

The background of the slide is a light green color with a repeating pattern of watermelon slices. Each slice is a quarter of a watermelon, showing a red interior with black seeds and a green rind. The slices are arranged in a grid-like pattern, with some slices partially cut off at the edges.

data analytics

UNIT-1

Exploratory Data Analysis and Visualization

feedback/corrections: vibha@pesu.pes.edu

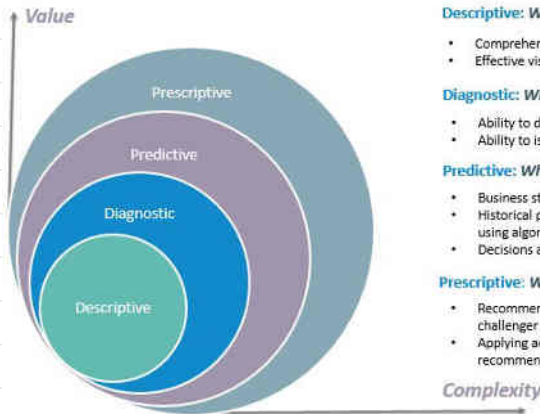
VIBHA MASTI

© vibhas notes 2021

Data Analytics

- 4 levels of analytics

4 types of Data Analytics



What is the data telling you?

Descriptive: What's happening in my business?

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: Why is it happening?

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: What's likely to happen?

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: What do I need to do?

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

- Data analytics life cycle

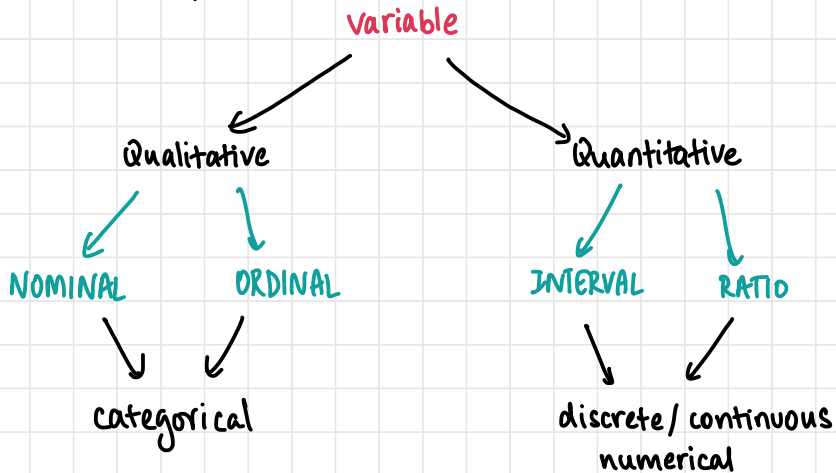


Data Sources

- Web data
- Transactions

Data

- Collection of attributes and objects
- Attributes: columns
Objects/records: rows
- Types of attributes
 - Nominal**: no ranking/order
 - ID, eye colour, zip codes
 - Ordinal**: order matters, no absolute difference between values
 - rankings, grades, height (tall, medium, short)
 - Interval**: differences meaningful, no absolute zero
 - calendar dates, °C, °F
 - Ratio**: all mathematical operations allowed
 - Kelvin, elapsed time



DATA REPRESENTATIONS

1. Structured

- described in a matrix/data structure format
- relational databases

2. Unstructured

- no fixed structure for the data
- documents, tweets, videos

3. Semi-Structured

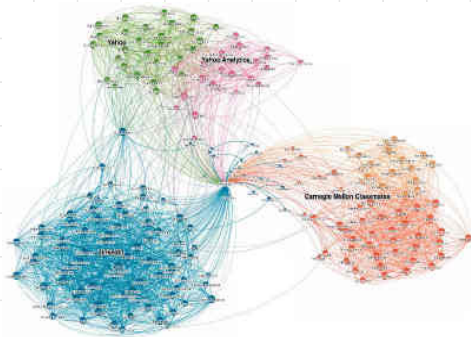
- combination of the two
- emails, XML

(a) Record

- Relational records
- Document data

(b) Graph and Network

- WWW
- Social networks



(c) Ordered

- video: sequence of images
- temporal data: time series
- sequential data: transactions, genetic sequence © vibhas notes 2021

(d) Spatial, Image and Multimedia

- Spatial data: maps
- Images and video

(e) Record data

- Collection of records (rows)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(f) Data Matrix

- Data objects are points in multi-dimensional space
- Each dimension one attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

(g) Document Data

- Each term is an attribute
- Value: number of occurrences of word in document
- Similarity between documents: difference in word occurrences (mod or squared)

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

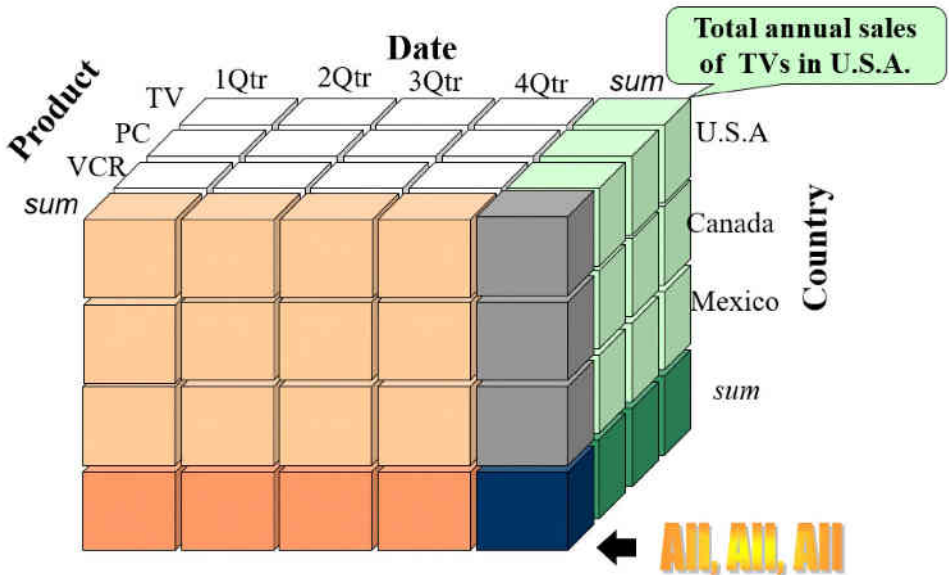
(h) Transaction Data

- Special kind of record
- Each record is transaction — set of items
- Eg: grocery store

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

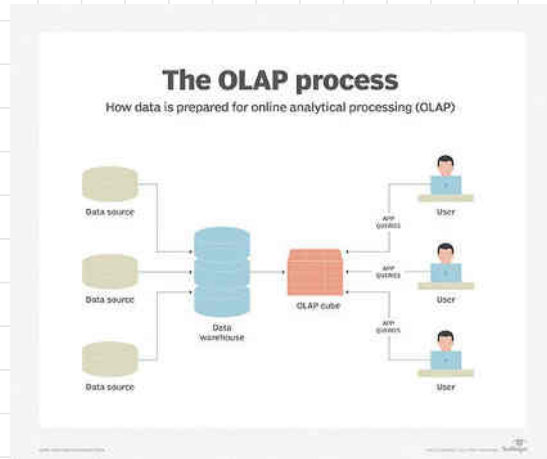
(i) Data Warehouse

- Subject-oriented, integrated, time-variant and non-volatile collection of data
- Specific collection of data for a domain
- Also called **data cube**



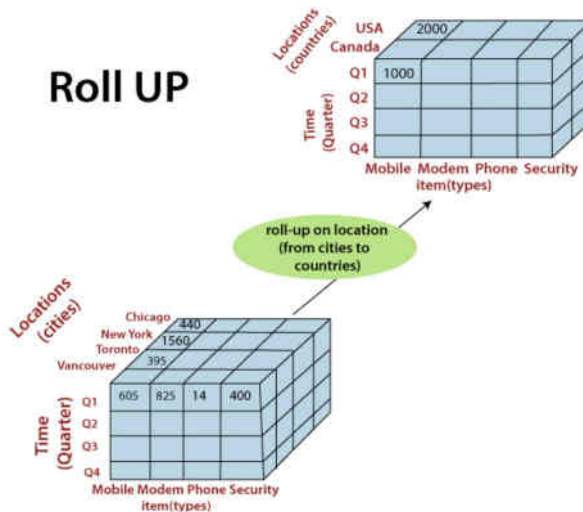
Typical OLAP (Online Analytical Processing) Operations

On data cubes



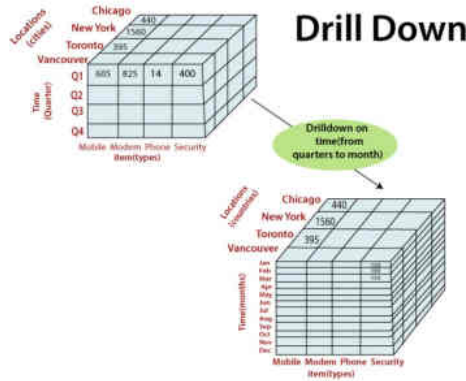
1. Roll up (drill up)

- Summarise / aggregate data by climbing up hierarchy
- Dimension reduction
- Eg: aggregate from day \rightarrow week \rightarrow month \rightarrow year



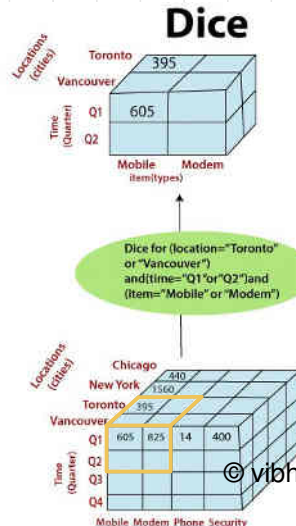
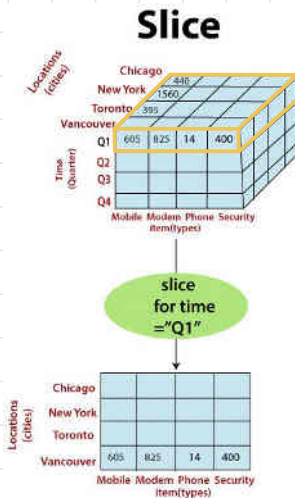
2. Drill down (roll down)

- From higher level summary to lower level summary
- Introduce dimensions and detail
- Finer granularity
- Eg: maps at continent → country → state → city → street levels
- Replace image with more detailed image based on level of zoom
- Must have data for finest granularity; cannot generate data down the ladder



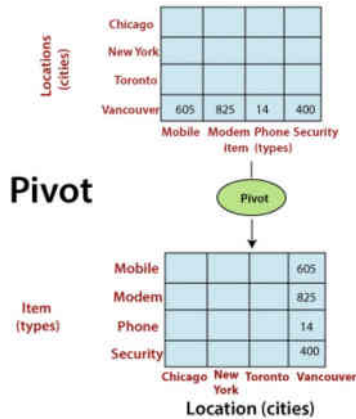
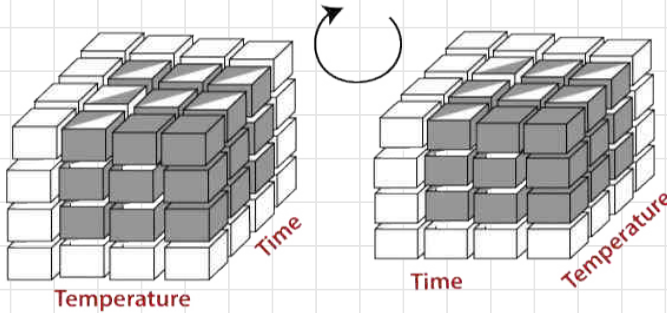
3. Slice and dice

