Abalone Final Project

Chloe Prowse & Nashita Khandaker

2025-12-14

Objective 1: Building a Predictive Model

```
# Upload the abalone dataset.

abalone <- read.csv("trainabalone.csv")

summary(abalone)
```

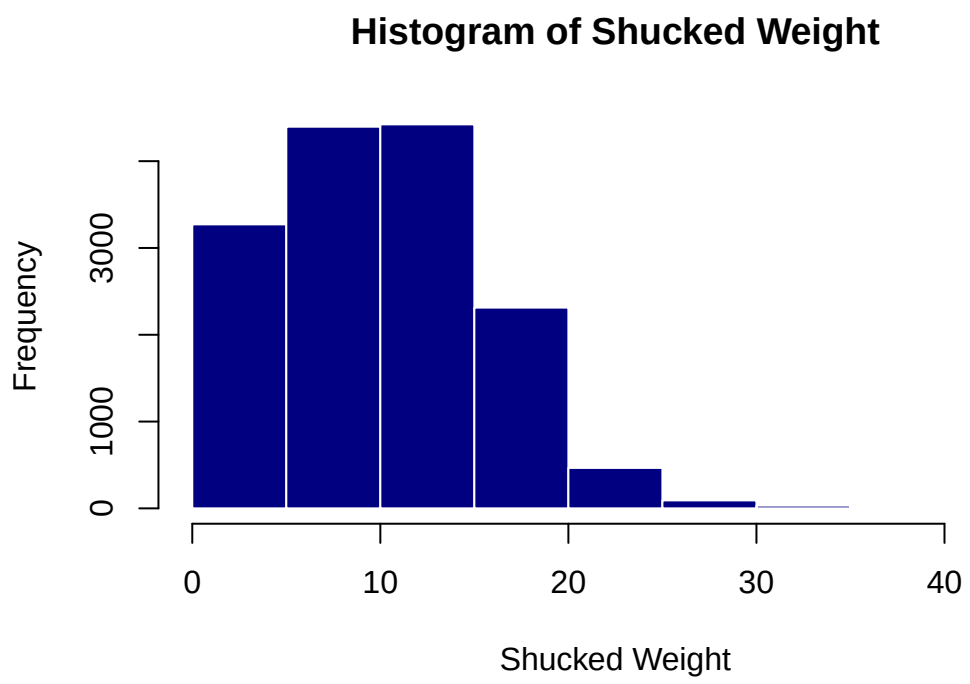
```
##        id            Sex                Diameter          Length
##  Min.   :    1   Length:15000       Min.   :0.0000   Min.   :0.200
##  1st Qu.: 3751   Class :character   1st Qu.:0.8875   1st Qu.:1.150
##  Median : 7500   Mode  :character   Median :1.0750   Median :1.375
##  Mean   : 7500                      Mean   :1.0221   Mean   :1.315
##  3rd Qu.:11250                      3rd Qu.:1.2000   3rd Qu.:1.538
##  Max.   :15000                      Max.   :1.5750   Max.   :2.038
##      Height           Weight         Shucked.Weight     Viscera.Weight
##  Min.   :0.0000   Min.   : 0.2268   Min.   : 0.09922   Min.   : 0.01417
##  1st Qu.:0.2875   1st Qu.:13.1967   1st Qu.: 5.66990   1st Qu.: 2.80660
##  Median :0.3625   Median :23.5584   Median : 9.86563   Median : 4.89029
##  Mean   :0.3467   Mean   :23.2434   Mean   :10.04032   Mean   : 5.01447
##  3rd Qu.:0.4125   3rd Qu.:32.1625   3rd Qu.:13.99048   3rd Qu.: 6.98815
##  Max.   :0.6000   Max.   :75.3246   Max.   :42.18406   Max.   :20.12814
##   Shell.Weight           Age
##  Min.   : 0.09922   Min.   : 1.000
##  1st Qu.: 3.82718   1st Qu.: 8.000
##  Median : 6.80388   Median :10.000
##  Mean   : 6.67329   Mean   : 9.985
##  3rd Qu.: 9.07184   3rd Qu.:11.000
##  Max.   :29.10076   Max.   :29.000
```

```
# I need to make sex as a factor.
abalone$Sex <- as.factor(abalone$Sex)

# To make sure there are no NA columns in the dataset
colSums(is.na(abalone))
```

```
##             id            Sex       Diameter         Length         Height
##              0              0              0              0              0
##         Weight Shucked.Weight Viscera.Weight   Shell.Weight            Age
##              0              0              0              0              0
```

I wanted to check the data to see if any of shucked weights were greater than the total weight. I found that 10 of the observations had a shucked weight greater than the total weight. They were also significantly bigger, averaging about 9-10 units bigger. This statistically does not make sense since shucked weight is just the weight of an abalone without its shell. I decided to remove these values from the dataset.

```
hist(abalone$Shucked.Weight,
     main = "Histogram of Shucked Weight",
     xlab = "Shucked Weight",
     col = "navy",
     border = "white")
```




```
sum(abalone$Shucked.Weight > abalone$Weight)
```

```
## [1] 10
```

```
abalone[abalone$Shucked.Weight > abalone$Weight, ]
```

