VIT-AP UNIVERSITY, ANDHRA PRADESH

CSE2047 – Data Analytics - Lab Sheet : 2

Academic year: 2020-2021 **Branch/ Class:** B.Tech/M.Tech
Semester: Fall **Date:**
Faculty Name: Prof. S.Gopikrishnan **School:** SCOPE
Student name: Valiveti Manikanta bhuvanesh **Reg. no.: 19BCD7088**

LAB 2

1. **USE DIABETES.CSV**

   df<-read.csv("Diabetes_Updated.csv")

```
> df<-read.csv("Diabetes_Updated.csv")
```

2. **DISPLAY THE DATAFRAME**

   print(df)

```
> print(df)
     id chol stab.glu hdl ratio glyhb   location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
1  1000  203       82  56   3.6  4.31 Buckingham  46 female     62    121 medium   118    59    NA    NA    29  38      720
2  1001  165       97  24   6.9  4.44 Buckingham  29 female     64    218  large   112    68    NA    NA    46  48      360
3  1002  228       92  37   6.2  4.64 Buckingham  58 female     61    256  large   190    92   185    92    49  57      180
4  1003   78       93  12   6.5  4.63 Buckingham  67   male     67    119  large   110    50    NA    NA    33  38      480
5  1005  249       90  28   8.9  7.72 Buckingham  64   male     68    183 medium   138    80    NA    NA    44  41      300
6  1008  248       94  69   3.6  4.81 Buckingham  34   male     71    190  large   132    86    NA    NA    36  42      195
7  1011  195       92  41   4.8  4.84 Buckingham  30   male     69    191 medium   161   112   161   112    46  49      720
8  1015  227       75  44   5.2  3.94 Buckingham  37   male     59    170 medium    NA    NA    NA    NA    34  39     1020
9  1016  177       87  49   3.6  4.84 Buckingham  45   male     69    166  large   160    80   128    86    34  40      300
10 1022  263       89  40   6.6  5.78 Buckingham  55 female     63    202  small   108    72    NA    NA    45  50      240
11 1024  242       82  54   4.5  4.77     Louisa  60 female     65    156 medium   130    90   130    90    39  45      300
12 1029  215      128  34   6.3  4.97     Louisa  38 female     58    195 medium   102    68    NA    NA    42  50       90
13 1030  238       75  36   6.6  4.47     Louisa  27 female     60    170 medium   130    80    NA    NA    35  41      720
14 1031  183       79  46   4.0  4.59     Louisa  40 female     59    165 medium    NA    NA    NA    NA    37  43       60
15 1035  191       76  30   6.4  4.67     Louisa  36   male     69    183 medium   100    66    NA    NA    36  40      225
16 1036  213       83  47   4.5  3.41     Louisa  33 female     65    157 medium   130    90   120    96    37  41      240
17 1037  255       78  38   6.7  4.33     Louisa  50 female     65    183 medium   130   100    NA    NA    37  43      180
18 1041  230      112  64   3.6  4.53     Louisa  20   male     67    159 medium   100    90    NA    NA    31  39     1440
```

3. **HOW MANY ROWS AND COLUMNS ARE THERE?**

   dim(df)

```
> dim(df)
[1] 403  19
```

4. **FIND OUT THE COLUMNS NAMES IN THE DATAFRAME**

   colnames(df)

```
> colnames(df)
 [1] "id"      "chol"    "stab.glu" "hdl"     "ratio"   "glyhb"   "location" "age"     "gender"  "height"  "weight"  "frame"   "bp.1s"   "bp.1d"   "bp.2s"
[16] "bp.2d"   "waist"   "hip"      "time.ppn"
```

