UNSW SYDNEY

Data Analysis Report

By.

Melvin Tjandradjaja

z5378817

ZZSC5855-Multivariate Analysis for Data Scientists

Lecture Time: Wednesday – 17.00 – 19.00

Lecturer's Name: Dr. Pavel Krivitsky

Date: 24 October 2022

1. Introduction

The motivation behind this project was to fulfill our client's 2 requirements in regard to abalone harvesting. The first requirement was to incorporate a statistical model that could predict an abalone's gender based on its length, diameter, and height measurement (obtained from their new generation calipers) into their new generation goggles. The second requirement was to develop an algorithm that could predict an abalone's shucked and viscera weight (based on its length, diameter, and height). Then use these predicted weights to determine the abalone's estimated dollar value (given the day's shucked and viscera weight dollar value per gram) and calculate the weights' prediction interval 90 percent of the time.

The steps taken to fulfill the client's requirements were as follows;

1. Cleaning the abalone data
2. Analyzing data points segregation and shape
3. Statistical model fittings for gender classification and prediction
4. Analyzing the models' accuracies
5. Selecting the model with the highest gender prediction accuracy
6. Statistical model fittings for shucked and viscera weight (interior measurements) prediction
7. Analyzing the models' prediction errors
8. Including the models into the algorithm to determine an abalone's estimated dollar value and calculate its interior measurements' prediction interval

Section two, three, and four of this report will explain these steps in greater details.

2. Cleaning the Dataset

Our data cleaning process involved discovery and removal of invalid data entries and outliers.

3. Analyzing Data Points Segregation and Shape

Before proceeding to model fitting and prediction analysis for an abalone gender classification, we decided to perform the following tasks;

- Investigate data points segregation
- Observe the average difference on the abalone populations

Figure 1.1, 1.2, and 1.3 are plots we generated to analyze the segregation and they focused on length, diameter, and height measurement (exterior measurements) data points.

