PRÀCTICA 2. Tipologia i Cicle de vida de les dades

Manel Benavides Palos (manelbenavides)
Fèlix Ribera Forés (friberaf)
Màster Universitari en Ciència de Dades

Pregunta 1

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

Per a la realització d'aquesta pràctica s'ha escollit el **dataset Titànic** proposat a l'enunciat, provinent de la web Kaggle [1].

Aquest dataset facilita dades sobre els passatgers del titànic tals com l'edat, sexe i preu del bitllet adquirit entre d'altres. A més, també es pot observar si la persona és un supervivent de la tragèdia.

Per tant, aquest dataset és important ja que **permet trobar una correlació** entre la probabilitat de sobreviure a la tragèdia i les diferents característiques (ja siguin pròpies de la persona o del tipus d'adquisició que van fer a l'hora de pujar al vaixell).

La pregunta que es pretén respondre en aquesta pràctica sobre el dataset de Titànic és:

Quines són les característiques més rellevants a l'hora de predir si una persona sobreviu a aquesta tragèdia?

Per a respondre a aquesta pregunta, es defineixen les variables que es poden trobar al dataset:

- **PassengerId:** Identificador únic.
- **Name:** Nom.
- **Survived:** Variable que indica la supervivència després de la tragèdia:
  - 0: No va sobreviure
  - 1: Sí va sobreviure
- **Pclass:** Classe en la que viatjava el passatger.
  - 1: Primera classe.
  - 2: Segona classe.
  - 3: Tercera classe.
- **Sex:** Sexe.
- **Age:** Edat en anys.
- **Sibsp:** Número de germans i parelles a bord.
- **Parch:** Número de pares i fills a bord.
- **Ticket:** Número del bitllet.
- **Fare:** Preu del bitllet.
- **Cabin:** Codi identificador del camarot.

- **Embarked:** Port en el que van embarcar.
    - C: Cherbourg.
    - Q: Queenstown.
    - S: Southampton.

Pregunta 2

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

A partir de la variable *cabin*, que és una variable que no aporta informació a nivell estadístic, es crea dues variables discretes de les quals sí es pot analitzar i mirar d'extreure algún patró sobre la variable survived.

Aquestes variables es diran **cabinType** i **cabinNumber** i contindran la lletra del codi del camarot al que pertany el passatger i el nombre del camarot respectivament. D'aquesta manera, un passatger que té el camarot C27 el seu cabinType serà 'C' i el cabinNumber serà 27.

A més de la variable **cabin**, hi ha tres variables més que no aporten informació a nivell estadístic a l'hora d'analitzar les dades. Aquestes variables són: **Name**, **Ticket** i **PassengerId**. Per tant, aquestes quatre variables seran eliminades de cara a l'anàlisi en una **fase de reducció**.

Sobre el dataset, s'ha hagut de tractar certes variables per tal de discretitzar-les per trobar un sentit més adient a les variables. Aquestes variables discretes són **Pclass**, **Sex** i **Embarked**. A més, la variable **cabinType** que ja s'ha creat com a variable discreta serà tractada com a tal.

Per altra banda, les variables **Age**, **Sibsp**, **Parch** i **Fare** seran tractades con a variables contínues. A més, la variable **cabinNumber** que ja s'ha creat com a variable contínua serà tractada com a tal.

Pregunta 3

Neteja de les dades.

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

Inicialment, s'ha avaluat l'aparició de zeros al dataset. Aquests zeros apareixen a les variables **Survived**, **Sibsp**, **Parch** i **Fare**. La gestió d'aquests zeros a les variables consisteix a deixar-los tal i com es troben, ja que aquests tenen un sentit coherent i no són errors o absència de valors, sino que signifiquen que la persona no ha sobreviscut (a la variable Survived), que la persona no te pares/fills (variable Parch) o germans (variable Sibsp) a bord, o que la persona va obtenir un bitllet de manera gratuïta (variable Fare).

