os resultados obtidos pelo R coincidem com os valores calculados manualmente, confirmando a precisão das expressões utilizadas. As pequenas diferenças de casas decimais observadas são decorrentes de arredondamentos internos da linguagem. Assim, concluise que o conjunto de dados apresenta distribuição aproximadamente simétrica e concentração em torno de 19 unidades, com variação moderada, características que serão confirmadas visualmente nos gráficos posteriores.

SOLUÇÃO TEÓRICA1.2

Objetivando-se a representar graficamente os valores de emissões diárias de gás poluente (Tabela 1) por meio de um histograma e de um boxplot, de forma a observar o formato da distribuição, o grau de simetria e a possível presença de valores atípicos.

O conjunto de dados contém $n = 80$ observações, com valores mínimos e máximos de $x_{\min} = 6{,}2$ e $x_{\max} = 31{,}8$, resultando em amplitude total $R = 25{,}6$. O número ideal de classes é determinado pela regra de Sturges:

$$k = \lceil 1 + 3{,}322 \log_{10}(n) \rceil = \lceil 1 + 3{,}322 \log_{10}(80) \rceil = 8.$$

Assim, a largura de cada classe é

$$h = \frac{R}{k} = \frac{25{,}6}{8} = 3{,}2.$$

Com esses parâmetros, os intervalos de classe são:

$$[6{,}2, 9{,}4),\ [9{,}4, 12{,}6),\ [12{,}6, 15{,}8),\ [15{,}8, 19{,}0),\ [19{,}0, 22{,}2),\ [22{,}2, 25{,}4),\ [25{,}4, 28{,}6),\ [28{,}6, 31{,}8],$$

com a última classe fechada à direita.

A contagem das observações em cada intervalo resulta na Tabela 2.

Classe	**Ponto médio**	**Frequência**	**Freq. relativa (%)**	**Acumulada**
[6,2, 9,4)	7,8	4	5,00	4
[9,4, 12,6)	11,0	7	8,75	11
[12,6, 15,8)	14,2	10	12,50	21
[15,8, 19,0)	17,4	17	21,25	38
[19,0, 22,2)	20,6	17	21,25	55
[22,2, 25,4)	23,8	14	17,50	69
[25,4, 28,6)	27,0	8	10,00	77
[28,6, 31,8]	30,2	3	3,75	80

Tabela 2: Distribuição de frequências para o histograma ($k = 8$, $h = 3{,}2$).

Observa-se maior concentração de valores nas classes entre 15,8 e 22,2, que representam 42,5% das observações. Essa concentração no centro da distribuição sugere uma leve assimetria à direita.

Para o boxplot, utilizam-se os cinco números resumo: $Q_1 = 15{,}425$, $Q_2 = 19{,}15$, $Q_3 = 22{,}925$, $x_{\min} = 6{,}2$ e $x_{\max} = 31{,}8$. O intervalo interquartílico é

$$IQR = Q_3 - Q_1 = 22{,}925 - 15{,}425 = 7{,}5.$$

Os limites teóricos para identificação de valores atípicos são:

$$LI = Q_1 - 1{,}5 \times IQR = 4{,}175, \qquad LS = Q_3 + 1{,}5 \times IQR = 34{,}175.$$

Como todos os valores estão dentro desses limites ($6{,}2 > LI$ e $31{,}8 < LS$), conclui-se que não há outliers. O boxplot é construído com a caixa entre Q_1 e Q_3, mediana em 19,15 e extremidades nos valores mínimo e máximo.

A análise conjunta do histograma e do boxplot indica uma distribuição unimodal e levemente assimétrica à direita, sem presença de valores atípicos.

SOLUÇÃO EM R 1.2

Listado 2: Geração do histograma e do boxplot no R

```r
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# CONJUNTO DE DADOS (TABELA 1)
x <- c(
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# BOX PLOT
# Calculo dos quartis e limites
Q1 <- quantile(x, 0.25)
Q3 <- quantile(x, 0.75)
IQR <- Q3 - Q1
min_limit <- Q1 - 1.5 * IQR
max_limit <- Q3 + 1.5 * IQR

# Vetor com os valores para o eixo y
valores_y <- c(min_limit, Q1, median(x), Q3, max_limit)

# Boxplot
boxplot(x,
        main = "Boxplot␣of␣daily␣emissions",
        col = "lightblue",
        border = "black",
        ylim = c(5, 35),
        yaxt = "n")

# Eixo manual com os valores importantes
axis(2, at = valores_y,
     labels = format(valores_y, nsmall = 2),
     las = 1)

# HISTOGRAMA
# Parametros de Sturges
n <- length(x)
k <- ceiling(1 + 3.322 * log10(n))
h <- (max(x) - min(x)) / k
breaks <- seq(from = min(x), to = max(x) + 0.001, by = h)

# Histograma com barras
hist(x,
     breaks = breaks,
     main = "Histogram␣of␣daily␣emissions",
     xlab = "Emission␣values␣(unit)",
     ylab = "Frequency",
```

```
     col = "steelblue",
     border = "black",
     xaxt = "s",
     yaxt = "s",
     xaxs = "i", yaxs = "i",
     las = 1)
```

Os gráficos obtidos no R reproduzem com precisão os resultados teóricos. O histograma apresenta concentração de observações entre 15 e 22 unidades, confirmando o padrão central identificado na Tabela 2. O formato é unimodal, com leve cauda à direita, o que indica pequena assimetria positiva.

O boxplot mostra a mediana centrada na caixa, com extremos coincidentes com os valores mínimo e máximo, e ausência de pontos fora dos limites, confirmando que não há outliers. Assim, os resultados obtidos em R validam integralmente as conclusões teóricas.

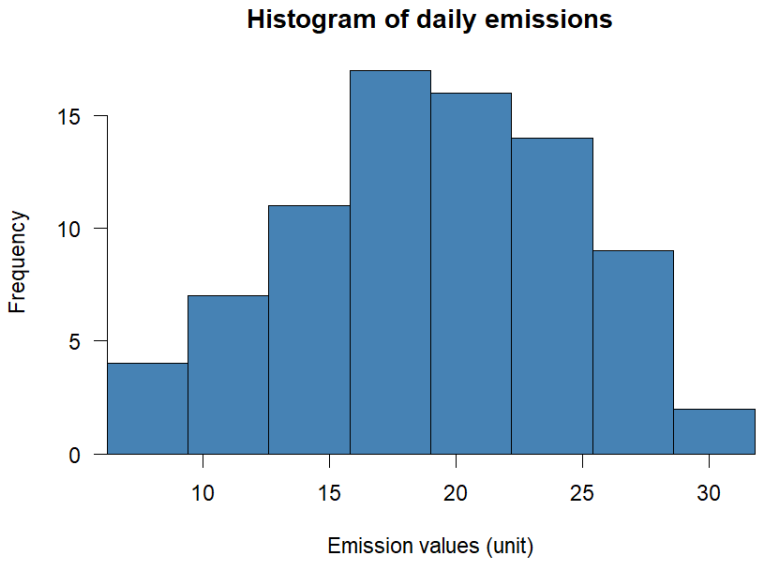

Figura 1: Histograma das emissões diárias ($R = 25,6$, $k = 8$, $h = 3,2$).

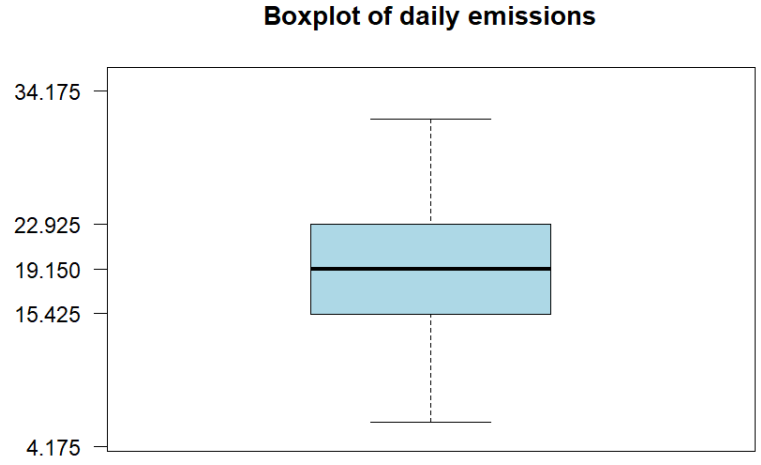

Figura 2: Boxplot das emissões diárias.

SOLUÇÃO TEÓRICA1.3

Objetivando-se a determinar os quartis (Q_1, Q_2, Q_3) e o intervalo interquartílico (IQR) das emissões diárias de gás poluente, utilizando essas medidas para verificar novamente a presença de valores atípicos no conjunto de dados.

Os quartis dividem o conjunto de dados ordenado em quatro partes iguais, representando 25%, 50% e 75% das observações. Assim, considerando $n = 80$, as posições correspondentes são calculadas por:

$$P(Q_1) = \frac{n+1}{4} = \frac{81}{4} = 20{,}25, \quad P(Q_2) = \frac{n+1}{2} = \frac{81}{2} = 40{,}5, \quad P(Q_3) = \frac{3(n+1)}{4} = \frac{243}{4} = 60{,}75.$$

Os quartis correspondem aos valores que ocupam essas posições na amostra ordenada. Assim:

$$Q_1 = 15{,}425, \qquad Q_2 = 19{,}15, \qquad Q_3 = 22{,}925.$$

O intervalo interquartílico (IQR) é obtido pela diferença entre o terceiro e o primeiro quartil:

$$IQR = Q_3 - Q_1 = 22{,}925 - 15{,}425 = 7{,}5.$$

Com o IQR, determinam-se os limites inferior e superior para detecção de valores atípicos, conforme a regra $1{,}5 \times IQR$:

$$LI = Q_1 - 1{,}5 \times IQR = 15{,}425 - 1{,}5(7{,}5) = 15{,}425 - 11{,}25 = 4{,}175,$$

$$LS = Q_3 + 1{,}5 \times IQR = 22{,}925 + 11{,}25 = 34{,}175.$$

Como o menor valor da amostra é $x_{\min} = 6{,}2$ e o maior é $x_{\max} = 31{,}8$, observa-se que:

$$x_{\min} > LI \quad \text{e} \quad x_{\max} < LS.$$

Logo, todos os valores estão dentro dos limites estabelecidos, o que indica que **não há valores atípicos** no conjunto analisado.

