Ablehulu Debebe

Harvard Data Science Boot Camp

December 28,2021

Background

The Titanic was a British ocean liner that struck an iceberg and sank on its maiden voyage in 1912 from the United Kingdom to New York. More than 1,500 of the estimated 2,224 passengers and crew died in the accident, making this one of the largest maritime disasters ever outside of war. The ship carried a wide range of passengers of all ages and both genders, from luxury travelers in first-class to immigrants in the lower classes. However, not all passengers were equally likely to survive the accident. We use real data about a selection of 891 passengers to learn who was on the Titanic and which passengers were more likely to survive.

Libraries, Options, and Data

Install **titanic** package.

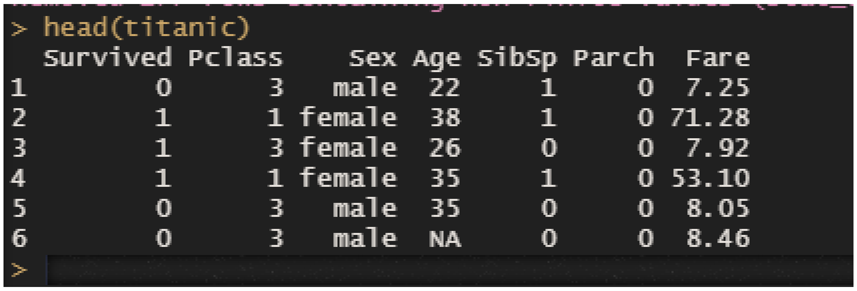
Defined the `titanic` dataset starting from the **titanic** library with the following code:

```
options(digits = 3)     # report 3 significant digits

library(tidyverse)

library(titanic)
```

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

Variable Types

1. Survived → Non-Ordinal Categorical
2. Pclass → Ordinal Categorical
3. Sex → Non-Ordinal Categorical
4. SibSP → Discrete
5. Parch → Discrete
6. Fare → Continuous

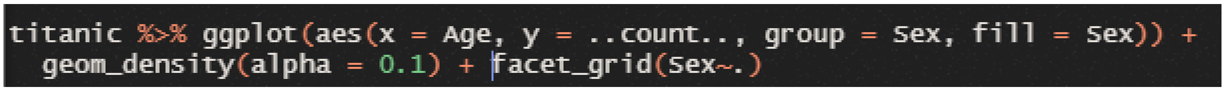


```
> head(titanic)
  Survived Pclass    Sex Age SibSp Parch  Fare
1        0      3   male  22     1     0  7.25
2        1      1 female  38     1     0 71.28
3        1      3 female  26     0     0  7.92
4        1      1 female  35     1     0 53.10
5        0      3   male  35     0     0  8.05
6        0      3   male  NA     0     0  8.46
>
```

Analysis

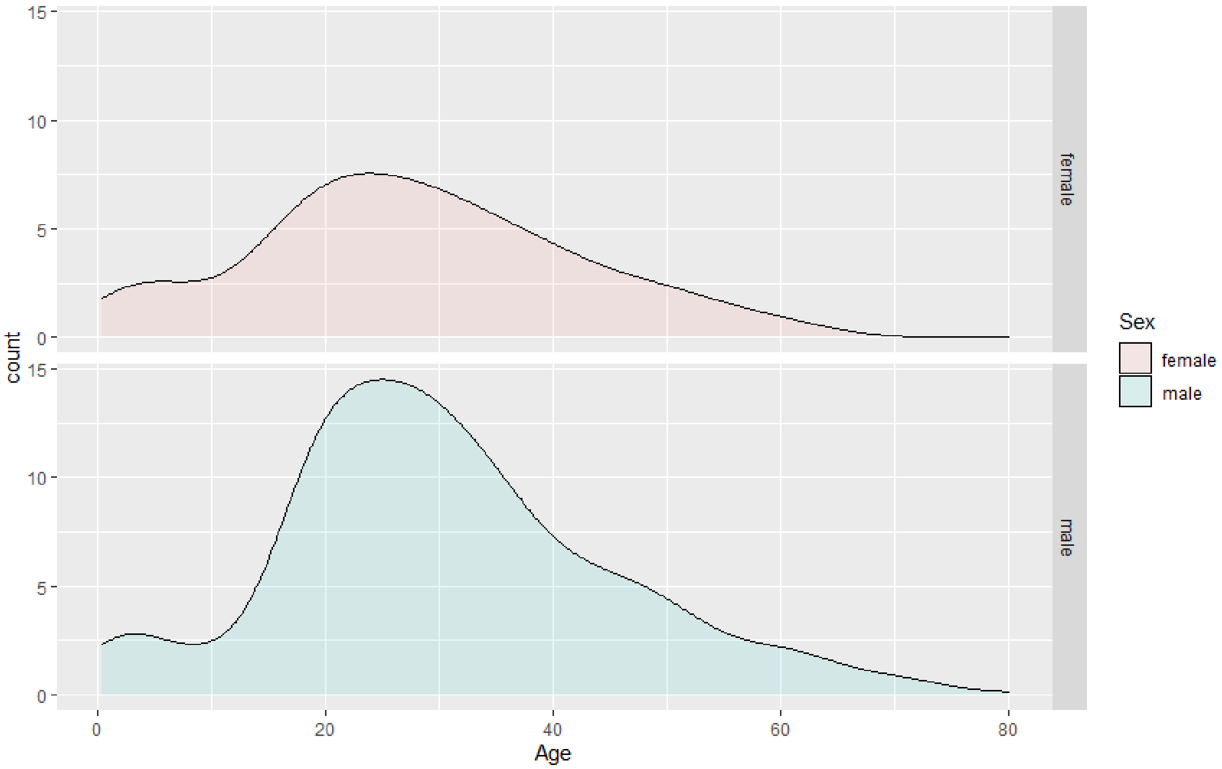
Demographics of Titanic Passengers- analysis done using density plots of age grouped by sex. Analysis with a combination of faceting, alpha blending, stacking and using variable counts on the y-axis.