També s'ha fet una cerca d'elements buits a les variables (tant string buida "", com NA). El resultat ha estat que sobre **les 891 files que conté el dataset** hi ha 2 casos a la variable Embarked, 687 casos a la variable Cabin i 177 casos a la variable Age. La gestió que es farà en aquests casos serà la de **no eliminar les files de les variables qualitatives que contenen valors buits, ni tampoc omplir-los**, però a l'hora d'avaluar cada una d'aquestes variables, no es tindrà en compte les files que continguin un element buit. No obstant, **per a la variable quantitativa Age, s'imputa el valor de la mitjana a cadascun dels valors buits**.

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

S'ha realitzat un **estudi sobre els valors extrems** (outliers) per cada una de les variables contínues que existeixen al dataset, mencionades a la pregunta 2. Aquest estudi es pot veure a la Figura 1.

Tot i que es pot apreciar que existeixen valor extrems i que aquests provoquen una distorsió a nivell estadístic de la interpretació de la variable, els valors no es poden eliminar, ja que no tindria cap sentit real eliminar-los. Per exemple, en la variable Parch, tot el que no sigui un 0 és considerat un outlier. Si s'agafa el sentit real de la variable, no seria coherent excloure de l'estudi totes les persones que tinguin un pare/mare/fill a bord del vaixell.

Com el què es vol és fer un estudi prenent una interpretació coherent de les dades, **la gestió d'aquests valors extrems en tots els casos que s'han avaluat és de no eliminar-los**, ja que es tracta de valors coherents i que no són un error.

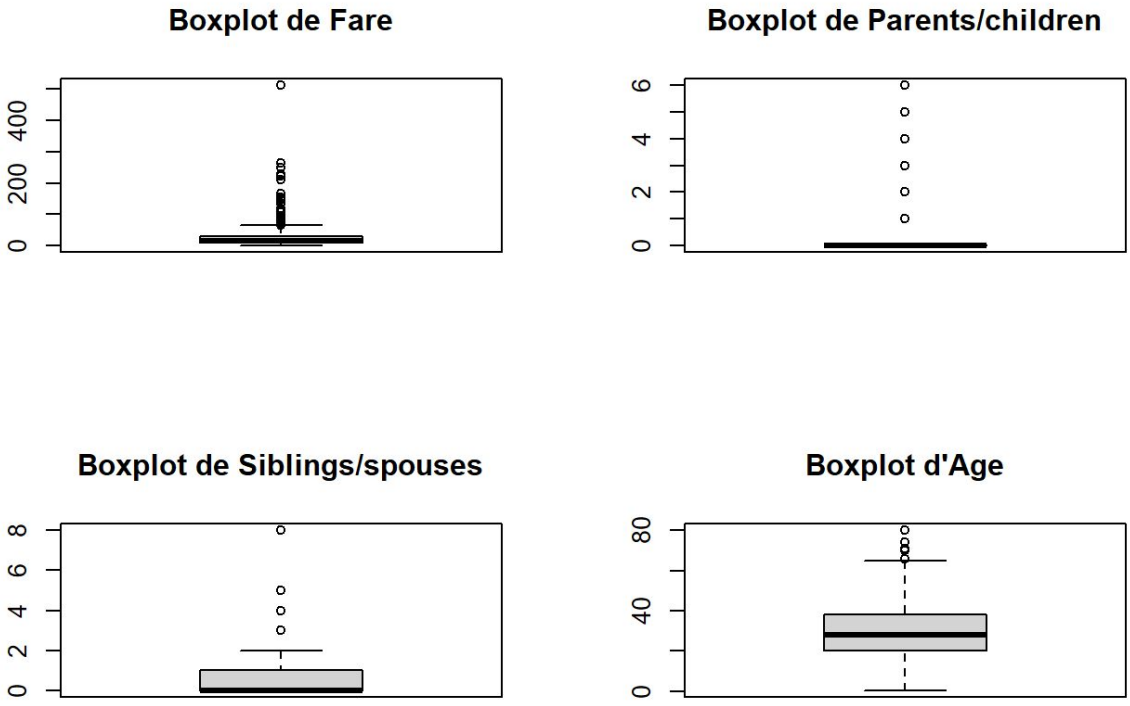

Figura 1. Estudi de valors extrems

Pregunta 4

Anàlisi de les dades.

