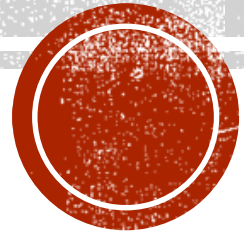
PENGENALAN KEPADA PROSES KDD DAN PERLOMBONGAN DATA

STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik,

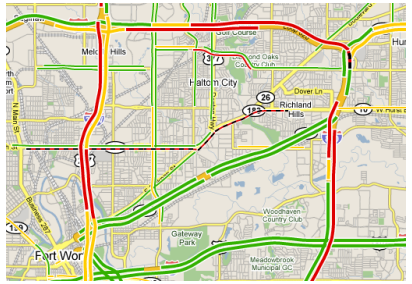
Universiti Kebangsaan Malaysia

PENGENALAN:

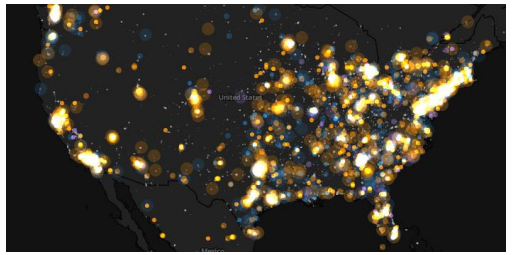
- Era kini, data terdapat di mana-mana.
- Data mudah diperoleh dengan kuantiti yang besar selari dengan perkembangan teknologi revolusi industri ke-4.
- Data juga datang dari pelbagai sumber yang berbeza.
- Data yang besar ini sukar untuk difahami.
- Jika tidak dianalisis, data adalah kurang/tidak bermakna.
- Analisis perlu dijalankan untuk mencungkil maklumat yang berguna dan menjawab sebarang persoalan.
- Tugas ahli statistik, saintis data, data analitik, jurutera data dan lain-lain.
- **Artikel:** Which Jobs Earn The Highest Salaries In Malaysia?

Keselamatan Cyber

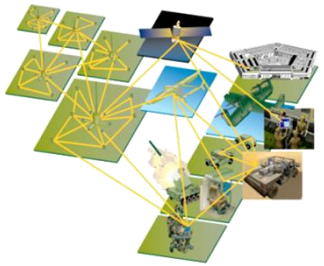
E-Dagang



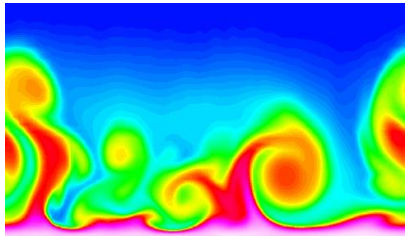
Corak Trafik



Jaringan Sosial: (Facebook, Twitter, dll)



Jaringan Sensor



Simulasi Komputer

PERLOMBONGAN DATA: PERFEKTIF INDUSTRI

- Di industri, data yang besar ini disimpan dalam gudang data, awan data, tasik data dll (*data warehoused, cloud data, lake data, etc*).

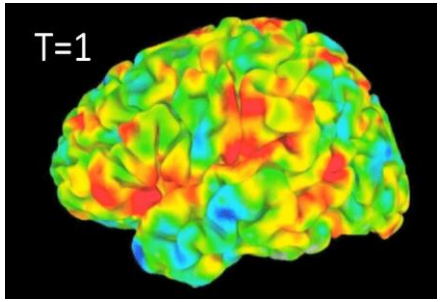
- Contoh Data:
  - Data laman sesawang (Web data):
    - Google mempunyai Peta Bytes bagi web data.
    - Facebook mempunyai berbillion pelanggan aktif.
  - Data pembelian (runcit, e-dagang):
    - Jutaan transaksi pembelian di Pasaraya setiap hari.
    - Jutaan pelanggan membuat pembelian online di Amazon.com dan Shoope setiap hari.
  - Data Bank dan transaksi Kad Kredit.
  - Data kesihatan rakyat Malaysia (contoh: MySejahtera)

- Persaingan antara industri lebih mencabar.
  - Perlu jalankan servis dan pengurusan yang lebih baik.
  - Memahami sentimen kehendak pasaran.
  - Menghasilkan produk yang lebih berkualiti dan murah.
  - Semua ini memerlukan maklumat dari data & analisis statistik.

