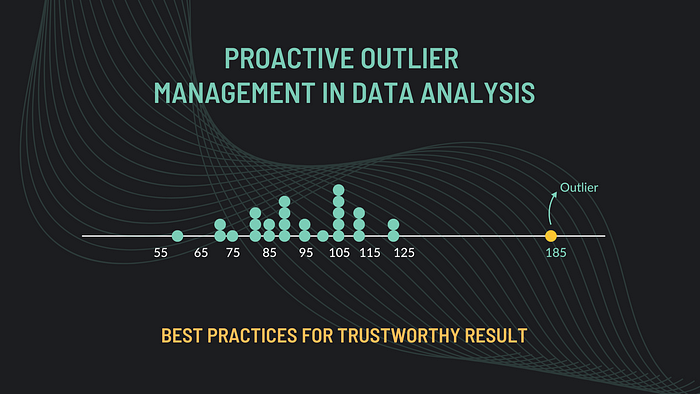
**Outliers** are data points that significantly differ from the majority of the data. They can be unusually high or low compared to the rest of the data.

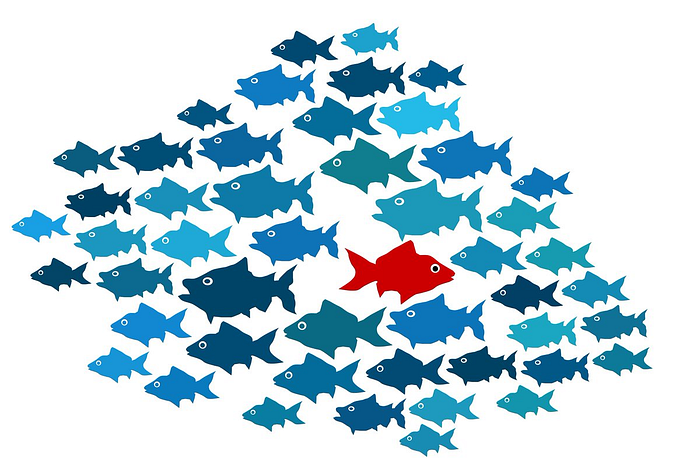
**Why handle?**

 **Distortion of Analysis**: Outliers can skew statistical analyses and affect the mean, variance, and other statistical measures.

 **Model Performance**: In machine learning, outliers can negatively impact model performance, leading to biased predictions or overfitting.



**Outliers**

Outliers are data points that differ significantly from most data points in a dataset. They are values outside a particular variable's expected or usual range. Outliers can occur for various reasons, such as measurement errors, data entry errors, sensor malfunctions, or genuinely unusual observations in the underlying data-generating process. Outliers can distort statistical measures, such as mean and standard deviation, leading to inaccurate data summaries and misleading conclusions. 

**Impact of Outliers on Data Analysis and Modeling:**

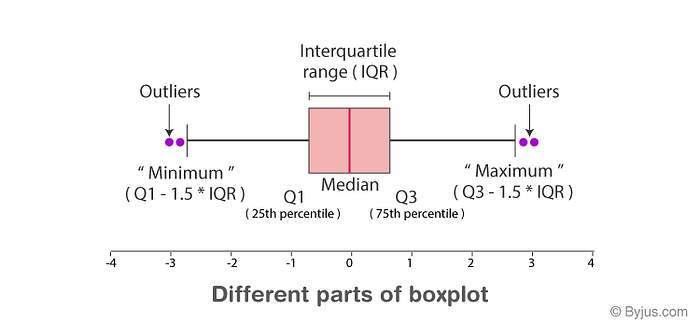
Outliers can have a significant impact on data analysis and modeling results. They can skew statistical measures, leading to inaccurate central tendency and dispersion estimates. For example, suppose a dataset of employee salaries contains a few highly high wages due to executive bonuses. In that case, the mean salary can be significantly inflated, leading to an overestimated average salary for the organization. Similarly, if a dataset of housing prices contains meager prices due to data entry errors, the median price may not accurately represent the typical price of houses in that area.

