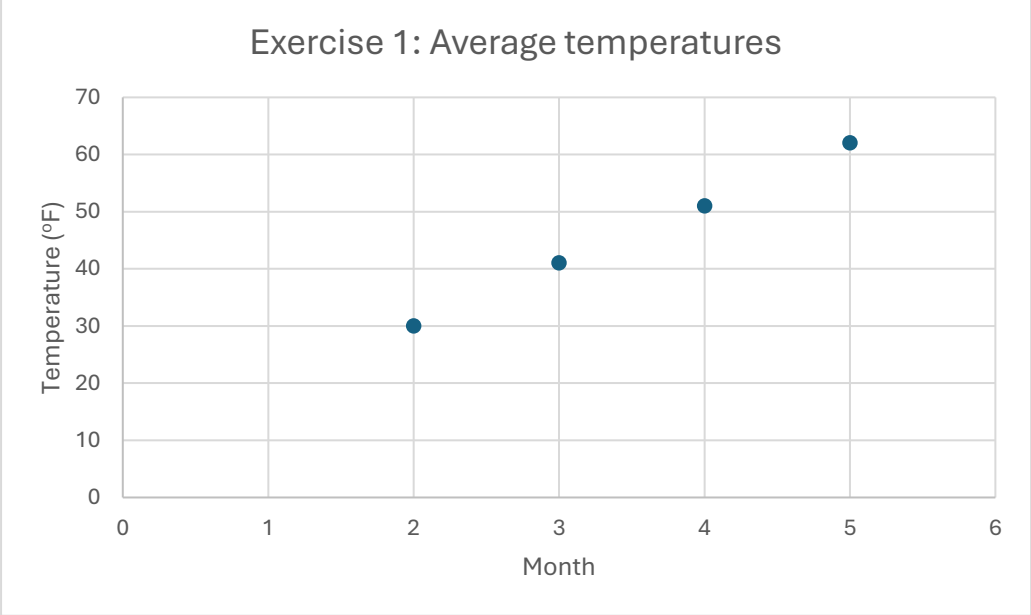
Problem set 03 – Solutions

Exercise 1.

a) The month should be the explanatory variable as it's the one that causes the change in the response variable "Temperature".

b) Scatterplot:



The relationship between the two variables is **very strong.** The response is **linear** with a **positive** association, and there is almost perfect correlation.

Exercise 2.

a) **Age**: Explanatory, **Weight:** Response
b) Explore the relationship (both variables could be explanatory or response)
c) **Number of bedrooms:** Explanatory, **Rental Price:** Response
d) **Amount of sugar:** Explanatory, **Sweetness:** Response
e) Explore the relationship

Exercise 3.

a) Conceptually, we expect a higher IQ to lead to better reading ability and a lower IQ to have worse reading ability. Indeed, the plot shows a **positive** association.
b) The bottom four points lie below the main trend, deviating from the pattern. These children have moderate IQ but the poorest reading scores.
c) The association is **roughly linear** but **moderately weak.** There is a lot of scatter.

