EXPLORACIÓN DATA FRAME TITANIC

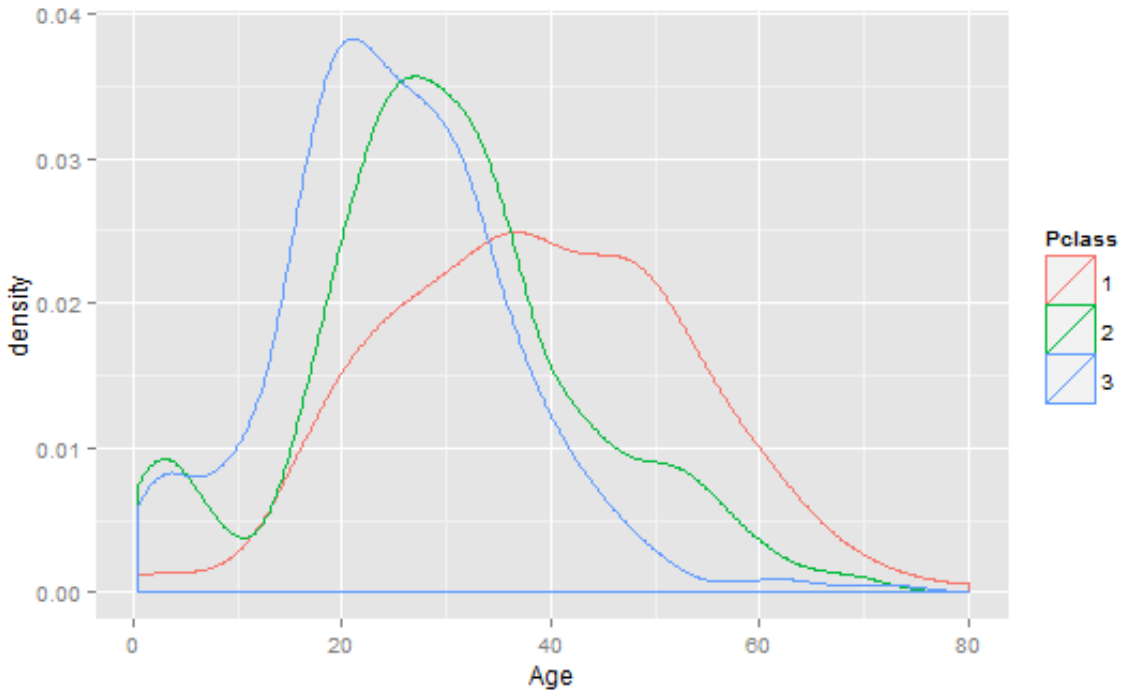
Federico Fernández
Jorge Ramírez

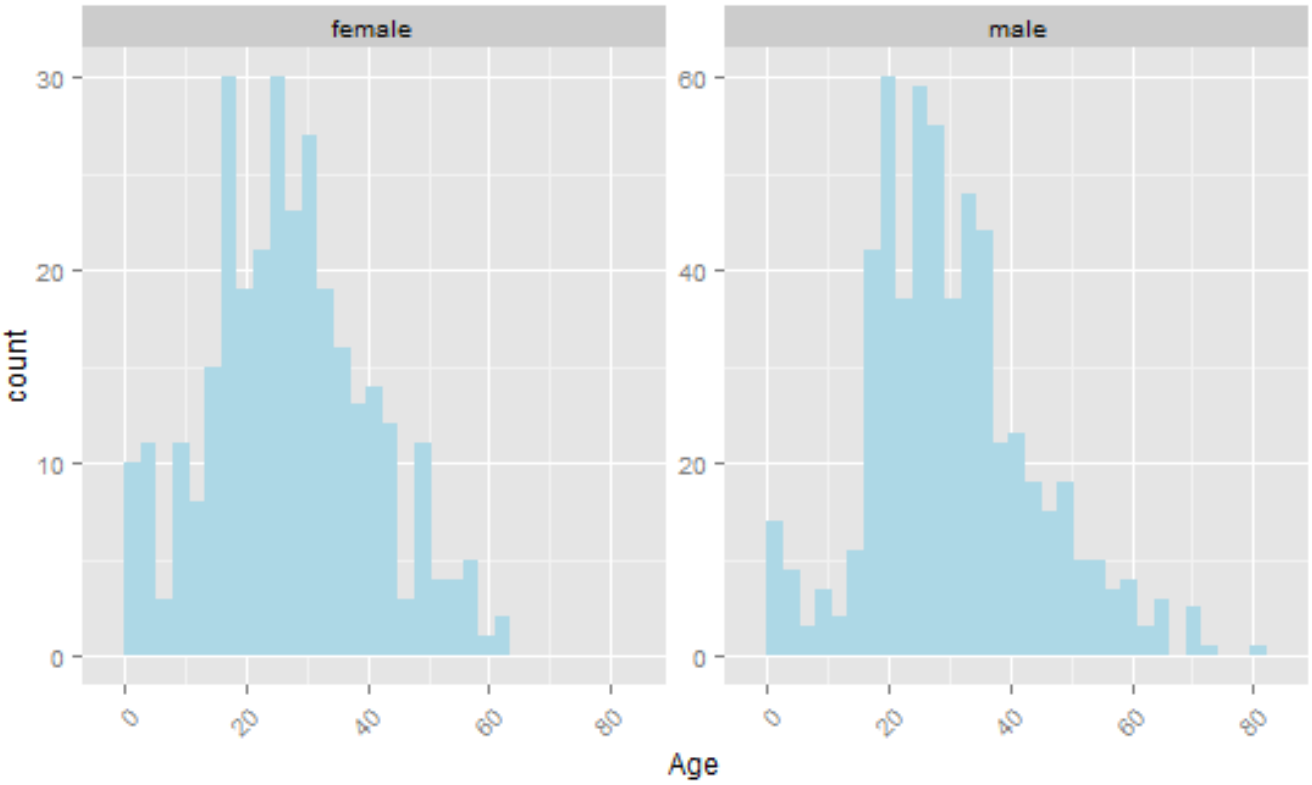
Summer Camp 2014

Data Adaptation

Antes de poder analizar los datos, ha sido necesario realizar un proceso de adaptación de los mismos. El proceso que se ha seguido es el siguiente:

1. Se han convertido a tipo factor las columnas correspondientes a la clase en que viajan los pasajeros (*passenger class*, *Pclass*) y si han sobrevivido o no (columna *Survived*).
2. Hay entradas del *dataframe* cuyo campo *Edad* (*Age*) estaba en blanco. Para darles un valor que introduzca el menor error posible, se ha optado por distinguir entre *Pclass* y *Sex*, ya que son las variables en función de las que más varía la edad, como muestran las siguientes gráficas:





3. De este modo, se ha asignado a cada grupo de las seis posibles combinaciones entre *Pclass* y *Sex*, un valor de *Age* correspondiente a la media de ese grupo más la desviación típica multiplicada por un número aleatorio entre -1 y 1 (siguiendo una distribución uniforme).
4. Por último, se han eliminado las entradas del *dataframe* en las que el campo *Embarked* estaba en blanco (solo eran 2 entradas en una muestra de 891 observaciones), y se ha refactorizado dicho campo.

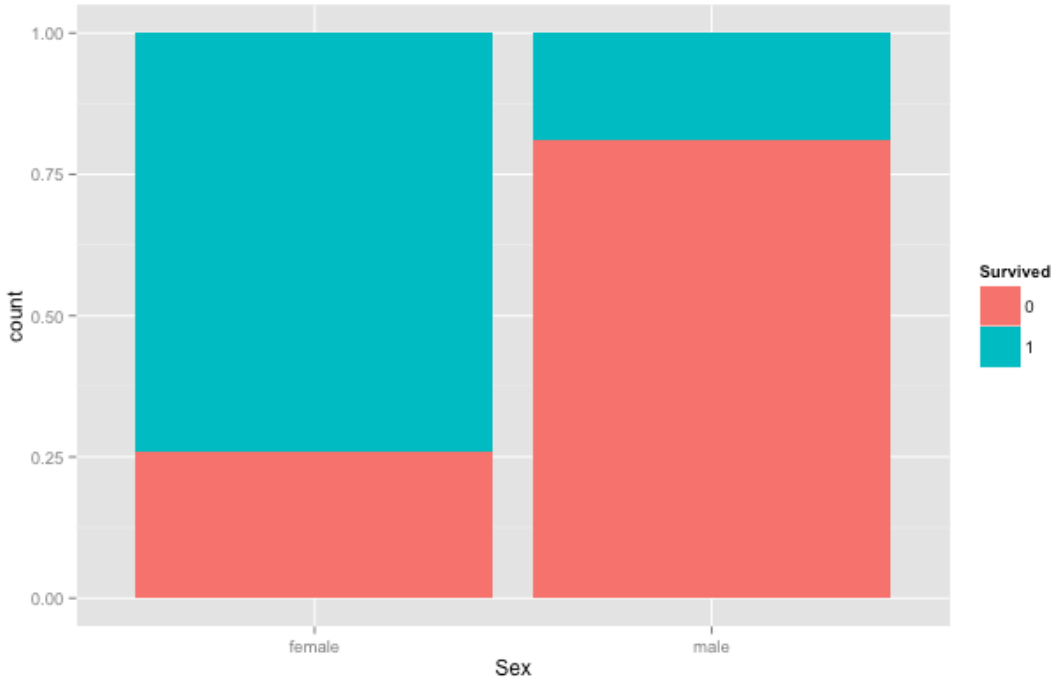
Análisis de la exploración

Realizamos un análisis de los datos con el objetivo de identificar las variables clave de cara a la supervivencia de los pasajeros.

Observamos como el sexo y la clase en la que viajaban los pasajeros son los factores más determinantes en su supervivencia y por ello realizaremos un estudio más a fondo.

