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

El dataset conté les dades dels passatgers del creuer titànic que va naufragar a causa d’un iceberg. El dataset és interessant perquè esta fraccionat en una part de traint i test per tal d’entrenar models de predicció que puguin predir quins factors van fer que alguns passatgers sobrevisquessin el titànic. El data set conte els següents camps:

* PassangerId: número identificatiu del passatger
* Survived: Si va sobreviure (1) o no (0)
* Pclass: Amb quina classe viatjava
* Name: Nom del passatger
* Sexe
* Edat
* Sibsp : Numero de germans i/o parella que tenien a bord
* Parch : Número de pares i/o fills que tenien a bord
* Ticket: numero de ticket
* Cabin: numero de cabina
* Embarked: Port d’embarcament a la nau

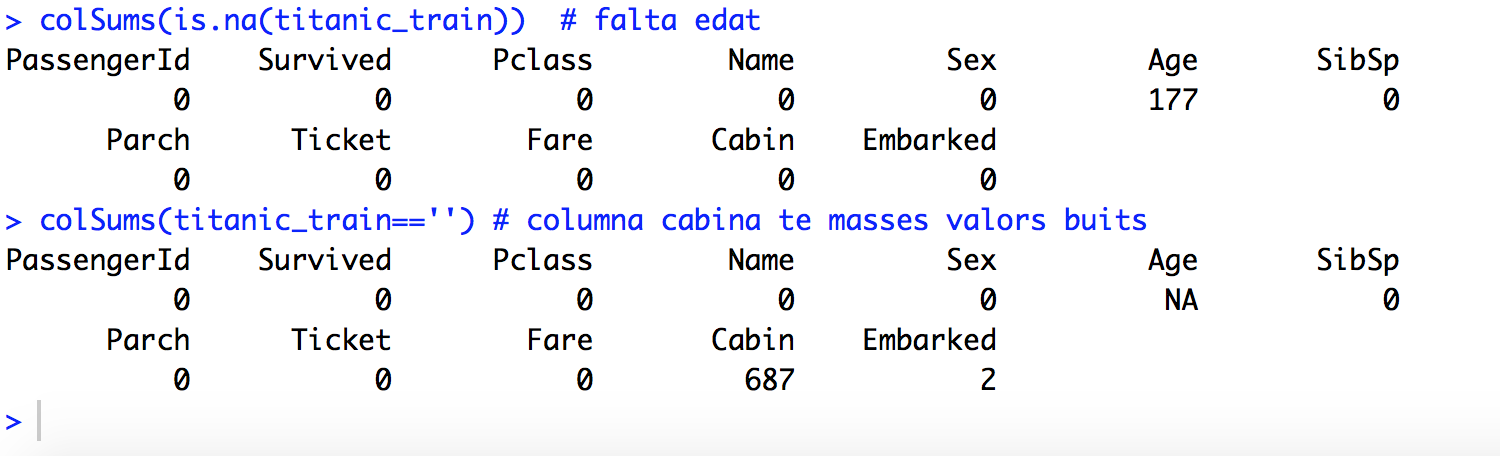
1. **Selecció de dades**

De les variables exposades amb anterioritat per tal de fer l’anàlisi prescindim de PassanferId, ticket i name ja que segur que no ens aporten cap mena d’informació rellevant per determinar si el passatger va sobreviure o no

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

Observem que per una banda tenim 177 registres de la edat que són nulls i per l’altra banda tenim 687 registres que no contenen cabina i 2 registres que no contenen port d’embarcament. Les estratègies a seguir són diferents:

* Edat: Val la pena aplicar o bé la mitjana o la mediana per tal de tenir valors representatius per a la edat. En aquest cas s’ha aplicat la mediana per tal de tenir valors menys afectats per els extrems.
* Cabina: Tenint en compte que el dataset té 891 entrades els 687 registres que falten representen un 77% de les dades. El més prudent és eliminar aquest camp ja que no pot aportar informació valuosa
* Port d’embarcament: Al tractarse d’una variable catergorica amb tants pocs registres buits el millor es posar el port d’embarcament amb més registres a les dades buides.

