Module 4 Missing Data Assignment

Laura Boggs

June 6, 2019

Load the necessary libraries and load in the data.

#install.packages("mice")  
#install.packages("VIM")  
library(tidyverse, quietly = TRUE)

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

## v ggplot2 3.1.0 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## 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(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

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

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?

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()  
## .. )

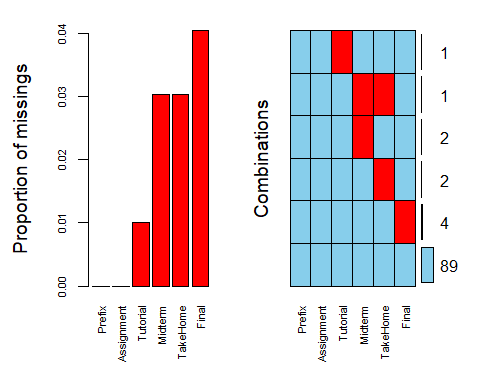
summary(grades)

## 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

Within the dataset grades there is one missing value in the variable tutorial, there are three missing values in the variable midterm, there are three missing values in the variable take home and there are four missing values in the variable final.

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?

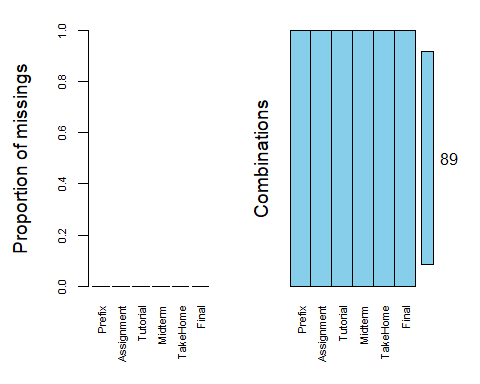
vim\_plot = aggr(grades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



After visualizing the missingness we see that for most of the missing values there does not appear to be systematic missingness except for one case. One student is missing both their midterm and their take home.

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

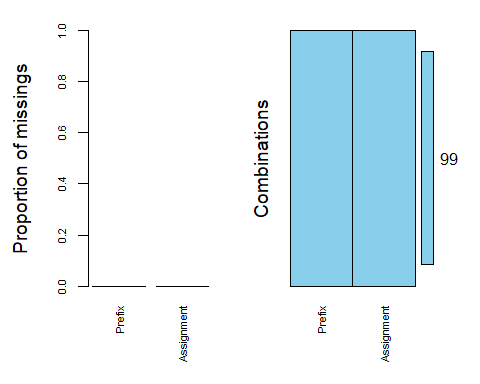
RowWiseGrades = grades %>% drop\_na()   
vim\_plot2 = aggr(RowWiseGrades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



After using row wise deletion to delete all of the rows that contained missing values there are 89 rows, there were originally 99 rows.

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?

ColumnWiseGrades = grades %>% select(-Tutorial, -Midterm, -TakeHome, -Final)   
vim\_plot3 = aggr(ColumnWiseGrades, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



After using column wise deletion to delete all of the columns that contained missing values there are 2 columns, there were originally 6 columns.

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

After doing row wise deletion and column wise deletion to remove the missing values it is clear that between the two options row wise deletion is preferable for this dataset. Even though row wise deletion did remove useful data, the column wise deletion removed almost all of the data. It would be very difficult to successfully analyze this data after using column wise deletion because there were only two columns left. Therefore for this dataset row wise deletion should be used over column wise deletion.

Task 6: Use the code below to impute the missing values in the dataset using the mice package.

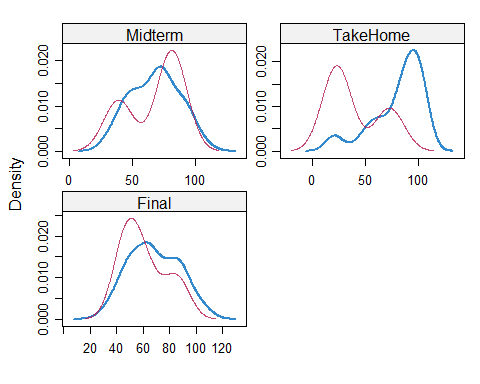
grades\_imp = mice(grades, m=1, method = "pmm", seed = 12345)

##   
## 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 : 67.68   
## 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.

There are many potential issues that could be encountered when working with missing data. Depending on the method used to remove the missing values there are different issues that the analyst might face. If they are using a deletion method (row wise or column wise) an issue that they might face is deleting valuable data that is essential to the analysis. Another issue that they might face is after deleting the missing values they might not have enough data to correctly see the pattern within the data. If they are using imputation depending on what they replace the missing values with it could potentially alter the data and the subsequent analysis. For example if they replace the missing values with the mean for that particular variable if the mean is significantly different from the actual value it could effect the pattern within the data and the analysis of the data. If they replace the missing value using a hot deck (substitute from similar observation randomly) or using a cold deck (substitute from similar observation without randomness) if the actual value is not close to the value from the similar observation it could effect the data and the analysis. Imputation may not be advisable if the most of the values in that column or row are missing. For example if there is a dataset that has 1,000 rows and 10 columns and one of the rows has missing values for 9 of the columns if we use imputation 90% of the data from that row is not the actual value and could therefore effect the data. In this example row wise deletion would be the best option because there would still be 999 rows of usable data.