Titanic Survival Analysis

Prasham Bhuta

June 2, 2020

Titanic exercise for R

```
library(titanic)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
options(digits = 3)

# defining the dataset
titanic <- titanic_train %>%
    select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%
    mutate(Survived = factor(Survived), Pclass = factor(Pclass),
           Sex = factor(Sex))

summary(titanic)
```

```
##  Survived Pclass      Sex           Age           SibSp          Parch
##  0:549    1:216   female:314   Min.   : 0.4   Min.   :0.00   Min.   :0.00
##  1:342    2:184   male  :577   1st Qu.:20.1   1st Qu.:0.00   1st Qu.:0.00
##           3:491                Median :28.0   Median :0.00   Median :0.00
##                                Mean   :29.7   Mean   :0.52   Mean   :0.38
##                                3rd Qu.:38.0   3rd Qu.:1.00   3rd Qu.:0.00
##                                Max.   :80.0   Max.   :8.00   Max.   :6.00
##                                NA's   :177
##       Fare
##  Min.   :  0
##  1st Qu.:  8
##  Median : 14
```

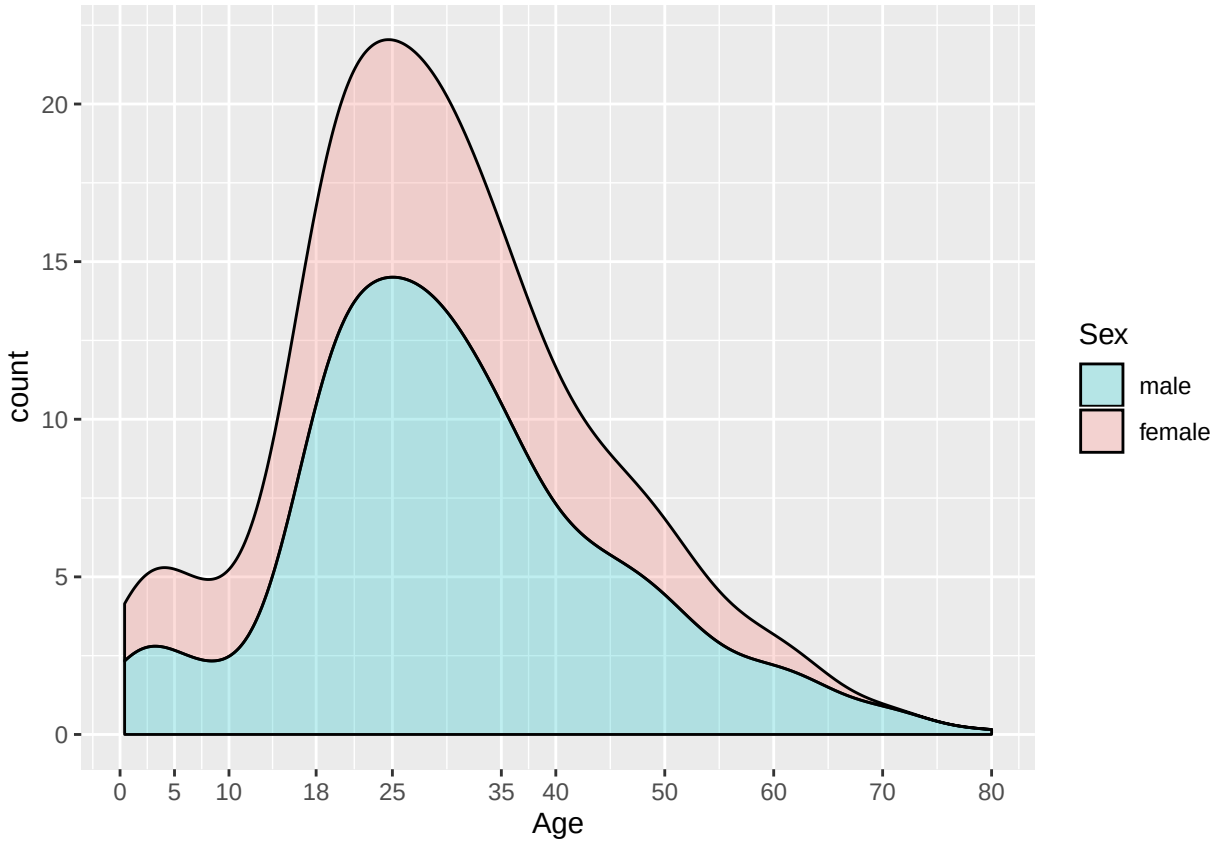
```
##  Mean   : 32
##  3rd Qu.: 31
##  Max.   :512
##
```

```
str(titanic)
```

