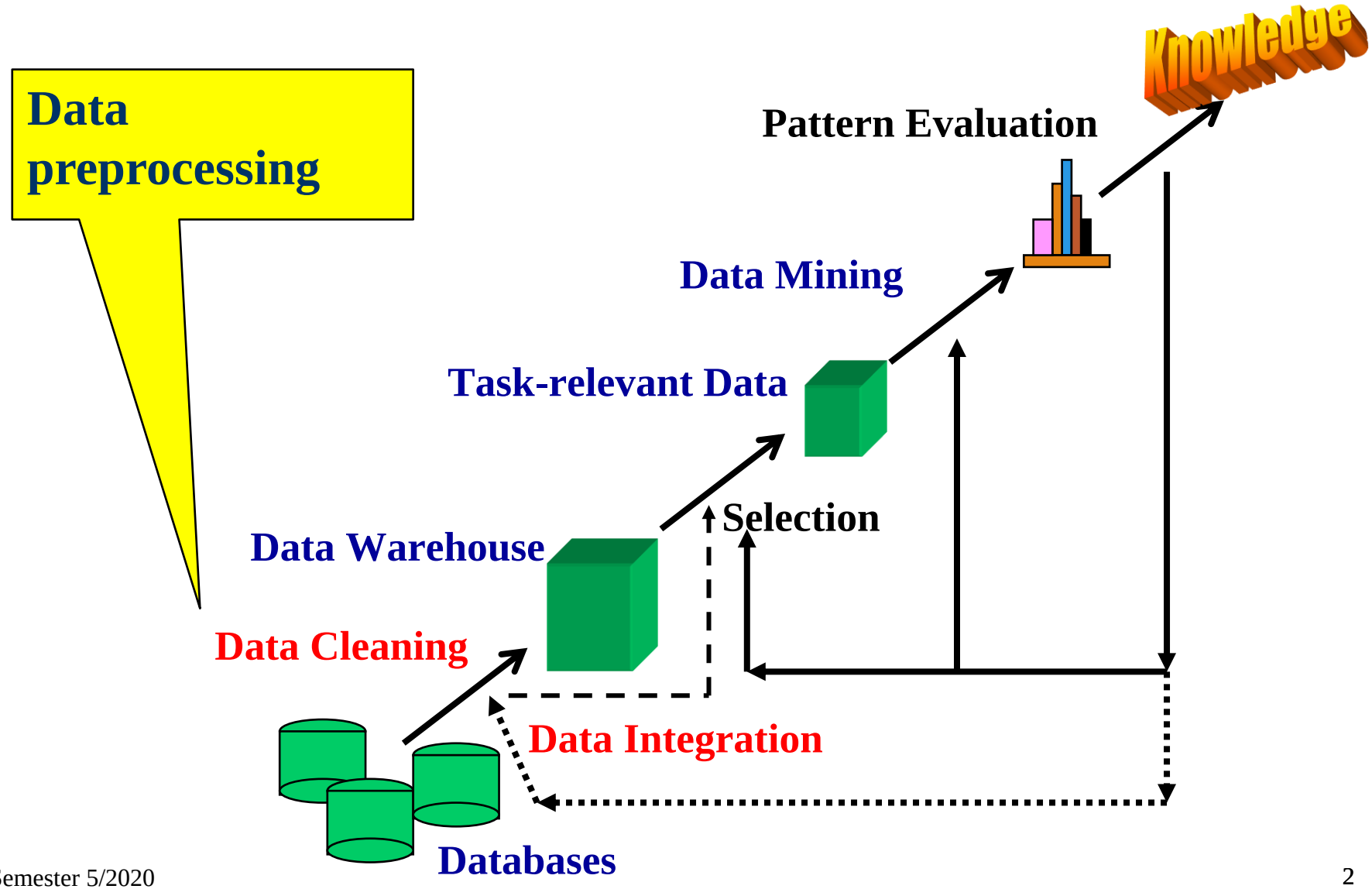
Data Science – Semester 4 – 2022/2023

INTRODUCTION TO Machine Learning

Lecture 3
Data Preprocessing

KDD process


Knowledge
Pattern Evaluation
Data Mining
Task-relevant Data
Selection
Data Warehouse
Data Cleaning
Data Integration
Databases
Data preprocessing

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Summary

Data Quality: Why preprocess the data?

