

# Ejercicio 3.1: Detección de datos perdidos

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   1.6.0
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr    1.3.1
v purrr    1.2.0

-- Conflicts ----- tidyverse_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
)
```

## 1. Cuantifica el número de datos perdidos

```
summary(data)
```

```

student_id  hours_work_week hours_study_week attendance_pct
SID1000: 1   Min.    : 0.00   Min.    : 0.000   Min.    : 65.00
SID1001: 1   1st Qu.: 7.00   1st Qu.: 6.700   1st Qu.: 81.50
SID1002: 1   Median  :12.65   Median  : 9.400   Median  : 86.30
SID1003: 1   Mean    :12.77   Mean    : 9.682   Mean    : 86.45
SID1004: 1   3rd Qu.:18.65   3rd Qu.:13.200   3rd Qu.: 92.60
SID1005: 1   Max.    :35.70   Max.    :22.400   Max.    :100.00
(Other):294                               NA's    :23      NA's    :15

      gpa          exam_score           program       study_mode
Min.   : 0.350   Min.   : 7.30   IT Support      :53   Hybrid   : 86
1st Qu.: 5.202   1st Qu.: 42.70  Network Administration:72  On-campus:156
Median  : 6.210   Median  : 58.20  Software Engineering :78   Online    : 45
Mean    : 6.075   Mean    : 57.48  Web Development   :90   NA's     : 13
3rd Qu.: 7.093   3rd Qu.: 71.22  NA's            : 7
Max.   :10.000   Max.   :100.00
NA's    :64        NA's   :12

      shift
Afternoon:119
Evening   : 73
Morning   :108

```

```
# Porcentaje total de datos faltantes
pct_miss(data)
```

[1] 4.962963

```
# Número de datos faltantes en todo el dataset
n_miss(data)
```

[1] 134

```
# Porcentaje de filas completas si borramos todas las filas que contienen al menos un dato faltante
pct_complete_case(data)
```

[1] 64.66667

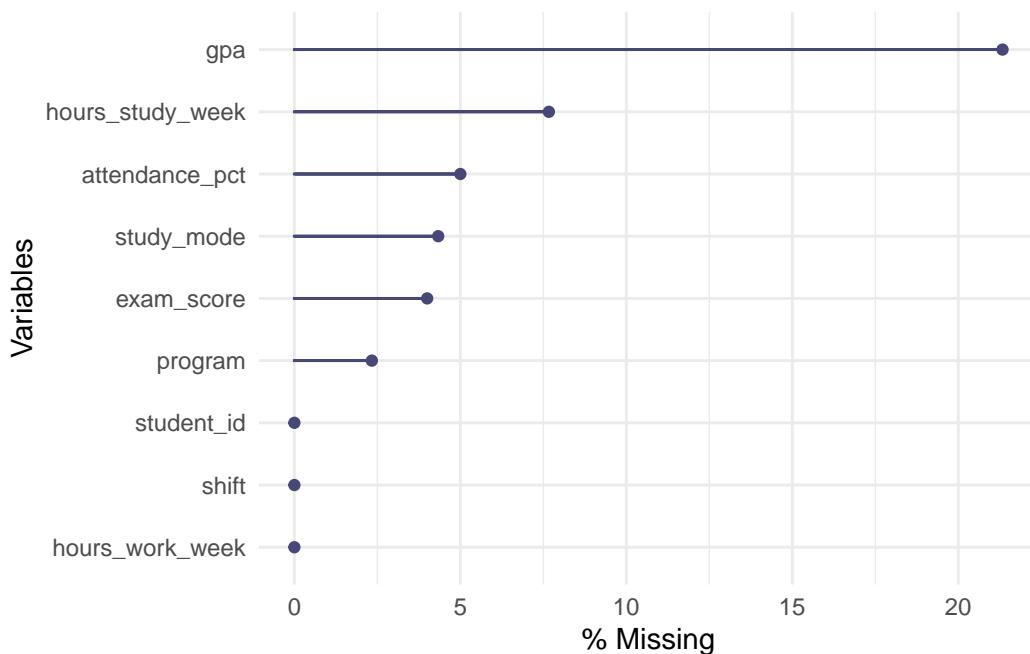
```
# Número de filas completas si borramos todas las filas que contienen al menos un dato faltante
n_case_complete(data)
```

```
[1] 194
```

El % total de datos faltantes es de 5.9% que hacen un total de 134 celdas faltantes, pero si lo contabilizamos en filas completas, sólo nos quedarían 194 observaciones que hacen el 65% del total de la base de datos que son 300 observaciones, por tanto, debemos hacer un tratamiento de datos faltantes para poder analizar la base de datos actual.

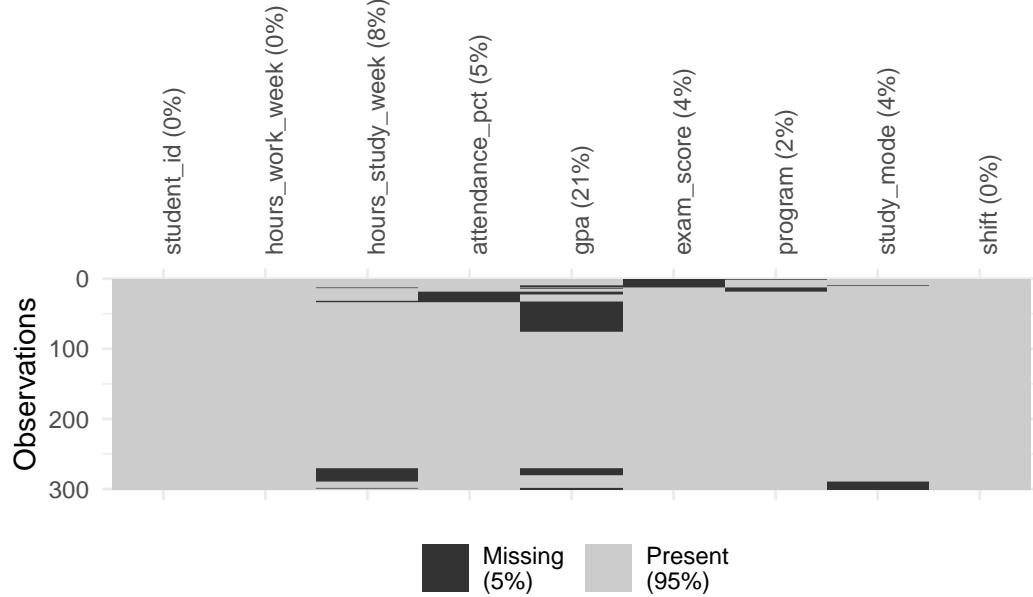
## 2. Visualiza los datos perdidos de forma global

