

# Ejercicio 3.3

Silvia Pineda

## Carga de Datos y Librerías

```
library(naniar)
library(tidyverse)

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr    1.3.1
v purrr    1.1.0

-- Conflicts -----
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become non-conflicting
```

```
library(rio)

data <- read.csv("students_FP.csv",
  na.strings = c("", "NA", "NaN", "NULL"),
  stringsAsFactors = TRUE
)
```

## Imputación múltiple

```
library(mice)
```

```
Attaching package: 'mice'
```

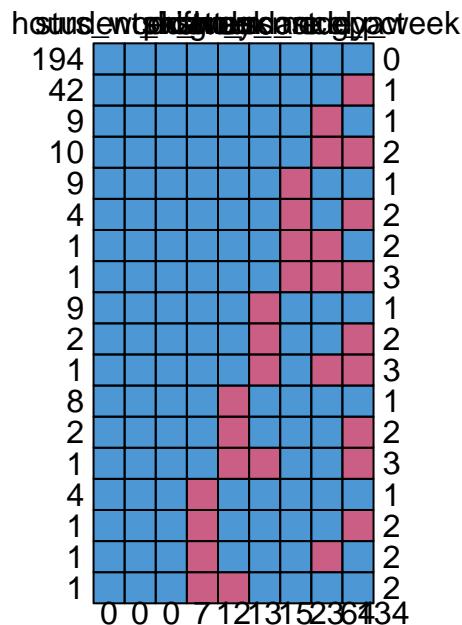
```
The following object is masked from 'package:stats':
```

```
filter
```

```
The following objects are masked from 'package:base':
```

```
cbind, rbind
```

```
md.pattern(data)
```



|     | student_id | hours_work_week | shift | program | exam_score | study_mode |
|-----|------------|-----------------|-------|---------|------------|------------|
| 194 | 1          |                 | 1     | 1       | 1          | 1          |
| 42  | 1          |                 | 1     | 1       | 1          | 1          |
| 9   | 1          |                 | 1     | 1       | 1          | 1          |
| 10  | 1          |                 | 1     | 1       | 1          | 1          |
| 9   | 1          |                 | 1     | 1       | 1          | 1          |
| 4   | 1          |                 | 1     | 1       | 1          | 1          |
| 1   | 1          |                 | 1     | 1       | 1          | 1          |

|     |                |                  |     |     |    |   |
|-----|----------------|------------------|-----|-----|----|---|
| 1   | 1              | 1                | 1   | 1   | 1  | 1 |
| 9   | 1              | 1                | 1   | 1   | 1  | 0 |
| 2   | 1              | 1                | 1   | 1   | 1  | 0 |
| 1   | 1              | 1                | 1   | 1   | 1  | 0 |
| 8   | 1              | 1                | 1   | 1   | 0  | 1 |
| 2   | 1              | 1                | 1   | 1   | 0  | 1 |
| 1   | 1              | 1                | 1   | 1   | 0  | 0 |
| 4   | 1              | 1                | 1   | 0   | 1  | 1 |
| 1   | 1              | 1                | 1   | 0   | 1  | 1 |
| 1   | 1              | 1                | 1   | 0   | 1  | 1 |
| 1   | 1              | 1                | 1   | 0   | 0  | 1 |
| 0   | 0              | 0                | 7   | 12  | 13 |   |
|     | attendance_pct | hours_study_week | gpa |     |    |   |
| 194 | 1              | 1                | 1   | 0   |    |   |
| 42  | 1              | 1                | 0   | 1   |    |   |
| 9   | 1              | 0                | 1   | 1   |    |   |
| 10  | 1              | 0                | 0   | 2   |    |   |
| 9   | 0              | 1                | 1   | 1   |    |   |
| 4   | 0              | 1                | 0   | 2   |    |   |
| 1   | 0              | 0                | 1   | 2   |    |   |
| 1   | 0              | 0                | 0   | 3   |    |   |
| 9   | 1              | 1                | 1   | 1   |    |   |
| 2   | 1              | 1                | 0   | 2   |    |   |
| 1   | 1              | 0                | 0   | 3   |    |   |
| 8   | 1              | 1                | 1   | 1   |    |   |
| 2   | 1              | 1                | 0   | 2   |    |   |
| 1   | 1              | 1                | 0   | 3   |    |   |
| 4   | 1              | 1                | 1   | 1   |    |   |
| 1   | 1              | 1                | 0   | 2   |    |   |
| 1   | 1              | 0                | 1   | 2   |    |   |
| 1   | 1              | 1                | 1   | 2   |    |   |
| 15  |                | 23               | 64  | 134 |    |   |