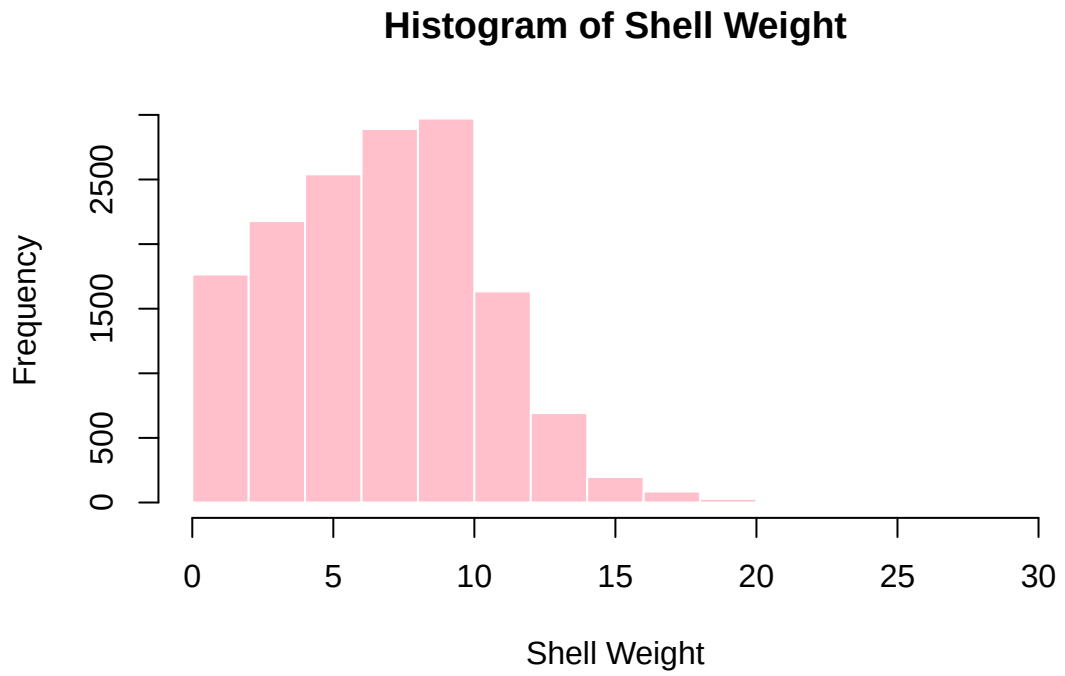
```
##          id Sex Diameter Length Height    Weight Shucked.Weight Viscera.Weight
## 3280   3280   I   0.5750 0.7000 0.1375 2.9908722     14.0330025      0.5244657
## 3950   3950   I   0.5125 0.6625 0.1625 2.9908722     15.5922250      0.8930092
## 6560   6560   I   0.7625 1.0125 0.2625 8.9300925     11.6091203      1.4883488
## 8550   8550   I   0.5250 0.7000 0.1750 2.9908722     14.0330025      0.5244657
## 10327 10327   I   0.2500 0.3500 0.0500 0.2976698      0.4110678      0.0850485
## 11645 11645   I   0.5125 0.6875 0.2000 2.8632995     14.0330025      0.7370870
## 12833 12833   I   0.4375 0.6000 0.1375 2.9908722     14.0330025      0.4677668
## 13129 13129   I   0.5000 0.6875 0.1750 2.9908722     14.0330025      0.3260192
## 13588 13588   M   0.4500 0.6125 0.1625 2.9908722     13.1399932      0.3968930
```

```
## 13752 13752   M   0.5000 0.6625 0.1750 2.9908722     14.0330025      0.5244657
##       Shell.Weight Age
## 3280     0.8504850   5
## 3950     0.9780577   6
## 6560     2.1262125   6
## 8550     0.8504850   6
## 10327    0.1417475   3
## 11645    0.9638830   6
## 12833    0.5669900   5
## 13129    0.8930092   5
## 13588    0.6945628   6
## 13752    0.9922325   3
```

```r
# Filtering out the observations where shucked weight is greater
abalone <- abalone %>%
  filter(Shucked.Weight <= Weight)
```

I wanted to also check if any of the shell weights were bigger than the total weight. None of the shell weights were bigger than the total weight, so I moved forward with checking other variables.

```r
hist(abalone$Shell.Weight,
     main = "Histogram of Shell Weight",
     xlab = "Shell Weight",
     col = "pink",
     border = "white")
```



```r
sum(abalone$Shell.Weight > abalone$Weight)
```

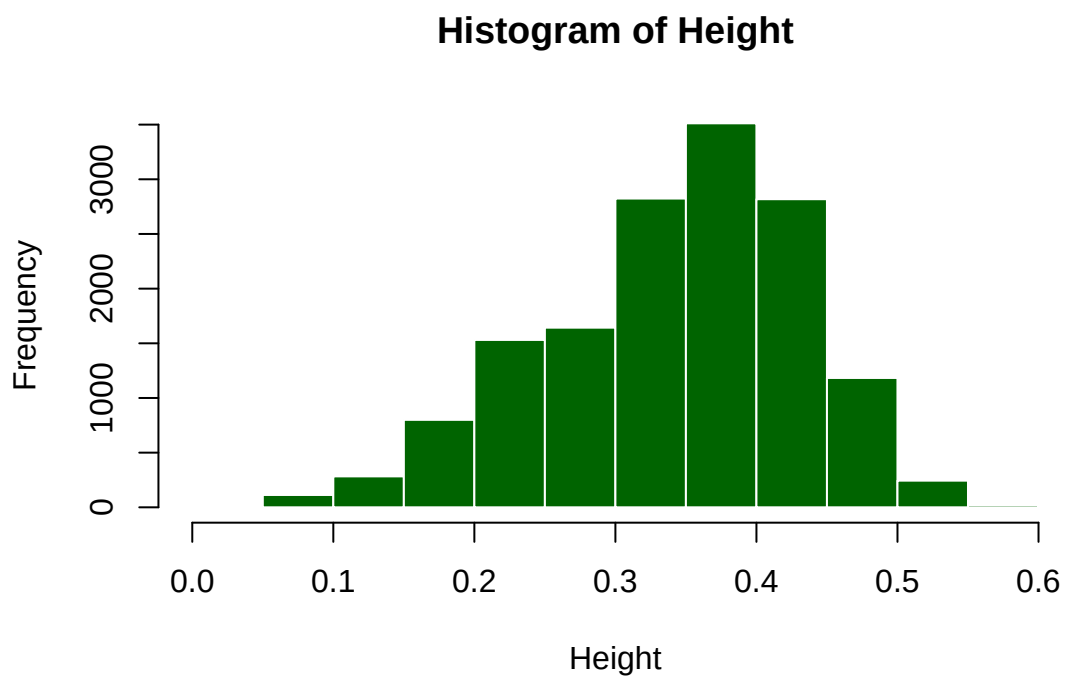
```
## [1] 0
```

```r
abalone[abalone$Shell.Weight > abalone$Weight, ]
```

```
##  [1] id             Sex            Diameter       Length         Height
##  [6] Weight         Shucked.Weight Viscera.Weight Shell.Weight   Age
## <0 rows> (or 0-length row.names)
```

