PART B - LESSON 2 - PART 1: The Effects of Transforming Data on Spread and Center

In this video, we will be talking about the effects of transforming data on spread and Center. Suppose I take a sample of five University students from Winnipeg and we recorded their weights and pounds, what is the mean weight and what is the standard deviation?

You should find out that the mean is equal to 135.6 and you should find the standard deviation is equal to 31.75. Remember that the mean and standard deviation are respectively measures of center and spread. Now let's see what happens to these values when we transform the data.

Suppose that during the winter, each person wears an extra five pounds of clothes, under these new conditions what is the new mean and standard deviation? To calculate the new mean we can add 5 to each individual value, then we could calculate the new mean from this data.

You should find that new X bar is equal to 140.6. Another way we could have calculated this is by adding 5 to the original mean, this also gives us an answer of 140.6. Conceptually, let me show you why we could do this. Here I have plotted the data on the number line and here's the original mean, by adding 5 to each data point, we shift the entire distribution towards the right, the new distribution looks like the old distribution except it is shifted to the right by 5.

Remember that the mean is the balance point of a data set, so because the new distribution looks a lot like the old distribution, the new mean must be in the same kind of the position, so in fact adding 5 pounds to each data actually adds 5 pounds to the mean as well and this gives us a new X bar of 140.6.

Now let's see what happens when we try to determine the new standard deviation, instead of using the formula to calculate the new standard deviation, we can also think about this conceptually. The standard deviation is a measure of spread for the data set, by adding 5 pounds to each data point, we see that the spread for the new data set hasn't even changed, so the spread for both distributions are the same and as a result the new standard deviation is equal to the original standard deviation.

When we are dealing with transforming data there are some guidelines that we can follow. We see that measures of center are affected by addition, subtraction and multiplication and division. So measures of center are basically affected by every mathematical operation.

If you have forgotten, measures of center include the mode, median and mean, for measures of spread we saw that they aren't affected by additive terms which indirectly means they are also not affected by subtractive terms. However, you will see that measures of spread are affected by multiplication and division.

The measures of spread we have talked about include things like range and standard deviation. To demonstrate these guidelines, let's use the same example we worked with, suppose that in order to stay hydrated, the students drink 2.5 milliliters of water

for every pound they weigh plus 750 milliliters of water a day. What is the new mean and standard deviation for the amount of water consumed every day.

We had previously calculated that the mean and standard deviation for the weights of the people in this data set to be equal to 135.6 pounds and 31.75 pounds. To use the guidelines that we just talked about, we need to analyze the question, it says that these students drink 2.5 milliliters of water for every pound they weigh, this suggest a multiplicative term and multiplication affects both the mean and standard deviation.

The question also says plus 750 milliliters of water a day, this suggest that we have an additive term and addition only affects measures of center. So to find the new X bar, we will have 135.6 times 2.5 plus 750 which gives us a value of 1,089. To find the new standard deviation, we would have 31.75 times 2.5 which gives us a value of 79.38, so based on these calculations, we can actually extract a formula for the transformation of each type of measure.

Measures of Center can be represented by this formula where the new center is equal to the old center times a multiplicative constant plus an addition constant.

On the other hand, measures of spread can be represented by this formula where the new spread is equal to the old spread times a multiplicative constant.

PART B - LESSON 2 - PART 2: The Effects of Outliers on Spread and Centre

In this video, we will be talking about the effects of outliers on spread and Center. An outlier can be defined as the data value that is numerically distant from the data set. An outlier is a data point that falls outside the main patterns of data points and it can be the largest value in a given data set or it can be the smallest value in a given data set. We will go through a couple of examples, so you can see what you mean by this, so if I had a histogram that looks like this, we can see that this point is numerically distant from the data set because of this, this data value can be classified as an outlier.

