

## Activitat grupal.

### Pràctica 2: Tipologia i cicle de vida de les dades

- 1. Descripció del data set.** Perquè és important i quina pregunta/problema pretén respondre?

*El dataset ofereix informació de 303 pacients a través de 14 variables.*

*L'objectiu es veure si es pot construir un model predictiu per tal d'explicar les causes d'un atac de cor.*

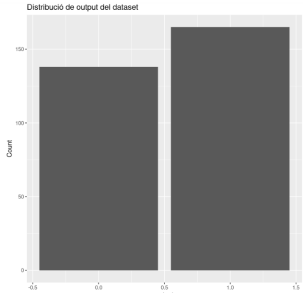
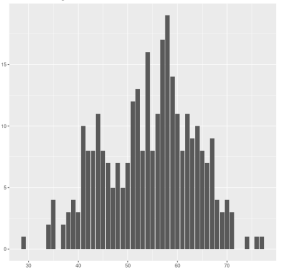
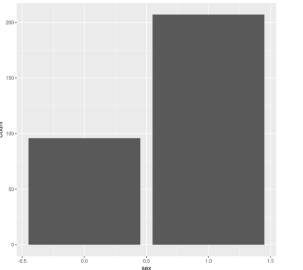
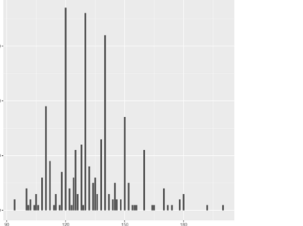
*Hem seleccionat la base de dades de Kaggle proposades. Heart attack analysis prediction dataset [link](#)*

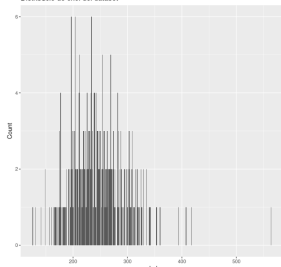
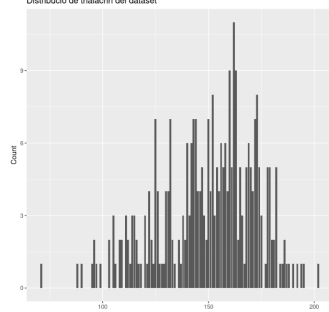
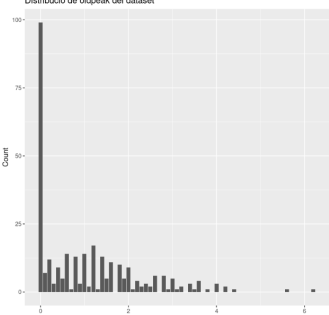
*L'anàlisi l'hem realitzat amb R.*

Variable	Descripció	tipus
Age	Edat del pacient	contínua
Sex	Sexe del pacient	Dicotòmica (0= Home; 1= Dona)
cp (chest pain type)	dolor al pit	1. Typical angina 2. Atypical angina 3. Non-angina pain 4. Asymptomatic.
Tstbps.- resting blood pressure (mm Hg)	Pressió sanguínia en repós	Contínua
Chol colesterol mg/dl via BMI sensor	Colesterol via sensor IBM	Contínua
fbs : (fasting blood sugar > 120 mg/dl)	Sucre en sang	Dicotòmica (1 = true; 0 = false)
rest_ecg : resting electrocardiographic results	Resultats electrocardiogràfics en repos	Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular

		<i>hypertrophy by Estes' criteria</i>
<i>thalach : maximum heart rate achieved</i>		<i>Continua</i>
<i>exang: exercise induced angina</i>		<i>Dicotòmica (1 = yes; 0 = no)</i>
<i>oldpeak: previous peak</i>		
<i>slp: slope</i>		
<i>caa: number of major vessels (0-3)</i>		<i>0-3</i>
<i>thall: thal rate</i>		
<i>Output</i>		<i>dicotòmica (0= less chance of heart attack 1= more chance of heart attack)</i>

