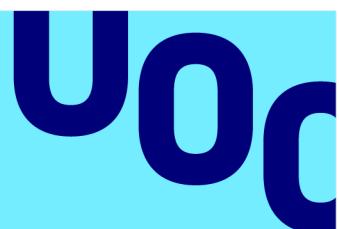


Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?



M2.851 - Tipología y ciclo de vida de los datos aula 2

ALUMNO: FRANCISCO JAVIER GHERSI LÁZARO

ALUMNO: JOSÉ MARÍA ARROYO SÁNCHEZ

PROFESOR: MARÍA ISABEL GUITART HORMIGO

PROFESOR: DANIEL ROMERO PEREZ

PRACTICA 2: limpieza y análisis

1 Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?	3
2 Integración y selección.	3
3 Limpieza de los datos.....	12
4 Análisis de los datos.	13
5 Conclusiones del estudio.	23
6 Aplicación Shiny.....	23

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

El conjunto de datos que hemos escogido para hacer el estudio es el siguiente: **DATOS DE ECOCARDIAGRAFÍA DE ESTRÉS**. El conjunto de datos stressEcho está disponible en la página web del Departamento de Bioestadística de la Universidad de Vanderbilt. En concreto, se han obtenido de la página web: <https://biostat.app.vumc.org/wiki/Main/DataSets>.

Hemos escogido este tema porque nos atrae mucho todo lo relacionado con los ecocardiogramas y queremos estudiar la variable **ecog** en función del resto de variables para poder extraer conclusiones, como se verá durante este trabajo.

El **ecocardiograma de estrés** es una prueba utilizada frecuentemente en cardiología que proporciona información en tiempo real acerca del comportamiento tanto del ventrículo izquierdo como de las válvulas en situación de estrés y permite compararlo con el estudio basal en reposo.

Existen diferentes formas de provocar el estrés. La elección de la técnica la sienta el médico que solicita la prueba. La forma más fisiológica es la realización de un ecocardiograma durante el esfuerzo físico que puede realizarse en una cinta rodante similar a la que se utiliza en los gimnasios o en una bicicleta estática. Cuando el paciente no puede realizar ningún tipo de esfuerzo por presentar una limitación física o por su edad se prefiere el ecocardiograma de estrés farmacológico, bien con dobutamina o bien con un vasodilatador (adenosina o dipiridamol).

2 Integración y selección.

Ahora procedemos a cargar el dataset seleccionado llamado **stressEcho**:

Podemos ver que tenemos un dataset con 32 variables y 558 observaciones por cada una de ellas. A continuación, realizamos una explicación de cada una de estas variables:

Variables numéricas:

- bhr: frecuencia cardíaca basal.
- basebp: presión arterial basal.
- basedp: producto doble basal (= bhr x basebp).
- pkhr: frecuencia cardíaca máxima.
- sbp: presión arterial sistólica.
- dp: producto doble (= pkhr x sbp).
- dose: dosis de dobutamina administrada.
- maxhr: frecuencia cardíaca máxima.
- pctMphr: % de frecuencia cardíaca máxima predicha alcanzada.
- mbp: presión arterial máxima.
- dpmaxdo: doble producto en dosis máxima de dobutamina.
- dobldose: dosis de dobutamina a la que se produjo el producto doble máximo.
- age: edad el paciente.
- baseEF: fracción de eyección cardíaca basal (una medida de la eficiencia de bombeo del corazón).
- dobEF: fracción de eyección de dobutamina.

Variables categóricas dicotómicas:

- gender: género ("male" o "female").
- chestpain: dolor torácico experimentado (0 = sí).
- restwma: anomalía del movimiento de la pared en reposo (0 = sí).
- posSE: ecocardiograma de estrés positivo (0 = sí).
- newMI: nuevo infarto de miocardio o ataque cardiaco (0 = sí).
- newPTCA: angioplastia reciente (0 = sí).
- newCABG: cirugía de derivación reciente (0 = sí).
- death: murió (0 = sí).
- hxofHT: historial de hipertensión (0 = sí).

- hxofDM: historial de diabetes (0 = sí).
- hxofMI: historial de ataque al corazón (0 = sí).
- hxofPTCA: historia de angioplastia (0 = sí).
- hxofCABG: historial de cirugía de bypass (0 = sí).
- any.event: death, newMI, newPTCA o newCABG (0 = sí).

Variables categóricas politómicas:

- hxofCig: historial de fumar ("non-smoker", "moderate" o "heavy").
- ecg: diagnóstico basal de ecocardiograma ("normal", "equivocal" o "MI").

Después de este resumen de las variables podemos ver que algunas de las variables son dicotómicas y politómicas, es decir, algunas están codificadas como texto en la base de datos, pero solo cogen 2 o 3 valores diferentes, y, otras aunque están codificadas como numéricas también cogen únicamente dos valores 0 o 1. Estas variables son las siguientes:

Variables dicotómicas codificadas como texto:

- gender.

Variables dicotómicas codificadas como numéricas:

- chestpain, restwma, posSE, newMI, newPTCA, newCABG, death, hxofHT, hxofDM, hxofMI, hxofPTCA, hxofCABG, any.event.

Variables politómicas:

- hxofCig, ecg.

A continuación utilizamos la función **skim()** para obtener un resumen de las principales estadísticas descriptivas del conjunto de datos. Esta función permite visualizar rápidamente la distribución de los datos, la presencia de valores ausentes, el tipo de datos y la cantidad de valores únicos en cada variable:

Variable type: character

skim_variablen_missingcomplete_rateminmaxemptyn_uniquewhitespace

gender	0	1	4	6	0	2	0	0
hxofCig	0	1	5	10	0	3	0	0
ecg	0	1	2	9	0	3	0	0

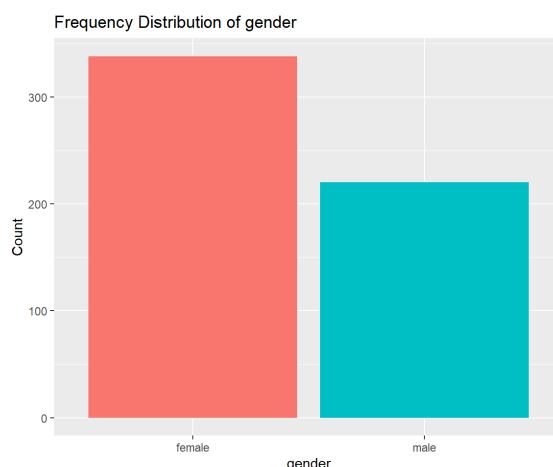
Variable type: numeric

	skim_variablen_missingcomplete_rate	mean	sd	p0	p25	p50	p75	p100	hist
...1	0	1	279.50	161.22	1	140.25	279.5	418.75	558
bhr	0	1	75.29	15.42	42	64.00	74.0	84.00	210
basebp	0	1	135.32	20.77	85	120.00	133.0	150.00	203
basedp	0	1	10181.312579.7550008400.00	9792.0	11663.2527300	11663.2527300	11663.2527300	11663.2527300	11663.2527300
pkhr	0	1	120.55	22.57	52	106.25	122.0	135.00	210
sbp	0	1	146.92	36.53	40	120.00	141.0	170.00	309
dp	0	1	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114	17633.845220.53510014033.0017060.020644.5045114
dose	0	1	33.75	8.13	10	30.00	40.0	40.00	40
maxhr	0	1	119.37	21.91	58	104.25	120.0	133.00	200
petMphr	0	1	78.57	15.12	38	69.00	78.0	88.00	133
mbp	0	1	156.00	31.71	84	133.25	150.0	175.75	309
dpmaxdo	0	1	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114	18549.884901.43713015260.0018118.021239.0045114
dobdose	0	1	30.24	9.54	5	20.00	30.0	40.00	40
age	0	1	67.34	12.05	26	60.00	69.0	75.00	93
baseEF	0	1	55.60	10.32	20	52.00	57.0	62.00	83
dobEF	0	1	65.24	11.76	23	62.00	67.0	73.00	94
chestpain	0	1	0.31	0.46	0	0.00	0.0	1.00	1
restwma	0	1	0.46	0.50	0	0.00	0.0	1.00	1
posSE	0	1	0.24	0.43	0	0.00	0.0	0.00	1
newMI	0	1	0.05	0.22	0	0.00	0.0	0.00	1
newPTCA	0	1	0.05	0.21	0	0.00	0.0	0.00	1
newCABG	0	1	0.06	0.24	0	0.00	0.0	0.00	1
death	0	1	0.04	0.20	0	0.00	0.0	0.00	1
hxofHT	0	1	0.70	0.46	0	0.00	1.0	1.00	1
hxofDM	0	1	0.37	0.48	0	0.00	0.0	1.00	1

	skim_variablen_missingcomplete_rate	mean	sd	p0	p25	p50	p75	p100	hist
hxofMI	0	1	0.28	0.45	0	0.00	0.0	1.00	1
hxofPTCA	0	1	0.07	0.26	0	0.00	0.0	0.00	1
hxofCABG	0	1	0.16	0.36	0	0.00	0.0	0.00	1
any.event	0	1	0.16	0.37	0	0.00	0.0	0.00	1

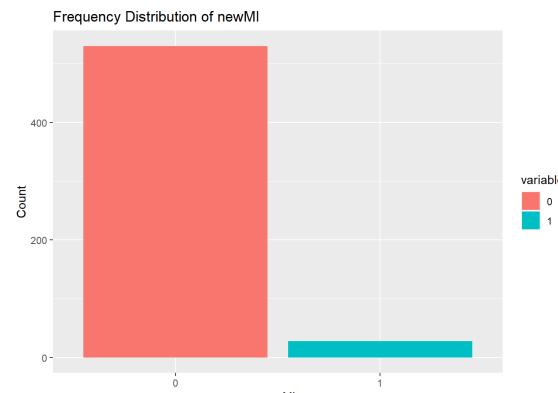
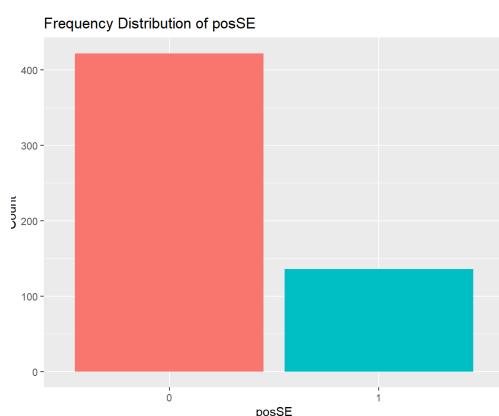
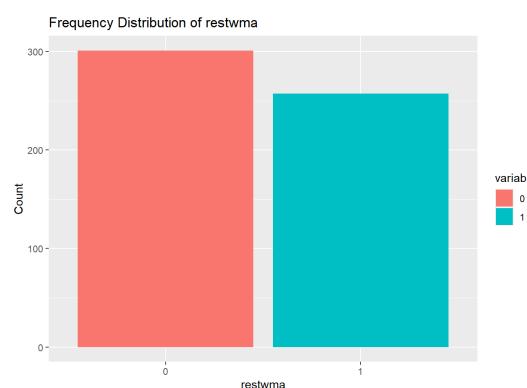
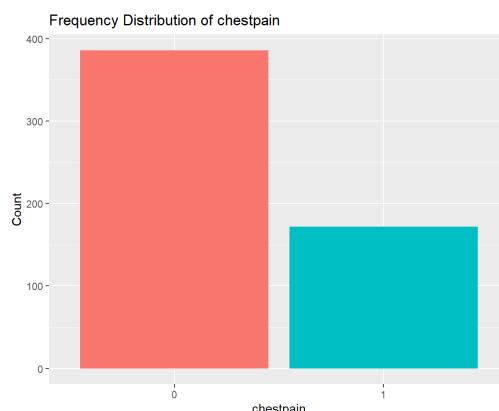
2.1 Variables dicotómicas codificadas como texto.

