* 1. Pick three attributes whose data types may present problems due to missing values, implausible values, or duplicate values during later analysis and transform them to numeric or nominal format.  How much data is corrupted or missing? (2 points)

In the Group Data Set , We try to identify if there are any missing values in the student alcohol distribution, and we identify there are no missing values, and there are no duplication of values. However, there said to be some nominal values where we managed to make it with numeric values, which we considered to identify the summary of all the fields, and find there is no duplication of the content provide. As a matter of fact , in nut shell , trusting the data completely is wrong, so we ideloized some points with the results which we found in this, and we found some in consistency in the data, and where we try to make over or fill the data using the mean and mode value to get the accurate results which help the person to take a decision.   
  
R Code:   
  
d1=read.table("Student-por.csv",sep=";",header=TRUE)

anyNA(d1)

duplicated(student\_por)

summary(student\_por)  
  
Output:  
  
> summary(student\_por)

school sex age address famsize

Length:649 Length:649 Min. :15.00 Length:649 Length:649

Class :character Class :character 1st Qu.:16.00 Class :character Class :character

Mode :character Mode :character Median :17.00 Mode :character Mode :character

Mean :16.74

3rd Qu.:18.00

Max. :22.00

Pstatus Medu Fedu Mjob Fjob

Length:649 Min. :0.000 Min. :0.000 Length:649 Length:649

Class :character 1st Qu.:2.000 1st Qu.:1.000 Class :character Class :character

Mode :character Median :2.000 Median :2.000 Mode :character Mode :character

Mean :2.515 Mean :2.307

3rd Qu.:4.000 3rd Qu.:3.000

Max. :4.000 Max. :4.000

reason guardian traveltime studytime failures

Length:649 Length:649 Min. :1.000 Min. :1.000 Min. :0.0000

Class :character Class :character 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:0.0000

Mode :character Mode :character Median :1.000 Median :2.000 Median :0.0000

Mean :1.569 Mean :1.931 Mean :0.2219

3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:0.0000

Max. :4.000 Max. :4.000 Max. :3.0000

schoolsup famsup paid activities nursery

Length:649 Length:649 Length:649 Length:649 Length:649

Class :character Class :character Class :character Class :character Class :character

Mode :character Mode :character Mode :character Mode :character Mode :character

higher internet romantic famrel freetime

Length:649 Length:649 Length:649 Min. :1.000 Min. :1.00

Class :character Class :character Class :character 1st Qu.:4.000 1st Qu.:3.00

Mode :character Mode :character Mode :character Median :4.000 Median :3.00

Mean :3.931 Mean :3.18

3rd Qu.:5.000 3rd Qu.:4.00

Max. :5.000 Max. :5.00

goout Dalc Walc health absences G1

Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000 Min. : 0.000 Min. : 0.0

1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:2.000 1st Qu.: 0.000 1st Qu.:10.0

Median :3.000 Median :1.000 Median :2.00 Median :4.000 Median : 2.000 Median :11.0

Mean :3.185 Mean :1.502 Mean :2.28 Mean :3.536 Mean : 3.659 Mean :11.4

3rd Qu.:4.000 3rd Qu.:2.000 3rd Qu.:3.00 3rd Qu.:5.000 3rd Qu.: 6.000 3rd Qu.:13.0

Max. :5.000 Max. :5.000 Max. :5.00 Max. :5.000 Max. :32.000 Max. :19.0

G2 G3

Min. : 0.00 Min. : 0.00

1st Qu.:10.00 1st Qu.:10.00

Median :11.00 Median :12.00

Mean :11.57 Mean :11.91

3rd Qu.:13.00 3rd Qu.:14.00

Max. :19.00 Max. :19.00

By this results, we inferred this that we can bin the results to categories, to make a best understanding about of data, and make them meaning ful even if the values are missing, so that it becomes easy to make over with the data. Taking three attributes which considered I nominal value so we moved the dat to numeric to get a meaning results which is easy in understanding, and the code is been added below.  
  
  
R Code:   
is.integer(student\_por)

is.null(student\_por)

is.character(student\_por)

is.numeric(student\_por)

is.factor(student\_por)  
student\_por$internet=as.numeric(student\_por$internet)

is.numeric(student\_por$internet)

student\_por$sex=as.numeric(student\_por$sex)

is.numeric(student\_por$sex)

student\_por$Pstatus=as.numeric(student\_por$Pstatus)

is.numeric(student\_por$Pstatus)  
  
Output :   
  
> is.integer(student\_por)

[1] FALSE

> is.null(student\_por)

[1] FALSE

> is.character(student\_por)

[1] FALSE

> is.numeric(student\_por)

[1] FALSE

> is.factor(student\_por)

[1] FALSE

> student\_por$internet=as.numeric(student\_por$internet)

Warning message:

NAs introduced by coercion

> is.numeric(student\_por$internet)

[1] TRUE

> student\_por$sex=as.numeric(student\_por$sex)

Warning message:

NAs introduced by coercion

> is.numeric(student\_por$sex)

[1] TRUE

> student\_por$Pstatus=as.numeric(student\_por$Pstatus)

Warning message:

NAs introduced by coercion

> is.numeric(student\_por$Pstatus)

[1] TRUE

In order to identify the data is ranging in as per expected value , we also identify by selecting couple of random attributes and we identify that all values are ranging as per expected.   
  
