hw2

2022-10-12

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
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         1.0.0
## 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()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(ggthemes)
```

```r
#Reading in data
setwd("~/Downloads/homework-2 4")
abalone <- read_csv(file = "data/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.
```

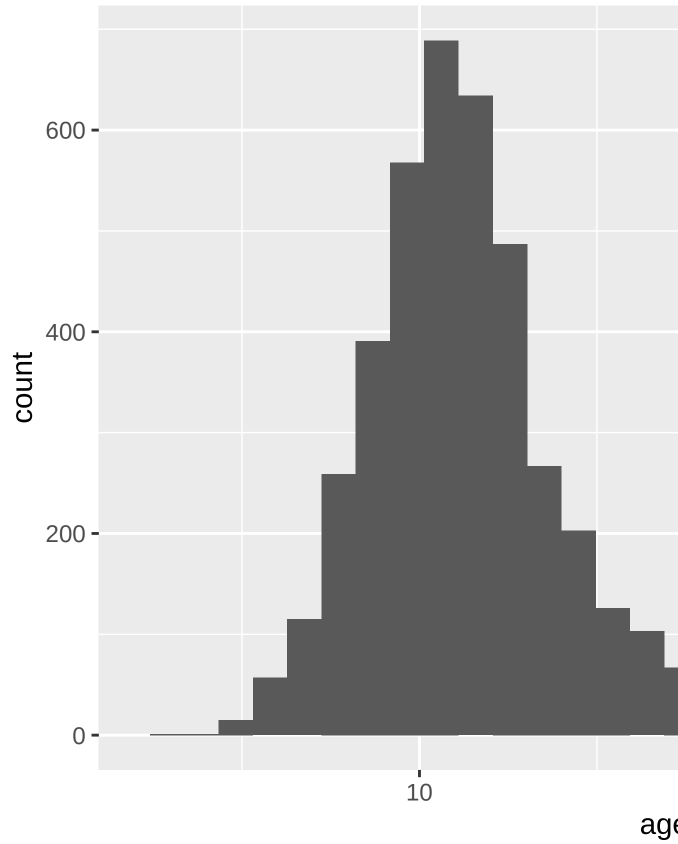
```
head(abalone)
```

```
## # A tibble: 6 x 9
##   type  longest_shell diameter height whole_weight shuck~1 visce~2 shell~3 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
## # ... with abbreviated variable names 1: shucked_weight, 2: viscera_weight,
## #   3: shell_weight
```

Question 1

```
#Calculating the number of rings then adding the age column to the data set
abalone$age <- abalone$rings + 1.5
```

```
abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30)
```



Creating a histogram to describe the distribution of Age
It seens that the histogram is slightly positively skewed. This means that most abalones are less than 20 years old.

Question 2

```
set.seed(0426)

#Splitting data
abalone_split <- initial_split(abalone, prop = 0.80,
                               strata = age)

#Testing and training data
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
#Creating a recipe for abalone
simple_abalone_recipe <-
  recipe(age ~ ., data = abalone_train)
```

We shouldn't use rings because age actually depends on the value of rings

```r
#Dummy coding categorical predictors
abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
   step_dummy(all_nominal_predictors()) %>%
  step_rm(contains('rings')) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight + #Interacation terms
                  longest_shell:diameter +
                  shucked_weight:shell_weight) %>%
  step_scale(all_predictors()) %>%
  step_center(all_predictors())

names(abalone)
```

```
## [1] "type"           "longest_shell"  "diameter"       "height"
## [5] "whole_weight"   "shucked_weight" "viscera_weight" "shell_weight"
## [9] "rings"          "age"
```

Question 4

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

Question 5

```r
#Setting up empty workflow
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
lm_wflow
```

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 5 Recipe Steps
##
## * step_dummy()
## * step_rm()
## * step_interact()
## * step_scale()
## * step_center()
##
## -- Model -----------------------------------------------------------------------
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

```r
#Adding the model previously created
lm_fit <- fit(lm_wflow, abalone_train)
lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy()
```

```
## # A tibble: 14 x 5
##    term                           estimate std.error statistic  p.value
##    <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                      11.4      0.0372   307.    0
##  2 longest_shell                     0.425    0.283      1.50  1.34e- 1
##  3 diameter                          2.03     0.312      6.50  9.36e-11
##  4 height                            0.489    0.0965     5.07  4.13e- 7
##  5 whole_weight                      4.48     0.386     11.6   1.37e-30
##  6 shucked_weight                   -4.13     0.247    -16.8   1.50e-60
##  7 viscera_weight                   -0.782    0.157     -5.00  6.16e- 7
##  8 shell_weight                      1.56     0.213      7.33  2.91e-13
##  9 type_I                           -0.886    0.116     -7.67  2.29e-14
## 10 type_M                           -0.236    0.103     -2.30  2.13e- 2
## 11 type_I_x_shucked_weight           0.438    0.0872     5.02  5.36e- 7
## 12 type_M_x_shucked_weight           0.279    0.108      2.59  9.70e- 3
## 13 longest_shell_x_diameter         -2.64     0.402     -6.57  5.84e-11
## 14 shucked_weight_x_shell_weight    -0.118    0.201     -0.587 5.57e- 1
```

Question 6

```r
#Predicting age of the hypothetical female abalone
predict(lm_fit, data.frame(longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, shuc
```

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

Our predicted age for this female abalone is about 24 years.

```r
#Generates predicted values for each observation in abalone_train
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
abalone_train_res %>%
  head()
```

Question 7

```
## # A tibble: 6 x 1
##   .pred
```

```
##   <dbl>
## 1  9.49
## 2  9.30
## 3 10.0
## 4 10.9
## 5  5.92
## 6  8.60
```

```r
#Attaching columns to actual age observations
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
abalone_train_res %>% head
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.49   8.5
## 2  9.30   9.5
## 3 10.0    9.5
## 4 10.9    9.5
## 5  5.92   5.5
## 6  8.60   8.5
```

```r
#Creating a metric set
library(yardstick)
abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.14
## 2 rsq     standard       0.557
## 3 mae     standard       1.54
```

Our RMSE = 2.144, MAE = 1.536, and R^2 value is 0.557108. This R^2 value indicates that about 56% of the variability in the response can be explained by this linear regression model. This can mean that our model is not the worst model for the abalone data, but it would also not be the best model.