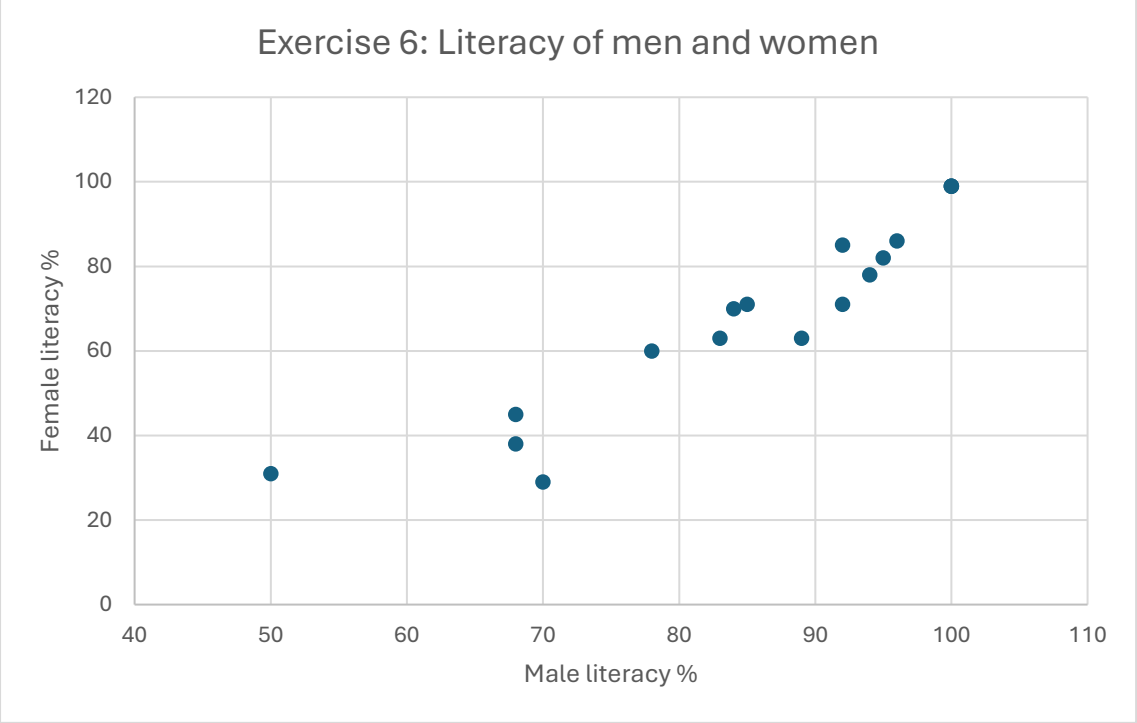
Exercise 4.

a) 1954: **~50%**, 1974: **~-30%**
b) 1981: **~15%**
c) There is no clear pattern. It can be considered **negative** but since there is a lot of data scattering, the relationship is **weak** at best with no linearity observed. The pattern appears to be more **random**.

Exercise 5.

a) The "stair-step" pattern is caused because the response variable (y-axis) can take only integer numbers
b) Yes, there is a **positive** trend, but the relationship is **weak**
c) The outlier is the (4,10) point. The child clearly overestimates his reading ability but ends up having the lowest score.

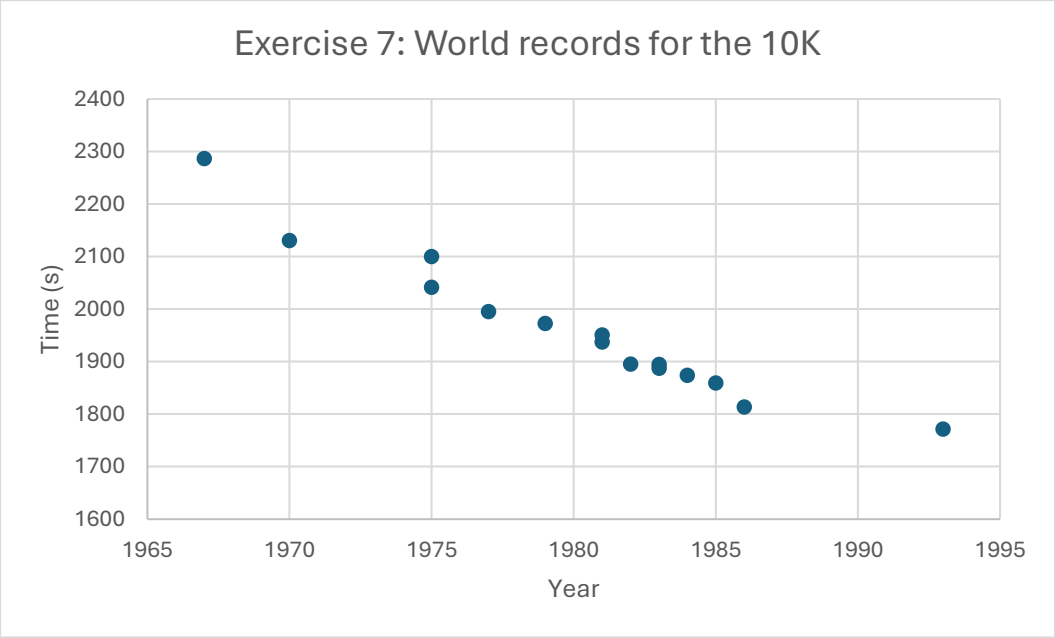
Exercise 6.



The relationship is **positive,** quite **linear**, and **strong.**

Yes, there are some identical observations (e.g. 99, 100) that overlap and the clear outlier here is **Yemen**.