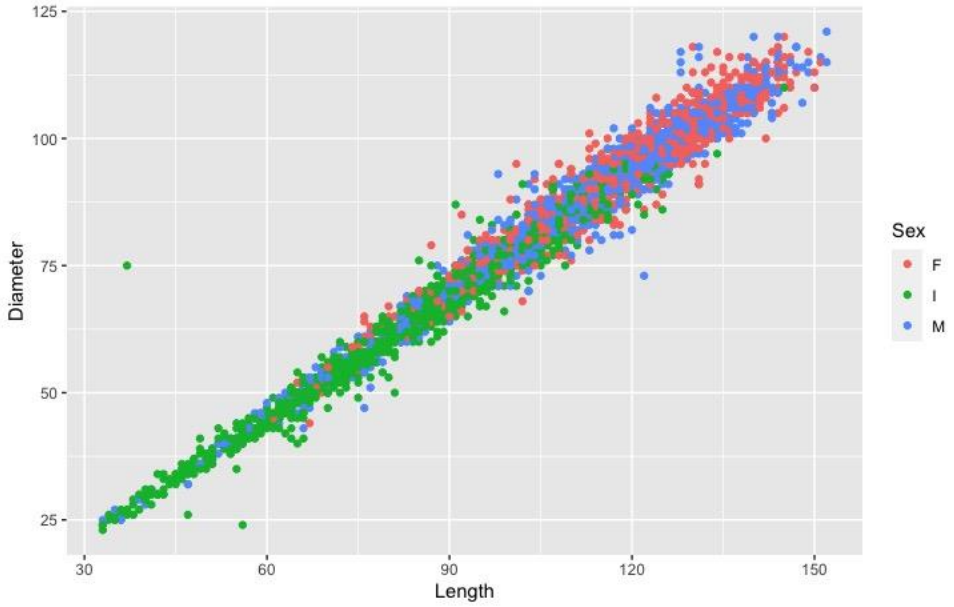

Fig. 1.1 Abalone Length vs. Diameter

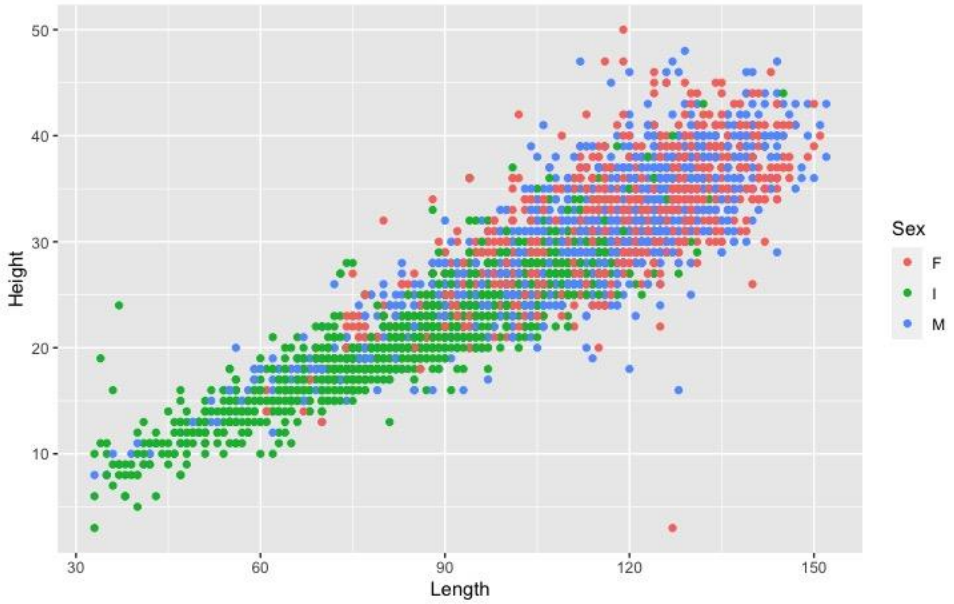

Fig. 1.2 Abalone Length vs. Height

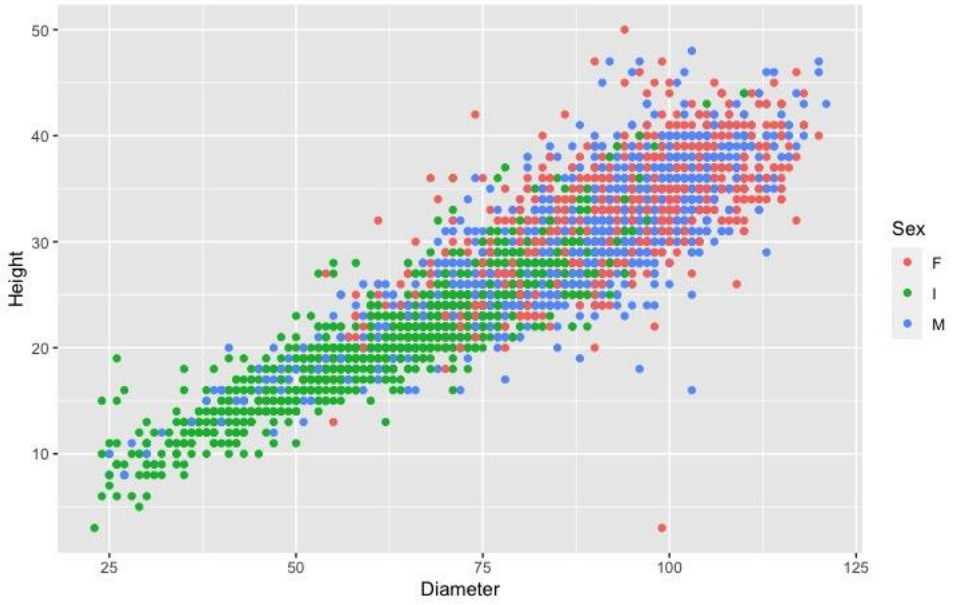

Fig. 1.3 Abalone Diameter vs. Height

The plots seem to show low segregation between the abalone populations.

Based on these observations, we proceeded in performing tests to observe the differences between the populations' average exterior measurements.

Our tests indicated the smallest differences to occur between female and male abalones' exterior measurements.

TABLE I shows the average female and male average exterior measurements and their differences.

TABLE I. AVERAGE EXTERIOR MEASUREMENTS AND DIFFERENCES

	Length	**Diameter**	**Height**
Mean Female	115.08	90.33	31.13
Mean Male	111.24	86.99	29.86
Difference	3.84	3.34	1.27

These very small differences could potentially lead to low accuracies in terms of predicting an abalone's gender based on new exterior measurements data.

Before proceeding to model fitting and prediction analysis for an abalone's interior measurements, we decided to observe the following relationships;

- Shucked weight and length
- Shucked weight and diameter
- Shucked weight and height
- Viscera weight and length
- Viscera weight and diameter
- Viscera weight and height

Figure 1.4, 1.5, 1.6, 1.7, 1.8, and 1.9 are plots we generated to analyze the data shape in order to determine the relationships.

