Statistics for Engineers
TI0111

Homework I

Nomes completos:

Leonardo Dayvison Silva de Almeida Teixeira

André Venâncio Sousa Grangeiro Filho

Números de matrícula:

566868

564717

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

Item 1, Questão 1

Para o cálculo das medidas centrais, utilizaremos as seguintes equações:

- Média:

$$\bar{x}_n = \frac{1}{N}(x_1 + x_2 + \cdots + x_n) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Figura 1: Cálculo da média

- Mediana:

▷ The sample size N is a odd number

⇝ The sample median is the number in position $\frac{N+1}{2}$.

▷ The sample size N is an even number

⇝ The sample median is the average between $\frac{N}{2}$ and $\frac{N}{2}+1$.

where N is the sample size.

Figura 2: Cálculo da mediana

◦ Desvio padrão:

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

Figura 3: Cálculo do desvio padrão

Vale ressaltar que essa é a variância, mas o desvio padrão é dado por:

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

Item 2, Questão 1

Estes foram os gráficos gerados em R:

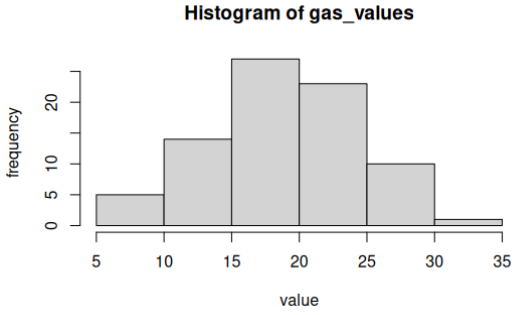

Figura 4: Histograma dividido em bins

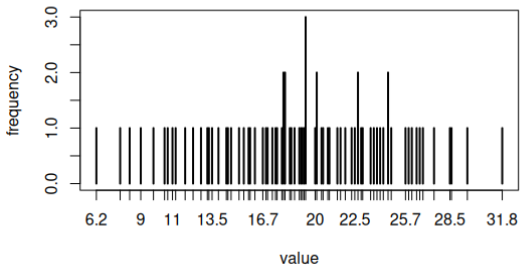

Figura 5: Histograma detalhado

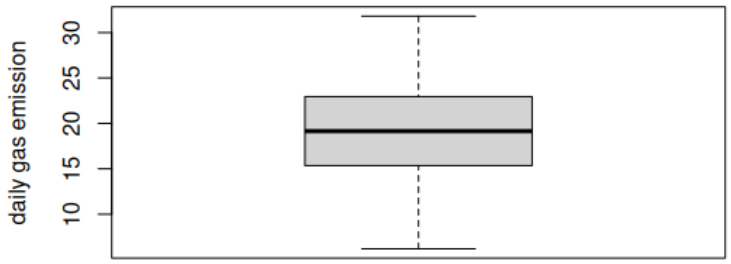

Figura 6: Boxplot

Podemos perceber que os dados são distribuídos simetricamente, concentrados nos valores médios e diminuindo conforme olhamos para os extremos, assemelhando-se a uma curva gaussiana nos histogramas e mostrando um formato simétrico nas hastes do boxplot. Olhando para o boxplot, podemos concluir que não obtemos valores atípicos, visto que não há indicadores de outliers, ou seja, amostras que fogem dos limites de suas hastes. Ou seja, os dados são simétricos e não apresentam valores atípicos.

Listado 1: Solution of item 2, exercise 1

```
# histogram
hist(gas_values, xlab="value", ylab="frequency") # for better
    visualization
plot(freq_table, xlab="value", ylab="frequency") # more detailed

# boxplot
boxplot(gas_values, ylab="daily␣gas␣emission")
```

Item 3, Questão 1

Para calcularmos os quartis, vamos utilizar as seguintes equações: (omitimos o segundo quartil, pois este terá o mesmo cálculo da mediana).

