



Módulo	Estadística para Ciencia de Datos
Nombre y apellidos	Richard Douglas Grijalba
Fecha entrega	07 abril 2024

Caso práctico

Para responder, tienes que enviar tu respuesta a través de una tutoría adjuntando un archivo **Word o PDF**. La plantilla puede descargarla en el apartado de RECURSOS.

Caso práctico final

Este es el último caso práctico de este módulo. Para realizarlo, **ha de escogerse una BBDD** para realizar una exploración estadística propia de un científico de datos. Para ello, se pueden seguir los siguientes pasos para desarrollar el proyecto:

PASO UNO

Definir un solo objetivo para el estudio con una BBDD. En este paso, se va a definir un objetivo de estudio (es muy importante que solo sea uno). Se definirá siguiendo estos puntos:

- ¿Qué problema se quiere solucionar con estos datos?
- ¿Qué significan las variables?
- ¿Qué tipo de variables hay?
- Definir un objetivo que ayude a solucionar el problema.

PASO DOS

Exploración de datos:

- Crear los gráficos más apropiados.
- Interpretar los gráficos.
- Encontrar los primeros indicios y sacar las preconclusiones.
- Listar por orden de importancia los indicios que han desvelado los gráficos.

PASO TRES

Ahora es momento de decidir si las preconclusiones son ciertas o no. Apoyarse en la estadística inferencial y del diseño de experimentos.

- Encontrar las técnicas más apropiadas para corroborar las preconclusiones con la ayuda de un mapa.
- Diseñar la metodología de análisis.
- Aplicar esta metodología.
- Resumir los resultados.

****1.1. ¿Qué problema se quiere solucionar con estos datos?****

En este dataset que corresponde a la información del desempeño académico de un grupo de estudiantes, se busca determinar la influencia de ciertos factores sobre sus resultados académicos.

****1.2. ¿Qué significan las variables?****

****Descripción del Dataset****

Este corresponde a un dataset de los resultados académicos de un grupo estudiantil, por lo que tendremos información de una población en la que se busca determinar ciertos resultados y cuáles factores influyen sobre estos. Este dataset puede ser encontrado en [www.kaggle.com / Math Students](https://www.kaggle.com/mathstudents)

****Características****

Características o columnas incluidas en este dataset: Ese conjunto de datos contenga una variedad de características incluye:

- * Escuela : (binaria: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- * Edad del estudiante (numérico: de 15 a 22)
- * Género/ sexo (binario: "F" - femenino o "M" - masculino)
- * Dirección (binario: "U" - urbano o "R" - rural)
- * Famsize (tamaño de la familia) (binario: "LE3" - menor o igual a 3 o "GT3" - mayor que 3)
- * Pstatus (estado de convivencia de los padres) ("T" - viviendo juntos o "A" - separados)
- * Medu (nivel educativo de la madre) (numérico: 0 - ninguna, 1 - educación primaria (4º grado), 2 - de 5º a 9º grado, 3 - educación secundaria o 4 - educación superior)
- * Fedu (nivel educativo del padre) (numérico: 0 - ninguna, 1 - educación primaria (4º grado), 2 - de 5º a 9º grado, 3 - educación secundaria o 4 - educación superior)
- * Mjob (trabajo de la madre) (nominal: "profesor(a)", relacionado con la "salud", "servicios" civiles (por ejemplo, administrativo o policial), "en_casa" u "otro")
- * Fjob (trabajo del padre) (nominal: "profesor(a)", relacionado con la "salud", "servicios" civiles (por ejemplo, administrativo o policial), "en_casa" u "otro")
- * Reason (razón para elegir la escuela)
- * Traveltime (tiempo de viaje a la escuela) (numérico: 1 - <15 min., 2 - de 15 a 30 min., 3 - de 30 min. a 1 hora, o 4 - >1 hora)
- * Studytime (tiempo de estudio semanal) (numérico: 1 - <2 horas, 2 - de 2 a 5 horas, 3 - de 5 a 10 horas, o 4 - >10 horas)
- * Failures (número de fallos en cursos anteriores)
- * Schoolsup (apoyo educativo adicional)
- * Famrel (calidad de las relaciones familiares)
- * Freetime (tiempo libre después de la escuela)
- * Goout (frecuencia de salir con amigos)
- * Dalc (consumo de alcohol en días laborales)
- * Walc (consumo de alcohol en fines de semana)
- * Health (estado de salud actual)
- * Absences (número de ausencias escolares) (numérico: de 0 a 93)
- * G1, G2, G3 (notas en diferentes momentos del año escolar)

****1.3. ¿Qué tipo de variables hay?****

tenemos variables numericas, binomiales (si o no) y del tipo string o categoricas con varias opciones.