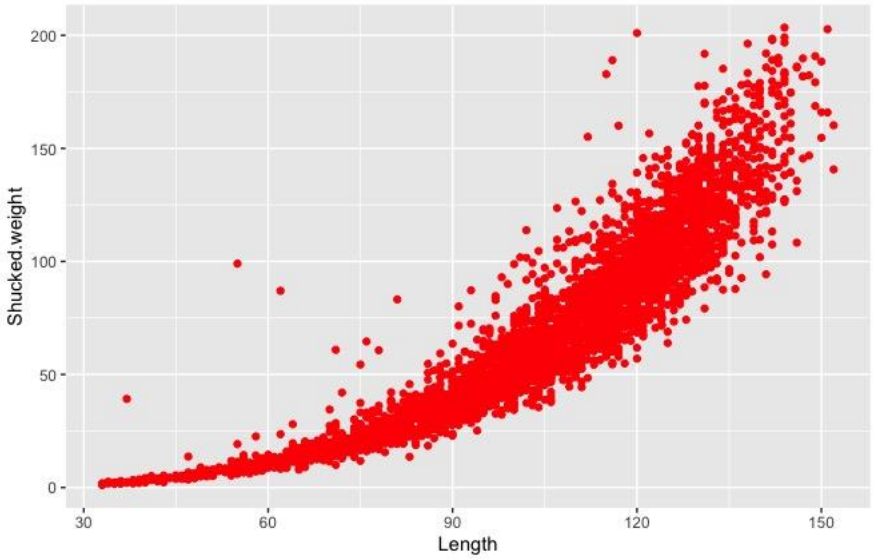

Fig. 1.4 Abalone Length vs. Shucked Weight

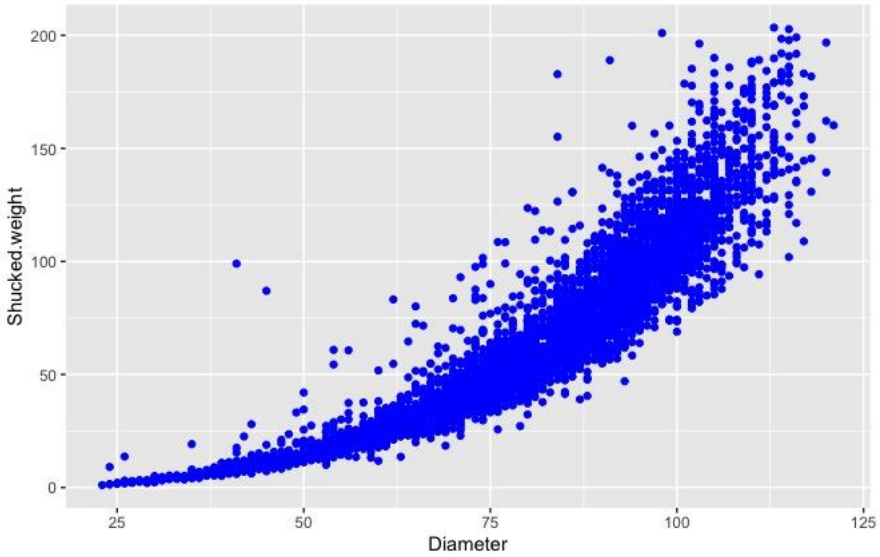

Fig. 1.5 Abalone Diameter vs. Shucked Weight

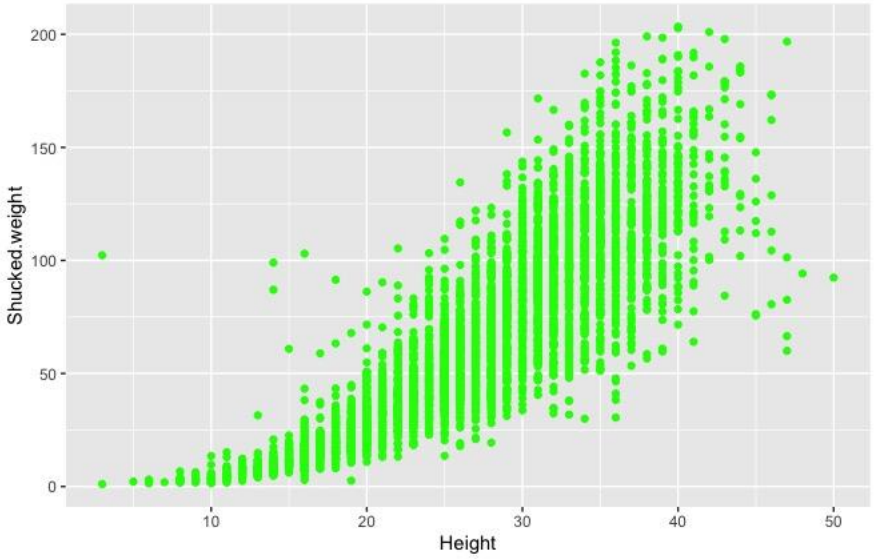

Fig. 1.6 Abalone Height vs. Shucked Weight

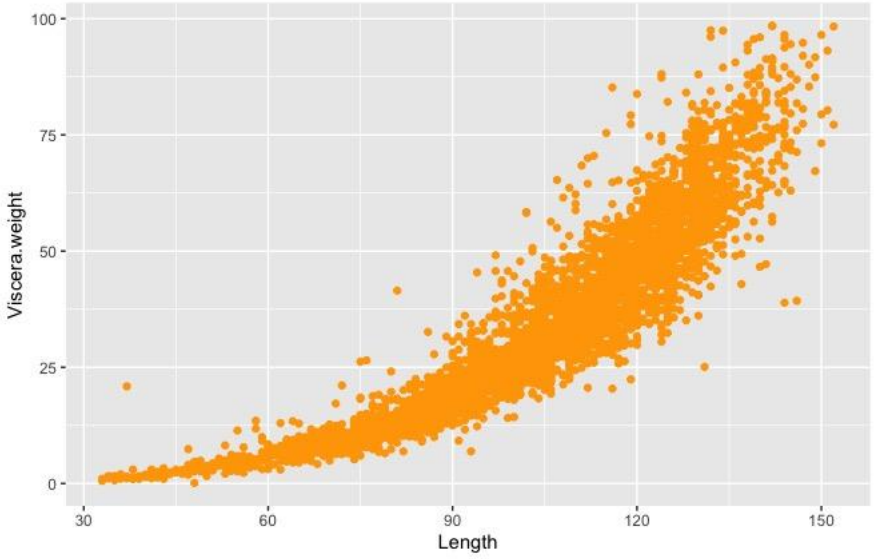

Fig. 1.7 Abalone Length vs. Viscera Weight

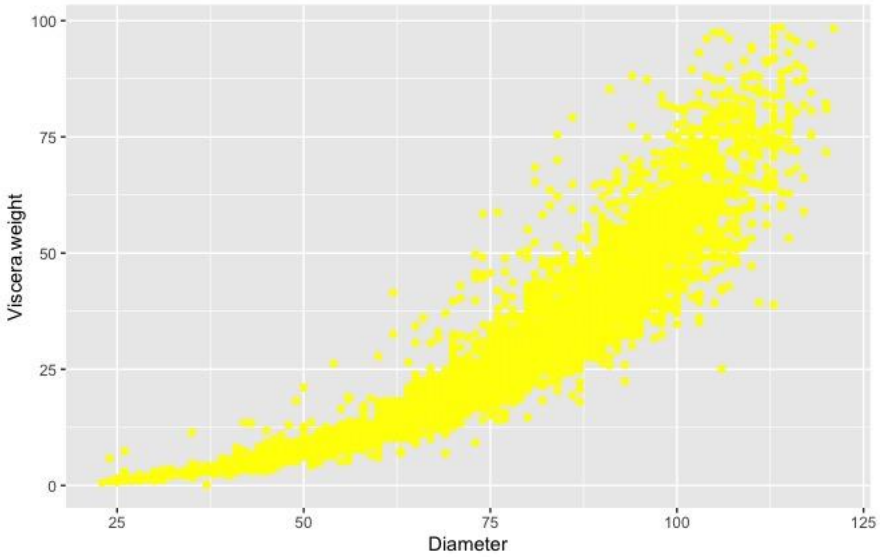

Fig. 1.8. Abalone Diameter vs. Viscera Weight

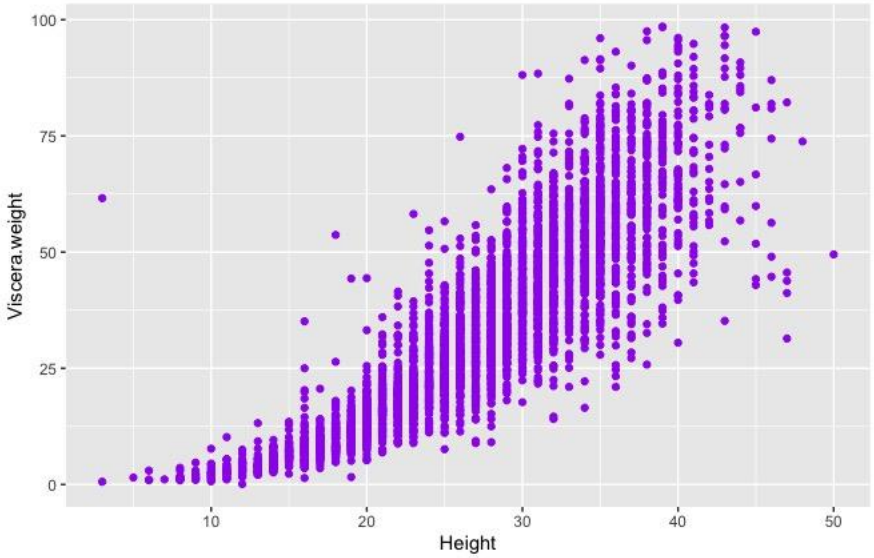

Fig. 1.8. Abalone Height vs. Viscera Weight

The upward curvy data shape on all the plots seems to indicate a logarithmic relationship between the interior and exterior measurements.

Based on these observations, we proceeded with using logarithm on the statistical models required to develop the algorithm to estimate an abalone's value and calculate its interior measurements' prediction interval.

4. Model Fittings and Gender Prediction Accuracies

Before selecting an appropriate model for gender prediction, we ran a test to see if the abalone populations were multivariate normal. Since the test result indicated that they are not, we selected Support Vector Machine (SVM) to handle the gender classification task.

We fitted SVM with Sigmoid, Linear, and Radial, obtained the models' prediction accuracy values, and populated TABLE II with them.

TABLE II. MODELS' PREDICTION ACCURACIES

Model	Accuracy(%)
SVM with Sigmoid	51.98
SVM with Linear	52.29
SVM with Radial	52.15

All the models involved seemed to have relatively low accuracies when predicting an abalone's gender given new exterior measurements data.

4. Model Fittings and Interior Measurements Observation

The model fitted for determining an abalone's estimated dollar value was Multivariate Linear Model (MLM) with logarithm. MLM with logarithm was selected because the relationship between the dependent variables (interior measurements) and the independent variables (exterior measurements) was logarithmic.

The models fitted for calculating an abalone's interior measurements prediction interval were Multiple Linear Regression (MLR) models with logarithm. These models were selected for the same reason as the MLM.

We conducted a prediction significance level test on the exterior measurements for the MLM and populated TABLE III with the level values.

TABLE III. EXTERIOR MEASUREMENTS' SIGNIFICANCE LEVELS

Variable	Shucked Weight Prediction Significance Level	Viscera Weight Prediction Significance Level
Length	3	3
Diameter	3	3
Height	3	3

The significant levels in descending order were 3, 2, and 1. Levels below 1 was considered insignificant.

As shown in TABLE III, the exterior measurements were at level 3 indicating the highest significance level in predicting the interior measurements.

After the MLM and the MLM models' fitting process were completed, we obtained the model's prediction error values and populated TABLE IV and TABLE V with them.

TABLE IV. MLM PREDICTION ERRORS

MLM	Prediction Error
Shucked Weight	0.0544
Viscera Weight	0.0553

TABLE V. MLR MODELS PREDICTION ERRORS

MLR	Prediction Error
Shucked Weight	0.0544
Viscera Weight	0.0553

As shown in TABLE IV and V, the MLM's and MLR model's prediction errors in predicting an abalone's interior measurements were low.

5. Conclusion

In this project, an abalone observation dataset was used to evaluate 3 SVM models in predicting an abalone's gender. The evaluation process was to determine their prediction accuracies with new observation data and select the model with the highest accuracy.

As mentioned in section 3, the prediction accuracy for each of the SVM models were relatively low. These occurrences were likely attributed to the fact that the models had to learn from data with low data points segregation and small differences between female and male average exterior measurements as mentioned in section 2. As a result, they poorly discern a pattern in the data to differentiate the abalones' genders.

We believed that the low prediction accuracies can be improved by perhaps adding an abalone characteristic information that absolutely differentiate between male, female, and infant abalones into the current dataset. Under the circumstances where the prediction accuracies were high, we would have recommended an SVM model with the highest accuracy value and lowest number of support vectors to be incorporated into the new generation diving equipment set.

The abalone observation dataset was also utilized to evaluate the MLM and MLR models in predicting an abalone's interior measurements. The evaluation process was to determine the exterior measurements' significance in predicting the interior measurements. The process also involved analyzing the MLM's and the MLR models' prediction errors.

As mentioned in section 4, the exterior measurements had the highest significant level which signifies a very high influence in the predictions. Section 4 also mentioned low prediction errors for the MLM and MLR models. This indicates the model were performing well in predicting the interior measurements. These findings justified incorporating the MLM and MLR models into the value estimator algorithms that calculates an abalone's estimated dollar value and its interior measurements' prediction interval.