- Slice the cube
- Project and select



4. Pivot (rotate)

- Reorient the cube without changing it
- 3D to series of 2D planes



5. Drill across

- Involving more than one fact table / data cube

6. Drill through

- Drill down through bottom of the cube to its backend relational tables
- DB queries (SQL)

Types of Data Measurement Scales

1. Nominal Scale

- qualitative data
- categorical variables

2. Ordinal scale

- value of data from ordered set
- numerical or categorical

3. Interval scale

- value chosen from interval set
- no absolute zero
- Eg: temp in Celsius, IQ

4. Ratio scale

- ratios can be meaningfully computed
- absolute zero

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male, female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Interval vs Ratio

Features	Interval scale	Ratio scale
Variable property	All variables measured in an interval scale can be added, subtracted, and multiplied. You cannot calculate a ratio between them.	Ratio scale has all the characteristics of an interval scale, in addition, to be able to calculate ratios. That is, you can leverage numbers on the scale against 0.
Absolute Point Zero	Zero-point in an interval scale is arbitrary. For example, the temperature can be below 0 degrees Celsius and into negative temperatures.	The ratio scale has an absolute zero or character of origin. Height and weight cannot be zero or below zero. (Zero means 'nothing'.)
Calculation	Statistically, in an interval scale, the arithmetic mean is calculated.	Statistically, in a ratio scale, the geometric or harmonic mean is calculated.
Measurement	Interval scale can measure size and magnitude as multiple factors of a defined unit.	Ratio scale can measure size and magnitude as a factor of one defined unit in terms of another.
Example	A classic example of an interval scale is the temperature in Celsius. The difference in temperature between 50 degrees and 60 degrees is 10 degrees; this is the same difference between 70 degrees and 80 degrees.	Classic examples of a ratio scale are any variable that possesses an absolute zero characteristic, like age, weight, height, or sales figures.

Operations on Data

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., $new_value = f(old_value)$ where f is a <u>monotonic function</u> . always	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Interval	$new_value = a * old_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$new_value = a * old_value$	Length can be measured in meters or feet.

monotonically increasing: if $x_1 > x_2$, $f(x_1) \geq f(x_2)$

DATA TYPES

1. Cross-Sectional Study

- data from many variables of interest at the same time

2. Time-Series Data

- data collected for single variable over several intervals
- longitudinal study: same parameter over multiple timestamps (similar)

3. Panel Data

- data collected for multiple variables (dimensions) over several time intervals
- also called longitudinal
- Eg: panel of health panel tests

Exploratory Data Analysis

- Preliminary exploration of data
- Summary statistics, visualisation

Summary Statistics

- Frequency, location, spread
- Location - mean, spread - standard deviation
- Calculated in a single pass through data (usually)

TYPES of DESCRIPTIVE STATS

1. Organise Data

(a) Tables

- frequency distributions
- relative frequency distributions

(b) Graphs

- bar chart, histogram
- stem & leaf plot, box & whisker plot
- frequency polygon

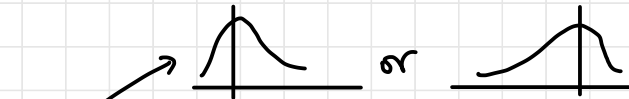
2. Summarising Data

(a) Central Tendency

- mean (pop: μ , sample: \bar{x}) $\mu \triangleq E[x]$
- median
- mode (unimodal, multimodal)
- percentile

(b) Variation

- range
- IQR
- variance
- standard deviation
- skew $\triangleq E[(x-\mu)^3]$
- kurtosis $\triangleq E[(x-\mu)^4]$



MEAN

Sample

Population

Symbol

\bar{x}

μ

Formula

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x$$

VARIANCE

Sample

Population

Symbol

s^2

σ^2

Formula

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Bessel's correction

PERCENTILE

$$P_x \approx \frac{x(n+1)}{100} \quad \text{for } x^{\text{th}} \text{ percentile}$$

Normal DISTRIBUTION

$$X \sim N(\mu, \sigma^2)$$

pdf:
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Any value is $\sim 68\%$ likely to be within 1 standard deviation of the mean $(\mu - \sigma, \mu + \sigma)$
- $\sim 95\%$ likely to be within 2 standard deviations of the mean $(\mu - 2\sigma, \mu + 2\sigma)$
- $\sim 99.7\%$ likely to be within 3 standard deviations of the mean $(\mu - 3\sigma, \mu + 3\sigma)$

Multidimensional - covariance matrix

$$\text{variance} = \Sigma = E((\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T)$$

Σ = covariance matrix (for multivariate Gaussians)

For 2 dimensions

$$\Sigma = E\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix}\right)$$

$$\Sigma = E \begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) \\ (x_1 - \mu_1)(x_2 - \mu_2) & (x_2 - \mu_2)^2 \end{bmatrix}_{2 \times 2}$$

$$= \begin{bmatrix} E(x_1 - \mu_1)^2 & E((x_1 - \mu_1)(x_2 - \mu_2)) \\ E((x_1 - \mu_1)(x_2 - \mu_2)) & E(x_2 - \mu_2)^2 \end{bmatrix}_{2 \times 2}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

If x_1 & x_2 are statistically independent,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Chebyshev's Theorem

Probability of finding a randomly selected value in an interval $\mu \pm k\sigma$ is at least $1 - \frac{1}{k^2}$

- Q: Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000?

$$\bar{x} = 12000 \quad s = 2000$$

$$\bar{x} \pm 2s \quad \Rightarrow \quad 1 - \frac{1}{2^2} = \text{at least } \frac{3}{4}$$

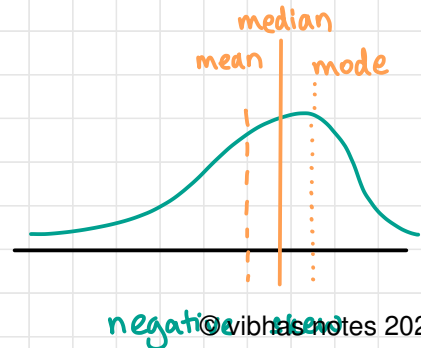
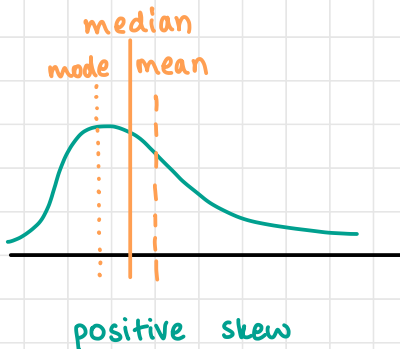
degrees of freedom

- No. of independent variables in the model
- If we are to construct a set of n numbers with a given mean, we have the freedom to choose $n-1$ of them
- The n^{th} one must be fixed (dependent on the other $n-1$ numbers)
- Degrees of freedom = $n-1$
- If there are n observations in the sample and k parameters estimated from it, there are $n-k$ degrees of freedom

measures OF SHAPE

1. Skewness

- measure of symmetry in a dataset (3rd moment)
- Symmetric dataset: equal proportion of data lying in the intervals $(\mu - k\sigma, \mu)$ and $(\mu, \mu + k\sigma)$, where k is some positive constant



- Pearson's moment coefficient for skewness for a dataset with n observations is (σ can be estimated with s)

$$g_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n\sigma^3}$$

- If g_1 is positive, it indicates positive skewness and if it is negative, it indicates negative skewness
- Following formula usually used for a sample with n observations

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

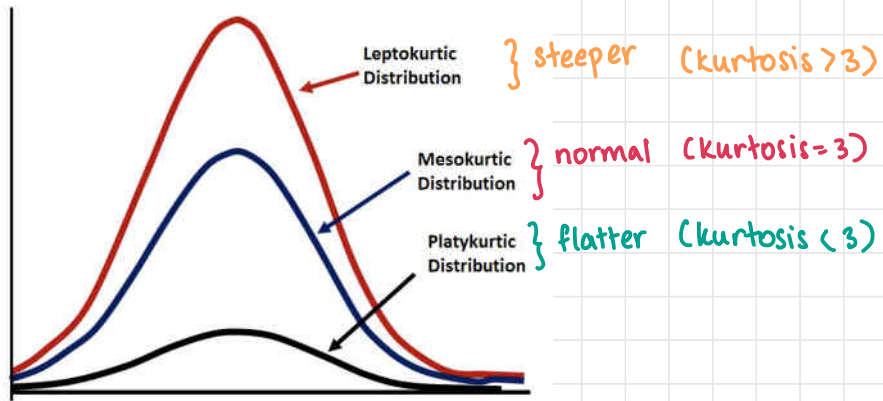
- As n increases, $\frac{\sqrt{n(n-1)}}{n-2}$ converges to 1 and the value of $G_1 \rightarrow g_1$

2. Kurtosis

- Measure of shape of the tail (4th moment); whether the tail of the data is heavy or light
- Measured by equation (σ can be estimated with s)

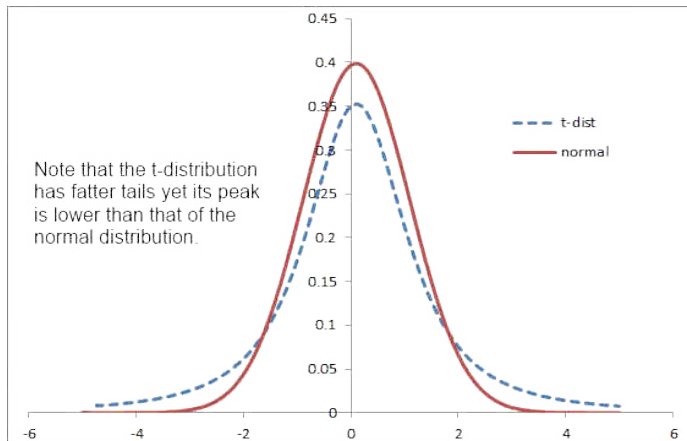
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n\sigma^4}$$

- Types of kurtosis



source: datavedas.com

- Difference between variance and kurtosis: kurtosis tells us how much of the "weight" of data lies in the tails, as opposed to the middle of the distribution
- kurtosis generally measured to compare curves with same mean and variance



source: riskprep.com

EXCESS KURTOSIS

- How much more kurtosis present in a graph in comparison to a normal distribution
- Leptokurtic: negative and Platykurtic: positive

$$\text{Excess kurtosis} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n\sigma^4} - 3$$

Population & Sample

- Population: set of all possible observations for a given context of a problem
- Sample: representative subset of the population

types of sampling

(a) PROBABILISTIC SAMPLING

1) Simple random sampling

- assign numbers to members of population and select randomly
- with or without replacement (finite/infinite population)
- good for small population

ADVANTAGES

easier, low error, no prior information required

DISADVANTAGES

can be biased, not proportionate, hard to scale

2) Stratified random sampling

- population proportion reflected in sample
- divide population into strata/groups (gender, hair colour, age etc)