A continuación nos vamos a centrar en estudiar las **variables dicotómicas** codificadas como texto:

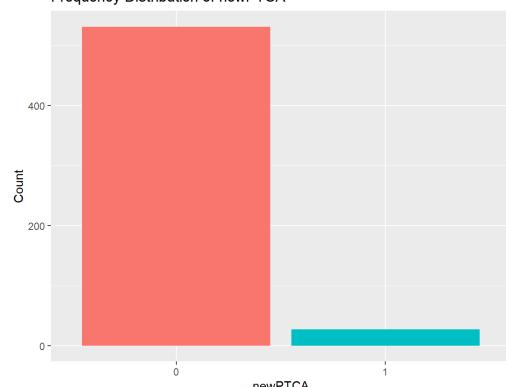


2.2 Variables dicotómicas codificadas como numéricas.

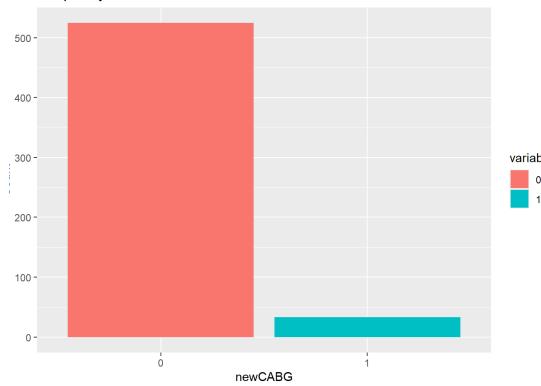
A continuación nos vamos a centrar en estudiar las **variables dicotómicas** codificadas como numéricas:



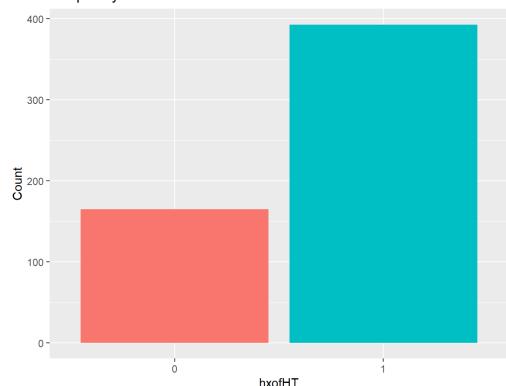
Frequency Distribution of newPTCA



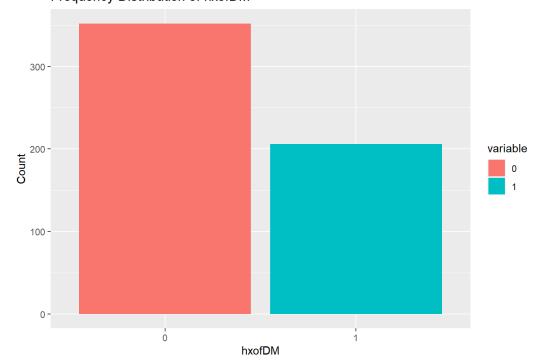
Frequency Distribution of newCABG



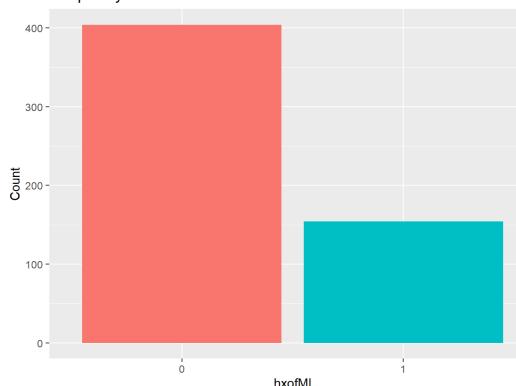
Frequency Distribution of hxofHT



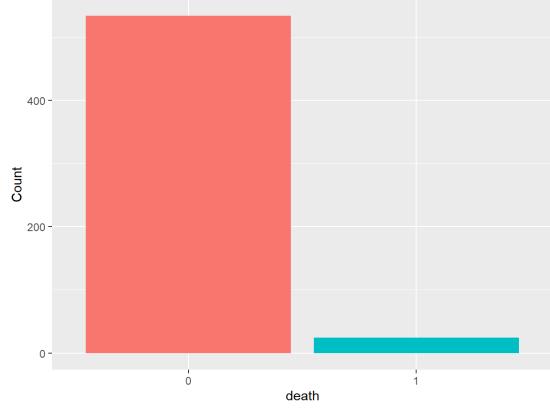
Frequency Distribution of hxofDM



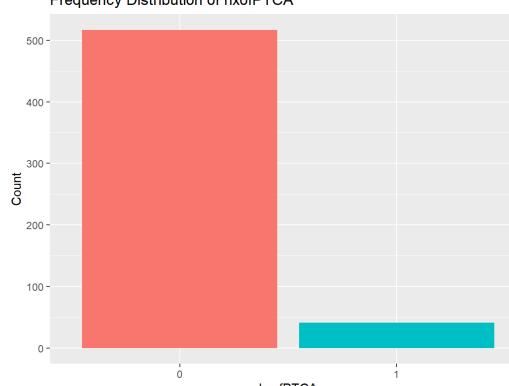
Frequency Distribution of hxofMI



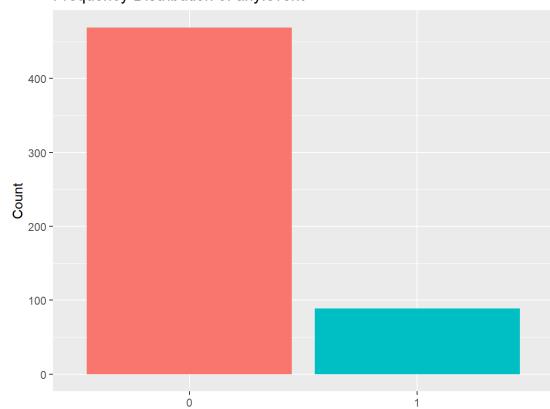
Frequency Distribution of death



Frequency Distribution of hxofPTCA

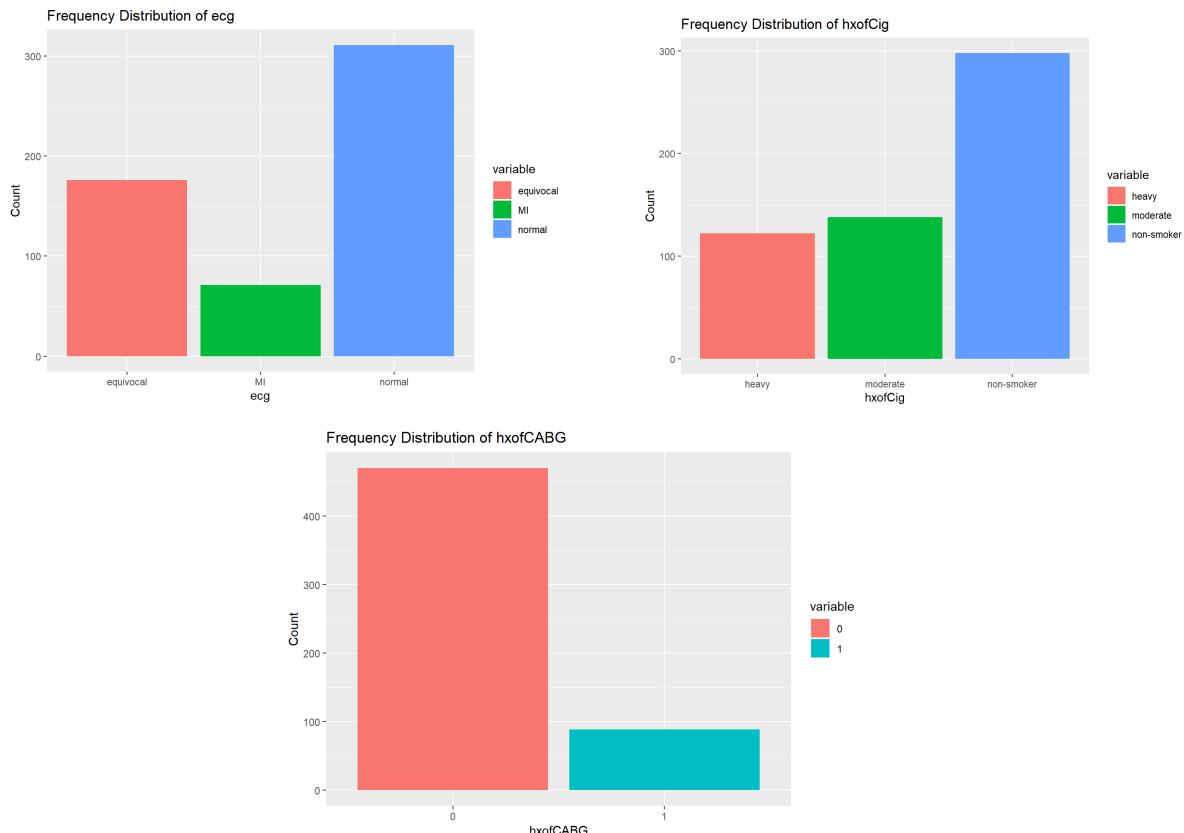


Frequency Distribution of any.event



2.3 Variables politómicas codificadas como texto.

A continuación nos vamos a centrar en estudiar las **variables politómicas** codificadas como texto:



2.4 Variables numéricas.

A continuación, nos vamos a centrar en estudiar las **variables numéricas**. En primer lugar, miraremos su estructura y buscaremos los estadísticos más importantes:

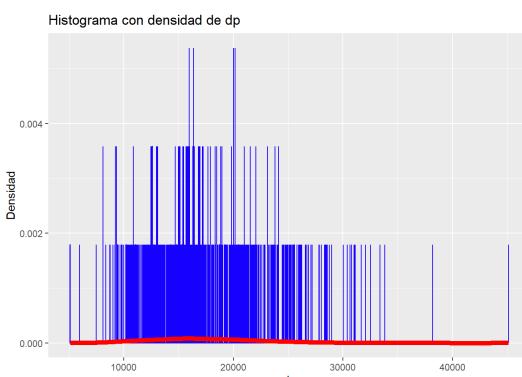
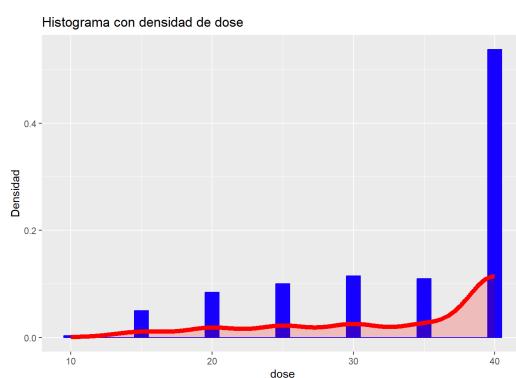
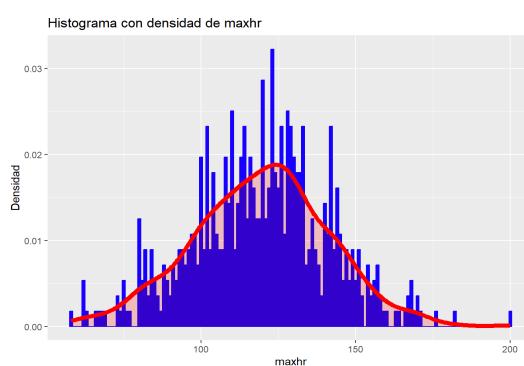
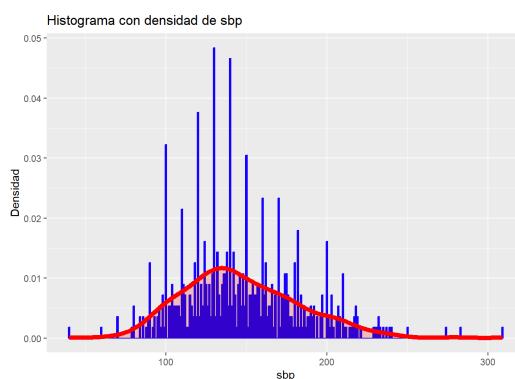
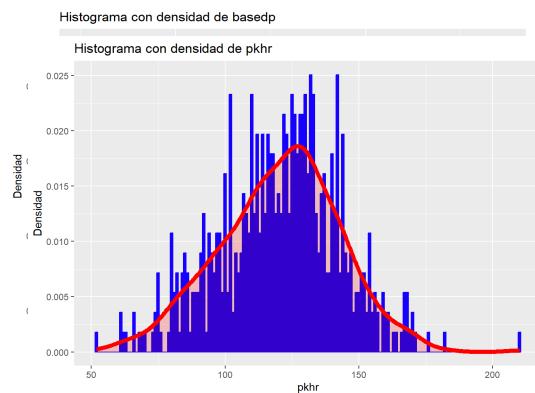
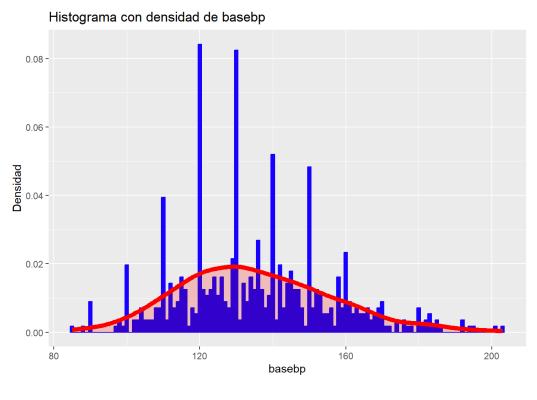
```

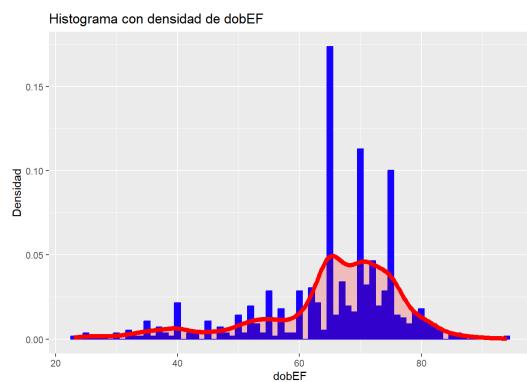
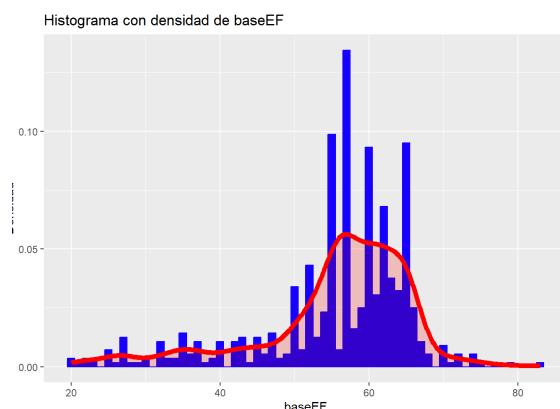
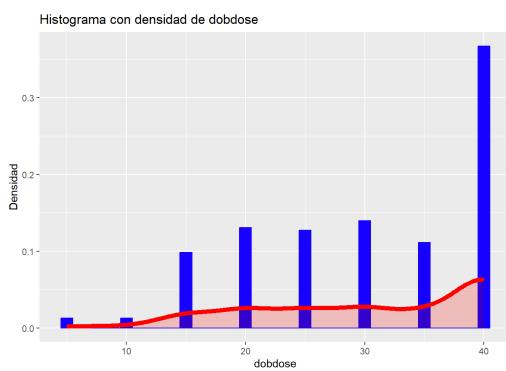
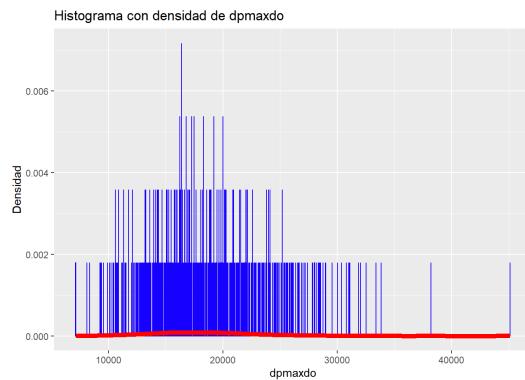
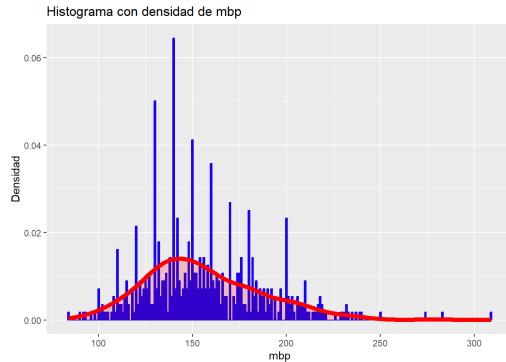
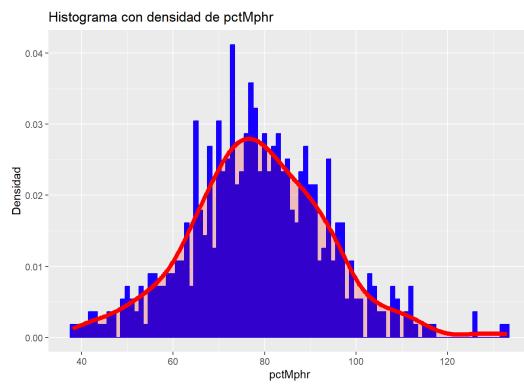
##      bhr        basebp      basedp      pkhr
##  Min.   : 42.00  Min.   : 85.0  Min.   : 5000  Min.   : 52.0
##  1st Qu.: 64.00  1st Qu.:120.0  1st Qu.: 8400  1st Qu.:106.2
##  Median : 74.00  Median :133.0  Median : 9792  Median :122.0
##  Mean   : 75.29  Mean   :135.3  Mean   :10181  Mean   :120.6
##  3rd Qu.: 84.00  3rd Qu.:150.0  3rd Qu.:11663  3rd Qu.:135.0
##  Max.   :210.00  Max.   :203.0  Max.   :27300  Max.   :210.0
##
##      sbp         dp        dose      maxhr
##  Min.   : 40.0  Min.   : 5100  Min.   :10.00  Min.   : 58.0
##  1st Qu.:120.0  1st Qu.:14033  1st Qu.:30.00  1st Qu.:104.2
##  Median :141.0  Median :17060  Median :40.00  Median :120.0
##  Mean   :146.9  Mean   :17634  Mean   :33.75  Mean   :119.4
##  3rd Qu.:170.0  3rd Qu.:20645  3rd Qu.:40.00  3rd Qu.:133.0
##  Max.   :309.0  Max.   :45114  Max.   :40.00  Max.   :200.0
##
##      pctMphr      mbp      dpmaxdo      dobdoze
##  Min.   : 38.00  Min.   : 84.0  Min.   : 7130  Min.   : 5.00
##  1st Qu.: 69.00  1st Qu.:133.2  1st Qu.:15260  1st Qu.:20.00
##  Median : 78.00  Median :150.0  Median :18118  Median :30.00
##  Mean   : 78.57  Mean   :156.0  Mean   :18550  Mean   :30.24
##  3rd Qu.: 88.00  3rd Qu.:175.8  3rd Qu.:21239  3rd Qu.:40.00

```

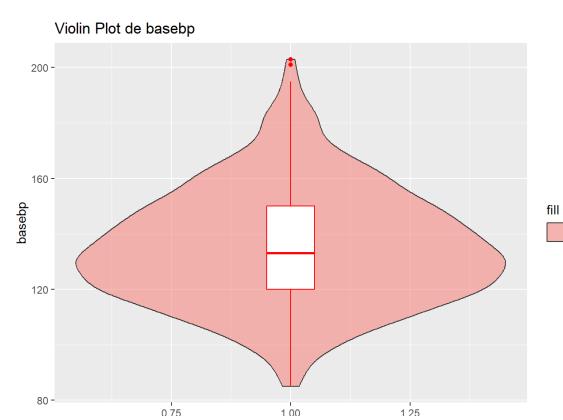
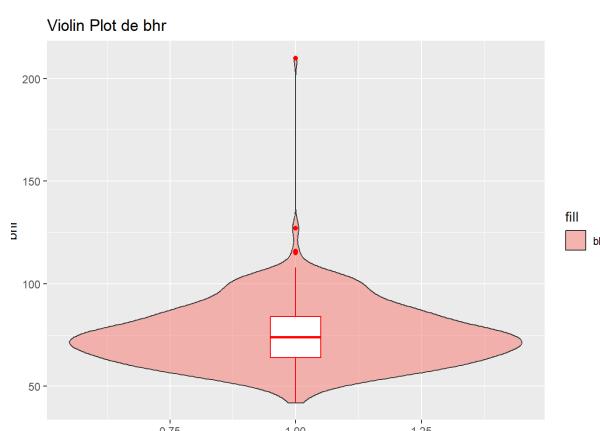
```
## Max.    :133.00   Max.    :309.0   Max.    :45114   Max.    :40.00
## age      baseEF    dobef
## Min.    :26.00   Min.    :20.0   Min.    :23.00
## 1st Qu.:60.00   1st Qu.:52.0   1st Qu.:62.00
## Median  :69.00   Median  :57.0   Median  :67.00
## Mean    :67.34   Mean    :55.6   Mean    :65.24
## 3rd Qu.:75.00   3rd Qu.:62.0   3rd Qu.:73.00
## Max.    :93.00   Max.    :83.0   Max.    :94.00
```

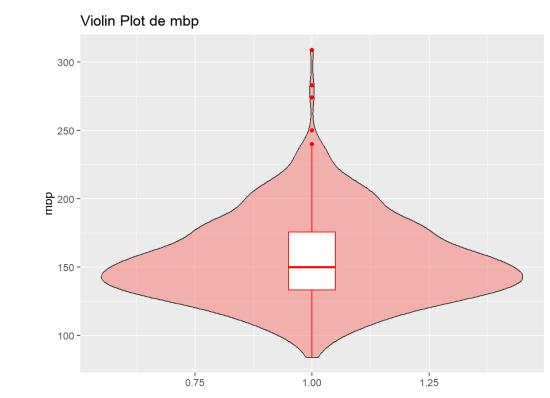
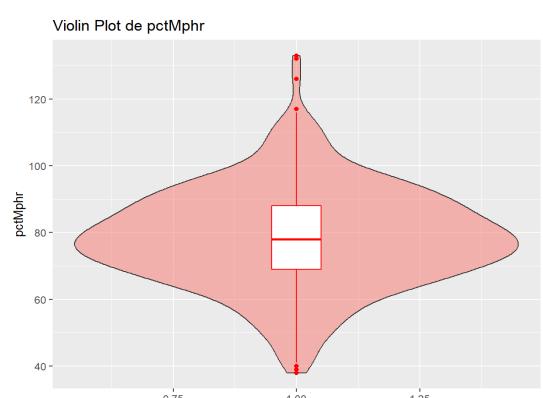
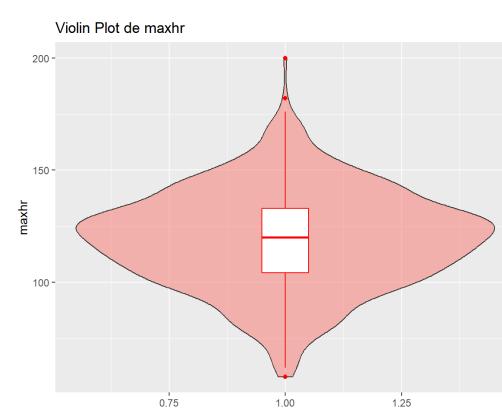
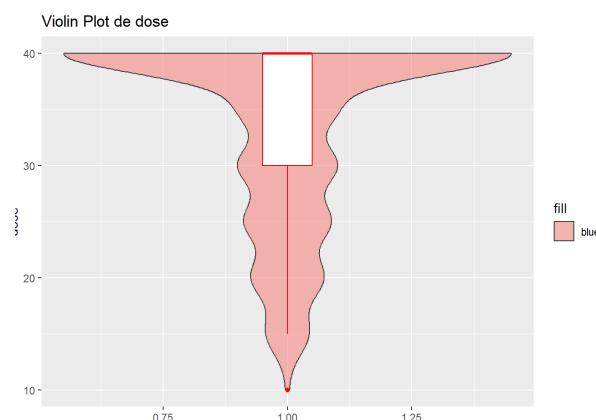
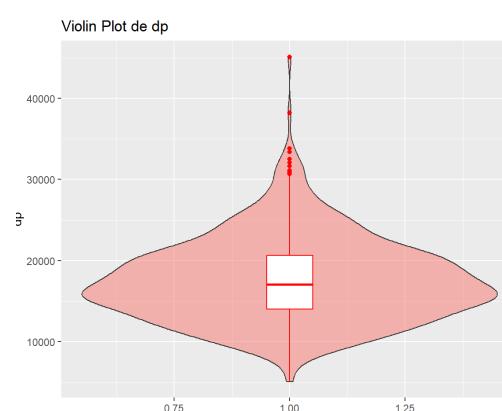
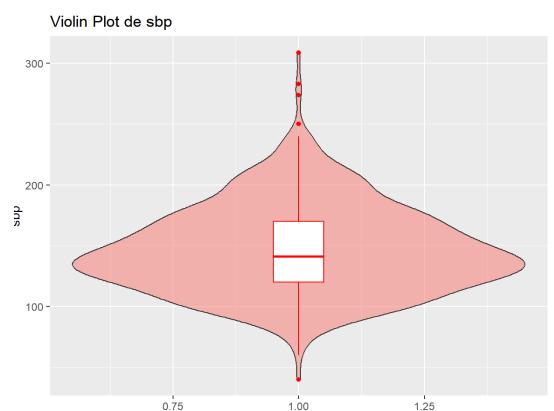
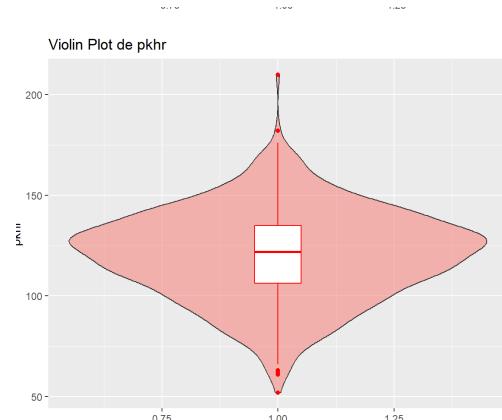
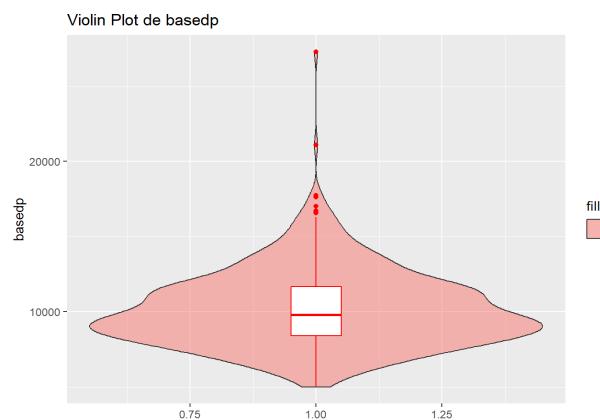
Para hacernos una idea visual del comportamiento de estas variables numéricas haremos el **histograma de densidad** para cada una de ellas:

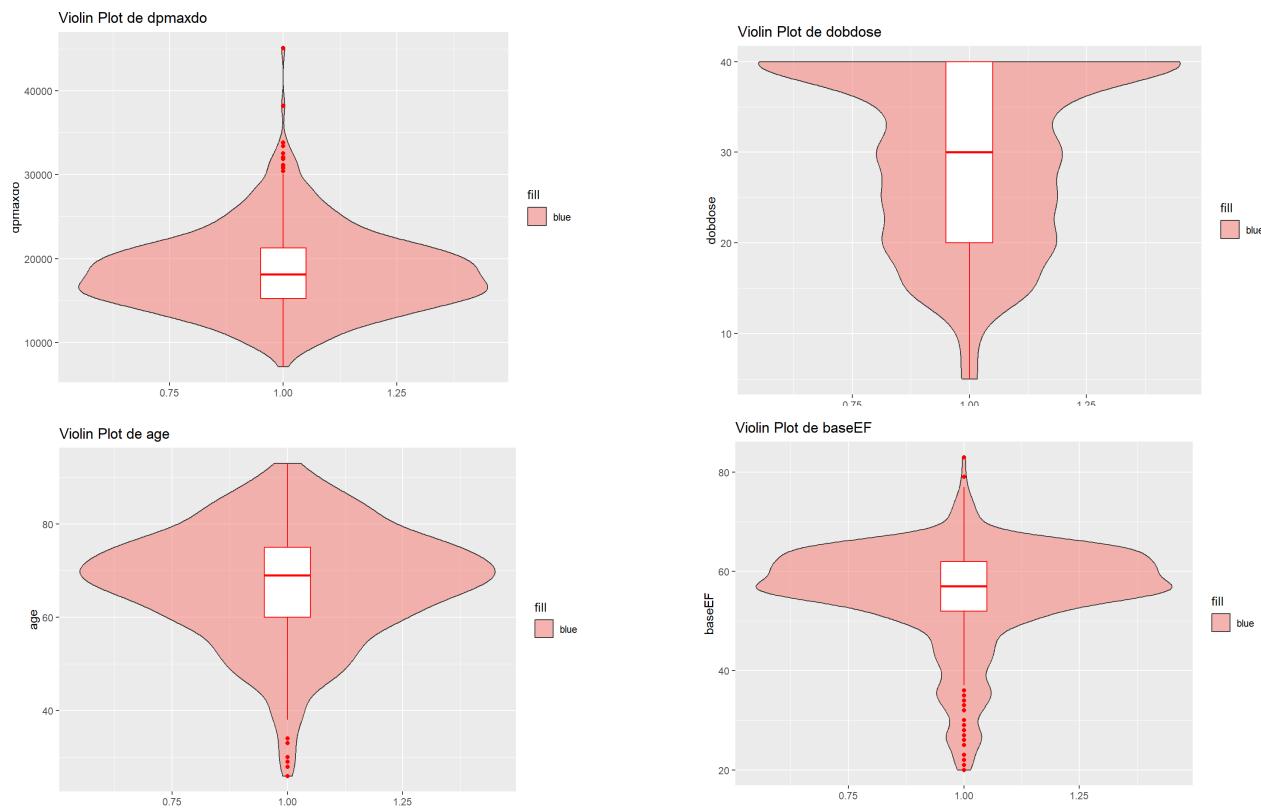




Para hacernos una idea visual de la distribución de estas variables numéricas haremos un **boxplot-violín** para cada una de ellas, **nótese que representamos los valores atípicos**:





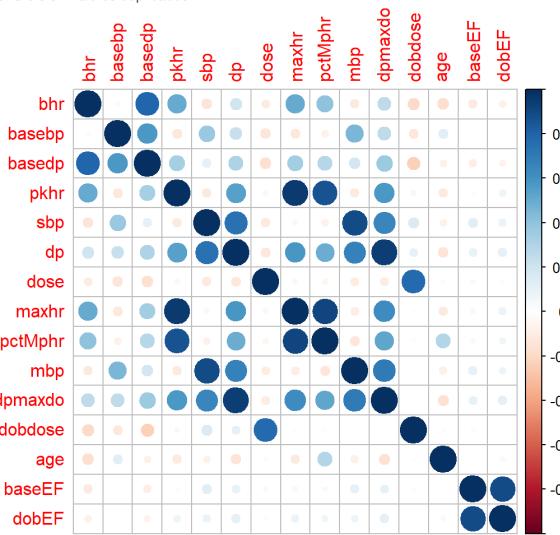


Por último, vamos aver la **matriz de correlación** entre estas variables numéricas:

Matriz de correlación

	bhr	basebp	basedp	pkhr	sbp	dp	dose	maxhr	pctMphr	mbp	dpmaxdo	dobdose	age	baseEF	dobEF
bhr	1.000000-0.0228474	0.7946821	0.5069987-0.1444254	0.1993087-0.1052755	0.5044911	0.4074955-0.1160365	0.2529296-0.1973112-0.1775550-0.1116504-0.00797389								
basebp	-0.0228474	1.000000	0.5783270-0.1260488	0.3722908	0.2208140-0.1387677	-0.1238062-0.0709128	0.4544641	0.2525293-0.1238648	0.1284350	0.0048881-0.0011008					
basedp	0.7946821	0.5783270	1.0000000	0.3393872	0.3018262-0.1691731	0.3402048	0.2886518	0.1809016	0.3663922-0.2375921-0.0778380-0.0901115-0.0660357						
pkhr	0.5069987-0.1260488	0.3393872	0.1000000	-0.1173338	0.5485216-0.0434947	0.9544381	0.8662407-0.1171976	0.5771125-0.0409584-0.1151532	0.0238051	0.0718384					
sbp	-0.1444254	0.3722908	0.1068736-0.1173338	1.0000000	0.7475026-0.1167759-0.0456711-0.0765864	0.8982150	0.6567105	0.1559768-0.0726594	0.1210188	0.0854158					
dp	0.1993087	0.2208140	0.3018262	0.5485216	0.7475026	1.0000000-0.1220713	0.5834510	0.4942528	0.6730694	0.9440458	0.1020321-0.1491650	0.0996224	0.1088093		
dose	-0.1052755-0.1387677	-0.1691731-0.0434947	-0.1167759-0.1220713	1.0000000	-0.0431522	-0.0526006-0.1029160-0.1032096	0.7763776	-0.0264825-0.0389213	0.0133808						
maxhr	0.5044911-0.1238062	0.3402048	0.9544381-0.0456711	0.5834510	-0.0431522	1.0000000	0.9109132-0.0959646	0.6226925	0.0195113-0.1142203	0.0411094	0.0834485				
pctMphr	0.4074955-0.0709128	0.2886518	0.8662407-0.0765864	0.4942528	-0.0526006	0.9109132	1.0000000-0.1322988	0.5257254	0.0185913	0.2976045	0.0377297	0.0626851			
mbp	-0.1160364	0.4544641	0.1809016-0.1171976	0.8982150	-0.6730694-0.1029160-0.0959646-0.1322988	1.0000000	0.7051445-0.0013065-0.0829208	0.1098756	0.0775935						
dpmaxdo	0.2529296	0.2525293	0.3663922	0.5771125	0.6567105	0.9440458-0.1032096	0.6226925	0.5257254	0.7051445	1.0000000	0.0082080-0.1594466	0.1024968	0.1152387		
dobdose	-0.1973112-0.1238648	-0.2375921-0.0409584	0.1559768	0.1020321	0.7763776	0.0195113	0.0185913-0.0013065	0.0082080	0.1000000	-0.0123673-0.0204589	0.0046796				
age	-0.1775550	0.1284350	-0.0778380	-0.1151532	-0.1171976	0.1210188	0.0969213	0.0411094	0.0377297	0.1098756	0.0204589	0.0036384	1.0000000	0.8997335	
baseEF	-0.1116504	0.0048881	0.0448881	0.0511115	0.0238051	0.1210188	0.0969213	0.0411094	0.0377297	0.1098756	0.0204589	0.0036384	1.0000000	0.8997335	
dobEF	-0.0797389	-0.0011008-0.0660357	0.0718384	0.0854158	0.05133808	0.0133808	0.0834485	0.0626851	0.0775935	0.1152387	0.0046796-0.0321174	0.8997335	1.0000000		

Al ser esta tabla difícil de interpretar vamos a crear el **gráfico de correlación** y también vamos a agrupar las variables con un coeficiente de correlación igual o superior a 0.5 sin valores duplicados:



3 Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos?

Para comprobar valores ausentes en todo el conjunto de datos stressEcho, excluyendo los valores 0 en variables dicotómicas numéricas, primero identificaremos cuáles son las columnas con variables dicotómicas. Luego, verificaremos la presencia de valores ausentes en las columnas no dicotómicas y, para las dicotómicas, solo consideraremos NA como valor ausente. Como la **variable 1 corresponde al id del paciente** la eliminaremos de nuestro conjunto de datos:

Realizado esto, podemos afirmar que **no existen valores ausentes**.

3.2 Identifica y gestiona los valores extremos.

Para identificar y gestionar los **valores extremos** en un conjunto de datos, una técnica común es utilizar el **método de los cuartiles**, también conocido como el método IQR (rango intercuartil). Esto implica identificar valores que son significativamente más bajos que el primer cuartil o significativamente más altos que el tercer cuartil.

Variable Valores Extremos

	Valores Extremos
bhr	116, 115, 210, 127, 127
basebp	201, 203
basedp	16704, 17710, 17748, 17000, 16562, 17604, 16600, 27300, 16686, 21082
pkhr	182, 62, 61, 210, 61, 63, 52
sbp	309, 283, 40, 274, 250
dp	31122, 31096, 45114, 38205, 33840, 30810, 31671, 31020, 32058, 33400, 30660, 32518
dose	10, 10
maxhr	182, 200, 58
pctMphr	126, 126, 133, 117, 39, 132, 40, 38
mbp	309, 283, 240, 274, 250
dpmaxdo	31122, 31096, 45114, 38205, 33840, 30810, 31866, 31020, 32058, 33400, 31080, 32518, 30400
age	34, 29, 26, 28, 29, 30, 34, 33, 34
baseEF	27, 32, 20, 34, 22, 23, 32, 36, 30, 34, 35, 35, 35, 36, 20, 21, 25, 35, 79, 27, 32, 35, 25, 29, 30, 33, 22, 23, 27, 46
dobEF	32, 35, 36, 32, 83, 35, 33, 27, 28, 32, 35, 25, 25, 27, 26, 27, 27
dobdose	32, 40, 40, 32, 37, 40, 27, 34, 37, 43, 35, 35, 43, 25, 40, 35, 40, 40, 45, 32, 45, 40, 36, 45, 94, 33, 30, 28, 35, 51
dp	40, 40, 39, 45, 42, 37, 90, 30, 35, 37, 38, 45, 45, 40, 25, 42, 40, 40, 35, 26, 38, 23

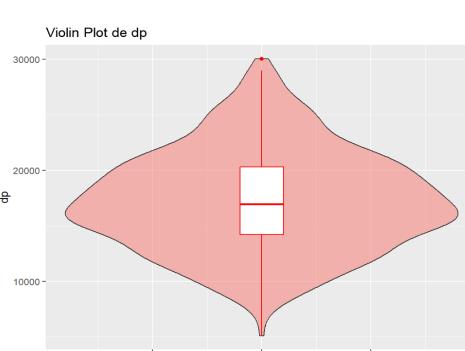
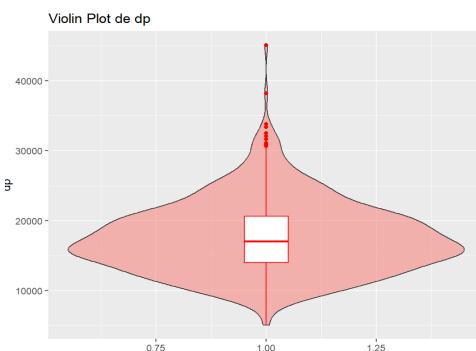
Número de Valores Extremos	Rango de Valores No Extremos	Extremos
5	42 - 108	
2	85 - 195	
10	5000 - 16280	
7	66 - 176	
5	60 - 240	
12	5100 - 30400	
2	15 - 40	
3	62 - 176	
8	41 - 116	
5	84 - 239	
13	7130 - 30044	
9	38 - 93	
46	37 - 77	
47	47 - 89	

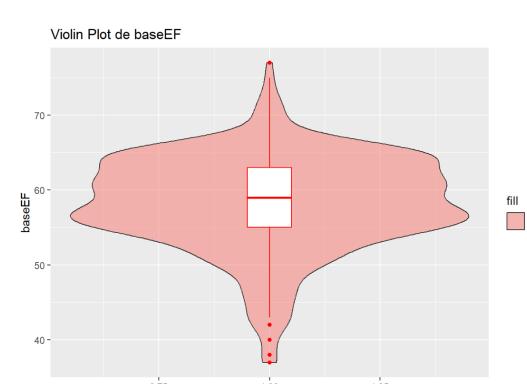
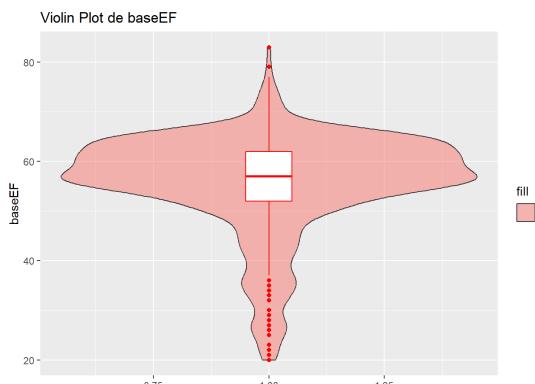
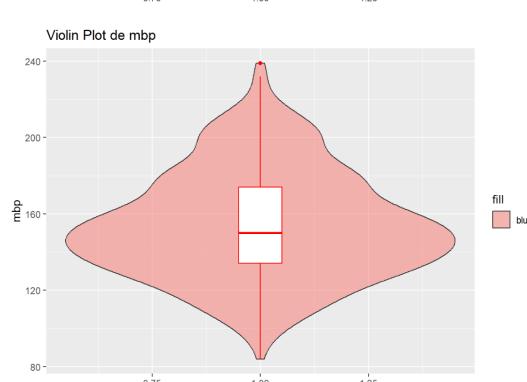
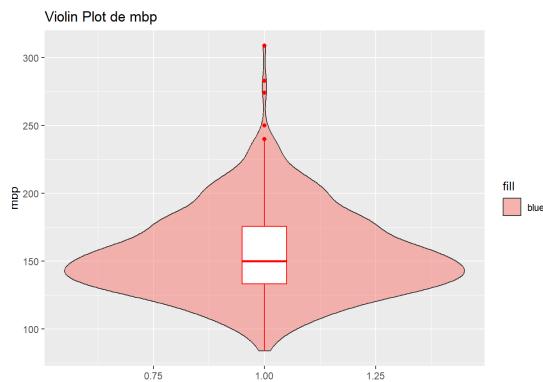
Hemos podido observar que casi todas las variables presentan valores atípicos. En este sentido, optaremos por eliminar todas estas observaciones extremas (excepto las de la **variable age**) ya que nuestro conjunto de datos es muy grande. Esto lo vamos a realizar con la certeza de que no van a influir en los resultados finales, pues consideramos que hay información de sobra.

Procedemos a **eliminar los valores extremos**:

```
## Se han eliminado 101 filas del conjunto de datos original.
```

Ahora volvemos a realizar la distribución de las variables mediante un **boxplot-violín** después de eliminar los valores extremos y podemos visualizar un cambio muy sustancial en casi todas ellas, menos en las variables **dose y dobdose**. A modo de muestra presentamos una selección de parámetros, todas las gráficas se encuentran en los recursos del repositorio de Github:





4 Análisis de los datos.

4.1 Comprobación de la normalidad y homogeneidad de la varianza de las variables.

Para comprobar la **normalidad de los datos** del conjunto de datos **stressEcho_no_outliers**, se pueden utilizar pruebas estadísticas como la **prueba de normalidad de Shapiro-Wilk** o la **prueba de Kolmogorov-Smirnov**. Para la **homogeneidad de la varianza**, se pueden utilizar pruebas estadísticas como la **prueba de Levene** o la **prueba de Bartlett**.

Comprobación de la normalidad de los datos:

El objetivo de este análisis es evaluar la normalidad de las variables numéricas sin valores ausentes y no categóricas del conjunto de datos **stressEcho_no_outliers**. Para lograr esto, se utiliza la **prueba de Shapiro-Wilk**:

```
## [1] "maxhr"    "pctMphr"
## Las siguientes variables cumplen con el requisito de normalidad: maxhr, pctMphr
```

Comprobación de la homogeneidad de varianza de los datos:

El objetivo de este análisis es evaluar la homogeneidad de la varianza de las variables numéricas sin valores ausentes y no categóricas del conjunto de datos **stressEcho_no_outliers**. Para lograr esto, se utiliza la **prueba de Bartlett**:

```
## [1] "bhr"       "basebp"    "basedp"    "pkhr"      "sbp"       "dp"        "dose"
## [8] "maxhr"     "pctMphr"   "mbp"       "dpmaxdo"   "age"       "dobEF"
## Las siguientes variables cumplen con el requisito de homogeneidad de la varianza: bhr, basebp, basedp, pkhr, sbp, dp, dose, maxhr, pctMphr, mbp, dpmaxdo, age, dobEF
```

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

Ahora queremos analizar la **variable ecg** en función de diferentes grupos de variables. Para llevar acabo esto hemos pensado en los siguientes grupos:

- 1.- **gender**: comparar la variable ECG entre hombres y mujeres.
- 2.- **chestpain**: comparar la variable ECG solo entre pacientes con y sin dolor en el pecho.
- 3.- **maxhr**: comparar la variable ECG solo entre pacientes con frecuencia cardíaca máxima.

4.3 Aplicación de pruebas estadísticas.

4.3.1 Variable gender.

Para hacer un estudio entre las **variables gender y ecg**, vamos a utilizar una **prueba de chi-cuadrado** para evaluar si hay una asociación significativa entre ambas variables.

La **prueba de chi-cuadrado** es una prueba estadística que se utiliza para evaluar la asociación entre dos variables categóricas. En este caso, la variable gender es categórica y la variable ecg es categórica también.

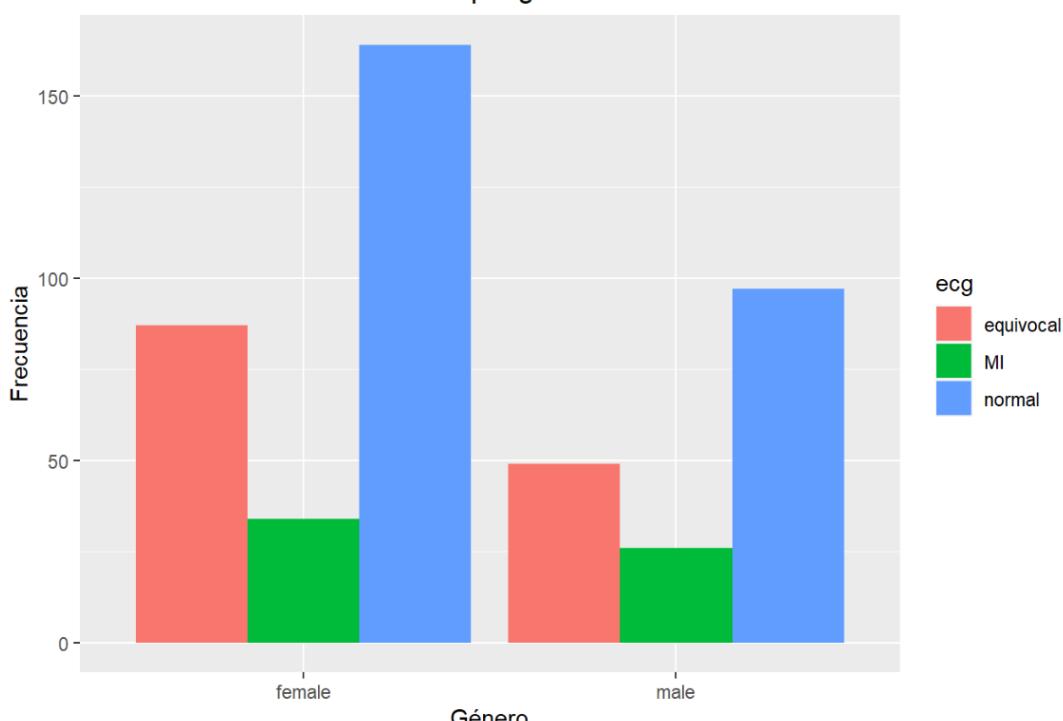
Antes de aplicar la prueba de chi-cuadrado, nos tenemos que asegurar de que se cumplen los supuestos:

- 1.- Las observaciones sean independientes.
- 2.- Las celdas de la tabla de contingencia tengan un tamaño mínimo de 5.
- 3.- No haya valores ausentes en los datos.

Una vez que se cumplen estos supuestos, podremos aplicar la prueba de chi-cuadrado y evaluar si hay una asociación significativa entre las variables. Si el valor de **p** es menor que el **nivel de significancia establecido** (por ejemplo, 0.05), entonces se puede concluir que hay una asociación significativa entre las variables.

En nuestro caso, ya sabemos que se cumplen los supuestos por los estudios realizados anteriormente, con lo cual procedemos a realizar la **prueba chi-cuadrado**:

Distribución de la variable ECG por género

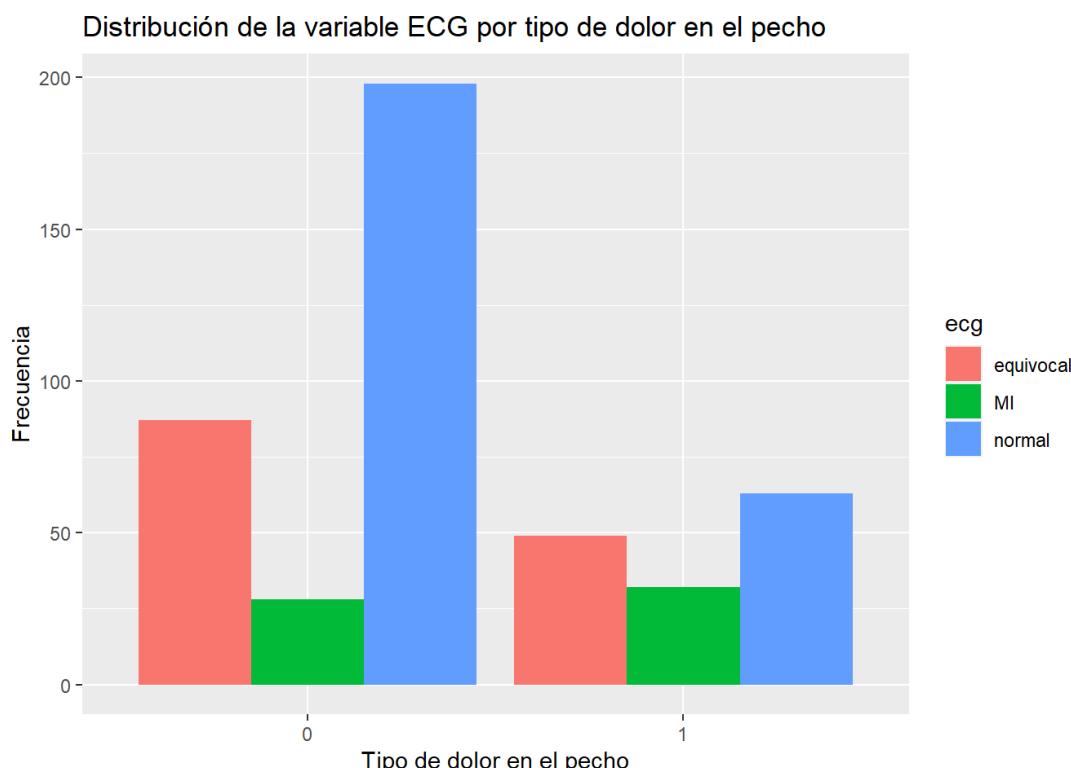


```
## Pearson's Chi-squared test
## data: contingency_table
## X-squared = 1.004, df = 2, p-value = 0.6053
```

Según los resultados obtenidos, la estadística de prueba es 1.004, los grados de libertad son 2 y el valor p es 0.6053. Con un **valor p mayor a 0.05**, no hay suficiente evidencia para rechazar la hipótesis nula. Esto significa que **no hay una asociación significativa entre las variables gender y ecg**.

4.3.2 Variable chestpain.

Para hacer un estudio entre las **variables chestpain y ecg** también vamos a utilizar una **prueba de chi-cuadrado** para evaluar si hay una asociación significativa entre ambas variables:



```
## Pearson's Chi-squared test
## data: contingency_table
## X-squared = 21.101, df = 2, p-value = 2.618e-05
```

Según los resultados obtenidos, la estadística de prueba es 21.101, los grados de libertad son 2 y el valor p es **2.618e-05 (muy cercano a cero)**. Con un valor p tan bajo, podemos concluir que hay suficiente evidencia para rechazar la hipótesis nula y afirmar que **hay una asociación significativa entre las variables chestpain y ecg**.

4.3.3 Variable maxhr.

Para hacer un estudio entre las **variables maxhr y ecg** vamos a utilizar la **prueba de análisis de varianza (ANOVA) de un factor**, ya que la variable ecg tiene 3 niveles. Si tuviese menos niveles podríamos haber utilizado directamente la **prueba t de Student para muestras independientes**.

El **ANOVA (Análisis de Varianza) de un factor** es una técnica estadística que se utiliza para evaluar si hay diferencias significativas en la media de una variable continua entre dos o más grupos definidos por una variable categórica. En el **ANOVA de un factor**, se compara la variación entre los grupos con la variación dentro de los grupos para determinar si la variación entre los grupos es significativamente mayor que la variación dentro de los grupos.

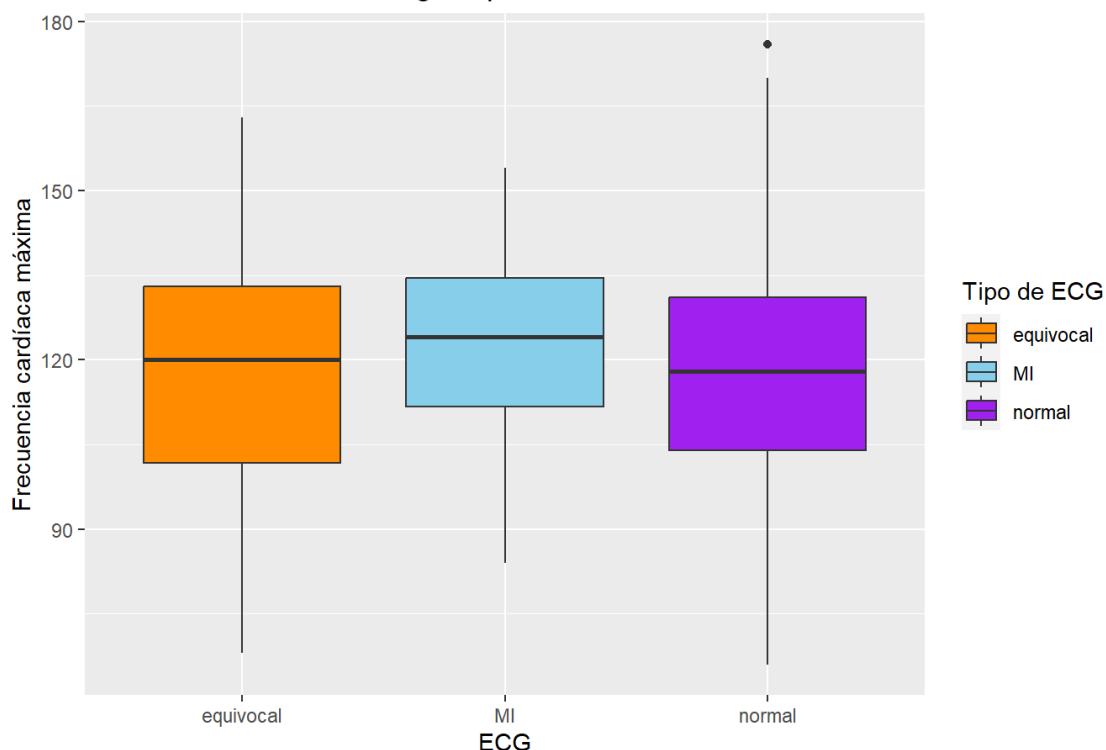
La **hipótesis nula del ANOVA de un factor** es que no hay diferencias significativas entre las medias de los grupos, mientras que la **hipótesis alternativa** es que al menos una de las medias de los grupos es diferente. Si se rechaza la hipótesis nula, se puede concluir que hay al menos un grupo cuya media es significativamente diferente a la media de los otros grupos.

Para realizar un **ANOVA de un factor**, se deben cumplir los supuestos:

- 1.- Las variables tengan una distribución normal.
- 2.- La varianza sea igual en todos los grupos.

En nuestro caso, ya sabemos que se cumplen los supuestos por los estudios realizados anteriormente, con lo cual procedemos a realizar la **prueba de análisis de varianza (ANOVA) de un factor**:

Frec. cardíaca máxima según tipo ECG



```
##          Df Sum Sq Mean Sq F value Pr(>F)
## ecg       2   1307   653.6   1.582  0.207
## Residuals 454 187534   413.1
```

La **tabla ANOVA** muestra que hay dos grados de libertad para `ecg` y 454 grados de libertad para `Residuals`. La suma de cuadrados para `ecg` es 1307 y la suma de cuadrados para `Residuals` es 187534. La media de cuadrados para `ecg` es 653.6 y la media de cuadrados para `Residuals` es 413.1.

El valor F calculado para `ecg` es 1.582, y el **valor p asociado con la prueba F es 0.207**. El valor p indica que no hay suficiente evidencia para rechazar la hipótesis nula de que **no hay efecto significativo de `ecg` en la variable respuesta**.

4.4 Análisis de regresión.

En este apartado del estudio queremos predecir la **variable categórica politómica ecg** en función de todas las variables numéricas. Para poder hacerlo tenemos que realizar un **modelo de regresión logística multinomial**.

Un **modelo de regresión logística multinomial** es una técnica estadística utilizada para predecir la probabilidad de que una observación pertenezca a cada una de las categorías de una variable categórica con tres o más niveles. En otras palabras, **es un modelo de clasificación multiclas** que se utiliza para predecir la probabilidad de que una observación pertenezca a cada una de las categorías de la variable dependiente.

El modelo se basa en la **función softmax**, que es una generalización de la función logística utilizada en el modelo de regresión logística binomial. La **función softmax** transforma la suma ponderada de las variables predictoras en una distribución de probabilidad sobre las categorías de la variable dependiente. Cada categoría tiene su propia ecuación de regresión logística, que se utiliza para modelar la probabilidad de que la observación pertenezca a esa categoría.

Para **ajustar el modelo de regresión logística multinomial**, primero debemos dividir los datos en conjuntos de entrenamiento y prueba:

Luego, podemos ajustar un modelo de regresión logística multinomial utilizando todas las variables predictoras:

```
## # weights:  99 (64 variable)
## initial value 352.654545
## iter  10 value 311.312120
## iter  20 value 302.085434
## iter  30 value 277.071559
## iter  40 value 255.929174
## iter  50 value 251.717147
## iter  60 value 250.576344
## iter  70 value 250.501763
## final value 250.493222
## converged
```

Podemos obtener un resumen del modelo utilizando la función `summary()`:

```
## Call:
## multinom(formula = ecg ~ bhr + basebp + basedp + pkhr + sbp +
##           dp + dose + maxhr + pctMphr + mbp + dpmaxdo + dobdose + age +
##           baseEF + dobEF + gender + chestpain + restwma + posSE + newMI +
##           newPTCA + newCABG + death + hxofHT + hxofDM + hxofMI + hxofPTCA +
##           hxofCABG + any.event + hxofCig, data = train)
##
## Coefficients:
##             (Intercept)      bhr     basebp     basedp      pkhr
## MI      -1.0738813  0.09095364  0.07194113 -0.0007076526 -0.03391809
## normal -0.9358411 -0.04440382 -0.01654479  0.0003091083 -0.04887421
##             sbp       dp      dose     maxhr    pctMphr
## MI      -0.04737363  0.0001808833 -0.04351670 -0.08397724  0.057554314
## normal -0.06771093  0.0005572269 -0.01586921  0.09875918 -0.001537847
##             mbp     dpmaxdo   dobdose      age     baseEF     dobEF
## MI      -0.04305116  0.0005712649  0.03113900 -0.008583352  0.09138636 -0.08436750
## normal  0.10493777 -0.0008784037  0.01957791 -0.035288634  0.02700928 -0.01194826
##             gendermale chestpain1   restwmal     posSE1     newMI1  newPTCA1
## MI      0.4826911  0.7124413 -0.003058847  0.09588896 -27.5131591 -26.836872
## normal  0.1099999 -0.6260942  0.376427362 -0.88330393 -0.2153325 -1.463152
##             newCABG1    death1   hxofHT1   hxofDM1   hxofMI1 hxofPTCA1
## MI      -25.3542849 -50.256958  0.1037987 -0.62271566  0.73991136  0.3592016
## normal -0.2815125  1.014971 -0.1200665  0.06426226  0.08943254  0.3755483
```

```

##      hxofCABG1 any.event1 hxofCigmoderate hxofCignon-smoker
## MI     -1.4724274 29.078987      1.0128789      0.95301497
## normal -0.3460733 1.040342      0.0871139      0.04749324
##
## Std. Errors:
##      (Intercept)      bhr      basebp      basedp      pkhr       sbp
## MI     9.854402e-05 0.008589467 0.012665511 1.318634e-04 0.01709659 0.01889383
## normal 6.245477e-05 0.006235497 0.008346668 8.841589e-05 0.01073444 0.01200734
##      dp      dose      maxhr      pctMphr      mbp      dpmaxdo
## MI     0.0001855178 0.01585544 0.01839668 0.010130256 0.01615801 0.0001685763
## normal 0.0001317153 0.01056842 0.01106544 0.006280737 0.01053816 0.0001178029
##      dobDose      age      baseEF      dobEF gendermale      chestpain1
## MI     0.013868655 0.01790907 0.012607894 0.01657760 0.0005767397 0.0005207717
## normal 0.008946569 0.01139380 0.008984815 0.01050193 0.0002722605 0.0002204501
##      restwmal      posSE1      newMI1      newPTCA1      newCABG1
## MI     0.0006036170 0.0004083918 2.376158e-05 0.0002112625 0.0003043162
## normal 0.0003924279 0.0002190812 4.627685e-05 0.0001105744 0.0001750793
##      death1      hxofHT1      hxofDM1      hxofMI1      hxofPTCA1
## MI     1.550237e-13 0.0002895437 0.0002738295 0.0003653043 9.429108e-05
## normal 2.366644e-05 0.0001806553 0.0002759498 0.0002527662 1.095496e-04
##      hxofCABG1 any.event1 hxofCigmoderate hxofCignon-smoker
## MI     0.0002727764 0.0004400747 0.0002027765 0.0003003470
## normal 0.0002774801 0.0002127304 0.0001241857 0.0001633511
##
## Residual Deviance: 500.9864
## AIC: 628.9864

```

Esto nos muestra los coeficientes estimados para cada variable predictoras, así como las estadísticas de ajuste del modelo.

Para evaluar el rendimiento del modelo, podemos utilizar el conjunto de prueba y **calcular la matriz de confusión y la precisión:**

```

## [1] equivocal MI      normal      normal      normal      normal      normal
## [8] normal      normal      normal      normal      normal      normal      equivocal
## [15] MI      equivocal      equivocal      MI      equivocal      normal      normal
## [22] normal      normal      equivocal      normal      normal      normal      normal
## [29] normal      normal      equivocal      normal      normal      normal      equivocal
## [36] normal      normal      normal      normal      normal      normal      normal
## [43] MI      normal      normal      normal      normal      normal      equivocal
## [50] normal      MI      equivocal      equivocal      equivocal      normal      normal
## [57] normal      normal      normal      equivocal      normal      normal      normal
## [64] normal      equivocal      normal      equivocal      normal      equivocal      normal
## [71] normal      equivocal      normal      normal      normal      normal      normal
## [78] normal      MI      MI      normal      normal      normal      normal
## [85] normal      equivocal      normal      normal      normal      normal      normal
## [92] equivocal      normal      MI      equivocal      normal      equivocal      equivocal
## [99] normal      equivocal      normal      MI      normal      normal      normal
## [106] normal      normal      normal      normal      normal      normal      equivocal
## [113] normal      normal      normal      normal      normal      normal      normal
## [120] normal      normal      normal      normal      normal      normal      normal
## [127] normal      normal      equivocal      normal      MI      MI      normal
## [134] normal      normal      normal

```

```
## Levels: equivocal MI normal
##          predicciones
##          equivocal MI normal
## equivocal     11  2    27
## MI           1   6    11
## normal       13  3    62
## [1] 0.5808824
```

La **matriz de confusión** nos muestra la cantidad de predicciones correctas e incorrectas para cada categoría de la variable dependiente. La **precisión** es la proporción de predicciones correctas en relación al total de predicciones. En nuestro modelo la **precisión es del 58.09%**.

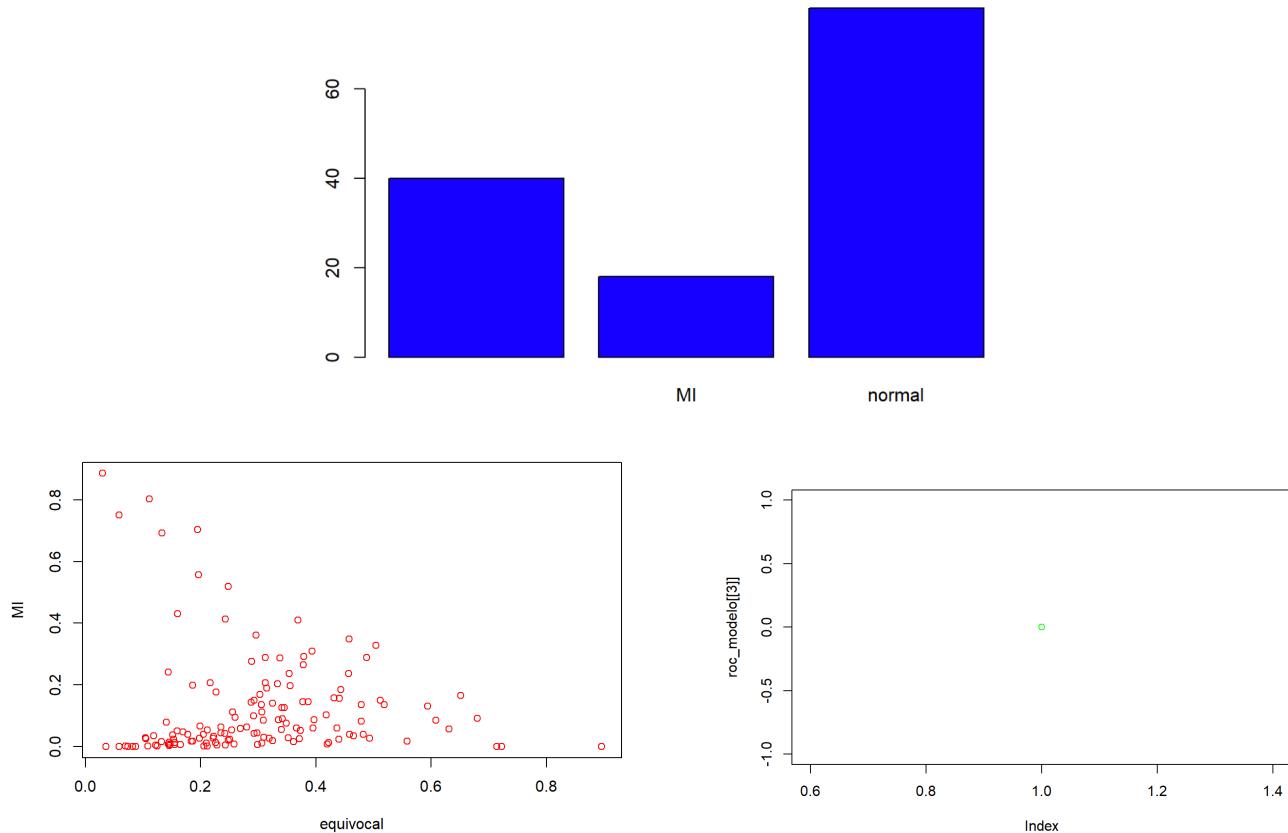
Además, podemos utilizar la **función confusionMatrix()** del paquete caret para obtener una evaluación más detallada del rendimiento del modelo:

```
## Confusion Matrix and Statistics
##          Reference
## Prediction equivocal MI normal
## equivocal     11  1    13
## MI           2   6    3
## normal       27 11    62
## Overall Statistics
## Accuracy : 0.5809
## 95% CI : (0.4933, 0.6649)
## No Information Rate : 0.5735
## P-Value [Acc > NIR] : 0.4672
##
## Kappa : 0.1838
## Mcnemar's Test P-Value : 0.0203
##
## Statistics by Class:
##          Class: equivocal Class: MI Class: normal
## Sensitivity      0.27500  0.33333  0.7949
## Specificity      0.85417  0.95763  0.3448
## Pos Pred Value   0.44000  0.54545  0.6200
## Neg Pred Value   0.73874  0.90400  0.5556
## Prevalence        0.29412  0.13235  0.5735
## Detection Rate   0.08088  0.04412  0.4559
## Detection Prevalence 0.18382  0.08088  0.7353
## Balanced Accuracy 0.56458  0.64548  0.5698
```

La matriz de confusión y las estadísticas indican que **el modelo tiene una precisión del 58.09%, con un intervalo de confianza del 95% entre 49.33% y 66.49%**.

La **tasa de error de clasificación es del 41.91%**. Los valores de sensibilidad y especificidad para cada clase indican que el modelo tiene una mayor capacidad para detectar la clase “normal” con una especificidad del 95.76%.

Ahora vamos a crear una **curva ROC (Receiver Operating Characteristic)** para evaluar la capacidad del modelo para discriminar entre las diferentes clases. Para ello, podemos utilizar la **función roc()** del paquete pROC:



4.5 Árbol de decisión.

