Data Transformation Techniques

A video lesson on various methods to preprocess and enhance data for analysis and model performance.

Logarithmic Transformation

01 Improves performance of models that assume normally distributed data

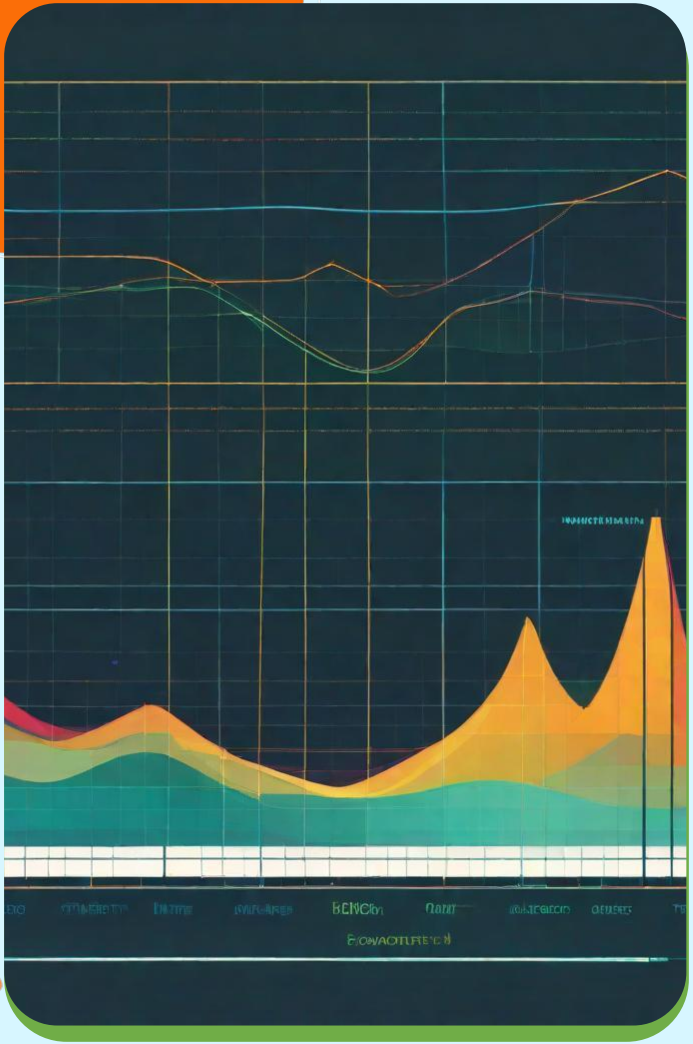
02 Useful for skewed or long-tailed data

