Data Preprocessing

IT 326: Data Mining

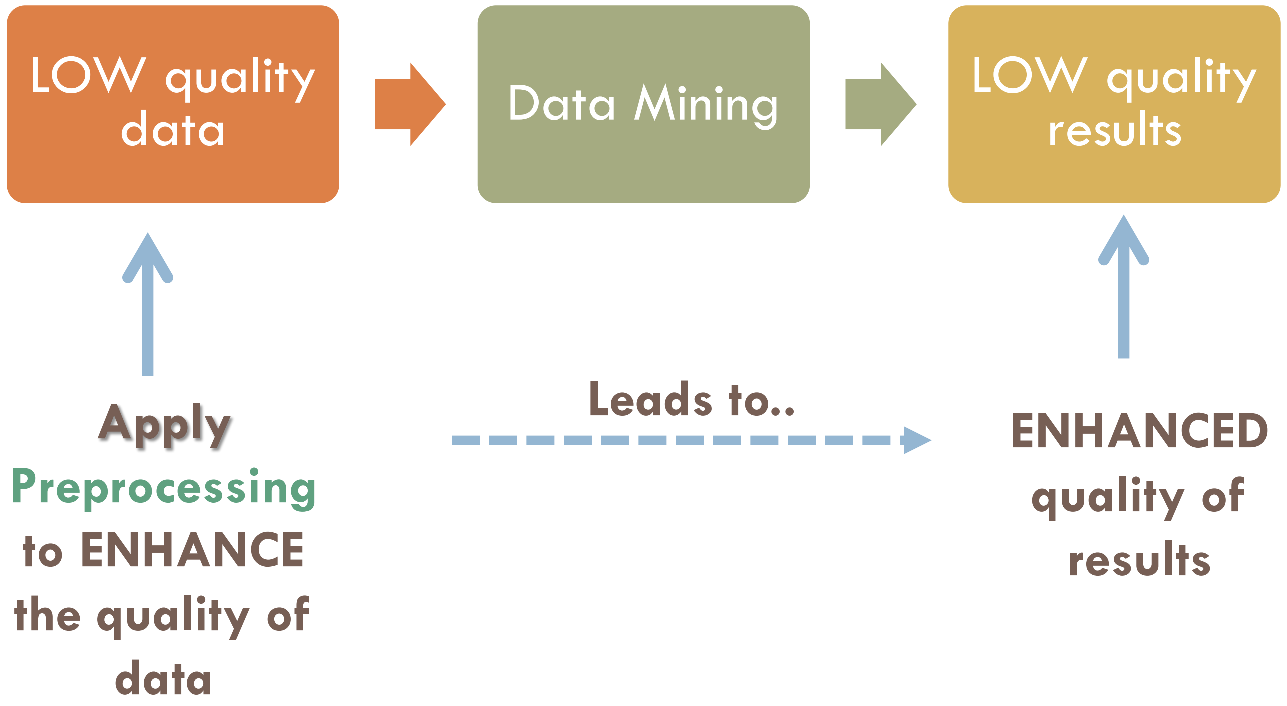
First semester 2021

Chapter 3, "Data Mining: Concepts and Techniques" (3rd ed.)

Outline

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

Data Preprocessing: Why Preprocess the Data?

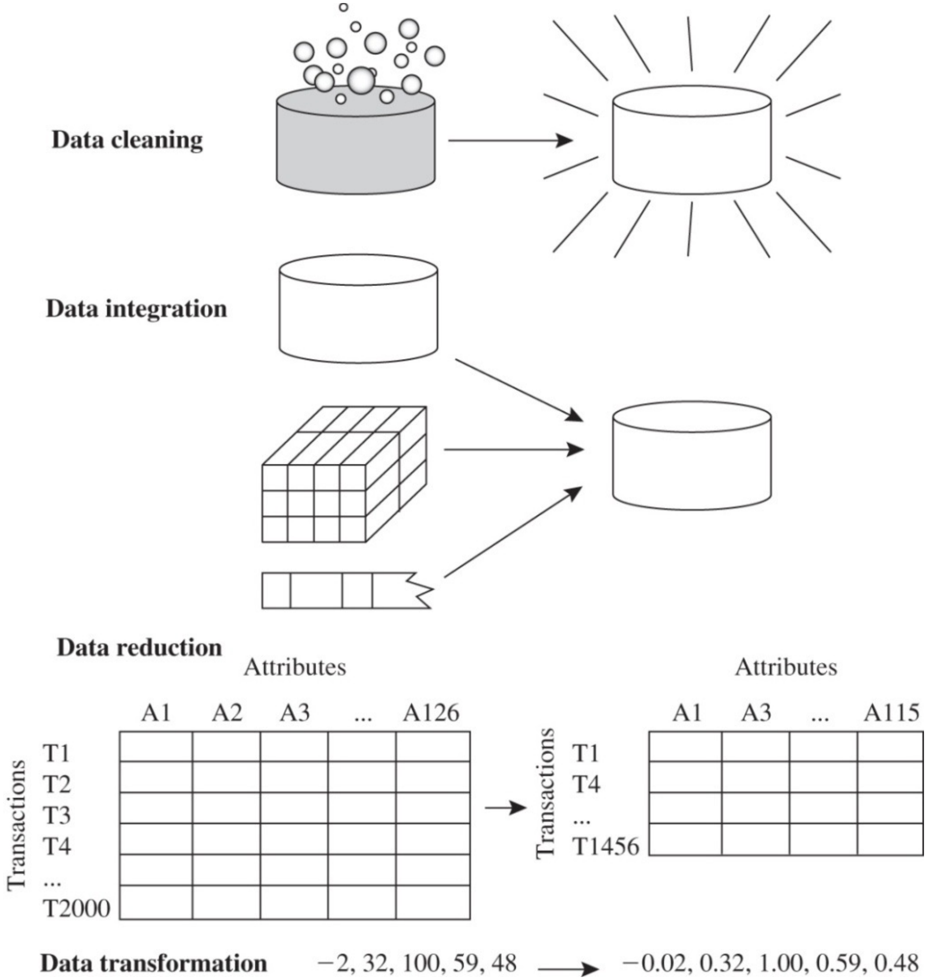


Data Quality

- Elements defining data quality:
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: inconsistent naming, coding, format … ⇒ متشابهة
  - Timeliness: timely updated?
  - Believability: how much the data are trusted by users?
  - Interpretability: how easy the data are understood? *

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 or files.
- **Data reduction**
  - Dimensionality reduction. ↓ attribute
  - Numerosity reduction. – ↓ obj
- **Data transformation**
  - Normalization. num
  - Concept hierarchy generation. nom
  - Discretization num → nom



Data Cleaning

- Data in the Real World is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error.
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. ⇒ Ex. without Details
    - e.g., Occupation=" " (missing data) *
  - Noisy: containing noise, errors, or outliers.
    - e.g., Salary="−10" (an error) *
  - Inconsistent: containing contradictories in codes or names, e.g., = يكون فيه تناقض
    - Age="42", Birthday="03/07/2010".
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?

→ missing or wrong

Dirty Data

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values (1)

Invalid values (2)

Misfielded values

must city

Misspellings

Uniqueness

Formats

Attribute dependencies

Incomplete (Missing) Values

- Data is not always available.
  - e.g., many tuples have no recorded value for several attributes, such as customer income in sales data.
- Missing data may be due to:
  - equipment malfunction.
  - inconsistent with other recorded data and thus deleted.
  - data not entered due to misunderstanding.
  - certain data may not be considered important at the time of entry.
- Missing data may need to be inferred. ← طريقة حلها

How to Handle Missing Values?

