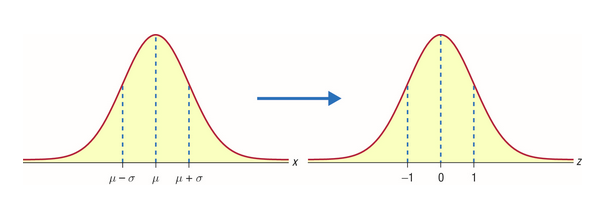
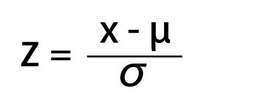
**Normal Distribution and Skewness**

Most of the machine learning models like Linear Regression assume that the data is normally distributed. If this assumption fails, the model fails to give accurate predictions.

# Normal Distribution

A normal distribution is **symmetric** about the mean and follows a **bell shaped curve**. And almost 99.7% of the values lies within 3 standard deviation. The mean, median and mode of a normal distribution are equal. A probability distribution with the mean 0 and standard deviation of 1 is known as standard normal distribution or Gaussian distribution. For any Normal distribution, we can convert it into Standard Normal distribution using the formula:



To understand the importance of converting Normal Distribution into Standard Normal Distribution, let’s suppose there are two students:*Ross and Rachel.*Ross scored 65 in the exam of paleontology and Rachel scored 80 in the fashion designing exam.

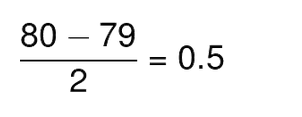
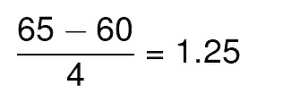
*Can we conclude that Rachel scored better than Ross?*

No**,** because the way people performed in paleontology may be different from the way people performed in fashion designing. The variability may not be the same here.

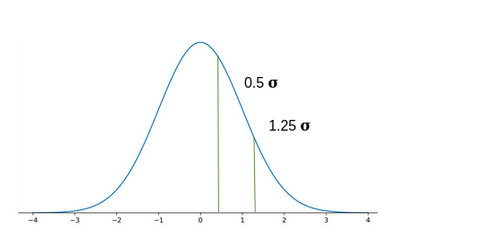
So, a direct comparison by just looking at the scores will not work.

Now, let’s say the paleontology marks follow a normal distribution with mean 60 and a standard deviation of 4. On the other hand, the fashion designing marks follow a normal distribution with mean 79 and standard deviation of 2.

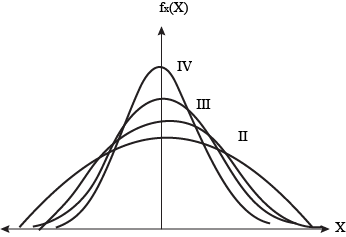
We will have to calculate the**z score** by standardization of both these distributions:



Thus, Ross scored 1.25 standard deviations above the mean score while Rachel scored only 0.5 standard deviations above the mean score. Hence we can say that *Ross Performed better than Rachel.*

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Screenshot-from-2020-03-06-17-33-00.png)

**Note**: One important point to note that **sharpness of the curve defines the variance**. Curve IV is normally distributed as compared to other thus having least variance.