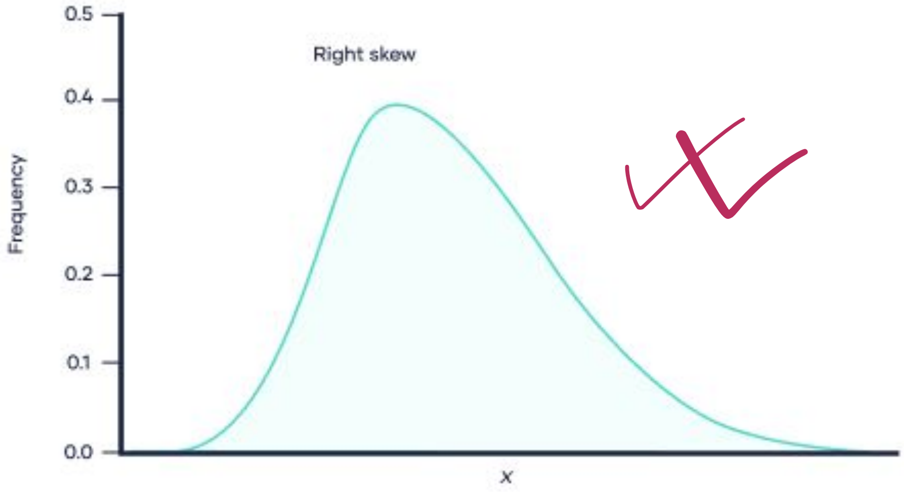
03 Makes distribution more symmetrical

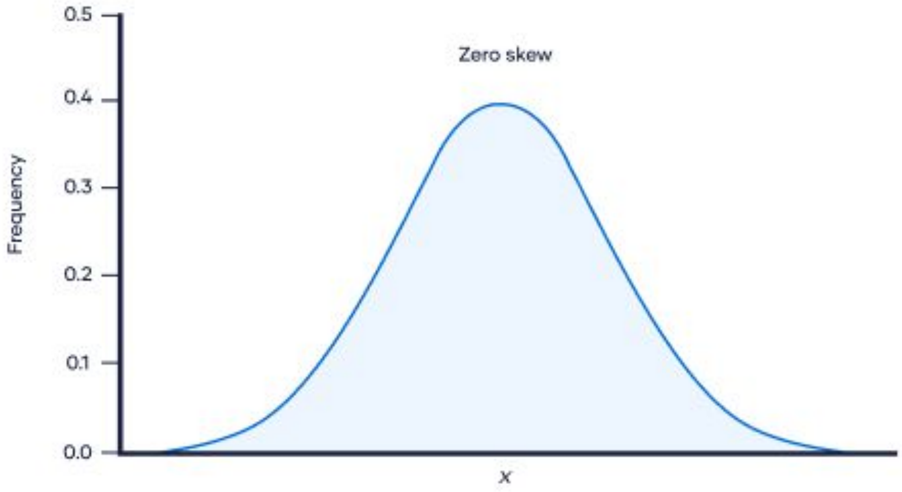
04 Example: Normalizing stock prices in financial data analysis

Box-Cox Transformation

- Generalization of logarithmic transformation
- Adjusts skewness of the data
- Effective for data that doesn't fit well with simple logarithmic transformation
- Commonly used in economics and environmental science




Right skew
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
x


Zero skew
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
x

Binning

01 Simplifies complex data structures and reduces noise

02 Divide continuous variable into intervals or bins

03 Replace values with bin number or representative value

04 Useful in machine learning algorithms that perform better with discretized data

Age	Age Binning	Final Category
63	3	Senior
62	3	Senior
51	3	Senior
60	3	Senior
32	2	Middle-aged
40	2	Middle-aged
57	3	Senior
64	3	Senior
34	2	Middle-aged
20	1	Young
63	3	Senior
19	1	Young
44	2	Middle-aged
55	3	Senior
41	2	Middle-aged
51	3	Senior
23	1	Young
60	3	Senior

One-Hot Encoding

01 Essential for algorithms that can't handle categorical data directly

02 Convert categorical variables to binary format

03 Each category becomes a separate binary feature

04 Example: Representing presence or absence of words in natural language processing

Social Class	Encoded
Rich	3
Rich	3
Rich	3
Rich	3
Middle class	2
Poor	1
Middle class	2
Rich	3
Rich	3
Rich	3
Rich	3
Middle class	2
Rich	3
Poor	1
Poor	1
Rich	3
Middle class	2
Middle class	2

Creating Dummies

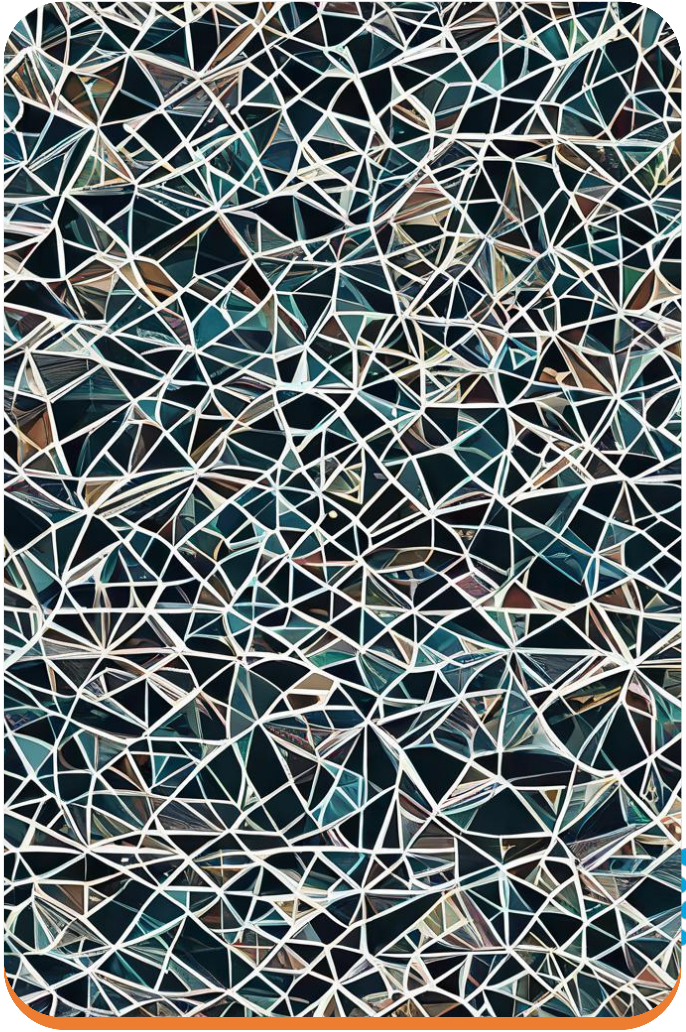
01 Effective in regression analysis with categorical predictors