I wanted to check if any of the heights were zero.There were 12 observations that had a height of zero but they were all infant abalones. Since abalones are generally very small, it can be hard to accurately measure the height. We kept these values in the data because they represent a very small portion of the data and aren't drastically different from the other indeterminant height values. They also represented 12 out of about 10,000 observations.This means they won't likely affect the data.

```r
hist(abalone$Height,
     main = "Histogram of Height",
     xlab = "Height",
     col = "darkgreen",
     border = "white")
```




```r
sum(abalone$Height == 0)
```

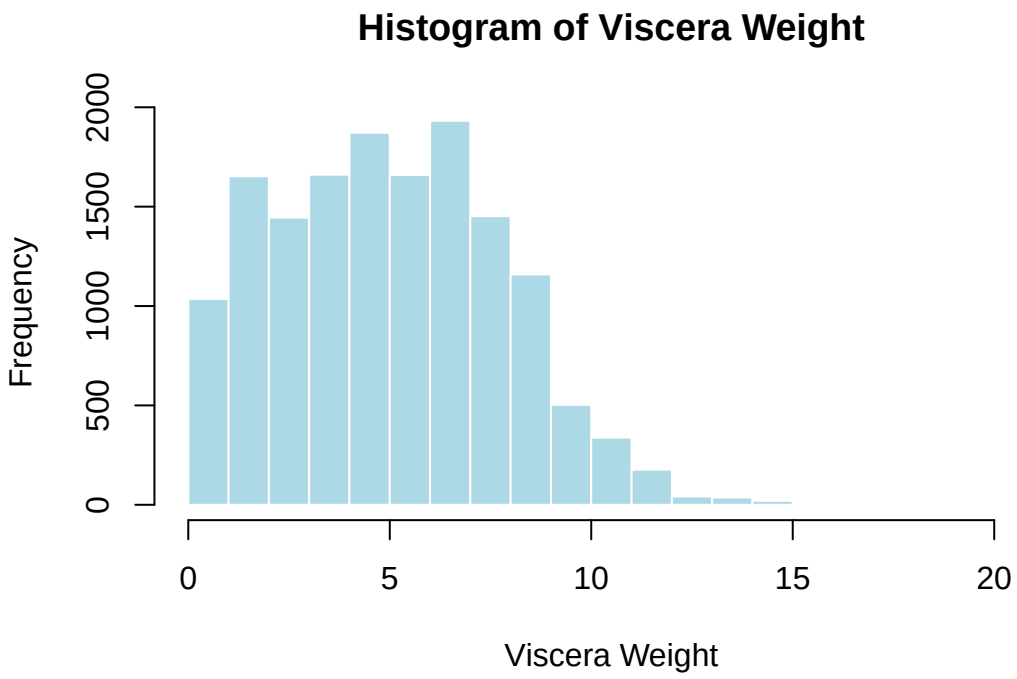
```
## [1] 12
```

```r
abalone[abalone$Height == 0, ]
```

```
##          id Sex Diameter Length Height   Weight Shucked.Weight Viscera.Weight
## 1864   1864   I   0.4250 0.5750      0 1.020582      0.5386405      0.3118445
## 2543   2543   I   0.5000 0.6750      0 2.721552      1.2615527      0.5811648
## 2926   2926   I   0.5000 0.6875      0 2.806601      1.0489315      0.6378637
## 2981   2981   I   0.2000 0.4500      0 0.963883      0.2693202      0.1842718
## 4658   4660   I   0.4250 0.5875      0 1.800193      0.8079608      0.3968930
## 5012   5014   I   0.6125 0.8125      0 4.819415      1.8427175      1.0914557
## 6200   6202   I   0.5000 0.6000      0 2.168737      0.7937860      0.3968930
## 6472   6474   I   0.3625 0.5125      0 1.133980      0.5102910      0.1275728
## 6887   6890   I   0.5250 0.7125      0 3.146794      1.1906790      0.6378637
## 12907 12914   I   0.4875 0.6875      0 2.353009      1.1056305      0.8079608
## 12980 12987   I   0.6500 0.8750      0 5.074560      1.9986397      0.9780577
## 14165 14175   I   0.7500 0.9875      0 7.682715      3.1184450      1.1339800
##       Shell.Weight Age
## 1864     0.3685435   4
## 2543     0.8504850   4
## 2926     0.8504850   4
## 2981     0.2267960   4
## 4658     0.5669900   3
## 5012     1.4174750   4
## 6200     0.6378637   4
## 6472     0.3118445   4
## 6887     0.9071840   5
## 12907    0.6945628   5
## 12980    1.7009700   6
## 14165    2.1403872   6
```

I wanted to also check if any of the viscera weights were bigger than the total weight. None of the viscera weights were bigger than the total weight, so I moved forward with creating a model.

```r
hist(abalone$Viscera.Weight,
     main = "Histogram of Viscera Weight",
     xlab = "Viscera Weight",
     col = "lightblue",
     border = "white")
```




```
sum(abalone$Viscera.Weight > abalone$Weight)
```

```
## [1] 0
```

```
abalone[abalone$Viscera.Weight > abalone$Weight, ]
```

