unit

Handling Missing data

① Remove all missing rows

② (or) Replace with generic substitute values (mean / median / s.d.)

/ most frequent if value is categorical.

③ Imputation :- Estimate a probability model for the missing variable & replace the missing value with one or more samples from probability model.

→ Types of Missing Data

Data may be missing for a variety of reasons.
→ corrupt during its transfer or storage

① MCAR – Missing Completely At Random
– prob of an observation being missing does not depend on observed or unobserved measurements.

Ex :- movie rating from users, since some movies are more popular than others, some movies may not have ratings = $\boxed{\text{NOT MCAR}}$

② MAR → given the observed data, the probability that data is missing does not depend on unobserved data

Ex :-

Y	Gender	Race	Income
	M	Asian	888
	F	Italian	
	M	American	

If missing data depends on Gender/Race then it is MAR

If not its MCAR.

Missing Data & R

→ **Dealing with Outliers**

- → Trimming - Remove outliers
- → Winsorization - Shrink outliers
- → Robustness - Keep the outliers & analyze data using a robust procedure.

Detecting Outliers

- → (1) values below the alpha percentile (α) [or above] $100 - \alpha$ percentile
- → (2) values more than c times std. deviation from the mean.
  - → follows 1st technique (when we assume data is gaussian)

→ Issue: outliers can affect mean and other calculations.

→ So to avoid this remove most Extreme

(or) just use percentiles (more robust)

(Lecture-8) DVA

→ **Data Transformations: Skewness & Power Transformations**

Data is generally drawn from a highly skewed distribution and that is not well described by a common distribution.

→ A simple transformation may map the data to a form that is well described by common distribution.

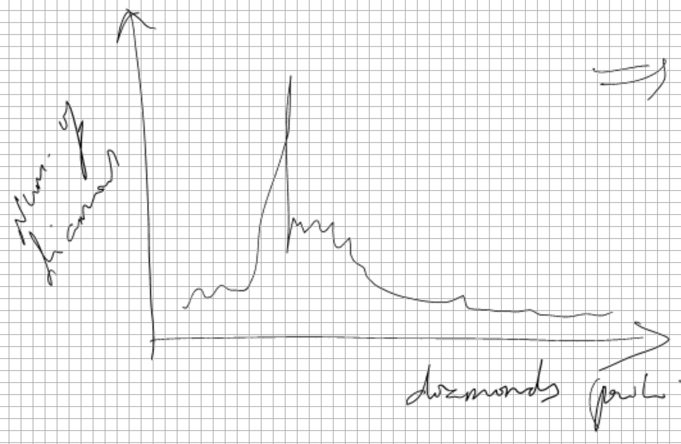
↳ Once transformed a suitable model can then be fitted to data.

→ Power Transformation family (Read from video lecture)

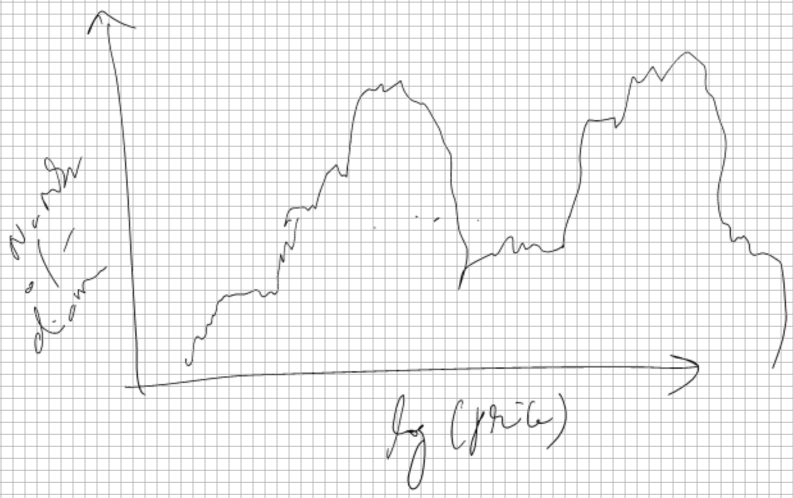
$$f_\lambda(x) = \begin{cases} (x^\lambda - 1)/\lambda & \lambda > 0 \\ \log x & \lambda = 0 \qquad x > 0 \\ -(x^{-\lambda} - 1)/\lambda & \lambda < 0 \end{cases}$$

{Ex} Data transformation helps to form **Insights**.

⇒ As price ↑ no. of diamonds ↓



Transform data



⇒ Transformed data looks like it is **Bi-modal** in nature.