VIT-AP UNIVERSITY, ANDHRA PRADESH

CSE2047 – Data Analytics - Lab Sheet : 4

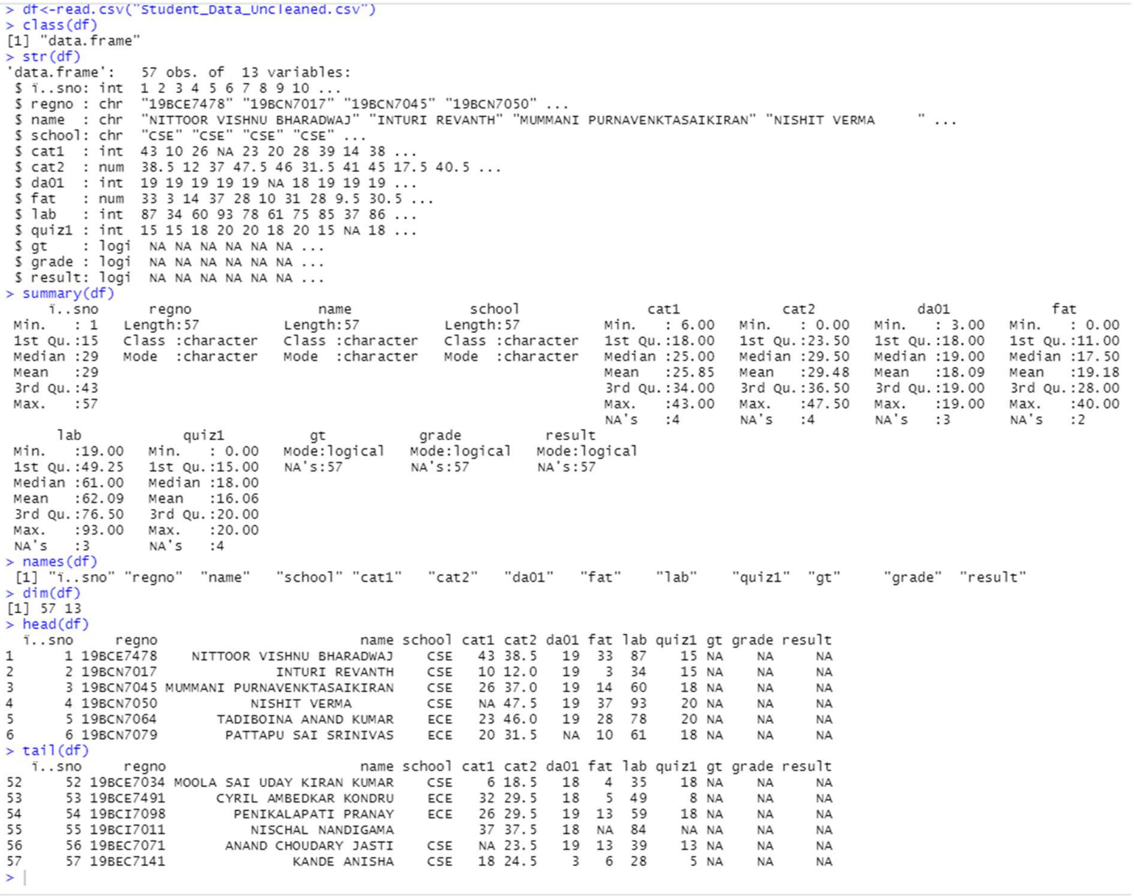
Academic year: 2020-2021 **Branch/ Class:** B.Tech/M.Tech
Semester: Fall **Date:**
Faculty Name: Prof. S.Gopikrishnan **School:** SCOPE
Student name: Valiveti Manikanta bhuvanesh **Reg. no.: 19BCD7088**

LAB 4 (Data Cleaning and Imputation)

Questions:

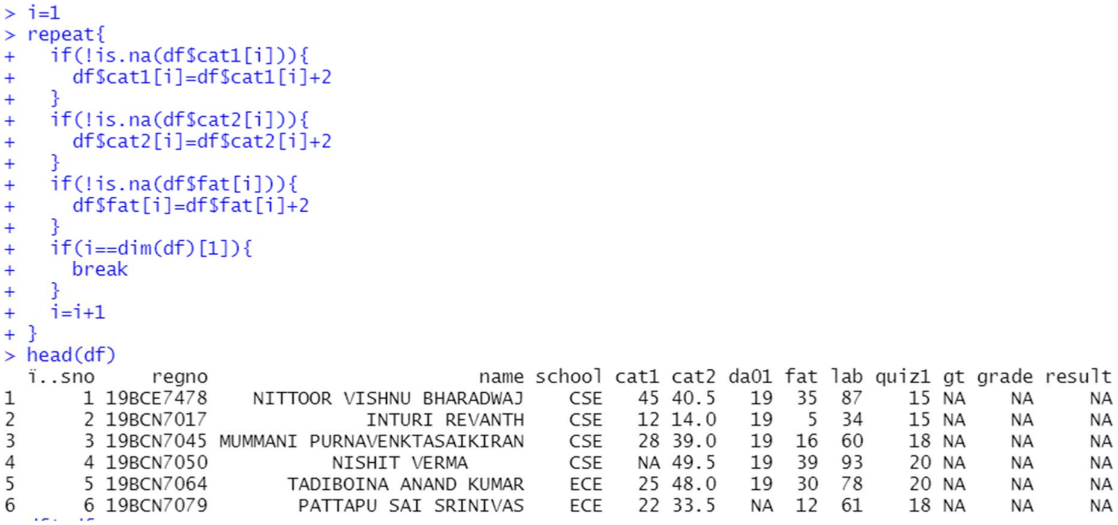
(Use Student_Data.csv)

