DOMS | Page No.
Date / /

U4-(BI)

Data preparation
↳ cleansing & transforming raw data prior to processing & analysis.

Steps of Data Preprocessing –
1) Data cleaning (filling missing data, smooth noise, resolve inconsistencies
2) Data Integration - (put together due to from diff src resolve conflicts)
3) Data Transformation (normalize, aggregate, generalize.
4) Data Reduction (reduce representation (dimensions)
5) Data Discretization. (reduce numerical values of continous attribute into intervals).

Data cleaning -
make invalid data valid by correcting it by:
1) filling missing values.
2) unified data format
3) convert to network data.
4) remove noise, identify & correctly remove outliers.
5) correct inconsistent data.

Incomplete data → lack of data
→ presence of missing values
→ techniques to adapt.
1) Elimination 2) Inspection 3) Identification 4) Substitution

~~Data affected by Noise~~ (Handle missing values) →
- ~~noise~~ ignore the record
- fill missing values manually.
- use global constant
- use attribute avg, mean, median, mode.

DOMS | Page No.
Date / /

* Data effected by noise
- noise is random error or variance in measured variables values.
- data collection, human error, data transmission errors.
- random disturbance within value.

1) Numeric Noise → plot a boxplot, scatter plot or dispersion
2) categorical data → clustering. datapoints away from all cluster

Data smoothing techniques –
1) Binny → bin median, bin max, bin bandance. 2) Regression - factor values
3) outlier analyses.

* Data Transformation.
• transforming or consolidation of data into appropriate form.
Techniques –
• Standarization (Normalization).
↳ entire data set values have partular property is transform it to have a particular distbution or centred around a particular value.
techniques.

A) Decimal Scaling –
transformation $x_{ij} = \frac{x_{ij}}{10^n}$ n = smallest values that range from [-1, 1]

B) Min Max Scaling –
values are tranformed to fin a certain scale/range. [0, 1]

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} (x_{newmax} - x_{newmin}) + x_{newmin}.$$

eg → $x = 5$ $x_{max} = 6$ $x_{min} = 2$.

$$x' = \frac{5-2}{8-2} (1-0) + 0 \quad = \frac{3}{6}(1) + 1 = \boxed{0.5}.$$

c) Z-score transformⁿ → based on mean & standard deviation

$$x' = \frac{x - \bar{x}}{\sigma} \quad \text{when} \quad \sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{x - 1}}$$

DOMS | Page No.
Date / /

2) Feature Extraction
1) Data Reduction → sampling → random sampling.
→ sampling with replacement.
→ sampling w/o replacement.
→ stratified sampling.
2) Feature Selection → redⁿ in no. of attributed by selection & projection.

Filter	Wrapper	embedded.
• generic method no ML algo	• evaluates specific ML algo	• done by observing each iteration
• fast	• slow (high computation)	• medium.
• less prove to overfitting.	• high chance of overfitting	• used to reduce overfitting
• ↳ correlation based selection.	• eg - forward selection, backward elimination, stepwise selection, etc.	• penalizing coefficient
• chisquare, test, ANOVA, EG.		• eg → LASSO, elastic net, ridge regression, etc.

• more features, more complex the model, diff to train & interpret.

* Principal Component Analysis.

• unsupervised ML algo
• used for dimensionality reduction.
• converts observation of correlated function into set of linearly features
• new transformed features → principal component.
• popular tool → EDA & predictive modeling
• technique to draw strong patterns from reducing variance.
• Appln → Image preprocessing, movie recommendation system.
• contains imp variables & deletes / drops least important variables.

Steps -
1) standardization → $z = \frac{x - \mu}{\sigma}$

2) Covariance matrix computation -
- Calculate strength of joint variability of 2 more variable how much they change in redata to each other.

