MICE

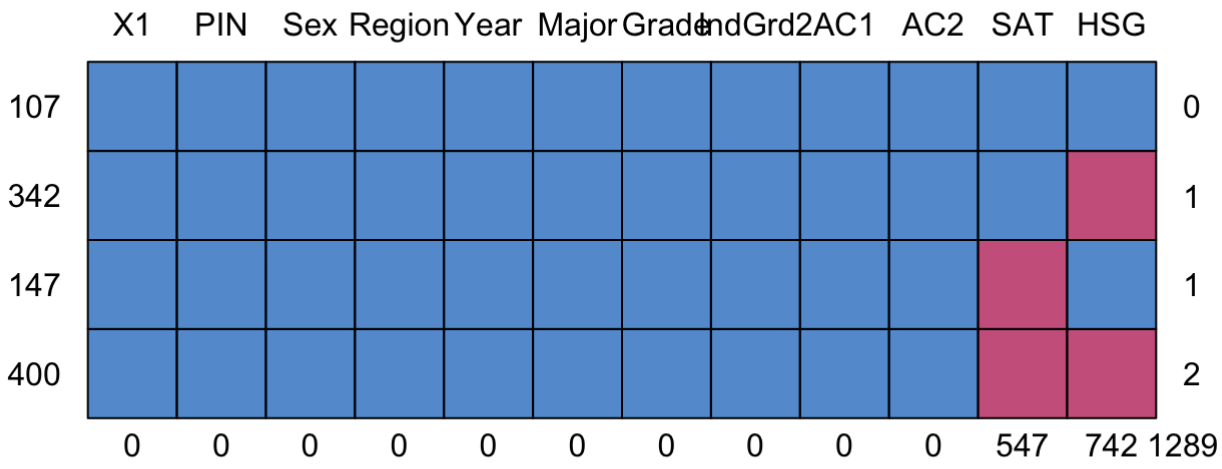
Talha Ahmad Farooqui

3/13/2021

Plots

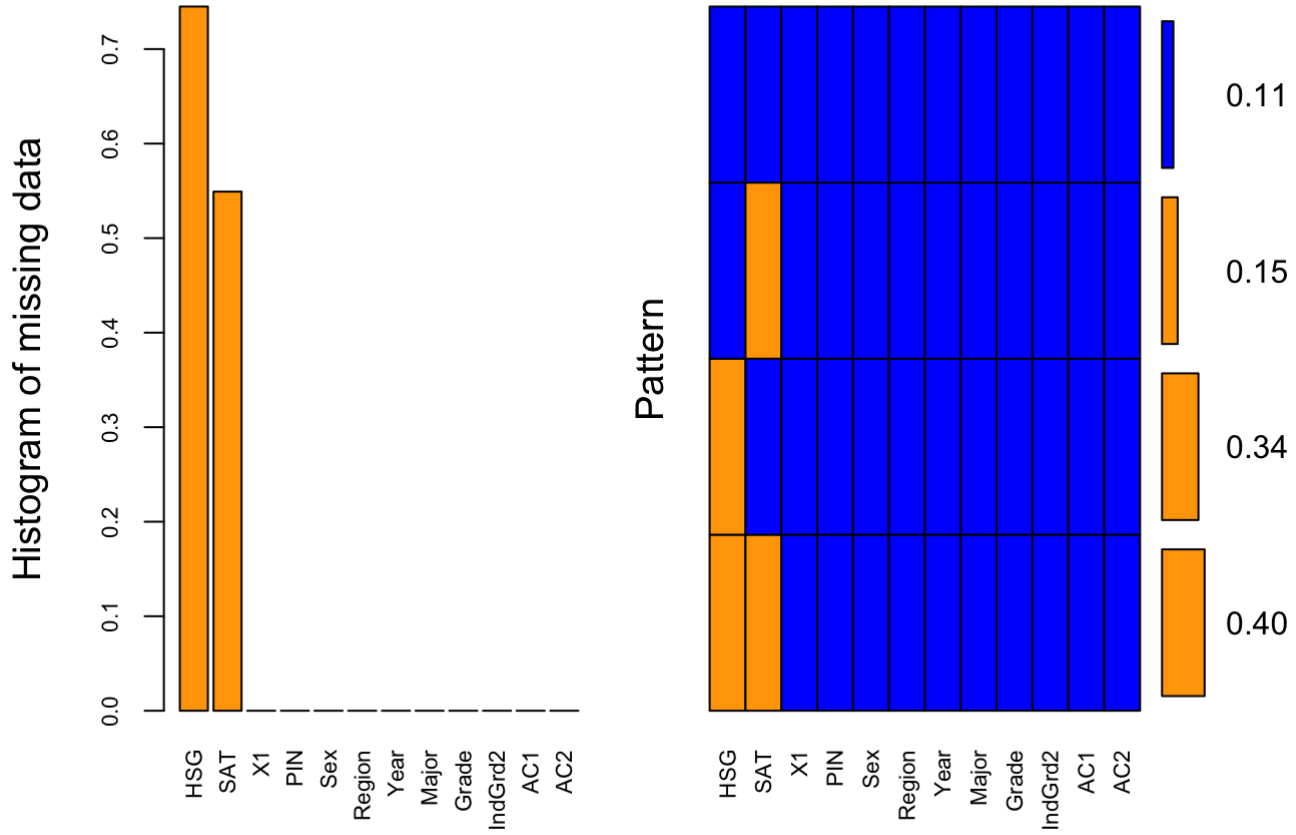
Missing data pattern plot

```
md.pattern(data)
```



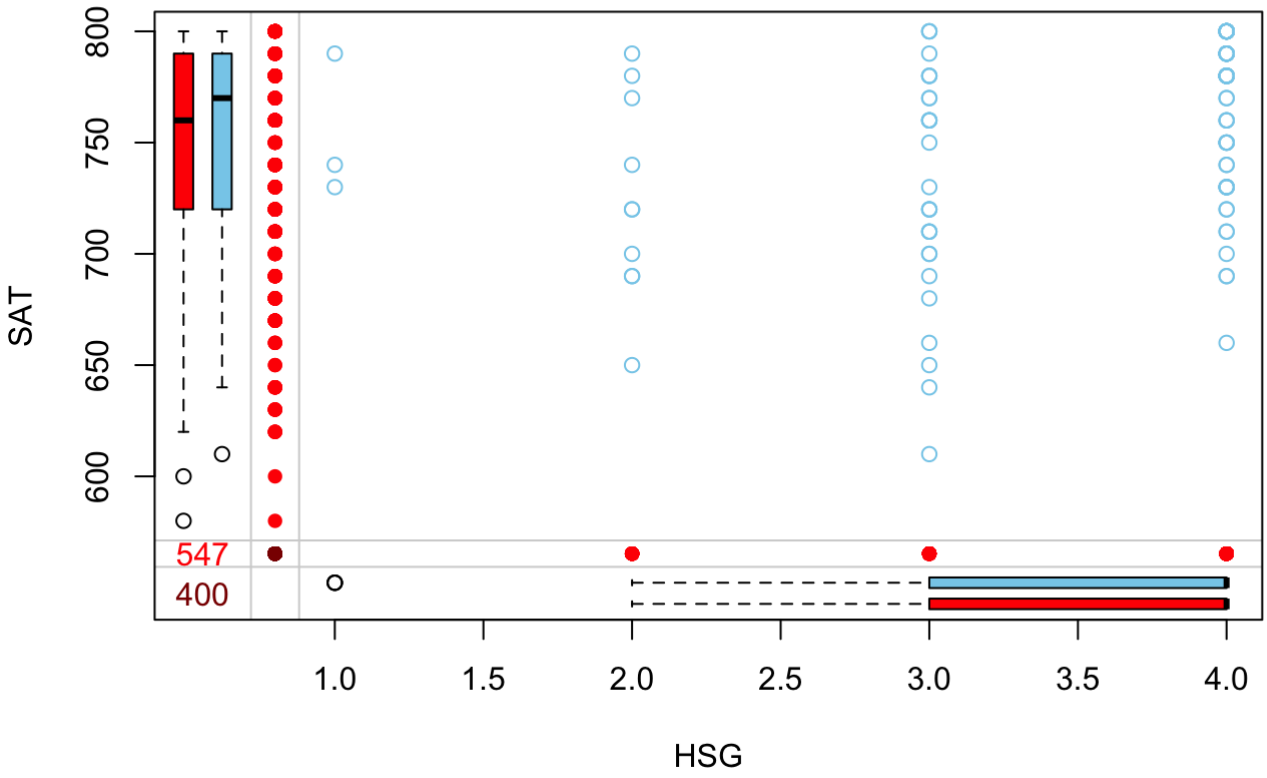

```
##     X1 PIN Sex Region Year Major Grade IndGrd2 AC1 AC2 SAT HSG
## 107  1   1   1      1    1     1     1       1   1   1   1   1    0
## 342  1   1   1      1    1     1     1       1   1   1   1   0    1
## 147  1   1   1      1    1     1     1       1   1   1   0   1    1
## 400  1   1   1      1    1     1     1       1   1   1   0   0    2
##      0   0   0      0    0     0     0       0   0   0 547 742 1289
```

```
#
#The output tells us that 107 samples are complete, 342 samples miss only the , 147 samp
les miss only the value and so on.

