DATA PRE-PROCESSING

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Nowadays, data is always readily available and it exist in large quantities (big data).
- Data can be obtained from a variety of different sources.
- Issues that arise: missing data problems, inconsistent data, too many and almost the same attributes/variables, outlier problem, and etc.
- Such problems affect the quality of the data.
- Data with low quality will lead to low quality data mining results.
- The data needs to be corrected to improve the quality of the data in turn improving the quality of statistical analysis and data mining.
- The process of improving/correcting the data is known as the Data Pre-processing.

DATA PRE-PROCESSING METHODS:

1. Data Integration: Combining data from multiple sources, add new attributes, removing of inappropriate attributes.
2. Data Cleaning: Manage missing data, correct inconsistent data, and manage outliers.
3. Data Reduction: Reduce the size of data through reduction of dimensions or reduction of the amount (numerosity) of data.
4. Data Transformation: Scaling the data, discretizing, and normalizing data distribution.

These techniques are not mutually exclusive, they can occur simultaneously in the same procedure.

DATA QUALITY:

- Data is defined as having quality if it meets the requirements of its use.

- Several factors that measure the quality of data

i) Accuracy

ii) Complete

iii) Uniqueness

iv) Consistent

v) Timeliness

vi) Trusted

vii) Interpretable.

EXAMPLE OF SITUATION:

- Suppose you are a manager in a company selling electronics parts.
- You are assigned to analyse the sales data for the company's branches.
- You find that the database system for branch-1 records error values, illogical data and inconsistent data for product sales record data.
- In addition, you need to get data from other branch databases to combine with the data of other branches.
- What do you need to do?

EXAMPLES OF DATA WITH POOR QUALITY:

	state.of.res	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	is.employed.fix1	Median.Income	gp	income.lt.30K	age.range
1	Alabama	1063014	F	TRUE	82000	Married	TRUE	Rented	FALSE	2	43	employed	52371	0.93506	FALSE	(25,65]
2	Alabama	1192089	M		49000	Married	TRUE	Homeowner free and clear	FALSE	2	77	missing	52371	0.1162411	FALSE	(65,Inf]
3	Allabama	16551	F		7000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	46	missing	52371	0.9906832	TRUE	(25,65]
4	Alabama	1079878	F		37200	Divorced/Separated	TRUE	Homeowner with mortgage/loan	FALSE	1	62	missing	52371	0.187356	FALSE	(25,65]
5	Alabama	502705	M	TRUE	70000	Married	FALSE	Rented	FALSE	4	37	employed	52371	0.8490238	FALSE	(25,65]
6	Alabama	674271	M	FALSE	0	Married	TRUE	Rented	TRUE	1	54	not employed	52371	0.3295085	TRUE	(25,65]
7	Alabama	15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear	FALSE	1	70	employed	52371	0.5097943	TRUE	(65,Inf]
8	Alabama	467335	M	TRUE	42600	Never Married	FALSE	Rented	FALSE	1	330	employed	52371	0.3253978	FALSE	(25,65]
9	Alabama	462569	M		22000	Widowed	TRUE	Homeowner free and clear	FALSE	0	89	missing	52371	0.5089611	TRUE	(65,Inf]
10	Alabama	1216026	M		9600	Never Married	FALSE	Rented	FALSE	6	50	missing	52371	0.5748651	TRUE	(25,65]
11	Alabama	1036358	F	TRUE	44500	Divorced/Separated	TRUE	Rented	TRUE	1	48	employed	52371	0.1778035	FALSE	(25,65]
12	Alabama	884334	M	TRUE	51000	Married	TRUE	Rented	FALSE	2	52	employed	52371	0.7030886	FALSE	(25,65]
13	A.laska	415575	M		0	Never Married	TRUE			NA	63	missing	44191	0.9561312	TRUE	(25,65]
14	Alaska	416144	F	TRUE	82000	Divorced/Separated	TRUE	Homeowner with mortgage/loan	FALSE	2	44	employed	44191	0.3066583	FALSE	(25,65]
15	Arizona	1096606	M	TRUE	52500	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	50	employed	65720	0.4211012	FALSE	(25,65]
16	Arizona	692445	M	TRUE	140000	Married	TRUE	Homeowner with mortgage/loan	FALSE	5	48	employed	65720	0.5417526	FALSE	(25,65]
17	Arizona	68013	M		-10000	Divorced/Separated	FALSE			NA	28	missing	65720	0.6294096	TRUE	(25,65]
18	Arizona	940084	M	TRUE	53000	Never Married	TRUE	Homeowner with mortgage/loan	FALSE	2	29	employed	65720	0.3583108	FALSE	(25,65]
19	Arizona	492072	F	TRUE	80000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	49	employed	65720	0.4468186	FALSE	(25,65]
20	Arizona	870909	F		4000	Married	TRUE	Homeowner free and clear	FALSE	2	57	missing	65720	0.5014896	TRUE	(25,65]
21	Arizona	1372296	F		62000	Widowed	TRUE	Homeowner free and clear	TRUE	1	62	missing	65720	0.3694147	FALSE	(25,65]
22	Arizona	958271	F	TRUE	180000	Divorced/Separated	TRUE	Rented	FALSE	1	39	employed	65720	0.3879025	FALSE	(25,65]
23	Arizona	498048	M	TRUE	95000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	60	employed	65720	0.7556033	FALSE	(25,65]
24	Arizona	211330	F		12200	Divorced/Separated	TRUE	Homeowner free and clear	FALSE	1	78	missing	65720	0.5814859	TRUE	(65,Inf]
25	Arizona	399150	M	TRUE	50000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	38	employed	65720	0.1404324	FALSE	(25,65]
26	Arizona	291564	F		28100	Widowed	TRUE	Homeowner free and clear	FALSE	1	75	missing	65720	0.002267708	TRUE	(65,Inf]
27	Arkansas	748153	F	TRUE	34200	Divorced/Separated	TRUE	Homeowner free and clear	FALSE	1	580	employed	48484	0.8591835	FALSE	(25,65]
28	Arkansas	1269051	F		137600	Widowed	TRUE	Homeowner with mortgage/loan	FALSE	1	69	missing	48484	0.6374044	FALSE	(65,Inf]
29	Arkansas	874159	F	TRUE	-7500	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	47	employed	48484	0.7697323	TRUE	(25,65]
30	Arkansas	1200487	M		0	Never Married	FALSE			NA	36	missing	48484	0.9784344	TRUE	(25,65]
31	Arkansas	253015	M	TRUE	30000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	35	employed	48484	0.5135767	FALSE	(25,65]
32	Selangor	399930	M	TRUE	55000	Divorced/Separated	TRUE	Homeowner with mortgage/loan	FALSE	2	42	employed	48484	0.7644437	FALSE	(25,65]
33	Arkansas	961665	M		0	Never Married	FALSE			NA	45	missing	48484	0.4410671	TRUE	(25,65]
34	Arkansas	356688	F	TRUE	27000	Never Married	FALSE	Rented	FALSE	1	26	employed	48484	0.6573675	TRUE	(25,65]
35	Arkansas	1358975	F	TRUE	92000	Divorced/Separated	TRUE	Homeowner with mortgage/loan	FALSE	1	46	employed	48484	0.8214495	FALSE	(25,65]
36	Arkansas	55992	F		0	Married	TRUE	Rented	FALSE	1	38	missing	48484	0.2685703	TRUE	(25,65]
37	Arkansas	1079462	F	TRUE	9500	Never Married	TRUE	Rented	FALSE	2	36	employed	48484	0.6756802	TRUE	(25,65]
38	Arkansas	1305771	F	TRUE	14400	Never Married	TRUE	Rented	TRUE	1	31	employed	48484	0.8590834	TRUE	(25,65]
39	Arkansas	450221	M	TRUE	15800	Married	FALSE	Homeowner with mortgage/loan	FALSE	2	64	employed	48484	0.2423167	TRUE	(25,65]
40	California	799565	M		1600	Never Married	FALSE			NA	23	missing	39832	0.2802194	TRUE	[0,25]