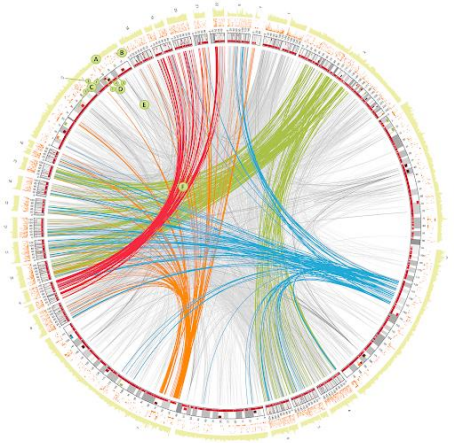
- Komputer hari ini lebih canggih dan murah untuk mengadaptasi senario era data masa kini.

PERLOMBONGAN DATA: PERFEKTIF SAINTIFIK

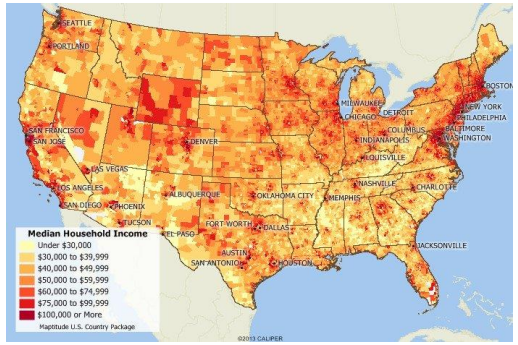
- Dari sudut saintifik juga, data dicerap dan dikumpul dengan sangat pantas.
- Contoh Data:
  - Data Satelit:
    - NASA EOSDIS mencerap lebih dari peta-bit data berkaitan sains bumi setiap tahun.
  - Teleskop astronomi:
    - Data cerapan langit (*Sky survey data*).
  - Data biologi intensiti tinggi:
    - Data Jujukan DNA.
    - Data Ekspresi Gene.
    - Data MRI.
    - Data Genom.
  - Data Demografi:
    - Data Pendapatan.
    - Data Profil Penduduk.



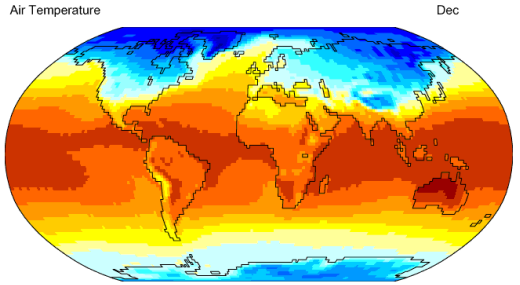
Data MRI dari otak



Data Genom



Data Demografi



Data suhu permukaan bumi

PERLOMBONGAN DATA:

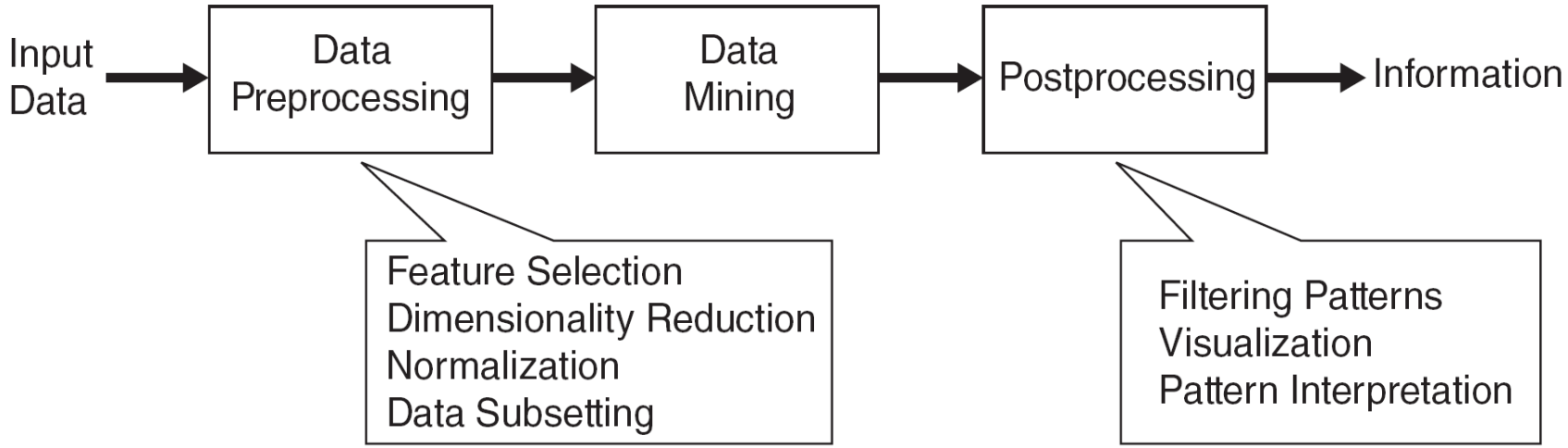
- Perlombongan Data ialah kaedah-kaedah analisis yang digunakan dalam proses "knowledge discovery in databases" atau KDD.

