Class 7

R – Data frame

and

Import and Export Datasets

Data Frame

- It is a 2D array like structure
- It’s a tabular data referred as a data frame in R
- Each column name considered as variable
- Used to store data table
- It is the special case of the list in which each component has equal length

Data Frame - Characteristics

- Column names should not be empty
- Row names should be unique
- Values can be a numeric, factor or character type
- Each column should contain same number of data items
- The data.frame() function used for creating data frame in R

Data Frame - Example

```
> #Creation of Data frame
> mydata <- data.frame(seq_id=c("Seq_1","Seq_2","Seq_3"),
+                      seq_name = c("DREB","ALS","WRKY"),
+                      seq_bp = c("ATCG","TGCA","CATG"),
+                      seq_length = c(100,200,300))
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG        100
2  Seq_2      ALS   TGCA        200
3  Seq_3     WRKY   CATG        300
```

Data Frame - Structure

```
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG         100
2  Seq_2      ALS   TGCA         200
3  Seq_3     WRKY   CATG         300
> str(mydata)
'data.frame':   3 obs. of  4 variables:
 $ seq_id    : chr  "Seq_1" "Seq_2" "Seq_3"
 $ seq_name  : chr  "DREB" "ALS" "WRKY"
 $ seq_bp    : chr  "ATCG" "TGCA" "CATG"
 $ seq_length: num  100 200 300
```

Data Frame – Extracting Data

- Extract specific column from a data frame using the column name
- Extract the specific rows from data frame
- Extract data from specific rows and columns

Data Frame – Extracting Data

Extract specific column from a data frame using the column name

```
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG         100
2  Seq_2      ALS   TGCA         200
3  Seq_3     WRKY   CATG         300
> #Retrieve Based on the column
> ext_data <- data.frame(mydata$seq_id,mydata$seq_name)
> print(ext_data)
  mydata.seq_id mydata.seq_name
1         Seq_1            DREB
2         Seq_2             ALS
3         Seq_3            WRKY
```

Data Frame – Extracting Data

Extract the specific rows from data frame

```
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG        100
2  Seq_2      ALS   TGCA        200
3  Seq_3     WRKY   CATG        300
> #Extract Based on the row
> print(mydata[1:2,])
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG        100
2  Seq_2      ALS   TGCA        200
```

Data Frame – Extracting Data

Extract data from specific rows and columns

```
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG         100
2  Seq_2      ALS   TGCA         200
3  Seq_3     WRKY   CATG         300
> #Extract from specific row and coloumn
> print(mydata[1,3])
[1] "ATCG"
```

Data Frame – How to check?

is.data.frame()

```
> mydata
  seq_id seq_name seq_bp seq_length
1  Seq_1     DREB   ATCG        100
2  Seq_2      ALS   TGCA        200
3  Seq_3     WRKY   CATG        300
> #Check Data frame
> is.data.frame(mydata)
[1] TRUE
```

Import Datasets

Import Functions

```
read.table()

read.csv()

read.csv2()

read.delim()

read.delim2()
```

Export Functions

```
write.table()

write.csv()

write.csv2()
```

Subset Functions

- Useful for extracting data based on the expression value.

Basic Syntax:-

```
subset(x,subset,…)
```

Handling Excel Files

- R can read excel files using some of specific packages.
- Few Packages:-

    xlsx

    readxl

    readxlsb

install.packages("xlsx")

Reading Excel File

```
read.xlsx(
  file,
  sheetIndex,
  sheetName = NULL,
  rowIndex = NULL,
  startRow = NULL,
  endRow = NULL,
  colIndex = NULL,
  as.data.frame = TRUE,
  header = TRUE,
  colClasses = NA,
  keepFormulas = FALSE,
  encoding = "unknown",
  password = NULL,
  ...
)
```