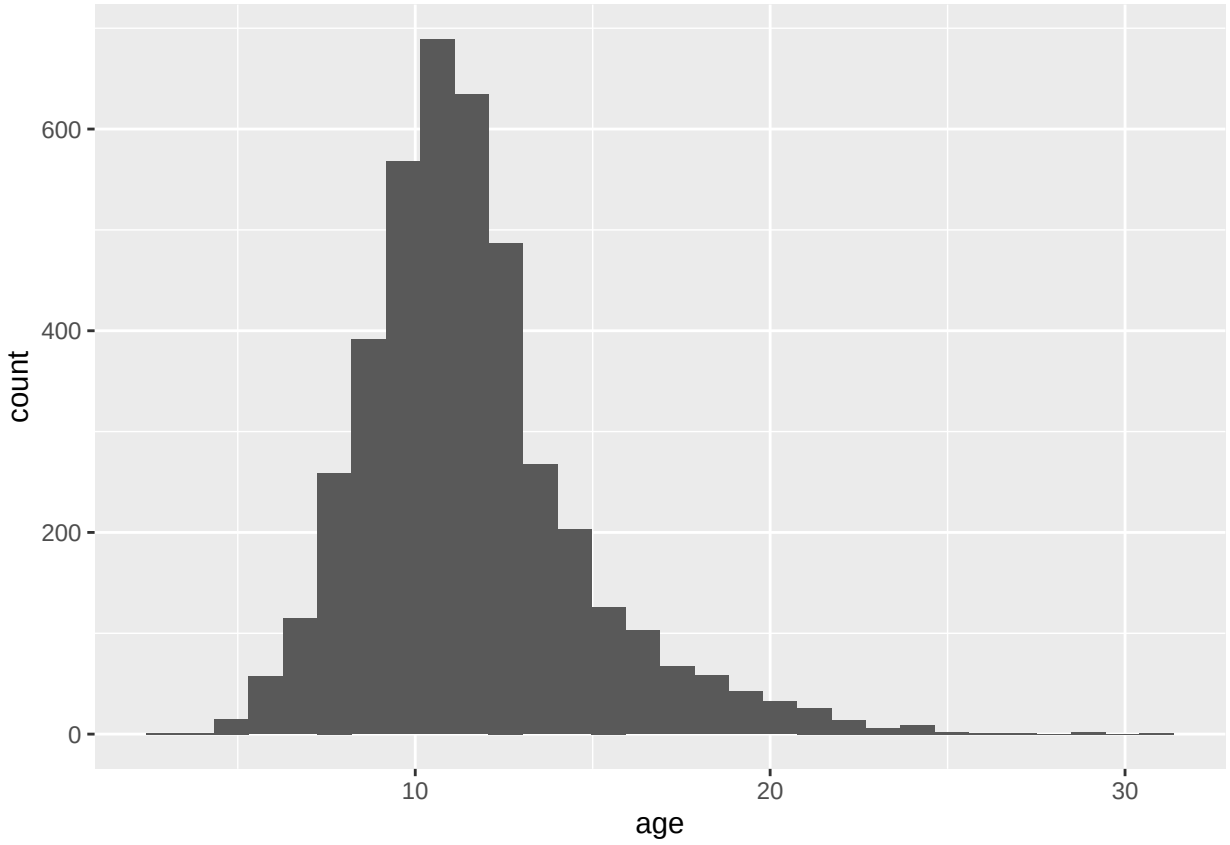
PSTAT 131 HW 2

Eric

4/5/2022

Question 1

```
abalone$age <- abalone$rings + 1.5
ggplot(abalone, aes(x = age)) + geom_histogram(bins = 30)
```




```
mean(abalone$age)
```

```
## [1] 11.43368
```

The distribution of age seems to be almost normally distributed (skewed right a bit), with mean at around 11.433. Most of the abalone ages in this dataset hover around the 10s, and it seems most abalone here do not live past their 20s.

Question 2

```
set.seed(1234)
abalone_split <- initial_split(abalone, prop = 0.80, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

I split the training/testing data 80/20 because I want a good amount of training data to fit my model, but also want to leave a good amount of testing data so that I can properly calculate the model's approximate performance.

Question 3

```
abalone_recipe <- recipe(age ~ ., data = subset(abalone_train, select = -rings)) %>%
  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_normalize(all_predictors())
```

We should not use rings to predict age because both are highly correlated with each other, since we created the age column using the rings variable. Our model may pick up on the pattern and give very good accuracy (maybe even an R^2 of 1).

Question 4

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

Question 5

```
abalone_workflow <- workflow() %>% add_model(abalone_lm) %>% add_recipe(abalone_recipe)
```

Question 6

```
abalone_fit <- fit(abalone_workflow, abalone_train)
onepredict_abalone <- predict(abalone_fit, data.frame(type = "F", longest_shell = 0.5,
                                                      diameter = 0.1, height = 0.3,
                                                      whole_weight = 4,
                                                      shucked_weight = 1,
                                                      viscera_weight = 2,
                                                      shell_weight = 1))
onepredict_abalone
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.5
```

The predicted age of an abalone with the given attributes is 23.488.

Question 7

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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.558
## 2 rmse    standard       2.15
## 3 mae     standard       1.55
```

The *R^2* value is 0.558, the RMSE is 2.15, and the MAE is 1.55. The *R^2* value tells us that our linear regression model was able to explain 55.8% of the variation in the outcome, which means the model is moderately strong.