Categóricas:

- * sexo
- * dirección
- * tamaño de la familia
- * estado de convivencia de los padres
- * trabajo de la madre
- * trabajo del padre
- * razón para elegir esta escuela
- * tutor
- * apoyo escolar
- * apoyo familiar
- * actividades extracurriculares
- * quiere cursar educación superior
- * internet
- * relacion/romántico

Numéricas:

- * edad
- * Medu (educación de la madre)
- * Fedu (educación del padre)
- * tiempo de viaje
- * tiempo de estudio
- * fracasos
- * relación familiar
- * tiempo libre salir
- * consumo de alcohol en días laborables
- * consumo de alcohol en fines de semana
- * salud
- * ausencias

las cuales se describen estadisticamente mas adelante, por medio de la exploracion de los datos

****1.4. Definir un objetivo que ayude a solucionar el problema.****

Obtener una relacion entre los factores de exito de los estudiantes.

En este caso se determina que cada estudiante tiene las mismas oportunidades para obtener un grado academico o desarrollarse, sin embargo por medio de los resultados estadisticos se buscara determinar que factor puede favorecer a los estudiantes un mejor desempeño academico.

Tenemos estudiantes mujeres y hombres, Puede este factor influir en los resultados? *Se realizara un grafico y tabla por genero.

****Los Estudiantes que dedican mas tiempo de estudio tienen menos perdidas de materias, se australian igual que aquellos que dedican menos tiempo de horas de estudio?****

PASO DOS

****2.Exploración de datos:****

****2.1. Crear los gráficos más apropiados.****

****2.2. Interpretar los gráficos.****

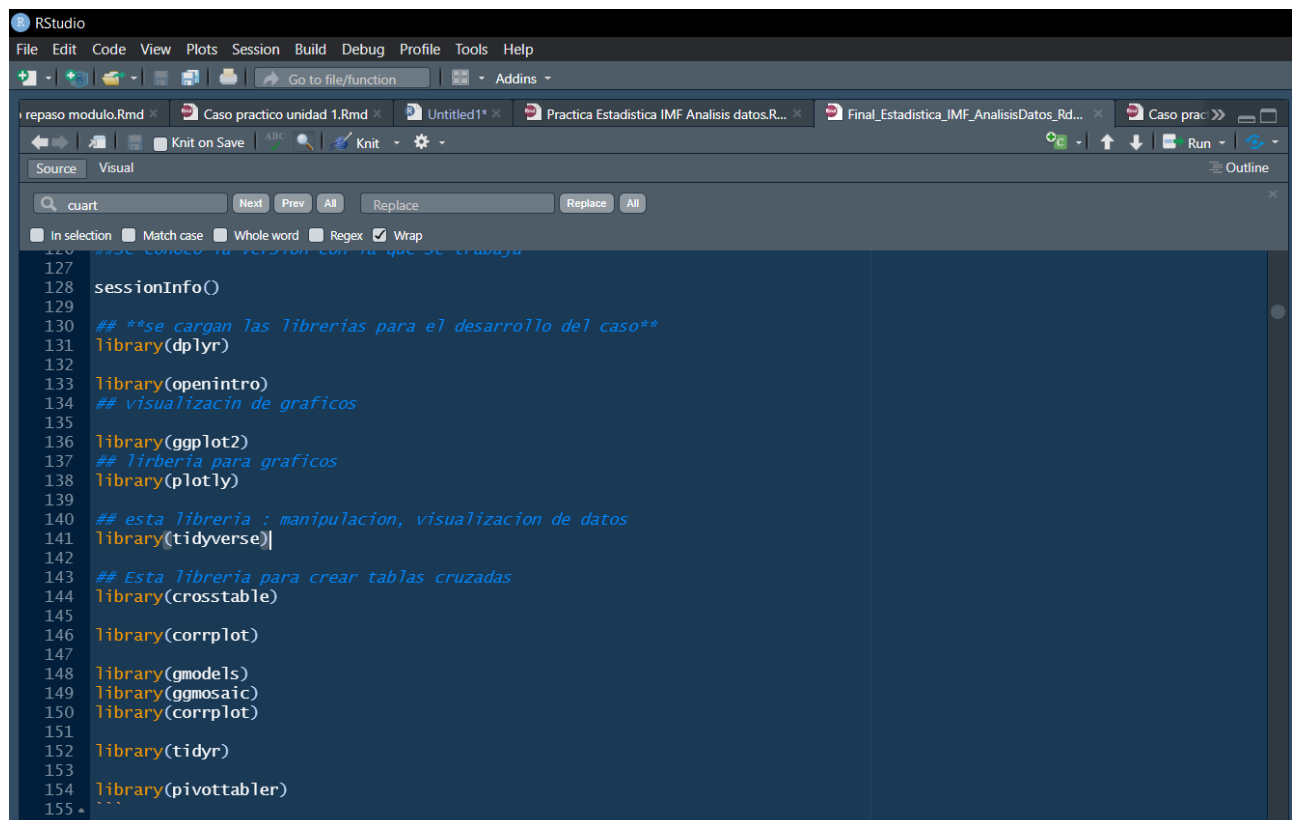
****2.3. Encontrar los primeros indicios y sacar las preconclusiones.****

****2.4. Listar por orden de importancia los indicios que han desvelado los gráficos.****

.

