PSTAT 131 HW2

Eric Liu

```
library(tidyverse)
library(tidymodels)
data <- read_csv(file = "data/abalone.csv")
head(data)
```

```
## # A tibble: 6 x 9
##   type  longest_shell diameter height whole_weight shucked_weight viscera_weight
##   <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M             0.455    0.365  0.095        0.514         0.224          0.101
## 2 M             0.35     0.265  0.09         0.226         0.0995         0.0485
## 3 F             0.53     0.42   0.135        0.677         0.256          0.142
## 4 M             0.44     0.365  0.125        0.516         0.216          0.114
## 5 I             0.33     0.255  0.08         0.205         0.0895         0.0395
## 6 I             0.425    0.3    0.095        0.352         0.141          0.0775
## # ... with 2 more variables: shell_weight <dbl>, rings <dbl>
```

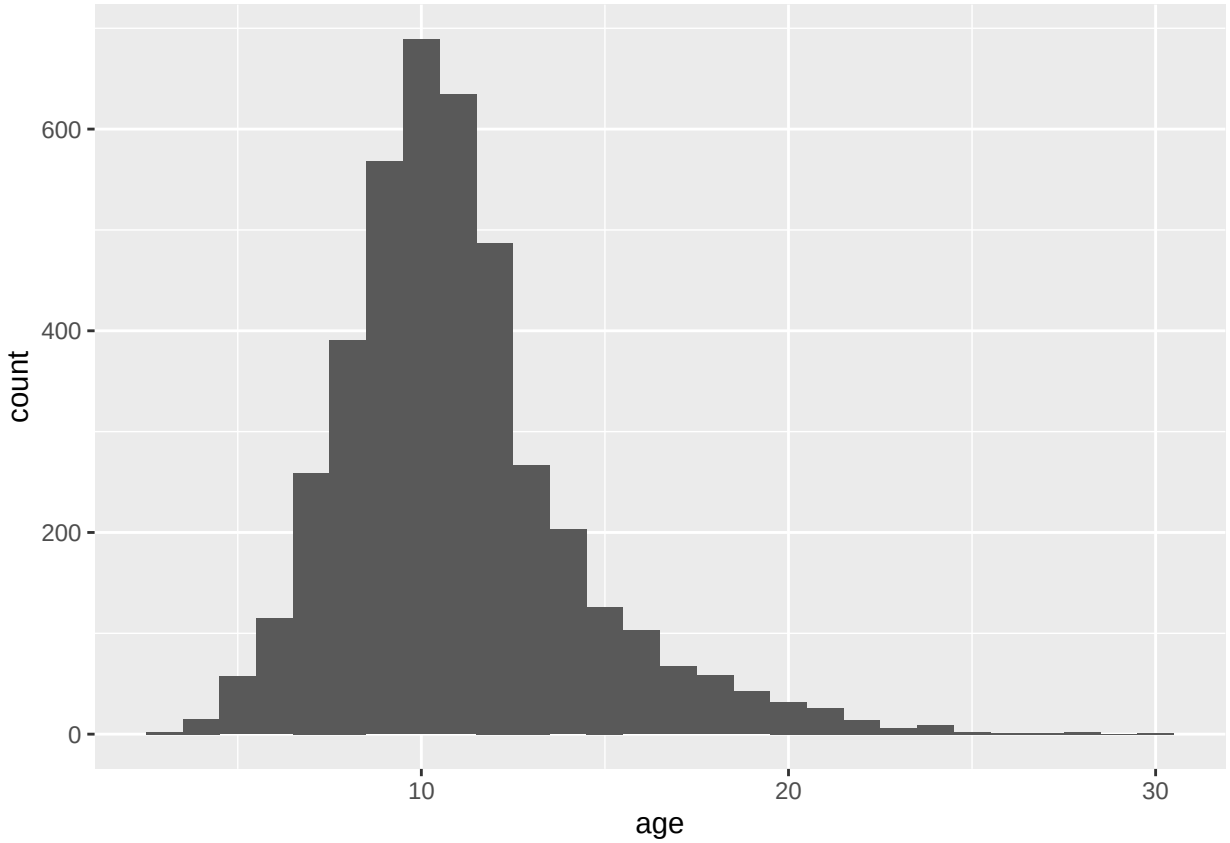
Question 1

```
data$age = data$rings + 1.5
head(data)
```

```
## # A tibble: 6 x 10
##   type  longest_shell diameter height whole_weight shucked_weight viscera_weight
##   <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M             0.455    0.365  0.095        0.514         0.224          0.101
## 2 M             0.35     0.265  0.09         0.226         0.0995         0.0485
## 3 F             0.53     0.42   0.135        0.677         0.256          0.142
## 4 M             0.44     0.365  0.125        0.516         0.216          0.114
## 5 I             0.33     0.255  0.08         0.205         0.0895         0.0395
## 6 I             0.425    0.3    0.095        0.352         0.141          0.0775
## # ... with 3 more variables: shell_weight <dbl>, rings <dbl>, age <dbl>
```