DATA INTEGRATION:

- Data Integration is the process of combining data from multiple sources.
- Referring to the case of an Electronics company, suppose you need to get data from different databases.
- Although the data from different databases may have the same information, but it might be represent in different name, attributes and etc.

i) Inconsistent attribute names:

- Example: the attribute of customer identity in branch-1 referred as "customer id", while in branch-2 database referred as "cust id".

ii) Inconsistent attribute values:

- Example: For the "Customer Name" attribute, the nominee is recorded as "W. Bill" in branch-1, while in for branch-2, they record it as " William Bill ".

DATA INTEGRATION:

- Apart from that, some information from the some database may contains too many attributes.
- Too many attributes can make data mining analysis difficult/confusing.
- Some algorithms are also difficult to run on data with high dimension.
- Basically, domain knowledge is required to determine which attributes should be retained and which can be removed.
- This procedures will make statistical analysis and data mining more efficient.

EXAMPLE OF DATA INTEGRATION:

	A	B	C	D
1	Item	Feb sales	Mar sales	Apr sales
2	Sweets	$140	$220	$160
3	Biscuits	$220	$190	$200
4	Ice-cream	$310	$320	$170

AZ report

	A	B	C	D
1	Item	Jan sales	Feb sales	Mar sales
2	Sweets	$100	$220	$320
3	Cakes	$250	$310	$280
4	Ice-cream	$110	$140	$190

... IL report

	A	B	C	D	E
1	Item	Jan sales	Feb sales	Mar sales	Apr sales
2	Sweets	$250	$140	$190	$200
3	Bisquites	$100	$310	$280	$170
4	Ice-cream	$110	$220	$320	$160
5	Cakes	$110	$140	$190	$340

... NY report

DATA CLEANING:

- Data cleaning consist of 3 main aspects :

i) Manage missing data.

ii) Fix inconsistent data.

iii) Manage outliers.

- If the data analyzed is “dirty", statistical analysis and data mining results are questionable, inaccurate or meaningless.

EXAMPLE OF DATA CLEANING:

Dirty Data

FirstName	Surname	CompanyName	Address1	Town
peter	jones	jones café	80 riverways	manchester
lisa sefton			76 the avenue	leicester
a baker		bakery baker ltd	7 main road	reading berkshire
Richard	Evans1	Richard's Treats	9 charles Street	Bracknel
Alex		The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights		Gillingham
Janine		The Janine Way	10 Fleet Place	Bracknell
Katherine	Bolton	Bolton Foods	bond Street	
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
emma	w	The Write Way	280 Bath rd	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh
Dave	Smith	Dave's Gift	po box	Leigh Lancs

Un-Standardised

Missing or misspelled

Duplications

Clean Data

FirstName	Surname	CompanyName	Address1	Town
Peter	Jones	Jones Café	80 Riverways	Manchester
Lisa	Sefton		76 The Avenue	Leicester
A	Baker	Bakery Baker Ltd	7 Main Road	Reading
Richard	Evans	Richard's Treats	9 charles Street	Bracknell
Alex	Froy	The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights	25 Camel Lane	Gillingham
Janine	Hutton	The Janine Way	10 Fleet Place	Bracknell
Katherine	Bolton	Bolton Foods	bond Street	London
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh

Correctly Standardised

Populated and Corrected

Duplications Removed

DATA REDUCTION:

- Data reduction is required to present a large data in a smaller form, yet it still retain information that is almost identical to the original data

- Data reduction consist of two main approaches:

i) Dimensional Data Reduction.

ii) Numerosity Data Reduction.

- Data reduction also aims to make data mining analysis more efficient
- Data mining algorithm will be more efficient.
- The results of the analysis will also be easier to interpret.

DIMENSIONAL DATA REDUCTION:

- The simplest method in reducing data dimension is by removing inappropriate variables.

