**SAVEETHA SCHOOL OF ENGINEERING**

**SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES**

**ITA 0443 - STATISTICS WITH R PROGRAMMING FOR REAL TIME PROBLEM**

**DAY 2 – LAB EXERCISES**

**Reg No: 192124100**

**Name:Mary Sushmija.X.M**

**IMPLEMENTATION OF VECTOR RECYCLING, APPLY FAMILY & RECURSION**

**1. Demonstrate Vector Recycling in R.**

**A:**

**> vec1=1:6**

**> vec2=1:2**

**> print(vec1+vec2)**

**[1] 2 4 4 6 6 8**

**B:**

**> vec1=20:25**

**> vec2=4:6**

**> print(vec1+vec2)**

**[1] 24 26 28 27 29 31**

**C:**

**> vec1=10:14**

**> vec2=3:5**

**> print(vec1+vec2)**

**[1] 13 15 17 16 18**

**2. Demonstrate the usage of apply function in R**

**> my\_matrix <- matrix((1:12), nrow=3)**

**> print(my\_matrix)**

**[,1] [,2] [,3] [,4]**

**[1,] 1 4 7 10**

**[2,] 2 5 8 11**

**[3,] 3 6 9 12**

**> print(apply(my\_matrix, 1, max))**

**[1] 10 11 12**

**3. Demonstrate the usage of lapply function in R**

**> names <- c("priyanl", "abhiraj", "pawananjani", "sudhanshu", "devraj")**

**> print("original data:")**

**[1] "original data:"**

**> names**

**[1] "priyanl" "abhiraj" "pawananjani" "sudhanshu" "devraj"**

**> print("data after lapply():")**

**[1] "data after lapply():"**

**> lapply(names, toupper)**

**[[1]]**

**[1] "PRIYANL"**

**[[2]]**

**[1] "ABHIRAJ"**

**[[3]]**

**[1] "PAWANANJANI"**

**[[4]]**

**[1] "SUDHANSHU"**

**[[5]]**

**[1] "DEVRAJ"**

**4. Demonstrate the usage of sapply function in R**

**> sample\_data<- data.frame( x=c(1,2,3,4,5,6),y=c(3,2,4,2,34,5))**

**> print("original data:")**

**[1] "original data:"**

**> sample\_data**

**x y**

**1 1 3**

**2 2 2**

**3 3 4**

**4 4 2**

**5 5 34**

**6 6 5**

**> print("data after sapply():")**

**[1] "data after sapply():"**

**> sapply(sample\_data,max)**

**x y**

**6 34**

**5. Demonstrate the usage of tapply function in R**

**> df <- data.frame(team=c('A','A','A','A','B','B','B','B'),**

**+ position=c('G','G','F','F','G','G','F','F'),**

**+ points=c(14,19,13,8,15,15,17,19),**

**+ assists=c(4,3,3,5,9,14,15,12))**

**> df**

**team position points assists**

**1 A G 14 4**

**2 A G 19 3**

**3 A F 13 3**

**4 A F 8 5**

**5 B G 15 9**

**6 B G 15 14**

**7 B F 17 15**

**8 B F 19 12**

**6. Demonstrate the usage of mapply function in R**

**> A = list(c(1,2,3,4))**

**> B = list(c(2,5,1,6))**

**> result = mapply(sum,A,B)**

**> print(result)**

**[1] 24**

**7. Sum of Natural Numbers using Recursion**

**> sum\_natural <- function(n) {**

**+ if (n <= 1) {**

**+ return(n)**

**+ } else {**

**+ return(n + sum\_natural(n - 1))**

**+ }**

**+ }**

**> result <- sum\_natural(5)**

**> print(result)**

**[1] 15**

**8. Write a program to generate Fibonacci sequence using Recursion in R**

**> fibonacci <- numeric()**

**> fibonacci[1] <- fibonacci[2] <- 1**

**> for (i in 3:10) fibonacci[i] <- fibonacci[i - 2] + fibonacci[i - 1]**

**> print("first fibonacci numbers:")**

**[1] "first fibonacci numbers:"**

**> print(fibonacci)**

**[1] 1 1 2 3 5 8 13 21 34 55**

**9. Write a program to find factorial of a number in R using recursion.**

**> factorial <- function(n) {**

**+ if (n <= 1) {**

**+ return(1)**

**+ } else {**

**+ return(n \* factorial(n - 1))**

**+ }**

**+ }**

**> result <- factorial(5)**

**> print(result)**

**[1] 120**

**CREATION AND MANIPULATION OF DATAFRAMES IN R**

**Exercise 1**

Consider two vectors: x=seq(1,43,along.with=Id)

y=seq(-20,0,along.with=Id)

Create a data frame ‘df’ as shown below.

