Titanic - Learning from the Disaster

Mireia Mora, Jose Antonio Montero

04/12/2020

1. Descripció del Dataset

Lectura de dades.

En primer lloc, llegiu el fitxer de dades i verifiqueu que els tipus de dades són interpretats correctament.

Si s'escau, feu les conversions de tipus que siguin oportunes.

```
#install.packages("stringr")
library(stringr)
## LLegim ele csv amb la sentencia read.csv i fem summary
setwd("D:/UOC_ML/S1-Tipologia i Cicle Vida Dades/PAC/PRAC2")
titanic <- read.csv("train_titanic.csv", dec=".", stringsAsFactors = FALSE)
summary(titanic)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
#tipus de cada variable
sapply(titanic,class)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   "integer"   "integer"   "integer" "character" "character"   "numeric"
```

```
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   "integer"   "integer" "character"   "numeric" "character" "character"
```

```
#a few data
head(as.matrix(titanic),3)
```

```
##      PassengerId Survived Pclass
## [1,] "  1"       "0"      "3"
## [2,] "  2"       "1"      "1"
## [3,] "  3"       "1"      "3"
##      Name                                                 Sex      Age
## [1,] "Braund, Mr. Owen Harris"                            "male"   "22.00"
## [2,] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "female" "38.00"
## [3,] "Heikkinen, Miss. Laina"                             "female" "26.00"
##      SibSp Parch Ticket             Fare       Cabin Embarked
## [1,] "1"   "0"   "A/5 21171"        "  7.2500" ""    "S"
## [2,] "1"   "0"   "PC 17599"         " 71.2833" "C85" "C"
## [3,] "0"   "0"   "STON/O2. 3101282" "  7.9250" ""    "S"
```

```
#llegim el dataset de test
titanic_test <- read.csv("test_titanic.csv", dec=".", stringsAsFactors = FALSE)
summary(titanic_test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age             SibSp            Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked
##  Min.   :  0.000   Length:418         Length:418
##  1st Qu.:  7.896   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

```
#tipus de cada variable
sapply(titanic_test,class)
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##   "integer"   "integer" "character" "character"   "numeric"   "integer"
##       Parch      Ticket        Fare       Cabin    Embarked
##   "integer" "character"   "numeric" "character" "character"
```

```r
#a few data
head(as.matrix(titanic_test),3)
```

```
##      PassengerId Pclass Name                                  Sex      Age
## [1,] " 892"      "3"    "Kelly, Mr. James"                    "male"   "34.50"
## [2,] " 893"      "3"    "Wilkes, Mrs. James (Ellen Needs)" "female" "47.00"
## [3,] " 894"      "2"    "Myles, Mr. Thomas Francis"           "male"   "62.00"
##      SibSp Parch Ticket   Fare       Cabin Embarked
## [1,] "0"   "0"   "330911" "  7.8292" ""    "Q"
## [2,] "1"   "0"   "363272" "  7.0000" ""    "S"
## [3,] "0"   "0"   "240276" "  9.6875" ""    "Q"
```

Descripció del dataset. ¿Per qué és important i quin/es preguntes/problema pretend respondre?

El dataset del Titanic és un dels més popular en l'anàlisis de dades per l'impacte que va tenir el succés.

Principalment està orientat a **calcular la probabilitat de supervivència dels passatgers en funció de les seves característiques (edat, clase social, preu del passatge, familia, etc)** però també dona lloc a fer-se preguntes de l'estil **si el preu del passatge dels que sobreviuen és més gran que el preu del passatge dels que moren**, per tant orientarem els nostres anàlisis per a respondre aquest tipus de preguntes.

2. Integració i selecció de les dades de interés a analitzar

En aquest apartat també farem screening i creació de noves variables / discretització.

```
#estructura
str(titanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
#gráficos

#par(mfrow=c(2,2))
par(mfrow=c(3,3))


#1) Boxplot de Price -> quantitativa continua
#boxplot(childseats$Advertising,main="Advertising Box plot", col="gray")
```

3. Neteja de dades

3.1 Reducció.

Dintre de la neteja de dades una de les opcions és la reducció que consisteix a eliminar variables redundants o que no estan relacionades amb el fet que es vol analitzar.

