HW2

Roy Zhang

2022-10-06

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.1
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(magrittr)
```

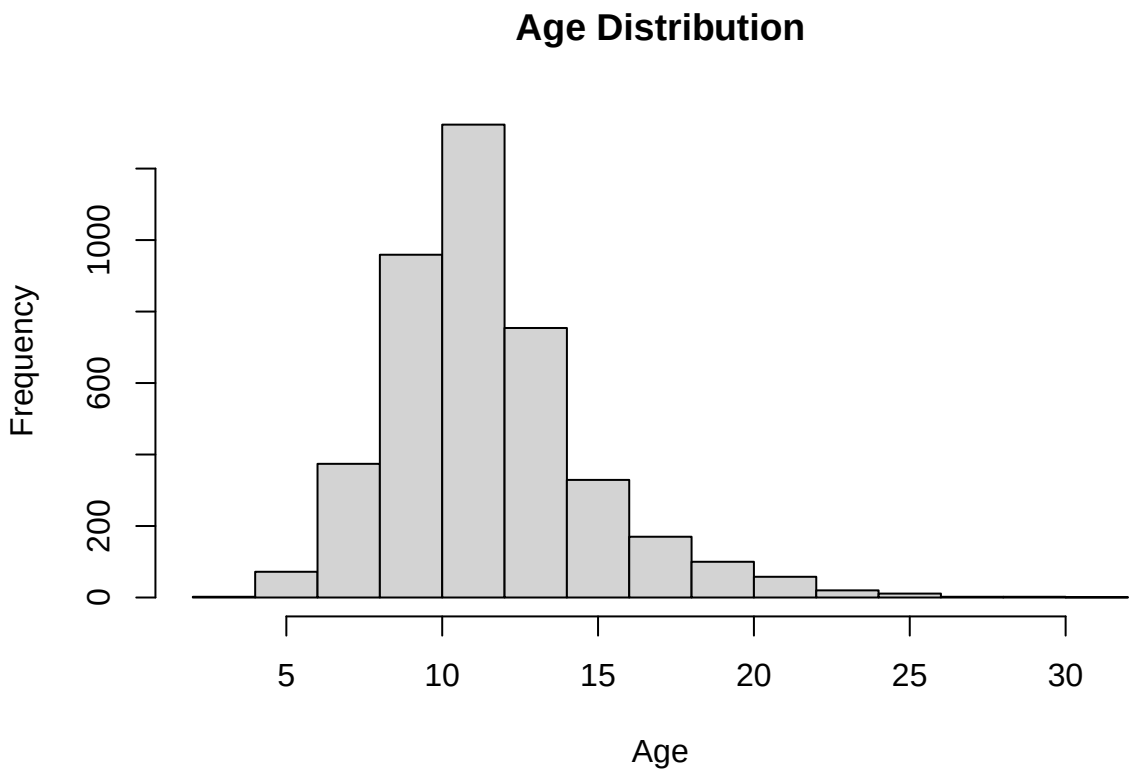
```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(yardstick)
abalone = read_csv("abalone.csv")
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 1.

```
abalone %<>% mutate(age = abalone$rings + 1.5)
hist(abalone$age, main = "Age Distribution", xlab = "Age")
```



The distribution of the data is skewed to the right. The range of the data is between 0 and 30. Most of the data points are between 8 and 14.

Question 2.

```
set.seed(1)
split = initial_split(abalone, strata = age)
training_set = training(split)
testing_set = testing(split)
```

Question 3.

```
abalone_recipe <- recipe(age ~ ., data = select(training_set, -rings)) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~starts_with("type"):shucked_weight) %>%
  step_interact(terms = ~longest_shell:diameter) %>%
```

```
  step_interact(terms = ~shucked_weight:shell_weight) %>%
  step_normalize(all_predictors())
```

Question 4.

```
lr_object = set_engine(linear_reg(), "lm")
```

Question 5.

```
abalone_workflow = workflow() %>%
  add_model(lr_object) %>%
  add_recipe(abalone_recipe)
```

Question 6.

```
abalone_fit = fit(abalone_workflow, training_set)
hypothetical_prediction = abalone_fit %>%
  predict(tibble(type = "F", longest_shell = 0.50, diameter = 0.10,
                 height = 0.30, whole_weight = 4, shucked_weight = 1,
                 viscera_weight = 2, shell_weight = 1))
```

The predicted age of the hypothetical abalone is 24.486188.

Question 7.

```
assessment_set = metric_set(rsq, rmse, mae)
training_prediction = predict(abalone_fit, training_set)
predicted_vs_actual = bind_cols(predicted = training_prediction$.pred,
                                actual = training_set$age)
assessment_outcome = assessment_set(predicted_vs_actual, truth = actual,
                                    estimate = predicted)
```

The R-squared value is 0.5553094, the RMSE is 2.1781885, and the MAE is 1.5568521.

According to the R-squared value, 55.5309356% of the variability of the abalone's age can be explained by this regression. model.