

PENJELMAAN DATA DAN PENDISKRETAN

STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- Dalam penjelmaan data, data diubahsuai menjadi bentuk yang lebih sesuai sebelum analisis perlombongan data dijalankan.
- Beberapa teknik penjelmaan data:

1. Penormalan:

- Melibatkan proses menskalakan semula nilai-nilai atribut.

Contoh:

- Proses menskalakan nilai atribut data asal (0–100) kepada julat yang lebih kecil, (0–1). Terutama bagi data dengan pelbagai atribut yang tidak sama unit ukuran.
- Proses menjadikan data tertabur secara taburan Normal.
- Kebanyakan kaedah dalam statistik dan perlombongan data memerlukan andaian kenormalan data.

contoh: model Regresi.



PENGENALAN:

2. Pendiskretan:

- Proses menjelmakan nilai-nilai atribut (**contoh:** umur) kepada nilai dalam bentuk selang (**contoh:** 0–10, 11–20, 20–40)
- Atau dalam bentuk konseptual (**contoh:** kanak-kanak, remaja, dewasa, warga emas).

3. Penjelmaan Atribut:

- Atribut baru dibentuk daripada gabungan atau penjelmaan atribut-atribut yang telah ada dalam data.

4. Pelicinan (*smoothing*) dan lain-lain kaedah penjelmaan.




PENORMALAN:

- Bertujuan untuk menskalakan semula nilai-nilai atribut ataupun menjadikan taburan data menghampiri taburan Normal:

i. Penormalan Min-Max:

- Melibatkan penjelmaan linear terhadap nilai-nilai data.
- Misalkan \min_x dan \max_x ialah nilai minimum dan nilai maksimum bagi data/atribut X.
- Kita mahu menskalakan data X daripada julat $[\min_x, \max_x]$ kepada julat $[\text{baru_min}_x, \text{baru_max}_x]$.
- Penormalan ini akan menjelmakan semua data/nilai atribut X kepada data atribut V dalam selang $[\text{baru_min}_x, \text{baru_max}_x]$.
- Ini dibuat menerusi persamaan:

$$V = \frac{[X - \min(X)] \times [\text{baru_max}(X) - \text{baru_min}(X)]}{\max(X) - \min(X)} + \text{baru_min}(X)$$


ii. Penormalan skor-Z:

- Juga dikenali sebagai penormalan min-sifar.
- Data/nilai atribut X dijelmakan kepada data skor-Z menerusi persamaan:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

- dengan μ_X dan σ_X ialah min dan sisihan piawai bagi atribut X .
- Jika μ_X dan σ_X tidak diketahui, ianya akan dianggarkan daripada sampel.
- Nilai p/ubah Z akan mempunyai min 0 dan sisihan piawai 1.
- Kaedah ini lebih sesuai dari penormalan min-max jika terdapat data pencil.



iii. Penormalan berdasarkan penskalaan perpuluhan:

- Menjelmakan data dengan memindahkan titik perpuluhan bagi nilai atribut X .
- **Contoh:** 3600 dijemakan kepada 0.36.
- Bilangan titik perpuluhan yang dipindahkan bergantung pada nilai mutlak maksimum X .
- Ini dibuat menerusi persamaan:

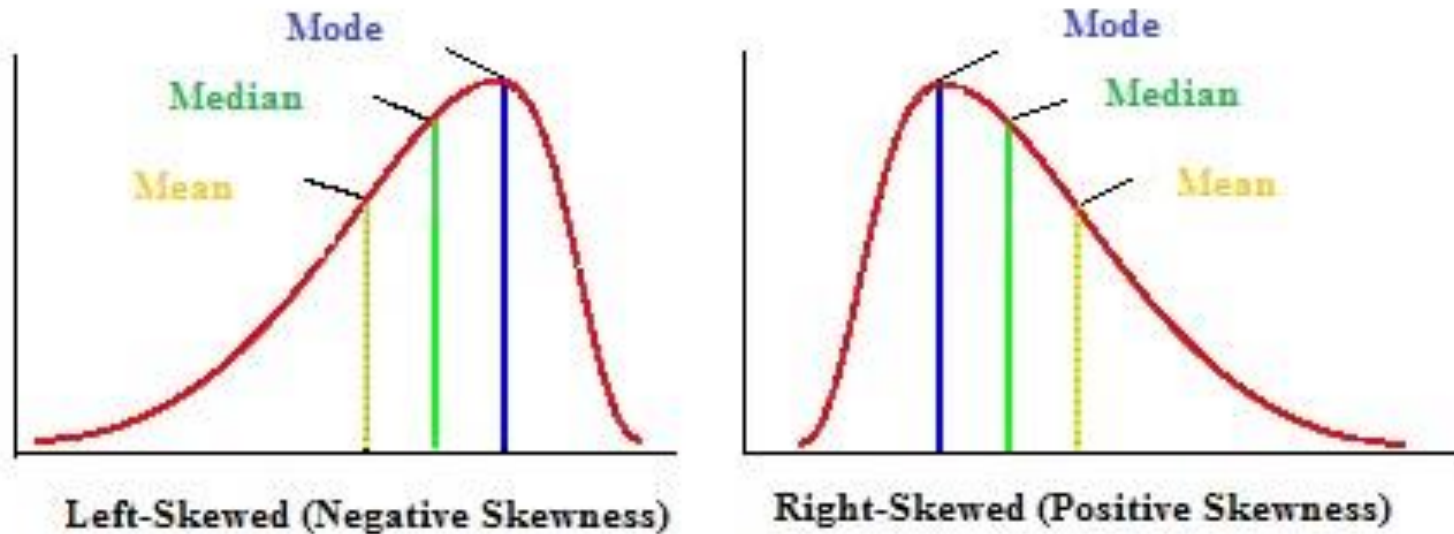
$$v'_i = \frac{v_i}{10^j}$$

- j ialah integer terkecil sedemikian hingga $\max(v'_i) < 1$.



iv. Penormalan taburan data:

- Penjelmaan jenis ini perlu dijalankan jika data adalah pincang ke kanan (positif) atau ke kiri (negatif).



- Penjelmaan ini melibatkan fungsi matematik terhadap setiap nilai data.



- Beberapa fungsi matematik yang digunakan dalam penormalan taburan data ialah:

a) Logaritma:

- Penjelmaan menerusi fungsi **$\log(x)$** sesuai jika varians bagi data didapati meningkat terhadap min data.
- Ianya juga sesuai untuk data kadar pertumbuhan yang biasanya mempunyai taburan eksponen.

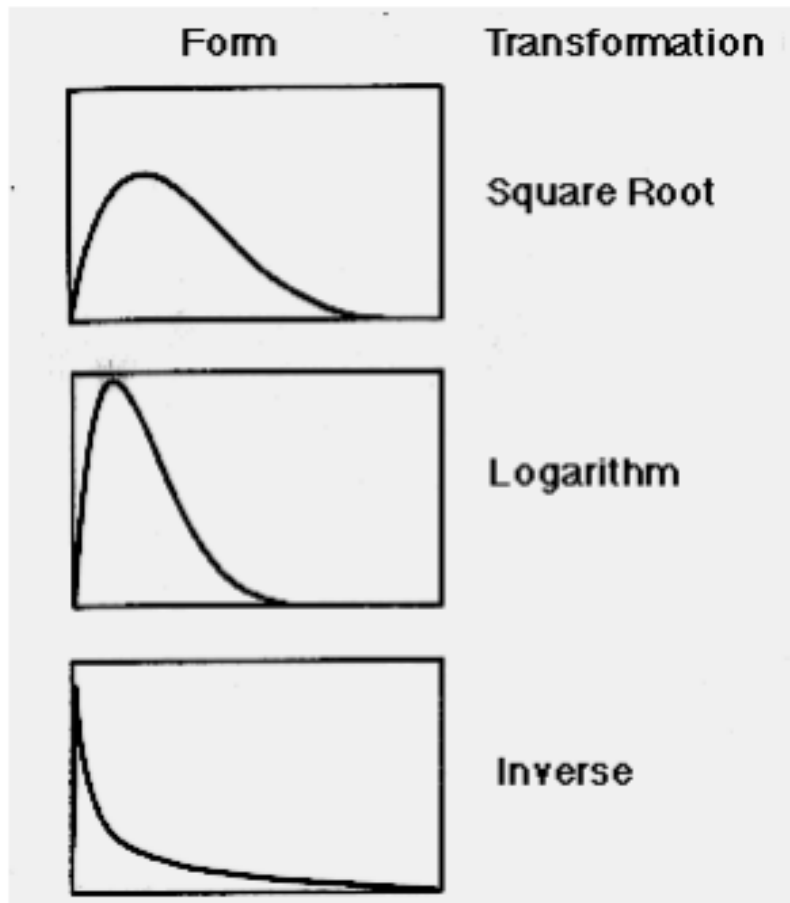
b) Jika logaritma tidak sesuai, beberapa fungsi lain boleh dicuba:

- **Penjelmaan Salingan ($1/x$).**
- **Penjelmaan Punca Kuasa Dua ($x^{1/2}$).**
- **Penjelmaan Arcsine ($\text{asin}(x)$):** dikenali sebagai penjelmaan sudut dan berguna untuk data jenis peratusan dan perkadaran yang tidak tertabur secara Normal.

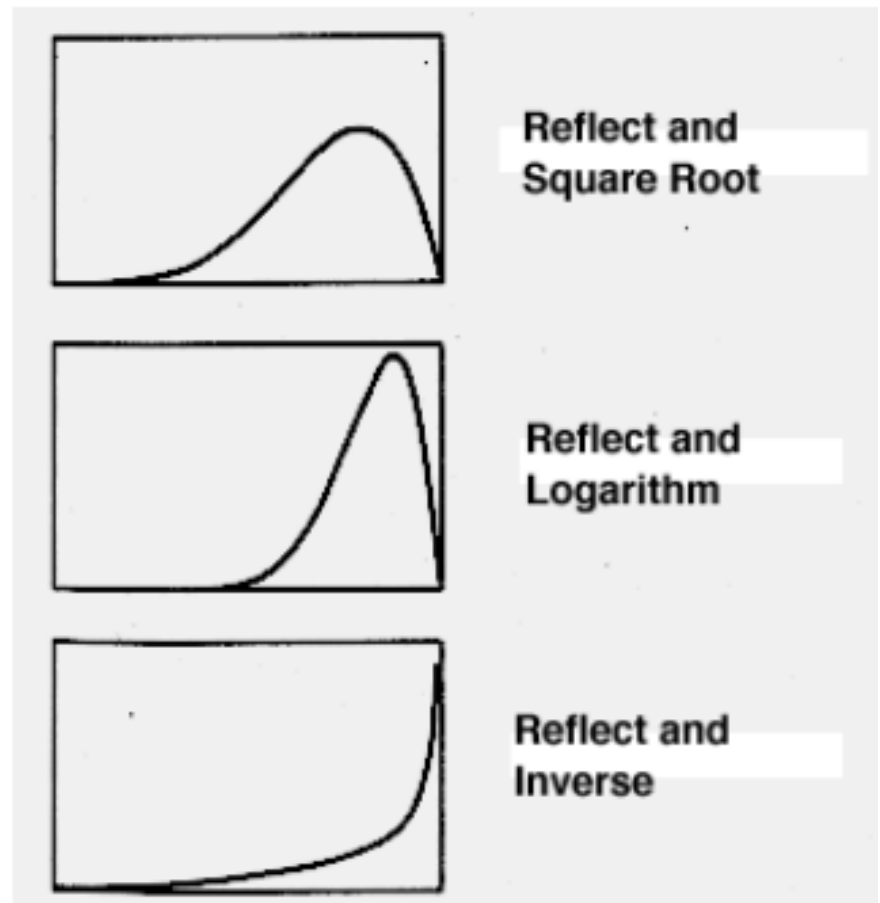


- Rajah tersebut mencadangkan fungsi matematik yang sesuai bergantung kepada darjah kepencongan bagi data taburan asal.

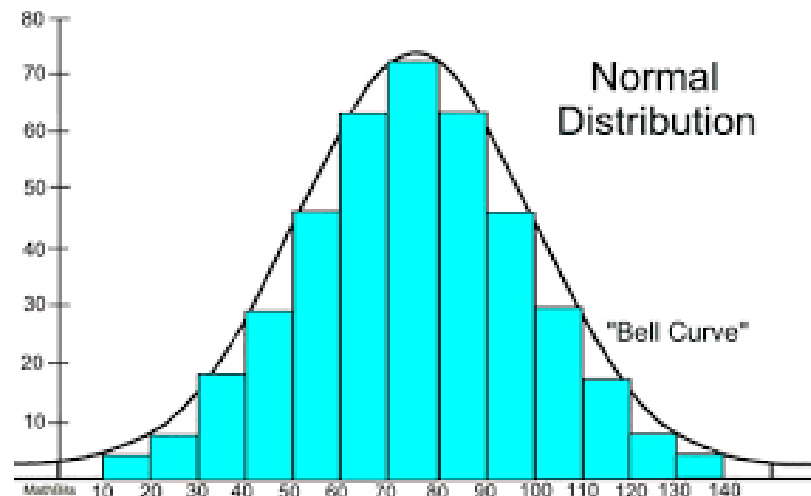
Positively skewed data



Negatively skewed data



- Data yang pincang ke kiri (negatif) memerlukan penjelmaan terpantul (*reflected transformation*).
- Bermaksud, setiap data perlu dipantulkan sebelum penjelmaan dibuat.
- Pantulan p/ubah ini dibuat menerusi pembentukan p/ubah baru dengan nilai suatu pemalar, k ditolakkan dengan data asal.
- Pemalar k dikira dengan menambahkan 1 kepada nilai terbesar p/ubah asal, $k = (\max(x) + 1)$.
- Seterusnya, p/ubah terpantul (P) dihitung menerusi: $P = k - X$
- Penormalan taburan data bertujuan menjadikan data menghampiri bentuk taburan Normal.



- Tabachnick & Fidell (2007) dan Howell (2007) memberikan tatacara berikut untuk penjelmaan data berdasarkan kepencongan taburan data asal.

Data Taburan Asal	Teknik Penjelmaan yang dicadangkan
Kepencongan Positif Sederhana	Kuasa-Dua, $Y = X^2$
Kepencongan Positif Ketara	Logarithma, $Y = \log_{10}(X)$
Kepencongan Negatif Sederhana	Punca Kuasa-Dua, $Y = \sqrt{k - X}$
Kepencongan Negatif Ketara	Logarithma, $Y = \log_{10}(k - X)$

* Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston: Allyn and Bacon.

* Howell, D. C. (2007). Statistical methods for psychology (6th ed.). Belmont, CA: Thomson Wadsworth.



MENILAI KENORMALAN DATA:

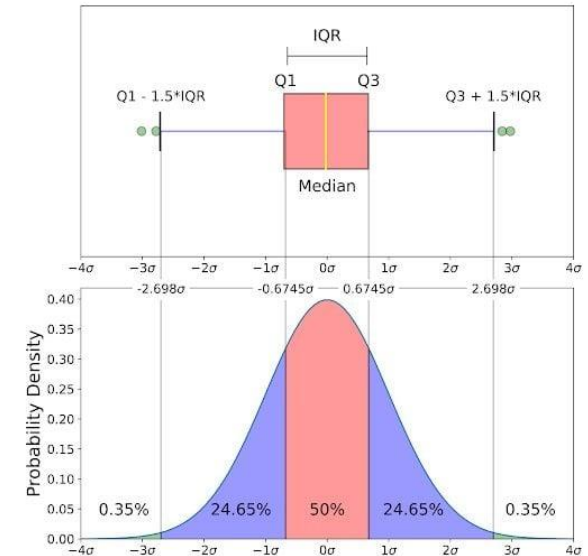
i. Histogram dan plot kotak.

ii. Plot Normal Kuantil.

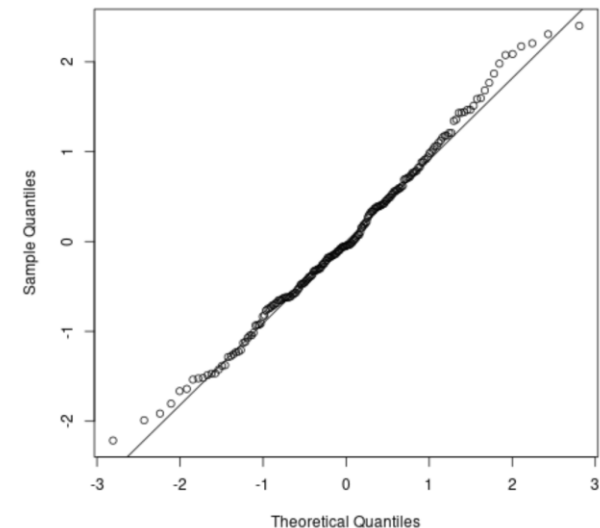
- juga dikenali sebagai Plot Kebarangkalian Normal.

iii. Ujian Kebagusan penyuaian:

- i) Ujian Kolmogorov-Smirnov.
- ii) Ujian Shapiro-Wilk.
- iii) Ujian Anderson-Darling.

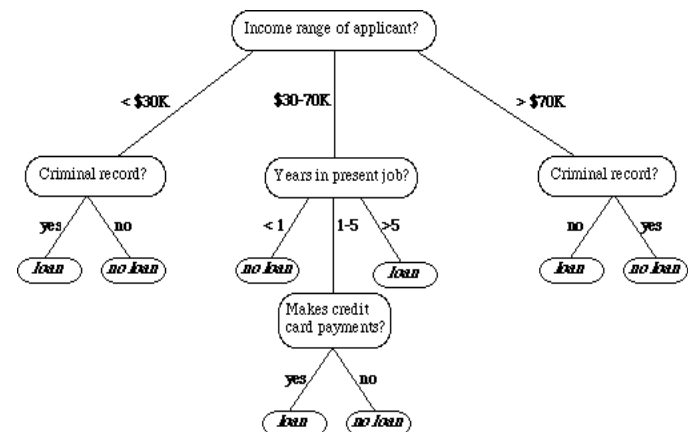
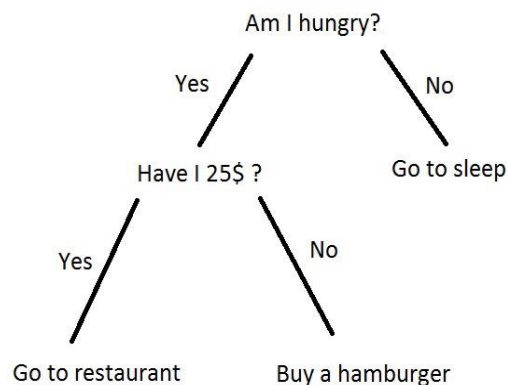


Normal Q-Q Plot



PENDISKRETAN:

- **Pendiskretan:** Membahagikan data atribut kepada beberapa selang.
- Data dalam bentuk selang akan digunakan untuk menggantikan data sebenar.
- Sebahagian kaedah perlombongan data hanya boleh dijalankan terhadap data diskrit. **Contoh:** Pokok-keputusan (*decision trees*).
- Pendiskretan merupakan pendekatan dalam penurunan data untuk menjadikan algoritma/al-Khwarizmi perlombongan data lebih efisien.
- Pendiskretan boleh dilakukan secara berulang terhadap atribut.



- Menerusi pendiskretatan, atribut dalam bentuk selanjar akan ditukarkan kepada atribut dalam bentuk diskrit ataupun selang.

Contoh:

Selanjar: Jumlah pendapatan, $1000 < X < 10000$.

Selang: 1000-2000, 2000-3000, >3000.

Diskrit/berkategori: 1=pendapatan rendah, 2=pendapatan sederhana, 3=pendapatan tinggi

- Tujuan pendiskretan ialah menurunkan bilangan nilai atribut selanjar dengan mengumpulkannya kepada bilangan b selang/bin.
- Isu penting dalam pendiskretan ialah bagaimana untuk memilih bilangan selang/bin.



- **Dua pendekatan:** pendekatan terselia (*supervised*) dan pendekatan tidak terselia (*unsupervised*).
- **Pendekatan tidak terselia:** Tiada label kelas diketahui. Selang pendiskretan boleh dijalankan terus terhadap data.
- **Pendekatan terselia:** Jika label kelas diketahui, kaedah pendiskretan perlu memanfaatkan maklumat ini, dan al-Khwarizmi/algorithm model terselia boleh digunakan.
- Kaedah pendiskretan perlu memaksimumkan ketergantungan antara nilai atribut dengan label kelas dan meminimumkan kehilangan maklumat.



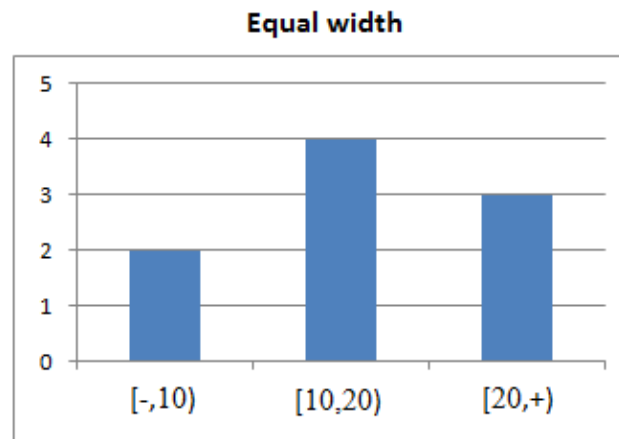
PENDISKRETAN TAK TERSELIA:

i. Pendiskretan Data menerusi pengetahuan Domain:

- Dibuat secara manual.
- Namun, saintis data perlulah mempunyai hujah yang sesuai berkaitan pembahagian selang tersebut.

ii. Pendiskretan Sama-Lebar (*Equal-width*):

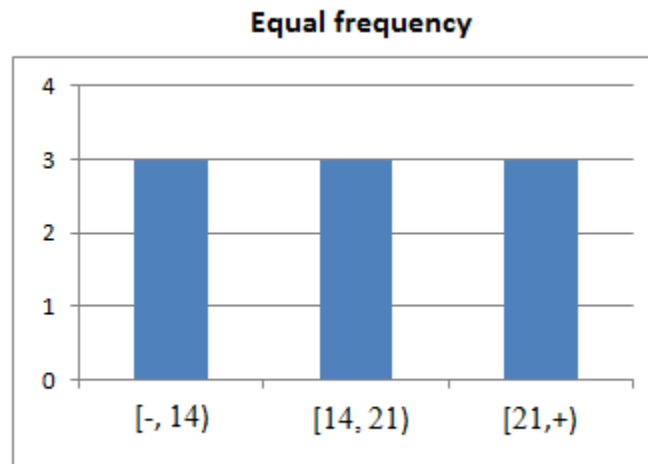
- Al-khwarizmi ini menggunakan maklumat minimum (A) dan maksimum (B) bagi atribut, X_i .
- Seterusnya, lebar selang pendiskretan dibuat menerusi: $W = (B - A)/N$.



PENDISKRETAN TAK TERSELIA:

iii. Pendiskretan Sama-Kekerapan (*Equal-frequency*):

- Al-khwarizmi ini menggunakan maklumat minimum (A) dan maksimum (B) bagi atribut, X_i .
- Seterusnya semua nilai X_i ditertibkan dalam susunan menaik.
- Lebar selang pendiskretan ditentukan berdasarkan bilangan cerapan yang sama dalam setiap selang.



PENDISKRETAN TERSELIA:

- Pendiskretan terselia mengambilkira maklumat kelas dalam set data.

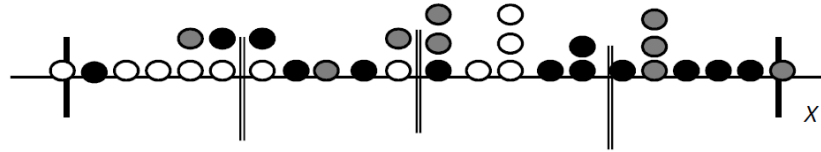


Fig. 6.2 Distribution of values belonging to three classes {white, gray, black} over variable X .

- Pelbagai al-khwarizmi pendiskretan terselia boleh dijalankan menggunakan R:
 - i) Pendiskretan menerusi Al-khwarizmi Chi2.
 - ii) Pendiskretan menerusi Al-khwarizmi ChiMerge.
 - iii) Pendiskretan menerusi Al-khwarizmi Atas-Bawah (*Top-down*).
 - iv) Pendiskretan menerusi Al-khwarizmi MDLP (*Minimum Description Length Principle*).
 - v) Dan banyak lagi.



PELBAGAI AL-KHWARIZMI PENDISKRETAN:

Equal Width Discretizer	EqualWidth
Equal Frequency Discretizer	EqualFrequency
<i>No name specified</i>	Chou91
Adaptive Quantizer	AQ
Discretizer 2	D2
ChiMerge	ChiMerge
One-Rule Discretizer	1R
Iterative Dichotomizer 3 Discretizer	ID3
Minimum Description Length Principle	MDLP
Valley	Valley
Class-Attribute Dependent Discretizer	CADD
ReliefF Discretizer	ReliefF
Class-driven Statistical Discretizer	StatDisc
<i>No name specified</i>	NBIterative
Boolean Reasoning Discretizer	BRDisc
Minimum Description Length Discretizer	MDL-Disc
Bayesian Discretizer	Bayesian
<i>No name specified</i>	Friedman96
Cluster Analysis Discretizer	ClusterAnalysis
Zeta	Zeta
Distance-based Discretizer	Distance
Finite Mixture Model Discretizer	FMM

<i>No name specified</i>	Butterworth04
<i>No name specified</i>	Zhang04
Khiops	Khiops
Class-Attribute Interdependence Maximization	CAIM
Extended Chi2	Extended Chi2
Heterogeneity Discretizer	Heter-Disc
Unsupervised Correlation Preserving Discretizer	UCPD
<i>No name specified</i>	Multi-MDL
Difference Similitude Set Theory Discretizer	DSST
Multivariate Interdependent Discretizer	MIDCA
MODL	MODL
Information Theoretic Fuzzy Partitioning	ITFP
<i>No name specified</i>	Wu06
Fast Independent Component Analysis	FastICA
Linear Program Relaxation	LP-Relaxation
Hellinger-Based Discretizer	HellingerBD
Distribution Index-Based Discretizer	DIBD
Wrapper Estimation of Distribution Algorithm	WEDA
Clustering + Rought Sets Discretizer	Cluster-RS-Disc
Interval Distance Discretizer	IDD
Class-Attribute Contingency Coefficient	CACC
Rectified Chi2	Rectified Chi2

