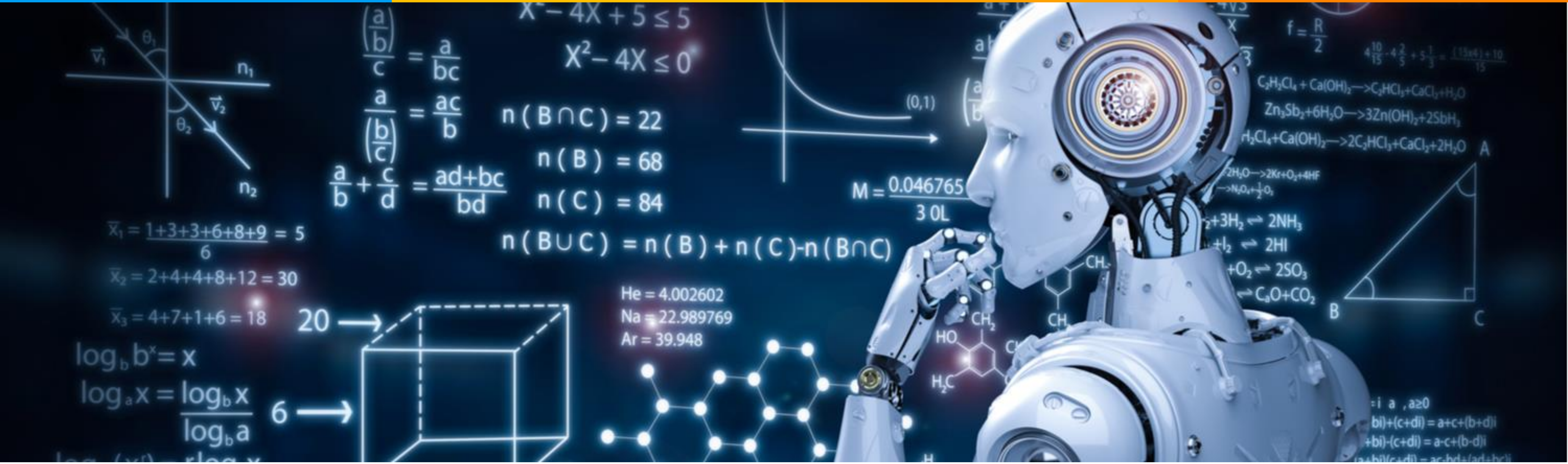
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML



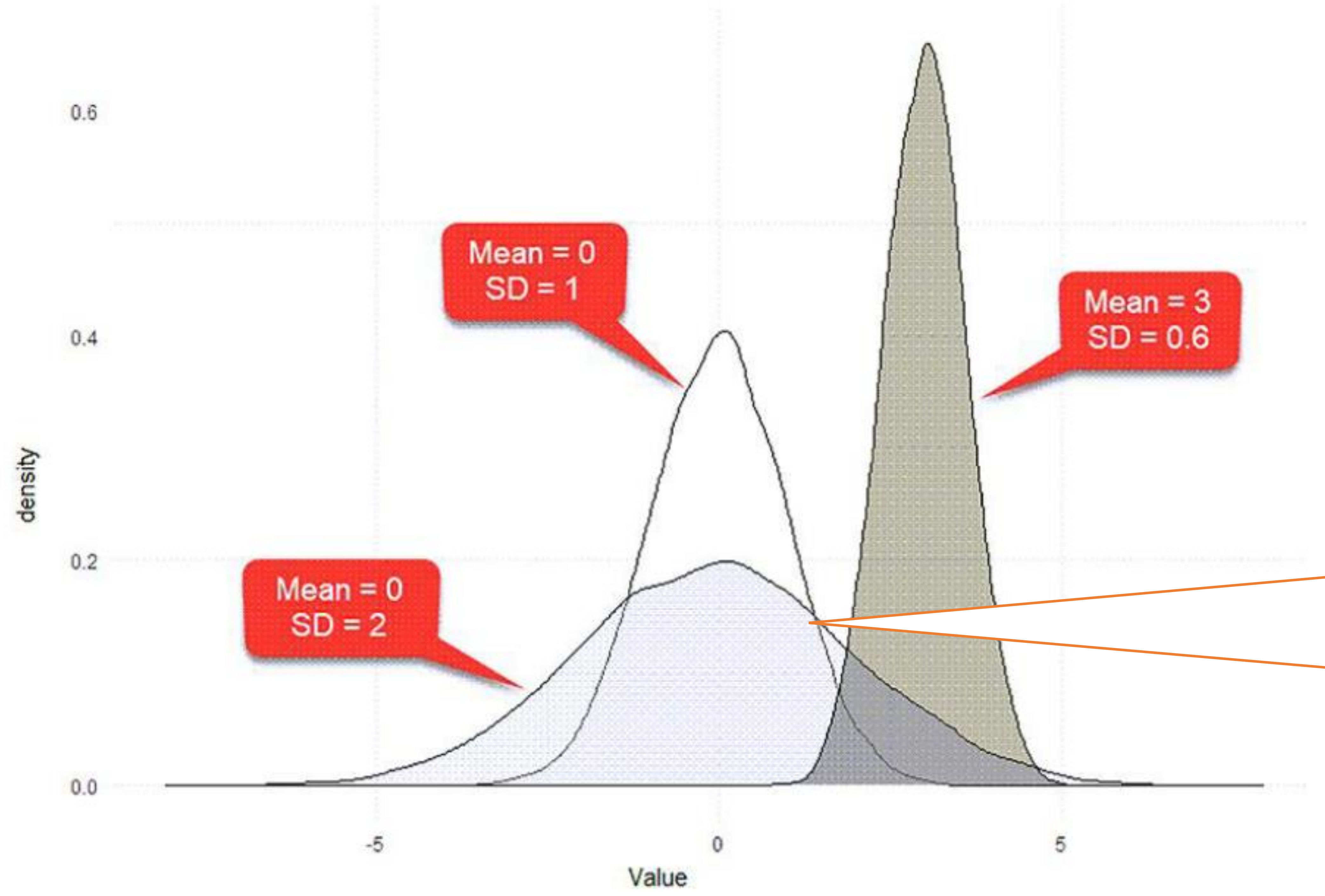
Lecture 09: Outliers & Feature Transformations

Outliers handled 3 ways

- Remove from dataset
- Make more “palatable” with appropriate scaling and/or transformation
- Analyze with specialized techniques that focus on detecting & predicting outliers

Standard deviation method

- Works best for Gaussian distributions $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$
- A univariate Gaussian



Why does height decrease as the distribution becomes wide?