5. **ACCESS THE AGE COLUMN.**

   df$age

```
> df$age
  [1] 46 29 58 67 64 34 30 37 45 55 60 38 27 40 36 33 50 20 36 62 70 47 38 66 24 41 37 48 43 40 42 52 61 61 25 47 35 46 57 70 22 52 36 43 72 37 54 60 40 55 76 43 65 45 70 20
 [57] 62 92 49 44 74 36 51 38 31 28 22 71 76 91 40 23 20 40 52 76 46 48 22 58 34 61 40 28 53 67 51 49 65 54 38 64 41 67 27 21 41 47 61 65 28 41 37 50 57 28 31 83 79 68 32 26
[113] 36 53 19 63 58 53 50 41 48 59 34 63 23 21 23 36 71 64 43 31 44 60 43 48 56 55 49 58 33 48 66 59 45 52 76 36 41 20 50 43 82 35 47 75 62 31 50 39 33 58 81 27 47 33 67 42
[169] 21 51 27 51 71 50 54 59 59 40 58 72 66 23 42 43 75 65 34 37 61 36 45 68 57 41 68 40 79 62 63 55 55 27 66 63 78 68 31 64 40 61 28 34 63 55 26 36 40 45 68 82 60 30 41 54
[225] 72 47 50 51 45 38 20 44 63 50 44 48 41 29 76 69 26 70 25 42 56 31 31 27 73 32 19 71 27 31 20 31 62 44 36 36 47 30 63 48 65 59 37 78 23 38 38 41 29 49 23 29 40 38 40 29
[281] 78 50 23 60 40 60 40 30 21 63 63 43 46 64 56 35 59 22 43 26 41 43 20 28 30 66 20 32 38 61 26 74 72 21 36 42 66 34 43 57 45 44 27 63 65 30 28 41 31 33 66 28 25 26 40 38
[337] 30 52 22 51 45 53 21 53 37 34 30 74 36 45 35 50 27 52 42 39 73 28 53 49 55 37 60 56 84 20 80 60 80 29 43 63 37 20 44 54 58 35 52 60 43 59 33 37 40 38 32 60 30 42 52 59
[393] 78 51 25 37 54 89 53 51 29 41 68
```

6. DISPLAY THE NUMBER OF PEOPLE WHOSE AGE IS GREATER THAN 40.

dim(subset(df,age>40))[1]

```
> dim(subset(df,age>40))[1]
[1] 243
```

7. FIND OUT THE FEMALE DIABETIC PATIENTS OF AGE > 30

subset(df,gender=="female" & age>30)

```
> subset(df,gender=="female" & age>30)
     id chol stab.glu hdl ratio glyhb    location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
1  1000  203       82  56   3.6  4.31 Buckingham  46 female     62    121 medium   118    59    NA    NA    29  38      720
3  1002  228       92  37   6.2  4.64 Buckingham  58 female     61    256  large   190    92   185    92    49  57      180
10 1022  263       89  40   6.6  5.78 Buckingham  55 female     63    202  small   108    72    NA    NA    45  50      240
11 1024  242       82  54   4.5  4.77     Louisa  60 female     65    156 medium   130    90   130    90    39  45      300
12 1029  215      128  34   6.3  4.97     Louisa  38 female     58    195 medium   102    68    NA    NA    42  50       90
14 1031  183       79  46   4.0  4.59     Louisa  40 female     59    165 medium    NA    NA    NA    NA    37  43       60
16 1036  213       83  47   4.5  3.41     Louisa  33 female     65    157 medium   130    90   120    96    37  41      240
17 1037  255       78  38   6.7  4.33     Louisa  50 female     65    183 medium   130   100    NA    NA    37  43      180
20 1250  196      206  41   4.8 11.24 Buckingham  62 female     65    196  large   178    90    NA    NA    46  51      540
23 1254  203      299  43   4.7 12.74 Buckingham  38 female     69    288  large   136    83    NA    NA    48  55      240
24 1256  281       92  41   6.9  5.56 Buckingham  66 female     62    185  large   158    88   160    88    48  44      285
26 1277  179       80  92   1.9  4.18 Buckingham  41 female     72    118  small   144   112    NA    NA    28  36      780
29 1282  254       84  52   4.9  4.52 Buckingham  43 female     62    145 medium   125    70    NA    NA    31  38      720
31 1301  177      101  36   4.9  5.11 Buckingham  42 female     65    174 medium   146    94   139    89    37  40      540
34 1305  182       85  37   4.9  5.66 Buckingham  61 female     69    174 medium   176    86   180    90    49  43      330
36 1312  183       81  60   3.1  4.03 Buckingham  47 female     66    186 medium   140    97    NA    NA    39  44      780
42 1321  218       68  46   4.7  3.89 Buckingham  52 female     62    170 medium   142    79    NA    NA    40  43      720
45 1500  213       76  40   5.3  5.96 Buckingham  72 female     59    137  large   130    60    NA    NA    40  40       90
```

