# #Mengindex Struktur Data

## #1. Pengindeksan mengikut nombor dan nama.

### #1.1 sampel vektor

v = c(1,4,4,3,2,2,3)

v[c(2,4,6)]

v[c(2,1,6,4)]

### #1.2 sampel bingkai data

data = read.table(header=T, text='

subject sex size

1 M 7

2 F 6

3 F 9

4 M 11')

#### #1.2.1 ambil data baris 1&2, untuk kolum 1&3

data[c(1,2),c(1,3)]

#### #1.2.2 ambil data baris 2&3, untuk kolum "sex" dan "size"

data[c(2,3),c("sex","size")]

## #2. Pengindeksan melalui vektor Boolean.

### #2.1 sampel vektor V > 2

v[v>2]

### #2.2 bingkai data wanita dan size > 6

v3 = data$sex == "F" & data$size > 6

data[v3,]

## #3. Pengindeksan negatif.

### #3.1 vektor

#buang unsur ke 3 dan 6

v[-c(3,6)]

### #3.2 bingkai data

#kekalkan data selain baris 2&3 dan kolum "sex"

data[-c(2,3),-2]

## #4. Pengekstrakan bersyarat

#vektor

x = 11:30

#ambil nombor yang lebih besar dari 12 dan kurang dari 20

x[x>12 & x<20]

#ambil nombor bukan 15 dan bukan 20

x[x!=15&x!=20]

#------------------------------------------------------

# #Subset bagi struktur data

## #1. Subset mengikut nombor dan nama.

### #1.1 kekalkan nombor subset >3

subset(v, v<3)

### #1.2 data aksara

t = c("small","small","large","small","moderate")

#keluarkan data "small"

subset(t, t!="small")

### #1.3 data frame

subset(data,size>6&sex=="M")

## #2. Subset bagi baris dan lajur tertentu.

subset(data, subject>2, select=c("sex","size"))

#Latihan ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#pilih data dgn size > 6 & <10 serta kekalkan lajur subject & sex

subset(data,size>6&size<10,

select=c("subject", "sex"))

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## #3. Subset berdasarkan Operator Logik (AND).

### #3.1 pilih data dgn subject <3 dan wanita

subset(data, subject<3 & sex=="F")

## #4. Subset berdasarkan Operator Logik (OR).

### #4.1 pilih data dgn subject <3 atau wanita

subset(data, subject<3 | sex=="F")

## #5. Subset data dengan syarat penjelmaan.

subset(data, log2(size)>3)

log2(data$size)

#------------------------------------------------------

# #Mendapat maklumat asas daripada data

data(mtcars)

mtcars

## #1. Menyenaraikan nama pembolehubah-pembolehubah dalam set data.

names(mtcars)

## #2. Dapatkan maklumat tentang pembolehubah. (variable)

ls() #---to visualize list of set variables

### #2.1 to remove variable(pembolehubah)

rm(v3)

ls()

## #3. Dapatkan maklumat tentang saiz dan struktur data.

str(mtcars)

class(mtcars)

nrow(mtcars)

ncol(mtcars)

dim(mtcars)

## #4. Lihat n baris pertama bagi set data.

head(mtcars, 10)

## #5. Dapatkan jumlah data lenyap.

colSums(is.na(mtcars))

#----------------------------------------------------

# #Pembundaran nombor

## #1. Membundarkan kepada integer terdekat.

x = seq(-2.5,2.5, by=0.03)

round(x,1)

## #2. Membundarkan ke sempadan atas integer.

ceiling(x)

## #3. Membundarkan ke sempadan bawah integer.

floor(x)

## #4. Membundarkan kepada bilangan titik perpuluhan tertentu.

xp = c(0.001,0.07,1.2,44.0233,738.1111,27.998)

round(xp,digits=2)

#----------------------------------------------------

# #Pengisihan (sorting)

## #1. Pengisihan mengikut tertib menaik.

v = rnorm(10)

sort(v)

## #2. Pengisihan mengikut tertib menurun.

sort(v, decreasing=T)

## #3. Pengisihan dalam bingkai data sepadan dengan beberapa ciri tertentu.

attach(mtcars) #supaya boleh call column tanpa '$'

### #3.1 isih mengikut susunan p/ubah mpg

newdata = mtcars[order(mpg),]

### #3.2 isih mengikut susunan p/ubah mpg & cyl

newdata2 = mtcars[order(cyl,mpg),]

newdata2

#Latihan ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#isih data susunan menurun p/ubah cly & susnan meningkat p/ubah hp

newdata3 = mtcars[order(-cyl,hp),]

newdata3

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#----------------------------------------------------