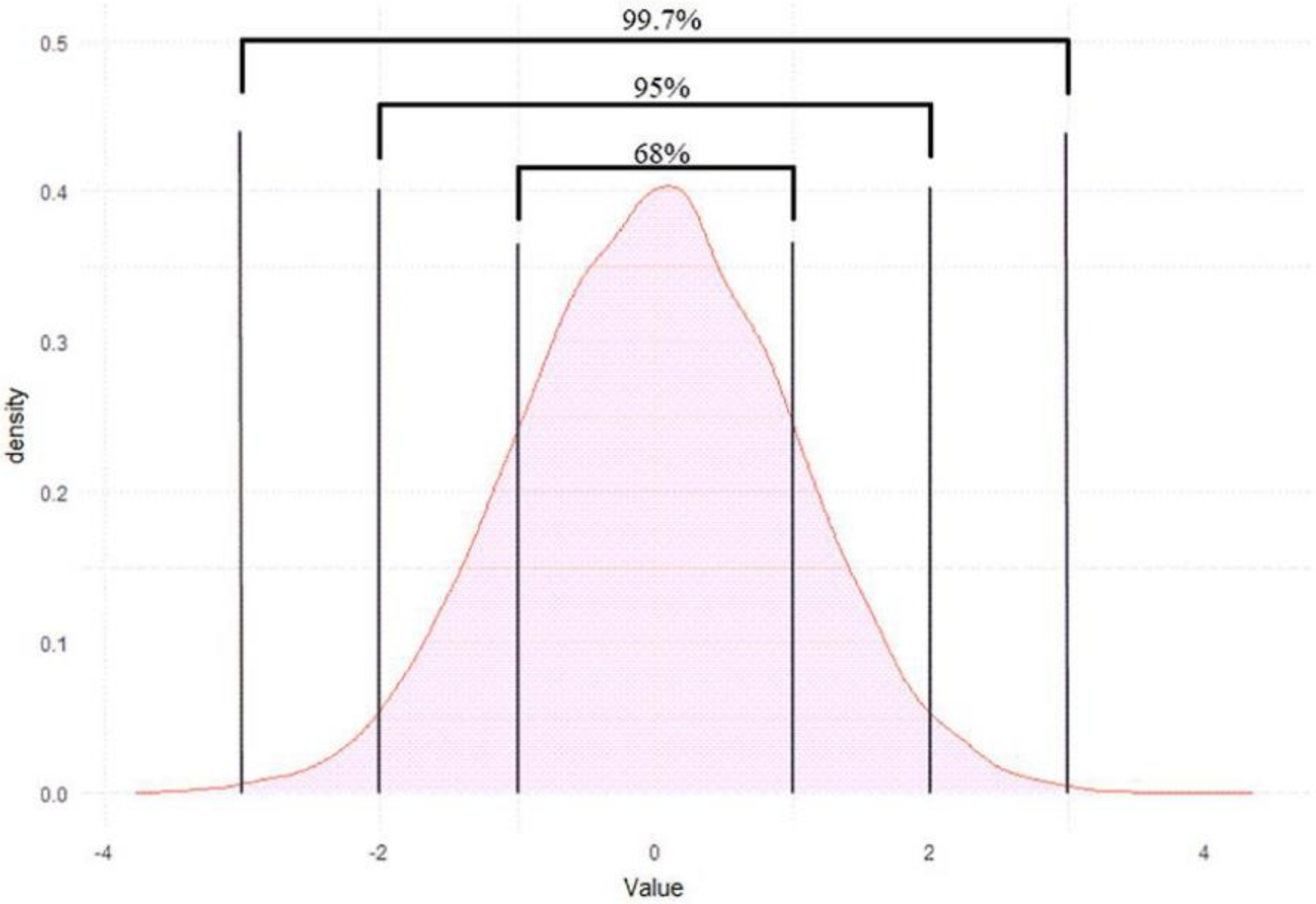
Standard deviation method (contd.)

- Standard deviation is the typical deviation of feature value from mean

$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \mu)^2} = \frac{\|x - \mu\mathbf{1}\|}{\sqrt{n}}$$

- https://www.geogebra.org/calculator/ve2earrn

Empirical Formula for Gaussian Distribution



Outlier v/s anomaly

- Used interchangeably
- Difference between outlier and anomaly
  - Outlier is determined by technical considerations
  - Anomaly is additionally determined by business considerations
- Outlier is generally
  - 2.5 Standard deviations away
  - 3 standard deviations away
  - Or some SD per our choice suiting the problem

Removing outliers during data preprocessing

- Data farther than 3 standard deviation from mean
  - Impacts mean – pulling it in the direction of outlier
  - Skews prediction
- How to deal?
  - Delete first
  - Apply Standard Scaler for the rest

Convention to treat any feature transformation as function

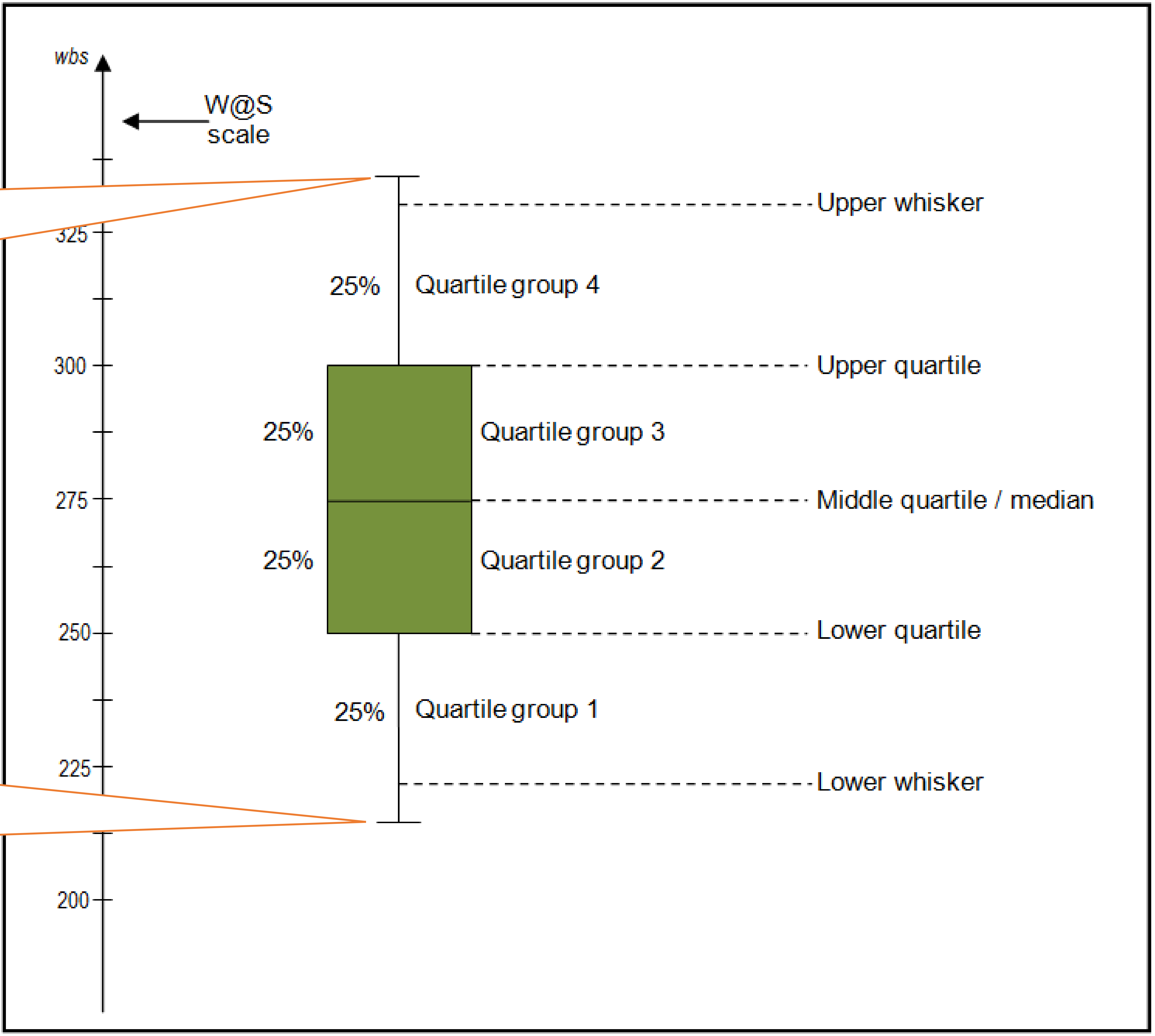
$$\phi(x) = z = \frac{x - \mu}{\sigma}$$

Retaining outliers

- Box plot analysis
- Very extreme values can be removed
- Then Robust Scaler

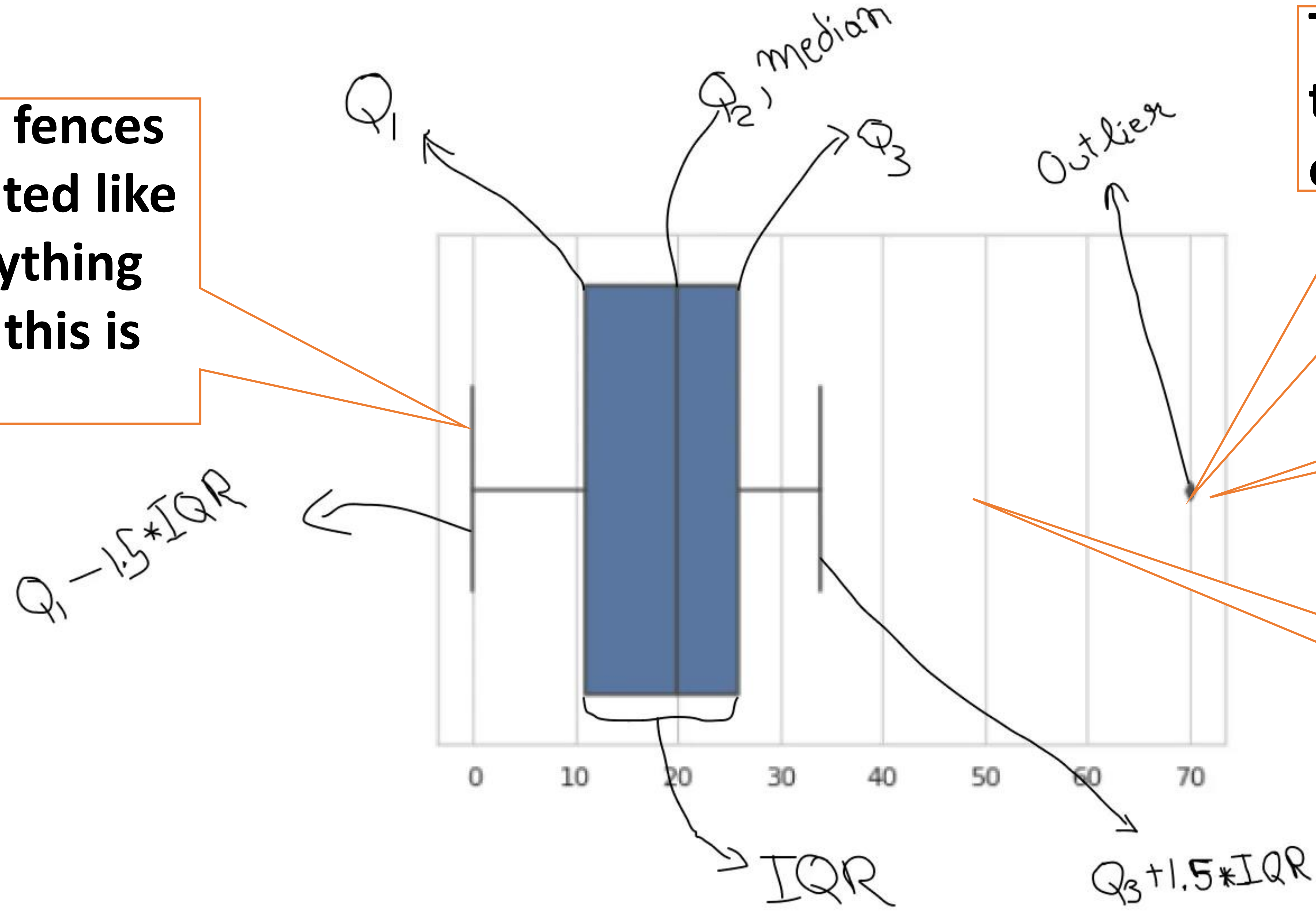
Do we care about max in the context of outliers?

Do we care about min in the context of outliers?



Retaining outliers – Box Plots

Internal fences are created like this. Anything outside this is outlier



This is max, but this is also outlier

We MAY decide to delete this

We may retain points from this region

Retaining outliers – Why 1.5 IQR?

1.0

```
Lower Bound:

= Q1 - 1 * IQR
= Q1 - 1 * (Q3 - Q1)
= -0.675σ - 1 * (0.675 - [-0.675])σ
= -0.675σ - 1 * 1.35σ
= -2.025σ

Upper Bound:

= Q3 + 1 * IQR
= Q3 + 1 * (Q3 - Q1)
= 0.675σ + 1 * (0.675 - [-0.675])σ
= 0.675σ + 1 * 1.35σ
= 2.025σ
```

1.5

```
Lower Bound:

= Q1 - 1.5 * IQR
= Q1 - 1.5 * (Q3 - Q1)
= -0.675σ - 1.5 * (0.675 - [-0.675])σ
= -0.675σ - 1.5 * 1.35σ
= -2.7σ

Upper Bound:

= Q3 + 1.5 * IQR
= Q3 + 1.5 * (Q3 - Q1)
= 0.675σ + 1.5 * (0.675 - [-0.675])σ
= 0.675σ + 1.5 * 1.35σ
= 2.7σ
```

- 1.7 IQR = 3 Standard deviation
- 1.5 IQR = 2.7 SD captures 99.65% data in Gaussian

Retaining outliers during data pre-processing

- Apply Robust Scaler and retain data

$$\phi(x) = \frac{x - Q2}{Q3 - Q1}$$

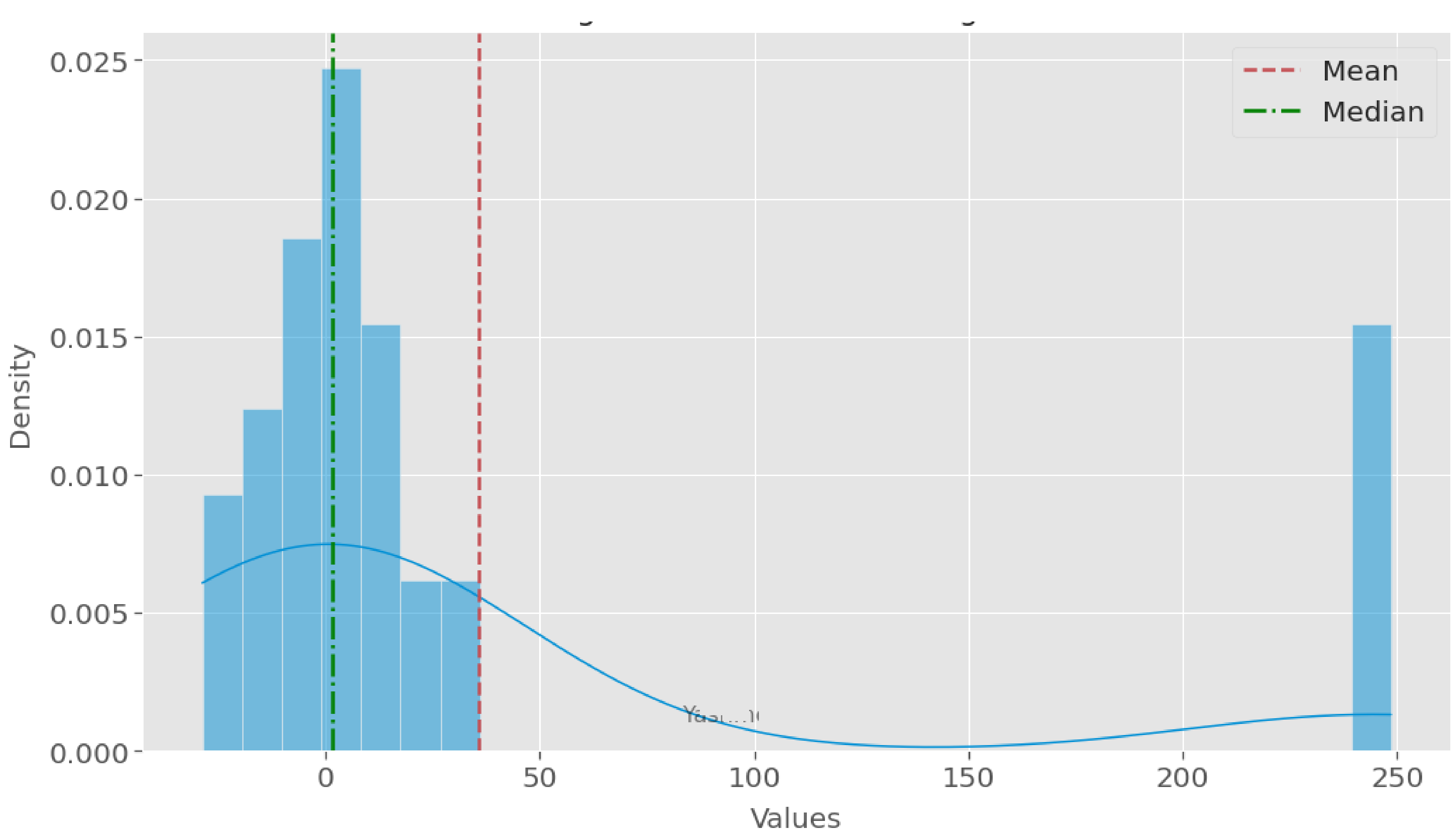
- Q2 = Median, Q3-Q1 = IQR
- Median is not severely affected by the outlier

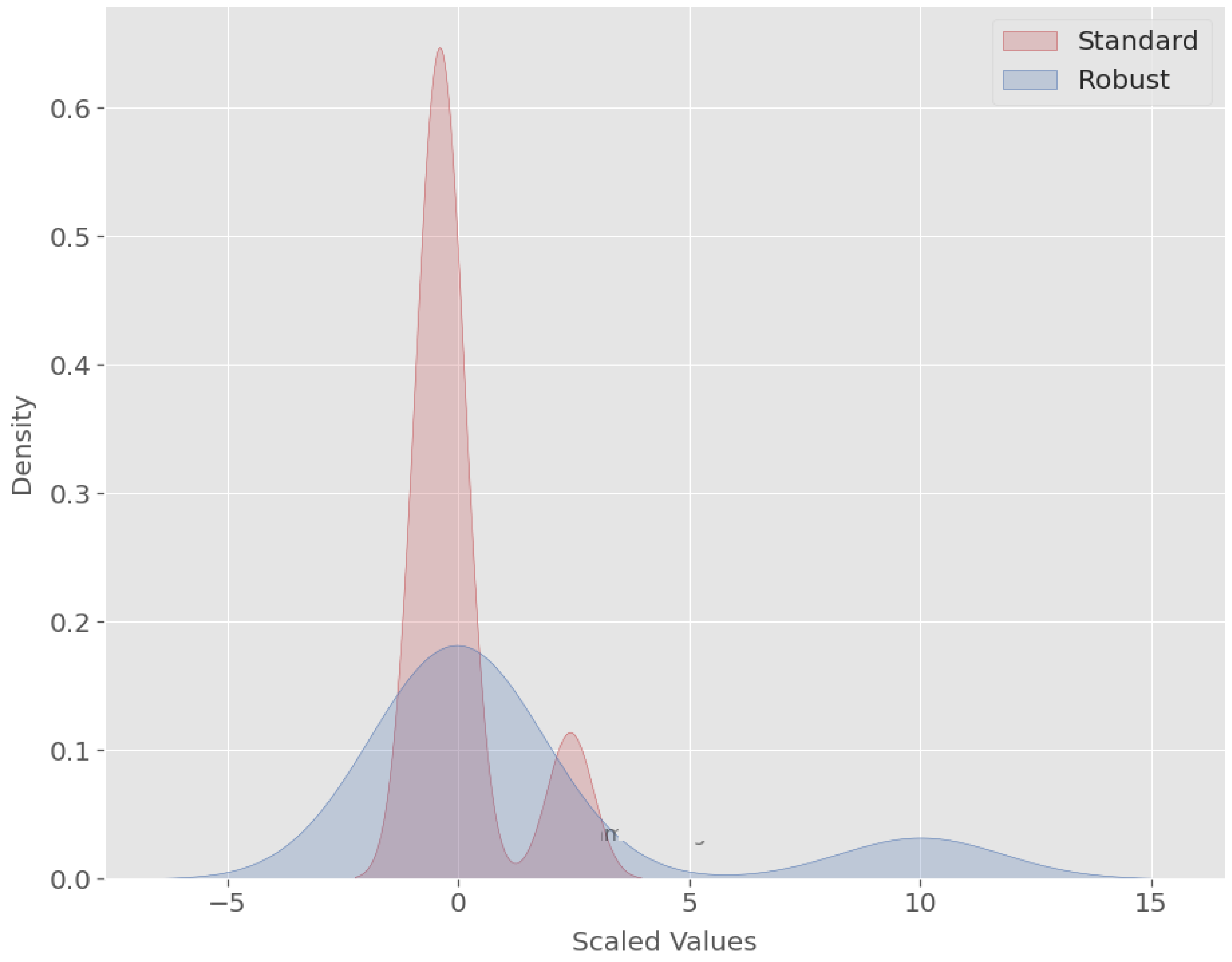
Why Robust Scaler?

	No outliers	With outliers
Min	-28.72	-28.72
Max	32.45	248.51
Range	61.17	277.23
Mean	0.92	35.71
Median	-0.38	1.50
Standard Deviation	15.47	87.64
IQR	17.23	24.30

	Standard	Robust
Min	-0.75	-1.24
Max	2.46	10.17
Range	3.21	11.41
Mean	-0.00	1.41
Median	-0.40	0.00
Standard Deviation	1.01	3.61
IQR	0.28	1.00

- Robust Scaler provides wide range for feature
- Standard Scaler shrinks the range due to outliers
- More variance in scaled data is good for prediction


Mean
Median
Density
Values
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250


Standard
Robust
Density
Scaled Values
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−5
0
5
10
15

Sample sessional problem

- A dataset has 3 features
- Outliers were deleted on first two features
- Standard scaler was applied on the first feature
- MinMax Scaler on second feature
- Third feature ranges between 1500 to 10,000
  - Robust Scaler was applied

Are 3 features ML ready now?

- Clue:
  - What is the range of Standard, Minmax & Robust Scaler?

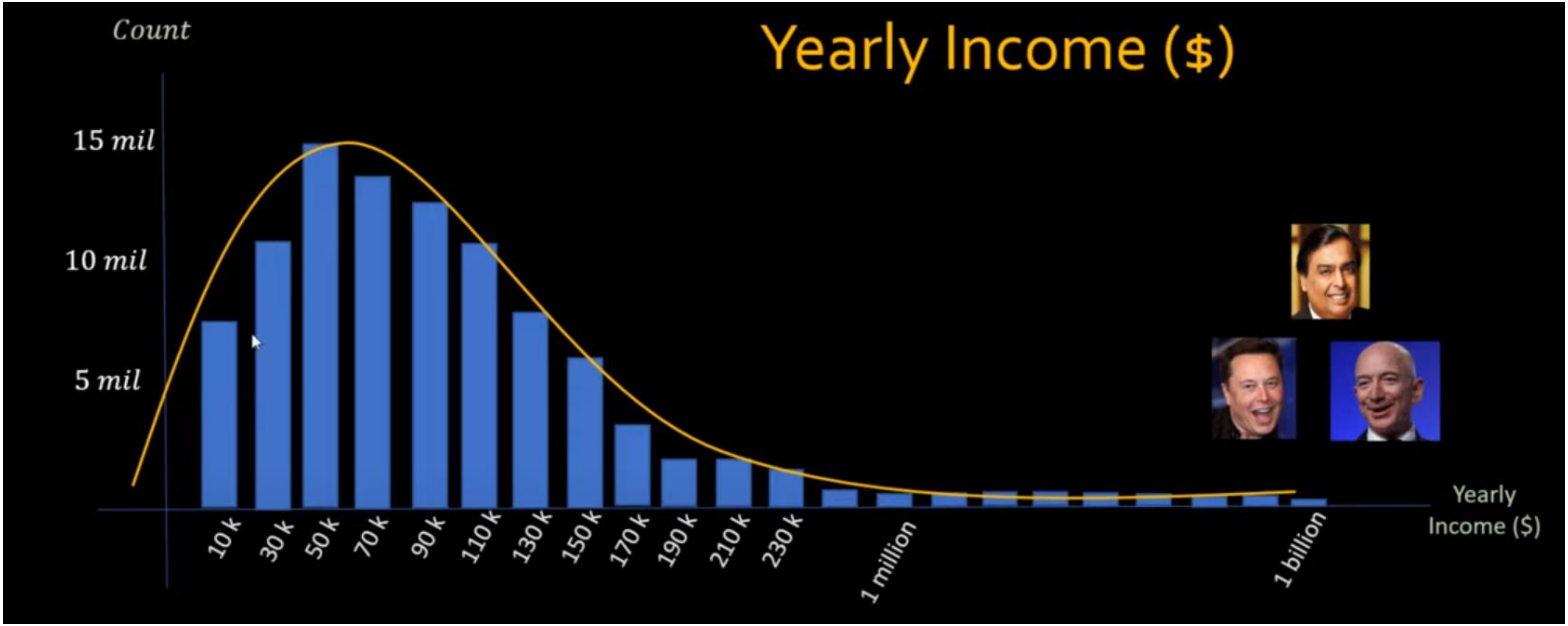
MinMaxScaler

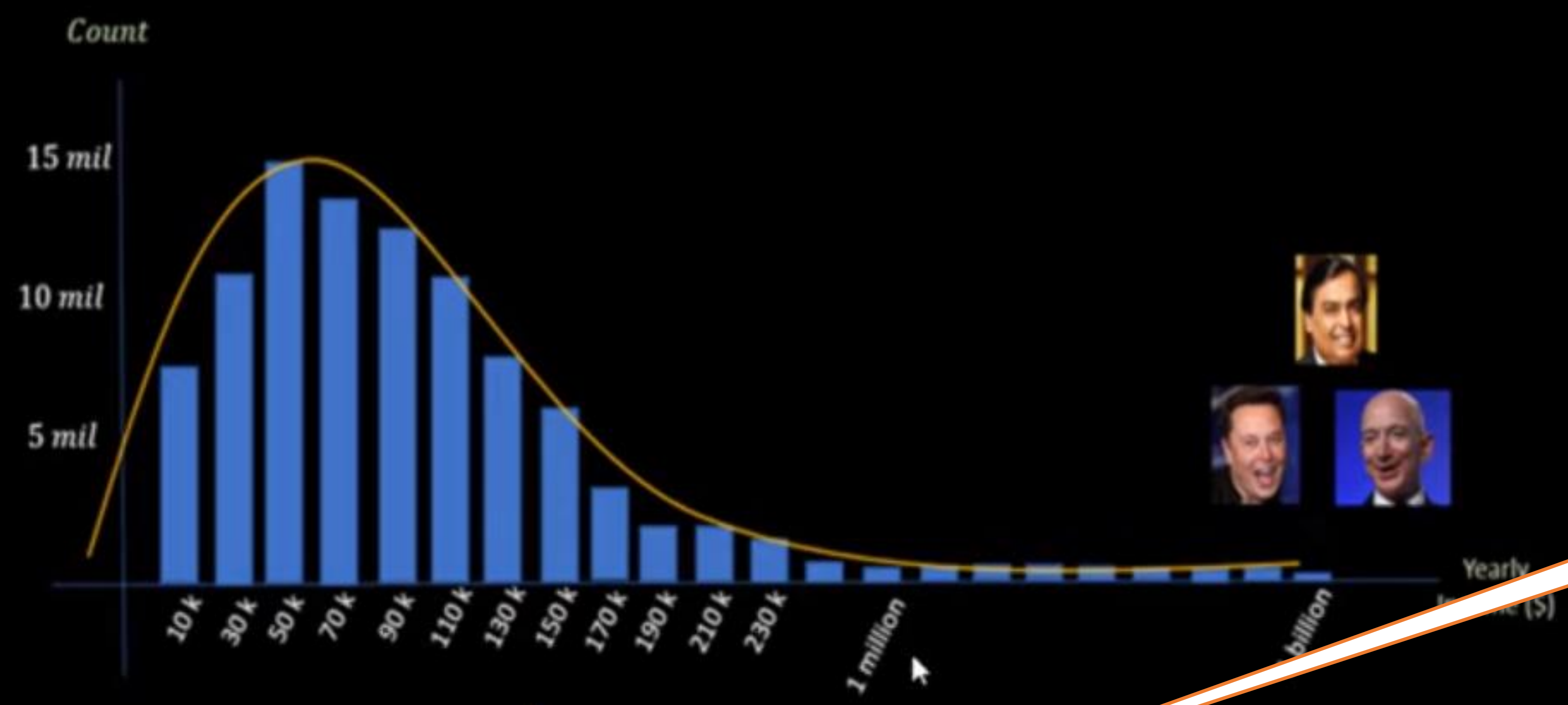
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Compress data in 0-1 range
- Applied to images
- Compress inliers in 0-0.5 when outliers are present
- https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Distribution Transformations

