Problem set 02 – Solutions

Exercise 1.

Mean = $81,875

Median = $35,000

Everyone but the owner earns less than the mean wage.

In this example, the median value is more indicative of the salary distribution among the employees.

Exercise 2.

Mean changes to $98,750 while the median remains unchanged.

Exercise 3.

Stemplots:

Women		**Men**	
6	0	0	0
9	0	3	0 0 0 0
11	5	4	5
12	0 0 0 0 0 0 0	6	0 0 0
15	0 0 0 0	7	5
17	0	9	0 0 0 5
18	0 0 0 0 0 0 0 0 0 0	12	0 0 0 0 0 0 0
20	0	15	0 0
24	0 0 0	18	0
36	0	20	0 0
		23	0
		24	0 0
		30	0

Five-number summary

	Women	**Men**
Minimum	60	0
Q1	120	60
Median	175	120
Q3	180	150
Maximum	360	300

Mean and standard deviation

	Women	**Men**
Mean ($\bar{x}$)	165.16	117.17
Standard Deviation (σ)	56.51	74.23

a) Usually, the mean and standard deviation are most appropriate for distributions that are approximately symmetric without any strong outliers. Strong outliers tend to shift these parameters, thus complicating the distributions.

   For skewed or distributions with outliers, the median and IQR are generally more robust and better reflect a typical value and spread. Here, the stemplots we created earlier are very useful in identifying outliers, and since the distributions show skew and possible outliers, the median/IQR are safer summaries.

b) IQR = Q3 – Q1

Women: $IQR_{Women} = 180 - 120 = \mathbf{60}$

Men: $IQR_{Men} = 150 - 60 = \mathbf{90}$

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

Women

$1.5 \times IQR_{Women} = \mathbf{90}$

$Q1 - 90 = \mathbf{30}$ and $Q3 + 90 = \mathbf{270}$

So, 360 is an outlier ($>[Q3 + 1.5IQR_{Women}]$)

Men:

1.5 x IQR_{Men} = **135**

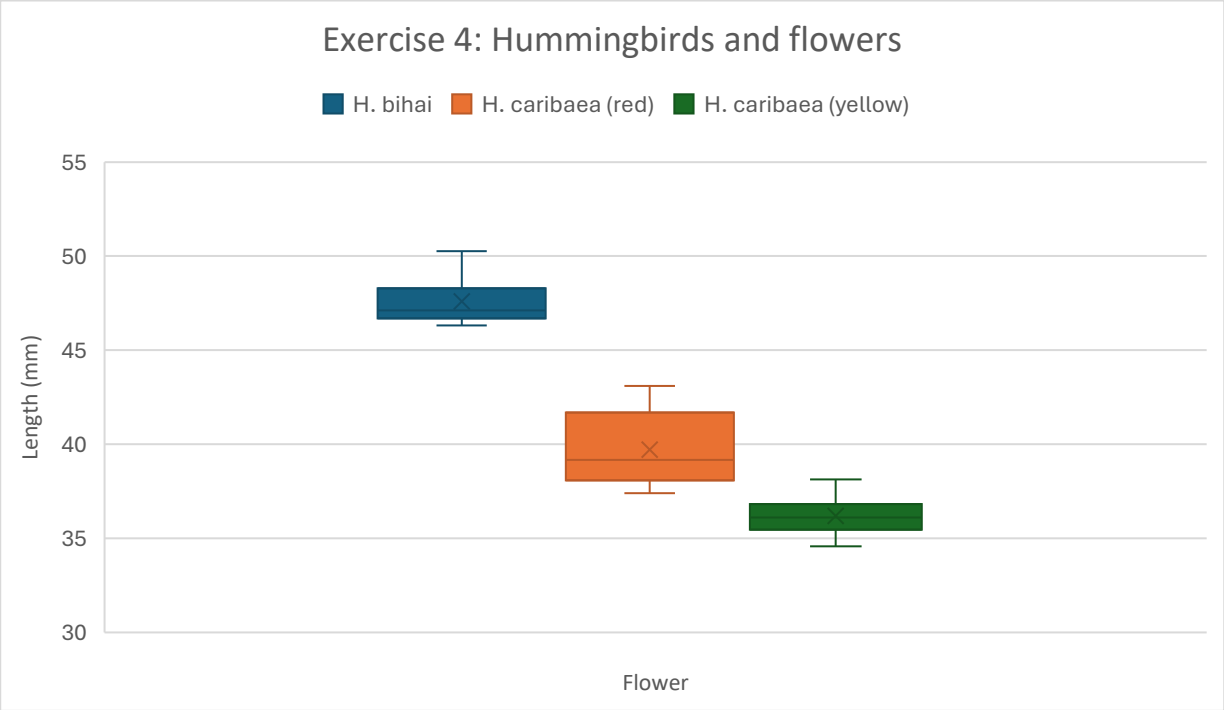
Q1 – 135 = **-75** and Q3 + 135 = **285**

