Pembersihan Data

install.packages(“tinytex”) tinytex::install\_tinytex()

# Pembersihan Data

#1. Kenalpasti corak data-data lenyap library(mice)

MData = read.csv(file.choose(), sep = “;”) head(MData, 10)

md.pattern(MData) # Kenal pasti lokasi data lenyap str(MData) summary(MData)

#2. Keluarkan cerapan yang mengandungi data lenyap MData2 = MData[complete.cases(MData),] MData2

#2.1 Lihat cerapan yang mempunyai data lenyap MData[!complete.cases(MData),]

#3. Lengkapkan data lenyap secara manual

MDataindus indus.fix = edit(MData$indu) # Tak tukar ori data indus.fix

#4. Gunakan sukatan memusat sebagai anggaran terhadap data lenyap attach(MData)

par(mfrow =c(1,3)) hist(crim) #tak simetri hist(indus) #tak simteri hist(medv) # tak simetri hist(rm) # simetri

#4.2 Untuk data taburan bersifat pincang/bukan simetri: median boleh digunakan. # Kenal pasti median data

#crim median.crim = median(crim, na.rm=T) median.crim crim.fix = ifelse(is.na(crim), median.crim, crim)

par(mfrow = c(1,2)) hist(crim, main=“Bentuk taburan data asal”) hist(crim.fix, main=“Bentuk taburan data dengan anggaran median”)

#indus median.indus = median(indus, na.rm=T) median.crim indus.fix = ifelse(is.na(indus), median.indus, indus)

par(mfrow = c(1,2)) hist(indus, main=“Bentuk taburan data asal”) hist(indus.fix, main=“Bentuk taburan data dengan anggaran median”)

#medv median.medv = median(medv, na.rm=T) median.medv medv.fix = ifelse(is.na(medv), median.medv, medv)

par(mfrow = c(1,2)) hist(medv, main=“Bentuk taburan data asal”) hist(medv.fix, main=“Bentuk taburan data dengan anggaran median”)

#4.3 Untuk data taburan normal/simetri dengan nilai berangka: nilai min boleh digunakan.

#rm mean.rm = mean(rm, na.rm=T) mean.rm rm.fix = ifelse(is.na(rm), mean.rm, rm)

par(mfrow = c(1,2)) hist(rm, main=“Bentuk taburan data asal”) hist(rm.fix, main=“Bentuk taburan data dengan anggaran median”)