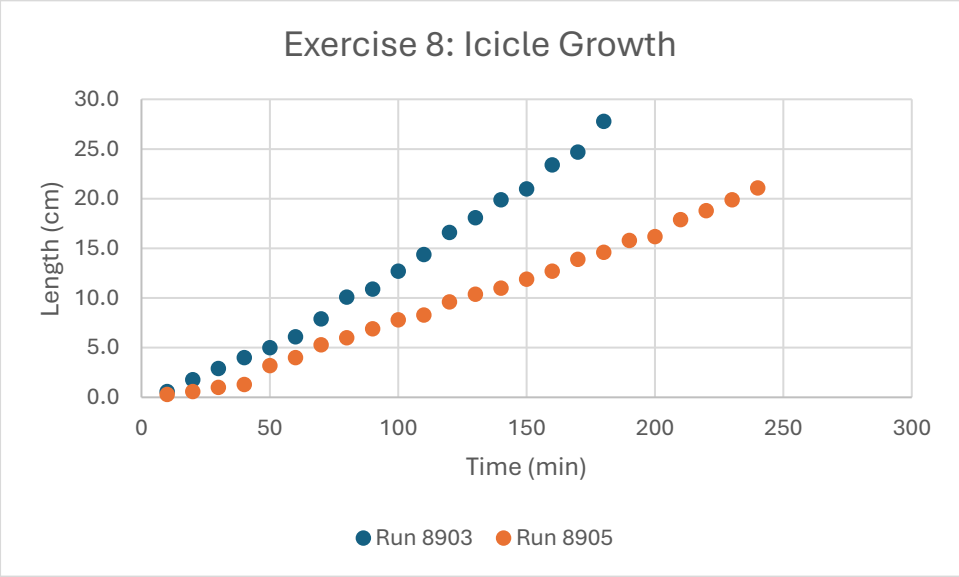
Exercise 7.



The relationship is **negative,** quite **linear,** and **strong**. There has been a steady improvement in the world record over time until the latest years, where there was much slower progress.

Exercise 8.

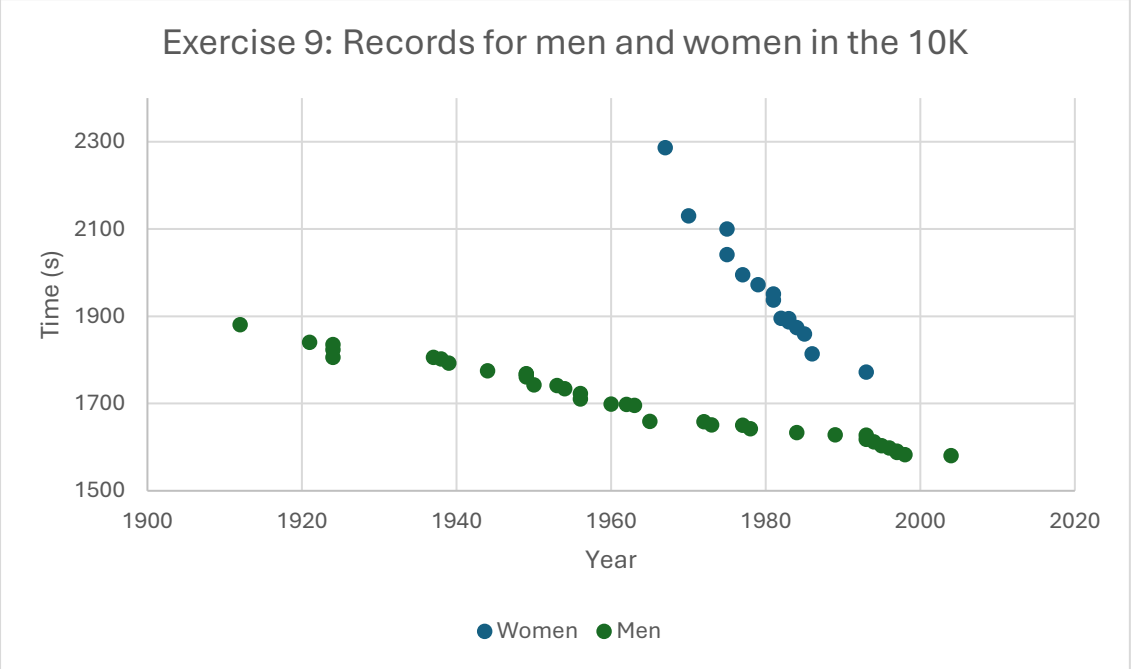
a) Scatterplot:



b) Both runs show an increase in length over time. Run 8905 had a higher water flow but a slower growth than run 8903, resulting in shorter icicles, whereas run 8903 produced longer icicles. It means that higher flow makes the water more difficult to freeze.

