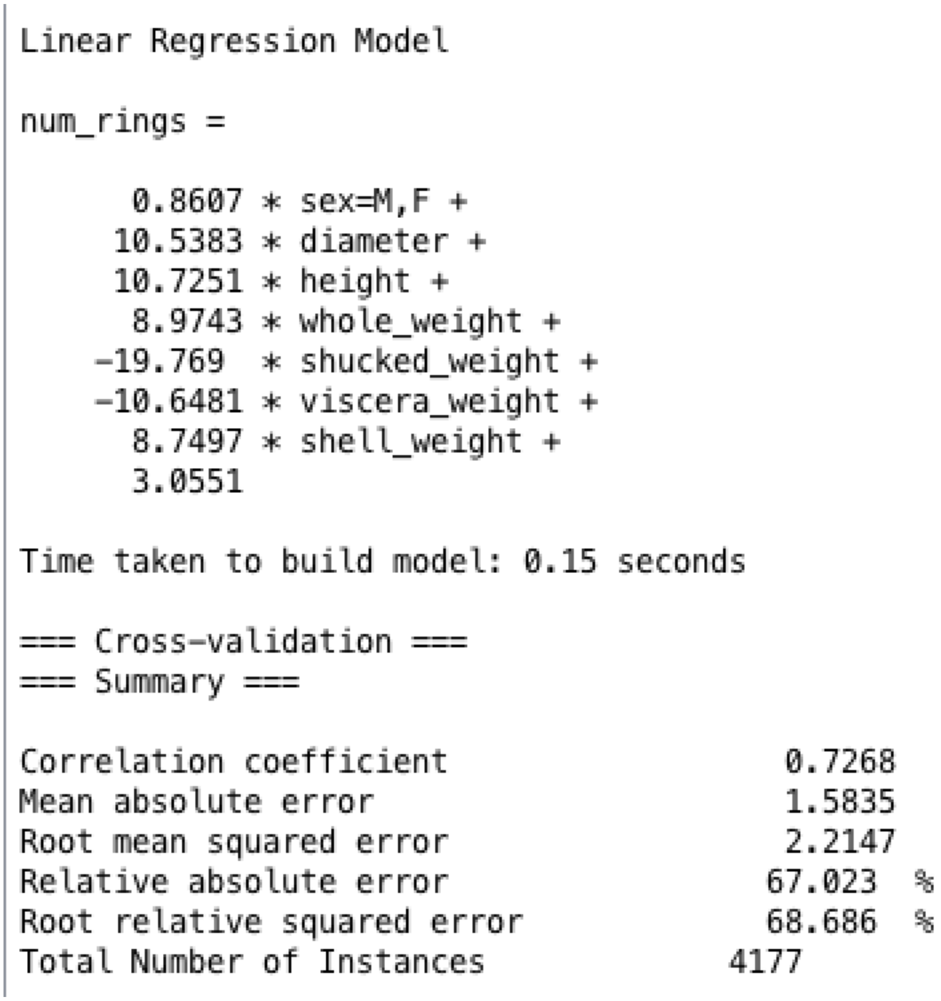
CSCI 585: HW5
USC ID:6386461387
Name: Prakhar Sethi
Tools Used: WEKA, KNIME & RAPIDMINER

Question 1: What is the MAE (mean absolute error)?

