Fish Market Project

Trinity Miller

2023-03-21

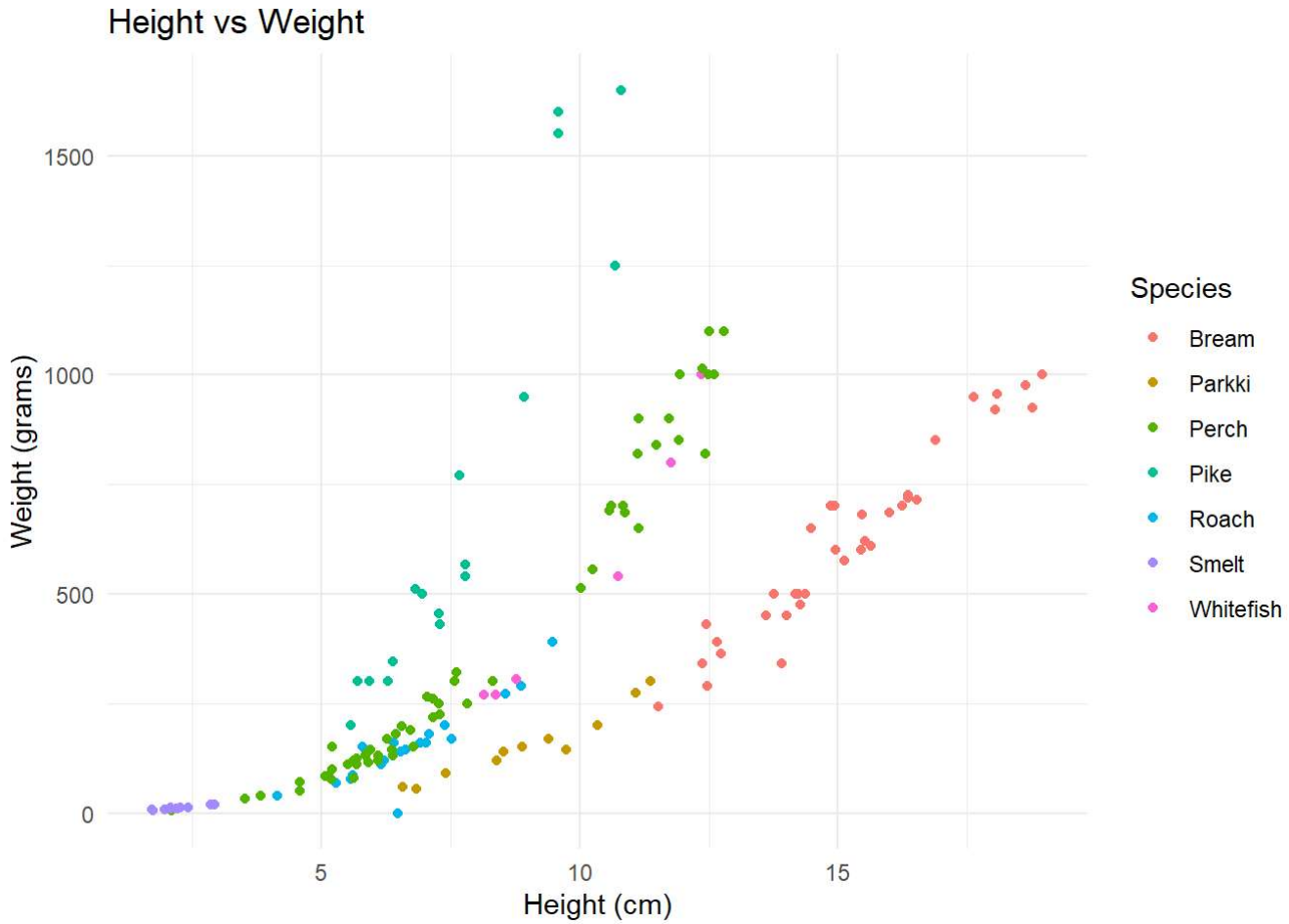
Statement of Purpose:

This data set is a record of 7 common different fish species in fish market sales. The source of this data set does not specify where this is collected from, but says it was uploaded around for years ago. The variables of the data set include the species name of the fish (Species), the weight of the fish in grams (Weight), the vertical length in cm (Length1), diagonal length in cm (Length2), cross length in cm (Length3), height in cm (Height), and diagonal width in cm (Width). My response variable will be the weight of the fish.This information could be helpful for a fisherman that might want to know the weight of a fish they are catching or size in order to use the correct kind of net.