```
gg_miss_var(data, show_pct = TRUE)
```



Tenemos 6 variables de un total de 9 con datos faltantes, de las cuales `gpa` es la que mayor % tiene con alrededor de un 20%, luego `hours_study_week` tiene casi un 10% y finalmente `exam_score`, `attendance_pct`, `study_mode` y `program` con un 5% aproximadamente.

```
vis_miss(data, cluster=TRUE) +  
  theme(axis.text.x = element_text(angle = 90))
```

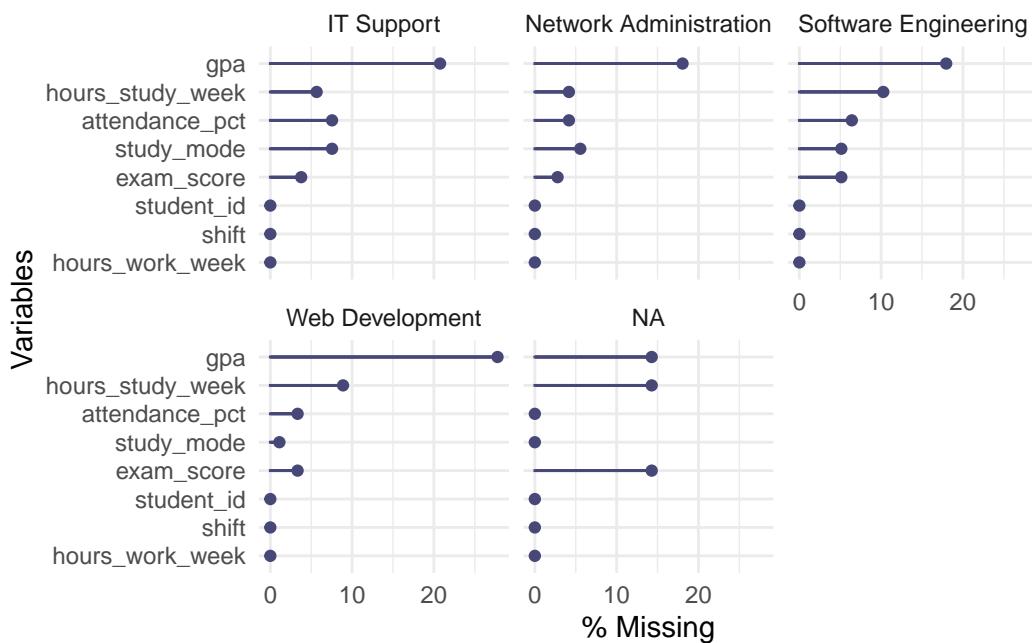


Las variables que parecen que clusterizan los datos faltantes podrían indicar MAR/MNAR y las que no se ve tan claro podrían ser MCAR.

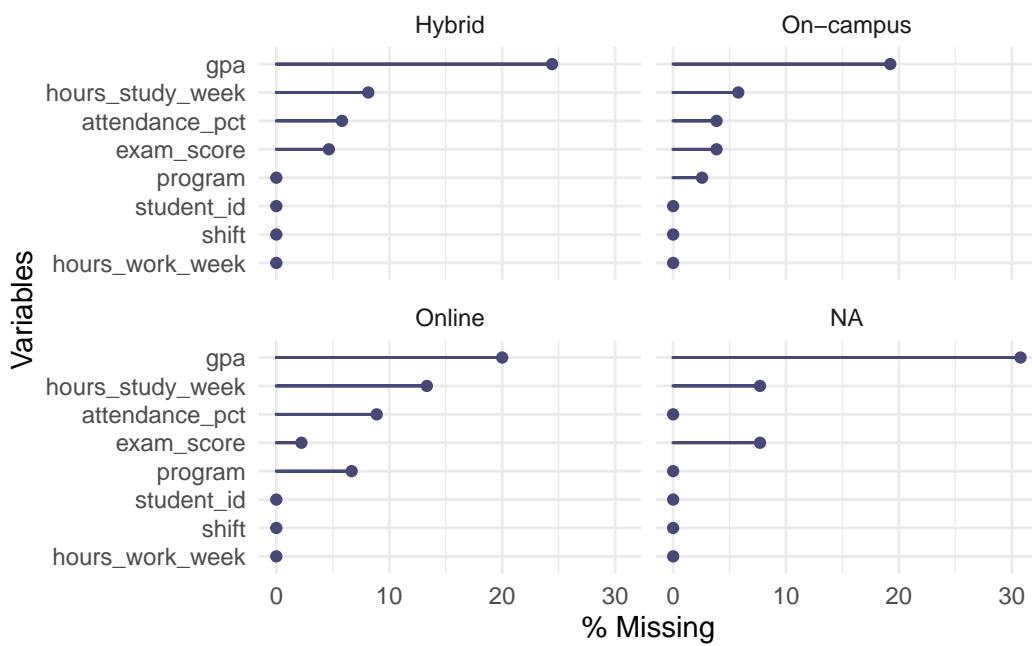
### 3. Clasifica en MAR, MNAR y MCAR los datos perdidos.

Primero visualizamos por subgrupos de las variables cualitativas

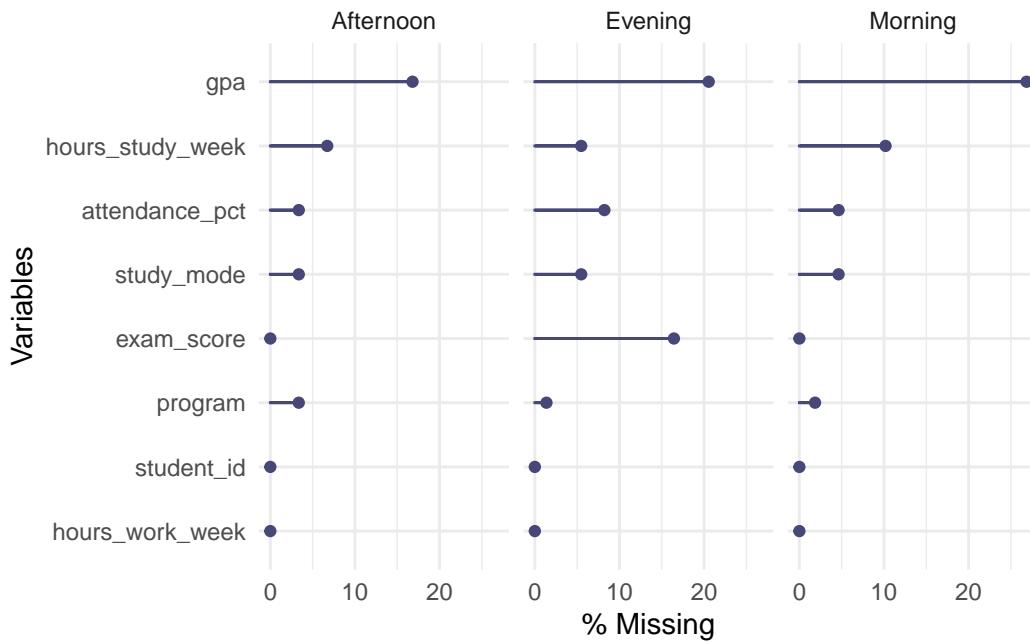
```
gg_miss_var(data, show_pct = TRUE, facet = program)
```