- real data is **noisy, incomplete and inconsistent**
  - Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

- Measures for **data quality:** A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Quality: Why preprocess the data?

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

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	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

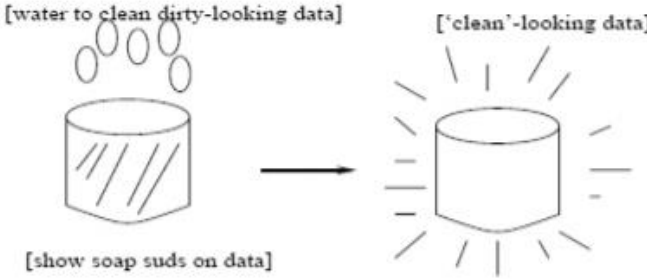
A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

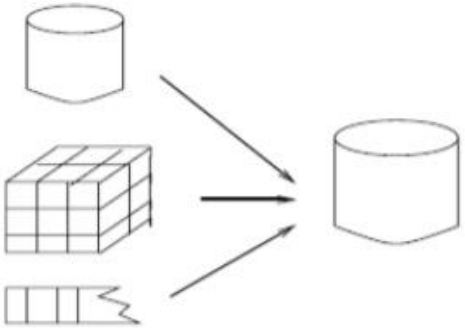
Major tasks in data preprocessing

Data Cleaning



- **Fill in missing values**
- **smooth noisy data**
- **identify or remove outliers, and resolve inconsistencies**

Data Integration



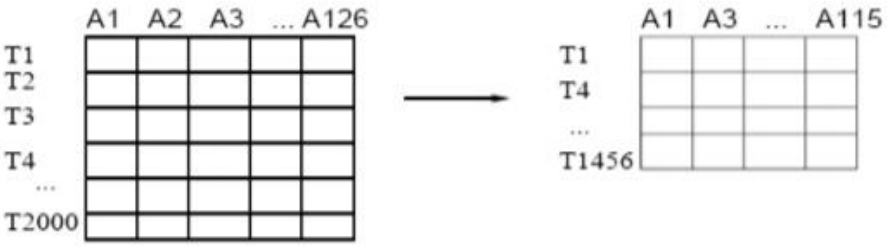
Integration of multiple databases, data cubes, or files

Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

e.g. normalization

Data Reduction



Dimensionality reduction

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- **Data Cleaning**
- Data Integration
- Data Reduction
- Data Transformation
- Summary

Data Cleaning

- **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., *Occupation*=“ ” (missing data)
- **Noisy**: containing noise, errors, or outliers
  - e.g., *Salary*=“−10” (an error)
- **Inconsistent**: containing discrepancies in codes or names
  - Was rating “1, 2, 3”, now rating “A, B, C”
  - *Age*=“42”, *Birthday*=“03/07/2010”
  - discrepancy between duplicate records
- **Intentional** (e.g., *disguised missing* data)
  - Jan. 1 as everyone’s birthday?