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

Per a aquest apartat, el dataset se separa en els diferents grups que existeixen per a cada una de les variables discretes. Així doncs, obtindrem tres grups diferents per Pclass:

```
# Agrupació per classe.
df.p1 <- df[df$Pclass == "1",]
df.p2 <- df[df$Pclass == "2",]
df.p3 <- df[df$Pclass == "3",]
```

Obtindrem tres grups diferents per Embarked:

```
#Agrupació per port d'embarc
df.emb_s <- df[df$Embarked == "S",]
df.emb_q <- df[df$Embarked == "Q",]
df.emb_c <- df[df$Embarked == "C",]
```

Obtindrem vuit grups diferents per cabinType:

```
# Agrupació per cabinType
df.ctype_a <- df[df$cabinType == "A",]
df.ctype_b <- df[df$cabinType == "B",]
df.ctype_c <- df[df$cabinType == "C",]
df.ctype_d <- df[df$cabinType == "D",]
df.ctype_e <- df[df$cabinType == "E",]
df.ctype_f <- df[df$cabinType == "F",]
df.ctype_g <- df[df$cabinType == "G",]
df.ctype_h <- df[df$cabinType == "T",]
```

Obtindrem dos grups diferents per Sex:

```
# Agrupació per sex
df.male <- df[df$Sex == "male",]
df.female <- df[df$Sex == "female",]
```

- **Comprovació de la normalitat i homogeneïtat de la variança.**

Per tal de comprovar la normalitat de la població que conforma les variables contínues (quantitatives) es durà a terme la prova de normalitat d'Anderson-Darling.

D'aquesta manera, es comprova que el p-valor de cada una de les proves per cada columna ha de ser superior al nivell prefixat d'alfa=0.05 per tal de considerar que la variable segueix una distribució normal.

```
alpha = 0.05
col.names = colnames(df)
for (i in 1:ncol(df)) {
  if (is.integer(df[,i]) | is.numeric(df[,i])) {
    p_val = ad.test(df[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i]) cat('\n')
    }
  }
}
```

Donat que les dades no compleixen una distribució normal, per a comprovar la homogeneïtat de la variança es farà mitjançant el test de Fligner-Killeen. Es comprova la homogeneïtat de la variança de la variable sortida Survived contra cada una de les variables quantitatives.

```
fligner.test(Survived ~ Fare, data = df)
fligner.test(Survived ~ Parch, data = df)
fligner.test(Survived ~ SibSp, data = df)
fligner.test(Survived ~ Age, data = df)
```

D'on s'obtenen els resultats de p-value:
Survived-Fare = 0.299,
Survived-Parch =0.008,
Survived-SibSp =0.001,
Survived-Age = 0.81

Per tant, no es compleix la condició d'homogeneïtat de la variança en els casos de Survived-ParCh i Survived-SibSp, mentre si es compleix en els casos de Survived-Fare i Survived-Age.

- **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.**

Es vol conèixer com afecta cada una de les variables quantitatives sobre la probabilitat de sobreviure o no a la tragèdia. Per a aconseguir això, s'ha elaborat una **prova de correlacions** entre la variable de sortida *Survived* i cada una de les variables quantitatives.

```
df_num <- df[, sapply(df, is.numeric)]
corrplot(cor(df_num), method="color", addCoef.col = "black")
```

L'explicació dels resultats obtinguts serà exposada a l'apartat següent.

La segona prova realitzada sobre les dades pretén respondre a la hipòtesi: **Hi ha alguna diferència entre ser home o dona a l'hora de sobreviure?** En altres paraules, es va portar a la pràctica "Les dones i els nens primer"?

Per a fer això, es realitza un **contrast d'hipòtesi**, presentat a continuació.

