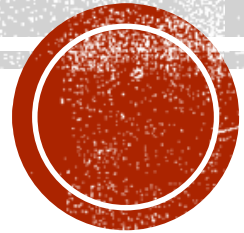
Teknik-teknik Asas Jelajahan Data Menggunakan Pengaturcaraan R

STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik,

Universiti Kebangsaan Malaysia

MENGINDEKS STRUKTUR DATA:

- Bertujuan untuk mengekstrak beberapa data daripada keseluruhan struktur data.
- Elemen-elemen daripada vektor, matriks atau bingkai data boleh diekstrak menggunakan pengindeksan berangka ataupun vektor Boolean.

- Antara teknik-teknik pengindeksan yang penting:

1. Pengindeksan mengikut nombor dan nama.
2. Pengindeksan melalui vektor Boolean.
3. Pengindeksan negatif.
4. Pengekstrakan bersyarat

SUBSET BAGI STRUKTUR DATA:

- Bertujuan untuk mendapatkan subset elemen dalam vektor, matriks ataupun dalam bingkai data.
- Objektifnya adalah sama seperti untuk pengindeksan dalam data. Namun, teknik ini mempunyai sedikit perbezaan dari segi kaedah pengekstrakan data.

- Antara teknik-teknik subset data yang penting:

1. Subset mengikut nombor dan nama.
2. Subset bagi baris dan lajur tertentu.
3. Subset berdasarkan Operator Logik (AND).
4. Subset berdasarkan Operator Logik (OR).
5. Subset data dengan syarat penjelmaan.

MAKLUMAT ASAS BERKAITAN DATA:

- Bertujuan untuk mendapatkan maklumat-maklumat asas berkaitan karakter dan sifat data.

- Antara teknik-teknik yang boleh digunakan:

1. Menyenaraikan nama pembolehubah-pembolehubah dalam set data.
2. Dapatkan maklumat tentang pembolehubah.
3. Dapatkan maklumat tentang saiz dan struktur data.
4. Lihat n baris pertama bagi set data.
5. Dapatkan jumlah data lenyap.

PEMBUNDARAN NOMBOR:

- Pembundaran bermaksud menjadikan nombor dalam bentuk yang lebih mudah tetapi mengekalkan nilainya hampir sama dengan nombor tersebut.
- Hasilnya kurang tepat, tetapi nombor tersebut akan lebih mudah digunakan untuk analisis.

- Antara teknik-teknik pembundaran data yang penting:

1. Membundarkan kepada integer terdekat.
2. Membundarkan ke sempadan atas integer.
3. Membundarkan ke sempadan bawah integer.
4. Membundarkan kepada bilangan titik perpuluhan tertentu.

PENGISIHAN:

- Pengisihan ialah proses menyusun data ke dalam susunan yang bermakna supaya kita boleh menganalisis data dengan lebih berkesan.
- Pengisihan melibatkan proses menyusun elemen-elemen dalam vektor, matriks atau bingkai data dalam tertib menaik atau menurun.

- Antara teknik-teknik pengisihan data yang penting:

1. Pengisihan mengikut tertib menaik.
2. Pengisihan mengikut tertib menurun.
3. Pengisihan dalam bingkai data sepadan dengan beberapa ciri tertentu.

TERTIB PERAWAKAN:

- Bertujuan untuk merawakkan tertib dalam struktur data.
- Kaedah ini banyak digunakan dalam teknik persampelan semula, pemilihan data ujian & data latihan, kajian simulasi, dan lain-lain.

- Antara teknik-teknik tertib perawakan ialah:

1. Perawakan tertib vektor.
2. Perawakan tertib bingkai data.

ARAS DALAM PEMBOLEH UBAH FAKTOR:

- Bertujuan untuk mengatur aras bagi suatu pembolehubah faktor dalam set data.

- Antara teknik-teknik manipulasi aras bagi pemboleh ubah faktor:

1. Membina pembolehubah faktor baharu.
2. Mengtakrif pembolehubah faktor bertertib.
3. Namakan semula aras faktor.
4. Menambah dan menurunkan aras dalam pembolehubah faktor.

JUJUKAN DALAM BLOK:

- Bertujuan untuk mendapatkan ringkasan statistik data dalam beberapa saiz blok tertentu.
- Contoh: Kita ingin mendapatkan purata bagi 4 nombor pertama, 4 nombor seterusnya, dan seterusnya.

- Antara teknik-teknik jujukan dalam blok:

1. Mentakrifkan saiz blok.
2. Menggantikan nilai yang terkurang dengan NA.
3. Membina matriks blok baris.
4. Menghitung ukuran statistik lajur.
5. Jalankan Pengekodan Panjang (*Run Length Encoding*).

PERNYATAAN IF ELSE DAN NESTED IF ELSE:

- Pernyataan ifelse( ) memberikan jalan untuk mentakrifkan pernyataan bersyarat dalam analisis data.
- Pernyataan Nested ifelse( ) pula merujuk kepada pernyataan berganda bagi pernyataan ifelse.

- Antara penggunaan teknik-teknik asas yang berkaitan:

1. Pernyataan mudah ifelse( ).
2. Fungsi ifelse( ) terhadap pemboleh ubah kualitatif (aksara).
3. Pernyataan Nested ifelse( ).

