

## CIS 2334 Semester Project

### Part 3

The marine biologists research team are satisfied with the excel application you have developed, which helped them greatly understand the abalone across the country. On top of the analysis, you have done in part 2, the scientists are keen to find some underlying patterns from the abalone data. In other words, the research team wants to build mathematical models for the abalone data, which reveal the fundamental relationships among the variables in the abalone data. As an expert in business analytics, you have the perfect skill set for this task.

To build a solid model, you need to go through the following steps and finalize your model in the end.

#### Task 1. Prepare the dataset

Firstly, you need to prepare the data for building the model. In classic data modeling tasks, you only use a portion of the data to train your model – this portion of the data is called the training set; the rest of the data are used to evaluate the performance of your model – this is called the test set.

What you need to do:

- Create a new excel file called “Firstname\_Lastname\_DataModeling.xlsx”.
- Name your current worksheet “Original Data”.
- Copy the data in your “Personal Data” worksheet from your semester project part 2 and paste the data set in the “Original Data” worksheet.
- Create a new worksheet called “Training set” and copy the first 2/3 of the data from the “Original data” and paste them here.
- Create a new worksheet called “Test set” and copy the rest of 1/3 of the data from the “Original data” and paste them here.

#### Task 2. Find relationships among variables in stacked data

Before modeling the data, you need to have a better understanding of the relationship among variables. The research team have specified a set of numerical variables that they care the most. They are listed in the table below. In particular, scientists are mostly interested in the rings of the abalone, since it tells the age of the abalone.

Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
--------	----------	--------	--------------	----------------	----------------	--------------	-------

What you need to do:

- Create a new worksheet called “Stacked data analysis”
- Use the “Training set”. Explore and create histograms for different variables listed above and then pick 3 most interesting histograms and describe the characteristics of each of them.

- c. Use the "Training set". Create a box plot for Shucked\_weight, Viscera\_weight and Shell\_weight and describe characteristic for each of the variable in the plot.
- d. Use the "Training set". Explore and create scatter plots for different variables listed above, then pick 5 most interesting scatter plots and describe the characteristics of each of them.
- e. Use the "Training set". Calculate the correlation between **every pair** of the variables listed above. Identify the top-5-strong correlated variables. Apply conditional formatting on your computed results that indicates top-5-strong correlations.
- f. Use scatter plots to demonstrate the strong correlated variables. Describe your findings.

### Task 3. Build regression models for stacked data

Since you have revealed the top-5-strong correlated variables, you need to build regression models that describe the data relationship mathematically.

What you need to do:

- a. Create a new worksheet called "Regressions for stacked data"
- b. Use the "Training set". Build a regression model for the variables that has the strongest correlation.
- c. Explicate your regression equations. Explain the coefficients, interceptions in your models.
- d. Use the "Test set". Compute the mean squared error for the regression model you have built.
- e. Use the "Training set". Build a regression model for the variables that has the fifth strongest correlation.
- f. Explicate your regression equations. Explain the coefficients, interceptions in your models.
- g. Use the "Test set". Compute the mean squared error for the regression model you have built.
- h. Compare the mean squared error between the two regression models. Describe your findings.

### Task 4. Create unstacked data

It is very important to look at the different genders separately and see if the relationships are different for different genders.

What you need to do:

- a. Create a new worksheet called "Unstacked Training set".
- b. Unstack the Training set, separating male, female, and infant.
- c. Create a new worksheet called "Unstacked Test set".
- d. Unstack the Test set, separating male, female, and infant.

### Task 5. Find relationships among variables in unstacked data

What you need to do:

- a. Create a new worksheet called "Unstacked data analysis".

- b. Use each gender's data in the "Unstacked Training set" to create the same set of histograms as in the task 2 step b.
- c. Compare the stacked data histograms against each gender's histogram. Is there any difference? If so, describe them.
- d. Use the whole\_weight variable of each gender in the "Unstacked Training set" and for all genders in the "Training set" to create a box plot (four boxes – whole\_weight for female, whole\_weight for male, whole\_weight for infant, and whole\_weight for all). Describe your findings from the plot.
- g. Use each gender's data in the "Unstacked Training set" to compute the same correlation matrix as in the task 2 step e.
- h. Compare the four correlation matrixes (one for each gender and one from the task 2 step e). Describe the value differences for the top-5-strong correlated variables identified in the task 2 step e.
- i. Use each gender's data in the "Unstacked Training set" to create the same set of scatter plots as in the task 2 step f.
- j. Compare the stacked data scatter plots against each gender's scatter plots. Is there any difference? If so, describe them.

## Task 6. Build regression models for unstacked data

Next, you need to build regression models on the unstacked data and compare them with the models with the stacked data.

What you need to do:

- a. Create a new worksheet called "Regressions for unstacked data"
- b. Build regression models on the same variables as in task 3 step b but use each gender's data in "Unstacked Training set".
- c. Explicate your regression equations. Explain the coefficients, interceptions in your models.
- d. Use the "Unstacked Test set". Compute the mean squared error for the regression models you have built.
- e. Build regression models on the same variables as in task 3 step e but use each gender's data in "Unstacked Training set".
- f. Explicate your regression equations. Explain the coefficients, interceptions in your models.
- g. Use the "Unstacked Test set". Compute the mean squared error for the regression models you have built.
- h. Compare the mean squared errors between the stacked data regression models and the unstacked data regression models. Describe your findings.

## Task 7. Build one-variable regression models for Abalone "Rings"

The rings on the abalone indicate it's age. The most interesting problem that the research team found is how to predict the abalone's age using the other measurements in the data. You believe you can build good regression models to do the prediction.

Now, you need to find the best predictor for abalone's age. This is a trial and error process.

What you need to do:

- a. Create a new worksheet called "Single variable regression for Rings"
- b. Explore different variables for regression models of "Rings". You could choose to use stacked data or unstacked data.
- c. Examine each models mean squared error. The smaller the errors are the better the prediction model you have.
- d. Decide the best regression model(s). Explicate your regression equations. Explain the coefficients, interceptions in your models. Report the mean squared error of the best model.

### Task 8. Build two-variable regression models for Abalone "Rings"

Now you want to only focus on the **stacked data**, but merely using one variable in the regression model for the "Rings" is not good enough. You decided to create a two-variable regression model. Instead of using the build-in regression analysis, you decided to use the solver to derive the regression formular.

What you need to do:

- a. Create a new worksheet called "Two variable regression for Rings"
- b. Normalize your "Training set" and "Test set" according to the course slides.
- c. Select two explanatory variables (independent variables).
- d. Explicit the general expression of your regression equation.
- e. Initialize your regression coefficients randomly.
- f. Compute the initial predictions of "Rings" for the "Test set".
- g. Compute the mean squared error for the "Test set".
- h. Minimize the mean squared error using the solver by changing the regression coefficients.
- i. Explore different two explanatory variables (independent variables) and redo the step d trough h.
- j. Compare each model's mean squared error. The smaller the errors are the better the prediction model you have.
- k. Decide the best two-variable regression model(s). Explicate your regression equations. Explain the coefficients, interceptions in your models. Report the mean squared error of the best model.