HW3_JiangShu

JiangShu

2022-10-19

same family member's ticket id is the same

Convert survived and pclass to factor; reorder survived to make "Yes" comes first

```
titanic$survived <- as.factor(titanic$survived)
titanic$survived <- factor(titanic$survived, levels=c("Yes","No"))
titanic$pclass <- as.factor(titanic$pclass)
```

Question 1

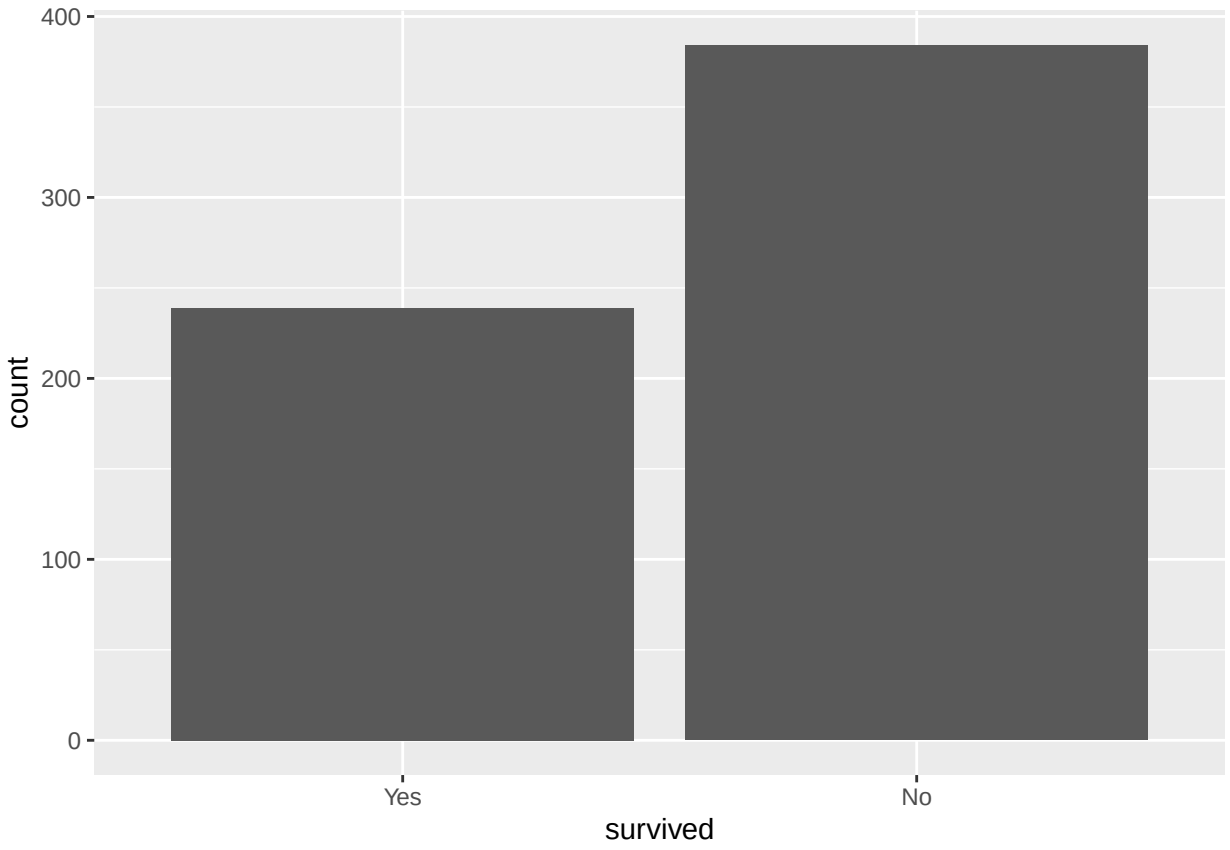
```
set.seed(1112)

titanic_split <- initial_split(titanic, prop = 0.70,
                               strata = survived)
titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)
```

- The total sample size is 891. The training data set size is set to be 70% of the total sample size, which gives 623 observations, while the other 30% of total sample data comprises test size and gives 268 observations.
- There are missing values in predictors like "age", "cabin","embarked". I don't think cabin embarked would cause issue since I don't think it is related to survived at all. For embarked, there're only 2 missing values, so we could fill them with the most common value. For age there might be some impacts, so I might drop rows with missing age. I also notice there are some 0 in "fare" column, which should be investigated cause a free ticket would impact our model on surviving.
- The outcome "survived" might be depend on predictors like "pclass", where first class passenger has a higher surviving probability. We would want to ensure the split data still follows normal distribution so that the training model we are using won't have extreme values. For instance, if not using stratified sampling, there's some chance that R randomly selects most of the survived case into training data, which makes the model we generated tends to predict every case to be survived.

Question 2

```
titanic_train %>%
  ggplot(aes(x = survived)) +
  geom_bar()
```



- The number of survived is less than the number of not survived, even though I used stratified sampling. This means the chance of survival in this titanic wreck is rather low.

Question 3