AGREGAT DATA BERDASARKAN KUMPULAN:

- Pengagregatan data ialah proses menggabungkan data dan mempersembahkannya dalam format deskriptif.
- Untuk melakukan pengagregatan, kita perlu menentukan tiga perkara dalam kod pengaturcaraan:

i) Data yang ingin kita agregatkan.

ii) Pembolehubah untuk dikumpulkan dalam kumpulan tertentu.

iii) Penghitungan yang akan digunakan kepada kumpulan tertentu.

- Antara teknik-teknik pengagregatan data:

1. Pengagregatan satu pemboleh ubah & kumpulan berdasarkan satu pemboleh ubah.
2. Pengagregatan satu pemboleh ubah & kumpulan berdasarkan pemboleh ubah berganda.
3. Pengagregatan pemboleh ubah berganda & kumpulan berdasarkan satu pemboleh ubah.
4. Pengagregatan pemboleh ubah berganda & kumpulan berdasarkan pemboleh ubah berganda.

PENGGELUNGAN (LOOPING) DALAM R:

- Gelung *(loop)* digunakan dalam pengaturcaraan untuk mengulang blok kod tertentu.

- Antara teknik-teknik penggelungan yang penting dalam R:

1. Fungsi Apply.
2. Fungsi Lapply.
3. Fungsi Sapply.
4. For Loop
5. While Loop

- Beberapa konsep penggelungan dalam R:

1. Break Keyword
2. Next Keyword

JELAJAHAN DATA: PAKEJ DPLYR

- "dplyr" merupakan pakej yang sangat baik untuk proses manipulasi, pembersihan dan memperilhalkan data berstruktur atau tidak berstruktur.
- "dplyr" terdiri daripada pelbagai fungsi yang melakukan operasi manipulasi data seperti menggunakan penapis *(filter)*, memilih lajur tertentu, mengisih data, menambah atau memadam lajur, mengagregatkan data, dan lain-lain.

- Beberapa fungsi penting dalam pakej dplyr:

dplyr Function	Description	Equivalent SQL
select()	Selecting columns (variables)	SELECT
filter()	Filter (subset) rows.	WHERE
group_by()	Group the data	GROUP BY
summarise()	Summarise (or aggregate) data	-
arrange()	Sort the data	ORDER BY
join()	Joining data frames (tables)	JOIN
mutate()	Creating New Variables	COLUMN ALIAS

JELAJAHAN DATA: PAKEJ DPLYR

- Beberapa teknik penting untuk penerokaan dan manipulasi data menggunakan pakej dplyr adalah seperti berikut:

1. Pemilihan rawak N baris.
2. Pemilihan rawak pecahan/peratusan baris.
3. Menyusun semula pembolehubah.
4. Menamakan semula pembolehubah.
5. Menapis baris.
6. Pemililihan kriteria berganda.
7. Syarat 'AND' dalam pemilihan kriteria.
8. Syarat 'OR' dalam pemilihan kriteria.
9. Syarat NOT.
10. Syarat CONTAINS.

JELAJAHAN DATA: PAKEJ DPLYR

- Beberapa teknik penting untuk penerokaan dan manipulasi data menggunakan pakej dplyr adalah seperti berikut:

11. Memperihalkan pemboleh ubah terpilih.
12. Memperihalkan pemboleh ubah berganda.
13. Memperihalkan data berdasarkan fungsi tersuai *(Custom functions)*.
14. Memperihalkan semua pemboleh ubah berangka.
15. Menyusun data menerusi pemboleh ubah berganda.
16. Operator Pipe %>%.
17. Memperihalkan data menerusi pemboleh ubah berkategori.
18. Penapisan data dalam pemboleh ubah berkategori.

JELAJAHAN DATA: PAKEJ DPLYR

- Beberapa teknik penting untuk penerokaan dan manipulasi data menggunakan pakej dplyr adalah seperti berikut:

20. Memperihalkan, mengkumpulkan dan menyusun data secara bersama.
21. Memilih kumpulan yang menjana nilai tertinggi antara beberapa pembolehubah tertentu.
22. Menghitung nilai kumulatif bagi pemboleh ubah.
23. Operasi ROW WISE.
24. Menghitung nilai-nilai persentil.
25. Dan banyak lagi

RUJUKAN:

- Chang, W. (2020). *R graphics Cookbook. 2nd edition.* O'Reilly Media.
- Davies, T.M. (2016). *The Book of R: A First Course in Programming and Statistics 1st Edition*. No Starch Press.
- Grolemund, G., Wickham, H. (2014). *Hands-On Programming with R: Write Your Own Functions and Simulations*. O'Reilly Media.
- Jones, E., Harden, S., Crawley, M. J. (2022). *The R Book, 3rd Edition*. Wiley.
- Long, J.D., Teetor, P. (2019). *R Cookbook, 2nd Edition*. O'Reilly Media.
- Xie, Y. (2020). *R Markdown Cookbook, 1st Edition*. Routledge
- Wickham, G., Grolemund, G. (2023). *R for Data Science*. *2nd edition.* O'Reilly Media.

TOPIK SETERUSNYA:

Pra-pemprosesan Data