```
##Esborrarem les variables Cabin,Ticket,Name,PassengerId
##Una vegada tiguem calculada la IsAlone també esborrarem Parch,SibSp i FamilySize

titanic = titanic[,!(names(titanic) %in% c("Ticket","Cabin","Name","PassengerId"))]
titanic_test = titanic_test[,!(names(titanic_test) %in% c("Ticket","Cabin","Name","PassengerId"))]

#titanic = titanic[,!(names(titanic) %in% c("Parch","SibSp","FamilySize"))]
#titanic_test = titanic_test[,!(names(titanic_test) %in% c("Parch","SibSp","FamilySize"))]
```

3.2 Identificació i tractament de valors extrems.

Tractament de outliers.

```
#crearem grafics boxplot per a cada una de les variables regressores
#que ens interesen
```

3.3 Dades perdudes - missing data.

Les dades contenen ceros o elements buits? Com gestionaries cadascun d'aquests casos?.

Anem a evaluar quines variables tenen valors nulls o no informats (amb "") i les completarem i discretitzarem si escau.

Además, de cara a fer prediccions i sobretot si volem fer regressió logística és adequat que les variables convertides **quedin com a tipus factor** en R.

```
#
#missing values
#

colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0      177        0        0        0        0
```

```
#a train tenim bàsicament null values a Age, 177

colSums(titanic=="")
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0       NA        0        0        0        2
```

```
#a train tenim valors perduts no nulls a Cabin i a Embarked

colSums(is.na(titanic_test))
```

```
##   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0       86        0        0        1        0
```

```
#a train tenim bàsicament null values a Age 86 i 1 a Fare

colSums(titanic_test=="")
```

```
##   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
```

```
##        0        0       NA        0        0       NA        0
```

```
#a train tenim valors perduts no nulls a Cabin

#
# Embarked
#

#Al set de traing tenim 2 missing values que son "", no son pas NA
#Els completem amb el valor més freqüent
#Discretitzem Embarked - convertim Embarked a numeric (0-2 son 3 valors)

#Crearem una funció on passem una columna de dataframe i ens la retorna discretitzada

clean_Embarked <- function (cp) {
  #asignamos los 2 "" el valor més freqüent freq_embarked
  freq_embarked <- tail(names(sort(table(cp))), 1)
  cp[cp==""] <- freq_embarked
  #discretitzem
  cp[cp=='S'] <- 0
  cp[cp=='C'] <- 1
  cp[cp=='Q'] <- 2
  cp<-as.factor(cp)
  return(cp)
}

#Train
titanic %>% count(Embarked)
```

```
## # A tibble: 4 x 2
##   Embarked     n
##   <chr>    <int>
## 1 ""           2
## 2 "C"        168
## 3 "Q"         77
## 4 "S"        644
```

```
titanic$Embarked <- clean_Embarked(titanic$Embarked)
titanic %>% count(Embarked)
```

```
## # A tibble: 3 x 2
##   Embarked     n
##   <fct>    <int>
## 1 0          646
## 2 1          168
## 3 2           77
```

```
#Test
titanic_test %>% count(Embarked)
```

```
## # A tibble: 3 x 2
##   Embarked     n
##   <chr>    <int>
## 1 C          102
## 2 Q           46
## 3 S          270
```

```r
titanic_test$Embarked <- clean_Embarked(titanic_test$Embarked)
titanic_test %>% count(Embarked)
```

```
## # A tibble: 3 x 2
##   Embarked     n
##   <fct>    <int>
## 1 0          270
## 2 1          102
## 3 2           46
```

```r
#
# Fare - crearem una nova variable però deixem l'original per a fer un test de hipòtesis posterior
#

#Al set de test tenim 1 NA
#Els completem amb el valor més freqüent
#Discretitzem Fare - creem una rang per Fare de 4 convertim Fare a factor (Q1-Q4 son 4 valors)

clean_Fare <- function (cp) {
  #com a paràmetre rep una columna d'un dataframe que serà l'afectada
  #asignamos els nulls o  "" el valor més freqüent freq_fare
  fare_embarked <- tail(names(sort(table(cp))), 1)
  #cp[cp==""] <- fare_embarked
  cp[is.na(cp)] <- fare_embarked
  #hem de convertit a numeric per poder fer els rangs
  cp <- as.numeric(cp)
  #discretitzem en base a generar un rang de 4 buckets bassat en quartiles
  #la funció CutQ genera els rangs i ja els assigna segons el valor
  fare_quartiles <- CutQ(cp)
  #és una variable tipus factor que assignem a la variable de sortida
  cp <- fare_quartiles
  return(cp)
}

