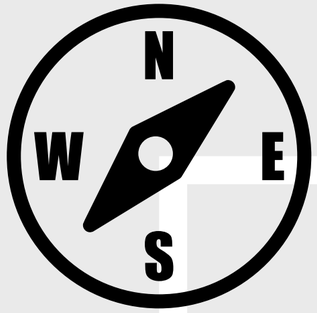
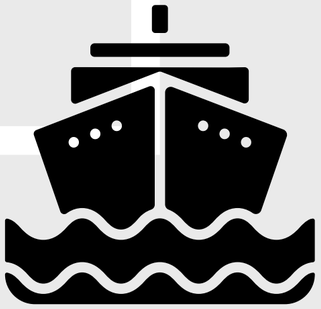


EDA ON TITANIC DATASET



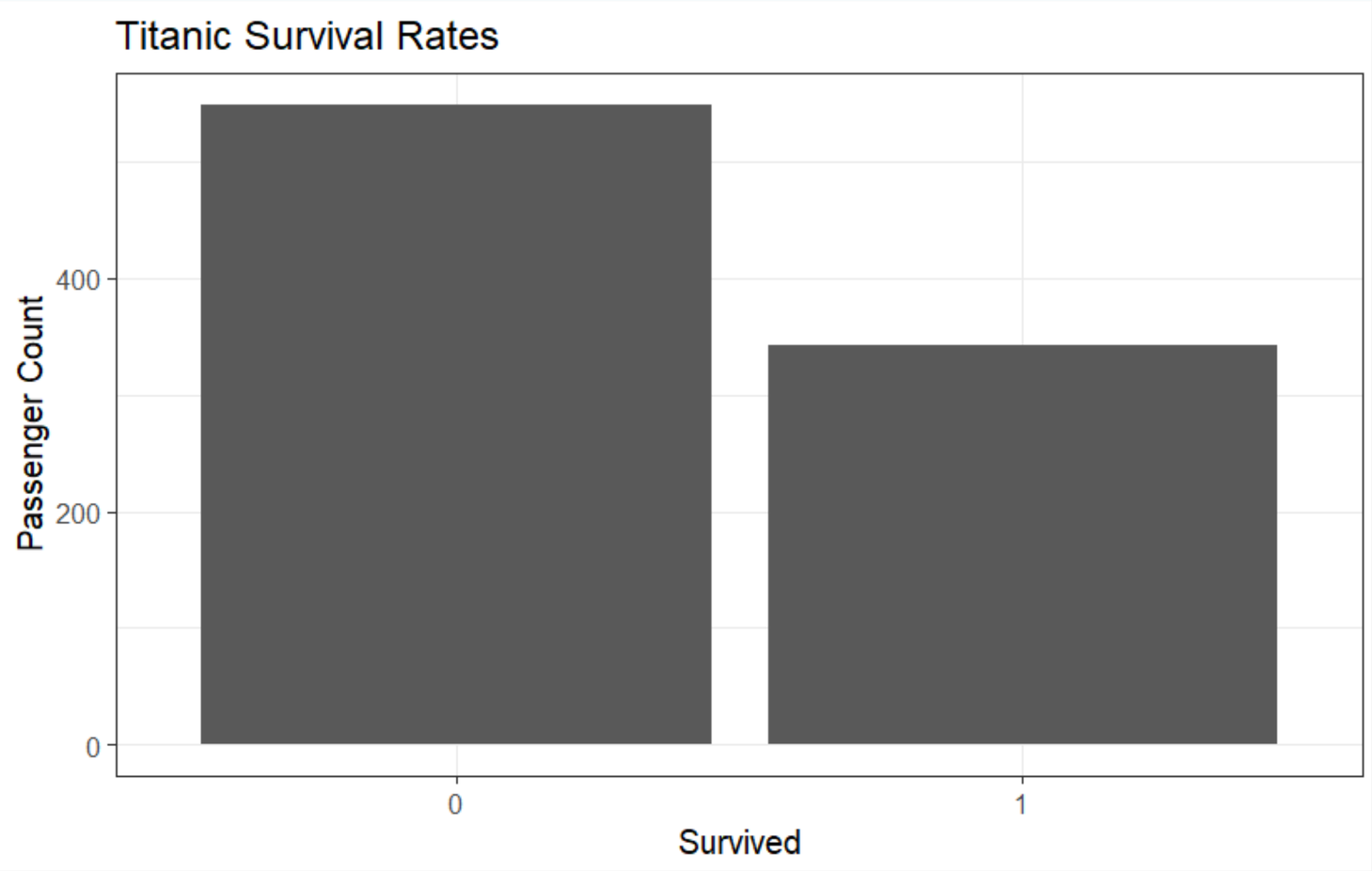
Data Dictionary

survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	Number of siblings / spouses	
	aboard the Titanic	
parch	Number of parents / children	
	aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Questions answered