So, 300 is an outlier ($>[Q3 + 1.5IQR_{Women}]$)

By removing the outlier from each group:

	Women	**Men**
Mean ($\bar{x}$)	158.45	110.86
Standard Deviation (σ)	43.65	66.88

Exercise 4.



Five-number summary

	H. binai	**H. caribaea (red)**	**H. caribaea (yellow)**
Minimum	46.32	37.40	34.57
Q1	46.37	38.09	35.57

Median	47.12	39.16	36.11
Q3	48.20	41.58	36.80
Maximum	50.26	43.09	38.13

The three varieties are very well separated with no overlap between the medians. *H. binai* are longer with a narrow with *H. caribaea (red)* having the widest spread and *H. caribaea (yellow)* being the shortest with the tightest spread and the most symmetrical distribution.

The boxplots are very useful here, as the three groups are very well separated, and the extraction of results for this data analysis is easy.

Exercise 5.

a) Height: $= 1 \times 1 = 1$
b) $0.35 - 0 = 0.35$ (35%)
c) $0.65 - 0.35 = 0.30$ (30%)

Exercise 6.

a) The area of the uniform distribution must always account for 1, so height $= \frac{1}{4} = 0.25$
b) $1 \times 0.25 = 0.25$ (25%)
c) $(2.5 - 0.5) \times 0.25 = 0.5$ (50%)

Exercise 7.

a) $\mu = 0.5$, median $= 0.5$
b) $Q1 = 0.25$, $Q3 = 0.75$

Note

A uniform distribution is *always* symmetric, so the mean, the median, and the quartiles are evenly spaced, and the values are the same.

Exercise 8.

We know that for a normal distribution:

σ: 68%, 2σ: 95%, 3σ: 99.7%

and for this exercise, we have:

$\mu = 336$ days, $\sigma = 3$ days or it can be written as 336 ± 3 days

a) For 99.7% pregnancies:

$336 \pm 3\sigma$, so $336 - 3$ x $3 =$ **327** and $336 + 3$ x $3 =$ **345** so the range is **[327, 345]**

Thus, almost all horse pregnancies are between 327 and 345 days.

