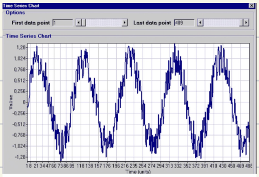
Umoto YU 1155167057

BMEG 3105 Lec 6 Scribing

Part I. Data Cleaning

- Why?
  - ↳ Data quality problems
    - ↳ Ex.:
      - ↳ 1. Noise
        - ↳ Definition: Modification of original values
        - ↳ Solution: Denoise Data
          - ↳ Ex.:

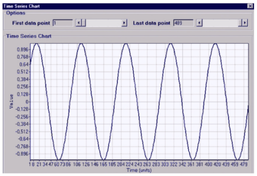


First data point 1

Last data point 489

A sine wave with noise

Denoise →

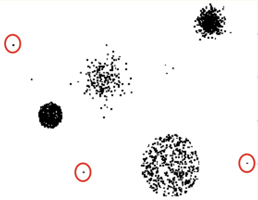


First data point 1

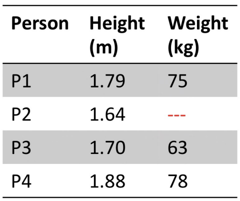
Last data point 489

The denoised sine wave

      - ↳ 2. Outliers
        - ↳ Definition: Data objects obviously different than most of others in data set
        - ↳ Solution: Remove outliers
          - ↳ Ex.

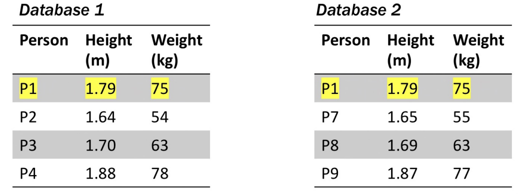


      - ↳ 3. Missing values
        - ↳ Reasons:
          - ↳ 1. Information not collected
          - ↳ 2. Attributes not appliable to all
        - ↳ Solution: Handling missing values
          - ↳ 1. Eliminate Data Objects
          - ↳ 2. Estimate missing values
          - ↳ 3. Ignore missing values during analysis
          - ↳ 4. replace with possible values



Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	---
P3	1.70	63
P4	1.88	78

- ↳4. Duplicate data
  - ↳Definition: Dataset include (almost) duplicated data objects
    - ↳Ex.:



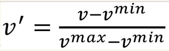
Database 1

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Database 2

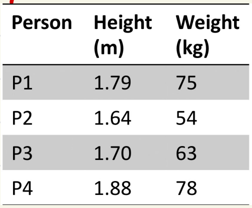
Person	Height (m)	Weight (kg)
P1	1.79	75
P7	1.65	55
P8	1.69	63
P9	1.87	77

  - ↳How?
    - ↳mostly from merging data from heterogenous sources
  - ↳Solution:
    - ↳Remove duplicates
- ↳5. Unnormalized data
  - ↳Definition: Attributes not on similar level of measurement
  - ↳Solution: Normalization
    - ↳Min-max normalization:



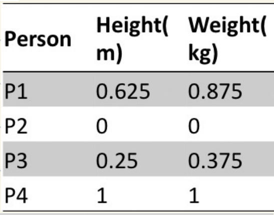
$$v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

      - ↳Ex.:



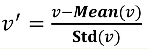
Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

$\xrightarrow{\text{Min-max}}$ Normalization



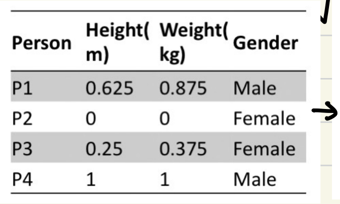
Person	Height(m)	Weight(kg)
P1	0.625	0.875
P2	0	0
P3	0.25	0.375
P4	1	1

    - ↳Z-score normalization:



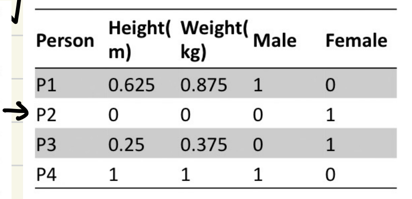
$$v' = \frac{v - Mean(v)}{\text{Std}(v)}$$

- ↳6. Categorical data
  - ↳Solution: one-hot encoding
    - ↳Ex.:



Person	Height(m)	Weight(kg)	Gender
P1	0.625	0.875	Male
P2	0	0	Female
P3	0.25	0.375	Female
P4	1	1	Male

→



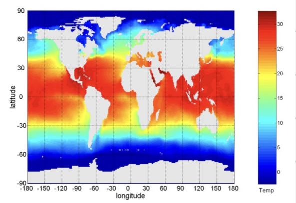
Person	Height(m)	Weight(kg)	Male	Female
P1	0.625	0.875	1	0
P2	0	0	0	1
P3	0.25	0.375	0	1
P4	1	1	1	0

Part II. Data Exploration

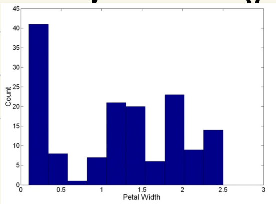
- Summary Statistics
  - ↳Definition: numbers that summerize properties of data
  - ↳Measure of location:
    - ↳mean
      - ↳ sensitive to outliers
      - ↳ $mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$
    - ↳median
      - ↳ $median(x) = \begin{cases} x_{(r+1)} & if\ m\ is\ odd, i.e., \ \ m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even,\ i.e.,\ m = 2r \end{cases}$
  - ↳Measure of spread
    - ↳range
      - ↳ Definition: Difference between max & min
    - ↳variance/standard deviation
      - ↳ ➢$variance(x) = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - mean(x)\right)^2$
    - ↳Median absolute deviation
      - ↳sensitive to outliers
      - ↳ $median(|x_1 - mean(x)|, \ldots, |x_m - mean(x)|)$
    - ↳Interquatile range
      - ↳sensitive to outliers
      - ↳ $x_{75\%} - x_{25\%}$
  - ↳Percentiles
    - ↳p-th percentile
      - ↳Definition: value of x such that p% of observed values of <u>x</u> are less than x_p
        - ↳x: ordinal/continuos attribute
    - ↳p=50
      - ↳means: x_p close to the median value.

↳ Frequency
↳ Definition: percentage of time value occurs in data set
↳ Mode
↳ Definition: most frequent attribute value
↳ usually used with categorical data

- Exploratory visualization
↳ Definition: conversion of data into visual/tabular format
↳ Why?
↳ to analyse & report the characteristics & relationships of data
↳ Ex.:



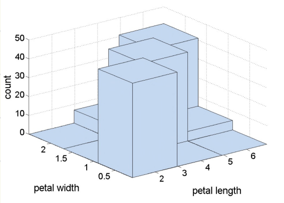
↳ powerful & appealing
↳ Because:
↳ 1. We are good at analysing visually presented data
2. can detect general patterns & trends
3. can detect outliers & unusual patterns
↳ Common techniques
↳ 1. Histograms
↳ shows: distribution of single variable value
↳ How?
↳ Divide values into bins, create a bar plot
↳ Height of bar ⇒ number of objects
↳ Shape of histogram ⇒ number of bins
↳ Ex.



↳2-d histograms

↳shows: joint distribution of 2 attributes' values

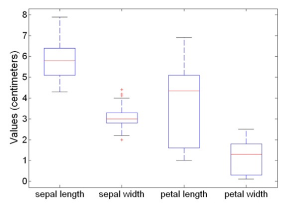
↳Ex.

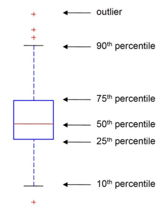


↳2. Box plots

↳for displaying & comparing data distribution

↳Ex.


Values (centimeters)
sepal length
sepal width
petal length
petal width


outlier
90th percentile
75th percentile
50th percentile
25th percentile
10th percentile