Mod4\_Assign1-Quiz1\_MissingData

Grace Williams

2023-06-06

# Libraries: For this assignment you will need the following libraries: tidyverse, VIM, and mice.

# Before beginning the assignment tasks, you should read-in the data for the assignment into a data frame called grades. This data contains grade information from an engineering course. The dataset was originally used to investigate how student performance in the course would be predictive of student grades on the “Final” exam. The “Prefix” column is a surrogate for enrollment year in the engineering program. Smaller values imply older (more mature?) students.

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(VIM)

## Loading required package: colorspace  
## Loading required package: grid  
## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,  
## which was just loaded, will retire in October 2023.  
## Please refer to R-spatial evolution reports for details, especially  
## https://r-spatial.org/r/2023/05/15/evolution4.html.  
## It may be desirable to make the sf package available;  
## package maintainers should consider adding sf to Suggests:.  
## The sp package is now running under evolution status 2  
## (status 2 uses the sf package in place of rgdal)  
## VIM is ready to use.  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues  
##   
## Attaching package: 'VIM'  
##   
## The following object is masked from 'package:datasets':  
##   
## sleep

library(mice)

##   
## Attaching package: 'mice'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(readr)  
library(ggplot2)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(naniar)  
library(skimr)

##   
## Attaching package: 'skimr'  
##   
## The following object is masked from 'package:naniar':  
##   
## n\_complete

grades <- read\_csv("class-grades.csv")

## Rows: 99 Columns: 6  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (6): Prefix, Assignment, Tutorial, Midterm, TakeHome, Final  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Task 1: For which variables is there missing data?

is.na(grades)

## Prefix Assignment Tutorial Midterm TakeHome Final  
## [1,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [5,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [6,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [7,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [8,] FALSE FALSE FALSE TRUE TRUE FALSE  
## [9,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [10,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [11,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [12,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [13,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [14,] FALSE FALSE TRUE FALSE FALSE FALSE  
## [15,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [16,] FALSE FALSE FALSE TRUE FALSE FALSE  
## [17,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [18,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [19,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [20,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [21,] FALSE FALSE FALSE FALSE FALSE TRUE  
## [22,] FALSE FALSE FALSE FALSE TRUE FALSE  
## [23,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [24,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [25,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [26,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [27,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [28,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [29,] FALSE FALSE FALSE FALSE FALSE TRUE  
## [30,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [31,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [32,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [33,] FALSE FALSE FALSE FALSE TRUE FALSE  
## [34,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [35,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [36,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [37,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [38,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [39,] FALSE FALSE FALSE FALSE FALSE TRUE  
## [40,] FALSE FALSE FALSE TRUE FALSE FALSE  
## [41,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [42,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [43,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [44,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [45,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [46,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [47,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [48,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [49,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [50,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [51,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [52,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [53,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [54,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [55,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [56,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [57,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [58,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [59,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [60,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [61,] FALSE FALSE FALSE FALSE FALSE TRUE  
## [62,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [63,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [64,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [65,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [66,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [67,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [68,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [69,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [70,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [71,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [72,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [73,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [74,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [75,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [76,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [77,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [78,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [79,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [80,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [81,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [82,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [83,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [84,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [85,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [86,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [87,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [88,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [89,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [90,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [91,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [92,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [93,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [94,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [95,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [96,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [97,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [98,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [99,] FALSE FALSE FALSE FALSE FALSE FALSE

skim(grades)

Data summary

|  |  |
| --- | --- |
| Name | grades |
| Number of rows | 99 |
| Number of columns | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

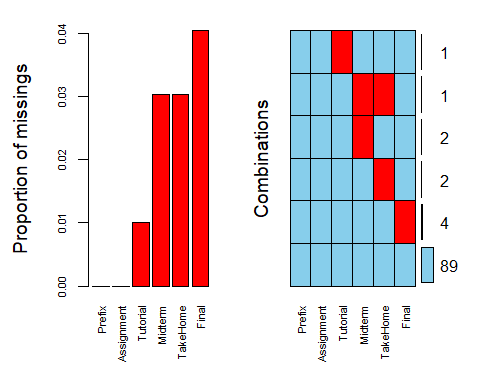
**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Prefix | 0 | 1.00 | 7.31 | 0.93 | 4.00 | 7.00 | 8.00 | 8.00 | 8.00 | ▁▁▁▅▇ |
| Assignment | 0 | 1.00 | 85.49 | 12.60 | 28.14 | 80.88 | 89.94 | 95.00 | 100.83 | ▁▁▁▅▇ |
| Tutorial | 1 | 0.99 | 89.79 | 15.14 | 34.09 | 83.93 | 93.37 | 100.56 | 112.58 | ▁▂▁▇▆ |
| Midterm | 3 | 0.97 | 67.70 | 19.43 | 28.12 | 52.50 | 69.38 | 81.56 | 110.00 | ▃▅▇▅▂ |
| TakeHome | 3 | 0.97 | 81.12 | 23.95 | 16.91 | 69.90 | 88.42 | 99.07 | 108.89 | ▁▁▂▅▇ |
| Final | 4 | 0.96 | 68.23 | 18.82 | 28.06 | 52.92 | 66.11 | 83.61 | 108.89 | ▂▇▇▇▂ |

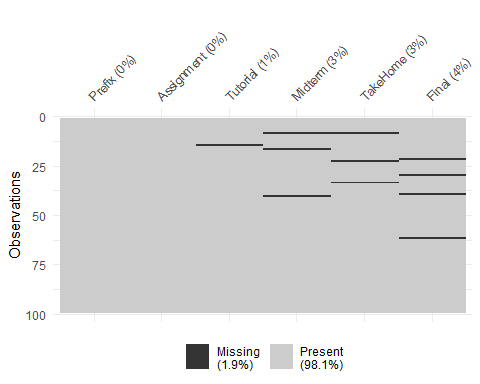
## Task 2: Use the VIM package to visualize missingness. Are there any students that are missing multiple pieces of data?

## Yes

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



vis\_miss(grades)



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

grades\_rowdel = grades %>% drop\_na()

## 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?

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

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

### With column deletion we lose whole variables, row deletion only removes portions of the variables that have missing data.

## Task 6 Use the “mice” package to impute the missing data in this dataset. Set “m = 5” and “method =”pmm”” in the imputation. You MUST use a random number seed of 123 for this imputation. What is the mean value for the “Final” variable after imputation?

### 68.33

set.seed(123)  
imp\_final = mice(grades, m=5, method='pmm', printFlag=FALSE)  
summary(imp\_final)

## Class: mids  
## Number of multiple imputations: 5   
## 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

final\_complete = complete(imp\_final)  
summary(final\_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.: 52.81   
## Median :8.000 Median : 89.94 Median : 93.10 Median : 70.00   
## Mean :7.313 Mean : 85.49 Mean : 89.76 Mean : 68.31   
## 3rd Qu.:8.000 3rd Qu.: 95.00 3rd Qu.:100.55 3rd Qu.: 82.81   
## 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.: 53.33   
## Median : 88.89 Median : 66.11   
## Mean : 80.80 Mean : 68.33   
## 3rd Qu.: 99.07 3rd Qu.: 83.61   
## Max. :108.89 Max. :108.89