Solution: The Mean Absoluter Error of the given dataset is 1.5835

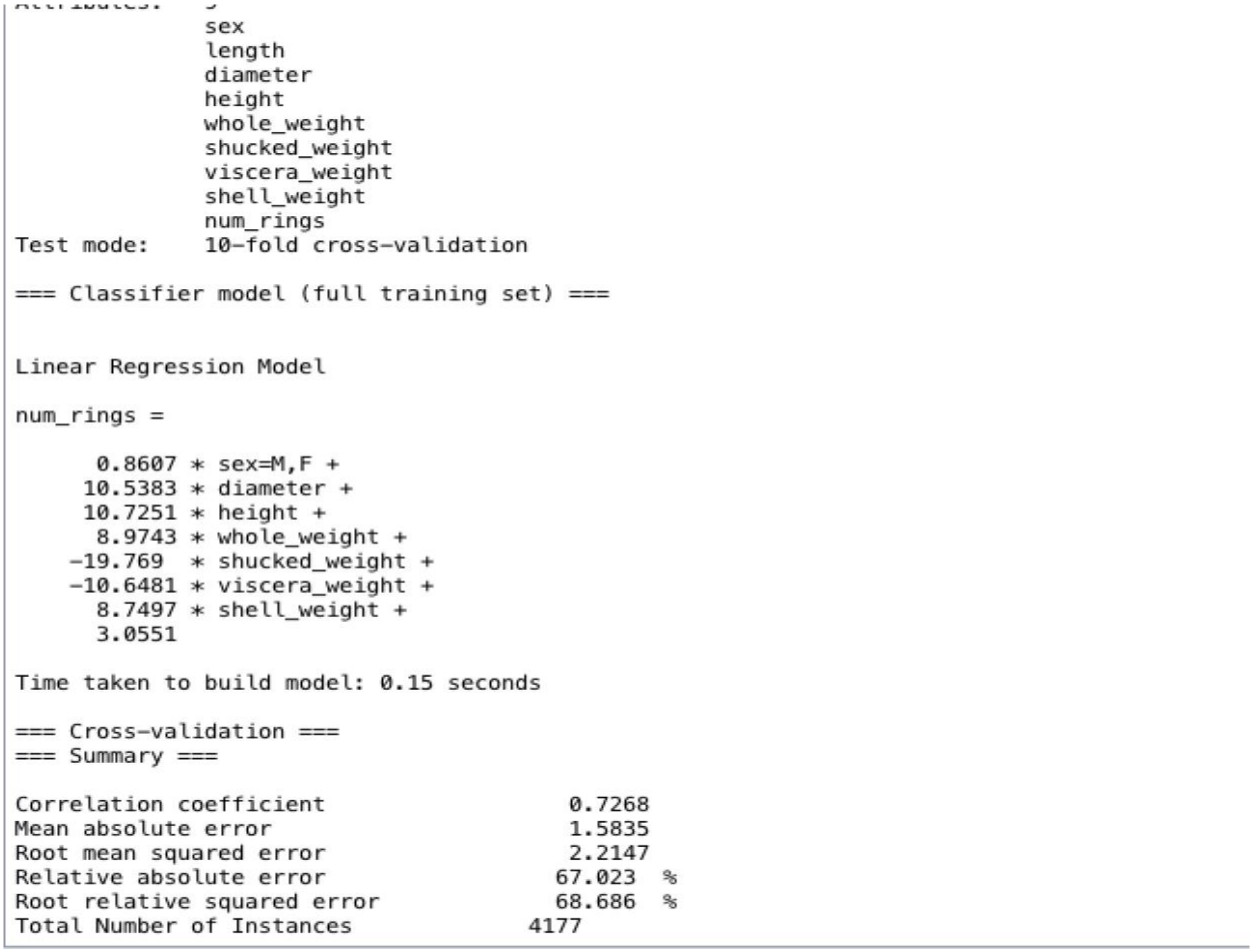


```
Linear Regression Model

num_rings =

      0.8607 * sex=M,F +
     10.5383 * diameter +
     10.7251 * height +
      8.9743 * whole_weight +
    -19.769  * shucked_weight +
    -10.6481 * viscera_weight +
      8.7497 * shell_weight +
      3.0551

Time taken to build model: 0.15 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7268
Mean absolute error                      1.5835
Root mean squared error                  2.2147
Relative absolute error                 67.023  %
Root relative squared error             68.686  %
Total Number of Instances             4177
```

Question 2: What is the equation for num_rings? You need to provide an equation, in the form of y=f(x_0,x_1,x_2...), using the output that WEKA provides. As you know, this is the result of the 'mining' - for a new shell, we can predict num_rings, by measuring the other params and inputting them into our equation.

Solution:
Equation:
num_rings= 0.8607 * sex=M,F + 10.5383 * diameter + 10.7251 * height +8.9743 * whole_weight + (-19.769) * shucked_weight +(-10.6481) * viscera_weight + 8.7497 * shell_weight + 3.0551

If we want to determine the “num_rings” based on the given data just “Plug in” the values and find it easily.



