

PRA-PEMROSESAN DATA

STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- Data pada masa kini adalah mudah diperolehi dan ianya adalah dalam kuantiti yang besar.
- Data juga boleh diperoleh dari pelbagai sumber yang berbeza.
- **Wujud situasi:** masalah data yang lenyap, data tidak konsisten, atribut/pemboleh ubah yang terlalu banyak dan hampir sama, masalah data pencil (*outliers*), dan lain-lain.
- Permasalahan tersebut memberi kesan kepada kualiti data.
- Data dengan kualiti rendah akan membawa kepada hasil perlombongan data yang rendah kualiti.
- Data ini perlu dibaiki untuk meningkatkan kualiti data seterusnya meningkatkan kualiti analisis statistik dan perlombongan data.
- Proses ini dikenali sebagai kaedah Pra-pemprosesan Data atau teknik Kejuruteraan Fitur.



- Beberapa teknik kaedah Pra-pemprosesan Data:

i) **Integrasi Data:** Menggabungkan data dari pelbagai sumber, memasukan atribut baru, penyingkiran atribut yang tidak sesuai.

ii) **Pembersihan Data:** Mengurus data lenyap, membetulkan data yang tidak konsisten, dan menguruskan data pencil.

iii) **Penurunan Data:** Mengurangkan saiz data menerusi pengurangan dimensi, pengurangan amaun (*numerosity*) data ataupun pengagregatan data.

iv) **Penjelmaan Data:** Menskalakan data, pendiskretan data, menormalkan taburan data.

- Teknik-teknik ini bukanlah saling eksklusif, ianya boleh berlaku serentak dalam proses yang sama.

- **Contoh:**

- Pembersihan data juga melibatkan penjelmaan data.
- Data integrasi juga melibatkan data yang tidak konsisten (pembersihan data).
- Data integrasi juga melibatkan proses penurunan data.



KUALITI DATA:

- Suatu data dikatakan berkualiti jika ianya memenuhi keperluan kegunaannya.
- Beberapa faktor yang mengukur kualiti data:
 - i) *Ketepatan*
 - ii) *Lengkap*
 - iii) *Konsisten*
 - iv) *Ketepatan Masa*
 - v) *Boleh dipercayai*
 - vi) *Boleh ditafsir.*



CONTOH SITUASI:

- Misalkan anda ialah pengurus di sebuah syarikat menjual barang Elektronik.
- Anda ditugaskan untuk menganalisis data jualan bagi cawangan-cawangan syarikat.
- Anda mendapati sistem pangkalan data bagi cawangan-1 merekodkan nilai-nilai ralat, data yang tidak logik dan data yang tidak konsisten bagi data rekod jualan produk.
- Disamping itu, anda perlu mendapatkan data daripada pangkalan data cawangan lain untuk menggabungkan dengan data cawangan-cawangan lain.
- Apa yang perlu anda lakukan?



CONTOH DATA “TIDAK BERKUALITI”:

	state.of.res	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	is.employed.fixl	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]



1. INTEGRASI DATA:

- **Integrasi Data** ialah proses menggabungkan data dari pelbagai sumber.
- Merujuk kepada kes syarikat Elektronik, iaitu anda perlu mendapatkan data daripada pangkalan data yang berbeza.
- Data dari pangkalan data yang berbeza mungkin mempunyai data yang sama, namun dengan atribut nama yang berbeza.

i) Nama atribut yang tidak konsisten:

- **Contoh:** atribut bagi no pelanggan bagi pangkalan data cawangan-1 ialah *customer id* , manakala bagi pangkalan data cawangan-2 ialah *cust id*.

ii) Nilai atribut yang tidak konsisten:

- **Contoh:** Bagi atribut “Nama Pelanggan”, penama direkod sebagai “W. Bill” pangkalan data cawangan-1, manakala dalam bagi pangkalan data cawangan-2 ialah “William Bill”.





1. INTEGRASI DATA:

- Selain itu, anda juga mungkin mendapati maklumat dari pangkalan data juga mengandungi terlalu banyak atribut.
- Terlalu banyak atribut boleh menjadikan analisis perlombongan data sukar/keliru.
- Malah, beberapa al-Khwarizmi (*algorithm*) juga sukar untuk dijalankan terhadap data berdimensi tinggi.
- Pengetahuan domain diperlukan untuk menentukan atribut yang sepatutnya dikekalkan dan yang boleh dikeluarkan.
- Ini akan menjadikan analisis statistik dan perlombongan data lebih efisien.