#Train
titanic <- titanic %>%
  mutate(Fare_disc =  clean_Fare(titanic$Fare)
  )
#titanic$Fare <- clean_Fare(titanic$Fare)
titanic %>% count(Fare_disc)
```

```
## # A tibble: 4 x 2
##   Fare_disc     n
##   <fct>     <int>
## 1 Q1          223
## 2 Q2          224
## 3 Q3          222
## 4 Q4          222
```

```r
#Test
titanic_test <- titanic_test %>%
  mutate(Fare_disc =  clean_Fare(titanic_test$Fare)
  )
#titanic_test$Fare <- clean_Fare(titanic_test$Fare)
titanic_test %>% count(Fare_disc)
```

```
## # A tibble: 4 x 2
##   Fare_disc     n
##   <fct>     <int>
## 1 Q1          115
## 2 Q2           96
## 3 Q3          102
## 4 Q4          105
```

Pels cas del missing values de la variable Age anem a fer una mica de tractament especial.

Al tractar-se d'una variable cuantitativa continua ens interessa **omplir el missing values (que no son poc) d'una forma acurada** i per altre banda, pensant en els anàlisis posteriors **ens interessa discretitzar aquesta variable**.

Per tant procedirem de la següent manera:

1. predicció de valors fent servir altres features correlades (Age, Sex, Pclass), agafarem per cada combinació de Pclass-Sex la mediana del valor de Age, i aquest serà el que assignarem per a totes les combinacinos de Pclass-Sex que tinguin missing values o NA
2. creem una nova feature AgeBand (5 intervals de edat)
3. substituim Age per AgeBand
4. finalment podem esborrar la AgeBand

```
#creem la función per al tractament de la variable Age

clean_Age <- function (df) {
  #com a paràmetre rebem un data-frame complet perque necessitarem varies columnes

  #Necessitem com a pas previ discretitzar la variable Sex
  df$Sex[df$Sex=='male'] <- 0
  df$Sex[df$Sex=='female'] <- 1
  df$Sex<-as.integer(df$Sex)

  #1.matriu pels guessed values de age segons Sex i Pclass
  pred_age <- matrix(nrow = 3, ncol = 2)
  #2.càlcul de les medianes per a cada combinació de Pclass (i)
  for(i in 1:3) {
    for(j in 1:2) {
      df_ages <- subset(df,Pclass==i & Sex==j-1)
      pred_age[i,j]  <- median(df_ages$Age,na.rm=T)
      }
  }
  #3.canviem els NA per a cada combinació de Pclass i Sex pel valor que hem calculat abans
  for(i in 1:3) {
    for(j in 1:2) {
      df$Age[is.na(df$Age) & df$Pclass ==i & df$Sex == j-1] <- pred_age[i,j]
    }
  }

  #4.Creem els intervals de Age
  df$Age_grouping <- cut(df$Age, breaks=c(0,16,32,48,64,100,140), right = FALSE, labels = FALSE)
  #5.Assignem els intervals com a valor de Age
  df$Age <- df$Age_grouping
  #6.Esborrem la variable intermitja Age_grouping
  df = df[,!(names(df) %in% c("Age_grouping"))]
```

```
  return(df)
}

#Train
titanic <- clean_Age(titanic)
titanic %>% count(Age)
```

```
## # A tibble: 5 x 2
##     Age     n
##   <int> <int>
## 1     1    83
## 2     2   492
## 3     3   227
## 4     4    76
## 5     5    13
```

```
head(titanic)
```

```
##   Survived Pclass Sex Age SibSp Parch    Fare Embarked Fare_disc
## 1        0      3   0   2     1     0  7.2500        0        Q1
## 2        1      1   1   3     1     0 71.2833        1        Q4
## 3        1      3   1   2     0     0  7.9250        0        Q2
## 4        1      1   1   3     1     0 53.1000        0        Q4
## 5        0      3   0   3     0     0  8.0500        0        Q2
## 6        0      3   0   2     0     0  8.4583        2        Q2
```