## Techniques for Detecting Outliers:

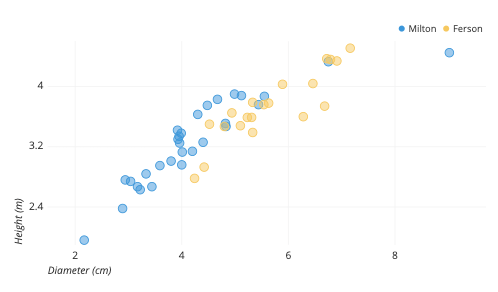
Handling outliers is an essential step in the data preprocessing stage in data science. Several standard techniques can deal with outliers, depending on the specific dataset, the analysis or modeling technique, and the underlying domain knowledge. Let’s explore some of these techniques in detail:

The first step in dealing with outliers is to detect them in the dataset. Several statistical techniques can be used for outlier detection, such as:

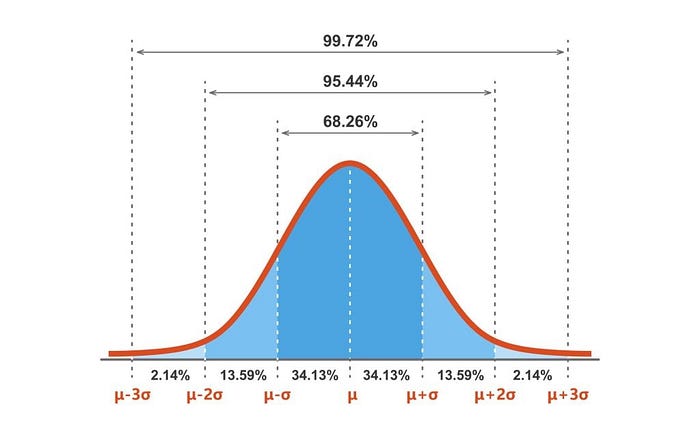
* **Box plots:** Box plots, also known as whisker plots, can visually display the data distribution and identify potential outliers. In a box plot, the box represents the interquartile range (IQR), which contains the middle 50% of the data, and the whiskers represent the data points within 1.5 times the IQR from the upper and lower quartiles. Data points outside this range are considered potential outliers and are plotted as individual points.



* **Scatter plots:** Scatter plots plot two variables against each other and can be used to identify outliers in bivariate datasets. Outliers are data points that deviate significantly from the general pattern of the scatter plot.



* **Z-score:** The Z-score indicates how far a data point is from the mean of the dataset in terms of standard deviations. It can be calculated for each data point, and those with a Z-score above a certain threshold (e.g., 2 or 3) can be flagged as potential outliers.



## Handling Techniques:

* **Imputation:** Imputation involves replacing or filling in the outlier values with estimated or imputed values. This can be done using various methods, such as replacing outliers with the mean, median, or mode of the dataset, or using more advanced techniques, such as k-nearest neighbors imputation or regression imputation, where imputed values are estimated based on the values of neighboring data points or by regressing on other variables in the dataset.
* **Transformation:** Transformation involves applying mathematical transformations to the data to reduce the impact of outliers. For example, taking the data's logarithm, square root, or cube root can compress the data's scale and reduce the influence of extreme values. However, transformation should be done with caution, as it can also affect the interpretation and validity of the results.
* **Winsorizing:** Winsorizing involves capping or truncating the extreme values at a certain threshold. For example, values above a certain percentile (e.g., 95th percentile) can be set to the value of the percentile, and values below a certain percentile (e.g., 5th percentile) can be set to the value of the percentile. This can be done to reduce the impact of extreme values without completely removing them from the dataset.
* **Data partitioning:** Another approach is to partition the data into different subsets based on the presence of outliers. For example, one can create separate subsets of data with and without outliers and analyze or model them separately. This can provide insights into the differences in data patterns or model performance with and without outliers.