STAT FINAL PROJECT REPORT

Analysis of Abalone dataset

Team Members:

Dhavani Avu
Pragathi Tummala
Shreya Nagaram

Professor:

Dr. Isuru Dassanayake

May 11th, 2023

ABSTRACT:

The Abalone dataset is a collection of measurements obtained from abalone, a type of marine snail. The data set contains a variety of physical measurements, including weights, heights, diameters, and lengths of various abalone components. The dataset consists of 4177 instances, with each instance representing a single abalone. It includes both numerical and categorical attributes, such as sex and the number of rings. The sex of the abalone is classified as male, female, or infant (coded as M, F, and I, respectively). The dataset presents intriguing challenges because of the diversity of abalone properties and the requirement to manage both numerical and categorical information. It has been used to investigate various machine learning algorithms, such as clustering methods, random forests, and decision trees. Researchers and professionals interested in marine biology, pattern recognition, and predictive modeling can benefit greatly from the Abalone dataset.

INTRODUCTION:

The Abalone dataset is widely used in machine learning and data analysis to understand the biology and ecology of abalone, a species of marine snail. The dataset comprises physical measurements of abalone specimens, including length, diameter, height, and weight, and a categorical attribute indicating their sex. The age of abalone can be determined by counting the rings in their shell. The dataset is a valuable tool for researchers to estimate age and sex, experiment with clustering algorithms, and explore research applications in marine biology, pattern recognition, and machine learning.
This project uses the Abalone dataset to answer research questions through visualizations and statistical models such as regression models, decision trees, clustering models, and various hypotheses tests.

RESEARCH QUESTIONS:

1) How well can we predict the age of an abalone based on its physical measurements?
2) Is there any significance between age and sex?
3) Is there any significant relation between gender of the abalone and other physical parameters?
4) Identify distinct groups of abalone based on similar characteristics?

DATA:

The Abalone dataset, sourced from the UCI Machine Learning Repository, contains information on the physical characteristics of abalone such as sex, height, length, diameter, and weight measurements, including whole weight, shucked weight, viscera weight, shell weight, and rings which represent age. Preprocessing was conducted to maintain data integrity, including checking for missing values and ensuring consistent data types.

MATERIALS AND METHODS:

LINEAR REGRESSION:

This model is used to answer the first research question. MLR stands for Multiple Linear Regression, which is a statistical technique used to model the relationship between a dependent variable (Y) and two or more independent variables (X1, X2, X3, ..., Xn). In our case we considered Rings as the dependent variable and other physical parameters as independent variables. We have used multiple linear models by including variable selection, interaction terms and fit best regression model.

POLYNOMIAL REGRESSION:

Polynomial regression models the relationship between the independent variable and the dependent variable as an nth degree polynomial, allowing for non-linear relationships to be captured. In R, this can be performed using the lm() function and the poly() function to create polynomial terms. The 1st and 2nd models of polynomial regression create terms of degree 1, 2, and 3 for each independent variable, with the second model removing some variables for more significant results.

PRINCIPAL COMPONENT ANALYSIS:

PCA (Principal Component Analysis) is a statistical method used to reduce the dimensionality of a dataset while retaining as much of the variation in the data as possible. We can perform a Principal Component Analysis (PCA) to transform the correlated variables into a set of uncorrelated principal components. Then, we can use these principal components as predictors in a linear regression model. In our case we have used PCA for a model with high multicollinearity between the predictor variables and obtained a model which is with good accuracy than the previous model.

DECISION TREE:

Decision tree is deployed to answer the second part of the second research question. In order to determine if the physical measurements of an abalone can be used to classify its sex, we employ a decision tree (refer to Figure 2.5). The decision tree consists of multiple nodes, each representing a splitting criterion, a specific feature or attribute, and a class or category. The tree includes a root node and four-leaf nodes, which provide the final classification outcomes.

RANDOM FOREST:

We trained a regression tree model (Fig 2.6 to 2.8) using the training data to make predictions, where the target variable was the 'rings' column and the other variables served as predictors. We evaluated the model's performance by calculating the Mean Squared

Error and conducted cross-validation. Additionally, we determined the importance of variables using this model, revealing that shell weight is the most significant physical feature for predicting abalone characteristics.

CHI- SQUARE TEST, MEAN SQUARED ERROR:

We performed chi-square tests on categorical variables representing age and sex to investigate if there is a significant relationship between them. Two tests were conducted: one including infants and one excluding them. The results indicated a highly significant association between age and sex in both cases. This suggests that there is a significant relationship between age and sex, regardless of the inclusion or exclusion of infants. In regression tree modeling, mean squared error (MSE) is calculated before and after pruning to assess if there is an improvement. The value of MSE remains the same before and after pruning in the decision tree model. A lower value of MSE indicates a better model with an optimal number of trees.

RESULTS:

Solution for Q1:

To address the research question and develop a prediction model, regression analysis, regression tree, and decision tree methods can be utilized. Initially, a correlation plot was generated to examine the relationships among variables. As the dataset describes abalone age as 1.5+ Rings, we adjusted the Rings column accordingly. For the current analysis, it is advisable to exclude the 'Sex' variable from the model due to its categorical nature and lack of continuity or physical measurement.

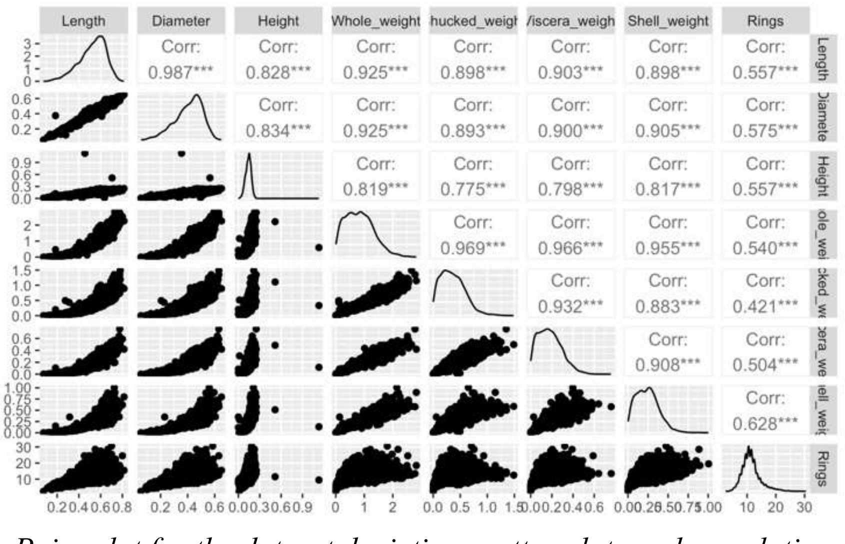


Fig 1.1. Pairs plot for the dataset depicting scatter plots and correlation.