# # Perawakan

## #1. Perawakan tertib vektor.

v = 11:20

vi2 = sample(v)

vi2

## #2. Perawakan tertib bingkai data.

data3 = data.frame(label=letters[1:5],

number=11:15)

data4 = data3[sample(1:nrow(data3)),]

data4

#----------------------------------------------------

# #aras dalam p/ubah faktor (rename classes)

## #1. Membina pembolehubah faktor baharu.

fi = c("alpha","beta","gamma","alpha","beta")

xf = as.factor(fi)

class(xf)

xf

## #2. Mengtakrif pembolehubah faktor bertertib.

ses = c("low", "middle", "low", "low", "low", "middle", "low", "middle", "middle", "high","high")

ses2 = as.factor(ses)

ses.f = factor(ses, levels=c("low","middle","high"))

ses.f

### #2.2 bingkai data

mtcars$cyl = as.factor(mtcars$cyl)

mtcars$gear = factor(mtcars$gear, levels=c("3","4","5"))

mtcars$gear

str(mtcars)

## #3. Namakan semula aras faktor.

library(plyr) #use "plyr" library

xf2 = revalue(xf, c("beta"="delta","gamma"="zeta"))

xf2

## #4. Menambah dan menurunkan aras dalam pembolehubah faktor.

ses.f = factor(ses.f, levels=c(levels(ses.f), "very high"))

ses.f

levels(ses.f)

rm(ses.f)

newdata = droplevels(ses.f)

newdata

#----------------------------------------------------

# #jujukan dalam blok

## #1. Mentakrifkan saiz blok.

x = rnorm(22)\*100

x

### #1.1 blok saiz 4

block = ceiling(length(x)/4)\*4

block #24 values needed for block size 4 required

## #2. Menggantikan nilai yang terkurang dengan NA.

x[block] = NA

x[23]

## #3. Membina matriks blok baris.

xm = matrix(x,nrow=4)

xm

## #4. Menghitung ukuran statistik lajur.

colMeans(xm,na.rm=T)

## #5. Jalankan Pengekodan Panjang (Run Length Encoding).

# hitung bilangan nilai seragam dalam jujukan

vr = c("A","A","A","B","B","B","B","X","L","C","C","B","C","C","C")

vlr = rle(vr)

vlr #results will be A 3 times, B 4 times, X once, L once....

test = as.factor(vr)

summary(test)

#----------------------------------------------------

# #ifelse dan nested ifelse

## #1. Pernyataan mudah ifelse( ).

mydata = data.frame(x1=seq(1,20,by=2),

x2=sample(100:200,10,FALSE),

x3=LETTERS[1:10])

mydata

### #1.1 bina binari p/ubah y (-1 atau 0)

#jika x2>150, maka y=-1

#jika sebaliknya y=0

mydata$y = ifelse(mydata$x2>150, -1, 0)

mydata

## #2. Fungsi ifelse( ) terhadap pemboleh ubah kualitatif (aksara).

### #2.1 bina data y2

#jika x3 ialah A & D, x1 darab dengan 2

#selainnya x1 darab dengan 3

mydata$y2 = ifelse(mydata$x3=="A"|mydata$x3=="D", mydata$x1\*2, mydata$x1\*3)

mydata$y2 = ifelse(mydata$x3%in%c("A","D"), mydata$x1\*2, mydata$x1\*3)

mydata

## #3. Pernyataan Nested ifelse( ).

### #3.1 bina data y3 x1 darab dengan 5 jika x3 ialah nilai A, B, dan D

#jika nilai x3 ialah C&H x1 darab 10

#selainnya x2 darab dengan 20

mydata$y3 = ifelse(mydata$x3%in%c("A","B","D"), mydata$x1\*5,

ifelse(mydata$x3%in%c("C","H"), mydata$x1\*10,

mydata$x2\*20)

)

mydata

#----------------------------------------------------

# #aggregate data

## #1. Pengagregatan satu pemboleh ubah & kumpulan berdasarkan satu pemboleh ubah.

df = data.frame(team=c('A','A','A','B','B','B','C','C'),

event=c('E','E','W','W','W','W','W','W'),

point=c(1,3,3,4,5,7,7,9),

level=c(7,7,8,3,2,7,14,13))

#hitung purata point berdasarkan team

aggregate(point~team,data=df,FUN=mean)

## #2. Pengagregatan satu pemboleh ubah & kumpulan berdasarkan pemboleh ubah berganda.

### #2.1 hitung purata point berdasarkan team dan event

aggregate(point~team+event, data=df, FUN=mean)

### #2.2 sishan piawai

aggregate(point~team+event, data=df, FUN=sd)

## #3. Pengagregatan pemboleh ubah berganda & kumpulan berdasarkan satu pemboleh ubah.

#hitung purata point dan level berdasarkan team

aggregate(cbind(point,level)~team, data=df, FUN=mean)

## #4. Pengagregatan pemboleh ubah berganda & kumpulan berdasarkan pemboleh ubah

berganda.