```
gg_miss_var(data, show_pct = TRUE, facet = study_mode)
```



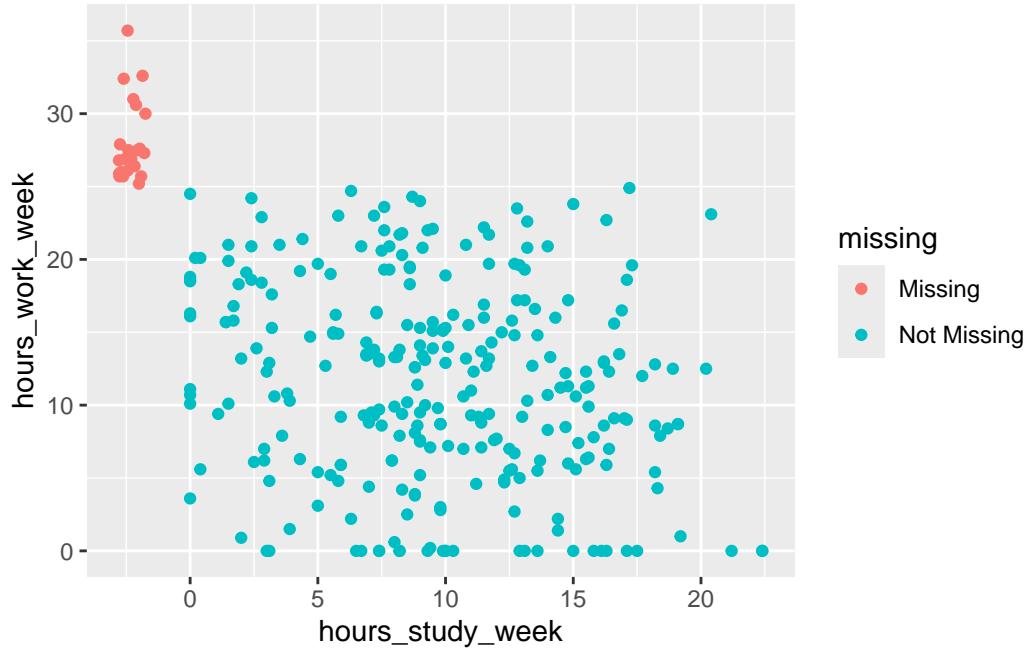
```
gg_miss_var(data, show_pct = TRUE, facet = shift)
```



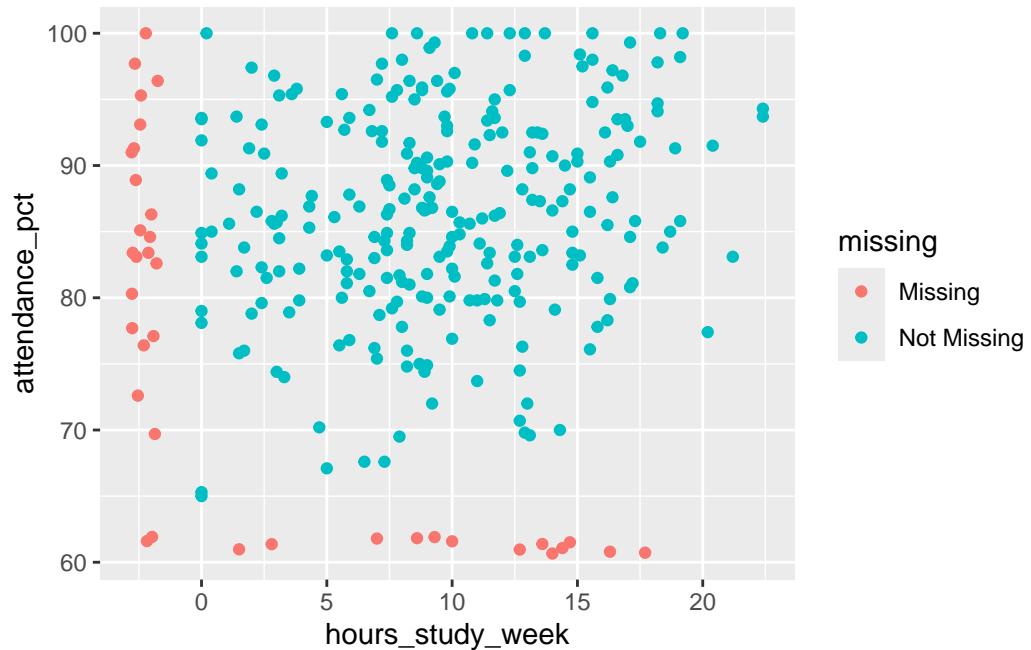
- Los NA de **exam\_score** corresponden al turno de evening, aunque hay gente en el turno de evening que si tiene dato. Podría ser un **MAR**, si por ejemplo es un profesor que todavía no ha subido las notas. Otra cosa es