

▼ Nombre: Celic Gabriel Hernández Archundia

Matrícula: 2877240

```
%%capture  
!pip install rpy2==3.5.1
```

```
%load_ext rpy2.ipython
```

```
from google.colab import drive  
drive.mount('/content/drive', force_remount=True)
```

```
Mounted at /content/drive
```

▼ Mounted at [/content/drive](#)

```
%%R  
install.packages("tidyverse")
```

```
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: =
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: =
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:

WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: downloaded 688 KB

WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:

WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: The downloaded source packages are in
'/tmp/Rtmp8Bnw12/downloaded_packages'
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:
```

```
%%R
```

```
library(readr)
library(magrittr)
library(dplyr)
```

```
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]:
Attaching package: 'dplyr'
```

```
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
WARNING: rpy2.rinterface_lib.callbacks:R[write to console]: The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

```
%%R
```

```
data <- read.csv("drive/MyDrive/Tecmilenio/Big Data/student-mat.csv")
```

```
%%R
```

```
data
```

```

117 2 2 11 13 14
118 5 0 13 14 13
119 5 20 9 7 8
120 4 6 14 13 13
121 1 2 16 15 15
122 5 6 16 14 15
123 5 2 13 13 13
124 5 18 14 11 13
125 5 0 8 7 8
126 5 0 13 13 12
127 1 0 7 10 11
128 5 2 7 8 9
129 4 0 7 4 0
130 4 8 18 18 18
131 5 0 12 0 0
132 4 0 8 0 0
133 5 12 10 13 12
134 5 16 12 11 11
135 5 0 9 0 0
136 5 0 11 0 0
137 5 0 10 0 0
138 5 0 4 0 0
139 5 0 14 12 12
140 5 0 16 16 15
141 3 0 7 9 0
142 2 8 9 9 9
143 5 2 9 11 11
144 1 2 14 14 13
145 12 10

```

%%R

tail(data)

```

      school sex age address famsize Pstatus Medu Fedu      Mjob      Fjob reason
390      MS  F  18      U    GT3      T   1   1    other    other course
391      MS  M  20      U    LE3      A   2   2  services  services course
392      MS  M  17      U    LE3      T   3   1  services  services course
393      MS  M  21      R    GT3      T   1   1    other    other course
394      MS  M  18      R    LE3      T   3   2  services  other course
395      MS  M  19      U    LE3      T   1   1    other    at_home course

      guardian traveltime studytime failures schoolsup famsup paid activities
390    mother          2          2          1        no    no    no          yes
391    other          1          2          2        no    yes  yes          no
392    mother          2          1          0        no    no    no          no
393    other          1          1          3        no    no    no          no
394    mother          3          1          0        no    no    no          no
395    father          1          1          0        no    no    no          no

      nursery higher internet romantic famrel freetime goout Dalc Walc health
390     yes     yes      no      no      1        1      1      1      1      5
391     yes     yes      no      no      5        5      4      4      5      4
392     no     yes     yes     no      2        4      5      3      4      2
393     no     yes     no      no      5        5      3      3      3      3
394     no     yes     yes     no      4        4      1      3      4      5
395     yes     yes     yes     no      3        2      3      3      3      5

      absences G1 G2 G3
390         0  6  5  0
391        11  9  9  9
392         3 14 16 16
393         3 10  8  7
394         0 11 12 10
395         5  8  9  9

```

%%R

length(rownames(data))

[1] 395

```
%%R
length(colnames(data))

[1] 33

%%R
colnames(data)

[1] "school"      "sex"         "age"         "address"     "famsize"
[6] "Pstatus"    "Medu"        "Fedu"        "Mjob"        "Fjob"
[11] "reason"     "guardian"    "traveltime"  "studytime"   "failures"
[16] "schoolsup"  "famsup"      "paid"        "activities"  "nursery"
[21] "higher"     "internet"    "romantic"    "famrel"      "freetime"
[26] "goout"      "Walc"        "Walc"        "health"      "absences"
[31] "G1"         "G2"          "G3"
```

## ▼ Ejercicio 0. Visualización previa de datos

