# Lesson 4 - Outliers

Outliers are an interesting part of statistics. An outlier is a piece of data which is significantly above or below all the other pieces of data in a set. In general, we don't include these data points since they don't follow the overall trend of the data.

For example, consider a data set that was recording the amount of time it took to run from one end of the gymnasium to the other end.

| **Student 1** | **Student 2** | **Student 3** | **Student 4** | **Student 5** | **Student 6** | **Student 7** | **Student 8** | **Student 9** | **Student 10** |
|---|---|---|---|---|---|---|---|---|---|
| 4.5s | 3.6s | 3.8s | 4.1s | 12.6s | 4.4s | 4.2s | 3.8s | 4.0s | 4.4s |
| **Student 11** | **Student 12** | **Student 13** | **Student 14** | **Student 15** | **Student 16** | **Student 17** | **Student 18** | **Student 19** | **Student 20** |
| 3.9s | 4.2s | 3.8s | 3.5s | 4.0s | 4.1s | 14.2s | 3.9s | 3.4s | 4.6s |

If we calculate the mean for the data set, we get:

$(4.5 + 3.6 + 3.8 + 4.1 + 12.6 + 4.4 + 4.2 + 3.8 + 4.0 + 4.4 + 3.9 + 4.2 + 3.8 + 3.5 + 4.0 + 4.1 + 14.2 + 3.9 + 3.4 + 4.6) \div 20$

$= 4.95$ .

Most of the data points took between 3.5 and 4.5 seconds. Which values are well outside of this range? It is a good idea to eliminate these data points so that we get a better idea of what the real mean should be.

| Student 1 | Student 2 | Student 3 | Student 4 | Student 5 | Student 6 | Student 7 | Student 8 | Student 9 | Student 10 |
|---|---|---|---|---|---|---|---|---|---|
| 4.5s | 3.6s | 3.8s | 4.1s | 12.6s | 4.4s | 4.2s | 3.8s | 4.0s | 4.4s |
| **Student 11** | **Student 12** | **Student 13** | **Student 14** | **Student 15** | **Student 16** | **Student 17** | **Student 18** | **Student 19** | **Student 20** |
| 3.9s | 4.2s | 3.8s | 3.5s | 4.0s | 4.1s | 14.2s | 3.9s | 3.4s | 4.6s |

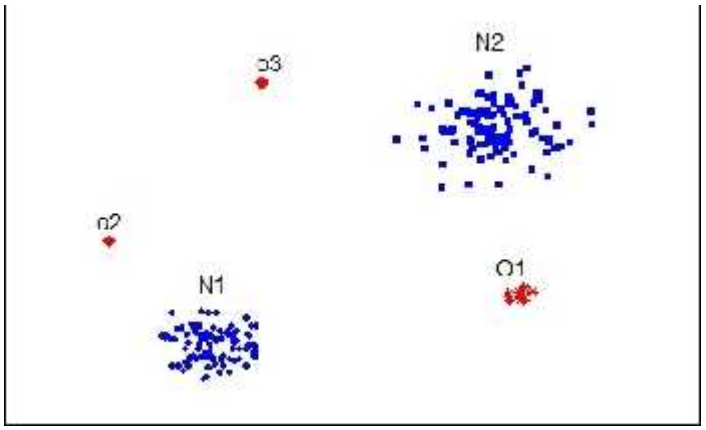
If we eliminate the students outside of the range we get:

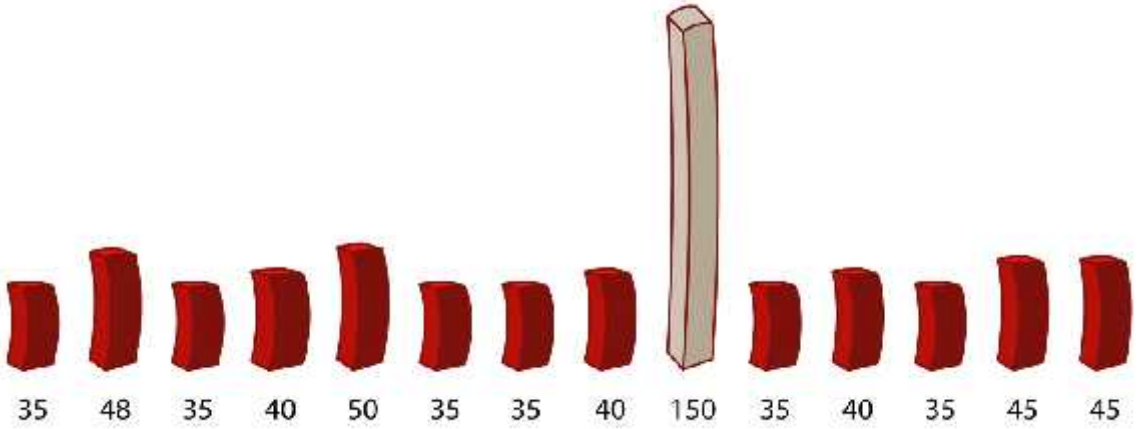
$$(4.5 + 3.6 + 3.8 + 4.1 + 4.4 + 4.2 + 3.8 + 4.0 + 4.4 + 3.9 + 4.2 + 3.8 + 3.5 + 4.0 + 4.1 + 3.9 + 3.4 + 4.6) \div 18$$

