PEMBERSIHAN DATA

STQD6414 PERLOMBONGAN DATA

Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- **Pembersihan data** melibatkan proses:

i) Menguruskan data-data lenyap.

ii) Menguruskan data-data yang tidak konsisten.

iii) Menguruskan data pencil (*outlier*).

- Jika pengguna tahu data adalah 'tidak bersih/kotor', mereka tidak akan mempercayai hasil perlombongan data yang dibentangkan.

- Data 'kotor' boleh menyebabkan kekeliruan/kesukaran dalam prosedur perlombongan data, juga menghasilkan keputusan yang tidak boleh dipercayai.

- Data lenyap berpunca dari pelbagai faktor, antaranya:

i) Kesilapan individu semasa pemasukan/perekodan data.

- Contoh: Faktor manusia.

ii) Berlaku kerosakan alat perekod data.

- Contoh: data meteorologi.

iii) Keengganan pelanggan untuk memberikan maklumat.

- Contoh: Data kaji selidik, Data Banci.

iv) Data tersebut memang tidak wujud.

- Contoh: Atribut no lesen memandu, sebahagian responden tiada lesen memandu.

MENGURUSKAN DATA-DATA LENYAP:

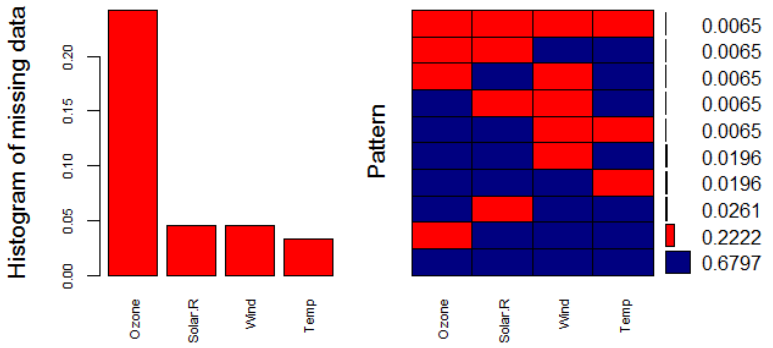
- Beberapa pendekatan boleh digunakan dalam mengurus data-data lenyap:

1. Kenalpasti data-data lenyap dan coraknya:

- Bertujuan untuk mendapatkan gambaran awal berkenaan data lenyap dalam set data.

2. Keluarkan cerapan yang mengandungi data lenyap:

- Jika data kita besar dan bilangan data lenyap adalah kecil, kaedah ini sesuai.
- Namun, tidak efektif jika data lenyap adalah agak banyak.
- Cerapan yang mengandungi data lenyap mungkin mengandungi maklumat yang penting dalam atribut lain. Data lenyap perlu dianggarkan.
- Namun, sebahagian data lenyap memang tidak sesuai dianggar dan ianya perlu dikeluarkan. Ini bergantung kepada pengetahuan domain penganalisis.

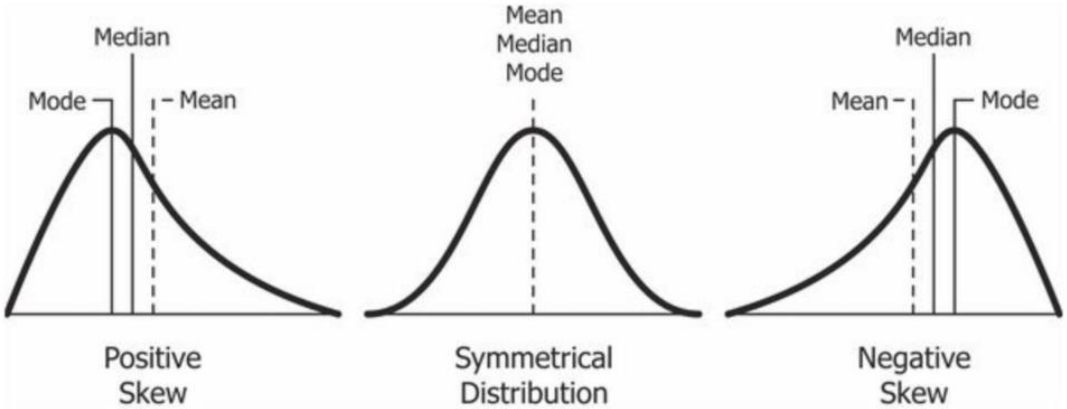


3. Lengkapkan data lenyap secara manual:

- Memerlukan pengetahuan domain (dalam bidang) berkenaan data.
- Jika data menunjukkan trend yang jelas, kaedah ini sesuai digunakan.
- Anggaran berdasarkan nilai sebelum & selepas.

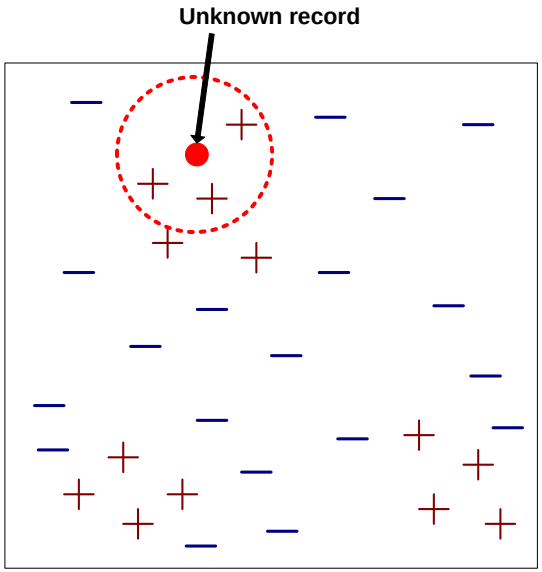
4. Gunakan sukatan memusat sebagai anggaran terhadap data lenyap (atribut yang sama):

- Untuk data taburan normal/simetri dengan nilai berangka: nilai min boleh digunakan.
- Untuk data taburan simetri dengan nilai bukan angka: nilai mod boleh digunakan.
- Untuk data taburan bersifat pincang/bukan simetri: median boleh digunakan.
- Anggaran berdasarkan data bagi keseluruhan lajur/atribut yang terlibat.



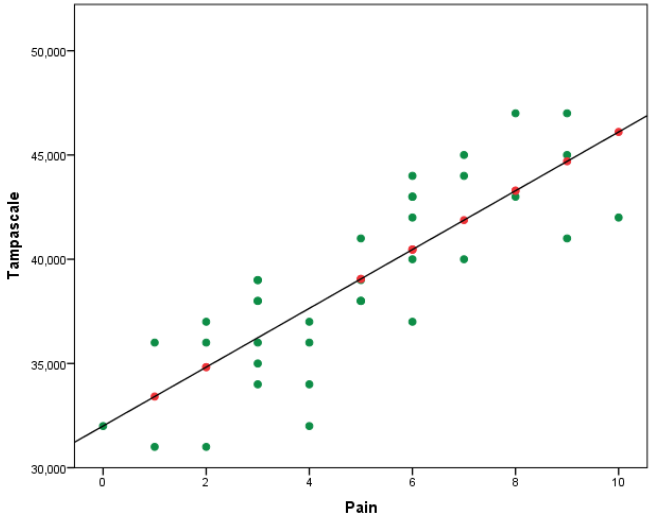
5. Gunakan maklumat k-jiran terdekat sebagai anggaran terhadap data lenyap:

- Kenalpasti k-titik cerapan yang paling hampir dengan data lenyap.
- Guna maklumat bagi jiran-jiran terdekat sebagai anggaran terhadap data lenyap.



6. Anggaran data lenyap menerusi pelbagai kaedah imputasi statistik:

- Imputasi tunggal, imputasi berganda.
- Berdasarkan kaedah regression, Pemadanan Peramal Min, Bayesian, Multivariat, dll. Kaedah-kaedah ini boleh dijalankan menggunakan pelbagai pakej R.



ANGGARAN DATA-DATA LENYAP MENERUSI R:

- Terdapat beberapa pakej dalam R yang boleh digunakan untuk berurusan dengan data lenyap.

- Antaranya:

i) mice
ii) Amelia
iii) missForest
iv) Hmisc
v) Mi
vi) missMDA

BERURUSAN DENGAN DATA LENYAP MENERUSI PAKEJ MICE:

- mice merujuk kepada "*Multivariate Imputation via Chained Equations*".
- Ianya menjalankan imputasi berganda terhadap data lenyap berdasarkan model peramal.

- Misalkan kita ada pemboleh ubah $X_1, X_2 \ldots . X_k$.
- Jika X_1 mempunyai data-data lenyap, maka p/ubah X_2, X_3 ... X_k akan digunakan dalam model peramal untuk menganggar nilai-nilai X_1.
- Seterusnya, data-data lenyap dalam X_1 akan digantikan dengan nilai-nilai anggaran yang diperoleh daripada model.
- Sama juga, jika X_2 mempunyai data-data lenyap, maka p/ubah X_1, X_3 ... X_k akan digunakan dalam model peramal untuk menganggar nilai-nilai X_2.

- Antara model peramal yang boleh digunakan dalam pakej mice:

i) PMM (*Predictive Mean Matching*): untuk p/ubah berangka.

ii) Logreg (*Logistic Regression*): untuk p/ubah dedua (dengan 2 aras)

iii) Polyreg (*Bayesian polytomous regression*): untuk p/ubah faktor (>= 2 aras)

iv) Proportional odds model: untuk p/ubah bertertib (>= 2 aras)

v) Dan lain-lain.

##rujuk Package 'mice' - R Project untuk mendalami pelbagai kaedah dalam pakej mice.

BERURUSAN DENGAN DATA LENYAP MENERUSI PAKEJ AMELIA:

- Pakej ini dinamakan sempena nama penulis dan perintis American Aviation, iaitu Amelia Earhart.
- Amelia merupakan juruterbang wanita yang telah membuat cubaan mencipta rekod sebagai wanita pertama terbang di seluruh dunia secara solo di Lautan Atlantik pada tahun 1932.
- Namun, beliau didapati hilang.
- Tidak ada bukti sama ada dia masih hidup ataupun mati.

- Pakej ini menggunakan kaedah pensampelan butstrap (*bootstrap*) dan al-khwarizmi Pemaksimuman-Jangkaan (*Expectation-Maximization*) untuk meramal data-data lenyap dalam set data.
- Langkah 1: Pensampelan butstrap

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

- Langkah 2: Imputasi berdasarkan al-khwarizmi Pemaksimuman-Jangkaan

#Rujuk Package 'Amelia' - R

untuk mendalami kaedah-kaedah dalam pakej Amelia.

#imputation	Row num	Var 1	Var 2	Var 3
1	1	9	3	4
1	2	4	3	2
1	3	4	2	5
2	1	9	3	4
2	2	4	4	2
2	3	2	2	3
3	1	9	3	4
3	2	4	2	2
3	3	2	2	4

BERURUSAN DENGAN DATA TIDAK KONSISTEN:

- Data tidak konsisten perlu diperbaiki ataupun disingkirkan agar ianya tidak memberi kesan buruk kepada proses perlombongan data.

1. Mengenalpasti data tidak konsisten:

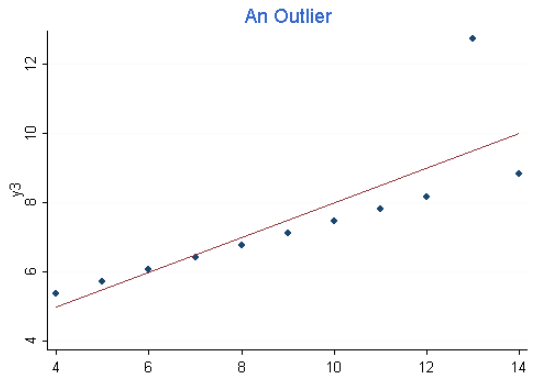
- Ini perlu dijalankan apabila data digabungkan dari beberapa sumber/file.
- Contoh: Nama seseorang mungkin dinyatakan dengan penuh dalam satu sumber data, sedangkan sumber lain hanya mengandungi awal dan nama akhir sahaja.

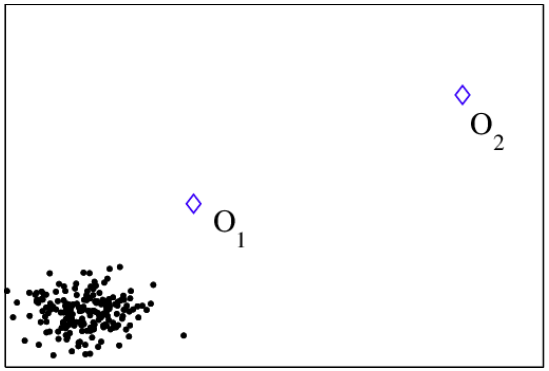
2. Pengetahuan Domain:

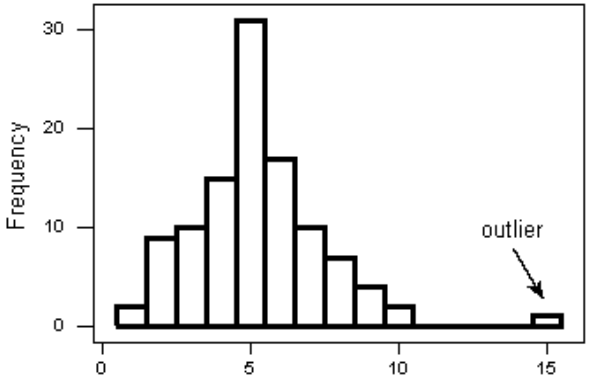
- Pengetahuan yang luas domain (bidang) yang dikaji sangat membantu dalam membetulkan sata-data yang tidak konsisten.
- Contoh: jika atribut daerah ialah "Pasir Mas" maka Negeri tidak boleh "Selangor".

BERURUSAN DENGAN DATA PENCIL:

- Data pencil ialah cerapan yang terletak pada jarak yang agak jauh dari kebanyakan data.





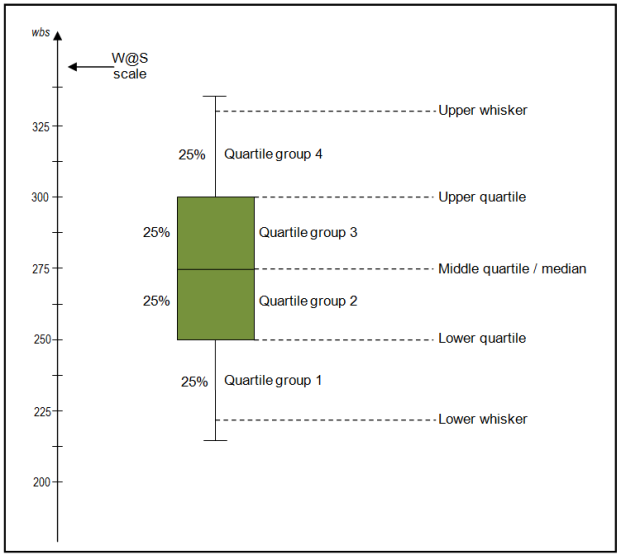


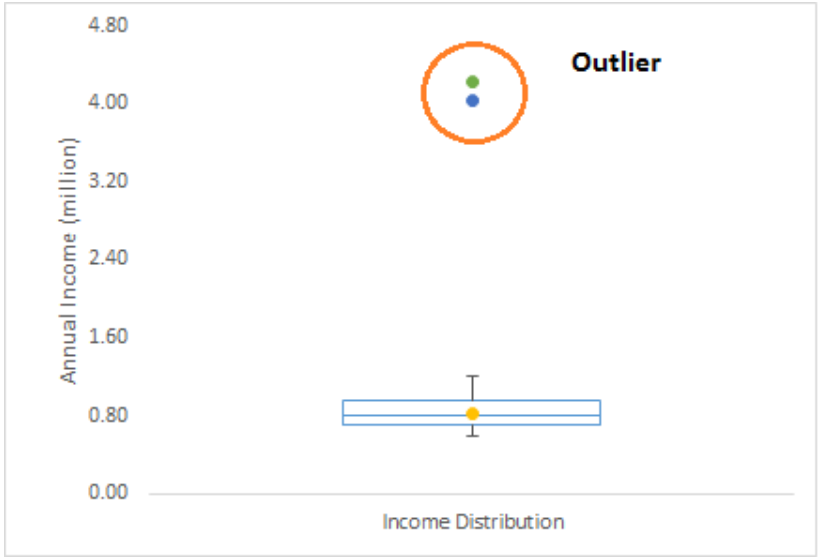
- Data pencil boleh berlaku disebabkan oleh ralat pemasukan data, kerosakan alat atau kesalahan mengambil cerapan.
- Namun, jika tidak berlaku sebarang kesalahan dalam perekodan data, data pencil merupakan maklumat yang sangat penting (kes jarang berlaku).
- Penipuan kad kredit, banjir besar, pendapatan jutawan, dll.
- Data pencil boleh mempengaruhi ketepatan perlombongan data jika ianya tidak dikenalpasti dan dikendalikannya dengan sewajarnya.