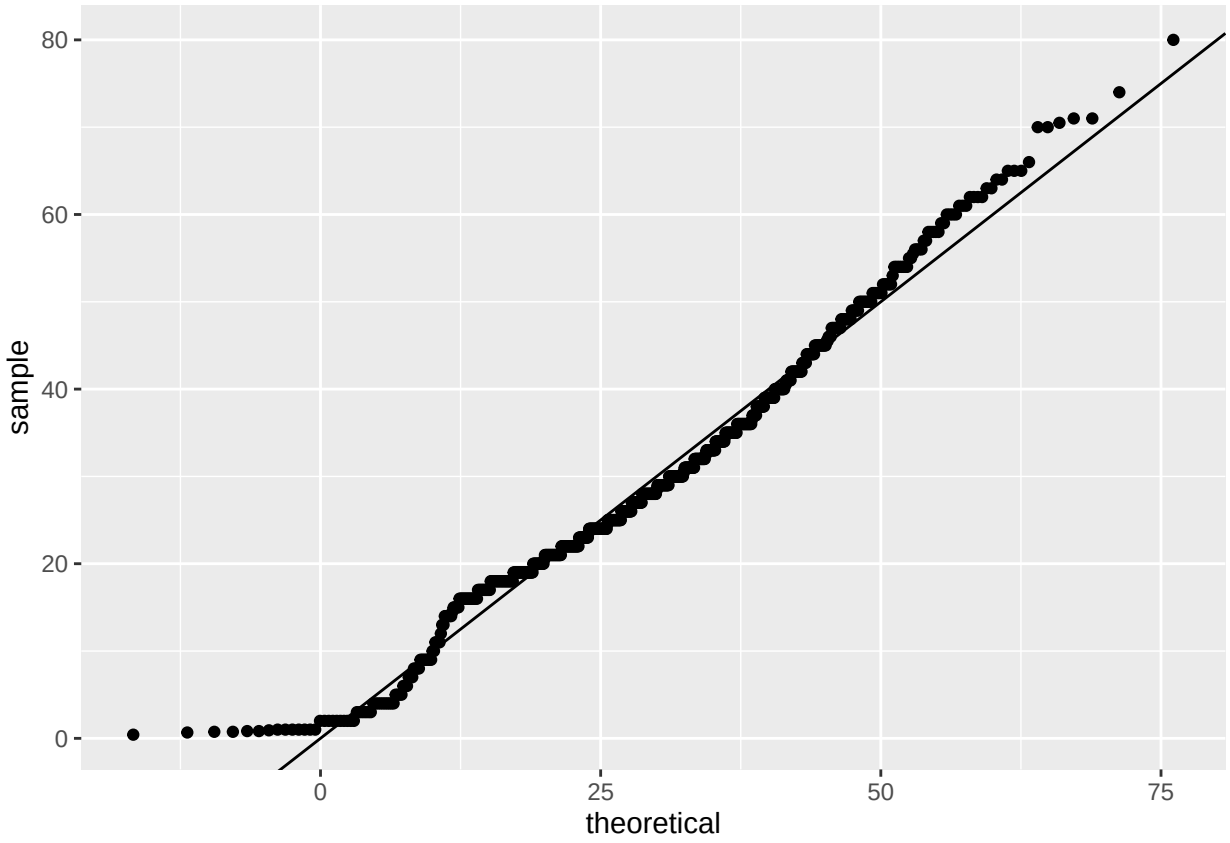
```
## 'data.frame':    891 obs. of  7 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 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 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
# density plot
titanic %>%
    ggplot(aes(Age, y= ..count.., fill = Sex)) + geom_density(alpha = 0.25,
                                                  position = "stack") +
    guides(fill = guide_legend(reverse = TRUE)) +
    scale_x_continuous(breaks = c(0,5,10,18,25,35,40,50,60,70,80))
```

