Homework I

Nome completo: Lorena de Carvalho Gonçalves e Maria Lissa Rodrigues Costa

Numero de matricula: 567096 e 565974

LINK REPOSITÓRIO GIT

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 gás 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

1. As medidas de tendência central são média, moda e mediana. Para média, temos que somar todos os dados e dividir pelo número total de elementos. A fórmula matemática para a média é:

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

A somatória de todos os dados é:

$$\sum_{i=1}^{80} x_i = 1521.7$$

Sabendo que temos 80 elementos, aplicamos os valores à fórmula:

$$media = \frac{1521.7}{80}$$

Portanto, o valor final da média é:

$$media = 19.02125$$

Para encontrar a mediana, é necessário a organização dos dados em ordem crescente ou decrescente. Nessa questão vamos fazer uma organização crescente:

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

Tabela 2: Emissões diárias de gás poluente em ordem crescente.

Sabe-se que a mediana é o elemento do meio, e, como temos dois elementos no meio, 40 e 41, fazemos a média dos dados dessas posições:

$$mediana = \frac{19.1 + 19.2}{2} = 19.15$$

Em seguida, pela análise do gráfico, descobrimos a moda, ou seja, o valor com maior frequência:

$$moda = 19.4$$

Após isso, vamos descobrir as medidas de dispersão. Em relação a amplitude, é só subtrair o maior elemento pelo menor, como já colocamos a tabela em ordem crescente, descobrimos facilmente esse valor:

$$amplitude = 31.8 - 6.2 = 25.6$$

A variância amostral mede a dispersão dos dados em relação à média. A fórmula é:

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

Onde a soma dos quadrados das diferenças ($\sum_{i=1}^{n}(x_i - \bar{x})^2$) é aproximadamente 2436.4739 e o número de amostras $n = 80$. Substituindo os valores na fórmula:

$$s^2 = \frac{2436.4739}{80-1} = \frac{2436.4739}{79}$$

O resultado da variância amostral é:

$$s^2 \approx 30.8414$$

O desvio padrão é a raiz quadrada da variância, indicando o grau de dispersão na mesma unidade dos dados. A fórmula é:

$$s = \sqrt{s^2}$$

Utilizando o valor da variância calculado anteriormente:

$$s = \sqrt{30.8414}$$

O resultado do desvio padrão amostral é:

$$s \approx 5.5535$$

O coeficiente de variação é uma medida de dispersão relativa, expressa em porcentagem, que permite comparar a variabilidade entre diferentes conjuntos de dados. A fórmula é:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Usando os valores do desvio padrão ($s \approx 5.5535$) e da média ($\bar{x} \approx 19.0213$):

$$CV = \frac{5.5535}{19.0213} \times 100\%$$

O resultado do coeficiente de variação é:

$$CV \approx 29.20\%$$

Tabela 3: Medidas Estatísticas Descritivas

Medida	**Valor**
Medidas de Tendência Central	
Média	19.02125
Mediana	19.15
Moda	19.4
Medidas de Dispersão	
Amplitude	25.6
Variância	30.8414
Desvio Padrão	5.5535
Coeficiente de Variação	29.20%

Em relação aos valores de Tendência Central, sabe-se que servem para resumir os dados em valores típicos. Nota-se que esses valores estão em torno de 19, o que se assemelha com os dados recebidos.

Em relação aos valores de dispersão, eles ajudam a entender o espalhamento dos dados, se referindo a quantidade de variação desses deles. Pelo, Coeficiente de Variação, percebe-se, que ele está moderadamente disperso. Ademais, observa-se que existe uma variação considerável dos dados, mediante a análise das outras medidas de dispersão.

Listado 1: Código para questão 1, item 1.

```
daily_emissions <- 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)

media <- mean(daily_emissions)
cat("m dia:", media, "\n")

mediana <- median(daily_emissions)
cat("mediana:", mediana, "\n")

tabela_freq <- table(daily_emissions)
moda <- names(tabela_freq)[tabela_freq == max(tabela_freq)]
cat("moda:", moda, "\n")

range <- max(daily_emissions) - min(daily_emissions)
cat("amplitude:", range, "\n")

sd <- sd(daily_emissions)
cat("desvio␣padr o:", sd, "\n")
```

```r
var <- var(daily_emissions)
cat("vari ncia:", var, "\n")

coef_variation <- (sd(daily_emissions) / mean(daily_emissions)) * 100
cat("coeficiente de varia   o:", coef_variation, "\n")
```

2.

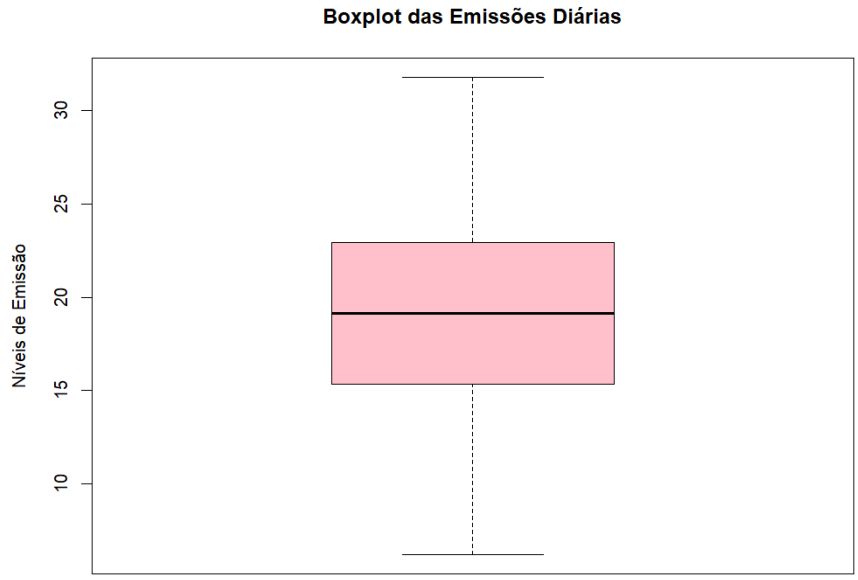

Figura 1: Boxplot para questão 1.