ADVANTAGES

enhanced representation, more scalable and efficient

DISADVANTAGES

classification error, time consuming, expensive

- **example:** a student council surveys 50 students by getting random samples of 25 juniors and 25 seniors

3) Systematic Sampling

- Find the k^{th} value

ADVANTAGES

easy to select, evenly spread sample, cost effective

DISADVANTAGES

biased, no equal chance, ignored elements

- **example:** a principal takes an alphabetised list of students and picks every fourth student from a random starting point
- **example:** time series analysis, sampling pixels from an image

4) Cluster Sampling

- Population divided into non-overlapping areas (clusters)
- Each cluster microcosm of population

ADVANTAGES

convenient for geographically dispersed populations, simplified administration, economical, feasible

DISADVANTAGES

less efficient statistically, higher sampling error, more problems

- **example:** airline company randomly selects 5 flights and surveys everyone on them

(b) NON-PROBABILISTIC SAMPLING

1) Convenience / Accidental

- subjects for sampling easily available
- when population not clearly defined

ADVANTAGES

easy to select, saves time and money

DISADVANTAGES

biased, sampling errors, cannot generalise

2) Judgemental Sampling

- researcher chooses / is related to sample based on their judgement

ADVANTAGES

minimum time

DISADVANTAGES

selection bias, sample size

3) Quota Sampling

- non-probability equivalent of stratified
- till quota is met

ADVANTAGES

minimum time

DISADVANTAGES

bias

4) Snowball Sampling

- for rare characteristic / difficulty
- from initial subject, referrals

ADVANTAGES

lowers cost

DISADVANTAGES

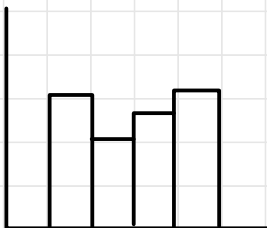
bias

- Note: **sampling in DSP** — downsampling vs upsampling

— DATA VISUALISATION

1. Histogram

- Lose data while aggregating



review: bin size

$$h = 2 \frac{IQR(x)}{\sqrt[3]{n}}$$

2. Stem- and-Leaf Plot

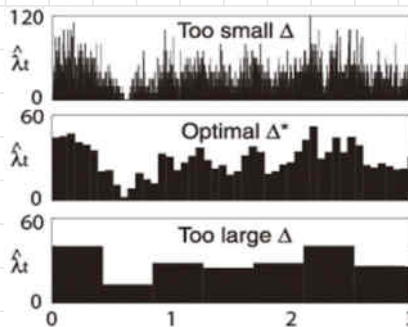
- Like a histogram on its side with no loss of information

stem	leaf	
1	2 3	→ 12, 13
2	1 7	→ 21, 27
3	3 4 5 7	→ 33, 34, 35, 37
4	0 0 1	→ 40, 40, 41

↑ ↑
tens' place ones' place

HISTOGRAM

- Optimal bin size important

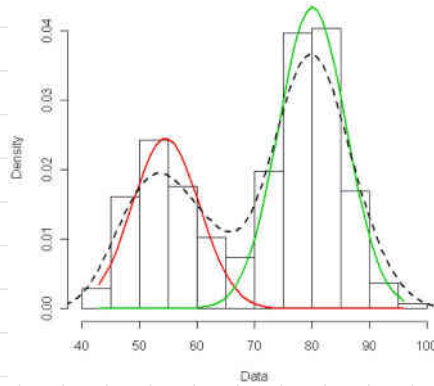


<http://toyozumilab.brain.riken.jp/hideaki/res/histogram.html>

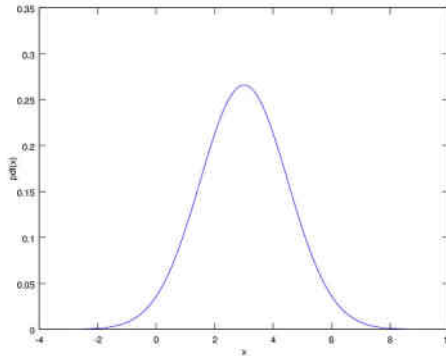
- Helps analyse shape

Bimodal Function

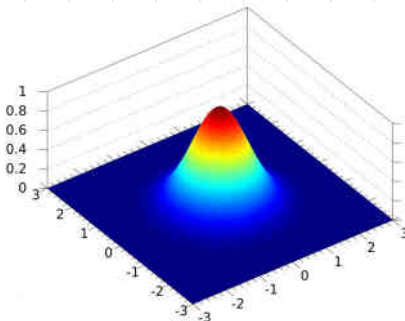
- Mixture of 2 Gaussians



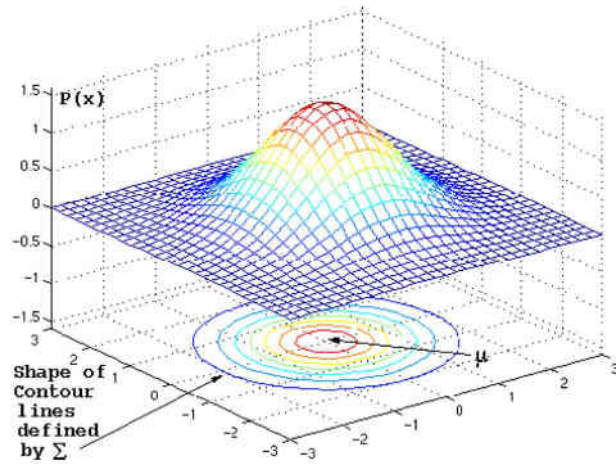
1. Univariate Gaussian



2. Bivariate Gaussian

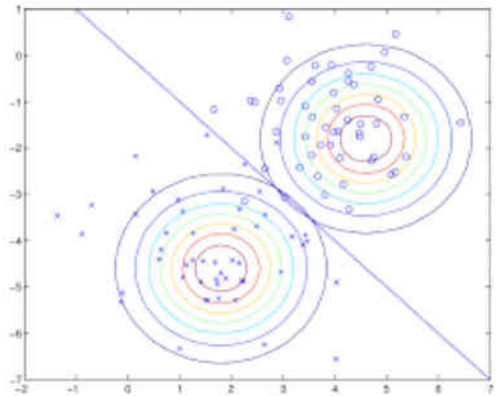
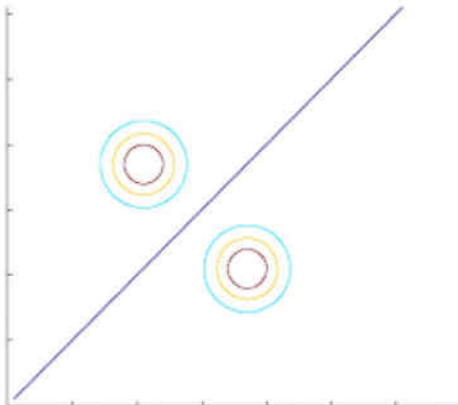


- Sections of bivariate Gaussian

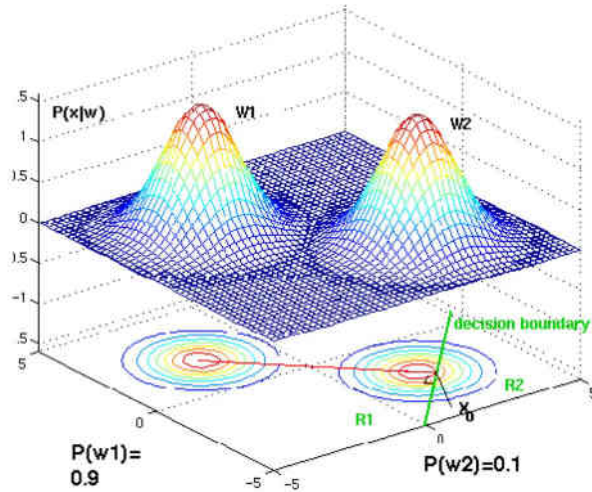


Boundaries in Higher Dimensions

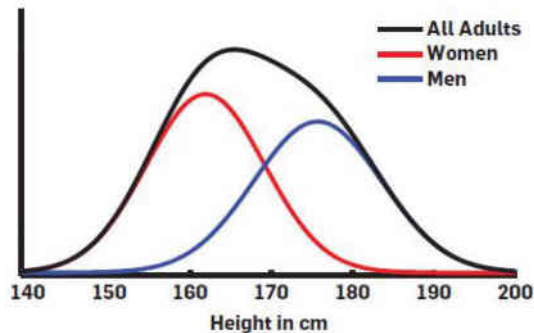
- Equiprobable



- Not equiprobable



Indistinguishable Mixture



- Read more on slides — 07 Visualization and R packages

TESTS for BIMODALITY

- Necessary condition for bimodal functions
 - **Pearson's criterion**: $kurtosis - (skewness)^2 \leq 1$
 - Equality holds in extreme cases of two diracs
- To check if distribution is anything but unimodal
 - **Hartigan's Dip Test Statistic**
 - $p\text{-value} < 0.05$: significant multimodality
 - $0.05 < p\text{-value} < 0.10$: multimodality with marginal significant
 - **R**: mixtools, flexmix, mclust, mcclust

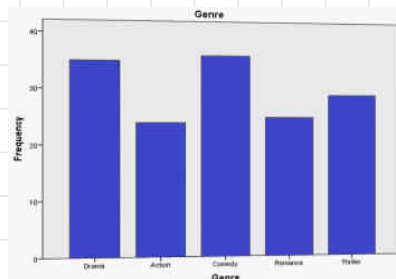
Ogive Curves

- Cumulative histograms are called ogive curves
- Eg: covid19india.org cumulative cases

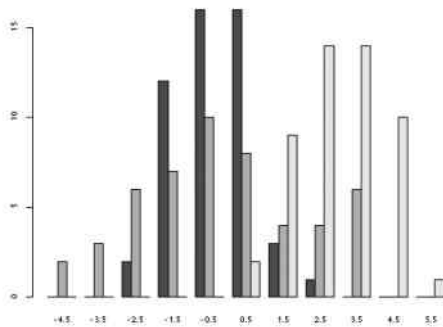
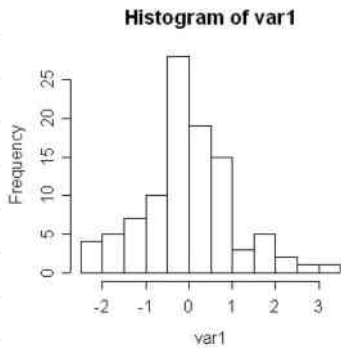


Bar Chart

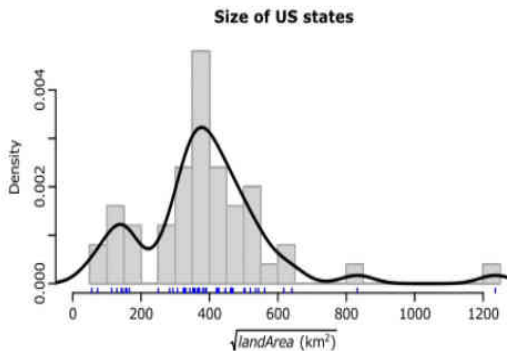
- Histograms for categorical data



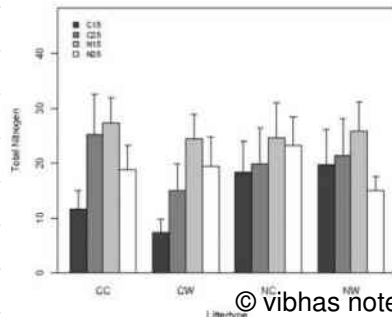
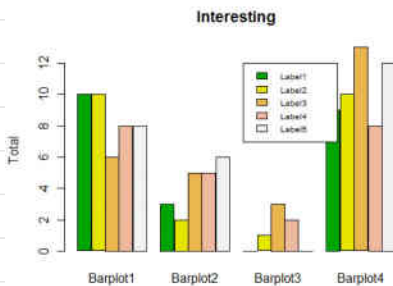
hist and multihist



