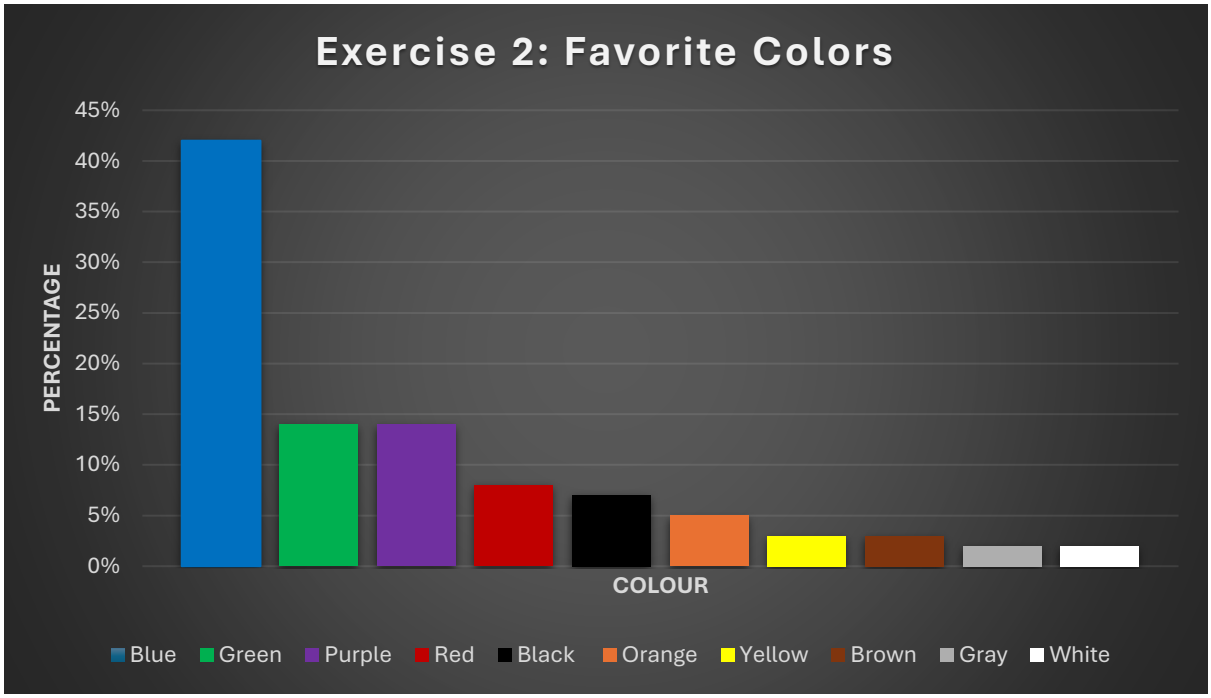
Problem set 01 – Solutions

Exercise 1.

a) Quantitative
b) Categorical
c) Categorical
d) Quantitative
e) Quantitative
f) Categorical

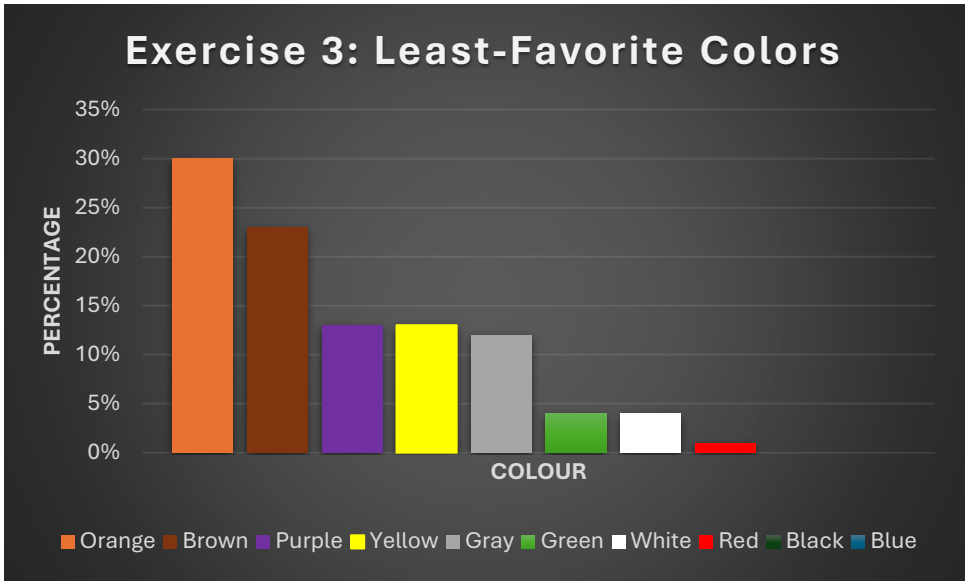
Exercise 2.



Blue was by far the most popular choice.

70% of the people in the survey chose 3 out of 10 colours.

Exercise 3.



Even if two options are 0%, they *must* be included in the graph.

Orange was the most frequent answer chosen by 30% of the participants, followed by brown at 23% and together, they account for more than half of the least-favorite colour. Black and blue were not disliked by any participant.

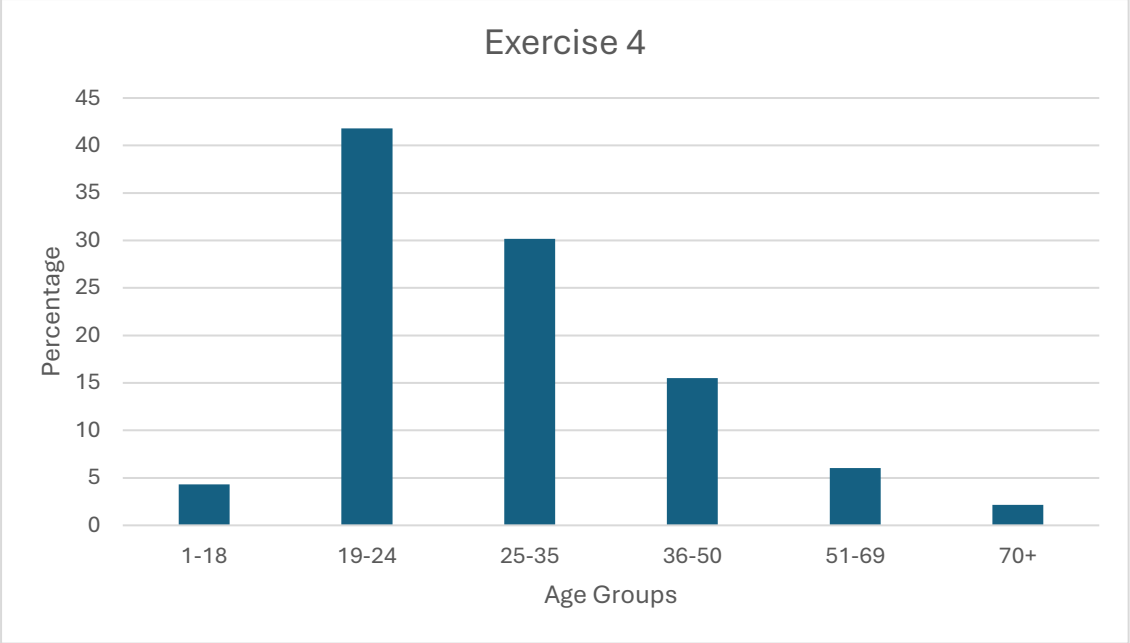
Exercise 4.

a) $\boldsymbol{n = \sum_1^6 x_i \Leftrightarrow n = 232}$

$Percentage = \frac{x_i}{n}\, x\, 100$

Age Groups	1-18	19-24	25-35	36-50	51-69	70+
Count	10	97	70	36	14	5
Percentage	**4.31%**	**41.81%**	**30.17%**	**15.52%**	**6.03%**	**2.16%**

b) Age has been sorted into groups, and it is considered here **categorical** data