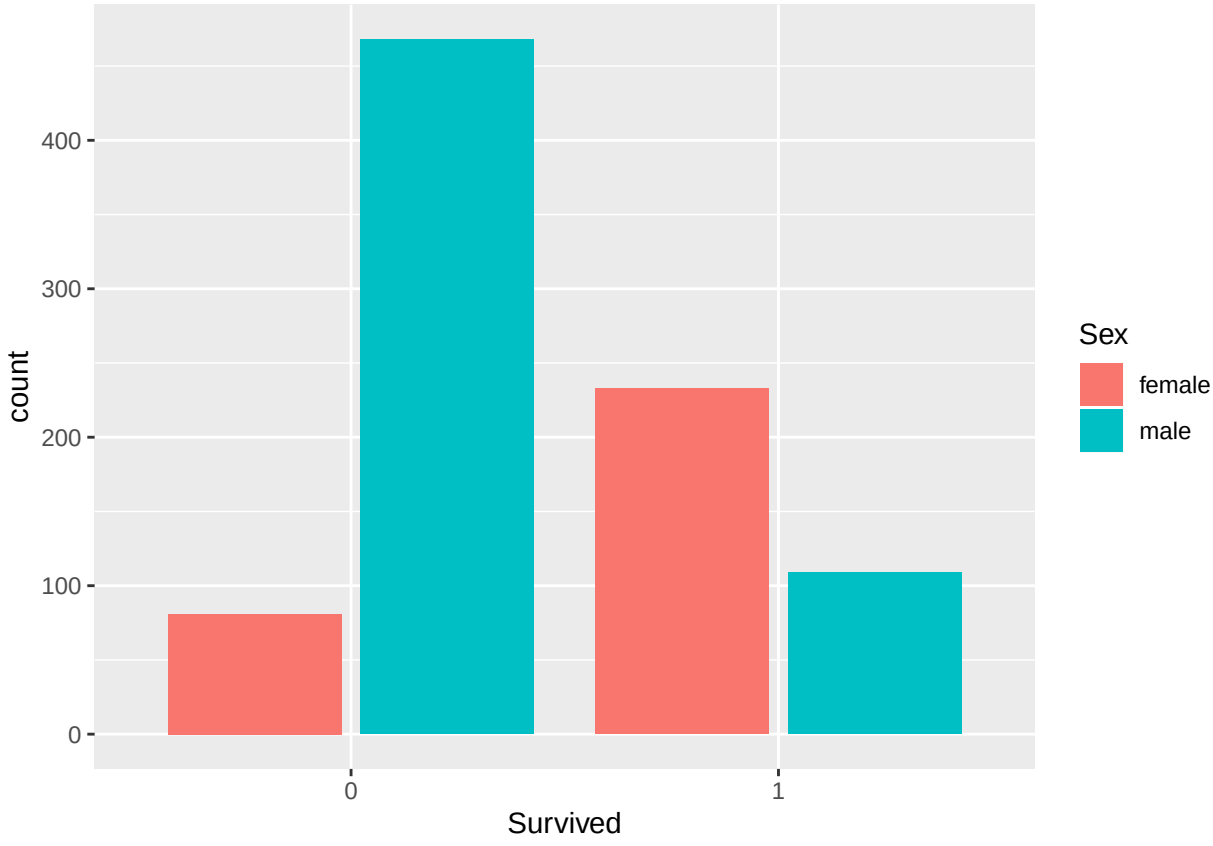
```
## Warning: Removed 177 rows containing non-finite values (stat_density).
```