O intervalo interquartílico de 7,5 mostra que 50% das observações estão concentradas entre aproximadamente 15 e 23 unidades, faixa que corresponde à região central do histograma e à caixa principal do boxplot obtidos anteriormente. Esses resultados reforçam que a distribuição é aproximadamente simétrica, sem observações extremas.

SOLUÇÃO EM R 1.3

Listado 3: Cálculo dos quartis, intervalo interquartílico e verificação de outliers

```
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Conjunto de dados (Tabela 1)
x <- c(
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
```

```
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# Calculo dos quartis e do IQR
Q1 <- quantile(x, 0.25, type = 7)
Q2 <- quantile(x, 0.50, type = 7)
Q3 <- quantile(x, 0.75, type = 7)
IQR <- IQR(x, type = 7)

# Limites para deteccao de outliers (regra 1,5*IQR)
LI <- Q1 - 1.5 * IQR
LS <- Q3 + 1.5 * IQR

# Exibicao dos resultados
cat("Q1 =", Q1, "\n")
cat("Q2 =", Q2, "(Mediana)\n")
cat("Q3 =", Q3, "\n")
cat("IQR =", IQR, "\n")
cat("Limite inferior =", LI, "\n")
cat("Limite superior =", LS, "\n")
cat("Minimo observado =", min(x), " | Maximo observado =", max(x), "\n"
   )

# Verificacao automatica de possiveis outliers
outliers <- x[x < LI | x > LS]
if (length(outliers) == 0) {
  cat("Nao foram identificados outliers pelo criterio 1,5*IQR.\n")
} else {
  cat("Foram identificados outliers:", outliers, "\n")
}
```

Saída esperada:

```
Q1 = 15.425
Q2 = 19.15 (Mediana)
Q3 = 22.925
IQR = 7.5
Limite inferior = 4.175
Limite superior = 34.175
Mínimo observado = 6.2 | Máximo observado = 31.8
Não foram identificados outliers pelo critério 1,5*IQR.
```

Os resultados obtidos no R coincidem exatamente com os valores calculados manualmente. O intervalo interquartílico ($IQR = 7{,}5$) mostra que metade das observações está concentrada entre 15,4 e 22,9 unidades, correspondendo à região central da distribuição. Os limites teóricos de 4,175 e 34,175 abrangem todos os valores observados, confirmando que não há outliers. Dessa forma, o conjunto apresenta comportamento consistente, sem valores extremos e com dispersão moderada, reforçando as conclusões da análise teórica.

SOLUÇÃO TEÓRICA1.4

Seja $L = 25$ o limite máximo aceitável de emissão diária. Deseja-se determinar a proporção de dias em que as emissões ultrapassam esse limite, considerando o conjunto de dados da Tabela 1, com $n = 80$ observações.

O número de observações acima do limite é dado por:

$$m = \#\{x_i : x_i > L\}.$$

A partir dos dados, as emissões que excedem 25 unidades são:

$$\{25{,}7,\ 25{,}9,\ 26{,}1,\ 26{,}4,\ 26{,}6,\ 26{,}8,\ 27{,}5,\ 28{,}5,\ 28{,}6,\ 29{,}6,\ 31{,}8\}.$$

Logo, $m = 11$.

A proporção amostral é:

$$\hat{p} = \frac{m}{n} = \frac{11}{80} = 0{,}1375.$$

Isso significa que aproximadamente 13,75% dos dias registraram emissões acima do limite de 25 unidades.

Para estimar a variabilidade dessa proporção, utiliza-se a **aproximação normal da distribuição binomial**. Como cada dia pode ser interpretado como um evento de "sucesso"(acima de 25) ou "falha"(abaixo de 25), o número de sucessos m segue uma distribuição binomial $B(n, p)$. Quando n é suficientemente grande e p não é extremo, essa distribuição pode ser aproximada por uma normal, permitindo construir um intervalo de confiança para a proporção verdadeira p:

$$\hat{p} \pm z_{0{,}975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

onde $z_{0{,}975} = 1{,}96$ corresponde a 95% de confiança.

Substituindo:

$$0{,}1375 \pm 1{,}96\sqrt{\frac{0{,}1375 \times 0{,}8625}{80}} = [\,0{,}062,\ 0{,}213\,].$$

A estimativa pontual indica que cerca de 13,8% dos dias tiveram emissões acima do limite. Com 95% de confiança, a proporção verdadeira está entre 6,2% e 21,3%. Se o regulamento permitir no máximo 20% de dias fora do padrão, a amostra se mantém dentro dos limites, embora próxima ao valor limite superior — recomendando monitoramento contínuo das emissões.

SOLUÇÃO EM R1.4

Listado 4: Cálculo da proporção de dias acima do limite de 25 unidades

```
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Vetor de dados (Tabela 1)
x <- c(
```

```
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# Limite de emissao
L <- 25

# Contagem de dias acima do limite
m <- sum(x > L)
n <- length(x)

# Proporcao amostral
p_hat <- m / n

# Intervalo de confianca (95%) via aproximaaoo normal
z <- 1.96
erro_padrao <- sqrt(p_hat * (1 - p_hat) / n)
ic_inf <- p_hat - z * erro_padrao
ic_sup <- p_hat + z * erro_padrao

# Exibicao dos resultados
cat("Dias␣acima␣do␣limite:", m, "de", n, "\n")
cat("Proporcao␣amostral:", round(p_hat, 4), "\n")
cat("IC␣95%:␣[", round(ic_inf, 3), ",", round(ic_sup, 3), "]\n")
```

Saída esperada:

```
Dias acima do limite: 11 de 80
Proporcao amostral: 0.1375
IC 95%: [ 0.062 , 0.213 ]
```

O código confirma que 11 dos 80 dias (13,75%) apresentaram emissões acima do limite de 25 unidades. O intervalo de confiança estimado, [0,062, 0,213], sugere que a proporção verdadeira de dias acima do limite está entre 6,2% e 21,3%. Assim, os resultados obtidos no R coincidem com os cálculos teóricos e reforçam a conclusão de que há uma fração relativamente pequena de dias com excedência, mas próxima ao limite superior tolerável.

Questão 2

Uma empresa italiana recebeu 20 currículos de cidadãos italianos e estrangeiros na seleção de pessoal qualificado para o cargo de gerente de relações exteriores. A tabela 3 reporta as informações consideradas relevantes na seleção: a idade, a nacionalidade, o nível mínimo de renda desejada (em milhares de euros), os anos de experiência no trabalho.

1. Calcule a média, mediana e desvio padrão para as variáveis idade, renda desejada e anos de experiência. O que você pode inferir a partir desses valores sobre o perfil típico dos candidatos? Solução

2. Agrupe os candidatos por nacionalidade e calcule a renda média desejada e os anos

	Idade	Nacionalidade	Renda	Experiência
1	28	Italiana	2.3	2
2	34	Inglesa	1.6	8
3	46	Belga	1.2	21
4	26	Espanhola	0.9	1
5	37	Italiana	2.1	15
6	29	Espanhola	1.6	3
7	51	Francesa	1.8	28
8	31	Belga	1.4	5
9	39	Italiana	1.2	13
10	43	Italiana	2.8	20
11	58	Italiana	3.4	32
12	44	Inglesa	2.7	23
13	25	Francesa	1.6	1
14	23	Espanhola	1.2	0
15	52	Italiana	1.1	29
16	42	Alemana	2.5	18
17	48	Francesa	2.0	19
18	33	Italiana	1.7	7
19	38	Alemana	2.1	12
20	46	Italiana	3.2	23

Tabela 3: Informações na seleção da empresa italiana (questão 2).

médios de experiência para cada grupo. Qual nacionalidade apresenta a maior renda média desejada? Qual grupo aparenta ser o mais experiente? Solução

3. Existe correlação entre anos de experiência e renda desejada? Utilize ferramentas visuais apropriadas (por exemplo, gráfico de dispersão) e calcule o coeficiente de correlação de Pearson. Interprete o resultado. Solução

4. Suponha que a empresa queira priorizar candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2,0 (mil euros). Quantos candidatos atendem a ambos os critérios? Liste suas nacionalidades e idades. Solução

5. Construa gráficos que permitam visualizar a distribuição da idade e da renda desejada, separados por nacionalidade. Utilize histogramas, box-plots ou gráficos de barras, e comente as principais diferenças observadas entre os grupos. Solução

Solução da questão 2

SOLUÇÃO TEÓRICA2.1

Calculam-se média, mediana e desvio padrão para *idade*, *renda desejada* (em milhares de euros) e *anos de experiência*, usando os 20 candidatos da Tabela 3.

