**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

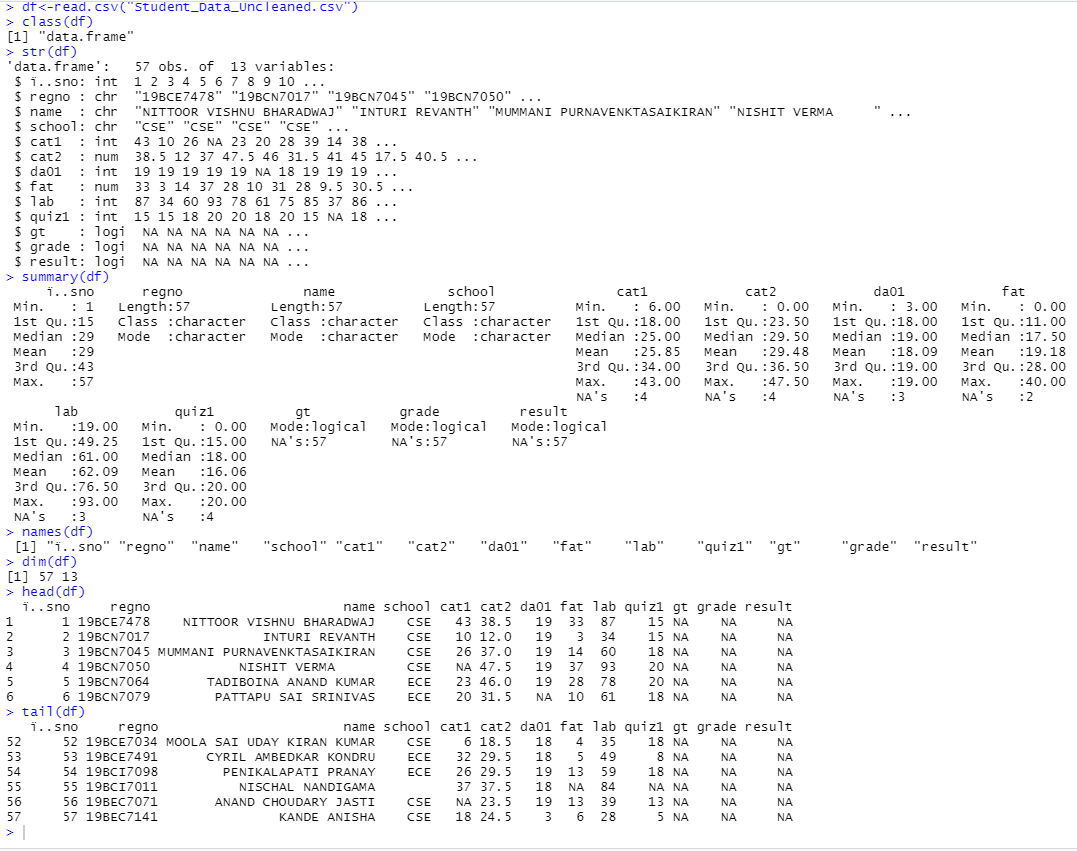
**Student name:** Valiveti Manikanta bhuvanesh **Reg. no.: 19BCD7088**

LAB 4 (Data Cleaning and Imputation)

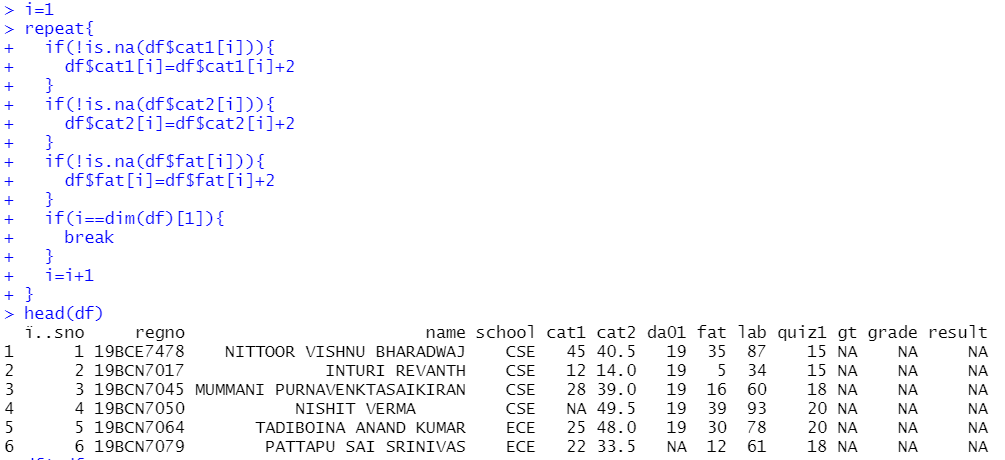
Questions:

(Use Student\_Data.csv)

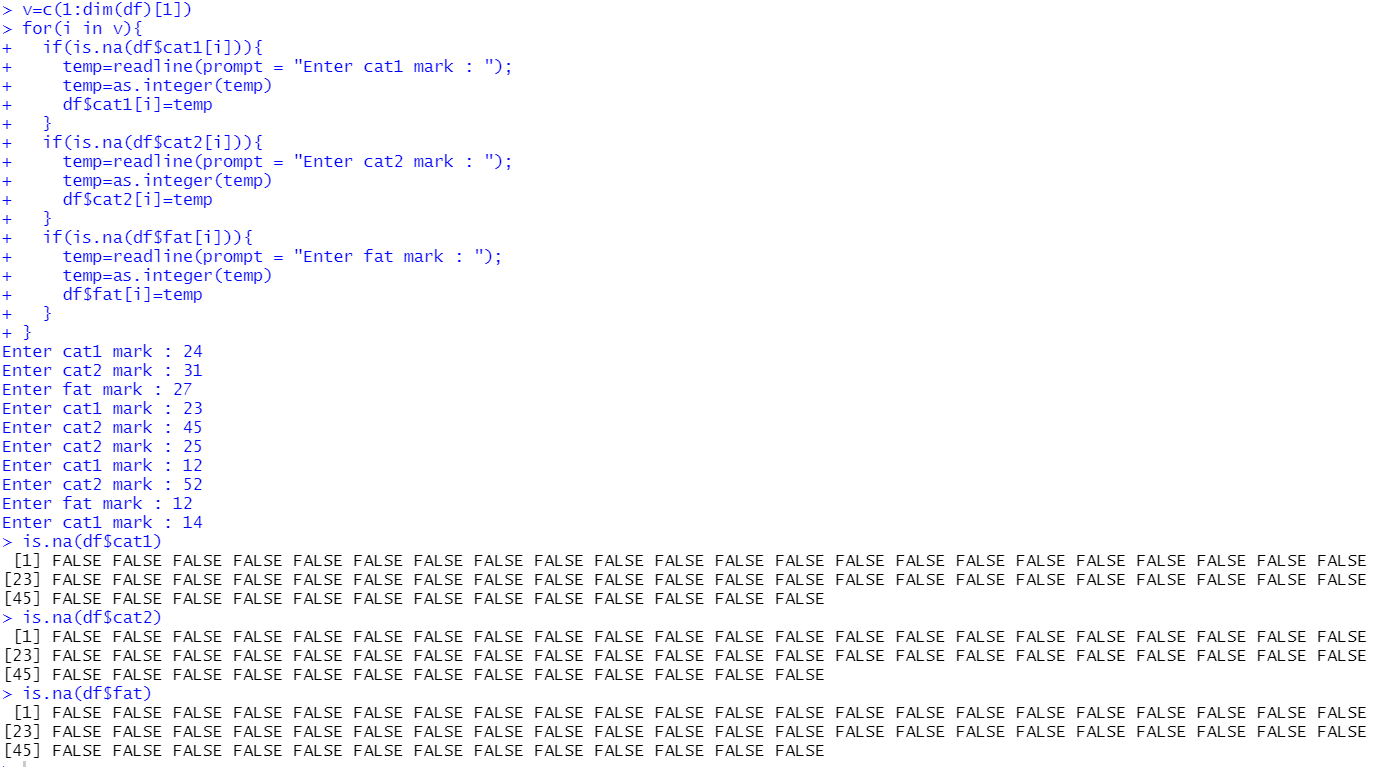
1. Do preliminary observations. (head, str…)