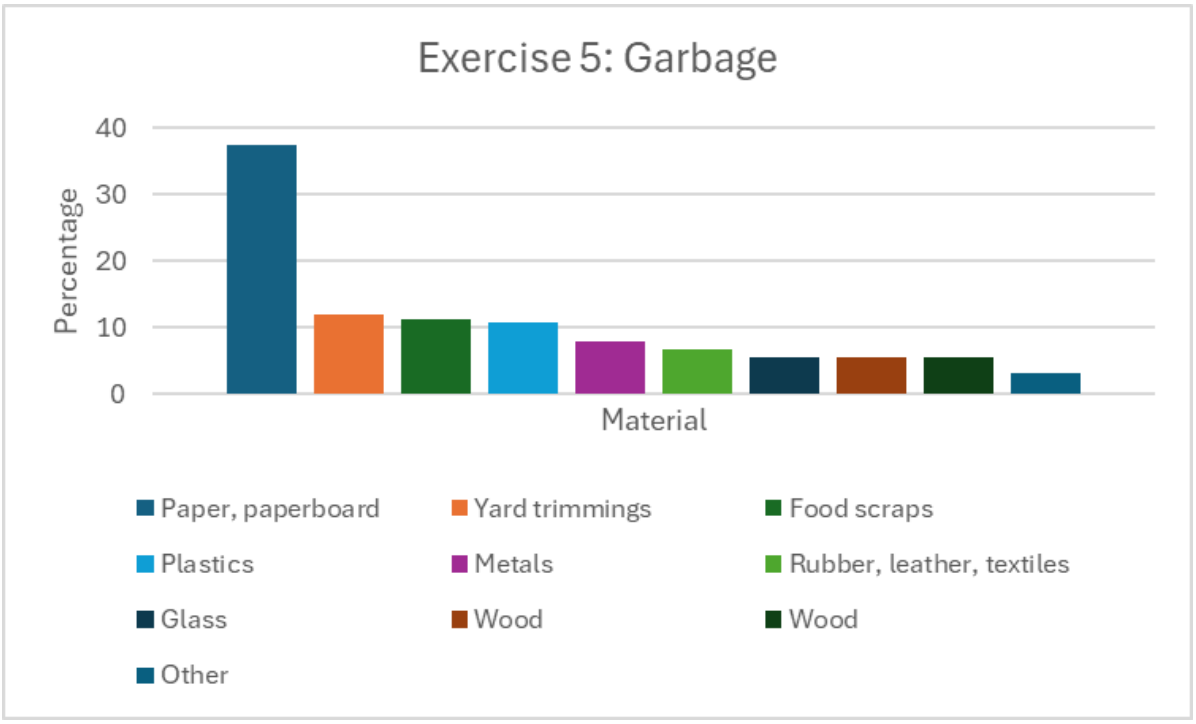


c) The survey was more focused on younger ages (i.e., 19-35, ~72% of the total survey)
d) A histogram is used for **quantitative** data where the bars represent continuous intervals in numbers, where the order matters. In a bar graph, we can sort the data according to our preference, and we use it because we have sorted the age groups and the data is **categorical**

Exercise 5.

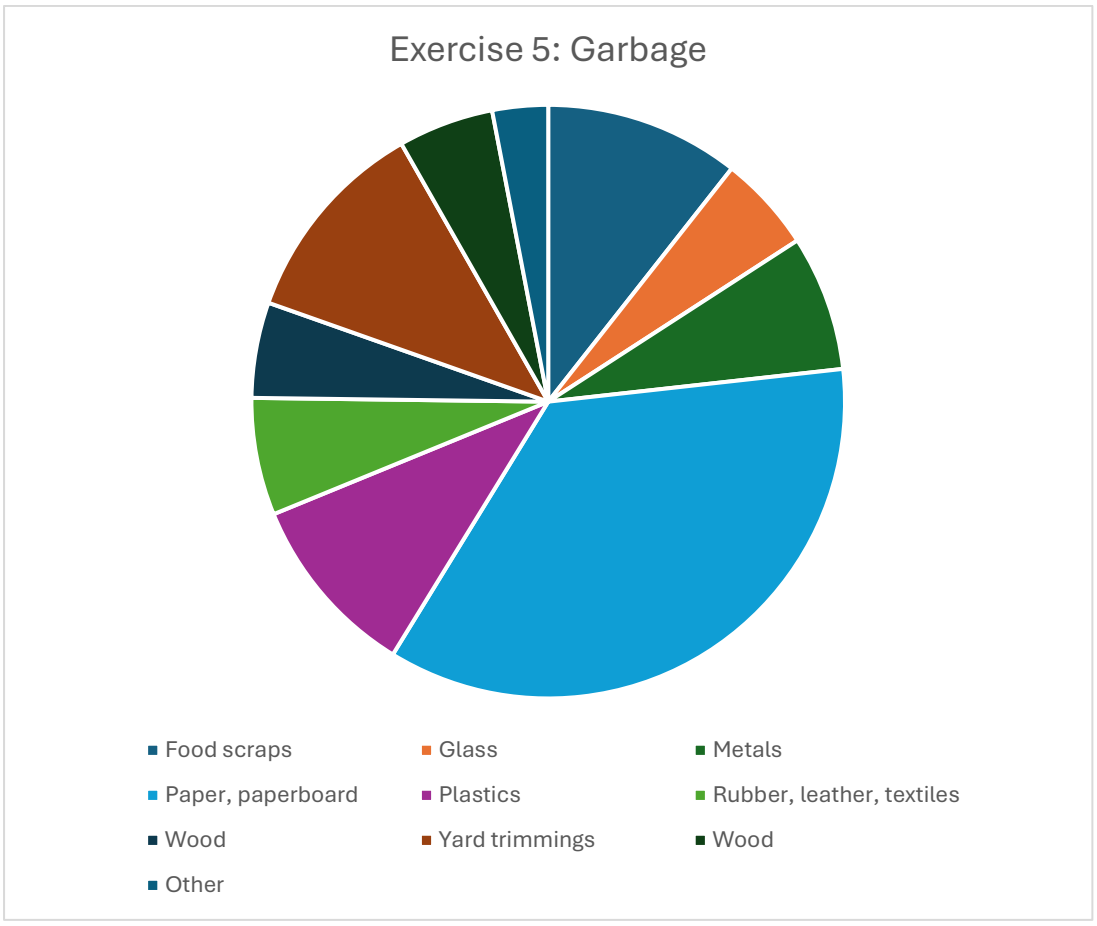
a) Sum: **Million Tons:** 244, **Total percentage:** 105.5%. The sum of the total percentage is more than 100 because of the rounding error

b) After reordering:



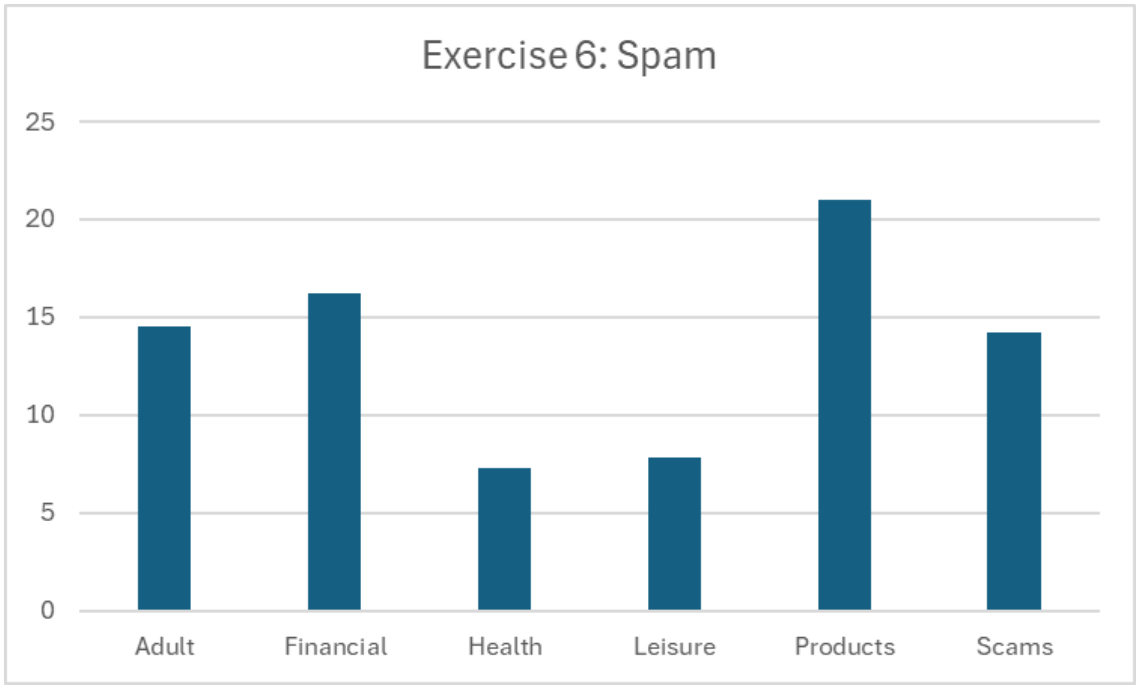
It becomes clearer which are the main contributors to the total waste.

