Solución Ejercicio

UEB

2025-05-29

Ejercicio Práctico

Este ejercicio consta de diversas partes en un intento de simular lo que se lleva a cabo en un estudio real. Se ha simplicado para hacerlo más practicable por lo que no hace falta que os agobiéis si algo no os cuadre del todo. De lo que se trata es que veamos como aplicar las distintas técnicas que hemos estudiado, de forma integral, en un problema de análisis de datos.

Los datos

La demora entre el comienzo de los síntomas y el ingreso hospitalario es un factor que determina la mortalidad del infarto agudo de miocardio (IAM). Se estudian 426 sujetos que acuden al servicio de urgencias de 5 hospitales por dolor torácico , recogiendo el tiempo entre los primeros síntomas y la llegada al hospital y una serie de variables sociodemográficas. Se está interesado en estimar el retraso prehospitalario y determinar las variables asociadas.

- DEMORA Minutos desde el inicio de los síntomas hasta llegar al hospital
- HOSPITAL
- EDAD
- SEXO 0 Hombre 1 Mujer
- NACIONAL 0 Español 1 Extranjero
- MEDIO1 Lugar inicio síntomas 0 Domicilio 1 Fuera del domicilio
- DISTANC Distància en isòcronas (0,1,2,3, . . .)
- REMITIDO Enviado por un sanitario 0 Si 1 No
- NSE Nivel socioeconómico 0 Bajo 1 Medio 2 Alto
- COHABIT Convivencia 0 Vive en comapñia 1 vive solo
- CARDIOP Diagnóstico previo de cardiopatia 0 No 1 Si
- DOLOR Nivel de dolor 0 intenso 1 moderado 2 ligero

- NOCHE Aparición nocturna 0 Dia 1 Noche

- AMBULAN Acude a l'hospital en ambulancia 0 Ambulancia 1 Otros medios
- imc: Index de masa corporal
- sbp: presión sistólica
- dbp: presión distólica

Los datos los podéis encontrar en los ficheros de Stata demora.dta, de Excel demora.xls y de texto plano separado por comas demora.csv

Preprocesado de datos

- Crea una nueva base de datos que contenga únicamente las variables con las que se va a trabajar: demora, edad, noche, ambulan y dolor

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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 error
```

```
dat <- demora %>% select(demora,edad,noche,ambulan,dolor)
```

- De cuantos individuos y variables dispone ahora la base de datos?

```
dat %>% dim()
```

```
## [1] 426   5
```

- Verifica de que tipo es cada una de las variables.

```
lapply(dat, class)
```

```
## $demora
## [1] "integer"
##
## $edad
## [1] "integer"
##
## $noche
```

```
## [1] "integer"
##
## $ambulan
## [1] "integer"
##
## $dolor
## [1] "integer"
```

- Asigna las variables: noche, ambulan y dolor a factor con los niveles y etiquetas correspondientes.

```
dat <- dat %>%
  mutate(noche = factor(noche, 0:1, c("Dia", "Noche"))) %>%
  mutate(ambulan = factor(ambulan, 0:1, c("Ambulancia", "Otros medios"))) %>%
  mutate(dolor = factor(dolor, 0:2, c("Intenso", "Moderado", "Ligero")))
```

Los análisis

Apartado a

Se está interesado en verificar la calidad de los datos y describir la muestra de estudio

- **Realizar un resumen numérico de las variables demora y edad que incluya medidas de tendencia central y de dispersión. Interpretar los resultados**

```
require(summarytools)
```

```
## Loading required package: summarytools
```

```
##
## Attaching package: 'summarytools'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

```
dfSummary(dat %>% select(demora,edad))
```

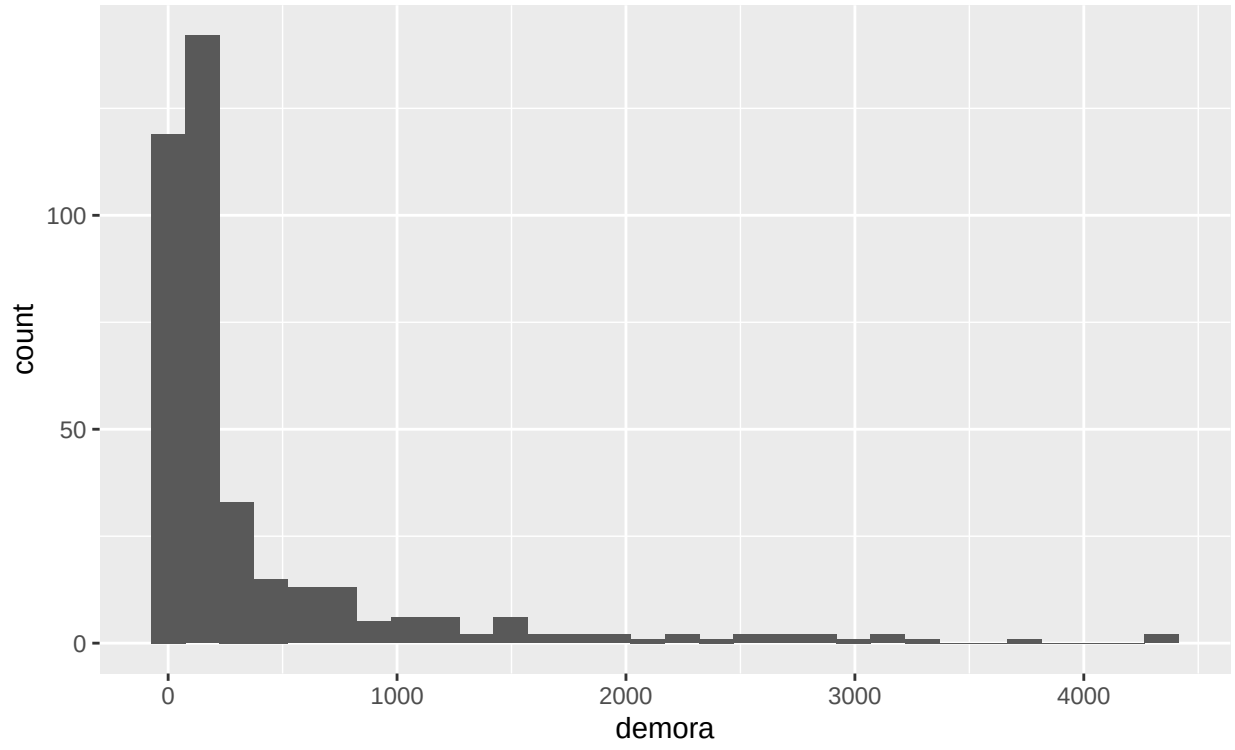
```
## Data Frame Summary
## dat
## Dimensions: 426 x 2
## Duplicates: 56
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values              Freqs (% of Valid)    Graph                 Valid      M
## ---- ----------- --------------------------- --------------------- --------------------- ---------- -
## 1    demora      Mean (sd) : 381.4 (676.7)   139 distinct values   :                     383        4
##      [integer]   min < med < max:                                  :                     (89.9%)
##                  4 < 120 < 4345                                    :
##                  IQR (CV) : 255 (1.8)                              :
##                                                                    : . .
```

```
##
## 2    edad        Mean (sd) : 62.4 (17.7)     57 distinct values                  :       426        0
##      [integer]   min < med < max:                                                    :       (100.0%)
##                  0 < 66 < 91                                                  : : :
##                  IQR (CV) : 17 (0.3)                                        : : : :
##                                                                        .     . . : : : : :
## ---------------------------------------------------------------------------------------------------------
```

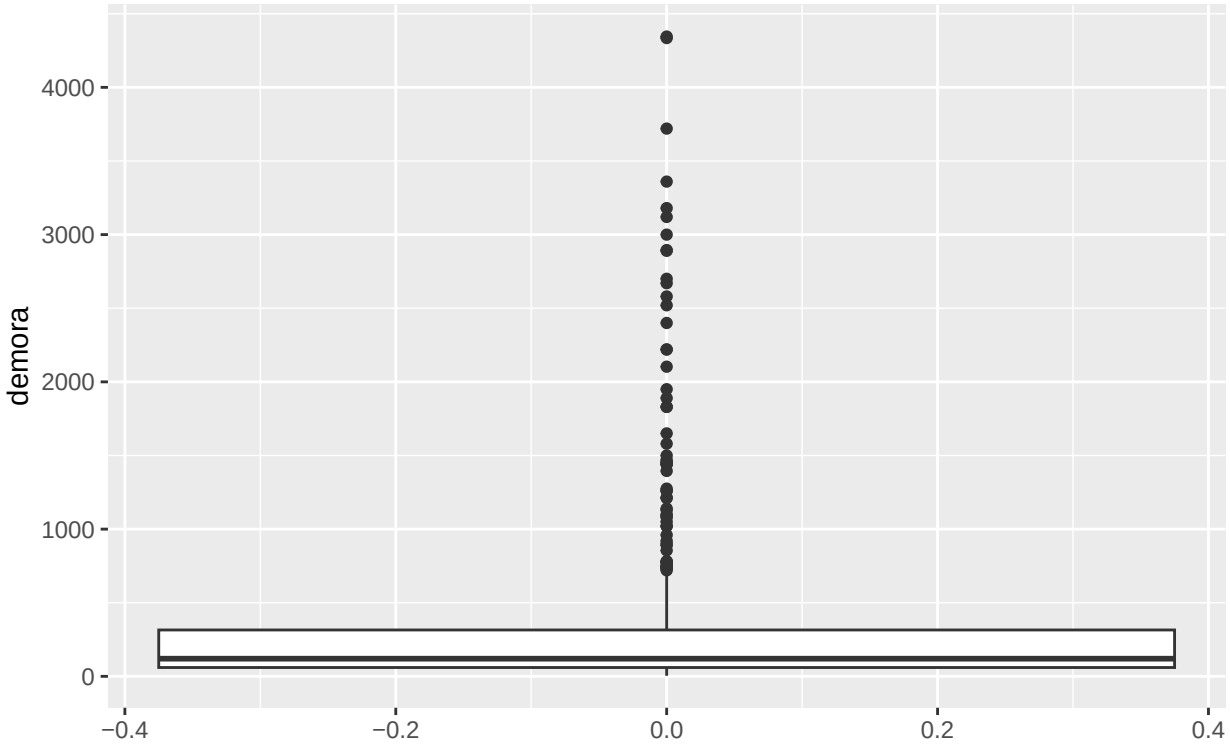
- **Que medida es más adecuada para describir los datos, media o mediana?**
- **Que nos indica el percentil 50%? i el 100%?**
- **Que variable tiene más dispersión?**
- **Realizar un resumen gráfico de las variables demora y edad. Interpretar los resultados**

```
ggplot(dat, aes(x = demora) ) +
  geom_histogram()
```

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

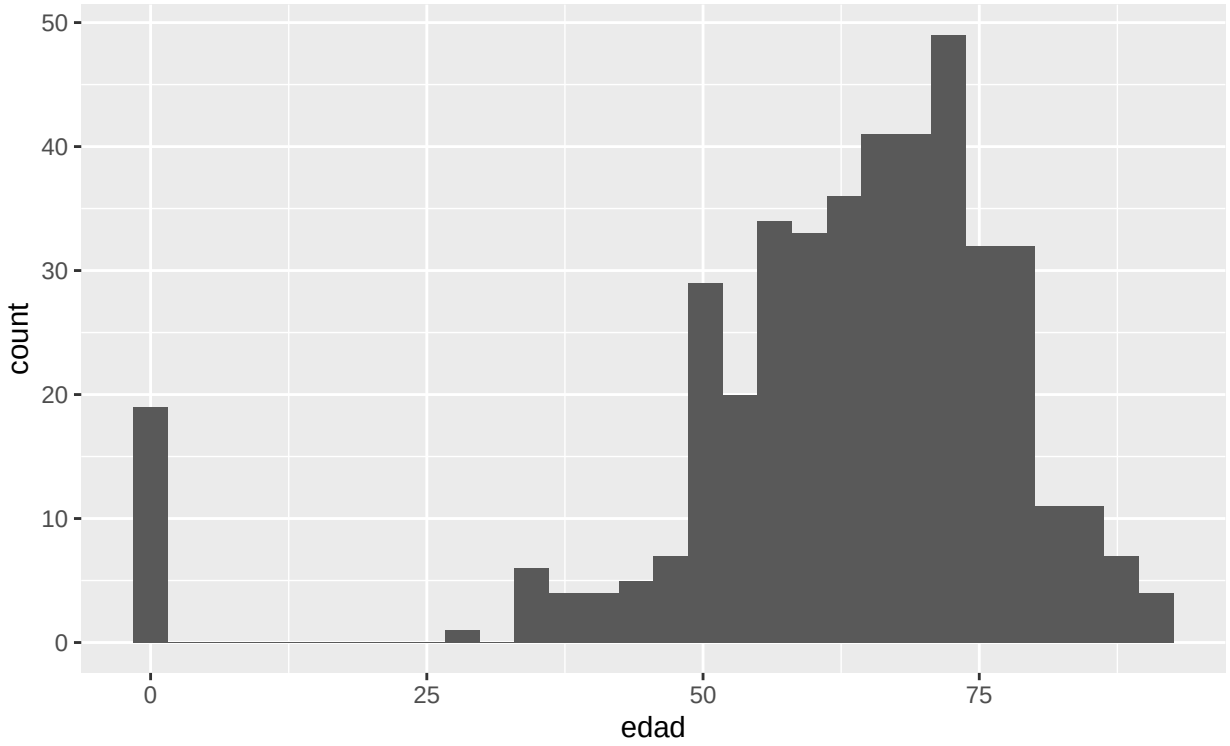



```
ggplot(dat, aes(y = demora) ) +
  geom_boxplot()
```

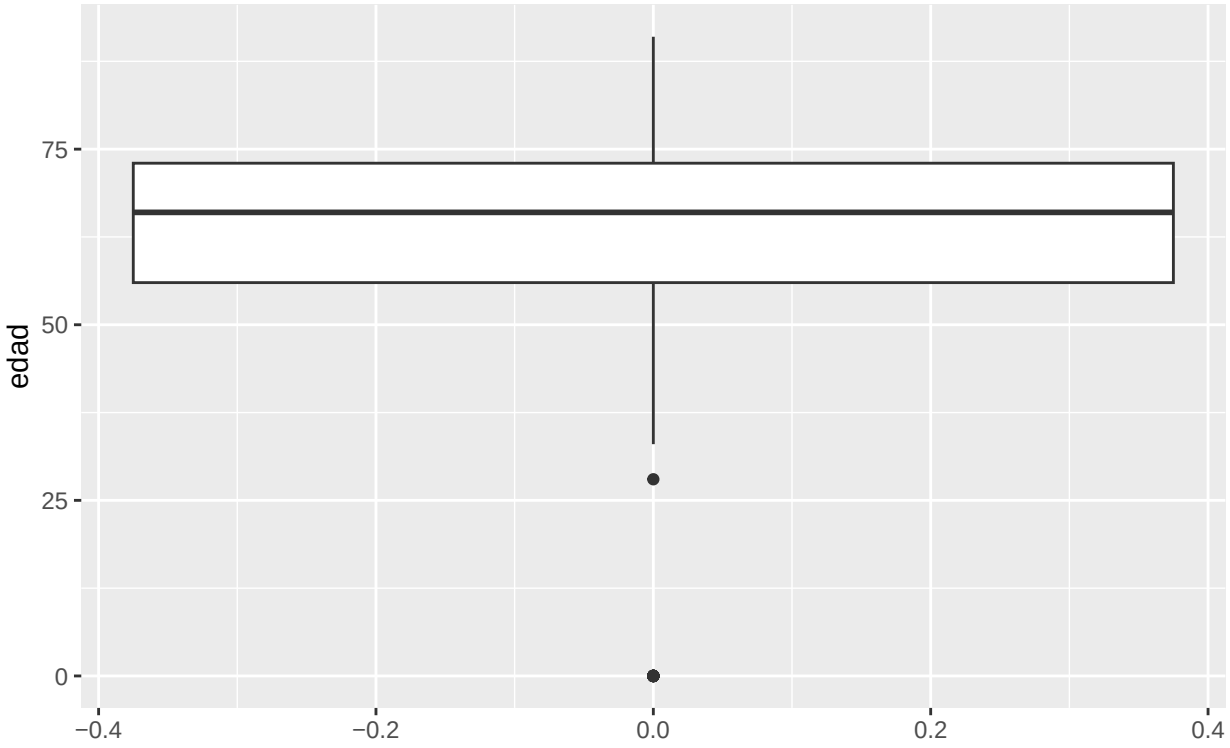



```
ggplot(dat, aes(x = edad) ) +
  geom_histogram()
```

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




```
ggplot(dat, aes(y = edad) ) +
  geom_boxplot()
```



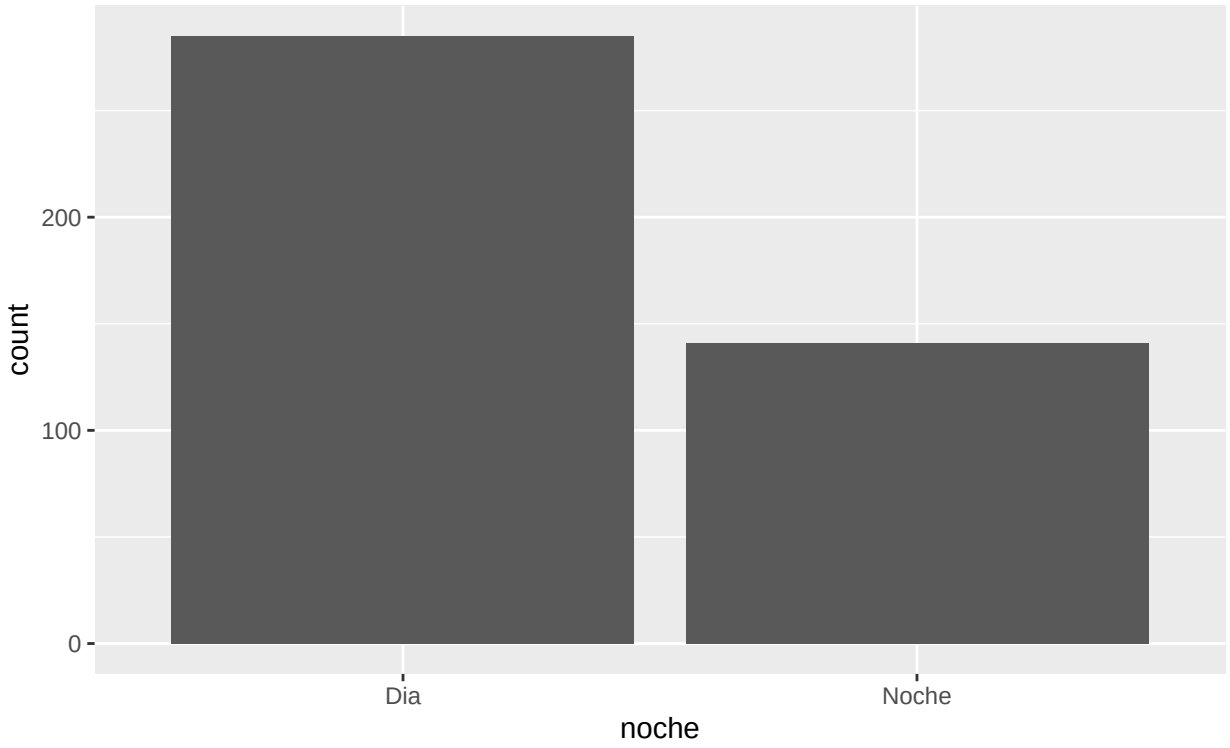
- **Son simétricas las variables?**

- **Existe algun outlier?**
- **Realizar un resumen numérico y gráfico de las variables noche, dolor y ambulan. Interpreta los resultados**

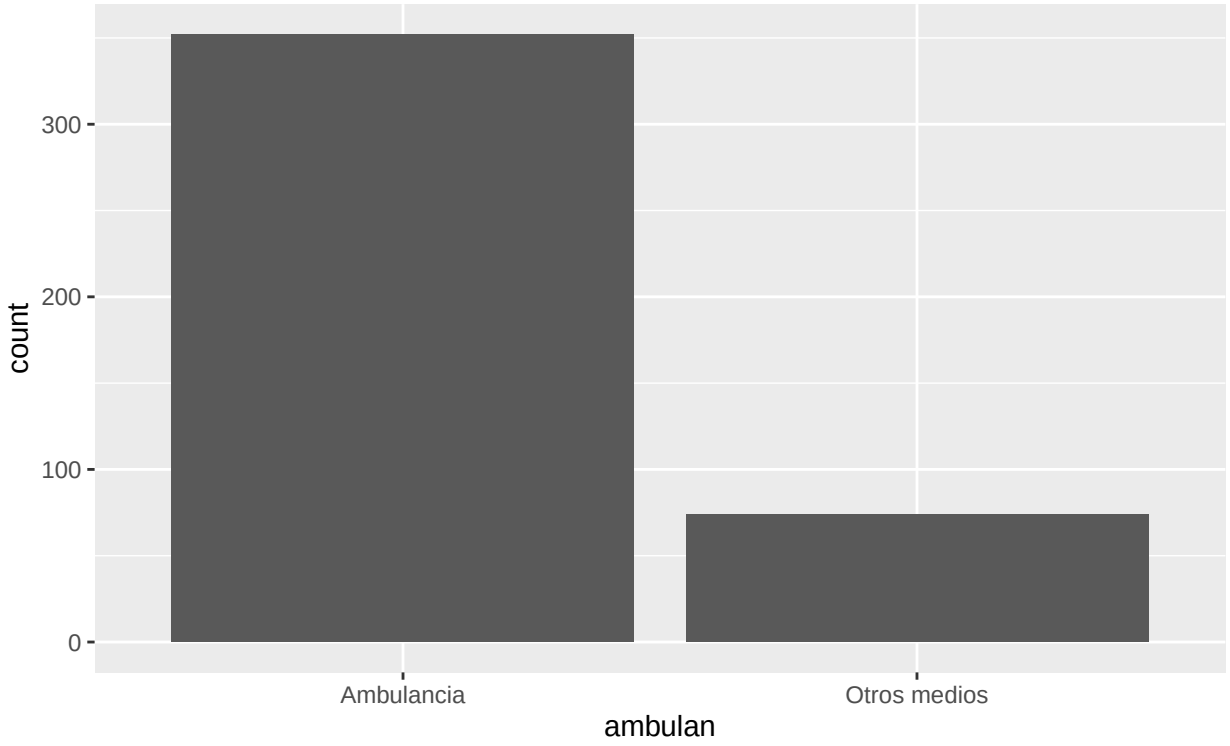
```
dfSummary(dat %>% select(noche,dolor, ambulan))
```

```
## Data Frame Summary
## dat
## Dimensions: 426 x 3
## Duplicates: 414
##
## ------------------------------------------------------------------------------------------------------
## No   Variable   Stats / Values    Freqs (% of Valid)   Graph                Valid      Missing
## ---- ---------- ----------------- -------------------- -------------------- ---------- ---------
## 1    noche      1. Dia            285 (66.9%)          IIIIIIIIIIIII        426        0
##      [factor]   2. Noche          141 (33.1%)          IIIIII               (100.0%)   (0.0%)
##
## 2    dolor      1. Intenso        205 (48.1%)          IIIIIIIII            426        0
##      [factor]   2. Moderado       183 (43.0%)          IIIIIIII             (100.0%)   (0.0%)
##                 3. Ligero          38 ( 8.9%)          I
##
## 3    ambulan    1. Ambulancia     352 (82.6%)          IIIIIIIIIIIIIIII     426        0
##      [factor]   2. Otros medios    74 (17.4%)          III                  (100.0%)   (0.0%)
## ------------------------------------------------------------------------------------------------------
```

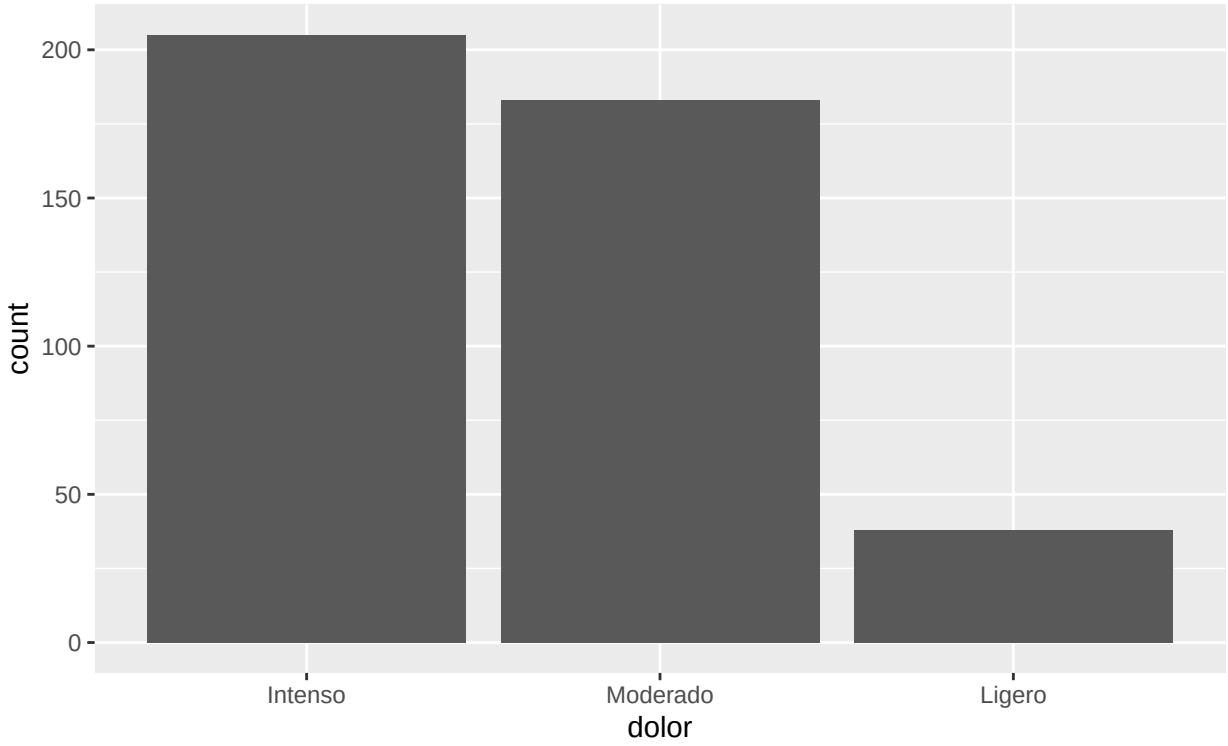
```
ggplot(dat, aes(x = noche) ) +
  geom_bar()
```



```
ggplot(dat, aes(x = ambulan) ) +
  geom_bar()
```

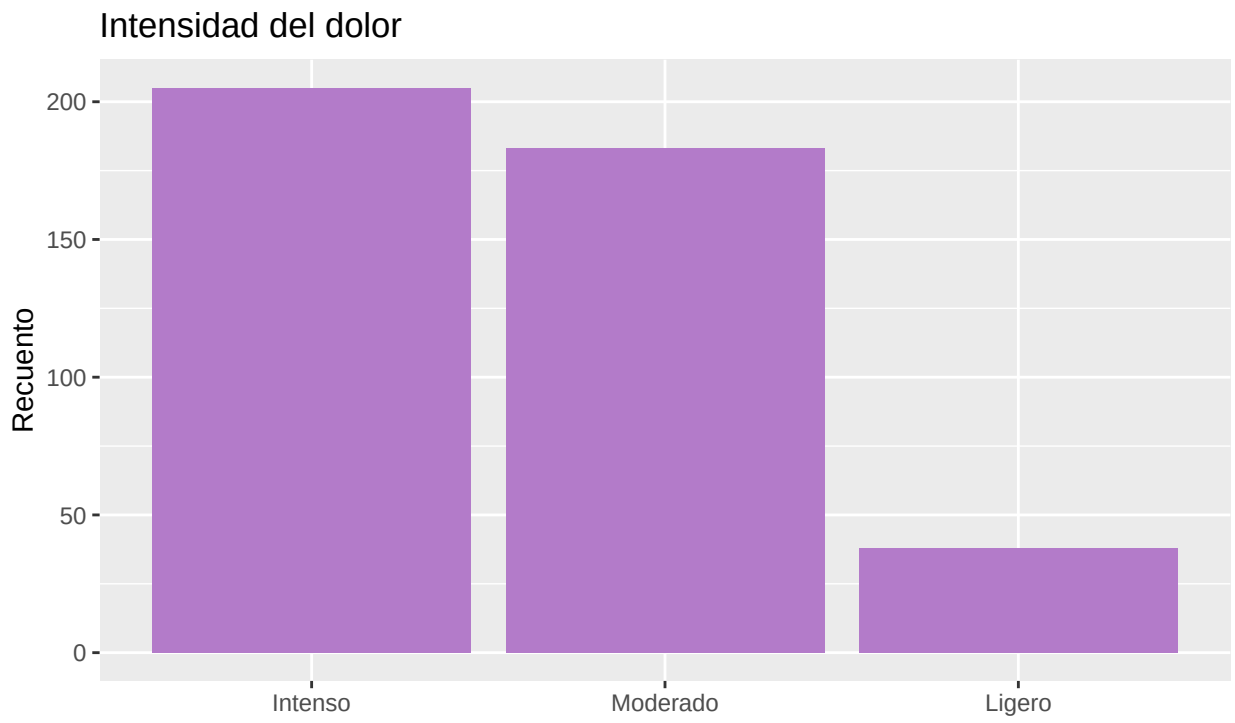



```
ggplot(dat, aes(x = dolor) ) +
  geom_bar()
```



- (OPCIONAL): Cambiar el color de las barras. Añadir un título al gráfico y modificar las etiquetas de los ejes.

```
ggplot(dat, aes(x = dolor)) +
  geom_bar(fill = c( "#b37bc9") ) +
  ggtitle("Intensidad del dolor") +
  xlab("") +
  ylab("Recuento")
```



Apartado b

Se está interesado en conocer la relación entre distintas variables.

- **Se espera que las variables ambulan (que mide como han acudido los pacientes al hospital) y dolor esten relacionadas. Realizar un análisis numérico y gráfico e interpretar los resultados.**

```
require(gmodels)
```

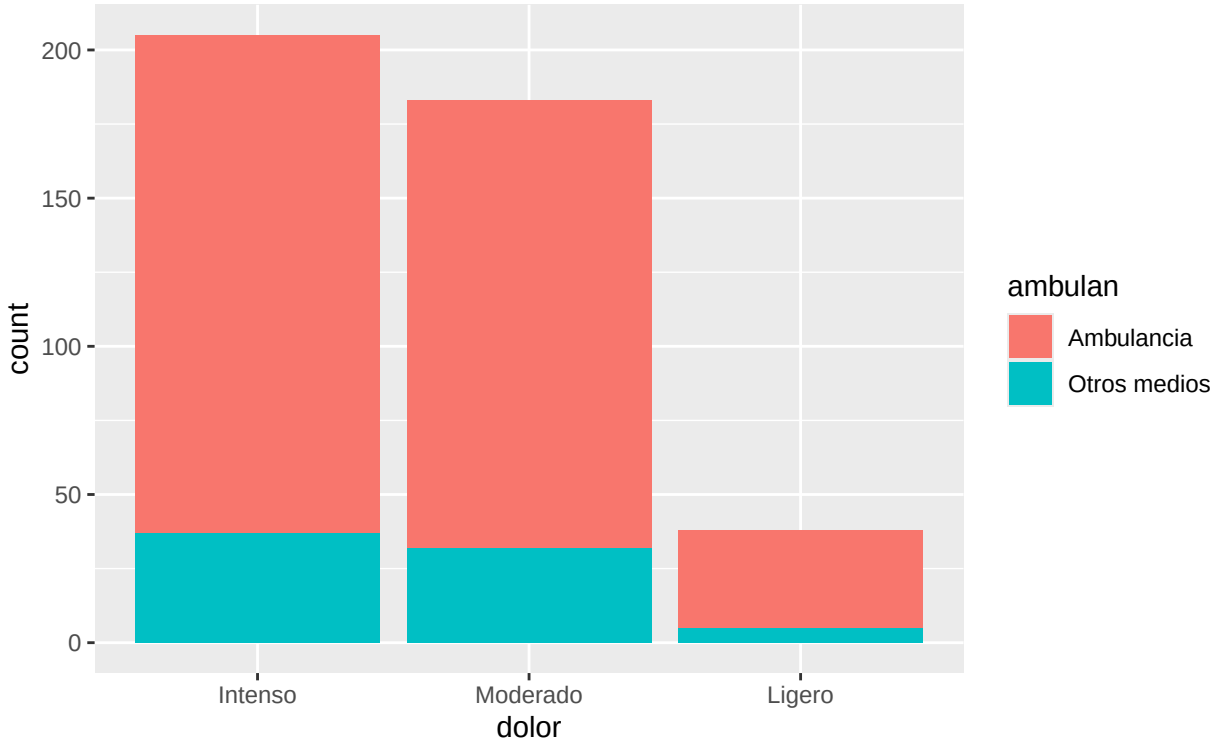
```
## Loading required package: gmodels
```

```
CrossTable(dat$dolor, dat$ambulan, prop.chisq = F, prop.c = F, prop.r = F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  426
##
##
```

```
##              | dat$ambulan
##    dat$dolor |   Ambulancia | Otros medios |    Row Total |
## -------------|--------------|--------------|--------------|
##      Intenso |          168 |           37 |          205 |
##              |        0.394 |        0.087 |              |
## -------------|--------------|--------------|--------------|
##     Moderado |          151 |           32 |          183 |
##              |        0.354 |        0.075 |              |
## -------------|--------------|--------------|--------------|
##       Ligero |           33 |            5 |           38 |
##              |        0.077 |        0.012 |              |
## -------------|--------------|--------------|--------------|
## Column Total |          352 |           74 |          426 |
## -------------|--------------|--------------|--------------|
##
##
```

```
ggplot(dat, aes(x = dolor, fill = ambulan) ) +
  geom_bar()
```



- **El equipo investigador cree que la demora en acudir al hospital y el momento en el que aparece el dolor (variable noche) están relacionadas.**
  - **Realizar el resum numérico de la variable 'demora' según els grupo de la variable 'noche'. Comentar los resultados**

```r
library(dplyr)
dat %>%
group_by(noche) %>%
summarize(median(demora, na.rm = T))
```

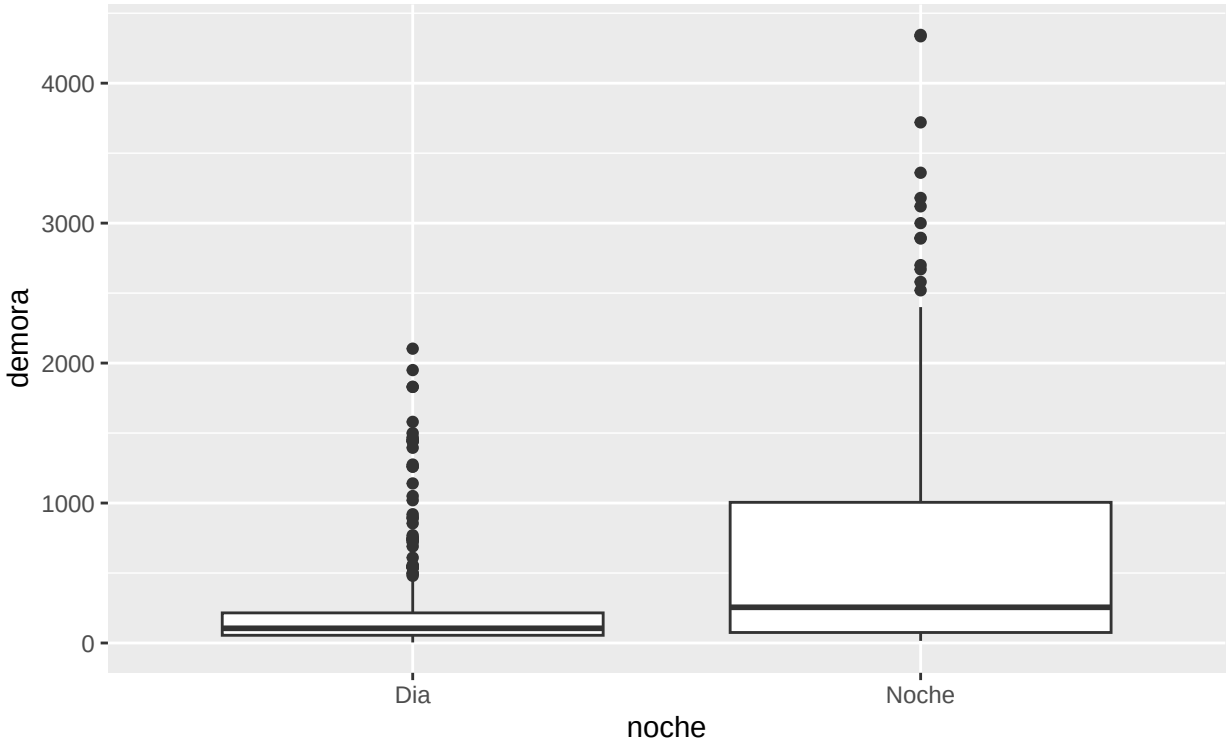
```
## # A tibble: 2 x 2
##   noche `median(demora, na.rm = T)`
##   <fct>                       <dbl>
## 1 Dia                           105
## 2 Noche                         255
```

```r
dat %>%
group_by(noche) %>%
summarize(IQR(demora, na.rm = T))
```

```
## # A tibble: 2 x 2
##   noche `IQR(demora, na.rm = T)`
##   <fct>                    <dbl>
## 1 Dia                        160
## 2 Noche                      930
```

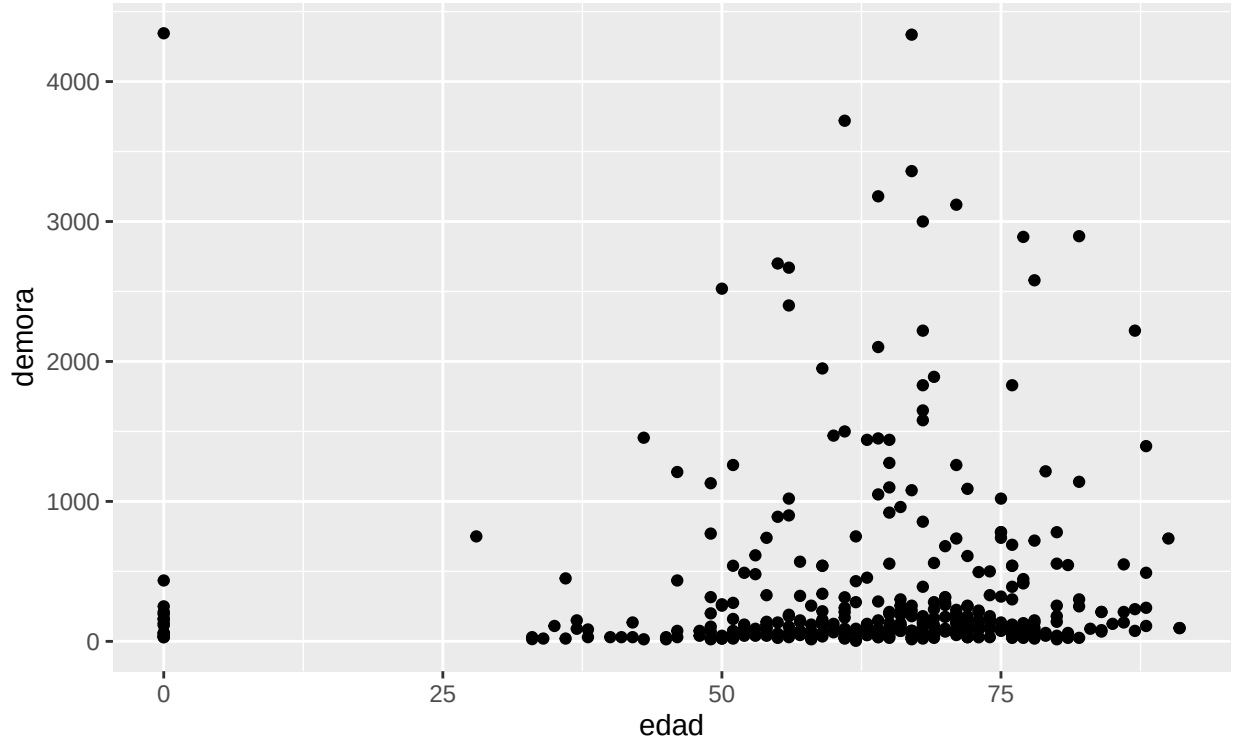
- **Realizar el resum gráfico de la variable 'demora' según els grupo de la variable 'noche'. Comentar los resultados**

```r
ggplot(dat, aes(x = noche, y = demora) ) +
  geom_boxplot()
```



- **El equipo investigador cree que la demora en acudir al hospital y la edad están relacionadas.**
  - **Realizar el resum gráfico de la variable 'demora' y 'edad'. Comentar los resultados**

```
ggplot(dat, aes(x = edad, y = demora) ) +
  geom_point()
```



- **Calcula el coeficiente de correlación de Pearson y Spearman para las variables 'demora' y 'edad' e interpreta el resultado.**

```
cor(dat$edad,dat$demora, use = "complete.obs")
```

```
## [1] 0.02967572
```

```
cor(dat$edad,dat$demora, use = "complete.obs", method = "spearman")
```

```
## [1] 0.1554909
```

Apartado c

- **Realiza el resumen gráfico para todas las parejas de variables posibles (Pista: usar funcion ggpairs)**

```
require(GGally)
```

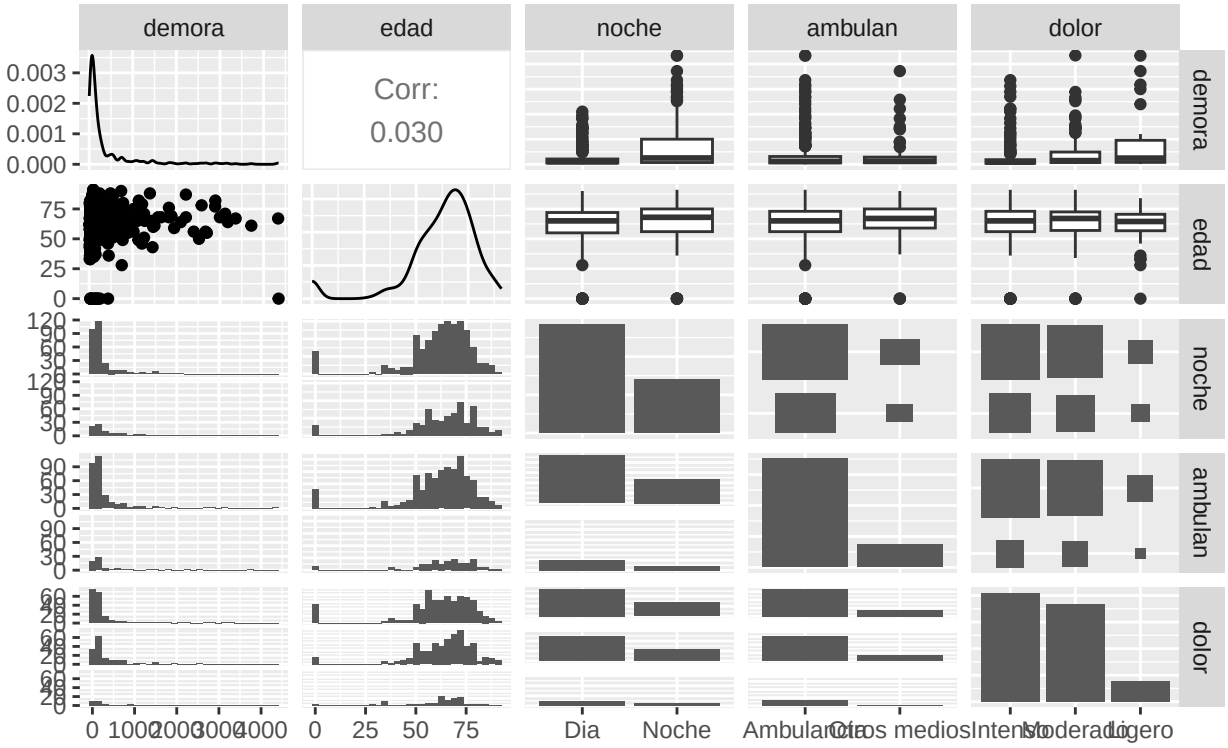
```
## Loading required package: GGally
```

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

```
ggpairs(dat)
```

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

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



- **¿Crees que hay algún gráfico que no sea útil? ¿Por qué?**