Statistics for Engineers
TI0111

Homework I

Full name: Arthur Lopes Pamplona, Ana Livia Cordeiro.
Registration number: 570357, 567183 (respectively).

Question 1

The daily emissions of a pollutant gas from an industrial plant were recorded 80 times, using a specific unit of measurement. The data obtained are given in Table 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				

Table 1: Daily emissions of polluting gas (problem 1).

1. Calculate the central tendency measures (mean, median, and mode) and the dispersion measures (range, variance, standard deviation, and coefficient of variation) for the dataset in Table 1. Interpret the results

Answear by Ana

Objective: The main objective of this item is to summarize the dataset using measures of central tendency, and understand the information provided by them; through measures of dispersion, and to finally note how much the measures of central tendency deviate from the majority of the data.

Measures of central tendency

- Mean
  It is the sum of all data, divided by the total number of data points.

  Mean found on calculator: 19.02

  Mean found in R: 19.02125

- Median
  It is the central number when the data is arranged in ascending order.

  Median found in R and manually: 19.15

- Mode
  It is the most repeated number in the dataset.

  Mode found in R and manually: 19.4

Measures of dispersion

- Range
  Identify the largest and smallest number and calculate the difference between them. In this case:

$$\text{largest} = 31.8$$
$$\text{smallest} = 6.2$$
$$\text{difference} = 31.8 - 6.2 = 25.6$$

- Variance
  It is calculated as follows:

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

Where:

▷ s^2 is the variance

▷ x_i is the value of each data point

▷ $\bar{x}$ is the mean of the data

▷ n is the total number of data points

Variance found in R: 30.84144

- Standard Deviation
  Its the square root of the variance, and shows back the real unity.

  $\sqrt{30.84} = 5.553507$ units

- Coefficient of variation
  It is calculated as follows:

$$CV = \frac{s}{\bar{x}} \cdot 100 \tag{2}$$

Where:

▷ s is the standard deviation

▷ $\bar{x}$ is the mean of the data

Coefficient of variation found in R: 29.19633

Interpretation: It is noted that the measures of central tendency are very close(around 19), which indicates homogeneity in the centralization of the data. However, the magnitude of the dispersion measures is also noticeable, which indicates that most of the data deviates from the center. In summary, it can be affirmed that the data are symmetrical and dispersed.

Code 1: Question 1, Item 1

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

mean(x)
median(x)
moda <- function(x) {
uniq <- unique(x)
uniq[which.max(tabulate(match(x, uniq)))]
}
moda(x)

#range
range(x)
amplitude<- max(x) - min(x)
amplitude

#variance
var(x)

