**Data Visualization :**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data visualization translates complex data sets into visual formats that are easier for the human brain to comprehend. This can include a variety of visual tools such as:

* **Charts**: Bar charts, line charts, pie charts, etc.
* **Graphs**: Scatter plots, histograms, etc.
* **Maps**: Geographic maps, heat maps, etc.
* **Dashboards**: Interactive platforms that combine multiple visualizations.

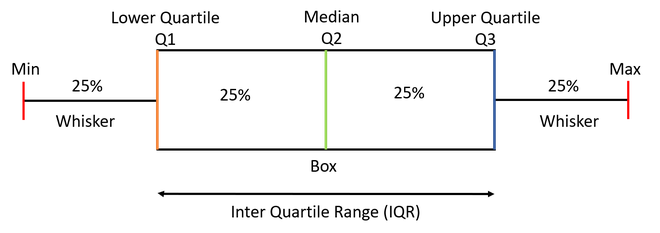
# Box Plot

# Box Plot is a graphical method to visualize data distribution for gaining insights and making informed decisions. Box plot is a type of chart that depicts a group of numerical data through their quartiles.

### **Elements of Box Plot**

A box plot gives a five-number summary of a set of data which is-

* **Minimum** – It is the minimum value in the dataset excluding the outliers.
* **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
* **Median (Q2)**– It is the mid-point of the dataset. Half of the values lie below it and half above.
* **Third Quartile (Q3)**– 75% of the data lies below the Third (Upper) Quartile.
* **Maximum**– It is the maximum value in the dataset excluding the outliers.



## **Data Preprocessing :**

<https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/>

# Distribution :

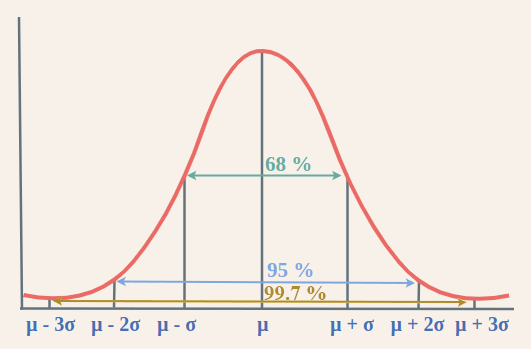
# A  **Distribution**represents the predicted outcomes of various values for a given data. Probability distributions occur in a variety of forms and sizes, each with its own set of characteristics such as mean, median, mode, skewness, standard deviation, kurtosis, etc.

# Normal Distribution

The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by its bell-shaped curve. It is one of the most important distributions in statistics due to its wide range of applications in natural and social sciences. Here are some key characteristics and properties of the normal distribution:

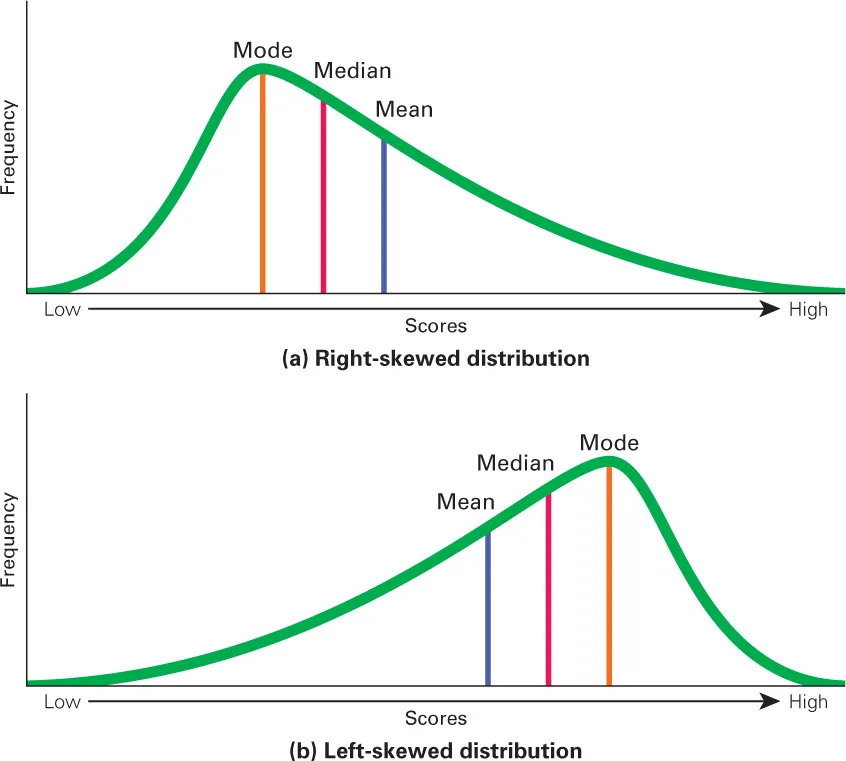
### Key Characteristics of Normal Distribution:

1. **Symmetry**: The normal distribution is symmetric around its mean, meaning that the left and right sides of the distribution are mirror images of each other.
2. **Mean, Median, and Mode**: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.
3. **Bell-shaped Curve**: The distribution has a bell-shaped curve with a single peak at the mean. The tails of the curve approach the horizontal axis asymptotically, but never actually touch it.
4. **Defined by Two Parameters**: The normal distribution is defined by two parameters:
   * **Mean (μ)**: Determines the location of the center of the distribution.
   * **Standard Deviation (σ)**: Determines the spread or width of the distribution. A larger standard deviation results in a wider and flatter curve, while a smaller standard deviation results in a narrower and taller curve.

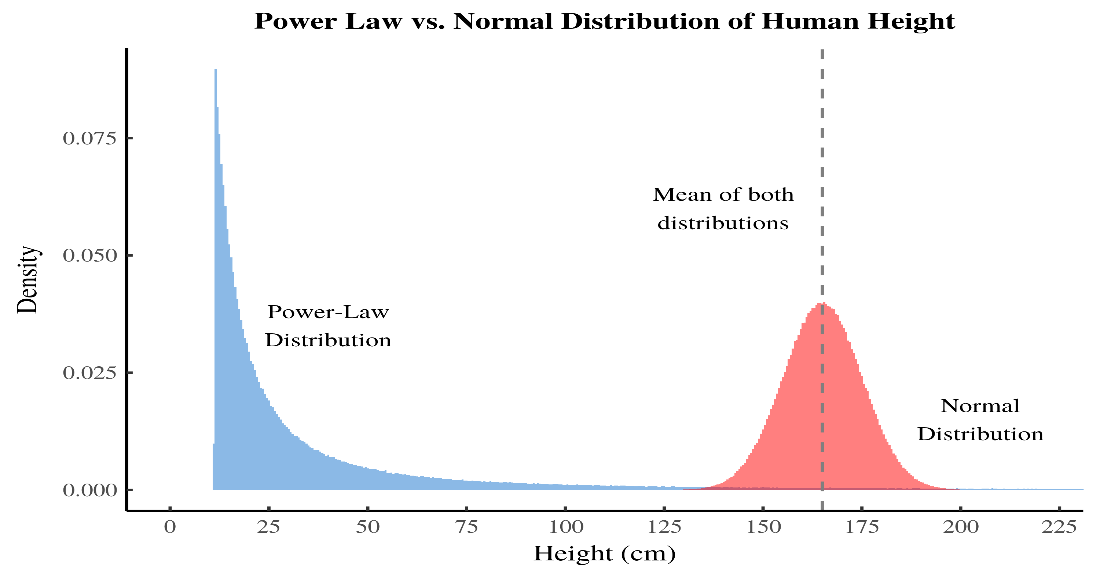


# Lognormal Distribution

The log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. In other words, if XXX is a random variable with a normal distribution, then Y=eXY = e^XY=eX follows a log-normal distribution. The log-normal distribution is used to model data that is positively skewed and spans several orders of magnitude, such as income, stock prices, and natural phenomena like the size of particles in the atmosphere.

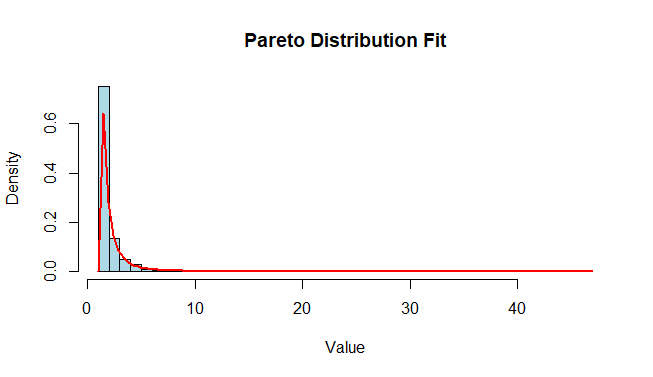


**Power low distribution:**A power-law distribution is a type of probability distribution that has the form P(x)∝x−αP(x) \propto x^{-\alpha}P(x)∝x−α, where α\alphaα is a positive constant known as the exponent or scaling parameter. Power-law distributions are characterized by the property that large events are rare but significant, and small events are common. They are used to model various natural and social phenomena where a few large events have a disproportionate impact.



**Pareto distribution:**

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a continuous probability distribution that is often used to model the distribution of wealth, income, and other phenomena where a small number of occurrences contribute to a large proportion of the effect. This distribution is characterized by its "heavy tail," which indicates that large values are more likely than they would be in other distributions like the normal distribution.



# Ploynomial Regression

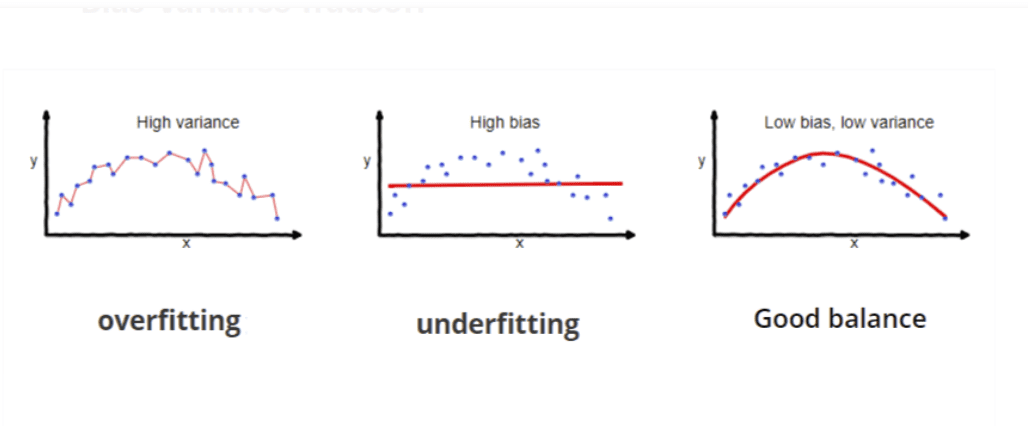
Polynomial regression is a type of regression analysis used in statistics and machine learning when the relationship between the independent variable (input) and the dependent variable (output) is not linear. While simple linear regression models the relationship as a straight line, polynomial regression allows for more flexibility by fitting a polynomial equation to the data.

When the relationship between the variables is better represented by a curve rather than a straight line, polynomial regression can capture the non-linear patterns in the data.

## **Regularization in Machine Learning**

Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting. The commonly used [regularization techniques](https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/) are :

1. Lasso Regularization – L1 Regularization
2. Ridge Regularization – L2 Regularizatio



### **Lasso Regression**

A regression model which uses the **L1 Regularization**technique is called **LASSO(Least Absolute Shrinkage and Selection Operator)** regression. **Lasso Regression** adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function(L). Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model.

*where,*

* ***m****– Number of Features*
* ***n****– Number of Examples*
* ***y\_i****– Actual Target Value*
* ***y\_i(hat)****– Predicted Target Value*

### **Ridge Regression**

A regression model that uses the **L2 regularization** technique is called **Ridge regression**. **Ridge regression** adds the “squared magnitude” of the coefficient as a penalty term to the loss function(L).