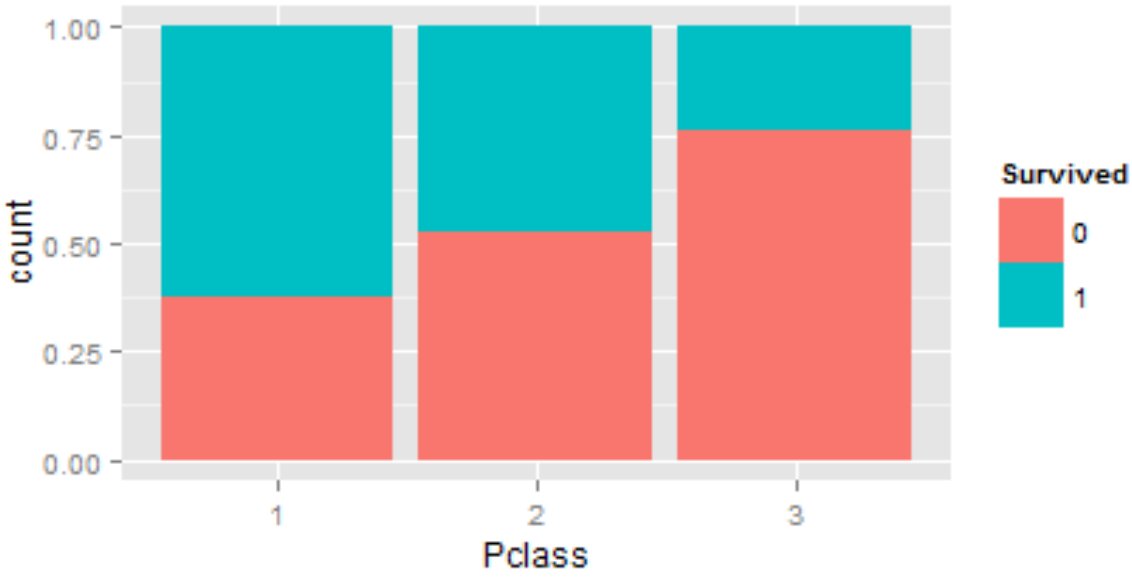
SEXO

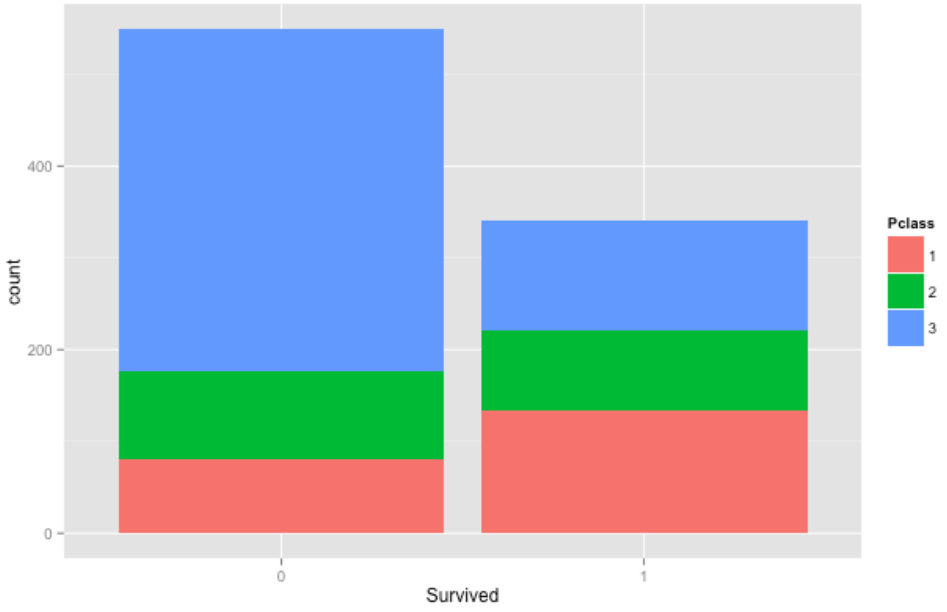
El sexo aparece como un variable clave para determinar la supervivencia de los pasajeros, observándose una tasa de mortalidad en los hombres(77%) mucho mayor que en las mujeres(25%).



CLASE

También se observa como variable determinante en la supervivencia la pertenencia a las distintas clases. La primera clase presenta una tasa de mortalidad del 37.5% , la segunda del 50% y la tercera del 75%.

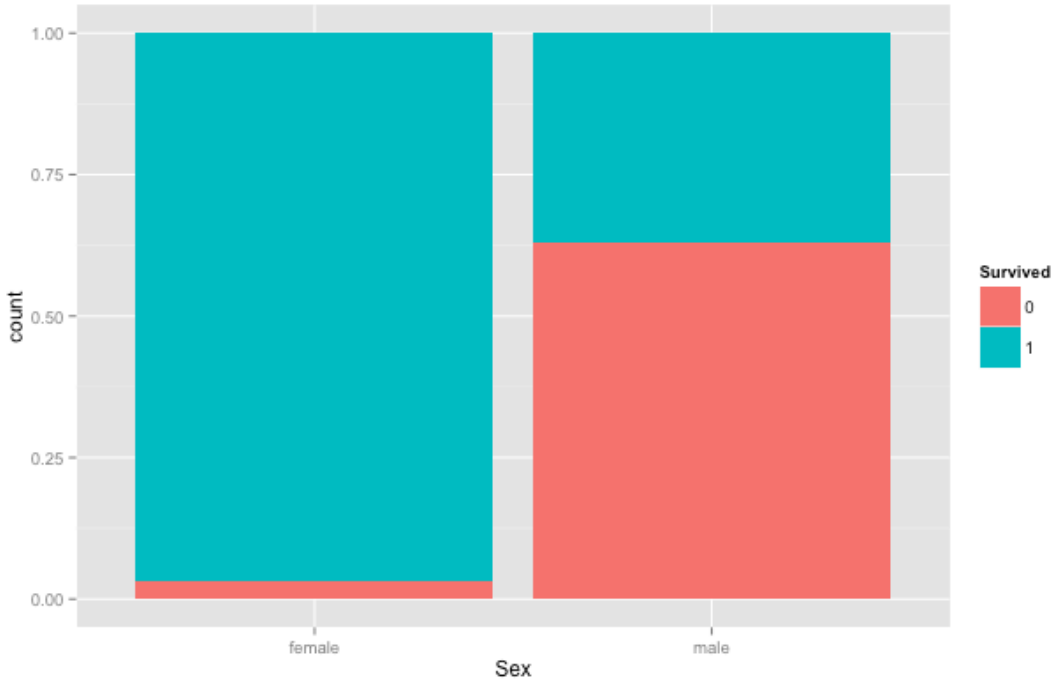




Para una mejor exploración estudiaremos distintas variables para cada una de las clases.