- Primeiro quartil:

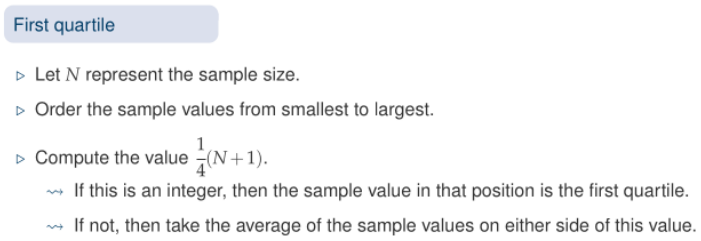

Figura 7: Cálculo do primeiro quartil

- Terceiro quartil:

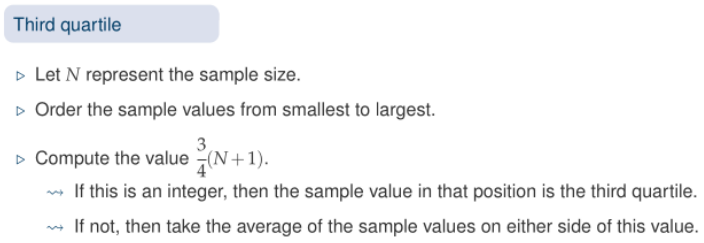

Figura 8: Cálculo do terceiro quartil

Item 4 ,Questão 1

Em nosso dataset, temos um total de 11 medições em que a emissão de gases obtida estava acima do limite estipulado pela questão. Como a maioria das medições não ultrapassou esse limite, o comportamento geral está em conformidade com o padrão regulatório; porém, idealmente, essa ocorrência deveria ser minimizada, o que faz compensar uma análise mais detalhada do que poderia ter acontecido em tais dias.

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 2 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 2: 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

Item 1, Questão 2

Para cada variável listada no item 1, calcularemos alguns valores, dados pelas seguintes equações:

Para a variável idade:

- Média:

$$\mu = \frac{28 + 34 + 46 + 26 + ... + 38 + 46}{20} = 38.65 \tag{1}$$

- Mediana:

$$\frac{20 + 1}{2} = 10.5 \tag{2}$$

Ordenando o vetor de idades usando R, podemos verificar que a mediana é a seguinte média:

- Desvio-padrão:

$$\begin{aligned} \sigma^2 &= \frac{\sum(x_i - \mu)^2}{N} \\ \sigma^2 &= \frac{(28 - 38.65)^2 + (34 - 38.65)^2 + \cdots + (46 - 38.65)^2}{20} \\ \sigma^2 &= \frac{1872.95}{20} \\ \sigma^2 &= 93.6475 \\ \sigma &= \sqrt{93.6475} \\ \sigma &\approx 9.677 \end{aligned} \tag{3}$$

Importante ressaltar que, tanto aqui, quanto no R, utilizamos a variância populacional, dado que utilizamos toda a população de dados adquirida.

Para a variável Renda iremos realizar cálculos análogos aos feitos com a variável anterior:

- Média:

$$\mu = \frac{2.3 + 1.6 + 1.2 + 0.9 + ... + 2.1 + 3.2}{20} = 1.92 \tag{4}$$

- Mediana: Analogamente à variável anterior:

- Desvio-padrão:

$$\begin{aligned}\sigma^2 &= \frac{(2.3-1.92)^2+(1.6-1.92)^2+\cdots+(3.2-1.92)^2}{20}\\ \sigma^2 &= \frac{9.672}{20}\\ \sigma^2 &= 0.4836\\ \sigma &= \sqrt{0.4836}\\ \sigma &\approx 0.695\end{aligned} \tag{5}$$

O mesmo processo para Experiência:

- Média:

$$\mu = \frac{2+8+21+1+...+12+23}{20} = 14 \tag{6}$$

- Mediana: Analogamente à variável anterior:

- Desvio-padrão:

$$\begin{aligned}\sigma^2 &= \frac{(2-14)^2+(8-14)^2+\cdots+(23-14)^2}{20}\\ \sigma^2 &= \frac{2004}{20}\\ \sigma^2 &= 100.2\\ \sigma &= \sqrt{100.2}\\ \sigma &\approx 10.010\end{aligned} \tag{7}$$

Podemos verificar que há candidatos de idades muito variáveis. É intuitivo que, com a grande amplitude de idades, há grande amplitude de experiência, dado que pessoas mais velhas terão mais experiência de trabalho, mas os dados também comprovam esse fato através do alto desvio padrão. Entretanto, as rendas são bastante similares, estando em torno de 2 mil euros.