```
%%R
#select(data, famsize, traveltime, studytime)
data %>%
select(1:10, age) %>% # Esta es la forma correcta de escribirlo
head(n=10)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
1	GP	F	18	U	GT3	A	4	4	at_home	teacher
2	GP	F	17	U	GT3	T	1	1	at_home	other
3	GP	F	15	U	LE3	T	1	1	at_home	other
4	GP	F	15	U	GT3	T	4	2	health	services
5	GP	F	16	U	GT3	T	3	3	other	other
6	GP	M	16	U	LE3	T	4	3	services	other
7	GP	M	16	U	LE3	T	2	2	other	other
8	GP	F	17	U	GT3	A	4	4	other	teacher
9	GP	M	15	U	LE3	A	3	2	services	other
10	GP	M	15	U	GT3	T	3	4	other	other

## ▼ Ejercicio 1. Alumnos que salen de casa regularmente ( $\geq 3$ ) y toman regularmente los fines de semana

```
%%R
#FILTER Sirve para depurar valores en columnas
data %>%
filter(goout >= 3 & Walc >= 3) %>%
select(1:9, goout, Walc) %>%
head(n=100)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	goout	Walc
1	GP	M	15	U	LE3	T	4	4	health	3	3
2	GP	M	17	U	GT3	T	3	2	services	5	3
79	GP	M	18	U	GT3	T	2	2	other	4	5
80	GP	M	17	U	LE3	T	4	3	health	5	4
81	GP	M	17	R	LE3	A	4	4	teacher	3	3
82	GP	M	22	U	GT3	T	3	1	services	5	5
83	GP	M	18	R	LE3	T	3	3	other	3	3
84	GP	M	18	U	GT3	T	3	2	services	5	4
85	GP	M	16	U	GT3	T	3	3	at_home	3	3
86	GP	M	18	U	GT3	T	2	1	services	5	5
87	GP	M	18	R	LE3	A	3	4	other	5	4
88	GP	M	17	U	GT3	T	3	1	services	4	4
89	GP	M	18	U	GT3	T	4	2	health	5	3
90	GP	F	19	U	GT3	T	3	3	other	5	3
91	GP	F	18	U	GT3	T	2	3	other	5	3

92	GP	F	17	U	LE3	T	2	2	services	4	3
93	GP	M	18	U	GT3	T	4	4	teacher	4	4
94	GP	M	17	U	LE3	A	4	1	services	4	4
95	GP	M	17	U	LE3	A	3	2	teacher	4	4
96	GP	M	18	U	GT3	T	2	1	services	4	3
97	GP	M	17	U	GT3	T	3	3	health	3	3
98	GP	F	19	U	GT3	T	4	4	health	4	3
99	GP	F	19	U	LE3	T	1	1	at_home	3	3
100	GP	F	17	R	GT3	T	3	4	at_home	4	5

## Ejercicio 2. Alumnos que salen con amigos regularmente ( $\geq 3$ ) y toman regularmente cada semana

```
%%R
data %>%
  filter(goout >= 3 & Dalc >= 3) %>%
  select(1:9, goout, Dalc) %>%
  head(n=100)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	goout	Dalc
1	GP	M	16	U	GT3	T	4	4	teacher	5	5
2	GP	M	15	U	LE3	A	4	2	health	5	3
3	GP	F	15	U	LE3	A	3	3	other	4	4
4	GP	F	16	U	GT3	T	1	1	services	5	5
5	GP	M	15	U	GT3	A	4	4	other	3	5
6	GP	M	16	U	LE3	A	4	4	teacher	3	3
7	GP	M	16	U	GT3	T	4	4	services	5	5
8	GP	M	15	R	GT3	T	4	4	other	5	3
9	GP	M	15	U	GT3	T	3	4	services	5	3
10	GP	F	16	U	LE3	T	1	1	at_home	4	3
11	GP	M	16	U	LE3	T	2	1	at_home	4	3
12	GP	M	16	U	GT3	T	3	3	other	5	4
13	GP	M	17	U	LE3	T	4	3	teacher	4	4
14	GP	M	16	R	GT3	T	4	2	teacher	3	3
15	GP	M	17	U	GT3	T	1	2	at_home	4	4
16	GP	M	16	R	LE3	T	3	3	teacher	4	3
17	GP	F	17	U	GT3	T	3	4	at_home	3	3
18	GP	M	17	U	LE3	T	4				

