PRÀCTICA 2
M2.951 Tipologia i cicle de les dades

Gener 2023

Pau Casanova Pedrol

1. Descripció del dataset.

S'ha escollit dataset "Heart Attack Analysis & Prediction", proposat en l'enunciat, ja que compleix perfectament les característiques necessàries per a la realització de tots els exercicis de la pràctica. Aquest és l'enllaç de Kaggle des don s'ha descarregat:

https:/, /www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Aquest conjunt conté 303 observacions de pacients amb 14 atributs de diversa tipologia: categòrics, numèrics o binaris. Les variables que comprén són les següents:

-Age: Edat del pacient

-Sex: sexe del pacient

-cp: tipus de dolor al pit

- -0 = angina típica
- -1 = angina atípica
- -2 = dolor no anginal
- -3 = assimptomàtic

-trtbps: pressió sanguínia en repòs.

-chol: colesterol en sang en mg/dl

-fbs: sucre en sang en dejú (>120 mg/dl)

- -0 = veritable
- -1 = fals

-rest_ecg: resultats electrocardiograma en repòs

- -0 = normals
- -1 = mostra anormalitat d'ona T
- -2 = mostra probable hipertrofia al ventricle esquerre

-thalach: màxim ritme cardíac assolit

-exang: angina induïda per exercici

- -1=si
- -0=no

-old peak: depressió ST induïda per exercici

-slp: pendent del segment ST en exercici

- -0= en pujada
- -1= plana
- -2= en baixada

-caa = nombre de vasos principals (0-3)

-thall: talassèmia

- -0 = null
- -1 = fixed defect
- -2 = normal
- -3 = reversable defect

-output: diagnòstic d'infart (estat angiogràfic de la malaltia)

- 0: <50% reducció del diàmetre (menor risc de malaltia cardíaca)
- -1: >50% reducció del diàmetre (major risc de malaltia cardíaca)

El tema que tracta aquest conjunt de dades és molt rellevant ja que comprendre els factors que influeixen en les possibilitats de tenir un atac de cor és clau de cara a millorar la prevenció de l'infart, que és una de les principals causes de mort en els països occidentals. A més, els anàlisis estadístics i el desenvolupament de models de classificació o regressió són aspectes fonamentals del desenvolupament i la investigació de la medicina actual.

Aquestes són algunes de les preguntes que podríem respondre per mitjà de l'anàlisi d'aquest conjunt de dades:

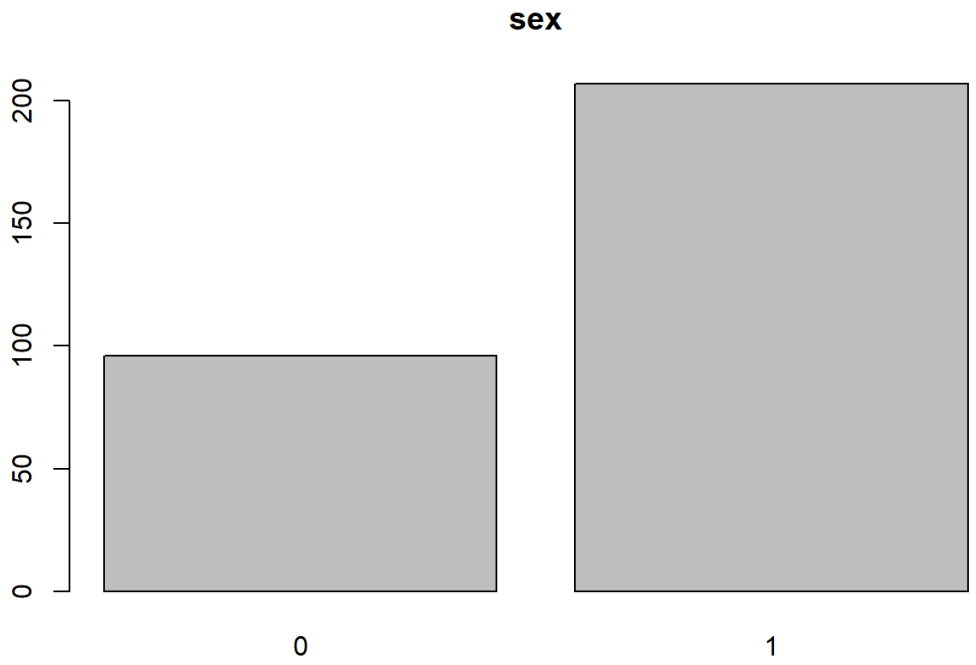
-La diferència en el risc de patir una malaltia entre homes i dones cardíaca és estadísticament significativa?

