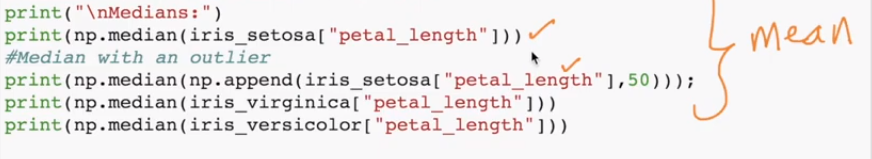
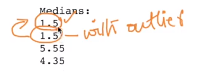
**Median:**

Unlike Mean, the median **don’t get corrupted too much with presence of outliers.** The value of medium is almost similar to mean, below image shows median without outlier and with outlier.

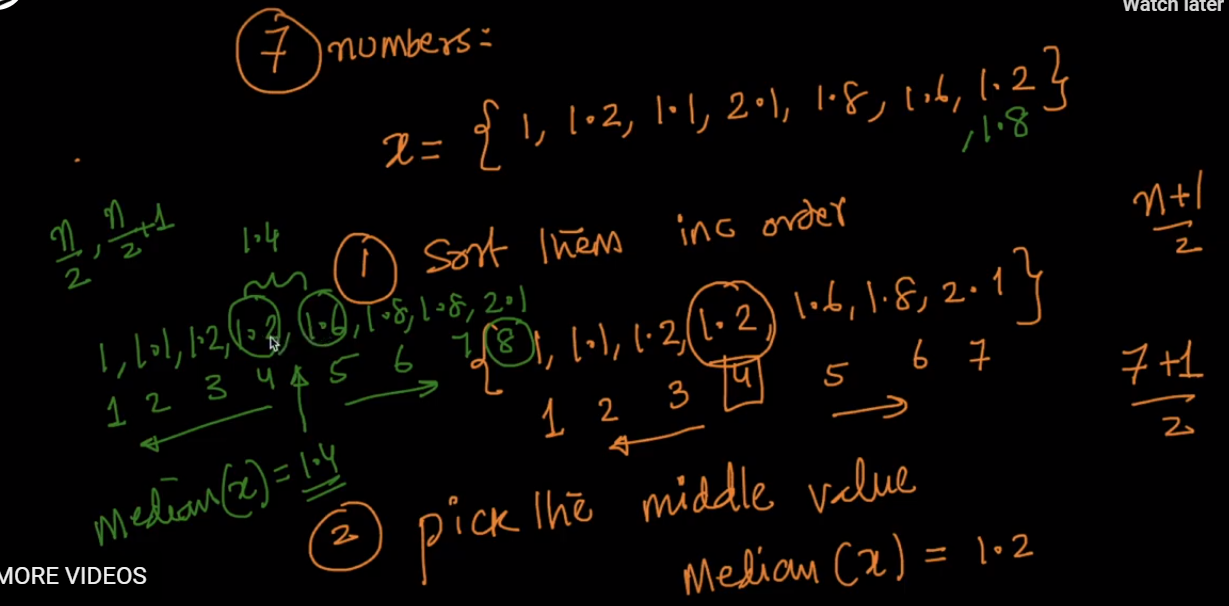
As we can see both are generating same result/median as 1.5 for setossa, that means outliers don’t have much effect on Median.





**How to calculate Median:**

Below fig shows how to calculate median for even no of data set and odd no of data set.



