PRAC 2: Tipologia i cicle de vida de les dades: Neteja i Validació de les dades.  
Autor: Ivan Puig

Data: 16/12/2018

*Link:* [*https://github.com/ipuiglob/Data\_Cleansing*](https://github.com/ipuiglob/Data_Cleansing)

*Wiki:* [*https://github.com/ipuiglob/Data\_Cleansing/wiki/Home-of-Data-Cleansing*](https://github.com/ipuiglob/Data_Cleansing/wiki/Home-of-Data-Cleansing)

Descripció de la pràctica:

L’objectiu d’aquesta activitat serà el tractament d’un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com).

Seguint les principals etapes d’un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

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

L’objectiu de la pràctica és identificar i analitzar quines eren les variables o factors a tenir en compte i que varen ser crucials per a la supervivència o no dels passatgers del Titànic i la seva tripulació.

El fet de viatjar a tercera classe era un factor determinat per a la supervivència? I el ser dona de segona classe? O tripulant? Són totes aquestes preguntes les que mirarem de donar resposta analitzant el corresponent dataset.

Evidentment, l’estudi i anàlisis del dataset no farà tornar a la vida a tots els essers humans que varen morir aquella nit del 14 d’abril de 1912, però si ens donarà un coneixement i ens permetrà establir un perfil del supervivent d’aquella tragèdia.

Per començar donarem un cop d’ull al conjunt de dades descarregat de la pàgina de Kaggle. Es tracta d’un fitxer amb extensió CSV i editant-lo podem extreure la següent informació:

Camps del dataset

|  |  |  |
| --- | --- | --- |
| Variable | Definició | Valors |
| survival | Supervivent de la tragèdia | 0 = No, 1 = Sí |
| pclass | Classe del bitllet del passatger. | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sexe del passatger |  |
| Age | Edat del passatger |  |
| sibsp | # de germans / esposes a bord del Titànic |  |
| parch | # de pares / fills a bord del Titànic |  |
| ticket | Numero del bitllet del passatge |  |
| fare | Tarifa |  |
| cabin | Numero de la cabina |  |
| embarked | Port embarcament | C = Cherbourg,  Q = Queenstown,  S = Southampton |

1. **Integració i selecció de les dades d’interès a analitzar.**

Com ja hem comentat l’objectiu de la pràctica es identificar un perfil del supervivent; Analitzar quines eren les variables que més influïen en la supervivència o no d’un passatger. En definitiva, establir una classificació que, de la manera més genèrica possible, i tenint en compte algunes de les variables recollides en el dataset, ens permetés establir una classificació per grups de passatgers. Per a dur a terme aquesta classificació aplicarem un model de classificació i veurem quins grups de passatgers tenien més oportunitats de sobreviure i quins no.

Es per això que del conjunt total de les dades inicials, hem de extreure aquelles que no ens aportaran cap mena d’informació rellevant i quedar-nos amb les que si poden tenir un paper important dins de la anàlisis. Per exemple, el nombre de la cabina o el port on va embarcar no ens ajudarà en aquest sentit, igual que les relacions de parentesc entre d’altres.

És per aquest motiu que ens hem de centrar en les variables qualitatives i que a la llarga podem utilitzar per agrupar grups d’individus. Les variables que utilitzarem per a construir el model seran la “classe”, “l’edat” i “sexe” en una primera instancia, per a més endavant afegir d’altres variables al nostre estudi si fos necessari. També utilitzarem per al nostre model la variable “Survived” com a variable classificatòria o també anomenada discriminant i necessària per a la construcció d’arbre de classificació.

1. **Neteja de les dades.**

Donem un primer com d’ull al nostre dataset (train.csv). Recordem que utilitzarem R per a la construcció del nostre model.

Obrirem R i carregarem el conjunt de dades Titànic al nostre espai de treball (workspace). Per fer-ho utilitzarem la funció en R “read.csv()” especificant la ruta d’accés i el nom del fitxer:

> titanic <- read.csv(file="C:/Temp/train.csv",head=TRUE,sep=",")

*\*El paràmetre head especifica que el dataset incorpora a la primera fila el nom dels camps i el paràmetre sep el caràcter separador que delimita els diferents camps.*

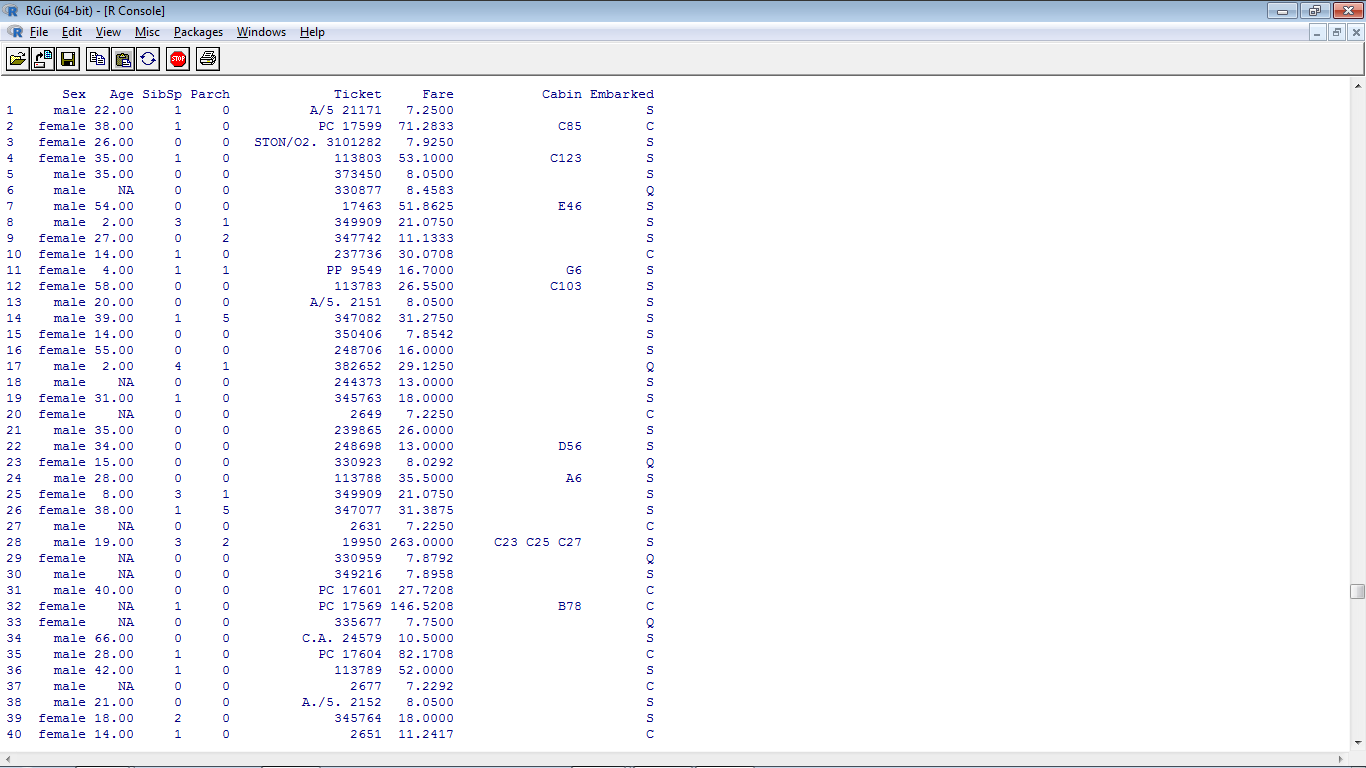
Una vegada executada la comanda ja podem fer una primera aproximació a les dades llistant el contingut de la variable Titànic:

> titanic

O

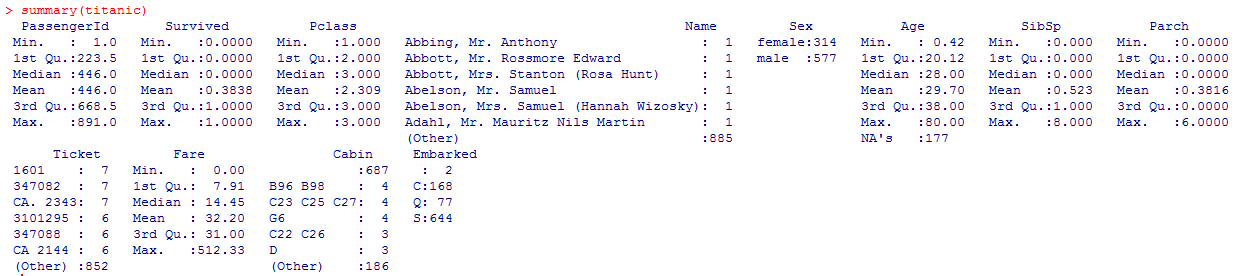
> Head(titanic)





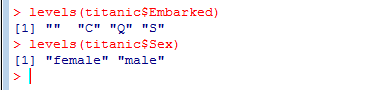
Recordem que podem utilitzar algunes comandes addicionals de R que ens facilitaran la comprensió de les dades i com s’agrupen. Per exemple, la comanda “summary()”, o comandes com “str()” o “lapply()”.