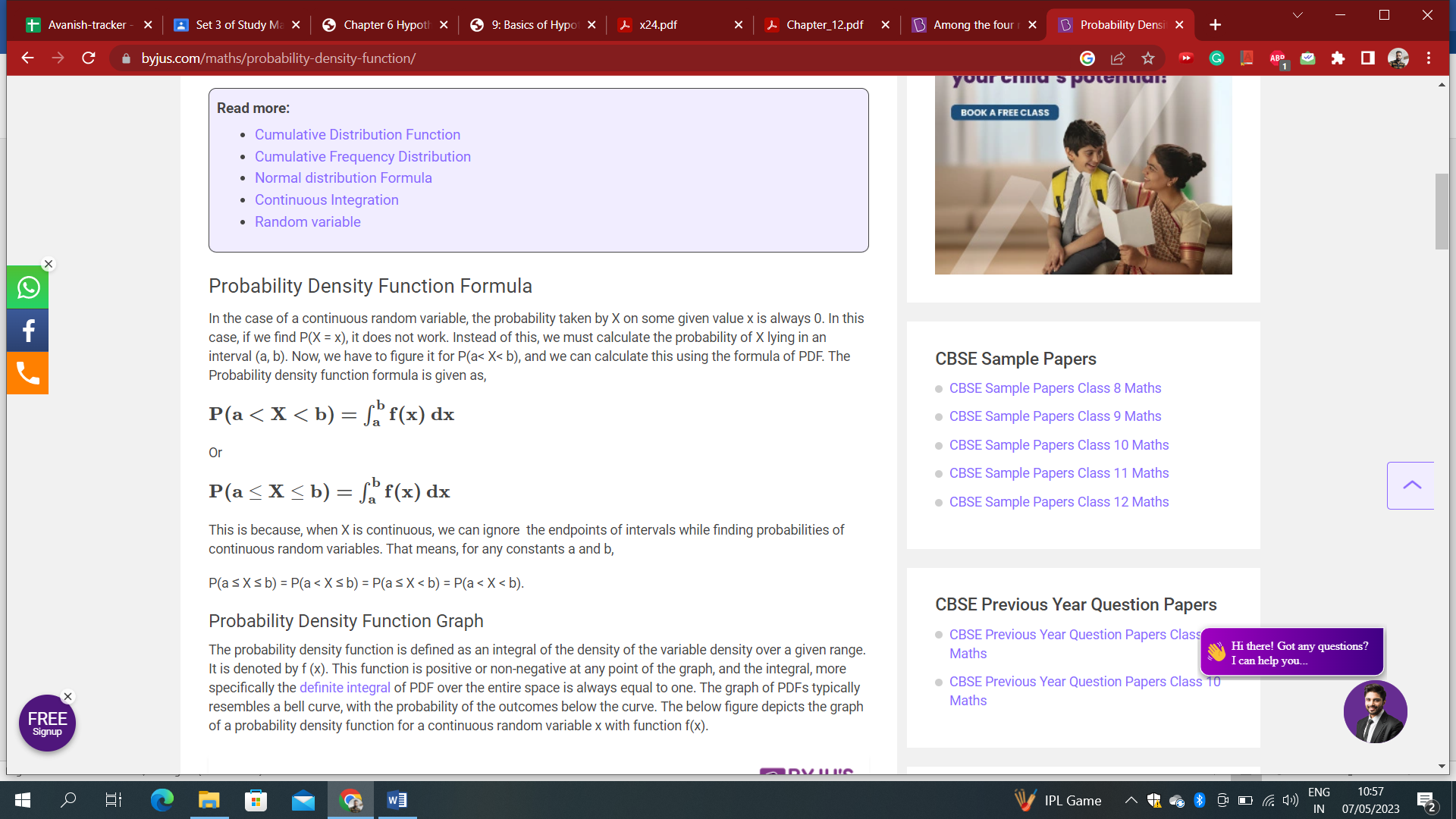
# Probability Density Function

# In probability theory, a probability density function (PDF) is used to define the continuous random variable’s probability coming within a distinct range of values, as opposed to taking on any one value. The function explains the probability density function of normal distribution and how mean and deviation exists.

# Often it is referred to as cumulative distribution function or sometimes as [probability mass function](https://byjus.com/maths/probability-mass-function/)(PMF). However, the actual truth is PDF (probability density function) is defined for continuous random variables, whereas PMF (probability mass function) is defined for discrete random variables.

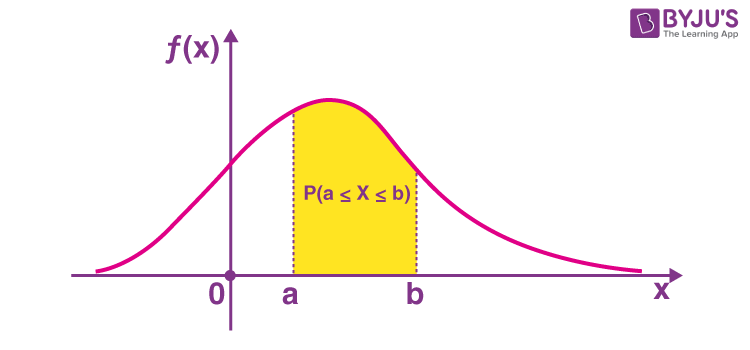
## **Formula**

In the case of a continuous random variable, the probability taken by X on some given value x is always 0. In this case, if we find P(X = x), it does not work. Instead of this, we must calculate the probability of X lying in an interval (a, b). Now, we have to figure it for P(a< X< b), and we can calculate this using the formula of PDF. The Probability density function formula is given as,



This is because, when X is continuous, we can ignore the endpoints of intervals while finding probabilities of continuous random variables. That means, for any constants a and b,

P(a ≤ X ≤ b) = P(a < X ≤ b) = P(a ≤ X < b) = P(a < X < b).



# Skewness

Data is skewed when its distribution curve is asymmetrical (as compared to a normal distribution curve that is perfectly symmetrical) and *skewness* is the measure of the asymmetry. In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the [normal distribution](https://www.analyticsvidhya.com/blog/2020/04/statistics-data-science-normal-distribution/?utm_source=blog&utm_medium=what-is-skewness-statistics). Well, the normal distribution is the probability distribution without any skewness. The skewness for a normal distribution is 0. You can look at the image below, which shows symmetrical distribution that’s a normal distribution, and you can see that it is symmetrical on both sides of the dashed line.

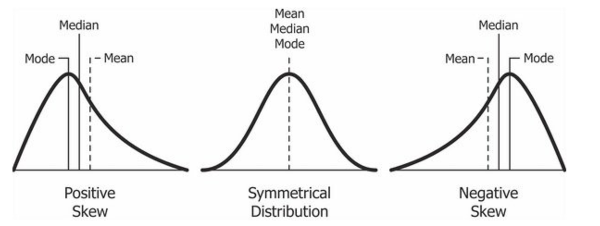
There are 2 different types of skews in data:

* left(negative) skew
* right(positive) skew

The probability distribution with its tail on the right side is a positively skewed distribution, and the one with its tail on the left side is a negatively skewed distribution. For a positive skewness mean > median > mode while for a negative skewness mean < median < mode.