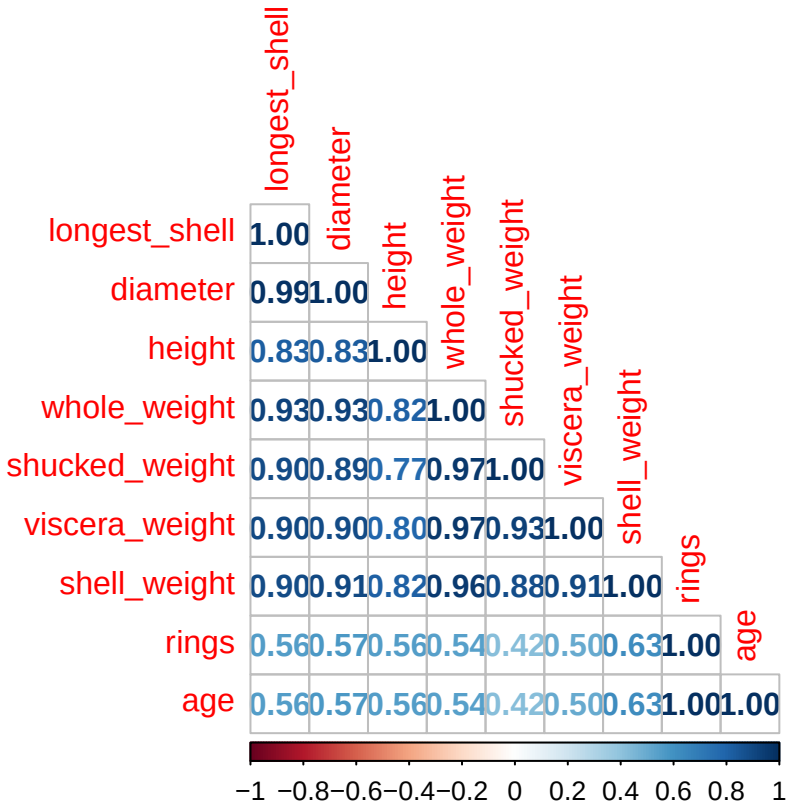
```
library(ggplot2)
library(corrplot)

ggplot(data) + geom_histogram(mapping = aes(x=age), binwidth = 1)
```




```
data %>%
  select(is.numeric) %>%
  cor() %>%
  corrplot(method = "number", type = "lower")
```

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```



From the histogram, we can see that most abalones have age in 8-12 years, and ten years old abalones are the most. In addition, from the correlation plot, we can see that age is positively correlated with all other numeric variables, and the correlation magnitude with each variable is very close. Among them except rings, shell weight is the most positively correlated, and shucked weight is the least positively correlated.

Question 2

```
set.seed(0)

data_split <- initial_split(data, prop = 0.8, strata = age)

data_train <- training(data_split)
data_test <- testing(data_split)
```

Question 3

```
data_recipe <-
  recipe(age ~ ., data = data_train) %>%
  update_role(rings, new_role = "rings") %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ type_M:shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
```

```
  step_interact(terms = ~ shucked_weight:shell_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())

summary(data_recipe)
```

```
## # A tibble: 10 x 4
##    variable       type    role      source
##    <chr>          <chr>   <chr>     <chr>
##  1 type           nominal predictor original
##  2 longest_shell  numeric predictor original
##  3 diameter       numeric predictor original
##  4 height         numeric predictor original
##  5 whole_weight   numeric predictor original
##  6 shucked_weight numeric predictor original
##  7 viscera_weight numeric predictor original
##  8 shell_weight   numeric predictor original
##  9 rings          numeric rings     original
## 10 age            numeric outcome   original
```

We shouldn't use rings to predict age. Since rings plus 1.5 gives age, rings can be seen as the outcome, so we cannot use the outcome as predictor to predict the outcome. In addition, based on the context of this dataset, we want to find a way to obtained age without knowing rings. Therefore, I used update_role function to make rings not predictor.

Question 4

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

Question 5

```
lm_workflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(data_recipe)

lm_fit <- fit(lm_workflow, data_train)

tidy(lm_fit)
```

```
## # A tibble: 13 x 5
##    term                        estimate std.error statistic  p.value
##    <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                   11.4       0.0377   303.    0
##  2 longest_shell                  0.687     0.286      2.40  1.64e- 2
##  3 diameter                       2.50      0.312      8.00  1.70e-15
##  4 height                         0.218     0.0703     3.10  1.93e- 3
##  5 whole_weight                   5.17      0.408     12.7   6.94e-36
```

```
##  6 shucked_weight                -4.00      0.245    -16.3   1.44e-57
##  7 viscera_weight                -1.04      0.162     -6.44  1.35e-10
##  8 shell_weight                   1.82      0.215      8.45  4.21e-17
##  9 type_I                        -0.364     0.0581    -6.27  4.20e-10
## 10 type_M                        -0.0251    0.0963    -0.260 7.95e- 1
## 11 type_M_x_shucked_weight        0.0437    0.101      0.434 6.64e- 1
## 12 longest_shell_x_diameter      -3.31      0.386     -8.59  1.31e-17
## 13 shucked_weight_x_shell_weight -0.428     0.202     -2.12  3.44e- 2
```

Question 6

```
data_point = data.frame(type="F", longest_shell=0.5, diameter=0.1, height=0.3, whole_weight=4, shucked_

predict(lm_fit, data_point)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.0
```

Question 7

```
library(yardstick)
data_metrics <- metric_set(rsq, rmse, mae)

data_train_res <- predict(lm_fit, new_data = data_train %>% select(-age))
data_train_res <- bind_cols(data_train_res, data_train %>% select(age))

data_metrics(data_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.555
## 2 rmse    standard       2.18
## 3 mae     standard       1.57
```

The obtained R^2 is 0.555, RMSE is 2.18, and MAE is 1.57.
R^2 measures the proportion of variability in the outcome that can be explained by the regression. The obtained R^2 is 0.555, which indicates that the trained linear regression model explains 55.5% of the variability in the outcome of the training data.