c) Pie chart:

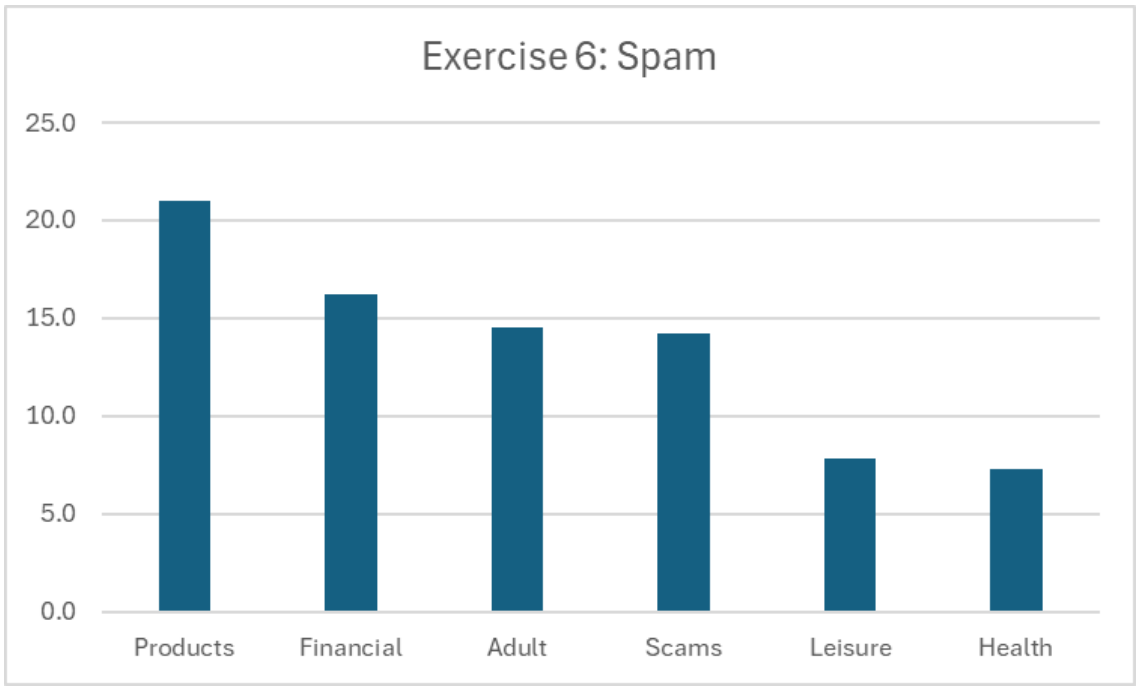


Exercise 6.

Alphabetical order:



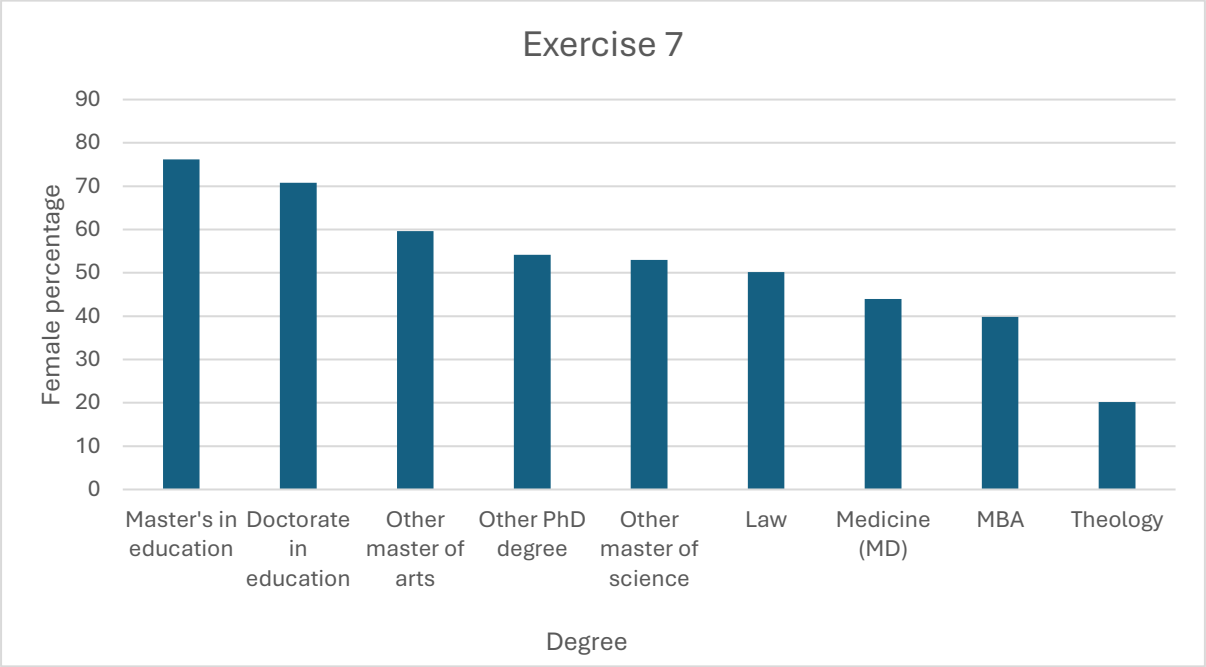
Tallest to shortest order:



Exercise 7.

a) The percentage refers to the value **within** the specific variable (i.e., the degree). The categories are independent, and they do not sum to 100%. Plotting a pie chart would be very misleading and a bar graph is the ideal choice to compare the percentages across different degrees.

b) Bar graph sorted by highest to lowest value:



Exercise 8.

a) According to the stemplot, **Alaska: 5.7%**, **Florida: 17.6%**

b) Ignoring the two outliers from above, the distribution is symmetric, centered around ~ 13% (more values are located there) with a spread of 8.5% – 15.6%.

Exercise 9.

Stemplot:

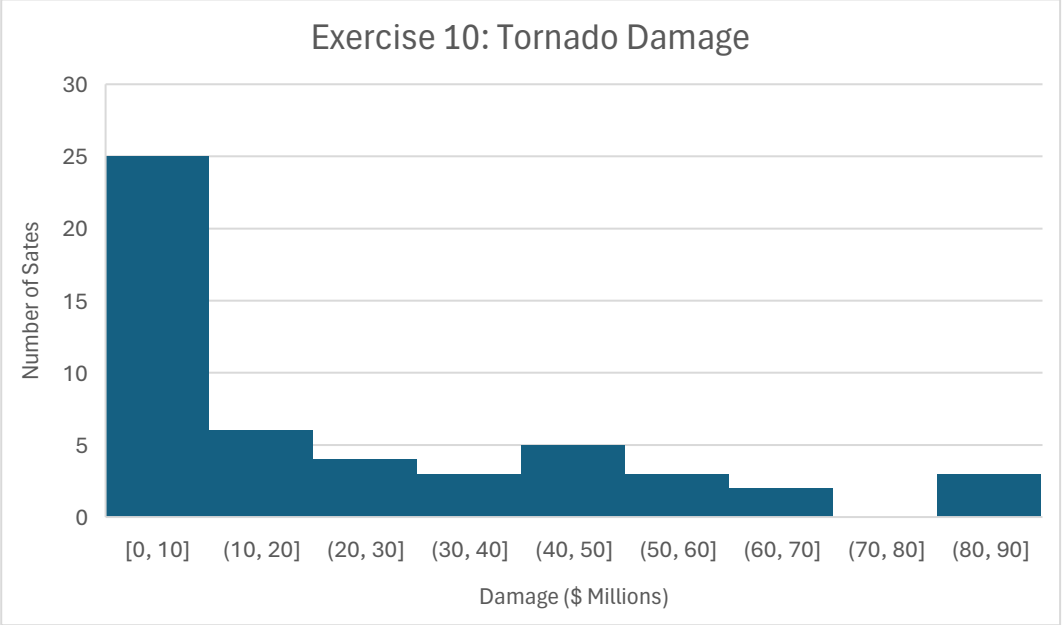
7	8
9	5 6
10	3
11	2
13	4
14	1 5 7 8
15	3 8
17	2 2
20	0
25	5
27	1
35	9

The distribution is right-skewed, centered around 140-150 mg/dl. Outlier: 359 mg/dl. The goal of the glucose level between 90-130 mg/dl is only satisfied by four people, so the group is not doing well.

