missingdata\_m4\_cohen

#necessary libaries  
library(tidyverse, quiet=TRUE)

## Warning: package 'tidyverse' was built under R version 3.5.3

## -- 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

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'stringr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

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

library(VIM)

## Warning: package 'VIM' was built under R version 3.5.3

## Loading required package: colorspace

## Warning: package 'colorspace' was built under R version 3.5.3

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.3

##   
## 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)

## Warning: package 'mice' was built under R version 3.5.3

## 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

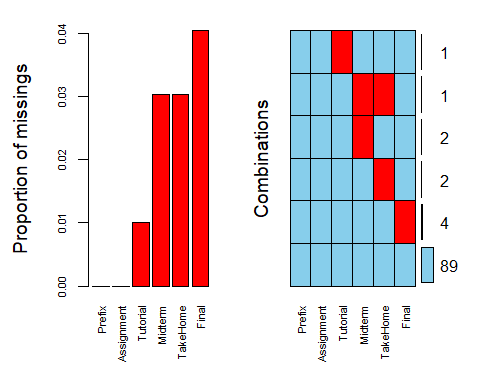
#read data in  
grades = read\_csv("class-grades.csv")

## Parsed with column specification:  
## cols(  
## Prefix = col\_double(),  
## Assignment = col\_double(),  
## Tutorial = col\_double(),  
## Midterm = col\_double(),  
## TakeHome = col\_double(),  
## Final = col\_double()  
## )

#str(grades)

# Task 1 and 2

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



Final is missing in 4% of observations. In the 4 observations it is missing, no other variables have missing values.

TakeHome is missing in 3% of observations. In 2 observations it is missing there are no other variables with missing values. In 1 observation it is missing, Midterm is also missing.

Midterm is missing in 3% of observations. In 2 observations it is missing there are no other variables with missing values. In 1 observation it is missing, TakeHome is also missing.

Tutorial is missing in 1% of observations. In the observation it is missing, no other variables have missing values.

Prefix and Assignment have no missing values.

There doesn’t seem to be any meaningful pattern to the missingness. Tutorial and Midterm missing values only were present when Prefix was 7, but Final and TakeHome missing values didn’t have a pattern.

# Task 3

Remove rows with missing data.

grades\_delrow = na.omit(grades)  
#str(grades\_delrow)

After deleting rows that contained any missing values, there were 6 rows and 89 rows.

# Task 4

Remove columns with missing data.

grades\_delcolumn = grades %>% select(Prefix, Assignment)   
str(grades\_delcolumn)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 99 obs. of 2 variables:  
## $ Prefix : num 5 8 8 7 8 7 8 7 8 7 ...  
## $ Assignment: num 57.1 95 83.7 81.2 91.3 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Prefix = col\_double(),  
## .. Assignment = col\_double(),  
## .. Tutorial = col\_double(),  
## .. Midterm = col\_double(),  
## .. TakeHome = col\_double(),  
## .. Final = col\_double()  
## .. )

After deleting columns that contained any missing values, there were 2 columns and 99 rows.

# Task 5

Is row-wise or column-wise deletion of missing preferable in this case? I think column-wise deletion should not be done here. We deleted the varible we’re trying to predict! The variable with the most missing values was Final and fewer than 5% of observations were missing Final grades. This seems like a very small amount of cases that were missing and it does not justify throwing out the 95% of cases that we did have data for.

# Task 6 and 7

Impute missing values in grades.

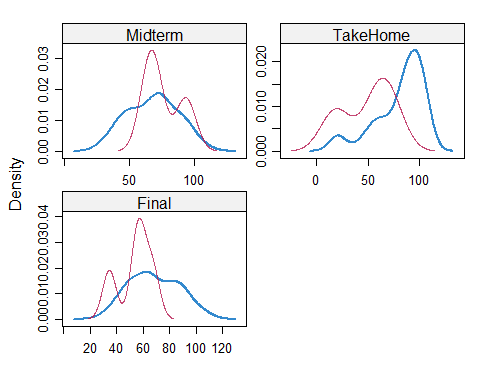
gradesnoprefix = grades %>% select(-Prefix)  
grades\_imp = mice(gradesnoprefix, m=1, method = "pmm", seed = 12345)

##   
## 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:  
## Assignment Tutorial Midterm TakeHome Final   
## "" "pmm" "pmm" "pmm" "pmm"   
## PredictorMatrix:  
## Assignment Tutorial Midterm TakeHome Final  
## Assignment 0 1 1 1 1  
## Tutorial 1 0 1 1 1  
## Midterm 1 1 0 1 1  
## TakeHome 1 1 1 0 1  
## Final 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)

## Assignment Tutorial Midterm TakeHome   
## Min. : 28.14 Min. : 34.09 Min. : 28.12 Min. : 16.91   
## 1st Qu.: 80.88 1st Qu.: 84.69 1st Qu.: 52.81 1st Qu.: 63.98   
## Median : 89.94 Median : 93.64 Median : 69.38 Median : 87.96   
## Mean : 85.49 Mean : 89.87 Mean : 67.94 Mean : 80.16   
## 3rd Qu.: 95.00 3rd Qu.:100.55 3rd Qu.: 81.88 3rd Qu.: 98.42   
## Max. :100.83 Max. :112.58 Max. :110.00 Max. :108.89   
## Final   
## Min. : 28.06   
## 1st Qu.: 52.91   
## Median : 65.56   
## Mean : 67.64   
## 3rd Qu.: 83.19   
## Max. :108.89

summary(grades\_delrow)

## Prefix Assignment Tutorial Midterm   
## Min. :4.000 Min. : 28.14 Min. : 34.09 Min. : 28.12   
## 1st Qu.:7.000 1st Qu.: 82.45 1st Qu.: 82.93 1st Qu.: 52.50   
## Median :8.000 Median : 89.94 Median : 92.93 Median : 70.00   
## Mean :7.303 Mean : 85.97 Mean : 89.40 Mean : 68.14   
## 3rd Qu.:8.000 3rd Qu.: 95.05 3rd Qu.:100.58 3rd Qu.: 83.12   
## Max. :8.000 Max. :100.83 Max. :112.58 Max. :110.00   
## TakeHome Final   
## Min. : 16.91 Min. : 28.06   
## 1st Qu.: 64.07 1st Qu.: 52.50   
## Median : 87.96 Median : 65.56   
## Mean : 81.03 Mean : 67.97   
## 3rd Qu.: 97.78 3rd Qu.: 83.06   
## Max. :108.89 Max. :108.89

I compared the summary of the dataset with imputed values to the summary of the dataset where rows with missing data were removed and the summary statistics are very similar, so I think imputing is fine to do here with the mice function. I couldn’t tell much from the density plot, because only TakeHome had a distribution that seemed to have a the same shape for the original and imputed data.

I don’t think I’d be comfortable imputing more than 5% of the values in any variable. Probably there is some best practice guidelines to follow regarding what percent of values are okay to impute. In the Data Cleaning reading there was good guidance on what should be considered an outlier (add 1.5% of the interquartile to the third quartile or subtracting 1.5% of the interquartile range from the first quartile). Impututing just feels like shaky ground. I think in a lot of cases, I’d rather omit the observations entirely because when data is missing, there is often a reason. Here, if a test grade is missing, wouldn’t it be a zero? Or wouldn’t the instructor maybe let another grade count more? So instead of imputing TakeHome, if an instructor let Midterm count more, then TakeHome should equal the Midterm grade. I also think if any data is imputed, it should be stated in results.