Per exemple, veure un resum de les dades contingudes al dataset amb “summary()”:



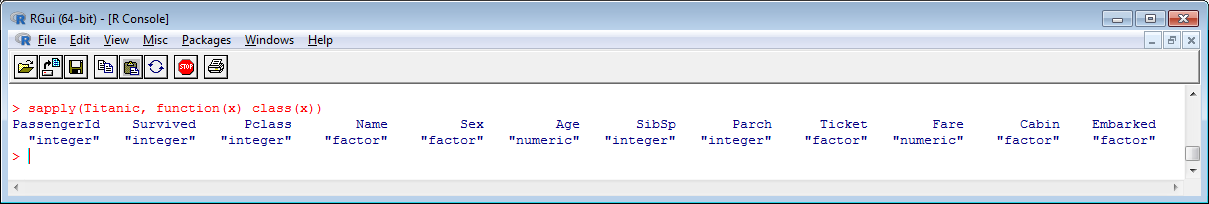
O aplicant la funció “levels()” que ens permet veure els diferents valors d’un atribut.

> levels(Titanic$Embarked) o > levels(Titanic$Sex)



Però el que realment ens interessa és conèixer els tipus de dades assignats a cada camp del nostre dataset. Per fer-ho executarem la comanda “sapply()”:

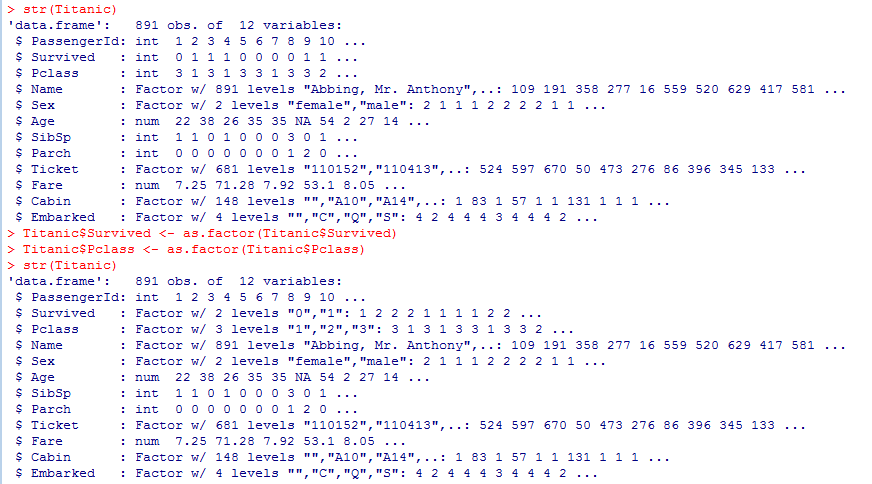
> sapply(titanic, function(x) class(x))



Arribats a aquest punt, és important també quan estem davant d’un dataset, poder modificar els tipus assignats a les variables. En moltes ocasions ens pot interessar que un valor que inicialment ha estat identificat com a numèric sigui transformat a un altre tipus.

En el nostre cas particular es dona en un parell de camps com són el camp “classe” (1,2,3) o el camp “supervivent” (0 o 1). Aquest dos camps són identificats com a enters inicialment, però és evident que estem davant de camps o variables categòriques i que per tant cal que modifiquem. Per a fer-ho utilitzarem instruccions com:

> levels(titanic$Pclass)



Creiem que es una factor a tenir en consideració dins del tractament inicial de les dades i el seu posterior anàlisis donat que en funció del tupis de dada podrem realitzar unes operacions o unes altres.

* 1. **Les dades contenen zeros o elements buits? Com gestionaries aquests casos?**

Identifiquem si tenim valors nuls o no informats al nostre dataset amb les comandes de R següents:

> sapply(titanic, function(x) sum(is.na(x)))



Veiem que per al camp edat (Age) tenim 177 valors a nul. En aquets sentit, si decidíssim eliminar els registres estaríem desaprofitant informació i la precisió del nostre estudi es veuria afectada. Per altra banda podríem mirar d’establir un valor mitjà (mitjana d’edat) en funció del conjunt total de les dades, però al igual que en cas anterior estaríem desvirtuant les dades del conjunt.

Tot i això, en la comunitat de Kaggle[[1]](#footnote-1) s’ofereixen diferents alternatives a aquest problema, i una de les que m’ha semblat força interessant és la que es comenta a continuació:

library(mice)

init = mice(titanic, maxit=0)

predM = init$predictorMatrix

predM[, c("Pclass”,"Sex")]=0

imp<-mice(titanic, m=5, predictorMatrix = predM)

titanic <- complete(imp)

View(titanic)

El que fa aquest conjunt d’instruccions (concretament les funcions de la llibreria **mice**) és deduir, basant-se en d’altres variables del dataset, el possible valor faltant. Per exemple, podríem mirar d’obtenir la edat en funció de les variables Classe o Sexe, o qualsevol altre del dataset per exemple (força absurd en el nostre cas però mol útil per a altres exemple).

Per al cas que ens ocupa el que farem finalment es no tenir en compte els valors nuls alhora de fer els càlculs de mitjanes o de normalitat i variància. Per exemple, com veurem més endavant la mitjana serà calculada sense tenir en compte els valors nuls amb la instrucció següent: > mean(titanic\_net$Age, na.rm=TRUE)

* 1. **Identificació i tractament de valors extrems.**

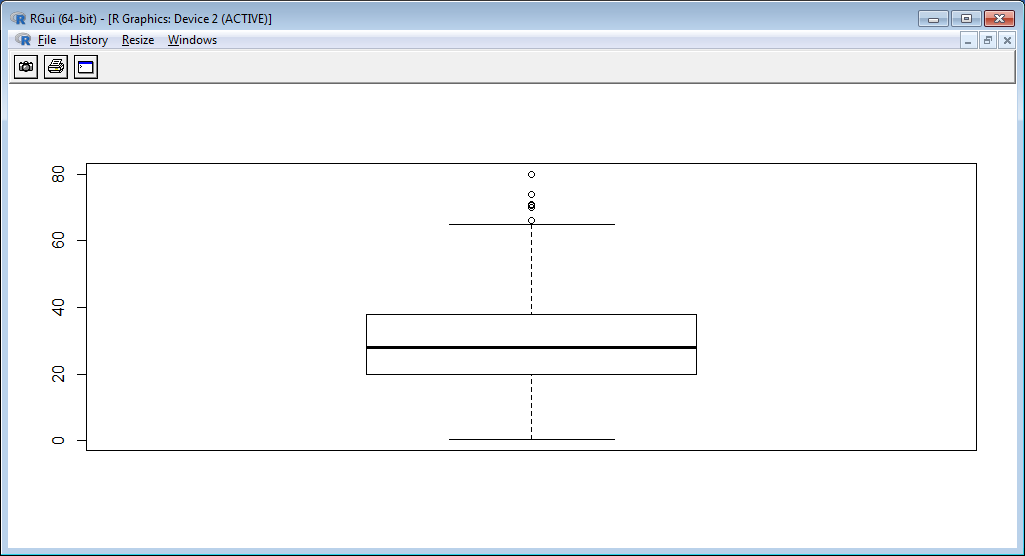
El que farem tot seguit és identificar dins del dataset els valors que poden ser extrems i que poden esbiaixar els resultats del nostre anàlisis posterior.