A. Code →



```
titanic %>% ggplot(aes(x = Age, y = ..count.., group = Sex, fill = Sex)) +
  geom_density(alpha = 0.1) + facet_grid(Sex~.)
```

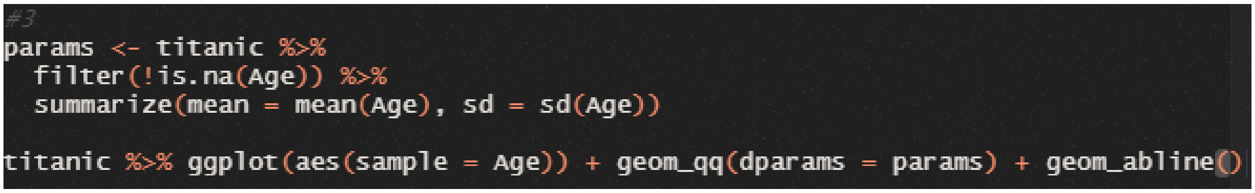
B. Density-plot →



C. Analysis → I was able to find out that females and males had the same general shape of age distribution, the age distribution was bimodal, with one mode around 25 years of age and a second smaller mode around 5 years of age and a second smaller mode around 5 years of age. The count of males of age 40 was higher than the count of females of age 40. The proportion of males age 18-35 was higher than the proportion of females age 18-35. The proportion of females under age 17 was higher than the proportion of males under age 17.

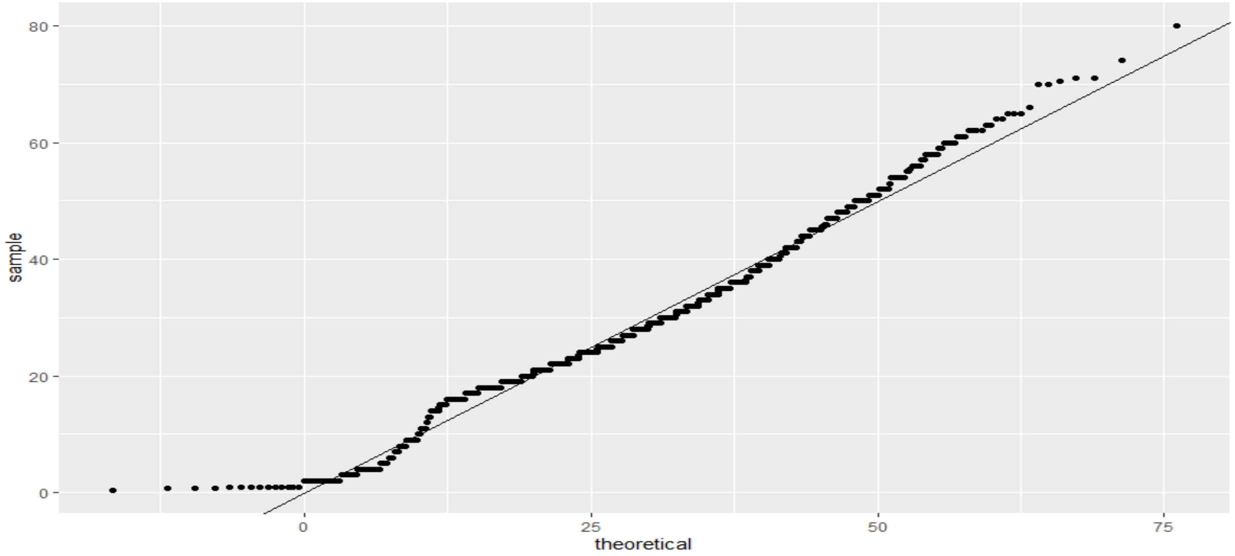
QQ-plot of Age Distribution- QQ_plot of passenger age and identity line with geom_abline(). Filtered individuals with an age of NA.

