HW2

Question 1

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

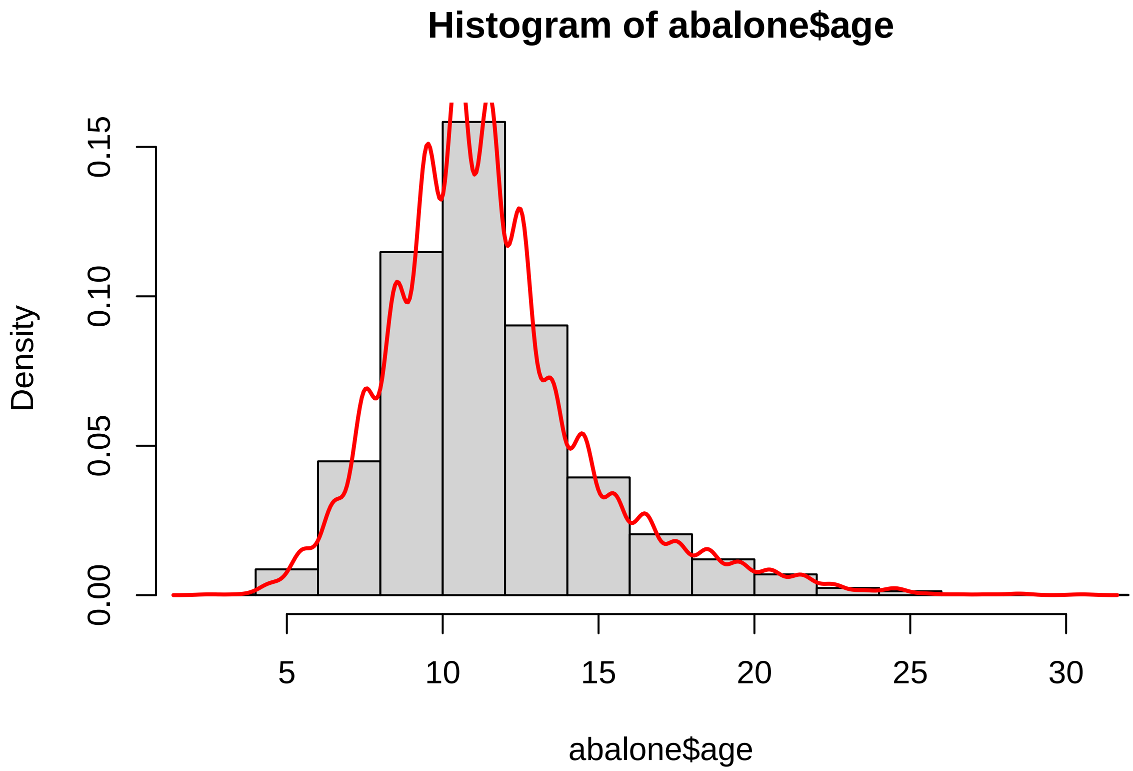
```
## -- Attaching packages -------------------------------------- tidymodels 0.2.0 --

## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0

## -- 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()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
abalone <- read.csv("~/Downloads/homework-2/data/abalone.csv")
View(abalone)
abalone <- abalone %>%
  mutate(age = rings + 1.5)

View(abalone)
hist(abalone$age, freq = FALSE)
lines(density(abalone$age), lwd = 2, col = 'red')
```


Histogram of abalone$age

The distribution is relatively right skewed, which implies that the mean of age is greater than the median of age.

Question 2

```
set.seed(4177)

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

Question 3

```
rec <- recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight + viscera_

abalone_recipe <- rec %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(~type:shucked_weight) %>%
  step_interact(~longest_shell:diameter) %>%
  step_interact(~shucked_weight:shell_weight) %>%
  step_normalize(all_predictors())
```

Rings should not be used to predict age because it was already used to calculate age. Including the variable will only mess with the results as we would be adding something that has already been accounted for. (It has been removed in the steps for Question 1)

Question 4

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

Question 5

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

Question 6

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

```
## Warning: Interaction specification failed for: ~type:shucked_weight. No
## interactions will be created.
```

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

```
## # A tibble: 12 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                    11.4       0.0374   306.    0
##  2 longest_shell                   0.819     0.274      2.99  2.81e- 3
##  3 diameter                        2.53      0.301      8.39  7.36e-17
##  4 height                          0.261     0.0696     3.75  1.80e- 4
##  5 whole_weight                    5.29      0.414     12.8   1.41e-36
##  6 shucked_weight                 -4.05      0.244    -16.6   1.03e-59
##  7 viscera_weight                 -1.14      0.160     -7.11  1.44e-12
##  8 shell_weight                    1.62      0.212      7.66  2.40e-14
##  9 type_I                         -0.348     0.0540    -6.44  1.33e-10
## 10 type_M                         -0.0228    0.0444    -0.513 6.08e- 1
## 11 longest_shell_x_diameter       -3.49      0.380     -9.17  8.04e-20
## 12 shucked_weight_x_shell_weight  -0.213     0.198     -1.08  2.81e- 1
```

```
female_ab <- data.frame(type = "F",longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight =

predict(lm_fit, new_data = female_ab)
```

```
## # A tibble: 1 x 1
##    .pred
##    <dbl>
## 1  22.2
```

Question 7

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

```
## # A tibble: 6 x 1
##    .pred
##    <dbl>
## 1  9.27
## 2  8.32
## 3 10.1
## 4  9.99
## 5  6.39
## 6  5.79
```

```
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  9.27   8.5
## 2  8.32   8.5
## 3 10.1    8.5
## 4  9.99   9.5
## 5  6.39   6.5
## 6  5.79   6.5
```

```
rmse(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##    .metric .estimator .estimate
##    <chr>   <chr>          <dbl>
## 1 rmse     standard        2.16
```

```r
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       2.16
## 2 rsq     standard       0.555
## 3 mae     standard       1.56
```