Now in this data set, we see that the number 9,000 is significantly larger than all of the other data points, so this data value can be classified as an outlier, and in this data set, the outlier is the number 3 because it is significantly smaller than all the other data points, in other words, it is numerically distant from the entire data set. Sometimes this

outlier and a data set may not be obvious. In another video, we will show you how you can calculate outliers.

Outliers can be thought of as data points that a very atypical and surprising because outliers are numerically distant from a data set they can affect measures of center and spread. Suppose a researcher decided to record the temperature of Winnipeg on July 1st for 7 years straight and got theses results, we can clearly see that negative 350 degrees Celsius is an outlier because it is not a typical observation especially during the summer if we use this data to calculate the mean, we get negative 28 degree Celsius.

Obviously, we know that the typical temperature around this time is very warm and around positive 20 to 30 not negative 28. We got this result because of the outlier, the outlier was involved in our calculations, which gave us a skewed result, therefore we see that the mean is affected by the presence of outliers. To show how outliers affect measures of center and spread, I will compare their calculations with the outlier, and calculations without the outlier, so with the outlier we get a mean of negative 28 and with the outlier excluded from the data set, you should find that we get a mean of 25.667.

Now let's see what happens to the median. First, we'll have to numerically order the data, we can clearly see that 26 is in the middle of the data set, so the median is equal to 26, without the outlier you should find that we get a median 28.25. Now the mode refers to the most frequently appearing data value.

In this data set, the number 31 appears the most, so the mode is equal to 31 and without the outlier the mode is still equal to 31. Let's look at the range, the range is equal to the maximum minus the minimum, with the outlier, you should find that the range is equal to 381 and without the outlier, you should find that the range is equal to 16.

Now let's go over how each of these measures of center and spread respond to outliers. We had previously discussed that the mean was affected by the presence of outliers. When we included the outlier in our calculations, we saw that it really skewed the results. In contrast, we say that the median and the mode are resistant to the presence of an outlier because the presence of an outlier doesn't change their values as much as the mean does. The median only cares about the middle of a data set and the mode only cares about how frequent our data value appears.

Now let's look at the range. We see that if an outlier is present, it can change the value of the range very drastically, this is because an outlier can either be the maximum value of a data set or the minimum of a data set, and the range is always equal to the maximum minus the minimum, so the outlier will always be involved in the calculations, and a result it really affects the value of the range just like how it affects the value of the mean.

Lastly, we will look at the standard deviation. Since the mean is contained in the formula for the standard deviation, and since the mean is affected by outliers, by default the standard deviation is also affected by the presence of outliers.

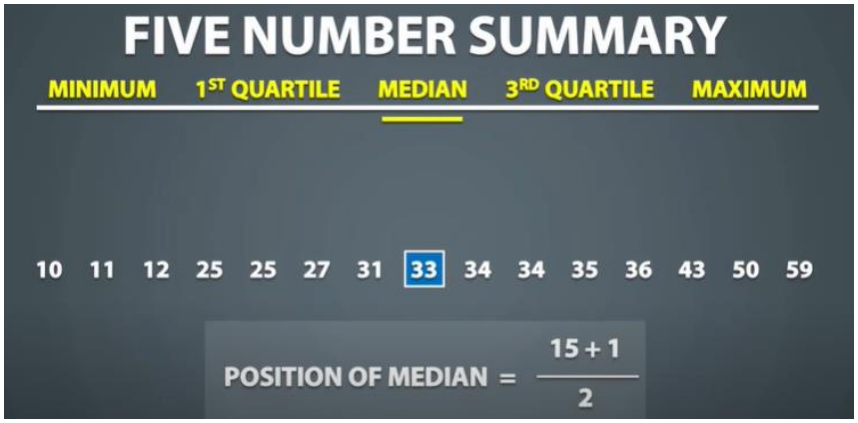
Part B Lesson 2 Part 3: The Five Number Summary, Boxplots, and Outliers

In this video, we will be looking at the five number summary, box plots and outliers.

