hw02

Mason Ma

10/3/2022

Question 1

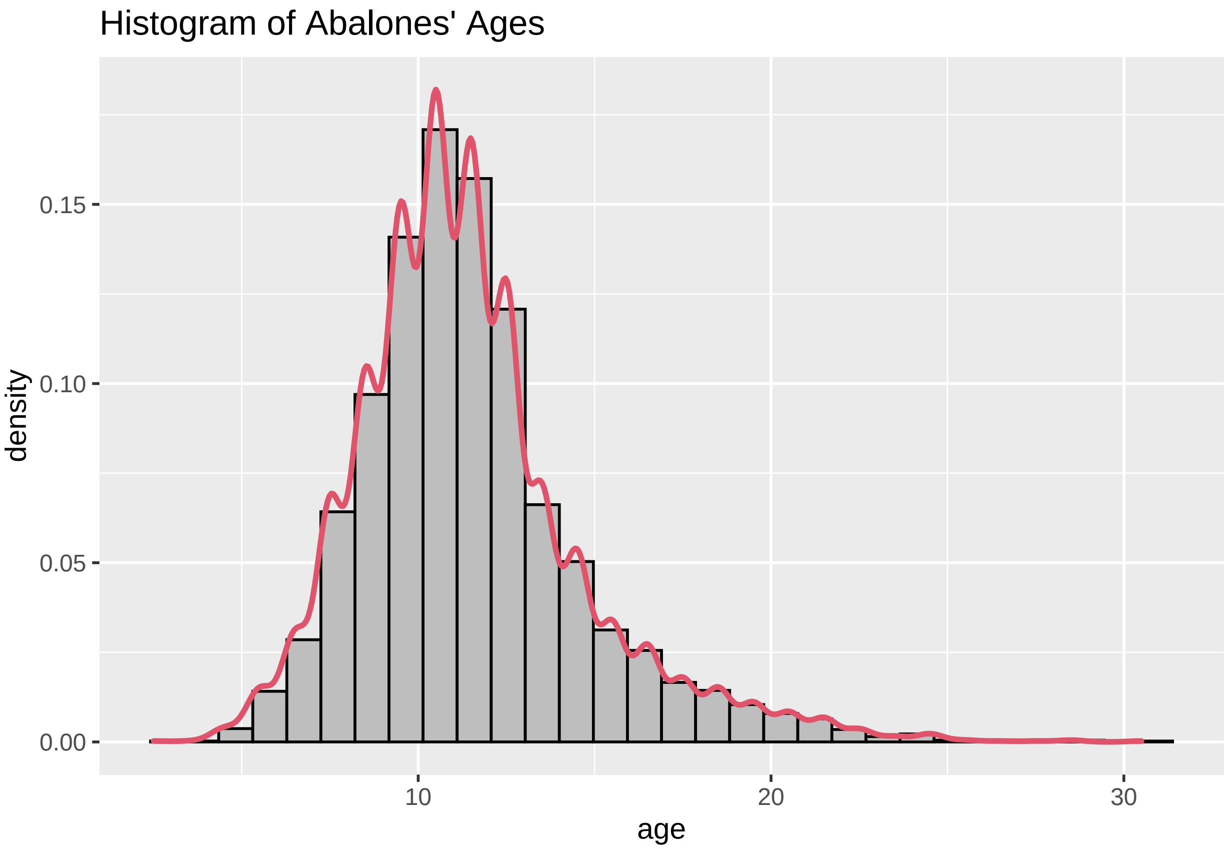
```
library(tidyverse)
```

```
## -- 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   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
abalone<-read.csv("/Users/mason/Desktop/homework-2/data/abalone.csv")

my_data<-mutate(abalone,age = rings+1.5)
age_plot<-ggplot(my_data,aes(x = age))+ geom_histogram(aes(y = ..density..),colour = 1, fill = "grey")
print(age_plot+ggtitle("Histogram of Abalones' Ages"))
```

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



From the output of the histogram of the Abalones' Ages, we can see that this plot appears to be relatively normally distributed with a longer tail on the right side–(right-skewed). Most of the data fall within the range between 7 and 15.

Question 2

```
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 tidymodels_prefer() to resolve common conflicts.
```

```
set.seed(713)
abalone_split <- initial_split(my_data, prop = 0.80,strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
library(tidymodels)
simple_abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
  step_rm(rings) %>%
## not include rings to predict age
  step_dummy(all_nominal_predictors()) %>%
## dummy code any categorical predictors
  step_interact(terms = ~ starts_with("type"):shucked_weight
                +longest_shell:diameter
                +shucked_weight:shell_weight) %>%
## create interactions
  step_normalize(all_predictors())
## center all predictors and scale all predictors.
```

Since the age is calculated as the number of rings plus 1.5, then if we include rings to predict age, there will be a rather strong relationship between rings and age. In that case, those remaining predictors would be meaningless and the model will lose its usage at all.

Question 4

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

Question 5

```
simple_abalone_wflow <- workflow() %>%
## set up an empty workflow
  add_model(lm_model) %>%
## add the model you created in Question 4
  add_recipe(simple_abalone_recipe)
## add the recipe that you created in Question 3

simple_abalone_fit <- fit(simple_abalone_wflow, abalone_train)
```

Question 6

```
## Use your fit() object to predict the age of a female abalone
## with these parameters below.
female_abalone <- tibble(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)
```

```
predict(simple_abalone_fit, new_data = female_abalone)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.3
```

We can see the predicted value of this hypothetical female abalone is 23.347 years.

Question 7

```
library(yardstick)
metric_group<-metric_set(rmse, rsq, mae)
## Create a metric set that includes R^2, RMSE, and MAE.
simple_abalone_predict <- predict(simple_abalone_fit, abalone_train) %>%
 bind_cols(abalone_train %>% select(age))
## Use predict() and bind_cols() to create a tibble of your model's predicted values from the training
metric_group(simple_abalone_predict, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.16
## 2 rsq     standard       0.558
## 3 mae     standard       1.55
```

```
## apply your metric set to the tibble, report the results, and interpret the R^2 value.
```

From the output above, we see the R-squared value (rsq) is 0.5575, which means that only about 55.75% of the variation in the dataset could be explained by out model. This is probably because the relationship between the outcome variable (age) and the predictors may not be linear. If we aim to lower the error, we can try to use other model to predict.