****llamado / carga de las librerías para el desarrollo de este caso****

```
```{r}
```



The screenshot shows the RStudio interface with a code editor window. The code is as follows:

```
126 #se conoce la version con la que se trabaja
127
128 sessionInfo()
129
130 ## **se cargan las librerías para el desarrollo del caso**
131 library(dplyr)
132
133 library(openintro)
134 ## visualización de gráficos
135
136 library(ggplot2)
137 ## librería para gráficos
138 library(plotly)
139
140 ## esta librería : manipulación, visualización de datos
141 library(tidyverse)
142
143 ## Esta librería para crear tablas cruzadas
144 library(crosstable)
145
146 library(corrplot)
147
148 library(gmodels)
149 library(ggmosaic)
150 library(corrplot)
151
152 library(tidyr)
153
154 library(pivottabler)
155
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

reparo modulo.Rmd x Caso practico unidad 1.Rmd x Untitled1\* x Practica Estadística IMF Analisis datos.R... x Final\_Estadística\_IMF\_AnalisisDatos\_Rd... x Caso prac x

Knit on Save Knit Run

Source Visual

Search: quart Next Prev All Replace Replace All

☐ In selection ☐ Match case ☐ Whole word ☐ Regex ☒ Wrap

```

157 ### Exploracion de datos:
158 **Se procede a llamar el dataset previamente descrito, en este caso se realiza la exploracion de los datos de una forma estaditica**
159
160 ```{r}
161 df1 <- read.csv("student-mat.csv", sep=";", stringsAsFactors=T)
162
163
164
165 **Breve descripcion del dataset**
166
167 ```{r}
168 summary(df1)
169 ```

```

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob
GP:349	F:208	Min.:15.0	R: 88	GT3:281	A: 41	Min.:0.000	Min.:0.000	at_home : 59
MS: 46	M:187	1st Qu.:16.0	U:307	LE3:114	T:354	1st Qu.:2.000	1st Qu.:2.000	health : 34
		Median :17.0				Median :3.000	Median :2.000	other :141
		Mean :16.7				Mean :2.749	Mean :2.522	services:103
		3rd Qu.:18.0				3rd Qu.:4.000	3rd Qu.:3.000	teacher : 58
		Max.:22.0				Max.:4.000	Max.:4.000	

Fjob	reason	guardian	traveltime	studytime	failures	schoolsup
at_home : 20	course :145	father: 90	Min.:1.000	Min.:1.000	Min.:0.0000	no :344
health : 18	home :109	mother:273	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:0.0000	yes: 51
other :217	other : 36	other : 32	Median :1.000	Median :2.000	Median :0.0000	
services:111	reputation:105		Mean :1.448	Mean :2.035	Mean :0.3342	
teacher : 29			3rd Qu.:2.000	3rd Qu.:2.000	3rd Qu.:0.0000	
			Max.:4.000	Max.:4.000	Max.:3.0000	

famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime
no :153	no :214	no :194	no : 81	no : 20	no : 66	no :263	Min.:1.000	Min.:1.000
yes:242	yes:181	yes:201	yes:214	yes:275	yes:220	yes:122	1st Qu.:4.000	1st Qu.:2.000

141:19 Chunk 2: R Markdown

## Exploracion general de los datos y dataset

```
Mostrar el número de columnas en df1
num_columnas <- ncol(df1)
print(paste("Número de columnas:", num_columnas))
```

```
[1] "Número de columnas: 33"
```

```
Mostar el número de filas en df1
num_filas <- nrow(df1)
print(paste("Número de filas:", num_filas))
```

```
[1] "Número de filas: 395"
```

```
Mostrar las primeras 10 filas de df1
head_10 <- head(df1, 10)
print(head_10)
```

```
Ahora vamos a ver la estructura del dataframe
str(df1)
```

```
'data.frame': 395 obs. of 33 variables:
$ school : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
$ sex : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
$ age : int 18 17 15 15 16 16 16 17 15 15 ...
$ address : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
$ famsize : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
$ Pstatus : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
$ Medu : int 4 1 1 4 3 4 2 4 3 3 ...
$ Fedu : int 4 1 1 2 3 3 2 4 2 4 ...
$ Mjob : Factor w/ 5 levels "at_home","health",...: 1 1 1 2 3 4 3 3 4 3 ...
$ Fjob : Factor w/ 5 levels "at_home","health",...: 5 3 3 4 3 3 3 5 3 3 ...
$ reason : Factor w/ 4 levels "course","home",...: 1 1 3 2 2 4 2 2 2 2 ...
$ guardian : Factor w/ 3 levels "father","mother",...: 2 1 2 2 1 2 2 2 2 2 ...
$ traveltime : int 2 1 1 1 1 1 1 2 1 1 ...
$ studytime : int 2 2 2 3 2 2 2 2 2 2 ...
$ failures : int 0 0 3 0 0 0 0 0 0 0 ...
$ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
$ famsup : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
$ paid : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
$ activities : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
$ nursery : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
$ higher : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
$ internet : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
$ romantic : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
$ famrel : int 4 5 4 3 4 5 4 4 4 5 ...
$ freetime : int 3 3 3 2 3 4 4 1 2 5 ...
$ goout : int 4 3 2 2 2 2 4 4 2 1 ...
$ Dalc : int 1 1 2 1 1 1 1 1 1 1 ...
$ Walc : int 1 1 3 1 2 2 1 1 1 1 ...
$ health : int 3 3 3 5 5 5 3 1 1 5 ...
$ absences : int 6 4 10 2 4 10 0 6 0 0 ...
$ G1 : int 5 5 7 15 6 15 12 6 16 14 ...
$ G2 : int 6 5 8 14 10 15 12 5 18 15 ...
$ G3 : int 6 6 10 15 10 15 11 6 19 15 ...
```

```
Contar los valores nulos en cada columna del dataframe
num_valores_nulos <- colSums(is.na(df1))
```

```
Imprimir el resultado
print(num_valores_nulos)
```

```
school sex age address famsize Pstatus Medu
0 0 0 0 0 0 0
Fedu Mjob Fjob reason guardian traveltime studytime
0 0 0 0 0 0 0
failures schoolsup famsup paid activities nursery higher
0 0 0 0 0 0 0
internet romantic famrel freetime goout Dalc Walc
0 0 0 0 0 0 0
health absences G1 G2 G3
0 0 0 0 0
```

```
Calculo de Rango Interquartilico en Variables Edad y Ausencias
revision de las variables
summary(df1$age)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
15.0 16.0 17.0 16.7 18.0 22.0
```

```
#rango intercuartil
iqr <- IQR(df1$age)
print(paste("Rango intercuartilico de Edad es de:",iqr))
```

```
[1] "Rango intercuartilico de Edad es de: 2"
```

```
revision de las variables
summary(df1$studytime)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 1.000 2.000 2.035 2.000 4.000
```

```
#rango intercuartil
iqr2 <- IQR(df1$studytime)
print(paste("Rango intercuartilico de Tiempo de Estudio es de:",iqr2))
```

```
[1] "Rango intercuartilico de Tiempo de Estudio es de: 1"
```

```
revision de las variables
CrossTable(df1$age)
```



```
##
##
Cell Contents
|-----|
| N |
| N / Table Total |
|-----|
##
##
Total Observations in Table: 395
##
##
| 15 | 16 | 17 | 18 | 19 |
|-----|-----|-----|-----|-----|
| 82 | 104 | 98 | 82 | 24 |
| 0.208 | 0.263 | 0.248 | 0.208 | 0.061 |
|-----|-----|-----|-----|-----|
##
##
| 20 | 21 | 22 |
|-----|-----|-----|
| 3 | 1 | 1 |
| 0.008 | 0.003 | 0.003 |
|-----|-----|-----|
##
##
##
##
```

```
revision de las variables
summary(df1$sex)
```

```
F M
208 187
```

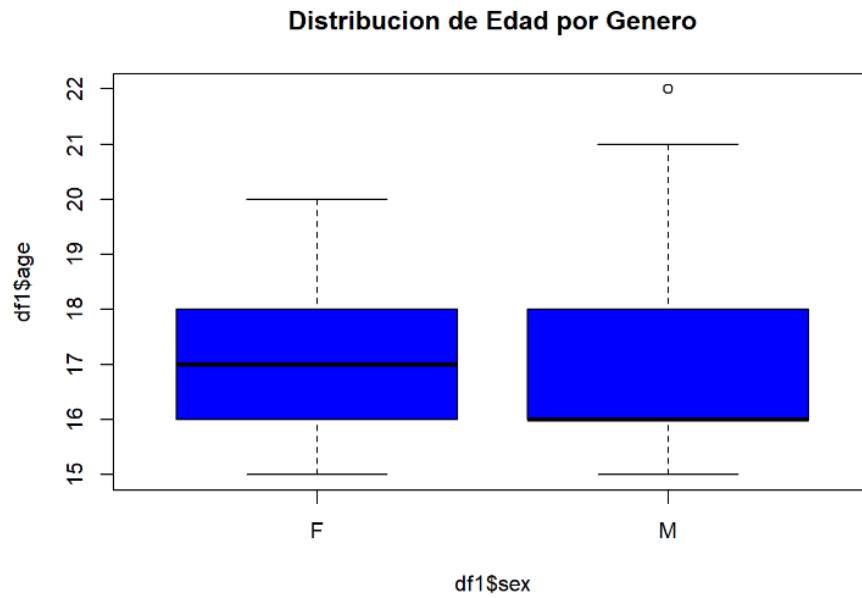
```
revision de las variables
Crear tabla resumen
tabla_resumen1 <- table(df1$school, df1$sex)

Mostrar la tabla resumen
print(tabla_resumen1)
```

