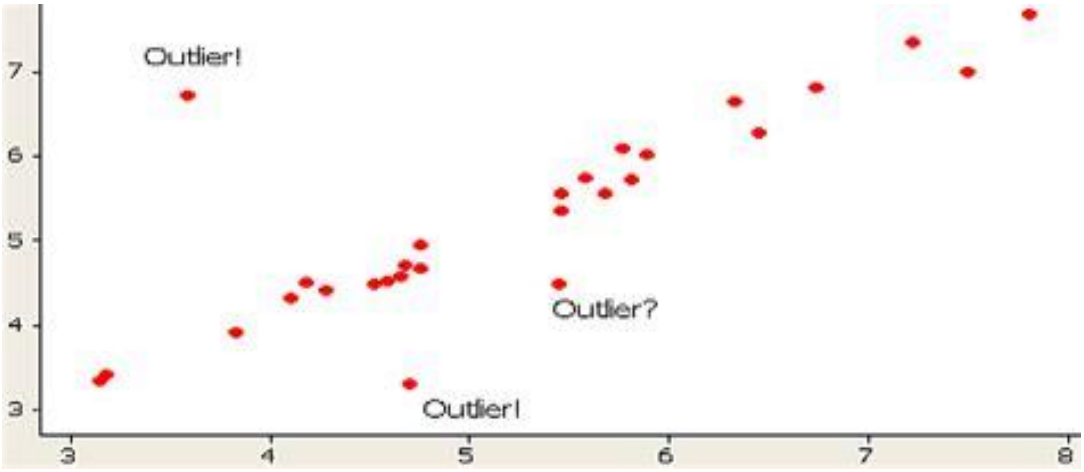
Outlier in the dataset

It is defined as the thing or object that is different from the rest of things.
Another definition is that the outliers are observations in the dataset that deviates significantly from the rest of the data.

In any data science project detecting and removing the outliers are very important, as they can have a significant impact on the any statistical technique such as mean, median, standard deviation and ml models. Outliers sometimes indicate errors or anomalies in the dataset.

The visual representation of the outlier in the form of the graphs is as follows:



It is very important to detect the outlier in the data because it has an impact on the result of the dataset, let take an example of the heights of the students of class 12. The heights of the students are as follows:

Heights
5
5.2
5.5
5.3
5.6
8.3
5.7
5.5

Now in the above data we notice that the height that is 8.3 is at very extreme and is considered as outlier. Now we find the meaning of the data with and without the outlier values that is.

With the outlier values

Mean = $\frac{5+5.2+5.5+5.3+5.6+8.3+5.7+5.5}{8}$

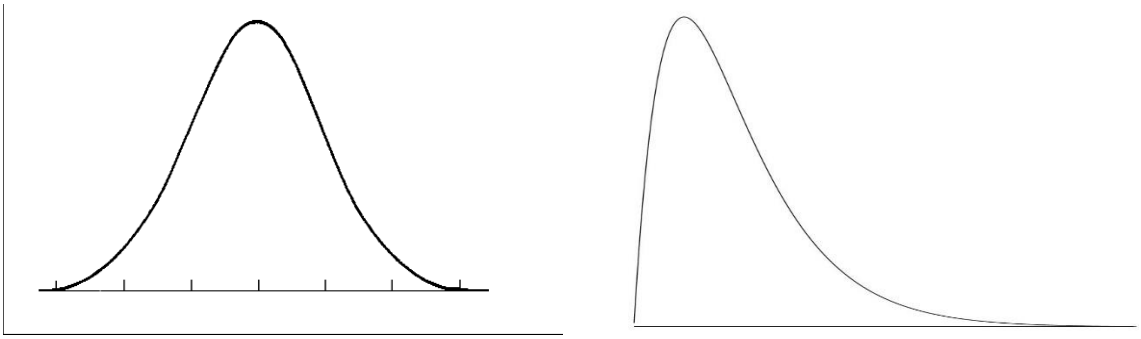
Mean = 5.76.

Without the outlier value

Mean = $\frac{5+5.2+5.5+5.3+5.6+5.7+5.5}{7}$

Mean = 5.40.

Both the results show significant differences. With this we say that the outlier can make the normal data skewed.



One of the curves is normal and one is skewed.

Types of the outliers in the dataset that are as follows:

1. Univariate outliers
2. Multivariate outliers
3. Global outliers
4. Contextual outliers
5. Collective outliers

- **Univariate:** Single data points that deviate significantly from the overall distribution of a single variable. (e.g., Income in a salary survey having one entry much higher than all others).
- **Multivariate:** Data points that appear normal for each variable individually but seem unusual when considered together across multiple variables. (e.g., A customer profile with very high income and low age might be odd).

- **Global:** Data points that stand out from the entire dataset regardless of any specific context. (e.g., Extremely high sales figure on a national holiday).
- **Contextual:** Data points that are outliers within a specific group or context within the data. (e.g., A sudden drop in website traffic during peak hours on a specific day).
- **Collective:** A group of data points that collectively deviate from the overall pattern in the dataset. (e.g., A cluster of customer profiles with similar unusual characteristics).

How to identify outlier and how to remove them?

Visual identification

1. Box plot.
2. Histogram.
3. Z score method.

- **Remove them** (select specific specific range of values for the columns)
- **Transform them** (transform the range of values log transformation and many more)
- **Impute them** (replace the missing values of the column with the mean, median and mode of the dataset)
- **Ml models** (use the type of the ml model that are robust outlier does not affect them)