POWER TRANSFORMERS

What are Power Transformers?

- reshapes numerical variables to follow a uniform or Gaussian distribution, making them more normally distributed.

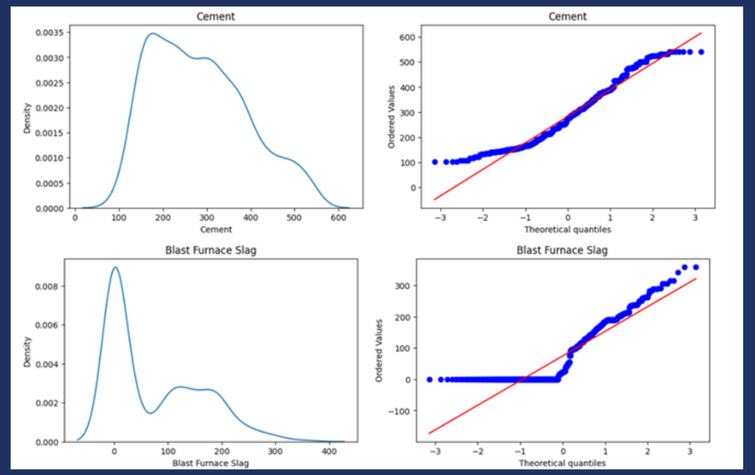
The Data set used:

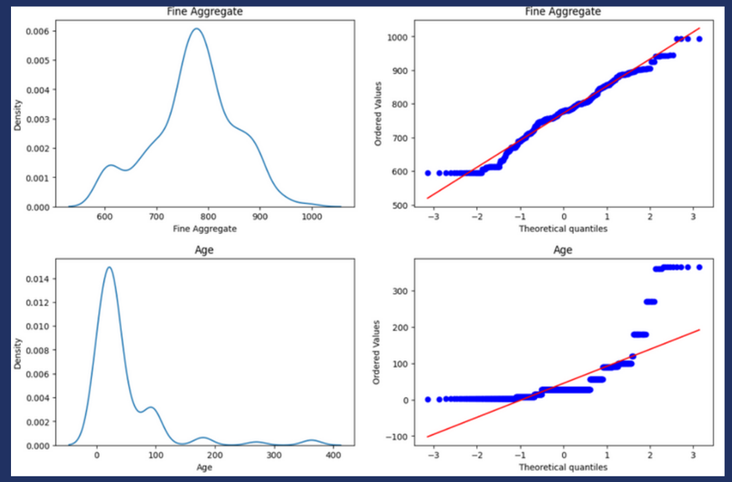
“concrete_data.csv” is a classic and widely used resource in machine learning and civil engineering research, the dataset contains 1,030 instances, with no missing values.

It presents a highly nonlinear relationship between the mixture composition, curing time, and resulting concrete strength, making it a challenging and insightful regression problem.

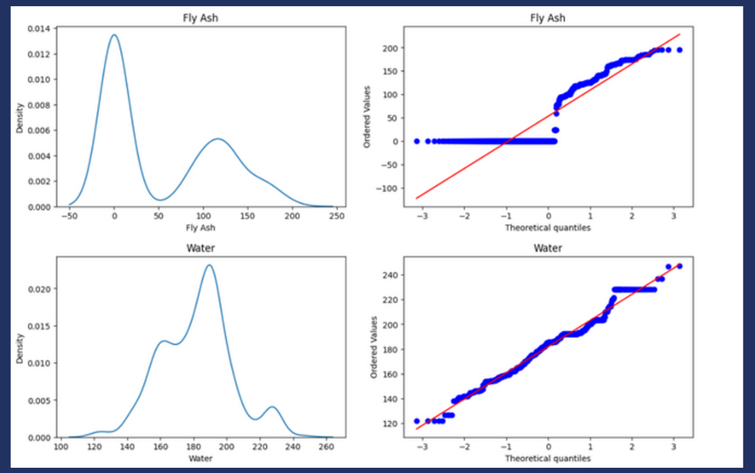
Since many of the input features and even the target variable are not normally distributed and can be highly skewed, applying a power transformation helps stabilize variance, reduce skewness, and make the data more Gaussian-like.

