Statistics for Engineers
TI0111

Homework I

Nome completo: Matheus Parente Reis e Vinicius Alexandre Gomes do Nascimento

Numero de matricula: 571954 e 568594

Repositório: Link

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.
2. Construa um histograma e um boxplot para os dados de emissões. Os dados parecem estar simetricamente distribuídos? Existem valores atípicos?
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.
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 da questão 1

A questão pede que analisemos os dados sobre emissões diárias de um gás poluente, analisando o comportamento geral das emissões e suas medidas estatísticas (media, variância, amplitude,etc). Também utilizando representações gráficas para a analise e verificando se as emissões se adequam a um limite estabelecido.

1. Tomar as seguintes fórmulas para calcular as medidas solicitadas:
   onde n é o tamanho da lista.

   - Medidas de tendência central

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

   $$\text{posição da mediana} = P_{m_d} = \frac{n+1}{2}$$

   caso a posição da mediana não seja um inteiro, tomamos a mediana como a média dos valores adjacentes. Obs: precisa ser calculada

   $$moda = m_o = \text{valor na lista com maior número de repetições}$$

   - Medidas de dispersão

   $$amplitude = \text{diferença entre o maior e menor dado} = x_{max} - x_{min}$$

   $$\text{variância} = \sigma^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n}$$

   $$\text{desvio padrão} = \sqrt{\sigma^2} = \sigma$$

   $$\text{coeficiente de variação} = \frac{\sigma}{m_e}$$

   Tabelando os valores encontrados.

Média	Mediana	Moda	Variância	Desvio Padrão	Coeficiente de variação
19.02	19.15	19.4	30.45	5.52	0.29

Tabela 2: Tabela Medidas

Os resultados expressam a média de emissão do gás é de 19.02, com as amostras variando cerca de 29% desse valor. A emissão mais comum é 19.4, registrado 3 vezes nos dados. Os valores de mediana e media sendo próximos também refletem que os dados não possuem muitos outliers.

Agora, tratando-se da análise em R: Primeiramente, os dados foram copiados direto da tarefa para um **.txt** e carregados como um `dataframe`(chamado de **dados**) por meio da função **scan()**:

```
# dados da questao
dados <- scan("dados/emissoes.txt")
```

Logo após isso, para as medidas de tendência central, como média e mediana, utilizou-se as funções nativas do R para calculá-las. Para moda, por não apresentar uma função nativa para valores quantitativos, criou-se uma função que traduzia o *dataframe* em uma tabela, retornando aquela chave com maior número de repetições.

Para melhor apresentação dos resultados, criou-se um *dataframe* com esses valores e seus respectivos nomes:

```
# dados de tendencia central
media <- mean(dados)
mediana <- median(dados)
moda <- function(x){
    tabela <- table(x)
    as.numeric(names(tabela)[tabela == max(tabela)])
}
moda <- moda(dados)

sumario_central <- data.frame(
    Media = media,
    Mediana = mediana,
    Moda = moda
)
sumario_central
```

Ao final, temos como saída:

```
   Media Mediana Moda
19.02125   19.15 19.4
```

No que diz respeito às medidas de dispersão, a variância e o desvio padrão foram calculados por meio das funções nativas do R, enquanto a amplitude e o coeficiente de variação foram obtidos, respectivamente, pela diferença entre os valores máximo e mínimo da amostra e pela razão entre o desvio padrão e a média, multiplicada por 100, representando-o em forma percentual.

```r
# dados de dispersao
amplitude <- range(dados)[2] - range(dados)[1]
variancia <- var(dados)
desvio_padrao <- sd(dados)
coef_variacao <- (desvio_padrao/media) * 100

sumario_dispersao <- data.frame(
    Amplitude = amplitude,
    Variancia = variancia,
    `Desvio Padrao` = desvio_padrao,
    `Coef. de Variacao` = coef_variacao
)
sumario_dispersao
```

E como saída:

```
Amplitude Variancia Desvio.Padrao Coef..de.Variacao
     25.6  30.84144      5.553507           29.19633
```