$$Cov(x_1, x_2) = \frac{\sum (x_{ij} - \bar{x_1})(x_{2i} - \bar{x_2})}{n-1}$$

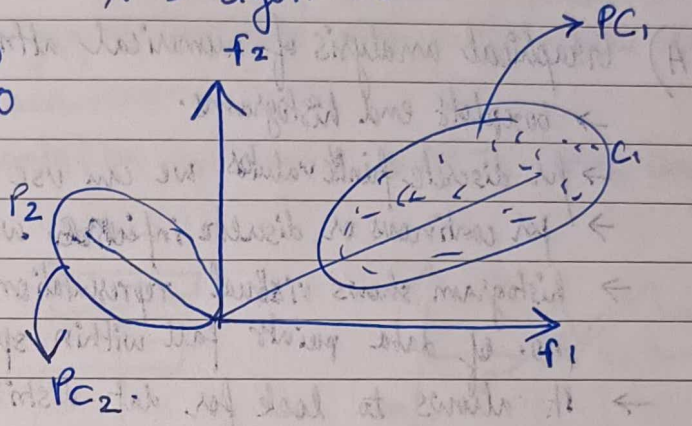
DOMS | Page No.
Date / /

>1 = ↑↑ 0 = no rel^n.
<1 = ↑↓.

3) Calculate Eigen values & Eigen vectors.

$AX = \lambda X$ λ = eigen vector
X = eigen vector.

$(A - \lambda I) X = 0$
$|A - \lambda I| = 0$
$\lambda \Rightarrow$
$X \Rightarrow$



3) Data Discretization - reduction in number of values through discretization & aggregation.
- decreases number of distinct values.
- continous values are connected to categorical attributes
  eg s → weekly expenditure. → 101, 103.5, 45, 35.
  instead we put it in range low (0-20), high (100-400), etc.

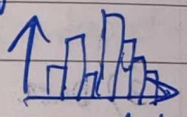
techniques → 1) Subjective Subdivision (based on experience & judgement of experts)
2) subdivison into classes (automated classes based on equal size or work)
3) Hierarchial discrization (applied to categorical attributes)

Methods - 1) Binning → freq
→ equal width (list).
2) Histogram Analysis.
3) Deacision Tree Analysis
4) correletion Analysis

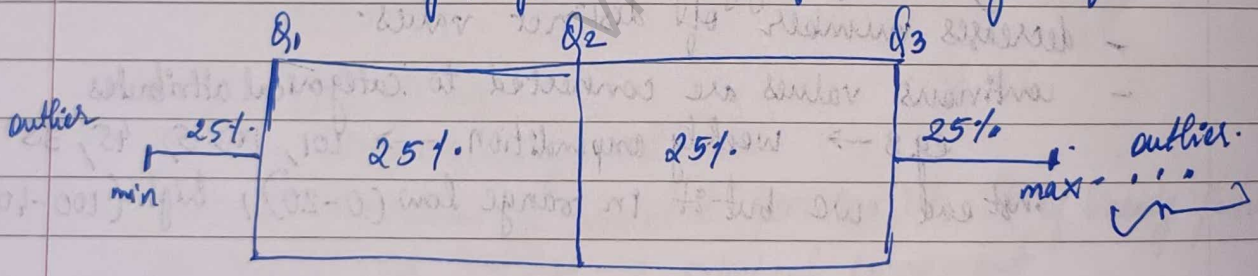
Data exploration -

3 phase → 1) Univariate → study each attribute's property are investigated
2) Bivariate → pair of attributes are considered
3) Multivariate → rel" holding b/w subject of attributes are in

A) Graphical analysis of numerical attributes -
→ boxplots and histograms.
→ for discrete finite values we can use bar chart.
→ for continous or discrete infinite we can use histogram by grouping values into range
→ histogram shows visual representation of numerial data by showing no. of data points fall within specified range of values (bins)
→ it allows to look for dat distribution



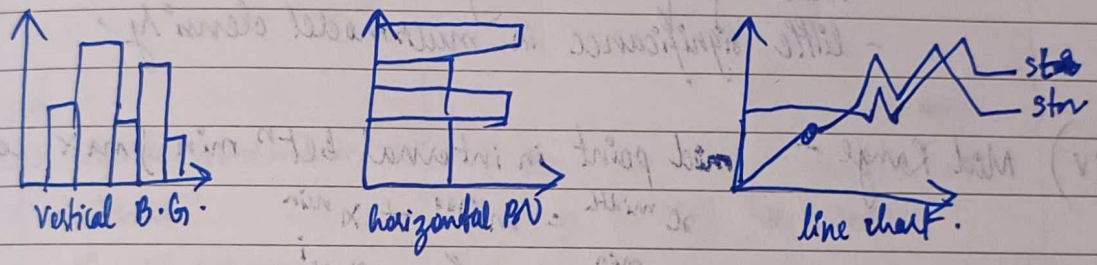
→ Box plot → summerizes data statistics, demostrate like locality, typical skewness of data through quartile