Several variables exhibit significant linear relationships, such as length and diameter, and whole weight and shucked weight. Moreover, certain variables, including rings, demonstrate skewed distributions. To visualize the correlation matrix, we will generate a plot. Darker plots in the figure indicate stronger correlations between variables.

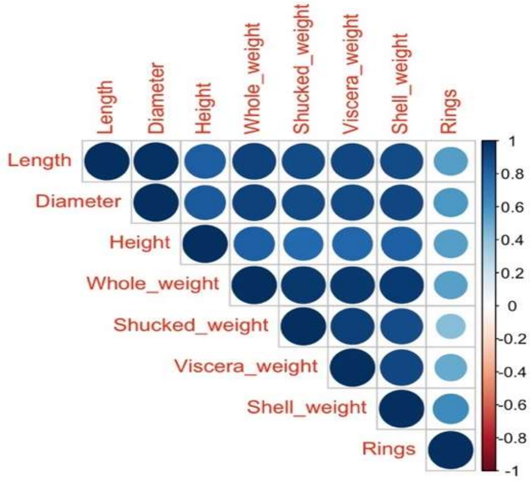


Fig 1.2. Correlation Matrix

To examine the relationship between the predictor variables and the response variable, a scatter plot was created. The scatter plot reveals that, with the exception of height, all other variables display a linear relationship with no outliers.

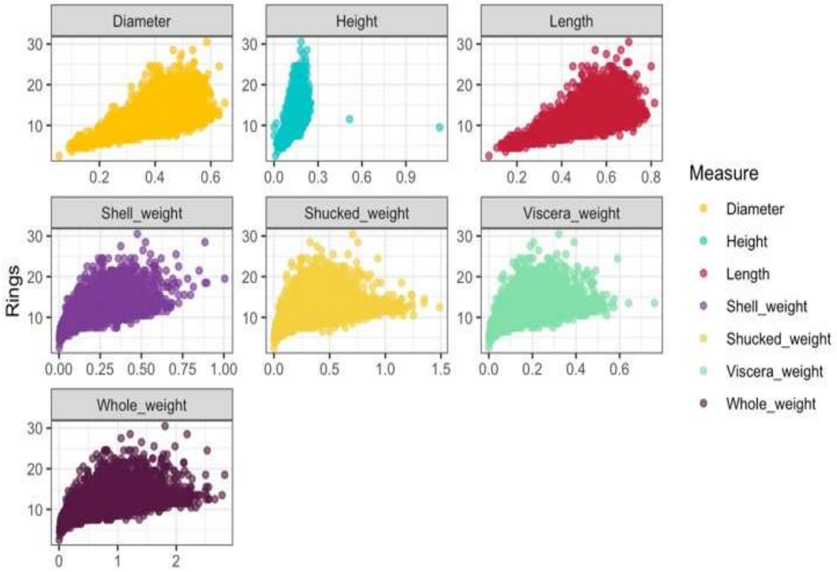
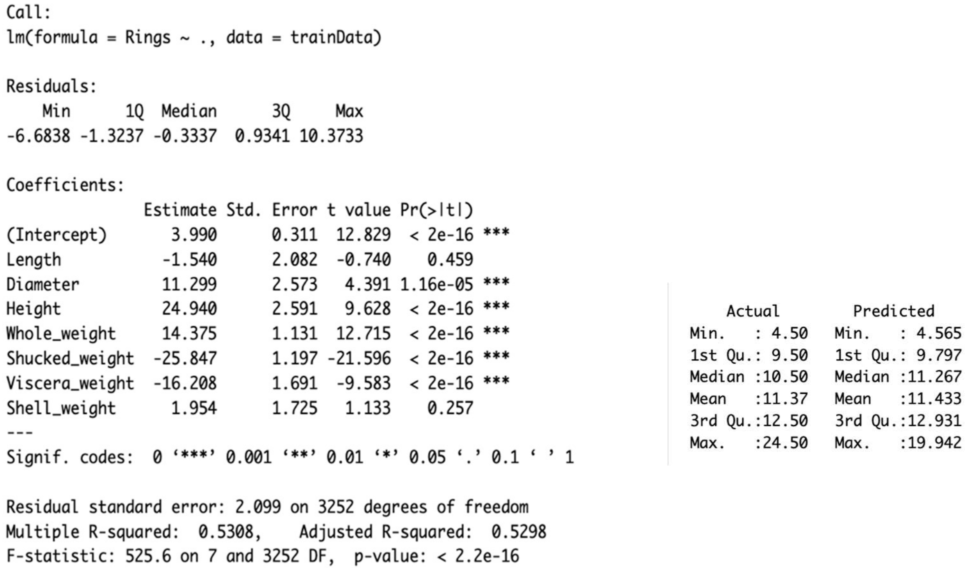


Fig 1.3. Scatter plots for each physical measurement of the abalone.

Subsequently, the dataset was cleansed by removing outliers. Additionally, the cleaned training set was examined for any missing values, and none were found. Following this, the data was divided into a training set (80%) and a test set (20%). A model was then constructed using the training set as outlined below.



```
Call:
lm(formula = Rings ~ ., data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6838 -1.3237 -0.3337  0.9341 10.3733 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.990      0.311  12.829  < 2e-16 ***
Length           -1.540      2.082  -0.740    0.459    
Diameter         11.299      2.573   4.391 1.16e-05 ***
Height           24.940      2.591   9.628  < 2e-16 ***
Whole_weight     14.375      1.131  12.715  < 2e-16 ***
Shucked_weight  -25.847      1.197 -21.596  < 2e-16 ***
Viscera_weight  -16.208      1.691  -9.583  < 2e-16 ***
Shell_weight      1.954      1.725   1.133    0.257    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.099 on 3252 degrees of freedom
Multiple R-squared:  0.5308,	Adjusted R-squared:  0.5298 
F-statistic: 525.6 on 7 and 3252 DF,  p-value: < 2.2e-16
```

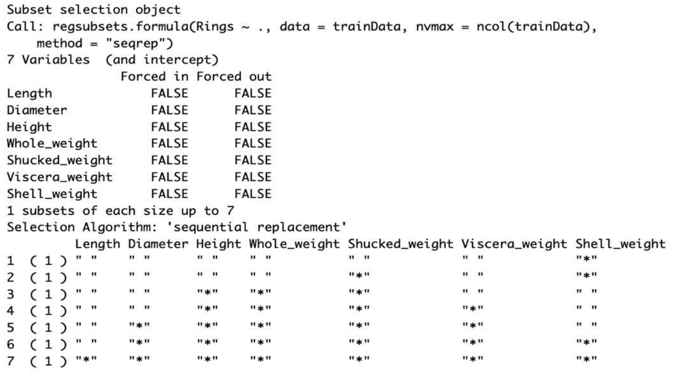
```
     Actual         Predicted     
 Min.   : 4.50   Min.   : 4.565  
 1st Qu.: 9.50   1st Qu.: 9.797  
 Median :10.50   Median :11.267  
 Mean   :11.37   Mean   :11.433  
 3rd Qu.:12.50   3rd Qu.:12.931  
 Max.   :24.50   Max.   :19.942  
```

Fig 1.4. Summary of regression model 1

During this stage, we applied a multiple linear regression (MLR) model to the training data using the lm function from the stats package. The associated p-value of the F-statistic was found to be less than 2.2e-16, indicating a significant relationship between at least one predictor and the response variable.