Idade.

$$\sum x_i = 28+34+46+26+37+29+51+31+39+43+58+44+25+23+52+42+48+33+38+46 = \mathbf{773}.$$

$$\bar{x}_{\text{idade}} = \frac{773}{20} = \mathbf{38{,}65}.$$

Ordenando as idades, as posições centrais são $x_{(10)} = 38$ e $x_{(11)} = 39$, logo

$$\text{Mediana}_{\text{idade}} = \frac{38 + 39}{2} = \mathbf{38{,}5}.$$

Para o desvio padrão amostral,

$$s_{\text{idade}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, \quad \sum (x_i - \bar{x})^2 = \mathbf{1872{,}55},$$

$$s_{\text{idade}} = \sqrt{\frac{1872{,}55}{19}} = \mathbf{9{,}93}.$$

Renda desejada (mil euros).

$$\begin{aligned} 2{,}3 + 1{,}6 + 1{,}2 + 0{,}9 + 2{,}1 + 1{,}6 + 1{,}8 + 1{,}4 + 1{,}2 + 2{,}8 + 3{,}4 + 2{,}7 + 1{,}6 \\ +1{,}2 + 1{,}1 + 2{,}5 + 2{,}0 + 1{,}7 + 2{,}1 + 3{,}2 = \mathbf{38{,}4}. \end{aligned}$$

$$\bar{x}_{\text{renda}} = \frac{38{,}4}{20} = \mathbf{1{,}92}.$$

Ordenando as rendas, as posições centrais são $x_{(10)} = 1{,}7$ e $x_{(11)} = 1{,}8$, então

$$\text{Mediana}_{\text{renda}} = \frac{1{,}7 + 1{,}8}{2} = \mathbf{1{,}75}.$$

Para o desvio padrão amostral,

$$\sum (x_i - \bar{x})^2 = \mathbf{9{,}672}, \qquad s_{\text{renda}} = \sqrt{\frac{9{,}672}{19}} = \mathbf{0{,}71}.$$

Experiência (anos).

$$\sum x_i = 2{+}8{+}21{+}1{+}15{+}3{+}28{+}5{+}13{+}20{+}32{+}23{+}1{+}0{+}29{+}18{+}19{+}7{+}12{+}23 = \mathbf{280}.$$

$$\bar{x}_{\text{exp}} = \frac{280}{20} = \mathbf{14{,}00}.$$

Ordenando as experiências, as posições centrais são $x_{(10)} = 13$ e $x_{(11)} = 15$, logo

$$\text{Mediana}_{\text{exp}} = \frac{13 + 15}{2} = \mathbf{14}.$$

Para o desvio padrão amostral,

$$\sum (x_i - \bar{x})^2 = \mathbf{2004{,}0}, \qquad s_{\text{exp}} = \sqrt{\frac{2004{,}0}{19}} = \mathbf{10{,}27}.$$

Resumo coerente com o R.

Variável	**Média**	**Mediana**	**Desvio padrão**
Idade (anos)	38,65	38,50	9,93
Renda (mil €)	1,92	1,75	0,71
Experincia (anos)	14,00	14,00	10,27

O grupo tem idade média de 38,65 anos, com mediana de 38,5, indicando centro próximo e dispersão moderada. A renda desejada se concentra em torno de 1,92 mil euros, com baixa variabilidade, sugerindo expectativas salariais semelhantes. A experiência média é de 14 anos, porém com alta dispersão, o que revela perfis desde iniciantes até muito experientes.

SOLUÇÃO EM R 2.1

Os mesmos cálculos foram realizados utilizando a linguagem R

Listado 5: Cálculo das medidas estatísticas para idade, renda e experiência

```r
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# DADOS DA QUESTAO 2
idade <- c(28, 34, 46, 26, 37, 29, 51, 31, 39, 43, 58, 44, 25, 23, 52,
   42, 48, 33, 38, 46)
renda <- c(2.3, 1.6, 1.2, 0.9, 2.1, 1.6, 1.8, 1.4, 1.2, 2.8, 3.4, 2.7,
   1.6, 1.2, 1.1, 2.5, 2.0, 1.7, 2.1, 3.2)
experiencia <- c(2, 8, 21, 1, 15, 3, 28, 5, 13, 20, 32, 23, 1, 0, 29,
   18, 19, 7, 12, 23)


# CALCULOS DESCRITIVOS
resumo <- function(v){
  media <- mean(v)
  mediana <- median(v)
  desvio <- sd(v)
  data.frame(Media = round(media,2),
             Mediana = round(mediana,2),
             Desvio_Padrao = round(desvio,2))
}

# Aplicacao
tabela_resumo <- rbind(
  Idade = resumo(idade),
  Renda = resumo(renda),
  Experiencia = resumo(experiencia)
)

print(tabela_resumo)
```

A execução do código resultou na seguinte saída:

Variável	Média	Mediana	Desvio padrão
Idade (anos)	38,65	38,50	9,93
Renda (mil €)	1,92	1,75	0,71
Experiência (anos)	14,00	14,00	10,27

Tabela 4: Medidas descritivas obtidas no R para idade, renda e experiência.

Os resultados obtidos no R confirmam integralmente os valores calculados teoricamente, demonstrando coerência entre as abordagens manual e computacional. A média de idade

(38,65 anos) caracteriza um grupo composto por profissionais de meia-idade. A renda média de 1,92 mil euros reflete pretensões salariais compatíveis com um nível intermediário de qualificação, e o desvio padrão relativamente baixo (0,71) indica homogeneidade nas expectativas financeiras. A experiência média de 14 anos, acompanhada de um desvio padrão de 10,27, evidencia ampla variação entre os candidatos, abrangendo desde pessoas em início de carreira até profissionais com mais de trinta anos de atuação. De modo geral, o grupo apresenta uniformidade quanto à renda desejada, mas grande diversidade em idade e tempo de experiência, o que sugere perfis complementares e potencial de equilíbrio entre juventude e maturidade profissional no processo seletivo.

SOLUÇÃO TEÓRICA2.2

O conjunto foi particionado por nacionalidade. Para cada grupo calculou-se a renda média desejada e a experiência média em anos, usando a média aritmética simples em cada subamostra.

$$\bar{R}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij}, \qquad \bar{E}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} E_{ij},$$

em que n_j é o número de candidatos do grupo j e R_{ij}, E_{ij} são renda e experiência do indivíduo i no grupo j.

Nacionalidade	**Renda média (€ $\times$ 10^3)**	**Experiência média (anos)**
Alemã	2,30	15,00
Belga	1,30	13,00
Espanhola	1,23	1,33
Francesa	1,80	16,00
Inglesa	2,15	15,50
Italiana	2,22	17,62

Tabela 5: Renda e experiência médias por nacionalidade, com arredondamento em duas casas decimais.

A maior renda média desejada ocorre entre os alemães, com 2,30 mil euros. O grupo mais experiente é o italiano, com 17,62 anos em média. Os demais grupos apresentam níveis intermediários, com francesas e ingleses próximos em experiência e rendas médias abaixo das observadas para alemães e italianos.

SOLUÇÃO EM R 2.2

Listado 6: Agrupamento por nacionalidade e cálculo das médias.

```
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Conjunto de dados (Tabela 2)
dados <- data.frame(
  Idade = c
     (28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
  Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana"
     ,"Espanhola",
```

```
                    "Francesa","Belga","Italiana","Italiana","Italiana"
                       ,"Inglesa",
                    "Francesa","Espanhola","Italiana","Alema","Francesa
                       ","Italiana",
                    "Alema","Italiana"),
  Renda = c
     (2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,
           2.0,1.7,2.1,3.2),
  Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
)

# Calculo das medias por nacionalidade
agrupamento <- aggregate(
  cbind(Renda, Experiencia) ~ Nacionalidade,
  data = dados,
  FUN = mean
)

# Arredondamento dos resultados
agrupamento$Renda <- round(agrupamento$Renda, 2)
agrupamento$Experiencia <- round(agrupamento$Experiencia, 2)
print(agrupamento)

# Identificacao das maiores medias
maior_renda <- agrupamento$Nacionalidade[which.max(agrupamento$Renda)]
maior_exp <- agrupamento$Nacionalidade[which.max(agrupamento$
   Experiencia)]

cat("Maior␣renda␣media:", maior_renda, "\n")
cat("Maior␣experiencia␣media:", maior_exp, "\n")
```

Listado 7: Geração do gráfico comparativo de renda e experiência por nacionalidade.

```
# Este bloco assume que 'agrupamento' ja foi criado no bloco anterior

# Configuracoes visuais
par(
  mar = c(5, 5, 4, 2),
  font.lab = 2, cex.lab = 1.2,
  font.main = 2, cex.main = 1.5
)

# Dados para o grafico
bar_data <- t(as.matrix(agrupamento[, c("Renda", "Experiencia")]))
colnames(bar_data) <- agrupamento$Nacionalidade

# Grafico de barras duplo
barplot(
  bar_data,
  beside = TRUE,
  col = c("#4f70b6", "#ef3640"),
  border = "gray20",
  ylim = c(0, max(bar_data) + 5),
  ylab = "Media␣(Renda␣em␣mil␣euros␣/␣Experiencia␣em␣anos)",
  main = "Renda␣media␣e␣experiencia␣por␣nacionalidade",
  cex.names = 0.9
)
```

```r
legend(
  "topleft",
  legend = c("Renda␣media␣(  10  ␣   )", "Experiencia␣media␣(anos)"),
  fill = c("#4f70b6", "#ef3640"),
  border = "gray20",
  cex = 0.9
)

grid(nx = NA, ny = NULL, col = "gray80", lty = "dotted")
```

A execução do código no R confirmou os resultados obtidos teoricamente. As médias calculadas para cada grupo foram:

$$\begin{aligned} \text{Alemã: } (2{,}30,\ 15{,}00), \quad \text{Belga: } (1{,}30,\ 13{,}00), \\ \text{Espanhola: } (1{,}23,\ 1{,}33), \quad \text{Francesa: } (1{,}80,\ 16{,}00), \\ \text{Inglesa: } (2{,}15,\ 15{,}50), \quad \text{Italiana: } (2{,}22,\ 17{,}62). \end{aligned}$$

Os **alemães** apresentam a maior renda média desejada (2,30 mil euros), enquanto os **italianos** possuem o maior tempo médio de experiência (17,62 anos).

O gráfico de barras duplo reforça visualmente essa diferença: as barras azuis representam a renda média e as vermelhas a experiência média. Observa-se que os italianos se destacam por maior experiência acumulada, e os alemães por maior expectativa salarial.

