Assignment 10: Surviving the Titanic

Alyssa Garcia

2022-07-20

Exercise 1

i.

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

ii.

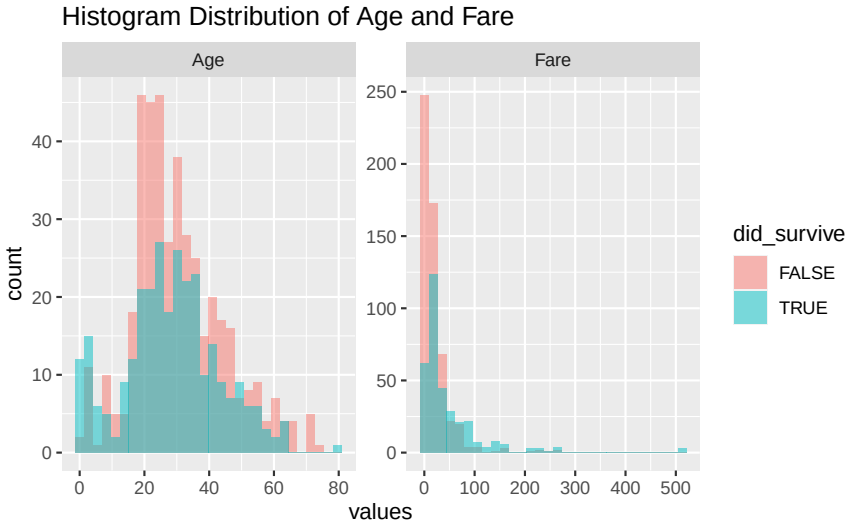
```
train_df <- train_df %>%
  mutate(
    did_survive = as.logical(Survived)
  )
```

Exercise 2

```
train_df %>%
  pivot_longer(
    cols = Age | Fare,
    names_to = "names",
    values_to = "values"
) %>%
  ggplot() +
  geom_histogram(aes(x = values, fill = did_survive),
                 position = "identity",
                 alpha = 0.5) +
  facet_wrap(~names, scales = "free") +
  labs(
    title = "Histogram Distribution of Age and Fare"
)
```

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

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



By examining the Fare histogram, we can see that as fare increased the number of survivors also increased. By examining our Age histogram, we can see as age increased the number of non-survivors increased. The differences between these variables should be helpful for predicting who survived.

Exercise 3

```
train_df %>%
  pivot_longer(
    cols = Pclass | Sex | Parch |SibSp,
    names_to = "names",
    values_to = "values") %>%
  ggplot() +
  geom_bar(aes(x = values, fill = did_survive),
           position = "identity",
           alpha = 0.5) +
  facet_wrap(~ names, scales = "free") +
  labs(
    title = "Bar Chart Distriution for Pclass, Sex, Parch, and SibSp"
  )
```

As the variable "Parch" increases, the number of survivors decreases. As the variable "Pclass", increases the non-survivors increase. This makes sense since the 3rd class was the lowest class and were unable to get onto the lifeboats. The variable "Sex" indicates that more females survived, while most non-survivors were male. This makes sense as women were called to board the lifeboats rather than males. As the variable "SipSp" increases the number of survivors and non-survivors decreases.

The categorical variables related to "did_survive" make sense given what I know about the Titanic's sinking. The variables that might be most useful for predicting survival is "Sex", "Pclass", and "Parch" since these variables include women and children, whom were the most likely to be saved.

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")
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## Please use `unite()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

The interaction between gender and passenger class seem to affect survival, as the mosaic plot reflects more women were saved in comparison to men.

Exercise 5

i.

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

count	missing	fraction_missing
891	177	0.1986532

ii.

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

iii. There is no missing data for the age_imputed variable within the train_imputed dataset.

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(
    count = n(),
    total_correct = sum(correct),
    accuracy = total_correct/count
  )
```

count	total_correct	accuracy
891	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
  )

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

```
## [1] 0.2065095 0.2060561
```

ii.

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

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

```
## [1] 0.1874299 0.1847229
```

iii.

From the three models created, the model with the most accurate validation error is model_3.