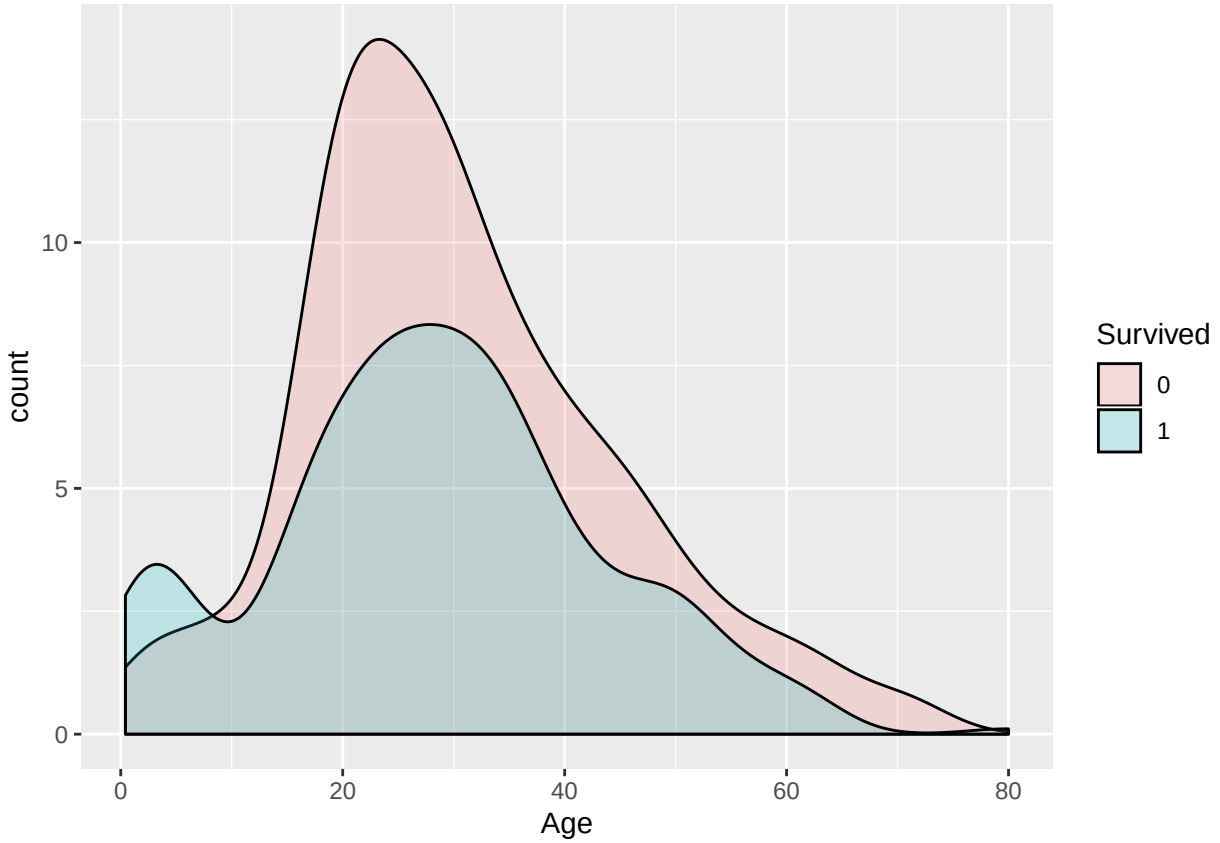
```r
# qqplot
params <- titanic %>% filter(!is.na(Age)) %>% summarize(mean = mean(Age),
                                                        sd = sd(Age))


titanic %>% filter(!is.na(Age)) %>%
    ggplot(aes(sample = Age)) + geom_qq(dparams = params) +geom_abline()
```




```r
# Barplot
titanic %>% ggplot(aes(Survived, fill = Sex)) + geom_bar(position = 'dodge2')
```




```
# Density plot for survival by Age
titanic %>% filter(!is.na(Age)) %>% ggplot(aes(Age, y=..count..,
                                               fill=Survived)) +
    geom_density(alpha = 0.2)
```