Exercise 10.

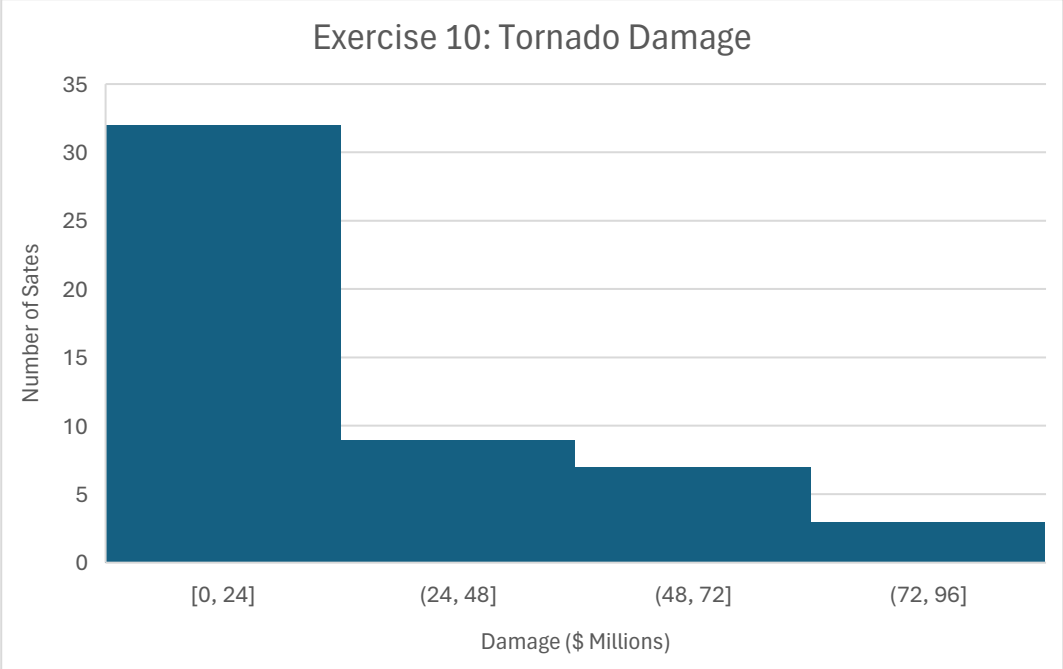
a) After sorting the table from highest to lowest damage:
Top five states: Texas, Minnesota, Oklahoma, Missouri, Illinois
Bottom five states: Alaska, Puerto Rico, Rhode Island, Nevada, Vermont

b) *Make sure you put the bin width at **10***



Strongly right skewed, centered between 15-20 million $, spread: very wide range. Texas, Minnesota, and Oklahoma are very clear **outliers.**

c) Default graph:

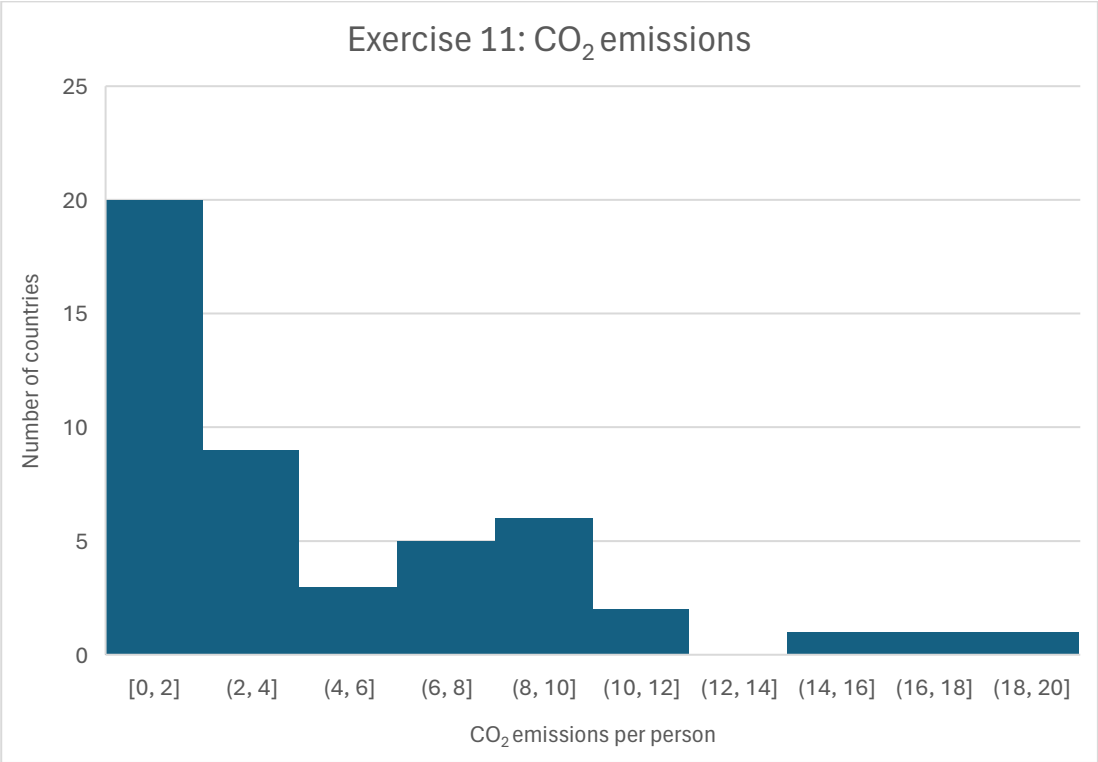


The first graph with intervals of 10 paints a clearer picture of our data and helps us to draw conclusions better and allow us a fairer comparison.

Exercise 11.

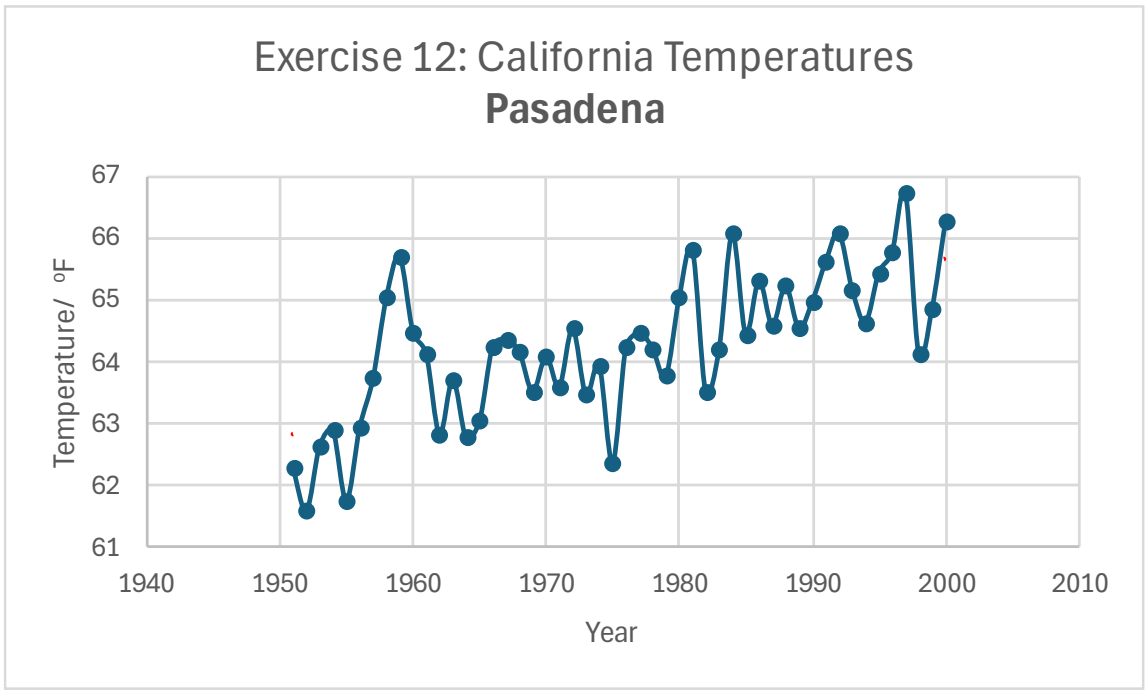
a) The statistical analysis will lead to inaccuracies due to countries with higher populations will always have the highest CO_2 emissions