8. FIND OUT THE DETAILS OF PATIENTS WHO ARE NOT FROM LOUSIA.

subset(df,location !="Louisa")

```
> subset(df,location !="Louisa")
     id chol stab.glu hdl ratio glyhb    location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
1  1000  203       82  56   3.6  4.31 Buckingham  46 female     62    121 medium   118    59    NA    NA    29  38      720
2  1001  165       97  24   6.9  4.44 Buckingham  29 female     64    218  large   112    68    NA    NA    46  48      360
3  1002  228       92  37   6.2  4.64 Buckingham  58 female     61    256  large   190    92   185    92    49  57      180
4  1003   78       93  12   6.5  4.63 Buckingham  67   male     67    119  large   110    50    NA    NA    33  38      480
5  1005  249       90  28   8.9  7.72 Buckingham  64   male     68    183 medium   138    80    NA    NA    44  41      300
6  1008  248       94  69   3.6  4.81 Buckingham  34   male     71    190  large   132    86    NA    NA    36  42      195
7  1011  195       92  41   4.8  4.84 Buckingham  30   male     69    191 medium   161   112   161   112    46  49      720
8  1015  227       75  44   5.2  3.94 Buckingham  37   male     59    170 medium    NA    NA    NA    NA    34  39     1020
9  1016  177       87  49   3.6  4.84 Buckingham  45   male     69    166  large   160    80   128    86    34  40      300
10 1022  263       89  40   6.6  5.78 Buckingham  55 female     63    202  small   108    72    NA    NA    45  50      240
20 1250  196      206  41   4.8 11.24 Buckingham  62 female     65    196  large   178    90    NA    NA    46  51      540
21 1252  186       97  50   3.7  6.49 Buckingham  70   male     67    178  large   148    88   148    84    42  41     1020
22 1253  234       65  76   3.1  4.67 Buckingham  47   male     67    230  large   137   100   149   110    45  46      480
23 1254  203      299  43   4.7 12.74 Buckingham  38 female     69    288  large   136    83    NA    NA    48  55      240
24 1256  281       92  41   6.9  5.56 Buckingham  66 female     62    185  large   158    88   160    88    48  44      285
25 1271  228       66  45   5.1  4.61 Buckingham  24 female     61    113 medium   100    70   110    70    33  38      210
26 1277  179       80  92   1.9  4.18 Buckingham  41 female     72    118  small   144   112    NA    NA    28  36      780
27 1280  232       87  30   7.7  5.10 Buckingham  37   male     68    252  large   140    95    NA    NA    43  47      420
28 1281   NA       74  NA    NA  4.28 Buckingham  48   male     68    100  small   120    85    NA    NA    27  33      510
```

