HW2-Camilla Zhai

2022-10-12

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## 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
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

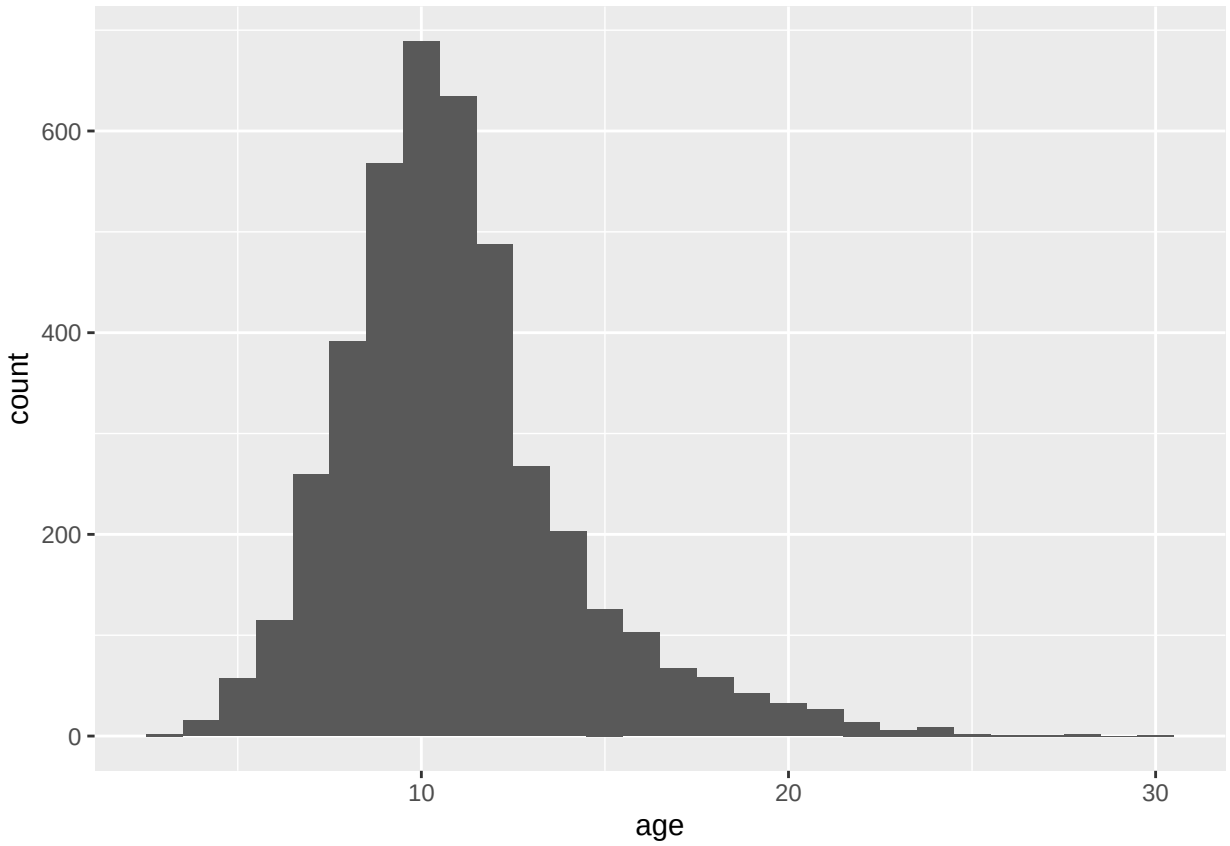
```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      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()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```
abalone_info <- read_csv(file='/Users/weiqizhai/Desktop/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:

```
#calculate and add variable age to the dataset
new_abalone<-abalone_info %>%mutate(age=rings+1.5) %>% select(everything(), age)
#Creating a histogram of age
age_hist <- new_abalone %>% ggplot(aes(x=age))+geom_histogram(binwidth = 1)
age_hist
```



From the histogram above, age is basically normally distributed with a slight right skew. The majority of abalone are between the ages of 8 and 13 years.

Question 2:

```
#data splitting
set.seed(1018)
abalone_split<- initial_split(new_abalone, prop=0.80,strata=age)
A_train <- training(abalone_split)
A_test <- testing(abalone_split)
```

Question 3:

```
A_recipe <- recipe(age~type+longest_shell+diameter+height+whole_weight+shucked_weight
+viscera_weight+shell_weight,data=A_train) %>% step_dummy(all_nominal_predictors())

A_recipe %>% step_interact(~type:shucked_weight+longest_shell:diameter+shucked_weight:shell_weight) %>%


## Recipe
##
```

```
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with type:shucked_weight + longest_shell:diameter + shu...
## Centering and scaling for all_numeric_predictors()
```

Because our goal is to see if abalone age can be accurately predicted using information other than rings. As a result, we should not use rings as a predictor variable; instead, we should build a model using other variables from the dataset to determine their relationship with the response variable (age); using rings as a predictor variable fails our goal.

Question 4:

```
m1 <- linear_reg() %>% set_engine("lm") #specify "lm" model engine
```

Question 5:

```
wflow1 <- workflow() %>% add_recipe(A_recipe) %>% add_model(m1) # setting up a workflow
fitted_m <- fit(wflow1, A_train) #fit the linear model to the training set
```

Question 6:

```
test_case <- tibble(type='F',longest_shell=0.50, diameter=0.1,height=0.3,
                    whole_weight=4, shucked_weight=1,viscera_weight=2,shell_weight=1)
predict(fitted_m, test_case)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  13.3
```

Question 7:

```
abalone_metrics <- metric_set(rsq,rmse,mae)
abalone_train  <- predict(fitted_m,new_data=A_train %>% select(-age))
abalone_train <- bind_cols(abalone_train, A_train %>% select(age))
abalone_metrics(abalone_train, truth=age, estimate=.pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.533
## 2 rmse    standard       2.20
## 3 mae     standard       1.59
```

The R squared value is 0.5327499, which is low. It reveals that about 53% of the variability observed in the target variable (age) is explained by the regression model