Measures of Central Tendency:

One value will be used to characterize or summarize an entire data set. In the case of numerical data, it's thought to represent the center or middle of the values.

Some data sets are very large with a wide range of values, so trying to characterize all of the values with one value is a pretty bold undertaking!

Our textbook discusses three such measures of center:

Mean

Median

Mode

The *Mean* of a list of numerical data $\{x_1, x_2, x_3, \ldots, x_n\}$ **is defined to be** $\dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$.

The formula is sometimes written in abbreviated form as $\dfrac{\sum x}{n}$**, where the Greek letter** $\sum$ **is the abbreviation for add them up.**

If the data set represents a sample, the mean value is abbreviated as $\bar{x}$ **and called the sample mean.**

If the data set represents an entire population, the mean value is abbreviated with the Greek letter μ **and called the population mean.**

Examples of Mean Calculations:

1. $\{1, 2, 2, 3\}$

$$\frac{1+2+2+3}{4} = \frac{8}{4} = \boxed{2}$$

Sometimes the value of the mean will be an actual value in the data set.

2. $\{1, 2, 2, 3, 4\}$

$$\frac{1+2+2+3+4}{5} = \frac{12}{5} = \boxed{2.4}$$

Sometimes the value of the mean won't be an actual value in the data set.

If you give a rounded value for the mean, always round to one more decimal place than the data values.

3.

0	2 3
1	0 3 9
2	1 1

$$\frac{2+3+10+13+19+21+21}{7}=\frac{89}{7}=\boxed{12.7}$$

4.

Value	Frequency
1	**11**
2	**13**
3	**22**
Total	**46**

$$\frac{1+1+\cdots+1+2+2+\cdots+2+3+3+\cdots+3}{46}=$$

$$\frac{1\cdot 11+2\cdot 13+3\cdot 22}{46}=\frac{11+26+66}{46}=\boxed{2.2}$$

The repeated additions can be done by multiplying each value with its frequency.

To determine the *Median* of a list of numerical data, you first arrange the values in order.

If there is an odd number of values in the data set, the median is the middle value.

If there is an even number of values in the data set, the median is the average of the middle two values.

Examples of Median Calculations:

1. $\{5, 2, 6, 3, 8\}$

2,3,5,6,8

The middle value is 5, so the median is 5.

When the number of values is odd, the median value will definitely be an actual value in the data set.

2. $\{5, 2, 6, 3, 8, 10\}$

2,3,5,6,8,10

The middle two values are 5 and 6, so the median is $\frac{5+6}{2} = \boxed{5.5}$.

When the number of values is even, the median value might not be an actual value in the data set.

<u>There is another method for finding the median that works well for larger data sets.</u>

If the data set has *n* values, put the values in order.

If *n* is odd, the median is the value in the $\frac{n+1}{2}$ **position.**

If *n* is even, the median is the average of the values in the $\frac{n}{2}$ **and** $\frac{n}{2}+1$ **positions.**

More Examples of Median Calculations:

3. $\{5,2,6,3,8\}$

$2,3,5,6,8$

$n=5$

So the median is the value in position $\frac{5+1}{2}=3$, and the third value is 5.

4. $\{5,2,6,3,8,10\}$

$2,3,5,6,8,10$

$n=6$

So the median is the average of the values in positions $\frac{6}{2}=3$ and $\frac{6}{2}+1=4$. The values n the third and fourth positions are 5 and 6, so the median is $\frac{5+6}{2}=\boxed{5.5}$.

5.

3	2	3	5	8	9	9	
4	0	3	4	7	9	9	
5	1	1	2	2	8	9	
6	2	2	7	7	8	9	9

$n = 25$

So the median is the value in position $\frac{25+1}{2} = 13$, so the median is 51.

6.

3	2	3	5	8	9	9
4	0	3	4	7	9	9
5	1	1	2	2	8	9
6	2	2	7	7	8	9

$n = 24$

So the median is the average of the values in positions $\frac{24}{2} = 12$ and $\frac{24}{2} + 1 = 13$. These two values are 49 and 51, so the median is $\frac{49+51}{2} = \boxed{50}$.

7.

Value	Frequency
1	12
2	13
3	22
Total	47

$n = 47$

So the median is the value in position $\frac{47+1}{2} = 24$, so the median is 2.

8.

Value	Frequency
1	12
2	11
3	23
Total	46

$n = 46$

So the median is the average of the values in positions $\frac{46}{2} = 23$ and $\frac{46}{2} + 1 = 24$. These two values are 2 and 3, so the median is $\frac{2+3}{2} = \boxed{2.5}$.