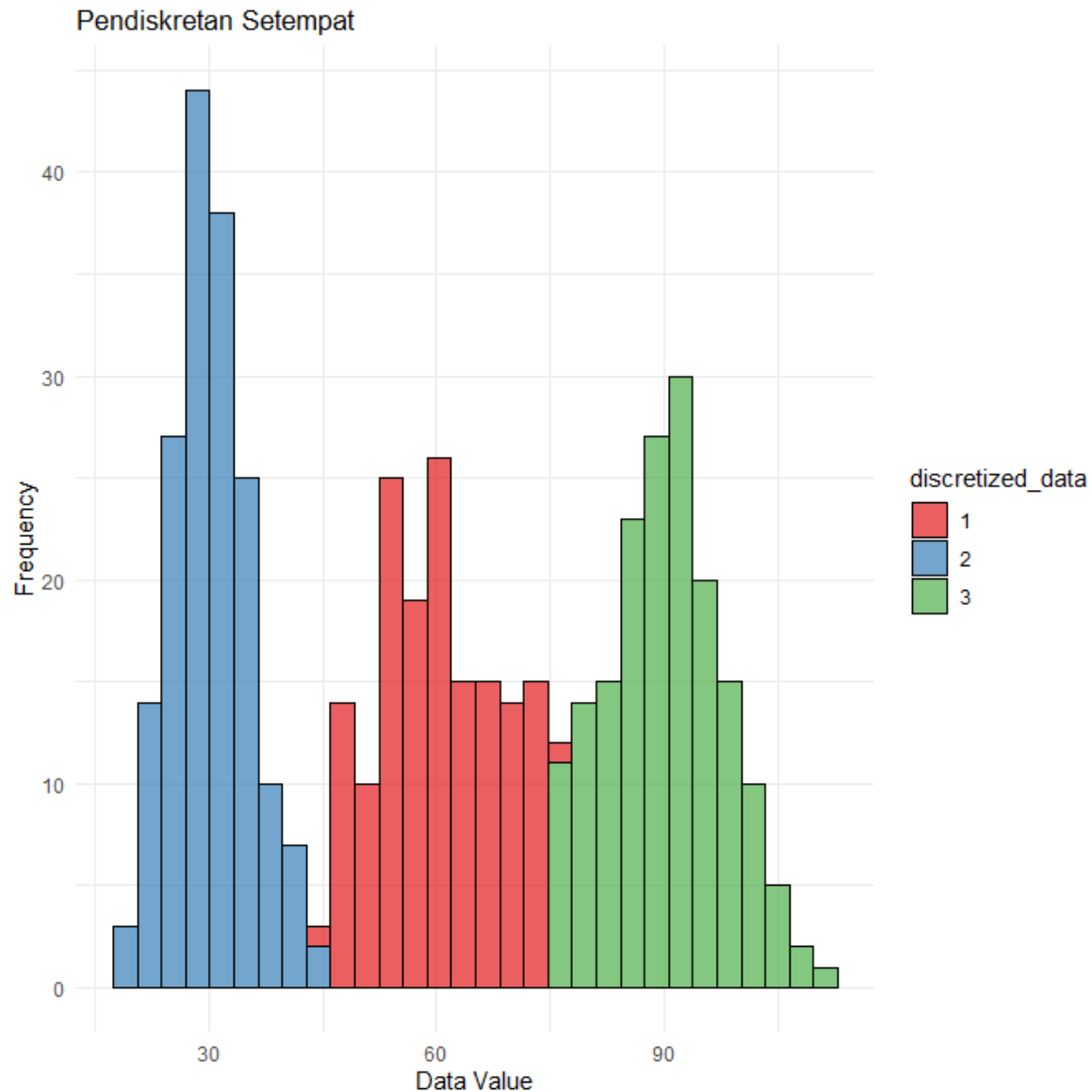


PENDISKRETAN SETEMPAT:

- Pendiskritan setempat (*local discretization*) merujuk kepada kaedah pendiskretan data yang mempertimbangkan ciri setempat dalam taburan data.
- Pendekatan ini berguna apabila kita berurusan dengan data yang bersifat heterogen, dengan segmen data yang berbeza memerlukan strategi pendiskretan yang berbeza.
- **Contoh:** Dalam set data dengan ketumpatan taburan data yang berbeza-beza, pendiskretan setempat perlu untuk membina; i) bin-bin yang lebih kecil di kawasan yang mempunyai ketumpatan data yang tinggi, dan ii) bin-bin yang lebih besar di kawasan ketumpatan data yang rendah.
- Dengan menyesuaikan proses pendiskretan kepada konteks setempat, kaedah ini boleh membantu mengekalkan corak dan hubungan penting dalam data.