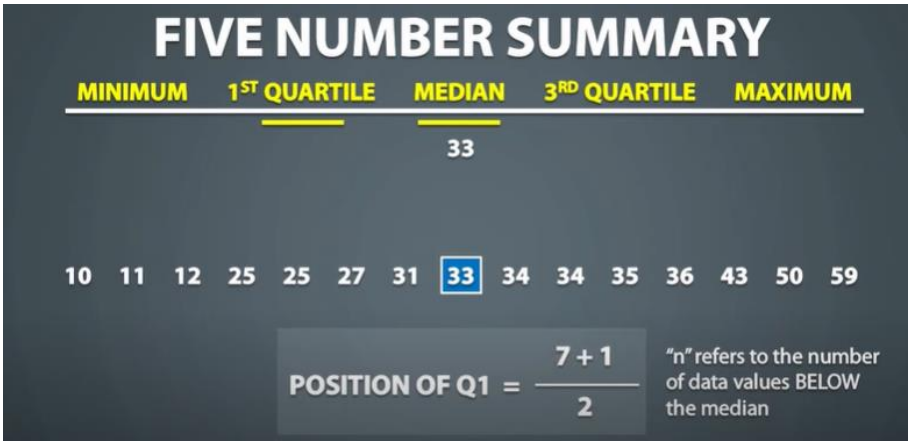
The five number summary gives us a way to describe a distribution using only five numbers, these five numbers include the minimum, first quartile, median, third quartile, and the maximum. So if we took a sample and measured some random quantitative variable, we could order these values from smallest to largest and use the five number summary to describe the distribution.

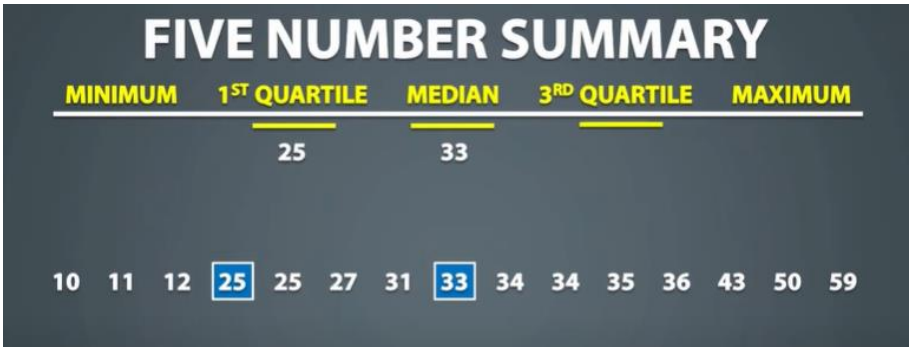
The minimum is the smallest value in a data set and the maximum is the largest value in a data set. The median is the middle data value, it is a point at which 50% of the data values are below the median and 50% of the data values are larger than the median. Now the median of the bottom half is called the first quartile, it is a position where 25% of the data values are below it and 75% of the data values are larger than it. The first quartile is essentially the median of the median, the same thing can be said for the third quartile.

The median of the top half gives us the third quartile and it is a position where 75% of the data values are below it and 25% of the data values are larger than it. The five number summary also gives us a way to divide the data into four equal quarters, so let's determined the five number summary for the following data set. We'll start with the median. To find the median, you can look for it visually and you should find that the median is equal to 33. You can also use the formula to find the position of the median, and we find that it is in the eighth position which is equal to 33.

