hw2

2022-10-06

Set up:

```
## ── Attaching packages ─────────────────────────────────────── tidymodels 1.0.0 ──
```

```
## ✓ broom        1.0.1      ✓ recipes      1.0.1
## ✓ dials        1.0.0      ✓ rsample      1.1.0
## ✓ dplyr        1.0.10     ✓ tibble       3.1.8
## ✓ ggplot2      3.3.6      ✓ tidyr        1.2.1
## ✓ infer        1.0.3      ✓ tune         1.0.0
## ✓ modeldata    1.0.1      ✓ workflows    1.1.0
## ✓ parsnip      1.0.1      ✓ workflowsets 1.0.0
## ✓ purrr        0.3.4      ✓ yardstick    1.1.0
```

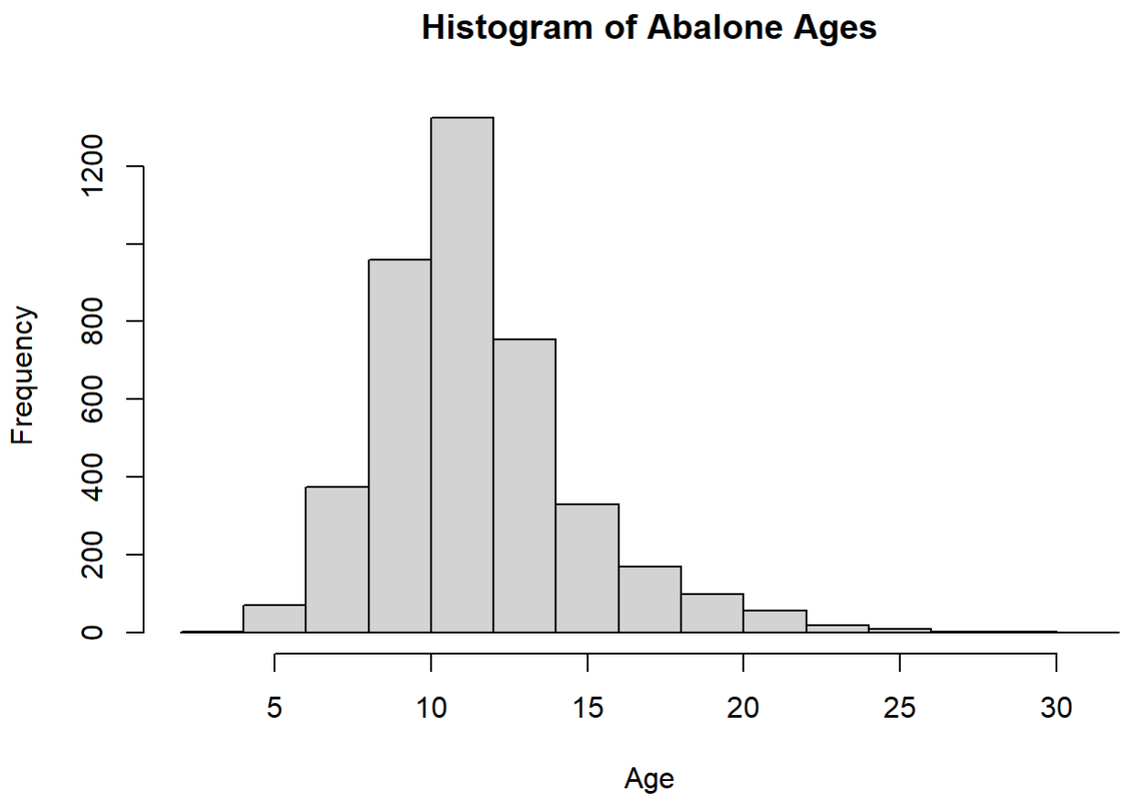
```
## ── Conflicts ────────────────────────────────────────── tidymodels_conflicts() ──
## ✖ purrr::discard() masks scales::discard()
## ✖ dplyr::filter()  masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## ✖ recipes::step()  masks stats::step()
## • Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## ✓ stringr 1.4.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ readr::col_factor() masks scales::col_factor()
## ✖ purrr::discard()    masks scales::discard()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ stringr::fixed()    masks recipes::fixed()
## ✖ dplyr::lag()        masks stats::lag()
## ✖ readr::spec()       masks yardstick::spec()
## 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.
```

```
## # A tibble: 4,177 × 9
##    type  longest_shell diameter height whole_wei…¹ shuck…² visce…³ shell…⁴ rings
##    <chr>         <dbl>    <dbl>  <dbl>       <dbl>   <dbl>   <dbl>   <dbl> <dbl>
##  1 M             0.455    0.365  0.095       0.514  0.224   0.101    0.15     15
##  2 M             0.35     0.265  0.09        0.226  0.0995  0.0485   0.07      7
##  3 F             0.53     0.42   0.135       0.677  0.256   0.142    0.21      9
##  4 M             0.44     0.365  0.125       0.516  0.216   0.114    0.155    10
##  5 I             0.33     0.255  0.08        0.205  0.0895  0.0395   0.055     7
##  6 I             0.425    0.3    0.095       0.352  0.141   0.0775   0.12      8
##  7 F             0.53     0.415  0.15        0.778  0.237   0.142    0.33     20
##  8 F             0.545    0.425  0.125       0.768  0.294   0.150    0.26     16
##  9 M             0.475    0.37   0.125       0.509  0.216   0.112    0.165     9
## 10 F             0.55     0.44   0.15        0.894  0.314   0.151    0.32     19
## # … with 4,167 more rows, and abbreviated variable names ¹whole_weight,
## #   ²shucked_weight, ³viscera_weight, ⁴shell_weight
```

Question 1:

```
ages <- abalone %>%
  mutate(age = rings + 1.5)
hist(ages$age, main = "Histogram of Abalone Ages",
     xlab = "Age")
```




```
ages
```

```
## # A tibble: 4,177 × 10
##    type  longest_sh…¹ diame…² height whole…³ shuck…⁴ visce…⁵ shell…⁶ rings   age
##    <chr>        <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>
##  1 M            0.455   0.365  0.095   0.514  0.224   0.101    0.15     15  16.5
##  2 M            0.35    0.265  0.09    0.226  0.0995  0.0485   0.07      7   8.5
##  3 F            0.53    0.42   0.135   0.677  0.256   0.142    0.21      9  10.5
##  4 M            0.44    0.365  0.125   0.516  0.216   0.114    0.155    10  11.5
##  5 I            0.33    0.255  0.08    0.205  0.0895  0.0395   0.055     7   8.5
##  6 I            0.425   0.3    0.095   0.352  0.141   0.0775   0.12      8   9.5
##  7 F            0.53    0.415  0.15    0.778  0.237   0.142    0.33     20  21.5
##  8 F            0.545   0.425  0.125   0.768  0.294   0.150    0.26     16  17.5
##  9 M            0.475   0.37   0.125   0.509  0.216   0.112    0.165     9  10.5
## 10 F            0.55    0.44   0.15    0.894  0.314   0.151    0.32     19  20.5
## # … with 4,167 more rows, and abbreviated variable names ¹longest_shell,
## #   ²diameter, ³whole_weight, ⁴shucked_weight, ⁵viscera_weight, ⁶shell_weight
```

In this histogram, we see that the distribution of age is mostly centered around 8 to 14 years old. There are very few abalone ages that are older than 15 and there are extremely older than 25 years old. The most common ages are between 10 and 12 years old as an unimodal, right-skewed distribution.

Question 2

```
set.seed(1500)

abalone_split <- initial_split(ages, prop = 0.80,
                               strata = age)

train = training(abalone_split)
test = testing(abalone_split)
```

Question 3

```
abalone_recipe <- recipe(age ~ ., data = train) %>%
  step_rm(rings) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight +
                  longest_shell:diameter +
                  shucked_weight:shell_weight) %>%
  step_scale(all_numeric_predictors()) %>%
  step_center(all_numeric_predictors())
abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          9
##
## Operations:
##
## Variables removed rings
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight + longest_shell...
## Scaling for all_numeric_predictors()
## Centering for all_numeric_predictors()
```

I shouldn't include rings to predict age since rings are directly related to age as we would just add 1.5 years to rings and we would automatically know the age of the abalone.

Question 4

```
lm_model <- linear_reg() %>%
  set_engine("lm")
lm_model
```

```
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

Question 5

```
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)

lm_fit <- fit(lm_wflow, train)
lm_fit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 14 × 5
##    term                            estimate std.error  statistic  p.value
##    <chr>                              <dbl>     <dbl>      <dbl>    <dbl>
##  1 (Intercept)                     11.5          0.0373 307.       0
##  2 longest_shell                    0.578        0.283    2.04     4.12e- 2
##  3 diameter                         2.20         0.307    7.17     9.43e-13
##  4 height                           0.207        0.0686   3.02     2.54e- 3
##  5 whole_weight                     5.02         0.388   12.9      2.51e-37
##  6 shucked_weight                  -4.44         0.253  -17.5      7.22e-66
##  7 viscera_weight                  -0.914        0.155   -5.89     4.34e- 9
##  8 shell_weight                     1.52         0.212    7.18     8.45e-13
##  9 type_I                          -0.951        0.117   -8.16     4.84e-16
## 10 type_M                          -0.283        0.105   -2.70     6.97e- 3
## 11 type_I_x_shucked_weight          0.531        0.0882   6.02     1.90e- 9
## 12 type_M_x_shucked_weight          0.294        0.112    2.64     8.41e- 3
## 13 longest_shell_x_diameter        -2.92         0.398   -7.34     2.61e-13
## 14 shucked_weight_x_shell_weight    0.0000418    0.204    0.000205 1.00e+ 0
```

```
lm_fit
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 5 Recipe Steps
##
## • step_rm()
## • step_dummy()
## • step_interact()
## • step_scale()
## • step_center()
##
## -- Model -----------------------------------------------------------------------
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##                   (Intercept)                  longest_shell
##                     1.145e+01                      5.779e-01
##                      diameter                         height
##                     2.204e+00                      2.071e-01
##                   whole_weight                 shucked_weight
##                     5.020e+00                     -4.442e+00
##                  viscera_weight                   shell_weight
##                     -9.142e-01                      1.524e+00
##                         type_I                         type_M
##                     -9.512e-01                     -2.832e-01
##       type_I_x_shucked_weight        type_M_x_shucked_weight
##                      5.310e-01                      2.942e-01
##      longest_shell_x_diameter  shucked_weight_x_shell_weight
##                     -2.920e+00                      4.176e-05
```

Question 6

```
gather_data <- data.frame(type = "F", longest_shell = 0.50, diameter = 0.10,
                          height = 0.30, whole_weight = 4, shucked_weight = 1,
                          viscera_weight = 2, shell_weight = 1, rings = 0)

predict(lm_fit, new_data = gather_data)
```

```
## # A tibble: 1 × 1
##   .pred
##   <dbl>
## 1  24.0
```

Based on the information given, the abalone age is predicted to be around 24 years old.

Question 7

```
multi_metric <- metric_set(rsq, rmse, mae)

train_res <- predict(lm_fit, new_data = train %>%
                       select(-age))

train_res <- bind_cols(train_res, train %>%
                         select(age))

multi_metric(train_res, truth = age,
               estimate = .pred)
```

```
## # A tibble: 3 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.551
## 2 rmse    standard       2.15
## 3 mae     standard       1.55
```

Based on the results, we see that the R squared value is around 0.551. This means that there is a correlation (although not an extremely strong one, but still a strong one) between our predicted values and response values.