1. Do preliminary observations. (head, str…)



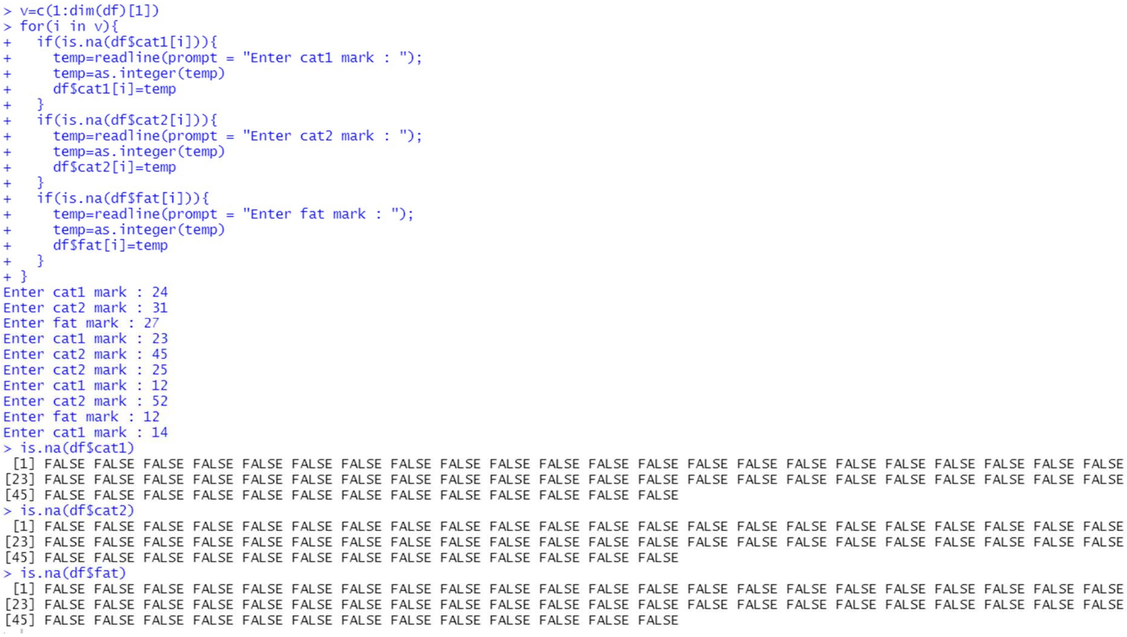
```
> df<-read.csv("Student_Data_Uncleaned.csv")
> class(df)
[1] "data.frame"
> str(df)
'data.frame':	57 obs. of  13 variables:
 $ ï..sno: int  1 2 3 4 5 6 7 8 9 10 ...
 $ regno : chr  "19BCE7478" "19BCN7017" "19BCN7045" "19BCN7050" ...
 $ name  : chr  "NITTOOR VISHNU BHARADWAJ" "INTURI REVANTH" "MUMMANI PURNAVENKTASAIKIRAN" "NISHIT VERMA      " ...
 $ school: chr  "CSE" "CSE" "CSE" "CSE" ...
 $ cat1  : int  43 10 26 NA 23 20 28 39 14 38 ...
 $ cat2  : num  38.5 12 37 47.5 46 31.5 41 45 17.5 40.5 ...
 $ da01  : int  19 19 19 19 19 NA 18 19 19 19 ...
 $ fat   : num  33 3 14 37 28 10 31 28 9.5 30.5 ...
 $ lab   : int  87 34 60 93 78 61 75 85 37 86 ...
 $ quiz1 : int  15 15 18 20 20 18 20 15 NA 18 ...
 $ gt    : logi  NA NA NA NA NA NA ...
 $ grade : logi  NA NA NA NA NA NA ...
 $ result: logi  NA NA NA NA NA NA ...
> summary(df)
     ï..sno       regno               name              school               cat1            cat2            da01             fat       
 Min.   : 1   Length:57          Length:57          Length:57          Min.   : 6.00   Min.   : 0.00   Min.   : 3.00   Min.   : 0.00  
 1st Qu.:15   Class :character   Class :character   Class :character   1st Qu.:18.00   1st Qu.:23.50   1st Qu.:18.00   1st Qu.:11.00  
 Median :29   Mode  :character   Mode  :character   Mode  :character   Median :25.00   Median :29.50   Median :19.00   Median :17.50  
 Mean   :29                                                            Mean   :25.85   Mean   :29.48   Mean   :18.09   Mean   :19.18  
 3rd Qu.:43                                                            3rd Qu.:34.00   3rd Qu.:36.50   3rd Qu.:19.00   3rd Qu.:28.00  
 Max.   :57                                                            Max.   :43.00   Max.   :47.50   Max.   :19.00   Max.   :40.00  
                                                                       NA's   :4       NA's   :4       NA's   :3       NA's   :2      
      lab            quiz1            gt            grade          result       
 Min.   :19.00   Min.   : 0.00   Mode:logical   Mode:logical   Mode:logical  
 1st Qu.:49.25   1st Qu.:15.00   NA's:57        NA's:57        NA's:57       
 Median :61.00   Median :18.00                                                
 Mean   :62.09   Mean   :16.06                                                
 3rd Qu.:76.50   3rd Qu.:20.00                                                
 Max.   :93.00   Max.   :20.00                                                
 NA's   :3       NA's   :4                                                    
> names(df)
 [1] "ï..sno" "regno"  "name"   "school" "cat1"   "cat2"   "da01"   "fat"    "lab"    "quiz1"  "gt"     "grade"  "result"
> dim(df)
[1] 57 13
> head(df)
  ï..sno     regno                        name school cat1 cat2 da01 fat lab quiz1 gt grade result
1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   43 38.5   19  33  87    15 NA    NA     NA
2      2 19BCN7017              INTURI REVANTH    CSE   10 12.0   19   3  34    15 NA    NA     NA
3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   26 37.0   19  14  60    18 NA    NA     NA
4      4 19BCN7050          NISHIT VERMA          CSE   NA 47.5   19  37  93    20 NA    NA     NA
5      5 19BCN7064       TADIBOINA ANAND KUMAR    ECE   23 46.0   19  28  78    20 NA    NA     NA
6      6 19BCN7079        PATTAPU SAI SRINIVAS    ECE   20 31.5   NA  10  61    18 NA    NA     NA
> tail(df)
   ï..sno     regno                      name school cat1 cat2 da01 fat lab quiz1 gt grade result
52     52 19BCE7034 MOOLA SAI UDAY KIRAN KUMAR    CSE    6 18.5   18   4  35    18 NA    NA     NA
53     53 19BCE7491      CYRIL AMBEDKAR KONDRU    ECE   32 29.5   18   5  49     8 NA    NA     NA
54     54 19BCI7098        PENIKALAPATI PRANAY    ECE   26 29.5   19  13  59    18 NA    NA     NA
55     55 19BCI7011          NISCHAL NANDIGAMA           37 37.5   18  NA  84    NA NA    NA     NA
56     56 19BEC7071       ANAND CHOUDARY JASTI    CSE   NA 23.5   19  13  39    13 NA    NA     NA
57     57 19BEC7141               KANDE ANISHA    CSE   18 24.5    3   6  28     5 NA    NA     NA
> |
```

2. Using repeat loop add 2 to all CAT1, CAT2 and FAT columns.



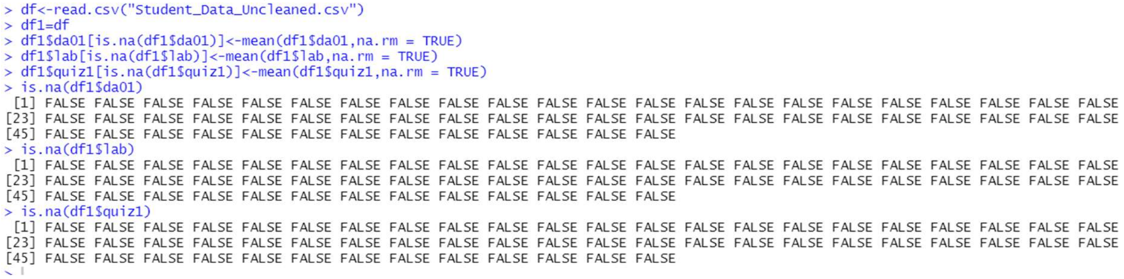
```
> i=1
> repeat{
+   if(!is.na(df$cat1[i])){
+     df$cat1[i]=df$cat1[i]+2
+   }
+   if(!is.na(df$cat2[i])){
+     df$cat2[i]=df$cat2[i]+2
+   }
+   if(!is.na(df$fat[i])){
+     df$fat[i]=df$fat[i]+2
+   }
+   if(i==dim(df)[1]){
+     break
+   }
+   i=i+1
+ }
> head(df)
  ï..sno     regno                        name school cat1 cat2 da01 fat lab quiz1 gt grade result
1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5   19  35  87    15 NA    NA     NA
2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0   19   5  34    15 NA    NA     NA
3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0   19  16  60    18 NA    NA     NA
4      4 19BCN7050                NISHIT VERMA    CSE   NA 49.5   19  39  93    20 NA    NA     NA
5      5 19BCN7064         TADIBOINA ANAND KUMAR    ECE   25 48.0   19  30  78    20 NA    NA     NA
6      6 19BCN7079          PATTAPU SAI SRINIVAS    ECE   22 33.5   NA  12  61    18 NA    NA     NA
```

