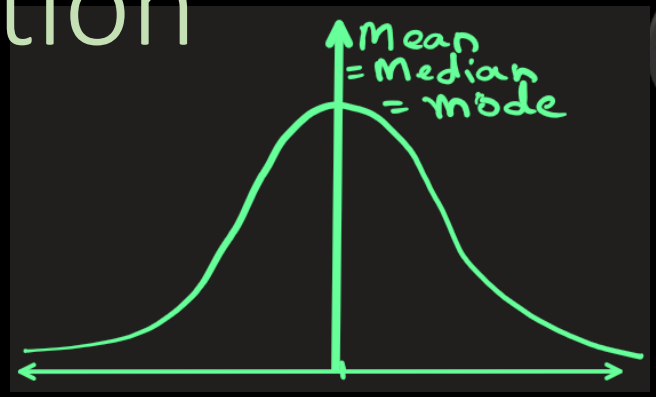
Skewness

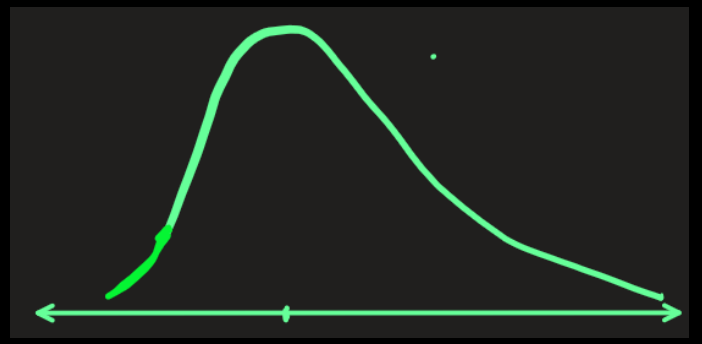
What is skewness ?
What are some examples of skewness ?
How do we measure skewness ?
Why do we care about skewness ?

Skewness: Definition

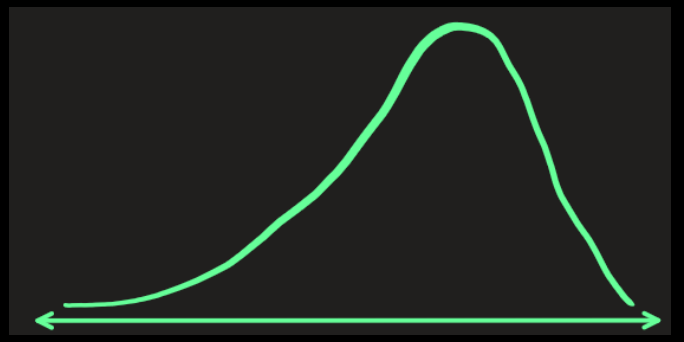
A data with symmetrical distribution has mean = mode= median.
Another name for this is normal distribution.



But sometimes the data is not symmetrical as shown in figure:
The curve can be skewed – either to the right or to the left.

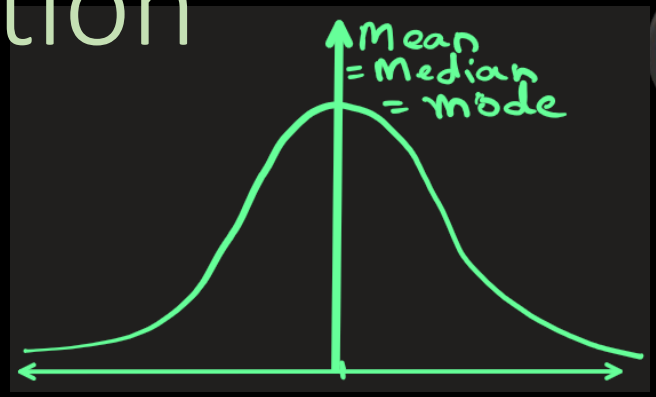


A **skewness** measures how much the data is far from being symmetrical



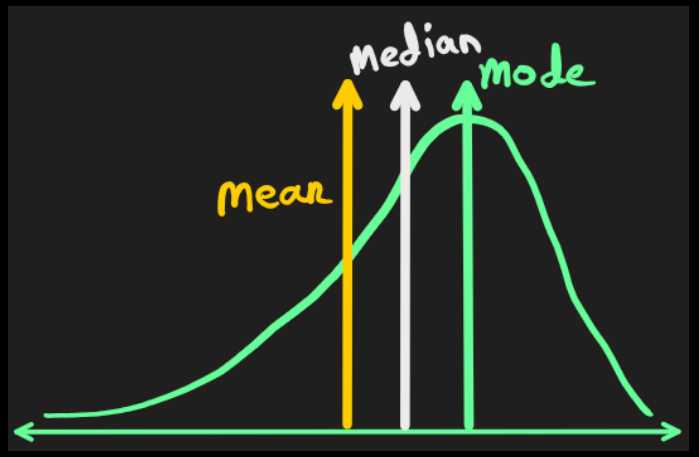
Skewness: Definition

A **skewness of zero** suggests a symmetrical distribution.
Here **mean = median = mode**



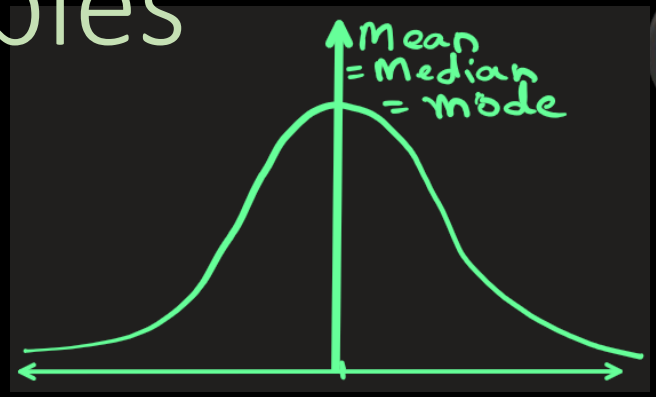
A **positive skewness (right skew)** indicates that the tail of the distribution extends further to the **right**, with more data points clustered on the **left side**. It also indicates presence of outliers towards right side
Here **mean > median > mode**

A **negative skewness (left skew)** indicates that the tail extends further to the **left**, with more data points clustered on the **right side**. It also indicates presence of outliers towards left side
Here **mean < median < mode**



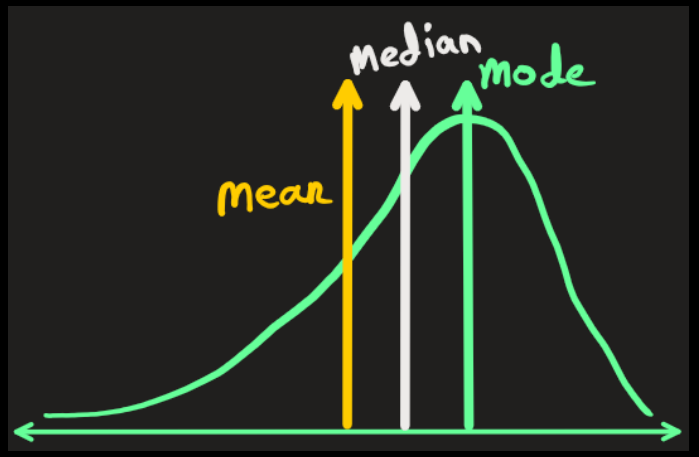
Skewness: Examples

Example of 0 skewedness:
The height distribution in a most countries follows symmetrical distribution.



Examples of right skewedness :
1) Income Distribution: A large portion of the population earns moderate incomes, while a smaller number of individuals earn extremely high incomes, creating a right-skewed distribution.
2) Test Scores: In education, most students might score around the average, but a few exceptional students may achieve very high scores, resulting in a right-skewed distribution of test scores.

Examples of left skewedness :
1) Death from natural cause: Most of death from natural causes (heart disease, cancer, etc.) happen at older ages, with fewer cases happening at younger ages.
2) Scores on an easy test: If students take test that is easy, then most will score high marks, with very few scoring low marks.



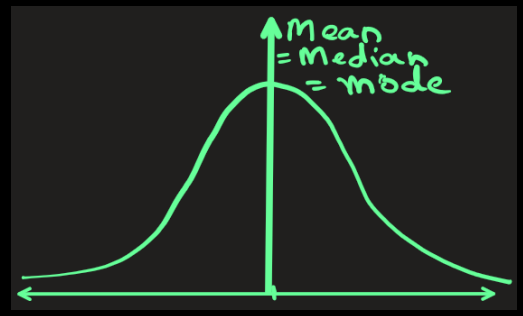
How Do We Measure Skewness?

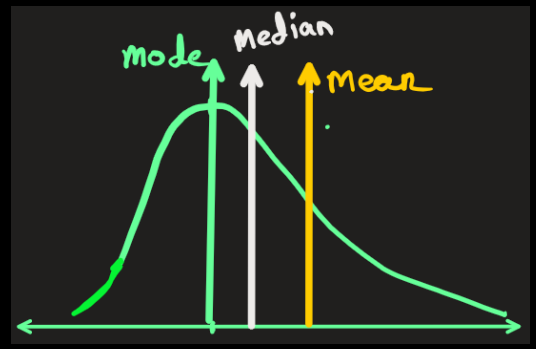
There are multiple formulas, but one common approach is **Pearson's Moment Coefficient of Skewness**:

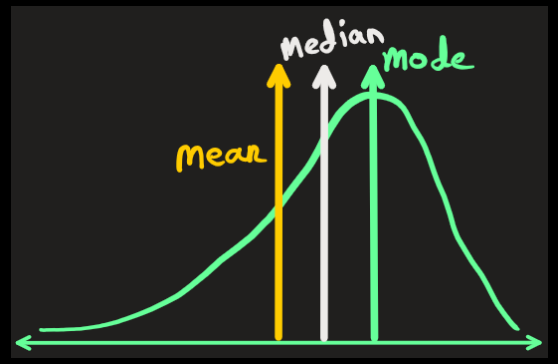
$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3/n}{(\sum(x_i - \bar{x})^2/n)^{3/2}}$$

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

- x_i = each data point
- $\bar{x}$ = sample mean
- s = sample standard deviation
- n = number of observations



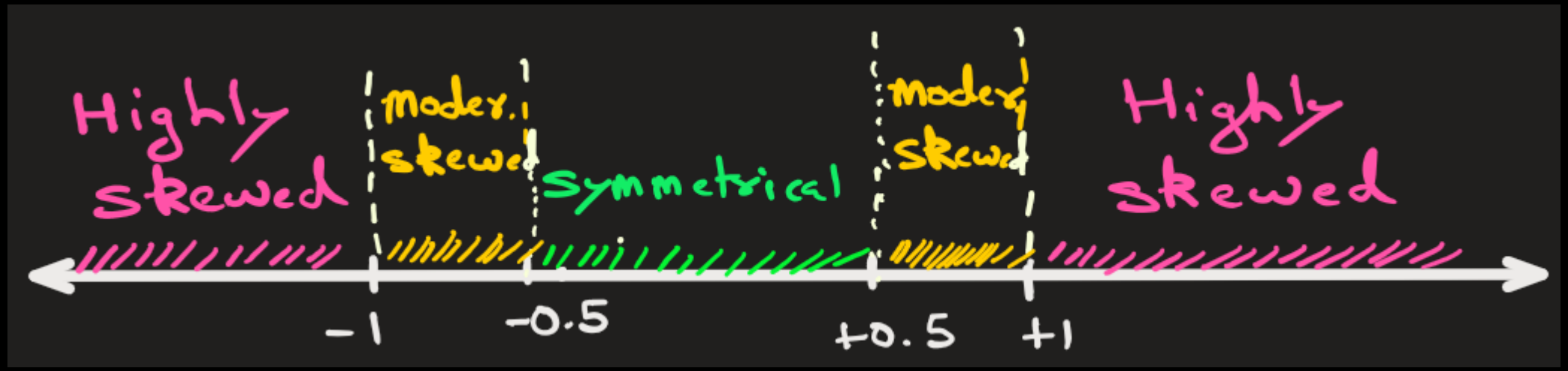




Skewness meaning:

While there isn't a single universally agreed-upon scale, here's a general guideline:

- **-0.5 to 0.5**: The distribution is considered approximately **symmetrical**.
- **-1 to -0.5 or 0.5 to 1**: The distribution is **moderately skewed**.
- **Values beyond -1 and 1**: The distribution is considered **highly skewed**.



Why Skewness matters

1. Helps Detect Non-Normality

Many algorithms (like Linear Regression, Logistic Regression, PCA) assume the data—or residuals—are normally distributed.
If data is highly skewed, this assumption is violated, leading to **biased model** parameters and hence poor predictions

2. Impacts Feature Scaling

Skewed features can distort mean and standard deviation, which are used in StandardScaler or Z-score normalization.
Solution: Apply log, square-root, or Box–Cox transformations to reduce skewness.

3. Improves Model Performance

Models like tree-based ones (Decision Tree, Random Forest, XGBoost) are less sensitive, but linear or distance-based models (KNN, SVM) can perform better after skewness correction.

STOP

Problem Statement: Predicting House Prices — When the Model Gets "Fooled" by Skewness

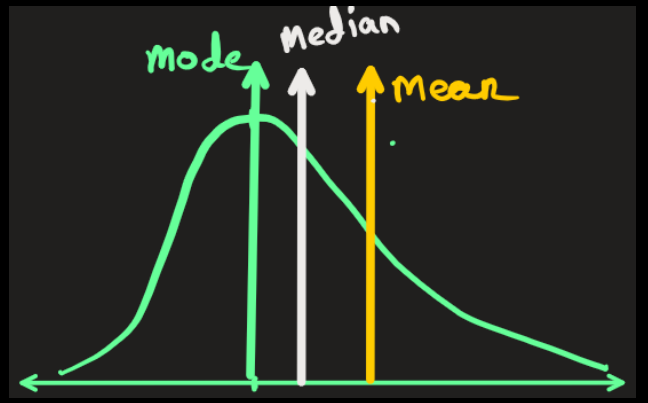
You are building a **machine learning model** to predict **house prices** in a city.
Your dataset has 10,000 houses with various features — area, number of rooms, location, etc.
The **target variable** is the *price of the house*.
Before modeling, you plot the prices and notice something interesting:
Most houses are between ₹30–₹80 lakhs
A few luxury villas cost ₹5–₹10 crores
The **average price** is ₹1.2 crore — but that's *not* what most houses cost.

The Problem

You train a **Linear Regression model** directly on the raw price data. The result you obtain is

- Training error: ✅ Low
- Validation error: ❌ Very high

The model severely **overestimates** the prices of mid-range homes and **underestimates** luxury homes.
You check the residuals (errors) and see they are **not symmetric** — they have a **long right tail**.
In other words, **the target variable is positively skewed.**



Solution Using Skewness

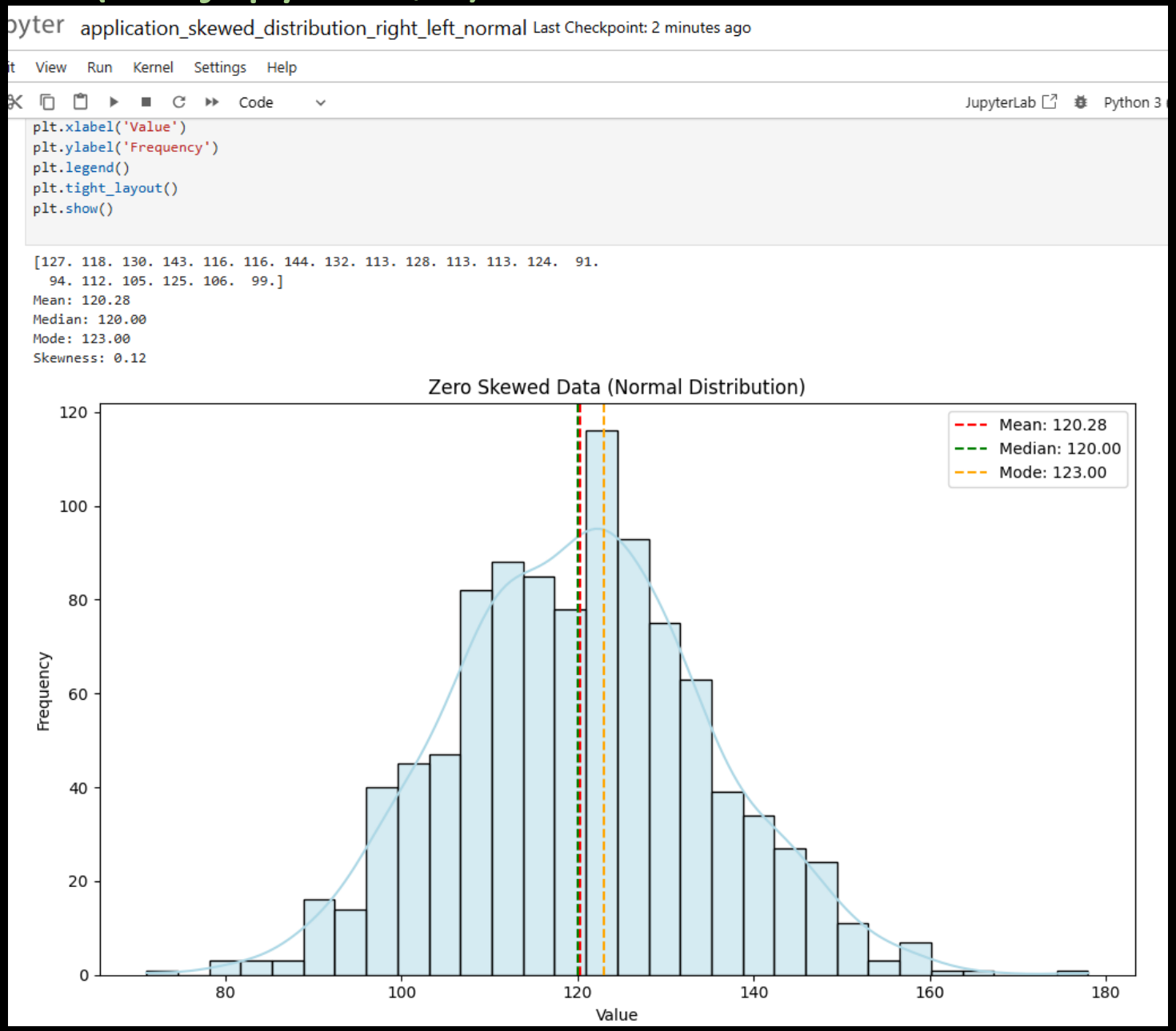
To fix this, you apply a **log transformation** to the target variable: $y' = \log(y)$
Then you retrain the model and the result is:

- Residuals become more **symmetric (skewness ≈ 0)**
- Model predictions improve dramatically
- Validation RMSE drops by 30–40%

After prediction, you simply reverse the transformation using:

$$\hat{y} = e^{\hat{y}'}$$

Example of 0 skewed data (see jupyter n/b):



Example of right skewed data (see jupyter n/b):