Per a fer-ho, podem fer representacions gràfiques de caixes o mitjançant funcions com la “boxplot.stats()”:

> boxplot.stats(titanic$Age)$out



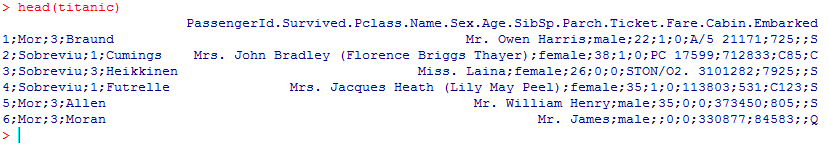
O de forma gràfica: > boxplot(titanic$Age)



Com podem observar, no tenim valors extrems donat que tenir passatgers amb 70 anys entre dins de lo normal tot i la època. Podem aplicar el mateix principi a la resta de camps del nostre dataset.

***\*Nota de l’autor****: En el cas del dataset del Titànic, tenim un cert nombre de variables del conjunt que són categòriques (Pclass, Sex, etc) i per tant tan no els hi podem aplicar aquest tipus de tractament. Tampoc ho podem fer amb variables del tipus Cabin o Ticket donat que són valors discrets i no continus.*

Per últim no voldria deixar de comentar que pot ser de gran ajuda re-nombrar els valors de certs valors del dataset per a fer-los més entenedors. Per exemple, el camp Supervivent conté un 0 si el passatger va morir i un 1 si va sobreviure. Creiem que podria facilitar la feina fer alguna tasca de conversió/transformació sobre el camp i en comptes d’utilitzar nombres enters (0 o 1) poguéssim fer referencia a “sobreviu” o “mor” (text).



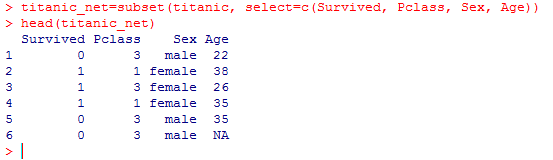
1. **Anàlisi de les dades.**
   1. **Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).**

Com ja hem comentat en apartats anteriors, no utilitzarem tots els camps del dataset. Els identificadors com per exemple, el nom de la persona o el número de bitllet no són significatius per al nostre estudi. Si en canvi l’edat per exemple o la classe en la que viatjava el passatger.

Per tant, per a la construcció del nostre model usarem els camps “classe”, “l’edat” i “sexe”. També utilitzarem per al nostre model la variable “Survived” com a variable classificatòria o també anomenada discriminant.

Utilitzarem la funció “subset()” per a crear un nou dataset amb els atributs i les dades que necessitem; Aquest nou dataset serà anomenat “titanic\_net” i és el que utilitzarem per a fer les següents operacions:

> titanic\_net=subset(titanic, select=c(Survived, Pclass, Sex, Age))



*\*****Nota de l’autor****: En aquest punt cal també fer referencia a la utilitat de treballar amb subconjunts del conjunt principal. D’aquesta manera reduïm l’scope del problema i podem focalitzar-nos en el que realment pot ser d’interès.*

*La creació d’aquests subsets no esta tan sols destinada a reduir el nombre d’atributs, també podem utilitzar les funcions per a extreure dades que compleixin una determinada condició.*

> titanic\_homes=subset(titanic, select=c(Survived, Pclass, Sex, Age), Sex==male)

*o*

> titanic\_dones1=subset(titanic, select=c(Survived, Pclass, Sex, Age), Sex==female & Pclass==1)

* 1. **Comprovació de la normalitat i homogeneïtat de la variància.**

Comencem per calcular la mitjana i la desviació estàndard per després comprovar i representar la normalitat. Les funcions a utilitzar seran “mean()”, “sd()”, i “rnorm()” respectivament.

> Mean(titanic\_net$Age, na.rm=TRUE)

> sd(titanic\_net$Age, na.rm=TRUE)



O també altres càlculs d’interès com “la mediana” o el rang de valors d’un atribut:

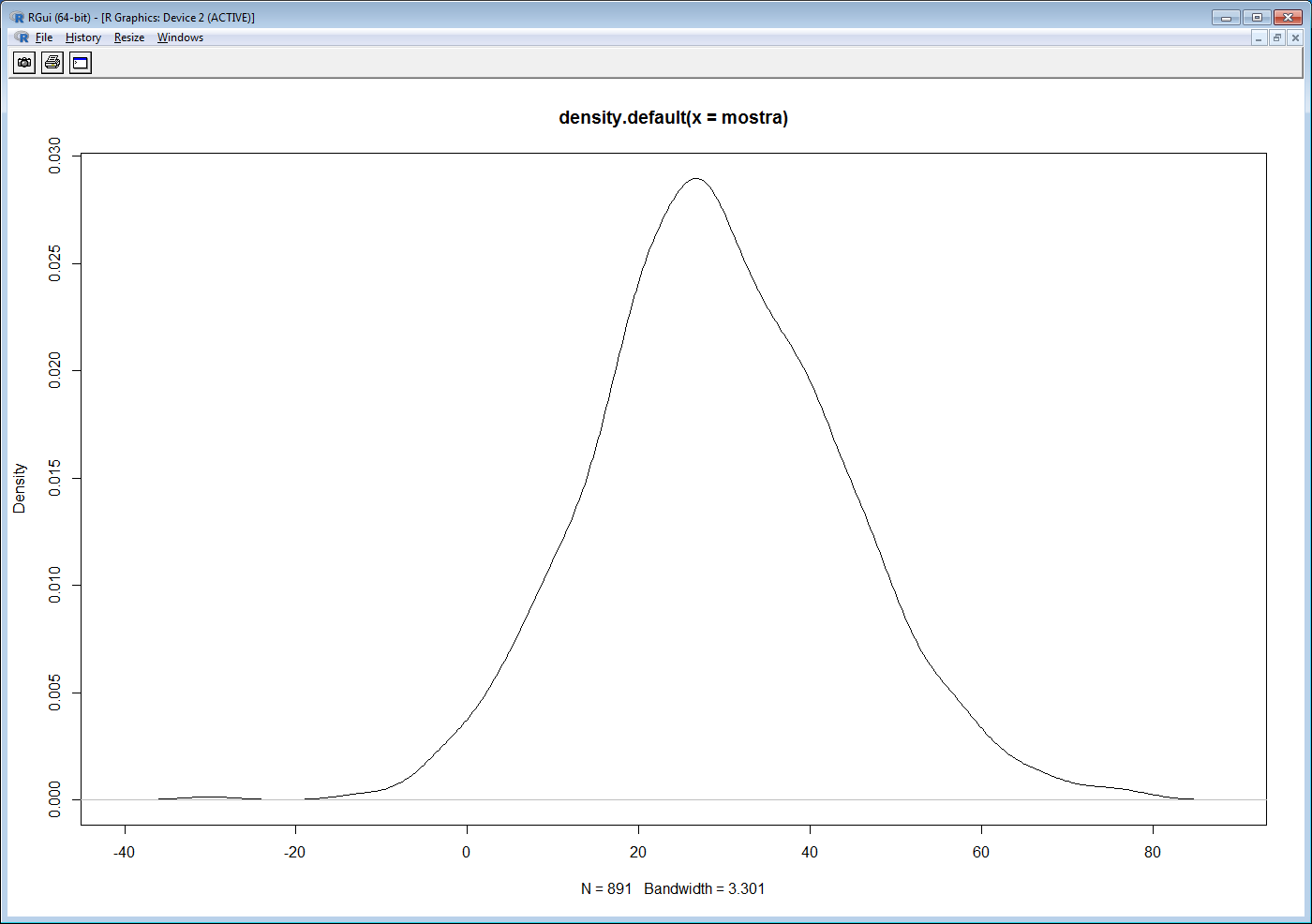
> median(titanic\_net$Age, na.rm=TRUE)

> range(titanic\_net$Age, na.rm=TRUE)

Les funcions són moltes i variades però passem finalment a representar la normalitat amb la funció “rnorm()”:

> mostra <- rnorm(titanic\_net$Age, mean(titanic\_net$Age, na.rm=TRUE), sd(titanic\_net$Age, na.rm=TRUE))



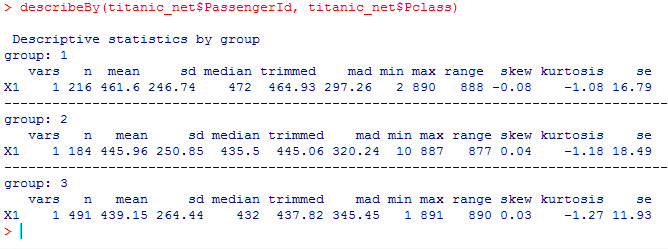


Com podem veure la densitat es força simètrica, centrada sobre el valor 27 que coincideix pràcticament amb la mitjana (29).