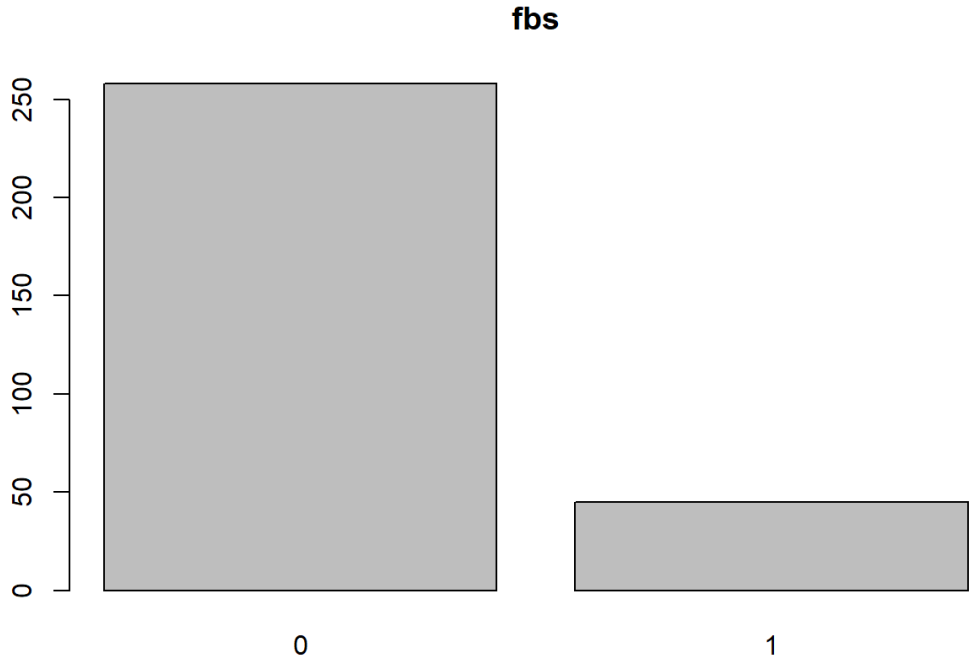
-Hi ha alguna correlació entre les variables nivell de colesterol i talassèmia.

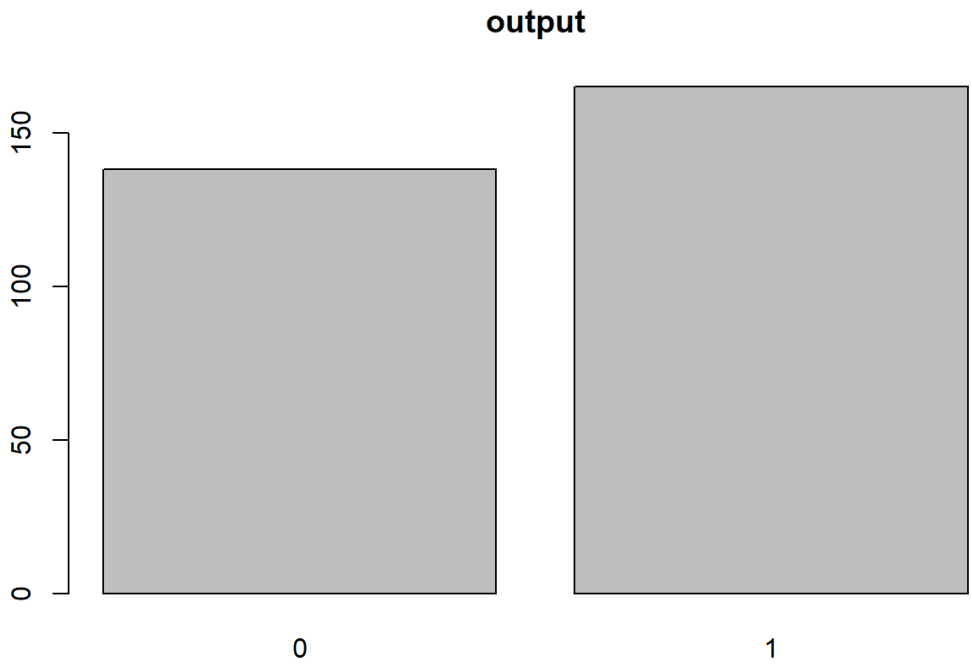
-La variable numèrica pressió sanguínia té una relació significativa amb la variable output?

Amb aquestes comprovacions i algunes altres, podem mirar de comprendre millor les causes de les malalties cardíaques i trobar amb més facilitat les maneres de prevenir-les.

