## Daniel Sanchez

## BAN 502

## Module 4 Assignment 1

## Dr. Stephen Hill

## June 8th, 2020

library(tidyverse)  
library(VIM)  
library(mice)

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

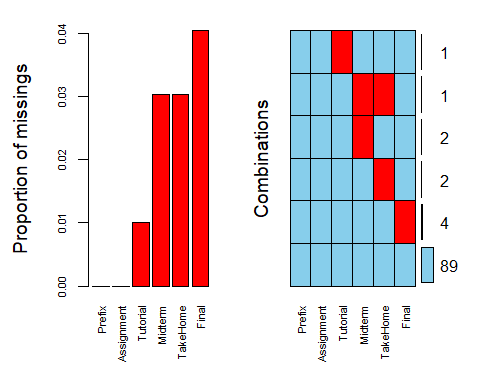
## Prefix Assignment Tutorial Midterm   
## Min. :4.000 Min. : 28.14 Min. : 34.09 Min. : 28.12   
## 1st Qu.:7.000 1st Qu.: 80.88 1st Qu.: 83.93 1st Qu.: 52.50   
## Median :8.000 Median : 89.94 Median : 93.37 Median : 69.38   
## Mean :7.313 Mean : 85.49 Mean : 89.79 Mean : 67.70   
## 3rd Qu.:8.000 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 1: How much data is missing and in what variables?**

By examining the summary of the “grades” dataset, we can find missing data in the following columns: Tutorials (1 missing value), Midterm (1 missing value), TakeHome(3 missing values), and Final(4 missing values).

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

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



The VIM plot shows that 1 student has two missing variables (Mideterm and TakeHome). Multiple students have missing values for either Midterm or TakeHome. The most missingness is found in the Final variable.

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

grades\_rowwise = grades %>% drop\_na(Tutorial,Midterm,TakeHome,Final)  
ncol(grades\_rowwise)#Columns

## [1] 6

nrow(grades\_rowwise)#Rows

## [1] 89

Row-wise deletion removed 10 rows from the original data set, while retaining the same number of columns. We are left with 89 rows of data.

**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?**

grades\_columnwise = grades %>% select(-Tutorial,-Midterm,-TakeHome,-Final)  
ncol(grades\_columnwise)#Columns

## [1] 2

nrow(grades\_columnwise)#Rows

## [1] 99

Column-wise deletion removed 4 variables from the original data set. We are left with only 2 columns of data, with the same number of rows (99).

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

For this dataset, row-wise deletion would be preferable. This approach left most of the data set intact. This contrasts with the column-wise deletion of missing data. The column-wise approach left only Prefix and Assignment, which are variables that would provide no insight on their own.

**Task 6 Impute the missing values in the dataset using the mice package.**

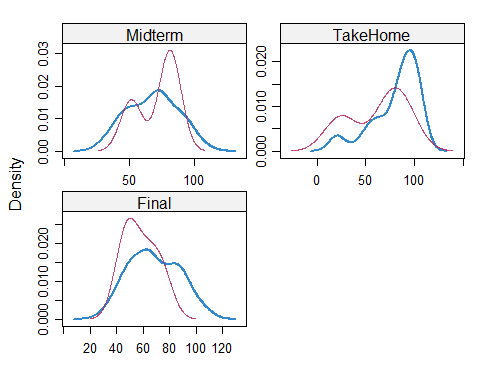
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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

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)



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.**

Missing data can arise in a variety of ways. A large amount of missing data could be indicative of a poor design in the data collection model itself. Being limited to the data set on hand, the type of missing data and its potential influence on the analysis should determine how it is handled. If one variable contains so much missing data that it results in having no impact on the analysis, that variable should be removed. If missing data is scattered across many different variables, column-wise deletion is most likely not the best option.

Another issue can result from the labeling of missing data in the source file. For example, a document might list null values, zeroes, or “NA” to represent the same concept. It would be important to understand what each labeling means before moving to exclude or impute the data.

Since imputing is basically data-replacement or data-substitution, it seems like it should be used sparingly. In a scenario where a critical variable is missing a large amount of data, imputation may not be the best option. Instead, it might help to identify a surrogate in the data set that could adequately represent the variable in question.