IQR = Q3 - Q1
Lower limit = Q1 - 1.5 × IQR.
Upper limit = Q3 + 1.5 × IQR.
Q1 ⇒ Median of 0 to Q2 (median) 25%.
Q3 ⇒ median of Q2 to n (-) 75%.
outliers by opt of lower & Upper limit.

b) Graphical analysis of categorical attributes-

→ most natural representation vertical bar chart.
→ vertical bar chart value indicates categories & respective frequencies.
→ sometimes horizontal bar chart is also used.
→ pie chart is used to show empirical density function values in %.
→ some times line graphs are used when one wants to see how values change over time.
→ Adv of horizontal bar chart → easier to display long labels.



c) Measuring Central Tendency - (describe main loc" statistics).

i) Mean - org of data points $\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- strongly affected by extreme values (outliers) hence not robust.
- deviation - sum of diff b/w each values & mean = 0 $\sum^{n}(x_i - \mu) = 0$.

- it minimizes sum of squared deviations from constant ref. values.
$\sum^{n}(x_i - \mu)^2 = \min_c \sum_{i=1}^{m}(x_i - c)^2$

- It minimizes sum of weighted mean → $\mu = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots w_n} = \frac{\sum w_i x_i}{\sum w_i}$.

x → shipping cost of product A
w → no. of A products shipped.

10/17

ii) Median - central value of all observations.

$x^{med} = \frac{x_{m+1}}{2}$ if m.n odd $\quad x^{med} = \frac{x_{m/2} + x_{m+2/2}}{2}$

- if affected by no. of elements in sample due to so is more robust mean as not affected by outliers.
- suitable measure for arithmetic distributions.
- if data is not concentrated at central part median if no values

iii) Mode - value corresponding to peak of density curve is most freq occuring values.
- little significance in multimodel density.

iv) Mid Range - mid point in internal betn min & max value

$x^{midth} = x^{max} + x^{min}$

$x^{min} = x^{i} \max x^{i}$

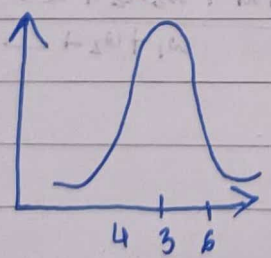
$x^{min} = \min x^{i}$

- not robust as effected by outliers.

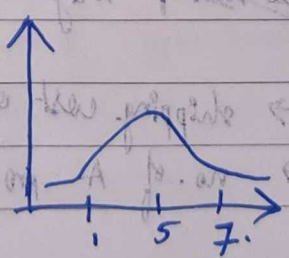
v) Geometric Mean - m^{th} root of product of m observations of attribute

$g^{geom} = m\sqrt{x_1 x_2 \cdots x_m}$

I d) Measures of dispersion for numeric attributes-
dispersion of data represent based of variability expressed by observation wrt to central values (location called in context)



less dispersion
(4, 6).



(1, 7).
more dispersion.

DOMS | Page No.
Date / /

i) Range = diff b/w 2 extreme obs. ∴ range = $x_{max} - x_{min}$

ii) Variance = measure of variability of data ∴ simple variance = $(\sigma^2) = \frac{\Sigma(x-\mu)^2}{n-1}$

pop variance $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$

iii) Standard Deviation (σ) (s) – positive spot of variance.

$$\sigma = s = \sqrt{\frac{(x-\mu)^2}{n-1}}$$

std	variance.
- measures of dispersion from mean.	- how far nos are from average
- sqrt of variance	- avg in diff from mean.
- spread pt^n nos in dataset	- avg degree to which each part diff from mean
- same unit of data	- squared unit as data or percent.
- $\sqrt{\frac{(x-\bar{x})^2}{n}}$	$= \frac{(x-\bar{x})^2}{n}$

iv) Mean Absoluted Deviation (Deviation or spread). $Si = \sum_{i=1}^{m} \frac{(x_i - \bar{x})}{m}$

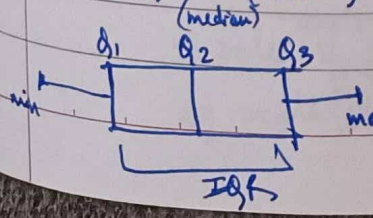
v) Quartile Deviation (LQR) $Q_3 - Q_1$

vi) Coefficient of variation.