Passem ara a calcular la homogeneïtat de la variància; Utilitzarem les mateixes variables, edat i classe, i veiem si existeix alguna uniformitat en les dades per aquests dos atributs:

*\*****Nota de l’autor****: per a la utilització de la funció “describeBy()” cal prèviament instal·lar la llibreria “psych”*

> describeBy(titanic\_net$Age, titanic\_net$Pclass)

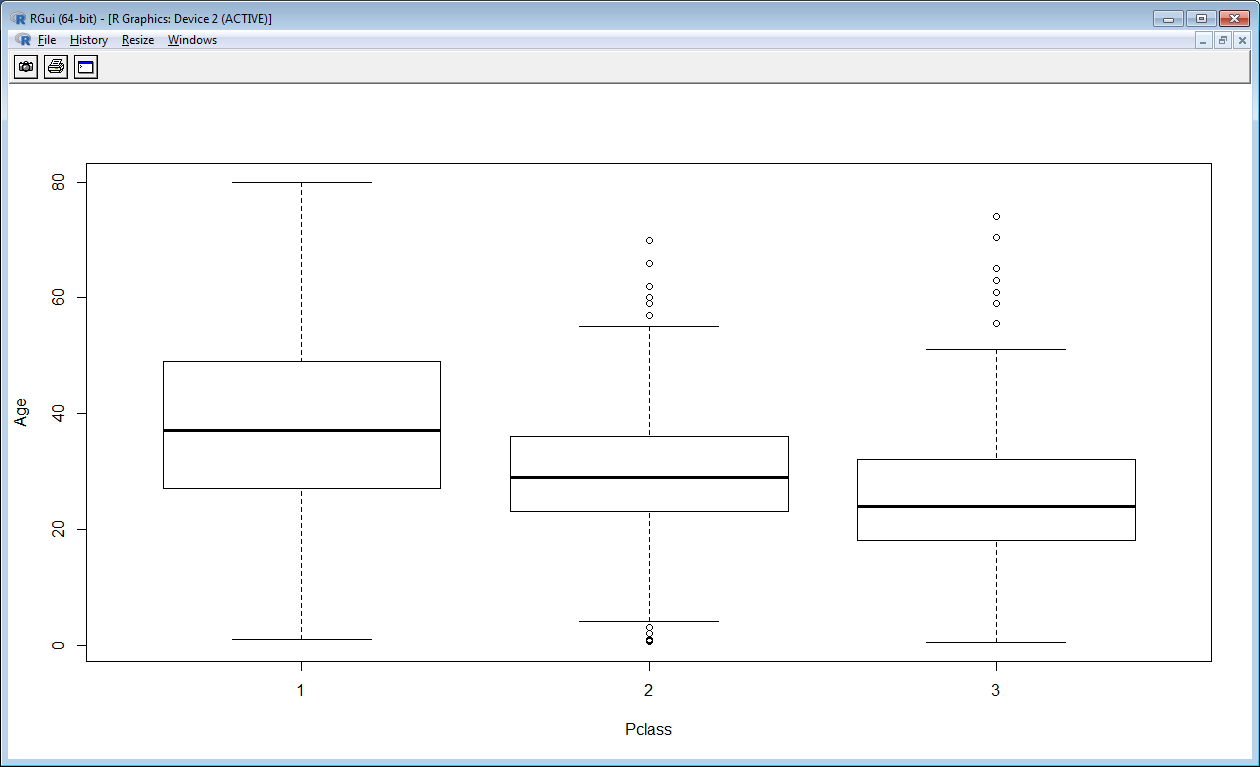


Per a evitar la pèrdua d’informació crearem un nou subset (titanic\_class) amb les dades necessàries per a representar la homogeneïtat de la variància.

> titanic\_class <- subset(titanic\_net, Pclass==1 | Pclass==2 | Pclass==3)

Tot seguit representem visualment la informació mitjançant un gràfic de caixes.

> plot(Age ~ Pclass, data = titanic\_class)

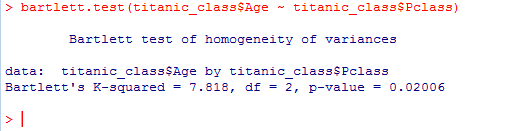


L’anàlisi de la gràfica ja ens indica que hi ha certa homogeneïtat en la edat dels passatger per a cadascuna de les classes. Sí és cert que els passatgers de primera classe, la seva edat mitjana era més alta, però en conjunt estem davant d’un grup força homogeni i amb un valors extrems dins de la normalitat. En definitiva, estem dins d’una distribució normal.

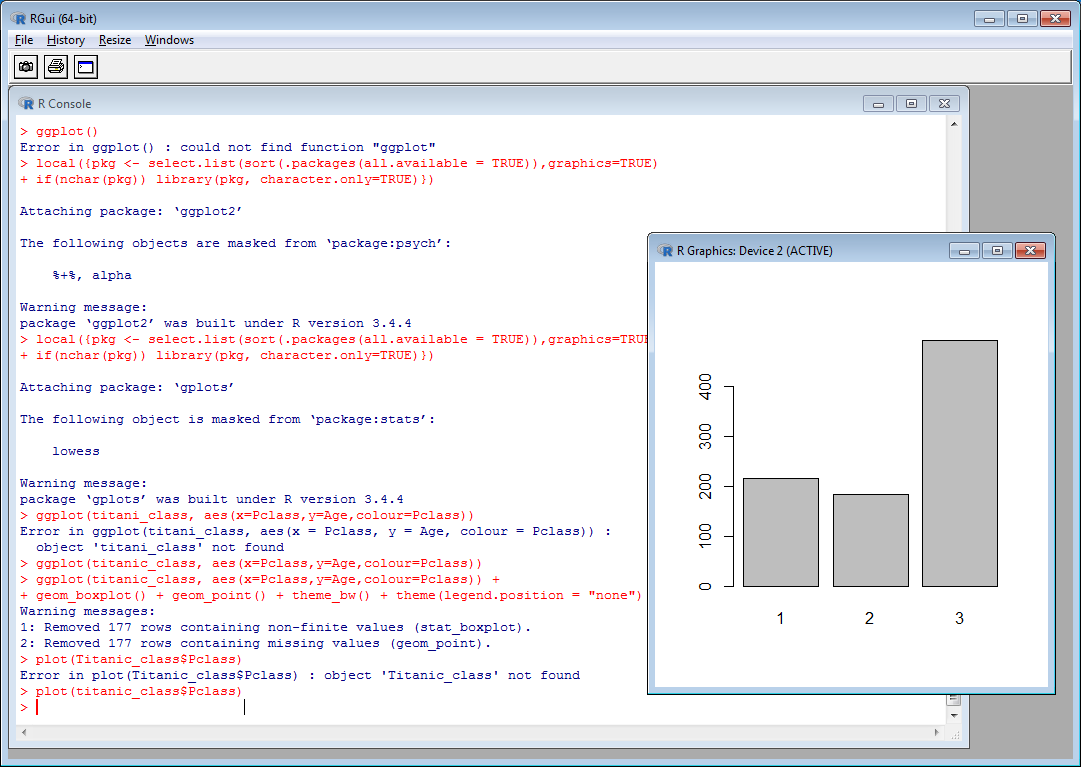
És ara quan hauríem d’aplicar per exemple el test de Bartlet o el F-test.