A. Code →



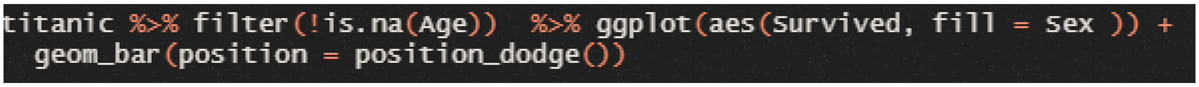
```
#3
params <- titanic %>%
  filter(!is.na(Age)) %>%
  summarize(mean = mean(Age), sd = sd(Age))

titanic %>% ggplot(aes(sample = Age)) + geom_qq(dparams = params) + geom_abline()
```

B. QQ-plot →



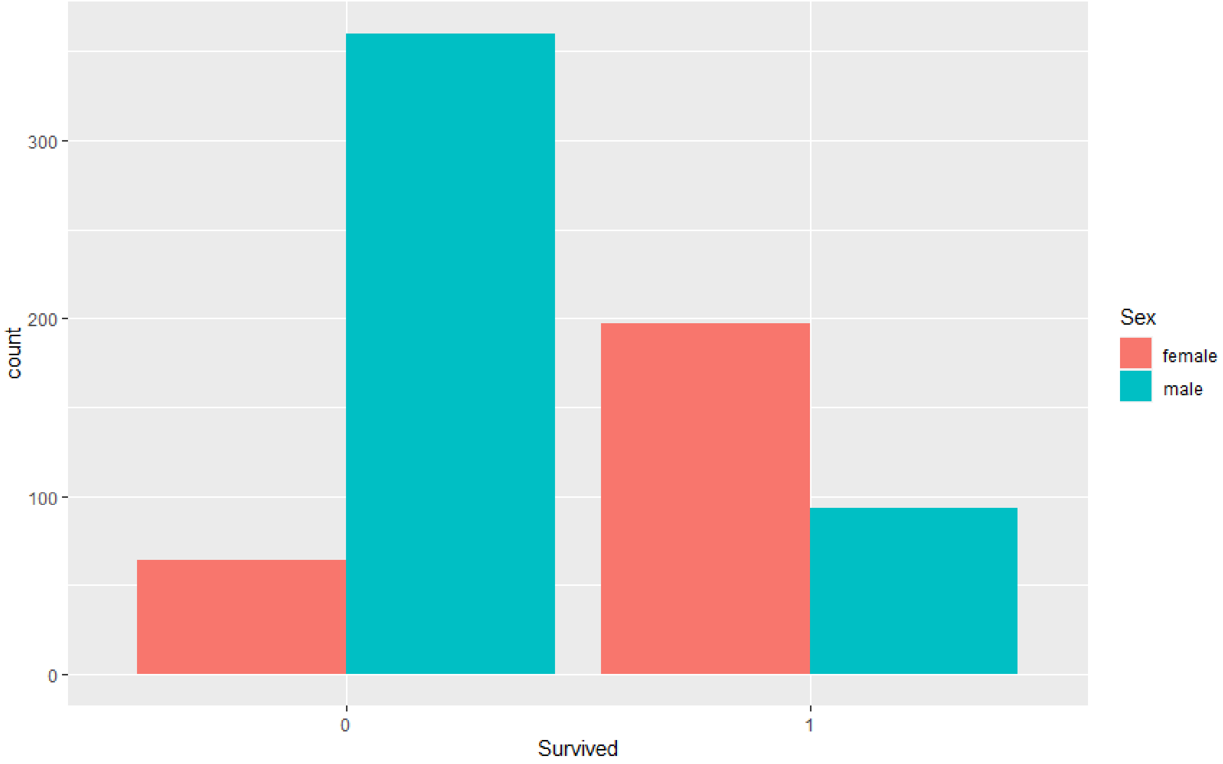
Survival by Sex- analysis done using barplots of the Survived and Sex variables using geom_bar(). Incorporated position = position_dodge() to make separate bars for each group.