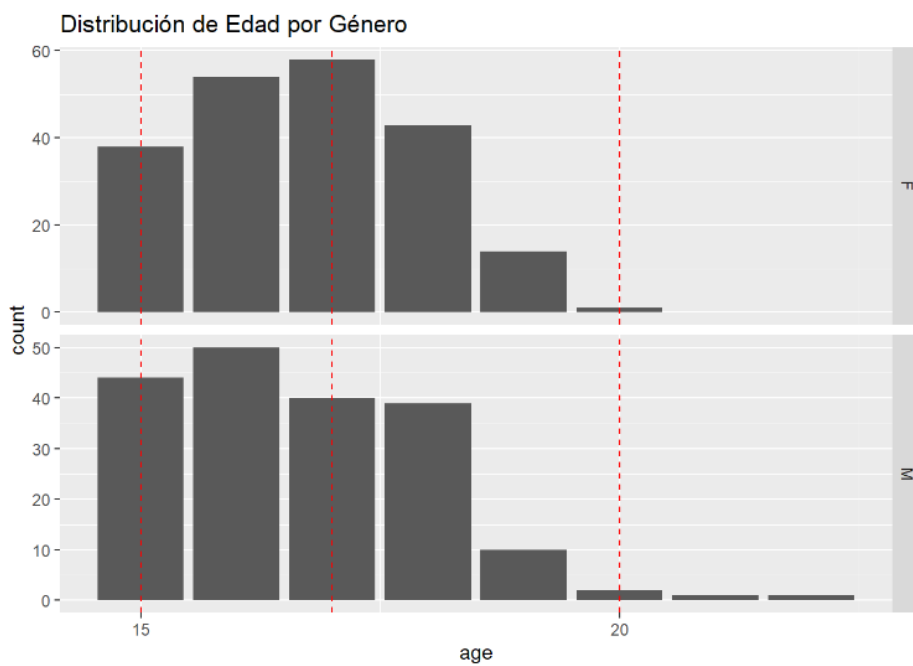
```
##
F M
GP 183 166
MS 25 21
```

## 2.1. Crear los gráficos más apropiados.

```
boxplot(df1$age ~ df1$sex, col = "blue",
main = "Distribucion de Edad por Genero")
```



```
df1 %>%
 ggplot() +
 aes(x = age) +
 geom_bar() +
 geom_vline(xintercept = c(15,17, 20),
 col = "red",
 linetype = "dashed") +
 facet_grid(sex ~ .,
 scales = "free_y") +
 scale_x_continuous(breaks = seq(0, 100, 5)) +
 labs(title = "Distribución de Edad por Género")
```



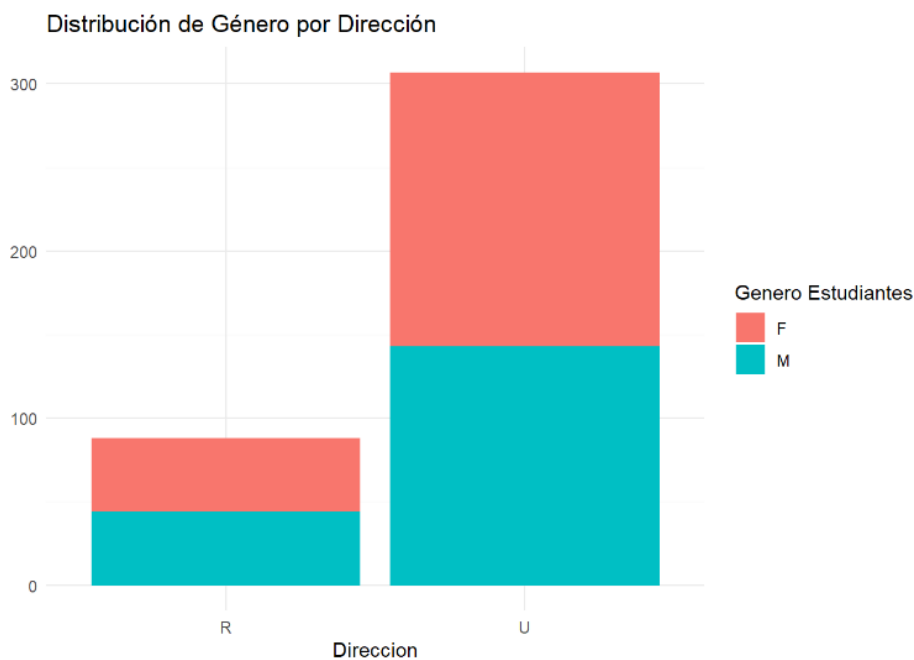
```
tabla_resumen <- table(df1$address, df1$sex)
```

```
Mostrar la tabla resumen
print(tabla_resumen)
```

