PRACTICA 2

Neteja i validació de les dades

Jordi Orriols i Jordi Aballó

Dels dos datasets proposats per escollir (Red Wine Quality I el del Titanic) hem trobat mes interesant treballar amb el del Titanic, ja que si mes endavant tenim els coneixements podria ser interesant participar en la competició per tal de aprendre encara més.

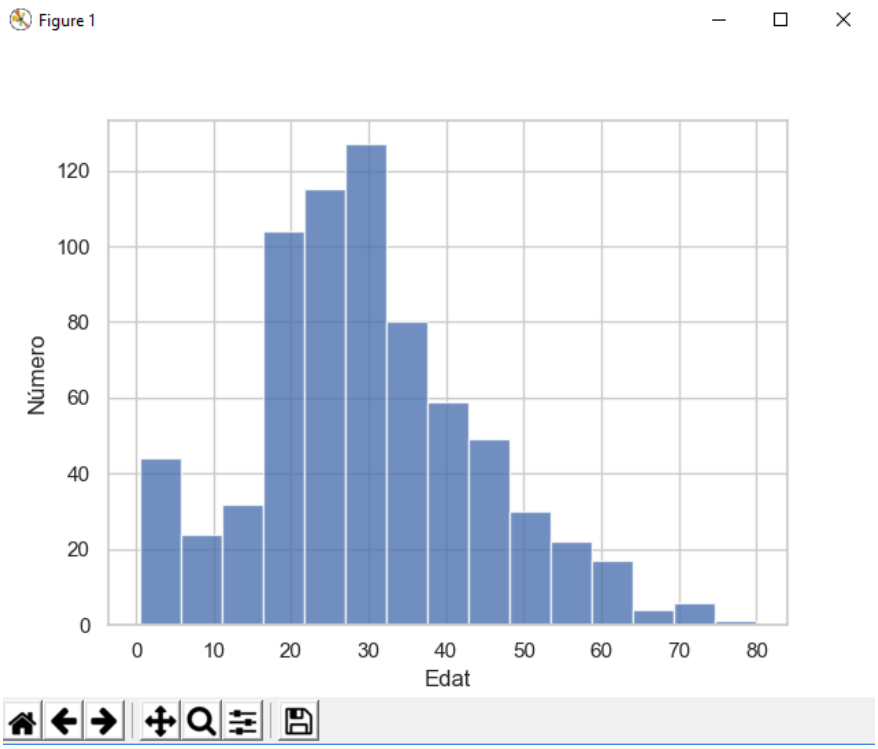
El dataset que tenim es important ja que recompila dades dels passatgers a bord del titànic i es composa dels següents camps:

- Survival: un binari que pot ser 0 en el cas que no hagi sobreviscut o 1 que significaria que si
- Pclass: es la classe dels passatgers, que pot ser 1a, 2na o 3a classe.
- Sex: el sexe del passatger
- Age: edat que te en anys
- Sibsp: numero de familiars directes que te a bord del vaixell
- Parch: numero de pares o fills que te a bord
- Ticket: numero del tiquet
- Faré: la tarifa de cada passatger
- Cabin: el numero de cabina
- Embarked: ens descriu el port d'embarc, tenim les següents opcions (C = Cherbourg, Q = Queenstown, S = Southampton)

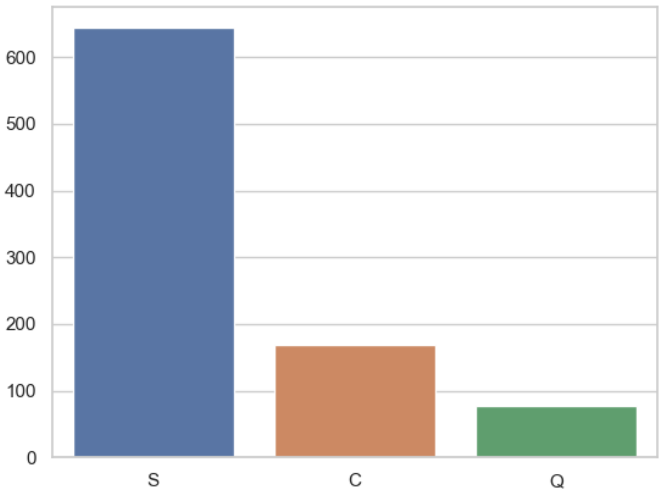
Per tal de analitzar les dades, hem anat comentant el codi per tal de que cada execució que fem es pugui interpretar

Per tal d'observar el data set podem fer un primer sampling mostrant totes les entrades nulls que tenim d'edat (la primer mètrica) i veiem que tenim 177 Valors a null.