#hitung purata point dan level berdasarkan team dan event

aggregate(cbind(point,level)~team+event, data=df, FUN=mean)

#----------------------------------------------------

# #Penggelungan

## #Fungsi Apply

dat = data.frame(x=c(1:5,81),

z=c(1,1,99,0,1,0),

y=5\*c(1:6))

dat

#cari nilai maksimum bagi setiap baris

apply(dat,1,max)

#cari nilai maksimum bagi setiap lajur

#for loop

x=NULL

for(i in 1:ncol(dat))

{x[i] = max(dat[i])}

x

#----------------------------------------------------

# #Pakej DPLYR

library(dplyr)

mydata = read.csv(file.choose())

head(mydata)

## #1. Pemilihan rawak N baris.

sample\_n(mydata, 15)

## #2. Pemilihan rawak pecahan/peratusan baris.

sample\_frac(mydata, 0.5)

## #3. Menyusun semula pembolehubah.

mydata5 = select(mydata,State,Y2002,Index,everything())

mydata5

## #4. Menamakan semula pembolehubah.

mydata6 = rename(mydata, IndexXY=Index,

USState=State)

mydata6

## #5. Menapis baris.

#pilih subset data untuk index "A"

mydata7 = filter(mydata, Index=="A")

mydata7

## #6. Pemililihan kriteria berganda.

mydata8 = filter(mydata, Index%in%c("A","C"))

mydata8

## #7. Syarat 'AND' dalam pemilihan kriteria.

mydata9 = filter(mydata,

Index%in%c("A","C") & Y2002>=130000)

mydata9

## #8. Syarat 'OR' dalam pemilihan kriteria.

mydata10 = filter(mydata,

Index%in%c("A","C") | Y2002>=130000)

mydata10

## #9. Syarat NOT.

#ambil data bukan Index A,C,M & N

mydata11 = filter(mydata,

!Index%in%c("A","C","M","N"))

mydata11

## #10. Syarat CONTAINS.

#teknik digunakan untuk cari padanan data

#pattern matching

#dapatkan data dengan syarat p/ubah untuk State yang ada pattern nama Ar

mydata12 = filter(mydata, grepl("Ar", State))

mydata12

## #11. Memperihalkan pemboleh ubah terpilih.

#hitung min & median by p/ubah 2015

summarise(mydata, Y2015\_mean=mean(Y2015),

Y2015\_median=median(Y2015)

)

## #12. Memperihalkan pemboleh ubah berganda.

summarise\_at(mydata, vars(Y2005,Y2006),

list(means=mean, medians=median)

)

## #13. Memperihalkan data berdasarkan fungsi tersuai (Custom functions).

summarise\_at(mydata, vars(Y2005,Y2006),

function(x) log(sd(x-mean(x)))

)

## #14. Memperihalkan semua pemboleh ubah berangka.

summarise\_if(mydata, is.numeric,

list(means=mean, medians=median)

)

## #15. Menyusun data menerusi pemboleh ubah berganda.

### #15.1 Susunan menaik

arrange(mydata, Index, Y2011)

### #15.2 Susunan menurun (index)

arrange(mydata, desc(Index), Y2011)

## #16. Operator Pipe %>%.

#berguna untuk tulis sub-queries

#gabungkan beberapa fungsi secara serentak

dt2 = mydata%>%select(Index,State,Y2002)%>%

sample\_n(10)

dt2

## #17. Memperihalkan data menerusi pemboleh ubah berkategori.

#hitung min p/ubah Y2011 & Y2012 mengikut kumpulan p/ubah index

tdata = mydata%>%group\_by(Index)%>%summarise\_at(vars(Y2011,Y2012),list(means=mean,variances=var))

tdata

## #18. Penapisan data dalam pemboleh ubah berkategori.

tdata2 = mydata%>%select(Index,Y2015)%>%

filter(Index%in%c("A","C","I"))%>%

group\_by(Index)%>%

do(arrange(.,desc(Y2015)))%>%

slice(1) #to select nth number of row

tdata2

#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