Effect of an outlier NAME:

We will investigate how an outlier can affect the mean, standard deviation, median, and quartiles of a set of data. The data we will use is shown below. You may assume it is sample (*not* population) data.

42	50	55	60	60	65	67	70	125

1. Use a calculator to find the mean and standard deviation of this set of numbers. Round to two decimal places.

2. Find the median and the quartiles of this set of numbers. Use these to write the Five Number Summary for this set of numbers. There is no need to round.

3. Which of the numbers in this set would be considered an outlier? Why?

Now let’s find the same information except we’ll exclude the outlier from the list. So now we will work with the numbers below.

42	50	55	60	60	65	67	70

4. Use a calculator to find the mean and standard deviation of this truncated set of numbers. (Simply edit the list of numbers in the calculator to delete the value 125, and then re-run the **1-Var Stats** command.) Round to two decimal places.

5. Find the median and the quartiles of this truncated set of numbers. Use these to write the Five Number Summary for this set of numbers. There is no need to round.

It so happens that the mean and standard deviation of a set of numbers are much more affected by outliers than are the median and quartiles. We say that the median and quartiles are **resistant** whereas the mean and standard deviation are *not*. Let's see!

6a. Notice the standard deviation of the first set (including the outlier 125) is much larger than that of the second, truncated set. Explain why this is so using what you know about standard deviation.

6b. Compare the means of the original and truncated sets. What do you notice?

7. Fill in the table below to organize and compare the quartiles and medians of the original and truncated sets. What do you notice? Do they differ as much as the means and standard deviations differ?

	Q_1	**Median**	Q_3
Set 1			
Set 2 (truncated)			

8. Suppose you are compiling data about the sale prices of local houses for an upcoming realty event. Your local area has mostly moderate houses with a few (expensive) mansions on the outskirts of town. If you want to convince people that house prices are high and it might be a good time to sell (using you, of course, as their agent), would you use a mean or a median to figure the center of this data set? Explain why.