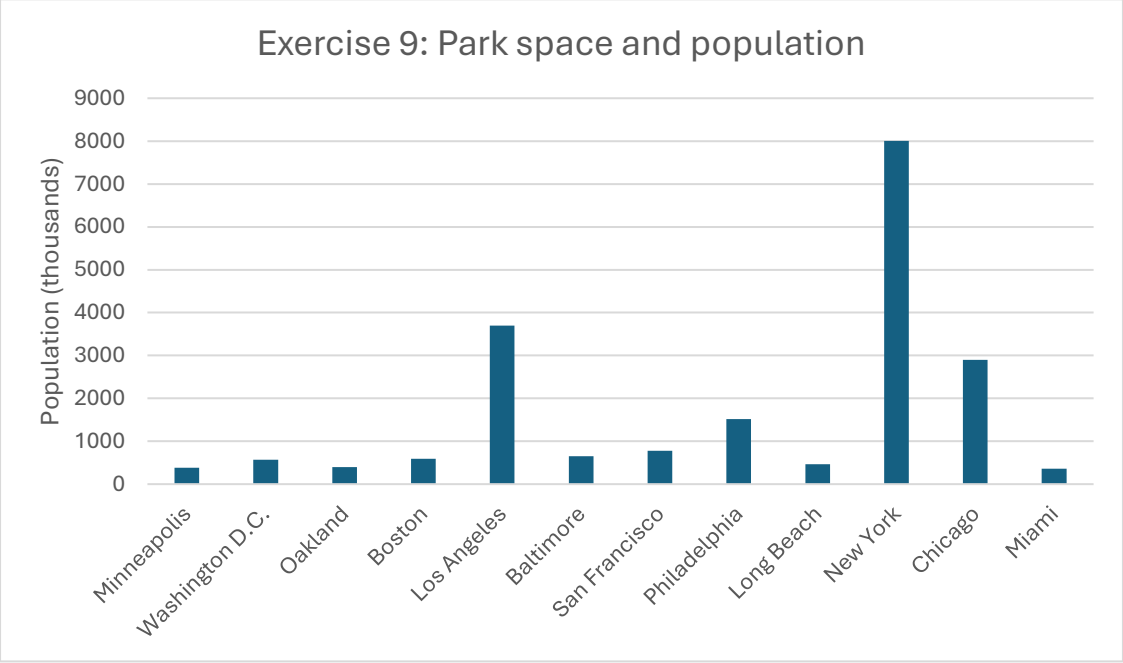
b) We want >339 days:

339 is $\mu + \sigma$ (68% of the population)

100% - 68% = 32%, divided by 2 as we want only the positive part so **16%**

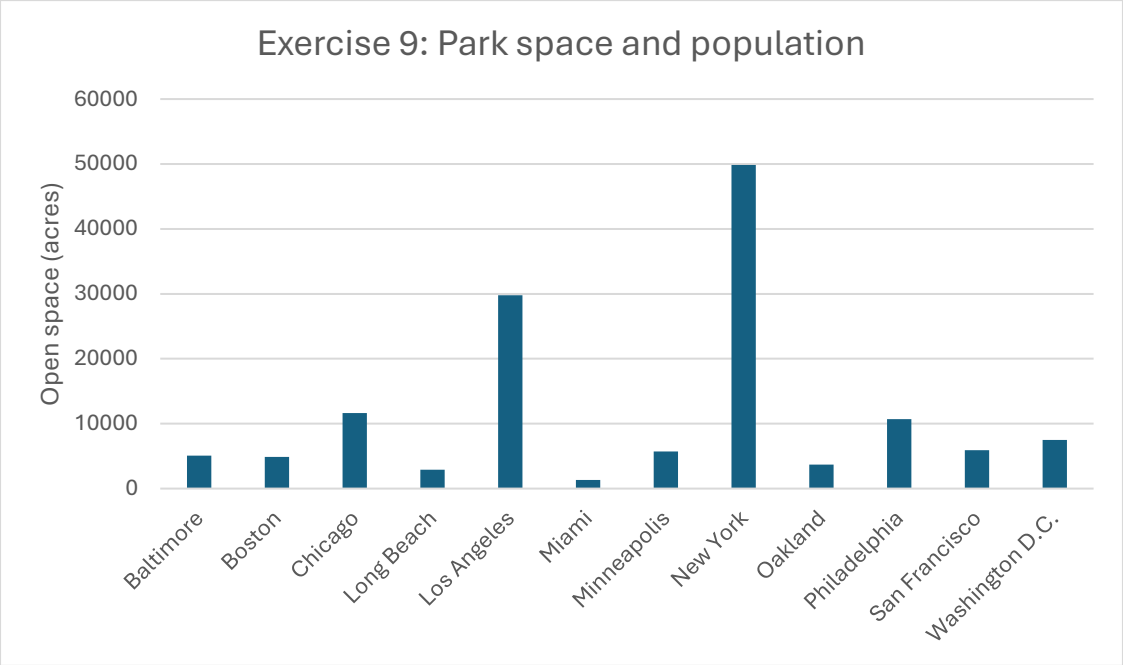
Exercise 9.

a) For the population graph



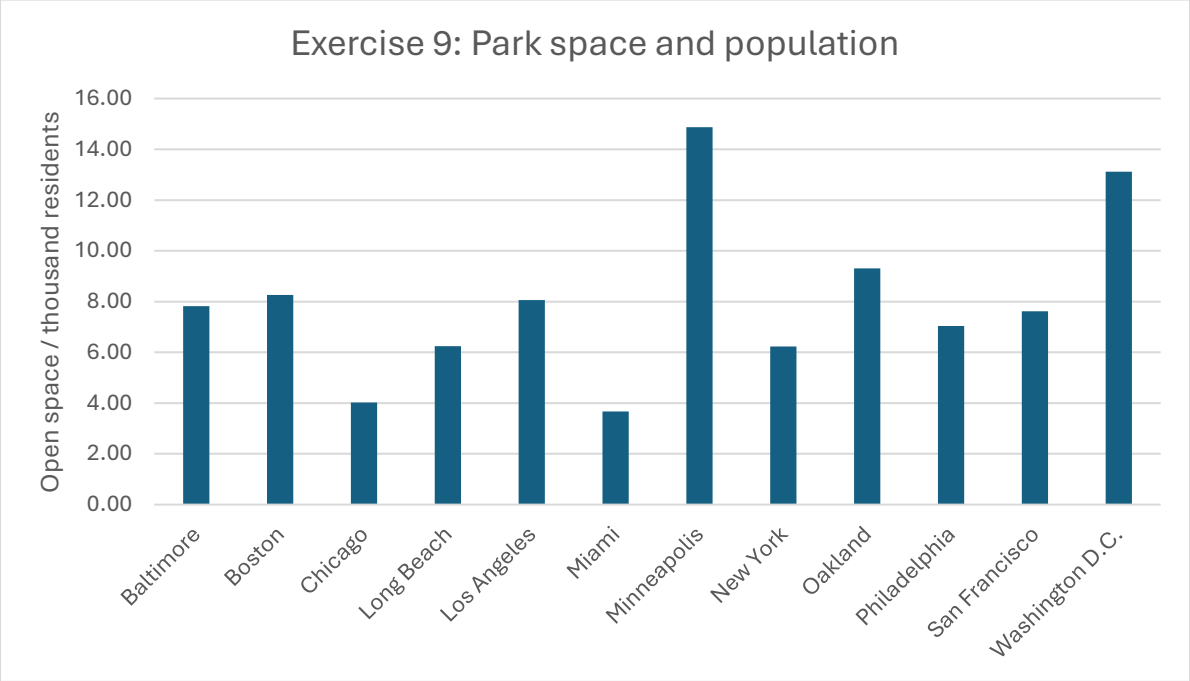
New York has the highest population, followed by Los Angeles and Chicago third.

b) For the open space graph



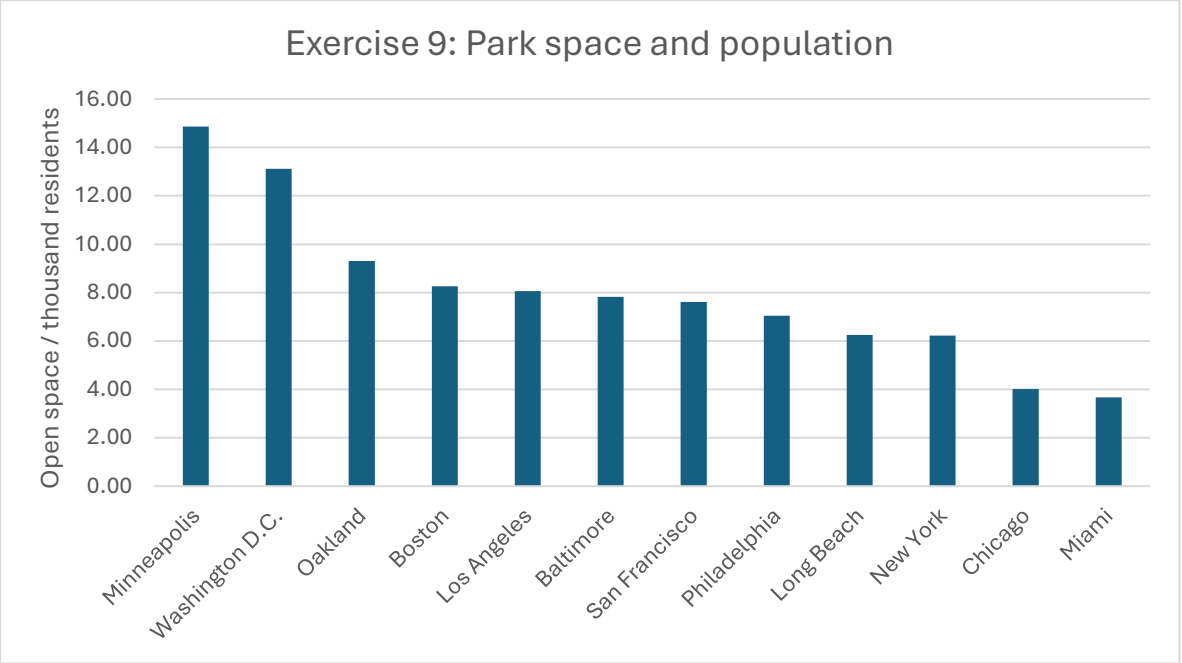
The same trend as with the population graph can be observed.

c) We divide open space by population
d) Bar graph of acres of open space per thousand residents



This is the true value we are interested in. Now, we see that Minneapolis has the highest open space per thousand residents, followed by Washington D.C. The three largest cities from before have dropped in the rankings.

e) Bar graph ordered from highest to lowest open space per thousand residents.



f) The last graph makes comparisons and rankings among the U.S. states much easier, where we can immediately define which is larger and draw conclusions *between* the states. However, by ordering the bar graph in alphabetical order, we can directly pinpoint the U.S. state we are interested in. Overall, it depends on what we are actually looking for.

Exercise 10.