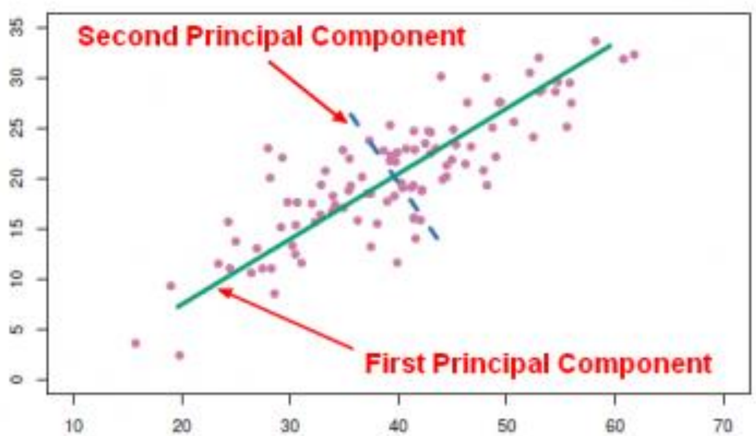
- Statistical dimension reduction techniques involve the process of forming new variables (smaller dimensions) that describe information that is almost similar to the original data.
- Principal Component Analysis, Wavelet Transformation, Factor Analysis, and etc.

- The construction of new variables that involves the aggregation of several appropriate variables is also a technique of data reduction.

EXAMPLE OF DATA DIMENSION REDUCTION:

i) Principal Component Analysis.

ii) Removal of inappropriate variables.



	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1	0.777483	0.747555	0.745291	0.818301	0.796642	0.690015	0.561432	0.702765	0.80520
x2	0.777483	1	0.733936	0.623458	0.754961	0.699861	0.567189	0.46811	0.579661	0.71280
x3	0.747555	0.733936	1	0.591841	0.697472	0.641457	0.529001	0.481284	0.536544	0.64495
x4	0.745291	0.623458	0.591841	1	0.668066	0.62058	0.493015	0.399857	0.501061	0.65653
x5	0.818301	0.754961	0.697472	0.668066	1	0.734173	0.625786	0.506842	0.627085	0.77692
x6	0.796642	0.699861	0.641457	0.62058	0.734173	1	0.588516	0.465064	0.596105	0.74475
x7	0.690015	0.567189	0.529001	0.493015	0.625786	0.588516	1	0.575315	0.653577	0.63495
x8	0.561432	0.46811	0.481284	0.399857	0.506842	0.465064	0.575315	1	0.489172	0.48503
x9	0.702765	0.579661	0.536544	0.501061	0.627085	0.596105	0.653577	0.489172	1	0.62294
x10	0.805206	0.712806	0.644959	0.656534	0.776928	0.744755	0.634956	0.485031	0.622942	1

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1						0.690015	0.561432		
x2	0.777483						0.567189	0.46811		
x3	0.747555						0.529001	0.481284		
x4	0.745291						0.493015	0.399857		
x5	0.818301						0.625786	0.506842		
x6	0.796642						0.588516	0.465064		
x7	0.690015						1	0.575315		
x8	0.561432						0.575315	1		
x9	0.702765						0.653577	0.489172		
x10	0.805206						0.634956	0.485031		

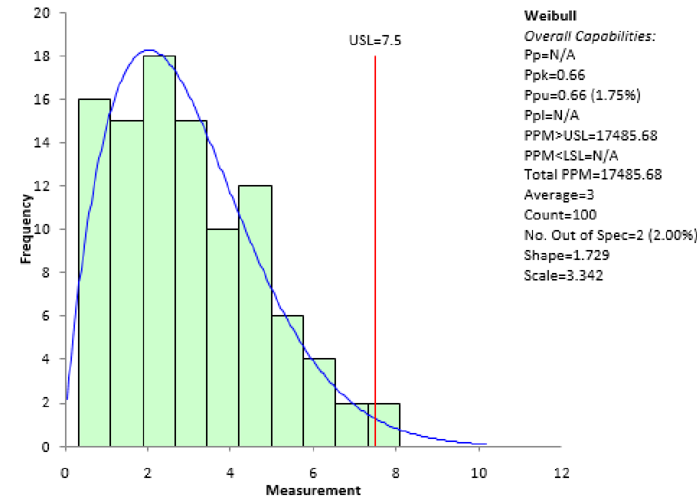
NUMEROSITY DATA REDUCTION:

- The data will be replaced with the following alternative forms:

i) Parametric Model:

Example:

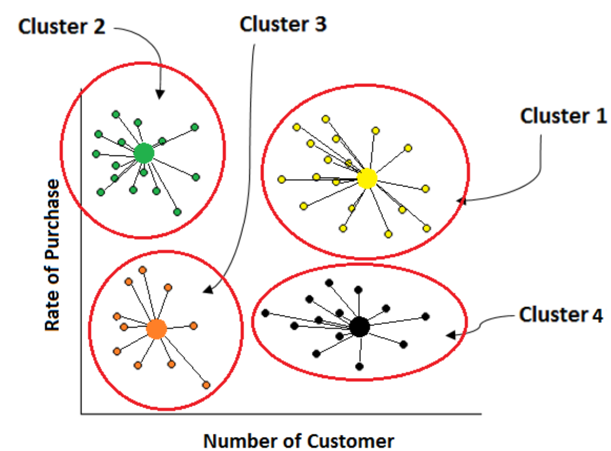
regression, log-linear model, distribution model, and etc.



ii) Non-parametric Model:

Example:

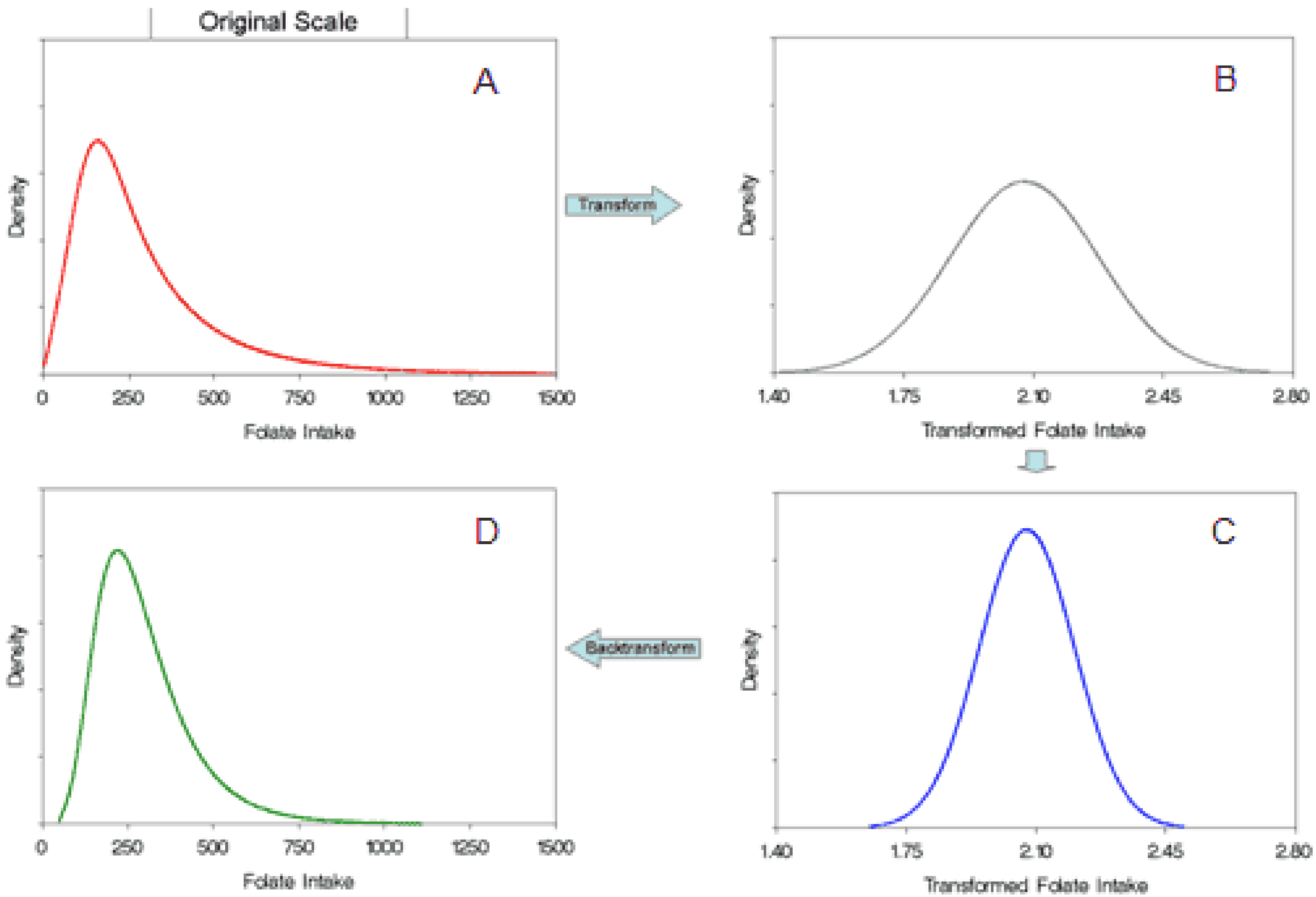
histograms, clustering, resampling technique.