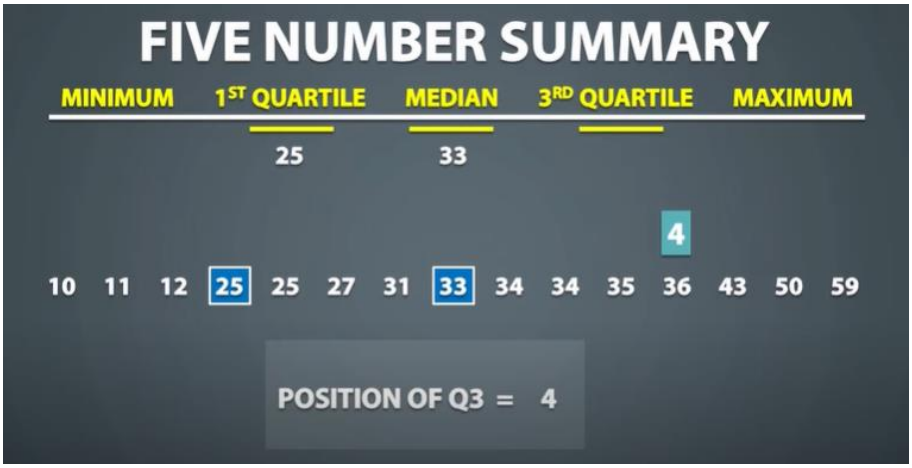


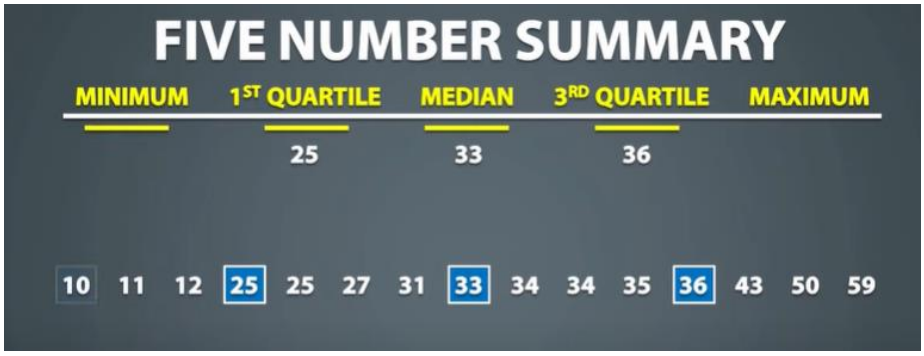
Now to find the first quartile, we can use the same formula to find the position of q1 except this time "n" refers to the number of data values below the median, there are seven data points below the median, so "n" is equal to 7 and we find that the first quartile is on the fourth position, so we count to the fourth position and we see that q1 is equal to 25.

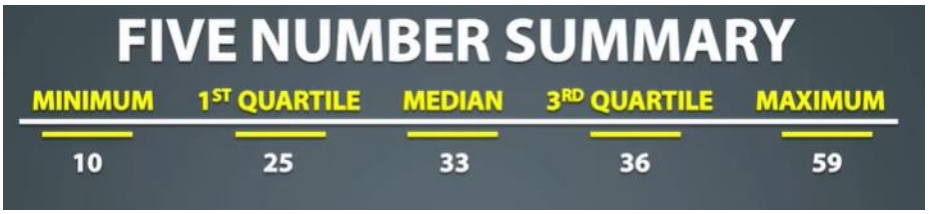




To find the 3rd quartile, we will do the same sort of thing except "n" refers to the number of data values that are above the median, there will always be symmetry, so you should find that q3 is also in position four, so we count four positions above the median and we find that q3 is equal to 36.

