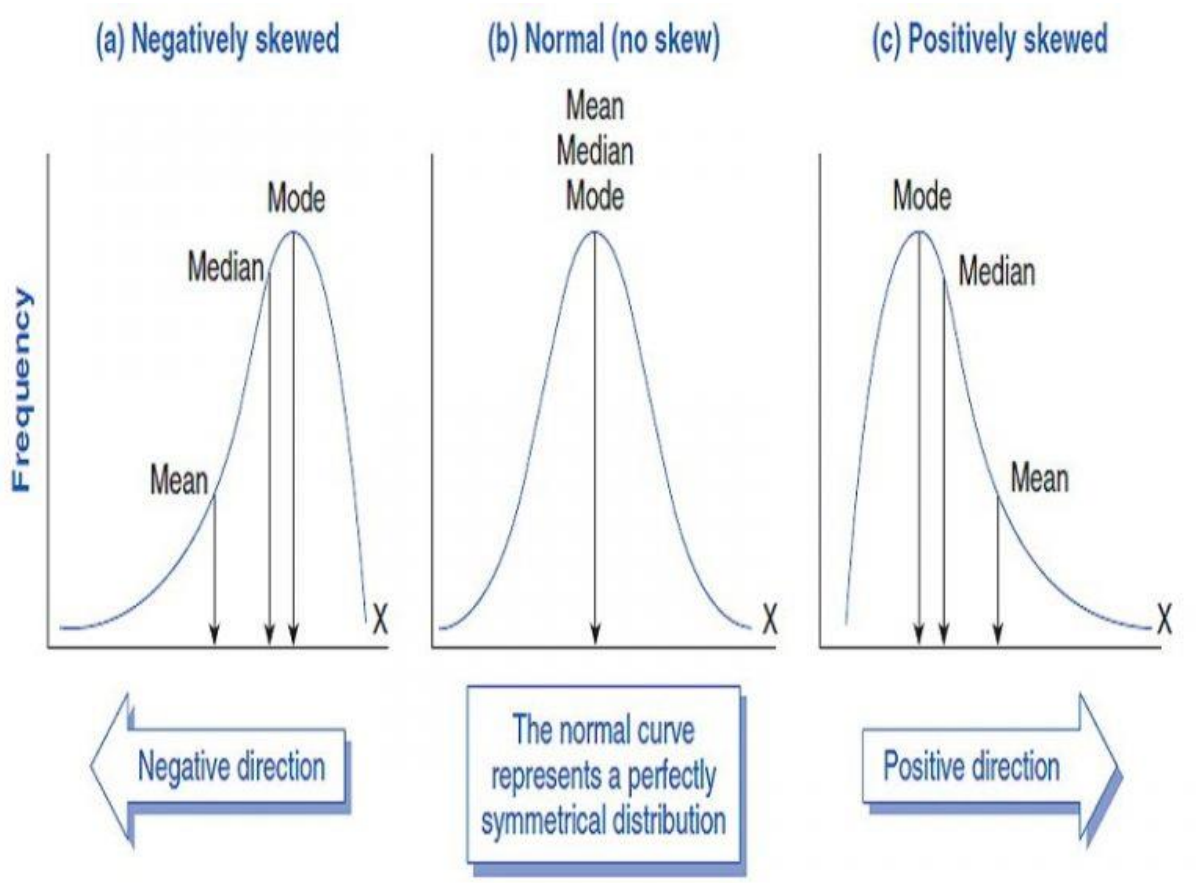
SKEWNESS

Definition:

- Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.
- A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak.

Types of Skewness:



Positive Skewness: When the tail of the distribution extends towards the right side, it's called positively skewed. Imagine a distribution where most data points are clustered on the left, and a few outliers stretch the right tail.

Condition for positive skewness = **Mean > Median >Mode**

Negative Skewness: Conversely, when the tail extends towards the left side, it's negatively skewed.

Condition for negative skewness = **Mode > Median > Mean**

Zero Skewness: When a distribution has zero skew, it is symmetrical. Its left and right sides are mirror images. Normal distributions have zero skew, but they're not the only distributions with zero skew. Any symmetrical distribution, such as a uniform distribution or some bimodal (two-peak) distributions, will also have zero skew.

Condition for zero skewness is **Mean = Mode = Median**

Formula:

$$skewness = \frac{1}{n}\frac{\sum_{j=1}^{n}(x_j - \bar{\bar{x}})^3}{s^3}$$

Where s is Standard Deviation.