Subsequently, we calculated the difference between the actual and predicted values. The results indicate that the actual and predicted values are relatively close, suggesting the significance of this model. However, to enhance its performance, we can consider eliminating insignificant variables.

To achieve this, we employed a hybrid selection technique and constructed a model. Furthermore, we generated a plot (Fig 1.6) to determine the optimal number of variables to include in the model.



```
Subset selection object
Call: regsubsets.formula(Rings ~ ., data = trainData, nvmax = ncol(trainData), 
    method = "seqrep")
7 Variables  (and intercept)
               Forced in Forced out
Length             FALSE      FALSE
Diameter           FALSE      FALSE
Height             FALSE      FALSE
Whole_weight       FALSE      FALSE
Shucked_weight     FALSE      FALSE
Viscera_weight     FALSE      FALSE
Shell_weight       FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: 'sequential replacement'
         Length Diameter Height Whole_weight Shucked_weight Viscera_weight Shell_weight
1  ( 1 ) " "    " "      " "    " "          " "            " "            "*"         
2  ( 1 ) " "    " "      " "    " "          "*"            " "            "*"         
3  ( 1 ) " "    " "      "*"    "*"          "*"            " "            " "         
4  ( 1 ) " "    " "      "*"    "*"          "*"            "*"            " "         
5  ( 1 ) " "    "*"      "*"    "*"          "*"            "*"            " "         
6  ( 1 ) " "    "*"      "*"    "*"          "*"            "*"            "*"         
7  ( 1 ) "*"    "*"      "*"    "*"          "*"            "*"            "*"         
```

Fig 1.6. Summary of regression model 1.

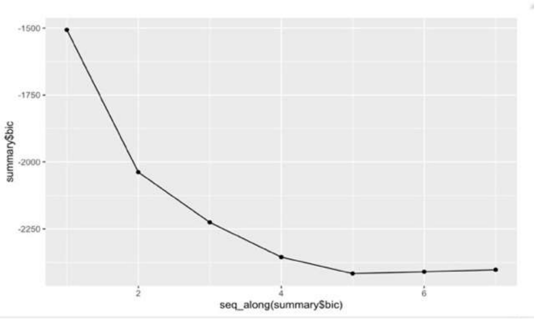
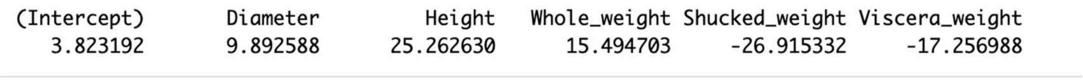


Fig:1.7

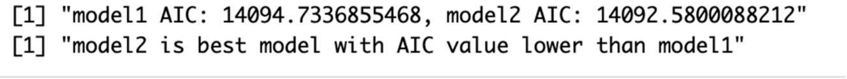
According to the plot analysis, the point where the Bayesian Information Criterion (BIC) is the lowest is at 5 variables. Hence, we will select these 5 variables as they offer the most optimal model fit based on the BIC criterion.



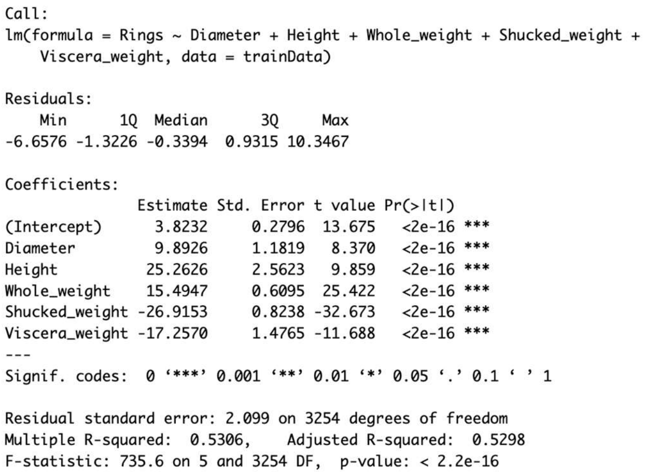
```
 (Intercept)       Diameter         Height   Whole_weight Shucked_weight Viscera_weight 
    3.823192       9.892588      25.262630      15.494703     -26.915332     -17.256988 
```

Fig 1.8.

After identifying the five variables (Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight) as the most influential based on the plot, a regression model was constructed using only these variables. Upon examining the results, it is apparent that the outcome remains unchanged. However, the insignificant variables have been eliminated from the model.



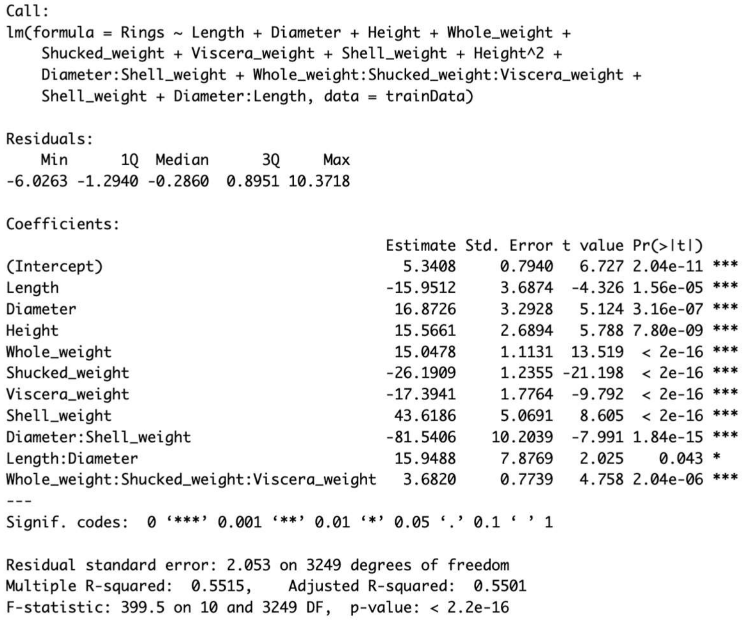
```
[1] "model1 AIC: 14094.7336855468, model2 AIC: 14092.5800088212"
[1] "model2 is best model with AIC value lower than model1"
```



```
Call:
lm(formula = Rings ~ Diameter + Height + Whole_weight + Shucked_weight + 
    Viscera_weight, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6576 -1.3226 -0.3394  0.9315 10.3467 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.8232     0.2796  13.675   <2e-16 ***
Diameter         9.8926     1.1819   8.370   <2e-16 ***
Height          25.2626     2.5623   9.859   <2e-16 ***
Whole_weight    15.4947     0.6095  25.422   <2e-16 ***
Shucked_weight -26.9153     0.8238 -32.673   <2e-16 ***
Viscera_weight -17.2570     1.4765 -11.688   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.099 on 3254 degrees of freedom
Multiple R-squared:  0.5306,	Adjusted R-squared:  0.5298 
F-statistic: 735.6 on 5 and 3254 DF,  p-value: < 2.2e-16
```

Fig 1.9. Model-2

In comparison to the previous model, the revised model (model2) yielded slightly improved results. The removal of insignificant variables contributed to this enhancement. However, to confirm which model is superior, we can utilize the Akaike Information Criterion (AIC), where a lower value indicates better performance. As observed in the output below, model2 exhibits a lower AIC value, indicating its superiority over the initial model.