que en ese turno los profesores no hayan subido las notas de aquellos que han suspendido, entonces si sería un **MNAR** porque es algo no observable.

## Segundo visualizamos por las variables cuantitativas

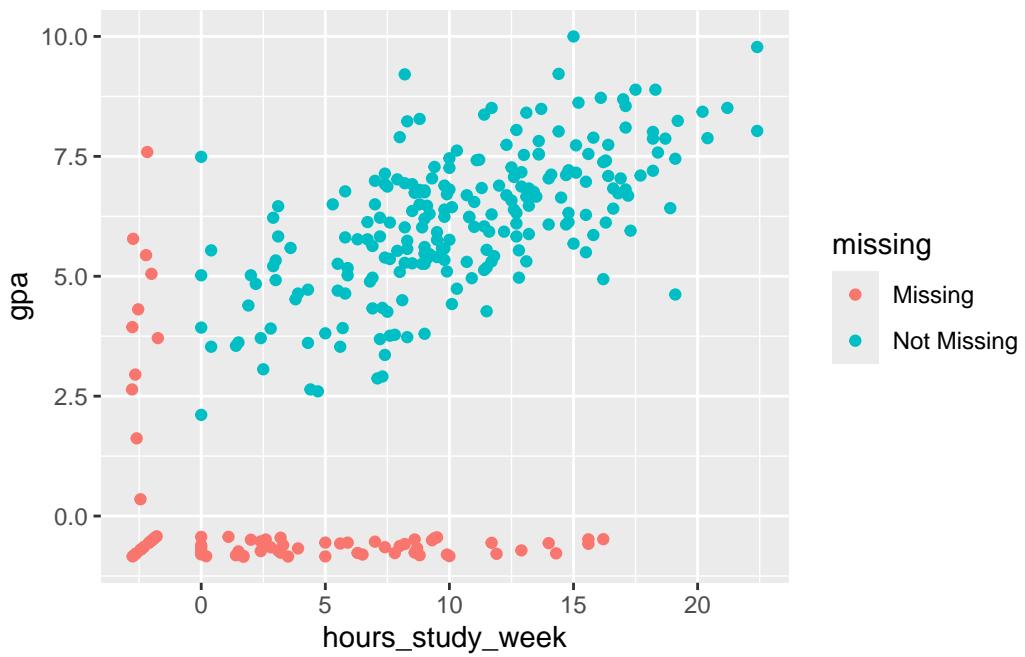
```
ggplot(data = data, aes (x = hours_study_week ,  
y = hours_work_week )) + geom_miss_point()
```



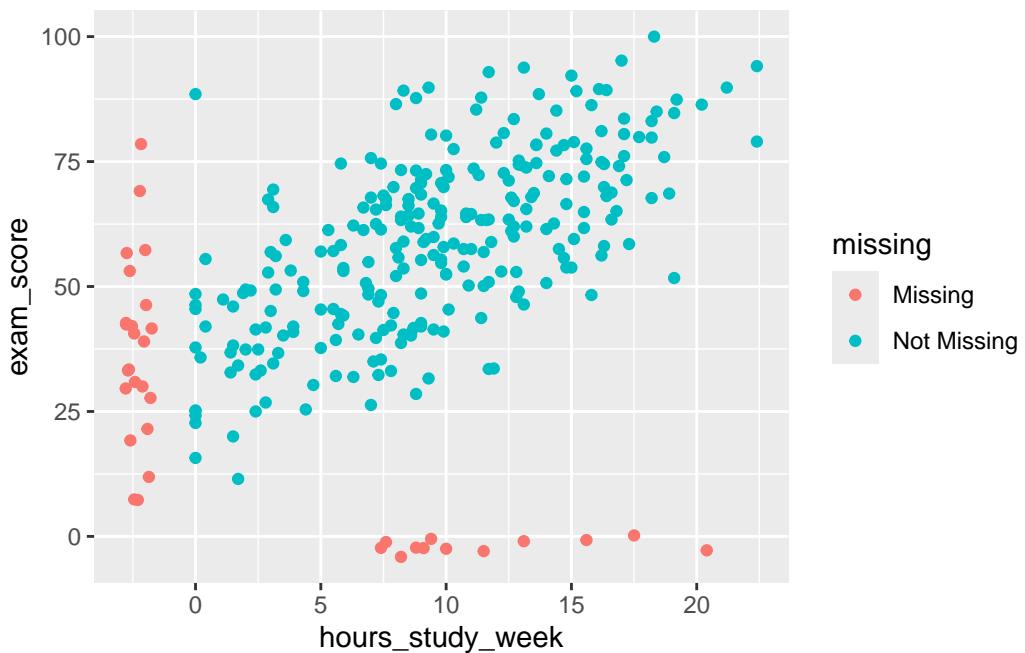
```
ggplot(data = data, aes (x = hours_study_week ,