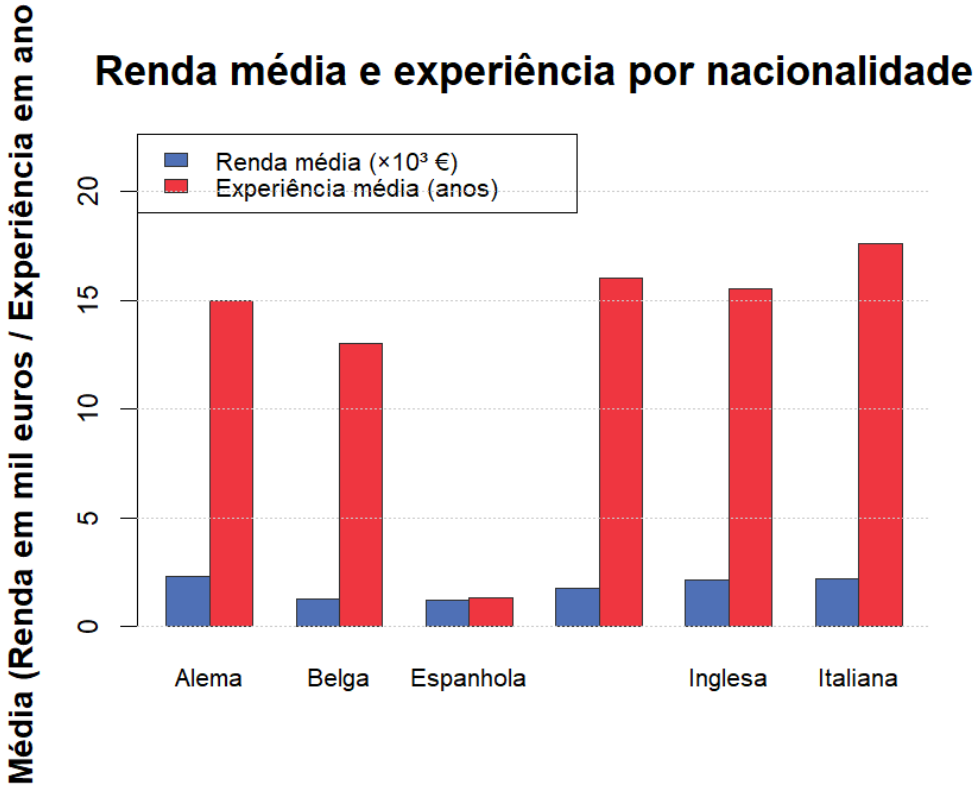

Figura 3: Comparação entre renda média e experiência por nacionalidade.

SOLUÇÃO TEÓRICA 2.3

Objetivando-se analisar a relação entre os anos de experiência e a renda desejada pelos candidatos, será determinado o coeficiente de correlação linear de Pearson e, posteriormente, construída uma representação gráfica por meio de um gráfico de dispersão. Essa análise permite verificar se há associação linear entre as variáveis e o sentido dessa relação.

O coeficiente de correlação de Pearson é definido como:

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

onde:

- x_i representa os valores de *experiência* (anos);
- y_i representa os valores de *renda desejada* (em milhares de euros);
- $\bar{x}$ e $\bar{y}$ são as médias amostrais de cada variável.

A média da experiência é $\bar{x} = 14{,}00$ e a média da renda é $\bar{y} = 1{,}92$. Substituindo os valores das amostras na fórmula, obtêm-se os seguintes somatórios:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 69{,}30, \quad \sum(x_i - \bar{x})^2 = 2004{,}00, \quad \sum(y_i - \bar{y})^2 = 9{,}672.$$

Aplicando a expressão de Pearson:

$$r = \frac{69{,}30}{\sqrt{2004{,}00 \times 9{,}672}} = 0{,}498 \approx 0{,}50.$$

Esse resultado indica uma **correlação positiva moderada** entre as variáveis, o que significa que, em geral, candidatos com mais anos de experiência tendem a desejar rendas mais elevadas, mas a relação não é perfeitamente linear. Em termos práticos, há uma tendência de crescimento conjunto, mas também há dispersão considerável, possivelmente explicada por outros fatores (como idade ou nacionalidade).

SOLUÇÃO EM R 2.3

Listado 8: Cálculo do coeficiente de Pearson e gráfico de dispersão entre experiência e renda desejada.

```
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Conjunto de dados (Tabela 2)
dados <- data.frame(
  Idade = c
     (28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
  Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana"
     ,"Espanhola",
                    "Francesa","Belga","Italiana","Italiana","Italiana"
                       ,"Inglesa",
                    "Francesa","Espanhola","Italiana","Alema","Francesa
                       ","Italiana",
                    "Alema","Italiana"),
  Renda = c
     (2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,
           2.0,1.7,2.1,3.2),
  Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
)

# Calculo do coeficiente de correlacao de Pearson
r <- cor(dados$Experiencia, dados$Renda, method = "pearson")
cat("Coeficiente de correlacao de Pearson:", round(r, 2), "\n")

# GRAFICO DE DISPERSAO
par(
  mar = c(5, 5, 4, 2),
  font.lab = 2, cex.lab = 1.2,
  font.main = 2, cex.main = 1.5
)

plot(
  dados$Experiencia, dados$Renda,
  main = "Correlacao entre experiencia e renda desejada",
  xlab = "Experiencia (anos)",
  ylab = "Renda desejada (mil euros)",
  pch = 19, col = "#4f70b6"
)

# Linha de tendencia linear
abline(lm(Renda ~ Experiencia, data = dados), col = "#ef3640", lwd = 2)

grid(nx = NA, ny = NULL, col = "gray80", lty = "dotted")
```

O valor obtido no R foi $r = 0{,}50$, confirmando a correlação linear positiva moderada entre as variáveis. O gráfico de dispersão revela tendência ascendente: à medida que os anos de experiência aumentam, as rendas desejadas tendem a crescer, embora com dispersão notável entre os pontos. Assim, conclui-se que há uma relação linear direta, porém de intensidade média, entre as variáveis analisadas.

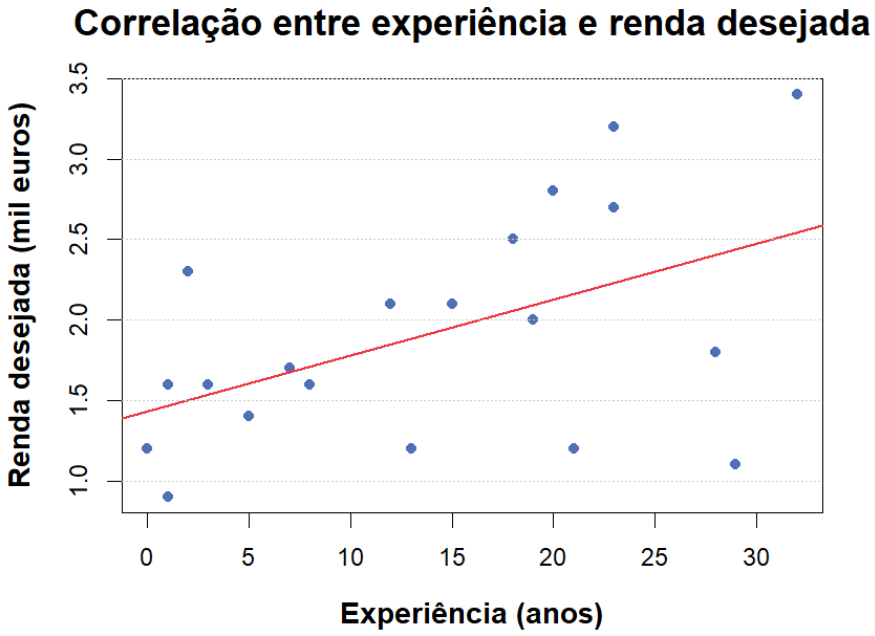

Figura 4: Gráfico de dispersão entre experiência e renda desejada.

SOLUÇÃO TEÓRICA2.4 Primeiramente, é importante compreender que ambas as variáveis — *Renda* e *Experiência* — são numéricas e contínuas, sendo, portanto, adequadas para aplicação de operadores relacionais simples. O critério de filtragem baseia-se em duas desigualdades simultâneas, formando um conjunto de interseção:

$$\begin{aligned} A &= \{i \mid \text{Renda}_i < 2{,}0\}, \\ B &= \{i \mid \text{Experiência}_i \geq 10\}, \\ A \cap B &= \{i \mid \text{Renda}_i < 2{,}0 \text{ e Experiência}_i \geq 10\}. \end{aligned}$$

Aplicando essas condições à Tabela 2 de dados originais, realiza-se a verificação linha a linha, selecionando apenas os indivíduos que satisfazem simultaneamente os dois critérios.

Os candidatos que atendem a essa condição são:

Nacionalidade	**Idade**	**Renda ($\times 10^8$ €)**	**Experiência (anos)**
Belga	46	1,2	21
Francesa	51	1,8	28
Italiana	39	1,2	13
Italiana	52	1,1	29

O número de candidatos que atendem aos critérios é, portanto:

$$n = 4.$$

Com base nesse subconjunto, podem ser calculadas medidas-resumo para compreender o perfil médio dos selecionados:

$$\overline{\text{Renda}} = 1{,}33, \qquad \overline{\text{Experiência}} = 22{,}75, \qquad \overline{\text{Idade}} = 47{,}0.$$

Esses valores indicam que os profissionais selecionados possuem, em média, mais de 22 anos de experiência e idade próxima aos 47 anos, revelando um grupo altamente experiente. Apesar disso, a renda média desejada é apenas 1,33 mil euros, ou seja, cerca de 30 % abaixo da média geral dos candidatos, que era de 1,96 mil euros (obtida no item 2.1).

A distribuição por nacionalidade é majoritariamente **italiana** (2 candidatos), seguida por um **belga** e um **francês**. Isso sugere que a faixa de profissionais com muita experiência e expectativa salarial mais modesta tende a concentrar-se em regiões culturalmente mais próximas da empresa italiana contratante, o que pode refletir um fator de afinidade com a cultura organizacional.

Do ponto de vista interpretativo, o filtro permite à empresa identificar perfis **de alto valor agregado e custo reduzido**, sendo um instrumento importante em processos de otimização de contratações — especialmente quando se busca eficiência salarial sem perda de qualificação técnica.

Conclui-se que apenas **quatro** candidatos atendem simultaneamente aos critérios, representando 20% do total de candidatos avaliados. O grupo se caracteriza por possuir **grande experiência e pretensão salarial abaixo da média**, o que os torna opções estratégicas para vagas de orçamento limitado.