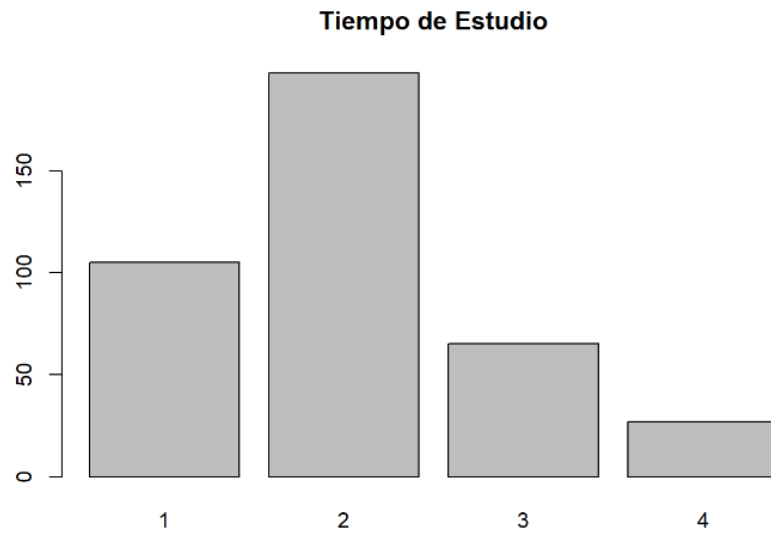
```

F M
R 44 44
U 164 143
```

```
df1 %>%
 ggplot() +
 geom_bar(aes(x = address, fill = sex)) +
 xlab("Dirección") +
 ylab(NULL) +
 labs(fill = "Genero Estudiantes") +
 labs(title = "Distribución de Género por Dirección") +
 theme_minimal()
```



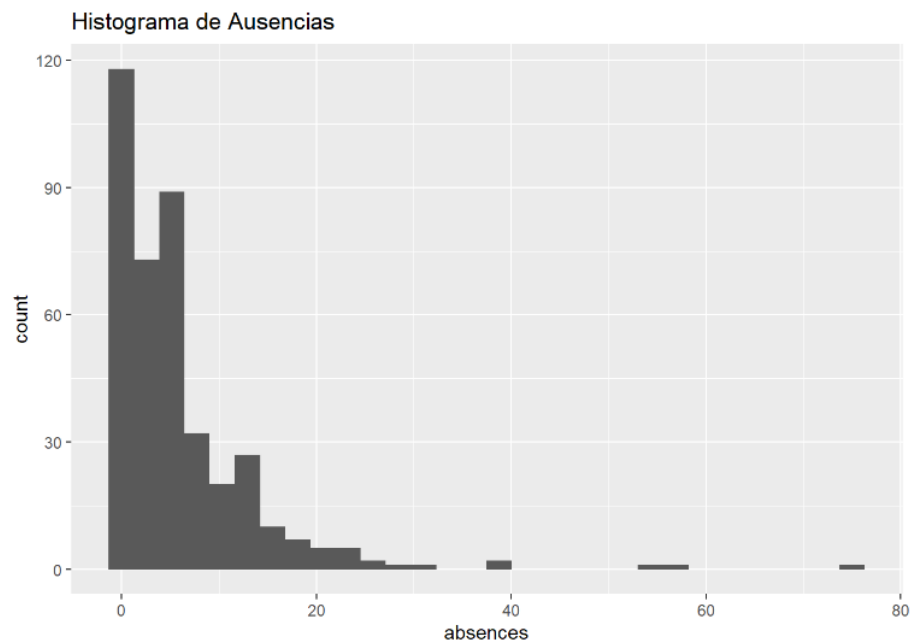
```
barplot(table(df1$studytime),main = "Tiempo de Estudio")
```