A) Code →



```
titanic %>% filter(!is.na(Age))  %>% ggplot(aes(Survived, fill = Sex )) +
  geom_bar(position = position_dodge())
```

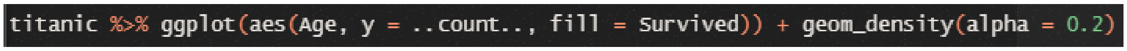
B) Bar-plot →



C) Analysis → I was able to find out that less than half of the passengers survived. Most of the survivors were female and most of the females survived.

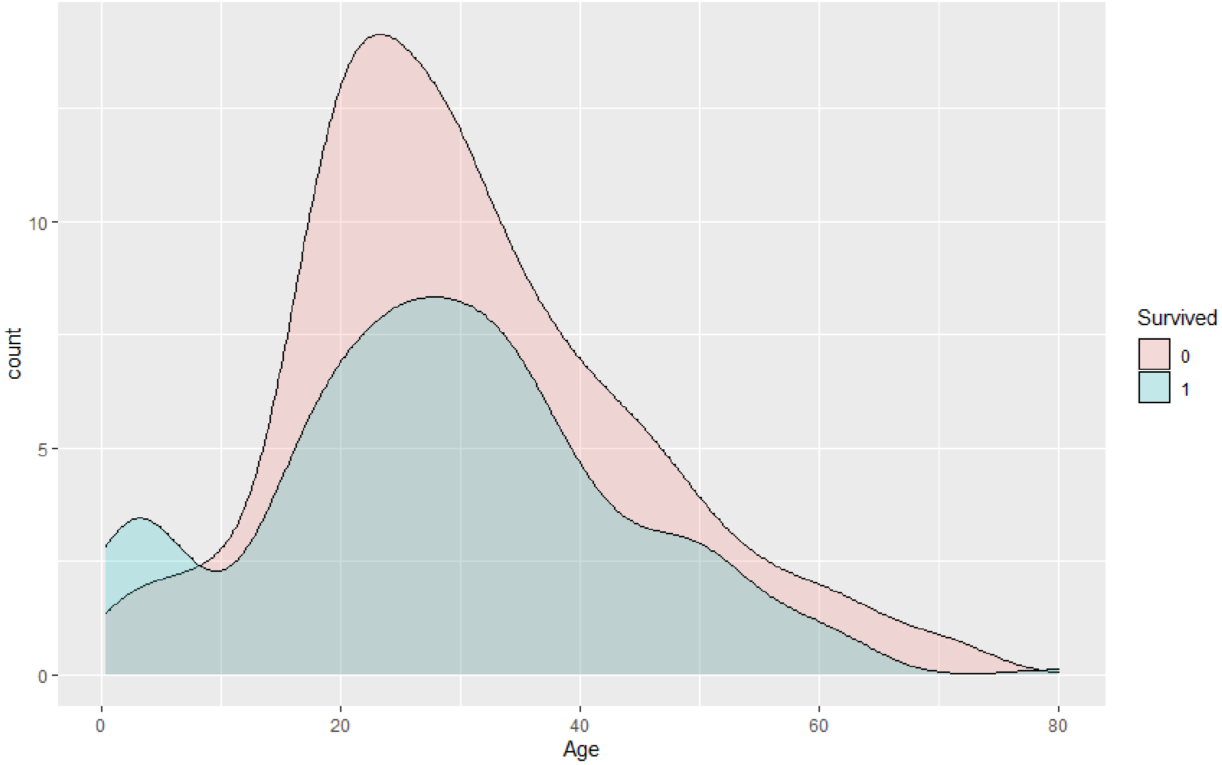
Survival by Age- analysis done using density plot of age filled by survival status. Set y-axis to count and integrate alpha blending, alpha = 0.2.

A. Code →



```
titanic %>% ggplot(aes(Age, y = ..count.., fill = Survived)) + geom_density(alpha = 0.2)
```

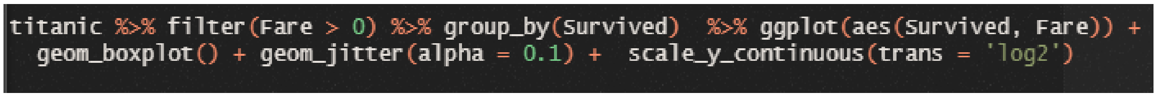
B. Density-plot →



C. Analysis → I was able to find out that the 0 - 8 age group is more likely to survive than die. Age group 18 - 30 had the most deaths. Age group 70 - 80 had the highest proportion of deaths.