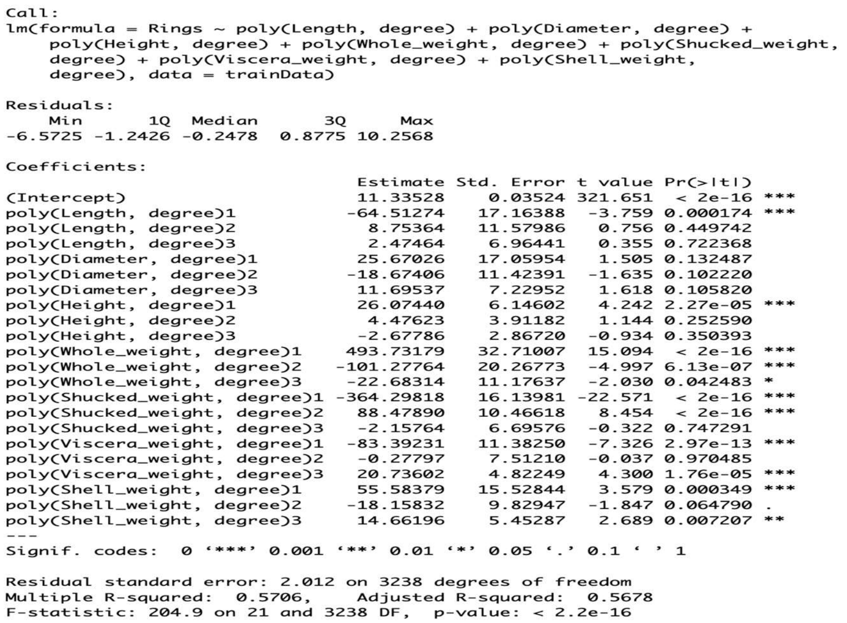
```
Call:
lm(formula = Rings ~ Length + Diameter + Height + Whole_weight +
    Shucked_weight + Viscera_weight + Shell_weight + Height^2 +
    Diameter:Shell_weight + Whole_weight:Shucked_weight:Viscera_weight +
    Shell_weight + Diameter:Length, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0263 -1.2940 -0.2860  0.8951 10.3718

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  5.3408     0.7940   6.727 2.04e-11 ***
Length                                     -15.9512     3.6874  -4.326 1.56e-05 ***
Diameter                                    16.8726     3.2928   5.124 3.16e-07 ***
Height                                      15.5661     2.6894   5.788 7.80e-09 ***
Whole_weight                                15.0478     1.1131  13.519  < 2e-16 ***
Shucked_weight                             -26.1909     1.2355 -21.198  < 2e-16 ***
Viscera_weight                             -17.3941     1.7764  -9.792  < 2e-16 ***
Shell_weight                                43.6186     5.0691   8.605  < 2e-16 ***
Diameter:Shell_weight                      -81.5406    10.2039  -7.991 1.84e-15 ***
Length:Diameter                             15.9488     7.8769   2.025    0.043 *
Whole_weight:Shucked_weight:Viscera_weight   3.6820     0.7739   4.758 2.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.053 on 3249 degrees of freedom
Multiple R-squared:  0.5515,    Adjusted R-squared:  0.5501
F-statistic: 399.5 on 10 and 3249 DF,  p-value: < 2.2e-16
```

Fig 1.10. Model-3

In the model, the interaction term between "Diameter" and "Shell_weight" was included, as well as the interaction between "Diameter" and each of "Whole_weight", "Shucked_weight", "Viscera_weight", and "Shell_weight". These interactions aim to capture the combined effects of these variables on the response variable "Rings". Additionally, the quadratic term of "Height" was incorporated to account for any potential nonlinear relationship between "Height" and "Rings".

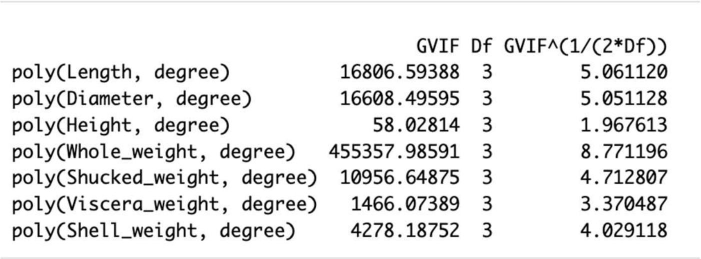
The performance of this model surpasses that of the previous one. We assessed the accuracy of the model by comparing the actual and predicted values, which yielded favorable results.

Now, let's proceed with polynomial regression. In the following results, we can observe some non-significant variables. To address this issue, we will employ the VIF (Variance Inflation Factor) method and remove those variables with high VIF numbers. Subsequently, we will build another model, as depicted in Figure 1.12.