1. **Ignore the tuple**: usually done when class label is missing (when doing classification)—not effective when the percentage of missing values per attribute varies considerably. big Data لو كانت <= الغالبية من الداتا

2. **Fill in the missing value :** ✓
   - **Manually**: time consuming + infeasible (large data and many missing values) => أروح أبحث عنهم
   - Use **a global constant** (such as a label like "Unknown" or $-\infty$ or "NA") *
   - Use **the central tendency** for the attribute (e.g., the mean or median)
   - Use **the attribute mean/median** for all samples belonging to the same class.
   - Use **the most probable value.**

=> أفضل شي لكن تأخذ وقت

prediction => other attr help me to know the missing value
classification

طريقة الاختيار على حسب التطبيق
=> ايش يستخدم طريقة لإيجاد ال missing value

Noisy Data ⇒ error or varaince

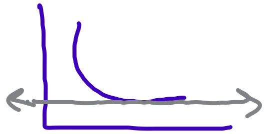
- ① ②
  Noise: random error or variance in a measured variable.

- Incorrect attribute values may be due to:
  - ① faulty data collection instruments.
  - ② data entry problems.
  - ③ data transmission problems.

- Smooth out the data to remove noise.

How to solve it?

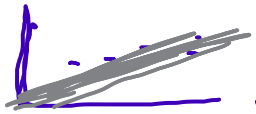
Smothing →



make value
مقرب

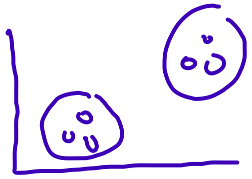
How to Handle Noisy Data?

- Binning
  - First, sort data and partition into (equal-frequency) bins.
  - Then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - Smooth by fitting the data into regression functions.



⇒ نحاول نمثل النقاط المختلفة بخط

- Outlier Analysis
  - Detect and remove outliers using clustering.
    - Values that fall outside of the set of clusters may be considered outliers.
  - Combined computer and human inspection.



Example: Binning To Handle Noise

أهم خطوة

importance Step

1 Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

minimum maximum

2 Partition into (equal-frequency) bins:

نشوف أقرب طرفين

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

3 or

Smoothing by bin means:

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

$\frac{4+8+15}{3} = 9$

نقسم على عددهم

Smoothing by bin boundaries:

① نشوف الباوندريس

- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

سهلة الطرق
وواضحة

Binning methods for data smoothing.

2 ways

①

equal-frequency: It divides the range into N intervals, each containing approximately same number of samples.

3 / 3 / 3

②

equal-width: It divides the range into *N* intervals of equal size.

Width= (max-min/#bins) number of bins

Example: Width=(34-4/3)=10

[4-13], [14-23] , [24-34]

- Bin1: 4, 8 [4-13]
- Bin2: 15, 21, 21 [14-23]
- Bin3: 24, 25, 28, 34 [24-34]

Data Integration

- **Data integration:**
  - The merging of data from multiple sources into a coherent store.

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

- **Challenges:**
  - Entity identification problem: How to match schemas and objects from different sources?
  - Redundancy and Correlation Analysis: Are any attributes correlated?

Entity Identification Problem

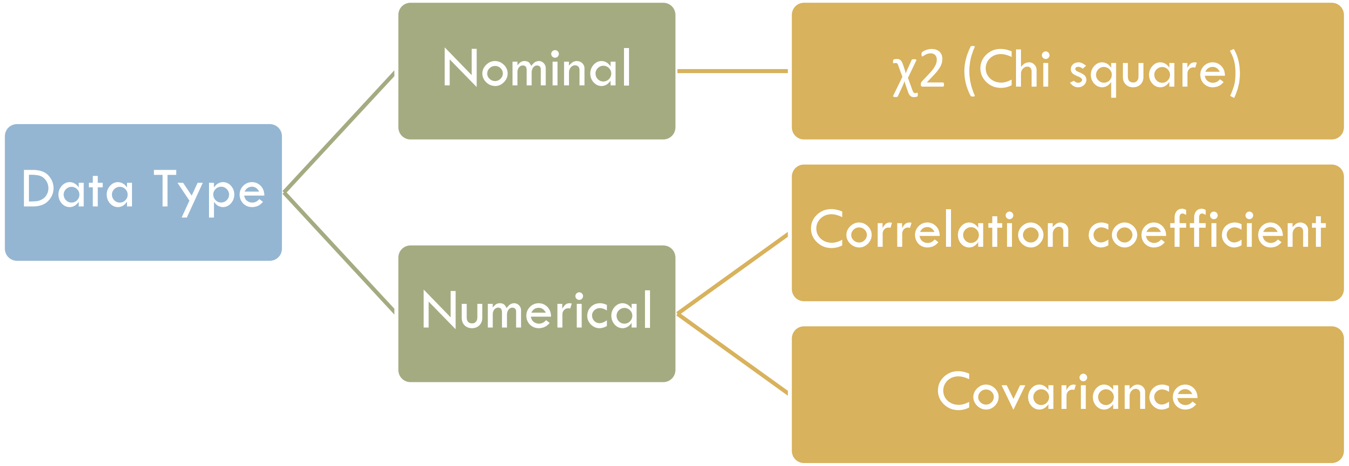
- How can equivalent real-world entities from multiple data sources be matched up?
- Same attribute or object may have different names in different databases.
  - Attribute name : e.g., A.cust-id ≡ B.cust-#.
  - Attribute values : e.g., Bill Clinton ≡ William Clinton
  - If both kept → redundancy
- Schema integration and object matching are tricky:
  - Integrate metadata from different sources.
  - Match equivalent real-world entities from multiple sources.
  - Detecting and resolving data value conflicts.
    - Possible reasons: different representations, different scales.
  - Metadata can be used to avoid errors in schema integration.
    - Examples of metadata for attributes : name, meaning, data type, and range of values permitted.

Attribute Redundancy and Correlation Analysis

- Redundant data occur often when integration of multiple databases.
- Causes of redundancy:
  - An attribute may be redundant if it can be "derived" from another attribute or set of attributes.
  - Inconsistencies in attribute naming can also cause redundancies in the resulting dataset.
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

Handling Redundancy with Correlation Analysis

- Some redundancies can be detected by correlation analysis.
- Given two attributes, correlation analysis can measure how strongly one attribute implies the other, based on the available data.
- Each type of data has special type of correlation measure.



Correlation Analysis (Nominal Data): Chi-Square

- **X^2 (chi-square) test:**
  - Observed value is the actual count & Expected value is the expected frequency.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \qquad \chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

- Expected values are calculated using :

$$e_{ij} = \frac{count\ (A = a_i) \times count(B = \boldsymbol{b}_j)}{n}$$

- The larger the X^2 value, the highest the correlation.
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count.

Example: Chi-Square

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}},$$

Dataset

Gender	Preferred reading
Male	Fiction
Female	Not fiction
Female	Fiction
..	…
..	…
..	…
Female	Fiction

n=1500

Contingency table

	male	**Female**	**Sum (row)**
fiction	250 (90)	200 (360)	450
Not fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- The expected frequencies(*numbers in parentheses*):
  - $e_{11} = \frac{count\ (male)\times count(fiction)}{1500} = \frac{300\times 450}{1500} = 90$
- X2 (chi-square) calculation to test correlation between “preferred reading” and “gender” :

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Example: Chi-Square cont.

- Degree of freedom (df) = **df** = (r-1)(c-1) where r is the number of rows and c is the number of columns.
  - df= (2-1)(2-1) = 1
- From the below table, the X^2 rejected value for 0.001 significant level is 10.828, then the results show that "preferred reading" and "gender" are strongly correlated in the given group of people.

Null hypothesis:
H_0: the two attributes are independent
α= 0.001
Critical value= 10.827
X^2 : chi-square test statistic
If X^2 > the critical value ➔ H_0 is rejected.

