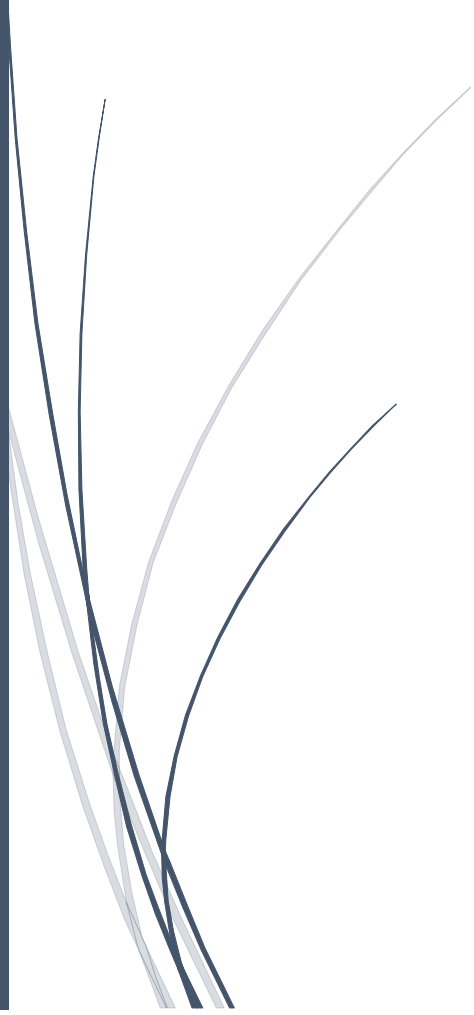
7-1-2019

PRÁCTICA 2:

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS



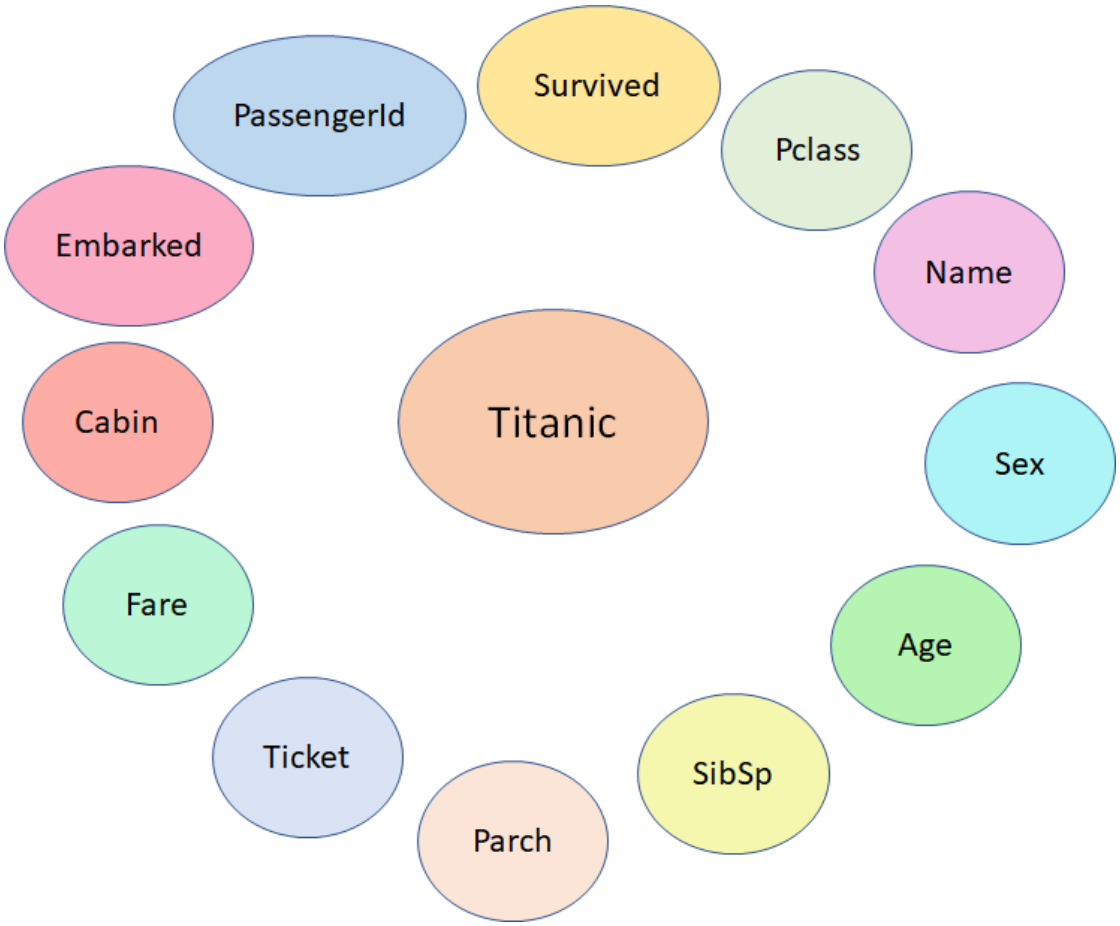
MARTA MARTÍNEZ ROMAY
UOC

Indice

1. Descripción del dataset 2
2. Integración y selección de los datos de interés a analizar. 3
3. Limpieza de los datos. 4
   3.1. Ceros y elementos vacíos. 4
   3.2. Valores extremos (outliers). 6
   3.3. Exportación de los datos preprocesados. 8
4. Análisis de datos. 9
   4.1. Selección de los grupos de datos que se quieren analizar/comparar. 9
   4.2. Comprobación de la normalidad y homogeneidad de la varianza. 9
   4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. 11
5. Representación de los resultados a partir de tablas y gráficas. 14
6. Resolución del problema. 16
7. Código en R. 16

1. Descripción del dataset

El dataset elegido es “datosTitanic” obtenido a partir del enlace en Kaggle (https://www.kaggle.com/c/titanic). Este contiene datos de los pasajeros del Titanic que fue uno de los naufragios más terribles de la historia. Este dataset está formado por 12 variables (columnas o atributos) y 891 pasajeros (filas o nº de muestras).



Vamos a explicar las variables del dataset:

- **PassegerId**: es simplemente un contador de pasajeros del 1 al 891.
- **Survived**: esta variable toma dos valores e indica si el pasajero sobrevivió. (0=”No”, 1=”Si”).
- **pClass**: clase del ticket. 1=1st (clase alta), 2=2nd (clase media) y 3=3rd (clase baja).
- **Name**: nombre completo del pasajero.
- **Sex**: sexo del pasajero (Female o Male).
- **Age**: edad del pasajero.

- **SibSp**: número de hermanos/hermanas, hermanastros/hermanastros y marido o esposa del pasajero que también iban a bordo. (No contarían los amantes y los novios).
- **Parch**: número de hijas, hijos, padre y madre del pasajero a bordo del Titanic. Para los niños que viajaban con la niñera esta variable sería 0.
- **Ticket:** El número del ticket del pasajero.
- **Fare:** Es la tarifa del pasajero en dólares.
- **Cabin**: Número de la cabina.
- **Embarked**: el puerto en el que embarcó el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

Este dataset nos permite analizar cuál es el grupo de personas que más probabilidad tienen de sobrevivir (entendemos que fueron los niños y las mujeres), podremos ver si existió algún motivo más que influyó en la supervivencia de los pasajeros con la finalidad de maximizar el número de supervivientes en otros naufragios.

El objetivo principal de este estudio es predecir si el pasajero sobrevivió según su edad, sexo, clase del ticket que haya comprado, etc…
Este conjunto de datos también se puede utilizar para la agrupación en clústeres y tal vez encontrar diferentes clústeres de pasajeros dentro de él.

Al final del estudio tendremos un conjunto de datos de test que le aplicaremos algún algoritmo después de ser entrenado con el conjunto de datos inicial (datos de entrenamiento) para poder predecir si sobrevivieron o no al naufragio esos pasajeros.

2. Integración y selección de los datos de interés a analizar.

Para llevar a cabo nuestro principal objetivo que es predecir si un pasajero sobrevive o no, nos interesa la variable respuesta “Survived”. Todos los atributos son características del pasajero correspondiente del Titanic, por tanto, en un primer momento nos interesan para el análisis.

Si nos fijamos en el tipo de variable que es cada una :

```
> sapply(data, function(x) class(x))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp
 Parch
  "integer"   "integer"   "integer"    "factor"    "factor"   "numeric"   "integer"   "in
teger"
     Ticket        Fare       Cabin    Embarked
   "factor"   "numeric"    "factor"    "factor"
```

Tenemos variables categóricas que toman un número finito de valores numéricos: **Survived, Pclass, SibSp, Parch y PassengerId.**

Las variables de tipo factor serían: **Name, Sex, Ticket, Cabin y Embarked**.

Y las únicas variables continuas del dataset serían: **Age y Fare**.

Podríamos eliminar las columnas de "**PassengerId**", "**Ticket"** y **'Name'** ya que no van a tener impacto en la variable que queremos predecir ('Survived').

Quizás el nombre podría tener relevancia para definir el género del pasajero pero en esta práctica no va a ser caso de estudio.

Es bastante obvio que estos datos no interesan para el análisis, pero quizás haya otras no tan obvias que iremos descubriendo a medida que vayamos realizando el análisis de los datos. Mediante técnicas de reducción de las dimensiones ya eliminaremos las variables que no sean significativas para el análisis y que no afecten en la respuesta.

3. Limpieza de los datos.

3.1. Ceros y elementos vacíos.

Previo a la lectura de los datos y si nos fijamos en el archivo podemos ver que hay campos que no tienen datos. Cuando leemos los datos con R estos datos vacíos aparecen como NA, que se denominan valores erróneos o perdidos.

Veamos cuantos valores vacíos hay y en qué atributos:

```
> sapply(data,function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin Embarked
       0        0        0      177        0        0        0      687        2
```

Es decir, hay 177 pasajeros de los que no se conoce la edad, 687 valores de la cabina perdidos y 2 pasajeros que no sabemos el puerto donde embarcaron. Es lógico que la mayoría de estos sean personas que no han sobrevivido al naufragio, y no se ha podido conocer este dato sobre ellos. De los que sobrevivieron, pueden ser valores que se hayan perdido en la toma de los datos, errores, etc…

Llegados a este punto y con la finalidad de tener la muestra preparada para la analítica tenemos que decidir qué hacer con estos datos. Lo ideal sería intentar conseguir estos datos faltantes, hablando con la persona que recopiló los datos o realizando de nuevo la extracción de ellos. En este caso eso no es posible, por tanto, tenemos las siguientes opciones:

1) **Eliminar las muestras que contengan estos valores erróneos:**

```
> #Eliminar los registros con valores erróneos.
> datasinNA<- na.omit(data)
> dim(datasinNA)
[1] 183   9
```

Pasamos así de tener 891 muestras a 183 muestras (eliminando las muestras en las que haya algún valor erróneo). Vemos así que el total de pasajeros con alguna variable nula es **708**.

Si eliminamos estos registros estamos perdiendo información de una variable importante (y un número alto de registros, 708). Tendríamos la muestra sesgada de forma que los resultados finales no serán fiables ni realistas.

Por los motivos previamente expuestos, no vamos a eliminar los registros y vamos a trabajar con aproximaciones de estos datos.

2) **Reemplazar los datos erróneos.**

Lo que estaríamos haciendo aquí sería aproximar el dato inexistente por otro valor lo más preciso posible.

En el caso de la variable 'Age' podemos hacerlo sustituyendo el valor por la media del resto de variables (la mediana en caso de no ser una variable continua). Así las edades vacías las sustituimos por la **edad media** del resto de muestras:

```
> media= mean(data$Age, na.rm = TRUE)
> data$Age<- replace(data$Age, is.na(data$Age), ceiling(media))
> media
[1] 29.69912
```

También podemos aplicar el algoritmo de **los k vecinos más próximos** que consiste en agrupar las muestras según la similitud del valor del resto de variables. Es decir, si queremos predecir la variable 'Age' le asignaremos el valor que tengan los k vecinos más próximos (en los que se parezcan más los valores del resto de las variables). Estamos suponiendo entonces que nuestros datos guardan cierta relación entre ellos, lo cual parece tener sentido.

Aplicamos así el algoritmo en las 3 variables con errores y comprobamos que efectivamente ya no tenemos valores NaN en ninguna variable.

```
> suppressWarnings(suppressMessages(library(VIM)))
> data$Age <- kNN(data)$Age
> sapply(data, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin Embarked
       0        0        0        0        0        0        0      687        2
> data$Cabin<-kNN(data)$Cabin
> sapply(data, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin Embarked
       0        0        0        0        0        0        0        0        2
> data$Embarked<-kNN(data)$Embarked
> sapply(data, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin Embarked
       0        0        0        0        0        0        0        0        0
```

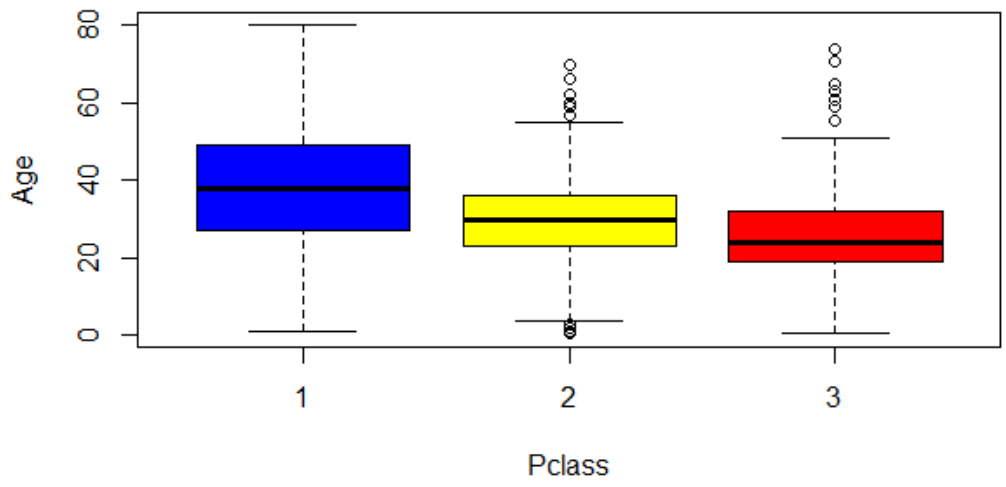
Nos quedamos así con esta última opción de sustituir los valores perdidos ya que, tiene sentido que los pasajeros tengan cierta relación en la edad. Además, de esta forma representamos mucho mejor la realidad.

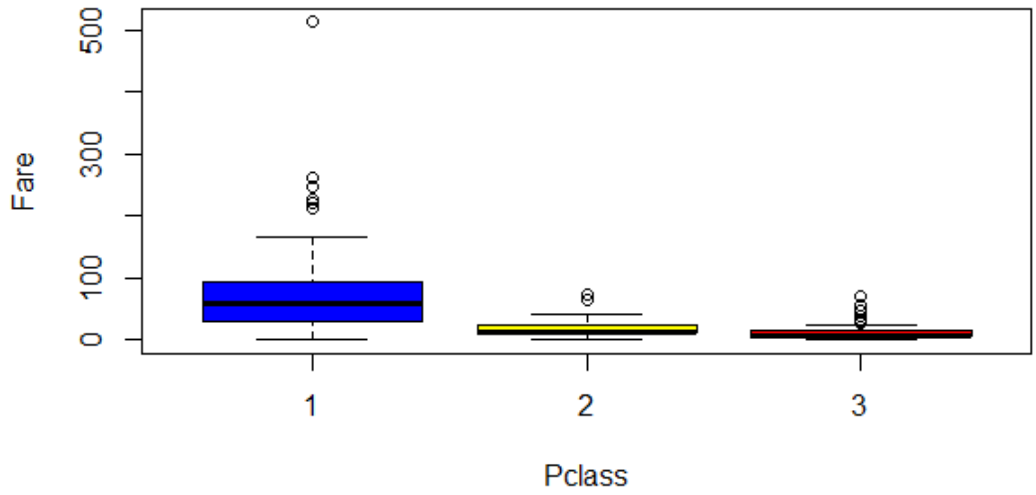
3.2. Valores extremos (outliers).

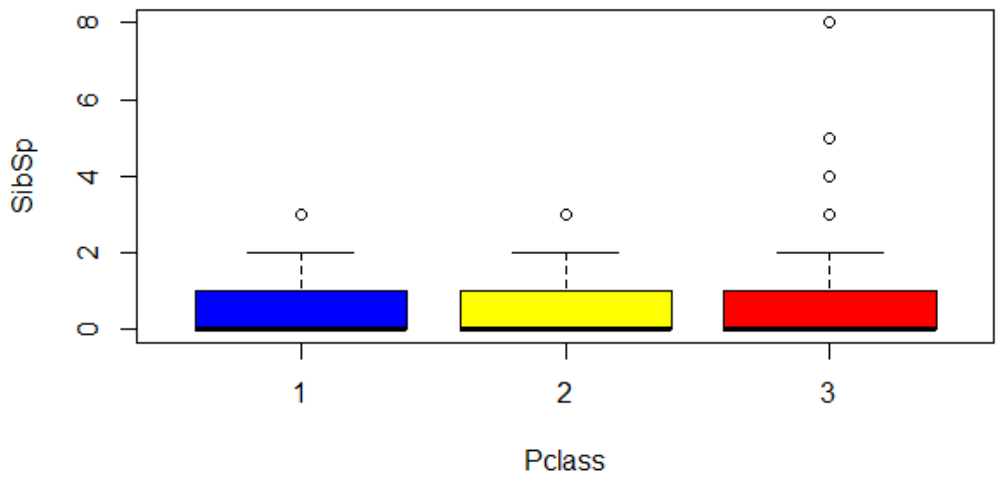
Los valores extremos (**valores extremos/outliers**) son aquellos que se consideran diferentes a los otros datos de la muestra. A la hora de analizar el modelo que incluye estos datos, debemos ver cuales se separan mucho del comportamiento esperable bajo el modelo. Puede ser que estos datos sean erróneos y estén mal tomados, en este caso sería conveniente eliminarlos o modificarlos o puede que sean correctos y sean interesante porque si se ignoran, puede haber cambios importantes en las conclusiones obtenidas del estudio.

Como siempre se ven mejor las cosas gráficamente vamos a representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico.

Un diagrama de cajas es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles.







Si nos fijamos en los gráficos vemos así los valores extremos según la clase del billete de cada pasajero (Pclass).

En el segundo gráfico parece lógico que en la tercera clase el precio del billete (Fare) sea más bajo que en la primera. Aparece algún outlier sobre todo en la primera clase.

En el boxplot de la edad también podemos ver que hay más gente adulta viajando en primera clase que en clases más bajas y en esta clase no existen valores extremos de la variable 'Age'.

Si utilizamos la función boxplots.stats() de R, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
> boxplot.stats(data$Age)$out
 [1] 66.0 65.0 71.0 70.5 63.0 65.0 64.0 65.0 63.0 71.0 64.0 80.0 70.0 70.0 74.0
> boxplot.stats(data$SibSp)$out
 [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3 4 8
[41] 4 3 4 8 4 8
> boxplot.stats(data$Fare)$out
  [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
  [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
 [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
 [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
 [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
 [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
 [49] 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
 [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
 [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
 [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
 [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
 [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
 [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
[105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
[113]  89.1042 164.8667  69.5500  83.1583
```

Estos datos, como es lógico, coinciden con los valores que aparecen en los boxplot.
Aunque estos datos sean anómalos en esta muestra porque son poco comunes y diferentes del resto no son cosas incoherentes.
Tiene sentido que la edad de un pasajero sea de 74 años, que viaje con 2 hermanos, 2 herman astros y con su mujer por ejemplo (dando un total de SibSp=5) o incluso una familia muy gran de en la que viajen los 7 hermanos con sus respectivas mujeres daría un valor de SibSp=8.

La tarifa de cada pasajero varía según la clase del billete, por tanto, también puede ser que de ntro de los de primera clase haya tarifas altas, incluso hay ciertas cabinas que podrían ser de lu jo más caras de lo normal (512,33):

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
259	1	1	female	35.00	0	0	512.3292	B101	C
680	1	1	male	36.00	0	1	512.3292	B51 B53 B55	C
738	1	1	male	35.00	0	0	512.3292	B101	C

Nos quedamos así con todos los valores extremos ya que parece que son aceptables y situacio nes perfectamente válidas.

3.3. Exportación de los datos preprocesados.

Después de los procedimientos de integración, validación y limpieza guardamos estos en un nuevo fichero denominado datos_titanic_limpieza.

```
#Exportación de los datos limpios en .csv
write.csv(data, "datos_titanic_limpieza.csv")
```

4. Análisis de datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar.

En este apartado se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. De todos modos, no se utilizarán todos estos grupos para los posteriores análisis estadísticos.

Podríamos agrupar por la variable pClass, teniendo así los pasajeros de la primera, segunda y tercera clase. Podríamos utilizarlo para comparar los sobrevivientes de las distintas clases y ver cuáles tienen más probabilidad de sobrevivir.

Del mismo modo agrupamos por Género ('Sex') y puerto en el que embarcaron los pasajeros ('Embarked').

```
# Agrupación por clase
pasajeros1clase<-data[data$Pclass==1,]
pasajeros2clase <- data[data$Pclass==2,]
pasajeros3clase <- data[data$Pclass==3,]

# Agrupación por género
Hombres<-data[data$Sex=='male',]
Mujeres<- data[data$Sex=='female',]

#Agrupación por puerto en el que embarcaron.
PuertoC<-data[data$Embarked=='C',]
PuertoQ <- data[data$Embarked=='Q',]
PuertoS <- data[data$embarked=='S',]
```

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

Es necesario realizar la comprobación de si nuestras muestras siguen una distribución normal ya que hay algunos análisis que solamente produce resultados fiables para poblaciones normales, por ejemplo, el ANOVA (Análisis de la varianza).

Para analizar el cumplimiento de la hipótesis de normalidad se pueden aplicar diferentes métodos: test de tipo Kolmogorov-Smirnov, el test de Shapiro-Wilk (para muestras pequeñas), un test ji-cuadrado (para variables cualitativas) o el test de Anderson-Darling.

Para nuestro conjunto de datos vamos a determinar si las variables cuantitativas, es decir, Pclass,Age,SibSp,Parch y Fare cumplen el supuesto de normalidad.

- Primero aplicamos el test de **Kolmogorov-Smirnov** para ver si la edad sigue o no una distribución normal.

```
> testKs

        One-sample Kolmogorov-Smirnov test

data:  data$Age
D = 0.96553, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Vemos que el p-value es <0,05 por tanto rechazamos la hipótesis de que los datos siguen una distribución normal.

Si hacemos lo mismo con todas las variables llegamos a la misma conclusión.

- La segunda prueba que se usará será la prueba de **Anderson Darling.** La función asociada a esta prueba es ad.test del paquete goftest y tiene los mismos argumentos de la prueba ks.

Las hipótesis para la prueba de Anderson-Darling son:

H0: Los datos siguen una distribución especificada (en este caso la normal).

H1: Los datos no siguen una distribución especificada (la normal).

Aplicamos el test a todas las variables de mis datos para ver cuales siguen una distribución normal y cuales no:

```
> library(nortest)
> alpha = 0.05
> col.names = colnames(data)
> for (i in 1:ncol(data))
+ { if (i == 1) cat("Variables que no siguen una distribución normal:\n")
+   if (is.integer(data[,i]) | is.numeric(data[,i])) {
+     p_val = ad.test(data[,i])$p.value
+     if (p_val < alpha) {
+       cat(col.names[i])
+   # Format output
+       if (i < ncol(data) - 1) cat(", ")
+       if (i %% 3 == 0) cat("\n")
+     }
+   }
+ }
Variables que no siguen una distribución normal:
Survived, Pclass, Age, SibSp, Parch,
Fare,
```

Vemos así que los p-valores son más pequeños que Alpha=0,05 y tenemos pruebas significativas para rechazar la hipótesis nula, es decir, ninguna de las variables sigua una distribución normal.

El test de Shapiro–Wilk o la prueba de Anderson-Darling son alternativas más potentes.

Conviene tener en cuenta que la prueba Kolmogórov-Smirnov es más sensible a los valores cercanos a la mediana que a los extremos de la distribución. La prueba de Anderson-Darling proporciona igual sensibilidad con valores extremos.

Para estudiar la **homogeneidad** de las varianzas se pueden utilizar varios tipos de test: por un lado el **test de Levene** o el test de **Fligner-Killeen.** El primero de ellos se suele utilizar con muestras distribuidas normalmente por tanto nos vamos a decantar por el segundo que funciona de forma óptima para datos que no siguen una distribución normal.

El test de **Fligner-Killeen** comprueba la homogeneidad de la varianza, en la que la hipótesis nula del test afirma que las varianzas de todos los grupos son iguales.

Vamos a aplicarle el test a los grupos formados por los pasajeros que embarcaron en los distintos puertos y ver si las varianzas de estos 3 grupos (Q, C y S) son iguales.

```
> fligner.test(data$Survived ~ data$Embarked, data= data)

        Fligner-Killeen test of homogeneity of variances

data:  data$Survived by data$Embarked
Fligner-Killeen:med chi-squared = 6.7468, df = 2, p-value = 0.03427

>
```

Rechazamos la hipótesis nula de que las varianzas son homogéneas ya que el p-value < 0,05.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

- **Regresión logística.**
  En el caso de mis datos, la variable respuesta (‘Survived’) es binaria o dicotómica (toma los valores 0 o 1), por tanto ,no podemos aplicar una regresión lineal normal. La variable binaria sigue una distribución de Bernoulli y la media será la probabilidad de éxito (,es decir, 1).

  Queremos estudiar la influencia de las variables explicativas en la variable respuesta ‘Survived’. Elegimos así 7 variables explicativas relacionadas con el pasajero y que pensamos que pueden tener efecto sobre la probabilidad de sobrevivir o no al naufragio. En concreto, las variables explicativas serían las siguientes:
  Pclass , Sex, Sibsp, Pach, Fare y Age.

  Aplicamos así el modelo de regresión logística sobre las variables explicativas.

```
> mod_log=glm(Survived~Pclass+Sexo+SibSp+Parch+Fare+Age,data=data,family=binomial(link="logit"))
> summary(mod_log)

Call:
glm(formula = Survived ~ Pclass + Sexo + SibSp + Parch + Fare +
    Age, family = binomial(link = "logit"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8775  -0.5973  -0.3797   0.6148   2.5181

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.868683   0.578446  10.146  < 2e-16 ***
Pclass      -1.291093   0.150317  -8.589  < 2e-16 ***
Sexomale    -2.735816   0.201794 -13.557  < 2e-16 ***
SibSp       -0.448269   0.113476  -3.950 7.80e-05 ***
Parch       -0.087585   0.120864  -0.725    0.469
Fare         0.001824   0.002325   0.785    0.433
Age         -0.053375   0.007985  -6.684 2.32e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  766.27  on 884  degrees of freedom
AIC: 780.27

Number of Fisher Scoring iterations: 5
```

Cabe observar que las variables Parch y Fare presentan coeficientes muy poco significativos comparados con el resto, por lo que podemos pensar que estas variables no afectan a la variable respuesta y sería lógico plantear un modelo simplificado sin estas variables. Comparamos ambos modelos mediante el método basado en la deviance:

```
> mod_log_simplif=glm(Survived~Pclass+Sexo+SibSp+Age,data=data,family=binomial(link="logit"))
> summary(mod_log_simplif)

Call:
glm(formula = Survived ~ Pclass + Sexo + SibSp + Age, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8749  -0.5912  -0.3785   0.6151   2.5319

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.036148   0.521665  11.571  < 2e-16 ***
Pclass      -1.352469   0.130574 -10.358  < 2e-16 ***
Sexomale    -2.718565   0.197615 -13.757  < 2e-16 ***
SibSp       -0.459550   0.108037  -4.254 2.10e-05 ***
Age         -0.053869   0.007953  -6.774 1.26e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  767.26  on 886  degrees of freedom
AIC: 777.26

Number of Fisher Scoring iterations: 5
```

```
> anova(mod_log_simplif,mod_log)
Analysis of Deviance Table

Model 1: Survived ~ Pclass + Sexo + SibSp + Age
Model 2: Survived ~ Pclass + Sexo + SibSp + Parch + Fare + Age
  Resid. Df Resid. Dev Df Deviance
1       886     767.26
2       884     766.27  2  0.98786
```

La función anova no nos proporciona directamente el nivel crítico, así que lo calculamos en base a la diferencia de deviance's y los grados de libertad:

```
> 1-pchisq(0.98786,2)
[1] 0.6102235
```

Como el nivel crítico es alto se puede aceptar el modelo simplificado en el cual solo se incluyen las variables explicativas Pclass,Sexo,Age y SibSp.

También se pueden utilizar procesos forward o backward de selección de variables explicativas. Utilizando estos obtenemos el mismo resultado de modelo de regresión logístico:

```
>
> step(mod_log)
Start:  AIC=780.27
Survived ~ Pclass + Sexo + SibSp + Parch + Fare + Age

         Df Deviance     AIC
- Parch   1   766.80  778.80
- Fare    1   766.92  778.92
<none>        766.27  780.27
- SibSp   1   785.57  797.57
- Age     1   816.75  828.75
- Pclass  1   843.35  855.35
- Sexo    1   997.98 1009.98

Step:  AIC=778.8
Survived ~ Pclass + Sexo + SibSp + Fare + Age

         Df Deviance     AIC
- Fare    1   767.26  777.26
<none>        766.80  778.80
- SibSp   1   790.43  800.43
- Age     1   817.36  827.36
- Pclass  1   847.77  857.77
- Sexo    1  1002.20 1012.20

Step:  AIC=777.26
Survived ~ Pclass + Sexo + SibSp + Age

         Df Deviance     AIC
<none>        767.26  777.26
- SibSp   1   790.48  798.48
- Age     1   819.32  827.32
- Pclass  1   898.24  906.24
- Sexo    1  1006.54 1014.54

Call:  glm(formula = Survived ~ Pclass + Sexo + SibSp + Age, family = binomial(link = "logit"),
    data = data)

Coefficients:
(Intercept)       Pclass     Sexomale        SibSp          Age
    6.03615     -1.35247     -2.71857     -0.45955     -0.05387

Degrees of Freedom: 890 Total (i.e. Null);  886 Residual
Null Deviance:      1187
Residual Deviance: 767.3        AIC: 777.3
```

Por último, vamos a predecir si los pasajeros del dataset “datos_test” sobrevivieron o no , a partir del modelo de regresión logística:

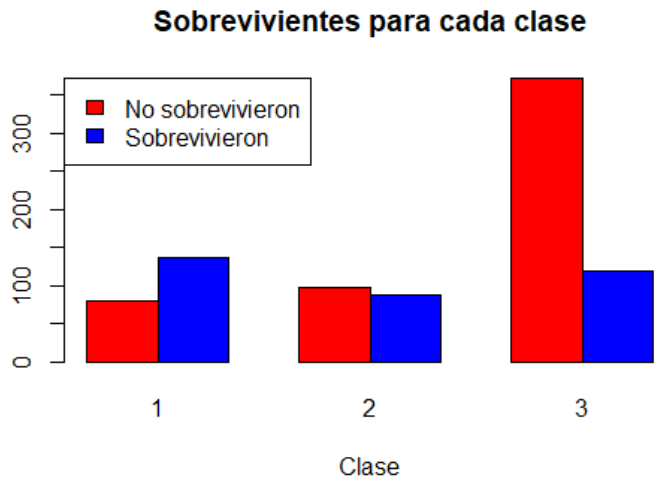
```
prediccion<-round(predict(mod_log_simplif,newdata=datos_test,type="response"))
prediccion
#Añadimos la columna de la predicción.
datos_test["Survived"]<-prediccion
```

Así, mi dataset de datos de test ahora tiene una columna más ‘Survived’ que será la variable que nos indicará si sobrevivió el pasajero (1) o no (0).

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
892	3	Kelly, Mr. James	male	34.50	0	0	330911	7.8292	F G63	Q	0
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.00	1	0	363272	7.0000	G6	S	0
894	2	Myles, Mr. Thomas Francis	male	62.00	0	0	240276	9.6875	D	Q	0
895	3	Wirz, Mr. Albert	male	27.00	0	0	315154	8.6625	F G63	S	0
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.00	1	1	3101298	12.2875	F E57	S	1
897	3	Svensson, Mr. Johan Cervin	male	14.00	0	0	7538	9.2250	F	S	0
898	3	Connolly, Miss. Kate	female	30.00	0	0	330972	7.6292	C106	Q	1

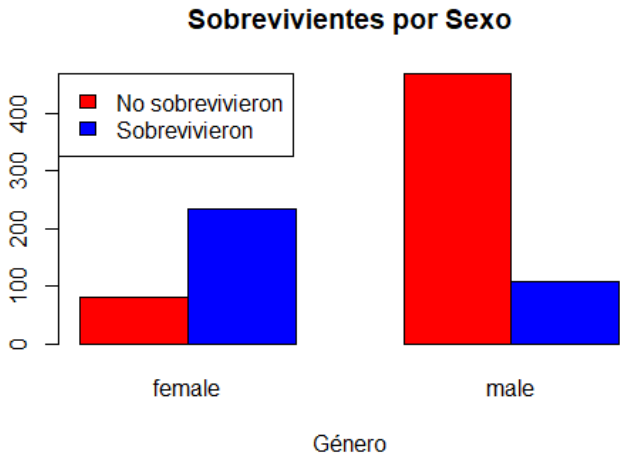
5. Representación de los resultados a partir de tablas y gráficas.

Representando los datos a partir de gráficos podemos ver la diferencia de pasajeros que sobrevivieron al naufragio en las distintas clases:

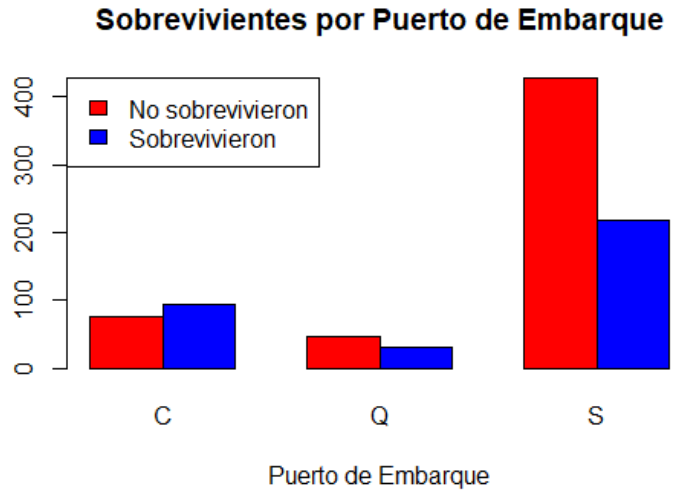


Como era de esperar, los de la clase más baja (tercera) sobrevivieron muchos menos pasajeros bien por la ubicación de las cabinas o porque los que tenían prioridad para salvarse eran los de las clases más altas (sobre todo niños y mujeres).

Si hacemos el mismo gráfico dividiendo por Sexo:

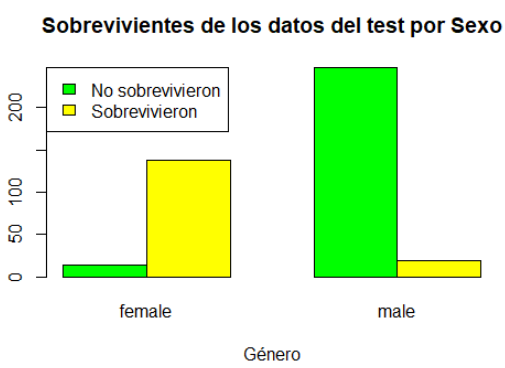


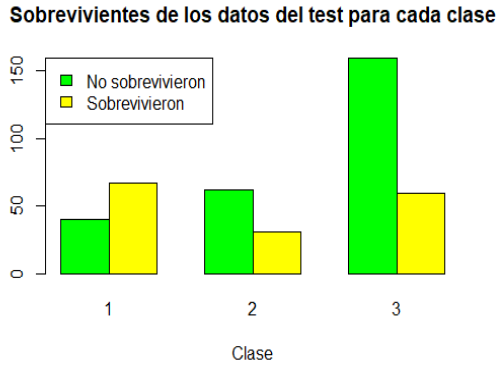
Corrobora lo que ya esperábamos y es que sobrevivieron muchas más mujeres que hombres.



Si comparamos los 3 puertos de embarque de los pasajeros, vemos que los que más murieron fueron los que embarcaron en el puerto S aunque también es lógico, ya que es el puerto donde más gente embarcó.

Vamos a representar estos gráficos para los datos del data_test, es decir, las predicciones de supervivencia realizadas según el modelo de regresión logística:





Parece que siguen el mismo patrón que los datos de partida lo cual indica que las predicciones parecen acertadas en la gran mayoría de los casos.

Así, siguen siendo los de la tercera clase los pasajeros que más mueren y los de la primera clase los que menos.

La diferencia entre los hombres y mujeres que sobreviven es todavía más notable en las predicciones.

6. Resolución del problema.

El objetivo del estudio era predecir la variable 'Survived', es decir, dada una muestra da pasajeros conseguir predecir si el pasajero sobrevivió al naufragio del Titanic o no. Como podemos ver en el apartado anterior, esto está resuelto.

Además, mediante un análisis de regresión logística vimos las variables que influyen significativamente en si un pasajero sobrevive o no. Llegamos a la conclusión así que las variables que más afectan a la hora de decidir si sobrevive son: la edad, el sexo, la clase del ticket y el número de hermanos/hermanas, hermanastros/hermanastros y marido o esposa del pasajero que también vayan a bordo.

7. Código en R.

Podemos acceder al código en R desde el siguiente enlace de GitHub:

https://github.com/marmartnz/Practica-2-Tipologia/tree/master/code

Además, tenemos los datos del dataset "datosTitanic" y los datos del test en el siguiente enlace. También están los csv's de salida con las predicciones:

https://github.com/marmartnz/Practica-2-Tipologia/tree/master/data