2. Plot dos gráficos

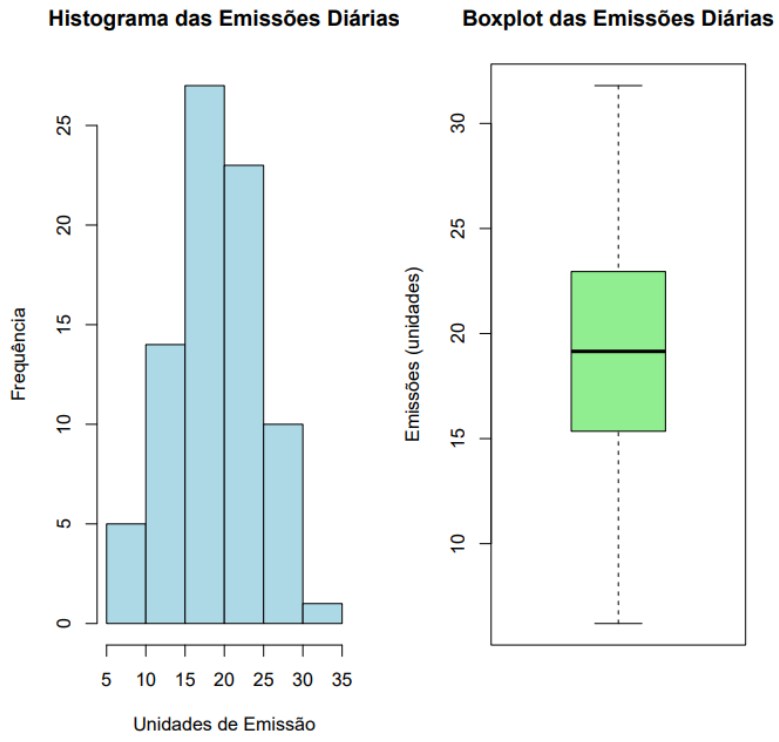

Figura 1: Gráficos dados de emissão

Os dados estão bem distribuídos, como pode ser visto no boxplot, onde a mediana esta bem no meio do gráfico, mas os valores mínimos são um pouco mais presentes que os máximos, denotado pelos extremos esquerdo e direito do histograma. Não existem outliers nos dados, pois seriam denotados como pontos isolados no boxplot.

Sobre os gráficos apresentados, eles foram feitos em R criando um **.pdf** em uma pasta específica do projeto que armazena os gráficos. As funções utilizadas para criar esses gráficos são todas nativas e de simples implementação, tendo somente como parâmetro os dados a serem utilizados, legendas e cores:

```r
# cria um canva para plotar os dois graficos
pdf("graficos/graficos_questao1.pdf")
par(mfrow = c(1,2))

# Histograma
hist(dados,
     main = "Histograma das Emissoes Diarias",
     xlab = "Unidades de Emissao",
     ylab = "Frequencia",
     col = "lightblue",
     border = "black"
)

# Boxplot
boxplot(dados,
        main = "Boxplot das Emissoes Diarias",
        ylab = "Emissoes (unidades)",
        col = "lightgreen",
        border = "black"
)

# fecha e salva o pdf
dev.off()
```

3. As fórmulas utilizadas para determinar a posição de cada um dos quartis e o IQR:

$$P_{Q_1} = \frac{1}{4}(n+1)$$

$$P_{Q_2} = P_{m_d} = \frac{n+1}{2}$$

$$P_{Q_3} = \frac{3}{4}(n+1)$$

$$IQR = Q_3 - Q_1$$

Caso as posições não sejam inteiros os valores dos quartis sera o da posição mais próxima ou a media das posições adjacentes. com os valores dos quartis podemos

Q_1	Q_2/m_d	Q_3	IQR
15.2	19.15	23	7.8

Tabela 3: Tabela Quartis

encontrar os limites superior e inferior, onde fora deles todos valores são outiliers.

$$L_{superior} = Q_3 + 1.5IQR = 34.7$$

$$L_{inferior} = Q_1 - 1.5IQR = 3.5$$

Pode se observar que nos dados não ha amostras nem acima do limite superior e nem abaixo do limite inferior, com isso podemos concluir afirmar a conclusão de que não há outliers nos dados.

Afim de comprovar esses dados, foi feito o seguinte script em R:

```r
# calculo dos quartis
q1 <- as.numeric(quantile(dados, probs = c(1/4)))
q2 <- as.numeric(quantile(dados, probs = c(2/4)))
q3 <- as.numeric(quantile(dados, probs = c(3/4)))

# calculo do intervalo interquartil
iqr <- IQR(dados)

# calculo de limites por meio do IQR
limite_inferior <- q1 - (1.5 * iqr)
limite_superior <- q3 + (1.5 * iqr)

outliers <- dados[dados < limite_inferior | dados > limite_
    superior]

sumario_quartis <- data.frame(
    Q1 = q1,
    Q2 = q2,
    Q3 = q3,
    `Limite Sup.` = limite_superior,
    `Limite Inf.` = limite_inferior,
    `N de outliers` = length(outliers)
)
sumario_quartis
```

Nele cada uns dos quartis foram calculados individualmente por meio da função `quantile()`, o IQR foi calculado pela função de mesmo nome e os limites foram calculados pela mesma fórmula apresentada no início da resolução.