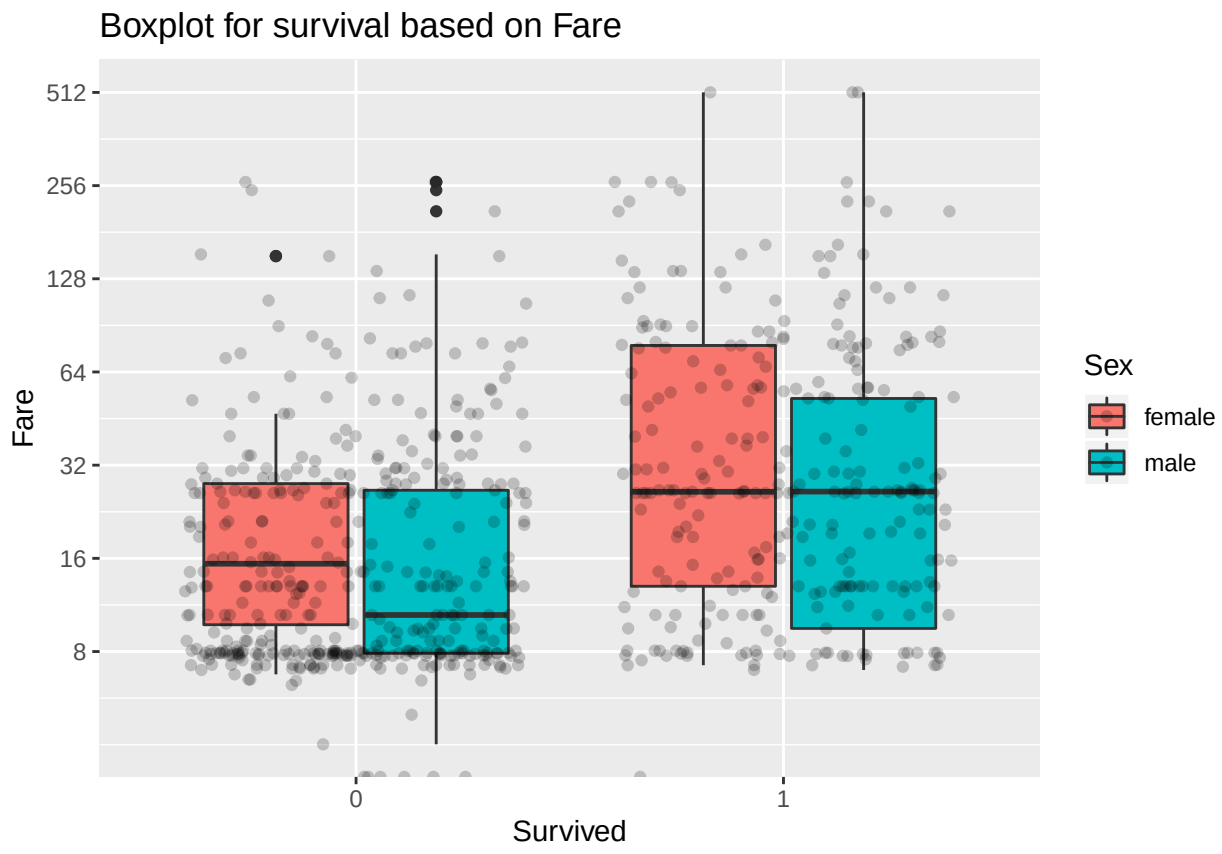



```
# Box plot for survival by Fare
titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    ggplot(aes(Survived, Fare, fill=Sex)) + geom_boxplot() +
    scale_y_continuous(trans = "log2", breaks = c(8,16,32,64,128,256,512)) +
    geom_jitter(alpha = 0.2) +
    ggtitle("Boxplot for survival based on Fare")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

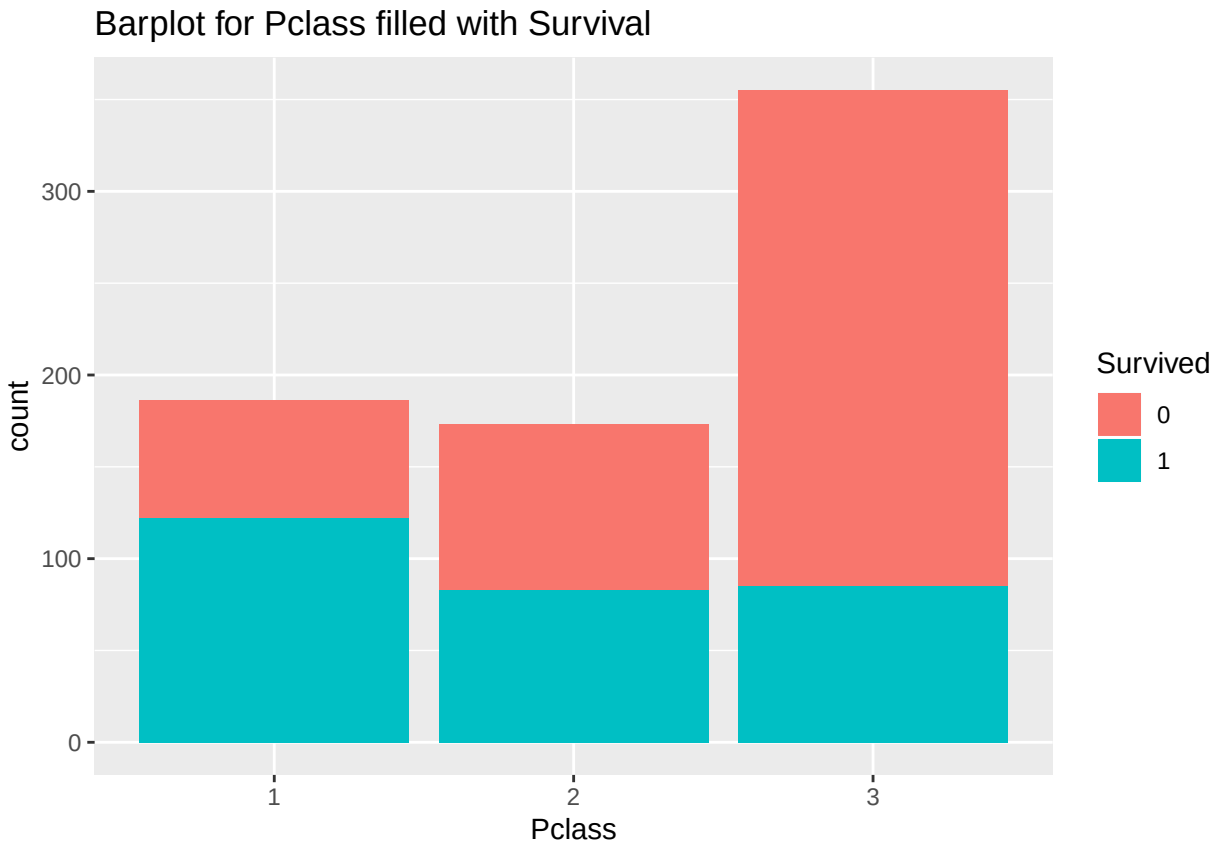
```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```