```
#Test
titanic_test <- clean_Age(titanic_test)
titanic_test %>% count(Age)
```

```
## # A tibble: 5 x 2
##     Age     n
##   <int> <int>
## 1     1    32
## 2     2   251
## 3     3    91
## 4     4    39
## 5     5     5
```

```
head(titanic_test)
```

```
##   Pclass Sex Age SibSp Parch    Fare Embarked Fare_disc
## 1      3   0   3     0     0  7.8292        2        Q1
## 2      3   1   3     1     0  7.0000        0        Q1
## 3      2   0   4     0     0  9.6875        2        Q2
## 4      3   0   2     0     0  8.6625        0        Q2
## 5      3   1   2     1     1 12.2875        0        Q2
## 6      3   0   1     0     0  9.2250        0        Q2
```

4. Anàlisis de de les dades

4.1 Selecció de dades.

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

Els grups els tenim clarament identificats en el sentit de que farem servir el cojunt de train (titanic) per a generar els models, i el conjunt de test (titanic_test) per a provar-lo o ajustar-lo, i obtenir la seva bondat (accuracy).

Per centrant-nos en els diferents tipus d'anàlisis i en concret si volem fer un contrast de hipótesis per a validar si la mija dels preus (fare) dels que sobreviuen és més gran i igual que la dels que moren, crearem dos grups entorn a la variable preu : els preus dels que sobreviuen i els preus dels que moren

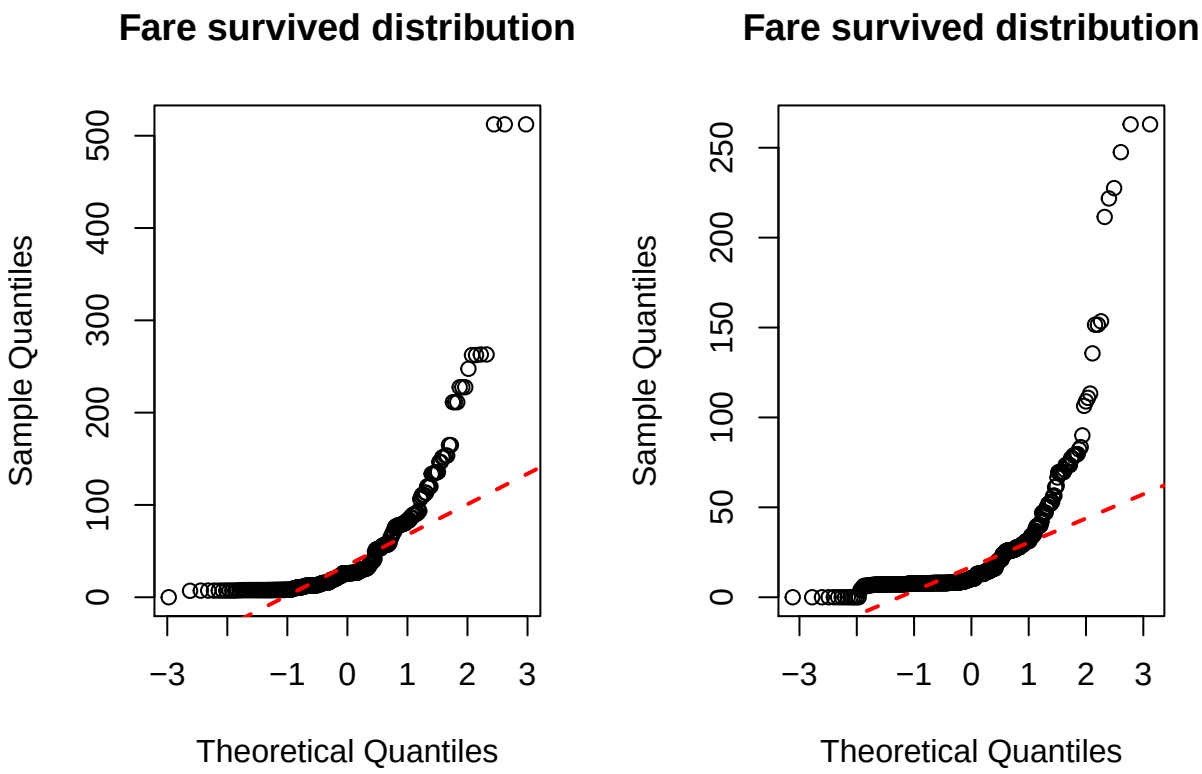
```
fare_vius = titanic$Fare[titanic$Survived==1]
fare_morts = titanic$Fare[titanic$Survived==0]
```

4.2 Comprovació de la normalitat i homegeneitat de la variança.

Farem inicialment aquestes comprovacions per la variable Fare que és la que voldrem involucrar en el contrast de hipòtesis.

Per la comprovació de normalitat ho farem de manera visual fent servir la funció qnorm

