**Anomaly / Outlier**

**What is an Outlier?**

An outlier is a data point in a data set that is distant from all other observations.

A data point that lies outside the overall distribution of dataset (95% of customer behaviour / claims / spending nature of customer)

**Many people get confused between Extreme values & Outliers.**

**What is an Extreme Value?**

An Extreme value is just a minimum or a maximum, it need not be much different from of the data.

**What is difference between Extreme value & Outlier?**

An Extreme value is just a minimum or a maximum, it need not be much different from the data & a point that is far away from the other points called as outlier.

**What is the reason for an outlier to exist in dataset?**

- Variability in the data

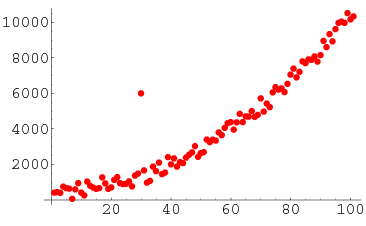
- An Experimental measurement errors

**How can we Identify an outlier?**

- Using Box plots

- Using Scatter plot

- Using Z score



**Why Do We Care About Anomalies?**

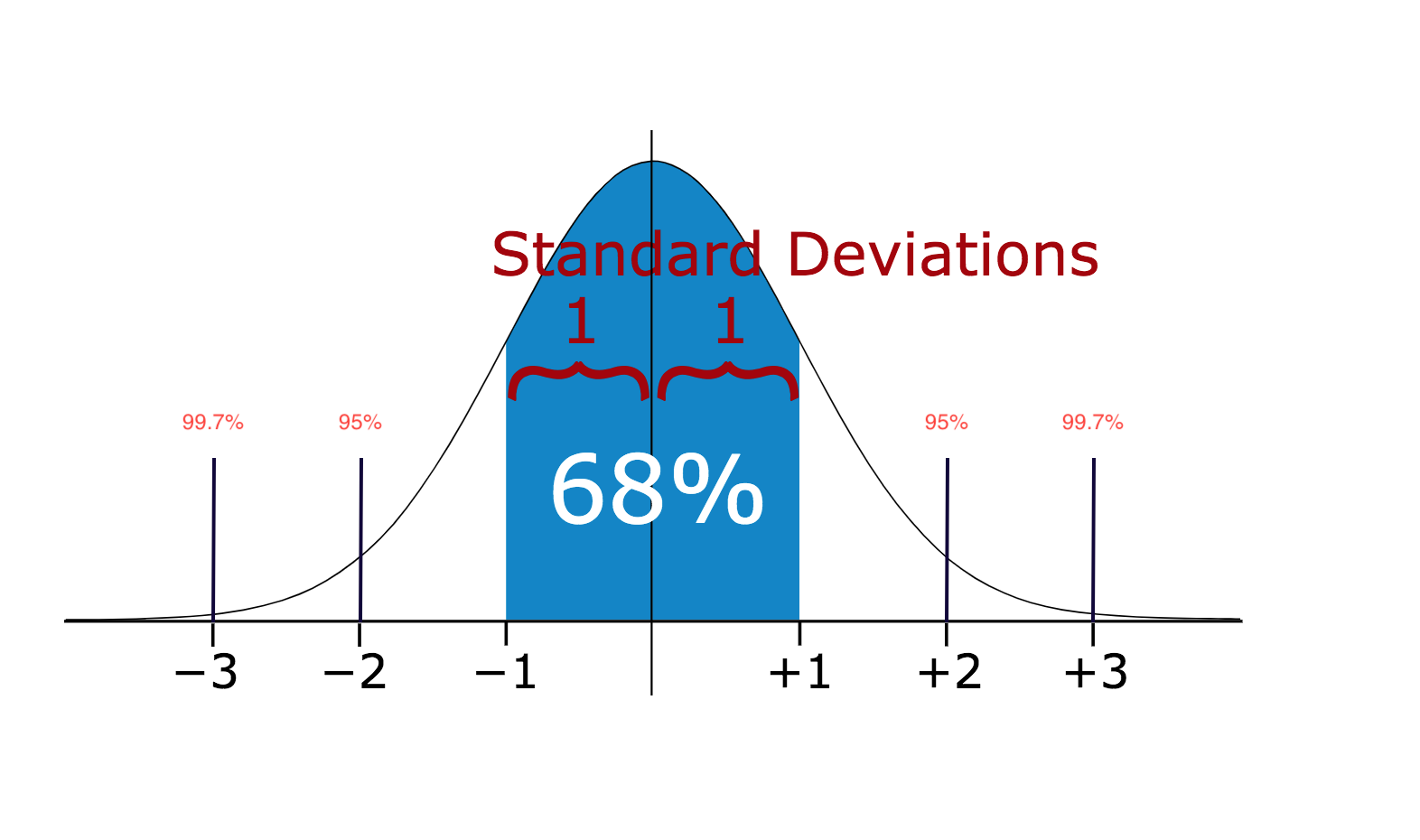
Detecting outliers or anomalies is one of the core problems in data mining. The emerging expansion and continued growth of data and the spread of IoT devices, make us rethink the way we approach anomalies and the use cases that can be built by looking at those anomalies.

