# Missing Data

### Miguel Fernandez

### BAN-502

# Import libraries  
library(tidyverse)  
library(VIM)  
library(mice)

# Read in data  
grades <- read\_csv("class-grades.csv")

#### Task 1

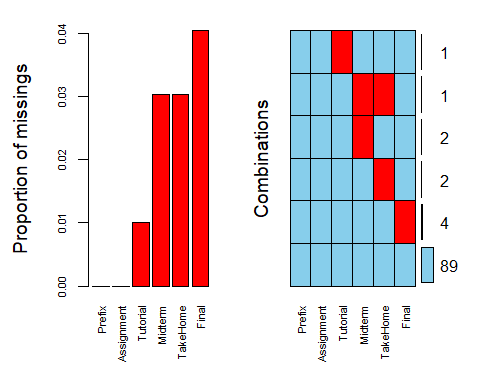
# Summary of data  
summary(grades)

## 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 summary above reveals that of the 99 observations in the data set, we are missing 1 value in Tutorial, 3 values in Midterm, 3 values in TakeHome and 4 values in Final.

#### Task 2

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



The aggregation plot above highlights missing data in the data set grades. We can see that there is one case where an observation is missing two variables, MidTerm and TakeHome. Apart from that one case, there are only a few instances of an observation missing a single variable.

#### Task 3

# Row-wise deletion  
drop.row <- grades %>%  
 drop\_na()  
nrow(drop.row)

## [1] 89

After performing row-wise deletion of NA values, we have a data set with 89 rows.

#### Task 4

# Column-wise deletion  
drop.col <- grades[ , colSums(is.na(grades)) == 0]  
length(drop.col)

## [1] 2

Using the column-wise approach to removing NA values, the data set is left with 2 columns of data, Prefix and Assignment.

#### Task 5

The row-wise approach is the most appropriate method for this data set. This is because only 10 rows of data are removed while the column-wise method removes 4 variables of data, leaving very little for analysis. If the columns that were removed contained more than 70 or 80 percent missing values, then removing those columns may be the best approach. However, in this case, there were only a few missing values from those columns so there is no need to remove the entire column.

#### Task 6

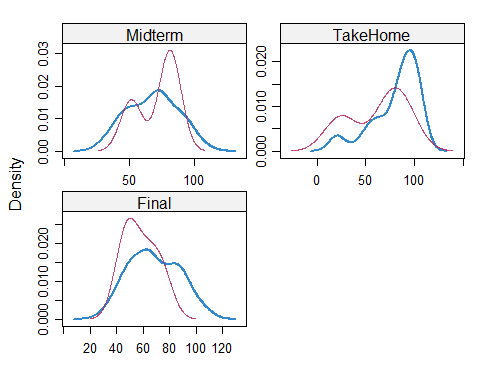
# Perform imputation on data set  
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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

# Create plot of imputed data  
densityplot(grades\_imp)



# Merge imputed data with original data set  
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

Missing data can have a tremendous impact on data analysis and modeling. Therefore, knowing how to handle missing values is extemely important. For example, missing data can cause misleading results. If values are not present when the model is trained, the model could miss out on detecting additional patterns. When new data is presented to the model, crucial insights will be missed. Also, several model algorithms will not function properly when fed missing data.  
Imputation is the process whereby missing values are predicted to fill in the data set. There are several methods that can be used including calculating the mean, median, or mode, as well as using regression techniques. However, this may not always be the best approach. For example, using mean imputation would calculate one value to be used to replace all missing values regardless of the relationship between this variable and all other variables. Or there are significant outliers present in the variable, those values will skew the imputed mean. If the data set you are working with is relatively small, using the regression approach may not be suitable because there may not be enough data for the model to make an informed decision.