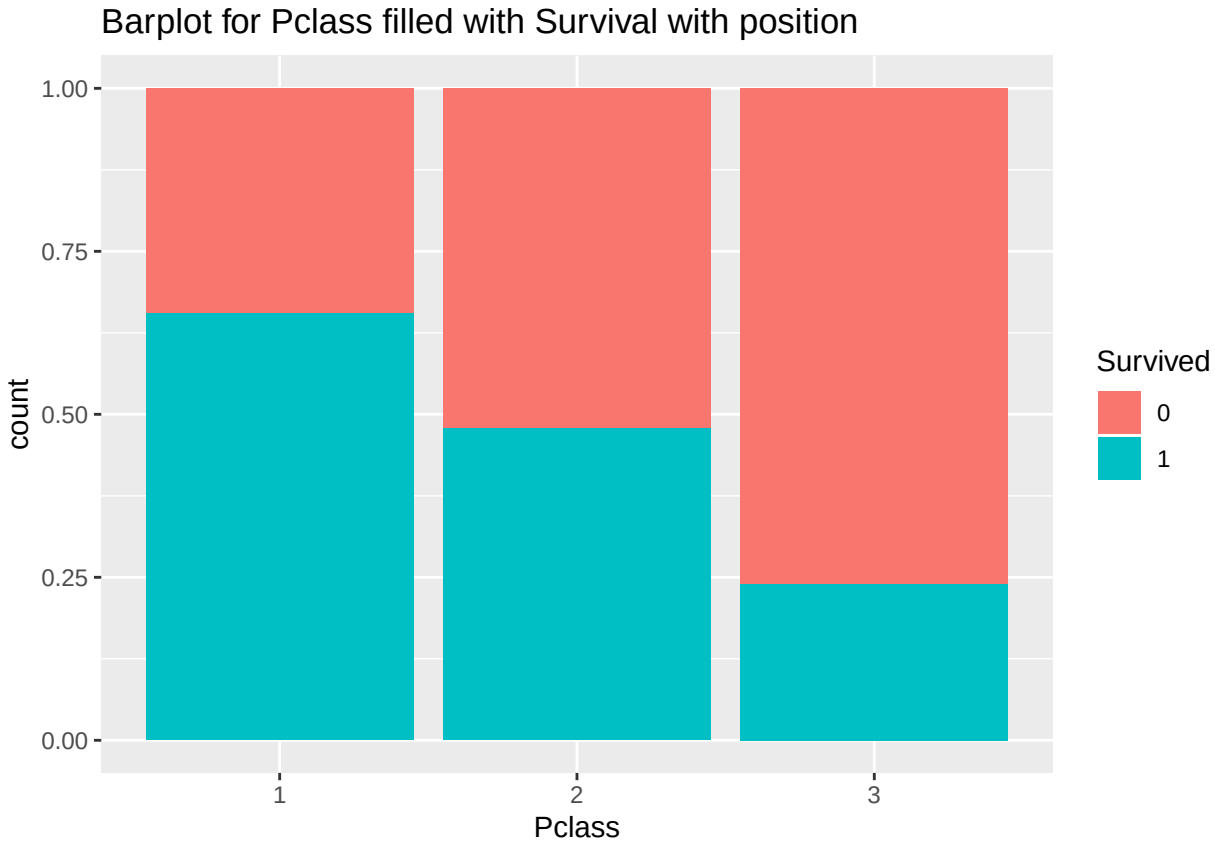

```
median_fare <- titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    group_by(Survived) %>% summarise(median = median(Fare))

print(median_fare)
```

```
## # A tibble: 2 x 2
##   Survived median
##   <fct>     <dbl>
## 1 0          11.9
## 2 1          26.2
```