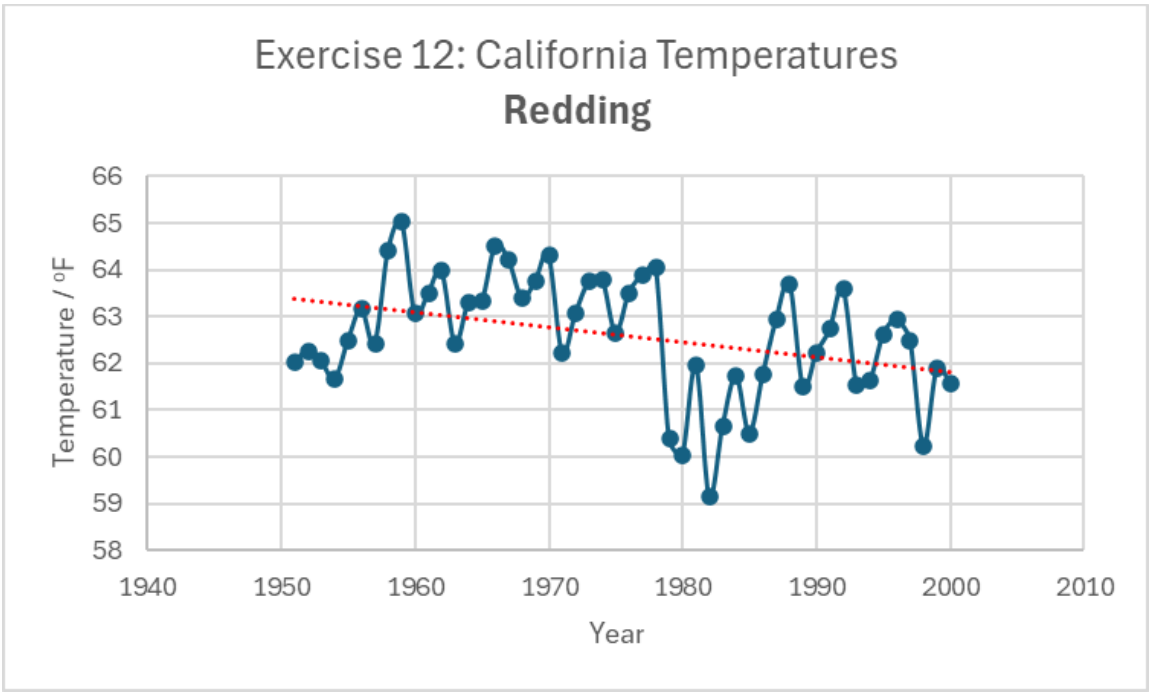
b) *Set the bin width at* ***2***



Strongly right skewed, centered around 2 tons, wide spread with 0-19.9 tons. Outliers: Australia, Canada, and USA.

Exercise 12.

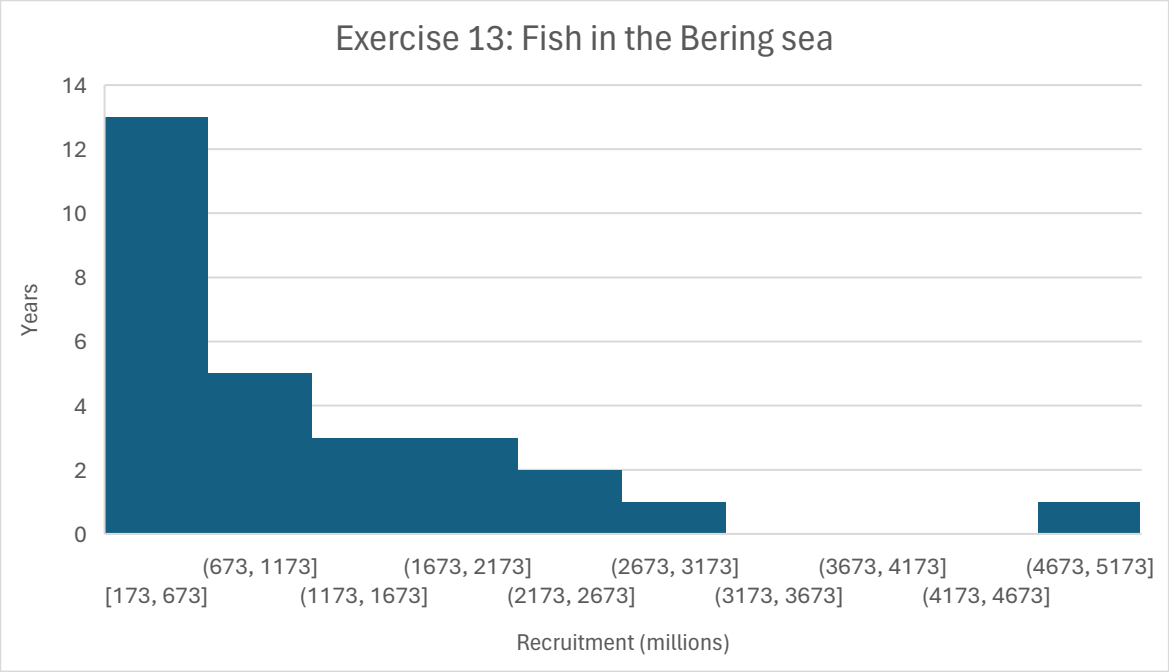




We make separate timeplots of the two cities and then we plot the trendline. Both timeplots show random fluctuations of the temperature over the years, but after plotting the trendline, we see an **upward** trend for **Pasadena** and a **downward** trend for **Redding,** mainly due to dropped temperature in the mid-1980s. The downward trend in the Redding graph is why discussions of climate change can bring disagreement.

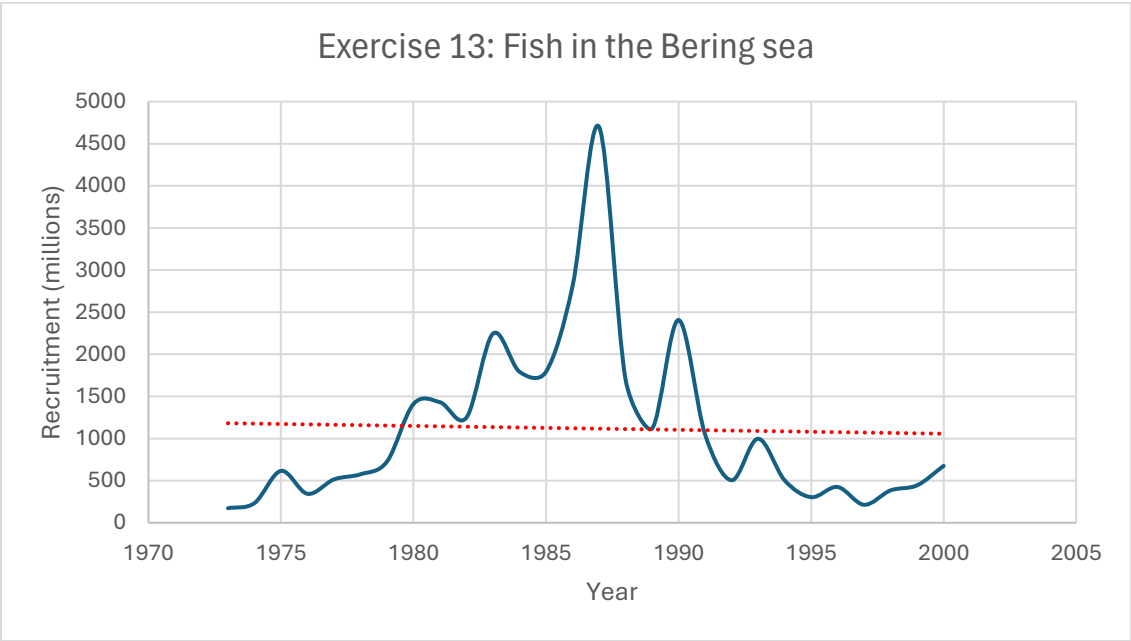
Exercise 13.

