R Notebook

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## 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
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## -- 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      1.0.1     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
##
## corrplot 0.92 loaded
```

Question 1:

code from Lab 2

```
# Changing my working directory (for some reason, read_csv won't work without this step)
setwd("~/Downloads/homework-2")
abalone_data <- read_csv(file = "data/abalone.csv")
```

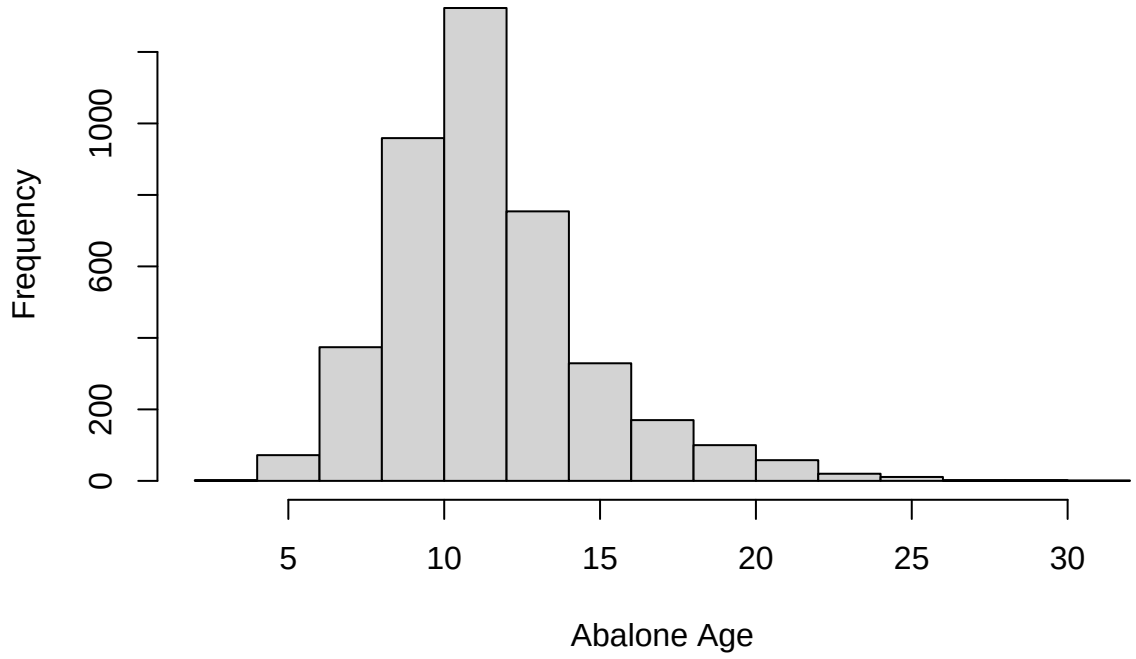
```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this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

```r
# Creating "age" variable and binding it to abalone_data data frame
age <- abalone_data[9] + 1.5

abalone_data_age <- cbind(age, abalone_data)

colnames(abalone_data_age)[1] <- "age"

# Creating a histogram to assess distribution of "age" variable
hist(abalone_data_age$age, xlab = "Abalone Age", main = "")
```



Based off the histogram, we can see that abalone between the ages of 10 and 12 are most prevalent in this dataset. Abalone between the age of 12 and 14, as well as 8 and 10 are also observed frequently. Abalone older than 14 and younger than 8 are observed rarely.

Question 2:

code from Lab 2

```r
# Splitting the abalone data, with 80% of the data going to the training set, and 20% of the data going
abalone_split <- abalone_data_age %>% initial_split(strata = age, prop = 0.80)

abalone_train <- testing(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3:

code from Lab 2

```
# Dummy coding categorical predictors, creating interactions between variables, scaling all predictors,
abalone_steps <- recipe(age ~ ., data = abalone_train) %>% step_rm(rings) %>% step_dummy(all_nominal_pr
```

Question 4:

code from lab 2

```
# Creating lm object, storing in lm_abalone_model
lm_abalone_model <- linear_reg() %>% set_engine("lm")
```

Question 5:

code from Lab 2

```
# Setting up empty workflow, adding lm_abalone_model, and adding abalone_steps
abalone_workflow <- workflow() %>% add_recipe(abalone_steps) %>% add_model(lm_abalone_model)
```

Question 6:

code from Lab 2

```
# Fitting linear model to abalone training data set
abalone_fit <- fit(abalone_workflow, abalone_train)

# Creating data frame to hold inputs we want to use for our prediction
predict_input <- data.frame(type = "F", longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_wei

# Predicting age of hypothetical abalone with inputs from predict_input
predict(abalone_fit, new_data = predict_input)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  18.1
```

Question 7:

code from Lab 2

```
# Creating data set of predicted values of age versus actual values of age
abalone_train_results <- predict(abalone_fit, abalone_train) %>% bind_cols(abalone_train %>% dplyr::sel
head(abalone_train_results)
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.61   8.5
## 2 10.3   10.5
## 3 12.5   12.5
## 4 10.8   12.5
## 5 12.1   15.5
## 6  6.22   6.5
```

```r
# Creating a metric set that includes R squared, RMSE, and MAE
abalone_metrics <- metric_set(rmse, rsq, mae)

# Generating R squared, RMSE, and MAE values of the model
abalone_metrics(abalone_train_results, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.07
## 2 rsq     standard       0.580
## 3 mae     standard       1.51
```

The RMSE value was 2.0734525, the R squared value was 0.5800694, and the MAE value was 1.5068976. The R squared value tells us that approximately 58% of the variability that is observed in the "age" variable can be explained by our model.