1. What was the survival rate?
2. What was the survival rate by gender?
3. What was the survival rate by class of ticket?
4. What was the survival rate by class of ticket and gender?
5. What is the distribution of passenger ages?
6. What are the survival rates by age?
7. What is the survival rates by age when segmented by gender and class of ticket?

```r
ggplot(titanic, aes(x = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates")
```



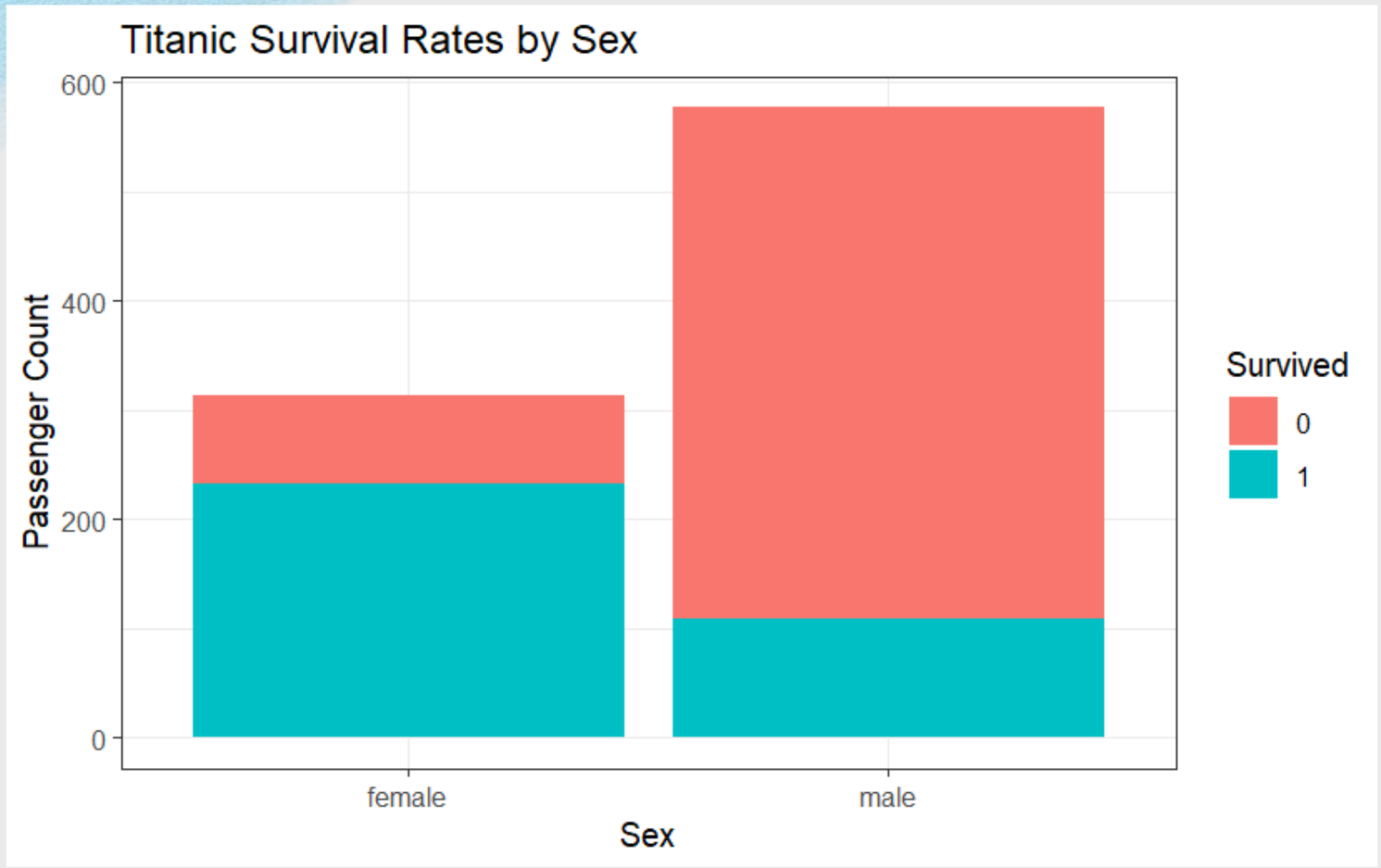
0 - No, 1 - Yes

Observation 1:

The number of people who perished on the Titanic is more than those who survived.