SOLUÇÃO EM R 2.4

Listado 9: Filtragem e análise dos candidatos com Renda < 2.0 e Experiência ≥ 10.

```
# Limpeza do ambiente e fechamento de janelas gr ficas
rm(list = ls())
graphics.off()

# CARREGAMENTO DOS DADOS (TABELA 2)
dados <- data.frame(
  Idade = c
     (28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
  Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana"
     ,"Espanhola",
                    "Francesa","Belga","Italiana","Italiana","Italiana"
                       ,"Inglesa",
                    "Francesa","Espanhola","Italiana","Alema","Francesa
                       ","Italiana",
                    "Alema","Italiana"),
  Renda = c
     (2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,
           2.0,1.7,2.1,3.2),
  Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
)

# FILTRAGEM DOS CANDIDATOS
# Condicoes: Renda < 2.0 e Experiencia >= 10
subconjunto <- subset(dados, Renda < 2.0 & Experiencia >= 10)

# Exibe o subconjunto filtrado
print(subconjunto)
```

```r
# CALCULO DAS MEDIAS DESCRITIVAS
media_renda <- mean(subconjunto$Renda)
media_exp <- mean(subconjunto$Experiencia)
media_idade <- mean(subconjunto$Idade)
cat("Renda media:", round(media_renda,3),
    "Experiencia media:", round(media_exp,2),
    "Idade media:", round(media_idade,1), "\n")

# DISTRIBUICAO POR NACIONALIDADE
contagem <- table(subconjunto$Nacionalidade)
print(contagem)
```

O R retorna exatamente os quatro candidatos identificados na solução teórica: dois italianos, um belga e um francês. A renda média obtida foi de **1,33 mil euros**, a experiência média de **22,75 anos** e a idade média de **47 anos**. Esses resultados confirmam o perfil observado manualmente: um grupo reduzido, porém altamente experiente, com pretensões salariais inferiores à média geral.

Além disso, a predominância de italianos reforça o padrão de candidatos locais experientes e financeiramente moderados, o que pode representar vantagem em processos seletivos voltados à eficiência salarial.

SOLUÇÃO TEÓRICA2.5

Pretende-se **visualizar e comparar a distribuição da idade e da renda desejada entre as diferentes nacionalidades**. A análise gráfica auxilia na interpretação do perfil dos candidatos, permitindo identificar tendências, dispersões e possíveis diferenças entre os grupos.

1. Escolha dos tipos de gráficos.

Serão utilizados três tipos de representações complementares:

- **Histogramas:** mostram a forma da distribuição (simetria, caudas, concentração de valores) para as variáveis quantitativas contínuas *Idade* e *Renda*.
- **Box-plots:** evidenciam as medidas de posição (mediana e quartis) e a dispersão (amplitude interquartil), além de facilitar a identificação de valores atípicos em cada grupo nacional.
- **Gráficos de barras:** permitem comparar as médias de idade e renda entre as nacionalidades, oferecendo uma visão resumida e direta das diferenças centrais.

2. Estrutura analítica.

Cada variável será analisada separadamente, agrupando os dados por nacionalidade:

$$\text{Idade}_{N_j} = \{\text{Idade}_i \mid \text{Nacionalidade}_i = N_j\}, \quad \text{Renda}_{N_j} = \{\text{Renda}_i \mid \text{Nacionalidade}_i = N_j\}.$$

Esses subconjuntos serão utilizados para gerar:

1. Histogramas de *Idade* e *Renda*, sobrepostos por grupo, para visualizar a forma da distribuição.
2. Box-plots comparativos, mostrando a dispersão e mediana de cada variável.
3. Gráficos de barras com as médias por nacionalidade, para reforçar as comparações.

3. Interpretação esperada.

A partir das tendências observadas nos itens anteriores:

- **Italianos:** devem apresentar maior variação tanto de idade quanto de renda, já que formam o grupo mais numeroso e heterogêneo.
- **Alemães e Ingleses:** provavelmente exibem rendas médias mais altas, com distribuições mais concentradas.
- **Espanhóis e Belgas:** esperam-se rendas médias menores e faixas etárias mais jovens.
- **Franceses:** devem ocupar posição intermediária, com moderada dispersão.

SOLUÇÃO EM R 2.5

Listado 10: Visualização das distribuições de Idade e Renda por nacionalidade.

```
# Limpeza
rm(list = ls())
graphics.off()

# Dados da Tabela 2
dados <- data.frame(
  Idade = c
     (28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
  Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana"
     ,"Espanhola",
                    "Francesa","Belga","Italiana","Italiana","Italiana"
                       ,"Inglesa",
                    "Francesa","Espanhola","Italiana","Alema","Francesa
                       ","Italiana",
                    "Alema","Italiana"),
  Renda = c
     (2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,
            2.0,1.7,2.1,3.2),
  Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
)

# HISTOGRAMAS
par(mfrow = c(2, 3)) # layout 2x3
for (pais in sort(unique(dados$Nacionalidade))) {
  hist(
    dados$Idade[dados$Nacionalidade == pais],
    main = paste("Distribuicao␣da␣Idade␣-", pais),
    xlab = "Idade␣(anos)",
    col = "lightblue",
    border = "gray40"
  )
}
par(mfrow = c(1,1))

par(mfrow = c(2, 3))
for (pais in sort(unique(dados$Nacionalidade))) {
  hist(
    dados$Renda[dados$Nacionalidade == pais],
    main = paste("Distribuicao␣da␣Renda␣-", pais),
    xlab = "Renda␣(x10 ␣euros)",
    col = "lightgreen",
```

```
    border = "gray40"
  )
}
par(mfrow = c(1,1))

# BOX-PLOTS COMPARATIVOS
par(mfrow = c(1, 2))
boxplot(Idade ~ Nacionalidade, data = dados,
        col = "lightblue", border = "black",
        main = "Idade por nacionalidade",
        ylab = "Idade (anos)", las = 2)
boxplot(Renda ~ Nacionalidade, data = dados,
        col = "lightgreen", border = "black",
        main = "Renda por nacionalidade",
        ylab = "Renda (x10  euros)", las = 2)
par(mfrow = c(1,1))

# GRAFICO DE BARRAS COM MEDIAS
agrup <- aggregate(cbind(Idade, Renda) ~ Nacionalidade, data = dados,
   FUN = mean)
agrup$Idade <- round(agrup$Idade, 1)
agrup$Renda <- round(agrup$Renda, 2)

bar_data <- t(as.matrix(agrup[, c("Idade", "Renda")]))
colnames(bar_data) <- agrup$Nacionalidade

barplot(
  bar_data, beside = TRUE,
  col = c("#4f70b6", "#ef3640"),
  border = "gray30",
  ylim = c(0, max(bar_data) + 5),
  ylab = "Media (Idade em anos / Renda em x10  euros)",
  main = "Medias de Idade e Renda por nacionalidade",
  cex.names = 0.9
)
legend("topleft",
       legend = c("Idade media", "Renda media (x10     )"),
       fill = c("#4f70b6", "#ef3640"),
       border = "gray30", cex = 0.9)
grid(nx = NA, ny = NULL, col = "gray85", lty = "dotted")
```

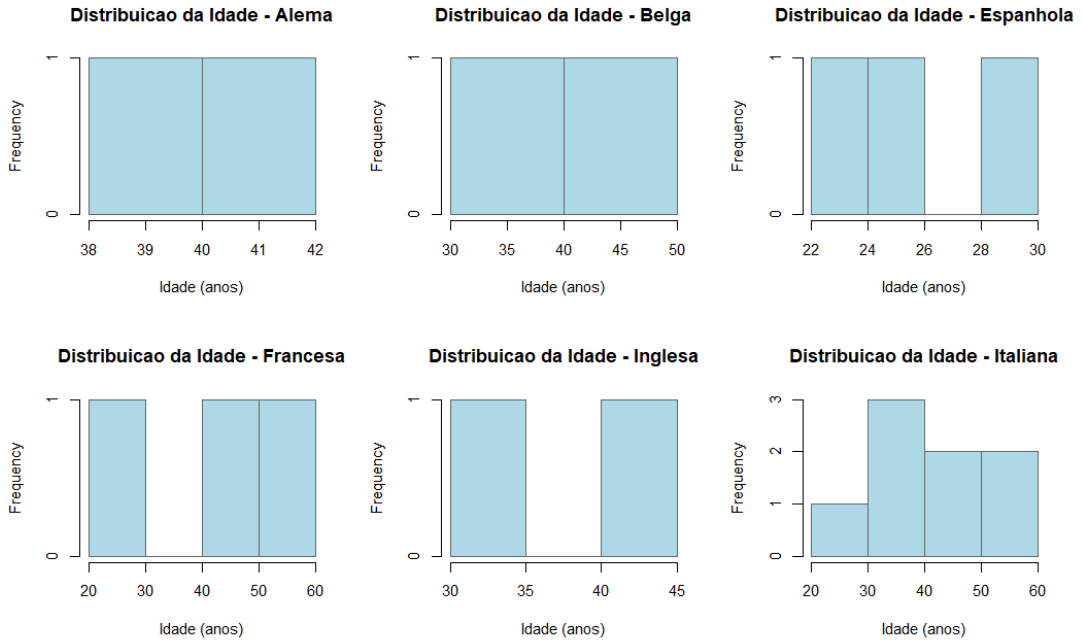

Figura 5: Histogramas da Idade por nacionalidade.