<i>variables</i>	<i>clase</i>
<i>age</i>	<i>integer</i>
<i>sex</i>	<i>integer</i>
<i>cp</i>	<i>integer</i>
<i>trtbps</i>	<i>integer</i>
<i>chol</i>	<i>integer</i>
<i>fbs</i>	<i>integer</i>
<i>restecg</i>	<i>integer</i>
<i>thalachh</i>	<i>integer</i>
<i>exng</i>	<i>integer</i>
<i>oldpeak</i>	<i>numeric</i>
<i>slp</i>	<i>integer</i>
<i>caa</i>	<i>integer</i>
<i>thall</i>	<i>integer</i>
<i>output</i>	<i>integer</i>

	<p>Distribució de la variable output:</p> <p>El percentatge de pacients amb output positiu és de 54.45545 mentre que el percentatge de pacients amb output negatiu és de 45.54455</p>
	<p><i>Distribució de la variable edat</i></p>
	<p><i>Distribució de la variable sexe</i></p> <p>El percentatge de pacients dones és de 68.31683 mentre que el percentatge de pacients amb output homes és de 31.68317</p>
	<p><i>Distribució de la variable trtbps</i></p>

 <p>Distribució de chol del dataset</p>	<p><i>Distribució de la variable chol</i></p>
 <p>Distribució de thalachh del dataset</p>	<p><i>Distribució de la variable thalachh</i></p>
 <p>Distribució de oldpeak del dataset</p>	

Observem com és distribueixen les variables gràficament:

## 2. Integració i selecció de les dades d'interès a analitzar.

*Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.*

*Hem fraccionat el dataset quan ha sigut necessari per realitzar les anàlisis.*

## 3. Neteja de les dades.

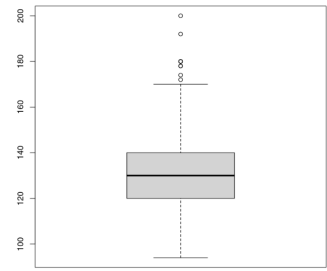
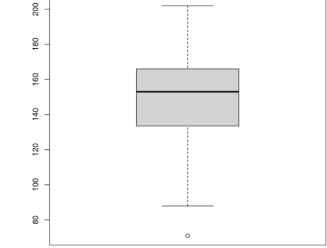
3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

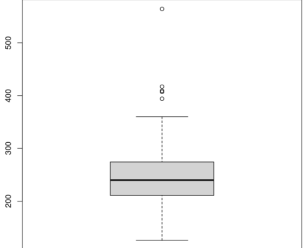
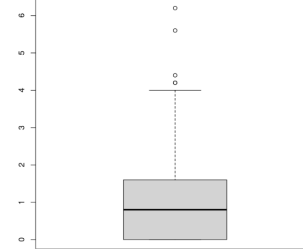
*En general no hem trobat valors NA.*

*La variable oldpeak té molts valors zero, tenim el dubte de si són valors perduts, però segurament és que en el període anterior no tenien dades i per això decidim no fer res al respecte*

3.2. Identifica i gestiona els valors extrems.

*Busquem valors extrems (outliers) en les columnes trtbps, thalachh, chol i oldpeak dibuixant uns gràfics boxplot*

	<p><i>La suma de files amb outliers a la columna trtbps és de 13</i></p>
	<p><i>La suma de les files amb outliers a la columna thalachh és de 3</i></p>

	<p><i>La suma de les files amb outliers a la columna chol és de 8</i></p>
	<p><i>La suma de les files amb outliers a la columna oldpeak és de 8</i></p>