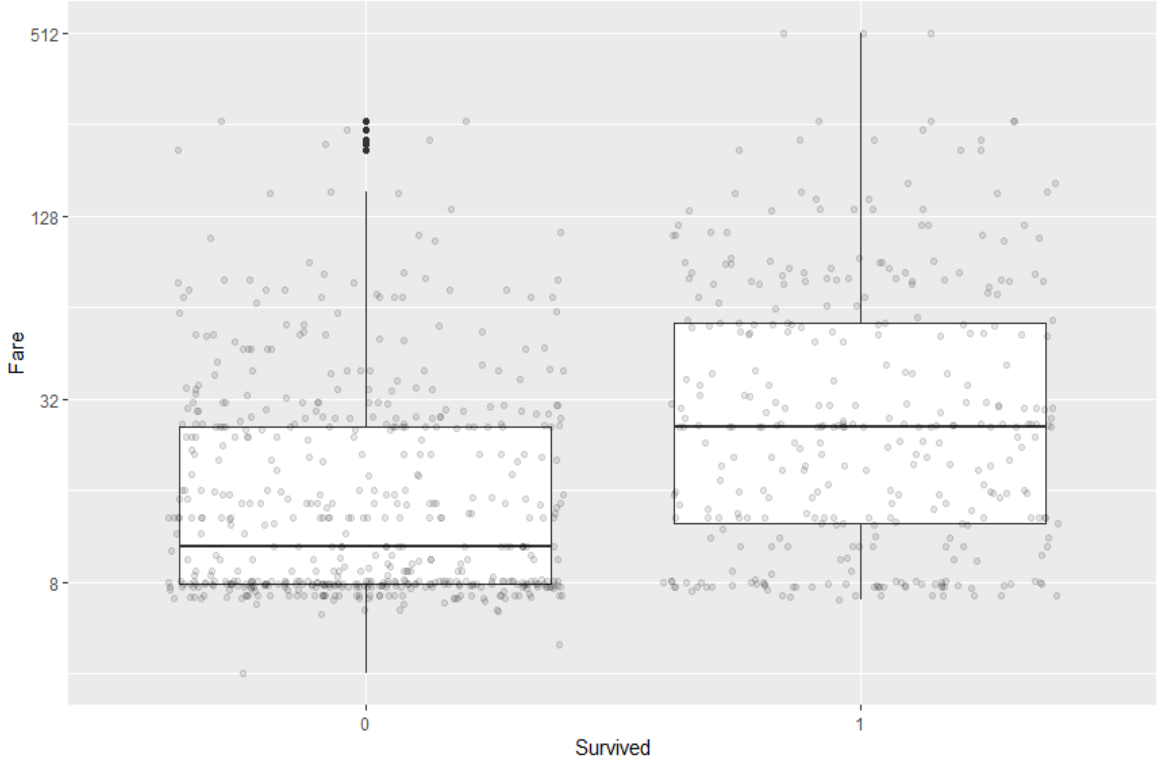
Survival by Fare- analysis done using boxplot of fare grouped by survival status. Implemented log 2 transformation of fares for better readability. Included jitter and alpha blending for better data analysis.

A. Code →



```
titanic %>% filter(Fare > 0) %>% group_by(Survived)  %>% ggplot(aes(Survived, Fare)) +
  geom_boxplot() + geom_jitter(alpha = 0.1) +  scale_y_continuous(trans = 'log2')
```

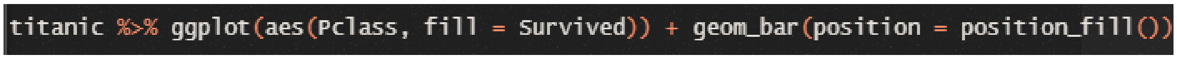
B. Density-plot →



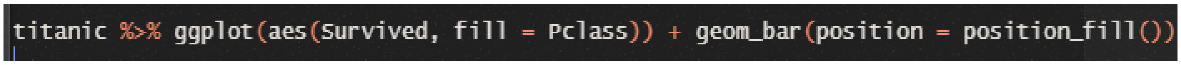
C. Analysis → I was able to find out that passengers who survived generally paid higher fares than those who did not survive. The median fare was lower for passengers who did not survive. Most individuals who paid a fare around $8 did not survive.

Survival by Passenger Class- analysis done using a basic bar plot of passenger class filled by survival, same plot using the argument position = position_fill() and a barplot of survival filled by passenger class.