Apliquem tot segui el test de **Bartlett**:

> bartlett.test(titanic\_class$Age ~ titanic\_class$Pclass)

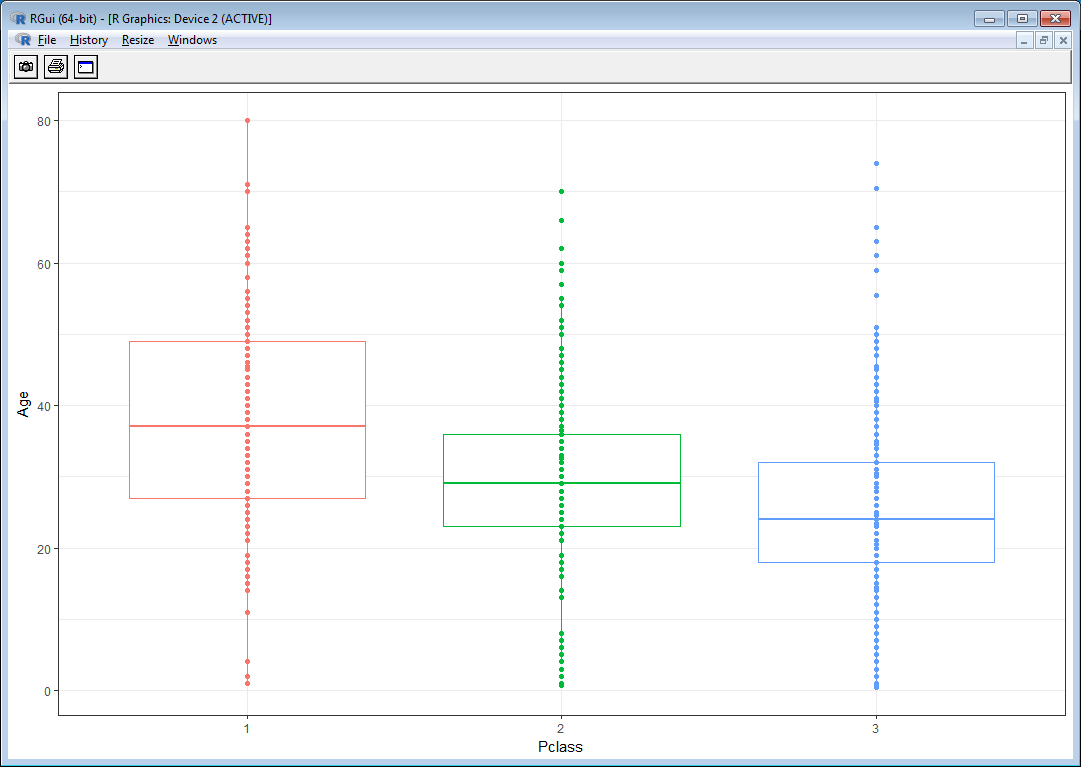


El test no troba diferencies significatives tot i que hem de tenir en compte la mida dels diferents grups (passatgers per cada classe) amb els que treballem; recordem de nou que el nombre de passatgers de tercera és pràcticament el doble dels de primera i segona junts.



Per últim representarem les dades de nou amb les funcions de la llibreria “ggplot”:

> ggplot(titanic\_class, aes(x=Pclass,y=Age,colour=Pclass)) + geom\_boxplot() + geom\_point() + theme\_bw() + theme(legend.position = "none")



* 1. **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.**

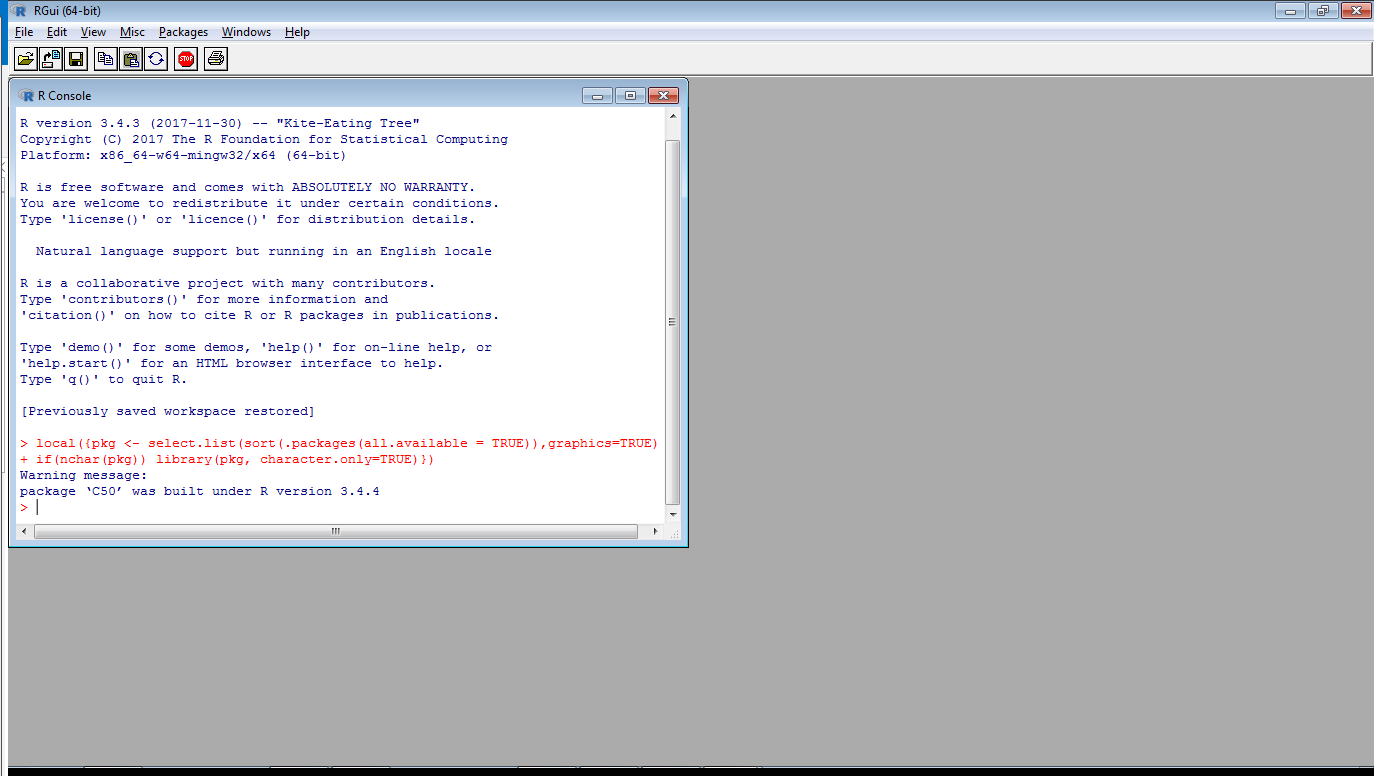
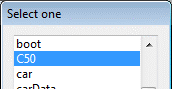
Com ja hem comentat al llarg de la pràctica, l’objectiu final és determinar quines variable eren crucials o no per a la supervivència a la tragèdia del Titànic i poder així obtenir uns perfils de passatger (o classificació) amb més oportunitats de sobreviure.

Saber per exemple si, un passatger home que viatjava a tercera classe tenia més possibilitats de sobreviure que un de primera, o una dona de segona classe de mitjana edat també ho era, és l’objectiu d’aquest anàlisis.

Les variables que utilitzarem per a construir el model de classificació seran la classe, l’edat i sexe. I utilitzarem la variable “Survived” com a variable classificatòria o també anomenada discriminant.

Estem fent referencia a un model de classificació i per tant tenim diferents opcions a R per a obtenir models classificatoris, però ens decidirem per la utilització del paquet C5.0 que és possiblement el més estes per a aquest tipus d’anàlisis de dades i que a més permet per exemple fer poda de l’arbre de classificació resultant.

El primer que caldrà fer és carregar la llibreria corresponent. Per fer-ho seleccionarem del menú “Packages” l’opció “Load Package” i escollirem de la llista l’opció C50.



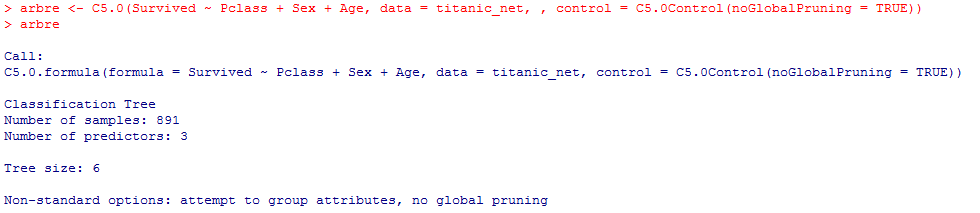
Dins del paquet C5.0 tenim disponibles tot un seguit de funcions que ens permeten generar un arbre de classificació. En el nostre cas particular aplicarem la funció C5.0 basant-nos en les variable classe, sexe i edat com ja hem comentat abans.

***\*Nota de l’autor****: El camp “Survided” cal transformar-lo de tipus enter a factor donat que el valor 0 correspon a No-Supervivent i el valor a 1 a Supervivent. Per fer aquest pas aplicarem la comanda següent:*

> titanic\_net$Survived <- as.factor(titanic\_net$Survived)

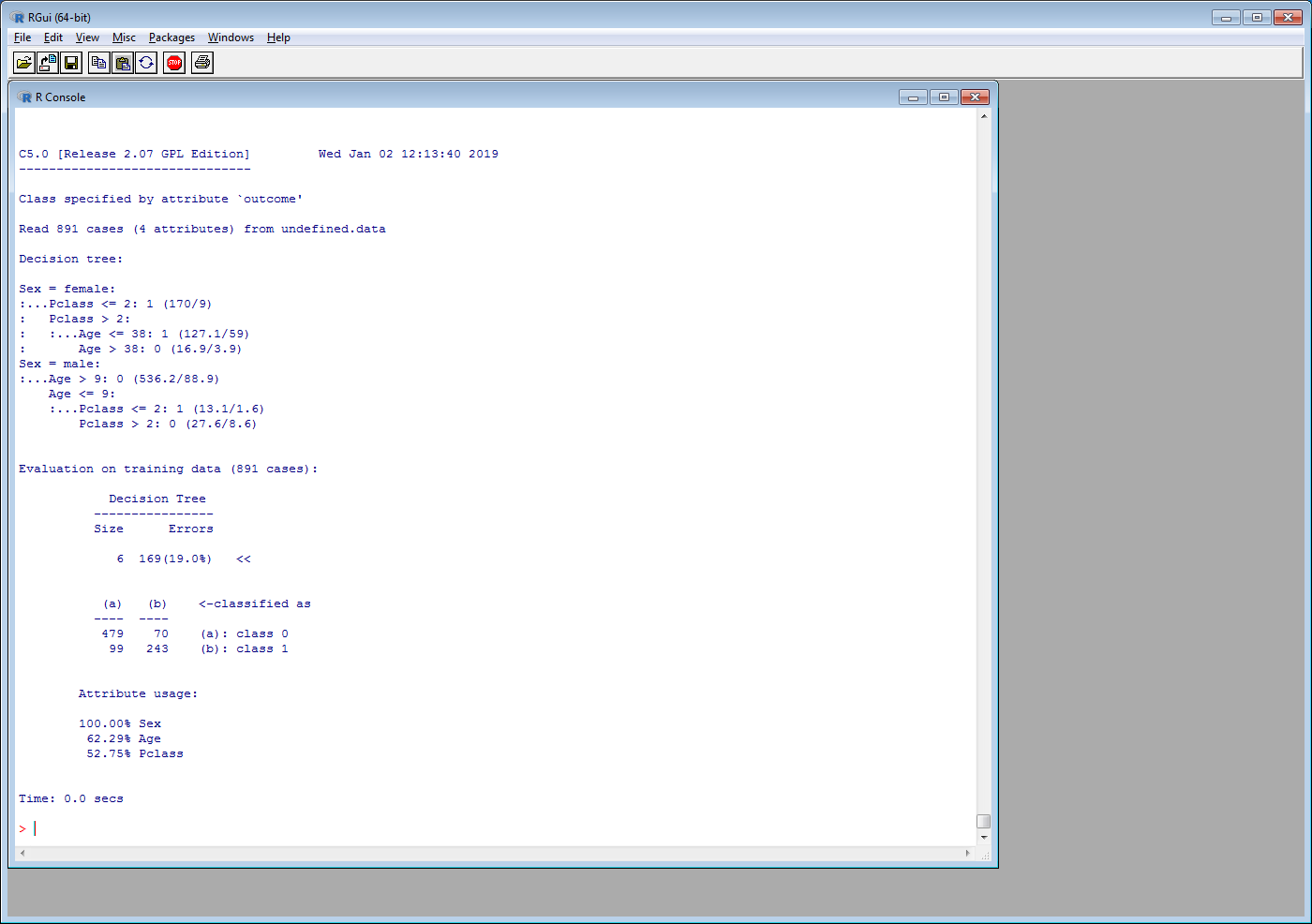
La funció en R queda de la següent manera:

> arbre <- C5.0(Survived ~ Pclass + Sex + Age, data = titanic\_net, control = C5.0Control(noGlobalPruning = TRUE))  
> arbre



Una vegada tenim l’arbre de classificació construït ja podem iniciar el nostre anàlisis. Per fer-ho podem començar per veure el pes de les variables i alguns indicadors més addicionals amb la funció “summary()”

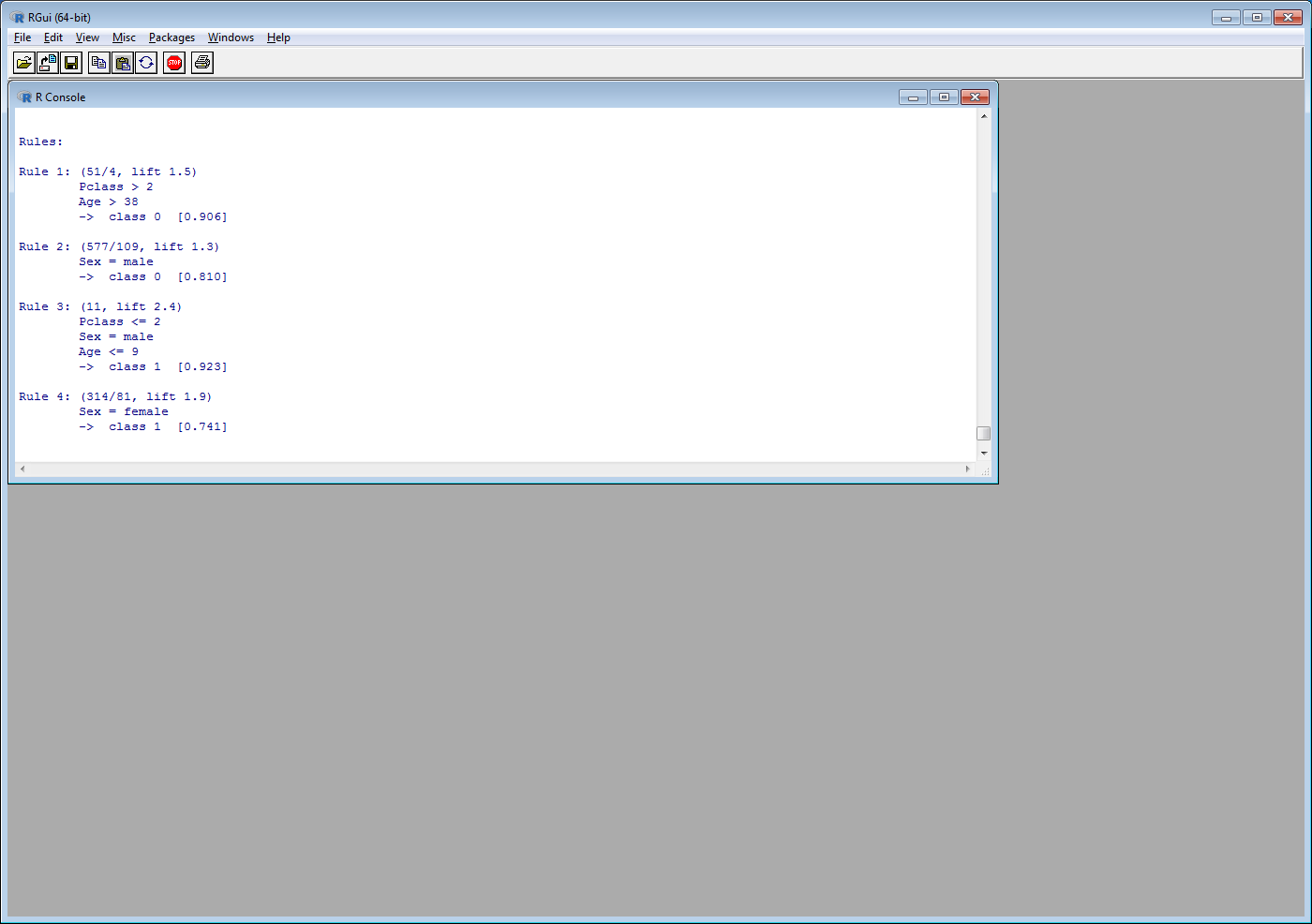
> summary(arbre)



