# Module 4 - Missing Data Assignment

## Brian Libby

### Answer to Questions (See relevant sections below for analysis)

Task 1:  
1. How much data is missing and in what variables?  
There are a total of 11 missing pieces of data, spread over 4 variables and 10 rows, in the data set. The variables with missing data are Tutorial (1), Midterm (3), TakeHome (3), and Final (4).

Task 2:  
1. Does there appear to be systematic missingness? In other words, are there students that are mising multiple pieces of data?  
**There is one student that is missing mutliple pieces of data. This student is missing data in Midterm and TakeHome. All other students with missing data are only missing one piece of data.**

Task 3:  
1. Use row-wise deletion of missing values to create a new data frame. How many rows remain in this data frame?  
**After deleting rows with missing data, there are 89 rows of data remaining.**

Task 4:  
1. 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?  
**After deleting columns with missing data, there are only 2 columns of data remaining.**

Task 5:  
1. Which approach (Task 3 or Task 4) seems preferable for this dataset? Briefly discuss your answer.  
**The row-wise approach from Task 3 seems preferable for this dataset. Row-wise deletion only removed ten observations from the data. While this could be significant, column-wise deletion removed all of the data from four variables that are important in predicting grades on the final exam.**

Task 7:  
1. Briefly discuss potential issues that could be encountered when working with missing data.  
**There are several potential issues that could be encountered when working with missing data. First, there are several packages in R that cannot handle missing data. Users will receive errors if they try to use these packages with missing data. Potential issues exist when users decide to remove/replace missing data. Row-wise deletion will remove the entire row of data if one or more variables have missing data. Issues arise because usable data has been removed from the data set. Column-wise deletion will remove an entire variable from the data set. Issues arise because an important predictor variable may be removed from the data set, thus altering the effectiveness of a model. Imputation will replace missing data. This method can use the mean of similar reords, regression, maximum likelihood, or multiple imputation to replace missing data.**  
2. Describe situations where imputation may not be advisable.  
**Imputation may not be advisable when examining results of critical studies, like a clinical trial. Biases introduced through imputation could skew the results of these studies in a favorable way, when the true results might indicate the opposite. It would be better to remove the missing data and report results on the limited data set. Researchers may then discusss the effects the missing data may have had on the results of the study.**  
**The resulting charts in Task 6 support this. The data for Midterm is biased towards higher grades after imputation while TakeHome is skewed in the opposite direction. The results for the Final are also significantly sekewed by the imputation of the missing data.**

### Load Libraries

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

### Load Data

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 - How Much Missing Data and In What Variables

str(grades)

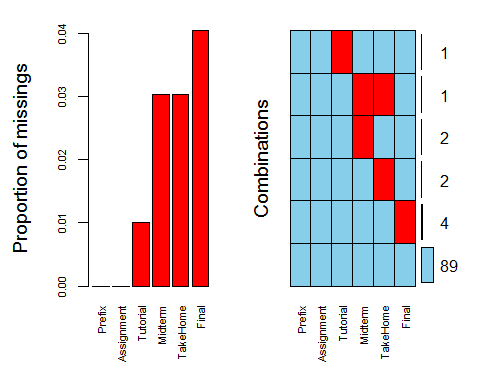
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 99 obs. of 6 variables:  
## $ Prefix : num 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num 57.1 95 83.7 81.2 91.3 ...  
## $ Tutorial : num 34.1 105.5 83.2 96.1 93.6 ...  
## $ Midterm : num 64.4 67.5 30 49.4 95 ...  
## $ TakeHome : num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

### Task 2 - Visualize Missingness with VIM

vim\_plot <- aggr(grades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7) #note: the cex.axis reduces size of text on x-axis so labels fit better



### Task 3 - Row-wise Deletion of Missing Data

grades\_row <- grades %>% drop\_na()  
nrow(grades\_row)

## [1] 89

### Task 4 - Column-wise Deletion of Missing Data

grades\_column <- grades %>% select(-Tutorial, -Midterm, -TakeHome, -Final)  
ncol(grades\_column)

## [1] 2

### Task 6 - Impute Missing Data Using Mice Package

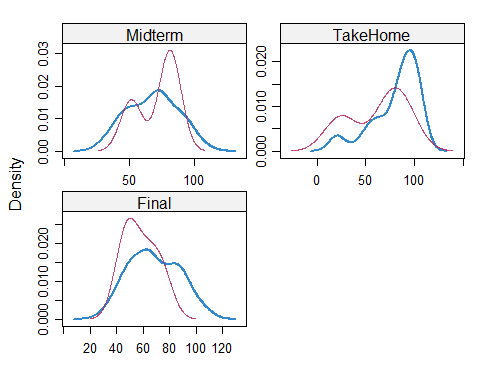
grades\_imp <- mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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