se = read.csv("2016 School Explorer.csv", header = TRUE)

#Convert 'N/A' strings to NA

se[ se == "N/A" ] = NA

#Remove dollar signs

se$School.Income.Estimate = as.numeric(gsub('[$,]', '', se$School.Income.Estimate))

#Removing Percent symbols

percent\_cols =dplyr::select(se,Percent.Asian,Percent.Black,Percent.Black...Hispanic,Percent.ELL,Percent.Hispanic,Percent.of.Students.Chronically.Absent,Percent.White,Supportive.Environment..,Rigorous.Instruction..,Collaborative.Teachers..,Effective.School.Leadership..,Trust..,Strong.Family.Community.Ties..,Student.Attendance.Rate)

percent\_cols = apply(percent\_cols, 2, function(y) as.numeric(gsub("%", "", y)))

se$Percent.Asian = percent\_cols[, 1]

se$Percent.Black = percent\_cols[, 2]

se$Percent.Black...Hispanic = percent\_cols[, 3]

se$Percent.ELL = percent\_cols[, 4]

se$Percent.Hispanic = percent\_cols[, 5]

se$Percent.of.Students.Chronically.Absent = percent\_cols[, 6]

se$Percent.White = percent\_cols[, 7]

se$Supportive.Environment.. = percent\_cols[, 8]

se$Rigorous.Instruction.. = percent\_cols[, 9]

se$Collaborative.Teachers.. = percent\_cols[, 10]

se$Effective.School.Leadership.. = percent\_cols[, 11]

se$Trust.. = percent\_cols[, 12]

se$Strong.Family.Community.Ties.. = percent\_cols[, 13]

se$Student.Attendance.Rate = percent\_cols[, 14]

# after doing a colSums(is.na(se)), I found that there were over 100 rows with missing values across the following 4 columns, so I removed them.

se = se[!is.na(se$Rigorous.Instruction.Rating) ,]

se = se[!is.na(se$Student.Achievement.Rating) ,]

se = se[!is.na(se$Trust.Rating) ,]

se = se[!is.na(se$Supportive.Environment.Rating) ,]

# Here I handle all na values in School income by assigning them the average income of their corresponding district.

Districts = se$District

means = aggregate(.~Districts, data=se, mean)

income\_means = means$School.Income.Estimate

income = se$School.Income.Estimate

d = se$District

for (x in 1:length(income))

if (is.na(income[x]))

income[x] = income\_means[d[x]]

se$School.Income.Estimate = income

#converting decimal data from <fctr> to <dbl>

se$Economic.Need.Index = as.numeric(as.character(se$Economic.Need.Index))

se$Average.ELA.Proficiency = as.numeric(as.character(se$Average.ELA.Proficiency))

se$Average.Math.Proficiency = as.numeric(as.character(se$Average.Math.Proficiency))

#Creating dummy variables for categorical data

se$Community.School.= as.numeric(se$Community.School.)

se$Rigorous.Instruction.Rating = as.numeric(se$Rigorous.Instruction.Rating)

se$Collaborative.Teachers.Rating = as.numeric(se$Collaborative.Teachers.Rating)

se$Supportive.Environment.Rating = as.numeric(se$Supportive.Environment.Rating)

se$Effective.School.Leadership.Rating = as.numeric(se$Effective.School.Leadership.Rating)

se$Strong.Family.Community.Ties.Rating = as.numeric(se$Strong.Family.Community.Ties.Rating)

se$Trust.Rating = as.numeric(se$Trust.Rating)

se$Student.Achievement.Rating = as.numeric(se$Student.Achievement.Rating)