y = attendance_pct )) + geom_miss_point()
```



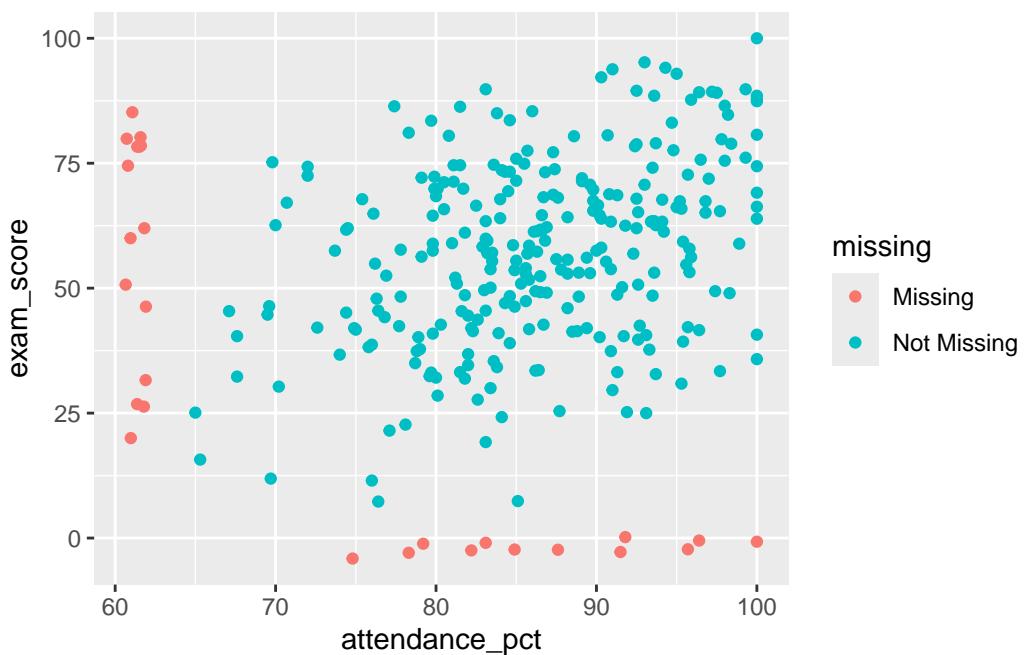
```
ggplot(data = data, aes (x = hours_study_week ,  
y =gpa )) + geom_miss_point()
```



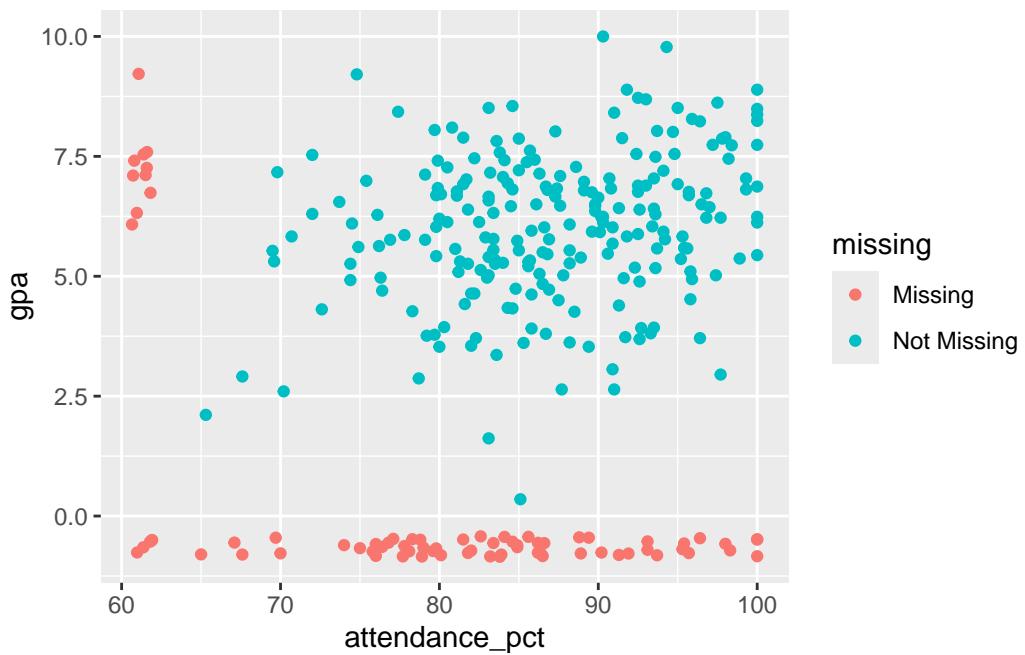
```
ggplot(data = data, aes (x = hours_study_week ,  
y =exam_score )) + geom_miss_point()
```



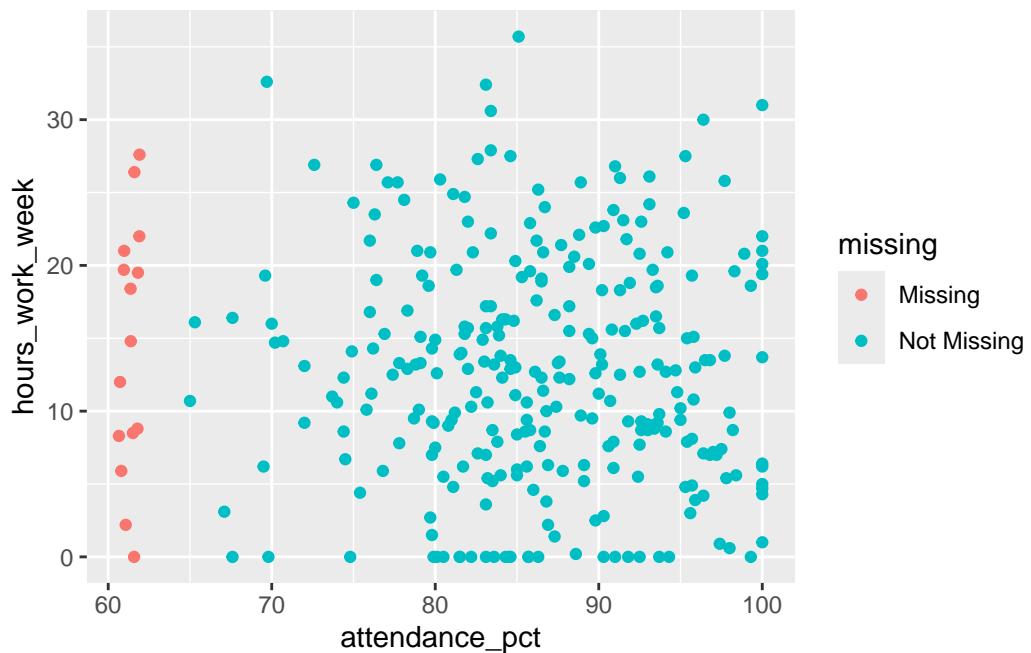
```
ggplot(data = data, aes (x = attendance_pct ,