DATA TRANSFORMATION:

- The process of transforming the data into a simpler and more appropriate form which corresponds to the data mining technique to be used.
- Among the methods of data transformation are Data Normalization and Data Discretization.
- Some data mining methods such as regression models and statistical tests require the assumption of normality against the data.
- If the assumptions of normality are not met, regression analysis will give inaccurate results.
- In addition, methods such as neural networks and clustering (distance-based algorithm) require data to be in the range of [0.0, 1.0].
- Thus, through the transformation method, the original data can be transformed to a normal distribution and also scaled to a certain range, such as [0.0, 1.0].

EXAMPLE OF DATA TRANSFORMATION: NORMALIZING DATA



DATA TRANSFORMATION:

- Discretization aims transform the data into a simpler form (within a certain range).
- Data that goes through the discretionary process is more "rough" than the original data.
- However, it still provides the same information, in accordance with the analysis conducted.

- Example:
- The data for the age attribute of the recorded customers are ranged from 10 to 100 years.
- Through discretization, data for age can be categorized into adolescents (10–30), adults (31–60) and seniors (> 60).

EXAMPLE OF DATA TRANSFORMATION: DISCRETIZATION,

- Original Data:

	years employed	yearly income	position	gender	took holidays	:rience in the indu	name
1	13.000	42000.000	office worker	male	0	12.000	Mark
2	3.000	37000.000	technical staff	female	0	4.000	Michelle
3	5.000	36000.000	technical staff	male	0	8.000	Andy
4	15.000	46000.000	office worker	male	1	17.000	Bob
5	2.000	42000.000	office worker	female	1	15.000	Delilah
6	10.000	41000.000	office worker	female	1	14.000	Marlene
7	5.000	33000.000	technical staff	male	0	5.000	Oli
8	12.000	32000.000	technical staff	male	1	12.000	Tom
9	10.000	39000.000	office worker	female	0	14.000	Tanya
10	12.000	43000.000	office worker	female	1	17.000	Rebeccah
11	1.000	37000.000	technical staff	female	0	1.000	Gill
12	14.000	42000.000	office worker	male	0	16.000	Hank

- Data for the variables "years employed" & "yearly income" were transformed through discretization.

