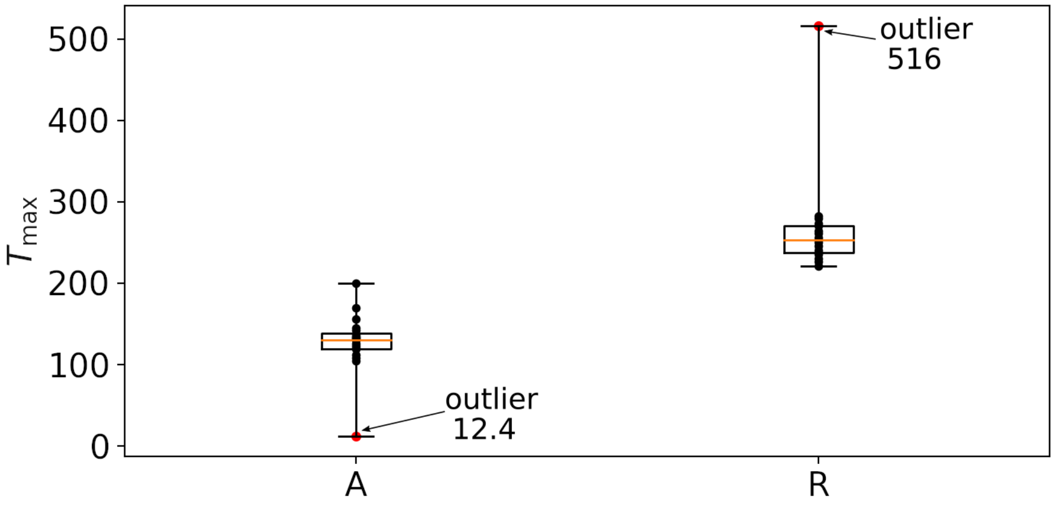
Question 1: The one about drug therapy

(a) Construct a side-by-side box plot and identify outliers.

First, we need summaries for two sets(A and R), respectively.

	Minimum	Q1	Median	Q3	Maximum
$T_{\max}(A)$	12.4	119.5	130	139	200
$T_{\max}(R)$	221	238	253	271	516

Then, we can create boxplots for them. Here, the maximum and minimum are used as the ends of the whiskers.

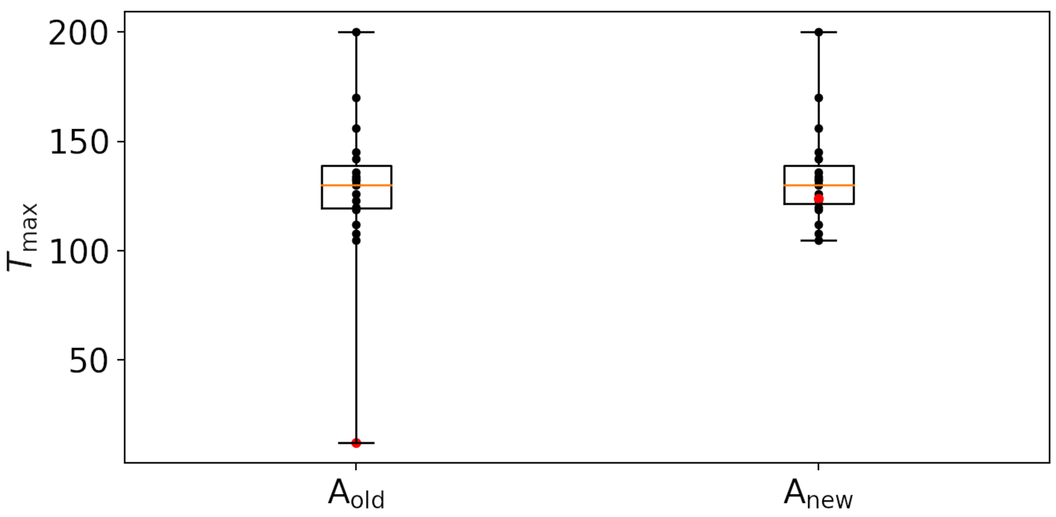


As you can see, the outlier in A set is 12.4, and the outlier in R set is 516.

(b) Calculate $\overline{x}$ and s^2 for the data of set A.

- $\overline{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{105+126+120+\cdots+170}{19} = 128.07$
- $$s^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1} = \frac{(105-128.07)^2+(126-128.07)^2+\cdots+(170-128.07)^2}{18} = 1282.68$$

(c) Assume that the outlier of set A is the result of a misplaced decimal point. Correct the error by deleting the decimal and see what changes this makes in your box plot. Recompute $\overline{x}$ and s^2 , using the correct data point, and compare your results to those of part b.



- $\overline{x} = 133.95$
- $s^2 = 503.83$

After correcting the decimal point of the outlier, the new $\overline{x}$ is larger than the old $\overline{x}$ because there is no small value. And the new s^2 is smaller than the old s^2 because the spread of distribution becomes small.

(d) Is there an outlier in set R? If so, is there an obvious legitimate reason to delete it from the data set.

- Yes, there is an outlier(516) in set R, as you can see in the first figure.<\li>
- I think this outlier should be deleted because 516 is much larger than other values. This value is not possible for healthy adults in comparison with other data points.

Question 2: The data sets below are temperature readings from two different sensors (a Celsius sensor and a Fahrenheit sensor). Which sensor is better by picking the one with the least variance.

Intuitively, it is not appropriate for directly comparing these two sensors by standard deviation because they have different units. Therefore, it is better to use coefficient of variation(CV) to compare them. The definition of CV is: $\text{CV} = \frac{S}{\overline{X}} \times 100\%$

First, we need to calculate S and $\overline{X}$ for them:

- $\overline{X}_{\text{Celsius}} = 20$
- $S_{\text{Celsius}} = 15.81$
- $\overline{X}_{\text{Fahrenheit}} = 68$
- $S_{\text{Fahrenheit}} = 28.46$

Finally, we get

- $\text{CV}_{\text{Celsius}} = 79\%$
- $\text{CV}_{\text{Fahrenheit}} = 42\%$

The better sensor is the Fahrenheit one.

But, this is totally wrong!

The coefficient of variation may not have any meaning for data on an "interval scale". In other words, the zero degree Celsius or the zero degree Fahrenheit do not represent the absence of thermal energy. We need something which is a "ratio scale", that is, scale that has a meaningful zero and hence allow relative comparison of two measurements. For temperature, the only choice is Kelvin, in which, the 0 K stands for the complete absence of thermal energy.

- Celsius to Kelvin: [0, 10, 20, 30, 40] → [273.15, 283.15, 293.15, 303.15, 313.15]
- Fahrenheit to Kelvin: [32, 50, 68, 86, 104] → [273.15, 283.15, 293.15, 303.15, 313.15]

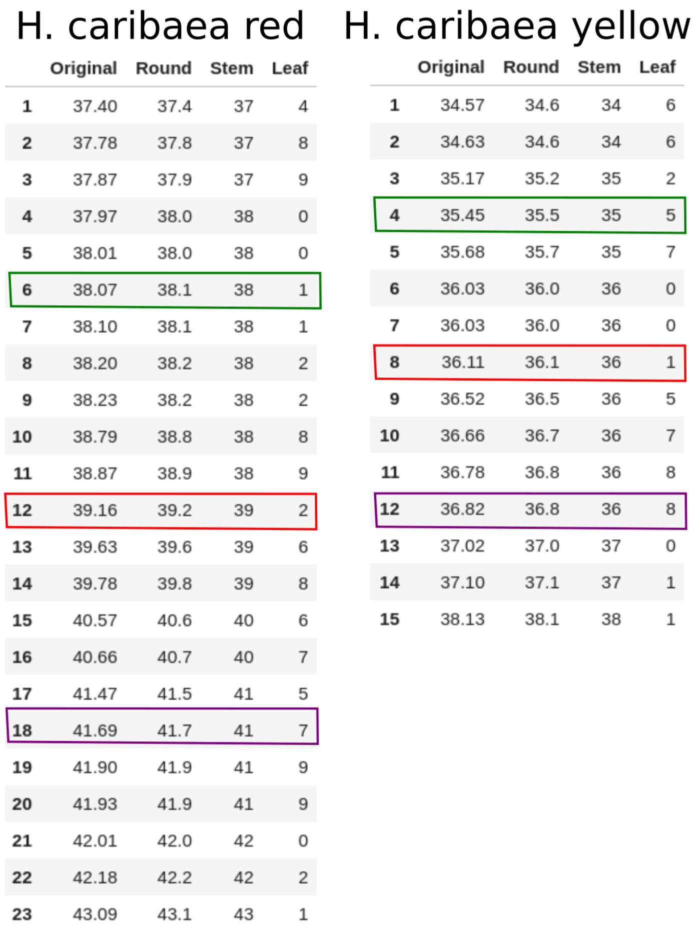
As you can see, these two sensors are same, and their mean and standard deviation:

- $\overline{x} = 293.15\ \mathrm{K}$
- $S = 15.81\ \mathrm{K}$

Question 3: Coevolution

(a) Make a back-to-back stemplot to compare the two samples. That is, use one set of stems with two sets of leaves, one to the right and one to the left of the stems. (Draw a line on either side of the stems to separate stems and leaves.) Order both sets of leaves from smallest at the stem to largest away from the stem. (Please round the data to the first decimal place. Use non-fractional part as the stem, first decimal place as leaf. Ex: 41.93 → 41.9 → stem = 41, leaf = 9; 38.79 → 38.8, stem = 38, leaf = 8).

First, I create tables for each dataset for the convenience of making stemplot.



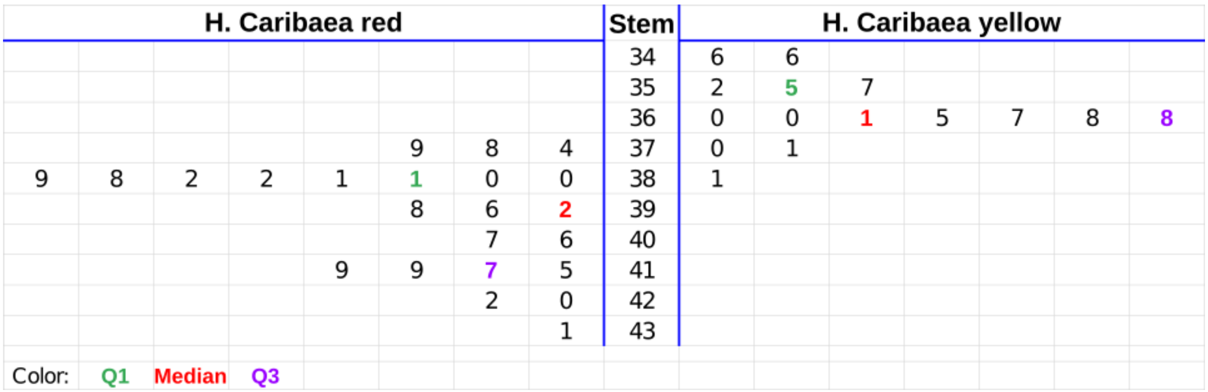
H. caribaea red

	Original	Round	Stem	Leaf
1	37.40	37.4	37	4
2	37.78	37.8	37	8
3	37.87	37.9	37	9
4	37.97	38.0	38	0
5	38.01	38.0	38	0
6	38.07	38.1	38	1
7	38.10	38.1	38	1
8	38.20	38.2	38	2
9	38.23	38.2	38	2
10	38.79	38.8	38	8
11	38.87	38.9	38	9
12	39.16	39.2	39	2
13	39.63	39.6	39	6
14	39.78	39.8	39	8
15	40.57	40.6	40	6
16	40.66	40.7	40	7
17	41.47	41.5	41	5
18	41.69	41.7	41	7
19	41.90	41.9	41	9
20	41.93	41.9	41	9
21	42.01	42.0	42	0
22	42.18	42.2	42	2
23	43.09	43.1	43	1

H. caribaea yellow

	Original	Round	Stem	Leaf
1	34.57	34.6	34	6
2	34.63	34.6	34	6
3	35.17	35.2	35	2
4	35.45	35.5	35	5
5	35.68	35.7	35	7
6	36.03	36.0	36	0
7	36.03	36.0	36	0
8	36.11	36.1	36	1
9	36.52	36.5	36	5
10	36.66	36.7	36	7
11	36.78	36.8	36	8
12	36.82	36.8	36	8
13	37.02	37.0	37	0
14	37.10	37.1	37	1
15	38.13	38.1	38	1

Then, we can create the stemplot



H. Caribaea red								Stem	H. Caribaea yellow						
								34	6	6					
								35	2	5	7				
								36	0	0	1	5	7	8	8
					9	8	4	37	0	1					
9	8	2	2	1	1	0	0	38	1						
					8	6	2	39							
						7	6	40							
				9	9	7	5	41							
						2	0	42							
							1	43							
Color:	Q1	Median	Q3												

(b) What are the most important differences among the two varieties of flower?

From the stemplot, we found that the length of the red flower is generally longer than the length of the yellow flower.

(c) Find Q1, Median and Q3 for each group.

As you can see from the above dataframe or the stem plot.

	Q1	Median	Q3
H. caribaea red	38.07	39.16	41.69
H. caribaea yellow	35.45	36.11	36.82