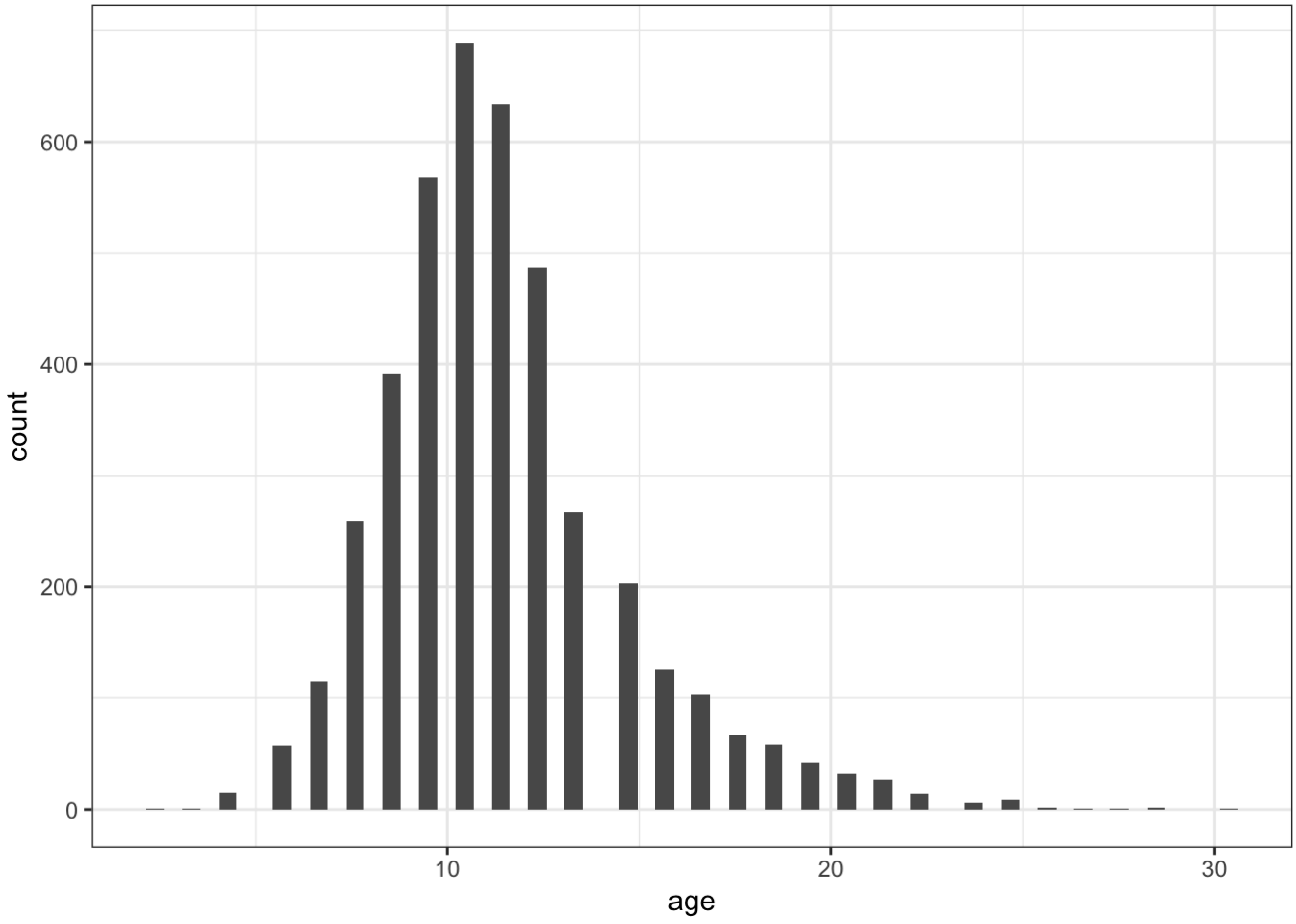
Homework2

Safiya Alavi

10/6/2022

Question 1

```
setwd("/Users/safiyaalavi/Desktop/PSTAT 131/Homework 2/homework-2/data")
abalone_df <- read_csv("abalone.csv")
age <- abalone_df[,9]+1.5
abalone_df <- cbind(abalone_df, age)
names(abalone_df)[10] <- "age"
abalone_df %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 60) +
  theme_bw()
```



The distribution of age predictor in the abalone data is clearly normally distributed. The distribution has positive skewness, or in other words the data is skewed to the right. The mean looks to be about 11.

Question 2

```
set.seed(3435)
abalone_split <- initial_split(abalone_df, prop = 3/4, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
simple_abalone_recipe <-
  recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight
+ viscera_weight + shell_weight, data = abalone_train) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight + longest_shell:diameter +
 shucked_weight:shell_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())

simple_abalone_recipe
```

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

We do not use rings to predict age because age is directly based on rings. Age is equal to 1.5 plus the value of rings, hence we cannot use rings to predict age because the information is the same.

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 × 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     11.4      0.0382    299.   0
##  2 longest_shell                    0.485    0.288       1.68 9.22e- 2
##  3 diameter                         1.96     0.313       6.24 4.91e-10
##  4 height                           0.218    0.0698      3.12 1.80e- 3
##  5 whole_weight                     4.22     0.406      10.4  6.66e-25
##  6 shucked_weight                  -3.85     0.264     -14.6  1.06e-46
##  7 viscera_weight                  -0.817    0.160      -5.11 3.49e- 7
##  8 shell_weight                     1.95     0.224       8.69 5.57e-18
##  9 type_I                          -0.934    0.118      -7.90 3.94e-15
## 10 type_M                          -0.243    0.106      -2.29 2.21e- 2
## 11 type_I_x_shucked_weight          0.486    0.0890      5.46 5.23e- 8
## 12 type_M_x_shucked_weight          0.317    0.113       2.81 4.99e- 3
## 13 longest_shell_x_diameter        -2.62     0.404      -6.50 9.40e-11
## 14 shucked_weight_x_shell_weight   -0.272    0.204      -1.34 1.82e- 1
```

```
new_abalone <- tibble(type = "F", longest_shell = 0.50, diameter = 0.1, height = 0.3, wh
ole_weight = 4, shucked_weight = 1, viscera_weight = 2, shell_weight = 1, rings = 0)
predict(lm_fit, new_data = new_abalone)
```

```
## # A tibble: 1 × 1
##   .pred
##   <dbl>
## 1  22.6
```

Question 7

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

abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.13
## 2 rsq     standard       0.558
## 3 mae     standard       1.54
```