3. Use for loop to get Not Available data from user for CAT1, CAT2and FAT and update it into CSV



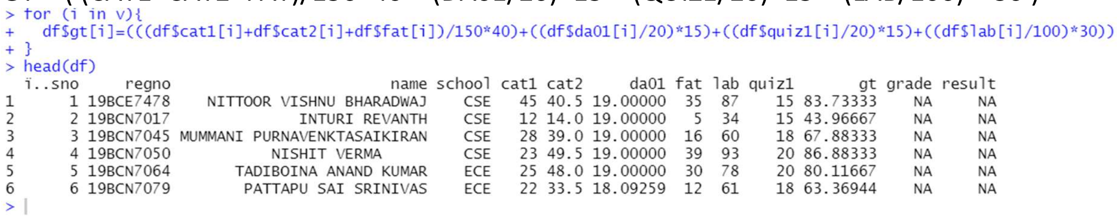
```
> v=c(1:dim(df)[1])
> for(i in v){
+   if(is.na(df$cat1[i])){
+     temp=readline(prompt = "Enter cat1 mark : ");
+     temp=as.integer(temp)
+     df$cat1[i]=temp
+   }
+   if(is.na(df$cat2[i])){
+     temp=readline(prompt = "Enter cat2 mark : ");
+     temp=as.integer(temp)
+     df$cat2[i]=temp
+   }
+   if(is.na(df$fat[i])){
+     temp=readline(prompt = "Enter fat mark : ");
+     temp=as.integer(temp)
+     df$fat[i]=temp
+   }
+ }
Enter cat1 mark : 24
Enter cat2 mark : 31
Enter fat mark : 27
Enter cat1 mark : 23
Enter cat2 mark : 45
Enter cat2 mark : 25
Enter cat1 mark : 12
Enter cat2 mark : 52
Enter fat mark : 12
Enter cat1 mark : 14
> is.na(df$cat1)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> is.na(df$cat2)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> is.na(df$fat)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

4. Use mean value to replace Not available data in DA01, QUIZ1 and LAB



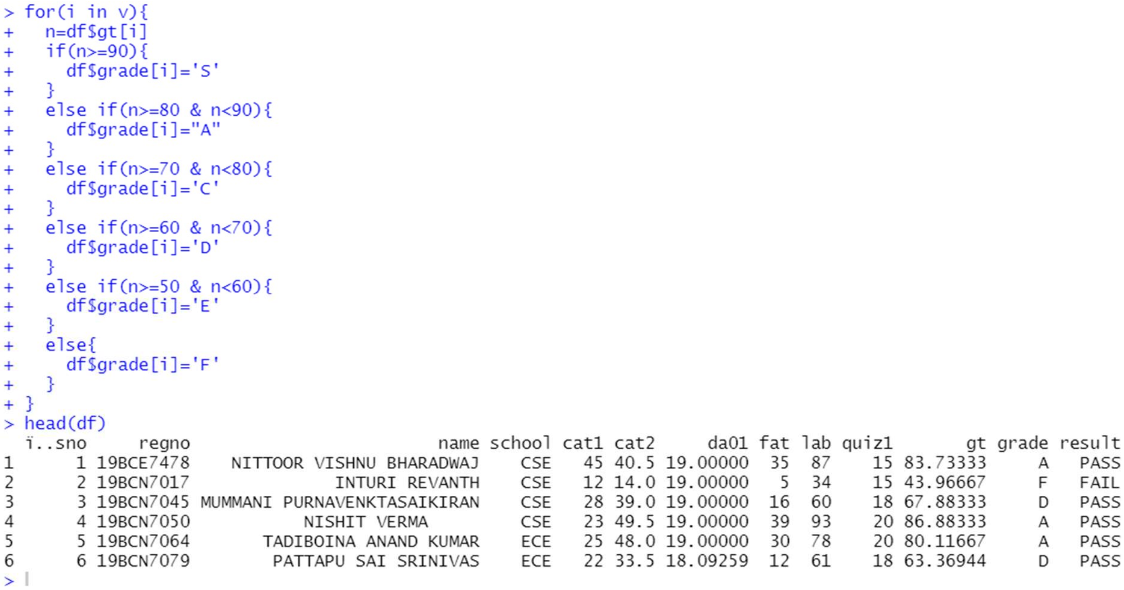
```
> df<-read.csv("Student_Data_Uncleaned.csv")
> df1=df
> df1$da01[is.na(df1$da01)]<-mean(df1$da01,na.rm = TRUE)
> df1$lab[is.na(df1$lab)]<-mean(df1$lab,na.rm = TRUE)
> df1$quiz1[is.na(df1$quiz1)]<-mean(df1$quiz1,na.rm = TRUE)
> is.na(df1$da01)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> is.na(df1$lab)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> is.na(df1$quiz1)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

5. Find the Grant Total (GT) for all students and update it into the CSV file.
   GT = ((CAT1+CAT2+FAT)/150*40 + (DA01/20)*15 + (QUIZ1/20)*15 + (LAB/100) * 30)



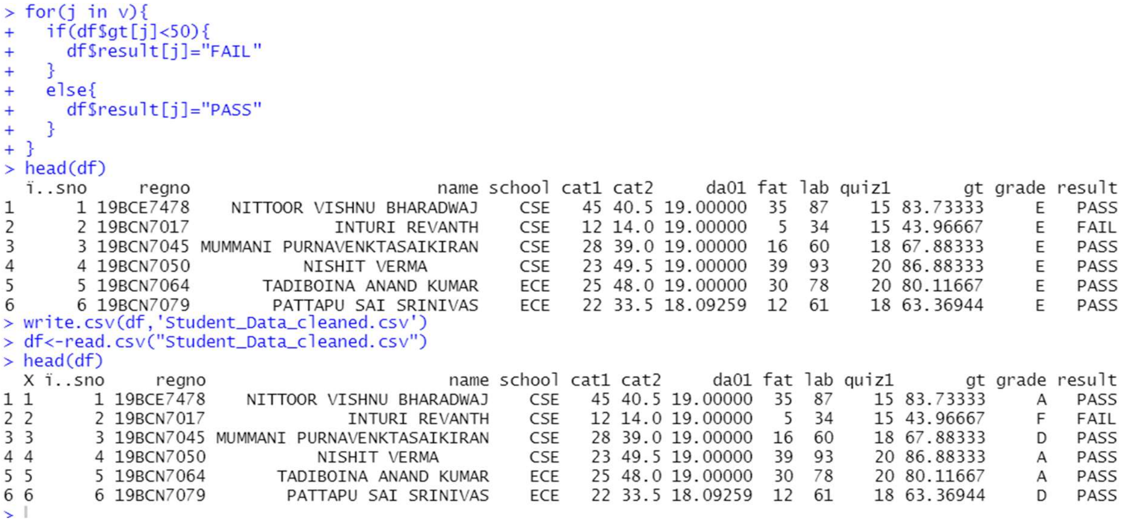
```
> for (i in v){
+   df$gt[i]=(((df$cat1[i]+df$cat2[i]+df$fat[i])/150*40)+((df$da01[i]/20)*15)+((df$quiz1[i]/20)*15)+((df$lab[i]/100)*30))
+ }
> head(df)
  ï..sno     regno                          name school cat1 cat2     da01 fat lab quiz1       gt grade result
1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333    NA     NA
2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667    NA     NA
3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333    NA     NA
4      4 19BCN7050                NISHIT VERMA    CSE   23 49.5 19.00000  39  93    20 86.88333    NA     NA
5      5 19BCN7064        TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667    NA     NA
6      6 19BCN7079         PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944    NA     NA
> 
```