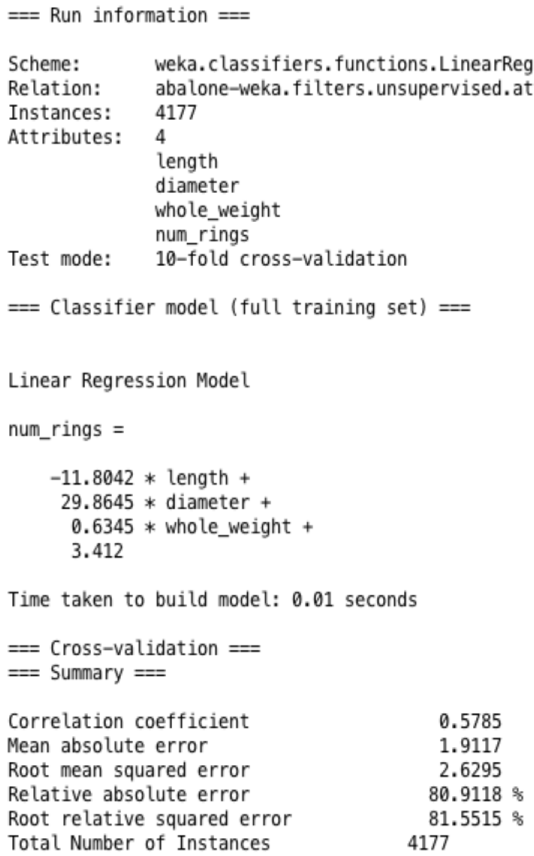
```
              sex
              length
              diameter
              height
              whole_weight
              shucked_weight
              viscera_weight
              shell_weight
              num_rings
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

num_rings =

      0.8607 * sex=M,F +
     10.5383 * diameter +
     10.7251 * height +
      8.9743 * whole_weight +
    -19.769  * shucked_weight +
    -10.6481 * viscera_weight +
      8.7497 * shell_weight +
      3.0551

Time taken to build model: 0.15 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7268
Mean absolute error                      1.5835
Root mean squared error                  2.2147
Relative absolute error                 67.023  %
Root relative squared error             68.686  %
Total Number of Instances             4177
```

Question 3: Re-open the dataset, and keep only these columns: length, diameter, whole_weight, num_rings. Do the linear regression again. What is the equation now? The idea is that now, we would only need do just 3 measurements, to predict num_rings.

Solution: The mean absolute error with above given column is 1.9117 and the equation is:

num_rings= (-11.8042) * length + 29.8645 * diameter + 0.6345 * whole_weight + 3.412

If we want to determine the "num_rings" based on the given data just "Plug in" the values of three columns and find it easily.



```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     abalone-weka.filters.unsupervised.attribute.Remove-R1,4,6-8
Instances:    4177
Attributes:   4
              length
              diameter
              whole_weight
              num_rings
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

num_rings =

    -11.8042 * length +
     29.8645 * diameter +
      0.6345 * whole_weight +
      3.412

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5785
Mean absolute error                      1.9117
Root mean squared error                  2.6295
Relative absolute error                 80.9118 %
Root relative squared error             81.5515 %
Total Number of Instances             4177
```

Question 4: What is the linear equation now in KNIME? Compare this to WEKA's output - what parameters have similar coefficients (where they differ by 0.5 atmost)?