CONTOH INTEGRASI DATA:

	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
◀ ▶ ...		<u>NY report</u>			



2. PEMBERSIHAN DATA:

- Pembersihan Data melibatkan aspek:
 - i) Mengurus data lenyap.
 - ii) Membaiki data yang tidak konsisten.
 - iii) Mengurus data pencil (*outliers*).
- Jika data yang dianalisis adalah “tidak bersih”, analisis statistik & perlombongan data adalah diragui/tidak tepat ataupun tidak memberi makna.



CONTOH PEMBERSIHAN DATA:

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	Bracknell
Alex		The Alex Centre	13-15 athol street	Bournemouth
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	Bournemouth
Derren	Knight0	Derrens' Delights	25 Camel Lane	Gillingham
Janine	Hulton	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



3. PENURUNAN DATA:

- **Penurunan Data** diperlukan untuk mempersembahkan data dalam bentuk yang lebih kecil, namun masih mengekalkan maklumat yang hampir sama dengan data asal
- Penurunan Data terdiri daripada:
 - i) Penurunan Dimensi Data
 - ii) Penurunan amaun (*Numerositi*) Data.
- Penurunan data juga bertujuan untuk menjadikan analisis perlombongan data lebih efisien.
- Al-Khwarizmi akan menjadi lebih cekap terhadap data yang berdimensi lebih rendah.
- Hasil analisis juga akan menjadi lebih mudah untuk ditafsir.



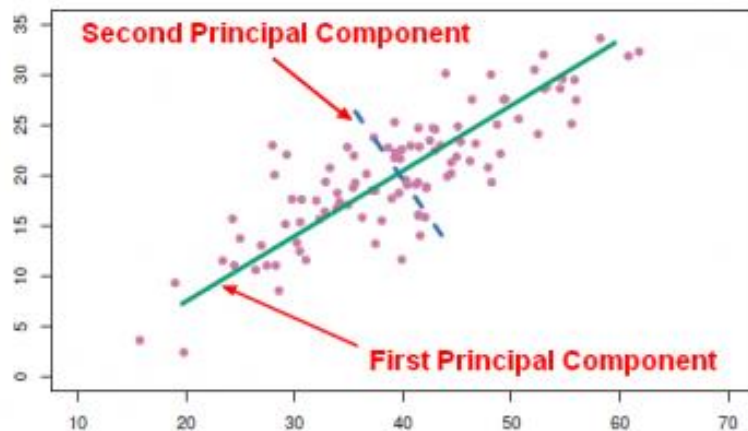
3.1 PENURUNAN DIMENSI DATA:

- Penurunan dimensi melibatkan proses membentuk p/ubah baru (dimensi lebih kecil) yang menerangkan hampir keseluruhan maklumat data asal.
- (Analisis Komponen Utama, Penjelmaan Wavelet, Analisis Faktor, dan lain-lain)
- Penyingkiran p/ubah yang tidak sesuai juga merupakan proses penurunan dimensi data (diterangkan dalam integrasi data).
- Penurunan dimensi data juga boleh dibuat dengan pembinaan p/ubah baru yang melibatkan pengagregatan beberapa p/ubah lain.



CONTOH PENURUNAN DIMENSI DATA:

- i) Analisis Komponen Utama.
- ii) Penyingkiran p/ubah yang tidak sesuai.



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

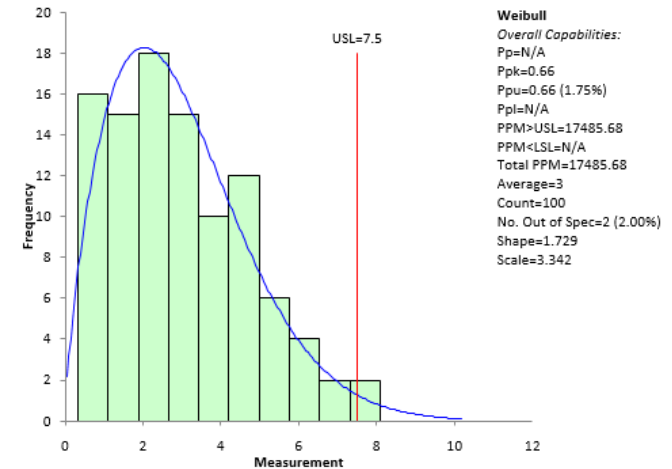


3.2 PENURUNAN AMAUN (NUMEROSITI) DATA:

- Data akan digantikan dengan bentuk alternatif berikut:

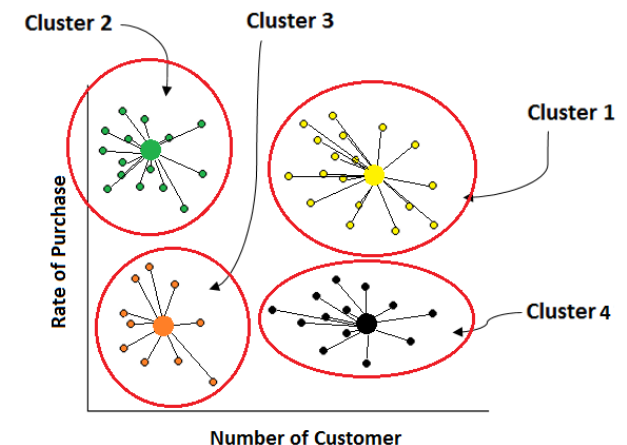
i) Model Berparameter:

- **contoh:** regression, model log-linear, dan lain-lain.



ii) Model tak berparameter:

- **contoh:** histograms, kluster, pensampelan semula.

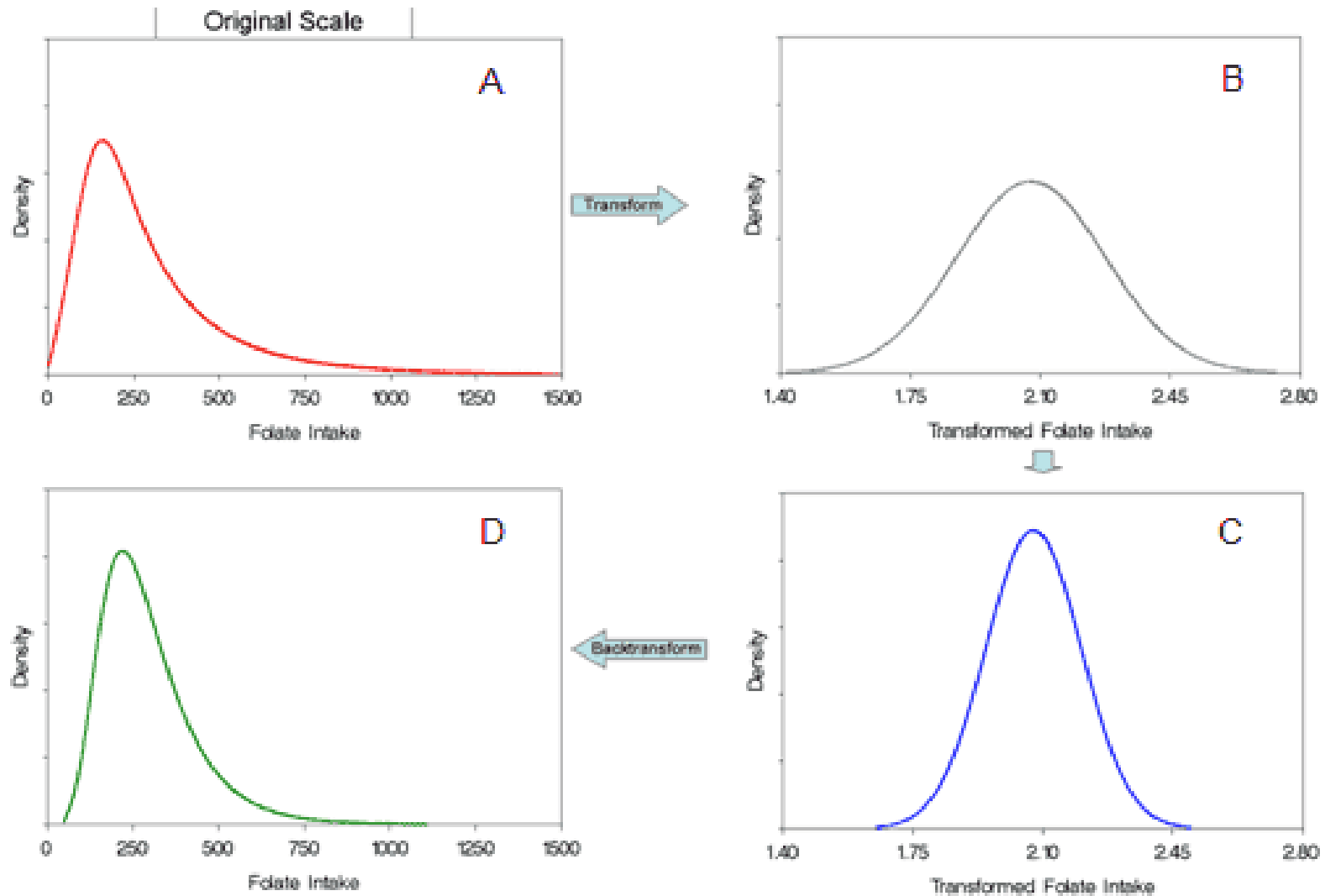


4. PENJELMAAN DATA:

- Menjelmakan data kepada bentuk yang lebih ringkas dan bersesuaian dengan analisis perlombongan data yang akan dijalankan.
- Antara kaedah penjelmaan data ialah menerusi Penormalan dan Pendiskretan Data.
- Sebahagian analisis perlombongan data seperti model regresi memerlukan andaian kenormalan terhadap data.
- Jika andaian kenormalan tidak dipenuhi, analisis regresi akan memberikan hasil yang tidak tepat.
- Disamping itu, kaedah seperti rangkaian neural dan pengkelompokan (al-Khwarizmi berasaskan jarak) memerlukan data dalam julat $[0.0, 1.0]$.
- Maka, menerusi kaedah penjelmaan, data asal akan dijelmakan kepada taburan normal dan juga perlu diskalakan kepada julat tertentu, seperti $[0.0, 1.0]$.



CONTOH PENJELMAAN DATA: PENORMALAN



4. PENJELMAAN DATA:

- Pendiskretan dibuat untuk menjelmakan data kepada bentuk yang lebih ringkas (dalam julat tertentu).
- Data yang melalui proses pendiskretan adalah lebih “kasar” daripada data asal.
- Namun, ianya masih boleh memberikan maklumat yang sama, sesuai dengan analisis yang dijalankan.
- **Contoh:** data bagi atribut umur pelanggan yang direkodkan ialah antara 10 hingga 100 tahun.
- Menerusi pendiskretan, data umur boleh dikategorikan kepada remaja (10-30), dewasa (31-60) dan warga emas (>60).



CONTOH PENJELMAAN DATA: PENDISKRETAN

■ Data Asal:

	years employed	yearly income	position	gender	took holidays	experience 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 p/ubah “years employed” & “yearly income” yang dijemakan menerusi pendiskretan:

	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



RINGKASAN:

Rajah menunjukkan ringkasan bagi kaedah-kaedah pra pemprosesan data yang dibincangkan dalam topik ini.

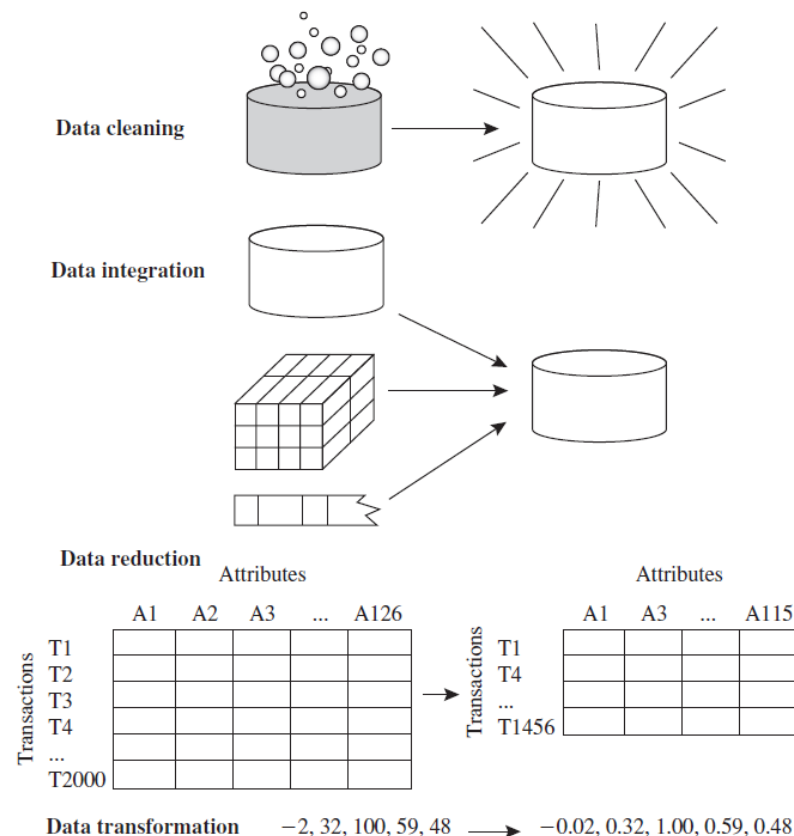


Figure 3.1 Forms of data preprocessing.



RUJUKAN:

- 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.



TOPIK SETERUSNYA:

Integrasi Data