Para visualização dos *outliers*, selecionou-se aqueles dados que eram inferiores e superiores aos limites determinados. Chegando na seguinte saída:

```
    Q1    Q2     Q3 Limite.Sup. Limite.Inf. N..de.outliers
15.425 19.15 22.925       34.175       4.175              0
```

4. Fórmula para encontrar a proporção

$$p = \frac{dias\ acima\ do\ limite}{dias\ totais}$$

A partir dos dados se encontra que 11 dias ficaram acima do limite de emissão diário

$$p = \frac{11}{80} = 0.138$$

Aproximadamente 14% dos dias ficaram acima do limite permitido. Com isso, se conclui que o comportamento geral de emissões se mantêm dentro do limite.

Em R:

```r
# supondo limite maximo de 25
excederam <- dados[dados > 25]
proporcao <- length(excederam)/length(dados) * 100
proporcao
```

Para calcular essa porcentagem, foram selecionados todos aqueles dados que superam 25 em unidade de emissão. O número de amostras selecionadas foi dividida pelo número de amostras totais. Multiplicasse por 100 para obter esse valor em porcentagem e como saída:

```
[1] 13.75
```

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 4 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.

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

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?

2. Agrupe os candidatos por nacionalidade e calcule a renda média desejada e os anos médios de experiência para cada grupo. Qual nacionalidade apresenta a maior renda média desejada? Qual grupo aparenta ser o mais experiente?

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.

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.

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 da questão 2

A atividade tem base na obtenção de dados estatísticos a partir da tabela dada, extraindo dados uteis para a analise dos candidatos e entender correlações entre as diferentes informações.

1. Vamos começar calculando a média, mediana e desvio padrão teóricos para cada variável de acordo com as fórmulas:

$$Média = m_e = \frac{1}{n} \cdot \sum_{i}^{n} x_i$$

$$\text{Posição } Mediana = P_{m_d} = \frac{n+1}{2}$$

Essa fórmula nos da a posição da mediana dentro da lista de valores ordenados. Caso a posição não seja um número inteiro é tirada a media com os as posições adjacentes.

$$Desvio\ Padrão = \sigma = \sqrt{\frac{1}{n} \cdot \sum_{i}^{n} (x_i - m_e)^2}$$

Onde nosso m_e sera a média para aquela informação.

- Resultados

	Idade	Renda	Experiencia
M_e	38.65	1.92	14
M_d	40.5	1.9	16.5
σ	9.68	0.7	10

Tabela 5: Tabela com Informações Médias

- Conclusão

  Com os dados obtidos é possível concluir que em média os candidatos possuem 38.65 (aproximadamente 39 anos), esperam 1.92 mil euros de salario e possuem 14 anos de experiência. Além disso se observa que não existem outliers, pois as médias sempre estão próximas das medianas, mesmo com os dados possuindo uma boa dispersão indicada pelos desvios padrão.

- Código R

  Para que seja mais facilmente analisado, os dados da tabela foram transformados em um *dataframe*:

```r
# transformando a tabela em dataframe
dados <- data.frame(
  idade = c(28, 34,..., 38, 46),
  nacionalidade = c("Italiana", "Inglesa",..., "Alemana", "
     Italiana"),
  renda_desejada = c(2.3, 1.6,..., 2.1, 3.2),
  experiencia_anos = c(2, 8,..., 12, 23)
)
dados
```

  Eis o *dataframe*:

```
   idade nacionalidade renda_desejada experiencia_anos
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
```

Para os dados de Média, Mediana e e Desvio Padrão foram usadas as funções nativas do R. Entregando como parâmetro os respectivos valores de Idade, Renda e Experiência. Assim como anteriormente, utilizou-se outro *dataframe* focado em visualizar os dados no terminal:

```
media_idade <- mean(dados$idade)
mediana_idade <- median(dados$idade)
desvio_idade <- sd(dados$idade)

media_renda <- mean(dados$renda_desejada)
mediana_renda <- median(dados$renda_desejada)
desvio_renda <- sd(dados$renda_desejada)

media_experiencia <- mean(dados$experiencia_anos)
mediana_experiencia <- median(dados$experiencia_anos)
desvio_renda <- sd(dados$experiencia_anos)

tabela <- data.frame(
  Medida = c("Media", "Mediana", "Desvio Padrao"),
  Idade = c(media_idade, mediana_idade, desvio_idade),
  Renda = c(media_renda, mediana_renda, desvio_renda),
  Experiencia = c(media_experiencia, mediana_experiencia,
     desvio_experiencia)
)
tabela
```

Gerando como saída a seguinte tabela:

```
       Medida    Idade      Renda Experiencia
        Media 38.6500 1.9200000    14.00000
      Mediana 38.5000 1.7500000    14.00000
Desvio Padrao  9.9275 0.7134792    10.27004
```