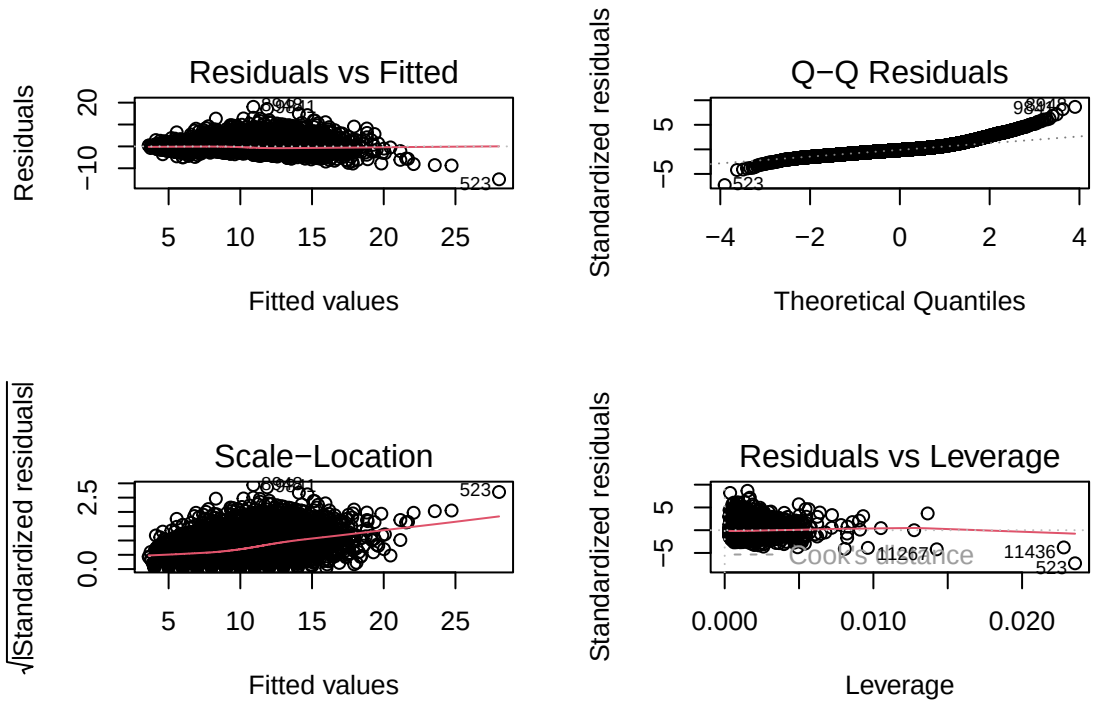
```
##  [1] id             Sex            Diameter       Length         Height
##  [6] Weight         Shucked.Weight Viscera.Weight Shell.Weight   Age
## <0 rows> (or 0-length row.names)
```

Many of the histograms were right skewed but that is to be expected with biological data. In this instance, doing a log transformation made the model more complex and it didn't improve MAE (I ran this on a different R script). The model without a transformation performs well, even with the skewness present.

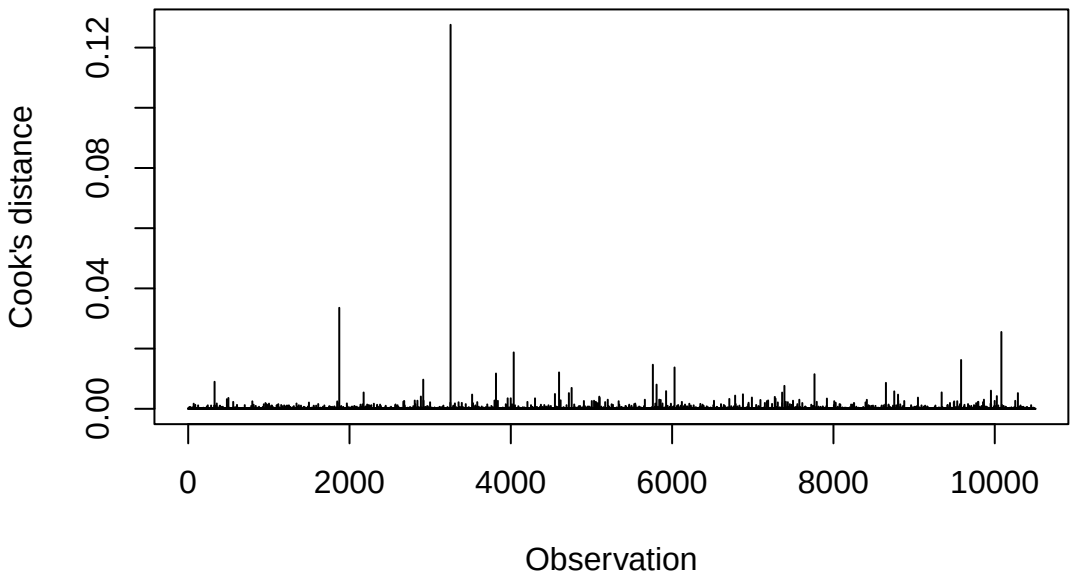
```
set.seed(123)

# Split the data 70/30 for train and test
n <- nrow(abalone)
train_index <- sample(1:n, size = 0.7 * n)

train_data <- abalone[train_index, ]
test_data  <- abalone[-train_index, ]

# This is the base model for the dataset
abalone_model <- lm(Age ~ Sex + Length + Diameter + Height + Weight + Shucked.Weight +
                      Viscera.Weight + Shell.Weight, data = train_data)

summary(abalone_model)
```

```
##
## Call:
## lm(formula = Age ~ Sex + Length + Diameter + Height + Weight +
##     Shucked.Weight + Viscera.Weight + Shell.Weight, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0410  -1.1828  -0.2969   0.7499  18.0920
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.83708    0.20422  18.789   <2e-16 ***
## SexI           -1.08215    0.06830 -15.844   <2e-16 ***
## SexM           -0.11747    0.05023  -2.339   0.0194 *
## Length          0.77709    0.56328   1.380   0.1677
## Diameter        1.09063    0.70454   1.548   0.1217
## Height         10.29232    0.72138  14.268   <2e-16 ***
## Weight          0.24361    0.01458  16.704   <2e-16 ***
## Shucked.Weight -0.67730    0.01755 -38.599   <2e-16 ***
## Viscera.Weight -0.35819    0.03249 -11.026   <2e-16 ***
## Shell.Weight    0.56633    0.02731  20.740   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.093 on 10483 degrees of freedom
## Multiple R-squared:  0.5939, Adjusted R-squared:  0.5935
## F-statistic:  1703 on 9 and 10483 DF,  p-value: < 2.2e-16
```

```r
# To look at the assumptions for the base model
par(mfrow = c(2, 2))
plot(abalone_model)
```




```
par(mfrow = c(1, 1))

plot(cooks.distance(abalone_model),
     type = "h",
     ylab = "Cook's distance",
     xlab = "Observation")
```