Use of Skewness in Data Science:

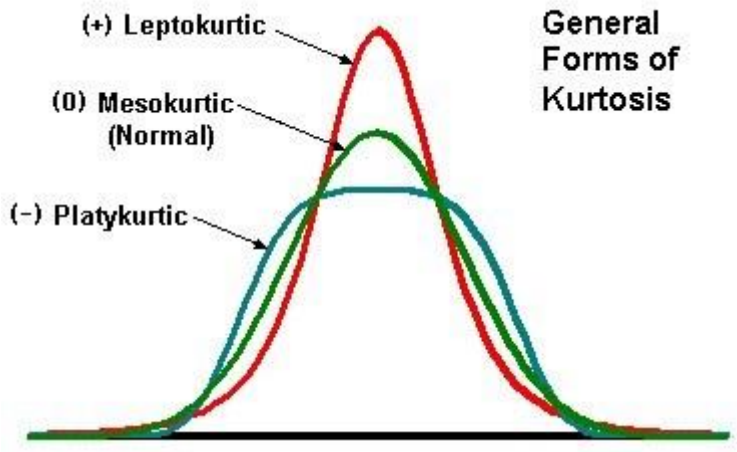
- Understanding skewness helps analysts decide how data looks and what methods to use. It's crucial for accurate data analysis.
- Skewness impacts various statistical techniques, such as regression models and hypothesis testing.
- When dealing with skewed data, transformations (like power, log, or exponential transformations) can be applied to make the distribution more symmetric.

KURTOSIS

Definition:

- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution.
- In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.
- Along with skewness, kurtosis is an important descriptive statistic of data distribution. However, the two concepts must not be confused with each other. Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.

- In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high risk for an investment because it indicates high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.



Types of Kurtosis:

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values, as well as values close to zero.

1. Mesokurtic (kurtosis=3)

Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. This means that if the data follows a normal distribution, it follows a mesokurtic distribution.

2. Leptokurtic (kurtosis>3)

Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.

3. Platykurtic (kurtosis<3)

A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the

investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.

Formula:

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^4$$

Importance of kurtosis in Data Science:

- **Outlier Detection:** High kurtosis indicates the presence of outliers or extreme values. Data scientists can use kurtosis to identify and handle outliers appropriately.
- **Model Assumptions:** Some statistical models assume that the data follows a normal distribution. Kurtosis helps us assess whether the data deviates significantly from normality.
- **Choosing Analysis Techniques:** Depending on the kurtosis value, data analysts can select appropriate analysis techniques. For instance, t-tests and ANOVA assume normality, so understanding kurtosis guides our choice of tests.

Skewness and Kurtosis output for placement dataset:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	108	67.3034	66.3347	66.3586	72.1006	62.2782	277649
Median	108	67	65	66	71	62	265000
Mode	1	62	63	65	60	56.7	300000
Q1:25%	54.5	60.6	60.9	61	60	57.945	240000
Q2:50%	108	67	65	66	71	62	265000
Q3:75%	161.5	75.7	73	72	83.5	66.255	300000
Q4:100%	215	89.4	91.15	88.5	98	77.89	390000
IQR	107	15.1	12.1	11	23.5	8.31	60000
1.5rule	160.5	22.65	18.15	16.5	35.25	12.465	90000
Lesser	-106	37.95	42.75	44.5	24.75	45.48	150000
Greater	322	98.35	91.15	88.5	118.75	78.72	390000
Min	1	40.89	42.75	50	50	51.21	200000
Max	215	89.4	91.15	88.5	98	77.89	390000
kurtosis	-1.2	-0.60751	0.0869008	-0.0974897	-1.08858	-0.470723	-0.239837
skew	0	-0.132649	0.162611	0.204164	0.282308	0.313576	0.8067

Kurtosis:

- Kurtosis value for ssc_p is less than 3, so it is platykurtic.
- Kurtosis value for hsc_p is less than 3, so it is platykurtic.
- Kurtosis value for degree_p is less than 3, so it is platykurtic.
- Kurtosis value for etest_p is less than 3, so it is platykurtic.
- Kurtosis value for mba_p is less than 3, so it is platykurtic.
- Kurtosis value for salary is less than 3, so it is platykurtic.

Skewness:

- When comparing the mean, median and mode values of ssc_p column, Mean>Median>Mode, so it is negative skew.
- When comparing the mean, median and mode values of hsc_p column, Mean>Median>Mode, so it is negative skew.
- When comparing the mean, median and mode values of degree_p column, Mean>Median>Mode, so it is negative skew.
- When comparing the mean, median and mode values of etest_p column, Mean>Median>Mode, so it is negative skew.
- When comparing the mean, median and mode values of mba_p column, Mean>Median>Mode, so it is negative skew.
- When comparing the mean, median and mode values of salary column, Mean>Median>Mode, so it is negative skew.