PENGESANAN DATA PENCIL:

i) Pendekatan Univariat (satu p/ubah):

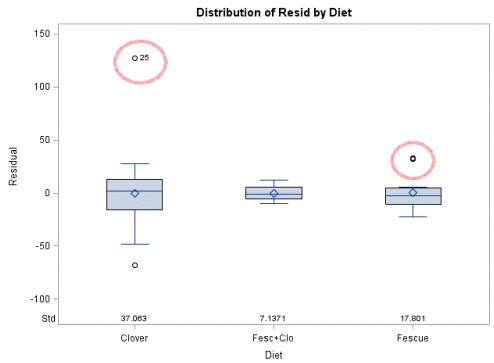
- Gunakan Plot Kotak (boxplot).
- Bagi p/ubah selanjar univariat, data pencil ialah cerapan-cerapan yang berada diluar 1.5 x *IQR* (*Inter Quartile Range*).
- IQR ialah perbezaan antara 75th dan 25th quartil.

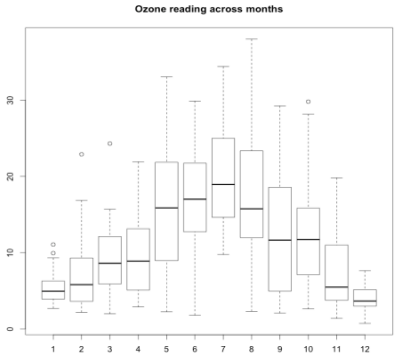


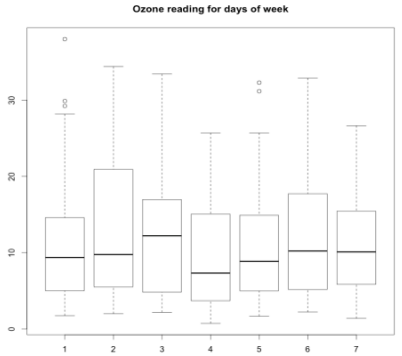


ii) Pendekatan Bivariat (2 p/ubah (X dan Y)):

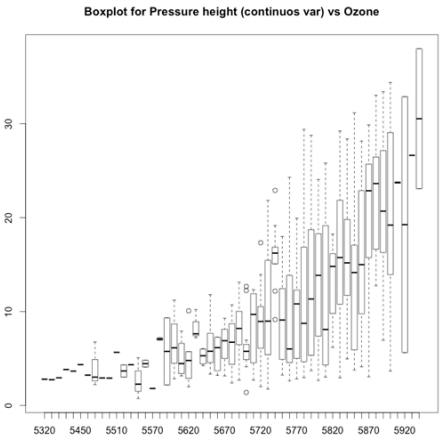
- Jika p/ubah X ialah berkategori (aras) dan Y ialah selanjar, gunakan boxplot.