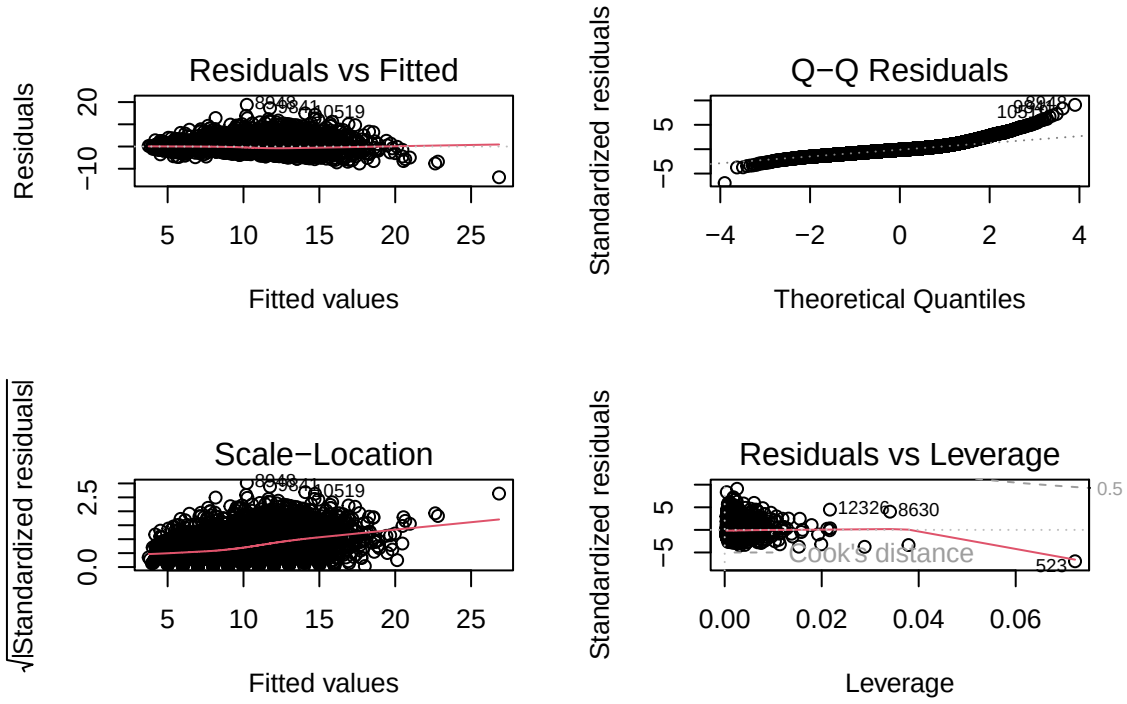
This model shows that every variable is significant except length and diameter.For the next model, I will include an interaction between length and diameter to see if they will be significant together. I will also include other interactions that scientifically make sense. Cook's D shows an influential observation. When I looked into the observation, it was the sex indeterminant. It had some larger values for its features compared to the other indeterminants. It did not appear to be an error and seemed to be something that could happen, so it was kept with the model.

```
# I created another model to compare MAE
# I used interactions between variables that would make sense
abalone_interact <- lm(Age ~ Sex * Weight + Length * Diameter + Weight * Shell.Weight +
                         Height + Shucked.Weight + Viscera.Weight, data = train_data)


# Length multiplied by Diameter gives us the overall shell size
# Weight multiplied by Shell.Weight give us the total mass with what the shell contributes
# Sex multiplied by Weight gives us how Weight differs between male, females, and indeterminant

summary(abalone_interact)
```

