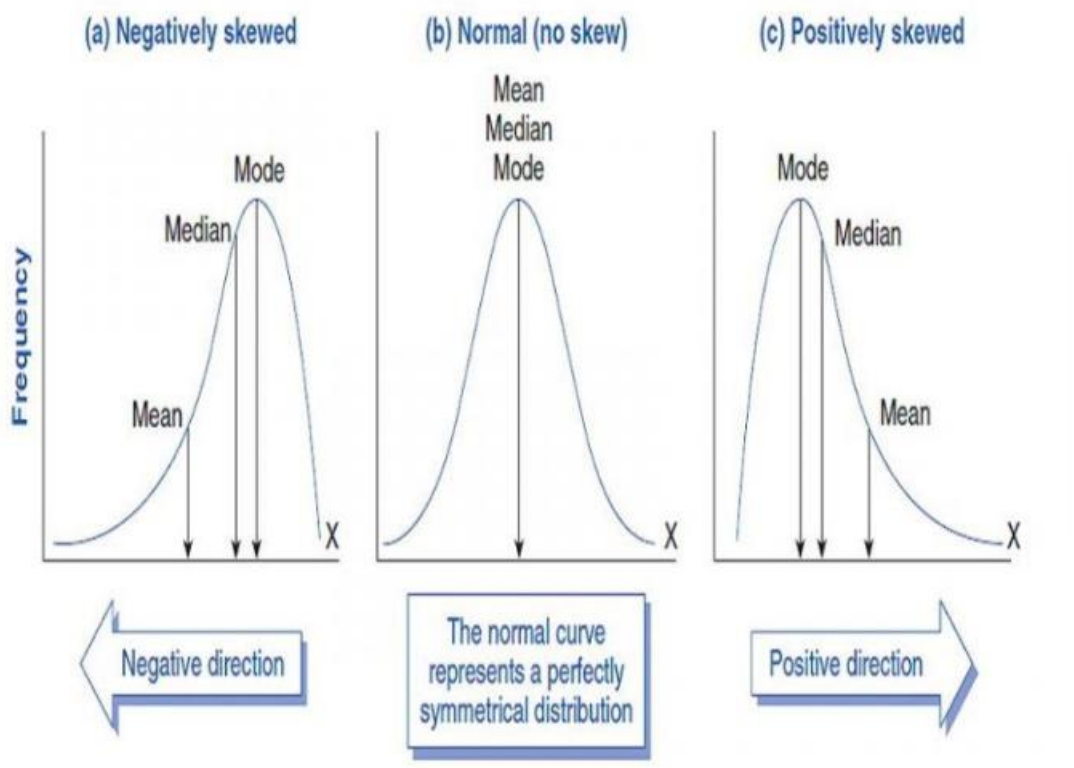
Skewness

- 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 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 of zero skewness is Mean = Median = Mode

The Formula of Skewness is:

$$\text{Skewness} = \frac{\sum \left(x - \bar{x}\right)^3}{(n - 1) \cdot S^3}$$

Where:
S: standard deviation
$\bar{X}$: Mean

Use of skewness in Data Science

- Skewness helps analysts to 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.

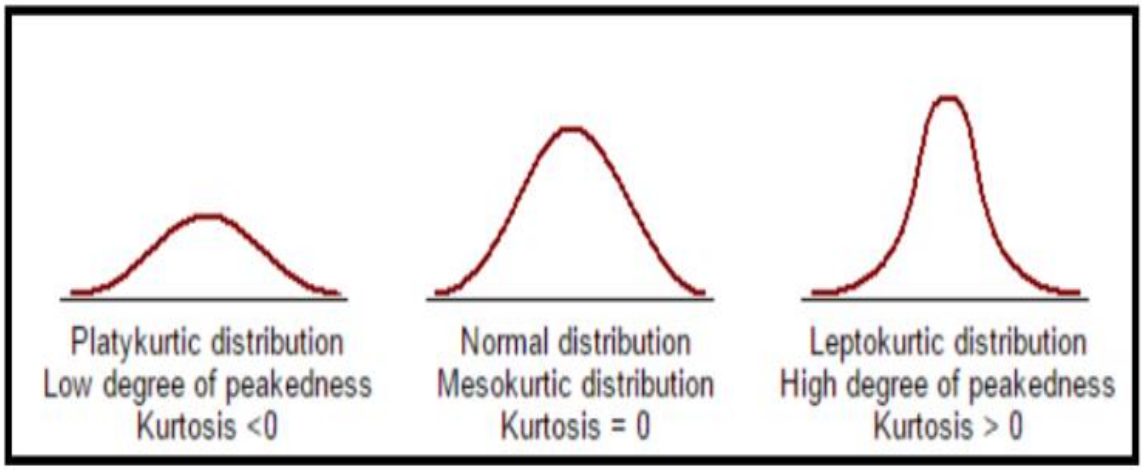
Kurtosis

Kurtosis is a statistical number that tells us if a distribution is taller or shorter than a normal distribution. If a distribution is similar to the normal distribution, the Kurtosis value is 0. If Kurtosis is greater than 0, then it has a higher peak compared to the normal distribution. If Kurtosis is less than 0, then it is flatter than a normal distribution.

There are three types of distributions:

Leptokurtic: Sharply peaked with fat tails, and less variable.
Mesokurtic: Medium peaked

Platykurtic: Flattest peak and highly dispersed.



The Formula of kurtosis is:

$$\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n-1) \cdot S^4}$$

Where:
S: standard deviation $\bar{X}$: Mean

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	108.0	67.303395	66.334744	66.358558	72.100558	62.278186	277648.648649
Median	108.0	67.0	65.0	66.0	71.0	62.0	265000.0
Mode	1	62.0	63.0	65.0	60.0	56.7	300000.0
Q1:25%	54.5	60.6	60.9	61.0	60.0	57.945	240000.0
Q2:50%	108.0	67.0	65.0	66.0	71.0	62.0	265000.0
Q3:75%	161.5	75.7	73.0	72.0	83.5	66.255	300000.0
Q4:100%	215.0	89.4	91.15	88.5	98.0	77.89	390000.0
IQR	107.0	15.1	12.1	11.0	23.5	8.31	60000.0
1.5rule	160.5	22.65	18.15	16.5	35.25	12.465	90000.0
Lesser	-106.0	37.95	42.75	44.5	24.75	45.48	150000.0
Greater	322.0	98.35	91.15	88.5	118.75	78.72	390000.0
Min	1	40.89	42.75	50.0	50.0	51.21	200000.0
Max	215	89.4	91.15	88.5	98.0	77.89	390000.0
kurtosis	-1.2	-0.60751	0.086901	-0.09749	-1.08858	-0.470723	-0.239837
skew	0.0	-0.132649	0.162611	0.204164	0.282308	0.313576	0.8067

Summary

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.

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.