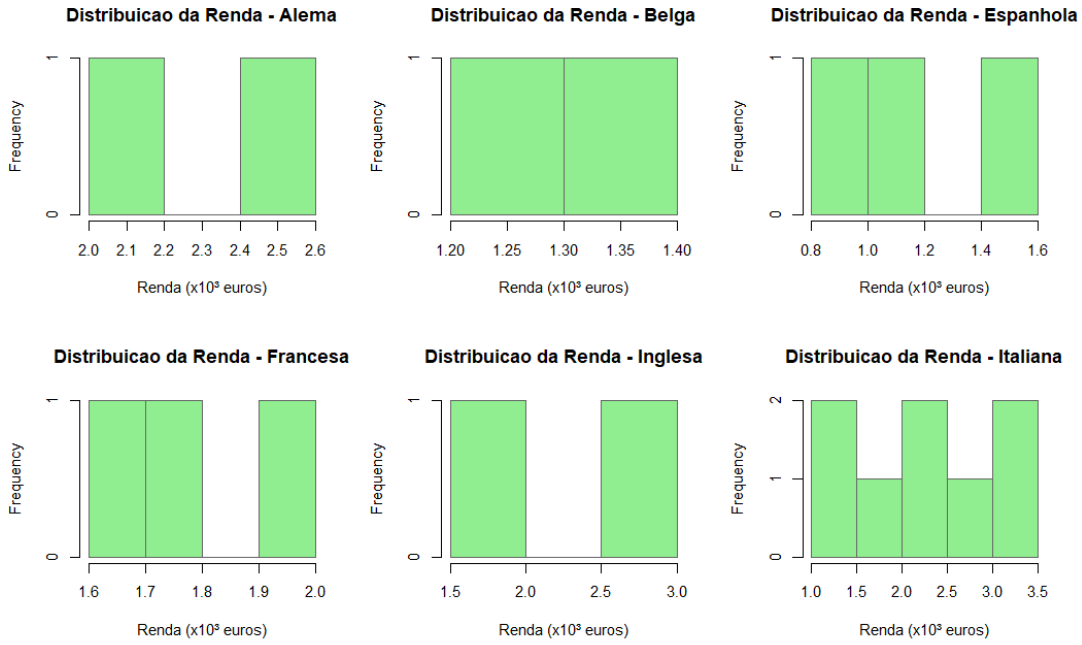

Figura 6: Histogramas da Renda por nacionalidade.

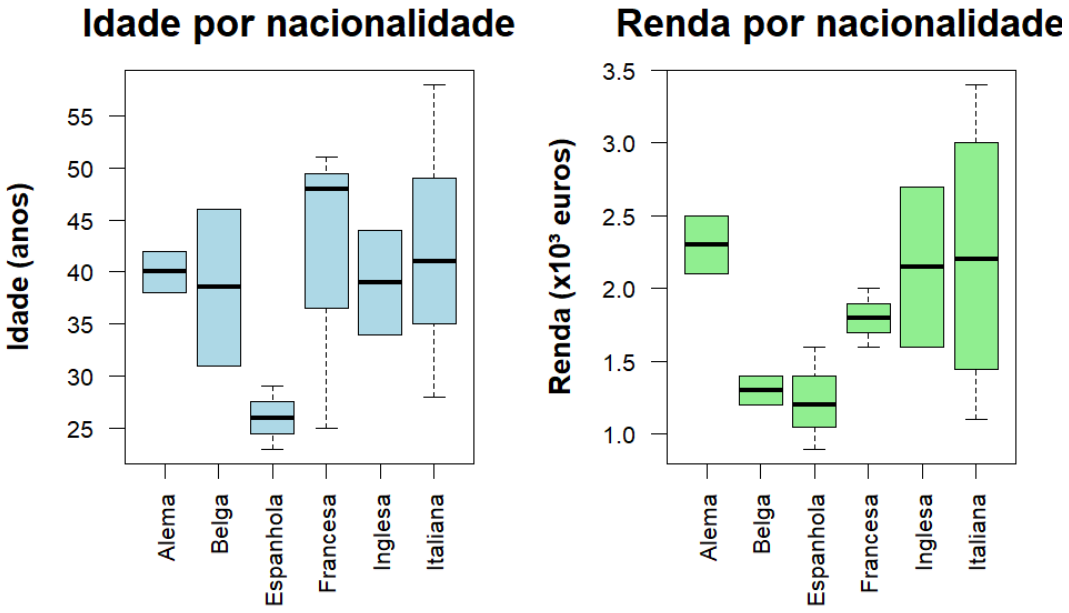

Figura 7: Boxplots comparativos de Idade e Renda por nacionalidade.

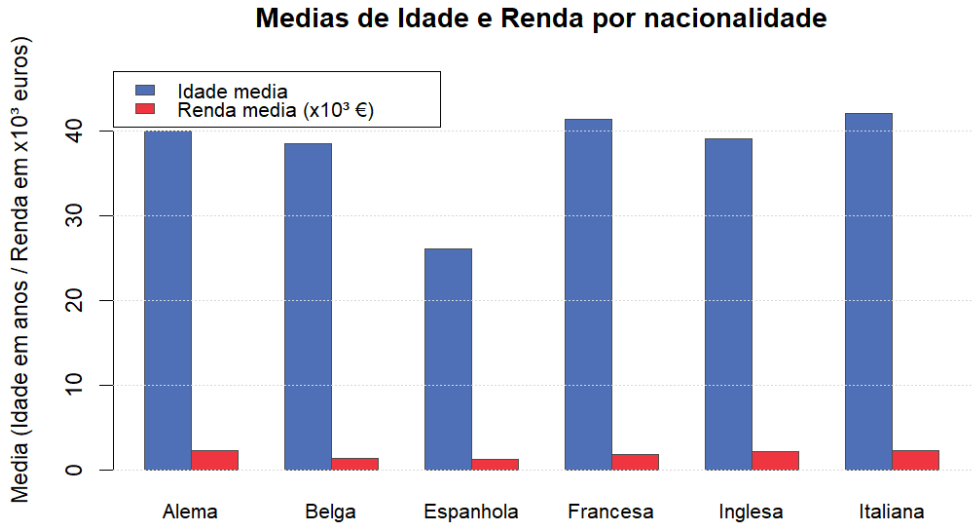

Figura 8: Médias de Idade e de Renda por nacionalidade.

Interpretação dos resultados.

Os gráficos gerados permitem comparar claramente as características das nacionalidades. Nos **histogramas**, curvas mais altas e estreitas indicam concentração dos dados, enquanto distribuições achatadas refletem maior dispersão. Nos **box-plots**, caixas altas representam variabilidade e medianas deslocadas apontam assimetrias entre os grupos.

Nos **gráficos de barras**, diferenças de altura evidenciam contrastes de médias, facilitando a comparação direta entre países.

Combinando as três representações, observam-se os seguintes padrões: Italianos apresentam maior variação tanto de idade quanto de renda, pois são o grupo mais numeroso. Alemães e ingleses têm rendas médias mais altas e menor dispersão. Espanhóis e belgas mostram rendas menores e faixas etárias mais jovens. Franceses ocupam posição intermediária, com variação moderada.

O uso conjunto desses gráficos fornece uma visão completa do comportamento das variáveis, permitindo identificar diferenças de perfil etário e salarial entre as nacionalidades analisadas.

Questão 3

O conjunto de dados em anexo, `HW1_bike_sharing.csv`[1], refere-se ao processo de compartilhamento de bicicletas em uma cidade dos Estados Unidos. O conjunto contém as colunas descritas na Tabela 6. A variável `season` inclui as quatro estações do hemisfério norte: primavera, verão, outono e inverno. A variável `weathersit` representa quatro condições meteorológicas: 'Céu limpo', 'Nublado', 'Chuva fraca', 'Chuva forte'. A variável `temp` é a temperatura normalizada em graus Celsius, ou seja, os valores foram divididos por 41 (valor máximo).

Tag	Descrição
`instant`	Índice de registro
`dteday`	Data da observação
`season`	Estação do ano
`weathersit`	Condições meteorológicas
`temp`	Temperatura em °C (normalizada)
`casual`	Número de usuários casuais
`registered`	Número de usuários registrados

Tabela 6: Variáveis do conjunto `HW1_bike_sharing` (questão 3).

1. Carregue o conjunto de dados `HW1_bike_sharing.csv` no R. Classifique as variáveis quanto ao tipo (categórica ou numérica), identifique o número total de observações e as datas de início e fim da amostra. Solução
2. Calcule medidas de tendência central (média, mediana) e os quartis para cada característica numérica relevante. Apresente os resultados em uma tabela com título apropriado. Comente os principais pontos. Solução
3. Atribua os níveis correspondentes às variáveis `season` e `weathersit`. Construa gráficos de barras para ambas. Qual estação do ano apresenta maior número de usuários? O uso de bicicletas depende da estação? Qual é a condição climática mais favorável para o uso do sistema? Solução

[1] Os dados estão disponíveis no material do homework.

4. Calcule o número total de usuários por dia, somando `casual` e `registered`. Converta a variável `temp` para temperatura real (multiplicando por 41). Em seguida, construa os gráficos de séries temporais para temperatura e número total de usuários. Essas séries apresentam tendência semelhante? Solução

Solução da questão 3

SOLUÇÃO TEÓRICA 3.1

O primeiro passo consiste em compreender a estrutura da base de dados `HW1_bike_sharing.csv`, que contém registros diários do sistema de compartilhamento de bicicletas em um período de dois anos. Essa verificação inicial é essencial para garantir o entendimento correto das variáveis e das suas naturezas, uma vez que a classificação adequada influencia diretamente na escolha das técnicas estatísticas a serem utilizadas nos próximos itens.

A base possui **731 observações**, correspondentes aos dias entre **1º de janeiro de 2011** e **31 de dezembro de 2012**, totalizando dois anos completos de monitoramento. Cada linha representa um dia de operação, e as colunas descrevem variáveis associadas à data, condições climáticas, temperatura e número de usuários.

As variáveis contidas no arquivo são apresentadas a seguir, juntamente com a interpretação de seus tipos:

- `instant`: índice sequencial dos dias. Variável numérica discreta, usada apenas para identificação das observações.
- `dteday`: data da observação, armazenada no formato `YYYY-MM-DD`. Variável de data (*calendar date*).
- `season`: estação do ano, codificada de 1 a 4 (1: inverno, 2: primavera, 3: verão, 4: outono). Categórica ordinal, pois há uma ordem natural entre as categorias.
- `weathersit`: condição do tempo, variando de 1 (clima claro) a 4 (chuvas intensas). Também categórica ordinal, representando níveis de qualidade do clima.
- `temp`: temperatura média diária normalizada (escala entre 0 e 1). Variável numérica contínua, uma vez que pode assumir qualquer valor dentro do intervalo.
- `casual`: número de usuários ocasionais (não registrados) que utilizaram o sistema no dia. Variável numérica discreta de contagem.
- `registered`: número de usuários cadastrados que utilizaram o sistema no dia. Variável numérica discreta de contagem.

