# Missing Data Assignment

## BAN502 Predictive Analytics

### Susan Wiggins

Libraries needed for the assignment:

library(tidyverse)

## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.1 v dplyr 0.8.1  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

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

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

library(mice)

## 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

Reading in the data set grades:

library(readr)  
class\_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()  
## )

View(class\_grades)  
  
grades = data.frame(class\_grades)  
  
str(grades)

## '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 ...

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?

BAsed on the summary of the data for the class\_grades data set there is one value missing in the Tutorial variable, three values missing in the Midterm variable, 3 values missing in the TakeHome variable, and four values missing in the Final variable. Seems like a very small porportion of the data is missing.

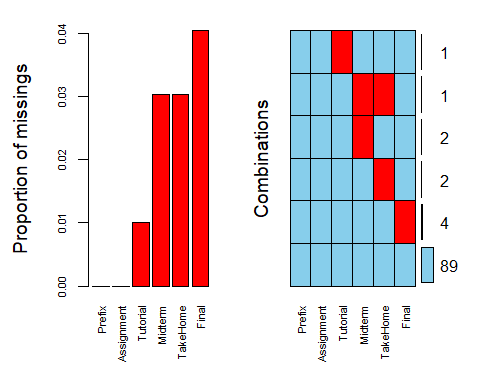
### Task 2

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?

Based on the Mid-term and TakeHome there is an overlap of data to show mutiple pieces of data missing in combinations of one category. Tutorial is missing about 1% of data, Midterm and TakeHome is missing 3% of the data, and Final is missing 4% of the data.

View Missingness:

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? The number of row remaining in the data fame is 89 rows.

If we were to do row-wise deletion of the missing variable.

row\_deleteGrades = grades %>% drop\_na(Tutorial,Midterm, TakeHome,Final)

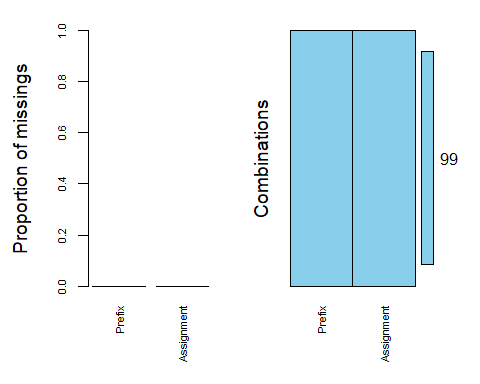
### Task 4

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?

The number of columns that still appear in the data set is 2 columns which are Prefix and Assignment.

Column-wise deletion of the variable.

column\_deleteGrades = grades %>% select(-Tutorial,-Midterm,-TakeHome,-Final)  
  
  
vim\_plot = aggr(column\_deleteGrades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



### Task 5

Which approach (Task 3 or Task 4) seems preferable for this dataset? Brieï¬y discuss your answer.

I believe the approach for task 3 for deleting rows of data would be a much better approach due to the factor that you ar removing a lot of viable data to use for analysis of the data set for grades. Have to be careful with these approaches could be destructive in the data set, looks like the column deletion could be removing a variable that is needed in analysis process.

### Task 6

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

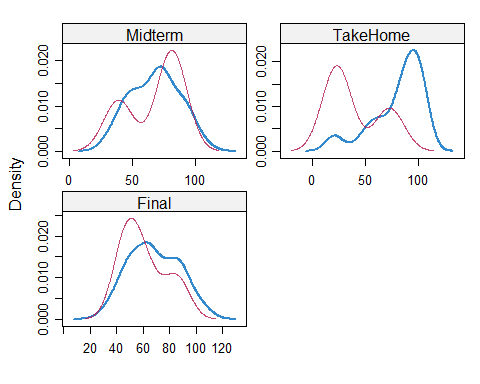
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345) #in line above: m=1 -> runs one imputation, seed sets the random number seed to get repeatable results

##   
## 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)



#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 : 67.68   
## 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.: 63.98 1st Qu.: 52.09   
## Median : 87.96 Median : 65.56   
## Mean : 79.87 Mean : 67.92   
## 3rd Qu.: 98.42 3rd Qu.: 83.61   
## 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.

Imputation is bascially coming up with an estimate of the missing values in the data set. In the density plots above you can see that the red line is the imputed Midterm and does not aline well with the actual values, neither does the TakeHome: however, the red line for imputed does aline with the structure of the data in Final. Based on the imputation it might not be feasible to try this method a better possiblity might be the row deletion method. Possibly you could also throw out a variable that has a lot of missingness, but in this case there is not a great deal of missing data.