```
#1) Diagrama de punts de fare_vius i fare_morts
par(mfrow=c(1,2))

qqnorm(fare_vius,main="Fare survived distribution");qqline(fare_vius, col = 2,lwd=2,lty=2)

qqnorm(fare_morts,main="Fare survived distribution");qqline(fare_morts, col = 2,lwd=2,lty=2)
```



Encara que tenim força punts fora de la línea recta podriem dir que la majora s'agrupen al voltat d'ella per tant donarem per supossat el factor de normalitat, encara que amb dubtes. Amb un conjunt de dades gran podriem arrivar a assumir el factor de normalitat però amb només 891 observaciones aquesta afirmació queda en entredit. De totes formes farem el contrast supossant normalitat.

Per la comprovació de la homocedasticitat podem fer servir la funció var.test de R.

```
#Comprovem homocedasticitat - variances iguals
var.test(x = fare_vius,y = fare_morts)
```

```
##
##  F test to compare two variances
##
## data:  fare_vius and fare_morts
## F = 4.5017, num df = 341, denom df = 548, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.725366 5.463382
## sample estimates:
## ratio of variances
##           4.501697
```

De la sortida de var.test podem veure que **el ratio of variances que és de 4.50 està dintre del interval de confiança de 95%**, per tant el test no troba diferencies significatives entre les variances d'ambdos grups.

4.3 Aplicació proves estadístiques.

Aplicació de proves estadístiques per a comparar els grups de dades. En funció de les dades i l'objectiu de l'estudi, aplicar proves de contrast de hipòtesis, correlacions, regressions, etc. Aplicar almenys 3 mètodes d'anàlisis diferents.

4.3.1 Contrastos de hipotesis.

Podem acceptar que el preu del passatge (Fare) és més gran en els que sobreviuen que en els que moren? Validarem aquest fet amb un contrast de hipòtesis per a la diferència de 2 mitjanes amb els suposits de normalitat i homocedasticitat, és a dir, aplicarem una t de Student.

La hipòtesi nul.la H0 seria establir que **la mitjana de preus del passatge dels que sobreviuen es igual a la dels que no sobreviuen.**

H0 : mean(Fare sobreviuen) = mean(Fare moren)

ó

H0 : mean(Fare sobreviuen) - mean(Fare moren) = 0

La hipòtesi alternativa H1 representa que s'ha produit algun canvi respecte la situació descrita per la hipòtesis nul.la. En aquest cas establirem el fet que realment volem comprovar al final, que la mitjana del preu del passatge del que sobreviuen és més gran que la dels que moren (és un contrast unilateral per la dreta).

H1 : mean(Fare sobreviuen) > mean(Fare moren)

ó

H1 : mean(Fare sobreviuen) - mean(Fare moren) > 0

Encara que podem utilitzar directament la funció t.test de R ens farem una funció propia que ens implementi el càlcul del nostre contrast

```
#Contruim una funció que ens faci el contrast de la diferencia de
#mitjanes de 2 mostres en mode unilateral dret

contrast_dif_mitjana_2_mostres <- function (p_mostra1,p_mostra2, p_alfa,p_tipus) {

  #p_mostra1 és la mostra 1
  #p_mostra2 és la mostra 2
  #p_alfa és el nivell de significació

  #H0 : mean(p_mostra1) - mean(p_mostra2) = 0 - hipòtesis nul.la
  #H1 : mean(p_mostra1) - mean(p_mostra2) > 0 - hipòtesis alternativa