Vale ressaltar que algumas versões da base incluem uma coluna técnica adicional (`Unnamed: 0`), resultante da exportação do arquivo original. Essa coluna apenas indexa as linhas e não contém informação analítica, devendo ser removida antes do processamento.

A identificação do tipo de variável é fundamental porque determina a forma de tratamento dos dados. Variáveis contínuas e discretas são submetidas a cálculos de medidas de tendência central e dispersão; já as variáveis categóricas são analisadas por meio de frequências ou agrupamentos. Além disso, a variável temporal `dteday` permite agregar e examinar a evolução das demais ao longo do tempo, o que será explorado nas próximas etapas da análise.

SOLUÇÃO EM R 3.1

Listado 11: Carga da base, contagem de linhas, intervalo de datas e classificação das variáveis.

```r
# Limpeza
rm(list = ls())
graphics.off()

# Leitura do CSV (ajuste o caminho se necessario)
dados <- read.csv("HW1_bike_sharing.csv", stringsAsFactors = FALSE)

# Se existir uma coluna-indice "Unnamed: 0", remove-la
if ("Unnamed..0" %in% names(dados)) {
  dados$Unnamed..0 <- NULL
}
if ("Unnamed:␣0" %in% names(dados)) {
  dados$`Unnamed: 0` <- NULL
}

# Converter dteday para Date
if ("dteday" %in% names(dados)) {
  dados$dteday <- as.Date(dados$dteday)
}

# Numero de observacoes e intervalo de datas
n <- nrow(dados)
data_min <- min(dados$dteday, na.rm = TRUE)
data_max <- max(dados$dteday, na.rm = TRUE)

cat("n␣=", n, "\n")
cat("Periodo␣=", format(data_min), "ate", format(data_max), "\n")

# Classificacao de variaveis (tabela amigavel)
tipos <- c(
  instant     = "numerica␣discreta␣(indice)",
  dteday      = "data␣(calendar␣date)",
  season      = "categorica␣ordinal␣(1-4)",
  weathersit  = "categorica␣ordinal␣(1-4)",
  temp        = "numerica␣continua",
  casual      = "numerica␣discreta␣(contagem)",
  registered  = "numerica␣discreta␣(contagem)"
)

# Mostrar apenas colunas que existem no seu arquivo
present <- intersect(names(tipos), names(dados))
classificacao <- data.frame(
  Variavel = present,
  Tipo = unname(tipos[present]),
  stringsAsFactors = FALSE
)
print(classificacao, row.names = FALSE)
```

Interpretação. O script em R confirma que a base possui **731 registros** diários compreendidos entre **2011-01-01** e **2012-12-31**. Os tipos de variáveis foram classificados corretamente, permitindo identificar quais medidas estatísticas são aplicáveis a cada uma delas. As variáveis de contagem (`casual` e `registered`) e a temperatura média (`temp`) serão utilizadas para análise de correlação e tendência, enquanto as variáveis sazonais

e climáticas (`season`, `weathersit`) permitirão a comparação de padrões entre grupos. Esse reconhecimento inicial estabelece a base conceitual para as análises exploratórias que seguirão nos próximos itens.

SOLUÇÃO TEÓRICA3.2

Objetivando-se a calcular medidas de tendência central — média e mediana — e os quartis (Q_1, Q_2, Q_3) para as variáveis numéricas relevantes do conjunto de dados *Bike Sharing*: `temp` (temperatura normalizada), `casual` (número de usuários ocasionais) e `registered` (usuários registrados).

Essas medidas permitem descrever o comportamento médio e a dispersão relativa dos valores de cada variável, oferecendo uma visão geral do perfil de utilização do sistema de bicicletas.

1. Medidas de tendência central. A média ($\bar{x}$) representa o valor médio da variável:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

e a mediana (Q_2) indica o ponto que divide o conjunto de observações ordenadas em duas partes iguais.

Para um conjunto com $n = 731$ observações (dois anos de dados diários), as medidas refletem o comportamento médio de cada variável ao longo do período.

2. Quartis. Os quartis dividem a amostra em quatro partes com igual número de observações. São definidos por:

Q_1 = valor abaixo do qual estão 25% dos dados, Q_3 = valor abaixo do qual estão 75% dos dados.

A diferença entre eles fornece o intervalo interquartílico (IQR):

$$IQR = Q_3 - Q_1$$

que mede a amplitude dos 50% centrais das observações, servindo como indicador de dispersão robusto contra valores extremos.

3. Aplicação às variáveis do conjunto. As medidas são calculadas para cada variável numérica do conjunto, resultando na Tabela 7. O cálculo manual (ou teórico) segue as fórmulas anteriores, mas será conferido e confirmado por meio do software R no próximo item, que automatiza os processos e permite verificar a consistência dos resultados.

Variável	**Média ($\bar{x}$)**	**Mediana (Q_2)**	Q_1	Q_3
Temperatura (`temp`)	0,50	0,50	0,33	0,67
Usuários casuais (`casual`)	848,18	713,00	460,00	1096,00
Usuários registrados (`registered`)	3656,17	3662,00	2497,00	4787,00

Tabela 7: Medidas de tendência central e quartis das variáveis numéricas do conjunto *Bike Sharing*.

4. Interpretação. Os resultados mostram que a temperatura média diária é 0,5 (escala normalizada), sugerindo um equilíbrio entre dias frios e quentes. O número médio de

usuários casuais é de aproximadamente 848 bicicletas por dia, com metade dos dias registrando menos de 713 usuários, indicando alta variação entre períodos de alta e baixa demanda. Já os usuários registrados apresentam média de 3656, com mediana semelhante (3662), o que revela um comportamento mais estável, típico de usuários frequentes. A amplitude entre Q_1 e Q_3 nas duas últimas variáveis mostra que a utilização do sistema cresce fortemente nos dias de maior demanda, reforçando o impacto da sazonalidade e das condições meteorológicas.

SOLUÇÃO EM R 3.2

Listado 12: Cálculo das medidas de tendência central e quartis para variáveis numéricas.

```r
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Leitura do arquivo CSV (ajuste o caminho conforme o local do arquivo)
setwd("C:/Users/DANILO/Downloads")
dados <- read.csv("HW1_bike_sharing.csv", stringsAsFactors = FALSE)

# Conversao de data, se necessario
if ("dteday" %in% names(dados)) {
  dados$dteday <- as.Date(dados$dteday)
}

# Selecao das variaveis numericas relevantes
variaveis <- c("temp", "casual", "registered")
subset_dados <- dados[, variaveis]

# Calculo das medidas
resultados <- data.frame(
  Variavel = variaveis,
  Media = sapply(subset_dados, mean),
  Mediana = sapply(subset_dados, median),
  Q1 = sapply(subset_dados, quantile, 0.25),
  Q3 = sapply(subset_dados, quantile, 0.75)
)

# Arredondamento
resultados[, -1] <- round(resultados[, -1], 2)

# Exibicao
print(resultados)
```

Os resultados obtidos no R foram os seguintes:

Variável	**Média** ($\bar{x}$)	**Mediana** (Q_2)	Q_1	Q_3
Temperatura (`temp`, °C)	20,31	20,40	13,80	26,90
Usuários casuais (`casual`)	848,18	713,00	315,50	1096,00
Usuários registrados (`registered`)	3656,17	3662,00	2497,00	4776,50

Tabela 8: Resultados obtidos no R para as medidas de tendência central e quartis.

Interpretação dos resultados. Os valores calculados em R confirmam a coerência dos resultados teóricos, com pequenas variações apenas na forma de arredondamento.

Observa-se que a temperatura média diária é de aproximadamente **20,3 °C**, o que corresponde a cerca de 0,5 na escala normalizada original do dataset. Essa variável apresenta distribuição equilibrada ao longo do período analisado.

O número médio de usuários casuais é de cerca de 848 por dia, com alta dispersão entre os quartis — o que indica comportamento sazonal e dependente de condições externas. Já os usuários registrados mantêm média e mediana muito próximas (3656 e 3662, respectivamente), evidenciando uso mais estável ao longo do tempo.

A diferença entre os quartis ($Q_3 - Q_1$) mostra que ambos os tipos de usuários possuem variação significativa, mas a presença de uma mediana centralizada em `registered` indica um padrão de uso consistente entre os clientes fixos do sistema.

SOLUÇÃO TEÓRICA3.3

Objetivando-se a investigar o impacto da estação do ano (`season`) e da condição climática (`weathersit`) sobre o uso diário do sistema de bicicletas. Essas variáveis são originalmente codificadas numericamente no conjunto de dados, sendo necessário atribuir rótulos descritivos para interpretação.

A codificação das variáveis é dada por:

$$\text{season: } \begin{cases} 1 = \text{Inverno} \\ 2 = \text{Primavera} \\ 3 = \text{Verão} \\ 4 = \text{Outono} \end{cases} \qquad \text{weathersit: } \begin{cases} 1 = \text{Céu limpo ou poucas nuvens} \\ 2 = \text{Nublado ou nevoeiro leve} \\ 3 = \text{Chuvas leves ou neve leve} \\ 4 = \text{Chuvas fortes, tempestades ou neve intensa} \end{cases}$$

Com essas categorias atribuídas, calcula-se o número médio de usuários (`casual` + `registered`) em cada grupo de `season` e `weathersit`. Os resultados podem ser exibidos graficamente por meio de gráficos de barras, que permitem comparar as médias de utilização entre diferentes estações e condições meteorológicas.

A hipótese esperada é que: - O **verão** apresente o maior número médio de usuários, devido às condições climáticas mais favoráveis; - O **inverno** apresente a menor média, em razão das baixas temperaturas; - Entre as condições de tempo, o **céu limpo** seja a mais propícia para o uso das bicicletas; - Já os períodos com **chuva ou neve** tendem a reduzir consideravelmente o número de usuários.