#standard deviation
desvio<-sqrt(var(x))
desvio

#coefficient of variation
cv<-(desvio/mean(x))*100
cv
```

2. Create a histogram and a boxplot for the emission data. Do the data appear to be symmetrically distributed? Are there any outliers?

Answear by Arthur

In order to create a histogram its necessary to organize the data and divide into sections, creating the x section of the graph, and in these sections sum their frequencies to draw the y section of the graph.

For the data given, a division of sections of 3 is appropriate, the minimum value is 6.2 and the max value is 31.8, so the range of the graph is between 6 and 33.

Table 2: Frequencies of each interval

$[6, 9)$	3
$[9, 12)$	7
$[12, 15)$	9
$[15, 18)$	11
$[18, 21)$	21
$[21, 24)$	12
$[24, 27)$	11
$[27, 30)$	4
$[30, 33]$	1

Now creating the histogram:

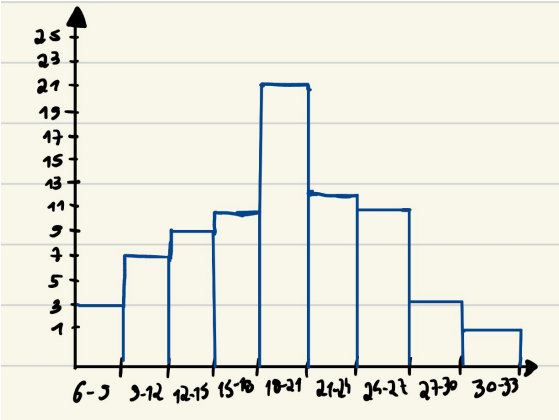

Figure 1: Histogram drew on tablet

R Histogram:

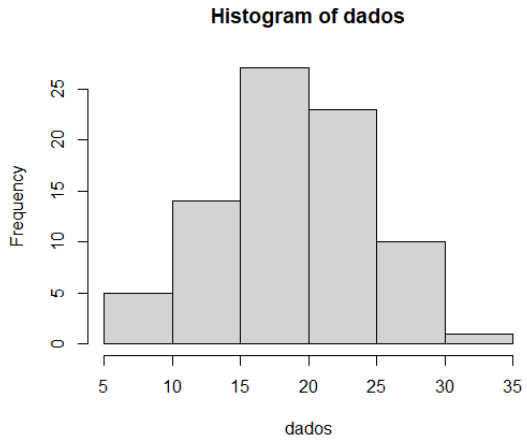

Figure 2: R's histogram

Comparing to R, the manual approach of sectioning in 3 gives more detail about the data.

The next step is to create the boxplot of the emission data, to create the boxplot its necessary to divide the data in quartiles, the first quartile is calculated as shown:

$$\textbf{Q1} = wi + (N/4 - Ni - 1)/li$$

Which:
wi = inferior limit of the quartile class, N/4 = total divided by 4, Ni-1 = acumulated frequency before the quartile class, li = frequency density

Creating the table of acumulative frequencies:

Class	Frequency	Accumulated frequency
$[6, 9)$	3	3
$[9, 12)$	7	10
$[12, 15)$	9	19
$[15, 18)$	11	30
$[18, 21)$	21	51
$[21, 24)$	12	63
$[24, 27)$	11	74
$[27, 30)$	4	78
$[30, 33]$	1	79

Table 3: Adding the accumulated frequency

- Calculating the first quartile:

N/4 is 79/4 = 19.75, falling into the [15,18) class, with wi = 15 and Ni-1 = 19
The frequency density of the class is 11/3.

$$\textbf{Q1} = 15 + (19.75\text{-}19)/(11/3) = 15.21.$$

The median is: Q2 = wi + (N/2 - Ni-1)/li;

N/2 = 79/2 = 39.5, fitting into the (18,21] class, with wi=18 and Ni-1 = 30, the frequency density is 21/3 = 7. So the median is:

$$\mathbf{Q2} = 18 + (39.5\text{-}30)/7 = 19.36$$

The third quartile is: Q3 = wi + (3N/4 - Ni-1)/li.

3N/4 = 3.79/4 = 59.25, falling into the (18,21] class, with wi=18 and Ni-1 = 30
The frequency density is 21/3 = 7.

$$\mathbf{Q3} = 18 + (59.25\text{-}30)/7 = 22.17$$

Now the **inferior limit** is: Q1 - 1.5x(Q3-Q1) = 15.21 - 1.5x(22.17-15.21) = 4.77
The **superior limit** is: Q3 + 1.5(Q3-Q1) = 22.17 + 1.5x(22.17-15.21) = 32.61

Drawing the boxplot:

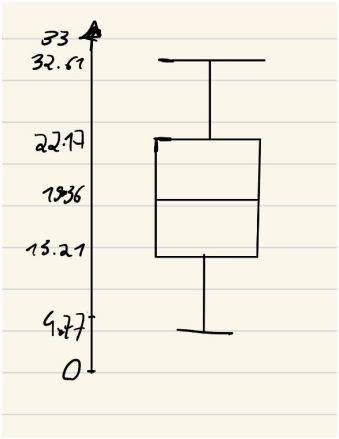

Figure 3: Boxplot drew on tablet

Comparing with R: The boxplots are similar in both approaches.

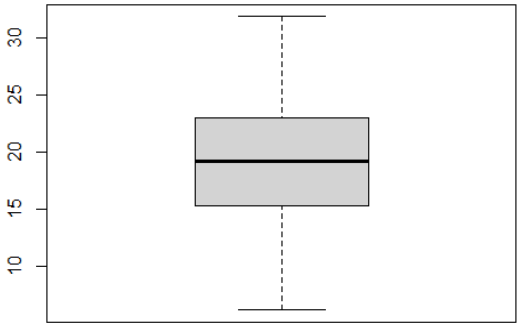

Figure 4: R's boxplot

At last, the data doesnt show any outliers, following a Gauss curve pattern.

Interpretation: On this item the knowledge for graphics build is very important, the quantitative data shown in the problem can be viewed and interpreted by these graphics much easier than the table given.

Code 2: Question 2, Item 2

