Integrasi Data

1. Import data ke dalam R

```
data(iris)
data("precip")
```

```
rm(iris)
rm(precip) #remove data
```

1.2 Excel Data

jenis .xlsx

```
library(openxlsx)
data.ex = read.xlsx(xlsxFile = "G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/Big Mart Da
                    sheet =1,
                    startRow = 1)
head(data.ex,10)
```

```
##    Item_Identifier Item_Weight Item_Fat_Content Item_Visibility
## 1            FDA15       9.300          Low Fat        0.016047
## 2            DRC01       5.920          Regular        0.019278
## 3            FDN15      17.500          Low Fat        0.016760
## 4            FDX07      19.200          Regular        0.000000
## 5            NCD19       8.930          Low Fat        0.000000
## 6            FDP36      10.395          Regular        0.000000
## 7            FDO10      13.650          Regular        0.012741
## 8            FDP10          NA          Low Fat        0.127470
## 9            FDH17      16.200          Regular        0.016687
## 10           FDU28      19.200          Regular        0.094450
##                 Item_Type Item_MRP Outlet_Identifier Outlet_Establishment_Year
## 1                   Dairy 249.8092            OUT049                      1999
## 2             Soft Drinks  48.2692            OUT018                      2009
## 3                    Meat 141.6180            OUT049                      1999
## 4  Fruits and Vegetables 182.0950            OUT010                      1998
## 5               Household  53.8614            OUT013                      1987
## 6            Baking Goods  51.4008            OUT018                      2009
## 7             Snack Foods  57.6588            OUT013                      1987
## 8             Snack Foods 107.7622            OUT027                      1985
## 9            Frozen Foods  96.9726            OUT045                      2002
## 10           Frozen Foods 187.8214            OUT017                      2007
##    Outlet_Size Outlet_Location_Type       Outlet_Type Item_Outlet_Sales
## 1       Medium               Tier 1 Supermarket Type1         3735.1380
```

```
## 2       Medium              Tier 3 Supermarket Type2          443.4228
## 3       Medium              Tier 1 Supermarket Type1         2097.2700
## 4         <NA>              Tier 3     Grocery Store          732.3800
## 5         High              Tier 3 Supermarket Type1          994.7052
## 6       Medium              Tier 3 Supermarket Type2          556.6088
## 7         High              Tier 3 Supermarket Type1          343.5528
## 8       Medium              Tier 3 Supermarket Type3         4022.7640
## 9         <NA>              Tier 2 Supermarket Type1         1076.5990
## 10        <NA>              Tier 2 Supermarket Type1         4710.5350
```

1.3 jenis .csv

```
Data3 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/custdata3.csv", header = T
head(Data3,10)
```

```
##      X state.of.res  custid sex is.employed income       marital.stat
## 1  215      Florida   46791   M          NA  22700            Married
## 2  444     Michigan   36825   M        TRUE  17500            Married
## 3  949     Virginia  415060   F       FALSE  16000 Divorced/Separated
## 4  122   California 1159665   M        TRUE  37000            Married
## 5  994    Wisconsin 1131536   M        TRUE  80000            Married
## 6   58   California  819436   F          NA  56800            Widowed
## 7  536       Nevada 1082333   M        TRUE  45000 Divorced/Separated
## 8  557   New Jersey  968478   F        TRUE  46800 Divorced/Separated
## 9  639     New York 1185185   M          NA  19000            Married
## 10 590   New Mexico  864947   F          NA  22700 Divorced/Separated
##    health.ins                housing.type recent.move num.vehicles age
## 1        TRUE     Homeowner free and clear       FALSE            2  67
## 2       FALSE                      Rented       FALSE            2  35
## 3        TRUE Homeowner with mortgage/loan       FALSE            1  60
## 4        TRUE Homeowner with mortgage/loan       FALSE            4  46
## 5        TRUE Homeowner with mortgage/loan       FALSE            4  57
## 6        TRUE Homeowner with mortgage/loan       FALSE            1  81
## 7        TRUE     Homeowner free and clear       FALSE            2  50
## 8        TRUE Homeowner with mortgage/loan       FALSE            3  48
## 9        TRUE Homeowner with mortgage/loan       FALSE            1  71
## 10       TRUE     Homeowner free and clear       FALSE            1  67
##    is.employed.fix1 Median.Income        gp income.lt.30K age.range Income
## 1           missing         56895 0.5949839          TRUE  (65,Inf]  45000
## 2          employed         62634 0.5550284          TRUE   (25,65]  60000
## 3      not employed         53914 0.1122820          TRUE   (25,65]     NA
## 4          employed         39832 0.2773083         FALSE   (25,65] 402000
## 5          employed         41073 0.1444786         FALSE   (25,65]     NA
## 6           missing         39832 0.3011925         FALSE  (65,Inf] 120300
## 7          employed         52498 0.5055806         FALSE   (25,65]     NA
## 8          employed         68187 0.1249902         FALSE   (25,65]  23800
## 9           missing         44819 0.1120956          TRUE  (65,Inf]     NA
## 10          missing         68071 0.3309783          TRUE  (65,Inf]  30900
```