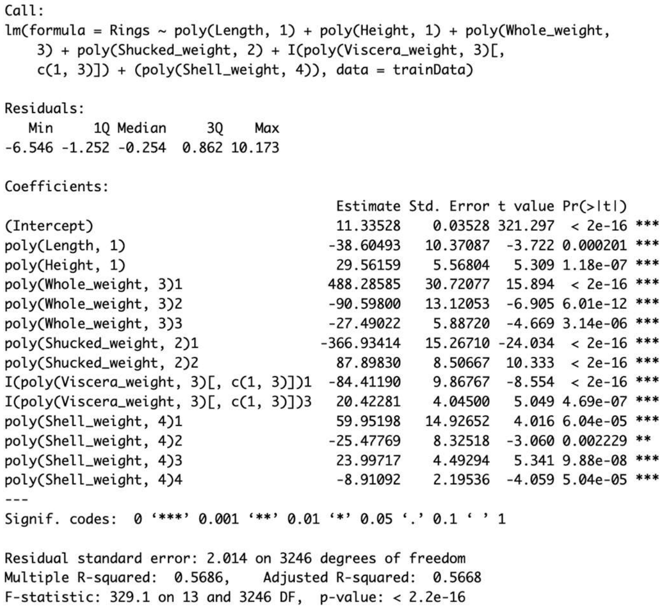
```
Call:
lm(formula = Rings ~ poly(Length, degree) + poly(Diameter, degree) +
    poly(Height, degree) + poly(Whole_weight, degree) + poly(Shucked_weight,
    degree) + poly(Viscera_weight, degree) + poly(Shell_weight,
    degree), data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5725 -1.2426 -0.2478  0.8775 10.2568

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     11.33528    0.03524 321.651  < 2e-16 ***
poly(Length, degree)1          -64.51274   17.16388  -3.759 0.000174 ***
poly(Length, degree)2            8.75364   11.57986   0.756 0.449742
poly(Length, degree)3            2.47464    6.96441   0.355 0.722368
poly(Diameter, degree)1         25.67026   17.05954   1.505 0.132487
poly(Diameter, degree)2        -18.67406   11.42391  -1.635 0.102220
poly(Diameter, degree)3         11.69537    7.22952   1.618 0.105820
poly(Height, degree)1           26.07440    6.14602   4.242 2.27e-05 ***
poly(Height, degree)2            4.47623    3.91182   1.144 0.252590
poly(Height, degree)3           -2.67786    2.86720  -0.934 0.350393
poly(Whole_weight, degree)1    493.73179   32.71007  15.094  < 2e-16 ***
poly(Whole_weight, degree)2   -101.27764   20.26773  -4.997 6.13e-07 ***
poly(Whole_weight, degree)3    -22.68314   11.17637  -2.030 0.042483 *
poly(Shucked_weight, degree)1 -364.29818   16.13981 -22.571  < 2e-16 ***
poly(Shucked_weight, degree)2   88.47890   10.46618   8.454  < 2e-16 ***
poly(Shucked_weight, degree)3   -2.15764    6.69576  -0.322 0.747291
poly(Viscera_weight, degree)1  -83.39231   11.38250  -7.326 2.97e-13 ***
poly(Viscera_weight, degree)2   -0.27797    7.51210  -0.037 0.970485
poly(Viscera_weight, degree)3   20.73602    4.82249   4.300 1.76e-05 ***
poly(Shell_weight, degree)1     55.58379   15.52844   3.579 0.000349 ***
poly(Shell_weight, degree)2    -18.15832    9.82947  -1.847 0.064790 .
poly(Shell_weight, degree)3     14.66196    5.45287   2.689 0.007207 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.012 on 3238 degrees of freedom
Multiple R-squared:  0.5706,	Adjusted R-squared:  0.5678
F-statistic: 204.9 on 21 and 3238 DF,  p-value: < 2.2e-16
```

Fig 1.11. Model 4



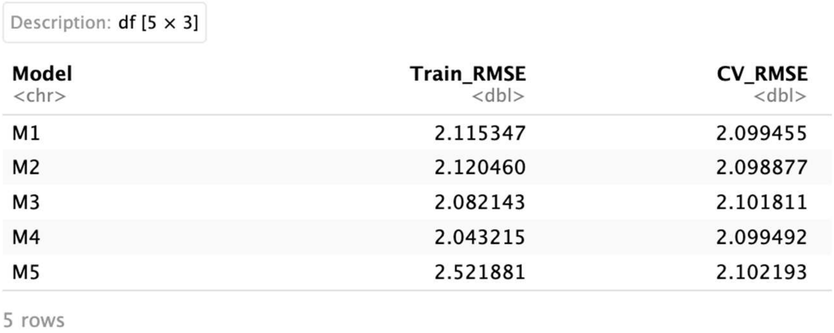
```
                                   GVIF Df GVIF^(1/(2*Df))
poly(Length, degree)        16806.59388  3        5.061120
poly(Diameter, degree)      16608.49595  3        5.051128
poly(Height, degree)           58.02814  3        1.967613
poly(Whole_weight, degree)  455357.98591 3        8.771196
poly(Shucked_weight, degree) 10956.64875 3        4.712807
poly(Viscera_weight, degree)  1466.07389 3        3.370487
poly(Shell_weight, degree)    4278.18752 3        4.029118
```

Fig 1.12. VIF



```
Call:
lm(formula = Rings ~ poly(Length, 1) + poly(Height, 1) + poly(Whole_weight,
    3) + poly(Shucked_weight, 2) + I(poly(Viscera_weight, 3)[,
    c(1, 3)]) + (poly(Shell_weight, 4)), data = trainData)

Residuals:
   Min     1Q Median     3Q    Max
-6.546 -1.252 -0.254  0.862 10.173

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          11.33528    0.03528 321.297  < 2e-16 ***
poly(Length, 1)                     -38.60493   10.37087  -3.722 0.000201 ***
poly(Height, 1)                      29.56159    5.56804   5.309 1.18e-07 ***
poly(Whole_weight, 3)1              488.28585   30.72077  15.894  < 2e-16 ***
poly(Whole_weight, 3)2              -90.59800   13.12053  -6.905 6.01e-12 ***
poly(Whole_weight, 3)3              -27.49022    5.88720  -4.669 3.14e-06 ***
poly(Shucked_weight, 2)1           -366.93414   15.26710 -24.034  < 2e-16 ***
poly(Shucked_weight, 2)2             87.89830    8.50667  10.333  < 2e-16 ***
I(poly(Viscera_weight, 3)[, c(1, 3)])1  -84.41190    9.86767  -8.554  < 2e-16 ***
I(poly(Viscera_weight, 3)[, c(1, 3)])3   20.42281    4.04500   5.049 4.69e-07 ***
poly(Shell_weight, 4)1               59.95198   14.92652   4.016 6.04e-05 ***
poly(Shell_weight, 4)2              -25.47769    8.32518  -3.060 0.002229 **
poly(Shell_weight, 4)3               23.99717    4.49294   5.341 9.88e-08 ***
poly(Shell_weight, 4)4               -8.91092    2.19536  -4.059 5.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.014 on 3246 degrees of freedom
Multiple R-squared:  0.5686,	Adjusted R-squared:  0.5668
F-statistic: 329.1 on 13 and 3246 DF,  p-value: < 2.2e-16
```

Fig 1.13 model5



Description: df [5 × 3]

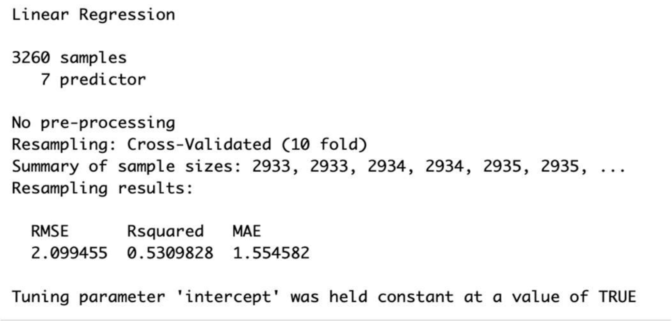
Model <chr>	Train_RMSE <dbl>	CV_RMSE <dbl>
M1	2.115347	2.099455
M2	2.120460	2.098877
M3	2.082143	2.101811
M4	2.043215	2.099492
M5	2.521881	2.102193

5 rows

Fig 1.14 Cross validation result.

The above is the combined table of the MSE variables. Based on the model 5 output, the adjusted R-squared value is 0.5668, model 4 is 0.57 and RSE values are also better for model4. But model4 have more insignificant variables and AIC is better for model5. So assume model4 is better indicating that the model explains about 57% of the variance in the Rings Variable based on the independent variable.

Subsequently, additional analysis was conducted using plots, specifically figures 1.15 and 1.16, for all the best models under consideration. The results were then compared using cross-validation. Based on this evaluation, it was determined that model5 outperformed the other models and is therefore considered the best model for the given scenario.



```
Linear Regression

3260 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2933, 2933, 2934, 2934, 2935, 2935, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.099455  0.5309828  1.554582

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Fig 1.15 Cross validation result.

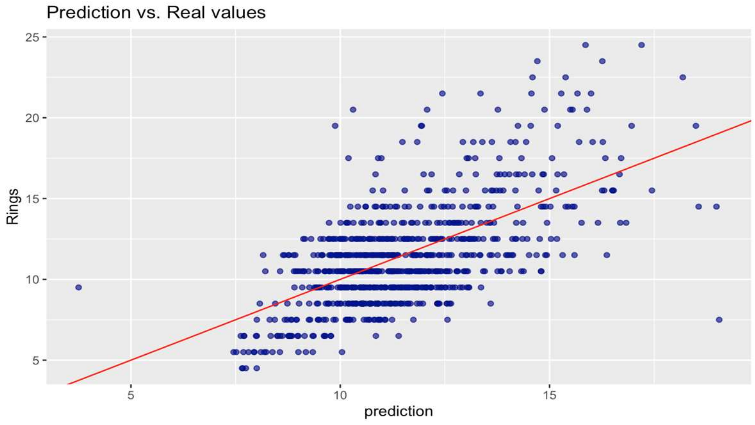


Fig 1.16. Prediction vs real values

The visualization below provides a means to assess the performance of the linear regression model. The proximity of the data points to the red line indicates the quality of the model, with closer alignment indicating a better fit. Based on the visualization, it can be concluded that the model is performing well and yielding satisfactory results.

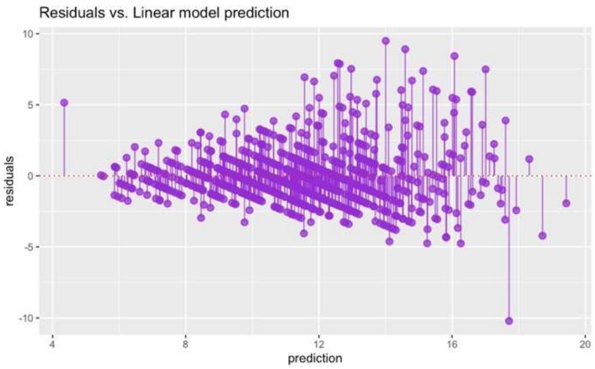
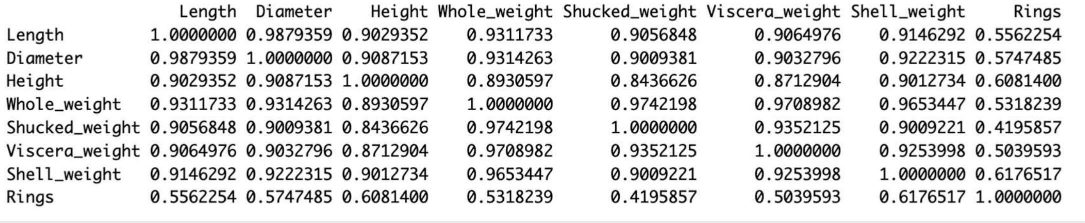


Fig 1.17. Prediction vs real values

The preceding plot illustrates the residuals, which represent the discrepancies between the actual and predicted values, plotted against the predictions generated by the linear model. A more scattered distribution of the residuals is indicative of a better-performing model. Based on this observation, it can be concluded that the model demonstrates satisfactory performance. Given the presence of multicollinearity, as evidenced by the VIF numbers and correlation plots, we will employ Principal Component Analysis (PCA) to address this issue.



```
                 Length  Diameter    Height Whole_weight Shucked_weight Viscera_weight Shell_weight     Rings
Length        1.0000000 0.9879359 0.9029352    0.9311733      0.9056848      0.9064976    0.9146292 0.5562254
Diameter      0.9879359 1.0000000 0.9087153    0.9314263      0.9009381      0.9032796    0.9222315 0.5747485
Height        0.9029352 0.9087153 1.0000000    0.8930597      0.8436626      0.8712904    0.9012734 0.6081400
Whole_weight  0.9311733 0.9314263 0.8930597    1.0000000      0.9742198      0.9708982    0.9653447 0.5318239
Shucked_weight 0.9056848 0.9009381 0.8436626    0.9742198      1.0000000      0.9352125    0.9009221 0.4195857
Viscera_weight 0.9064976 0.9032796 0.8712904    0.9708982      0.9352125      1.0000000    0.9253998 0.5039593
Shell_weight  0.9146292 0.9222315 0.9012734    0.9653447      0.9009221      0.9253998    1.0000000 0.6176517
Rings         0.5562254 0.5747485 0.6081400    0.5318239      0.4195857      0.5039593    0.6176517 1.0000000
```

Fig 1.18. Correlation matrix.

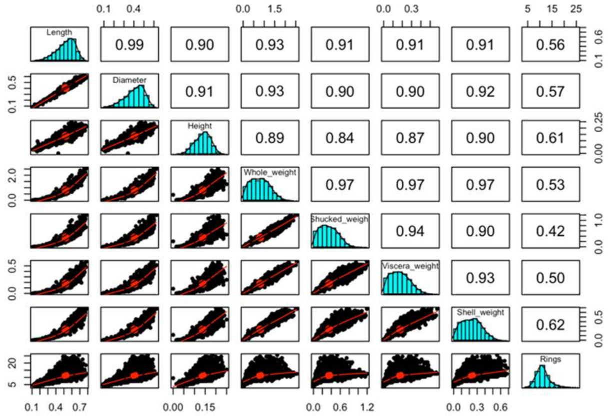


Fig 1.19 Pairs panels for abalone

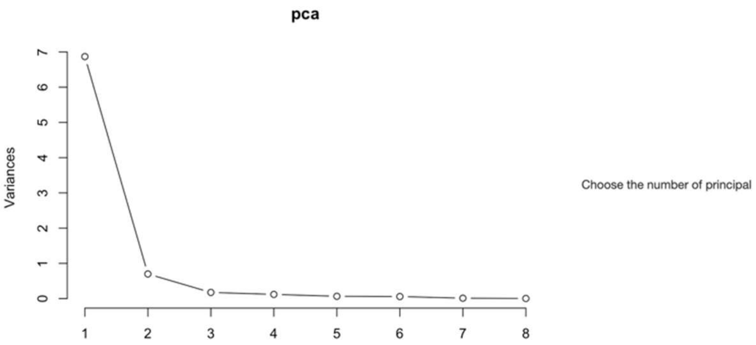
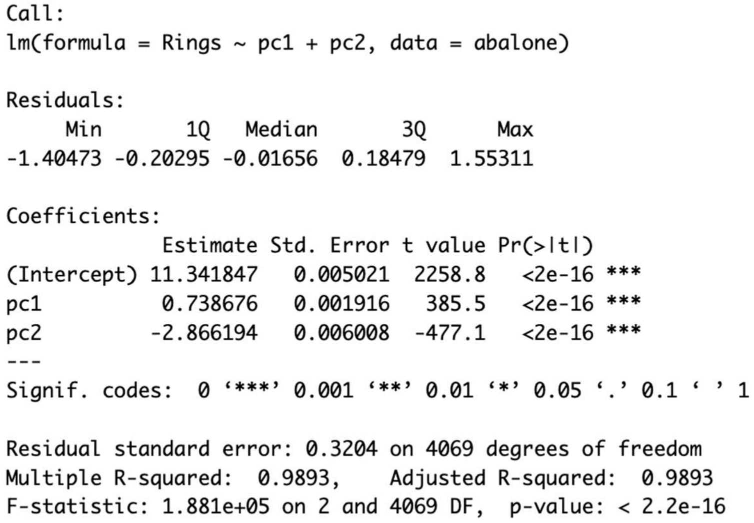


Fig 1.20. Plot to visualize the variance.



