ACTIVIDAD 2 - UD2

Code

Actividad 2 - UD2

Carga el dataset **titanic** con el comando y haz un sumario - Transforma las variables que sean factores en factores. Si son ordinales usa *ordered* para crear factores ordenados. Recuerda usar lapply. - Representa en un qplot la edad frente a la tarifa, y en un segundo qplot lo mismo pero con la clave de la clase en la que viajaban. Represéntalo a su vez factorizándolo (i.e. aplicando faceting) por *Sex* y *Embarked* - Pinta un boxplot de la edad agrupado según *Sex* - Pinta un barplot que represente la supervivencia en cada *Pclass* coloreando las barras según esta - Pinta la supervivencia en función de la categoría *Sex*, ¿qué observas? - Pinta la supervivencia en función de la categoría *Pclass*, ¿qué se aprecia? - Crea un histograma de *Age*, ¿qué observas? - Crea una agrupación de los datos usando `dplyr` de *Sex* y *Pclass*. Haz sumarios de media, conteo y mediana. ¿Qué podrías decir de los resultados? ¿Hay suficiente muestra para sacar conclusiones en todas las categorías creadas al combinar la edad y la clase?

Cargamos primro las librerías necesarias

Hide

```
library(ggplot2)
library(tidyr)
library(dplyr)
library("RColorBrewer")
```

Hide

```
df <- read.csv("C:/Users/Usuario/Desktop/Máster_IMF/Ejercicios_R/train.csv", heade
r = TRUE, sep = ",",quote = "\"",comment.char = "",encoding = "UTF-8")
```

Hide

```
df
```

PassengerId <int>	Survived <int>	Pclass <int>
1	0	3
2	1	1
3	1	3
4	1	1
5	0	3
6	0	3
7	0	1
8	0	3
9	1	3

PassengerId <int>	Survived <int>	Pclass <int>
10	1	2

1-10 of 891 rows | 1-3 of 12 columns Previous **1** 2 3 4 5 6 … 90 Next

Hide

```
head(df)
```

	PassengerId <int>	Survived <int>	Pclass <int>
1	1	0	3
2	2	1	1
3	3	1	3
4	4	1	1
5	5	0	3
6	6	0	3

6 rows | 1-4 of 12 columns

Hide

```
str(df)
```

```
'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 (Floren
ce Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily Ma
y Peel)" ...
 $ 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" ...
```

Hide

```
summary(df)
```

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

NA's   :177
    Cabin             Embarked
 Length:891         Length:891
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Hemos decidido quitar los valores cero del dataframe

Hide

```
df <- na.omit(df)
```

Hide

```
df
```

	PassengerId <int>	Survived <int>	Pclass <int>
1	1	0	3
2	2	1	1
3	3	1	3
4	4	1	1
5	5	0	3
7	7	0	1
8	8	0	3
9	9	1	3

	PassengerId <int>	Survived <int>	Pclass <int>
10	10	1	2
11	11	1	3

1-10 of 714 rows | 1-4 of 12 columns Previous **1** 2 3 4 5 6 … 72 Next

Hide

```
NA
```

Hide

```
dim(df)
```

```
[1] 714  12
```

Hide

```
summary(df)
```

```
  PassengerId       Survived          Pclass          Name                Sex
Age             SibSp            Parch           Ticket               Fare
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:714         Length:714
Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   Length:714         Min.   :  0.0
0
 1st Qu.:222.2   1st Qu.:0.0000   1st Qu.:1.000   Class :character   Class :charac
ter   1st Qu.:20.12   1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Q
u.:  8.05
 Median :445.0   Median :0.0000   Median :2.000   Mode  :character   Mode  :charac
ter   Median :28.00   Median :0.0000   Median :0.0000   Mode  :character   Median
: 15.74
 Mean   :448.6   Mean   :0.4062   Mean   :2.237
Mean   :29.70   Mean   :0.5126   Mean   :0.4314                      Mean   : 34.6
9
 3rd Qu.:677.8   3rd Qu.:1.0000   3rd Qu.:3.000
3rd Qu.:38.00   3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.: 33.3
8
 Max.   :891.0   Max.   :1.0000   Max.   :3.000
Max.   :80.00   Max.   :5.0000   Max.   :6.0000                      Max.   :512.3
3
    Cabin             Embarked
 Length:714         Length:714
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Hide

```
colnames(df)
```

```
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"
"SibSp"       "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
```

Transforma las variables que sean factores en factores. Si son ordinales usa *ordered* para crear factores ordenados. Recuerda usar lapply.

Hide

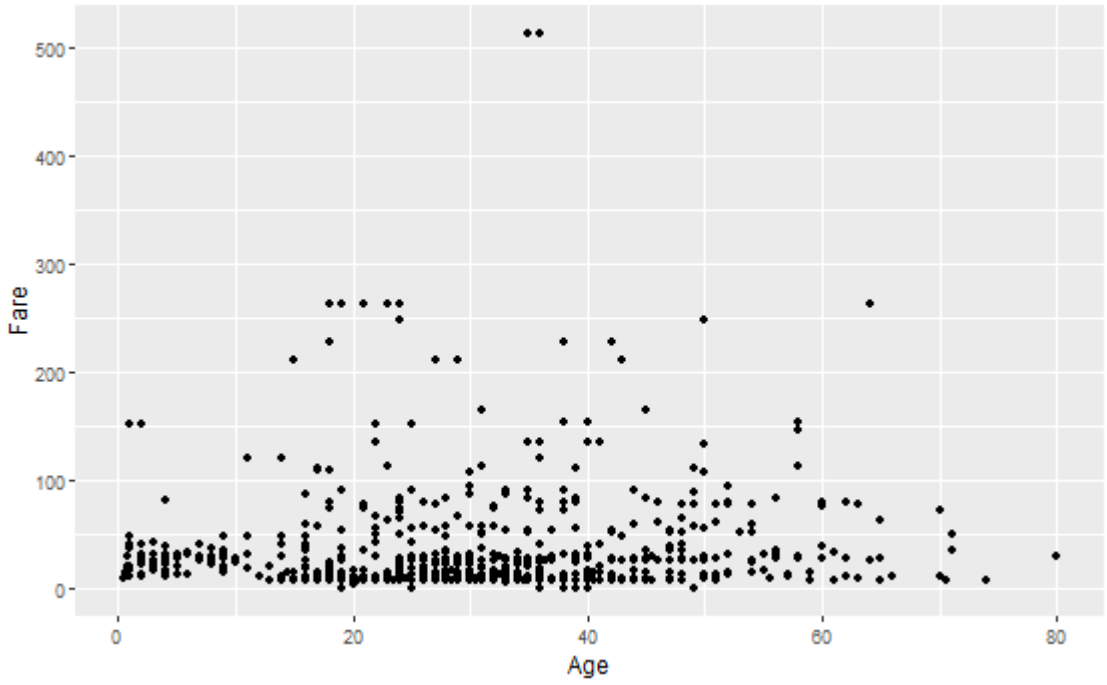
```
x<-df

columnas<- c("Survived","Sex" ,"Embarked","Pclass")

x[columnas] <- lapply(x[columnas], factor)

str(x)
```

```
'data.frame':   714 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 7 8 9 10 11 ...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 2 2 ...
 $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 1 3 3 2 3 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Floren
ce Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily Ma
y Peel)" ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 1 1 1 ...
 $ Age        : num  22 38 26 35 35 54 2 27 14 4 ...
 $ SibSp      : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch      : int  0 0 0 0 0 0 1 2 0 1 ...
 $ 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   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 4 4 4 2 4 ...
 - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
  ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
```

- Representa en un qplot la edad frente a la tarifa, y en un segundo qplot lo mismo pero con la clave de la clase en la que viajaban. Represéntalo a su vez factorizándolo (i.e. aplicando faceting) por *Sex* y *Embarked*

Hide

```
ggplot(data = df) +
  geom_point(mapping = aes(x = Age , y =Fare))
```



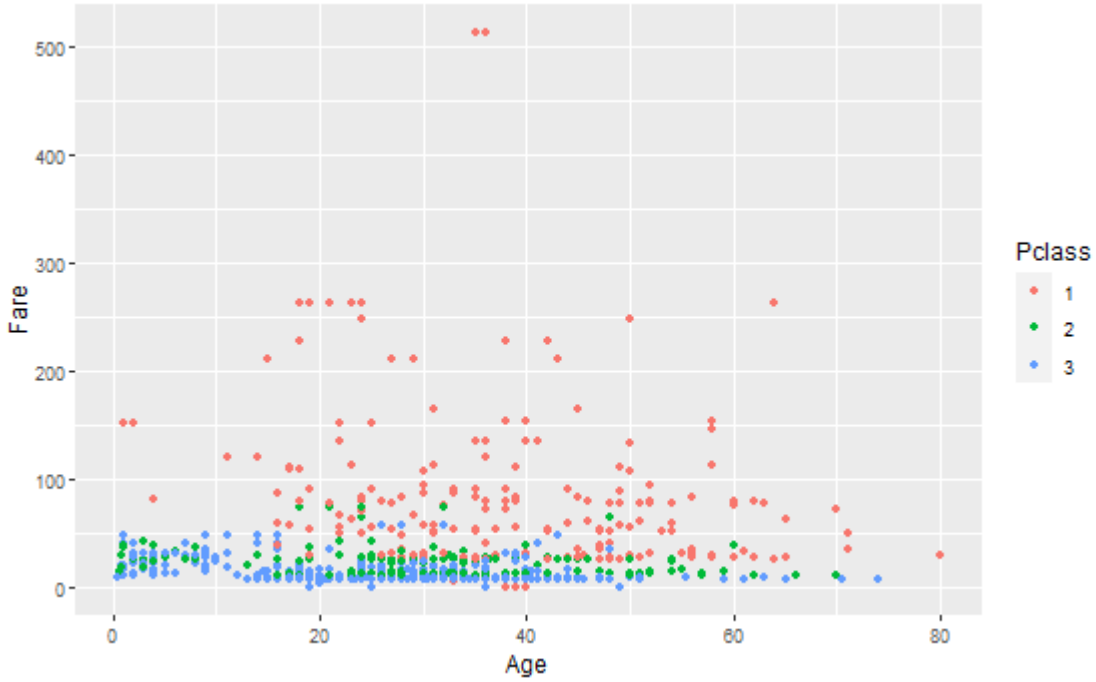
Hide

```
NA
NA
```

En el gráfico podemos observar que hay gran cantidad de personas que viajan en clase baja ya que el precio del billete es pequeño, la mayoría de los puntos estan abajo, donde fare es más pequeño, sin importar la edad. La edad no está relacionada con el precio del billete. Aunque nos da información de que hay mucha más gente viajando en tercera clase que en primera. Hay una nube de puntos más densa abajo, fare más pequeño.

Hide

```
ggplot(df, aes(x = Age, y = Fare, colour = Pclass))+ geom_point(size = 1.5)
```



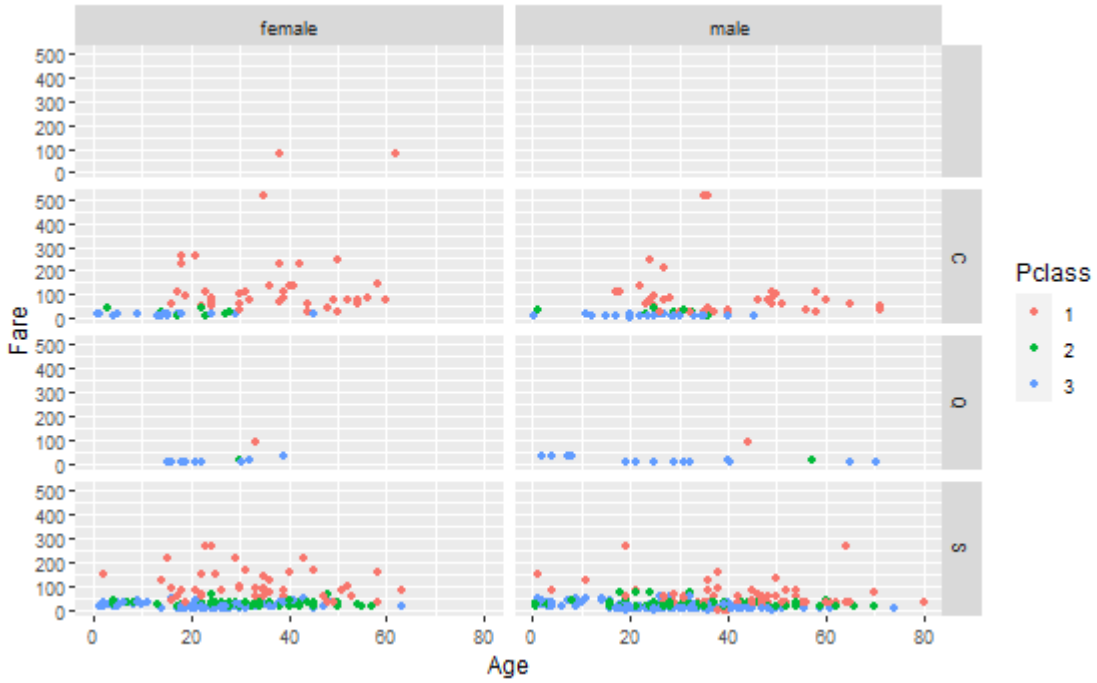
Hide

```
NA
NA
```

Con esta gráfica se confirma lo dicho en la primera. La mayoría de las personas viajan en tercera clase, y después hay una diferencia considerable de precios entre los que viajan en primera clase. Hay dos puntos arriba del todo juntos que no sabemos si son outlier. Habría que estudiarlos.

Hide

```
qplot(data = df, x = Age, y = Fare, color = Pclass, facets = Embarked~Sex)
```



Hide

```
NA
NA
NA
```

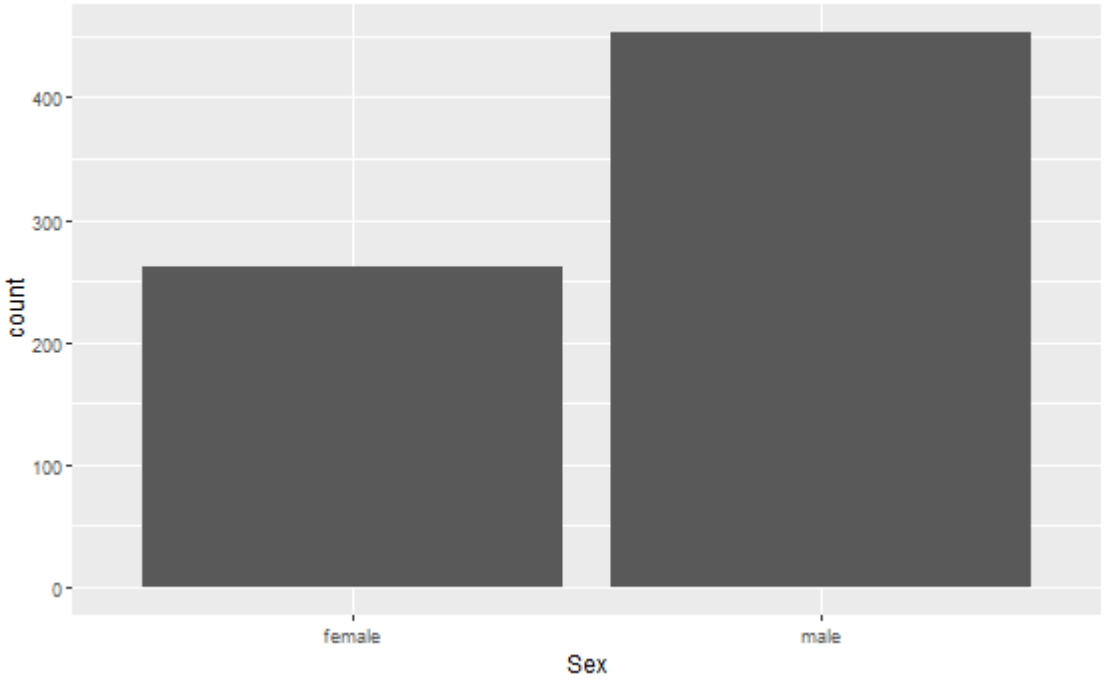
Aquí embarked C = Cherburgo, Q = Queenstown, S = Southampton

Según la ruta del Titanic, salió de Southampton, después paró en Queenstown y por último en Cherburgo. Vemos que a simple vista parece haber más hombres que mujeres en todos los grupos.Lo que vemos es que la mayoría de la gente embarcó en Southampton, personas de las 3 clases. Es normal porque el Titanic partió de allí. Después en la escala de Queenstown se subieron muy pocos, todos de 3 clase menos dos de primera. Y la tercera parada, en Francia ya, también se sube más gente, no tanta como dsde el puerto de partida en Reino Unido. Aunque se ve claramente que hay una gran mayoría de primera clase. Y después hay dos puntos (dentro de mujeres) que no tienen embarque. No sabemos si son outlier.Se deberían estudiar con más detalle.

Para comprobar si había más hombres podemos hacer lo siguiente

Hide

```
ggplot(data=df, aes(x=Sex))+
    geom_bar(stat="count")
```

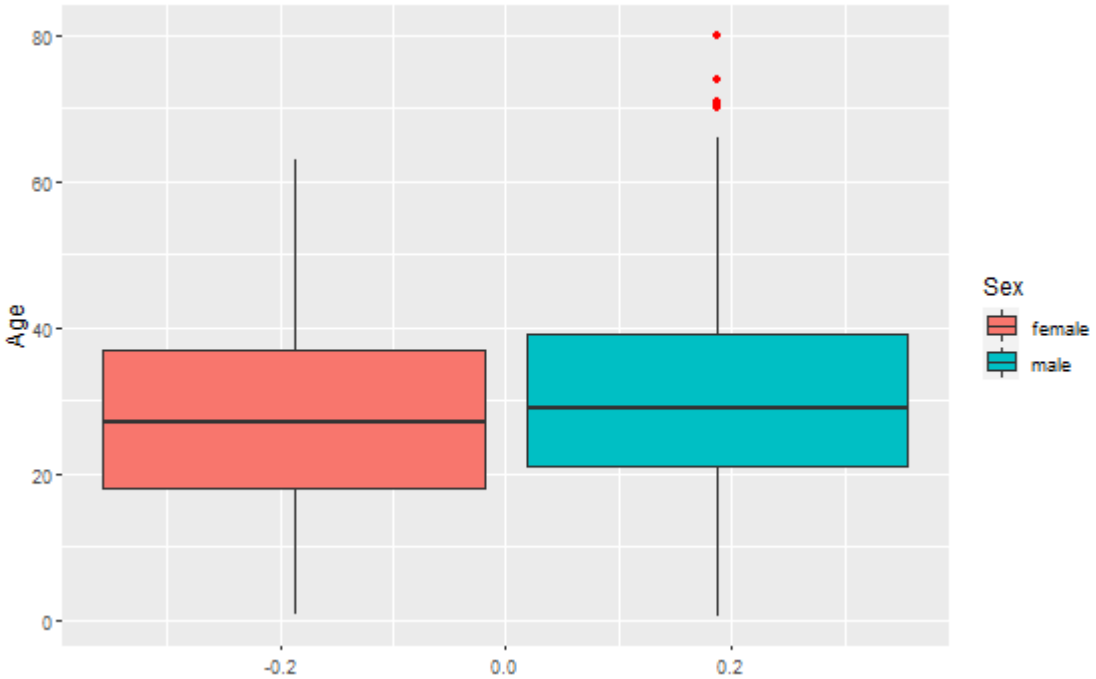


Se ve con la gráfica que si, había más hombres.

Pinta un boxplot de la edad agrupado según *Sex*

Hide

```
ggplot(df,aes(Age))+geom_boxplot(aes(fill=Sex),outlier.colour = "red")+ coord_flip
()
```

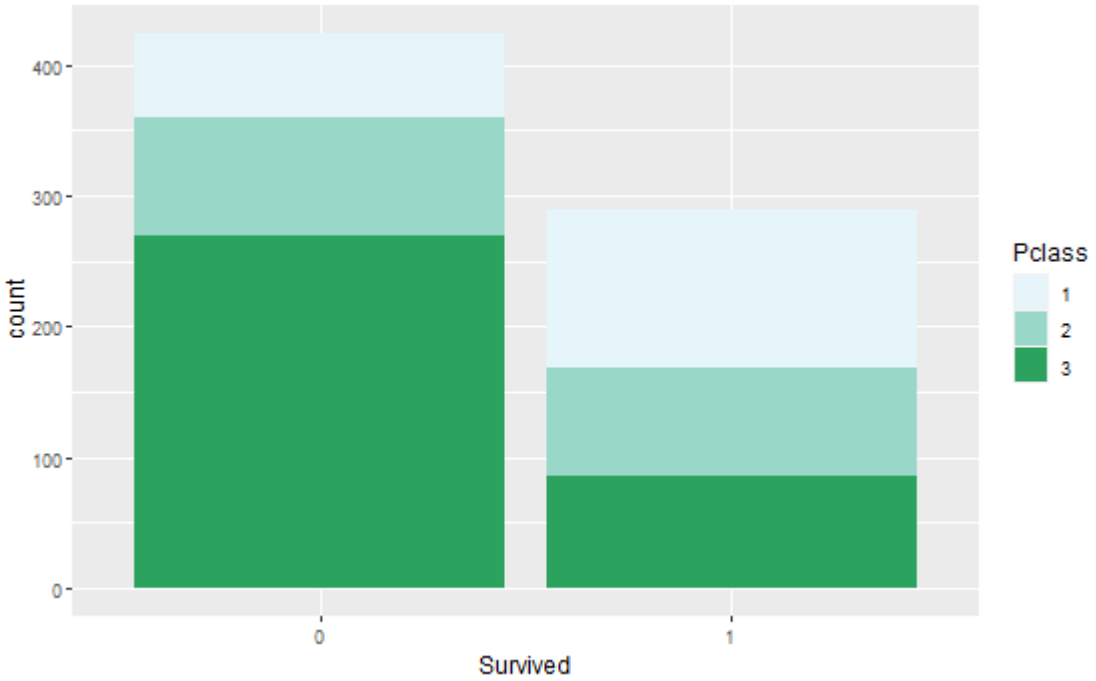


La edad es muy similar, las medias son muy similares entre hombres y mujeres. En este gráfico podemos apreciar que en hombres tenemos outlier, que se deberían estudiar para ver si se quitan o no.

Pinta un barplot que represente la supervivencia en cada *Pclass* coloreando las barras según ésta.

Hide

```
ggplot(data=df, aes(x=Survived,fill=Pclass)) +
    geom_bar(stat="count")+
  scale_fill_manual(values=brewer.pal(n=3,name= "BuGn"))
```



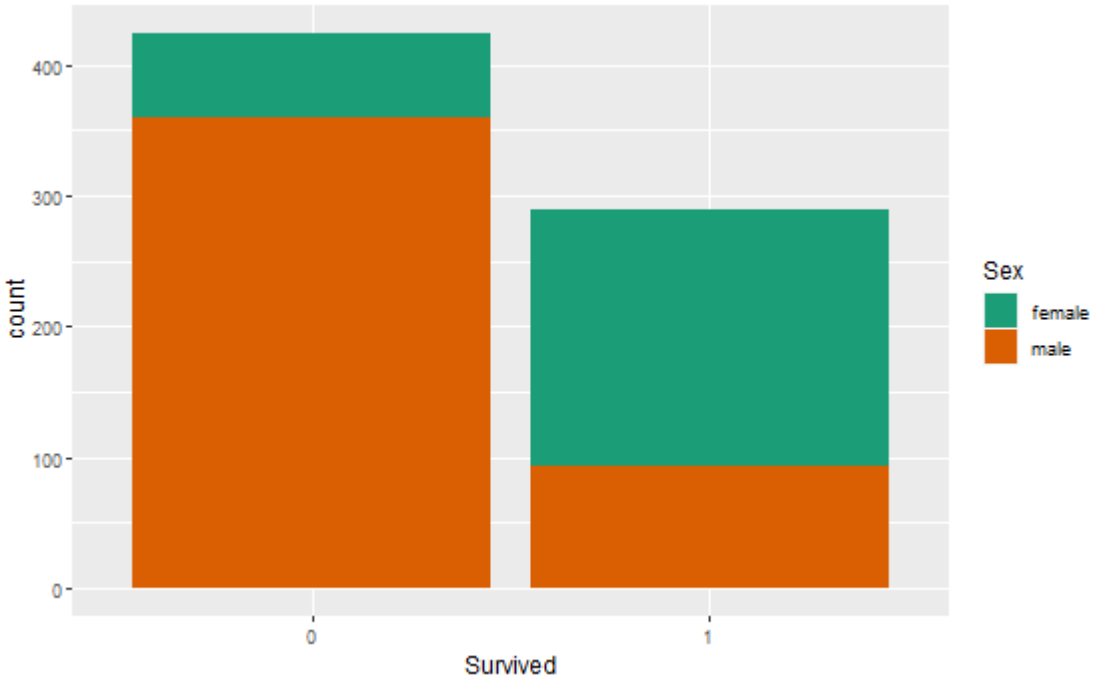
Vemos que hay muchos menos supervivientes que no supervivientes en la tercera clase. En la segunda clase supervivientes y no supervivientes parecen iguales. Y en la primera clase el número de suprvivientes es bastante mayor que el número de no supervivientes.

Pinta la supervivencia en función de la categoría *Sex*, ¿qué observas?

Hide

```
ggplot(data=df, aes(x=Survived,fill=Sex)) +
    geom_bar(stat="count") +
  scale_fill_manual(values=brewer.pal(n=2,name= "Dark2"))
```

```
Warning: minimal value for n is 3, returning requested palette with 3 different le
vels
```

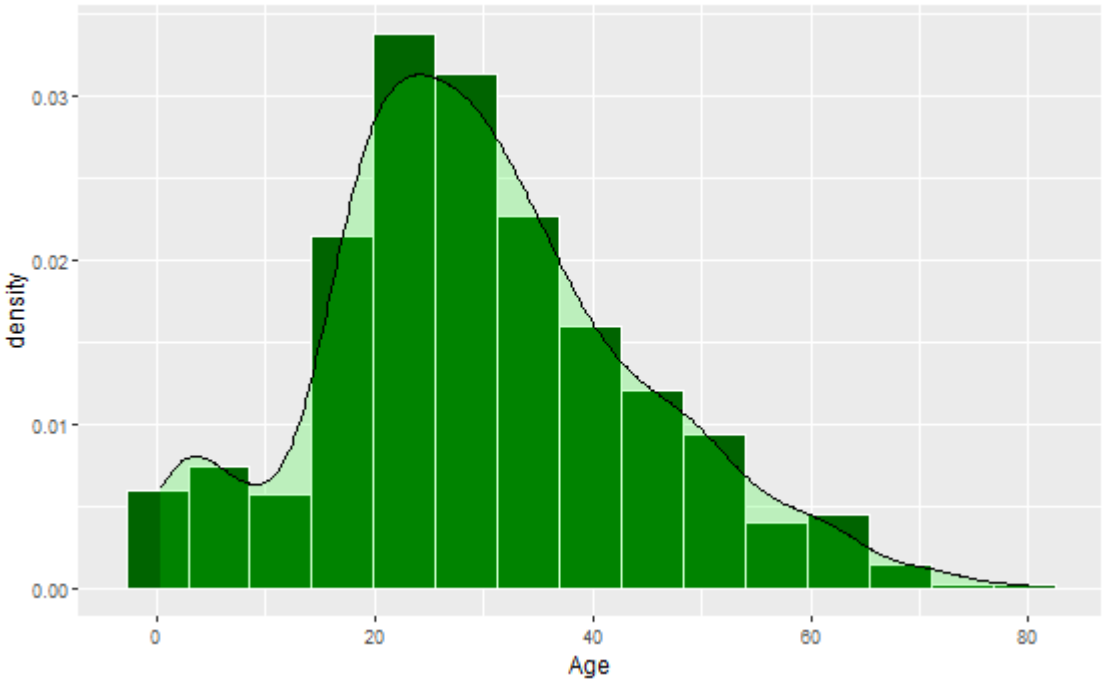


Hombres: más muertos que vivos. Mujeres alrevés, sobrevivieron más mujeres que las que murieron.

Crea un histograma de *Age*, ¿qué observas?

Hide

```
ggplot(df, aes(x = Age)) + geom_histogram(aes(y =..density..),bins = 15,colour="white", fill="darkgreen")+