Listado 2: Solution of item 1, exercise 2

```
media_idade <- mean(idade)
media_renda <- mean(renda)
media_experiencia <- mean(experiencia)

mediana_idade <- median(idade)
mediana_renda <- median(renda)
mediana_experiencia <- median(experiencia)

dp_idade <- sd(idade)
dp_renda <- sd(renda)
dp_experiencia <- sd(experiencia)
```

Item 2, Questão 2

Agora, vamos agrupar os candidatos (seus índices) por nacionalidade:
Italiana = {1, 5, 9, 10, 11, 15, 18, 20}
Inglesa = {2, 12}
Belga = {3, 8}
Espanhola = {4, 6, 14}
Francesa = {7, 13, 17}
Alemã = {16, 19}

Calculando a renda média desejada para cada nacionalidade:

- Italiana:
$$\mu = \frac{2.3 + 2.1 + 1.2 + 2.8 + 3.4 + 1.1 + 1.7 + 3.2}{8} = 2.225 \tag{8}$$
- Inglesa:
$$\mu = \frac{1.6 + 2.7}{2} = 2.15 \tag{9}$$
- Belga:
$$\mu = \frac{1.2 + 1.4}{2} = 1.3 \tag{10}$$
- Espanhola:
$$\mu = \frac{0.9 + 1.6 + 1.2}{3} = 1.23 \tag{11}$$
- Francesa:
$$\mu = \frac{1.8 + 1.6 + 2.0}{3} = 1.86 \tag{12}$$
- Alemã:
$$\mu = \frac{2.5 + 2.1}{2} = 2.3 \tag{13}$$

Podemos perceber que, de acordo com as médias calculadas, a nacionalidade alemã é a que espera maior renda. Entretanto, a italiana está bem próxima e, por ter mais amostras, o que pode torná-la mais representativa, podemos concluir que a italiana é mais "exigente"nessa pesquisa.
Para a experiência média, por questão de brevidade, iremos omitir os cálculos, que serão análogos aos feitos anteriormente.

- Italiana: 17.625
- Inglesa: 15.5
- Belga: 13
- Espanhola: 1.33
- Francesa: 16

◦ Alemã: 15

Podemos, facilmente, concluir que a nacionalidade Italiana é a mais experiente.

Listado 3: Solution of item 2, exercise 2

```
df <- data.frame(idade, nacionalidade, renda, experiencia)

nac_media_renda <- tapply(df$renda, df$nacionalidade, mean)
nac_media_experiencia <- tapply(df$experiencia, df$nacionalidade, mean)

i_maior_renda <- which.max(nac_media_renda)
nac_maior_renda <- names(nac_media_renda[i_maior_renda])
print(paste("Nacionalidade com maior renda:", nac_maior_renda))

i_maior_experiencia <- which.max(nac_media_experiencia)
nac_maior_experiencia <- names(nac_media_experiencia[i_maior_experiencia
    ])
print(paste("Nacionalidade com maior experiencia:", nac_maior_experiencia
    ))
```

Item 3, Questão 2

Observe o gráfico de dispersão que relaciona a renda desejada com a experiência dos candidatos: Podemos observar que parece haver certa correlação, visto que valores de