```
Call:
lm(formula = Rings ~ pc1 + pc2, data = abalone)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.40473 -0.20295 -0.01656  0.18479  1.55311 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.341847   0.005021  2258.8   <2e-16 ***
pc1          0.738676   0.001916   385.5   <2e-16 ***
pc2         -2.866194   0.006008  -477.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3204 on 4069 degrees of freedom
Multiple R-squared:  0.9893,	Adjusted R-squared:  0.9893 
F-statistic: 1.881e+05 on 2 and 4069 DF,  p-value: < 2.2e-16
```

Fig 1.21 Plot to visualize the variance.

After examining the PCA values and referring to the plot, we have chosen the first two principal components (PCs) for our model. Using these components, we constructed a new model as described previously. The model demonstrates a remarkable 98 percent accuracy in predicting the values of the predictor variables.

Now let's develop a regression tree with rings as response variable.

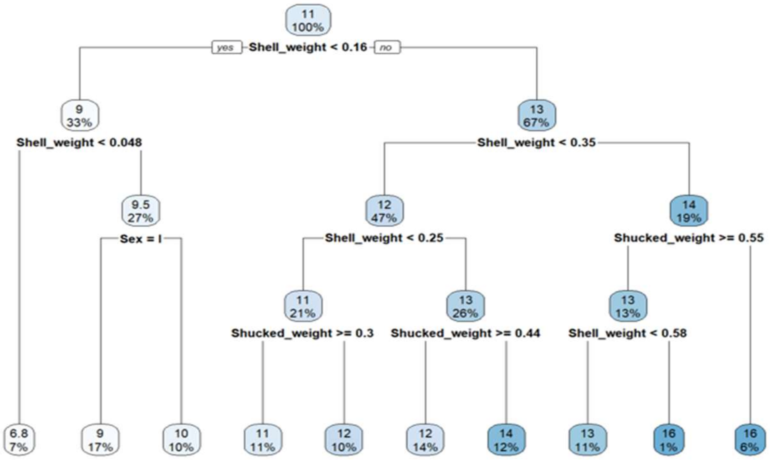


Fig 1.22 Regression tree

The variable importance plot from the random forest model reveals that "Shell_weight" is the most influential predictor for determining the number of rings in abalone. "Diameter" and "Height" also have notable importance. "Length" and "Whole_weight" contribute moderately. "Viscera_weight," "Shucked_weight," and "Sex" have relatively low importance. The plot provides insights into the significant predictors for predicting abalone rings.

Solution for Q2: To address the research question, we have introduced a new variable called "age group" and conducted a chi-square test to examine the relationship between the age group and sex variables. In order to ensure equal representation, we modified the dataset to have an equal number of rows for each gender. Furthermore, we assessed the relationship between age and sex using a box plot and observed that there is minimal disparity between males and females, as illustrated below.

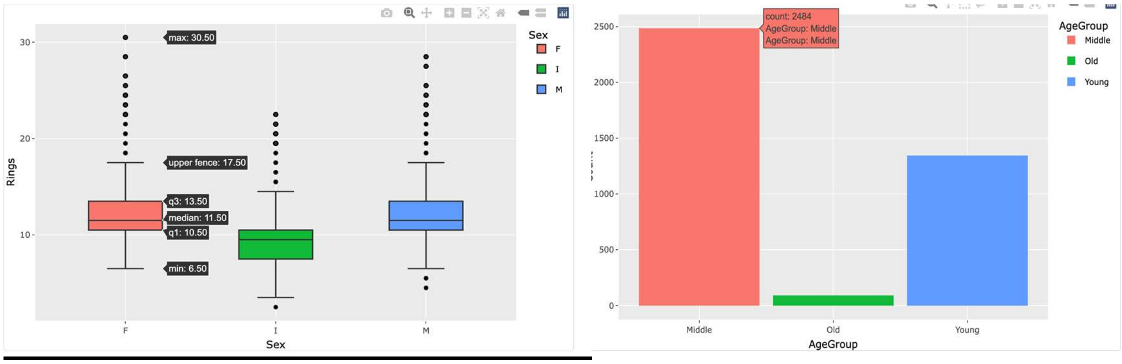
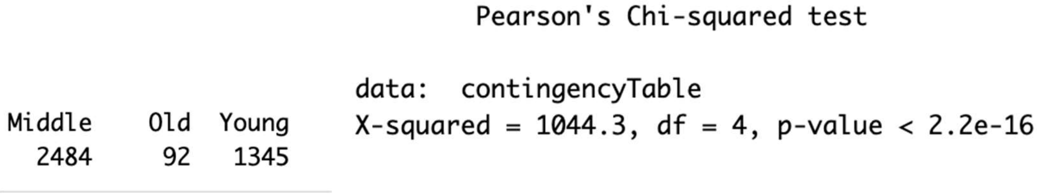


Fig 2.1 (a) Box plot: Age and sex (b) Age Group Bar graph

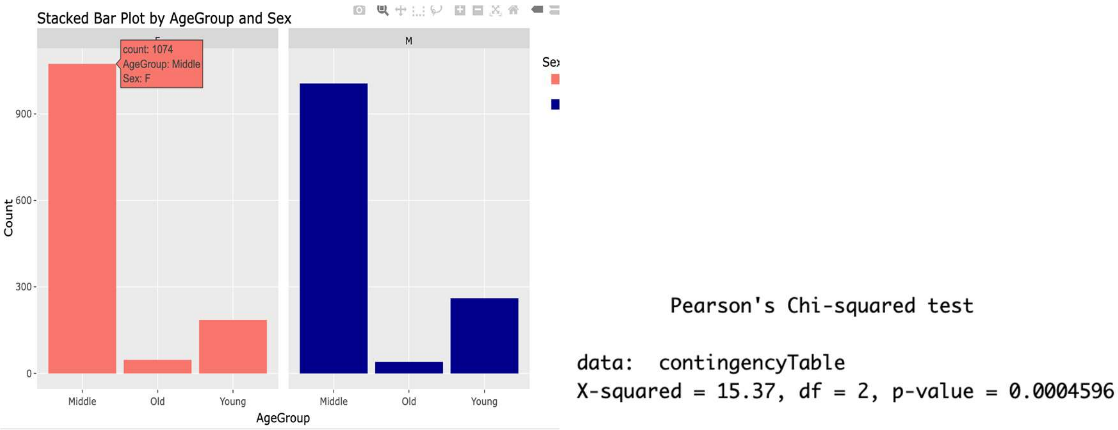
Subsequently, we created a new column called "Age" that consists of three distinct age groups. We then generated a count plot to visualize the distribution of observations within each age group, as depicted below.