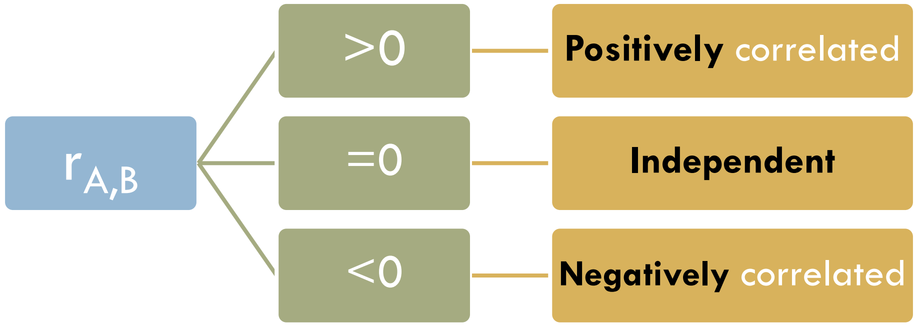
	Probability level (alpha)					
df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Correlation Analysis (Numeric Data): Correlation Coefficient

- **Correlation coefficient:**

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

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_ib_i)$ is the sum of the AB cross-product.



Correlation Analysis (Numeric Data): Covariance

- **Covariance:**

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

where n is the number of tuples, and $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B

- Covariance is similar to correlation:

Correlation Coefficient

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

- Covariance can be simplified as :

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}.$$

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

Correlation Analysis (Numeric Data): Covariance

- Positive covariance: If A and B both tend to be larger than their expected values THEN Cov(A,B) > 0 → they rise together
- Negative covariance: If A is larger than its expected value, while B is smaller than its expected value THEN Cov(A,B) < 0.
- Independence: If A and B are independent then Cov(A,B) = 0.

- But the converse is not true.
  - Cov(A,B) = 0 does NOT imply that A and B are independent.
  - Some pairs of random variables may have a covariance of 0 but are not independent.
  - Only under some additional assumptions does a covariance of 0 imply independence.

Example: Covariance

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

Stock prices observed at five time points for AllElectronics and HighTech company.

Time point	AllElectronics	HighTech
T1	6	20
T2	5	10
T3	4	14
T4	3	5
T5	2	5

1 $$E(AllElectronics) = \frac{(6) + (5) + (4) + (3) + (2)}{5} = 4$$

2 $$E(HighTech) = \frac{(20) + (10) + (14) + (5) + (5)}{5} = 10.80$$

3 $$cov(AllElectronics, HighTech) =$$

$$\frac{(6\times20) + (5\times10) + (4\times14) + (3\times5) + (2\times5)}{5} - (4\times10.80) =$$

$$50.2 - 43.2 = \mathbf{7}$$

Data Transformation

- Data are transformed or consolidated into forms appropriate for mining.
  - the resulting mining process may be more efficient, and the patterns found may be easier to understand.
- Attribute Transformation: A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.

Data Transformation

Original data	Transformed data by label encoding
A	0
B	1
D	3
C	2
B	1
C	2
:	:

Encoding

Name	Birthday		Birthday
John	31/12/1990	------->	90's
Mery	15/10/1978	------->	70's
Alice	19/04/2000	------->	00's
Mark	01/11/1997	------->	90's
Alex	15/03/2000	------->	00's
Peter	01/12/1983	------->	80's
Calvin	05/05/1995	------->	90's
Roxane	03/08/1948	------->	40's
Anne	05/09/1992	------->	90's
Paul	14/11/1992	------->	90's