The *Mode* of a list of numerical or descriptive data is defined to be the most frequently occurring value.

Unlike the textbook, I don't allow for two modes(bimodal), or three modes(trimodal), etc.. For me either there is one most frequently occurring value or there isn't.

Mode Examples:

1. $\{1, 2, 2, 3\}$

The mode is 2.

2. $\{1, 1, 2, 2, 3\}$

There is no mode.

Sometimes a data set won't have a mode.

3. $\{red, blue, green, green\}$

The mode is *green*.

Means and Medians only work for numerical data, but this isn't the case for Modes.

4.

3	2	3	5	8	9	9	
4	0	3	4	7	9	9	
5	1	1	2	2	8	9	
6	2	2	7	7	7	9	9

The mode is 67.

5.

Value	Frequency
1	12
2	11
3	23
Total	46

The mode is 3.

6.

Value	Frequency
Small	10
Medium	11
Large	40
X-Large	23
Total	84

The mode is Large.

See the link Measures of Center.

One more!

The following ordered stem-and-leaf plot shows the temperatures of patients who registered at a clinic on a particular day.

97	2 8
98	1 3 3 5 6 7 7 7
99	0 0 0 4
100	0 2 4 4 6
101	5 9
102	3 3
103	1

Key: 98 | 3 = 98.3°F

1. How many patients registered at the clinic that day?

24

2. What is the mode of the temperatures?

No mode

3. What is the median of the temperatures?

The average of the 12th and 13th values is $\frac{99.0+99.0}{2} = \boxed{99.0°}$

4. If the sum of the temperatures is $2392°F$, then what is the mean of the temperatures? $\frac{2392}{24} = \boxed{99.67°}$

Box Plots or Box and Whisker Plots:

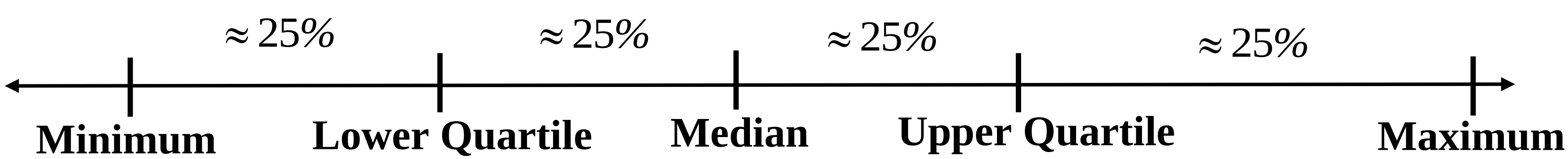
A Box Plot is a visual presentation of five numbers associated with a data set called the 5-Number Summary.

The five numbers are the Minimum, the Lower Quartile(Q_1), the Median(Q_2), the Upper Quartile(Q_3), and the Maximum.

The *Median* divides the data set roughly in half separating the lower 50% from the upper 50%.

The *Lower Quartile* is the median of the lower 50%, roughly dividing the data values into a lower 25% and an upper 75%.

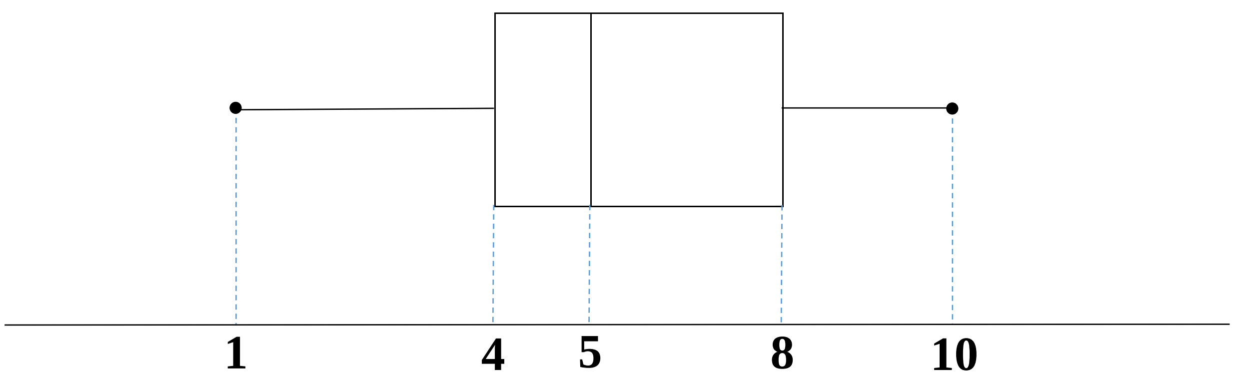
The *Upper Quartile* is the median of the upper 50%, roughly dividing the data values into a lower 75% and upper 25%.