- Secara khusus, ianya bertujuan:

i) Mencungkil maklumat & menjawab sebarang persoalan.

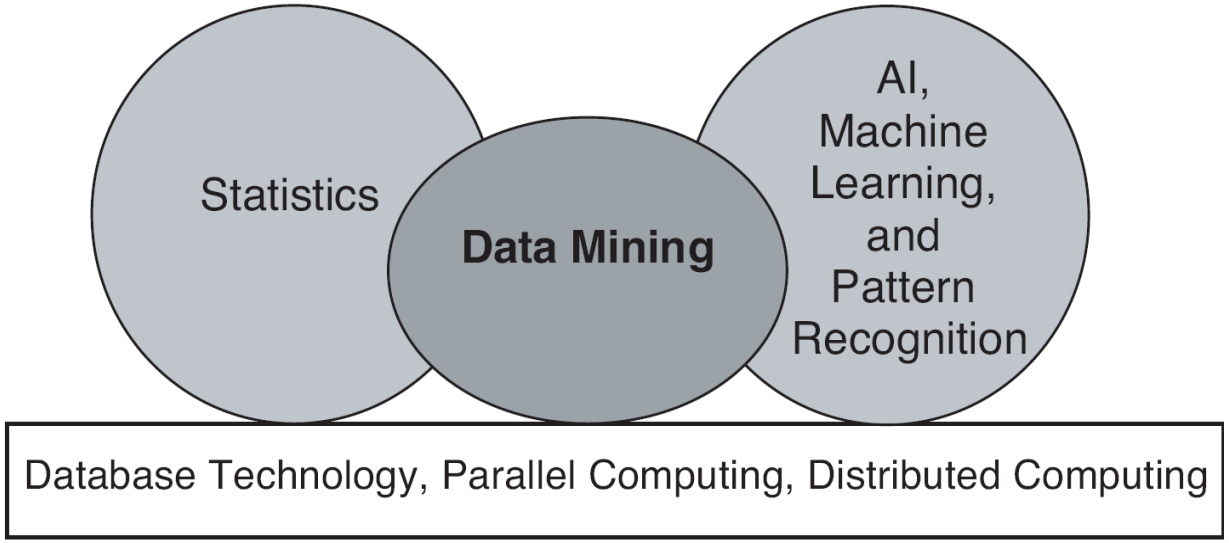
ii) Memodelkan data & meramal nilai p/ubah masa hadapan.

iii) Mengenal pasti corak & trend dalam data.



PERLOMBONGAN DATA:

- Sebahagian besar ilmu dan teknik perlombongan data terbit dari bidang ilmu Statistik, pembelajaran mesin (*machine learning*), kecerdasan buatan (*AI, Artificial Intelligence*), pengecaman corak (*pattern recognition*) dan sistem pangkalan data (*database systems*).



CONTOH APLIKASI:

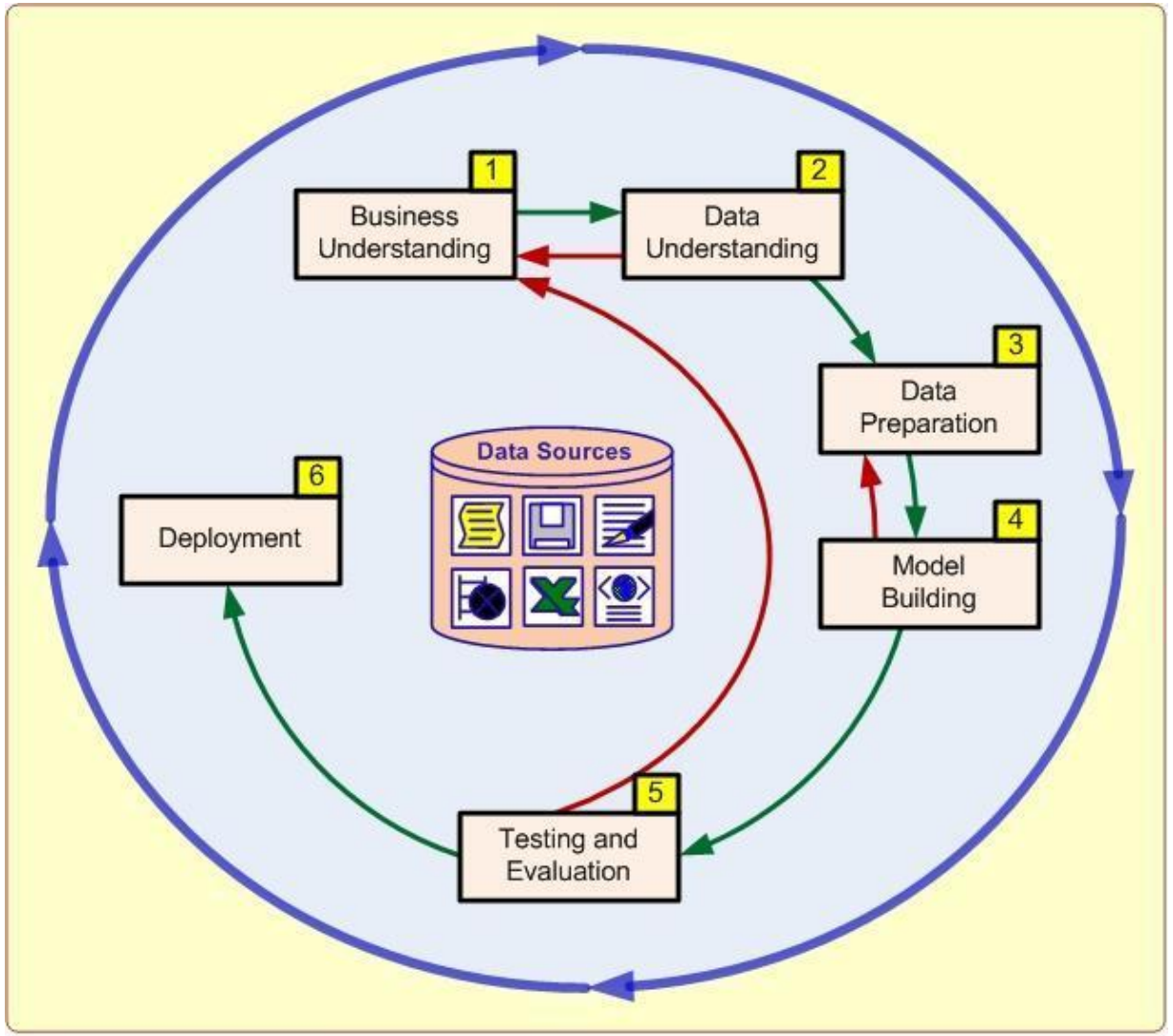
- Perbankan (kelulusan pinjaman/kad kredit):
  - Meramal kategori pelanggan yang 'bagus' berdasarkan rekod pelanggan-pelangan lama.
- Pengurusan hubungan pelanggan:
  - Mengenalpasti pelanggan-pelanggan yang berpotensi untuk berpindah keluar dari servis (contoh: risiko peralihan pelanggan celcom kepada maxis).
- Pemasaran sasaran:
  - Mengenal pasti sasaran promosi kepada kumpulan tertentu.
- Pengesanan penipuan:
  - Telekomunikasi, transaksi kewangan.
- Pembuatan dan pengeluaran:
  - Menyesuaikan sistem secara automatik apabila parameter proses berubah.
- Perubatan:
  - Menganalisis sejarah penyakit pesakit, cari hubungan antara penyakit
  - Mengecam sifat penyakit dan keberkesanan rawatan.

PROSES KDD:

- Formulasi Masalah.
- Pengumpulan & Pemahaman Data.
- Pra-Pemposesan Data:
  - Pembersihan Data.
  - Integrasi Data.
  - Penjelmaan Data.
  - Penurunan Data.
- Memilih kaedah/model statistik yang sesuai dan jalankan analisis perlombongan data.
- Penilaian Hasil dan Pengvisualan.