No hay ninguna observación con todas las variables missing.

```
impData <- mice(select(data,-student_id),m=5,maxit=50,seed=500)
```

| iter | imp | variable         |                |     |            |         |            |  |
|------|-----|------------------|----------------|-----|------------|---------|------------|--|
| 1    | 1   | hours_study_week | attendance_pct | gpa | exam_score | program | study_mode |  |
| 1    | 2   | hours_study_week | attendance_pct | gpa | exam_score | program | study_mode |  |
| 1    | 3   | hours_study_week | attendance_pct | gpa | exam_score | program | study_mode |  |











```

44  4 hours_study_week attendance_pct gpa exam_score program study_mode
44  5 hours_study_week attendance_pct gpa exam_score program study_mode
45  1 hours_study_week attendance_pct gpa exam_score program study_mode
45  2 hours_study_week attendance_pct gpa exam_score program study_mode
45  3 hours_study_week attendance_pct gpa exam_score program study_mode
45  4 hours_study_week attendance_pct gpa exam_score program study_mode
45  5 hours_study_week attendance_pct gpa exam_score program study_mode
46  1 hours_study_week attendance_pct gpa exam_score program study_mode
46  2 hours_study_week attendance_pct gpa exam_score program study_mode
46  3 hours_study_week attendance_pct gpa exam_score program study_mode
46  4 hours_study_week attendance_pct gpa exam_score program study_mode
46  5 hours_study_week attendance_pct gpa exam_score program study_mode
47  1 hours_study_week attendance_pct gpa exam_score program study_mode
47  2 hours_study_week attendance_pct gpa exam_score program study_mode
47  3 hours_study_week attendance_pct gpa exam_score program study_mode
47  4 hours_study_week attendance_pct gpa exam_score program study_mode
47  5 hours_study_week attendance_pct gpa exam_score program study_mode
48  1 hours_study_week attendance_pct gpa exam_score program study_mode
48  2 hours_study_week attendance_pct gpa exam_score program study_mode
48  3 hours_study_week attendance_pct gpa exam_score program study_mode
48  4 hours_study_week attendance_pct gpa exam_score program study_mode
48  5 hours_study_week attendance_pct gpa exam_score program study_mode
49  1 hours_study_week attendance_pct gpa exam_score program study_mode
49  2 hours_study_week attendance_pct gpa exam_score program study_mode
49  3 hours_study_week attendance_pct gpa exam_score program study_mode
49  4 hours_study_week attendance_pct gpa exam_score program study_mode
49  5 hours_study_week attendance_pct gpa exam_score program study_mode
50  1 hours_study_week attendance_pct gpa exam_score program study_mode
50  2 hours_study_week attendance_pct gpa exam_score program study_mode
50  3 hours_study_week attendance_pct gpa exam_score program study_mode
50  4 hours_study_week attendance_pct gpa exam_score program study_mode
50  5 hours_study_week attendance_pct gpa exam_score program study_mode