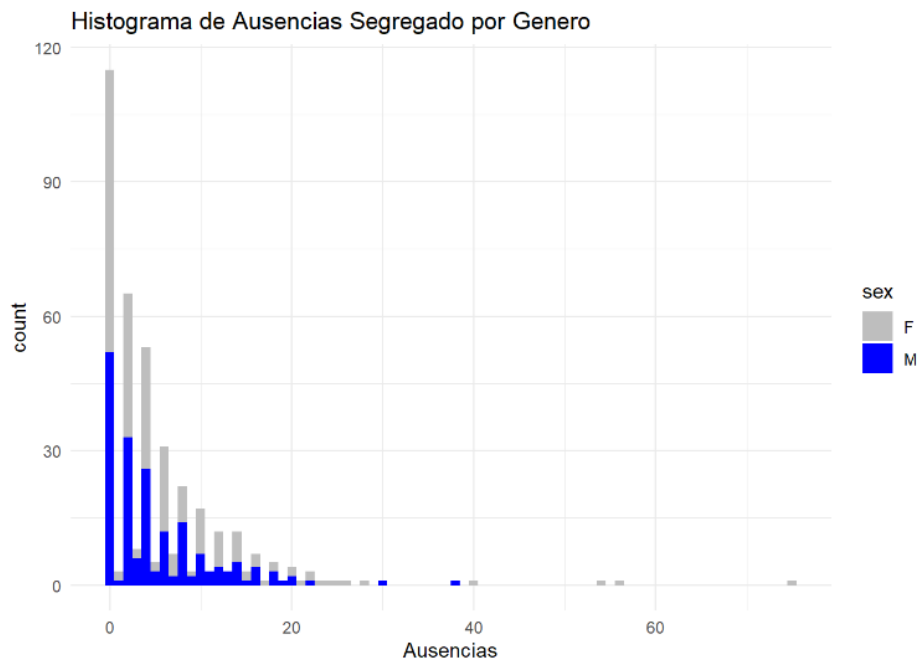


#### Tiempo de estudio

Aquellos estudiantes que dedican tiempo al estudio por lo general le dedican un tiempo de 2 horas

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





### 3.1. Encontrar las técnicas más apropiadas para corroborar las preconclusiones con la ayuda de un mapa.

Vamos a realizar un grafico de correlacion, encontrar alguna insidencia entre algunas carateristicas.

Ademas se procede a realizar calculo de intervalos de confianza.

### 3.2. Diseñar la metodología de análisis.

Tomar el dataset y revisar las variables, aquellas que se requieren transformar se les aplica una tecnica de **Feature engineering**, se procede a buscar las relaciones entre variables, buscar o calcular los intervalos de confianza.

**Feature engineering** Ademas de los resultados previamente vistos, se procede a realizar una tecnica de feature engineering, en la que se va agregar una columna denoinada target, la cual indica si el estudiante pasa o no pasa el grado o materia. En este caso vamos a tener una un Faul o Success.

Segun parte de las indicaciones del dataset si estudiante logra una nota final de mayr igual a 10 entonces "pasa" el grado o materia, de lo contrario pierde el "fail".

```
Agregar una nueva columna llamada 'resultado' basada en la columna 'G3'
df1 <- df1 %>%
 mutate(resultado = ifelse(G3 < 10, "fail", "success"))

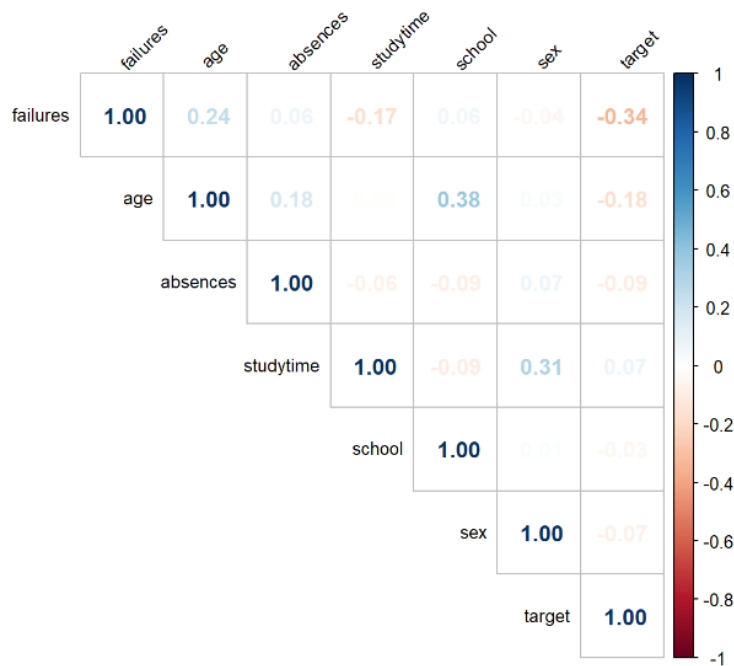
Se muestra el resultado del cambio realizado
colnames(df1)
```

```
[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" "Dalc" "Walc" "health" "absences"
[31] "G1" "G2" "G3" "resultado"
```

### 3.3. Aplicar esta metodología.

Se muestra un grafico en el que se hace una relacion entre variables

```
df3 %>%
 select(failures, age, absences, studytime, school, sex, target) %>%
 cor() %>%
 corplot(method = "number",
 type = "upper",
 tl.cex = 0.8,
 tl.srt = 45,
 tl.col = "black")
```



```
modelo_genero <- lm(G3~sex, data = df3)
summary(modelo_genero)
```

```
##
Call:
lm(formula = G3 ~ sex, data = df3)
##
Residuals:
Min 1Q Median 3Q Max
-10.9144 -1.9663 0.0856 3.0856 9.0856
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.9144 0.3337 32.712 <2e-16 ***
sex -0.9481 0.4598 -2.062 0.0399 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 4.563 on 393 degrees of freedom
Multiple R-squared: 0.0107, Adjusted R-squared: 0.008186
F-statistic: 4.252 on 1 and 393 DF, p-value: 0.03987
```

## Calculo de los Intervalos de Confianza

```
calculo de intervalo de confianza
t.test(df1$age, conf.level =0.95)$conf.int
```

```
[1] 16.56998 16.82243
attr(,"conf.level")
[1] 0.95
```