Discretization

Name	Salary		Centered Salary	Centered & Scaled Salary
John	1,500 €	------>	14.30	0.0199
Mery	1,234 €	------>	-251.70	-0.3504
Alice	2,211 €	------>	725.30	1.0098
Mark	2,159 €	------>	673.30	0.9374
Alex	2,689 €	------>	1203.30	1.6753
Peter	958 €	------>	-527.70	-0.7347
Calvin	823 €	------>	-662.70	-0.9226
Roxane	1,897 €	------>	411.30	0.5726
Anne	627 €	------>	-858.70	-1.1955
Paul	759 €	------>	-726.70	-1.0117
Mean	*1,485.70 €*			
Std Dev	*718.27 €*			

Normalization

InvoiceNo	Description	...	Quantity	InvoiceDate	UnitPrice	CustomerID
536365	WHITE HANGING HEART T-LIGHT HOLDER	...	6	12/1/18	2.55	17850
536366	WHITE METAL LANTERN	...	6	23/1/18	3.39	17850
536367	CREAM CUPID HEARTS COAT HANGER	...	8	18/2/18	2.75	17850
536368	KNITTED UNION FLAG HOT WATER BOTTLE	...	6	15/1/18	3.39	12583
536369	RED WOOLLY HOTTIE WHITE HEART.	...	6	10/3/18	3.39	12583
...	...	...	...		...	...
739411	JAM MAKING SET WITH JARS	...	4	25/2/18	4.29	13047

Aggregate instances by customer ID

CustomerID	Count_CustomerID	...	Sum_Quantity	Min_Date	Avg_UnitPrice
17850	3	...	20	12/1/18	2.90
12583	2		12	15/1/18	3.39
...	...	...	...	...	...
13047	1	...	4	25/2/18	4.29

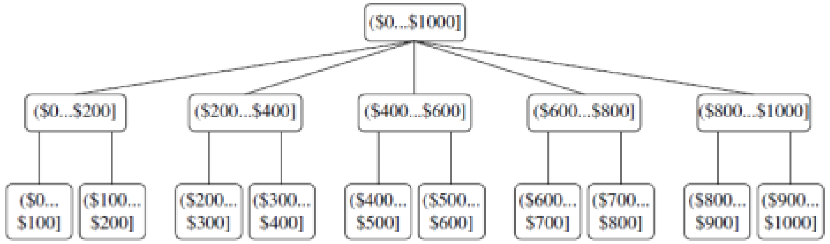
Aggregation

Data Transformation: Strategies

- Smoothing: Remove noise from data.
- Attribute construction: New attributes constructed from the given ones. New attributes are added to help the mining process.
- Aggregation: Summary or aggregation operations are applied to the data.
  - e.g., daily sales data may be aggregated so as to compute monthly and annual total amounts.
- Normalization: where the attribute data are scaled so as to fall within a smaller, specified range. such as [−1.0 to 1.0], or [0.0 to 1.0].

Data Transformation: Strategies

- **Discretization**: Raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).
  - The labels can be recursively organized into higher-level concepts, resulting in **a concept hierarchy** for the numeric attribute.



A concept hierarchy for the attribute *price*, where an interval ($X . . . $Y] denotes the range from $X (exclusive) to $Y (inclusive).

More than one concept hierarchy can be defined for the same attribute to accommodate the needs of various users.

- **Concept hierarchy generation for nominal data**: replacing low level concepts by higher level concepts
  - i.e. attributes such as street can be generalized to higher-level concepts, like city or country.

Data Transformation by Normalization

- The measurement unit used can affect the data analysis.
  - For example, changing measurement units from meters to inches for height (2.5 cm= 1 inches), or from kilograms to pounds for weight(1 kg=2.2 pounds), may lead to very different results.
- In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or "weight."
  - To help avoid dependence on the choice of measurement units, the data should be normalized or standardized.
  - This involves transforming the data to fall within a smaller or common range such as [-1, 1] or [0.0, 1.0].
- Normalizing the data attempts to give all attributes an equal weight.
  - For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes).

Data Transformation by Normalization

- **Min-max normalization**: to [*new_minA, new_maxA*]

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

Example: Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0].

