# Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

David Fernández Álvarez y Sara Robisco Cavite

Diciembre 2022

## Contents



---

# Introducción

---

## Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

## Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

## Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

### Descripción de la práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la Práctica 1 o bien cualquier dataset libre disponible en Kaggle https://www.kaggle.com.

Un ejemplo de dataset con el que podéis trabajar es el "Heart Attack Analysis & Prediction dataset": https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-predictiondataset

Importante: si se elige un dataset diferente al propuesto es importante que este contenga una amplia variedad de datos numéricos y categóricos para poder realizar un análisis más rico y poder responder a las diferentes preguntas planteadas en el enunciado de la práctica.

## Tareas a realizar

### Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Para describir el dataset de una forma visual, cargamos las librerías ggplot2 y dplry.

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

Ahora cargamos el fichero de datos.

```
dataset <- read.csv('../dataset/detecciones_ondas_gravitacionales.csv',stringsAsFactors = FALSE)
filas=dim(dataset)[1]
```

Para describir el conjunto de datos en profundidad vamos a comenzar verificando su estructura:

```
str(dataset)
```

```
## 'data.frame':    119 obs. of  36 variables:
##  $ name                          : chr  "GW200322_091133" "GW200316_215756" "GW200311_115853" "GW20
##  $ version                       : chr  "v1" "v1" "v1" "v1" ...
##  $ release                       : chr  "GWTC-3-confident" "GWTC-3-confident" "GWTC-3-confident" "GW
##  $ gps                           : num  1.27e+09 1.27e+09 1.27e+09 1.27e+09 1.27e+09 ...
##  $ mass_1                        : num  34 13.1 34.2 NA 36.4 28.3 37.8 19.3 40 38.9 ...
##  $ mass_1_upper                  : num  48 10.2 6.4 NA 11.2 17.1 8.7 5 6.9 14.1 ...
##  $ mass_1_lower                  : num  -18 -2.9 -3.8 NA -9.6 -7.7 -8.5 -3 -4.5 -8.6 ...
##  $ mass_2                        : num  14 7.8 27.7 NA 13.8 14.8 20 14 32.5 27.9 ...
##  $ mass_2_upper                  : num  16.8 1.9 4.1 NA 7.2 6.5 8.1 2.8 5 9.2 ...
```

```
##  $ mass_2_lower                   : num  -8.7 -2.9 -5.9 NA -3.3 -6.4 -5.7 -3.5 -7.2 -9 ...
##  $ network_snr                    : num  6 10.3 17.8 9.2 7.1 7.8 10.8 12.5 20 8.5 ...
##  $ network_snr_upper              : num  1.7 0.4 0.2 NA 0.5 0.4 0.3 0.3 0.2 0.3 ...
##  $ network_snr_lower              : num  -1.2 -0.7 -0.2 NA -0.5 -0.6 -0.4 -0.4 -0.2 -0.5 ...
##  $ distance                       : int  3600 1120 1170 NA 5400 2100 1480 1150 1710 4000 ...
##  $ distance_upper                 : int  7000 470 280 NA 2700 1700 1020 510 490 2800 ...
##  $ distance_lower                 : int  -2000 -440 -400 NA -2600 -1100 -700 -530 -640 -2200 ...
##  $ chi_eff                        : num  0.24 0.13 -0.02 NA 0.65 0.32 0.01 -0.12 0.1 -0.07 ...
##  $ chi_eff_upper                  : num  0.45 0.27 0.16 NA 0.17 0.28 0.25 0.17 0.15 0.27 ...
##  $ chi_eff_lower                  : num  -0.51 -0.1 -0.2 NA -0.21 -0.46 -0.26 -0.28 -0.15 -0.33 ...
##  $ total_mass                     : num  55 21.2 61.9 NA 50.6 43.9 57.8 33.5 72.2 67 ...
##  $ total_mass_upper               : num  37 7.2 5.3 NA 10.9 11.8 9.6 3.6 7.2 17 ...
##  $ total_mass_lower               : num  -27 -2 -4.2 NA -8.5 -7.5 -6.9 -3 -5.1 -12 ...
##  $ chirp_mass                     : num  15.5 8.75 26.6 NA 19 17.5 23.4 14.2 31.1 28.2 ...
##  $ chirp_mass_upper               : num  15.7 0.62 2.4 NA 4.8 3.5 4.7 1.5 3.2 7.3 ...
##  $ chirp_mass_lower               : num  -3.7 -0.55 -2 NA -2.8 -3 -3 -1.4 -2.6 -5.1 ...
##  $ detector_frame_chirp_mass      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ detector_frame_chirp_mass_upper: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ detector_frame_chirp_mass_lower: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ redshift                       : num  0.6 0.22 0.23 NA 0.83 0.38 0.28 0.22 0.32 0.66 ...
##  $ redshift_upper                 : num  0.84 0.08 0.05 NA 0.32 0.24 0.16 0.09 0.08 0.36 ...
##  $ redshift_lower                 : num  -0.3 -0.08 -0.07 NA -0.35 -0.18 -0.12 -0.1 -0.11 -0.31 ...
##  $ false_alarm_rate               : chr  "140" "â‰¤ 1.0e-05" "â‰¤ 1.0e-05" "1.3" ...
##  $ p_astro                        : chr  "0.62" "â‰¥ 0.99" "â‰¥ 0.99" "0.19" ...
##  $ final_mass                     : chr  "53" "20.2" "59.0" "" ...
##  $ final_mass_upper               : num  38 7.4 4.8 NA 11.1 12.3 8.9 3.5 6.6 16 ...
##  $ final_mass_lower               : num  -26 -1.9 -3.9 NA -7.7 -6.9 -6.6 -2.8 -4.7 -11 ...
```

Observamos que tenemos 119 registros corespondientes con datos de ondas gravitacionales y 36 variables que los caracterizan. A continuación describimos las variables:

**name** cadena de caracteres con el identificador de la detección de la onda gravitacional.

**version** versión de la detección. Se revisan periódicamente.

**release** datos de la comunicación de la detección, si es confirmada, si es descartada...

**gps** fecha y hora de la detección en formato GPS.

**mass__1** masa del primer objeto en masas solares.

**mass__1__upper** valor máximo del rángo de error de la masa del primer objeto.

**mass__1__lower** valor mínimo del rángo de error de la masa del primer objeto.

**mass__2** masa del segundo objeto en masas solares.

**mass__2__upper** valor máximo del rángo de error de la masa del segundo objeto.

**mass__2__lower** valor mínimo del rángo de error de la masa del segundo objeto.

**network__snr** ratio señal/ruido en la red.

**metwork__snr__upper** valor máximo del rángo de error del ratio señal/ruido en la red.

**metwork__snr__lower** valor mínimo del rángo de error del ratio señal/ruido en la red.

**distance** distancia a la que se ha producido la colisión, en Megapársecs.

**distance__upper** valor máximo del rángo de error de la distancia.

**distance__lower** valor mínimo del rángo de error de la distancia.

**chi_eff** correlación de campo z de las fusiones de agujeros negros binarios.

**chi_eff_upper** valor máximo del rángo de error de la correlación de campo.

**chi_eff_lower** valor mínimo del rángo de error de la correlación de campo.

**total_mass** masa total de ambos cuerpos. Medida en masas solares.

**total_mass_upper** valor máximo del rángo de error de la masa total.

**total_mass_lower** valor mínimo del rángo de error de la masa total.

**chirp_mass** masa efectiva de un sistema binario. Medida en masas solares.

**chirp_mass_upper** valor máximo del rángo de error de la masa efectiva.

**chirp_mass_lower** valor mínimo del rángo de error de la masa efectiva.

**detector_Frame_Chirp_Mass** marco del detector de la masa efectiva. Medida en masas solares.

**detector_Frame_Chirp_mass_upper** valor máximo del rángo de error del marco del detector de la masa efectiva.

**detector_Frame_Chirp_mass_lower** valor mínimo del rángo de error del marco del detector de la masa efectiva.

**redshift** corrimiento al rojo, marca la velocidad a la que se alejan de nosotros.

**redshift_upper** valor máximo del rángo de error del corrimiento al rojo.

**redshift_lower** valor mínimo del rángo de error del corrimiento al rojo.

**false_Alarm_Rate** tasa de falsa alarma. La medida es años elevado a -1.

**p_astro** probabilidad de que el evento tenga un origen astrofísico.

**final_mass** masa final del objeto resultante tras la colisión. Medida en masas solares.

**final_mass_upper** valor máximo del rángo de error de la masa final.

**final_mass_lower** valor mínimo del rángo de error de la masa final.

Observamos que tenemos seis variables de tipo caracter: tres tienen el tipo adecuado, pero hay otras tres que deberían ser de tipo numérico: false_alarm_rate, p_astro y final_mass. Esto debemos corregirlo, para ello los transformaremos en numéricos:

```
dataset$false_alarm_rate <- as.numeric(dataset$false_alarm_rate)
```

```
## Warning: NAs introducidos por coerción
```

```
dataset$p_astro <- as.numeric(dataset$p_astro)
```

```
## Warning: NAs introducidos por coerción
```

```
dataset$final_mass <- as.numeric(dataset$final_mass)
```

```
## Warning: NAs introducidos por coerción
```

Ahora mostramos cómo queda el análisis estadístico:

```
summary(dataset)
```

```
##      name              version            release              gps
##  Length:119         Length:119         Length:119         Min.   :1.126e+09
##  Class :character   Class :character   Class :character   1st Qu.:1.240e+09
##  Mode  :character   Mode  :character   Mode  :character   Median :1.249e+09
##                                                           Mean   :1.236e+09
```

```
##                                                          3rd Qu.:1.261e+09
##                                                          Max.   :1.269e+09
##
##      mass_1          mass_1_upper      mass_1_lower          mass_2
##  Min.   :  1.46   Min.   :  0.12   Min.   :-33.000   Min.   : 1.17
##  1st Qu.: 20.80   1st Qu.:  5.60   1st Qu.: -9.600   1st Qu.: 8.20
##  Median : 35.40   Median :  9.80   Median : -6.000   Median :23.30
##  Mean   : 34.95   Mean   : 13.13   Mean   : -7.668   Mean   :21.62
##  3rd Qu.: 42.20   3rd Qu.: 14.10   3rd Qu.: -3.200   3rd Qu.:29.00
##  Max.   :105.50   Max.   :104.00   Max.   : -0.100   Max.   :76.00
##  NA's   :26       NA's   :26       NA's   :26        NA's   :26
##   mass_2_upper      mass_2_lower       network_snr    network_snr_upper
##  Min.   : 0.070   Min.   :-36.500   Min.   : 6.00   Min.   :0.1000
##  1st Qu.: 2.200   1st Qu.: -9.300   1st Qu.: 9.10   1st Qu.:0.2000
##  Median : 5.500   Median : -5.900   Median :10.70   Median :0.3000
##  Mean   : 6.985   Mean   : -6.971   Mean   :12.01   Mean   :0.3349
##  3rd Qu.: 9.300   3rd Qu.: -2.400   3rd Qu.:13.15   3rd Qu.:0.4000
##  Max.   :27.100   Max.   : -0.060   Max.   :33.00   Max.   :1.7000
##  NA's   :26       NA's   :26                        NA's   :33
##  network_snr_lower    distance    distance_upper distance_lower
##  Min.   :-1.2000   Min.   :  40   Min.   :   7   Min.   :-4290.0
##  1st Qu.:-0.6000   1st Qu.: 930   1st Qu.: 340   1st Qu.:-1500.0
##  Median :-0.4000   Median :1620   Median : 770   Median : -650.0
##  Mean   :-0.4837   Mean   :2098   Mean   :1371   Mean   : -995.8
##  3rd Qu.:-0.3000   3rd Qu.:3280   3rd Qu.:1930   3rd Qu.: -380.0
##  Max.   :-0.2000   Max.   :8280   Max.   :7000   Max.   :  -15.0
##  NA's   :33        NA's   :26     NA's   :26     NA's   :26
##     chi_eff          chi_eff_upper     chi_eff_lower      total_mass
##  Min.   :-0.29000   Min.   :0.0200   Min.   :-0.510   Min.   :  3.40
##  1st Qu.:-0.03000   1st Qu.:0.1500   1st Qu.:-0.310   1st Qu.: 31.62
##  Median : 0.05000   Median :0.2100   Median :-0.220   Median : 58.75
##  Mean   : 0.07989   Mean   :0.2173   Mean   :-0.229   Mean   : 57.90
##  3rd Qu.: 0.15000   3rd Qu.:0.2600   3rd Qu.:-0.130   3rd Qu.: 74.65
##  Max.   : 0.68000   Max.   :0.5000   Max.   :-0.010   Max.   :182.30
##  NA's   :26         NA's   :26       NA's   :26       NA's   :37
##  total_mass_upper total_mass_lower    chirp_mass     chirp_mass_upper
##  Min.   :  0.30   Min.   :-35.700   Min.   : 1.186   Min.   : 0.001
##  1st Qu.:  4.20   1st Qu.:-12.000   1st Qu.: 9.000   1st Qu.: 0.660
##  Median :  9.45   Median : -7.650   Median :24.400   Median : 3.500
##  Mean   : 13.71   Mean   : -8.512   Mean   :23.095   Mean   : 4.720
##  3rd Qu.: 17.82   3rd Qu.: -2.800   3rd Qu.:30.900   3rd Qu.: 7.300
##  Max.   :100.00   Max.   : -0.100   Max.   :76.000   Max.   :23.000
##  NA's   :37       NA's   :37        NA's   :26       NA's   :26
##  chirp_mass_lower  detector_frame_chirp_mass detector_frame_chirp_mass_upper
##  Min.   :-17.400   Min.   : 0.900            Min.   : 6.00
##  1st Qu.: -5.000   1st Qu.: 1.508            1st Qu.:12.55
##  Median : -2.600   Median : 3.075            Median :19.10
##  Mean   : -3.400   Mean   : 8.742            Mean   :19.10
##  3rd Qu.: -0.600   3rd Qu.: 6.317            3rd Qu.:25.65
##  Max.   : -0.001   Max.   :49.800            Max.   :32.20
##  NA's   :26        NA's   :103               NA's   :117
##  detector_frame_chirp_mass_lower    redshift     redshift_upper
##  Min.   :-12.4                   Min.   :0.010   Min.   :0.0000
##  1st Qu.:-10.5                   1st Qu.:0.190   1st Qu.:0.0600
```

```
##  Median : -8.6                  Median :0.300   Median :0.1200
##  Mean   : -8.6                  Mean   :0.362   Mean   :0.1843
##  3rd Qu.: -6.7                  3rd Qu.:0.550   3rd Qu.:0.2600
##  Max.   : -4.8                  Max.   :1.180   Max.   :0.8400
##  NA's   :117                    NA's   :26      NA's   :26
##  redshift_lower    false_alarm_rate       p_astro         final_mass
##  Min.   :-0.5300   Min.   :  0.00001   Min.   :0.0500   Min.   :  7.20
##  1st Qu.:-0.2200   1st Qu.:  0.03700   1st Qu.:0.7700   1st Qu.: 32.15
##  Median :-0.1200   Median :  0.55000   Median :0.9700   Median : 56.30
##  Mean   :-0.1526   Mean   :  4.78454   Mean   :0.8574   Mean   : 55.32
##  3rd Qu.:-0.0700   3rd Qu.:  4.43750   3rd Qu.:1.0000   3rd Qu.: 69.70
##  Max.   : 0.0000   Max.   :140.00000   Max.   :1.0000   Max.   :172.90
##  NA's   :26        NA's   :37          NA's   :38       NA's   :28
##  final_mass_upper final_mass_lower
##  Min.   :  1.30   Min.   :-33.600
##  1st Qu.:  4.05   1st Qu.:-11.000
##  Median :  8.60   Median : -6.600
##  Mean   : 12.65   Mean   : -7.763
##  3rd Qu.: 16.00   3rd Qu.: -2.750
##  Max.   :100.00   Max.   : -0.660
##  NA's   :28       NA's   :28
```

La importancia de este conjunto de datos radica en nuestra curiosidad por conocer más a fondo los datos que componen las detecciones de ondas gravitacionales detectadas por el consorcio LIGO, VIRGO y KAGRA, tanto las confirmadas como las rechazadas. La idea es aprender más de estos fenómenos gracias a sus datos.

Con estos datos queremos intentar responder algunas preguntas:

- ¿Qué intervalos de masas de objetos son los más detectados?
- ¿Hay periodos del año donde haya más probabilidad de detecciones? Si es así ¿De qué región del espacio provienen?
- ¿Qué hace que una señal se considere buena o se descarte?
- ¿Cuáles son las detecciones más cercanas? ¿Y las más lejanas?

## Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Aquí podemos hacer un análisis de los datos relevantes como los que hemos visto en la teoría para decidir qué datos tomar. Pero antes necesitamos hacer los pasos posteriores.

## Limpieza de los datos.

### ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Veamos las estadísticas de valores nulos:

```
colSums(is.na(dataset))
```

```
##                         name                      version
##                            0                            0
##                      release                          gps
##                            0                            0
##                       mass_1                 mass_1_upper
##                           26                           26
```

```
##                     mass_1_lower                           mass_2
##                               26                               26
##                     mass_2_upper                     mass_2_lower
##                               26                               26
##                      network_snr                  network_snr_upper
##                                0                               33
##                 network_snr_lower                         distance
##                               33                               26
##                    distance_upper                   distance_lower
##                               26                               26
##                          chi_eff                    chi_eff_upper
##                               26                               26
##                    chi_eff_lower                       total_mass
##                               26                               37
##                  total_mass_upper                 total_mass_lower
##                               37                               37
##                       chirp_mass                 chirp_mass_upper
##                               26                               26
##                  chirp_mass_lower       detector_frame_chirp_mass
##                               26                              103
## detector_frame_chirp_mass_upper detector_frame_chirp_mass_lower
##                              117                              117
##                         redshift                    redshift_upper
##                               26                               26
##                    redshift_lower                 false_alarm_rate
##                               26                               37
##                           p_astro                       final_mass
##                               38                               28
##                  final_mass_upper                 final_mass_lower
##                               28                               28
```

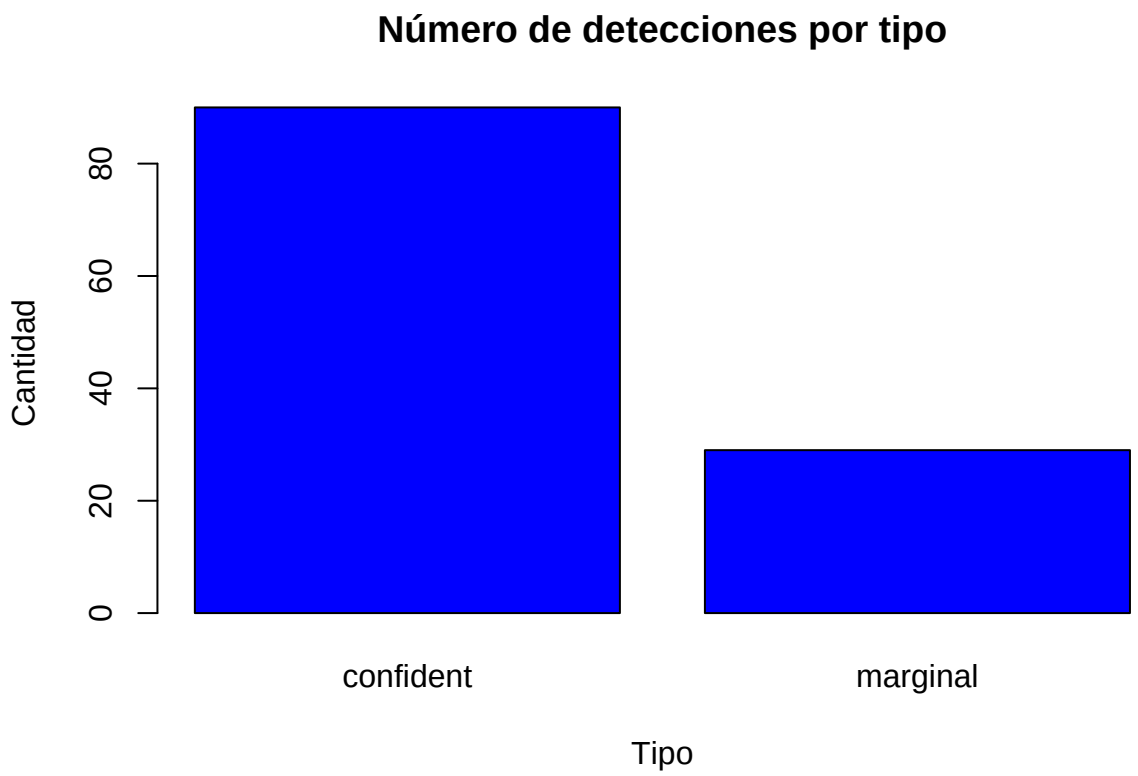
En función a las dos tablas obtenidas vamos viendo qué valores no podemos usar debido a su enorme cantidad de valores vacíos. Por ejemplo: detector_frame_chirp_mass, detector_frame_chirp_mass_upper y detector_frame_chirp_mass_lower tienen casi todos sus valores nulos. Por este motivo descartaremos estas columnas. Al ser datos del propio detector no son críticos y no afectarán a nuestro resultado final.

Sobre el resto de valores nulos, podemos observar que coinciden las cantidades en muchos valores ¿Podrá corresponderse con los datos de las ondas gravitacionales descartadas? Realmente tenemos un campo en el que se indica si la detección es buena o no, ese es el campo release. Debemos transformarlo para poder clasificar las detecciones entre confirmadas o no y así poder sacar mejores conclusiones.Vamos a meter ese valor en una nueva variable:

```
library(stringr)
dataset$tipo <- ifelse(str_detect(dataset$release, "confident"), "confident", "marginal")
```

Veamos ahora cuántas son detecciones de ondas gravitacionales confirmadas y cuántas no. Lo haremos mostrando un gráfico:

```
plot(factor(dataset$tipo),main="Número de detecciones por tipo",xlab="Tipo", ylab="Cantidad",col = "blu
```




Tenemos unas 29 detecciones de tipo marginal, hemos tomado además como marginales aquellas que no estaban etiquetadas. Parece que los valores nulos se corresponden con estas detecciones.

**Identifica y gestiona los valores extremos.**

## Análisis de los datos.

**Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)**

**Comprobación de la normalidad y homogeneidad de la varianza.**

**Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

## Representación de los resultados a partir de tablas y gráficas.

Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica. Lo dejo aquí para acordarnos de poner todas las gráficas y tablas posibles.

## Resolución del problema.

## Vídeo.