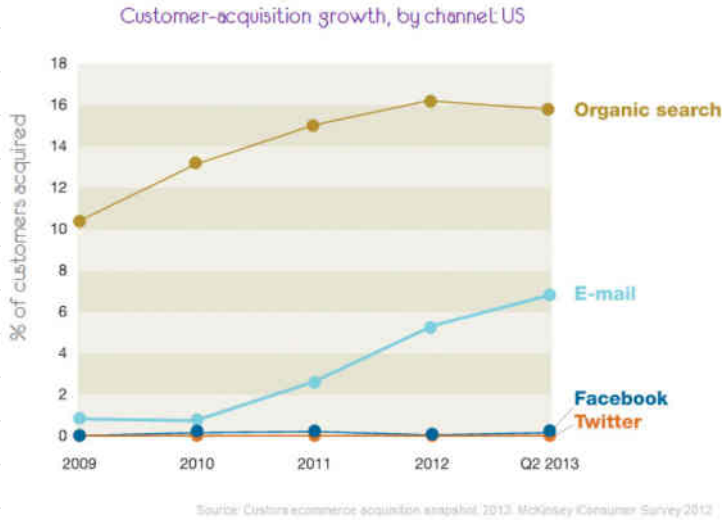
hist and density



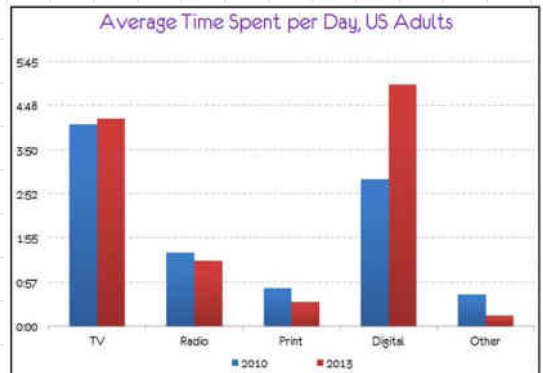
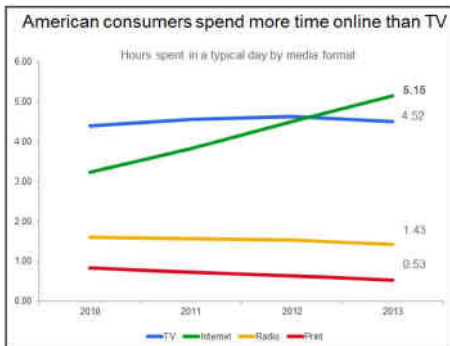
barplot



Labels and Legends

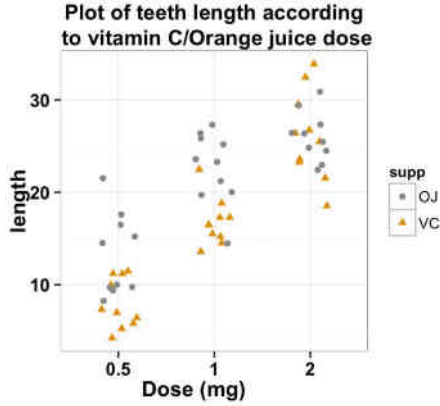


Horizontal vs Vertical

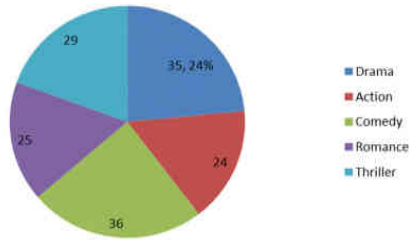


- Review SDS from sem 3 for more

stripchart

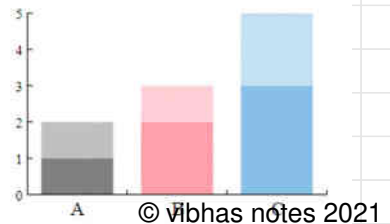
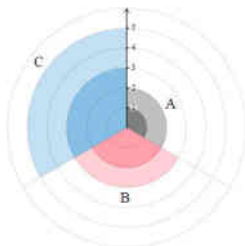


pie chart

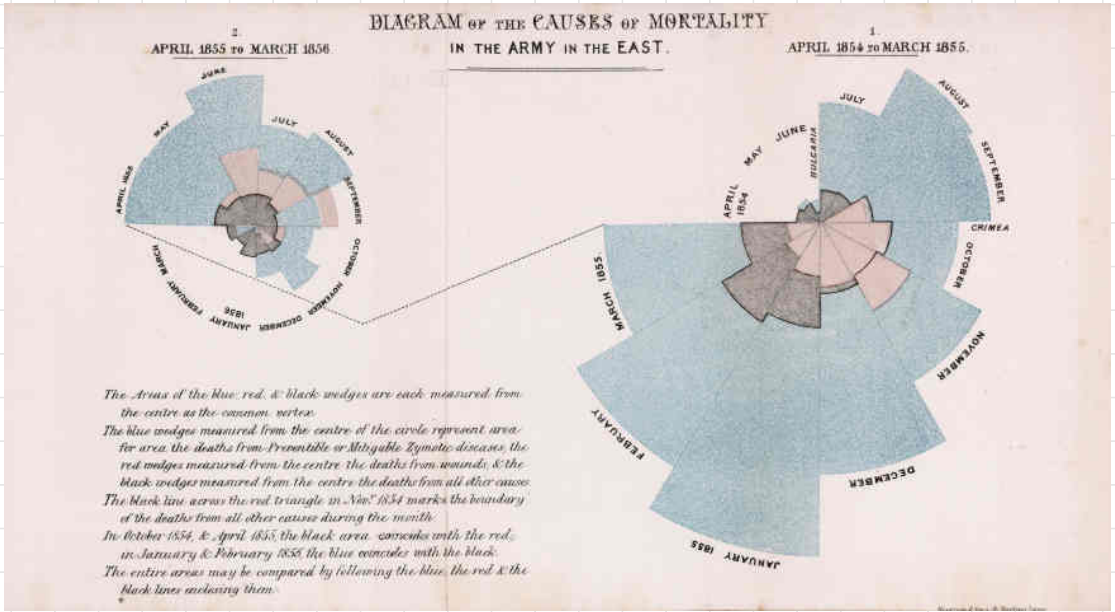


Coxcomb Chart

- polar area chart or roses
- Florence Nightingale
- radius corresponds to a magnitude of the category



- Causes of mortality prepared by Florence Nightingale
- Zoom and explore:



source: wikipedia

Scatter Plot

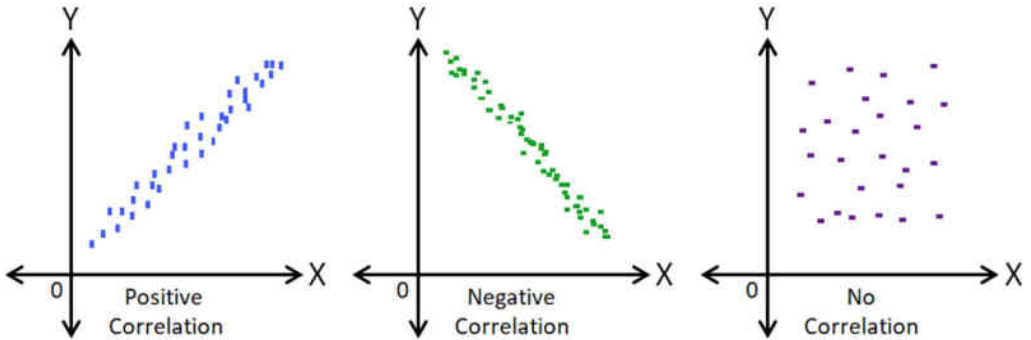
r = coefficient of correlation (Pearson)

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

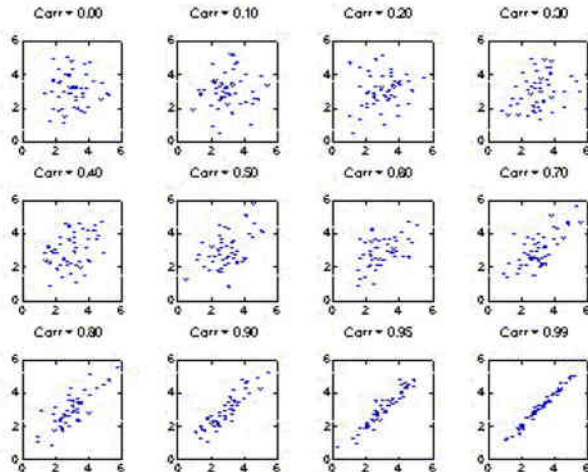
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Scatter Plots & Correlation Examples



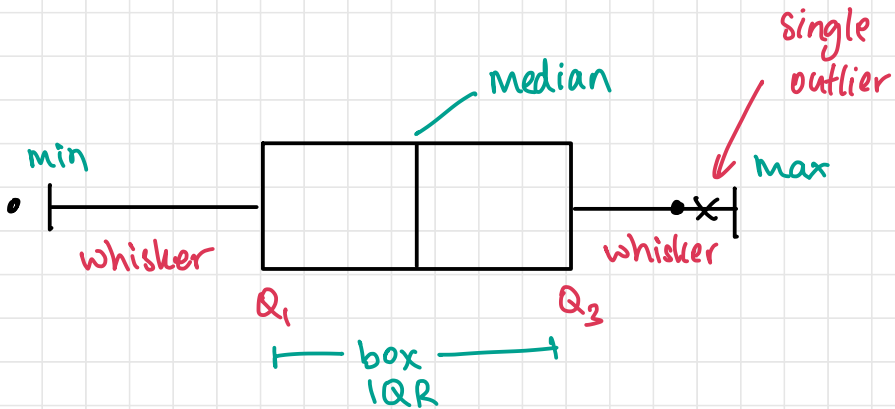
Notion of Correlation



$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

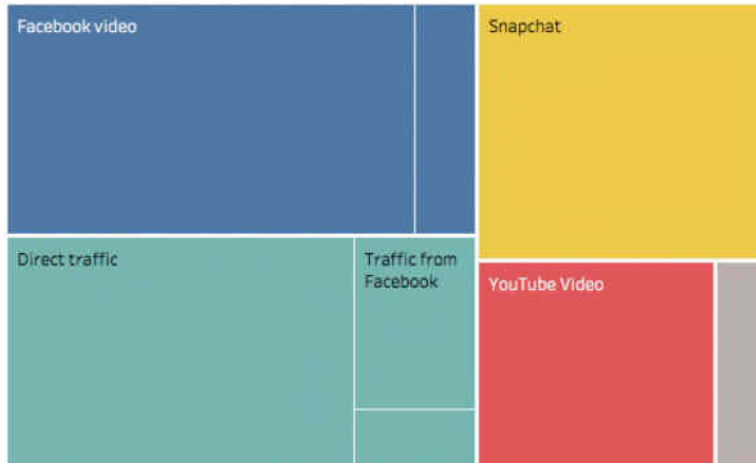
Box Plot

- Limit = $Q_3 + 1.5 IQR$ and $Q_1 - 1.5 IQR$ (outliers)



Tree Map

Where people consumed BuzzFeed content in 2015



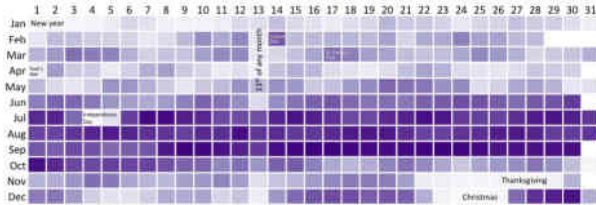
- Look at xkcd package in R

BIRTHDAYS IN THE US AND IN INDIA

This visualisation shows the **popularity of birthdays in the US** between 1973 – 1999. Dark colours indicate more popular birthdays. Light colours are less popular.

