Feature Scaling

In different columns, the scale of the number could be drastically different.
We scale down all the columns to same scale so that ML algos do not face any problem.

Need:
- To improve the performance of distance based algos (bigger numerical columns dominates more).
- Improve in Optimized technique
- Numerical stability(deep learning)

Univariate scaling:

Standardization:

(X-mean)/SD.
Try to make data mean centred and within SD of 1 without changing covariance.
So if data is positively correlated, it will try to keep it positively correlated.
Therefore not necessary that data would be circular after standardization.
It does not work with negative values.
It will preserve the distribution of given column.
Outliers remain as it is.

Minmax Scaling:

X'i=(Xi-Xmin)/(Xmax-Xmin)

Always results will be between 0 and 1.
Distribution is preserved with distance between relative points.
Highly sensitive to outliers.
Dependency on min and max values.
0 values would be shifted.
It is reversible.

When to use what?(standardization vs normalization)

When features follows normal distribution, standardization could be used.
Algorithms which prefers mean centre data, standardization could be used.
When one knows that data is bounded, (eg. Image based data 0-1) normalization is used.
Working with algos, which does not accept negative values, normalization is used.

Log transformation with minimal scaling could be used when outliers are present.

Robust scaler:

X'i=(Xi-median)/(IQR)

Distribution is preserved.
Used when you have heavy outlier.

Max Absolute Scaler:

When data has lot of zeroes(sparsity) and zero has a meaning so zeros cannot be change to some other value.

X'i= Xi/max(|X|)

Range(-1 to 1) for data being negative as well as positive
Range(0 to 1) for data being 0 to 1.

Preserves sign. Relative distance is preserved. Zero values remain unchanged. Reversible. Impact of outlier remains.

Multivariate Scaling:

L2 normalization:

Row based.
Find Euclidean distance between origin and given points(L2 norm). Divide this distance with those points. Now all the rows are treated as vectors. These vectors have equal distance from origin. If plotted from the circle.

L1 normalization:
Find Manhattan distance, everything else is same. If plotted, diamond is formed.
If data is sparse or if ml algos are ranking based, we L1 normalization.

Used when rows are more important, treated as vectors (eg. Recommender systems)
Distribution will not be same of columns after scaling, irreversible.

Parameter/ Criteria	Standardization	Min-Max Scaling	Robust Scaling	Max Absolute Scaling
Formula	(X - μ) / σ	(X - Min) / (Max - Min)	(X - Median) / IQR	X / MaxAbs(X)
Centering	Centers data around mean (0)	No centering; shifts to start at 0	Centres around median	No centering; retains centre
Scaling	Scales by std deviation	Scales to [0,1] or range	Scales by IQR	Scales relative to max abs value
Sensitivity to Outliers	Yes, affected by mean & std	Yes, affected by min & max	No, uses median & IQR	Yes, depends on max abs
Range of Scaled Data	No fixed range (~[-3,3])	Fixed, [0,1] or [-1,1]	No fixed range, varies with IQR	Typically [-1,1]

Impact on Distribution Shape	Maintains shape	Maintains shape	Maintains shape	Maintains shape
Suitability for Sparse Data	Not suitable	Not suitable (shifts zeros)	Not suitable	Suitable
Preservation of Zero	Shifted unless mean=0	Shifted unless min=0	Shifted unless median=0	Zero preserved
Common Use Cases	Gaussian data, PCA, clustering	Neural networks, image processing	Outliers, robust ML, finance	Text processing, sparse data
Interpretability	Interpretation in std deviations	Direct due to fixed range	Interpretation in median & IQR	Interpretation in proportion to max
Impact on Feature Importance	Equalizes based on variance	Equalizes based on range	Equalizes based on IQR	Normalizes based on max abs value
Algorithmic Suitability	Algorithms assuming Gaussian features	Algorithms sensitive to feature scale	Algorithms robust to outliers	Algorithms preserving magnitude/sign
Invariance to Transformations	Not invariant to multiplicative transforms	Invariant to linear transforms	Not invariant to multiplicative transforms	Invariant to scaling transforms
Mathematical Properties	Linear transformation	Linear transformation	Non-linear if outliers present	Linear transformation
Parameter Dependency	Depends on mean & std	Depends on min & max	Depends on median & IQR	Depends on max abs value
Effect on Data Ordering	Maintains order	Maintains order	Maintains order	Maintains order
Typical Value Range Post-Scaling	Mostly within [-3,3], but outliers possible	[0,1] or user-defined range	Varies; central values ~[-1,1]	[-1,1], but extremes may exceed if outliers
Ease of Reversibility	Easy to reverse with mean & std	Easy to reverse with min & max	Reversible with median & IQR	Easy to reverse with max abs value

Parameter/Criteria	L2 Normalization	L1 Normalization
Formula		

Centering	No centering; does not alter the mean of the data	No centering; does not alter the mean of the data
Scaling	Scales the feature vector so its Euclidean norm (L2 norm) is 1	Scales the feature vector so its Manhattan norm (L1 norm) is 1
Sensitivity to Outliers	Moderately sensitive; outliers affect the norm, thus the scaling	Less sensitive; since it sums absolute values, the impact of outliers is diluted
Range of Scaled Data	No fixed range; the vector's length is 1, but individual elements can vary	No fixed range; elements are scaled relative to the sum of absolute values
Impact on Distribution Shape	Changes the magnitude but not the direction of data points in feature space	Changes the magnitude but not the direction of data points in feature space
Suitability for Sparse Data	Suitable; does not alter zero entries	Suitable; does not alter zero entries

L2 Normalization

$$\frac{X}{\|X\|_2} \text{ where } \|X\|_2 = \sqrt{\sum x_i^2}$$

L1 Normalization

$$\frac{X}{\|X\|_2} \text{ where } \|X\|_2 = \sqrt{\sum x_i^2}$$