Observation 2:

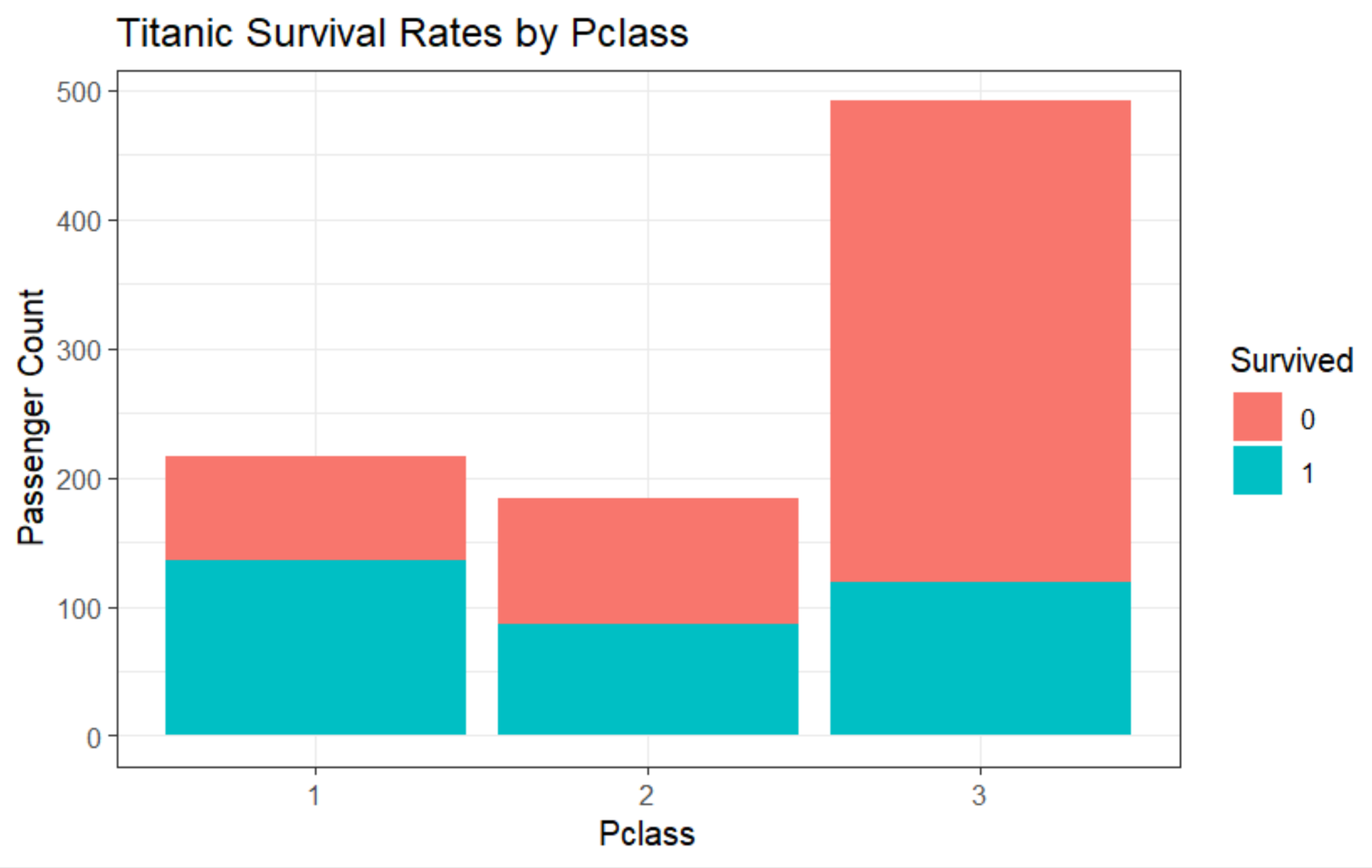
Females survived disproportionately more than males did on the Titanic.
It also tells us that there were twice as many males approximately than there were females in the Titanic




```
ggplot(titanic, aes(x = Sex, fill = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Sex")
```