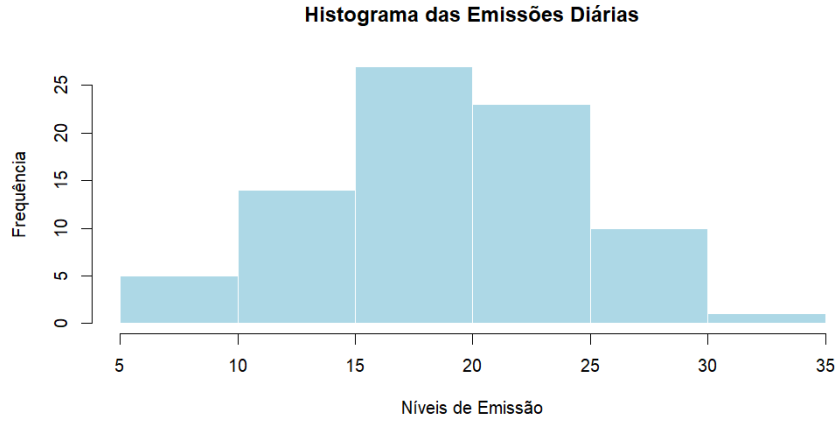

Figura 2: Histograma para questão 1.

Pelo Boxplox, nota-se que está bem simetrico, a mediana e os quartis estão bem distintos. Além disso, nota-se que não existem "outliners", ou seja, dados extremos que não alteram as condições dos quartis, todos os dados estão dentro do escopo dos bigodes.

Em relação ao histograma, também temos uma simetria, os valores centrais (15 até 25) são os de maiores frequências. Diante disso, observa-se que os valores mais à direita e à esquerda são os de menores frequência.

Listado 2: Código para questão 1, item 2

```
daily_emissions <- 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)
# N o esque a de aumentar o Painel de Gr ficos para caberem os plots
boxplot(daily_emissions,
        main = "Boxplot das Emiss es Di rias",
        ylab = "N veis de Emiss o",
        col = "pink")

hist(daily_emissions,
     main = "Histograma das Emiss es Di rias",
     xlab = "N veis de Emiss o",
     ylab = "Frequ ncia",
     col = "lightblue",
     border = "white")
```

3. Sabe-se que os quartis são os valores que dividem os dados, em ordem crescente, em quatro partes. No item 1, já encontramos o Q_2, visto que é a mediana. Assim,

$Q_2 = 19.15$. Para encontrar a posição dos quartis Q_1 e Q_3, n sendo o número de dados:

$$Q_1pos = \frac{n+1}{4} = \frac{80+1}{4} = 20.25$$

$$Q_3pos = \frac{3 \cdot (n+1)}{4} = \frac{3 \cdot (80+1)}{4} = 60.75$$

Como as posições deram valores não-inteiros, utiliza-se os dois números inteiros que vêm antes e depois do valor decimal. Logo, os valores de quartis serão a média dos dois valores dessas posições. Portanto, a posição de Q_1 é 20 e 21 e de Q_3 é 60 e 61:

$$Q_1 = \frac{15.2 + 15.5}{2} = 15.35$$

$$Q_3 = \frac{22.9 + 23.0}{2} = 22.95$$

Intervalo interquartil representa a diferença entre Q_3 e Q_1. Assim:

$$IQR = Q_3 - Q_1 = 22.95 - 15.35 = 7.6$$

Sabe-se que valores atípicos são valores que estão fora do escopo dos bigodes de um gráfico boxplot. Esses intervalos são $1.5 \cdot IQR$ acima de Q_3 e abaixo de Q_1. Assim:

$$Q_1 - 1.5 \cdot IQR = 15.35 - 1.5 \cdot 7.6 = 3.95$$

Não existe dado igual ou menor que 3.95 registrado na questão.

$$Q_3 + 1.5 \cdot IQR = 22.95 + 1.5 \cdot 7.6 = 34.35$$

Não existe dado igual ou maior que 34.35 registrado na questão.

Listado 3: Código para questão 1, item 3

```
daily_emissions <- 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)

quantile(daily_emissions)
cat("\n")

iqr <- IQR(daily_emissions)
cat("intervalo interquartil:", iqr, "\n")
```

4. Em primeira análise, é necessário contar os dias em que os dados estão acima de 25. São 11 dias que esse resultado foi excedido. A proporção em relação aos dias totais:

$$\frac{11}{80} = 0.1375$$

O comportamento geral das emissões estaria em conformidade com esse padrão regulatório, visto que foram poucos dias que excediram 25, apenas 13.75% dos dias totais.

Listado 4: Código para questão 1, item 4

```
daily_emissions <- 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)

v25 <- sum(daily_emissions > 25)

total <- length(daily_emissions)

proporcao <- v25 / total
cat("propor  o de dias que a planta excedeu 25 unidades:", proporcao, "
   \n")
```

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.

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.

