DATA ANALYSIS

DATA TRANSFORMATIONS

COMMON TRANSFORMATIONS

Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

If the raw data does not meet the requirements, we can either:

- abandon the model
- attempt to **transform** the data

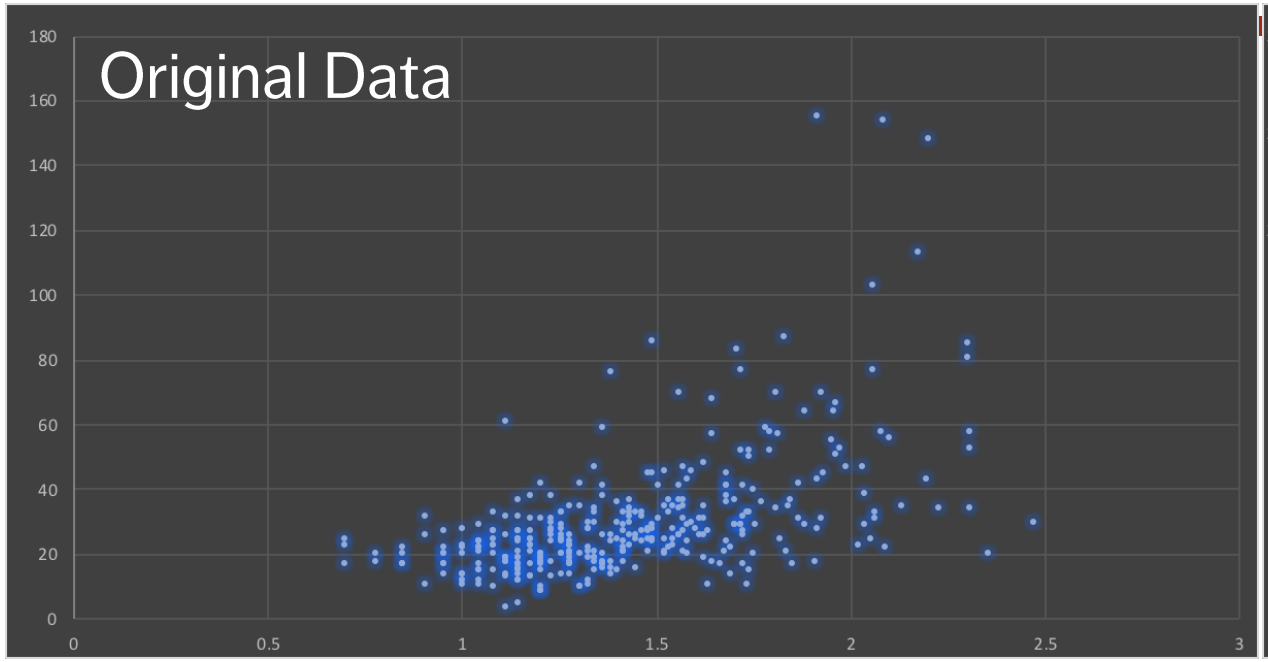
The second approach requires an **inverse transformation** to be able to draw conclusions about the **original data**.