1. How to handle missing data?

- **Ignore the tuple**: not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually**: tedious + infeasible?
- **Fill in it automatically with**
  - **a global constant** : e.g., "unknown", a new class?!
  - **A measure of central tendency:** the attribute mean for all samples belonging to the same class
  - **the most probable value**: inference-based such as Bayesian formula or decision tree (popular strategy)

2. How to handle noisy data?

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression:** smooth by fitting the data into regression functions
- **Clustering:** detect and remove outliers

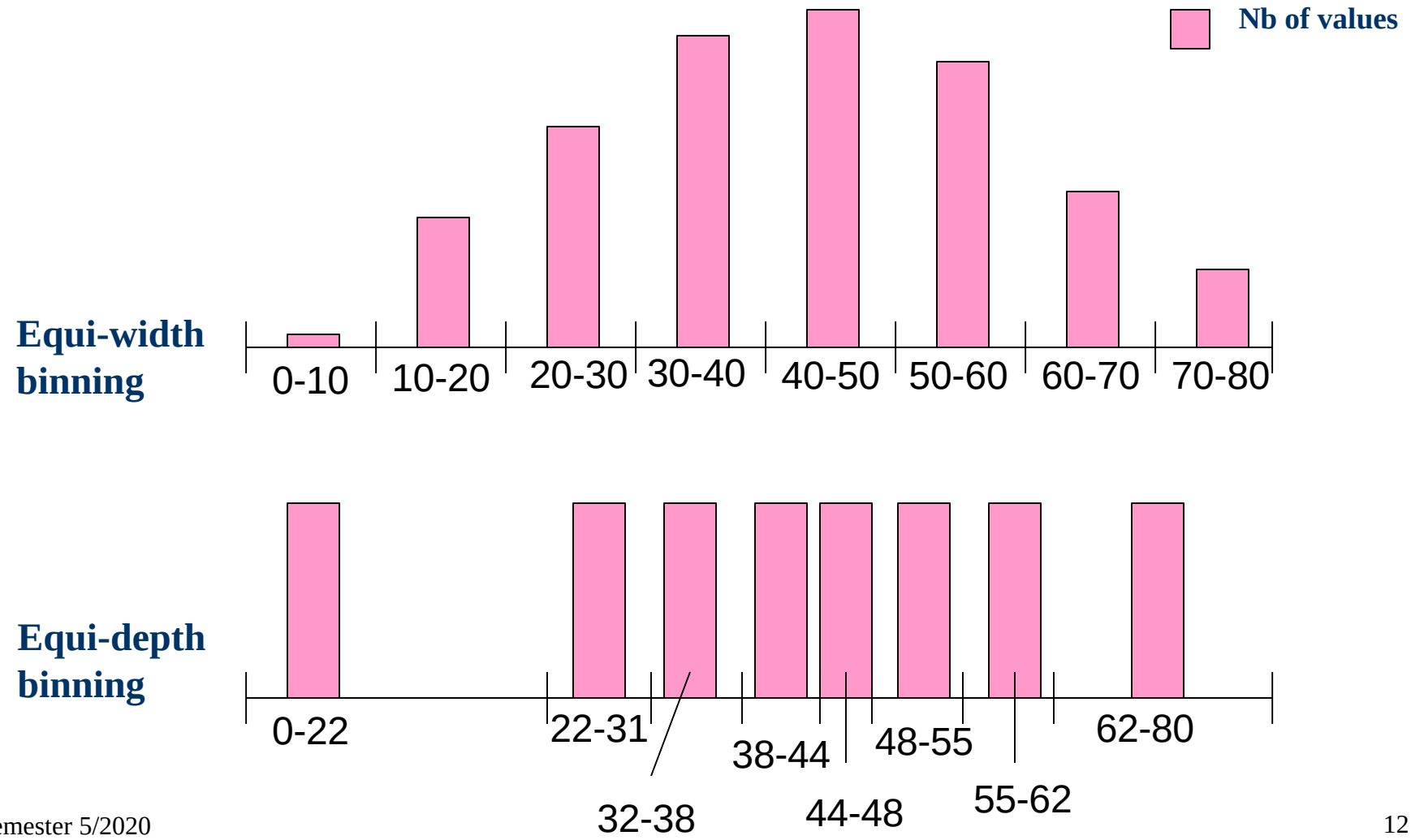
2. How to handle noisy data? (Binning)

- **Equal-width (distance) partitioning:**
  - Divides the range into N intervals of equal size: uniform grid
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well.