It's interesting that there are **fewer births on holidays** – almost as if doctors and hospitals do not wish to be disturbed during these days. Since 60% of the births in this period were C-sections, this does offer some flexibility.

But it's the parents too. Notice how **fewer children are born on the 13th** of any month? Superstition, perhaps? April 1st appears to be a day to avoid too, while Feb 14th – Valentine's Day – is a favourite.



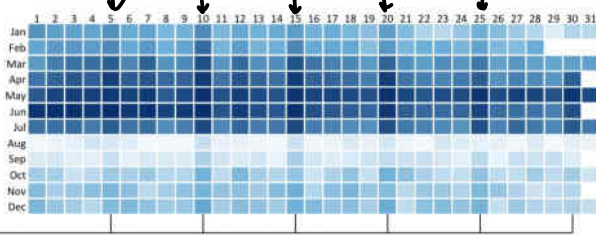
U.S. Birthdays

Most birthdays are in Jul – Sep, roughly 9 months after the winter holidays

Shown alongside is the **popularity of birthdays in India** between 2007 – 2012, for about 10 million students. Dark colours indicate more popular birthdays. Light colours are less popular.

We see a very different pattern here. Almost **no one is born in August**. A lot of births are also clustered around the months of May and June, just before schools open – and given that this data is based on school records, perhaps there is reason to suspect that these numbers are faked.

It's also suspicious that a surprisingly large number of people have birthdays on the 5th, the 10th, the 15th, the 20th etc of the month. Perhaps, when faking numbers, it is **easier to fake round numbers**.



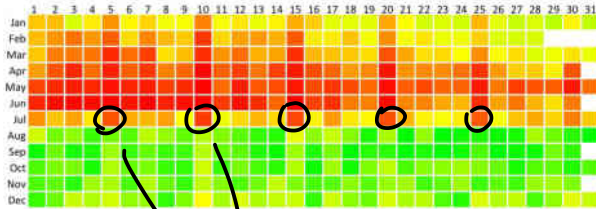
Indian Birthdays

Most birthdays are in Apr-June, while almost no one is born in August.

This rush to get children into school has an adverse impact on their marks. You can see that those "born" on the 5th, the 10th, the 15th, etc. have lower marks – most likely because these are younger children who have been taken to school earlier than their peers.

Similarly, those "born" in the first half of May have relatively lower marks. June the 1st is a particularly bad day. This is the most common birthday according to the records. (More common than Jan 1st, which is the second most common.) It also has the lowest marks on average.

Source: Tamil Nadu & Karnataka State Board examination results, 2006 - 2012



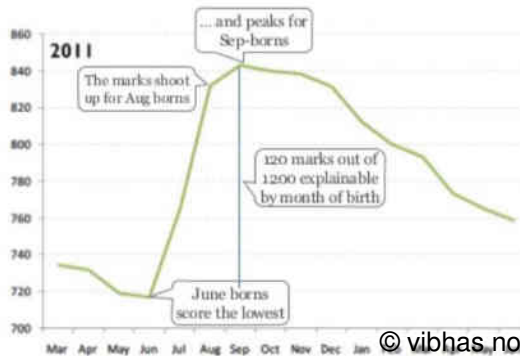
Indian Marks

Those "born" just before school opens seem to have lower marks

more likely to be fake birthdates

Results of TN X Std Boards

Based on the results of the 20 lakh students taking the Class XII exams at Tamil Nadu over the last 3 years, it appears that the month you were born in can make a difference of as much as 120 marks out of 1,200.



Restaurant Data Collected over Years

- in thousands of rupees
- Units 1, 2 & 3: poor performance on Wednesdays; Unit 4: overall poor performance

<https://gramener.com/restaurant/unit>

DAILY SALES CALENDAR MAP

UNIT1 (Rs. '000)



UNIT2 (Rs. '000)



What influences Class VIII marks?

Impact ▾ Across ▾ Subject ▾

Gender ▾ State ▾ Below poverty ▾ Siblings ▾

These results are based on 185,348 students across India.

The factors that influence the marks the **most** (and their impact in marks) are:

1	Father edu	+11.0%
2	Father occupation	+10.6%
3	Mother edu	+10.5%
4	Mother occupation	+9.1%
5	Help in household	+5.2%

The factors that influence the marks the **least** (and their impact in marks) are:

1	Gender	+0.7%
2	Distance	+0.8%
3	Private tuition	+1.3%
4	Watch TV	+2.1%
5	Siblings	+2.4%

The top influencers by subject are:

- **Maths %:** Father occupation, Father edu, Computer use, Mother occupation, ...
- **Reading %:** Father edu, Mother edu, Father occupation, Mother occupation, ...
- **Science %:** Father occupation, Mother edu, Father edu, Mother occupation, ...
- **Social %:** Father occupation, Mother occupation, Father edu, Mother edu, ...

Factor	Total %	Maths %	Reading %	Science %	Social %
Gender	0.7%	0.2%	1.9%	0.1%	0.5%
Age	4.1%	3.1%	8.0%	3.3%	2.9%
Siblings	2.4%	1.5%	8.3%	2.2%	0.8%
Father edu	11.0%	6.6%	18.8%	9.9%	7.9%
Mother edu	10.5%	4.3%	18.4%	10.3%	7.7%
Father occupation	10.6%	8.7%	17.4%	10.3%	8.5%
Mother occupation	9.1%	5.7%	14.9%	7.3%	7.9%
Below poverty	3.3%	1.8%	5.7%	2.6%	3.0%
Use calculator	2.7%	0.9%	5.0%	2.7%	2.1%
Use dictionary	3.3%	1.3%	6.4%	3.1%	2.2%
Read other books	3.1%	0.8%	6.4%	2.5%	2.6%
# Books	4.7%	2.5%	8.2%	4.1%	3.8%
Distance	0.8%	2.2%	1.2%	0.9%	0.9%
Computer use	3.4%	6.5%	5.9%	3.2%	4.1%
Library use	2.8%	2.9%	5.6%	2.7%	3.3%
Private tuition	1.3%	1.3%	2.1%	1.2%	0.3%
Watch TV	2.1%	1.5%	5.3%	2.0%	1.1%
Read magazine	3.4%	1.6%	7.2%	2.3%	2.4%
Read a book	3.7%	2.0%	7.1%	3.5%	2.5%
Play games	3.3%	3.0%	4.7%	4.0%	2.9%
Help in household	5.2%	4.6%	5.7%	5.0%	6.5%

Based on data from the [National Achievement Survey for Class VIII](#)

• Ethically: credit public source before using

Data Cleaning

- incomplete
- noisy
- inconsistent
- intentional

1. Incomplete / Missing Data

- Due to many reasons
- How to fix: interpolation

2. Noisy Data

- NLP to fix inconsistent naming convention
- Solutions
 - **Binning**: sort & partition data into equal-frequency bins
* smooth by bin means, smooth by bin median, smooth by bin boundaries
 - **Regression**
 - **Clustering**
 - **combined computer & human inspection**

DATA CLEANING AS A PROCESS

1. Data Discrepancy Detection

- use metadata, check field overloading, uniqueness rule, consecutive rule
- use commercial tools
 - **Data scrubbing**: simple domain knowledge (postal codes, spell checks)
 - **Data auditing**: analyse data to discover © rules and detect violations

2. Data Migration and Integration

- Data migration tools
- ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations via GUI
- Integration of the two processes

Exercise

- Explore how binning, clustering and regression are used in handling noisy data.
- Is combined computer and human inspection of noisy data a better way of handling the noisy data? Give reasons.
- Explain the process of data cleaning.

Important Characteristics of Data

1. Dimensionality

- number of attributes
- high dimensionality requires high volume of data to prevent overfitting

2. Sparsity

- presence (yes/no) counts
- eg: what do people sketch vs do people sketch

3. Resolution

- time scale (yearly, monthly, daily etc)
- patterns depend on scale

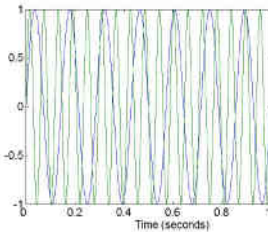
4. Size

- type of analysis may depend on size of data

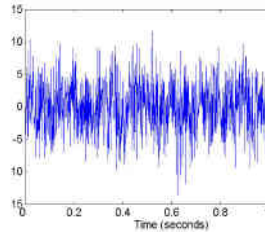
DATA QUALITY

- Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Interpretability
- Poor data negatively affects data preprocessing efforts

— Noise



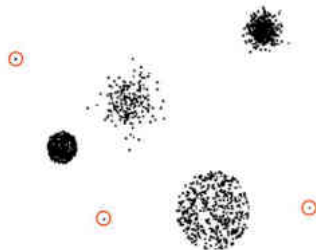
Two Sine Waves



Two Sine Waves + Noise

— Outliers

- Can be due to noise or can be genuine (eg: cases of fraud)



— Missing Values

1. Missing Completely at Random (MCAR)

- Missingness of value unrelated to attribute
- Fill in values based on attributes (mean/median) — unrealistic
- Analysis maybe unbiased overall

2. Missing at Random (MAR)

- Missingness related to other variables
- Fill in values based on other attributes — often realistic
- Analysis almost always produces bias
- Eg: unable to measure weight using scale on soft surfaces (missing values related to surface) and fill based on height

3. Missing Not at Random (MNAR or NMAR)

- Missingness related to unobserved measurements
 - Did not record, therefore missing
 - Censored data
 - Informative or non-ignorable missingness
 - Must understand why; find more data, what ifs
- When dataset given, cannot tell if MCAR, MAR, MNAR
- Multiple imputation notes:
<http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html>

Solutions

1. MCAR

- Delete rows
 - if small fraction of rows
 - can be ignored
- Delete columns
 - if small fraction of attributes
- Pairwise deletion
 - compute mean, variance and covariance with another variable
 - works reasonably well if the multivariate normal assumption holds
 - variable must have relation with another attribute
- mean imputation (only for MCAR)

2. MAR

- Regression imputation
 - $n-1$ variables related to n^{th} variable
 - unbiased estimates of mean under MCAR
 - unbiased under MAR if factors that influence missingness included
 - can expect false positives and spurious correlation
- Stochastic regression imputation
 - same as LR but adds random constant/residual to prediction

- Last observed carried forward (LOCF) and Baseline Observation carried Forward (BOCF) and worst observation carried forward (WOCF)
 - healthcare applications
 - yield biased estimates even under MCAR
 - only used if assumptions are scientifically justified
 - eg: factory output predictions
- Use of multiple imputation
 - mice, amelia in R
 - <https://stefvanbuuren.name/fimd/sec-simplesolutions.html>

3. MNAR

- Model missing values explicitly

DUPLICATE DATA

- Can occur while merging data from multiple sources
- Duplicate rows or almost duplicate rows
- Eg: same person with multiple email addresses
- When should duplicate data not be removed?
 - video processing: transmission of frames with duplicates for safety
 - must compress/delete duplicates at receiver's end

similarity measures

- Numerical measure of how alike two data objects/rows are
- Higher value when objects are more alike
- Typically falls in the interval $[0,1]$

dissimilarity measures

- Lower when objects more alike
- Distance measures (recall: word count example)
- Upper limit varies

Proximity refers to a similarity or dissimilarity

MAJOR TASKS IN DATA PREPROCESSING

1. Data Cleaning

- Fill in missing values
- Smooth noisy data
- Identify/remove outliers
- Resolve inconsistencies (imputation, dropping etc.)