9. IF GLUCOSE LEVELS IN BLOOD IS > 7, DIAGNOSE AS DIABETIC BY ADDING A COLUMN TO THE DATA FRAME.

df$diabetic <- df$glyhb >= 7.0

```
> df$diabetic <- df$glyhb >= 7.0
> print(df)
     id chol stab.glu hdl ratio glyhb    location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn diabetic
1  1000  203       82  56   3.6  4.31 Buckingham  46 female     62    121 medium   118    59    NA    NA    29  38      720    FALSE
2  1001  165       97  24   6.9  4.44 Buckingham  29 female     64    218  large   112    68    NA    NA    46  48      360    FALSE
3  1002  228       92  37   6.2  4.64 Buckingham  58 female     61    256  large   190    92   185    92    49  57      180    FALSE
4  1003   78       93  12   6.5  4.63 Buckingham  67   male     67    119  large   110    50    NA    NA    33  38      480    FALSE
5  1005  249       90  28   8.9  7.72 Buckingham  64   male     68    183 medium   138    80    NA    NA    44  41      300     TRUE
6  1008  248       94  69   3.6  4.81 Buckingham  34   male     71    190  large   132    86    NA    NA    36  42      195    FALSE
7  1011  195       92  41   4.8  4.84 Buckingham  30   male     69    191 medium   161   112   161   112    46  49      720    FALSE
8  1015  227       75  44   5.2  3.94 Buckingham  37   male     59    170 medium    NA    NA    NA    NA    34  39     1020    FALSE
9  1016  177       87  49   3.6  4.84 Buckingham  45   male     69    166  large   160    80   128    86    34  40      300    FALSE
10 1022  263       89  40   6.6  5.78 Buckingham  55 female     63    202  small   108    72    NA    NA    45  50      240    FALSE
11 1024  242       82  54   4.5  4.77     Louisa  60 female     65    156 medium   130    90   130    90    39  45      300    FALSE
12 1029  215      128  34   6.3  4.97     Louisa  38 female     58    195 medium   102    68    NA    NA    42  50       90    FALSE
13 1030  238       75  36   6.6  4.47     Louisa  27 female     60    170 medium   130    80    NA    NA    35  41      720    FALSE
14 1031  183       79  46   4.0  4.59     Louisa  40 female     59    165 medium    NA    NA    NA    NA    37  43       60    FALSE
15 1035  191       76  30   6.4  4.67     Louisa  36   male     69    183 medium   100    66    NA    NA    36  40      225    FALSE
16 1036  213       83  47   4.5  3.41     Louisa  33 female     65    157 medium   130    90   120    96    37  41      240    FALSE
17 1037  255       78  38   6.7  4.33     Louisa  50 female     65    183 medium   130   100    NA    NA    37  43      180    FALSE
18 1041  230      112  64   3.6  4.53     Louisa  20   male     67    159 medium   100    90    NA    NA    31  39     1440    FALSE
```

10. WHICH FEMALE SUBJECTS FROM BUCKINGHAM ARE UNDER THE AGE OF 25?

subset(df,gender=="female" & age<25 & location == "Buckingham")

```
> subset(df,gender=="female" & age<25 & location == "Buckingham")
      id chol stab.glu hdl ratio glyhb    location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn diabetic
25  1271  228       66  45   5.1  4.61 Buckingham  24 female     61    113 medium   100    70   110    70    33  38      210    FALSE
41  1317  136       81  51   2.7  4.58 Buckingham  22 female     66    160  large   105    85    NA    NA    35  40      720    FALSE
56  2763  193      106  63   3.1  6.35 Buckingham  20 female     68    274  small   165   110   153   100    49  58       60    FALSE
67  2787  223       75  85   2.6  4.25 Buckingham  22 female     62    137 medium   120    70    NA    NA    28  35      960    FALSE
72  3250  164       86  40   4.1  5.23 Buckingham  23 female     69    245  large   126    75    NA    NA    44  47      420    FALSE
73  3750  170       69  64   2.7  4.39 Buckingham  20 female     64    161 medium   108    70    NA    NA    37  40      120    FALSE
79  4506  217       81  60   3.6  3.93 Buckingham  22 female     71    223 medium   120    75    NA    NA    46  50      210    FALSE
126 10001  132       99  34   3.9  4.01 Buckingham  21 female     65    169  large   112    62    NA    NA    39  43      180    FALSE
169 15264  187       84  64   2.9  4.40 Buckingham  21 female     63    158  small   138    88    NA    NA    39  43      180    FALSE
251 17790  146       79  41   3.6  4.76 Buckingham  19 female     60    135 medium   108    58    NA    NA    33  40      240    FALSE
255 17800  149       77  49   3.0  4.50 Buckingham  20 female     62    115  small   105    82    NA    NA    31  37      720    FALSE
275 20260  179       75  36   5.0  4.75 Buckingham  23 female     65    183 medium   120    80    NA    NA    43  45      720    FALSE
283 20279  147       78  42   3.5  4.67 Buckingham  23 female     61    185   <NA>   127    71    NA    NA    43  47      600    FALSE
```