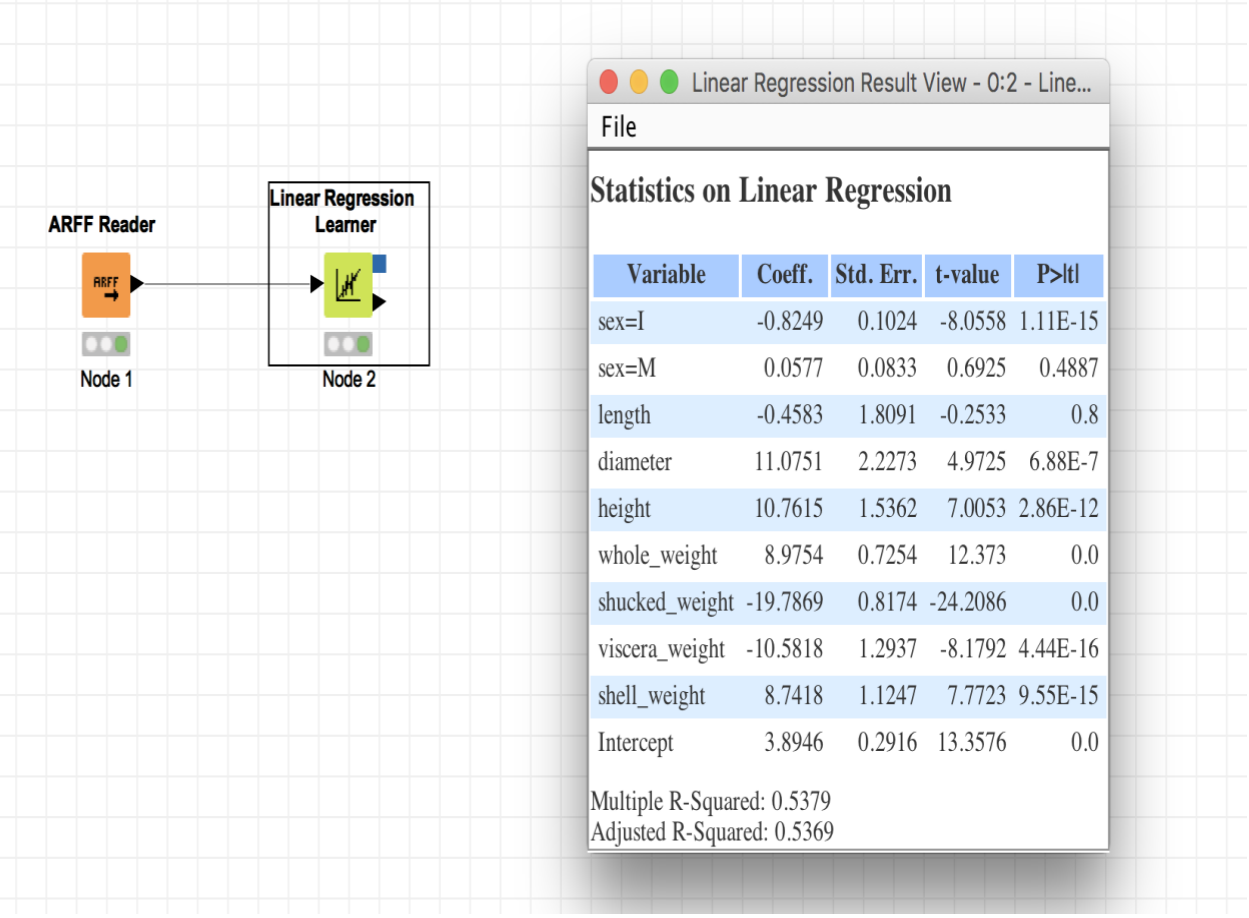
Solution:

num_rings: -0.8249* sex=I + 0.0577*sex=M+ (-0.4583)*length+ 11.0751 * diameter + 10.7615 * height +8.9754 * whole_weight + (-19.7869) * shucked_weight +(-10.5818) * viscera_weight + 8.7418 * shell_weight + 3.8946

Multiple R-Squared: 53.79% of the variance and quality judgements by this model with the other predictor variables.
Adjusted R-Squared: It is affected by the number of predictors compared to the number of test cases. Here, it is 53.69%

In comparison with WEKA output following parameters have similar coefficients (where they differ by 0.5 atmost):

Height
Whole_Weight
Shucked_Weight
Viscera_weight
Shell_weight


ARFF Reader
ARFF
Node 1
Linear Regression Learner
Node 2
Linear Regression Result View - 0:2 - Line...
File

Statistics on Linear Regression

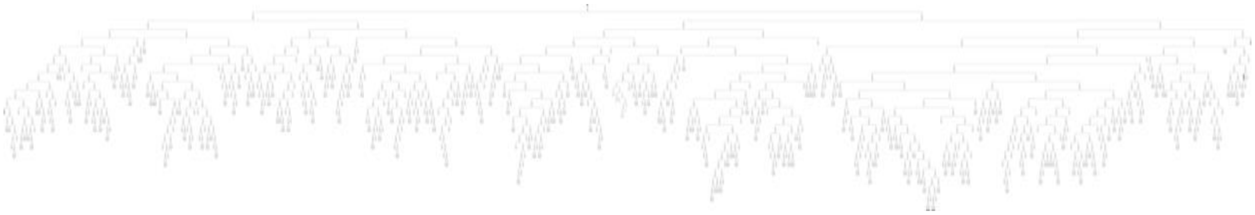
Variable	Coeff.	Std. Err.	t-value	P>|t|
sex=I	-0.8249	0.1024	-8.0558	1.11E-15
sex=M	0.0577	0.0833	0.6925	0.4887
length	-0.4583	1.8091	-0.2533	0.8
diameter	11.0751	2.2273	4.9725	6.88E-7
height	10.7615	1.5362	7.0053	2.86E-12
whole_weight	8.9754	0.7254	12.373	0.0
shucked_weight	-19.7869	0.8174	-24.2086	0.0
viscera_weight	-10.5818	1.2937	-8.1792	4.44E-16
shell_weight	8.7418	1.1247	7.7723	9.55E-15
Intercept	3.8946	0.2916	13.3576	0.0

Multiple R-Squared: 0.5379
Adjusted R-Squared: 0.5369

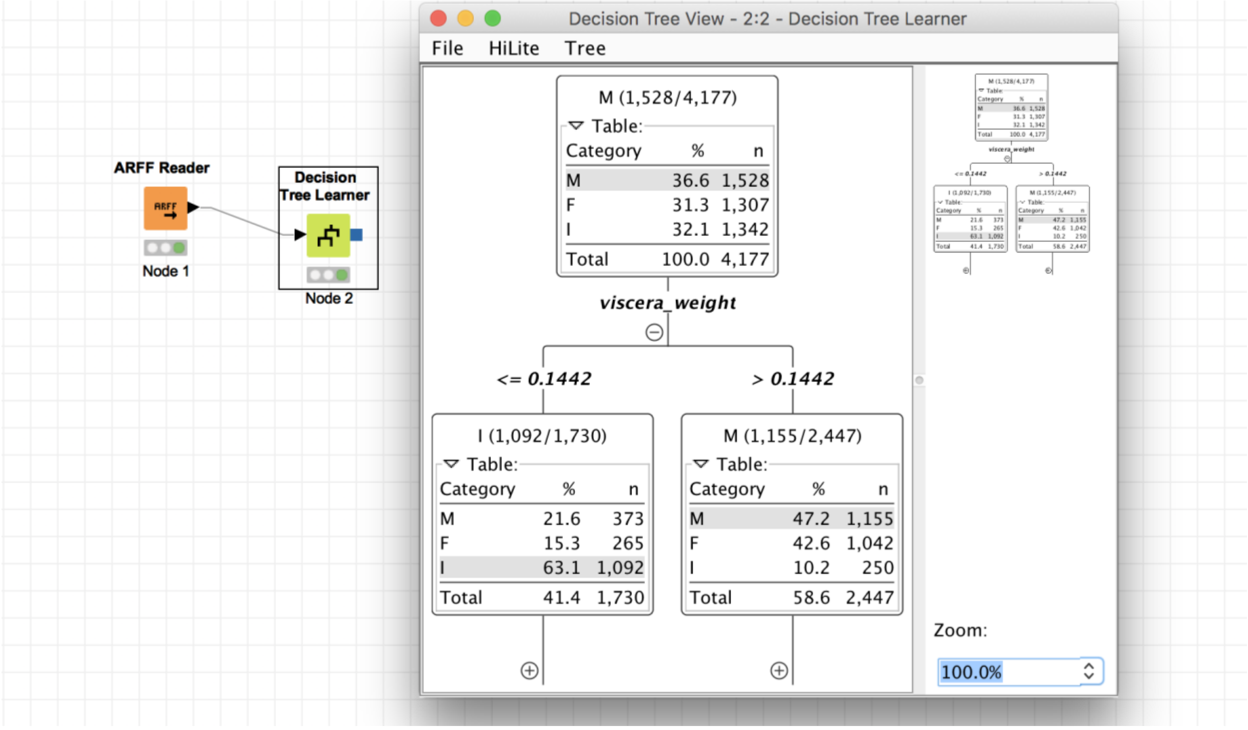
Question 5: Set up a 'Decision Tree Learner' predictor, where 'Sex' is the predicted variable. Note - think "simple" - no need to partition the data into training and test data, etc! Provide a snapshot (.jpg or .png) of the *entire* decision tree [OK if the nodes are too zoomed out and are therefore unreadable] - hint: look at the *right* side of the split-pane window.

