Understanding the shape of data is crucial while practicing data science. It helps to understand where the most information lies and analyze the outliers in a given data. In this article, we’ll learn about the shape of data, the importance of skewness, and kurtosis in statistics. The types of ***skew and kurtosis*** , Analyze the shape of data in the given dataset. Let’s first understand what skewness and kurtosis is.

one of the must-do steps in EDA is checking the shapes of distributions. Correctly identifying the shape influences many decisions later on in the project such as:

* Further preprocessing steps
* Whether or not perform outlier detection and maybe, removal
* Feature transformation or scaling steps
* Feature selection
* Algorithm selection

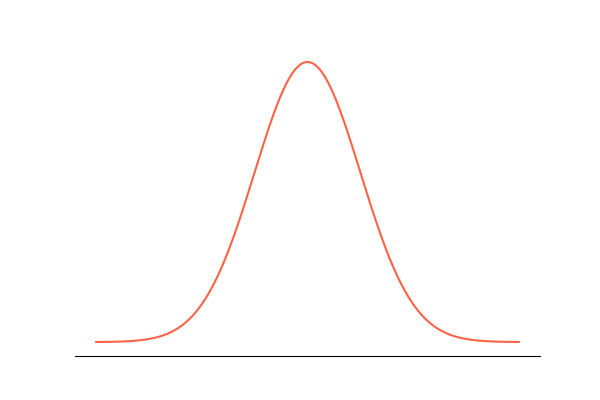
and so on. While there are visuals to do the task, you need more reliable metrics to quantify various characteristics of distributions. Two of such metrics are **skewness** and **kurtosis**. You can use them to assess the resemblance between your distributions and a perfect, normal distribution.

**What is Skewness?**

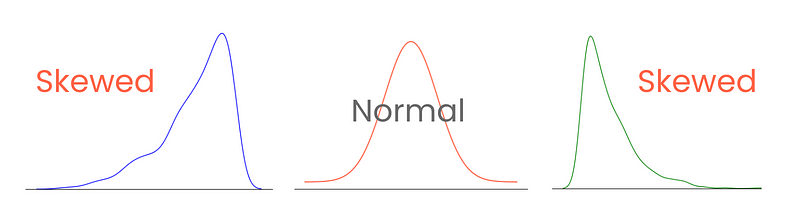
We see normal distribution everywhere: human body measurements, weights of objects, IQ scores, test results, or even at the gym

Besides being nature’s favorite distribution, it is universally loved by almost all machine learning algorithms. While some want it to improve and stabilize their performance, some outright refuse to work well with anything other than normal distribution (I am talking to you, linear models).

So, to satisfy the algorithms’ need for normalcy, we need a way to measure how similar or (dissimilar) our own distributions are compared to the perfect bell-shaped curve.



Let’s start with the tails. In a perfect normal distribution, the tails are equal in length. But, when there is asymmetry between the tails, giving it a leaned, squished-to-one-side look, we say it is skewed. And you guessed it, **we measure the extent of this asymmetry with skewness**.



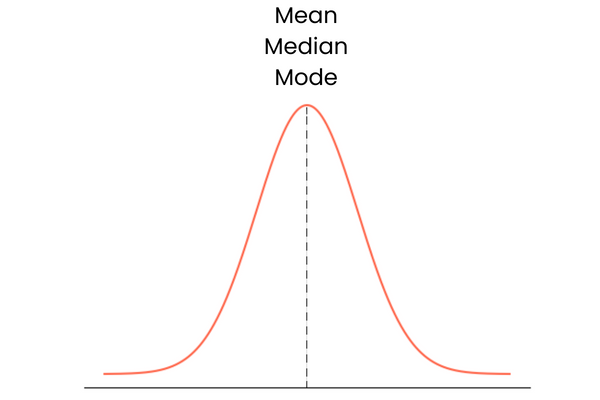
Correctly categorizing and measuring skewness provides insights into how values are spread around the mean and influence the choices of statistical techniques and data transformations. For example, highly skewed distributions might benefit from normalization or scaling techniques to make them resemble normal distribution. This would aid in model performance.

**Types of Skewness**

There are three types of skewness: positive, negative, and zero skewness.