a) First we plot a histogram



The distribution is right-skewed, centered around 600-800 million with a wide spread of 173-4700 million. There is a very noticeable outlier in 1987.

b) A time plot is more useful to analyze the trend over the years:
(Note that the axes are changing)



The trend appears to be a bit downward but not very steady with severe fluctuations over the time period with a big spike in 1987, the outlier as mentioned earlier.

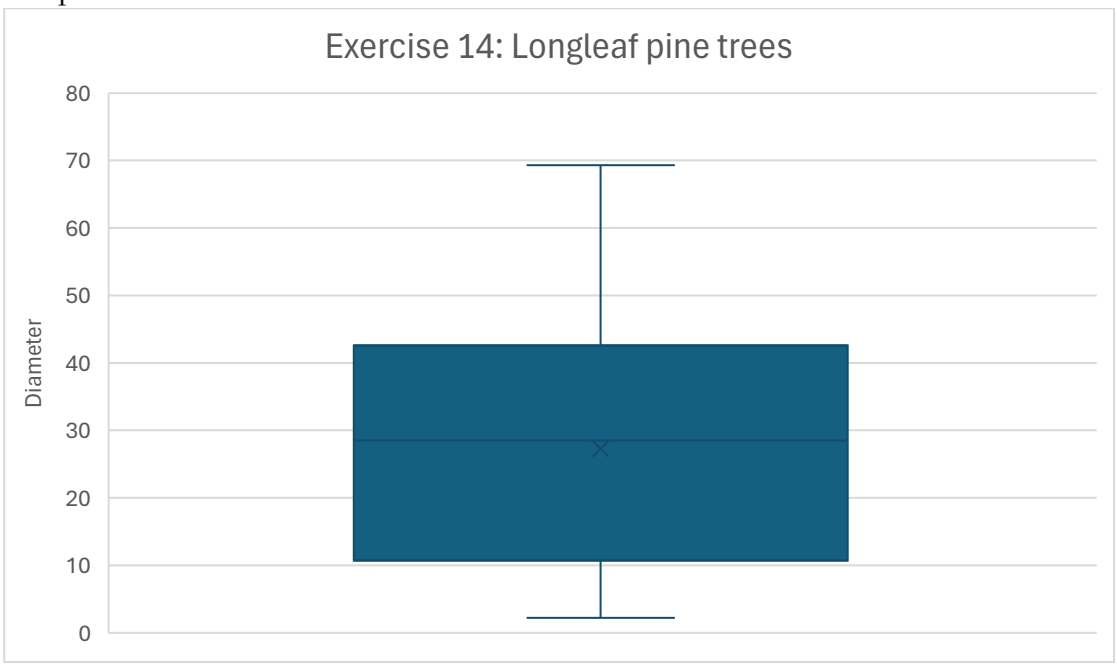
Note:
Histograms show the distribution, while timeplots reveal the fluctuation over the years.

Exercise 14.

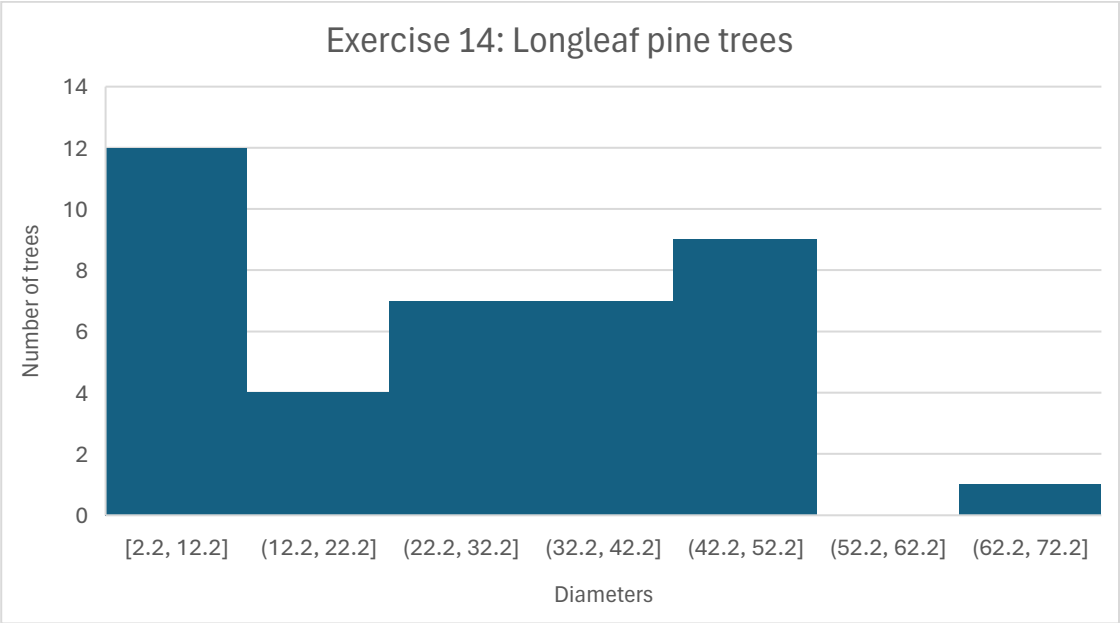
a) Five-number summary

Minimum: 2.2
Q1: 11.2
Median: 28.5
Q3: 41.2
Maximum: 28.5

b) Boxplot:



c) Histogram:



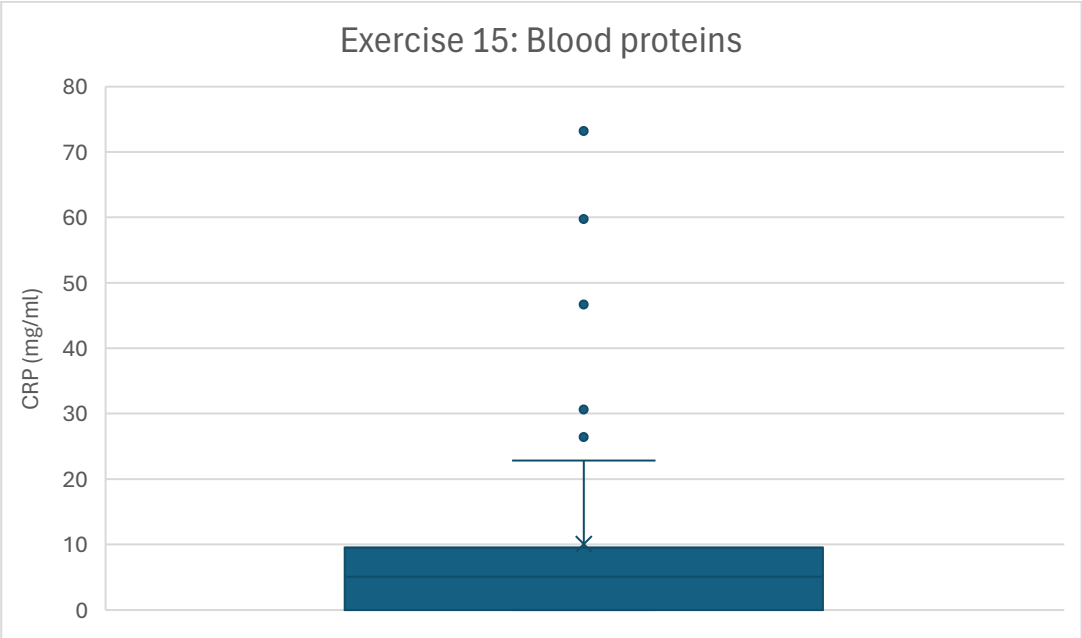
d) The distribution is **right-skewed** as evidenced by both plots but more visible in the histogram. The second graph reveals a strong outlier of 69.3 diameter, spread is wide. In this occasion, the histogram is more informative of the dataset.