Proses KDD merupakan proses yang bersifat berulang

PROSES KDD:

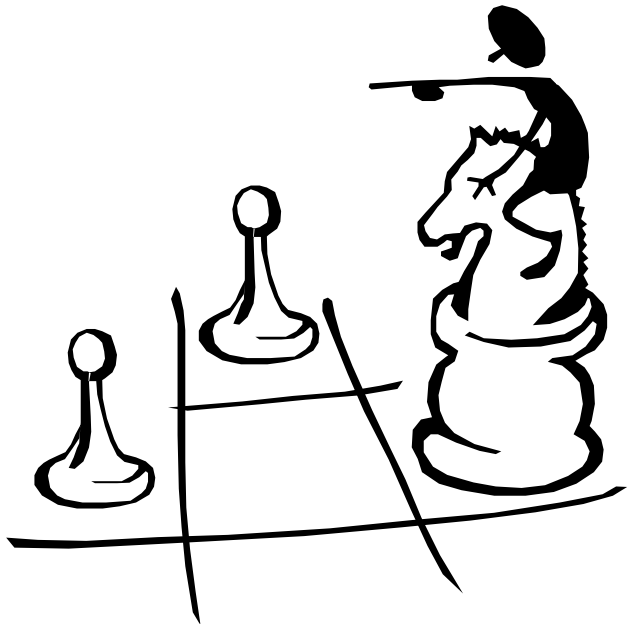


Source: Adapted from CRISP-DM.org.

GUDANG DATA & PERLOMBONGAN DATA:

- Gudang Data ialah kemampuan institusi untuk menyimpan maklumat dan data. (*memory*)

- Perlombongan Data pula membantu institusi membuat keputusan berdasarkan maklumat dari data. (*intelligence*)



KAEDAH UMUM PERLOMBONGAN DATA:

- **Kaedah Perihalan:**
  - Mengenalpasti corak yang boleh menerangkan data.

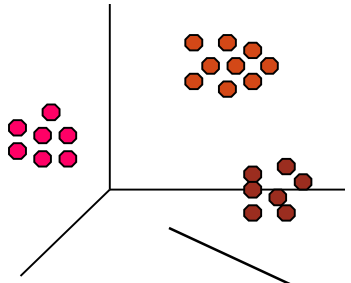
- **Kaedah Peramalan:**
  - Gunakan beberapa pemboleh ubah untuk meramal nilai masa hadapan bagi pemboleh ubah lain.

PERISIAN STATISTIK:

- Pengaturcaraan R:

- Phyton:

Contoh teknik-teknik Perlombongan Data:

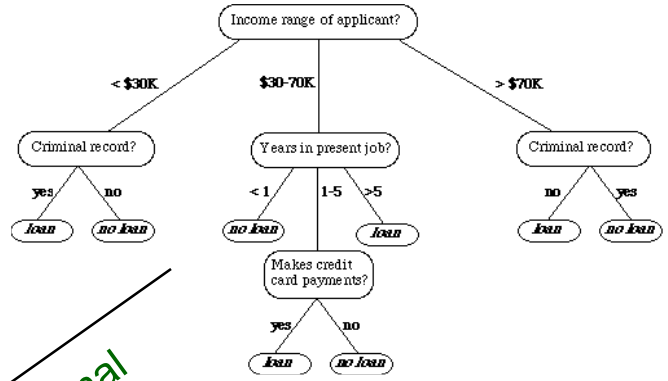


Pengkelompokan

Data

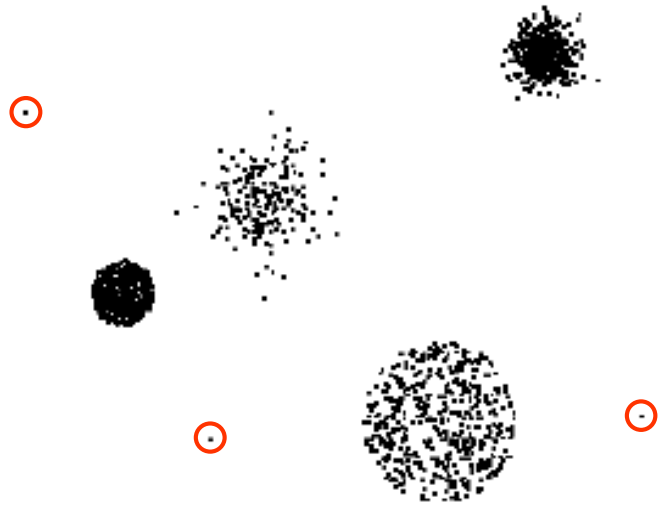
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Model Peramal



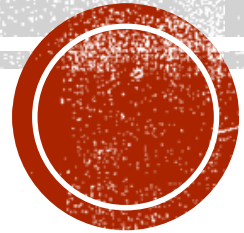
Aturan Sekutuan

Pengecaman Anomali



Asas Pengaturcaraan R

STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik,

Universiti Kebangsaan Malaysia

MENGAPA PERLU MENGGUNAKAN PERISIAN R ?