μ_1: mitja de la població masculina que sobreviu
μ_2: mitja de la població femenina que sobreviu

L'experiment planteja com a hipòtesi nul·la H_0 que la probabilitat de sobreviure sent de qualsevol d'ambdós sexes és la mateixa, mentre que H_1 indica que la probabilitat de sobreviure sent dona és major a la probabilitat de sobreviure sent home:

$$\mathbf{H_0: \mu_1 - \mu_2 = 0}$$
$$\mathbf{H_1: \mu_1 - \mu_2 < 0}$$

Prendrem un interval de confiança del 95%, per la qual cosa $\alpha = 0.05$.

```
df.male.survived <- df[df$Sex == "male",]$Survived
df.female.survived <- df[df$Sex == "female",]$Survived

t.test(df.male.survived, df.female.survived, alternative = "less")

ggplot(df[1:891,], aes(Age, fill = factor(Survived)))
  + geom_histogram()
  + facet_grid(.~Sex)
```

I com a resultat obtingut:

```
Welch Two Sample t-test
data: df.male.survived and df.female.survived
t = -18.672, df = 584.43, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0 95 percent
confidence interval:
-Inf -0.5043259
sample estimates: mean of x mean of y
```

```
0.1889081 0.7420382
```

El resultat serà interpretat en l'apartat següent.

Finalment, s'ha creat uns **models de regressió lineal** per tal de veure quin nivell de model de regressió s'és capaç de trobar a partir de les variables del dataset.

S'ha decidit crear 3 models diferents: En un primer model, s'avaluen només les variables quantitatives, en un segon model només les variables qualitatives i el tercer model pren com a variables d'entrada tant les variables quantitatives com les qualitatives. Aquests models s'han fet usant el codi següent:

```
model_1 <- lm(Survived ~ Age + Parch + SibSp + Fare + cabinNumber, data =
df)
model_2 <- lm(Survived ~ Pclass + Sex + Embarked + cabinType, data = df)
model_3 <- lm(Survived ~ Age + Parch + SibSp + Fare + cabinNumber + Pclass
+ Sex + Embarked + cabinType, data = df)

taula.coeficients <- matrix(c(
  "Quantitatives", summary(model_1)$r.squared,
  "Qualitatives", summary(model_2)$r.squared,
  "Mix", summary(model_3)$r.squared
  ),
  ncol = 2, byrow = TRUE
)
colnames(taula.coeficients) <- c("Model", "R^2")
taula.coeficients
```

Els resultats es presenten en l'apartat següent.

Pregunta 5

Representació dels resultats a partir de taules i gràfiques.

Pel què fa a la correlació, es pot observar quina **correlació** hi ha entre totes les variables quantitatives. Una correlació perfecta tindrà mòdul 1 i una correlació que no guarda cap relació tindrà mòdul 0. El signe només indica si la relació és directament proporcional o inversament proporcional. Els resultats es poden veure a la *Figura 2*.

