Homework2

```
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.0.5
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.2.0 --
```

```
## v broom        0.7.12     v recipes      0.2.0
## v dials        0.1.0      v rsample      0.1.1
## v dplyr        1.0.8      v tibble       3.1.6
## v ggplot2      3.3.5      v tidyr        1.2.0
## v infer        1.0.0      v tune         0.2.0
## v modeldata    0.1.1      v workflows    0.2.6
## v parsnip      0.2.1      v workflowsets 0.2.1
## v purrr        0.3.4      v yardstick    0.0.9
```

```
## Warning: package 'broom' was built under R version 4.0.5
```

```
## Warning: package 'dials' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'parsnip' was built under R version 4.0.5
```

```
## Warning: package 'recipes' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'tune' was built under R version 4.0.5
```

```
## Warning: package 'workflows' was built under R version 4.0.5
```

```
## Warning: package 'workflowsets' was built under R version 4.0.5
```

```
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```r
library(tidyverse)
```

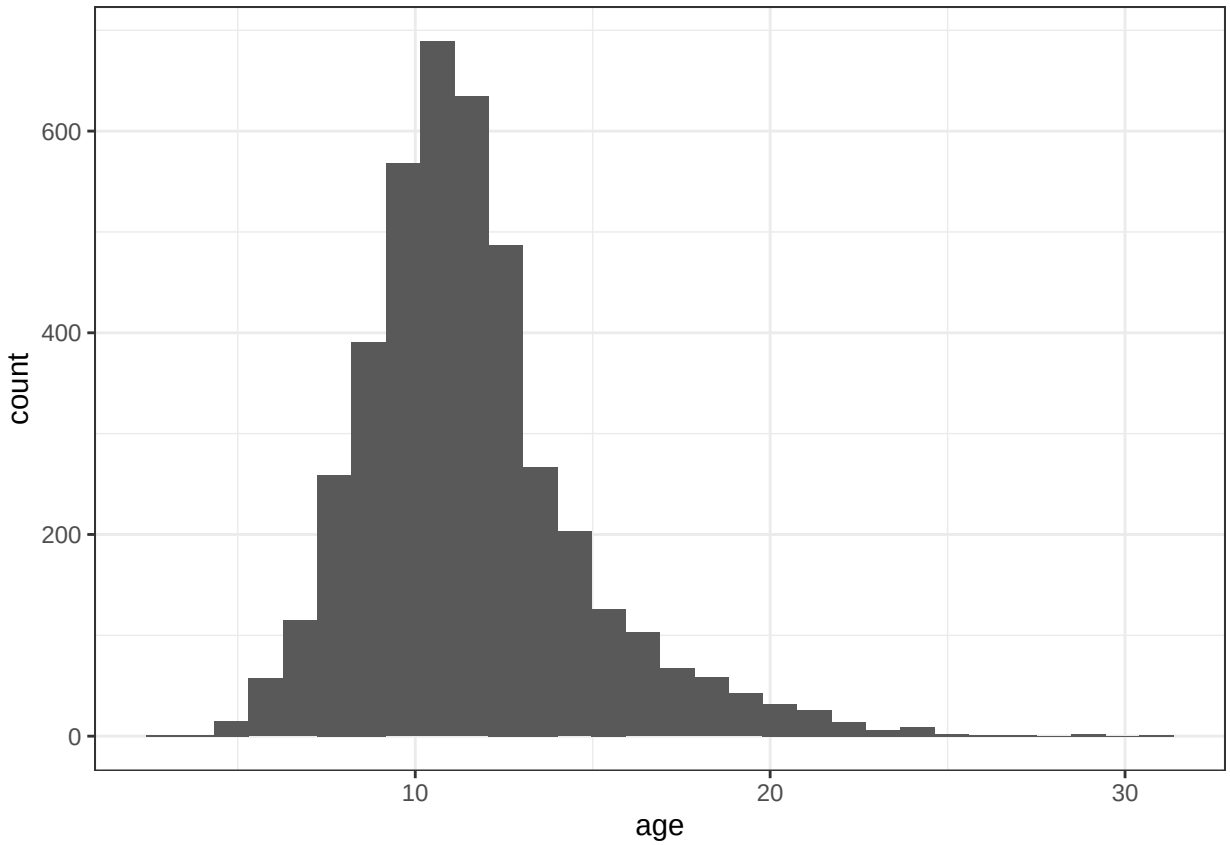
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v readr   2.1.2     v forcats 0.5.1
## v stringr 1.4.0

## Warning: package 'readr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x readr::spec()       masks yardstick::spec()
```

```r
abalone <- read.csv("abalone.csv")
```

Question 1

```r
# add age variable to the dataset
abalone <- abalone %>%
  mutate(age = rings + 1.5)
```

```r
# make a histogram to assess the distribution of age
abalone %>%
  ggplot(aes(x=age)) +
  geom_histogram(bins=30)+
  theme_bw()
```



The distribution of age seems unsymmetrical and is a bit of a right skewed. Also, the most count age is around 11.

Question 2

```
set.seed(1979)
abalone_split <- initial_split(abalone, prop = 0.8,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
simple_abalone_recipe <-
  recipe(age ~ ., data = abalone_train %>% select(-rings)) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(~ shucked_weight:starts_with("type") + longest_shell:diameter + shucked_weight:shell_wei
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
```

We should not use rings to predict age because we use rings to caculate age. When we use rings to predict age, our R^2 will becomes 1 even we do not include other predictors in our model. Therefore, our model is not meanful when we include rings.

Question 4

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

Question 5

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

Question 6

```
lm_fit <- fit(lm_wflow,abalone_train)
```

```
lm_fit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     11.4       0.0370  309.     0
##  2 longest_shell                    0.743     0.282     2.63   8.49e- 3
##  3 diameter                         2.03      0.309     6.58   5.49e-11
##  4 height                           0.202     0.0683    2.96   3.12e- 3
##  5 whole_weight                     5.64      0.399    14.1    4.89e-44
##  6 shucked_weight                  -4.67      0.258   -18.1    6.37e-70
##  7 viscera_weight                  -1.19      0.158    -7.54   5.82e-14
##  8 shell_weight                     1.41      0.215     6.59   5.12e-11
##  9 type_I                          -0.897     0.116    -7.76   1.10e-14
## 10 type_M                          -0.177     0.103    -1.73   8.45e- 2
## 11 shucked_weight_x_type_I          0.464     0.0867    5.36   8.96e- 8
## 12 shucked_weight_x_type_M          0.214     0.107     1.99   4.61e- 2
## 13 longest_shell_x_diameter        -2.84      0.404    -7.03   2.58e-12
## 14 shucked_weight_x_shell_weight   -0.0197    0.201    -0.0980 9.22e- 1
```

```
abalone2 <- data.frame(type = 'F',longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4
predict_age <- predict(lm_fit, abalone2)
predict_age
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.3
```

Question 7

```
library(yardstick)
abalone_metrics <- metric_set(rmse,rsq,mae)
```

```
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-rings,-age))
abalone_train_res %>% head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1  9.56
## 2  8.09
## 3  9.82
## 4 10.1
## 5 10.9
## 6  6.24
```

```
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  9.56   8.5
## 2  8.09   8.5
## 3  9.82   8.5
## 4 10.1    9.5
## 5 10.9    9.5
## 6  6.24   6.5
```

```
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       0.563
## 3 mae     standard       1.53
```

R^2 is 0.5633, which means 56.33% of the variability in Y(age) that can be explained by using predictor variables