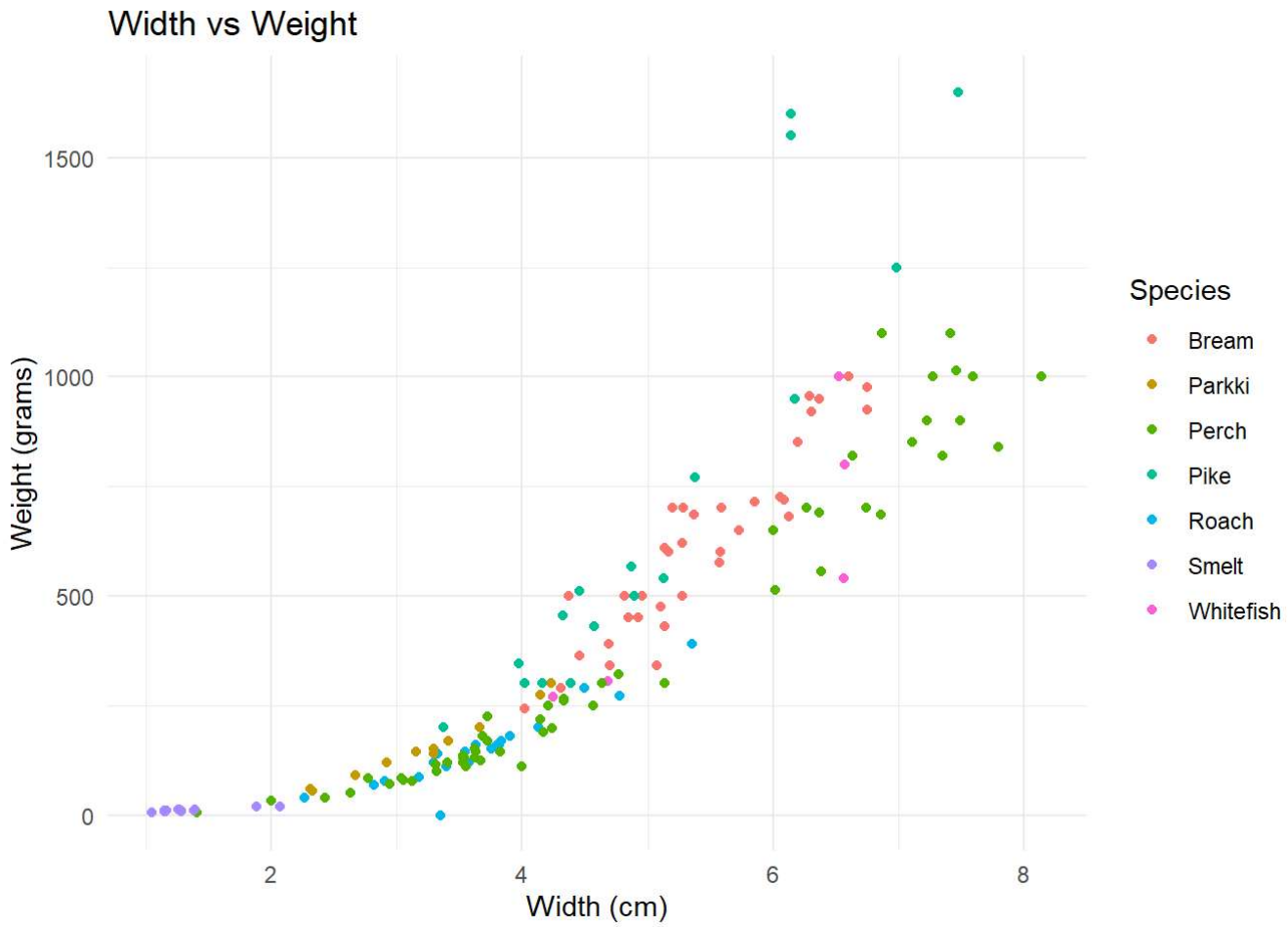
Executive Summary:

I predicted the weight of a fish with the variables included in this data set of 7 common fish species. This model would be useful for a fisherman trying to choose what kind of equipment they would need in order to catch a specific species of fish. I constructed and evaluated three different models depending on how the different variables related to the variable of weight. All of the visualizations included in this report show a polynomial relationship therefore I made my first model using only polynomial terms. This produced the first model, which was also the best model, with the least amount of error. As I tried to make the error lower by making the other two models, it just made the error worse. In my first model, one of the variables was not significant which is not great, so I changed that variable to not be a polynomial variable in the next model. Although this made the variable significant, it made the error higher, which is not better than the first model. Lastly, for my third model, I studied the visualization that compared height of the fish to the weight which shows the data splitting into three different distinct groups. Because of this, I decided to mutate my data to make those three distinct groups and made a model based off of that. All of the variables were significant, but there was even more error than the first and second model. I concluded after making these models, that although I could not lower the error from the first model it was not all that bad. My model predictions are off by about 29.96 grams, with the smallest fish being 0.0 grams and the largest being 1650 grams.

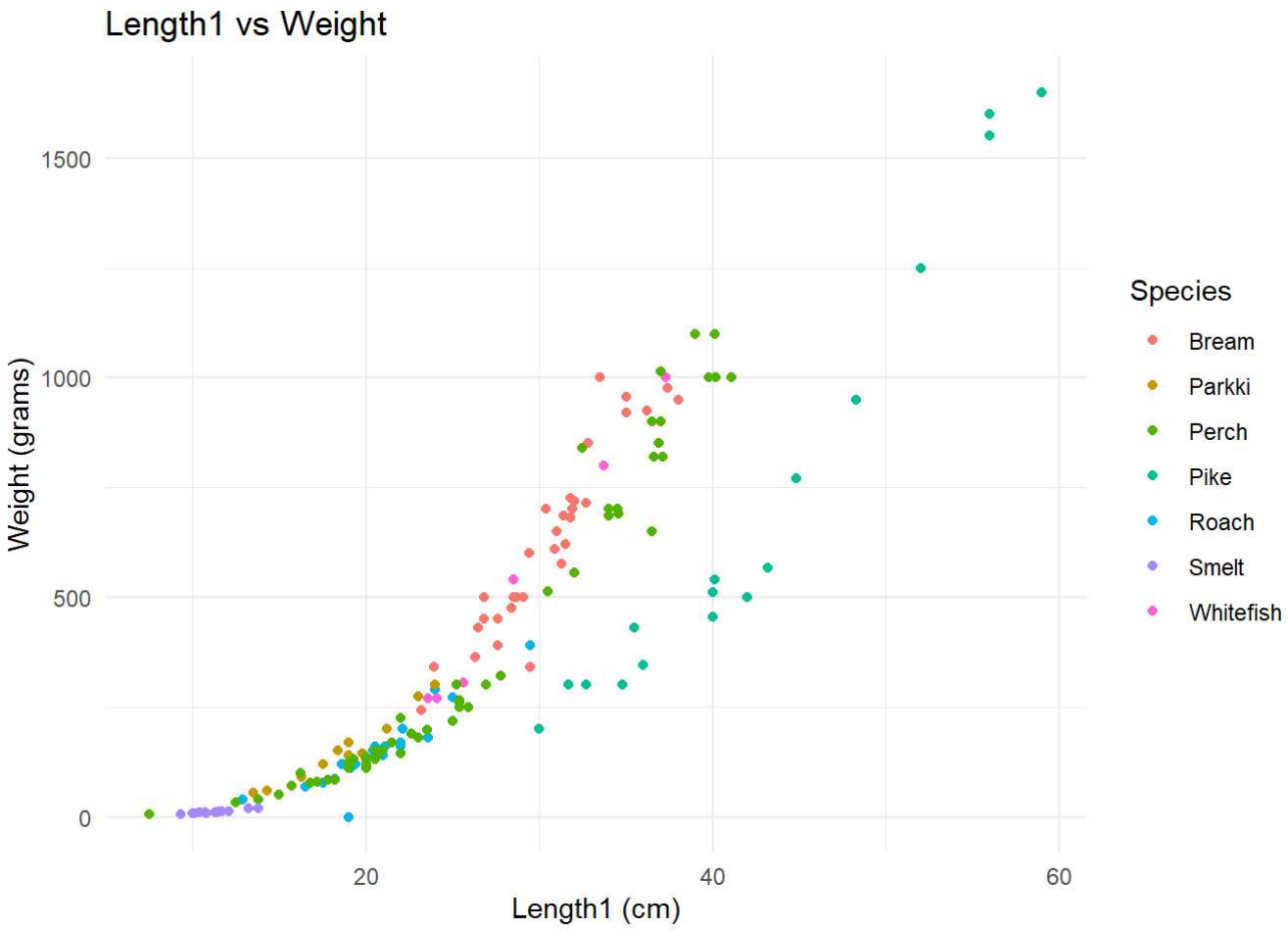
Exploratory Data Analysis:

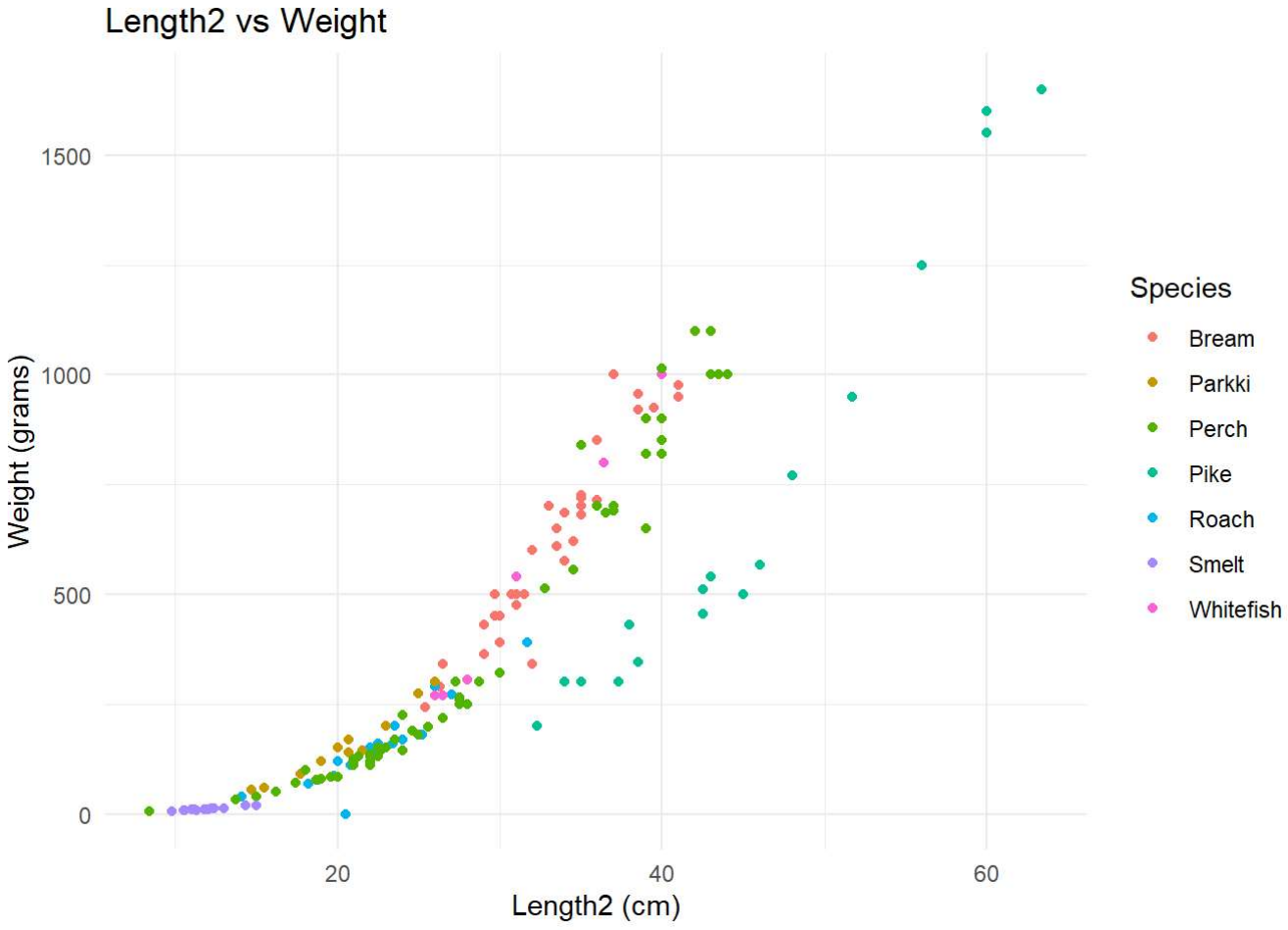


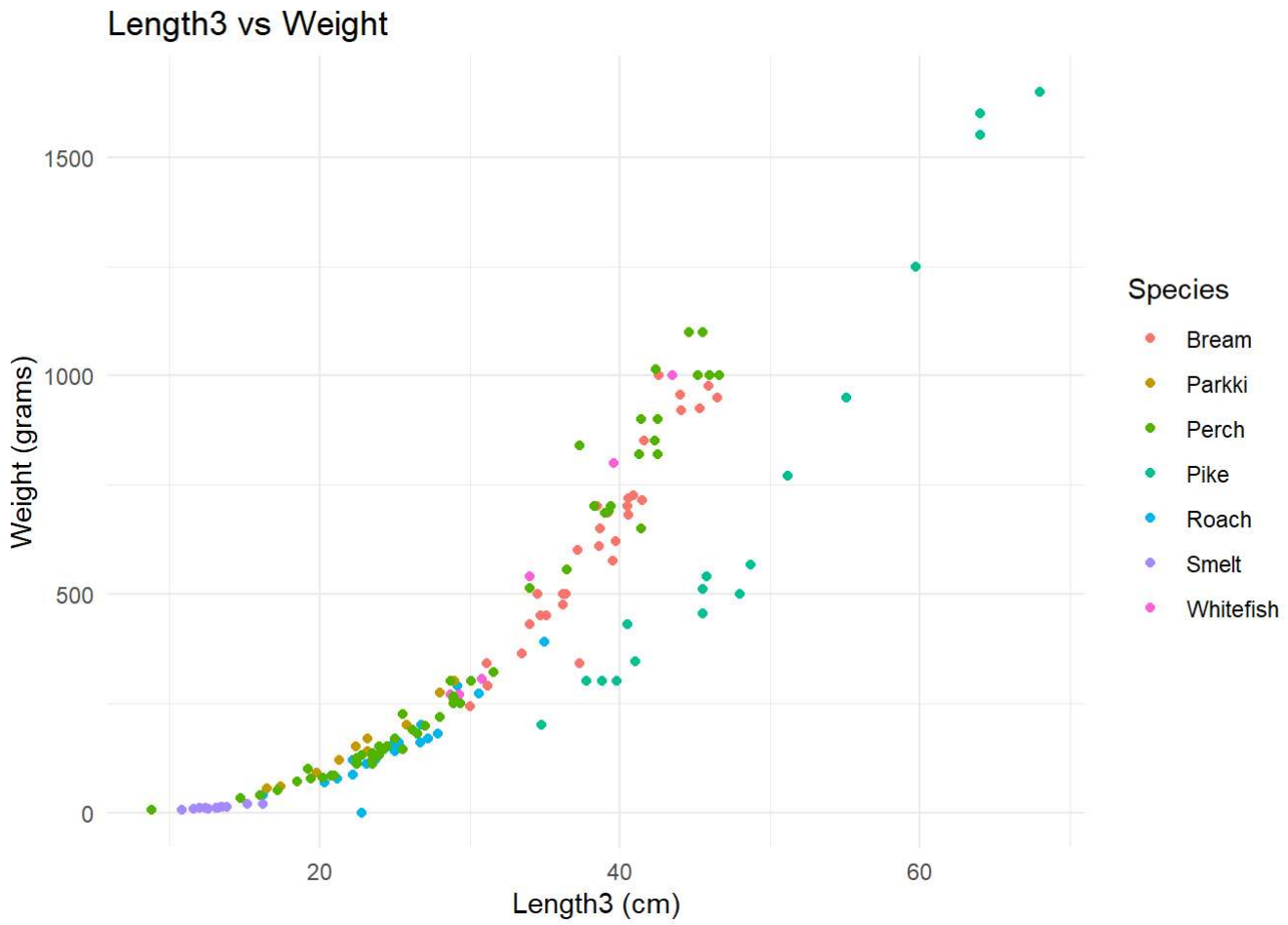
Note: Since there are about 3 different lines made of the points, we can say that some species definitely have the same comparison between weight and height.