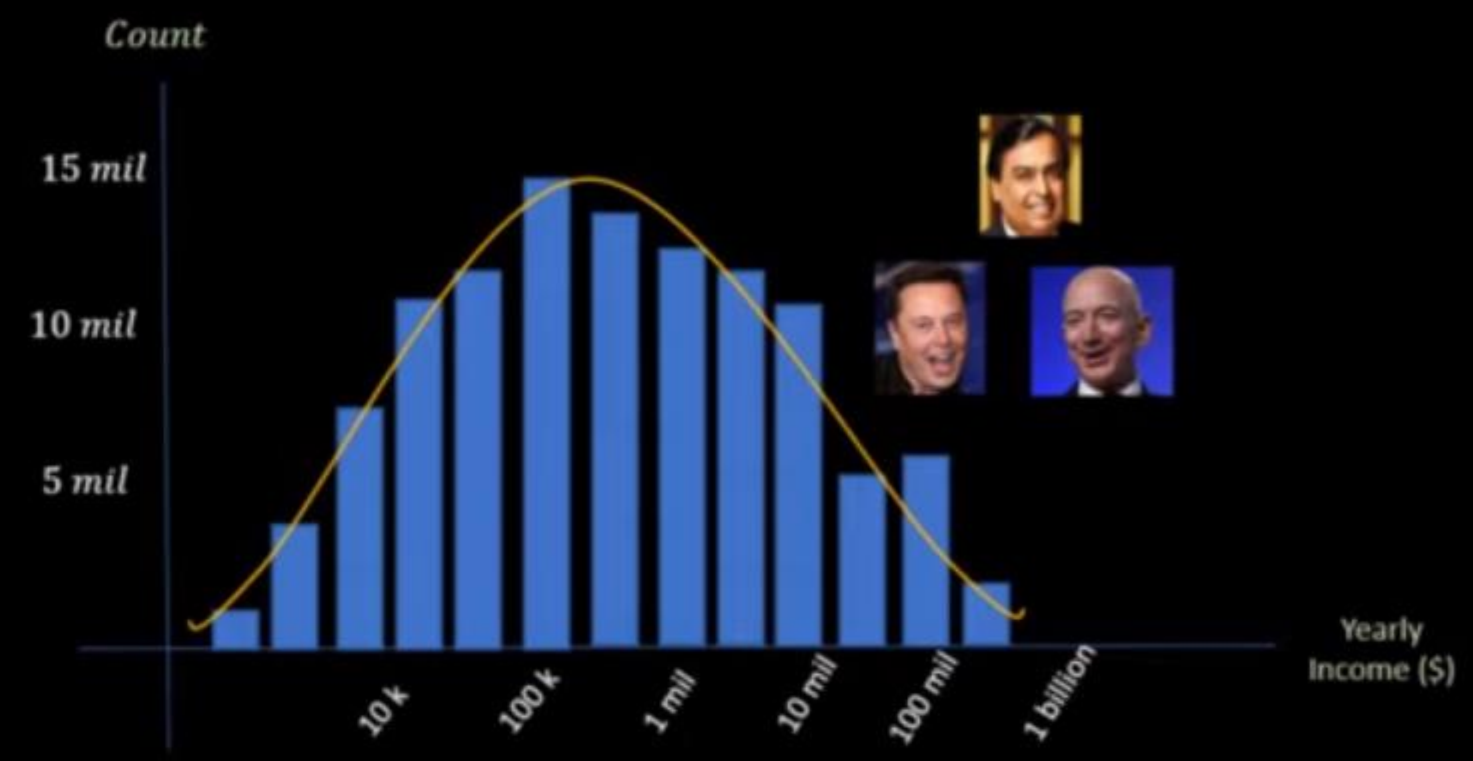
Log Normal Distribution



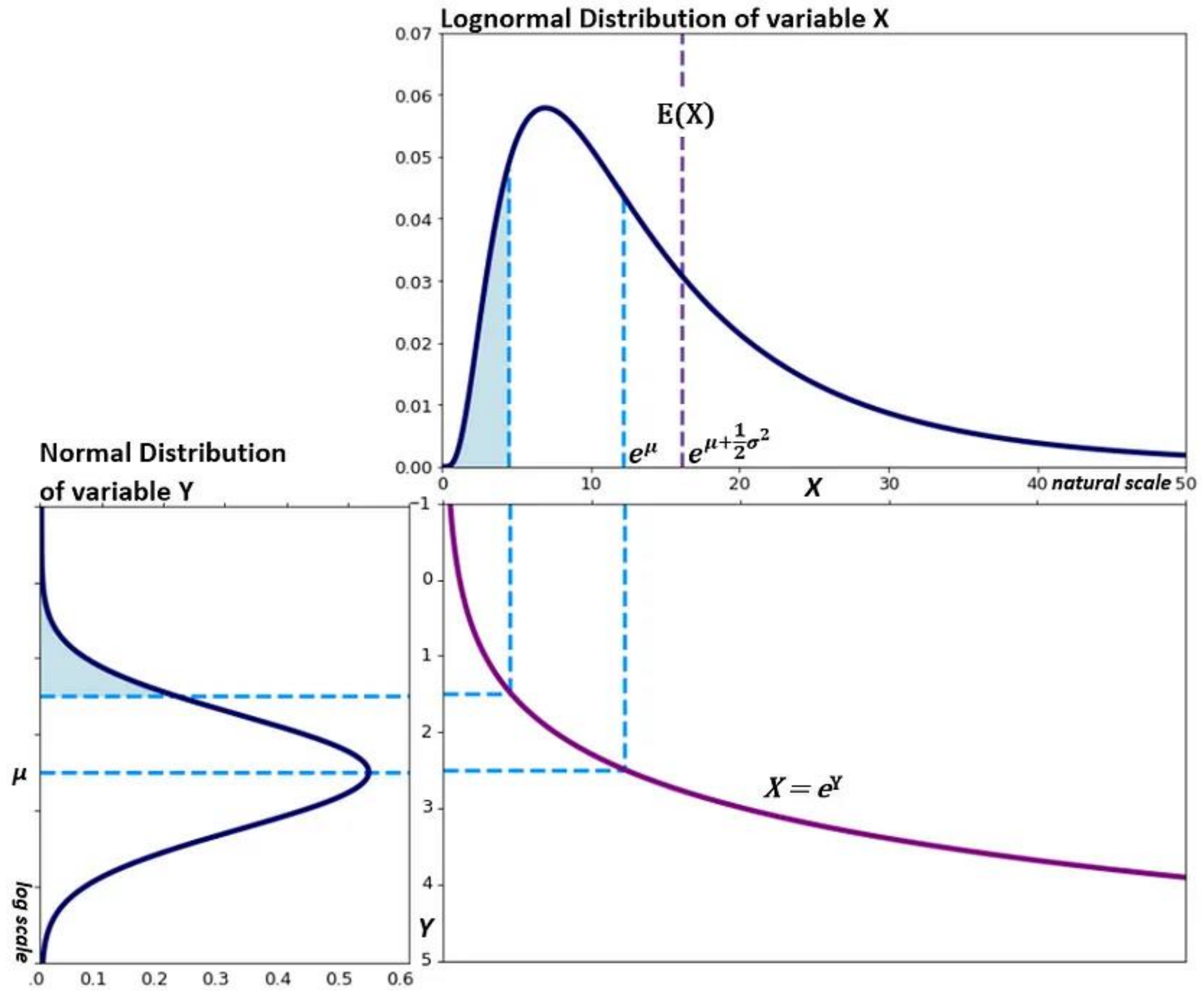


log(1+income)

log (income)