PENDISKRETAN SETEMPAT:



PENJELMAAN DATA MEMBENTUK ATRIBUT BARU:

- Atribut baru dibina daripada beberapa gabungan atau penjelmaan atribut-atribut yang sedia ada dalam data.
- **Contoh:**
 - i) atribut bagi “luas kawasan” boleh dibina dari nilai atribut “panjang” dan atribut “lebar” kawasan.
 - ii) atribut bagi “BMI” boleh dibina dari nilai atribut “berat” dan atribut “tinggi” individu.
 - iii) atribut bagi “pendapatan bersih” boleh dibina dari perjumlahan nilai semua atribut berkaitan “pendapatan” dan penolakan semua atribut berkaitan “hutang” individu.
- Pengetahuan domain sangat diperlukan untuk mentakrifkan hubungan yang betul.



- Atribut baru juga boleh dibentuk menerusi pelbagai hubungan matematik antara atribut-atribut dalam set data.
- Antara kaedah Penjelmaan Data Membentuk Atribut Baru:

i) **Penjelmaan Linear:**

- Teknik ini melibatkan penjelmaan algebra mudah seperti hasil tambah, purata, putaran, dll
- Misalkan $A = A_1, A_2, \dots, A_n$ ialah set atribut, dan misalkan $B = B_1, B_2, \dots, B_m$ ialah subset bagi set atribut lengkap A .
- Atribut baru Z boleh dibentuk menerusi penjelmaan linear berikut:

$$Z = r_1 B_1 + r_2 B_2 + \dots + r_M B_M$$



ii) Penjelmaan Data menerusi Pengekoden:

- Teknik ini digunakan untuk menjelmakan data berkategori kepadadata berangka supaya ia boleh digunakan dengan efisien dalam analisis perlombongan data.
- Sebahagian al-khawarizmi perlombongan data seperti pokok keputusan atau regresi linear memerlukan data dalam bentuk berangka.
- Maka, kaedah pengkodean merupakan pendekatan yang penting apabila berurusan dengan pembolehubah berkategori (**contoh:** warna atau label).
- Beberapa kaedah pengkodean:
 - i) Pengkodean Satu-Hot;
 - ii) Pengkodean Ordinal;
 - iii) Pengkodean Target;
 - iv) Pengkodean Kekerapan;
 - v) Dan banyak lagi.



ii) Penjelmaan Pangkat:

- Penjelmaan ini dijalankan bertujuan menggantikan nilai atribut berangka kepada nilai atribut bersifat pangkat.
- Atribut akan berubah menjadi atribut baru yang mengandungi nilai integer (pangkat, r_i) antara 1 hingga m (tertib meningkat atau menurun).
- Seterusnya, pangkat akan dijelmakan kepada data dalam bentuk skor Normal menerusi persamaan:

$$y_i = \Phi^{-1} \left(\frac{r_i - \frac{3}{8}}{m + \frac{1}{4}} \right)$$



iii) Penjelmaan Box-Cox:

- Penjelmaan Box-Cox bertujuan menjadikan atribut baru data dalam bentuk taburan hampir Normal menerusi persamaan:

$$y = \begin{cases} x^{\lambda-1} / \lambda, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$

- Nilai λ perlu dicari secara cuba jaya dengan menguji pelbagai nilai antara -3.0 to 3.0 .
- Nilai λ terbaik dipilih jika didapati taburan menghampiri normal.
- Namun, persamaan tersebut terhad kepada data bukan negatif. Data yang mempunyai nilai negatif, sedikit perubahan perlu dibuat terhadap rumus tersebut.



LAIN-LAIN KAEDAH PENJELMAAN

- Terdapat banyak lagi kaedah penjelmaan. Antara yang popular ialah:
 - i) Penjelmaan Pelicinan.
 - ii) Penjelmaan Penghampiran Polynomial.
 - iii) Penjelmaan Penghampiran Bukan Polynomial.
 - iv) Penjelmaan Wavelet.
 - v) Dan lain-lain.



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:

Penurunan Data