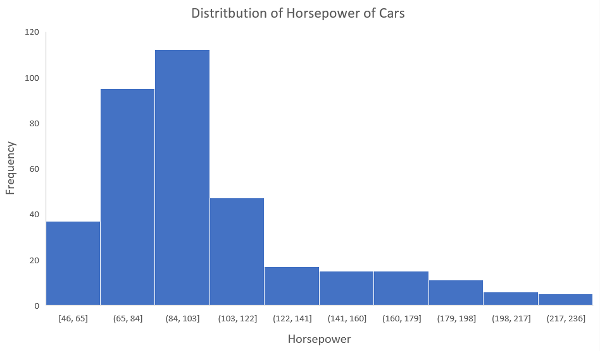
Now, the question is when we can say our data is moderately skewed or heavily skewed?

The thumb rule is: If the skewness is between -0.5 to +0.5 then we can say data is fairly symmetrical. If the skewness is between -1 to -0.5 or 0.5 to 1 then data is moderately skewed. And if the skewness is less than -1 and greater than +1 then our data is heavily skewed.



**Effects of skewed data:** Degrades the model’s ability (especially regression based models that work on the assumption that the distribution of the independent variable and the target variable are similar) to describe typical cases as it has to deal with rare cases on extreme values. i.e. right skewed data will predict better on data points with lower value as compared to those with higher values. Skewed data also does not work well with many statistical methods. However, ***tree based models are not affected***.

Let’s take a look at the below distribution for better understanding the effect of skewness. It is the distribution of horsepower of cars:



You can clearly see that the above distribution is positively skewed. Now, let’s say you want to use this as a feature for the model that will predict the mpg (miles per gallon) of a car.

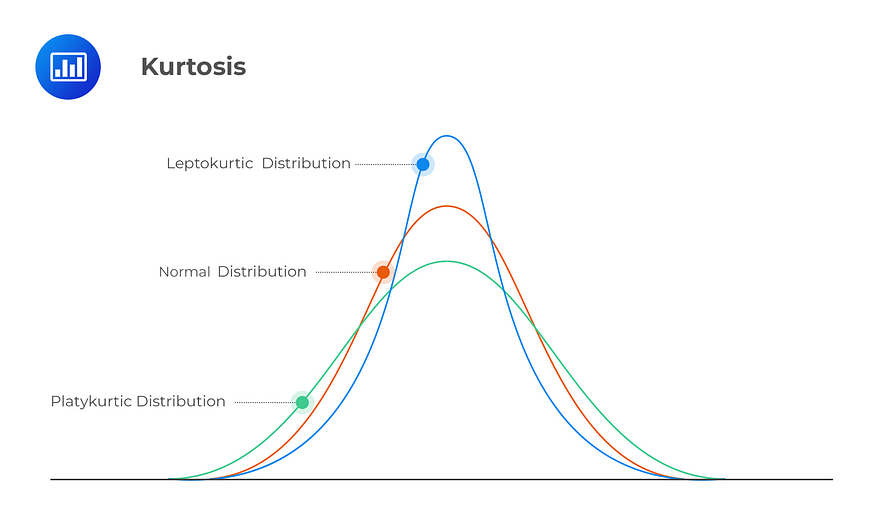
Since our data is positively skewed here, it means that it has a higher number of data points having low values, i.e., cars with less horsepower. So when we train our model on this data, it will perform better at predicting the mpg of cars with lower horsepower as compared to those with higher horsepower.

Also, skewness tells us about the direction of [outliers](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?utm_source=blog&utm_medium=what-is-skewness-statistics). You can see that our distribution is positively skewed, and most of the outliers are present on the right side of the distribution.

**Note:** The skewness does not tell us about the number of outliers. It only tells us the direction.

# Kurtosis

Kurtosis tell us about the tails behaviour. It is actually the measure of outliers present in the distribution.



Kurtosis are of three types:

* **Mesokurtic:** When the tails of the distribution is similar to the normal distribution then it is mesokurtic. The kurtosis for normal distribution is 3.
* **Leptokurtic:** If the kurtosis is greater than 3 then it is leptokurtic. In this case, the tails will be heavier than the normal distribution which means lots of outliers are present in the data. It can be recognized as thin bell shaped distribution with peak higher than normal distribution.
* **Platykurtic:** Kurtosis will be less than 3 which implies thinner tail or lack of outliers than normal distribution. In case of platykurtic, bell shaped distribution will be broader and peak will be lower than the mesokurtic.

**Skewed Data Handling:**

To ensure that the machine learning model capabilities is not affected, skewed data has to be transformed to approximate to a normal distribution. The method used to transform the skewed data depends on the characteristics of the data.

**To check for skew and kurtosis in data:**

df.skew().sort\_values(ascending=False)

df.kurtosis().sort\_values(ascending=False)

**Dealing with skew data:**

1.log transformation

2.Remove outliers

3.Normalize (min-max)

4.Box Cox transformation

**Skewness in target variable:**Use class balancing technique like under sampling, oversampling or SMOTE