CSC3031: Additional Worksheet

TJ McKinley (t.mckinley@exeter.ac.uk)

You should keep an R script file (or preferably an R Markdown file) for all your work. Your script file should be neat, readable and commented throughout (but be succinct). Since we are working with data sets you are encouraged to use the `tidyverse` philosophy for consistency throughout. It would also be good practice to put your answers together with your code as an R Markdown document to produce a final PDF or HTML to act as a future reference.

1. Download the `titanic.csv` data file from the ELE page. Data collected from the Titanic disaster describe the survival status of individual passengers on the Titanic (it does not contain information for the crew). The variables recorded are:

   - `pclass`: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd);
   - `survival`: (0 = No; 1 = Yes);
   - `name`;
   - `gender`: (Male; Female);
   - `age`: in years;
   - `fare`: Passenger Fare (in Pre-1970 British Pounds).

   You are interested in factors affecting the survival of individuals during the disaster. Firstly, read the data into R.

```
## load tidyverse
library(tidyverse)

## read data into R
titanic <- read_csv("titanic.csv")
```

   (a) Use a suitable R command (or commands) to return the number of rows and columns of the dataset.

```
## get dimensions
dim(titanic)
```

```
[1] 1309    6
```

```
## get rows and columns separately
nrow(titanic)
```

```
[1] 1309
```

```
ncol(titanic)
```

```
[1] 6
```

   (b) For each variable in the data set, identify whether the variable is:

   - explanatory or response;

- numeric or categorical.
- If it should be categorical, is it recognised as a 'factor' in R? If not, then turn it into a 'factor'. Ensure that it has the correct levels in the order defined by the data dictionary.
- Remove any variables that do not make sense as either the response, or as an explanatory variable.

```
## examine summary
summary(titanic)
```

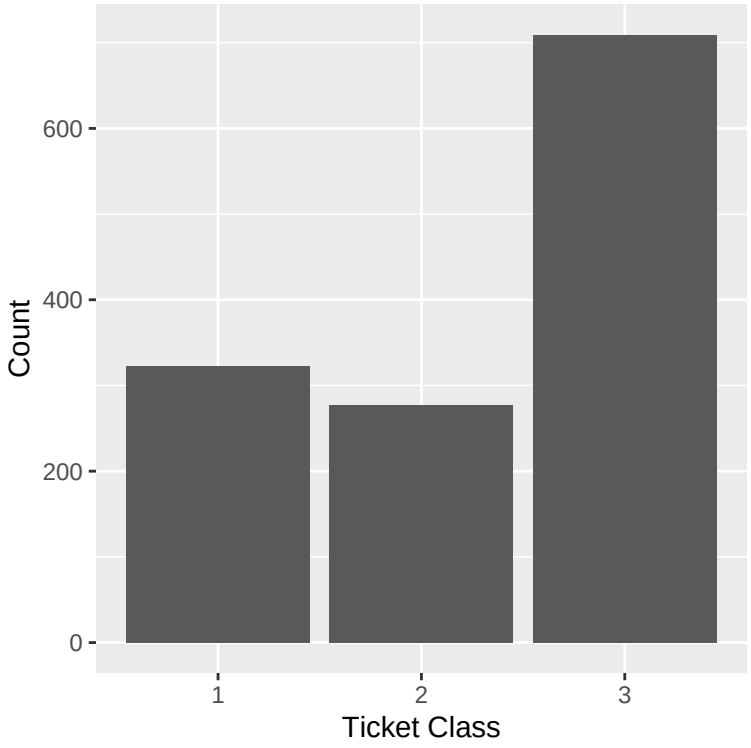
```
     pclass          survived         name              gender
 Min.   :1.000   Min.   :0.000   Length:1309        Length:1309
 1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
 Median :3.000   Median :0.000   Mode  :character   Mode  :character
 Mean   :2.295   Mean   :0.382
 3rd Qu.:3.000   3rd Qu.:1.000
 Max.   :3.000   Max.   :1.000

      age              fare
 Min.   : 0.1667   Min.   :  0.000
 1st Qu.:21.0000   1st Qu.:  7.896
 Median :28.0000   Median : 14.454
 Mean   :29.8811   Mean   : 33.295
 3rd Qu.:39.0000   3rd Qu.: 31.275
 Max.   :80.0000   Max.   :512.329
 NA's   :263       NA's   :1
```

Since the question is interested in *survival*, the obvious response variable is `survived`, the rest are explanatory variables here (except `name`, since this is not necessary for any analysis of these data). The variable `pclass` should be categorical, but has been read in as a discrete variable, as has `survived`; `gender` has been read in as a character, but is categorical and so should be converted into a `factor`; `age` and `fare` are both numeric and have been read in as such. Using `tidyverse` functions, these data can be tidied as below:

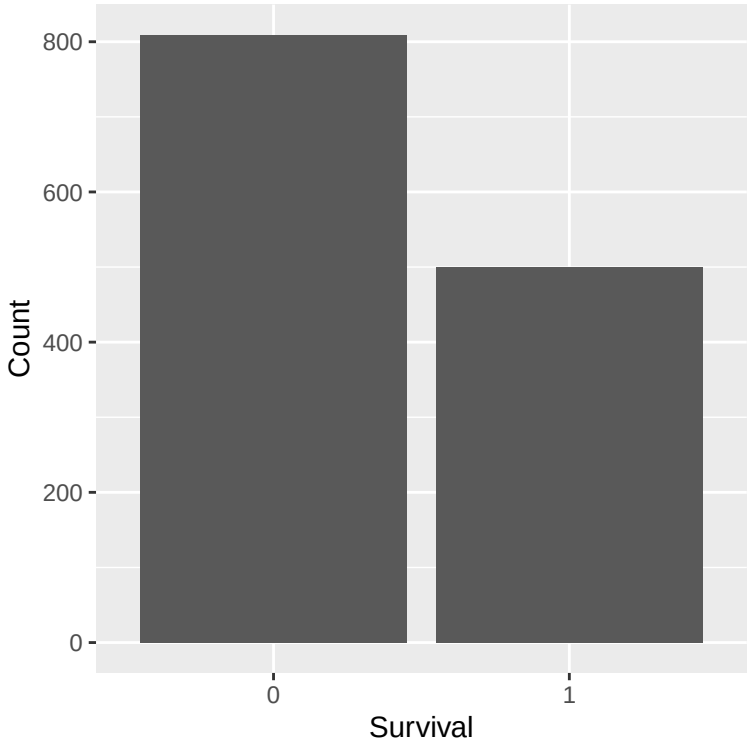
```
## tidy data
titanic <- titanic %>%
    select(-name) %>%
    mutate(
        pclass = factor(pclass),
        survived = factor(survived),
        gender = factor(gender)
    )
```

(c) For each categorical variable, produce a bar chart of counts in each category. For each numeric variable, produce a box-and-whisker plot and a kernel density plot. Discuss which you think is more informative for each variable and why.

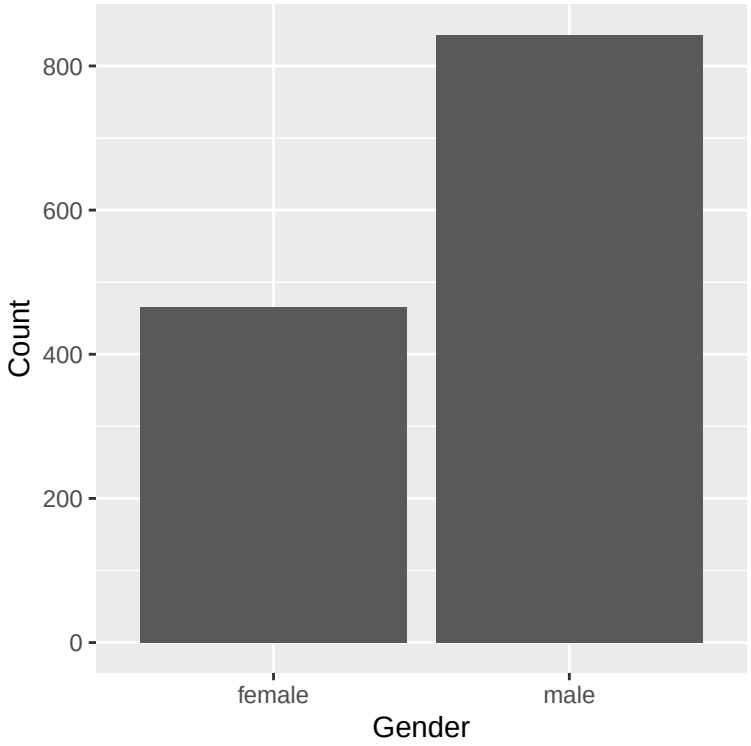
```
## produce a bar chart of the counts of passengers
## in each class
titanic %>%
    ggplot(aes(x = pclass)) +
        geom_bar() +
        xlab("Ticket Class") +
        ylab("Count")
```



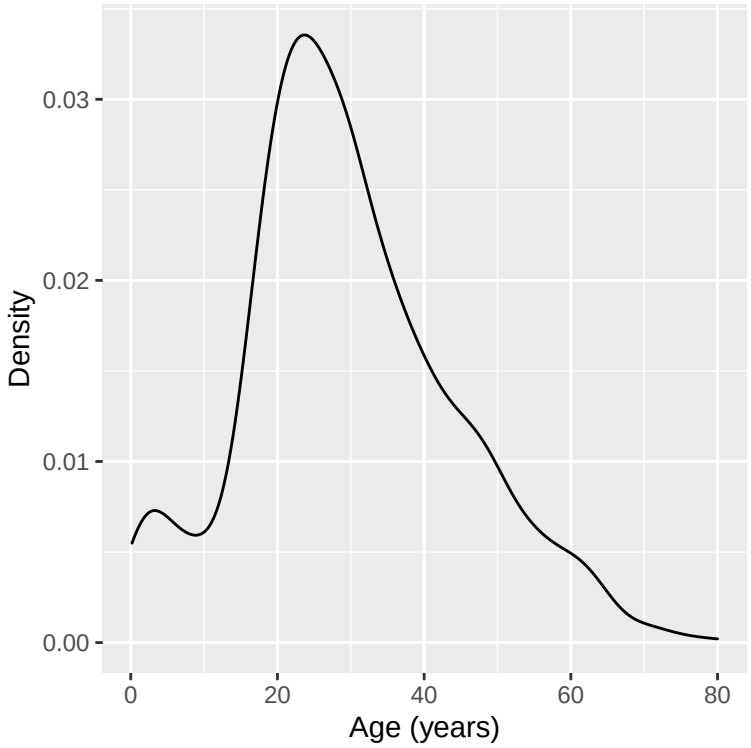

```
## produce a bar chart of the counts of survivors
titanic %>%
    ggplot(aes(x = survived)) +
        geom_bar() +
        xlab("Survival") +
        ylab("Count")
```



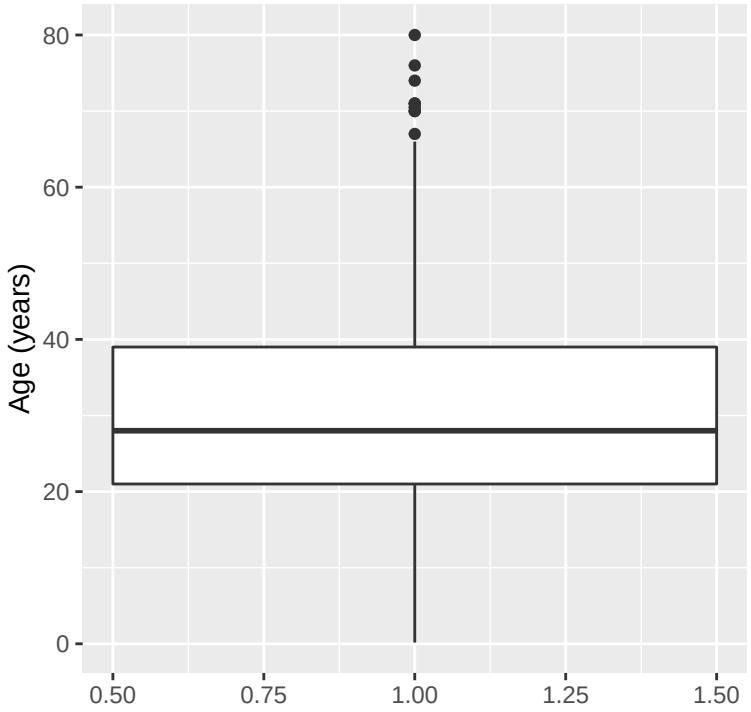

```
## produce a bar chart of the number of men and women
titanic %>%
    ggplot(aes(x = gender)) +
        geom_bar() +
        xlab("Gender") +
        ylab("Count")
```



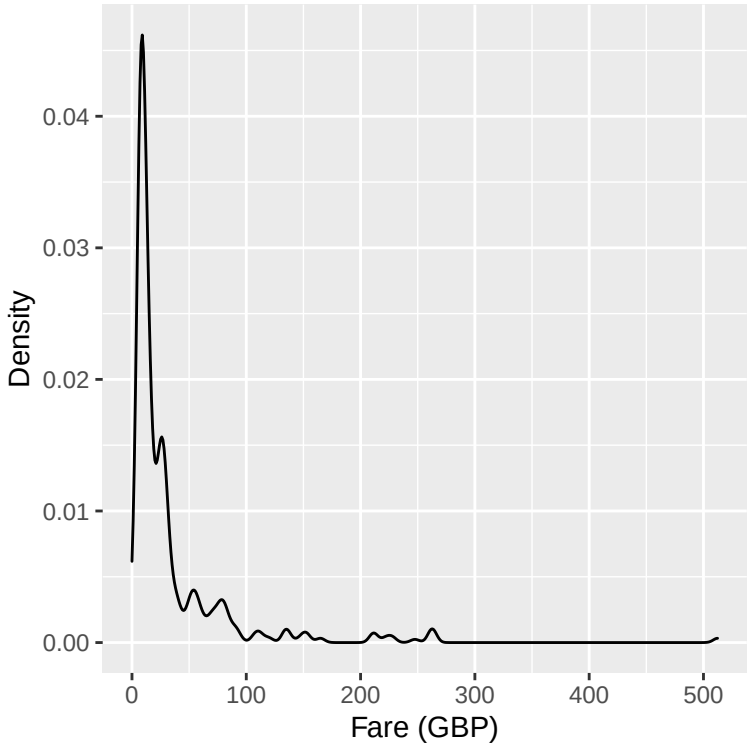

```
## produce a kernel density plot of age
titanic %>%
    ggplot(aes(x = age)) +
        geom_density() +
        xlab("Age (years)") +
        ylab("Density")
```



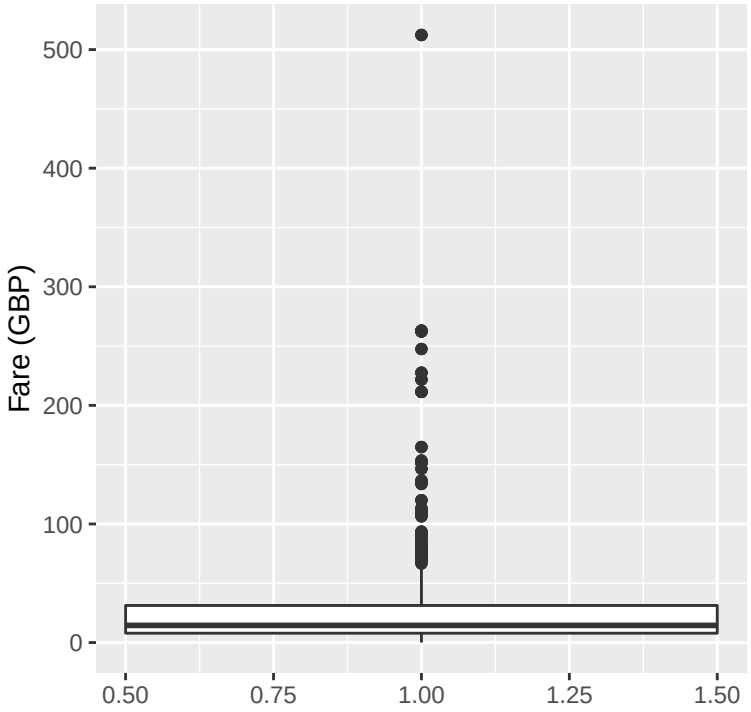

```
## produce a box-and-whisker plot of age
titanic %>%
    ggplot(aes(x = 1, y = age)) +
        geom_boxplot(width = 1) +
        ylab("Age (years)") +
        xlab("")
```




```
## produce a kernel density plot of fare
titanic %>%
    ggplot(aes(x = fare)) +
        geom_density() +
        xlab("Fare (GBP)") +
        ylab("Density")
```

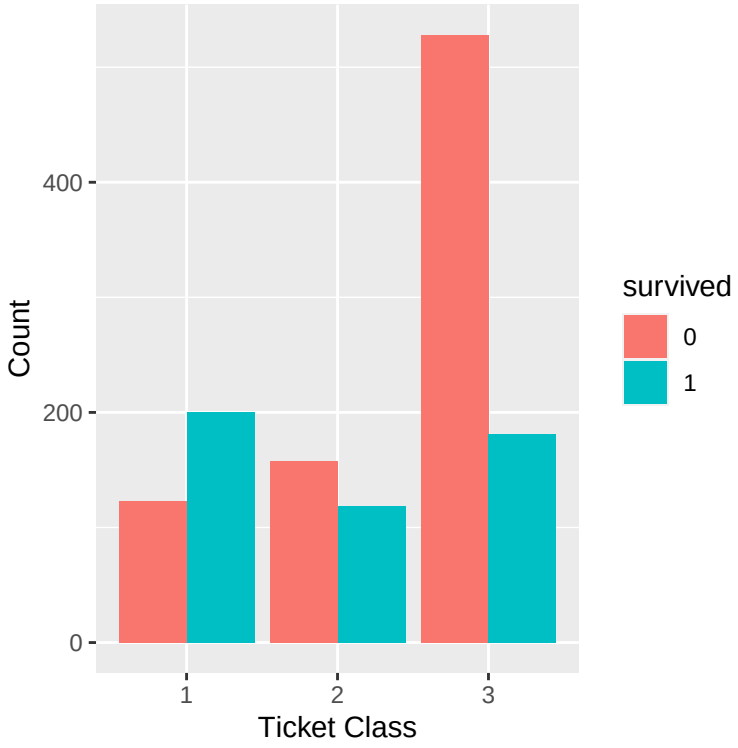



```
## produce a box-and-whisker plot of fare
titanic %>%
    ggplot(aes(x = 1, y = fare)) +
        geom_boxplot(width = 1) +
        ylab("Fare (GBP)") +
        xlab("")
```

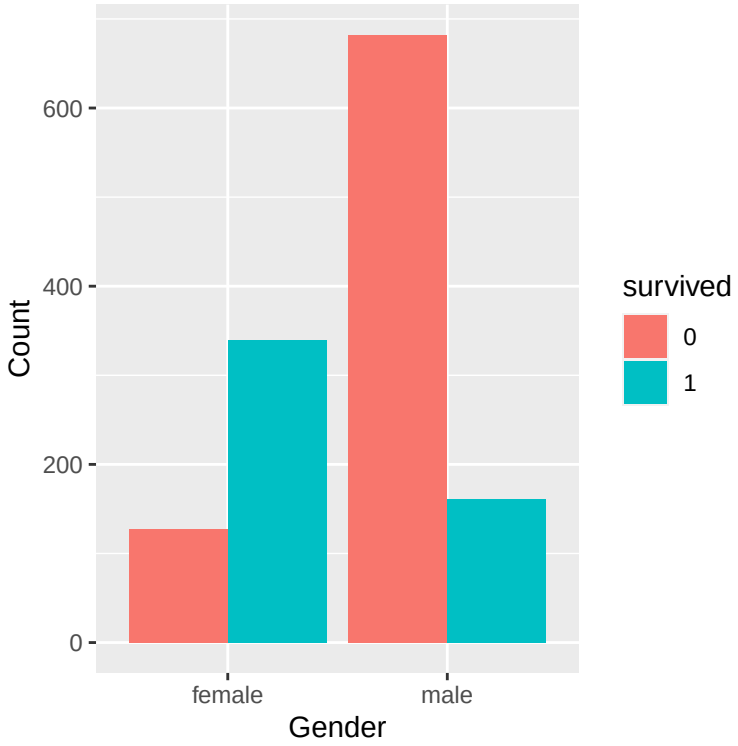


(d) Assuming the response variable is `survived`, produce a series of suitable plots exploring the relationship between the response and each explanatory variable in turn.

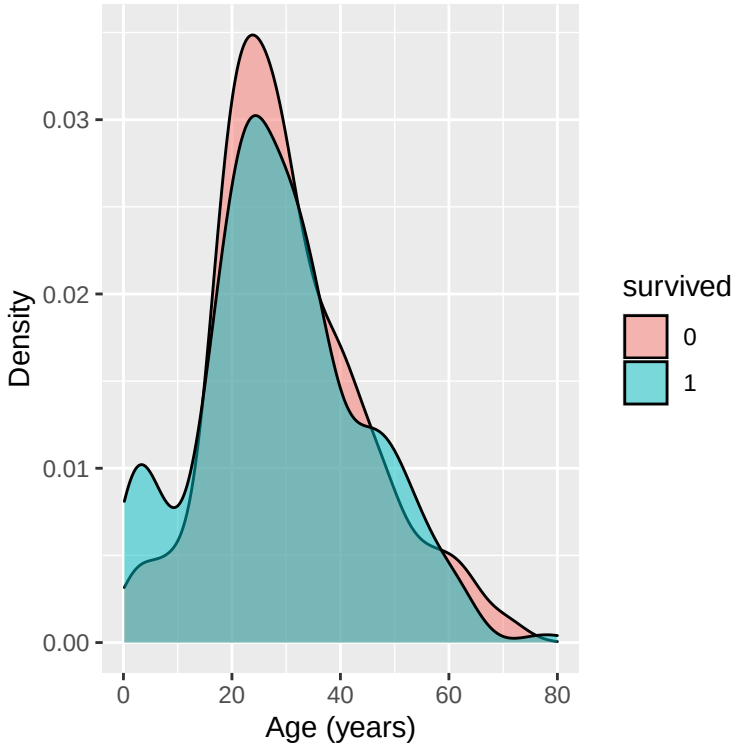
```
## produce a bar chart of the counts of passengers
## in each class
titanic %>%
    ggplot(aes(x = pclass, fill = survived)) +
        geom_bar(position = "dodge") +
        xlab("Ticket Class") +
        ylab("Count")
```



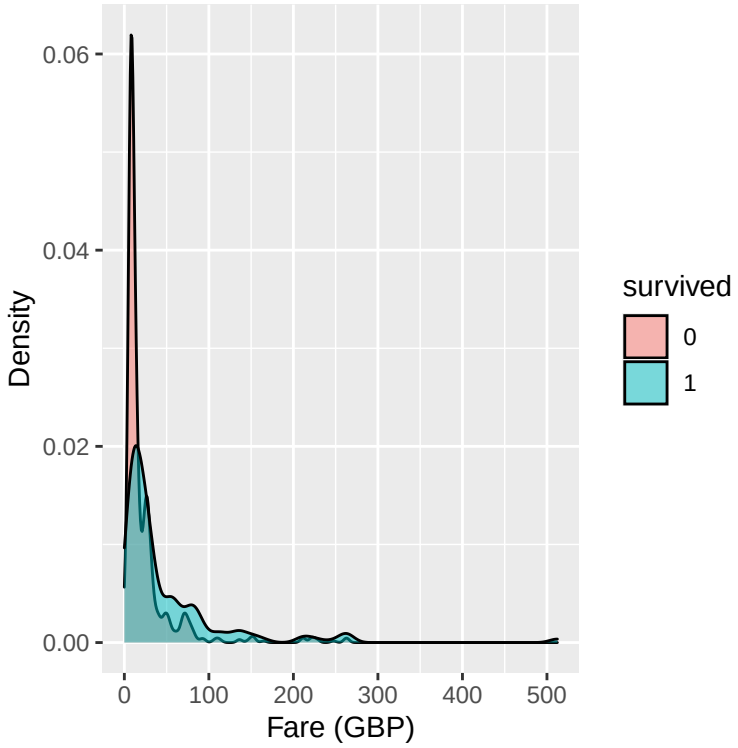

```
## produce a bar chart of the number of men and women
titanic %>%
    ggplot(aes(x = gender, fill = survived)) +
        geom_bar(position = "dodge") +
        xlab("Gender") +
        ylab("Count")
```




```
## produce a kernel density plot of age
titanic %>%
    ggplot(aes(x = age, fill = survived)) +
        geom_density(alpha = 0.5) +
        xlab("Age (years)") +
        ylab("Density")
```




```
## produce a kernel density plot of fare
titanic %>%
    ggplot(aes(x = fare, fill = survived)) +
        geom_density(alpha = 0.5) +
        xlab("Fare (GBP)") +
        ylab("Density")
```