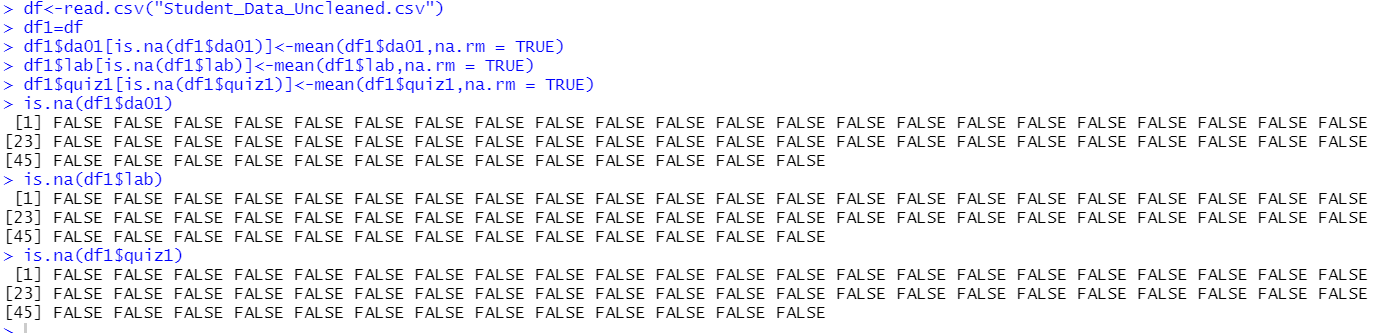
1. Using repeat loop add 2 to all CAT1, CAT2 and FAT columns.



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

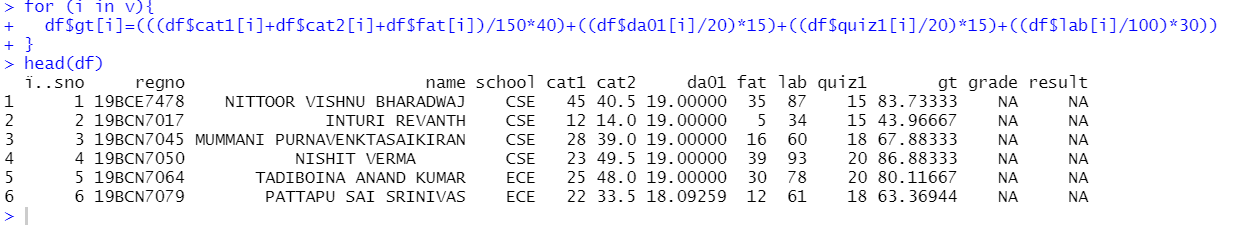


1. Use mean value to replace Not available data in DA01, QUIZ1 and LAB

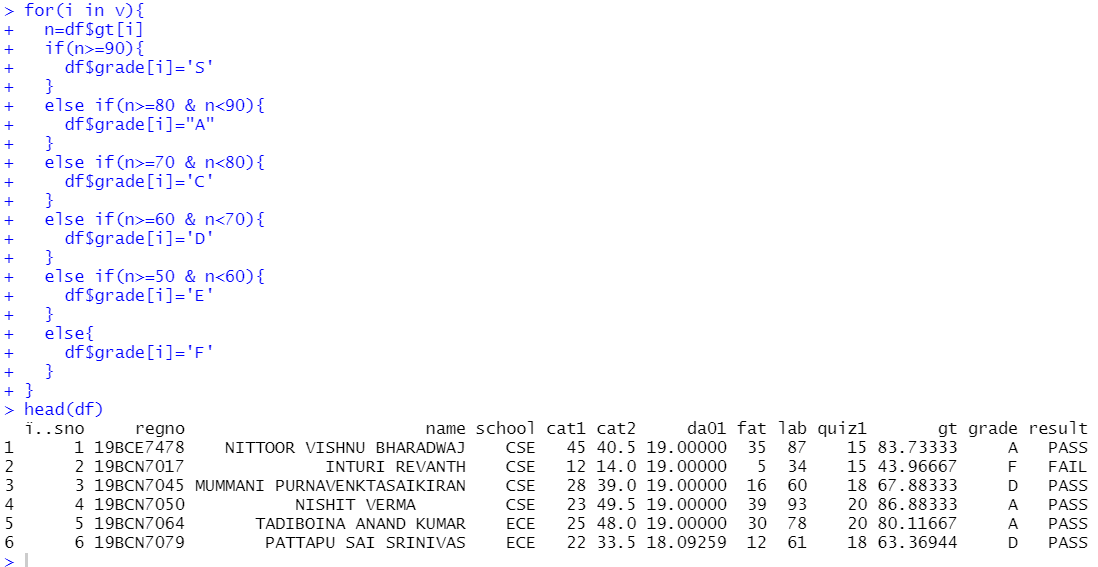


1. Find the Grant Total (GT) for all students and update it into the CSV file.

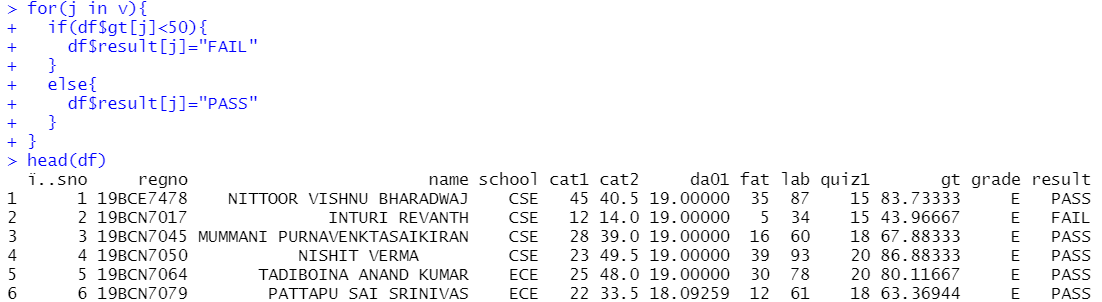
GT = ( (CAT1+CAT2+FAT)/150\*40 + (DA01/20)\*15 + (QUIZ1/20)\*15 + (LAB/100) \* 30 )