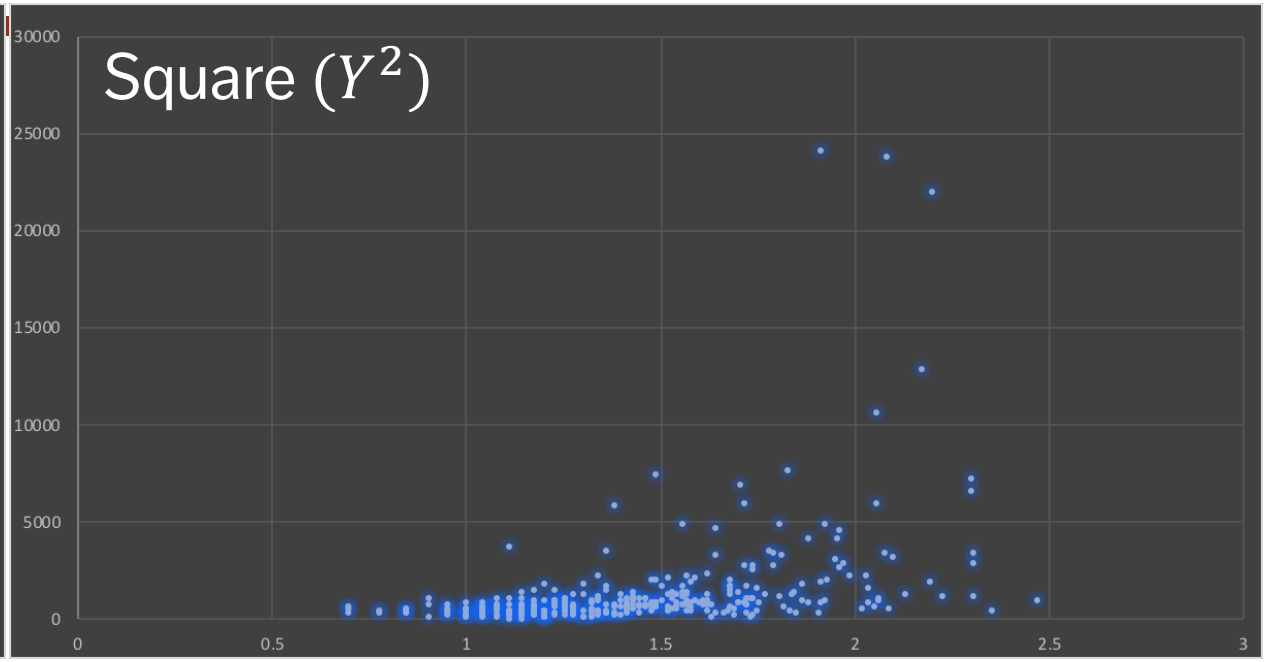
COMMON TRANSFORMATIONS

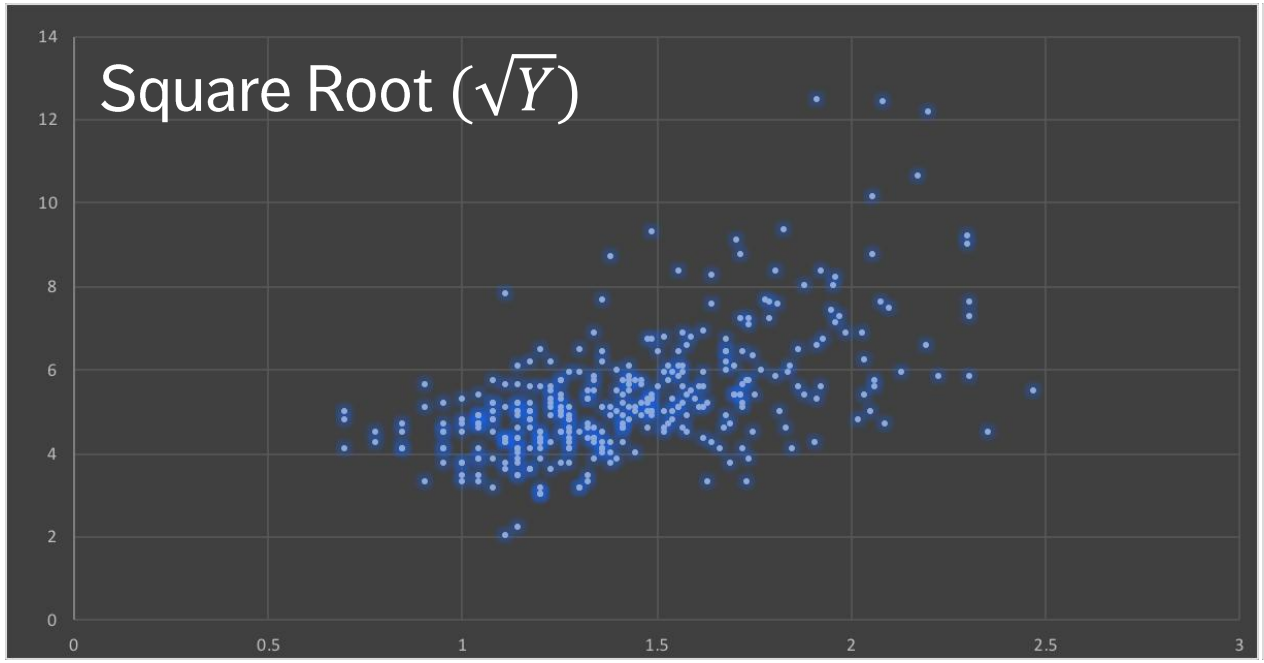
In the data analysis context, transformations are **monotonic:**

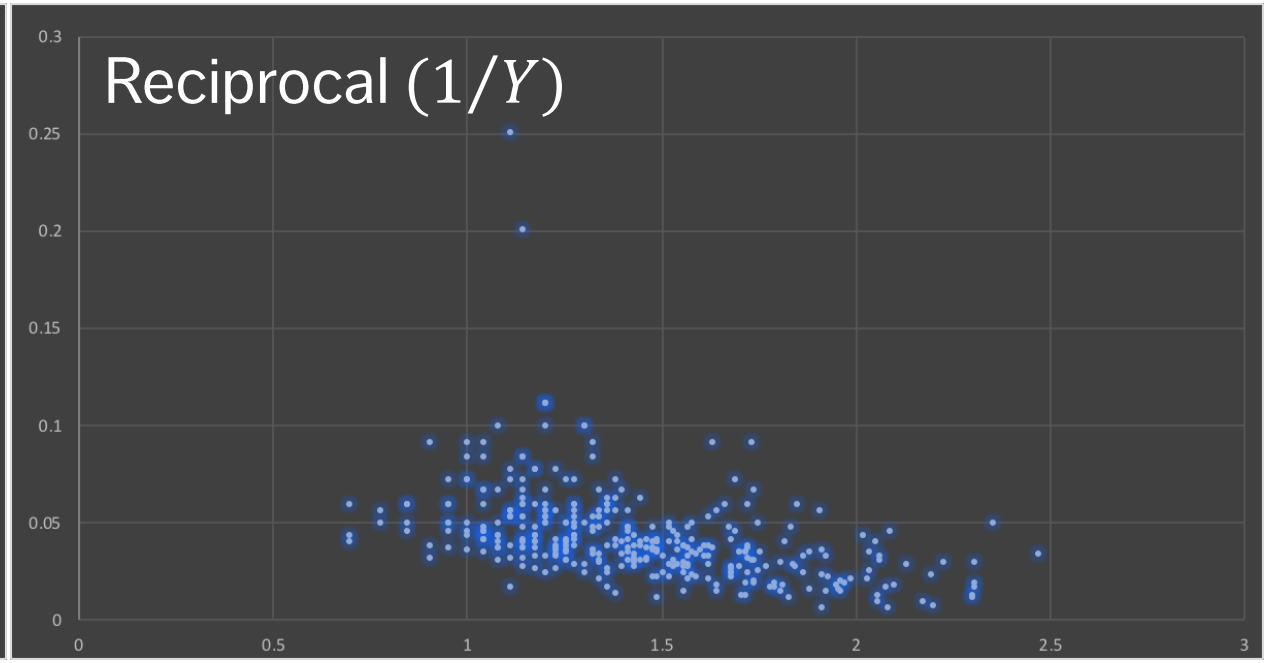
- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

