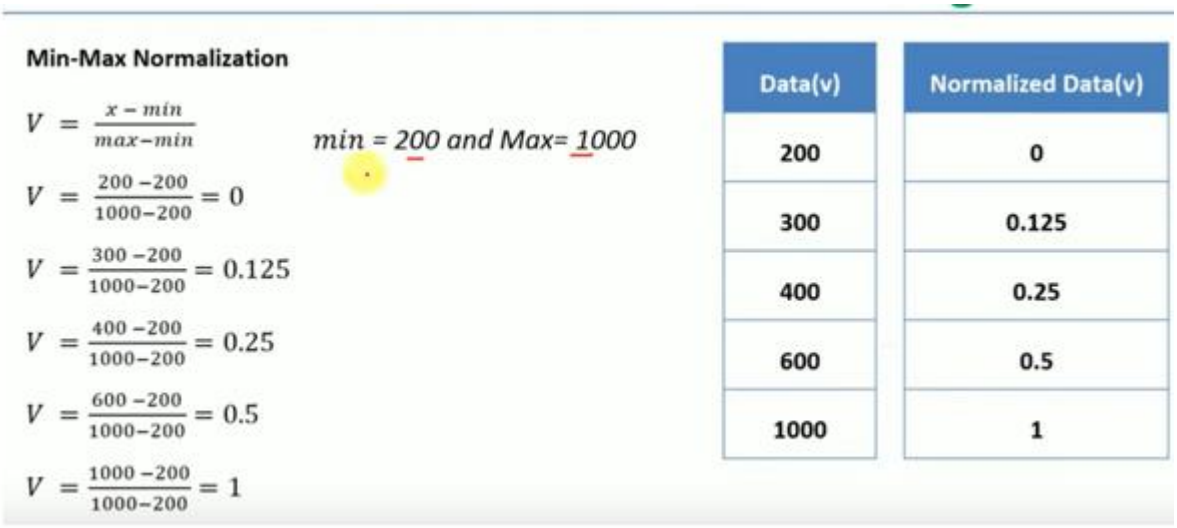
Data normalization or scaling

Data normalization is a process of transforming data into a consistent and comparable format, which can improve data quality, analysis, and integration.

Data scaling and normalization are two important processes that data scientists use to ensure that their data is ready for analysis. Scaling is the process of changing the range of data so that it is within a smaller range, such as from 0 to 1.

Min-max normalization

Min-max normalization is a simple technique that rescales the data values to a range between 0 and 1, using the minimum and maximum values of the original data. This technique preserves the relative order and distance of the data points, but it also reduces the variance and magnifies the effect of outliers. Min-max normalization is useful when the data has a fixed range, such as grades or percentages, but it can distort the data if there are extreme values or different scales.



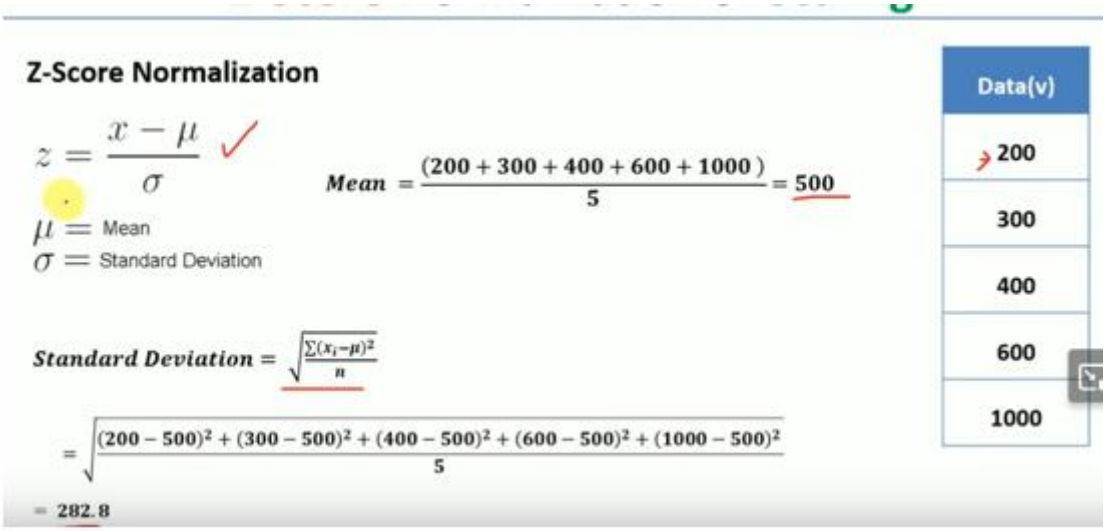
Data(v)	Normalized Data(v)
200	0
300	0.125
400	0.25
600	0.5
1000	1

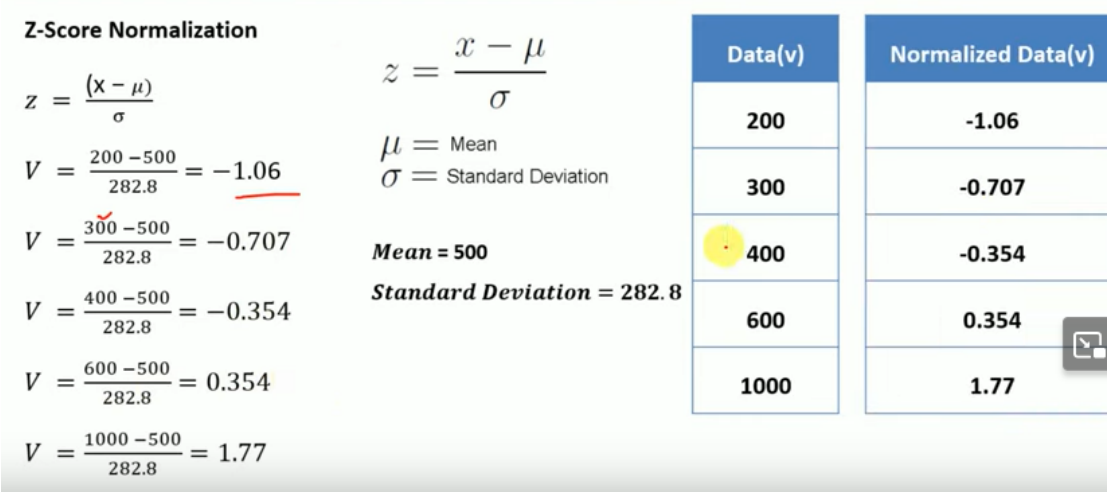
Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1. This technique **preserves the relative order and distance of the data points, but it also reduces the variance and magnifies the effect of outliers**.

Z-score normalization

Z-score normalization is a technique that standardizes the data values by subtracting the mean and dividing by the standard deviation of the original data. This technique transforms the data into a normal distribution with a mean of 0 and a standard deviation of 1, which makes it easier to compare and analyze different variables. However, z-score normalization also changes the original scale and range of the data, and it can be affected by outliers and skewness. Z-score normalization

is useful when the data has a Gaussian distribution or when the scale and range are not important, but it can be misleading if the data has a different distribution or if the outliers have meaning.





Standardization

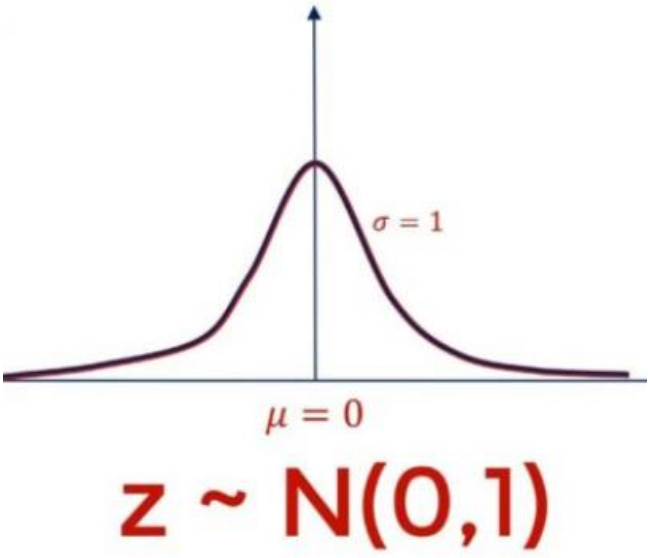
Standardization is another scaling method where the values are centered around mean with a unit standard deviation. It means if we will calculate mean and standard deviation of standard scores it will be 0 and 1 respectively.

$z=x-\mu/\sigma$

where μ= mean of the given distribution, σ = standard deviation of the given distribution

This Z is called standard score and it represents the number of standard deviations above or below the mean that a specific observation falls.

If we plot these standard scores it will be a normal distribution with the mean at 0 and the standard deviation equal to 1.

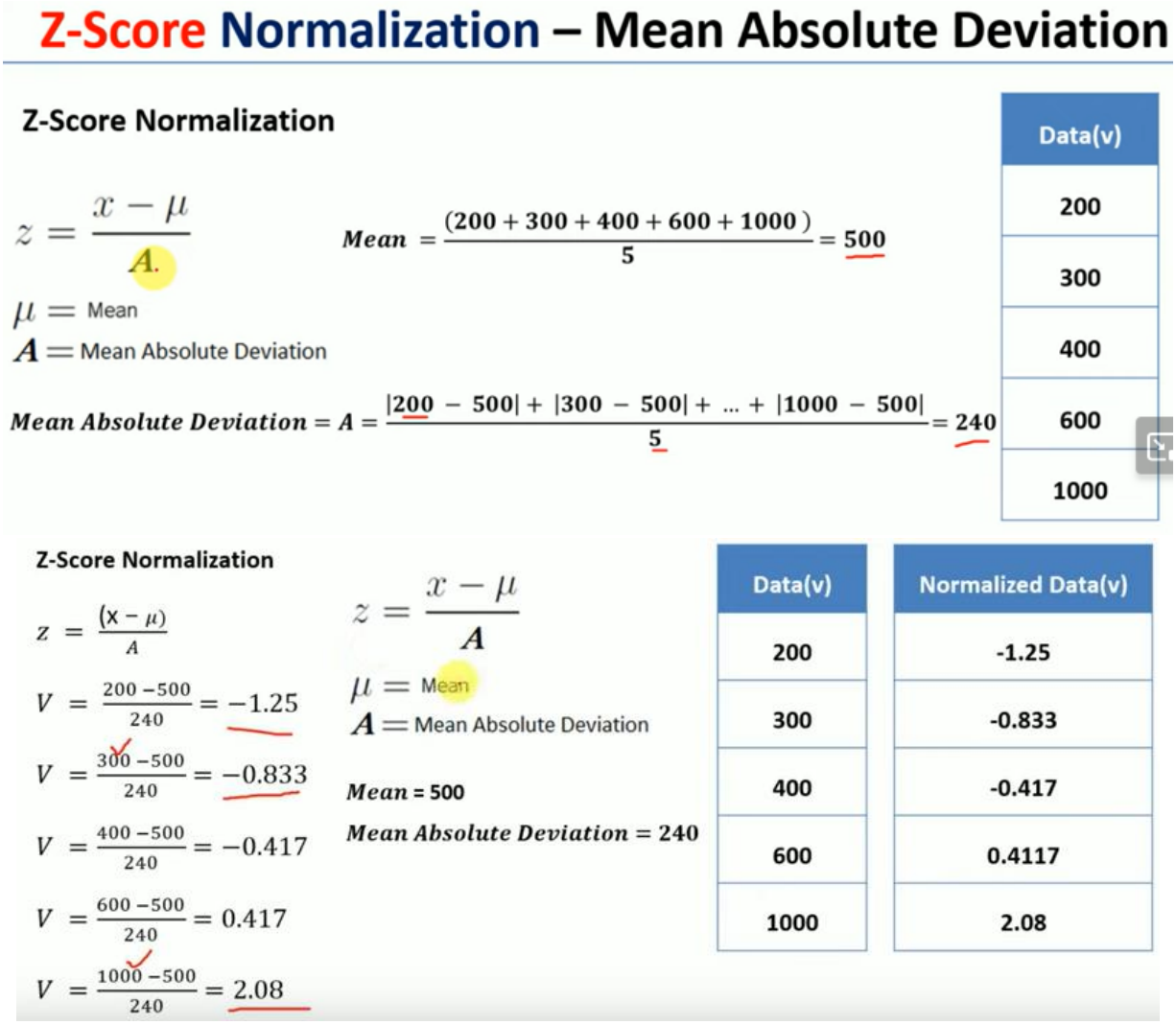


The standard deviation with mean = 0 and σ =1 is also known as standard normal distribution and is denoted by N(0,1).