6. Update the grade as per grade policy of our institution.



```
> for(i in v){
+   n=df$gt[i]
+   if(n>=90){
+     df$grade[i]='S'
+   }
+   else if(n>=80 & n<90){
+     df$grade[i]="A"
+   }
+   else if(n>=70 & n<80){
+     df$grade[i]='C'
+   }
+   else if(n>=60 & n<70){
+     df$grade[i]='D'
+   }
+   else if(n>=50 & n<60){
+     df$grade[i]='E'
+   }
+   else{
+     df$grade[i]='F'
+   }
+ }
> head(df)
  ï..sno     regno                          name school cat1 cat2     da01 fat lab quiz1       gt grade result
1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333     A   PASS
2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667     F   FAIL
3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333     D   PASS
4      4 19BCN7050                NISHIT VERMA    CSE   23 49.5 19.00000  39  93    20 86.88333     A   PASS
5      5 19BCN7064        TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667     A   PASS
6      6 19BCN7079         PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944     D   PASS
> 
```

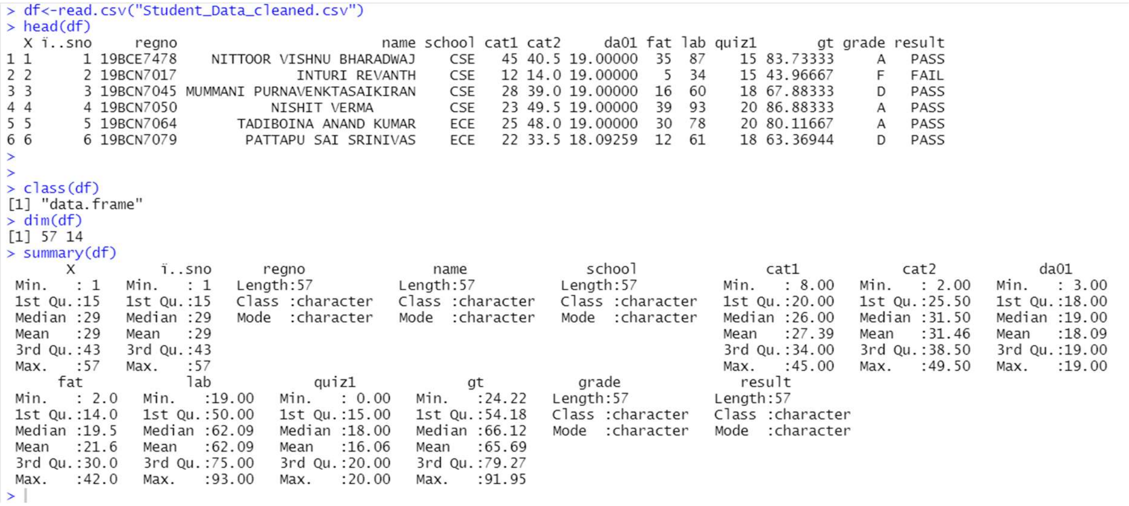
7. Update the result as “PASS” if their mark is greater than or equal to 50. Else result is “FAIL”



```
> for(j in v){
+   if(df$gt[j]<50){
+     df$result[j]="FAIL"
+   }
+   else{
+     df$result[j]="PASS"
+   }
+ }
> head(df)
  ï..sno     regno                          name school cat1 cat2     da01 fat lab quiz1       gt grade result
1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333     E   PASS
2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667     E   FAIL
3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333     E   PASS
4      4 19BCN7050                NISHIT VERMA    CSE   23 49.5 19.00000  39  93    20 86.88333     E   PASS
5      5 19BCN7064        TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667     E   PASS
6      6 19BCN7079         PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944     E   PASS
> write.csv(df,'Student_Data_cleaned.csv')
> df<-read.csv("Student_Data_cleaned.csv")
> head(df)
  X ï..sno     regno                          name school cat1 cat2     da01 fat lab quiz1       gt grade result
1 1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333     A   PASS
2 2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667     F   FAIL
3 3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333     D   PASS
4 4      4 19BCN7050                NISHIT VERMA    CSE   23 49.5 19.00000  39  93    20 86.88333     A   PASS
5 5      5 19BCN7064        TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667     A   PASS
6 6      6 19BCN7079         PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944     D   PASS
> 
```

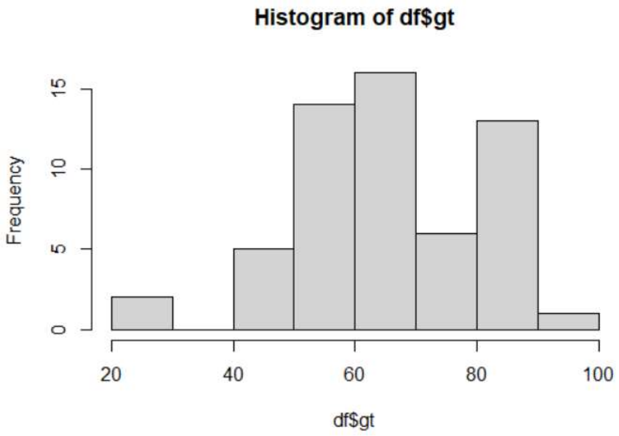
(Use Regression-Analysis-Data.csv)

1. Perform Exploratory Analysis

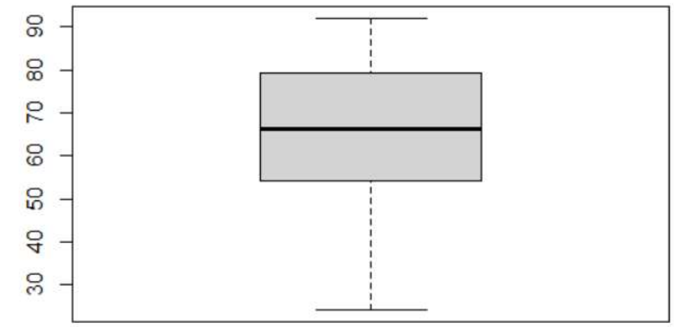


```
> df<-read.csv("Student_Data_cleaned.csv")
> head(df)
  X ï..sno     regno                        name school cat1 cat2     da01 fat lab quiz1       gt grade result
1 1      1 19BCE7478    NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333     A   PASS
2 2      2 19BCN7017              INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667     F   FAIL
3 3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333     D   PASS
4 4      4 19BCN7050            NISHIT VERMA        CSE   23 49.5 19.00000  39  93    20 86.88333     A   PASS
5 5      5 19BCN7064        TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667     A   PASS
6 6      6 19BCN7079         PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944     D   PASS
>
>
> class(df)
[1] "data.frame"
> dim(df)
[1] 57 14
> summary(df)
       X          ï..sno      regno               name              school               cat1            cat2            da01      
 Min.   : 1   Min.   : 1   Length:57          Length:57          Length:57          Min.   : 8.00   Min.   : 2.00   Min.   : 3.00  
 1st Qu.:15   1st Qu.:15   Class :character   Class :character   Class :character   1st Qu.:20.00   1st Qu.:25.50   1st Qu.:18.00  
 Median :29   Median :29   Mode  :character   Mode  :character   Mode  :character   Median :26.00   Median :31.50   Median :19.00  
 Mean   :29   Mean   :29                                                            Mean   :27.39   Mean   :31.46   Mean   :18.09  
 3rd Qu.:43   3rd Qu.:43                                                            3rd Qu.:34.00   3rd Qu.:38.50   3rd Qu.:19.00  
 Max.   :57   Max.   :57                                                            Max.   :45.00   Max.   :49.50   Max.   :19.00  
      fat            lab            quiz1             gt            grade              result         
 Min.   : 2.0   Min.   :19.00   Min.   : 0.00   Min.   :24.22   Length:57          Length:57         
 1st Qu.:14.0   1st Qu.:50.00   1st Qu.:15.00   1st Qu.:54.18   Class :character   Class :character  
 Median :19.5   Median :62.09   Median :18.00   Median :66.12   Mode  :character   Mode  :character  
 Mean   :21.6   Mean   :62.09   Mean   :16.06   Mean   :65.69                                        
 3rd Qu.:30.0   3rd Qu.:75.00   3rd Qu.:20.00   3rd Qu.:79.27                                        
 Max.   :42.0   Max.   :93.00   Max.   :20.00   Max.   :91.95                                        
> 
```