	years employed	yearly income	position	gender	took holidays	experience in the industry	name
1	≥ 8	≥ 39000	office worker	male	0	≥ 9	Mark
2	< 8	< 39000	technical staff	female	0	< 9	Michelle
3	< 8	< 39000	technical staff	male	0	< 9	Andy
4	≥ 8	≥ 39000	office worker	male	1	≥ 9	Bob
5	< 8	≥ 39000	office worker	female	1	≥ 9	Delilah
6	≥ 8	≥ 39000	office worker	female	1	≥ 9	Marlene
7	< 8	< 39000	technical staff	male	0	< 9	Oli
8	≥ 8	< 39000	technical staff	male	1	≥ 9	Tom
9	≥ 8	≥ 39000	office worker	female	0	≥ 9	Tanya
10	≥ 8	≥ 39000	office worker	female	1	≥ 9	Rebeccah
11	< 8	< 39000	technical staff	female	0	< 9	Gill
12	≥ 8	≥ 39000	office worker	male	0	≥ 9	Hank

SUMMARY:

The figure shows a summary of the data pre-processing methods discussed in this topic.

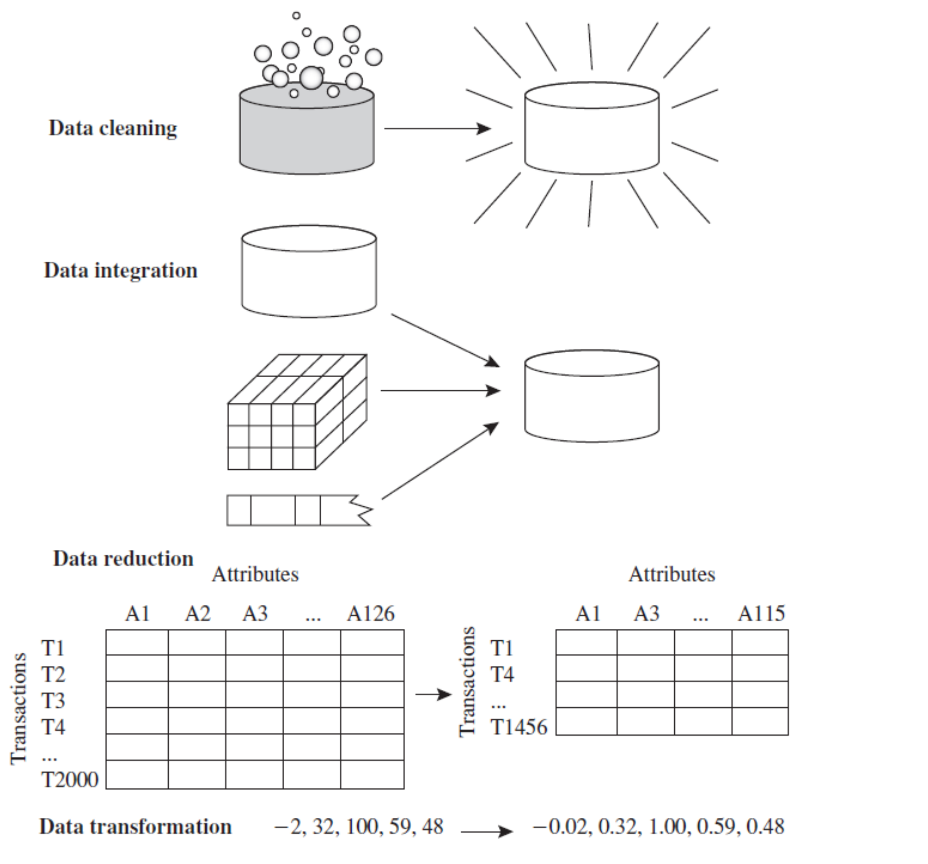


Figure 3.1 Forms of data preprocessing.

REFERENCES:

- Aggarwal, C.C. (2015). *Data Mining. The Textbook*. Springer, New York.
- García, S., Luengo, J., Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer, New York.
- Jafari, R. (2022). *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics 1st Edition*. Packt Publishing
- Kuhn, M., Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., Herrera, F. (2020). *Big Data Preprocessing*. Springer, Switzerland.

NEXT TOPIC:

Data Integration