Let's assume you and your friend study in different universities where the grading system is different. You get your score of 80 in a test. The mean grade of the class is 70 and a standard deviation of 5. Your friend got a grade of 610 and the mean grade of the class is 600 with a standard deviation of 50. How you are going to say who is performing better? As grade 80 can't be compared to 610.

Here comes the role of standardization as it allows us to compare the scores with different metrics directly and make a statement about them.

Your Z score = 80–70/5 = 2
It means you are 2 standard deviations above the average grade.
Your friend's Z score = 610–600/50 = 0.2
It means you are 0.2 standard deviations above the average grade.
By looking at standard scores you can clearly say that you are performing much better that he or she is the class.

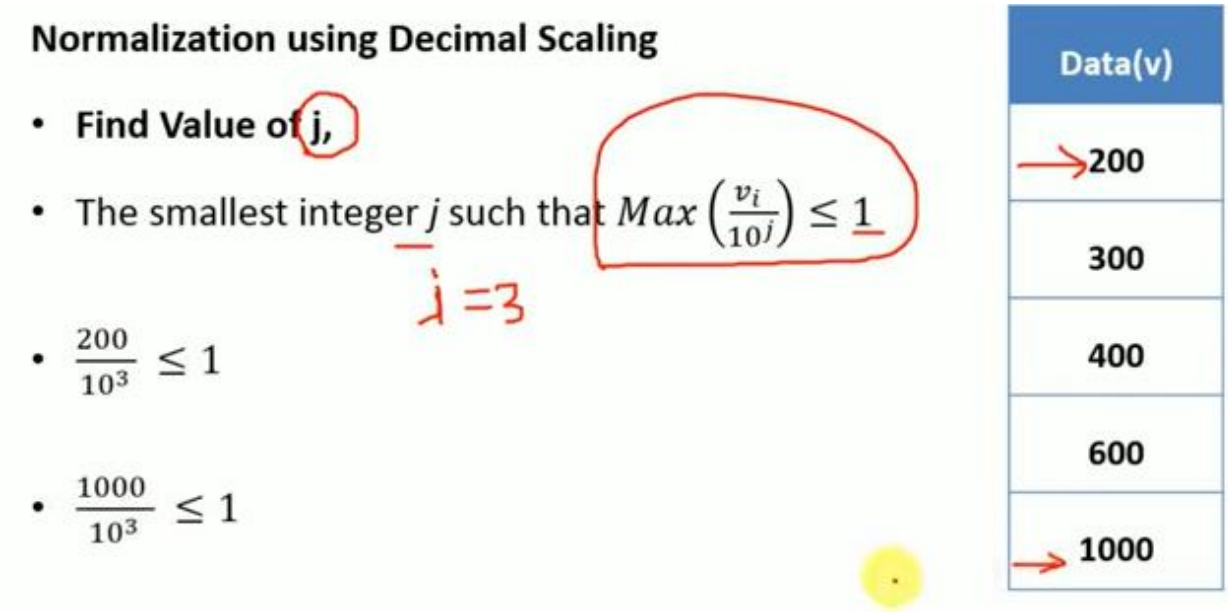


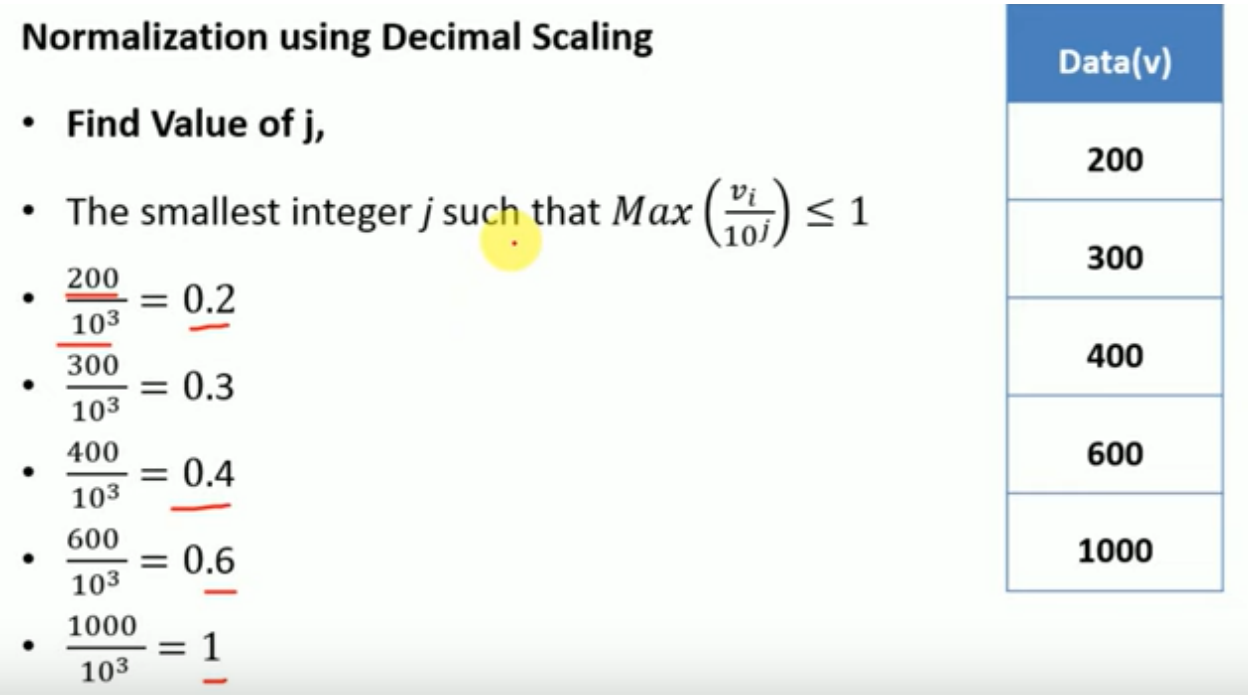
***Z*-scores and the Standard Normal Distribution**

Z-scores are distributed according to the standard normal distribution which has a mean of 0 and a standard deviation of 1. The standard normal distribution can range from
to , but extreme values are highly unlikely. According to the empirical rule, about 68% of all *z*-scores will be between -1 and 1 (standard deviations from mean), 95% will be between -2 and 2, and 99.7% will be between -3 and 3.

Decimal scaling normalization

Decimal scaling normalization is a technique that shifts the decimal point of the data values to reduce their magnitude, using a factor of 10. This technique preserves the relative order and proportion of the data points, but it also changes the scale and range of the data. Decimal scaling normalization is useful when the data has a large range of values or when the magnitude is not important, but it can be problematic if the data has a small range of values or if the scale and range have meaning.





Normalization is preferred over standardization when our data doesn't follow a normal distribution. It can be useful in those machine learning algorithms that do not assume any distribution of data like the k-nearest neighbor and neural networks.

Standardization is good to use when our data follows a normal distribution. It can be used in a machine learning algorithm where we make assumptions about the distribution of data like linear regression etc

Point to be noted that unlike normalization, standardization doesn't have a bounding range i.e. 0 to 1.

Log transformation normalization

Log transformation normalization is a technique that applies a logarithmic function to the data values, which reduces the skewness and compresses the range of the data. This technique makes the data more symmetric and homoscedastic, which can improve the performance of some statistical and machine learning models. However, log transformation normalization also changes the original scale and distribution of the data, and it can create negative values or lose information for zero or small values. Log transformation normalization is useful when the data has a skewed or exponential distribution or when the outliers are not important, but it can be inappropriate if the data has a linear or uniform distribution or if the outliers have meaning.

Log-transform decreases skew in some distributions, especially with large outliers. But, it may not be useful as well if the original distributed is not skewed. Also, **log transform may not be applied to some cases (negative values), but standardization is always applicable (except $\sigma=0$)**.

NB: Homoscedasticity, or homogeneity of variances, is an assumption of equal or similar variances in different groups being compared. This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities.