Variable Transformation

transformation of your variable through mathematical formulation

- Why do we require these mathematical formulation for transformation in Feature Engineering ?

⇒ The distribution of ol data through this gets converted into normal distribution.

The end goal is to normally distribute data.

Why normal distribution is required ?

⇒ Problem solving around normal distribution becomes easy.

Linear / logistic regression assume that data is normal distributed, if not then we have to normalize it.

Few machine learning algo works better on normally distributed data, and these transformer help us normally distribute the data.

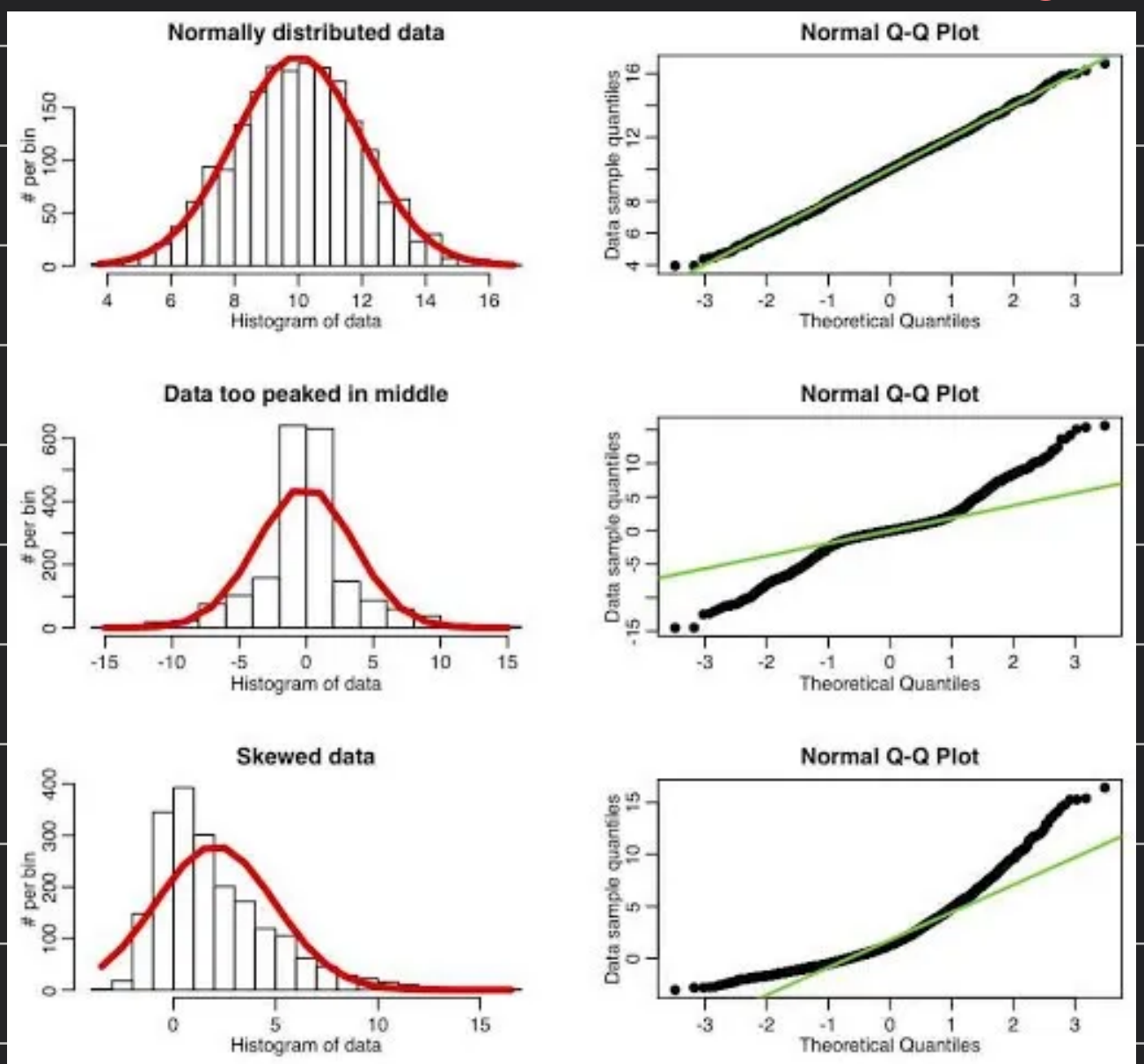
How to find if data is normally distributed ?

There are few ways

1.) sns.distplot ⇒ gives a distribution looking at which we can see if data is normally distributed

2.) pd.skew() ⇒ pandas skew function
if o/p ⇒ 0 → then Normally dist
o/p ⇒ -ve/+ve → then data is skewed

3.) QQ plots ⇒ The best way & reliable way



This is one way of reading QQ plots, other forms exist as well

▲ log Transform
We take log of all the value present

in our variable. Base 2 or 10 depends, the data Normal distributed not completly.

we can't apply this on -ve data & can only be applied on right skewed data.

But why Log?

log bring values such as 1, 10, 100, 1000 comes to an equivalent scale, so small values & very large values comes to an equivalent scale.

This is helpful in algo such as linear regression.

Additive scale gets converted to multiplicative scale.

⚠ Reciprocal Transform

$\frac{1}{x}$ ⇒ small value becomes big & big becomes small through this transformation

value of variable

⚠ Sq Transform

Done for left Skewed data, where you square all values in your variable

$x_i^1 \Rightarrow x_i^2$

Box - Cox Transform

cannot be used on -ve values, or 0

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$$

The exponent here is called λ (lambda) that varies over range of -5 to 5 & in the process of examining all values to find appropriate or optimal value for λ. This help in best approximation for our distribution

X ka power (λ) kitna shi rhega. we calculate that. we try all values in range of -5 to 5, 1.25, -1, 1, 2 etc.

we use two techniques for λ calculation

1) max Likelihood
2) Bayesian Statistics

Yeo - Johnson

An adjustment of Box - Cox transform,

which can be applied to -ve numbers.

$$x_i^{(\lambda)} = \begin{cases} [(x_i+1)^{\lambda} - 1]/\lambda & \text{if } \lambda \neq 0 \quad x_i \geq 0 \\ \ln(x_i+1) & \text{if } \lambda = 0 \quad x_i \geq 0 \\ -[(-x_i+1)^{2-\lambda} - 1]/(2-\lambda) & \text{if } \lambda \neq 2 \quad x_i < 0 \\ -\ln(-x_i+1) & \text{if } \lambda = 2 \quad x_i < 0 \end{cases}$$