Ch IV.
Data Preparation.

Data Validation.

- process of ensuring that data is accurate, consistent & conforms to certain rules.
- essential step in data preparation ~~to main~~
- helps to identify errors, inconsistences & anomalies in data.
- If collected data is noisy & inconsistent, it is not validated.
- can be performed manually or by using automated tools.
- method depends on complexity & volume of data.

Incomplete Data.

- missing values in dataset
- common issue.
- occurs due to various reasons such as human error while data collection, technical issues or intentional ommissions.
- important step because it can affect the quality & reliability of analysis.
- occur when no data value is stored in an observation.
- ~~Reasons~~ :- How to overcome it :

i) +) Elimination
discard all records for which the values of one or more attributes / features is missing.

ii) Inspection.
inspection of each missing value, carried out by experts.

iii) Identification
conventional values used to encode & identify missing values, eliminating the need to delete the records.

iv) Substitution

Noisy Data.

- meaningless data which cannot be interpreted by machines.
- data that contains random, irrelevant, variations errors, inconsitencies.
- caused by various factors such as measurement errors, data entry mistakes, other sources of disturbance during data collection.
- crucial part in data preparation.
- ensure to accuracy & reliability of model.
- outliers in the data.
- Numerical values → Box & Scatter Plots
  ↗ anomaly detection.
- Methods:

i) Binning
- works on sorted data to smooth it.
- data is divided into equal parts & various methods are used to complete the task.
- each segment is handled separately.
- all data in a segment can be replaced by its mean or boundary values can be used to complete task.
- methods of binning → smoothing by bin means, medians, ~~mod~~ bound

ii) Regression
- data is smoothened by fitting it into regression function.
- regression may be linear or multiple.

iii) Clustering.
- groups similar data in clusters.
- outliers may be undetected or fall outside.
- quality in input not adequate.

Data Transformation.

- converting data from one format to another.
- data is transformed into form appropriate for mining.
- important aspect of data management.
- can be simple or complex depending on reqd. changes from source to final data.
- mixture of manual & automated steps.
- Process.

1) Data discovery.
2) data mapping
3) Code generation
4) Code Execution
5) Data review.

Types.

- Batch Data
- Interactive Data.

1) Batch Data.
- Batch process.
- developers write code in a data integration tool.
- That code is executed on large volumes of data.
- applied to data warehousing.

2) Interactive Data.
- give business analyst direct access & interaction with large data sets through visual interface.
- Power BI.

Standardization.

- also called z-score normalization.

- data preprocessing technique
- used to transform numerical data is a standard sc
- involves rescaling the data that it has zero mean & unit variance.
- easier comparison & interpretation of data across diff. scales.
- Process:

1) Calculate mean of feature (μ)
2) Calculate S.D. of feature (σ)
3) for each point (x) in the feature, subtract mean ($x-\mu$) divide it by S.D.

$$\left(\frac{x-\mu}{\sigma}\right).$$

- resulting transformed data will have mean = 0 & S.D = 1
- Popular methods: • decimal scaling
  • min-max
  • z-index.

Feature Extraction.

- meaningful & relevant info is extracted from raw data
- It is then transformed into set of derived features.
- swap out attribute values for derived values through tra
- involves selecting & combining the original variables of d to create new representation that captures underlying patte of data.
- Overcomes the challenges of high-dimensional, noisy raw data by reducing dimensionality, removing noise & highlight important aspects.
- Techniques used: 1) PCA.
  2) Feature selection
  3) Linear Discriminant Analysis.

PCA

- Principal Component Analysis.
- used for analysing huge data sets that contain large amount of dimensions per observation.
- improves interpretability while preserving max info.
- enables visualization.
- stat. technique to reduce dimensionality.
- linearly transform data into a new 'co-ordinate system'.
- converts observations of correlated features into set of linearly uncorrelated features with orthogonal transformation

correlated → uncorrelated (ortho. trans)

- technique to draw strong patterns from data by reducing variance.
- used in exploratory data analysis for making predictive models

Steps:

1) Standardize the dataset.
   - by calculating mean & S.D for each feature.
   - z-score is used.
2) Calculate covariance matrix.

for given dataset: covariance matrix will be calculated as

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

For 3 variables, (x, y, z)

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

3) Calculate eigen vectors & eigen values.
4) Sort eigen vectors & values from highest eigen value to low
5) Select the no. of principal components.
6) Transform the original matrix

Feature Matrix x top k eigen vectors = Transformed Data.