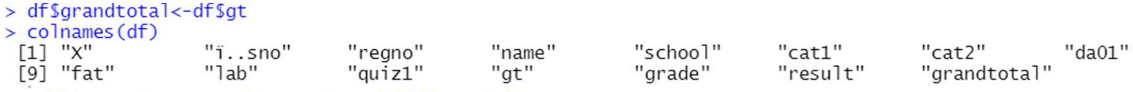
2. Perform visual Exploratory Analysis
   hist(df$gt)



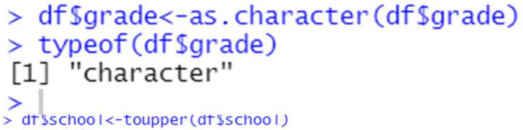
boxplot(df$gt)



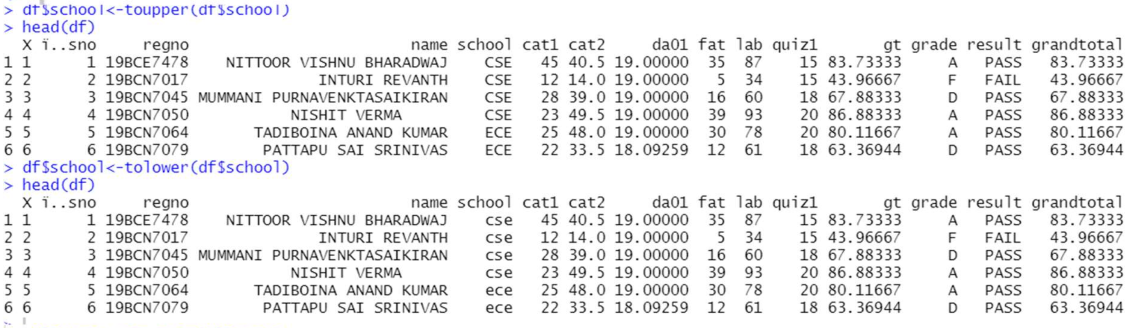
3. Perform Data cleaning operation and upload the corrected csv file (Practice all examples: https://dataanalyticsedge.com/2018/05/02/data-cleaning-using-r/)



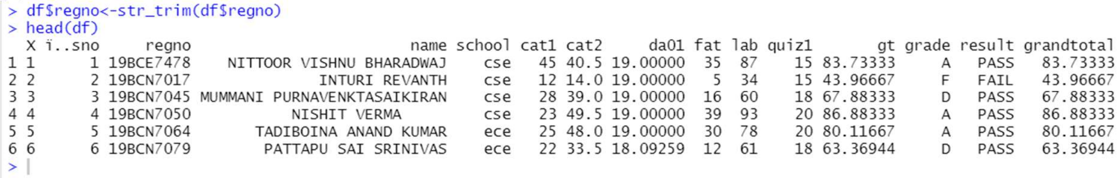
```
> df$grandtotal<-df$gt
> colnames(df)
 [1] "X"          "ï..sno"     "regno"      "name"       "school"     "cat1"       "cat2"       "da01"
 [9] "fat"        "lab"        "quiz1"      "gt"         "grade"      "result"     "grandtotal"
```



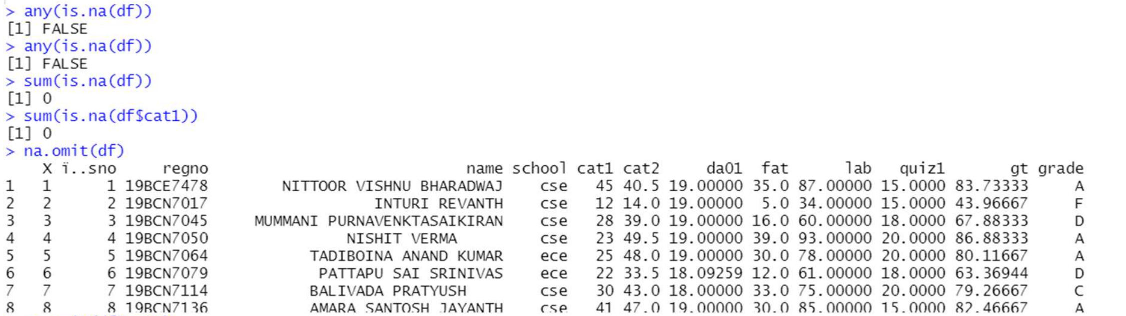
```
> df$grade<-as.character(df$grade)
> typeof(df$grade)
[1] "character"
>
```