0 - No, 1 - Yes

```
ggplot(titanic, aes(x = Pclass, fill = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Pclass")
```

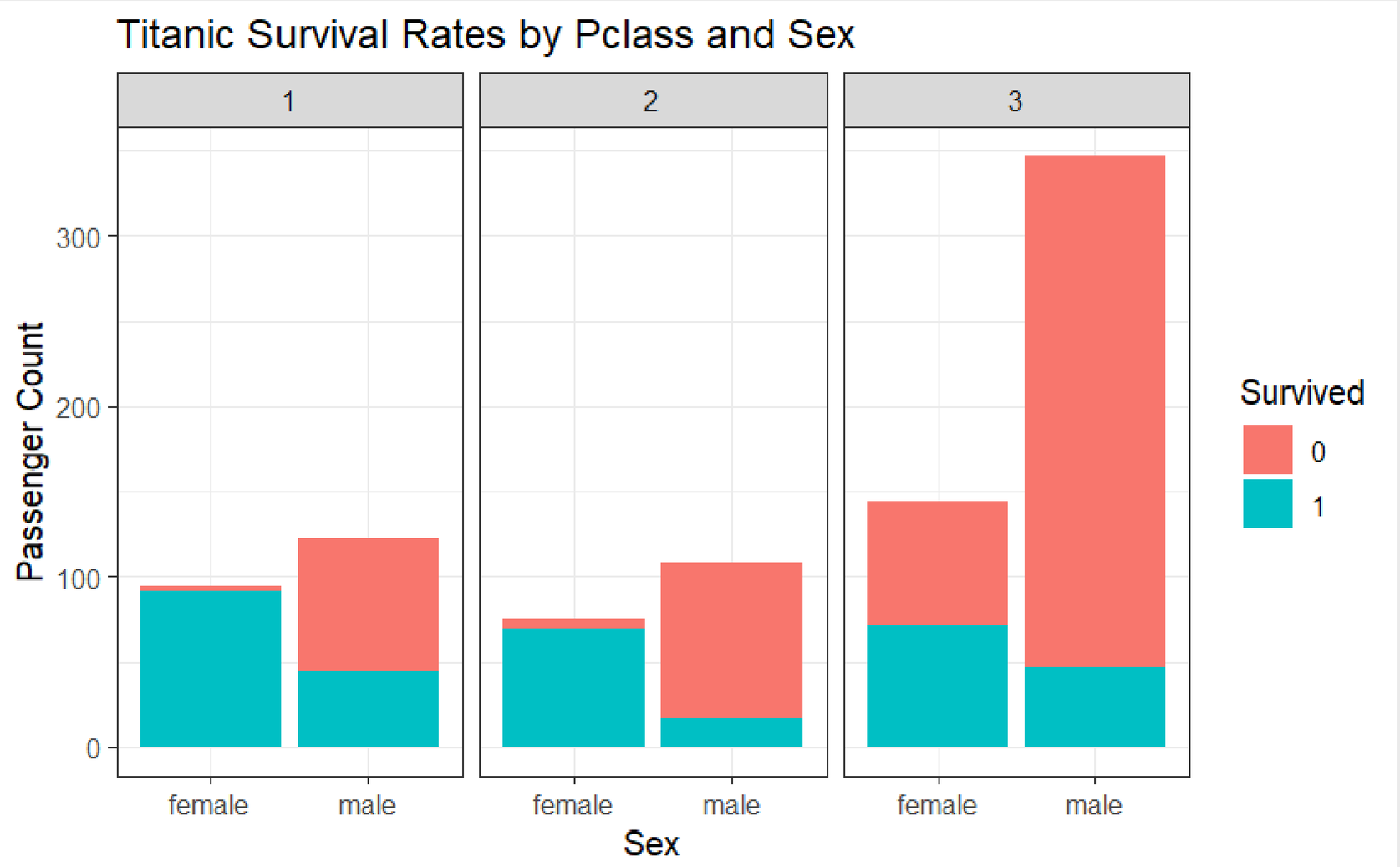


Observation 3:

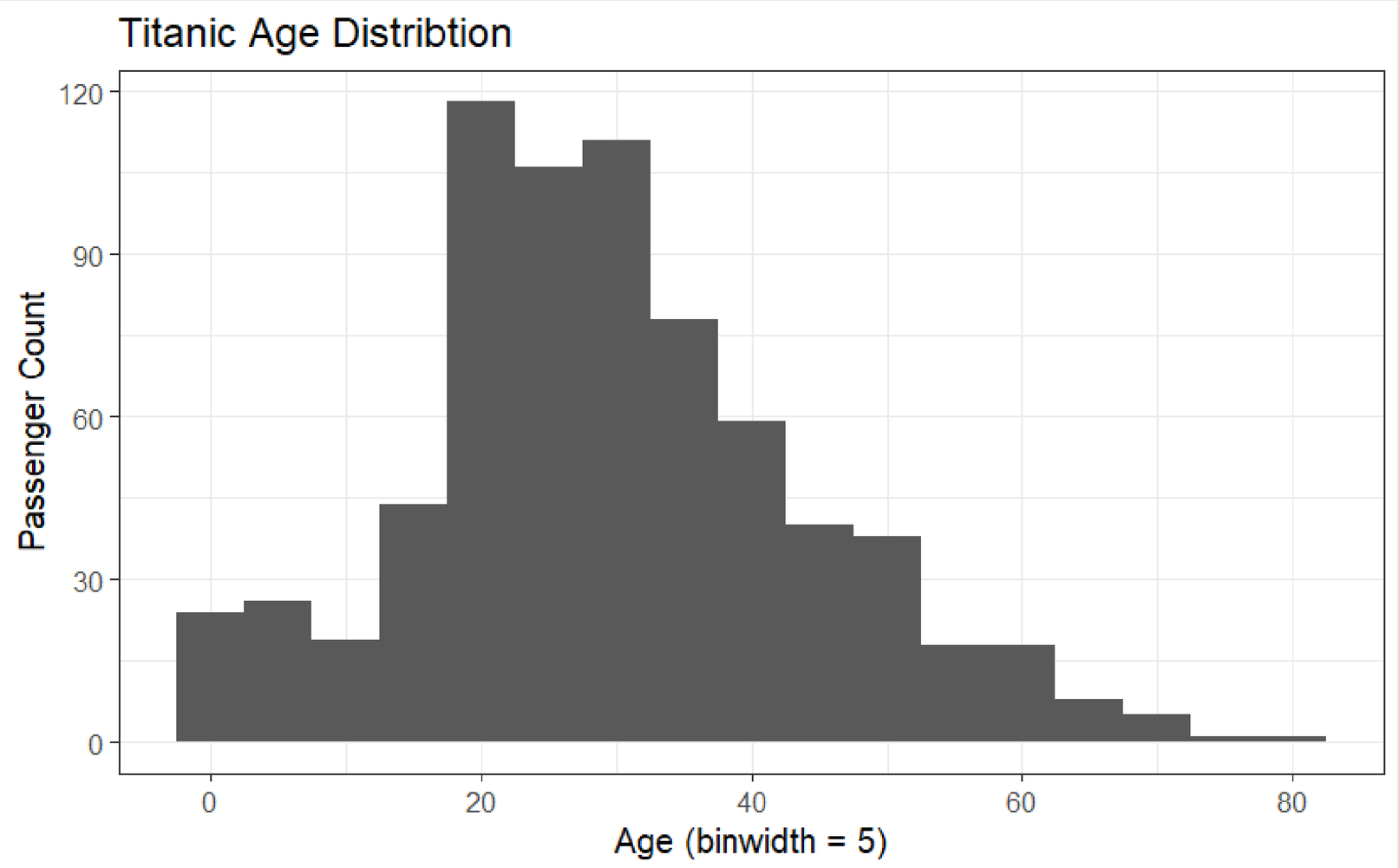
People in 3rd class had pretty poor survival rates, maybe 3:1 ratio.
And in 1st class reflexively, more than half survived.
There were less 2nd class passengers than either 1st or 3rd class passengers.

Observation 4:

Females disproportionately survived more than the males. In the 3rd class, each had approximately 50-50 chances of survival. Males in the 1st class had the highest levels of survival as compared to those in the 2nd and 3rd class.
Also, there are way more males in the 3rd class than there are females.




```
ggplot(titanic, aes(x = Sex, fill = Survived)) +
  theme_bw() +
  facet_wrap(~ Pclass) +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Pclass and Sex")
```

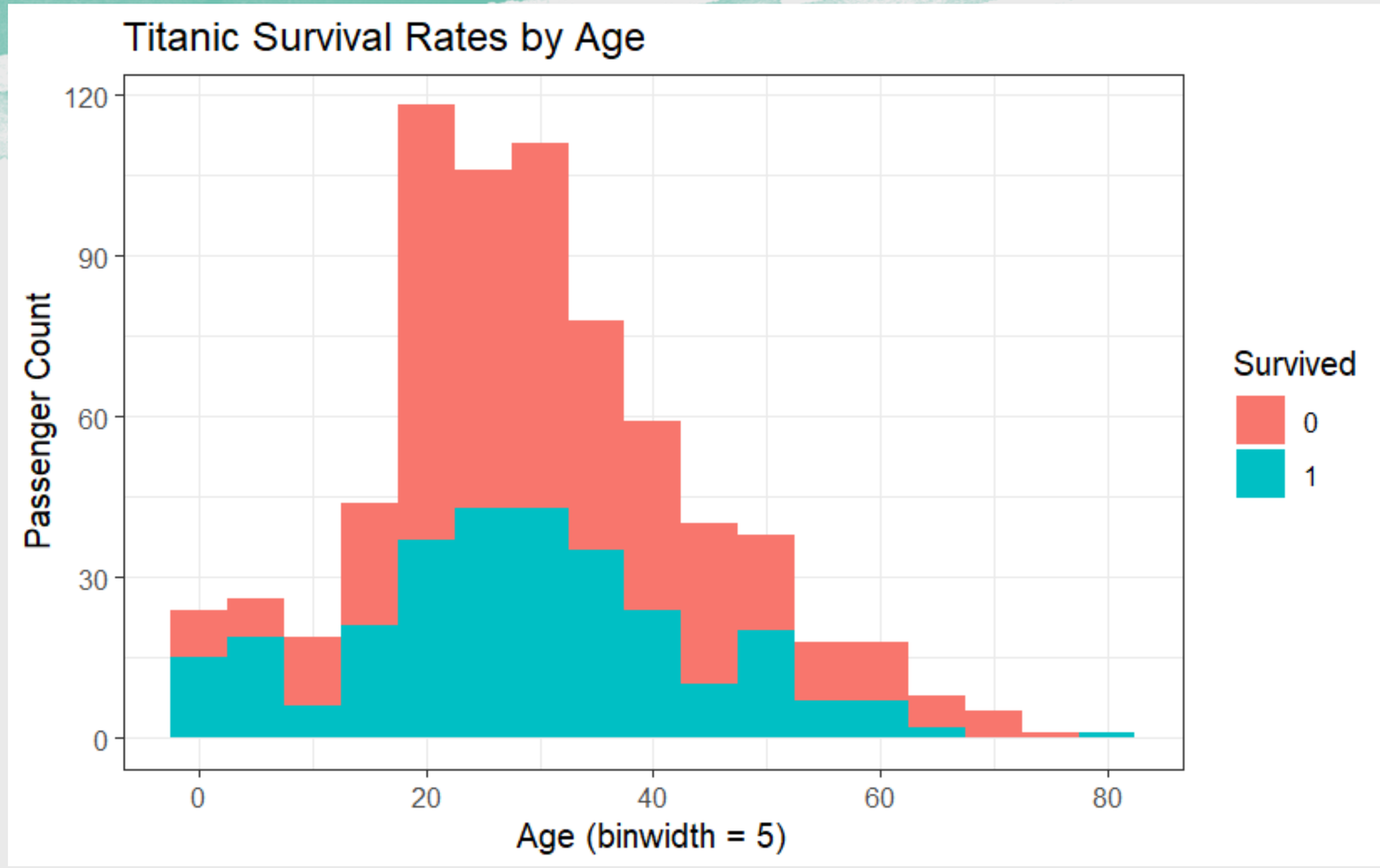


Observation 5:

Older passengers weren't that prevalent.
There were some children as well.
Many people belonged to the age group of 20-40.

```
ggplot(titanic, aes(x = Age)) +
  theme_bw() +
  geom_histogram(binwidth = 5) +
  labs(y = "Passenger Count",
       x = "Age (binwidth = 5)",
       title = "Titanic Age Distribtion")
```

```r
ggplot(titanic, aes(x = Age, fill = Survived)) +
  theme_bw() +
  geom_histogram(binwidth = 5) +
  labs(y = "Passenger Count",
       x = "Age (binwidth = 5)",
       title = "Titanic Survival Rates by Age")
```

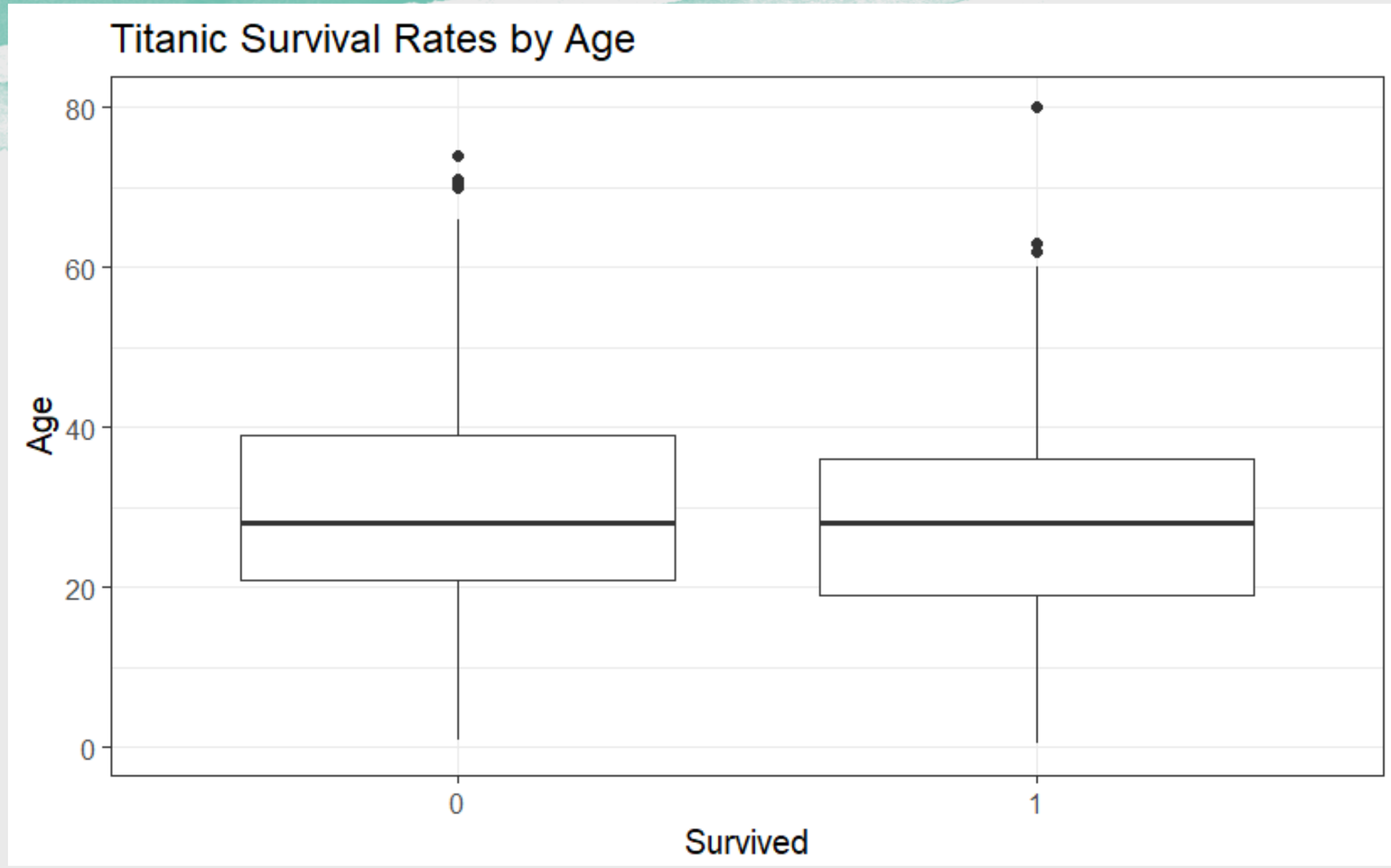


Observation 6:

More than half of the children at the younger end survived.
At the higher end (50+) of the spectrum, survivability is quite low.

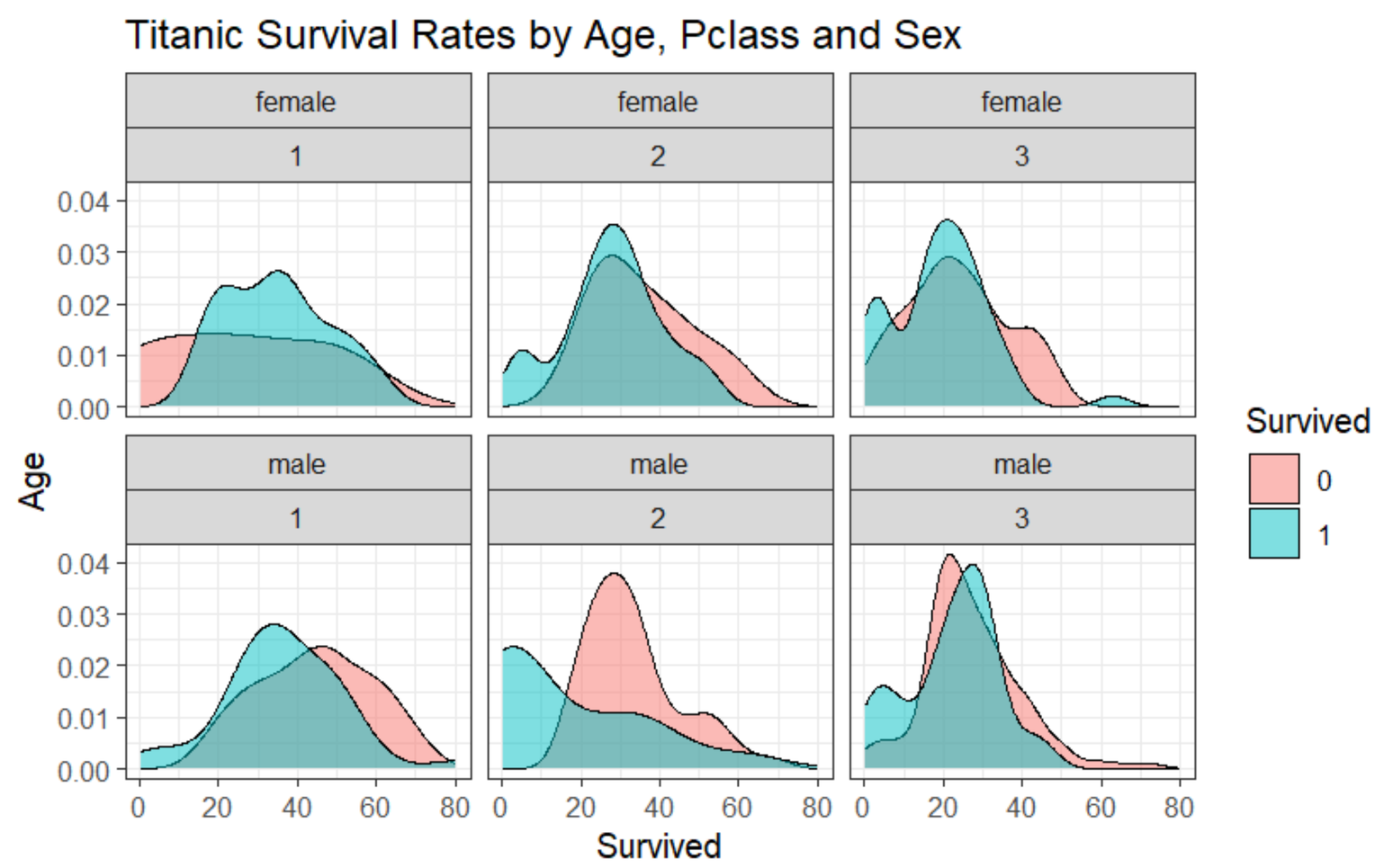
NOTE: Another great visualization for this question is the box-and-whisker plot.

```r
ggplot(titanic, aes(x = Survived, y = Age)) +
  theme_bw() +
  geom_boxplot() +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age")
```



Observation 6:

In general, people who survived tended to be younger than those who perished but only by a little bit.




```
ggplot(titanic, aes(x = Age, fill = Survived)) +
  theme_bw() +
  facet_wrap(Sex ~ Pclass) +
  geom_density(alpha = 0.5) +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age, Pclass and Sex")
```

Observation 7:

The youngest children in both females and males disproportionately survived in the 2nd class.

NOTE: We can do the same thing using Histograms instead.


```
ggplot(titanic, aes(x = Age, fill = Survived)) +
  theme_bw() +
  facet_wrap(Sex ~ Pclass) +
  geom_histogram(binwidth = 5) +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age, Pclass and Sex")
```

Observation 7:

The girls (younger females) have disproportionately survived in almost all the passenger classes.
All the boys (younger males) in the 1st and 2nd class survived.
For males, as you get older in 3rd class, the survival rates become really bad.

Thank you!

Priyanshi Negi
RA2111027010135