Tot seguit, mirem que respecte el total d'entrades tenim un 20% dels valors en blanc, per tant pot ser interssant veure distribuït per l'edat el numero de persones:



Veiem que no tenim un valor predominant, així que farem servir la mitja que es de 28 anys per tal de afegir els valor que tenim en blanc i poder utilitzar aquells registres per altres coses.

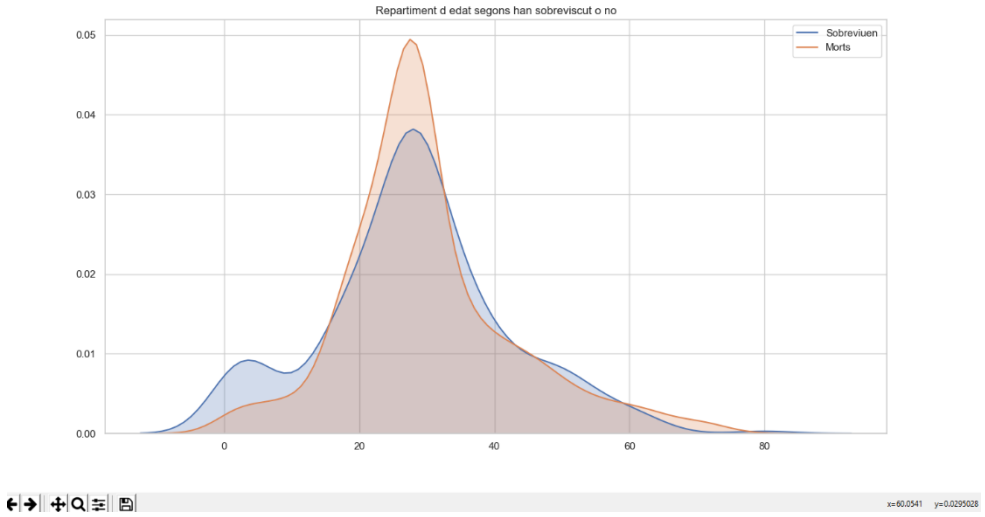


Si fem el matiex per el numero d'embarcaments, veiem que la majoria de gent ha embarcat a Sourthhampton i per tant, impotarema als 2 valors que nulls el valor de S.

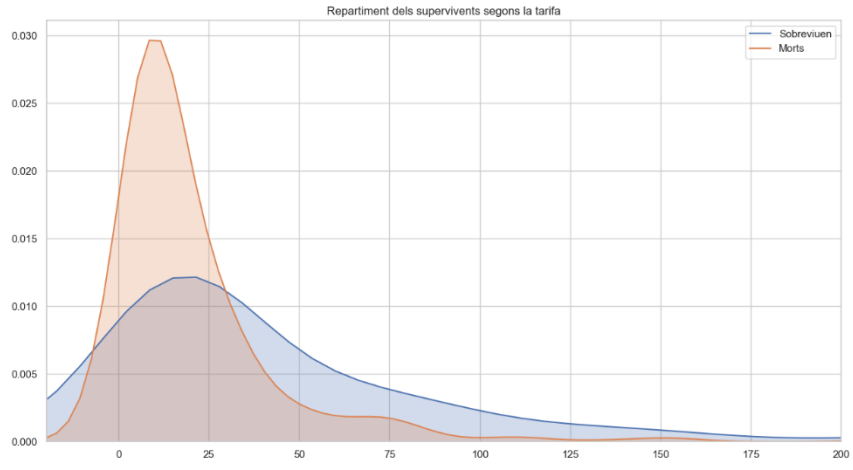
Tambe en el codi podem veure comentat totes les conclusion com que per exemple a cabina com a variable no ens serveix fer aquest últim mètode de data quality perquè, tal i com hem vist a teoria, hi ha masses valors nulls i no ens serviria per fer el imput, per tant farem un drop de aquella informació.