Veiem entre d’altres dades referents als registres analitzats (891 casos), la tassa d’error (19%), i fins i tot la descripció de les regles. Però centrem-nos-hi en el que de veritat ens interessa per al nostre anàlisis, és a dir, quines son les variables amb més pes del model i que han estat decisives per a la classificació. Com podem veure, l’atribut Sexe té un pes del 100% en el model de classificació, seguit per l’atribut Edat (50%) i finalment per l’atribut Classe (17%).

També es força interessant per a la anàlisis poder extreure les regles de classificació que han estat utilitzades per a la construcció del mode de classificació. Les comandes a utilitzar seran:

> arbre <- C5.0(Survived ~ Pclass + Sex + Age, data = titanic\_net, control = C5.0Control(noGlobalPruning = TRUE),rules = TRUE)  
> summary(arbre)



Com podem veure del resultat en pantalla una regla ens diu que si el passatger era dona la probabilitat de sobreviure era molt major (del 74%), per exemple; O que si eres un nen menor de 10 anys i viatjaves a segona o tercera classe tenies una probabilitat del 92% de sobreviure. Totes aquestes dades i el seu tractament són les que realment ens aporta informació dins del nostre conjunt de dades i que cal saber identificar pel seu valor afegit. Dades com el número de cabina o el preu del bitllet creiem que poden ser discriminats pel seu poc valor afegit.

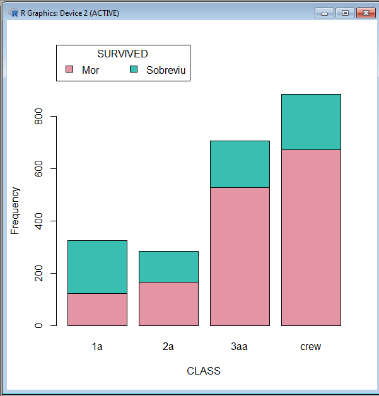
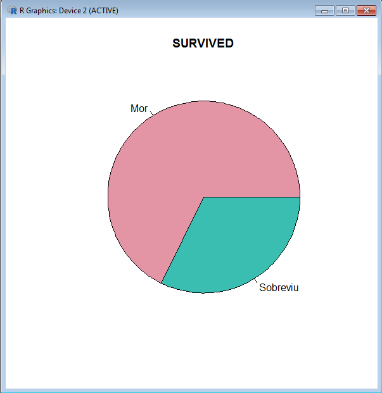
No entrarem gaire més en el model de classificació donat que el que realment ens interessa ja ho hem extret de la execució de les funcions anteriors i de la implementació de l’arbre sobre el nostre dataset.

Hem estat capaços de determinar quines són les variables “d’interès” i en quina proporció influeixen en cas de estudi del Titànic. Podríem seguir realitzant operacions molt més especifiques donat que R ens ofereix un ventall molt alt d’eines i funcions per a la anàlisis de dades; però creiem que amb els exemples anteriors queda cobert l’objectiu principal de l’exercici que és la d’identificar les variables d’interès en un conjunt de dades o dataset i el seu posterior anàlisis aplicant eines estadístiques.

1. **Representació dels resultats a partir de taules i gràfiques.**

Una manera molt útil d’analitzar les dades és utilitzant la representació gràfica que proporciona R. Per exemple, podem representar de forma ràpida en un diagrama de pastis la proporció de morts i supervivents o, com a segon exemple de representació gràfica, la proporció de morts i supervivents per cadascuna de les classes en un format de columnes:

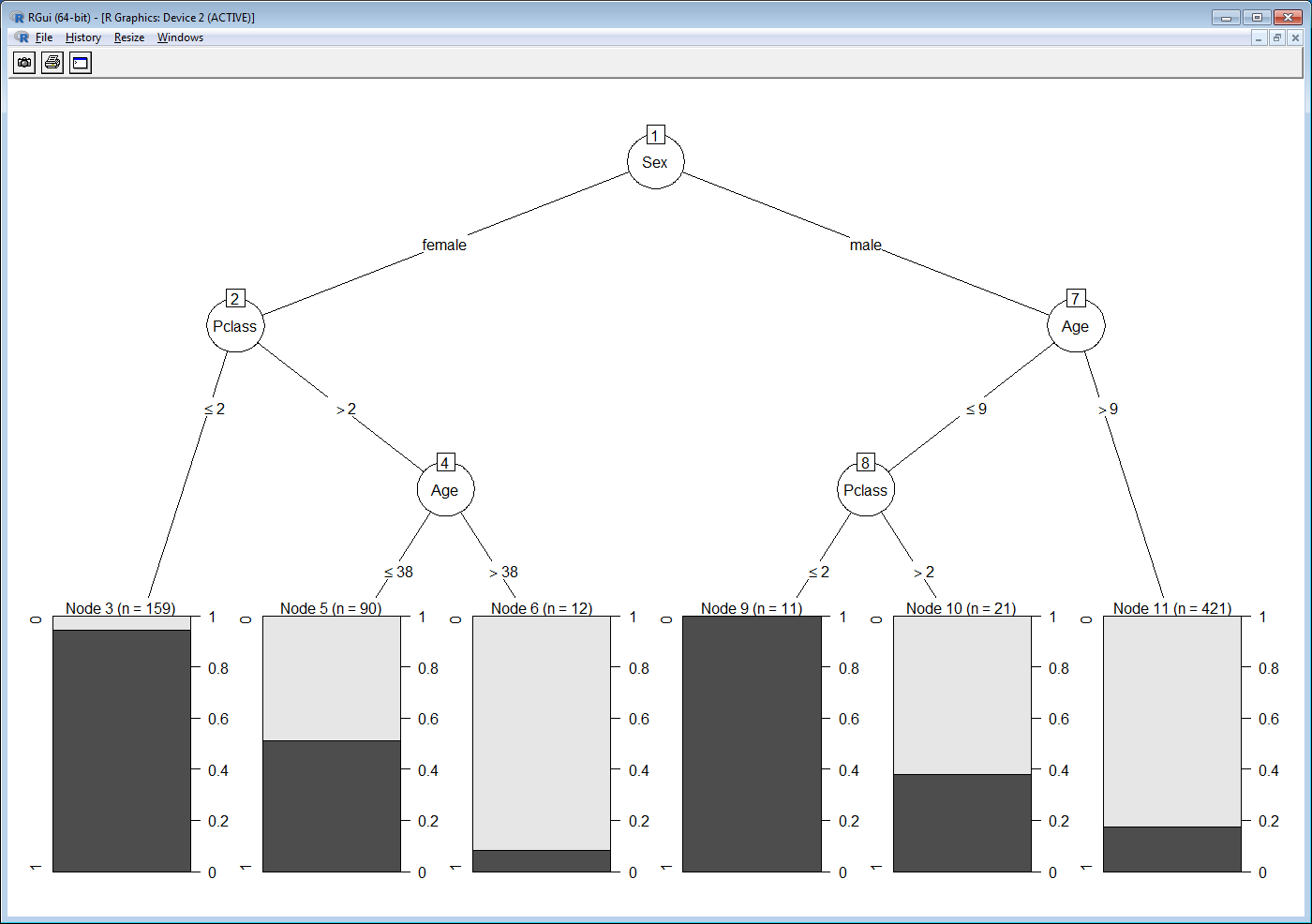
***\*Nota de l’autor****: En el cas de les dues gràfiques següents he utilitzat R-comander per a la representació de les dades donat que permet moltes més “customitzacions” alhora de representar les dades de forma visual (llegendes, colors, títols, etc.)*