Note: Based off this scatter plot, we can see that each species follows the same pattern with width vs weight.


Length1 vs Weight
Weight (grams)
1500
1000
500
0
20
40
60
Length1 (cm)
Species
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish


Length2 vs Weight
Weight (grams)
1500
1000
500
0
20
40
60
Length2 (cm)
Species
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish



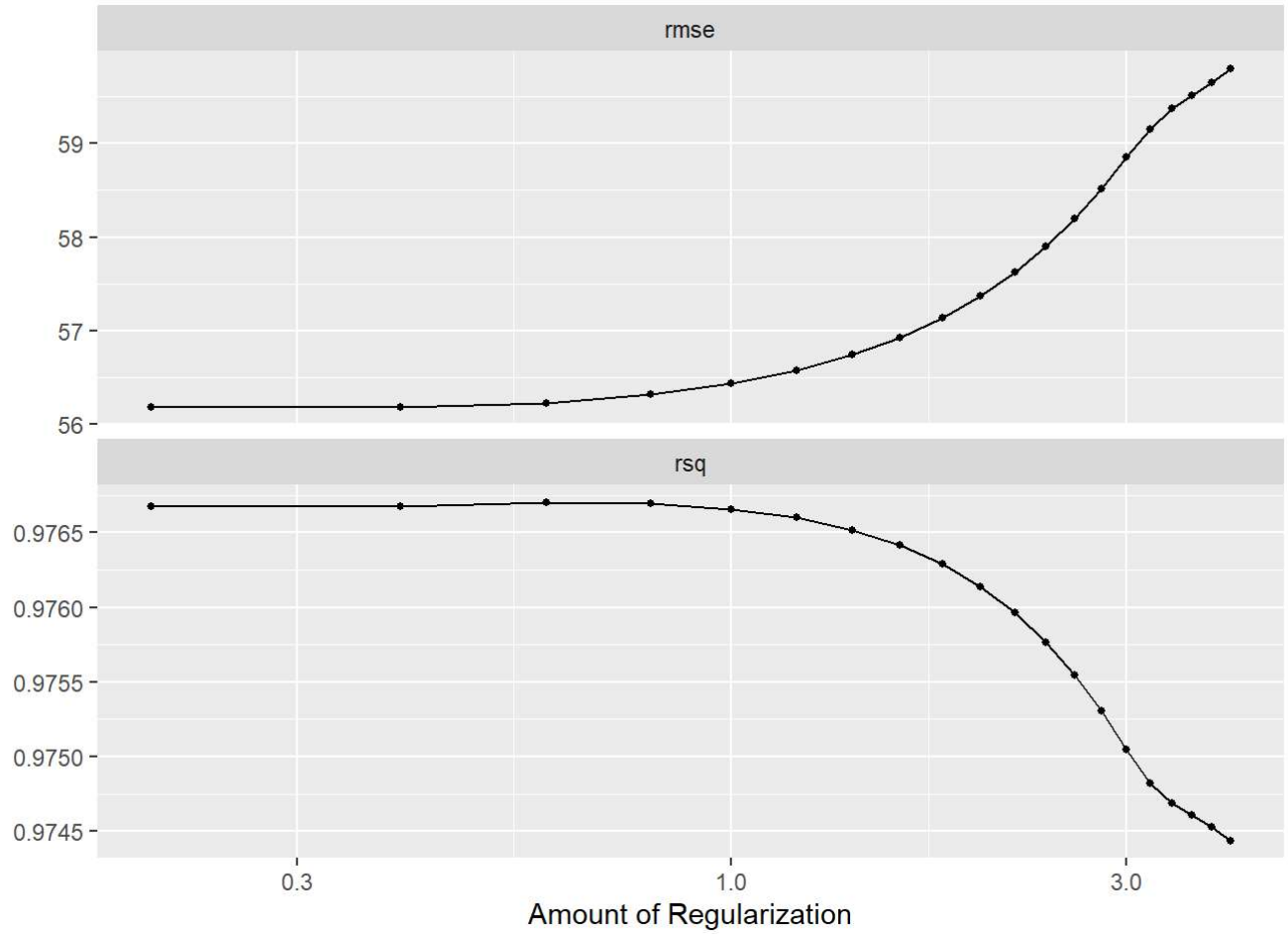
Note: for all the scatter plots of length vs weight, you can see that the all the species beside Pike are about the same.

Model Construction:

Model 1 Construction: Lasso Regression

My first model used Height, Width, Length1, and grouped_species as variables, with Width and Length1 being polynomial regression. I mutated my data to make grouped_species which consists of Pike on it's own, Perch, Whitefish, Roach and Smelt in Group1, and Bream and Parkki in Group2.

Note: Here I did lambda perimeter tuning in order to find the best value of lambda for my first model.



Note: I found the value 0.4 to be the best for lambda in my first model.

```
## # A tibble: 5 × 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1     0.4 rmse    standard    56.2    10    4.57 Preprocessor1_Model02
## 2     0.2 rmse    standard    56.2    10    4.57 Preprocessor1_Model01
## 3     0.6 rmse    standard    56.2    10    4.59 Preprocessor1_Model03
## 4     0.8 rmse    standard    56.3    10    4.60 Preprocessor1_Model04
## 5     1   rmse    standard    56.4    10    4.63 Preprocessor1_Model05
```

Note: This shows the coefficients for my variables in my first model.

```
## # A tibble: 8 × 3
##   term                   estimate penalty
##   <chr>                     <dbl>   <dbl>
## 1 (Intercept)               105.      0.4
## 2 Height                    174.      0.4
## 3 grouped_species_Group2    -44.0     0.4
## 4 grouped_species_Pike      -37.3     0.4
## 5 Width_poly_1              421.      0.4
## 6 Width_poly_2              471.      0.4
## 7 Length1_poly_1           2442.      0.4
## 8 Length1_poly_2           1106.      0.4
```

Note: This shows the RMSE for my first model.

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        33.7
```

Note: This shows the RSQ for my first model.

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.990
```

Model 1 Construction: Least Squares Regression

This model is a least squares model using Height, Width, Length1, and grouped_species. Width and Length1 are polynomial regression.

Note: This shows the RMSE and RSQ for my model

```
## # A tibble: 2 × 6
##   .metric .estimator   mean     n std_err .config
##   <chr>   <chr>       <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   56.1      10 4.56    Preprocessor1_Model1
## 2 rsq     standard    0.977    10 0.00508 Preprocessor1_Model1
```

Note: Shows the coeffients for my variables and their p-values.

```
## # A tibble: 1 × 12
##   r.squared adj.r.squa…¹ sigma stati…²  p.value    df logLik   AIC   BIC devia…³
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.981        0.979  53.5    800. 7.25e-92     7  -638. 1294. 1319. 317124.
## # … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names ¹adj.r.squared, ²statistic, ³deviance
```

```
## # A tibble: 8 × 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                83.2      56.2      1.48 1.42e- 1
## 2 Height                    187.       30.5      6.15 1.27e- 8
## 3 grouped_species_Group2    -52.1      17.1     -3.05 2.86e- 3
## 4 grouped_species_Pike      -38.0      15.3     -2.48 1.46e- 2
## 5 Width_poly_1              323.      297.       1.09 2.80e- 1
## 6 Width_poly_2              466.       78.0      5.97 2.92e- 8
## 7 Length1_poly_1           2453.      356.       6.89 3.55e-10
## 8 Length1_poly_2           1122.      110.      10.2  1.29e-17
```

Model 2 Construction: Least Squares Regression

This model is a least squares model using Height, Width, and Length1 as variables, with Width and Length1 being polynomial regression.

Note: Shows the RMSE and RSQ for my second model.

```
## # A tibble: 2 × 6
##   .metric .estimator   mean     n std_err .config
##   <chr>   <chr>       <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   61.8      10 5.38    Preprocessor1_Model1
## 2 rsq     standard    0.973    10 0.00592 Preprocessor1_Model1
```