- **Equal-depth (frequency) partitioning:**
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

2. How to handle noisy data? (Binning)

- Example: customer ages

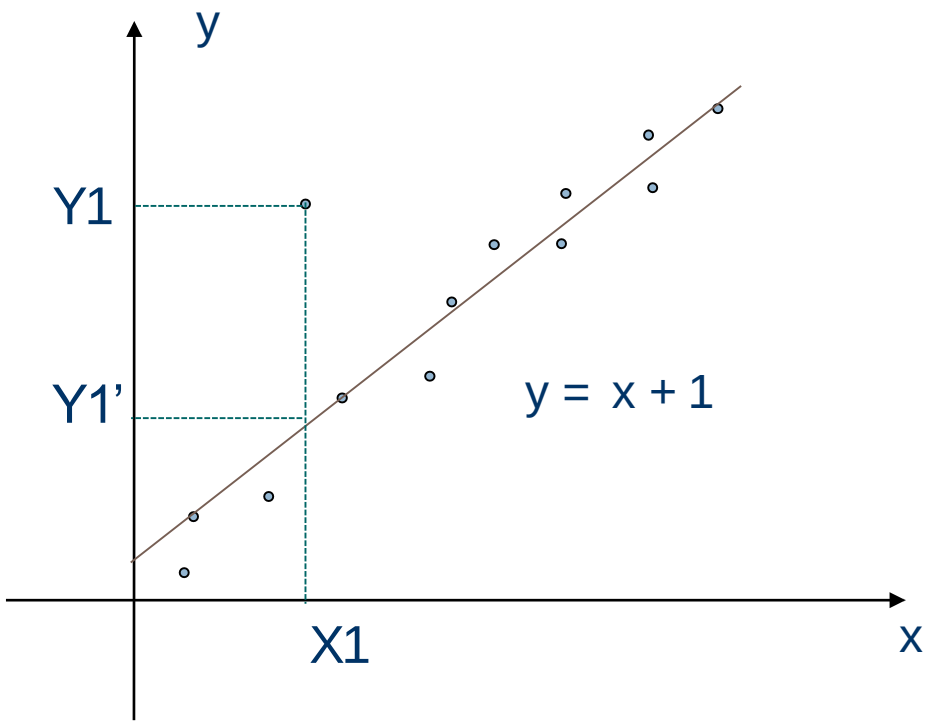


2. How to handle noisy data? (Binning)

- Example: Sorted price values 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into three (equi-depth) bins
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

2. How to handle noisy data? (Regression)

- Replace noisy or missing values by **predicted values**
- Requires model of attribute dependencies (maybe wrong!)
- Can be used for data
- smoothing or for handling missing data



Data cleaning as a process

- But data cleaning is a big job. How exactly does one proceed in tackling this task?

1. Start with identifying **data discrepancies**
   - poorly designed data entry forms that have many optional fields,
   - human error in data entry
   - Inconsistencies in data representations and inconsistent use of codes.

Data cleaning as a process

- How to identify **data discrepancies**?
  - Use domain knowledge, meta data
  - what are the data type and domain of each attribute?
  - What are the acceptable values for each attribute?
  - Use basic statistical data descriptions (find the mean, median, and mode values, symmetric vs skewed, find outliers, find dependencies between data)
  - Find inconsistent use of codes and any inconsistent data representations (e.g., “2010/12/25” and “25/12/2010” for date).