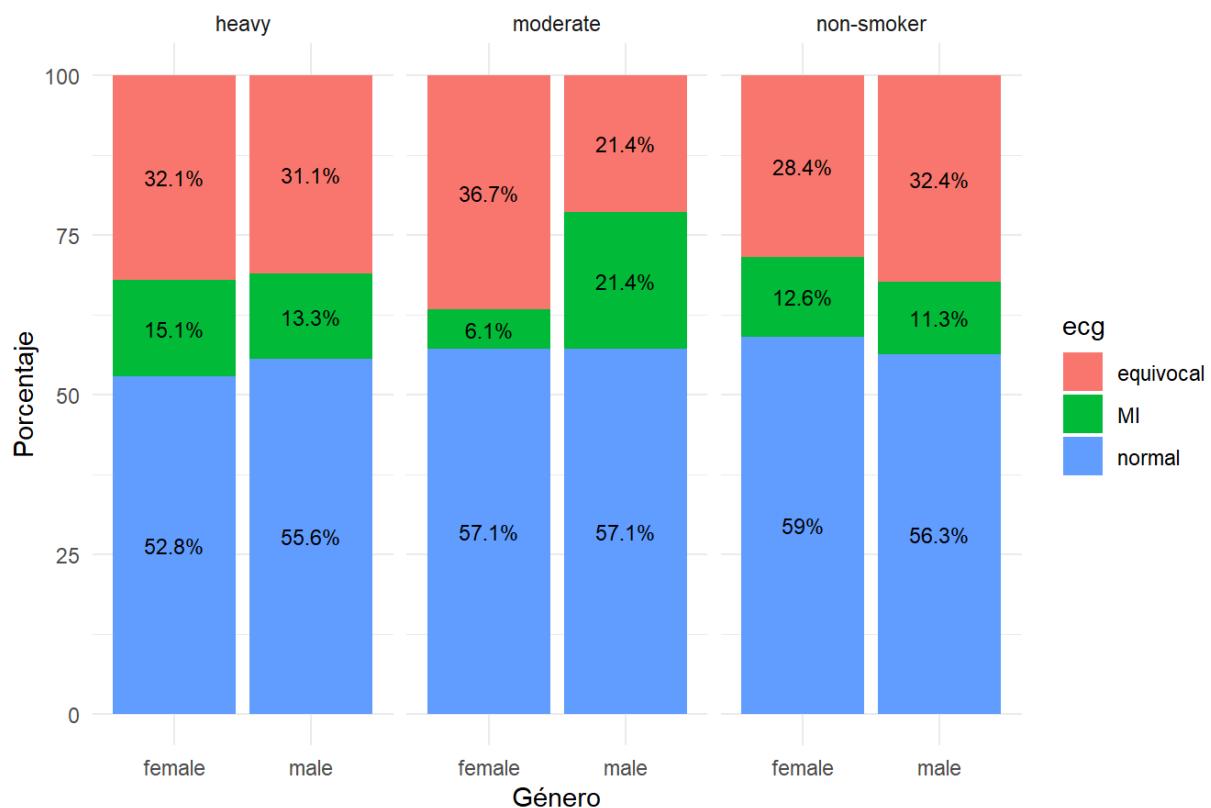
Los **árboles de decisión** son una herramienta de modelado estadístico que se utilizan para predecir el valor de una variable objetivo (también conocida como variable de respuesta) en función de una serie de variables predictoras (también conocidas como variables explicativas).

El árbol de decisión **es una estructura jerárquica similar a un diagrama de flujo**, donde cada nodo interno representa una pregunta o una condición sobre una variable predictor, y cada rama representa una posible respuesta o resultado de la pregunta o condición. Las hojas del árbol representan las predicciones finales para la variable objetivo.

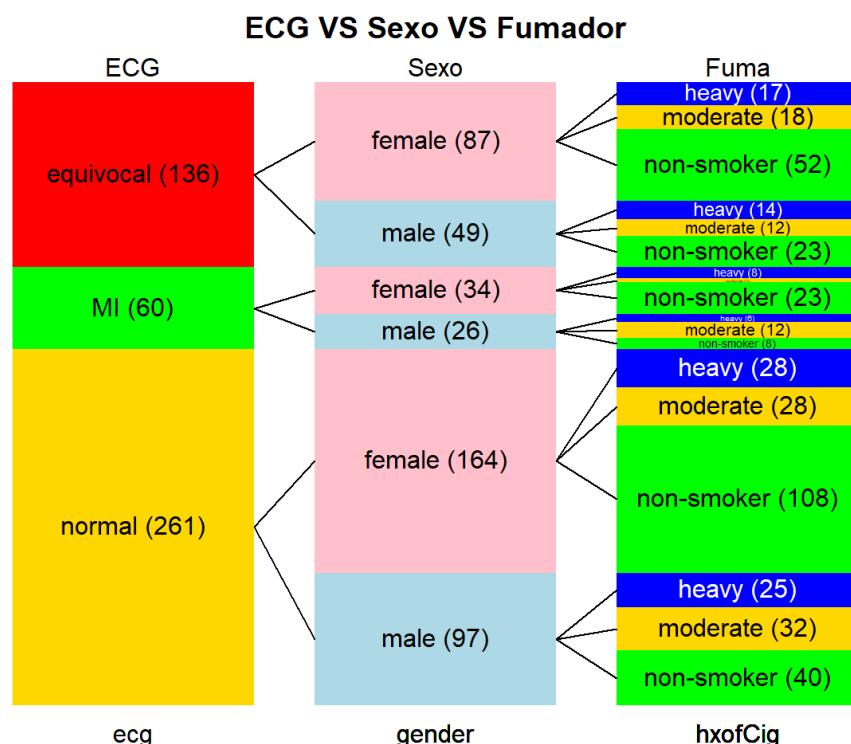
El **objetivo del árbol de decisión** es dividir el conjunto de datos en subconjuntos más pequeños y homogéneos en función de las variables predictoras, de manera que las predicciones sean lo más precisas posible. La división se realiza seleccionando la variable predictor que mejor separa los datos en términos de la variable objetivo, y luego dividiendo los datos en dos o más grupos en función de los valores posibles de la variable predictor.

Antes de realizar el árbol de decisión vamos a mostrar **la proporción de cada categoría de ecg dentro de cada género y nivel de hxofCig**, con los porcentajes etiquetados en cada segmento de las barras apiladas:

Gráfico de barras apiladas con porcentajes

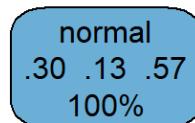


Ahora realizamos un **gráfico tipo sizetree** para entender mejor las **variables ecg, gender y hxofCig**:

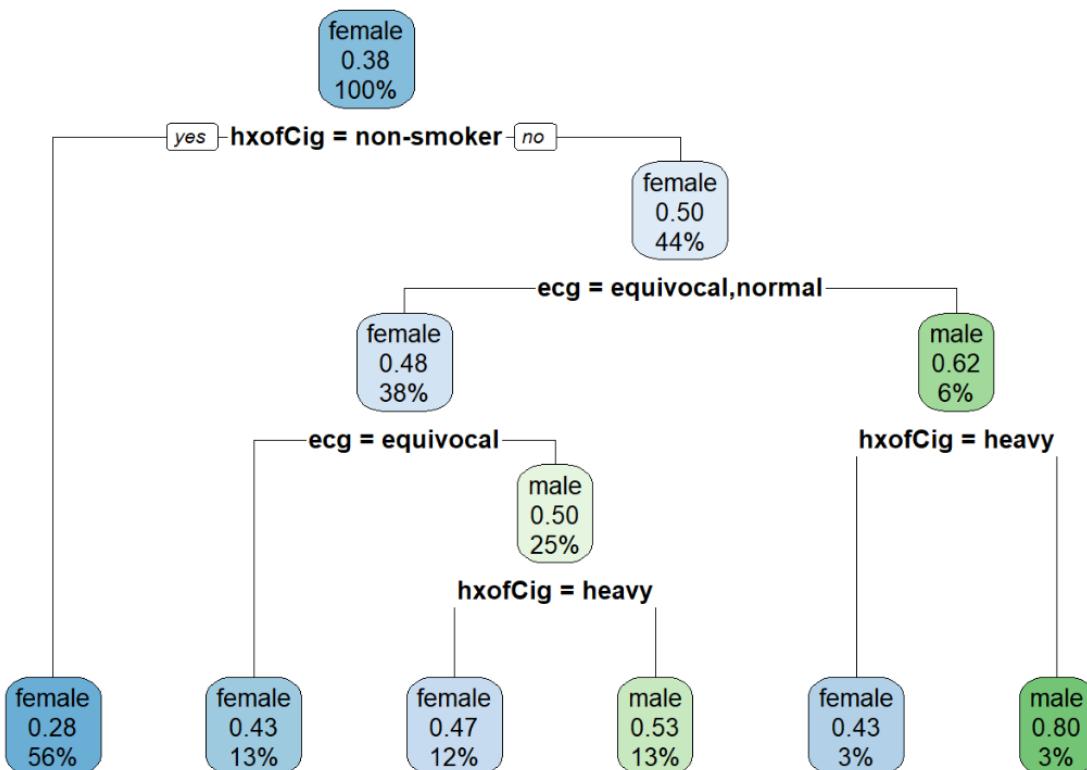


A continuación, vamos a buscar el mejor árbol de decisión entre las **variables ecg, gender y hxofCig**, no solo vamos a ajustar los modelos y calcular su precisión, sino que también vamos a visualizar cada uno de los árboles de decisión:

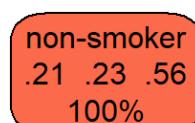
```
## Precisión del modelo con 'ecg' como respuesta: 0.571116
## Precisión del modelo con 'gender' como respuesta: 0.6520788
## Precisión del modelo con 'hxofCig' como respuesta: 0.5557987
## El mejor modelo basado en la precisión es: modelo_gender
## Árbol de decisión para 'ecg':
```



```
## Árbol de decisión para 'gender':
```



```
## Árbol de decisión para 'hxofCig':
```



El **mejor modelo de árbol de decisión** es el modelo_gender con una precisión del 0.65.

5 Conclusiones del estudio.

A continuación, vamos a mostrar todas las conclusiones que hemos podido extraer del **estudio del dataset stressEcho**:

5.1 Análisis descriptivo.

1.- El dataset contiene **32 variables y 558 observaciones**, de las cuales: 15 son variables numéricas, 14 son variables categóricas dicotómicas (1 está codificada como texto y 13 están codificadas como numéricas) y 2 son variables categóricas politómicas.

2.- Las variables con **correlaciones mayores a 0.5 sin duplicados** son: basedp, pkhr, maxhr, dp, pctMphr, dpmaxdo, mbp, dobdoce, dobEF, bhr, basebp, sbp, dose, baseEF.

3.- El dataset **no contiene valores vacíos**.

4.- El dataset **sí contiene valores extremos en diferentes variables**, sobre todo en las variables baseEF y dobEF. Hemos procedido a **eliminar 101 filas** para poder trabajar con un dataset sin valores extremos. Podríamos haber optado por otras técnicas como sustituir estos valores por la media o mediana, etc., pero al haber suficientes observaciones optamos por la eliminación de valores extremos.

5.2 Análisis inferencial.

1.- **Estudio de normalidad de las variables:** las únicas variables que cumplen con el requisito de normalidad son maxhr, pctMphr.

2.- **Estudio de la homogeneidad de varianza:** las variables que cumplen con el requisito de homogeneidad de la varianza son bhr, basebp, basedp, pkhr, sbp, dp, dose, maxhr, pctMphr, mbp, dpmaxdo, age, dobEF.

5.3 Aplicación de pruebas estadísticas.

1.- **Prueba chi-cuadrado:**

1.1.- **Variables gender y ecg:** hemos podido concluir que no hay una asociación significativa entre estas dos variables ya que el p-valor es mayor de 0.05.

1.2.- **Variables chestpain y ecg:** hemos podido concluir que sí hay una asociación significativa entre estas dos variables ya que el p-valor es menor de 0.05.

2.- **Prueba ANOVA de un factor:** se ha realizado entre las Variables maxhr y ecg. Hemos podido concluir que no hay un efecto significativo de ecg en maxhr ya que el p-valor es mayor de 0.05.

3.- **Análisis de regresión logística multinomial:** se ha realizado entre la variable ecg en función de todas las variables numéricas. Hemos podido concluir que la precisión del modelo es solo de un 58.09% con un intervalo de confianza del 95% entre 49.33% y 66.49%. La precisión del modelo es baja para poder decir que el modelo es bueno.

4.- **Árbol de decisión:** se ha realizado entre las variables gender, hxofCig y ecg. Hemos buscado el mejor árbol entre estas tres variables, y, el resultado obtenido es que el mejor modelo de árbol de decisión es el modelo_gender con una precisión del 65%.

6 Aplicación Shiny.

Durante la realización de esta práctica nos dimos cuenta que siempre tienes que “perder tiempo” mirando cualquier dataset.

Para intentar paliar este problema **hemos creado una aplicación Shiny** que realiza un análisis descriptivo de cualquier dataset en formato .csv. Dicha aplicación **es pública** y está alojada en el siguiente link:

https://arrocar.shinyapps.io/ANALISIS_DATOS/

También adjuntamos todo el código en la carpeta source de esta práctica.