Data Reduction

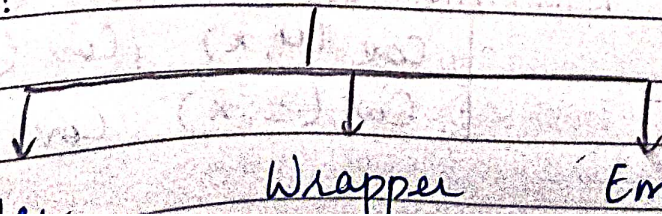
- process of reducing volume of dataset while preserving info.
- used where dataset is large & contains redundant info.
- dimensionality reduction
- select subset of data that are actually useful & relevant
- Binning

1) Feature selection.

- method of reducing I/P variable to your model ~~without~~
- done by using only relevant data & getting rid of noise.
- process of automatically choosing relevant features for ML model based on type of problem you are trying to solve.
- we do this by including important features without changing them..
- Role of feature selection.

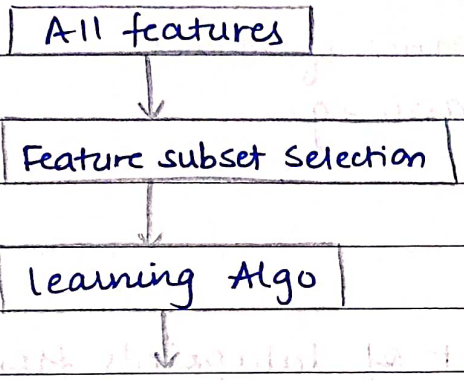
1) reduce dimensionality of feature space
2) speed up a learning algorithm
3) improve predictive accuracy of classification algorithm.
4) improve comprehensibility of feature selection.

- Types:



Filter.

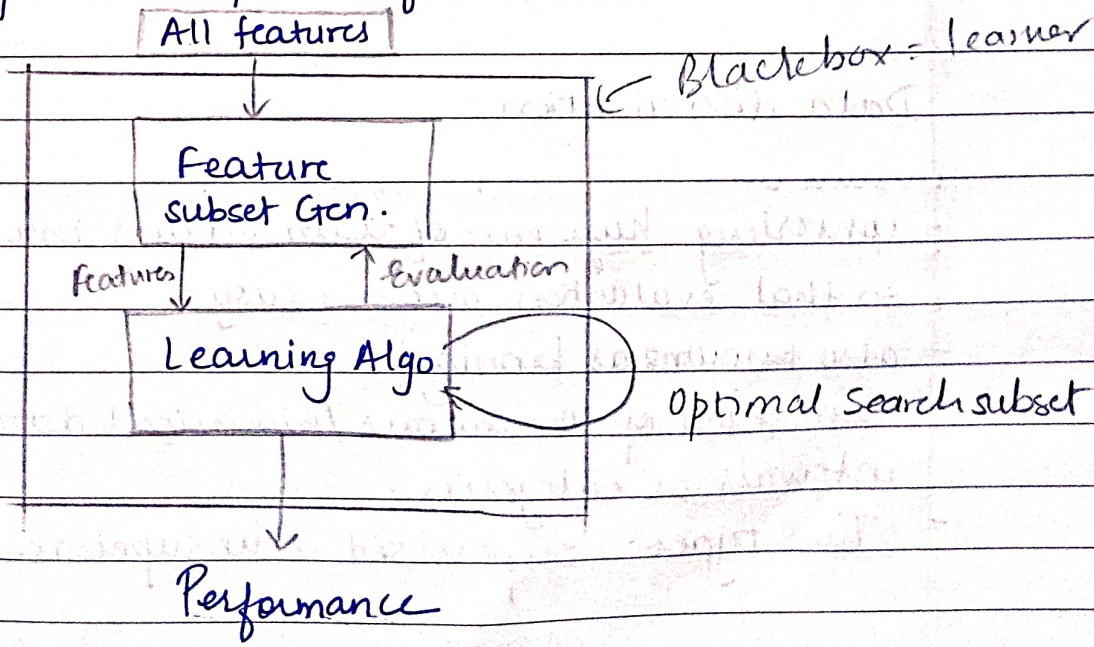
- generally used as preprocessing step.
- selection of features is independent of any ML algo.
- features are selected on the basis of their scores in various statistical tests for correlation with outcome variable.



- makes use of stat. methods to predict relationship between independent input & output variables, which assigns scores for each feature.
- CFS → ranks feature subsets according to correlation based heuristic evaluation function.

Wrapper

- learner is considered as a blackbox.
- interface of blackbox → used to score subsets of variables according to predictive power of learner



- feature selection Algo searches for a good feature subset.
- recursive feature elimination.

Embedded.

- similar to wrapper
- optimize performance of model.
- Random Forest, Lasso eg.

Sampling.

- selecting subset of datapoints from a large dataset.
- used in data analysis, research.
- picking random sample S from dataset D & search for frequent items in S.
- represents large dataset by a small random sample of data.
- methods - Random sampling, Cluster sampling etc.
- main goal is to obtain rep. sample that accurately reflects characteristics of population.
- allows researchers to draw conclusions about populations based on characteristics observed in the sample.
- based on classical inferential reasoning.

Data discretization.

- converting huge no. of data values into smaller ones so that evaluation becomes easy.
- also known as binning.
- conversion of continuous/numerical data into discrete intervals or categories.
- Two types: supervised & unsupervised.

- supervised : class data is used
- unsupervised : depends on way which operation proceeds.
- Eg. Data of ages of people : 5, 8, 9, 15, 25, 45, 60, 8, 10, 12, 3

Attribute	Age	Age	Age	Age
	3,5,8,9	10,12,15	25,45	60
After disc.	Children	Young	Mature	Old

- Techniques.

1) Binning. group huge amt of cont values into smaller values.
2) Cluster Analysis
clustering algo is executed by dividing values X into clusters.
3) Decision tree analysis.
4) Correlation analysis.
- linear regression used.
- supervised learning.

- Importance.

1) create heirarchy of concepts.
2) transform numeric data
3) ease evaluation & management of data
4) Minimize data loss.
5) produce better result

Heirarchy Generation for Categorical Data.

- Categorical values → discrete data.
- have finite no. of distinct values.
- Methods of heirarchy Generation

1) Partial ordering at schema level by users.
- eg. datawarehouse contains attributes : street, city, province, country
- user can easily form a heirarchy by specifying order of attribute
2) Explicit data grouping of portion of heirarchy.
- easily specify explicit groupings for small portion of intermediate level data.

Eg. After specifying area & country form heirarchy, user can define intermediate levels manually such as {India: Maharashtra: Pune} > SPPU.

3) Specification of set of attributes, not ordering.
- user specifies set of attributes but not their ordering.
- system then tries to automatically order it to construct meaningful concept.

Data Exploration.

- exploring datasets using statistical methods & graphic visualizations.
- done to gain better understanding of data.
- used to analyze & summarize datasets by using data visualization.
- used to uncover patterns in data & study them.
- looks for similarities, patterns & outliers to identify relationships between variables.
- creating a mental model of data & understanding relation between variables
- Phases:

1) Univariate (every attribute properties)
2) Bivariate (pairs of attributes & relationship bet^n them)
3) Multivariate (relationships within subset of attributes)

Univariate Analysis.

- simplest form of statistical analysis.
- can be descriptive
- only one variable is involved.
- explores each variable in the dataset separately.

- observes range & central tendencies of values.
- technique of comparing & analysing dependency of single predictor.
- does not deal with relationships & causes.
- takes data & provides summary & patterns.
- Steps:

1) Accessing dataset
2) Identify the variable that needs to be analysed
3) Identify questions to be answered through analysis.
4) Determine appropriate. U.A. technique to answer questions.

- dataset contains heterogeneous values & datatypes such as text, number, date, logical etc.
- log. graphical rep → starting point for analysis.

i) Graphical Analysis for Categorical Attributes.

- Category.
- has a specific value from limited selec. of values.
- natural rep. of categorical attributes → vertical barchart.
- Ways:

a) Histogram (Numerical Data)

- used to chart continuous frequency distribution.
- consists of erecting series of adjacent vertical rectangles on x axis.

bases = width of class interval
heights = area of rectangle = frequency of class.

- variate values → X axis
  frequencies → Y axis.

i) equal classes (classes of equal mag).

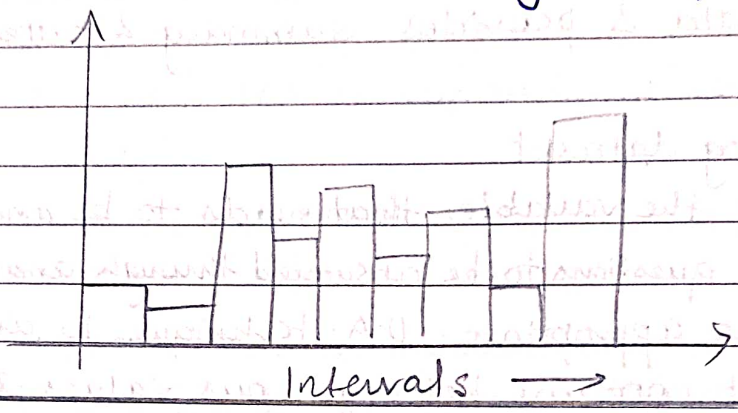
- class interval → x axis by section = mag. of class interval
- on each class interval erect a rectangle with height equal to corresponding frequency of class

Bar charts → categorical Data.

ii) Unequal classes.

- if classes are not uniform, different classes are rep. on x axis

Frequency of Density class = $\frac{\text{Freq of class}}{\text{Magnitude of class.}}$

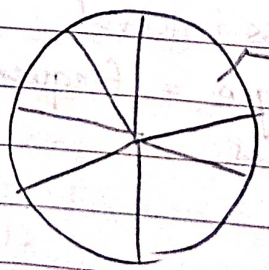


b) Pie-chart. (categorical Data).

- circle can be divided into parts to represent portions of different categories.
- such divided circle is called pie diagram.
- entire circle = 100% of data
  portion = Specific category.
- Steps.

1) Express each comp. value as % of total
2) Angle at centre is 360°, total magnitude of components = 360°.
3) Degree calculation

$$\frac{\text{component value} \times 360^\circ}{\text{Total value}}$$



Measures of Central Tendency.

- arithmetic measure which gives central value of set of observations.

1) Mean.

- average of data values.

$$\text{Sample mean} = \frac{\text{Sum of n obs}}{\text{No. of obs.}} = \frac{\Sigma x_i}{n}.$$

2) Median.

- value in middle of data set when it is arranged in A.O.
- when dataset has extreme values, median is prefered measure of central tendency.
- for odd obs.

7 observations → 26, 18, 27, 12, 14, 29, 19

12, 14, 18, 19, 26, 27, 29

19 → median.

- for even obs.

8 observations 7, 9, 5, 6, 1, 4, 12, 4.

1, 4, 5, 6, 7, 9, 12.

$$\text{Median} = \text{Avg. of mid. values} = \frac{6+7}{2} = \frac{13}{2} = 6.5.$$

3) Mode.

- value that occurs with greatest frequency.
- greatest freq. can occur at 2 or more values.
- 2 modes → bimodal
- more than 2 → multimodal.

4) Range.

- difference between highest & lowest values in a set.

Measures of Dispersion of Data.

5) **Variance.**

- difference bet^n value of each obs. (x) & the mean($\bar{x}$) squared.

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

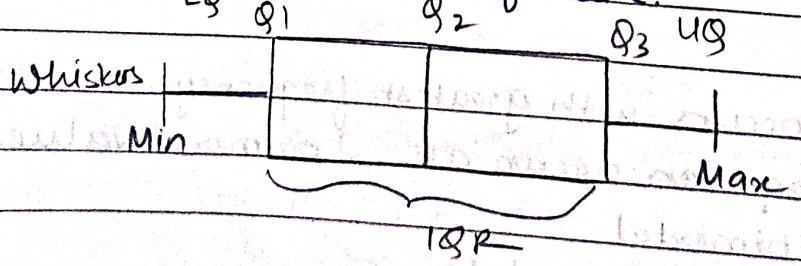
6) **Standard Deviation**

- positive square root of variance.
- represented by σ.

$$S.D = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} \quad / N.$$

Identification of Outliers:

- outliers: observation that lies on abnormal distance from other values in a distribution.
- Q1 - contains 25% of values ~~are~~.
  25% percent of values are smaller than Q3 & 75% are larger.
- Q2 - contains 50% of measurement
- a way to detect outliers is by using Box plots.
- Box-whisker plots - used for exploratory analysis.
  shows distribution of data.



