## Unit II

## 8. Converting to a different scale:

Converting data to a different scale in data analytics, also known as data scaling or data normalization, involves transforming the numerical values of a variable to a specific range or distribution. This process is commonly used to ensure that variables are on a comparable scale, improve the performance of certain algorithms, and make data more amenable to analysis. There are several scaling techniques used in data analytics:

1. Min-Max Scaling (Normalization):
    - Min-Max scaling scales the data to a specific range, typically [0, 1]. It transforms each data point 'x' to a new value 'x_scaled' using the following formula:

x_scaled = (x - min) / (max - min)

    - This method preserves the relative relationships between data points and is useful when the distribution of the data is roughly uniform or when a specific range is required for a particular algorithm.
2. Z-Score Standardization:
    - Z-score standardization scales the data to have a mean of 0 and a standard deviation of 1. It transforms each data point 'x' to a new value 'x_standardized' using the following formula:

x_standardized = (x - mean) / standard_deviation

    - This method is useful when the data follows a roughly normal distribution, and it allows comparisons across different variables with different scales.
3. Robust Scaling:
    - Robust scaling is similar to Z-score standardization but uses the median and the interquartile range (IQR) instead of the mean and standard deviation. It is less sensitive to outliers, making it a suitable choice when dealing with datasets containing extreme values.
4. Log Transformation:
    - Logarithmic transformation is used to stabilize the variance of data with a skewed distribution. Taking the logarithm of the data can convert it from multiplicative relationships to additive ones, making it more amenable to linear modeling.
5. Power Transformation:
    - Power transformation, like the Box-Cox transformation, is used to adjust the data's distribution by applying a power function. It helps to make the data more closely resemble a normal distribution and improve the performance of certain statistical models that assume normality.

6. Binning:
    - Binning involves dividing numerical data into discrete intervals or bins. It simplifies the data and can be helpful when dealing with continuous variables in classification or when analyzing patterns within specific ranges.