# Module 4 - Assignment 1 - Missing Data

## Ahrens, Hailey

#### Load Libraries

options(tidyverse.quiet=TRUE)  
library(tidyverse)

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

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

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

##   
## Attaching package: 'mice'

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

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

#### Read in data and name it “grades”

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 data is missing and in what variables?

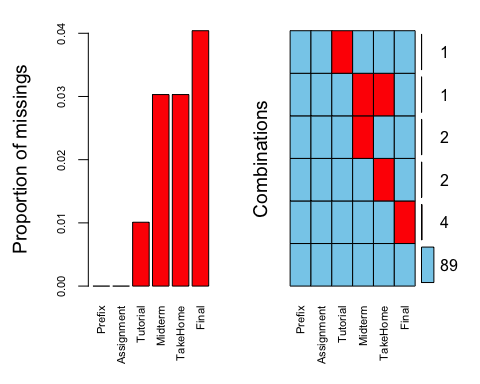
sum(is.na(grades))

## [1] 11

Data is missing in 10 rows- a total of 11 missing values within the Tutorial, Midterm, TakeHome, and Final variables.

### Task 2: View missingness with VIM package

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



vim\_plot

##   
## Missings in variables:  
## Variable Count  
## Tutorial 1  
## Midterm 3  
## TakeHome 3  
## Final 4

Does there appear to be systematic missingness? In other words, are there students that are missing multiple pieces of data?

There is 1 student missing more than 1 piece of data but the others that are missing data are only missing 1 piece of data.

### Task 3: Use row-wise deletion of missing values to create a new data frame.

rowgrades=grades%>%  
 drop\_na()

How many rows remain in the data frame? 89

### Task 4: Use column-wise deletion of missing values to create a new data frame.

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

How many columns remain in this data frame? 2

### Task 5: Which approach Task 3 or Task 4 seems preferable for this dataset?

I think Task 3 row-wise deletion seems better simply because we are only getting rid of a small part of the data vs 4 out of 6 columns of data.

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

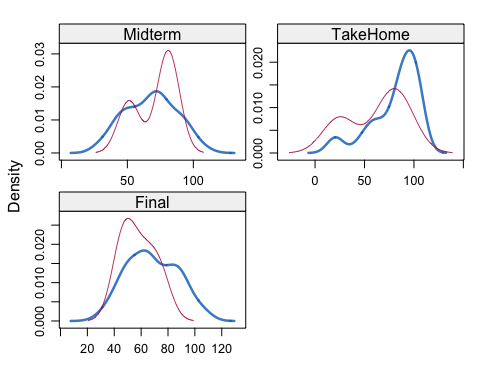
grades\_imp=mice(grades,m=1,method="pmm",seed=12345)

##   
## 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)



#red is imputed, blue is the original, only shows density plots when more than 1 value of the variable is imputed  
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 situtations where imputation may not be advisable.

An issue with row-wise deletion is that we deleted rows with missing data but those rows also contained values in other columns so we removed good data that could potentially be useful. An issue with column-wise deletion is that it gets rid of a column with any missing values so it can be destructive to the whole dataset if it is an important variable for analysis. A situation where imputation may not be advisable is using the mean option to replace a value within a categorical variable. The mean for other variables is also not going to be very accurate it is just a realistic guess.