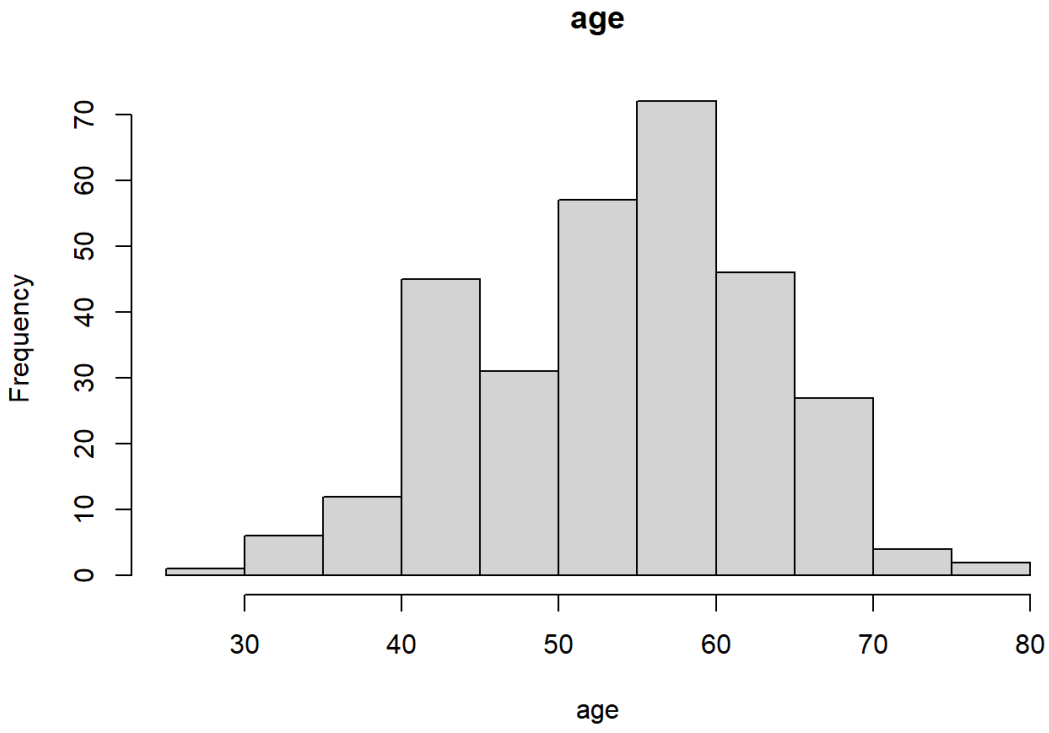
Per fer un primer cop d'ull a l'estructura de les dades, pot ser útil observar les freqüències d'algunes variables categòriques:

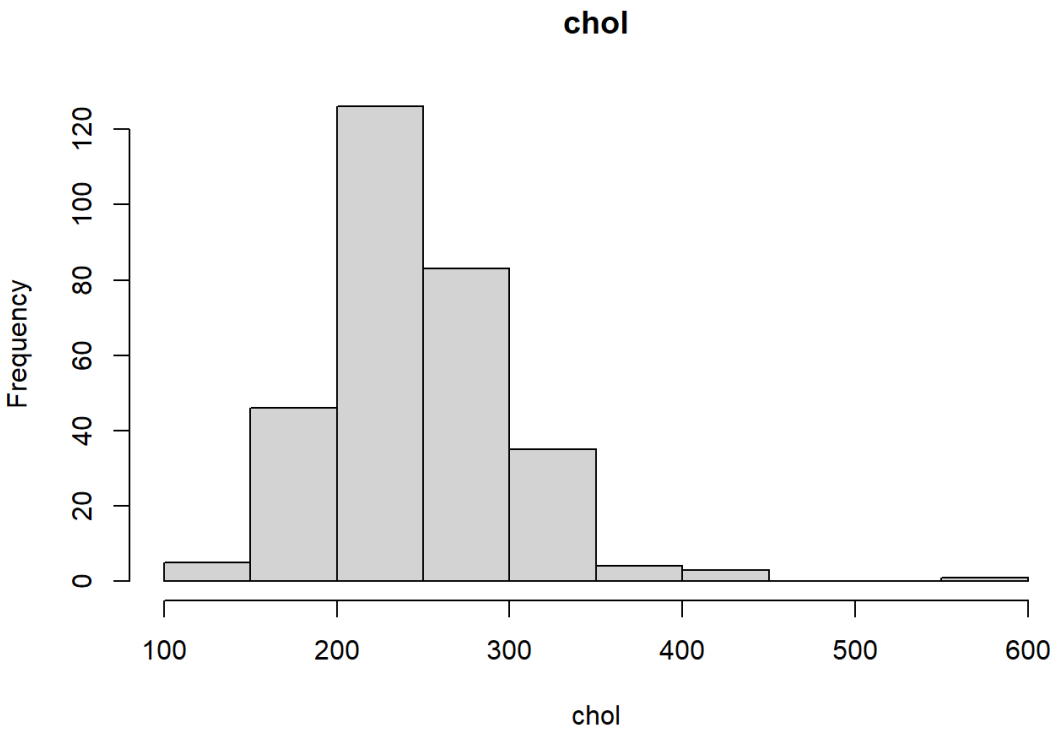


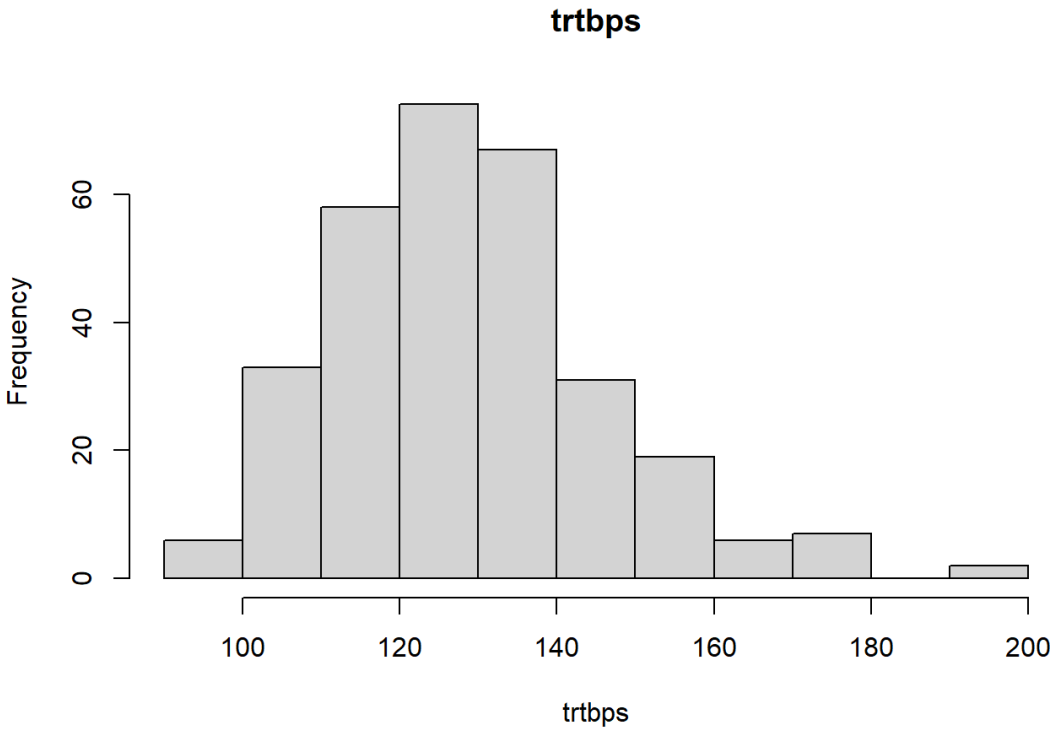




Per veure l'estructura de les variables numèriques, pot ser útil mostrar la seva distribució:




chol
Frequency
0
20
40
60
80
100
120
100
200
300
400
500
600
chol


trtbps
Frequency
0
20
40
60
100
120
140
160
180
200
trtbps

2. Integració i selecció

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.

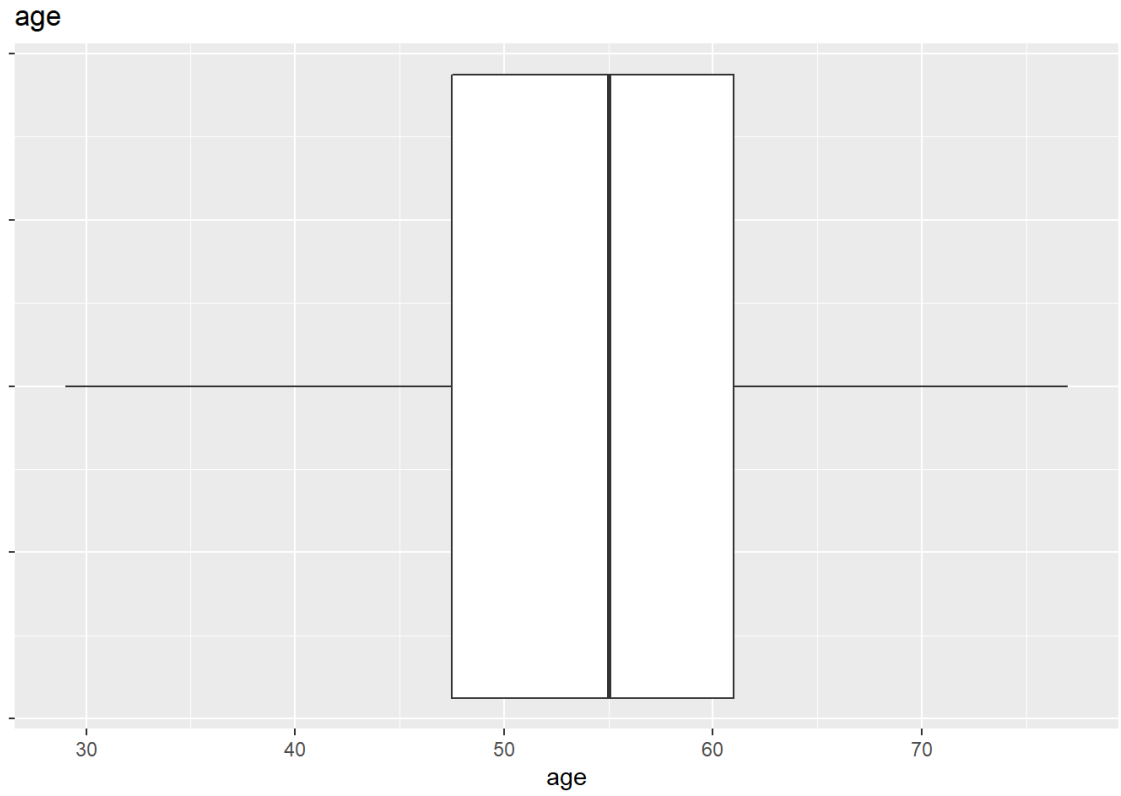
Com que no tenim accés a més dades dels mateixos pacients, ni a més observacions d'aquestes variables en altres pacients, no podem addicionar altres datasets a l'original. En principi tampoc té sentit fer una subselecció de les dades originals, ja que no sabem si són rellevants o no per a l'objectiu del nostre anàlisi, fonamentalment enfocat a analitzar possibles causes o factors de risc de les malalties cardíaques.

Tampoc hi ha variables molt correlacionades que no aportin informació addicional i de les que puguem prescindir, de manera que deixem el dataset original tal i com està.

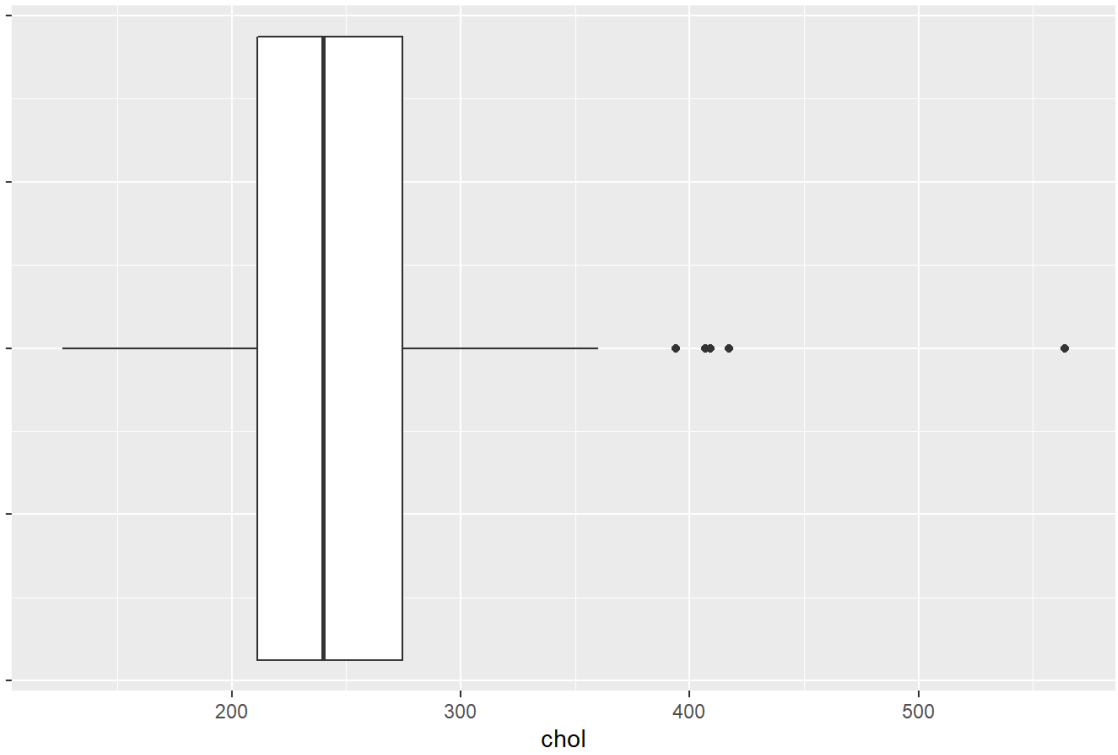
3. Outliers

Identifica i gestiona els valors extrems.

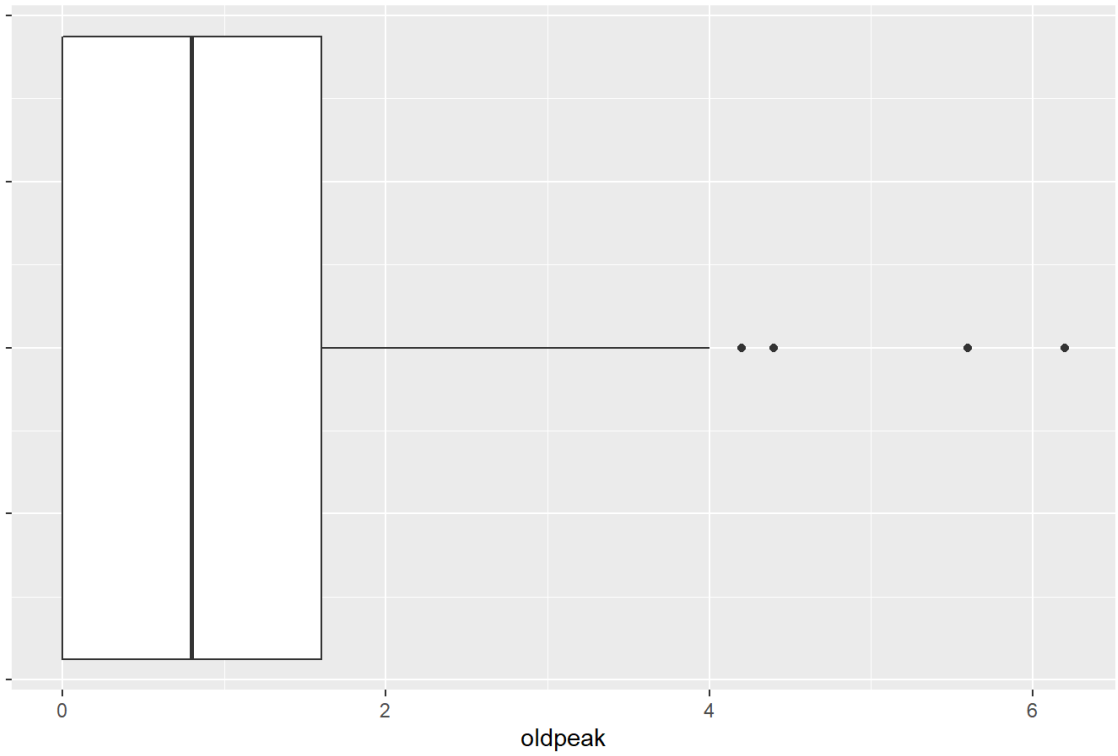
Per detectar visualment la presència d'outliers, es mostra un gràfic de caixa de les variables numèriques.

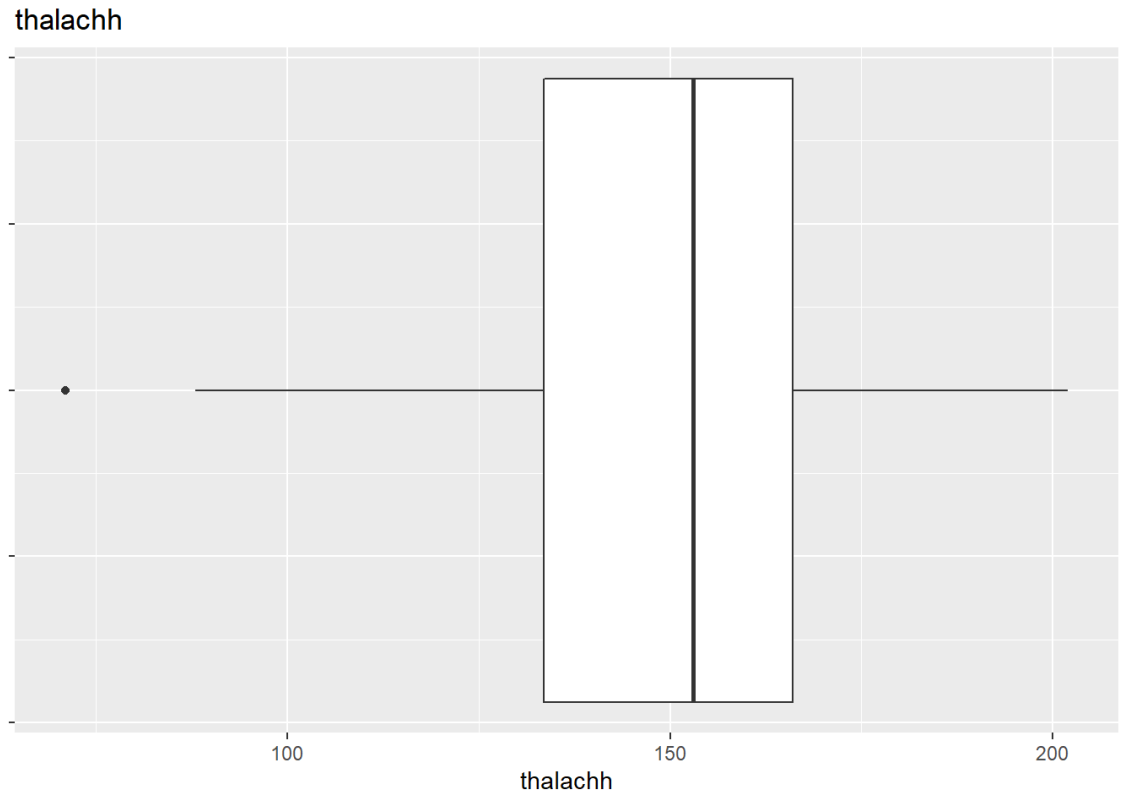


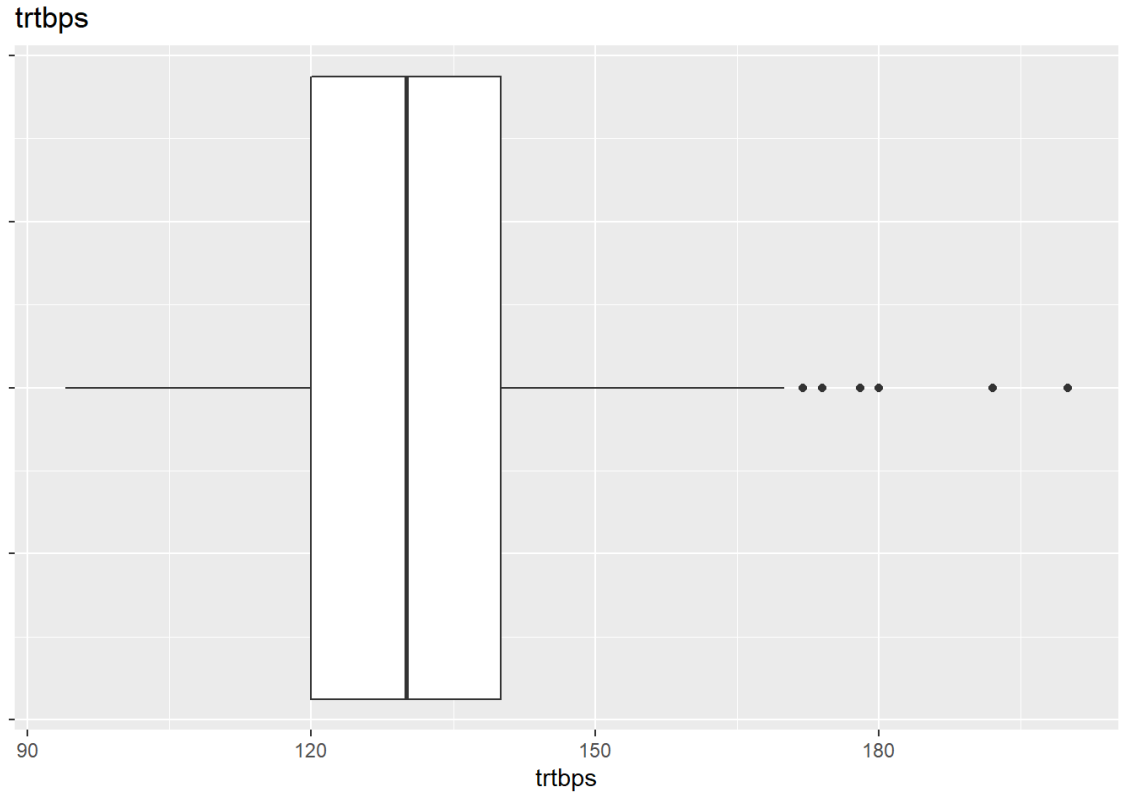
chol



oldpeak







En aquesta primera exploració, hem detectat ja que les variables trtbps (pressió sanguínia en repòs), chol(colesterol en sang), thalachh (ritme cardíac màxim) i oldpeak (depressió segment ST a l'electrocardiograma) tenen alguns outliers, ja que els punts que queden fora dels "bigotis" al gràfic són valors que estan per sota del primer quartil o per sobre del 3r quartil com a mínim 1,5 vegades el rang interquartílic.

En aquest cas, al tractar-se d'un tema mèdic sobre el qual no tenim coneixements previs, i per tant no sabem el significat d'aquests valors per a aquestes variables, nosaltres no podem determinar si el valor atípic es deu a un error del registre que s'hauria d'eliminar o a un cas real a tenir en compte. De cara a la comparació dels grups de dades del nostre anàlisi en el proper punt, es considera més raonable eliminar-los del dataset, ja que són pocs registres i no tindran un impacte significatiu en el resultat final. Un cop eliminats els registres del conjunt de dades, ens queden 294 observacions.

4. Anàlisi de les dades

4.1. Selecció de grups

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?)

Podem analitzar la variable objectiu des de diverses perspectives, comparant per sexe o per grup d'edat, i també mirant, per exemple, si la pressió sanguínia o el colesterol en sang del pacient tenen alguna influència a l'hora de desenvolupar malalties cardíaques. Com que la variable Age és una variable numèrica, farem una operació de discretització per tal de convertir-la en categòrica i poder comparar entre grups d'edat. Aquests són els tipus d'anàlisi que s'aplicaran per a realitzar les nostres comparacions:

-Per analitzar la variable objectiu 'output', que determina el nivell de risc de patir una malaltia cardíaca, segons el sexe, podem aplicar un test chi-squared.