	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	Alemã	2.5	18
17	48	Francesa	2.0	19
18	33	Italiana	1.7	7
19	38	Alemã	2.1	12
20	46	Italiana	3.2	23

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

Solução da questão 2

1. Para média, temos que somar todos os dados e dividir pelo número total de elementos. A fórmula matemática para a média é:

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

A somatória de todos os dados das variáveis é:

$$soma\_idade = \sum_{i=1}^{20} x_i = 773$$

$$soma\_renda\_desejada = \sum_{i=1}^{20} x_i = 38,4$$

$$soma\_anos\_experincia = \sum_{i=1}^{20} x_i = 280$$

Sabendo que temos 20 elementos, aplicamos os valores à fórmula:

$$media\_idade = \frac{773}{20} = 38.65$$

$$media\_renda\_desejada = \frac{38,4}{20} = 1.92$$

$$media\_anos\_experincia = \frac{280}{20} = 14$$

Para o cálculo da mediana, é necessária a ordenação crescente dos dados. Isso se mostraria dessa forma:

Idade = 23 25 26 28 29 31 33 34 37 38 39 42 43 44 46 46 48 51 52 58
Renda desejada = 0.9 1.1 1.2 1.2 1.2 1.4 1.6 1.6 1.6 1.7 1.8 2.0 2.1 2.1 2.3 2.5 2.7 2.8 3.2
Anos de experiência = 0 1 1 2 3 5 7 8 12 13 15 18 19 20 21 23 23 28 29 32

Sabe-se que a mediana é o elemento do meio, e, como temos dois elementos no meio, 10 e 11, fazemos a média dos dados dessas posições:

$$mediana\_idade = \frac{38 + 39}{2} = 38.5$$

$$mediana\_renda\_desejada = \frac{1.7 + 1.8}{2} = 1.75$$

$$mediana\_anos\_experiencia = \frac{13 + 15}{2} = 14$$

Para o cálculo do desvio padrão, é necessário calcular a variância de cada conjunto de dados. A variância amostral mede a dispersão dos dados em relação à média. A fórmula é:

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

Temos que $somavar\_idade = (\sum_{i=1}^{n}(x_i - \bar{x})^2) = 1872.55$, $somavar\_renda\_desejada = (\sum_{i=1}^{n}(x_i - \bar{x})^2) = 9.672$, $somavar\_idade = (\sum_{i=1}^{n}(x_i - \bar{x})^2) = 2004$ e o número de amostras $n = 20$. Substituindo os valores na fórmula:

$$s^2\_idade = \frac{1872.55}{20-1} = \frac{1872.55}{19} = 98.55526$$

$$s^2\_renda\_desejada = \frac{9.672}{20-1} = \frac{9.672}{19} = 0.5090526$$

$$s^2\_anos\_experiencia = \frac{2004}{20-1} = \frac{2004}{19} = 105.4737$$

Sabendo que o desvio padrão é a raiz quadrada dos valores da variância, temos que

$$dv\_idade = 9.9275$$

$$dv\_renda\_desejada = 0.7134792$$

$$dv\_anos\_experiencia = 10.27004$$

A partir desses dados obtidos, podemos inferir algumas informações sobre os candidatos. A média e a mediana da idade são bem próximas, respectivamente 38.65 e 38.5

anos, o que evidencia a prevalência de profissionais nessa faixa de idade. Além disso, o desvio padrão = 9.9275 demonstra uma grande variação nos dados, ou seja, amostras de valores variados amplamente. Em relação a renda, a média é maior que a mediana, respectivamente 1.92 e 1.75 mil euros, o que mostra que alguns candidatos tem uma renda desejada mais elevada em relação aos outros. Em relação aos anos de experiência, temos que a média é igual a mediana, ambas 14 anos, podemos observar perfeita simetria na distribuição dos dados, além de uma alta variabilidade, tendo candidatos com muitos anos de experiência ou com muito poucos, dado que o desvio padrão para essa variável é de 10.27004.

Em suma, é possível se observar, sobre o perfil típico dos candidatos, a grande presença de candidatos com idades próximas a 28, e que tem, em média 14 anos de experiência, ou seja, profissionais consolidados, com uma média de 1.92 mil euros de salário desejado.

Listado 5: Código para questão 2, item 1

```
#QUESTAO-02

    #ITEM-1
        #M DIA
            #idade
            i <- c(28, 34, 46, 26, 37, 29, 51, 31, 39, 43, 58, 44, 25,
                23, 52, 42, 48, 33, 38, 46)
            cat("m dia da idade =", mean(i), "\n")

            #renda
            r <- 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)
            cat("m dia da renda =", mean(r), "\n")

            #experiencia
            e <- c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
            cat("m dia da experiencia =", mean(e), "\n")

        #MEDIANA
            #idade
            cat("mediana da idade =", median(i), "\n")

            #renda
            cat("mediana da renda =", median(r), "\n")

            #experiencia
            cat("mediana da experiencia =", median(e), "\n")

        #DESVIO-PADRAO
            #idade
            cat("desvio p. da idade =", sd(i), "\n")

            #renda
            cat("desvio p. da renda =", sd(r), "\n")

            #experiencia
```

```
cat("desvio p. da experiencia =", sd(e), "\n")
```

2. Precisamos agrupar os candidatos por nacionalidade. A seguir, os candidatos de mesma nacionalidade estão agrupados.

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

Para calcularmos a média, como descrevemos no ítem acima, utilizamos a equação:

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

A somatória de todos os dados de renda de cada grupo de nacionalidade é:

$$soma\_renda\_ita = \sum_{i=1}^{8} x_i = 17.8$$

$$soma\_renda\_ing = \sum_{i=1}^{2} x_i = 4.3$$

$$soma\_renda\_bel = \sum_{i=1}^{2} x_i = 2.6$$

$$soma\_renda\_esp = \sum_{i=1}^{3} x_i = 3.7$$

$$soma\_renda\_fra = \sum_{i=1}^{3} x_i = 5.4$$

$$soma\_renda\_ale = \sum_{i=1}^{2} x_i = 4.6$$

A somatória de todos os dados de anos de experiência de cada grupo de nacionalidade é:

$$soma\_exp\_ita = \sum_{i=1}^{8} x_i = 141$$

$$soma\_exp\_ing = \sum_{i=1}^{2} x_i = 31$$

$$soma\_exp\_bel = \sum_{i=1}^{2} x_i = 26$$

$$soma\_exp\_esp = \sum_{i=1}^{3} x_i = 4$$

$$soma\_exp\_fra = \sum_{i=1}^{3} x_i = 48$$

$$soma\_exp\_ale = \sum_{i=1}^{2} x_i = 30$$

Sabendo que temos 20 elementos, aplicamos os valores à fórmula para a média da renda:

$$media\_renda\_ita = \frac{17.8}{8} = 2.225$$

$$media\_renda\_ing = \frac{4.3}{2} = 2.15$$

$$media\_renda\_bel = \frac{2.6}{2} = 1.3$$

$$media\_renda\_esp = \frac{3.7}{3} = 1.233333$$

$$media\_renda\_fra = \frac{5.4}{3} = 1.8$$

$$media\_renda\_ale = \frac{4.6}{2} = 2.3$$

Analogamente, para a experiência

$$media\_exp\_ita = \frac{141}{8} = 17.625$$

$$media\_exp\_ing = \frac{31}{2} = 15.5$$

$$media\_exp\_bel = \frac{26}{2} = 13$$

$$media\_exp\_esp = \frac{4}{3} = 1.333333$$

$$media\_exp\_fra = \frac{48}{3} = 16$$

$$media\_exp\_ale = \frac{30}{2} = 15$$

Analisando os dados de média obtidos para renda média e anos de experiência, podemos perceber que a nacionalidade que apresenta maior renda média desejada é a alemã. Além disso, o grupo que aparenta ser o mais experiente é o grupo dos italianos, já que eles possuem a maior média de anos de experiência dentre os grupos.

Listado 6: Código para questão 2, item 2

```
#QUESTAO-02

    #ITEM-02

        #grupo-italiano

        renda_italiana <- c(2.3,2.1,1.2,2.8,3.4,1.1,1.7,3.2)
        exp_italiana <- c(2,15,13,20,32,29,7,23)
        cat("renda m dia italiana =", mean(renda_italiana), "\n")
        cat("experiencia m dia italiana =", mean(exp_italiana), "\n")

        #grupo-ingles

        renda_inglesa <- c(1.6,2.7)
        exp_inglesa <- c(8,23)
        cat("renda m dia inglesa =", mean(renda_inglesa), "\n")
        cat("experiencia m dia inglesa =", mean(exp_inglesa), "\n")

        #grupo-belga

        renda_belga <- c(1.2,1.4)
        exp_belga <- c(21,5)
        cat("renda m dia belga =", mean(renda_belga), "\n")
        cat("experiencia m dia belga =", mean(exp_belga), "\n")

        #grupo-espanhol

        renda_espanhola <- c(0.9,1.6,1.2)
        exp_espanhola <- c(1,3,0)
        cat("renda m dia espanhola =", mean(renda_espanhola), "\n")
        cat("experiencia m dia espanhola =", mean(exp_espanhola), "\n")

        #grupo-frances
```

```
renda_francesa <- c(1.8,1.6,2.0)
exp_francesa <- c(28,1,19)
cat("renda m dia francesa =", mean(renda_francesa), "\n")
 cat("experiencia m dia francesa =", mean(exp_francesa), "\n")

#grupo-alemao

renda_alema <- c(2.5,2.1)
exp_alema <- c(18,12)
cat("renda m dia alema =", mean(renda_alema), "\n")
cat("experiencia m dia alema =", mean(exp_alema), "\n")
```

3. Para analisar a correlação, existente ou não, entre os anos de experiência e renda desejada, o gráfico de dispersão proporciona uma boa visualização dos dados.

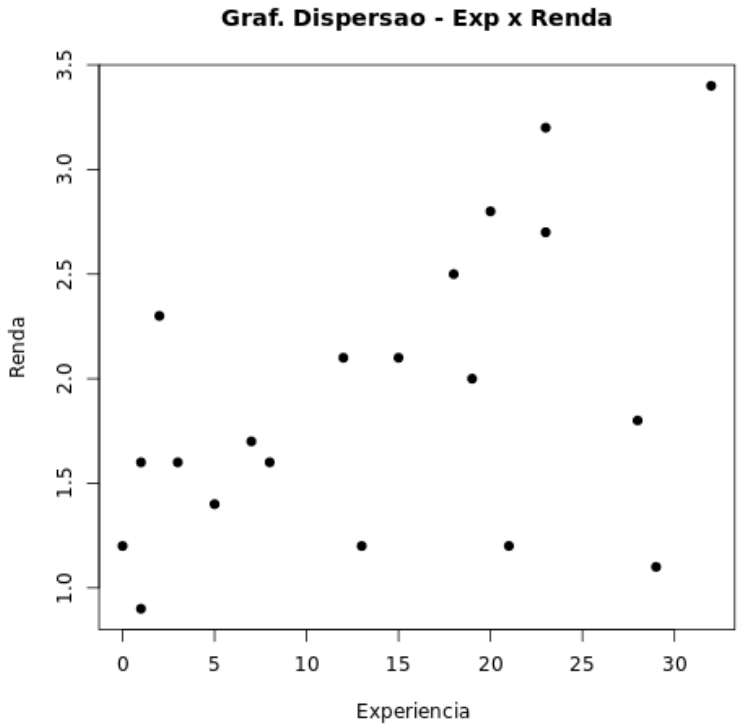

Figura 3: Gráfico de dispersão para questão 2.3

Para calcular o coeficiente de Pearson, temos que

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

equação em que X representa os dados da experiência e Y representa os dados da renda.