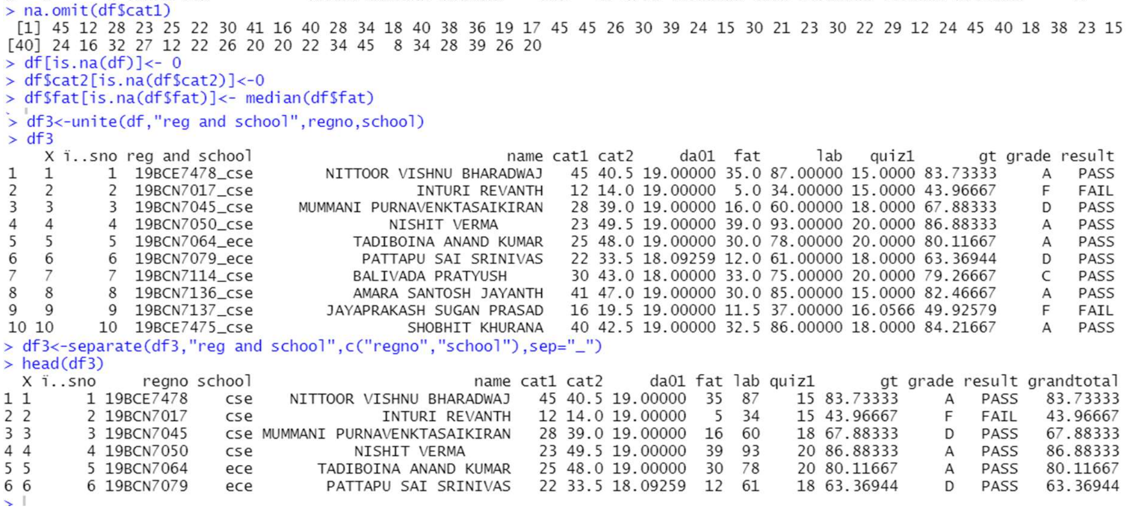
```
> df$school<-toupper(df$school)
> head(df)
  X ï..sno     regno                           name school cat1 cat2     da01 fat lab quiz1       gt grade result grandtotal
1 1      1 19BCE7478     NITTOOR VISHNU BHARADWAJ    CSE   45 40.5 19.00000  35  87    15 83.73333     A   PASS   83.73333
2 2      2 19BCN7017                INTURI REVANTH    CSE   12 14.0 19.00000   5  34    15 43.96667     F   FAIL   43.96667
3 3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    CSE   28 39.0 19.00000  16  60    18 67.88333     D   PASS   67.88333
4 4      4 19BCN7050              NISHIT VERMA      CSE   23 49.5 19.00000  39  93    20 86.88333     A   PASS   86.88333
5 5      5 19BCN7064       TADIBOINA ANAND KUMAR    ECE   25 48.0 19.00000  30  78    20 80.11667     A   PASS   80.11667
6 6      6 19BCN7079        PATTAPU SAI SRINIVAS    ECE   22 33.5 18.09259  12  61    18 63.36944     D   PASS   63.36944
> df$school<-tolower(df$school)
> head(df)
  X ï..sno     regno                           name school cat1 cat2     da01 fat lab quiz1       gt grade result grandtotal
1 1      1 19BCE7478     NITTOOR VISHNU BHARADWAJ    cse   45 40.5 19.00000  35  87    15 83.73333     A   PASS   83.73333
2 2      2 19BCN7017                INTURI REVANTH    cse   12 14.0 19.00000   5  34    15 43.96667     F   FAIL   43.96667
3 3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    cse   28 39.0 19.00000  16  60    18 67.88333     D   PASS   67.88333
4 4      4 19BCN7050              NISHIT VERMA      cse   23 49.5 19.00000  39  93    20 86.88333     A   PASS   86.88333
5 5      5 19BCN7064       TADIBOINA ANAND KUMAR    ece   25 48.0 19.00000  30  78    20 80.11667     A   PASS   80.11667
6 6      6 19BCN7079        PATTAPU SAI SRINIVAS    ece   22 33.5 18.09259  12  61    18 63.36944     D   PASS   63.36944
```



```
> df$regno<-str_trim(df$regno)
> head(df)
  X ï..sno     regno                           name school cat1 cat2     da01 fat lab quiz1       gt grade result grandtotal
1 1      1 19BCE7478     NITTOOR VISHNU BHARADWAJ    cse   45 40.5 19.00000  35  87    15 83.73333     A   PASS   83.73333
2 2      2 19BCN7017                INTURI REVANTH    cse   12 14.0 19.00000   5  34    15 43.96667     F   FAIL   43.96667
3 3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    cse   28 39.0 19.00000  16  60    18 67.88333     D   PASS   67.88333
4 4      4 19BCN7050              NISHIT VERMA      cse   23 49.5 19.00000  39  93    20 86.88333     A   PASS   86.88333
5 5      5 19BCN7064       TADIBOINA ANAND KUMAR    ece   25 48.0 19.00000  30  78    20 80.11667     A   PASS   80.11667
6 6      6 19BCN7079        PATTAPU SAI SRINIVAS    ece   22 33.5 18.09259  12  61    18 63.36944     D   PASS   63.36944
>
```



```
> any(is.na(df))
[1] FALSE
> any(is.na(df))
[1] FALSE
> sum(is.na(df))
[1] 0
> sum(is.na(df$cat1))
[1] 0
> na.omit(df)
   X ï..sno     regno                           name school cat1 cat2     da01  fat      lab   quiz1       gt grade
1  1      1 19BCE7478     NITTOOR VISHNU BHARADWAJ    cse   45 40.5 19.00000 35.0 87.00000 15.0000 83.73333     A
2  2      2 19BCN7017                INTURI REVANTH    cse   12 14.0 19.00000  5.0 34.00000 15.0000 43.96667     F
3  3      3 19BCN7045 MUMMANI PURNAVENKTASAIKIRAN    cse   28 39.0 19.00000 16.0 60.00000 18.0000 67.88333     D
4  4      4 19BCN7050              NISHIT VERMA      cse   23 49.5 19.00000 39.0 93.00000 20.0000 86.88333     A
5  5      5 19BCN7064       TADIBOINA ANAND KUMAR    ece   25 48.0 19.00000 30.0 78.00000 20.0000 80.11667     A
6  6      6 19BCN7079        PATTAPU SAI SRINIVAS    ece   22 33.5 18.09259 12.0 61.00000 18.0000 63.36944     D
7  7      7 19BCN7114       BALIVADA PRATYUSH        cse   30 43.0 18.00000 33.0 75.00000 20.0000 79.26667     C
8  8      8 19BCN7136       AMARA SANTOSH JAYANTH    cse   41 47.0 19.00000 30.0 85.00000 15.0000 82.46667     A
```



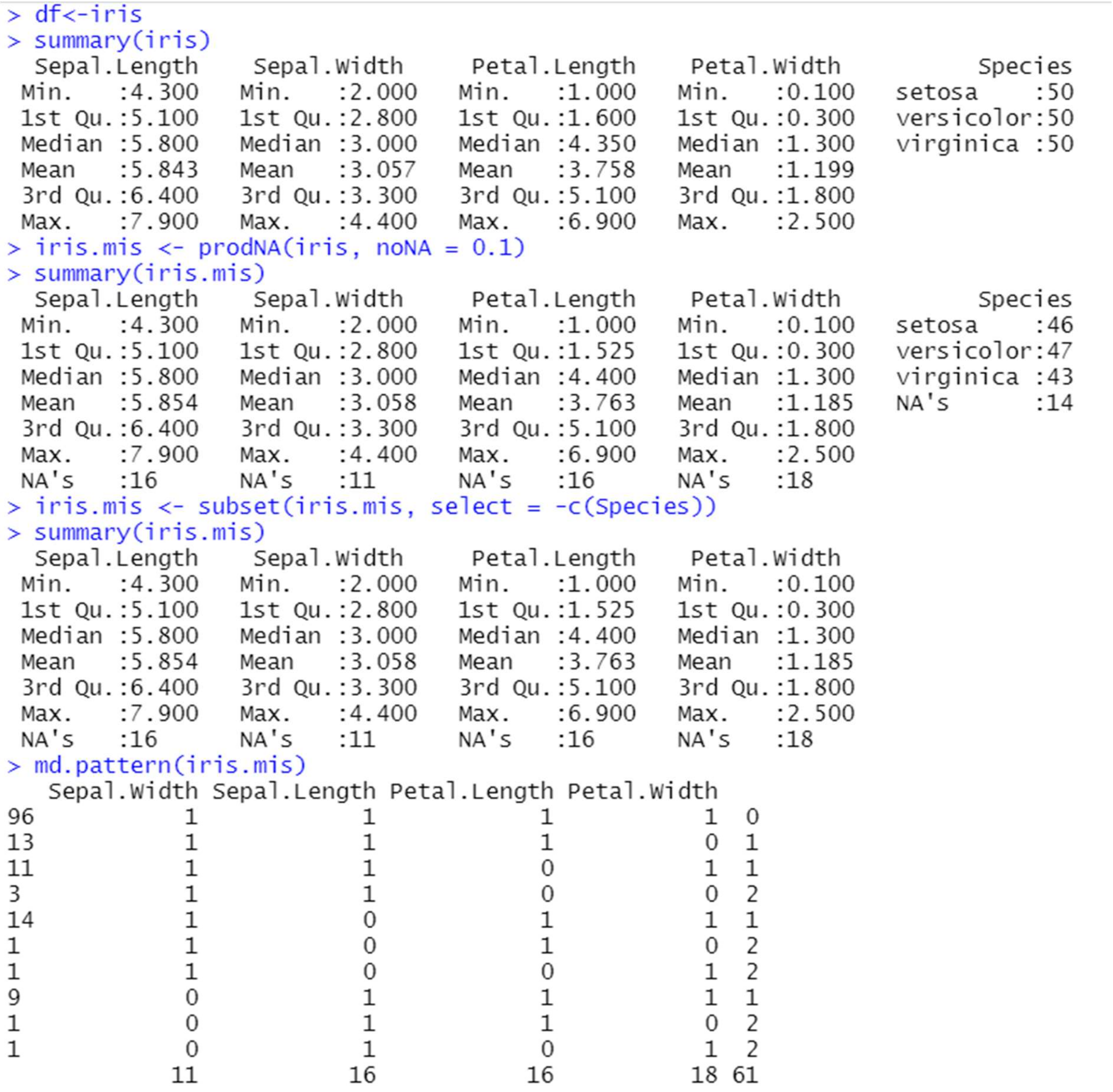
```
> na.omit(df$cat1)
 [1] 45 12 28 23 25 22 30 41 16 40 28 34 18 40 38 36 19 17 45 45 26 30 39 24 15 30 21 23 30 22 29 12 24 45 40 18 38 23 15
[40] 24 16 32 27 12 22 26 20 20 22 34 45  8 34 28 39 26 20
> df[is.na(df)]<- 0
> df$cat2[is.na(df$cat2)]<-0
> df$fat[is.na(df$fat)]<- median(df$fat)
> df3<-unite(df,"reg and school",regno,school)
> df3
    X ï..sno reg and school                           name cat1 cat2     da01  fat      lab   quiz1       gt grade result
1   1      1  19BCE7478_cse     NITTOOR VISHNU BHARADWAJ   45 40.5 19.00000 35.0 87.00000 15.0000 83.73333     A   PASS
2   2      2  19BCN7017_cse                INTURI REVANTH   12 14.0 19.00000  5.0 34.00000 15.0000 43.96667     F   FAIL
3   3      3  19BCN7045_cse MUMMANI PURNAVENKTASAIKIRAN   28 39.0 19.00000 16.0 60.00000 18.0000 67.88333     D   PASS
4   4      4  19BCN7050_cse              NISHIT VERMA     23 49.5 19.00000 39.0 93.00000 20.0000 86.88333     A   PASS
5   5      5  19BCN7064_ece       TADIBOINA ANAND KUMAR   25 48.0 19.00000 30.0 78.00000 20.0000 80.11667     A   PASS
6   6      6  19BCN7079_ece        PATTAPU SAI SRINIVAS   22 33.5 18.09259 12.0 61.00000 18.0000 63.36944     D   PASS
7   7      7  19BCN7114_cse       BALIVADA PRATYUSH       30 43.0 18.00000 33.0 75.00000 20.0000 79.26667     C   PASS
8   8      8  19BCN7136_cse       AMARA SANTOSH JAYANTH   41 47.0 19.00000 30.0 85.00000 15.0000 82.46667     A   PASS
9   9      9  19BCN7137_cse   JAYAPRAKASH SUGAN PRASAD   16 19.5 19.00000 11.5 37.00000 16.0566 49.92579     F   FAIL
10 10     10  19BCE7475_cse             SHOBHIT KHURANA   40 42.5 19.00000 32.5 86.00000 18.0000 84.21667     A   PASS
> df3<-separate(df3,"reg and school",c("regno","school"),sep="_")
> head(df3)
  X ï..sno     regno school                           name cat1 cat2     da01 fat lab quiz1       gt grade result grandtotal
1 1      1 19BCE7478    cse     NITTOOR VISHNU BHARADWAJ   45 40.5 19.00000  35  87    15 83.73333     A   PASS   83.73333
2 2      2 19BCN7017    cse                INTURI REVANTH   12 14.0 19.00000   5  34    15 43.96667     F   FAIL   43.96667
3 3      3 19BCN7045    cse MUMMANI PURNAVENKTASAIKIRAN   28 39.0 19.00000  16  60    18 67.88333     D   PASS   67.88333
4 4      4 19BCN7050    cse              NISHIT VERMA     23 49.5 19.00000  39  93    20 86.88333     A   PASS   86.88333
5 5      5 19BCN7064    ece       TADIBOINA ANAND KUMAR   25 48.0 19.00000  30  78    20 80.11667     A   PASS   80.11667
6 6      6 19BCN7079    ece        PATTAPU SAI SRINIVAS   22 33.5 18.09259  12  61    18 63.36944     D   PASS   63.36944
>
```

