## Missing Data - Module 4 Assignment 1

Load Libraries

options(tidyverse.quiet = TRUE)  
library(tidyverse)  
library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

library(mice)

##   
## Attaching package: 'mice'

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

grades = read\_csv("class-grades.csv")

## Parsed with column specification:  
## cols(  
## Prefix = col\_double(),  
## Assignment = col\_double(),  
## Tutorial = col\_double(),  
## Midterm = col\_double(),  
## TakeHome = col\_double(),  
## Final = col\_double()  
## )

str(grades)

## tibble [99 x 6] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Prefix : num [1:99] 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num [1:99] 57.1 95 83.7 81.2 91.3 ...  
## $ Tutorial : num [1:99] 34.1 105.5 83.2 96.1 93.6 ...  
## $ Midterm : num [1:99] 64.4 67.5 30 49.4 95 ...  
## $ TakeHome : num [1:99] 51.5 99.1 63.1 105.9 107.4 ...  
## $ Final : num [1:99] 52.5 68.3 48.9 80.6 73.9 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Prefix = col\_double(),  
## .. Assignment = col\_double(),  
## .. Tutorial = col\_double(),  
## .. Midterm = col\_double(),  
## .. TakeHome = col\_double(),  
## .. Final = col\_double()  
## .. )

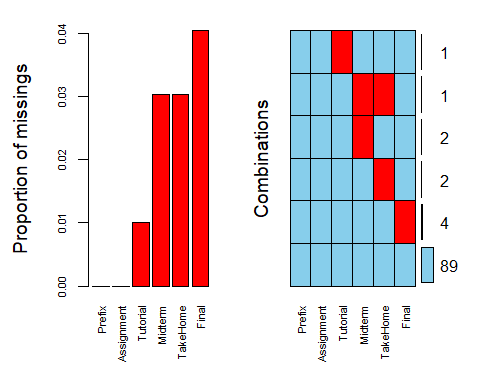
summary(grades)

## 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.: 83.93 1st Qu.: 52.50   
## Median :8.000 Median : 89.94 Median : 93.37 Median : 69.38   
## Mean :7.313 Mean : 85.49 Mean : 89.79 Mean : 67.70   
## 3rd Qu.:8.000 3rd Qu.: 95.00 3rd Qu.:100.56 3rd Qu.: 81.56   
## Max. :8.000 Max. :100.83 Max. :112.58 Max. :110.00   
## NA's :1 NA's :3   
## TakeHome Final   
## Min. : 16.91 Min. : 28.06   
## 1st Qu.: 69.91 1st Qu.: 52.91   
## Median : 88.42 Median : 66.11   
## Mean : 81.12 Mean : 68.23   
## 3rd Qu.: 99.07 3rd Qu.: 83.61   
## Max. :108.89 Max. :108.89   
## NA's :3 NA's :4

The variables Tutorial, Midterm, TakeHome and Final are missing 1,3,3,4 values respectively.

View missingness

vim\_plot = aggr(grades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



There is one case of systematic missingness with a student missingboth the Midterm and Takehome.

grades\_r = grades %>% drop\_na()

We have 89 rows of data after removing NA’s

Column-wise deletion of variables with NA’s.

grades\_c = grades %>% select(-c(Tutorial, Midterm, TakeHome, Final))   
str(grades\_c)

## tibble [99 x 2] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Prefix : num [1:99] 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num [1:99] 57.1 95 83.7 81.2 91.3 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Prefix = col\_double(),  
## .. Assignment = col\_double(),  
## .. Tutorial = col\_double(),  
## .. Midterm = col\_double(),  
## .. TakeHome = col\_double(),  
## .. Final = col\_double()  
## .. )

We are left with only two columns.

Row-wise deletion is better suited for this dataset. If we delete columns we are left with only one variable to analyze.

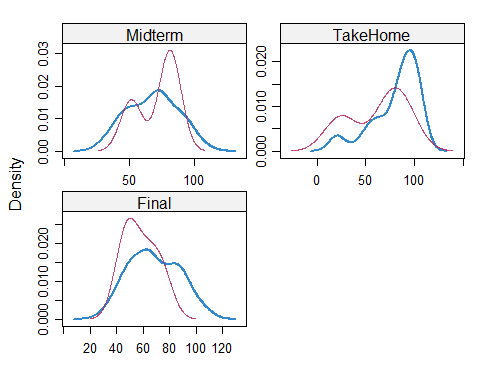
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## iter imp variable  
## 1 1 Tutorial Midterm TakeHome Final  
## 2 1 Tutorial Midterm TakeHome Final  
## 3 1 Tutorial Midterm TakeHome Final  
## 4 1 Tutorial Midterm TakeHome Final  
## 5 1 Tutorial Midterm TakeHome Final

#in line above: m=1 -> runs one imputation, seed sets the random number seed to get repeatable results  
summary(grades\_imp)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## Prefix Assignment Tutorial Midterm TakeHome Final   
## "" "" "pmm" "pmm" "pmm" "pmm"   
## PredictorMatrix:  
## Prefix Assignment Tutorial Midterm TakeHome Final  
## Prefix 0 1 1 1 1 1  
## Assignment 1 0 1 1 1 1  
## Tutorial 1 1 0 1 1 1  
## Midterm 1 1 1 0 1 1  
## TakeHome 1 1 1 1 0 1  
## Final 1 1 1 1 1 0

densityplot(grades\_imp)



#red imputed, blue original, only shows density plots when more than 1 value the variable was imputed  
#note that the density plots are fairly uninteresting given the small amount of missing data  
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

Missing data could be random or not random, imputation is not advisable when there is not at random missing data is present. Missing data can skew the model if to many rows are deleted.