02 Convert categorical variables to binary columns

03 1 indicates presence of category, 0 otherwise

04 Incorporate categorical data into models

City	city_newyork	city_mumbai	city_dhaka
Mumbai	0	1	0
Mumbai	0	1	0
Mumbai	0	1	0
New York	1	0	0
Dhaka	0	0	1
Mumbai	0	1	0
New York	1	0	0
New York	1	0	0
Dhaka	0	0	1
Mumbai	0	1	0
Dhaka	0	0	1
Dhaka	0	0	1
New York	1	0	0
Dhaka	0	0	1
Dhaka	0	0	1
Mumbai	0	1	0
New York	1	0	0
New York	1	0	0
Dhaka	0	0	1



Creating New Features

- Generate additional features from existing ones
- Capture complex patterns in the data
- Include feature scaling, polynomial features, or interaction terms
- Example: Extracting texture, color, or shape information from image data

Revenue	Cost	Profit (Revenue - Cost)
$45,95,88,246	$6,82,821	$45,89,05,425
$42,90,33,203	$3,29,177	$42,87,04,026
$22,36,47,334	$5,49,877	$22,30,97,457
$37,30,16,161	$7,36,802	$37,22,79,359
$60,99,36,002	$6,36,024	$60,92,99,978
$44,60,28,190	$6,73,985	$44,53,54,205
$23,26,00,839	$98,584	$23,25,02,255
$25,37,51,892	$1,75,367	$25,35,76,525
$86,21,56,728	$3,66,470	$86,17,90,258
$33,38,10,005	$15,169	$33,37,94,836
$85,91,15,517	$7,13,625	$85,84,01,892
$71,71,11,601	$5,32,559	$71,65,79,042
$5,17,44,907	$1,74,756	$5,15,70,151
$28,21,440	$35,913	$27,85,527
$81,76,96,952	$1,09,997	$81,75,86,955
$22,96,45,365	$3,25,467	$22,93,19,898
$78,53,65,269	$1,05,395	$78,52,59,874
$46,87,84,477	$1,69,370	$46,86,15,107
$76,50,56,461	$3,38,520	$76,47,17,941

Extracting Day, Month & Year

01 Reveal temporal patterns and trends

02 Extract components from date or timestamp variables

03 Useful for time series analysis and seasonal forecasting

04 Example: Extracting month or quarter from retail transaction dates

Date	Day	Month	Year
1/1/22	1	1	2022
2/1/22	2	1	2022
3/1/22	3	1	2022
4/1/22	4	1	2022
5/1/22	5	1	2022
6/1/22	6	1	2022
7/1/22	7	1	2022
8/1/22	8	1	2022
9/1/22	9	1	2022
10/1/22	10	1	2022
11/1/22	11	1	2022
12/1/22	12	1	2022
13/1/22	13	1	2022
14/1/22	14	1	2022
15/1/22	15	1	2022
16/1/22	16	1	2022
17/1/22	17	1	2022
18/1/22	18	1	2022
19/1/22	19	1	2022

Standardization (Z-Score)

01 Useful for algorithms sensitive to variable scale

02 Rescale data to have mean of 0 and standard deviation of 1

03 Comparable across different variables

04 Example: Comparing financial metrics using z-score standardization

Normalization (Min-Max Scale)