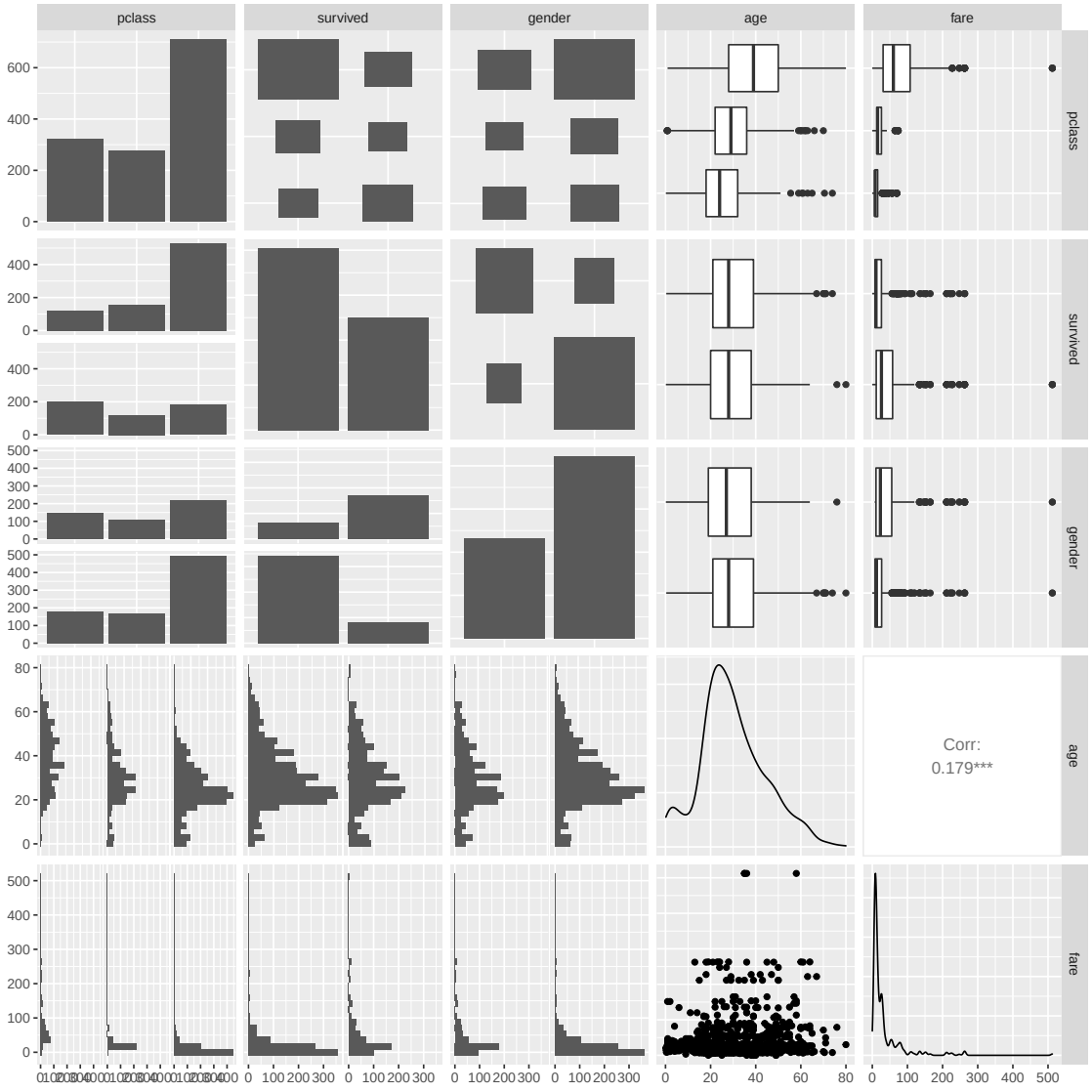


(e) Write a short passage summarising your thoughts about the observed patterns in the data, focussing on the potential relationships between the response and each explanatory variable.

It looks like the survival rates are higher for first and second class passengers, and women seem to have a higher survival rate that men. The kernel density plots of `age` suggest that children have higher survival rates, which can be seen by the lower mode in the survived group that is not present in the the density plot for those passengers that died. There also seems to be some relationship with fare, such that pasengers that paid a higher fare were more likely to survive, though this is effect might also be confounded with ticket class.

(f) Install the package `GGally`. This provides some additional functions for producing more complex plots, but built on `ggplot2`. Explore the website http://ggobi.github.io/ggally/, in particular the `ggpairs` function, and try to produce an informative multivariate summary plot of these data.

```
## load GGally library
library(GGally)

## produces a generalised pairs plot
titanic %>% ggpairs()
```


pclass
survived
gender
age
fare
Corr:
0.179***