11. WHAT IS THEIR AVERAGE GLYHB?

mean(df$glyhb,na.rm = TRUE)

```
> mean(df$glyhb,na.rm = TRUE)
[1] 5.589769
```

12. ARE ANY OF THEM DIABETIC?

subset(df,diabetic==TRUE)

```
> subset(df,diabetic==TRUE)
      id chol stab.glu hdl ratio glyhb   location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn diabetic
5   1005  249       90  28   8.9  7.72 Buckingham  64   male     68    183 medium   138    80     NA     NA    44  41      300     TRUE
20  1250  196      206  41   4.8 11.24 Buckingham  62 female     65    196  large   178    90     NA     NA    46  51      540     TRUE
23  1254  203      299  43   4.7 12.74 Buckingham  38 female     69    288  large   136    83     NA     NA    48  55      240     TRUE
33  1304  265      330  34   7.8 15.52 Buckingham  61   male     74    191 medium   170    88    168     80    39  41      225     TRUE
40  1316  182      206  43   4.2  7.91 Buckingham  70   male     69    214  large   158    90    160     96    45  48      840     TRUE
48  2004  128      223  24   5.3 10.90 Buckingham  60   male     67    196 medium   110    68     NA     NA    42  43      450     TRUE
55  2762  289      111  50   5.8  9.39 Buckingham  70 female     60    220 medium   126    80     NA     NA    51  54      780     TRUE
59  2773  237      233  58   4.1 13.70 Buckingham  49 female     62    189  large   130    90     NA     NA    43  47      195     TRUE
61  2775  296      262  60   4.9 10.93 Buckingham  74 female     63    183  large   159    99    160    103    42  48      300     TRUE
63  2778  443      185  23  19.3 14.31 Buckingham  51 female     70    235 medium   158    98    148     88    43  48      420     TRUE
68  2791  213      203  75   2.8 11.41 Buckingham  71 female     63    165 medium   150    80    145     80    34  42      960     TRUE
70  2794  232      184 114   2.0  8.40 Buckingham  91 female     61    127   <NA>   170    82     NA     NA    35  38      120     TRUE
76  4000  209      113  65   3.2  7.44 Buckingham  76 female     60    143  large   156    78    144     76    35  40     1200     TRUE
87  4760  218      182  54   4.0 10.55     Louisa  51 female     NA    215  large   139    69     NA     NA    42  53      720     TRUE
92  4771  249      197  44   5.7  9.17     Louisa  64 female     63    159 medium   151    85    148     79    33  41     1140     TRUE
98  4787  245      120  39   6.3  7.79     Louisa  47 female     63    156 medium   142   102    156    106    35  39      120     TRUE
100 4790  224      341  33   6.8 10.15     Louisa  65   male     67    197 medium   160    80    158     80    42  43      390     TRUE
105 4796  209      176  55   3.8  9.77     Louisa  57 female     61    150  small   115    68     NA     NA    36  39      780     TRUE
109 4805  292      235  55   5.3  7.87 Buckingham  79   male     70    165   <NA>   170    90    170    100    39  41      240     TRUE
```