Temos, então, que o coeficiente de Pearson (coeficiente de relação linear) = 0.4977672.

Podemos, então, a partir desse valor encontrado e da representação gráfica do gráfico de dispersão, observar que há uma relação, ainda que não absoluta, dos valores de experiência e de renda. No gráfico é possível reconhecer uma quantidade de valores de renda que aumentam em consonância com o aumento da experiência, mas que existem, também, uma quantidade considerável de valores que não seguem completamente esse padrão. Com o cálculo do coeficiente de Pearson, é mais claro a relação entre esses valores, já que ele mostra o grau de relação linear entre essas atribuições. O valor que encontramos foi de 0.4977672, o que representa uma correlação linear moderada, dado que valores aproximados de 0 representam a falta de correlação linear, e valores igual ou maiores que 1 representam grande correlação linear.

Listado 7: Código para questão 2, item 3

```
#QUESTAO-02

    #ITEM-03


        #renda
        r <- 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
        e <- c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)

        #plotagem-graf-dispersao
        plot(e, r, main="Graf. Dispersao - Exp x Renda",
        xlab="Experiencia ", ylab="Renda", pch=19)

        #corr. de Pearson
        cat("coeficiente de corr. Pearson da experiencia e da renda =",
            cor(e, r), "\n")
```

4. Analisando a tabela e os valores de restrição dados, de

$$experincia \geq 10$$

$$renda < 2k$$

em anos e em euros, respectivamente, podemos observar os candidatos de número 3, 7, 9 e 15 na tabela, ou seja, 4 candidatos atendem ambos os critérios, e estão listados abaixo com sua idade e nacionalidade.

46 anos, Belga
51 anos, Francesa
39 anos, Italiana
52 anos, Italiana

5. Para a questão 5, plotamos dois box plots para cada nacionalidade, em relação a idade e a renda de cada grupo, para vermos a distribuição dos dados e compará-los.

Os gráficos estão colocados abaixo. Por alguma razão, mesmo estando escrito e sendo compilado no código, a legenda para o boxplot dos franceses não apareceu no gráfico plotado, então o 5° boxplot de cada imagem é a representação do boxplot dos franceses.

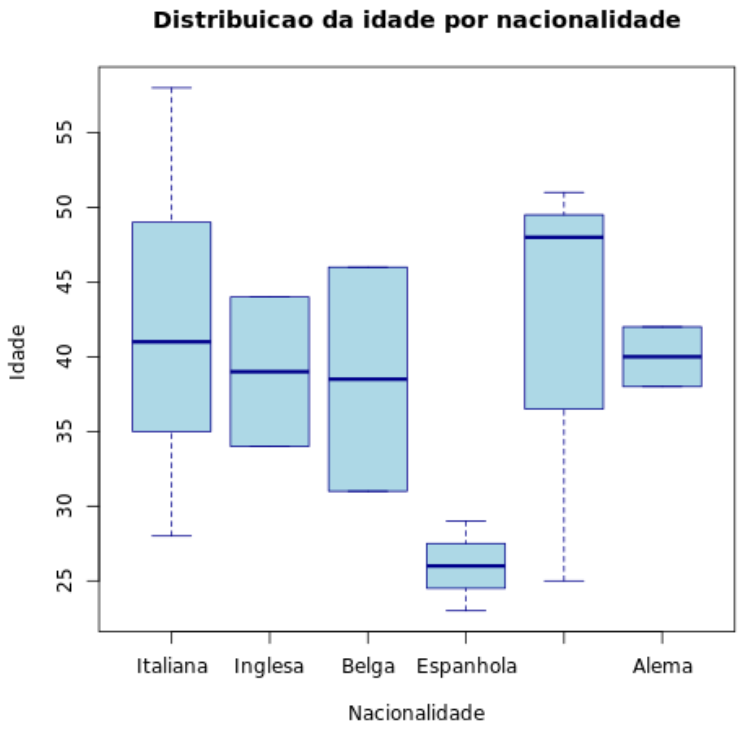

Figura 4: Distribuição da idade por nacionalidade para a questão 2.5

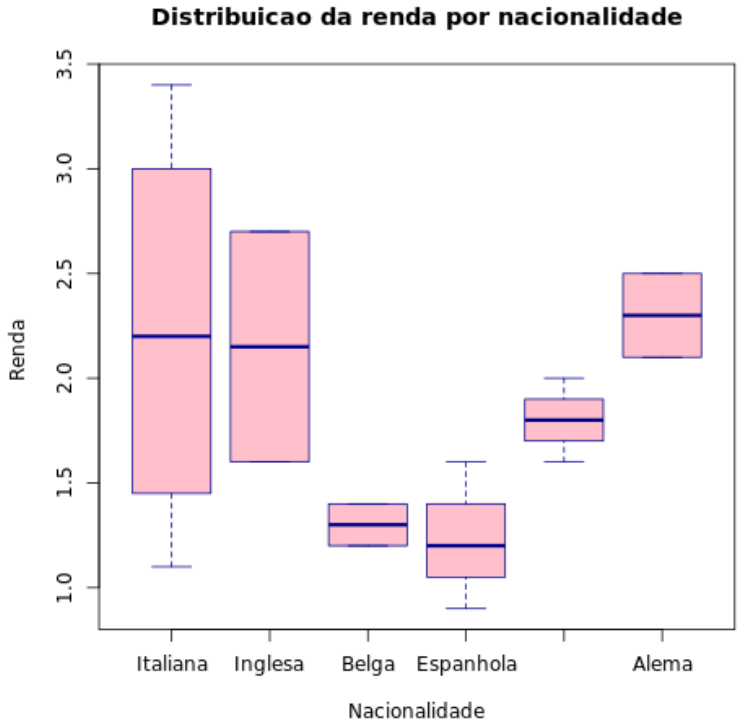

Figura 5: Distribuição da renda por nacionalidade para a questão 2.5

A partir desses gráficos, podemos aferir algumas informações sobre os dados de cada variável.