y =exam_score )) + geom_miss_point()
```



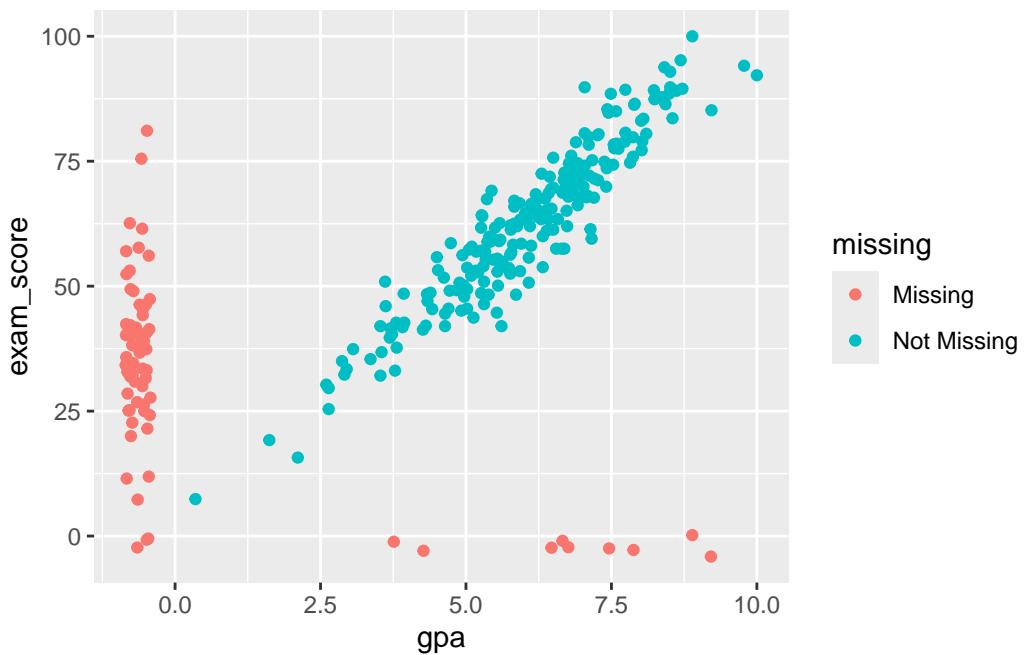
```
ggplot(data = data, aes (x = attendance_pct ,  
y =gpa )) + geom_miss_point()
```



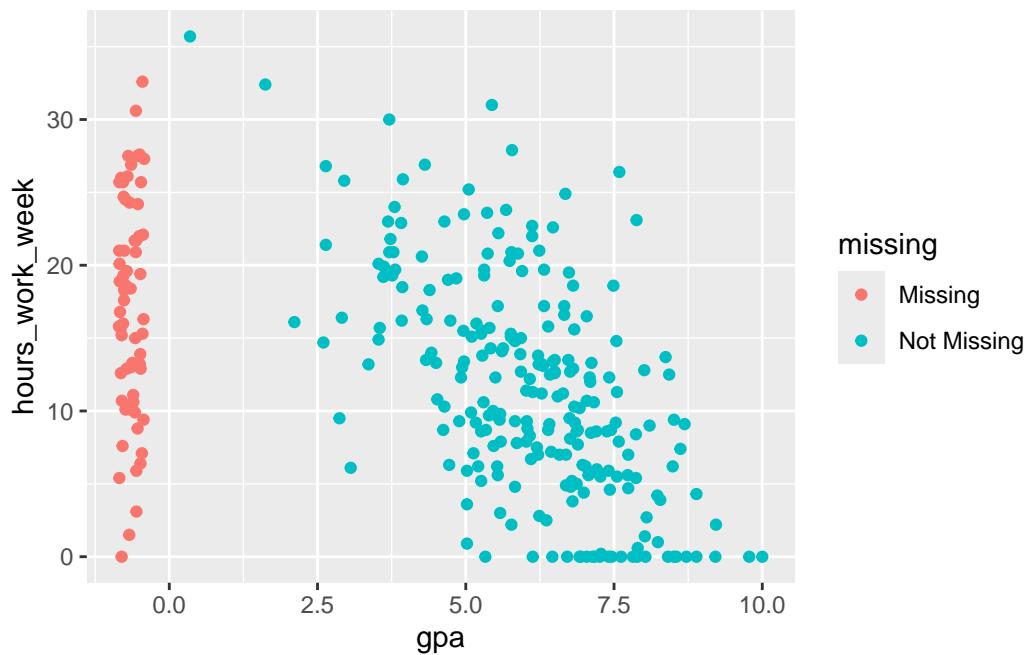
```
ggplot(data = data, aes (x = attendance_pct ,  
y = hours_work_week )) + geom_miss_point()
```



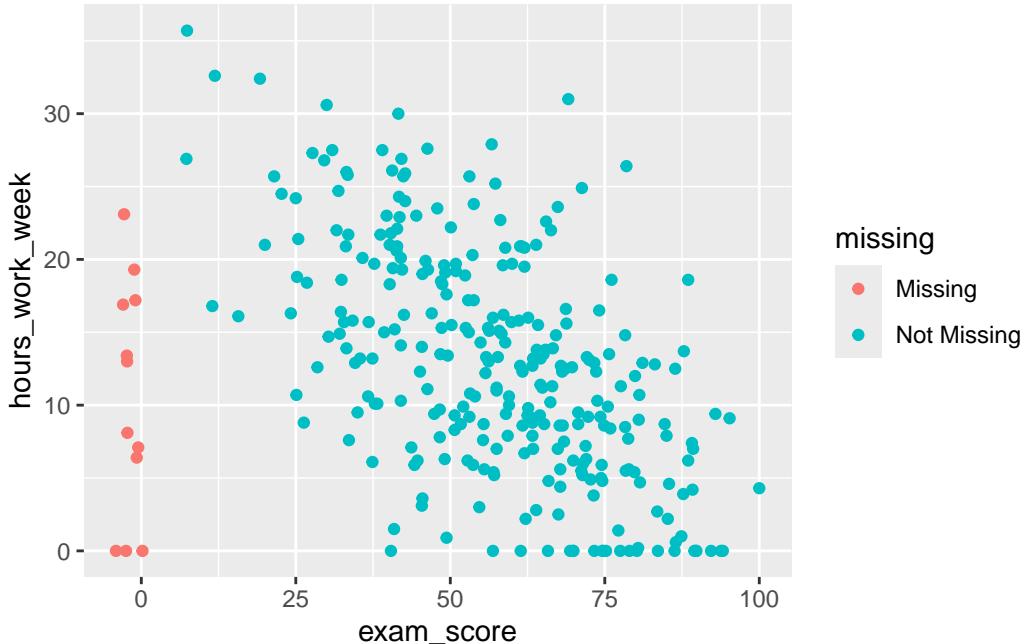
```
ggplot(data = data, aes (x = gpa ,
y = exam_score )) + geom_miss_point()
```