13. FIND OUT EACH COLUMN TYPE IN THE DATAFRAME

str(df)

```
> str(df)
'data.frame':   403 obs. of  20 variables:
 $ id      : int  1000 1001 1002 1003 1005 1008 1011 1015 1016 1022 ...
 $ chol    : int  203 165 228 78 249 248 195 227 177 263 ...
 $ stab.glu: int  82 97 92 93 90 94 92 75 87 89 ...
 $ hdl     : int  56 24 37 12 28 69 41 44 49 40 ...
 $ ratio   : num  3.6 6.9 6.2 6.5 8.9 ...
 $ glyhb   : num  4.31 4.44 4.64 4.63 7.72 ...
 $ location: chr  "Buckingham" "Buckingham" "Buckingham" "Buckingham" ...
 $ age     : int  46 29 58 67 64 34 30 37 45 55 ...
 $ gender  : chr  "female" "female" "female" "male" ...
 $ height  : int  62 64 61 67 68 71 69 59 69 63 ...
 $ weight  : int  121 218 256 119 183 190 191 170 166 202 ...
 $ frame   : chr  "medium" "large" "large" "large" ...
 $ bp.1s   : int  118 112 190 110 138 132 161 NA 160 108 ...
 $ bp.1d   : int  59 68 92 50 80 86 112 NA 80 72 ...
 $ bp.2s   : int  NA NA 185 NA NA NA 161 NA 128 NA ...
 $ bp.2d   : int  NA NA 92 NA NA NA 112 NA 86 NA ...
 $ waist   : int  29 46 49 33 44 36 46 34 34 45 ...
 $ hip     : int  38 48 57 38 41 42 49 39 40 50 ...
 $ time.ppn: int  720 360 180 480 300 195 720 1020 300 240 ...
 $ diabetic: logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
> summary(df)
```

14. PRODUCE THE SUMMARY OF THE DATAFRAME.

summary(df)

```
> summary(df)
       id              chol         stab.glu          hdl              ratio          glyhb          location              age       
 Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.500   Min.   : 2.68   Length:403         Min.   :19.00  
 1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.200   1st Qu.: 4.38   Class :character   1st Qu.:34.00  
 Median :15766   Median :204.0   Median : 89.0   Median : 46.00   Median : 4.200   Median : 4.84   Mode  :character   Median :45.00  
 Mean   :15978   Mean   :207.8   Mean   :106.7   Mean   : 50.45   Mean   : 4.522   Mean   : 5.59                      Mean   :46.85  
 3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00   3rd Qu.: 5.400   3rd Qu.: 5.60                      3rd Qu.:60.00  
 Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.300   Max.   :16.11                      Max.   :92.00  
                 NA's   :1                       NA's   :1        NA's   :1        NA's   :13                                        
    gender              height          weight          frame              bp.1s           bp.1d           bp.2s      
 Length:403         Min.   :52.00   Min.   : 99.0   Length:403         Min.   : 90.0   Min.   : 48.00   Min.   :110.0  
 Class :character   1st Qu.:63.00   1st Qu.:151.0   Class :character   1st Qu.:121.2   1st Qu.: 75.00   1st Qu.:138.0  
 Mode  :character   Median :66.00   Median :172.5   Mode  :character   Median :136.0   Median : 82.00   Median :149.0  
                    Mean   :66.02   Mean   :177.6                      Mean   :136.9   Mean   : 83.32   Mean   :152.4  
                    3rd Qu.:69.00   3rd Qu.:200.0                      3rd Qu.:146.8   3rd Qu.: 90.00   3rd Qu.:161.0  
                    Max.   :76.00   Max.   :325.0                      Max.   :250.0   Max.   :124.00   Max.   :238.0  
                    NA's   :5       NA's   :1                          NA's   :5       NA's   :5        NA's   :262    
     bp.2d            waist             hip            time.ppn       diabetic      
 Min.   : 60.00   Min.   :26.0   Min.   :30.00   Min.   :   5.0   Mode :logical  
 1st Qu.: 84.00   1st Qu.:33.0   1st Qu.:39.00   1st Qu.:  90.0   FALSE:330      
 Median : 92.00   Median :37.0   Median :42.00   Median : 240.0   TRUE :60       
 Mean   : 92.52   Mean   :37.9   Mean   :43.04   Mean   : 341.2   NA's :13       
 3rd Qu.:100.00   3rd Qu.:41.0   3rd Qu.:46.00   3rd Qu.: 517.5                  
 Max.   :124.00   Max.   :56.0   Max.   :64.00   Max.   :1560.0                  
 NA's   :262      NA's   :2      NA's   :2       NA's   :3                       
```