2. Data Integration

- Integration of multiple databases, cubes, files

3. Data Reduction

- Dimensionality reduction (reduce no. of features)
- Numerosity reduction (reduce no. of rows — rarely done)
- Data compression

4. Data Transformation and Discretisation

- Normalisation (most common — $[0,1]$)
- Concept hierarchy generation
- Discretisation (create intervals from continuous data)

Schematic recap

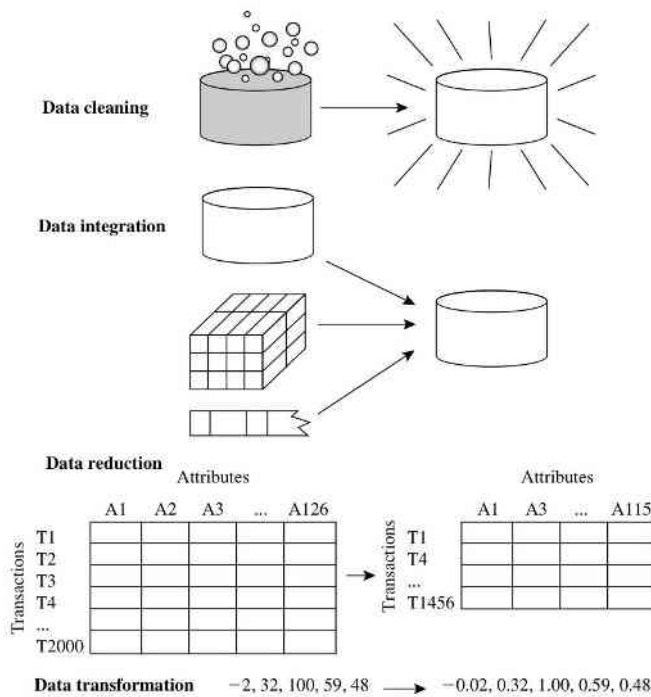


Figure 3.1 Forms of data preprocessing.

source: R1, fig 3.1

Data Integration

- Main issues: inconsistency and redundancy
- Semantic heterogeneity and structure of data

1. Entity Identification Problem

- Eg: use NLP to map common nicknames to full names (Bill = William, Bob = Robert, Dick = Richard, Liz = Elizabeth etc.)

2. Detecting and Resolving Data Value Conflicts

- For same real world entity, attribute values from different sources different
- Eg: measurements in metric system vs imperial system

Handling Redundancy

1. Object Identification

- same attribute or object may have different names
- identify row or column

2. Derivable Data

- one attribute maybe derived from another attribute
- eg: marks and grade, annual salary and monthly salary
- Redundancies can be detected using **correlation analysis** and **covariance analysis**
- Improve data mining speed by carefully integrating data

CORRELATION ANALYSIS

For CATEGORICAL DATA

- χ^2 test (chi-squared test) performed (recall Stats unit 4)
- Null hypothesis H_0 = the two variables are independent
- Alternate hypothesis H_a = the two variables are not independent

$$\chi^2 \text{ statistic} = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

k = degrees of freedom

n = no. of possible outcomes

O_i = observed no. of trials

E_i = expected no. of trials (if H_0 true)

- Larger χ^2 value, more likely to be correlated
- Cells that contribute most: large $|O_i - E_i|$ value
- Can be used for categorical values where O_i and E_i are frequencies and not fractions/percentages
- Important: correlation does not imply causation
 - spurious correlation & hidden variables
 - eg: no. of hospitals & no. of McDonald's joints \rightarrow due to high population

Q: Find if liking Science Fiction is related to playing chess

Observed Values

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Sum(col.)	300	1200	1500

$$E_{ij} = \frac{\text{sum}(A=a_i) \times \text{sum}(B=b_j)}{N}$$

$$E_{11} = \frac{300 \times 450}{1500} = 90$$

Expected values

	Play chess	Not Play chess	Sum (row)
Like Sci-Fi	90	360	450
Not Like Sci-Fi	210	840	1050
Sum (col)	300	1200	1500

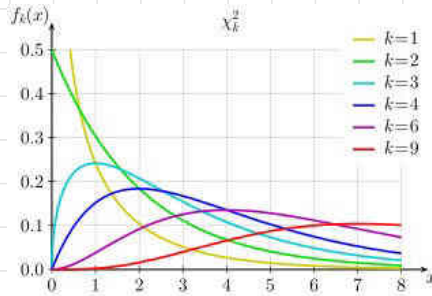
$$\chi_1^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(360-200)^2}{360} + \frac{(840-1000)^2}{840}$$

$$= 507.94$$

$k = \text{degrees of freedom} = (\text{no_of_rows} - 1) (\text{no_of_columns} - 1)$

$k = 1$

χ^2 vs Degrees of Freedom

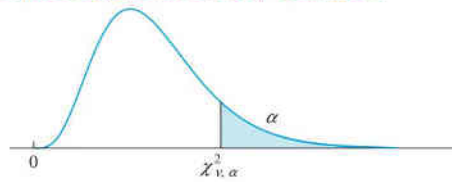


source: statistics how to

Using χ^2_1 table: $507.94 > 7.879 \Rightarrow \alpha < 0.005$

$\therefore H_0$ fails and H_a is true \Rightarrow the two are related

TABLE A.7 Upper percentage points for the χ^2 distribution



ν	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.707	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

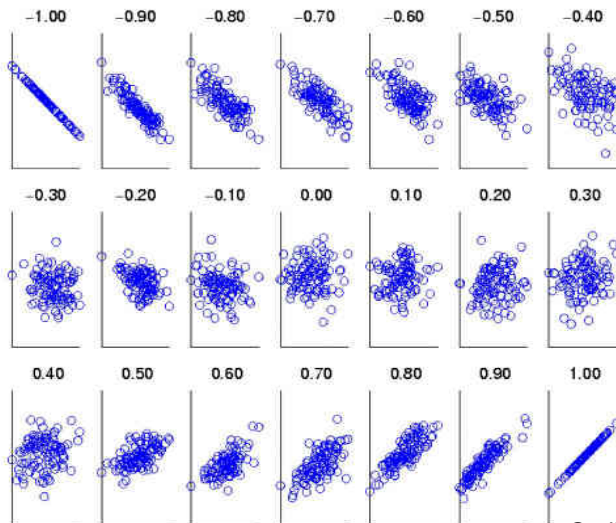
For NUMERICAL DATA

1. Pearson's correlation coefficient

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1) \sigma_A \sigma_B}$$

$$r_{A,B} = \frac{\sum_{i=1}^n [a_i b_i] - n \bar{A} \bar{B}}{(n-1) \sigma_A \sigma_B}$$

- if $r > 0$: positive correlation
if $r < 0$: negative correlation
- Visually evaluating correlation (ρ or r)



Correlation as Linear Relationship

$$a'_k = \frac{(a_k - \bar{A})}{\sigma_A} \quad b'_k = \frac{(b_k - \bar{B})}{\sigma_B}$$

$$\text{Correlation (A,B)} = A^T B$$

- inner product of vectors

2. Covariance

- Similar to correlation

$$\text{COV (A,B)} = E((A-\bar{A})(B-\bar{B})) = \sum_{i=1}^n \frac{(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- Relationship with Pearson's correlation coefficient

$$r_{A,B} = \frac{\text{COV(A,B)}}{\sigma_A \sigma_B}$$

Types of Covariance

1. Positive Covariance

- $\text{COV}_{A,B} > 0$
- Both A & B tend to be larger than their expected values

2. Negative Covariance

- $\text{COV}_{A,B} < 0$
- Both A & B tend to be smaller than their expected values

3. Independence

- $\text{Cov}_{A,B} = 0$
 - Converse not true; $\text{Cov}_{A,B} = 0 \not\Rightarrow$ independence
 - Additional assumptions (data follows multivariate normal distribution) required to imply independence
-
- Covariance alone cannot tell us the magnitude of correlation (no fixed range to decide what is high or low)

Note:

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu}$$

Q: Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

In other words: they are related, but are they t vely or $-$ vely correlated?

$$\bar{a} = \frac{2+3+5+4+6}{5} = 4$$

$$\bar{b} = \frac{5+8+10+11+14}{5} = 9.6$$

$$\text{Covariance} = \sum_{i=1}^5 \frac{(a_i - \bar{a})(b_i - \bar{b})}{5} = \frac{20}{5} = 4$$

\therefore they rise together as $\text{Cov}(A,B) > 0$

DATA REDUCTION

1. Dimensionality Reduction

- Wavelet transforms
- Principle Component Analysis
- Feature subset selection, Feature creation

2. Numerosity Reduction

- Regression and log-linear models
- Histograms, clustering, sampling
- Data cube aggregation

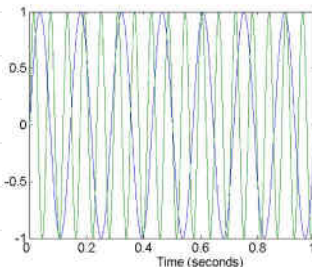
3. Data compression

DIMENSIONALITY REDUCTION

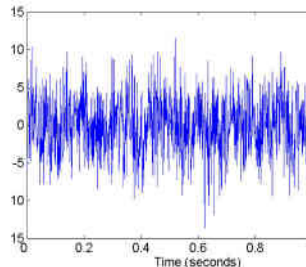
- Eliminate irrelevant features, reduce noise
- Prevent overfitting
- Avoid the "curse of dimensionality"
- Reduce time & space required for data mining

(a) Mapping Data to a New Space

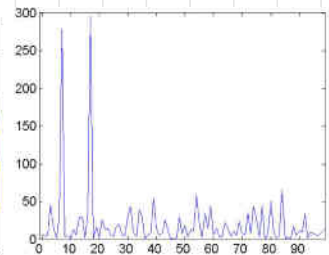
- Fourier transform (time \rightarrow frequency)
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise



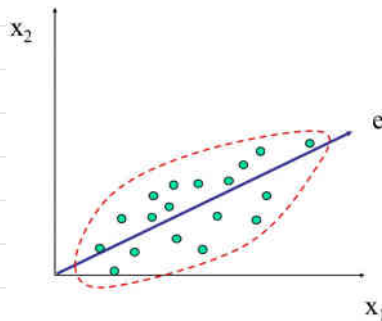
© vibhas notes 2021
Frequency

Wavelet Transformation

- Discrete wavelet transform (DWT) for linear signal processing
- Store only a small fraction of the strongest wavelet coefficients (compressed approximation)
- Similar to Discrete Fourier Transform (DFT)
- Will study in detail later

(b) Principal Component Analysis (PCA)

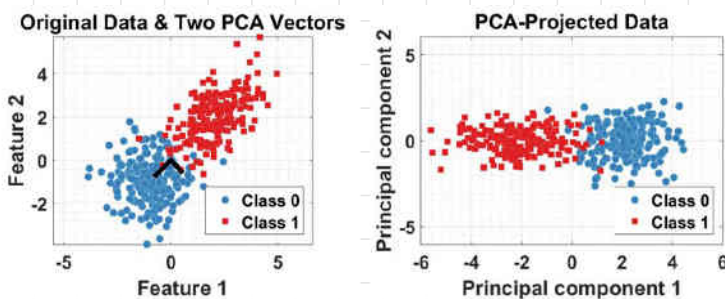
- Find a projection (eigenvector) that the original data can be projected onto with the least error
- Projection that captures the largest amount of variation in data
- Reduction in dimensionality



- Find eigenvectors of the covariance matrix which will define the new space

Steps

1. Normalise input data
2. Compute $k \leq n$ orthonormal (unit) vectors
3. Each input vector is a linear combination of the k principal component vectors
4. Principal components sorted in order of decreasing importance or strength
5. Size of data can be reduced by eliminating the weak components (low variance)



<https://www.researchgate.net/profile/Nicholas-Czarnek/publication/320410861/figure/fig7/AS:551041819447302@1508390015760/Example-application-of-principal-component-analysis-to-simple-synthetic-data-The-black.png>

PCA Example

- http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

	x	y
Data =	2.5	2.4
	0.5	0.7
	2.2	2.9
	1.9	2.2
	3.1	3.0
	2.3	2.7
	2	1.6
	1	1.1
	1.5	1.6
	1.1	0.9

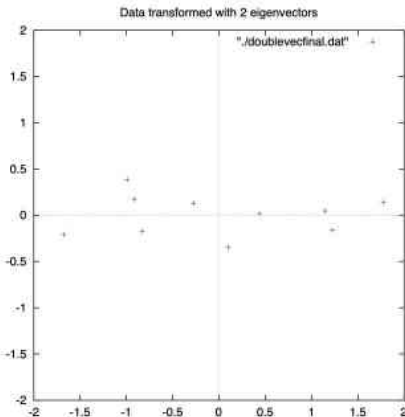
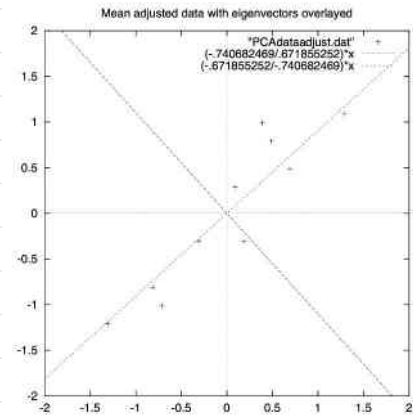
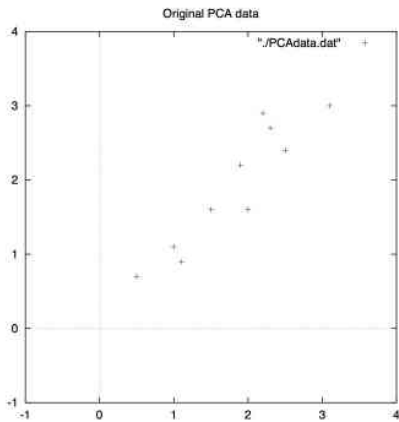
	x	y
DataAdjust =	.69	.49
	-1.31	-1.21
	.39	.99
	.09	.29
	1.29	1.09
	.49	.79
	.19	-.31
	-.81	-.81
	-.31	-.31
	-.71	-1.01

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$





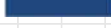
$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix} \rightarrow \text{unit eigenvectors}$$

Choosing an Appropriate Axis



PCA Using R

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

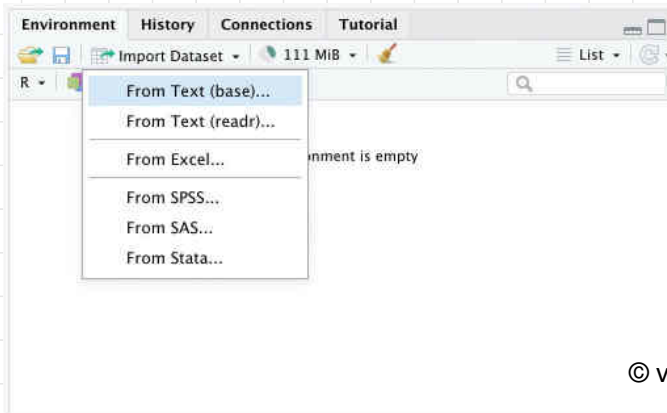
	Active individuals
	Active variables
	Supplementary quantitative variables
	Supplementary qualitative variable
	Supplementary individuals

1. Download the dataset from here:

<http://factominer.free.fr/factomethods/datasets/decathlon.txt>

2. Import it into RStudio (note: first column name is missing; add it to the txt file before importing, or download it from here)

https://drive.google.com/file/d/1OAg99W6zseJgCSGSKRog_KZfEXHh_I30/view?usp=sharing



3. Make sure you select **Heading**

Import Dataset

Name: decathlon

Encoding: Automatic

Heading: Yes No

Row names: Automatic

Separator: Tab

Decimal: Period

Quote: Double quote ("")

Comment: None

na.strings: NA

Strings as factors

Input File

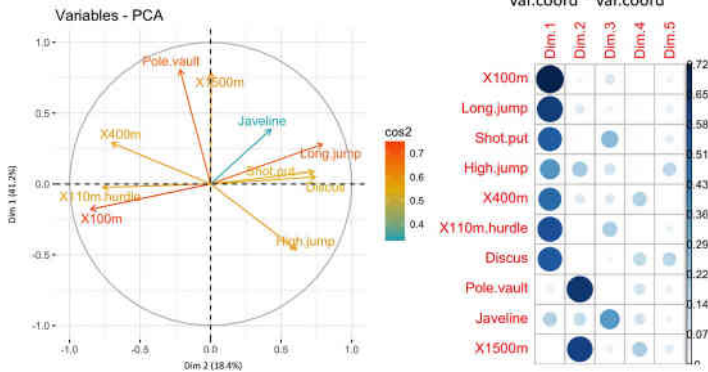
```
"Name" = "100m" = "Long_jump" = "Shot_put" = "High_jump"
"SEBRLE" = 11.04 = 7.58 = 14.83 = 2.07 = 49.81 = 14.69
"CLAY" = 10.76 = 7.4 = 14.26 = 1.86 = 49.37 = 14.85 =
"KARPOV" = 11.02 = 7.3 = 14.77 = 2.04 = 48.37 = 14.09
"BERNARD" = 11.02 = 7.23 = 14.25 = 1.92 = 48.93 = 14.99
"YURKOV" = 11.34 = 7.09 = 15.19 = 2.1 = 50.42 = 15.31
"WARNEKERS" = 11.11 = 7.6 = 14.31 = 1.98 = 48.68 = 14.23
"ZSIVOCZKY" = 11.13 = 7.3 = 13.48 = 2.01 = 48.62 = 14.1
"McMULLEN" = 10.83 = 7.31 = 13.76 = 2.13 = 49.91 = 14.3
"MARTINEAU" = 11.64 = 6.81 = 14.57 = 1.95 = 50.14 = 14.
"HERNU" = 11.37 = 7.56 = 14.41 = 1.86 = 51.1 = 15.06 =
"BARRAS" = 11.33 = 6.97 = 14.09 = 1.95 = 49.48 = 14.48
"NDOL" = 11.33 = 7.27 = 12.68 = 1.98 = 49.2 = 15.29 =
"BOURGUIGNON" = 11.36 = 6.8 = 13.46 = 1.86 = 51.16 = 15
```

Data Frame

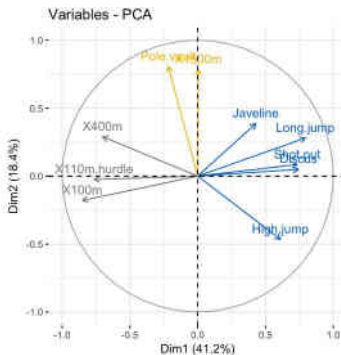
Name	X100m	Long_jump	Shot_put	High_jump	X400m
SEBRLE	11.04	7.58	14.83	2.07	49.81
CLAY	10.76	7.40	14.26	1.86	49.37
KARPOV	11.02	7.30	14.77	2.04	48.37
BERNARD	11.02	7.23	14.25	1.92	48.93
YURKOV	11.34	7.09	15.19	2.10	50.42
WARNEKERS	11.11	7.60	14.31	1.98	48.68
ZSIVOCZKY	11.13	7.30	13.48	2.01	48.62
McMULLEN	10.83	7.31	13.76	2.13	49.91
MARTINEAU	11.64	6.81	14.57	1.95	50.14
HERNU	11.37	7.56	14.41	1.86	51.10
BARRAS	11.33	6.97	14.09	1.95	49.48
NDOL	11.33	7.27	12.68	1.98	49.20
BOURGUIGNON	11.36	6.80	13.46	1.86	51.16

Import Cancel

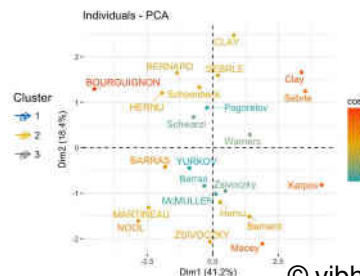
• Correlation circle



• Which events are similar?



Which athletes are similar?