 geom_density(alpha=.2, fill="green")
```



Hide

```
NA
NA
```

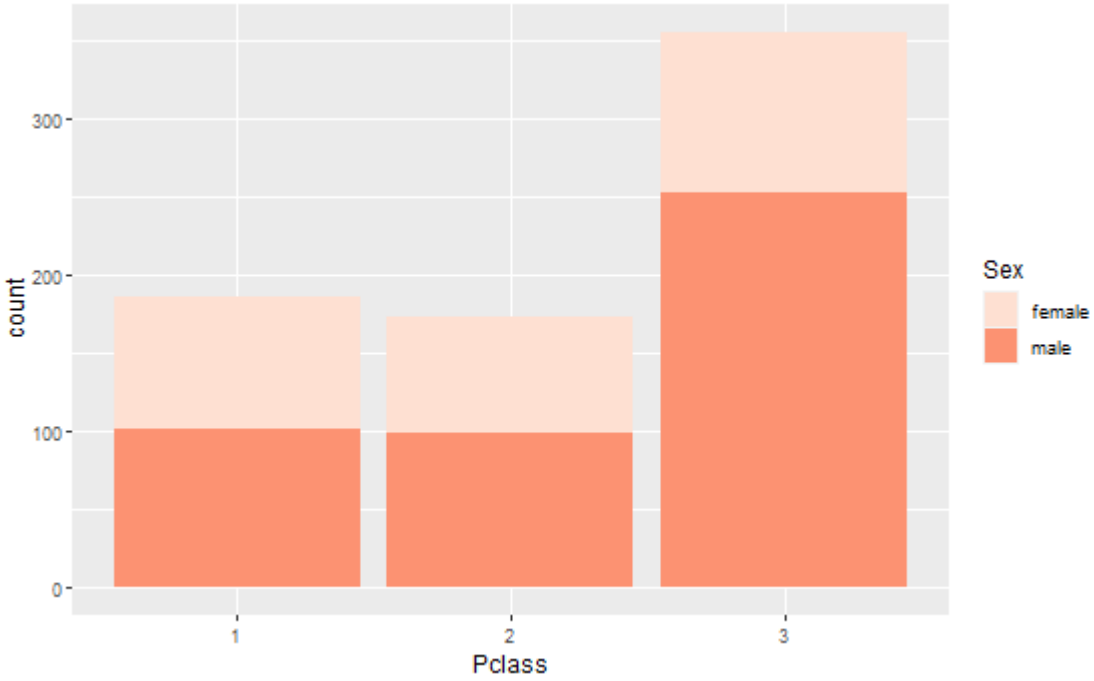
La gran parte de los pasajeros están entre los 18-40 años aproximadamente. Se ve una anomalía al principio de la curva,hay bastantes niños en el barco. Es normal porque los pasajeros están entre 20-40 años, lógico que tengan niños pequeños. Si nos fijamos en las primeras gráficas del ejercicio, la maoría van en clase 3ª y alguno en segunda.No es exactamente una distribución normal.

Crea una agrupación de los datos usando `dplyr` de *Sex* y *Pclass*. Haz sumarios de media, conteo y mediana. ¿Qué podrías decir de los resultados? ¿Hay suficiente muestra para sacar conclusiones en todas las categorías creadas al combinar la edad y la clase?

Hide

```
ggplot(data=df, aes(x=Pclass,fill=Sex))+
    geom_bar(stat="count")+
  scale_fill_manual(values=brewer.pal(n=2,name= "Reds"))
```

```
Warning: minimal value for n is 3, returning requested palette with 3 different le
vels
```



En los hombres hay mucha más variación dentro de las clases. hay muchos más hombres en 3ª. En 2ª y en 1ª más menos igual. En mujeres hay más en 3ª clase pero la diferencia con las otras dos es mucho menos pronunciada que en los hombres

Hide

```
df %>% group_by(Pclass)%>% summarise(mean=mean(Age),n=n())
```

Pclass <fctr>	**mean** <dbl>	**n** <int>
1	38.23344	186
2	29.87763	173
3	25.14062	355

3 rows

Aquí vemos que en conjunto, la media de edad es mayor en la primera clase, después en segunda y por último tercera. Y lo de siempre, viajan más en tercera que en el resto, con mucha diferencia.

Hide

```
df%>% group_by(Pclass)%>% summarise(median=median(Age),n())
```

Pclass <fctr>	median <dbl>	n() <int>
1	37	186
2	29	173
3	24	355

3 rows

Hide

```
NA
```

Con la mediana lo mismo que con la media.

Hide

```
df %>%group_by(Sex)%>%summarise(mean=mean(Age),n=n())
```

Sex <fctr>	mean <dbl>	n <int>
female	27.91571	261
male	30.72664	453

2 rows

Hide

```
NA
```

Entre hombres y mujeres la media es muy similar.

Hide

```
genero_clase <- df %>%
 group_by(Sex, Pclass) %>% summarize(n(), media =mean(Age), mediana = median(Age))
```

```
`summarise()` has grouped output by 'Sex'. You can override using the `.groups` ar
gument.
```

Hide

```
genero_clase
```

Sex <fctr>	Pclass <fctr>	n() <int>	media <dbl>	mediana <dbl>
female	1	85	34.61176	35.0
female	2	74	28.72297	28.0
female	3	102	21.75000	21.5
male	1	101	41.28139	40.0
male	2	99	30.74071	30.0
male	3	253	26.50759	25.0

6 rows

Hide

```
NA
NA
```

Conclusiones: Más hombres que mujeres. La edad de los hombres es mayor, según la media y mediana, en las tres clases, que la de las mujeres. La edad de las personas que viajan en primera clase es mayor que ls que viajan en segunda clase y la edad menos es la delas personas que viajan en tercera clase, tanto para hombres como para mujeres.