>df

Id Letter x y

1 1 a 1.000000 -20.000000

2 1 b 4.818182 -18.181818

3 1 c 8.636364 -16.363636

4 2 a 12.454545 -14.545455

5 2 b 16.272727 -12.727273

6 2 c 20.090909 -10.909091

7 3 a 23.909091 -9.090909

8 3 b 27.727273 -7.272727

9 3 c 31.545455 -5.454545

10 4 a 35.363636 -3.636364

11 4 b 39.181818 -1.818182

12 4 c 43.000000 0.000000

**CODE:**

**> Id <- rep(seq(1, 4), each=3)**

**> Letter <- rep(letters[1:3], 4)**

**>**

**> Id <- rep(seq(1, 4), each=3)**

**> Letter <- rep(letters[1:3], 4)**

**> x <- seq(1, 43, along.with = Id)**

**> y <- seq(-20, 0, along.with = Id)**

**> df <- data.frame(Id, Letter, x, y)**

**> print(df)**

**Id Letter x y**

**1 1 a 1.000000 -20.000000**

**2 1 b 4.818182 -18.181818**

**3 1 c 8.636364 -16.363636**

**4 2 a 12.454545 -14.545455**

**5 2 b 16.272727 -12.727273**

**6 2 c 20.090909 -10.909091**

**7 3 a 23.909091 -9.090909**

**8 3 b 27.727273 -7.272727**

**9 3 c 31.545455 -5.454545**

**10 4 a 35.363636 -3.636364**

**11 4 b 39.181818 -1.818182**

**12 4 c 43.000000 0.000000**

**Exercise 2**

Using the data frame ‘df’ in Exercise1, Construct the following data frame. Id x.ay.ax.by.bx.cy.c 1 1 1.00000 -20.000000 4.818182 -18.181818 8.636364 -16.363636 4 2 12.45455 -14.545455 16.272727 -12.727273 20.090909 -10.909091 7 3 23.90909 -9.090909 27.727273 -7.272727 31.545455 -5.454545 10 4 35.36364 -3.636364 39.181818 -1.818182 43.000000 0.000000

**CODE:**

**> print(df)**

**Id Letter x y**

**1 1 a 1.000000 -20.000000**

**2 1 b 4.818182 -18.181818**

**3 1 c 8.636364 -16.363636**

**4 2 a 12.454545 -14.545455**

**5 2 b 16.272727 -12.727273**

**6 2 c 20.090909 -10.909091**

**7 3 a 23.909091 -9.090909**

**8 3 b 27.727273 -7.272727**

**9 3 c 31.545455 -5.454545**

**10 4 a 35.363636 -3.636364**

**11 4 b 39.181818 -1.818182**

**12 4 c 43.000000 0.000000**

**Exercise 3**

Create two data frame df1 and df2:

> df1

Id Age

1. 1 14
2. 2 12
3. 3 15
4. 4 10

> df2

Id Sex Code

1. 1 F a
2. 2 M b
3. 3 M c
4. 4 F d

From df1 and df2 create M:

>M

Id Age Sex Code

1. 1 14 F a
2. 2 12 M b
3. 3 15 M c 4 4 10 F d

**CODE:**

**> df1 <- data.frame(Id = 1:4, Age = c(14, 12, 15, 10))**

**> df2 <- data.frame(Id = 1:4, Sex = c("F", "M", "M", "F"), Code = c("a", "b", "c", "d"))**

**> M <- merge(df1, df2, by = "Id")**

**> df3 <- data.frame(id2 = c(4, 3, 2, 1), score = c(100, 98, 94, 99))**

**> N <- merge(M, df3, by.x = "Id", by.y = "id2")**

## **Exercise 4**

Create a data frame df3:

> df3 id2 score 1 4 100

1. 3 98
2. 2 94
3. 1 99

From M (used in Exercise-3) and df3 create N:

Id Age Sex Code score

1. 1 14 F a 99
2. 2 12 M b 94
3. 3 15 M c 98 4 4 10 F d 100

**CODE:**

**> N <- N[, c("Id", "Age", "score")]**

**> values <- unlist(N)**

**> ind <- rep(c("Id", "Age", "score"), each = nrow(N))**

**> result <- data.frame(values, ind)**

## **Exercise 5**

Consider the previous one data frame N:

1) Remove the variables Sex and Code

2) From N, create a data frame:

values ind

1. 1 Id
2. 2 Id
3. 3 Id
4. 4 Id
5. 14 Age
6. 12 Age
7. 15 Age
8. 10 Age
9. 99 score
10. 94 score
11. 98 score
12. 100 score

**CODE**

**> N <- N[, c("Age", "score")]**

**> values <- c(N$Age, N$score)**

**> ind <- c(rep("Id", length(N$Age)), rep("score", length(N$score)))**

**> new\_df <- data.frame(values, ind)**

## **Exercise 6**

For this exercise, we’ll use the (built-in) dataset trees.

1. Make sure the object is a data frame, if not change it to a data frame.
2. Create a new data frame A:

>A

Girth Height Volume

mean\_tree 13.24839 76 30.17097

min\_tree 8.30000 63 10.20000

max\_tree 20.60000 87 77.00000

sum\_tree 410.70000 2356 935.30000

**CODE:**

**> mean\_tree <- apply(trees[, 1:3], 2, mean)**

**> min\_tree <- apply(trees[, 1:3], 2, min)**

**> max\_tree <- apply(trees[, 1:3], 2, max)**

**> sum\_tree <- apply(trees[, 1:3], 2, sum)**

**> A <- data.frame(**

**+ Girth = c(mean\_tree[1], min\_tree[1], max\_tree[1], sum\_tree[1]),**

**+ Height = c(mean\_tree[2], min\_tree[2], max\_tree[2], sum\_tree[2]),**

**+ Volume = c(mean\_tree[3], min\_tree[3], max\_tree[3], sum\_tree[3]),**

**+ row.names = c("mean\_tree", "min\_tree", "max\_tree", "sum\_tree")**

**+ )**

**> # Print data frame A**

## **Exercise 7**

Consider the data frame A:

1. Order the entire data frame by the first column.

+ A <- data.frame(V1 = c(4, 2, 6, 8),

+ V2 = c(3, 1, 7, 5),

+ V3 = c(9, 11, 10, 12),

+ V4 = c(15, 13, 14, 16))

1. Rename the row names as follows: mean, min, max, tree

> A <- A[order(A[, 1]), ]

> rownames(A) <- c("mean", "min", "max", "tree")

## **Exercise 8**

Create an empty data frame with column types:

>df

IntsLogicals Doubles Characters

(or 0-length row.names)

## **CODE:**

> df <- data.frame(Ints = integer(),

+ Logicals = logical(),

+ Doubles = double(),

+ Characters = character(),

+ stringsAsFactors = FALSE)

## **Exercise 9**

Create a data frame XY

X=c(1,2,3,1,4,5,2)

Y=c(0,3,2,0,5,9,3)

> XY

X Y

1. 1 0
2. 2 3
3. 3 2
4. 1 0
5. 4 5
6. 5 9
7. 2 3
8. look at duplicated elements using a provided R function.

**CODE:**

**> XY <- data.frame(X=c(1,2,3,1,4,5,2), Y=c(0,3,2,0,5,9,3))**

**> duplicated\_rows <- duplicated(XY)**

**> duplicated\_XY <- XY[duplicated\_rows, ]**

**> print(duplicated\_XY)**

**X Y**

**4 1 0**

**7 2 3**

1. keep only the unique lines on XY using a provided R function.

CODE:

> XY <- data.frame(X=c(1,2,3,1,4,5,2), Y=c(0,3,2,0,5,9,3))

> unique\_XY <- unique(XY)

> print(unique\_XY)

X Y

1 1 0

2 2 3

3 3 2

5 4 5

6 5 9

## **Exercise 10**

Use the (built-in) dataset Titanic.