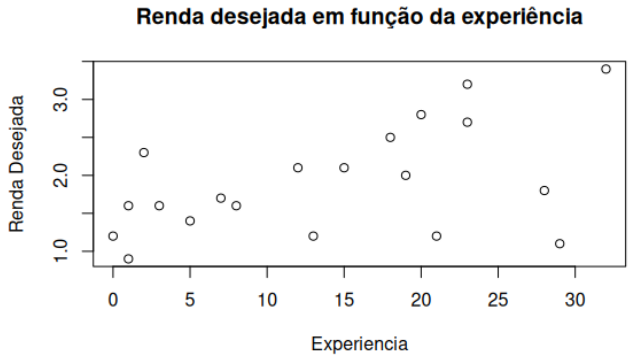

Figura 9: Gráfico de dispersão renda x experiência

renda mais altos encontram-se em candidatos com experiência também mais avançada. Utilizando a ferramenta do R, podemos calcular o coeficiente de Pearson, que mede a correlação linear entre duas variáveis, resultando em um número entre 0 e 1, sendo que se -1 $<$ x $<$ 0, há relação inversa, se 0 $<$ x $<$ 1, há relação direta e, se x $=$ 0, não há correlação linear. Com isso, descobrimos que o coeficiente de Pearson das variáveis em questão é 0.4977672. Isso nos confirma que, ainda que não seja completamente direta, as duas variáveis estão, sim, correlacionadas. Com isso, concluimos que candidatos com mais anos de experiência tendem a esperar uma renda maior.

Listado 4: Solution of item 3, exercise 2

```
pearson <- cor(df$renda, df$experiencia)
print(paste("Coeficiente de correlacao de Pearson entre experiencia e
    renda:", pearson))
plot(df$experiencia, df$renda, main = "Renda desejada em funcao da
    experiencia", xlab = "Experiencia", ylab = "Renda Desejada")
```

Item 4, Questão 2

Os candidatos que cumpram os requisitos descritos no enunciado somam 4 candidatos. Seus respectivos pares "Nacionalidade x Idade"são:

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

Listado 5: Solution of item 4, exercise 2

```
selecionados <- subset(df, df$renda < 2 & df$experiencia >= 10)

num_selecionados <- nrow(selecionados)
print(paste("Numero de candidatos selecionados:", num_selecionados))

print(selecionados$nacionalidade)
print(selecionados$idade)
```

Item 5, Questão 2

Optamos por utilizar boxplots e histogramas. Eis os gráficos plotados:

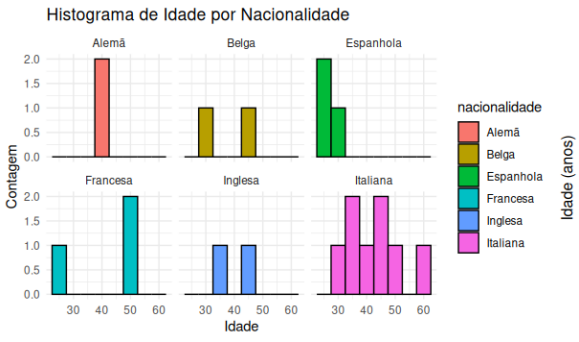

Figura 10: Histograma Idade x Nacionalidade

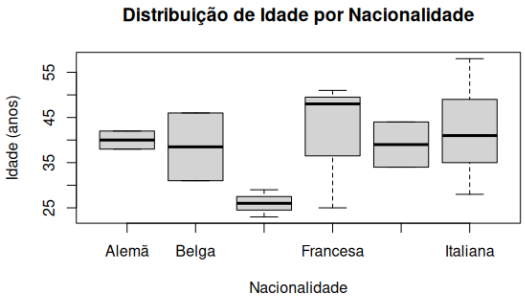

Figura 11: Boxplot Idade x Nacionalidade

Podemos perceber que o boxplot nos dá informação mais precisa acerca, principalmente, da simetria dos dados. Em contraste, o histograma nos informa melhor acerca da quantidade absoluta de candidatos dentro de certa faixa etária. São gráficos diferentes, mas que se complementam para uma boa análise.

Listado 6: Solution of item 5, exercise 2

```
boxplot(idade ~ nacionalidade,
        data = df,
        main = "Distribui  o␣de␣Idade␣por␣Nacionalidade",
        xlab = "Nacionalidade",
        ylab = "Idade␣(anos)"
)

ggplot(df, aes(x = idade, fill = nacionalidade)) + geom_histogram(
    binwidth = 5, color = "black") + facet_wrap(~ nacionalidade) + labs(
    title = "Histograma␣de␣Idade␣por␣Nacionalidade", x = "Idade", y = "
    Contagem") + theme_minimal()
```

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 3. 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 3: 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.