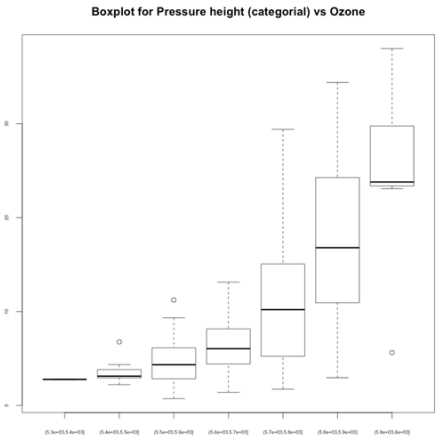




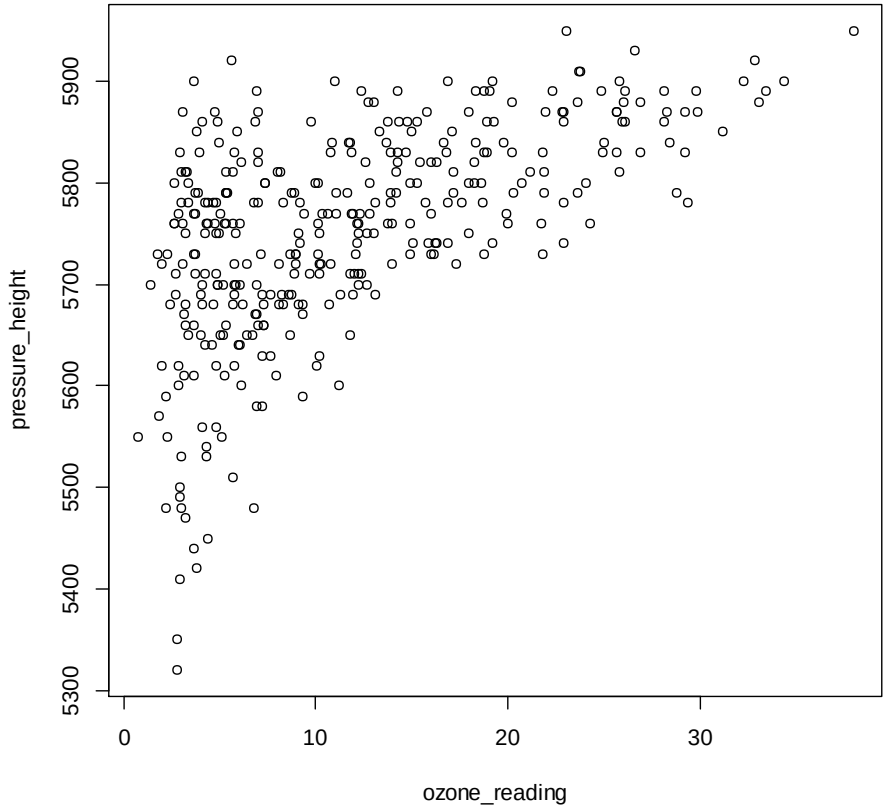


- Jika X ialah selanjar, Y ialah selanjar, masih boleh gunakan plot kotak.
- Namun, cuba jelmakan X kepada bentuk berkategori jika variasinya terlalu besar.





- Jika *X* adalah selanjar, *Y* juga selanjar, pendekatan lain ialah gunakan plot serakan (*scatter plot*).



- Apabila data pencil telah dikenalpasti, tugas Saintis Data adalah menyiasat sama ada ianya adalah data yang salah ataupun maklumat yang jarang berlaku (maklumat sangat penting).

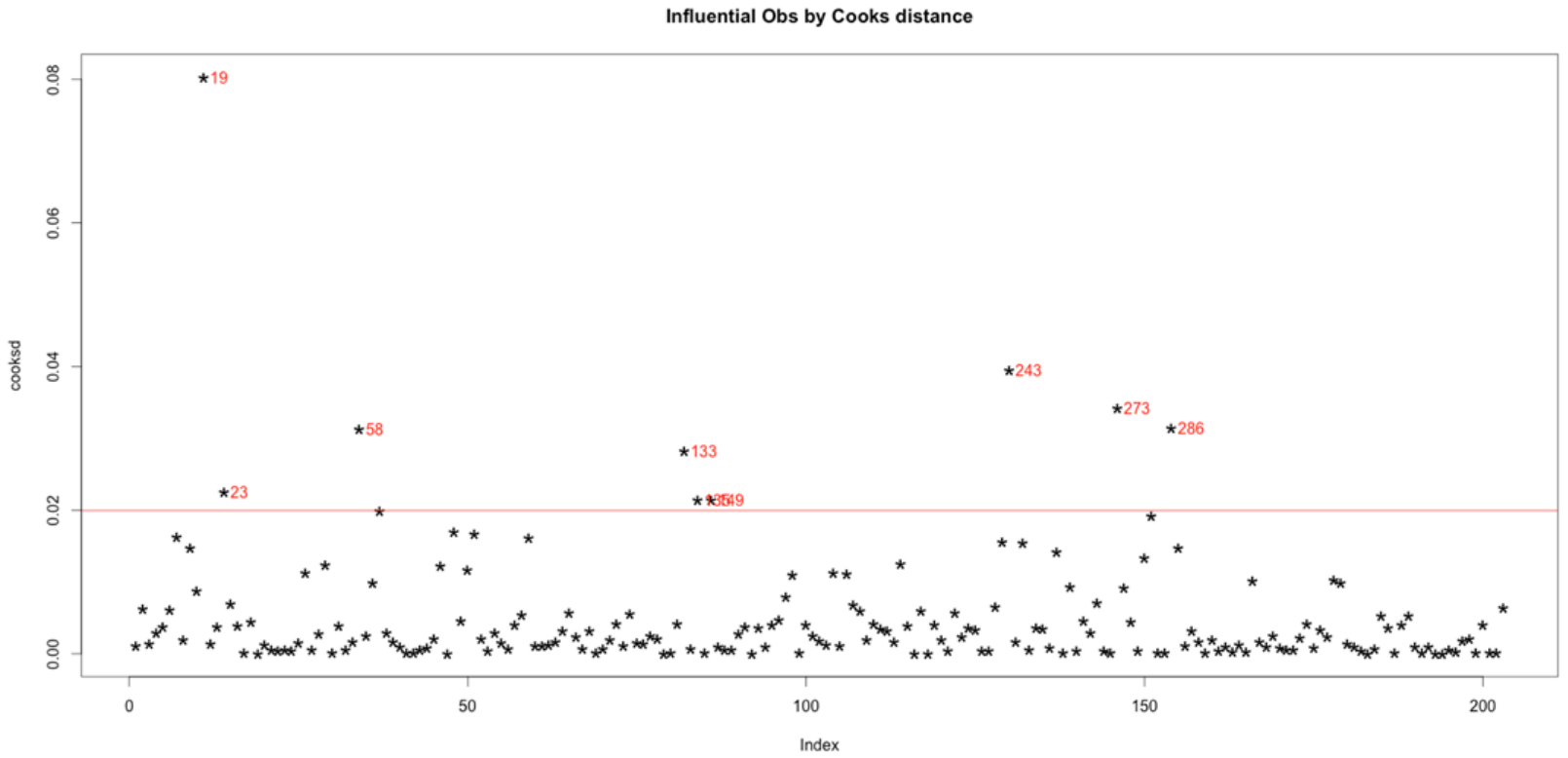
ii) Pendekatan Multivariat (kes terselia):