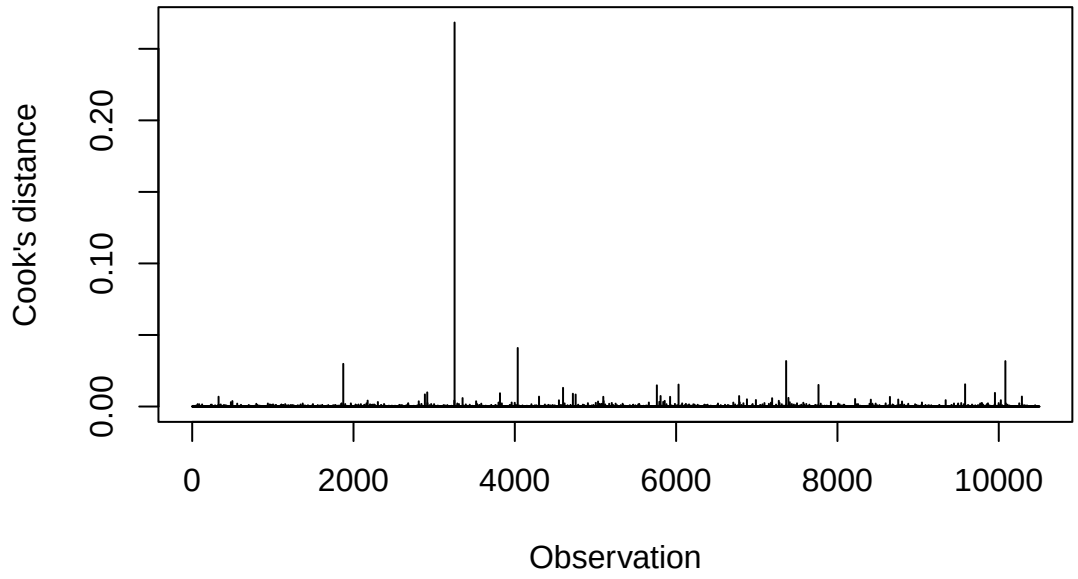
```
##
## Call:
## lm(formula = Age ~ Sex * Weight + Length * Diameter + Weight *
##     Shell.Weight + Height + Shucked.Weight + Viscera.Weight,
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8432  -1.1903  -0.2656   0.7472  18.7806
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.0612211  0.4477420  11.304  < 2e-16 ***
## SexI                   -1.9882109  0.1797768 -11.059  < 2e-16 ***
## SexM                   -0.4990660  0.1505190  -3.316 0.000917 ***
## Weight                  0.3419747  0.0173671  19.691  < 2e-16 ***
## Length                  0.7049759  0.6514462   1.082 0.279202
## Diameter                1.6199027  0.8552420   1.894 0.058241 .
## Shell.Weight            0.8110959  0.0369642  21.943  < 2e-16 ***
## Height                  7.1424310  0.7446471   9.592  < 2e-16 ***
## Shucked.Weight         -0.6525519  0.0175366 -37.211  < 2e-16 ***
## Viscera.Weight         -0.2939757  0.0326022  -9.017  < 2e-16 ***
## SexI:Weight             0.0595811  0.0087703   6.793 1.15e-11 ***
## SexM:Weight             0.0122517  0.0048198   2.542 0.011039 *
## Length:Diameter        -2.7742158  0.4448063  -6.237 4.64e-10 ***
## Weight:Shell.Weight    -0.0053251  0.0007125  -7.474 8.39e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.069 on 10479 degrees of freedom
## Multiple R-squared:  0.6035, Adjusted R-squared:  0.603
## F-statistic:  1227 on 13 and 10479 DF,  p-value: < 2.2e-16
```

```r
# To look at the assumptions for the interaction model
par(mfrow = c(2, 2))
plot(abalone_interact)
```



```r
par(mfrow = c(1, 1))

plot(cooks.distance(abalone_interact),
     type = "h",
     ylab = "Cook's distance",
     xlab = "Observation")
```



This model shows that interaction for variables length and diameter made them much more significant. The other interactions added were also significant, so we continued on with using this model. Cook's D shows an influential observation. When I looked into the observation, it was the sex indeterminant. It had some larger values for its features compared to the other indeterminants. It did not appear to be an error and seemed to be something that could happen, so it was kept with the model.

```r
# To calculate MAE, so we can compare the models
mae <- function(actual, predicted) {
  mean(abs(actual - predicted)) }

# Testing the base and interaction model
pred_abalone <- predict(abalone_model, test_data)
pred_abalone_interact <- predict(abalone_interact, test_data)

# To get the MAE for each model
mae_abalone <- mae(test_data$Age, pred_abalone)
mae_abalone_interact <- mae(test_data$Age, pred_abalone_interact)

# To compare the two models to find which is a better fit
MAE <- data.frame(Model = c("Abalone Base Model", "Abalone Interaction Model"),
                  Test_MAE = c(mae_abalone, mae_abalone_interact))
```

```
MAE
```

```
##                         Model Test_MAE
## 1        Abalone Base Model 1.406604
## 2 Abalone Interaction Model 1.389691
```

The linear regression model with interactions has a lower MAE, so we will proceed with using the interaction model to do predictions for the competition set.

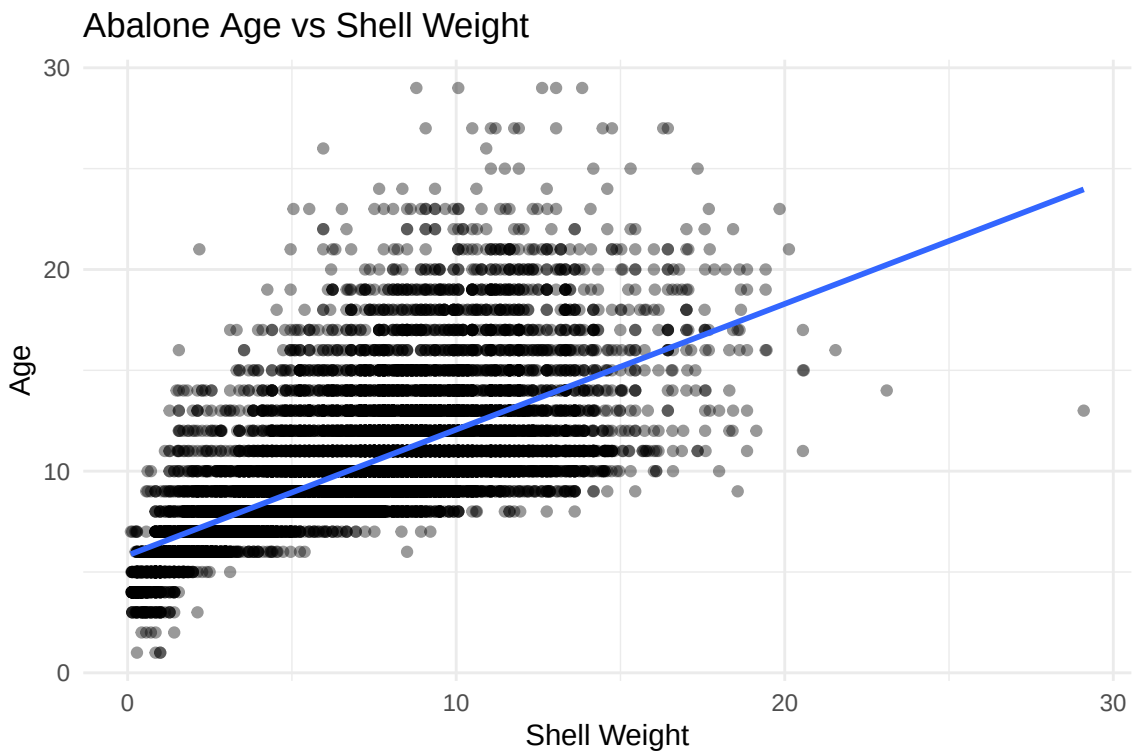
```
# I need to fit the model so that it is now trained on all the
# data to predict for the competition set

final_abalone_interact <- lm(Age ~ Sex * Weight + Length * Diameter + Weight * Shell.Weight
                             + Height + Shucked.Weight + Viscera.Weight, data = abalone)

competition_data <- read.csv("competition.csv")

competition_data$Sex <- as.factor(competition_data$Sex)

# To generate the predictions
competition_predictions <- predict(final_abalone_interact, newdata = competition_data)

# To make sure iD is included in the CSV with Age
submission <- data.frame(iD  = competition_data$id, Age = competition_predictions)

# Writing the predictions in a CSV file
write.csv(submission,"abalone_predictions_project.csv", row.names = FALSE)
```