Women: 579 ± 49

Men: 565 ± 55

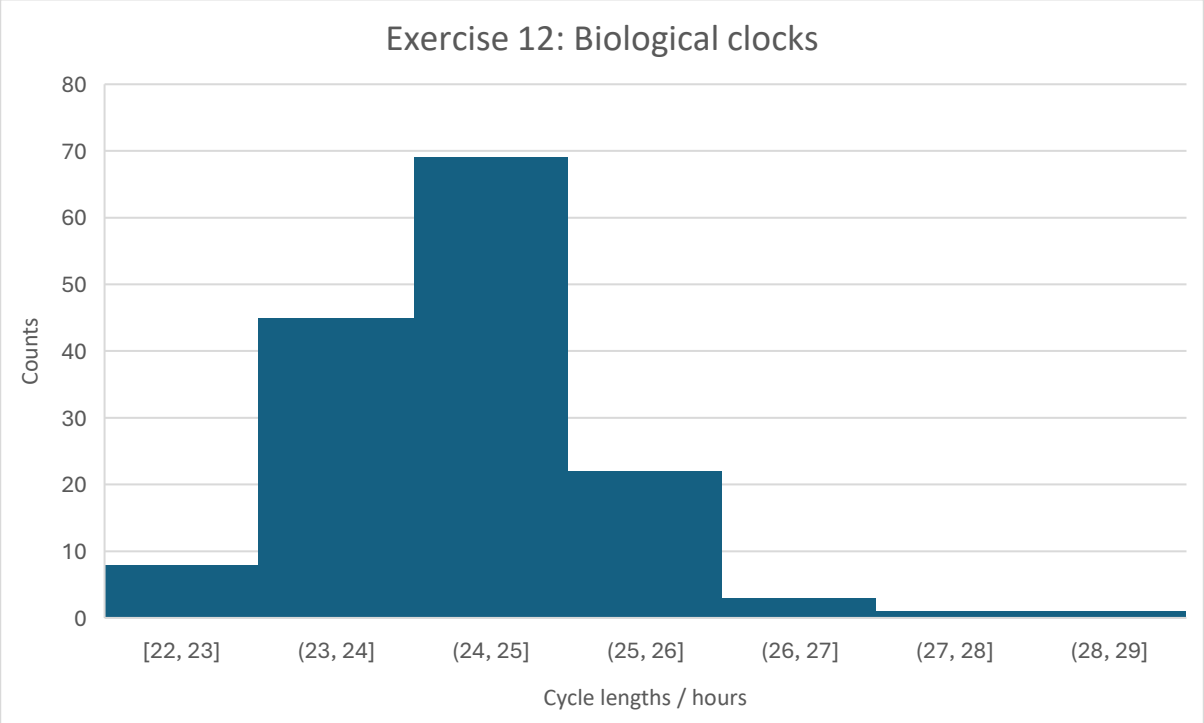
The statement that women score about 14 points higher than men on the exam is **partially true.** Indeed, judging from the mean values, the difference is 14, but the two graphs are overlapping and we can see that some men have scored higher points than some women. Thus, the term is a bit exaggerated and it should be corrected to *women score higher than men on average.*

Exercise 11.

Total (n): 149

Five-number summary

Minimum	22.00
Q1	23.74
Median	24.31
Q3	24.82
Maximum	28.55



The distribution is **slightly right-skewed** because of two outliers, mainly. The distribution is tight around the median of 24.31 with a very small spread.

Exercise 12.

Gender: Categorical

Age: Quantitative

Income: Quantitative

Preference: Categorical

Exercise 13.

Mode: The most frequent value

Median: The "middle" value

This happens because many students are 19 years old but fewer than half (because of the median).

Exercise 14.

3^{rd} graders: 75 ± 10

6^{th} graders: 82 ± 11

We need 100 ± 20

Equation: $x_{new} = a + bx$

a) For 3^{rd} graders:

$$b_{new} = \frac{\sigma_{new}}{\sigma_{old}} = \frac{20}{10} \Leftrightarrow b_{new} = 2$$

$$a_{new} = \mu_{new} - b\mu_{old} = 100 - 2x75 \Leftrightarrow a_{new} = -50$$

$$x_{new} = -50 + 2x$$

b) Similarly for 6^{th} graders:

$$b = 1.818$$

$$a = -49.08$$

$$x_{new} = -49.08 + 1.818x$$

c) David (3^{rd} grader)
$x_{new} = -50 + 2x78 \Leftrightarrow$
$x_{new} = 106$

Nancy (6^{th} grader)
$x_{new} = -49.08 + 1.818x78 \Leftrightarrow$
$x_{new} = 92.72$

d) *Note: Because the transformation of the distributions is linear, using both new and old distributions will give the same result.*

3rd graders

$$z_{3rd} = \frac{x - \mu}{\sigma} = \frac{\mathbf{106 - 100}}{\mathbf{20}} \Leftrightarrow$$

$$\boldsymbol{z_{3rd} = 0.3}$$

6th graders

Similarly:

$$z_{6th} = \frac{92.72 - 100}{20} = \frac{\mathbf{106 - 100}}{\mathbf{20}} \Leftrightarrow$$

$$\boldsymbol{z_{6th} = -0.364}$$

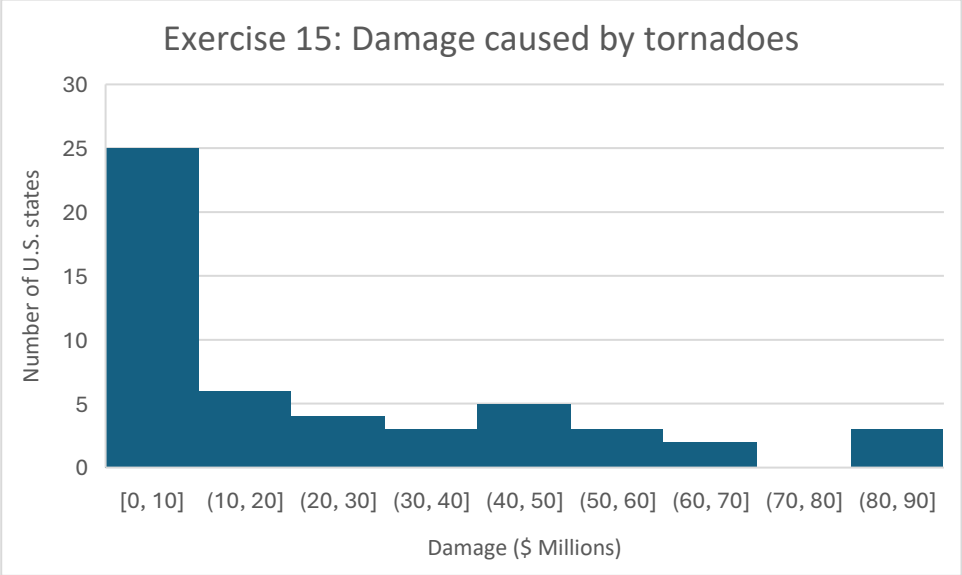
By checking Table A in Pages T-2 and T-3 of the book we find out that the percentage for scores below 78 are:

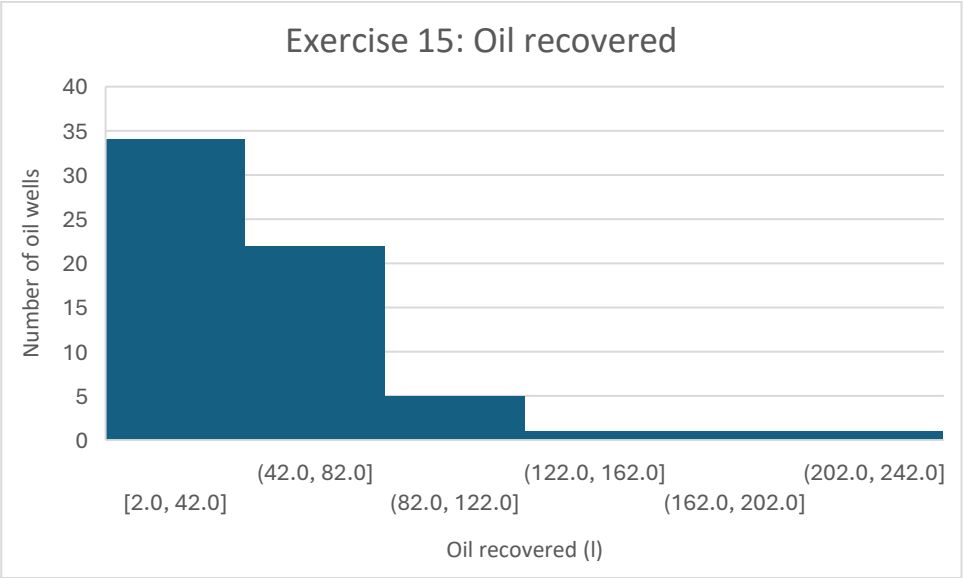
For 3rd graders: **61.8%**

For 3rd graders: **35.8%**

Exercise 15.

Looking at both histograms, we see that most data is centered towards lower values with longer tails to the right (less counts in higher values)




Exercise 15: Oil recovered
Number of oil wells
40
35
30
25
20
15
10
5
0
[2.0, 42.0]
(42.0, 82.0]
(82.0, 122.0]
(122.0, 162.0]
(162.0, 202.0]
(202.0, 242.0]
Oil recovered (l)