```
ggplot(data = data, aes (x = gpa ,  
y = hours_work_week )) + geom_miss_point()
```



```
ggplot(data = data, aes (x = exam_score ,  
y = hours_work_week )) + geom_miss_point()
```

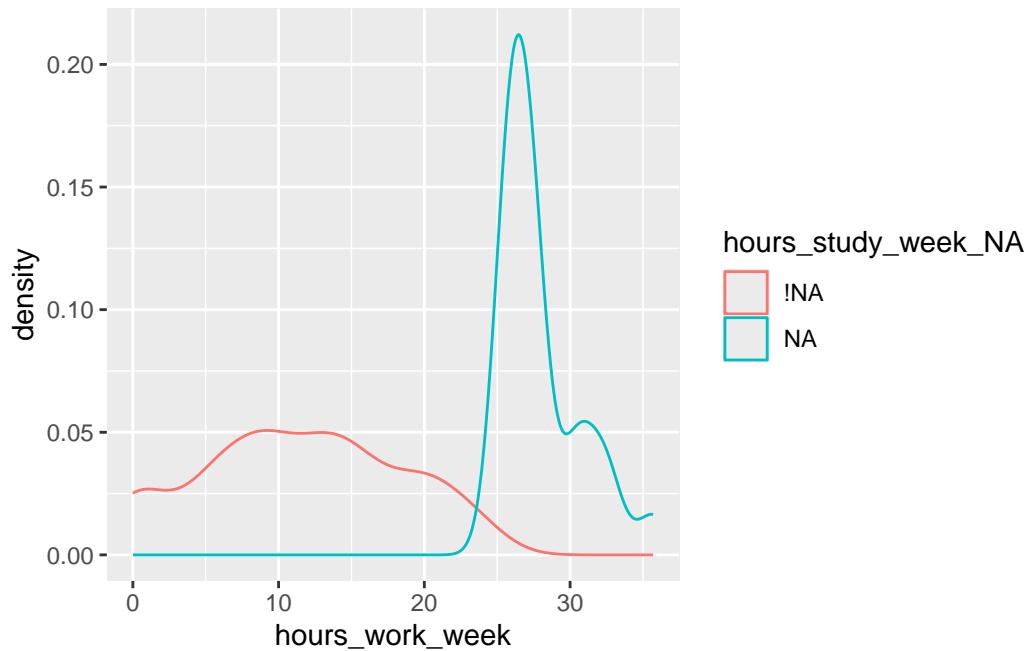


- Los NA de `hours_study_week` corresponden a valores bajos de `exam_score`, valores bajos de `gpa` y valores  $> 25$  para `hours_work_week`, esta asociación corresponde a variables observadas (rendimiento y horas de trabajo) y por tanto corresponde a un **MAR**.
- Los NA de `attendance_pct` corresponden a valores altos de `gpa`, pero no sabemos cuál podría ser la razón, por ejemplo exención de asistencia por programa de honores, prácticas externas), por tanto sería indicativo de un **MNAR**.
- Los NA de `gpa` corresponden mayoritariamente a valores con datos inferiores de `hours_study_week` y con datos menores de `exam_score`. Por tanto, este patrón podría corresponder a un patrón **MNAR** ya que quizás corresponda a valores bajos de `gpa`.

### Tercero, con las variables shadow podemos verificar

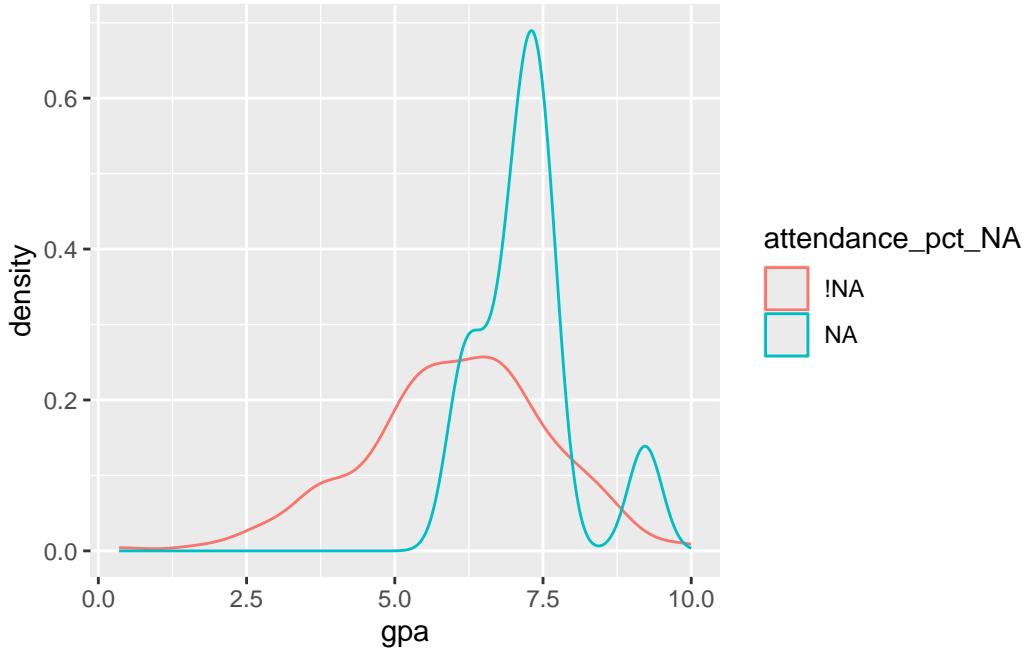
```
shadowed_data <- data %>%
bind_shadow()