-Per analitzar si hi ha diferències significatives en la variable pressió sanguínia (trtbps) per grups d'edat, aplicarem un test d'anàlisi de la variancia ANOVA.

-Per analitzar si la variable numèrica contínua chol té una relació amb la variable objectiu, que és dicotòmica, s'hauria de veure primer si la variable dependent té una distribució normal, i, en cas afirmatiu podríem aplicar el test estadístic t-student, que serviria per comparar si les mitjanes del nivell de colesterol són significativamenthi ha diferents en el grup que té el risc de malaltia cardíaca baix i en el de risc alt.

4.2. Comprovació de la normalitat i homogeneïtat de la variància

S'han aplicat tests de Kolmogorov-Smirnov i de Shapiro-Wilk a cadascuna de les variables numèriques que s'han fet servir.

Per la variable chol hem obtingut un valor p que ens permet assumir tant la seva normalitat com la seva homoscedasticitat, però de la variable de la pressió sanguínia només en podem assumir l'homoscedasticitat.

4.3. Proves estadístiques

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.

En primer lloc, volem comprovar si hi ha diferències significatives en el risc de malaltia cardíaca entre sexes, és a dir comparar diferències en una variable categòrica entre grups definits per una variable categòrica,i per tant s'aplica el test khi quadrat. Els resultats del test ens mostren un valor p de 6.018e-07, molt per sota del nivell de significació establert en 0.05. Per tant, podem rebutjar la hipòtesi nul·la i concloure que el parell de variables avaluat té una associació estadísticament significativa, és a dir, que la diferència en el risc de patir malalties cardiovasculars entre sexes és significativa.

```
chisq.test(sex_output_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sex_output_table
## X-squared = 24.907, df = 1, p-value = 6.018e-07
```


```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sex_output_table
## X-squared = 24.907, df = 1, p-value = 6.018e-07
```


En segon lloc fem servir un test d'anàlisi de la variància per analitzar les diferències de la pressió sanguínia entre grups d'edat. Amb aquest test ANOVA comparem les variàncies de la variable numèrica trtbps entre cada grup d'edat de la variable age_group.

```
heart2.aov <- aov(trtbps ~ age_group, data = heart2)
summary(heart2.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## age_group     3   6527  2175.6   8.275 2.67e-05 ***
## Residuals   290  76249   262.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observant els resultats del test a la taula, podem veure que el valor p és inferior a 0.05 i per tant podem concloure que la diferència en la pressió sanguínia entre grups és estadísticament significativa.

Per últim, tenint en compte que podem assumir la normalitat de la variable 'chol', aplicarem un test t-student per comprovar si té una influència significativa en el risc de desenvolupar malalties cardíaques. El valor p que s'obté és major que el nivell de significació del 5%, per tant amb les dades disponibles no s'observa que hi hagi diferències estadísticament significatives entre els pacients amb risc alt i els pacients amb risc baix pel que fa al nivell de colesterol.

```
t.test(chol ~ output, data = heart2)
```

```
##
##  Welch Two Sample t-test
##
## data:  chol by output
## t = -1.8671, df = 277.37, p-value = 0.06294
## alternative hypothesis: true difference in means between group Risc alt and group Risc baix is not equal to 0
## 95 percent confidence interval:
##  -20.5729016   0.5440532
## sample estimates:
##  mean in group Risc alt mean in group Risc baix
##                239.1840                249.1985
```

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.

En l'anàlisi exploratori inicial del conjunt de dades, hem generat diverses visualitzacions per poder observar diferents aspectes de les variables originals. Hem creat gràfics de barres de les variables categòriques per analitzar visualment les seves freqüències, i histogrames i diagrames de caixa de les variables numèriques per veure la seva distribució i variància.

Les imatges generades es poden trobar a la carpeta 'images' del repositori GitHub de la pràctica.

6. Resolució del problema

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

A partir de les proves estadístiques realitzades, algunes de les conclusions que podem extreure són que hi ha diferències significatives entre sexes pel que fa al risc de desenvolupar malalties cardíaques, o que la pressió sanguínia és variable entre els grups d'edat.

En canvi, sorprenentment, pel que fa a la relació entre el nivell de colesterol i el risc de malaltia, no hem obtingut resultats que ens permetin establir que existeix una diferència significativa.

Aquestes conclusions poden permetre tenir un enfocament més adequat alhora de tractar o investigar les malalties del cor i millorar la seva prevenció.

7. Codi

Cal adjuntar el codi, preferiblement en R, amb el que s' ha realitzat la neteja, analisi i representacio de les dades. Si ho preferiu, tambe podeu treballar en Python.

El codi en R utilitzat per a la realització de la pràctica es pot trobar en format html i rmd a la carpeta source del repositori GitHub de la pràctica.

Contribucions	Signatura
Investigació prèvia	PCP
Redacció de les respostes	PCP
Desenvolupament del codi	PCP
Participació al vídeo	PCP