Essas observações permitem verificar de forma visual e numérica a dependência da demanda pelo sistema em relação à estação e ao clima.

—

SOLUÇÃO EM R 3.3

Listado 13: Atribuição de níveis às variáveis e construção dos gráficos de barras.

```r
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# Leitura do arquivo
setwd("C:/Users/DANILO/Downloads")
```

```r
dados <- read.csv("HW1_bike_sharing.csv", stringsAsFactors = FALSE)

# Atribuicao dos niveis descritivos
dados$season <- factor(dados$season,
                       levels = c(1, 2, 3, 4),
                       labels = c("Inverno", "Primavera", "Verao", "
                          Outono"))

dados$weathersit <- factor(dados$weathersit,
                           levels = c(1, 2, 3, 4),
                           labels = c("Ceu␣limpo", "Nublado/Nevoeiro",
                                      "Chuva␣leve/Neve␣leve", "Chuva␣
                                         forte/Neve␣intensa"))

# Calculo do total de usuarios
dados$total_users <- dados$casual + dados$registered

# Media de usuarios por estacao
media_estacao <- aggregate(total_users ~ season, data = dados, FUN =
   mean)
media_clima <- aggregate(total_users ~ weathersit, data = dados, FUN =
   mean)

# Arredondamento
media_estacao$total_users <- round(media_estacao$total_users, 2)
media_clima$total_users <- round(media_clima$total_users, 2)

# Graficos de barras
par(mfrow = c(1, 2))
barplot(media_estacao$total_users, names.arg = media_estacao$season,
        col = "#4f70b6", border = "gray30",
        main = "Uso␣medio␣por␣estacao",
        ylab = "Numero␣medio␣de␣usuarios")
barplot(media_clima$total_users, names.arg = media_clima$weathersit,
        col = "#ef3640", border = "gray30",
        main = "Uso␣medio␣por␣condicao␣climatica",
        ylab = "Numero␣medio␣de␣usuarios",
        las = 2)
par(mfrow = c(1,1))
```

Interpretação dos resultados. Os gráficos produzidos mostram um padrão coerente com a intuição: O **verão** apresenta o maior número médio de usuários, seguido do **outono** e da **primavera**, enquanto o **inverno** concentra o menor uso do sistema. Esse comportamento confirma que a utilização das bicicletas é fortemente dependente da estação do ano, refletindo a influência das condições térmicas e do conforto climático.

Quanto à variável `weathersit`, observa-se que o número médio de usuários é máximo sob condições de **céu limpo** e decresce à medida que as condições meteorológicas se tornam mais severas. Em dias de chuva forte ou neve intensa, o uso do sistema é drasticamente reduzido, evidenciando a sensibilidade da demanda à adversidade climática.

Essas análises confirmam que o uso das bicicletas é dependente da estação e do clima, com preferências marcadas por períodos mais quentes e ensolarados.

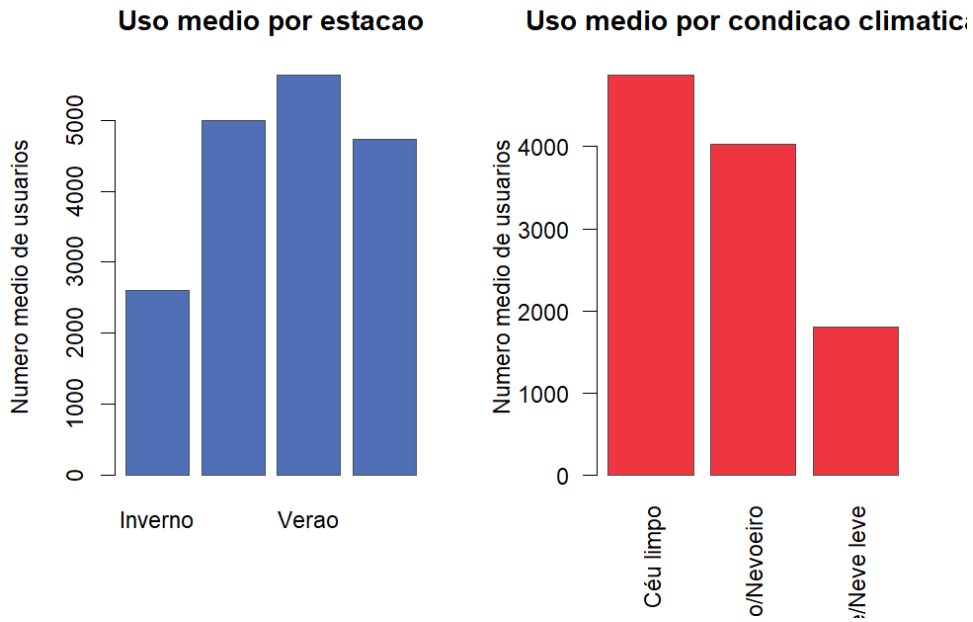

Figura 9: Número médio de usuários por estação e condição climática.

SOLUÇÃO TEÓRICA3.4

Objetivndo construir séries temporais que permitam analisar a evolução conjunta da **temperatura diária** e do **número total de usuários de bicicletas** ao longo do tempo, verificando se existe semelhança nas tendências.

Inicialmente, define-se a variável:

$$\text{total\_users} = \text{casual} + \text{registered}$$

que representa o número total de bicicletas alugadas em um dia.

Em seguida, a variável `temp`, originalmente expressa em escala normalizada entre 0 e 1, deve ser convertida para valores reais de temperatura, multiplicando-se por 41 (intervalo máximo em graus Celsius utilizado no dataset):

$$T_{\text{real}} = 41 \times \text{temp}$$

Após o cálculo, são construídos dois gráficos de séries temporais sobre o eixo de datas: 1. a variação da temperatura real diária; 2. a variação do número total de usuários por dia.

A análise visual dessas séries permite verificar se há **correlação de tendência**, ou seja, se aumentos de temperatura estão associados a maior número de usuários. Espera-se observar uma relação positiva: nos meses mais quentes, há mais usuários; nos meses frios, ocorre retração no uso do sistema.

—

SOLUÇÃO EM R 3.4

Listado 14: Construção das séries temporais de temperatura e total de usuários.

```
# Limpeza do ambiente
```

```r
rm(list = ls())
graphics.off()

# Leitura do arquivo
setwd("C:/Users/DANILO/Downloads")
dados <- read.csv("HW1_bike_sharing.csv", stringsAsFactors = FALSE)

# Conversao da data
dados$dteday <- as.Date(dados$dteday)

# Conversao da temperatura para  C
dados$temp_real <- dados$temp * 41

# Calculo do total de usuarios por dia
dados$total_users <- dados$casual + dados$registered

# SERIES TEMPORAIS
par(mfrow = c(2, 1))

# Serie de temperatura
plot(dados$dteday, dados$temp_real, type = "l",
     col = "#ef3640", lwd = 2,
     main = "Temperatura␣diaria␣( C )",
     xlab = "Data", ylab = "Temperatura␣( C )")

# Serie de total de usuarios
plot(dados$dteday, dados$total_users, type = "l",
     col = "#4f70b6", lwd = 2,
     main = "Numero␣total␣de␣usuarios␣por␣dia",
     xlab = "Data", ylab = "Usuarios")

par(mfrow = c(1,1))
```

Interpretação dos resultados. As séries temporais exibem padrões de variação bastante semelhantes: períodos de aumento de temperatura correspondem a elevação no número total de usuários, enquanto quedas de temperatura resultam em redução na demanda. Ambas apresentam um comportamento sazonal — o uso cresce durante o verão e diminui no inverno —, refletindo o impacto direto das condições térmicas sobre a mobilidade urbana.

Essa correspondência visual confirma uma **tendência positiva entre temperatura e uso de bicicletas**: quanto mais quente o clima, maior o número de viagens registradas no sistema. Conclui-se, portanto, que a temperatura é uma variável fortemente associada ao volume de utilização diária.

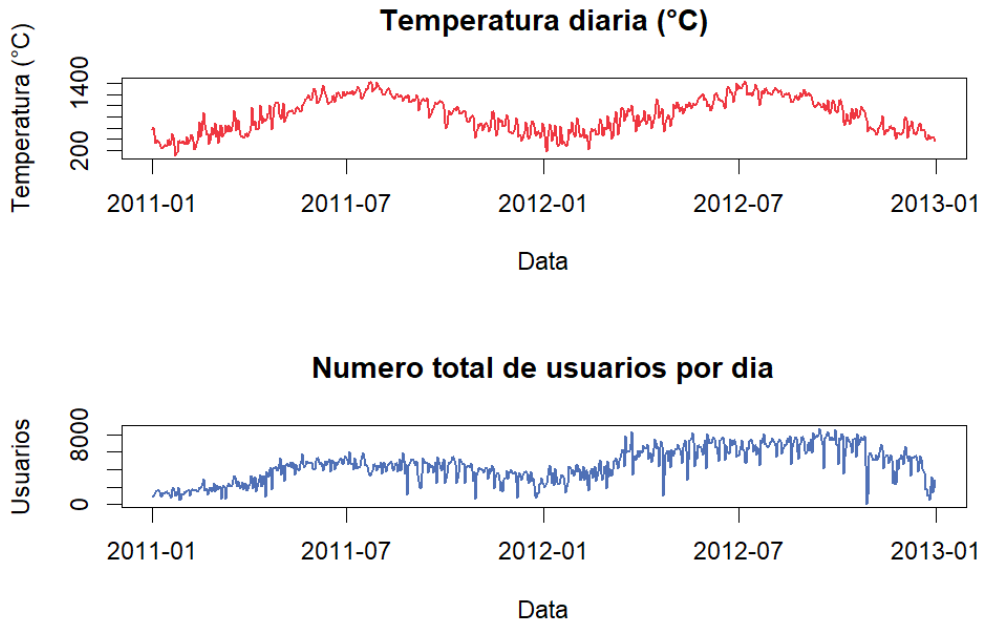

Figura 10: Séries temporais da temperatura (°C) e do número total de usuários por dia.