# Module 4 - Missing Data

## Cooper, Sarah

library(tidyverse)

library(VIM)

library(mice)

class\_grades <- read\_csv("C:/Users/Sarah/Downloads/class-grades.csv")

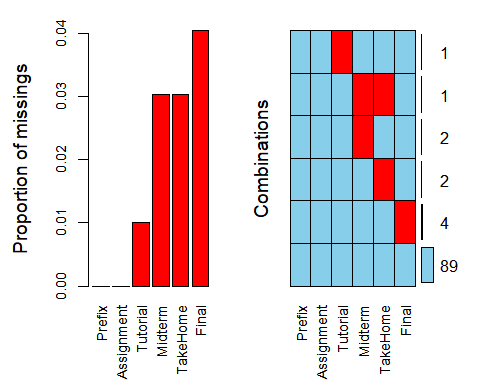
grades <- class\_grades

## Task 1

*There are 11 variables denoted with an “NA” in the headers of Tutorial, Midterm, TakeHome, and Final.*

## Task 2

vim\_plot = aggr(grades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.9)



*There is minimal systematic missingness present in this dataset. Only one student is missing both Midterm and TakeHome.*

## Task 3

RowWiseDeletion = grades %>% drop\_na(Midterm)

*The new dataset now contains 96 rows.*

## Task 4

ColumnWiseDeletion = grades %>% select(-Final)

*The new dataset now contains 5 columns.*

## Task 5

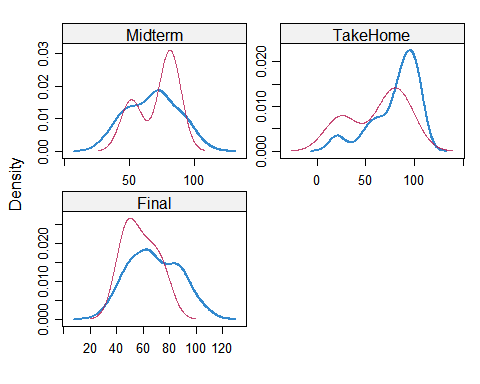
*I think the row-wise deletion method for this dataset is more applicable. Since the column-wise method deletes the variable that the dataset was hoping to predict (Final) I would suggest foregoing this method and using row-wise instead.*

## Task 6

grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

summary(grades\_imp)

densityplot(grades\_imp)



grades\_complete = complete(grades\_imp)  
summary(grades\_complete)

## Prefix Assignment Tutorial Midterm   
## Min. :4.000 Min. : 28.14 Min. : 34.09 Min. : 28.12   
## 1st Qu.:7.000 1st Qu.: 80.88 1st Qu.: 84.69 1st Qu.: 52.50   
## Median :8.000 Median : 89.94 Median : 93.10 Median : 69.38   
## Mean :7.313 Mean : 85.49 Mean : 89.76 Mean : 67.80   
## 3rd Qu.:8.000 3rd Qu.: 95.00 3rd Qu.:100.55 3rd Qu.: 81.88   
## Max. :8.000 Max. :100.83 Max. :112.58 Max. :110.00   
## TakeHome Final   
## Min. : 16.91 Min. : 28.06   
## 1st Qu.: 67.96 1st Qu.: 52.09   
## Median : 87.96 Median : 65.56   
## Mean : 80.54 Mean : 67.81   
## 3rd Qu.: 98.42 3rd Qu.: 83.19   
## Max. :108.89 Max. :108.89

## Task 7

*Issues can be encountered when attempting to deal with missing data. By using the row-wise or column-wise deletion method we could potentially be tossing valuable data by trying to rid of invaluable data. It could funnel our results perhaps making the dataset findings less than reliable. By imputating, you have a much better chance of obtaining a result that is more representative. Taking the mean of the information you do have to fill in the information that you don’t is far preferred. I think there is a smaller risk factor in this approach.*