- Bagi data yang terdiri dari beberapa p/ubah, ianya perlu dilihat secara menyeluruh untuk menentukan data pencil.
- Bagi kes data terselia, data pencil boleh dikenalpasti menerusi pendekatan jarak-Cook.
- Jarak-Cook dihitung menerusi formula berikut:

$$D_i = \frac{\sum_{j=i}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p \times MSE}$$

- Model regresi linear berganda disuaikan bagi data p/ubah sambutan Y terhadap p/ubah $X_1, X_3 \ldots X_k$
- $\hat{Y}_j$ ialah nilai tersuai ke-j apabila mengambilkira nilai semua cerapan.
- $\hat{Y}_{j(i)}$ ialah nilai tersuai ke-j apabila cerapan i tidak diambilkira.
- *MSE* ialah min ralat kuasa dua.
- p ialah bilangan pekali model regresi.

- Secara umumnya, cerapan-cerapan yang mempunyai nilai jarak-Cook yang 4 kali ganda lebih besar daripada min jarak-Cook akan diklasifikasikan sebagai cerapan yang berpengaruh.

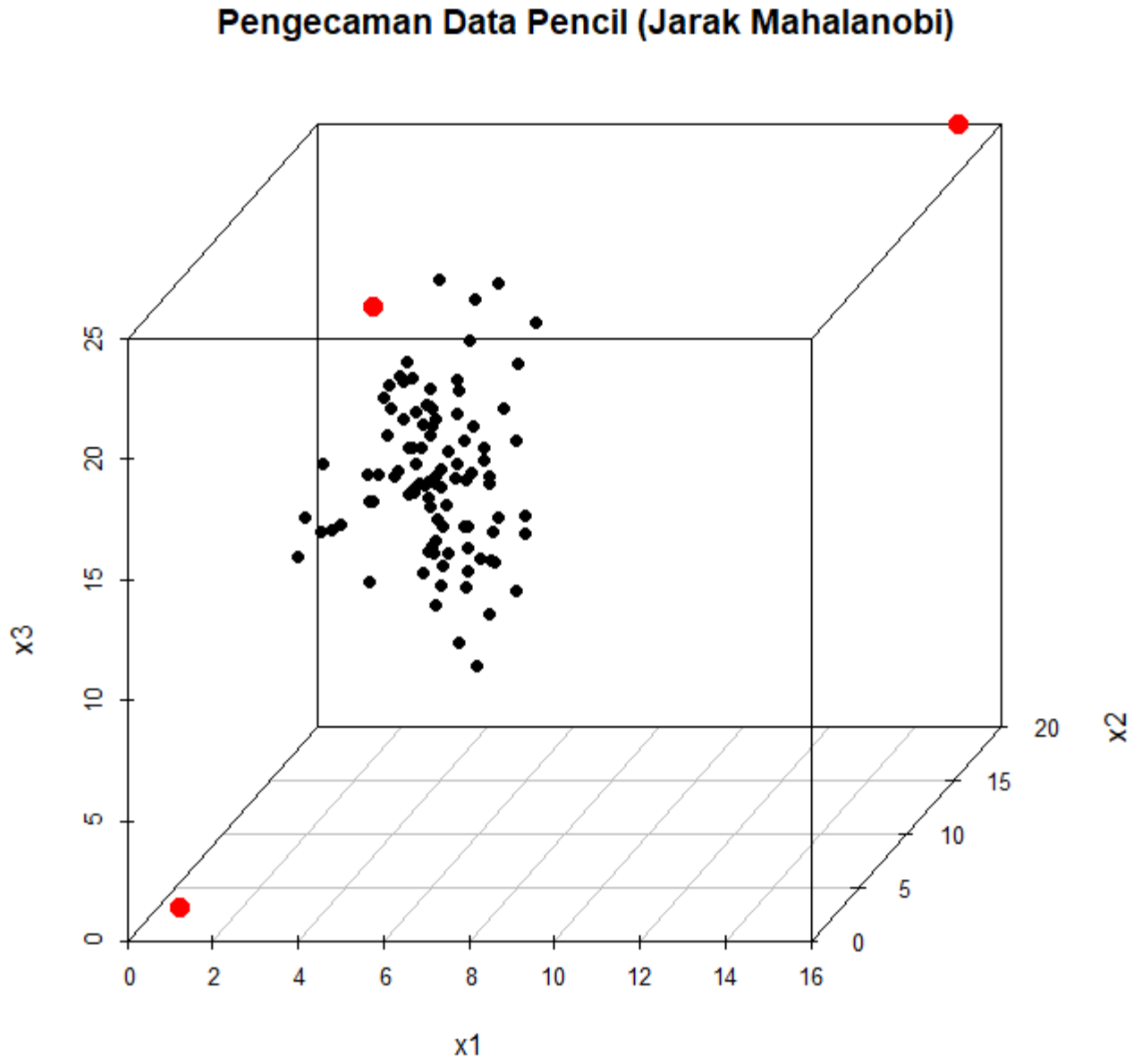


iii) Pendekatan Multivariat (kes tak terselia):

- Bagi kes data tak terselia, kaedah berasaskan jarak seperti jarak Mahalanobi seringkali digunakan.
- Jarak Mahalanobi mengukur sejauh mana sesuatu cerapan berbeza daripada min data dalam ruang multivariat menerusi formula berikut:

$$D_M = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- Iaitu $\mathbf{x}$ ialah data multivariat $\left[x_1, x_2, \ldots, x_p\right]^T$.
- $\boldsymbol{\mu}$ ialah min vektor.
- $\mathbf{S}$ ialah kovarians matriks.
- Jarak Mahalanobis yang lebih besar menunjukkan bahawa satu titik data adalah lebih jauh daripada min taburan data.