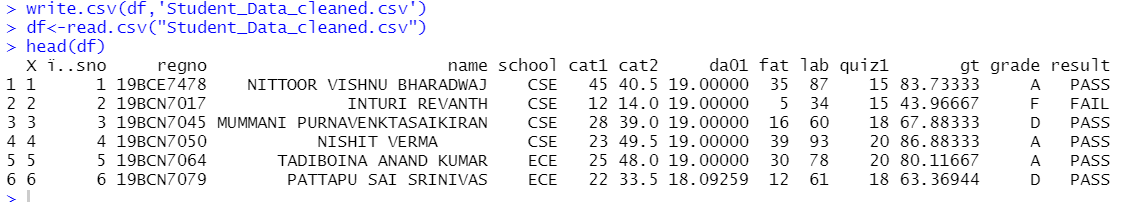


1. Update the grade as per grade policy of our institution.



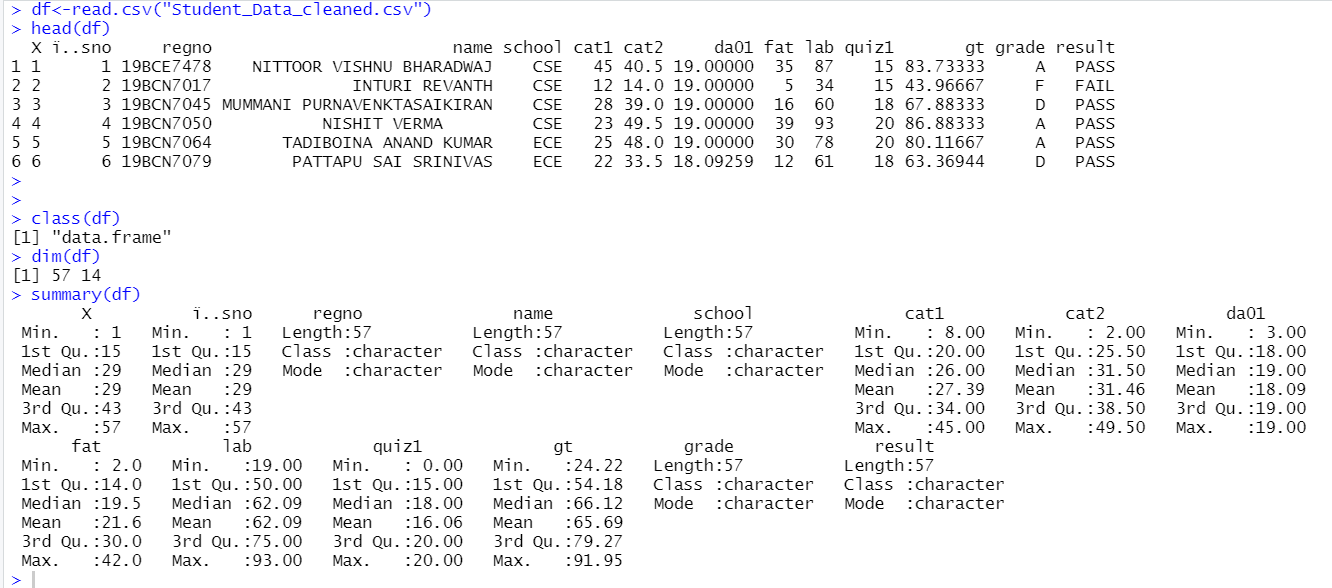
1. Update the result as “PASS” if their mark is greater than or equal to 50. Else result is “FAIL”





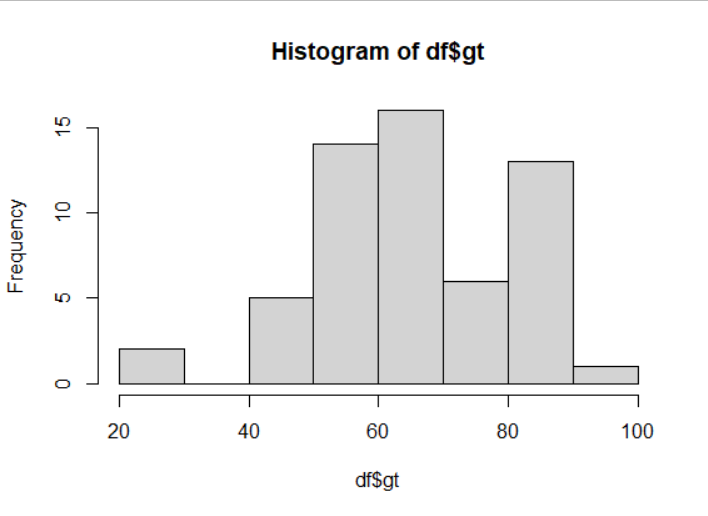
(Use Regression-Analysis-Data.csv)