Exercise 15.

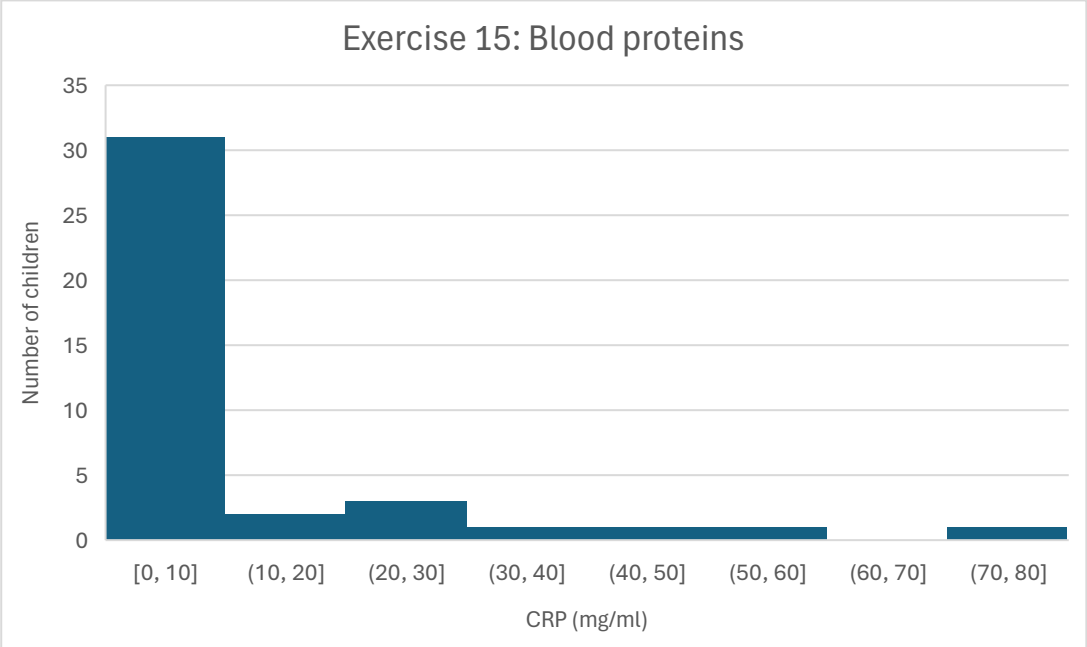
a) Five-number summary

Minimum:	0
Q1:	0
Median:	5.085
Q3:	9.42
Maximum:	73.20

b) Boxplot



c) Histogram



d) The distribution is extremely **right-skewed** mainly because of a large number of children having 0.00 mg/ml of CRP. Both plots can evidence this, but the histogram conveys the distribution better.

Exercise 16

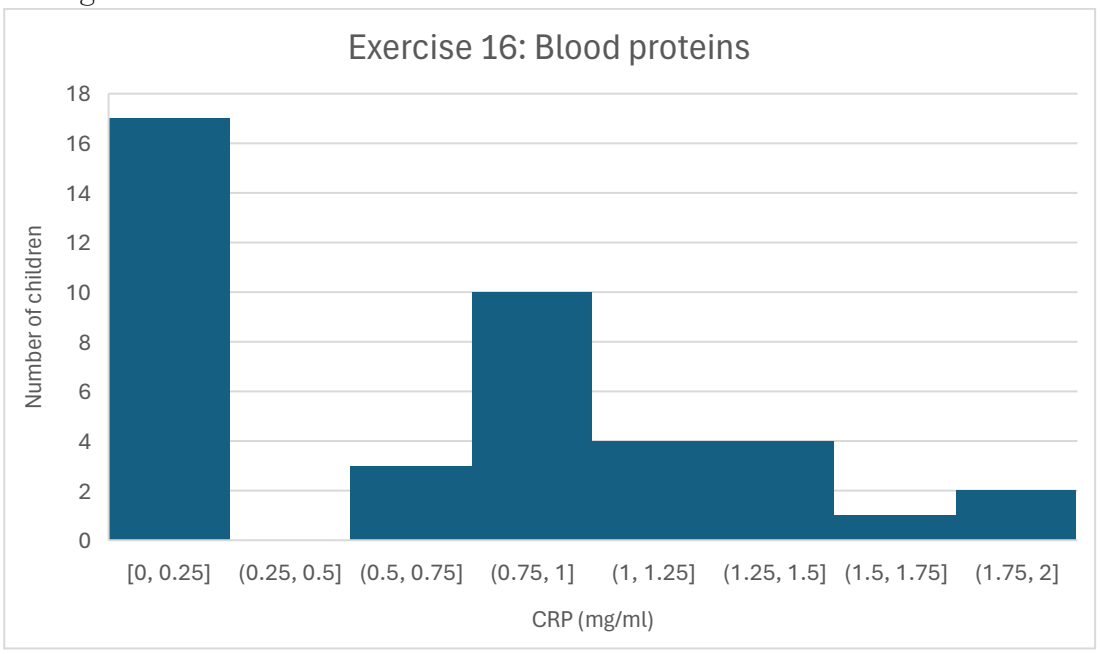
By adding 1 to all the previous values and then taking the logarithm, we present the new values and plots:

a) Five-number summary

Minimum: 0
Q1: 0
Median: 0.78
Q3: 1.02
Maximum: 1.87

b) Boxplot

c) Histogram



d) The distribution is still **right-skewed** but not as extreme as in *Exercise 15*. This is further evidenced by the new five-number summary, with the maximum number being closer to the median, because the range is now smaller.

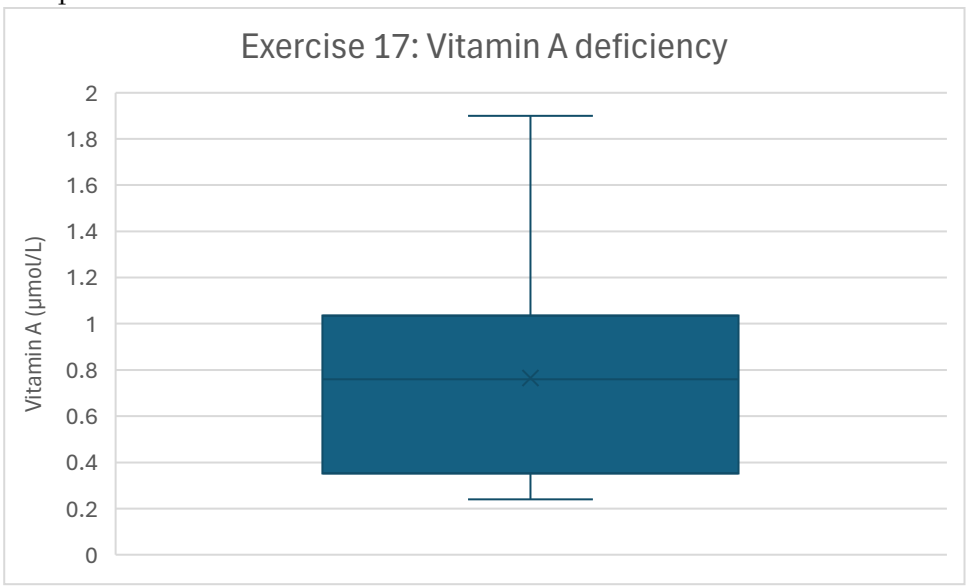
Therefore, the logarithmic transformation can improve the interpretation of data analysis.