Data cleaning as a process

2. Correct data inconsistencies
   - Some errors can be corrected manually
   - Apply data transformation techniques

3. Iterate
   - Some transformations may introduce more discrepancies

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- **Data Integration**
- Data Reduction
- Data Transformation
- Summary

Data Integration

- **Data integration**: Combines data from multiple sources into a coherent store
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
  - Schema integration: e.g., A.cust-id ≡ B.cust-#
- **Redundancy and Correlation Analysis**
  - An attribute (such as annual revenue) may be redundant if it can be "derived" from another attribute or set of attributes.
  - Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.
  - can be detected with correlation analysis and covariance analysis

Redundancy example

- We have a data set having three attributes: **person_name, is_male, is_female**.
- is_male is 1 if the corresponding person is a male else it is 0 .
- is_female is 1 if the corresponding person is a female else 0

person_name	is_male	is_female
Aman	1	0
Abhinav	1	0
Ashutosh	1	0
Dishi	0	1
Abhishek	1	0
Avantika	1	0
Ayushi	0	1

HIGHLY CORRELATED ATTRIBUTES

One attribute can be removed without any information loss. As one attribue can easily determine the other.

Detecting Redundancy

- Redundancies can be detected using following methods
  - **X^2 Test** (Used for nominal Data or categorical or qualitative data)
  - **Correlation coefficient and covariance** (Used for numeric Data or quantitative data)

Correlation Analysis (Nominal data)

- **X^2 (chi-square) test:**
- Let there are two attributes A and B in a data set. A contingency table is made for representing data tuples.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Where observed values are the actual count and expected values are the count obtained from contingency table joint events.
- The χ^2 checks the hypothesis that A and B are independent. If this **hypothesis can be rejected, we can say that A and B are statistically correlated** and one of them (either A or B) can be discarded.

Correlation Analysis (Example)

1. Contingency table summarizing the relationships between gender and buying different types of pets

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

- The **aim** of the test is to conclude whether the two variables (gender and choice of pet) are related to each other.

- **Null hypothesis H0**: there is no relation between the variables
  - **If χ^2 <= critical value, then H0 holds true**

Correlation Analysis (Example)

2. Table of calculated (expected) values $\frac{row\ total * column\ total}{grand\ total}$

The expected values table :

	dog	cat	bird	total
men	223.87343533	266.00834492	240.11821975	730
women	217.12656467	257.99165508	232.88178025	708
total	441	524	473	1438

- The **aim** of the test is to conclude whether the two variables (gender and choice of pet) are related to each other.

- **Null hypothesis H0**: there is no relation between the variables
  - **If χ^2 <= critical value, then H0 holds true**

Correlation Analysis (Example)

3. Chi-square table $\frac{(Observed\_value - Calculated\_value)^2}{Calculated\_value}$

The chi-square table:

	observed (o)	calculated (c)	(o-c)^2 / c
	207	223.87343533	1.2717579435607573
	282	266.00834492	0.9613722161954465
	241	240.11821975	0.003238139990850831
	234	217.12656467	1.3112758457617977
	242	257.99165508	0.991245364156322
	232	232.88178025	0.0033387601600580606
Total		χ^2	**4.542228269825232**

Correlation Analysis (Example)

4. Find critical value of chi-square:

- **degrees of freedom** for the dataset: **(no. of rows – 1) * (no. of columns – 1) = (2-1) * (3-1) = 2**
- Now, let us look at the table and find the value corresponding to **2** degrees of freedom and **0.05** significance factor
- The tabular or critical value of chi-square here is **5.991 > 4.54**

$$critical\ value\ of\ \chi^2\ >=\ calculated\ value\ of\ \chi^2$$

H0 is **accepted**, that is, the variables ***do not*** have a significant relation.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