Sobre a distribuiçãp da idade, os italianos e os belgas possuem os maiores IQR (Interquartile Range) entre os dados mostrados, ou seja, maior amplitude dos valores dos dados coletados. Os espanhóis são visivelmente o grupo mais novo, com idades bem abaixo dos valores das outras nacionalidades. Os franceses tem uma maior dispersão, dado que existe um valor outlier muito menor que o resto dos dados.

Sobre a distribuição da renda desejada, podemos observar novamente a grande amplitude de renda do grupo italiano, há maior variabilidade nos dados. Os belgas e os espanhóis são os que possuem a renda desejada mais baixa, enquanto os belgas e os franceses possuem a renda de valores mais concentrados.

Listado 8: Código para questão 2, item 5

```
#QUESTAO-02

        #ITEM-05
            idade_italiana <- c(28, 37, 39, 43, 58, 52, 33, 46)
            idade_inglesa <- c(34, 44)
```

```r
idade_belga <- c(46, 31)
idade_espanhola <- c(26, 29, 23)
idade_francesa <- c(51, 25, 48)
idade_alema <- c(42, 38)

#boxplot para a idade
boxplot(idade_italiana, idade_inglesa, idade_belga,
        idade_espanhola, idade_francesa, idade_alema,
        names = c("Italiana", "Inglesa", "Belga", "Espanhola"
            , "Francesa", "Alema"),
        main = "Distribuicao da idade por nacionalidade",
        xlab = "Nacionalidade",
        ylab = "Idade ",
        col = "lightblue",
        border = "darkblue")


renda_italiana <- c(2.3,2.1,1.2,2.8,3.4,1.1,1.7,3.2)
renda_inglesa <- c(1.6,2.7)
renda_belga <- c(1.2,1.4)
renda_espanhola <- c(0.9,1.6,1.2)
renda_francesa <- c(1.8,1.6,2.0)
renda_alema <- c(2.5,2.1)


#boxplot para a renda desejada
boxplot(renda_italiana, renda_inglesa, renda_belga,
        renda_espanhola, renda_francesa, renda_alema,
        names = c("Italiana", "Inglesa", "Belga", "Espanhola"
            , "Francesa", "Alema"),
        main = "Distribuicao da renda por nacionalidade",
        xlab = "Nacionalidade",
        ylab = "Idade ",
        col = "pink",
        border = "darkblue")
```

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 5. 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).

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

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

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 5: Variáveis do conjunto `HW1_bike_sharing` (questão 3).

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?

Solução da questão 3

1. LINK DAS AMOSTRAS
Nota-se que existem 7 variáveis: instant, dteday, season, weathersit, temp, casual, e registered. Analisando cada uma delas:

INSTANT: númerica discreta;

DTEDAY: categórica nominal;

SEASON: categórica nominal;

WEATHERSIT: categórica nominal;

TEMP: númerica contínua;

CASUAL: númerica discreta;

REGISTERED: númerica discreta.
Para descobrir o número de observações, basta analisar o indíce, que indica o número de dados. Logo, são 731 observações, entre os dias 01/01/2011 a 31/12/2012.

Listado 9: Código para questão 3, item 1

```
library(tidyverse)

dados_bike <- read_csv("HW1_bike_sharing.csv")

total_observacoes <- nrow(dados_bike)

dados_bike$dteday <- as_date(dados_bike$dteday)

data_inicio <- min(dados_bike$dteday)
data_fim <- max(dados_bike$dteday)

colunas_numericas <- c()
colunas_categoricas <- c()

categorias_em_numero <- c("season", "yr", "mnth", "hr", "holiday", "
    weekday", "workingday", "weathersit")

for (coluna in names(dados_bike)) {

  if (coluna %in% categorias_em_numero) {
    colunas_categoricas <- c(colunas_categoricas, coluna)

  } else if (is.numeric(dados_bike[[coluna]])) {
    colunas_numericas <- c(colunas_numericas, coluna)

  } else {
    colunas_categoricas <- c(colunas_categoricas, coluna)
  }
}


cat("N mero␣Total␣de␣Observa  es:", total_observacoes, "\n\n")

cat("Per odo␣da␣Amostra:\n")
cat("Data␣de␣In cio:", format(data_inicio, "%d/%m/%Y"), "\n")
cat("Data␣de␣Fim:", format(data_fim, "%d/%m/%Y"), "\n\n")

cat("Classifica  o␣das␣Vari veis:\n")
cat("Vari veis␣Num ricas:\n", paste("␣-", colunas_numericas, collapse =
     "\n"), "\n\n")

cat("Vari veis␣Categ ricas:\n", paste("␣-", colunas_categoricas,
    collapse = "\n"), "\n")
```

2. Podemos calcular a média dos 731 dados dispostos de cada variável:

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

Assim, fazemos a somatória dos dados de cada variável e dividimos pelo número total de amostras. Temos:

$$media_{temp} = 20.31$$

$$media_{casual} = 848.2$$

$$media_{registered} = 3656$$

Em relação a mediana e os quartis, podemos descobrir suas determinadas posições:

$$Q_1 pos = \frac{n+1}{4} = \frac{731+1}{4} = 183$$

$$Q_2 pos = \frac{n+1}{2} = \frac{731+1}{2} = 366$$

$$Q_3 pos = \frac{3 \cdot (n+1)}{4} = \frac{3 \cdot (731+1)}{4} = 549$$

Em seguida, organizamos os dados em ordem crescente e descobrimos os valores que estão nessas posições. Temos:

- TEMP:

$$Q_1 = 13.80$$

$$Q_2 = 20.40$$

$$Q_3 = 26.90$$

- CASUAL:

$$Q_1 = 315.5$$

$$Q_2 = 713.0$$

$$Q_3 = 1096.0$$

- REGISTERED:

$$Q_1 = 2497$$

$$Q_2 = 3662$$

$$Q_3 = 4776$$

Tabela 6: Medidas Descritivas para Uso de Bicicletas

Estatística	**temp**	**casual**	**registered**
Mínimo	2.40	2.0	20
1º Quartil (Q1)	13.80	315.5	2497
Mediana	20.40	713.0	3662
Média	20.31	848.2	3656
3º Quartil (Q3)	26.90	1096.0	4776
Máximo	35.30	3410.0	6946

O número de usuários registrados é drasticamente maior do que o de usuários casuais. A média de usuários registrados (3656) é mais de 4 vezes superior à média de usuários casuais (848.2). Isso indica que a base do serviço é sustentada por usuários frequentes e fidelizados.
O uso é consistente e previsível. A média (3656) e a mediana (3662) são quase idênticas. Isso sugere uma distribuição de dados simétrica, típica de um comportamento rotineiro, como pessoas usando as bicicletas para ir ao trabalho ou à universidade todos os dias.
Assim como os usuários registrados, a temperatura também apresenta uma distribuição muito simétrica, com a média (20.31°C) e a mediana (20.40°C) sendo praticamente iguais. Ademais, a maior parte das observações (50% dos dias, entre o 1º e o 3º quartil) ocorreu em uma faixa de temperatura bastante agradável para atividades ao ar livre, entre 13.8°C e 26.9°C.

Listado 10: Código para questão 3, item 2

```
library(tidyverse)
library(knitr)

nome_arquivo <- "HW1_bike_sharing.csv"
dados_bikes <- read_csv(nome_arquivo)


dados_analise <- dados_bikes |>
  select(temp, casual, registered)

sumario_estatistico <- summary(dados_analise)

tabela_resultados <- kable(sumario_estatistico,
                           caption = "Tabela 1: Medidas Descritivas para
                               Uso de Bicicletas",
                           format = "pipe")

print(tabela_resultados)
```

3. A variável season representa as estações do ano no hemisfério norte em números. Podemos atribuir os valores aos níveis (estações) de modo que:

1 - Primavera
2 - Verão
3 - Outono
4 - Inverno

A variável weathersit representa quatro condições metereológicas possíveis, e podemos atribuir os valores de modo que:

1 - Céu limpo
2 - Nublado
3 - Chuva fraca
4 - Chuva forte

Utilizando as funções table e barplot do R, foram plotados os seguintes gráficos de barra para ambas as variáveis em observação:

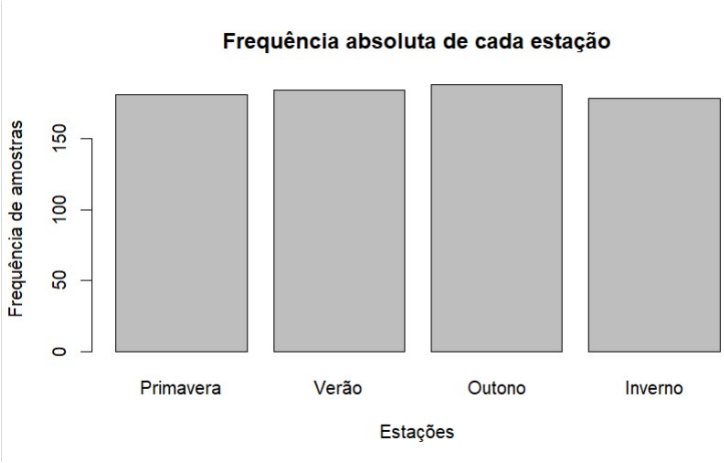

Figura 6: Distribuição das estações para a questão 3.3

No gráfico abaixo, não foi inserido a parte de chuva forte, já que não havia nenhuma amostra com esse valor de weathersit, e o RStudio sinalizava erro na plotagem do gráfico.

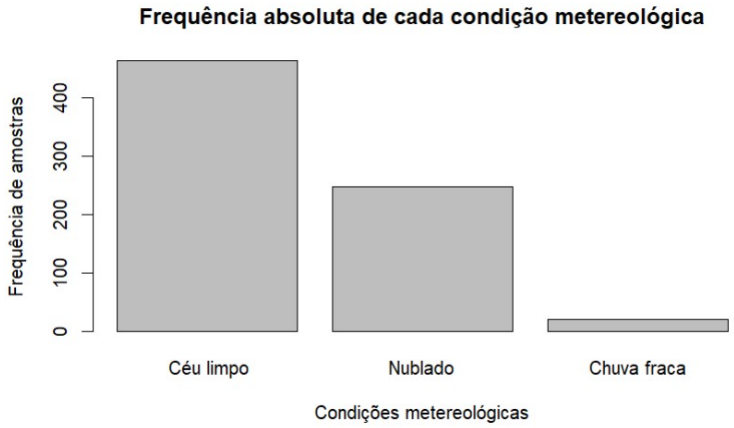

Figura 7: Distribuição das condições metereológicas para a questão 3.3

Após observação dos valores numéricos e do gráfico das estações, podemos observar a semelhança da quantidade de amostras para cada estação do ano. A estação que apresenta o maior número de usuários é o outono, mesmo que as outras estações não se afastem muito do valor desse evento, o que nos leva a acreditar que o uso de biciletas não depende da estação, dado que esses valores para cada estação são muito semelhantes entre si e não demonstram preferência de estação significativa.

