## Missing Data

### Anthony Sumter

**Libraries**

#install.packages("tidyverse","mice", "VIM")  
options(tidyverse.quiet = TRUE)  
library(tidyverse)  
library(mice)

## Warning: package 'mice' was built under R version 3.6.2

## Loading required package: lattice

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

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

library(VIM)

## Warning: package 'VIM' was built under R version 3.6.2

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.6.2

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

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

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

## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

## 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

**Dataset**

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

#### Task 1: How much data is missing and in what variables?

There are 11 pieces of data (NA) that are missing in variables Tutorial, Midterm, TakeHome, and Final.

str(grades)

## 'data.frame': 99 obs. of 6 variables:  
## $ Prefix : int 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 ...

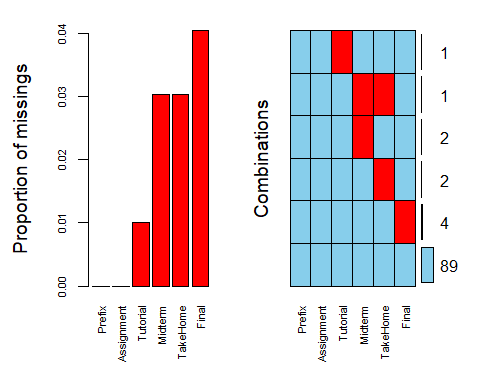
summary(grades)

## 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

#### Task 2: Use the VIM package to visualize missingness. Does there appear to be systematic missingness? In other words, are there students that are mising multiple pieces of data?

It does appear that there is systematic missingness as in the below visual it shows that students are missing data in the Midterm and TakeHome variables.

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



#### Task 3: Use row-wise deletion of missing values to create a new data frame. How many rows remain in this data frame?

In the new data frame of grades2 there are a total of 89 rows that remain.

grades2 = grades %>% drop\_na(Tutorial, Midterm, TakeHome, Final)  
str(grades2)

## 'data.frame': 89 obs. of 6 variables:  
## $ Prefix : int 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(grades2)

## 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

#### Task 4: Use column-wise deletion of missing values to create a new data frame (from the original data frame not from the data frame created in Task 3). How many columns remain in this data frame?

Using column-wise deletion only 2 columns would remain in the new data frame of grades3.

grades3 = grades %>% select(-Tutorial, -Midterm, -TakeHome, -Final)   
str(grades3)

## 'data.frame': 99 obs. of 2 variables:  
## $ Prefix : int 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num 57.1 95 83.7 81.2 91.3 ...

summary(grades3)

## 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

#### Task 5: Which approach (Task 3 or Task 4) seems preferable for this dataset? Briefly discuss your answer.

Task 3 would be more preferable because there would more data to use in the type of analysis that would be conducted. Using task 4 would be limited in terms of a deeper analysis because you would only be using 2 variables compared to 6.

#### Task 6 Use the code below to impute the missing values in the dataset using the mice package.

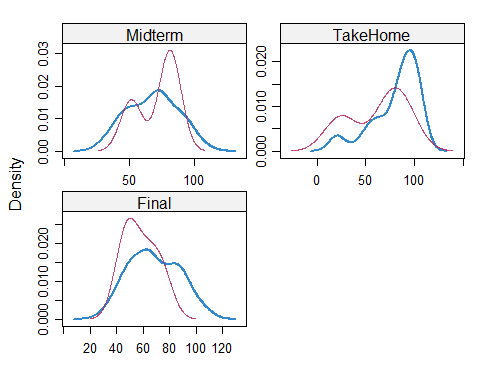
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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: Briefly discuss potential issues that could be encountered when working with missing data. Describe situations where imputation may not be advisable.

Potential issues that could arise when working with missing data is that it could lessen the statistical impact of the overall study and lead to biased opinions that doesn’t tell the full story of the data. Situations where imputation may not be advisable is if there is a majority of data missing or if data is missing not at random.