Exercise 9.

a) We add the data for men to the plot from Exercise 7:



Both datasets have a **negative,** fairly **linear,** and **strong** relationship. Men have a longer history in competition, and the record improved steadily over time. Women started later, but the improvement over time has been more rapid.

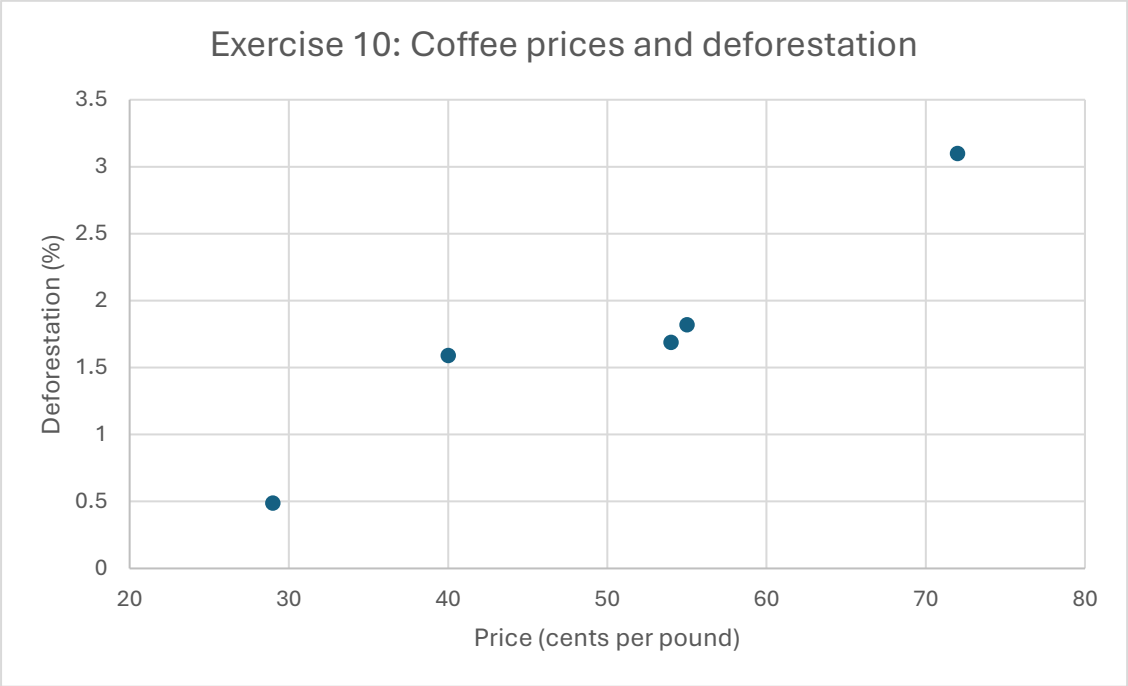
b) While we don't have enough data for the world record of women beyond 1993, judging from the pattern of the scatterplot, it appears that, indeed, over time, women have managed to close the gap.

Note

Since the correlation is very strong in both patterns, we can safely predict the pattern of the plot (i.e., extrapolate)

Exercise 10.

a) Scatterplot:



The explanatory variable is the **price,** as it causes further deforestation. The association is **positive, linear,** and **strong**.

b) Correlation equation:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

We compute

$\bar{x} = 50$, $s_x = 16.325$

$\bar{y} = 1.74$, $s_y = 0.928$

And using the above equation, we determine:

$r = 0.955$

The value confirms our observation for the strong association described in part a).

c) Use the CORREL function in Excel.

Exercise 11.

a) We calculate the mean value and the standard deviation for both variables and afterwards, we calculate the correlation through the equation written above.

For x: first test score, and y: final exam score:

$\bar{x} = 145.50$, $s_x = 15.372$

$\bar{y} = 156.25$, $s_y = 14.079$

Thus, $r = -0.201$

b) Yes, the value of **-0.201** proves that the relationship between the two variables is **very weak**. The correlation value is closer to zero, which means that the first test scores are very weakly associated with the final test.