```
dados = 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
)
hist(dados)
boxplot(dados)
```

3. [Determine the quartiles (Q1, Q2, Q3) and the interquartile range (IQR. Use these values to support your analysis regarding the presence of outliers

Answear by Arthur

Quartiles are values that divide the ordered dataset into four equal parts. Based on the calculations performed in Item 2 for the boxplot construction (using grouped data in frequency classes), the quartiles found were:

- **Q1 (First Quartile):** 15.21
- **Q2 (Second Quartile/Median):** 19.36
- **Q3 (Third Quartile):** 22.17

The **Interquartile Range (IQR)** is the difference between the third and first quartiles, representing the dispersion of the central 50

$$IQR = Q3 - Q1 = 22.17 - 15.21 = \mathbf{6.96} \tag{3}$$

Analysis of Outliers (Atypical Values):
To identify outliers, the $1.5 \times IQR$ rule is used. The lower and upper "fences" (or "limits") are calculated. Any value outside these limits is considered an outlier.

- **Lower Limit** $= Q1 - (1.5 \times IQR) = 15.21 - (1.5 \times 6.96) = 15.21 - 10.44 = \mathbf{4.77}$
- **Upper Limit** $= Q3 + (1.5 \times IQR) = 22.17 + (1.5 \times 6.96) = 22.17 + 10.44 = \mathbf{32.61}$

Now, the extreme values of the original (ungrouped) dataset are compared with these limits:

- **Minimum Value (original)** $= 6.2$ (which is *greater* than the lower limit of 4.77)
- **Maximum Value (original)** $= 31.8$ (which is *less* than the upper limit of 32.61)

Interpretation: Since no value in the original dataset is below 4.77 or above 32.61, it can be concluded that, according to the IQR criterion applied to the grouped data, the dataset has no outliers. This analysis reinforces the conclusion from Item 2.

4. Suppose the maximum acceptable daily limit for emissions is 25 units. What is the proportion of days on which the plant exceeded this limit? Would the general behavior of the emissions be in compliance with this regulatory standard?

Answear by Ana

Objective: Identify and select the samples that do not fit an established standard.

In the 80 days, there are 11 in which the maximum emissions limit was exceeded. Therefore, the proportion (P) of days on which the plant exceeded the emissions limits is:

$$P = \frac{11}{80} = 0.1375 \tag{4}$$

Converting the proportion to percentage:

$$P \cdot 100 = 13.75\% \tag{5}$$

Conclusion: It is noted that 86.25% of the data are within the estimated standard; It can be inferred that the general behavior of the emissions is in compliance with the maximum daily limit.

Question 2

An Italian company received 20 resumes from Italian and foreign citizens during the selection of qualified personnel for the position of foreign relations manager. Table 4 reports the information considered relevant for the selection: age, nationality, minimum desired income (in thousands of euros), and years of work experience.

	Age	Nationality	Income	Experience
1	28	Italian	2.3	2
2	34	English	1.6	8
3	46	Belgian	1.2	21
4	26	Spanish	0.9	1
5	37	Italian	2.1	15
6	29	Spanish	1.6	3
7	51	French	1.8	28
8	31	Belgian	1.4	5
9	39	Italian	1.2	13
10	43	Italian	2.8	20
11	58	Italian	3.4	32
12	44	English	2.7	23
13	25	French	1.6	1
14	23	Spanish	1.2	0
15	52	Italian	1.1	29
16	42	German	2.5	18
17	48	French	2.0	19
18	33	Italian	1.7	7
19	38	German	2.1	12
20	46	Italian	3.2	23

Table 4: Information from the Italian company's selection (question 2).

1. Calculate the mean, median, and standard deviation for the variables age, desired income, and years of experience. What can you infer from these values about the typical profile of the candidates?

Answear by Arthur
In order to calculate the mean, its needed to sum all values, and divide them by the number of values. Since Age,Income, and Years are quantitative data, this can be done without problems.

Starting with **Age**:

The sum of all ages: $28 + 34 + 46 + 26 + 37 + 29 + 51 + 31 + 39 + 43 + 58 + 44 + 25 + 23 + 52 + 42 + 48 + 33 + 38 + 46$ is 773, and the number of candidates is 20, so the mean of the age data is $773/20 = \mathbf{38.65}$.

Now for the income mean:

The incomes are abbreviated from thousand of euros, so 2.3 = 2300 euros. Performing the conversion on all income numbers, the sum is: $2300 + 1600 + 1200 + 900 + 2100 + 1600 + 1800 + 1400 + 1200 + 2800 + 3400 + 2700 + 1600 + 1200 + 1100 + 2500 + 2000 + 1700 + 2100 + 3200 = 38400$, dividing by 20, the mean for the income data is **1920** euros. The Years of experience mean:

Years of experience is also a quantitative data, so the sum is:$2 + 8 + 21 + 1 + 15 + 3 + 28 + 5 + 13 + 20 + 32 + 23 + 1 + 0 + 29 + 18 + 19 + 7 + 12 + 23 = 280$, dividing by 20, the result for the mean is **14** years.

Now, to calculate the median,

The median is calculated by listing all of the quantitative data in a crescent order, and because of that, the value in the middle do not deform from outliers.

For the Age, the ordered list is: 23, 25, 26, 28, 29, 31, 33, 34, 37, 38, 39, 42, 43, 44, 46, 46, 48, 51, 52, 58. Since the total is 20, the median is between the 10th and 11th position, so find this, we sum both values and divide by 2. The result is: 38+39/2 = 38.5 .

The ordered income is: 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, 3.4 . The median can be calculated similar to the age, so 1.7+1.8/2 = 1.75 .

At last, the ordered years of experience is: 0, 1, 1, 2, 3, 5, 7, 8, 12, 13, 15, 18, 19, 20, 21, 23, 23, 28, 29, 32. The median is 13+15/2 = 14 .

Finally, to calculate the standard deviation, we continue from the mean. This process involves finding the variance first, which is the average of the squared differences from the mean. The standard deviation is simply the square root of the variance.