Objective 2 : Exploratory Analysis

The purpose for this section is to explore which variables of abalone are strongly associated with age. Instead of focusing on prediction accuracy, this section will go into exploratory analysis and interpretability. We are doing this so that we can better understand growth patterns and what helps determine an abalones age.

Age vs Shell.Weight plot

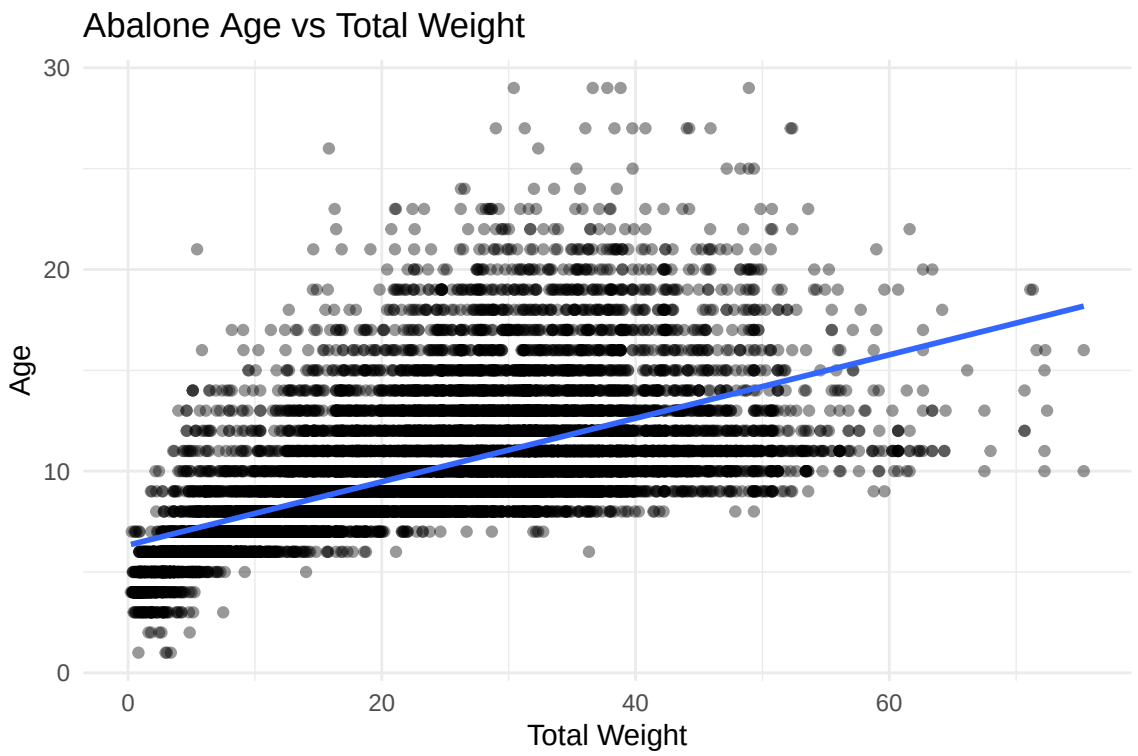
```
ggplot(abalone, aes(x = Shell.Weight, y = Age)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Abalone Age vs Shell Weight",
    x = "Shell Weight",
    y = "Age"
  ) +
  theme_minimal()
```



This scatterplot shows that there is a strong positive relationship between shell weight and age. This indicates that abalones with heavier shells tend to be older. The trend appears to be relatively consistent across the range of shell weights. This overall suggests that shell weight is a good indicator of growth and aging for an abalone.