Let’s start with the last one. A distribution with zero skewness has the following characteristics:

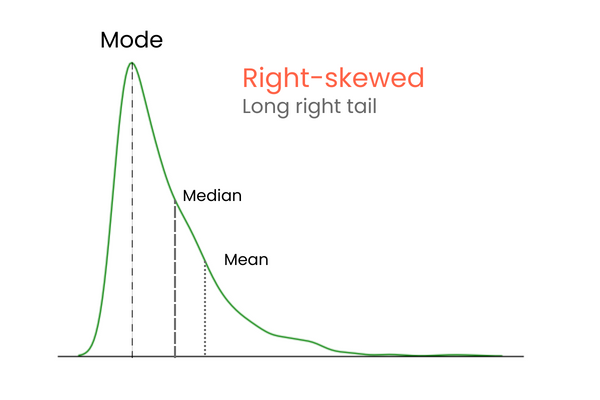
* Symmetric distribution with values evenly centered around the mean.
* No skew, lean or tail to either side.
* The mean, median, and mode are all at the center point.



In practice, mean, median, and mode may not form a perfect overlapping straight line. They may be slightly away from each other but the difference would be too small to matter.

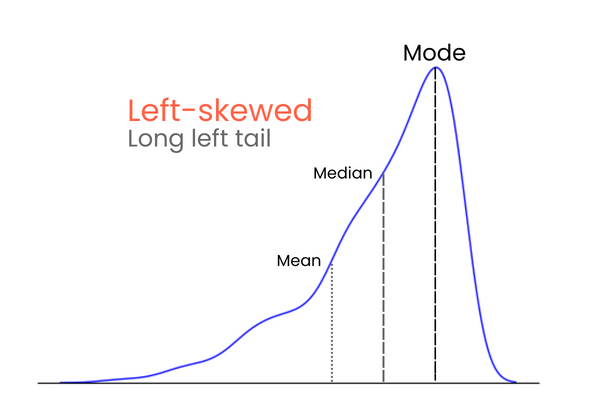
In a distribution with positive skewness (right-skewed):

* The right tail of the distribution is longer or fatter than the left.
* The mean is greater than the median, and the mode is less than both mean and median.
* Lower values are clustered in the “hill” of the distribution, while extreme values are in the long right tail.
* It is also known as right-skewed distribution.



In a distribution with negative skewness (left-skewed):

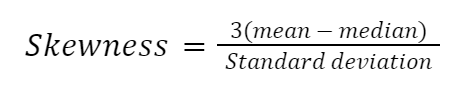
* The left tail of the distribution is longer or fatter than the right.
* The mean is less than the median, and the mode is greater than both mean and median.
* Higher values are clustered in the “hill” of the distribution, while extreme values are in the long left tail.
* It is also known as left-skewed distribution.



To remember the differences between positive and negative skewness, think of it this way: if you want to increase the mean of a distribution, you should add much higher values than the mean to the distribution. To lower the mean, you should do the opposite — introduce much lower values than the mean to the distribution. So, if the majority of the extreme values is higher than the mean, the skewness will be positive because they increase the mean. If the majority of extreme values is smaller than the mean, the skewness is negative because they decrease the mean.

**How to Calculate Skewness in Python**

There are many ways to calculate skewness, but the simplest one is Pearson’s second skewness coefficient, also known as median skewness.



Let’s implement the formula manually in Python:

import numpy as np

import pandas as pd

import seaborn as sns

# Example dataset

diamonds = sns.load\_dataset("diamonds")

diamond\_prices = diamonds["price"]

mean\_price = diamond\_prices.mean()

median\_price = diamond\_prices.median()

std = diamond\_prices.std()

skewness = (3 \* (mean\_price - median\_price)) / std

>>> print(

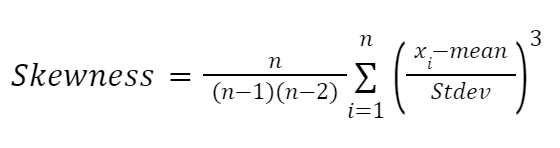
f"The Pierson's second skewness score of diamond prices distribution is {skewness:.5f}"

)

The Pierson's second skewness score of diamond prices distribution is 1.15189

[**Powered By**](https://www.datacamp.com/datalab)

Another formula highly influenced by the works of Karl Pearson is the moment-based formula to approximate skewness. It is more reliable and given as follows:



Here:

* n represents the number of values in a distribution
* x\_i denotes each data point

While all formulas to approximate skewness return different scores, their differences are too small to be significant or change the categorization of the skew. For example, all methods we have used today leverage different formulas under the hood, but the results are very close.