Since this is a set of 20 resumes (a **sample**), we use the sample standard deviation formula (s), which divides by $n - 1$ (i.e., $20 - 1 = 19$).

The Sample Variance (s^2) formula is:

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

And the Sample Standard Deviation (s) is:

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

Standard Deviation of Age:

Mean ($\bar{x}$): 38.65 years
Sum of Squares ($\sum(x_i - \bar{x})^2$): $(28 - 38.65)^2 + ... + (46 - 38.65)^2 = 1872.55$
Variance (s^2): $\frac{1872.55}{19} \approx 98.555$

- **Standard Deviation (s):** $\sqrt{98.555} \approx$ **9.93 years**

Standard Deviation of Income (in thousands of euros):

Mean ($\bar{x}$): 1.92 (or 1920 euros)
Sum of Squares ($\sum(x_i - \bar{x})^2$): $(2.3 - 1.92)^2 + ... + (3.2 - 1.92)^2 = 9.672$
Variance (s^2): $\frac{9.672}{19} \approx 0.509$

- **Standard Deviation (s):** $\sqrt{0.509} \approx$ **0.713** (or **713 euros**)

Standard Deviation of Years of Experience:

Mean ($\bar{x}$): 14.0 years
Sum of Squares ($\sum(x_i - \bar{x})^2$): $(2 - 14.0)^2 + ... + (23 - 14.0)^2 = 2004.0$
Variance (s^2): $\frac{2004.0}{19} \approx 105.474$

- **Standard Deviation (s):** $\sqrt{105.474} \approx$ **10.27 years**

Inference about the candidate profile:

Based on the measures of central tendency (Mean and Median), the typical candidate profile is:

- **Age:** Mean of 38.65 years and Median of 38.5 years.
- **Income:** Mean of 1920 euros and Median of 1750 euros.
- **Experience:** Mean and Median of 14 years.

The medians (38.5 years, 1750 euros, 14 years) are probably the best indicators of the "typical" candidate, as they are not affected by extreme values (like the 58-year-old candidate or the one with 0 years of experience).

The standard deviations tell us about the diversity of the group:

- **Age ($s \approx 9.93$ years):** The dispersion is moderate. It indicates that, although the mean is 38.65, there is a considerable variety of candidates, from the youngest (23) to the oldest (58).
- **Income ($s \approx 713$ euros):** The dispersion is relatively low compared to the mean (1920 euros). This suggests that most candidates have similar salary expectations, revolving around 1750-1920 euros.
- **Experience ($s \approx 10.27$ years):** This is a *very high* standard deviation, especially when compared to the mean of 14 years. This reveals that the group is extremely heterogeneous in terms of experience, containing both candidates at the beginning of their careers (0, 1, 2 years) and very senior professionals (28, 29, 32 years).

Conclusion: The "typical" profile is a middle-aged (38.5) and mid-career (14 years) professional, with a salary expectation of 1750 euros.

2. Group the candidates by nationality and calculate the average desired income and average years of experience for each group. Which nationality has the highest average desired income? Which group appears to be the most experienced?

Answear by Arthur
In total we have 6 different nationalities: Italian, English, Belgic, Spanish, French and German.

The 8 **Italians** candidates have an average desired income of$(2.3 + 2.1 + 1.2 + 2.8 + 3.4 + 1.1 + 1.7 + 3.2)/8 = 17.8/8 =$ **2.225** or 2225 euros, and the average experience is:

$(2 + 15 + 13 + 20 + 32 + 29 + 7 + 23)/8 = 141/8 =$ **17.65 years**.

For the two **English** candidates, the mean desired income are (1.6+2.7)/2 = 4.3/2 = **2.15** or 2150 euros, with an average experience of (8+23)/2 = 31/2 = **15.5 years**.

The two **Belgian** candidates have an average income of (1.2+1.4)/2 = 2.6/2 = **1.3** or 1300 euros, and an average experience of (21+5)/2 = 26/2 = **13 years**.

For the three **Spanish** candidates, the average income is (0.9+1.6+1.2)/3 = 3.7/3 ≈ **1.233** or 1233 euros, and the average experience is (1+3+0)/3 = 4/3 ≈ **1.33 years**.

The three **French** candidates have an average income of (1.8+1.6+2.0)/3 = 5.4/3 = **1.8** or 1800 euros, and an average experience of (28+1+19)/3 = 48/3 = **16 years**.

Finally, the two **German** candidates have an average income of (2.5+2.1)/2 = 4.6/2 = **2.3** or 2300 euros, with an average experience of (18+12)/2 = 30/2 = **15 years**.

According to the calculations, the **German** candidates have the highest average desired income at 2300 euros, followed very closely by the Italians (2225 euros).

3. Is there a correlation between years of experience and desired income? Use appropriate visual tools (e.g., scatter plot) and calculate the Pearson correlation coefficient. Interpret the result.

Answear by Ana
Objective: Use the correct tools to identify a possible correlation between two variables.

- Building a scatter plot in R:

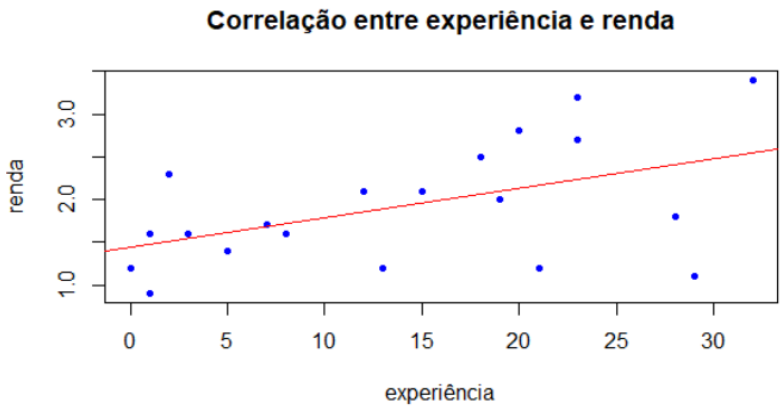

Figure 5: Question 2, Item 3

It is noted that there is no pattern regarding the position of the points. This indicates considerable dispersion among the data in question.
By calculating the Pearson coefficient, we can be more certain of the level of dispersion.

- Calculation of the Pearson correlation coefficient

$$P = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} \tag{6}$$

Where:

▷ y_i are the values of the second sample under analysis

▷ $\bar{y}$ is the mean of the second sample under analysis

Output in R: 0.4977672

The Pearson coefficient indicates a positive correlation between the data, meaning they tend to grow together. However, this correlation is moderate, given that the Pearson coefficient ranges from -1 to 1. And as indicated in the graph, the dispersion between the data is quite considerable.

Code 3: Question 2, Item 3