- Data pencil boleh dikenal pasti dengan menetapkan ambang berdasarkan taburan khi kuasa dua.
- Jika D_M^2 melebihi nilai kritikal daripada taburan khi kuasa dua dengan k darjah kebebasan (k ialah bilangan pembolehubah), titik boleh dianggap sebagai data pencil.


Pengecaman Data Pencil (Jarak Mahalanobi)
x3
0
5
10
15
20
25
x1
0
2
4
6
8
10
12
14
16
x2
0
5
10
15
20

MERAWAT DATA PENCIL:

- Selepas data-data pencil dikenalpasti, ianya perlu dirawat menerusi:

i) Jika data pencil ialah data ralat (error):

- Boleh dibuang (jadikan ianya data lenyap).
- Seterusnya, nilai data tersebut boleh dianggar semula menerusi kaedah imputasi.

ii) Jika data pencil ialah data sebenar (data jarang berlaku/mempunyai maklumat penting):

- Ianya perlu dikekalkan dalam data.
- Kaedah statistik yang khusus mungkin perlu digunakan untuk menganalisis data jenis ini (Statistik Teguh/*Robust Statistics*).
- Sebahagian kaedah perlombongan data boleh digunakan terhadap data yang mengandungi data pencil (pengkelompokan/*clustering*)

TUGASAN:

Diberi set data custdata5.csv:

i) Terangkan berkenaan latar belakang data dan statistik ringkas data.

ii) Terangkan berkenaan corak dan sifat-sifat data lenyap.

iii) Gunakan teknik-teknik yang sesuai untuk menganggar data lenyap bagi setiap atribut.

iv) Berikan alasan yang wajar mengapa teknik-teknik tersebut digunakan untuk anggaran data lenyap.

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:

Penjelmaan Data dan Pendiskretan