  #Calculem tamany,mitjanes i desviacions típiques d'ambues mostres
  n_us = length(p_mostra1)
  n_nous = length(p_mostra2)

  mean_us = mean(p_mostra1)
  mean_nous = mean(p_mostra2)

  dev_tipica_us = sqrt(sum((p_mostra1-mean_us)^2)/(n_us-1))
  dev_tipica_nous = sqrt(sum((p_mostra2-mean_nous)^2)/(n_nous-1))

  #calculem estadístic t
  #ditribució t-student amb n_us+n_nous - 2 graus de llibertat (398)
```

```
  s = sqrt(((n_us-1)*dev_tipica_us^2+((n_nous-1)*dev_tipica_nous^2))/(n_us+n_nous-2))

  s_error_std = s * sqrt((1/n_us)+(1/n_nous))

  t = (mean_us - mean_nous) / s_error_std

  #i finalment el p-value tenint en compte la distribució de t
  #i la hipotesis alternativa
  p_valor = case_when (p_tipus == 'uniesquerra' ~ pt(t,n_us+n_nous-2),
                       p_tipus == 'unidreta' ~ pt(-t,n_us+n_nous-2),
                       p_tipus == 'bidireccional' ~ 2*pt(t,n_us+n_nous-2)
                       )

  #si p_value >= nivell de significació p_alfa, acceptarem la H0
  #si p_value < nivell de significació p_alfa, rebutjarem la H0


  if (p_valor >= p_alfa) {
    ic = c(t,p_valor,'Acceptem Hipòtesi nul.la')
  }
  else
  {
    ic = c(t,p_valor,'Rebutgem Hipòtesi nul.la')
  }

  return(ic)

}

c_mitjanes = contrast_dif_mitjana_2_mostres(fare_vius,fare_morts,0.5,'unidreta')
c_mitjanes
```

```
## [1] "7.93919166087105"         "3.06009467096209e-15"
## [3] "Rebutgem Hipòtesi nul.la"
```

```
#[1] t(estadístic) = "7.93919166087105"         p_valor = "3.06009467096209e-15"
#[3] Resultat : "Rebutgem Hipòtesi nul.la"

#el p_value es pot dir que és infim i per tant més petit que 0.05 així que rebutjariem
#la Hipotesis nul.la


#Comprovació amb t_test
t.test( fare_vius, fare_morts, # dues mostres
        alternative = "greater", # contraste per resta de mitjanes
        paired = FALSE, # muestras independientes
        var.equal = TRUE,  # se supone homocedasticidad
        conf.level=0.95)
```