```
experiencia <- c(2, 8, 21, 1, 15, 3, 28, 5, 13, 20, 32, 23, 1, 0, 29,
18, 19, 7, 12, 23)
renda <- c(2.3, 1.6, 1.2, 0.9, 2.1, 1.6, 1.8, 1.4, 1.2, 2.8, 3.4,
    2.7,1.6, 1.2, 1.1, 2.5, 2.0, 1.7, 2.1, 3.2)

plot(experiencia, renda, main = "Correlation␣between␣experience␣and␣
    income",
xlab = "experience", ylab = "income", col = "blue", pch=20)
abline(lm(renda~experiencia), col = "red")
cor(experiencia, renda)
```

4. Suppose the company wants to prioritize candidates with at least 10 years of experience and a desired income below 2.0 (thousand euros). How many candidates meet both criteria? List their nationalities and ages.

Answear by Ana
There are four candidates who meet the given criteria.

AGE	NATIONALITY	INCOME	EXPERIENCE
46	Belgian	1.2	21
51	French	1.8	28
39	Italian	1.2	13
52	Italian	1.1	29

Table 1: Question 2, Item 4

5. Construct graphs that allow visualizing the distribution of age and desired income, separated by nationality. Use histograms, box-plots, or bar charts, and

comment on the main differences observed between the groups.

Answear by Ana
Objective: Construction of graphs that allow comparing certain classes in relation to a main variable.

Age-Nationality Distribution: The faceted histogram is chosen to represent the age distribution by nationality groups, as it allows us to compare the differences between groups in relation to a continuous variable, as is the case with age ranges.

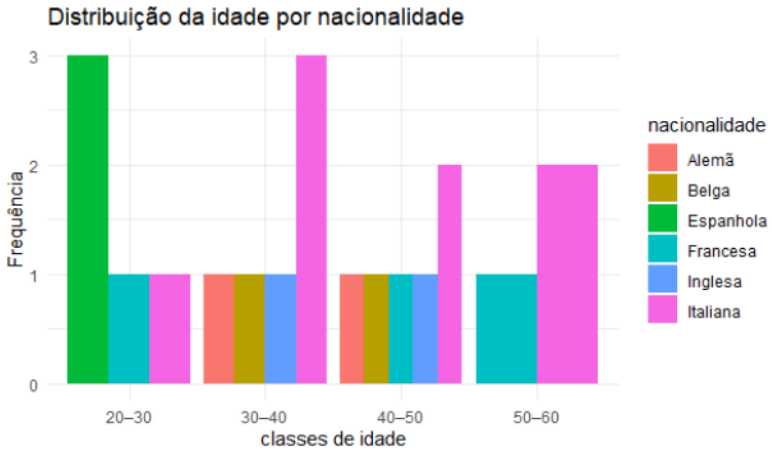

Figure 6: Question 2, Item 5

- Differences between age groups

- All individuals of Spanish nationality are in the youngest age class.
- The Italian nationality is the only one present in all classes, with most of them in the 30-40 class.
- The French nationality is present in all classes, except 30-40.
- All individuals of German nationality are between 30 and 50 years old, as are individuals of Belgian and English nationality.

Income-Nationality Distribution: As it is again a numerical variable, the boxplot is an excellent option for quick comparison, as it summarizes the data at its central points.

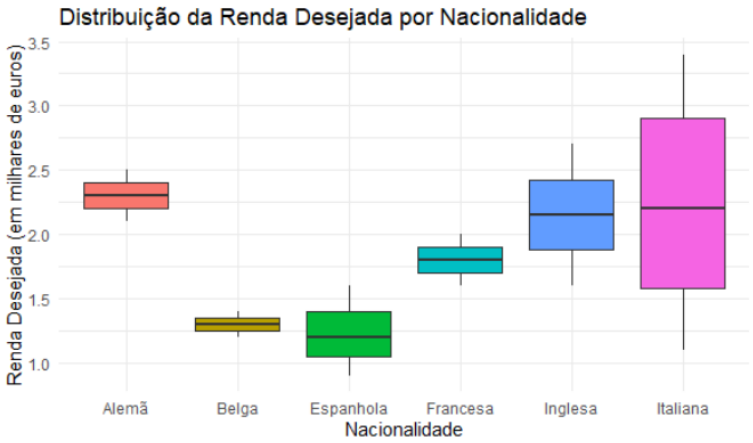

Figure 7: Question 2, Item 5

- Differences between income groups

- The Italian nationality group shows the greatest dispersion in desired income, as well as the highest maximum desired income. The group's central tendency for income coincides with the central tendency of the English and German nationality groups.
- The Belgian nationality group shows the lowest dispersion in desired income; the group's central tendency for income coincides with the central tendency of the Spanish nationality group.
- The Spanish nationality group shows the lowest minimum desired income.

Code 4: Question 2, Item 5