$CV = 100 \cdot \frac{\sigma}{\mu}$ ration of same st. deviation to sample mean expressed at %

o) Identification of Outliers. 1. using z-index $z\ score = \frac{x-\bar{x}}{\sigma}$

2. box. plots (box & whiskers plot).



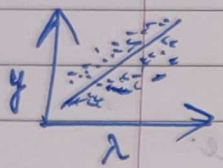
		Lower bound	Upper bound
Inner outliers	* 1.5	$Q_1 - (1.5 \times IQR)$	$Q_3 + (1.5 \times IQR)$
outer outliers	* 3	$Q_1 - (3 \times IQR)$	$Q_3 + (3 \times IQR)$

2] Bivariate Analysis -

- helpful in testing simple hypotheses of association.
- 3 distinguish cases that can occur.
  - i) both attribute is numeric
  - ii) one attribute to numeric one is categorized
  - iii) both attributes are categorical

b) Graphical Analysis -

i) Scatter Plot

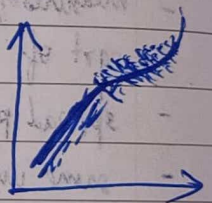


- intuitive representation of rel"ship b/w 2 numeric attributes
- pair of values as treated as pts & plotted on cartesian graph.

ii) loss curve (local regression)

- based on scatter plot, used for both numeric

states from scatterplot, adds a curve to assess ~~fractional~~ rel.



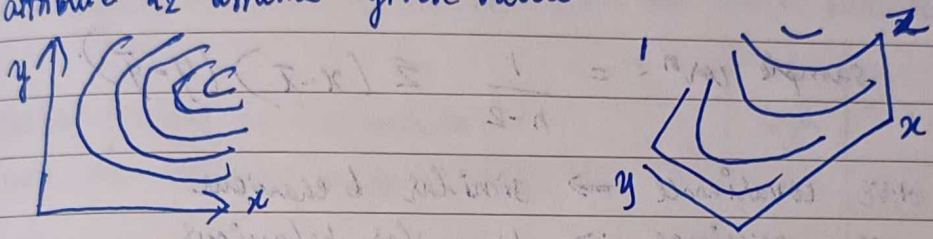
trend where can be attend obtained using local regression testing rel^n. curve is regulated by 2 parameters.

1) degree $\lambda > 0$ of polynomial that represents are usually 1 or 2 is assigned to λ.

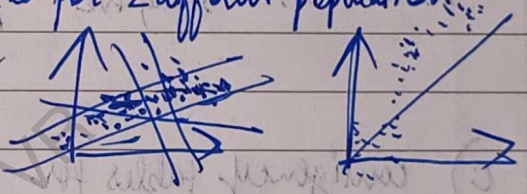
2) usually regularization constant of α 70 regulates size of neighbourhood if $\alpha \uparrow$, neighborhood $\uparrow$, more obs. used to calculate regression.

α in range of 1/4 to 1 is chosen normally

iii) Level Curves (identical to control lines as graph/maps):-
- further development of scatter plot again numeric attribute.
- highlight value by 3rd numerical attribute az as attribute aj & ak placed
- connecting to each other by points on plot share values of 3rd an excess of plot attribute obtains curved brem representating geometric locus of pride for attribute a_2 assumes given value.