Note: Shows the coefficients for my variables and their p-values from my second model.

```
## # A tibble: 1 × 12
##   r.squared adj.r.squa…¹ sigma stati…²  p.value    df logLik   AIC   BIC devia…³
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.976        0.975  58.6    928. 5.73e-90     5  -650. 1314. 1334. 388133.
## # … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names ¹adj.r.squared, ²statistic, ³deviance
```

```
## # A tibble: 6 × 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        177.     20.3       8.71 2.94e-14
## 2 Height             119.      9.55     12.5  5.80e-23
## 3 Width_poly_1      1460.    204.        7.16 8.54e-11
## 4 Width_poly_2       503.     84.3       5.97 2.82e- 8
## 5 Length1_poly_1    1653.    170.        9.75 1.17e-16
## 6 Length1_poly_2    1170.    108.       10.9  2.68e-19
```

Model 3 Construction: Least Squares Regression

This model is a least squares model using grouped_species, Height, Width, and Length1 as variables with Length1 being polynomial regression.

Note: Shows the RMSE and RSQ for my third model.

```
## # A tibble: 2 × 6
##   .metric .estimator   mean     n std_err .config
##   <chr>   <chr>       <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   64.5      10 4.23    Preprocessor1_Model1
## 2 rsq     standard    0.971    10 0.00457 Preprocessor1_Model1
```

Note: Shows the coefficiants for my variables and their p-values for my third model.

```
## # A tibble: 1 × 12
##   r.squared adj.r.squa…¹ sigma stati…²  p.value    df logLik   AIC   BIC devia…³
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.974        0.973  61.2    709. 1.31e-86     6  -655. 1325. 1348. 418905.
## # … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names ¹adj.r.squared, ²statistic, ³deviance
```

```
## # A tibble: 7 × 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)              -161.       86.2     -1.86 6.49e- 2
## 2 Height                    213.       34.5      6.17 1.10e- 8
## 3 Width                      76.3      30.0      2.54 1.23e- 2
## 4 grouped_species_Group2    -69.6      19.3     -3.61 4.55e- 4
## 5 grouped_species_Pike      -31.1      17.5     -1.78 7.81e- 2
## 6 Length1_poly_1           1798.      388.       4.64 9.53e- 6
## 7 Length1_poly_2           1560.       93.8     16.6  5.02e-32
```

Model Interpretation and Inference:

By constructing several models, I found that there are some variables that seemed to have no effect on weight. These variables were Length2 and Length3. Although these had to significance when it came to predicting weight, it was still important to try and use them in the models. In the future, we could possibly look into trying many interaction terms to see if they show any kind of significance that way. On the other hand, the variables Height, Length1, Species, and Width seemed to definitely improve all my models in some way. Due to the polynomial regression shown in all of the scatter plots in the exploratory data section, I made all of the significant terms in my best model polynomial variables first. Then after seeing the p values, I noticed that Height was not significant with being polynomial so I changed them back to being regular linear regression. Also in the scatter plot of Weight vs Height, you can notice that the species group off into 3 distinct groups, which is the reason why I mutated my data to have different groups of species based off of that graph. By doing this, my RMSE went done which shows there is less error in my model. The formula for the model is Width = 105.4 + 174 * Height + 44 * Group2 - 37.3 * Pike + 421.1 * Width + 470.5 * Width^2 + 2442.4 * Length1 + 1106.3 * Length1^2. When I created the recipe for this formula, I scaled all the predictors so that I could interpret which variables had the most influence on weight. This model shows that Length1 had the most influence on predicting weight, followed by Length1^2 because they have the largest coefficients. All the variables in my first model were significant except for Width, with there p-values being Height = 1.3e^-8, Group2 = 2.9 e^-3, Pike = 1.5e^-2, Width = 2.8e^-1, Width^2= 2.9e^-8, Length1 = 3.5e^-10, Length1^2 = 1.3e^-17.

Conclusion:

In conclusion, I made a lot of interesting discoveries while make several models. The one thing I found is that although some of my p-values were not significant (< .05) the variables still made a good contribution to my models by making the RMSE lower and the RSQ higher. I also found that by doing my exploratory data analysis, that when the variables are compared to each other through scatter plots they seem to show a polynomial regression. When I made my first model, I made them all polynomial regressions at first, but the p-values became insignificant. So although they looked polynomial on the scatter plot, it actually made my model worse. When I changed them back to linear regression, my p-values became significant again and my RMSE went down and my

RSQ went up. Lastly, I discovered that by grouping off the species, I was able to improve my model. By leaving the species all on their own, I was over fitting the training data. This was a result of including all the variables as polynomial regression.