****

* 1. **Identificació de valors extrems**

Mirant a les dues variables contiunes que tenim observem coses diferents. Primerament per l’edat veiem que tot i tenir algun valor extrem la distribució sembla ser prou realista i per tant s’ha decitat no tocar res:

****

Per altra banda la varia fares (tarifes) s’ha vist que hi ha valors extrems que fa que la variable quedi desajustada:



Ens aquest cas al tractar-se de pocs valors el que s’ha decidit és que els valors més grans de 300 seran iguals a 300. El nou histograma queda de la següent manera:

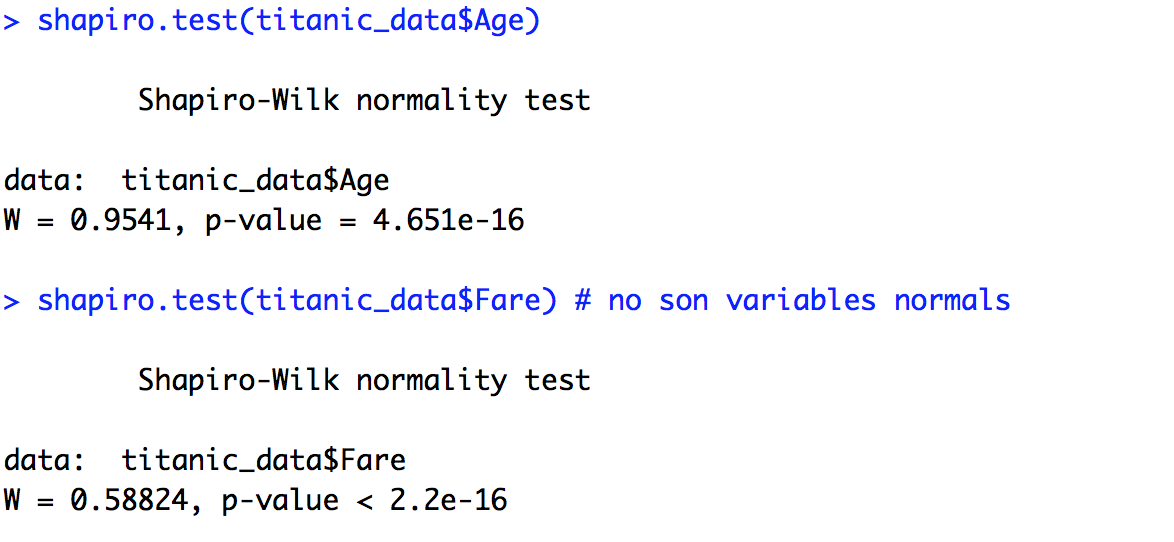
****

1. **Anàlisi de dades**
   1. **Selecció de dades**

Les dades seleccionades seran les d’entrenament tal i com s’explica en els apartats 1 i 2

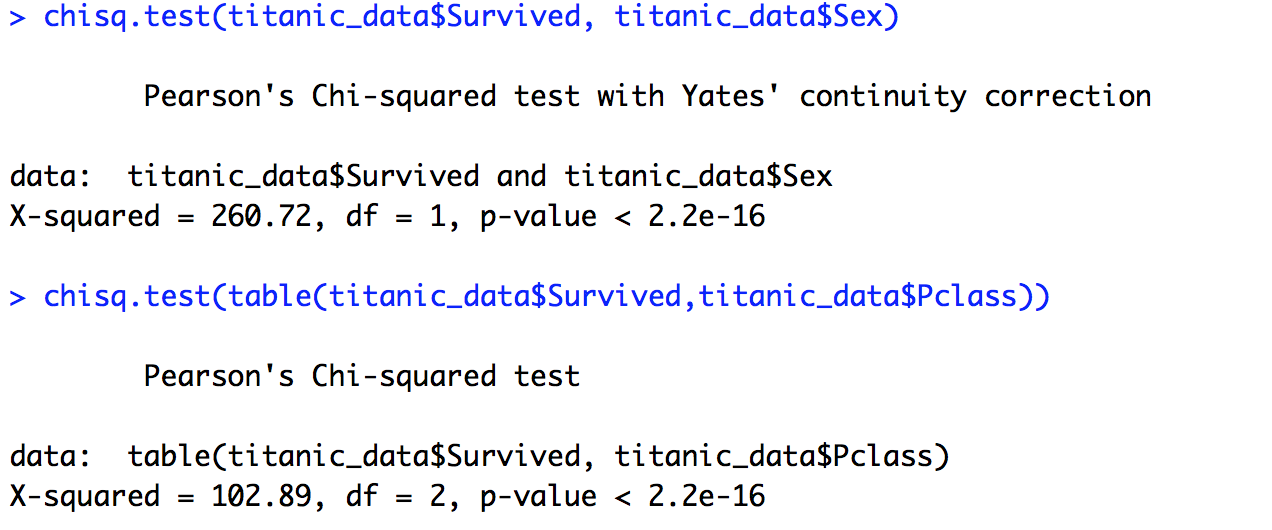
* 1. **Comprovació de la normalitat**

Les úniques dades continues (edat i tarifa) no són normals. Això ho hem mirat amb un test de shapiro



