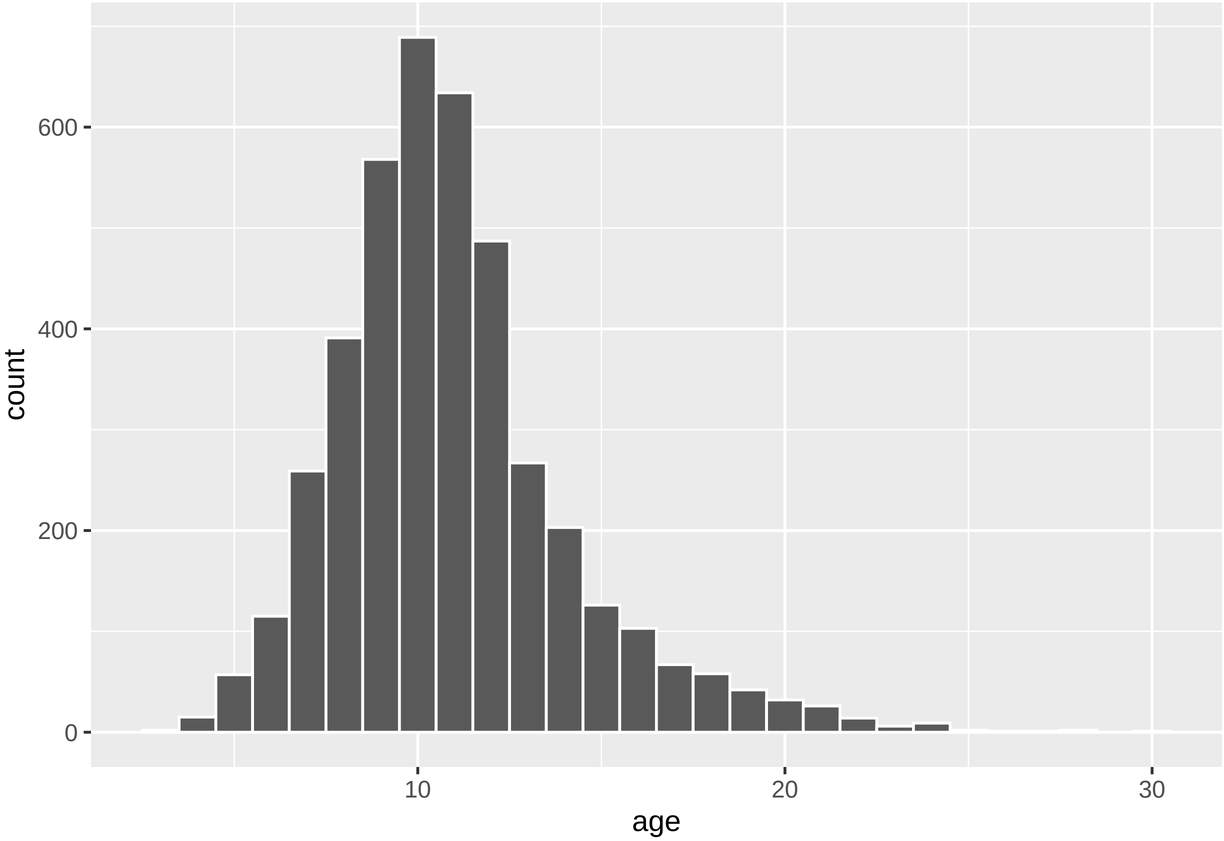
Pstat231HW2

Wentao Yu

2022-10-04

```
data1 <- read.csv('/Users/wentaoyu/Documents/UCSB File/Stats/Pstat131/Pstat131/homework-2/homework-2/da
#data1 <- read.csv('D:/Github/Pstat131/homework-2/homework-2/data/abalone.csv')
age <- data1[9]+1.5 # extract the rings column since age = rings+1.5
data2 <- cbind(data1,age) # add a new column into the dataframe
names(data2)[10] <- 'age' # rename the new column as age
ggplot(data2, aes(age))+
  geom_histogram(col='white', binwidth = 1) # plot the age column using histogram to make it access.
```



Question 1
The distribution of age is more likely a normal distribution with positive skew.

```
set.seed(4177) # set seed to make sure the output is stable
data3 = subset(data2, select = -c(rings)) # new dataframe exclude rings
abalone_split <- initial_split(data3, prop = 0.80) # split the data set, then what is the appropriate p
abalone_training <- training(abalone_split) # this is training data set
```

```
abalone_testing <- testing(abalone_split) # this is testing data set
```

Question 2

```
#simple_abalone_recipe <- recipe(age~.,data = abalone_training)
abalone_recipe <- recipe(age~., data = abalone_training) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ shucked_weight:starts_with('type')+
                  diameter:longest_shell+
                  shell_weight:shucked_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
summary(abalone_recipe)
```

Question 3

```
## # A tibble: 9 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 age            numeric outcome   original
```

```
lm_model <- linear_reg() %>%
  set_engine('lm') # create and store a linear regression object
lm_model
```

Question 4

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

```
lm_workflow <- workflow() %>% # set up a new workflow
  add_model(lm_model) %>% # add the linear model from question 4
  add_recipe(abalone_recipe) # add the recipe from question 3
lm_workflow
```

Question 5

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: linear_reg()
## 
## -- Preprocessor ----------------------------------------------------------------
## 4 Recipe Steps
## 
```

```
## * step_dummy()
## * step_interact()
## * step_center()
## * step_scale()
##
## -- Model -----------------------------------------------------------------------
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

```
#create a new dataframe including the question conditions
type <- c('F')
longest_shell <- 0.50
diameter <- 0.10
height <- 0.30
whole_weight <- 4
shucked_weight <- 1
viscera_weight <- 2
shell_weight <- 1
hypo1 <- data.frame(type, longest_shell, diameter, height, whole_weight, shucked_weight, viscera_weight
lm_fit <- fit(lm_workflow, abalone_training) # fit the training data
predict(lm_fit, hypo1) # predict the age using fitted training data from a new dataframe
```

Question 6

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.4
```

```
abalone_training_res <- predict(lm_fit, new_data = abalone_training %>% select(-age))
abalone_training_res <- bind_cols(abalone_training_res, abalone_training %>% select(age))
abalone_matrics <- metric_set(rsq, rmse, mae)
abalone_matrics(abalone_training_res,truth = age, estimate = .pred)
```

Question 7

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.550
## 2 rmse    standard       2.14
## 3 mae     standard       1.55
```

Required For 231 Students

Question 8 $\mathrm{Var}(\hat{f}(x_0))$ and $Bias(\hat{f}(x_0))^2$ are the reproducible errors.

$\mathrm{Var}(\epsilon)$ is the irreducible error.

Question 9

Question 10