1. Make sure the object is a data frame, if not change it to a data frame.

CODE:

> data(Titanic)

> df <- as.data.frame(Titanic)

1. Define a data frame with value 1st in Class variable, and value NO in Survived variable and variables Sex, Age and Freq.

Sex Age Freq

1 Male Child 0

5 Female Child 0

9 Male Adult 118

13 Female Adult 4

CODE:

> df\_filtered <- df[df$Class == "1st" & df$Survived == "No", c("Sex", "Age", "Freq")]

> df\_filtered

Sex Age Freq

1 Male Child 0

5 Female Child 0

9 Male Adult 118

13 Female Adult 4

MERGING DATAFRAMES

## **Exercise 11 a)**

Create the following dataframes to merge:

buildings<- data.frame(location=c(1, 2, 3), name=c("building1", "building2","building3"))

data <- data.frame(survey=c(1,1,1,2,2,2),location=c(1,2,3,2,3,1),efficiency=c(51,64,70,7,80,58))

The dataframes, *buildings*and *data*have a common key variable called, “location”.

Use the merge() function to merge the two dataframes by “location”, into a new dataframe,“buildingStats”.

CODE:  
> buildings <- data.frame(location=c(1, 2, 3), name=c("building1", "building2","building3"))

> data <- data.frame(survey=c(1,1,1,2,2,2),location=c(1,2,3,2,3,1),efficiency=c(51,64,70,7,80,58))

> buildingStats <- merge(buildings, data, by="location")buildingStats <- merge((buildings, data, by="location")

## **Exercise 11 b)**

Give the dataframes different key variable names:

buildings<- data.frame(location=c(1, 2, 3), name=c("building1","building2", "building3"))

data <- data.frame(survey=c(1,1,1,2,2,2), LocationID=c(1,2,3,2,3,1), efficiency=c(51,64,70,71,80,58))

The dataframes, buildings and data have corresponding variables called, location, and LocationID. Use the merge() function to merge the columns of the two dataframes by the corresponding variables.

CODE:  
> buildings <- data.frame(location=c(1, 2, 3), name=c("building1","building2", "building3"))

> data <- data.frame(survey=c(1,1,1,2,2,2), LocationID=c(1,2,3,2,3,1), efficiency=c(51,64,70,71,80,58))

> buildingStats <- merge(buildings, data, by.x="location", by.y="LocationID")

DIFFERENT TYPES OF MERGE IN R

**Exercise 12a)InnerJoin:**

The R merge() function automatically joins the frames by common variable names. In that case, demonstrate how you would perform the merge in **Exercise 11a** without specifying the key variable.

**Exercise 12b)OuterJoin:**

Merge the two dataframes from **Exercise 11a**. Use the “all=” parameter in the merge() function to return all records from both tables. Also, merge with the key variable, “location”.

**Exercise 12c)Left Join:**

Merge the two dataframes from **Exercise 11a**, and return all rows from the left table. Specify the matching key from **Exercise 11a.**

**Exercise 12d)Right Join:**

Merge the two dataframes from **Exercise 11a,** and return all rows from the right table. Use the matching key from **Exercise 11a** to return matching rows from the left table.

**Exercise 12e)Cross Join:**

Merge the two dataframes from **Exercise 11a**, into a “Cross Join” with each row of “buildings” matched to each row of “data”. What new column names are created in “buildingStats”?

CODE:

> buildingStats <- merge(buildings, data)

> buildingStats <- merge(buildings, data, by="location", all=TRUE)

> buildingStats <- merge(buildings, data, by="location", all.x=TRUE)

> buildingStats <- merge(buildings, data, by="location", all.y=TRUE)

buildingStats <- merge(buildings, data)

### **Exercise 13MergingDataframe rows:**

To join two data frames (datasets) vertically, use the rbind function. The two data frames must have the same variables, but they do not have to be in the same order.

Merge the rows of the following two dataframes:

buildings<- data.frame(location=c(1, 2, 3), name=c("building1",

"building2", "building3"))

buildings2 <- data.frame(location=c(5, 4, 6), name=c("building5", "building4", "building6"))

Also, specify the new dataframe as, “allBuidings”.