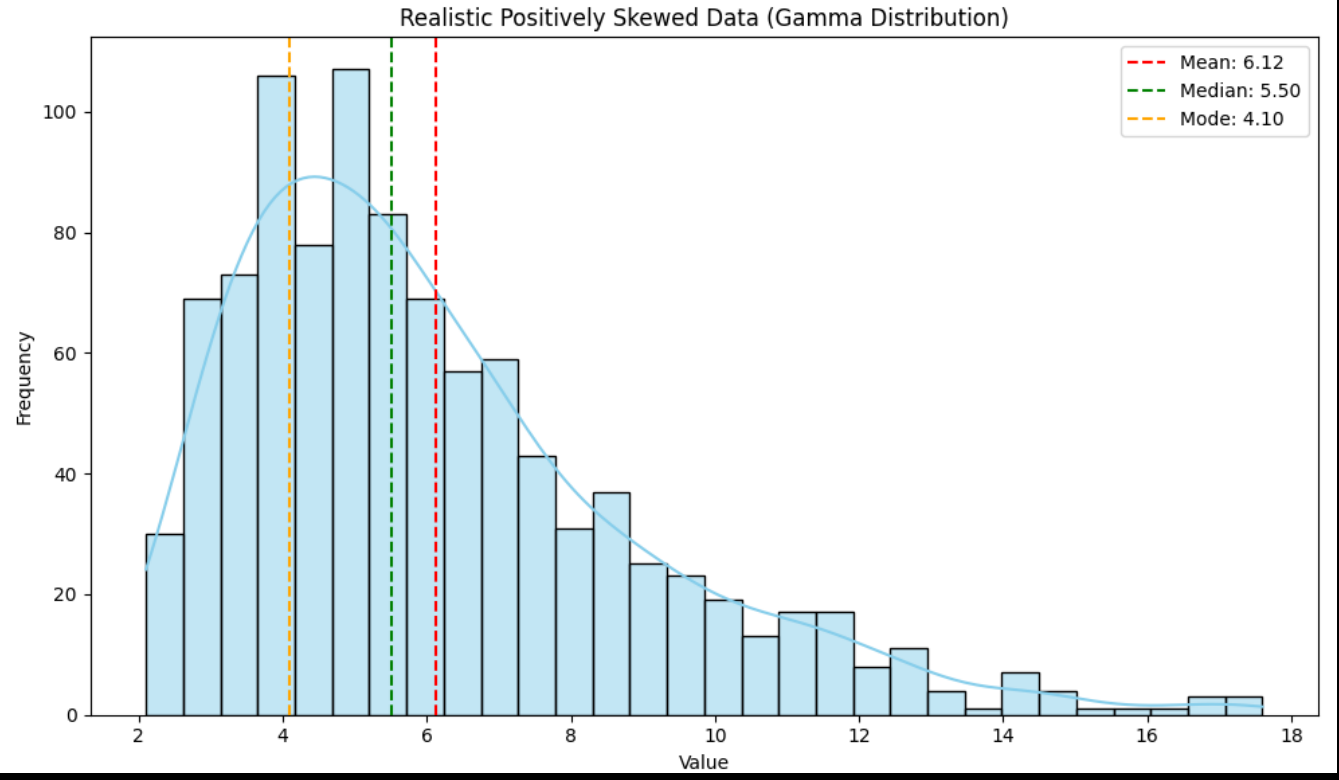
yter application_skewed_distribution_right_left_normal Last Checkpoint: 3 minutes ago

View Run Kernel Settings Help

Code JupyterLab Python

```
plt.title('Realistic Positively Skewed Data (Gamma Distribution)')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.tight_layout()
plt.show()
```

```
[ 6.8  5.   4.8  4.8 11.3  7.7  4.3  6.9  6.   2.4  3.3  6.2 10.7  4.8
  4.1  5.6  4.   4.6  5.3  3.3]
Mean: 6.12
Median: 5.50
Mode: 4.10
Skewness: 1.20
```



Example of left skewed data (see jupyter n/b):

yter application_skewed_distribution_right_left_normal Last Checkpoint: 4 minutes ago

View Run Kernel Settings Help

Markdown JupyterLab Python 3

```
plt.show()
```

```
[71. 82. 83. 83. 43. 65. 87. 70. 76. 98. 92. 74. 46. 83. 87. 78. 88. 84.
 80. 93.]
Mean: 75.02
Median: 79.00
Mode: 89.00
Skewness: -1.20
```