```
##
##  Two Sample t-test
##
## data:  fare_vius and fare_morts
## t = 7.9392, df = 889, p-value = 3.06e-15
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  20.82763      Inf
## sample estimates:
## mean of x mean of y
##  48.39541  22.11789
```

4.3.1.1 Interpretar contratos de hipotesis. Com el **p-value (3.06e-15, que és la probabilitat del resultat del estadístic t quan la hipòtesis nul.la és certa) és més PETIT que el nivell d'acceptació (0,05) llavors REBUTGEM la Hipòtesis nul.la** per tant això vol dir que confirmem que els preus del passatges del que sobreviuen és més gran el preu dels que moren.

Per una altre banda, la sortida de la funció t.test ens està dient que **el p_value NO està dintre de l'interval d'acceptació de la hipòtesi nul.la**, per tant ens porta a rebutjar-la.

Així mateix, el fet de que **l'estadístic de contrast sigui gran (7.9392)** fa que estigui allunyat del cero (zona on la distribució normal estàndard concentra una probabilitat més gran), per tant **poc probable sota la hipòtesis nul.la.** El fet d'haver plantejar una hipòtesi alternativa unilateral per la dreta fa també que aquest fet de tenir un valor positiu per l'estadístic de contrast, aquest sigui més probable sota la alternativa.

4.3.1.2 Interpretar correlacions.

4.3.1.3. Interpretar regressió logaritmica

4.3.1.4. Interpretar altres models

5. Visualització

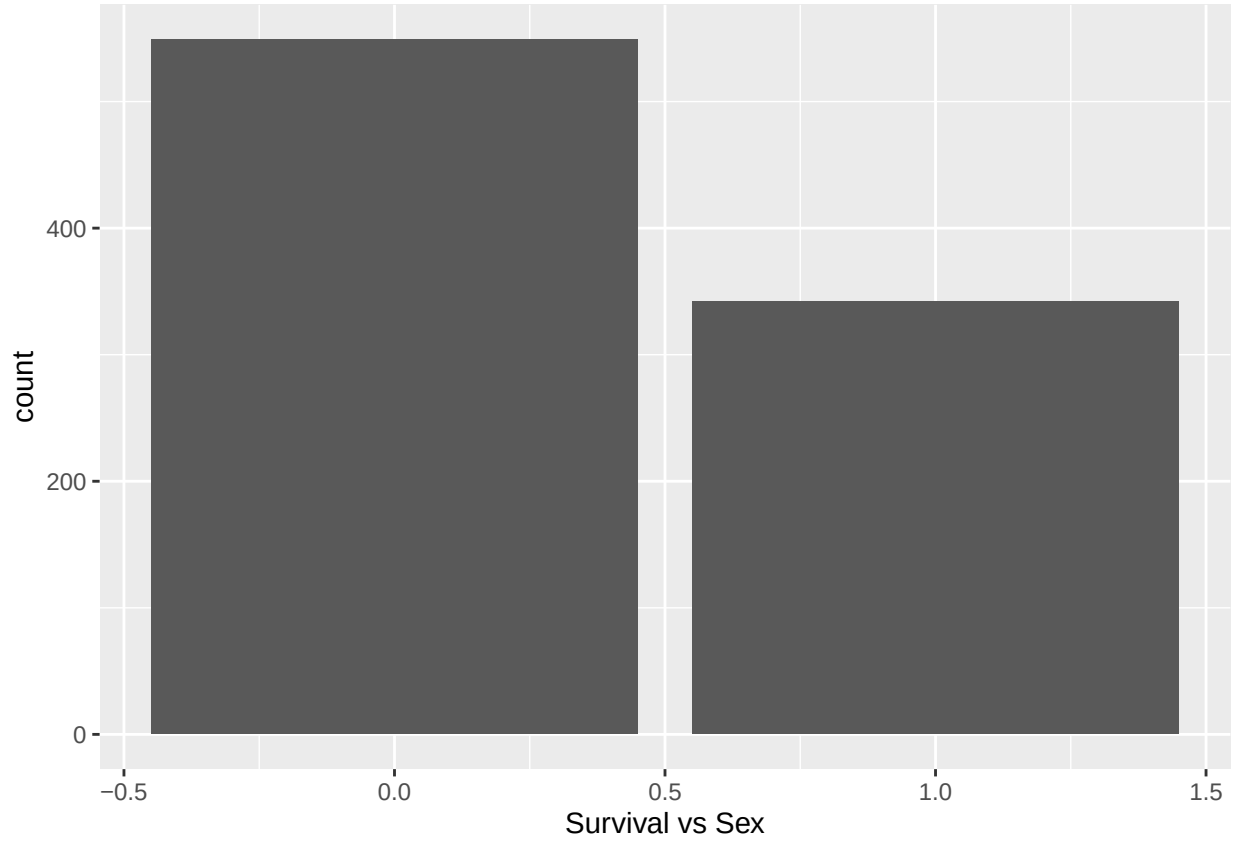
Representació dels resultats a partir de taules i gràfiques.