Exercise 12.

a) Similarly to the previous exercise:
$\bar{x} = 156$, $s_x = 11.916$
$\bar{y} = 156.25$, $s_y = 14.079$
Thus, **$r = 0.519$**

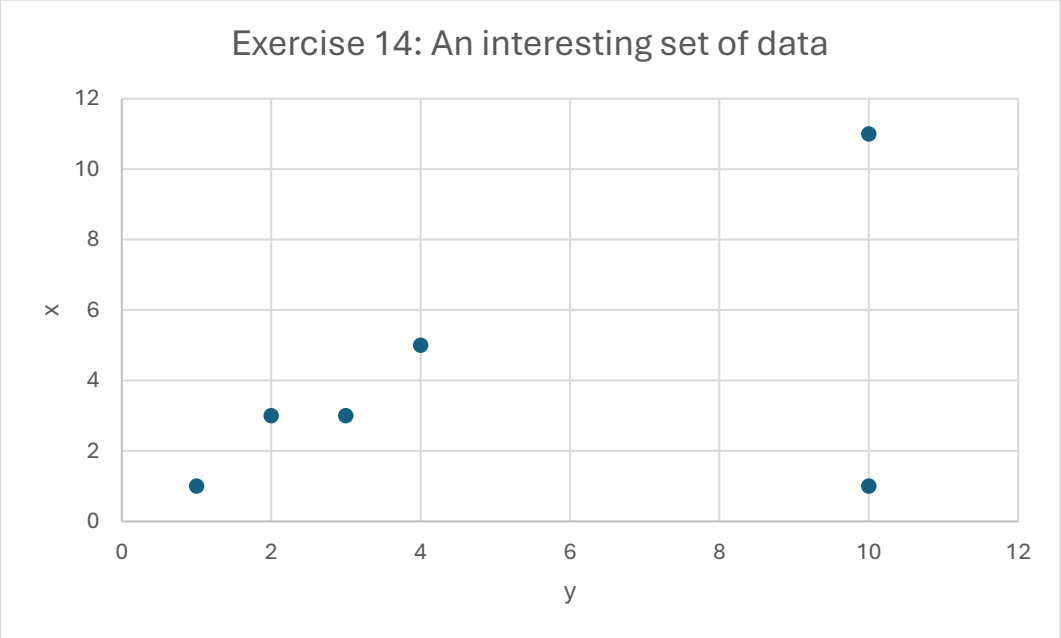
b) The new correlation supports the claim that the relationship is stronger between the second test scores and the final exam, showing a moderately **positive** association. Thus, higher second test scores have a tendency to be higher also in the final exam.

Exercise 13.

In Exercise 3, we described the association as roughly linear but moderately weak, so the correlation is closest to **$r = 0.6$**. ($r = 0.1$ would be too weak – almost complete scatter and $r = 0.9$ would be very strong, almost perfect linearity)

Exercise 14.