1º

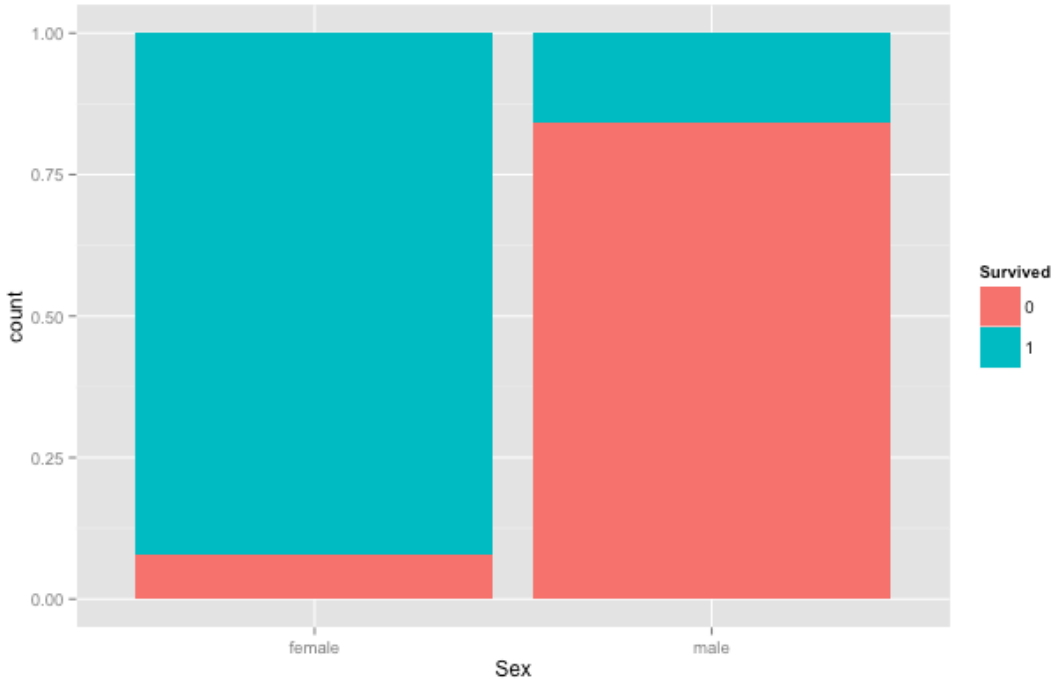
Comenzamos analizando la primera clase:



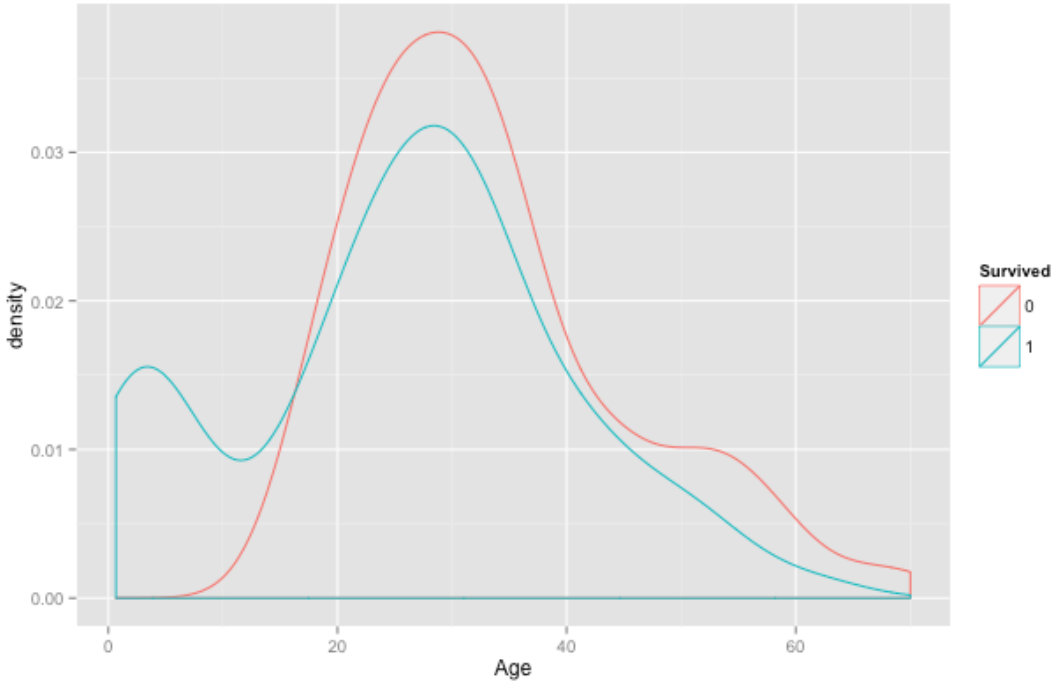
De nuevo confirmamos como variable clave el sexo de los pasajeros viendo como prácticamente el 100% de las mujeres en primera sobrevivieron.

2º

Con la segunda clase ocurre lo mismo que en con respecto al sexo.

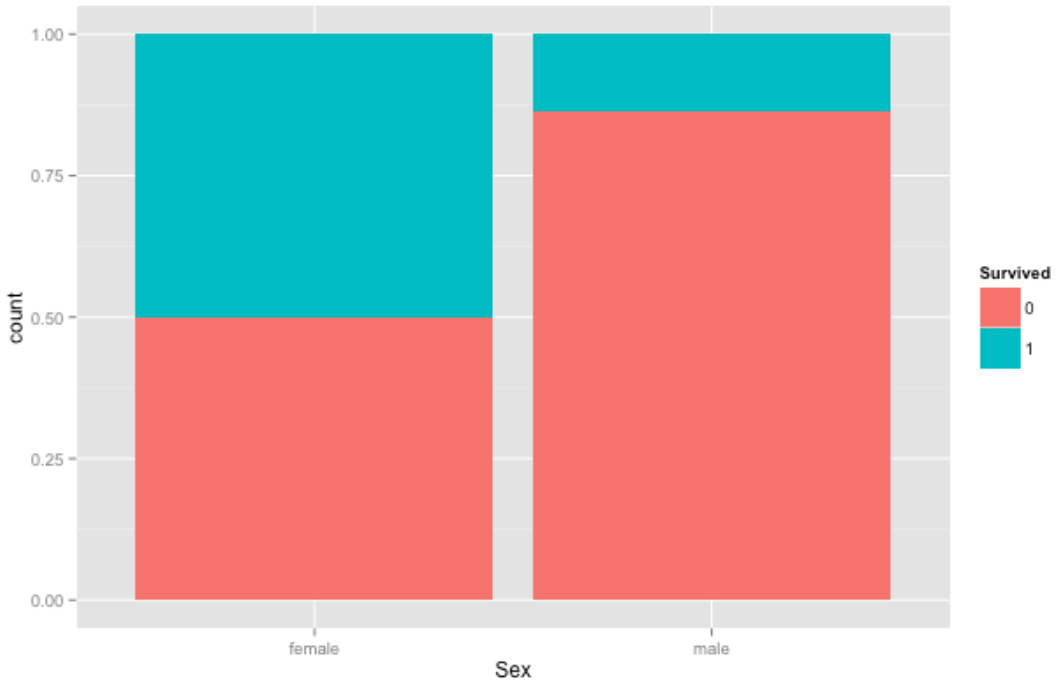


Con respecto a la edad, descubrimos un pico de supervivencia fuera de lo común que corresponde con los niños.



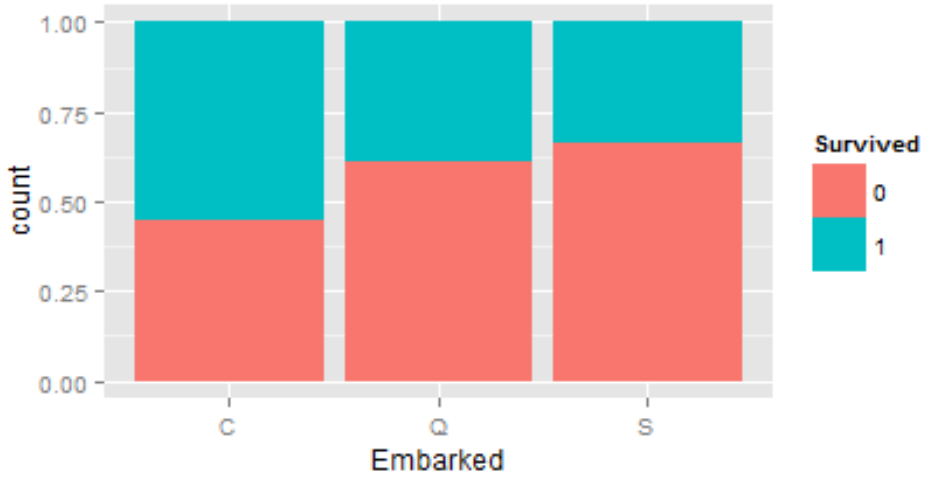
3º

Por último en la tercera clase tenemos datos similares a los encontrados en las otras clases, pero con una tasa de mortalidad mucho mayor propia de esta clase.



PUERTO

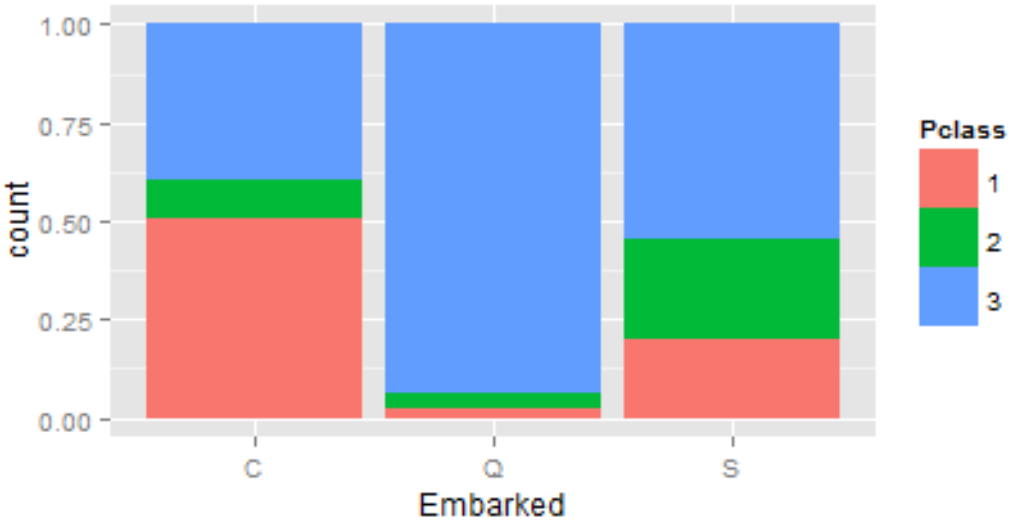
Parece que también existe una relación entre la supervivencia de los pasajeros (variable *Survived*) y el puerto en que estos embarcaron (variable *Embarked*).
Como muestra la siguiente gráfica, el porcentaje de pasajeros que no sobreviven y habían embarcado en Southampton (S) es mayor que los que no sobrevivieron y habían embarcado en Queenstown (Q), y este, a su vez, es mayor que los que no sobrevivieron y habían embarcado en Cherbourg (C).



Pensamos que, quizá, existe una relación entre la clase en que viajaban los pasajeros (*Pclass*) y el puerto en que embarcaron (*Embarked*).

Si esto es así, igual que *Embarked* está relacionado con *Survived*, tendría sentido lo que ya se ha demostrado: que existe una relación entre *Pclass* y la supervivencia o no de los pasajeros (*Survived*).

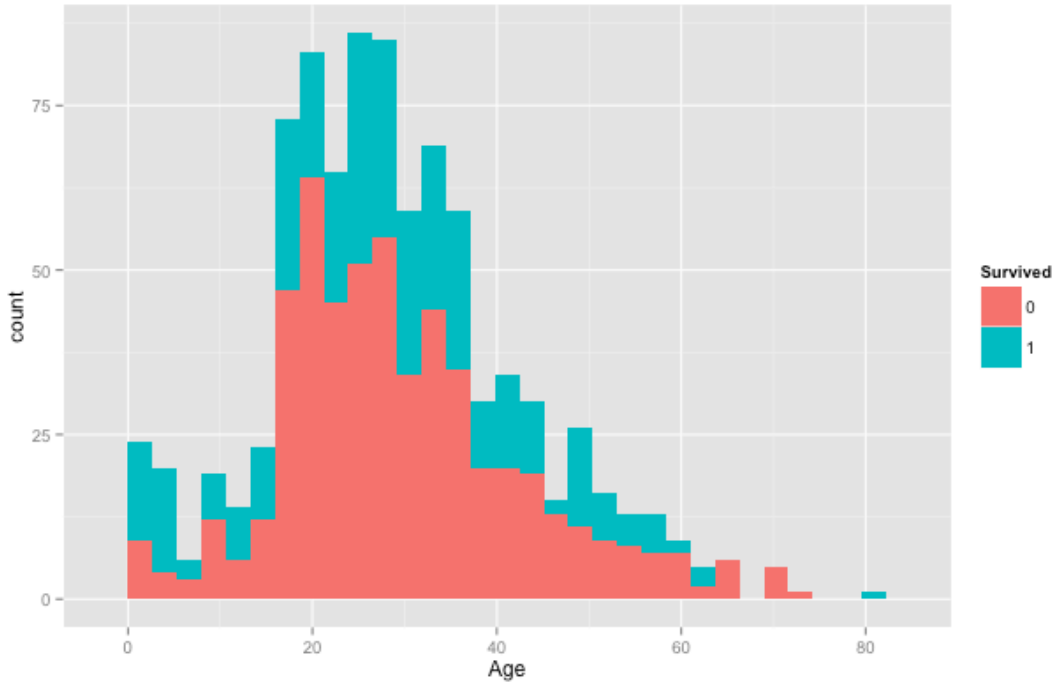
Como muestra la siguiente gráfica, esto en efecto es así, y existe una relación más que notable entre las variables *Pclass* y *Embarked*.

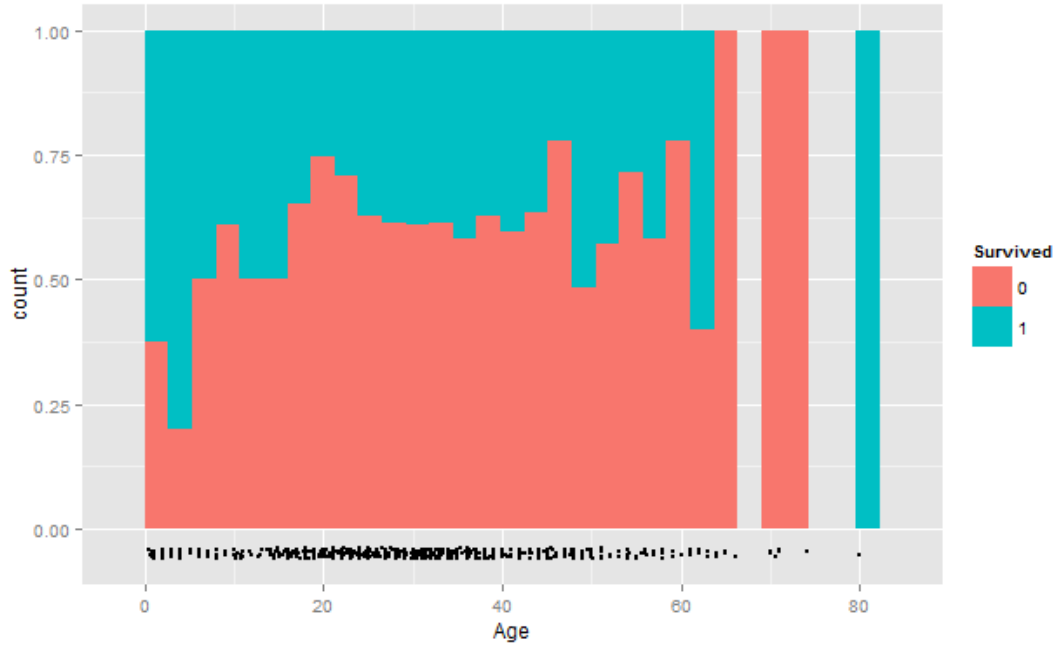


Los pasajeros que embarcaron en Queenstown eran de tercera clase en más de un 90%. Por otro lado, los que embarcaron en Cherbourg lo hicieron en primera clase en un 50% de los casos. Sin embargo, los que embarcaron en Southampton pertenecían a tercera clase en más de la mitad de los casos.