#Aggregations for missing/imputed values
aggr_plot <- aggr(data, col=c('blue','orange'), numbers=TRUE, sortVars=TRUE, labels=name
s(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##  Variable     Count
##       HSG 0.7449799
##       SAT 0.5491968
##        X1 0.0000000
##       PIN 0.0000000
##       Sex 0.0000000
##    Region 0.0000000
##      Year 0.0000000
##     Major 0.0000000
##     Grade 0.0000000
##   IndGrd2 0.0000000
##       AC1 0.0000000
##       AC2 0.0000000
```

```
#The plot helps us understanding that almost 40% of the samples are missing HSG and SAT
 information,
#34% are missing the HSG value, and the remaining ones show other missing patterns. T
#hrough this approach the situation looks a bit clearer in my opinion.
#Another (hopefully) helpful visual approach is a special box plot
marginplot(data[c(8,12)])
```



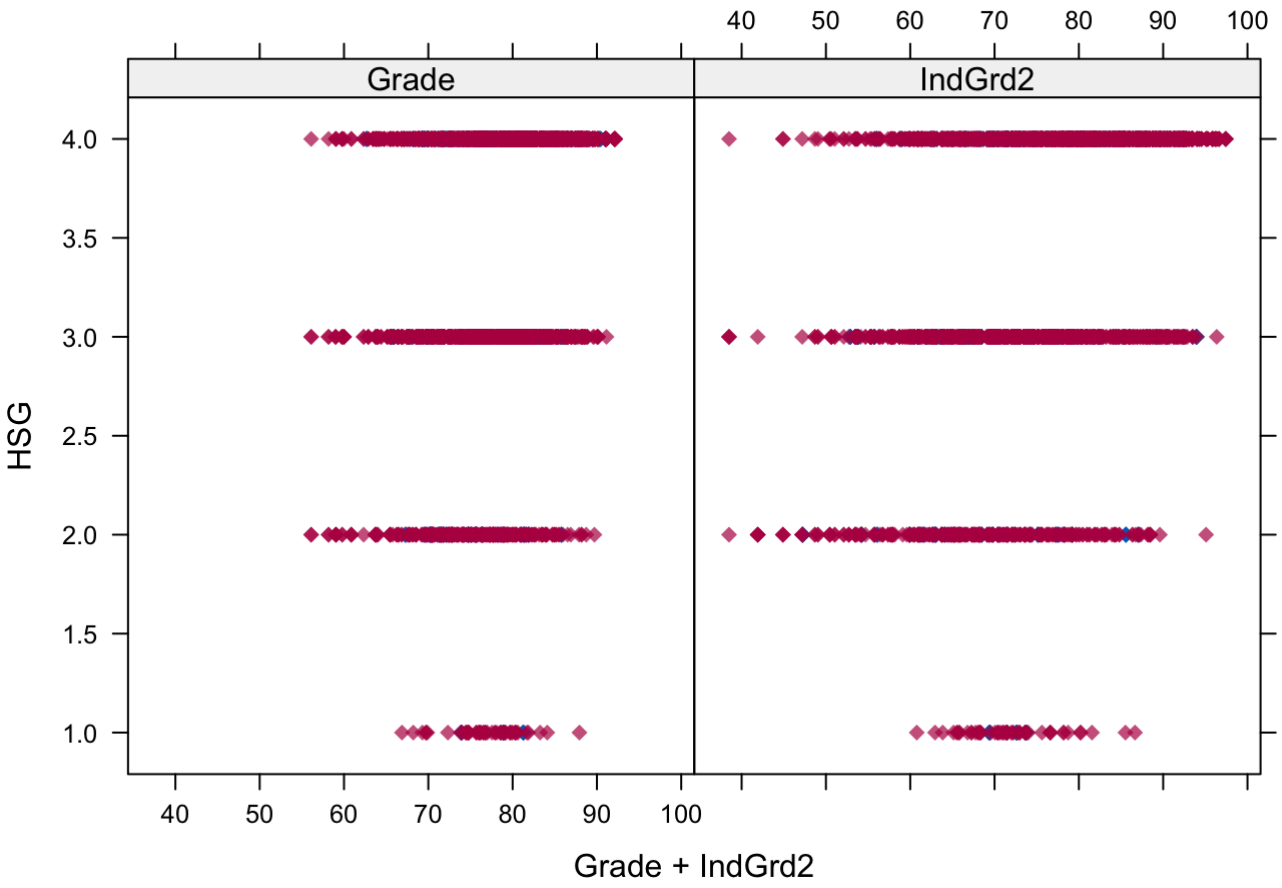

```
# Obviously here we are constrained at plotting 2 variables at a time only, but neverthe
less we can gather some interesting insights.
# The red box plot on the left shows the distribution of SAT with HSG missing while the
 blue box plot shows the distribution of the remaining datapoints. Likewhise for the HSG
box plots at the bottom of the graph.
# If our assumption of QR data is correct, then we expect the red and blue box plots to
 be very similar.


#The mice() function takes care of the imputing process
summary(tempData)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##      X1     PIN     Sex  Region    Year   Major   Grade     HSG IndGrd2     AC1
##      ""      ""      ""      ""      ""      ""      ""   "pmm"      ""      ""
##     AC2     SAT
##      ""   "pmm"
## PredictorMatrix:
##        X1 PIN Sex Region Year Major Grade HSG IndGrd2 AC1 AC2 SAT
## X1      0   0   0      0    0     0     1   1       1   0   0   1
## PIN     1   0   0      0    0     0     1   1       1   0   0   1
## Sex     1   0   0      0    0     0     1   1       1   0   0   1
## Region  1   0   0      0    0     0     1   1       1   0   0   1
## Year    1   0   0      0    0     0     1   1       1   0   0   1
## Major   1   0   0      0    0     0     1   1       1   0   0   1
## Number of logged events:  7
##   it im dep     meth    out
## 1  0  0     constant    PIN
## 2  0  0     constant    Sex
## 3  0  0     constant Region
## 4  0  0     constant   Year
## 5  0  0     constant  Major
## 6  0  0     constant    AC1
```

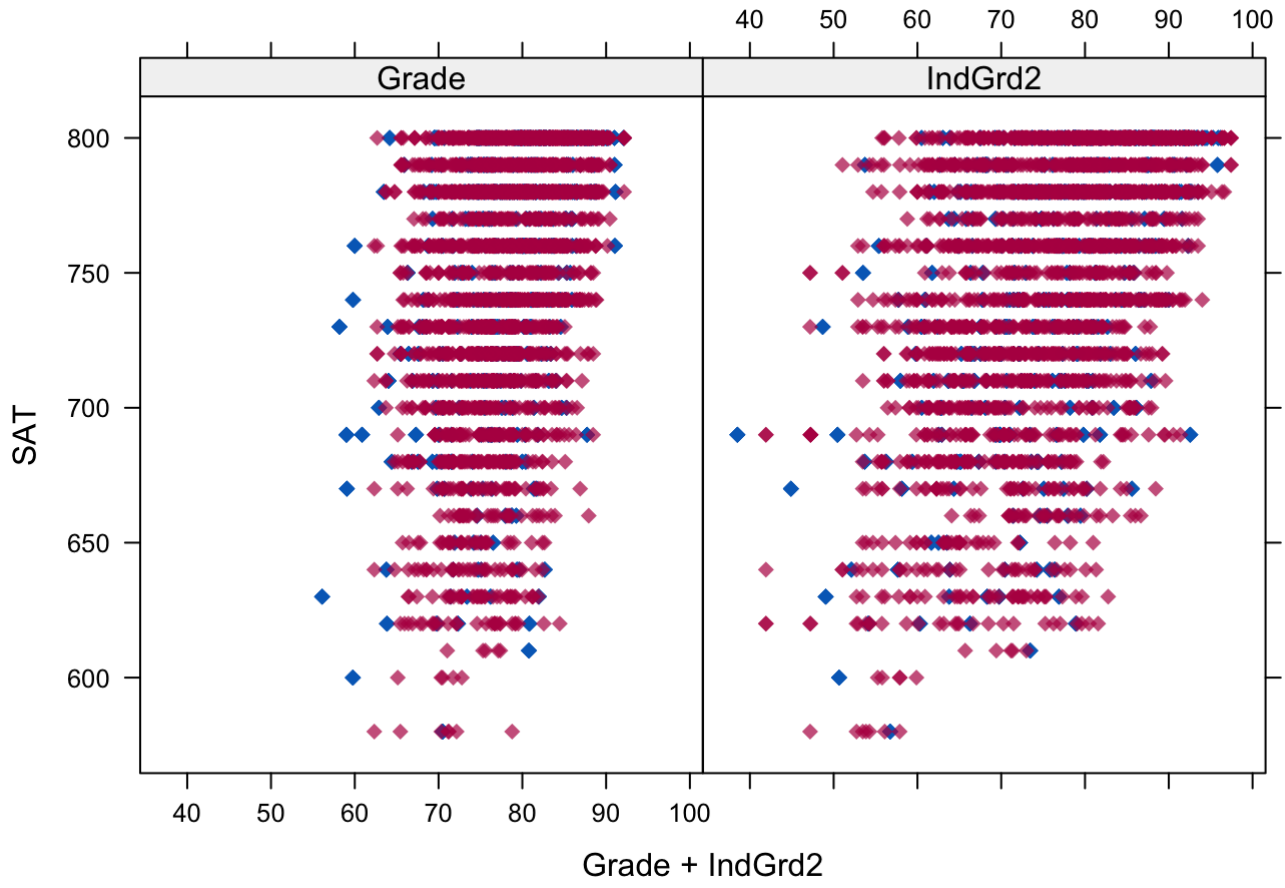
```
#A couple of notes on the parameters:

# m=5 refers to the number of imputed datasets. Five is the default value.
# meth='pmm' refers to the imputation method. In this case we are using predictive mean
 matching as imputation method. Other imputation methods can be used, type methods(mice)
for a list of the available imputation methods.
# If you would like to check the imputed data, for instance for the variable SAT, you ne
ed to enter the following line of code

SAT_imp <-tempData$imp$SAT

#The output shows the imputed data for each observation (first column left) within each
 imputed dataset (first row at the top).
#If you need to check the imputation method used for each variable, mice makes it very e
asy to do
tempData$meth
```

```
##      X1     PIN     Sex  Region    Year   Major   Grade     HSG IndGrd2     AC1
##      ""      ""      ""      ""      ""      ""      ""   "pmm"      ""      ""
##     AC2     SAT
##      ""   "pmm"
```

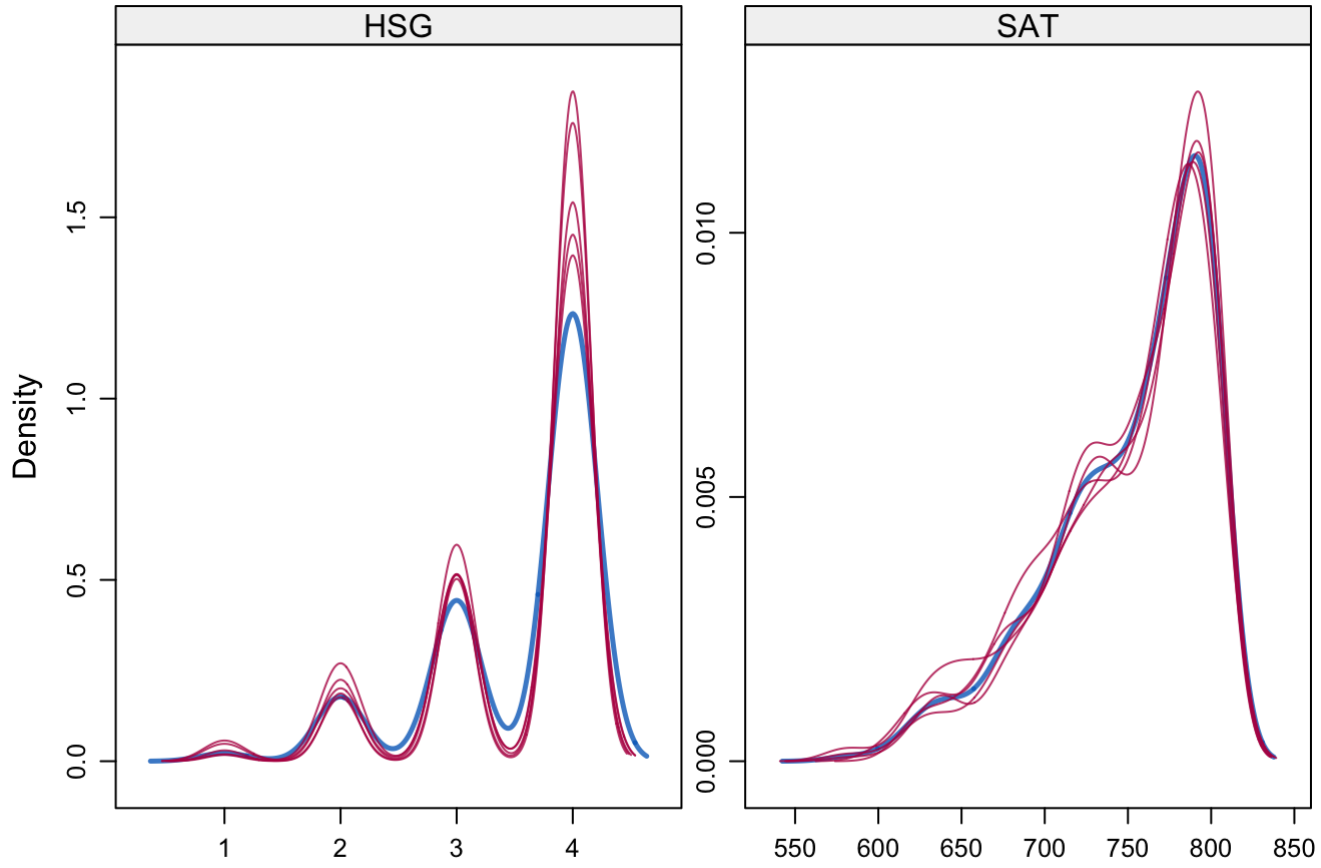
```
#Now we can get back the completed dataset using the complete() function. It is almost p
lain English:
completedData <- complete(tempData,1)

#The missing values have been replaced with the imputed values in the first of the five
 datasets.
#If you wish to use another one, just change the second parameter in the complete() func
tion.

#Inspecting the distribution of original and imputed data
#Let's compare the distributions of original and imputed data using a some useful plots.
#First of all we can use a scatterplot and plot HSG/SAT against the Grade variables
xyplot(tempData,HSG ~ Grade + IndGrd2,pch=18,cex=1)
```




```
xyplot(tempData,SAT ~ Grade + IndGrd2,pch=18,cex=1)
```




```
#What we would like to see is that the shape of the magenta points (imputed) matches the
shape of the blue ones (observed).
#The matching shape tells us that the imputed values are indeed “plausible values”.
#Another helpful plot is the density plot:
densityplot(tempData)
```




```
# The density of the imputed data for each imputed dataset is showed in magenta while th
e density of the observed data is showed in blue.
#Again, under our previous assumptions we expect the distributions to be similar.
```