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

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

Item 1, Questão 3

Utilizamos o seguinte código para carregar o conjunto de dados. (A fim de evitar falhas, é importante que os arquivos code.R e HW1_bike_sharing.csv estejam na mesma pasta):

Listado 7: Solution of exercise 3

```
bike_sharing <- read.csv("HW1_bike_sharing.csv")
```

Analisando as variáveis fornecidas, podemos classificar `season`, `weathersit` e `dteday` como categóricas, visto que elas representam "etiquetas"para os elementos a que pertencem, enquanto `instant`, `temp`, `casual` e `registered` são variáveis numéricas (ou quantitativas), pois são valores aos quais podem ser aplicadas operações matemáticas de modo que o resultado represente uma informação útil.

Além disso, ao observar o documento fornecido, é possível identificar, por meio dos íces e das datas de registro, que foram realizadas 731 observações ao longo do período de 01/01/2011 até 31/12/2012.

Item 2, Questão 3

Como a variável `instant` representa apenas o índice de registro, ela não foi considerada relevante para os calculos propostos. Assim, começamos os cálculos pela variável da temperatura normalizada, partindo, depois, para os números de usuários casuais e registrados: Temperatura(normalizada):

- Média: Calculamos a soma e dividimos pela quantidade total de elementos:

$$\mu = \frac{14,1 + 14,9 + 8,1 + ... + 10,5 + 8,8}{731} = \frac{14847,5}{731} = 20,31122 \tag{14}$$

- Mediana: Obtemos a posição da mediana:

$$\frac{N+1}{2} = \frac{731+1}{2} = 366 \tag{15}$$

Ao obsevar o valor que se encontra na posição 366 no vetor de temperaturas ordenado, obtemos mediana= $20,4$.

- Quartil 1: De maneira similar à mediana, utilizamos a seguinte fórmula para indicar a posição do valor que representa o primeiro quartil:

$$(N+1) \times \frac{1}{4} = 183 \tag{16}$$

Depois, analisando o vetor ordenado, verificamos o resultado $13,8$ nessa posição.

- Quartil 3: Por fim, descobrimos a posição do terceiro quartil por meio da fórmula:

$$(N+1) \times \frac{3}{4} = 550 \tag{17}$$

Assim, vemos que o número $26,9$ se encontra nessa posição no vetor ordenado e, portanto, é o valor de Q3.

Número de usuários casuais: A fim de evitar redundância, omitiremos certos passos dos cálculos que vêm a seguir. Isso se dá pelo fato de que, como as três características possuem a mesma quantidade de elementos, seus quartis e mediana ocupam a mesma posição no respectivo vetor ordenado de cada uma delas.

- Média: Usando a fórmula clássica:

$$\mu = \frac{331 + 131 + 120 + 108 + ... + 364 + 439}{731} = \frac{620017}{731} = 848,1765 \tag{18}$$

- Mediana = 713
- Quartil 1 = 315
- Quartil 3 = 1097

Número de usuários registrados:

- Média: Novamente, somamos todos os elementos e dividimos pelo número de ocorrências:

$$\mu = \frac{654 + 670 + 1229 + 1454 + ... + 1432 + 2290}{731} = \frac{2672662}{731} = 3656,172 \tag{19}$$

- Mediana = 3662
- Quartil 1 = 2493
- Quartil 3 = 4970

Com esses valores, podemos formar a tabela:

Analisando a tabela, pode-se observar que tanto a temperatura quanto o número de usuários registrados possuem o valor da média próximo ao valor da mediana, o que indica uma distribuição mais uniforme dos valores desses objetos. Enquanto isso, para o número de usuários casuais a média é bem maior do que a mediana, indicando a existência de certos outlier positivos, valores que aumentam a média, mas não afetam a mediana.