## Ejercicio 3. Alumnos que salen con amigos y toman regularmente cada día

```
%%R
data %>%
  filter(goout >= 3 & (Walc >=3 | Dalc >=3)) %>%
  select(1:9, goout, Walc, Dalc) %>%
  head(n=100)
```

## Ejercicio 4. ¿Los alumnos del ejercicio 1 son los mismos que los del ejercicio 2? ¿Por qué?

No, no son los mismos porque no todos los alumnos salen y toman el fin de semana o entre semana con regularidad. Habrá unos que toman en ambos, fin de semana y entre semana, pero no es una regla que aplique para todos.

▼ Ejercicio 5. Alumnos que no quieren continuar sus estudios pero sus padres sí cuentan con un alto grado de estudios ( $\geq 3$ )

```
%%R
colnames(data)

%%R
data %>%
  filter(higher == "no" & (Medu >=3 & Fedu >=3)) %>%
  select(1:9, higher, Medu, Fedu)
```

▼ Ejercicio 6. Alumnos que quieren continuar con sus estudios pero sus padres no cuentan con un alto grado de estudios ( $\leq 3$ )

```
%%R
data %>%
  filter(higher == "yes" & (Medu <=3 & Fedu <=3)) %>%
  select(1:9, higher, Medu, Fedu)
```

▼ Ejercicio 7. Alumnos provenientes de una zona rural que no cuentan con internet en casa

```
%%R
colnames(data)

%%R
data %>%
  filter(address == "R" & (internet == "no")) %>%
  select(1:9, address, internet)
```

▼ Ejercicio 8. Promedio de calificaciones G1, G2, G3 (mostrar primeros 50 estudiantes)

```
%%R
colnames(data)

%%R
data %>%
  select(1:9, G1, G2, G3) %>%
  mutate(prom = (G1+G2+G3)/3) %>%
  arrange(desc(prom)) %>%
  head(n = 50)
```

▼ Ejercicio 9. Número de alumnos menores de edad y número de alumnos mayores de edad

```
%%R
data %>%
  select(age) %>%
```

```
filter(age >= 18) %>%
count(age)

# Mayor de edad

  age  n
1  18 82
2  19 24
3  20  3
4  21  1
5  22  1
```

```
%%R
data %>%
filter(age < 18) %>%
rownames() %>%
length() -> menoresEdad
```

```
data %>%
filter(age >= 18) %>%
rownames() %>%
length() -> mayoresEdad
```

```
c(menoresEdad, mayoresEdad)

[1] 284 111
```

```
%%R
data %>%
select(age) %>%
filter(age < 18) %>%
count(age)
```

```
# Menor de edad
```

```
  age  n
1  15 82
2  16 104
3  17 98
```

▼ Ejercicio 10. ¿Qué grupo representa un mayor número de relaciones románticas, los hombres o las mujeres?

```
%%R
data %>%
group_by(sex) %>%
count(romantic) %>%
filter(romantic == "yes")

# A tibble: 2 × 3
# Groups:   sex [2]
  sex  romantic     n
<chr> <chr>   <int>
1 F      yes     79
2 M      yes     53
```

▼ Texto extra

```
%%R
#MUTATE Sirve para crear una nueva columna de datos
```

```
data %>%
select(age, G1, G2,G3) %>% # %>% significa algo así como "después"
mutate(prom =(G1+G2+G3)/3) %>%
filter(prom > 10)

%%R
#ARRANGE Ordena valores de un dataframe
data %>%
select(age, G1)
mutate(prom = (G1+G2+G3)/3) %>%
filter(prom > 10) %>%
arrange(desc(prom))
```

---

✓ 0 s se ejecutó 19:16