For odd

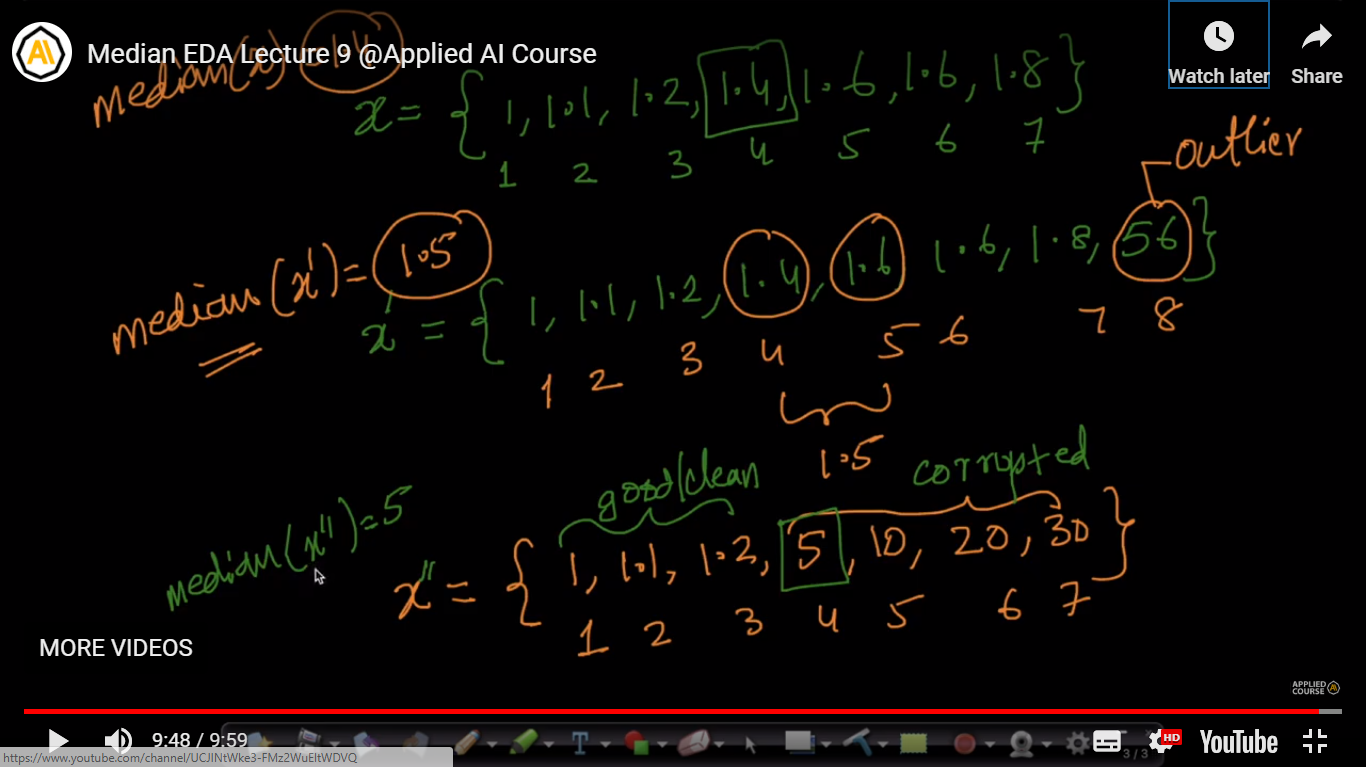
For even

**Why do Median don’t get corrupted with presence of outliers**

Since in median we are giving the value present in middle of the data set, in below fig we can see median(x’) with one outlier, it’s giving median which is not much differ from original median.

Now if the counts of outliers/corrupted become more than the 50% of original data set then median will also get corrupted, ex in below fig we can see median(x’’) contains 3 data set and 4 outliers/corrupted data now in this case we get median as 5 which is corrupted median.

**So we can say if corrupted data is more than 50% of good/clean data, only then median gets corrupted.**



**Some comments:**

1. Highly “skewed” data will have differing means and medians.

Skewed data is data where most of the scores are clustered near one extreme and relatively few near the other. Since means are more sensitive to extreme values than medians, the mean will be closer to the more populated extreme than the median.

1. 1. The mean is the more commonly used measure of the two. Moreover, it is the basis of many advanced statistical methods. For example, the mean is needed to calculate the standard deviation, which is the most prominent measure to assess the variability in a set of data. For a lot of analysis, the mean is very useful. Indeed, if you’re trying to understand data that falls under a normal curve, the mean can tell you a lot of information, because it gives you an overall average score for the group. And it is also needed for many statistical testing procedures, e.g. for the t-test.   
     
   2. The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value so median would be a better measure of central tendency in this situation.  
     
   Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values