1.4 jenis .txt

```
Data4 = read.table("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/car_prices.txt", header=
head(Data4,10)
```

```
##        x
## 1  25000
## 2  18000
## 3  22000
## 4  27000
## 5  35000
## 6  29000
## 7  31000
## 8  27000
## 9  24000
## 10 26000
```

2. Teknik integrasi Data dari sumber/format berbeza

2.1 Integrasi data yang berlainan attribut

```
mydata1 = read.table("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/mydata1.txt", header=
mydata2 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/mydata2.csv", header=TRU

mydata3 = cbind(mydata1, mydata2)
mydata3new = mydata3[, -c(7,8)]
```

2.2 Integrasi data nama attribut yang tak konsisten

```
mydata5 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/mydata5.csv", header=T)
load("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/mydata4.RData")

mydata6 = merge(mydata4,mydata5, by.x="ID", by.y="IDPerson")
```

2.3 Integrasi data nama saiz tak sama (INNER_JOIN)

```
mydata7 = mydata5[1:10,]
mydata8 = merge(mydata4, mydata7, by.x="ID", by.y="IDPerson")
```

semua data yang tak sepadan akan dikeluarkan

3. jika nak kekalkan data yang tak sepadan

data yang tak sepadan akan ditaarof sebagai NA (FULL_JOIN)

```
mydata9 = merge(mydata4, mydata7, by.x="ID", by.y="IDPerson", all=T)
```

3.1 menamakan semula atribut

4 ubah suai nilai data yang tak konsisten

4.1 ubah suai secara manual

```
# mydata11 = edit(mydata10)
```

4.2 ubah suai data tak konsisten (huruf besar dan kecil)

```
dataM1 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dataM1.csv", header=T)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Taarifkan nama atribut yang nak diubah suaikan

```
city_name = function(city) { city = tolower(city) #tukarkan semua ayat ke huruf
city = trimws(city) #buang semua spacing
city = gsub("+", "", city) #ganti dengan hanya 1 spacing
city = tools::toTitleCase(city) #format ke Title case
return(city) }
```

```
dataM1$City = sapply(dataM1$City, city_name)
```

4.3 ubah suai data tak konsisten (ejaan singkat dan ejaan penuh)

```
dataM2 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dataM2.csv", header=T)
```

Petakan singkat ke ejaan penuh

```
city_mapping = list('NY'='New York', 'LA'='Los Angeles', 'CHI'='Chicago')
```

bina fungsi untuk ubah suai nama singkat ke penuh

```
standard_city_name = function(city){ if(city%in%names(city_mapping)){ return(city_mapping[[city]])} else
```

gunakan fungsi terhadap data

```
dataM2$City = sapply(dataM2$City, standard_city_name)
```

5. Buang data yang berulang (redundant)

```
dataM3 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dataM3.csv", header=T, s
```

```
dataM3_NoDup = dataM3%>%distinct(dataM3$id, .keep_all=T)
```

6. eksport data (save data)

6.1 save file R

```
getwd()
```

```
## [1] "G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Rmd_File"
```

```
#save(dataM3_NoDup, file="dataM3_NoDup.RData")
```

6.2 save file.csv

```
#write.csv(mydata11, file='mydata11.csv')
```

6.3 save file txt

```
#write.table(Data4, file='mydata11.txt', sep='\t')
```

7. Kesan duplikasi data dengan id yang sama

```
dataM4 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dataM4.csv", header=T)
```

```
duplikasi_ID = dataM4[duplicated(dataM4$id) | duplicated(dataM4$id, fromLast=T),]
head(duplikasi_ID,10)
```

```
##    id name age gender income
## 2 102  Bob  60      M  63755
## 5 105  Eva  36      F  55489
## 6 102  Bob  60      M  63755
## 9 105  Eva  36      F  55489
```