EDAD

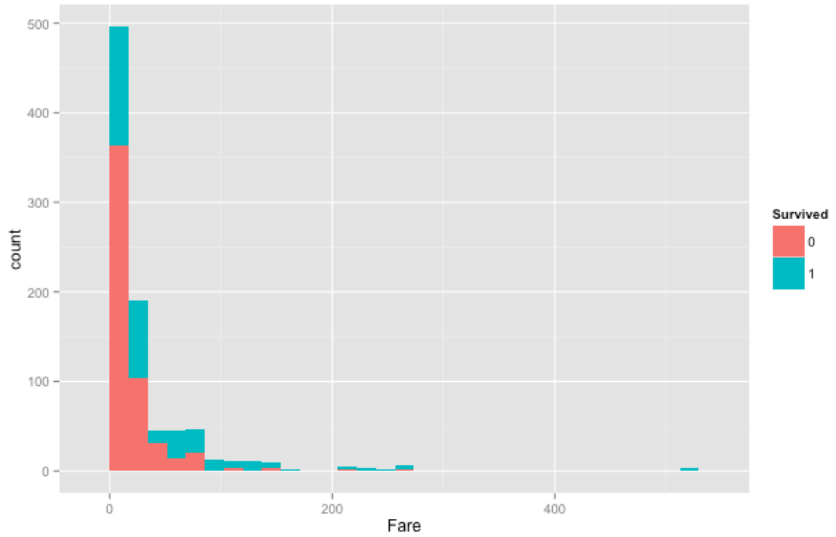
Con respecto a la edad la mayoría de los pasajeros pertenecen a un grupo de edad, el cual presenta el mayor número de supervivencias y de muertes. Como destacable se observa el pico de supervivencia por parte de los niños que será una variable a tener en cuenta.



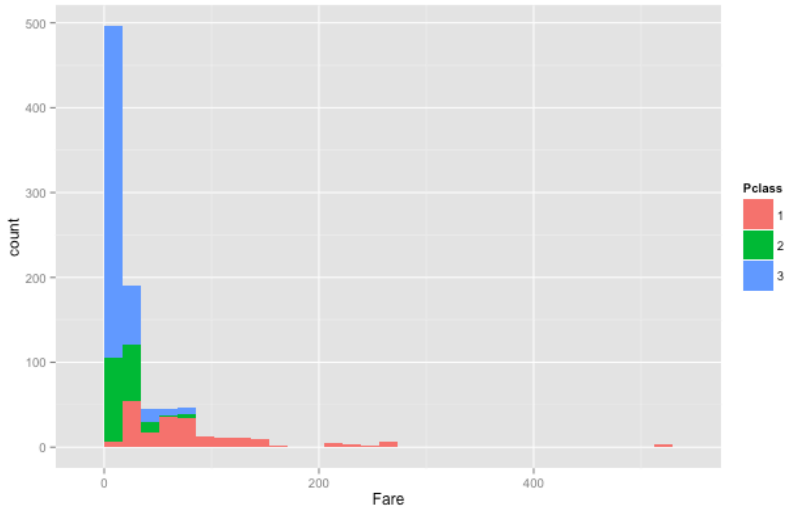


PRECIO DEL BILLETE

Parece que existe una relación entre si un pasajero sobrevivió o no y el precio que había pagado por su billete. Lo muestra la siguiente gráfica:

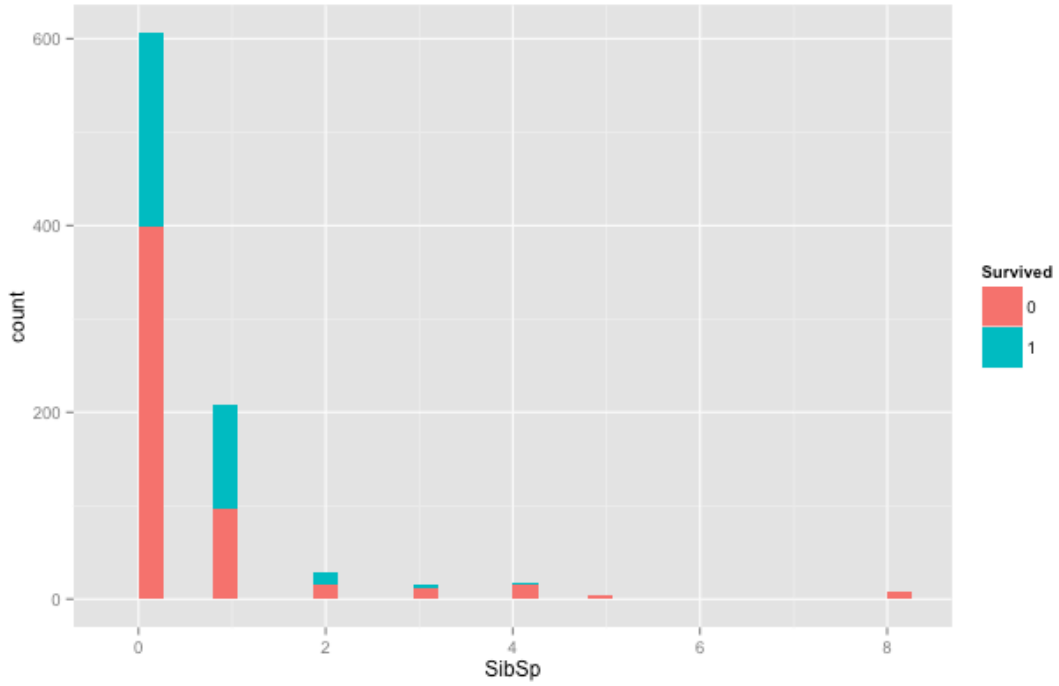


Sin embargo, esto no es más que el reflejo de la dependencia de la supervivencia con la clase en que viajaban. Precisamente, en la siguiente gráfica se observa que la zona correspondiente a tercera clase (color azul) se corresponde casi exactamente con el área de pasajeros que no sobrevivieron de la gráfica anterior (color rojo). Como ya se indicó anteriormente: la tasa de supervivencia es menor en tercera clase.