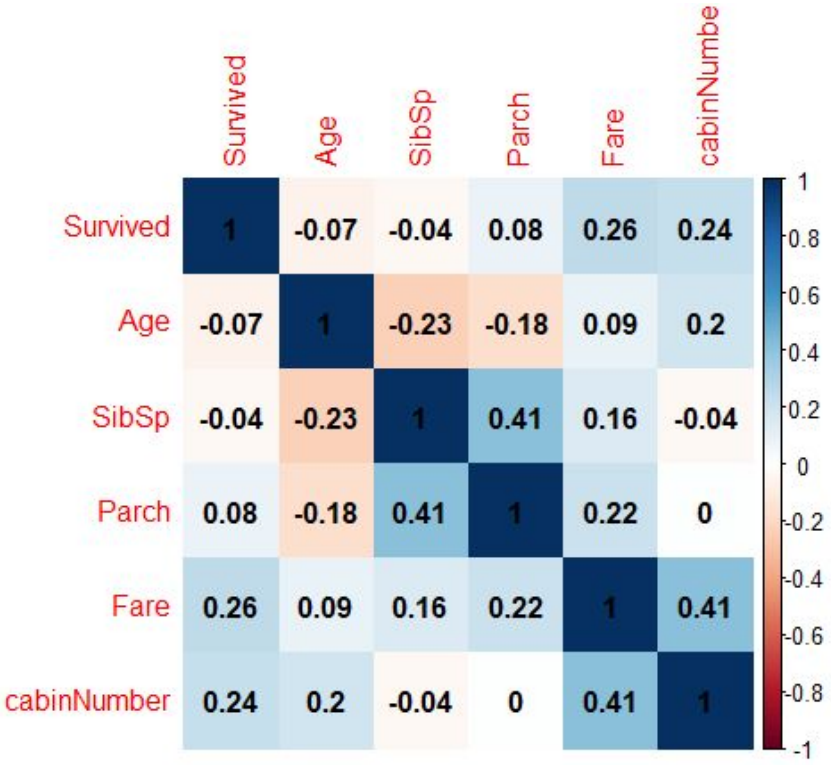

Figura 2. Correlació entre variables quantitatives

Es pot observar que existeix certa relació entre Survived i les variables Fare i cabinNumber. Aquesta relació es pot interpretar com que la gent que va pagar més pel seu bitllet (generalment gent de més poder adquisitiu) es va salvar en major proporció, mentre que la gent que va pagar menys pel seu bitllet (en general gent de menys poder adquisitiu) es va salvar en major proporció.

Una altra correlació interessant es troba en les variables Parch i Fare, on sembla que les famílies (valor alt de Parch - Parents/children a bord) van pagar més per bitllet, entenent que necessiten més comoditats ja que van amb gent gran o nens.

Per altra banda, hi ha una forta relació entre les variables Parch i Sibsp, donant a veure que hi havia un gran nombre de famílies que anaven tant amb germans com amb esposes i pares i fills.

A més, sembla que les famílies eren de gent de mitjana edat amb nens petits. Així ho indica la correlació de signe negatiu entre les variables Parch i Sibsp amb la variable Age.

Finalment, una altra correlació que salta a la vista és la variable Fare amb la variable CabinNumber, on sembla que els camarots amb nombre més elevat deuen ser més luxoses i per això el preu pagat per elles és més car.

En la prova de **contrast d'hipòtesi** s'ha volgut comprovar si les dones es salven amb més percentatge que els homes.

Donat que el p-valor ha estat menor a 0.05, la hipòtesi alternativa desbanca la hipòtesi nul·la, per tant es pot assegurar que les dones se salven en més percentatge que els homes.

A la *Figura 3* es pot comprovar que aquesta afirmació és certa.