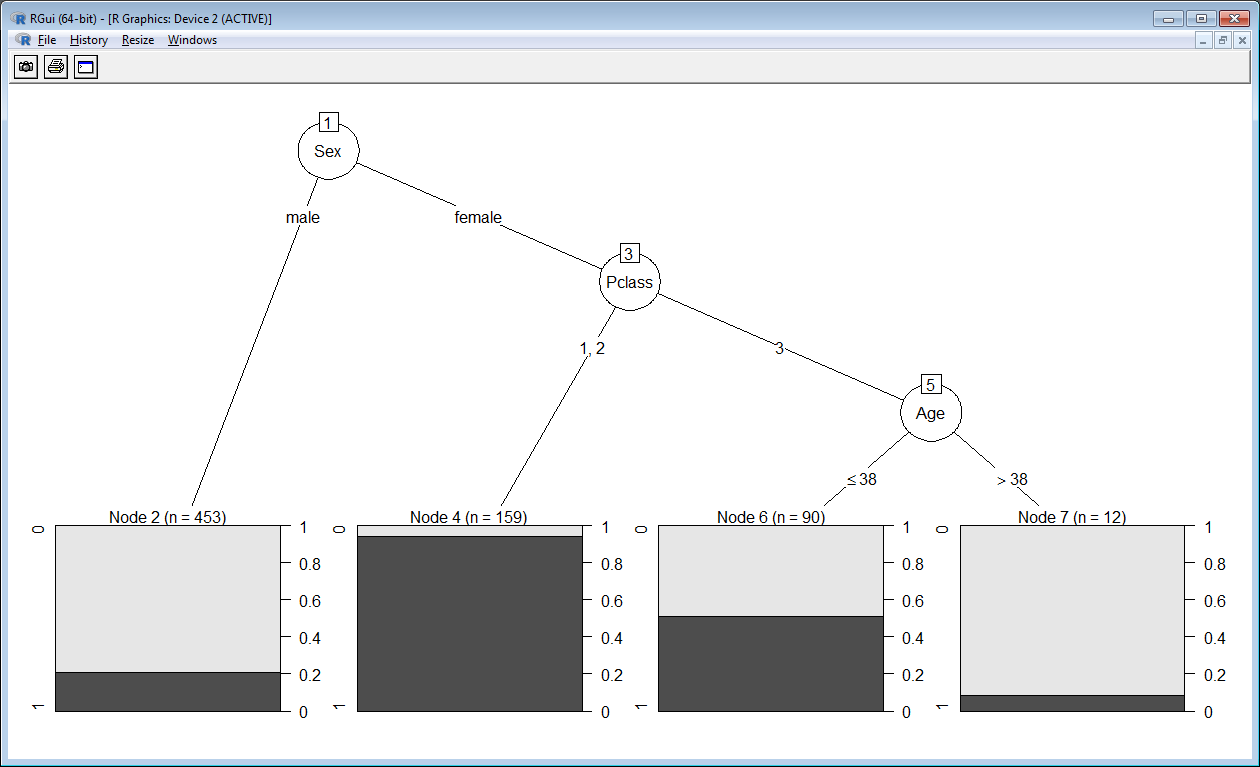
Un ràpid cop d’ull a les representacions ja ens permet establir un primer anàlisis i veure que en funció de la classe en la que es viatjava al Titànic ja era prou significatiu respecte a la supervivència o no supervivència a la tragèdia d’aquell passatger.

Per altra banda, i centrant-nos en el nostre cas particular de l’arbre de classificació, també som capaços de representar la informació obtinguda de la anàlisis anterior de forma visual.

La comanda a utilitzar serà: > plot(arbre)



O aplicant un mètode de poda a l’arbre de classificació:



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

Com ja hem anat analitzant en els diferents apartats de l’exercici i per no estendrem gaire més, el coneixement que traiem de la anàlisis de les dades i la construcció del model és que els factors a tenir en compte en la supervivència dels passatgers del Titànic son condicionades bàsicament pel sexe del passatger, la classe (si viatjava en 3ª classe, segona o primera) i la edat.

Per exemple, analitzant la representació gràfica de l’arbre obtingut veiem que pràcticament totes les dones de primera classe varen sobreviure, o que pràcticament tots els homes que viatjaven al Titànic varen morir.

L’estudi i el posterior anàlisis de les dades del conjunt Titànic ens han donat un major coneixement i ens ha permets establir certs patrons, que no és res més que l’objectiu final de l’anàlisi d’un conjunt de dades massives. És per tot això que creiem que donem resposta a les preguntes plantejades a l’exercici i hem identificat el valor real de les dades fent prèviament un treball de recollida, neteja i validació d’aquestes dades.

1. **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.**

Al llarg de l’exercici ja hem anat inclouen el codi utilitzat per a cada operació, de totes formes aquí recollim totes les instruccions executades:

*#Lectura del fitxer csv I càrrega de les dades a un dataframe*

> Titanic <- read.csv(file="C:/Temp/train.csv",head=TRUE,sep=",")

*#funcions per inspeccionar les dades*

> Titanic

> Summary(Titanic)

> levels(Titanic$Embarked) o > levels(Titanic$Sex)

> sapply(Titanic, function(x) class(x))

> sapply(titanic, function(x) sum(is.na(x)))

*#Funció que ens calcula la mitjana sense tenir en compte els valors nuls(NA)*

> mean(titanic\_net$Age, na.rm=TRUE)

*#identificació dels valors extrems*

> boxplot.stats(titanic$Age)$out

*#representació d’un diagrama de caixes de la variable edat*

> boxplot(titanic$Age)

*#Creació d’un subset del dataset principal*

> titanic\_net=subset(titanic, select=c(Survived, Pclass, Sex, Age))

*#Algunes funcions estadístiques*

> median(titanic\_net$Age, na.rm=TRUE)

> range(titanic\_net$Age, na.rm=TRUE)

*#Calcul de la normalitat*

> mostra <- rnorm(titanic\_net$Age, mean(titanic\_net$Age, na.rm=TRUE), sd(titanic\_net$Age, na.rm=TRUE))

*#Instrucció que converteix de numèric a factor un camp del dataset*

*> titanic\_net$Survived <- as.factor(titanic\_net$Survived)*

*#homogeneïtat de la variància*

*> describeBy(titanic\_net$Age, titanic\_net$Pclass)*

*> titanic\_class <- subset(titanic\_net, Pclass==1 | Pclass==2 | Pclass==3)*

*> plot(Age ~ Pclass, data = titanic\_class)*

*#Test de Bartlett:*

*> bartlett.test(titanic\_class$Age ~ titanic\_class$Pclass)*

*> ggplot(titanic\_class, aes(x=Pclass,y=Age,colour=Pclass)) + geom\_boxplot() + geom\_point() + theme\_bw() + theme(legend.position = "none")*

*#Funció que crea un arbre de classificació*

> arbre <- C5.0(Survived ~ Class + Sex + Age, data = titanic\_net, control = C5.0Control(noGlobalPruning = TRUE))

*#Visualització de les dades obtingudes de l’arbre de classificació*

> arbre

> summary(arbre)

*#Representació de les regles de l’arbre de classificació*

> arbre <- C5.0(Survived ~ Pclass + Sex + Age, data = titanic\_net, control = C5.0Control(noGlobalPruning = TRUE),rules = TRUE)

> summary(arbre)

*#representació gràfica de l’arbre de classificació*

> plot(arbre)

1. *https://www.kaggle.com/c/titanic/discussion/3189* [↑](#footnote-ref-1)