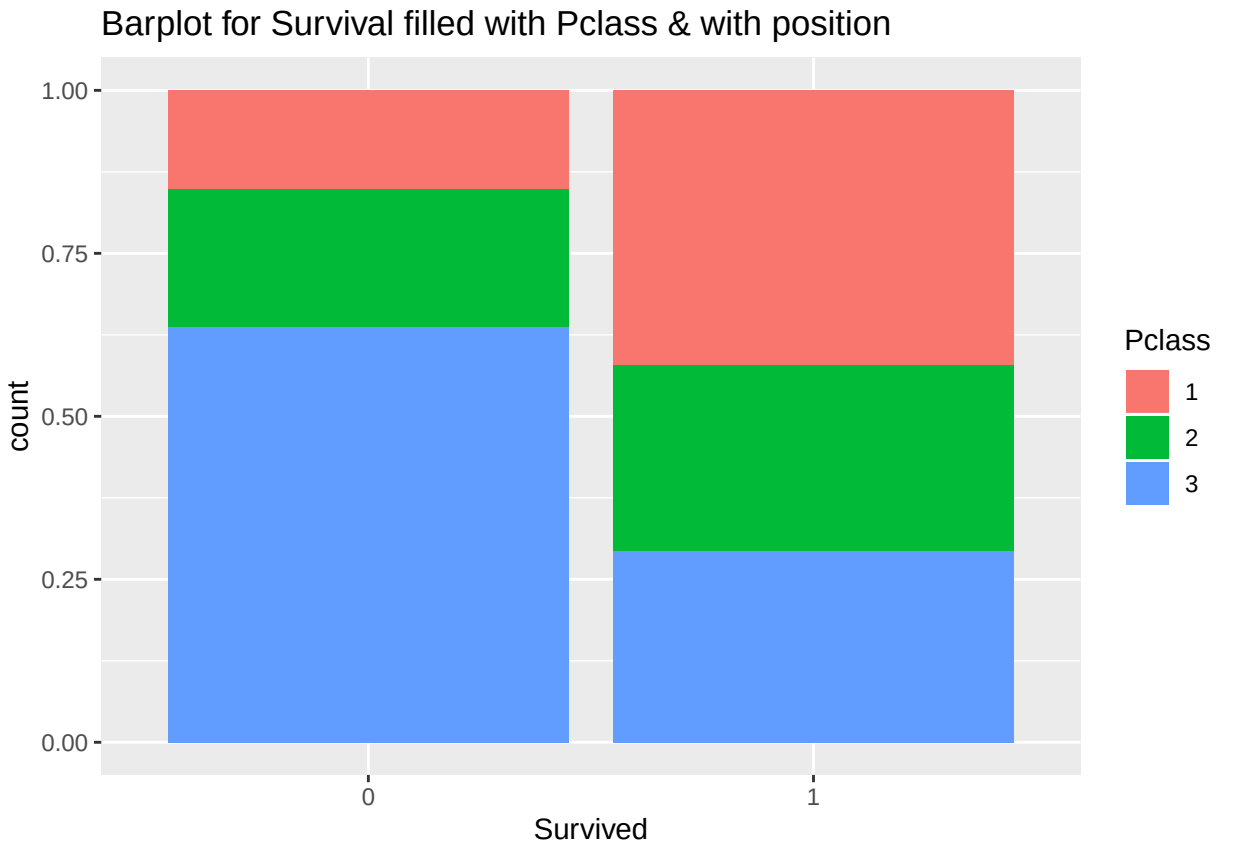
```
# Barplot for survival by passenger Class
titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    ggplot(aes(Pclass, fill=Survived)) +
    geom_bar() +
    ggtitle("Barplot for Pclass filled with Survival")
```




```
# Barplot2 with 'position' argument
titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    ggplot(aes(Pclass, fill=Survived)) +
    geom_bar(position = position_fill()) +
    ggtitle("Barplot for Pclass filled with Survival with position")
```