Exercise 17

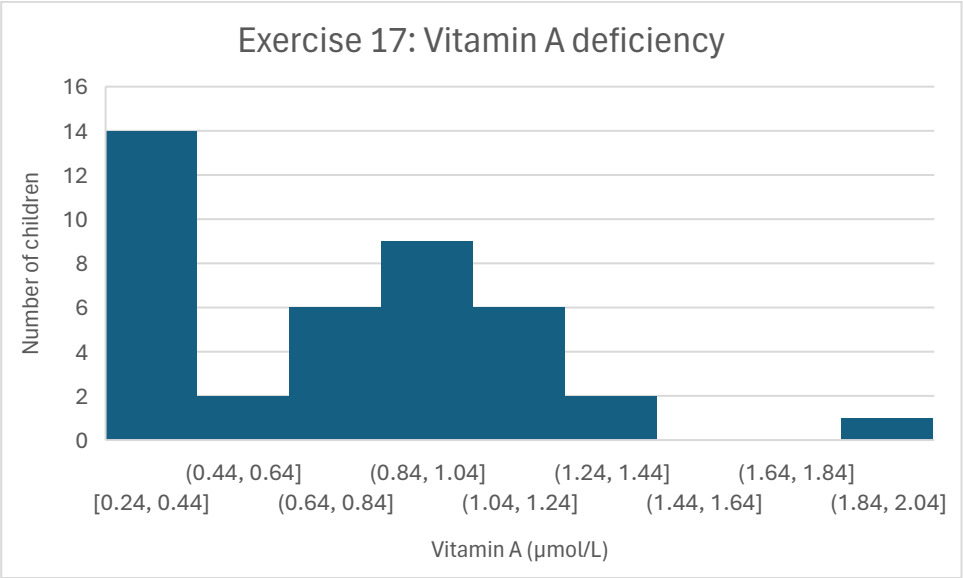
a) Five-number summary

Minimum:	0.24
Q1:	0.36
Median:	0.76
Q3:	1.03
Maximum:	1.90

b) Boxplot



c) Histogram



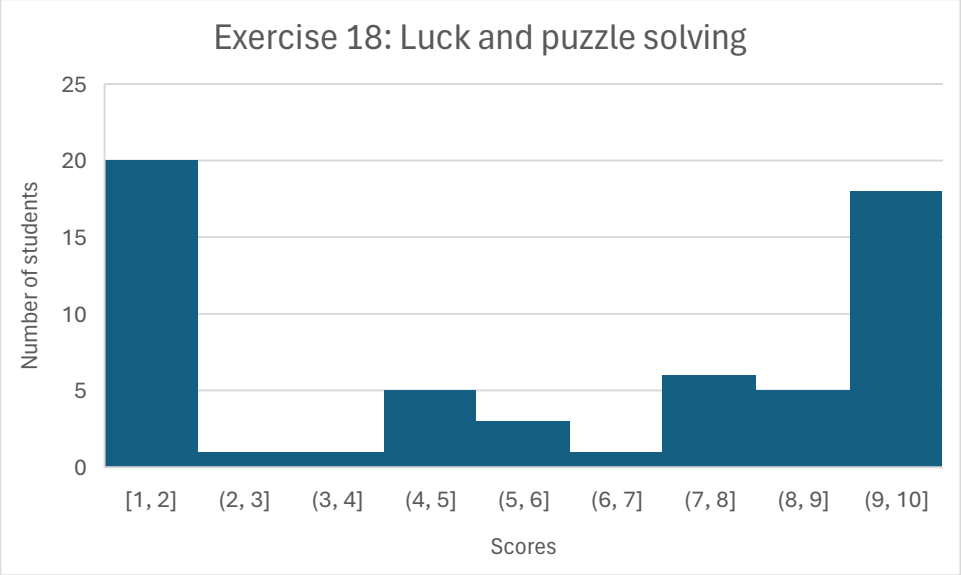
d) Overall, the distribution is relatively strong, **right-skewed** where the data indicate a predominantly low vitamin A status among the children, with only a few showing normal or elevated levels.

Exercise 18

Five-number summary

Minimum:	1
Q1:	2.3
Median:	6.5
Q3:	9.4
Maximum:	10

Mean	5.9
St. Dev.	3.77



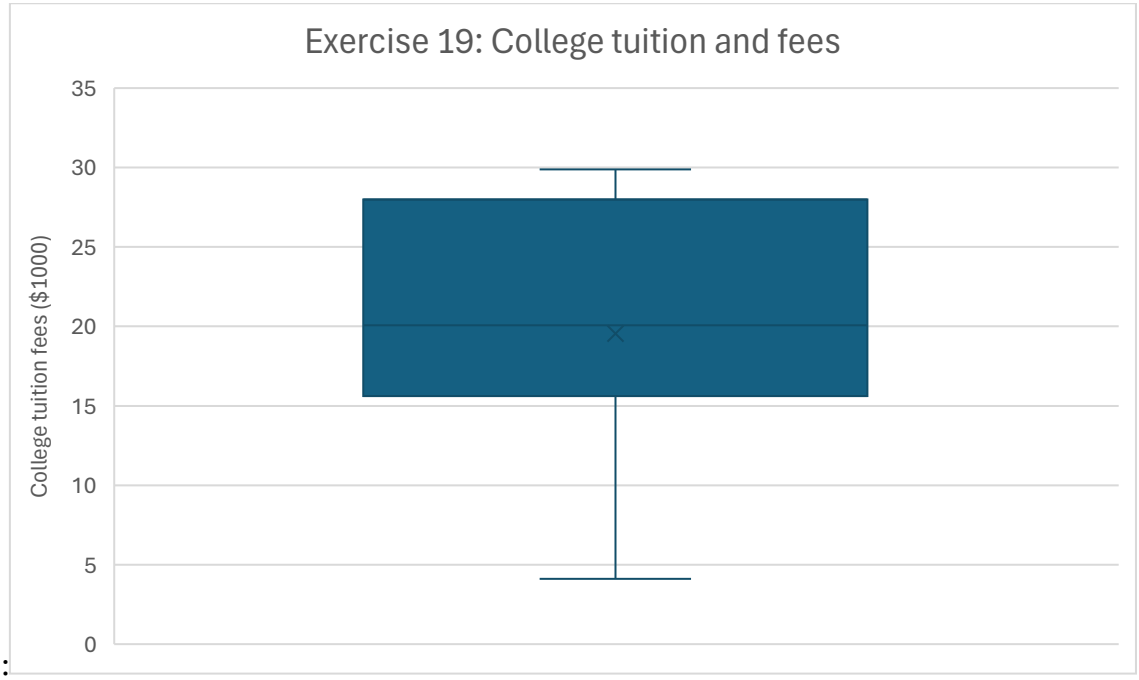
This indicates a **U-shaped (bimodal)** distribution, with peaks at both ends and a valley in the middle. The five-number summary suggests a shift towards being more "unlucky" (median: 6.5), but the dataset (and subsequently the histogram) clearly show the polarized distribution with most answers being either 1 or 10.

Exercise 19

a) Five-number summary

Minimum:	4.123
Q1:	15.826
Median:	20.072
Q3:	27.931
Maximum:	29.875

Boxplot



:

The boxplot and the five-number summary miss completely the bimodal nature of the distribution as depicted in the histogram, especially on the high end. The boxplot shows a continuous spread of tuition fees being relatively expensive, as opposed to the histogram which depicts three very distinct groups of schools.

Exercise 20

b) Five-number summary

Minimum:	5.7
Q1:	11.7
Median:	12.75
Q3:	13.48
Maximum:	17.6

c) $IQR = Q3 - Q1$

$IQR = 13.48 - 11.7 = \mathbf{1.78}$

IQR Rule: An outlier is defined when a value is lower than $Q1 - 1.5IQR$ or higher than $Q3 + 1.5IQR$

1.5 x IQR = **2.67**

Q1 – 2.67 = **9.03** and Q3 + 2.67 = **16.15**

Therefore, Alaska, with 5.7% and Florida, with 17.6%, which are less than 9.03% and more than 16.15%, are clearly identified as outliers by the 1.5xIQR rule. The only other state that is identified is the one with 8.5% which is also lower than 9.03%.