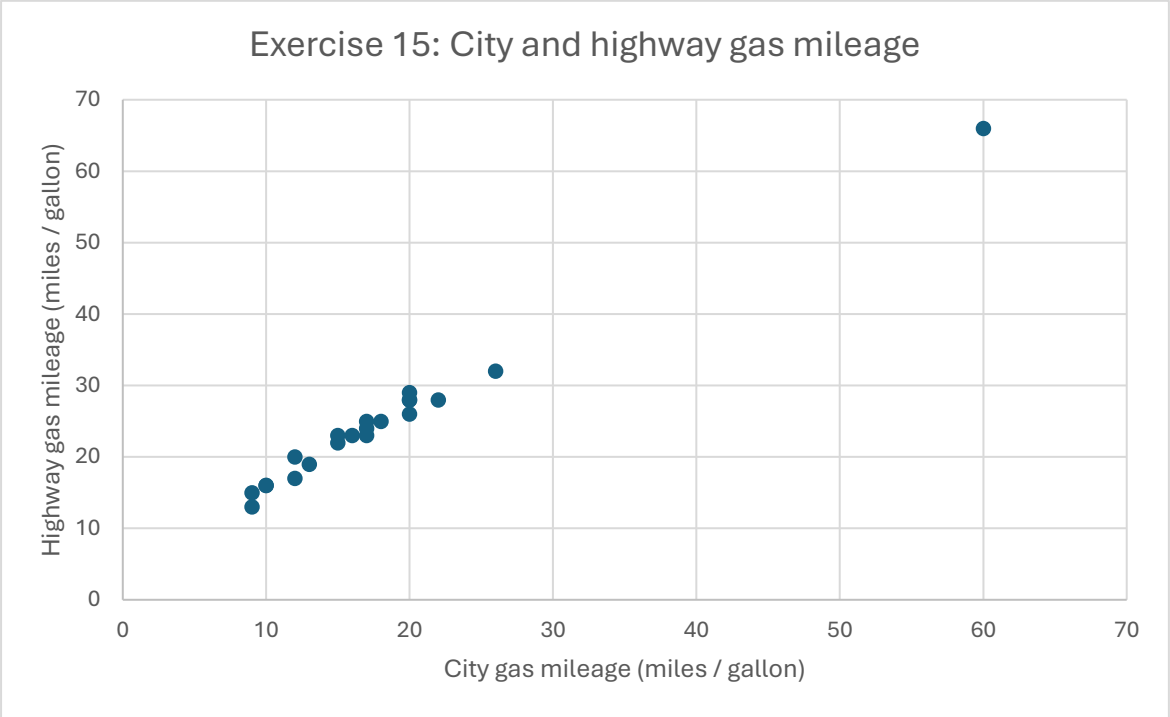
Scatterplot:



Through the Excel function, we determine that **$r = 0.481$** close to 0.5. The (1, 10) outlier causes the drop in the correlation value. If we remove this point, the correlation becomes **$r = 0.993$,** thus a perfect correlation. This shows that even a single outlier can cause a big change at the correlation number.

Exercise 15.

a) Scatterplot



Yes, the Honda Insight "extends" the linear pattern of the two-seater cars. Just because it's far from the main cluster of points doesn't mean it is an outlier.

b) With Honda Insight, **r = 0.994**
Without Honda Insight, **r = 0.977**
Honda Insight strengthens the already strong association because the point lies within the *slope*. In other words, it aligns very well with the trend so that's why it increases the correlation.

Exercise 16.

Equation of the least-squares regression line

$\hat{y} = b_0 + b_1x$

with **slope:** $b_1 = r\frac{s_y}{s_x}$

and **intercept:** $b_0 = \bar{y} - b_1\bar{x}$

We want to show that when $\mathbf{x} = \bar{\mathbf{x}}$, then $\hat{\mathbf{y}} = \bar{\mathbf{y}}$

$\hat{y} = b_0 + b_1\bar{x}$ *Substituting the intercept equation*

$\hat{y} = \bar{y} - b_1\bar{x} + b_1\bar{x}$ *The two* $b_1\bar{x}$ *factors cancel out*

Therefore, $\hat{\mathbf{y}} = \bar{\mathbf{y}}$

Exercise 17.

a) We calculate the mean value and the standard deviation for time and the icicle growth of the run 8903, so we can determine the correlation.
$\bar{x} = \mathbf{95}$, $s_x = \mathbf{53.385}$
$\bar{y} = \mathbf{12.7}$, $s_y = \mathbf{8.496}$
Thus, $\mathbf{r = 0.996}$

b) The values for x remain unaffected
For 1 inch = 2.54 cm, we divide all y values with 2.54 and:
$\bar{y} = \mathbf{4.98}$, $s_y = \mathbf{3.35}$
Also, the correlation, **r,** does not change

c) <u>For y in cm:</u>
$$b_1 = r\frac{s_y}{s_x} \Leftrightarrow b_1 = 0.996\frac{8.496}{53.385} \Leftrightarrow \boldsymbol{b_1 = 0.159}$$
<u>For y in inches:</u>
$$b_1 = 0.996\frac{3.35}{53.385} \Leftrightarrow \boldsymbol{b_1 = 0.063}$$
Alternatively, we can divide b_1 in cm by 2.54

Exercise 18.