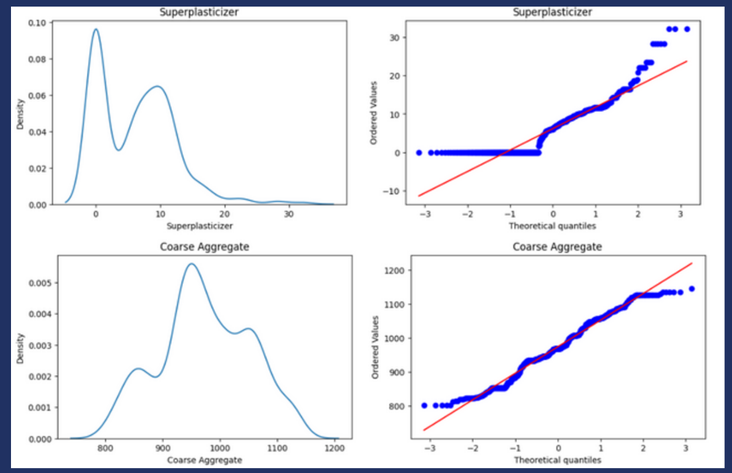
Problem





The features are not normally distributed.

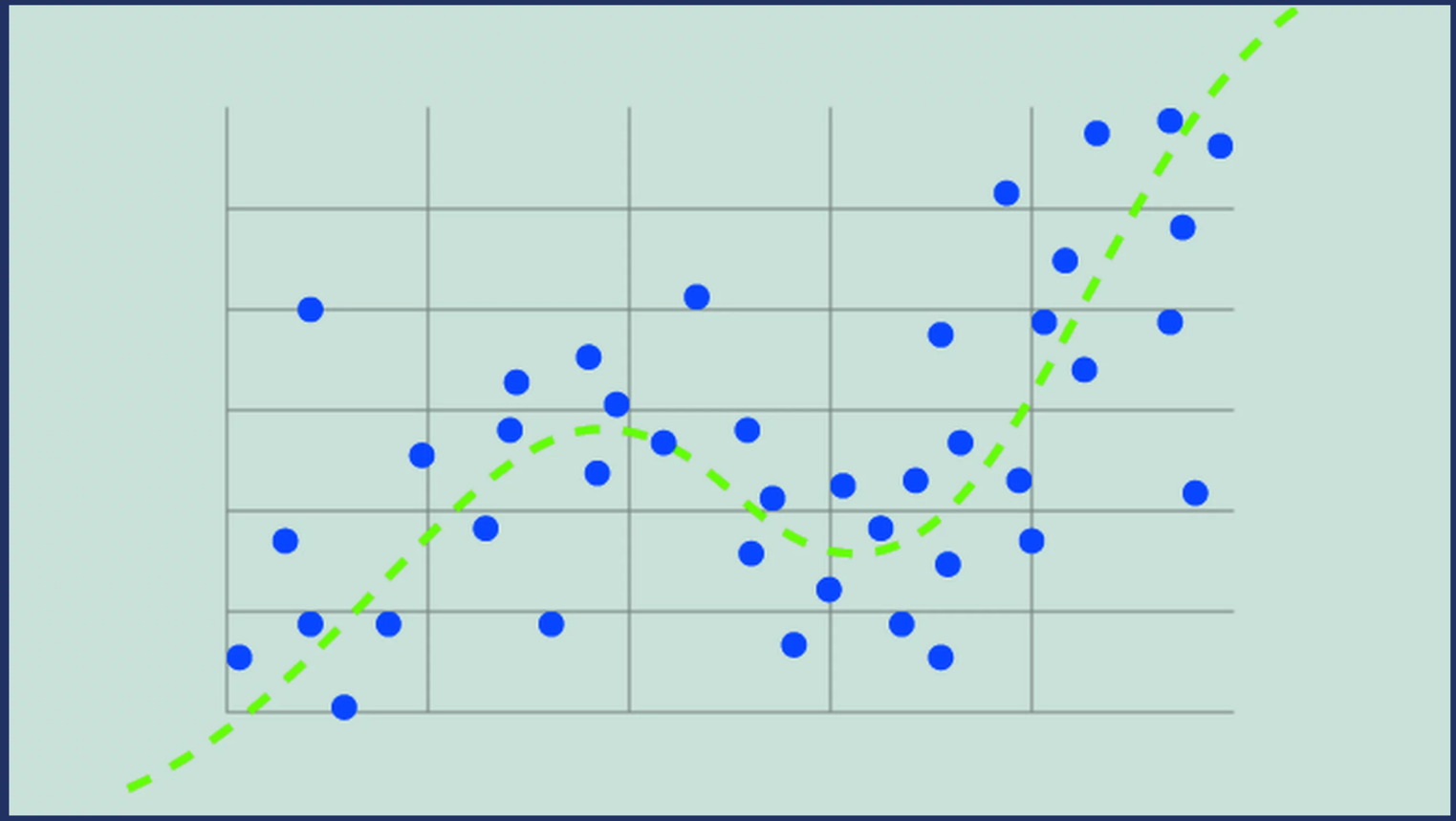




Shows skewness, multimodal peaks and heavy tails.

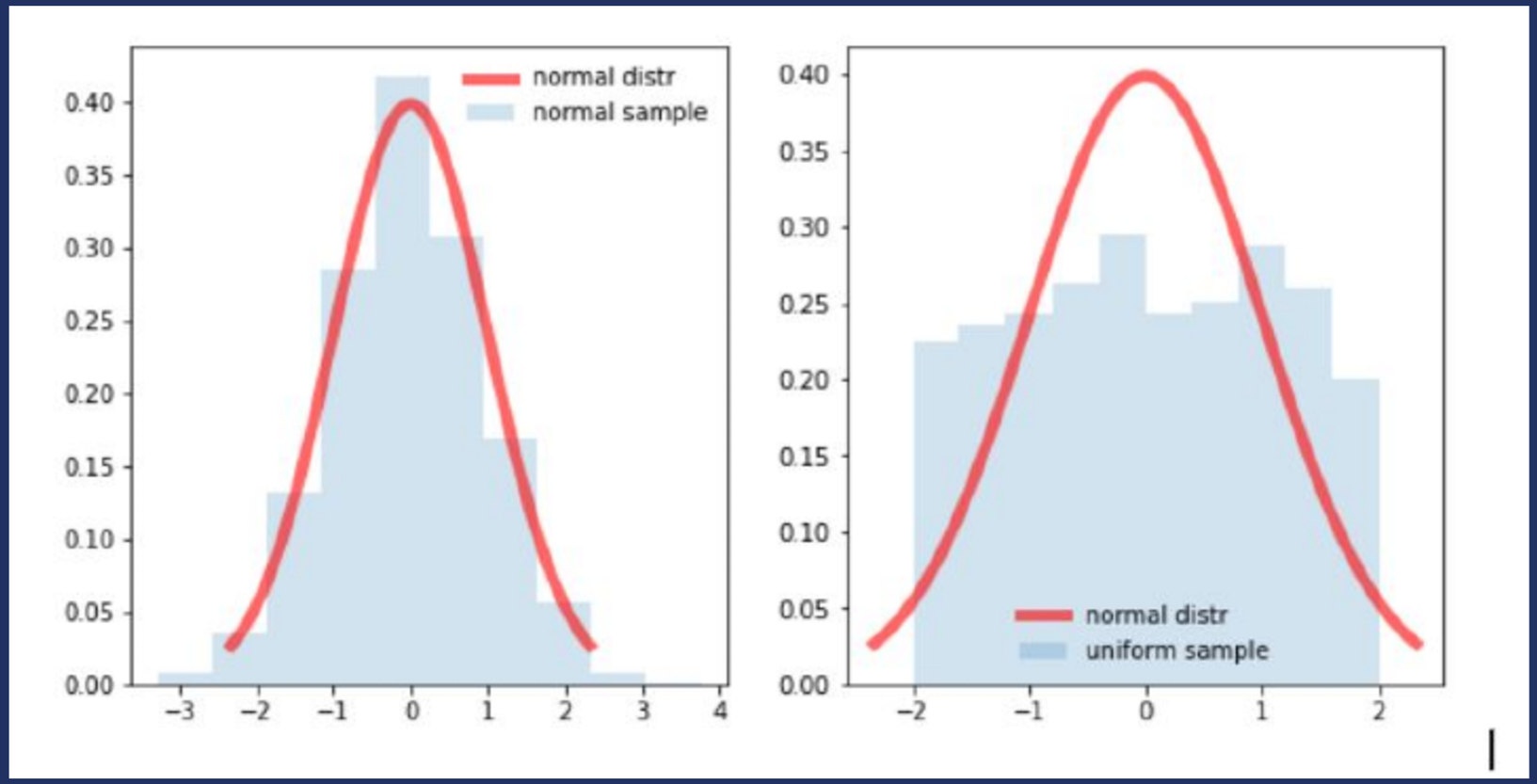
Applications of power transformation:

Improving Regression Models



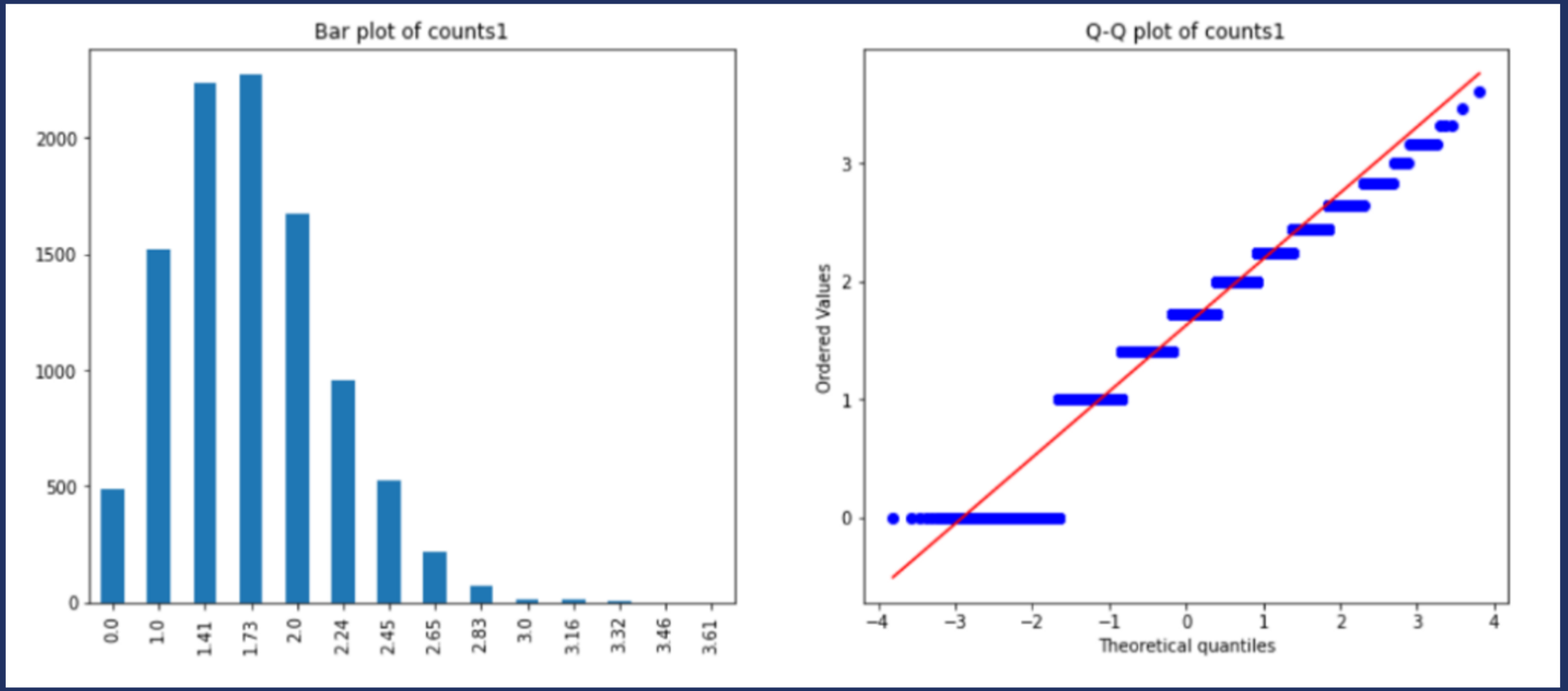
Applications of power transformation:

Enhancing Feature Normality



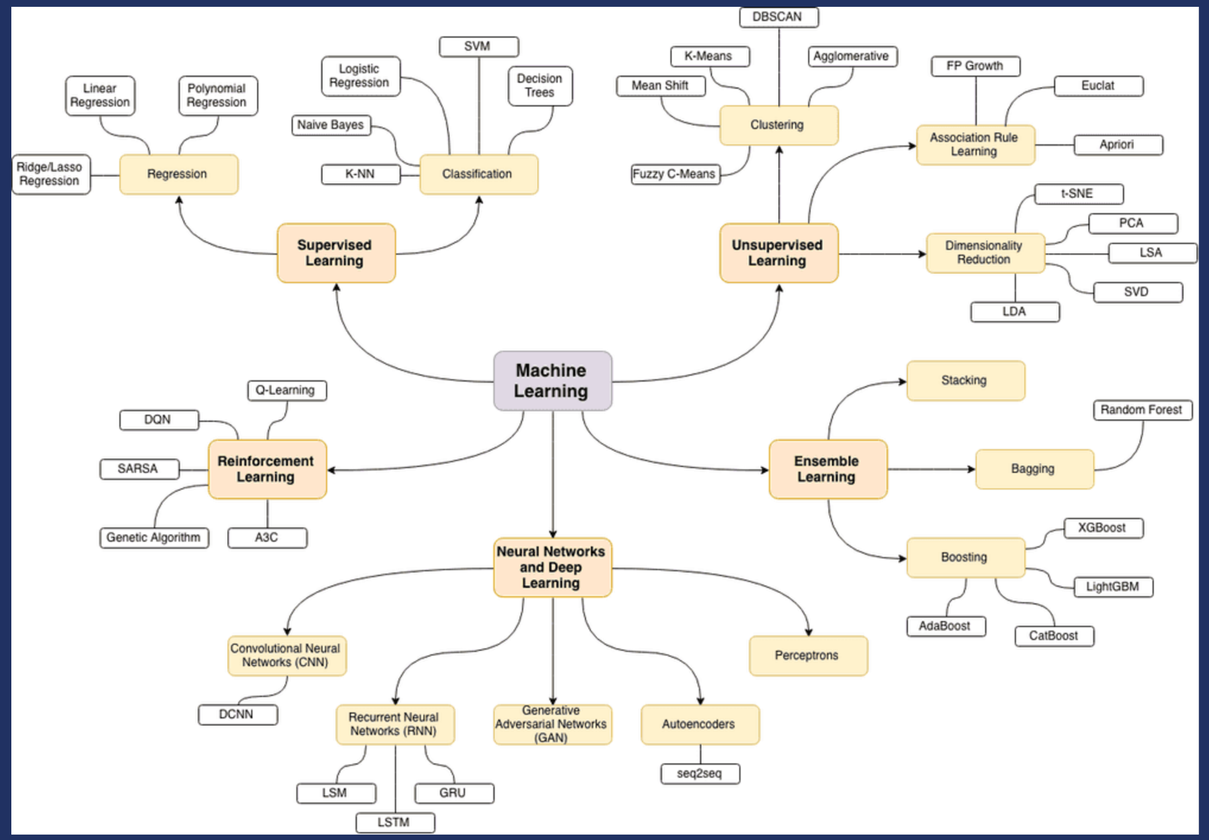
Applications of power transformation:

Stabilizing Variance



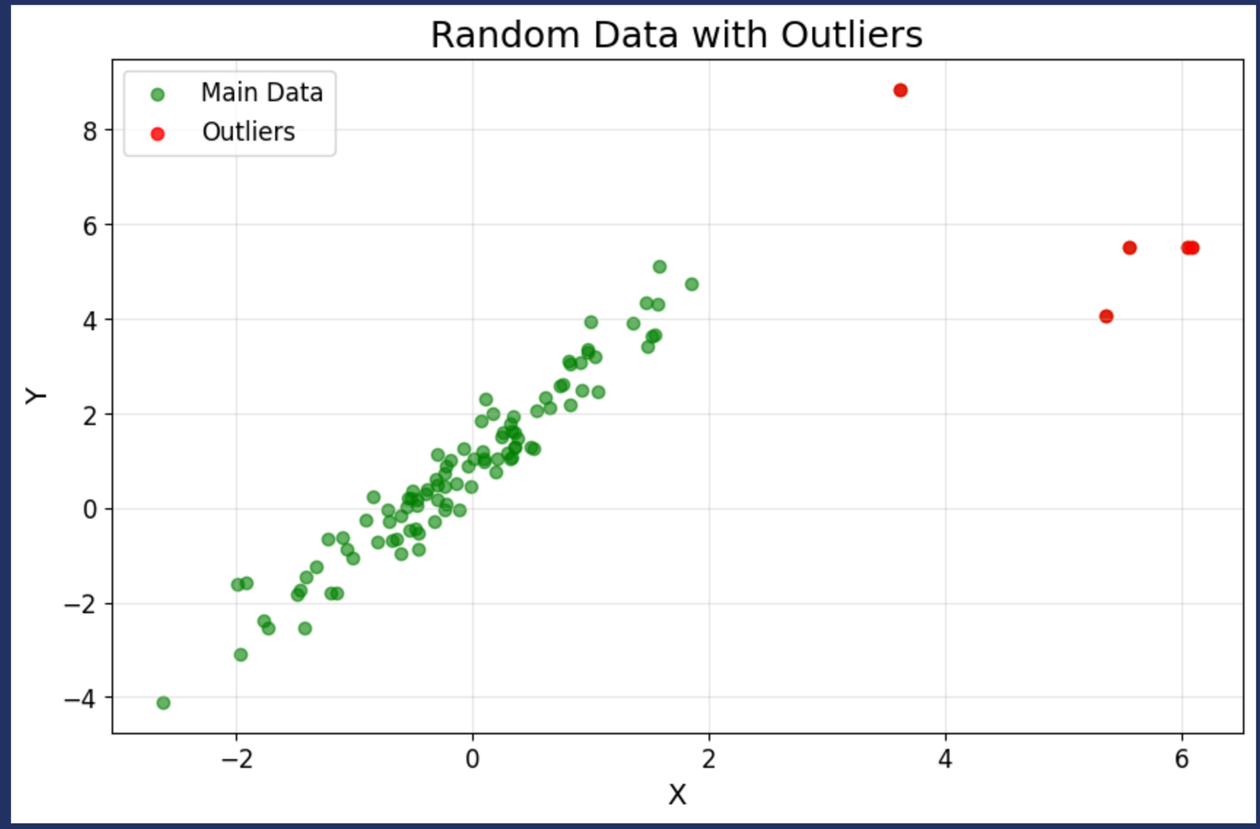
Applications of power transformation:

Improving Machine Learning Algorithms



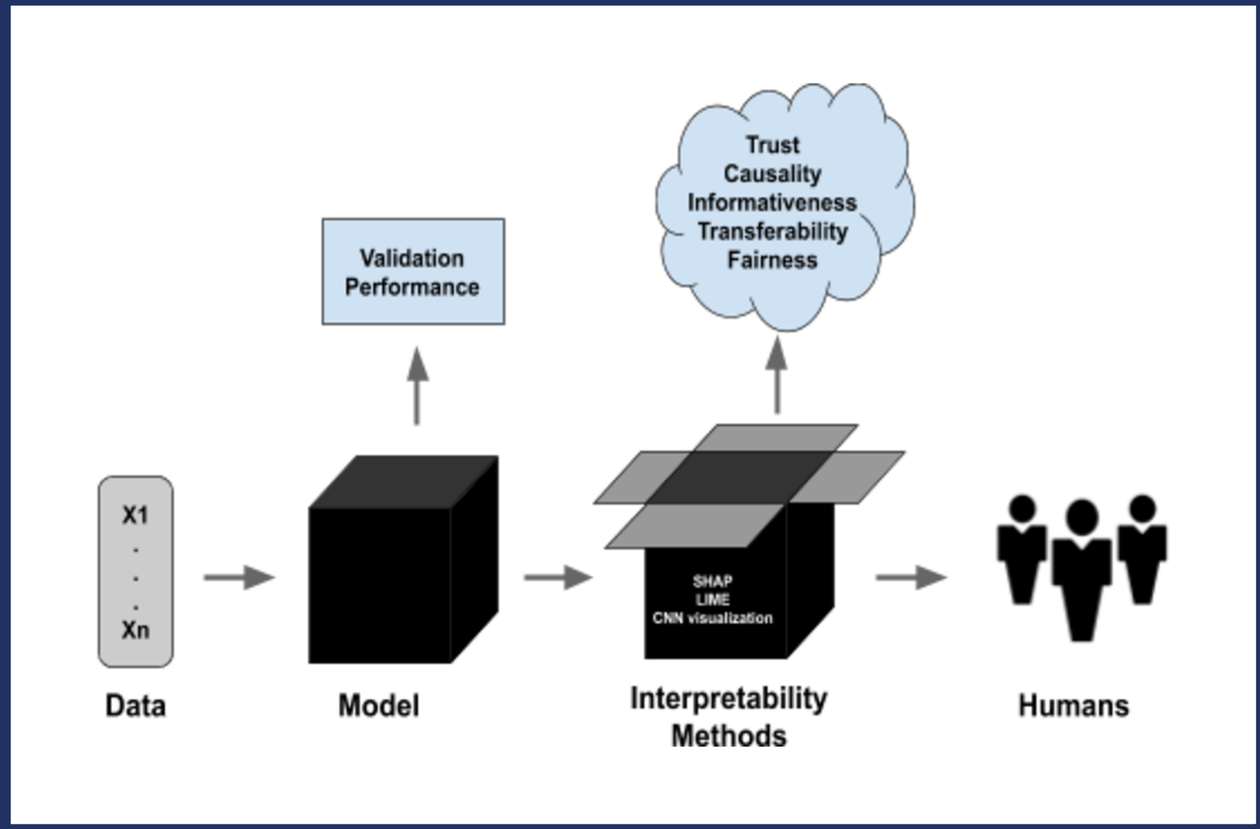
Applications of power transformation:

Handling Outliers



Applications of power transformation:

Enhancing Interpretability of Data



Pros:

Reduces Skewness. Makes the data more ‘normal-like’

Stabilizes Variance. Prevents outliers from dominating the data.

Better interpretability. Patterns and relationships can become clearer.

Improves model performance. Regression, k-NN, clustering often work much better.

Works on many data types. Box-cox is good for positive values while Yeo-Johnson is for both positive and negative.

CONS:

- Box-Cox is **limited** to positive data and Yeo-Johnson is needed for negative/zero values
- Potential to **over-fit** or **introduce noise if misapplied**
- **Not required** for tree-based models
- Adds **computational overhead** and **complexity** to preprocessing pipelines
- **May not always** yield interpretable transformed features

THANK YOU!

“If you torture the data long enough, it will confess"

-Ronald H. Coase