ggplot2 Titanic Exercise

Chinpei Tang

May 21, 2016

You've watched the movie Titanic by James Cameron (1997) again and after a good portion of sobbing you decide to investigate whether you'd have a chance of surviving this disaster.

To start your investigation, you decide to do some exploratory visualization with ggplot(). First, load the ggplot2 library.

```
library("ggplot2")
```

The data from Datacamp is copied and pasted, and imported to a csv file.

```
setwd("C:/Users/Chinpei/Documents/GitHub/Springboard_FDS/ggplot2_titanic")
titanic = read.csv("titanic.csv", header=T)
```

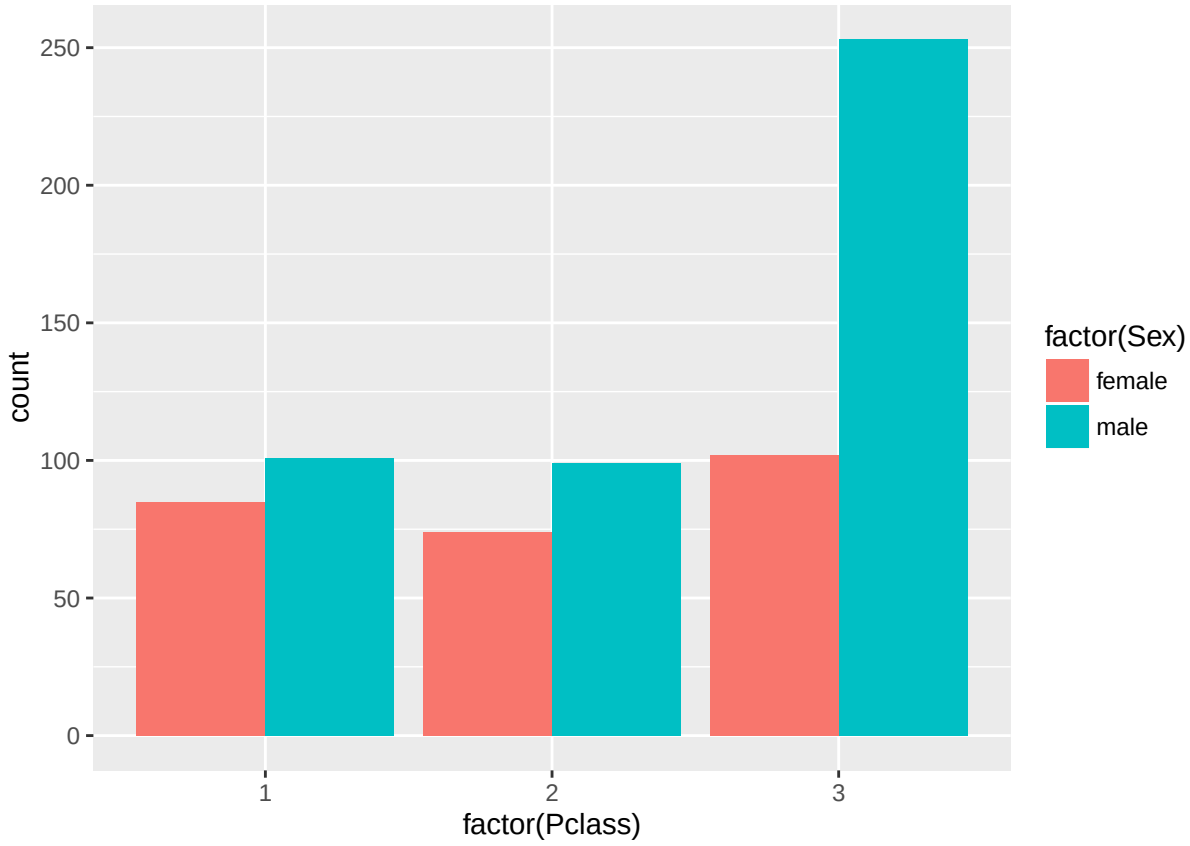
Have a look at the str() of the titanic dataset, which has been loaded into your workspace. Looks like the data is pretty tidy!

```
str(titanic)
```

```
## 'data.frame':    714 obs. of  5 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived: int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 1 3 3 2 3 ...
##  $ 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 ...
```