```
titanic %>% count(Sex)
```

```
## # A tibble: 2 x 2
##     Sex     n
##   <int> <int>
## 1     0   577
## 2     1   314
```

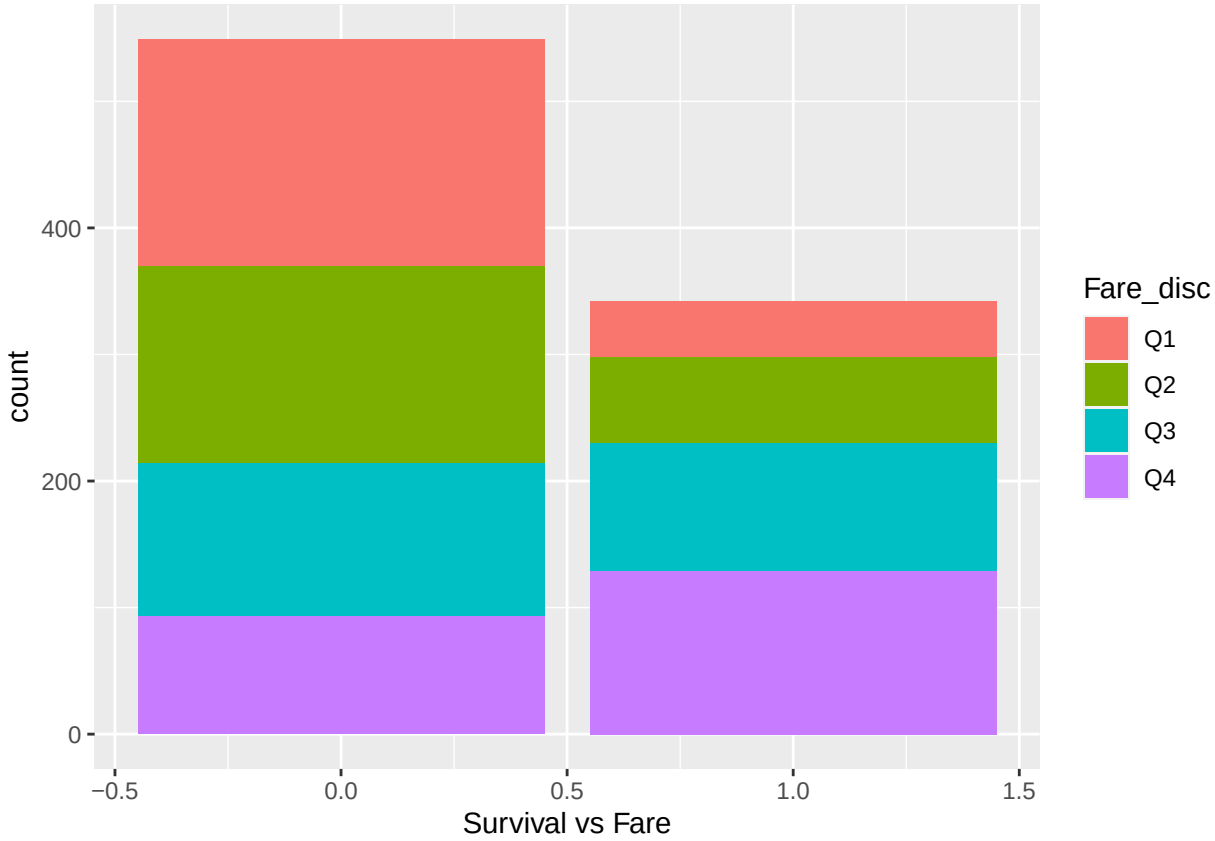
```
# 0 - Male, 1 - Female

#Survived by Sex
ggplot(titanic,aes(x=Survived, fill=Sex))+
  geom_histogram(stat = "count")+
  labs(x = "Survival vs Sex")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
#Survived by Fare Discretized
ggplot(titanic,aes(x=Survived, fill=Fare_disc))+
  geom_histogram(stat = "count")+
  labs(x = "Survival vs Fare")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



El gràfic amb les franges/banding de Fare ens diu clarament que dels que sobreviuen (dreta) la majora estan a les franges altes de preu de passatge, aspecte contrari dels que NO sobreviuen.

6. Conclussions

Resolució del problema. A partir dels resultats obtinguts ¿quines son les conclussions? ¿els resultats permeten respondre el problema?

1. Sobre la pregunta sobre si el preu del passatge dels que sobreviuen és més gran que el preu dels que moren hem plantejat el següent contrants de hipòtesis

$$\begin{cases} H_0: & \mu_0(Faresobreviuen) - \mu_1(Faremoren) = 0 \\ H_1: & \mu_0(Faresobreviuen) - \mu_1(Faremoren) > 0 \end{cases}$$

a on establint un nivell de significació de 0.05 **el p_value obtingut és ínfim i per tant més petit que 0.05, així que REBUTGEM la H0**, i això vol dir en que estem d'acord amb el postulat de la hipòtesis alternativa pel que fa al preu del passatge dels que sobreviuen és més gran que el preu dels que moren.