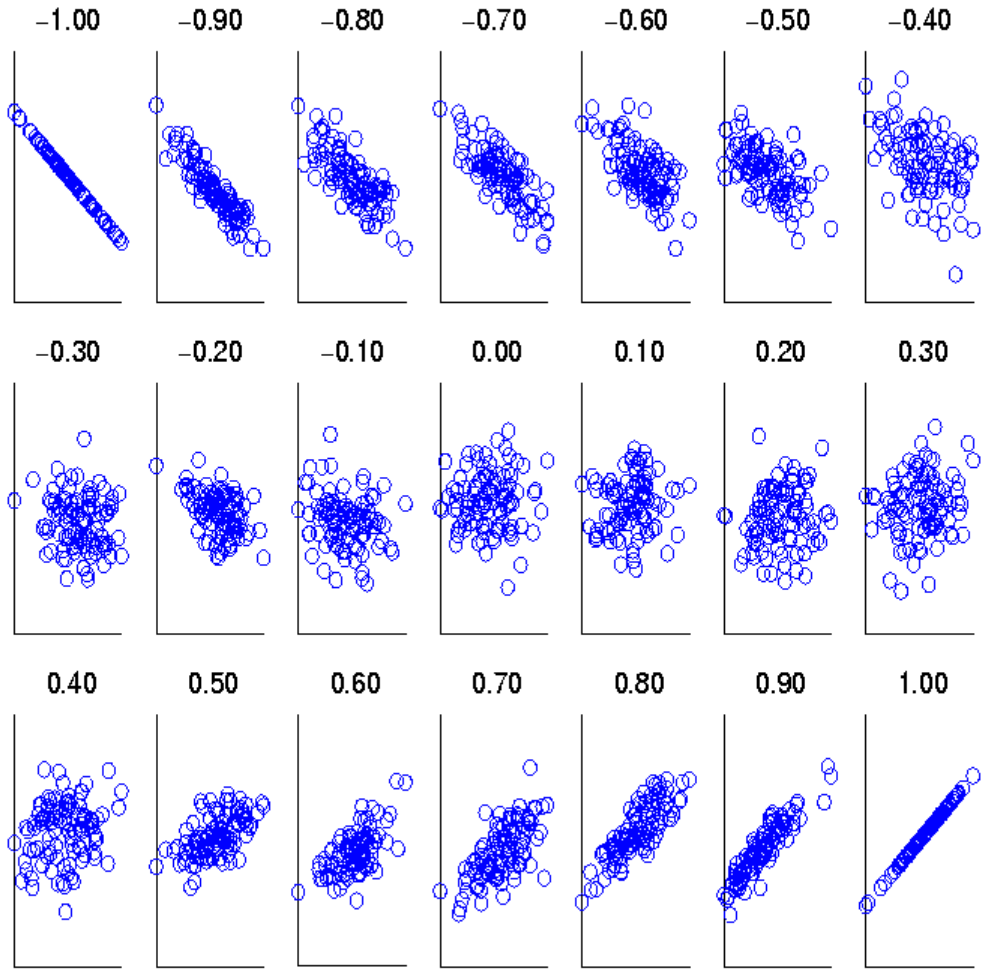
Correlation coefficient (Numerical data)

- Correlation coefficient (also called **Pearson's product moment coefficient**):
  - measures linear correlation between two variables *X* and *Y*.
  - It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

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

- where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- **Data Transformation**
- Summary

Data Transformation

Motivation

- Can we compare these attribute values?

- For Example: Compare following two records
  - (5.9 ft, 50 Kg)
  - (4.6 ft, 55 Kg)

    Vs.
  - (5.9 ft, 50 Kg)
  - (5.6 ft, 56 Kg)

- We need Data Transformation to makes different dimension(attribute) records comparable …

Data Transformation

- A function that **maps the entire set of values of a given attribute to a new set of replacement** values s.t. each old value can be identified with one of the new values
- Methods:
  - **Smoothing**: Remove noise from data
  - **Aggregation**: Summarization, data cube construction
  - **Normalization**: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling

Data Transformation (Aggregation)

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose:
  - Data reduction: Reduce the number of attributes or objects
  - Change of scale: Cities aggregated into regions, states, countries

Data Transformation (Normalization)

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- **Data Reduction**
- Data Transformation
- Summary

Data Reduction

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

- **Why data reduction?** — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

- **Dimensionality reduction**, e.g., remove unimportant attributes
  - Principle Component Analysis or Singular Value
  - Wavelet Transform (used for image compression)

- **Numerosity reduction** (some simply call it: Data Reduction)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation

- **Data compression**

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression