**Random Variable**

What is a random variable?

Any variable (placeholder to store something)

Why is it important?

Every feature in a dataset is essentially a random variable

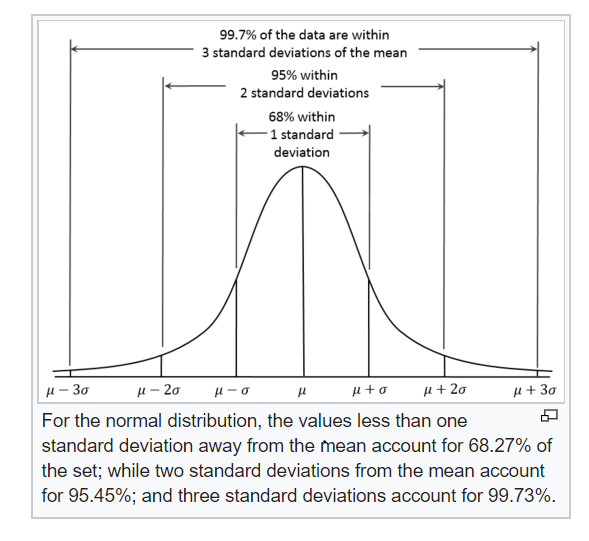
Types of Random Variable:

1. Numerical variable -> e.g. Age (numerical values)
   1. Discrete Random variable -> takes the values of a whole number e.g. Number of bank accounts a person holds, population of a state
   2. Continuous Random variable -> takes any value within a range e.g. Height of different people, salary
2. Categorical variable -> e.g. Gender (distinct number of categories)

**Gaussian Distribution ( Normal Distribution)**

Plotted for a continuous random variable

Empirical formula for Gaussian distribution (forms bell curve) -> specifies the distribution of the data as shown below:



**Measures of central tendency**

Video source: <https://www.youtube.com/watch?v=GvftKv9uctk&list=PLZoTAELRMXVMhVyr3Ri9IQ-t5QPBtxzJO&index=5>

1. Mean(µ)-> Mean is greatly affected by outliers
2. Median( Median helps us to lessen the impact of outliers when denoting the central value)
3. Mode-> Denotes the number with the highest frequency

We handle the missing values by taking a mean of the other numbers for attributes like Age(which are independent) in feature selection