Observando o segundo gráfico, nota-se a discrepância visível da quantidade de amostras para cada estação metereológica, com mais da metade das ocorrências acontecendo no clima de céu aberto e nenhuma instância na condição de chuva forte. A partir disso, podemos afirmar que a condição metereológica mais propícia para o uso das bicicletas compartilhadas é a de céu aberto.

Listado 11: Código para questão 3, item 3

```
#QUESTAO-03
    #ITEM-03

      library(tidyverse)

      dados_bike <- read_csv("HW1_bike_sharing.csv")


      #para gerar graf. barra para seasons
      var_seasons <- dados_bike[, 4]

      freq_seasons <- table(var_seasons)

      print(freq_seasons)
```

```
barplot(freq_seasons, main="Frequencia absoluta de cada esta ao",
        xlab="Esta  es",
        ylab="Frequ ncia de amostras",
        names.arg = c("Primavera", "Ver ao", "Outono", "Inverno"))

#para gerar gr f. barra para weathersit
var_weathersit <- dados_bike[, 5]

freq_weathersit <- table(var_weathersit)

print( freq_weathersit)

#nao adicionei o nome chuva forte porque ela nao tem nenhuma
    instancia, entao o rstudio sinalizava erro
barplot(freq_weathersit, main="Frequ ncia absoluta de cada
    condi  o metereol gica",
        xlab="Condi  es metereol gicas",
        ylab="Frequ ncia de amostras",
        names.arg = c(" C u limpo", "Nublado", "Chuva fraca"))
```

4. Juntando os valores de usuários casuais e usuários registrados para cada dia, e transformando a temperatura em valores em modelo real, conseguimos plotar gráficos de série temporal que demonstram as variações de temperatura e de número de usuários por dia. Os gráficos estão mostrados abaixo:

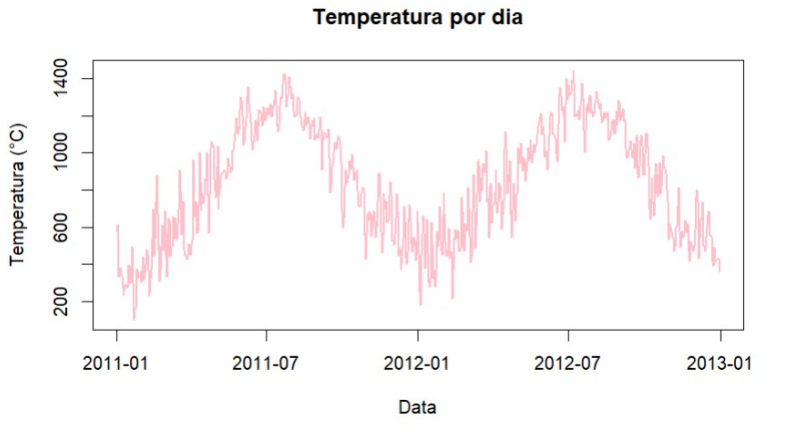

Figura 8: Gráfico de série temporal para a temperatura

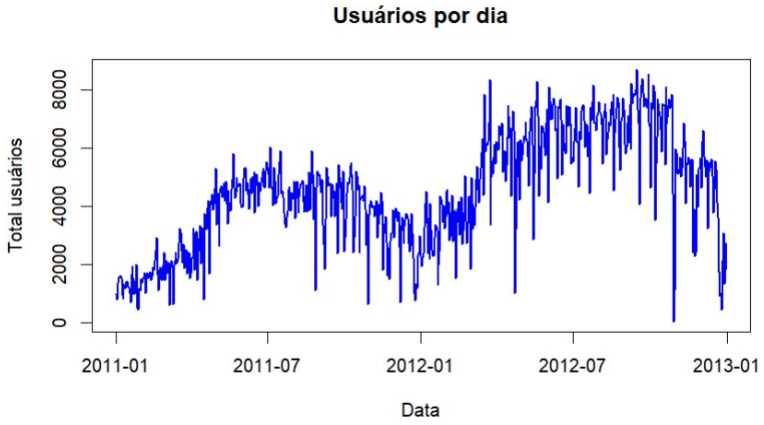

Figura 9: Gráfico de série temporal para o número total de usuários

Analisando o gráfico proposto, é possível observar uma tendência não vigente, mas presente de modo notável, do aumento de usuários das bicicletas em dias em que a temperatura está mais elevada, e a diminuição desse número quando a temperatura cai.

Listado 12: Código para questão 3, item 4

```
#QUESTAO-03
    #ITEM-04

      library(tidyverse)

      dados_bike <- read_csv("HW1_bike_sharing.csv")

      #transforma as colunas em vetor
      usuarios_dia <- as.numeric(unlist(dados_bike[, 7])) + as.numeric(
          unlist(dados_bike[, 8]))
      temp_real <- as.numeric(unlist(dados_bike[, 6])) * 41
      datas <- as.Date(dados_bike[[3]], format = "%Y-%m-%d")

      #gr fico da temperatura
          plot(datas, temp_real, type = "l", col = "pink", lwd = 2,
           main = "Temperatura␣por␣dia", xlab = "Data", ylab = "
               Temperatura␣( C )")
      par(mfrow = c(1, 1))

       #gr fico dos usu rios
      plot(datas, usuarios_dia, type = "l", col = "blue", lwd = 2,
           main = "Usu rios␣por␣dia", xlab = "Data", ylab = "Total␣
               usu rios")
```