Once you calculate skewness, you can categorize the extent of the skew:

* (-0.5, 0.5) — low or approximately symmetric.
* (-1, -0.5) U (0.5, 1) — moderately skewed.
* Beyond -1 and 1 — Highly skewed.

**What is Skewness?**

**Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.**

**Positive skewness indicates a longer tail on the right side of the distribution, while negative**[**skewness indicates**](https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/)**a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.**

**Depending on the model,**[**skewness**](https://www.analyticsvidhya.com/blog/2022/09/importance-of-skewness-kurtosis-co-efficient-of-variation/)**in the values of a specific independent variable (feature) may violate model assumptions or diminish the interpretation of feature importance.**

**A probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data exhibits skewness, which is a measure of asymmetry in statistics.**

**A skewed data set, typical values fall between the first quartile (Q1) and the third quartile (Q3).**

**The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.**

**In a symmetrically distributed dataset, both the left-hand side and the right-hand side have an equal number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.**

**Types of Skewness**

**Positive Skewed or Right-Skewed  (Positive Skewness)**

**In statistics, a positively skewed or right-**[**skewed distribution**](https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/)**has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.**

**In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.**

**Extreme positive skewness is not desirable for a distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.**

**Negative Skewed or Left-Skewed (Negative Skewness)**

**A distribution with a long left tail, known as negatively skewed or left-skewed, stands in complete contrast to a positively skewed distribution. skewness and kurtosis in statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.**

**In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.**

**Median is the middle value, and mode is the most frequent value. Due to an unbalanced distribution, the median will be higher than the mean.**

**How to Calculate the Skewness Coefficient?**

**Various methods can calculate**[**skewness**](https://github.com/manvendra7/Skewness-and-kurtosis/blob/master/SKEWNESS_AND_KURTOSIS_(1).ipynb)**, with Pearson’s coefficient being the most commonly used method.**

**Pearson’s first coefficient of skewness  
To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.**

**As Pearson’s correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of -1 to +1. That accurately shows the range of the correlation values.**

**Pearson’s first coefficient of skewness is helping if the data present high mode. However, if the data exhibits low mode or multiple modes, it is preferable not to use Pearson’s first coefficient, and instead, Pearson’s second coefficient may be superior, as it does not depend on the mode.**

**Pearson’s second coefficient of skewness  
subtract the median from the *mean*, *multiply the difference by 3, and divide the product by the standard deviation.***

**Rule of thumb:**

* **For skewness values between -0.5 and 0.5, the data exhibit approximate symmetry.**
* **Skewness values within the range of -1 and -0.5 (negative skewed) or 0.5 and 1(positive skewed) indicate slightly skewed data distributions.**
* **Data with skewness values less than -1 (negative skewed) or greater than 1 (positive skewed) are considered highly skewed.**

**What is Kurtosis?**

**Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.**

**Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.**

**The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.**

**Peakedness in a data distribution is the degree to which data values are concentrated around the mean. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.**

**In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are 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.**

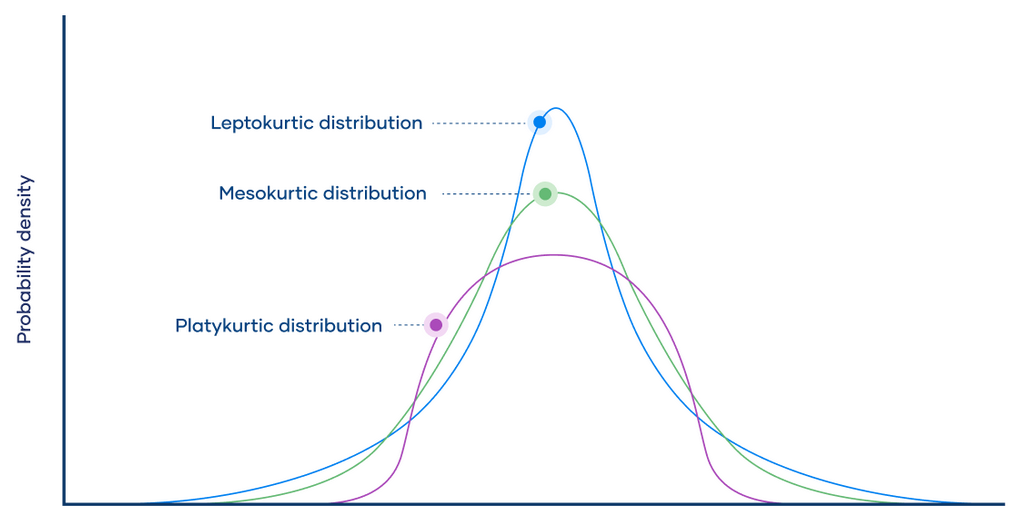
**What is Excess Kurtosis?**

**In statistics and probability theory, researchers use excess kurtosis to compare the kurtosis coefficient with that of a normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution). Since normal**[**distributions**](https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/)**have a kurtosis of 3, excess kurtosis is calculated by subtracting kurtosis by 3.**

**Excess kurtosis  =  Kurt – 3**

**Types of Kurtosis**

**Kurtosis is a statistical measure that describes the shape of a probability distribution’s tails relative to its peak. There are three main types of kurtosis:**

****

1. **Mesokurtic: A distribution with mesokurtic kurtosis has a similar peak and tail shape as the normal distribution. It has a kurtosis value of around 0, indicating that its tails are neither too heavy nor too light compared to a normal distribution.**
2. **Leptokurtic: A distribution with leptokurtic kurtosis has heavier tails and a sharper peak than the normal distribution. It has a positive kurtosis value, indicating that it has more extreme outliers than a normal distribution. This type of distribution is often associated with higher peakedness and a greater probability of extreme values.**
3. **Platykurtic: A distribution with platykurtic kurtosis has lighter tails and a flatter peak than the normal distribution. It has a negative kurtosis value, indicating that it has fewer extreme outliers than a normal distribution. This type of distribution is often associated with less peakedness and a lower probability of extreme values.**

**Skewness and Kurtosis Formula**

**Skewness and kurtosis are two statistical measures that describe the shape of a distribution. Let’s look at skewness and kurtosis formula in the next section!**

**Skewness Formula**

**Skewness measures the asymmetry of a distribution. A symmetrical distribution has a skewness of zero. Positive skewness indicates that the right tail of the distribution is longer or fatter than the left tail, while negative skewness indicates the opposite.**

**The formula for skewness (often denoted by 𝛾1*γ*1​) for a sample is:**

***γ*1​=(*n*−1)(*n*−2)*n*​∑*i*=1*n*​(*sxi*​−*x*ˉ​)3**

**Where:**

* **𝑛*n* is the number of observations in the sample**
* **𝑥𝑖*xi*​ is the ith observation**
* **𝑥ˉ*x*ˉ is the sample mean**
* **𝑠*s* is the sample standard deviation**

**Kurtosis Formula**

**Kurtosis measures the peakedness or flatness of a distribution relative to the normal distribution. A normal distribution has a kurtosis of 3, known as the excess kurtosis. Deviations from this value indicate how much the distribution deviates from the normal, with positive excess kurtosis indicating a more peaked distribution and negative excess kurtosis indicating a flatter one.**

**The formula for kurtosis (often denoted by 𝛾2*γ*2​) for a sample is:**

***γ*2​=(*n*−1)(*n*−2)(*n*−3)*n*(*n*+1)​∑*i*=1*n*​(*sxi*​−*x*ˉ​)4−(*n*−2)(*n*−3)3(*n*−1)2​**

**Where:**

* **𝑛*n* is the number of observations in the sample**
* **𝑥𝑖*xi*​ is the ith observation**
* **𝑥ˉ*x*ˉ is the sample mean**
* **𝑠*s* is the sample standard deviation**

**These formulas give the sample skewness and kurtosis. For population skewness and kurtosis, the divisor 𝑛*n* in the formulas is replaced with 𝑛−1*n*−1 and 𝑛−2*n*−2, respectively.**

**Difference Between Skewness and Kurtosis**

| **Skewness** | **Kurtosis** |
| --- | --- |
| **Skewness measures the asymmetry of a probability distribution** | **Kurtosis measures the tailedness or peakedness of a probability distribution** |
| **Positive skew indicates a right-skewed distribution, with the tail extending to the right** | **Positive kurtosis indicates a distribution with heavier tails, often referred to as “leptokurtic”** |
| **Negative skew indicates a left-skewed distribution, with the tail extending to the left** | **Negative kurtosis indicates a distribution with lighter tails, often referred to as “platykurtic”** |
| **A skewness value of zero indicates a symmetric distribution** | **A kurtosis value of zero indicates a distribution similar to the normal distribution, often referred to as “mesokurtic”** |
| **Used to identify the direction and degree of asymmetry** | **Used to identify the presence of outliers or extreme values** |
| **Sensitive to changes in the tails of the distribution** | **Sensitive to changes in the center and shoulders of the distribution** |
| **Commonly used in fields such as economics, finance, and social sciences** | **Commonly used in statistics, engineering, and physical sciences** |
| **Examples: income distribution, stock returns** | **Examples: particle physics, image processing** |

**Conclusion**

**Skewness and Kurtosis naturally complement each other in analyzing data distributions.**[**Skewness**](https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/)**, which measures the symmetry or asymmetry of data distribution, helps us understand if the data is pushed towards one side or the other. For instance, positive skewness indicates a distribution pushed towards the right side, while negative skewness implies a distribution pushed towards the left side. On the other hand, Kurtosis helps determine whether the data exhibits a heavy-tailed or light-tailed distribution. By incorporating both**[**Skewness and Kurtosis**](https://www.analyticsvidhya.com/blog/2022/09/importance-of-skewness-kurtosis-co-efficient-of-variation/)**into our analysis, we gain a more comprehensive understanding of the shape and characteristics of the data.**

**Skewness indicates the degree of tilt in data, whether it leans towards the left or right, exposing any asymmetry present. A positive skew indicates a tail extending towards the right, whereas a negative skew leans in the opposite direction.**

**Kurtosis, on the other hand, focuses on the distribution’s peaks and tails.**

**Skewed data may cause the tail region to act as an outlier for the statistical model, and such outliers can adversely impact the performance of the model, particularly in regression-based models. Some statistical models are robust to outliers like Tree-based models, but it will limit the possibility of trying other models. So there is a necessity to transform the skewed data to be close enough to a Normal distribution.**

**What is Kurtosis and its Types?**

While skewness focuses on the spread (tails) of normal distribution, **kurtosis** focuses more on the height.**It tells us how peaked or flat our normal (or normal-like) distribution is**. The term, which means curved or arched from Greek, was first coined by, unsurprisingly, from the British mathematician Karl Pearson (he spent his life studying probability distributions).

High kurtosis indicates:

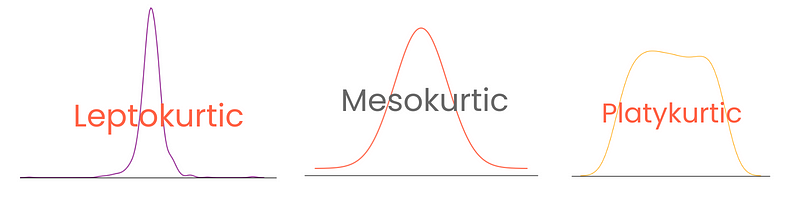
* Sharp peakedness in the distribution’s center.
* More values concentrated around the mean than normal distribution.
* Heavier tails because of a higher concentration of extreme values or outliers in tails.
* Greater likelihood of extreme events.

On the other hand, low kurtosis indicates:

* Flat peak.
* Fewer values concentrated around the mean but still more than normal distribution.
* Lighter tails.
* Lower likelihood of extreme events.

Depending on the degree, distributions have three types of kurtosis:

1. **Mesokurtic distribution** (kurtosis = 3, excess kurtosis = 0): perfect normal distribution or very close to it.
2. **Leptokurtic distribution** (kurtosis > 3, excess kurtosis > 0): sharp peak, heavy tails
3. **Platykurtic distribution** (kurtosis < 3, excess kurtosis < 0): flat peak, light tails



Note that here, **excess kurtosis** is defined as kurtosis - 3, treating the kurtosis of normal distribution as 0. This way, kurtosis scores are more interpretable.

[Understanding Skewness And Kurtosis And How to Plot Them | DataCamp](https://www.datacamp.com/tutorial/understanding-skewness-and-kurtosis)