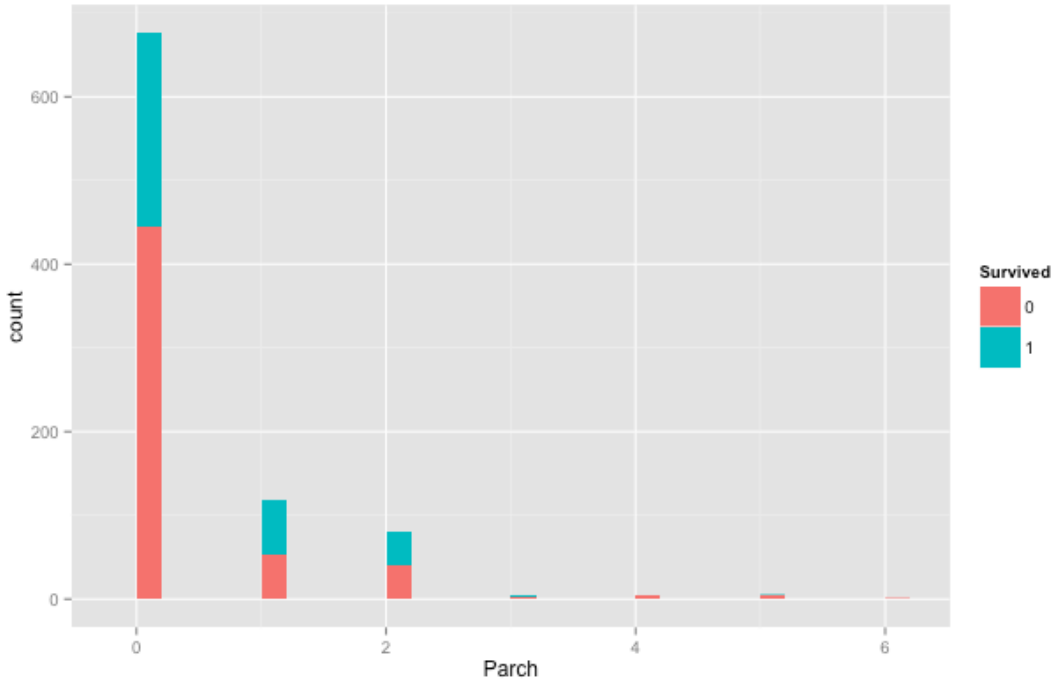
EXISTENCIA DE HERMANOS/ESPOSAS EN EL BARCO

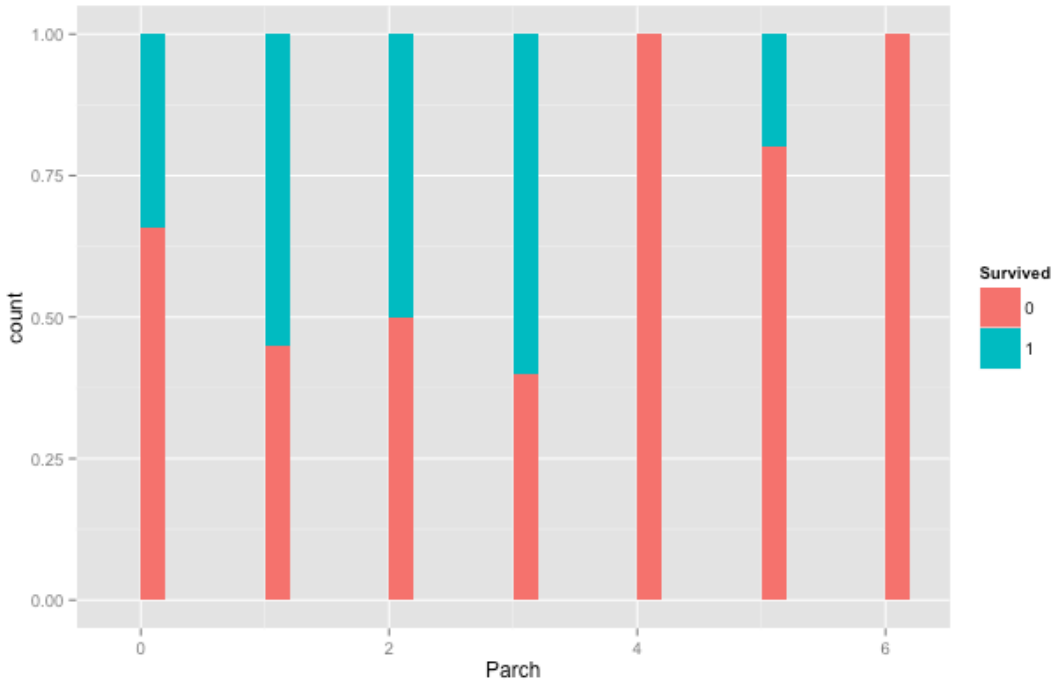
La variable de existencia de hermanos y esposas/maridos en el barco no resulta una variable que nos proporcione demasiada información ya que sus valores se encuentran cercanos al 50% en las distintas opciones.



EXISTENCIA DE PADRES/HIJOS EN EL BARCO

Por último exploramos la variable de existencia de padres o hijos en el barco y del mismo modo no la consideramos transcendente ya que sus valores se encuentran cercanos al 50% en las opciones que tienen un número de casos considerable.





Conclusiones de la exploración

Tras la exploración concluimos como factores determinantes para la supervivencia:

-El sexo debido a que las mujeres presentan una tasa de mortalidad mucho menor. Seguramente por la preferencia al tomar los botes de escape.

-La clase ya que la gente de 1º presenta menor tasa de mortalidad de nuevo al tener prioridad frente a las otras clases a la hora de salvarse.

-La edad ya que de la misma forma que con las mujeres los niños tienen prioridad para tratar escapar.