* 1. **Aplicació de proves estadístiques**

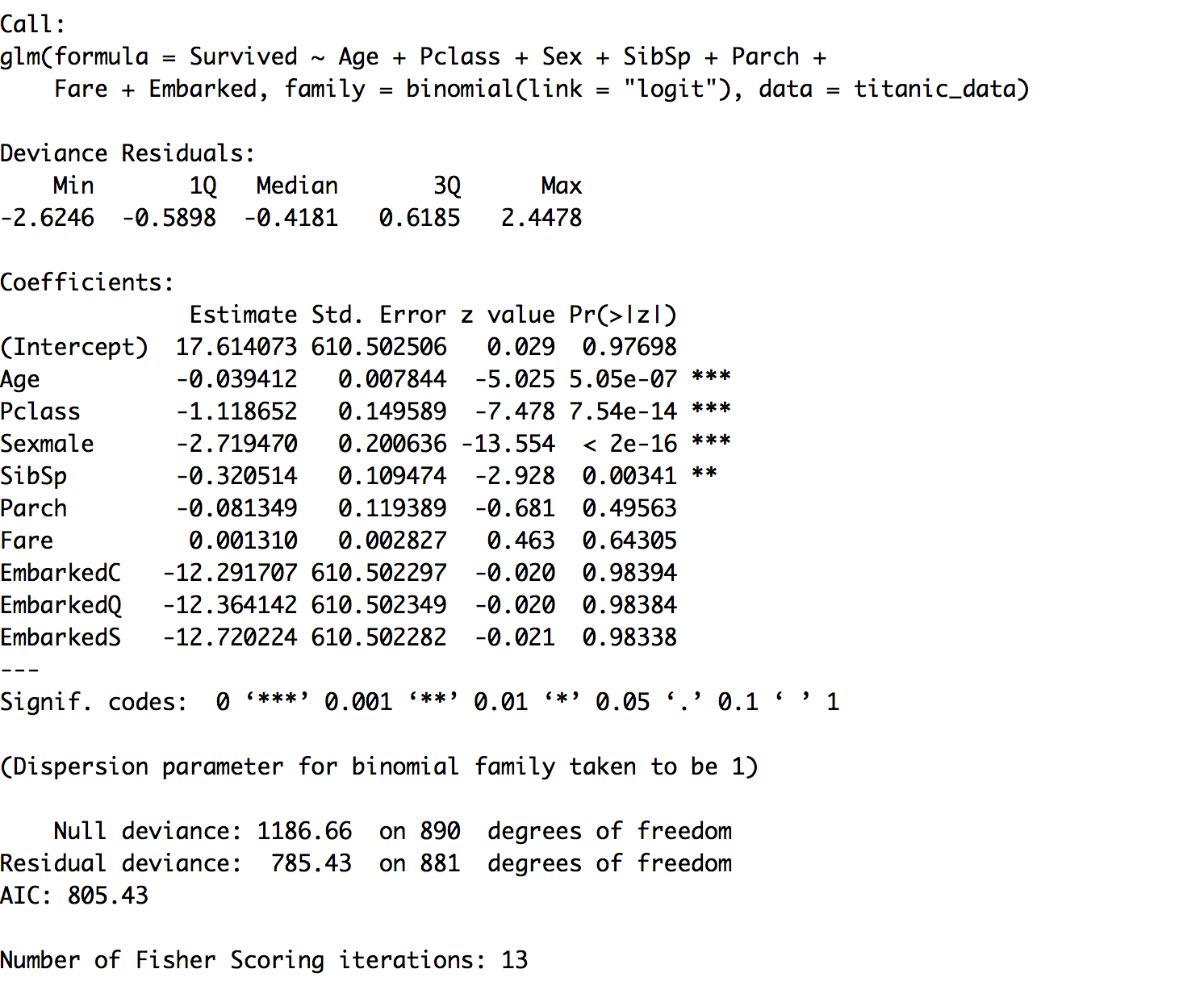
Abans de provar de fer un model predictor verificarem dos grans tòpics al voltant de la tragèdia del titànic. Aquests són que les dones es van salvar i que també ho van fer els passatgers que anaven amb primera classe. Per comprovar això hem aplicat un test de chi quadrat per comprovar la correlació entre aquestes variables i si el passatger va sobreviure o no.



Veiem que en els dos casos el p-valor és molt baix. D’aquest fet en podem extreure que la dependència de les dues variables anteriors són molt elevades i que per tant els tòpics es proven certs.

1. **Anàlisi de dades**

Al marge de les correlacions anteriors s’ha realitzat un model de regressió lineal i els resultats han sigut els següents:



Veiem que els valors residuals de desviació són molt elevats. Aquest model no és prou bo per explicar la supervivència dels passatgers del titànic

1. **Els resultats permeten respondre el problema?**

Els resultats obtinguts no ens permeten resoldre el problema. Per tal de resoldre aquest problema s’haurien d’aplicar altres tipus de models que ens permetessin respondre a les preguntes proposades. Un bon exemple seria una xarxa neuronal.