  
R Code:  
  
range(student\_por$school)

range(student\_por$sex)

range(student\_por$age)

range(student\_por$address)

range(student\_por$famsize)

range(student\_por$Pstatus)

range(student\_por$Medu)  
  
Output:   
  
> range(student\_por$school)

[1] "GP" "MS"

> range(student\_por$sex)

[1] NA NA

> range(student\_por$age)

[1] 15 22

> range(student\_por$address)

[1] "R" "U"

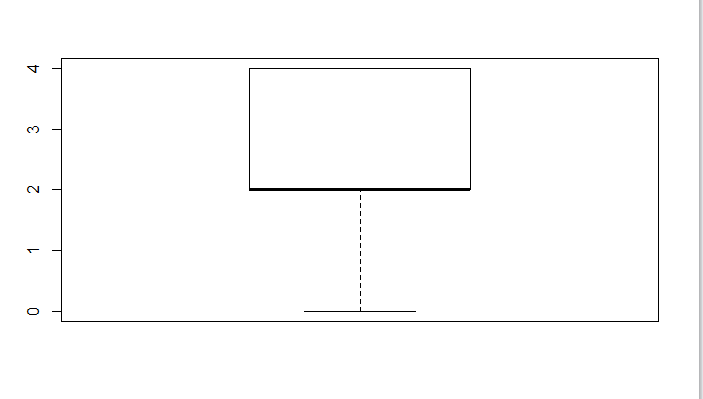
> range(student\_por$famsize)

[1] "GT3" "LE3"

1. range(student\_por$Medu)
2. [1] 0 4

2. If you have missing values, implausible values, or duplicate values, fill them in.  Explain your decisions and your strategy and show your code. (3 points)

As per the Data Set been idolized, we came to a confirmation that there is no missing values in the data field, neither there are no duplication as mentioned in above code and output. If there is any kind of data missing and for that we would able to fill using, then we would find the median wit removing missing value.   
  
Ex: Age   
  
median(Student\_por$age, na.rm=T)  
Then replace with missing value with median value sometimes with mean or mode as well  
  
student\_por$age=ifelse(is.na(student\_por$age),17,student\_por$age)   
  
  
R Code:  
> boxplot(student\_por$Medu)

duplicated(student\_por) > duplicated(student\_por)  
  
  
Ouput:   
  


[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[17] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[33] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[65] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[81] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[113] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[129] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[161] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[209] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[225] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[257] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[273] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[305] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[321] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[353] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[369] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[401] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[417] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[449] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[465] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[497] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[513] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[545] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[561] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[577] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[593] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[609] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

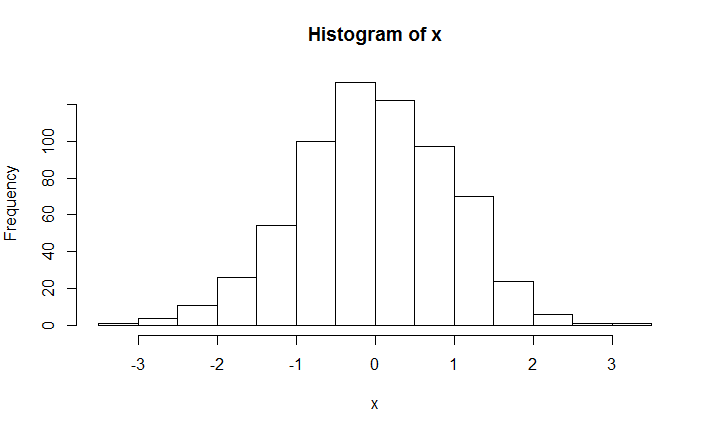
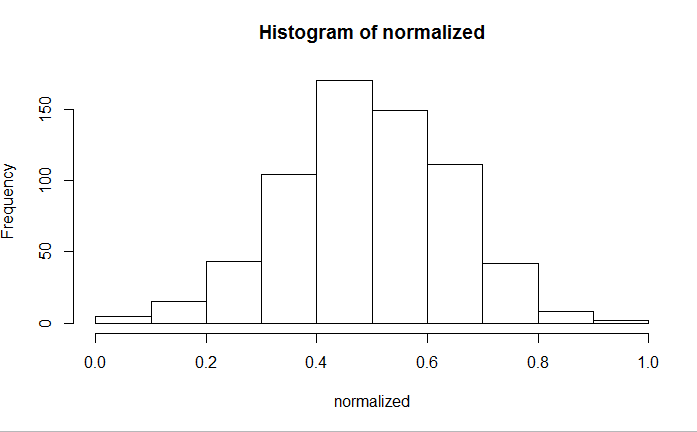
[625] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[641] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

3.Normalize the numeric attributes you have selected.  What benefit will this provide for your analysis? (2 points)

Normalize the attributes would always help to identify and analyze the content in what plot or in what region in what in , and what best make can be made out of it; how ever we identified the normalization of numeric attributes will help in identifying the nearest plot and make management to make the decision on required results and make best for the organization. It helps in what best region and range it places in.   
  
R Code:

|  |
| --- |
| > x = rnorm(student\_por$G1)  > normalized = (x-min(x))/(max(x)-min(x))  > hist(x)  > hist(normalized) |
|  |
| |  | | --- | | > | |

  
  
  
  
4.Now that you have a better dataset, list three (3) important organizational analytical questions that management could ask about the dataset in order to help them make business or organizational decisions. (3 points)  
  
Management would always to look at the data in all directions which would help the organization to gain best benefits from this, and for this , if we try to bin the data, and make it in a categorical format it enhances for management to make the best out of it. For this DataSet, we found that there are couple of categories which can be made out of it. Mainly we identified what management would necessary look , so for that we categorized into two parts for some of the data, as one is family, where student and family would be taken in consideration, and other one is school regarding absences etc., would be taken , so that big long data would come short and where binning of data would make management to make best decision for organization benefits.

Include your R code (or Excel screenshots) with the results of your transformation, normalization, and replacement attempts.