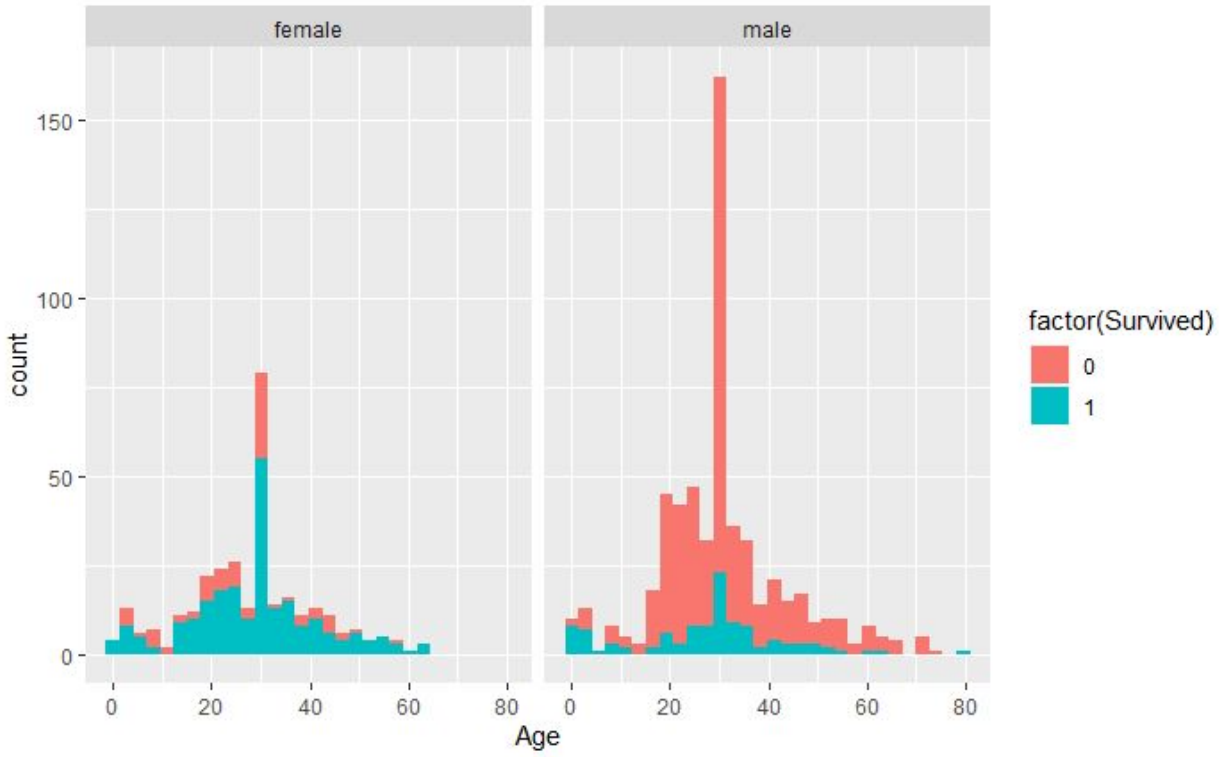

Figura 3. Survived en funció del sexe

Tal i com es veu a la figura, el percentatge de dones que sobreviuen a la tragèdia és clarament superior al percentatge d'homes que sobreviuen a la mateixa tragèdia. D'aquesta manera, es pot observar que es va complir la expressió "Les dones i els nens primer".

Pel què fa als **models de regressió lineal** que s'han creat, el valor a observar per contrastar entre ells i decidir quin és millor és l'R^2, i pot ser observat a la Figura 4.

```
     Model           R^2
[1,] "Quantitatives" "0.112346236189579"
[2,] "Qualitatives"  "0.387144392223383"
[3,] "Mix"           "0.413430912164883"
```

Figura 4. Comparativa de models

Es pot veure que amb una R2 de 0.41, el millor model és el que engloba totes les variables d'entrada. Tot i això, no s'ha aconseguit obtenir un resultat significativament bo. Per a aconseguir un model millor, es podria fer un altre tipus de model (per exemple un Random Forest o un SVM entre d'altres) o la creació de diverses noves variables a partir d'altres de conegudes (per exemple, utilitzar el nom que figura al dataset per extreure el títol de la persona, sigui Mr, Mrs, Ms, Dr, etc) i observar si aquest té una correlació superior amb la variable *Survived*.

També, a més dels tests estadístics de l'apartat anterior, s'ha volgut **observar com és la probabilitat de supervivència en funció de les diferents variables qualitatives**, i el resultat es pot observar a la *Figura 5*.