2. De inicio é necessário agrupar os candidatos por nacionalidade e obter as médias.

   - Italiana

$$renda\ media = \frac{2.3 + 2.1 + ... + 3.2}{8} = 2.25 \text{ mil euros}$$

$$experiencia\ media = \frac{2 + 15 + 13 + ... + 23}{8} = 17.6 = 17\ anos$$

   - Inglesa

$$renda\ media = \frac{1.6 + 2.7}{2} = 2.15 \text{ mil euros}$$

$$experiencia\ media = \frac{8 + 23}{2} = 15.5 = 15\ anos$$

	Idade	Nacionalidade	Renda	Experiência
1	28	Italiana	2.3	2
2	37	Italiana	2.1	15
3	39	Italiana	1.2	13
4	43	Italiana	2.8	20
5	58	Italiana	3.4	32
6	52	Italiana	1.1	29
7	33	Italiana	1.7	7
8	46	Italiana	3.2	23

Tabela 6: Candidatos Italianos

	Idade	Nacionalidade	Renda	Experiência
1	34	Inglesa	1.6	8
2	44	Inglesa	2.7	23

Tabela 7: Candidatos Ingleses

- Belga

	Idade	Nacionalidade	Renda	Experiência
1	46	Belga	1.2	21
2	31	Belga	1.4	5

Tabela 8: Candidatos Belgas

$$renda\ media = \frac{1.2 + 1.4}{2} = 1.3 \text{ mil euros}$$

$$experiencia\ media = \frac{21 + 5}{2} = 13\ anos$$

- Espanhola

	Idade	Nacionalidade	Renda	Experiência
1	26	Espanhola	0.9	1
2	29	Espanhola	1.6	3
3	39	Espanhola	1.2	0

Tabela 9: Candidatos Espanhóis

$$renda\ media = \frac{0.9 + 1.6 + 1.2}{3} = 1.23 \text{ mil euros}$$

$$experiencia\ media = \frac{1 + 3 + 0}{3} = 1.3 = 1\ ano$$

- Francesa

	Idade	Nacionalidade	Renda	Experiência
1	51	Francesa	1.8	28
2	25	Francesa	1.6	1
3	48	Francesa	2.0	19

Tabela 10: Candidatos Francesa

$$renda\ media = \frac{1.8 + 1.6 + 2.0}{3} = 1.8 \text{ mil euros}$$

$$experiencia\ media = \frac{28 + 1 + 19}{3} = 16 = 16\ anos$$

- Alemana

	Idade	Nacionalidade	Renda	Experiência
1	42	Alemana	2.5	18
2	38	Alemana	2.1	12

Tabela 11: Candidatos Alemães

$$renda\ media = \frac{2.5 + 2.1}{2} = 2.3 \text{ mil euros}$$

$$experiencia\ media = \frac{18 + 12}{2} = 15\ anos$$

Construindo uma tabela com os valores médios de cada nacionalidade.

	Nacionalidade	Renda média	Experiencia Média
1	Italiana	2.25	17
2	Inglesa	2.15	15
3	Belga	1.3	13
4	Espanhola	1.23	1
5	Francesa	1.8	16
6	Alemana	2.3	15

Tabela 12: Médias Nacionalidades

- Conclusão

  Entre todas nacionalidades podemos observar que os candidatos de nacionalidade alemana são os que possuem uma maior renda media desejada. Enquanto aqueles com mais anos de experiencia são de nacionalidade italiana.

- Código em R

  Para que fosse feito mais facilmente a divisão dos dados, utilizou-se a função `aggregate()`, que relaciona dois valores desejados no *dataframe*. Convenientemente, esse função aceita como parâmetro o que eu quero com os dados que, nesse caso, é a média(mean):

```
renda_media_por_nacionalidade = aggregate(renda_desejada ~
    nacionalidade,
                                          data = dados,
                                          FUN = mean
                                         )
experiencia_media_por_nacionalidade = aggregate(experiencia_
    anos ~ nacionalidade,
                                          data = dados,
                                          FUN = mean
                                         )
renda_media_por_nacionalidade
experiencia_media_por_nacionalidade
```

Com isso, criou-se outros dois *dataframe* listando as médias por nacionalidade:

```
nacionalidade renda_desejada
      Alemana       2.300000
        Belga       1.300000
    Espanhola       1.233333
     Francesa       1.800000
      Inglesa       2.150000
     Italiana       2.225000
nacionalidade experiencia_anos
      Alemana        15.000000
        Belga        13.000000
    Espanhola         1.333333
     Francesa        16.000000
      Inglesa        15.500000
     Italiana        17.625000
```

3. Analisando pelo gráfico e pelo coeficiente de correlação de pearson.
   O coeficiente de pearson é uma medida estatística que mede a relação entre duas variáveis distintas através da fórmula:

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i^n (y_i - \overline{y})^2}} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \cdot \sigma_y}$$

   Onde o termo de cima seria a covariância de x e y (a variância de x é um caso especial de cov(x,y) onde x=y). O valor de r varia de 1 a -1, com 1 significando uma forte correlação positiva, 0 não ha correlação alguma e -1 uma forte correlação negativa.

