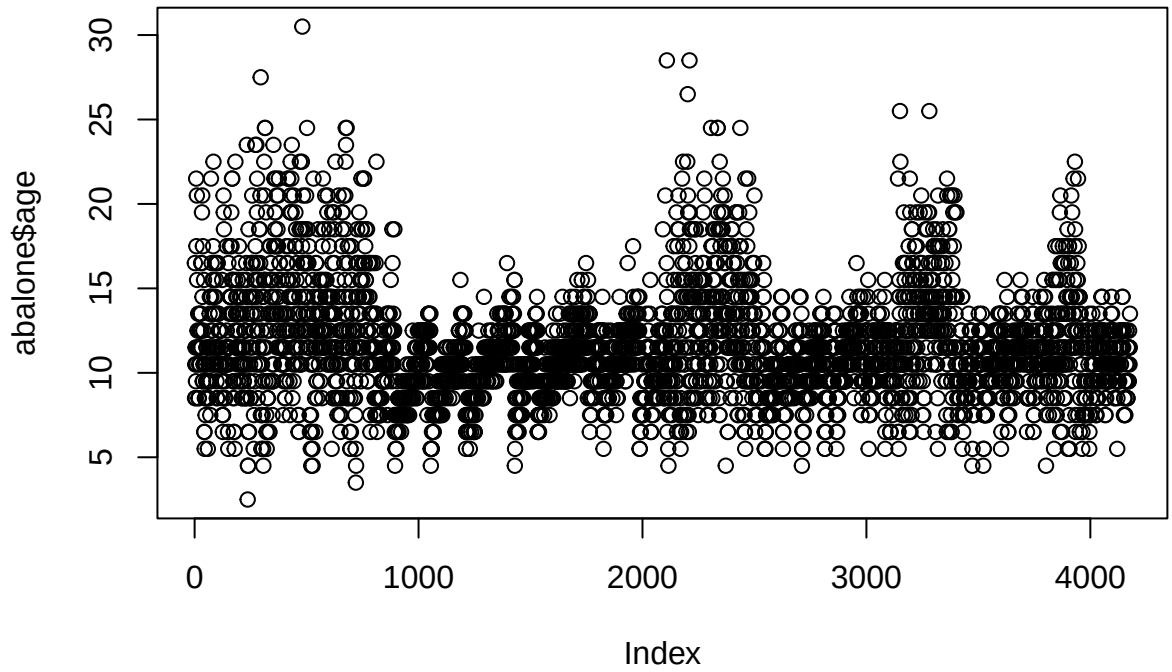
HW2

2022-10-04

1. Predict abalone age

```
abalone$age <- abalone$rings + 1.5
plot(abalone$age)
```



Assess and describe the distribution of age. Majority of their ages seem to be between 8 to 13.

2.

```
set.seed(1738)

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

3. Create a recipe predicting the outcome variable, age, with all other predictor variables. Explain why you shouldn't use rings to predict age.

```
abalone_recipe <- recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight
  step_dummy(all_nominal_predictors()) %>%
  step_interact(~starts_with("type"):shucked_weight) %>%
  step_interact(~longest_shell:diameter) %>%
  step_interact(~shucked_weight:shell_weight) %>%
  step_center(longest_shell, diameter, height, whole_weight, shucked_weight, viscera_weight, shell_weigh
  step_scale(longest_shell, diameter, height, whole_weight, shucked_weight, viscera_weight, shell_weight

abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight
## Interactions with longest_shell:diameter
## Interactions with shucked_weight:shell_weight
## Centering for longest_shell, diameter, height, whole_weight, ...
## Scaling for longest_shell, diameter, height, whole_weight, ...
```

Rings are directly correlated with age, so we can't use rings to predict age.

4. Create and store a linear regression object using the "lm" engine.

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

5. Workflow set up

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

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

6. Use your fit() object to predict the age of a hypothetical female abalone with longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, shucked_weight = 1, viscera_weight = 2, shell_weight = 1.

```
lm_fit <- fit(lm_wkflow, 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)                     18.3        1.02     17.8   1.06e-67
##  2 longest_shell                    0.0463     0.318     0.146 8.84e- 1
##  3 diameter                         2.35       0.353     6.67  3.00e-11
##  4 height                           0.578      0.103     5.59  2.42e- 8
##  5 whole_weight                     5.93       0.450    13.2   1.20e-38
##  6 shucked_weight                  -4.60       0.281   -16.4   8.09e-58
##  7 viscera_weight                  -1.18       0.173    -6.86  8.63e-12
##  8 shell_weight                     1.01       0.235     4.27  1.99e- 5
##  9 type_I                          -1.66       0.267    -6.23  5.41e-10
## 10 type_M                          -0.247      0.231    -1.07  2.86e- 1
## 11 type_I_x_shucked_weight          3.94       0.818     4.82  1.53e- 6
## 12 type_M_x_shucked_weight          0.856      0.472     1.81  6.99e- 2
## 13 longest_shell_x_diameter       -28.6        4.73     -6.05  1.63e- 9
## 14 shucked_weight_x_shell_weight   -1.03       1.91     -0.538 5.90e- 1
```

```
new_obs <- tibble(
  longest_shell = c(0.5), diameter = c(0.10), height = c(0.30), whole_weight = c(4), shucked_weight = c
  )
new_pred <- new_obs %>%
  bind_cols(lm_fit %>%
              predict(new_data = new_obs))
new_pred
```

```
## # A tibble: 1 x 9
##   longest_shell diameter height whole_weight shuck~1 visce~2 shell~3 type  .pred
##           <dbl>    <dbl>  <dbl>        <dbl>   <dbl>   <dbl>   <dbl> <chr> <dbl>
## 1           0.5      0.1    0.3            4       1       2       1 F      21.9
## # ... with abbreviated variable names 1: shucked_weight, 2: viscera_weight,
## #   3: shell_weight
```

Predicted Age: 21.94

7.

```
library(yardstick)
new_metric <- metric_set(rsq, rmse, mae)
preds <- abalone_train %>%
  select(age) %>%
  bind_cols(lm_fit %>%
              predict(abalone_train))
preds %>%
  new_metric(age, .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.554
## 2 rmse    standard       2.16
## 3 mae     standard       1.55
```

R^2: 0.55 Root Mean Square Error: 2.16 Mean Absolute Error: 1.55

R^2 depicts how well the predicted values fit the actual values. Since ours is about 0.55, that means about 55% of the observed variabilty is explained by our model.