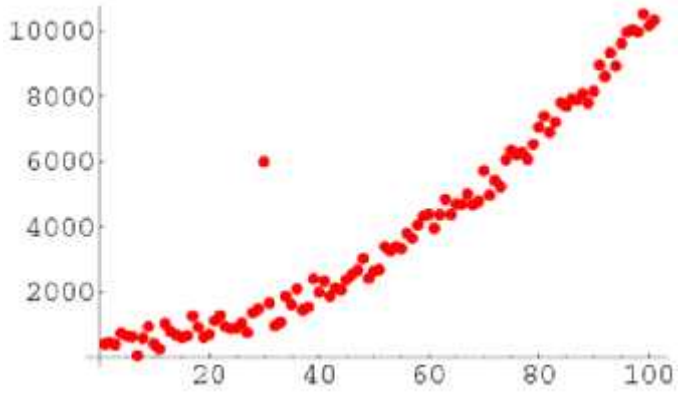
$$= 4.01.$$

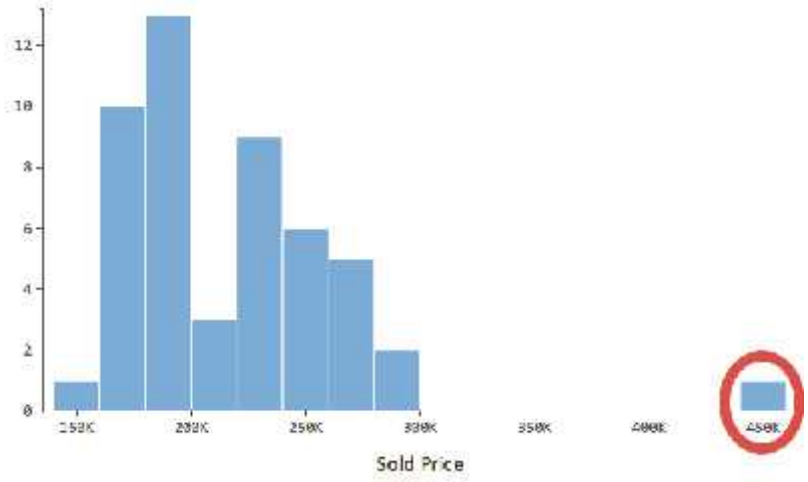
The reason why we eliminate outliers is to try and normalize the data so that it more truly resembles the average that we are seeing. In this case we saw a difference of almost 1 second.

## Outliers on Graphs

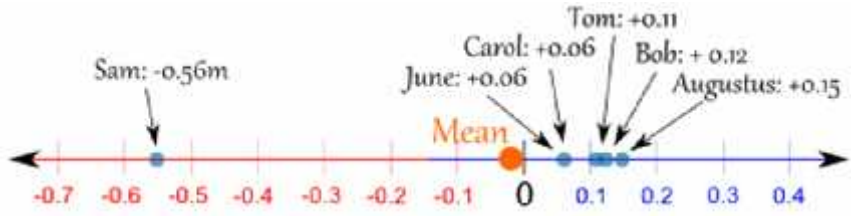


A new coach has been working with the Long Jump team this month, and the athletes' performance has changed. The number line below shows the results.




<u>The mean is:</u>

$= (0.15 + 0.11 + 0.06 + 0.06 + 0.12 - 0.56) \div 6$

$= -0.06 \div 6$

$= -0.01$

So, on average the performance went DOWN 0.01m.

<u>Let us try the results WITHOUT Sam:</u>

$= (0.15 + 0.11 + 0.06 + 0.06 + 0.12) \div 5$

$= 0.5 \div 5$

$= 0.1$

Now, on average the performance went UP 0.10m.

But is that fair? Can we just get rid of values we don't like?

Eliminating Outliers

Remember that the point of eliminating an outlier is to try and normalize your data. We can't just take out a value because we don't like it, we actually need to have a reason why it is not a useful piece of data.

**Useful Data Examples**

A worker at a company that manufactures gears for clocks is checking to see if the gears are within specifications. To pass, the gears must be within 0.2mm of their designed measurements. The worker looked at 5 gears which were: +0.004mm, -0.011mm, +0.45mm, +0.005mm, and -0.01mm.

*In this case, ALL of the measurements are important since any error will affect the parts coming out. This may mean that something needs to be fixed.*

The owner of a restaurant is keeping track of how much money they are making over the course of a week. On Monday's the restaurant makes about $2500, Tuesday they make $2700, Wednesday they make $2600, Thursday brings in $2800, Friday brings in $3200, Saturday they make $7800, and Sunday they make $3000.

*The increased amount of money can be explained because it is the weekend and it makes sense that the restaurant would earn more money.*

## Not Helpful Data Examples

Dave plays Basketball for his high school team and usually scores between 22 and 28 points per game. Last week he strained his shoulder baling hay and has only scored 8 points, 10 points, and 6 points over the last three games.

*In this case we shouldn't keep the data from his most recent games since he is injured and it is not a true reflection on how well he normally plays.*

A company is advertising to try and draw in new employees. There are currently 38 people that work for the company. 20 factory workers make $20,000 a year, 15 supervisors make $32,000 a year, and the three managers make $280,000, $450,000, and $600,000. The company has advertised that the average salary is more than $58,000 a year.

*While the amount of money advertised is true, it is also deceptive. The extremely large salaries of the managers is inflating the average income. Without the managers the average salary is only about $25,000.*