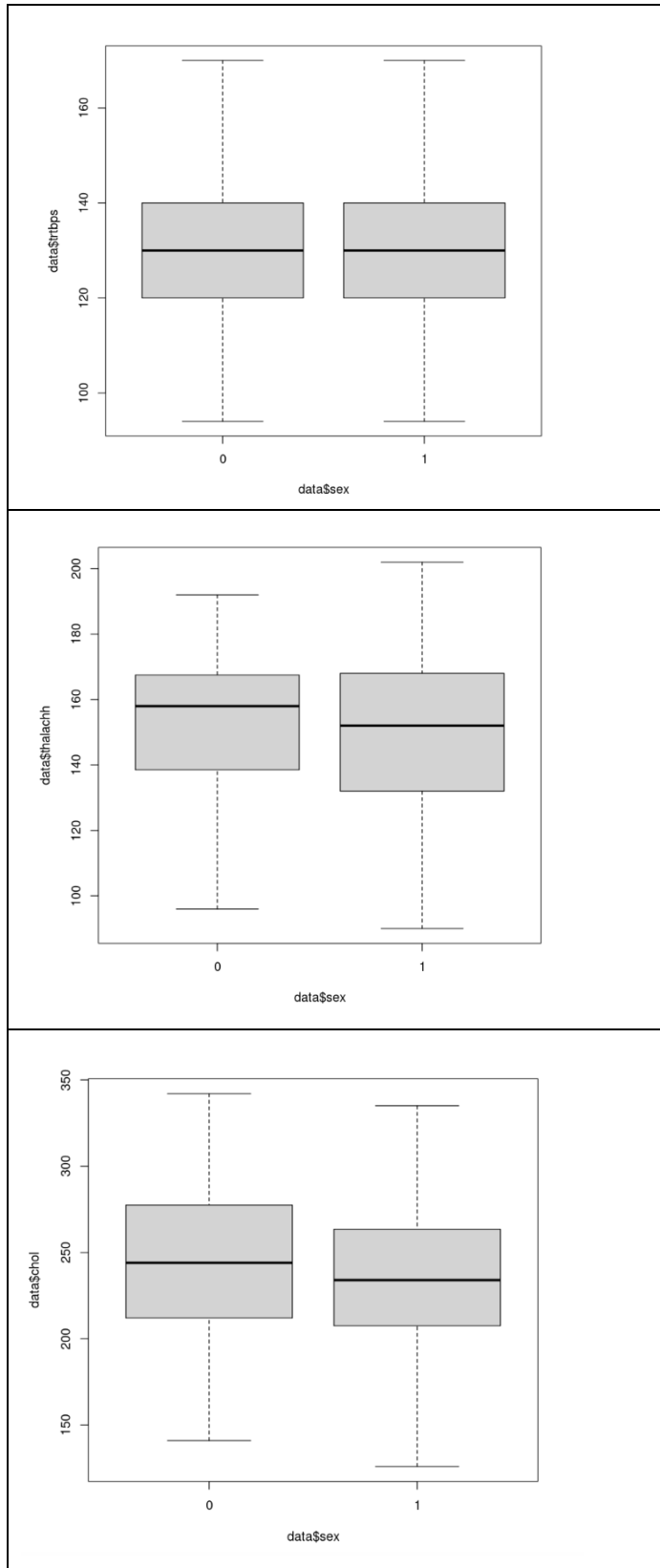
*En trobar outlier es pot procedir de diferents maneres, en funció de la causa de que hi hagi un valor que es distancia de la resta, es pot decidir d'imputar un valor constant o si s'escau es podria també d'assignar el valor de la mitjana per el grup al que pertanyi l'observació, per exemple es podria imputar el valor de la mitjana d'homes o dones.*