1. Perform Exploratory Analysis

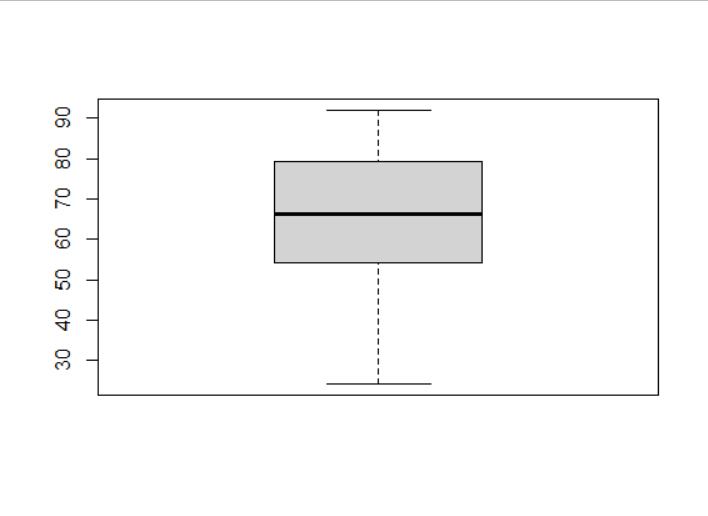


1. Perform visual Exploratory Analysis

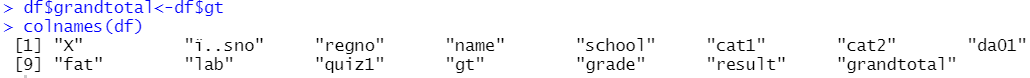
hist(df$gt)

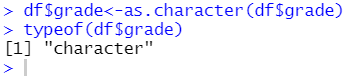


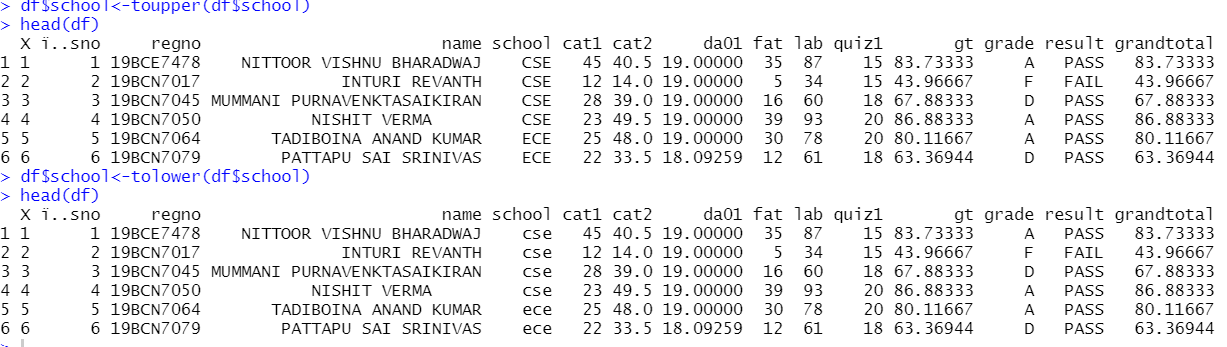
boxplot(df$gt)