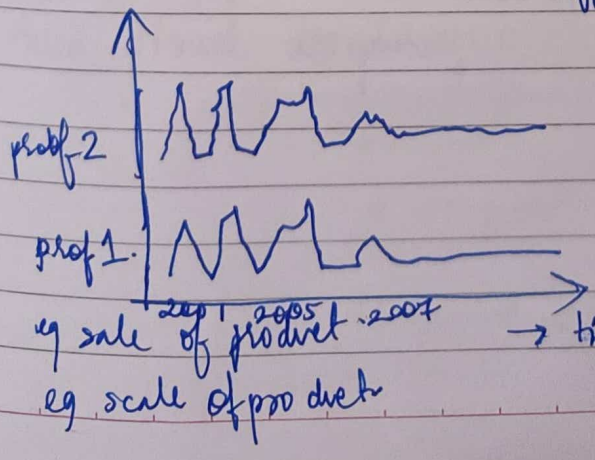
iv) QQ plots (Quantile - Quantile) -
- compares distribution of same attribute for 2 diff characteristics of population or for sample extracted for 2 different population.
- If samples are along same line



covariance	correlation
- extend to which 2 random variables change wrt to each other	- how strongly 2 random variables related
- means of correlation	- scaled from covariance
- $-\infty$ as to $+\infty$	- -1 to +1
- indicates direc" of linear relation	- directive of stream for linear relativity

v) Box Plot

vi) Time Series - collection - of same attribute relative different time instants



eg sale of product

eg scale of products

DOMS | Page No.
Date / /

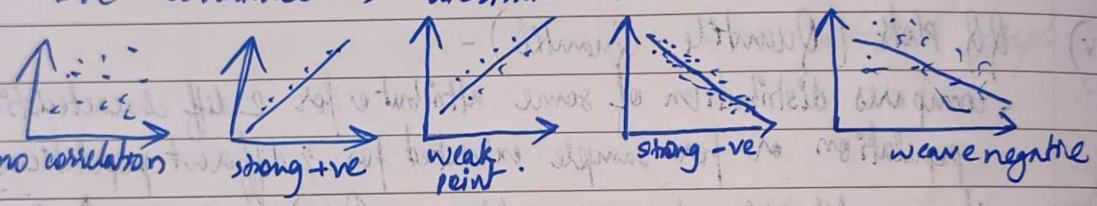
b) Measures of correlation for numeric attribute -
- covariance coefficient & covariance can be used to find variation b/w attributes.

i) Covariance -

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x})(y - \bar{y})}{n} = E\{(xy) - \bar{x}\bar{y}\}$$

$$\text{sample corr}^n = \frac{1}{n-2} \Sigma (x - \bar{x})(y - \bar{y})$$

+ve covariance → similar behaviour
-ve covariance → dissimilar behaviour



c) contigency tables for categorical attributes -
- tabular representation of categorical data, if represents frequencies for particular combination of values
- defined as a metric T whose generic elements T_{rs} indicate freq of pair of values $(x_{ij} = V_s) \wedge \{x_{ijk} = V_s)$

eg.

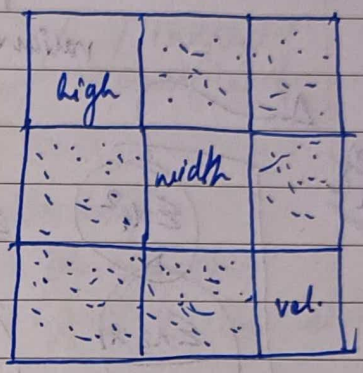
	Right hand	Lefthander
male	40	20
female	11	5.

- It summarizes into about data.

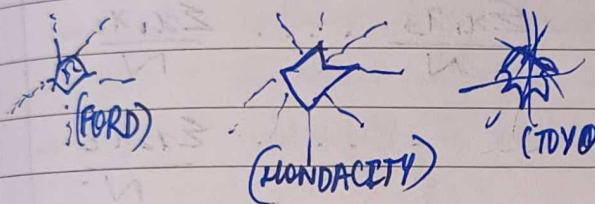
DOMS | Page No.
Date / /

3) Multivariate analysis
- extends bivariate analysis in order to assess relationship among multiple attributes in dataset

a) Graphical analysis -
i) matrix of scatter plot -
- roughly determines linear correlation b/w multiple variables.