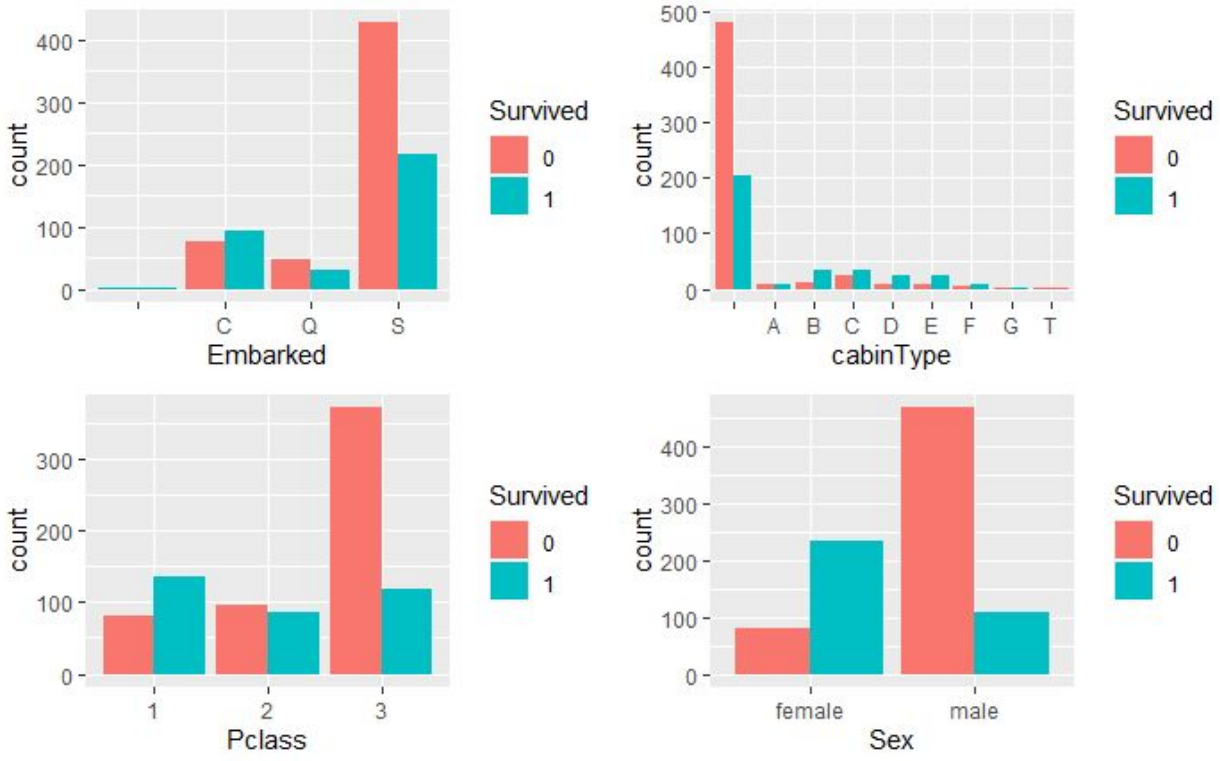

Figura 5. Variables qualitatives vs Survived

En la variable Embarked, es pot observar que la supervivència en la gent que va embarcar a C i Q (Cherbourg, Queenstown) és propera al 50%, mentre que la gent que va embarcar a S (Southampton) en major part no va sobreviure.

Pel què fa al tipus de cabina, es pot observar que la majoria de gent no tenia un camarot assignat. Tot i això, també es veu com la majoria de gent sense camarot assignat mor, mentre que de la gent que té camarot assignat, la meitat d'ells se salva. Aquesta dada té sentit amb la correlació entre Survived i Fare, on sembla ser que la gent de més poder adquisitiu va sobreviure, mentre que la gent més humil econòmicament no va aconseguir-ho.

Pel què fa a la classe en la què viatjaven els passatgers, es pot veure reforçada la idea de l'anterior paràgraf, doncs la gent tal i com va pujant la classe en la què es viatjava també incrementa la proporció de supervivència.

Finalment, tal i com ja s'ha comentat prèviament, es pot veure com la proporció de dones que sobreviuen és molt major a la proporció d'homes que sobreviuen. A més, com a dada extra, es pot extreure que viatjaven molts més homes que dones.

