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

*Tipología y ciclo de vida de los datos*

Fiorella Piriz Sapio y Rafael Pla Santo Tomas

02/01/2022

# • ÍNDICE

# 1. Introducción

El objetivo de esta práctica es analizar un dataset y realizar las siguientes tareas:

- Limpieza de los datos.

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

  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.

En esta práctica la hemos analizado el dataset GBSG2 sobre el que hemos realizado una limpieza de datos gestionando los outliers y los elementos vacíos para poder realizar distintos análisis centrados en los diferentes tratamientos aplicados para las pacientes. También se ha observado la normalidad y la homogeneidad de la varianza, así como aplicado diferentes pruebas estadísticas como correlaciones, regresiones y contraste de hipótesis para extraer conclusiones relativas al estado de cada paciente después de ser diagnosticadas y recibir los diferentes tratamientos.

# 2. Descripción del dataset

El dataset, cuyo título es “GBSG2” presenta información sobre un grupo de 686 pacientes que fueron diagnosticadas con cáncer de mama y se sometieron a cirugía para extraer el tumor, recibiendo distintos tratamientos en base a la recomendación médica para cada caso.

Resulta muy interesante conocer el tipo de cáncer, la cirugía que se escogió para eliminarlo y si la paciente sobrevivió al tratamiento pues puede indicar patrones capaces de predecir cuál es la mejor opción para cada tipo de cáncer, las posibilidades de recuperarse del mismo y relacionarlo con datos clínicos y personales de cada paciente, viendo así qué factores son los más influyentes en el tratamiento, incluyendo el tipo de vida media, los factores de riesgo y si se aplicó tratamiento hormonal o no.

Lo que nos ha inspirado para realizar este análisis ha sido un interés particular en cómo en la era del big data es posible utilizar tecnologías baratas y asequibles que pueden ahorrar años de estudios realizados mediante métodos obsoletos a la hora de obtener resultados que permitan el avance científico en campos donde es especialmente necesario.

El cáncer de mama no entiende de ideologías, clases sociales o nacionalidades, es una enfermedad compleja de la que aún no se conocen muchos aspectos cruciales y que puede afectar a cualquier mujer del mundo, este trabajo tiene como objetivo dar a conocer la enfermedad para que se invierta más en estudios y tratamientos con el fin de reducir al máximo la mortalidad que causa y también servir como un recordatorio a aquellas personas que la hayan padecido o conozcan a alguien en dicha situación de que es una lucha global en la que todos remamos en la misma dirección.

Antes de aplicar el análisis, hemos realizado una exploración inicial de los datos que en estadística se conoce como análisis univariante.

Este paso es fundamental pues nos permite obtener información relevante sobre los datos a tratar, conocer cuáles son los valores de los campos del conjunto y cómo se distribuyen.Por ello, para cada atributo de los datos hemos generado algunas gráficas que mostraremos a continuación.

** Las gráficas se han aplicado sobre los campos una vez realizada la limpieza de los mismos.

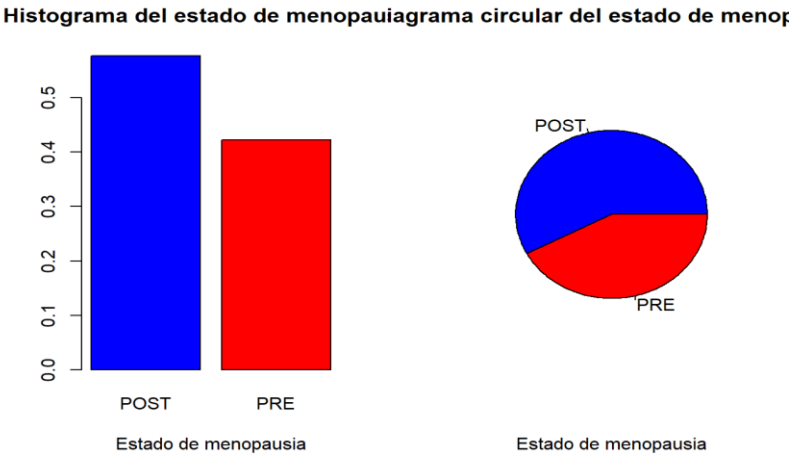
- **Exploración del conjunto.**

```
# Resumen del dataframe
summary(GBSG2)
# Imprimir los primeros registros del conjunto
head(GBSG2)
##   horTh age menostat tsize tgrade pnodes progrec estrec time cens
## 1    no  70     Post    21     II      3      48     66 1814    1
## 2   yes  56     Post    12     II      7      61     77 2018    1
# Inspeccionar el tipo de los campos
sapply(GBSG2, function(x) class(x))
## $horTh
## [1] "factor"
## $age
## [1] "integer"
## $menostat
## [1] "factor"
## $tsize
## [1] "integer"
## $tgrade
## [1] "ordered" "factor"
## $pnodes
```

```
## [1] "integer"
## $progrec
## [1] "integer"
## $estrec
## [1] "integer"
## $time
## [1] "integer"
## $cens
## [1] "integer"
```

- **Estado de menopausia (menostat).**

  Indica el estado de menopausia de la paciente en el momento de realización del estudio. Los valores del dataset original pueden ser 'Pre' (no tiene la menopausia) p 'Post' (tiene la menopausia).

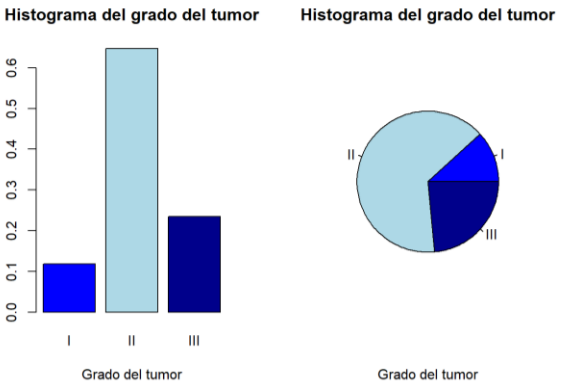



- **Tratamiento hormonal (horTh).**

  Puede tener el valor 'yes' si la paciente recibió tratamiento hormonal o 'no' si no lo recibió.

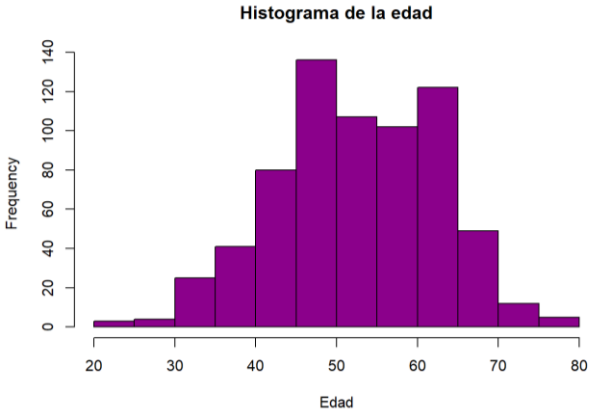
- **Grado del tumor (tgrade).**

  Hace referencia al grado del tumor que puede tener los valores 'I', 'II' y 'III'. En los análisis posteriores estudiaremos sí un grado de tumor mayor hace que la posibilidad de supervivencia sea menor.



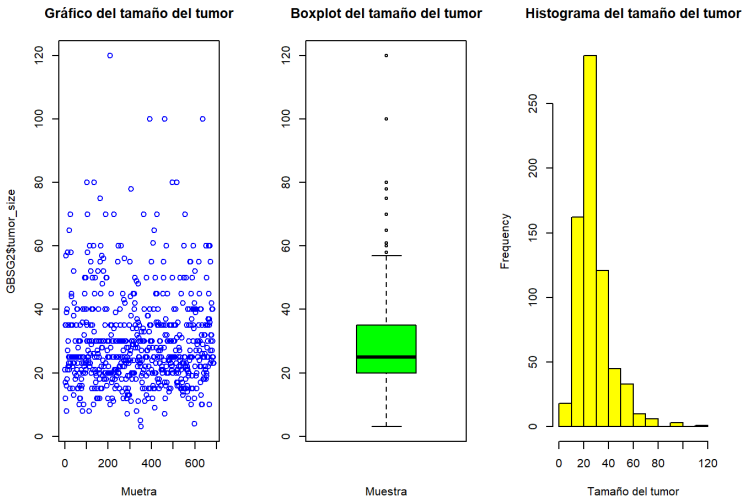
- **Edad del paciente (age).**

  Edad que tenían las pacientes en el momento de estudio. Es un valor numérico natural.




- **Tamaño del tumor (tsize).**

  Valor numérico que indica el tamaño del tumor en milímetros.

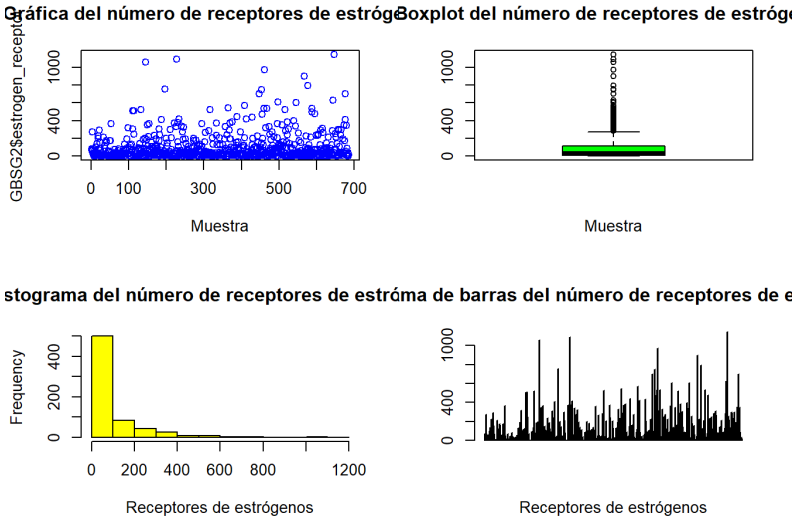



- **Número de nodos positivos.**

  Valor numérico con el número de nodos positivos o número de ganglios linfáticos que contienen cáncer.

- **Receptores de estrógenos (estrec).**

  Número de receptores de estrógenos medido en fmol (10-15 moles).

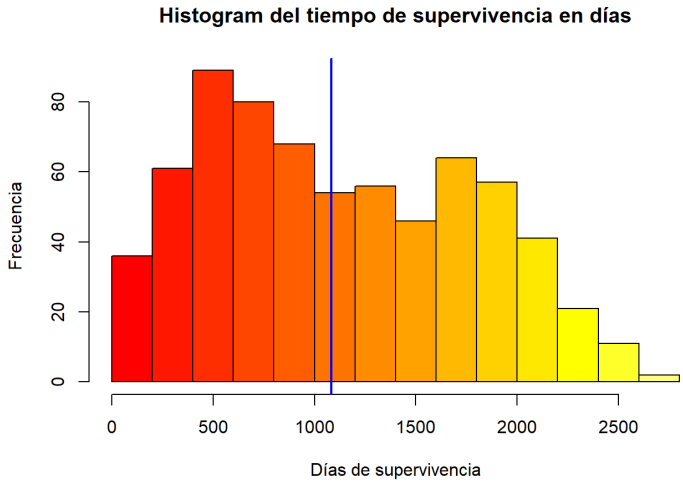



- **Receptores de progesterona (progrec).**

  Número de receptores de estrógenos medido en fmol (10-15 moles).

- **Tiempo de supervivencia (time).**

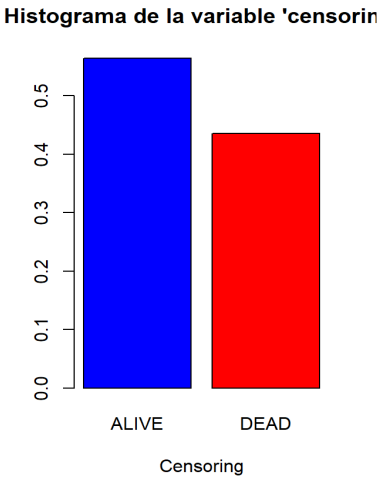
  Valor numérico que indica el número de días de supervivencia de la paciente hasta el momento de estudio. Si la paciente había muerto, ese tiempo indica exactamente el tiempo sobrevivido, pero si la paciente estaba viva, ese valor no aporta información exacta del tiempo de supervivencia.

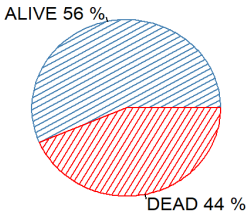



```
paste("Mediana de superviviencia:", median(GBSG2$survival_time))
## [1] "Mediana de superviviencia: 1084"
paste("Media de superviviencia:", mean(GBSG2$survival_time))
## [1] "Media de superviviencia: 1124.48979591837"
```

- **Indicador de censura (cens)**

  Puede tener un valor de 1 si la paciente ha muerto o un valor de 1 si no ha ocurrido el evento en el momento de estudio, por lo que no sabemos si después del estudio esa paciente murió o no.




# 2. Pregunta a resolver

Queremos saber cuáles son los factores que mayor importancia tienen en la supervivencia del paciente. **¿Qué factores tienen mayor influencia en la supervivencia?**

Algunas de las cosas que sería interesante comprobar son:

- ¿Hay más posibilidades de que una paciente de 50 años muera, que lo haga una de 25, es decir, influye la edad?
- ¿Depende el fallecimiento de las pacientes del estado o del tamaño del tumor?
- ¿Hay relación entre los valores de los receptores de estrógenos y progesterona?

# 3. Integración y selección

Debido a que los datos estaban correctamente integrados y tampoco era necesario seleccionar un grupo concreto de estos, hemos empleado los mismos datos que venían en el dataset original.

# 4. Limpieza de los datos

A continuación, listamos cada una de las transformaciones realizadas sobre el conjunto original con el fin de identificar registros incompletos o incorrectos con el fin de corregirlos o eliminarlos si se considera necesario. El objetivo final de esta etapa del ciclo de vida de los datos es mejorar la calidad

de los datos para poder generar modelos robustos y generalizables, por lo que es una de las fases más importante debido a la influencia que tiene sobre el resto de las fases.

Algunos de los métodos empleados en esta etapa son:

- **Eliminación de duplicados.**
- **Filtrado de datos.**
- **Conversión de campos.**
- **Manejo de datos perdidos (NA) o vacíos.**
- **Manejo de valores extremos (outliers).**

Transformaciones aplicadas:

❖ **Renombrar las columnas.**

El nombre de los atributos originales es bastante confuso y difícilmente identificable, por lo que se ha realizado una 'traducción' de los campos por nombres que aportan más información y facilitan la comprensión de los datos. Debemos tener en cuenta que unas buenas prácticas a la hora de trabajar con los datos y a la hora de diseñar el código, hacen que se ahorre mucho tiempo y que los resultados sean más claros.

```
####### LIMPIEZA DE LAS COLUMAS Y REGISSTROS DUPLICADOS ######

# Columnas originales

colnames(GBSG2)

##  [1] "horTh"    "age"      "menostat" "tsize"    "tgrade"   "pnodes"

##  [7] "progrec"  "estrec"   "time"     "cens"

# Renombrar las columnas por un nombre más relevante

colnames(GBSG2)    <-    c("hormonal_therapy",    "age",    "menopausal_status",
"tumor_size",    "tumor_grade",    "positive_nodes",    "progesterone_receptor",
"estrogen_receptor", "survival_time", "censoring")


colnames(GBSG2)

## [1] "hormonal_therapy"      "age"                "menopausal_status"

## [4] "tumor_size"            "tumor_grade"           "positive_nodes"

## [7] "progesterone_receptor" "estrogen_receptor"     "survival_time"

## [10] "censoring"
```

❖ **Eliminar duplicados.**

En este caso no parece haber duplicados porque los datos ya estaban procesados anteriormente y por eso tampoco existe un campo identificador de cada paciente. Sin embargo, en un caso como este, no tendría sentido tener dos registros con el mismo identificador de la paciente y habría que eliminar uno de ellos. Por norma general se elimina el primero que se introdujo pues es el más antiguo, pero depende del problema.

```
GBSG2 <- unique(GBSG2)
```

- **Eliminar valores vacíos o NA.**

A la hora de eliminar NAs, hay varias formas de hacerlo, siendo una de las más comunes en R usar la función na.omit() que elimina las filas que contienen NA. Sin embargo, hemos considerado que en este caso no es la mejor aproximación pues podemos querer aplicar cambios distintos dependiendo del campo. Es decir, en algunos casos sí que querríamos eliminar los registros donde un campo X es NA, pero en otros querríamos sustituir los valores NA de un campo Y por 0 o por 'no consta', por ejemplo.

```
GBSG2 <- GBSG2[GBSG2$hormonal_therapy != "" | ! is.na(GBSG2$hormonal_therapy),]

GBSG2 <- GBSG2[GBSG2$censoring != "" | ! is.na(GBSG2$censoring),]
```

Para ambos atributos queremos eliminar los registros con NA o "" pues no tiene sentido que esos campos estén incompletos y el análisis no sería correcto.

- **Conversión de tipos.**

```
GBSG2$menopausal_status <- factor(GBSG2$menopausal_status, levels = c("Post",
"Pre"), labels = c("POST", "PRE"))

GBSG2$hormonal_therapy <- factor(GBSG2$hormonal_therapy, levels = c("no",
"yes"), labels = c("NO", "YES"))

GBSG2$censoring <- factor(GBSG2$censoring, levels = c(0, 1), labels = c("ALIVE",
"DEAD"))
```

- Gestión de valores extremos.

No hemos considerado necesario eliminar ningún outlier porque no se tratan de errores en la creación de los datos, si no valores que realmente se dan en algunas pacientes y que nos podrían revelar patrones interesantes.

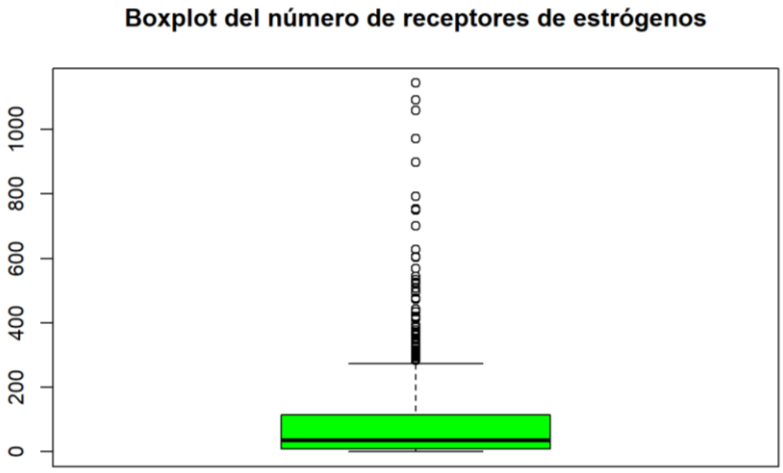
```
boxplot.stats(GBSG2$age)$out

## [1] 21
```

```
sort(boxplot.stats(GBSG2$tumor_size)$out)
##  [1]  58  58  58  60  60  60  60  60  60  60  60  60  60  60  61  65  65  70  70
## [20]  70  70  70  70  70  75  78  80  80  80  80 100 100 100 120
```

```
sort(boxplot.stats(GBSG2$positive_nodes)$out)

##  [1] 17 17 17 17 17 18 18 18 18 18 19 19 19 19 19 20 20 20 21 23 24 24 26 30 33
## [26] 35 36 38 51
```

```
sort(boxplot.stats(GBSG2$progesterone_receptor)$out)
##  [1]  320  323  324  328  340  340  345  345  345  349  350  356  360  364  364
## [16]  365  366  370  375  386  388  388  390  390  395  401  402  402  403  405
## [31]  406  408  408  412  422  423  431  432  437  462  472  488  502  505  525
```

```
## [46]  530  542  550  558  595  624  638  680  739  796  845  858  860  912  935
## [61]  980 1118 1152 1356 1490 1600 2380
sort(boxplot.stats(GBSG2$estrogen_receptor)$out)
##  [1]  284  286  287  288  288  293  294  298  299  300  304  306  307  312  315
## [16]  317  325  328  329  334  338  339  346  348  350  353  361  363  365  366
## [31]  369  371  372  378  386  386  394  412  413  418  419  435  442  472  477
## [46]  496  507  508  521  522  526  533  533  534  544  569  604  606  628  700
## [61]  701  749  753  792  898  972 1060 1091 1144
```




# • 4. Selección de subconjuntos

Hemos generado algunos grupos que podrían ser interesantes para otros análisis pero en este caso no se usarán porque el número de registros del conjunto no es demasiado grande y si además aplicamos un filtro será menor todavía, por lo que los resultados no serán demasiado precisos.

```
## Agrupación por edad
GBSG2_menor40 <- GBSG2 %>% dplyr::filter(age <= 40)
GBSG2_mayor40_menor65 <- GBSG2 %>% dplyr::filter(age > 40 & age < 65)
GBSG2_mayor65 <- GBSG2 %>% dplyr::filter(age >= 65)

## Agrupación por grado de tumor
GBSG2_gradoI <- GBSG2[GBSG2$tumor_grade == "I",]
GBSG2_gradoII <- GBSG2[GBSG2$tumor_grade == "II",]
GBSG2_gradoIII <- GBSG2[GBSG2$tumor_grade == "III",]

## Agrupación por grado de tumor
GBSG2_menopausal_post <- GBSG2 %>% dplyr::filter(menopausal_status == "POST")
GBSG2_menopausal_pre <- GBSG2 %>% dplyr::filter(menopausal_status == "PRE")

## Agrupación por tamaño de tumor
```

```
GBSG2_tumor_size_menor10 <- GBSG2[GBSG2$tumor_size <= 10,]
GBSG2_tumor_size_mayor10_menor40 <- GBSG2[GBSG2$tumor_size > 10 & GBSG2$tumor_size <= 40,]
GBSG2_tumor_size_mayor40 <- GBSG2[GBSG2$tumor_size > 40,]
```

# 5. Análisis de variables del dataset

Realizamos un análisis a continuación en que comprobamos las relaciones entre algunos de los atributos del conjunto.

Debemos tener en cuenta que el tamaño del dataset no es demasiado grande por lo que es entendible que no se cumpla el criterio de normalidad para la mayoría de los casos.

## 1-Dependencia, normalidad y homocedasticidad de las variables: age y censoring

En primer lugar, comprobamos la normalidad de las variables con el test mediante el test de shapiro y el test de Kolmogorov.

```
shapiro.test(GBSG2$age)

##  Shapiro-Wilk normality test

## data:  GBSG2$age

## W = 0.99079, p-value = 0.0002774



ks.test(GBSG2$age, pnorm, mean(GBSG2$age), sd(GBSG2$age))

##  One-sample Kolmogorov-Smirnov test

## data:  GBSG2$age

## D = 0.065642, p-value = 0.005415

## alternative hypothesis: two-sided
```

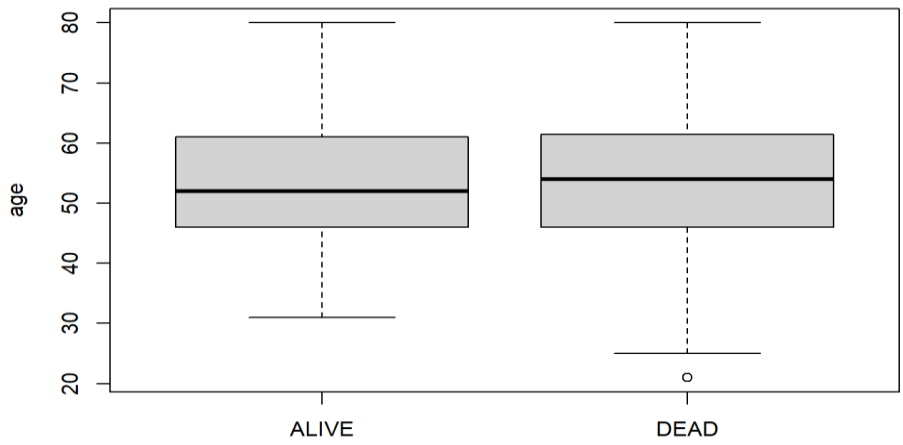
Vemos que la edad no sigue una distribución normal para ninguno de las dos pruebas. No sería necesario calcular la homocedasticidad, pero lo vamos a hacer para ver qué resultados se obtienen.

```
fligner.test(age ~ censoring, data = GBSG2)

##  Fligner-Killeen test of homogeneity of variances

## data:  age by censoring

## Fligner-Killeen:med chi-squared = 4.3452, df = 1, p-value = 0.03711
```

Por lo tanto, tampoco cumple con el criterio de homocedasticidad. Al ser una variable numérica y una categórica, y no se cumple la normalidad, hay que aplicar una prueba no paramétrica como Wilcoxon

```
# Al ser variable numérica y categórica, y no se cumple la normalidad, hay que aplicar la
prueba no paramétrica como Wilcoxon

wilcox.test(age ~ censoring, data = GBSG2)

##  Wilcoxon rank sum test with continuity correction

## data:  age by censoring

## W = 56815, p-value = 0.6857

## alternative hypothesis: true location shift is not equal to 0
```

Por tanto, podemos afirmar que no existe correlación entre las variables, lo que se puede comprobar en el siguiente diagrama ya que no hay diferencia de edades para un evento u otro.

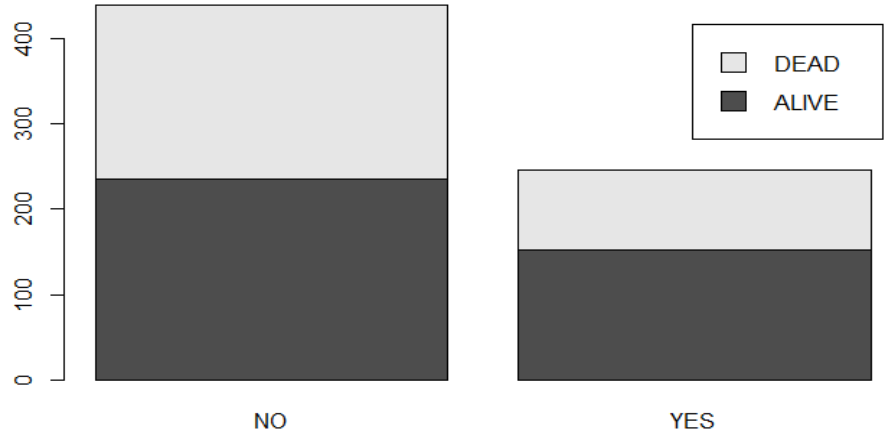



## ○ 2-Realizamos un test entre las variables hormonal_therapy y patient_tatus

Al ser las dos variables categóricas usamos el test de chi cuadrado para comprobar si la recepción de terapia hormonal influye en la censura.

```
tabla_relacion <- xtabs(~ censoring + hormonal_therapy , data = GBSG2)

chisq.test(tabla_relacion)$p.value

##  Pearson's Chi-squared test with Yates' continuity correction

## data:  tabla_relacion

## X-squared = 4.1714, df = 1, p-value = 0.04111
```

Como vemos el valor es menor que 0.05, por lo que sí que existe relación entre ambas variables.




Esta relación se observa en el gráfico superior donde vemos que para aquellas personas que no han recibido la terapia la posibilidad de no superar la enfermedad es ligeramente mayor, mientras que para las que sí la han recibido, la probabilidad es bastante menor.

### ○ 3-Realizamos la comprobación entre tumor_size y censoring

Queremos saber si el hecho de que el tumor sea más grande hace que la probabilidad de supervivencia sea menor o no. En primer lugar, comprobamos la normalidad y homocedasticidad de la variable tumor_size.

```
shapiro.test(GBSG2$tumor_size)

##  Shapiro-Wilk normality test

## data:  GBSG2$tumor_size

## W = 0.87236, p-value < 2.2e-16


fligner.test(tumor_size ~ censoring, data = GBSG2)

##  Fligner-Killeen test of homogeneity of variances

## data:  tumor_size by censoring

## Fligner-Killeen:med chi-squared = 1.5058, df = 1, p-value = 0.2198
```

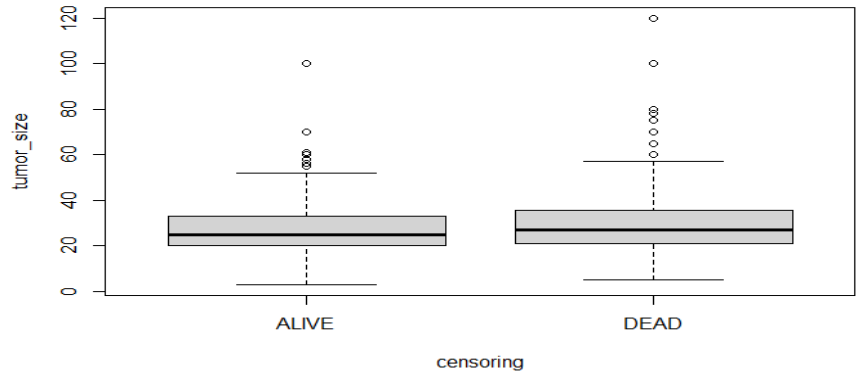
Resultado:

**La variable tumor_size no sigue una distribución normal según el test de Shapiro.**

**Las variables cumplen el criterio de homocedasticidad según el test de Fligner-Killeen.**

Al ser variable numérica y categórica, y no se cumple la normalidad, hay que aplicar la prueba no paramétrica como Wilcoxon.

```
wilcox.test(tumor_size ~ censoring, data = GBSG2)

##  Wilcoxon rank sum test with continuity correction

## data:  tumor_size by censoring

## W = 49710, p-value = 0.001525

## alternative hypothesis: true location shift is not equal to 0
```

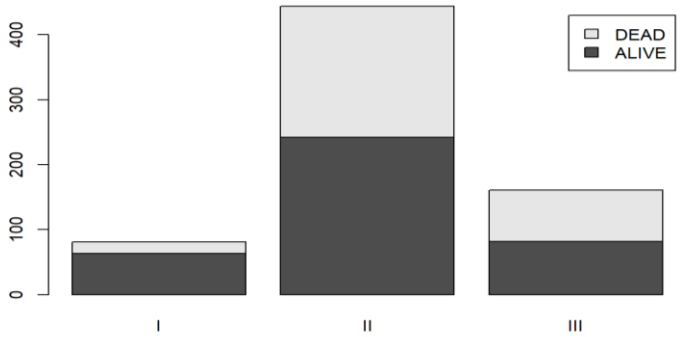



Tal y como se observa en el diagrama de cajas, el tamaño del tumor suele ser mayor en pacientes que no han superado la enfermedad.

- **4-Realizamos la comprobación de dependencia de las variables categóricas: tumor_stage y censoring.**

¿Una paciente con un estado del tumor III tiene menos posibilidades de supervivencia que una con estado I?

```
tabla_relacion <- xtabs(~ censoring + tumor_grade , data = GBSG2)

chisq.test(tabla_relacion)

##  Pearson's Chi-squared test

## data:  tabla_relacion

## X-squared = 17.662, df = 2, p-value = 0.0001462
```




Tanto la imagen como los resultados del test revelan que sí existe una dependencia entre las variables.

## ○ 5-Comprobamos dependencia, normalidad y homocedasticidad de las variables numéricas: progesterone_receptor y estrogen_receptor

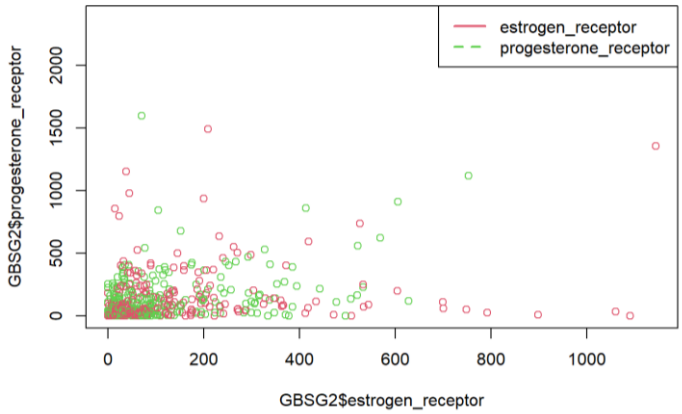
Además de conocer cómo influyen las variables independientes sobre la variable dependiente, es interesante conocer cómo se relacionan los atributos entre sí, como es el caso de los receptores de estrógenos y progesterona.

```
shapiro.test(GBSG2$progesterone_receptor)

##  Shapiro-Wilk normality test

## data:  GBSG2$progesterone_receptor

## W = 0.54349, p-value < 2.2e-16

shapiro.test(GBSG2$estrogen_receptor)

##  Shapiro-Wilk normality test

## data:  GBSG2$estrogen_receptor

## W = 0.63847, p-value < 2.2e-16

fligner.test(estrogen_receptor ~ progesterone_receptor, data = GBSG2)

##  Fligner-Killeen test of homogeneity of variances

## data:  estrogen_receptor by progesterone_receptor

## Fligner-Killeen:med chi-squared = 361.81, df = 241, p-value = 7.428e-07
```

Por lo tanto ni siguen una distribución normal ni cumplen los criterios de homocedasticidad, por lo que emplearemos la correlación de spearman para comprobar la dependencia de las variables.

```
cor.test(GBSG2$estrogen_receptor, GBSG2$progesterone_receptor, method="spearman")

##  Spearman's rank correlation rho

## data:  GBSG2$estrogen_receptor and GBSG2$progesterone_receptor

## S = 21638374, p-value < 2.2e-16

## alternative hypothesis: true rho is not equal to 0

## sample estimates:

##       rho

## 0.5978348
```

Por tanto, sí que existe correlación entre las variables, lo que se puede comprobar en el siguiente diagrama de puntos.




## ○ 6-Aplicamos la regresión logística y la regresión lineal

Regresión logística

```
modelo_glm <- glm(as.numeric(GBSG2$censoring == "ALIVE") ~ tumor_grade, data =
GBSG2, family = binomial("logit"))

calidad <- stepAIC(modelo_glm, direction = "backward")
```

Regresión lineal

```
modelo_lm <- lm(as.numeric(GBSG2$censoring == "ALIVE") ~ tumor_size + tumor_grade +
hormonal_therapy, data = GBSG2)

## Residual standard error: 0.4864 on 681 degrees of freedom

## Multiple R-squared:  0.04501,    Adjusted R-squared:  0.0394

## F-statistic: 8.025 on 4 and 681 DF,  p-value: 2.524e-06
```

# 6.Agradecimientos

*Agradecemos a la web https://www.rdocumentation.org por proporcionar el dataset completo, a https://www.datacamp.com por alojar en sus servidores parte de la documentación y al Bundesministerium für Gesundheit (https://www.bundesgesundheitsministerium.de) por hacer públicos los datos, así como por la recopilación de los mismos.*

# 7.Vídeo

# 8. Resolución del problema

Tras haber realizado los análisis pertinentes podemos concluir que quedan todavía muchas variables para ser descubiertas y agregadas al análisis, algo que es responsabilidad de los investigadores y los médicos, sin embargo, con las variables disponibles se demuestra que:

**1-** El tratamiento hormonal aumenta sustancialmente las posibilidades de supervivencia.

**2-**El grado del tumor es crucial para poder eliminarlo y que la paciente se recupere, así en los casos en los que el tumor pertenece al grado I las probabilidades de sobrevivir al tratamiento y sanar son mucho mayores que en los grados II y III.

**3-**El tamaño del tumor resulta ser un factor clave en una amplia mayoría de casos, siendo 19mm la frontera. Un tamaño menor aumenta significativamente las probabilidades de supervivencia, pero un tamaño mayor las reduce considerablemente.

# 9.Tabla de contribuciones

| Contribuciones | Firma |
|---|---|
| Investigación previa | FPS, RPS |
| Redacción de las respuestas | FPS, RPS |
| Desarrollo del código | FPS, RPS |
| Participación en el vídeo | FPS, RPS |

# 10. Bibliografía

https://rpubs.com/odenipinedo/survival-analysis-in-R https://r-coder.com/grafico-sectores-r/

https://picandoconr.wordpress.com/2016/06/09/paleta-de-colores/

https://seom.org/informacion-sobre-el-cancer/ique-es-la-medicina-de-precision

https://jllopisperez.com/2013/01/07/tema-21-analisis-de-supervivencia/#:~:text=En%20An%C3%A1lisis%20de%20supervivencia%20la,que%20entran%20en%20el%20estudio

https://www.fisterra.com/mbe/investiga/supervivencia/supervivencia.asp

https://es.wikipedia.org/wiki/An%C3%A1lisis_de_la_supervivencia

https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-40262007000100013

https://rpkgs.datanovia.com/survminer/reference/ggsurvplot.html

https://www.r-bloggers.com/2021/08/how-to-plot-categorical-data-in-r-quick-guide/

*Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.*

*M. Schumacher, G. Basert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L.A. Neumann and H.F. Rauschecker for the German Breast Cancer Study Group (1994), Randomized \(2\times2\) trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. Journal of Clinical Oncology, 12, 2086--2093.*