a) From the table, $n_{\text{F.t., 15-19}} = 3{,}388{,}000$

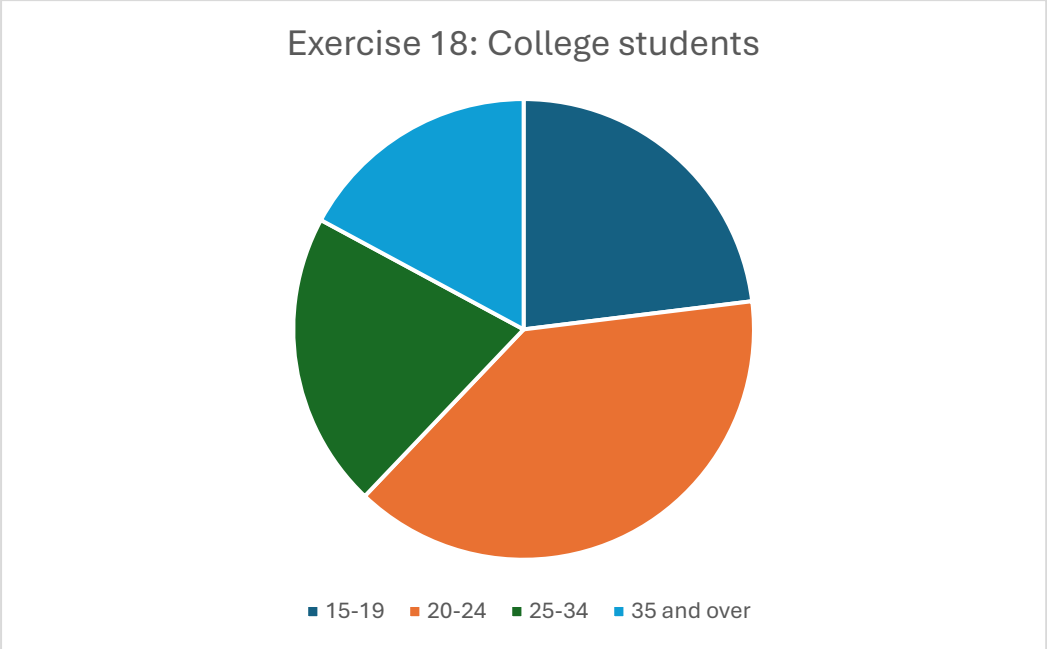
b) <u>We make the sum of the whole table</u>
Total number of U.S. students: $\mathbf{N = 16388}$
We also express the total as thousands

For the joint distribution, we divide each number by the sum

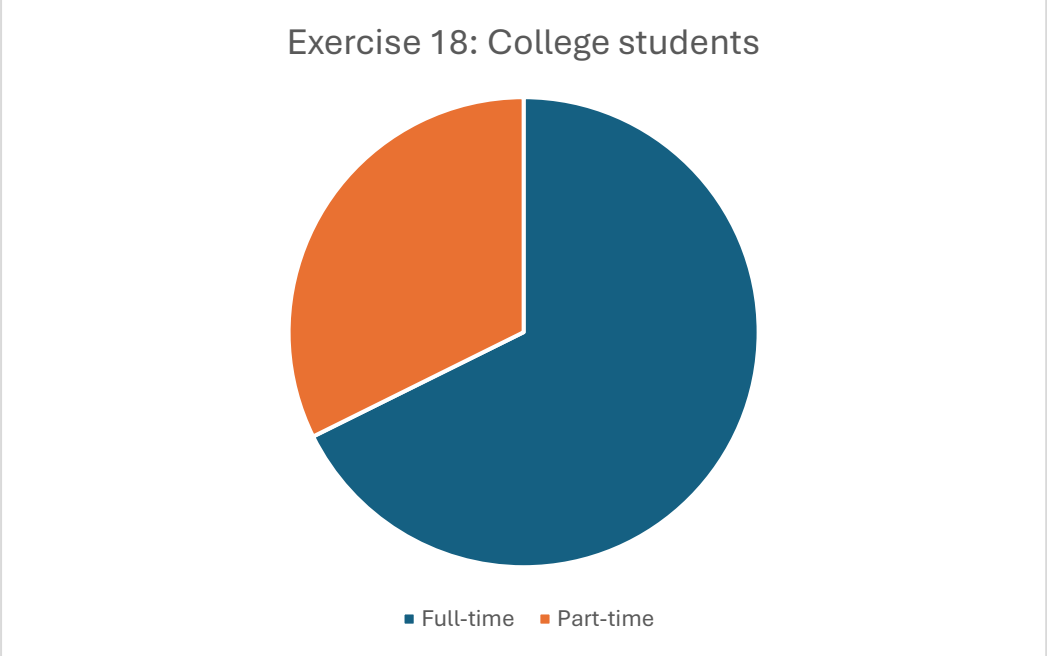
	Status			Joint distribution		
Age	**Full-time**	**Part-time**	**Total**	**Full-time (%)**	**Part-time (%)**	**Total (%)**
15-19	3388	389	3777	20.67	2.37	23.05
20-24	5238	1164	6402	31.96	7.10	39.07
25-34	1703	1699	3402	10.39	10.37	20.76
35 and over	762	2045	2807	4.65	12.48	17.13
Total	11091	5297	**16388**	67.68	32.32	**100.00**