Age vs Weight plot

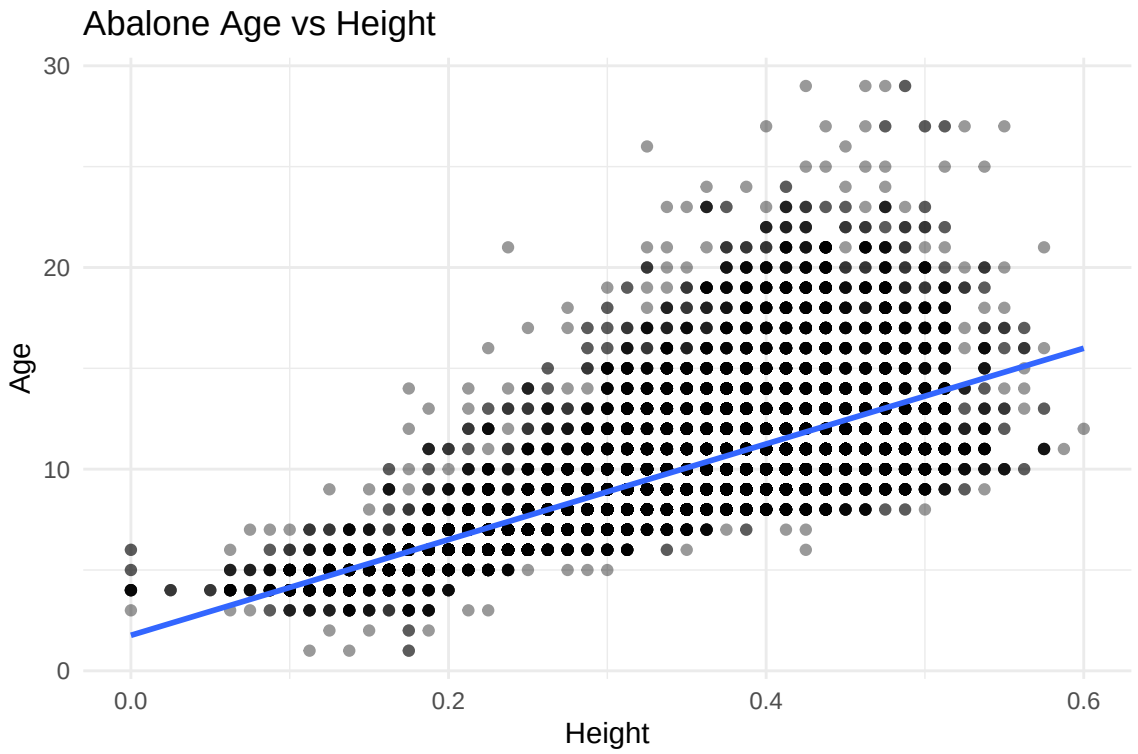
```
ggplot(abalone, aes(x = Weight, y = Age)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Abalone Age vs Total Weight",
    x = "Total Weight",
    y = "Age"
  ) +
  theme_minimal()
```



The scatterplot shows that the relationship between age and weight is positive. It appears that older abalones generally have a higher weight, but there is a considerable spread for weight at every age. The variability that we can see would suggest that weight may be influenced by other biological factors in addition to age.

Age vs Height plot

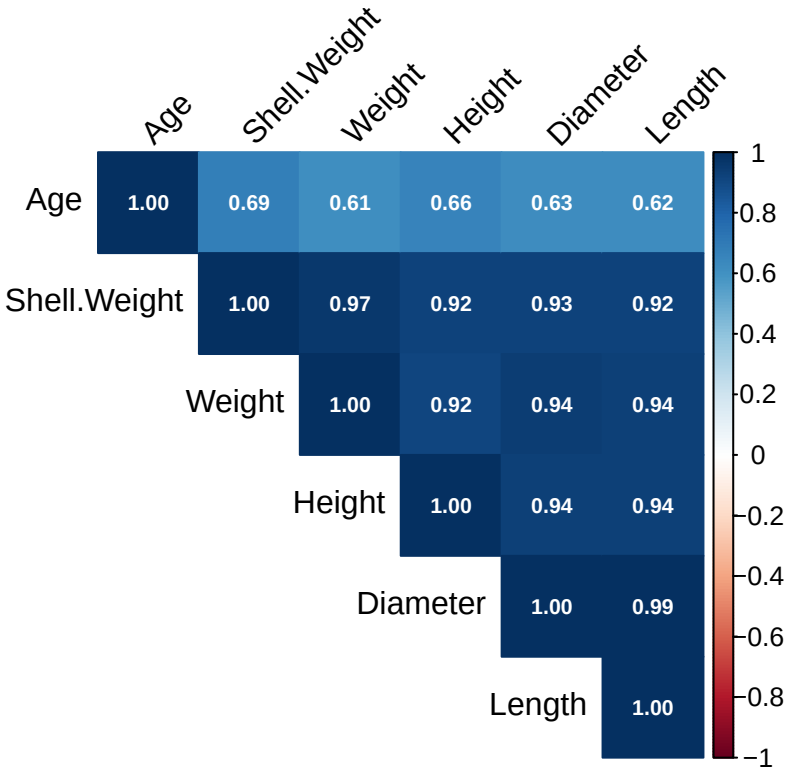
```
ggplot(abalone, aes(x = Height, y = Age)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Abalone Age vs Height",
    x = "Height",
    y = "Age"
  ) +
  theme_minimal()
```



The scatterplot shows that there is a positive relationship between abalone age and height. We can see as height increases, age tends to also increase. There is still some variability, but height appears to associate with growth and aging for an abalone.

Correlation Heat Map for variables

```
abalone_vars <- abalone %>%
  select(Age, Shell.Weight, Weight, Height, Diameter, Length)

cor_matrix <- cor(abalone_vars, use = "complete.obs")

corrplot(
  cor_matrix,
  method = "color",
  type = "upper",
  tl.col = "black",
  tl.srt = 45,
  addCoef.col = "white",
  number.cex = 0.7
)
```



The correlation heat map summarizes the pairwise relationships between age and other important features. Shell weight and height seem to have the strongest correlations with age, which we could see in the scatter plots above. Length and Diameter have moderate associations but not as much as other features. Also, length and diameter appear to highly correlated with other features, which would indicate that there may be overlap between the predictors for age. This would help explain why length and diameter by themselves were not significant in the model by themselves with the other weight variables.

These exploratory plots indicate that abalone age is strongly associated with shell related and mass based features. While length and diameter do increase with age, there strong correlations with other measurements make them not as significant when they are isolated. This supports my decision to include interactions for my model in objective 1. Overall, the findings help highlight the importance of growth and specifically shell development for a predictor of abalone age.

R Shiny App

The Shiny application is available at:
https://chloeprowse.shinyapps.io/abalone-shiny-app/