```

```
summary(impData)
```

Class: mids

Number of multiple imputations: 5

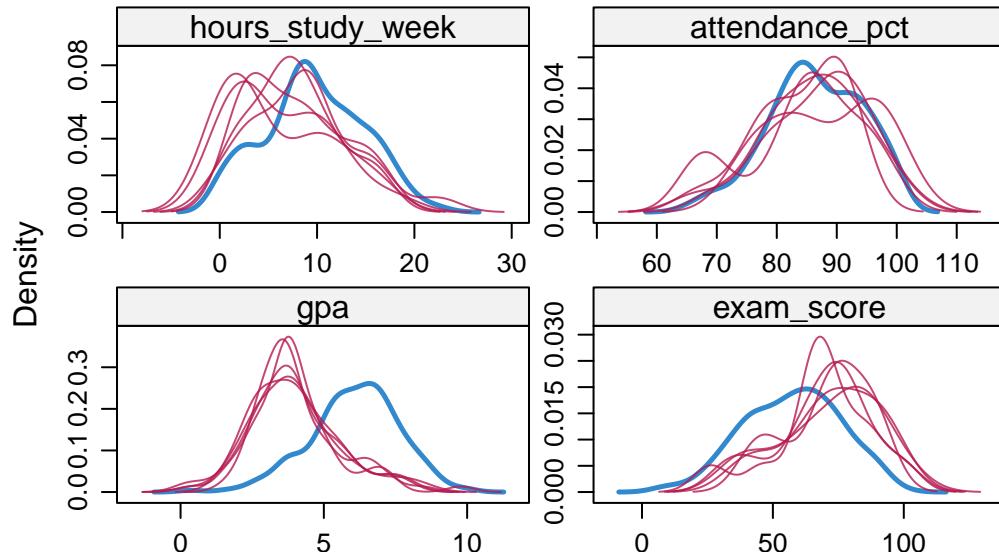
Imputation methods:

| hours_work_week | hours_study_week | attendance_pct | gpa   |
|-----------------|------------------|----------------|-------|
| ""              | "pmm"            | "pmm"          | "pmm" |
| exam_score      | program          | study_mode     | shift |
| "pmm"           | "polyreg"        | "polyreg"      | ""    |

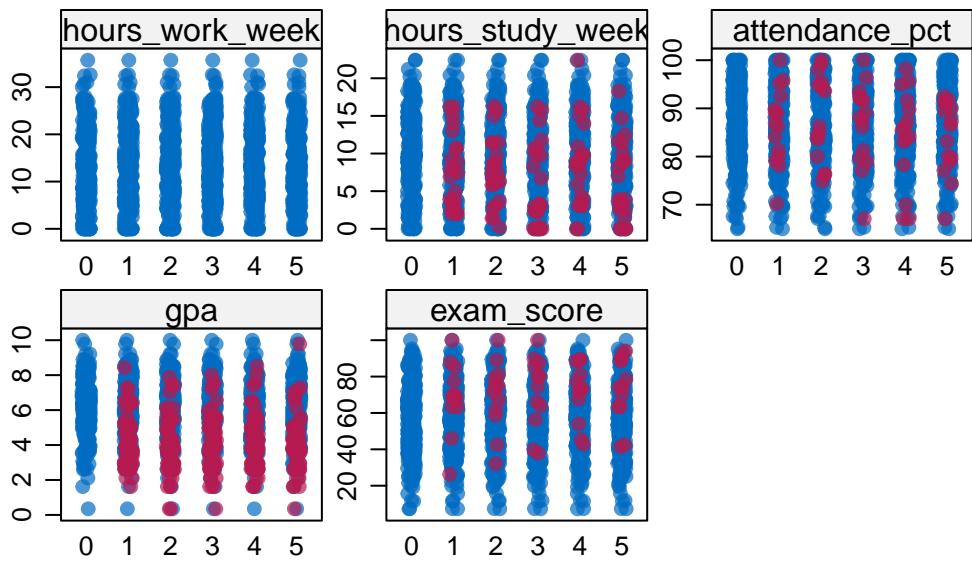
PredictorMatrix:

|                  | hours_work_week | hours_study_week | attendance_pct | gpa | exam_score |
|------------------|-----------------|------------------|----------------|-----|------------|
| hours_work_week  | 0               | 1                | 1              | 1   | 1          |
| hours_study_week | 1               | 0                | 1              | 1   | 1          |
| attendance_pct   | 1               | 1                | 0              | 1   | 1          |
| gpa              | 1               | 1                | 1              | 0   | 1          |
| exam_score       | 1               | 1                | 1              | 1   | 0          |
| program          | 1               | 1                | 1              | 1   | 1          |
|                  | program         | study_mode       | shift          |     |            |
| hours_work_week  | 1               | 1                | 1              |     |            |
| hours_study_week | 1               | 1                | 1              |     |            |
| attendance_pct   | 1               | 1                | 1              |     |            |
| gpa              | 1               | 1                | 1              |     |            |
| exam_score       | 1               | 1                | 1              |     |            |
| program          | 0               | 1                | 1              |     |            |

