Normalizing Data with Pandas



Pratheerth Padman FREELANCE DATA SCIENTIST



Module Summary



What does normalizing data mean?

Why is data normalization important?

Simple feature scaling

Min-max scaling

Z-score normalization



Normalizing Data - What and Why?



Data Normalization

Data normalization is the process of transforming your data by scaling each feature in a given dataset to a particular range, usually from 0 - 1.



Data Normalization

Age	Salary
20	45000
30	250000
40	150000
50	500000

Age	Salary
0.2	0.09
0.3	0.5
0.4	0.3
0.5	1

Before normalization



Why Normalize Data?

Age	Salary
20	45000
30	250000
40	150000
50	500000

Age	Salary
0.2	0.09
0.3	0.5
0.4	0.3
0.5	1

The two features - age and salary are on completely different scales

Say we apply a ML model - linear regression for example

Income will unduly influence the model due to its much higher value

In reality, this may or may not be true

To avoid this, data normalization!



Methods to Normalise Data

Simple feature scaling

Min-max scaling

Z-score scaling



Simple Feature Scaling



Simple Feature Scaling

$$Xnew = \frac{X}{Xmax}$$

Age	Salary
20	45000
30	250000
40	150000
50	500000

Age	Salary
0.4	0.09
0.6	0.5
0.8	0.3
1	1

Formula

Before normalization



Min-Max Scaling



Min-Max Scaling

$$Xnew = \frac{X - Xmin}{Xmax - Xmin}$$

Age	Salary
20	45000
30	250000
40	150000
50	500000

Age	Salary
0	0
0.33	0.45
0.66	0.23
1	1

Formula

Before normalization



Z-Score Normalization



What is Z-Score?



Number of standard deviations from the mean, a given data point is



Also known as the standard score



Ranges from -3 to + 3



Z-Score Normalization

$$Xnew = \frac{X - \mu}{\sigma}$$

Age	Salary
20	45000
30	250000
40	150000
50	500000

Age	Salary
-1.34	-1.13
-0.44	0.08
0.44	-0.511
1.34	1.56

Formula

Before normalization



Points to Consider for Data Normalization



Points to Consider for Data Normalization



Data normalization or feature scaling is not always required



Should be used when applied model uses distance calculations like KNN's, Linear regression etc.



Naïve Bayes, Decision trees etc. do not require data normalization



Technique to use depends on use case



Summary



Data normalization - transforming features to a common range

Required to avoid any feature to unduly influence the result

Simple feature scaling

Min Max scaling

Z-score scaling