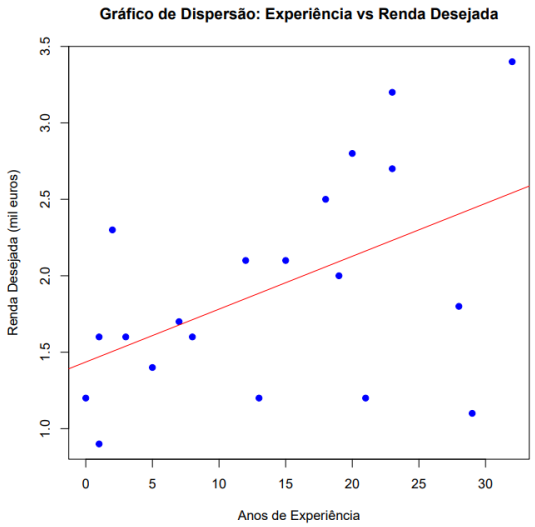

Figura 2: Gráfico de dispersão Renda x Expêriencia

Pelo gráfico observamos uma tendencia do aumento da renda desejada conforme os anos de experiencia. Contudo não é uma relação muito clara por ter algumas amostras discrepantes.
Através da fórmula do coeficiente de correlação:

$$r = \frac{3.47}{0.7 \cdot 10} = 0,5$$

Que confirma que as duas informações estão relacionadas positivamente.

Agora, tratando do código em R, a Correlação de Pearson é feita nativamente, ou seja, basta usar a função `cor()` com o parâmetro `"pearson"`. No mais, foi necessário escolher os dados de interesse:

```
correlacao_pearson <- cor(dados$experiencia_anos, dados$renda_desejada, method = "pearson")
correlacao_pearson
```

Para a construção do gráfico, foi o mesmo processual da primeira questão, somente com a adição de uma reta que traça a regressão linear entre os dados de renda desejada e experiência, facilitando a visualizar a correlação entre os dados:

```
pdf("graficos/graficos_questao2.pdf")

plot(dados$experiencia_anos, dados$renda_desejada,
     main = "Grafico de Dispersao: Experiencia vs Renda Desejada",
     xlab = "Anos de Experiencia",
     ylab = "Renda Desejada (mil euros)",
     pch = 19,
     col = "blue"
)

abline(lm(renda_desejada ~ experiencia_anos, data = dados), col = "red")
```

4. É necessário reconstruir a tabela cortando os funcionários que: possuem menos de 10 anos de experiência e que desejem renda igual ou acima de 2 mil euros.

	Idade	Nacionalidade	Renda	Experiência
1	46	Belga	1.2	21
2	51	Francesa	1.8	28
3	39	Italiana	1.2	13
4	52	Italiana	1.1	29

Tabela 13: Candidatos Condicionados

No código, seguiu-se a mesma lógica:

```
candidatos_priorizados <- dados[dados$renda_desejada < 2 &
   dados$experiencia_anos >= 10,]
candidatos_priorizados
```

Os candidatos foram selecionados dentre os dados com base nas condições desejadas: renda abaixo de 2 mil euros e 10 anos ou mais de experiência:

```
idade nacionalidade renda_desejada experiencia_anos
   46          Belga            1.2               21
   51       Francesa            1.8               28
   39       Italiana            1.2               13
   52       Italiana            1.1               29
```

5. Analisando os gráficos individualmente.

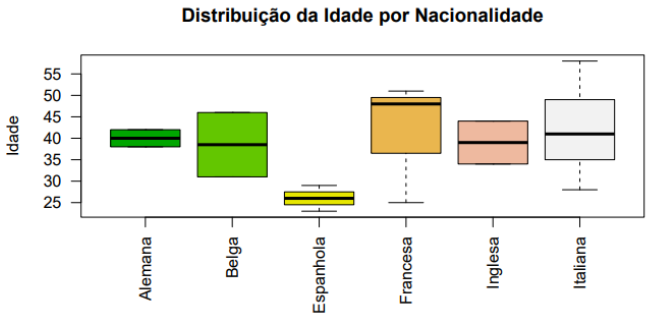

Figura 3: Gráfico idade por nacionalidade

Analisando cada nacionalidade é observado que:

- Os candidatos mais novos são espanhóis.
- Os italianos possuem maior dispersão em relação a idade, enquanto os alemães são os que possuem menos.
- Todos estão bem distribuídos em relação a idade, menos os franceses onde seus candidatos se concentram em idades mais altas, denotado pela mediana próxima do topo do boxplot.

