# Missing Data

## June 1st, 2020

### Stephen Kiser

library(tidyverse)

## -- Attaching packages ------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ---------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(mice)

##   
## Attaching package: 'mice'

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

library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

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

##   
## Attaching package: 'VIM'

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

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()  
## )

### Task 1

str(grades)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 99 obs. of 6 variables:  
## $ Prefix : num 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num 57.1 95 83.7 81.2 91.3 ...  
## $ Tutorial : num 34.1 105.5 83.2 96.1 93.6 ...  
## $ Midterm : num 64.4 67.5 30 49.4 95 ...  
## $ TakeHome : num 51.5 99.1 63.1 105.9 107.4 ...  
## $ Final : num 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()  
## .. )

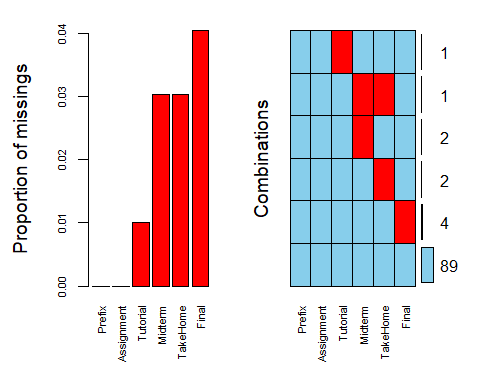
summary(grades)

## 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

There is 11 grades missing in total for this whole dataset. There is just one tutorial grade missing, 3 midterm exam grades, 3 takehome exam grades, and 4 final exam grades.

### Task 2

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



There are 4 grades missing the just the final exam grade. Then there are 2 missing grades for the takehome exam and there is also 2 missing grades for midterm grades. Then there is one tutorial grade missing and a combination of midterm and takehome exam grade missing. There is only one student missing multiple grades.

### Task 3

grades\_row <- grades %>% drop\_na(c(Tutorial, Midterm, TakeHome, Final))  
  
str(grades\_row)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 89 obs. of 6 variables:  
## $ Prefix : num 5 8 8 7 8 7 8 8 7 7 ...  
## $ Assignment: num 57.1 95 83.7 81.2 91.3 ...  
## $ Tutorial : num 34.1 105.5 83.2 96.1 93.6 ...  
## $ Midterm : num 64.4 67.5 30 49.4 95 ...  
## $ TakeHome : num 51.5 99.1 63.1 105.9 107.4 ...  
## $ Final : num 52.5 68.3 48.9 80.6 73.9 ...

summary(grades\_row)

## Prefix Assignment Tutorial Midterm   
## Min. :4.000 Min. : 28.14 Min. : 34.09 Min. : 28.12   
## 1st Qu.:7.000 1st Qu.: 82.45 1st Qu.: 82.93 1st Qu.: 52.50   
## Median :8.000 Median : 89.94 Median : 92.93 Median : 70.00   
## Mean :7.303 Mean : 85.97 Mean : 89.40 Mean : 68.14   
## 3rd Qu.:8.000 3rd Qu.: 95.05 3rd Qu.:100.58 3rd Qu.: 83.12   
## Max. :8.000 Max. :100.83 Max. :112.58 Max. :110.00   
## TakeHome Final   
## Min. : 16.91 Min. : 28.06   
## 1st Qu.: 64.07 1st Qu.: 52.50   
## Median : 87.96 Median : 65.56   
## Mean : 81.03 Mean : 67.97   
## 3rd Qu.: 97.78 3rd Qu.: 83.06   
## Max. :108.89 Max. :108.89

After doing a row-wise deletion of missing values there are 89 rows left from the original 99.

### Task 4

grades\_col <- grades %>% select(-c(Final, Midterm, TakeHome,Tutorial))  
  
str(grades\_col)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 99 obs. of 2 variables:  
## $ Prefix : num 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num 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()  
## .. )

summary(grades\_col)

## Prefix Assignment   
## Min. :4.000 Min. : 28.14   
## 1st Qu.:7.000 1st Qu.: 80.88   
## Median :8.000 Median : 89.94   
## Mean :7.313 Mean : 85.49   
## 3rd Qu.:8.000 3rd Qu.: 95.00   
## Max. :8.000 Max. :100.83

After doing the column-wise deletion only 2 columns were left in the dataset. These two columns were Prefix and Assignment grades. It kept the 99 observations.

### Task 5

After doing both the row-wise deletion and the column-wise deletion the most preferable to get rid of missing values is row-wise deletion. Even though we lost 10 observations with the row-wise deletion method we kept all 6 columns which is what we want to calculate in this engineering program. We want to be able to calculate the grades of students. The column-wise deletion method gets ride of to much of the data we are trying to calculate.

### Task 6

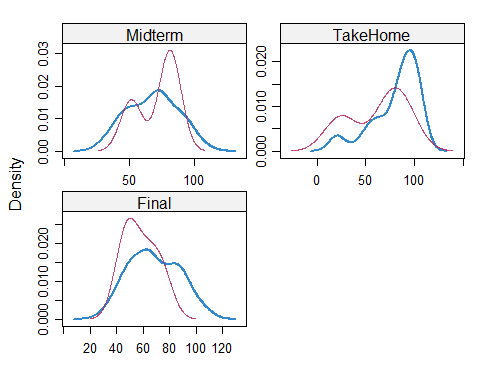
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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

### Task 7

The red graph is what is imputated for our grades by the mice function. It takes from our dataset to imputate these data values. As we see the midterm graph drastically changes with imputation. The second peak rises much higher than what our original graph gives us, and this could cause errors in our calculations. The final exam grade also has this issue but in the lower half of the grades. this could cause our average exam score to be lower than what it actually is. A time when imputation may not be advisable is when there is a lot of data missing.