Then \$73,000 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)

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

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

- **Normalization by decimal scaling:**

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

Example: Let A = -986 to 917.

Then the maximum absolute value is 986 and j=3 which normalize A to [-0.986 to 0.917].

Data Reduction

- **Data reduction** techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.
- Mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.
- Data reduction strategies include *dimensionality reduction*, *numerosity reduction*, and *data compression*.

Data Reduction: Dimensionality Reduction

- **Dimensionality reduction** is the process of reducing the number of attributes under consideration. Including:
- ***Data compression techniques*** such as *Wavelet transforms* and *principal components analysis (PCA)*, which transform or project the original data onto a smaller space.
- ***Attribute subset selection*** which removes irrelevant, weakly relevant, or redundant attributes or dimensions.
  - Irrelevant attributes: contain no information that is useful for the data mining task at hand.
  - Redundant attributes: duplicate much or all of the information contained in one or more other attributes.
  - Why ? to improve quality and efficiency of the mining process. Mining on a reduced set of attributes reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

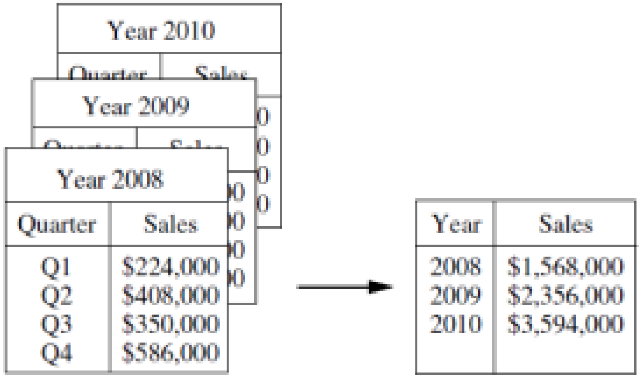
Data Reduction: Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation.
- Parametric methods:
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers).
  - Methods: Regression and Log-Linear Models.

Data Reduction: Numerosity Reduction

- **Non-parametric** methods: Do not assume models.
  - Histogram: Divide data into buckets and store average (sum) for each bucket
  - Clustering: Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only.
  - Sampling: obtaining a <u>small</u> sample s to represent the whole data set N.
    - *Key*: Choose a representative subset of the data.
  - Data cube aggregation: Data can be aggregated so that the resulting data summarize the data (smaller in volume), without loss of information necessary for the analysis task.

Data cube aggregation



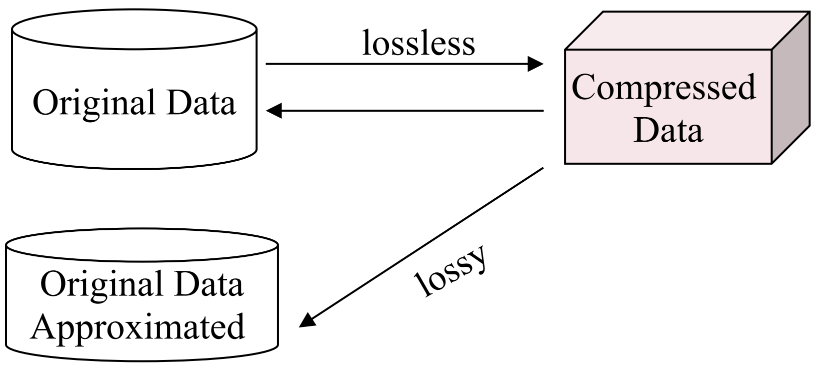
Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000

Data Reduction: Data Compression

- Obtain a reduced or "compressed" representation of the original data.
- Two types :
  - Lossless: if the original data can be reconstructed from the compressed data without any information loss.
  - Lossy: if we can reconstruct only an approximation of the original data.
- Dimensionality and numerosity reduction may also be considered as forms of 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, correlation analysis.
- Data transformation:
  - Normalization
  - Concept hierarchy generation
- Data reduction:
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression