HW2

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(readr)
library(yardstick)
```

Import dataset

```
abalone <- read_csv(file = 'abalone.csv')
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

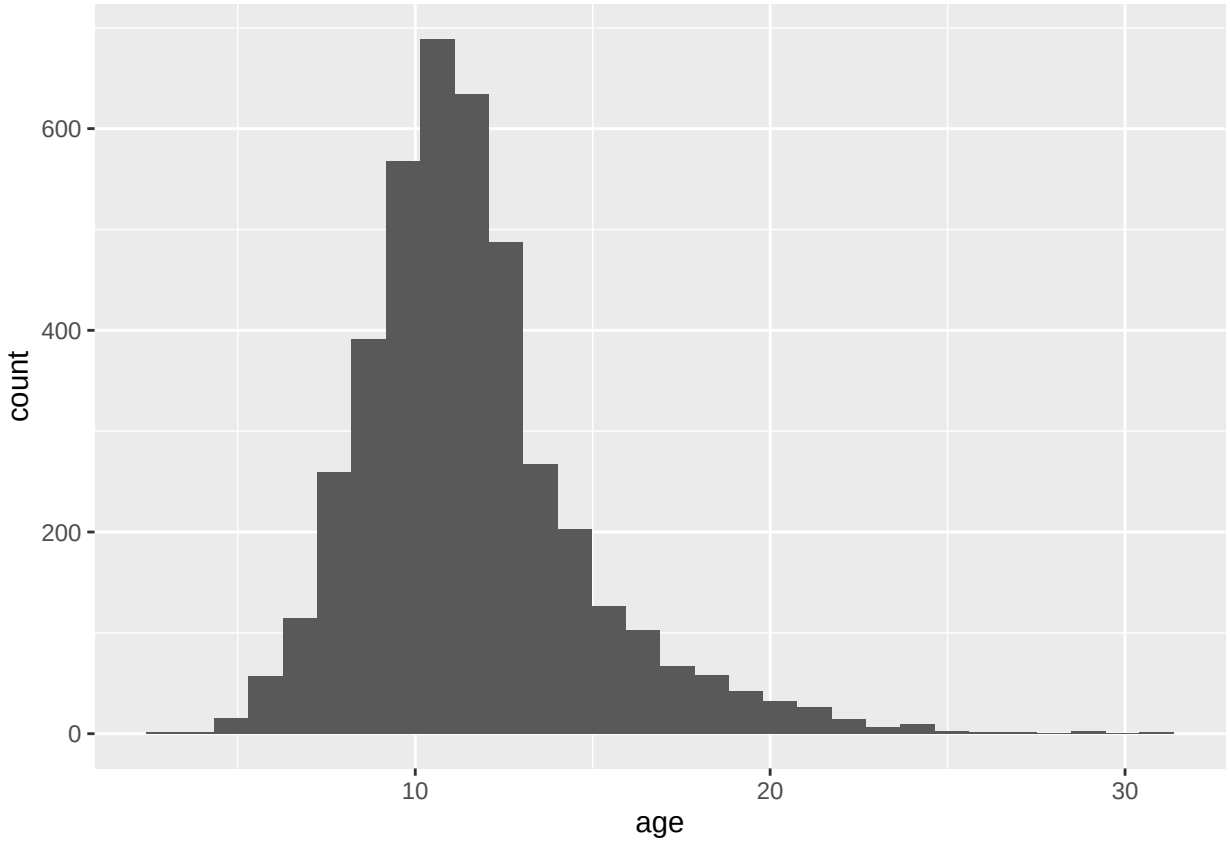
```
view(abalone)
```

Q1

```
age <- abalone$rings + 1.5
abalone_new <- cbind(abalone, age)
view(abalone_new)
```

```
ggplot(data = abalone_new, aes(x = age)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of age variable seems approximately normal to me. The data is right skewed.

Q2

```
set.seed(1234)
abalone_split <- initial_split(abalone_new, prop = 0.75, strata = age)
abalone_training <- training(abalone_split)
abalone_testing <- testing(abalone_split)
```

Q3

```
abalone_training_2 <- select(abalone_training, -rings)
recipe <- recipe(age ~ ., data = abalone_training_2) %>% step_dummy(type)

recipe <- step_interact(recipe, terms = ~ shucked_weight : starts_with('type'))
recipe <- step_interact(recipe, terms = ~ diameter : longest_shell)
recipe <- step_interact(recipe, terms = ~ shell_weight : shucked_weight)

recipe <- step_center(recipe, longest_shell, diameter, height, whole_weight, shucked_weight,
                      viscera_weight, shell_weight)
recipe <- step_scale(recipe, longest_shell, diameter, height, whole_weight, shucked_weight,
                     viscera_weight, shell_weight)
```

Rings variable should not be included because age variable is directly calculated from rings variable(linear relationship), so it is meaningless to include age while predicting rings.

Q4

```
lm_model <- linear_reg() %>% set_engine('lm') %>% set_mode('regression')
lm_model
```

```
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

Q5

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

Q6

```
lm_fit <- fit(lm_workflow, abalone_training_2)

type = c('F')
longest_shell = c(0.5)
diameter = c(0.1)
height = c(0.3)
whole_weight = c(4)
shucked_weight = c(1)
viscera_weight = c(2)
shell_weight = c(1)
```

```
predict_abalone <- data.frame(type, longest_shell, diameter, height, whole_weight,
                              shucked_weight, viscera_weight, shell_weight)
predict(lm_fit, new_data = predict_abalone)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.8
```

Q7

```
abalone_metric <- metric_set(rsq, rmse, mae)
abalone_training_3 <- select(abalone_training_2, -age)
abalone_training_pred <- predict(lm_fit, new_data = abalone_training_3)
abalone_training_age <- select(abalone_training_2, age)
abalone_training_combined <- bind_cols(abalone_training_pred, abalone_training_age)
abalone_training_combined
```

```
## # A tibble: 3,131 x 2
##    .pred   age
##    <dbl> <dbl>
##  1  9.33   9.5
##  2  9.81   8.5
##  3 10.1    9.5
##  4  5.81   6.5
##  5  5.94   5.5
##  6  8.63   8.5
##  7  7.74   7.5
##  8 12.6    9.5
##  9 11.3    9.5
## 10 10.1    8.5
## # ... with 3,121 more rows
```

```
abalone_metric(abalone_training_combined, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.556
## 2 rmse    standard       2.14
## 3 mae     standard       1.55
```

R-squared: 0.55627 RMSE: 2.13766 MAE: 1.54726 An R-square value of 0.55627 means that about 55.63% of the variation in age can be explained by the predictors.