## Estamos seguros en un 95% que la edad de Los estudiantes se ## encuentra entre 16.56 y 16.82

```
calculo de intervalo de confianza
t.test(df1$absences, conf.level =0.95)$conf.int
```

```
[1] 4.917192 6.500530
attr(,"conf.level")
[1] 0.95
```

## Estamos seguros en un 95% que la cantidad de ausencias de Los ## estudiantes se encuentra entre 4.91 y 6.5

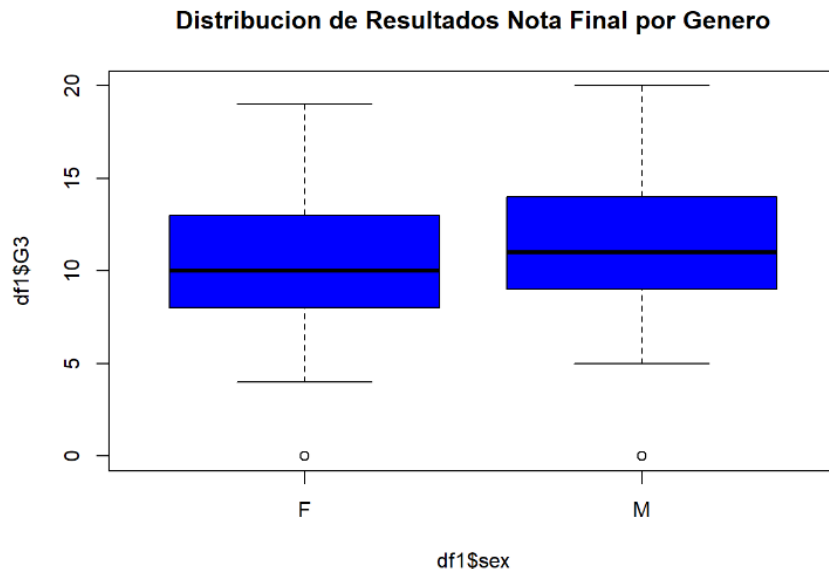
```
calculo de intervalo de confianza
t.test(df1$G3, conf.level =0.95)$conf.int
```

```
[1] 9.961992 10.868388
attr(,"conf.level")
[1] 0.95
```

## Estamos seguros en un 95% que Los resultados de Los estudiantes se encuentra entre 9.96 y 10.86



```
boxplot(df1$G3 ~ df1$sex, col = "blue",
main = "Distribucion de Resultados Nota Final por Genero")
```



```
qhpvt(df1, "sex", "G3", "n()")
```

	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
F	23	1	3	13	4	14	17	30	29	11	17	14	16	6	3	5	2		208
M	15		4	2	5	18	11	26	18	20	14	13	17	10	3	7	3	1	187
Total	38	1	7	15	9	32	28	56	47	31	31	27	33	16	6	12	5	1	395

### 3.4. Resumir los resultados.

Tenemos un dataset con una poblacion estudiantil de dos escuelas diferentes, ademas de otras características como genero, edad, ausencias, materias perdidas, actividades extracurriculares, trabajo del padre y madre.

Tenemos una columna llamada G3 la cual representa el resultado final de la nota del estudiante, se le procedio aplicar tecnica de **Feature engineering**, para transformar los valores para obtener una columna con un resultado el cual es de tipo binario = fail(0) o success (1), esto facilita el manejo de los datos y permite crear relaciones.

Se demuestra la existencia de una correlacion negativa entre la cantidad de horas de estudio y las ausencias, por lo que aquellos estuadintes que destinan mas horas en estudio presentan menos ausencias.

Existe una correlacion negativa entre la variable Target (FAIL / SUCCESS), aquellos alumnos que tienen mas ausencias pueden tener mas probabilidad de perder la materia.

Una alta correlacion negativa entre Target (FAIL/SUCCES) y aquellos alumnos que tienen Failures previos, a mayor failures (fallos) tienen mayor probabilidad de perder el grado/materia

Por medio del calculo de P-value se logra ver un valor de relevancia significativa p-value: 0.0003329, cuando relacionamos las variables entre la EDAD y la variable Target.

Tenemos una gran poblacion de estuadiantes provenientes de las zonas Urbanas.

La mayor poblacion de estudiantil de este caso de estudio se encuentra entre las edades de 15 a 17 años.

los resultados de las mujeres fueron ligeramente superiores (presentan mas Success que los hombres), pero en resultados de notas los hombre acumulan mejores notas mas altas (en los resultados de G3 las notas son mas altas que las mujeres)

Cuando se realiza los Pvalues para la realacion de los Failures y la columna G3 (nota final), se muestra una relacion significativa. Por lo tanto tener o presentar altos Failures insiden sobre el resultado de G3 (la nota final del curso)

Tambien tenemos un resultado de p-value: 0.03987 para la relacion de Genero (sex) contra el resulta de G3, por lo que tal parece en este caso de estudio el sexo infiere sobre el resultado.