```
install.packages("ggplot2")
library(ggplot2)

# ------------- histogram for age ---------------------
dadosA <- data.frame(
idade = c(28, 34, 46, 26, 37, 29, 51, 31, 39, 43, 58, 44, 25, 23, 52, 42,
    48, 33, 38, 46),
nacionalidade = c("Italian", "English", "Belgian", "Spanish",
"Italian", "Spanish", "French", "Belgian",
"Italian", "Italian", "Italian", "English",
"French", "Spanish", "Italian", "German",
"French", "Italian", "German", "Italian")
)

#age ranges
dadosA$classe_idade <- cut(dados$idade,
breaks = seq(20, 60, by = 10),
labels = c("20  30  ", "30  40  ", "40  50  ", "50  60  "),
include.lowest = TRUE)

# Grouped histogram
ggplot(dadosA, aes(x = classe_idade, fill = nacionalidade)) +
geom_bar(position = "dodge") +
labs( title = "Age distribution by nationality",
x = "age classes",
y = "Frequency"
) +
theme_minimal()
```

```r
# ------------- boxplot for income --------------------
dadosB <- data.frame(
renda_desejada = 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),
nacionalidade = c("Italian", "English", "Belgian", "Spanish",
"Italian", "Spanish", "French", "Belgian",
"Italian", "Italian", "Italian", "English",
"French", "Spanish", "Italian", "German",
"French", "Italian", "German", "Italian")
)

