HW2_JiangShu

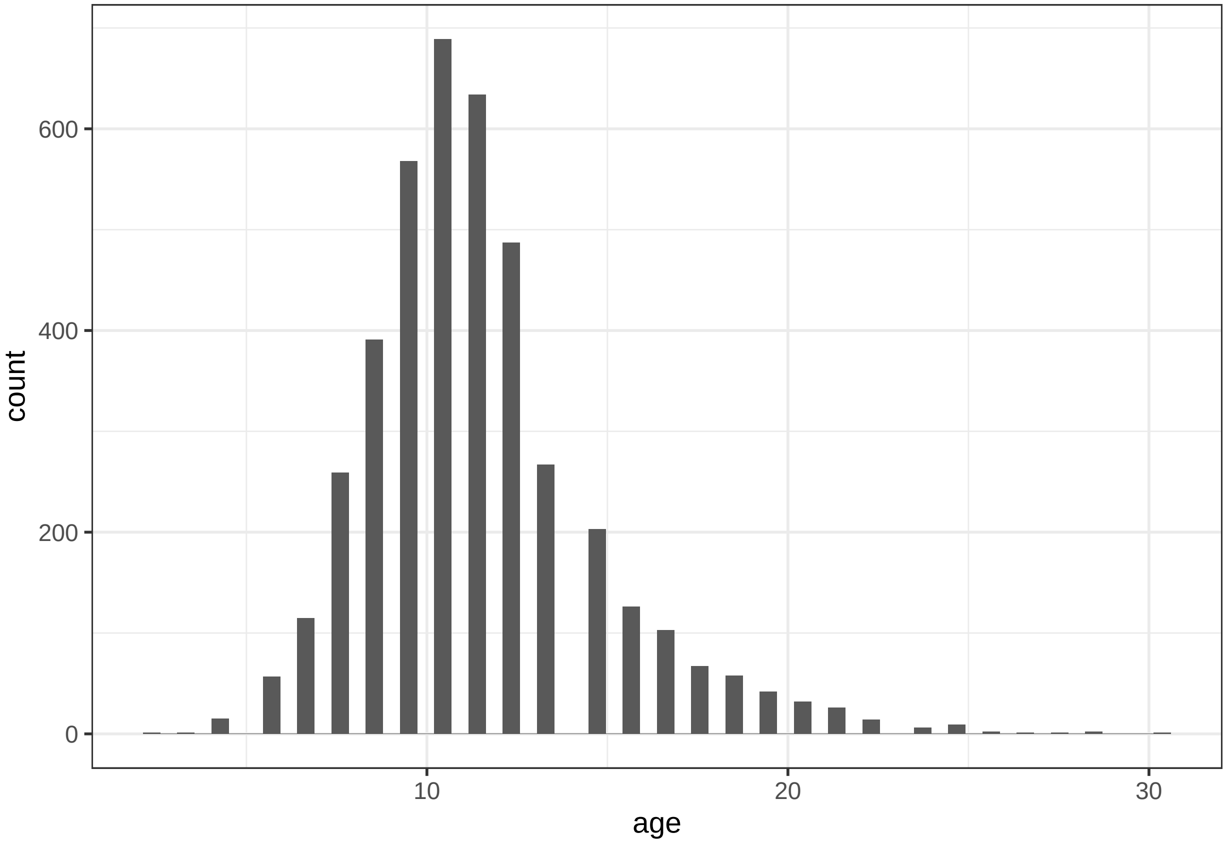
JIANG SHU

2022-10-07

Question 1

```
abalone_revised <- abalone %>%
  mutate(age = rings + 1.5)

abalone_revised %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 60) +
  theme_bw()
```



- We can see the distribution of abalone's age in this data set looks like normal distribution. Most abalones are about 10 years old with little abalone older than 25 or younger than 5.

Question 2

```
set.seed(1112)

abalone_split <- initial_split(abalone_revised, prop = 0.70,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

- The training data set size is set to be 70% of total sample size, which gives 2922 observations, while the other 30% of total sample data comprises test size and gives 1255 observations.

Question 3

```
abalone_train_revised <- abalone_train %>%
  select(-rings)


abalone_recipe <-
  recipe(age ~ ., data = abalone_train_revised) %>%
  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(all_predictors()) %>%
  step_scale(all_predictors())
```

- We shouldn't use rings to predict age because the actual age we assigned to abalone is based on rings, which is essentially our outcome variable.

Question 4

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

Question 5

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

Question 6

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

predict(lm_fit, new_data = 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))
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.4
```

- The predicted age for the female abalone with our model is 23.36107

Question 7

```r
abalone_train_res <- predict(lm_fit, new_data = abalone_train_revised %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train_revised %>% select(age))

abalone_metrics <- metric_set(rsq, rmse, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.566
## 2 rmse    standard       2.11
## 3 mae     standard       1.54
```

- Root mean squared error is 2.11. Mean absolute error is 1.54.We see the R^2 value is 0.566. Generally, R^2 value examine the goodness of fit of a linear regression model. R^2 of 0.566 means our model explains 56.6% of the variation in the response variable around its mean. This is not a particular good R^2 value, but it's also not that bad because we don't want make R^2 too large to have the problem of over fitting.