(Use iris.csv)

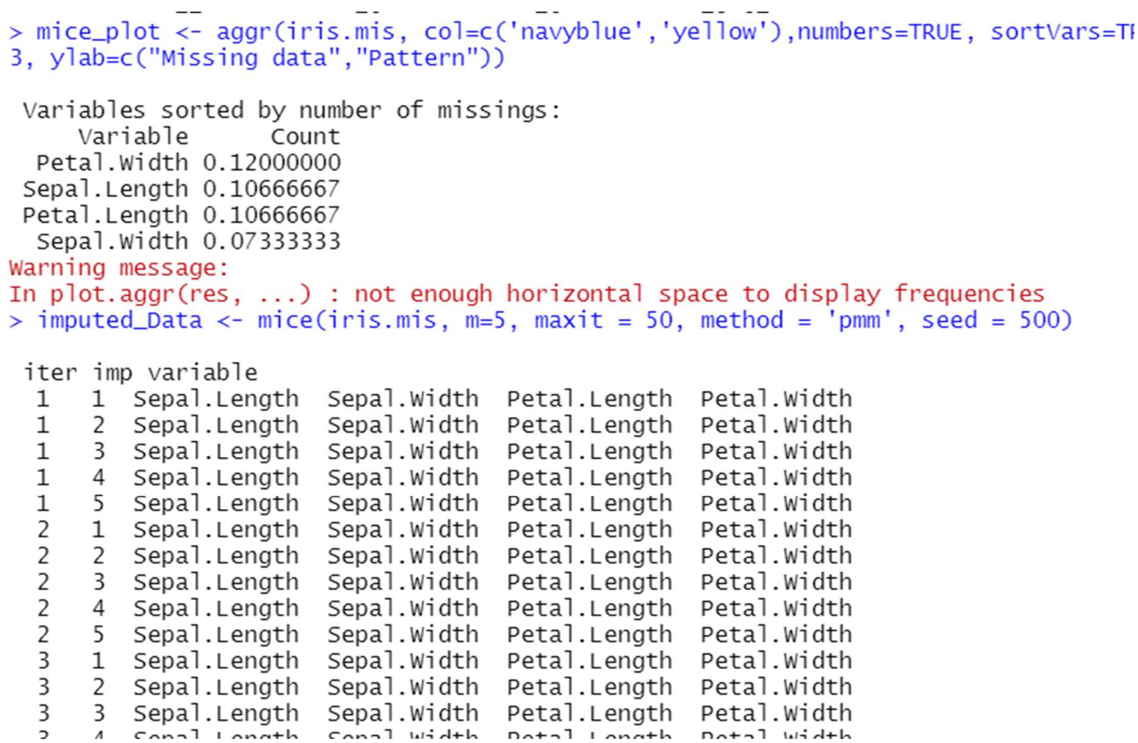
1. Perform Data Imputation using **Deletion**
2. Perform Data Imputation using **Mean/ Mode/ Median Imputation**
3. Perform Data Imputation using **Prediction Model**
4. Perform Data Imputation using **MICE Package**