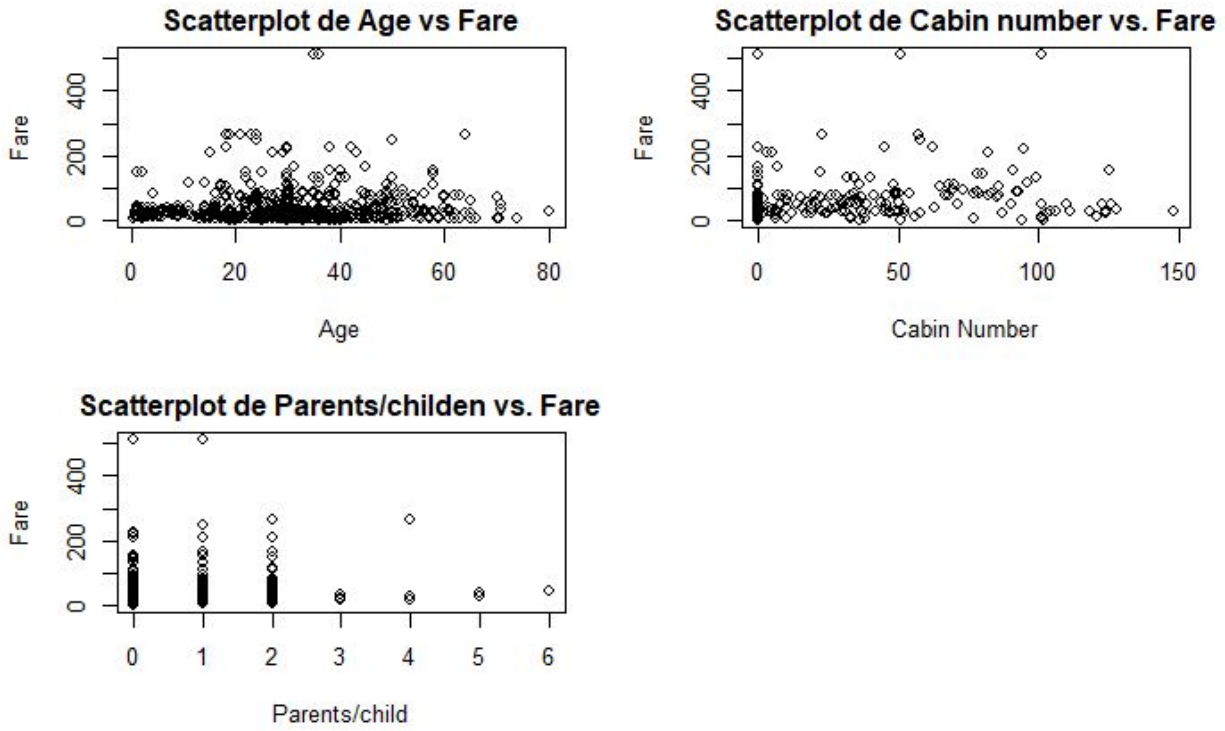

Figura 6. Preu pagat respecte a variables interessants

A partir de la *Figura 6* es pot veure com a mesura que augmenta l'edat fins a una mitjana edat (fins a uns 50-60 anys), també augmenta la tendència a pagar un bitllet més elevat (es pot

relacionar amb que a l'edat mencionada, una persona sol aconseguir la maduresa econòmica i per tant, més poder adquisitiu per a demandar més prestacions a un preu superior).

La relació entre el preu del bitllet i el nombre de camarot sembla no tenir gaire relació entre les dues variables.

Per altra banda, aparentment les famílies (Parents/children) amb més membres semblen requerir de prestacions superiors per a facilitar la seva estança, i per això hi ha una tendència a l'augment del preu del bitllet contrastat amb el nombre de Parch del dataset.

Pregunta 6

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

En aquest dataset s'ha pogut observar que hi havia moltes dades a partir de les quals poder treure informació i fer servir per a poder predir diferents tipus d'hipòtesis.

S'ha hagut de netejar les dades i prendre decisions de disseny en aquesta neteja, com ara el tractament de valors nuls i de valors extrems. Durant el procés de la pràctica, s'ha fet algún canvi de guió en aquest sentit, i s'ha observat que els resultats varien molt segons el tipus de decisió que es prengui sobre els valors nuls i els valors extrems. En el nostre cas, per tal de no llençar i desaprofitar molts exemples, no s'ha volgut eliminar les files en les quals hi faltava informació, tot i que en un dataset on la majoria de casos continguin informació, segurament el més purista seria eliminar aquestes files del dataset.

Durant el procés d'anàlisi s'ha observat que segons les característiques de les variables (normalitat i variança) s'ha hagut d'aplicar un tipus de test o altres, segons es pot haver seguit a la teoria de l'assignatura.

Finalment, s'han dut a terme unes proves estadístiques per tal de veure la relació entre diferents variables, que s'han interpretat a través de gràfiques i taules, i s'ha intentat trobar el sentit més acurat en aquests resultats, muntant un storytelling al seu voltant que facilita la comprensió del què va succeir durant una tragèdia tan famosa com la del naufragi del Titànic.

Pregunta 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.

El codi es pot trobar al següent enllaç de github:

https://github.com/friberaf/PAC2-analisis

TAULA DE CONTRIBUCIONS

Contribucions	Signa
Investigació prèvia	MBP, FRF
Redacció de les respostes	MBP, FRF
Desenvolupament del codi	MBP, FRF

Bibliografia

[1] https://www.kaggle.com/c/titanic/data?select=train.csv