Q2 is the median.

Upper Quartile - 75% score lies below UQ.

low Quartile - 25% score lie below LQ.

Median - midpoint.
Max. score - highest score including whiske outliers
Min score - lowest score " —— . —— "
Whiskers - represent score outside middle 50%.

Bivariate Analysis.

- one of the simplest forms of quant. analysis.
- involves analysing 2 variables for determining the relationship between them.
- to what extent does it become easier to predict a value of variable (dependent) if we know the other variable (independent).
- Types:

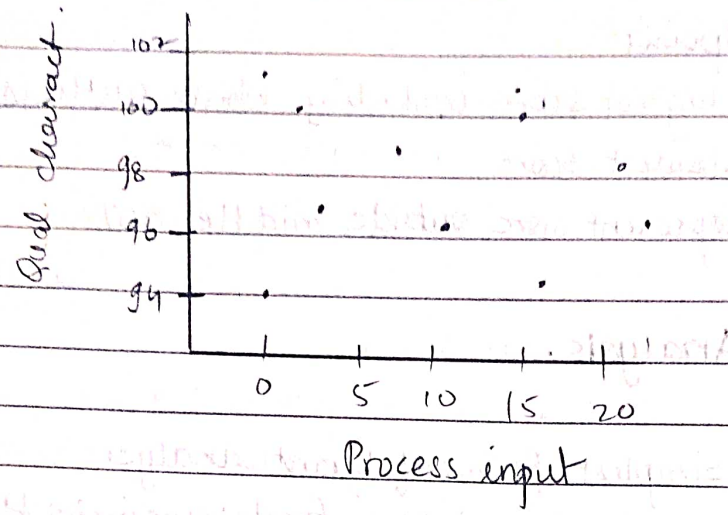
1) Numerical & Numerical. both are numerical values.
2) Categorical & Categorical both are categorical values.
3) Numerical & Categorical one is num & other is categorical

Graphical Analysis.

- for graphical analysis, it is imp. to know the type of variable.
- numerical variables - scatter plot.
- one categorical & other numerical - Box plot
- categorical variables - masaic plot.

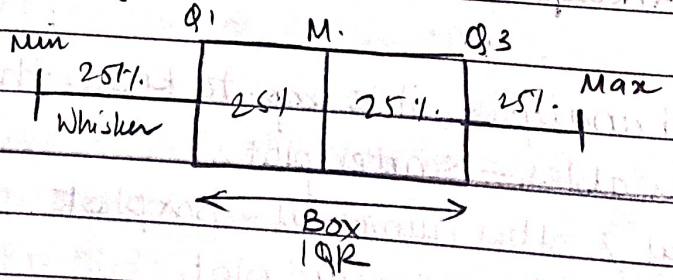
1) Scatter plot.

- displays value for 2 variables for a set of data.
- tool for analyzing relationship between 2 variables.
- one variable plotted on x axis & other on Y axis.
- pattern of intersection of points show relationship patterns.
- used to prove cause-&-effect relationships.



2) Box Plot.

- also called whisker plot
- shows the five number summary of data.
- 5 no. summary - max, min, Q1, Q3, median.
- we draw a box from Q1 to Q3.
- a vertical line goes through the box at median.
- exhibits group of data through their quartiles.
- shape of data can be visualized easily.
- comp. betn categories is easier.



