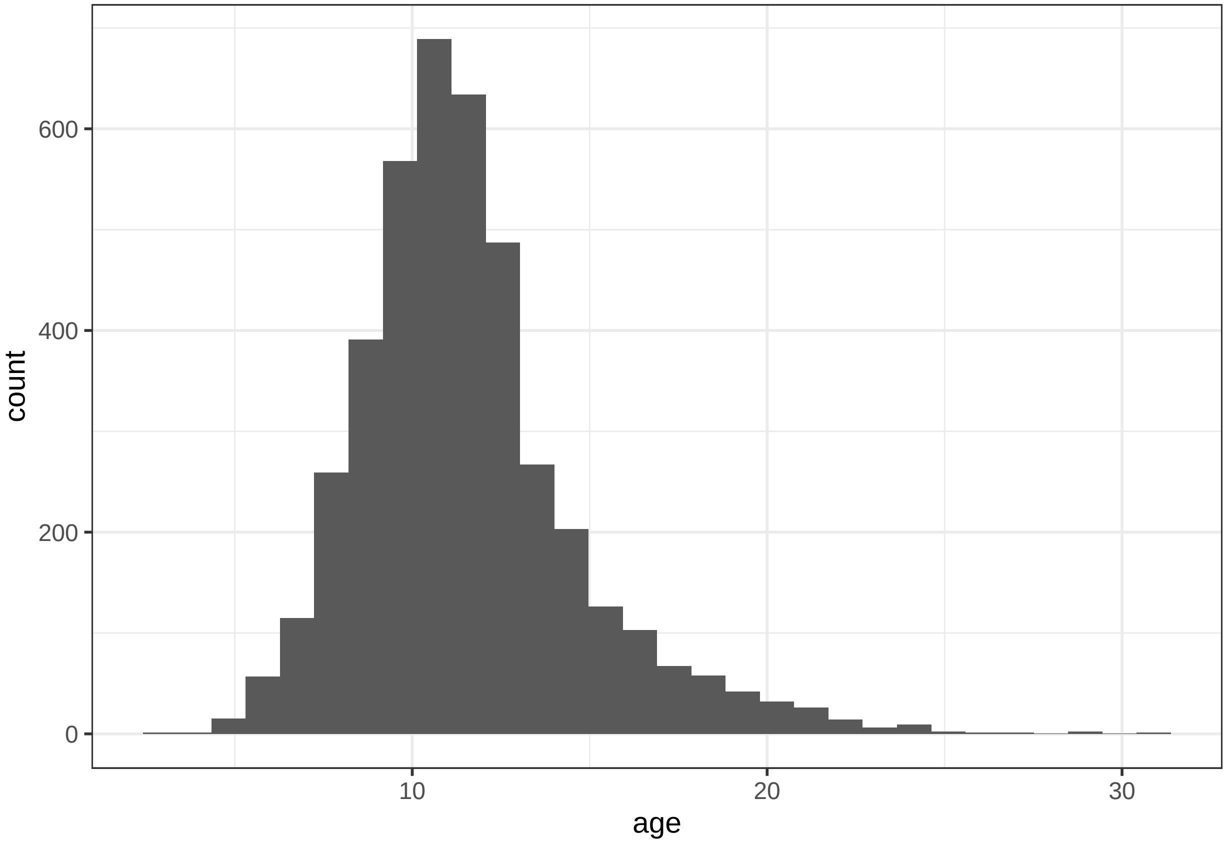
hw2

Jiaxuan Sun

2022-10-12

```
##Question1
abalone_added <- abalone %>%
  mutate(age = rings + 1.5)

abalone_added %>%
  ggplot(aes(x = age)) +
  geom_histogram() +
  theme_bw()
```



We can tell that the distribution of 'age' is approximately a normal distribution with μ being about 10 years old.

```
##Question2
set.seed(95)
```

```
abalone_split <- initial_split(abalone_added, prop = 0.80,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

```
##Question3
abalone_train_deleted <- abalone_train %>%
  select(-rings)

abalone_recipe <- recipe(age ~ ., data = abalone_train_deleted) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type_") : shucked_weight) %>%
  step_interact(terms = ~ longest_shell : diameter) %>%
  step_interact(terms = ~ shucked_weight : shell_weight) %>%
  step_center() %>%
  step_scale()
```

I should not use 'rings' to predict 'age' since we derived age values from rings' values, so it is already presented in our outcome.

```
##Question4
lm_model <- linear_reg() %>% set_engine("lm")
```

```
##Question5
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
```

```
##Question6
lm_fit <- fit(lm_wflow, abalone_train_deleted)

predict(lm_fit, new_data = data.frame(type = 'F', longest_shell = 0.50, diameter = 0.10, height = 0.30,
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.1
```

The predicted age of give abalone is 23.14435.

```
##Question7
abalone_metrics <- metric_set(rsq, rmse, mae)

abalone_train_res <- predict(lm_fit, new_data = abalone_train_deleted %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train_deleted %>% select(age))

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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.554
## 2 rmse    standard       2.13
## 3 mae     standard       1.53
```

The R^2 value is 0.55, root mean squared error value is 2.13, and mean absolute error value is 1.53. We can tell that 55% of the variability observed in 'age' can be explained by this regression model, although it means that this regression model does not explain all of the variation in the response variable around its mean, we are not supposed to get it close to 1 because the model would be subject to over-fitting.