```
                                          Pearson's Chi-squared test

                                 data:  contingencyTable
Middle    Old  Young             X-squared = 1044.3, df = 4, p-value < 2.2e-16
  2484     92   1345
```

Fig 2.2 Number of observations and chi-squared test

Following that, we conducted a chi-squared test, and the results indicated statistical significance. However, since the presence of the "infant" variable may introduce bias, we decided to exclude it and perform the test again. The outcomes of both tests will now be presented for examination.



```
        Pearson's Chi-squared test

data:  contingencyTable
X-squared = 15.37, df = 2, p-value = 0.0004596
```

Fig 2.3 (a) Bar plot by agegroup and sex (b) Chi-squared test

The p-value obtained from the Chi-Square Test is less than 0.05, indicating statistical significance. However, when compared to the previous model, the p-value is significantly smaller, and the effect size (squared value) is stronger in the previous model. Therefore, it can be concluded that the association between Age Group and Sex is relatively weak and less statistically significant. Additionally, the calculated Cramer's V value is 0.07668083, which further supports the notion of a weak association between the two variables. Overall, the results suggest that the inclusion of the "infant" variable in the chi-square model improves its performance, but it introduces bias due to the inherent differences in age among infants. Consequently, when considering the relationship between male and female, there is not a significant association between Age Group and Sex.

Solution for Q3

K-Prototypes clustering on a mixed dataset containing numerical and categorical variables. Sex is a categorical variable and other attributes are numerical. The numerical variables are first normalized using the scale function, and then the normalized data is combined with the categorical variables using the cbind function. K-Prototypes clustering is performed using the kproto function from the clustMixType package, specifying the number of clusters, random starts, lambda parameter, and maximum iterations. The resulting cluster labels and centers are extracted, and a cluster plot is created using the clusplot function with customized appearance parameters. The plot shows the clusters in a scatter plot, with color-coded points representing each observation's assigned cluster.

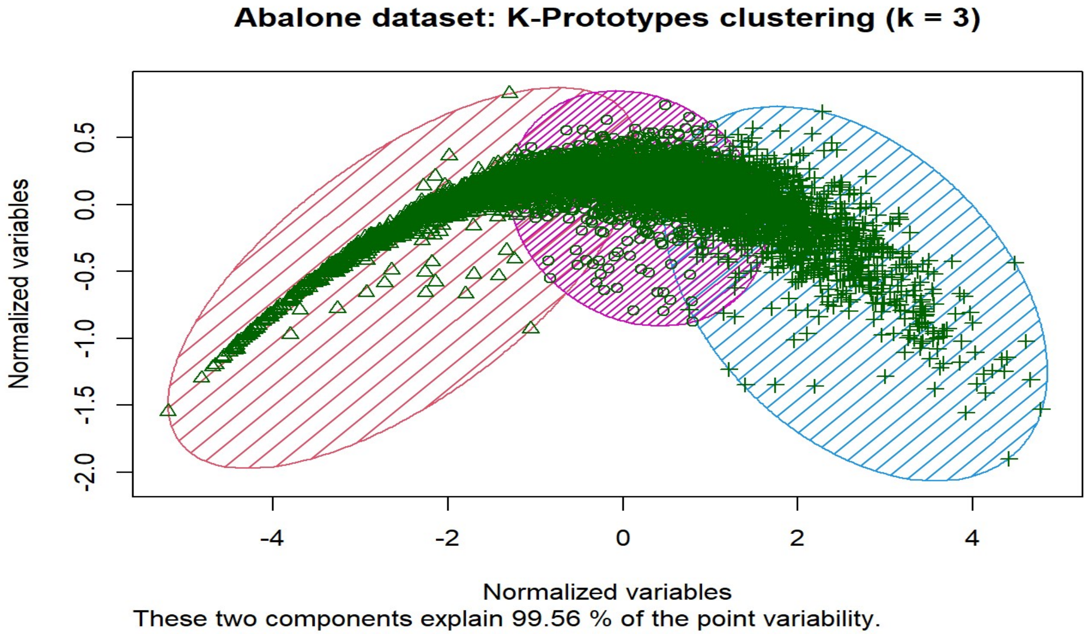


Fig 3.1 K-prototypes clustering (k=3)

CONCLUSION:

We carefully analyzed the abalone dataset using a variety of statistical methods, such as ANOVA, Chi-square tests, decision trees, random forests, regression analysis, mean square error evaluation, clustering, and regression analysis. With the help of these techniques, we addressed important issues like the ability to predict an abalone's age from physical measurements, the relationship between age and sex, the classification of an abalone's sex based on physical characteristics, and the separation of different abalone groups based on physical traits. Our results show a significant relationship between age and sex and that it is possible to determine an abalone's age from its morphological characteristics. Moreover, we successfully classified abalone sex based on physical attributes. These insights enhance our understanding of abalone characteristics and have practical implications for population management and resource allocation.

REFERENCES:

[1] *Abalone dataset*. UCI Machine Learning Repository: Abalone Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/abalone