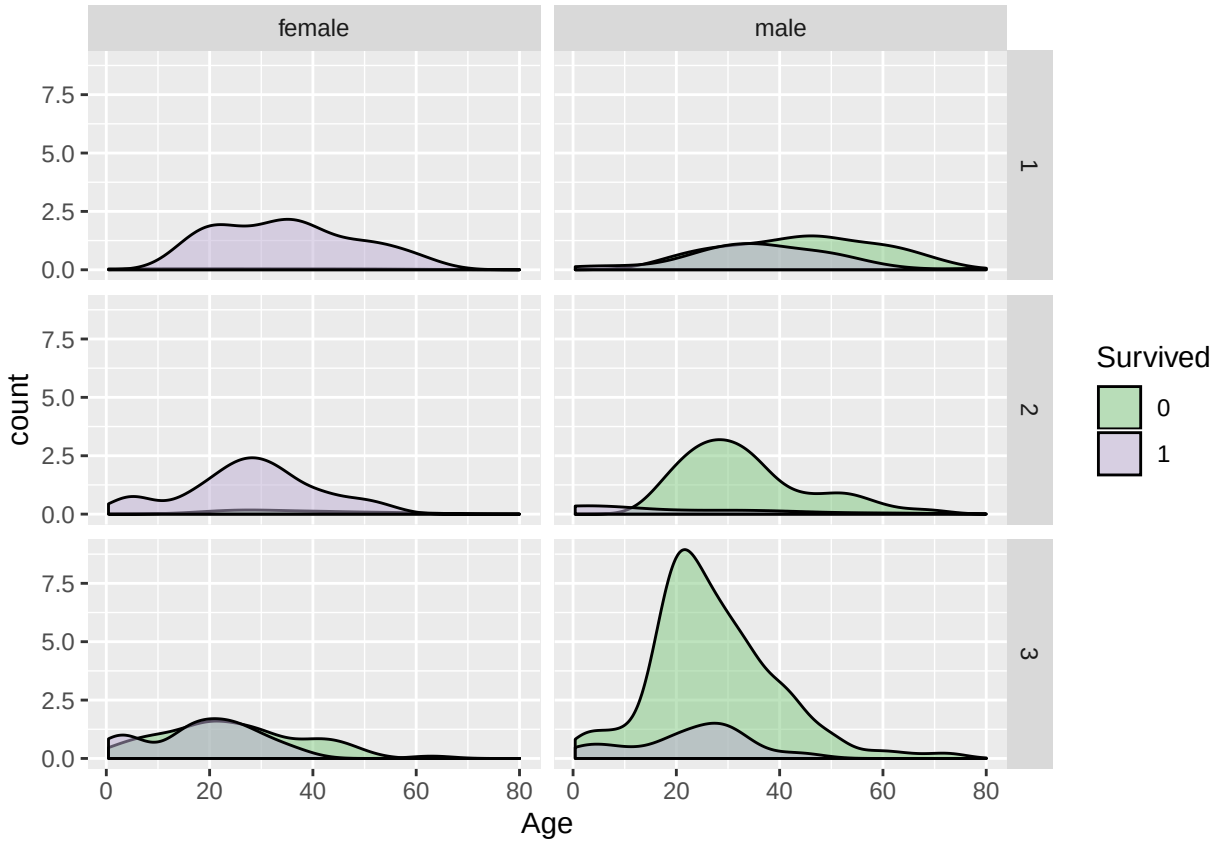

```
# Barplot3 for pclass with Survival as fill
titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    ggplot(aes(Survived, fill=Pclass)) +
    geom_bar(position = position_fill()) +
    ggtitle("Barplot for Survival filled with Pclass & with position")
```



Analysis

- first class or second class people had more chance of survival. Grid of density plot for age, filled by Survival, faceted by Sex & Pclass.

```
library(RColorBrewer)
titanic %>% filter(!is.na(Age) & !is.na(Fare)) %>%
    ggplot(aes(Age, y=..count.., fill=Survived)) +
    geom_density(alpha = 0.5) +
    facet_grid(Pclass ~ Sex) +
    scale_fill_brewer(palette = "Accent")
```



Analysis

- Third class males are the highest group on Titanic.
- Almost no male of 2nd class survived, except for children.