Solution:

DecisionTreeLearner.png



Screenshot of the model and collapse version of tree



M (1,528/4,177)

Table:

Category	%	n
M	36.6	1,528
F	31.3	1,307
I	32.1	1,342
Total	100.0	4,177

I (1,092/1,730)

Table:

Category	%	n
M	21.6	373
F	15.3	265
I	63.1	1,092
Total	41.4	1,730

M (1,155/2,447)

Table:

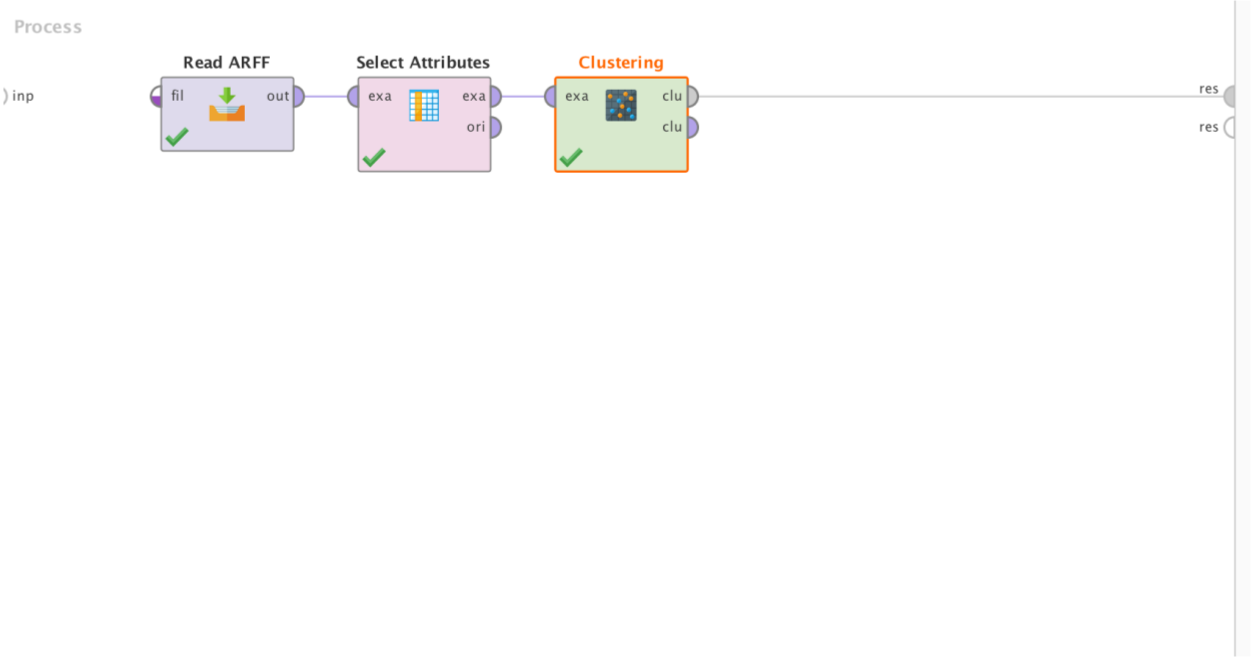
Category	%	n
M	47.2	1,155
F	42.6	1,042
I	10.2	250
Total	58.6	2,447

Bonus Question: a) Create 6 clusters out of the 4177 pieces of data (use a kMeans 'Clustering' node). Question: how many data points are in each cluster?

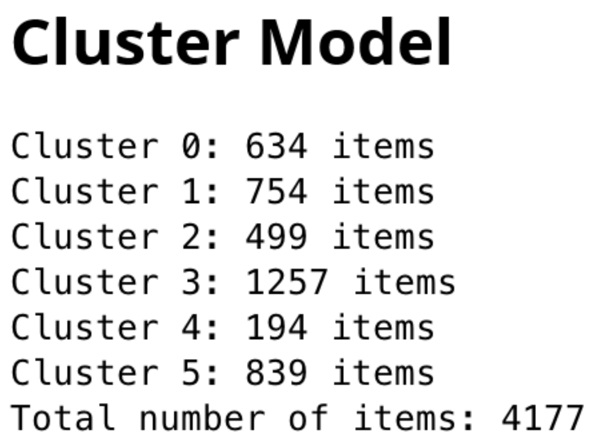
Solution:

Cluster 0: 634 items
Cluster 1: 754 items
Cluster 2: 499 items
Cluster 3: 1257 items
Cluster 4: 194 items
Cluster 5: 839 items
Total number of items: 4177