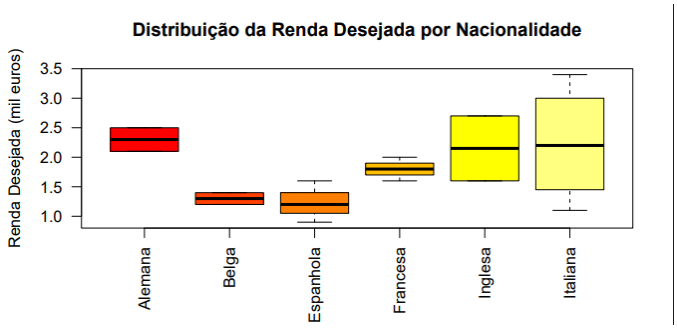

Figura 4: Gráfico renda por nacionalidade

Analisando cada nacionalidade é observado que:

- Os candidatos mais dispersos são os italianos.
- Os alemães possuem o piso salarial mais alto.
- Em todos os valores estão bem distribuídos, conforme a linha da mediana no centro dos boxplots.
- De todos os espanhóis são o com piso salarial mais baixo.

Novamente, o gráfico dessa questão não diverge do código das anteriores:

```
par(mfrow = c(2, 1))
boxplot(idade ~ nacionalidade,
        data = dados,
        main = "Distribuicao da Idade por Nacionalidade",
        xlab = "",
        ylab = "Idade",
        col = terrain.colors(6),
        las = 2
)

boxplot(renda_desejada ~ nacionalidade,
        data = dados,
        main = "Distribuicao da Renda Desejada por
           Nacionalidade",
        xlab = "",
        ylab = "Renda Desejada (mil euros)",
        col = heat.colors(6),
        las = 2
)
dev.off()
```

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 14. 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 14: 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.

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.

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?

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?

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

Solução da questão 3

Por se tratar de um volume maior de dados, a resolução dessa questão foi feita somente por meio de código em R. Em compensação, haverá uma discussão mais elaborada em relação ao código.

1. Primeiramente, todos os dados presentes no **.csv** foram carregados no código por meio da função:

```
dados <- read.csv("dados/HW1_bike_sharing.csv")
```

   Como o arquivo em questão não apresenta falhas de formatação, nenhum tratamento foi necessário, nem mesmo em relação a forma que os dados foram separados(com vírgula), como os decimais foram representados(com ponto final) ou sobre a posição do header(cabeçalho). Ou seja, bastou a utilização direta desse método.

   Para retornar o número total de observações, basta retornar o número de linhas que o *dataframe* tem:

```
numero_total_de_observacoes <-nrow(dados)
numero_total_de_observacoes
```

   Assim como observado ao abrir o arquivo pelo Excel, tem-se um total de **731 observações**.

   Logo após, utilizou-se as funções `min()` e `max()` na coluna de datas:

```
inicio_da_amostra <- min(dados$dteday)
inicio_da_amostra

fim_da_amostra <- max(dados$dteday)
fim_da_amostra
```

   Retornando:

```
[1] "2011-01-01"
[1] "2012-12-31"
```

   Determinando que a observação foi feita entre **01 de Janeiro de 2011** até **31 de Dezembro de 2012**, cerca de 02 anos.

   Agora, sobre as variáveis:

   Consideramos **Numérico** como aqueles valores que podem ser ordenados e somados(de maneira razoável). E o **Categórico** será considerado como aquilo que não é Numérico. Dessa forma, temos a seguinte distinção entre as *tags*:

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

Tabela 15: Variáveis do conjunto `HW1_bike_sharing` separadas por tipo.

Apesar de ser uma variável aparentemente numérica, `instant` foi considerada **categórica** por não existir uma razão lógica em somar, subtrair ou executar as demais possíveis operações sob esta variável.

2. Dentre as variáveis numéricas presentes na amostra, considera-se relevante para a análise os números de usuários casuais e registrados, além da temperatura, visto que é possível traçar um paralelo entre a medida de temperatura e número de usuários totais no dia, apesar de incerto:

```
variaveis_relevantes <- dados[c("temp", "casual", "registered"
    )]
```

Agora, afim de calcular os dados de tendência central e dispersão dessa amostra, utilizou-se a função `summary()` que calcula e organiza em um dataframe as informações: quartis, média e amplitude. Dessa forma, disponibilizando de maneira prática dados importante para discussão:

```
resumo_estatistico <- summary(variaveis_relevantes)
resumo_estatistico
```