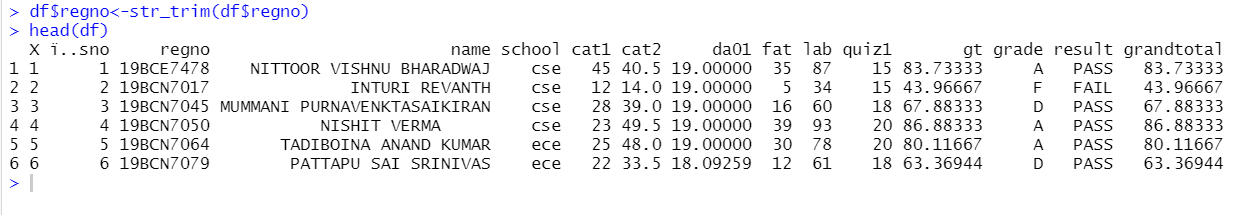


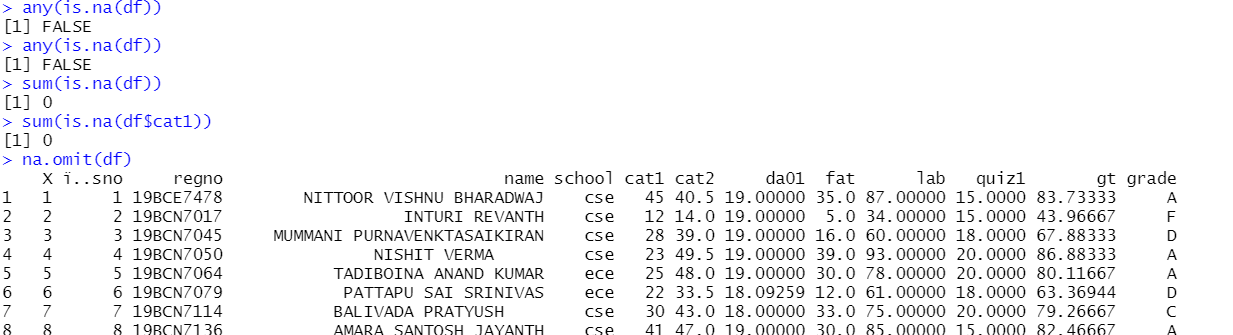
1. Perform Data cleaning operation and upload the corrected csv file (Practice all examples: <https://dataanalyticsedge.com/2018/05/02/data-cleaning-using-r/> )

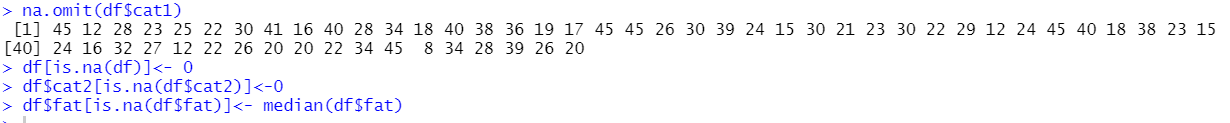


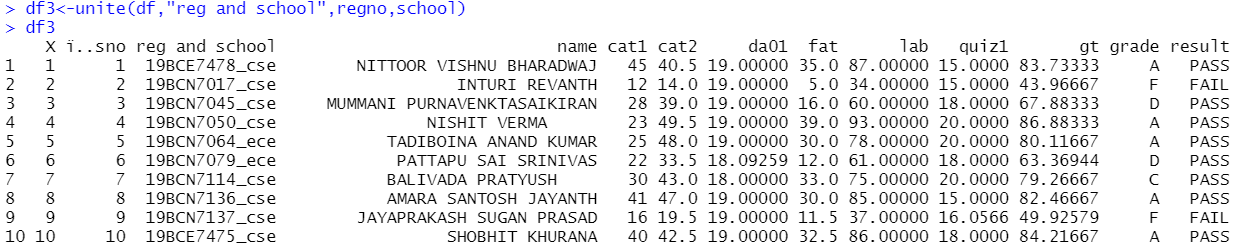


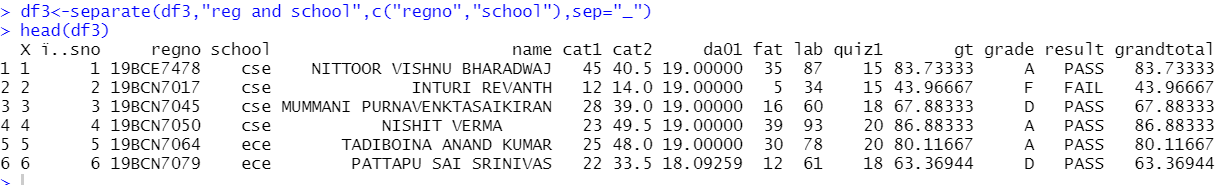












(Use iris.csv)

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)