01 Example: Normalizing pixel intensity values in image processing

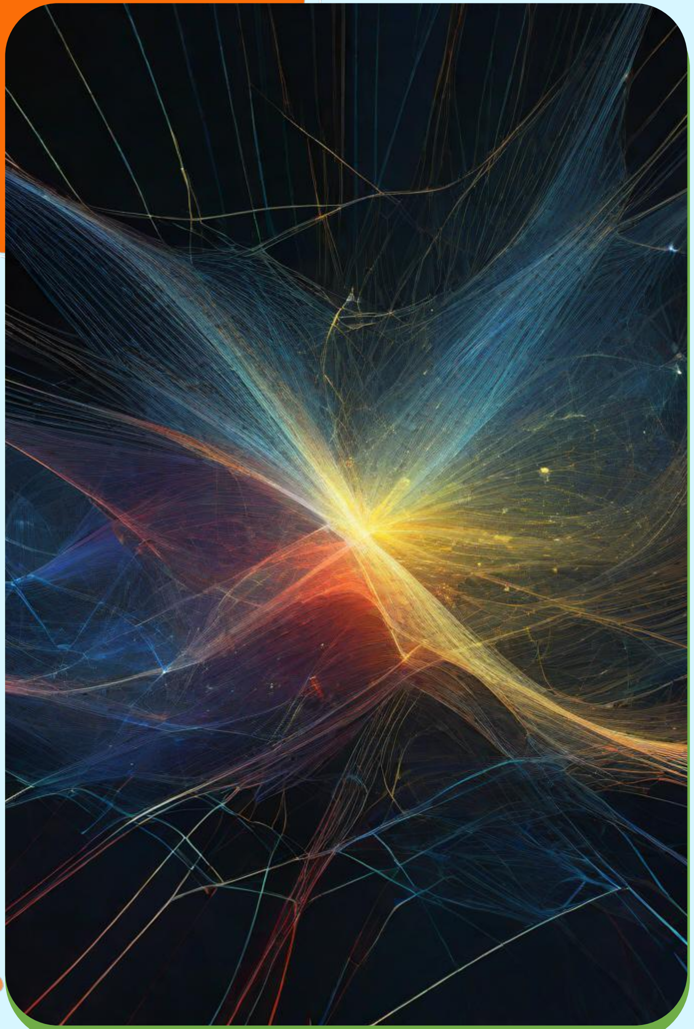
02 Scale data to fixed range, typically 0 to 1

03 Useful for non-Gaussian or non-normal distributions and algorithms expecting normalized data

Age	Income	AgeX	IncomeX
20	18920	0.2922	0.4344
38	20662	0.4894	0.3357
19	33296	0.2801	0.8616
18	18943	0.6806	0.1192
23	18719	0.0169	0.7461
33	24375	0.5094	0.4049
24	37590	0.5820	0.0263
33	26830	0.4237	0.8363
31	20943	0.9401	0.9991
25	21422	0.6790	0.2428
35	18729	0.7807	0.9924
31	18715	0.4867	0.1067
19	26685	0.7670	0.2291
36	22622	0.6743	0.8595
40	24814	0.4923	0.3672
23	30347	0.9634	0.9911
22	22443	0.9319	0.0248
23	33241	0.6804	0.2866
23	24189	0.5017	0.0663

PCA (Principal Component Analysis)

- Dimensionality reduction technique
- Transforms high dimensional data to lower-dimensional space
- Preserves most of the variance in the data
- Useful for reducing computational complexity and visualizing high dimensional data



Num 1	Num 2	Num 3	Num 4	Num 5	PCA
12	19	10	18	17	0.11239
11	18	18	14	12	0.16431
14	10	18	12	10	0.89520
13	11	10	19	13	0.15176
16	16	18	17	14	0.26125
19	12	13	16	12	0.92368
17	12	19	14	16	0.10278
16	13	17	14	10	0.63097
17	19	18	17	11	0.22122
13	15	17	15	10	0.13010
19	11	12	17	19	0.25666
12	10	14	13	17	0.32258
12	15	13	11	17	0.61531
16	16	17	13	11	0.78207
17	10	15	13	17	0.71200
10	12	19	12	15	0.70941
15	11	18	17	14	0.43906
17	10	16	16	12	0.81805
10	11	16	14	14	0.46370

Conclusion

01 Empower data scientists to extract valuable insights from diverse datasets

02 Data transformation techniques are essential for analysis and modeling

03 Enhance symmetry, handle categorical variables, and reduce dimensionality

Thank you for your time and attention 🙂