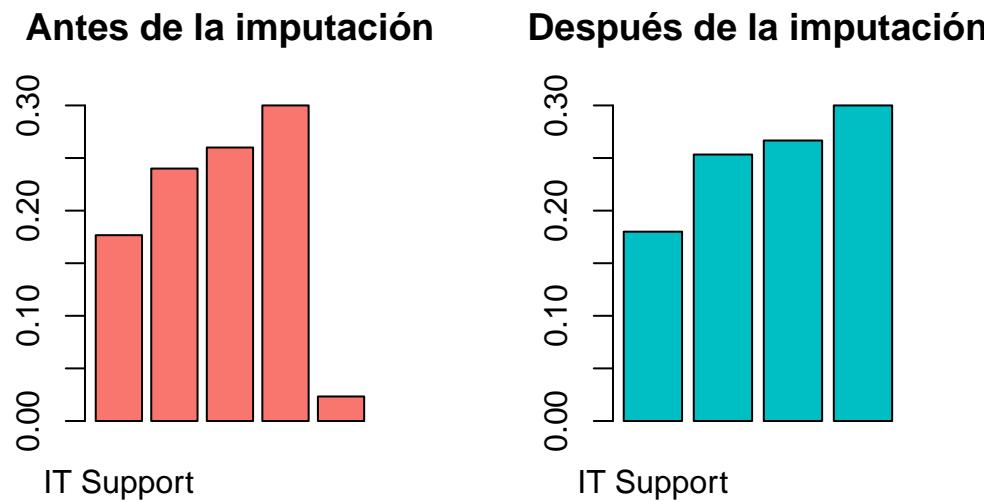
```
###Visualizar las variables cuantitativas  
densityplot(impData)
```



```
stripplot(impData, pch = 20, cex = 1.2)
```

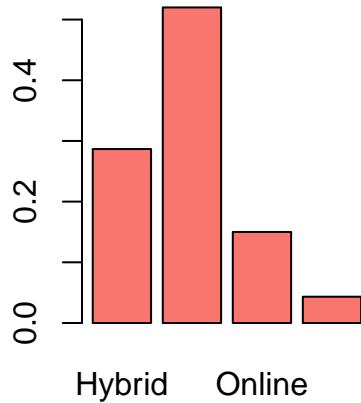


```
###Visualizar las variables cualitativas
completedData <- complete(impData,1)
par(mfrow = c(1, 2))
barplot(prop.table(table(data$program, useNA = "ifany")),
       main = "Antes de la imputación", col = "#F8766D" )
barplot(prop.table(table(completedData$program)),
       main = "Después de la imputación", col = "#00BFC4")
```

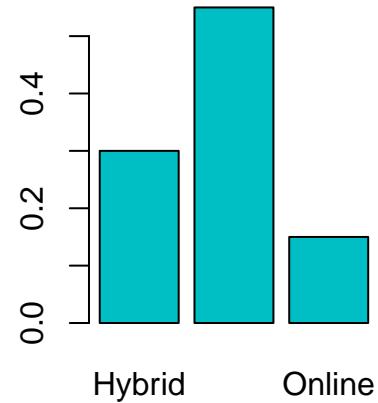


```
par(mfrow = c(1, 2))
barplot(prop.table(table(data$study_mode, useNA = "ifany")),
        main = "Antes de la imputación", col = "#F8766D" )
barplot(prop.table(table(completedData$study_mode)),
        main = "Después de la imputación", col = "#00BFC4")
```

**Antes de la imputación**



**Después de la imputación**



Excepto la variable attendance\_pct, ninguna variable se ha imputado demasiado bien, pero la que se ha imputado muy mal es la de gpa que todos los valores se han imputado por debajo del resto de la distribución, esto no necesariamente está mal, precisamente una hipótesis era que los datos faltantes de gpa era MNAR porque podrías estar relacionados con valores bajos de gpa, pero no explicado por el resto de variables. En este caso, al tener en cuenta de forma multivariante toda la base de datos, ha imputado los valores de la variable como bajos desplazando así toda la distribución.

Por otro lado las dos variables cualitativas parece que se imputa de forma correcta.