PSTAT131-HW2

2022-10-15

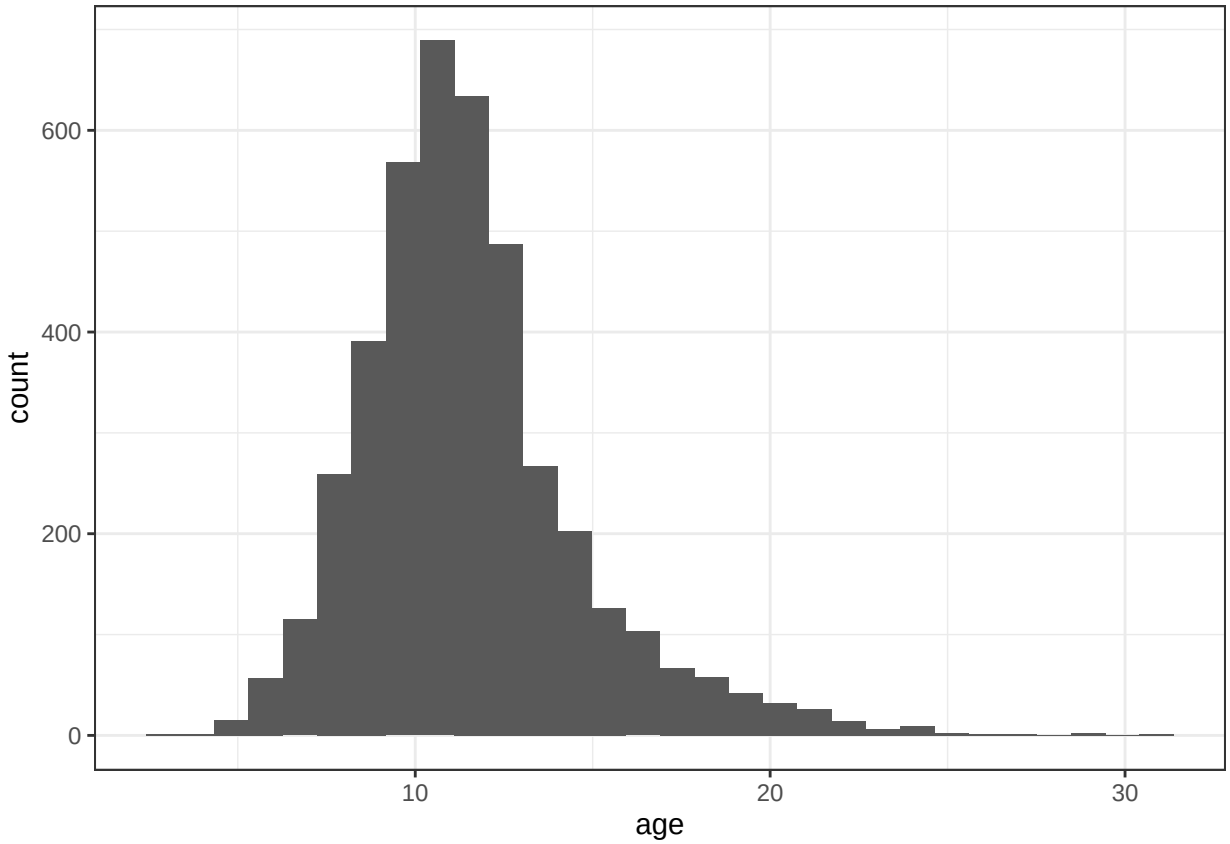
```
library(tidyverse)
library(tidymodels)
```

```
abalone <- read.csv("abalone.csv")
```

1

```
abalone["age"] <- abalone$rings + 1.5

abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30) +
  theme_bw()
```



Age seems to be more or less normally distributed where most of the data is around age = 11 and roughly half of the data is less than 11 and the other half is more than 11.

2

```
set.seed(1234)

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

3

```
abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
  step_dummy(all_nominal_predictors()) # %>%
  step_interact(terms = type ~ shucked_weight, longest_shell ~ diameter,
                shucked_weight ~ shell_weight) %>%
  step_center(.) %>%
  step_scale(.)
```

```
## `~`(longest_shell, diameter, steps = list(list(terms = type ~
##     shucked_weight, role = shucked_weight ~ shell_weight, trained = FALSE,
##     objects = NULL, sep = "_x_", skip = FALSE, id = "interact_4csRA"),
##     list(terms = list(), role = NA, trained = FALSE, means = NULL,
##         na_rm = TRUE, skip = FALSE, id = "center_Ta8Mo", case_weights = NULL),
##     list(terms = list(), role = NA, trained = FALSE, sds = NULL,
##         factor = 1, na_rm = TRUE, skip = FALSE, id = "scale_uiNZO",
##         case_weights = NULL)))
```

We should not use rings to predict age because the two variables would be very highly correlated. This is because age is equal to rings + 1.5.

4

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

5

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

6

```r
lm_fit <- fit(lm_wflow, abalone_train)

new_abalone <- data.frame(type = "F", longest_shell = 0.50, diameter = 0.10,
                  height =0.30, whole_weight = 4, shucked_weight = 1,
                  viscera_weight = 2, shell_weight = 1, rings = 0, age = 0)

predict(lm_fit, new_data = new_abalone)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  1.50
```

7

```r
abalone_metrics <- metric_set(rsq, rmse, mae)

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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard    1   e+ 0
## 2 rmse    standard    8.95e-13
## 3 mae     standard    8.94e-13
```

The R^2 means that the regression model is able to explain 100% of the variability observed in age.