If you get a normal distribution by applying a log function to a dataset then dataset is log normally distributed



Lognormal Distribution of variable X
E(X)
e^{μ}
$e^{\mu+\frac{1}{2}\sigma^2}$
X
natural scale
Normal Distribution of variable Y
μ
log scale
$X = e^Y$
Y

Generic Goal of Distribution Transforms

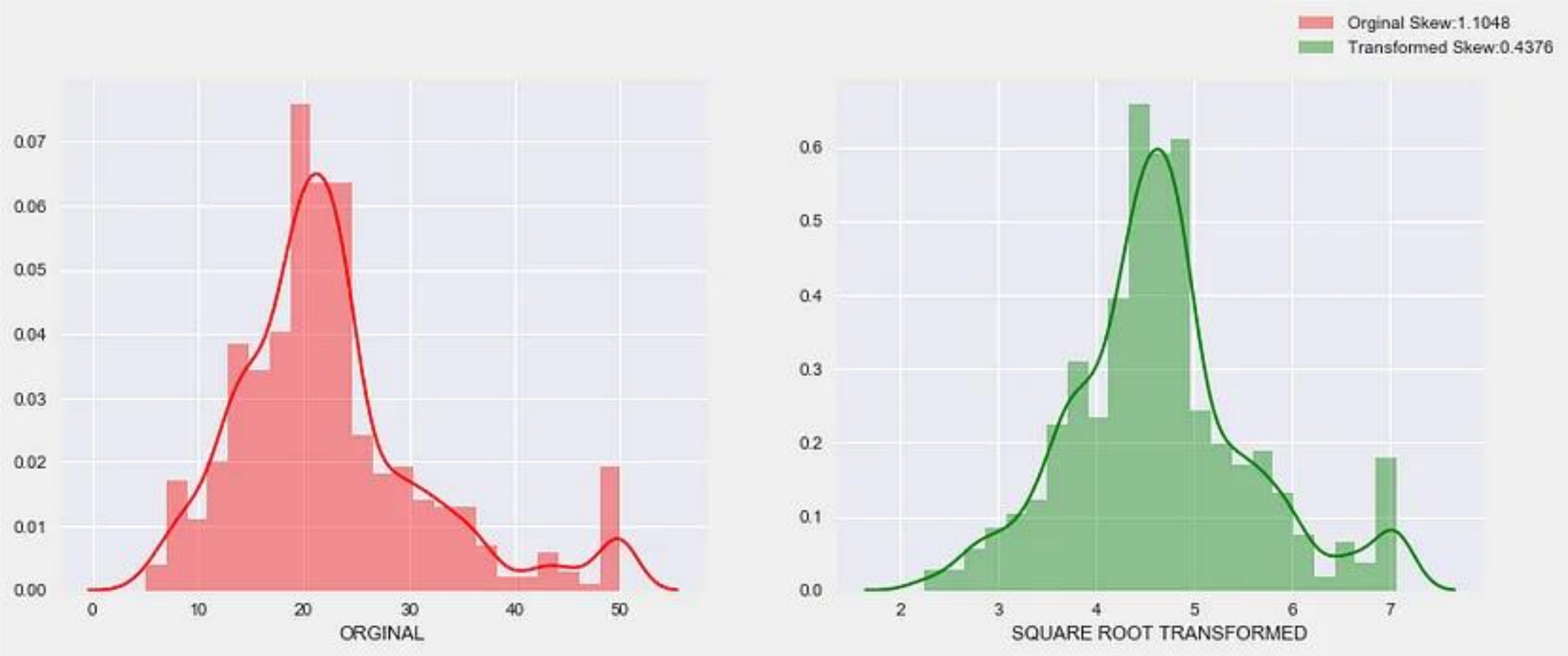
- Change the feature/target distribution to Gaussian
- Important for ML algorithms that have an underlying assumption that target variable is distributed normally
  - Linear Regression
  - Logistic Regression
- Theoretically features need not be Gaussian
- But numerically stable, faster convergence when features are Gaussian in a standardized range

Checking Gaussian-ness

- Visually with Searborn distplot()
- Deviation from symmetric Gaussian
  - pandas.skew(): 0, <0 or >0
- Visually with QQ plot – Very reliable (scipy.stats)
- Normality tests
- What to apply when?
  - Heavily right skewed -> Log Transformer
  - Left skewed – Square Transformer
- Function Transformer in Sklearn

Gaussian Transformation

- Square Root Transformation
  - For removing slight right skewedness
  - Weaker than log transformation



Power Transforms

- Log Transform is a member of family: Power Transform
- Variance stabilizing transformations
- Remove skewness, Creates Gaussian
- Power Transforms $\phi(x, \lambda)$
  - Box Cox Transforms (x > 0)
  - Yeo-Johnson Transforms (for any x)
- Lambda is hyper param. Tune for feature transformation

Power Transforms

Box Cox

Remember this

$$\phi(x,\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ log x, & \text{if } \lambda = 0 \end{cases} \qquad \lambda \in [-5, 5] \quad x > 0$$

Yeo Johnson

Don't even try to remember this

$$\phi(x,\lambda) = \begin{cases} \frac{(x+1)^{\lambda}-1}{\lambda}, & \text{if } x \geq 0 \; and \; \lambda \neq 0 \\ log(x+1), & \text{if } x \geq 0 \; and \; \lambda = 0 \\ -\frac{(-x+1)^{2-\lambda}-1}{2-\lambda} & \text{if } x < 0 \; and \; \lambda \neq 2 \\ -log(-x+1) & \text{if } x < 0 \; and \; \lambda = 2 \end{cases}$$


QUESTIONS

Literature Survey