**Definition**

**What is an Outlier?**

**While *Outliers*, are attributed to a rare chance and may not necessarily be fully explainable, *Outliers* in data can distort predictions and affect the accuracy, if you don’t detect and handle them.**

**The contentious decision to consider or discard an *outlier* needs to be taken at the time of building the model. *Outliers* can drastically bias/change the fit estimates and predictions. It is left to the best judgement of the analyst to decide whether treating *outliers* is necessary and how to go about it.**

**Treating or altering the *outlier*/extreme values in genuine observations is not a standard operating procedure. If a data point (or points) is excluded from the data analysis, this should be clearly stated on any subsequent report.**

**Outlier Analysis deals with identification and elimination of outliers.**

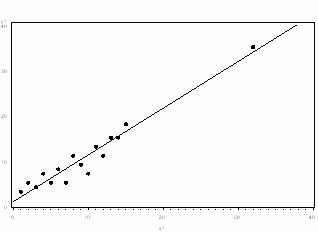
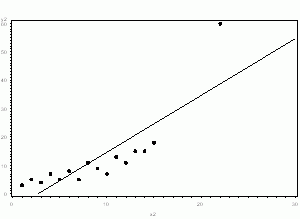
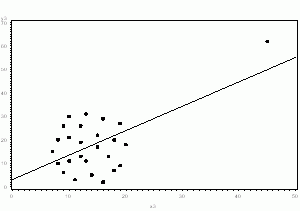
**Data: numerical data are suitable**

**Type of Data and or Technique/Method** (https://www.slideshare.net/dataminingtools/data-mining-outlier-analysis)

Some efficient algorithms for mining distance-based outliers are as follows:

* Cell-based algorithm
* Density-Based Local Outlier Detection
* Deviation-Based Outlier Detection with Sequential Exception Technique
* Distance-based Approaches

When are outliers dropped?

1. If it is obvious that the outlier is due to incorrectly entered or measured data, you should drop the outlier:
2. For example, I once analyzed a data set in which a woman’s weight was recorded as 19 lbs. I knew that was physically impossible. Her true weight was probably 91, 119, or 190 lbs, but since I didn’t know which one, I dropped the outlier.
3. This also applies to a situation in which you know the datum did not accurately measure what you intended. For example, if you are testing people’s reaction times to an event, but you saw that the participant is not paying attention and randomly hitting the response key, you know it is not an accurate measurement.
4. If the outlier does not change the results but does affect assumptions, you may drop the outlier. But note that in a footnote of your paper.
5. Neither the presence nor absence of the outlier in the graph below would change the regression line:
6. 
7. More commonly, the outlier affects both results and assumptions. In this situation, it is *not* legitimate to simply drop the outlier. You may run the analysis both with and without it, but you should state in at least a footnote the dropping of any such data points and how the results changed.
8. 
9. If the outlier *creates* a significant association, you *should* drop the outlier and *should not* report any significance from your analysis.
10. In the following graph, the relationship between X and Y is clearly created by the outlier. Without it, there is no relationship between X and Y, so the regression coefficient does not truly describe the effect of X on Y.
11. 

So in those cases where you shouldn’t drop the outlier, what do you do?

One option is to try a transformation. Square root and log transformations both pull in high numbers. This can make assumptions work better if the outlier is a dependent variable and can reduce the impact of a single point if the outlier is an independent variable.

Another option is to try a different model. This should be done with caution, but it may be that a non-linear model fits better. For example, in example 3, perhaps an exponential curve fits the data with the outlier intact.

**Starters List**

**Coursera:**

[**https://www.coursera.org/lecture/data-analytics-business-capstone/dealing-with-outliers-61QUP**](https://www.coursera.org/lecture/data-analytics-business-capstone/dealing-with-outliers-61QUP)

**Udemy:**

[**https://www.udemy.com/outlier-detection-techniques/**](https://www.udemy.com/outlier-detection-techniques/)