PSTAT 131 Homework 2

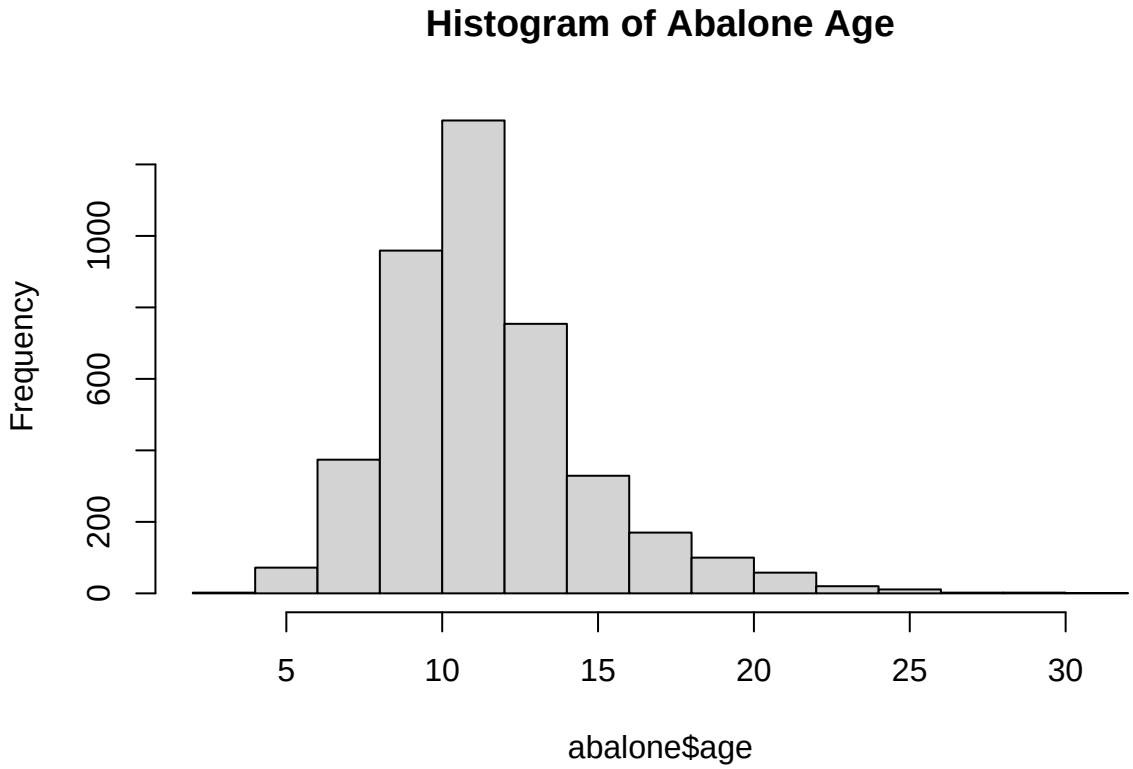
Logan Greenough

10/4/2022

Question 1

```
abalone <- abalone %>% mutate(age = rings + 1.5)

hist(abalone$age, main = "Histogram of Abalone Age")
```



Based on a histogram for abalone age, we can see that there is a skew to the right, with the majority of the observations being in the 10 to 15 year range.

Question 2

```r
set.seed(805)

abalone_split <- initial_split(abalone, prop = 0.7, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```r
#Step 1
simple_abalone_recipe <- recipe(age ~., data = abalone_train) %>% step_rm(rings) %>% step_dummy(type)

#Step 2
simple_abalone_recipe <- simple_abalone_recipe %>% step_interact( ~ type_M:shucked_weight + type_I:shuck

#Step 3
simple_abalone_recipe <- simple_abalone_recipe %>% step_center(all_predictors())

#Step 4
simple_abalone_recipe <- simple_abalone_recipe %>% step_scale(all_predictors())
```

We should not use rings to predict age because we are trying to create a model that will be able to predict the age of an abalone without the data for the number of rings. This is due to the fact that in order to know the number of rings, the abalone has to be cut open and the rings have to be counted under a microscope.

Question 4

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

Question 5

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

Question 6

```r
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      0.0393   291.    0
##  2 longest_shell                   0.627    0.299      2.10  3.60e- 2
##  3 diameter                        2.20     0.324      6.80  1.25e-11
##  4 height                          0.172    0.0711     2.42  1.56e- 2
##  5 whole_weight                    4.75     0.407     11.7   9.88e-31
##  6 shucked_weight                 -4.00     0.259    -15.4   9.78e-52
##  7 viscera_weight                 -0.858    0.164     -5.22  1.93e- 7
##  8 shell_weight                    1.77     0.221      8.01  1.63e-15
##  9 type_I                         -0.982    0.120     -8.17  4.68e-16
## 10 type_M                         -0.225    0.109     -2.07  3.89e- 2
## 11 type_M_x_shucked_weight         0.286    0.116      2.47  1.35e- 2
## 12 shucked_weight_x_type_I         0.554    0.0913     6.07  1.44e- 9
## 13 longest_shell_x_diameter       -3.20     0.425     -7.53  6.70e-14
## 14 shucked_weight_x_shell_weight  -0.215    0.216     -0.996 3.19e- 1
```

```
test_data = tibble(type = "F", longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, s

test_results <- predict(lm_fit, new_data = test_data)

test_results
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  24.5
```

Question 7

```
library(yardstick)

abalone_training_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))

abalone_training_res <- bind_cols(abalone_training_res, abalone_train %>% select(age))

abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_training_res, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.12
## 2 rsq     standard       0.560
## 3 mae     standard       1.53
```

Our model returns a RMSE of 2.117829, a MAE of 1.529558, and an R^2 of 0.560166. This can be interpreted as the regression model being able to explain 56% of the observed data and is about middle of the road when it comes to how well the model fits the data set.