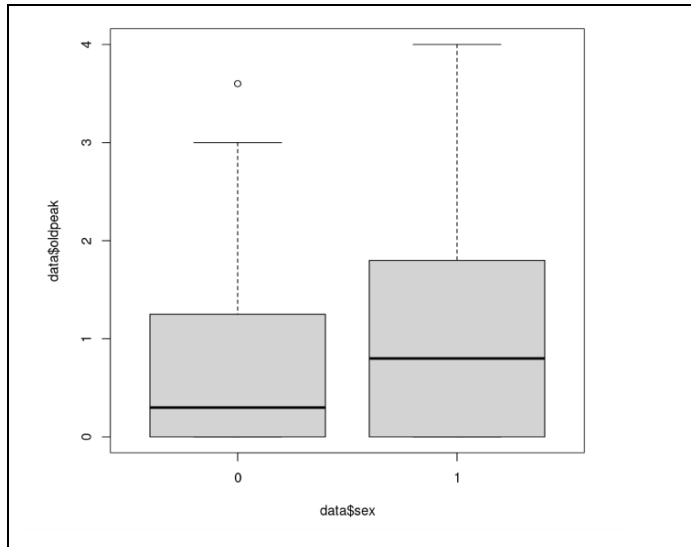
*Hi ha el mètode kNN (k-Nearest Neighbours) però es pot veure afectat per la tria de la k. També es pot suprimir el registre perquè els resultats no es vegin alterats per aquest valors erronis.*

*Hem decidit d'eliminar els registres amb valors perduts. Per fer-ho hem fet servir la comanda subset per descartar els casos extrems.*

*El dataset queda amb el nombre de files de 279*

*Observem el comportament d'aquestes variables discriminant per gènere:*







#### 4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?)

*Hem vist que al dataset hi ha variables contínues i variables categòriques: dicotòmiques i de 3 o 4 categories. I en aquest sentit hem orientat l'anàlisi.*

*A partir de la variable output que predir els que tenen risc d'atac de cor. En aquest sentit hem construït un model de regressió lineal.*

*Per gènere comparativa risc de patir un atac de cor homes o dones. Observant si hi ha independència entre les diferents variables en funció del gènere. Aquesta comparativa l'hem portat a terme amb tests de chi quadrat (o el test exacte de Fisher en cas de no ser apropiat aplicar el test de Chi Quadrat per tenir valors inferiors a 5 en alguna de les caselles de la taula de contingència del creuament de les variables 2 a 2. S'ha portat a terme amb les diferents variables categòriques o dicotòmiques en funció del gènere.*

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

*S'ha comprovat la normalitat de les variables amb el test de Normalitat de Shapiro-Wilk i Tenint que la  $H_0$  és que la variable segueix una distribució normal, si el p-valor és més petit que el nivell de significació a rebutgem la  $H_0$  i per tant no té una distribució normal, en el cas de tenir un p-valor més gran (com en la majoria de les variables que hem realitzat el test) no podem rebutjar la hipòtesi nul·la i per tant hem de concloure que la variable Chol segueix una distribució normal. Però donat el Teorema central del límit en el cas de tenir una  $n$  prou gran es pot considerar que segueix una distribució normal i es poden aplicar tècniques d'estadística paramètrica però els resultats seran poc robustos.*

*Per veure l'homogeneïtat de la variància es poden fer servir els tests: test de Levene, que s'aplica quan les dades segueixen una distribució normal, així com el test de Fligner-Killeen, que es tracta de l'alternativa no paramètrica, utilitzada quan les dades no compleixen amb la condició de normalitat.*

En el nostre cas hem procedit a fer un test d'hipòtesis de comparació de mitjanes entre homes i dones *F-test* de raó de variàncies de la variable que té distribució Normal chol en funció del gènere i no podem rebutjar la hipòtesi nul·la de que el quocient de les variàncies dels dos grups és diferent a zero, per tant podem dir que la variància és homogènia.

- 4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

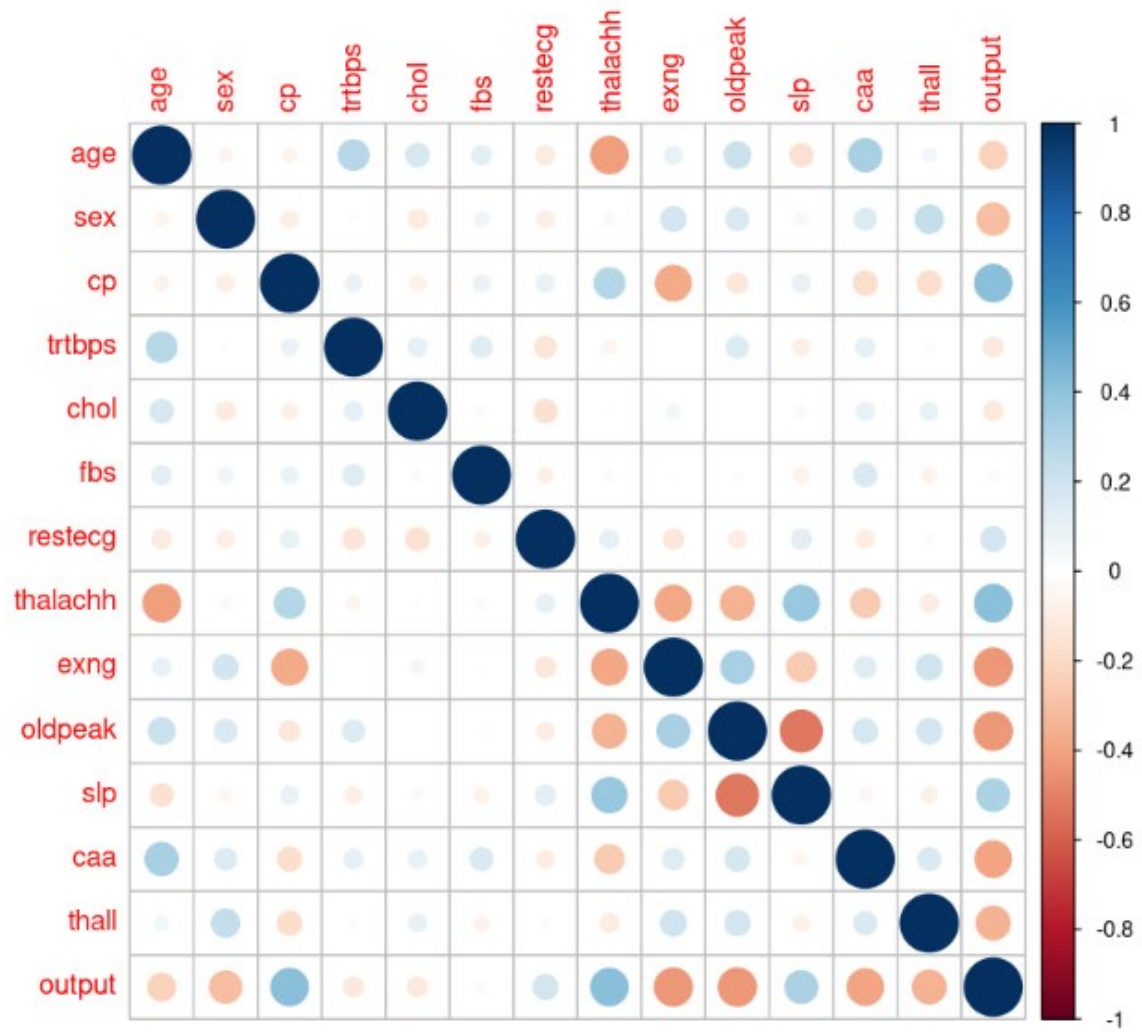
### 1. Contrastos d'hipòtesis.

Tests de Normalitat:		age		trtbps		chol		thalachh		oldpeak	
Només en el cas de la variable chol no es rebutja la $H_0$		-----		-----		-----		-----		-----	
		pvalue		0.0203		0.003591		0.1485		7.648e-05	
Homocedasticitat											

### 2. Correlació

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.00	-0.06	-0.06	0.28	0.16	0.11	-0.11	-0.42	0.09	0.21	-0.15	0.33	0.06	-0.23
sex	-0.06	1.00	-0.09	0.01	-0.11	0.06	-0.09	-0.03	0.18	0.16	-0.05	0.14	0.24	-0.31
cp	-0.06	-0.09	1.00	0.08	-0.07	0.08	0.10	0.28	-0.38	-0.12	0.09	-0.17	-0.17	0.41
trtbps	0.28	0.01	0.08	1.00	0.10	0.13	-0.14	-0.06	0.00	0.15	-0.08	0.11	-0.02	-0.12
chol	0.16	-0.11	-0.07	0.10	1.00	0.03	-0.16	-0.01	0.06	-0.01	0.03	0.09	0.09	-0.11
fbs	0.11	0.06	0.08	0.13	0.03	1.00	-0.08	-0.03	0.01	0.02	-0.07	0.16	-0.06	-0.03
restecg	-0.11	-0.09	0.10	-0.14	-0.16	-0.08	1.00	0.10	-0.12	-0.09	0.12	-0.09	0.03	0.18
thalachh	-0.42	-0.03	0.28	-0.06	-0.01	-0.03	0.10	1.00	-0.38	-0.34	0.37	-0.25	-0.10	0.42
exng	0.09	0.18	-0.38	0.00	0.06	0.01	-0.12	-0.38	1.00	0.32	-0.26	0.13	0.20	-0.43
oldpeak	0.21	0.16	-0.12	0.15	-0.01	0.02	-0.09	-0.34	0.32	1.00	-0.53	0.18	0.19	-0.43
slp	-0.15	-0.05	0.09	-0.08	0.03	-0.07	0.12	0.37	-0.26	-0.53	1.00	-0.05	-0.08	0.32
caa	0.33	0.14	-0.17	0.11	0.09	0.16	-0.09	-0.25	0.13	0.18	-0.05	1.00	0.15	-0.39
thall	0.06	0.24	-0.17	-0.02	0.09	-0.06	0.03	-0.10	0.20	0.19	-0.08	0.15	1.00	-0.34
output	-0.23	-0.31	0.41	-0.12	-0.11	-0.03	0.18	0.42	-0.43	-0.43	0.32	-0.39	-0.34	1.00

Podem observar que la major correlació negativa amb output és oldpeak (-0.43) i la major correlació positiva cp i thalachh (0.28) que estan allunyats de -1 o de 1



### 3. Regressió.

A partir de les observacions es pren un 0.8 de les observacions del dataset data i es guarda per fer entrenament i determinar el model (train) i el 0.2 restant es reserva per testejar el model predictiu (test)

```
lm(formula = output ~ age + sex + cp + trtbps + chol + fbs + restecg + thalachh + exng + oldpeak + slp + caa + thall, data = train)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.89275	-0.22510	0.05419	0.24653	0.87370

Coefficients:

	Estimate Std.	Error	t value	Pr(> t )
(Intercept)	0.7557434	0.3808685	1.984	0.048529 *
age	-0.0005069	0.0033506	-0.151	0.879906
sex	-0.2216493	0.0583599	-3.798	0.000191 ***
cp	0.1070723	0.0269074	3.979	9.52e-05 ***
trtbps	-0.0016189	0.0017129	-0.945	0.345692
chol	-0.0006088	0.0005743	-1.060	0.290354
fbs	0.0673672	0.0683451	0.986	0.325419
restecg	0.0285449	0.0491558	0.581	0.562063
thalachh	0.0034309	0.0013784	2.489	0.013586 *
exng	-0.1198486	0.0634863	-1.888	0.060433 .
oldpeak	-0.0646299	0.0293732	-2.200	0.028876 *
slp	0.0846552	0.0491295	1.723	0.086342 .
caa	-0.0970515	0.0264639	-3.667	0.000311 ***
thall	-0.1141214	0.0409612	-2.786	0.005823 **

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3653 on 210 degrees of freedom

Multiple R-squared: 0.4972, Adjusted R-squared: 0.466

F-statistic: 15.97 on 13 and 210 DF, p-value: < 2.2e-16

Veiem mitjançant la funció `lm` que les variables amb major explicació sobre la variable `output` son `sex`, `cp` i `caa`. El model no és massa bo donat el `R-squared` de 0.50 No obstant això intentarem predir noves incorporacions.

El model és majoritàriament correlacionat, la intersecció i les variables `sex`, `cp`, `thalachh`, `exng`, `oldpeak`, `slp`, `caa`, `thall` tenen uns valors estadísticament significatius, per això tornem a cridar la funció `lm` per només els coeficients correlacionats:

`lm(formula = output ~ sex + cp + thalachh + exng + oldpeak + slp + caa + thall, data = train)`

Residuals:

Min	1Q	Median	3Q	Max
-0.91758	-0.22544	0.06552	0.24343	0.90183

Coefficients:

	Estimate Std.	Error	t value	Pr(> t )
(Intercept)	0.408356	0.218259	1.871	0.062707 .
sex	-0.215509	0.056665	-3.803	0.000186 ***
cp	0.107618	0.026469	4.066	6.72e-05 ***
thalachh	0.003461	0.001254	2.760	0.006276 **
exng	-0.124359	0.062481	-1.990	0.047818 *
oldpeak	-0.069063	0.029070	-2.376	0.018393 *
slp	0.080529	0.048521	1.660	0.098439 .
caa	-0.099627	0.025788	-3.863	0.000148 ***
thall	-0.118345	0.040301	-2.937	0.003680 **

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3645 on 215 degrees of freedom

Multiple R-squared: 0.4874, Adjusted R-squared: 0.4683

F-statistic: 25.55 on 8 and 215 DF, p-value: < 2.2e-16

*Posem a prova el model calculat amb el subset test:*

| | *Real* | *Predicció* | *Diferencia?* |

|:---|----:|-----:|:-----|

|3 | 1| 1|No |

|10 | 1| 1|No |

|13 | 1| 1|No |

|20 | 1| 1|No |

|23 | 1| 1|No |

|50 | 1| 1|No |

|51 | 1| 1|No |

|56 | 1| 1|No |

|58 | 1| 1|No |

|59 | 1| 1|No |

|62 | 1| 1|No |

|63 | 1| 1|No |

|67 | 1| 1|No |

|70 | 1| 1|No |

|72 | 1| 1|No |

|73 | 1| 1|No |

|78 | 1| 1|No |

|80 | 1| 1|No |

|85 | 1| 1|No |

|100 | 1| 1|No |

|103 | 1| 1|No |

|114 | 1| 1|No |

|117 | 1| 1|No |

|125 | 1| 1|No |

|127 | 1| 1|No |

|128 | 1| 1|No |

|133 | 1| 1|No |

|134 | 1| 1|No |

|135 | 1| 1|No |

|136 | 1| 1|No |

|140 | 1| 0|Si |

|147 | 1| 1|No |

|148 | 1| 1|No |

155	1	1 No	
161	1	1 No	
163	1	1 No	
165	1	1 No	
177	o	o Si	
182	o	o Si	
191	o	o Si	
207	o	o Si	
214	o	o Si	
219	o	o Si	
223	o	1 Si	
228	o	o Si	
239	o	o Si	
240	o	o Si	
254	o	o Si	
257	o	o Si	
260	o	1 Si	
262	o	1 Si	
265	o	o Si	
271	o	o Si	
275	o	o Si	
276	o	o Si	
279	o	1 Si	

#### 4. Tests Chi Quadrat de variables qualitatives.

Number of factors: 2			Number of cases in table: 279		
table(data\$output, data\$sex) summary(table(data\$output, data\$sex))			<i>p-valor significatiu rebutgem Ho</i>		