c) Since we deal with percentages that have a total of 100%, the best graph would be a pie chart

For age:

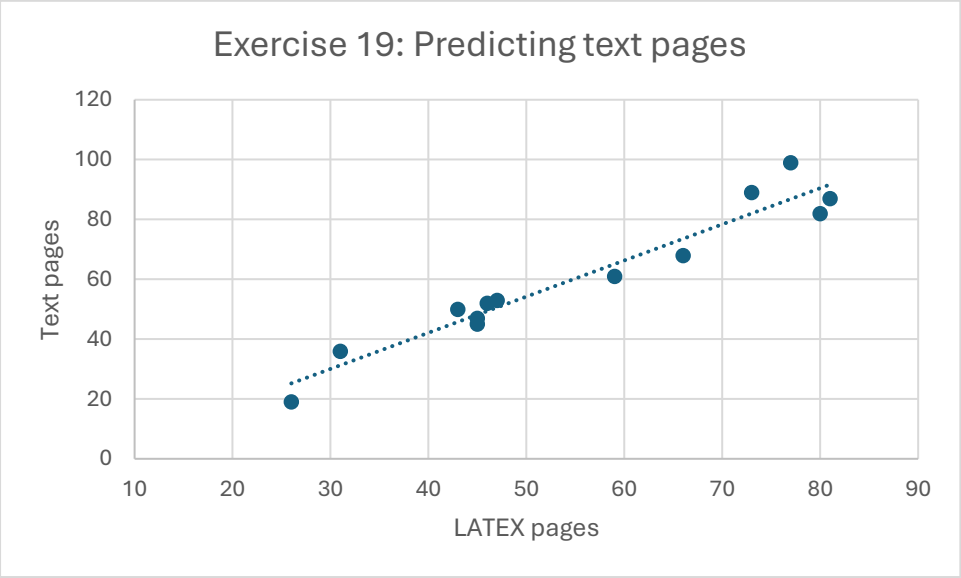


d) For status:



Exercise 19.

a)-b) We plot the data, and we add the trendline to the plot



Determine the equation of the least-squares regression line

$\hat{y} = b_0 + b_1 x$

$$b_1 = r\frac{s_y}{s_x}$$

$b_0 = \bar{y} - b_1\bar{x}$

We compute:
$\bar{x} = \mathbf{55.31}$, $s_x = \mathbf{18.598}$
$\bar{y} = \mathbf{60.62}$, $s_y = \mathbf{23.272}$
$\mathbf{r = 0.965}$

$$b_1 = r\frac{s_y}{s_x} = 0.965\frac{23.272}{18.598} \Leftrightarrow \boldsymbol{b_1 = 1.208}$$

$b_0 = 60.62 - 1.208 \times 55.31 \Leftrightarrow \mathbf{b_0 = -6.194}$

Therefore:

$\hat{y} = \mathbf{-6.194 + 1.208x}$

Note: We can verify the accuracy of the values by clicking on the "Display Equation on chart" option or using the SLOPE and INTERCEPT functions in Excel.

c) For $x = 62$
Using the equation above:
$y = -6.914 + 1.208 \times 62 \Leftrightarrow$ **y = 68.7 pages** (round up to **69 pages**)

d) The plot shows a **positive, linear,** and **strong** relationship, which is also evidenced by the calculated correlation factor ($r = 0.965$). The regression line equation can safely predict that for every LATEX page, we get ~1.21 more text pages.

Exercise 20.

The results support the idea that greater diversity is associated with less variation in population size, as indicated by the negative slope ($b_1 = -0.13$). However, the $r^2 = 0.34$ demonstrates a moderate to weak relationship. In other words, only 34% of the population is accounted for by the diversity mentioned in this context.