Screenshot of the model:



Screenshot of the output:



Cluster Model

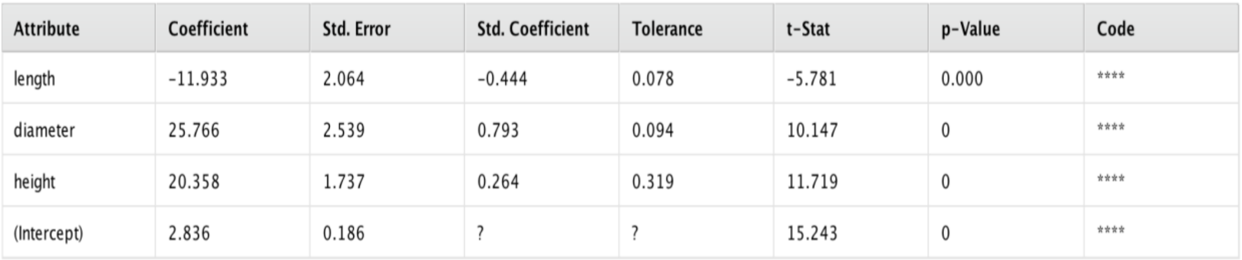
```
Cluster 0: 634 items
Cluster 1: 754 items
Cluster 2: 499 items
Cluster 3: 1257 items
Cluster 4: 194 items
Cluster 5: 839 items
Total number of items: 4177
```

Bonus Question: b) do a linear regression to predict num_rings, from length,diameter,height. Question: what is the equation?

Solution:
Equation:

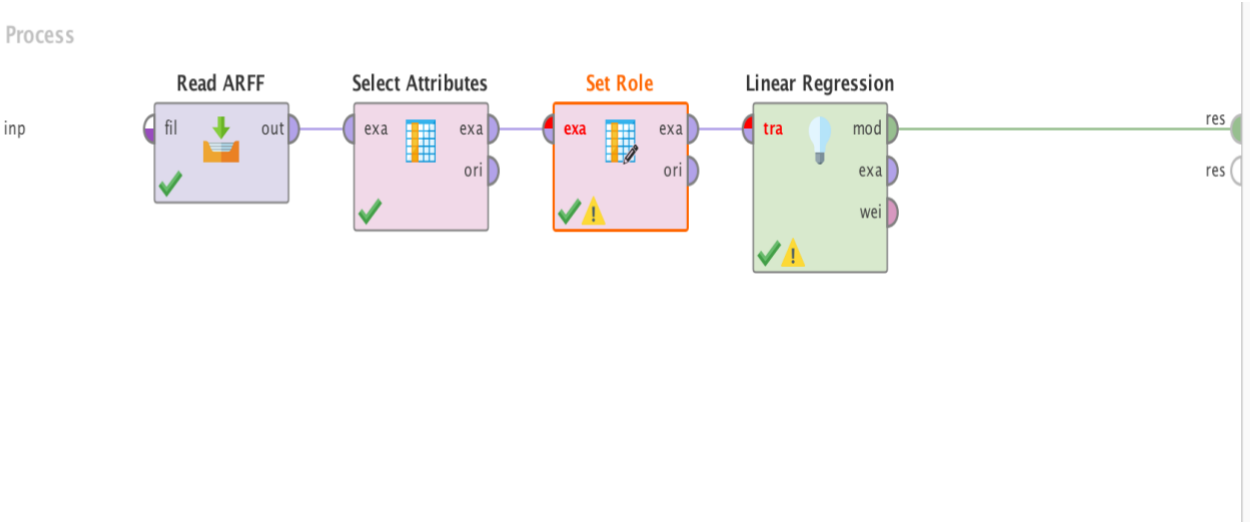
num_rings= (-11.933)*length+ 25.766 * diameter + 20.358 * height +2.836

Screenshot of Output:



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
length	-11.933	2.064	-0.444	0.078	-5.781	0.000	****
diameter	25.766	2.539	0.793	0.094	10.147	0	****
height	20.358	1.737	0.264	0.319	11.719	0	****
(Intercept)	2.836	0.186	?	?	15.243	0	****

Screenshot of model:


Process
inp
Read ARFF
fil
out
Select Attributes
exa
exa
ori
Set Role
exa
exa
ori
Linear Regression
tra
mod
exa
wei
res
res