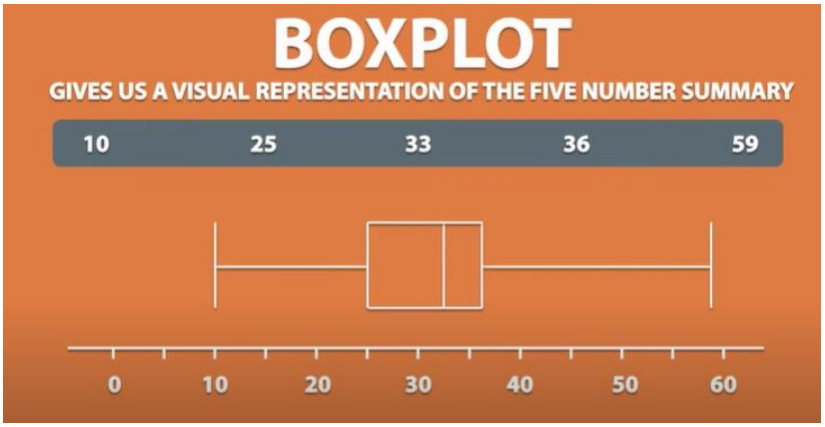




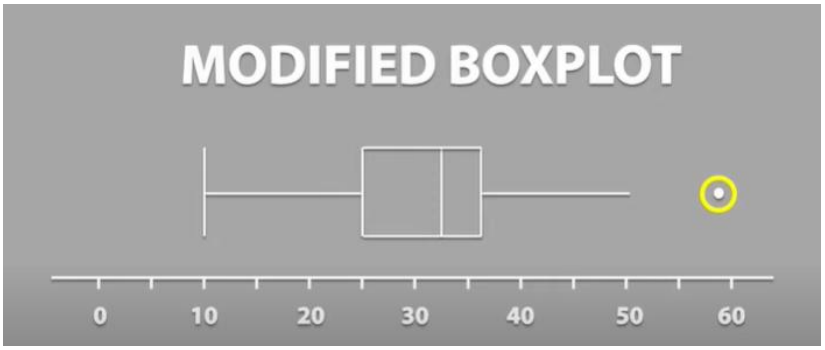


Now the minimum is the smallest number and the maximum is the largest number, so as a result this is our five number summary. We can then take these five numbers and make something called a box plot.

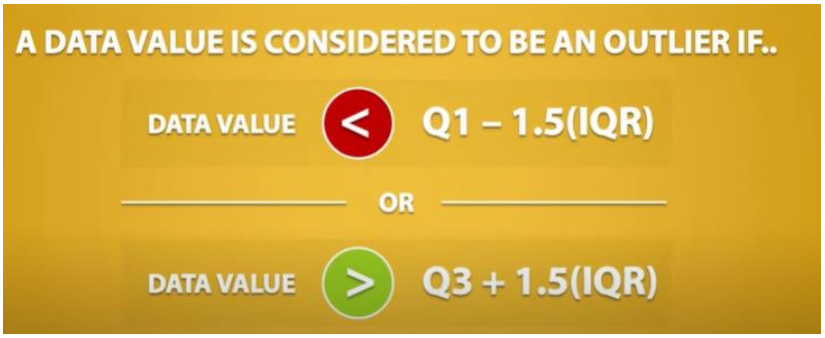
A box plot give us a visual representation of the five number summary, and it looks something like this. Each vertical line on the box plot represents a number from the five number summary. The horizontal line that extends out from the box are called whiskers and the actual box itself is called the interquartile range. The interquartile range refers to the middle 50% of an ordered set and it is equal to the third quartile minus the first quartile.



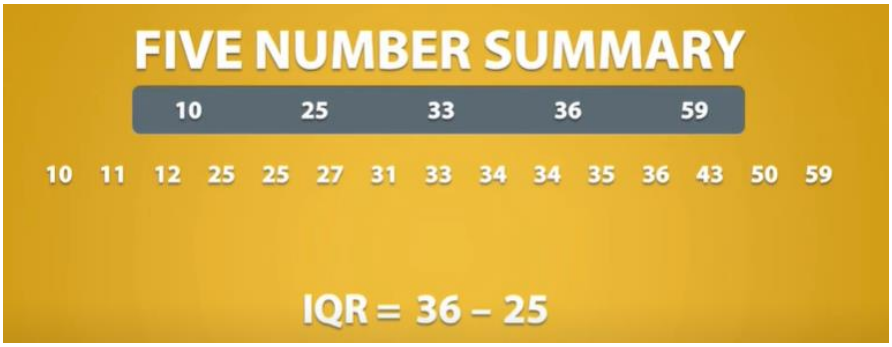
We can also have something called a modified box plot. It's like a regular box plot except that it accounts for outliers.



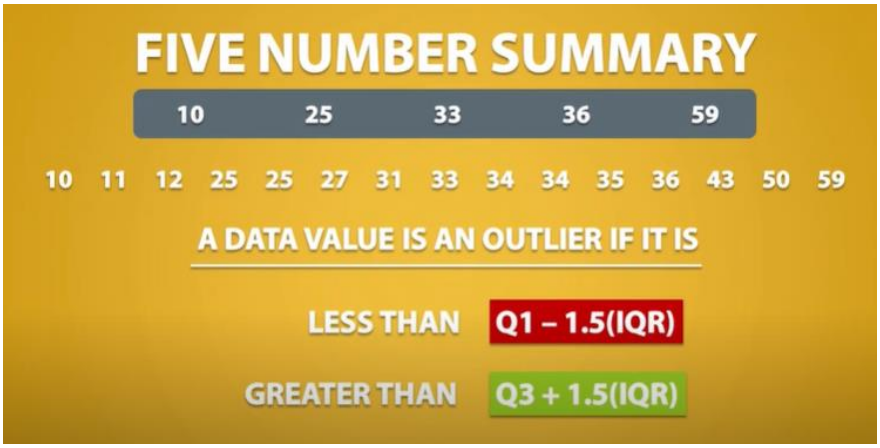
Sometimes outliers in a data set aren't that obvious. However, we can mathematically check if a data set has outliers in it. We say that a data value is considered to be an outlier if the data value is less than than Q1 minus 1.5 times the IQR or if data value is greater than Q3 plus 1.5 times the IQR.



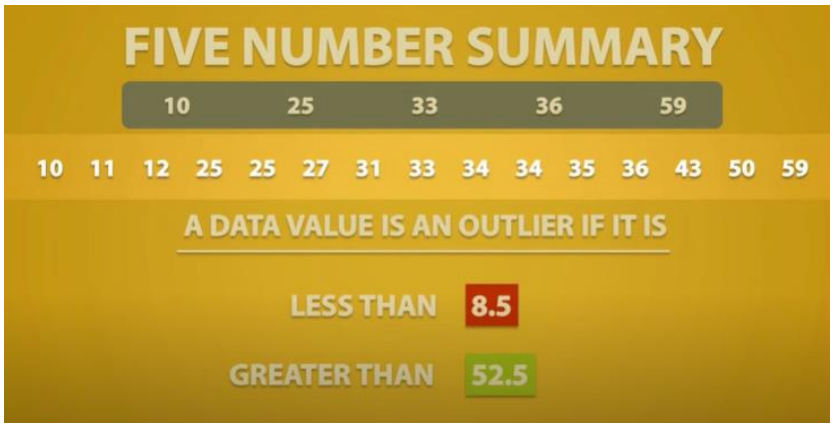
So if you remember the following data set, we had calculated the five number summary to be 10, 25, 33, 36 and 59. To check for outliers, we first need to calculate the interquartile range, we found that Q3 is equal to 36, and we found that Q1 is equal to 25. When we simplify this, we get an answer of 11.



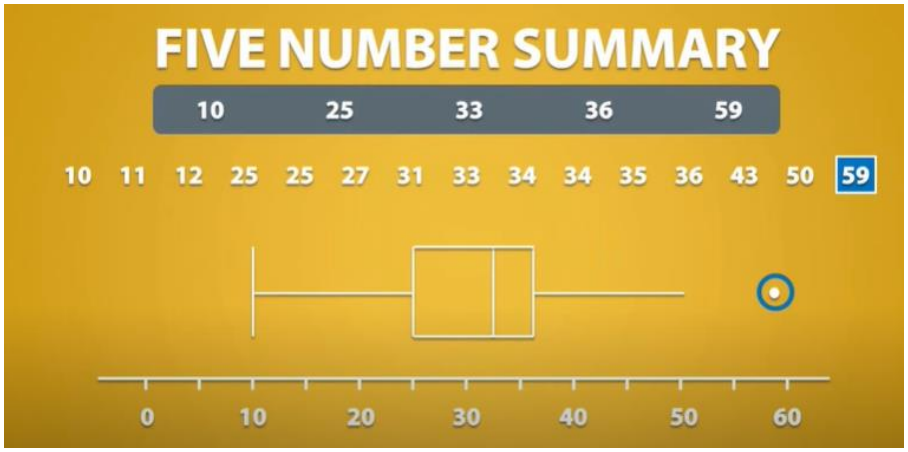
Now we said that a data value is an outlier if is it less than Q1 minus 1.5 times the IQR or if it is greater than Q3 plus 1.5 times the IQR.



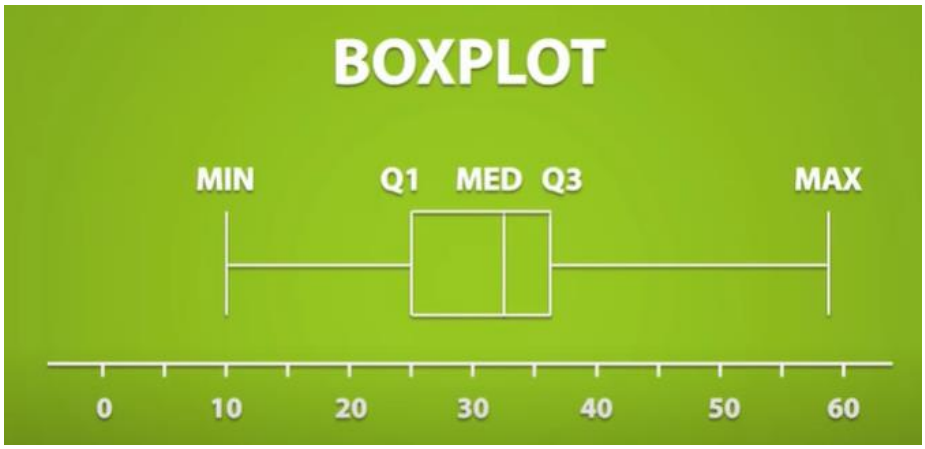
At this point we can start substituting values, Q1 is 25, Q3 is 36 and the IQR is 11 and so we say that a data value is considered to be an outlier if it is less than 8.5 or if it is greater than 52.5.



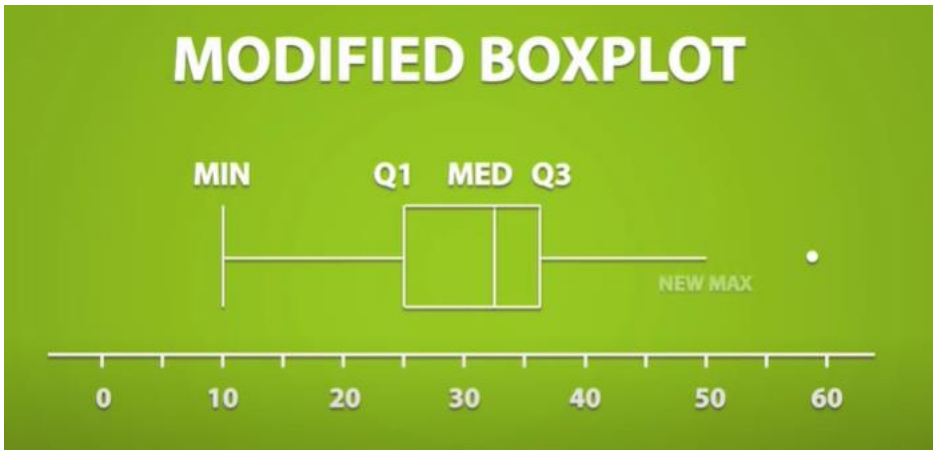
If we look at our data set, we see that no values are less than 8.5. However, we do have a value that is greater than 52.5, therefore we see that 59 is an outlier and so when we make a modified box plot, we write the outlier as a dot and the whisker will only extend to the new maximum. In this case, it is 50.