We now have smart watches and wristbands that can detect our heartbeats every few minutes. Detecting anomalies in the heartbeat data can help in predicting heart diseases. Anomalies in traffic patterns can help in predicting accidents. It can also be used to identify bottlenecks in network infrastructure and traffic between servers. Hence, the use cases and solution built on top of detecting anomalies are limitless.

Another reason why we need to detect anomalies is that when preparing datasets for machine learning models, it is really important to detect all the outliers and either get rid of them or analyse them to know why you had them there in the first place.

**Method 1 - Standard Deviation:**

In statistics, if a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations



Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are very likely to be anomalous or outliers.

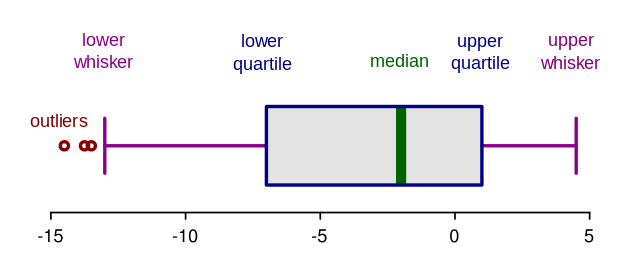
**Z score**

Z score indicates how many standard deviation away a data point

Calculate the **Z score = (X - m)/Sigma**, where m = mean, Sigma = standard deviation

**Method 2 - Boxplots:**

Box plots are a graphical depiction of numerical data through their quantiles. It is a very simple but effective way to visualize outliers. Think about the lower and upper whiskers as the boundaries of the data distribution. Any data points that show above or below the whiskers, can be considered outliers or anomalous.

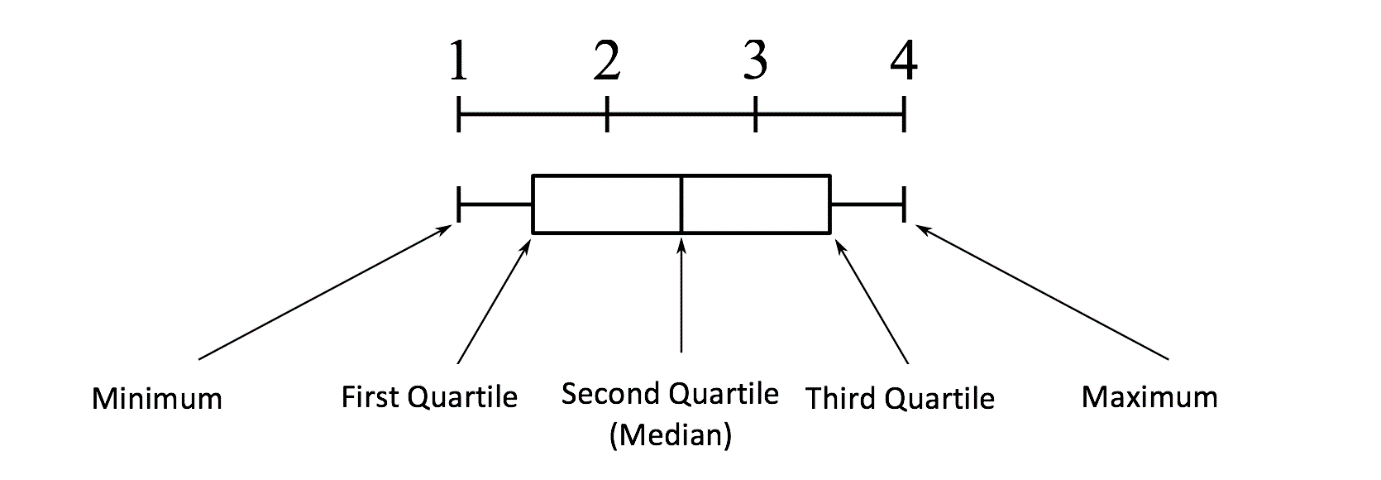


Box plots are a graphical depiction of numerical data through their quantiles. It is a very simple but effective way to visualize outliers. Think about the lower and upper whiskers as the boundaries of the data distribution. Any data points that show above or below the whiskers, can be considered outliers or anomalous.

**Boxplot Anatomy:**

The concept of the Interquartile Range (IQR) is used to build the boxplot graphs. IQR is a concept in statistics that is used to measure the statistical dispersion and data variability by dividing the dataset into quartiles.

In simple words, any dataset or any set of observations is divided into four defined intervals based upon the values of the data and how they compare to the entire dataset. A quartile is what divides the data into three points and four intervals.



Interquartile Range (IQR) is important because it is used to define the outliers. It is the difference between the third quartile and the first quartile (IQR = Q3 -Q1). Outliers in this case are defined as the observations that are below (Q1 − 1.5 x IQR) or boxplot lower whisker or above (Q3 + 1.5 x IQR) or boxplot upper whisker.