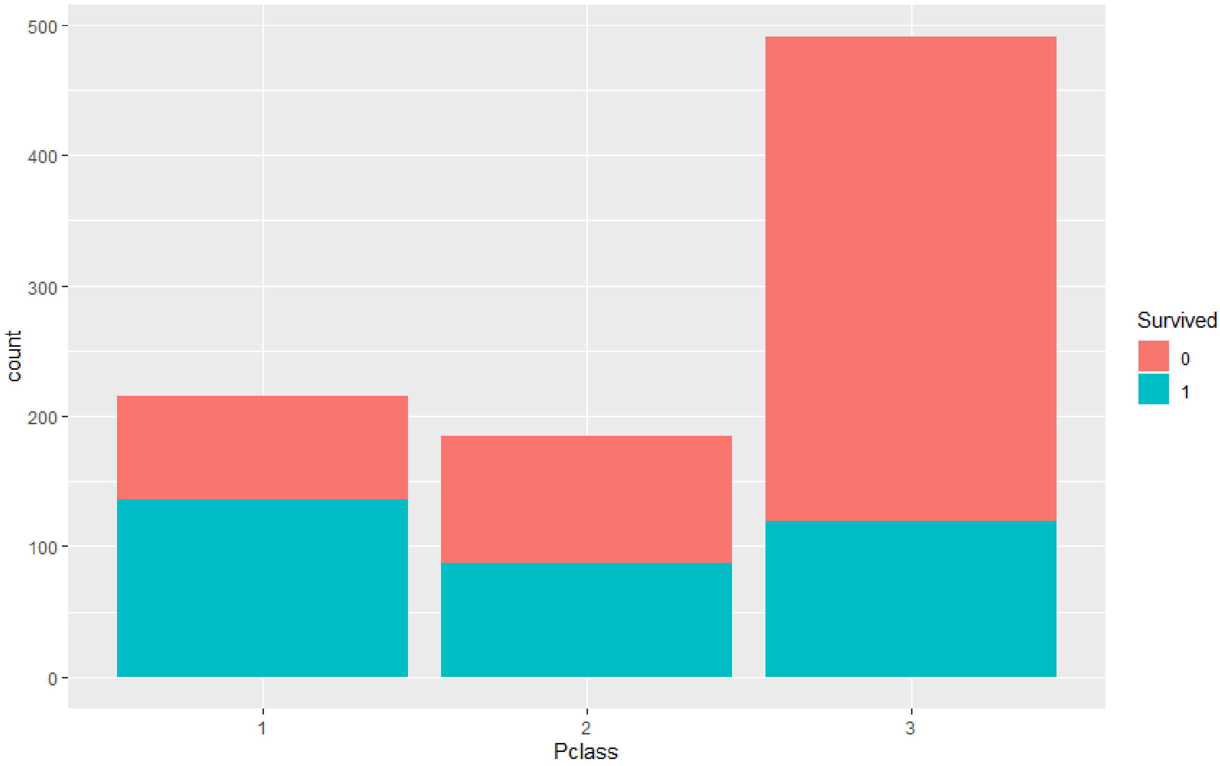
A. Code →



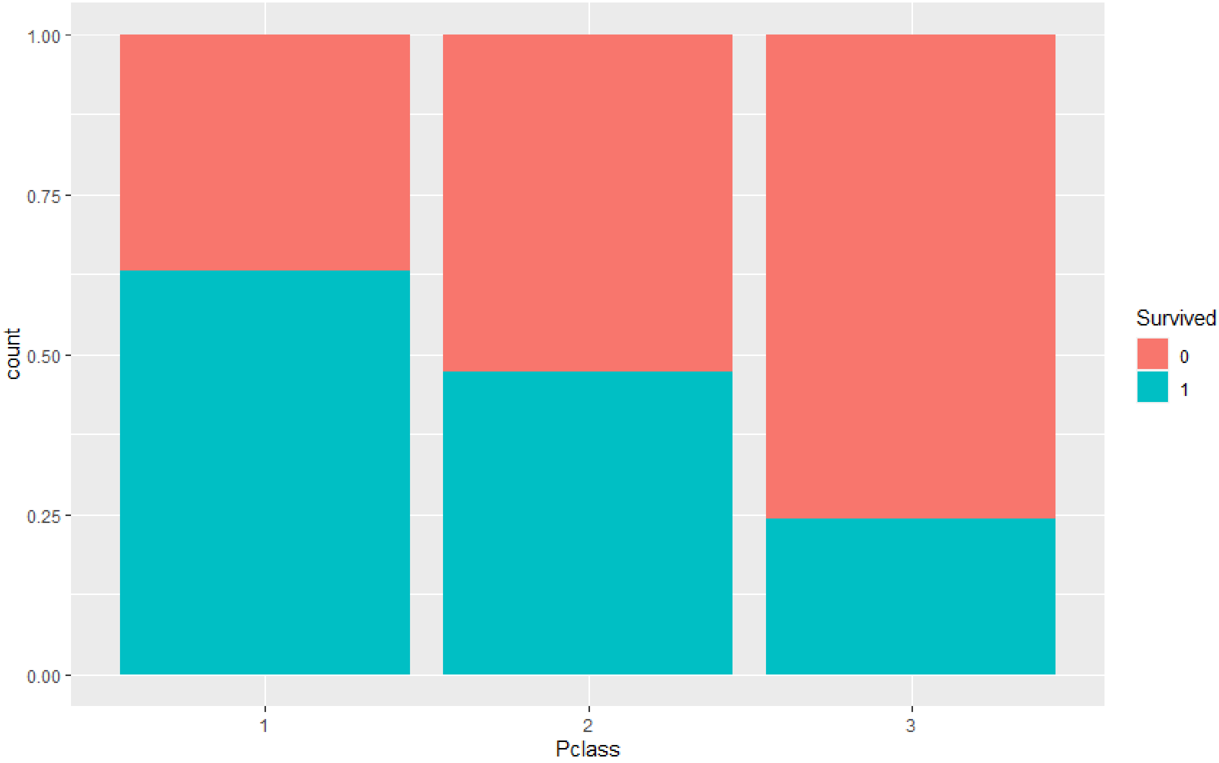
```
titanic %>% ggplot(aes(Pclass, fill = Survived)) + geom_bar(position = position_fill())
```