Retornando:

```
      temp           casual         registered
 Min.   : 2.40   Min.   :   2.0   Min.   :  20
 1st Qu.:13.80   1st Qu.: 315.5   1st Qu.:2497
 Median :20.40   Median : 713.0   Median :3662
 Mean   :20.31   Mean   : 848.2   Mean   :3656
 3rd Qu.:26.90   3rd Qu.:1096.0   3rd Qu.:4776
 Max.   :35.30   Max.   :3410.0   Max.   :6946
```

Em geral, podemos fazer a seguinte análise:

- A variável de temperatura apresenta uma amplitude relativamente alta, de 2.4 °C a 35.9 °C, que é o esperado, pois as amostras foram coletadas em todas as estações. Apesar disso, existe uma distribuição simétrica dos dados, visto que a média e mediana são bastante próximas. Sobre o Intervalo Interquartil(**Q1 - Q3 = 13.1 °C**), podemos considerá-la mediana, cerca de um terço do valor da amplitude.

- A variável de usuários casuais também é bem ampla, desde a casa das unidades, até a casa dos milhares. Mas, apesar disso, a mediana e média são relativamente próximas, mostrando dados levemente assimétricos. Ademais, uma média superior ao valor da mediana indica presença de outliers. O intervalo interquartil (**IQR = 780,5**) mostra que a metade central dos dados está concentrada em uma faixa menor que a amplitude total, indicando que a maior parte das observações se encontra em valores mais baixos.

- Seguindo a tendência, os valores de usuários registrado é bastante ampla, entre 20 e 6946 usuários. E, novamente, a mediana e média são bem próximas, mostrando simetria dos dados. Sobre o Intervalo Interquartil(**IQR = 2279**), concluí-se uma baixa acomodação das amostras em posições centrais.

3. Como os números de 1 a 4 foram utilizados para representar tanto as estações, como as condições metereológicas. Usamos a função `factor()` para atrelar o valor quantitativo da amostra ao seu respectivo qualitativo:

```
dados$season <- factor(dados$season,
                       levels = c(1, 2, 3, 4),
                       labels = c("Primavera", "Verao", "
                          Outono", "Inverno")
                  )
dados$weathersit <- factor(dados$weathersit,
                       levels = c(1, 2, 3, 4),
                       labels = c("Ceu␣Limpo", "Nublado", "
                          Chuva␣Fraca", "Chuva␣Forte")
                  )
```

Agora, afim de visualizar a estação com maior número de usuários, criou-se uma coluna com o número total de usuários por dia de observação:

```
dados$total_users <- dados$casual + dados$registered
```

Ademais, usando o `aggregate()`, atrela-se o número total de usuários a cada estação, fazendo a soma algébrica de todos os dados de certa estação:

```
total_por_estacao <- aggregate(total_users ~ season,
                               data = dados,
                               FUN = sum)
total_por_condicao <- aggregate(total_users ~ weathersit,
                               data = dados,
                               FUN = sum)
```

Adiante, novamente se inicia um **.pdf** e se utiliza das funções nativas para plotar os gráficos de barra:

```r
pdf("graficos/graficos_questao3.pdf")
par(mfrow = c(2, 1))
barplot(total_por_estacao$total_users,
        names.arg = total_por_estacao$season,
        main = "Total de Usuarios por Estacao do Ano",
        ylab = "Total de usuarios",
        xlab = "Estacoes",
        col = c("lightgreen", "yellow", "orange", "lightblue")
        )

barplot(total_por_condicao$total_users,
        names.arg = total_por_condicao$weathersit,
        main = "Total de Usuarios por Condicao Metereologica",
        ylab = "Total de usuarios",
        xlab = "Condicoes metereologicas",
        col = c("lightgreen", "yellow", "orange", "lightblue")
        )
```

Com isso, obtem-se os seguintes gráficos:

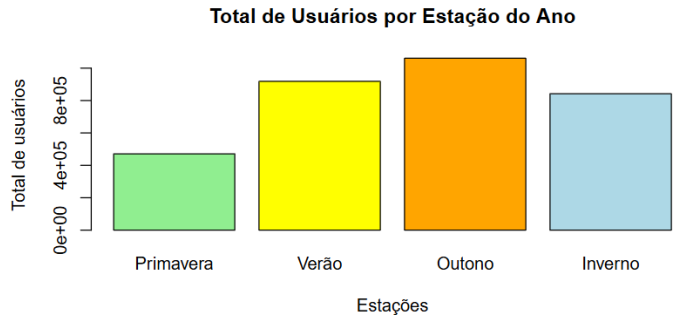

Figura 5: Total de usuários por Estação do Ano

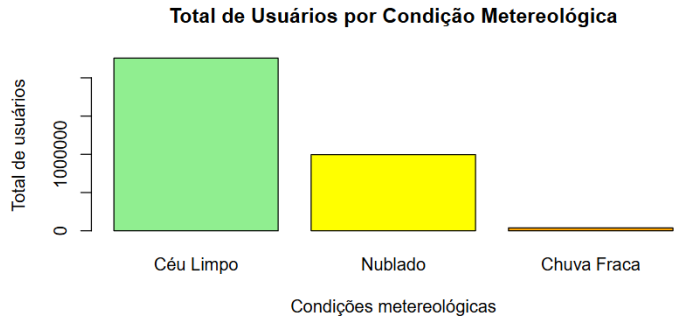

Figura 6: Total de usuários por Condição Metereológica

Destarte, concluí-se:

- A estação do ano com maior número de usuários é o outono.
- O uso de bicicletas parece pouco influenciado pela estação do ano, uma vez que os números de usuários no Verão, Outono e Inverno são relativamente próximos. Curiosamente, a Primavera, período em que seria esperado um aumento no número de usuários em relação ao Inverno, apresenta o menor número de registros, reforçando a ideia de que a estação não é um fator determinante para a utilização das bicicletas.
- A condição climática mais favorável é "Céu Limpo".

4. Apesar que o *HomeWork* tenha dito para multiplicar os valores da temperatura por 41, não foi considerado necessário, visto que as temperaturas obtidas no **.csv** variam de **2.4 a 35.9 °C**, valores totalmente plausíveis para uma região temperada.

   Para o plot dos valores, foi necessário transformar as datas em uma variável específica do R:

```
dados$dteday <- as.Date(dados$dteday)
```

   Agora, basta plotar as séries temporais desejadas:

```
par(mar = c(5, 4, 4, 5) + 0.1)
plot(dados$dteday, dados$total_users,
     type = "l",
     col = "blue",
     xlab = "Data",
     ylab = "Total de Usu rios",
     main = "Usu rios Totais ao Longo do Tempo"
)

plot(dados$dteday, dados$temp,
     type = "l",
     col = "red",
     xlab = "Data",
     ylab = "Temperatura",
     main = "Temperatura ao Longo do Tempo"
)
```

Chegando no seguinte resultado:

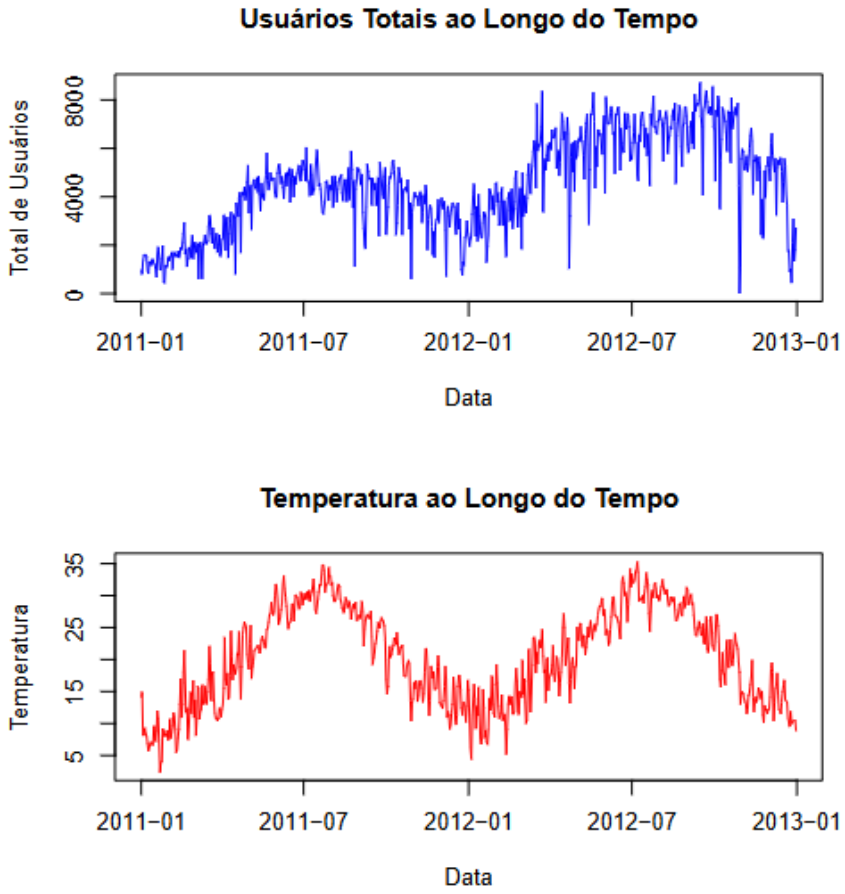

Figura 7: Relação entre temperatura e Número de Usuários

Comparando ambos os gráficos, percebe-se uma relação diretamente proporcional entre a temperatura e usuários totais. Existem alguns valores que não seguem essa relação, mas são justificadas pelo fato de serem dias chuvosos ou nublados.