# Boxplot
ggplot(dadosB, aes(x = nacionalidade, y = renda_desejada, fill =
nacionalidade)) +
geom_boxplot() +
labs(
title = "Distribution␣of␣Desired␣Income␣by␣Nationality",
x = "Nationality",
y = "Desired␣Income␣(in␣thousands␣of␣euros)"
) +
theme_minimal() +
theme(legend.position = "none")
```

Question 3

The attached dataset, `HW1_bike_sharing.csv`[a], refers to the bike-sharing process in a city in the United States. The set contains the columns described in Table 2. The `season` variable includes the four seasons of the northern hemisphere: spring, summer, autumn, and winter. The `weathersit` variable represents four weather conditions: 'Clear', 'Cloudy', 'Light rain', 'Heavy rain'. The `temp` variable is the temperature normalized in degrees Celsius, meaning the values were divided by 41 (maximum value).

Tag	Description
`instant`	Record index
`dteday`	Observation date
`season`	Season
`weathersit`	Weather conditions
`temp`	Temperature in °C (normalized)
`casual`	Number of casual users
`registered`	Number of registered users

Table 2: Variables from the `HW1_bike_sharing` set (question 3).

[a] The data is available in the homework materials.

1. Load the `HW1_bike_sharing.csv` dataset in R. Classify the variables by type (categorical or numerical), identify the total number of observations, and the start and end dates of the sample.

Answear by Arthur
The variables can be described as:

instant: Numerical (on the csv file it list the data in a crescent number)
dteday: Numerical (indicate the date of observation)
season: Categorical (Indicate one of the four seasons)
weathersit: Categorical (Which of the four weather conditions)
temp: Numerical (value in degrees Celsius)
casual: Numerical (value of the number of casual users)
registered: Numerical (value of registered users)

The file loaded in R indicate 731 observations, starting in 2011-01-01 and ending in 2012-12-31, which indicates 2 full years of observation.

2. Calculate measures of central tendency (mean, median) and quartiles for each relevant numerical characteristic. Present the results in a table with an appropriate title. Comment on the main points.

Answear by Ana
The concepts of mean and median are already known.

Quartiles are measures that segregate our data into four equal sectors, with 25% in each sector.

Being:
Q1: The first quartile, where up to 25% of the data is found.
Q2: The second quartile, where up to 50% of the data is found. It corresponds to the median.
Q3: The third quartile, where up to 75% of the data is found.

The interquartile range (IQR), given by:

$$IQR = Q3 - Q1 \tag{7}$$

is a good indicator of the dispersion among the data, with respect to centralization.

	TEMPERATURE	**CASUAL USERS**	**REGISTERED USERS**
MEAN	20.31	848.17	3656.2
MEDIAN	20.4	713	3662
Q1	13.8	315.5	2497
Q2	20.4	713	3662
Q3	26.9	1096	4776.5

Table 3: Question 3, Item 2

It is noted that:

- The largest difference between mean and median is found in the registered users' data, having a greater tendency to be asymmetrical.
- Over the days, the group with the greatest dispersion in users was the registered ones.

Code 5: Question 3, Item 2

```
install.packages("googlesheets4")
library(googlesheets4)

dados<-read_sheet("https://docs.google.com/spreadsheets/d/1YSj5hfQzoi_
    1thfWcpsdK7VC633TNiDJBg4er8P0r6k/edit?usp=sharing")
names(dados)

#-------measures of central tendency-----------
temp <- dados$temp
casuais <- dados$casual
registrados <- dados$registered

#means
mean(temp)
mean(casuais)
mean(registrados)

#medians
median(temp)
median(casuais)
median(registrados)

#quartiles
quantile(temp, probs = c(0.25, 0.50, 0.75))
quantile(casuais, probs = c(0.25, 0.50, 0.75))
quantile(registrados, probs = c(0.25, 0.50, 0.75))
```

3. Assign the corresponding levels to the `season` and `weathersit` variables. Construct bar charts for both. Which season has the highest number of users? Does bicycle use depend on the season? What is the most favorable weather condition for using the system?

Answear by Arthur

Assigning the season levels: 1 = Winter, 2 = Spring, 3 = Summer and 4 = Autumn. Of the weather levels: 1 = Clear Sky, 2 = Foggy, 3 = Light Rain, 4 = Heavy Rain.

Bar Plots made in R:

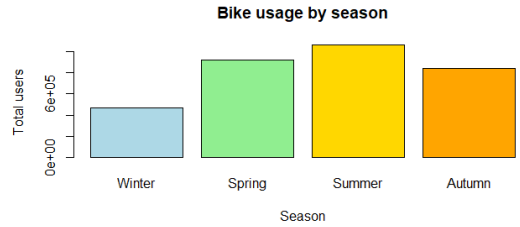

Figure 8: R season barplot

Visualising the graph, the numbers of users on Summer is the highest, and the Winter season the lowest, showing that bike sharing is highly dependent on seasons.

From the weather bar plot, its noticeable that bike share happens more in Clear Skies weather.

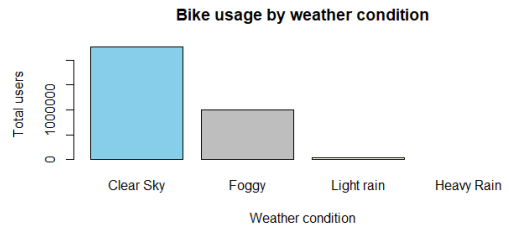

Figure 9: R weather barplot

Code 6: Question 3, Item 3