ggplot (data = shadowed_data,
       mapping = aes(x = hours_work_week,
                     colour = hours_study_week_NA)) + geom_density()
```



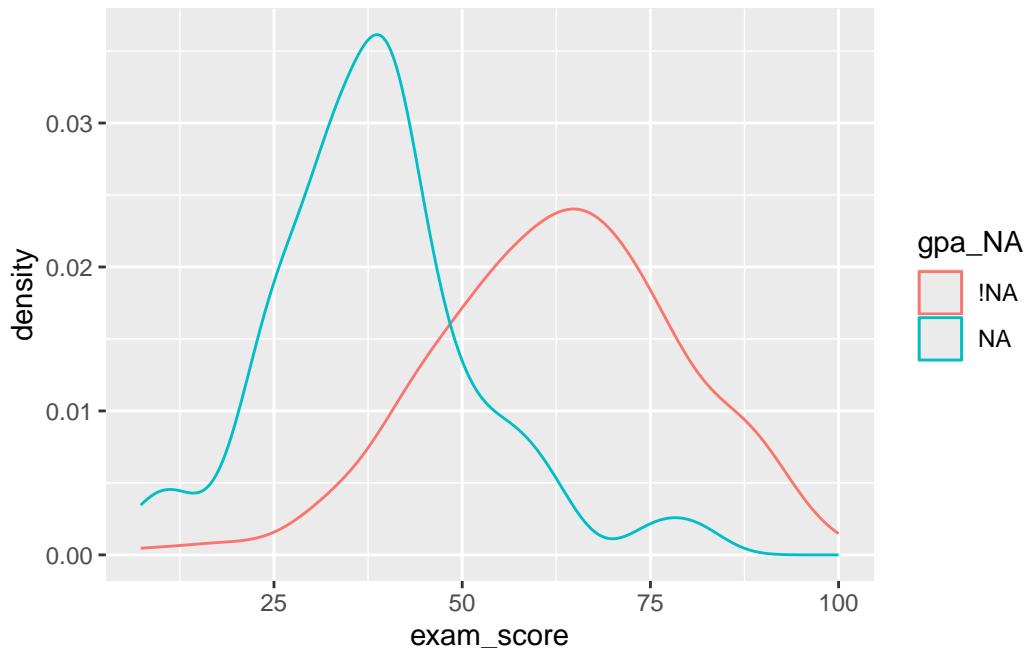
```
ggplot (data = shadowed_data,
        mapping = aes(x = gpa,
                      colour = attendance_pct_NA)) + geom_density()
```

Warning: Removed 64 rows containing non-finite outside the scale range  
(`stat\_density()`).

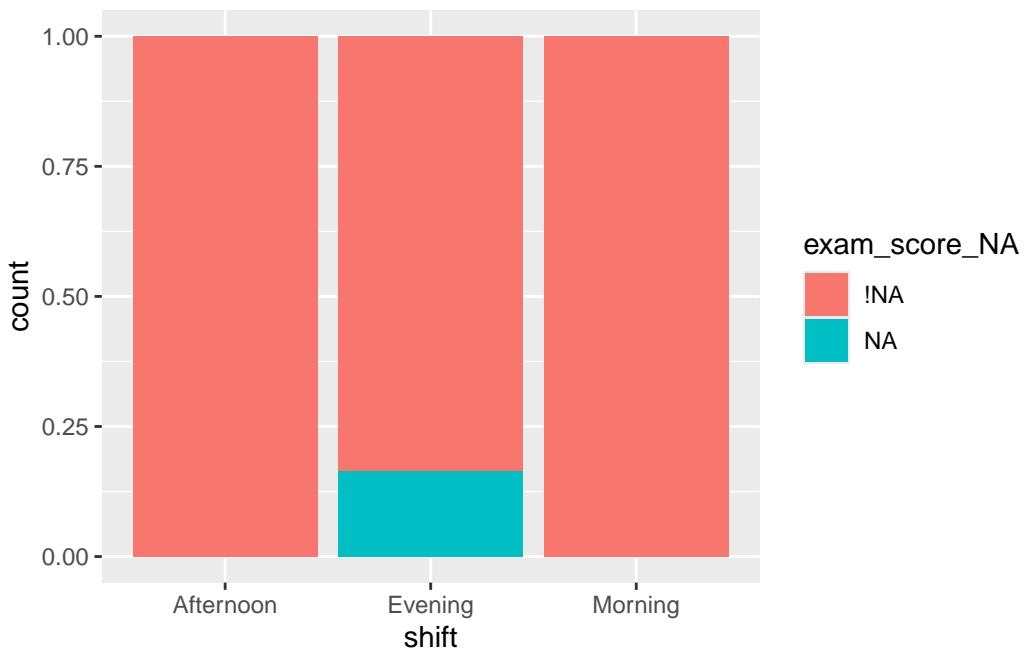


```
ggplot (data = shadowed_data,
        mapping = aes(x = exam_score ,
                      colour = gpa_NA)) + geom_density()
```

Warning: Removed 12 rows containing non-finite outside the scale range  
(`stat\_density()`).



```
ggplot(shadowed_data, aes(x = shift, fill = exam_score_NA)) +  
  geom_bar(position = "fill")
```



En el primero se ve un MAR claro, y en el segundo y tercero podría ser más indicativo de MNAR.

Los NA de `program`, `study_mode` asumiremos que son **MCAR**.

Los NA de `exam_score` son **MAR/MNAR**

Los NA de `hours_study_week` son **MAR**

Los NA de `attendance_pct` y `gpa` son **MNAR**