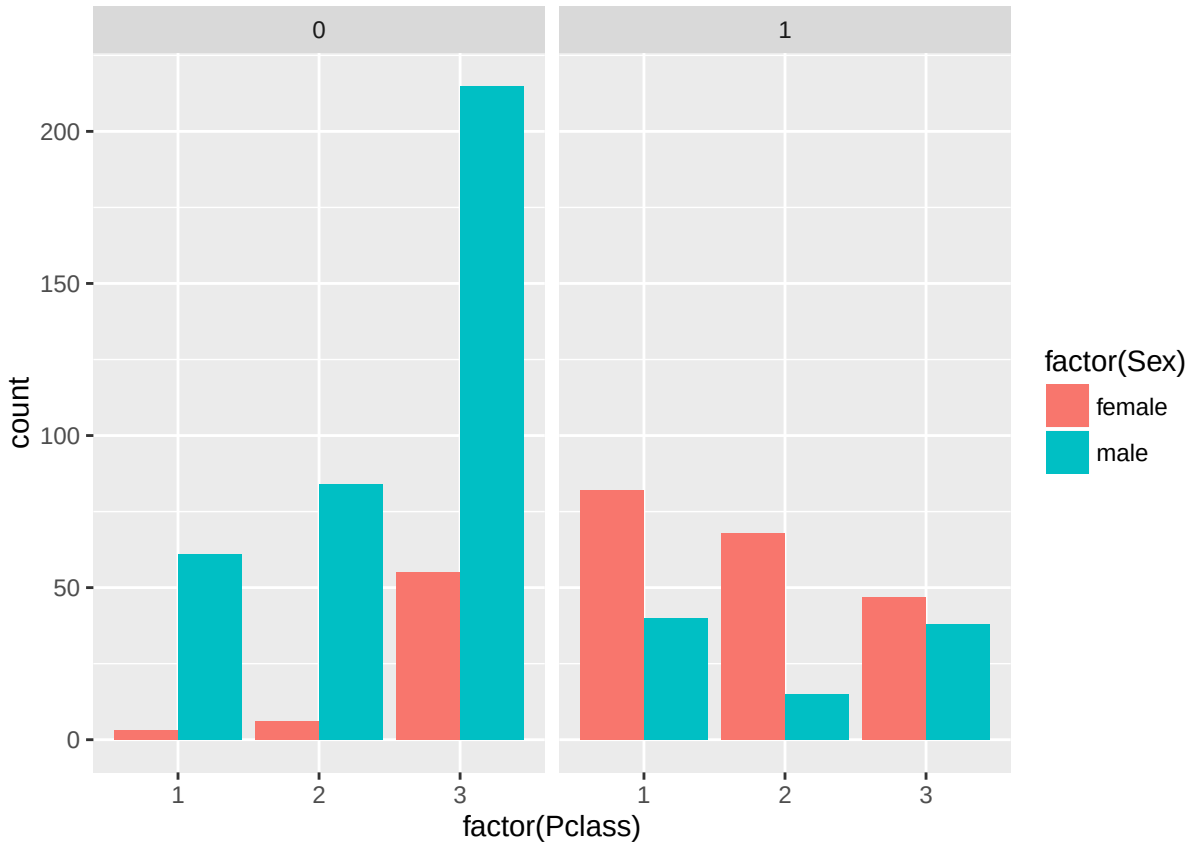
First, have a look at the distribution of sexes within the classes of the ship. Use ggplot() with the data layer set to titanic. Map factor(Pclass) onto the x axis, factor(Sex) onto fill and draw a dodgy bar plot using geom_bar(). By dodgy we mean that position has to be set to "dodge".

```
ggplot(titanic, aes(x = factor(Pclass), fill = factor(Sex))) +
  geom_bar(position = "dodge")
```