Per tal de veure els resultats de tant la creació de variables addicionals com a flag com les variables categòriques es interssant de veure el resultat de com tenim les dades actualment.

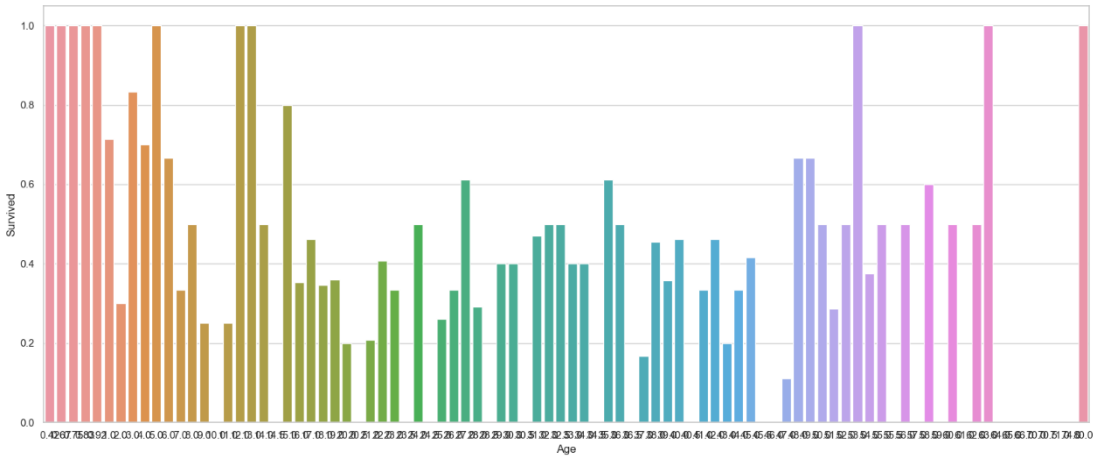
Tot seguit veiem la densitat de edat repartida entre la població i dividida en 2 series, els que han sobreviscut i els que no:



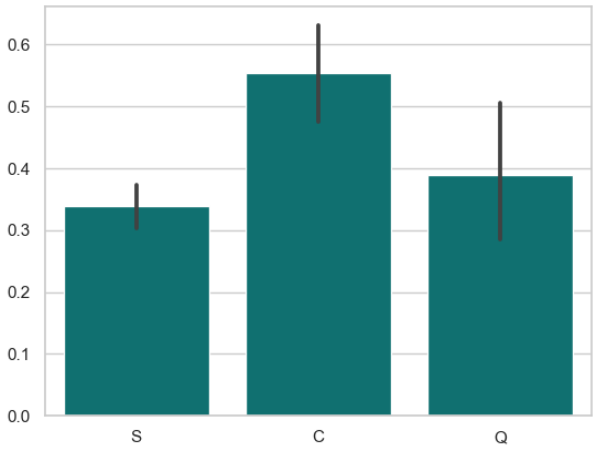
I també segons la tarifa que tenien, així podem començar a extreure conclusions, ja que els que van pagar una tarifa inferior veiem que la majoria han acabat morts:



Tot seguit he volgut veure com es distribuïa els supervivents respecte la edad que tenien, ja que veiem que tenim molts menors realment.



Finalment un gràfic que mostrem el numero de supervivents depenent del port on han embarcat.



Podem concloure que passatgers que van pujar a Cherbourg sembla que tenen un surival rate mes alt, i per altre banda, els de Southhampton mes baix. Podríem veure ja que també esta directament relacionat amb la classe o també amb el ordre d'assignació de les habitacions, ja que al embarcar primer poder tenien les habitacions mes a prop de coberta.

Contribuciones Firma

Investigació Prèvia	Jordi Orriols, Jordi Aballó
Redacció de les respostes	Jordi Orriols, Jordi Aballó
Desenvolupament del codi	Jordi Orriols, Jordi Aballó