- Ianya adalah percuma. Boleh dimuat turun dari internet.
- Antara perisian statistik yang paling baik buat masa ini.
- Mempunyai akses terhadap lebih daripada 10000 pakej untuk menjalankan pelbagai analisis statistik dan perlombongan data.
- Terdapat banyak forum/bengkel/kursus untuk membantu anda mendalami perisian R.
- Salah satu kemahiran yang sangat dicari oleh syarikat data analitik dan sains data.

BAGAIMANA UNTUK MEMASANG R (*INSTALL R*):

1. Lawati laman sesawang https://www.r-project.org/

- Atau gunakan carian google untuk "R programing download"

2. Muat turun Pemasang R (*R installer*)

3. Pasang perisian R dalam komputer anda.

ANTARA MUKA PERISIAN R (*R INTERFACE*):

1. Konsol R:

- Bahagian ini menunjukkan output bagi kod yang dijalankan.
- Anda juga boleh menulis kod secara terus di konsol ini.

2. Skrip R:

- Bahagian untuk menulis kod-kod pengaturcaraan.
- Untuk menjalankan kod-kod tersebut, hanya perlu sorotkan (*highlight*) kod-kod tersebut dan jalankan ia dibahagian konsol R.

3. Output Grafik:

- Bahagian ini menunjukkan graf-graf atau plot-plot yang dibina semasa menjalankan analisis data.

ASAS PENGHITUNGAN DALAM R:

- Sila taip perhitungan matematik tersebut dalam konsol R anda:

- 2 + 3
  > 5

- 6 / 3
  > 2

- (3*8)/(2*3)
  > 4

- log(12)
  > 1.07

- sqrt (121)
  > 11

- Penggunaan pemboleh ubah adalah sangat penting dalam penulisan kod.

- Misalkan pemboleh ubah x mewakili hasil tambah 7 dan 8. Ini boleh ditulis sebagai:

- ```
  x <- 8 + 7
  x
  > 15
  ```

- Bagi beberapa p/ubah:

- y<- 10/2
- z<- x*y
- z<- 75


  ```

JENIS-JENIS PEMBOLEH UBAH DALAM R:

1. **Pemboleh ubah Kualitatif:** P/ubah yang mengambil nilai bukan angka.

1.1 Pemboleh ubah Nominal:

- Nominal bermaksud "berkaitan nama."
- Nilai p/ubah nominal ialah simbol atau nama sesuatu kategori.

Contoh:

- Misalkan warna rambut dan status perkahwinan ialah p/ubah yang menerangkan data individu di kawasan A.
- P/ubah warna rambut setiap individu mengambil nilai: **hitam, perang, putih, kelabu.**
- P/ubah status perkahwinan mengambil nilai: **bujang, berkahwin, bercerai, ibu tunggal.**

1.1.1 Pemboleh ubah Dedua/Boolean (Binari):

- P/ubah nominal yang hanya mempunyai dua kategori (0, atau 1).

Contoh: Status merokok pesakit jantung. Mengambil nilai **0=tidak merokok, 1=merokok.**

1.2 Pemboleh ubah ordinal: P/ubah yang mengambil nilai berkategori yang boleh ditertibkan atau dipangkatkan.

Contoh: Gred pelajar: **A+, A-, A, B+, B, B-, C+, C, C-, D, D-, E.**

- Kepuasan Pelanggan: **0=sangat tidak puas hati, 1=kurang memuaskan, 2=sederhana, 3=memuaskan, 4=sangat memuaskan.**

2. Pemboleh ubah Berangka (numeric): P/ubah berangka bersifat kuantitatif. Mengambil nilai angka/nombor.

- Terbahagi kepada p/ubah diskrit & p/ubah selanjar.

2.1 Pemboleh Ubah Diskrit: P/ubah yang mengambil nilai terhingga (*finite*) atau tak terhingga boleh hitung (*countable infinite*).

- Dalam bentuk integer. ***0, 1, 2…..n***

Contoh: Bilangan anak, bilangan kereta, umur, dan lain-lain.

2.2 Pemboleh ubah Selanjar: P/ubah yang mengambil nilai tak terhingga (*infinite*).

- Dalam bentuk sebarang nombor nyata. Boleh mengambil sebarang nilai dalam selang, contoh: ***40<X<70***

Contoh: jumlah pendapatan, tinggi rakyat malaysia, suhu bilik, kelajuan angin.

KELAS-KELAS DALAM R:

- Semua yang dikodkan dalam R dikenali sebagai objek.

- Objek dalam R terdiri daripada 5 kelas berikut:

i) Aksara (*Character*)

ii) Angka/Nombor Nyata (*Numeric/Real Numbers*)

iii) Integer

iv) Logikal (Benar/ Salah)

v) Nombor Kompleks ($a+bi$)

BENTUK DATA DALAM R:

- Secara umum, R mempunyai 6 bentuk data iaitu:

i) Skalar
ii) Vektor
iii) Matriks
iv) Bingkai Data (*data frame*)
v) Senarai (*list*)
vi) Array

- **Skalar:**

- Vektor dengan satu unsur.

- **Vektor:**

- Data bagi p/ubah tunggal disimpan dalam bentuk vektor.
- Semua unsur dalam vektor adalah sama kelas.

- **Matriks:**

- Gabungan beberapa vektor akan membentuk matriks.
- Matriks ialah persembahan struktur data 2 dimensi.
- Ianya ditunjukkan menerusi set bilangan baris dan lajur.
- Namun, unsur dalam matrik hanya boleh mengandungi nombor nyata atau integer sahaja.

- **Bingkai Data (*Data Frame*):**

- Bingkai Data ialah persembahan struktur data 2 dimensi (sama seperti matriks).
- Namun unsur dalam bingkai data boleh terdiri daripada kelas yang berbeza.

- **Senarai (*List*):**

- Senarai ialah gabungan beberapa vektor ataupun p/ubah.
- Unsur dalam senarai boleh terdiri daripada kelas yang berbeza.

- **Array:**

- Matriks/bingkai data yang lebih dari 2 dimensi.

OPERATOR DALAM R:

i) Operator Aritmetik:

- operator yang terletak antara dua operan.

Operator	Penerangan	Operator	Penerangan
+	Penambahan	^ atau **	kuasa
-	Penolakan	x %% y	modulus (x mod y) Cth: 5%%2 ialah 1 (baki)
*	Pendaraban	x %/% y	Pembahagian integer Cth: 5%/%2 is 2
/	Bahagi	x %*% y	Pendaraban Matriks

ii) Operator Hubungan:

- operator yang digunakan untuk melakukan perbandingan antara dua pemboleh ubah.

Operator	Penerangan
<	Kurang daripada
<=	Kurang daripada atau sama dengan
>	Lebih besar daripada
>=	Lebih besar daripada Atau sama dengan
==	Sama dengan

iii) Operator Logik:

- operator untuk menggabungkan beberapa operator hubungan.

Operator	Penerangan
!=	Tidak sama dengan
!x	Bukan x
x | y	x atau y
x & y	x dan y
isTRUE(x)	Uji jika X ialah BENAR

FUNGSI TELAH TERSEDIA ADA DALAM R:

Fungsi Matematik:

```
abs(x)      # The absolute value of "x"
log(x),logb(),log10(),log2(),exp(),expm1(),log1p(),sqrt()   #Fairly obvious
cos(),sin(),tan(),acos(),asin(),atan(),atan2()        # Usual stuff
cosh(),sinh(),tanh(),acosh(),asinh(),atanh()          # Hyperbolic functions
union(),intersect(),setdiff(),setequal()              # Set operations
+,-,*,/,^,%%,%/%                                      # Arithmetic operators
<,>,<=,>=,==,!=                                       # Comparison operators

eigen()       # Computes eigenvalues and eigenvectors
deriv()       # Symbolic and algorithmic derivatives of simple expressions
integrate()   # Adaptive quadrature over a finite or infinite interval.
sqrt(),sum()
```

Fungsi Statistik:

```
cor.test()              # Perform correlation test
cumsum(); cumprod(); cummin(); cummax()# Cumulative functions
density(x)              # Compute kernel density estimates
ks.test()               # Performs one or two sample Kolmogorov-Smirnov tests
loess(), lowess()       # Scatter plot smoothing
mad()                   # Calculate median absolute deviation
mean(x), weighted.mean(x), median(x), min(x), max(x), quantile(x)
rnorm(), runif()  # Generate random data with Gaussian/uniform distribution
splinefun()             # Perform spline interpolation
smooth.spline()         # Fits a cubic smoothing spline
sd()                    # Calculate standard deviation
summary(x)              # Returns a summary of x: mean, min, max etc.
t.test()                # Student's t-test
var()                   # Calculate variance
sample()                # Random samples & permutations
ecdf()                  # Empirical Cumulative Distribution Function
qqplot()                # quantile-quantile plot
lm              # Fit liner model
glm             # Fit generalised linear model
nls             # non-linear (weighted) least-squares fitting
lqs             # "library(MASS)" resistant regression
optim           # general-purpose optimisation
optimize        # 1-dimensional optimisation
constrOptim     # Constrained optimisation
nlm             # Non-linear minimisation
nlminb          # More robust (non-)constrained non-linear minimisation
```

PLOT-PLOT ASAS DALAM R:

1. Histogram dan plot Ketumpatan.
2. Plot Kotak (Boxplot)
3. Plot Serakan (Scatter Plot)
4. Plot Q-Q.
5. Carta Pai

Dan banyak lagi.

PAKEJ DALAM R:

- Pakej-pakej dalam R mengandungi koleksi fungsi, data, kod yang khusus untuk analisis tertentu.
- Direktori di mana pakej disimpan dipanggil sebagi *library*.
- Pakej-pakej khas dalam R boleh dimuat turun secara percuma.

- **install.packages**("*nama package*")

Penggunaan pakej akan menjadikan analisis lebih mudah.

- **library(***nama package***)**

- Malah, pelbagai teknik-teknik analisis statistik dan perlombongan data yang lebih kompleks dapat dijalankan.

LATIHAN:

Nama Syarikat	Pendapatan Bulanan (RM)	Bilangan Pekerja	Kategori Perniagaan (1. Company = P, 2. Limited = B, 3. Private = S)	Modal Bulanan(RM)	Status Syarikat(1 = Large, 0 = Small)	Status Penarafan
ABC	15214.32	4	P	5000.00	0	A
Bookstore Timah	3126.60	3	P	1530.00	0	B
Prasa	5211.10	1	S	3211.00	0	B-
Delta	3000.00	1	P	1444.00	0	B
Alfa	12431.11	2	P	4372.11	0	A+
Gama	290000.00	1974	P	100000.00	1	A+
Bakeri Hasan	16321.00	5	S	6421.66	0	A
Shel	176342.11	100	B	54320.00	1	A+
DV	6251.99	4	S	4421.00	0	C
Viva	6011.23	4	S	5432.31	0	C-
Kedai Sate Ali	4321.67	2	S	1500.00	0	B-
Kedai Runcit Abu	6743.28	2	S	2000.00	0	B
Kilang Apel	167223.90	200	B	48761.00	1	A+
DM	26590.12	7	P	6000.00	0	A

LATIHAN:

Berdasarkan data yang diberikan, gunakan pengaturcaraan R untuk mendapatkan penyelesaian bagi arahan berikut:

1. Bina data untuk setiap lajur merujuk kepada pembolehubah yang berasingan.
2. Berdasarkan data dalam (1), gabungkan semua pembolehubah dalam bentuk bingkai data.
3. Berdasarkan data dalam (2), susun data untuk semua pembolehubah berdasarkan tingkat pendapatan (daripada pendapatan rendah kepada tinggi).
4. Berdasarkan data dalam (3), bina pembolehubah baharu yang menunjukkan syarikat yang mempunyai pendapatan melebihi RM10000.00 sebulan.
5. Berdasarkan data dalam (3), huraikan ringkasan statistik tentang data tersebut.
6. Bina plot data yang sesuai untuk memaparkan maklumat data dalam (3).

RUJUKAN:

- Aggarwal, C.C. (2015). *Data Mining. The Textbook*. Springer, New York.
- Castellani, B. C., Rajaram, R. (2022). *Big Data Mining and Complexity*. SAGE Publications.
- Han, J., Pei, J., Tong, H. (2022). *Data Mining: Concepts and Techniques. 4th edition*. Morgan Kaufmann.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning with application in R*. Springer, New York.
- Tan, P-G., Steinbach, M., Kumar, V. (2018). *Introduction to data mining*. 2nd edition. Pearson Education, Boston.
- Torgo, L. (2011). *Data mining with R. Learning with case studies*. Taylor & Francis Group, Boca Raton.

TOPIK SETERUSNYA:

Teknik-teknik Asas Jelajahan Data Menggunakan Pengaturcaraan R