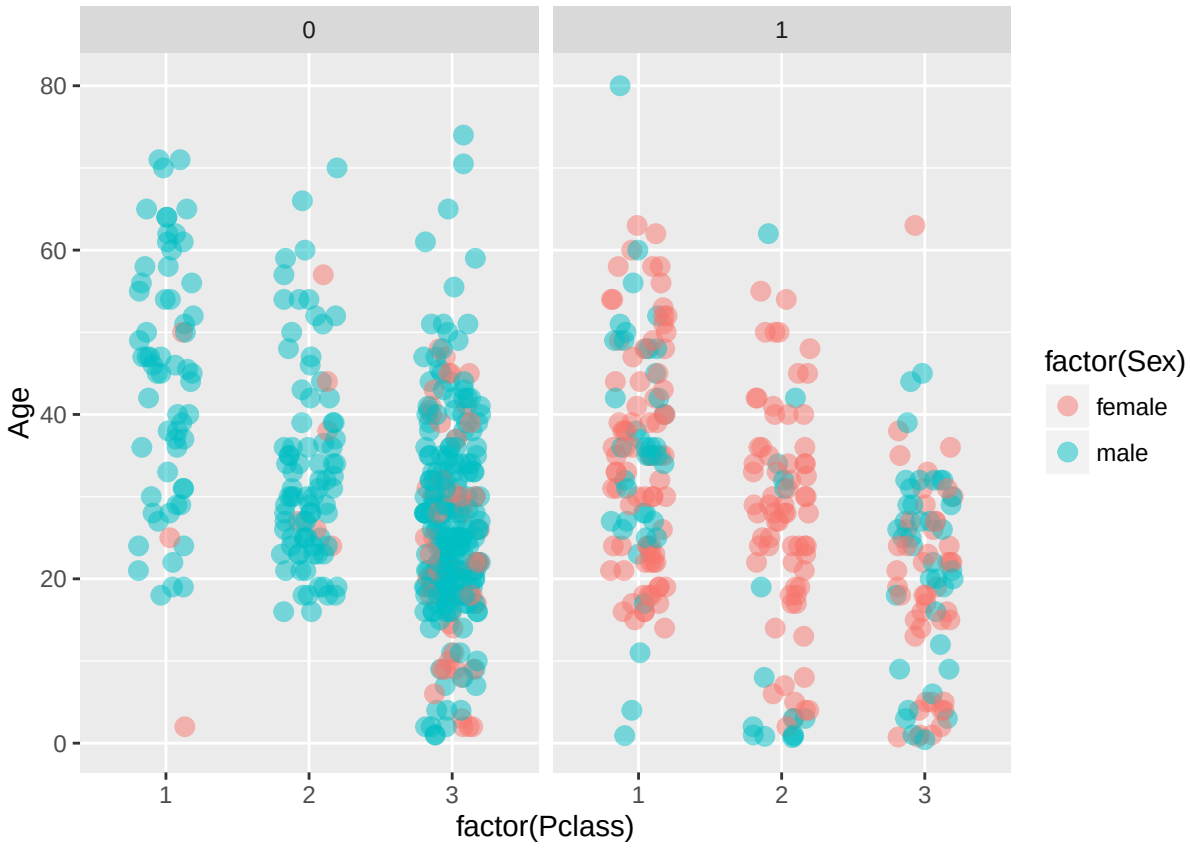
These bar plots won't help you estimate your chances of survival. Copy the previous bar plot, but this time add a facet_grid() layer: ". ~ Survived".

```
ggplot(titanic, aes(x = factor(Pclass), fill = factor(Sex))) +
  geom_bar(position = "dodge") +
  facet_grid(".~Survived")
```



You want to include the final variable in the plot, Age. Start off by copying the previous ggplot() command. Add a mapping of Age onto the y axis. Replace geom_bar() with geom_jitter() and set its attributes size to 3, alpha to 0.5 and position set to posn.j. Make sure that factor(Sex) is mapped onto col instead of fill this time to correctly color the scatter plots.

```
posn.j <- position_jitter(0.5, 0)
ggplot(titanic, aes(x = factor(Pclass), y = Age, col = factor(Sex))) +
  geom_jitter(size = 3, alpha = 0.5, position = posn.j) +
  facet_grid(".~Survived")
```


0
1
80
60
40
20
0
Age
1
2
3
1
2
3
factor(Pclass)
factor(Sex)
female
male