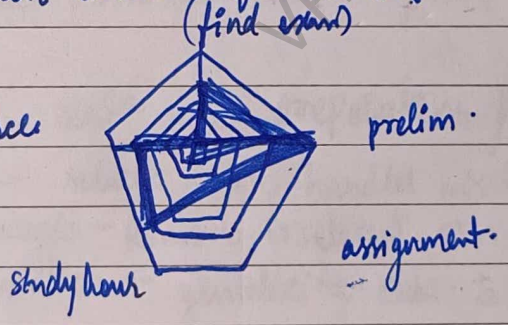


ii) starlist (rador chart or web chart)
cars are analyzed for some attributes.



iii) spired web charts (Rador charts)
- display data across unique elements.

eg-



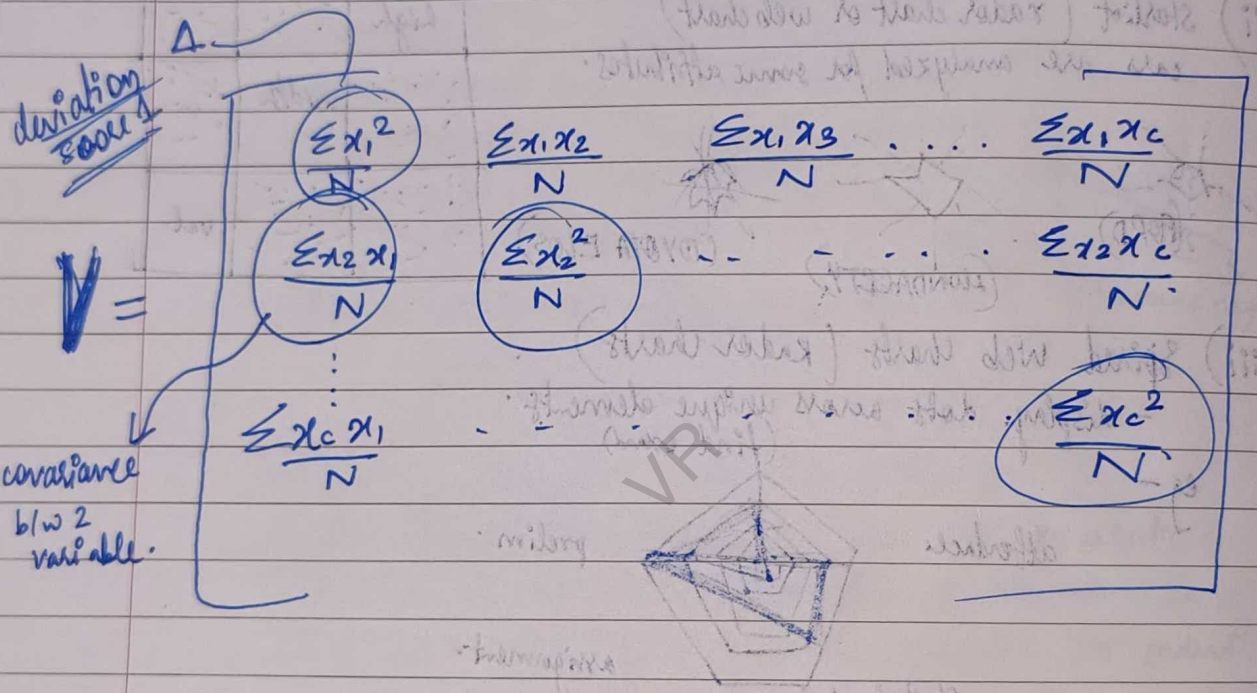
benefits -
1) make contractions of strength & deficiencies visible.
2) clearly display import categories
3) define full performance in each category.
4) valid visual decipher

DOMS | Page No.
Date / /

b) Measure of Correlation for Numerical Attributes –

Variance covariance matrix

- correlations & covariance matrix are calculated among pairs of attributes
- let V & R be 2nxn sq. matrix whose elements represent by covariance values & correlation values respectively.
- covariance matrix V contains on its diagonal sample covariance of each single value it called variance – covariance of matrix.



Univariate	Bivariate	Multivariate
• only summarize single variable at time	• only summarize two variables	• summarizes more than 2 variables
• don't deal with causes & relⁿ	• deal with cause & relⁿ & analysis is done	• doesn't deal with causes & relationships & analysis is done.
• doesn't contain any dependent variable	• does contain only one dependent variable	• similar to bivariate but it contains more than 2 variables
• main purpose is describe	• main purpose is to explain	• main purpose is to study relⁿ amongst them
• eg → height	• eg → temperature & ice scales in summer vacation	• eg → website views & relⁿ b/w variables