Suppose that the 5-Number Summary of a data set is

Minimum	Q_1	Q_2	Q_3	Maximum
1	4	5	8	10

Here's the Box Plot



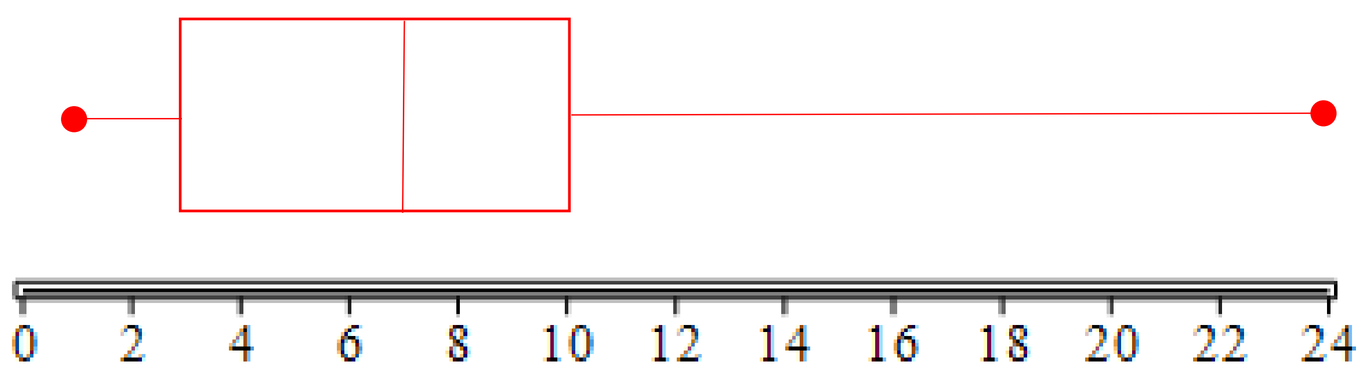
Find the 5-Number Summary, and create a Box Plot for the following Data Lists.

1. $\{3,8,6,1,24,10\}$

$Q_2 = 7$

	Q_1			Q_3	
1	**3**	**6**	**8**	**10**	**24**

Minimum	Q_1	Q_2	Q_3	Maximum
1	3	7	10	24

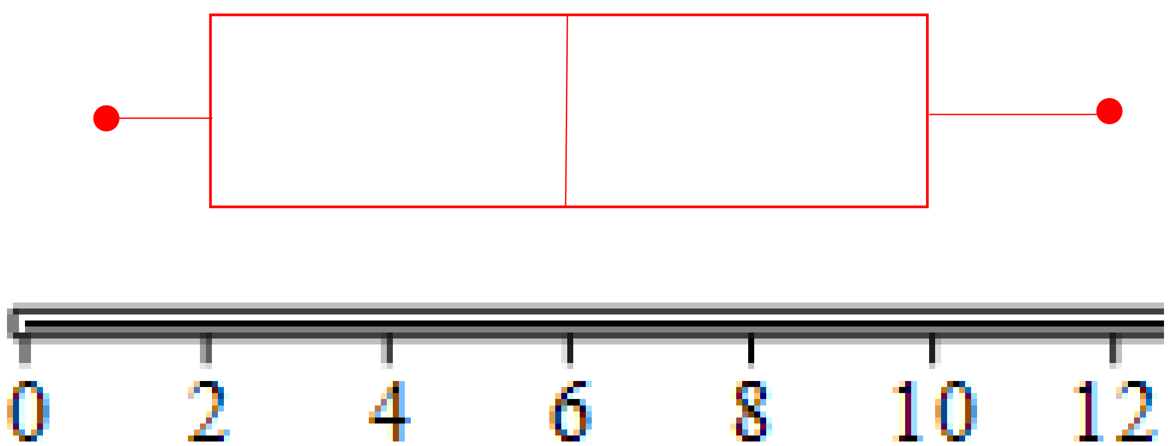


2. $\{3,8,6,1,12\}$

$Q_1 = 2$ $\qquad$ $Q_3 = 10$

		Q_2		
1	**3**	**6**	**8**	**12**

Minimum	Q_1	Q_2	Q_3	**Maximum**
1	2	6	10	12



The Interquartile Range(IQR):

The difference between the Upper Quartile(Q_3) and the Lower Quartile(Q_1) is called the Interquartile Range or *IQR*. $IQR = Q_3 - Q_1$**.**

It represents the width of an interval that roughly contains the middle 50% of the data values.

The *IQR* is used as a gauge for abnormality in a data set. Values in the data set that are smaller than $Q_1 - 1.5IQR$ **or larger than** $Q_3 + 1.5IQR$ **are considered unusual and are called outliers.**

For the data set $\{3,2,8,6,1,8,24,10\}$, determine the 5-number summary. Find the value of the *IQR*, and determine if any of the data values are outliers.
Make a fancy boxplot that indicates any outliers with an asterisk.

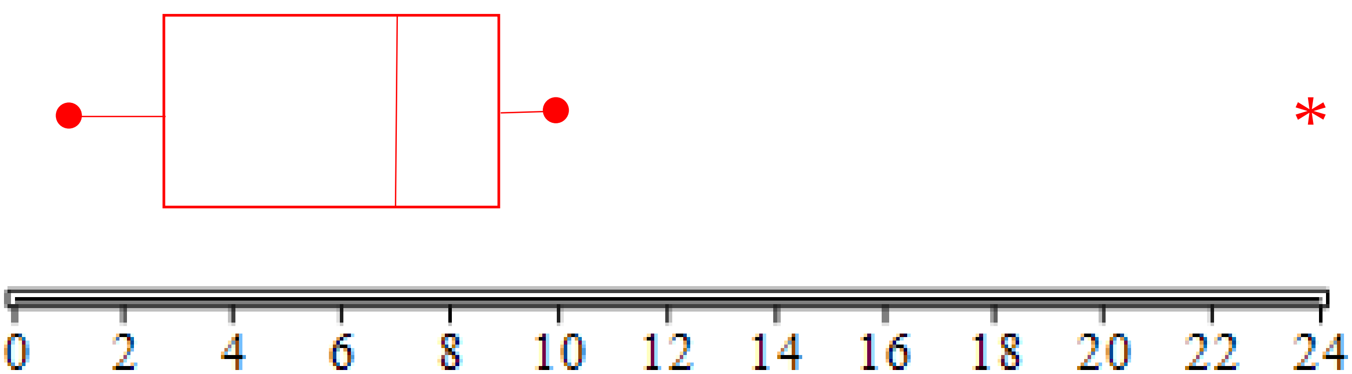
$Q_1 = 2.5$ $Q_2 = 7$ $Q_3 = 9$

1	**2**	**3**	**6**	**8**	**8**	**10**	**24**

Minimum	Q_1	Q_2	Q_3	**Maximum**
1	2.5	7	9	24

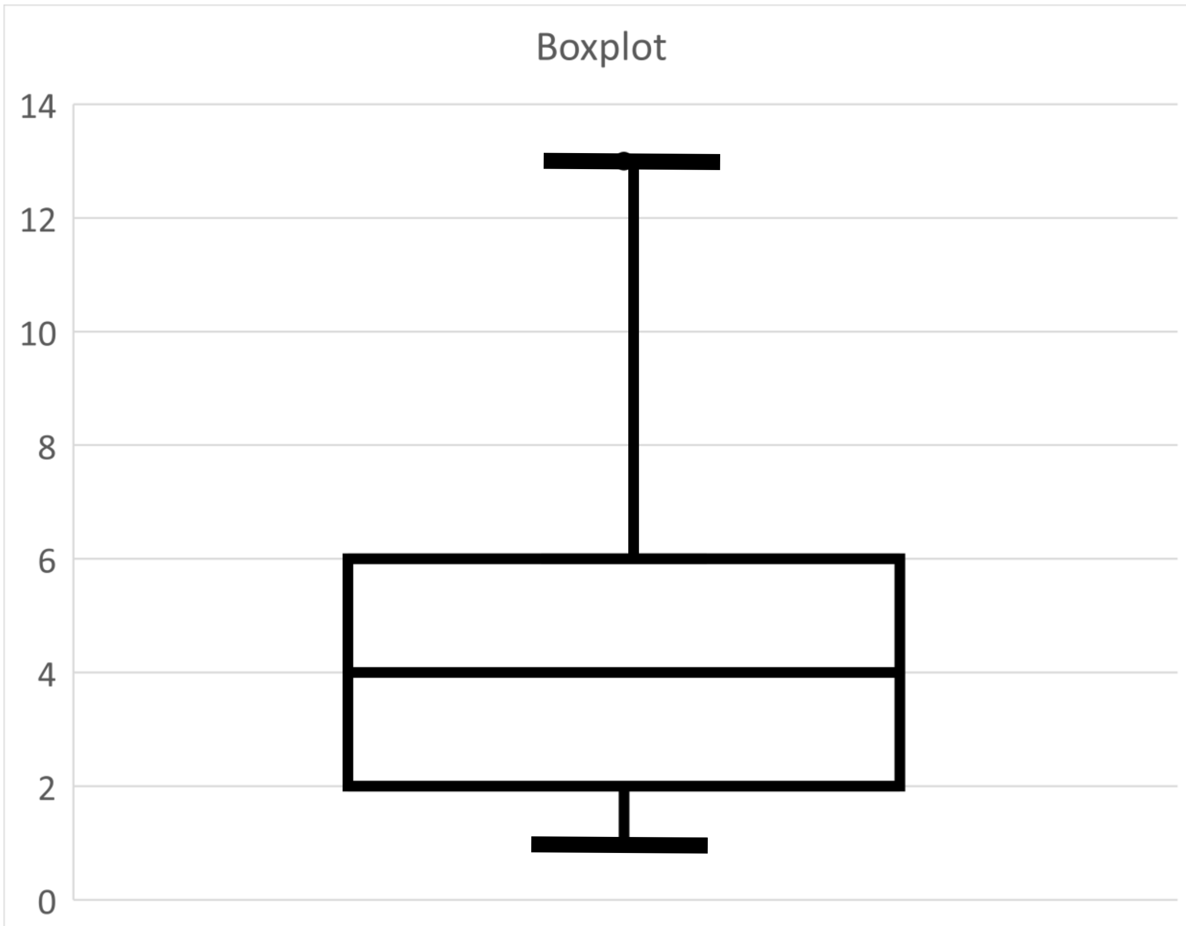
IQR = 9 - 2.5 = 6.5

1.5IQR = 9.75, so $Q_1 - 1.5IQR = -7.25$, and therefore there are no outliers below Q_1.
$Q_3 + 1.5IQR = 18.75$, and therefore 24 is the only outlier above Q_3.



A basic boxplot allows for the visual determination of the presence of outliers. Does the data set represented by the following basic boxplot contain any outliers?

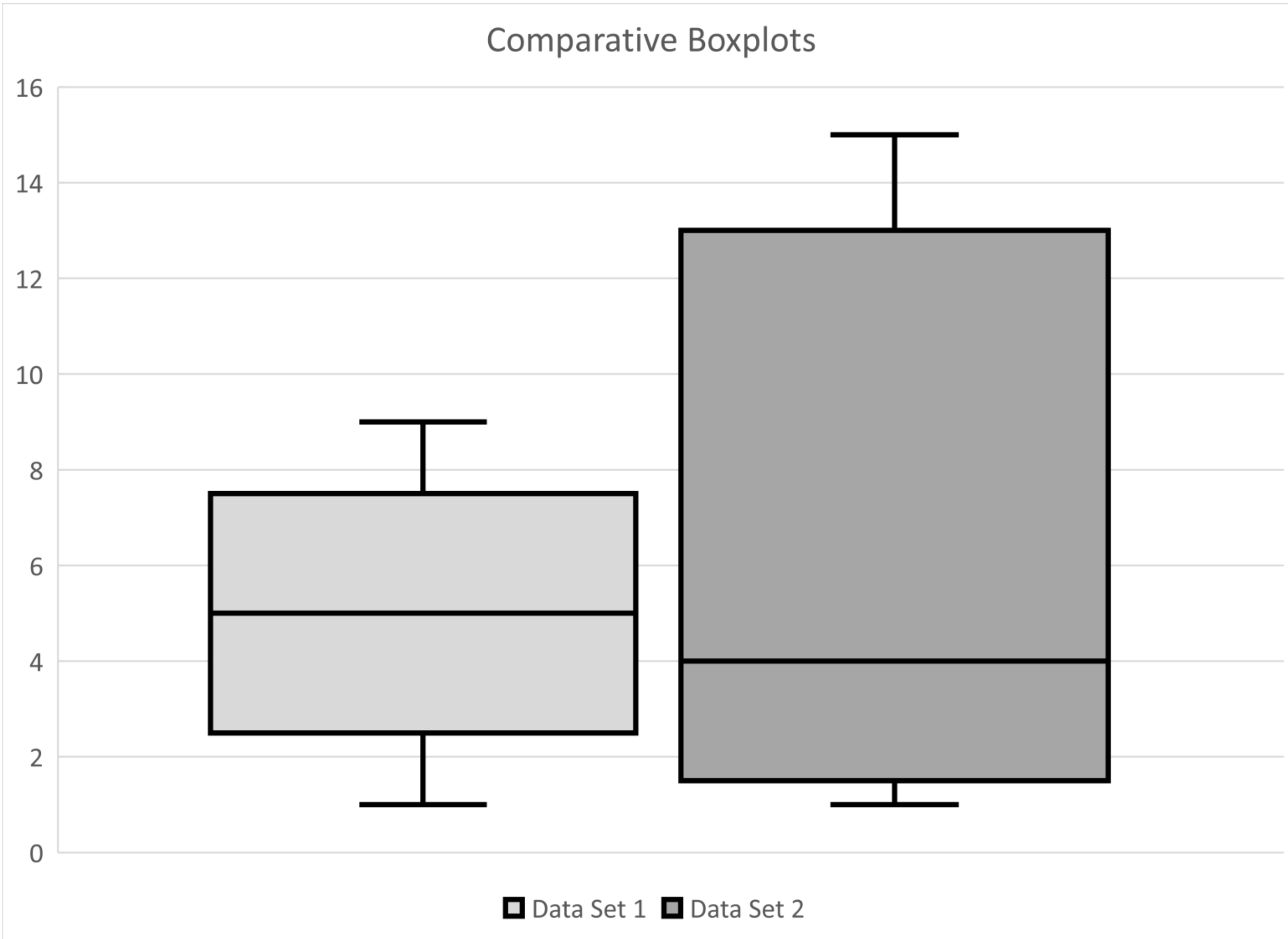
The IQR is 4, so 1.5IQR is 6. The whisker above Q_3 is definitely longer than 6, so there is at least one outlier in the upper whisker. The lower whisker is shorter than 6, so there are no outliers there.



Sometimes boxplots are drawn with a vertical scale.

<u>Comparative Boxplots:</u>

Used to compare two or more data sets using boxplots.



Which data set has the larger *IQR*?

Data Set 2

Which data set has the larger median?

Data Set 1

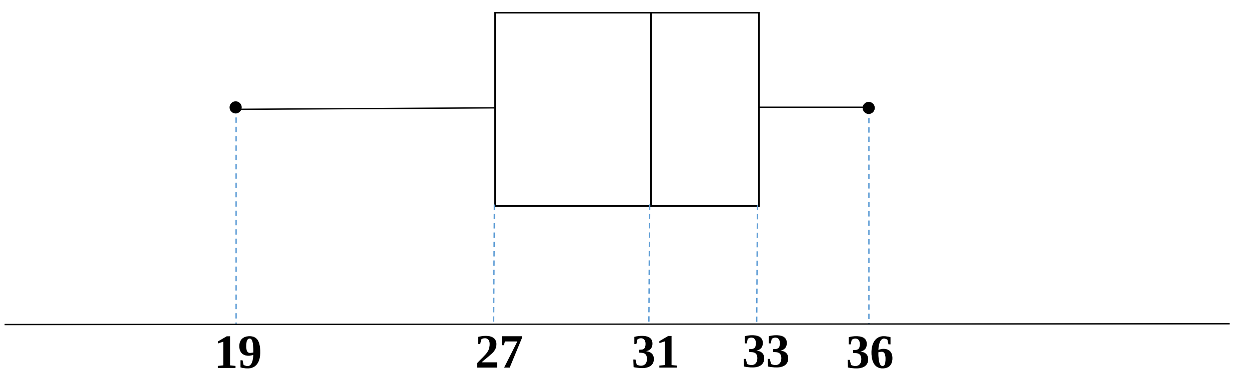
Which data set has the larger Q_1 value?

Data Set 1

One more Example!

Thirty singers entered a singing competition. Here are the 5-number summary and box-plot of the entrant's ages.

Minimum	Q_1	Q_2	Q_3	Maximum
19	27	31	33	36



1. What's the IQR of the data set?

$33 - 27 = \boxed{6}$

2. Does the data set have any outliers?

The outlier boundaries are 18 and 42, so no.

3. Is the data set more spread out in the ages below 31 or above 31?

Below 31