Module 2: Basic Data Manipulation using R

Assignment: Solution

Version 2.0

1. Given Data:

X = C(1,2,3,4,23,42,5,311,1)
Y = C(23,43,54,75,76,6,87,5,43,234,2)

a. Calculate the Mean of X and Y?

Ans. First create these two vectors namely X and Y in R.
X=c(1,2,3,4,23,42,5,311,1)
Y=c(23,43,54,75,76,6,87,5,43,234,2)

Now for calculating mean:
mean(X)
mean(Y)

```
Console ~/
> X=c(1,2,3,4,23,42,5,311,1)
> Y=c(23,43,54,75,76,6,87,5,43,234,2)
> X
[1]   1   2   3   4  23  42   5 311   1
> Y
 [1]  23  43  54  75  76   6  87   5  43 234   2
> mean(X)
[1] 43.55556
> mean(Y)
[1] 58.90909
>
```

Mean of X=43.55556 and Mean of Y=58.90909

b. Calculate the Standard Deviation of X and Y?

Ans. The command for calculating standard deviation is sd()
sd(X)
sd(Y)

```
Console ~/
> X=c(1,2,3,4,23,42,5,311,1)
> Y=c(23,43,54,75,76,6,87,5,43,234,2)
> X
[1]   1   2   3   4  23  42   5 311   1
> Y
 [1]  23  43  54  75  76   6  87   5  43 234   2
> sd(X)
[1] 101.24
> sd(Y)
[1] 65.39183
>
```

Standard deviation of X = 101.24 and for Y= 65.39183

c. Calculate the Sum of X and Y?

Ans. The command for calculating sum of a vector is sum()
sum(X)
sum(Y)

```
Console ~/
> X=c(1,2,3,4,23,42,5,311,1)
> Y=c(23,43,54,75,76,6,87,5,43,234,2)
> X
[1]   1   2   3   4  23  42   5 311   1
> Y
 [1]  23  43  54  75  76   6  87   5  43 234   2
> sum(X)
[1] 392
> sum(Y)
[1] 648
>
```

Sum of vector X is 392 and that of Y is 648

d. Create a custom function that can calculate the square of any number.

Ans. In R a custom function can be easily created by the following command:
Function_name = function(parameters) {function_definition}
Squarefun=function(x) {x*x}
Now this function will return square of a number x

```
Console ~/
> squarefun=function(x) {x*x}
> # Now this function can calcualte of any number x
>
> squarefun(2)
[1] 4
> squarefun(3)
[1] 9
> squarefun(6)
[1] 36
>
```

2. Import the following file formats into R:

→ sas7bdata dataset
→ spss dataset

Ans.

```
> install.packages("foreign")
Installing package into 'C:/Users/User/Documents/R/win-
library/3.0'

> library(foreign)
```

Importing a SAS7BDAT Dataset:

To import this file format it is essential to load the 'sas7bdat' package.

```
> library(sas7bdat)
Loading required package: chron
Warning message:
package 'sas7bdat' was built under R version 3.0.2

> data(sas7bdat.sources)
```

```
> head(sas7bdat.sources)
           filename            accessed uncompressed   gzip
1  depress.sas7bdat 2012-10-31 06:35:33        17408   2906
2 drugprob.sas7bdat 2012-10-31 06:37:41       705536 237408
3 drugtest.sas7bdat 2012-10-31 06:37:51       975872  57904
4  environ.sas7bdat 2012-10-31 06:37:53       156672  15076
5   event1.sas7bdat 2012-10-31 06:37:53         9216   1102
6   event2.sas7bdat 2012-10-31 06:37:55       115712  15357
   bzip2     xz
1   2384   2156
2 204523 183124
3  23365  35832
4   9192  10256
5   1175    992
6   9352  10520

1  http://wps.ablongman.com/wps/media/objects/1108/1135330/d
```

```
> depressdata<-read.sas7bdat("depress.sas7bdat")
```

```
> class(depressdata)
[1] "data.frame"
```

```
> names(depressdata)
 [1] "ID"      "AGE"     "IQ"      "ANXIETY" "DEPRESS" "SLEEP"   "SEX"
 [8] "LIFESAT" "WEIGHT"  "SATLIFE" "GENDER"  "SLEEP1"  "NEWIQ"   "NEWAGE"
```

```
> head(depressdata)
  ID AGE IQ ANXIETY DEPRESS SLEEP SEX LIFESAT WEIGHT SATLIFE GENDER SLEEP1
1  1  39 94       2       2     2   2       2    4.9       0      0      0
2  2  41 89       2       2     2   2       2    2.2       0      0      0
3  3  42 83       3       3     2   2       2    4.0       0      0      0
4  4  30 99       2       2     2   2       2   -2.6       0      0      0
5  5  35 94       2       1     1   2       1   -0.3       1      0      1
6  6  44 90     NaN       1     2   1       1    0.9       1      1      0
  NEWIQ  NEWAGE
1  2.21  1.5424
2 -2.79  3.5424
3 -8.79  4.5424
4  7.21 -7.4576
5  2.21 -2.4576
6 -1.79  6.5424
```

Importing a SPSS Dataset:

```
> Cancer<-read.spss("Cancer.sav")

> class(Cancer)
[1] "list"

> summary(Cancer)
         Length Class  Mode
ID       25     -none- numeric
TRT      25     -none- numeric
AGE      25     -none- numeric
WEIGHIN  25     -none- numeric
STAGE    25     -none- numeric
TOTALCIN 25     -none- numeric
TOTALCW2 25     -none- numeric
TOTALCW4 25     -none- numeric
TOTALCW6 25     -none- numeric


> head(Cancer)
$ID
 [1]  1  5  6  9 11 15 21 26 31 35 39 41 45  2 12 14 16 22 24 34 37 42 44 50
[25] 58

$TRT
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1

$AGE
 [1] 52 77 60 61 59 69 67 56 61 51 46 65 67 46 56 42 44 27 68 77 86 73 67 60
[25] 54

$WEIGHIN
 [1] 124.0 160.0 136.5 179.6 175.8 167.6 186.0 158.0 212.8 189.0 149.0 157.0
```

```
> tail(Cancer)
$WEIGHIN
 [1] 124.0 160.0 136.5 179.6 175.8 167.6 186.0 158.0 212.8 189.0 149.0 1
[16] 162.6 261.4 225.4 226.0 164.0 140.0 181.5 187.0 164.0 172.8

$STAGE
 [1] 2 1 4 1 2 1 1 3 1 1 4 1 1 2 4 1 2 1 4 2 1 0 1 2 4

$TOTALCIN
 [1]  6  9  7  6  6  6  6  6  6  6  7  6  8  7  6  4  6  6 12  5  6  8

$TOTALCW2
 [1]  6  6  9  7  7  6 11 11  9  4  8  6  8 16 10  6 11  7 11  7  7 11

$TOTALCW4
 [1]  6 10 17  9 16  6 11 15  6  8 11  9  9  9 11  8 11  6 12 13  7 16

$TOTALCW6
 [1]  7  9 19  3 13 11 10 15  8  7 11  6 10 10  9  7 14  6  9 12  7 NA
```

3. The following data is available in the LMS. It contains Credit card data that is generated in batches periodically. The first dataset contains data collected so far, whereas another dataset is generated in the last hour is to be combined with this data, so that the organization has the updated complete data.

Transaction_data

custID	gender	state	cardholder	balance	numTrans	numIntlTrans	creditLine	fraudRisk
1	1	35	1	3000	4	14	2	0
2	2	2	1	0	9	0	18	0
3	2	2	1	0	27	9	16	0
4	1	15	1	0	12	0	5	0
5	1	46	1	0	11	16	7	1
6	2	44	2	5546	21	0	13	0

Hour_transaction

custID	gender	state	cardholder	balance	numTrans	numIntlTrans	creditLine	fraudRisk
8	1	10	1	6016	20	3	6	0
9	2	32	1	2428	4	10	22	0
10	1	23	1	0	18	56	5	1

Problem: Import this data into R and Combine the data using any of the built-in functions present in R.

Ans.

```
> all_transactions <- rbind(transaction_data, hour_transaction)
> all_transactions
  custID gender state cardholder balance numTrans numIntlTrans creditLine fraudRisk
1      1      1    35          1    3000        4           14          2         0
2      2      2     2          1       0        9            0         18         0
3      3      2     2          1       0       27            9         16         0
4      4      1    15          1       0       12            0          5         0
5      5      1    46          1       0       11           16          7         1
6      6      2    44          2    5546       21            0         13         0
7      8      1    10          1    6016       20            3          6         0
8      9      2    32          1    2428        4           10         22         0
9     10      1    23          1       0       18           56          5         1
```