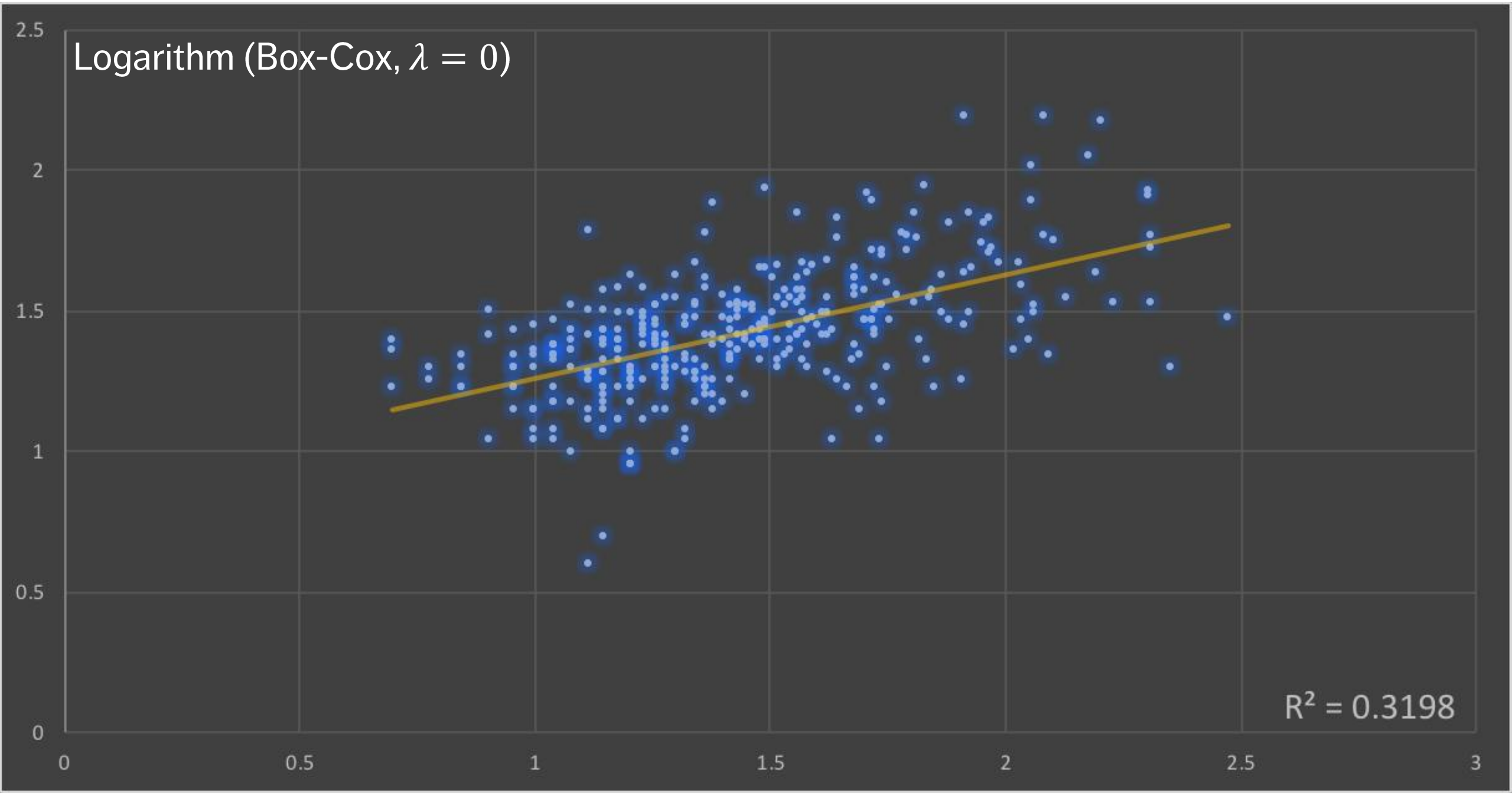
Transformations on X may achieve linearity, but usually at some price (correlations are not preserved, for instance). Transformations on Y can help with non-normality and unequal variance of error terms.


Original Data
180
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Square (Y^2)
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root ($\sqrt{Y}$)
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal ($1/Y$)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3


Logarithm (Box-Cox, $\lambda = 0$)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
$R^2 = 0.3198$

SCALING

Numeric variables may have different **scales** (i.e., weights and heights).

The variance of a large-range variable is typically greater than that of a small-range variable, introducing a bias (for instance).

Standardization creates a variable with mean 0 and std. dev. 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

Normalization creates a new variable in the range $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

DISCRETIZING

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to "*short*", "*average*", "*tall*", for instance).

Domain expertise can be used to determine the bins' limits (although that may introduce unconscious bias to the analyses)

In the absence of such expertise, limits can be set so that either

- the bins each contain the same number of observations
- the bins each have the same width
- the performance of some modeling tool is maximized

CREATING VARIABLES

New variables may need to be introduced:

- as **functional relationships** of some subset of available features
- because modeling tool may require **independence of observations**
- because modeling tool may require **independence of features**
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis)

Time dependencies ⟶ time series analysis (lags?)

Spatial dependencies ⟶ spatial analysis (neighbours?)

DATA ANALYSIS

DATA TRANSFORMATIONS