```
    file_path <- "C:\\Users\\Arthur␣Pamplona\\Documents\\HW1_bike_sharing
        .csv"
my_data <- read.csv(file_path)

head(my_data)
my_data$season <- factor(my_data$season,
                         levels = c(1, 2, 3, 4),
                         labels = c("Winter", "Spring", "Summer", "Autumn
                             "))
my_data$weathersit <- factor(my_data$weathersit,
                             levels = c(1, 2, 3, 4),
                             labels = c("Clear␣Sky", "Foggy", "Light␣rain
                                 ", "Heavy␣Rain"))
my_data$total_users <- my_data$casual + my_data$registered
users_by_season <- tapply(my_data$total_users, my_data$season, sum)
users_by_weather <- tapply(my_data$total_users, my_data$weathersit, sum)
barplot(users_by_season,
        main = "Bike␣usage␣by␣season",
        xlab = "Season",
        ylab = "Total␣users",
        col = c("lightblue", "lightgreen", "gold", "orange"))
barplot(users_by_weather,
        main = "Bike␣usage␣by␣weather␣condition",
        xlab = "Weather␣condition",
        ylab = "Total␣users",
        col = c("skyblue", "gray", "lightyellow", "lightpink"))
```

4. Calculate the total number of users per day by summing `casual` and `registered`. Convert the `temp` variable to real temperature (by multiplying by 41). Then, construct time series plots for temperature and total number of users. Do these series show a similar trend?

Answear by Ana

Objective: Identify a possible correlation between the data, through graph visualization.

Time series graphs allow visualizing the variation of a certain category over the collection time. By overlaying one graph on another, the possible correlation becomes evident, by paying attention to the moments of peak, fall, and the trend line.

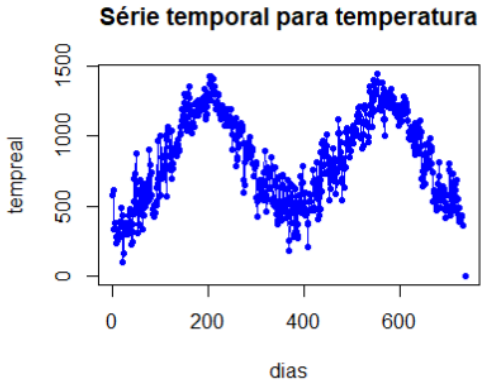

Figure 4.1: Question 3, Item 4

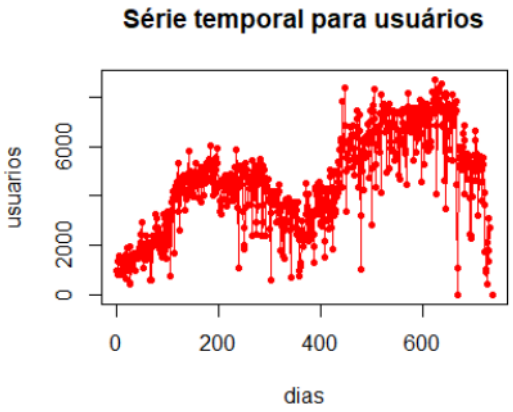

Figure 4.2: Question 3, Item 4

A clear similarity between the graphs is noted. The days with peak temperatures coincide with the days of peak users, as do the days of decline, indicating that both variables have similar trends.

Code 7: Question 3, Item 4

```
install.packages("googlesheets4")
library(googlesheets4)

dados<-read_sheet("https://docs.google.com/spreadsheets/d/1YSj5hfQzoi_
    lthfWcpsdK7VC633TNiDJBg4er8P0r6k/edit?usp=sharing")
names(dados)

#-------Graphs------------
usuarios <- dados$TotalDeUsuarios
tempreal <- dados$TemperaturaReal
dias <- c(1:737)

plot(dias, usuarios, main = "Time series for users", xlabel="days",
    ylabel="total users", pch=20, type="o", col="red")
plot(dias, tempreal, main = "Time series for temperature", xlabel="days",
     ylabel="temperature", pch=20, type="o", col="blue")
```

References 3

WIKIPEDIA. Pearson correlation coefficient. Wikipedia, the free encyclopedia. Available at ¡https://pt.wikipedia.org/wiki/Coeficiente_de_correlação_de_Pearson¿. Accessed on: 17 Oct. 2025.

FILHO, Mário. Quick Guide for Beginners in Time Series (Time Series) Mario Filho — Machine Learning. Available at ¡https://mariofilho.com/guia-rapido-para-iniciantes-em-series-temporais-time-series/¿. Accessed on: 19 Oct. 2025.