CLEAN DATA

REMOVE OUTLIERS

# 1. What are Outliers?

We all have heard of the idiom ‘odd one out which means something unusual in comparison to the others in a group.

Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations.

That means an outlier is vastly larger or smaller than the remaining values in the set.

# 2. Why do they Occur?

An outlier may occur due to the variability in the data, or due to experimental error/human error.

They may indicate an experimental error or heavy skewness in the data (heavy-tailed distribution).

# 3. What do they affect?

In statistics, we have three measures of central tendency namely Mean, Median, and Mode as help us describe the data.

Mean is the accurate measure to describe the data when we do not have any outliers present.

Median is used if there is an outlier in the dataset.

Mode is used if there is an outlier AND about ½ or more of the data is the same.

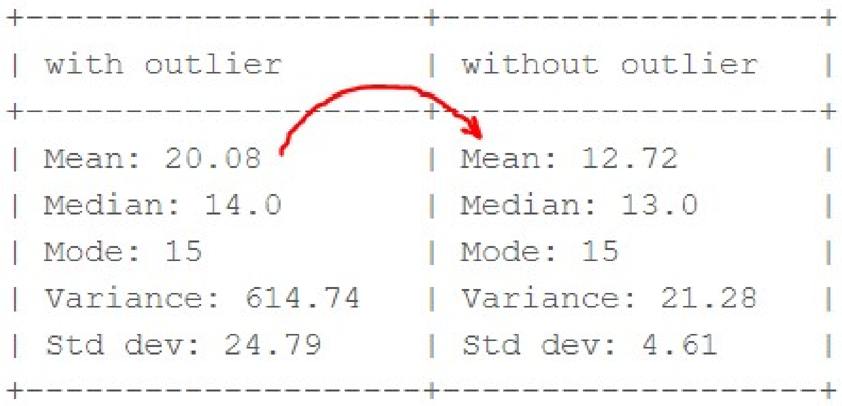
‘Mean’ is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.

## Example 1:

Consider a small dataset,

sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]

By looking at it, one can quickly say ‘101’ is an outlier that is much larger than the other values.



```
+--------------------+-------------------+
| with outlier       | without outlier   |
+--------------------+-------------------+
| Mean: 20.08        | Mean: 12.72       |
| Median: 14.0       | Median: 13.0      |
| Mode: 15           | Mode: 15          |
| Variance: 614.74   | Variance: 21.28   |
| Std dev: 24.79     | Std dev: 4.61     |
+--------------------+-------------------+
```

*From the above calculations, we can clearly say the Mean is more affected than the Median.*

## 4. Detecting Outliers

If our dataset is small, we can detect the outlier by just looking at the dataset.

But what if we have a huge dataset, how do we identify the outliers then?
We need to use visualization and mathematical techniques.

Below are some of the techniques of detecting outliers

- Boxplots
- Z-score
- Inter Quantile Range(IQR)