121HW2

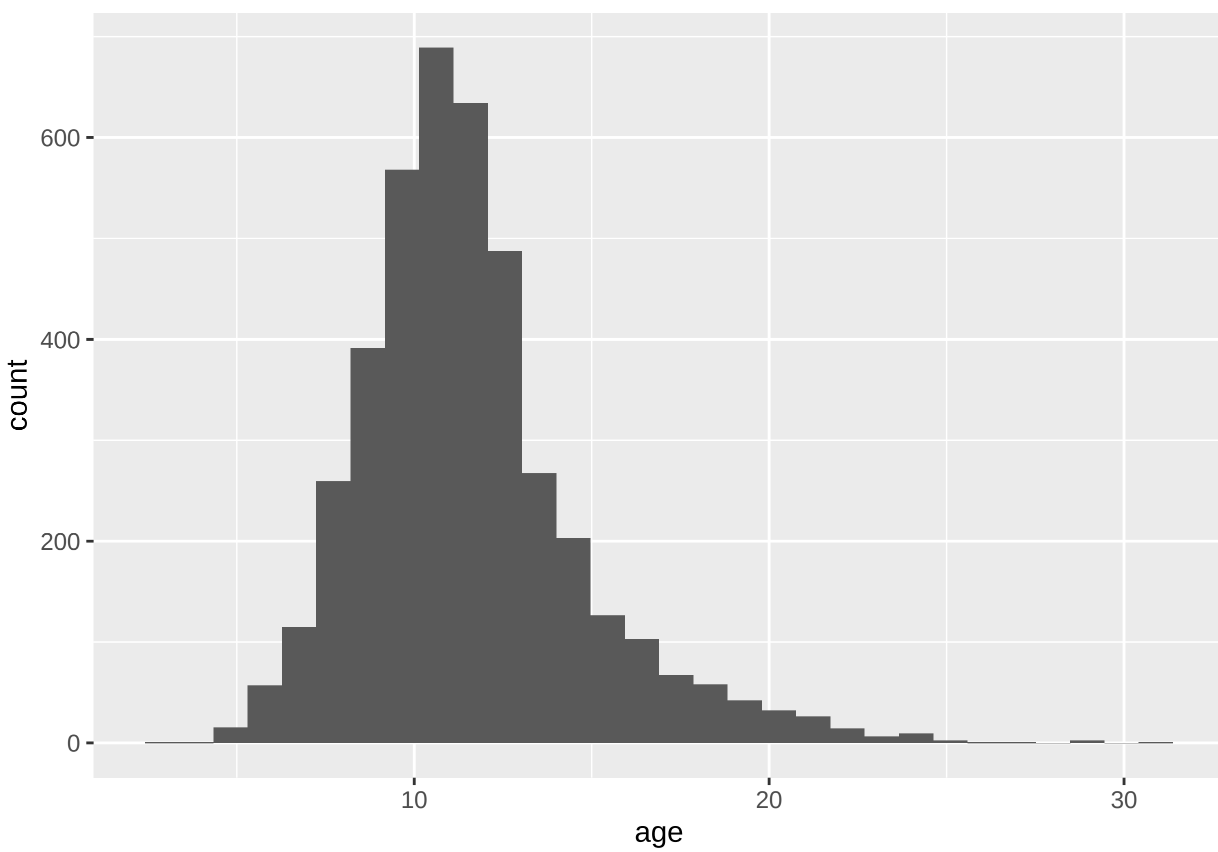
Linear Regression

Question 1

Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no `age` variable in the data set. Add `age` to the data set.

Assess and describe the distribution of `age`.

```
abalone['age'] <- abalone$rings + 1.5
abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins=30)
```



The distribution of age seems to be very slightly right skewed and unimodal with most of the abalone in the dataset seem to be around 10 years old. There are very few abalone below 5 and above 20 years old.

Question 2

Split the abalone data into a training set and a testing set. Use stratified sampling. You should decide on appropriate percentages for splitting the data.

```
set.seed(3435)
abalone_split <- initial_split(abalone, prop = 0.80,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

Using the **training** data, create a recipe predicting the outcome variable, `age`, with all other predictor variables. Note that you should not include `rings` to predict `age`. Explain why you shouldn't use `rings` to predict `age`.

Steps for your recipe:

1. dummy code any categorical predictors
2. create interactions between
   - `type` and `shucked_weight`,
   - `longest_shell` and `diameter`,
   - `shucked_weight` and `shell_weight`
3. center all predictors, and
4. scale all predictors.

You'll need to investigate the `tidymodels` documentation to find the appropriate step functions to use.

```
simple_abalone_recipe <-
  recipe(age ~ type+longest_shell+diameter+height+whole_weight+shucked_weight+viscera_weight+shell_weig
  step_dummy(all_nominal_predictors())  %>%
  step_interact(terms = ~ type_I:shucked_weight + type_M:shucked_weight + longest_shell:diameter + shuc
  step_normalize(all_predictors())
```

Age is directly dependent on rings so it wouldn't make sense to use that as a predictor since the age of an abalone is just the amount of rings +1.5. Also,if we are attempting to make a predictive model for the age of abalone, we cannot accurately identify the relationships between the response and predictors if rings is in the model.

Question 4

Create and store a linear regression object using the `"lm"` engine.

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

Question 5

Now:

1. set up an empty workflow,
2. add the model you created in Question 4, and
3. add the recipe that you created in Question 3.

```
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(simple_abalone_recipe)
lm_fit <- fit(lm_wflow, abalone_train)
lm_fit
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ------------------------------------------------------------------
## 3 Recipe Steps
##
## * step_dummy()
## * step_interact()
## * step_normalize()
##
## -- Model -------------------------------------------------------------------------
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##                  (Intercept)                 longest_shell
##                    11.423353                      0.591227
##                     diameter                        height
##                     2.064936                      0.235918
##                 whole_weight                shucked_weight
##                     4.288839                     -4.061214
##               viscera_weight                  shell_weight
##                    -0.791796                      1.735662
##                       type_I                        type_M
##                    -0.942109                     -0.238578
##       type_I_x_shucked_weight        shucked_weight_x_type_M
##                     0.524826                      0.293330
##      longest_shell_x_diameter  shucked_weight_x_shell_weight
##                    -2.753237                     -0.003297
```

Question 6

Use your `fit()` object to predict the age of a hypothetical female abalone with longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, shucked_weight = 1, viscera_weight = 2, shell_weight = 1.

```
x0 <- data.frame(type = "F",longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, shu
abalone_x0_predicted <- predict(lm_fit, new_data = x0)
abalone_x0_predicted
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.7
```

Using predict() the age of this hypothetical abalone is around 23.702.

Question 7

Now you want to assess your model's performance. To do this, use the `yardstick` package:

1. Create a metric set that includes R^2, RMSE (root mean squared error), and MAE (mean absolute error).
2. Use `predict()` and `bind_cols()` to create a tibble of your model's predicted values from the **training data** along with the actual observed ages (these are needed to assess your model's performance).
3. Finally, apply your metric set to the tibble, report the results, and interpret the R^2 value.

```
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
abalone_metrics <- metric_set(rmse, rsq, mae)
```

```
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.16
## 2 rsq     standard       0.551
## 3 mae     standard       1.55
```

Because the R^2 value is around .5513, or 55%, a little over half of the observed variation can be explained by our predictor variables. This indicates a moderately precise relationship for our model. I think it would be reasonable to use it as long as you acknowledge the unexplained variability.