	Temperatura (Normalizada)	N° usuários casuais	N° usuários registrados
Média	20,31121751	848,1764706	3656,172367
Mediana	20,4	713	3662
Q1	13,8	315	2493
Q3	26,9	1097	4790

Tabela 4: Média, mediana e quartis de variáveis significativas

Item 3, Questão 3

A partir dos dados fornecidos pela questão, podemos identificar os seguintes níveis para cada uma das variáveis:

- `season`: 1-Inverno; 2-Primavera; 3-Verão; 4-Outono.
- `weathersit`: 1-Céu limpo; 2-Nublado; 3- Chuva fraca; 4-Chuva forte.

Item 4, Questão 3

Inicialmente, somamos os números de usuários casuais e registrados de cada dia, a fim de obter o número total, e convertemos a temperatura para seu valor real (esta operação resultou em valores anormalmente grandes, os quais deveriam ser verificados). Tais passos foram realizados por meio do código 8:

Listado 8: Solution of exercise 3

```
total_users <- casual + registered
real_temp <- temp*41
```

Com desses dados, fomos capazes de criar as séries temporais da temperatura e do número de usuários utilizando o código e compará-los em busca de correlações. As imagens 13 e 12 foram geradas com excel, com o intuito de confirmar os gráficos obtidos com o R.

Listado 9: Solution of exercise 3

```
dates <- as.Date(dteday)
users_ts <- data.frame(Date=dates, Users=total_users)
temp_ts <- data.frame(Date=dates, Temp=real_temp)

ggplot(users_ts, aes(x = Date, y = Users)) + geom_line(color = "purple")
    + labs(title = "Total users time series", x = "Date", y = "Total Users
    ") + theme_minimal()
ggplot(temp_ts, aes(x = Date, y = Temp)) + geom_line(color = "red") +
    labs(title = "Temperature time series", x = "Date", y = "Temperature")
     + theme_minimal()
```

A partir das informações geradas, percebe-se a indicação de uma relação diretamente proporcional entre o aumento da temperatura e o uso de bicicletas. De fato, analisando o contexto geral, a hipótese que se forma é que a população provavelmente evita atividades ao ar livre em climas mais frios, quando as temperaturas as tornam desconfortáveis.

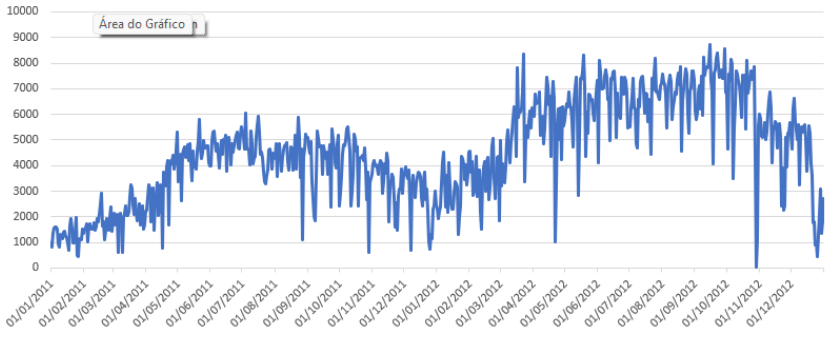

Figura 12: Número de usuários ao longo do tempo

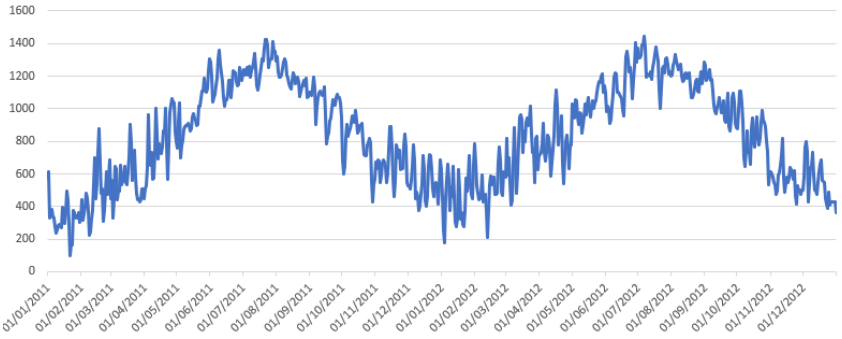

Figura 13: Temperatura (não normalizada) ao londo tempo