	0: Home	1:Dona	Test for independence of all factors: Chisq = 26.704, df = 1, p-value = 2.371e-07		
0:no risc	17	106			
1:risc	66	90			
table(data\$fbs, data\$sex) summary(table(data\$fbs, data\$sex))			<i>p-valor no significatiu, no podem rebutjar Ho</i>		
	0: Home	1:Dona	Test for independence of all factors: Chisq = 1.1741, df = 1, p-value = 0.2786		
0:false	74	165			
1:true	9	31			
table(data\$cp, data\$sex) summary(table(data\$cp, data\$sex))			<i>p-valor significatiu rebutgem Ho</i>		
	0: Home	1:Dona	Test for independence of all factors: Chisq = 9.148, df = 3, p-value = 0.02739		
0: typical angina	29	98			
1: atypical angina	18	31			
2: non-anginal pain	32	49			
3: asymptomatic	4	18			
table(data\$restecg , data\$sex) summary(table(data\$restecg, data\$sex)) test <- fisher.test(table(data\$restecg , data\$sex)) test			<i>La Taula té caselles inferiors a 4 o per això cal fer un test de Fisher (no paramètric) p-valor no significatiu</i>		
	0: Home	1:Dona	Fisher's Exact Test for Count Data  data: table(data\$slp, data\$sex) p-value = 0.6523 alternative hypothesis: two.sided		
0: normal	36	100			
1: ST-T wave abnormality	45	96			
2: left ventricular hypertrophy	2	0			
table(data\$thall , data\$sex) summary(table(data\$thall, data\$sex)) test <- fisher.test(table(data\$thall , data\$sex)) test			<i>La Taula té caselles inferiors a 4 o per això cal fer un test de Fisher (no paramètric) p-valor significatiu</i>		
	0: Home	1:Dona	Fisher's Exact Test for Count Data  data: table(data\$thall, data\$sex) p-value = 1.143e-11 alternative hypothesis: two.sided		
0:	1	1			
1:	1	16			
2:	72	85			
3:	9	94			
maximum heart rate achieved					
table(data\$exng , data\$sex) summary(table(data\$exng, data\$sex))			<i>p-valor significatiu rebutgem Ho</i>		
	0: Home	1:Dona	Test for independence of all factors: Chisq = 9.464, df = 1, p-value = 0.002096		
0:no	68	124			
1:sí	15	72			



*Com a conclusió podem veure que hi ha variables en les que el fet de ser home o dona no influeix en les observacions, no hi ha diferències significatives a les variables fbs, restecg. En canvi hi ha diferències significatives entre homes i dones a les variables output, cp, thall, exng.*

**5. Representació dels resultats a partir de taules i gràfiques.** Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

**6. Resolució del problema.** A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Podem predir a partir d'una regressió lineal dèbil (amb una R de 50) a partir de les variables:

*Sex, cp, thalachh, exng, oldpeak, slp, caa, thall*

*I respecte si els homes tenen més risc o menys que els dones de patir un infart hem pogut veure que només en algunes de les variables hi ha diferències: output, cp, thall, exng. I en canvi en les variables fbs, restecg el fet de ser home o dona no representa un canvi.*

**7. Codi.** Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python

**8. Vídeo.** Realitzar un breu vídeo explicatiu de la pràctica (màxim 10 minuts) on tots els integrants de l'equip expliquin amb les seves pròpies paraules el desenvolupament de la pràctica, basant-se en les preguntes de l'enunciat per a justificar i explicar el codi desenvolupat. Aquest vídeo s'haurà de lliurar a través d'un enllaç al Google Drive de la UOC (<https://drive.google.com/...>), juntament amb l'enllaç al repositori Git lliurat.

Contribucions	Signatura
Investigació prèvia	SGM, LTA
Redacció de respostes	SGM, LTA
Desenvolupament del codi	SGM, LTA

Participació al vídeo	SGM, LTA
-----------------------	----------

## Bibliografia

- Calvo M., Subirats L., Perez D. (2019). Introduccion a la limpieza y analisis de los datos. Editorial UOC.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with
- Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Tutorial de Github <https://guides.github.com/activities/hello-world>.
- Eina per a la realitzacio de grafiques: <https://www.data-to-viz.com/>