(c) Attribute Subset selection

1. Redundant Attributes

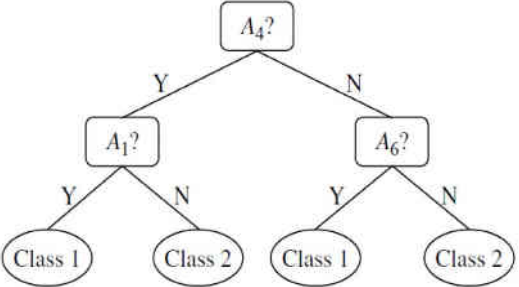
- eg: Sales tax and product price

2. Irrelevant Attributes

- eg: overfitting due to SRN-grade learning

Heuristic Search in Attribute Selection

- 2^d possible attribute combinations of d attributes (power set)
- Methods:
 - Best single attribute
 - Best step wise feature selection
 - Stepwise attribute elimination
 - Best combined attribute selection and elimination
 - Optimal branch and bound (elimination and backtracking)

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$

(d) Attribute Creation

- Create new attributes that capture info about dataset more effectively than
- Three methods
 - Attribute extraction
 - Mapping data to a new space
 - Attribute construction (Combining features, discretisation)

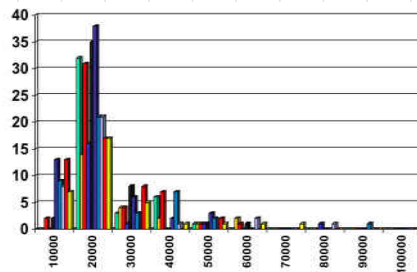
NUMEROSITY REDUCTION

(a) Parametric methods

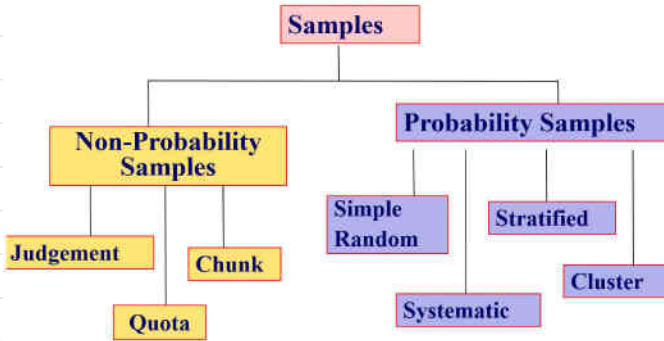
- Assume data fits a model and store only the parameters of the model
- eg: regression; store only the model and the outliers instead of storing all the points

(b) Non-parametric Methods

- Do not assume models
- eg: histograms, clustering, sampling



(c) Types of Sampling

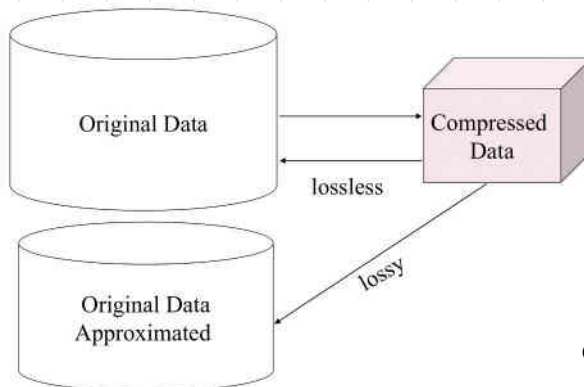


(d) Data Cube Aggregation



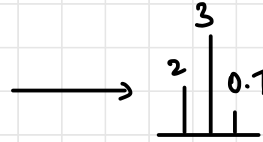
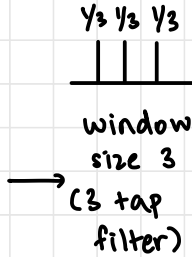
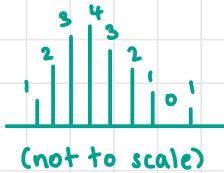
DATA COMPRESSION

- jpeg: Fourier Transform to remove high frequency formats
- jpeg 2000: similar but with Wavelet Transform



Data Transformation

- Smoothing (remove noise from data)
 - simple average
 - weighted average
 - Gaussian



- Attribute / feature construction
- Aggregation
- Normalisation
 - min-max
 - z-score
 - decimal scaling
- Discretisation

NORMALISATION

1. Min-Max Normalisation

- From $[min, max]$ to $[new-min, new-max]$

$$v' = \frac{v - min}{max - min} \times (new-max - new-min)$$

- Eg: scaling down test scores

2. Z-score Normalisation

$$v' = \frac{v - \mu}{\sigma}$$

$\mu = \text{mean}$, $\sigma = \text{std deviation}$ notes 2021

3. Normalisation by Decimal Scaling

$$v' = \frac{v}{10^j} \text{ where } j = \text{smallest int such that } \max(|v'|) \leq 1$$

DISCRETISATION

- Divide continuous values of attributes into discrete ranges (interval labels)
- Reduce data size
- Supervised vs unsupervised
Split (top-down) vs merge (bottom up)

typical methods

1. Binning: top-down split, unsupervised
2. Histogram analysis: top-down split, unsupervised
3. Clustering analysis: top-down or bottom up, unsupervised
4. Decision Tree analysis: top-down split, supervised
5. Correlation (eg: χ^2) analysis: bottom-up merge, unsupervised

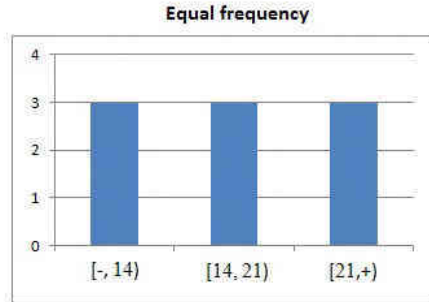
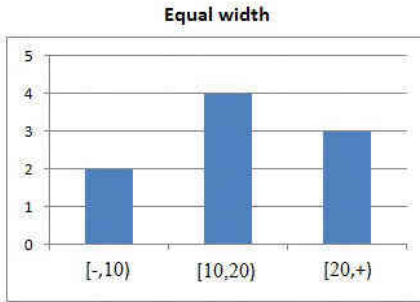
1. Binning

(a) Equal width

- range divided into N equal-sized intervals
- width = $\frac{(B-A)}{N}$ where B = highest, A = lowest, N = no. of bins
- outliers may dominate
- skewed data not handled well

(b) Equal depth

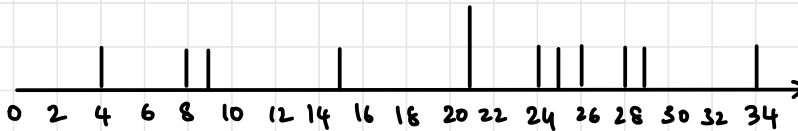
- N intervals, each containing approx. same no. of samples
- categorical can be tricky
- good data scaling



https://www.saedsayad.com/unsupervised_binning.htm

Q: Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equal-frequency bins $N=3$



total = 12 $N=3 \Rightarrow$ bin size = 4

B_1 : 4, 8, 9, 15

B_2 : 21, 24, 25

B_3 : 26, 28, 29, 34

Q: Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equal-width bins $N=3$

$$\text{bin size} = \frac{34-4+1}{3} = 10.33$$

$$B_1: [4, 14.33]$$

$$B_2: [14.34, 24.67]$$

$$B_3: [24.68, 34]$$

Q: Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Smooth by bin means, $N=3$

$$B_1: 9, 9, 9, 9$$

$$B_2: 23, 23, 23, 23$$

$$B_3: 29, 29, 29, 29$$

Q: Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

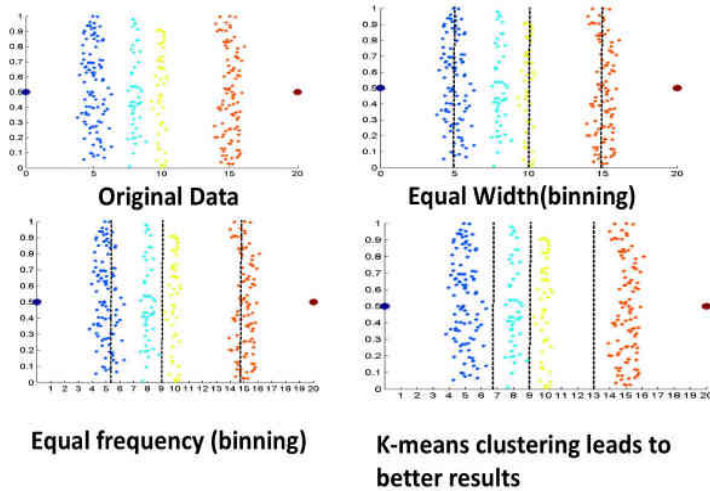
Smooth by bin boundaries, $N=3$

$$B_1: 4, 4, 4, 15$$

$$B_2: 21, 21, 25, 25$$

$$B_3: 26, 26, 26, 34$$

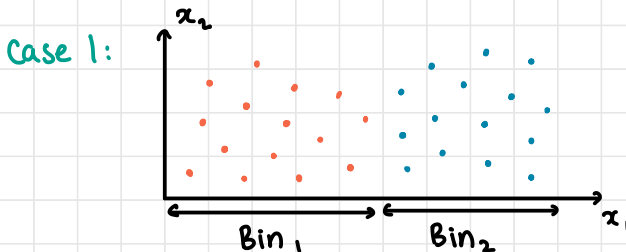
Discretisation without using Class Labels (Binning vs Clustering)



Discretisation by Classification and Correlation Analysis

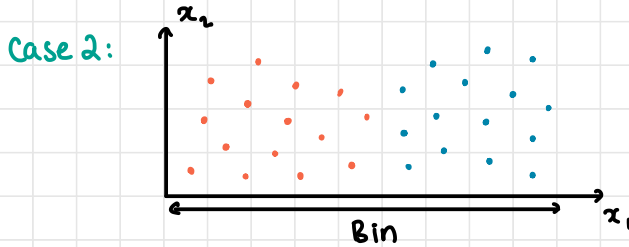
- Classification (eg: decision tree analysis)
- Supervised: given class labels
- Use entropy to determine split point (discretisation point)
- Top-down recursive split

$$E = \sum_{i=1}^n -P(c_i) \log_2(P(c_i))$$



$$\begin{aligned} \text{Entropy} (Bin_1) &= -P(\text{red}) \log_2(P(\text{red})) - P(\text{blue}) \log_2(P(\text{blue})) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(\text{Bin}_2) &= -P(\text{red})\log_2(P(\text{red})) - P(\text{blue})\log_2(P(\text{blue})) \\ &= 0 \leftarrow \text{ideal} \end{aligned}$$



$$\begin{aligned} \text{Entropy}(\text{Bin}) &= -P(\text{red})\log_2(P(\text{red})) - P(\text{blue})\log_2(P(\text{blue})) \\ &= 1 \end{aligned}$$

Correlation Analysis

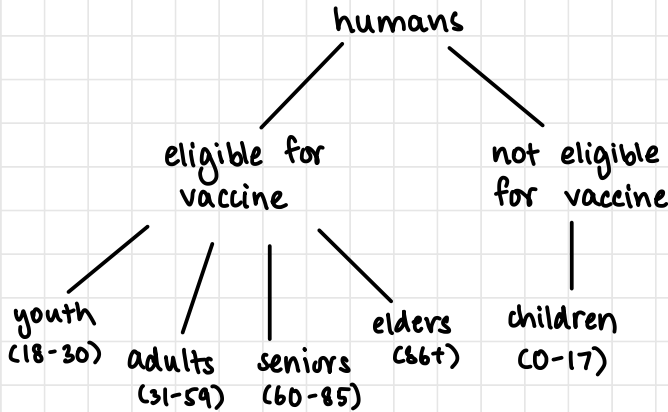
- Eg: Chi-merge : χ^2 discretisation
- Supervised: use class information
- Bottom-up merge: find best neighbouring intervals to merge (ones having similar distribution of classes — low χ^2)
- Recursive merge performed

Concept Hierarchy Generation

- Organises concepts (attribute) values hierarchically
- Drilling and rolling in data warehouses to view data in different levels of granularity
- Recursively reduce data

- Replace low level concepts (eg: age in years) to high level concepts (eg: youth, senior, child etc.)
- Explicitly defined : domain experts
- Automatically defined: both numeric and nominal data

Numeric Data



Nominal Data

- Order specified by experts

street < city < county < state < country

- Eg: {Urbana, Champaign, Chicago} < Illinois < USA
- Automatic: based on analysis of no. of distinct values placed at the lowest level of hierarchy

- Usually attribute with the most no. of distinct values is at the bottom of hierarchy

