Assignment 10: Surviving the Titanic

Sunehera Hasib

2022-04-17

Exercise 1

i.

```
train_df <- read_csv(file = "train.csv",
                     col_types= cols(
                       Pclass = col_character(),
                       SibSp = col_character(),
                       Parch = col_character()
                     )
)
```

ii.

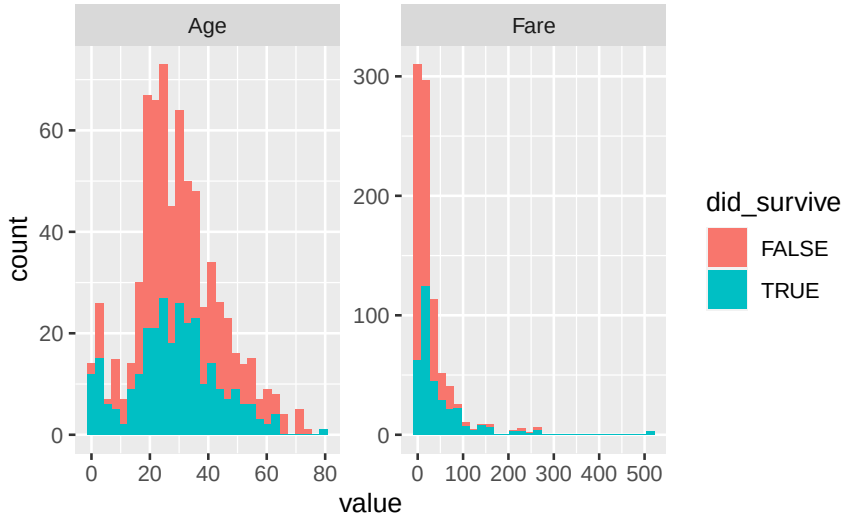
```
train_df <- train_df %>%
  mutate(did_survive = as.logical(Survived))
```

Exercise 2

```
train_df %>%
  pivot_longer(cols = c(Age,Fare), names_to="measurement", values_to = "value") %>%
  ggplot() +
  geom_histogram(aes(x = value, fill = did_survive)) +
  facet_wrap(~ measurement, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

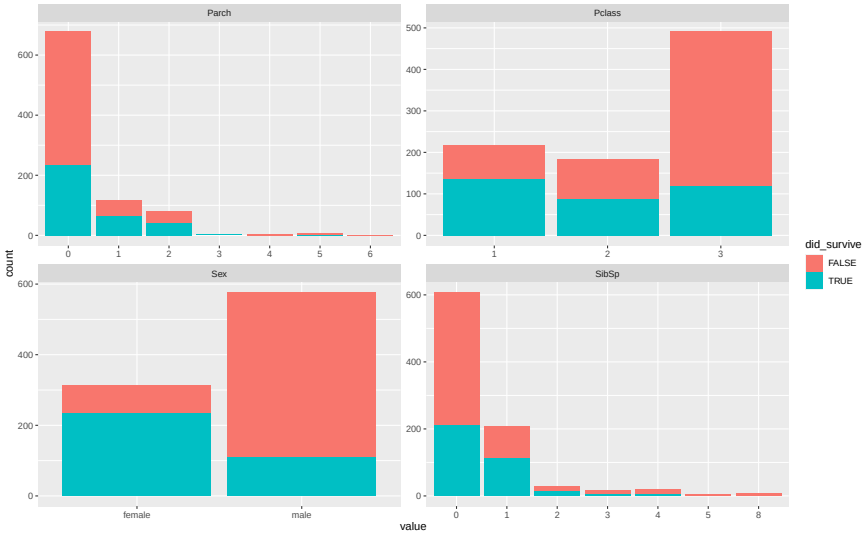
```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



In age there is a peak from the age of 20 - 40 who did not survive those younger.The same applies for survivors however, younger people had a higher survival rate and an outlier of older people of age >75 too had survived compared to people age < 75. On the other hand, people with lower fare did not survive and there is a peak on the lowest fares compared to people with higher fares. I think the fare and age will be helpful for predicting who survived.

Exercise 3

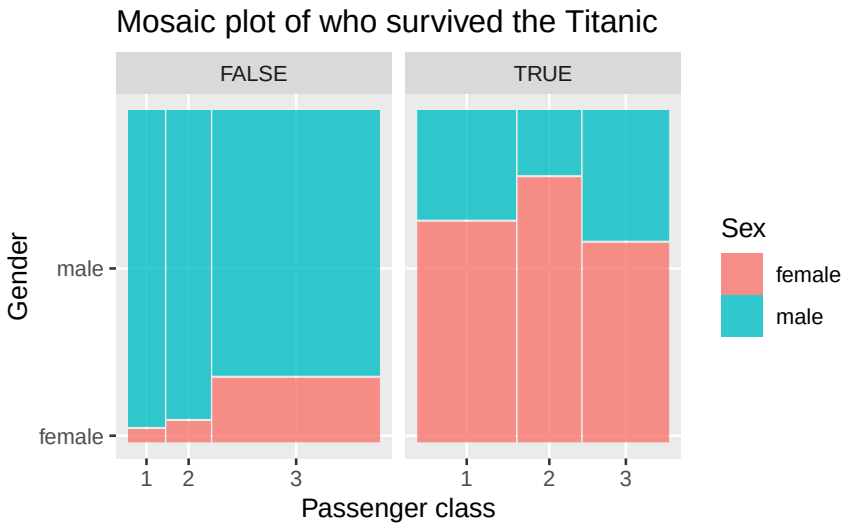
```
train_df %>%
  pivot_longer(cols = c(Pclass, Sex, Parch, SibSp), names_to="measurement", values_to = "value"
  ggplot() +
  geom_bar(aes(x = value, fill = did_survive)) +
  facet_wrap(~measurement, scales = "free")
```



In Pclass, people with Pclass = 1 survived the most, then Pclass = 2 and then Pclass =3, which higher the passenger class greater the survival rate. There were higher female survivors than male even though there were more male passengers. For both Parch and SibSp, since both are family members accompanying it can be deduced that lesser the count of family members, higher the chance that passenger did not survive. I think the variable Pclass is the most useful variable as Pclass = 1 had the highest count of survivors.

Exercise 4

```
train_df %>%
  ggplot() +
  geom_mosaic(mapping = aes(x = product(Sex, Pclass), fill = Sex)) +
  facet_grid(. ~ did_survive, scales="free") +
  labs(x = "Passenger class", y = "Gender", title = "Mosaic plot of who survived the Titanic")
```



Most passengers on Titanic were males, more females survived in all classes than males. There were more survivors in Pclass =1 for both genders. However, even with lowest Pclass = 3 survivors more women survived compared to men

Exercise 5

```
train_df %>%
  summarize(
    count = n(),
    missing = sum(is.na(Age)),
    fraction_missing = missing %% 100
  )
```

count	missing	fraction_missing
891	177	77

ii.

```
train_imputed <- train_df %>%
  mutate(
    age_imputed = if_else(
      condition = is.na(Age),
      true = median(Age, na.rm = TRUE),
      false = Age
    )
  )
```

iii.

```
train_imputed %>%
  summarize(
    count = n(),
    missing = sum(is.na(age_imputed)),
    fraction_missing = missing %% 100
  )
```

count	missing	fraction_missing
891	0	0

Exercise 6

i.

```
model_1 <- glm(
  Survived ~ age_imputed,
  family = binomial(),
  data = train_imputed
)
```

ii.

```
model_1_preds <- train_imputed %>%
  add_predictions(
    model_1,
    type = "response"
```

```
  ) %>%
  mutate(
    outcome = if_else(
      condition = pred > 0.5,
      true = 1,
      false = 0
    )
  )
```

iii.

```
model_1_preds %>%
  mutate(
    correct = if_else(
      condition = (Survived == outcome),
      true = 1,
      false = 0
    )
  ) %>%
  summarize(
    total_correct = sum(correct),
    accuracy = total_correct/n()
  )
```

total_correct	accuracy
549	0.6161616

Exercise 7

```
logistic_cv1 <- cv.glm(train_imputed, model_1, cost, K=5)
logistic_cv1$delta
```

```
## [1] 0.3838384 0.3838384
```

Exercise 8

i.

```
model_2 <- glm(
  Survived ~ age_imputed + SibSp + Pclass + Sex,
  family = binomial(),
  data = train_imputed
```

```
)



model_2_preds <- train_imputed %>%
  add_predictions(
    model_2,
    type = "response"
  ) %>%
  mutate(
    outcome = if_else(
      condition = pred > 0.5,
      true = 1,
      false = 0
    )
  )


model_2_preds %>%
  mutate(
    correct = if_else(
      condition = (Survived == outcome),
      true = 1,
      false = 0
    )
  ) %>%
  summarize(
    total_correct = sum(correct),
    accuracy = total_correct/n()
  )
```

total_correct	accuracy
715	0.8024691

```
logistic_cv2 <- cv.glm(train_imputed, model_2, cost, K=5)
logistic_cv2$delta
```

```
## [1] 0.1975309 0.1982148
```

ii.

```
model_3 <- glm(
  Survived ~ age_imputed * Pclass * Sex + SibSp,
  family = binomial(),
```

```
  data = train_imputed
)



model_3_preds <- train_imputed %>%
  add_predictions(
    model_3,
    type = "response"
  ) %>%
  mutate(
    outcome = if_else(
      condition = pred > 0.5,
      true = 1,
      false = 0
    )
  )


model_3_preds %>%
  mutate(
    correct = if_else(
      condition = (Survived == outcome),
      true = 1,
      false = 0
    )
  ) %>%
  summarize(
    total_correct = sum(correct),
    accuracy = total_correct/n()
  )
```

total_correct	accuracy
733	0.8226712

```
logistic_cv3 <- cv.glm(train_imputed, model_3, cost, K=5)
logistic_cv3$delta
```

```
## [1] 0.2031425 0.1966416
```

iii. In my opinion, the third interacting model had the highest accuracy and the most accurate validation error. As there are more explanatory variable there is a better chance of prediction and less chance of wrong prediction. Since predictive power of a model usually increases with its complexity, up to a point it should also make sense for the thirs model compared to the other two.

Bonus Exercise

i.

```
test_df <- read_csv(file = "test.csv",
                    col_types= cols(
                      Pclass = col_character(),
                      SibSp = col_character(),
                      Parch = col_character()
                    )
)

test_imputed <- test_df %>%
  mutate(
    age_imputed = if_else(
      condition = is.na(Age),
      true = median(Age, na.rm = TRUE),
      false = Age
    )
  )

test_imputed %>%
  summarize(
    count = n(),
    missing = sum(is.na(age_imputed)),
    fraction_missing = missing %% 100
  )
```

count	missing	fraction_missing
418	0	0

```
model_4_preds <- test_imputed %>%
  add_predictions(
    model_3,
    type = "response"
  ) %>%
  mutate(
    outcome = if_else(
      condition = pred > 0.5,
      true = 1,
      false = 0
    )
  )
```

```
colnames(model_4_preds)[colnames(model_4_preds)=="outcome"] <- "Survived"
test_df2 <- select(model_4_preds, PassengerId, Survived)

write_csv(test_df2,"test_survived.csv")
```