- Minimum : lowest value including outliers
- Maximum : highest value " "
- Q1 : 25% data lies below Q1
- Q3 : 75% data lies below Q3.
- Median : mid point of dataset.

$$IQR = Q3 - Q1.$$

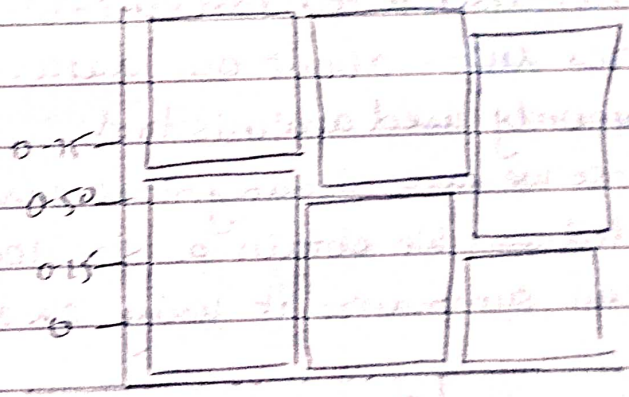
$$L.L = Q1 - 1.5\,IQR$$

$$U.L = Q3 + 1.5\,IQR$$

values below & above UL & LL are outliers.

3) Mosaic Plot:

- graphical rep
- allows visualization of joint distribution of categorical variables.
- similar to stacked bar chart.
- useful while analyzing contingency tables.
- visual representation of relationships / dependence bet" variables.
- help to identify patterns in data.
- Plot consists of ▭ tiles with w & h ∝ to freq. in contingency table.
- The tiles are vertically stacked to form columns.
- use colors to represent additional variables.



Measures of Correlation for Numerical Attributes.

- exactly 2 measures are made on each observation.
- 2 measurements are X & Y.
- Since X & Y are obtained from each observation, data for each obs. is (X, Y).
- Correlation - relationship between 2 or more objects/variables.
  - exists if one variable is related to other in some way.

Eg. V1 → no. of hunters
V2 → deep population
} as V1↑, V2↓.
negative correlation.
one ↑ other ↓.

- +ve correlation: 2 variables react in the same way.
  ↑ or ↓ together.
  Temp. → C & F has +ve correlation.
- correlation → strength of association bet^n two variables.
- <u>covariance.</u>
  extent to which change in one variable corresponds systematically to a change in another.
- correlation is standardised covariance

Contingency Tables for Categorical Attributes.

- ~~co~~ tabular representation of categorical data.
- usually shows frequencies for particular combinations of values of two discrete random variables.
- summarizes info. about our data.
- most commonly used analysis tool.
- Eg. Suppose we have 2 categorical variables: Gender & Handedness. We obtain a size 100 data.
  We can summarize it using 2 x 2 table

	Right handed	Left Handed	Total
Male	43	9	52
female	44	4	48
Total	87	13	100

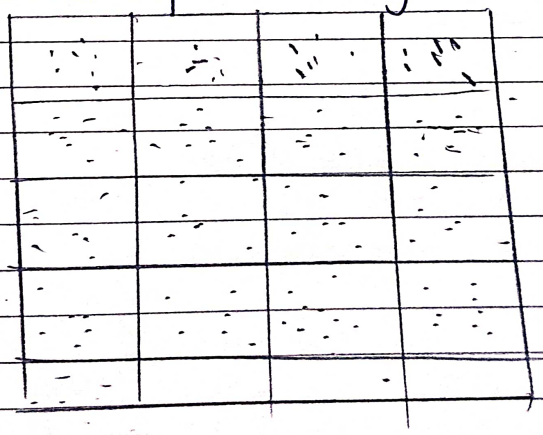
Multivariate Analysis.

- statistical analysis of multiple variables simultaneously to understand relationships, patterns & interactions among them.
- involves checking joint behaviours of elements to gain insights.

- wide range of stat techniques for analyzing complex datasets.
- commonly used techniques:
  1) PCA
  2) Factor Analysis
  3) Cluster analysis.
  4) Regression.

Graphical Analysis.

1) Scatter Plot Matrix
- we can look at all rel^ bet^ pairs of variables in group of plots
- describe relationship among 3 or more variables.
- ~~use~~ matrix of scatter plots
- plots all pairwise scatter between different variables.
- for K variables in data set,
  matrix → K rows & K columns
  each row & column represent single scatter plot.



2) Spiderweb Chart.
- represent multivariate data in 2D chart of 3 or more quant. variables.
- useful for rating item along 3 ~~axis~~ or more axes.

- One or more groups of values over similar / multiple common variables.
- do this by giving 1 axis - each variable.
- axis are arranged radially around a central point.