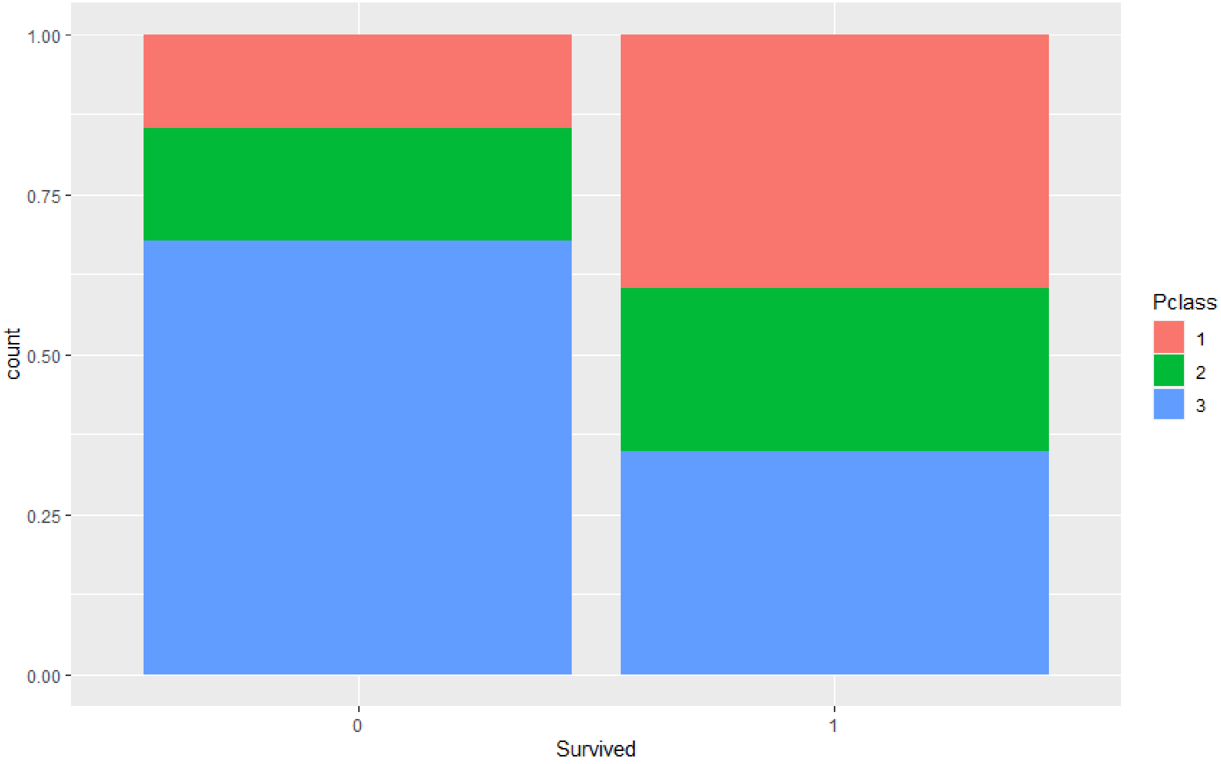


```
titanic %>% ggplot(aes(Survived, fill = Pclass)) + geom_bar(position = position_fill())
```



B. Bar-plot →

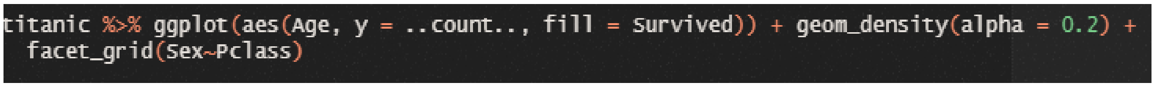

1.00
0.75
0.50
0.25
0.00
count
1
2
3
Pclass
Survived
0
1


1.00
0.75
0.50
0.25
0.00
count
0
1
Survived
Pclass
1
2
3

C. Analysis → I was able to find out that there were more third class passengers than passengers in the first two classes combined. Survival proportion was highest for first class passengers, followed by second class. Third-class had the lowest survival proportion. Most passengers in the first class survived. Most passengers in other classes did not survive. The majority of those who did not survive were from third class.

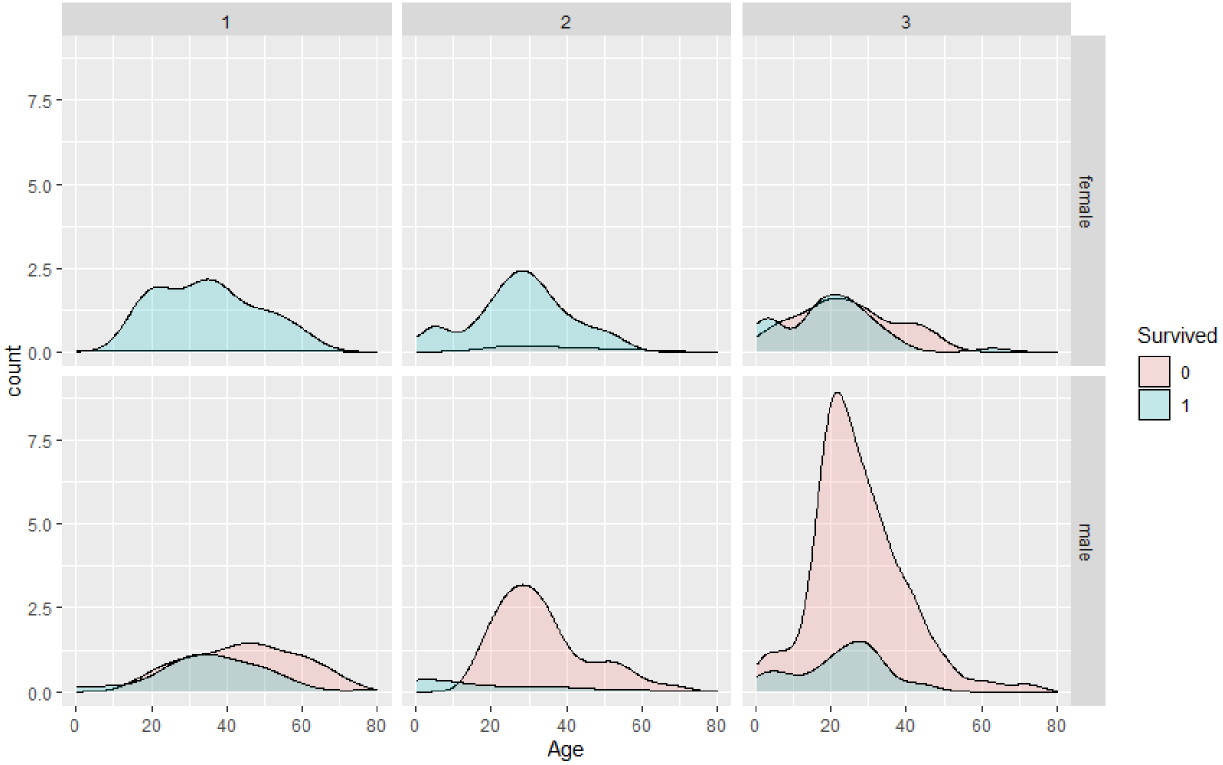
Survival by Age, Sex and Passenger Class- analysis done using a grid of density plots for age, filled by survival status, with count on the y-axis, faceted by sex and passenger class.

A. Code →



```
titanic %>% ggplot(aes(Age, y = ..count.., fill = Survived)) + geom_density(alpha = 0.2) +
  facet_grid(Sex~Pclass)
```

B. Grid density plot →



C. Analysis → I was able to find out that the largest group of passengers was third-class males. Most first-class and second-class females survived. Almost all second-class males did not survive, with the exception of children.