Understanding Z-Score Normalization

Z-score normalization, also known as standardization, is a technique used to transform data into a distribution with a mean of 0 and a standard deviation of 1. This method is particularly useful when comparing features that might have different units or scales. Below is a detailed explanation of how Z-score normalization works, illustrated with an example using the Titanic dataset.

# 1. Z-Score Formula

The Z-score for a data point x is calculated using the formula:  
  
z = (x - μ) / σ  
  
Where:  
- x is an individual data value.  
- μ is the mean (average) of the dataset.  
- σ is the standard deviation of the dataset.  
- z is the Z-score (normalized value).

# 2. Why is the Mean of Z-Scores 0?

When you calculate the Z-score for every data point in a dataset, you subtract the mean (μ) of the original data from each value (x). This operation shifts the entire dataset so that its new mean is 0.  
Here’s why:  
- Original Data: The mean (μ) is the central point of the data. When you subtract μ from each value, the new set of values (before dividing by σ) will have a mean of 0 because all values are centered around 0.  
- Z-Scores: After this subtraction, dividing by the standard deviation (σ) scales these centered values but doesn't change their mean. The result is that the mean of the Z-scores remains 0.

# 3. How Are Values Scaled by Standard Deviation?

After subtracting the mean, the next step in calculating the Z-score is dividing each value by the standard deviation (σ).  
  
- Standard Deviation (σ) measures how spread out the numbers in the dataset are around the mean. A larger σ indicates that the data points are more spread out.  
- Scaling: When you divide by σ, each data point is scaled relative to the spread of the data:  
 - If the original value is close to the mean, its Z-score will be close to 0.  
 - If the original value is far from the mean, its Z-score will be a larger positive or negative number, depending on whether the value is above or below the mean.

# 4. Interpretation of Z-Scores

- Z-Score of 0: This corresponds to a data value equal to the mean.  
- Positive Z-Score: Indicates a data value above the mean.  
- Negative Z-Score: Indicates a data value below the mean.  
- Magnitude of Z-Score: Indicates how many standard deviations a data point is from the mean.

# 5. Example with the Titanic Dataset

Suppose the mean age (μ) is 30 years, and the standard deviation (σ) is 12 years (these are hypothetical values).  
  
For an age of 22 years, the Z-score would be:  
z = (22 - 30) / 12 = -0.67  
  
This means that 22 is 0.67 standard deviations below the mean age.

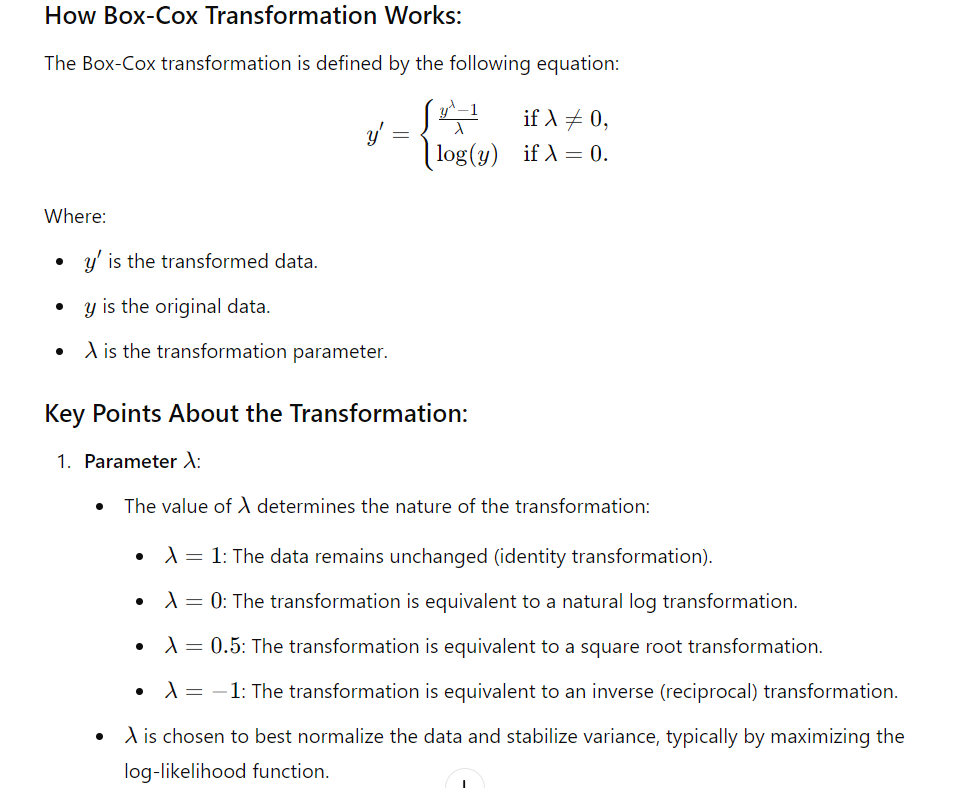
# 7. Conclusion

In summary, the mean of the Z-score normalized values is 0 because each original value has been centered by subtracting the mean. The values are scaled by the standard deviation, resulting in a normalized distribution where the magnitude of the Z-score reflects how far each value is from the mean in units of standard deviation.

Box-Cox transformation

The Box-Cox transformation is a statistical technique used to stabilize variance, make the data more normally distributed, and improve the validity of inferential statistics such as correlation and regression. Unlike simple log transformation, the Box-Cox transformation can handle a broader range of data types and offers a family of power transformations, allowing for more flexibility in transforming data.

**How Box-Cox Transformation Works:**





1. **Choosing λ**:
   * The optimal value of λ is usually determined by maximum likelihood estimation (MLE). This means finding the λ value that makes the transformed data as close to normally distributed as possible.
2. **Handling Negative Values**:
   * The Box-Cox transformation is defined only for positive values. If the data contains zero or negative values, a constant must be added to shift the data before applying the transformation.
3. **Inverse Transformation**:
   * The Box-Cox transformation is reversible, meaning you can transform the data back to its original scale using the inverse of the transformation.

**Benefits of the Box-Cox Transformation:**

* **Flexibility**: It accommodates a variety of data distributions by selecting the best λ.
* **Normality**: It can help in making the data more normally distributed, which is essential for many statistical techniques.
* **Variance Stabilization**: It reduces heteroscedasticity (variance that changes with the mean).

**Limitations:**

* **Positive Values Only**: The data must be strictly positive. If your data includes zeros or negative values, you need to shift them.
* **Complexity**: Determining and applying the optimal λ is more complex than simple transformations like log or square root.

**Practical Application:**

Box-Cox transformation is particularly useful in regression analysis, ANOVA, and other inferential statistics where the assumptions of normality and homoscedasticity (constant variance) are crucial.

**Output:**

1. **Optimal Lambda Value**: The output will include the optimal λvalue that was found using the data.
2. **Transformed Data**: The DataFrame df will contain both the original values and the Box-Cox transformed values.