### **Exercise 14**

Create a new dataframe, buildings3, that has variables not found in the previous dataframes.

buildings3 <- data.frame(location=c(7, 8, 9), name=c("building7", "building8", "building9"), startEfficiency=c(75,87,91))

Create a new buildings3 without the extra variables.

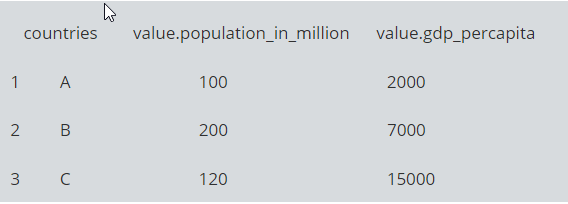
### **Exercise 15**

Instead of deleting the extra variables from buildings3. append the buildings, and buildings2 with the new variable in buildings3, **(from Exercise 14).** Set the new data in buildings and buildings2 , (**from Exercise 13)**, to NA.

**RESHAPE FUNCTION IN R**

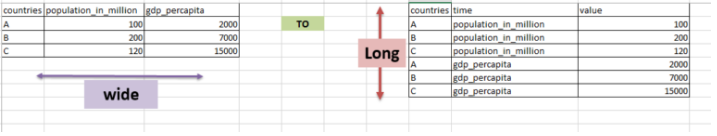
**Exercise: 16**

Construct the following data frame ‘country’.



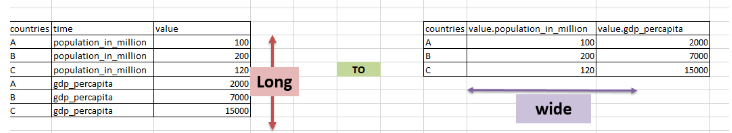
#### **Reshape in R from wide to long:**

Reshape the above data frame from wide to long format in R.



* data frame “country” is passed to reshape function
* idvar is the variable which need to be left unaltered which is “countries”
* varying are the ones that needs to converted from wide to long
* v.names are the values that should be against the times in the resultant [data frame](http://www.datasciencemadesimple.com/data-frame-in-r/).
* new.row.names is used to assign row names to the resultant dataset
* direction is, to which format the data needs to be transformed

1. **Reshape in R from long to wide:**



* data (country\_w\_to\_L) which is in long format,  is passed to reshape function
* idvar is the variable which need to be left unaltered, which is “countries”
* timevar are the variables that needs to converted to wide format
* v.names are the value variable
* direction is, to which format the data needs to be transformed
  1. **MELTING AND CASTING IN R**

**Exercises 17 :**

1. Melt airquality data set and display as a long – format data ?
2. Melt airquality data and specify month and day to be “ID variables” ?
3. Cast the molten airquality data set .
4. Use cast function appropriately and compute the average of Ozone, Solar.R , Wind and temperature per month ?

CODE:  
> data(airquality)

> airquality\_means <- aggregate(cbind(Ozone, Solar.R, Wind, Temp) ~ Month, data = airquality, mean)

> airquality\_means

Month Ozone Solar.R Wind Temp

1 5 24.12500 182.0417 11.504167 66.45833

2 6 29.44444 184.2222 12.177778 78.22222

3 7 59.11538 216.4231 8.523077 83.88462

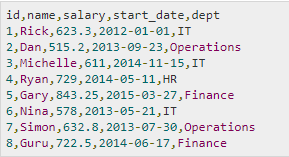
4 8 60.00000 173.0870 8.860870 83.69565

5 9 31.44828 168.2069 10.075862 76.89655

* + 1. **FILE MANUPULATION IN R**

**Exercise 18**

1. Consider the following data present. Create this file using windows notepad . Save the file as **input.csv** using the save As All files(\*.\*) option in notepad.



1. Use appropriate R commands to read **input.csv** file.

### Analyze the CSV File and compute the following.

### Get the maximum salary

### Get the details of the person with max salary

### Get all the people working in IT department

### Get the persons in IT department whose salary is greater than 600

### Get the people who joined on or after 2014

### Get the people who joined on or after 2014 and write the output onto a file called output.csv