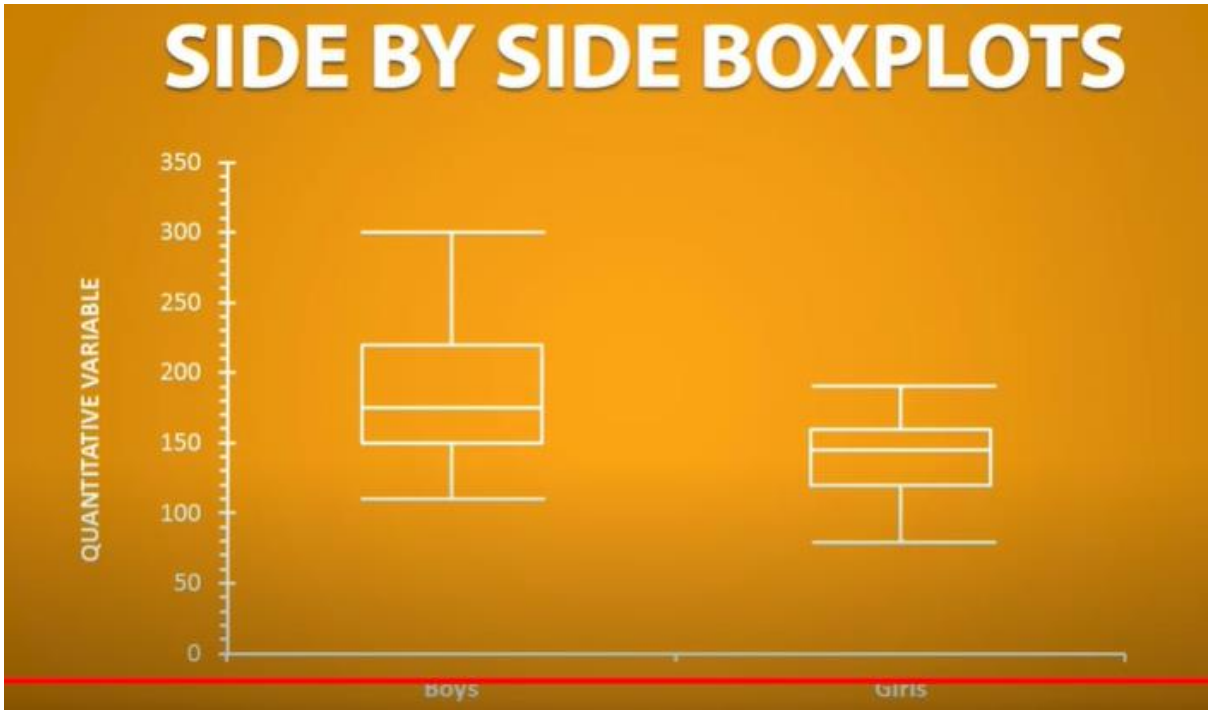
So to quickly recap, a regular box plot is drawn using the five number summary.



A modified box plot also uses the five number summary but it accounts for outliers, and if there are outliers, a whisker or both whiskers will extend only to the new minimum or new maximum.



Similar to back-to-back stem plots, we can have side by side box plots, by having them side by side, we can make easy mathematical and visual comparisons between two sets of data.


SIDE BY SIDE BOXPLOTS
QUANTITATIVE VARIABLE
350
300
250
200
150
100
50
0
Boys
Girls