   Ref: https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17

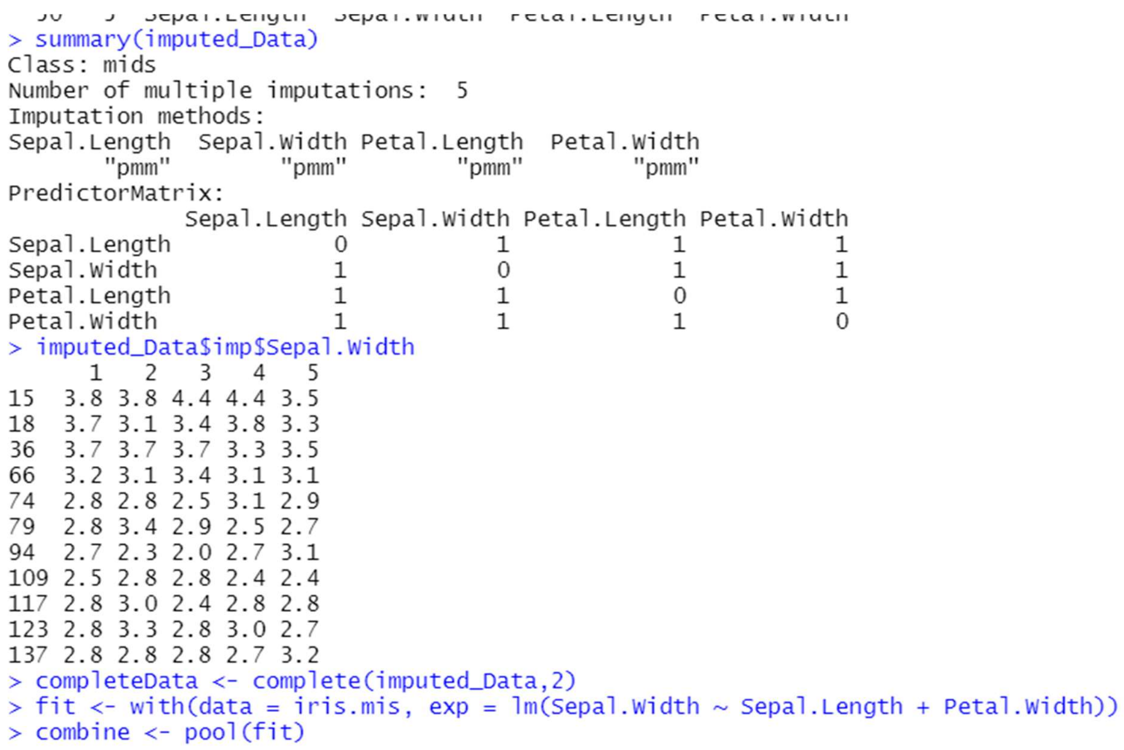
```
df<-iris

summary(iris)

iris.mis <- prodNA(iris, noNA = 0.1)

summary(iris.mis)

iris.mis <- subset(iris.mis, select = -c(Species))

summary(iris.mis)

md.pattern(iris.mis)

mice_plot <- aggr(iris.mis, col=c('navyblue','yellow'),numbers=TRUE,
sortVars=TRUE,labels=names(iris.mis), cex.axis=.7, gap=3, ylab=c("Missing data","Pattern"))

imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)

summary(imputed_Data)

imputed_Data$imp$Sepal.Width

completeData <- complete(imputed_Data,2)

fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

combine <- pool(fit)

summary(combine)
```



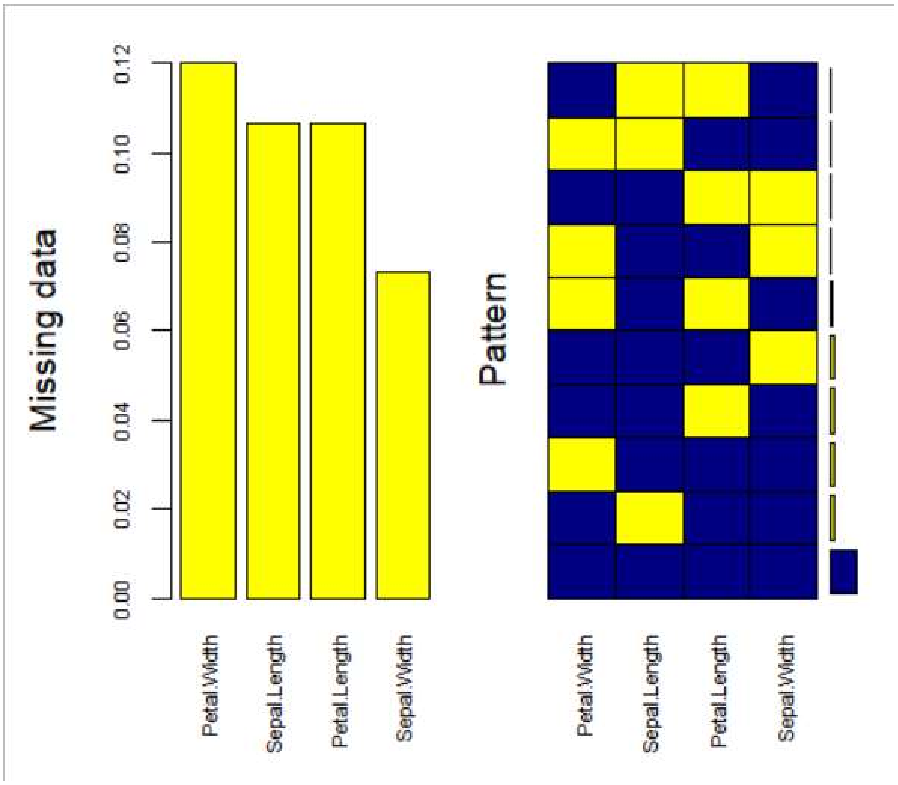
```
> df<-iris
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> iris.mis <- prodNA(iris, noNA = 0.1)
> summary(iris.mis)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :46
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.525   1st Qu.:0.300   versicolor:47
 Median :5.800   Median :3.000   Median :4.400   Median :1.300   virginica :43
 Mean   :5.854   Mean   :3.058   Mean   :3.763   Mean   :1.185   NA's      :14
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
 NA's   :16      NA's   :11      NA's   :16      NA's   :18
> iris.mis <- subset(iris.mis, select = -c(Species))
> summary(iris.mis)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.525   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.400   Median :1.300
 Mean   :5.854   Mean   :3.058   Mean   :3.763   Mean   :1.185
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
 NA's   :16      NA's   :11      NA's   :16      NA's   :18
> md.pattern(iris.mis)
   Sepal.Width Sepal.Length Petal.Length Petal.Width
96           1            1            1           1  0
13           1            1            1           0  1
11           1            1            0           1  1
3            1            1            0           0  2
14           1            0            1           1  1
1            1            0            1           0  2
1            1            0            0           1  2
9            0            1            1           1  1
1            0            1            1           0  2
1            0            1            0           1  2
            11           16           16          18 61
```



```
> mice_plot <- aggr(iris.mis, col=c('navyblue','yellow'),numbers=TRUE, sortVars=T
3, ylab=c("Missing data","Pattern"))

 Variables sorted by number of missings:
     Variable      Count
  Petal.Width 0.12000000
 Sepal.Length 0.10666667
 Petal.Length 0.10666667
  Sepal.Width 0.07333333
Warning message:
In plot.aggr(res, ...) : not enough horizontal space to display frequencies
> imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)

 iter imp variable
  1   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
```



```
  50   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
> summary(imputed_Data)
Class: mids
Number of multiple imputations:  5 
Imputation methods:
Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
       "pmm"        "pmm"        "pmm"        "pmm" 
PredictorMatrix:
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length            0           1            1           1
Sepal.Width             1           0            1           1
Petal.Length            1           1            0           1
Petal.Width             1           1            1           0
> imputed_Data$imp$Sepal.Width
      1   2   3   4   5
15  3.8 3.8 4.4 4.4 3.5
18  3.7 3.1 3.4 3.8 3.3
36  3.7 3.7 3.7 3.3 3.5
66  3.2 3.1 3.4 3.1 3.1
74  2.8 2.8 2.5 3.1 2.9
79  2.8 3.4 2.9 2.5 2.7
94  2.7 2.3 2.0 2.7 3.1
109 2.5 2.8 2.8 2.4 2.4
117 2.8 3.0 2.4 2.8 2.8
123 2.8 3.3 2.8 3.0 2.7
137 2.8 2.8 2.8 2.7 3.2
> completeData <- complete(imputed_Data,2)
> fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))
> combine <- pool(fit)
```


Missing data
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Petal.Width
Sepal.Length
Petal.Length
Sepal.Width
Pattern
Petal.Width
Sepal.Length
Petal.Length
Sepal.Width